MASTER EN SCIENCES INFORMATIQUES

L'extraction de données pour refléter les activités du marché immobilier namurois en temps réel et l'intégrer dans un système de support à la décision pour aider les collectivités locales à prendre des décisions informées

BARZIN, Félix

Award date:
2023

Awarding institution:
Universite de Namur

UNIVERSITÉ DE NAMUR
Faculté d'informatique
Année académique 2022–2023

L'extraction de données pour refléter les activités du marché immobilier namurois en temps réel et l'intégrer dans un système de support à la décision pour aider les collectivités locales à prendre des décisions informées.

BARZIN Félix

……...(Signature pour approbation du dépôt - REE art. 40)

Promoteur : VANHOOF Wim

Co-promoteur : YERNAUX Gonzague

Mémoire présenté en vue de l'obtention du grade de Master 60 en Sciences Informatiques

Faculté d'Informatique – Université de Namur
RUE GRANDGAGNAGE, 21 ● B-5000 NAMUR(BELGIUM)

Remerciements

Je souhaite exprimer ma gratitude envers les personnes qui ont contribué à la réalisation de ce mémoire. En particulier, je tiens à remercier Monsieur Gonzague Yernaux, doctorant à l'Université de Namur, pour son soutien, son écoute attentive, son expertise technique et ses relectures avisées. Depuis deux ans, son suivi a été une source de motivation.

Je tiens également à remercier les professeurs de l'Université de Namur et leurs équipes pédagogiques, qui m'ont fourni les outils nécessaires à la réussite de mes études universitaires.

Je remercie ma fiancée pour son soutien sans faille, ses encouragements et sa patience. Je la remercie chaleureusement aussi pour la relecture de ce travail. Elle a été une source d'inspiration et de motivation tout au long de cette aventure et je lui en suis infiniment reconnaissant.

Résumé

La récolte de données grâce aux méthodes de scraping est un procédé incontournable pour les études et la recherche en général. En effet, il permet de réaliser ce qu'aucune autre méthode ne propose : récolter des informations non structurées, de manière automatisée, en grand nombre, à un coût dérisoire et en temps réel. Nous verrons que la qualité des données peut parfois être questionnable, mais qu'il existe des moyens pour leur donner une certaine légitimité. Pour y parvenir, il est important de procéder à un traitement humain efficace tout au long du processus de récolte ainsi qu'une validation des résultats sur base d'une comparaison avec des sources conventionnelles et officielles. De plus, nous verrons que l'analyse exploratoire permet de comprendre les données, détecter les anomalies et les valeurs aberrantes.

Après certaines étapes d'analyse et de conception rigoureuse, nous serons en possession d'un jeu de données très complet et reflétant la réalité. Nous verrons alors qu'il existe de nombreuses applications pratiques comme, par exemple, la conception d'un système d'aide et de support à la décision utilisable et efficace.

Le scraping pose néanmoins des questions éthiques qui sont toujours en suspens. Il existe des éléments à prendre en considération afin de ne nuire à personne lorsqu'on souhaite procéder à une récolte de données. Nous aborderons cette thématique au cours de ce mémoire et évoquerons le besoin de définir un cadre légal explicite.

Mots-clés : web scraping, data mining, big data, traitement des données, analyse statistique, machine learning, temps réel, qualité des données, éthiques, système d'aide à la décision.

Abstract

Data scraping is an essential process for studies and research in general. In fact, it allows to achieve what no other method offers: collecting unstructured information in an automated way, in large numbers, at a low cost, and in real time. We will see that the quality of the data can sometimes be questionable, but there are ways to give them some legitimacy. To do this, it is important to proceed with effective human processing throughout the collection process, as well as validation of the results based on comparison with conventional and official sources. In addition, we will see that exploratory data analysis allows us to understand the data, detect anomalies and outliers.

After certain stages of rigorous analysis and design, we will have a very comprehensive dataset that reflects reality. We will then see that there are many practical applications, such as designing a usable and effective decision support system.

However, scraping raises ethical questions that are still unresolved. There are elements to consider avoiding harming anyone when collecting data. We will address this issue in this paper and discuss the need to define an explicit legal framework.

Keywords : web scraping, data mining, big data, data treatment, statistic analysis, machine learning, real-time data, data quality, human process, ethical questions, decision support system.

Table des matières

1. Introduction

1.1. Contexte

Les collectivités locales doivent avoir des données précises et à jour pour pouvoir prendre des décisions informées. Cela est particulièrement crucial pour le marché immobilier, qui est un secteur critique jouant un rôle vital dans la croissance et le développement des villes et régions. Le développement urbanistique a des effets sur le bien-être de la population. C'est pour cette raison que les preneurs de décisions ne devraient pas dépendre des méthodes traditionnelles de récolte de données pour agir (lentes, chères et biaisées).

Depuis quelques années, l'intérêt pour les techniques de minage de données est grandissant. On peut maintenant analyser des données en temps réel sur le marché immobilier et bien d'autres marchés. Les sources de ces données peuvent être variées (forums, réseaux sociaux, sites spécialisés, etc.) et la puissance du machine learning permet d'analyser un large volume de données. Ce qu'il nous manque, c'est un moyen de consulter, visualiser, comprendre et interpréter simplement les données et les résultats. C'est sur ce point que nous allons travailler.

Grâce aux techniques de minage et aux API open source, il est possible de construire un outil flexible, utilisable, évolutif et transparent. Cet outil doit permettre aux entités qui en ressentent le besoin d'avoir une base objective et non biaisée pour la prise de décision informée.

1.2. Motivations

Les données issues du marché immobilier sont rares, complexes et difficiles à analyser. En Belgique, les données relatives aux transactions immobilières sont fournies par le gouvernement et le bureau des notaires. Celles-ci sont filtrées, non contrôlables et diffusées quelques fois par an.

Les collectivités locales luttent pour garder un équilibre entre la demande pour un développement résidentiel, commercial et industriel. Face à des impératifs environnementaux et de qualité de vie, les décisions doivent se baser sur des éléments solides. Un système de support à la décision (SSD) flexible, utilisable, évolutif et transparent peut aider à prendre des décisions informées. Un système de support à la décision est un système informatique ayant pour but et utilité d'aider les décideurs à prendre des décisions en fournissant des informations, des analyses et des outils pour structurer et résoudre des problèmes complexes. Il peut inclure des outils d'analyse de données, des modèles de simulation, des systèmes experts et des interfaces graphiques conviviales pour permettre aux utilisateurs de visualiser les résultats et d'interagir avec les données de manière intuitive. Ces systèmes sont utilisés dans de nombreux domaines tels que la gestion d'entreprise, la finance, la santé, l'ingénierie et bien d'autres encore.

Les méthodes d'analyse traditionnelles du marché immobilier reposent sur des données statiques qui ont bien souvent un temps de retard, il est donc plus difficile de réagir en temps réel.

Les techniques de minage des données fournissent un aperçu des dynamiques réelles et cela permet aux collectivités de prendre des décisions qui s'alignent avec les tendances. C'est pourquoi il y a un besoin d'explorer comment ces données issues du minage peuvent être utilisées pour refléter les activités de marché en temps réel et servir de support pour la décision.

De nombreuses études axent leur réflexion autour du prix de vente. Beaucoup de recherches ont pour but de découvrir une sorte de « formule magique » permettant de prédire le prix de l'immobilier. Elles se concentrent également sur la recherche de la feature[1] la plus déterminante ou du modèle le plus précis.

Via ce mémoire, nous aimerions proposer une autre approche. Nous voudrions favoriser un développement durable et positionner l'humain, en tant qu'être, au centre des préoccupations. Afin de préserver un sentiment d'humanité dans la prise de décision, il est essentiel d'impliquer l'humain dans la réflexion durant le processus décisionnel. L'étude de Davidson-Hunt et al. montre qu'impliquer les communautés locales dans la prise de décision par rapport à la gestion des ressources naturelles permet d'obtenir le meilleur résultat possible pour l'environnement et pour les gens qui en dépendent ([1]). C'est le même raisonnement que nous voulons appliquer. Nous ne voulons pas qu'une formule magique dicte comment agir. Nous souhaitons donc un outil qui présente des données lisibles, compréhensibles, intelligibles. Nous aimerions qu'un être humain puisse trouver des corrélations lui-même et qu'il puisse élaborer sa propre réflexion sur une décision qui touche à la planification urbanistique (et donc à la vie des gens). Cela permettrait de contrer le manque de neutralité des algorithmes, qui sont biaisés par leurs créateurs, et les données qui les ont entrainés ([2]).

Il découle naturellement un besoin d'avoir une source de données non filtrée, contrôlable, objective, non biaisée et en temps réel. Parfois, les données publiques sont manipulées par des lobbies (banques, agences, entreprises de construction, les grands propriétaires, etc.). Mettre à disposition ces données sous la forme d'un outil utilisable et compréhensible est nécessaire pour aider les preneurs de décisions à agir de manière transparente.

1.3. Etat de l'art

Le but de cette revue littéraire est de fournir un aperçu des recherches existantes sur le minage de données pour l'analyse de marchés immobiliers afin d'apporter un support à la prise de décision. Il s'agit d'identifier les concepts clés, les méthodes et les découvertes par rapport à ce champ d'activité.

[1] En informatique et en apprentissage automatique (machine learning), une « feature » est une variable ou une caractéristique d'un ensemble de données qui est utilisée pour entraîner un modèle ou effectuer une analyse.

Le marché immobilier et ses dynamiques : une description du marché immobilier, de ses différents acteurs et de ses dynamiques.

Le marché immobilier est un système complexe et dynamique qui implique de multiples acteurs tels que les acheteurs, les vendeurs, les agents immobiliers et les investisseurs. Comprendre les dynamiques du marché immobilier est essentiel pour pouvoir prendre des décisions informées et pour pouvoir prédire les comportements qui en découlent, afin de développer des stratégies urbanistiques efficaces.

Un facteur clé qui affecte le marché immobilier est l'état général de l'économie ([3]). En période de croissance, la demande sur ce marché tend à croître et la compétition accrue entre les acheteurs conduit à une hausse des prix. L'effet inverse est observé en période de ralentissement de l'économie.

Un autre facteur qui affecte le marché de l'immobilier est la démographie. Les facteurs démographiques, tels que la taille de la population, l'âge, le niveau de revenu, etc., peuvent avoir un impact significatif sur la demande dans une zone en particulier. Par exemple, un afflux de jeunes travailleurs peut conduire à une hausse de la demande de location dans les zones urbaines ([4]). Dans certaines villes de Belgique, nous assistons à une hausse de la demande de colocation et cela tend à changer les habitudes et les dynamiques de vie dans ces zones (nouveaux types de baux, changement de type de voisinage, etc.).

Le marché immobilier est aussi influencé par la politique et ses règles. La définition d'un « zoning », les prescriptions urbanistiques, les taxes et autres peuvent influencer l'accessibilité à la propriété ou à la location pour certains segments de la population.

Par exemple, l'étude de Floetotto et al. [5], suggère que les crédits d'impôt pour l'achat d'une maison aux USA augmentent temporairement les prix des maisons et le volume des transactions, mais ont des effets négatifs sur le bien-être.

De manière générale, le marché immobilier est un système complexe qui est influencé par une grande variété de facteurs économiques, sociaux et politiques.

Les techniques résultantes du minage de données et leurs applications sur le marché immobilier

La revue systématique des modèles d'évaluation de masse du marché immobilier de Wang et al. [6] met en lumière certaines techniques :

Une première technique est l'analyse de régression pour l'analyse de données issues du marché de l'immobilier. Cette technique implique d'identifier la relation entre une ou plusieurs variables indépendantes comme la localisation, la taille, l'âge et une variable dépendante telle que la valeur du bien immobilier. Une autre technique fréquemment utilisée est l'analyse par cluster. Cette

technique implique l'identification de groupes de propriétés ayant des caractéristiques similaires telles que la localisation ou la taille. Cette analyse par cluster permet d'identifier des tendances.

Le machine learning (réseau neuronal, arbre de décision) est également beaucoup utilisé dans les analyses de marché immobilier. Ces techniques permettent d'analyser des grands jeux de données et identifier des patterns complexes ainsi que des relations qui ne sont pas facilement détectables par les méthodes de statistiques traditionnelles. L'étude de Mora-Garcia et al. [7] compare différents modèles de machine learning face aux modèles linéaires et détermine si un événement tel que le COVID-19 exerce une influence sur le prix de l'immobilier.

Les techniques résultantes du minage sont de plus en plus importantes pour analyser les dynamiques du marché immobilier, puisque le volume et la complexité des données augmentent.

La collecte de données en temps réel et le traitement

Voici un extrait de la revue littéraire « Applicabilité et fondements théoriques du scraping » [8] que nous avons réalisée préalablement à ce mémoire.

« Nous avons constaté que le scraping répond à de nombreux besoins en tant qu'outil de data mining. En effet, depuis que les humains laissent une empreinte digitale, il y a sur Internet une mine d'informations presque inépuisables.

C'est un outil relativement simple à construire mais relativement compliqué à maintenir car peu résilient : il demande un grand travail de la part de l'humain. La plus grande valeur ajoutée est apportée par l'humain lors de l'étape de sélection des sites Internet, de tri et d'analyse des données. L'analyse permet de révéler tout le potentiel informationnel de la donnée et tirer des conclusions de cette analyse est une prouesse qui demande une bonne connaissance dans les matières relatives à la data science. Le traitement humain est d'autant plus délicat que certaines préoccupations éthiques et légales ne sont pas encore totalement explorées. C'est donc un outil qu'il faut utiliser avec précaution et parcimonie (respect des sujets, respect du matériel qui héberge le site, ...). Des chercheurs seront parfois tentés de dire que « la fin justifie les moyens », mais c'est comme vouloir marcher sur un fil tendu entre deux falaises : cela demande un équilibre à toute épreuve. »

Les systèmes d'aide à la décision pour les collectivités locales : quels sont les différents systèmes existants et les bénéfices qu'ils offrent ?

Les collectivités font face à de multiples problèmes interconnectés et complexes qui requièrent une analyse précise avant de pouvoir prendre des décisions. Les systèmes de support à la décision (SSD) sont apparus comme étant des outils puissants pour faire face à ces défis. Les SSD sont des systèmes informatiques qui fournissent des outils interactifs et des modèles analytiques qui permettent d'apporter un support lors du processus décisionnel. Il en existe différents types offrant des bénéfices différents.

Par exemple, Balamurugan et al. [9] documente le processus de conceptualisation d'un tableau de bord incluant des graphiques et un système d'information géographique (SIG) permettant de montrer l'impact du COVID-19 sur le marché de l'immobilier. Les technologies SIG sont très pratiques. Elles permettent de visualiser des données spatiales de manière intuitive sur une carte.

Déjà en 1997, Timmermans [10] pointait le fait que les décisions urbanistiques doivent être éclairées et qu'un SSD flexible doté d'une bonne utilisabilité permet de faire une évaluation multicritère fort utile et transparente.

Les études expliquant qu'il existe de nombreuses applications utiles pour un SSD (planification urbaine, agriculture, gestion des ressources, …) sont nombreuses. Elles démontrent également l'importance d'incorporer des technologies avancées comme les intelligences artificielles.

Revues des recherches les plus pertinentes et récentes en lien avec le sujet du mémoire

De nombreuses études se concentrent sur l'évolution des prix ([11]) ou sur les différentes features les plus importantes pour déterminer un prix. Mais peu d'études visent à assurer un suivi du marché immobilier en vue d'améliorer la planification urbanistique dans le but d'améliorer/maintenir la qualité de vie (l'humain au centre du processus décisionnel).

Il existe des variantes. Beaucoup d'études cherchent à analyser l'impact que peut avoir un événement sur le marché. Par exemple, Grybauskas et al. [12] veut montrer quel est l'attribut d'un appartement qui est le plus susceptible d'influencer une révision de prix pendant un événement tel qu'une pandémie.

Pour conclure cette revue de littérature, nous pensons que ce mémoire comblera une lacune importante en analysant la conception et la mise en œuvre d'un SSD (Système de Surveillance de Données) flexible, évolutif et libre. Cela permettra aux utilisateurs non formés de s'informer de manière impartiale et de configurer leur propre tableau de bord afin de tirer leurs propres conclusions et d'identifier des corrélations.

1.4. Sous-questions de recherches

Comment est-ce que le minage de données peut-il être utilisé pour refléter les activités de marché en temps réel ?

Quelles sont les sources de données les plus appropriées pour l'analyse du marché immobilier utilisant les techniques de minage ?

Comment préserver le bien-être et le bien vivre dans le cadre du processus décisionnel de planification urbanistique ?

De quelle manière est-ce que les problèmes relatifs au nettoyage et à la qualité des données peut-il appréhender à la suite d'une collecte de données sur le marché immobilier ?

Comment est-ce que les techniques d'analyse spatiales et temporelles peuvent être utilisées pour fournir une meilleure compréhension des dynamiques du marché immobilier ?

Comment est-ce que les techniques de segmentation et de clustering peuvent être utilisées pour améliorer la précision des analyses du marché immobilier ?

Quels sont les éléments d'un SSD les plus efficaces pour les collectivités locales dans le contexte d'une prise de décision urbanistique ?

Comment les effets d'une décision peuvent-ils se refléter sur les performances du marché immobilier ? Comment peuvent-ils être évalués en utilisant des techniques de minage ?

Comment un outil de SIG peut-il être utilisé en amont d'une prise de décision dans le contexte d'une planification urbanistique ?

1.5. Objectifs

Cette recherche souhaite montrer qu'il est possible d'utiliser les techniques de minage de données pour créer un système d'aide et de support à la décision. Cela permettra aux décideurs de prendre des décisions informées tout en gardant l'humain au cœur du processus décisionnel. Les données récoltées se doivent de ne pas être biaisées et le système se doit d'être évolutif, utilisable et transparent. Le SSD est un outil utilisé pour prédire l'impact d'une décision ou mesurer le résultat d'une décision passée.

Le développement du SSD doit donc fournir une vue en temps réel pour dépeindre une image précise et complète du marché de l'immobilier.

Cette recherche se concentre sur une localisation géographique spécifique et dans un espace de temps donné. Il reposera uniquement sur des sources de données en accès libre et/ou extraites par un outil de scraping. Ces données seront la base de l'analyse.

Enfin, nous devrons être en mesure de proposer des exemples de recommandations pour des règles édictées par des collectivités locales.

1.6. Contribution de ce mémoire

Ce mémoire contribue à montrer qu'il est possible de créer un outil de scraping sans avoir besoin d'une infrastructure coûteuse dédiée, en utilisant uniquement des outils gratuits, grand public et libres de droit.

Il démontre également la possibilité de se réapproprier les données et l'interprétation que l'on en fait. Alors que la plupart des études sur ce sujet cherchent à trouver la meilleure combinaison de features ou le meilleur modèle pour prédire le prix, ce mémoire veut présenter un outil customisable et évolutif, piloté par un humain, pour aider celui-ci à prendre des décisions éclairées de son propre chef. Le fait de garder l'humain dans ce processus peut contribuer à garantir une vision durable et humaine pour préserver une bonne qualité de vie (pas l'efficacité à tout prix).

L'intégration d'un outil SIG dans un système web évolutif et customisable doit permettre aux utilisateurs du système d'avoir une meilleure compréhension des facteurs économiques et sociaux affectant le marché de l'immobilier et conduire à une meilleure prise de décision.

Ce mémoire contribue également à valider l'efficacité du scraping pour extraire des données utiles visant à une meilleure compréhension d'un marché.

2. Le scraping

2.1. Une définition

« Le web scraping est une technique d'extraction du contenu de sites Web, via un script ou un programme, dans le but de le transformer pour permettre son utilisation dans un autre contexte comme l'enrichissement de bases de données, le référencement ou l'exploration de données. »

(Wikipédia, 06/03/2023)

2.2. Les méthodes traditionnelles de collecte de données

Les méthodes traditionnelles sont :

- les enquêtes/sondages (en ligne, par téléphone, face-à-face) ;
- les observations/ enquêtes de terrain ;
- les focus groups ;
- l'analyse documentaire ;
- les études de cas ;
- etc.

Ces méthodes sont chronophages et imposent une charge de travail conséquente (et donc un coût important). Il est essentiel de savoir choisir la bonne méthode en fonction du besoin et des ressources disponibles.

2.3. Historique

En 1989, Berners-Lee crée le World Wide Web. A cette époque, les hyperliens, les URLs et les pages web formatées existaient déjà, mais les données n'étaient pas aussi bien présentées qu'aujourd'hui. Des scientifiques ont alors mis au point le web scraping (aussi appelé scraping). Les premiers robots crawlers avaient pour missions d'indexer le web. Aujourd'hui, il existe de nombreuses librairies offrant des fonctionnalités pour parser[2] et extraire le contenu d'une page web.

2.4. Les différents types de scraping

Voici les différents types de scraping qui peuvent être utilisés séparément ou de manière combinée :

- Le parsing HTML consiste en l'extraction des données HTML d'une page web statique.
- Le parsing DOM (Data Object Model)[3] consiste en l'extraction des données du DOM d'une page web dynamique.
- Le web crawling permet la navigation sur des pages web, le téléchargement de documents, la prise de captures d'écran et le remplissage de formulaires.

2.5. Applications existantes et potentielles

2.5.1.Statbel et l'indice des prix à la consommation

En Belgique, l'office de statistique (Statbel) utilise le scraping pour récolter plus d'informations afin d'accroître la précision de ses relevés. Les données sont ensuite utilisées pour calculer l'indice des prix à la consommation.

L'indice des prix à la consommation est une mesure mensuelle permettant de mesurer l'inflation et « *sert de base directe (…) à l'indexation des pensions, des allocations sociales, des barèmes fiscaux, des loyers et de certains salaires et traitements* » (Statbel, 2018) ([13]).

[2] Le parsing (ou l'analyse syntaxique en français) est le processus d'analyse d'une chaîne de caractères ou d'un ensemble de données structurées pour déterminer leur structure grammaticale et en extraire des informations. Dans le contexte de l'informatique, le parsing est souvent utilisé pour analyser des codes sources de programmes informatiques, des documents XML, des pages web ou des messages de protocoles de communication.

[3] Le DOM représente la page web sous la forme d'une hiérarchie d'objets permettant ainsi aux programmeurs de manipuler et de modifier le contenu et la structure de la page web en utilisant du code JavaScript.

Une soixantaine de scripts permettant de récolter des données sur des produits venant de segments aussi divers et variés que l'habillement, les hôtels, les véhicules d'occasion, les jeux vidéo, les supermarchés, etc.

Lors de son étude, Statbel remarque que les méthodes hédoniques (pour le calcul de régression) ne sont utilisées qu'exceptionnellement en Europe. Ce sont des méthodes qui demandent de nombreuses données. Grâce au scraping, il est possible de le faire sur certains groupes de produits. Il est rendu possible d'appliquer une équation de régression qui est à jour constamment.

À date de l'ouvrage, le segment des voitures d'occasion n'est pas inclus dans les indices de prix à la consommation (national et européen). Or, c'est une faute puisque le seuil de 1/000 des dépenses des ménages est dépassé. La raison est qu'il est difficile d'obtenir les données concernant ce segment. Avec le scraping, la donne change. Les prix sont disponibles sur les sites de petites annonces et peuvent être récupérés par des scripts. Grâce au scraping, le résultat est inclus dans l'indice européen depuis 2019 (concernant l'indice national, la Commission de l'indice devait encore remettre son avis).

2.5.2.Exemples d'applications potentielles

- Étudier la possibilité de mesurer le nombre de postes vacants ;
- surveiller la concurrence (prix et changement de prix, item ajouter/retirer dans le store) ;
- surveiller l'opinion publique (communautés en ligne, forums, reviews, etc.) ;
- surveiller les sites concurrents (quels sont les mots-clés et les liens utilisés) ;
- etc.

3. Ethique et légalité en Europe et en Belgique

3.1. Introduction

Il existe plusieurs raisons d'utiliser le scraping. On peut imaginer de nombreux cas « business » tels que de la surveillance ou du benchmarking. On peut aussi imaginer des usages frauduleux comme une copie d'un site Internet ou de son SEO (Search Engine Optimization). Ce sont des situations qui portent à croire que le web scraping ne devrait pas être autorisé.

À côté de cela, nous connaissons des histoires comme celle du site Internet « vitemadose », créé au tout début de la pandémie de COVID-19. Le site centralisait tous les lieux où il était possible de se faire vacciner afin d'optimiser au mieux les opérations. Pour y parvenir, il réalisait un scraping de tous les sites publics qui diffusaient de l'information sans concertations. Le site centralisait également toutes sortes d'informations provenant de diverses sources. La solution était rapide, efficace et apportée par un citoyen.

Nous avons également vu que le scraping répond à un grand nombre de besoins réels et est applicable à un grand nombre de cas d'utilisation possibles, utiles et nécessaires ([13]). Comme par exemple : suivre l'évolution de la qualité nutritionnelle des aliments de supermarchés, effectuer des recherches épidémiologiques ou encore organiser la planification de projets de santé publique. Nous avons également vu que Statbel l'utilise pour calculer des indices nationaux et européens.

3.2. Ethique

À qui appartiennent réellement les données publiques présentes sur le réseau Internet ? Que peut-on en faire ? Quelle est la limite ?

Il n'existe pas un cadre universel sur ce genre de questions, mais certaines règles de base de la recherche scientifique peuvent s'appliquer (voir notamment l'étude d'Emanuel et al. [14] qui a servi de base à Rennie et al. [15]). Mancosu et Vegetti [16] définissent un traitement éthique comme étant une recherche guidée par les principes de respect d'une personne, de bienfaisance et de justice.

Voici trois standards éthiques acceptés par la communauté scientifique :

- Préserver la vie privée et ne causer aucun tort ([16]).
- Respecter le cadre légal qui protège l'individu ([16]).
- Respecter les règles d'utilisation de la plateforme qui contient les données (le « TOS - Terms Of Service ») ([16]).

Les scientifiques sont d'accord sur le fait qu'il faut protéger l'individu mais ils nuancent tout de même :

- L'approbation et le consentement des utilisateurs (les sujets) ne sont pas requis puisque les données sont présentes sur un réseau public ([17], [18]).
- L'approbation et le consentement des utilisateurs (les sujets) ne sont pas requis puisque les chercheurs n'ont pas d'interaction avec les sujets ([17], [19]).
- Il faut veiller à ne pas apporter une charge excessive sur le système hôte ([19]) mais respecter le « TOS » est impossible car il est trop restrictif ([16]).
- Il faut tout mettre en œuvre pour organiser la désidentification des données, et si ce n'est pas possible, il faut limiter la publication des résultats.

Par souci d'explorer tous les points de vue, nous avons observé l'étude de Afzal Haque et al. [20] dont le but était de démontrer comment se défendre du scraping. Cette étude ne mesure pas son propos et s'arrête sur le fait qu'il est soi-disant interdit et que les sites Internet doivent tous s'en défendre. Ils ajoutent que les données ont de la valeur et que c'est donc du vol et une menace pour le business en ligne.

On remarque que cette question est abordée plus en profondeur dans certaines études. Nous pensons que le débat doit rester ouvert et que la synthèse des points cités ci-dessus constitue déjà une bonne approche.

En ce qui concerne la réalisation de ce mémoire, il est bon de noter :

- que nous ne causons aucun tort aux intérêts légitimes du producteur de la base de données ;
- que nous n'utilisons aucun subterfuge pour détourner d'éventuelles protections (pas de solveur de captcha, pas de proxy, …) ;
- que nous exécutons les scripts aux heures creuses (entre 00:00 A.M. et 05:00 A.M.) ;
- que nous n'exécutons pas les requêtes de manière intensive (1 à 2 secondes entre chaque requête pour un script) ;
- que notre but est l'étude et la recherche ;
- que nous ne cherchons pas à tirer profit ;
- que nous ne redistribuons pas l'information récoltée.

En somme, nous considérons que d'un point de vue éthique notre recherche est acceptable car elle observe un but louable et est réalisée de manière scientifique. Le résultat de celle-ci vise à proposer des solutions pour améliorer la planification urbanistique et préserver/améliorer le bien-être de la population.

3.3. Légalité

Régime légal du droit d'auteur

Dans le cadre du scraping, le régime du droit d'auteur en Belgique peut être interprété de la manière suivante :

Lorsqu'un site web contient des informations protégées par des droits d'auteur, telles que du texte, des images, des vidéos ou des données, le scraping de ces informations sans autorisation préalable peut constituer une violation des droits d'auteur. Le scraping peut être considéré comme une reproduction non autorisée de ces informations, qui sont protégées par le droit d'auteur.

Cependant, il est important de noter que le scraping de données publiques qui ne sont pas protégées par des droits d'auteur, telles que des informations d'ordre général sur les entreprises, les produits ou les services, peut être considéré comme légal.

A. La loi belge du 30 juin 1994 sur le droit d'auteur et les droits voisins ([21]) :

Elle définit les droits d'auteur et les droits voisins, qui protègent respectivement les œuvres originales de l'esprit (comme les livres, les films, les musiques et les logiciels) et les performances d'artistes-interprètes, etc. La loi établit également les conditions de protection, la durée de la protection, les exceptions aux droits d'auteur et les droits moraux des auteurs. Elle prévoit également des sanctions pénales et civiles pour les violations de droits d'auteur et de droits voisins.

B. La position de l'Office de la Propriété Intellectuelle de Belgique (OPRI) sur le web scraping (extrait) :

« Le SPF Economie mène actuellement des projets concrets pour exploiter des types prometteurs de mégadonnées : données de scanning du commerce de détail, données de téléphonie mobile, « webscraping » via les robots des sites internet. Elles constituent l'amorce d'une troisième révolution des données dans le domaine de la statistique, après les enquêtes auprès des citoyens et des entreprises et l'utilisation, plus récente, des fichiers administratifs. » (SPF Economie, 2018, p. 34) [22]

« *Par ailleurs, le SPF Economie s'investit pleinement dans l'exploitation des mégadonnées (big data) et l'utilisation du webscraping.* » (SPF Economie, 2018, p. 33) [22]

C. La position de la Commission européenne sur le scraping (extrait) :

« Web scraping will be performed in adherence with the principles of the European Statistics Code of Practice, and in compliance with intellectual property legislation (national copyright laws and the Database directive) (6).

The members of the ESS should use web scraped data solely for statistical purposes as laid down in regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics and the applicable national statistical legislation.

The principles guiding web scraping activities should be to maximize the benefits while minimizing any burden, risks or potential impacts arising from these activities. » (European Commission, 2020) [23]

The Belgian Database Act of 1998 [31]

En Belgique, la loi sur les bases de données est une loi qui vise à protéger les droits des producteurs de base de données. Cette loi est une transposition de la directive européenne n°96/9/CE. Cette loi confère aux producteurs de base de données des droits exclusifs, tels que le droit de reproduire, distribuer et communiquer au public tout ou partie de la base de données. Cela concerne les bases de données qui font l'objet d'une protection en vertu du droit d'auteur ou du droit sui generis ([24]).

Au niveau européen, la jurisprudence montre qu'il faut que le défendeur puisse prouver qu'il fait assez d'investissement pour protéger sa base de données. En effet, en 2015 aux Pays-Bas, Ryanair a perdu son procès contre un site agrégateur de tarif ([25]).

Règlement (UE) 2016/792 article 5.3 [26]

Une disposition du règlement européen prévoit et encadre le scraping dans un certain contexte. Le document de Statbel que nous avons déjà évoqué résume la situation :

« *Les unités statistiques qui communiquent des informations sur les produits inclus dans les dépenses monétaires de consommation finale des ménages coopèrent à la collecte ou à la communication des informations de base selon les besoins. Les unités statistiques sont tenues de transmettre des informations de base exactes et complètes aux organismes nationaux chargés du calcul des indices harmonisés.* » (Statbel, 2018, p. 7) [13]

RGPD

Si les données collectées ne contiennent pas de données personnelles (nom, année de naissance, adresse IP, …) alors RGPD ne s'applique pas.

4. Système de récolte des données du marché immobilier namurois en temps réel

Dans ce chapitre, nous allons présenter en détails le système de scraping que nous avons développé pour collecter les données du marché immobilier namurois en temps réel.

Nous expliquerons comment nous avons sélectionné les technologies et les outils utilisés pour collecter les données, comment celles-ci sont stockées et traitées, et comment le système permet de surveiller les tendances du marché immobilier en temps réel. Cette étape nous mènera à comprendre comment les données peuvent être intégrées dans un système de support à la décision, offrant ainsi une aide précieuse aux collectivités locales dans leurs prises de décisions.

4.1. Technologies

4.1.1.Le choix du langage

Lors de la réalisation de notre état de l'art, nous avons constaté que le langage Python était le plus utilisé par les chercheurs pour créer les scripts de collecte de données. Python est un langage de programmation de haut niveau, interprété et orienté objet. Il est connu notamment pour son utilisation dans l'analyse de données et l'intelligence artificielle.

Bien que Python semble être le choix qui s'impose, nos impératifs nous ont amenés vers un autre langage : le JavaScript. C'est également un langage de programmation orienté objet et il est principalement utilisé pour le développement web. À l'origine, il était essentiellement utilisé pour créer des effets dynamiques sur les pages web mais aujourd'hui, il est utilisé partout, tant pour le développement de jeux et d'applications mobiles que pour le développement de logiciels serveur.

Pour être précis, nous n'utilisons pas le Vanilla JS[4] mais le Typescript, afin de profiter des fonctionnalités telles que la vérification de type statique et la prise en charge des classes et des interfaces.

Nous justifions ce choix :

A. Simplicité

Développer un outil de scraping est quelque-chose de relativement complexe. Cela demande des connaissances multidisciplinaires (protocole http, développement frontend, développement backend, gestion de bases de données, HTML, DOM, infrastructures cloud, etc.).

Nous voulons limiter le nombre de technologies pour assurer une complémentarité/compatibilité des outils. Cela à l'avantage de rendre la construction de l'outil plus accessible et plus rapide.

B. L'écosystème de développement

Les interfaces web écrites en JavaScript sont efficaces. Ce langage de programmation est exécuté côté client. Son exécution directe dans le navigateur peut offrir une expérience utilisateur plus fluide et plus rapide qu'en Python, par exemple.

Le JavaScript est un langage de programmation très populaire et dispose d'une grande communauté de développeurs et de contributeurs actifs. Il existe donc une grande variété de frameworks.

C. La facilité de la mise à l'échelle des applications

D. La flexibilité

La récolte des données sur des sites dynamiques est aisée en JavaScript. Ce langage est capable de s'adapter à de nombreux types de projets. Il ne faut pas oublier que le JavaScript est conçu pour être exécuté côté client, c'est-à-dire directement dans le navigateur web de l'utilisateur. Le JavaScript peut interagir avec le DOM, ce qui permet de modifier dynamiquement le contenu et l'apparence des pages web sans avoir à recharger la page entière.

Python n'a pas cette manipulation native du DOM. Il existe des bibliothèques pour y parvenir (e.g. Beautiful Soup) mais elles ne sont pas aussi adaptées que le Javascript pour la manipulation directe du DOM dans le navigateur.

[4] Vanilla JS est un terme utilisé pour décrire le fait d'utiliser uniquement le langage JavaScript pur pour créer des applications web, sans recourir à des bibliothèques ou frameworks tiers tels que JQuery.

4.1.2.Constitution de la stack de développement

Tout d'abord, pour la collecte de données, nous avons utilisé la bibliothèque JavaScript open source **Puppeteer**. Elle permet de contrôler de manière programmée un navigateur web ayant un moteur Chromium. Elle permet d'effectuer des actions telles que la navigation vers des pages web, la saisie de formulaires, la gestion des cookies et surtout, elle permet de manipuler le DOM et d'interagir avec les éléments de la page web.

Pour le stockage, nous avons besoin d'une base de données qui soit facilement évolutive et qui ne nécessite pas un schéma de données rigide. Nous n'avons pas de contrôle sur la structure des données collectées, il est donc important d'avoir une solution flexible. Ainsi, nous avons choisi MongoDB, une base de données NoSQL capable de stocker des données non structurées et semi-structurées. MongoDB nous permet de stocker les données collectées sous forme de documents JSON, ce qui facilite leur traitement avec JavaScript.

Nous avons besoin de déclencher notre scraper quotidiennement à l'aide de requêtes http. Ainsi, nous avons choisi d'utiliser **Express.js** pour gérer efficacement ces requêtes. Ce framework minimaliste de développement d'applications web en JavaScript est construit sur Node.js, ce qui en fait un choix idéal pour la construction de notre API **RESTful**.

Enfin, il est important de respecter une méthode de versioning pour le code source. Le repository est sur **GitHub** et respecte le GitFlow, qui est une méthode populaire pour la gestion de versions de code avec git.

4.1.3.Déploiement

Nous abordons maintenant le sujet du déploiement de l'application sur l'environnement de production qui est hébergé sur Microsoft Azure. Grâce à notre statut d'étudiant à l'Université de Namur, nous avons pu créer un compte et obtenir 100 $ de crédit par an, ce qui nous a permis de nous familiariser avec les outils proposés par Microsoft Azure et de déterminer que cet outil convient à notre besoin. Les outils Azure que nous avons choisi d'utiliser sont :

- *Azure Container Instance* : outil qui permet de faire fonctionner des containers[5] et créer des groupes de containers. Nous avons le choix d'utiliser des containers Microsoft ou Docker.
- *Logic App* : c'est un service de workflow visuel qui permet de créer des processus automatisés pour les tâches métier et les intégrations de données. Il fournit une interface visuelle pour définir les étapes du workflow et peut être intégré avec une variété de services Azure et tiers pour créer des workflows puissants.

[5] « Container » est un mot anglais qui a été emprunté dans de nombreuses langues, y compris le français. En informatique, il est courant de l'utiliser pour désigner un conteneur, qui peut être une unité de déploiement.

Afin de faciliter le déploiement et la portabilité de notre application, nous avions la possibilité d'utiliser une VM Azure (machine virtuelle) ou les containers. Nous avons opté pour la conteneurisation Docker, qui s'est avérée être (trop) coûteuse.

Cependant, le choix du container est limité en raison de la nature de l'application et de son utilisation de la librairie de scraping que nous avons choisie. Puppeteer est une librairie qui fournit une API de haut niveau pour contrôler un « Headless Chrome » * ou un Chromium sur le « DevTools Protocol » **. Une application utilisant cette librairie requiert un support GDI ***. Or, les services Azure (qui tournent sur des serveurs Windows) ne peuvent pas faire tourner ce genre de logiciel pour des raisons de sécurité (l'accès aux fonctionnalités UI de Windows est dangereux).

* Cela signifie que la librairie peut faire tout ce que fait un navigateur Chromium sans pour autant avoir d'interface.

** C'est un protocole de communication qui permet aux développeurs d'interagir avec les outils de développement de Google Chrome et d'autres navigateurs web compatibles avec ce protocole. Le DevTools Protocol permet aux développeurs de contrôler le navigateur à distance, d'inspecter et de modifier le contenu de la page, de surveiller les événements de navigation, de collecter des informations de performance, de déboguer le code et d'effectuer d'autres tâches de développement web.

*** Le terme « GDI » signifie « Graphics Device Interface », qui est une interface de programmation d'application (API) fournie par le système d'exploitation Microsoft Windows pour gérer les graphiques et les images.

4.1.4.Outils de visualisation des données

MongoDB Atlas Charts

Outil de visualisation natif pour les données dans les bases de données MongoDB Atlas. Il permet de générer facilement des histogrammes, des graphiques circulaires (donuts) et autres types de graphiques.

Tableau Desktop

Notre statut d'étudiant à l'Université de Namur nous permet d'utiliser cet outil très puissant pour générer des graphiques. Nous l'utilisons principalement pour créer les cartes choroplèthe[6] et des diagrammes en boîte (boxplot).

Jupyter Notebook

[6] Une carte choroplèthe est une carte thématique dans laquelle les régions sont colorées d'une intensité variante en fonction des données.

Nous utilisons Python et des librairies telles que Pandas et Numpy pour générer des graphiques (scatter plot et cluster).

4.1.5.Autres outils

- Winston : librairie de gestion des logs.
- Visual Studio Code : environnement de développement léger (IDE – Interface Development Environment).
- Docker desktop : outil de gestion des containers et de leur versionning[7].
- Docker Hub : repository d'image docker
- Compass : système de gestion de base de données pour MongoDB.

4.2. Infrastructure et architecture du système

4.2.1.Infrastructure du système

Nous avons choisi le modèle PaaS (Platform as a Service), qui est une approche d'hébergement d'applications où le fournisseur de services cloud fournit une plateforme complète pour exécuter et gérer les applications. Puisque notre système de scraping doit fonctionner sept jours sur sept et que l'exécution d'un script peut durer jusqu'à 5 heures (en fonction de l'intensité à laquelle nous décidons d'envoyer les requêtes), nous nous sommes naturellement orientés vers ce type de solution cloud.

En effet, il aurait été très coûteux en argent et en temps d'investir dans un serveur physique. Le modèle PaaS nous permet donc une réduction des coûts d'infrastructure et de maintenance, car l'application est hébergée et gérée par le fournisseur de services. De ce fait, nous nous pouvons mettre de côté les soucis liés à l'évolutivité et la sécurité.

Nous avons utilisé Microsoft Azure, qui nous permet de se concentrer sur la création d'applications sans avoir à gérer l'infrastructure sous-jacente physique.

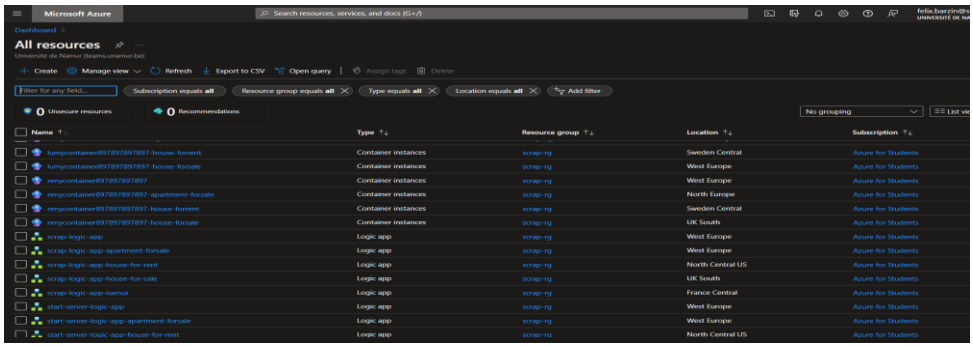


FIGURE 1 - Extrait d'une vue du tableau de bord de gestion des produits Azure

[7] Le versionning (ou versionnage) est le processus de numérotation et de suivi des différentes versions d'un logiciel ou d'un fichier.

4.2.2.Architecture du système

Description générale

Le composant principal est le scraper, qui récupère les données immobilières en temps réel à partir d'une source web. La récupération de données se fait à intervalle régulier tous les jours, grâce à un workflow ordonnancé par les services de Microsoft Azure. Les données collectées sont stockées dans une base de données centralisée pour une analyse ultérieure.

Le système est conçu pour être évolutif et extensible grâce à l'utilisation de fichiers de configuration pouvant être remplacés ou agrémentés facilement. Il permet également une intégration facile de nouvelles sources de données et une extension des fonctionnalités.

Diagramme d'architecture présentant les différents composants et leurs interactions

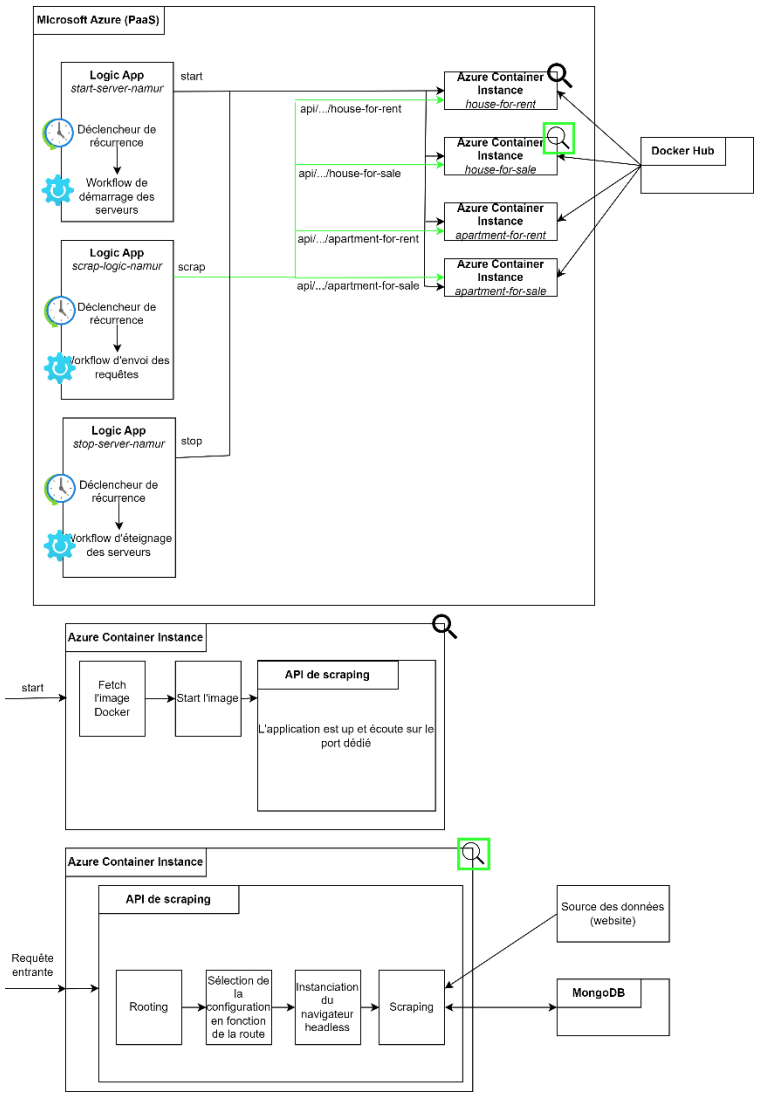


FIGURE 2 - Schéma d'infrastructure cloud

Description du workflow du PaaS Microsoft Azure

Le workflow général s'articule autour de trois « Logic App » qui offrent les fonctionnalités suivantes :

- La première (*start-server-namur*) allume les différents serveurs. Lorsqu'un serveur s'allume, il va chercher l'image Docker et il s'occupe de la mise en service.
- La deuxième (*scrap-logic-namur*) lance les requêtes sur les différents serveurs. Lorsque la première requête est envoyée vers une instance conteneurisée, notre API de scraping va commencer son travail : elle va commencer par instancier un navigateur en mode « headless », c'est-à-dire sans interface. Ensuite, la source de données sera récoltée en fonction de la route spécifiée.
- La troisième (*stop-server-namur*) va servir à éteindre les différents serveurs.

```
90        this._router.get('/immoweb/houses/for-sale', (req: Request, res: Response, next: NextFunction) => {
91          try {
92            this._logger.debug('GET /immoweb/houses/for-sale');
93            const province = req.query.province as string;
94
95            if (typeof province === 'undefined') {
96              res.status(404).json('404');
97            } else {
98              const result = this._controller.scrap(province, 'house-for-sale');
99              res.status(200).json(result);
100           }
101         }
102         catch (error) {
103           next(error);
104         }
105       });
106     }
107   }
```

FIGURE 3 - ScrapRouter.ts – La réception de la requête « /immoweb/houses/for-sale » entrainera la récolte de données des maisons à vendre pour la province passée en paramètre

Nous avons choisi d'articuler cela de cette manière pour limiter l'utilisation des outils très coûteux comme l'API Gateway et le Private Virtual Network, qui auraient pu permettre aux containers de s'éteindre eux-mêmes une fois le travail terminé.

Afin d'obtenir un maximum de récoltes, nous avons dû optimiser le temps durant lequel un serveur reste allumé. En effet, une fois qu'il a reçu son ordre de lancer la récolte, l'instanciation du navigateur va entraîner une consommation de 800 MB de mémoire et de l'utilisation de CPU (facturé par seconde d'utilisation).

Les quatre containers sont créés sur base de la même image Docker. Cependant, l'API contient une collection de routes différentes, et chacune est dédiée à un container en particulier. Donc, chaque container correspond à un type de données spécifique. Cette organisation facilite la maintenance car les logs sont bien découpés. Cela facilite aussi le débogage car les erreurs sont

fortement liées aux types de données. Par exemple, la propriété « monthlyCosts » existe pour les maisons en location mais n'existe pas pour les maisons à vendre. Si elle venait à changer et créer un bug, cela n'affecterait que les containers qui traitent les locations. Le découpage des scripts permet également de limiter la charge et les coûts dans l'éventualité où ceux-ci doivent être rejoués.

4.2.3.Présentation des composants de l'API de scraping

https://github.com/felixbarzin/scrap

dockerfile

Le fichier texte qui contient la série d'instructions pour la construction de l'image Docker. Il permet d'avoir un processus de création standardisé et reproductible.

Dans ce fichier, nous allons par exemple spécifier la version du moteur Chromium que nous désirons et le port sur lequel l'application écoute.

package.json

C'est le fichier de configuration qui décrit les dépendances du projet et le script de démarrage.

.env.[environnement]

C'est un fichier qui n'est pas versionné car il contient des informations sensibles. Son template se trouve dans le repository sous le nom de « *.env.example* ». C'est aussi dans ce fichier que nous devons renseigner les informations de connexion à la base de données, les credentials de login aux sites sources, etc.

C'est le fichier « Secret.ts » qui aura le rôle de parcourir cette configuration et de l'importer dans l'application.

app.ts

C'est un fichier qui contient la logique de démarrage de l'application. Il contient le système de routing et le setup de la base de données.

config.json

En fonction de la route empruntée, un container sera dédié à une récolte de données spécifique. Il récolte soit les données des appartements à vendre, soit les données des appartements en location, soit les données des maisons à vendre, soit les données des maisons en location. Dans ce fichier de configuration, nous retrouvons les informations dédiées pour que chaque container spécialisé puisse connaitre sa source de données.

ScrapRouter.ts

Ce fichier regroupe la configuration des routes de notre API. Par défaut, l'application écoute toutes les requêtes entrantes sur les routes commençant par « /api ». Un module appelé « MasterRouter » va venir se greffer à ce module de mapping de route par défaut pour ajouter des sous-routes. Le « MasterRouter » sera notre gestionnaire de route. Il va donc orienter les requêtes vers des contrôleurs spécifiques. Par exemple, il est configuré pour diriger les requêtes « /scrap » vers un contrôleur dédié au processus de récolte : le « ScrapController ». Par exemple : http://localhost:[PORT]/api/scrap/hello répondra avec un statut 200 si tout va bien.

ScrapController.ts

Ce contrôleur est le point d'entrée de l'API. En fonction de la requête et de ses paramètres, une route sera déterminée.

businessLogic/scrap.ts

Ce fichier est le cœur du script de scraping. C'est lui qui va articuler la recherche de données, la récolte, l'enregistrement, etc.

Que fait la couche logique métier (BL – Business Logic) ?

Elle utilise les capacités de Puppeteer pour créer un objet ayant les capacités d'un vrai navigateur Chromium (avec ou sans GUI - pour la production par exemple). Nous fournissons une URL qui pointe vers la page web contenant les données à récolter.

Cela étant fait, Puppeteer va nous produire un objet ayant les capacités d'un navigateur. Nous l'appellerons le « browser object ». Nous pouvons alors utiliser tout le potentiel d'un vrai navigateur grâce au « browser object » et initialiser la création d'une nouvelle page.

Le reste de la logique découle de l'analyse réalisée au préalable. En fonction de la page parcourue, il faudra par exemple se connecter ou accepter les cookies. Puisque Puppeteer navigue sur le site tel que le ferait un être humain, l'analyse préalable aura défini la majorité de la logique.

xxxDAO.ts

Les fichiers suffixés par « DAO » représentent les entités qui serviront à accéder et manipuler les entités de la base de données.

classifiedMapper.ts

Pour ajouter un niveau d'abstraction, il est nécessaire de créer une correspondance entre les entités stockées dans la base de données et les objets reconstruits à partir des informations présentes sur une page web. Cela permet d'assurer une cohérence et une lisibilité de la structure de données.

Ce processus de correspondance est appelé « mapping ». Nous n'enregistrons pas directement les objets reconstitués sur base des informations présentes sur une page web afin de garder un contrôle sur ce que nous enregistrons en base de données.

Nous faisons donc la distinction entre les *pages/entities* qui représentent les objets présents sur les pages web et les *Mongo/entities* qui représentent les objets effectivement sauvés en base de données.

xxxPage.ts

Les fichiers suffixés par « Page » contiennent la logique métier relatif à la page récoltée. Par exemple, LoginPage.ts contient le code qui va retrouver et compléter le formulaire de login.

selectors/xxx.json

Ce dossier contient tous les fichiers de configuration représentant les éléments du DOM parcourus par le scraper. Par exemple, SearchingPage.json va contenir l'élément du DOM (son JS Path) qui contient la liste des petites annonces :

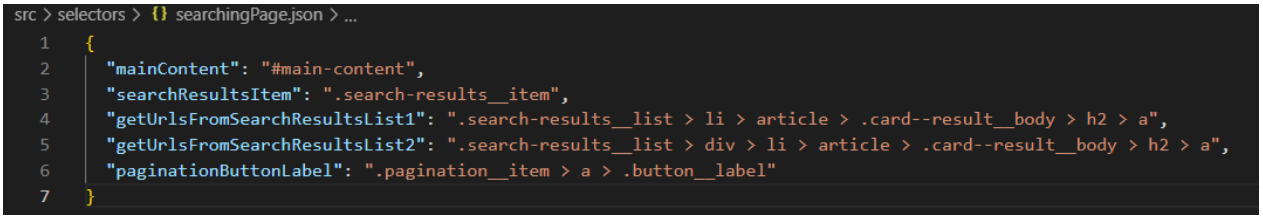


FIGURE 4 - searchingPage.json - Des exemples de JS Paths

Un JS Path ne doit jamais être écrit en dehors d'un fichier de configuration pour deux raisons. Premièrement, il est en effet susceptible d'être réutilisé à plusieurs endroits. Deuxièmement, la chaîne de caractères est compliquée et donc propice aux erreurs de frappe.

4.2.4.Détails d'implémentation et testing

Log

Puisque la source de données est susceptible d'évoluer à tout moment, nous sommes susceptibles d'avoir des récoltes qui ne sont pas complètes à 100%. En effet, il est possible que l'application ne soit pas toujours assez résiliente, et un bug provoquerait l'arrêt de l'exécution de la récolte.

Lorsque le scraper cherche un élément sur une page, il attend d'abord que la page soit entièrement chargée avant de parcourir le DOM. Si le scraper ne trouve pas l'élément recherché, un timeout est déclenché et nous enregistrons cet événement (message d'erreur) dans la base de

données. Nous pouvons ensuite analyser la page en question pour trouver l'élément qui a changé et apporter les modifications nécessaires au fichier de configuration ou au code source.

Verrou

Afin de s'assurer qu'une récolte n'est pas effectuée plusieurs fois par jour (cela génèrerait du trafic inutile sur la source), nous avons installé un système de verrou. Nous enregistrons dans la base de données une information qui nous indique si une récolte a déjà eu lieu ou non pour un jour donné.

Planning des récoltes

Afin de limiter au mieux notre impact sur la charge que peut subir la source, nous planifions la récolte aux heures creuses (aux alentours de 02:00AM). De plus, nous limitons la fréquence à une requête toutes les 1 à 2 secondes (par conteneur).

Mode debug

Lors de l'activité de programmation du scraper, il est intéressant de ne pas utiliser le mode « headless » de Chrome. Il faut alors préciser dans le fichier de configuration « .env.[ENV] » que l'on veut le mode debug. Grâce à cela, lorsque nous lançons l'application en local, nous pouvons voir un navigateur s'ouvrir et nous pouvons suivre le scraper parcourir les pages selon les algorithmes que nous avons codés.

4.3. Use case scenario pour une session de développement

Comment démarrer l'application ?

Pour lancer l'application, nous exécutons la commande suivante :

```
npm start
```

Cette commande démarre le serveur de développement, initialise un logger et récupère les variables d'environnement nécessaires. L'application est alors prête à fonctionner et écoute sur le port indiqué dans les variables d'environnement.

Quelle configuration pour les variables d'environnement ?

Avant de lancer l'application, nous nous assurons de configurer les variables d'environnement nécessaires. Ces variables comprennent les identifiants de connexion pour le site de petites annonces, la configuration de la base de données (chaîne de connexion et nom de la base de données) et le choix du port.

Comment débugger l'application ?

Pour débugger l'application, exécuter la commande suivante :

```
npm run dev
```

Cette commande démarre le serveur de développement et attache un débuggeur. L'application écoute sur le port spécifié dans les variables d'environnement. Pour déclencher un point d'arrêt, il suffit d'envoyer une requête http à un endpoint spécifique (par exemple, http://localhost:3000/api/scrap/...). Si un point d'arrêt est atteint, l'exécution passe en mode pas à pas.

La spécificité d'une application utilisant Puppeteer pour faire du scraping est qu'il existe deux contextes d'exécution. Nous pouvons distinguer Node.js qui exécute le code, et le navigateur instancié par Puppeteer. Le code qui est utilisé pour parcourir le site scrappé ne se debug pas avec des points d'arrêt mais avec l'instruction de débogage JavaScript « debugger; ». L'exécution du code s'arrêtera alors dans le navigateur.

Notons que l'application doit être compilée avant de pouvoir être débuguée.

Comment réaliser une récolte de données en local ?

Pour collecter des données une fois que l'application est en cours d'exécution en local, nous envoyons une requête paramétrée. Par exemple : http://localhost:[PORT]/api/scrap/immoweb?province=NAMUR

Comment créer une image Docker ?

1. Création du conteneur

Nous commençons par récupérer la solution depuis le repository GitHub et nous nous positionnons sur la branche master. Ensuite, Nous ajoutons le fichier « .env.dev » et nous renseignons les variables d'environnement nécessaires (pour rappel, ce fichier n'est pas versionné car il contient des informations sensibles). Après, nous spécifions le mode « production » dans les variables afin d'avoir une configuration headless, etc. Pour compiler l'application avec les configurations que nous venons de mettre en place, nous exécutons la commande suivante :

```
npm run build
```

Afin de créer une image de l'application et lui attribuer un tag, nous exécutons la commande suivante :

```
docker build -t felixbarzin/limo .
```

2. Nous poussons l'image sur le repo DockerHub via Docker Desktop

```
docker push felixbarzin/limo .
```

3. Nous testons l'image

Nous exécutons la commande suivante afin de lancer le conteneur sur le port 3000 :

```
docker run -d -p 3000/3000 limo
```

Nous ouvrons la console pour vérifier que l'application est bien en cours d'exécution et puis nous envoyons une requête à l'application pour vérifier que tout fonctionne correctement.

Comment déployer l'application sur Azure

- Créer une instance d'un container (Azure) Docker sur base de l'image qu'on a poussée sur le Docker Hub.
- Créer le workflow d'automatisation (Azure Logic App). Nous utiliserons le designer no code, comme illustré à la figure suivante.

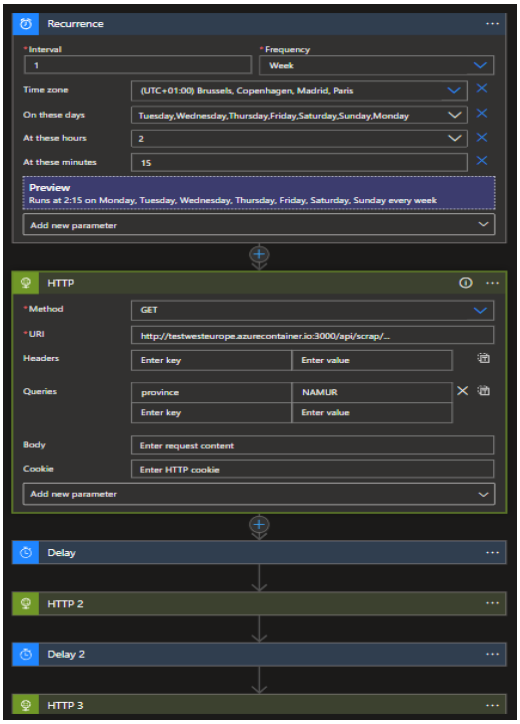


FIGURE 5 - Exemple d'un workflow sur une Logic App Azure

4.4. Limites et perspectives de ce système

4.4.1.Limites

Notre système présente plusieurs limites qu'il est important de prendre en compte.

1. Qualité des données

La qualité des données est la limite principale de cette méthode. Considérant que les données sont enregistrées à la source par des utilisateurs lambda, le risque d'erreur existe. Le traitement humain peut lisser ce risque mais il continuera d'exister.

Piste pour améliorer le système

Un moyen de limiter les erreurs présentes dans les petites annonces est de pouvoir traiter la version la plus à jour de la petite annonce. Pour cela, il faut donc revoir la méthode de sauvegarde des mises à jour des petites annonces que nous avons écrites. Actuellement, la source est collectée chaque jour et lorsqu'une référence connue est lue par le scraper, celui-ci vérifie si des changements ont été effectués. Lorsqu'une mise à jour est détectée, la petite annonce modifiée est entièrement sauvegardée à l'intérieur de la petite annonce originale.

Cette façon de faire ne pose pas de problème si on manipule les documents JSON via du JavaScript. En revanche, c'est un problème pour le traitement des informations avec nos outils de visualisation (MongoDB Charts et Tableau Desktop). Le document JSON n'est pas suffisamment plat pour être parcouru facilement. Nous ne pouvons donc pas profiter de ces mises à jour et nos outils consultent toujours la petite annonce d'origine.

Pour améliorer le système, il serait préférable d'enregistrer la petite annonce modifiée séparément et de faire une suppression logique de l'ancienne (booléen). L'aplatissage du document JSON présente l'avantage de conserver l'historique des modifications des petites annonces, si cela

est nécessaire, et permet de filtrer facilement les données avant de les transmettre aux outils de visualisation.

Notons aussi que les informations des petites annonces, même le plus à jour possible, ne sont pas forcément les informations finales. Par exemple, un prix affiché peut être négocié en face-à-face.

2. Sécurité de l'infrastructure

L'accès aux API pourrait être plus restrictif. Actuellement, nous avons une sécurité applicative (verrou) mais il serait mieux d'utiliser les outils proposés par Microsoft Azure. L'utilisation d'une API Gateway avec un Virtual Private Network devrait garantir un accès restreint à l'API de scraping. Cependant, ceux-ci ont un coût non négligeable et nous ne pouvons donc pas les utiliser avec le budget à notre disposition.

3. Technologie

Les librairies peuvent devenir obsolètes et créer de la dette technique[8]. De plus, nous ne sommes pas à l'abri de problèmes de performance ou de bugs.

4. Biais

Le traitement humain sur la clarification et le nettoyage des données peut entraîner des biais. Il est donc important de bien analyser les données récoltées et de choisir les outils de visualisation avec soin. Il ne faut pas oublier que même un choix de couleur dans un histogramme peut représenter un biais en soi.

5. Complexité

La mise en place de ce système demande des connaissances multidisciplinaires (programmation, base de données, infrastructure, expertise technique et analytique, …).

En prenant en compte ces limites et en travaillant à leur amélioration, nous pourrons optimiser le fonctionnement de notre système de scraping de petites annonces.

4.4.2.Perspectives

Le système peut être étendu pour fournir une vue plus complète du marché immobilier en ajoutant de nouvelles sources de données, comme les réseaux sociaux ou les données de géolocalisation.

[8] La dette technique est un concept en développement logiciel qui fait référence aux coûts cachés ou aux défauts de qualité d'un projet de développement qui se manifestent plus tard dans le processus de développement ou d'exploitation.

L’utilisation de l’analyse prédictive, comme le machine learning, peut aider à anticiper les tendances futures du marché immobilier.

De plus, le système peut être intégré à un système de support à la décision pour aider les planificateurs urbanistiques à prendre des décisions éclairées. Il peut également être étendu à d’autres marchés immobiliers, soit dans d’autres régions géographiques, soit dans d’autres secteurs de l’immobilier, tels que les biens commerciaux ou les propriétés résidentielles de luxe.

5. Collecte et prétraitement des données

5.1. Sources des données

La récolte de données est faite à partir d’un site Internet spécialisé dans les petites annonces de biens immobiliers, en vente ou en location. Le site est fiable et est une référence dans son domaine en Wallonie.

5.2. Qualité et nettoyage des données

5.2.1. Analyse rapide de la source des données

Le site Internet que nous avons ciblé pour notre étude propose une section dédiée aux petites annonces de biens immobiliers, qu’il s’agisse de ventes ou de locations. Bien que le site offre également d’autres services, nous nous concentrerons exclusivement sur cette section.

Les petites annonces sont accessibles gratuitement sur le site, mais certaines informations ne sont pas disponibles pour les utilisateurs anonymes. Pour visualiser l’ensemble des informations publiques, il est nécessaire de créer un compte et de se connecter. Le scraper devra être capable de se connecter.

L’analyse des données provenant du site Internet nécessite une approche de retro engineering de base de données. C’est en analysant les requêtes du trafic réseau que nous pouvons avoir une idée de ce à quoi ressemblent les objets et que nous pouvons reconstituer les relations entre les données.

5.2.2. Structure type des données extraites

Voici un extrait de la donnée récoltée. Le JSON complet fait plus de 1200 lignes et ne peut donc pas se trouver ici. Vous trouverez en annexe (*12.1. Exemple complet d’un enregistrement JSON*) un exemple complet d’un enregistrement JSON. Vous trouverez en annexe (*12.2. Détails des clés/valeurs*) une explication détaillée de chaque paire clé/valeur d’un enregistrement.

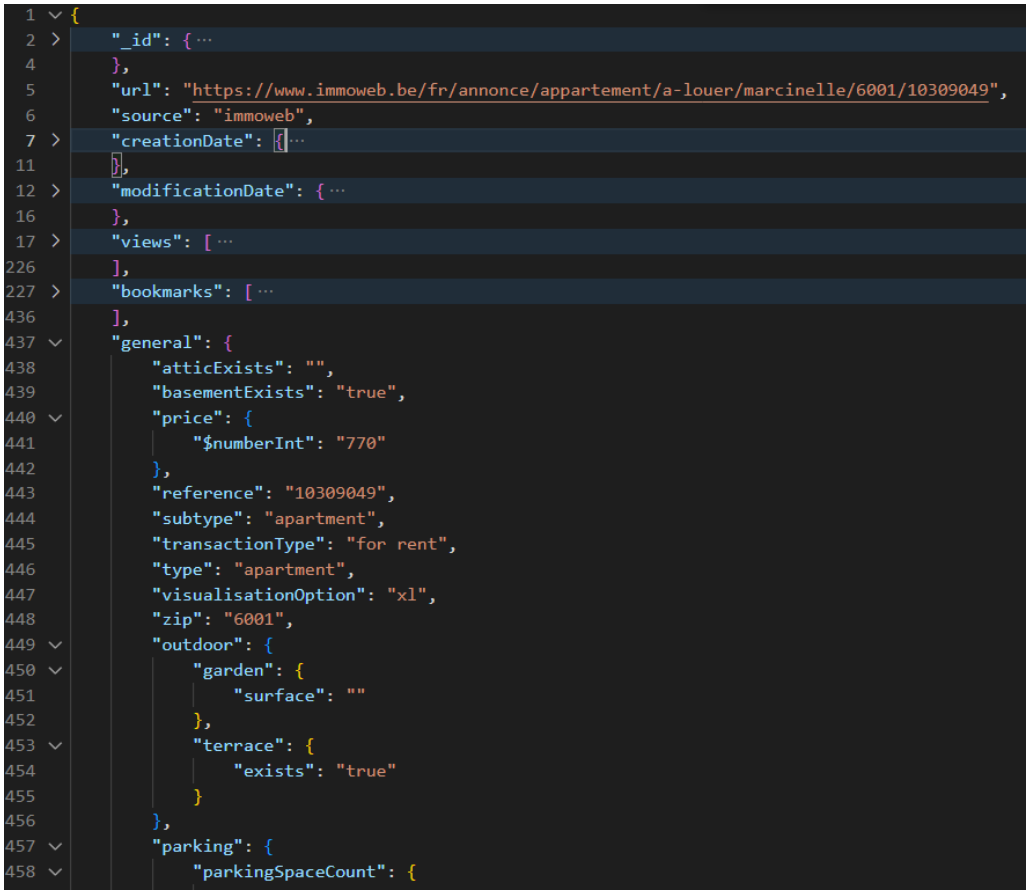


FIGURE 6 – Extrait d'un enregistrement

La structure des données extraites peut varier selon les annonces. Par exemple, une petite annonce pour une maison à vendre ne contiendra pas d'informations sur le coût de la location. Les enregistrements peuvent donc être classés en différentes catégories en se basant sur les propriétés « type d'immeuble » et « type de transaction ».

Type d'immeuble

Type principal	
Apartment	House
Sous-type	
Apartment	House
Duplex	Villa
Ground floor	Apartment block
Penthouse	Exceptional property
Loft	Mixed use building
Flat studio	Mansion
Triplex	Country cottage
Kot	Farmhouse

Service flat	Castle Town house Other property Bungalow Chalet Manor house

Type de transaction

Type de transaction	
For rent	For sale
Sous-type	
Rent regular	Buy regular Public sale Life annuity

Nous pouvons définir quatre grandes catégories en combinant ces différents types :

1. Les appartements en location
2. Les appartements à vendre
    a. Les appartements à vendre en vente régulière
    b. Les appartements à vendre en vente publique
    c. Les appartements à vendre en rente viagère
3. Les maisons en location
4. Les maisons à vendre
    a. Les maisons à vendre en vente régulière
    b. Les maisons à vendre en vente publique
    c. Les maisons à vendre en rente viagère

Les catégories qui vont nous intéresser pour ce mémoire sont les catégories 1 et 4.a.

5.2.3.Traitement des données

5.2.3.1. Format d'enregistrement et stockage

Considérant que nous ne pouvons pas contrôler la source des données, nous décidons de stocker les informations extraites dans une base de données résiliente face aux éventuels changements, notamment dans la structure des entités. Pour cela, une base de données NoSQL est préférable car elle offre une grande flexibilité en termes de schéma de données.

Nous avons opté pour le format JSON, qui est adapté à la structure de nos données. De plus, il a l'avantage d'être facilement manipulable et est facile à comprendre, ce qui convient parfaitement pour avoir plus d'aisance dans la maintenance.

5.2.3.2. Définition, classification et sous-classes théoriques

Dans cette section, nous allons discuter des différentes classes de biens immobiliers que nous pouvons rencontrer dans notre analyse, en nous concentrant sur les maisons et les appartements. Étant donné que ces deux types de propriétés peuvent être subdivisés en différents sous-types, nous devons définir les critères pertinents pour notre analyse qui nous permettront de distinguer les maisons des appartements de manière formelle.

Dans le contexte d'une analyse de petites annonces immobilières, il est essentiel de définir clairement les critères qui permettent de distinguer une maison d'un appartement. Cela nous permettra de déterminer les caractéristiques communes à chaque catégorie de biens et de les analyser en conséquence.

Classe « maison »

Une maison est une propriété immobilière généralement conçue pour servir de résidence principale ou secondaire pour une famille ou un ménage (achetée ou louée), et ce, pour y habiter de façon permanente ou à long terme. Elle comporte des caractéristiques essentielles telles qu'une cuisine, une salle de bain, une chambre et une salle à manger. Une maison est un bien autonome et individuel, distinct des autres propriétés immobilières voisines, avec sa propre entrée et ses limites clairement définies. Les caractéristiques physiques d'une maison indiquent qu'elle a au moins une façade et peut aller jusqu'à 4 façades. Notons que 4 façades signifie qu'aucune autre habitation n'est collée physiquement. Une maison est construite sur un terrain d'au moins 10 m². Cette valeur dépend des plans d'aménagements, des règlements communaux et des lotissements ([27]).

Une maison n'est pas un château.

Un château est un type de propriété plus grand, aux fonctions diverses et historiques. Les caractéristiques distinctives du château sont un grand nombre de pièces, des tours, des remparts, des douves, et autres caractéristiques architecturales complexes.

Une maison n'est pas un gîte.

Un gîte est un type de location de vacances qui est généralement loué pour une période plus courte. L'utilisation que l'on en fait ne correspond pas à l'utilisation d'une maison.

Une maison n'est pas une propriété exceptionnelle.

Les propriétés exceptionnelles peuvent être très différentes des maisons ordinaires en termes de prix, de caractéristiques physiques, de statut juridique ou d'autres facteurs.

En excluant les propriétés exceptionnelles de notre analyse, nous nous assurons que les propriétés que nous examinons sont comparables entre elles et qu'elles répondent aux critères spécifiques que nous avons définis pour une maison. Cela nous permet de faire des comparaisons plus précises et plus significatives entre les différentes maisons et d'obtenir des résultats plus fiables.

Une maison n'est pas une ferme.

Une ferme est généralement définie comme une propriété rurale utilisée à des fins agricoles. Les caractéristiques physiques et fonctionnelles sont différentes.

Une maison n'est pas un manoir.

Un manoir est généralement défini comme une grande et luxueuse résidence de campagne, souvent associée à une grande propriété. Les manoirs peuvent avoir des caractéristiques architecturales, des terrains et des équipements qui sont différents de ceux d'une maison ordinaire, comme des jardins formels, des terrains de golf, des piscines intérieures, des courts de tennis et d'autres équipements de luxe.

Une maison n'est pas une villa.

Une villa est un type de maison qui est souvent associé à une grande maison de vacances située dans un environnement pittoresque ou une station balnéaire. De plus, les villas ont souvent des caractéristiques spécifiques telles que des jardins luxuriants, des piscines ou des terrasses spacieuses, qui les distinguent des maisons ordinaires.

- ⇨ Pour résumer, la donnée enregistrée est une maison lorsque son type général est « maison » et que son sous-type est soit « maison », « maison de ville » (town house), « chalet », « bungalow », « maison de maître » (manor house).
- ⇨ Les sous-classes sont : « maison à vendre » et « maison à louer ». La définition de classe est la même. Seul le type de transaction change.
- ⇨ En sommes, nous pouvons construire ce filtre (requête MongoDB) :{ "general.price": { $gt: 0 }, "general.subtype": { $nin: ["apartment block", "other property", "mixed use building", "castle", "country cottage", "exceptional property", "farmhouse", "manor house", "mansion", "villa"] }, "general.transactionType": "for sale", "general.type": "house", "property.location.province": "Namur", "transaction.subtype": "BUY_REGULAR", $or: [{ "property.building.facadeCount": { $eq: null } }, { "property.building.facadeCount": { $gt: 0, $lte: 4 } }], "property.bedroomCount": { $gt: 0, $lte: 10 }}

Classe « appartement »

Un appartement est une unité de logement qui se compose généralement d'une ou plusieurs pièces dans un immeuble ou une résidence. Les appartements sont souvent situés dans des zones urbaines ou suburbaines et peuvent être loués ou achetés. Contrairement à une maison individuelle, les appartements partagent souvent des murs avec d'autres unités de logement et peuvent avoir des installations communes telles qu'une buanderie, une salle de sport, une piscine ou un espace de stationnement commun. Les appartements peuvent être de tailles différentes, allant des studios pour une personne à des unités à plusieurs chambres pour les familles.

Un appartement n'est pas une seigneurie (appartement de vie assistée).

Un appartement avec services est un type de logement conçu pour les personnes âgées ou les personnes handicapées qui souhaitent vivre de manière indépendante mais ayant besoin d'une aide pour les activités quotidiennes. Ces appartements offrent généralement des services supplémentaires tels que des repas, du ménage, du transport et des soins médicaux.

Un appartement n'est pas un kot.

Un kot est un logement étudiant, généralement situé à proximité d'un établissement d'enseignement supérieur, et pouvant être partagé avec d'autres étudiants. Les kots peuvent être meublés ou non et leur durée de location est souvent limitée à la durée des études. Les kots sont donc conçus pour répondre aux besoins spécifiques des étudiants, tandis que les appartements peuvent convenir à différentes catégories de personnes, qu'il s'agisse de familles, de couples ou de personnes seules.

- ⇨ Pour résumer, un enregistrement est un appartement lorsque son type général est « appartement » et que son sous-type est soit « appartement » (apartment), « duplex », « rez-de-chaussée » (ground floor), « studio » (flat), « penthouse », « loft », « triplex ».
- ⇨ Les sous-classes sont : « appartement à vendre » et « appartement à louer ». Le type de transaction est différent. Cela implique que l'appartement à louer n'a pas de prix de vente mais a un prix de location et des charges locatives. Le système de charges peut varier en fonction des règles de copropriété en vigueur. En général, chaque appartement est équipé d'un compteur électrique individuel qui mesure la consommation d'électricité de l'appartement et est géré par le locataire (il choisit son fournisseur, etc). Il existe les charges communes qui sont partagées (éclairage commun, ascenseur, production d'eau chaude, …). Ces frais communs sont répartis entre les occupants de chaque appartement, généralement en fonction de la taille de l'appartement.
- ⇨ En sommes, nous pouvons construire ce filtre (requête MongoDB) : { "general.price": { $gt: 0 }, "general.subtype": {$nin: ["service flat", "kot"] }, "general.transactionType": "for rent", "general.type": "apartment", "property.location.province": "Namur", "transaction.subtype": "RENT_REGULAR",}

Nous pourrions également définir les classes « maison de luxe » ou « kot », mais ces classes ne seront pas abordées au cours de la rédaction de ce mémoire.

5.2.3.3. Filtres et exclusions

Dans le cadre de la collecte de données, il est important de s'assurer que les valeurs collectées sont valides et utiles pour l'analyse. Pour ce faire, nous devons appliquer des filtres sur le jeu de données avant l'analyse pour ne garder que les données pertinentes. Par exemple, si l'analyse porte sur les maisons à vendre, nous filtrerons le type de transaction car les ventes en viager et par enchères ne nous intéressent pas.

Les valeurs

Il est important d'exclure les données erronées ou inutiles. Puisque les données récoltées sont des données entrées par un utilisateur lambda, il est important de faire un contrôle de la qualité et un nettoyage si besoin.

Par exemple, il peut être nécessaire de filtrer les annonces des maisons n'ayant pas de chambres, puisque notre classe théorique affirme qu'une maison doit avoir au moins une chambre.

Les types de transactions

Pour assurer l'homogénéité des données, nous avons décidé de travailler uniquement sur les ventes de type « BUY_REGULAR » (achat normal) et les locations de type « RENT_REGULAR » (location normale).

Nous avons choisi d'exclure les ventes de sous-type « LIFE_ANNUITY » (rente annuelle), qui correspondent à un régime de vente particulier impliquant la spéculation et dont les données ne sont pas comparables aux autres. De même, nous avons exclu les ventes de sous-type « PUBLIC_SALE », qui sont des enchères en ligne dont le montant final n'est pas déterminé à l'avance et dont les données ne sont pas comparables aux autres.

Les sous-types

Certains sous-types doivent également être exclus : les maisons de sous-type « bloc à appartement » (apartment block), « autres » (other property) et « bâtiment à usage multiple » (mixed use building).

Ces sous-types ne correspondent à aucune de nos classifications et ne sont donc pas comparables.

Autres

Exclusion des enregistrements dont la province est différente de « Namur ».

5.2.3.4. Dépersonnalisation

Pour garantir la confidentialité des utilisateurs, nous avons décidé de ne rien collecter qui pourrait être considéré comme étant à caractère personnel. Cependant, nous ne pouvons pas garantir que des informations n'ont pas été renseignées dans les champs libres. Nous partons du principe que les utilisateurs ont respecté l'usage des formulaires à leur disposition.

Nous avons également pris soin de ne pas collecter les photos, qui pourraient contenir des informations sensibles.

5.2.3.5. Trier les doublons

Un doublon peut se produire lorsque deux enregistrements différents contiennent des informations identiques ou presque. Est-ce que deux annonces identiques en tout point, à l'exception d'un mot dans la description doivent être considérées comme doublon ? Il est important de définir ce que nous considérons comme être un doublon ou non, afin d'avoir le jeu de données le plus juste possible.

Les cas de doublons possibles sont :

1. Une petite annonce arrive à échéance. Son propriétaire décide de poster une nouvelle annonce identique plutôt que de prolonger l'annonce existante.

Ce n'est clairement pas un cas qui doit se produire. Cependant, l'utilisateur étant libre de faire ce qu'il veut, nous devons l'envisager.

2. Un utilisateur décide de poster plusieurs fois la même annonce.

Ce n'est pas un cas d'utilisation normale de la plateforme. La fonctionnalité de recherche est performante, et l'utilisateur n'a pas besoin de trucs et astuces pour améliorer sa visibilité sur la plateforme de cette manière. En plus d'être dotée d'une fonctionnalité de recherche performante, la plateforme met à disposition des fonctionnalités pour les utilisateurs souhaitant améliorer leur visibilité (image plus grande, etc.).

3. Un utilisateur veut changer le prix de sa petite annonce, mais ne veut pas que les visiteurs le remarquent. Il décide alors de supprimer sa petite annonce et d'en créer une nouvelle.

C'est une technique courante des agences immobilières. Elle a plusieurs avantages en plus de cacher aux visiteurs que le prix a changé : la bannière « nouveau » sera affichée et les utilisateurs ayant des alertes recevront à nouveau la petite annonce dans leur boîte mail.

Une méthode d'identification et d'élimination des doublons

La méthode consisterait à faire tourner un job sur la base de données après une récolte et annoter les enregistrements étant des doublons potentiels. Le job aurait pour mission de comparer toutes les annonces entre elles et pour chaque propriété. Les champs textuels tels que la description devraient subir une analyse plus poussée pour déterminer quelle tolérance on accorderait afin de déterminer si une description est équivalente ou non.

Nous voyons immédiatement qu'un tel travail serait fastidieux. Or, comme nous l'avons constaté dans les cas énoncés 1 et 2 (et 3 dans une autre mesure), la probabilité de se retrouver avec des doublons est plutôt faible pour une simple raison : les petites annonces sont payantes, ce qui pourrait réduire l'intérêt pour un particulier (et pour une agence) de dupliquer son annonce.

Nous décidons alors que les doublons sont négligeables. Il existe tout de même une protection essentielle que nous avons mise en place. Pour éviter l'enregistrement de doublons, nous vérifions avant l'enregistrement des données que l'identifiant attribué à la petite annonce n'est pas déjà connu dans notre base de données. Pour rappel, nous avons un système qui enregistre les modifications apportées aux petites annonces et nous vérifions donc toujours systématiquement si une annonce n'existe pas déjà. Bien que cette méthode ne permette pas de détecter tous les cas de doublons, elle est la plus fiable et efficace que nous pouvons mettre en œuvre compte tenu de la complexité du problème.

5.3. Intégration et transformation des données

Intégration

À ce stade de l'étude, nous n'envisageons pas de procéder à une combinaison de données depuis des sources multiples. Notre source étant fiable et très riche en données, nous considérons notre jeu de données comme étant suffisant.

L'avantage de cette situation est que nous ne devons donc pas analyser la compatibilité des données, ainsi que leurs éventuelles inconsistances et conflits. Les tâches de nettoyage et de standardisation se voient donc simplifiées.

En revanche, nous utiliserons d'autres sources de données pour valider les résultats (*6. Validation des résultats*).

Transformation

La structure de la donnée que nous enregistrons est identique à la donnée présente sur la page web, à quelques exceptions près. En effet, nous ajoutons certaines métadonnées comme :

- L'URL : permet de faire des vérifications. En se rendant à l'URL indiquée, nous pouvons vérifier que les données présentes sur la page sont bien enregistrées, par exemple.

- La source : si nous voulons ajouter un deuxième site Internet comme source de données, nous pourrons facilement trier les données.

Nous avons également décidé d'organiser un système d'historisation de l'enregistrement. Effectivement, lors des collectes journalières, si nous détectons qu'un identifiant est déjà présent dans notre base de données, notre script de scraping va alors comparer chacune des informations. Si un changement est détecté, une copie de l'enregistrement modifié sera sauvegardée à l'intérieur de l'enregistrement courant.

5.4. Analyse exploratoire de données (AED)

5.4.1. Introduction

L'analyse exploratoire de données (AED) est une étape cruciale pour comprendre les caractéristiques du jeu de données et identifier les patterns et relations qui peuvent s'y trouver. Cette compréhension approfondie nous aidera à développer des hypothèses solides qui guideront notre analyse.

Dans le cadre du marché immobilier, la spéculation et les événements extérieurs exercent une influence directe sur les composants du marché. Par exemple, une loi restrictive sur le nombre de façades que peut posséder un bien à une influence directe sur un élément constituant un composant du marché. On peut facilement imaginer que la variable « nombre de façades » est importante puisqu'elle changerait non seulement le paysage urbain, mais aussi la façon dont vont cohabiter les gens. Le nombre de façades est donc une variable d'intérêt. À l'inverse, des variables comme « l'orientation de la terrasse », « l'accessibilité du grenier », « le nom de la propriété » ont peu ou pas d'intérêt.

En poursuivant cette réflexion, nous pouvons déterminer rapidement une première sélection de variables d'intérêts telles que le prix, la localisation, la taille de la propriété, le type de propriété, le nombre de chambres, etc.

Propriété	Description rapide
general.price	Le prix demandé
general.subtype	Le sous-type auquel appartient le bien annoncé
general. transactionType	Le type de transaction (louer ou acheter par exemple)
general.type	Le type du bien auquel appartient le bien annoncé
general.outdoor.garden.surface	La surface du jardin exprimée en mètre(s) carré(s)
general.outdoor.terrace.exists	Indique la présence ou non d'une terrasse
general.parking.parkingSpaceCount.indoor	Le nombre de parking intérieur

general.parking.parkingSpaceCount.outdoor	Le nombre de parking extérieur
general.bedroom.bedroomCount	Le nombre de chambres
general.building.condition	Décrit l'état du bâtiment selon une nomenclature
general.building.constructionYear	L'année de construction du bien
general.land.surface	La taille du terrain exprimée en mètre(s) carré(s)
general.enery.heatingType	Le type d'installation pour le chauffage
property.building.facadeCount	Indique le nombre de façades du bien
property.netHabitableSurface	Mesure de la surface nette habitable en mètre(s) carré(s)
property.land.sewerConnection	Indique si présence ou non d'une connexion aux égouts
property.energy.hasHeatPump	Indique si présence ou non d'une pompe à chaleur pour le chauffage
property.energy.hasPhotovoltaicPanels	Indique si présence ou non de panneaux photovoltaïques
transaction.certificates.epcScore	Indique le score du PEB
transaction.certificates .primaryEnergyConsumptionPerSqm	Indique le niveau de consommation d'énergie primaire par mètre(s) carré(s)

En réalisant des statistiques descriptives et des graphiques de visualisation sur ces variables sélectionnées, nous pourrons mieux comprendre la distribution des données, les valeurs extrêmes, les moyennes, les médianes et les modes. Cette approche nous permettra d'identifier les tendances, les relations et les anomalies qui peuvent influencer la prise de décision.

L'AED est donc un outil qui va nous permettre d'affiner notre jeu de données en identifiant les variables pertinentes et en détectant les problèmes de qualité des données telles que des valeurs aberrantes ou des données manquantes. En utilisant cette approche, nous pourrons améliorer notre compréhension des données.

Les jeux de données sont disponibles pour le corps enseignant sur simple demande à l'auteur de ce document.

5.4.2. AED de la sous-classe « maison à vendre »

Le jeu de données ayant servi de base pour cet AED s'appelle « *dataset house-for-sale.csv version V1 »*. Il concerne des récoltes quotidiennes ayant été réalisées entre le 01-01-2023 et le 30-04-2023, soit quatre mois de données.

Voici un récapitulatif des différents filtres et exclusions qui ont été appliqués pour produire ce jeu de données :

- Le prix doit être strictement plus grand que zéro ;
- Le sous-type ne peut pas être l'un de ces sous-types : apartment block, other property, mixed use building, castle, country cottage, exceptional property, farmhouse, manor house, mansion, villa ;
- Le type de transaction est une vente et le sous-type de transaction est de type « achat régulier » ;
- Le type général est « maison » ;
- Le bien se situe dans la province de Namur ;
- Si le nombre de façades est indiqué, il doit être compris entre 0 et 4 ;
- Le nombre de chambres doit être plus grand que zéro et plus petit que 10.

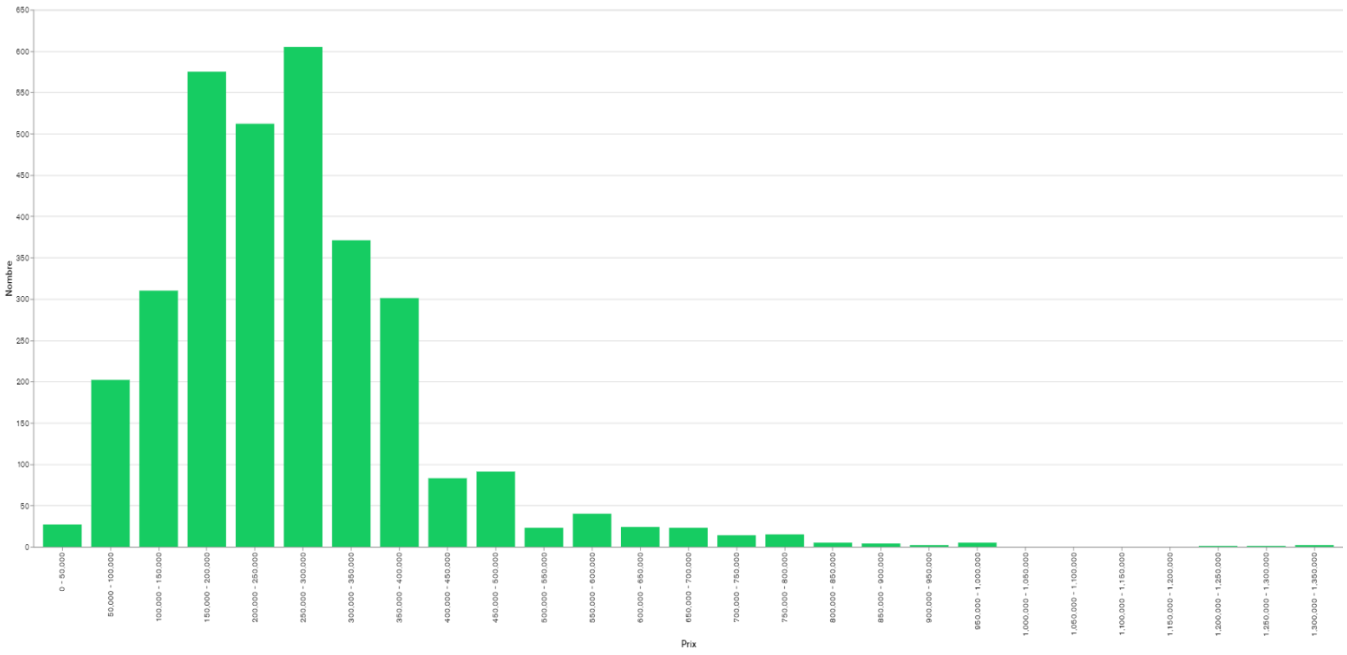
5.4.2.1. Distribution

Nous allons analyser les variables de manière indépendante pour comprendre leur distribution et ensuite leurs relations avec les autres variables.

Comprendre la distribution de la variable cible : le prix

Histogramme

L'histogramme que nous allons observer permet de visualiser la distribution des valeurs d'une variable, en l'occurrence le prix des maisons. Nous avons regroupé les prix par tranche de cinquante mille euros. L'histogramme montre une distribution asymétrique droite, donc avec une concentration des données sur la partie gauche.



Le mode, c'est-à-dire la valeur la plus fréquente, se situe dans la tranche 250.000 – 300.000 euros. Les tranches 150.000 – 200.000 euros et 200.000 – 250.000 euros sont également très fréquentes, ce qui indique que les prix sont répartis de manière plutôt équilibrée autour du mode. Nous observons que l'histogramme présente une longue queue vers la droite, ce qui indique qu'il y a un petit nombre de maisons vendues à des prix très élevés. Cette queue peut être causée par des valeurs extrêmes qui sont très éloignées de la moyenne. Il est important de prendre en compte ces valeurs extrêmes lors de cette analyse de la distribution des prix de vente des maisons.

Pour nous aider à comprendre la distribution, nous allons utiliser un boxplot et sa capacité à fournir un bon résumé statistique. Le boxplot est un outil graphique qui permet de visualiser la distribution d'une variable de manière synthétique. Il est composé d'une boîte représentant le premier et le troisième quartile, avec une barre à l'intérieur indiquant la médiane. Les moustaches qui s'étendent à partir de la boîte montrent l'étendue de la distribution, c'est-à-dire la plage de valeurs où se trouvent la plupart des données. Les valeurs aberrantes, qui sont les données situées en dehors de l'étendue de la distribution, sont représentées par des points isolés.

En utilisant un boxplot pour notre analyse, nous pouvons ainsi observer les quartiles et la médiane des prix de vente des maisons, ainsi que la présence de valeurs aberrantes. Cette information est très utile pour comprendre la distribution des prix et déterminer s'il y a des différences significatives entre les différents quartiles. En effet, si les moustaches sont très éloignées l'une de l'autre, cela peut indiquer la présence de valeurs aberrantes ou de sous-groupes différents dans notre population d'étude.

L'utilisation d'un boxplot est un moyen efficace pour fournir un résumé statistique de la distribution d'une variable. En l'occurrence, il nous permettra de mieux comprendre la distribution des prix de vente des maisons étudiées et de déterminer s'il y a des valeurs aberrantes ou des différences significatives entre les différents quartiles.

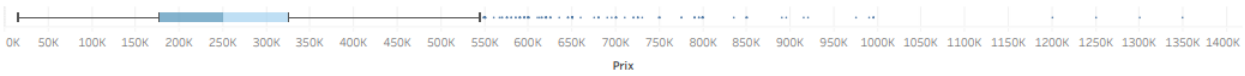


FIGURE 7 - Boxplot du prix de vente des maisons

Détails du boxplot

Barre supérieure	545.000
Charnière supérieure	325.000
Médiane	250.000
Charnière inférieure	177.310
Barre inférieure	15.000

Résumé statistique

Moyenne	266.229,60
Médiane	250.000
Maximum	1.349.000
Minimum	15.000

En examinant le boxplot, nous avons constaté la présence de valeurs aberrantes, confirmant nos observations précédentes sur l'histogramme. Cependant, après une inspection plus approfondie, nous avons identifié que certaines de ces valeurs étaient dues à des imprécisions introduites par les propriétaires de petites annonces. Par exemple, bien que les annonces soient classées comme « maison » (house), les descriptions et les autres variables semblent correspondre plutôt à des « blocs d'appartements » (apartment blocks), que nous avons exclus de notre description théorique de classe.

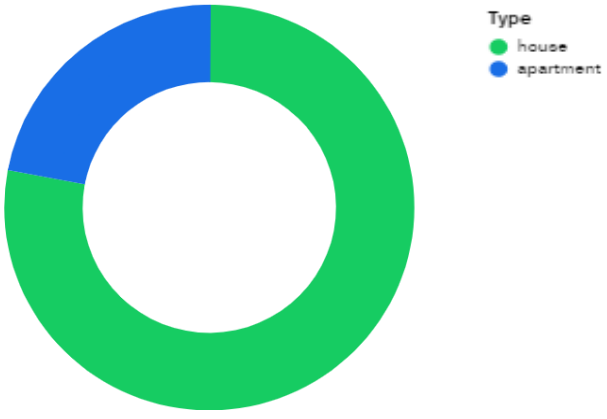
Cette analyse exploratoire de données nous a permis de détecter ces anomalies et de réfléchir à des solutions pour les traiter. Pour améliorer notre classification, nous prévoyons de créer des sous-catégories pour les maisons en fonction de leur prix. Pour déterminer des valeurs pertinentes pour effectuer cette segmentation, nous envisageons une analyse de cluster qui permettra de regrouper les maisons en fonction de leurs caractéristiques communes.

En fin de compte, mieux comprendre la distribution nous permet d'améliorer notre classification et d'éliminer les valeurs aberrantes introduites.

Exploration des variables catégoriques

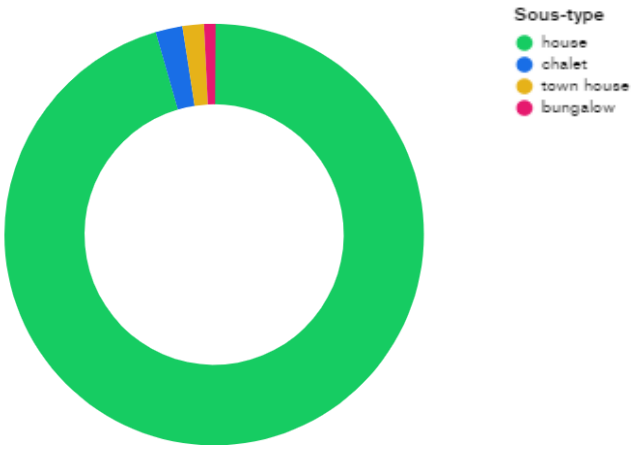
Le type de bien

Nous observons que 77.9% des enregistrements concernent des maisons à vendre et 22.1% concernent des appartements à vendre. Cette information nous apprend que les maisons sont plus prisées que les appartements dans la province de Namur. Nous ne pouvons pas tirer beaucoup de conclusions sur base de cette information, mais elle nous permet néanmoins de faire naître des questions et des hypothèses et de mieux appréhender notre jeu de données.



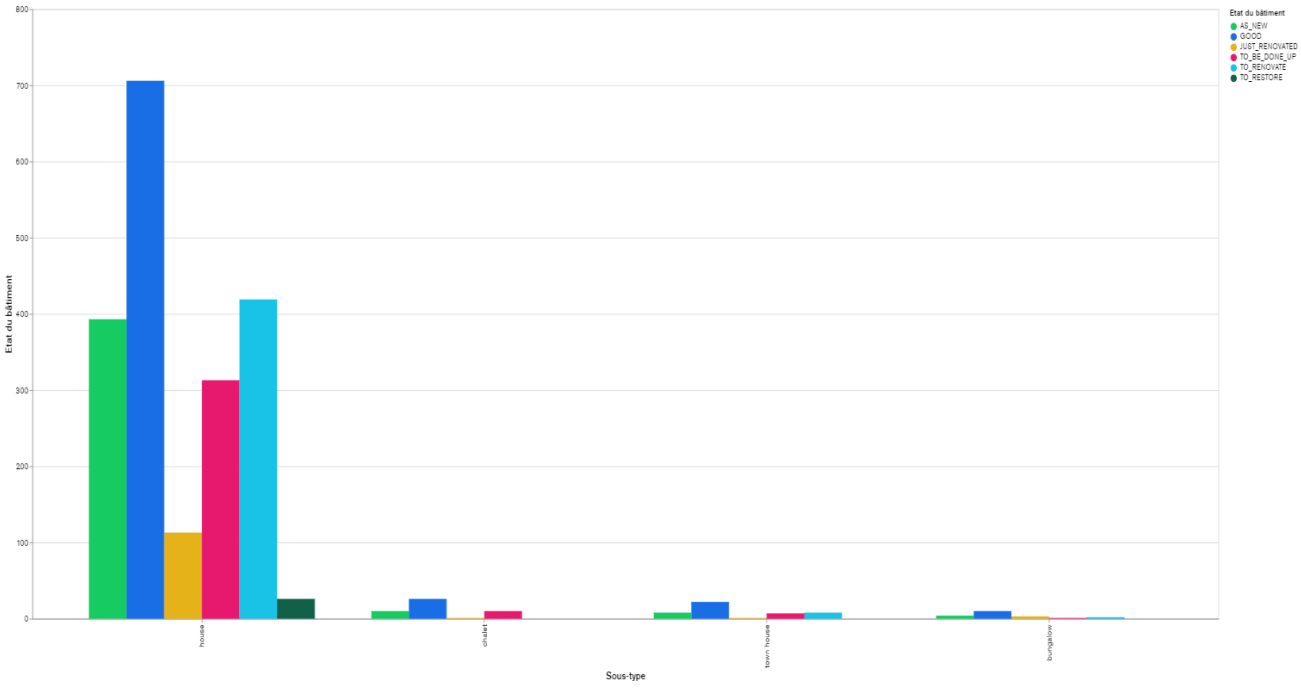
Le sous-type de bien

Parmi les maisons à vendre, nous retrouvons 0.9% de « bungalows », 1.7% de « maisons de ville » (town house), 2.1% de « chalets » et 95.4% de « maisons » (house). Il serait intéressant de croiser cette information avec une carte géographique pour savoir si des sous-types de biens sont répartis sur tout le territoire ou non.

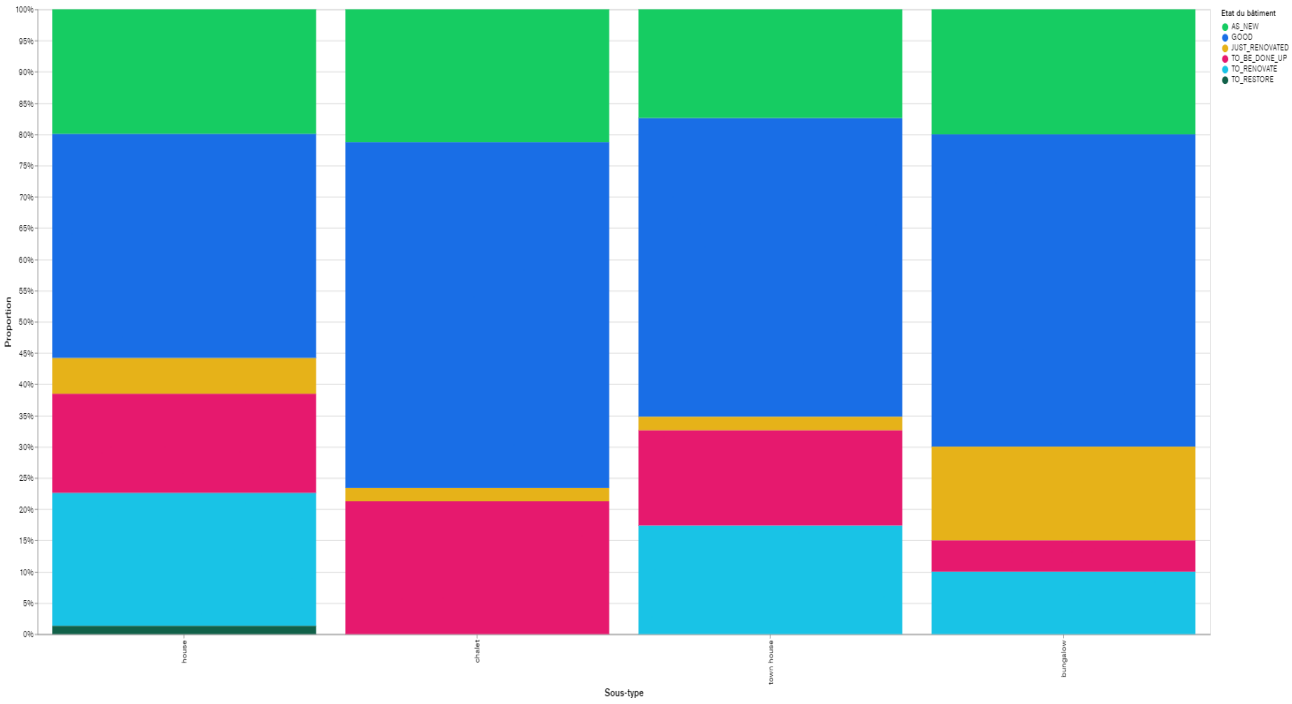


L'état du bâtiment

Pour chacun des sous-types de maison, l'état du bâtiment annoncé prédominant est « bon » (GOOD). Ensuite, on compte plus de maisons « à rénover » (TO_RENOVATE) que de maisons « comme neuves » (AS_NEW).



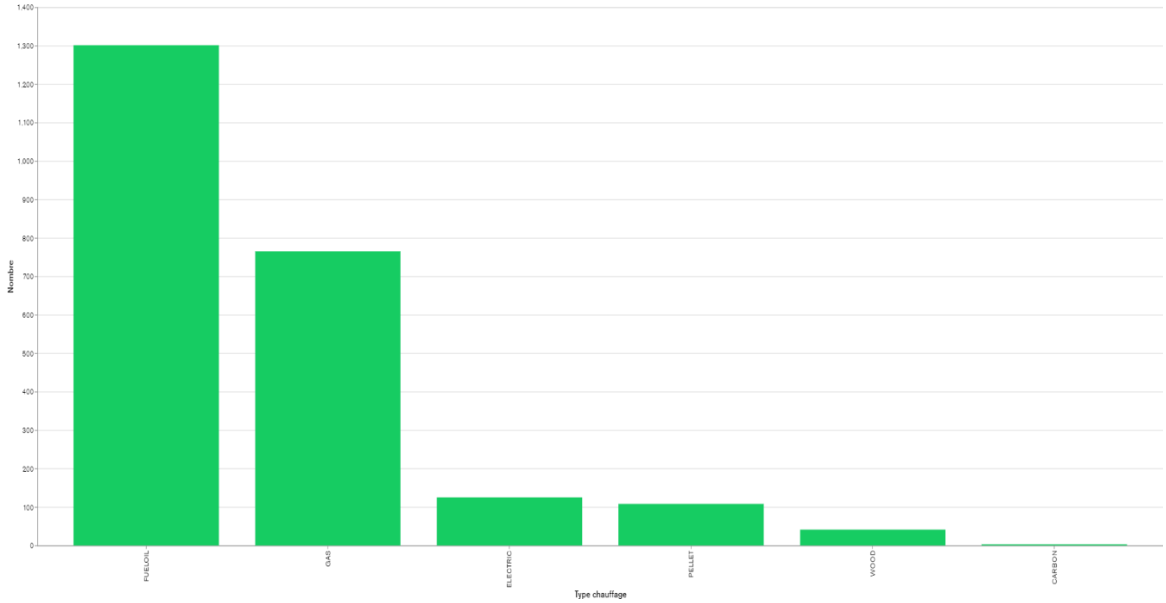
Puisque les sous-types sont peu représentés, nous allons utiliser un graphique en colonnes empilées pour visualiser les proportions relatives.



Proportionnellement, ce sont les bungalows qui sont dans le meilleur état (as new + good + just renovated). Viennent ensuite les chalets, les maisons de villes et puis seulement les maisons.

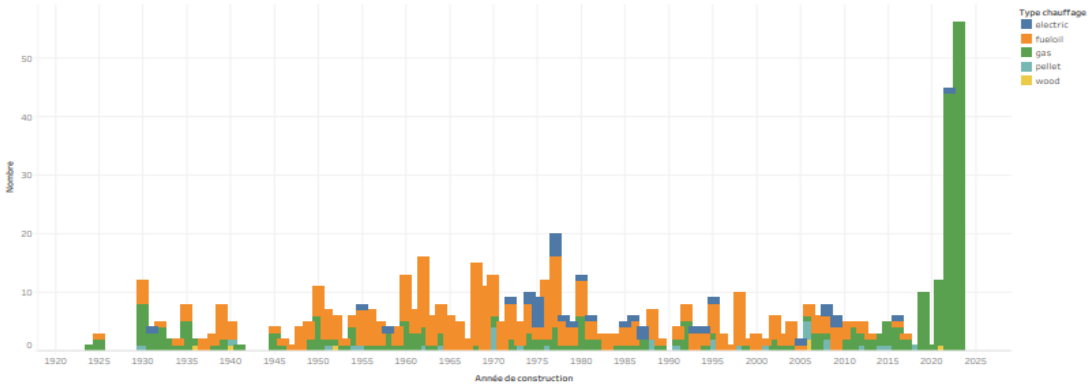
Système de chauffage

Le système de chauffage le plus répandu (le mode) dans les maisons en vente est le mazout, suivi du gaz.



Nous voulons tirer un peu plus d'informations de cette donnée. Nous pouvons, par exemple, la croiser avec l'année de construction du bâtiment.

Depuis 2019, toutes les nouvelles constructions sont équipées d'un chauffage au gaz. Nous émettons l'hypothèse que ceci est probablement dû à la prochaine règlementation visant à interdire les installations au mazout [28].



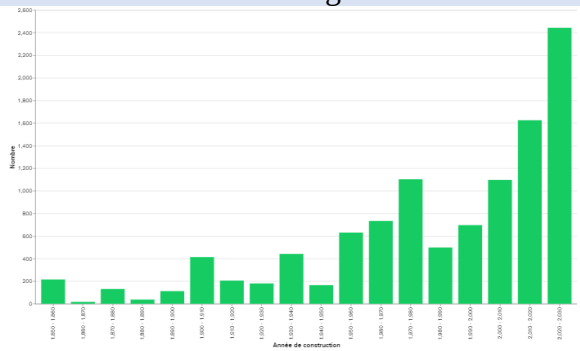
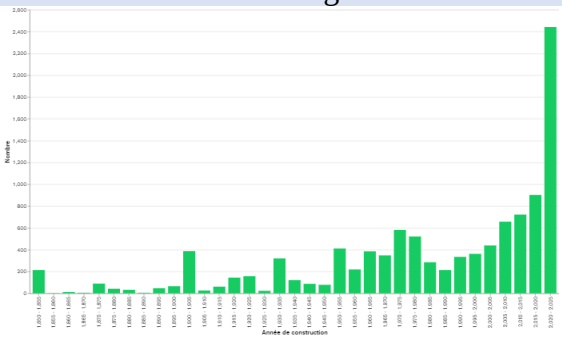
Exploration des variables continues

Maintenant que nous avons exploré les variables catégoriques, nous allons nous intéresser aux variables continues telles que les surfaces. L'analyse de ces variables nous permettra de mieux comprendre leur répartition dans notre ensemble de données.

Année de construction du bâtiment

Nous pouvons voir dans l'histogramme avec un binning[9] de 10 que l'année de construction la plus représentée est très récente, ce qui se confirme lorsque nous réduisons la taille du binning à 5.

L'histogramme montre que les valeurs sont concentrées du côté droit du graphique, ce qui indique une distribution asymétrique positive ou une distribution à gauche.

Binning 10	Binning 5
	

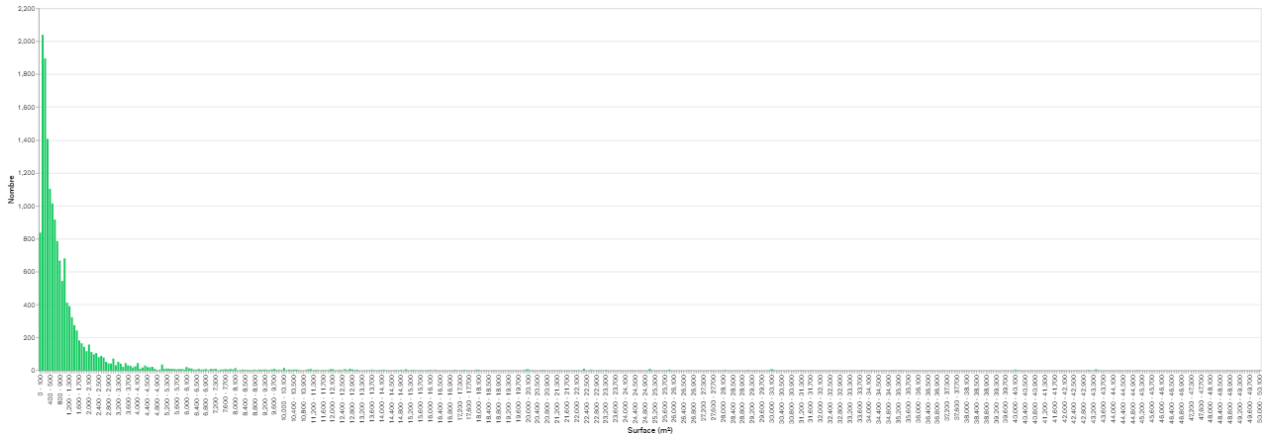
[9] Le binning, également appelé discrétisation, est une technique de prétraitement de données qui consiste à regrouper des valeurs continues en plusieurs catégories (ou bins) discrètes. Cette méthode est utilisée pour réduire le bruit dans les données, simplifier les calculs, faciliter la visualisation et l'analyse des données.

Cette observation suggère qu'un grand nombre de nouvelles maisons ont été mises sur le marché au cours des dernières années, ce qui témoigne d'une activité immobilière intense dans la région de Namur. L'analyse exploratoire de données révèle ici que la plupart des bâtiments mis en vente sont plutôt récents, toutefois, il convient de rester prudents dans nos conclusions et de poursuivre notre exploration des variables pour mieux comprendre la distribution.

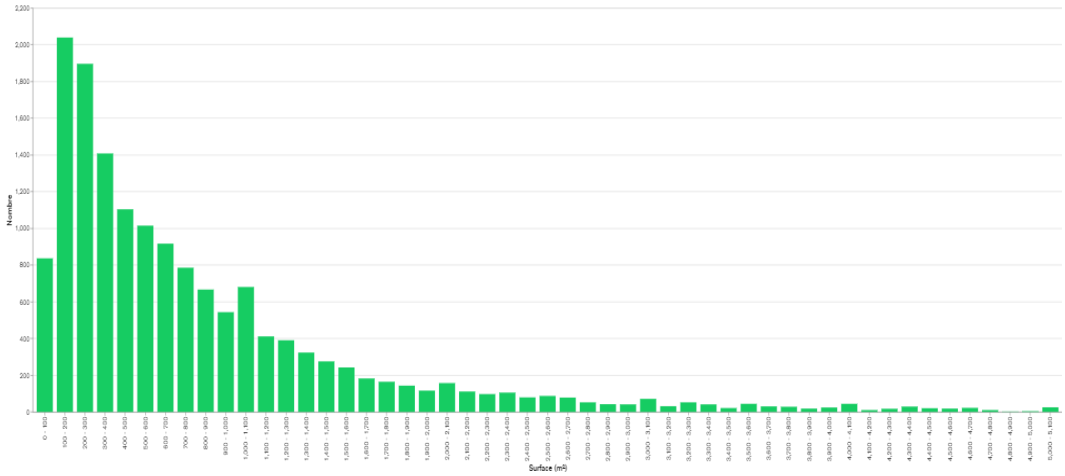
La surface du terrain

D'après l'histogramme, nous pouvons constater que la superficie des terrains des maisons à vendre est assez variable. La majorité des valeurs se concentrent sur la gauche, tandis qu'une longue queue s'étend sur la droite.

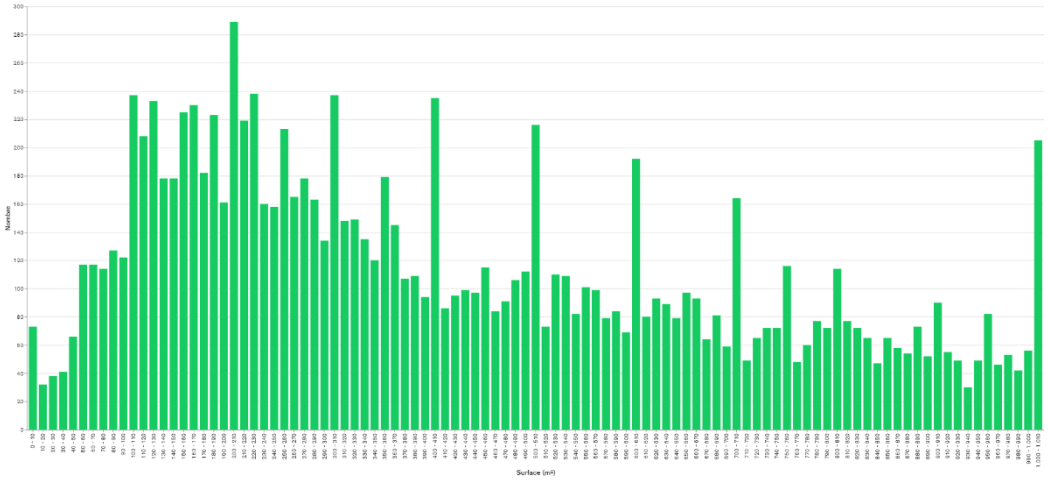
Nous remarquons qu'une petite annonce donne une propriété ayant une superficie de terrain d'un demi-kilomètre carré. En se basant sur notre classification théorique des maisons à vendre, une telle surface de terrain ne doit pas se retrouver dans la classe maison. C'est le genre de valeur qui correspondrait davantage à un château ou un domaine. Nous avons examiné plus attentivement cette petite annonce (identifiant 63b78204b2fe554d78a63290) et nous avons découvert qu'il s'agissait en réalité d'une maison de vacances située sur un domaine de 46 hectares. Il existe donc une imprécision entrée par le propriétaire de la petite annonce. Cette constatation nous a amenés à envisager la nécessité de revoir notre classification théorique et d'ajouter une limite de terrain pour notre définition de la classe maison.



Nous voulons également plus de détails sur la partie gauche de ce graphique. Nous allons donc réduire le bin size à 100 et limiter la taille de la surface du terrain maximum à 5000 pour obtenir un graphique plus lisible.



La catégorie 0-100 nous surprend car il semble y avoir beaucoup d'occurrence pour cette classe de surface. Nous allons procéder à un autre zoom, en réduisant le bin size à 10 et limiter la taille de la surface du terrain maximum à 1000 pour garder un graphique lisible.



La catégorie 0-10 nous surprend particulièrement car il semble y avoir relativement beaucoup d'occurrences pour une surface de maison aussi inhabituelle.

Dans la catégorie 0-10, nous observons des imprécisions entrées par l'utilisateur, comme par exemple, une maison de 132 mètres carrés habitables sur un terrain de 7 mètres carrés, ce qui voudrait dire qu'il s'agirait d'une maison à 18 étages. Les petites annonces présentant ce genre d'incohérence devront également être exclues.

Enfin, nous avons également remarqué une petite annonce (identifiant 63ba2479de8717d1353b4868) dans laquelle la superficie du terrain a été corrigée, passant de 9 mètres carrés à 900 mètres carrés. Bien que nous ayons cette correction dans notre jeu de données, nous ne savons pas y accéder avec les outils à notre disposition (*4.4.1. Limites*).

Un boxplot va nous permettre d'explorer encore davantage la distribution. Par souci de lisibilité de celui-ci, nous limitons la taille du terrain comprise entre 20 mètres carrés et 1,5 hectares.

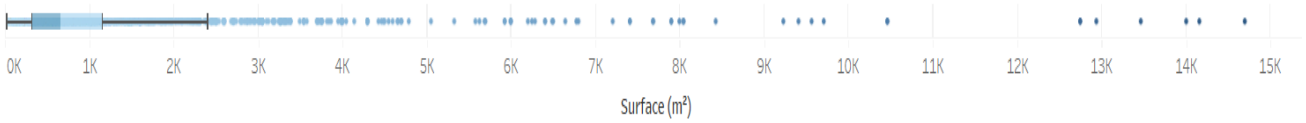


FIGURE 8 - Boxplot de la surface des terrains

Détails du boxplot

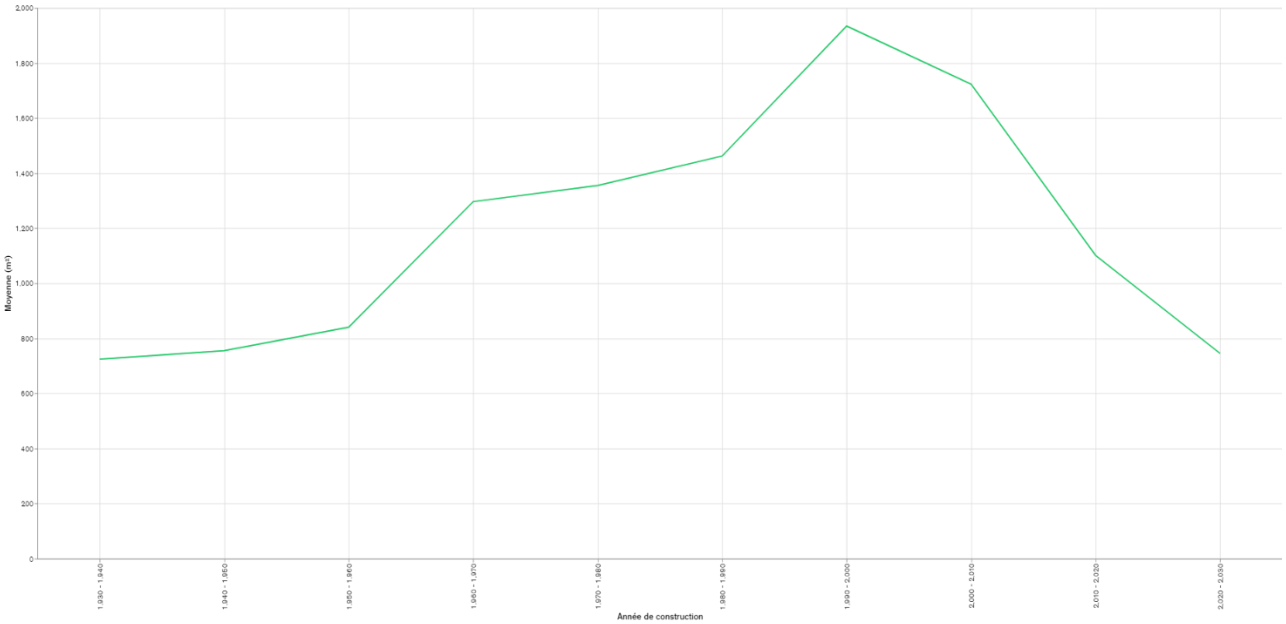
Barre supérieure	2.410
Charnière supérieure	1.159
Médiane	652,5
Charnière inférieure	320
Barre inférieure	20

Résumé statistique

Moyenne	990
Médiane	652,5
Maximum	14.693
Minimum	20

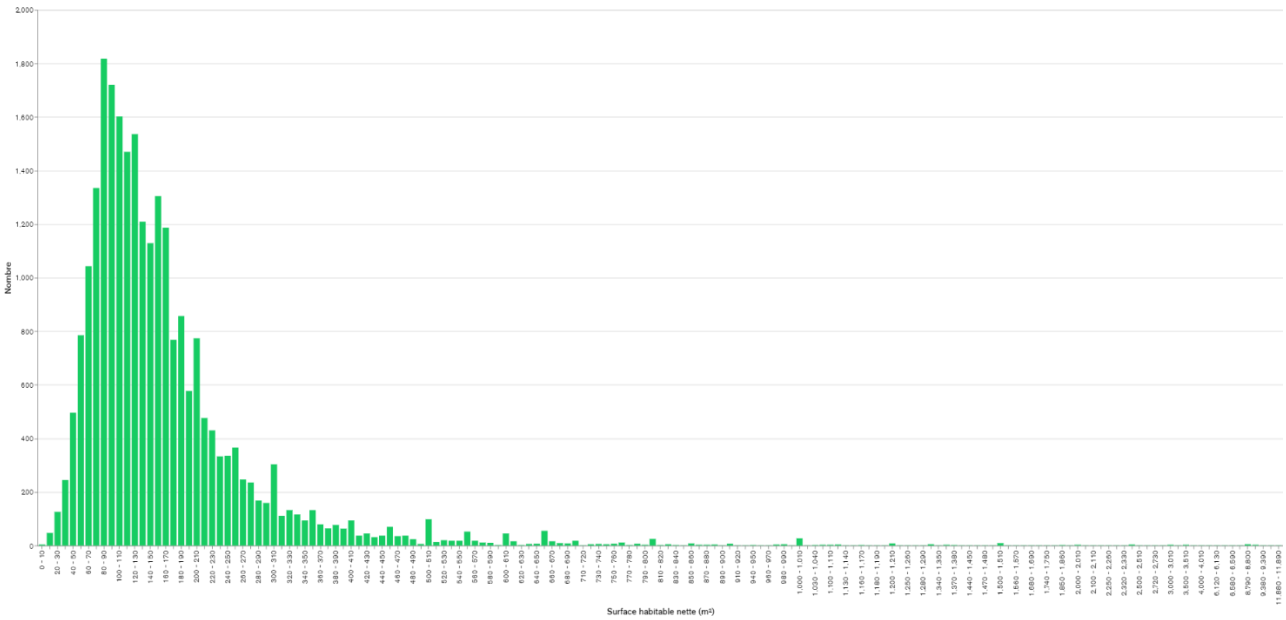
Au vu de cette tendance centrale exprimée par le boxplot, il sera probablement opportun d'affiner la classification lors de l'analyse. Par exemple, « maison à vendre » deviendrait « maison à vendre < 0.321k », « 0.321k < maison à vendre < 1.159k », etc.

Alors que la surface moyenne par maison augmentait jusque dans les années 1990, on observe, depuis ce moment, une diminution.



La surface nette habitable

Comme pour la distribution de la surface du terrain, l'histogramme révèle que les valeurs sont concentrées du côté gauche du graphique. On constate qu'il existe des valeurs aberrantes. Une longue queue s'étend sur la droite. Selon notre classification théorique des maisons à vendre, une maison ne devrait pas avoir une surface nette habitable de 500 mètres carrés ou plus. Cela ne correspondrait pas à notre définition.



Nous allons observer le boxplot en se basant sur la remarque précédente. Nous filtrons donc la valeur maximale à maximum 500. Pour rappel, une maison n'est pas un château ni un domaine.

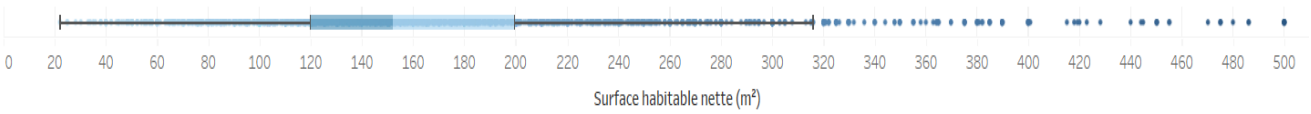


FIGURE 9 - Boxplot surface habitable nette

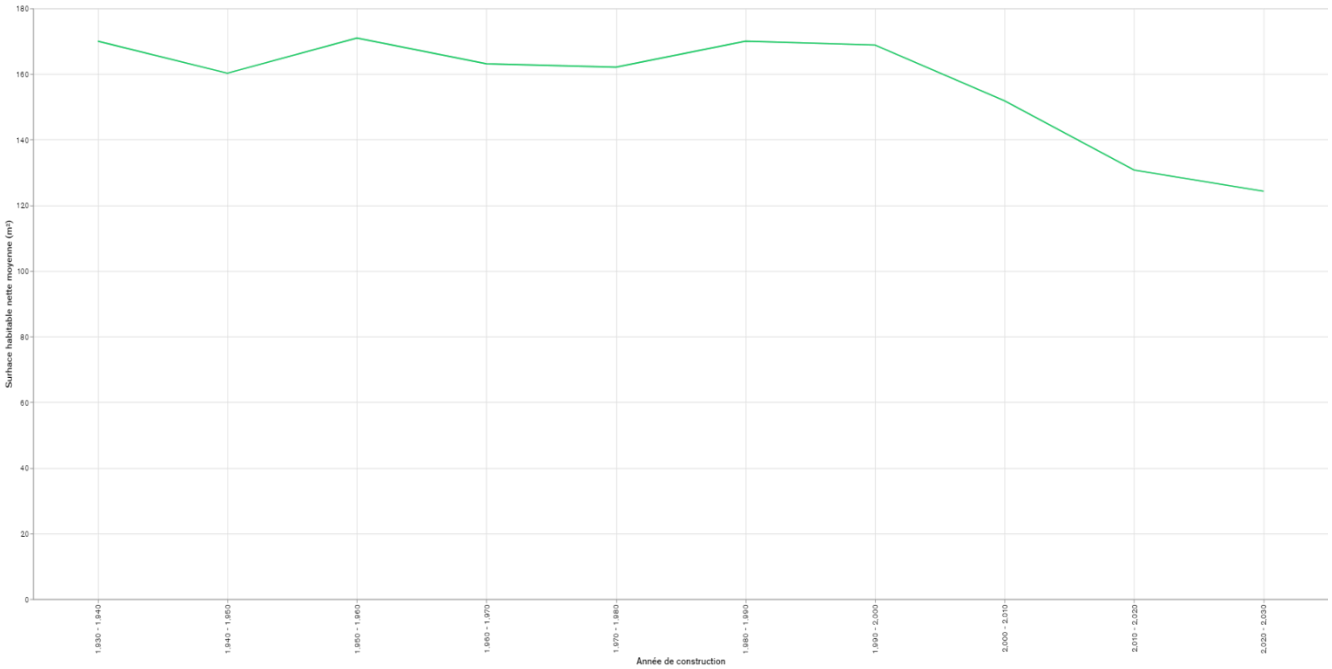
Détails du boxplot

Barre supérieure	316
Charnière supérieure	200
Médiane	152
Charnière inférieure	120
Barre inférieure	22

Résumé statistique

Moyenne	166,63
Médiane	152
Maximum	500
Minimum	22

Avec ce schéma en ligne discrète, on peut constater que la taille de la surface habitable nette est globalement en déclin. Puisque la pente est douce, on suppose qu'il ne s'agit pas d'un changement dans la méthode de calcul des surfaces habitables nettes, mais plutôt une diminution globale de la taille de la surface habitable nette des bâtiments.

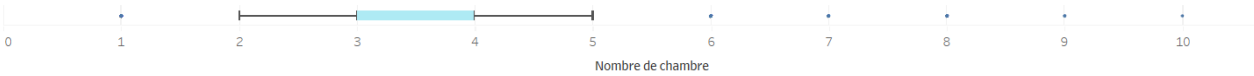


Exploration des variables discrètes

Le nombre de chambres

Le boxplot du nombre de chambres nous permet d'avoir une idée de la distribution de cette variable. Nous pouvons constater qu'il y a plusieurs outliers, c'est-à-dire des valeurs qui sont très éloignées de la moyenne. En général, cela peut être le résultat de données incorrectes ou de maisons avec des caractéristiques inhabituelles.

Par ailleurs, nous pouvons voir que la médiane est de 3 chambres, ce qui signifie que la plupart des maisons à vendre ont au moins 3 chambres.



Détails du boxplot

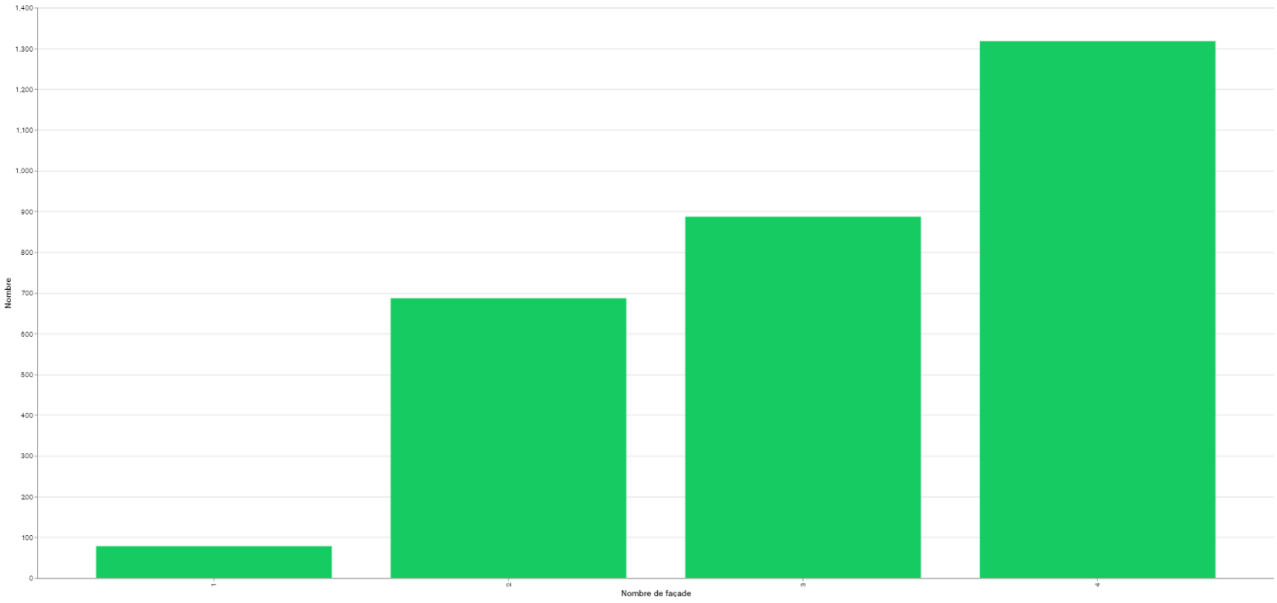
Barre supérieure	5
Charnière supérieure	4
Médiane	3
Charnière inférieure	3
Barre inférieure	2

Résumé statistique

Moyenne	3,33
Médiane	3
Maximum	10
Minimum	1

Le nombre de façades

Les maisons ayant quatre façades sont celles avec le plus d'occurrences (le mode). En d'autres mots, la majorité des maisons à vendre dans la province de Namur ont quatre façades.



On peut également ajouter que les nouvelles constructions qui apparaissent sur le marché ont majoritairement 3 façades. Dans les années 1950 à 2000, on trouvait surtout des maisons 4 façades. Est-ce une volonté urbanistique ou un résultat lié à l'économie et l'accessibilité au logement ?

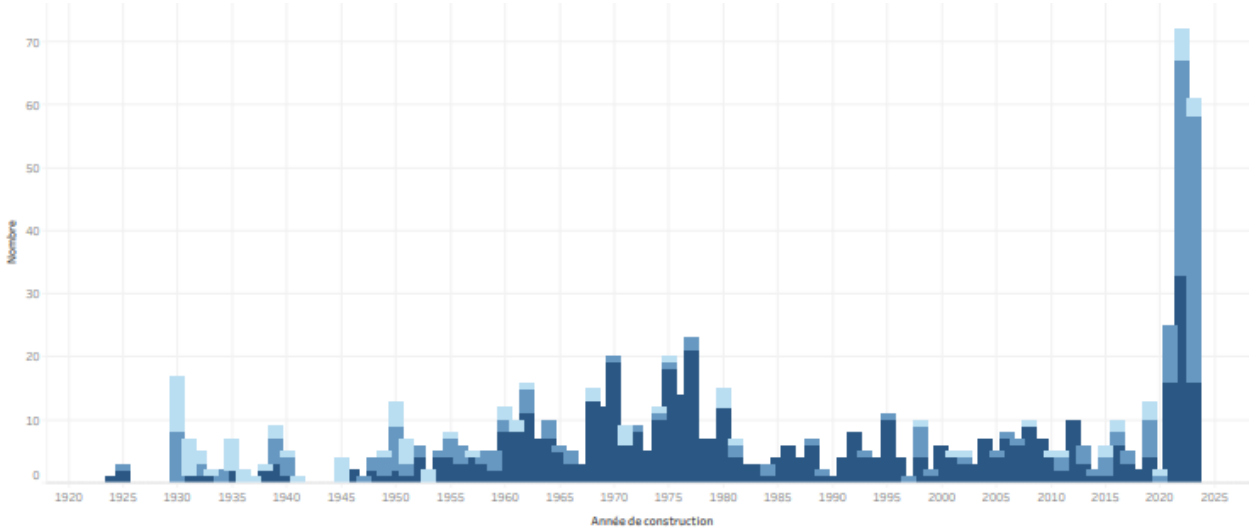
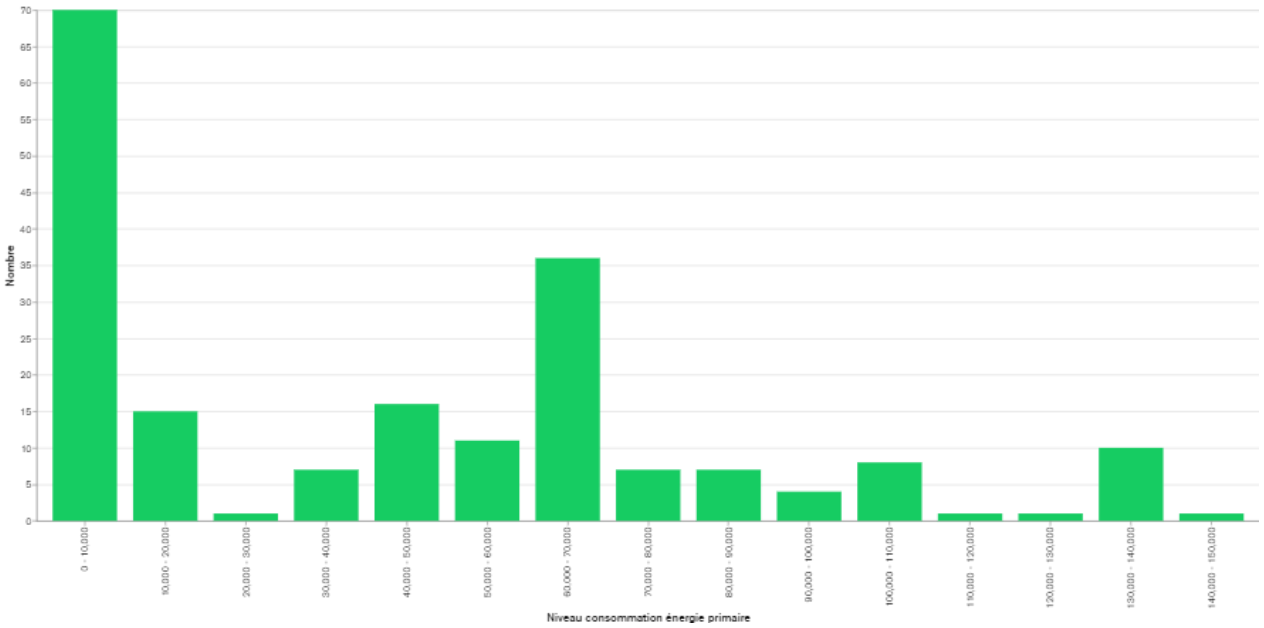


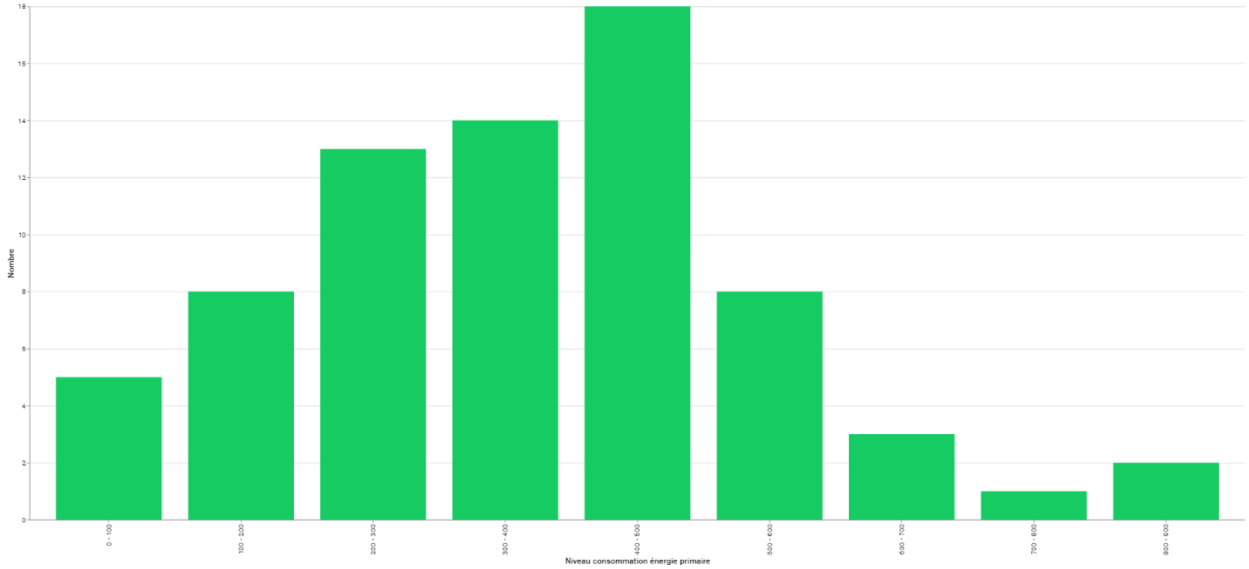
FIGURE 10 - Histogramme empilé du nombre de façades par année de construction

Le niveau de consommation d'énergie primaire

Nous constatons sur l'histogramme une certaine symétrie. Les valeurs du centre se répartissent autour de la classe 60 000 – 70 000. Cependant, le mode est la classe 0-10 000. Cette distribution est étrange, d'autant plus que c'est la première classe et que l'histogramme est symétrique en son centre.

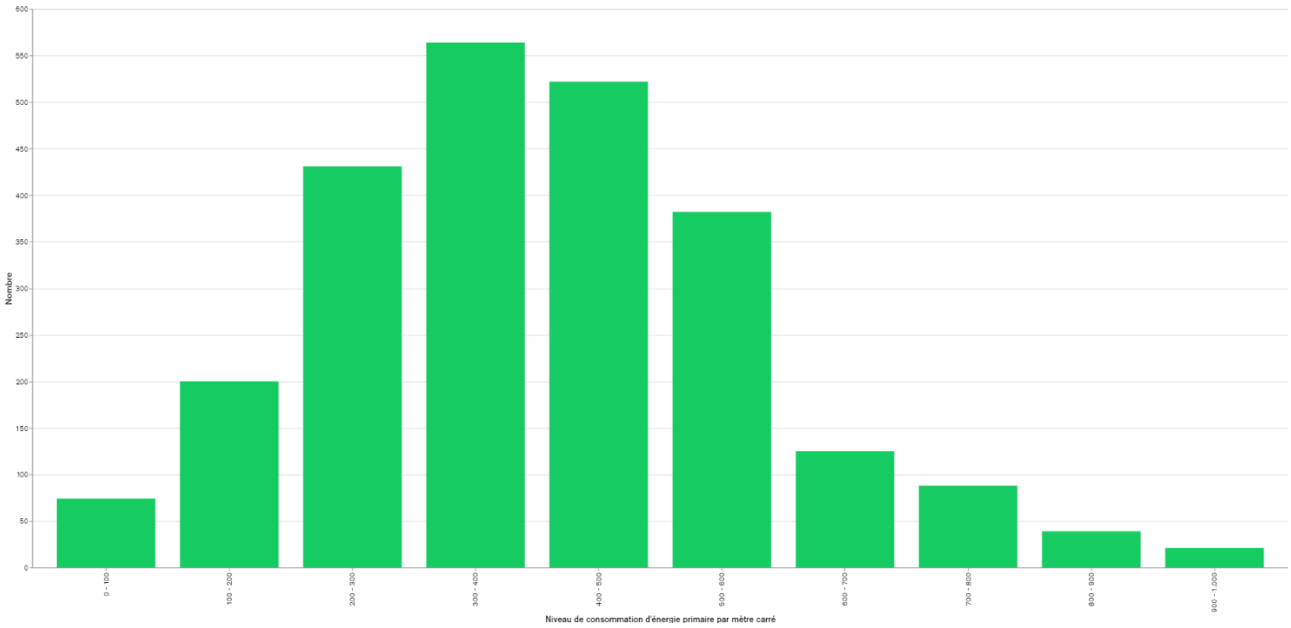


Nous décidons alors d'analyser plus en détail cette classe en faisant un zoom (et nous réduisons le bin size). Contrairement à ce qu'on aurait pu penser au départ, il existe une répartition au sein de cette classe.



Après analyse des enregistrements, il s'avère que la propriété « property.propertyCertificates.energyConsumptionLevel » soit comprise de deux manières par les utilisateurs (intentionnellement ou non). Certains créateurs de petites annonces vont renseigner le niveau de consommation d'énergie primaire annuelle, alors que d'autres indiquerons le niveau au mètre carré.

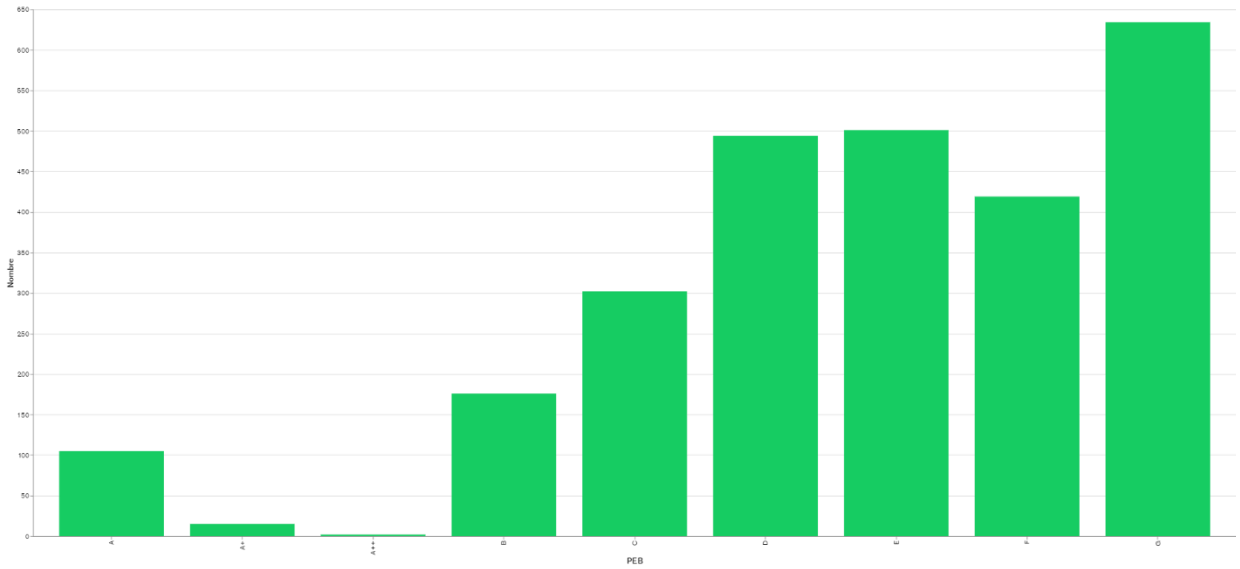
Nous utiliserons donc plutôt la variable « transaction.certificates.primaryEnergyConsumptionLevelPerSqm » dans laquelle on retrouve le niveau de consommation d'énergie primaire par mètre carré.



EpcScore (PEB – Performance Energetique des bâtiments)

Le score PEB est une mesure de la consommation d'énergie d'un bâtiment et de son impact sur l'environnement en termes d'émissions de CO2. Il est déterminé en fonction de plusieurs critères tels que l'isolation thermique, le système de chauffage, la ventilation, etc.

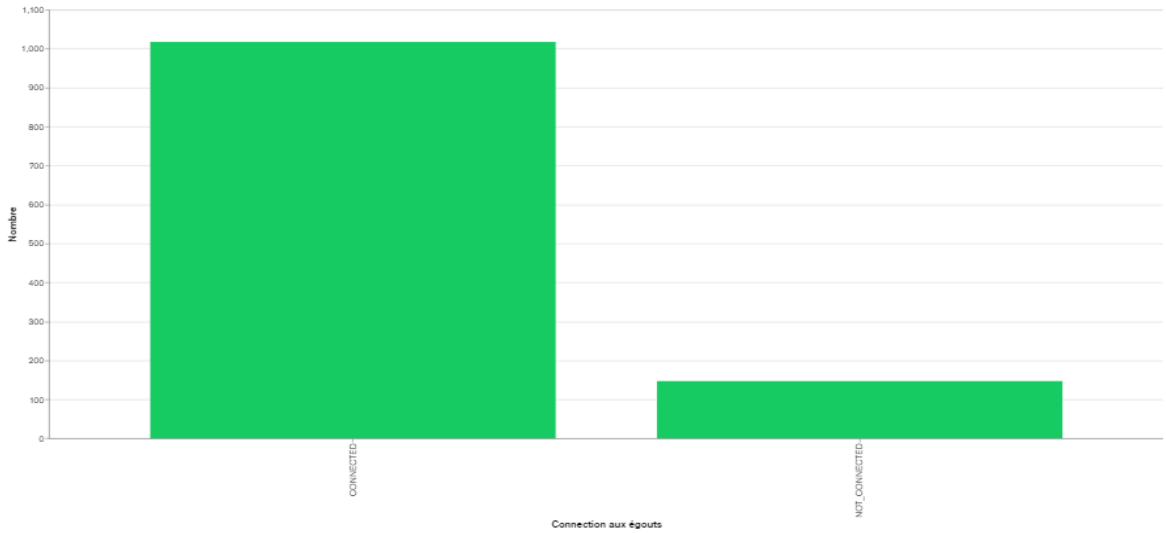
Dans cet histogramme, nous constatons que la plupart des maisons en vente sur le marché immobilier ont un score bas (G, F, E). Les maisons sont donc peu économes en énergie et ont une empreinte carbone plus grande. Nous devrons croiser cette information avec d'autres, comme le prix par exemple, pour voir s'il existe une corrélation.



Égouts

Le nombre de maisons raccordées aux égouts ou le nombre de maisons qui ne le sont pas est peut-être un facteur important à prendre en compte lors de l'achat d'une maison. Les maisons raccordées aux égouts sont généralement plus attrayantes car elles sont plus modernes et plus respectueuses de l'environnement que les maisons non raccordées qui utilisent souvent des fosses septiques.

Dans ce jeu de données, il serait intéressant d'analyser la répartition de ces deux catégories par rapport à d'autres variables telles que le prix, la surface de la maison ou la localisation géographique. Par exemple, il serait intéressant de savoir si les maisons raccordées aux égouts ont tendance à avoir un prix plus élevé ou si elles sont plus courantes dans certaines régions.



5.4.2.2. Tendance centrale

L'analyse des tendances centrales est une étape importante dans l'exploration des données. Nous allons observer des informations telles que les moyennes et les médianes afin de décrire la valeur la plus typique d'une variable et de comprendre comment les données sont distribuées autour de cette valeur. Pour rappel, nous avons vu précédemment les boxplots des prix, du nombre de chambre, la surface du terrain et la surface nette habitable.

La surface habitable nette par sous-type

Dans le graphique des surfaces habitables nettes par sous-type, nous constatons que la distribution des maisons de sous-type « maison » est très dispersée par rapport à celle des autres sous-types. Cette observation soulève la question de savoir pourquoi cela se produit.

Notre hypothèse est que de nombreux utilisateurs créent des annonces avec le sous-type « maison » par défaut sans tenir compte de la classification réelle de la propriété. Pour confirmer cette hypothèse, nous devons examiner les valeurs aberrantes dans notre jeu de données et les comparer à notre classification théorique.

À titre d'exemple, nous prenons la petite annonce dont l'identifiant est 63cb414ec33-fa45646782f50. Cette maison est annoncée avec une surface nette habitable de 580 m² et nous observons qu'elle est décrite comme une villa mais que le sous-type sélectionné est « maison » (house).

Pour un autre exemple, nous prenons la petite annonce dont l'identifiant est 63fab5294dc5-cc0e01aad7ba. En analysant les données, nous constatons qu'il s'agit d'un immeuble de rapport qui est destiné à faire plusieurs appartements.

Pour un dernier exemple, nous prenons la petite annonce dont l'identifiant est 63c612d15785-620102767b18. En lisant la description et en passant en revue les différentes propriétés, nous observons qu'il s'agit plutôt d'une « propriété d'exception » (exceptional property).

Ces annonces ont toutes été créées avec un sous-type inadéquat et ont en commun une surface nette habitable avec une valeur aberrante. Nous pouvons donc affiner notre classification en délimitant notre classification théorique des maisons avec une surface nette habitable maximale de 500 m², par exemple.

En conclusion, en examinant les valeurs aberrantes dans notre jeu de données, nous avons identifié une tendance à utiliser le sous-type « maison » par défaut, ce qui peut fausser la distribution des surfaces nettes habitables. En utilisant une classification plus précise, nous pouvons obtenir des résultats plus fiables et plus significatifs dans notre analyse des données immobilières.

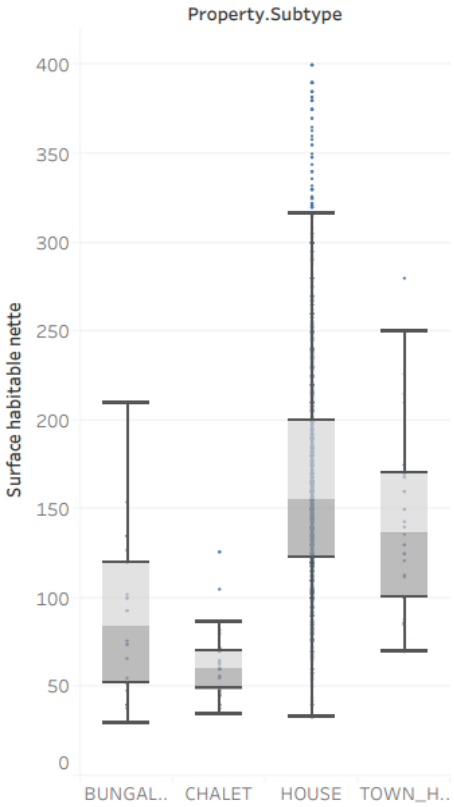


FIGURE 11 - Boxplots surface habitable nette par sous-type

Détails des boxplots

	BUNGALOW	CHALET	HOUSE	TOWN_HOUSE
Barre supérieure	210	87	316	250
Charnière supérieure	120	70	200	170
Médiane	84,5	60	155	136
Charnière inférieure	52	48,5	122	100
Barre inférieure	30	35	33	70

Résumés statistiques

	BUNGALOW	CHALET	HOUSE	TOWN_HOUSE
Moyenne	89,18	63,2	168,02	141,71
Médiane	84,5	60	155	136
Maximum	210	126	400	280
Minimum	30	35	33	70

Le nombre de chambres par sous-type

Les boxplots du nombre de chambres en fonction du sous-type de maison nous permettent de visualiser la distribution du nombre de chambres pour chaque sous-type. Nous pouvons ainsi observer des différences dans la médiane, la dispersion, ainsi que la présence d'éventuelles valeurs aberrantes.

Cependant, il est important de noter que ces observations ne sont que des tendances générales et qu'il peut y avoir des cas individuels qui ne suivent pas cette tendance. Les boxplots peuvent néanmoins nous aider à mieux comprendre la distribution des données pour chaque sous-type de maison.

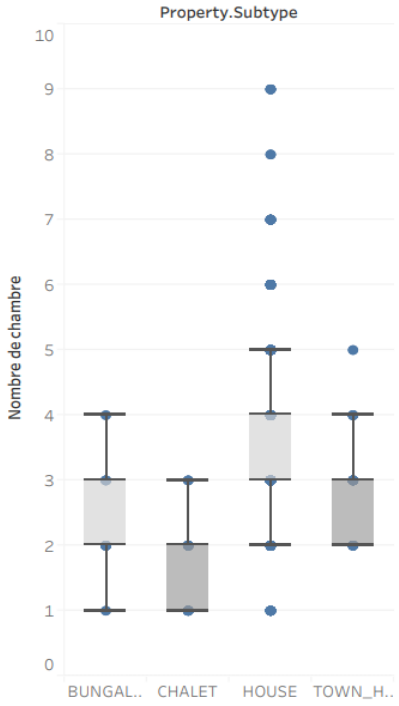


FIGURE 12 - Boxplots nombre de chambres par sous-type

Détails des boxplots

	BUNGALOW	CHALET	HOUSE	TOWN_HOUSE
Barre supérieure	4	3	5	4
Charnière supérieure	3	2	4	3
Médiane	2	2	3	3
Charnière inférieure	2	1	3	2
Barre inférieure	1	1	2	2

Résumés statistiques

	BUNGALOW	CHALET	HOUSE	TOWN_HOUSE
Moyenne	2,36	1,69	3,38	2,94
Médiane	2	2	3	3
Maximum	4	3	9	5
Minimum	1	1	1	2

La surface du terrain par sous-type

Les boxplots et les résumés statistiques nous communiquent des informations sur la distribution de la surface du terrain en fonction du sous-type de maison.

On remarque que les maisons de sous-type « chalet » ont la plus grande amplitude, allant de 60 m² à 6.200 m². Le résumé statistique montre que les maisons de type chalet ont la moyenne la plus élevée de surface de terrain, suivies des maisons et des bungalows, tandis que les maisons de ville ont la moyenne la plus basse. Cependant, la médiane montre un ordre différent, avec les bungalows ayant la surface médiane de terrain la plus élevée, suivis des maisons et des chalets. Les valeurs maximales sont également très élevées pour les maisons et les chalets, dépassant largement les valeurs des bungalows et des maisons de ville. Enfin, les valeurs minimales sont aussi très différentes entre les différents sous-types de maison, avec les maisons de ville ayant la surface de terrain minimale la plus basse.

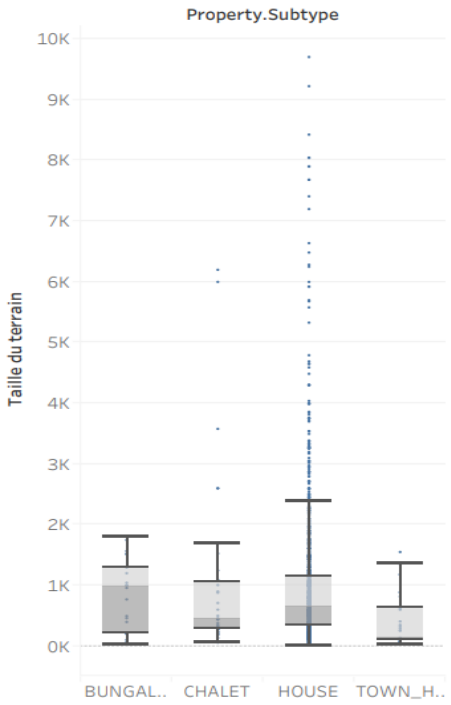


Figure 13 - Boxplots taille terrain par sous-type

Détails des boxplots

	BUNGALOW	CHALET	HOUSE	TOWN_HOUSE
Barre supérieure	1.801	1.684	2.390	1.364
Charnière supérieure	1.294	1.049	1.148,5	612,5
Médiane	979	448	650,5	136
Charnière inférieure	210	277,5	320	93
Barre inférieure	42	60	15	40

Résumés statistiques

	BUNGALOW	CHALET	HOUSE	TOWN_HOUSE
Moyenne	845,14	1.055,8	922,89	402,46
Médiane	979	448	650,5	136
Maximum	1.801	6.200	9.706	1.550
Minimum	42	60	15	40

5.4.2.3. Variabilité

Nous allons observer l'écart interquartile (RIQ), la variance et l'écart-type (standard deviation). Ce sont des mesures importantes de la variabilité des données, chacune offrant un aperçu différent de la dispersion des données autour de la moyenne.

Le RIQ est la différence entre le troisième quartile (Q3) et le premier quartile (Q1) de la distribution des données. Il représente la distance entre les valeurs centrales des données et ignore les valeurs extrêmes. Le RIQ est souvent utilisé pour identifier les valeurs aberrantes potentielles dans un ensemble de données. Il donne également une mesure plus précise de la dispersion.

La variance mesure la distance moyenne de chaque point de données par rapport à la moyenne de la distribution. C'est une mesure de la dispersion qui tient compte de chaque point de données de l'ensemble. Une variance élevée indique une grande dispersion des données (grande variabilité), tandis qu'une variance faible indique une faible dispersion (données homogènes).

L'écart-type est la racine carrée de la variance et mesure la distance moyenne des points de données par rapport à la moyenne de la distribution. Il s'agit d'une mesure de la dispersion qui est exprimée dans les mêmes unités que les données originales. Un écart-type élevé indique une grande dispersion des données, tandis qu'un écart-type faible indique une faible dispersion.

Pour réaliser ces calculs, nous utilisons la librairie « Numpy » de Python. Les scripts sont disponibles pour le corps enseignant par simple demande à l'auteur de ce mémoire.

Le prix

RIQ : 150.000

Les maisons observées dans la boîte (boxplot - entre Q1 et Q3, soit 50% des observations) ont des prix qui peuvent avoir un écart maximal de 150.000€.

Variance : 18554388967

La variance de 18554388967 pour la variable du prix de vente des maisons signifie que les valeurs de prix sont très dispersées autour de leur moyenne. Plus précisément, cela indique que la différence entre chaque valeur de prix et la moyenne est assez grande. En d'autres termes, il y a une grande variation dans les prix des maisons.

Prenons maintenant l'écart-type pour avoir une mesure dans la même unité que la variable elle-même (l'euro).

Écart-type : 136.214

L'écart-type de 136.214 pour la variable du prix de vente des maisons signifie que les valeurs de prix sont dispersées autour de leur moyenne d'environ 136.214 euros. Cela donne une idée de la variabilité des prix des maisons.

La surface du jardin

RIQ : 800

Les maisons observées dans la boîte (boxplot - entre Q1 et Q3, soit 50% des observations) ont des tailles de jardin qui peuvent avoir un écart maximal de 800 m².

Variance : 1804342

La variance de 1804342 pour la variable de la taille du jardin d'une maison signifie que les valeurs sont très dispersées autour de leur moyenne. Plus précisément, cela indique que la différence entre chaque valeur et la moyenne est assez grande. En d'autres termes, il y a une grande variation dans les tailles des jardins des maisons.

Écart-type : 1343

L'écart-type de 1343 pour la variable de la taille du jardin d'une maison signifie que les valeurs individuelles de la taille du jardin ont tendance à se situer à environ 1343 mètres carrés de la

moyenne de la taille du jardin. L'écart-type est une mesure de dispersion qui donne une idée de la variation ou de l'étendue des données autour de leur moyenne.

Un écart-type élevé indique que les données sont très dispersées autour de la moyenne, tandis qu'un écart-type faible indique que les données sont relativement regroupées autour de la moyenne. Dans ce cas, un écart-type de 1343 semble indiquer que la distribution des tailles de jardin est assez variable, mais moins que la variance ne le suggérait.

Il est important de garder à l'esprit que l'écart-type doit toujours être interprété en relation avec la moyenne et avec les valeurs individuelles de la variable étudiée.

La taille du terrain

RIQ : 859

Les maisons observées dans la boîte (boxplot - entre Q1 et Q3, soit 50% des observations) ont des tailles de terrain qui peuvent avoir un écart maximal de 859 m².

Variance : 99951632

La variance de 99951632 pour la variable de la taille du terrain d'une maison signifie que les valeurs sont très dispersées autour de leur moyenne. Plus précisément, cela indique que la différence entre chaque valeur et la moyenne est assez grande. En d'autres termes, il y a une grande variation dans les tailles des terrains des maisons.

Écart-type : 9997

Cela signifie que les valeurs de la taille du terrain sont assez éloignées de leur moyenne, en moyenne d'environ 9997 mètres carrés. Plus précisément, cela signifie que la plupart des valeurs de la taille du terrain se situent à environ 9997 mètres carrés de la moyenne. L'écart-type est une mesure de la dispersion des données par rapport à leur moyenne, et plus il est élevé, plus les valeurs sont dispersées.

La taille de la surface nette habitable

RIQ :

Les maisons observées dans la boîte (boxplot - entre Q1 et Q3, soit 50% des observations) ont des tailles de surface nette habitable qui peuvent avoir un écart maximal de 80 m².

Variance : 10981

La variance de 10981 pour la variable de la taille de la surface nette habitable d'une maison indique que les valeurs de la variable sont dispersées autour de la moyenne sur une étendue

relativement faible. Cela suggère que la plupart des maisons ont une taille de surface nette habitable similaire les unes aux autres.

Écart-type : 104

L'écart-type de 104 sur la variable de la taille de la surface nette habitable d'une maison indique que les valeurs sont relativement proches les unes des autres, avec peu de dispersion autour de la moyenne. Cela peut être interprété comme une indication que la taille de la surface nette habitable des maisons dans l'ensemble de données est relativement homogène, avec peu de variations importantes par rapport à la moyenne.

5.4.2.4. Régressions et corrélations (tests d'hypothèse)

Nous testerons des hypothèses sur les données afin de déterminer si deux variables sont corrélées. L'outil de visualisation que nous utiliserons est le scatter plot. C'est un type de graphique qui permet de représenter les relations entre deux variables quantitatives en les plaçant sur les axes X et Y. Nous l'utiliserons pour visualiser la corrélation ou l'association entre deux variables continues.

Lorsqu'on lit un scatter plot, on peut observer la dispersion des points sur le graphique. Si les points sont concentrés autour d'une ligne droite, cela indique une forte corrélation positive ou négative entre les deux variables. Si les points sont dispersés de manière uniforme sur le graphique, cela indique une faible ou aucune corrélation entre les deux variables.

Il est également possible de détecter des valeurs aberrantes ou des points qui sont très éloignés des autres points sur le graphique, ce qui peut influencer les résultats de l'analyse. Nous utilisons plusieurs librairies de Python. Les scripts sont disponibles pour le corps enseignant par simple demande à l'auteur de ce mémoire.

En ce qui concerne les variables qualitatives ou non continues, nous utiliserons plutôt le prix moyen et un histogramme.

Présentation des tests d'hypothèse

- Prix demandé

Cette propriété peut varier considérablement en fonction de divers facteurs tels que la taille du terrain, la taille de la surface nette habitable, la taille du jardin, le nombre de places de parking intérieures, le nombre de chambres, l'année de construction et l'état du bien immobilier.

- Nombre de chambres

Cette propriété peut varier selon la taille de la surface nette habitable et l'année de construction.

- Surface du jardin

Cette propriété peut varier en fonction de la taille de la propriété et l'année de construction.

- L'emplacement

Nous l'analyserons via une carte choroplèthe.

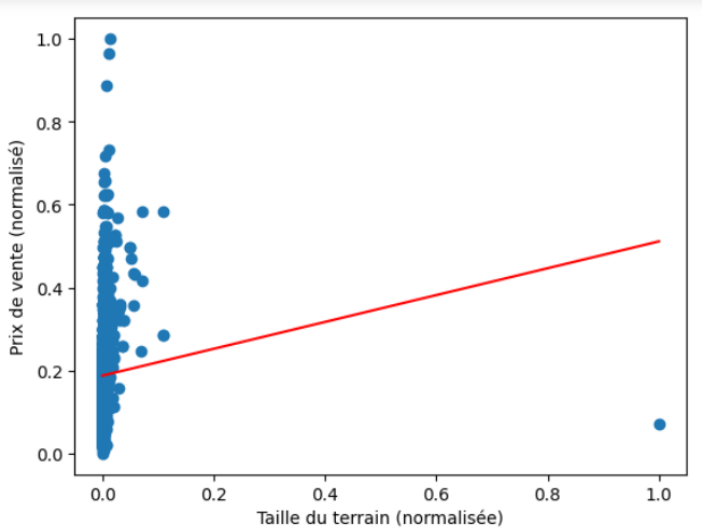
- Raccordement aux égouts

Cette propriété peut varier selon la situation géographique du bien immobilier. Nous l'explorerons avec une carte choroplèthe.

Prix demandé vs surface du terrain

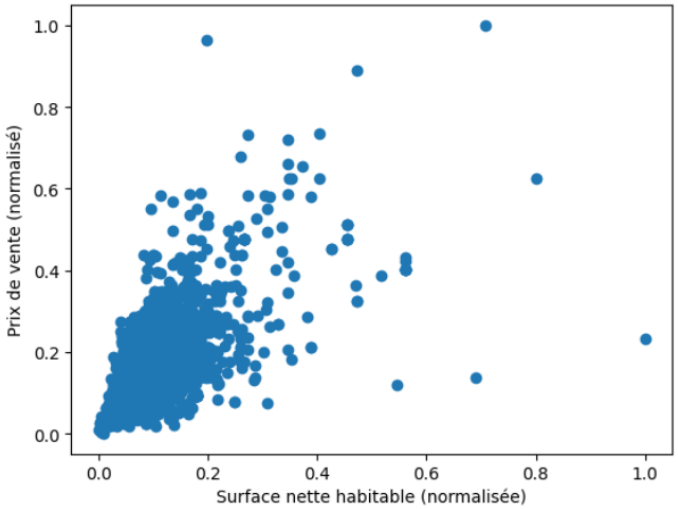
Puisque nous comparons un prix en euro et une surface de terrain en mètre(s) carré(s) (unités différentes), il est judicieux de normaliser les données sur les deux axes pour mieux visualiser la relation entre les variables et éviter d'avoir un axe qui domine l'autre.

Le scatter plot montre qu'une faible augmentation du terrain entraîne une grande augmentation du prix.



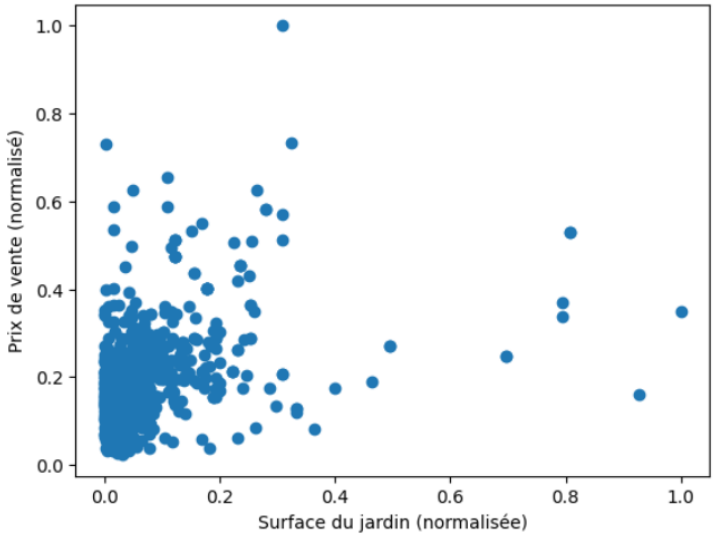
Prix demandé vs surface nette habitable

Le nuage de points semble suivre le tracé d'une droite. Nous distinguons donc une relation linéaire dans laquelle l'augmentation d'une variable entraîne à peu près au même rythme une augmentation de l'autre variable.



Prix demandé vs surface du jardin

On ne distingue pas de linéarité évidente entre ces observations. Le lien direct entre augmentation du prix et taille du jardin n'est pas évident.

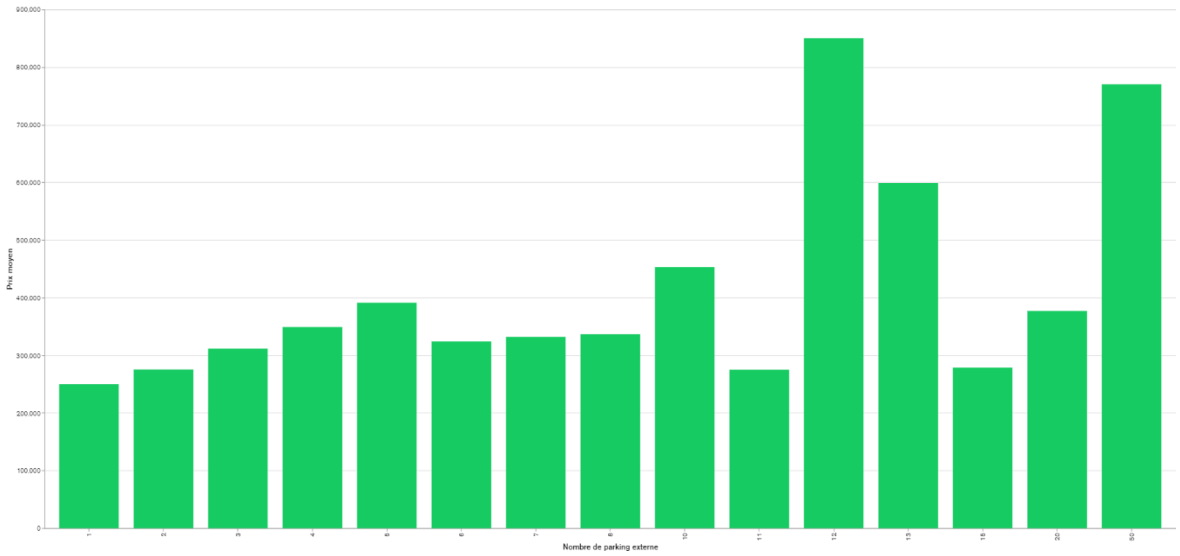


Prix demandé vs nombre de places de parking extérieur

Nous constatons d'abord qu'il existe des valeurs aberrantes. Selon notre classification théorique, une maison ne devrait pas avoir 20 ou 50 places de parking extérieures. Il s'agit, une nouvelle fois, de petites annonces qui ont un mauvais sous-type. Nous pourrions donc décider d'affiner notre classification théorique en disant qu'une maison a entre zéro et dix places de parking extérieures.

Les valeurs aberrantes mises à part, nous observons que le prix moyen tend à augmenter jusqu'à cinq garages extérieurs, puis se stabilise. Nous émettons l'hypothèse que cette

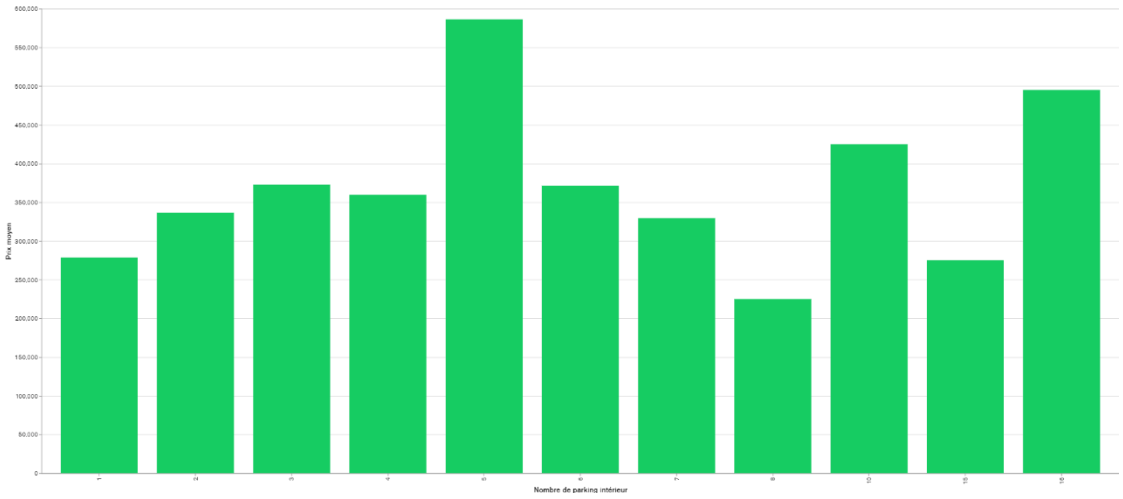
augmentation est probablement due à l'augmentation de la taille du terrain sous-jacente plutôt qu'à l'augmentation du nombre de places de parking.



Prix demandé vs nombre de places de parking intérieur

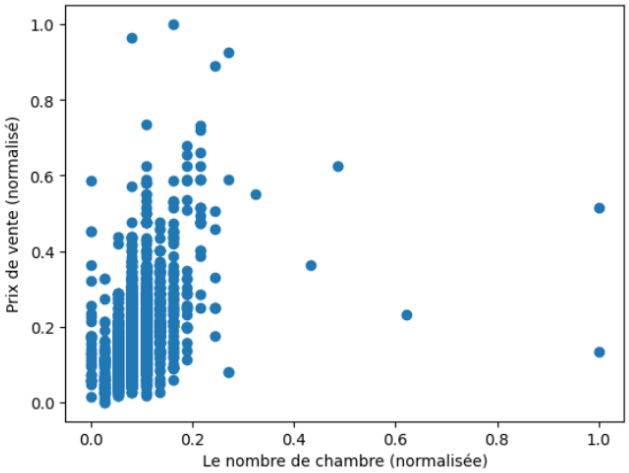
Nous constatons d'abord qu'il existe des valeurs aberrantes. Selon notre classification théorique, une maison ne devrait pas avoir plus de 10 places de parkings intérieures. Il s'agit, une nouvelle fois de petites annonces qui ont un mauvais sous-type. Nous pourrions donc décider d'affiner notre classification théorique en disant qu'une maison a entre zéro et un nombre déterminé de places de parking intérieures.

Les valeurs aberrantes mises à part, nous observons que le prix moyen est stable face à l'augmentation du nombre de places de parking intérieures. Autrement dit, cette variable ne semble pas influencer le prix directement de manière évidente. Nous notons cependant que le mode est étrangement élevé et mériterait une analyse plus approfondie des données qui le composent.

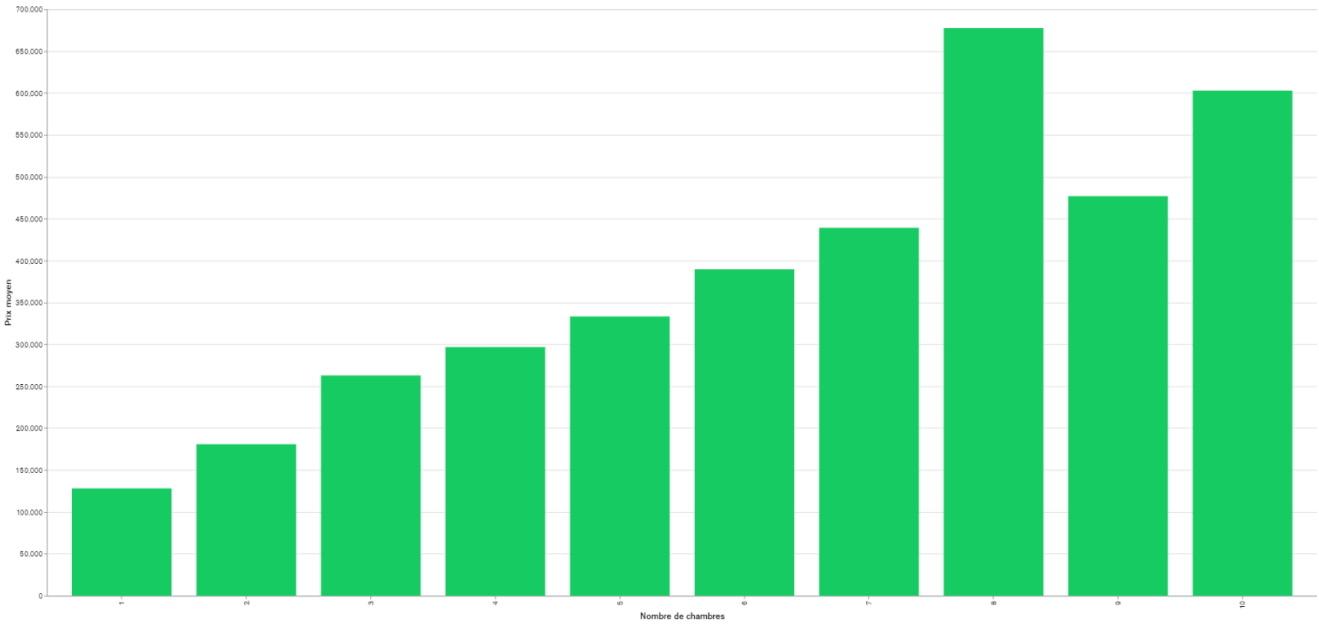


Prix demandé vs nombre de chambres

Les observations sont relativement dispersées, cependant, nous pouvons quand-même observer qu'une augmentation du nombre de chambres entraînera une augmentation du prix.



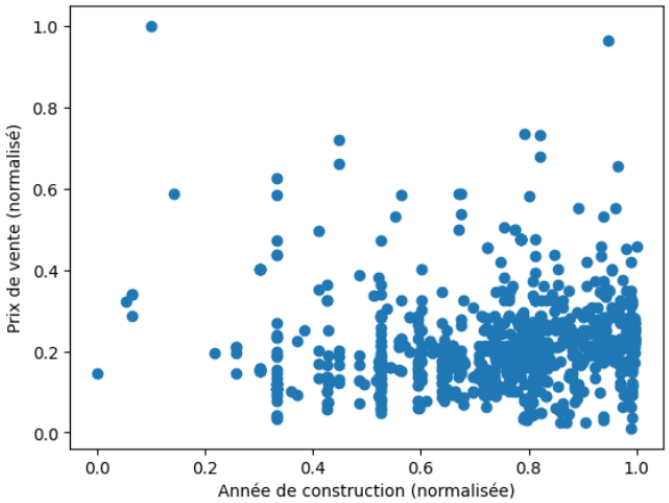
L'histogramme montre que le prix moyen d'une maison est plus élevé si le nombre de chambres est plus élevé.



Prix demandé vs l'année de construction

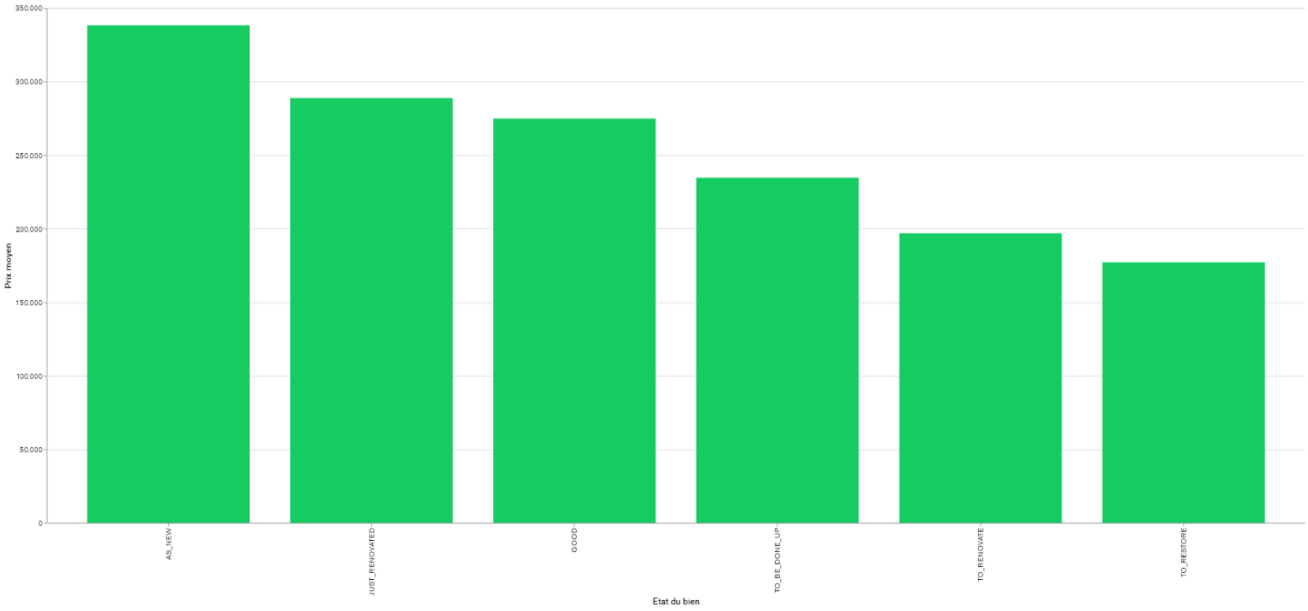
L'année de construction n'est pas une variable déterminante sur le prix. Les points sont dispersés partout sur le scatter plot, cela suggère qu'il n'y a pas de relation linéaire claire entre l'année de construction et le prix de vente. En d'autres termes, l'année de construction ne semble

pas être un facteur déterminant dans la détermination du prix de vente des biens immobiliers étudiés.



Prix demandé vs l'état du bien

Le prix moyen est plus élevé quand le bien est en bon état. On peut donc déduire que l'état du bien est une variable déterminante du prix. Cela est peu étonnant puisque nous pouvons expliquer cette observation par une hypothèse simple : le fait que les biens en bon état nécessitent moins de travaux de rénovation ou de réparation, ce qui tend à réduire les coûts pour l'acheteur potentiel et donc à justifier un prix plus élevé. De plus, les biens en bon état peuvent donner une impression de qualité supérieure, ce qui peut augmenter leur attractivité et donc leur valeur sur le marché. Il est important de prendre en compte l'état du bien lors de l'évaluation de son prix, que ce soit pour un acheteur potentiel ou pour un vendeur souhaitant fixer un prix de vente optimal.



Prix demandé vs emplacement du bien

L'emplacement joue un rôle sur la variation du prix. On constate que le prix moyen est plus bas dans la moitié basse de la province. En revanche, le prix moyen est plus élevé dans la moitié haute de la province et plus particulièrement au cœur de la capitale wallonne.

Nous pouvons voire dans la *figure 14* ci-dessous que les localités ayant le prix moyen le plus élevé sont Falaën (5522) et Florée (5334). Ces résultats sont à relativiser car nous avons, pour le moment, peu d'occurrences pour ces localités (*figure 15*).

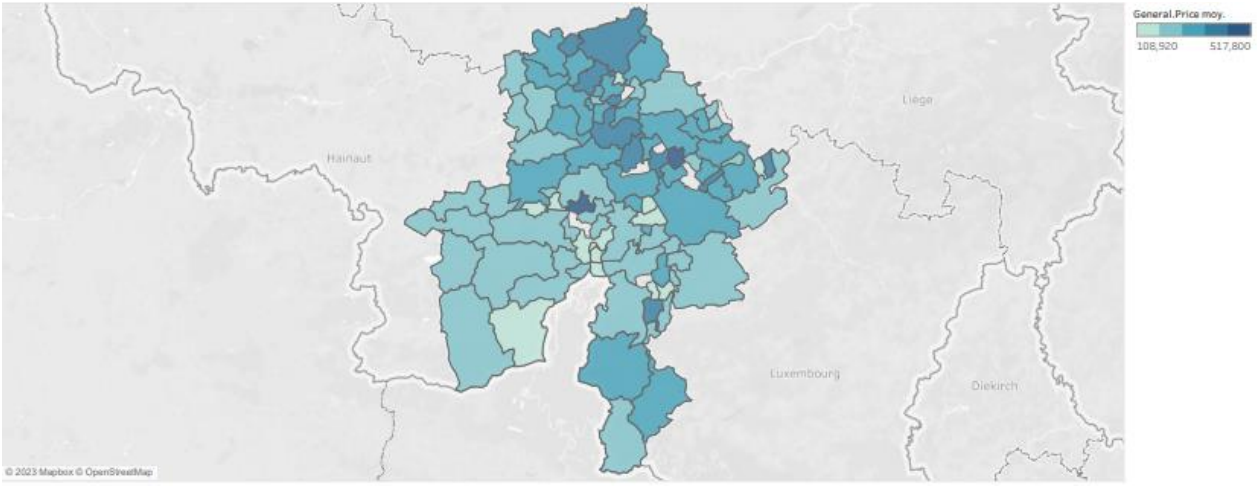


FIGURE 14 - Carte choroplèthe prix moyen par localité

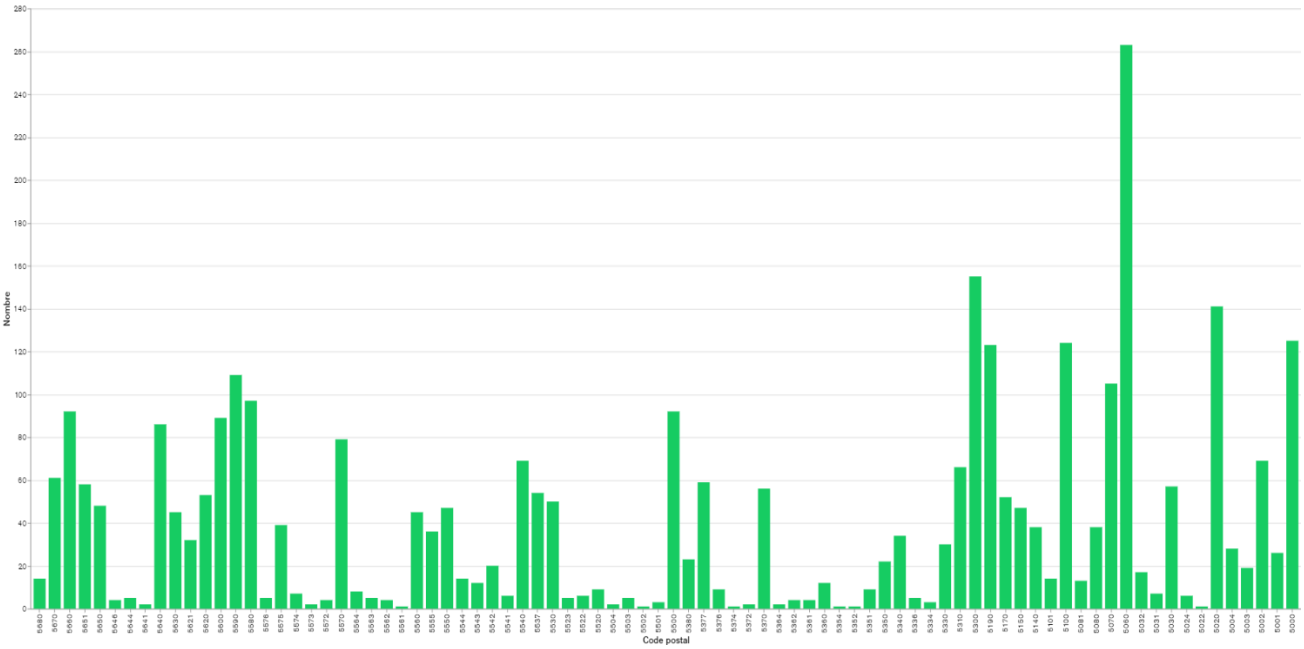
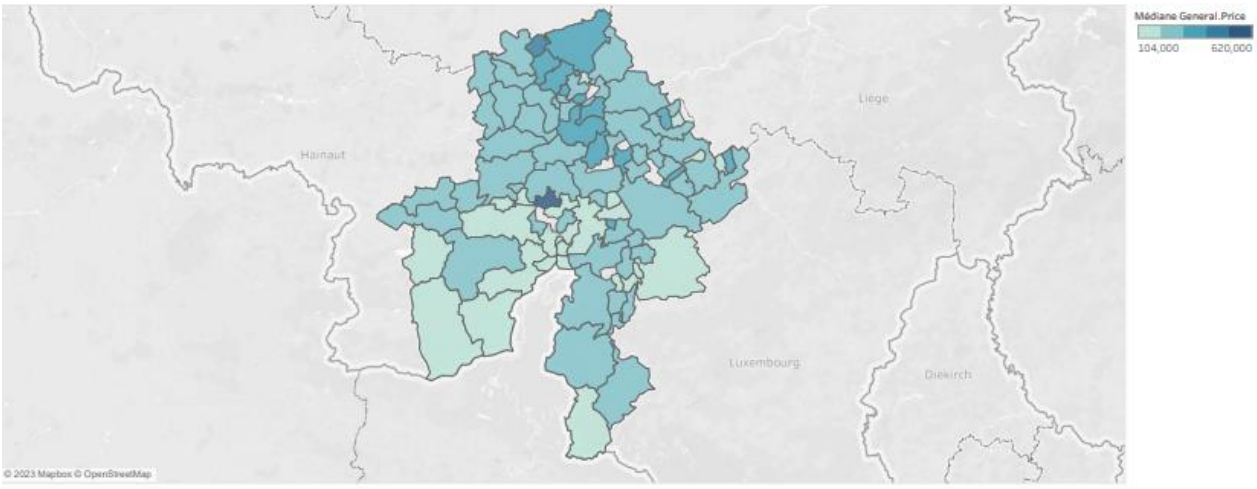


FIGURE 15 - Histogramme des occurrences par localité

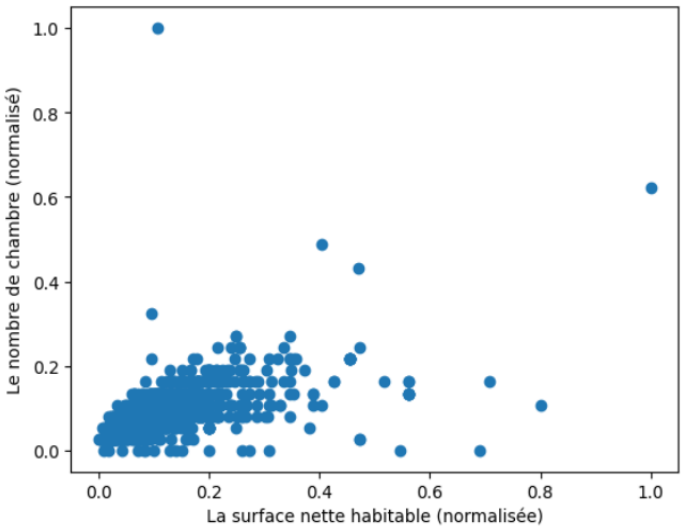
La dispersion du prix étant importante et asymétrique, il peut être intéressant de plutôt choisir la médiane comme référence. Dans ce cas, la moyenne peut être influencée par les valeurs aberrantes ou extrêmes. Par conséquent, l'utilisation de la médiane comme mesure de référence est plus appropriée, car elle n'est pas affectée par les valeurs extrêmes. La médiane est la valeur qui sépare la distribution en deux parties égales, et est donc moins sensible aux valeurs aberrantes. En somme, la médiane est une mesure de tendance centrale robuste qui est plus appropriée pour les distributions asymétriques et avec des valeurs extrêmes.

Donc, si on prend plutôt la médiane comme référence, on constate que les couleurs s'éclaircissent.



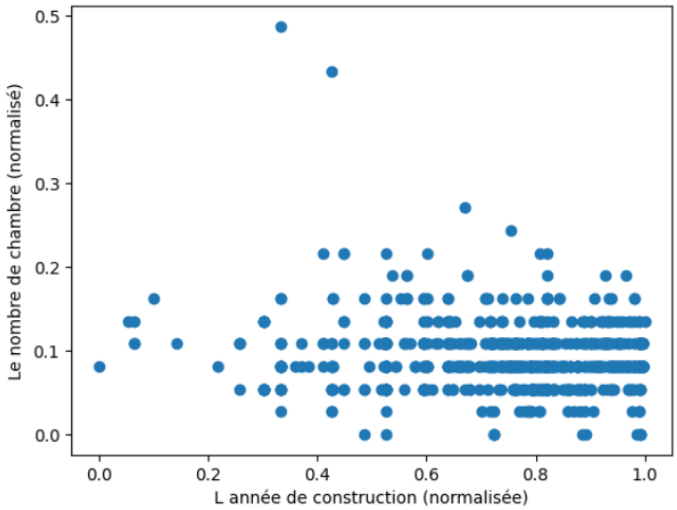
Nombre de chambres vs surface nette habitable

On peut dire que s'il y a une corrélation positive entre la surface nette habitable et le nombre de chambres, cela signifie que lorsque la surface nette habitable augmente, il est plus probable que le nombre de chambres augmente également.



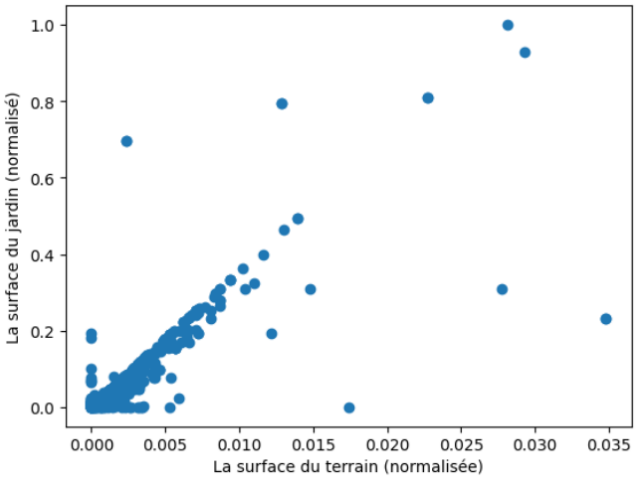
Nombre de chambres vs année de construction

Il n'y a pas de relation linéaire directe entre l'année de construction et le nombre de chambres, ce qui indique que l'année de construction n'a pas d'effet significatif sur le nombre de chambres dans les biens immobiliers étudiés. Les points sont dispersés sur tout le scatter plot, ce qui suggère qu'il n'y a pas de tendance ou de corrélation claire entre l'année de construction et le nombre de chambres. En somme, l'analyse montre que l'année de construction n'est pas une variable déterminante pour le nombre de chambres dans les biens immobiliers étudiés.



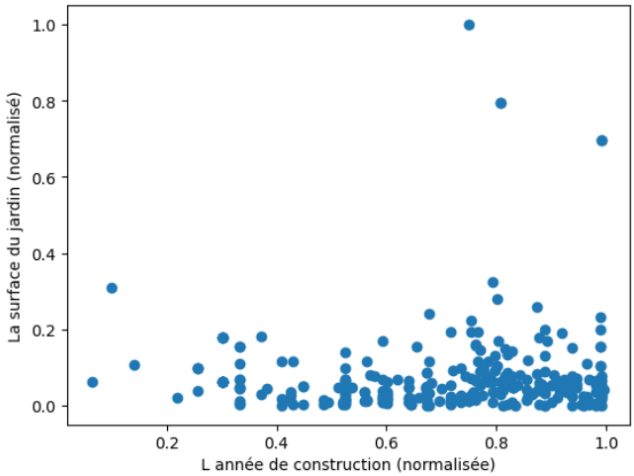
Surface du jardin vs taille de la propriété

L'analyse du scatter plot met en évidence une relation linéaire directe entre la surface du jardin et la surface du terrain. En effet, on peut constater que les points sont alignés selon une droite croissante, ce qui indique que plus la surface du terrain est grande, plus la surface du jardin est grande également. Cette corrélation positive entre ces deux variables suggère que la surface du terrain est un facteur important dans la détermination de la surface du jardin pour les biens immobiliers étudiés.



Surface du jardin vs année de construction

Il n'y a pas de relation linéaire directe entre l'année de construction et la surface du jardin, ce qui indique que l'année de construction n'a pas d'effet significatif sur la surface du jardin dans les biens immobiliers étudiés. Les points sont dispersés sur tout le scatter plot, ce qui suggère qu'il n'y a pas de tendance ou de corrélation claire. En somme, l'analyse montre que l'année de construction n'est pas une variable déterminante pour la surface du jardin dans les biens immobiliers étudiés.

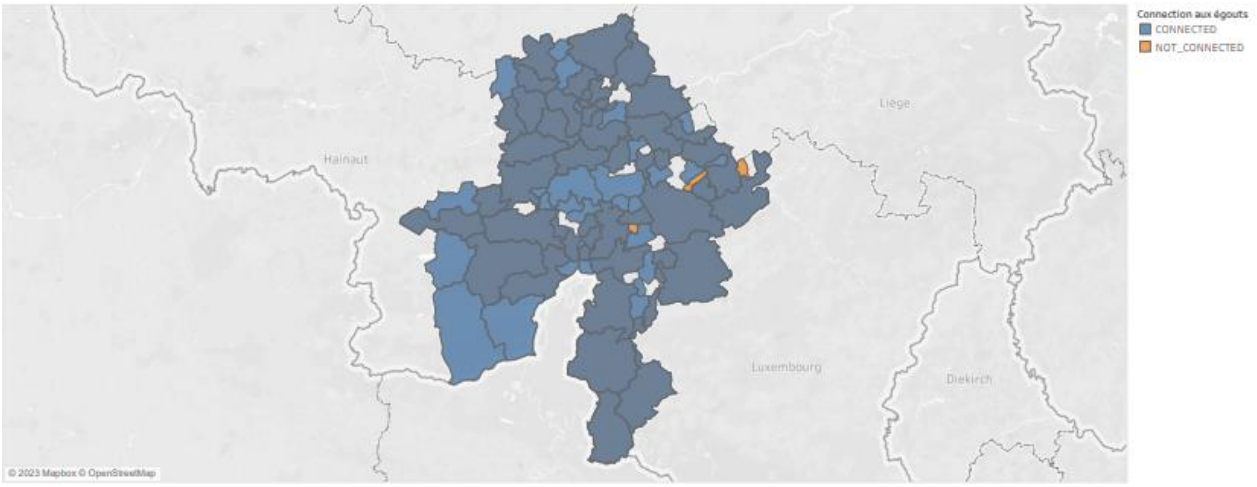


Raccordement aux égouts vs emplacement du bien

Le dégradé de couleur montre la proportion de maisons raccordées ou non raccordées au réseau d'égout dans chaque zone géographique.

Nous constatons que trois régions géographiques sont moins desservies. Le nombre de maisons non raccordées au réseau d'égout sont plus fréquentes. Sont-elles des zones isolées ? Rurales ? Il peut être intéressant d'obtenir une carte des zones humides (et/ou à risque d'inondation) à mettre en parallèle.

Il est important de noter que la corrélation ne signifie pas nécessairement une relation de cause à effet. D'autres facteurs peuvent influencer le raccordement au réseau d'égout, tels que l'âge de la maison ou la proximité des conduites d'égout. Il est donc important de prendre en compte d'autres facteurs pertinents lors de l'analyse des données et de l'interprétation des résultats de la carte choroplèthe.



5.4.2.5. Clusters

Générer des clusters va nous permettre de trouver et d'analyser des groupements d'observation qu'on ne voit pas forcément au premier regard. Dans les graphiques de clustering, les différents points représentent les observations individuelles du jeu de données. Chaque point est attribué à un cluster spécifique, qui est représenté par une couleur différente. Les points regroupés ensemble dans un même cluster sont censés avoir des caractéristiques similaires. La distance entre les points sur le graphique est une mesure de leur similarité, et plus la distance est grande, moins ils sont similaires.

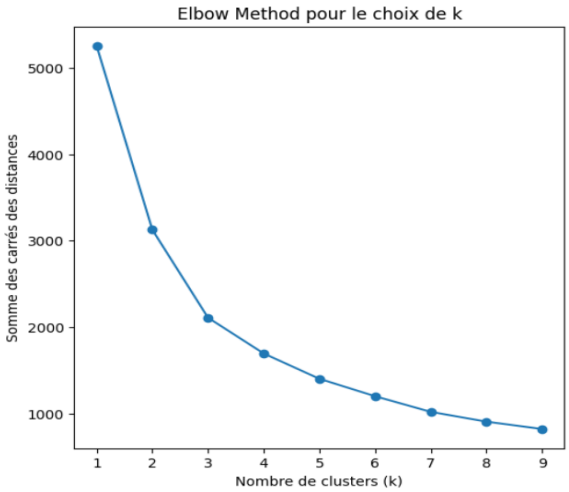
Nous avons remarqué lors des points précédents (de 4.4.3.1 à 4.4.3.4) que des variables importantes sont le prix, la surface nette habitable et le nombre de chambres. Nos clusters se baseront donc sur ces variables.

Cluster surface habitable nette et prix

⇨ Choisir le nombre approprié de clusters

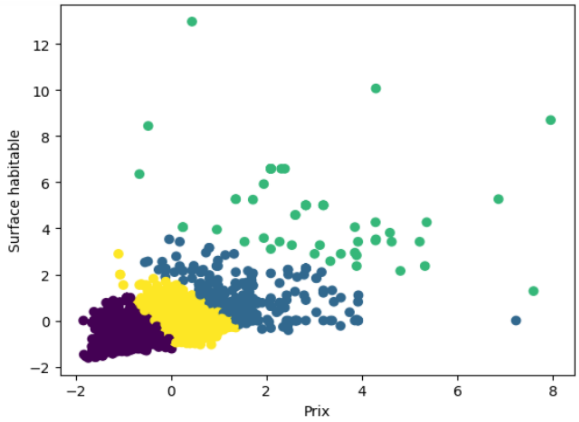
L'interprétation de l'elbow plot se fait visuellement en cherchant le point d'inflexion, c'est-à-dire le point où la courbe cesse de diminuer brusquement et devient plus plate. Ce point marque le nombre optimal de clusters à utiliser.

Ici, le coude est 3 ou 4. Nous choisirons 4.



⇨ *Les clusters*

Les variables ont été normalisées pour pouvoir être comparées. Nous pouvons maintenant identifier des groupes de maisons avec des caractéristiques similaires en termes de prix et de surface habitable.



Nous identifions donc 4 groupes :

	« Bas »	« Moyen-bas »	« Moyen-haut »	« Haut »
Nombre de maisons	1.092	1.231	252	53
Prix moyen	157.950,07 €	297.926,65 €	480.769,98 €	678.094,34 €
Surf. habitable moy.	131.68 m²	176,01 m²	272,09 m²	626,21 m²
Prix min.	15.000 €	115.000 €	190.750 €	175.000 €
Prix max.	269.000 €	450.000 €	1.250.000 €	1.349.000 €
Surf. min.	22 m²	77 m²	136 m²	297 m²
Surf. max.	270 m²	450 m²	510 m²	1.405 m²

Nous constatons que le groupe 4 contient des écarts très importants entre ces différents minimaux et maximaux. Il est probable que ce groupe contienne des observations à caractère particuliers, des valeurs aberrantes ou des erreurs.

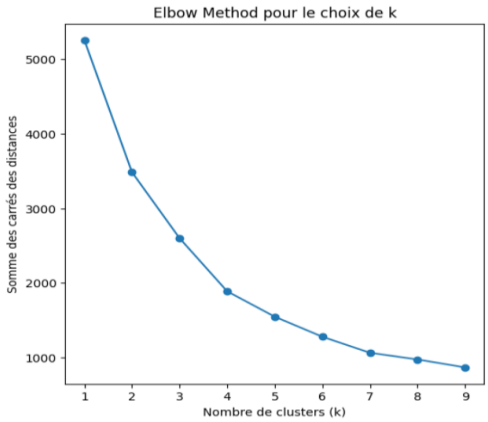
Ces résultats suggèrent qu'il existe une relation entre la surface habitable et le prix des maisons, avec une augmentation significative du prix moyen des maisons dans les groupes ayant une plus grande surface habitable. Les résultats montrent également que les maisons du groupe « Haut » ont des prix moyens considérablement plus élevés que les autres groupes, indiquant qu'il s'agit de maisons haut-de-gamme. Les groupes « Bas » et « Moyen-bas » sont susceptibles de contenir des maisons moins chères et plus petites, tandis que les groupes « Moyen-haut » et « Haut » sont plus susceptibles de contenir des maisons plus grandes et plus chères.

Cluster nombre de chambres et prix

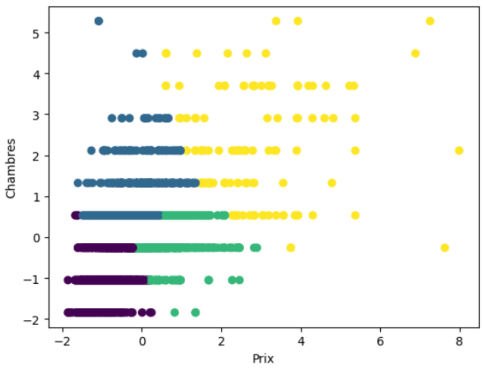
⇨ *Choisir le nombre approprié de clusters*

L'interprétation de l'elbow plot se fait visuellement en cherchant le point d'inflexion, c'est-à-dire le point où la courbe cesse de diminuer brusquement et devient plus plate. Ce point marque le nombre optimal de cluster à utiliser.

Ici, le coude est 4.



⇨ *Clusters*

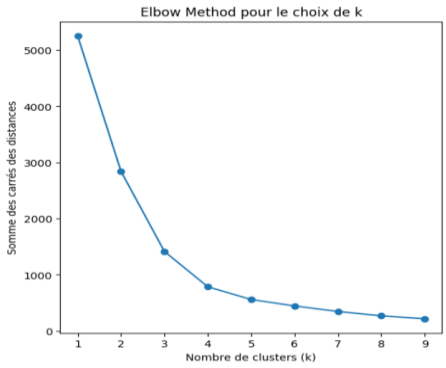


Nous identifions donc 4 groupes :

	« Bas-gauche »	« Bas-droite »	« Haut-gauche»	« Haut-droite »
Nombre de maisons	1.188	1.014	730	157
Prix moyen	165.772,38 €	338.573,96 €	250.322,02 €	633.099,24 €
Nb. chambre moy.	2,4	3,13	4,53	6,03
Prix min.	15.000 €	242.500 €	49.000 €	349.000 €
Prix max.	300.000 €	659.500 €	449.999 €	1.349.000 €
Nb. min.	1	1	4	3
Nb. max.	4	4	10	10

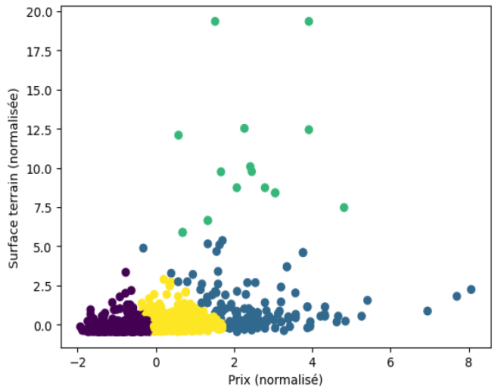
Cluster surface terrain et prix

⇨ *Choisir le nombre approprié de clusters*



⇨ *Clusters*

L'augmentation de la taille de la surface du terrain à un faible impact sur les deux premiers groupes, le troisième étant plus affecté. En ce qui concerne le quatrième groupe, une augmentation de la surface du terrain entraîne une augmentation du prix. Ce quatrième groupe mériterait d'être analysé plus en profondeur pour déterminer quel facteur est à l'origine de ce phénomène (localisation, zone urbaine, zone rurale, …).



Nous identifions donc 4 groupes :

	« Gauche »	« Milieu »	« Droite»	« Haut »
Nombre de maisons	1.295	1.233	149	17
Prix moyen	174.297,14 €	329.511,28 €	620.814,57 €	575.117,65 €
Surf. terrain moy.	588,85 m²	1.094,63 m²	3.558,09 m²	26.712,88 m²
Prix min.	15.000 €	225.000 €	230.000 €	350.000 €
Prix max.	255.000 €	499.000 €	1.349.000 €	915.000 €
Surf. min.	15 m²	16 m²	170 m²	16.000 m²
Surf. max.	9.567 m²	8.427 m²	14.693 m²	50.000 m²

Ces groupes suggèrent une corrélation entre la surface de terrain et le prix des biens immobiliers étudiés. Les groupes « gauche » et « milieu » correspondent à des maisons avec des surfaces de terrain relativement modestes, tandis que les groupes « droite » et « haut » correspondent à des maisons avec des surfaces de terrain plus grandes. Les maisons du groupe « gauche » ont un prix moyen inférieur à ceux des groupes « milieu », « droite » et « haut », qui ont des prix moyens de plus en plus élevés. Les maisons du groupe « haut » ont la plus grande surface de terrain moyenne et le prix moyen le plus élevé, ce qui suggère une forte corrélation entre la surface de terrain et le prix dans ce groupe. Les groupes « gauche » et « milieu » ont des surfaces de terrain moyennes plus modestes, mais il y a une variation importante des prix, ce qui peut suggérer l'importance d'autres facteurs, tels que l'emplacement, la taille de la maison ou l'âge de la construction.

5.4.2.6. Conclusion

Cette AED nous a permis de mieux comprendre le jeu de données « house-for-sale.csv version V1 » que nous avons construit sur la base de notre classification théorique de l'observation « maison » (*5.2.3.2. Définition, classification et sous-classes théoriques*). Nous avons extrait ce jeu de données en appliquant le filtre sur l'ensemble des récoltes (*5.2.3.2. Définition, classification et sous-classes théoriques*). Toutefois, cette analyse a ses limites, en raison des limites intrinsèques du jeu de données (*4.4.1. Limites*).

En ce qui concerne les principales conclusions de cette AED, nous avons observé une forte dispersion des variables des maisons à vendre, notamment le prix qui varie considérablement. Nous avons identifié certaines variables qui influencent la variation des prix, telles que la surface nette habitable ou le nombre de chambres.

L'analyse par cluster a révélé l'existence de groupes de variables similaires dans la distribution du prix, de la surface nette habitable et de la surface du terrain. Ces informations sont précieuses et serviront de base à une segmentation lors d'analyses futures.

En partant de nos hypothèses de base, nous avons appris qu'il n'y avait pas de lien évident entre l'année de construction et la taille du jardin. Cependant, nous avons utilisé des cartes choroplèthes qui nous ont révélé le lien qui existe entre le prix moyen et la zone géographique.

Nous avons également détecté des valeurs aberrantes et des erreurs dans certains scatter plots, qui ne contenaient que peu de valeurs. Cette AED nous permettra donc d'affiner notre jeu de données en appliquant des filtres, tels que la taille du terrain et la surface nette habitable, par exemple.

Pour les points qui manquent de représentativité, tels que le nombre d'occurrences par localité dans la carte choroplèthe, il sera intéressant de régénérer le graphique après avoir collecté plusieurs jours de données supplémentaires.

L'analyse par cluster nous a aussi montré certains groupes qui méritent une analyse plus approfondie en raison de leur faible nombre ou de leur position différente dans le graphique. Ces groupes présentent des sensibilités différentes à l'augmentation de la variable présente sur l'autre axe, comme le cluster de la surface du terrain et du prix, par exemple. Nous pourrons approfondir l'analyse de ces groupes dans des travaux de recherche futurs.

En somme, cette AED nous a fourni des informations précieuses sur les variables du jeu de données house-for-sale.csv version V1, et a également identifié des axes d'analyse pour des travaux futurs.

5.4.3. Conclusion générale

En conclusion, les analyses exploratoires de données sont un outil essentiel pour découvrir les caractéristiques et les relations dans les ensembles de données. Elles permettent de trouver des tendances, des anomalies et des relations qui peuvent servir de base à des analyses futures plus poussées.

Nous avons pu émettre des hypothèses et poser des questions. Nous avons également pu identifier les failles dans notre structure de données récoltées et les pistes d'amélioration. Nous avons pu confirmer la sélection d'outils de calculs et d'affichage statistique. Ces analyses sont donc essentielles pour préparer les données en vue d'analyses plus poussées et elles sont une étape importante pour comprendre les données.

Cette étape de l'analyse exploratoire doit donc être répétée pour chacune des classes que nous voulons étudier. Vous avez trouvé dans ce mémoire au point l'AED de la sous-classe « maison à vendre » et vous trouverez en annexe *12.2.1* l'AED de la sous-classe « appartement à louer ».

6. Validation des résultats

6.1. Processus de validation

Nous voulons identifier les sources officielles qui distribuent les informations sur le marché immobilier et nous voulons confronter nos résultats. Cette démarche permet de valider l'utilisation du scraping. Si les segments de marché couverts par les méthodes traditionnelles et le résultat du scraping produisent des données équivalentes, alors cela donnera de la légitimité au scraping pour donner des résultats sur les segments de marché non-couverts par les méthodes traditionnelles.

Par exemple, un segment couvert par l'office des notaires est l'augmentation du prix de vente des maisons par région. Nous allons couvrir ce même segment avec le scraping et nous pourrons alors comparer les résultats.

Nous avons identifié deux acteurs importants : *Statbel, l'office belge de statistique* et *l'association des notaires*

Statbel

Les informations publiées par Statbel sont, par exemple, des statistiques sur les ventes des maisons ou appartements : nombre et prix de vente par date, superficie et type de bâtiment.

Région / Année	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
	Nombre de ventes	Nombre de ventes	Nombre de ventes	Nombre de ventes	Nombre de ventes	Nombre de ventes	Nombre de ventes	Nombre de ventes	Nombre de ventes	Nombre de ventes	Nombre de ventes	Nombre de ventes	Nombre de ventes
Région flamande	63.566	68.748	64.800	64.669	77.353	60.597	74.728	77.603	81.188	99.334	75.603	90.696	.
Région de Bruxelles-Capitale	9.668	10.516	10.431	9.725	10.099	9.941	10.359	11.289	11.174	11.706	10.501	12.484	.
Région wallonne	32.459	34.049	31.961	30.726	32.896	32.979	33.977	34.959	37.964	38.627	36.395	40.142	.

FIGURE 16 - Tableau StatBel [29]

Nous pouvons également retrouver le prix médian des maisons quatre façades par municipalité, arrondissement, province et région.

Cet acteur est intéressant puisqu'il est gouvernemental et est donc en mesure de disposer des données à leur source. Cependant, cela peut être à la fois un point fort et un point faible, car l'accès aux informations dépend de la confiance que l'on accorde à cette organisation. Il convient de noter que les actes notariaux de moins de 100 ans ne sont pas accessibles librement dans les archives de l'État, conformément à l'article 62 de la loi sur la fonction notariale du 25 ventôse an XI [30].

Association des notaires

Via le site notaire.be, l'association des notaires publie régulièrement des informations sur le marché immobilier en Belgique. Nous pouvons trouver des informations par rapport aux ventes de maisons, hausses ou baisses de prix des appartements de 2 chambres.

6.2. Validation

Puisque Statbel a un temps de retard important sur le marché, nous ne pourrons pas l'utiliser pour valider nos informations. Les dernières informations publiées par Statbel remontent au 12 décembre 2022 et sont relatives au Q3 2022. En ce qui nous concerne, nous avons les informations du Q1 2023.

Nous rencontrons le même problème avec l'association des notaires. Tout comme pour Statbel, les informations ne sont pas publiées en temps réel. En date de cette rédaction (avril 2023), le document le plus récent est une analyse de l'année 2022.

7. Aide à la décision pour les collectivités locales

L'aide à la décision est un enjeu majeur pour les collectivités locales qui cherchent à anticiper les évolutions de leur marché et à prendre les décisions les plus pertinentes pour leur territoire. Dans ce contexte, la mise en place d'un système d'aide à la décision est une étape essentielle pour les aider à atteindre leurs objectifs. Un tel système doit être en mesure d'évaluer la performance du marché, de prévoir son activité future et de réaliser des analyses d'impact pour évaluer les conséquences de chaque décision prise. Dans ce chapitre, nous allons explorer et aborder des thématiques nécessaires et préalables à la création d'une plateforme web de type SSD.

7.1. Evaluation de la performance du marché immobilier

Un système d'aide et de support à la décision doit être en mesure d'offrir et de combiner divers outils et techniques pour permettre aux agents urbanistiques d'évaluer les performances du marché immobilier. Plusieurs indicateurs peuvent être utilisés à cet effet, tels que :

L'indice des prix

Il est utilisé pour suivre les changements de prix dans le temps. Les collectivités locales peuvent l'utiliser pour analyser les tendances de prix dans leur juridiction et identifier les zones de fortes demandes.

Le volume des biens en vente

Il est utilisé pour suivre le nombre de biens en vente sur une période donnée.

L'analyse spatiale

Les systèmes d'information géographique (SIG) pour visualiser et analyser les modèles spatiaux des biens immobiliers. Les collectivités locales peuvent l'utiliser pour identifier des zones qui nécessitent un redéveloppement ou une revitalisation.

La segmentation de marché

Consiste à regrouper les biens en fonction de certaines propriétés (emplacement, type, taille, etc.). Les collectivités locales peuvent utiliser cette méthode pour identifier des groupements de biens et analyser l'impact éventuel de leurs décisions.

L'analyse de la perception du public

Permet aux collectivités locales d'obtenir des commentaires et des réactions des citoyens, offrant ainsi des informations complémentaires aux statistiques. Cette approche permet d'identifier les zones d'inquiétude ou d'incompréhension.

7.2. Prévision de l'activité du marché immobilier

En outre, il est important de noter que l'évaluation de la performance du marché immobilier ne doit pas se limiter à l'analyse des données historiques, mais doit également inclure des prévisions futures. Les modèles de prévision peuvent aider les collectivités locales à planifier des projets futurs, à anticiper les fluctuations du marché et à prendre des décisions éclairées pour maximiser les avantages pour la communauté.

Il existe un grand nombre de techniques pour prédire l'activité sur le marché. Nous n'allons pas approfondir ces techniques, mais nous allons présenter brièvement certaines d'entre elles qui pourraient être utilisées.

Time Series Analysis

Utilisation de modèles statistiques pour analyser et prédire la tendance du marché à travers le temps. Cette technique pourrait être utilisée pour prédire de futures demandes et donc de futurs développements urbanistiques.

Artificial Neural Networks

Des algorithmes de machine learning qui sont capables d'apprendre des patterns complexes dans les données. Cela devrait permettre de pouvoir trouver des patterns et des relations qui sont difficiles à identifier autrement.

Expert Systems

Des programmes qui simulent la capacité de prise de décision humaine. À terme, un système expert pourrait fournir des recommandations personnalisées aux collectivités locales en fonction de leurs besoins et de leurs objectifs.

7.3. Analyse de l’impact du marché immobilier

Le marché immobilier est un élément crucial pour les collectivités locales, qui ont un rôle important à jouer dans sa gestion. En effet, l'augmentation de l'offre immobilière sur un territoire peut avoir un impact significatif sur plusieurs aspects de la vie locale, notamment sur les recettes fiscales, les infrastructures, l'impact social et environnemental.

En termes de recettes fiscales, il est important pour les collectivités locales de suivre l'évolution du marché immobilier afin d'évaluer la contribution de celui-ci à leurs flux de recettes et d'identifier les domaines de croissance potentielle. Cela leur permet de prendre des décisions justifiées en matière d'investissements et de politiques publiques, tout en veillant à la durabilité financière de la collectivité.

En ce qui concerne les infrastructures, il est essentiel que les collectivités locales surveillent de près l'évolution du paysage urbanistique pour s'assurer que les infrastructures, telles que les transports et autres services publics, soient en adéquation constante avec les besoins de la population. Des investissements dans les infrastructures sont nécessaires pour accompagner l'évolution de l'offre immobilière et pour garantir la qualité de vie des citoyens.

L'impact social de l'offre immobilière doit également être pris en compte. En effet, l'évolution de la proportion de maisons et d'appartements peut entraîner des conséquences importantes sur le type de population présente dans une juridiction, ainsi que sur l'accessibilité au logement. Les politiques et les programmes doivent être adaptés aux besoins de la population locale pour éviter les changements brutaux et garantir une cohésion sociale.

Enfin, l'impact environnemental est un autre aspect crucial à prendre en compte dans la gestion du marché immobilier. Les collectivités locales doivent être en mesure d'analyser l'impact qu'aura un changement de loi ou de réglementation sur l'environnement, tout en veillant à respecter les normes en vigueur à long terme. La durabilité environnementale est un enjeu majeur pour les collectivités locales, qui ont un rôle important à jouer dans la protection de l'environnement et la lutte contre le changement climatique.

En somme, la gestion du marché immobilier est un enjeu complexe qui nécessite une approche multidimensionnelle et une analyse approfondie de l'impact sur les différentes dimensions de la vie locale. Les collectivités locales ont un rôle crucial à jouer dans la mise en place de politiques et d'investissements durables, qui prennent en compte les besoins de la population, tout en garantissant la durabilité financière, sociale et environnementale de la collectivité.

7.4. Système de support d’aide à la décision

Comme vu dans les points précédents, l’activité du marché immobilier peut avoir un impact significatif sur le processus décisionnel des collectivités locales, car elle concerne des domaines

tels que la politique du logement, l'aménagement du territoire et le développement économique. Les collectivités locales doivent donc avoir accès à des informations opportunes et précises sur l'activité du marché immobilier pour prendre des décisions éclairées.

Les planificateurs urbanistiques ont différentes politiques à mener, telles que :

- politique d'habitat abordable ;
- politique d'aménagement du territoire ;
- politique d'aménagement des taxes sur la propriété ;
- politique d'aménagement des transports et des infrastructures ;
- politique environnementale ;
- etc.

Description de ce que pourrait être le système

Nous proposons un système de support d'aide à la décision personnalisé qui permettrait à chaque utilisateur de construire son propre tableau de bord en fonction de ses besoins et de ses préférences. L'interface serait conviviale et facile d'utilisation, avec une variété d'options disponibles pour afficher des données pertinentes telles que les indicateurs clés de performance, les tendances du marché immobilier, les projections de croissance de la population et les changements réglementaires importants. En outre, le système serait capable de fournir des alertes et des notifications en temps réel en cas de changements significatifs tels qu'une augmentation de l'offre ou de la demande de logements, des fluctuations des taux d'intérêt ou des évolutions réglementaires. Les alertes seraient personnalisables en fonction des besoins de chaque utilisateur, avec la possibilité de recevoir des notifications par e-mail ou via l'interface elle-même. Enfin, le système pourrait également inclure des outils de modélisation avancés pour aider les utilisateurs à explorer différents scénarios et à évaluer les conséquences de différentes décisions. Par exemple, les utilisateurs pourraient simuler les effets d'une nouvelle politique de logement sur le marché immobilier local ou prévoir les impacts financiers de différents choix d'investissement. Ces outils

permettraient aux utilisateurs de prendre des décisions éclairées et de mieux comprendre les impacts de leurs choix sur la collectivité locale.

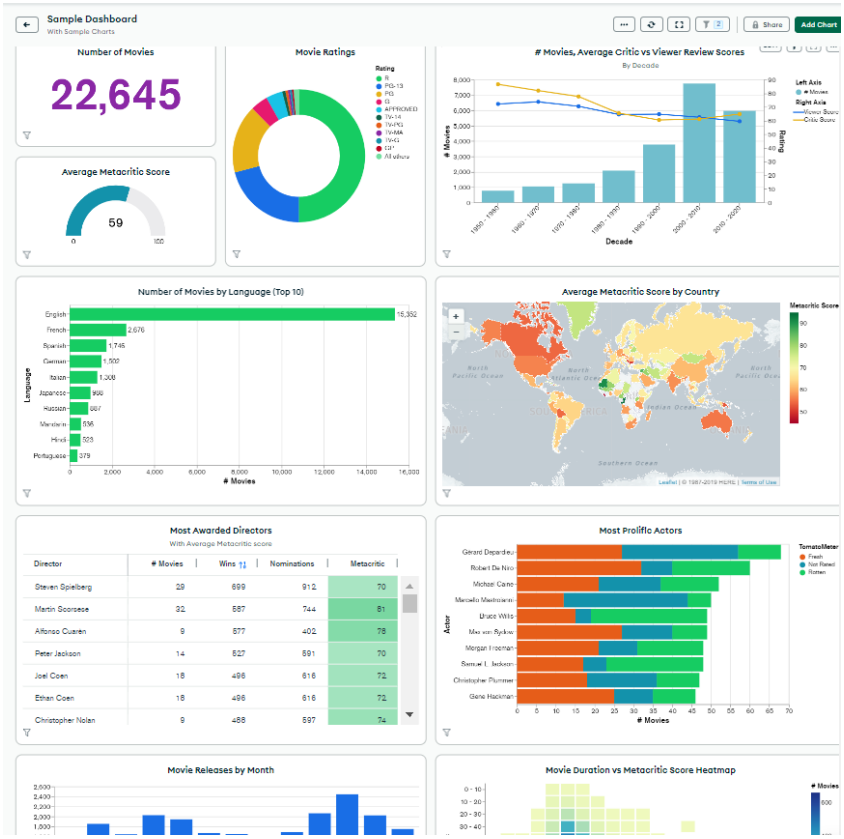


Figure 17 - Exemple de dashboard

7.4.1.Analyse croisée

Pour obtenir des données plus utiles, il est essentiel d'agréger des sources de données autres que la base de données des petites annonces. L'interface devrait donc permettre aux utilisateurs de croiser les informations provenant de différentes sources de données. Par exemple, il pourrait être possible de suivre l'évolution de la proportion croissante d'appartements dans une commune et son impact sur la qualité de l'air. Nous pourrions aussi suivre l'évolution du taux d'emploi dans une localité par rapport à l'évolution du prix de la location des appartements dans cette même localité à l'aide du système d'information géographique (SIG). Le système pourrait également permettre de suivre l'impact d'une hausse des taux d'intérêt bancaires sur le nombre de petites annonces en cours, ou encore de suivre l'évolution démographique par rapport aux proportions de types de bâtiments nouvellement construits.

Il existe un nombre impressionnant de combinaisons possibles à faire sur la base de toutes les informations disponibles. Par exemple, il pourrait être possible de croiser les informations sur la variation de la proportion des maisons unifamiliales avec les informations sur la qualité de l'air, les zones inondables, les informations sur les services publics (emplacement, nombre d'effectifs, etc.),

les informations sur l'emploi, les informations bancaires (taux d'emprunt, etc.), les informations sur les taux de criminalité, les informations « smart cities » ou encore les informations sur l'évolution démographique. La liste de combinaisons possibles est non exhaustive et dépendra des besoins de l'utilisateur et des sources de données disponibles.

7.4.2.Des exemples rapides de cas d'utilisation

Imaginons que la région Wallonne décide d'interdire la construction de nouvelle maison 4 façades.

Effectivement, si la Région wallonne décide d'interdire la construction de nouvelles maisons 4 façades, cela aura un impact direct sur le marché immobilier local. Les constructeurs ne pourront plus bâtir de nouvelles maisons répondant à ces critères, ce qui pourrait entraîner une diminution de l'offre de ce type de biens sur le marché.

Le planificateur urbanistique aura donc la tâche de suivre l'évolution des biens immobiliers en appliquant des filtres sur les données récupérées en temps réel. Il pourra définir des classes de biens en fonction de critères tels que la taille, le type de bâtiment, etc. et suivre leur évolution au fil du temps. Par exemple, il pourrait créer une première classe de maisons 4 façades et une deuxième classe pour les maisons de moins de 4 façades.

En utilisant les graphiques obtenus à partir de ces classes de biens, le planificateur pourra monitorer l'effet de l'interdiction de construction de nouvelles maisons 4 façades sur le marché immobilier local. Il pourra ainsi constater s'il y a des variations statistiques (médian, moyenne, etc.) pour les classes qu'il a lui-même constituées. S'il constate une diminution de l'offre de maisons 4 façades, il pourra proposer des solutions pour pallier ce manque, telles que l'encouragement de la construction de maisons de moins de 4 façades ou la promotion de la rénovation de maisons existantes.

En somme, le suivi de l'évolution des biens immobiliers en temps réel permet au planificateur urbanistique d'anticiper les effets d'une action politique sur le marché immobilier et d'inciter à la prise de décisions en réaction. Cela lui permet également d'adapter sa stratégie en fonction de l'évolution du marché et d'assurer une gestion efficace et durable de l'urbanisme local.

Imaginons que le planificateur urbanistique définisse des classes types pour les biens immobiliers de sa juridiction.

Une fois que le planificateur urbanistique a défini les classes types pour les biens immobiliers de sa juridiction, il peut créer un tableau de bord pour suivre leur évolution en temps réel. Il peut utiliser des indicateurs tels que le nombre de transactions, le prix moyen, la taille du bien, le type de bâtiment, etc. pour suivre l'évolution de chaque classe au fil du temps.

En utilisant des graphiques et des tableaux de bord, le planificateur urbanistique peut observer les tendances et les fluctuations dans chaque classe de biens immobiliers. S'il constate des changements significatifs dans l'un des indicateurs, tels qu'une augmentation des prix, il peut enquêter sur les causes de cette augmentation. Par exemple, il peut vérifier s'il y a eu une augmentation de la demande pour les biens immobiliers de cette classe ou s'il y a eu une baisse de l'offre de biens similaires sur le marché.

De plus, en plaçant des alertes dans le tableau de bord, le planificateur urbanistique peut être informé en temps réel des changements importants dans les données. Cela lui permettra d'agir rapidement et de prendre des décisions en réponse à ces changements.

En somme, la surveillance des classes types pour les biens immobiliers peut aider les planificateurs urbains à suivre l'évolution des tendances du marché immobilier et à prendre des décisions pour mieux gérer le développement urbain de leur juridiction.

Imaginons que la Région wallonne oblige les bailleurs à obtenir un PEB correct à partir d'une date donnée

L'obligation pour les bailleurs d'obtenir un PEB correct peut avoir un impact sur les prix des locations, car cela peut entraîner des coûts supplémentaires pour les propriétaires qui doivent réaliser des travaux pour se conformer à la nouvelle réglementation. En suivant l'évolution des prix des locations dans la région, le planificateur urbanistique pourra observer s'il y a une augmentation ou une diminution des loyers après la mise en place de la nouvelle réglementation. Si une augmentation significative des loyers est constatée, il pourra enquêter pour déterminer si cette augmentation est liée à la nouvelle réglementation ou à d'autres facteurs tels qu'une augmentation de la demande de logements dans la région. En fin de compte, l'analyse des données sur les prix des locations permettra au planificateur urbanistique de mieux comprendre les effets de la réglementation sur le marché locatif et d'ajuster si nécessaire les politiques publiques en conséquence.

Imaginons que l'on souhaite suivre l'évolution d'une catastrophe comme les inondations à Liège en 2022

En cas de catastrophe naturelle, comme les inondations qui ont touché Liège en 2022, le planificateur urbanistique peut suivre l'impact de l'événement sur les biens immobiliers environnants en utilisant des données en temps réel. Par exemple, il peut suivre les changements dans les prix de l'immobilier dans la région touchée, la variation du nombre de maisons en vente ou en location, la vitesse à laquelle les maisons sont réparées ou reconstruites, etc.

À partir de ces données, le planificateur urbanistique pourra déterminer les actions à mettre en œuvre pour aider les propriétaires affectés. Par exemple, s'il constate une baisse des prix de l'immobilier dans la région touchée, il peut recommander des aides financières pour les

propriétaires qui cherchent à vendre ou à louer leur maison. S'il constate que le nombre de maisons en vente a augmenté, il peut recommander des mesures pour stimuler la demande, comme des incitations fiscales pour les acheteurs.

En outre, en suivant l'impact de l'événement sur les biens immobiliers, le planificateur urbanistique peut aider à déterminer les politiques publiques à mettre en place pour prévenir ou atténuer les impacts de futurs événements similaires. Par exemple, il peut recommander des mesures de prévention ou de gestion des risques pour réduire les dégâts causés par les inondations dans la région.

Imaginons que nous voulons suivre l'évolution du type de building dans une région

Lorsque l'on suit l'évolution du type de bâtiment dans une région, il est important de considérer plusieurs facteurs. Si l'on constate une augmentation du nombre d'appartements construits mais que, simultanément, le prix augmente considérablement, il est important de comprendre pourquoi cela se produit.

Tout d'abord, il convient de se demander si les nouveaux appartements sont plus luxueux que les précédents. Si tel est le cas, cela pourrait expliquer l'augmentation des prix. En effet, des appartements plus haut de gamme offrent souvent plus de commodités et de prestations de qualité, ce qui justifie un prix plus élevé.

Ensuite, il est possible que l'offre de nouveaux appartements ne soit pas suffisante pour répondre à la demande croissante. Dans ce cas, les prix pourraient augmenter malgré l'augmentation de l'offre, car la demande est plus forte que l'offre. Cela peut se produire si la région connaît une croissance économique ou démographique rapide, ou si la région devient plus attractive pour les investisseurs immobiliers.

Enfin, il est également possible qu'il y ait un changement d'attractivité de la région ou d'une région voisine. Si une région voisine devient moins attractive en termes de prix, de commodités ou d'offre, cela peut augmenter la demande pour des appartements dans la région en question. Cela pourrait expliquer l'augmentation des prix malgré l'augmentation de l'offre.

En somme, suivre l'évolution du type de bâtiment dans une région est important pour comprendre les tendances du marché immobilier et anticiper les changements à venir. En examinant attentivement les facteurs qui influencent l'offre et la demande, on peut obtenir une vision claire de la situation et prendre les décisions appropriées.

8. Travaux futurs

Extension à d'autres régions

Effectivement, l'extension du système de scraping à d'autres régions peut être très utile pour répondre à des questions plus larges et avoir une vue d'ensemble plus complète. Cela permettrait également de comparer les données entre différentes régions et d'identifier des tendances et des disparités géographiques.

Pour étendre le système, il faudrait identifier les sources de données appropriées pour chaque région et adapter les scripts de scraping en fonction. Il est également important de tenir compte des différences régionales en matière de législation et de réglementation immobilière, car cela pourrait avoir un impact sur la disponibilité et la qualité des données.

Intégrer une source alternative de données

L'intégration de sources alternatives de données pourrait être une étape cruciale pour améliorer la qualité des données collectées par le scraping. Les réseaux sociaux, en particulier, peuvent être une source riche en informations pour les collectivités locales en quête d'informations sur le marché immobilier. Les commentaires et les vidéos publiés par les utilisateurs sur ces plateformes peuvent fournir une vue plus complète des tendances du marché immobilier, des préférences des acheteurs et des vendeurs, ainsi que des opinions sur les développements locaux et les projets de construction.

En intégrant ces données alternatives dans les systèmes de scraping existants, les collectivités locales pourraient bénéficier d'une vue plus complète et plus précise de l'état du marché immobilier, ce qui leur permettrait de prendre des décisions justifiées. Cependant, il est important de noter que l'intégration de ces sources de données alternatives nécessite également des compétences techniques supplémentaires et une analyse de données plus sophistiquée pour filtrer et interpréter les données collectées. Par conséquent, les collectivités locales devraient être prêtes à investir dans des ressources supplémentaires pour intégrer ces sources alternatives de données.

Utiliser des algorithmes de machine learning

L'utilisation d'algorithmes de machine learning est une piste prometteuse pour améliorer l'efficacité et la précision des systèmes de scraping en temps réel dans le domaine immobilier. En effet, ces algorithmes peuvent aider à détecter des modèles et des tendances dans les données immobilières collectées, ce qui peut aider les collectivités locales à prendre des décisions.

Par exemple, l'analyse prédictive pourrait être utilisée pour prévoir l'évolution des prix de l'immobilier dans une région donnée en fonction de diverses variables, telles que la demande, l'offre, l'évolution démographique et économique. De même, les algorithmes de clustering peuvent aider à identifier des groupes de biens immobiliers similaires en termes de caractéristiques et de prix, ce qui peut être utile pour la planification de l'aménagement du territoire.

Généraliser le stockage et centraliser les informations scrappées pour les recherches futures

En effet, généraliser le stockage en base de données pourrait être une solution efficace pour faciliter la collecte de données à partir de différents sites Internet. Au lieu de stocker les données dans des formats différents pour chaque site, les données pourraient être stockées dans une structure de base de données commune, permettant ainsi une analyse plus cohérente et une comparaison plus facile entre différents sites. De plus, une telle généralisation faciliterait l'entraînement d'algorithmes pour la collecte de données et la reconnaissance de motifs dans les sites web. Les algorithmes pourraient être entraînés sur une structure de base de données commune et ainsi être plus facilement transférables à différents sites Internet.

Qui plus est, pour des raisons économiques et écologiques, il serait intéressant de centraliser les bases de données issues du scraping afin qu'elles soient mises à disposition des chercheurs.

Définir un cadre légal explicite

Les lois sont obsolètes et peu claires sur la pratique du scraping. Sur base de considérations éthiques, il peut être intéressant de définir un cadre légal autour de la pratique du scraping.

Tableau de bord de gestion des scripts

Implémenter un tableau de bord de gestion des scripts qui permettrait de voir en un coup d'œil si un script a fonctionné ou non. Il devrait également proposer la possibilité d'obtenir un aperçu sur les avertissements et erreurs qui se produisent.

9. Discussion

L'analyse exploratoire de données a démontré qu'un jeu de données récolté grâce aux méthodes de scraping a de la valeur. Notre jeu de données n'est pas parfait, mais nous avons identifié les causes de cette imperfection (*4.4.1. Limites*). Cette démonstration implique que ces types de jeux de données peuvent devenir des alternatives tout-à-fait valables par rapport aux méthodes plus traditionnelles. Le choix d'une méthode ou d'une autre dépendra toujours des impératifs, besoins et ressources de l'étude, mais la méthode du scraping dispose d'avantages qu'aucune autre méthode ne peut proposer : en effet les données que nous avons collectées nous permettent d'obtenir une vue complète et étendue en temps réel.

Nous insistons fortement sur l'importance du traitement humain (analyse préalable, filtrage, nettoyage, etc.) car c'est l'étape clé pour obtenir une étude réussie. C'est une étape délicate dans laquelle la réflexion doit être poussée le plus loin possible, sous peine de générer un jeu de données inexploitable. Nous avons exploré plusieurs façons de valider cette réflexion. Une première façon de valider notre travail est de comparer certains de nos résultats avec les résultats officiels publiés par les institutions gouvernementales (quand c'est possible). Nous estimons que c'est une bonne pratique et cela permet de garantir la fiabilité de notre jeu de données en lui donnant une certaine légitimité.

Il existe une autre façon de valider notre travail, que nous avons apprise de manière empirique et à nos dépends. Nous avons commis l'erreur d'attendre d'avoir un jeu de données conséquent avant de commencer l'analyse exploratoire de données. Il aurait été plus pertinent de mettre à l'épreuve nos récoltes plus tôt pour nous permettre d'ajuster nos schémas de données, par exemple. En faisant cela, nous aurions pu découvrir que notre schéma de données (le document JSON enregistré dans base de données NoSQL) n'était pas assez plat que pour être parcouru facilement par nos outils de visualisation. L'erreur vient du fait que lors des prémices de l'analyse, nous avions envisagé de générer les graphiques dans une interface web avec des langages de programmation. Le schéma de données avait été pensé dans cette optique. Au fur et à mesure de l'avancement du travail, nous avons remarqué que nous n'aurions pas le temps de faire cela et qu'il fallait donc utiliser des outils de visualisations existants pour gagner du temps. Comme nous l'avons expliqué au point *4.4.1. Limites*, nous n'avons donc pas pu profiter des mises à jour enregistrées. Nous avons donc introduit une source d'imprécision. Pour améliorer les futures recherches sur le sujet, si nous devions refaire cette étude, nous n'attendrions pas d'avoir un jeu de données complet avant de commencer l'analyse exploratoire de données. Même si le jeu de données est petit, nous pourrions peut-être identifier les problèmes potentiels et les corriger en temps opportun. Un autre avantage d'une AED précoce est de détecter les informations redondantes ou en double. En effet, nous avons découvert des propriétés qui se chevauchent et auraient pu être fusionnées lors de l'enregistrement des données par le scraper dans notre base de données. Cette approche aurait permis une meilleure qualité des données et des résultats plus précis.

Notons tout de même que même avec un prétraitement parfait, nous ne serions pas à l'abri d'enregistrer des erreurs. En effet, nos sources sont des données web, et qui plus est, entrées par des utilisateurs lambda. Nous insinuons que le post-traitement a tout autant d'importance que le pré-traitement. La présence de doublons ou de valeurs aberrantes fera toujours partie du jeu de données et c'est pourquoi le filtrage est très important pour en assurer la qualité.

De plus, le web est en constante évolution et nous n'avons aucun contrôle sur la source de données : il est possible que nous rencontrions des problèmes en termes de disponibilité. Il sera toujours important d'assurer une bonne maintenance du scraper. Effectivement, les sites sources peuvent être régulièrement mis à jour, ce qui peut le rendre obsolète. Pour cette raison, il est essentiel de concevoir un code robuste, résilient et facile à déboguer.

Bien que les méthodes de scraping puissent être piégeuse et parfois difficiles à apprivoiser, il se pourrait que l'effort ne soit pas vain. Comme dit précédemment, le scraping est souvent la seule méthode pour obtenir des données en temps réel et en quantité impressionnante. Notre AED aura le mérite de démontrer qu'il est possible d'avoir des résultats cohérents et utilisables. Nous pouvons dès lors facilement imaginer qu'une version améliorée de notre jeu de données puisse servir à la création d'un SSD. Nous avons mis sur papier une ébauche de conception en ce sens, et nous imaginons qu'il puisse être un outil d'aide de prise à la décision puissant. La réalisation d'un POC (proof-of-concept) constitue une piste pour des travaux futurs. Nous pourrions imaginer un SSD

dans lequel on croiserait les données immobilières avec d'autres données en temps réel, telles que les données environnementales et il permettrait à ses utilisateurs d'obtenir une image plus complète de la situation immobilière dans une région donnée. Par exemple, si les données immobilières montrent une augmentation de l'offre de logements dans une région donnée, mais que les données environnementales montrent une augmentation de la pollution dans cette région, cela peut indiquer un problème potentiel pour les futurs résidents. De même, si les données immobilières montrent une baisse des prix des logements dans une région, mais que les données sur la qualité de vie montrent une détérioration de la sécurité dans cette région, cela peut indiquer un risque pour les futurs résidents.

Au cours de cette étude, nous n'avons fait qu'effleurer la surface des possibilités offertes par l'utilisation de techniques telles que le machine learning et d'autres outils de prédiction. Nous sommes conscients que l'évaluation de la performance du marché immobilier ne doit pas se limiter à l'analyse des données historiques et en temps réel, mais que la prévision de l'activité du marché immobilier sur la base de nos données récoltées pourrait constituer une étude à part entière. Nous pensons donc que l'exploration de ces techniques constitue une piste d'amélioration pour les futurs travaux de recherche. En outre, nous pourrions envisager d'explorer des sources de données supplémentaires telles que les données des réseaux sociaux, les données météorologiques et les données économiques pour améliorer la précision des prévisions du marché immobilier.

Cette étude permet de mieux comprendre le rôle du web scraping et sa valeur pour la construction d'une base de données. Nous récoltons quotidiennement des milliers d'annonces de manière automatique sur toute la province de Namur. Nous ne savons pas quel est le prix pour un tel résultat avec les méthodes traditionnelles (ni combien de temps cela prendrait), mais ce que nous savons en revanche, c'est le prix que nous avons payé pour notre méthode : une collecte quotidienne de données sur la vente de maisons et la location d'appartements coûte actuellement 0,35 euros, soit 64 euros pour 6 mois de données. Nous pensons pouvoir affirmer que c'est un prix dérisoire.

10. Conclusion

Dans ce mémoire, l'objectif principal était de montrer comment le scraping de données peut être utilisé pour refléter les activités du marché immobilier namurois en temps réel afin de l'intégrer dans un outil SSD et SIG. Pour atteindre cet objectif, nous avons d'abord examiné l'architecture du système de scraping, ainsi que les différentes technologies utilisées pour collecter, stocker et analyser les données. Ensuite, nous avons approfondi les méthodes de collecte de données, en explorant différentes techniques pour extraire des informations pertinentes des pages web. Nous avons également discuté des défis et des limites du scraping de données. Enfin, nous avons présenté les résultats obtenus, qui ont démontré que le scraping de données est capable de fournir des informations précieuses sur le marché immobilier namurois en temps réel. Nous avons également discuté de la manière dont ces résultats pourraient être utilisés pour aider les collectivités locales à

prendre des décisions éclairées en matière de politique immobilière grâce à des données récoltées en temps réel.

Nous avons montré que la collecte de données peut être automatisée grâce aux outils de scraping et aux technologies connexes, permettant ainsi d'obtenir des données de manière plus efficace et plus rapide. En outre, notre analyse exploratoire a démontré que les données récoltées peuvent être utilisées pour fournir des informations utiles sur le marché immobilier, telles que les prix des biens immobiliers, leur emplacement et leur état. Cependant, nous avons également souligné que la mise en place d'un système de surveillance du marché immobilier en temps réel par le scraping demande des compétences multidisciplinaires, notamment en programmation, en analyse de données et en gestion de bases de données.

En effet, notre étude a également identifié plusieurs pistes de recherche futures pour améliorer l'utilisation du scraping de données dans le domaine immobilier. L'une de ces pistes consiste à explorer l'utilisation de techniques de machine learning pour améliorer la qualité des données collectées et l'efficacité du scraping. Par exemple, il pourrait être possible d'utiliser des modèles de machine learning pour reconnaître les structures et les formats de données spécifiques sur les pages web, permettant ainsi de collecter plus efficacement les informations pertinentes. Enfin, nous avons aussi identifié la nécessité de concevoir des systèmes de visualisation et de prise de décision plus conviviaux pour les utilisateurs finaux. Cela permettrait de rendre les données collectées plus accessibles et de faciliter leur utilisation dans la prise de décision.

En conclusion, le scraping de données peut être une méthode très utile pour la recherche car elle permet d'obtenir un grand nombre de données en temps réel, rapidement et à moindre coût. Cette méthode permet également de collecter des données qui ne seraient pas facilement accessibles autrement, en particulier sur des sites web n’ayant pas de bases de données accessibles ou qui ne seraient pas structurées. Cependant, il est important de reconnaître les défis et les limites de cette méthode, tels que la nécessité de nettoyer et de filtrer les données collectées, ainsi que la complexité de la mise en place de systèmes de scraping en temps réel. Enfin, il est important de concevoir des systèmes de visualisation et de prise de décision plus conviviaux pour les utilisateurs finaux afin de maximiser l'impact de ces systèmes dans les décisions de politique immobilière.

11. Bibliographie – Références

[1] Berkes, Fikret & Davidson-Hunt, Iain. (2007). Communities and social enterprises in the age of globalization. Journal of Enterprising Communities: People and Places in the Global Economy.

[2] Mehrabi, Ninareh, Fred Morstatter, Nripsuta Ani Saxena, Kristina Lerman and A. G. Galstyan. “A Survey on Bias and Fairness in Machine Learning.” ACM Computing Surveys (CSUR) 54 (2019): 1 - 35.

[3] Trends - Le Vif. (2014, Mai 6). Cinq facteurs qui influencent les prix de l'immobilier. Récupéré sur https://trends.levif.be/: https://trends.levif.be/immo/cinq-facteurs-qui-influencent-les-prix-de-limmobilier/ (visite 01/02/2023)

[4] L'Echo. (2022, Septembre 26). La colocation pour survivre à l'inflation: comment ça marche? Récupéré sur https://lecho.be/: https://www.lecho.be/monargent/analyse/immobilier/la-colocation-pour-survivre-a-l-inflation-comment-ca-marche/10414127.html/ (visite 12/02/2023)

[5] Floetotto, Max & Kirker, Michael & Stroebel, Johannes. (2016). Government Intervention in the Housing Market: Who Wins, Who Loses?. Journal of Monetary Economics. 80. 10.1016/j.jmoneco.2016.04.005.

[6] Wang, Daikun & Jing, Li. (2019). Mass Appraisal Models of Real Estate in the 21st Century: A Systematic Literature Review. Sustainability. 11. 7006. 10.3390/su11247006.

[7] García, Raúl Tomás & Lopez, Maria Francisca & Pérez Sánchez, Raúl. (2022). Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. Land. 11. 2100. 10.3390/land11112100.

[8] État de l'art – Applicabilité et fondements théoriques du scraping – Félix Barzin – IDS 2021/2022

[9] Soundararaj, BalaMurugan & Pettit, Christopher & Lock, Oliver. (2022). Using Real-Time Dashboards to Monitor the Impact of Disruptive Events on Real Estate Market. Case of COVID-19 Pandemic in Australia. Computational Urban Science. 2. 14. 10.1007/s43762-022-00044-z.

[10] Fava, Davide & Guaragno, Graziella & Dall'Olio, Claudia. (2016). Decision Support Systems for Urban Planning. 10.1007/978-3-319-10425-6_5.

[11] Hromada, E.. (2015). Mapping of Real Estate Prices Using Data Mining Techniques. Procedia Engineering. 123. 233-240. 10.1016/j.proeng.2015.10.083.

[12] Grybauskas, A., Pilinkienė, V. & Stundžienė, A. Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic. *J Big Data* **8**, 105 (2021).

[13] Statbel. (2018, Février 8). Le webscraping, la collecte et le traitement de données en ligne pour l'indice des prix à la consommation. Récupéré sur statbel.fgov.be: https://statbel.fgov.be/fr/nouvelles/le-webscraping-la-collecte-et-le-traitement-de-donnees-en-ligne-pour-lindice-des-prix-la (visite 07/03/2023)

[14] Mutenherwa,F., Wassenaar, D.R., and Oliveira,T.(2019). Ethical Issues Associated with HIV Phylogenetics in HIV Transmission Dynamics Research: A Review of the Literature Using the Emanuel Framework. Developing World Bioethics, 19, 25–35.

[15] Rennie, Stuart & Buchbinder, Mara & Juengst, Eric & Brinkley-Rubinstein, Lauren & Blue, Colleen & Rosen, David. (2020). Scraping the Web for Public Health Gains: Ethical Considerations from a 'Big Data' Research Project on HIV and Incarceration. Public health ethics. 13. 111-121. 10.1093/phe/phaa006.

[16] Mancosu, Moreno & Vegetti, Federico. (2020). What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data. Social Media + Society. 6. 1-11. 10.1177/2056305120940703.

[17] Li, Jiawei & Xu, Qing & Shah, Neal & Mackey, Tim. (2019). A Machine Learning Approach for the Detection and Characterization of Illicit Drug Dealers on Instagram: Model Evaluation Study. Journal of Medical Internet Research. 21. 10.2196/13803.

[18] Raj, P. L. Joseph, Ch. Aswani Kumar and Mukul Rawat. "Automatic retrieval of updated information related to COVID-19 from web portals." (2020).

[19] Harrington, Richard & Adhikari, Vyas & Rayner, Mike & Scarborough, Peter. (2019). Nutrient composition databases in the age of big data: foodDB, a comprehensive, real-time database infrastructure. BMJ Open. 9. e026652. 10.1136/bmjopen-2018-026652.

[20] Singh, Sanjay & Haque, Afzalul. (2015). Anti-Scraping Application Development. 10.1109/ICACCI.2015.7275720.

[21] Moniteur Belge. (1994, 27 Juillet). Loi relative au droit d'auteur et aux droits voisins. Récupéré sur https:// ejustice.just.fgov.be/: https://www.ejustice.just.fgov.be/cgi_loi/loi_a.pl?language=fr&caller=list&cn=1994063035&la=f&fromtab=loi&sql=dt%3D%27loi%27&tri=dd%20as%20rank&rech=1&numero=1/ (visite 10/01/2023)

[22] SPF Economie. (2018, Février 9). Plaquette du SPF Economie. Récupéré sur economie.fgov.be: https://economie.fgov.be/fr/publicaties/plaquette-du-spf-economie (visite 04/02/2023)

[23] European Commission. (2020, Février 12). Item 04 - Web scraping policy. Récupéré sur https://cros-legacy.ec.europa.eu/: https://cros-legacy.ec.europa.eu/content/item-04-web-scraping-policy_en (visite 02/04/2023)

[24] Gouvernement Belge Economie. (2020, Octobre 27). La protection des bases de données par la propriété intellectuelle. Récupéré sur https://economie.fgov.be/: https://economie.fgov.be/fr/themes/propriete-intellectuelle/droits-de-pi/droits-dauteur-et-droits/droit-des-bases-de-donnees/la-protection-des-bases-de/ (visite 04/02/2023)

[25] Arrêt de la Cour (deuxième chambre) du 15 janvier 2015. Ryanair Ltd contre PR Aviation BV. Demande de décision préjudicielle, introduite par le Hoge Raad der Nederlanden. Renvoi préjudiciel – Directive 96/9/CE – Protection juridique des bases de données – Base de données qui n'est protégée ni par le droit d'auteur ni par le droit sui generis – Limitation contractuelle des droits des utilisateurs de la base de données. Affaire C-30/14.

[26] Règlement (UE) 2016/792 du Parlement européen et du Conseil du 11 mai 2016 relatif aux indices des prix à la consommation harmonisés et à l'indice des prix des logements, et abrogeant le règlement (CE) n° 2494/95 du Conseil (Texte présentant de l'intérêt pour l'EEE).

[27] WALLEX. (1998, Février 12). Décret modifiant le Code wallon de l'Aménagement du Territoire, de l'Urbanisme et du Patrimoine. Récupéré sur https://wallex.wallonie.be/: https://wallex.wallonie.be/de/contents/acts/7/7466/1.html/ (visite 08/02/2023)

[28] Wallonie. (2023, Février 1). Interdiction des chaudières au mazout : quelles sont les mesures prévues ? Récupéré sur https://www.wallonie.be/: https://www.wallonie.be/fr/actualites/interdiction-des-chaudieres-au-mazout-quelles-sont-les-mesures-prevues#:~:text=Dans%20les%20b%C3%A2timents%20neufs%2C%20l,au%201er%20janvier%202026 (visite 08/02/2023)

[29] Statbel. (2023, Février 5). Statistique des ventes de bâtiments. Récupéré sur https://bestat.statbel.fgov.be/: https://bestat.statbel.fgov.be/bestat/crosstable.xhtml?datasource=2a7528e9-8b3f-4ee5-ae67-8ebc33664d1f (visite 05/02/2023)

[30] Archives de l'Etat en Belgique. Archives notariales. Récupéré sur https://www.arch.be/: https://www.arch.be/index.php?l=fr&m=l-institution&r=que-conservons-nous&sr=modalites-de-consultation-specifiques&p=archives-notariales/ (visite 01/02/2023)

[31] Moniteur Belge. (1998, Novembre 14). Loi transposant en droit belge la directive européenne du 11 mars 1996 concernant la protection juridique des bases de données. Récupéré sur https://etaamb.openjustice.be/: https://etaamb.openjustice.be/fr/loi-du-31-aout-1998_n1998009789.html/ (visite 01/02/2023)

12. Annexes

12.1. Exemple complet d'un enregistrement JSON

Cet exemple est long de plus de mille lignes et s'étale donc sur plusieurs pages.


```
{
    "_id": {
        "$oid": "63b7777fb2fe554d78a62fe1"
    },
    "url": "https://www.immoweb.be/fr/annonce/maison/a-
vendre/beuzet/5030/10249342",
    "source": "immoweb",
    "creationDate": {
        "$date": "2023-01-06T01:21:03.300Z"
    },
    "modificationDate": {
        "$date": "2023-02-16T01:21:41.518Z"
    },
    "views": [
        {
            "count": 4870,
            "date": {
                "$date": "2023-01-06T01:21:03.335Z"
            }
        },
        {
            "count": 5008,
            "date": {
                "$date": "2023-01-07T01:21:49.928Z"
            }
        },
        {
            "count": 5095,
            "date": {
                "$date": "2023-01-08T01:21:21.825Z"
            }
        },
        {
            "count": 5196,
            "date": {
                "$date": "2023-01-09T01:21:54.399Z"
            }
        },
        {
            "count": 5356,
            "date": {
                "$date": "2023-01-10T01:21:42.556Z"
```

```
            }
        },
        {
            "count": 5472,
            "date": {
                "$date": "2023-01-11T01:21:31.220Z"
            }
        },
        {
            "count": 5551,
            "date": {
                "$date": "2023-01-12T01:21:14.981Z"
            }
        },
        {
            "count": 5644,
            "date": {
                "$date": "2023-01-13T01:21:50.496Z"
            }
        },
        {
            "count": 5790,
            "date": {
                "$date": "2023-01-14T01:22:03.947Z"
            }
        },
        {
            "count": 5864,
            "date": {
                "$date": "2023-01-15T01:22:27.668Z"
            }
        },
        {
            "count": 5953,
            "date": {
                "$date": "2023-01-16T01:20:59.002Z"
            }
        },
        {
            "count": 6086,
            "date": {
                "$date": "2023-01-17T01:21:31.561Z"
            }
        },
        {
            "count": 6191,
            "date": {
                "$date": "2023-01-18T01:21:50.499Z"
            }
```

```
        },
        {
            "count": 6358,
            "date": {
                "$date": "2023-01-20T01:22:05.336Z"
            }
        },
        {
            "count": 6435,
            "date": {
                "$date": "2023-01-21T01:21:31.239Z"
            }
        },
        {
            "count": 6503,
            "date": {
                "$date": "2023-01-22T01:22:51.177Z"
            }
        },
        {
            "count": 6573,
            "date": {
                "$date": "2023-01-23T01:21:57.153Z"
            }
        },
        {
            "count": 6691,
            "date": {
                "$date": "2023-01-24T01:22:35.885Z"
            }
        },
        {
            "count": 6794,
            "date": {
                "$date": "2023-01-25T01:21:51.482Z"
            }
        },
        {
            "count": 6909,
            "date": {
                "$date": "2023-01-26T01:21:49.853Z"
            }
        },
        {
            "count": 7002,
            "date": {
                "$date": "2023-01-27T01:21:29.229Z"
            }
        },
```

```
        {
            "count": 7106,
            "date": {
                "$date": "2023-01-28T01:21:22.909Z"
            }
        },
        {
            "count": 7194,
            "date": {
                "$date": "2023-01-29T01:21:13.159Z"
            }
        },
        {
            "count": 7303,
            "date": {
                "$date": "2023-01-30T01:21:23.076Z"
            }
        },
        {
            "count": 7423,
            "date": {
                "$date": "2023-01-31T01:26:21.066Z"
            }
        },
        {
            "count": 7512,
            "date": {
                "$date": "2023-02-01T01:23:24.147Z"
            }
        },
        {
            "count": 7624,
            "date": {
                "$date": "2023-02-02T01:22:28.218Z"
            }
        },
        {
            "count": 7727,
            "date": {
                "$date": "2023-02-03T01:22:53.795Z"
            }
        },
        {
            "count": 7808,
            "date": {
                "$date": "2023-02-04T01:22:17.415Z"
            }
        },
        {
```

```
            "count": 7867,
            "date": {
                "$date": "2023-02-05T01:23:57.263Z"
            }
        },
        {
            "count": 7977,
            "date": {
                "$date": "2023-02-06T01:21:19.902Z"
            }
        },
        {
            "count": 8080,
            "date": {
                "$date": "2023-02-07T01:23:17.622Z"
            }
        },
        {
            "count": 8173,
            "date": {
                "$date": "2023-02-08T01:23:39.981Z"
            }
        },
        {
            "count": 8244,
            "date": {
                "$date": "2023-02-09T01:21:47.945Z"
            }
        },
        {
            "count": 8849,
            "date": {
                "$date": "2023-02-16T01:21:41.518Z"
            }
        }
    ],
    "bookmarks": [
        {
            "count": 73,
            "date": {
                "$date": "2023-01-06T01:21:03.335Z"
            }
        },
        {
            "count": 78,
            "date": {
                "$date": "2023-01-07T01:21:49.928Z"
            }
        },
```

```
        {
            "count": 80,
            "date": {
                "$date": "2023-01-08T01:21:21.826Z"
            }
        },
        {
            "count": 82,
            "date": {
                "$date": "2023-01-09T01:21:54.399Z"
            }
        },
        {
            "count": 84,
            "date": {
                "$date": "2023-01-10T01:21:42.556Z"
            }
        },
        {
            "count": 86,
            "date": {
                "$date": "2023-01-11T01:21:31.220Z"
            }
        },
        {
            "count": 87,
            "date": {
                "$date": "2023-01-12T01:21:14.981Z"
            }
        },
        {
            "count": 89,
            "date": {
                "$date": "2023-01-13T01:21:50.496Z"
            }
        },
        {
            "count": 93,
            "date": {
                "$date": "2023-01-14T01:22:03.947Z"
            }
        },
        {
            "count": 94,
            "date": {
                "$date": "2023-01-15T01:22:27.668Z"
            }
        },
        {
```

```
            "count": 96,
            "date": {
                "$date": "2023-01-16T01:20:59.002Z"
            }
        },
        {
            "count": 97,
            "date": {
                "$date": "2023-01-17T01:21:31.561Z"
            }
        },
        {
            "count": 97,
            "date": {
                "$date": "2023-01-18T01:21:50.499Z"
            }
        },
        {
            "count": 101,
            "date": {
                "$date": "2023-01-20T01:22:05.336Z"
            }
        },
        {
            "count": 101,
            "date": {
                "$date": "2023-01-21T01:21:31.239Z"
            }
        },
        {
            "count": 103,
            "date": {
                "$date": "2023-01-22T01:22:51.177Z"
            }
        },
        {
            "count": 104,
            "date": {
                "$date": "2023-01-23T01:21:57.153Z"
            }
        },
        {
            "count": 106,
            "date": {
                "$date": "2023-01-24T01:22:35.885Z"
            }
        },
        {
            "count": 106,
```

```
            "date": {
                "$date": "2023-01-25T01:21:51.482Z"
            }
        },
        {
            "count": 108,
            "date": {
                "$date": "2023-01-26T01:21:49.853Z"
            }
        },
        {
            "count": 110,
            "date": {
                "$date": "2023-01-27T01:21:29.229Z"
            }
        },
        {
            "count": 111,
            "date": {
                "$date": "2023-01-28T01:21:22.909Z"
            }
        },
        {
            "count": 116,
            "date": {
                "$date": "2023-01-29T01:21:13.159Z"
            }
        },
        {
            "count": 120,
            "date": {
                "$date": "2023-01-30T01:21:23.076Z"
            }
        },
        {
            "count": 121,
            "date": {
                "$date": "2023-01-31T01:26:21.066Z"
            }
        },
        {
            "count": 121,
            "date": {
                "$date": "2023-02-01T01:23:24.147Z"
            }
        },
        {
            "count": 123,
            "date": {
```

```
            "$date": "2023-02-02T01:22:28.218Z"
        }
    },
    {
        "count": 126,
        "date": {
            "$date": "2023-02-03T01:22:53.795Z"
        }
    },
    {
        "count": 127,
        "date": {
            "$date": "2023-02-04T01:22:17.415Z"
        }
    },
    {
        "count": 127,
        "date": {
            "$date": "2023-02-05T01:23:57.263Z"
        }
    },
    {
        "count": 129,
        "date": {
            "$date": "2023-02-06T01:21:19.902Z"
        }
    },
    {
        "count": 129,
        "date": {
            "$date": "2023-02-07T01:23:17.622Z"
        }
    },
    {
        "count": 129,
        "date": {
            "$date": "2023-02-08T01:23:39.982Z"
        }
    },
    {
        "count": 131,
        "date": {
            "$date": "2023-02-09T01:21:47.945Z"
        }
    },
    {
        "count": 142,
        "date": {
            "$date": "2023-02-16T01:21:41.518Z"
```

```
            }
        }
    ],
    "general": {
        "atticExists": "",
        "basementExists": "true",
        "price": 315000,
        "reference": "10249342",
        "subtype": "house",
        "transactionType": "for sale",
        "type": "house",
        "visualisationOption": "xl",
        "zip": "5030",
        "outdoor": {
            "garden": {
                "surface": "450"
            },
            "terrace": {
                "exists": "true"
            }
        },
        "parking": {
            "parkingSpaceCount": {
                "indoor": "2",
                "outdoor": "6"
            }
        },
        "specificities": {
            "SME": {
                "office": {
                    "exists": "true"
                }
            }
        },
        "bedroom": {
            "bedroomCount": "4"
        },
        "building": {
            "condition": "just renovated",
            "constructionYear": "1909"
        },
        "certificates": {
            "primaryEnergyConsumptionLevel": "463"
        },
        "energy": {
            "heatingType": "fueloil"
        },
        "kitchen": {
            "type": "installed"
```

```
        },
        "land": {
            "surface": "1172"
        },
        "wellnessEquipment": {
            "hasSwimmingPool": ""
        }
    },
    "publisher": {
        "family": "agency",
        "id": "3692030",
        "name": "trevi tessier",
        "groupInfo": {
            "id": "",
            "name": ""
        },
        "networkInfo": {
            "id": "",
            "name": ""
        }
    },
    "property": {
        "bedrooms": [
            {
                "surface": 16
            },
            {
                "surface": 11
            },
            {
                "surface": 11
            },
            {
                "surface": 10
            }
        ],
        "roomCount": null,
        "building": {
            "condition": "JUST_RENOVATED",
            "constructionYear": 1909,
            "annexCount": null,
            "facadeCount": 4,
            "floorCount": null,
            "streetFacadeWidth": 15
        },
        "constructionPermit": {
            "totalBuildableGroundFloorSurface": 0,
            "constructionType": null,
            "floodZoneIconUrl": null,
```

```
            "floodZoneType": null,
            "hasObligationToConstruct": null,
            "hasPlotDivisionAuthorization": null,
            "hasPossiblePriorityPurchaseRight": null,
            "isBreachingUrbanPlanningRegulation": null,
            "isObtained": null,
            "urbanPlanningInformation": null
        },
        "energy": {
            "hasCollectiveWaterHeater": null,
            "hasDoubleGlazing": true,
            "hasHeatPump": null,
            "hasPhotovoltaicPanels": null,
            "hasThermicPanels": null,
            "heatingType": "FUELOIL",
            "performance": null
        },
        "kitchen": {
            "type": "INSTALLED",
            "hasDishwasher": null,
            "hasFreezer": null,
            "hasFridge": null,
            "hasMicroWaveOven": null,
            "hasOven": null,
            "hasSteamOven": null,
            "hasWashingMachine": null,
            "surface": null
        },
        "land": {
            "hasGasWaterElectricityConnection": true,
            "hasPlotToRear": null,
            "isFacingStreet": null,
            "isFlat": null,
            "isWooded": null,
            "landWidth": 30,
            "latestUseDesignation": null,
            "rearLand": null,
            "sewerConnection": null,
            "surface": 1172
        },
        "location": {
            "pointsOfInterests": [
                {
                    "distance": 185,
                    "type": "SCHOOL"
                },
                {
                    "distance": 0,
                    "type": "SHOPS"
```

```
            },
            {
                "distance": 68,
                "type": "TRANSPORT"
            }
        ],
        "approximated": null,
        "box": null,
        "country": "Belgique",
        "district": "Namur",
        "floor": null,
        "hasSeaView": null,
        "latitude": 50.5331025,
        "locality": "BEUZET",
        "longitude": 4.7416218,
        "number": "397",
        "placeName": null,
        "postalCode": "5030",
        "propertyName": null,
        "province": "Namur",
        "region": "Wallonie",
        "regionCode": "WALLONIE",
        "street": "Chaussée de Namur",
        "type": "RESIDENTIAL"
    },
    "propertyCertificates": {
        "builtPlanStatus": "NO",
        "globalInsulationLevel": null,
        "oilTankCertificateStatus": null,
        "primaryEnergyConsumptionLevel": 0
    },
    "specificities": {
        "SME": {
            "office": {
                "exists": ""
            }
        },
        "accessDoorCount": null,
        "ceilingHeight": null,
        "coveredBaysCount": null,
        "goodwillPrice": null,
        "hasGoodwill": null,
        "hasOffice": true,
        "hasReceptionDesk": null,
        "hasWorkspace": null,
        "internalFixturesAndFittings": null,
        "loadingBayCount": null,
        "loadingBayWithLiftingDeviceCount": null,
        "maxAvailableHeight": null,
```

```
            "minAvailableHeight": null,
            "miscellaneousFixturesAndFittings": null,
            "officeSurface": 17,
            "sectionalDoorCount": null,
            "shopWindowWidth": null,
            "showroomSurface": null,
            "slidingDoorCount": null,
            "totalFloorSurface": null,
            "workspaceSurface": null
        },
        "attic": {
            "isAccessibleWithFixedStairs": null,
            "isIsolated": null,
            "surface": null
        },
        "description": "Située le long de la chaussée de Namur à Beuzet, Trevi
Tessier vous propose à la vente une maison bourgeoise construite en 1909 et
rafraichie en 2020. La maison développe une superficie habitable de 164m² sur un
terrain de 11 ares 72 hors grenier et garage. Elle se compose comme suit: Au rez-
de-chaussée, un hall d'entrée, un séjour, une cuisine avec arrière cuisine, une
salle-de-bains avec douche, un loggia pouvant convenir de bureau ou de salle-de-
jeux, une buanderie et une véranda donnant accès au jardin. Au 1er étage, un hall
de nuit dessert 4 chambres (10 - 11 - 12 - 16m²). Le 2ème étage dispose d'un
grenier qui peut être transformé en suite parentale. Un garage avec une fosse
vient compléter cet espace de vie. La maison est située enrecule de la voirie et
offre un bel espace de stationnement privatif ainsi qu'un jardin, une terrasse
ouverte, une terrasse couverte et un terrain de pétanque. La maison permet
l'établissement d'un commerce ou d'une profession libérale. RC: 713€ permettant
la réduction des droits d'enregistrements réduits à 6% pour la première tranche.
CONFORT: CC au mazout, châssis DV en bois, citerne d'eau de pluie de 5.000L,
électricité conforme. PEB: Classe F - 463 kWh/m².an - 20220831018519. PRIX: Faire
offre à partir de 315.000€ sous réserve d'acceptation des propriétaires. VIDEO de
présentation disponible sur notre chaîne YouTube Trevi Tessier. DISPONIBILITE: A
l'acte. VISITES: Sur RDV avec notre agence au 081/34.13.55 ou par email
info@trevitessier.",
        "basementSurface": null,
        "bathroomCount": 1,
        "bedroomCount": 4,
        "bedroomSurface": null,
        "diningRoomSurface": null,
        "fireplaceCount": null,
        "fireplaceExists": false,
        "gardenOrientation": "SOUTH",
        "gardenSurface": 450,
        "habitableUnitCount": null,
        "hasAirConditioning": null,
        "hasArmoredDoor": null,
        "hasAttic": null,
        "hasBalcony": null,
```

```
        "hasBarbecue": null,
        "hasBasement": true,
        "hasCaretakerOrConcierge": null,
        "hasDiningRoom": null,
        "hasDisabledAccess": null,
        "hasDoorPhone": null,
        "hasDressingRoom": null,
        "hasFitnessRoom": null,
        "hasGarden": true,
        "hasHammam": null,
        "hasInternet": null,
        "hasJacuzzi": null,
        "hasLaundryRoom": null,
        "hasLift": null,
        "hasLivingRoom": true,
        "hasSauna": null,
        "hasSecureAccessAlarm": null,
        "hasSwimmingPool": null,
        "hasTVCable": null,
        "hasTennisCourt": null,
        "hasTerrace": true,
        "hasVisiophone": null,
        "isFirstOccupation": null,
        "isHolidayProperty": null,
        "laundryRoomSurface": null,
        "livingRoomSurface": 29,
        "monthlyCosts": null,
        "name": null,
        "netHabitableSurface": 164,
        "parkingCountClosedBox": null,
        "parkingCountIndoor": 2,
        "parkingCountOutdoor": 6,
        "showerRoomCount": 0,
        "subtype": "HOUSE",
        "terraceOrientation": null,
        "terraceSurface": 25,
        "title": "Maison bourgeoise avec garage, jardin et droits réduits",
        "toiletCount": 2,
        "type": "HOUSE"
    },
    "publication": {
        "creationDate": "2022-11-25T14:08:56.000+0000",
        "expirationDate": "2023-01-31T22:59:59.000+0000",
        "lastModificationDate": "2023-01-05T23:15:54.443+0000",
        "publisherId": null,
        "visualisationOption": "XL"
    },
    "statistics": {
        "bookmarkCount": 73,
```

```
        "viewCount": 4870,
        "alertPrice": null,
        "isAlertEmailSet": null,
        "rating": null
    },
    "customers": [
        {
            "family": "",
            "id": 3692030,
            "name": "Trevi Tessier",
            "contactHoursLandline": "anytime",
            "contactHoursMobile": "anytime",
            "email": "tessier@omniwebsites.be",
            "ipiNo": "508.080",
            "isOwner": true,
            "type": "AGENCY",
            "website": "http://www.trevitessier.be",
            "groupInfo": {
                "id": "",
                "name": ""
            },
            "networkInfo": {
                "id": "",
                "name": ""
            }
        }
    ],
    "transaction": {
        "certificates": {
            "primaryEnergyConsumptionLevel": null,
            "carbonEmission": 116,
            "epcDescription": null,
            "epcReference": "20220831018519",
            "epcScore": "F",
            "epcUrl": "https://static.immoweb.be/pics/peb/peb_f.png",
            "hasElectricalInstallationComplianceCertificate": null,
            "primaryEnergyConsumptionPerSqm": 463,
            "primaryEnergyConsumptionYearly": 90601
        },
        "investor": {
            "currentMonthlyRentalIncome": null,
            "currentReturnOnInvestment": null,
            "expectedMonthlyRentalIncome": null,
            "expectedMonthlyRentalIncomeDescription": null,
            "expectedReturnOnInvestment": null,
            "habitableUnitCount": null,
            "isInvestmentProperty": false,
            "occupancyRate": null
        },
```

```
        "rental": {
            "areBigPetsAllowed": null,
            "areSmallPetsAllowed": null,
            "isFurnished": null,
            "monthlyRentalCosts": null,
            "monthlyRentalPrice": null,
            "yearlyRentalPrice": null,
            "yearlyRentalPricePerSqm": null
        },
        "sale": {
            "lifeAnnuity": {
                "annuitantAges": []
            },
            "cadastralIncome": 713,
            "hasStartingPrice": false,
            "homeToBuild": null,
            "isFurnished": false,
            "isSubjectToVat": false,
            "oldPrice": null,
            "price": 315000,
            "pricePerSqm": null,
            "publicSale": null
        },
        "availabilityDate": "2023-03-01T00:00:00.000+0000",
        "availabilityPeriodType": "UPON_EXCHANGE_OF_DEEDS",
        "subtype": "BUY_REGULAR",
        "type": "FOR_SALE"
    },
    "updates": [
        {
            "url": "https://www.immoweb.be/fr/annonce/maison/a-
vendre/beuzet/5030/10249342",
            "source": "immoweb",
            "creationDate": {
                "$date": "2023-02-16T01:21:41.489Z"
            },
            "modificationDate": null,
            "views": [],
            "bookmarks": [],
            "general": {
                "atticExists": "",
                "basementExists": "true",
                "price": 315000,
                "reference": "10249342",
                "subtype": "house",
                "transactionType": "for sale",
                "type": "house",
                "visualisationOption": "xl",
                "zip": "5030",
```

```
            "outdoor": {
                "garden": {
                    "surface": "450"
                },
                "terrace": {
                    "exists": "true"
                }
            },
            "parking": {
                "parkingSpaceCount": {
                    "indoor": "2",
                    "outdoor": "6"
                }
            },
            "specificities": {
                "SME": {
                    "office": {
                        "exists": "true"
                    }
                }
            },
            "bedroom": {
                "bedroomCount": "4"
            },
            "building": {
                "condition": "just renovated",
                "constructionYear": "1909"
            },
            "certificates": {
                "primaryEnergyConsumptionLevel": "463"
            },
            "energy": {
                "heatingType": "fueloil"
            },
            "kitchen": {
                "type": "installed"
            },
            "land": {
                "surface": "1172"
            },
            "wellnessEquipment": {
                "hasSwimmingPool": ""
            }
        },
        "publisher": {
            "family": "agency",
            "id": "3692030",
            "name": "trevi tessier",
            "groupInfo": {
```

```
                "id": "",
                "name": ""
            },
            "networkInfo": {
                "id": "",
                "name": ""
            }
        },
        "property": {
            "bedrooms": [
                {
                    "surface": 16
                },
                {
                    "surface": 11
                },
                {
                    "surface": 11
                },
                {
                    "surface": 10
                }
            ],
            "roomCount": null,
            "building": {
                "condition": "JUST_RENOVATED",
                "constructionYear": 1909,
                "annexCount": null,
                "facadeCount": 4,
                "floorCount": null,
                "streetFacadeWidth": 15
            },
            "constructionPermit": {
                "totalBuildableGroundFloorSurface": 0,
                "constructionType": null,
                "floodZoneIconUrl": null,
                "floodZoneType": null,
                "hasObligationToConstruct": null,
                "hasPlotDivisionAuthorization": null,
                "hasPossiblePriorityPurchaseRight": null,
                "isBreachingUrbanPlanningRegulation": null,
                "isObtained": null,
                "urbanPlanningInformation": null
            },
            "energy": {
                "hasCollectiveWaterHeater": null,
                "hasDoubleGlazing": true,
                "hasHeatPump": null,
                "hasPhotovoltaicPanels": null,
```

```
            "hasThermicPanels": null,
            "heatingType": "FUELOIL",
            "performance": null
        },
        "kitchen": {
            "type": "INSTALLED",
            "hasDishwasher": null,
            "hasFreezer": null,
            "hasFridge": null,
            "hasMicroWaveOven": null,
            "hasOven": null,
            "hasSteamOven": null,
            "hasWashingMachine": null,
            "surface": null
        },
        "land": {
            "hasGasWaterElectricityConnection": true,
            "hasPlotToRear": null,
            "isFacingStreet": null,
            "isFlat": null,
            "isWooded": null,
            "landWidth": 30,
            "latestUseDesignation": null,
            "rearLand": null,
            "sewerConnection": null,
            "surface": 1172
        },
        "location": {
            "pointsOfInterests": [
                {
                    "distance": 185,
                    "type": "SCHOOL"
                },
                {
                    "distance": 0,
                    "type": "SHOPS"
                },
                {
                    "distance": 68,
                    "type": "TRANSPORT"
                }
            ],
            "approximated": null,
            "box": null,
            "country": "Belgique",
            "district": "Namur",
            "floor": null,
            "hasSeaView": null,
            "latitude": 50.5331025,
```

```
            "locality": "BEUZET",
            "longitude": 4.7416218,
            "number": "397",
            "placeName": null,
            "postalCode": "5030",
            "propertyName": null,
            "province": "Namur",
            "region": "Wallonie",
            "regionCode": "WALLONIE",
            "street": "Chaussée de Namur",
            "type": "RESIDENTIAL"
        },
        "propertyCertificates": {
            "builtPlanStatus": "NO",
            "globalInsulationLevel": null,
            "oilTankCertificateStatus": null,
            "primaryEnergyConsumptionLevel": 0
        },
        "specificities": {
            "SME": {
                "office": {
                    "exists": ""
                }
            },
            "accessDoorCount": null,
            "ceilingHeight": null,
            "coveredBaysCount": null,
            "goodwillPrice": null,
            "hasGoodwill": null,
            "hasOffice": true,
            "hasReceptionDesk": null,
            "hasWorkspace": null,
            "internalFixturesAndFittings": null,
            "loadingBayCount": null,
            "loadingBayWithLiftingDeviceCount": null,
            "maxAvailableHeight": null,
            "minAvailableHeight": null,
            "miscellaneousFixturesAndFittings": null,
            "officeSurface": 17,
            "sectionalDoorCount": null,
            "shopWindowWidth": null,
            "showroomSurface": null,
            "slidingDoorCount": null,
            "totalFloorSurface": null,
            "workspaceSurface": null
        },
        "attic": {
            "isAccessibleWithFixedStairs": null,
            "isIsolated": null,
```

```
                    "surface": null
                },
                "description": "Située le long de la chaussée de Namur à Beuzet, Trevi Tessier vous propose à la vente une maison bourgeoise construite en 1909 et rafraichie en 2020. La maison développe une superficie habitable de 164m² sur un terrain de 11 ares 72 hors grenier et garage. Elle se compose comme suit: Au rez-de-chaussée, un hall d'entrée, un séjour, une cuisine avec arrière cuisine, une salle-de-bains avec douche, un loggia pouvant convenir de bureau ou de salle-de-jeux, une buanderie et une véranda donnant accès au jardin. Au 1er étage, un hall de nuit dessert 4 chambres (10 - 11 - 12 - 16m²). Le 2ème étage dispose d'un grenier qui peut être transformé en suite parentale. Un garage avec une fosse vient compléter cet espace de vie. La maison est située enrecule de la voirie et offre un bel espace de stationnement privatif ainsi qu'un jardin, une terrasse ouverte, une terrasse couverte et un terrain de pétanque. La maison permet l'établissement d'un commerce ou d'une profession libérale. RC: 713€ permettant la réduction des droits d'enregistrements réduits à 6% pour la première tranche. CONFORT: CC au mazout, châssis DV en bois, citerne d'eau de pluie de 5.000L, électricité conforme. PEB: Classe F - 463 kWh/m².an - 20220831018519. PRIX: Faire offre à partir de 315.000€ sous réserve d'acceptation des propriétaires. VIDEO de présentation disponible sur notre chaîne YouTube Trevi Tessier. DISPONIBILITE: A l'acte. VISITES: Sur RDV avec notre agence au 081/34.13.55 ou par email info@trevitessier.",
                "basementSurface": null,
                "bathroomCount": 1,
                "bedroomCount": 4,
                "bedroomSurface": null,
                "diningRoomSurface": null,
                "fireplaceCount": null,
                "fireplaceExists": false,
                "gardenOrientation": "SOUTH",
                "gardenSurface": 450,
                "habitableUnitCount": null,
                "hasAirConditioning": null,
                "hasArmoredDoor": null,
                "hasAttic": null,
                "hasBalcony": null,
                "hasBarbecue": null,
                "hasBasement": true,
                "hasCaretakerOrConcierge": null,
                "hasDiningRoom": null,
                "hasDisabledAccess": null,
                "hasDoorPhone": null,
                "hasDressingRoom": null,
                "hasFitnessRoom": null,
                "hasGarden": true,
                "hasHammam": null,
                "hasInternet": null,
                "hasJacuzzi": null,
                "hasLaundryRoom": null,
```

```
                "hasLift": null,
                "hasLivingRoom": true,
                "hasSauna": null,
                "hasSecureAccessAlarm": null,
                "hasSwimmingPool": null,
                "hasTVCable": null,
                "hasTennisCourt": null,
                "hasTerrace": true,
                "hasVisiophone": null,
                "isFirstOccupation": null,
                "isHolidayProperty": null,
                "laundryRoomSurface": null,
                "livingRoomSurface": 29,
                "monthlyCosts": null,
                "name": null,
                "netHabitableSurface": 164,
                "parkingCountClosedBox": null,
                "parkingCountIndoor": 2,
                "parkingCountOutdoor": 6,
                "showerRoomCount": 0,
                "subtype": "HOUSE",
                "terraceOrientation": null,
                "terraceSurface": 25,
                "title": "Maison bourgeoise avec garage, jardin et droits
réduits",
                "toiletCount": 2,
                "type": "HOUSE"
            },
            "publication": {
                "creationDate": "2022-11-25T14:08:56.000+0000",
                "expirationDate": "2023-02-28T22:59:59.000+0000",
                "lastModificationDate": "2023-02-15T23:24:01.867+0000",
                "publisherId": null,
                "visualisationOption": "XL"
            },
            "statistics": {
                "bookmarkCount": 142,
                "viewCount": 8849,
                "alertPrice": null,
                "isAlertEmailSet": null,
                "rating": null
            },
            "customers": [
                {
                    "family": "",
                    "id": 3692030,
                    "name": "Trevi Tessier",
                    "contactHoursLandline": "anytime",
                    "contactHoursMobile": "anytime",
```

```
                "email": "tessier@omniwebsites.be",
                "ipiNo": "508.080",
                "isOwner": true,
                "type": "AGENCY",
                "website": "http://www.trevitessier.be",
                "groupInfo": {
                    "id": "",
                    "name": ""
                },
                "networkInfo": {
                    "id": "",
                    "name": ""
                }
            }
        ],
        "transaction": {
            "certificates": {
                "primaryEnergyConsumptionLevel": null,
                "carbonEmission": 116,
                "epcDescription": null,
                "epcReference": "20220831018519",
                "epcScore": "F",
                "epcUrl": "https://static.immoweb.be/pics/peb/peb_f.png",
                "hasElectricalInstallationComplianceCertificate": null,
                "primaryEnergyConsumptionPerSqm": 463,
                "primaryEnergyConsumptionYearly": 90601
            },
            "investor": {
                "currentMonthlyRentalIncome": null,
                "currentReturnOnInvestment": null,
                "expectedMonthlyRentalIncome": null,
                "expectedMonthlyRentalIncomeDescription": null,
                "expectedReturnOnInvestment": null,
                "habitableUnitCount": null,
                "isInvestmentProperty": false,
                "occupancyRate": null
            },
            "rental": {
                "areBigPetsAllowed": null,
                "areSmallPetsAllowed": null,
                "isFurnished": null,
                "monthlyRentalCosts": null,
                "monthlyRentalPrice": null,
                "yearlyRentalPrice": null,
                "yearlyRentalPricePerSqm": null
            },
            "sale": {
                "lifeAnnuity": {
                    "annuitantAges": []
```

```
                    },
                    "cadastralIncome": 713,
                    "hasStartingPrice": false,
                    "homeToBuild": null,
                    "isFurnished": false,
                    "isSubjectToVat": false,
                    "oldPrice": null,
                    "price": 315000,
                    "pricePerSqm": null,
                    "publicSale": null
                },
                "availabilityDate": "2023-03-01T00:00:00.000+0000",
                "availabilityPeriodType": "UPON_EXCHANGE_OF_DEEDS",
                "subtype": "BUY_REGULAR",
                "type": "FOR_SALE"
            }
        }
    ]
}
```


12.2. Détails des clés/valeurs

La table « classifieds » contient les petites-annonces. Certaines propriétés existent, mais ne sont pas / ne sont plus récoltées. Celles-ci sont suivies de la mention « n.a.». Elles ne sont dès lors pas utilisées par notre analyse.

classifieds : une petite annonce
|_ _ _ _ _ _id : est l'identifiant de l'enregistrement
|_ _ _ _ _ source : permet d'identifier la provenance de la donnée (le site scrappé)
|_ _ _ _ _ creationDate : date de l'enregistrement de la donnée
|_ _ _ _ _ modificaitonDate : date de la modification de la donnée
|_ _ _ _ _ views : tableau dans lequel on enregistre le nombre de vues de la petite-annonce lors de chaque récolte
|_ _ _ _ _ bookmarks : tableau dans lequel on enregistre le nombre de fois que la petite-annonce est sauvegardée
|_ _ _ _ _ **general :** contient un résumé d'une petite annonce
|_ _ _ _ _ _ _ _ _ _ atticExists : indique la présence ou non d'un grenier
|_ _ _ _ _ _ _ _ _ _ basementExists : indique la présence ou non d'une cave
|_ _ _ _ _ _ _ _ _ _ price : le prix demandé
|_ _ _ _ _ _ _ _ _ _ reference : la référence donnée par le site de petites-annonces
|_ _ _ _ _ _ _ _ _ _ subtype : le sous-type auquel appartient le bien annoncé
|_ _ _ _ _ _ _ _ _ _ transactionType : le type de transaction (louer ou acheter par exemple)
|_ _ _ _ _ _ _ _ _ _ type : le type du bien auquel appartient le bien annoncé
|_ _ _ _ _ _ _ _ _ _ zip : le code postal

|_ _ _ _ _ _ _ _ _ **outdoor**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ **garden**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ surface : la surface du jardin exprimée en (m²)
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ **terrace**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ exists : indique la présence ou non d'une terrasse
|_ _ _ _ _ _ _ _ _ **parking**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ **parkingSpaceCount**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ indoor : le nombre de parking intérieur
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ outdoor : le nombre de parking extérieur
|_ _ _ _ _ _ _ _ _ **specificities**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ **SME**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ **office**
|_ exists : indique la présence ou non de bureau pour une entreprise de petite à moyenne taille
|_ _ _ _ _ _ _ _ _ **bedroom**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ bedroomCount : le nombre de chambres
|_ _ _ _ _ _ _ _ _ **building**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ condition : le type de condition (e.g. « comme neuf »)
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ constructionYear : l'année de construction du bien
|_ _ _ _ _ _ _ _ _ **certificates**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ primaryEnergyConsumptionLevel : le niveau de consommation d'énergie primaire
|_ _ _ _ _ _ _ _ _ **energy**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ heatingType : le type de chauffage
|_ _ _ _ _ _ _ _ _ **kitchen**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ type : le type de cuisine
|_ _ _ _ _ _ _ _ _ **land**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ surface : la taille du terrain exprimée en mètre(s) carré(s)
|_ _ _ _ _ _ _ _ _ **wellnessEquipment**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ hasSwimmingPool : indique la présence ou non d'une piscine
|_ _ _ _ _ **publisher :** les informations sur le publieurs de la petite annonce
|_ _ _ _ _ _ _ _ _ family : le type de publieur (e.g. agency)
|_ _ _ _ _ _ _ _ _ id : l'identifiant du publieur
|_ _ _ _ _ _ _ _ _ name : le nom du publieur
|_ _ _ _ _ _ _ _ _ **groupInfo**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ id : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ name : n.a.
|_ _ _ _ _ _ _ _ _ **networkInfo**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ id : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ name : n.a.
|_ _ _ _ _ **property :** les informations sur le bien

|_ _ _ _ _ _ _ _ _ bedrooms : tableau qui contient les surfaces des différentes chambres
|_ _ _ _ _ _ _ _ _ roomCount : le nombre de chambres
|_ _ _ _ _ _ _ _ _ **building**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ condition : le type d'état du bâtiment (e.g. comme neuf)
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ constructionYear : l'année de construction
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ annexCount : le nombre d'annexes
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ facadeCount : le nombre de façades
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ floorCount : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ streetFacadeWidth : la largeur de la façade côté rue
|_ _ _ _ _ _ _ _ _ **constructionPermit**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ totalBuildableGroundFloorSurface : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ constructionType : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ floodZoneIconUrl : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ floodZoneType : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasObligationToConstruct : indique si il y a obligation de construire
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasPlotDivisionAuthorization : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasPossiblePriorityPurchaseRight : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ isBreachingUrbanPlanningRegulation : T n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ isObtained : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ urbanPlanningInformation : n.a.
|_ _ _ _ _ _ _ _ _ **energy**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasCollectiveWaterHeater : indique la présence ou non d'un chauffe-eau collectif
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasDoubleGlazing : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasHeatPump : indique la présence ou non d'une pompe à chaleur
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasPhotovoltaicPanels : indique la présence ou non de panneaux photovoltaïques
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasThermicPanels : indique la présence ou non de panneaux thermiques
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ heatingType : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ performance : n.a.
|_ _ _ _ _ _ _ _ _ **kitchen**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ type : le type de cuisine
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasDishwasher : indique la présence ou non d'un lave-vaisselle
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasFreezer : indique la présence ou non d'un réfrigérateur
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasFridge : indique la présence ou non d'un frigidaire
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasMicroWaveOven : indique la présence ou non d'un four à micro-ondes
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasOven : indique la présence ou non d'un four
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasSteamOven : indique la présence ou non d'un four à vapeur
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasWashingMachine : indique la présence ou non d'une machine à

laver
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ surface : la surface de la cuisine
|_ _ _ _ _ _ _ _ _ **land**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ hasGasWaterElectricityConnection : indique si présence d'eau, gaz et électricité
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ hasPlotRear : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ isFacingStreet : indique si le terrain fait face à la rue ou non
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ isFlat : indique si le terrain est plat ou non
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ isWooded : indique si le terrain est boisé ou non
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ landWidth : la largeur du terrain à rue exprimée en mètre
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ latestUseDesignation : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ rearLand : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ sewerConnection : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ surface : la surface du terrain exprimée en mètre(s) carré(s)
|_ _ _ _ _ _ _ _ _ **location**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ pointsOfInteresets : une liste d'objets représentants les points d'intérêt autour du bien (distance et type)
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ approximated : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ box : le numéro de la boite postale
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ country : pays dans lequel se trouve le bien
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ district : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ floor : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ hasSeaView : indique la présence ou non d'une vue sur mer
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ latitude : la latitude du bien
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ locality : la localité du bien
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ longitude : la longitude du bien
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ number : le numéro de maison
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ placeName : le lieu-dit
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ postalCode : le code postal du bien
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ propertyName : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ province : la province dans laquelle se trouve le bien
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ region : la région dans laquelle se trouve le bien
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ regionCode : le code de la région
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ street : le nom de la rue dans laquelle se trouve le bien
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ type : n.a.
|_ _ _ _ _ _ _ _ _ **propertyCertificates**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ builtPlanStatus : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ globalINsulationLevel : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ oilTankCertificateStatus : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ primaryEnergyConsumptionLevel : le niveau de consommation d'énergie primaire exprimé en kWh/an

|_ _ _ _ _ _ _ _ _ **specificities**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ **SME**
|_ **office**
|_ exists : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ accessDoorCount : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ ceilingHeight : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ coveredBaysCount : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ goodwillPrice : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasGoodwill : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasOffice : indique la présence ou non de bureau
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasReceptionDesk : indique la présence ou nom d'un bureau de réception
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasWorkspace : indique la présence ou non d'un espace de travail
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ internalFixturesAndFittings : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ loadingBayCount : le nombre de baies de chargement
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _loadingBayWithLiftingDeviceCount : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ maxAvailableHeight : taille de la hauteur maximale
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ minAvailableHeight : taille de la hauteur minimale
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ miscellaneousFixturesAndFittings : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ officeSurface : la surface de bureau
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ sectionalDoorCount : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ shopWindowWidth : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ showroomSurface : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ slidingDoorCount : nombre de portes coulissantes
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ totalFloorSurface : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ workspaceSurface : n.a.
|_ _ _ _ _ _ _ _ _ **attic**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ isAccessibleWithFixedStairs : indique si le grenier est accessible depuis des escaliers fixes
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ isIsolated : indique si le grenier est isolé
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ surface : la surface du grenier en mètre(s) carré(s)
|_ _ _ _ _ _ _ _ _ description : la description du bien
|_ _ _ _ _ _ _ _ _ basementSurface : la surface de la cave en mètre(s) carré(s)
|_ _ _ _ _ _ _ _ _ bathroomCount : le nombre de salles de bain
|_ _ _ _ _ _ _ _ _ bedroomCount : le nombre de chambres
|_ _ _ _ _ _ _ _ _ bedroomSurface : la surface de la chambre
|_ _ _ _ _ _ _ _ _ diningRoomSurface : la surface de la salle à manger
|_ _ _ _ _ _ _ _ _ fireplaceCount : le nombre de cheminées
|_ _ _ _ _ _ _ _ _ fireplaceExists : indique si une cheminée existe
|_ _ _ _ _ _ _ _ _ gardenOrientation : l'orientation du jardin
|_ _ _ _ _ _ _ _ _ gardenSurface : la surface du jardin

|_ _ _ _ _ _ _ _ _ habitableUnitCount : le nombre d'unité habitable
|_ _ _ _ _ _ _ _ _ hasAirConditionning : indique si présence ou non d'air conditionné
|_ _ _ _ _ _ _ _ _ hasArmoredDoor : indique si présence ou non d'une porte blindée
|_ _ _ _ _ _ _ _ _ hasAttic : indique si présence ou non d'un grenier
|_ _ _ _ _ _ _ _ _ hasBalcony : indique si présence ou non d'un balcon
|_ _ _ _ _ _ _ _ _ hasBarbecue : indique si présence ou non d'un barbecue
|_ _ _ _ _ _ _ _ _ hasBasement : indique si présence ou non d'une cave
|_ _ _ _ _ _ _ _ _ hasCaretakerOrConcierge : indique si présence ou non d'un concierge
|_ _ _ _ _ _ _ _ _ hasDiningRoom : indique si présence ou non d'une salle à manger
|_ _ _ _ _ _ _ _ _ hasDisabledAccess : indique si présence ou non d'un accès pour personnes à mobilité réduite
|_ _ _ _ _ _ _ _ _ hasDoorPhone : indique si présence ou non d'un parlophone
|_ _ _ _ _ _ _ _ _ hasDressingRoom : indique si présence ou non d'un dressing
|_ _ _ _ _ _ _ _ _ hasFitnessRoom : indique si présence ou non d'une salle de fitness
|_ _ _ _ _ _ _ _ _ hasGarden : indique si présence ou non d'un jardin
|_ _ _ _ _ _ _ _ _ hasHammam : indique si présence ou non d'un hammam
|_ _ _ _ _ _ _ _ _ hasInternet : indique si présence ou non d'Internet
|_ _ _ _ _ _ _ _ _ hasJacuzzi : indique si présence ou non d'un jacuzzi
|_ _ _ _ _ _ _ _ _ hasLaundryRoom : indique si présence ou non d'une buanderie
|_ _ _ _ _ _ _ _ _ hasLifts : indique si présence ou non d'ascenseur
|_ _ _ _ _ _ _ _ _ hasLivingRoom : indique si présence ou non de living room
|_ _ _ _ _ _ _ _ _ hasSauna : indique si présence ou non de sauna
|_ _ _ _ _ _ _ _ _ hasSecureAccessAlarm : indique si présence ou non d'une alarme
|_ _ _ _ _ _ _ _ _ hasSwimmingPool : indique si présence ou non d'une piscine
|_ _ _ _ _ _ _ _ _ hasTVCable : indique si présence ou non du câble de télévision
|_ _ _ _ _ _ _ _ _ hasTennisCourt : indique si présence ou non d'un court de tennis
|_ _ _ _ _ _ _ _ _ hasTerrace : indique si présence ou non d'une terrasse
|_ _ _ _ _ _ _ _ _ hasVisiophone : indique si présence ou non de visiophone
|_ _ _ _ _ _ _ _ _ isFirstOccupation : indique si c'est la première occupation ou non
|_ _ _ _ _ _ _ _ _ isHolidayProperty : indique si c'est une résidence de vacances
|_ _ _ _ _ _ _ _ _ laundryRoomSurface : la surface de la buanderie
|_ _ _ _ _ _ _ _ _ livingRoomSurface : la surface du living-room
|_ _ _ _ _ _ _ _ _ monthlyCosts : les charges liées à la location
|_ _ _ _ _ _ _ _ _ name : n.a.
|_ _ _ _ _ _ _ _ _ netHabitableSurface : la surface habitable nette
|_ _ _ _ _ _ _ _ _ parkingCountClosedBox : le nombre de parkings fermés
|_ _ _ _ _ _ _ _ _ parkingCountIndoor : le nombre de parkings intérieurs
|_ _ _ _ _ _ _ _ _ showerRoomCount : le nombre de salles de douche
|_ _ _ _ _ _ _ _ _ subtype : le sous-type du bien
|_ _ _ _ _ _ _ _ _ terraceOrientation : l'orientation de la terrasse

|_ _ _ _ _ _ _ _ _ terraceSurface : la surface de la terrasse
|_ _ _ _ _ _ _ _ _ title : le titre de la petite annonce
|_ _ _ _ _ _ _ _ _ toiletCount : le nombre de toilettes
|_ _ _ _ _ _ _ _ _ type : le type de bien
|_ _ _ _ _ **publication :** les informations méta de la publication
|_ _ _ _ _ _ _ _ _ creationDate : date de création de la publication
|_ _ _ _ _ _ _ _ _ expirationDate : date d'expiration de la publication
|_ _ _ _ _ _ _ _ _ lastModificaitonDate : date de la dernière modification de la publication
|_ _ _ _ _ _ _ _ _ publisherId : identifiant du publieur
|_ _ _ _ _ _ _ _ _ visualisationOption : n.a.
|_ _ _ _ _ **statistics :** les informations méta sur la publication
|_ _ _ _ _ _ _ _ _ bookmarkCount : nombre de fois ajouté aux favoris
|_ _ _ _ _ _ _ _ _ viewCount : le nombre de vues
|_ _ _ _ _ _ _ _ _ alertPrice : n.a.
|_ _ _ _ _ _ _ _ _ isAlertEmailSet : n.a.
|_ _ _ _ _ _ _ _ _ rating : n.a.
|_ _ _ _ _ **customers :** liste d'objet Client (= une personne qui achète un espace pour sa petite-annonce (personne privée, agent immobilier, …))
|_ _ _ _ _ **transaction :** les informations liées à la transaction
|_ _ _ _ _ _ _ _ _ **certificates**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ primaryEnergyConsumptionLevel : le niveau de consommation d'énergie primaire
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ carbonEmission : les émissions de carbone
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ epcDescription : description du certificat de performance énergétique
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ epcReference : référence du certificat de performance énergétique
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ epcScore : score du certificat de performance énergétique
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ epcUrl : url du certificat de performance énergétique
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ hasElectricalInstallationComplianceCertificate : présence ou non d'un certificat d'installation électrique
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ primaryEnergyConsumptionPerSqm : la consommation d'énergie primaire par mètre carré
|_ _ _ _ _ _ _ _ _ **investor**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ currentMonthlyRentalIncome : la rente mensuelle actuelle
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ currentReturnOnInvestmen : le retour sur investissement actuel
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ expectedMonthlyRentalIncome : la rente mensuelle attendue
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ expectedMonthlyRentalIncomeDescription : description de la rente mensuelle attendue
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ expectedReturnOnInvestment : le retour sur investissement attendu
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ habitableUnitCount : nombre d'unités habitables
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ isInvestmentProperty : indique si c'est un immeuble de rapport
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ occupancyRate : le taux d'occupation

|_ _ _ _ _ _ _ _ _ **rental**
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ areBigPetsAllowed : indique si oui ou non les gros animaux sont admis
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ areSmallPetsAllowed : indique si oui ou non les petits animaux sont admis
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ isFurnished : indique si oui ou non le bien est meublé
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ monthlyRentalCosts : le coût mensuel des charges
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ monthlyRentalPrice : le coût mensuel de la location
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ yearlyRentalPrice : n.a.
|_ _ _ _ _ _ _ _ _ _ _ _ _ _ _yearlyRentalPricePerSqm : n.a.
|_ _ _ _ _ _ _ _ _ availabilityDate : date de la disponibilité
|_ _ _ _ _ _ _ _ _ awailabilityPeriodType : quand le bien est-il disponible
|_ _ _ _ _ _ _ _ _ sale : n.a.
|_ _ _ _ _ _ _ _ _ subtype : le sous-type du bien
|_ _ _ _ _ _ _ _ _ type : le type du bien
|_ _ _ _ _ **updates** contient un objet « classified » pour historisation lorsqu'une petite-annonce est modifiée

classifieds.general.subtype : « kot », « house », « apartment », « villa », «apartment block », « duplex », « mixed use building », « ground floor », « flat studio », « exceptional property », « country cottage », « town house », « mansion », « penthouse », « chalet », « bungalow », « farmhouse », « other property », « loft », « triplex », « castle », « service flat », «manor house »

classifieds.general.transactionType : « for rent », « for sale »

classifieds.general.type : « house », « apartment »

classifieds.general.building.condition : « as new », « good », « to renovate », « to be done up », « just renovated », « to restore »

classifieds.publisher.family : « agency », « private », « notary », « property_developer », « agency_paying_with_ogone », « real_estate_agency », « company », « company_paying_with_ogone », « detached_house_builders »

classifieds.property.building.condition : « as new », …

classifieds.property.kitchen.type : n.a.

classifieds.property.location.type : n.a.

classifieds.property.location.subtype : n.a.

classifieds.transaction.availabilityPeriodType : « IMMEDIATELY », …

12.2. Supplément

12.2.1. ADE de la sous-classe « appartement à louer »

12.2.1.1. Distribution

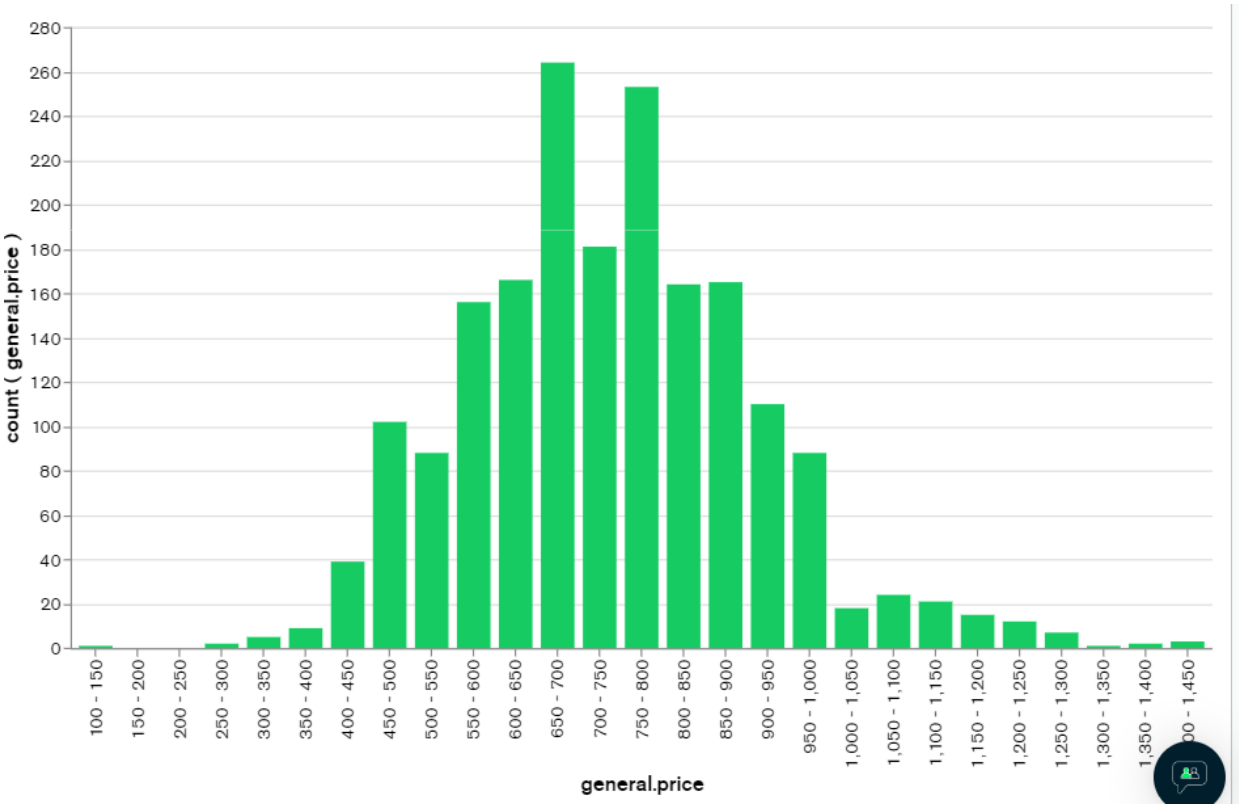
Nous allons analyser les variables de manière indépendante pour comprendre leurs distributions et ensuite leurs relations avec les autres variables.

Comprendre la distribution de la variable cible (le prix)

Histogramme

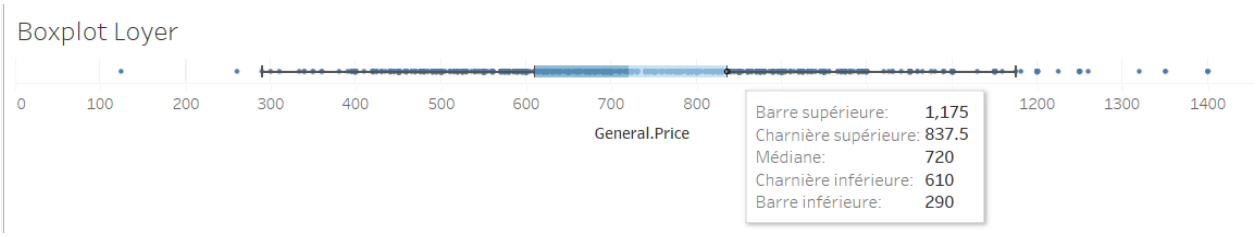
L'histogramme permet de montrer la distribution des valeurs d'une variable. Nous explorons ici la fréquence de chaque valeur de prix regroupée par tranche de cinquante euros (bin size 50). L'histogramme présente une distribution symétrique et donc normale, à priori.

Une distribution symétrique d'une variable signifie que la distribution est équilibrée de chaque côté de la moyenne, c'est-à-dire que la distribution est également répartie autour de sa moyenne, ce qui se traduit par une courbe en forme de cloche. Si votre histogramme montre une distribution symétrique, cela signifie que les prix sont répartis de manière égale autour de leur valeur moyenne. En d'autres termes, cela suggère que les appartements sont proposés à des prix raisonnables et que le marché est équilibré entre l'offre et la demande.



Les valeurs les plus fréquentes sont les tranches 650-700, 750-800 et 700-750.

Résumé statistique



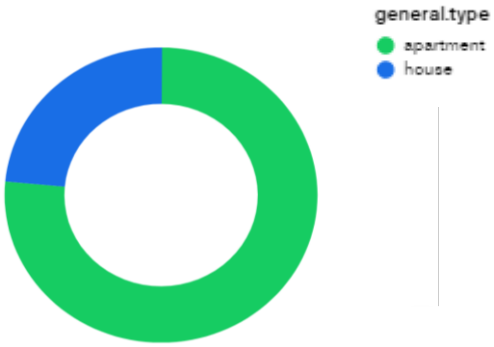
Le boxplot montre également une distribution symétrique et uniforme. Les écarts entre la médiane et les charnières inférieures et supérieures sont presque les mêmes. De plus, les écarts entre la médiane, la barre supérieure et la barre inférieure sont aussi presque les mêmes.

Les valeurs en dehors des limites déterminées par le boxplots seront analysées ultérieurement. Nous verrons si l'analyse par cluster peut identifier un groupe particulier d'appartements (grand, bien situé et/ou luxueux par exemple).

Exploration des variables catégoriques

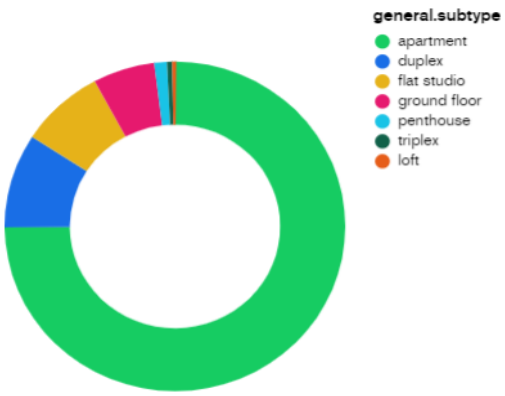
Le type de bien

Nous observons que 76.4% des enregistrements concernent des appartements à louer et 23.6% concernent des maisons à louer.



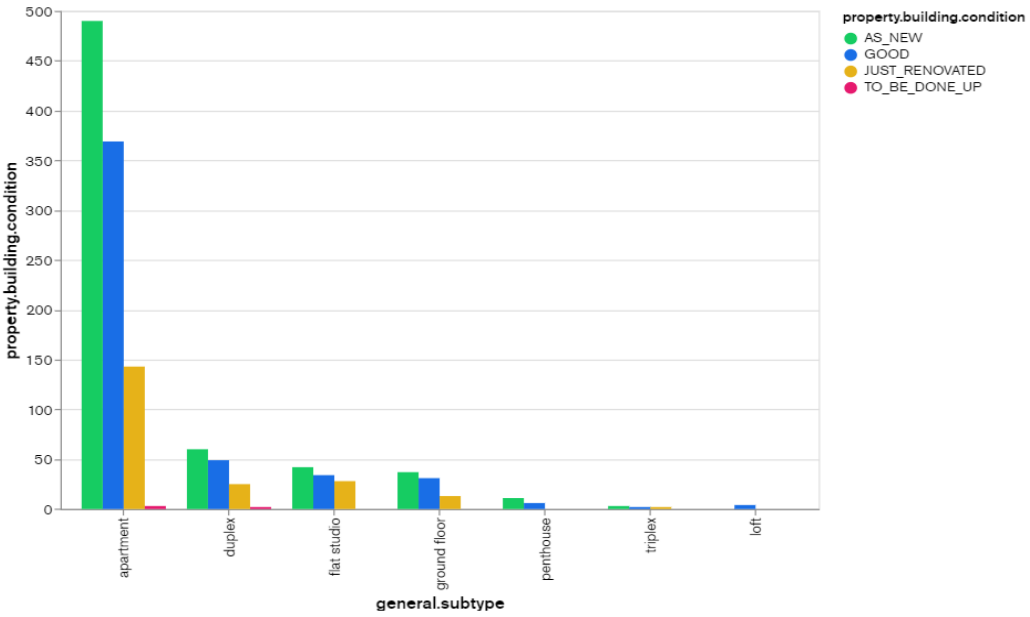
Le sous-type de bien

Parmi les appartements en location, nous retrouvons 0.4% de « loft », 0.5% de « triplex », 1.2% de « penthouse », 5.9% de « Rez-De-Chaussée » (ground floor), 8% de « studio » (flat studio), 9.2% de « duplex » et 74.8% « appartement ».



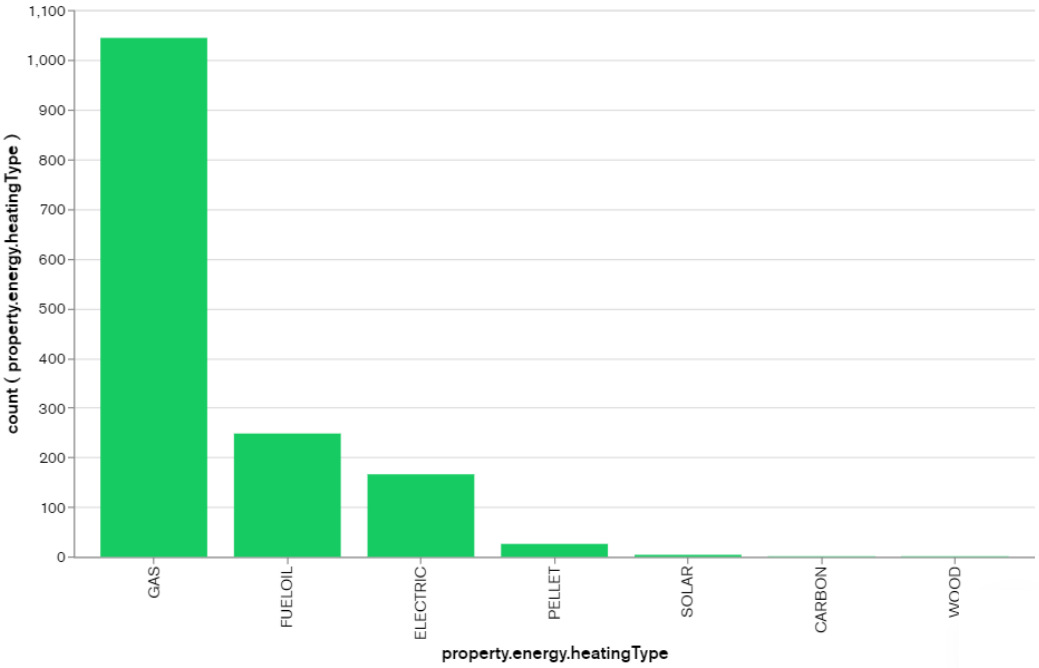
L'état du bien

L'état le plus fréquent est « comme neuf » (AS_NEW). Ensuite, on compte plus d'appartements en état « bon » (GOOD). L'état le moins fréquent, et vraiment peu représenté, est « à rénover » (TO_BE_DONE_UP).



Système de chauffage

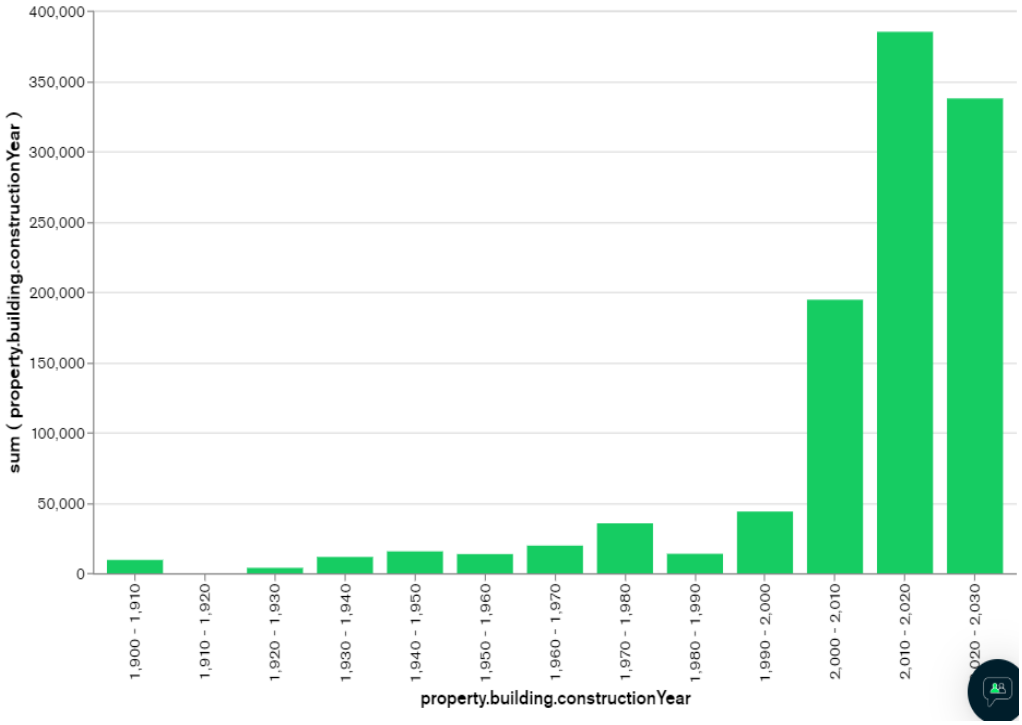
Le système de chauffage le plus répandu (le mode) est le gaz, suivi du mazout.



Exploration des variables continues

Année de construction du bâtiment

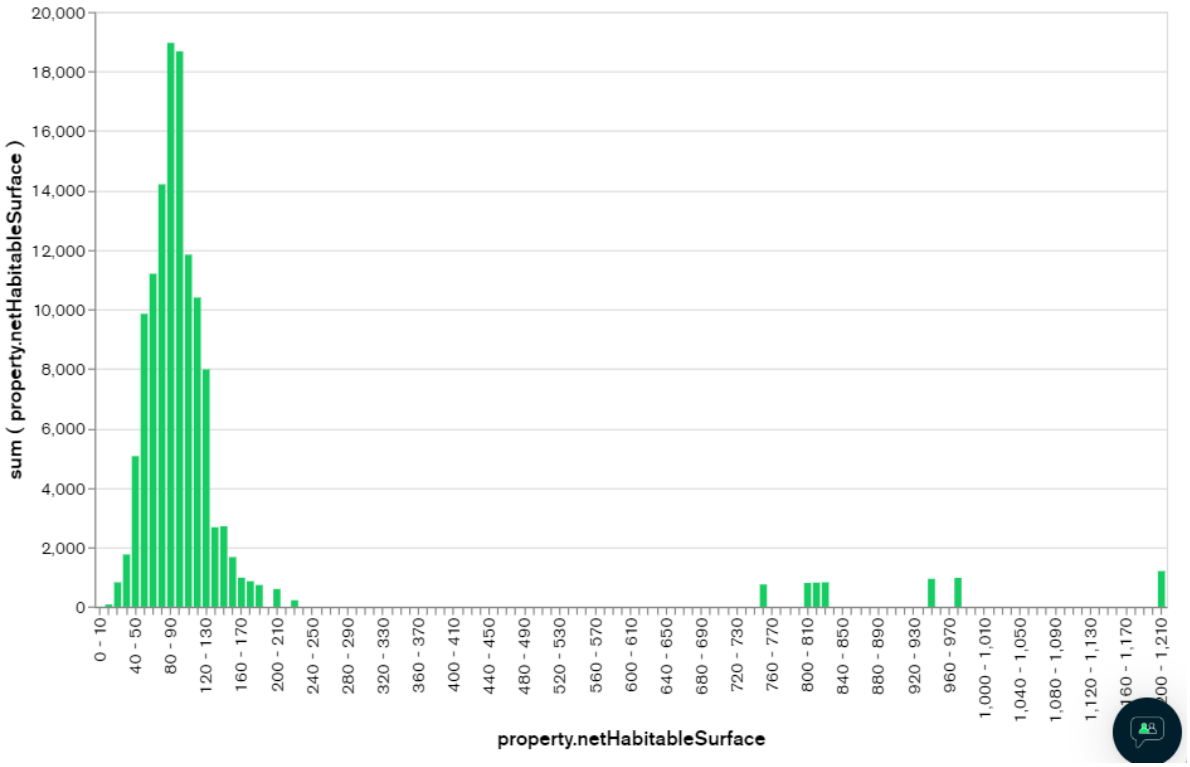
La classe modale est 2010-2020. Les trois tranches les plus fréquentes sont entre 2000 et 2023. Cela signifie que le marché de la location d'appartement concerne principalement des appartements construits récemment.



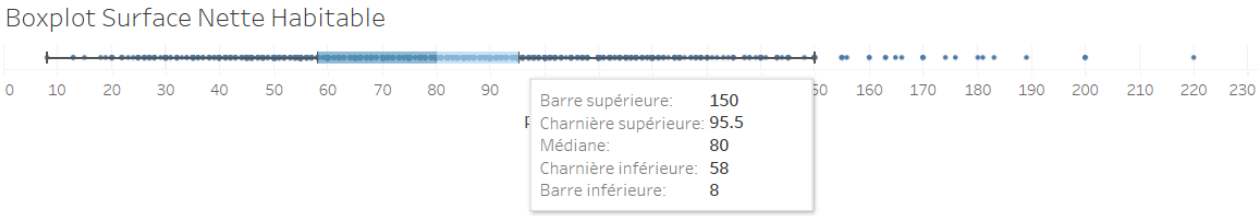
L'histogramme révèle que les valeurs sont concentrées du côté droit du graphique et diminuent en allant vers la gauche. On parle alors de distribution asymétrique positive ou d'une distribution à droite. Cela signifie que la majorité des biens mis en location sont dans des bâtiments qui ont été construits récemment.

La surface nette habitable

L'histogramme révèle que les valeurs sont concentrées du côté gauche. Cependant, les valeurs situées dans la partie droite du graphique semblent aberrantes. Après analyse, nous remarquons qu'il s'agit de petites annonces avec un mauvais sous-type. Il s'agit en réalité d'annonces regroupant plusieurs appartements et dont les surfaces nettes sont sommées.



Si l'on omet ces valeurs aberrantes et qu'on se limite à une taille de surface nette habitable de 250 m² pour un appartement, nous obtenons ce boxplot.



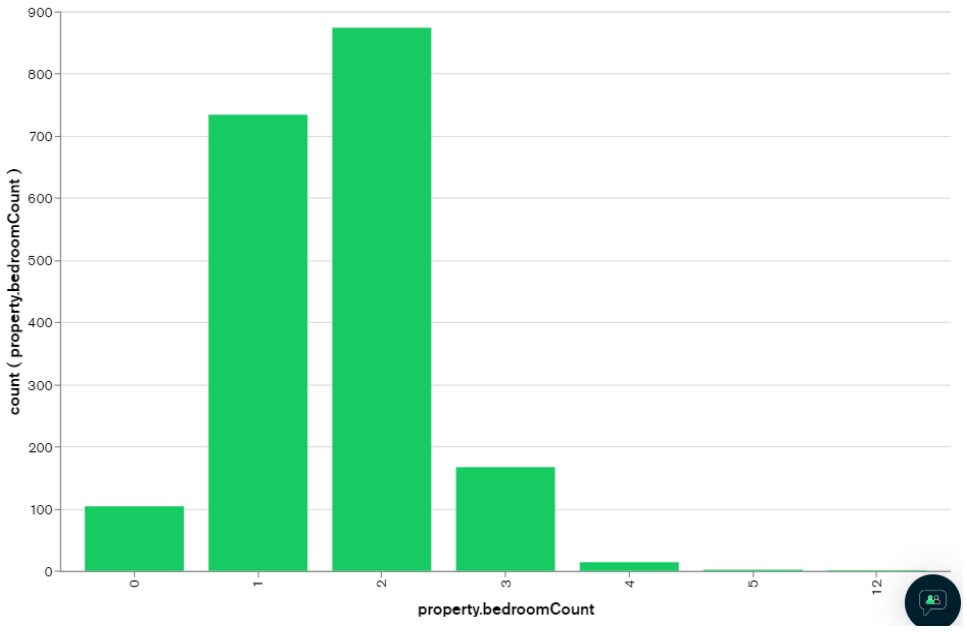
La distribution est plutôt asymétrique. La médiane est plus proche de la charnière supérieure que de la charnière inférieure. Un nombre important de valeurs hautes influencent et tirent donc la moyenne vers le haut.

Il sera intéressant de voir l'analyse avec cluster s'il existe des groupements et de voir leur évolution au cours du temps.

Exploration des variables discrètes

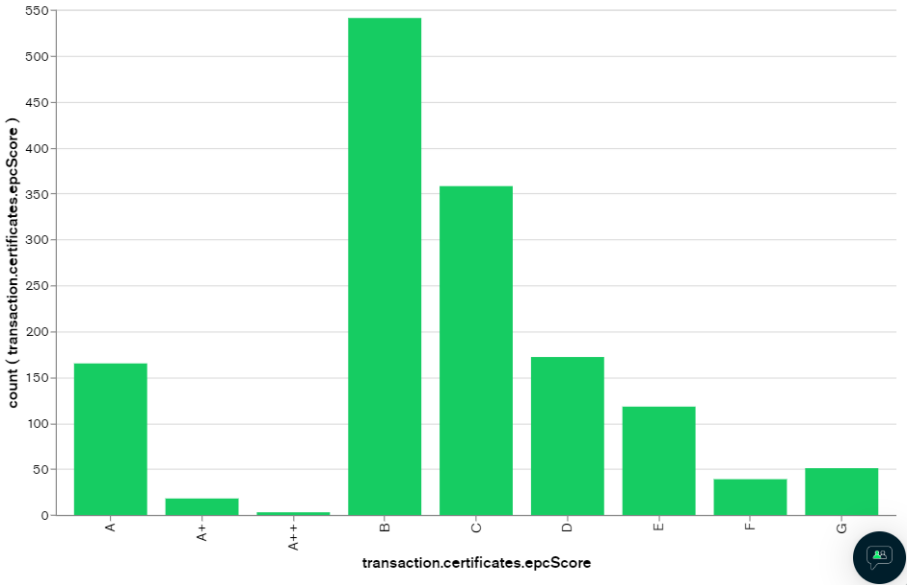
Le nombre de chambres

Le nombre de chambres le plus fréquent est trois. La barre tout à droite est une anomalie. Il s'agit d'une petite annonce qui concerne plusieurs appartements en même temps.



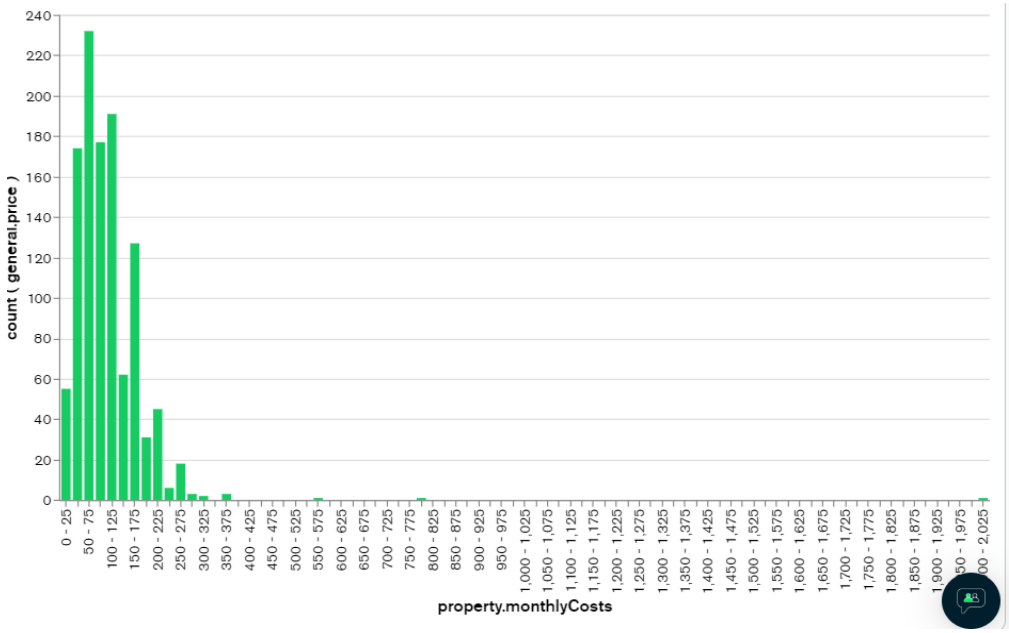
EpcScore

Les appartements sont globalement en bon état d'un point de vue certification énergie, puisque le mode est B et que les valeurs les plus fréquentes suivantes sont C, A(+,++).



monthlyCost (charges locatives)

L'histogramme présente une distribution à gauche. Il existe des valeurs dispersées à que nous devons analyser.



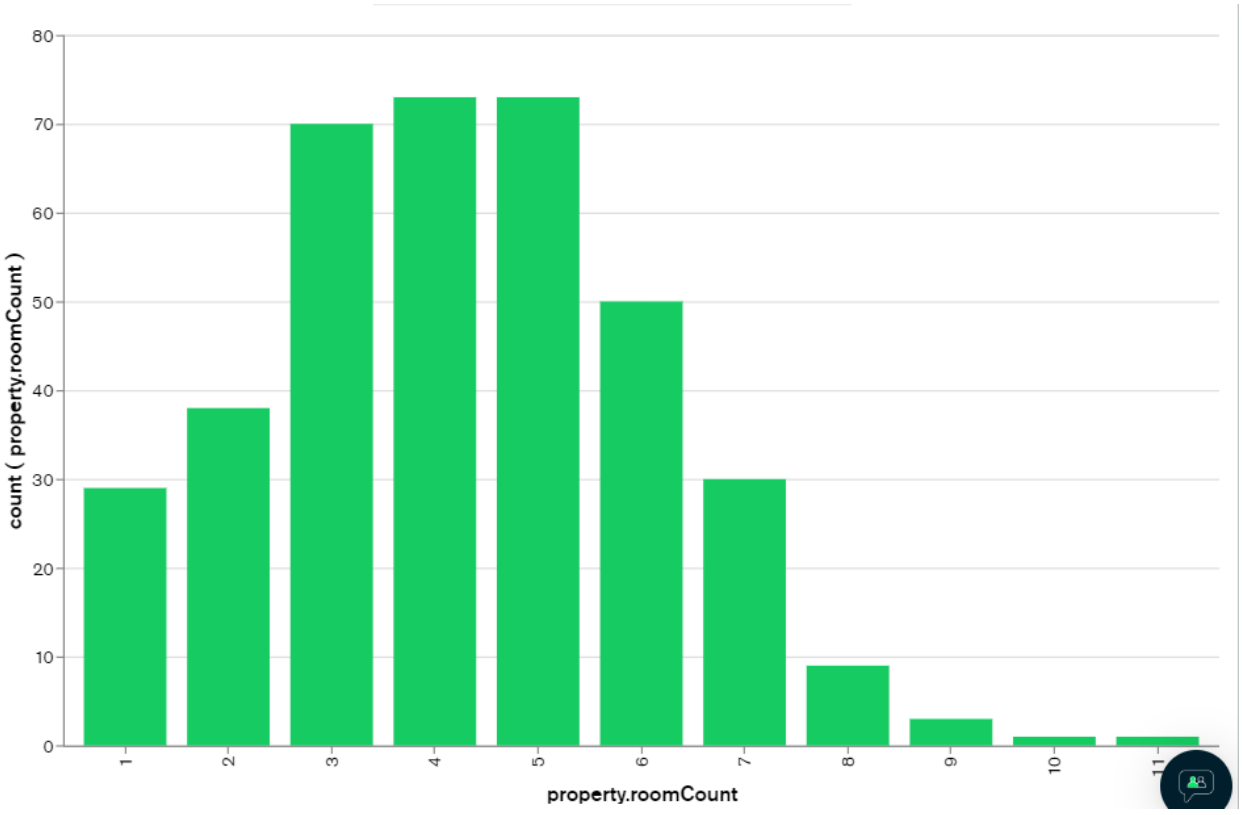
Il ressort qu'il s'agit de petites annonces erronées. Par exemple, l'appartement ayant des charges à 550€ est en réalité une seigneurie « appartement de service » (service flat). Le propriétaire de la petite annonce n'a pas sélectionné le bon sous-type, mais l'indique de manière textuelle dans la description.

Une autre affiche un prix de location à 485€ et des charges à 2010€. Cela n'est pas cohérent. Nous pouvons vérifier dans les mises-à-jour que c'était une erreur et le prix des charges locatives est en réalité de 25€.

Nous constatons une nouvelle fois qu'il est dommage que notre structure de données ne soit pas optimale et ne nous permette pas d'avoir les mises-à-jour pour la génération de graphiques. Pour pallier cela, nous prenons la décision d'ajouter un filtre sur les charges (voir filtre V2).

roomCount – le nombre de pièces pour un appartement

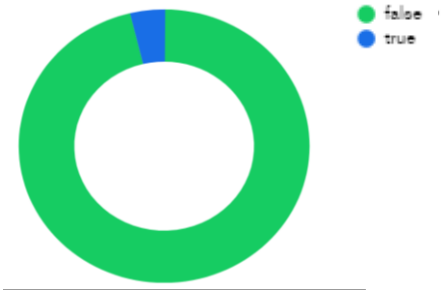
L'histogramme présente une distribution à gauche. Le mode est 5 et presque à égalité avec 4. L'analyse par cluster nous montrera au point (x) qu'on peut distinguer deux groupes d'appartements à l'intérieur de ces fréquences.



hasAirConditionning

Le donut nous permet d'avoir un aperçu rapide sur la distribution d'une donnée booléenne. Nous apprenons que la plupart des appartements n'ont pas de système de climatisation.

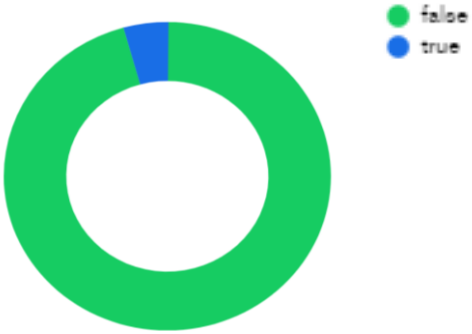
Il serait intéressant de confronter cette information avec d'autres graphiques. Nous pourrions peut-être remarquer qu'un groupe d'un cluster contient une quantité non négligeable des appartements ayant un système de climatisation.



hasCareTakerOrConcierge

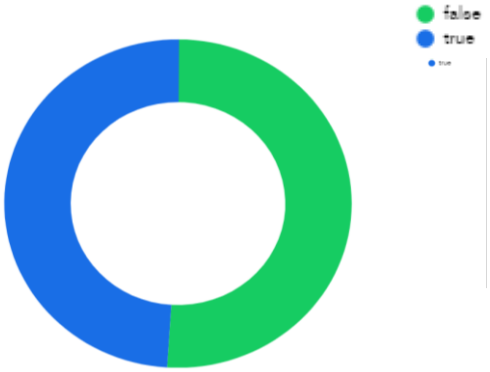
Le donut nous permet d'avoir un aperçu rapide sur la distribution d'une donnée booléenne. Nous apprenons que la plupart des appartements n'ont pas de concierge.

Il serait intéressant de confronter cette information avec d'autres graphiques. Nous pourrions peut-être remarquer qu'un groupe d'un cluster contient une quantité non négligeable des appartements ayant une conciergerie.



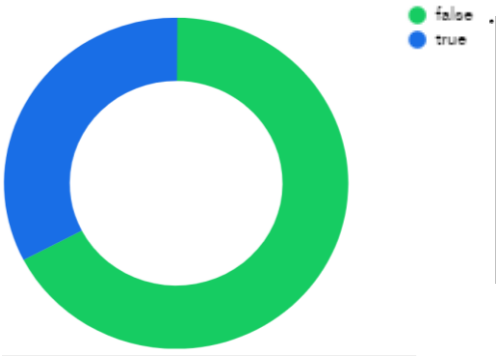
hasLift

Le donut nous permet d'avoir un aperçu rapide sur la distribution d'une donnée booléenne. Nous apprenons que la proportion d'appartement ayant un ascenseur et ceux n'en n'ayant pas est presque la même.



hasVisiophone

Nous apprenons que la majorité des appartements n'ont pas de visiophone.



isFirstOccupation

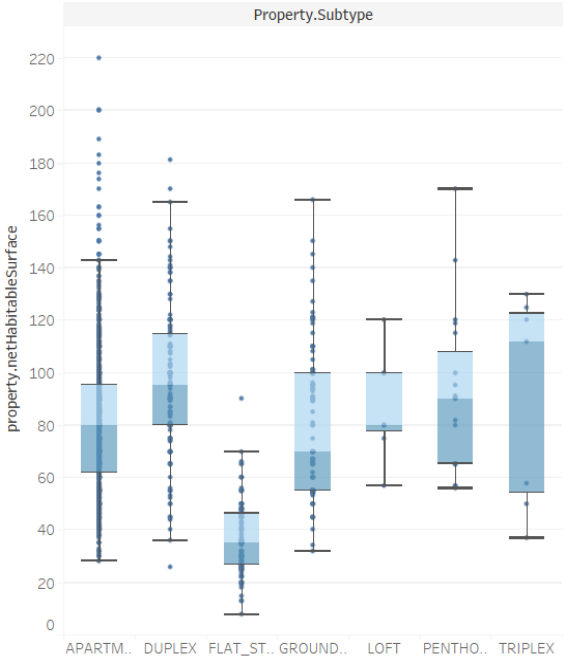
Nous n'avons pas cette valeur, mais elle aurait pu être intéressante à analyser comme étant une valeur déterminante du prix.

12.2.1.2. Tendance centrale

La surface habitable nette par sous-type

On constate une plus grande dispersion des données pour le sous-type « appartement » (apartment). On constate qu'il y a un certain nombre de données au-dessus de la barre supérieure. Cela peut indiquer qu'il y a une concentration de valeurs élevées dans le jeu de données qui sont éloignées de la majorité des autres données.

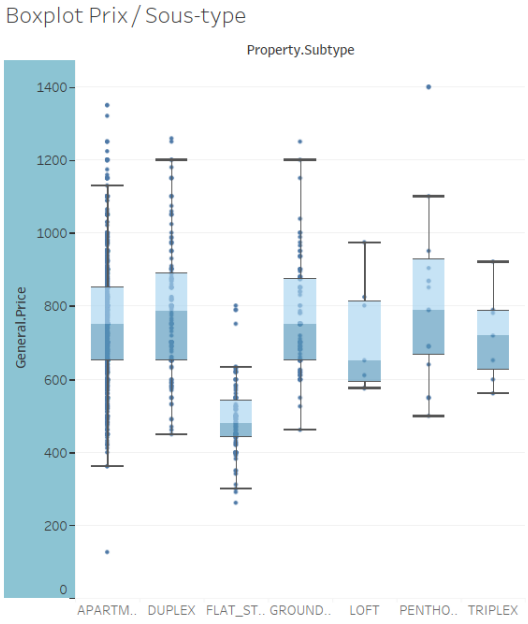
On voit également que le sous-type « rez-de-chaussée » (groundfloor) a une distribution fortement asymétrique. Cela se reflète dans la boxplot par une barre supérieure qui est beaucoup plus longue que la barre inférieure, indiquant une concentration de valeurs élevées.



Le prix par sous-type

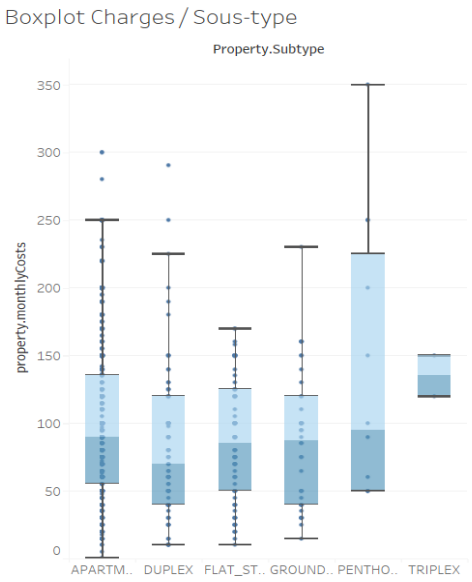
Les sous-types « appartement » (apartment) et « duplex » ont leurs barres basses et hautes fortement éloignées des charnières (de la boîte). Cela peut indiquer une forte asymétrie dans la distribution des données, où il y a un nombre important de valeurs extrêmes (valeurs très élevées ou très basses) qui sont très éloignées de la majorité des autres données du jeu de données.

Cela dit, cette asymétrie ne signifie pas que la moyenne et la médiane des données sont différentes. Puisque les barres hautes et basses sont éloignées à distances plutôt égales, la moyenne n'est pas vraiment biaisée par les valeurs extrêmes. Donc, la moyenne et la médiane sont deux bons représentants de la valeur centrale de l'échantillon.



Les charges par sous-type

Le montant des charges pour le sous-type « appartement » (apartment) est fortement dispersé. Il sera intéressant de chercher s'il existe des groupes d'appartements ayant des caractéristiques communes pour une tranche de charge donnée. Nous pourrions peut-être observer que les appartements ayant les charges les plus élevées sont majoritairement équipés d'ascenseur ou de climatisation ou d'un concierge. Également, une carte choroplèthe pourrait être intéressante afin de découvrir si certaines zones demandes plus de charges que d'autres.



12.2.1.3. Variabilité

Nous allons observer l'écart interquartile (RIQ), la variance et l'écart-type (standard deviation). Ce sont des mesures importantes de la variabilité des données, chacune offrant un aperçu différent de la dispersion des données autour de la moyenne.

Le RIQ est la différence entre le troisième quartile (Q3) et le premier quartile (Q1) de la distribution des données. Il représente la distance entre les valeurs centrales des données et ignore les valeurs extrêmes. Le RIQ est souvent utilisé pour identifier les valeurs aberrantes potentielles dans un ensemble de données. Il donne également une mesure plus précise de la dispersion.

La variance mesure la distance moyenne de chaque point de données par rapport à la moyenne de la distribution. C'est une mesure de la dispersion qui tient compte de chaque point de données de l'ensemble. Une variance élevée indique une grande dispersion des données (grande variabilité), tandis qu'une variance faible indique une faible dispersion (données homogènes).

L'écart-type est la racine carrée de la variance et mesure la distance moyenne des points de données par rapport à la moyenne de la distribution. Il s'agit d'une mesure de la dispersion qui est exprimée dans les mêmes unités que les données originales. Un écart-type élevé indique une grande dispersion des données, tandis qu'un écart-type faible indique une faible dispersion.

Le prix

RIQ : 226,25

Les appartements observés dans la boîte (boxplot - entre Q1 et Q3, soit 50% des observations) ont des prix de location qui peuvent avoir un écart maximal de 226,25€.

Variance : 28958

La variance de 28958 pour la variable du prix de location des appartements signifie que les valeurs de prix sont très dispersées autour de leur moyenne. Plus précisément, cela indique que la différence entre chaque valeur de prix et la moyenne est assez grande. En d'autres termes, il y a une grande variation dans les coûts des charges locatives des appartements. Les données sont très hétérogènes et il pourrait être difficile de trouver des tendances. Il pourrait donc être utile d'explorer davantage la distribution des données pour comprendre la raison de cette forte variance.

Prenons maintenant l'écart-type pour avoir une mesure dans la même unité que la variable elle-même (l'euro).

Écart-type : 170,17

L'écart-type de 170,17 pour la variable du prix de location des appartements signifie que les valeurs de prix sont dispersées autour de leur moyenne d'environ 170,17 euros. Cela donne une idée de la variabilité du prix de la location des appartements.

Les charges

RIQ : 38

Les appartements observés dans la boîte (boxplot - entre Q1 et Q3, soit 50% des observations) ont des coûts de charges locatives qui peuvent avoir un écart maximal de 38€.

Variance : 28958

La variance de 28958 pour la variable du prix de location des appartements signifie que les valeurs de prix sont relativement peu dispersées autour de leur moyenne. Plus précisément, cela indique que les charges locatives des appartements sont relativement similaires les unes aux autres, avec des écarts faibles par rapport à la moyenne. En d'autres termes, une variance faible indique que les données sont relativement homogènes.

Prenons maintenant l'écart-type pour avoir une mesure dans la même unité que la variable elle-même (l'euro).

Écart-type : 63,04

L'écart-type de 63,04 pour la variable du prix de location des appartements signifie que les valeurs de prix sont dispersées autour de leur moyenne d'environ 63,04 euros. Cela donne une idée de la variabilité du coût des charges locatives des appartements.

12.2.1.4. Régressions et corrélations

Nous testerons des hypothèses sur les données afin de déterminer si deux variables sont corrélées. L'outil de visualisation que nous utiliserons est le scatter plot. C'est un type de graphique qui permet de représenter les relations entre deux variables quantitatives en les plaçant sur les axes X et Y. Nous l'utiliserons pour visualiser la corrélation ou l'association entre deux variables continues.

Lorsqu'on lit un scatter plot, on peut observer la dispersion des points sur le graphique. Si les points sont concentrés autour d'une ligne droite, cela indique une forte corrélation positive ou négative entre les deux variables. Si les points sont dispersés de manière uniforme sur le graphique, cela indique une faible ou aucune corrélation entre les deux variables.

Il est également possible de détecter des valeurs aberrantes ou des points qui sont très éloignés des autres points sur le graphique, ce qui peut influencer les résultats de l'analyse. Nous utilisons plusieurs librairies de Python.

En ce qui concerne les variables qualitatives ou non continues, nous utiliserons plutôt la moyenne et/ou un histogramme.

Présentations des tests d'hypothèse

- Prix demandé

Cette propriété peut varier considérablement en fonction de divers facteurs tels que la surface nette habitable, la présence d'un ascenseur, la présence d'un concierge, la présence d'un système de climatisation, et enfin, l'emplacement géographique.

- La surface nette habitable

Cette propriété peut varier selon son emplacement géographique.

- Les charges locatives

Cette propriété peut varier selon son emplacement géographique.

- Présence ou non d'un système de climatisation

Cette propriété peut varier selon son emplacement géographique.

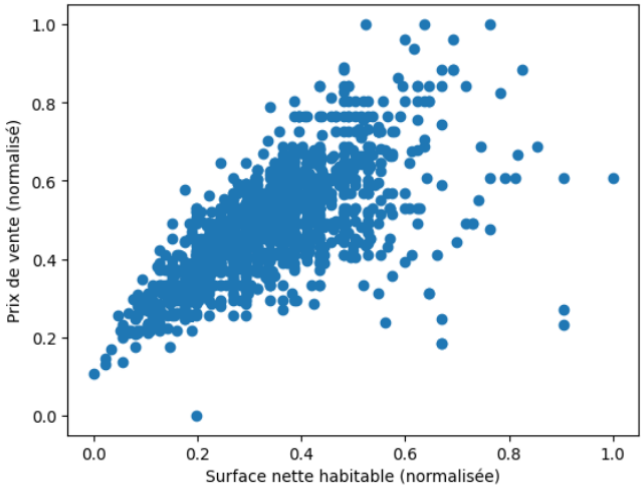
- Présence ou non d'une conciergerie

Cette propriété peut varier selon son emplacement géographique.

Prix demandé vs. Surface nette habitable

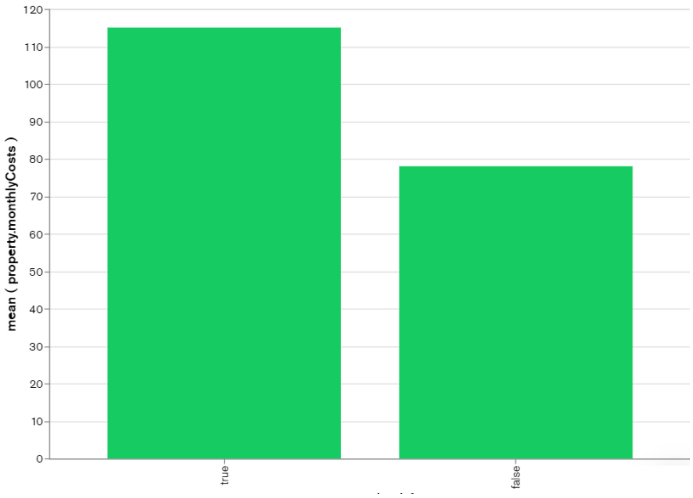
Puisque nous comparons un prix en euro et une surface habitable en mètre(s) carré(s) (unités différentes), il est judicieux de normaliser les données sur les deux axes pour mieux visualiser la relation entre les variables et éviter d'avoir un axe qui domine l'autre.

Le nuage de points semble suivre le tracé d'une droite. Nous distinguons donc une relation linéaire dans laquelle l'augmentation d'une variable entraine à peu près au même rythme une augmentation de l'autre variable.



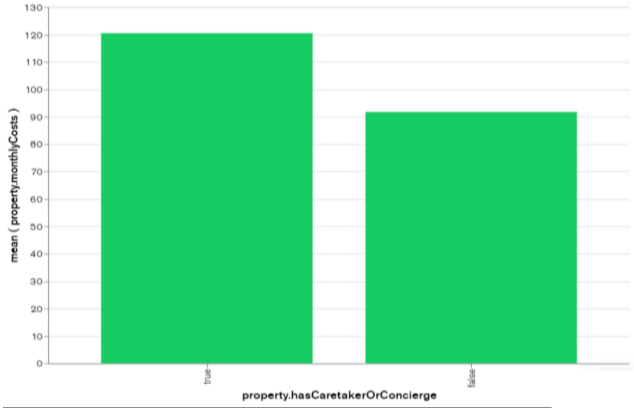
Prix demandé vs. Présence d'un ascenseur ou non

Le montant des charges locatives d'un appartement avec ascenseur est en moyenne plus élevé qu'un appartement sans ascenseur.



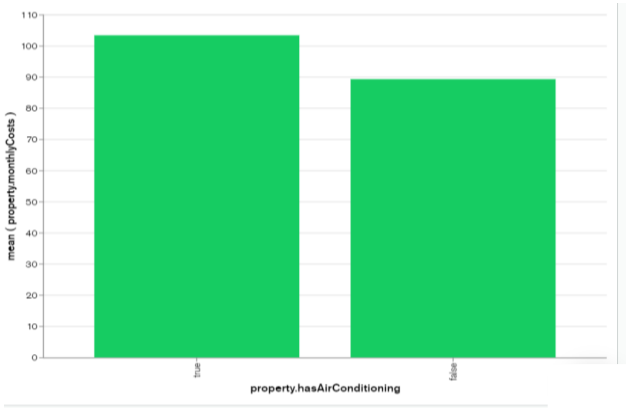
Prix demandé vs. Présence d'une conciergerie ou non

Le montant des charges locatives d'un appartement avec conciergerie est en moyenne plus élevé qu'un appartement sans conciergerie.



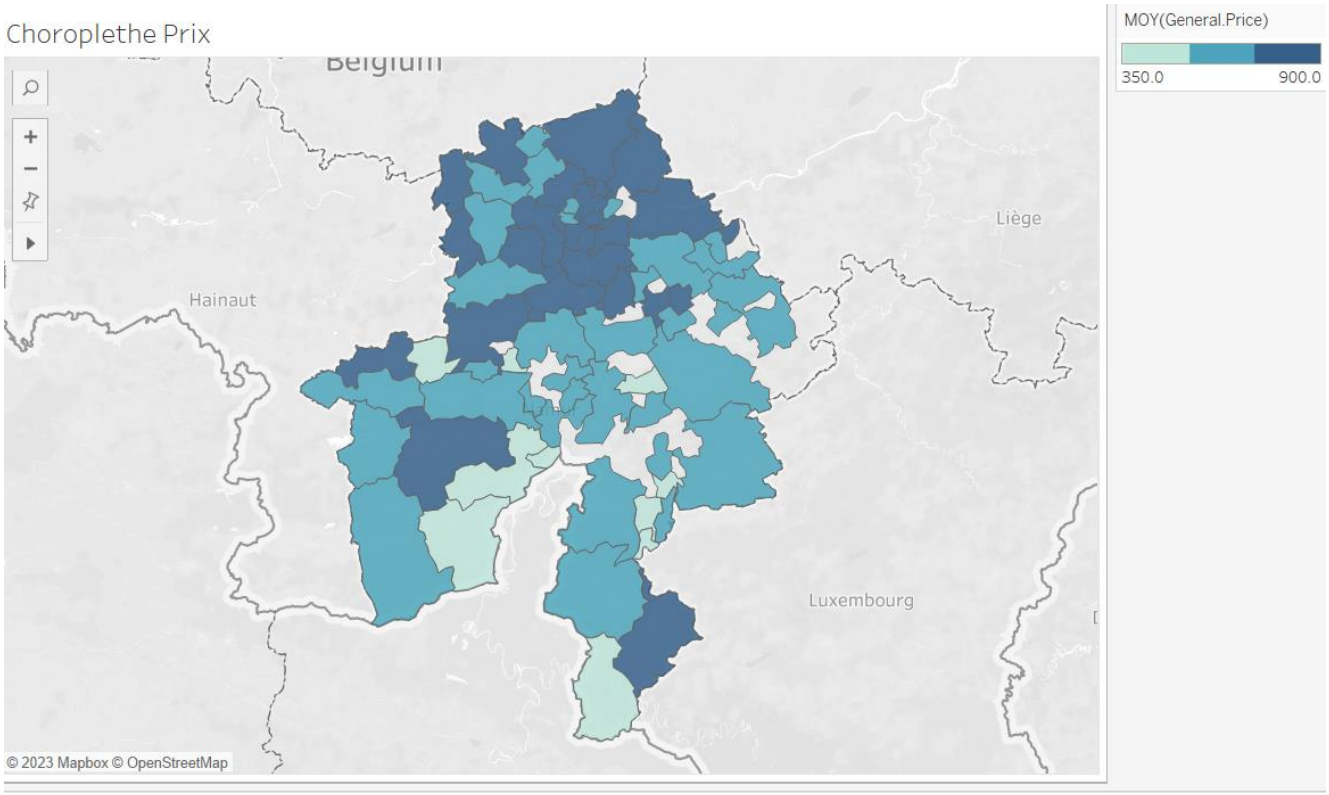
Prix demandé vs. Présence d'un système de climatisation

Le montant des charges locatives d'un appartement avec système de climatisation est en moyenne légèrement plus élevé qu'un appartement sans système de climatisation.



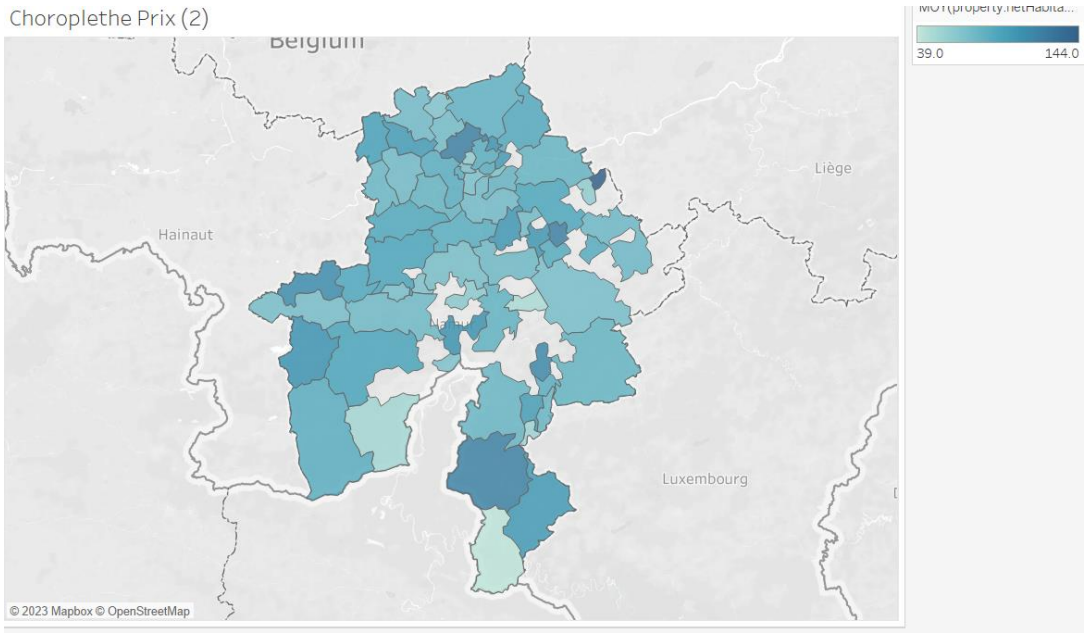
Prix demandé vs. Emplacement du bien

Le prix tant à être plus cher au nord de la province et les prix les moins chers sont le long de la frontière avec la France.



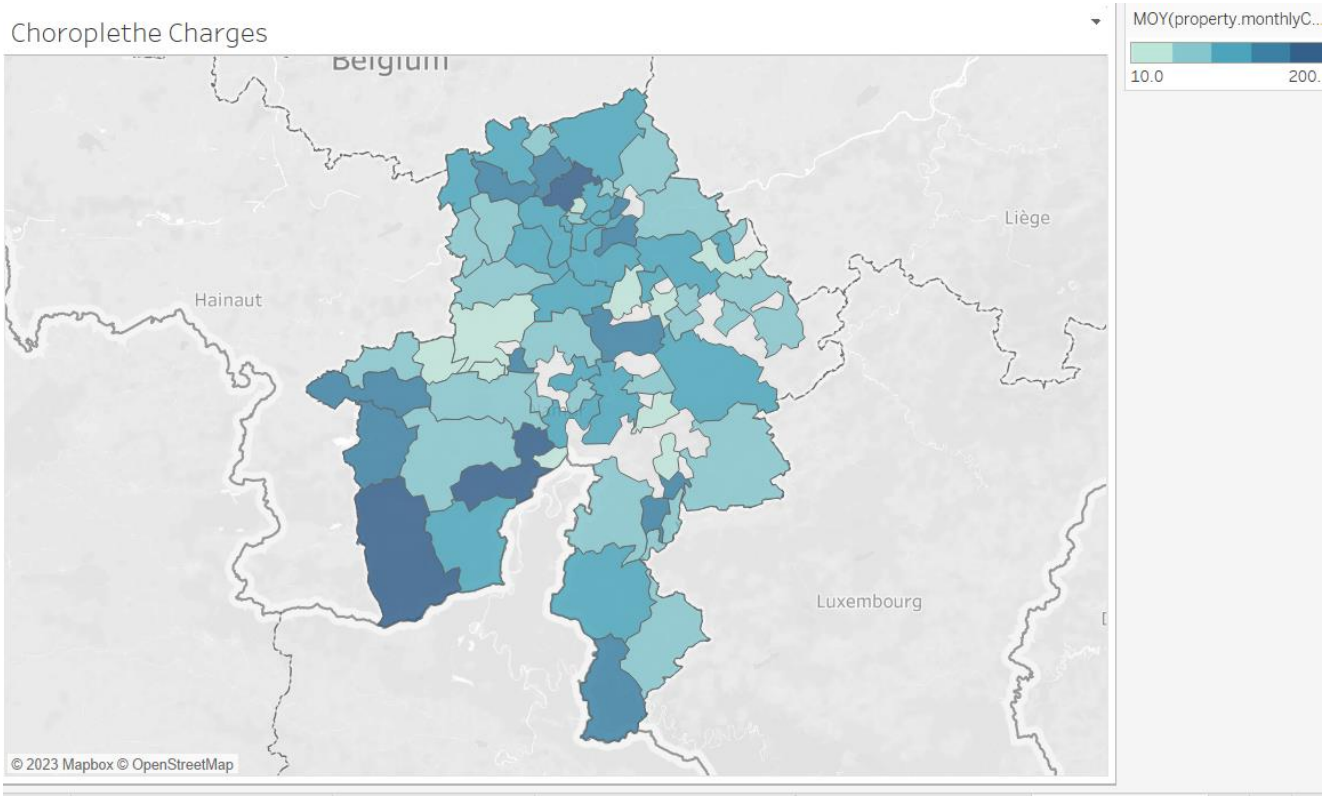
La surface nette habitable vs. Emplacement du bien

Les surfaces nettes habitables moyennes sont réparties sur l'ensemble de la province.



Les charges locatives vs. Emplacement du bien

Nous n'observons pas de démarcation nord-sud ni est-ouest. Cependant, nous pourrions toujours nous poser la question de savoir s'il existe une raison pour une telle hétérogénéité.

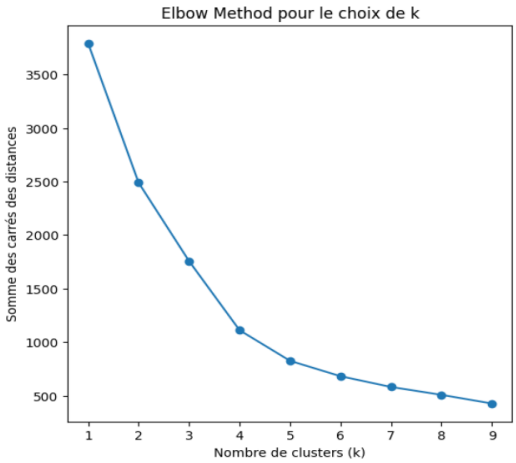


12.2.1.5. Clusters

Cluster roomCount et prix

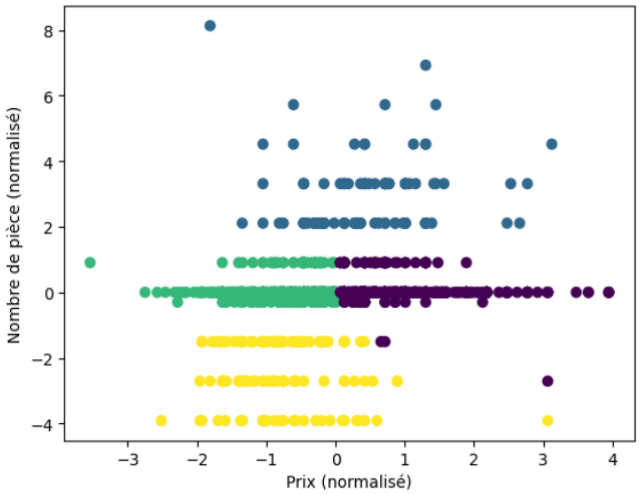
L'interprétation de l'elbow plot se fait visuellement en cherchant le point d'inflexion, c'est-à-dire le point où la courbe cesse de diminuer brusquement et devient plus plate. Ce point marque le nombre optimal de clusters à utiliser.

Ici, le coude est 4.



Les clusters

Les variables ont été normalisées pour pouvoir être comparées. Nous pouvons maintenant identifier des groupes d'appartements en location avec des caractéristiques similaires en termes de prix et de nombre de pièces.



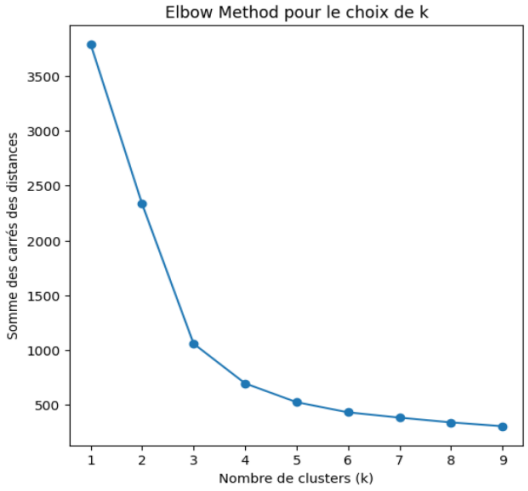
Nous avons quatre groupes :

	« Bas »	« Moyen-gauche »	« Moyen-droit »	« Haut »
Nombre d'appartements	134	850	818	94
Prix moyen	595,3 €	604,45 €	869,09 €	809,3 €
Nb. pièces moy.	2,29	4 ,27	4,27	6,7
Prix min.	300 €	125 €	740 €	420 €
Prix max.	1.250 €	730 €	1.400 €	1.260 €
Nb. pièce min.	1	4	2	6
Nb. pièce max.	3	5	5	11

Cluster surface habitable nette et prix

L'interprétation de l'elbow plot se fait visuellement en cherchant le point d'inflexion, c'est-à-dire le point où la courbe cesse de diminuer brusquement et devient plus plate. Ce point marque le nombre optimal de clusters à utiliser.

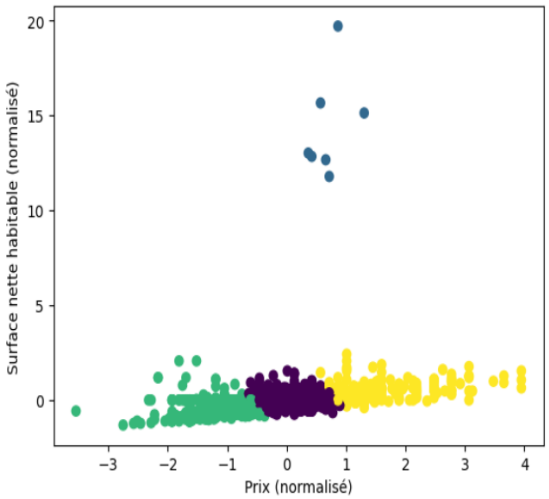
Ici, le coude est 3.



Les clusters

Les variables ont été normalisées pour pouvoir être comparées. Nous pouvons maintenant identifier des groupes d'appartements en location avec des caractéristiques similaires en termes de prix et surface nette habitable.

Ce clustering nous permet donc d'identifier 4 groupes, dont l'un semble être constitué de valeurs aberrantes (demandant une analyse des données dans le jeu de données).



Nous avons quatre groupes :

	« Gauche »	« Milieu»	« Droite »	« Haut »
Nombre d'appartements	609	912	368	7
Prix moyen	546,4 €	748, 69 €	975,69 €	847,14 €
Surf. nette moy.	60,48 m²	82,26 m²	106,05 m²	900,57 m²
Prix min.	125 €	620 €	825 €	790 €
Prix max.	665 €	880 €	1.400 €	950 €
Surf. nette min.	8 m²	40 m²	60 m²	752 m²
Surf. nette max.	200 m²	170 m²	220 m²	1.202 m²

12.2.1.6. Conclusion

Le jeu de données des appartements à louer est de bonne qualité. Les petites annonces sont probablement de meilleures qualités puisqu'elles sont majoritairement postées par des professionnels du métier.

L'analyse par cluster nous a montré que des groupes de variables similaires existaient dans la distribution du prix et de la surface nette habitable. Cette information est utile et permettra de servir de base à une segmentation lors d'analyses futures.

En partant de nos hypothèses de base, nous avons appris qu'il n'y avait pas de lien évident entre charges locatives et emplacement du bien.

Nous avons également utilisé des cartes choroplèthe qui nous ont révélé le lien qui existe entre prix moyen et zone géographique.

Nous avons aussi pu détecter des valeurs aberrantes ainsi que des erreurs. En effet, certains scatter plot ne contiennent que peu de valeurs. Cet ADE aura donc aussi le mérite d'offrir l'opportunité de revoir le filtrage des données à la base de l'extraction du jeu de données.