Analyse du marché immobilier français à partir de données issues du web scraping

Édouard & Élise | D.U. DATA ANALYTICS

1. Introduction

1.1 Contexte et enjeux du marché immobilier

Le marché immobilier français occupe une place très importante dans l'économie Française et dans les préoccupations des ménages. Les prix de l'immobilier, et plus particulièrement le prix au mètre carré, constituent un indicateur clé pour les acheteurs, les vendeurs, ainsi que pour les investisseurs.

Cependant, ce marché est caractérisé par une forte hétérogénéité, tant sur le plan géographique que structurel : le prix d'un bien dépend de nombreux facteurs tels que la localisation, la surface, le type de bien, les caractéristiques du logement ou encore sa performance énergétique (DPE).

Dans un contexte où l'accès aux données devient de plus en plus crucial, les plateformes d'annonces immobilières représentent une source d'information riche mais non structurée. Leur exploitation nécessite des compétences techniques spécifiques, notamment en web scraping et en analyse de données.

1.2 Objectifs du projet

Ce projet s'inscrit dans le cadre du **DU Data Analytics** et a pour objectif de mettre en pratique l'ensemble de la chaîne de traitement des données, depuis leur collecte jusqu'à leur visualisation.

Les objectifs principaux sont :

- apprendre et maîtriser les techniques de **web scraping** en Python ;
- nettoyer, structurer et enrichir des données issues du web ;
- réaliser une **analyse exploratoire** pertinente ;
- concevoir un **dashboard interactif** facilitant la compréhension des données.

La problématique centrale du projet est la suivante :

> **Comment le prix au mètre carré varie-t-il en fonction de la localisation, de la surface et du type de biens immobiliers en France ?**

Pour répondre à cette question, nous avons choisi de concentrer notre analyse sur les **20 plus grandes villes françaises** (Paris, Marseille, Lyon, Toulouse, Nice, Nantes, Montpellier, Strasbourg, Bordeaux, Lille, Rennes, Reims, Toulon, Saint Etienne, Le havre, Grenoble, Dijon, Angers, Nîmes, Clermont-Ferrand) afin de garantir à la fois la pertinence des résultats et la comparabilité des données.

2. État de l'art

2.1 Présentation du scraping et des outils Python

Le web scraping est une technique permettant **d'extraire automatiquement des données depuis des pages web**. Dans le domaine immobilier, cette méthode est largement utilisée pour collecter des annonces, analyser les tendances de prix ou comparer différents marchés locaux.

Les principales difficultés rencontrées dans notre projets sont :

- la structure HTML variable des pages ;
- la présence de mécanismes anti-bot (CAPTCHA, limitations de requêtes) ;
- la qualité hétérogène des données collectées.

2.2 Travaux similaires et études existantes

De nombreuses études de marché immobilier s'appuient sur des données issues de portails d'annonces ou d'organismes publics. Toutefois, ces analyses reposent souvent sur des données agrégées ou propriétaires.

Notre approche se distingue par :

- l'utilisation exclusive de données **scrapées** ;
- une analyse détaillée au niveau de l'annonce ;
- la mise à disposition d'un **outil interactif** (Streamlit) accessible au grand public.

3. Méthodologie

3.1 Description du site cible

Le site **paruvendu.fr** a été choisi comme source de données en raison de :

- la richesse des informations disponibles dans les annonces ;
- la couverture nationale du site ;
- l'accessibilité des pages de résultats par ville.

Le scraping a été réalisé sur les **cinq premières pages de résultats** pour chacune des **20 villes sélectionnées**, représentant environ 30 annonces par page.

3.2 Architecture technique

Le projet repose sur une architecture simple mais robuste :

- scripts Python pour le scraping et le nettoyage ;
- fichiers CSV intermédiaires (RAW et CLEAN) ;
- analyse exploratoire avec pandas, matplotlib et seaborn ;
- visualisation finale via un dashboard Streamlit.

Outils et bibliothèques Python utilisés

Le projet a été développé intégralement en Python et s'appuie sur un ensemble de bibliothèques complémentaires, chacune répondant à un besoin spécifique du pipeline de données.

Pour la **collecte des données**, nous avons utilisé :

- `requests`, afin d'effectuer les requêtes HTTP vers les pages du site cible ;
- `BeautifulSoup` (module `bs4`), pour parser et extraire les informations depuis le code HTML des pages de résultats et des pages détaillées des annonces ;
- `time` et `sleep`, pour contrôler la fréquence des requêtes et limiter les risques de blocage par des mécanismes anti-bot ;
- `json`, pour la gestion des fichiers de checkpoints permettant de reprendre le scraping en cas d'interruption ;
- `csv` et `os`, pour la gestion des fichiers et des chemins locaux.

Pour le **nettoyage et la structuration des données**, les bibliothèques suivantes ont été utilisées :

- `pandas`, pour la manipulation des DataFrames, le nettoyage, la transformation et l'agrégation des données ;
- `numpy`, pour les calculs numériques et statistiques ;
- `re` (expressions régulières), pour l'extraction et la standardisation des informations textuelles (prix, surfaces, caractéristiques des biens).

Pour l'**enrichissement géographique** des données :

- `geopy` (via le géocodeur `Nominatim`), utilisé afin de convertir certaines localisations textuelles en coordonnées géographiques exploitables (latitude et longitude).

Pour l'**analyse exploratoire et la visualisation des données** :

- `matplotlib`, comme bibliothèque de visualisation de base ;
- `seaborn`, pour la production de graphiques statistiques avancés (histogrammes, boxplots, scatterplots, heatmaps), facilitant l'interprétation des résultats.

Enfin, pour la **visualisation interactive et la restitution des résultats** :

- `Streamlit`, utilisé pour développer un dashboard interactif intégrant des filtres dynamiques et des visualisations exploratoires ;
- `pydeck`, pour la création de cartes de chaleur géographiques permettant d'analyser la répartition spatiale des prix immobiliers.

3.3 Stratégie de scraping et gestion des mécanismes anti-bot

Le scraping des **pages de résultats par ville** sur le site *paruvendu.fr* ne présentait pas de difficulté technique majeure. Ces pages regroupent plusieurs annonces (environ 30 par page) et permettent de collecter les informations principales telles que le titre, le prix, la description et certaines caractéristiques du bien.

En revanche, l'accès aux **pages individuelles des annonces** s'est révélé plus problématique. Cette étape était indispensable afin de récupérer des informations complémentaires, notamment la **localisation précise des biens**, nécessaire à la construction de visualisations géographiques telles que les cartes de chaleur. L'enchaînement de requêtes vers ces pages de détail a entraîné des **blocages fréquents du site**, liés à la détection d'un trafic automatisé (mécanismes anti-bot et CAPTCHA).

Afin de garantir la robustesse du processus de collecte, plusieurs stratégies ont été mises en place.

Tout d'abord, un **système de checkpoints** a été développé. Celui-ci repose sur la sauvegarde, dans un fichier JSON, de la position courante du scraping (indice de la ville et numéro de page). En cas d'interruption volontaire ou de blocage par le site, le scraping peut ainsi être **repris automatiquement** à partir du dernier point valide, sans perte des données déjà collectées.

Ensuite, une **détection explicite des pages CAPTCHA** a été implémentée. Le contenu HTML des réponses est analysé afin d'identifier des messages caractéristiques indiquant un trafic inhabituel ou la présence d'un CAPTCHA. Lorsqu'un tel cas est détecté, le scraping est immédiatement interrompu et l'état courant est sauvegardé via le checkpoint, évitant ainsi des requêtes répétées susceptibles d'aggraver le blocage.

Par ailleurs, une **limitation volontaire du nombre de requêtes par exécution** a été introduite, à travers un seuil maximal d'annonces collectées par run. Cette approche permet de répartir le scraping sur plusieurs exécutions distinctes, réduisant significativement la charge exercée sur le site cible.

Enfin, des **temps de pause contrôlés** ont été intégrés entre les requêtes, notamment lors de l'accès aux pages de détail des annonces. Ces délais visent à simuler un comportement utilisateur plus naturel et à diminuer la probabilité de déclenchement des mécanismes anti-bot.

L’ensemble de ces dispositifs a permis de mettre en place un **processus de scraping robuste, résilient et maîtrisé**, capable de fonctionner malgré les restrictions du site et de collecter un volume de données conséquent de manière fiable. Cette stratégie constitue un élément central du projet et illustre l’importance d’une approche méthodologique rigoureuse lors de la collecte de données web à grande échelle.

4. Résultats et analyses

4.1 Description du jeu de données

À l’issue du scraping :

- **2465 annonces** ont été collectées dans le fichier RAW, contenant 7 variables ;
- Après nettoyage et enrichissement, le fichier CLEAN contient **2145 annonces** et **18 variables**.

Les variables créées sont :

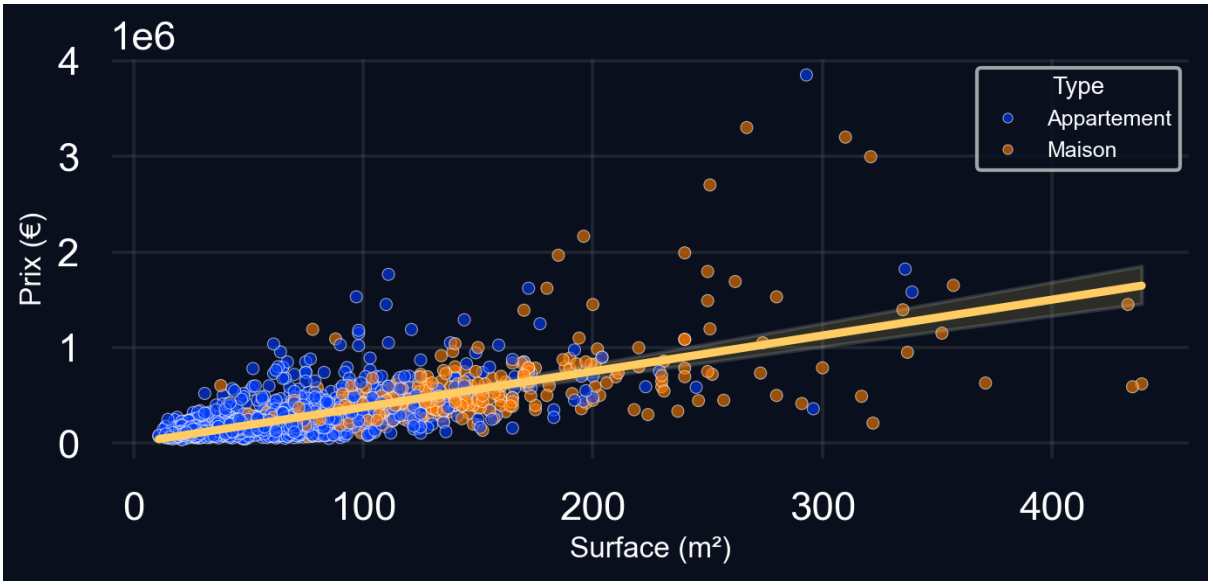
- Prix_m2 (regex de Prix)
- Surface_m2 (regex Titre)
- Type (variable catégorielle Appartement/Maison)
- Pièce (regex Détails)
- Chambre (regex Détails)
- Garage (regex Détails)
- Balcon (regex Détails)
- Ascenseur (regex Détails)
- Terrain_m2 (regex Détails)
- DPE (regex Détails)
- Latitude (nominatim sur Localisation)
- Longitude (nominatim sur Localisation)

4. Résultats et analyses

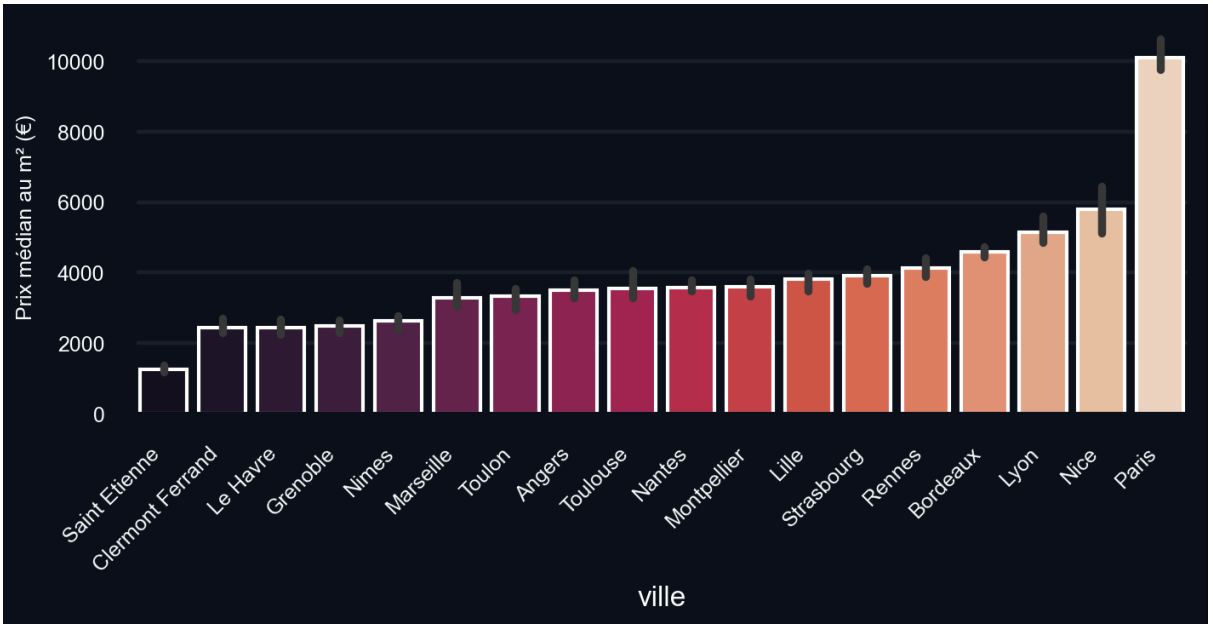
4.2 Statistiques clés

L’analyse descriptive du jeu de données met en évidence plusieurs éléments structurants du marché immobilier étudié.

Tout d’abord, les **prix de vente** présentent une **forte dispersion**, avec des écarts importants entre les biens les moins chers et les plus onéreux. Cette dispersion traduit l’hétérogénéité du marché, notamment en termes de localisation, de surface et de type de bien.



Le **prix au mètre carré** apparaît quant à lui comme un indicateur particulièrement pertinent pour comparer les biens entre eux. Les valeurs médianes varient fortement selon les villes, confirmant l'importance du facteur géographique dans la formation des prix immobiliers.



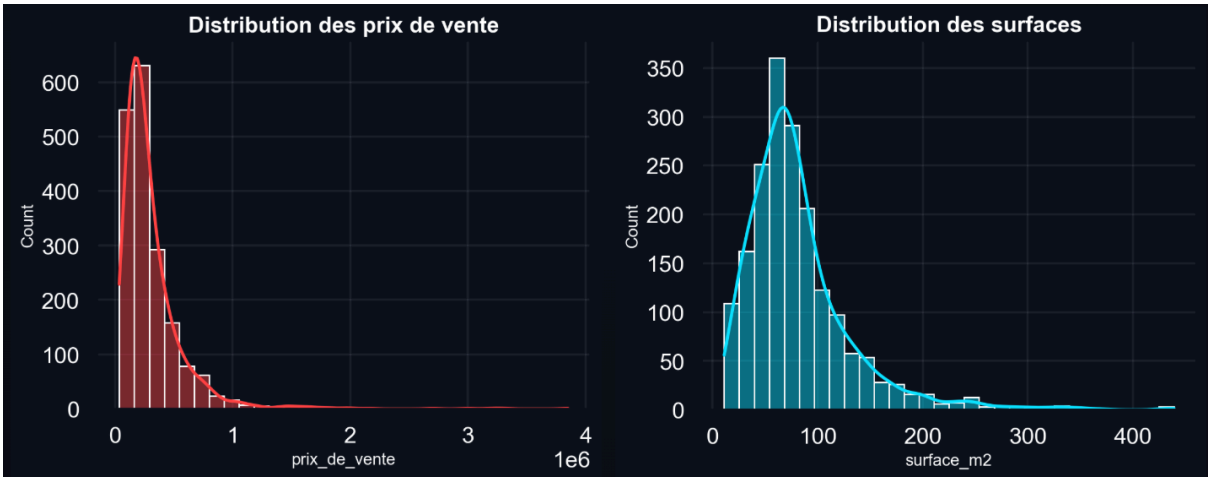
En revanche, la **surface médiane** des biens reste relativement comparable entre les grandes villes étudiées. Cette relative homogénéité suggère que les écarts de prix observés sont davantage liés à la localisation et aux caractéristiques qualitatives des biens qu'à des différences majeures de surface.

Ces premiers indicateurs fournissent une **vision globale du marché** et constituent une base essentielle pour les analyses exploratoires approfondies présentées par la suite.

4.3 Analyse des distributions

L'étude des distributions à l'aide d'**histogrammes** permet de mieux comprendre la structure des données.

La distribution des **prix de vente** est fortement **asymétrique à droite**. La majorité des biens se concentre dans une fourchette de prix intermédiaire, tandis qu'un nombre plus restreint d'annonces affiche des prix très élevés. Ces valeurs extrêmes correspondent principalement à des biens de grande surface ou situés dans des zones à forte attractivité, comme les grandes métropoles.



De manière similaire, la distribution des **surfaces** révèle une grande variabilité. Si la majorité des biens se situe autour de surfaces modestes à moyennes, une queue de distribution importante illustre la présence de biens de grande superficie, notamment des maisons individuelles.

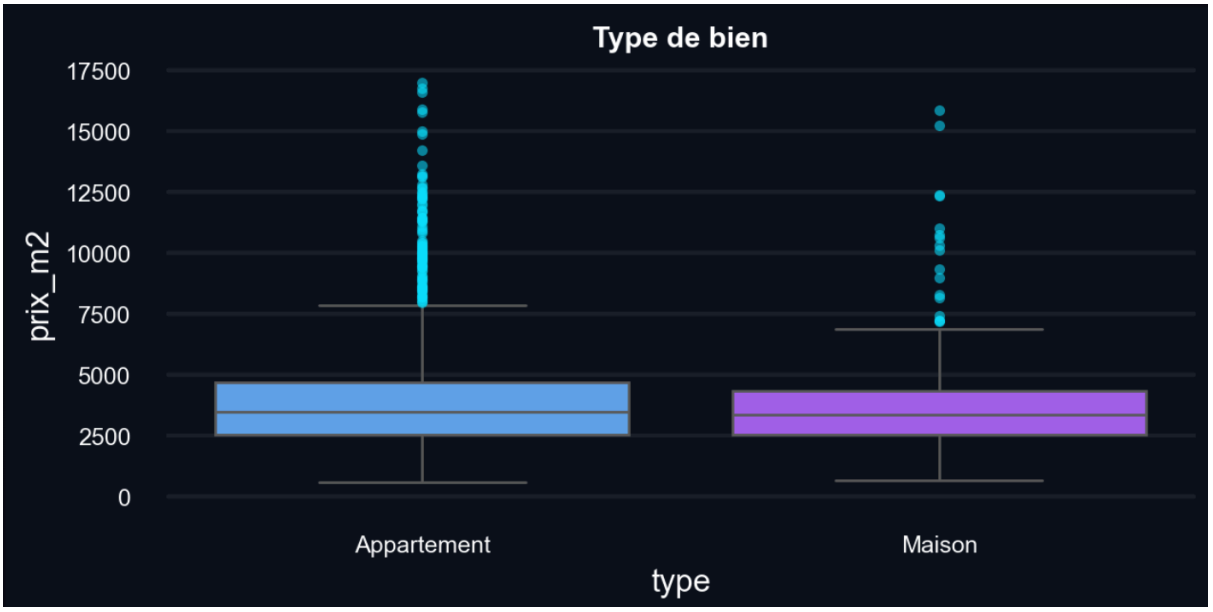
Les **boxplots** enrichissent cette analyse en facilitant les comparaisons entre groupes :

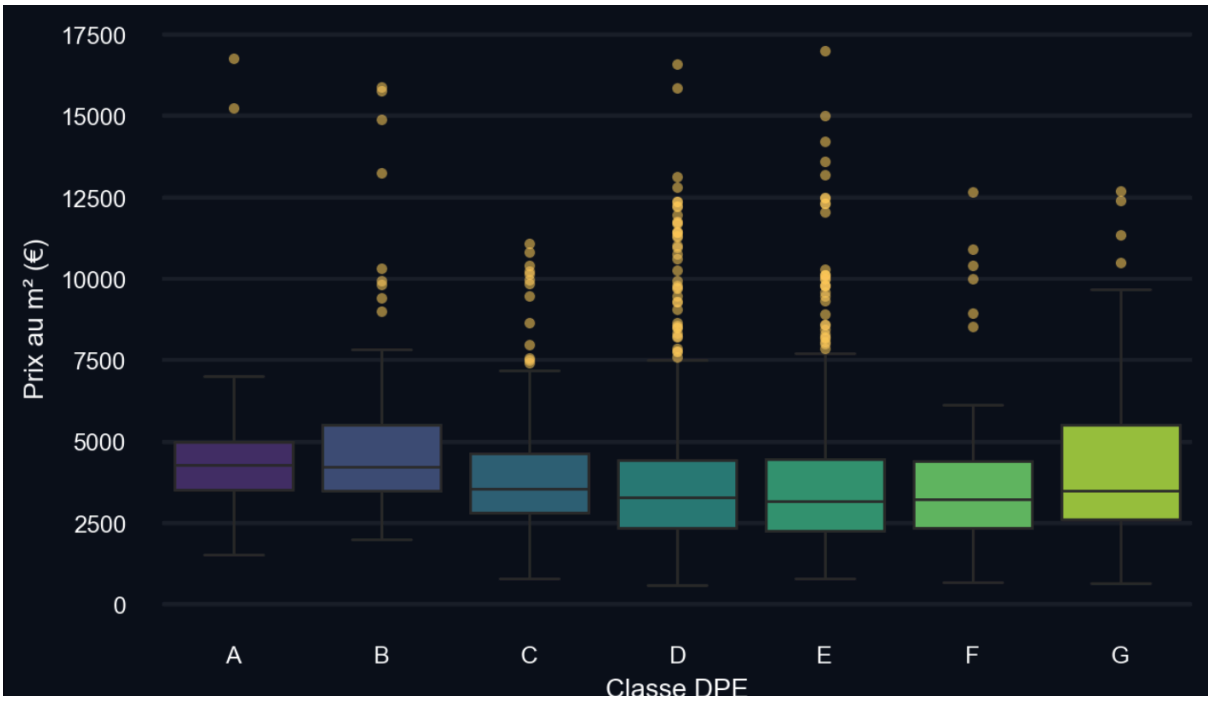
- **Type de bien** : les appartements présentent en moyenne un **prix au m² légèrement plus élevé** que les maisons, ce qui s'explique par leur localisation plus fréquente en centre-ville et dans les zones tendues.

- **Options** (balcon, garage, ascenseur) : la présence de ces équipements est généralement associée à des prix au m² plus élevés, bien que l'effet varie selon la ville et le type de bien.

- **Classe énergétique (DPE)** : les biens mieux classés (A, B, C) tendent à afficher des prix au m² plus élevés, traduisant une valorisation croissante des performances énergétiques sur le marché immobilier.

Ces visualisations permettent ainsi d'identifier des **différences structurelles claires** entre les catégories de biens.





4.4 Corrélation entre surface et prix

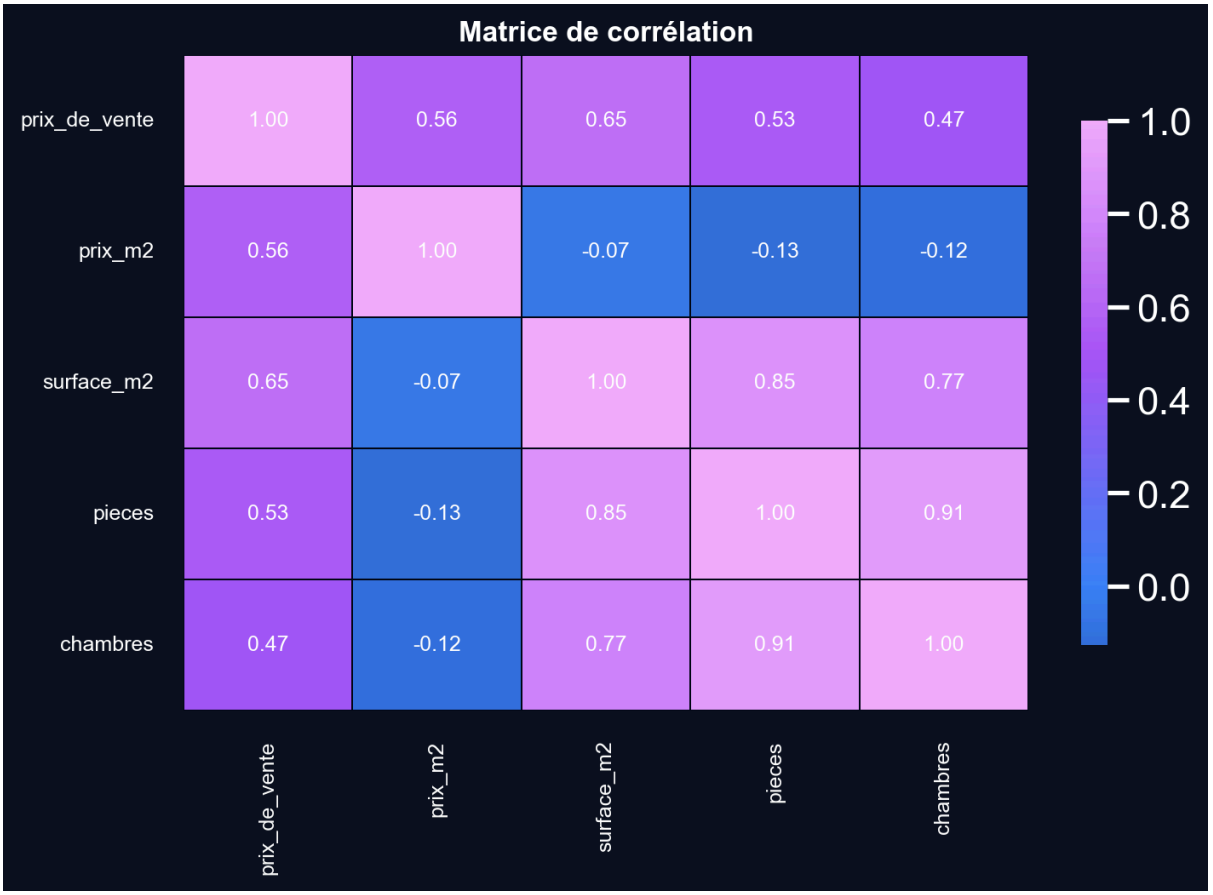
L'analyse des corrélations met en évidence une **relation positive marquée** entre la **surface** et le **prix de vente**. Plus la surface d'un bien augmente, plus son prix tend à être élevé, ce qui constitue un résultat attendu sur le marché immobilier.

Cependant, cette relation n'est pas strictement linéaire. Le **scatterplot**, complété par une **régression linéaire**, montre une dispersion croissante des prix pour les grandes surfaces. Cela suggère que d'autres facteurs jouent un rôle déterminant, notamment :

- le **type de bien** (appartement ou maison),

- la **localisation géographique**,
  et les **caractéristiques qualitatives** du logement.

La matrice de corrélation confirme également une forte corrélation entre la surface, le nombre de pièces et le nombre de chambres, tandis que le prix au m² apparaît moins directement corrélé à la surface. Ces résultats soulignent l'intérêt d'une analyse multivariée pour comprendre finement la formation des prix.



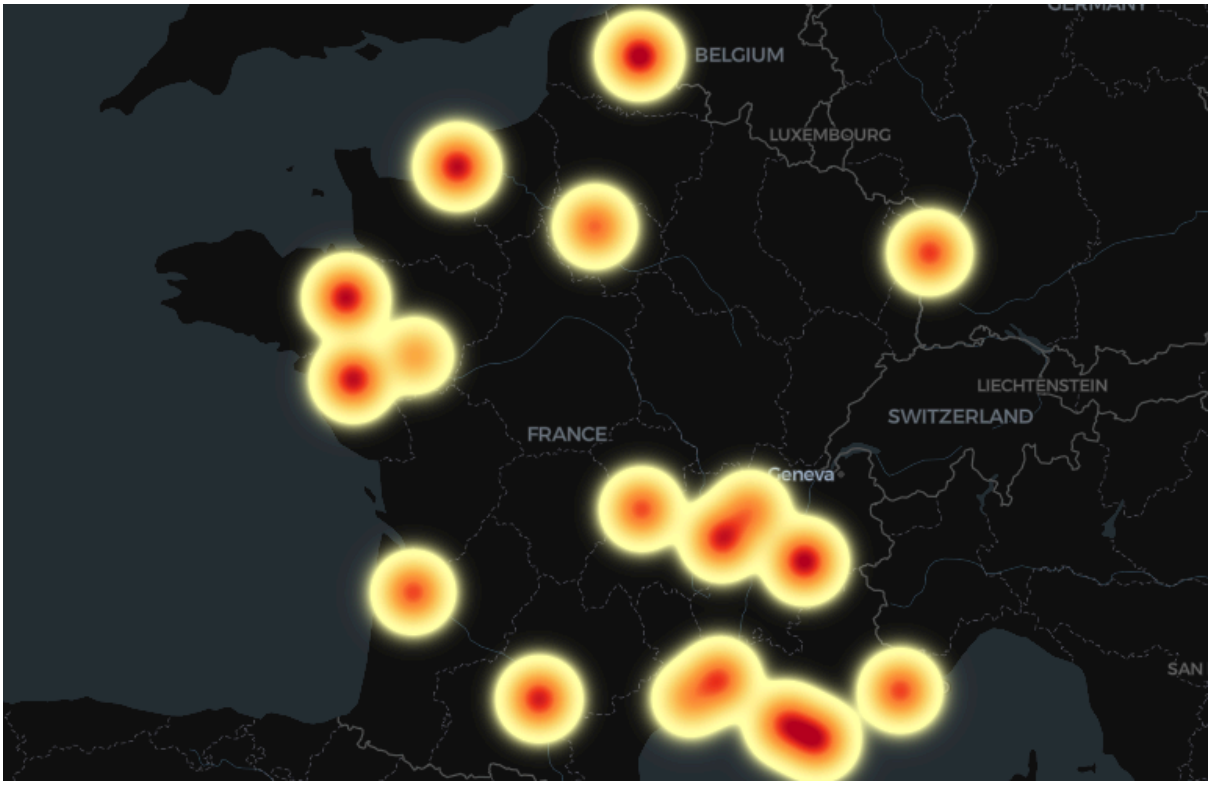
4.5 Analyse géographique

L'analyse géographique repose sur deux approches complémentaires : une **carte de chaleur** et une **comparaison des prix au m² par ville**.

La **heatmap** met en évidence une concentration des annonces et des niveaux de prix élevés autour des grandes agglomérations françaises. Les zones urbaines les plus attractives présentent une densité importante de biens et une pression immobilière marquée.

L'étude des **prix médians au m² par ville** révèle des **disparités géographiques significatives**. Les grandes métropoles, et en particulier Paris, affichent des niveaux de prix nettement supérieurs à ceux des villes de taille intermédiaire. Cette hiérarchie reflète à la fois la rareté du foncier, l'attractivité économique et la forte demande dans ces zones.

En l’absence d’une variable régionale explicite, le choix de se concentrer sur les **grandes villes comme proxy géographique** s’est avéré pertinent dans le cadre de cette étude exploratoire. Cette approche permet de capturer efficacement les grandes tendances spatiales du marché immobilier français.



5. Discussion et limites

5.1 Fiabilité des données et biais potentiels

Plusieurs limites doivent être prises en compte :

- les données proviennent d’un **seul site** ;
- nécessairement les prix de transaction réels ;
- le périmètre géographique est volontairement restreint ;
- Certaines informations peuvent être manquantes ou déclaratives.

Ces biais n’invalident pas l’analyse, mais doivent être intégrés dans l’interprétation des résultats.

5.2 Améliorations possibles

Plusieurs pistes d'amélioration peuvent être envisagées :

- intégrer d'autres plateformes immobilières ;
- étendre l'analyse à une dimension temporelle ;
- enrichir les données avec des sources publiques (INSEE, DVF) ;
- développer des modèles prédictifs de prix.

6. Conclusion

Ce projet a permis de mettre en œuvre l'ensemble des compétences clés du **Data Analytics**, depuis la collecte de données non structurées jusqu'à leur visualisation interactive.

L'analyse montre finalement que le prix au mètre carré dépend fortement de la localisation, du type de bien et de la surface, confirmant les tendances observées sur le marché immobilier français.