Rapport de projet : Python avancé web-scrapping

« There are 3 things that matter in property: location, location, location. »

Lord Harold Samuel (promoteur immobilier anglais). 1956

1. Introduction

Cet adage américain de 1956 illustre le fait que le plus important dans le prix d'un bien immobilier est son emplacement.

De façon générale, le prix d'un bien immobilier est fonction tout d'abord de son emplacement et ensuite de diverses caractéristiques qui vont venir impacter ce prix.

Plusieurs méthodes de valorisation d'un bien immobilier existent et sont couramment utilisées par les professionnels. La plus courante est la méthode hédonique, qui consiste à pondérer par un coefficient la part que représente une caractéristique du bien dans sa valeur finale.

Forme mathématique de la méthode hédonique :

Un modèle de régression linéaire typique

$$\ln(P\_i) = \beta\_0 + \beta\_1 \cdot [\!["Surface"]\!]\_i + \beta\_2 \cdot [\!["Type"]\!]\_i + \beta\_3 \cdot [\!["Localisation"]\!]\_i + \cdots + \beta\_n [caractéristique\ n] + \epsilon\_i$$

- $\ln(P_i)$: prix du bien $i$transformé en log (souvent pour stabiliser la variance)
- β: coefficients estimés
- ϵ_i: erreur aléatoire

D'autres méthodes existent, comme la méthode financière des cash-flows actualisés, la méthode des comparables, ou encore la méthode de la valeur vénale, la méthode par capitalisation.

Nous utilisons ici la méthode hédonique car elle constitue une bonne première approche de la façon dont les acheteurs et les vendeurs réfléchissent pour estimer le prix d'un bien. Cet aspect est le plus pertinent selon nous car nous allons travailler sur un corpus de bien immobiliers en vente, à destination des particuliers principalement.

On retrouve ainsi des articles d'information sur l'évaluation d'un bien en fonction de ses caractéristiques.

Article PAP :

Sur le site Particulier à Particulier (PAP), la recommandation de calcul du prix d'un bien se fait en multipliant la surface du bien habitable (surface Carrez) par le prix au mètre carré moyen du secteur.

En fonction des caractéristiques particulières du bien, sa valeur peut varier de + ou – 20% par rapport à la moyenne du secteur, mais rarement plus.

Si le bien a besoin d'être rafraichi, il faut prévoir une décote de 5/10%. S'il a été récemment rénové, une majoration de 15/20% peut s'appliquer.

L'étage joue un rôle important. Les appartements les plus recherchés se situent entre le 3eme et le 6eme étage, car ils sont réputés plus lumineux et plus calmes que les autres. S'il possède un ascenseur, on peut justifier d'une augmentation de prix de 2.5% par niveau.

Avoir un extérieur est valorisé, un bien avec une terrasse peut se vendre jusqu'à 10% plus cher qu'un bien sans terrasse. D'autres équipements font grimper le prix, comme une piscine, un jardin, un balcon.

Une belle vue a son importance. Les acheteurs préfèreront un bien orienté plein sud, avec une belle luminosité. Mais la surcote ne pourra pas dépasser 5%.

Si le bien a une vue dégagée, cela peut aller jusqu'à 10% de plus. Une vue très remarquable (sur un monument historique important) ou sur une montagne ou autre peut faire grimper la surcote à 15%. Mais elles sont rares.

La classe énergie a un impact : Les logements peuvent avoir une surcote de 10/15% s'ils ont une bonne classe énergétique (A ou B). Par rapport à des « passoires énergétiques » (F ou G).

La qualité du réseau internet peut aussi avoir un impact, du fait du développement du télétravail.

En moyenne, une maison se vendra plus cher qu'un appartement, surtout s'il n'y a pas de mitoyenneté avec d'autres édifices à côté.

url : https://www.pap.fr/vendeur/fixer-prix/quels-atouts-font-augmenter-le-prix-de-vente-de-votre-bien/a23776

On trouvera ainsi de nombreux tableaux d'évaluation de la valeur d'un bien en fonction de ses caractéristiques intrinsèques. Par exemple, ce tableau, trouvé sur le site logic-immo.com.

Facteur	**Description**	**Pondération moyenne**
Localisation	Ville, quartier, attractivité, bassin d'emploi, transports	**40 %**
Surface habitable	m² habitables (loi Carrez)	**25 %**
Type de bien	Appartement / maison / standing	**10 %**
État général	Neuf, bon état, travaux à prévoir	**8 %**
Performance énergétique (DPE)	Classe A → G	**7 %**
Caractéristiques annexes	Balcon, jardin, garage, étage, ascenseur	**5 %**
Contexte de marché	Taux d'intérêt, tension locale, offre/demande	**5 %**
TOTAL		**100 %**

Afin de vérifier la validité de cette méthode d'évaluation, nous allons effectuer une recherche sur un corpus d'annonces immobilières à destination des particuliers. Nous allons analyser et mettre en valeur le prix et les principales caractéristiques qui font évoluer le prix d'un bien immobilier.

Celons les informations que nous avons collectées, les 3 principales caractéristiques qui influencent la valeur d'un bien sont : sa localisation, sa surface et le type de bien dont il s'agit.

La localisation : il s'agit d'une variable discrète et catégorielle. Le bien peut se trouver dans une zone géographique ou une autre. Il n'y a pas de continuité dans cette variable et elle réfère à une caractéristique qui n'est pas quantifiable.

La surface : Il s'agit d'une variable continue et numérique.

Le type de bien : Il s'agit d'une variable discrète et catégorielle.

Le prix : Il s'agit d'une variable numérique et continue.

Notre problématique :

Comment le prix au mètre carré varie-t-il en fonction de la localisation, de la surface et du type de bien immobilier en France ?

Objectif :

Vérifier que ces 3 éléments ont bien un impact sur le prix moyen au mètre carré.

Afin de pourvoir remplir cet objectif, nous allons devoir collecter un certain nombre d'annonces immobilières et analyser ces données pour observer l'impact relatif de chaque élément. Pour cela, nous allons scrapper les annonces immobilières sur le territoire français, sélectionner les caractéristiques les plus pertinentes et faire différentes analyses statistiques permettant de mettre à jour les grandes tendances.

2. Etat de l'art

A/ Présentation du scrapping et des outils python :

Le webscrapping désigne les techniques d'extraction de contenu provenant d'un site internet.

Il s'agit généralement, dans l'acceptation du terme, de créer et d'exécuter des robots permettant de récupérer de l'information automatiquement.

La méthode consiste à envoyer des requêtes vers des sites internet, à récupérer du contenu qui est en grande majorité du HTML, puis à retraiter ces données pour en extraire les informations et structures de données pertinentes.

Le processus de scrapping se déroule en plusieurs étapes. Tout d'abord, les pages web ciblées sont récupérées à l'aide de requêtes HTTP. Ensuite, le code HTML est analysé afin d'identifier les balises contenant les informations d'intérêt. Enfin, les données extraites sont nettoyées, structurées et stockées dans des fichiers exploitables (CSV).

Il peut y avoir plusieurs variations des méthodes employées pour cette activité. Une méthode souvent utilisée se déroule en 2 temps :

1- Utilisation d'un crawler, qui va récupérer toutes les url pertinentes d'un site, cette méthode est appelée plutôt du web crawling.
2- Scrapping des pages récupérées via leurs url pour en extraire l'information pertinente.

C'est la technique que nous allons utiliser ici : le mode de fonctionnement des sites internet et les méthodes de protection anti-bots mises en place nous ont conduits à privilégier cette approche.

B/ Méthode utilisée :

La méthode utilisée dans ce travail est en 2 étapes : tout d'abord récupérer le contenu des pages en html, puis utiliser le package BeautifulSoup4 pour extraire les données intéressantes pour notre étude.

La récupération de données sur les sites se fait en utilisant l'outil playwright et request. Le nettoyage et la sélection dans les pages html avec beautiful soup.

3. Méthodologie

A/ Les sites cibles :

Nous nous sommes concentrés sur le site Logic-immo. Il s'agit d'une marque reconnue dans le secteur de l'immobilier.

url : www.logic-immo.com

Le site était le troisième site d'annonces immobilières en avril 2021 par ordre de fréquentation avec un peu plus de 5 millions de visites mensuelles, derrière seloger.com (8M) et leboncoin immobilier (14M). (Ces chiffres sont issus d'une enquête médiamétrie sur la fréquentation des sites d'annonces immobilières.

https://www.mediametrie.fr/sites/default/files/2021-06/2021%2006%2002%20CP%20Audience%20Internet%20Global%20Avril%202021.pdf

Top 10 de la sous-catégorie Immobilier
Audience Internet Global, avril 2021 - Copyright Médiamétrie//NetRatings

Rang	Brands (B) / Channels (C)	Visiteurs uniques mensuels	Couverture mensuelle (en % des Français 2 ans et plus)
1	**Leboncoin.fr Immo (C)** – T	14 762 000	23,4
2	**SeLoger (B)** – T ACPM/OJD	8 089 000	12,8
3	**Logic-immo.com (B)** – T ACPM/OJD	5 082 000	8,1
4	**Bien ici (B)**	4 382 000	6,9
5	**Particulier a Particulier (B)** – T	3 337 000	5,3
6	**Figaro Immo (C)** – T ACPM/OJD	2 688 000	4,3
7	**MeilleursAgents.com (B)**	2 101 000	3,3
8	**AVendreALouer (B)**	1 861 000	2,9
9	**Orpi (B)**	1 443 000	2,3
10	**Ouestfrance-immo.com (C)** – T ACPM/OJD	1 420 000	2,3

*NB : Les marques participant à la mesure par l'implémentation d'un Tag sur au moins 50 % de leur périmètre sont notifiées «**T**» ; celles dont le périmètre est intégralement taggué et certifié par l'ACPM (Alliance pour les Chiffres de la Presse et des Médias) sont notifiées «**T ACPM/OJD**».*

B/ Technologies utilisées :

Langage : Python, html

Librairies : Pandas, Streamlit, Folium, os, csv, bs4 (beautiful soup), Playwright, random, Time,

Nominatim, Pathlib, geopy, requests

Base de données : CSV

Outils : Git, Github (pour la collaboration et échanges de fichier), Google meet

C/ Stratégie de scrapping :

Le site propose plusieurs possibilités d'accès aux annonces immobilières.

1- Par la barre de recherche lors de l'arrivée sur le site.
2- Par département lorsqu'on descend en bas de la page d'accueil.

Nous avons décidé de tester deux approches de scrapping différentes : par l'utilisation de la barre de recherche et par l'accés aux annonces par département. Ce choix a été motivé d'une part par la volonté de tester plusieurs méthodes de scrapping et de l'autre par la volonté d'avoir des annonces prenant en compte d'une part les grandes villes, de l'autre le reste des départements.

1- Scrapping par la barre de recherche via la ville.

Le scrapping se fait via le package playwright, qui permet de récupérer les données et d'interagir avec le site. La difficulté principale de ce scrapping a été liée au fait que le site dispose d'un système de protection contre les requêtes automatisées, ce qui perturbe la recherche et la collecte d'information via un robot.

Pour éviter ce problème, nous avons découpé la collecte en 2 étapes : collecte des url de pages de recherche par ville, puis enregistrement de la page des annonces via un deuxième code.

Nous avons procédé en plusieurs étapes :

- Récupération manuelle d'une base de données des 100 villes les plus importantes de France en termes de population. Récupération du nom, du code postal et de la taille de population. Fichier csv : « 1-villes_france.csv ».

- Récupération de l'url des annonces de cette ville via un premier code : « 1-recup_url.py ». Ce code va entrer dans la barre de recherche, inscrire le code postal de la ville souhaitée (via 1-villes_france.csv), valider la première localité suggérée par le moteur de recherche du site, valider. Le site nous dirige alors sur la page des résultats de sa recherche et le code enregistre l'url de cette page dans le fichier csv : « 2-liste_url.csv ».

Ce code a posé différents problèmes pour être opérationnel : le site demande tout d'abord de valider l'installation de cookies sur la machine. Le script comprend donc des lignes d'acceptation des cookies.
De plus, lors de l'arrivée sur la barre de recherche, une fenêtre demande de choisir entre : reprendre la recherche précédente ou effectuer une nouvelle recherche. Le script comprend un code pour choisir une nouvelle recherche.
Enfin, le site comprend une sécurité « Datadome » qui bloque les requêtes automatisées. Afin de le contourner, nous avons dû effectuer cette étape en utilisant les données mobiles de notre téléphone portable et non une connexion fixe. En effet, la connexion par téléphone portable change l'adresse IP à chaque requête, ce qui empêche d'être identifié par Datadome.
De plus, l'utilisation de Playwright nous a permis d'utiliser une fonctionnalité de stimulation du comportement d'un être humain, ce qui réduit le risque de détection.

- Récupération de la page html : pour chaque url récupérée par ville dans le fichier « 2-liste_url », un nouveau code : « 2-copie_page_html.py » va entrer l'url dans un navigateur, aller sur la page et enregistrer l'html de cette page dans un fichier .txt . Il va donc enregistrer 100 fichiers text comprenant chacun la page html de l'une des villes. Les pages html enregistrées sont nommées : « page_logic_immo{i}.txt ».

- Récupération des informations sur la page html : pour chaque page html récupérée, le code : « 3-extract_du_html.py » utilise beautiful soup pour extraire les informations importantes (type de bien, prix, pièces, nombre de chambres, surface, étage, adresse, description, agence). Il les enregistre alors dans un fichier : « 3-annonces.csv ».

  Ici, nous n'avons pas eu de problème de protection du site.

- Enfin, pour les besoins du travail, le code « 4-formatage_annonces.py » va récupérer le fichier csv : « 3-annonces.csv » et le mettre en forme pour pouvoir le fusionner avec le csv extrait de la recherche par département. Les données formatées sont stockées dans le fichier « annonces_france-2.csv ».

2- <u>Scrapping par départements :</u>

Dans le cadre de ce projet, les données immobilières ont été collectées par web scrapping à partir du site Logic-Immo, une plateforme spécialisée dans les annonces de vente de biens immobiliers en France. La collecte a débuté par le scrapping de la page suivante : https://www.logic-immo.com/index/departements-vente.
Cette page regroupe l'ensemble des départements français et fournit, pour chacun d'eux, une URL dédiée listant les annonces de biens à vendre dans le département concerné. Ce point d'entrée a été choisi afin de garantir une couverture géographique nationale et homogène dès le début du processus.

Une fois cette page scrappée, l'ensemble des URLs correspondant aux départements a été extrait et stocké. Ces liens ont ensuite servi de base pour la seconde étape du scrapping,

consistant à collecter les annonces immobilières associées à chaque département. Pour chaque URL départementale, seules les annonces de maisons et d'appartements à vendre ont été récupérées, sur les quatre premières pages de résultats.

Le choix de travailler à l'échelle du département, plutôt qu'à celle des villes, répond à un objectif de représentativité des prix immobiliers en France. En effet, un scrapping centré uniquement sur certaines villes, et notamment sur les centres-villes, aurait pu conduire à une surévaluation du prix moyen au mètre carré, ces zones étant généralement plus chères que les zones périphériques ou rurales. L'approche départementale permet ainsi de lisser cet effet et d'obtenir une vision plus globale du marché immobilier français.

Il est toutefois important de noter que, malgré ce choix méthodologique, une part significative des annonces collectées provient des principales villes de chaque département, ce qui reflète la réalité du marché immobilier, où l'offre est majoritairement concentrée dans les zones urbaines. Ce biais reste néanmoins cohérent avec l'objectif du projet, qui vise à analyser les dynamiques réelles du marché immobilier en France.

Cette méthodologie de collecte constitue la base de l'analyse exploratoire menée par la suite et permet d'étudier les prix, les surfaces et les prix au mètre carré à différentes échelles géographiques, notamment nationale, régionale, départementale et locale.

À l'issue de la phase de scrapping, les données issues des différentes collectes ont été regroupées au sein d'un fichier fusionné unique. Cette étape visait à consolider l'ensemble des annonces immobilières dans un seul jeu de données exploitable pour l'analyse. La fusion a été réalisée à partir de deux fichiers CSV distincts, correspondant aux différentes techniques de scrapping.

Afin de garantir l'intégrité du jeu de données final, une attention particulière a été portée à la structure des fichiers, et notamment à l'ordre et au nom des colonnes. Les colonnes ont été harmonisées et réordonnées de manière identique dans les deux fichiers avant la fusion, afin d'éviter toute incohérence ou décalage de données lors de la concaténation. Cette étape était indispensable pour s'assurer que chaque variable corresponde correctement à son attribut et pour prévenir toute altération du dataset final.

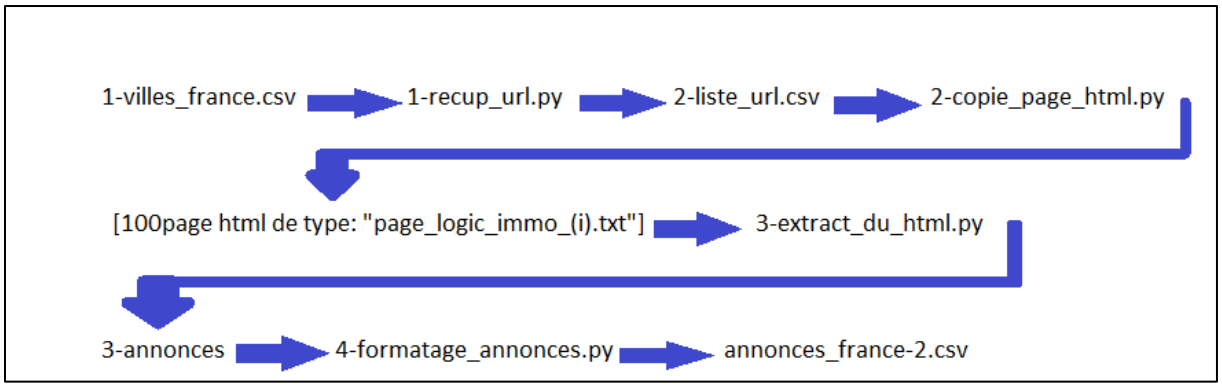
D/ Architecture fonctionnelle :

1/ collecte de données :

A- Par villes :

Nous collectons manuellement, via excel, un fichier csv contenant les 100 plus grandes villes de France en termes de population « 1-villes_france.csv ».

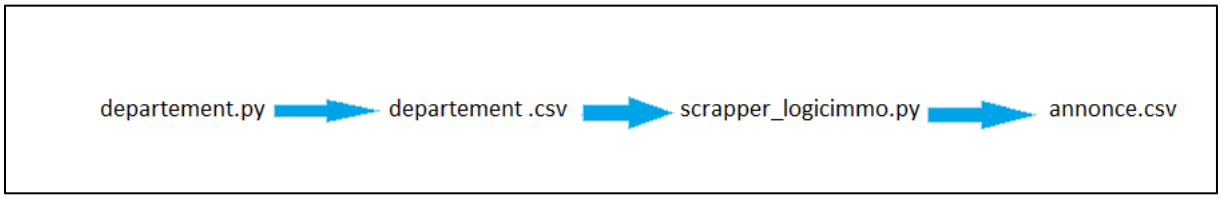
Nous effectuons alors 4 opérations, afin d'obtenir un fichier csv de données pouvant être exploitées « annonces_france-2.csv ».



B- Par département :

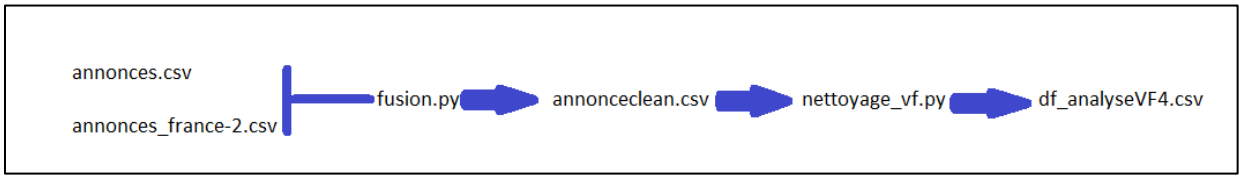
Nous lançons une application « departement.py » qui permet de récupérer les url des annonces par départements. Cette application stocke les url récupérées dans le fichier « departement.csv ».

Nous effectuons alors un scrapping, via « scrapper_logicimmo.py » permettant de créer un csv de donnés exploitables « annonce.csv ».



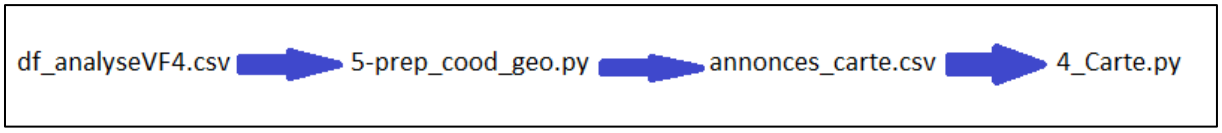
2/ Mise en forme

Nous fusionnons les 2 csv obtenus et les nettoyons pour obtenir le fichier exploitable final « df_analyseVF4.csv ».



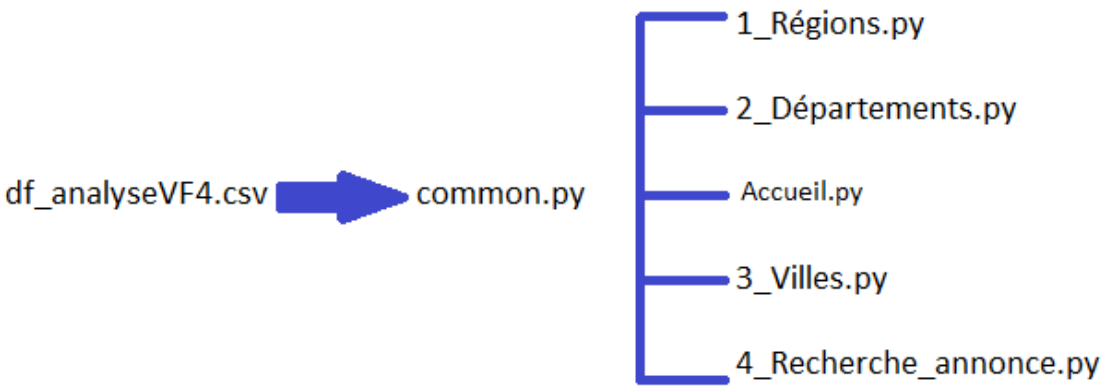
3/ Création de la carte interactive :

A partir de « df_analyseVF4.csv », nous créons un fichier permettant d'obtenir les coordonnés de latitude et longitude « annonces_carte.csv ». Nous pouvons alors créer la carte interactive « 6-visualisation-cartographique ».



4/ Création de l'application interactive

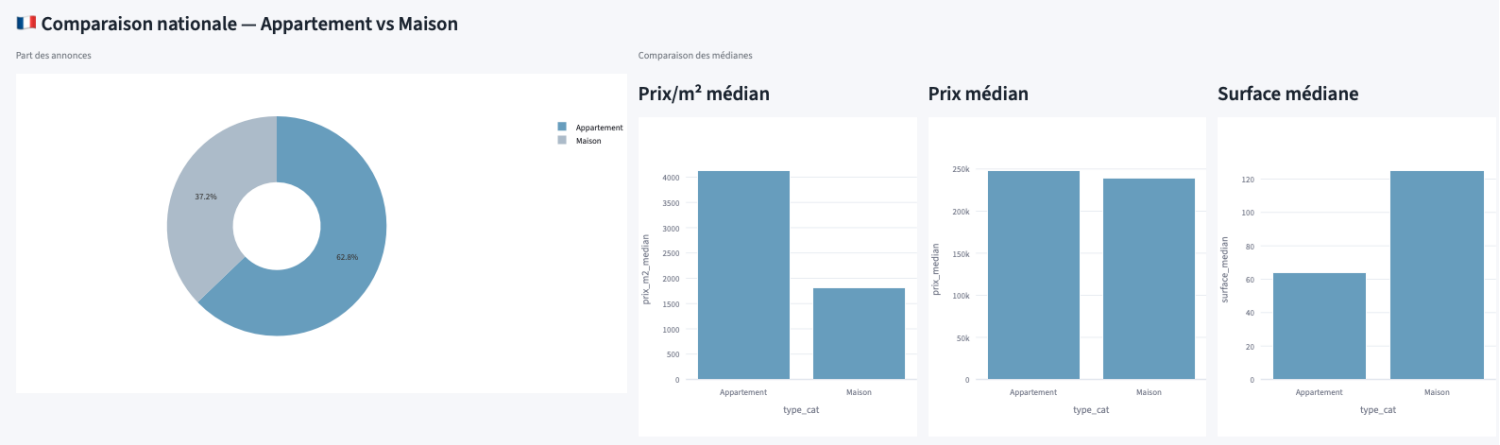
A partir de « df_analyseVF4.csv », nous créons les 2 applications permettant de visualiser les tendances immobilières.



4. Résultats et analyses

Cette partie a pour objectif de répondre à la problématique suivante : comment le prix au mètre carré varie-t-il en fonction de la localisation, de la surface et du type de bien immobilier en France ?

L'analyse repose sur un corpus de 8 274 annonces immobilières, réparties sur 2 928 villes, 103 départements et 18 régions, assurant une couverture géographique large et représentative du territoire français.

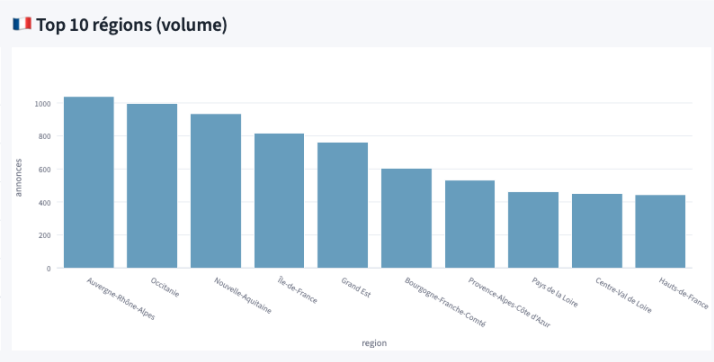


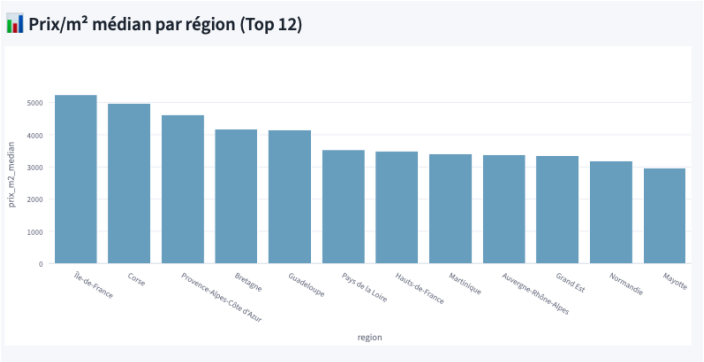
À l'échelle nationale, les indicateurs globaux montrent un prix médian de 245 000 €, une surface médiane de 80 m² et un prix au mètre carré médian de 3 459 €/m². Ces valeurs

traduisent un marché immobilier globalement tendu, mais elles constituent des moyennes qui masquent de fortes disparités spatiales et structurelles, mises en évidence par les analyses suivantes.

La répartition des annonces par type de bien révèle une prédominance nette des appartements, qui représentent environ 63 % des annonces, contre 37 % pour les maisons, les autres types de biens (comme les duplex) restant marginaux. Cette structure reflète la concentration de l'offre immobilière dans les zones urbaines, où le logement collectif est dominant. La comparaison nationale entre appartements et maisons met en évidence des différences structurelles marquées. Les appartements affichent un prix au mètre carré médian significativement plus élevé que les maisons. Ce phénomène s'explique principalement par leur localisation majoritaire dans les centres urbains, où la pression foncière est plus forte. À l'inverse, les maisons présentent des surfaces médianes nettement supérieures, ce qui conduit à des prix totaux médians parfois proches, voire inférieurs, à ceux des appartements. Les boxplots montrent également une dispersion plus importante des prix au mètre carré pour les appartements, traduisant une forte hétérogénéité des localisations, allant des centres-villes très prisés aux zones périphériques. Ainsi, le type de bien apparaît comme un facteur structurant du prix, mais il ne suffit pas à expliquer seul l'ensemble des écarts observés.

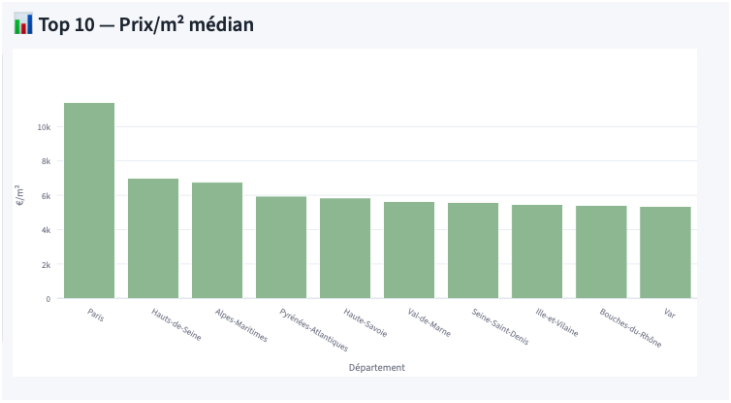
L'analyse de la distribution nationale du prix au mètre carré met en évidence une distribution fortement asymétrique à droite. La majorité des annonces se concentre dans une fourchette comprise entre 2 000 et 5 000 €/m², tandis qu'une longue traîne de valeurs élevées apparaît, avec des prix dépassant 20 000 €/m². Ces valeurs extrêmes correspondent à des marchés très spécifiques, notamment dans les grandes métropoles et les zones dites premium. Cette distribution confirme l'existence de marchés immobiliers locaux très contrastés et souligne le rôle central de la localisation dans la formation des prix.





L'analyse régionale confirme cette influence majeure de la localisation. Certaines régions se distinguent par un volume d'annonces élevé, notamment Auvergne-Rhône-Alpes, Occitanie, Nouvelle-Aquitaine et Île-de-France, qui combinent une forte population, une diversité de territoires et une activité immobilière soutenue. L'étude des prix au mètre carré médians met en évidence une hiérarchie régionale claire. L'Île-de-France se distingue nettement avec les niveaux de prix les plus élevés, dépassant 5 200 €/m², suivie par certaines régions du sud et du littoral, telles que la Corse et la Provence-Alpes-Côte d'Azur. À l'inverse, des régions plus rurales ou industrielles présentent des prix au mètre carré plus modérés. Les boxplots

interrégionaux montrent également une variabilité intra-régionale parfois importante, en particulier dans les régions combinant grandes métropoles et zones rurales, traduisant la coexistence de marchés très hétérogènes au sein d'un même territoire.



À une échelle encore plus fine, l'analyse départementale renforce ces constats. Les départements les plus chers sont majoritairement situés en Île-de-France, notamment Paris, les Hauts-de-Seine, le Val-de-Marne et la Seine-Saint-Denis, ainsi que dans certaines zones littorales ou alpines. Paris se distingue très nettement, avec un prix au mètre carré médian supérieur à 11 000 €/m², largement au-dessus des autres départements. L'analyse des volumes d'annonces par département permet par ailleurs d'identifier les territoires à forte activité immobilière et de distinguer les départements chers mais peu liquides de ceux combinant prix élevés et forte dynamique de marché.

L'analyse à l'échelle des villes met en évidence le rôle structurant des grandes métropoles françaises, telles que Paris, Marseille, Bordeaux, Toulouse ou Strasbourg, qui concentrent naturellement une part importante des annonces. À l'intérieur d'une même ville, les histogrammes et boxplots montrent une forte dispersion des prix au mètre carré, révélant l'existence de marchés immobiliers très hétérogènes. Le cas de Paris illustre particulièrement bien cette fragmentation spatiale. L'analyse par arrondissement met en évidence des écarts de prix très marqués, avec des niveaux particulièrement élevés dans les arrondissements centraux et de l'ouest, où les médianes dépassent fréquemment 15 000 €/m², tandis que les arrondissements périphériques présentent des niveaux plus modérés, généralement compris entre 7 000 et 10 000 €/m². La dispersion observée au sein même des arrondissements souligne l'existence de micro-marchés à une échelle très fine.

Enfin, l'étude des relations entre prix, surface et prix au mètre carré met en évidence des mécanismes économiques cohérents. Les graphiques de dispersion montrent une corrélation positive entre la surface et le prix total, ce qui est attendu d'un point de vue économique. En revanche, la relation entre la surface et le prix au mètre carré est inverse : les petites surfaces présentent généralement un prix au mètre carré plus élevé, tandis que les grandes surfaces

bénéficient d'un effet de décote. Ce phénomène est particulièrement marqué pour les appartements, notamment dans les zones urbaines denses, où la demande pour les petites surfaces est forte.

Dans l'ensemble, ces résultats montrent que le prix au mètre carré en France varie fortement selon la localisation, à toutes les échelles géographiques, mais également en fonction du type de bien et de la surface. La localisation apparaît comme le facteur le plus déterminant, tandis que la surface et le type de bien modulent les niveaux de prix observés. Cette analyse confirme la nécessité d'une approche multi-échelles pour comprendre les dynamiques du marché immobilier français et apporte une réponse structurée et cohérente à la problématique du projet.

5. Discussion et limites

Bien que l'analyse réalisée permette de mettre en évidence des tendances claires du marché immobilier français, plusieurs limites doivent être prises en compte afin d'interpréter correctement les résultats. Les données utilisées proviennent exclusivement d'annonces immobilières en ligne et correspondent à des prix de mise en vente, et non à des prix de transaction réels. Ces prix peuvent donc différer des montants effectivement pratiqués après négociation. Par ailleurs, certaines annonces peuvent contenir des informations incomplètes ou imprécises, ce qui constitue une source potentielle de biais malgré les étapes de nettoyage réalisées.

La géolocalisation des biens repose sur l'utilisation d'une API externe, dont la précision dépend de la qualité des adresses renseignées dans les annonces. Dans certains cas, cette localisation peut être approximative, ce qui limite la précision des analyses spatiales fines, notamment à l'échelle intra-urbaine.

L'analyse est également contrainte par l'absence de certaines variables explicatives clés. En particulier, la date de publication des annonces n'a pas pu être intégrée, ce qui empêche toute analyse temporelle du marché immobilier, comme l'étude de l'évolution des prix dans le temps, la comparaison de différentes périodes économiques ou l'identification de marchés en tension. De même, l'absence d'information sur la durée de publication des annonces limite l'analyse de la liquidité du marché et la distinction entre biens surévalués et biens attractifs.

Il convient également de souligner que les prix utilisés dans cette étude correspondent à des prix de mise en vente et non à des prix de transaction réels. En pratique, le prix affiché dans une annonce est souvent renégocié avant la vente finale, avec des écarts variables selon la localisation et la tension du marché. Les résultats obtenus reflètent donc davantage les attentes des vendeurs que les prix effectivement pratiqués, ce qui constitue une limite à prendre en compte dans l'interprétation des analyses.

D'autres variables importantes, telles que l'année de construction du bien, la présence d'un extérieur (balcon, terrasse ou jardin) ou encore l'étage et la présence d'un ascenseur pour les

appartements, n’ont pas pu être exploitées. Leur intégration aurait permis d’affiner l’analyse des déterminants du prix au mètre carré et de mieux comprendre certaines différences observées entre territoires et types de biens.

Enfin, plusieurs axes d’amélioration peuvent être envisagés, notamment l’élargissement du nombre de sites sources, l’automatisation de la mise à jour des données, l’intégration de données socio-économiques complémentaires et une analyse plus fine à l’échelle des quartiers. Malgré ces limites, le travail réalisé fournit une base solide pour analyser les variations du prix au mètre carré en France et ouvre des perspectives intéressantes pour des analyses futures plus approfondies.

6. Conclusion

L'objectif de ce projet était d'analyser la variation du prix au mètre carré des biens immobiliers en France en fonction de la localisation, de la surface et du type de bien, à partir de données issues d'annonces immobilières en ligne. L'exploitation d'un jeu de données couvrant plus de 8 000 annonces réparties sur l'ensemble du territoire permet d'apporter des éléments de réponse clairs et cohérents à cette problématique.

Les résultats montrent que la localisation est le facteur prédominant dans la formation des prix immobiliers. À toutes les échelles étudiées, des écarts significatifs de prix au mètre carré apparaissent entre territoires, traduisant une forte hétérogénéité spatiale du marché immobilier français. Les régions et zones urbaines les plus attractives présentent des niveaux de prix élevés, tandis que les territoires plus ruraux affichent des prix plus modérés. L'analyse intra-urbaine, notamment dans le cas de Paris, met en évidence des disparités marquées entre quartiers, confirmant l'importance d'une localisation fine dans l'analyse des prix.

Le type de bien influence également de manière significative les niveaux de prix observés. Les appartements présentent en moyenne un prix au mètre carré plus élevé que les maisons, en raison de leur concentration dans les centres urbains. À l'inverse, les maisons se caractérisent par des surfaces plus importantes, conduisant à un prix au mètre carré plus faible malgré des prix totaux parfois comparables. Ces résultats soulignent la nécessité d'analyser conjointement la typologie du bien et son implantation géographique.

La surface du bien joue enfin un rôle structurant dans la formation des prix. Si le prix total augmente logiquement avec la surface, le prix au mètre carré diminue généralement pour les biens de grande taille, traduisant un effet de décote. À l'inverse, les petites surfaces présentent une surcote, particulièrement marquée dans les zones urbaines denses, où la pression de la demande est plus forte.

En définitive, cette étude montre que le prix au mètre carré résulte de l'interaction entre plusieurs facteurs, la localisation demeurant le plus déterminant.
Ce travail ouvre des perspectives d'approfondissement, notamment par l'intégration d'une dimension temporelle permettant d'analyser l'évolution des prix et la tension du marché, ainsi que par l'ajout de variables structurelles et socio-économiques susceptibles d'enrichir la compréhension des dynamiques immobilières et de servir de base à des modèles prédictifs.