UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

# Analyse du marché immobilier

## Comment le prix au mètre carré varie-t-il en fonction du type de bien et de la localisation au sein des grandes métropoles françaises ?

**Rédigé par**

Mahamane Ousmane Maiga Imane

Ngirabanzi Elodie

# Introduction

Le marché immobilier est un domaine complexe où les prix fluctuent constamment en fonction de la localisation, de l'état du bien ou de la conjoncture économique. Pour un acheteur ou un analyste, obtenir une vision claire et instantanée du marché est difficile car l'information est dispersée sur de multiples portails web.

Ce projet a pour ambition de combler ce manque en développant un outil capable de capturer l'état actuel du marché. Notre but est de créer une chaîne de traitement automatisée qui part de la collecte de l'annonce en ligne jusqu'à sa visualisation sur une carte interactive. Concrètement, il s'agit de récupérer des annonces de vente dans dix grandes métropoles françaises, de les nettoyer, et de les rendre intelligibles à travers un tableau de bord.

## 1. État de l'art : Méthodologie de collecte automatisée

Pour récupérer des données sur le web, plusieurs stratégies existent, chacune avec ses avantages et ses inconvénients :

- L'approche par Web Scraping classique (Parsing HTML) est la méthode la plus courante. Elle consiste à écrire un programme qui "lit" le code visuel d'une page web pour y trouver des informations. Bien que populaire (via des outils comme *BeautifulSoup*), cette technique est fragile. Il suffit que le site change la couleur ou la position d'un prix pour que le programme cesse de fonctionner.

- L'exploitation de données structurées (Flux JSON) est la stratégie privilégiée dans ce projet. En contournant l'interface graphique pour accéder directement à la source des données (API), nous récupérons des tableaux informatiques propres et non du texte en vrac. Cette méthode élimine la majorité des erreurs de conversion (ex: confusion entre un code postal et un prix) et garantit une fiabilité maximale des indicateurs statistiques que nous allons calculer par la suite.

## 2. Méthodologie : Architecture du pipeline de données (ETL)

Notre démarche repose sur une chaîne de traitement séquentielle de type **ETL** (*Extract, Transform, Load*), allant de l'interrogation des serveurs jusqu'à la constitution du jeu de données final. L'objectif était de construire un processus robuste, capable de s'adapter aux changements fréquents des données web.

### 2.1. Collecte des données via API

Pour l'acquisition des annonces, nous avons écarté les méthodes de simulation de navigateur (type Selenium), souvent lentes et instables. Nous avons privilégié une approche plus directe en interrogeant l'API du site *Immobilier Notaires* via la librairie Python "*requests*".

**Identification du point d'entrée (Endpoint API)**

Avant de développer le script d'automatisation, une phase d'exploration manuelle a été nécessaire pour localiser la source des données. En utilisant les outils de développement du navigateur (Inspecteur Web), nous avons analysé le trafic réseau dans l'onglet *Network*.

En filtrant les requêtes de type *XHR/Fetch* lors d'une recherche sur le site, nous avons isolé l'appel asynchrone responsable du chargement des annonces. L'analyse de la réponse a confirmé qu'il s'agissait d'un flux JSON structuré, accessible via l'URL "https://www.immobilier.notaires.fr/pub-services/inotr-www-annonces/v1/annonces".

Une fois l'URL identifiée, nous avons développé un script Python capable de l'interroger dynamiquement. Ce dernier repose sur une **double boucle** :

- **Itération géographique :** Le script parcourt étape par étape une liste définie de dix métropoles en associant nom de la ville et code départemental.

- **Gestion de la pagination :** Pour chaque ville, une boucle parcourt les résultats page par page. À chaque tour, le script modifie le point de départ de la recherche (paramètre offset) pour ignorer les annonces déjà récupérées et charger le lot suivant. À chaque appel, les données reçues sont instantanément filtrées (exclusion des biens hors vente et surfaces < 9 m²) avant d'être stockées.

Notre objectif final étant de comparer les métropoles sur une base équitable de **50 annonces exhaustives**, nous avons dû anticiper la perte de données non conformes lors de la phase de nettoyage (parkings, terrains, saisies incomplètes).

Pour garantir l'atteinte de cette cible, nous avons configuré l'arrêt de la boucle de collecte à un volume brut de 80 annonces par ville. Ce surplus agit comme une marge de sécurité. Il permet de compenser le taux de rejet lors du nettoyage ultérieur et assure qu'il restera systématiquement au moins 50 observations exploitables pour constituer l'échantillon final. Une fois ce cycle de collecte terminé pour les dix villes, l'ensemble des données brutes est immédiatement exporté dans un premier fichier de sauvegarde (*data_brute.csv)*.

## 2.2. Normalisation et nettoyage des données (Data Cleaning)

Une fois les données brutes récupérées et stockées en mémoire, l'étape de nettoyage a été réalisée intégralement avec la bibliothèque *pandas*.
Sur le plan méthodologique, nous avons fait le choix de **ne pas utiliser d'expressions régulières (Regex)**. Contrairement au scraping de pages HTML où il faut extraire l'information au milieu de texte brut, nous disposons ici de données déjà semi-structurées (JSON). L'usage des fonctions natives de Pandas s'est avéré plus adéquat :

- **Uniformisation des formats numériques :** La méthode *pd.to_numeric (..., errors='coerce')* a été utilisée pour convertir les colonnes "Prix" et "Surface". Elle transforme automatiquement les erreurs de saisie ou les formats invalides en valeurs nulles (*NaN)* sans bloquer l'exécution du script.

- **Création d'indicateurs :** C'est à cette étape que nous calculons la **variable clé** du projet, le **Prix au m2**, et que nous harmonisons les catégories de biens (Maison ou Appartement).

### 2.3 Projection géographique

Pour permettre l'analyse cartographique, nous avons enrichi le jeu de données en convertissant les adresses postales en coordonnées GPS (latitude, longitude) via l'API *Nominatim* (librairie *geopy*). Afin de respecter les limitations techniques de ce service et d'éviter un bannissement de notre adresse IP, nous avons implémenté un limiteur de débit (Rate Limiter) qui impose une pause d'une seconde entre chaque requête de géocodage. Au terme de ces deux dernières opérations (nettoyage et géolocalisation), un fichier intermédiaire est généré (*data_cleaning.csv*). Ce fichier contient l'ensemble des annonces valides et géolocalisées (environ 80 par ville), prêtes pour la sélection finale.

### 2.4 Constitution de l'échantillon final

Afin de pouvoir comparer les villes de manière juste, il fallait qu'elles aient toutes le même poids dans l'analyse. Comme le nettoyage a laissé des quantités d'annonces inégales d'une ville à l'autre, nous avons harmonisé le tout. Nous avons sélectionné au hasard 50 annonces pour chaque métropole. Cela permet d'éviter qu'une ville ne soit sur-représentée par rapport aux autres. Ce résultat final est enregistré dans le fichier *dataset_final.csv*, qui servira de base unique pour tous nos graphiques.

## 3. Analyse du marché immobilier

Une fois la base de données constituée, on passe ensuite aux statistiques descriptives et à la visualisation des tendances du marché.

### 3.1 Analyse de la répartition des biens par ville

L'analyse de la structure de notre base de données révèle une hétérogénéité marquée dans la distribution géographique et typologique des biens. Pour rappel, les données proviennent d'une extraction aléatoire des 50 premières annonces du site *Notaires_Immobilier*. Cette méthode d'acquisition influe directement sur l'équilibre statistique de l'échantillon.

On distingue notamment des villes où il n'y a aucune annonce concernant les appartements (Rennes), ce qui limite l'analyse de cette ville aux maisons individuelles. Les métropoles de Bordeaux et Lille comptent une seule annonce pour les appartements et 5 pour Toulouse.

| Ville | Appartement | Maison | Total |
|---|---|---|---|
| Bordeaux | 1 | 49 | 50 |
| Lille | 1 | 49 | 50 |
| Lyon | 28 | 22 | 50 |
| Marseille | 23 | 27 | 50 |
| Montpellier | 14 | 36 | 50 |
| Nice | 33 | 17 | 50 |
| Paris | 47 | 3 | 50 |
| Rennes | 0 | 50 | 50 |
| Rouen | 11 | 39 | 50 |
| Toulouse | 5 | 45 | 50 |

## 3.2 Analyse comparative des prix par ville et par type de bien

L'examen des données collectées permet d'établir une hiérarchie claire de l'accessibilité immobilière selon les métropoles. En analysant les prix moyens et médians, nous observons des écarts de valeur considérables qui redéfinissent la notion de pouvoir d'achat immobilier selon la zone géographique.

### Les maisons : Un grand écart entre le Nord et le Sud-Est

Le marché des maisons révèle des contrastes saisissants. Les villes de Lille, Rouen et Toulouse se positionnent comme les plus abordables de notre échantillon. À l'opposé, Paris et Nice affichent des prix extrêmement élevés. Pour comparer le pouvoir d'achat de manière illustrative, un investisseur pourrait acquérir en moyenne sept maisons à Lille pour le prix d'une seule propriété à Nice.

| Ville | Type | Prix_Moyen | Prix_Median | Nombre_Total |
|---|---|---|---|---|
| Lille | Maison | 181022.86 | 150000.0 | 49 |
| Rouen | Maison | 224989.72 | 209000.0 | 39 |
| Toulouse | Maison | 274624.67 | 195000.0 | 45 |
| Lyon | Maison | 389307.64 | 387500.0 | 22 |
| Rennes | Maison | 425906.0 | 363000.0 | 50 |
| Montpellier | Maison | 433551.22 | 382039.0 | 36 |
| Bordeaux | Maison | 573753.65 | 380000.0 | 49 |
| Marseille | Maison | 1001920.78 | 480777.0 | 27 |
| Paris | Maison | 1294666.67 | 1444000.0 | 3 |
| Nice | Maison | 1356873.94 | 802000.0 | 17 |

### Les appartements : Une accessibilité à deux vitesses

Le segment des appartements confirme cette tendance géographique. Si Lille et Toulouse demeurent les villes les plus accessibles avec des budgets moyens inférieurs à 150 k€, le

marché parisien évolue dans une toute autre dimension. Ainsi, Lille et Toulouse sont les points d'entrée les plus aisés pour les ménages souhaitant acquérir une propriété.

La capitale (Paris) demande un effort financier près de 10 fois supérieur (910k € en moyenne) par rapport aux villes les moins chères de l'étude.

| Ville | Type | Prix_Moyen | Prix_Median | Nombre_Total |
|---|---|---|---|---|
| Lille | Appartement | 95000.0 | 95000.0 | 1 |
| Toulouse | Appartement | 147071.6 | 127358.0 | 5 |
| Bordeaux | Appartement | 150000.0 | 150000.0 | 1 |
| Rouen | Appartement | 175158.18 | 150000.0 | 11 |
| Marseille | Appartement | 200787.39 | 155000.0 | 23 |
| Montpellier | Appartement | 220517.36 | 202000.0 | 14 |
| Lyon | Appartement | 235352.79 | 223500.0 | 28 |
| Nice | Appartement | 420935.61 | 300000.0 | 33 |
| Paris | Appartement | 910122.23 | 690000.0 | 47 |

Ce classement met en évidence une France immobilière à plusieurs vitesses. Alors que le centre et le nord (Rouen, Lille) ainsi que le sud-ouest (Toulouse) offrent des opportunités d'achat à des coûts modérés, l'axe Paris-Côte d'Azur (Nice) présente des barrières à l'entrée très élevées, rendant l'accession à la propriété nettement plus complexe pour les ménages standards.

### 3.3 Analyse du prix au mètre carré (m²)

L'analyse du prix moyen au m² est l'indicateur le plus précis pour comparer l'attractivité réelle des territoires, indépendamment de la taille des biens. Les résultats obtenus confirment une polarisation extrême du marché français, où les prix varient de **1 618,78 €/m²** à Lille jusqu'à **11 269,25 €/m²** à Paris.

L'échantillon permet de distinguer clairement trois catégories de villes :

Les villes accessibles ***(< 2 500 €/m²)*** : Lille, Toulouse et Rouen se positionnent comme les secteurs les plus abordables. Ces métropoles offrent des opportunités d'investissement et d'accession à la propriété à des coûts modérés.

Les métropoles régionales établies ***(3 000 à 4 000 €/m²)*** : Ce groupe intermédiaire comprend Bordeaux, Lyon et Marseille. Il représente un marché mature où la demande reste forte mais plus équilibrée que dans la capitale.

Le marché Premium ***(> 6 000 €/m²)*** : Nice et Paris s'extraient totalement des standards nationaux, avec des tarifs reflétant leur statut de zones à forte tension immobilière et touristique.

**Analyse de la distribution : Moyenne vs Médiane**

Un phénomène statistique est observé de manière systématique dans toutes les villes étudiées : la moyenne est systématiquement supérieure à la médiane. Par exemple, à Marseille, le prix moyen s'élève à 3 891 €/m² contre une médiane de 3 465 €/m².

Cet écart est révélateur de la structure des annonces : la présence de quelques biens d'exception avec des prix au m² très élevés, tire la moyenne vers le haut.

| Ville_Recherche | Prix_Moyen_au_m2 | Prix_Median_au_m2 |
|---|---|---|
| Lille | 1618.78 | 1399.64 |
| Toulouse | 1899.53 | 1728.74 |
| Rouen | 2007.4 | 1783.75 |
| Rennes | 3013.63 | 2731.27 |
| Bordeaux | 3047.23 | 2728.12 |
| Montpellier | 3073.22 | 2987.22 |
| Lyon | 3325.92 | 3104.51 |
| Marseille | 3891.14 | 3465.22 |
| Nice | 6210.67 | 5515.14 |
| Paris | 11269.25 | 10746.88 |

## 3.4 Analyse de la corrélation entre la surface et le prix

Au-delà de la localisation géographique, nous avons cherché à mesurer l'influence directe de la dimension physique du bien sur son prix de vente. Pour ce faire, nous avons calculé le coefficient de corrélation de Pearson entre la surface et le prix.

Le coefficient obtenu est de **0,5472**. Ce chiffre nous indique dans un premier temps une corrélation positive : plus la surface d'un logement augmente, plus son prix de vente a tendance à s'élever. Cependant, un score de 0,55 est considéré comme une corrélation modérée. Si le coefficient était proche de 1, nous pourrions conclure que la surface est le seul déterminant du prix.

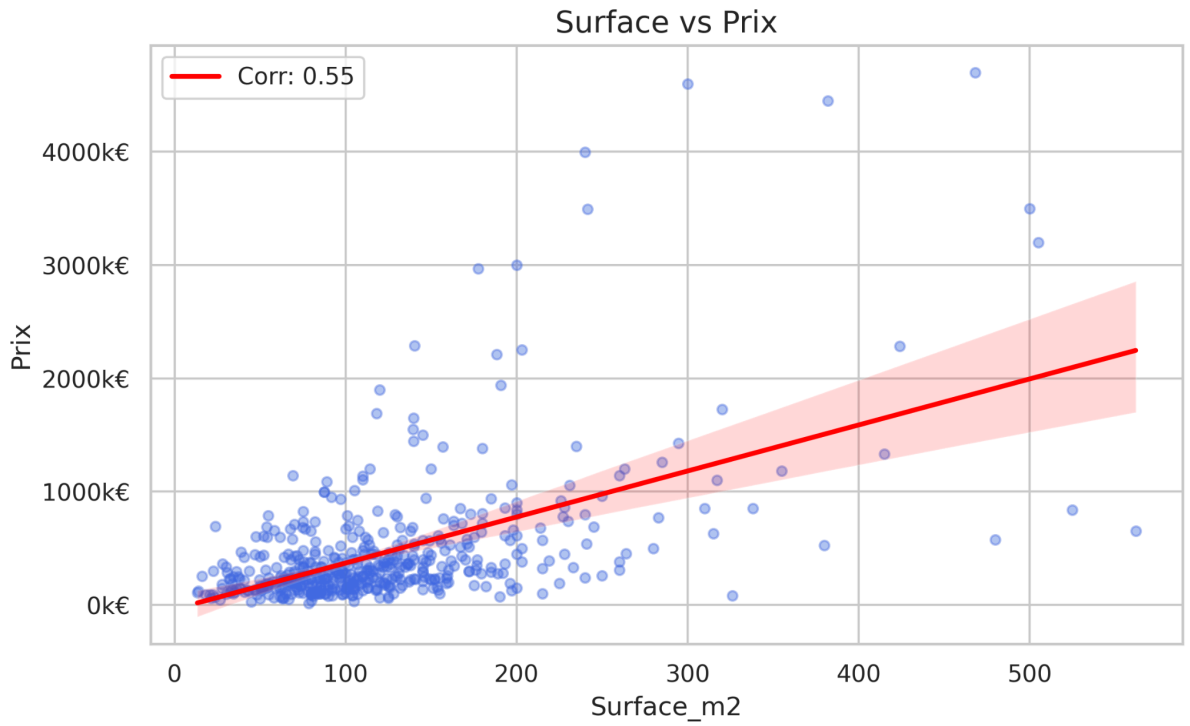
Le fait que ce score ne soit pas plus proche de l'unité indique que la surface n'est qu'une composante de l'équation. D'autres variables, que nous avons observées précédemment, viennent "bruiter" cette relation :

- L'effet "Ville" : Un petit appartement à Paris peut coûter bien plus cher qu'une grande maison à Rouen, ce qui affaiblit la corrélation globale sur l'ensemble du territoire.

- L'état du bien : La présence d'une rénovation récente, d'un jardin ou d'une terrasse peut justifier un prix élevé malgré une surface réduite.

- La typologie : Le prix au mètre carré diffère souvent entre un studio (plus cher au m²) et une grande propriété.

Si la surface reste un facteur prédictif important, elle ne peut expliquer à elle seule la valeur d'un bien immobilier dans notre échantillon.



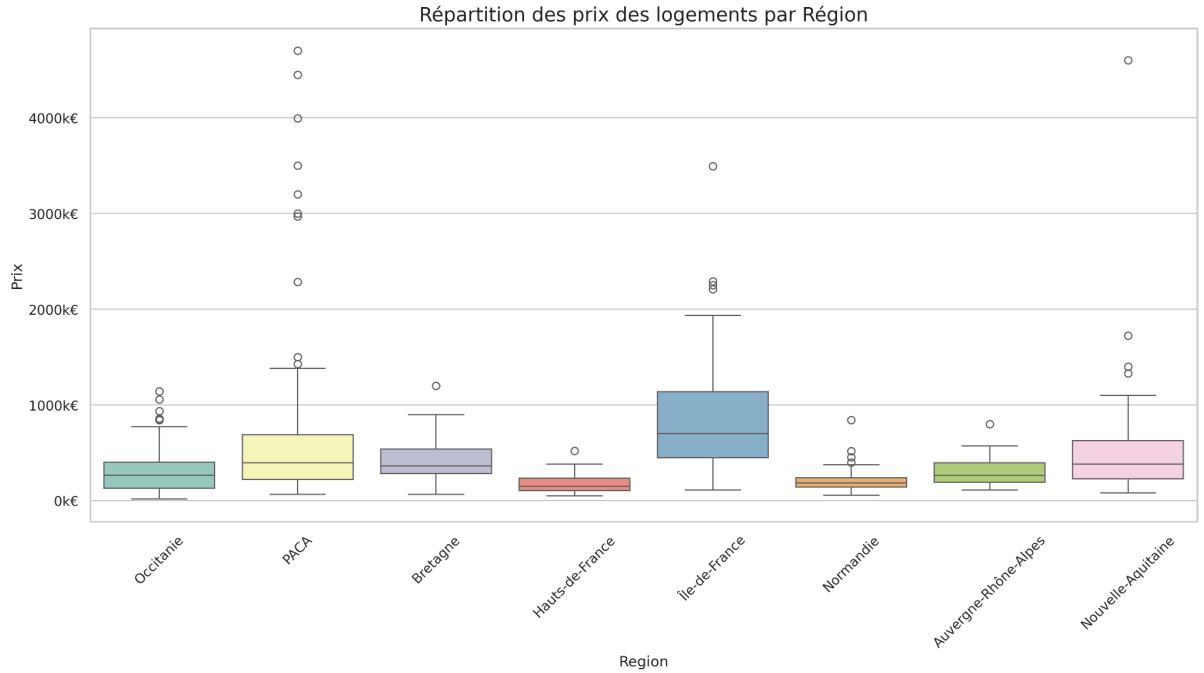

## 3.5 Analyse des tendances observées

Afin d'obtenir une vision macroéconomique de notre échantillon, les données ont été agrégées par grandes régions administratives. L'examen des distributions de prix par région révèle des comportements de marché très hétérogènes.

En Île-de-France et PACA, on remarque que les boîtes à moustaches sont situées plus haut que les autres, ce qui confirme que ces régions sont les plus onéreuses. On voit également des "outliers" situés très hauts, ce qui indique la présence de biens d'exception (plusieurs millions d'euros).

Pour les Hauts-de-France et la Normandie, les boîtes sont plus basses et plus "serrées", indiquant des prix plus abordables et plus homogènes.

Enfin en Nouvelle-Aquitaine, notez le point très élevé tout en haut ; il s'agit d'une valeur extrême (probablement une propriété de luxe à Bordeaux) qui tire la moyenne vers le haut, alors que la médiane reste plus basse.




Cette analyse régionale confirme une fracture territoriale. Le marché se divise entre des régions "homogènes et accessibles" (Nord, Normandie, Bretagne) et des régions "exclusives et hétérogènes" (Paris, Sud-Est). Pour un investisseur ou un acheteur, le risque financier et la barrière à l'entrée sont donc radicalement différents selon la région ciblée, le marché d'Île-de-France exigeant un capital de départ près de cinq fois supérieur à celui de la Normandie pour un bien médian.

# 4. Discussion et limites

## 4.1 Fiabilité des données et biais potentiels

L'analyse menée sur cet échantillon présente plusieurs limites méthodologiques qu'il convient de souligner pour interpréter les résultats avec la prudence nécessaire :

- *Taille de l'échantillon* : Avec seulement 50 annonces collectées, la représentativité statistique est limitée. Ce faible volume explique l'absence de certains types de biens dans certaines villes (par exemple, l'absence d'appartements à Rennes).

- *Biais de sélection* (Extraction aléatoire) : L'extraction des "50 premières annonces" ne garantit pas une répartition homogène. Les résultats reflètent davantage l'état du stock disponible sur le site Notaires Immobilier à un instant t plutôt qu'une réalité structurelle du marché immobilier national.

- *Variables manquantes* : Le prix immobilier est influencé par des facteurs non capturés ici, tels que le Diagnostic de Performance Énergétique (DPE), la présence d'un extérieur (balcon, jardin), la proximité des transports ou l'état général du bien (travaux à prévoir). Ces variables "cachées" expliquent pourquoi la corrélation entre surface et prix (0,55) n'est que modérée.

Le script de scraping développé nécessite un temps d'exécution significatif, imposé notamment par des délais d'attente (sleep timers) nécessaires pour contourner les mécanismes anti-bots et ne pas surcharger le serveur cible. Cette contrainte temporelle rend impossible une collecte en temps réel lors du lancement du Dashboard. C'est pourquoi nous avons opté pour une architecture dissociée : les données sont pré-collectées et stockées dans des fichiers CSV statiques (dataset_final.csv), garantissant ainsi une fluidité immédiate pour l'utilisateur final de l'application Streamlit.

### 4.2 Améliorations possibles

Pour accroître la robustesse de cette étude, plusieurs pistes d'amélioration pourraient être explorées notamment l'élargissement du dataset. En effet, il serait intéressant d'automatiser la collecte sur plusieurs milliers d'annonces et sur plusieurs plateformes (SeLoger, Leboncoin) pour lisser les biais spécifiques à un seul site.

Une analyse temporelle sur plusieurs mois en intégrant la date de publication pour observer l'évolution des prix et identifier des tendances saisonnières. Et enfin créer de nouvelles variables, comme la distance par rapport au centre-ville ou le prix moyen du quartier, pour affiner les modèles de corrélation.

## Conclusion

Ce projet d'analyse de données immobilières a permis de mettre en lumière la complexité et la segmentation du marché français. À travers les différentes étapes de traitement, du nettoyage des données à la visualisation, plusieurs enseignements clés se dégagent :

- La fracture territoriale : Il existe un rapport de 1 à 7 entre les prix des villes les plus accessibles (Lille, Rouen) et les marchés premium (Paris, Nice).

- La prédominance du facteur géographique : Bien que la surface soit un déterminant logique du prix, la localisation reste le levier principal de valorisation d'un bien.

- L'hétérogénéité des marchés : Les écarts entre moyennes et médianes, particulièrement en région PACA et Île-de-France, révèlent des marchés tirés vers le haut par des biens d'exception.

Au-delà de cette analyse descriptive, ce travail pose les bases d'une démarche d'analyse prédictive. Une suite logique serait d'utiliser ces données pour entraîner un modèle d'apprentissage automatique (Machine Learning) capable d'estimer la valeur d'un bien en fonction de ses caractéristiques. À terme, un tel outil pourrait servir d'aide à la décision pour des investisseurs ou des particuliers souhaitant évaluer la cohérence d'un prix de vente par rapport au marché local.