



# Rapport PSID

Projet Durabil'immo

*présenté et soutenu par*

**Florine PHILIPPE**  
**Niraiksan RAVIMOHAN**  
**Fanuel MEHARI**

le 28 avril 2025

# Contents

<b>1</b>	<b>Présentation de l'application</b>	<b>3</b>
1.1	Personas et fonctionnalités . . . . .	3
<b>2</b>	<b><i>Analytics</i></b>	<b>5</b>
2.1	Objectif . . . . .	5
2.2	Présentation du jeu de données . . . . .	6
2.2.1	Présentation des données . . . . .	6
2.2.2	Biais . . . . .	8
2.3	TraITEMENT sur le jeu de données . . . . .	9
2.3.1	Extraction d'informations supplémentaires . . . . .	9
2.3.2	Colonnes modifiées . . . . .	11
2.4	Choix de l'analyse . . . . .	11
2.5	Analyse . . . . .	12
2.5.1	Analyse générale . . . . .	12
2.5.2	Analyse des biens . . . . .	14
2.5.3	Tendances du marché par ville . . . . .	16
2.5.4	Analyse des annonces . . . . .	18
2.5.5	Carte interactive . . . . .	21
<b>3</b>	<b><i>Machine learning</i></b>	<b>29</b>
3.1	Exploration des idées . . . . .	29
3.2	Principe retenu . . . . .	30
3.3	Nettoyage et transformation des données . . . . .	30
3.4	Sélection des modèles de <i>clustering</i> . . . . .	30
3.4.1	Méthode . . . . .	30
3.4.2	Modèles de <i>clustering</i> sélectionnés . . . . .	31
3.5	Entraînement des modèles et prédictions . . . . .	32
3.6	<i>Threats to validity</i> . . . . .	32
3.7	Pistes d'amélioration et ouverture . . . . .	32
3.8	Conclusion . . . . .	33
<b>Webographie</b>		<b>33</b>

# Chapter 1

## Présentation de l'application

Notre application est conçue pour aider les agents immobiliers à se recentrer sur leur cœur de métier en optimisant leur temps de prospection sur le terrain.

Notre application offre un support complet pour collecter et organiser les informations, permettant ainsi aux agents de maximiser leur présence sur le terrain. Inspirée par des cours spécialisés et une analyse approfondie des concurrents, notre solution combine des fonctionnalités éprouvées avec des innovations uniques pour répondre aux besoins non satisfaits des agents immobiliers. Pour cela, notre application propose plusieurs fonctions :

1. Permettre à l'agent immobilier de connaître parfaitement son secteur géographique
2. Anticiper les besoins des clients voulant acheter ou vendre un bien

L'objectif est d'améliorer l'efficacité et l'expérience des agents, en leur fournissant un outil convivial et complet qui les aide à exceller dans leur métier.

### 1.1 Personas et fonctionnalités

Agent immobilier

1. Prospection : Intégration de cartes pour localiser les biens sur une carte interactive.
2. Dashboard : Fournit un tableau de bord avec des graphiques, des courbes et des tableaux pour suivre et gérer l'inventaire des biens, l'activité, et offre des statistiques de marché.
3. Gestion des biens immobiliers : Permet d'ajouter, modifier, supprimer des annonces de biens, télécharger des photos, vidéos, et offre une alerte pour les nouveaux biens.
4. Gestion des clients (prospects) : Permet la gestion de clients potentiels, suit leurs préférences, et offre des rappels et notifications.
5. Gestion des offres et des négociations : Inclut le suivi des offres, des négociations et la documentation des transactions.
6. Gestion de projets potentiels : Enregistre les projets futurs des clients avec des rappels.

7. Agenda et planification : Offre un calendrier intégré avec des tâches, des rappels et des alertes.
8. Documents : Permet le suivi des documents et contrats, la création de supports marketing personnalisés, et la signature électronique.
9. Communication : Inclut des fonctionnalités de partage sur les réseaux sociaux, de messagerie instantanée, et d'envoi de notifications aux clients.
10. Formation et support : Offre des ressources de formation et un support client.
11. Outils de recherche avancée : Propose une recherche par emplacement, prix, type de bien, et d'autres critères pertinents.

# Chapter 2

## *Analytics*

### 2.1 Objectif

Notre objectif est d'analyser les annonces d'un site proposant des ventes de biens immobiliers. Cette analyse permettra à l'agent immobilier de :

#### 1. Connaître le marché local en temps réel

En analysant les annonces, l'agent immobilier peut savoir quels types de biens sont disponibles à quel prix et dans quels quartiers. Cela lui permet de mieux comprendre l'offre actuelle. L'ancienneté des annonces peut aussi être un indicateur précieux pour savoir quels biens sont plus difficiles à vendre.

#### 2. Suivre l'évolution du marché

Certains quartiers peuvent devenir plus populaires avec le temps, ou des zones peuvent subir des changements (nouvelles constructions, rénovations, etc.). Cela donne un aperçu précieux sur où il faut investir ou se concentrer.

#### 3. Déetecter les biens attrayants

Certaines annonces reçoivent plus de *likes* ou d'intérêt. Cela peut aider l'agent immobilier à identifier les biens qui captent l'attention des acheteurs potentiels, ce qui pourrait être un bon signe pour savoir où concentrer ses efforts.

#### 4. Identifier les biens qui peinent à se vendre

Une annonce qui reste en ligne trop longtemps peut signaler un problème de prix, de description, ou d'emplacement. Cela permet à l'agent de mieux comprendre les raisons possibles des difficultés de vente.

#### 5. Observer la concurrence

En analysant les autres agences, l'agent peut voir quelles sont celles qui sont les plus présentes sur le marché et quelles stratégies elles adoptent. Cela permet d'adapter sa propre approche pour rester compétitif.

#### 6. Améliorer ses propres annonces

En observant ce qui fonctionne chez les autres, l'agent immobilier peut affiner ses propres annonces en améliorant des aspects comme les titres, les photos, ou les descriptions pour mieux capter l'attention des acheteurs potentiels.

## 2.2 Présentation du jeu de données

### 2.2.1 Présentation des données

Notre jeu de données regroupe des données d'annonces immobilières. Ces annonces sont extraites du site leboncoin [leboncoin]. Pour récupérer ces données, nous avons effectué du *web scraping* via un script Python. L'extraction des annonces en ligne a eu lieu le 30 mars 2025. Notre choix de scraper les données s'explique par le fait que ce type de données est peu accessible.

L'extraction porte sur les annonces immobilières du département des Hauts-de-Seine. Nous avons sélectionné un seul département car le temps d'extraction serait trop élevé pour une zone plus grande. Finalement, nous avons obtenu 1381 lignes dans notre jeu de données. Ces dernières concernent des annonces publiées en ligne entre novembre 2022 et mars 2025.

Les colonnes de notre fichier CSV, décrites dans le tableau Table 2.1, contiennent des informations sur :

#### 1. L'annonce

Exemples : url, description, boosté ou non (mise en avant payante sur la plateforme), nombre de mises en favori, s'il y a une visite virtuelle jointe à l'annonce...

#### 2. La personne qui a publié l'annonce

Exemples : nom, SIREN, type de vendeur/se (professionnel/le/particulier), type de mandat...

#### 3. Le bien

Exemples : localisation (ville, latitude et longitude), description, prix, nombre de m<sup>2</sup>, nombre de pièces, lien de la visite virtuelle...

Table 2.1: Description du jeu de données : nom, type et description des colonnes

Début de table		
Colonne	Type	Description
list_id	entier	Identifiant de l'annonce
url	chaîne de caractères	URL de l'annonce sur le site leboncoin
price	réel	Prix fixé par le/la vendeur/se
body	chaîne de caractères	Description de l'annonce
subject	chaîne de caractères	Titre de l'annonce
first_publication_date	date	Date de publication de l'annonce
index_date	date	Date de la dernière modification de l'annonce
status	étiquette ('active')	Présence de l'annonce en ligne ('active' : présente en ligne)
nb_images	entier	Nombre d'images dans l'annonce
country_id	étiquette ('FR')	Identifiant du pays du bien
region_id	entier	Identifiant de la région du bien
region_name	chaîne de caractères	Nom de la région du bien
department_id	entier	Numéro du département du bien
city	chaîne de caractères	Nom de la ville du bien
zipcode	entier	Code postal du bien
lat	réel	Latitude de l'adresse du bien
lng	réel	Longitude de l'adresse du bien
type	étiquette ('pro' ou 'private')	Type de vendeur/se (professionnel/le ou particulier)
name	chaîne de caractères	Nom du/de la vendeur/se
siren	entier	Numéro SIREN de l'agent ou de l'agence immobilière
has_phone	booléen	Présence d'un numéro de téléphone dans l'annonce
is_boosted	booléen	Vrai si l'annonce est boostée (mise en avant payante sur la plateforme), faux sinon
favorites	entier	Nombre de mises en favori de l'annonce
square	réel	Superficie en mètres carrés du bien
land_plot_surface	réel	Superficie en mètres carrés du terrain du bien
rooms	entier	Nombre de pièces du bien
bedrooms	entier	Nombre de chambres du bien
nb_bathrooms	entier	Nombre de salles de bain du bien
nb_shower_room	entier	Nombre de salles d'eau du bien

Fin de la table 2.1		
Colonne	Type	Description
energy_rate	étiquette (de 'a' à 'g')	Classement du bien suite à la partie Énergie du diagnostic de performance énergétique ('a' : extrêmement performant à 'g' : extrêmement peu performant)
ges	étiquette (de 'a' à 'g')	Classement du bien suite à la partie Climat du diagnostic de performance énergétique ('a' : peu d'émission de gaz à effet de serre à 'g' : émission très importante)
heating_type	étiquette ('communal' ou 'individual')	Type de chauffage du bien
heating_mode	étiquette ('electric', 'fuel', 'gas', 'solar' ou 'other')	Mode de chauffage de bien
elevator	booléen	Nombre d'ascenseurs du bien
fees_at_the_expanse_of	étiquette ('buyer', 'seller' ou 'buyer_and_seller')	Destinataire des frais
fai_included	booléen	Vrai si les frais d'agence sont inclus, faux sinon
mandate_type	étiquette ('simple' ou 'exclusive')	Type de mandat de l'agent immobilier
price_per_square_meter	réel	Prix par mètres carrés du bien
immo_sell_type	étiquette ('new' ou 'old')	Ancienneté du bien
nb_floors	entier	Nombre d'étages du bien
nb_parkings	entier	Nombre de parkings du bien
building_year	entier	Date de construction du bien
virtual_tour	chaîne de caractères	URL de la visite virtuelle du bien
old_price	réel	Ancien prix du bien
annual_charges	réel	Charges annuelles du bien
orientation	étiquette ('north', 'west', 'east', 'south', 'north-west', 'north-east', 'south-west', 'south-east')	Orientation du bien
is_virtual_tour	booléen	Vrai si l'annonce contient une visite virtuelle, faux sinon

## 2.2.2 Biais

Cependant, l'analyse de notre jeu de données peut provoquer des biais. Les biais sont :

1. Les annonces récupérées n'étant que celles du site leboncoin, elles ne représentent pas réellement l'activité immobilière de la zone sélectionnée. En plus, les données sont sur une période restreinte (2022 à 2025) et donc pas représentative de la tendance du marché sur le long terme.
2. Les annonces du jeu de données correspondent à des biens pas encore vendus. Après l'achat d'un bien, son annonce n'est plus présente sur leboncoin. Juger le succès d'une annonce de manière précise s'avère donc difficile puisque les comparaisons entre les annonces de biens vendus et non vendus est impossible.
3. Compte tenu de la lenteur de l'extraction des données, le jeu de données contient 1381 annonces. Le nombre peu élevé d'annonces permet difficilement d'observer une tendance fiable.
4. Le prix proposé dans une annonce est le prix fixé par le/la vendeur/se. Ce n'est pas forcément le prix de vente.

Notre analyse a été faite en prenant en compte la présence de ces biais.

## 2.3 Traitement sur le jeu de données

### 2.3.1 Extraction d'informations supplémentaires

En observant nos données, nous avons remarqué que :

1. Plusieurs annonces avaient des valeurs manquantes dans certaines colonnes mais ces valeurs pouvaient être présentes dans la description de l'annonce (colonne body)
2. Plusieurs informations importantes étaient présentes dans la description de l'annonce mais ne correspondaient à aucune colonne du jeu de données. Par exemple, la description, décrite dans la Table 2.3.1, indique la présence de transports en commun ("RER A à 5mn à pied") ou d'écoles ("proximité des écoles") à proximité du bien. Nous avons donc décidé d'ajouter des colonnes correspondant à ce type de données dans notre fichier CSV. Ces dernières sont décrites dans la Table 2.3.1.

Pour extraire ces données, la solution a été d'utiliser un LLM, auquel on enverrait la description et qui nous renverrait les informations que nous recherchons. Nous avons donc écrit un programme Python réalisant cela. Nous avons utilisé la librairie ScrapeGraphAI [ScrapeGraphAI] qui permet de demander une requête, qui concerne une source de données que l'on passe en paramètre, à un LLM. Dans notre cas, la source de données est la description du bien. Dans notre code, nous allons boucler sur tout le fichier CSV et pour chaque annonce :

1. Nous récupérons les noms des colonnes ayant des valeurs manquantes
2. Nous envoyons le *prompt* au LLM. Ce *prompt* contient :
  - (a) Les noms des colonnes ayant des valeurs manquantes
  - (b) Le format de réponse à respecter (format JSON)

Exemple de description
<p>Triangle d'Or Rueil-Malmaison, RER A à 5mn à pied, proximité des écoles, bords de seine et au calme absolu ! Venez découvrir notre agréable maison de famille de 134m<sup>2</sup> au sol, dont 83m<sup>2</sup> habitables. Distribuée sur 3 niveaux elle offre 3 chambres, un garage aménageable de 32m<sup>2</sup> et un petit jardin. Poussez la porte et c'est un espace de vie lumineux et modulable selon vos envies, qui vous offre un premier plateau de 32m<sup>2</sup>, actuellement aménagé en un séjour traversant de 22m<sup>2</sup>, une entrée de 5m<sup>2</sup> et une cuisine équipée de 6m<sup>2</sup>. Un wc invité et un accès au garage complètent ce niveau.</p> <p>Les espaces nuits se situent en étages avec au premier niveau, 2 belles chambres, un dressing et une salle de bain avec wc. Le niveau supérieur offre une troisième chambre avec divers rangements sur le palier. Le bien propose également un grand garage de 32m<sup>2</sup> aménageable selon vos projets, comprenant le local technique, un point d'eau aménagé en salle de douche et un accès au jardin de 16m<sup>2</sup>. SES ATOUTS : quartier très prisé et calme. Toutes les commodités accessibles à 5mn à pied, une maison très bien entretenue où l'on se sent bien, parquet impeccable, toiture et façades en très bon état, double vitrage, possibilité d'aménagement à votre goût. N'hésitez plus, venez la visiter !</p>

Table 2.2: Description d'un bien : exemple d'une valeur de la colonne body du jeu de données

- (c) Une description des formats attendus pour certaines valeurs (exemple : la colonne energy\_rate ne peut avoir que les caractères allant de 'a' à 'g' comme valeurs)
3. Nous ajoutons, dans notre fichier, les valeurs récupérées via le LLM

Pour le choix du LLM, nous avons essayé plusieurs méthodes :

## 1. Local

Nous avons installé en local, via le logiciel ollama [Ollama], plusieurs LLMs (llama3.2:1b, phi3:mini, gemma1.1:2b et mistral0.3:7b). Cependant, les informations extraites étaient insuffisantes ou inexactes et certaines informations de la description n'étaient pas récupérées. De plus, une requête prenait, en général, plusieurs minutes à s'exécuter, ce qui est lent puisqu'il faudrait exécuter 1381 requêtes (une requête par ligne).

## 2. Distant

Nous avons utilisé l'API Mistral [Mistral AI], qui est gratuite, et le modèle open-mistral-nemo 24.07 [Mistral Nemo], car ce dernier est performant dans un contexte multilingue et notamment en français. Des points de vue de la qualité des réponses (assez d'informations sont récupérées et les informations récupérées sont exactes) et de la rapidité d'exécution (une requête s'exécute en général en quelques secondes), cette méthode est plus performante que l'utilisation de modèles en

Colonne	Type	Description
transport_exists_nearby	booléen	Vrai si des transports en commun sont à proximité du bien, faux sinon
school_exists_nearby	booléen	Vrai si des écoles sont à proximité du bien, faux sinon
medical_service_exists_nearby	booléen	Vrai si des services médicaux sont à proximité du bien, faux sinon
centre_of_town_exists_nearby	booléen	Vrai si le centre-ville est à proximité du bien, faux sinon
nb_square_meter_basement	réel	Superficie en mètres carrés du sous-sol
nb_square_meter_balcony	réel	Superficie en mètres carrés de la terrasse

Table 2.3: Description des colonnes ajoutées dans le jeu de données : nom, type et description des colonnes

local. Nous avons donc utilisé l'API Mistral et le modèle open-mistral-nemo 24.07.

### 2.3.2 Colonnes modifiées

Nous avons également effectué d'autres modifications :

1. À la suite de l'extraction de données via le LLM, certaines données n'étaient pas dans le bon format. Par exemple, pour la colonne land\_plot\_surface, les valeurs attendues doivent être des nombres réels. Cependant, il pouvait arriver que le LLM retourne un nombre réel suivi des caractères "m<sup>2</sup>". Dans ces cas-là, nous avons simplement retiré les caractères en trop pour conserver seulement le nombre.
2. Une colonne contenant des valeurs booléennes (colonne fai\_included) représentait la valeur "vrai" par le chiffre 1 et la valeur "faux" par le chiffre 2. Pour faciliter l'analyse, nous avons remplacé les valeurs pour les convertir dans un format 1 pour "vrai" et 0 pour "faux".
3. Les annonces de viagers ont été supprimées car elles concernent un autre marché que celui que l'on souhaite analyser. Les conserver fausserait les prix et donc notre analyse.
4. Une colonne inutilisée (colonne elevator) a été supprimée.

## 2.4 Choix de l'analyse

Nous avons eu plusieurs idées d'analyse :

1. Analyse des biens uniquement
2. Analyse des annonces

Nous avons choisi d'effectuer une analyse des annonces puisqu'après une étude de notre jeu de données, nous avons réalisé que l'analyse centrée exclusivement sur les biens n'était pas adaptée à nos données. L'analyse des annonces en elles-mêmes plutôt que sur les biens s'est donc avérée être plus appropriée : nous avons estimé que des informations pertinentes pour l'agent immobilier pouvaient en être tirées.

## 2.5 Analyse

### 2.5.1 Analyse générale

Dans cette section, nous allons effectuer une analyse générale des annonces disponibles.

#### Répartition des annonces par type de vente

Ce graphique (Figure 2.1) représente la répartition des annonces par type de vente.

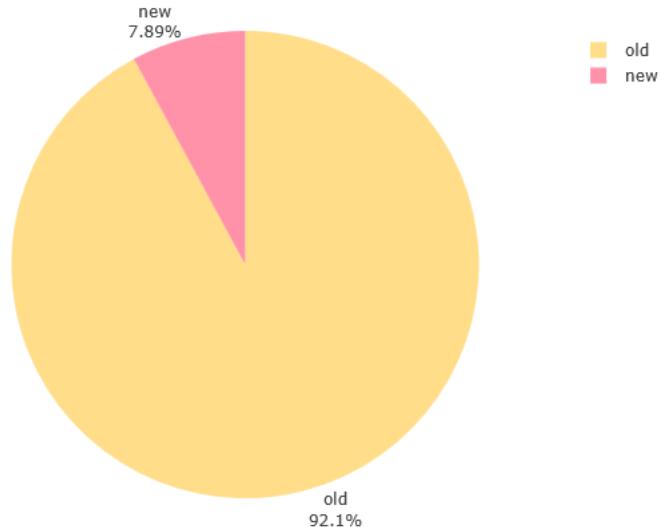


Figure 2.1: Répartition des annonces par type de vente

#### Utilité du graphique :

Ce graphique permet à l'agent immobilier de visualiser le type de biens (ancien (*old*) ou neuf (*new*)) le plus présent sur le marché.

#### Observation :

Ici, les biens anciens représentent 92 % des annonces, contre seulement 8 % pour les biens neufs. Cela traduit une prédominance du marché ancien et peut aussi traduire un faible volume de construction dans le secteur. D'autre part, les biens neufs, plus rares, peuvent constituer un segment de différenciation à exploiter.

### Répartition des annonces par type de vendeur/se

Ce graphique (Figure 2.2) représente la répartition des annonces selon le type de vendeur/se (professionnel/le ou particulier).

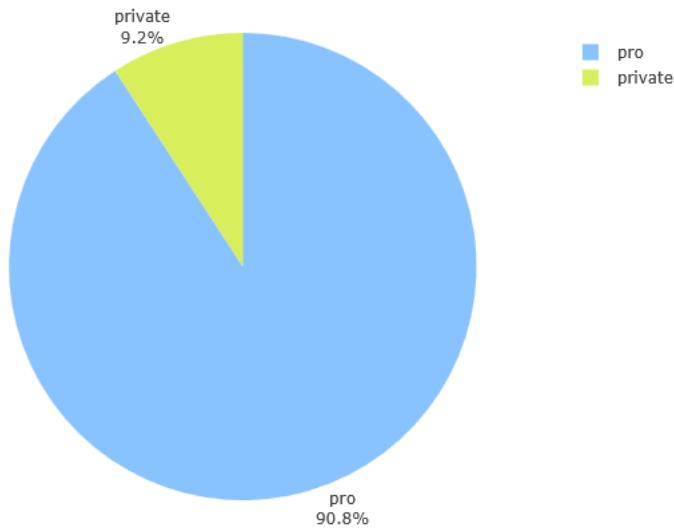


Figure 2.2: Répartition des annonces par type de vendeur/se

#### Utilité du graphique :

Il offre une vision claire de la concurrence et des tendances du marché.

#### Observation :

Ici, 90 % des annonces sont publiées par des professionnel/les et seulement 10 % par des particuliers. Cela indique que les professionnel/les contrôlent largement le marché, ce qui peut signifier une concurrence accrue entre agents immobiliers. Cela suggère qu'il est essentiel de se différencier.

### Dates de publication des annonces

Ce graphique (Figure 2.3) place dans le temps les dates de publication des annonces.

#### Utilité du graphique :

Il permet à l'agent immobilier de savoir si une annonce est présente en ligne depuis longtemps ou non. Si c'est le cas, nous pouvons estimer que l'offre n'est pas assez convaincante pour de futur/es acheteurs/ses et nécessite donc d'être améliorée.

#### Observation :

La plus ancienne annonce a été publiée le 11 novembre 2022 alors que la plus récente l'a été le 28 mars 2025. La date médiane est le 26 février 2025 et on remarque que la plupart des annonces sont récentes par rapport à la date d'extraction puisque 50

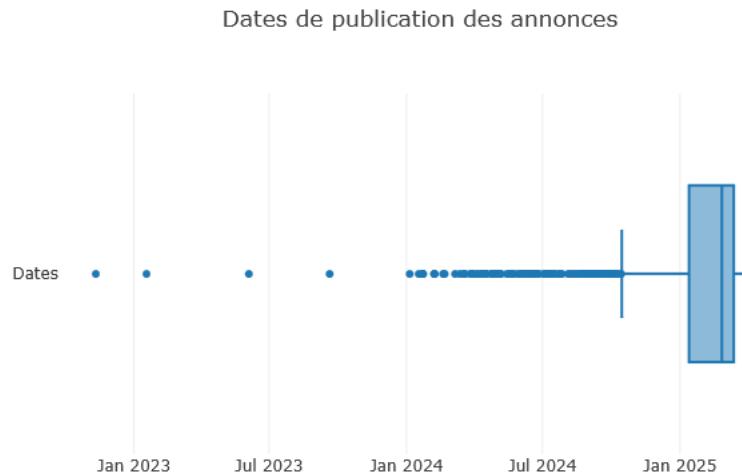


Figure 2.3: Dates de publication des annonces

% de celles-ci sont apparues entre le 13 janvier et le 14 mars 2025. De plus, des valeurs extrêmes sont présentes. En effet, seulement 4 annonces furent publiées avant 2024. Malgré leur durée d'accessibilité en ligne plus élevée, celles-ci n'ont pas trouvé preneur/se.

## 2.5.2 Analyse des biens

Dans cette section, nous allons effectuer une analyse des biens.

### Regroupement des biens similaires

Ce *biplot* (Figure 2.4) représente des groupes similaires de biens. Nous avons effectué une ACP sur les variables et une classification *K-means*. Nous avons conservé les dimensions 1 et 2 de l'ACP :

1. Dim1 est liée aux variables square, room et bedrooms et reflète les caractéristiques physiques des biens.
2. Dim2 est basée sur energy\_rate et ges et reflète la performance énergétique des biens.

### Utilité du graphique :

Ce *biplot* permet de segmenter les biens immobiliers selon leurs caractéristiques principales, en identifiant des groupes similaires.

### Observations :

1. Le cluster 0 regroupe des grands biens avec de nombreuses pièces mais une performance énergétique variable. Le cluster 1 contient des biens plus petits avec

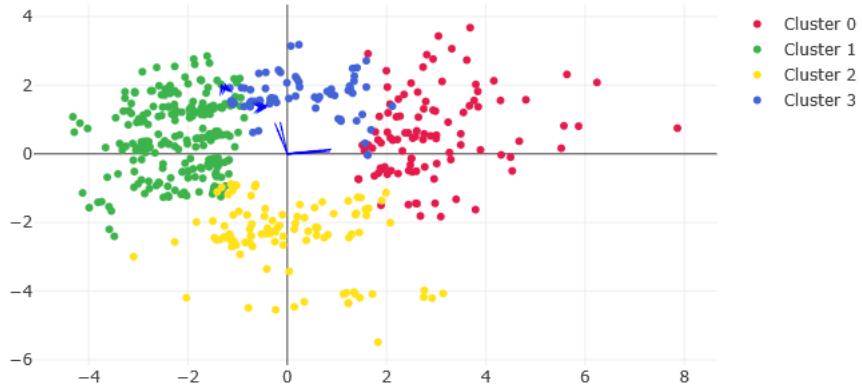


Figure 2.4: Regroupement des biens similaires

moins de pièces et une performance énergétique variée. Le cluster 3 inclut des biens avec une haute performance énergétique, bien que leur taille varie. Enfin, le cluster 2 correspond à des biens de taille variable et de faible performance énergétique.

2. Dans le graphique de répartition des *clusters* (Figure 2.5), nous pouvons observer que la majorité des biens (45 %) est de petite taille avec une performance énergétique variable, ce qui correspond à une forte demande pour des biens abordables. 21 % des biens ont une faible consommation énergétique et une taille variable, ce qui correspond à un marché de niche pour les clients soucieux de l'environnement. 20 % des biens sont grands mais avec une performance énergétique variable, ce qui peut attirer les familles, mais nécessite de travailler sur leur efficacité énergétique. Enfin, 13 % des biens consomment beaucoup d'énergie, ce qui peut rendre leur vente plus difficile.

### Années de construction par ville

#### Utilité du graphique :

Ce graphique (Figure 2.6) permet d'observer, pour chaque ville, si les biens à vendre sont plutôt anciens ou récents.

#### Observation :

Dans la majorité des villes, la plupart des maisons en vente ont été construites entre 1900 et 2000. Certaines communes, comme Bagneux, Colombes ou Nanterre, présentent également des biens plus anciens, datant parfois de 1800 à 1850. On observe aussi, dans la plupart des villes, la présence de biens construits après 2000, mais en très faible proportion. Cela indique que le marché immobilier est principalement composé de logements anciens, ce qui peut refléter un parc immobilier relativement âgé, avec

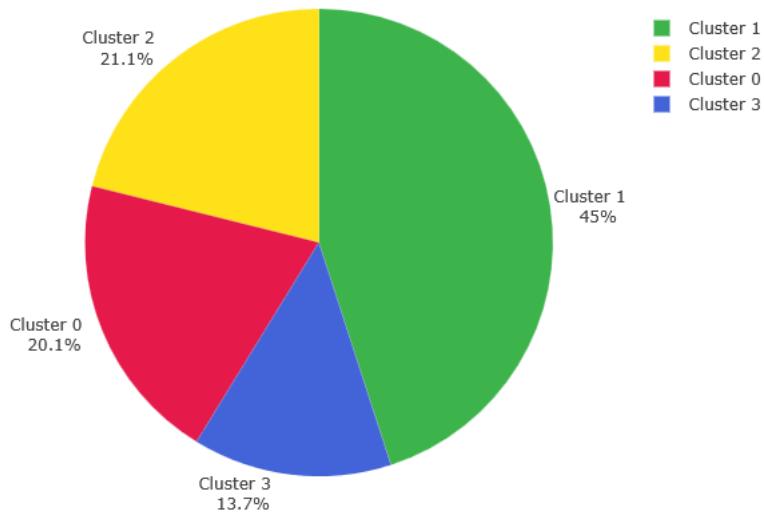


Figure 2.5: Répartition des *clusters*

potentiellement des besoins en rénovation ou en mise aux normes. La faible présence de biens récents suggère un rythme de construction neuve peu soutenu dans ces zones, ce qui peut impacter l'offre disponible pour les acheteurs recherchant des logements modernes.

### Corrélation entre le prix annoncé et les autres variables

#### Utilité du graphique :

Ce graphique (Figure 2.7) permet de comprendre comment les différentes caractéristiques d'un bien influencent le prix d'annonce.

#### Observation :

Les variables les plus corrélées avec le prix d'annonce sont bedrooms, room et square, avec une corrélation positive. Cela signifie que le prix augmente à mesure que la taille de la maison croît. Cependant, cela paraît étrange, car d'autres variables semblent moins corrélées avec le prix d'annonce. Cela peut être normal, ou pourrait indiquer que les prix fixés dans les annonces ne sont pas représentatifs. En effet, le prix devrait normalement être influencé par d'autres facteurs tels que le quartier, le score de consommation énergétique, le score des services, etc.

### 2.5.3 Tendances du marché par ville

Dans cette section, nous allons effectuer une analyse des tendances du marché par ville.

#### Nombre d'annonces par ville

#### Utilité du graphique :

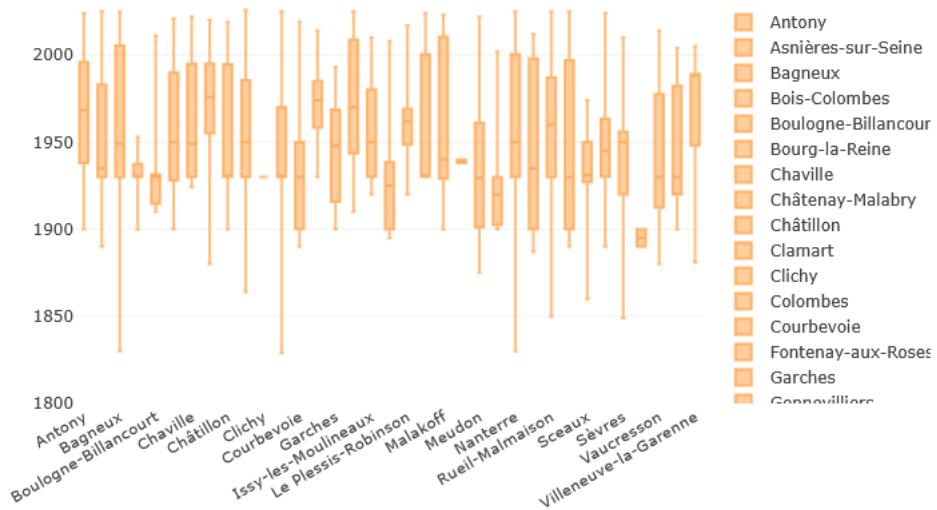


Figure 2.6: Années de construction par ville

Ce graphique (Figure 2.11) permet d'identifier les villes où l'offre immobilière est la plus concentrée, celles qui sont les plus actives en termes de diffusion d'annonces. Il aide ainsi à repérer les zones dynamiques du marché.

### Observation :

On observe que les villes comptant le plus grand nombre d'annonces sont Colombes, Rueil-Malmaison, Antony, Clamart et Nanterre. Cela indique que, dans le département des Hauts-de-Seine (92), ces communes représentent actuellement les zones les plus dynamiques en termes d'offre immobilière. Pour un agent immobilier, cela peut représenter un intérêt stratégique de concentrer davantage ses efforts sur ces secteurs porteurs.

### Offres récentes et favoris

#### Utilité du graphique :

Ce graphique (Figure 2.9) permet d'identifier les villes les plus dynamiques sur une période donnée et d'évaluer si ces zones ont suscité de l'intérêt de la part des clients sur cette période.

### Observation :

Au cours des 15 derniers jours, de nombreuses annonces ont été publiées à Rueil-Malmaison, mais la réactivité des clients n'a pas été au rendez-vous, avec peu d'interactions. En revanche, à Clamart, bien que de nombreuses annonces aient également été publiées, la réactivité des clients a été plus forte, faisant de cette ville celle qui a généré le plus d'intérêt en moyenne durant cette période. Quant à Antony, les annonces n'ont pas rencontré de succès.

Sur les 12 derniers mois (Figure 2.10), les annonces des villes Ville-d'Avray, Gennevilliers, Sèvres et Vanves ont plutôt bien fonctionné.

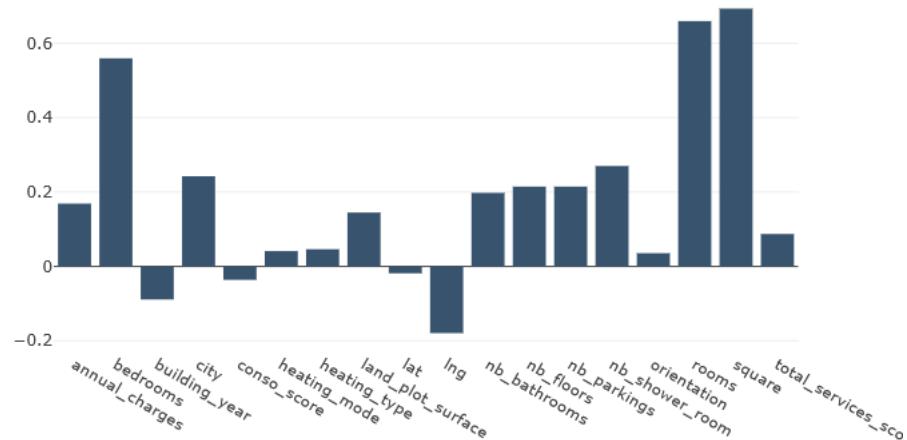


Figure 2.7: Corrélations entre le prix annoncé et les autres variables

Ce sont ces villes à succès que l'agent immobilier doit surveiller de près pour orienter ses actions de prospection.

### Analyse des concurrents

#### Utilité du graphique :

Ce graphique (Figure 2.11) permet d'analyser l'activité des agences en comparant le nombre d'annonces, offrant ainsi des *insights* sur la concurrence et permettant d'ajuster les stratégies commerciales.

#### Observation :

Les agences les plus présentes sont SARL Davidson, Peclers Immobilier, Maisons Pierre et JDC Conseil. L'agent immobilier doit donc analyser ces concurrents afin de se différencier en offrant des services de meilleure qualité.

### 2.5.4 Analyse des annonces

Dans cette section, nous allons effectuer une analyse des annonces. Avant l'analyse, nous devons préciser plusieurs éléments :

1. Comme le jeu de données ne possède pas des annonces de biens déjà vendus, nous évaluons le succès d'une annonce à son nombre de mises en favori.
2. Analyser le nombre de mises en favori des annonces sans prendre en compte leur durée passée en ligne provoquerait un biais. En effet, une annonce publiée antérieurement à une autre risque d'obtenir plus de mises en favori que cette dernière, compte tenu de sa durée d'accessibilité plus élevée. Ainsi, nous avons réparti les données par mois afin de corriger ce biais. Cependant, cette correction

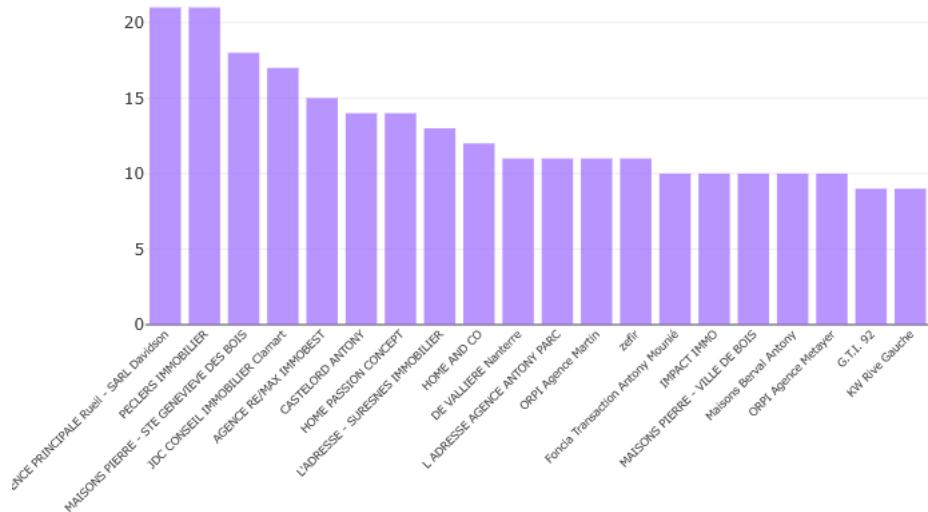


Figure 2.8: Nombre d'annonces par ville

est partielle puisqu'il subsiste une différence de temps entre la date de publication en début et en fin de mois.

3. Pour les graphiques de cette section, seuls les mois ayant chaque type d'annonces souhaité sont conservées. Par exemple, pour le graphique "Distribution moyenne de mises en favori par annonce" (2.5.4), les mois ayant seulement des annonces boostées ou seulement des annonces non boostées ne sont pas gardés, la comparaison entre les annonces boostées et non boostées étant impossible.

### Distribution moyenne de mises en favori par annonce selon le boost

Ce graphique (Figure 2.12) représente, par mois, le nombre moyen de mises en favori par annonce selon si elle est boostée ou non. Une annonce boostée est une annonce qui est mise en avant par la plateforme en échange d'un paiement. Le bien concerné est ainsi censé trouver preneur/se plus rapidement.

#### Utilité du graphique :

Ce graphique permet à l'agent immobilier de savoir si "booster" une annonce a réellement une influence sur son succès.

#### Observation :

Nous constatons que, paradoxalement, une annonce boostée a souvent moins de succès qu'une annonce non boostée. En effet, sur les 14 mois sélectionnés, huit voient le nombre moyen de mises en favori par annonce non boostée dépasser celui par annonce boostée. Dans certains cas (3/2024 ou 5/2024), la différence est très élevée (exemple 3/2024 : boostée = 35,3 mises en favori & non boostée = 284 mises en favori). Il semble donc qu'il n'est pas utile, pour l'offre, de "booster" son annonce. Cependant, l'absence des annonces de biens vendus nuance ce résultat. Par exemple, en mars 2024, il ne serait pas surprenant d'imaginer que si les biens boostés déjà vendus étaient pris

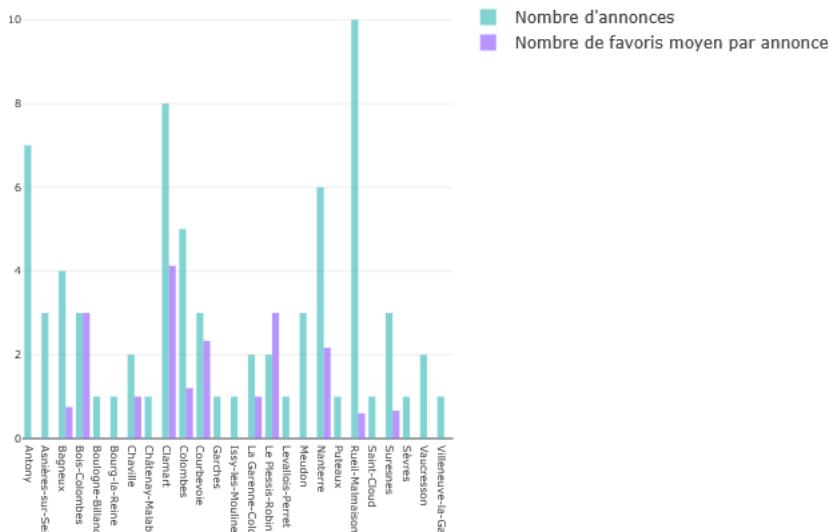


Figure 2.9: Offres récentes et favoris sur les 15 derniers jours

en compte, le nombre moyen de mises en favori soit beaucoup plus élevé.

### Distribution moyenne de mises en favori par type de vendeur/se

Ce graphique (Figure 2.13) représente, par mois, le nombre moyen de mises en favori par annonce selon le type de vendeur/se (professionnel/le ou particulier).

### Utilité du graphique :

Ce graphique permet à l'agent immobilier de savoir si le fait d'être un/e professionnelle favorise le succès d'une annonce.

### Observation :

Nous constatons que, sur les trois mois sélectionnés, une annonce d'un particulier a toujours plus de succès qu'une annonce d'un/e professionnel/le. Cependant, l'absence des annonces de biens vendus pourrait nous laisser penser que les biens proposés par des professionnels/lles ont, en moyenne, moins de mises en favori car, parmi ces biens, beaucoup ont trouvé preneur/se alors que ceux des particuliers sont populaires virtuellement mais très peu achetés. De plus, toutes les annonces antérieures à 2025 (non présentes sur le graphique) sont des annonces professionnelles. Cela peut signifier soit que les biens proposés par des particuliers ont tous trouvé preneur/se, soit que les annonces professionnelles sont majoritaires par rapport à celles des particuliers. Dans la première situation, cela indiquerait à l'agent immobilier qu'être professionnel/le n'est pas une garantie de succès dans la vente d'un bien. Le deuxième cas pourrait simplement être la conséquence d'une activité de ventes plus grande chez les agents, ce qui semblerait logique et n'indiquerait pas forcément un manque d'attractivité pour les annonces professionnelles.

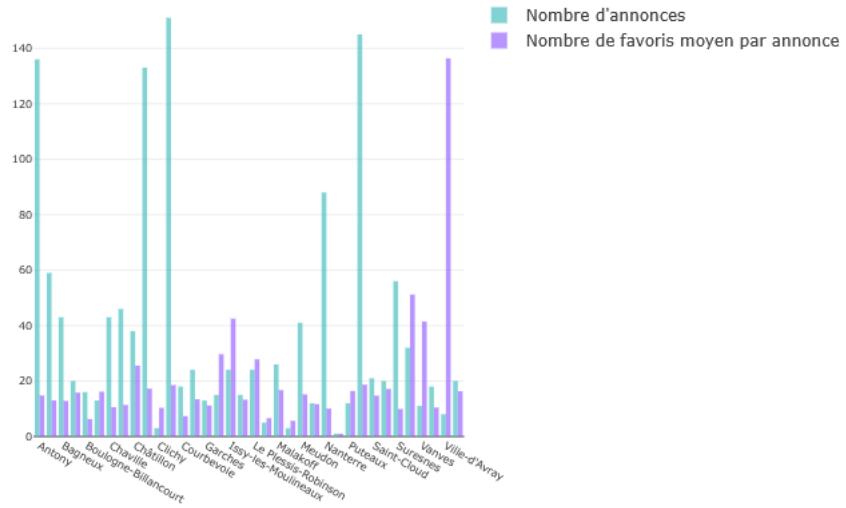


Figure 2.10: Offres récentes et favoris sur les 12 derniers mois

### 2.5.5 Carte interactive

La carte interactive, respectivement dans les affichages par villes, par zones (*clusters*) et par zones (*heatmap*) dans les Figures 2.14, 2.15 et 2.16, permet à l'agent immobilier d'accéder à une carte géographique centrée sur son territoire, les Hauts-de-Seine.

Cette carte propose deux modes de visualisation :

#### 1. Le mode "Annonces" (Figure 2.14)

Il affiche le volume d'annonces par commune. Les communes sont colorées selon leur niveau d'activité :

- (a) Les teintes qui se rapprochent du rouge représentent des volumes élevés d'annonces
- (b) Les teintes qui se rapprochent du vert indiquent une faible activité

Cela permet de :

- (a) Repérer les zones très actives comme Colombes, Nanterre, Rueil-Malmaison, Antony et Clamart, où l'offre est abondante.
- (b) Observer les zones sous-exploitées, notamment les villes de la petite couronne, où très peu d'annonces sont présentes. Ces zones représentent une opportunité forte de prise de mandat.

#### 2. Le mode "Prix" (Figure 2.17)

Il affiche le prix moyen au  $\text{m}^2$  par secteur. Les communes sont colorées selon leur niveau d'activité :

- (a) Les teintes qui se rapprochent du rouge représentent des prix élevés
- (b) Les teintes qui se rapprochent du vert indiquent des prix bas

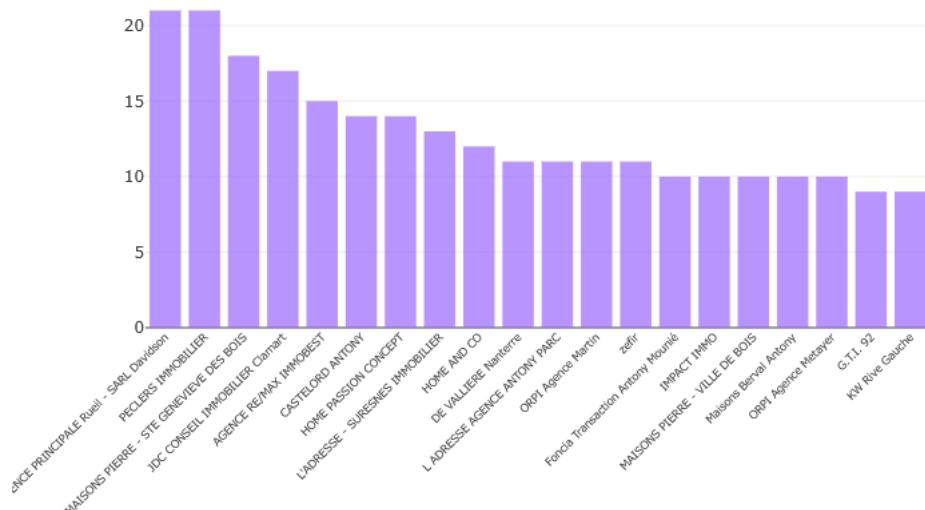


Figure 2.11: Nombre d'annonces par agence

Nous observons donc que :

- (a) Villeneuve-la-Garenne ressort comme la ville la moins chère.
- (b) Levallois-Perret ressort comme la ville la plus chère avec un prix record de 11 009 €/m<sup>2</sup>, mais très peu d'annonces (seulement trois).

On observe un gradient de prix classique : plus on se rapproche de Paris, plus les prix augmentent.

L'utilité ici est que le croisement de ces deux vues permette de comparer offre et attractivité. Tout cela permet aux agents d'identifier en un coup d'œil les zones en tension, les poches d'opportunités ou de blocage.

La carte possède plusieurs filtres :

1. Un filtre (Figure 2.18) permet d'afficher les biens ayant généré le plus de *likes* (jusqu'à 1554 favoris à Suresnes). On observe :
  - (a) Une activité diffuse mais modérée dans le sud (100 à 600 *likes*).
  - (b) Des "pépites" centrales comme à Sèvres, Saint-Cloud, Rueil-Malmaison.

Cela permet à l'agent de :

- (a) Identifier les types de biens qui attirent l'attention.
  - (b) Comprendre ce qui déclenche l'engagement (prix, qualité de la présentation, type de bien, quartier).
  - (c) Reproduire les recettes gagnantes pour ses propres annonces.
2. Un filtre (Figure 2.19) permet d'afficher les annonces restées longtemps en ligne (parfois plus de 400 jours à Montrouge ou à Clamart). On en observe une forte

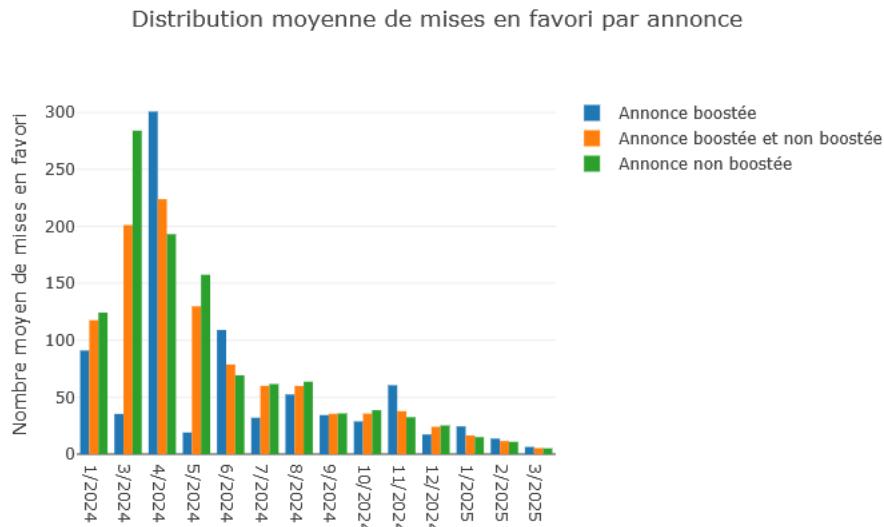


Figure 2.12: Distribution moyenne de mises en favori par annonce selon le boost

concentration à Antony et Fontenay-aux-Roses, dans le sud des Hauts de Seine, ce qui indique un désalignement avec la demande (prix trop élevés, mauvaise valorisation, etc.).

Cela permet à l'agent de :

- Repérer les biens surestimés à reprendre.
- Proposer au vendeur d'adapter sa stratégie : repositionnement du prix, amélioration de l'annonce, renforcement du discours client.

En cliquant sur une commune (Figure 2.20), un panneau latéral affiche :

- Le nombre d'annonces
- Le prix moyen au  $m^2$
- Et potentiellement d'autres données (indications sur la ville, type de biens, évolution...)

De plus, l'agent peut ajouter ses propres observations locales, comme :

- Des projets d'urbanisme
- De nouvelles infrastructures
- Sa perception de la qualité de vie

Cette couche qualitative enrichit la donnée brute et permet de mieux contextualiser les chiffres.

Pour guider les agents dans l'interprétation des données, nous avons intégré un module de questionnement stratégique, avec des questions comme :

- Pourquoi certains biens stagnent-ils ?

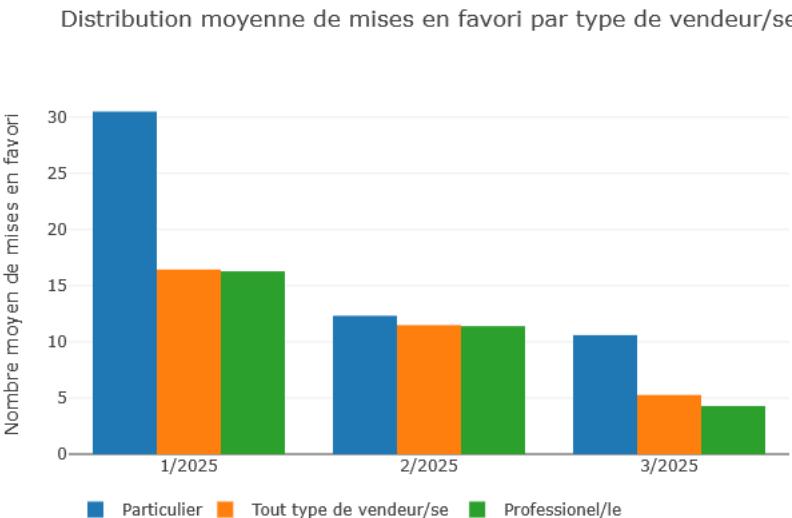


Figure 2.13: Distribution moyenne de mises en favori par type de vendeur/se

2. Pourquoi cette zone est-elle saturée ?
3. Pourquoi ce bien est-il populaire ?

Ces questions sont autant de points de départ pour une analyse terrain fine et un discours argumenté face aux clients. Les observations tirées de l'analyse, présentes sous la carte (Figure 2.21), ont une dimension immédiatement opérationnelle :

1. La rareté de l'offre en petite couronne représente un argument fort pour convaincre un vendeur.
2. Les biens populaires sont des modèles à suivre pour booster la présentation de ses annonces.
3. Les biens stagnants sont des opportunités de se positionner comme solution face à des propriétaires déçus.
4. Les zones à forte activité représentent des besoins de se différencier par une expertise locale renforcée.

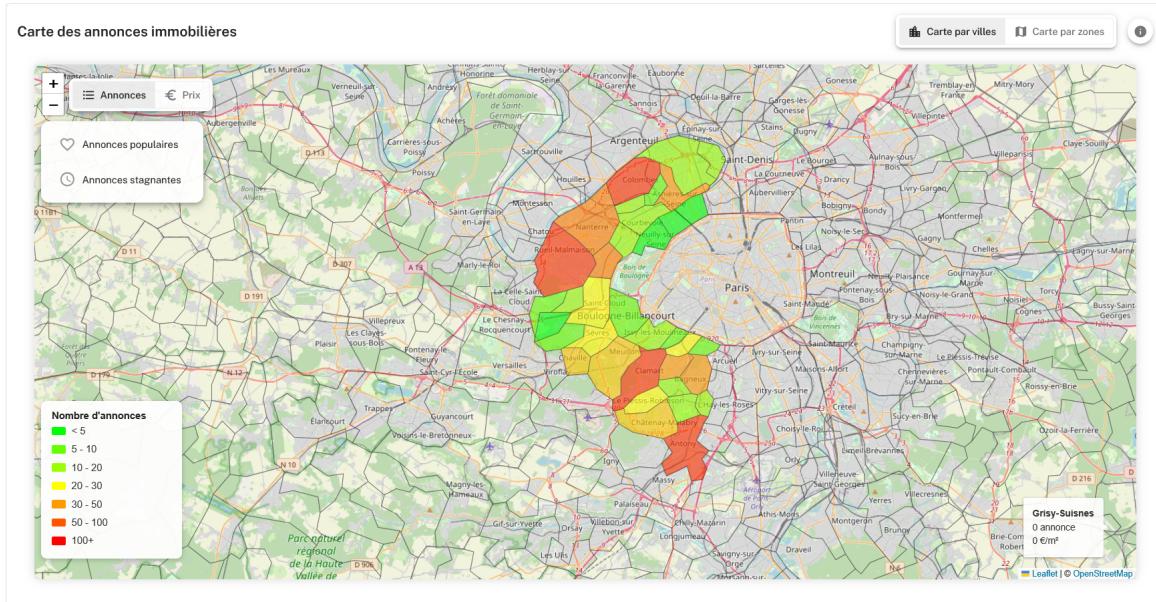


Figure 2.14: Carte interactive par villes : mode "Annonces"

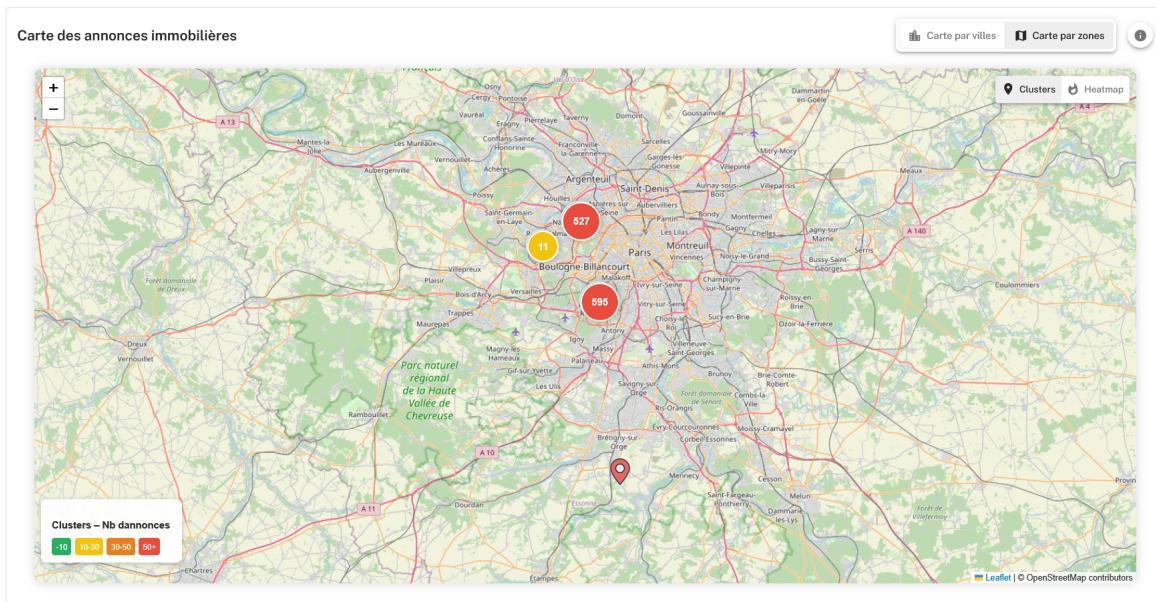


Figure 2.15: Carte interactive par zones : clusters

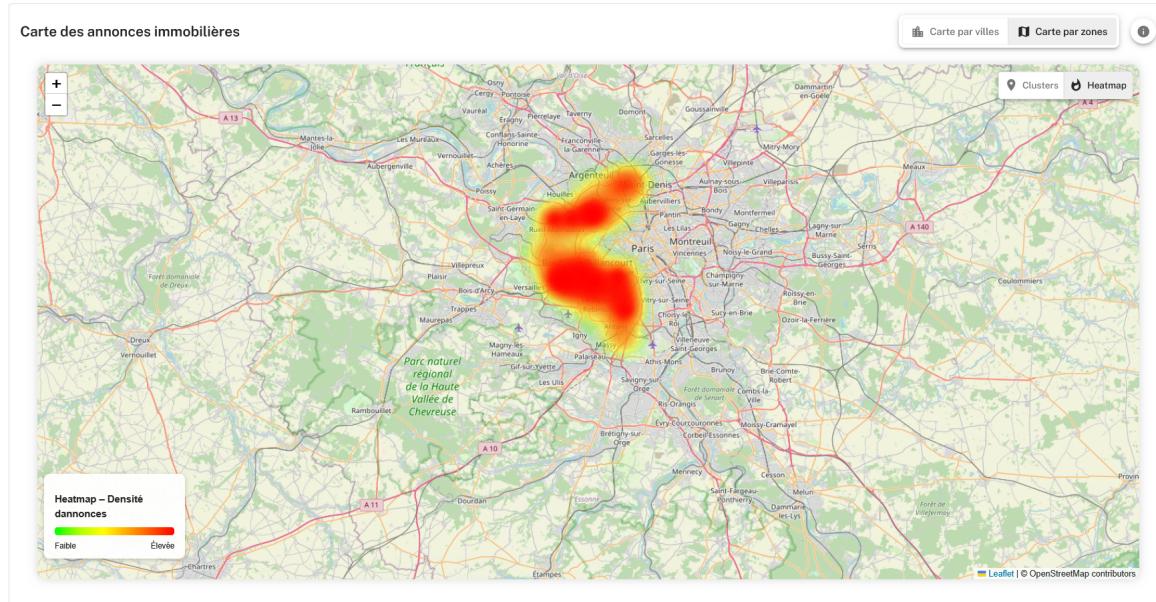


Figure 2.16: Carte interactive par zones : heatmap

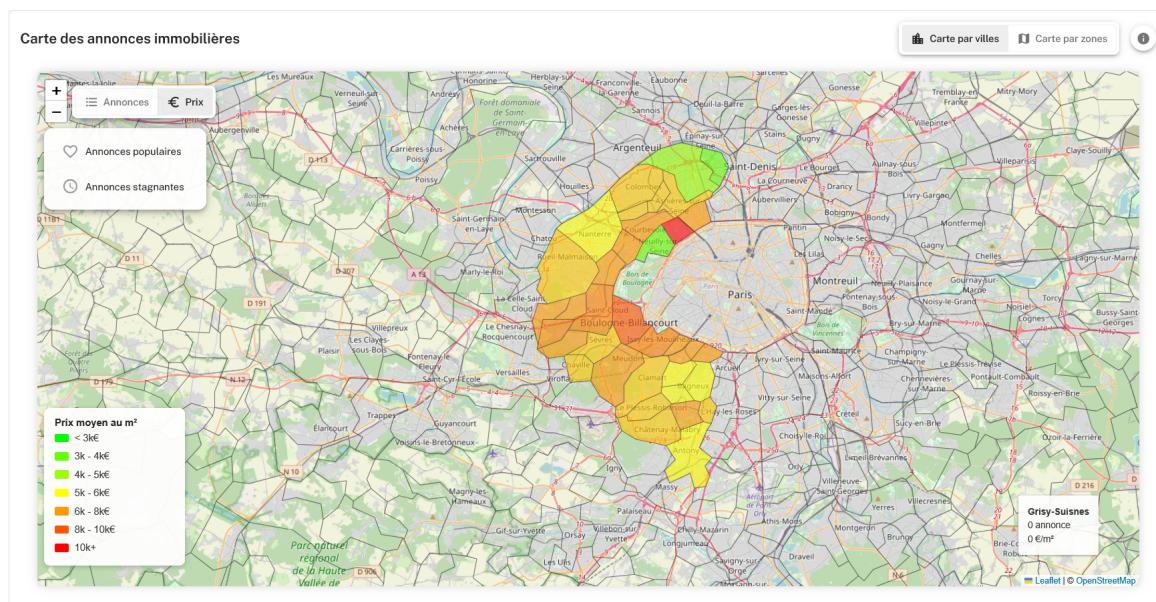


Figure 2.17: Carte interactive par villes : mode "Prix"

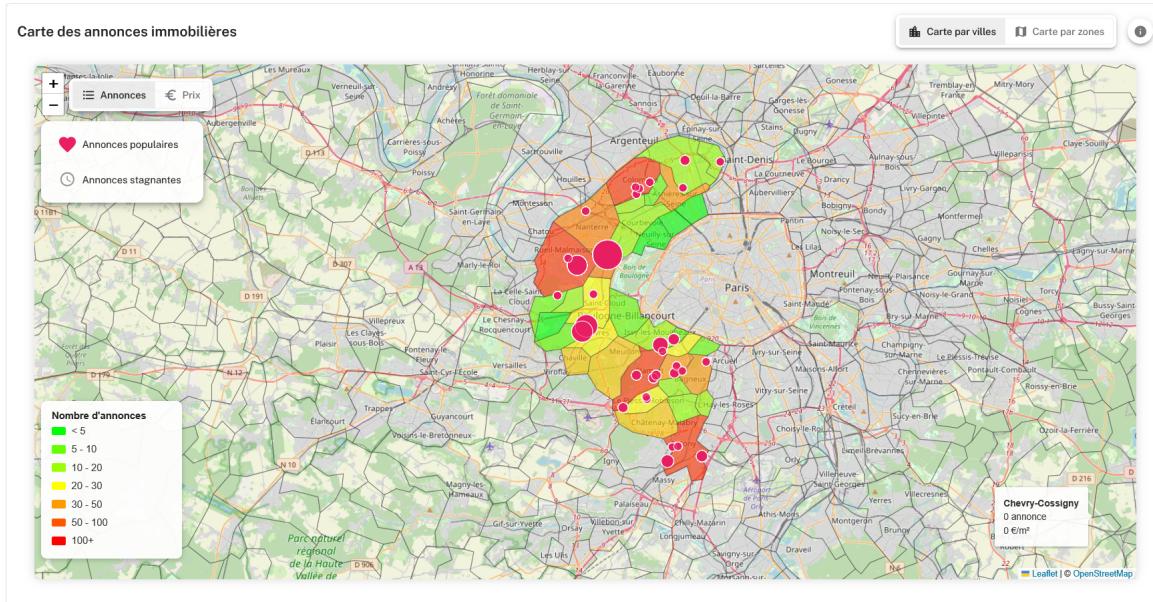


Figure 2.18: Carte interactive : filtre "Annonces populaires"

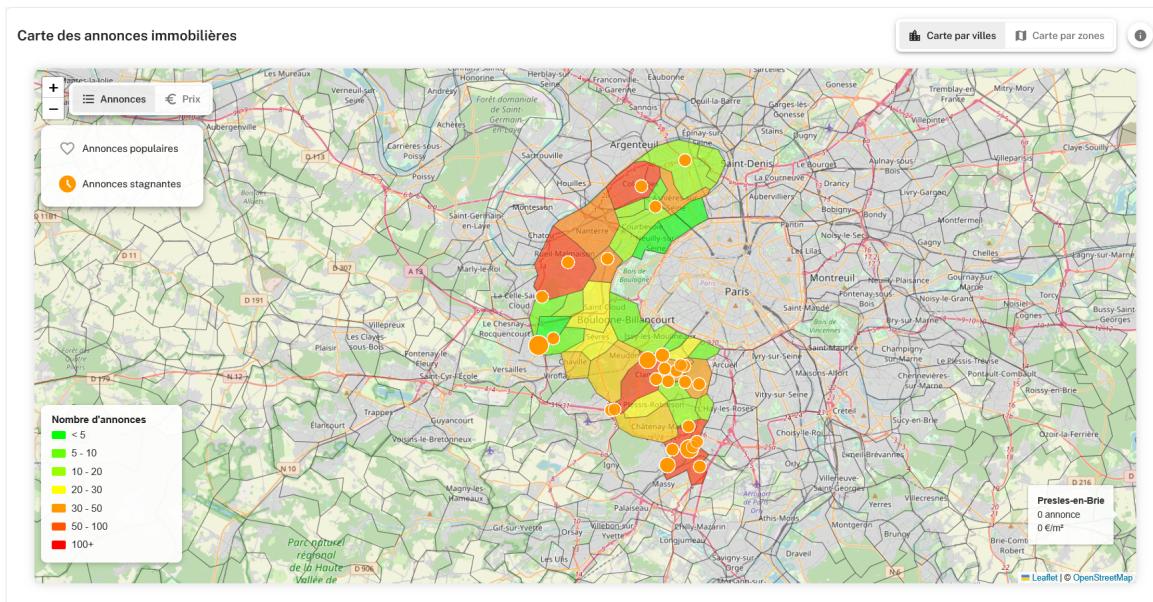


Figure 2.19: Carte interactive : filtre "Annonces stagnerantes"

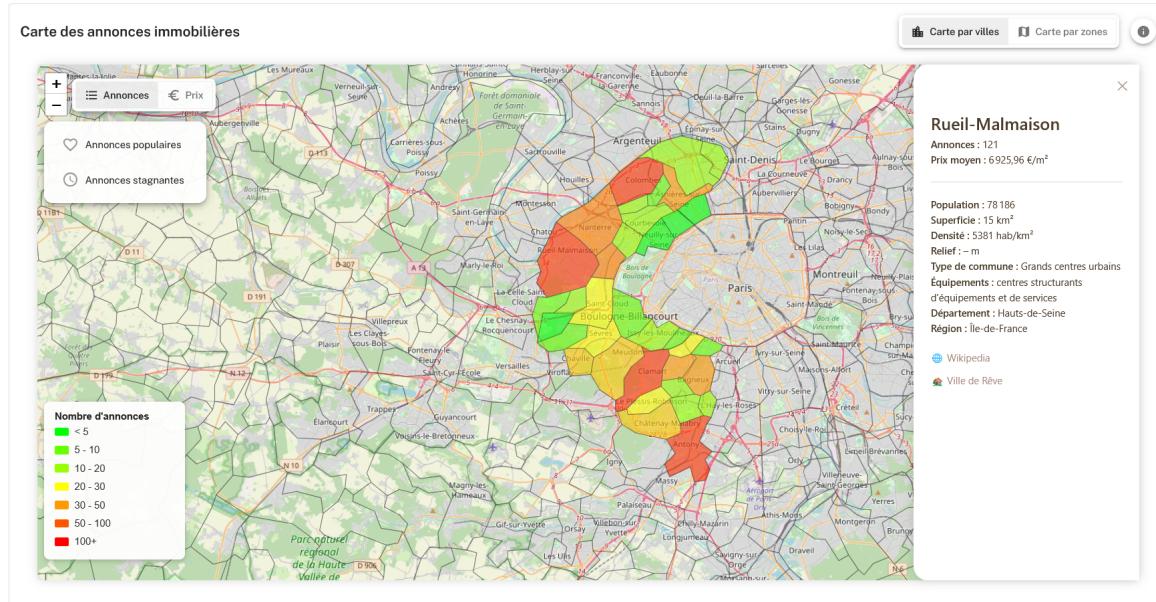


Figure 2.20: Carte interactive : informations détaillées pour une commune



Figure 2.21: Conseils stratégiques : conseils issus de notre analyse

# Chapter 3

## *Machine learning*

Dans le cadre de notre projet de *machine learning*, nous avons choisi de poursuivre le développement de l'application *web* commencée lors de notre projet de M1. Cette application est destinée aux agents immobiliers. Pour cela, nous avons exploré plusieurs pistes de travail en nous appuyant sur notre jeu de données existant que nous avons présenté dans la partie précédente.

### 3.1 Exploration des idées

Nous avons d'abord envisagé plusieurs idées basées sur notre jeu de données initial. La première piste consistait à prédire le prix d'un bien immobilier en fonction de ses caractéristiques. Cette approche, basée sur des méthodes de prédiction supervisée, avait pour but d'aider l'agent immobilier dans l'estimation rapide de la valeur des biens. Toutefois, nous avons constaté que notre base de données n'était pas adaptée à cette tâche, car elle ne contenait que des annonces encore disponibles à la vente. L'absence de données sur les biens effectivement vendus introduisait ainsi un biais important, empêchant une validation correcte du modèle.

Une seconde piste que nous avons explorée était l'évaluation de la qualité des annonces, en réalisant une classification basée sur le nombre de fois qu'une annonce avait été ajoutée en favoris par des utilisateurs. Cette méthode visait à fournir un indicateur qualitatif aux agents immobiliers pour améliorer leurs annonces. Cependant, après analyse, nous avons jugé cette approche insuffisante : le nombre de favoris ne représente pas directement la qualité d'une annonce et ne fournit pas de pistes précises d'amélioration. De plus, avec les avancées des modèles de langage de grande taille (LLM), il est désormais possible de générer automatiquement des descriptions attractives pour les annonces, ce qui limite encore la pertinence de cette méthode.

Finalement, nous avons retenu une troisième idée : la recommandation de biens immobiliers à un acheteur potentiel en fonction de critères pondérés qu'il définit lui-même. Cette approche, reposant sur des techniques de clustering non supervisé, nous est apparue plus pertinente à la fois pour la valeur ajoutée qu'elle apporte à l'utilisateur final et pour la faisabilité technique avec notre jeu de données.

## 3.2 Principe retenu

Le projet repose donc sur la recommandation de biens immobiliers en fonction des préférences exprimées par l'acquéreur. L'idée est de minimiser la distance entre les critères pondérés définis par l'utilisateur et les caractéristiques des annonces présentes dans notre base de données.

Cependant, comparer les critères utilisateur avec l'ensemble des 1000 annonces disponibles à chaque requête serait extrêmement coûteux en termes de temps de calcul. Pour optimiser ce processus, nous avons mis en place une approche en deux étapes. Dans un premier temps, nous effectuons un clustering sur les annonces immobilières afin de les regrouper selon leurs similarités. Dans un second temps, à partir des critères fournis par l'utilisateur, nous identifions le cluster le plus proche et nous réalisons une recherche de similarité parmi les annonces de ce cluster. Cela nous permet de classer rapidement les biens les plus pertinents.

## 3.3 Nettoyage et transformation des données

Les données utilisées pour ce projet proviennent d'un scraping des annonces immobilières publiées sur le site Le Bon Coin, en se concentrant pour l'instant sur le département des Hauts-de-Seine. Pour des informations plus détaillées concernant la collecte de données et le scraping, nous renvoyons à la partie dédiée à l'analyse des données.

Lors du traitement de ces données, nous avons récupéré toutes les caractéristiques explicitement mentionnées dans les critères des annonces. Cependant, certaines informations importantes, telles que la présence de services à proximité du bien ou la taille du sous-sol, étaient uniquement présentes dans les descriptions textuelles. Afin de ne pas perdre ces données précieuses, nous avons utilisé un modèle de langage (LLM) pour extraire automatiquement les informations pertinentes des descriptions et compléter notre jeu de données. La méthodologie employée est décrite dans la section 2.3.1.

## 3.4 Sélection des modèles de *clustering*

Notre projet repose sur un mécanisme d'attribution dynamique des modèles de machine learning, basé sur les variables renseignées par l'utilisateur. Concrètement, un modèle spécifique est sélectionné en fonction des critères sur lesquels l'utilisateur souhaite obtenir des recommandations.

### 3.4.1 Méthode

Pour cela, notre objectif initial était de fournir un filtre dans l'IHM reposant sur l'ensemble des variables sélectionnées lors du prétraitement des données. L'utilisateur aurait ainsi pu sélectionner toutes les variables qu'il souhaitait sans restriction sur l'ensemble des variables de départ. Toutefois, nous avons rapidement constaté que les utilisateurs pouvaient souhaiter ne renseigner qu'une partie de ces variables, ce qui engendrait un très grand nombre de combinaisons possibles. Chaque combinaison nécessitait potentiellement l'entraînement d'un modèle spécifique, rendant le projet difficilement soutenable en termes de complexité et de maintenance.

Afin de limiter cette explosion combinatoire, nous avons pris la décision de réduire le nombre de variables utilisées. Cette première réduction a permis de diminuer le nombre de modèles à entraîner, mais sans pour autant améliorer les performances, qui demeuraient globalement insatisfaisantes.

Pour comprendre cette situation, nous avons mené une analyse approfondie des corrélations entre les variables. Nous avons constaté que l'ajout de certaines variables dégradait la qualité du clustering, au lieu de l'améliorer.

Nous avons alors opté pour une démarche itérative : en retirant certaines variables et en intégrant d'autres, nous avons observé systématiquement l'impact de chaque modification sur les scores de performance. Celles-ci ont été évaluées selon deux indicateurs : le Silhouette Score et le Davies–Bouldin Index. Pour chaque cas, nous avons retenu les algorithmes offrant les meilleures performances. Cette approche nous a permis d'identifier clairement les variables qui contribuaient positivement à la qualité des modèles, ainsi que celles qui, au contraire, la pénalisaient.

### 3.4.2 Modèles de *clustering* sélectionnés

À l'issue de ce processus, nous avons retenu uniquement les variables les plus pertinentes. La version finale de notre modèle repose ainsi sur quatre variables obligatoires et trois variables facultatives, assurant un bon compromis entre performance, flexibilité et complexité de mise en œuvre. Afin de couvrir l'ensemble des combinaisons possibles de variables facultatives, nous avons donc déterminé qu'il était nécessaire de construire 15 modèles différents, correspondant aux combinaisons de 1 à 3 variables. En incluant également le modèle de référence, dans lequel aucune variable facultative n'est sélectionnée, nous avons ainsi élaboré un total de 16 modèles. Les scores des modèles sont présentés dans le tableau suivant (Figure 3.1) :

Cas	CAH		KMEANS	
	Sihoulette	Davies B	Sihoulette	Davies B
Only required var (Nb cluster : 3)	0.354652674014893	0.842812501109508	0.363617301	1.009916975
Trans (Nb cluster : 3)	0.3961879438079913	1.1116203955404809	0.380223801	1.135532884
Trans_Sch (Nb cluster : 3)	0.317892740546666	1.1101594149789764	0.314221437	1.251974666
Trans_Med (Nb cluster : 3)	0.36633784700460875	1.3529334522907108	0.311915621	1.398943556
Trans_Sch_Med (Nb cluster : 3)	0.32954834301904906	1.4101903403528544	0.271090293	1.538930521
Sch (Nb cluster : 3)	0.3128910726774541	0.9920096624993763	0.330434547	1.203626320
Sch_Med (Nb cluster : 4)	0.32888151745453076	1.0860477692263728	0.34226839851777907	1.1799764606118495
Med	(Nb cluster : 4) 0.38278335103071526	(Nb cluster : 4) 0.8673091680246258	[Nb cluster : 3] 0.369046130	[Nb cluster : 3] 1.187744338

Figure 3.1: Scores des modèles : Silhouette Score et Davies–Bouldin Index

### 3.5 Entrainement des modèles et prédictions

Lorsqu'un utilisateur renseigne ses critères de recherche, ceux-ci sont transformés pour être comparables à un nouveau bien immobilier fictif. L'objectif est alors de prédire dans quel cluster existant ce bien serait le plus susceptible de se trouver.

En fonction des critères fournis par l'utilisateur, le système sélectionne automatiquement le modèle de clustering adapté parmi ceux préalablement construits. Selon les cas, nous utilisons soit l'algorithme K-Means, soit la Classification Ascendante Hiérarchique (CAH) et nous faisons également varier le nombre de clusters K.

L'algorithme K-Means, implémenté dans la bibliothèque Scikit-learn, possède une fonction native de prédiction qui permet d'attribuer facilement un nouveau point à un cluster existant. En revanche, l'approche CAH ne dispose pas d'une fonction de prédiction directe. Pour pallier cette limitation, nous avons mis en œuvre une méthode complémentaire de prédiction/classification (KNclassifier) pour estimer l'appartenance au cluster.

### 3.6 *Threats to validity*

Notre approche présente plusieurs limites.

Tout d'abord, les données utilisées proviennent exclusivement d'annonces du site Le Bon Coin pour le département des Hauts-de-Seine, ce qui limite leur représentativité à l'échelle nationale. De plus, seules les annonces actives ont été collectées, introduisant un biais de sélection qui peut affecter la pertinence des modèles.

L'extraction automatique d'informations via un modèle de langage expose également à des risques d'erreurs, notamment lorsque les descriptions textuelles sont ambiguës ou incomplètes.

Concernant les variables, la sélection manuelle, bien que guidée par des scores de clustering, reste subjective et peut avoir écarté certaines variables importantes. Enfin, la réduction du nombre de variables facultatives, nécessaire pour limiter la complexité, restreint la capacité du modèle à capturer toute la diversité des préférences utilisateurs.

### 3.7 Pistes d'amélioration et ouverture

En termes d'amélioration, nous envisageons à l'avenir de rendre les modèles de machine learning dynamiques en fonction des pondérations données par l'utilisateur. De plus, l'utilisation d'outils comme FAISS (Facebook AI Similarity Search) ou Annoy (développé par Spotify) pourrait permettre d'optimiser davantage la recherche de similarité, soit sur la totalité des annonces, soit à l'intérieur des clusters. Nous prévoyons de comparer les performances de ces différentes approches.

### 3.8 Conclusion

Nous avons ainsi choisi d'approfondir notre travail en développant plusieurs modèles adaptés aux différentes configurations, plutôt que de nous limiter à un modèle unique. Notre démarche visait à construire des modèles à la fois performants, grâce à une évaluation rigoureuse basée sur des indicateurs de qualité, et complémentaires, afin de couvrir un large éventail de besoins utilisateurs. L'objectif final était de proposer une approche réaliste et robuste, fidèle aux contraintes et à la diversité des situations rencontrées en conditions réelles.

# Webographie

[leboncoin] [leboncoin.fr](http://leboncoin.fr)

[ScrapeGraphAI] [scrapegraph-ai.readthedocs.io/en/latest/](https://scrapegraph-ai.readthedocs.io/en/latest/)

[Ollama] [ollama.com](https://ollama.com)

[Mistral AI] [mistral.ai](https://mistral.ai)

[Mistral Nemo] [mistral.ai/news/mistral-nemo](https://mistral.ai/news/mistral-nemo)