

SEGMENTATION DU MARCHE IMMOBILIER

FRANÇAIS



Nathan WOHL
Robin KHATIB
Alexandru NITESCU

Partie 0 : Approche méthodologique et choix du périmètre

Ce projet consiste à segmenter le marché immobilier français à partir des données DVF (Demandes de Valeurs Foncières). Avant toute analyse, il est indispensable d'examiner la structure de la base et d'évaluer sa cohérence. En effet, même si DVF est une source officielle produite par l'État, les données sont brutes, très hétérogènes et comportent des transactions de nature différentes (terrains, parcelles agricoles, dépendances, parkings, lots multiples, etc.).

L'exploration préliminaire du fichier montre que DVF recense chaque transaction lot par lot, ce qui entraîne des doublons logiques, des valeurs manquantes et de nombreux enregistrements correspondant à des biens non comparables. De plus, la base contient un volume important de transactions qui ne relèvent pas de l'immobilier résidentiel : terrains nus, bois et prés, garages, dépendances ou ventes atypiques. Conserver l'ensemble de ces catégories rendrait toute segmentation incohérente, les variables clés (surface habitable, pièces, prix/m²) n'étant pas définissables pour ces biens.

Comme l'objectif de ce projet est d'identifier des groupes homogènes de biens immobiliers partageant des caractéristiques similaires, il est logique de restreindre l'analyse aux biens pour lesquels ces variables ont un sens. On limite donc le périmètre aux seuls biens résidentiels bâtis, à savoir :

- Maisons
- Appartement

Ce choix garantit une cohérence statistique (unité de mesure et distributions des variables comparables), une cohérence métier (focus sur le marché résidentiel), une stabilité des algorithmes de clustering, et des profils de clusters lisibles.

Par ailleurs, pour assurer une base cohérente et éviter les erreurs liées aux évolutions réglementaires ou aux effets conjoncturels, l'analyse est limitée à l'année 2024, qui constitue la dernière année complète disponible. Cela garantit l'homogénéité temporelle, la cohérence des prix et la réduction du volume de données.

Partie 1 : Chargement et pré filtrage

Le fichier DVF 2024 contient plusieurs millions de lignes et ne peut pas être chargé en mémoire d'un seul bloc. On utilise donc un chargement par chunks (blocs successifs). Cette méthode est indispensable pour manipuler ce type de quantité de données.

Avec ce chargement nous appliquons un premier pré filtrage pour éliminer immédiatement les transactions qui ne sont pas pertinentes pour la suite du projet. Seules les lignes correspondant

à une vente (“Nature mutation = Vente”) sont conservées, tandis que les autres types d’opérations (échanges, VEFA, mutations non tarifées, etc.) sont exclues.

Ici on garde seulement les colonnes essentielles à l’analyse d’un point de vue métier. On a principalement des variables liées :

- Au **prix de vente**,
- Aux **surfaces** et aux **nombres de pièces**,
- Au **type de bien**,
- A la **localisation** (commune, code postal, département, voie),
- A la **date de transaction**.

Ce pré filtrage conduit à environ 3,2 millions d’enregistrements.

Partie 2 : Nettoyage des données

Après le pré filtrage des seules transactions de type « Vente », nous avons appliqué un nettoyage approfondi pour obtenir une base cohérente et exploitable pour la segmentation. Notre fichier DVF 2024 pré filtré contenait **3 218 183 enregistrements**, une grande partie correspondait à des biens non comparables ou à des transactions atypiques.

Nous avons nettoyé en suivant ces étapes :

1. **Conversion des variables numériques** : les colonnes “Valeur foncière”, “Surface réelle bâtie”, “Nombre de pièces principales” et “Surface terrain” ont été normalisées et converties en types numériques exploitables.
2. **Filtrage du périmètre résidentiel** : Seuls les enregistrements dont le type est Maison ou Appartement ont été conservés.
3. **Suppression des ventes multi-lots** : Les transactions impliquant plusieurs lots génèrent plusieurs lignes dans DVF et elles ne peuvent pas être interprétées comme des biens individuels.
4. **Elimination des valeurs aberrantes** : Les transactions à un prix très faible (moins de 5 000€), une surface bâtie inférieure à 10 m², ou un prix au mètre carré incohérent ont été retirées.

À l’issue de ces opérations, la base finale contient **921 191 transactions**, soit environ **29 %** des données initialement pré filtrées. Cette base nettoyée, homogène et cohérente, constitue un bon fondement pour les étapes suivantes.

Partie 3 : Feature engineering

L'objectif du feature engineering est d'enrichir la base nettoyée avec des variables informatives permettant aux algorithmes de clustering de mieux identifier des groupes pertinents. En plus de certaines variables conservées (surfaces réelle bâtie, nombre de pièces principales), plusieurs nouvelles variables ont été construites.

3.1 Variables individuelles

Plusieurs nouvelles variables ont été construites à partir des attributs de chaque transaction :

- **Prix au mètre carré**
- **Logarithme du prix au mètre carré** : permet de réduire l'impact des valeurs extrêmes et de stabiliser la distribution.
- **Le ratio surface terrain / surface bâtie** : différencie les biens urbains (ratio faible) des biens ruraux (ratio élevé)
- **Encodage numérique du type de bien** : transforme la variable catégorielle en variable numérique.

3.2 Variables temporelles

On intègre la dimension temporelle à partir de la date de mutation :

- Mois de la transaction,
- Trimestre de la transaction.

Ces variables permettent de capter d'éventuels effets saisonniers ou des dynamiques temporelles du marché immobilier.

3.3 Indicateurs de marché local

Le marché immobilier est fortement dépendant de la localisation, on construit plusieurs indicateurs de contexte à différentes échelles géographique :

À l'échelle de la commune :

- Prix médian,
- Surface médiane,
- Volume de transactions,
- Part de maison parmi les ventes.

À l'échelle du code postal :

- Prix médian au mètre carré.

À l'échelle de la voie :

- Prix médian au mètre carré

Grace à cette approche cela nous permet de capturer le contexte de marché local dans lequel s'inscrit chaque bien.

3.4 Encodage des variables catégorielles

Pour rendre exploitables les variables qualitatives tout en conservant leur information métier nous avons procédé à :

- un **target encoding** qui est appliqué aux variables de localisation (commune, code postal, département, voie) ;
- un **frequency encoding** qui est appliqué au type de voie afin de représenter la fréquence relative des différents type d'adresses.

On conserve les colonnes textuelles brutes quand c'est pertinent, mais les variables numériques issues de ces encodages sont les features finales.

Partie 4 : Réduction dimensionnelle

Notre base comporte un ensemble de variables numériques décrivant à la fois les caractéristiques des biens et les dynamiques locales du marché. Ces variables incluent des indicateurs de prix, de surface, de ratios, ainsi que des statistiques agrégées à différentes échelles géographiques. Pour obtenir un espace plus stable, décorrélé et adapté à la segmentation, nous avons appliqué une réduction dimensionnelle par Analyse en Composantes Principales (PCA).

Ici la PCA est utilisé comme étape de pré-traitement, et non dans un objectif d'interprétation économique des composantes. On veut décorréliser les variables pour **supprimer les potentielles redondances d'information, stabiliser les distances entre observation et réduire le bruit statistique**. En somme, la PCA permet de préparer un espace de représentation cohérent et exploitable pour les prochaines étapes.

4.1 Standardisation des variables

Avant l'application de la PCA, l'ensemble des variables numériques sélectionnées est standardisé. Cette étape est indispensable, car les variables utilisées présentent des échelles très différentes (prix, volumes, ratios, fréquences). La standardisation permet de centrer et réduire les variables.

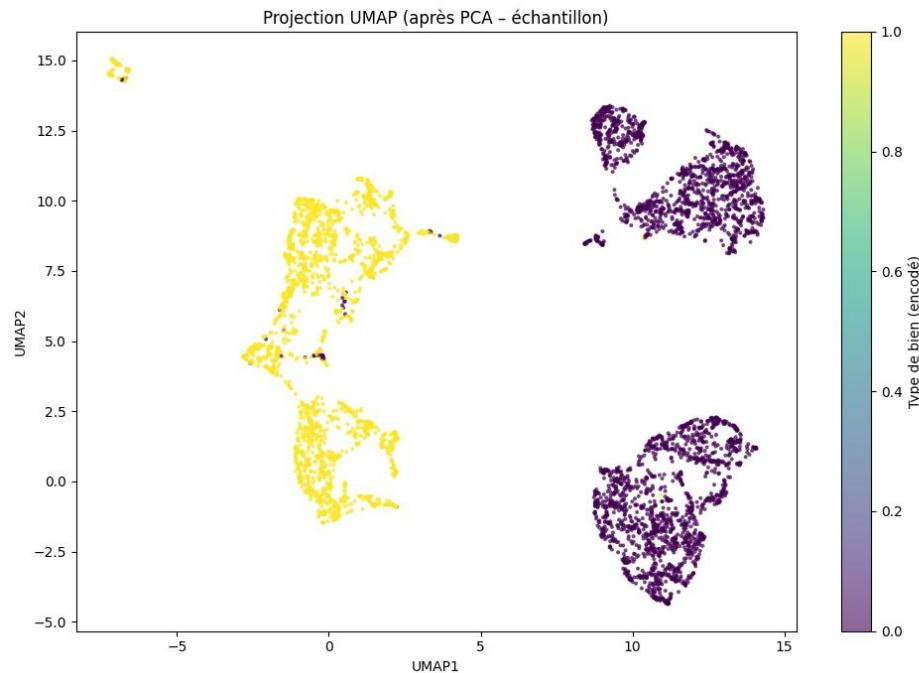
4.2 PCA complète

Après standardisation, une première PCA est appliquée en conservant l'ensemble des composantes possibles. Le nombre de variables numériques utilisées est de 17, la PCA est donc réalisée avec 17 composantes principales, ce qui correspond au maximum autorisé par les données. Dans ce cas, la PCA restitue logiquement 100 % de la variance totale.

Cette PCA complète ne vise pas à réduire fortement la dimension, mais à effectuer un **changement de base orthogonale**, ce qui permet de projeter les données dans un espace décorrélé.

4.3 Visualisation

Nous avons réalisé une visualisation exploratoire grâce à UMAP. Cette projection est appliquée à partir des données préalablement projetées par la PCA. Pour des raisons de complexité et de lisibilité nous avons réalisé cette visualisation sur un sous-échantillon.



La figure obtenue montre une séparation nette entre les maisons (jaune) et les appartements (violet). Les appartements forment des groupes relativement compacts, traduisant une certaine homogénéité des biens, tandis que les maisons apparaissent plus dispersées, reflétant une plus grande diversité en termes de surfaces, de prix et de contexte géographique. On observe également plusieurs zones de densité au sein de chaque type de bien, suggérant l'existence de sous-structures internes.

Partie 5 : Clustering

Le but de cette partie est d'identifier des segments de marché homogènes au sein des transactions immobilières françaises de 2024.

5.1 Choix méthodologique

Compte tenu du volume de données conséquent, le clustering est réalisé à l'aide de **MiniBatch K-Means**, qui permet de traiter l'ensemble de la base tout en conservant une bonne stabilité des

résultats. Le clustering est effectué sur les **3 premières composantes principales**, issues de la PCA validée en Partie 4. Cette projection permet de conserver l'essentiel de l'information structurante tout en limitant le bruit. Aucune séparation a priori entre maisons et appartements n'est imposée : le modèle est volontairement global, afin de vérifier si la typologie des biens émerge naturellement des données.

5.2 Nombre de Cluster et Score silhouette

5.2.1 Modèle principale

Le modèle principal repose sur un **K-Means global avec k = 6**, conformément à l'énoncé (5 à 7 clusters attendus).

- Silhouette score (global) : 0.331** (calculé sur un échantillon de 20 000 observations)

Ce score, bien que modéré, est cohérent avec la nature complexe du marché immobilier, où les transitions entre segments sont progressives.

Modèle	Nombre de Clusters	PCA utilisées	Silhouette score
MiniBatch K-Means	6	PC 1 à 3	0.331

5.2.2 Tests par types de biens

Des tests en plus ont été réalisés par type de bien :

Appartements :

- k testé: 3 à 6
- k optimal: 4
- Silhouette max: 0.301

Maisons :

- k testé: 3 à 6
- k optimal: 3
- Silhouette max: 0.330

Sous-ensemble	K testé	Silhouette
Appartements	3	0.268
Appartements	4	0.301
Appartements	5	0.248
Maisons	3	0.330
Maisons	4	0.264
Maisons	5	0.253

5.3 Profils statistiques et clusters

L'interprétation des clusters repose sur les statistiques descriptives par cluster, calculées sur les variables métier principales :

- prix au m² (moyenne et médiane)
- surface habitable
- nombre de pièces
- ratio terrain / bâti
- part maisons / appartements
- indicateurs de prix locaux (commune, code postal, département)

Ces résultats permettent d'identifier **six profils** de marché distincts, allant de segments urbains très chers à des segments ruraux accessibles.

Cluster	Prix €/ m ² moyen	Prix €/ m ² médian	Surface moy.(m ²)	Pièce moy.	% Maisons	Nb biens
0	10 211	9 757	56.8	2.5	5%	35 975
1	3 140	2 750	55.2	2.5	14%	221 830
2	4 103	3 640	122.7	4.8	95%	107 723
3	1 729	1 607	101.8	4.2	96%	286 870
4	2 492	2 203	85.3	3.6	68%	149 305
5	6 614	5 592	55.3	2.5	14%	119 488

Ce tableau est un échantillon d'un tableau contenant plus de variables généré par le code (cluster_profiles.csv).

Afin de comparer visuellement les profils moyens par cluster sur plusieurs dimensions clés, nous utilisons un diagramme radar :

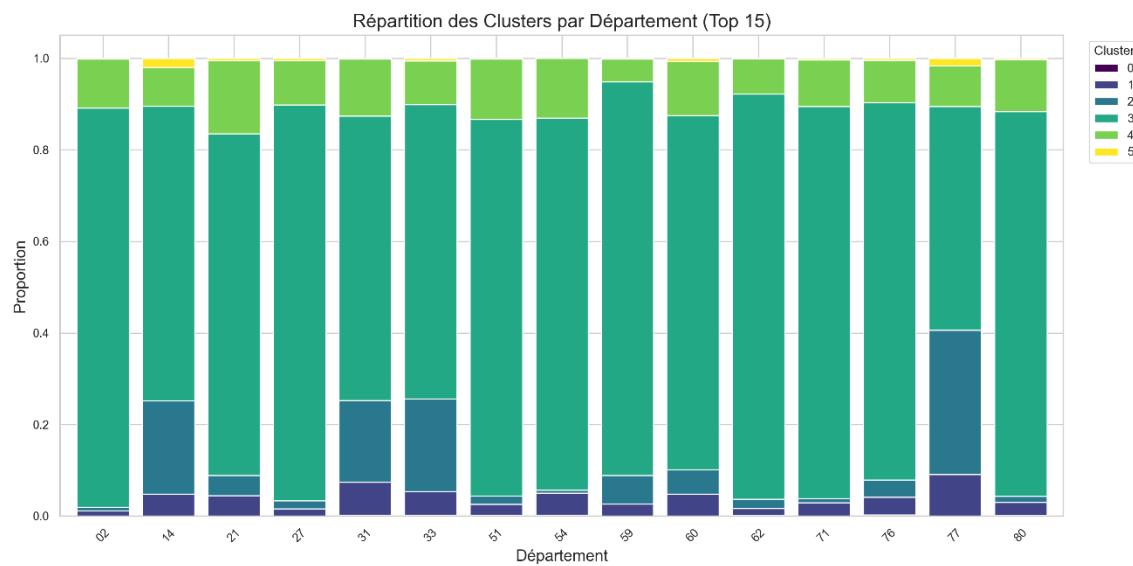


Ce graph mets en évidence des profils clairement différenciés, opposant des segments urbains à prix élevé et surfaces réduites à des segments majoritairement composés de maisons, avec de plus grandes surfaces.

5.4 Visualisation et validation géographique

5.4.1 Distributions Géographique

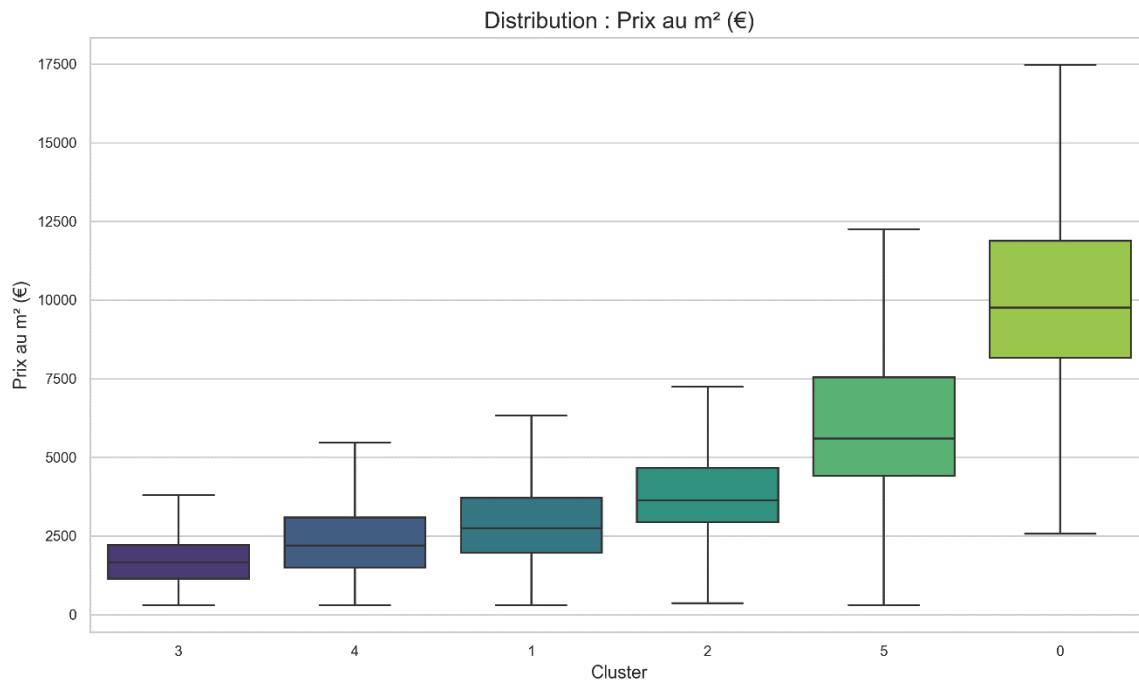
Pour valider la cohérence spatiale des segments identifiés, nous analysons la répartition des clusters par département. Cette visualisation permet de vérifier si les segments obtenus présentent une logique territoriale identifiables.



On observe une spécialisation géographique marquée des clusters. Certains segments dominent largement dans les zones rurales (comme le Cluster 3), tandis que le Cluster 0 est quasi-exclusif à Paris et sa petite couronne.

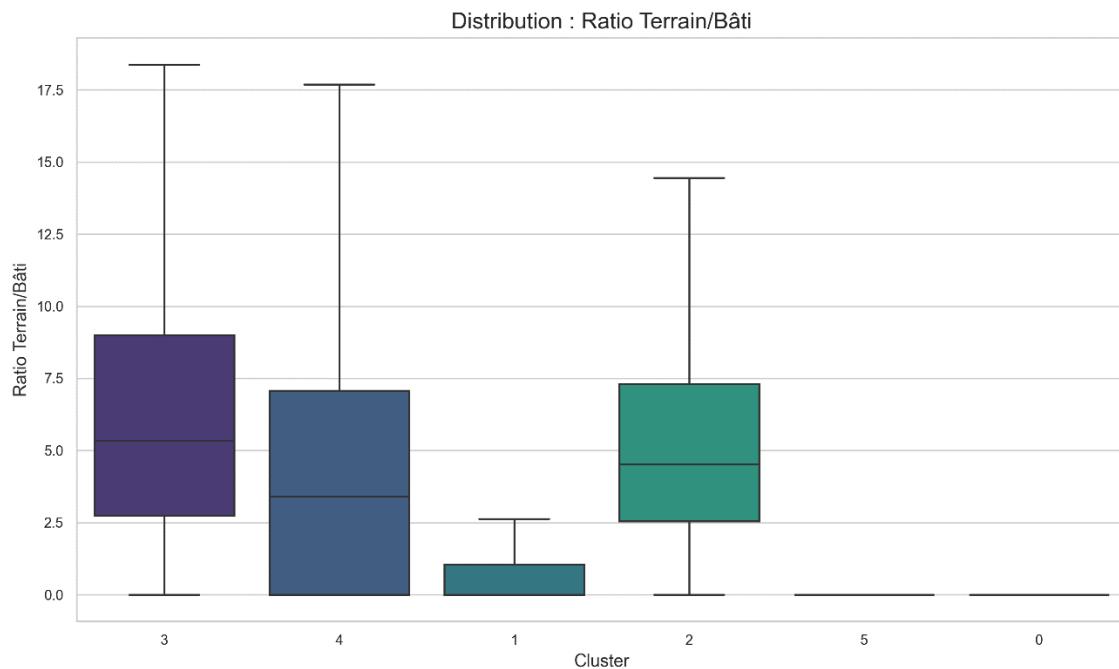
5.4.2 Comparaison des distributions des variables clés

- La distribution du prix au mètre carré par cluster met en évidence une hiérarchie claire entre les segments.



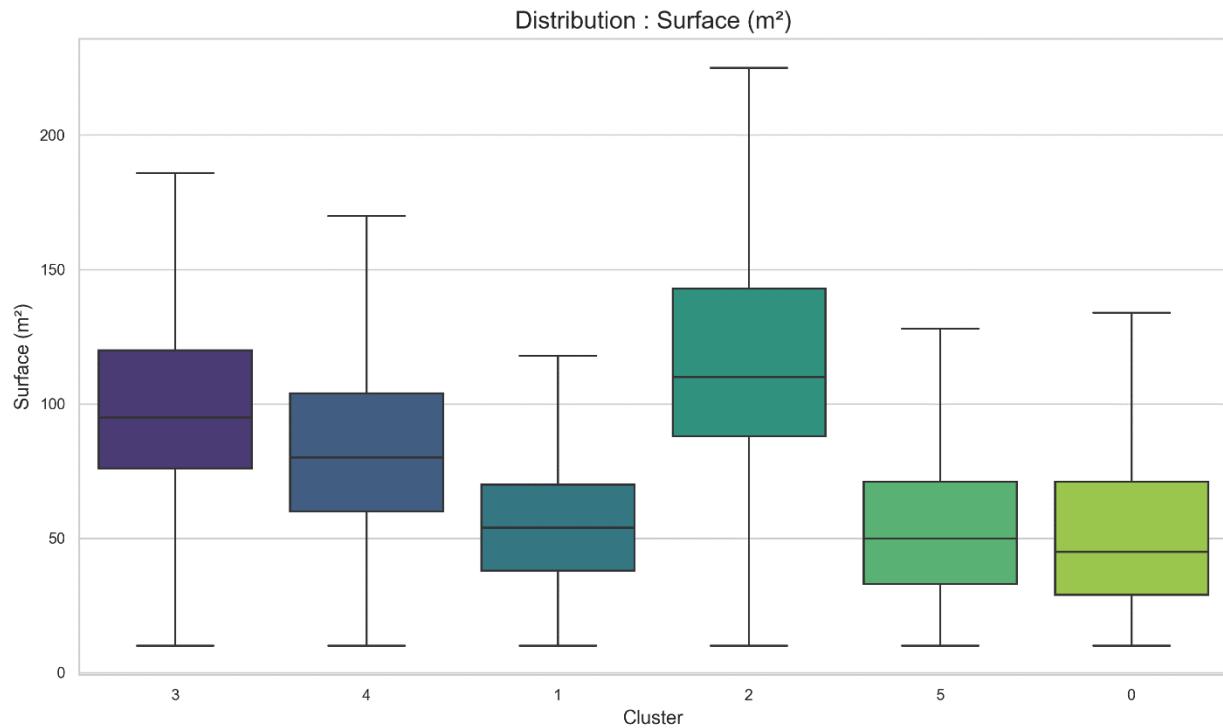
Les clusters se distinguent nettement par leur niveau de prix allant de segments à bas prix à des segments premium. La dispersion observée reflète la diversité des marchés locaux à l'intérieur de chaque segment.

- Le ratio Terrain/bâti permet de différencier les biens urbains denses des biens disposant de grandes parcelles.



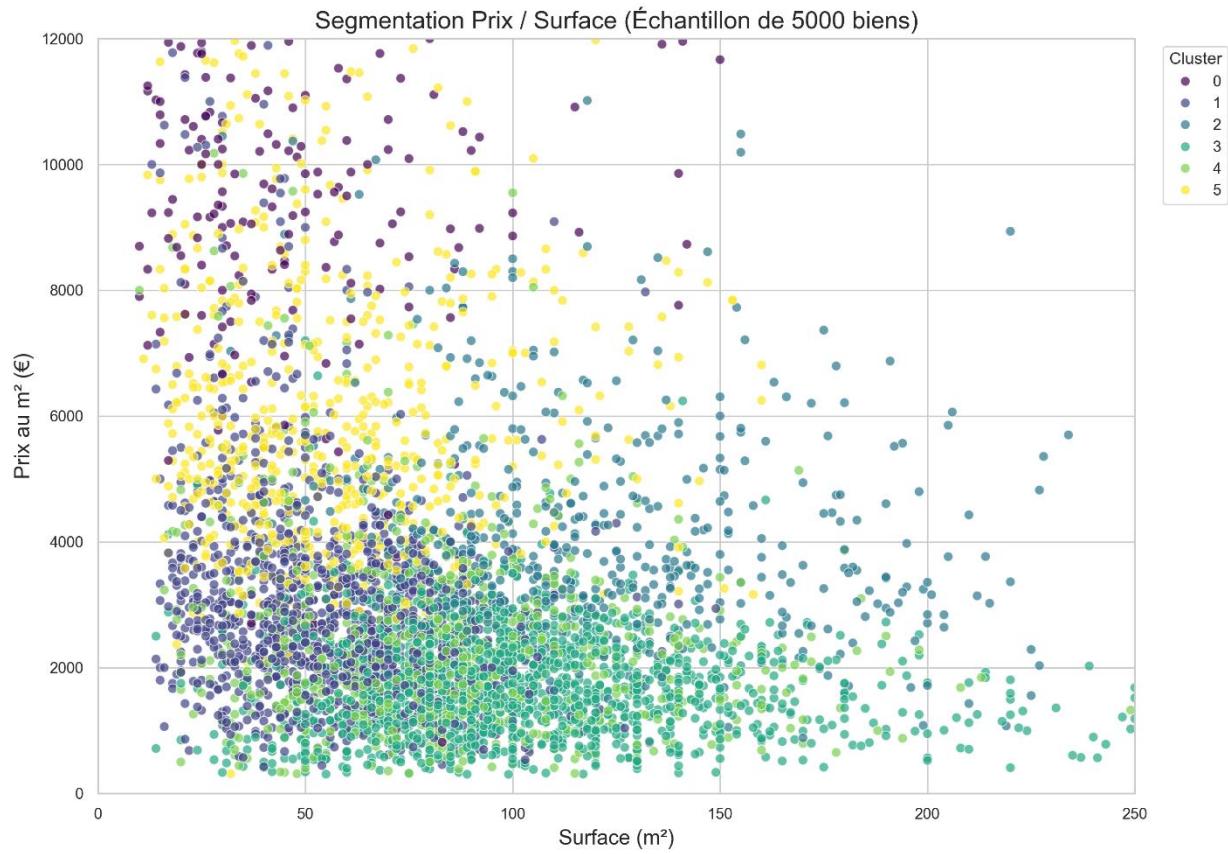
Les clusters dominés par les maisons présentent des ratios significativement plus élevés, tandis que les clusters majoritairement composés d'appartements affichent des valeurs proches de zéro.

- La surface habitable constitue un critère structurant de la segmentation.



Les clusters se répartissent clairement entre petites surfaces urbaines et grandes surfaces résidentielles, ce qui confirme la capacité du modèle à capter des différences structurelles de taille des biens.

- Une projection bidimensionnelle prix au mètre carré / surface permet d'illustrer visuellement la séparation des segments sur des variables directement interprétables.



Bien que les clusters se chevauchent partiellement, des zones de concentration apparaissent clairement, confirmant que la segmentation repose sur les combinaisons cohérentes de prix et de surface plutôt que sur une séparation artificielle.

Partie 6 : Détection d'opportunités d'investissement

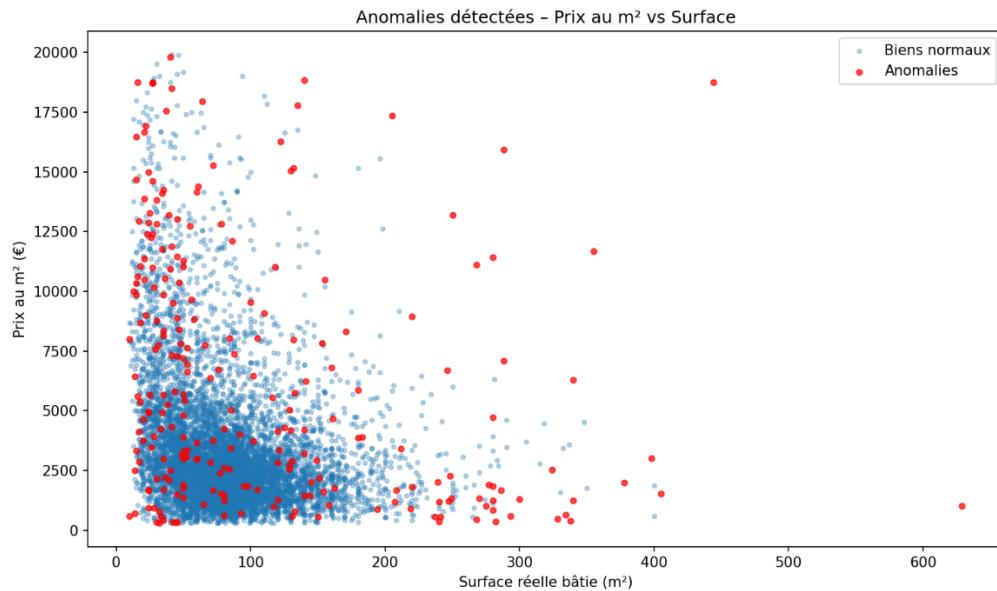
Après la segmentation du marché immobilier, nous avons fait une détection d'anomalies pour identifier des biens atypiques au sein de chaque cluster. On a utilisé pour cela un modèle Isolation Forest qui a été entraîné séparément sur chaque cluster, en utilisant des variables structurelles (prix au m^2 , surface, nombre de pièces, ratio terrain/bâti) et un indicateur de contexte local (prix médian communal).

Grace à ça on peut identifier des biens qui s'écartent beaucoup du comportement moyen de leur segment de marché, sans biaiser l'analyse par des différences structurelles entre type de bien (maison vs appartement ou zones urbaines vs rurales).

Nous observons un taux d'anomalies de 3 %, un volume réaliste qui capture les biens singuliers sans rejeter massivement les données. D'un point de vue opérationnel, ces biens sortant de la norme du cluster sont les candidats idéaux pour une stratégie d'investissement. L'analyse se

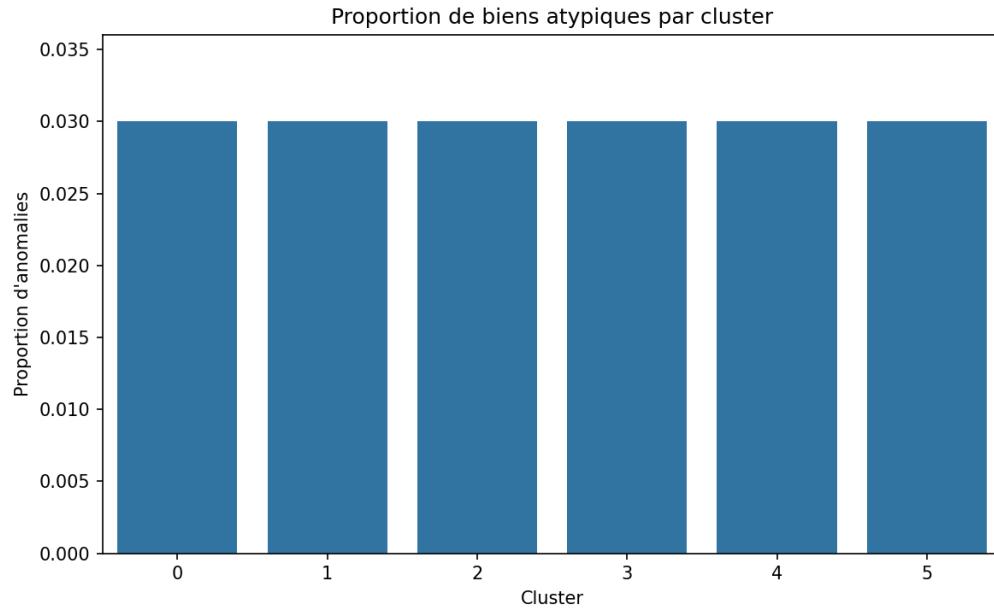
concentre spécifiquement sur **les anomalies basses (prix inférieur au modèle)**, qui signalent de potentielles opportunités d'achat.

Figure 6.1 : Scatter prix/surface



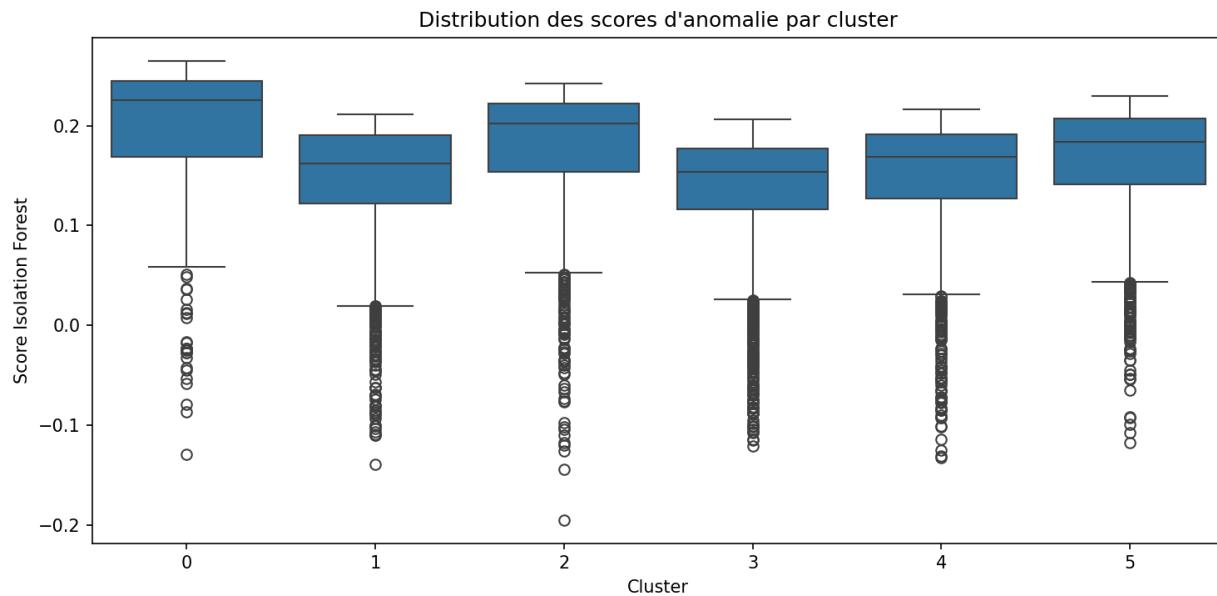
Les biens atypiques se distinguent par un prix au m² anormalement élevé ou faible au regard de leur surface, relativement aux biens de leur cluster.

Figure 6.2 : Proportion d'anomalie par cluster



La proportion d'anomalies reste stable entre les clusters, ce qui confirme que la détection n'est pas biaisée par un segment particulier du marché.

Figure 6.3 : Score d'anomalie



Les distributions montrent une variabilité interne aux clusters, indiquant que l'atypicité est définie localement et non globalement.

Tableau : Détection des anomalies extrême

Cluster	Commune	Type	Surface	Prix m ²	Moyenne Cluster	Décote
0	Paris 14	Appartement	99 m ²	303€	10 211€	97.0%
0	Paris 16	Appartement	36 m ²	306€	10 211€	97.0%
0	Paris 10	Appartement	64 m ²	312€	10 211€	96.9%
0	Paris 16	Appartement	59 m ²	314€	10 211€	96.6%
0	Paris 08	Appartement	35 m ²	314€	10 211€	96.9%

Ce tableau présente les anomalies les plus fortes détectées par l'algorithme Isolation Forest (scores d'anomalie les plus élevés).

Nous observons des décotes extrêmes (> 95 %) sur le Cluster 0 (Paris), avec des prix au m² autour de 300 €, contre une moyenne de 10 211 €. D'un point de vue Data Science, ces points sont correctement identifiés comme des *outliers* majeurs. D'un point de vue Métier, un tel écart de prix ne correspond pas à une simple opportunité de marché mais signale probablement :

1. **Des transactions atypiques** : Ventes en viager rachats de parts indivises ou cessions intra-familiales.
2. **Des artefacts de données** : Erreurs de saisie dans la base DVF originale ou confusion sur la nature du lot.

Conclusion sur la méthode : Ce résultat valide la capacité de l'Isolation Forest à isoler purement et simplement les transactions qui ne respectent pas la logique de marché standard. Pour une utilisation industrielle, ces alertes serviraient de filtre de "Data Quality" avant d'être transmises aux équipes d'investissement.