ANALYSE DE L'ÉVOLUTION DES PRIX DE L'IMMOBILIER À PARIS

GIULIA GOVERNATORI - ESN DATA

LES OBJECTIFS DE CETTE ANALYSE

Les Plus Beaux Logis de Paris

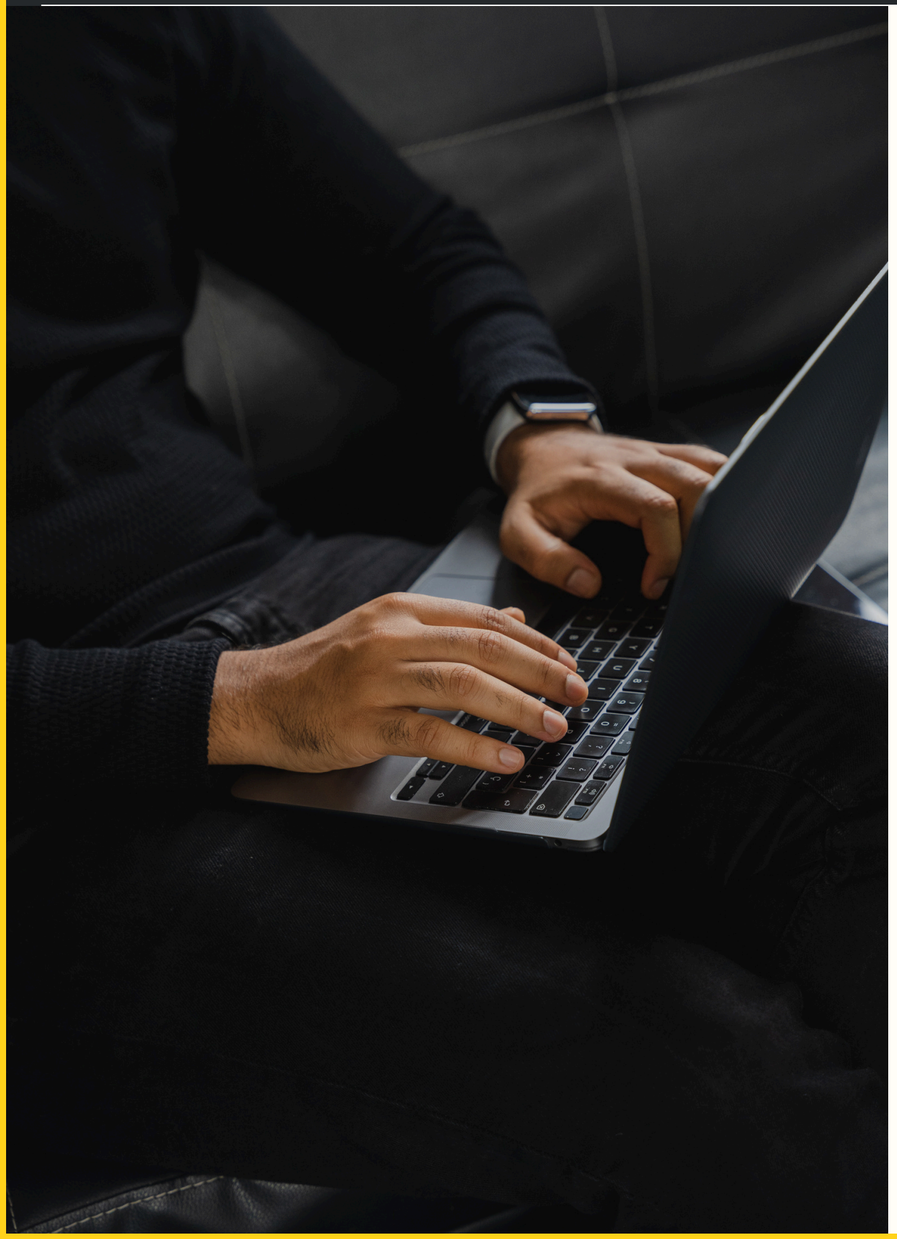
Les Plus Beaux Logis de Paris nous a sollicités pour mener une analyse approfondie de son portefeuille immobilier.

Deux objectifs principaux ont été définis:

1. Identifier **les biens les moins rentables** et prédire leur valorisation future afin d'ajuster la **stratégie d'investissement.**

2. Développer **un algorithme de classification automatique** pour distinguer les appartements des locaux commerciaux, sur la base de leurs caractéristiques, notamment leur **prix**.

I. ANALYSE DU MARCHÉ IMMOBILIER (2017–2021)



DONNÉES UTILISÉES

L'analyse repose sur deux sources principales :
- Un historique des ventes immobilières à Paris entre 2017 et 2021 (**df_histo_immo**)
- Le portefeuille d'actifs immobiliers de l'entreprise en 2022 (**df_actifs**)

Un nouveau fichier, mis à jour, relatif au portefeuille a récemment été transmis.

👉 **Deux prédictions distinctes** ont donc été réalisées, une pour chaque fichier.

Les données reçues étaient de bonne qualité:

aucune valeur manquante
ni anomalie majeure n'a été détectée.

📌 Seuls quelques outliers ont été supprimés afin d'améliorer la robustesse du modèle :

- Méthode du Z-score (**seuil : 3**)
- Méthode de l'IQR (**seuil : 2**)

TENDANCES GLOBALES

Pour comparer les biens, nous avons introduit une métrique clé:

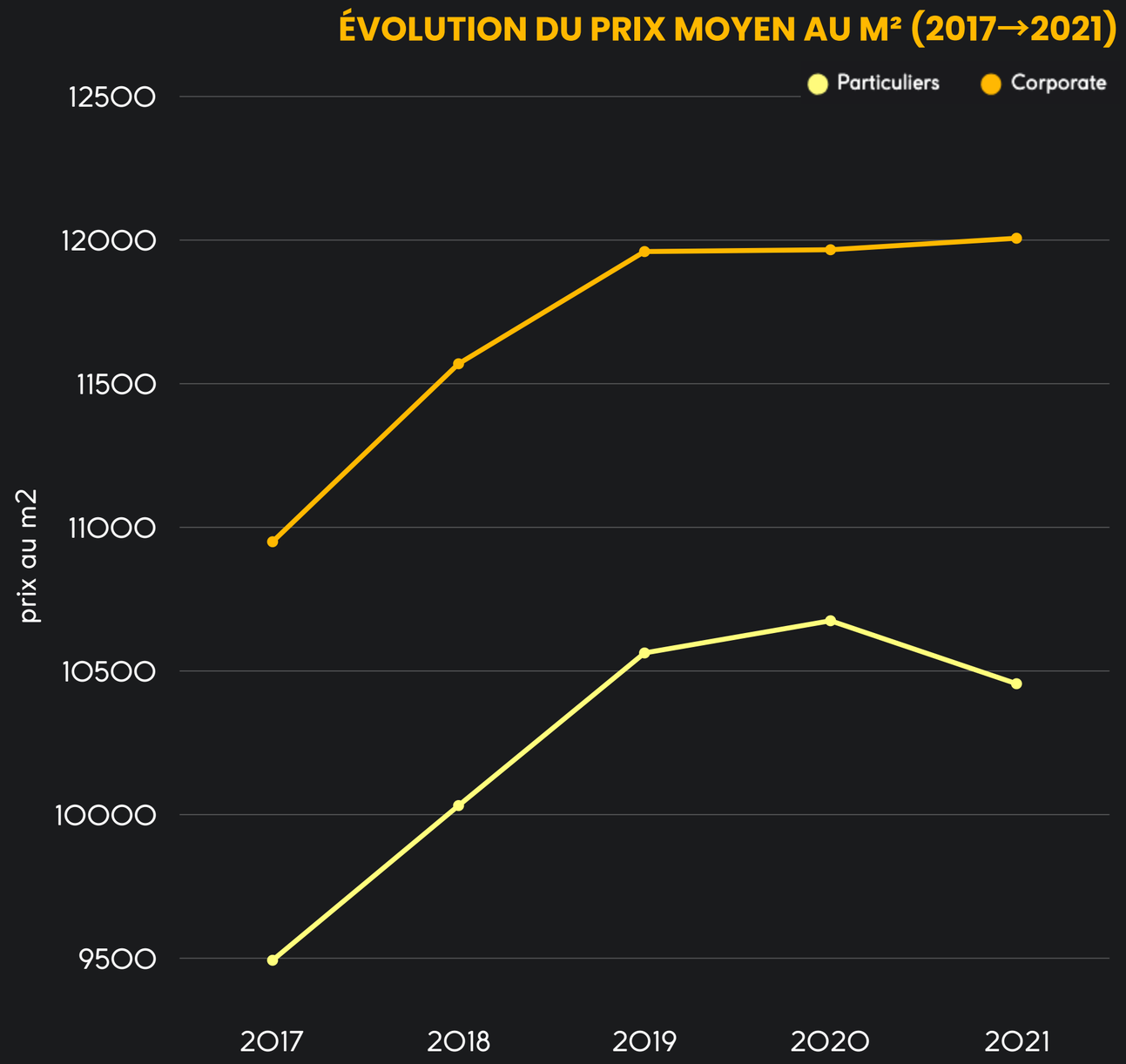
👉 le **prix au mètre carré**, calculé pour chaque bien.

Les données ont ensuite été agrégées **par année** afin d'analyser l'évolution des prix moyens à Paris.

📈 Résultat:

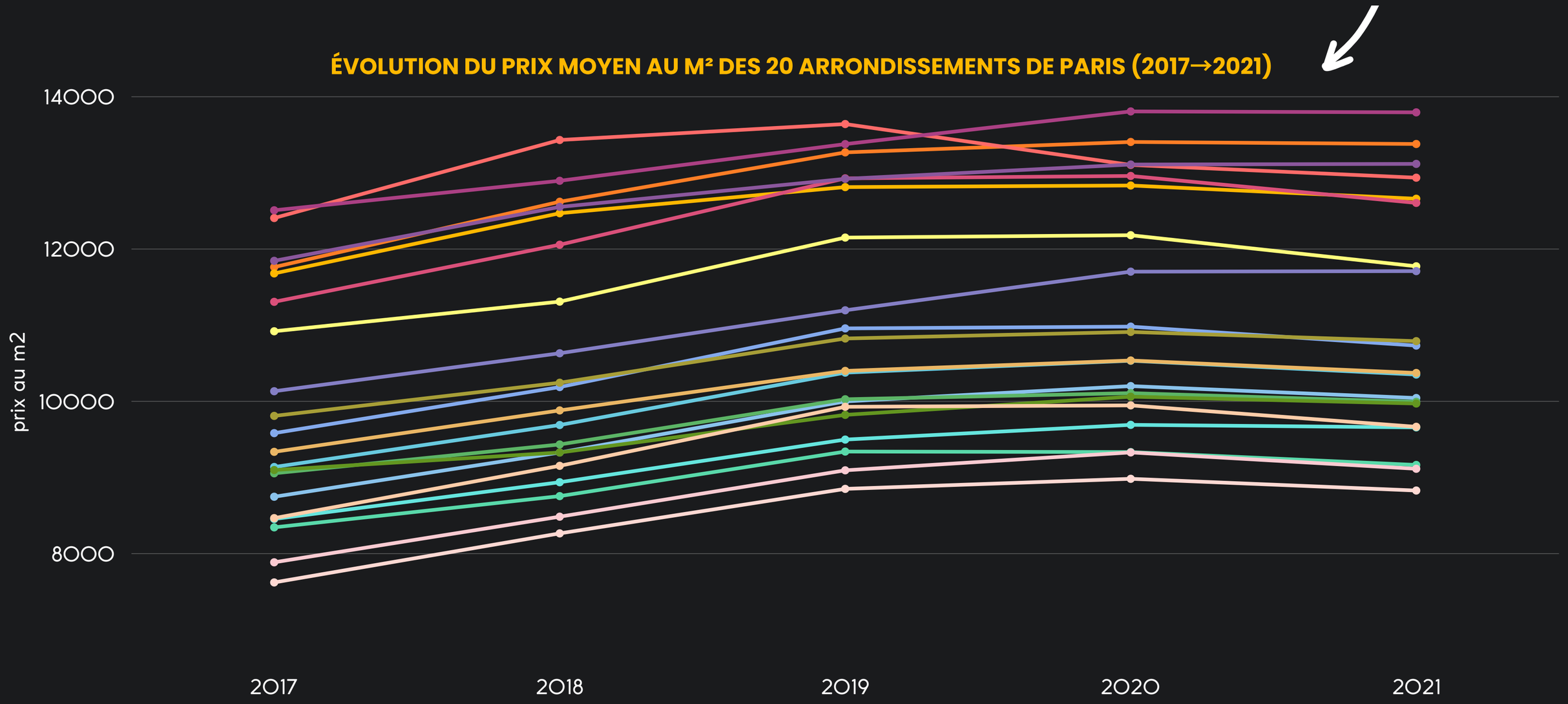
Une **hausse** générale entre 2017 et 2020, suivie d'un léger **recul** en 2021.

Ici, on peut voir la différence selon le **type de bien**.



VARIATIONS PAR ARRONDISSEMENT

… et ici, les différences selon les **arrondissements.**



REPRÉSENTATIVITÉ DES DONNÉES

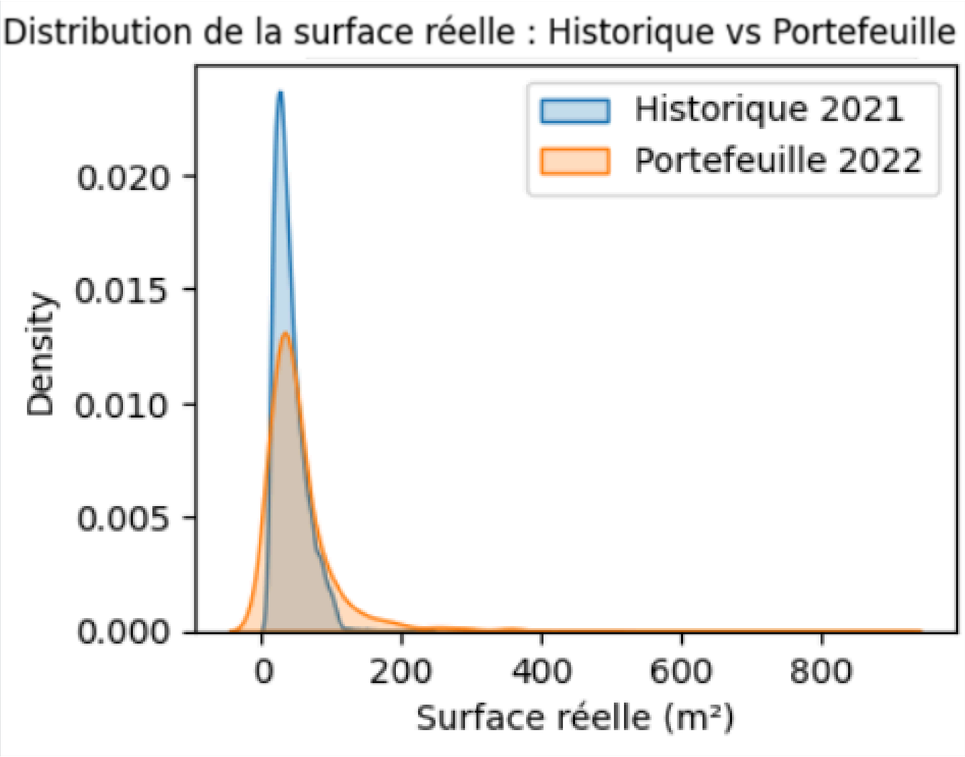
🔹 Distribution des surfaces:

Le portefeuille 2 présente des surfaces beaucoup plus extrêmes (ouliers à droite):

- ***Skewness : 6,02 | Kurtosis : 57,74***

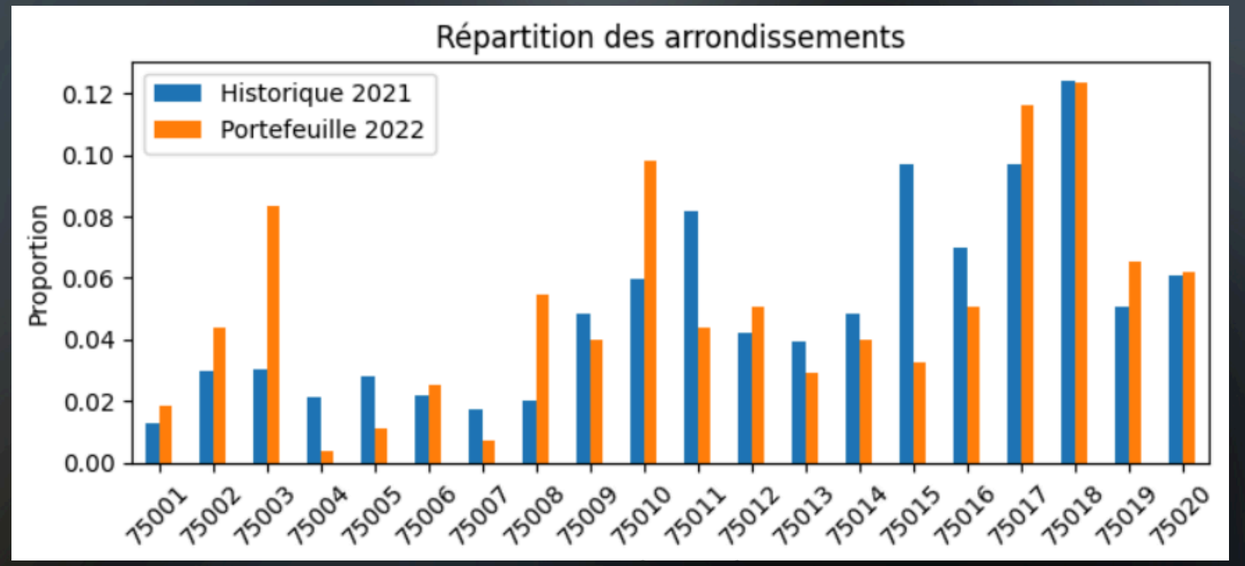
(distribution de la surface réelle du 2eme fichier du portefeuille 2022)

👉 Ce qui peut augmenter l'incertitude pour certains biens très atypiques, et expliquer en partie pourquoi la seconde prédiction est nettement plus élevée.



🔹 Répartition géographique:

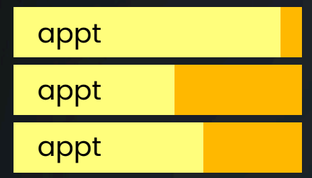
Certaines zones sont sous-représentées dans l'historique, ce qui peut affecter la précision dans ces arrondissements.



🔹 Répartition des types de biens:

Les biens du portefeuille sont plus diversifiés que ceux de l'historique :

- Historique 2021: **92,9%** appt | **7,1%** locaux
- Portefeuille 1: **56,0%** appt | **44,0%** locaux
- Portefeuille 2: **66,1%** appt | **33,9%** locaux



CORRÉLATIONS CLÉS:

1. TEMPORELLE

Pearson

Le coefficient de corrélation de Pearson est de **0,904,** ce qui indique une forte corrélation positive entre la *date de mutation* (c'est-à-dire le temps qui passe) et le prix au mètre carré.

2. SURFACE RÉELLE

Le coefficient de corrélation de Pearson est de **0,980**: cela signifie que, de manière générale, plus un appartement est grand, plus son prix de vente est élevé.

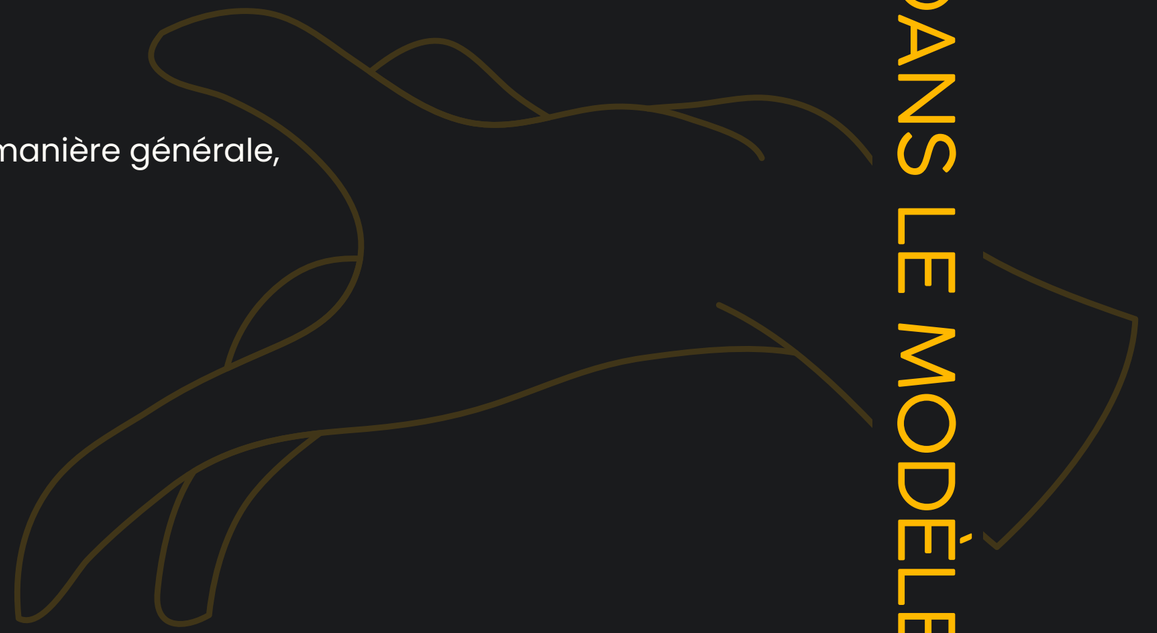
3. GÉOGRAPHIQUE

Graphiques

Comme nous avons pu le constater dans le graphique en courbes représentant les 20 arrondissements.

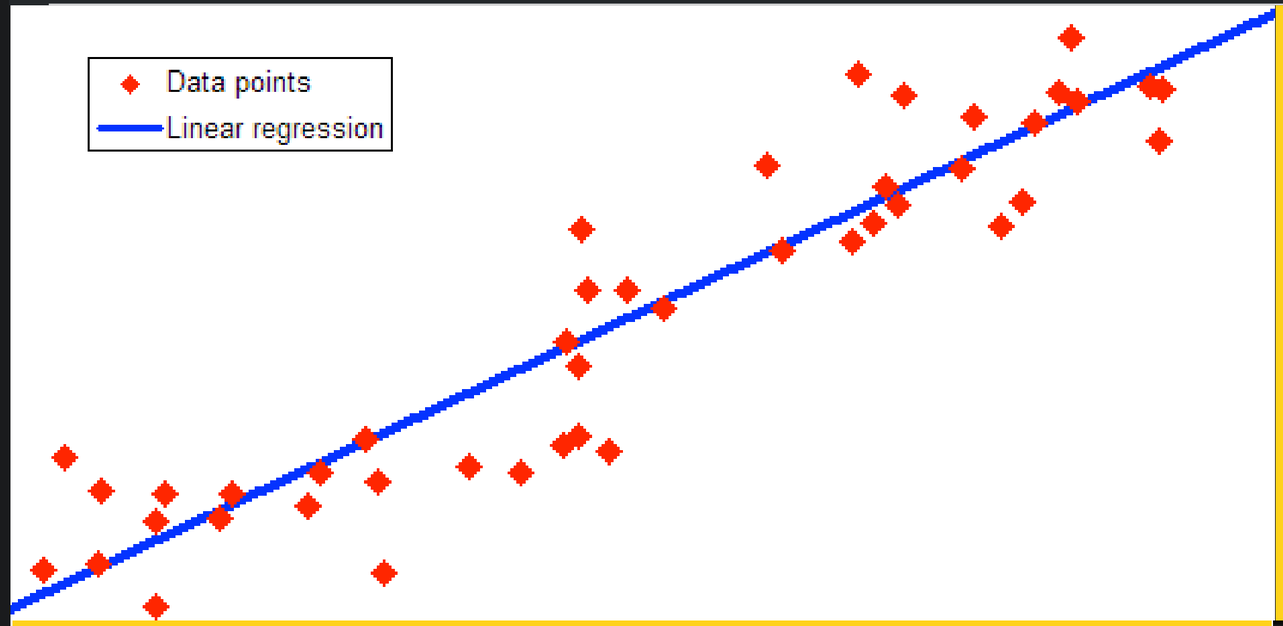
4. TYPE DE BIEN

Comme nous avons pu le voir dans le graphique précédent.



INCLUS DANS LE MODÈLE

II. MÉTHODOLOGIE DE LA RÉGRESSION LINÉAIRE



Régression Linéaire:

La régression linéaire est un modèle qui cherche à tracer une **ligne droite** qui permet de prédire au mieux les données. Cette ligne **sert à faire des prédictions.**

Mais « *au mieux* », ça veut dire quoi? Ça veut dire avec **le moins d'erreurs possible** entre les valeurs prédites et les valeurs réelles.

Techniquement, elle minimise l'erreur quadratique moyenne (**MSE** - *Mean Squared Error*), c'est-à-dire la distance (au carré) entre les points réels et ceux prédits.

Cependant, ce modèle présente des limitations, notamment une sensibilité aux valeurs aberrantes, et il suppose que la relation entre les variables est linéaire.

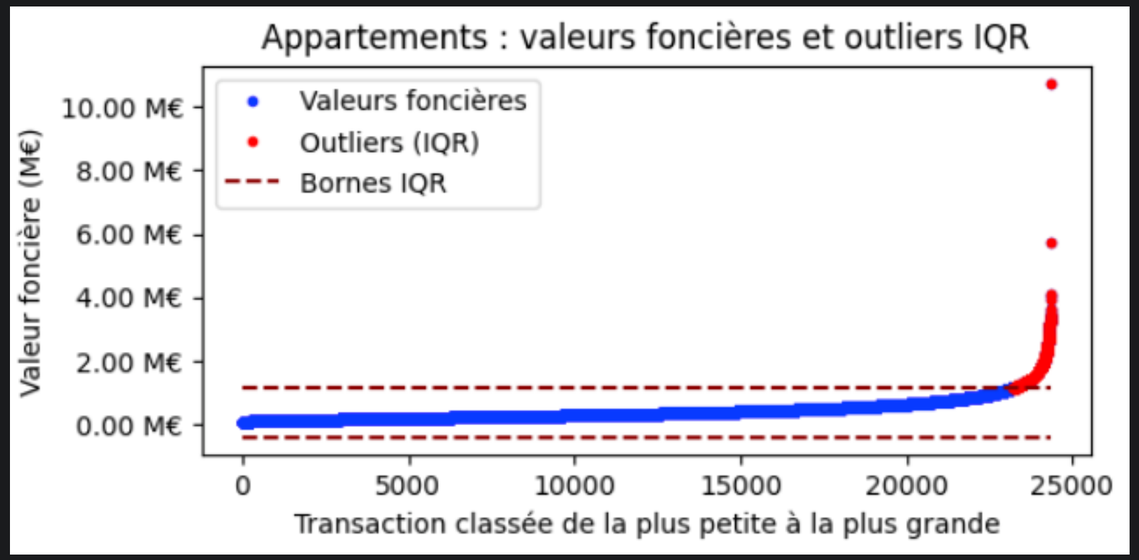
Minimiser les biais et garantir des prédictions fiables:

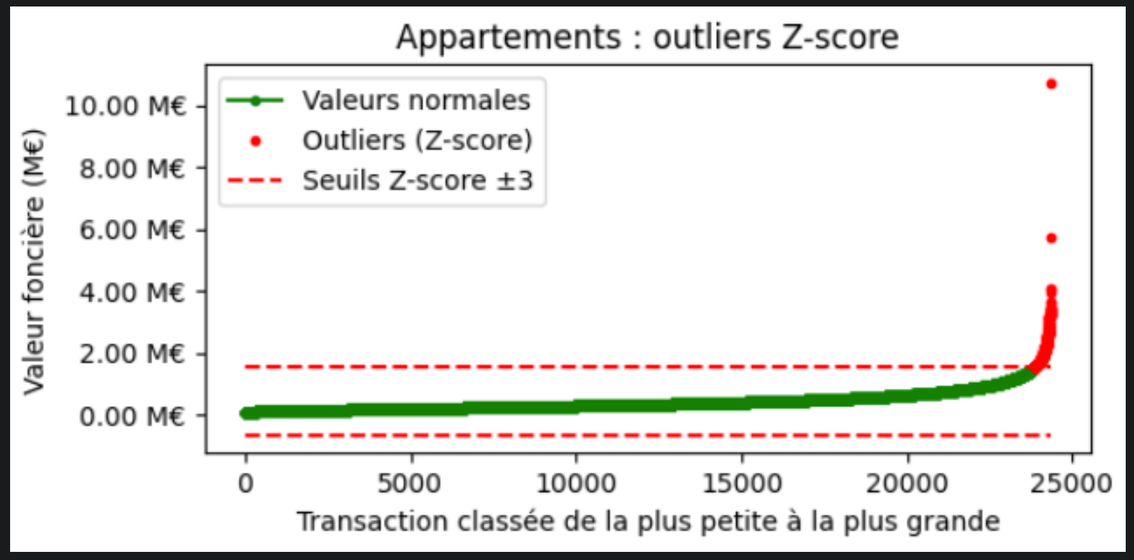
Pour garantir un modèle robuste et fiable, **quatre actions clés** ont été mises en place:

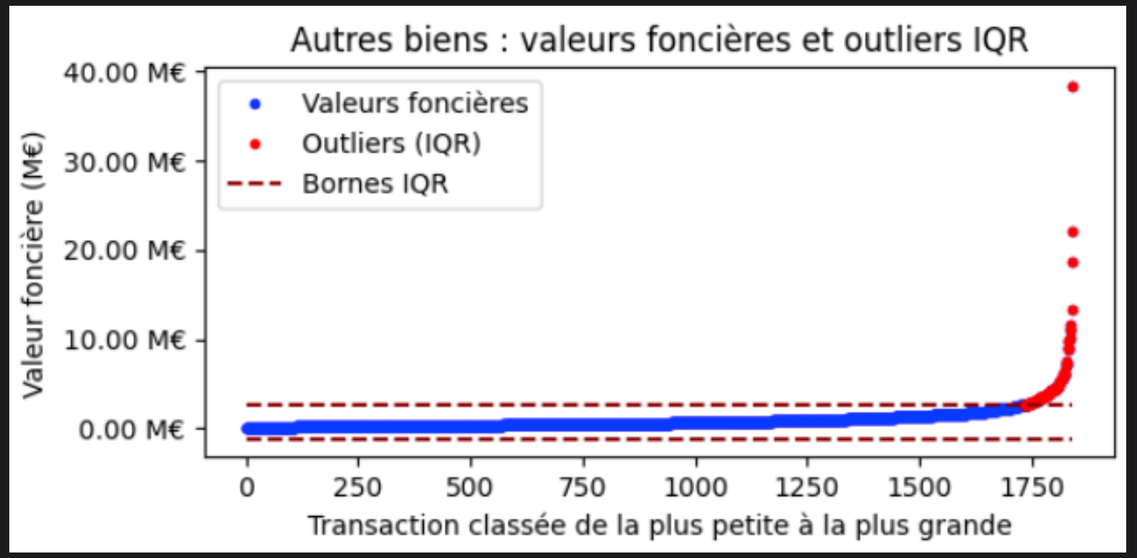
- **1. Nettoyage des outliers**: suppression des valeurs aberrantes via Z-score et IQR
- **2. Transformation logarithmique:** normalisation de la variable cible pour un apprentissage optimal.
- **3. Segmentation par type de bien** modèles distincts pour les appartements et autres biens
- **4. Séparation temporelle:** données jusqu'en 2020 pour l'entraînement, 2021 pour le test.

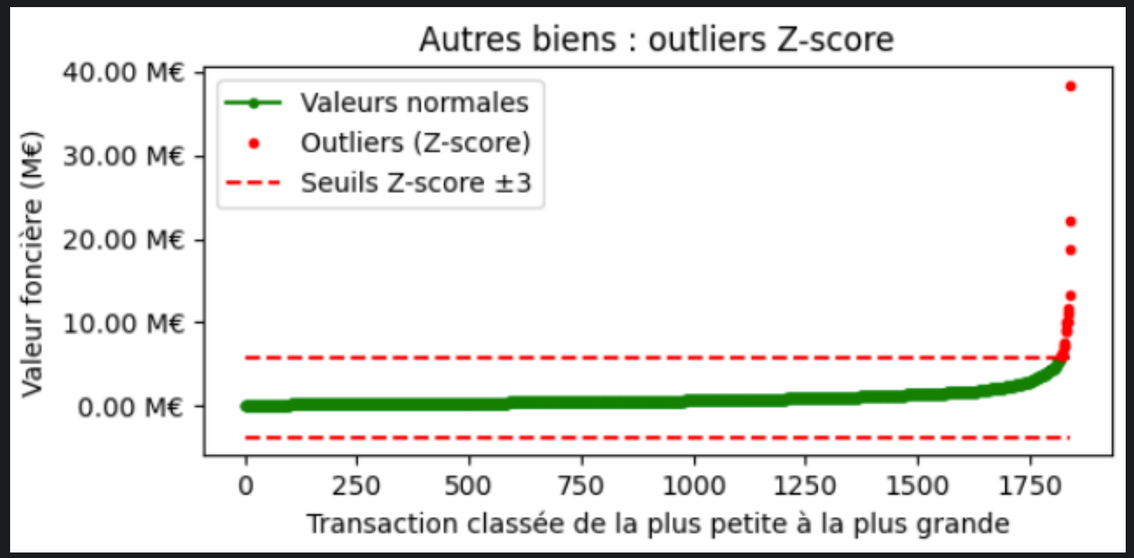
NETTOYAGE DES OUTLIERS

```
Appartements : éliminés 1253 sur 24353 transactions (5.15 %)
Autres biens : éliminés 117 sur 1843 transactions (6.35 %)
```







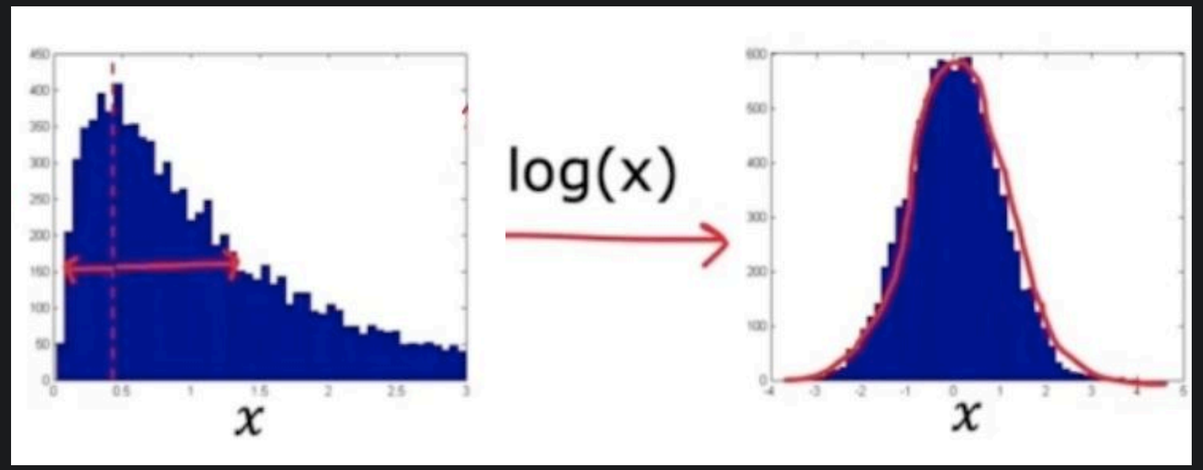


TRANSFORMATION LOGARITHMIQUE (DE LA VARIABLE CIBLE)

La transformation logarithmique permet de rendre la distribution de la variable cible plus normale et symétrique, ce qui améliore l'efficacité de la régression linéaire, parce qu'elle fonctionne mieux quand les données sont normalement distribués.

Cela réduit l'impact des valeurs extrêmes et facilite l'apprentissage du modèle.

Après la prédiction, les valeurs sont de nouveau converties à leur échelle d'origine pour une interprétation correcte.



Cela permet de rendre la distribution initiale aussi proche que possible d'une distribution normale.

La transformation logarithmique n'est pas adaptée à tous les types de données:

Si les données contiennent des zéros ou des valeurs négatives, elle ne peut pas être appliquée directement.

Ce que fait la transformation logarithmique:

1. Réduit l'asymétrie de la variable cible
2. Rend la relation entre les variables plus linéaire
3. Stabilise la variance: moins outliers
4. Améliore la performance du modèle

AVANT: valeurs de la variable **x** → [2000, 2500, 3000, 12000, 15000]
Distribution: ●●●● ●———●———●

APRÈS: **log(x)** → [7.6, 7.8, 8.0, 8.3, 9.4, 9.6]
Distribution: ●●●●●●●●● ●

🎯 SÉPARATION TEMPORELLE DES DONNÉES: jusqu'à 2020 → *train* - 2021 → *test*

📌 Pourquoi ce choix est pertinent?

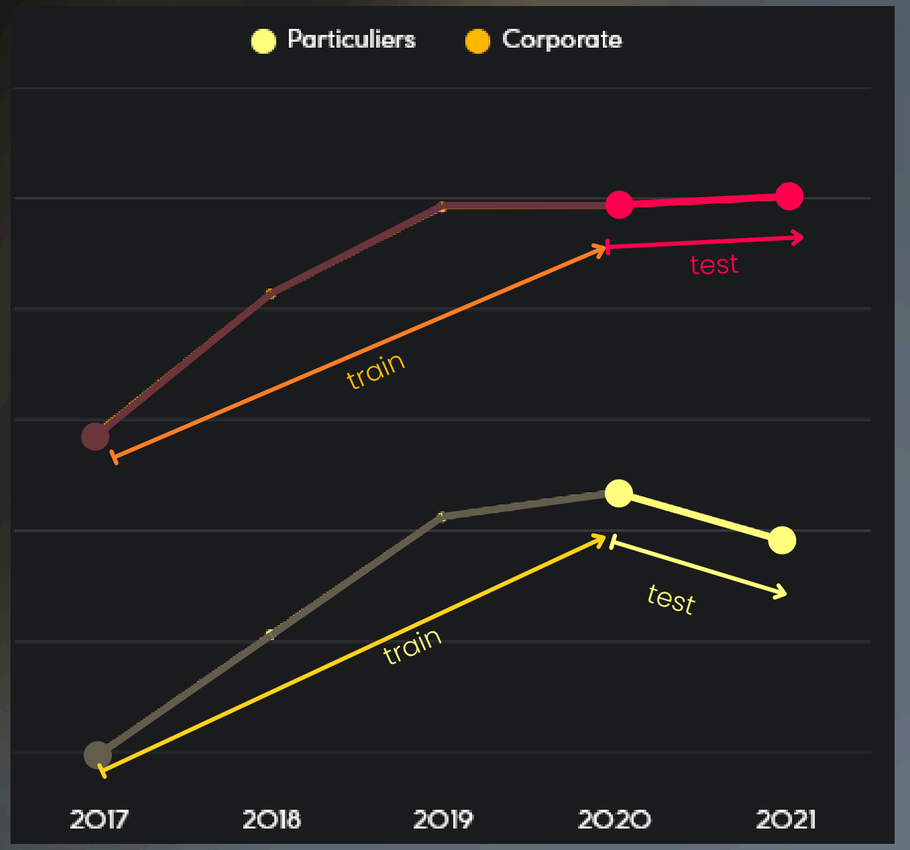
- **Simule un vrai scénario de prédiction**
  → On utilise le passé pour prévoir le futur.
- **Respecte la chronologie**
  → Le modèle ne voit que ce qui était disponible au moment de la prédiction.

📌 Et 2021? Un cas particulier...

→ **Oui,** car le modèle est **entraîné** sur une tendance «***normale***»,
mais il est **testé** sur une année ***atypique***: **2021**.

➡ Le marché y a connu un **<u>comportement inhabituel</u>**
(possible effet post-Covid, réajustement économique...)

🧪 C'est donc **un vrai stress test** pour le modèle.
👉 Cela soulève une question importante:



*Peut-il rester **fiable***
*face à un changement de **contexte**?*

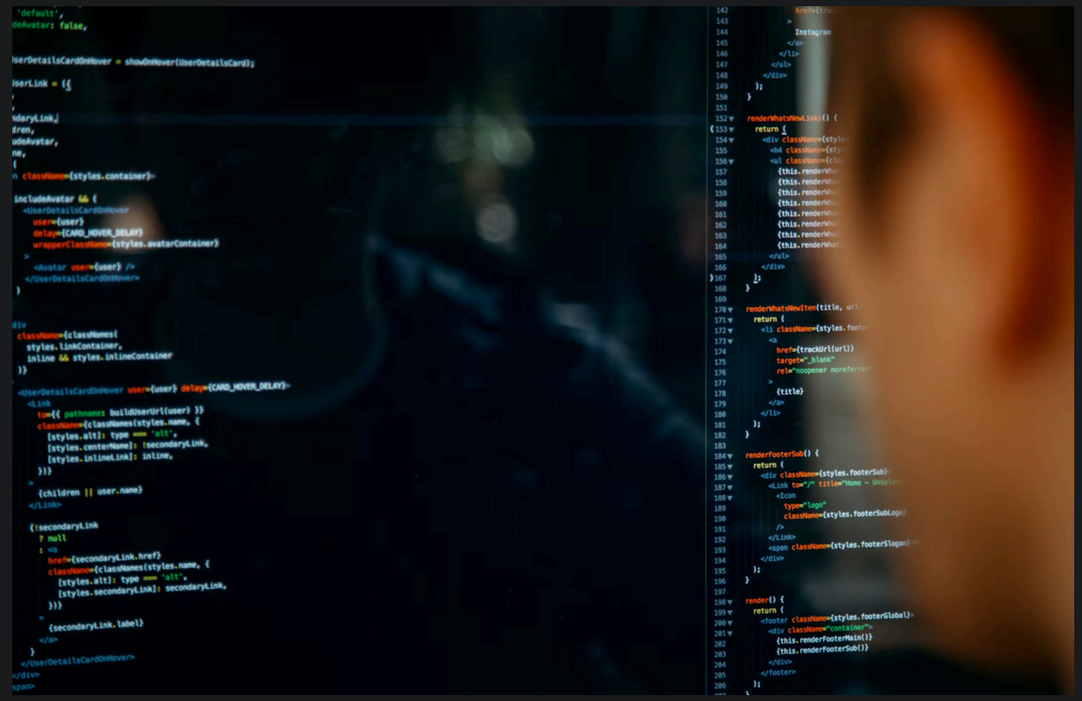
📌 Le modèle reste-t-il fiable malgré tout?

✅ **Oui** , s'il s'agit de tester sa capacité à généraliser à un **contexte nouveau** .
Si l'erreur augmente, ce n'est pas un échec, mais un signal utile.
Le modèle sera plus flexible et capable de s'adapter à des contextes nouveaux.

ENCODAGE ET VARIABLES

Nous avons appliqué un ***one-hot encoding*** aux variables code_postal. Cette méthode transforme chaque valeur unique d'une catégorie en une colonne binaire (**0** ou **1**).

Chaque code postal distinct devient une nouvelle colonne, et un "**1**" est attribué à la ligne correspondant à ce code, tandis que les autres colonnes auront un "**0**".



La *"machine"* a besoin d'une *"traduction"*.

Un example:

code_postal_75006	code_postal_75007	code_postal_75008	code_postal_75009	code_postal_75010	code_postal_75011	code_postal_75012	code_postal_75013
False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	False

III. RÉSULTATS DES PRÉDICTIONS

VALORISATION (AU 31/12/2022) :

fichier 1 : *Corporate* = **62,4 %**
fichier 2 : *Corporate* = **98,5 %**

** de la valorisation*

fichier 1 : *Corporate* = **44 %**
fichier 2 : *Corporate* = **33,9 %**

** des actifs 2022*

→ Portefeuille 2022 - fichier 1 :

Particuliers (***154*** *biens*) : **67,5 M€** | *Corporate* (***121*** *biens*) : **111,9 M€**

→ Portefeuille 2022 - fichier 2 :

Particuliers (***310*** *biens*) : **788,5 M€** | *Corporate* (***159*** *biens*) : **52 454,8 M€**

Pourquoi une valeur nettement plus élevée pour le deuxieme fichier?

✅ Portefeuille beaucoup plus volumineux (275 articles VS 469 articles)
✅ Forte proportion de locaux commerciaux (7,1 % vs 33,9 %)
✅ Biens aux caractéristiques atypiques (très grandes surfaces, forte asymétrie)
→ Risque accru de surestimation ponctuelle

🔎 Ces facteurs peuvent expliquer une estimation globale bien plus élevée, tout en restant cohérente avec la logique du modèle.

PERFORMANCES DES MODÈLES

🔍 Performances des modèles

SEGMENT	MAPE (%) TRAIN/TEST	MAE (LOG) Train / Test	RMSE (LOG) Train / Test
Appartements	1,05 % / 1,08 %	0,13 / 0,14	0,17 / 0,16
Locaux	1,47 % / 1,48 %	0,19 / 0,20	0,24 / 0,24

💡 Erreur moyenne pondérée:

Ces résultats montrent une excellente généralisation du modèle avec une faible marge d'erreur.

1,08 % train
1,11 % test

MÉTHODOLOGIE K-MEANS

🎯 Contexte & besoin métier:

Besoin exprimé par Louise:
automatiser la qualification des actifs.

🛠️ Préparation des données:

- Calcul du prix au m² pour chaque bien.
- Nettoyage des valeurs manquantes.
- Suppression des colonnes non pertinentes à la segmentation.

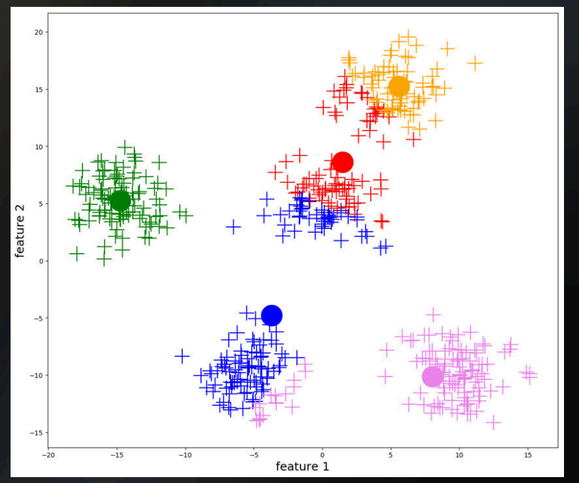
📊 Clustering K-Means:

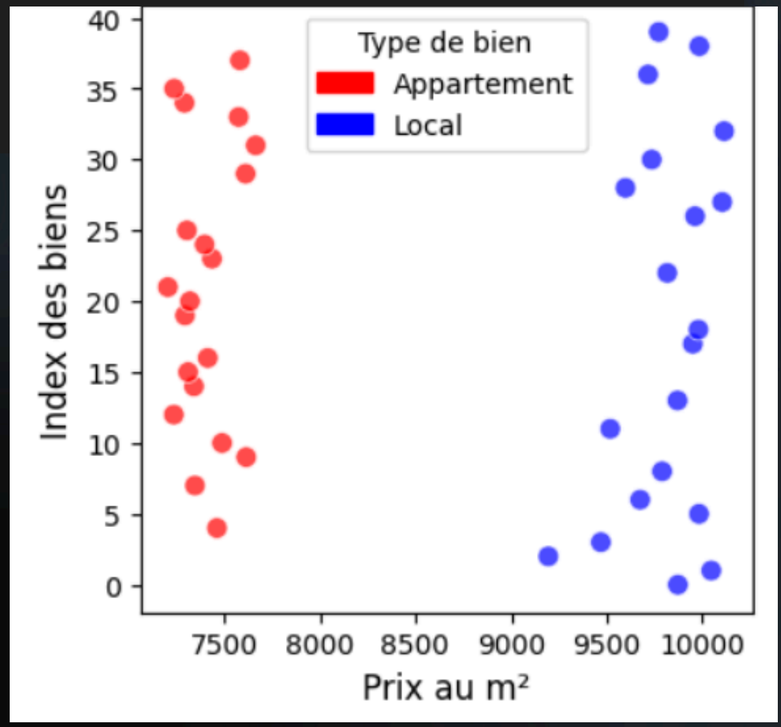
- Algorithme appliqué sur la variable “prix au m²”
- Nombre de clusters : 2
- Moyenne de chaque groupe calculée pour différencier les segments.

```python
# Calcul du prix au mètre carré
df_kmeans["prix_m2"] = (
    df_kmeans["valeur_fonciere"] / df_kmeans["surface_reelle"]
)

# Suppression des colonnes inutiles pour l’analyse
df_kmeans_clustering = df_kmeans.drop(
    columns=[
        "valeur_fonciere",
        "surface_reelle",
        "nom_commune"
    ]
)
```

example de comment le k-means fonctionne:





LE K-MEANS EN DÉTAIL

🎯 Attribution des labels aux groupes K-Means:

- Après analyse des moyennes de chaque cluster **(1)** doivent etre attribues **(2)**:
  - Cluster à faible prix/m² → majoritairement des **appartements**
  - Cluster à prix élevé/m² → correspond principalement à des **locaux commerciaux**

🧠 Ces correspondances permettent de relier les groupes identifiés automatiquement à les catégories *Appartements vs Locaux*.

(1)

```
# Moyenne du prix au m² par cluster
df_kmeans_clustering.groupby("cluster")["prix_m2"].mean()
```

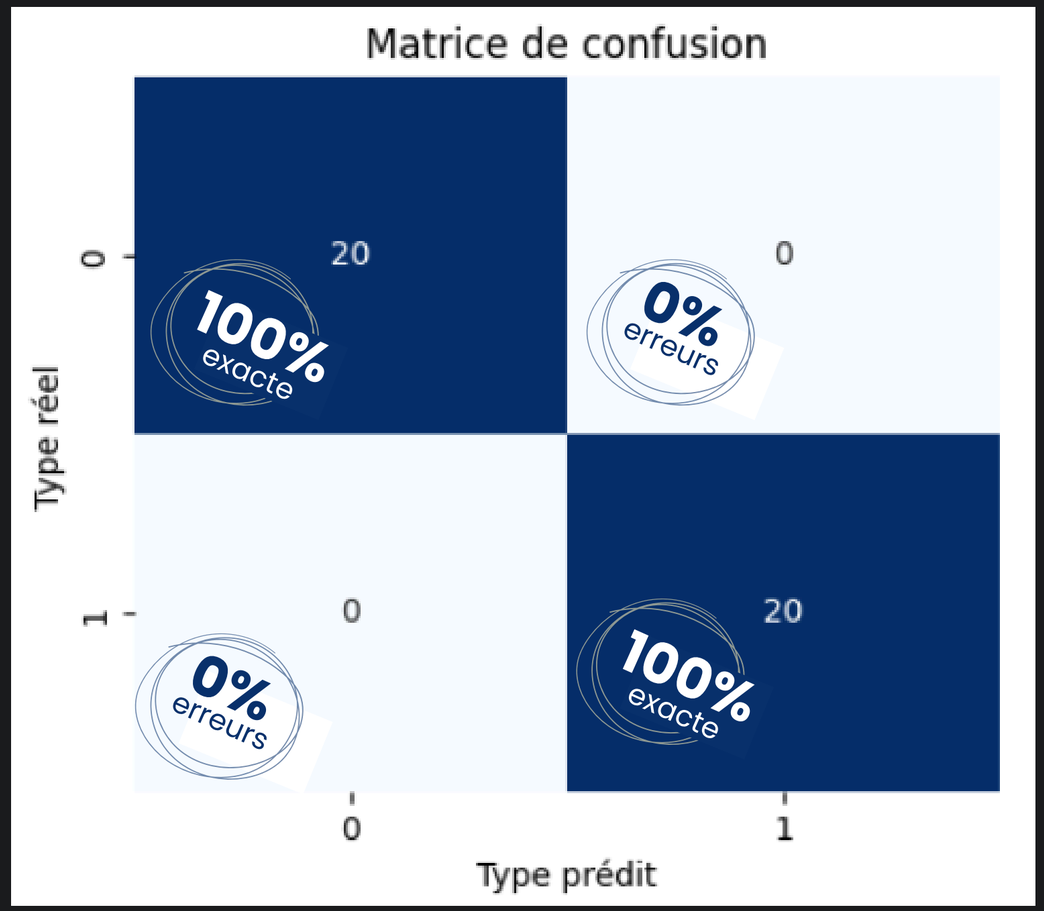
cluster	prix_m2
0	7408.78
1	9806.92

(2)

```
# Dictionnaire de correspondance
mapping = {
    0: "Appartement",
    1: "Local industriel, commercial ou assimilé"
}
df_kmeans_clustering["type_bien_pred"] = df_kmeans_clustering["cluster"].map(mapping)
```

RÉSULTATS DU K-MEANS:
une classification parfaite

🧠 Matrice de confusion:



Tous les indicateurs montrent **une performance parfaite**.

🔹 Précision (1.00) :

parmi tous les biens que le modèle a classés comme Appartement ou Local, 100% étaient corrects. **Cela évite les faux positifs.**

🔹 Rappel (1.00) :

parmi tous les vrais Appartements ou Locaux présents dans les données, 100% ont été correctement identifiés par le modèle. **Cela évite les faux négatifs.**

👉 Ces deux mesures sont **complémentaires**: la *précision* vérifie la fiabilité des prédictions, le *rappel* vérifie si le modèle oublie des cas.

🔹 F1-score (1.00) :

combine *précision* et *rappel* en une seule mesure.
Un score parfait indique **un excellent équilibre entre les deux.**

🔹 Accuracy (100 %) :

le modèle a **tout bien prédit**, sans aucune erreur.

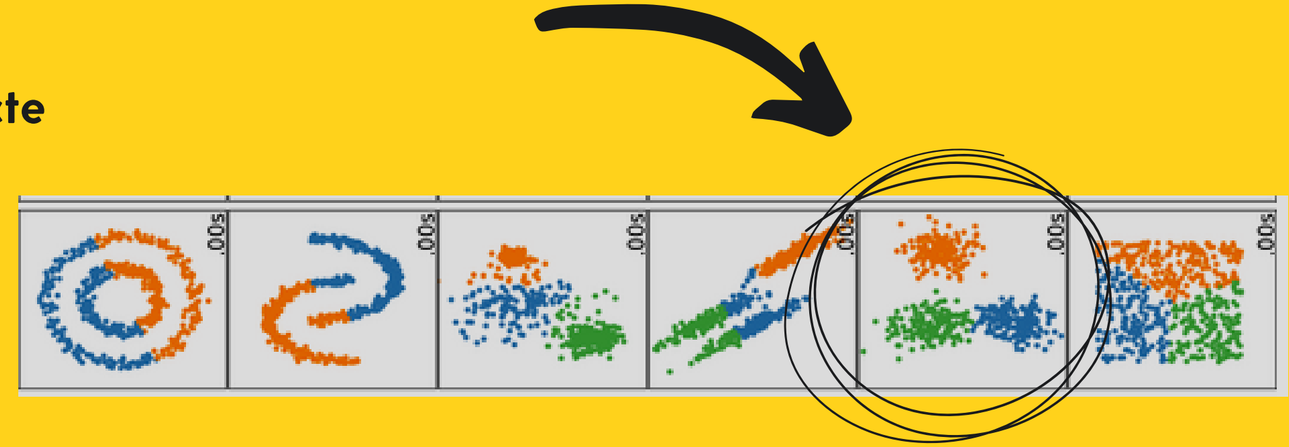
🔹 Matrice de confusion:

tous les biens sont correctement classés,
aucune confusion entre les catégories.

⚠ LIMITES ET RISQUES DU K-MEANS

Le K-Means est efficace quand...

- ✓ **Les groupes sont bien séparés**
- ✓ **Chaque groupe forme une "bulle" compacte**
- ✓ **La densité des groupes est homogène**



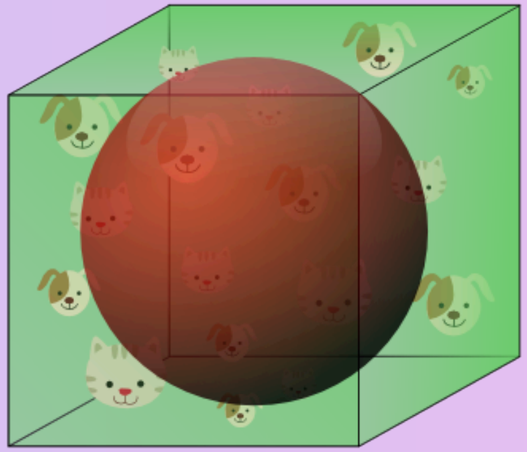
Mais le K-Means présente plusieurs limites :

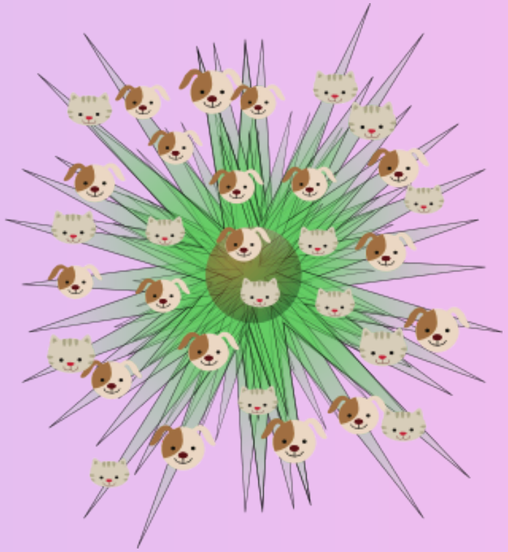
- **Choix arbitraire de k:** → le nombre de clusters doit être fixé à l'avance, ce qui peut biaiser l'analyse.
- **Résultats stochastiques:** → les résultats peuvent changer à chaque exécution selon l'initialisation des centres.
- **Sensibilité aux outliers:** → un seul point extrême peut fortement perturber la formation des clusters.
- **Formes complexes:** → difficile à interpréter quand les données ne sont pas bien séparées ou forment des formes complexes.

🔮 *"CURSE OF DIMENSIONALITY"* 🔮

Plus on ajoute de variables, plus les distances deviennent moins discriminantes, car:

- **Les points semblent tous équidistants** → le modèle n'arrive plus à regrouper efficacement.
- **La densité des données s'aplatit** → les bulles n'existent plus clairement.
- **Le bruit augmente** → le K-Means détecte des patterns non pertinents.





Résultat: le K-Means perd en efficacité et en fiabilité à mesure que le nombre de dimensions augmente.

📌 RECOMMANDATIONS STRATÉGIQUES:

🏷️ Ciblage des ventes:

Vendre en priorité les actifs appartenant au segment le moins porteur, selon les prédictions.
Vendre les 310 appartements, maintenir les 159 locaux commerciaux.

📈 Enrichissement des données:

Ajouter des variables comme l'étage, l'année de construction ou l'état général du bien.

⚙️ Amélioration des modèles:

Tester des algorithmes plus avancés (Random Forest, XGBoost) pour gagner en précision.

🧠 Montée en compétence:

Former les équipes à l'interprétation des résultats pour une meilleure prise de décision.

MERCI je reste à votre écoute pour toute remarque ou question.

Giulia Governatori