

---

# Les Plus Beaux Logis de Paris

## Partie 1



## Les Plus Beaux Logis de Paris

*Analysez l'évolution des prix de l'immobilier avec Python*

*Aurelia*

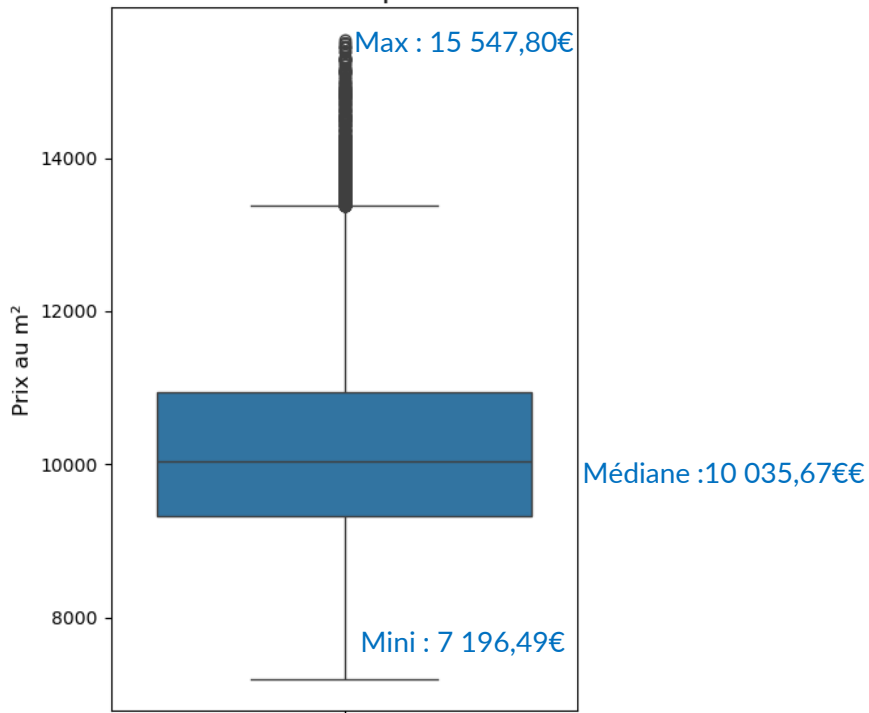
*De Infanti*

*28/03/2025*

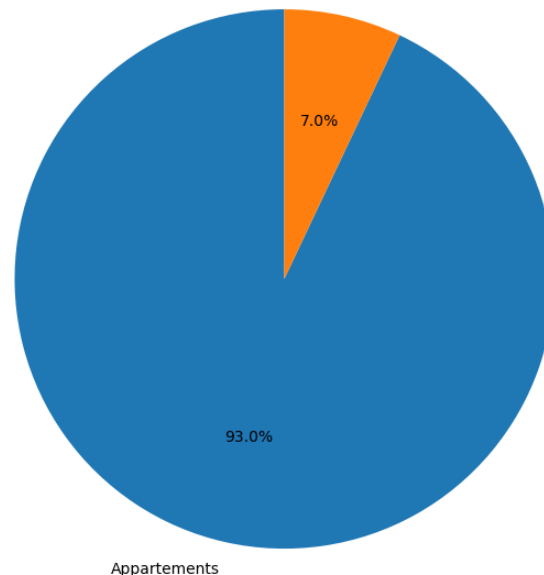
# I. Analyse du marché de l'immobilier

Entre le 02/01/2017 et 31/12/2021

Distribution du prix au m<sup>2</sup>



Répartition des transactions immobilières  
Locaux commerciaux



Nombre de transactions :

Appartements : 24 353

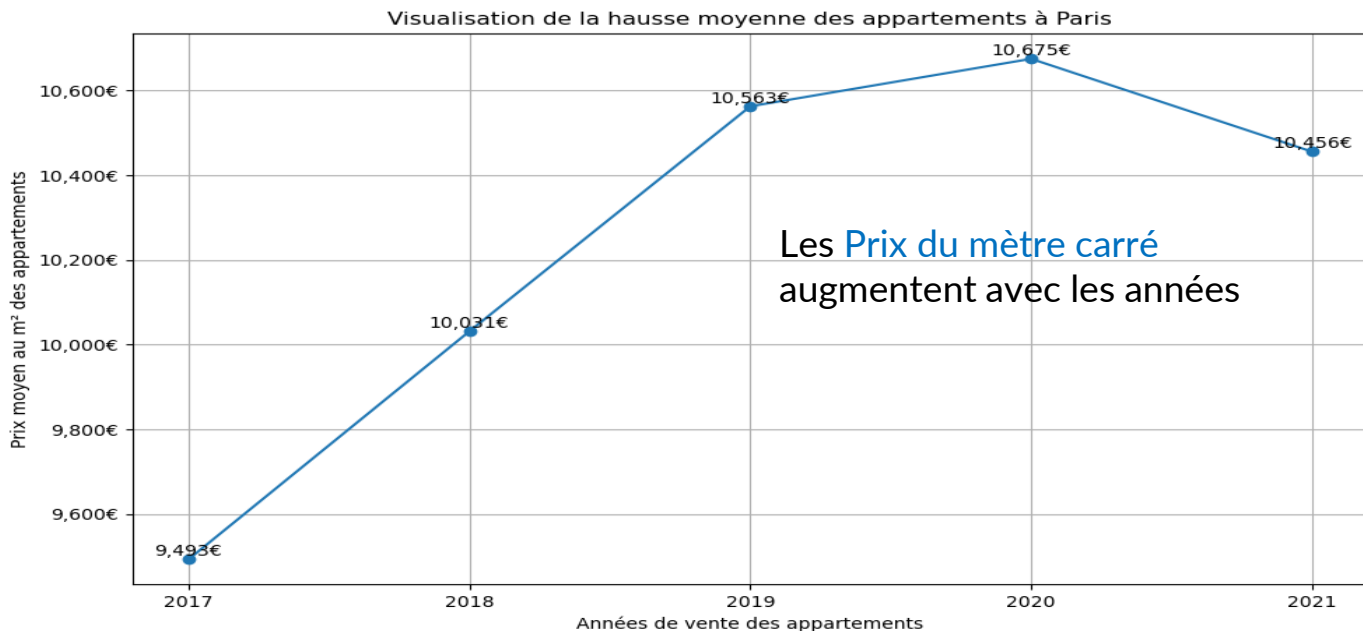
Locaux commerciaux : 1 843

# I. Analyse du marché de l'immobilier

## APPARTEMENTS

En 2017, le prix moyen est le plus bas et la surface moyenne est la plus élevée

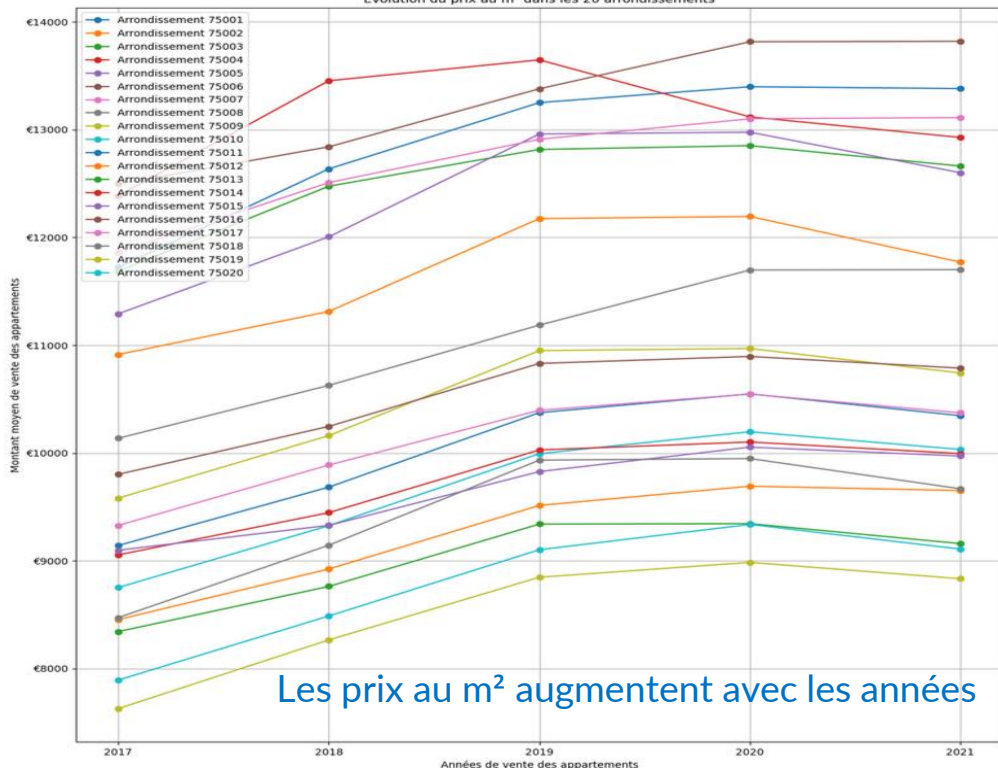
Dès 2020 le prix diminue. Est-ce dû au COVID ?



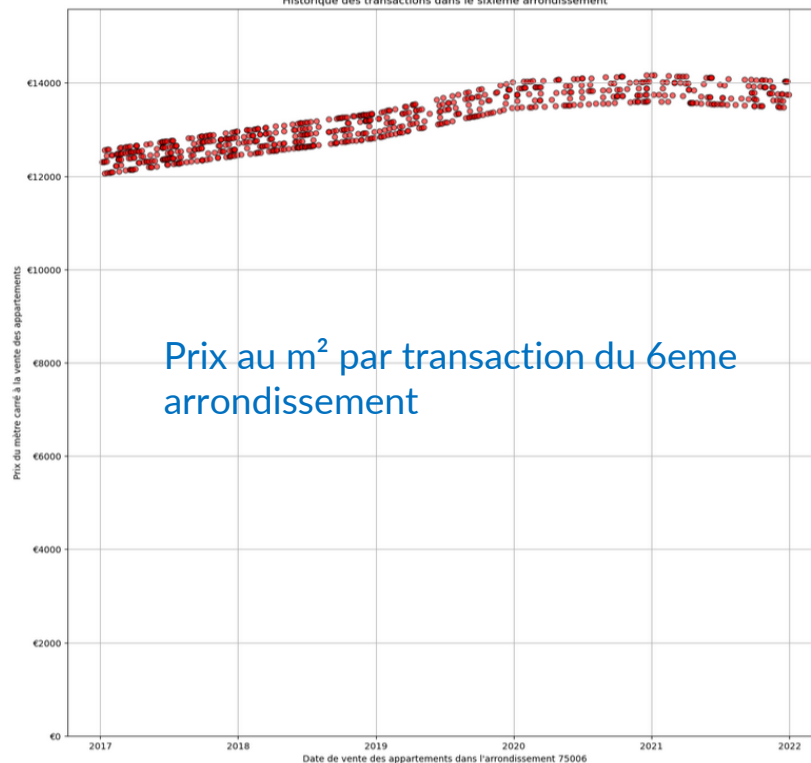
Annee	Prix moyen	Surface moyenne
2017	9 492,86 €	44.63 m2
2018	10 031,40 €	44.27 m2
2019	10 562,71 €	43.36 m2
2020	10 674,87 €	42.90 m2
2021	10 455,60 €	43.48 m2

# I. Analyse du marché de l'immobilier

Evolution du prix au m<sup>2</sup> dans les 20 arrondissements



Historique des transactions dans le sixième arrondissement



# I. Analyse du marché de l'immobilier

Le 25/03/2025 dans le 19<sup>eme</sup>

Prix au m<sup>2</sup> Loyer au m<sup>2</sup>

## Prix immobilier dans le 19<sup>ème</sup> arrondissement de Paris (75019)

Estimations de prix MeilleursAgents au 1 mars 2025. [Comprendre nos prix](#)



Moyenne du prix du mètre carré en 2021 dans le 19<sup>ème</sup> :  
**7408€**

Ceci confirme donc que les prix augmentent en fonction des années

## Évolution du prix de l'immobilier dans le 19<sup>ème</sup> arrondissement de Paris

### Évolution du prix des appartements

1 mois	3 mois	1 an	2 ans	5 ans	10 ans
- 0.2%	+ 0.9%	+ 0.7%	- 9.3%	- 14.0%	+ 16.4%

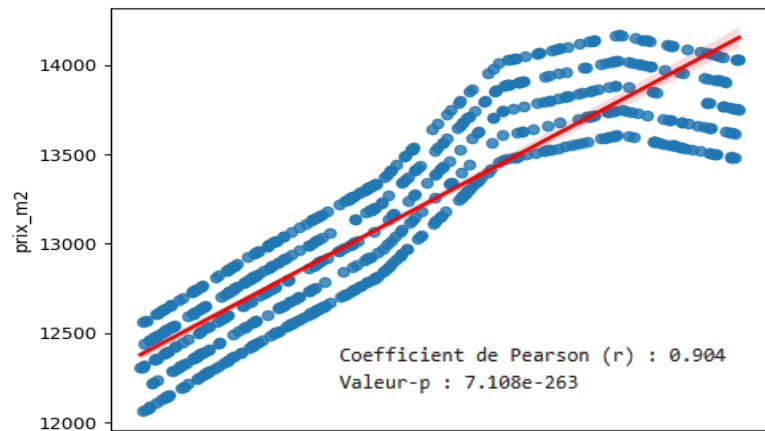


Source : Données MeilleursAgents et données publiques (Notaires, INSEE)

Source : <https://www.meilleursagents.com/prix-immobilier/paris-19eme-arrondissement-75019/>

# I. Analyse du marché de l'immobilier

Relation entre le prix au m<sup>2</sup> et la date dans le 6ème arrondissement

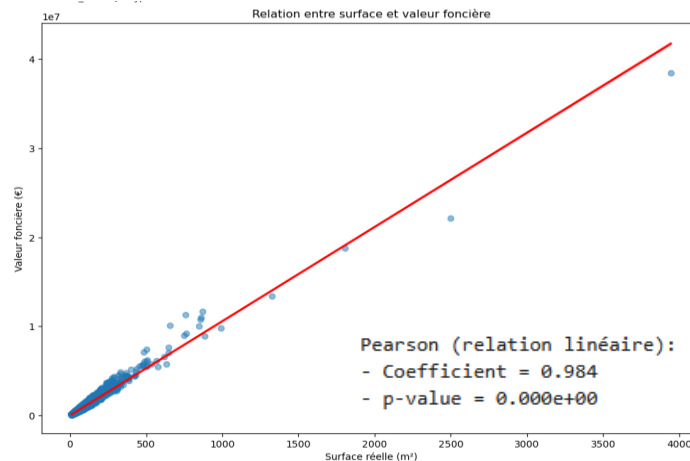


Le coefficient de corrélation est de **0.904** : relation linéaire très forte, avec une pvalue est **inférieure à 0.05**

Confirmation de la corrélation :

**Le temps exerce une influence sur le prix.**

Relation entre la valeur foncière et la surface



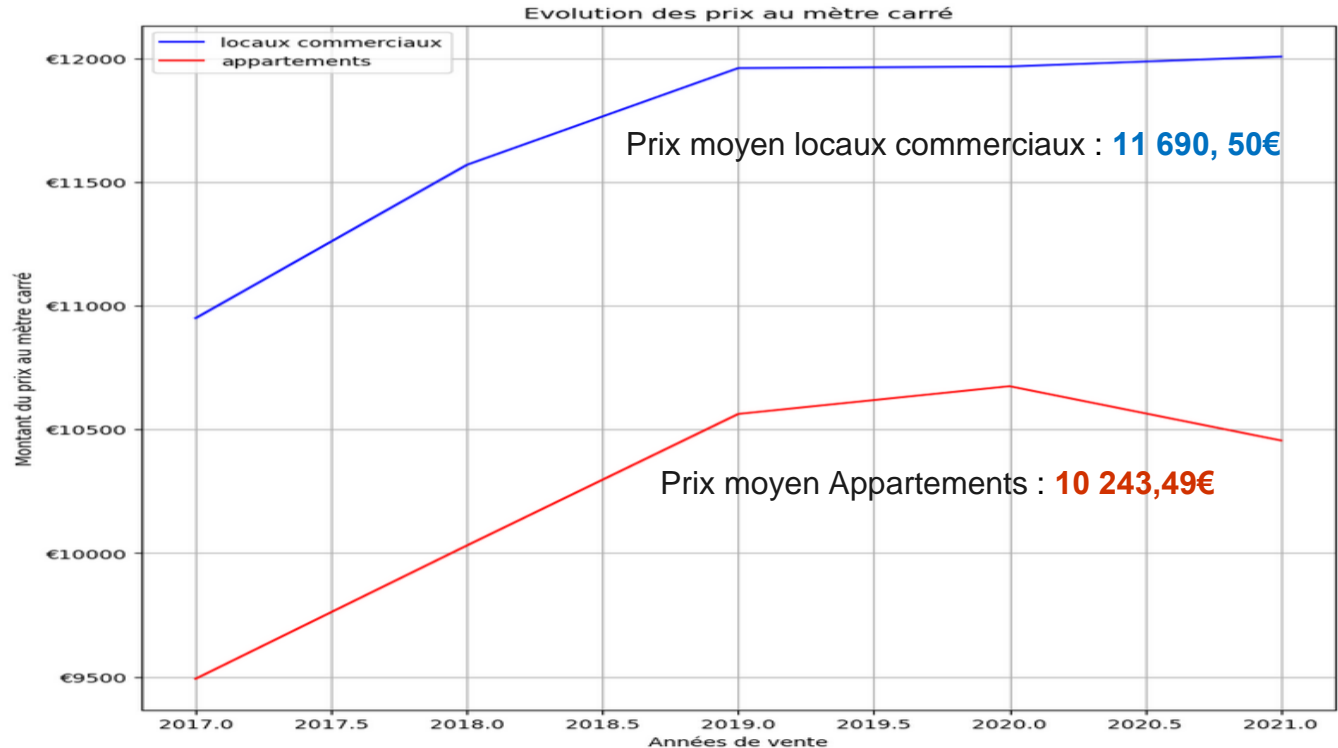
Le coefficient de corrélation est de **0.984** : relation linéaire très forte, avec une pvalue est **inférieure à 0.05** .

Confirmation de la corrélation :

**La surface explique une grande partie de la variation de la valeur foncière**

# I. Analyse du marché de l'immobilier

Les **locaux commerciaux** n'ont pas connu de baisse significative en 2020.





## II. Méthodologie suivie

Analyse des besoins

### Objectif d'analyse SMART :

Détermine le but spécifique que cherche à atteindre l'étude statistique, guidant la direction de la recherche.

Collecte des données

### Importation des données

- Données quantitatives : valeurs foncières et surface réelle
- Données qualitatives : type de logement, code postal et date

Préparation des données

### Nettoyer les données

valeurs manquantes, correction erreurs, suppression doublons, traitement données aberrantes, suppression de colonnes inutiles

Visualisation des données

Utilisation des librairies comme 'matplotlib'

- Création de colonnes : prix au mètre carré, prix moyen, surface moyenne et Création de table : appartement, locaux industriels, arrondissement
- Visualisation des cohérences

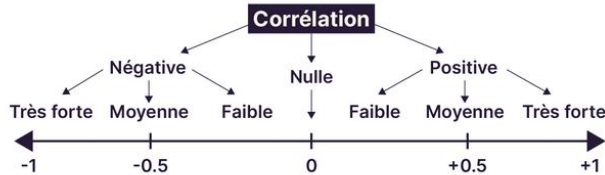
# II. Méthodologie suivie

## Test de corrélation

### Choix du test Pearson

Le test de corrélation de Pearson mesure la force et la direction de la relation linéaire entre deux variables quantitatives.

Il est représenté par le coefficient de corrélation  $r$ , qui varie entre -1 et 1.



Test  
quantitatif

Deux variables  
quantitatives

Test  
paramétrique

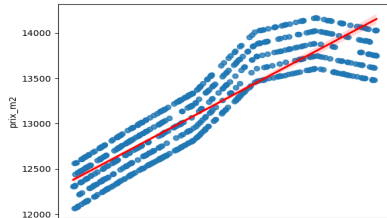
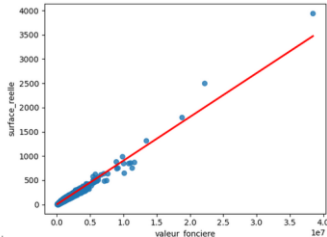
Test non  
paramétrique

Test  
Pearson

Test  
Spearman

**Pourquoi "paramétrique" ?**

Le test Pearson utilise des calculs mathématiques qui fonctionnent mieux quand les points sont répartis symétriquement autour d'une moyenne.



### Caractéristiques clés :

- Mesure la relation linéaire entre deux variables continues.
- Suppose une distribution normale des données.
- Requiert des données quantitatives (intervalle ou ratio).
- Sensible aux valeurs aberrantes.

## II. Méthodologie suivie

df\_locaux\_commerc

	annee	Type de local	Prix moyen au m <sup>2</sup>
1	2017	4	10,949.91
3	2018	4	11,569.50
5	2019	4	11,960.13
7	2020	4	11,966.47
9	2021	4	12,006.49

df\_appartements

	annee	Type de local	Prix moyen au m <sup>2</sup>
0	2017	2	9,492.86
2	2018	2	10,031.40
4	2019	2	10,562.71
6	2020	2	10,674.87
8	2021	2	10,455.60

```
prix_moyen_locaux_comm = round(df_locaux_commerc['Prix moyen au m2'].mean(),2)
prix_moyen_locaux_comm
```

11690.5

```
prix_moyen_appart = round(df_appartements['Prix moyen au m2'].mean(),2)
prix_moyen_appart
```

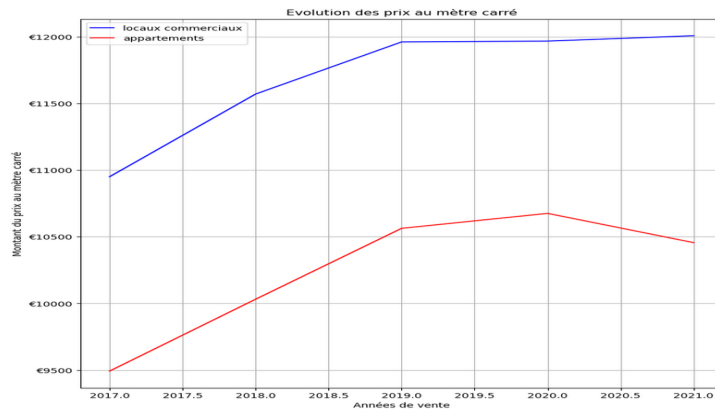
10243.49

### Analyse des locaux industriels, commerciaux et assimilés

Création d'un dataset ne contenant que les locaux commerciaux  
Comparaison des prix au mètre carré entre les appartements/locaux commerciaux

#### Choix :

Graphique pour visualisation  
Prix moyen mètre carré par année



## II. Méthodologie suivie

*Pour prédire le prix du m2, nous aurons besoin de :*

la surface du bien immobilier :

plus la surface est grande, plus le prix augmente

la date considérée :

plus le temps passe plus le prix augmente

la localisation (code\_postal) :

le prix augmente selon l'arrondissement

le type de bien :

les prix des locaux commerciaux sont plus élevés

One hot encoder

annee	2017	2018	2019	2020	2021
2017	1	0	0	0	0
2018	0	1	0	0	0
2019	0	0	1	0	0
2020	0	0	0	1	0
2021	0	0	0	0	1
2019	0	0	1	0	0

*Entraînement de l'algorithme de régression linéaire*

On prépare les données en transformant les colonnes catégoriques du code postal et du type de local grâce au **one hot encoder** (sklearn) / `get_dummies` (pandas) – voir exemple avec les années ci-dessus

*Préparation des données*

Transformation des données

Variables prédictives = surface réelles, années, code postal, code local

Variable cible = valeurs foncières

## II. Méthodologie suivie

On utilise le `train_test_split` pour prélever un tiers de nos données (33%) et les garder de côté.

```
# On sépare le jeu de données entre échantillons d'apprentissage et de test
# La valeur y à trouver est la valeur foncière

# Données déjà prétraitées (X_processed) et cible (y)
X_train, X_test, y_train, y_test = train_test_split(
    X_processed, # Features prétraitées (après encodage)
    y, # Variable cible (valeur_foncière)
    test_size=0.33,
    random_state=42, # Pour reproductibilité
    shuffle=True # Mélange aléatoire des données (désactiver pour séries temporelles)
)
```

Nous avons entraîné notre algorithme sur le reste des données.

Puis mesurer notre erreur moyenne en pourcentage de la valeur foncière.

Notre algorithme fait donc **3 %** d'erreur en moyenne sur la prédiction de la valeur foncière.

### Mes conclusions sur ce résultat :

- Augmentation d'une unité de **surface réelle** = augmentation moyenne de 10 602.36 de la valeur foncière
- Augmentation d'une unité de **année** = augmentation moyenne de 14 127.25 de la valeur foncière
- Augmentation d'une unité de **code postal** = diminution moyenne de 9 172.71 de la valeur foncière
- Augmentation d'une unité de **code type local** = augmentation moyenne de 45 354.73 de la valeur foncière

```
from sklearn.linear_model import LinearRegression
reg = LinearRegression()
# On entraîne l'algorithme ci-dessous et on effectue la prédiction
reg.fit(X, y)

# et on obtient directement un score
print(reg.score(X, y))

# ainsi que les coefficients a, b, c de la régression linéaire
print(reg.coef_)

0.9777254877206208
[10602.35997897 14127.24838172 -9172.7143504 45354.7343238 ]
```

# III. Résultat des prédictions

## Prédiction définitive pour le client méthodologie suivie

### Collecte des données :

Nous avons récupéré le fichier avec le **portefeuille des actifs de la société**.

### Préparations les données :

**Date demandée** : 31 décembre 2022

Utilisation de la surface réelle et non surface carrez.

Nombre de transaction : **275**

### Nettoyer les données... (voir p9)

On réutilise les mêmes fonctions pour faire le **one hot encoding** (voir p12)

Conversion du dataframe en objet = pour les données ('surface\_reelle', 'annee', 'code\_postal', 'code\_type\_local')

### Notre dataframe est prêt à être utilisé par notre algorithme de prédiction.

Application du même préprocesseur utilisé lors de l'entraînement (voir p13).

On ajoute les prédictions au DataFrame original.

### Vérifications

# III. Résultat des prédictions

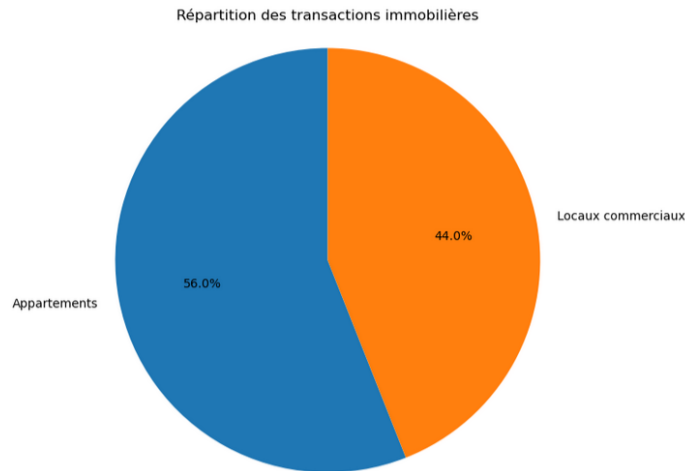
## Prédiction définitive pour le client

La valorisation du segment **particulier** est de 82,54 M€

La valorisation du segment **corporate** est de 93,90 M€

Un petit échantillon (275) pourrait ne pas être représentatif de l'ensemble du marché.

Le segment **corporate** représente donc une part plus importante du portefeuille immobilier bien qu'il représente 44% des biens immobiliers.



---

# Les Plus Beaux Logis de Paris

## Partie 2





## Les Plus Beaux Logis de Paris

*Analysez l'évolution des prix de l'immobilier avec Python*

*Aurelia*

*De Infanti*

*28/03/2025*

# I. Méthodologie suivie

## Classification des données issues du jeu de test

Pour **labellisé automatiquement** les biens immobiliers comme étant :  
soit des Appartements  
soit des Local industriel, commercial ou assimilé



Utilisation de l'algorithme du **Kmeans** qui va rechercher 2 centroïdes à travers les données.

Pour que l'algorithme fonctionne, nous avons préparé les données en :

- supprimant les dimensions inutiles
- se concentrant sur le facteur discriminant entre les 2 types de biens immobiliers: la différence dans le prix au mètre carré.

### Application des transformations nécessaires.

Calculer le prix au mètre carré (en divisant la valeur foncière par la surface).

Retirer ces colonnes car nous avons déjà l'information qu'elles contiennent dans la dimension prix au mètre carré.

Toutes les données sont de l'année 2021. Nous avons retiré cette dimension

### *Calcul du prix au mètre carré*

```
: # calculer le prix au mètre carré
df_aclasse['prix_m2'] = df_aclasse['valeur_fonciere']/df_aclasse['surface_reelle']
df_aclasse.head()
```

	valeur_fonciere	code_postal	nom_commune	surface_reelle	prix_m2
0	868,687.08	75019	Paris 19e Arrondissement	88	9,871.44
1	452,050.76	75019	Paris 19e Arrondissement	45	10,045.57
2	193,088.65	75019	Paris 19e Arrondissement	21	9,194.70
3	303,012.55	75019	Paris 19e Arrondissement	32	9,469.14
4	149,272.20	75019	Paris 19e Arrondissement	20	7,463.61

# I. Méthodologie suivie

## Prix au mètre carré du 19<sup>eme</sup> arrondissement :

Dimensions utilisées pour attribuer les prix au mètre carré les plus élevés dans un département aux locaux commerciaux, et les prix les plus bas aux appartements.

```
# Préparer Les données pour le clustering
X = df_aclasse[['prix_m2']].values

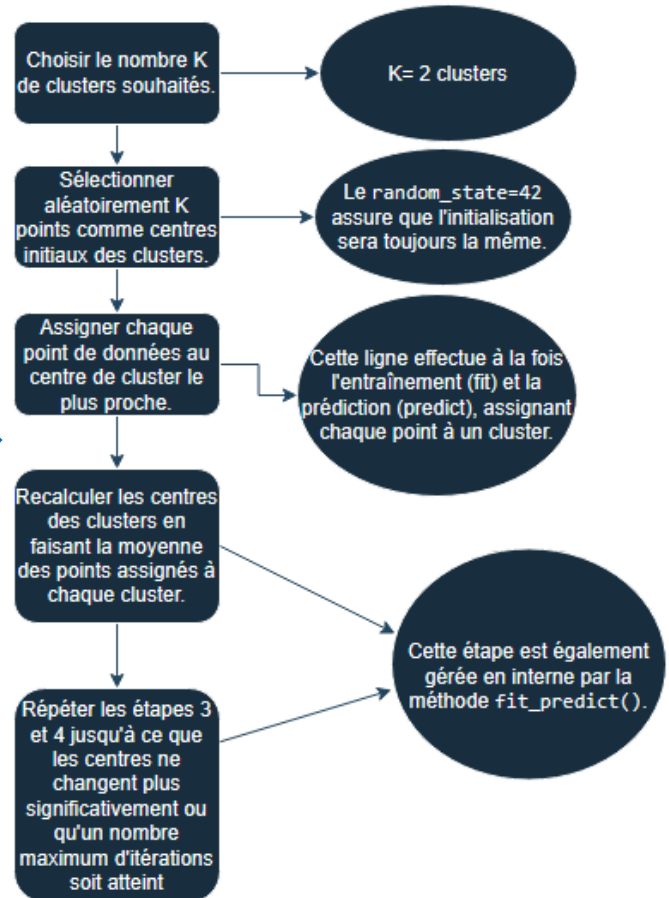
# Normaliser les données
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Appliquer l'algorithme KMeans
kmeans = KMeans(n_clusters=2, random_state=42)
df_aclasse['code_type_local'] = kmeans.fit_predict(X_scaled)

# Identifier le cluster avec le prix moyen le plus élevé comme "Local industriel, commercial ou assimilé"
prix_moyen_par_cluster = df_aclasse.groupby('code_type_local')['prix_m2'].mean()
cluster_commercial = prix_moyen_par_cluster.idxmax()
```

traduction

	prix_m2 mean	min	max
type_local			
Appartement	7,408.78	7,207.22	7,666.07
Local industriel, commercial ou assimilé	9,806.92	9,194.70	10,113.20



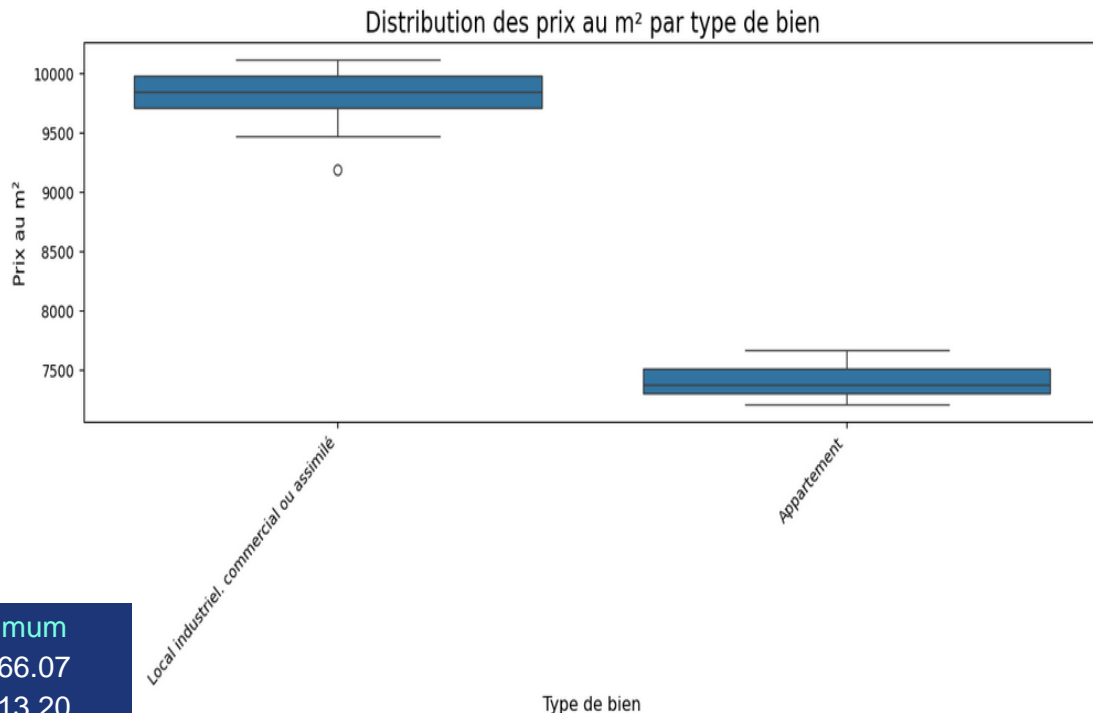
## II. Résultat de la classification

### Classification des données issues du jeu de test

- Année 2021
- Arrondissement 19eme

### Nous avons :


- ✓ Obtenue notre prédiction
- ✓ Changer les labels
- ✓ Remplacer les valeurs à 0 par Local industriel, commercial ou assimilé et les valeurs à +1 par Appartement.



Type_local	moyenne	minimum	maximum
Appartement	7,408.78	7,207.22	7,666.07
Local commercial	9,806.92	9,194.70	10,113.20

## II. Résultat de la classification

### Conclusions sur l'analyse et les limites de l'exercice :

- L'algorithme **KMeans** a permis de classer les biens en 2 catégories : Appartements (code\_type\_local = 2) et Locaux industriels, commerciaux ou assimilés (code\_type\_local = 4).
- Les résultats montrent que les biens avec des prix au m<sup>2</sup> élevés ont été **correctement** attribués à la catégorie des locaux commerciaux, tandis que les biens avec des prix au m<sup>2</sup> plus bas ont été classés comme appartements.
- Prix moyen appartements : **7 404.78** euros
- Prix moyen locaux commerciaux : **9 806.92** euros
-  Certains biens avec un prix au m<sup>2</sup> élevé pourraient être des appartements luxueux plutôt que des locaux commerciaux.
- **KMeans** suppose que les clusters sont convexes et bien séparés, ce qui peut ne pas être le cas dans des données réelles où il existe un chevauchement entre les catégories.
- En **conclusion**, bien que cet exercice fournisse une approche intéressante pour différencier appartements et locaux commerciaux à partir du prix au mètre carré, il reste simplifié et pourrait être enrichi par une analyse plus approfondie et multidimensionnelle.
- **Invertir locaux commerciaux = plus rentable**
- **Les prédictions** : Ne tiennent pas compte des changements tel que catastrophes climatiques, épidémies, économiques...