Projet DALAS

Agence Immobilière

Master DAC 2023 - 2024

Par Jean-Marc ZHUANG et Salwa MUJAHID

Projet dirigé par Laure SOULIER

Sommaire

Introduction

Dans ce projet, nous sommes agents immobiliers et nous souhaitons comprendre la prédiction des prix des appartements à Paris afin de nous permettre d'obtenir des résultats optimales. Nous avons collecté et analysé des données provenant de plusieurs sources, notamment Century21, housedata de Kaggle, et immo-data. En nous focalisant principalement sur les données les plus pertinentes issues de immo-data, nous avons nettoyé et préparé le jeu de données pour en extraire des informations importantes.

Les axes principaux :

1. Prédiction de prix selon les caractéristiques d'un appartement
2. Prédiction de prix pour les prochaines années (inflation ?)

Jeux de données

Century21

Century21 (https://www.century21.fr/) est un site de ventes d'appartements. Le scraping depuis le site était assez simple, car nous trouvons la même architecture dans la liste d'appartements. Nous nous focalisons principalement sur la ville de PARIS.

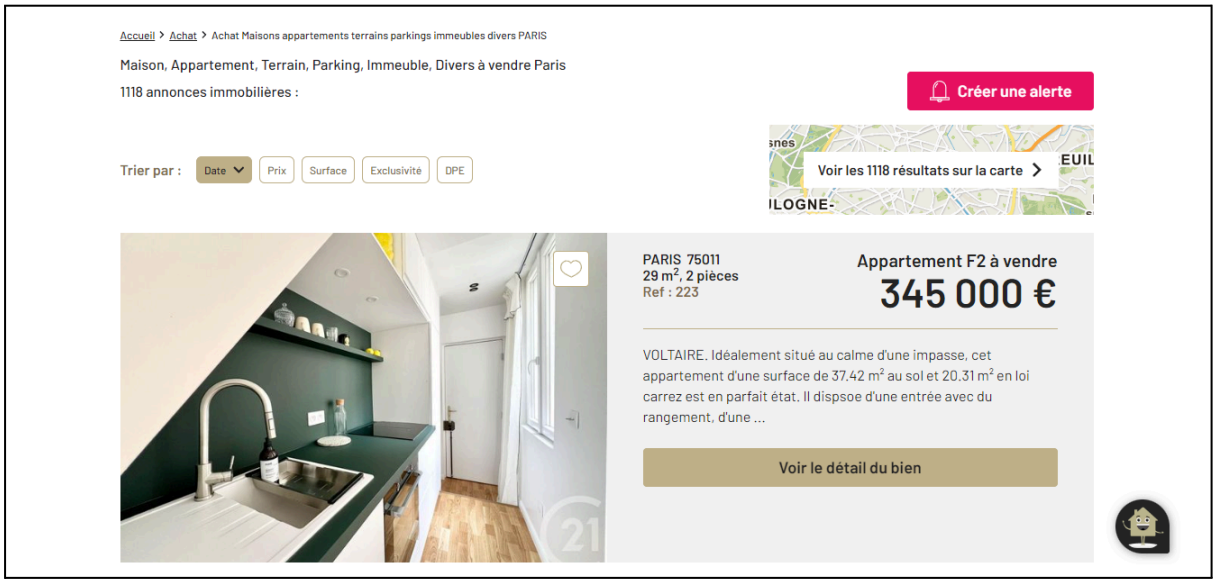


***Image 1** : Capture d'écran du site Century21 des appartements à PARIS*

	Ville	Arrondissement	Surface (m2)	Pièce(s)	Prix (€)
0	PARIS	75008	9.00	1	536
1	PARIS	75019	49.70	2	1446
2	PARIS	75016	42.87	1	1480
3	PARIS	75013	84.30	4	2550
4	PARIS	75015	45.28	2	1650

***Table 1** : Les 5 premières lignes du tableau construit à partir des données de Century21*

housedata

Nous avons également importé les données housedata depuis Kaggle. Il s'agit de données sur les informations d'appartements aux Etats-Unis, ce qui est moins intéressant, car nous voulons nous focaliser sur les appartements à Paris. Nous pouvons remarquer que les données sur la surface sont plus détaillées, ce qui peut peut-être influencer le prix.

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	\
0	2014-05-02 00:00:00	313000.0	3.0	1.50	1340	7912	
1	2014-05-02 00:00:00	2384000.0	5.0	2.50	3650	9050	
2	2014-05-02 00:00:00	342000.0	3.0	2.00	1930	11947	
3	2014-05-02 00:00:00	420000.0	3.0	2.25	2000	8030	

```
4  2014-05-02 00:00:00   550000.0       4.0       2.50         1940      10500

   floors  waterfront  view  condition  sqft_above  sqft_basement  yr_built  \
0     1.5           0     0          3        1340              0      1955
1     2.0           0     4          5        3370            280      1921
2     1.0           0     0          4        1930              0      1966
3     1.0           0     0          4        1000           1000      1963
4     1.0           0     0          4        1140            800      1976

   yr_renovated                    street       city  statezip country
0          2005      18810 Densmore Ave N  Shoreline  WA 98133     USA
1             0           709 W Blaine St    Seattle  WA 98119     USA
2             0  26206-26214 143rd Ave SE       Kent  WA 98042     USA
3             0           857 170th Pl NE   Bellevue  WA 98008     USA
4          1992         9105 170th Ave NE    Redmond  WA 98052     USA
```

***Table 2** : Les 5 premières lignes du tableau housedata*

immo-data

immo-data (https://www.immo-data.fr/) est un site qui permet d'analyser le marché de l'immobilier, donc de consulter les ventes d'appartements passés avec quelques informations sur une zone géographique précise. Dans notre cas, nous centrons la zone à PARIS pour récupérer les données.

Pour construire notre tableau, les données sont récupérées comme suit : Nous avons entré manuellement en recherche chaque arrondissement, fait un zoom sur ces zones et récupéré les données de plusieurs centaines d'appartements de manière automatique. Nous n'avons pas appliqué de filtre pour la récupération de données.

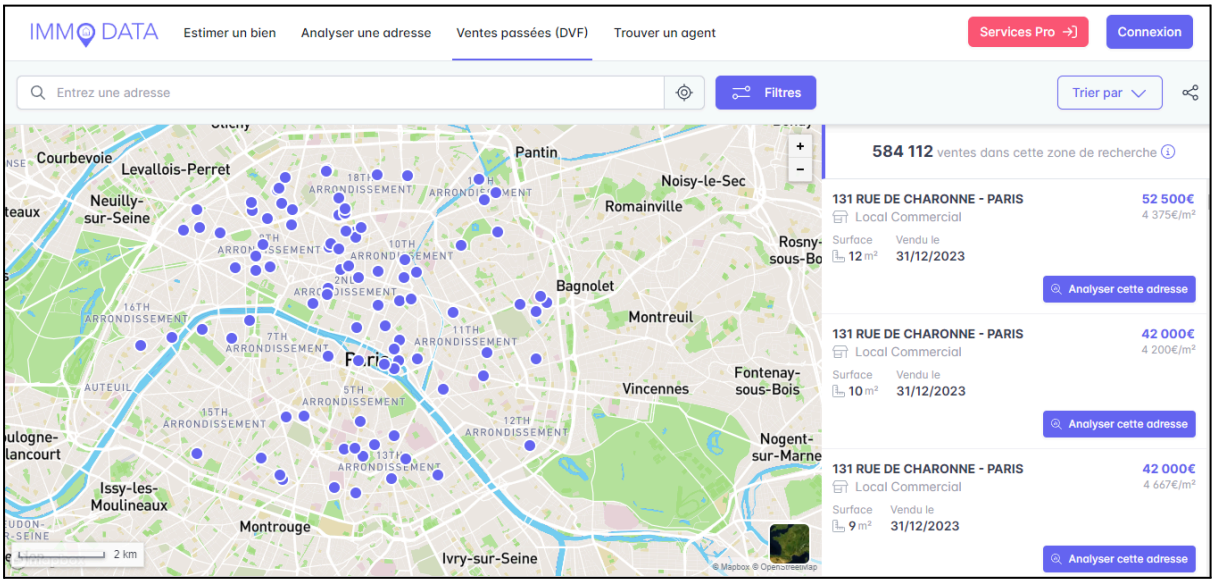


***Image 2** : Capture d'écran du site immo-data centré sur Paris*

```
                Adresse  Ville  Arrondissement        Type  Prix (€)  \
```

0	270 RUE SAINT-HONORÉ	PARIS		75001	Appartement	750400
1	186 RUE DE RIVOLI	PARIS		75001	Appartement	330000
2	23 RUE DE RICHELIEU	PARIS		75001	Appartement	360100
3	27 RUE DE RICHELIEU	PARIS		75001	Appartement	286123
4	272 RUE SAINT-HONORÉ	PARIS		75001	Appartement	411636

	Prix mensuel (€)	Pièce(s)	Surface (m2)	Date de vente
0	14431	2	52	22/05/2023
1	14348	1	23	28/04/2023
2	15657	1	23	29/03/2023
3	11005	1	26	10/02/2023
4	12864	2	32	28/12/2022

***Table 3** : Les 5 premières lignes du tableau construit à partir des données de immo-data*

Finalement, nous n'avons utilisé que les données de immo-data, car c'est celui qui contient les informations les plus pertinentes. Peut-être que nous utiliserons les autres pour confirmer les résultats ou obtenir plus d'informations.

Exploration des données

Le tableau construit

	Adresse	Ville	Arrondissement	Type	Prix (€)	Prix mensuel (€)	Pièce(s)	Surface (m2)	Date de vente
0	270 RUE SAINT-HONORÉ	PARIS	75001	Appartement	750400	14431	2	52	22/05/2023
1	186 RUE DE RIVOLI	PARIS	75001	Appartement	330000	14348	1	23	28/04/2023
2	23 RUE DE RICHELIEU	PARIS	75001	Appartement	360100	15657	1	23	29/03/2023
3	27 RUE DE RICHELIEU	PARIS	75001	Appartement	286123	11005	1	26	10/02/2023
4	272 RUE SAINT-HONORÉ	PARIS	75001	Appartement	411636	12864	2	32	28/12/2022

***Table 4** : Les 5 premières lignes du tableau construit à partir des données de immo-data*

Nous avons effectué plusieurs modifications sur le tableau après l'avoir examiné. Nous avons retiré la colonne ville, car nous n'avons qu'une seule valeur qui est "PARIS" sur lequel nous avons centré nos données. Nous avons également retiré les lignes aux valeurs manquantes, car elles ne sont pas nombreuses et nous préférons éviter de remplir les lignes ou faire d'autres modifications. Les valeurs à 0 pour certaines features ont également été retirées, car ces données ne sont pas pertinentes pour nos analyses. Pour finir, les valeurs catégorielles ont été encodées et les dates ont été transformées en valeurs numériques pour permettre de mieux les manipuler et les représenter dans les graphes.

Distribution de données

Distribution des Arrondissements et Villes

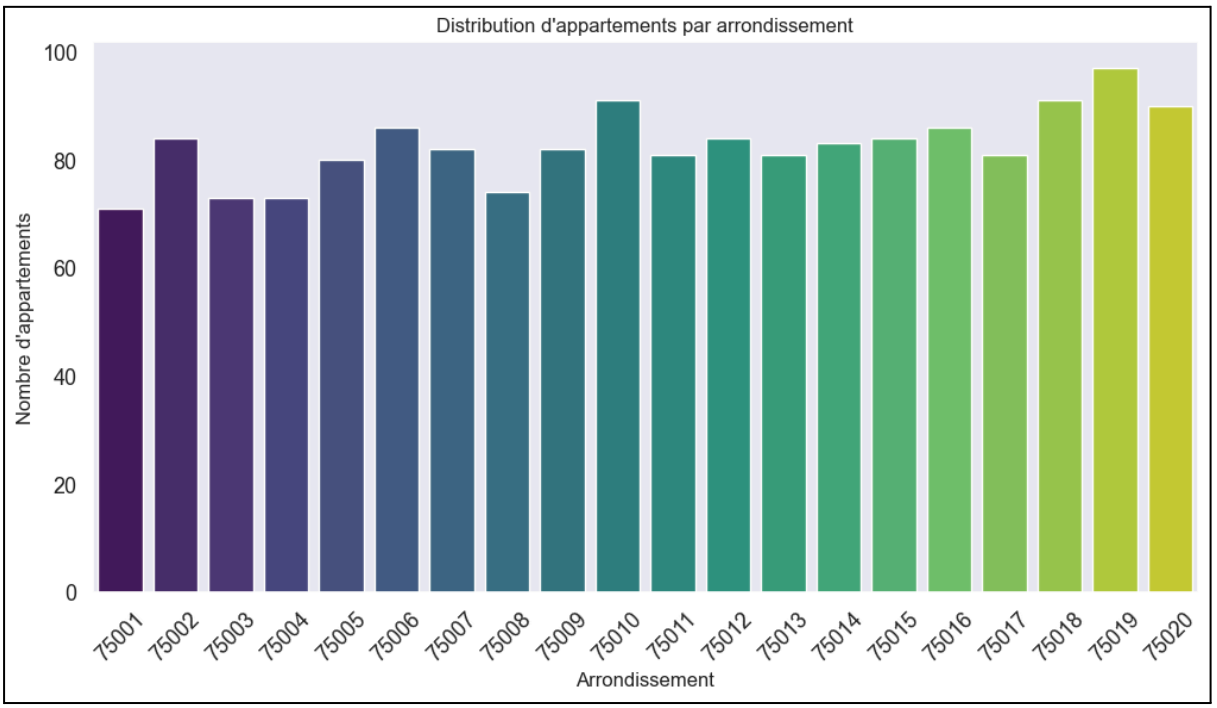


***Figure 1** : Histogramme des distributions d'arrondissement*

Nous avons manuellement récupéré les arrondissements dans ***Figure 1***, d'où le fait qu'ils soient distribués de manière équilibrée. Cela permet également de faire de bonnes prédictions par rapport aux arrondissements.

Distribution des Prix

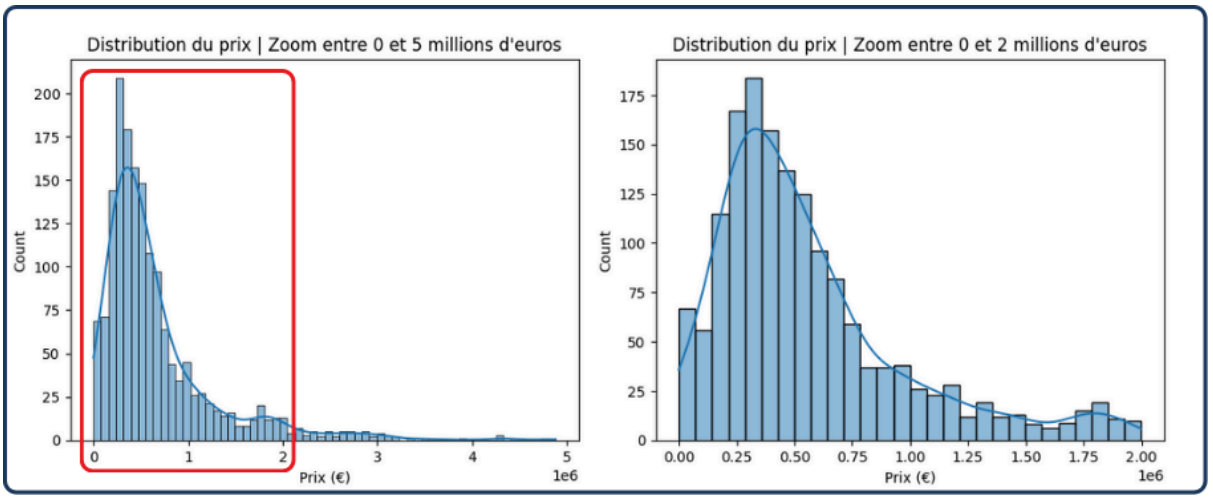


Figure 2 *: Histogrammes sur la distribution des prix total*

D'après ***Image 3***, la distribution de prix dans ***Figure 3*** suit une distribution Log Normale et celle du prix mensuel dans ***Figure 2*** suit une distribution normale. Il faut cependant, voir de plus près l'analyse du prix mensuel par rapport au prix total. C'est ce qu'on a analysé dans la partie précédente.

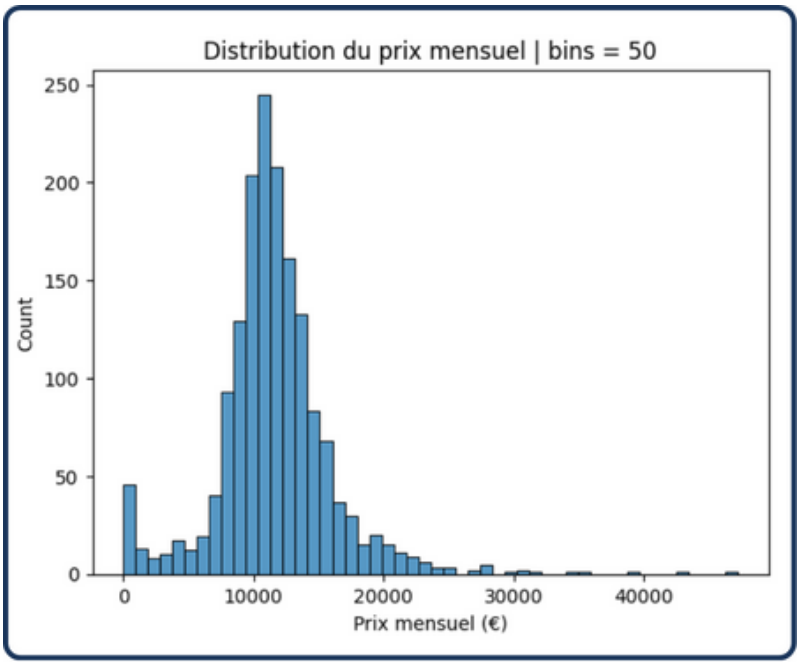


Figure 3 *: Histogrammes sur la distribution des prix mensuel*

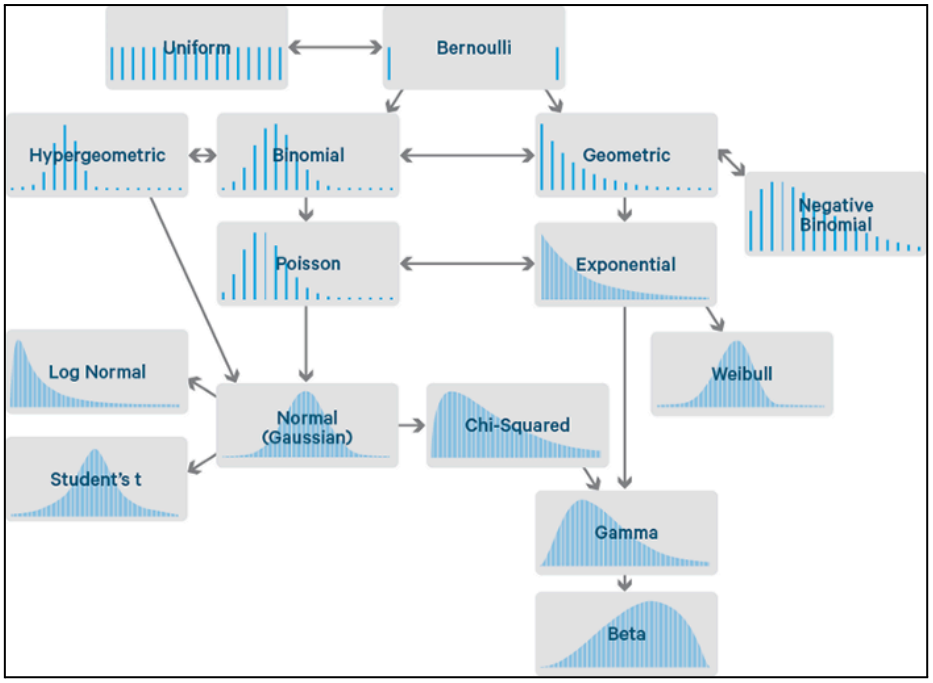


Image 3 *: Liste de distributions*

Distribution de Pièces et Surface

On peut observer dans la ***Figure 4*** et ***Figure 5***, que la distribution du nombre de pièces et de la surface suivent une distribution Log Normale.

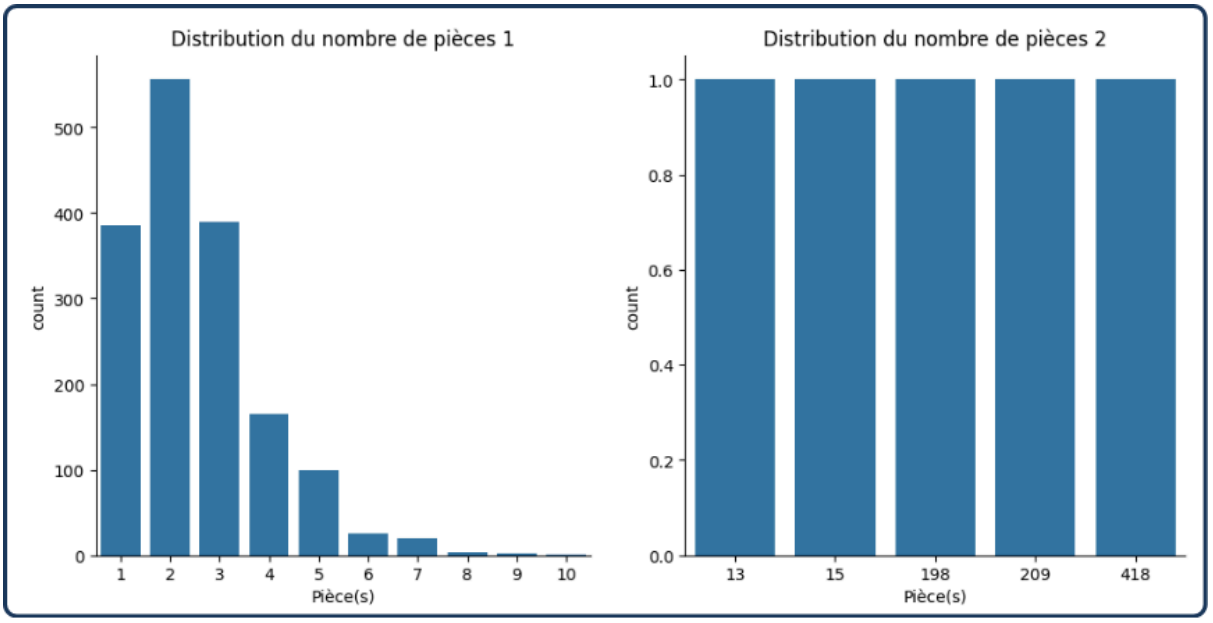


Figure 4 *: Histogrammes de la distribution du nombre de pièces*

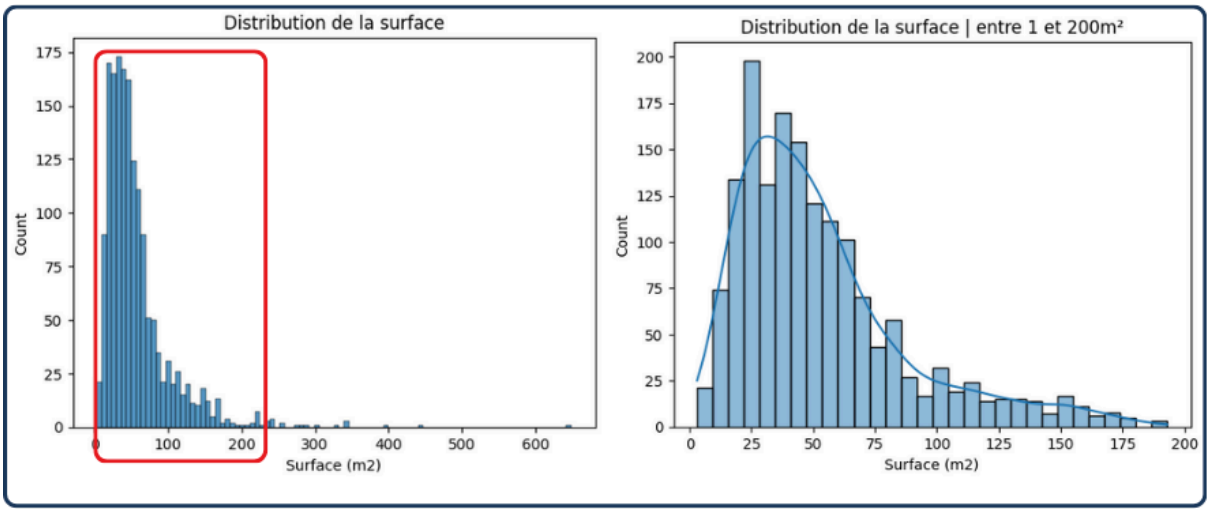


Figure 5 *: Histogrammes de la distribution de la surface*

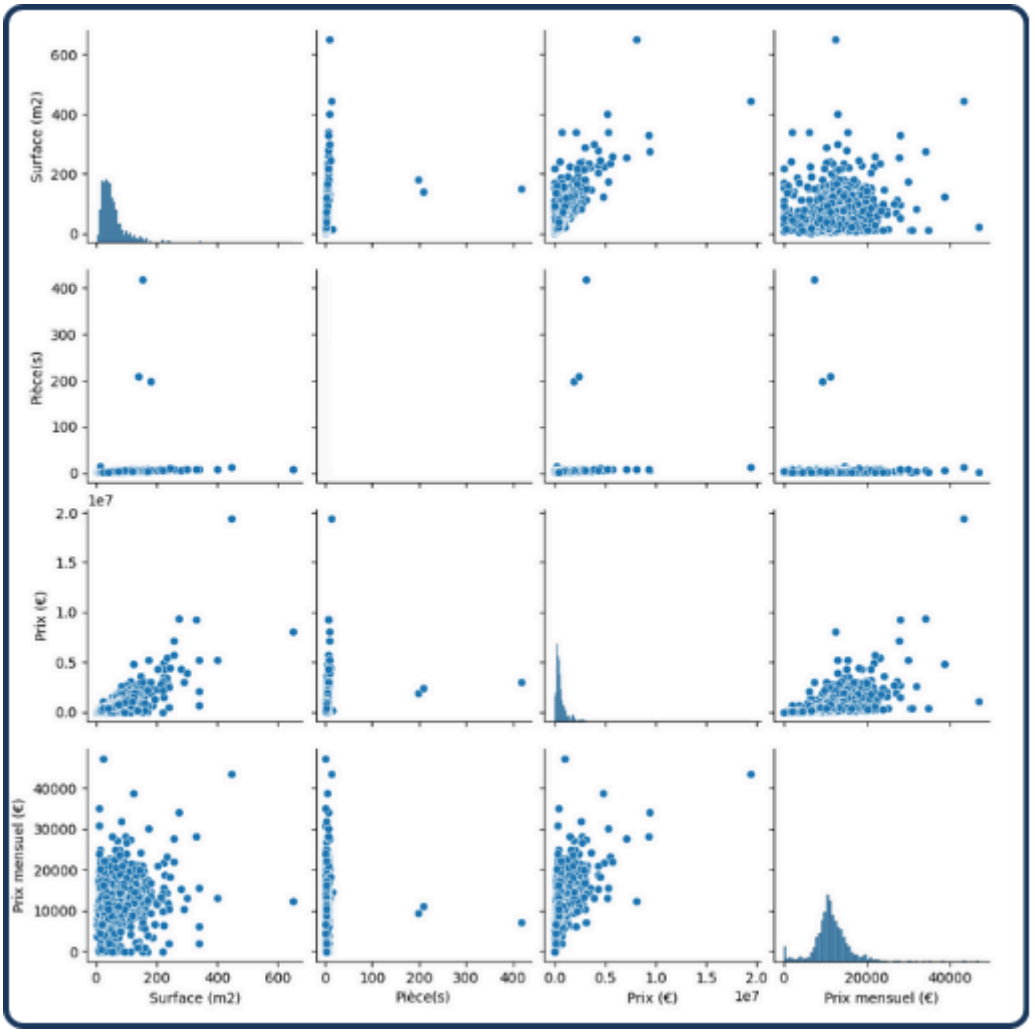


Figure 6 *: Pair Plot des features de prix (mensuel et total), de nombre de pièces et de surface avant d'avoir retiré les valeurs aberrantes*

Comme nous pouvons observer dans ***Figure 6***, il est difficile de représenter les valeurs avec les nuages de points à cause des valeurs aberrantes. Nous montrerons nos résultats après l'étape de la détection et de suppressions de valeurs aberrantes plus loin.

Distribution de Dates

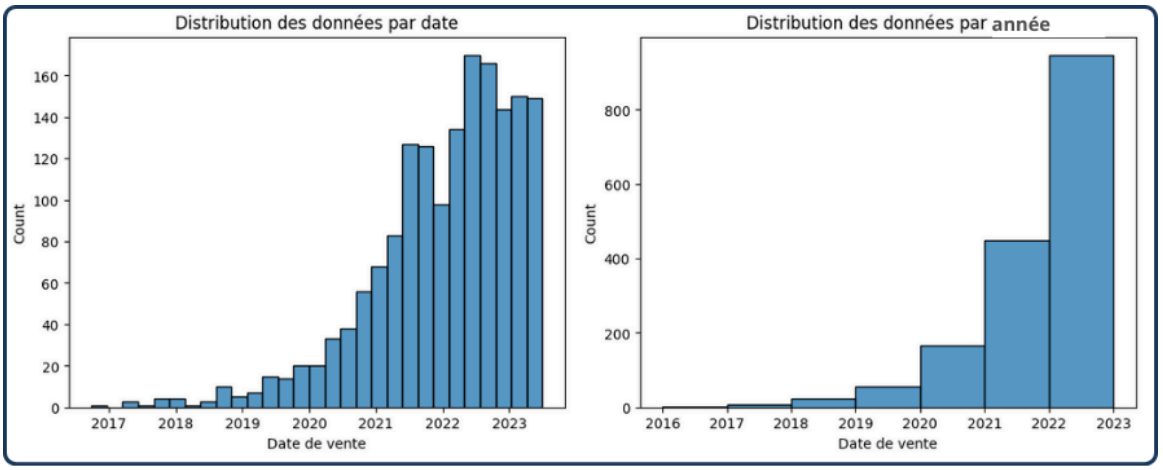


Figure 7 *: Histogrammes sur la distribution des dates*

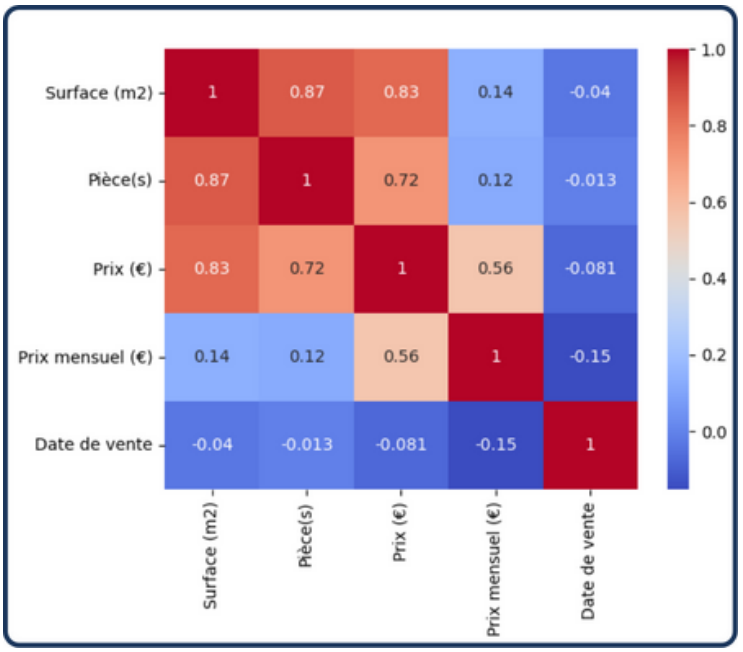


Figure 8 *: Matrice de corrélation entre les features sauf l'arrondissement, la ville et le type*

Deux des dates avaient des valeurs étranges, donc nous les avons retirées.

Il est difficile de faire la prédiction en fonction des dates comme on peut voir dans ***Figure 8*** dans laquelle la date n'est pas corrélée avec les autres features, car la distribution des dates n'est pas uniforme comme montré dans ***Figure 7***. En effet, nous avons plus de dates récentes que d'anciennes.

Distribution des Types

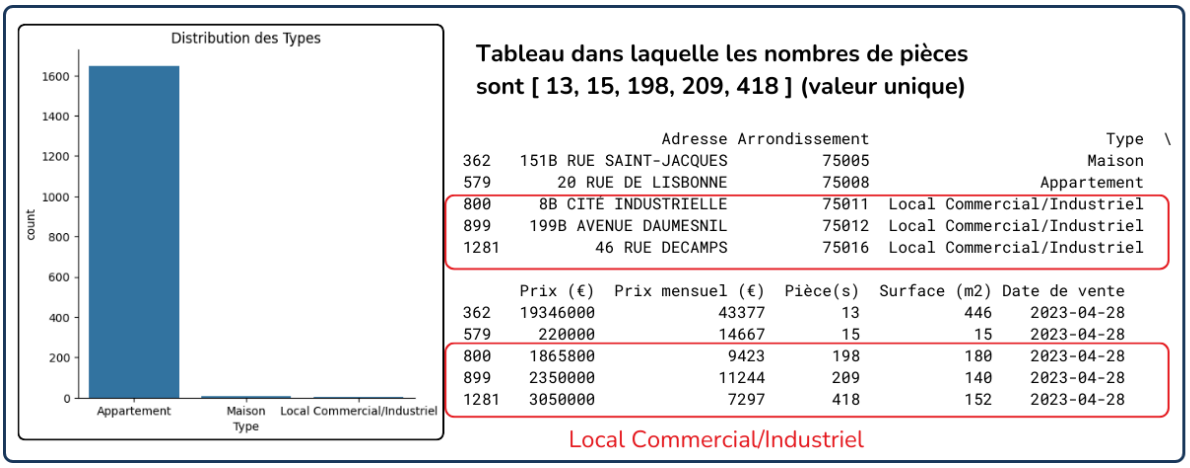


Tableau dans laquelle les nombres de pièces sont [13, 15, 198, 209, 418] (valeur unique)

	Adresse	Arrondissement	Type
362	151B RUE SAINT-JACQUES	75005	Maison
579	20 RUE DE LISBONNE	75008	Appartement
800	8B CITÉ INDUSTRIELLE	75011	Local Commercial/Industriel
899	199B AVENUE DAUMESNIL	75012	Local Commercial/Industriel
1281	46 RUE DECAMPS	75016	Local Commercial/Industriel

	Prix (€)	Prix mensuel (€)	Pièce(s)	Surface (m2)	Date de vente
362	19346000	43377	13	446	2023-04-28
579	220000	14667	15	15	2023-04-28
800	1865800	9423	198	180	2023-04-28
899	2350000	11244	209	140	2023-04-28
1281	3050000	7297	418	152	2023-04-28

Local Commercial/Industriel

***Figure 9** : Histogramme de la distribution des types d'appartements et exemples de données avec pour feature Pièce(s) les valeurs [13, 15, 198, 209, 418]*

D'après ***Figure 9***, les 3 appartements de type 'Local Commercial/Industriel' ont une surface de plus de 100 pièces et entre 100 et 200 m².

Nous avons retiré la colonne "Type" ainsi que les lignes contenant les valeurs autres que "Appartement", car selon la distribution présentée dans l'histogramme dans ***Figure 9***, nous n'avons que très peu de valeurs, ce qui est peu significatif. Une autre alternative serait de récupérer plus de données contenant ces valeurs, mais nous avons décidé de supprimer à la place.

Corrélations entre les variables

Analyse des variables 2 à 2

D'après ***Figure 10***, pour comparer les features 2 à 2, les features les moins corrélés avec toutes les autres features sont le prix mensuel et l'arrondissement. Plus précisément, la matrice montre que le prix mensuel est assez corrélé au prix, mais pas aux autres features. Pour en être sûrs, nous avons analysé le prix mensuel et le prix total.

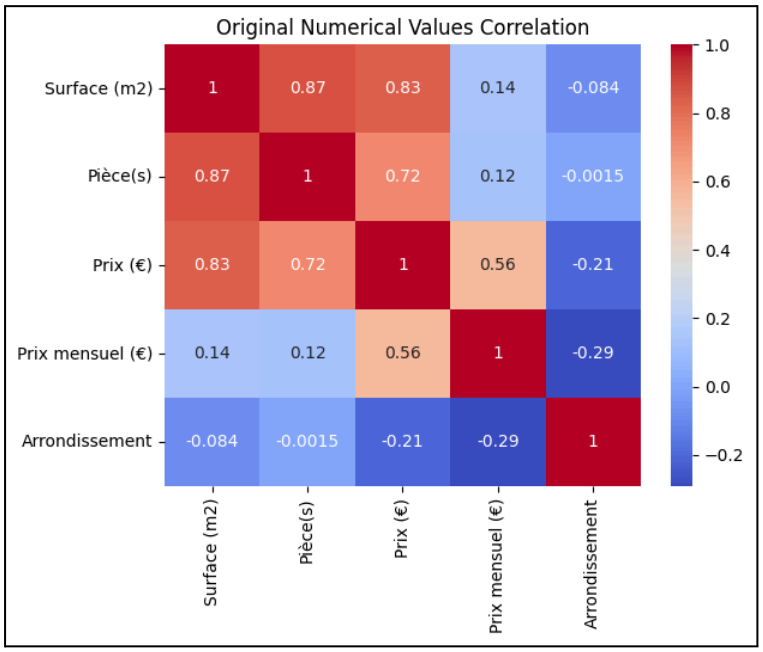


Figure 10 *: Matrice de corrélation entre les features sauf la ville, le type et la date de vente*

Analyse du prix

Dans la ***Figure 11***, nous observons que même si la courbe paraît suivre une droite de régression, nous avons beaucoup de variations, ce qui complique la prédiction du prix mensuel selon le prix original. Ceci peut être dû au fait que les prix mensuels ne sont pas tous payés sur le même nombre de mois pour les appartements.

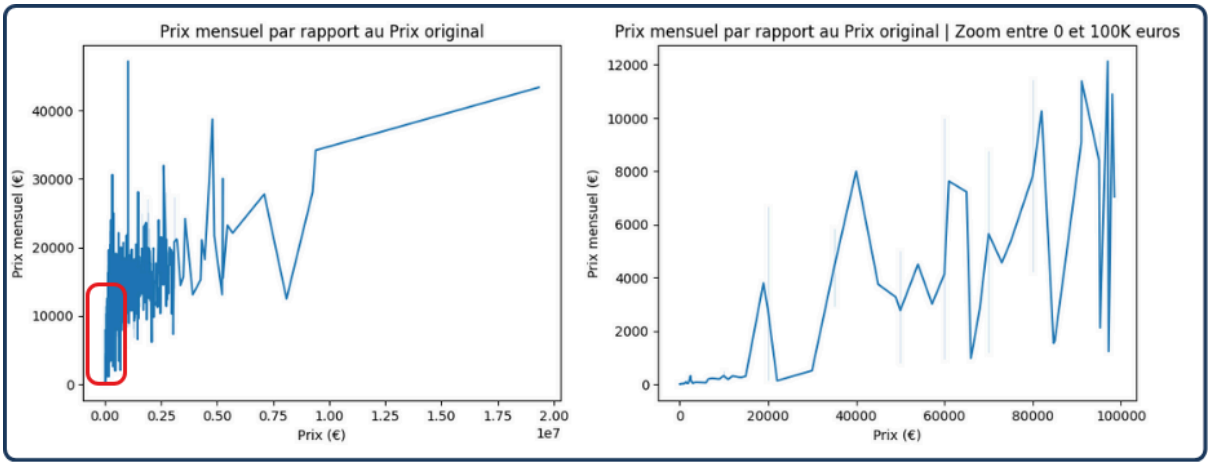


Figure 11 *: Courbe représentant le prix mensuel par rapport au prix total*

Analyse par arrondissement

Dans la ***Figure 12***, nous pouvons distinguer les arrondissements différents selon la surface et le prix. Sachant que la distribution des appartements que nous avons collectés par arrondissement est uniforme, nous pouvons nous fier au prix moyen par arrondissement présenté dans l'histogramme de la ***Figure 13***.

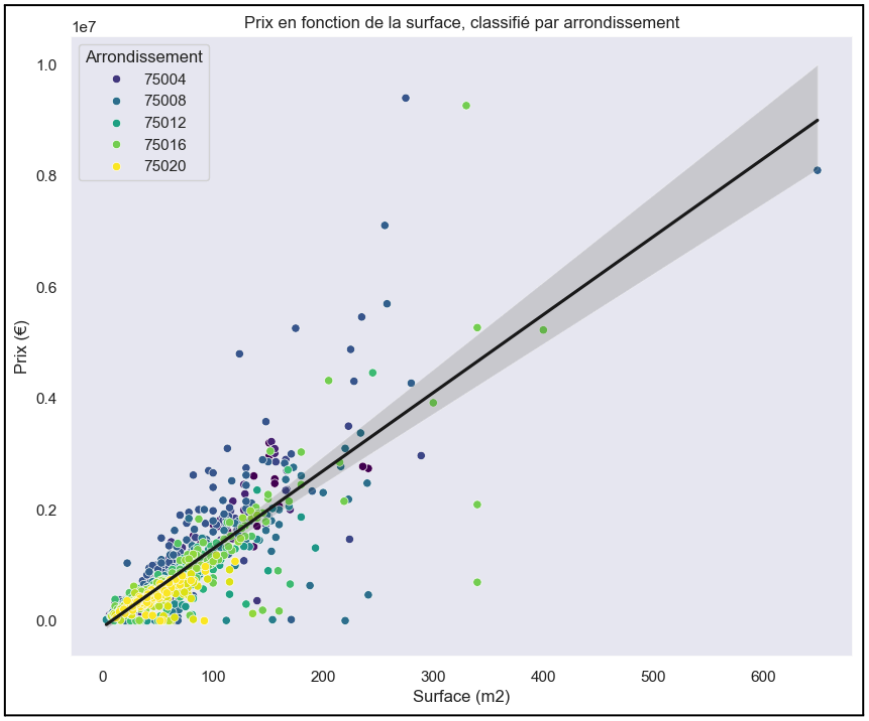


Figure 12 *: Courbe représentant le prix et la surface selon les différentes zones géographiques (arrondissements)*

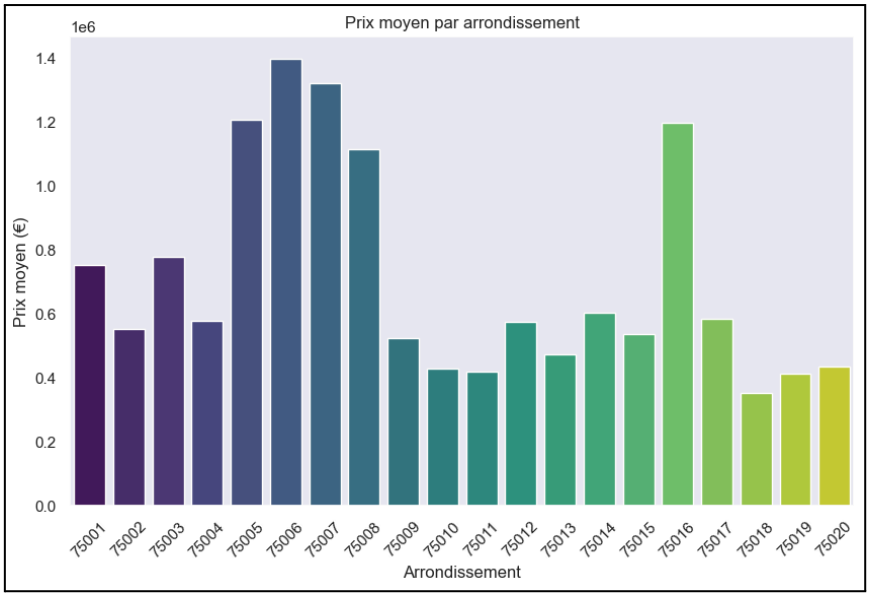


Figure 13 *: Histogramme représentant le prix moyen d'appartements selon les différentes zones géographiques (arrondissements)*

Preprocessing

Normalisation des données

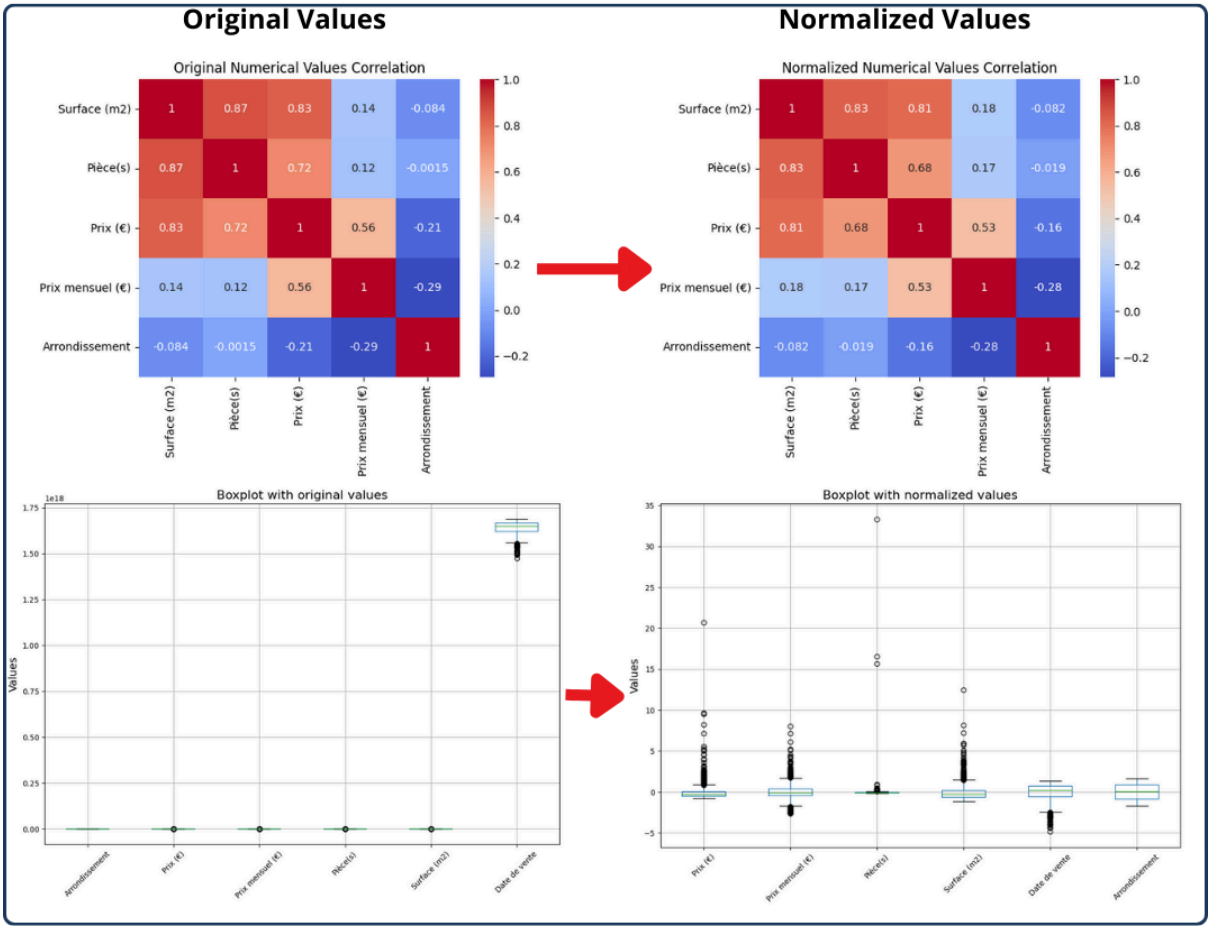


Figure 14 *: Graphiques permettant d'analyser l'effet de la normalisation*

Selon ***Figure 14***, la normalisation, plus précisément la standardisation, n'apporte pas beaucoup de modifications d'après les matrices de corrélations, mais permet de visualiser la distribution dans les box plot. Cependant, ce n'est pas une bonne idée de normaliser les données pour la prédiction.

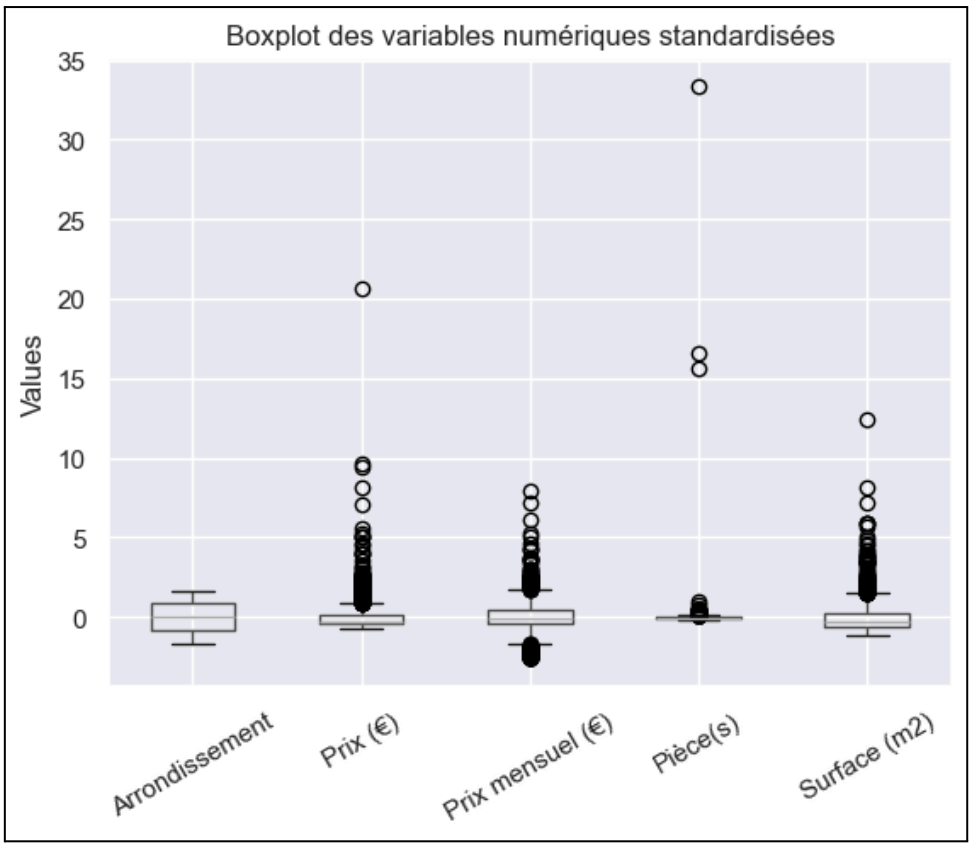
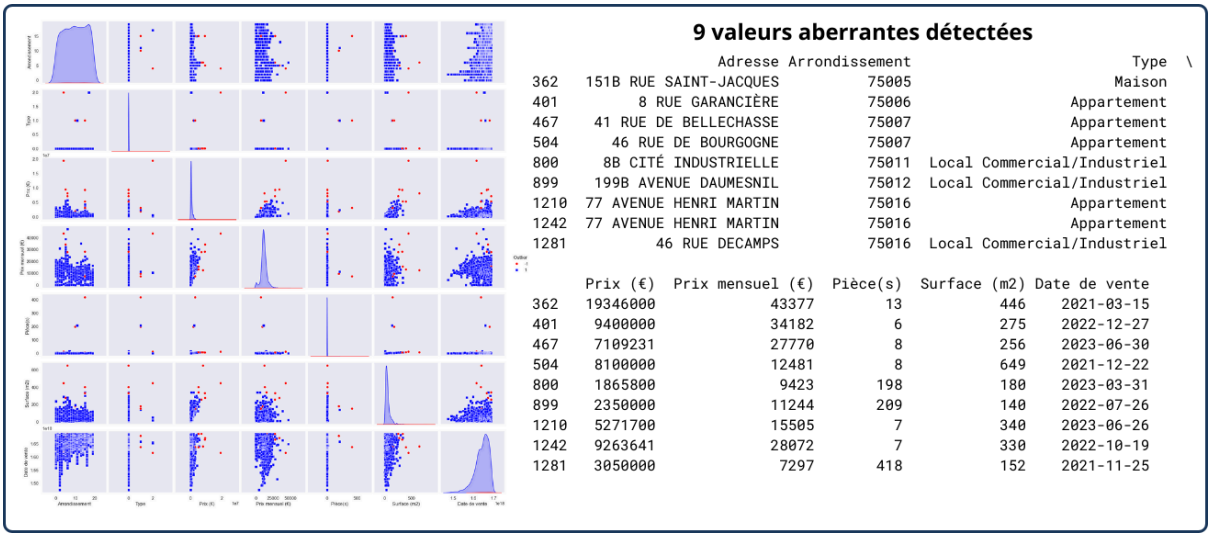


***Figure 15** : Boxplot des variables normalisés sans la Date de vente*

Dans la ***Figure 15***, Les variables "Arrondissement" et "Pièces" ont des plages de valeurs relativement étroites et moins de valeurs aberrantes, ce qui suggère une distribution plus serrée des données. En revanche, "Prix (€)", "Prix mensuel (€)" et "Surface (m²)" montrent une plus grande variabilité avec de nombreuses valeurs aberrantes, indiquant des écarts importants dans les données. La médiane semble relativement basse pour les catégories "Prix" et "Surface", ce qui pourrait indiquer une asymétrie dans la distribution des prix et des tailles des surfaces.

Détection de valeurs aberrantes



9 valeurs aberrantes détectées

	Adresse	Arrondissement	Type
362	151B RUE SAINT-JACQUES	75005	Maison
401	8 RUE GARANCIÈRE	75006	Appartement
467	41 RUE DE BELLECHASSE	75007	Appartement
504	46 RUE DE BOURGOGNE	75007	Appartement
800	8B CITÉ INDUSTRIELLE	75011	Local Commercial/Industriel
899	199B AVENUE DAUMESNIL	75012	Local Commercial/Industriel
1210	77 AVENUE HENRI MARTIN	75016	Appartement
1242	77 AVENUE HENRI MARTIN	75016	Appartement
1281	46 RUE DECAMPS	75016	Local Commercial/Industriel

	Prix (€)	Prix mensuel (€)	Pièce(s)	Surface (m2)	Date de vente
362	19346000	43377	13	446	2021-03-15
401	9400000	34182	6	275	2022-12-27
467	7109231	27770	8	256	2023-06-30
504	8100000	12481	8	649	2021-12-22
800	1865800	9423	198	180	2023-03-31
899	2350000	11244	209	140	2022-07-26
1210	5271700	15505	7	340	2023-06-26
1242	9263641	28072	7	330	2022-10-19
1281	3050000	7297	418	152	2021-11-25

Figure 16 *: Pair Plot indiquant les emplacements des valeurs aberrantes*

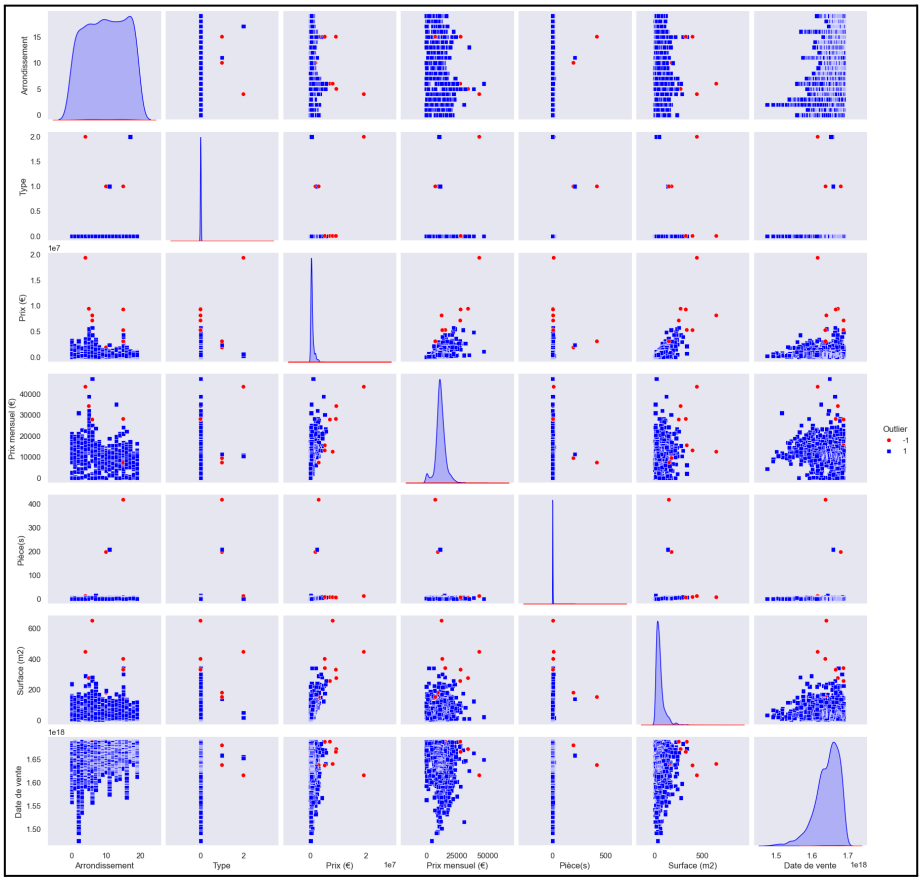


Figure 17 *: Pair Plot indiquant les emplacements des valeurs aberrantes*

Dans la ***Figure 17***, les points bleus représentent les données standard, tandis que les points rouges signalent des valeurs aberrantes. Il y a des distributions asymétriques pour le prix et la surface, comme en témoignent les longues queues sur les histogrammes de densité correspondants. Les valeurs aberrantes sont particulièrement notables dans la relation entre le nombre de pièces et les autres variables, ce qui suggère des anomalies spécifiques, comme des appartements exceptionnellement grands ou chers par rapport au reste de l'échantillon.

Nous pouvons constater des différences sur l'affichage dans les Pair Plot dans la ***Figure 18***, qui permet de faciliter également les prédictions, c'est-à-dire qu'il est plus facile de tracer des relations de régression lorsqu'on se débarrasse des valeurs aberrantes.

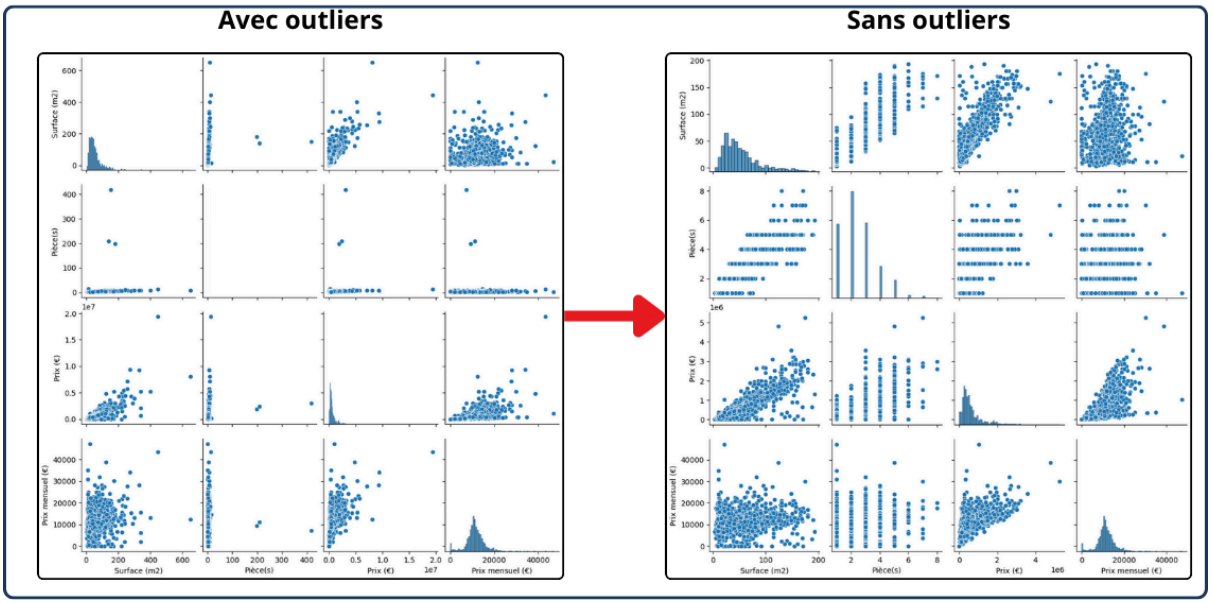


Figure 18 *: Pair Plot des features de prix (mensuel et total), de nombre de pièces et de surface avant et après avoir retiré les valeurs aberrantes*

Retirer certaines colonnes

Nous retirons les colonnes “Types” et “Adresses”, car “Types” ne contient que très peu de valeurs autres que “Appartement” et nous n’avons pas l’intention de faire des analyses sur les adresses.

Feature importance

D’après ***Figure 19***, la variable qui influence le plus sur la prédiction de prix est le prix mensuel. Comme nous l'avons vu précédemment, il est difficile de prédire le prix mensuel en fonction du prix total, donc lorsque nous le retirons, nous obtenons les résultats dans le graphe.

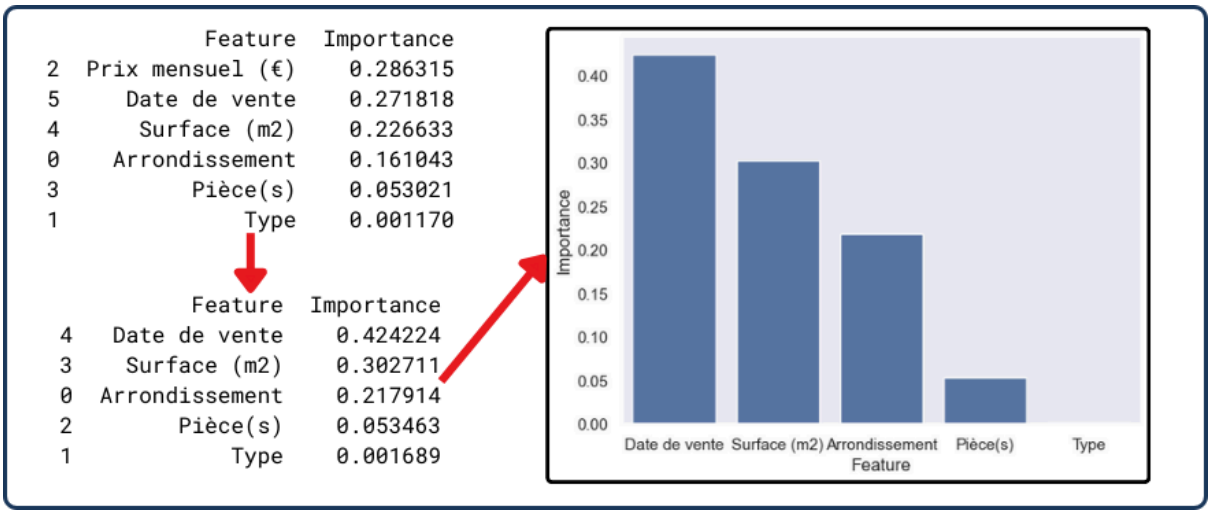


Figure 19 *: Histogramme de l’importance des variables pour la prédictions de prix*

Nous pouvons faire plusieurs constatations :

- Le prix mensuel ne sert à rien comme on l’a analysé avant, donc on le retire des variables les plus importants
- C’est étonnant que l’arrondissement affecte plus le prix que le nombre de pièces
- Il n’est pas étonnant que plus récemment une maison est vendue, plus elle est chère. On y trouve également ce phénomène pour les mêmes appartements.

Réduction de dimensions

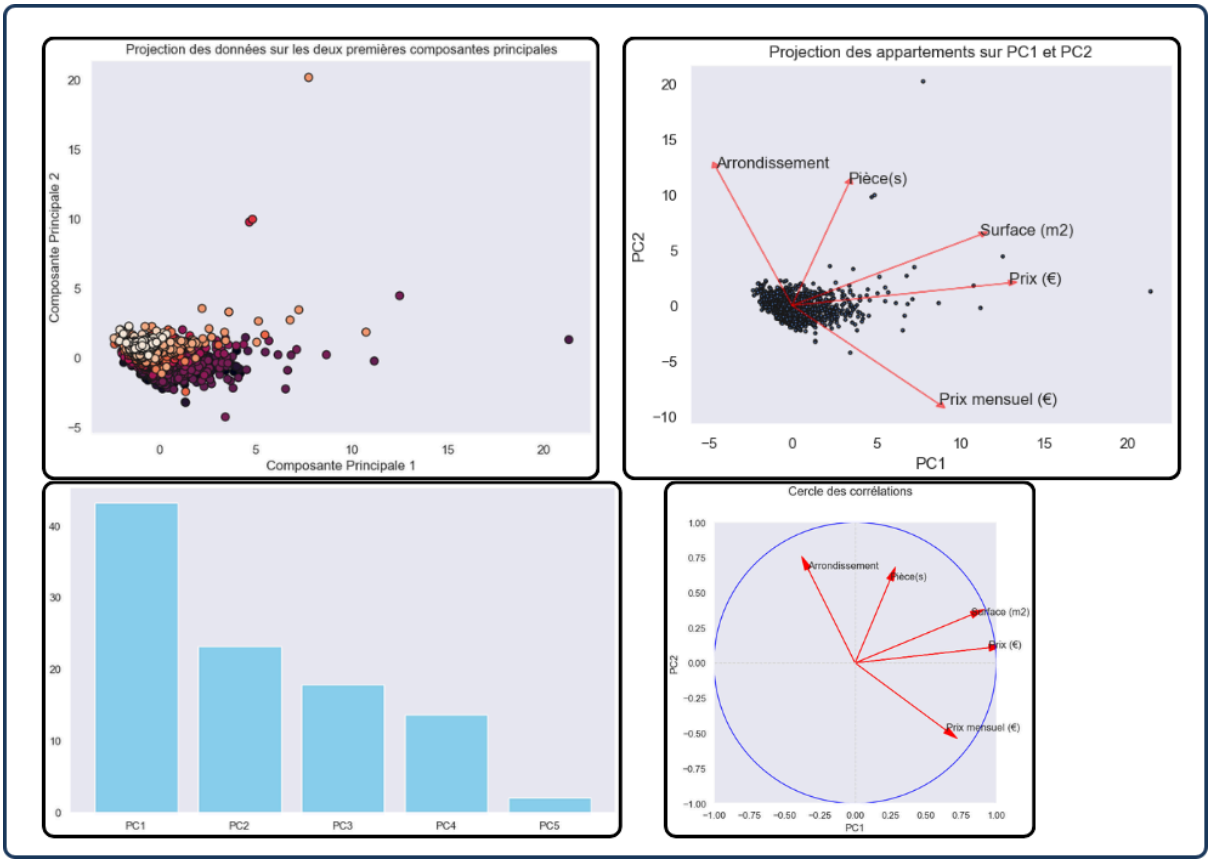


Figure 20 *: Différents graphiques permettant d'analyser la projection PCA sur 2 dimensions dont un histogramme sur la proportion de variance de chaque composant*

Avec les 2 premières composantes, il n'y a pas assez de variance pour distinguer les appartements en fonction de son arrondissement comme on peut le voir dans la **Figure 20**.

D'après l'histogramme de la proportion de variance, PCA semble être efficace puisque les trois premières composantes expliquent déjà 84% de la variance totale, et les quatre premières en expliquent 98%. Cela signifie que les données peuvent être réduites à 3 ou 4 dimensions tout en conservant la majorité de l'information.

D'après le cercle de corrélation, le prix d'achat et la surface habitable sont étroitement liés à PC1, indiquant que généralement, une plus grande surface implique un prix plus élevé pour les biens immobiliers. Le loyer mensuel est aussi lié au PC1, mais dans une moindre mesure, suggérant que d'autres éléments influencent également le loyer au-delà de la surface de la propriété. D'autre part, l'arrondissement montre une corrélation plus marquée avec la deuxième composante principale (PC2), ce qui suggère qu'il affecte les valeurs immobilières d'une manière qui ne se résume pas uniquement à une augmentation de la

surface ou du prix. Le nombre de pièces a une influence sur les deux composantes, bien que moins prépondérante que la surface ou le prix. Toutes ces variables sont relativement proches du cercle de corrélation, ce qui indique qu'elles sont assez bien représentées par ces deux composantes principales, bien qu'une certaine quantité d'informations soit perdue, comme le montre le fait qu'aucune variable n'est parfaitement alignée sur le cercle.

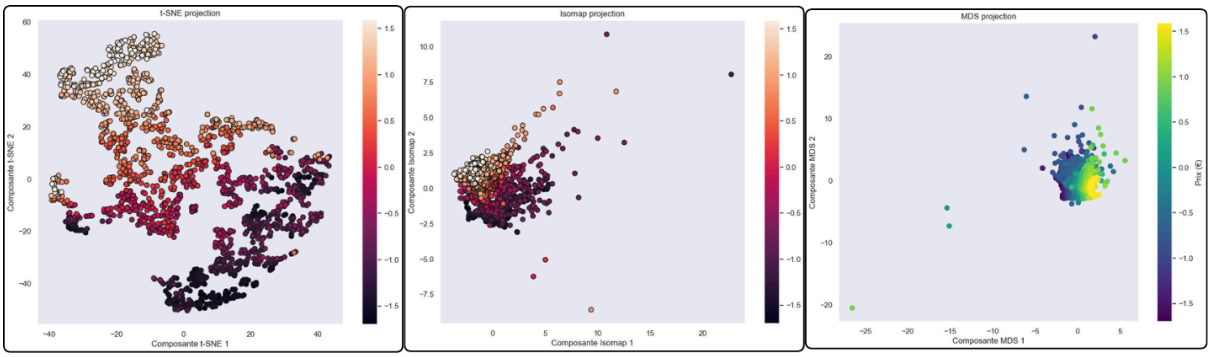


Figure 21 *: Différents techniques de réductions de dimensions sur les 5 variables (Arrondissement, Pièce(s), Surface (m2), Prix (€), Prix mensuel (€))*

D'après ***Figure 21***, t-SNE à l'air de mieux capturer la variance des données en 2 composantes principales quand on le compare à PCA, MDS et Isomap. L'isomap ne capture pas correctement la variance de chaque donnée. Changer le nombre de composantes principales ne modifie pas beaucoup le scatterplot. Le MDS ne capture pas correctement la variance des données en 2 composantes principales et prend beaucoup de temps.

Modelling

Prédiction du prix

	Variables	VIF
0	const	1.000000
1	Surface (m2)	1.039003
2	Pièce(s)	1.033087
3	Arrondissement	1.007166

***Table 5** : Valeurs de VIF par variable*

Si les valeurs du VIF sont supérieur à 5-10, il y a une multicolinéarité, comme on peut le constater dans la ***Table 5***, ils sont tous autour de 1 donc pas de problème de multicolinéarité

OLS Regression Results

Dep. Variable:	Q("Prix (€)")	R-squared:	0.670
Model:	OLS	Adj. R-squared:	0.669
Method:	Least Squares	F-statistic:	52.43
Date:	Thu, 23 May 2024	Prob (F-statistic):	2.28e-32
Time:	03:36:07	Log-Likelihood:	-1430.6
No. Observations:	1654	AIC:	2869.
Df Residuals:	1650	BIC:	2891.
Df Model:	3		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	8.52e-14	0.014	5.89e-12	1.000	-0.028	0.028
Q("Surface (m2)")	0.8024	0.084	9.520	0.000	0.637	0.968
Q("Pièce(s)")	0.0210	0.029	0.726	0.468	-0.036	0.078
Q("Arrondissement")	-0.0909	0.011	-8.501	0.000	-0.112	-0.070

Omnibus:	2581.668	Durbin-Watson:	1.822
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3595570.070
Skew:	9.095	Prob(JB):	0.00
Kurtosis:	230.688	Cond. No.	1.22

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

***Table 6** : Résumé du modèle*

	df	sum_sq	mean_sq	F	PR(>F)
Q("Surface (m2)")	1.0	1093.760234	1093.760234	3304.389065	0.000000e+00
Q("Pièce(s)")	1.0	0.517396	0.517396	1.563118	
Q("Arrondissement")	1.0	13.568649	13.568649	40.992617	
Residual	1650.0	546.153722	0.331002	NaN	

```
Q("Pièce(s)")        2.113865e-01
Q("Arrondissement")  1.986956e-10
Residual                     NaN
```

***Table 7** : Résultat du test ANOVA*

Les variables Surface et Arrondissement sont des variables explicatives significatives pour le prix des appartements, comme on peut le constater avec une p-value nulle et une extrêmement petite respectivement. Contrairement à la variable Pièces qui possède une p-value de 0.21 > 0.05, ce qui signifie que cette variable n'apporte pas d'information supplémentaire significative pour expliquer la variance des prix des appartements.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Q("Prix (€)")   R-squared:                       0.658
Model:                            OLS   Adj. R-squared:                  0.657
Method:                 Least Squares   F-statistic:                     32.44
Date:                Thu, 23 May 2024   Prob (F-statistic):           3.93e-20
Time:                        03:36:08   Log-Likelihood:                -1126.4
No. Observations:                1157   AIC:                             2261.
Df Residuals:                    1153   BIC:                             2281.
Df Model:                           3
Covariance Type:                  HC3
================================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                0.0051      0.018      0.285      0.776      -0.030       0.040
Q("Surface (m2)")        0.8184      0.106      7.749      0.000       0.611       1.025
Q("Pièce(s)")            0.0197      0.030      0.657      0.511      -0.039       0.078
Q("Arrondissement")     -0.0911      0.014     -6.469      0.000      -0.119      -0.063
==============================================================================
Omnibus:                     1799.294   Durbin-Watson:                   1.997
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1977608.339
Skew:                           8.907   Prob(JB):                         0.00
Kurtosis:                     204.754   Cond. No.                         1.29
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
```

***Table 8** : Résumé du modèle*

<u>Significativité du Modèle</u>

Le coefficient de détermination (R^2) est de 0.658, ce qui signifie que 65.8% de la variance du prix des appartements est expliquée par les variables incluses dans le modèle. Le R^2 ajusté, qui prend en compte le nombre de prédicteurs, est de 0.657, suggérant que le modèle a une bonne capacité explicative avec un ajustement minimal pour le nombre de variables. La statistique F est de 32.44, avec une p-value associée très faible (3.93e-20), confirmant que le modèle global est statistiquement significatif.

<u>Interprétation des Coefficients</u>

L'intercept a une p-value de 0.776, ce qui n'est pas statistiquement significatif. Cela signifie que, selon ce modèle, la constante n'ajoute pas de valeur significative à la prédiction du prix des appartements. Le coefficient de la surface (Q("Surface (m2)")) a une p-value de 0.000, indiquant qu'il est hautement significatif. Cela montre que la surface en mètres carrés est un facteur important pour expliquer les variations du prix des appartements. Le coefficient du nombre de pièces (Q("Pièce(s)")) a une p-value de 0.511, ce qui n'est pas statistiquement significatif. Cela suggère que, dans ce modèle, le nombre de pièces n'a pas un impact significatif sur le prix des appartements. En revanche, le coefficient de l'arrondissement (Q("Arrondissement")) a une p-value de 0.000, indiquant qu'il est également un facteur significatif. Un arrondissement plus élevé est associé à une diminution du prix des appartements, toutes choses égales par ailleurs.

Diagnostics du Modèle

Les diagnostics du modèle révèlent quelques préoccupations. La statistique Omnibus et sa p-value associée indiquent que les résidus ne sont pas normalement distribués, ce qui pourrait suggérer des problèmes avec les hypothèses du modèle. La statistique de Durbin-Watson est de 1.997, suggérant une faible autocorrélation des résidus, ce qui est favorable pour l'indépendance des erreurs. La statistique Jarque-Bera est extrêmement élevée, indiquant que les résidus s'écartent fortement d'une distribution normale. Cela est corroboré par la kurtosis de 204.754, qui suggère une distribution des résidus avec des queues très lourdes, signalant la présence de nombreux outliers. Enfin, le numéro de condition est de 1.29, ce qui est relativement faible et n'indique pas de problèmes de multicolinéarité dans ce modèle.

```
Mean Squared Error: 0.3801101286920795
Mean Absolute Error: 0.21152228324706956
R²: 0.7271950333566979
Mean Absolute Percentage Error: 1.8449259205727384% accuracy
```

***Table 9** : Résultats de résidus (coût)*

Après avoir trouvé un modèle linéaire pour notre jeu de données sur les données quantitatives, on va faire intervenir la date.
On veut pouvoir prédire l'évolution d'un bien par rapport au temps, par exemple quel sera le prix d'une maison de 30m carré qui sera vendu en 2050.
Pour cela, on va créer une nouvelle variable quantitative basée sur les dates que nous avons.

Prédiction avec la date

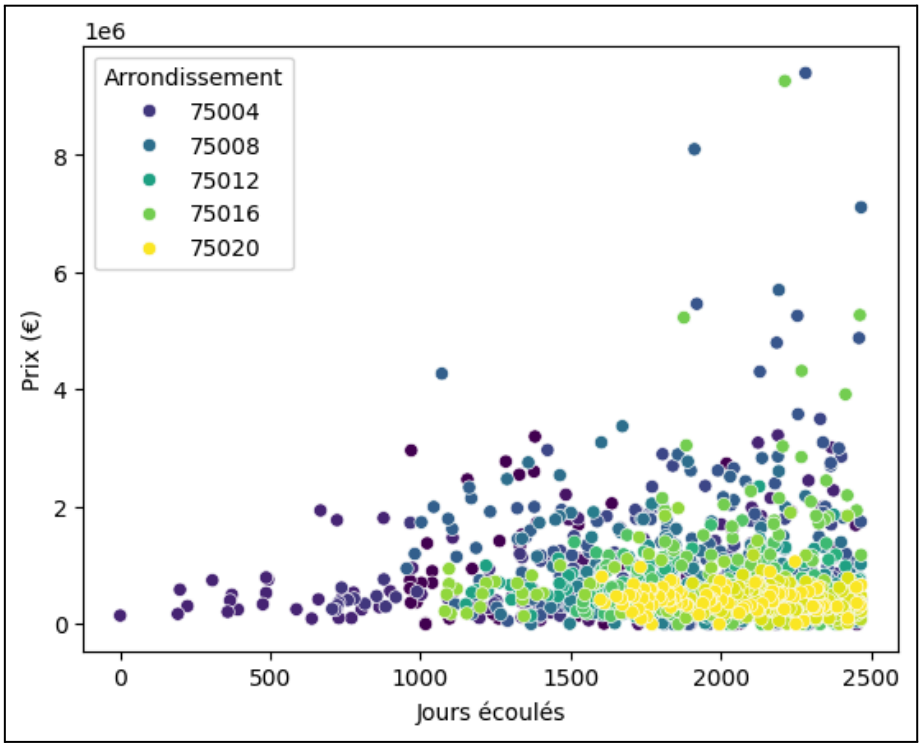


***Figure 22** : Nuages de points représentant le prix en fonction des jours écoulées (inflation) par zone géographique (arrondissement)*

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Q("Prix (€)")   R-squared:                       0.706
Model:                            OLS   Adj. R-squared:                  0.705
Method:                 Least Squares   F-statistic:                     99.98
Date:                Thu, 23 May 2024   Prob (F-statistic):           4.46e-73
Time:                        03:36:09   Log-Likelihood:                -16601.
No. Observations:                1155   AIC:                         3.321e+04
Df Residuals:                    1150   BIC:                         3.324e+04
Df Model:                           4
Covariance Type:                  HC3
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                     1.17e+09   1.64e+08      7.135      0.000    8.48e+08    1.49e+09
Q("Pièce(s)")               1819.9126   4347.926      0.419      0.676   -6701.867    1.03e+04
Q("Surface (m2)")           1.373e+04   1076.752     12.748      0.000    1.16e+04    1.58e+04
Q("Jours écoulés")            38.8819     32.636      1.191      0.233     -25.083     102.846
Q("Arrondissement")         -1.56e+04   2185.817     -7.135      0.000   -1.99e+04   -1.13e+04
==============================================================================
Omnibus:                      869.162   Durbin-Watson:                   1.978
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           152208.270
Skew:                           2.536   Prob(JB):                         0.00
Kurtosis:                      59.009   Cond. No.                     1.05e+09
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
```

```
[2] The condition number is large, 1.05e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Table 10 *: Résumé du modèle*

On constate un nombre de conditions très large, et en faisant des tests ce n'est pas un problème de multicolinéarité mais plutôt de data scaling. Nous ne l'avons pas fait pour mieux visualiser les prix qui ont été prédit par le modèle

```
Mean Squared Error: 380490.19945142383
Mean Absolute Error: 187849.39037857708
R²: 0.7483681258075996
Mean Absolute Percentage Error: 13.94390680854607% accuracy
```

Table 11 *: Résultats de résidus (coût)*

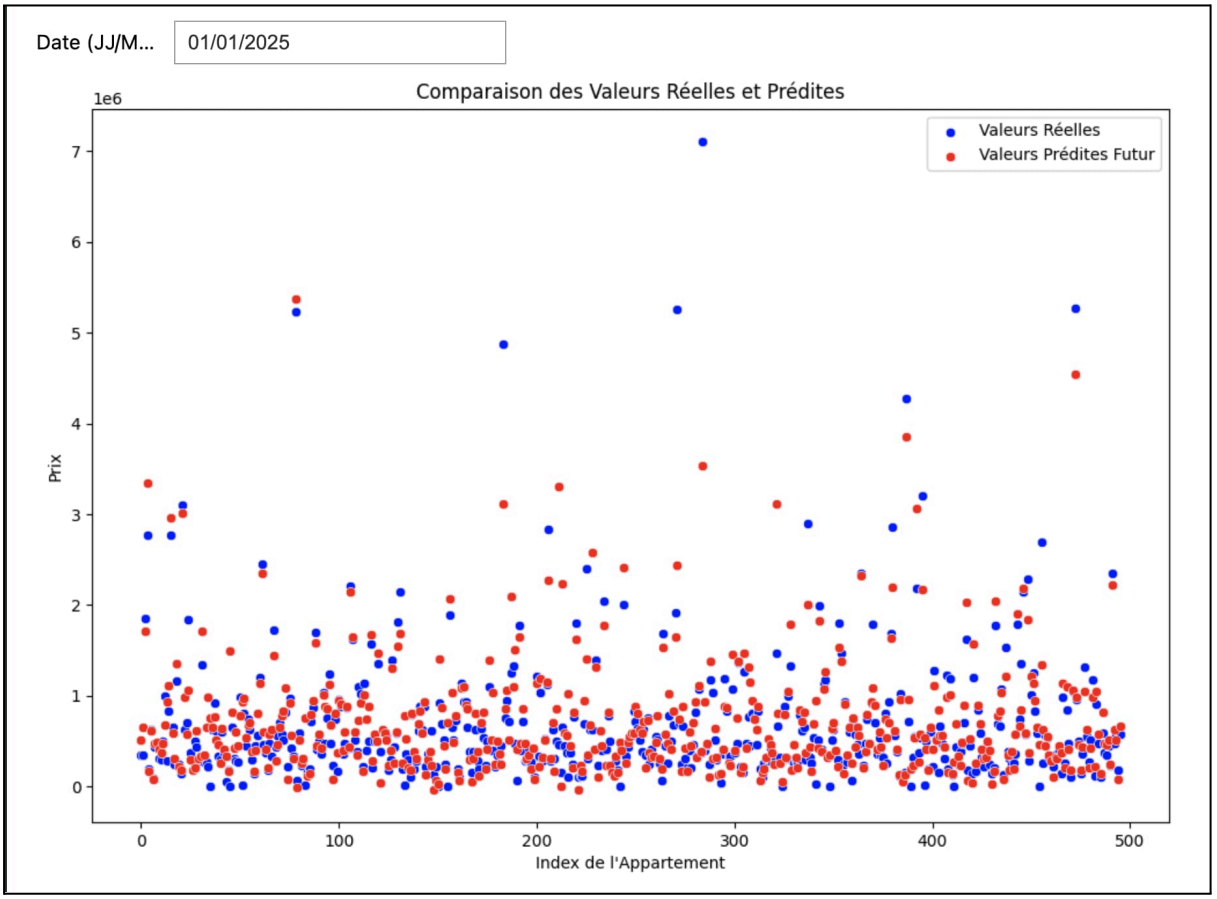


Figure 23 *: Comparaison des Valeurs Réelles et Prédites*

Conclusion

Nos expérimentations ont montré que nous avons pu prédire le prix d'un appartement à Paris en utilisant immo-data contenant les informations les plus importantes, soit le prix et les dates de ventes ainsi que d'autres caractéristiques pour évaluer le marché de l'immobilier. Les prédictions donnent plutôt de bons résultats. Nos résultats confirment également l'inflation sur le prix pour les prochaines années.

Références

- Site de immo-data : https://www.immo-data.fr/
- Site de Century21 : https://www.century21.fr/
- Site de housedata de Kaggle : https://www.kaggle.com/datasets/shree1992/housedata