UNIVERSITÉ SAINT JEAN
Saint Jean Ingénieur

RAPPORT
DATASCIENCE
MASTER I

ÉTUDE DE L'ÉVALUATION DES BIENS IMMOBILIERS

Zone : Sindian Dist., New Taipei City, Taiwan.
Date de collecte des données : 2012 - 2013

Étudiant :
OLONGO ONDIGUI JAMES WILLIAM

Enseignant :
Pr. NGUEFACK TSAGUE GEORGES

Année académique : 2022 - 2023

Table des matières

Introduction

Le domaine de l'immobilier joue un rôle très important dans notre société. Le problème de l'évaluation des biens immobilières est un problème pour les investisseurs qui y souhaitent une meilleure transparence et une rentabilité dans le temps de leurs biens. Mais souvent complexe car peut varier suivant le temps ce qui l'associe à des problème de régression, cela ne facilite pas la tâche pour ces derniers.

Il est devenu évident que pour vraiment comprendre les évaluations immobilières, il faut comprendre quels sont les paramètres qui peuvent impacter sur une fluctuation de ces prix, de manière direct et indirecte. Ces paramètres qui influencent sur l'évaluation immobilière sont extrêmement complexes. Parfois juste à cause d'un accès compliqué à la localité, ou peu de commerce dans la zone, voir même la localisation au site prix peuvent entraîner une variation.

Il est donc actuellement compliqué de spéculer sur cette évaluation. C'est pourquoi des analystes de données collectent la plupart du temps des données liées à des localités spécifique et essayent de manipuler ces données pour pourvoir comprendre cette variation et prédire comment cela pourrait se comporter dans le temps. Souvent, les ensembles données utilisées sont très maigres en ce sens qu'ils ne contiennent des données sur une période d'un an voir deux ans, il est clair que cela ne représente pas beaucoup de données pour une localité, ou il n'y a pas assez de variable pour mener à bien l'étude, cela pourrait en gros biaiser nos rapports si l'on considère que les données sur l'évaluation immobilières peuvent être très bruité.

En particulier, nous nous concentrerons sur un ensemble de données collecté dans la localité de Sindian Dist., New Taipei City, Taiwan durant l'année 2012 à 2013 et notre travail consistera à démontrer si des facteurs tels que l'âge d'une maison, sa localisation, sa proximité à des magasins peuvent influencer sur les prix immobiliers. Mais tout d'abord nous décrirons les données afin d'assurer la bonne compréhension lors des manipulations, ensuite nous définirons les objectifs de nos travaux puis nous montrerons les résultats pour afin terminer avec une discussions de nos résultats.

1. Description des données

Notre jeu de donnée provient du marché de l'évaluation dans le district de Sindian, New Taipei City, Taiwan collecté durant l'année 2012 et 2013. L'ensemble de données ont été divisés de manière aléatoire en un ensemble de données de formation (2/3 des échantillons) et un ensemble de données de test (1/3 des échantillons).

Informations sur les attributs

Colonne	Type	Libellé
X1	Float	Date de la transaction
X2	Float	Age de la maison
X3	Float	Distance de la station de métro la plus proche
X4	Int	Nombre de magasins à proximité à pied dans le cercle d'habitation
X5	Float	Coordonnées géographiques(Latitude)
X6	Float	Cote géographiques(Longitude)
Y	Float	Prix de l'unité de surface
X7	String	Concaténation de X5 et X6

2. Objectifs

L'évaluation boursières n'est pas un domaine qui relève du hasard, juste à cause d'un petits facteurs tel que la zone immobilières peut voir ses prix chuter ou augmenter et ainsi pour comprendre comment cette fluctuation de prix selon une zone à une autre, notre travail consistera après avoir manipuler ce jeu de donnée qui a été mis à notre disposition de déterminer les indices d'une variation des prix immobiliers.

Il sera question en effet de pouvoir voir comment l'âge des maisons dans une zone peut fluctuer sur l'indice immobilier, comment la présence des magasins dans une zone influencent ces prix, la proximité d'une station de métro peut jouer un rôle capital sur cette variation et enfin si les prix immobiliers peuvent varier selon la zone géographique.

3. Résultats Descriptifs

a. Résumé des données statistique

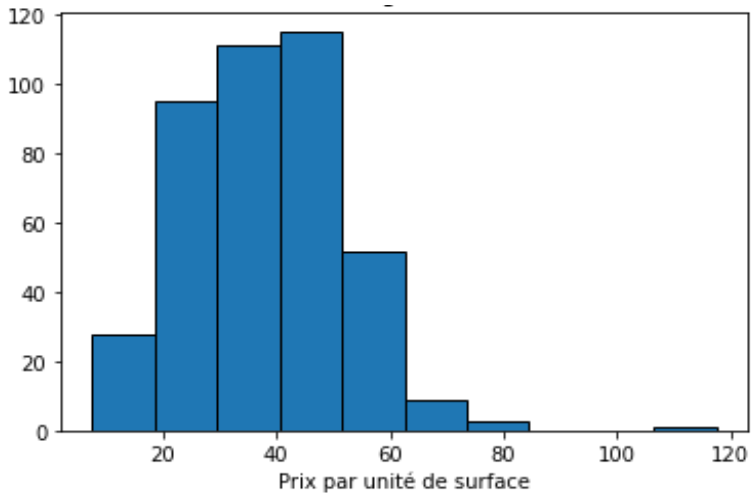
Résumé de certaines données statistiques telles que le centile, la moyenne et la norme des valeurs numériques de la série

	X1	X2	X3	X4	Y
COUNT	414.00	414.00	414.00	414.00	414.00
MOYENNE	2013.14	17.71	1083.88	4.09	37.98
ÉCART TYPE	0.28	11.39	1262.10	2.94	13.60
MINIMUM	2012.66	0.00	23.38	0.00	7.60
25%	2012.91	9.02	289.32	1.00	27.70
50%	2013.16	16.10	492.23	4.00	38.45
75%	2013.41	28.15	1454.27	6.00	46.60
100%	2013.58	43.80	6488.02	10.00	117.50

b. Distribution de la variable d'intérêt (Prix par unité de surface)

Compte tenu du fait que nous intéressons à la valeur d'un bien immobilier, il est pertinent d'observer l'intervalles de de variations des prix par unité de surface pour avoir une idée du marché.

Graphique : Histogramme de la répartition des prix par unité de surface



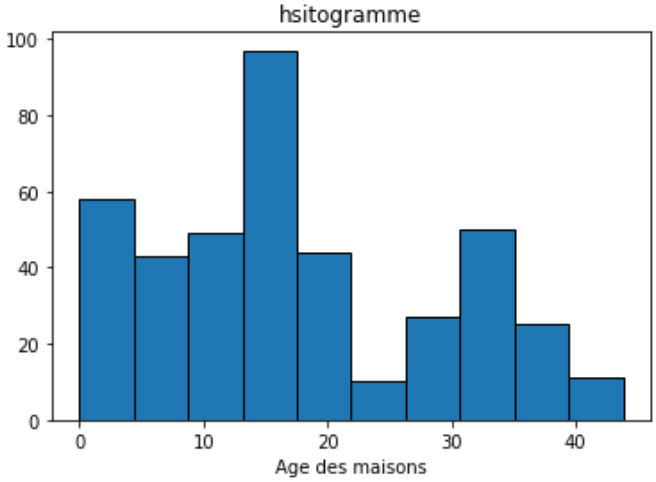
Source : Real estate valuation are collected from Sindian Dist 2012-2013

Nous observons donc que Le prix de l'unité de surface moyen est de 37.98 et représente presque 120 habitations / 414, mais les logements moyen oscillent entre 30 et 50. Tout du moins nous notons aussi des prix élevé allant jusqu'à 117.5 mais ne représente qu'à peine 3 habitations et aussi une minimum de 7.60 représentant à peine 30 habitations.

c. Distribution de la variable de l'âge de la maison

L'observation de la distribution des âges des maisons serait intéressante afin d'avoir une idée d'ensemble sur la répartition des âges des maisons

Graphique : Histogramme de la répartition des âges des maisons



Source : Real estate valuation are collected from Sindian Dist 2012-2013

L'âge des maisons varient entre 0 à 43 ans, avec l'âge moyen des maisons tournant autour de 18ans.
Nous observons une répartition quasiment en part égale entre des maisons de moins de de 5ans, 10 ans et 35 ans

d. Distribution de la variable nombre de magasin à proximité

Tableau : Nombre de magasin à proximité

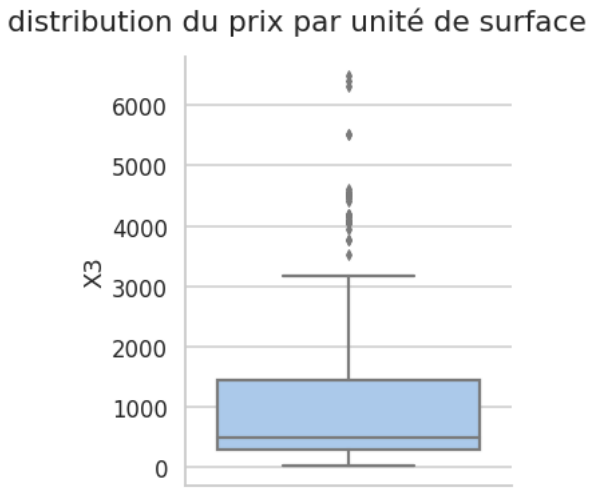
	X4 (NOMBRE DE MAGAZIN À PROXIMITÉ)
MOYENNE	4.09
ÉCART TYPE	2.94
MINIMUM	0.00
25%	1.00
50%	4.00
75%	6.00
100%	10.00

Source : Real estate valuation are collected from Sindian Dist 2012-2013

En moyenne nous recensons 4 magasins par secteurs d'habitations, mais il existe bien des zones ayant un pic de magasin allant jusqu'à 10 et des zones d'habitions n'ayant quasiment aucun magasins.

e. Distributions sur la variable distance jusqu'à la station la plus proche

Graphique : Catplot sur la variable Distance jusqu'à la station la plus proche



Source : Real estate valuation are collected from Sindian Dist 2012-2013

Nous observons à partir de ce catplot que plusieurs habitations sont à moyenne à 1000m de la station de métro et avec des habitations plus proche du métro à des distance de moins de 300m et d'autres plus à plus de 6000m, mais on observe une plus forte concentration des maisons à des distance du métro entre l'intervalle 300m à 1500m.

4. Résultats Analytiques

Afin de mieux comprendre la variation des prix immobiliers de notre jeu de données, nous effectuerons une série de tests statiques, pour observer comment agit les variables et à la fin nous pourrons apporter une discussion de ces derniers.
Nos tests analytiques seront centrés sur la variable d'intérêt celle du prix de l'unité de surface et nous but sera de chercher à comprendre quel rôle joue les variables tels que : l'âge de la maison, le nombre de magasin, la distance du métro la plus proche et la position géogrpahique sur celle-ci.

a. Prix de l'unité de surface par rapport à l'âge de la maison

Nous commencer par grouper le prix par unité de surface avec l'âge de maison afin d'observer le comportement de celle-ci et de déterminer si l'un joue un rôle sur l'autre.

Tableau : Comparaison du prix par unité de surface par rapport à l'âge de ma maison

	Moyenne	Minimum	Maximum
X2			
0.0	54.13	37.9	73.6
1.0	50.70	50.7	50.7
1.1	49.78	45.1	54.4
1.5	48.70	47.7	49.7
1.7	50.40	50.4	50.4
…	…	…	…
40.9	54.35	41.0	67.7
41.3	47.90	35.1	60.7
41.4	63.30	63.3	63.3
42.7	35.30	35.3	35.3
43.8	42.70	42.7	42.7

Source : Real estate valuation are collected from Sindian Dist 2012-2013

b. Prix de l'unité de surface par rapport aux nombres de magasin

Ici nous jouons avec le nombre de magasin à proximité pour observer le comportement qui en résulte.

Tableau : Comparaison du prix par unité de surface par rapport au nombre de magasin

	Moyenne	Minimum	Maximum
X4			
0	26.46	11.6	55.3
1	31.63	11.2	117.5

2	31.41	20.9	50.5
3	29.53	17.7	61.5
4	37.47	21.8	62.9
5	44.72	22.8	60.7
6	46.95	7.6	73.6
7	43.84	25.0	62.1
8	44.69	26.5	67.7
9	51.73	32.4	78.3
10	48.43	37.9	61.9

Source : Real estate valuation are collected from Sindian Dist 2012-2013

c. Prix de l'unité de surface par rapport à la station de métro la plus proche

Afin de bien mener notre étude nous allons aussi grouper la prix de l'unité de surface avec la distance de la station de métro la plus proche et de voir s'il y a causalité entre les deux.

Tableau : Comparaison du prix de l'unité de surface par rapport à la distance de métro la plus proche

	Moyenne	Minimum	Maximum
X3			
23.38	48.70	47.70	49.7
49.66	57.30	56.80	57.8
56.47	56.66	53.50	62.1
57.58	42.70	42.70	42.7
82.88	46.60	46.60	46.6
…	…	…	…
4605.74	13.40	13.40	13.4
5512.03	18.10	17.40	18.8
6306.15	15.00	15.00	15.0
6396.28	12.20	12.20	12.2
6488.02	11.20	11.20	11.2

Source : Real estate valuation are collected from Sindian Dist 2012-2013

d. Prix de l'unité de surface par rapport à la zone géographique

Tableau : Comparaison du prix par unité par rapport à la distance la localisation

	Moyenne	Minimum	Maximum
X3			
(24,93207 ; 121,51597)	29.30	29.30	29.30
(24,93363 ; 121,51158)	45.10	45.1	45.1
(24,94155 ; 121,50381)	16.45	15.90	20.0
(24,9746 ; 121,53046)	117.5	117.5	117.5
…	…	…	…
(24,998 ; 121,5155)	41.20	41.20	41.20
(25,01459 ; 121,51816)	27.30	27.30	27.30

Source : Real estate valuation are collected from Sindian Dist 2012-2013

e. Test de normalité

Notre objectifs ici est de vérifier si certaines des variables suivent une normale

- **Variable prix par unité de surface**

```
        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.97275, p-value = 5.411e-07
```

Source : Code R

p-value renvoyé n'est pas supérieur à 0.05 pour rester sur l'hypothèse nulle d'une distribution normale

- **Variable âge de la maison**

```
$X2

        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.94674, p-value = 4.798e-11
```

Source : Code R

p-value renvoyé n'est pas supérieur à 0.05 pour rester sur l'hypothèse nulle d'une distribution normale

- **Variable nombre de métro à proximité des habitations**

```
$X3

        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.7381, p-value < 2.2e-16
```

Source : Code R

p-value renvoyé n'est pas supérieur à 0.05 pour rester sur l'hypothèse nulle d'une distribution normale

- **Variable nombre de magasins à proximité des zones d'habitation**

```
$X4

        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.93737, p-value = 3.461e-12
```

p-value renvoyé n'est pas supérieur à 0.05 pour rester sur l'hypothèse nulle d'une distribution normale

f. Test de comparaison des moyennes (cas du prix de l'unités de surface et le nombre de magasin)

Nous voulons vérifier si le prix de l'unité de surface et le nombre de magasin dépendent ou non du fruit du hasard, de ce fait nous allons effectuer un test de student.

- **Comparons ces deux échantillons pour voir s'ils sont significativement différents**

```
pValue est de  : 1.6978283752001522e-249
```

Source : Code Source Python

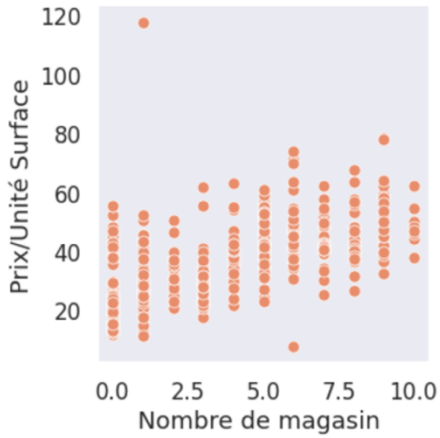
La valeur renvoyée de p-value est inférieur à 0.01 donc il y a une différence très significative.

g. Nuage de point entre des variables

Nous allons représenter certaines variables suivant le nuage de point afin de présenter la mesure de deux variables liées. Le nuage de points est particulièrement utile lorsque les valeurs des variables sur l'axe des y dépendent des valeurs de la variable de l'axe des x.

- **Prix/Unité Surface et nombre de magasins**

Figure : Nuage de points: Prix/Unité Surface et nombre de magasins

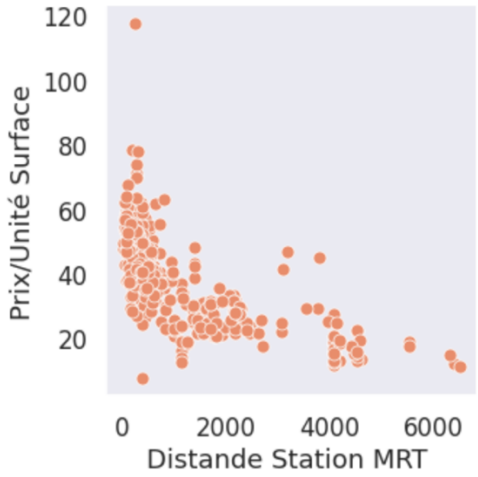


Source : Code Python

A partir de ce graphe ci-dessous, nous observons que la relation entre ces variables deux variable sont linéaire, et aussi nous notons que certaines points sont très peu dispersés.

- **Prix/Unité Surface et station de métro la plus proche**

Figure : Prix/Unité Surface et station de métro la plus métro

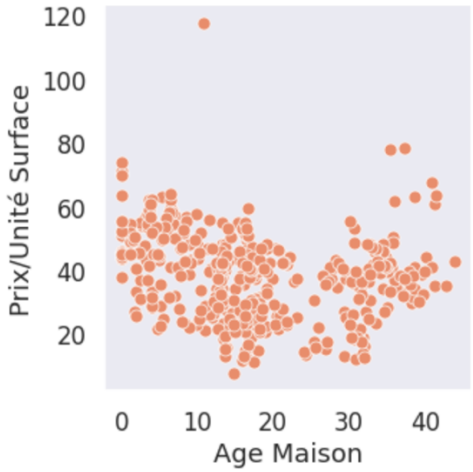


Source : Code Python

A partir de ce graphe ci-dessous, nous observons que la relation entre ces variables deux variable est une relation négative ou inverse , et nous notons aussi que certaines points sont très peu dispersés.

- **Prix/Unité Surface et âge de la maison**

Figure : Nuage de point Prix/Unité Surface et âge de la maison

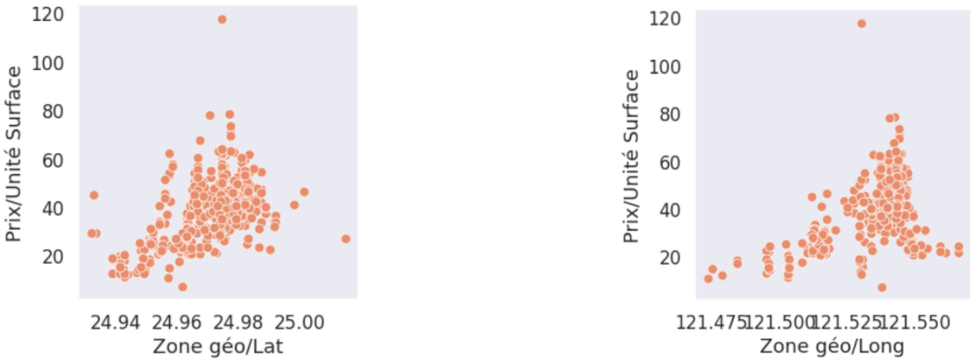


Source : Code Python

A partir de ce graphe ci-dessous, nous observons que la relation entre ces variables deux variable est une relation négative ou inverse , et nous notons aussi que certaines points sont très peu dispersés.

- **Prix/Unité Surface et zone géographique**

Figure : Nuage de point entre le prix par unité de surface et les coordonnées géographique



Source : Code Python

Ces graphes nous permettent de conclure que la relation entre et les coordonnées géographique et la variable d'intérêt n'est pas très dispersés.

5. Discussions

Les séries de tests de statistiques effectuées dans la section suivante nous ont permis de faire des manipulations toute autour de la variable d'intérêt « prix par unité de surface », le but de ces tests étaient de comprendre l'évaluation immobilière des maisons dans ce district. Nous avons joués avec des variables tels que : le nombre de magasins dans le cercle d'habitation, la proximité d'une station de métro, l'âge de la maison et la position géographique.

Nous apporterons une analyse personnelle de ces manipulations et nous essayerons de déterminer quels paramètres influences le plus sur la variable d'intérêt.

a. Cas 1 : Impact de l'âge de la maison sur le prix

Ici nous pourrions conclure au premier abord qu'une maison est plus chère lorsqu'elle est nouvellement construite, car nous observé dans le tableau de la section 4.a que les maisons ayant l'un des plus grand maximum est une maison de moins de 1 an avec une moyenne assez élevé à 54.13, mais nous observons dans le nuage de point de la section 4.e (3^{e} graphe) que la relation entre ces deux variables est une relation négative. Cela se confirme vu que la maison la plus chère est une maison de 10 ans et nous avons aussi des maison de 40 ans avec de même que des maisons de moins de 2 ans

Il en va de soi que l'âge de la maison ne joue pas un grand rôle sur le prix par unité de mesure, et même si il joue un rôle il est très minime. il faudrait essayer d'observer si d'autre variable ne joue pas un impact plus considérable.

b. Cas 2 : Impact du nombre de magasin sur le prix

Nous observons très vite à partir du tableau 4.b que les prix sont grandissant avec le nombre de magasins dans le zone d'habitation.

Les habitations ayant plus de 6 magasins à proximité recensent des plus grands prix par unité de surface, et cela se confirme aussi en observant le nuage de point de la section 4.g (première figure) qui permet de nous montrer que ces deux variables sont linéaire et très peu dispersé, et même le test normalité bien qu'il ne soit pas très concluant il obtient il valeur plus proche de 0,05 que les autres variables.

Constatant ces faits, il est clair que ce paramètre impacte fortement sur la variable d'intérêt, et sa valeur faible entraine aussi une faible valorisation du bien immobilier.

Mais nous observons une habitation avec un prix de l'unité de surface à 117.5 mais qui n'a qu'un magasin dans le secteur, cela nous permet de comprendre que cette variable n'est pas la seule à impacter sur le prix, il faudra aussi qu'on observe d'autre paramètre comme la zone géographique.

c. Cas 3 : Impact de la zone géographique sur le prix.

Il va de soi que la zone géographique joue un rôle plus ou moins directe dans l'estimation d'une habitations. Une augmentation du prix pourrait être le cas d'une zone urbaine et une baisse de prix pourrait être un quartier en périphérie.

Les nuages de points obtenus sur les graphes 4.g (4e graphe) nous confirment cette hypothèse suivant laquelle la zone géographique pourrait jouer un rôle, car la concentration peu dispersée des points marque une corrélation entre ces derniers.

Mais cet impact n'est pas très considérable, ou tout au moins ne pourra pas à lui seul impacter sur cette fluctuation de prix, d'autres paramètres plus important peuvent jouer ce rôle.

d. Cas 4 : Impact d'une proximité du métro sur le prix

Logiquement, la présence d'une station de métro proche d'une habitation pourra faciliter une hausse des prix immobiliers, car cela facilite la mobilités des habitants.

Le tableau obtenu à la section 4.e essaye de nous orienter dans ce sens car nous constatons fortement que les zones d'habitation de moins de 100m du métro sont celle ayant des prix moyen élevé, mais ceux à plus de 4000m ont les prix les plus bas.

Logiquement, il est clair que cette hypothèse est confirmé et cela se renforce avec le nuage de point obtenu à la section 4.g (2e graphe) qui nous une relation négative entre ces deux variables, donc une augmentation de la distance entraine une dévaluation du prix par unité de surface.

Conclusion et recommandations

L'évaluation immobilière est un problème de régression observé dans la majorité des pays du monde. Il est fort de constater que le district Sindian, New Taipei City de Taiwan en est la preuve concrète de ce problème. Malgré des commissions de régulations, des textes de lois qui tendent à protéger les investisseurs, la problématique de la variation des prix immobiliers restent toujours d'actualité, ce qui nous fait assister à des pics et des chutes de prix d'un secteur à un autre dépendamment d'une durée donnée créant ainsi une balance inégalitaire sur le marché.

Ce travail a consisté à déterminer les causes impactant la fluctuation des prix immobiliers dans un secteur, analysant le cas d'une circonscription bien précise. Le but principal était de dégager les facteurs pouvant jouer de manière directe et indirecte sur cette variation. Plus en détails il était question de déterminer quel impact l'âge d'une maison, le nombre de magasin à proximité d'une localité, la position géographique et la distance d'une station de métro la plus proche a sur une évaluation immobilière et si possible une estimation d'un bien immobiliers dans cette zone.

Afin de mener à bien nos objectifs, nous avons procédé par une décomposition du problème, d'un côté ressortir les résultats descriptifs univarié et d'autre part une les résultats analytiques bivarié effectués en utilisant les données historiques du marché de l'évaluation des biens immobiliers du district Sindian Dist., New Taipei City, Taiwan collecté entre 2012 et 2013.

Les résultats descriptifs univariés nous ont permis d'avoir une connaissance globale des données, notamment avoir des informations tels que la répartition de l'âge des maisons dans le district, le nombre moyen de magasin par secteur, le positionnement des maisons dans cette localité et ainsi que la habitation plus ou moins proche des stations de métro, ces analyses basiques à priori, nous ont permis de biens diriger nos travaux sur la partie de l'analyse bivarié afin de desceller les relations de ces variables autour de la variable d'intérêt.

Les résultats analytique quant à eux ont été d'une importance capitale, car nous ont permis de ressortir le lien qui existe ces différents données et la variable d'intérêt qui est le prix par unité de surface. Nous avons pu conclure que l'âge d'une maison n'avait pas un impact directe sur son prix, probablement parce que d'autres paramètres pouvaient s'y ajouter notamment la zone géographique ou nous pouvons être dans une zone urbaine ou dans une zone en périphérie. D'autres facteurs comme le nombre de magasins dans un secteur avaient un plus fort impact sur cette fluctuation, constatant ainsi qu'une forte augmentation de magasins dans une zone augmente aussi le prix moyen par unité de surface et afin un la proximité à une station de métro qui n'était pas à négliger, les zones étant à moins de 200m d'une station avait un prix plus élevé.

Parvenu au terme de notre étude il en ressort que dans le cas du district ou a porté notre étude, certains facteurs suscités n'impactent pas sur l'évaluation immobilière mais d'autres impactent de manière directe ou plus ou moins indirecte. Cela revient à confirmer l'hypothèse de départ selon laquelle l'évaluation immobilière serait liée à un problème de régression. Mais l'ensemble des variables manipulé au cours de notre étude comporte des exceptions pas moins négligeable, ce qui laisse penser qu'ils existent d'autres variables aussi pertinentes qui n'ont pas été prises

en compte, notamment des variables comme le type d'habitation (maison résidentielle, Immeuble, …), configuration d'une maison (nombres de chambres, nombres de douches), qualité du logement.

Recommandations

- **Investisseurs**

  - Prendre attache avec des experts dans le domaine (agents immobiliers, expert data scientiste) pour une évaluation du bien immobilier que l'on souhaiterait acquérir.
  - Une fois le bien acquis, s'il y a lieu de faire des rénovations, il est important d'en faire, cela augmente la valeur immobilière

- **Autorités**

  - Mettre un accent plus pointillé sur la régulation des biens immobiliers en vue de la stabilisation et une développement du secteur immobilier.

- **DataScientists**

  - Établir des modèles concrets prenant en paramètre le maximum de variable pour pouvoir ressortir une bonne évaluation

Référence

Site web et articles

VISUALISATION DES DONNÉES : Nuage de point
https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch9/scatter-nuages/5214827-fra.htm

ANALYSES DE DONNÉES ET DATAVIZ : Test de student
https://sites.google.com/view/aide-python/statistiques/test-de-student

VALUE REAL ESTATE INVESTMENT PROPERTY
https://www.investopedia.com/articles/mortgages-real-estate/11/valuing-real-estate.asp

COURS STATISTIQUES MASTER I – DATA SCIENCE
Auteur : Pr. NGUEFACK TSAGUE GEORGES

INTRODUCTION À PYTHON
https://courspython.com/introduction-python.html

INITIEZ-VOUS AU LANGAGE R POUR ANALYSER VOS DONNÉES
https://openclassrooms.com/fr/courses/4525256-initiez-vous-au-langage-r-pour-analyser-vos-donnees

DATA REAL ESTATE VALUATION USING LINEAR REGRESSION
https://www.kaggle.com/code/mahyamahjoob/real-estate-valuation-using-linear-regression/notebook

Appendices

Code Python

```
#importation de panda pour la manipulation des données
import pandas as pd

#importation de numpy et matplotlib pour une facile manipulation des
données
import numpy as np
import matplotlib.pyplot as plt

# importation de seaborn pour la graphique
import seaborn as sns


# téléchargement de la base de données excel

data = pd.read_excel(r"RealEstateValuation.xlsx", index_col = 0)


# Aperçu de notre jeu de donnée
print(data)
```

```
              X1    X2          X3  X4       X5          X6     Y  \
No
1    2012.916667  32.0    84.87882  10  24.98298  121.54024  37.9
2    2012.916667  19.5   306.59470   9  24.98034  121.53951  42.2
3    2013.583333  13.3   561.98450   5  24.98746  121.54391  47.3
4    2013.500000  13.3   561.98450   5  24.98746  121.54391  54.8
5    2012.833333   5.0   390.56840   5  24.97937  121.54245  43.1
..           ...   ...         ...  ..       ...        ...   ...
410  2013.000000  13.7  4082.01500   0  24.94155  121.50381  15.4
411  2012.666667   5.6    90.45606   9  24.97433  121.54310  50.0
412  2013.250000  18.8   390.96960   7  24.97923  121.53986  40.6
413  2013.000000   8.1   104.81010   5  24.96674  121.54067  52.5
414  2013.500000   6.5    90.45606   9  24.97433  121.54310  63.9

                       X7
No
1    (24,98298;121,54024)
2    (24,98034;121,53951)
3    (24,98746;121,54391)
4    (24,98746;121,54391)
5    (24,97937;121,54245)
..                    ...
410  (24,94155;121,50381)
411   (24,97433;121,5431)
```

```
412  (24,97923;121,53986)
413  (24,96674;121,54067)
414   (24,97433;121,5431)

[414 rows x 8 columns]
```

```
# informations sur les métas données du jeu de donnée


informations = data.info()
print(informations)
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 414 entries, 1 to 414
Data columns (total 8 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   X1      414 non-null    float64
 1   X2      414 non-null    float64
 2   X3      414 non-null    float64
 3   X4      414 non-null    int64
 4   X5      414 non-null    float64
 5   X6      414 non-null    float64
 6   Y       414 non-null    float64
 7   X7      414 non-null    object
dtypes: float64(6), int64(1), object(1)
memory usage: 29.1+ KB
None
```

```
# Résumé de certaines données statistiques telles que le centile, la
moyenne et la norme des valeurs numériques de la série
resume = data.describe()
print(resume)
```

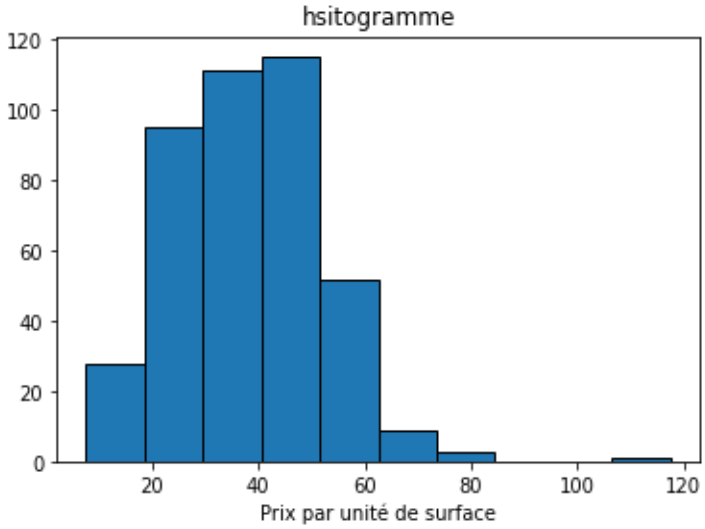
```
                X1          X2           X3          X4          X5   \
count   414.000000  414.000000   414.000000  414.000000  414.000000
mean   2013.148953   17.712560  1083.885689    4.094203   24.969030
std       0.281995   11.392485  1262.109595    2.945562    0.012410
min    2012.666667    0.000000    23.382840    0.000000   24.932070
25%    2012.916667    9.025000   289.324800    1.000000   24.963000
50%    2013.166667   16.100000   492.231300    4.000000   24.971100
75%    2013.416667   28.150000  1454.279000    6.000000   24.977455
max    2013.583333   43.800000  6488.021000   10.000000   25.014590

               X6          Y
count  414.000000  414.000000
mean   121.533361   37.980193
std      0.015347   13.606488
min    121.473530    7.600000
25%    121.528085   27.700000
50%    121.538630   38.450000
```

```
75%    121.543305   46.600000
max    121.566270  117.500000


# Statistiques groupées

# prix de l'unité de surface par l'age de la maison

priceUnitByAge = data.groupby('X2')['Y'].agg([np.mean, np.median, np.min,
np.max])
print(priceUnitByAge)

           mean  amin  amax
X2
0.0   54.135294  37.9  73.6
1.0   50.700000  50.7  50.7
1.1   49.780000  45.1  54.4
1.5   48.700000  47.7  49.7
1.7   50.400000  50.4  50.4
...         ...   ...   ...
40.9  54.350000  41.0  67.7
41.3  47.900000  35.1  60.7
41.4  63.300000  63.3  63.3
42.7  35.300000  35.3  35.3
43.8  42.700000  42.7  42.7

[236 rows x 3 columns]

# Statistiques groupées

# prix de l'unité de surface par nombre de magasin à proximité à pieds

priceUnitByCountShop = data.groupby('X4')['Y'].agg([np.mean, np.median,
np.min, np.max])
print(priceUnitByAge)

         mean  amin   amax
X4
0   26.462687  11.6   55.3
1   31.839130  11.2  117.5
2   31.412500  20.9   50.5
3   29.536957  17.7   61.5
4   37.474194  21.8   62.9
5   44.729851  22.8   60.7
6   46.951351   7.6   73.6
7   43.848387  25.0   62.1
8   44.696667  26.5   67.7
9   51.732000  32.4   78.3
10  48.430000  37.9   61.9

# Statistiques groupées
```

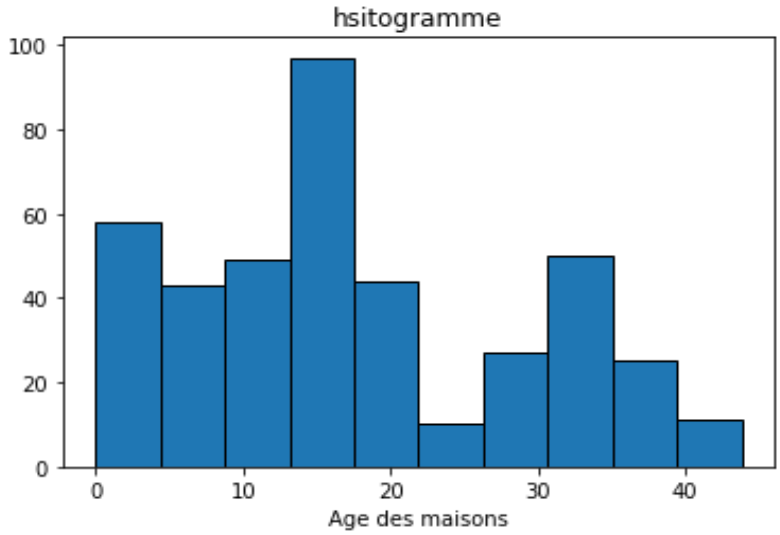
```
# prix de l'unité de surface par rapport à a distance à la station de
metro

priceUnitByDistanceMRT = data.groupby('X3')['Y'].agg([np.mean,np.median,
np.min, np.max])


print(priceUnitByAge)

                mean  amin  amax
X3
23.38284    48.700000  47.7  49.7
49.66105    57.300000  56.8  57.8
56.47425    56.666667  53.5  62.1
57.58945    42.700000  42.7  42.7
82.88643    46.600000  46.6  46.6
...               ...   ...   ...
4605.74900  13.400000  13.4  13.4
5512.03800  18.100000  17.4  18.8
6306.15300  15.000000  15.0  15.0
6396.28300  12.200000  12.2  12.2
6488.02100  11.200000  11.2  11.2

[259 rows x 3 columns]


# prix de l'unité de surface par rapport à la zone géographique (Lat,Lon)

priceByLocation = data.groupby('X7')['Y'].agg([np.mean,np.median, np.min,
np.max])


print(priceByLocation)

                           mean  median  amin  amax
X7
(24,93207;121,51597)  29.300000   29.30  29.3  29.3
(24,93293;121,51203)  45.100000   45.10  45.1  45.1
(24,93363;121,51158)  29.300000   29.30  29.3  29.3
(24,93885;121,50383)  16.933333   18.60  13.0  19.2
(24,94155;121,50381)  16.457143   15.90  12.8  20.0
...                         ...     ...   ...   ...
(24,99156;121,53406)  33.250000   33.25  32.2  34.3
(24,99176;121,53456)  35.400000   35.40  34.1  36.7
(24,998;121,5155)     41.200000   41.20  41.2  41.2
(25,00115;121,51776)  46.600000   46.60  46.6  46.6
(25,01459;121,51816)  27.300000   27.30  27.3  27.3

[259 rows x 4 columns]
```

```
#les dates de transactions des maisons selons leur ages

dateTransactionsByAgeofHouse = data.groupby('X1')['X2'].agg([np.mean,
np.median, np.min, np.max])


print(dateTransactionsByAgeofHouse)

                  mean  median  amin  amax
X1
2012.666667  18.623333   17.90   0.0  37.1
2012.750000  15.525926   14.10   0.0  38.0
2012.833333  13.203226   12.70   0.0  34.9
2012.916667  19.421053   17.25   2.0  40.9
2013.000000  17.814286   16.70   1.0  39.6
2013.083333  20.873913   16.70   0.0  42.7
2013.166667  18.960000   16.20   1.1  43.8
2013.250000  17.437500   16.35   0.0  40.9
2013.333333  16.555172   14.80   0.0  39.7
2013.416667  16.810345   16.40   0.0  40.1
2013.500000  16.465957   13.90   4.0  38.3
2013.583333  21.208696   18.10   2.6  35.9


# Quantiles

# 1- Quartile d'ordre 1 de l'age des maison


Q1 = np.quantile(data['X2'], 0.25)
print(Q1)

9.025

# Quantiles

# 2- Tous les quartiles du prix de l'unité de surface

QY = np.quantile(data['X2'], [0, .25, .5, .75, 1])


print(QY)

[ 0.     9.025 16.1   28.15  43.8  ]
```

```
#Distribution de la variable d'interet(Prix par unité de surface)
# histogramme

plt.hist(data['Y'], edgecolor='black')
plt.xlabel("Prix par unité de surface")
plt.title('hsitogramme')
plt.show()
```



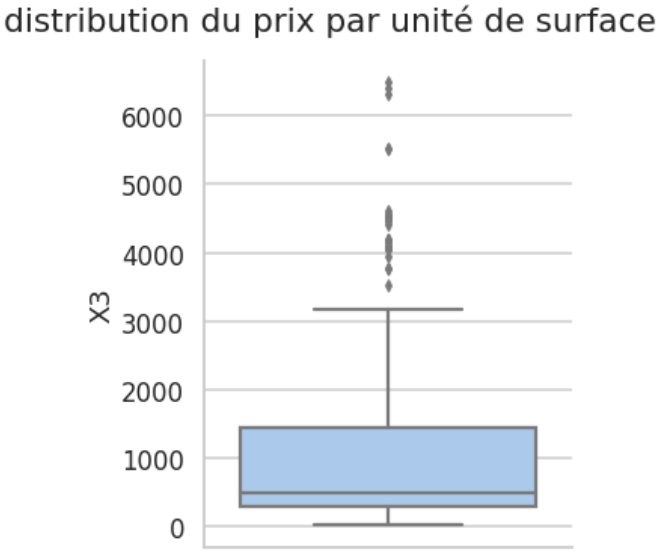

```
#Distribution de la variable des ages des maisons
# histogramme

plt.hist(data['X2'], edgecolor='black')
plt.xlabel("Age des maisons")
plt.title('hsitogramme')
plt.show()
```



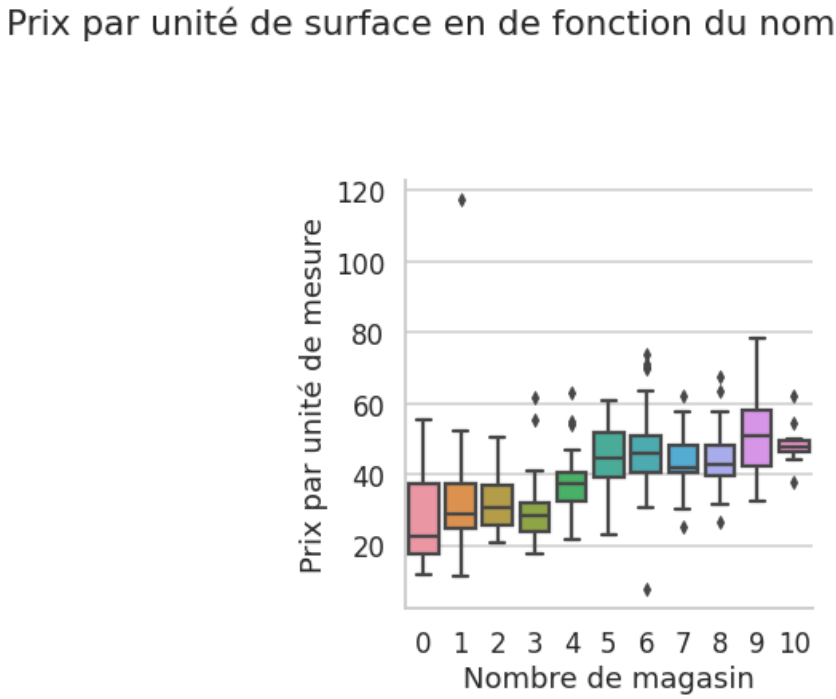
```
# Distribution de la variable du nombre de magasin à proximité des
magasins
# Statistique
resumeMagz = data['X4'].describe()
print(resumeMagz)
```

```
count    414.000000
mean       4.094203
std        2.945562
min        0.000000
25%        1.000000
50%        4.000000
75%        6.000000
max       10.000000
Name: X4, dtype: float64
```

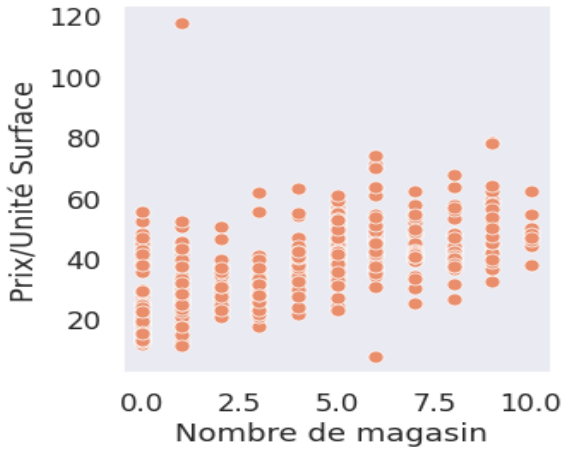
```
#Distributions de la distance de metro le plus proche

sns.set_style("whitegrid")
sns.set_context("talk")
sns.set_palette("pastel")


g1 = sns.catplot(y='X3', data = data, kind = 'box')
g1.fig.suptitle('Station de metro la plis proche', y =1.05)
plt.show()
```



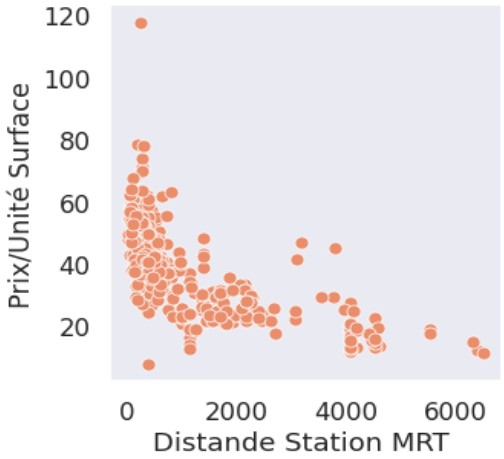
```
#distributions du prix par unité de surface par le nombre de magasin
sns.set_style("whitegrid")
sns.set_context("talk")
sns.set_palette("pastel")


g2 = sns.catplot(x ="X4", y = "Y", data = data, kind = "box" )
g2.fig.suptitle("Prix par unité de surface en de fonction du nombre de
magasin", y =1.25)
g2.set(xlabel = "Nombre de magasin", ylabel = "Prix par unité de mesure")
plt.show()
```



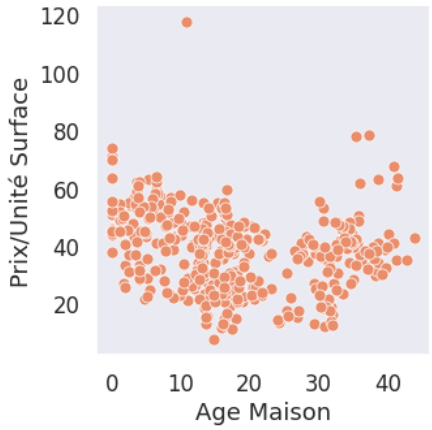

```
# Lien entre variables

sns.set_style('dark')
sns.set_context("talk")
sns.set_palette("flare")

#  nuage de points: Prix/Unité Surface et nombre de magasins

g3 = sns.relplot(x = "X4", y = "Y", data = data , kind = "scatter")
g3.set(xlabel ="Nombre de magasin " ,ylabel = "Prix/Unité Surface")
plt.show()
```



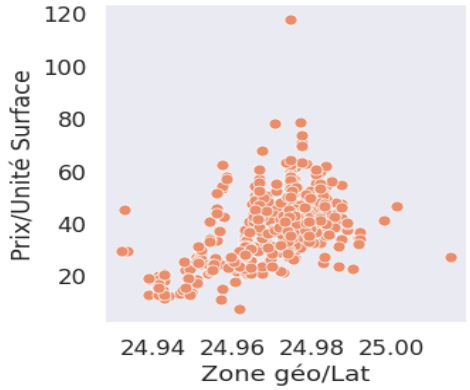

```
# Lien entre variables


sns.set_style('dark')
sns.set_context("talk")
sns.set_palette("flare")

#  nuage de points: Prix/Unité Surface et station de métro la plus métro

g4 = sns.relplot(x = "X3", y = "Y", data = data , kind = "scatter")
g4.set(xlabel ="Distande Station MRT " ,ylabel = "Prix/Unité Surface")
plt.show()
```

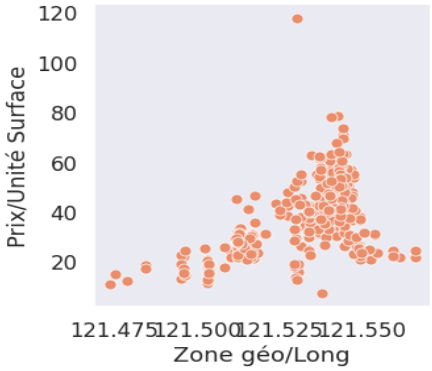


```
# Lien entre variables

sns.set_style('dark')
sns.set_context("talk")
sns.set_palette("flare")

#  nuage de points: Prix/Unité Surface et age maison

g5 = sns.relplot(x = "X2", y = "Y", data = data , kind = "scatter")
g5.set(xlabel ="Age Maison" ,ylabel = "Prix/Unité Surface")
plt.show()
```



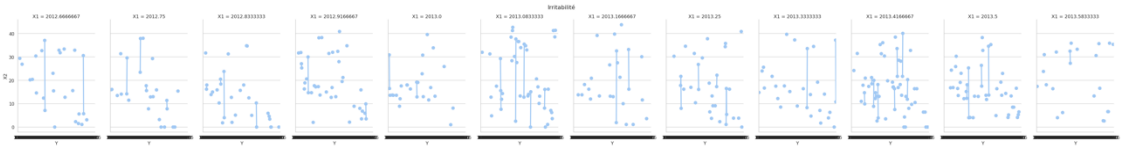

```
# Lien entre variables

sns.set_style('dark')
sns.set_context("talk")
sns.set_palette("flare")

#  nuage de points: Prix/Unité Surface et zone géographique

g6 = sns.relplot(x = "X5 ", y = "Y", data = data , kind = "scatter")
g6.set(xlabel ="Zone géo/Lat" ,ylabel = "Prix/Unité Surface")

g7 = sns.relplot(x = "X6", y = "Y", data = data , kind = "scatter")
g7.set(xlabel ="Zone géo/Long" ,ylabel = "Prix/Unité Surface")
plt.show()
```






```
# poids moyen a la naissance des enfants nés de mère irrités

ax = sns.catplot('Y', 'X2', data = data, kind = 'point', col = 'X1', ci =
None)
ax.fig.set_figheight(8.7)
ax.fig.suptitle("Irritabilité")
plt.show()
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36:
FutureWarning: Pass the following variables as keyword args: x, y. From
version 0.12, the only valid positional argument will be `data`, and
passing other arguments without an explicit keyword will result in an
error or misinterpretation.
  warnings.warn(
```



```
from scipy.stats import shapiro

p1 = shapiro(data['Y'])

print("pValue de X4 : "+ str(p1))
```

```
pValue de X4 : ShapiroResult(statistic=0.972750186920166,
pvalue=5.412278483163391e-07)
```

```
#Comparaison deux échantillons(age de la maison et prix de l'unité de
surface) pour voir s'ils sont significativement différents


import scipy.stats as stats

copyY = np.copy(data['X2'])
copyX4 = np.copy(data['X3'])

# y = stats.ttest_ind(copyY,copyX4)

# print("pValue est de  : "+ str(y[1]))
```

```
result = stats.kstest(copyY, copyX4)

print(result)
```

```
KstestResult(statistic=0.9951690821256038, pvalue=1.3800844769330233e-242)
```

```
from scipy.stats import anderson

x = np.copy(data['X2'])

anderson(x, dist='norm')
```

```
AndersonResult(statistic=7.327941717351109, critical_values=array([0.571,
0.65 , 0.78 , 0.909, 1.082]), significance_level=array([15. , 10. ,  5. ,
2.5,  1. ]))
```