

Ecole Nationale d'Ingénieurs de Carthage – TUNIS

المدرسة الوطنية للمهندسين بقرطاج



المدرسة الوطنية للمهندسين بقرطاج

Ecole Nationale d'Ingénieurs de Carthage

Analyse des Données

Rapport de mini projet Analyse de données

Spécialité : Génie Informatique

Analyse du Marché Immobilier de New York City

Réalisé par

Jihen BOUKHADHRA, Tasnim MAAMOURI, Yosra SAIDI,

Ines ZAIRI

Encadrant académique : Mme. Feriel Ben Nasr

Année Universitaire 2023 – 2024

Ecole Nationale d'Ingénieurs de Carthage-TUNIS

Tél/Fax : (+216) 71 940 699 / (+216) 71 940 775

Table des matières

Table des figures	ii
Introduction Générale	1
1 Identification des individus et des variables.	2
1.1 Description des données :	2
1.2 Exploration :	3
2 Analyse factorielle des Correspondances (AFC)	6
2.1 Le tableau de contingence	6
2.2 Les vecteurs de poids des lignes et colonnes	9
2.3 Les matrices des profils des lignes et colonnes	11
3 Analyse en Composantes Principales (ACP)	17
3.1 Etude préliminaire précédant l'ACP	17
3.2 Les représentations graphiques de l'étude ACP	21
4 Classification	26
4.1 Classification Ascendante Hiérarchique (CAH)	26
4.2 La méthode k-means	28
4.3 K-means après élimination des individus	35
4.4 Comparaisons entre CAH et K-means	37
Conclusion Générale	39

Table des figures

1.1	Affichage du dataset	3
1.2	Affichage de dimension : nombre de lignes et nombre de colonnes	4
1.3	Accès à une colonne	5
2.1	Tableau de contingence	6
2.2	Tableau de contingence en fréquences	7
2.3	Profils lignes	7
2.4	Profils colonnes	7
2.5	Tableau de fréquences théoriques attendues	8
2.6	Statistique du chi carré	9
2.7	Matrice de frequence	9
2.8	Affichage des vecteurs de poids des colonnes	10
2.9	Affichage des vecteurs de poids des lignes	10
2.10	Matrcie profils de colonnes	11
2.11	Matrcie profils de lignes	12
2.12	Affichage de chaque profil de ligne	13
2.13	Affichage de matrice de profil de ligne et de colonne moyen	15
3.1	Matrice de corrélation	17
3.2	Standardisation (réduction-centrage)	18
3.3	Data frame	18
3.4	Moyenne, écart-type et matrice de corrélation après standardisation	20
3.5	Affichage des valeurs propres en ordre décroissant	21
3.6	Pourcentage de variance expliquée par chacun des axes factoriels	21
3.7	Diagramme des valeurs propres	22
3.8	Diagramme des pourcentages cumulés	23
3.9	Projection des individus sur les deux premiers axes factoriels	24
3.10	La contribution des variables aux deux axes principaux	25
4.1	Le dendrogramme	27

4.2	La partition obtenue avec CAH	27
4.3	Corrélation entre les variables	29
4.4	Répartition des individus sur 4 clusters	30
4.5	Répartition des individus sur 4 clusters(suite)	31
4.6	Méthode du coude pour le choix du nombre de clusters	33
4.7	Les moyennes de caractéristiques de chaque cluster	34
4.8	Le clustering de données	35
4.9	Méthode du coude pour le choix du nombre de clusters après modification	36
4.10	Les moyennes de caractéristiques de chaque cluster après modification	36
4.11	Le clustering de données après modification	37

Introduction Générale

Notre projet s'inscrit dans le cadre d'une analyse de données visant à explorer en profondeur le marché immobilier de New York City. En tant que l'un des marchés les plus dynamiques et complexes au monde, le secteur immobilier de cette métropole emblématique est caractérisé par ses quartiers diversifiés, ses prix variés et ses tendances en constante évolution. Comprendre ce marché est essentiel pour les investisseurs, les agents immobiliers et les chercheurs souhaitant saisir les opportunités qu'il offre. Dans ce projet, nous nous plongeons dans l'analyse de données pour découvrir des insights précieux sur le marché immobilier de New York City.

Objectif de l'Analyse :

L'objectif principal de cette analyse est de comprendre les tendances, les schémas et les facteurs qui influencent le marché immobilier de New York City. Nous chercherons à répondre à des questions telles que :

- o Quels sont les types de propriétés les plus courants à New York City ?
- o Existe-t-il une corrélation entre le nombre de chambres et le nombre de salles de bain dans les propriétés ?
- o Y a-t-il des zones géographiques spécifiques qui présentent des caractéristiques immobilières particulières ?
- o Comment la taille de la propriété influe-t-elle sur son prix ?

Ce rapport est structuré en quatre chapitres, chacun correspondant à une phase distincte de notre analyse. Nous commençons par l'identification des individus et des variables présents dans notre dataset. Ensuite, nous explorons l'Analyse des Composantes Principales (ACP) pour identifier les principales sources de variation dans nos données. Nous poursuivons avec une analyse de la Classification Ascendante Hiérarchique (CAH) afin de regrouper les propriétés similaires et de comprendre les clusters présents sur le marché. Enfin, nous concluons avec une interprétation détaillée des résultats et la formulation de recommandations stratégiques basées sur nos conclusions.

Identification des individus et des variables.

Nous disposons d'un ensemble de données comprenant 500 individus représentant des propriétés à travers New York City. Chaque individu correspond à une maison est caractérisé par 8 variables.

1.1 Description des données :

Chaque maison est caractérisé par 6 variables quantitatives et 2 variables qualitatives.

- variables quantitatives :

Prix : Cette variable quantitative représente le prix de chaque maison.

Nombre de Chambres (BED) : Il s'agit d'une variable quantitative indiquant le nombre de chambres dans la propriété.

Nombre de Salles de bain (BATH) : Cette variable quantitative indique le nombre de salles de bain dans la propriété.

Surface de la Propriété (PROPETYSQFT) : Il s'agit d'une variable quantitative indiquant la superficie de la propriété en pieds carrés.

Latitude : Cette variable quantitative représente la coordonnée géographique de latitude de la propriété, fournissant ainsi des informations sur son emplacement.

Longitude : Il s'agit d'une variable quantitative représentant la coordonnée géographique de longitude de la propriété, offrant également des informations sur son emplacement.

- Variables qualitatives :

Type : Cette variable qualitative indique le type de propriété, offrant une perspective plus détaillée sur sa catégorie, tel que maison familiale, individuelle, appartement, etc.

Cette variable comporte 11 modalités, les suivantes : 'Coming Soon', 'Condo for Sale', 'Contingent', 'Co-Op for Sale', 'For Sale', 'Foreclosure', 'House for Sale', 'Land for Sale', 'Multi-family home for sale', 'Pending', 'Townhouse for Sale'.

Ville de la Propriété(SUBLOCALITY) : Cette variable qualitative indique le territoire exacte

de chaque maison à New York.

Cette variable comporte 11 modalités, les suivantes : 'Bronx County', 'Brooklyn', 'Kings County', 'Manhattan', 'New York County', 'New York', 'Queens', 'Queens County', 'Richmond County', 'Staten Island', 'The Bronx'.

1.2 Exploration :

La figure 1.1 présente les caractéristiques de notre dataset formé par 500 individus, 8 variables dont 6 quantitatives et 2 qualitatives.

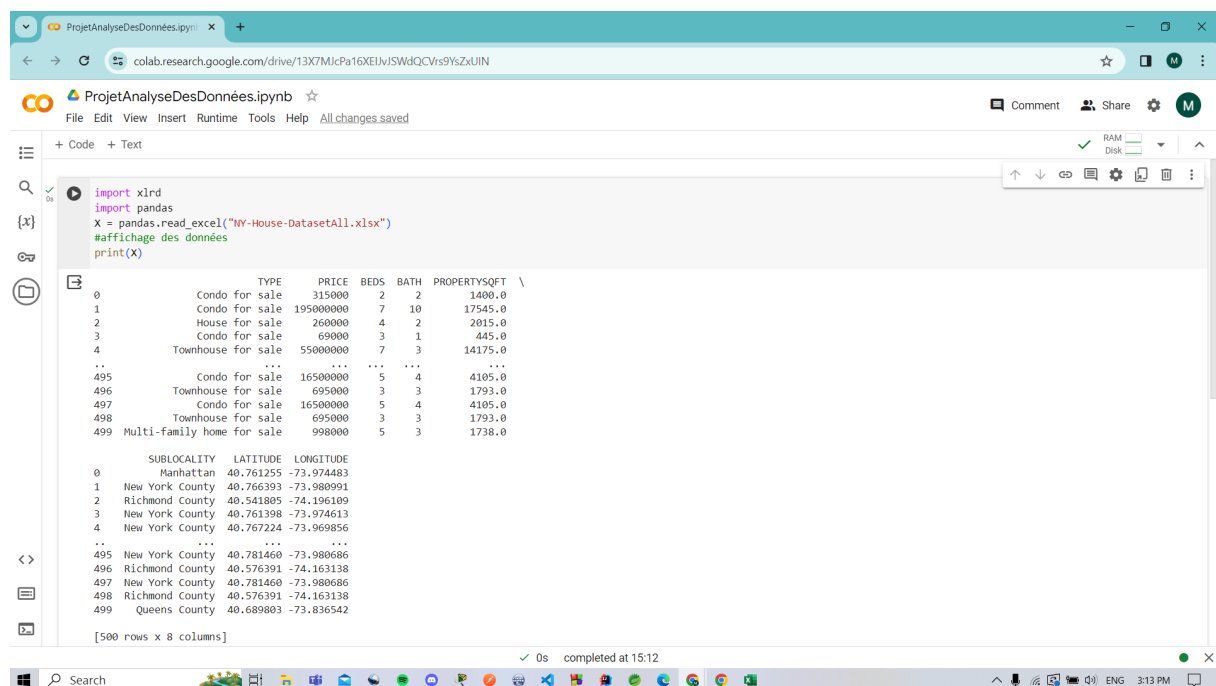


FIGURE 1.1: Affichage du dataset

La figure 1.2 présente la dimension du notre dataset (500,8) : 500 lignes(individus) et 8 colonnes(variables) et une description des données :

- moyenne.
- fréquence.
- écart-type.
- min et max de chaque variable.
- 1er, 2ème(médiane) et 3ème quartile.

```
(500, 8)
500
8
```

	TYPE	PRICE	BEDS	BATH	PROPERTYSQFT	SUBLOCALITY \
0	Condo for sale	315000	2	2	1400.0	Manhattan
1	Condo for sale	195000000	7	10	17545.0	New York County
2	House for sale	260000	4	2	2015.0	Richmond County
3	Condo for sale	69000	3	1	445.0	New York County
4	Townhouse for sale	55000000	7	3	14175.0	New York County

	LATITUDE	LONGITUDE
0	40.761255	-73.974483
1	40.766393	-73.980991
2	40.541805	-74.196109
3	40.761398	-73.974613
4	40.767224	-73.969856

	TYPE	PRICE	BEDS	BATH	PROPERTYSQFT \
count	500	5.000000e+02	500.00000	500.000000	500.000000
unique	11	NaN	NaN	NaN	NaN
top	House for sale	NaN	NaN	NaN	NaN
freq	130	NaN	NaN	NaN	NaN
mean	NaN	6.754716e+06	3.58400	2.512000	2255.276437
std	NaN	9.651201e+07	2.18552	1.635606	2410.902139
min	NaN	2.494000e+03	1.00000	1.000000	230.000000
25%	NaN	4.250000e+05	2.00000	1.750000	1234.500000
50%	NaN	7.999990e+05	3.00000	2.000000	2184.207862
75%	NaN	1.300000e+06	4.00000	3.000000	2184.207862
max	NaN	2.147484e+09	15.00000	16.000000	33000.000000

	SUBLOCALITY	LATITUDE	LONGITUDE
count	500	500.000000	500.000000
unique	12	NaN	NaN
top	Richmond County	NaN	NaN
freq	85	NaN	NaN
mean	NaN	40.701616	-73.952099
std	NaN	0.096501	0.118240
min	NaN	40.501623	-74.246109
25%	NaN	40.621172	-74.001605
50%	NaN	40.699734	-73.946550
75%	NaN	40.773236	-73.858960
max	NaN	40.907370	-73.712559

FIGURE 1.2: Affichage de dimension : nombre de lignes et nombre de colonnes

La figure ci-dessous présente un exemple d'accès à une colonne (PRICE).

```
#accès à une colonne ou plusieurs
print(X.PRICE)
print(X.PRICE.mean())
```

0	315000
1	195000000
2	260000
3	69000
4	55000000
	...
495	16500000
496	695000
497	16500000
498	695000
499	998000

Name: PRICE, Length: 500, dtype: int64
6754715.812

FIGURE 1.3: Accès à une colonne

Analyse factorielle des Correspondances

(AFC)

L'analyse factorielle des correspondances (AFC) est une méthode exploratoire d'analyse des tableaux de contingences, c'est-à-dire aux tableaux de comptages obtenus par le croisement de deux variables nominales.

2.1 Le tableau de contingence

La figure 2.1 présente le tableau de contingence, c'est une méthode de représentation de données issues d'un comptage permettant d'estimer la dépendance entre deux caractères. Elle consiste à croiser deux caractères de notre dataset "NY-Housing" (Type et Sublocality) en dénombrant l'effectif correspondant à la conjonction « Type » et « Sublocality ».

Les effectifs partiels sont rassemblés dans un tableau à double entrée, par ligne pour le premier caractère, et par colonne en fonction du second caractère.

TYPE	Bronx County	Brooklyn	East Bronx	Kings County	Manhattan	New York	New York Cou	Queens	Queens Count	Richmond Co	Staten Island	The Bronx
Co-op for sale	5	17	1	1	1	28	1	29	4	0	1	15
Coming Soon	0	0	0	2	0	0	0	0	0	0	0	0
Condo for sale	1	14	0	3	3	25	8	6	0	3	7	2
Contingent	0	0	0	0	0	0	0	0	0	8	0	0
For sale	0	1	0	3	0	12	3	0	0	0	0	0
Foreclosure	0	0	0	0	0	1	1	0	1	2	0	0
House for sale	15	0	0	25	0	4	5	0	38	43	0	0
Land for sale	7	1	0	12	0	1	0	0	6	5	0	0
Multi-family h	18	0	0	22	0	2	0	0	27	14	3	0
Pending	0	0	0	3	0	1	1	2	1	5	0	0
Townhouse fo	1	0	0	7	0	0	13	0	4	5	0	0

FIGURE 2.1: Tableau de contingence

La figure 2.2 présente le tableau de contingence en fréquences en divisant les valeurs de chaque valeur sur l'effectif total.

TYPE	Bronx County	Brooklyn	East Bronx	Kings County	Manhattan	New York	New York Cou	Queens	Queens Count	Richmond Co	Staten Island	The Bronx
Co-op for sale	0.01	0.034	0.002	0.002	0.002	0.056	0.002	0.058	0.008	0	0.002	0.03
Coming Soon	0	0	0	0.004	0	0	0	0	0	0	0	0
Condo for sale	0.002	0.028	0	0.006	0.006	0.05	0.016	0.012	0	0.006	0.014	0.004
Contingent	0	0	0	0	0	0	0	0	0	0.016	0	0
For sale	0	0.002	0	0.006	0	0.024	0.006	0	0	0	0	0
Foreclosure	0	0	0	0	0	0.002	0.002	0	0.002	0.004	0	0
House for sale	0.03	0	0	0.05	0	0.008	0.01	0	0.076	0.086	0	0
Land for sale	0.014	0.002	0	0.024	0	0.002	0	0	0.012	0.01	0	0
Multi-family ho	0.036	0	0	0.044	0	0.004	0	0	0.054	0.028	0.006	0
Pending	0	0	0	0.006	0	0.002	0.002	0.004	0.002	0.01	0	0
Townhouse fo	0.002	0	0	0.014	0	0	0.026	0	0.008	0.01	0	0

FIGURE 2.2: Tableau de contingence en fréquences

La figure 2.3 présente le tableau de profils lignes qui décrit les caractéristiques de la variable qualitative TYPE.

TYPE	Bronx County	Brooklyn	East Bronx	Kings County	Manhattan	New York	New York Cou	Queens	Queens Count	Richmond Co	Staten Island	The Bronx
Co-op for sale	0.04854369	0.16504854	0.00970874	0.00970874	0.00970874	0.27184466	0.00970874	0.2815534	0.03883495	0	0.00970874	0.14563107
Coming Soon	0	0	0	1	0	0	0	0	0	0	0	0
Condo for sale	0.01388889	0.19444444	0	0.04166667	0.04166667	0.34722222	0.11111111	0.08333333	0	0.04166667	0.09722222	0.02777778
Contingent	0	0	0	0	0	0	0	0	0	1	0	0
For sale	0	0.05263158	0	0.15789474	0	0.63157895	0.15789474	0	0	0	0	0
Foreclosure	0	0	0	0	0	0.2	0.2	0	0.2	0.4	0	0
House for sale	0.11538462	0	0	0.19230769	0	0.03076923	0.03846154	0	0.29230769	0.33076923	0	0
Land for sale	0.21875	0.03125	0	0.375	0	0.03125	0	0	0.1875	0.15625	0	0
Multi-family ho	0.20930233	0	0	0.25581395	0	0.02325581	0	0	0.31395349	0.1627907	0.03488372	0
Pending	0	0	0	0.23076923	0	0.07692308	0.07692308	0.15384615	0.07692308	0.38461538	0	0
Townhouse fo	0.03333333	0	0	0.23333333	0	0	0.43333333	0	0.13333333	0.16666667	0	0

FIGURE 2.3: Profils lignes

La figure 2.4 présente le tableau de profils colonnes qui décrit les caractéristiques de la variable qualitative SUBLOCALITY.

TYPE	Bronx County	Brooklyn	East Bronx	Kings County	Manhattan	New York	New York Cou	Queens	Queens Count	Richmond Co	Staten Island	The Bronx
Co-op for sale	0.10638298	0.51515152	1	0.01282051	0.25	0.37837838	0.03125	0.78378378	0.04938272	0	0.09090909	0.88235294
Coming Soon	0	0	0	0.02564103	0	0	0	0	0	0	0	0
Condo for sale	0.0212766	0.42424242	0	0.03846154	0.75	0.33783784	0.25	0.16216216	0	0.03529412	0.63636364	0.11764706
Contingent	0	0	0	0	0	0	0	0	0	0.09411765	0	0
For sale	0	0.03030303	0	0.03846154	0	0.16216216	0.09375	0	0	0	0	0
Foreclosure	0	0	0	0	0	0.01351351	0.03125	0	0.01234568	0.02352941	0	0
House for sale	0.31914894	0	0	0.32051282	0	0.05405405	0.15625	0	0.4691358	0.50588235	0	0
Land for sale	0.14893617	0.03030303	0	0.15384615	0	0.01351351	0	0	0.07407407	0.05882353	0	0
Multi-family ho	0.38297872	0	0	0.28205128	0	0.02702703	0	0	0.33333333	0.16470588	0.27272727	0
Pending	0	0	0	0.03846154	0	0.01351351	0.03125	0.05405405	0.01234568	0.05882353	0	0
Townhouse fo	0.0212766	0	0	0.08974359	0	0	0.40625	0	0.04938272	0.05882353	0	0

FIGURE 2.4: Profils colonnes

Test du chi-deux :

Le test du chi2 (prononcé "khi-deux") est une méthode statistique utilisée pour déterminer si l'association entre deux variables catégorielles est statistiquement significative. Il est couramment utilisé pour analyser les données de contingence, où les données sont organisées dans un tableau de contingence (également appelé tableau de croisement), qui montre la fréquence de chaque combinaison de catégories pour deux variables.

Suite à l'application du test chi2 sur notre dataset on obtient la valeur suivante : chi-deux = 620.5640189197744.

$$p\text{-value} = 4.2908371766749314e-76$$

Si la valeur de la statistique de test chi2 est grande et la p-value associée est petite (généralement inférieure à un seuil de signification, comme 0,05), on rejette l'hypothèse nulle selon laquelle les variables sont indépendantes, ce qui suggère qu'il existe une association significative entre les deux variables.

$$\text{Degré de liberté} = (11-1)*(11-1) = 100.$$

chi-deux » p-value : donc les deux variables sont dépendantes.

Le tableau de fréquences théoriques attendues représente les fréquences que l'on s'attendrait à observer dans chaque cellule du tableau de contingence.

```
Fréquences théoriques attendues :
[[9.6820e+00 6.7980e+00 1.6068e+01 8.2400e-01 1.5244e+01 6.5920e+00
 7.6220e+00 1.6686e+01 1.7510e+01 2.2660e+00 3.7080e+00]
 [1.8800e-01 1.3200e-01 3.1200e-01 1.6000e-02 2.9600e-01 1.2800e-01
 1.4800e-01 3.2400e-01 3.4000e-01 4.4000e-02 7.2000e-02]
 [6.7680e+00 4.7520e+00 1.1232e+01 5.7600e-01 1.0656e+01 4.6080e+00
 5.3280e+00 1.1664e+01 1.2240e+01 1.5840e+00 2.5920e+00]
 [7.5200e-01 5.2800e-01 1.2480e+00 6.4000e-02 1.1840e+00 5.1200e-01
 5.9200e-01 1.2960e+00 1.3600e+00 1.7600e-01 2.8800e-01]
 [1.7860e+00 1.2540e+00 2.9640e+00 1.5200e-01 2.8120e+00 1.2160e+00
 1.4060e+00 3.0780e+00 3.2300e+00 4.1800e-01 6.8400e-01]
 [4.7000e-01 3.3000e-01 7.8000e-01 4.0000e-02 7.4000e-01 3.2000e-01
 3.7000e-01 8.1000e-01 8.5000e-01 1.1000e-01 1.8000e-01]
 [1.2220e+01 8.5800e+00 2.0280e+01 1.0400e+00 1.9240e+01 8.3200e+00
 9.6200e+00 2.1060e+01 2.2100e+01 2.8600e+00 4.6800e+00]
 [3.0080e+00 2.1120e+00 4.9920e+00 2.5600e-01 4.7360e+00 2.0480e+00
 2.3680e+00 5.1840e+00 5.4400e+00 7.0400e-01 1.1520e+00]
 [8.0840e+00 5.6760e+00 1.3416e+01 6.8800e-01 1.2728e+01 5.5040e+00
 6.3640e+00 1.3932e+01 1.4620e+01 1.8920e+00 3.0960e+00]
 [1.2220e+00 8.5800e-01 2.0280e+00 1.0400e-01 1.9240e+00 8.3200e-01
 9.6200e-01 2.1060e+00 2.2100e+00 2.8600e-01 4.6800e-01]
 [2.8200e+00 1.9800e+00 4.6800e+00 2.4000e-01 4.4400e+00 1.9200e+00
 2.2200e+00 4.8600e+00 5.1000e+00 6.6000e-01 1.0800e+00]]
```

FIGURE 2.5: Tableau de fréquences théoriques attendues

Test du chi-carré : Le test du chi-carré est un moyen statistique de déterminer les différences entre ce qui était attendu et ce qui a été observé dans une ou plusieurs catégories.

La figure 2.6 présente les statistique du chi carré.

La statistique du khi-carré mesure la différence globale entre les effectifs de cellules observés et les effectifs attendus si les proportions de colonne étaient identiques d'une colonne à l'autre. Plus la valeur de la statistique du khi-carré est élevée, plus la différence entre les effectifs de cellules observés et théoriques est importante, et plus il apparaît que les proportions de colonne ne sont pas égales, que l'hypothèse d'indépendance est incorrecte et, par conséquent, que les variables TYPE et SUBLOCALITY sont liées.

TYPE	Bronx County	Brooklyn	East Bronx	Kings County	Manhattan	New York	New York Cou	Queens	Queens Count	Richmond Cot	Staten Island	The Bronx
Co-op for sale	0.10638298	0.51515152	1	0.01282051	0.25	0.37837838	0.03125	0.78378378	0.04938272	0	0.09090909	0.88235294
Coming Soon	0	0	0	0.02564103	0	0	0	0	0	0	0	0
Condo for sale	0.0212766	0.42424242	0	0.03846154	0.75	0.33783784	0.25	0.16216216	0	0.03529412	0.63636364	0.11764706
Contingent	0	0	0	0	0	0	0	0	0	0.09411765	0	0
For sale	0	0.03030303	0	0.03846154	0	0.16216216	0.09375	0	0	0	0	0
Foreclosure	0	0	0	0	0	0.01351351	0.03125	0	0.01234568	0.02352941	0	0
House for sale	0.31914894	0	0	0.32051282	0	0.05405405	0.15625	0	0.4691358	0.50588235	0	0
Land for sale	0.14893617	0.03030303	0	0.15384615	0	0.01351351	0	0	0.07407407	0.05882353	0	0
Multi-family h	0.38297872	0	0	0.28205128	0	0.02702703	0	0	0.33333333	0.16470588	0.27272727	0
Pending	0	0	0	0.03846154	0	0.01351351	0.03125	0.05405405	0.01234568	0.05882353	0	0
Townhouse fo	0.0212766	0	0	0.08974359	0	0	0.40625	0	0.04938272	0.05882353	0	0

FIGURE 2.6: Statistique du chi carré

La matrice de fréquence est un outil utilisé dans notre Dataset pour représenter la fréquence d'occurrence de chaque modalité d'une variable dans un ensemble de données(dans notre cas on a choisir la variable qualitative TYPE). Cette matrice peut être utilisée pour visualiser la distribution des données et identifier les tendances ou les modèles.

TYPE	Bronx County	Brooklyn	East Bronx	Kings County	Manhattan	New York	New York Cou	Queens	Queens Count	Richmond Cot	Staten Island	The Bronx
Co-op for sale	0.10638298	0.51515152	1	0.01282051	0.25	0.37837838	0.03125	0.78378378	0.04938272	0	0.09090909	0.88235294
Coming Soon	0	0	0	0.02564103	0	0	0	0	0	0	0	0
Condo for sale	0.0212766	0.42424242	0	0.03846154	0.75	0.33783784	0.25	0.16216216	0	0.03529412	0.63636364	0.11764706
Contingent	0	0	0	0	0	0	0	0	0	0.09411765	0	0
For sale	0	0.03030303	0	0.03846154	0	0.16216216	0.09375	0	0	0	0	0
Foreclosure	0	0	0	0	0	0.01351351	0.03125	0	0.01234568	0.02352941	0	0
House for sale	0.31914894	0	0	0.32051282	0	0.05405405	0.15625	0	0.4691358	0.50588235	0	0
Land for sale	0.14893617	0.03030303	0	0.15384615	0	0.01351351	0	0	0.07407407	0.05882353	0	0
Multi-family h	0.38297872	0	0	0.28205128	0	0.02702703	0	0	0.33333333	0.16470588	0.27272727	0
Pending	0	0	0	0.03846154	0	0.01351351	0.03125	0.05405405	0.01234568	0.05882353	0	0
Townhouse fo	0.0212766	0	0	0.08974359	0	0	0.40625	0	0.04938272	0.05882353	0	0

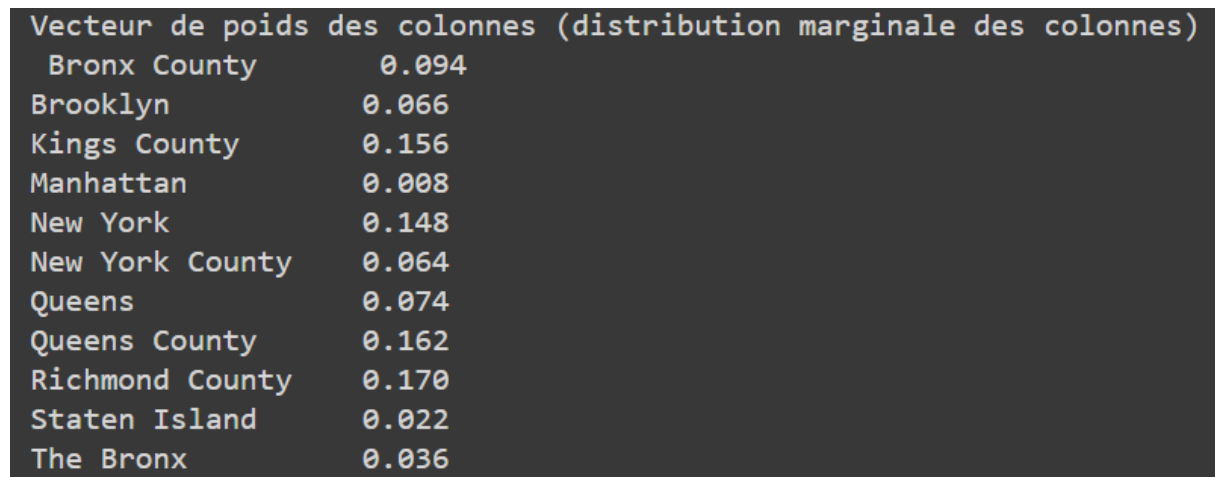
FIGURE 2.7: Matrice de fréquence

2.2 Les vecteurs de poids des lignes et colonnes

Les vecteurs de poids des colonnes représentent la distribution marginale des colonnes dans la matrice de fréquence. Chaque élément du vecteur correspond à la somme des fréquences des modalités pour chaque colonne de la matrice de fréquence. En d'autres termes, ces vecteurs donnent une indication de la fréquence totale de chaque modalité dans l'ensemble des observations pour une variable(Dans notre cas la variable SUBLOCALITY) .

À partir de la figure 2.8 des vecteurs de poids des colonnes on peut conclure que les modalités les plus fréquentes pour la variable SUBLOCALITY sont : Richmond County ,Queens County ,Kings County et New York.

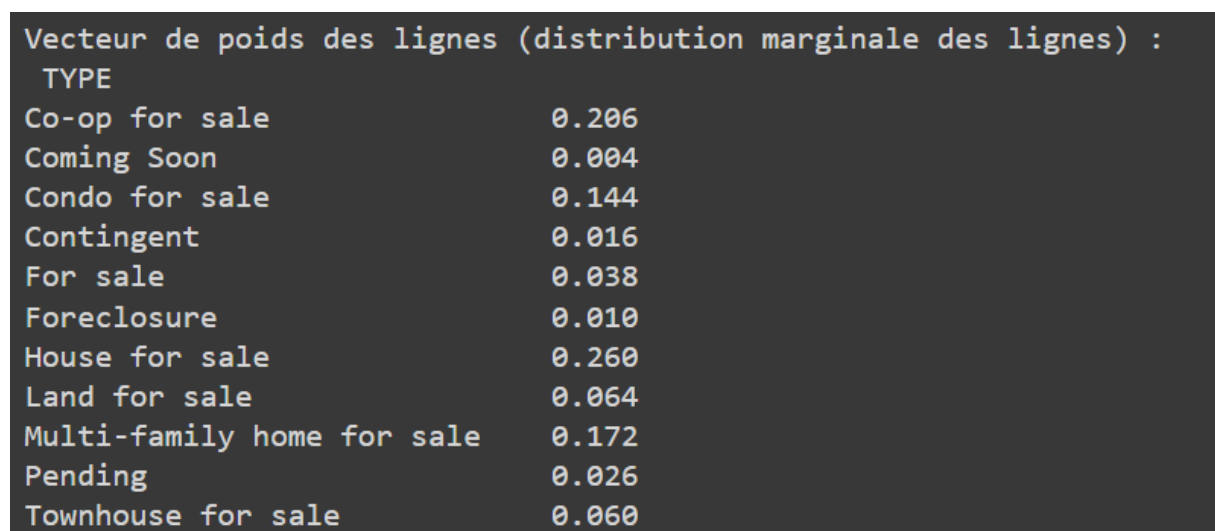
les modalités les plus fréquentes pour la variable SUBLOCALITY sont :Manhattan,Staten Island et The Bronx.



Vecteur de poids des colonnes (distribution marginale des colonnes)	
Bronx County	0.094
Brooklyn	0.066
Kings County	0.156
Manhattan	0.008
New York	0.148
New York County	0.064
Queens	0.074
Queens County	0.162
Richmond County	0.170
Staten Island	0.022
The Bronx	0.036

FIGURE 2.8: Affichage des vecteurs de poids des colonnes

Les vecteurs de poids des lignes sont des vecteurs numériques utilisés dans l'analyse de notre dataset pour donner un poids différent à chaque ligne ou observation dans un ensemble de données. Ils permettent de prendre en compte la contribution de chaque observation dans l'analyse des données .



Vecteur de poids des lignes (distribution marginale des lignes) :	
TYPE	
Co-op for sale	0.206
Coming Soon	0.004
Condo for sale	0.144
Contingent	0.016
For sale	0.038
Foreclosure	0.010
House for sale	0.260
Land for sale	0.064
Multi-family home for sale	0.172
Pending	0.026
Townhouse for sale	0.060

FIGURE 2.9: Affichage des vecteurs de poids des lignes

2.3 Les matrices des profils des lignes et colonnes

La figure ci-dessous présente la matrice de profil colonne.

La matrice des profils colonnes C est obtenue en divisant chaque colonne i de matrice des fréquences F par son poids fi.

En examinant cette matrice de profil colonne on peut remarquer que : - 51.51% des individus ayant un type de maison "Co-op for sale" se trouvent dans la SUBLOCALITY "Brooklyn". Cela peut indiquer une forte demande de ce type de maison dans cette région. - une proportion relativement faible de 1.23% pour les individus ayant comme type "Pending" et sublocality "Queens County", cela peut indiquer une demande très faible aux maisons de type "Pending" dans cette région.

TYPE	Bronx County	Brooklyn	East Bronx	Kings County	Manhattan	New York	New York Cou	Queens	Queens Count	Richmond Co	Staten Island	The Bronx
Co-op for sale	0.10638298	0.51515152	1	0.01282051	0.25	0.37837838	0.03125	0.78378378	0.04938272	0	0.09090909	0.88235294
Coming Soon	0	0	0	0.02564103	0	0	0	0	0	0	0	0
Condo for sale	0.0212766	0.42424242	0	0.03846154	0.75	0.33783784	0.25	0.16216216	0	0.03529412	0.63636364	0.11764706
Contingent	0	0	0	0	0	0	0	0	0	0.09411765	0	0
For sale	0	0.03030303	0	0.03846154	0	0.16216216	0.09375	0	0	0	0	0
Foreclosure	0	0	0	0	0	0.01351351	0.03125	0	0.01234568	0.02352941	0	0
House for sale	0.31914894	0	0	0.32051282	0	0.05405405	0.15625	0	0.4691358	0.50588235	0	0
Land for sale	0.14893617	0.03030303	0	0.15384615	0	0.01351351	0	0	0.07407407	0.05882353	0	0
Multi-family ho	0.38297872	0	0	0.28205128	0	0.02702703	0	0	0.33333333	0.16470588	0.27272727	0
Pending	0	0	0	0.03846154	0	0.01351351	0.03125	0.05405405	0.01234568	0.05882353	0	0
Townhouse fo	0.0212766	0	0	0.08974359	0	0	0.40625	0	0.04938272	0.05882353	0	0

FIGURE 2.10: Matrcie profils de colonnes

La figure ci-dessous présente la matrice de profil ligne.

La matrice des profils lignes L est obtenue en divisant chaque ligne i de matrice des fréquences F par son poids fi.

Elle est utilisé pour représenter la distribution des modalités des variables qualitatives(TYPE et SUBLOCALITY) pour chaque individu dans notre dataset.

En examinant cette matrice de profil ligne on peut remarquer que :

- 43.33% des individus ayant un type de maison "Townhouse for sale" se trouvent dans la région "New York County". Donc on peut dire que La région de New York County semble être un endroit populaire pour les townhouses en vente parmi les autres région. -seulement 0.97% des individus ayant un type de maison "Co-op for sale" se trouvent dans les régions "Sublocality" (East Bronx, Kings County, Manhattan, New York County et Staten Island). Donc on peut dire que le type "Co-op for sale" est relativement rare dans ces différentes régions. Cela pourrait

indiquer une disponibilité limitée de ce type dans ces zones .

TYPE	Bronx County	Brooklyn	East Bronx	Kings County	Manhattan	New York	New York Cou	Queens	Queens Count	Richmond Co	Staten Island	The Bronx
Co-op for sale	0.04854369	0.16504854	0.00970874	0.00970874	0.00970874	0.27184466	0.00970874	0.2815534	0.03883495	0	0.00970874	0.14563107
Coming Soon	0	0	0	1	0	0	0	0	0	0	0	0
Condo for sale	0.01388889	0.19444444	0	0.04166667	0.04166667	0.34722222	0.11111111	0.08333333	0	0.04166667	0.09722222	0.02777778
Contingent	0	0	0	0	0	0	0	0	0	1	0	0
For sale	0	0.05263158	0	0.15789474	0	0.63157895	0.15789474	0	0	0	0	0
Foreclosure	0	0	0	0	0	0.2	0.2	0	0.2	0.4	0	0
House for sale	0.11538462	0	0	0.19230769	0	0.03076923	0.03846154	0	0.29230769	0.33076923	0	0
Land for sale	0.21875	0.03125	0	0.375	0	0.03125	0	0	0.1875	0.15625	0	0
Multi-family hc	0.20930233	0	0	0.25581395	0	0.02325581	0	0	0.31395349	0.1627907	0.03488372	0
Pending	0	0	0	0.23076923	0	0.07692308	0.07692308	0.15384615	0.07692308	0.38461538	0	0
Townhouse fo	0.03333333	0	0	0.23333333	0	0	0.43333333	0	0.13333333	0.16666667	0	0

FIGURE 2.11: Matrcie profils de lignes

La figure 2.12 présente l’affichage de chaque profil de ligne. Dans le domaine de l’analyse de données, le terme "profil de ligne" peut faire référence à une visualisation utilisée pour représenter les données le long d’une ligne, c’est-à-dire, pour représenter chaque modalité d’une variable qualitative (dans notre cas c’est TYPE) en fonction de toutes les autres modalités de l’autre variable qualitative (dans notre cas c’est SYBLOCALITY). Cela peut être pertinent dans le contexte d’essayer de comprendre les dépendances entre les modalités des deux variables qualitatives. A titre d’exemple, selon notre cas, on peut citer ceci :

- Presque 100% des accommodations à vendre sont situées à Richmond County. Ceci peut-être expliqué par le fait que Richmond County est une ile éloignée du centre de NY où les habitants veulent passer seulement leurs vacances et revenir à leurs villes de départ.
- Pour la modalité ‘Multi-Family Home for Sale’, le plus grand pourcentage (35%) est celui avec la modalité ‘ Kings County’. Ceci peut-être expliqué par le fait que le comté de Kings est le comté le plus peuplé de l’Etat de New York, et le second comté des États-Unis au niveau de la densité de population, après Manhattan.
- Pour la modalité ‘ Pending’, le pourcentage le plus élevé (30%) est celui avec la modalité ‘ Queens County’. Ceci peut-être expliqué par le fait que le Queens est un arrondissement très étendu et une grand partie n’est pas spécialement intéressante pour les touristes. Il y a cependant quelques endroits qui valent le détour lorsque vous avez déjà vu tout ce que Manhattan a à offrir.

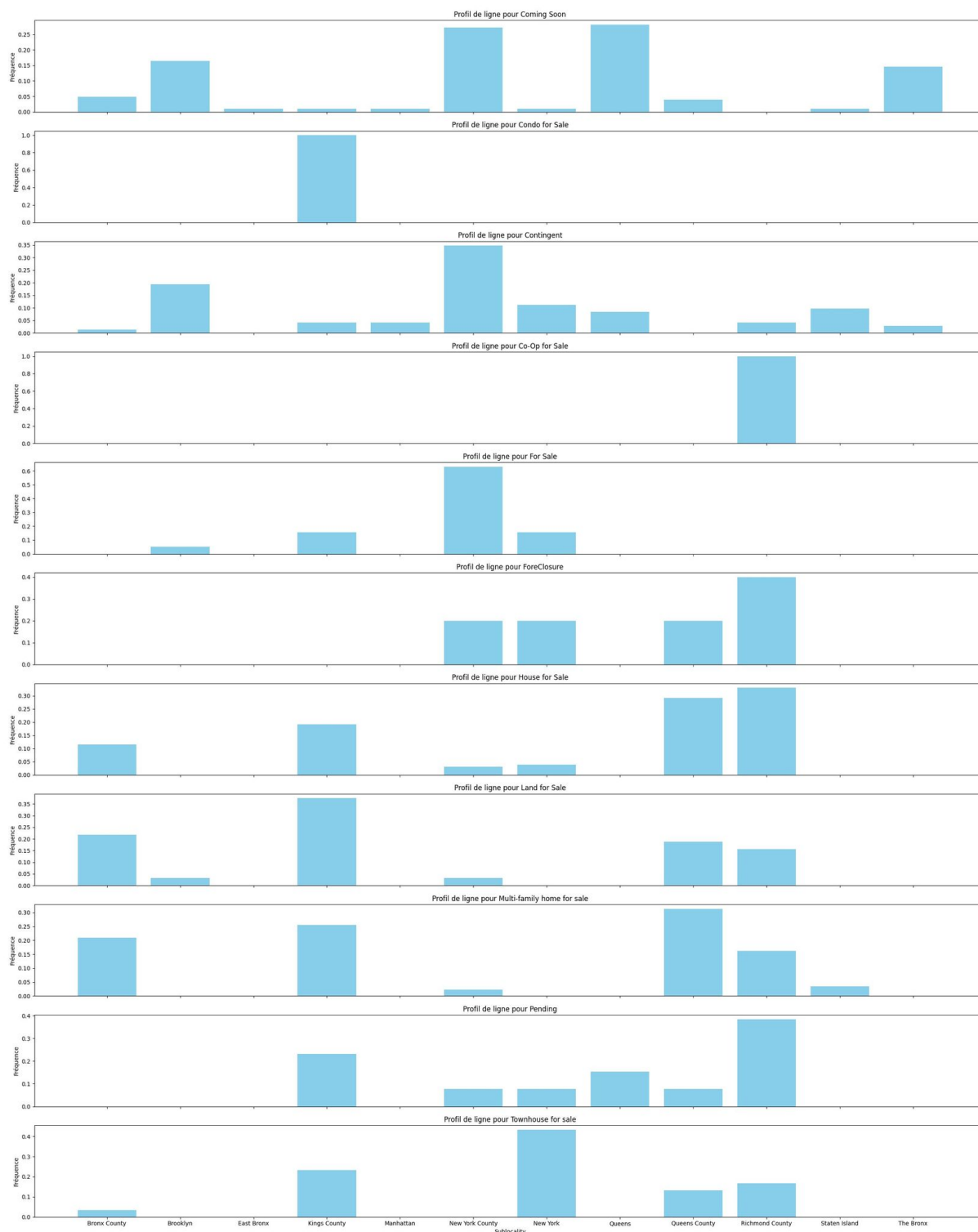


FIGURE 2.12: Affichage de chaque profil de ligne

Le **profil de ligne moyen** est utilisé dans notre dataset pour représenter la tendance centrale des variables qualitatives dans une matrice de profil ligne.

La figure ci-dessous présente le profil de ligne moyen de la variable "SUBLOCALITY", en examinant cette figure on peut remarquer que :

- Environ 17% des individus ont la modalité "Richmond County" pour la variable "Sublocality" qui indique une représentation modérée de cette région dans l'ensemble de données.
- Seulement 0.2% des observations ont la modalité "East Bronx" pour la variable "Sublocality". Ceci indique une faible représentation de cette région dans notre dataset, ce qui peut avoir des implications pour l'analyse et l'interprétation des résultats.

Le profil de colonne moyen est utile pour évaluer la similarité ou la différence entre les modalités de variable qualitative "TYPE" de notre dataset.

D'après la figure de profil colonne on peut remarquer que :

- Environ 17% des individus ont la modalité "multi-family home for sale" pour la variable "TYPE", ce qui indique que "Multi-family home for sale" est une modalité relativement courante parmi les individus dans l'ensemble de données. Cela signifie qu'un nombre important d'individus ont ce type.
- 0.4% pour la modalité "Coming Soon" de la variable "TYPE" indique que cette variable est très rare dans notre dataset.

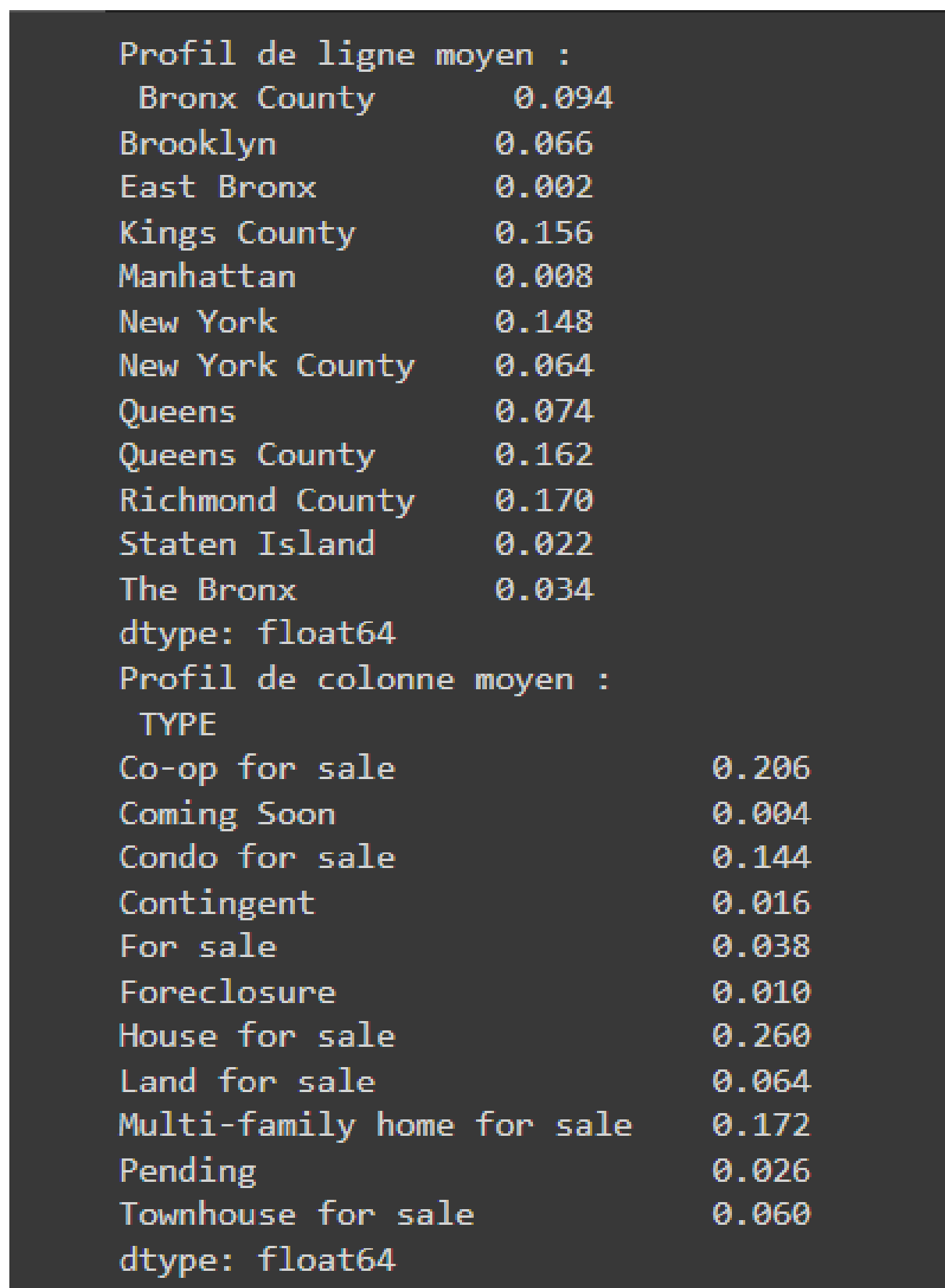


FIGURE 2.13: Affichage de matrice de profil de ligne et de colonne moyen

Distances de chi2 entre deux modalités :

Permet de quantifier la dissimilarité entre deux modalités :comme le montre la figure , la distance de χ^2 entre les deux modalités "Co-op for sale" et "Coming Soon" de la variable qualitative "TYPE" est égal à 8.588477 et celle entre les deux modalités "Coming Soon" et "Land For Sale" est de l'ordre de 3.39508.Cela indique que "Co-op for sale" et "Coming Soon" sont plus différentes en termes de distribution des fréquences que "Coming Soon" et "Land For Sale". En d'autres termes, il y a une plus grande dissimilarité entre les premières modalités qu'entre les secondes.

En conclusion, l'Analyse Factorielle des Correspondances (AFC) s'est avérée être un outil puissant pour explorer et comprendre la structure de nos données catégorielles. En réduisant la dimensionnalité et en mettant en évidence les relations entre les catégories, l'AFC nous a permis de visualiser clairement les associations entre les variables(TYPE et SUBLOCALITY). En identifiant les associations les plus fortes, nous avons pu mettre en lumière les aspects les plus saillants de nos données. Globalement, l'AFC a été un outil essentiel pour analyser et interpréter nos données catégorielles de manière efficace.

Analyse en Composantes Principales (ACP)

L'objectif est d'obtenir une représentation des individus avec le minimum de déformation c.a.d en conservant au mieux la distance entre les individus

3.1 Etude préliminaire précédant l'ACP

Matrice de corrélation : Une matrice de corrélation est une table qui montre les corrélations entre plusieurs variables dans un ensemble de données. Une corrélation mesure la relation statistique entre deux variables. Elle peut être positive (les variables augmentent ou diminuent ensemble) ou négative (les variables évoluent dans des directions opposées).

La matrice de corrélation montre les corrélations entre chaque paire de variables, avec une valeur de corrélation allant de -1 à 1. Une valeur de 1 indique une corrélation positive parfaite, -1 une corrélation négative parfaite et 0 indique aucune corrélation linéaire.

D'après la figure :

- les variables fortement coorelées positivement sont : "Beds" et "Price","Bath" et "Beds" , "Bath" et "PropertySqft". - Les variables "Longitude" et "Beds","Latitude" et "Beds" sont faiblement corrélées.

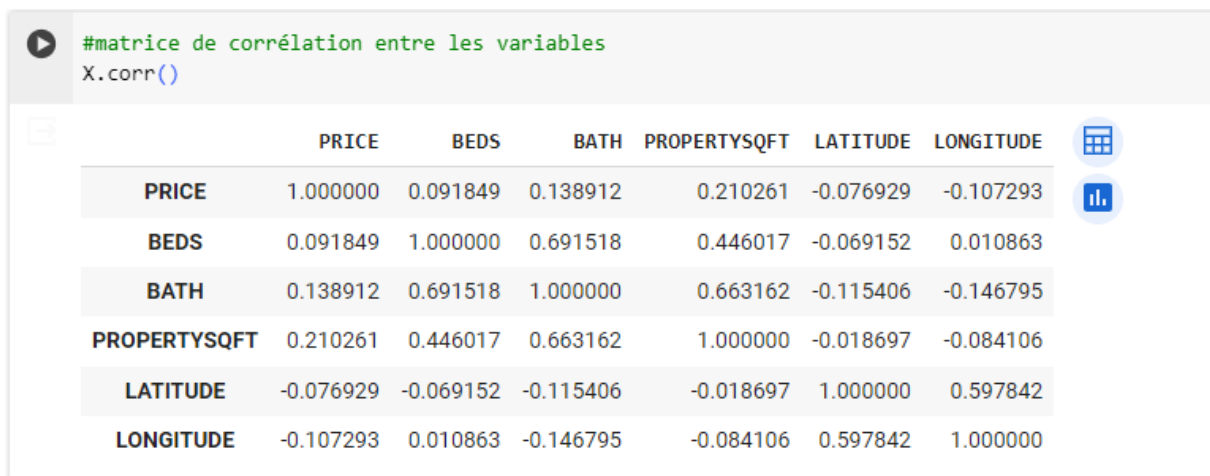


FIGURE 3.1: Matrice de corrélation

La figure ci-dessous présente standardisation des données (réduction-centrage).

Centrage : Le centrage consiste à déplacer la distribution des valeurs d'une variable afin que sa moyenne soit égale à zéro. Cela signifie soustraire la moyenne de chaque valeur de la variable.

Réduction : La réduction met à l'échelle la distribution des valeurs d'une variable pour avoir une dispersion homogène. Cela signifie diviser chaque valeur de la variable par son écart-type.

La standardisation permet de rendre les variables comparables en termes de distribution et de dispersion, facilitant ainsi l'interprétation des données et la comparaison entre différentes variables.

```
#scikit-learn
import sklearn
from sklearn.preprocessing import StandardScaler

sc = StandardScaler()
#transformation - centrage-réduction
Z = sc.fit_transform(X)
print(Z)
```

```
[[-0.06679132 -0.72549622 -0.31334736 -0.35510898  0.61863641 -0.18950647]
 [ 1.95243895  1.56458023  4.58270516  6.3482611  0.67193779 -0.244598 ]
 [-0.06736177  0.19053436 -0.31334736 -0.09976227 -1.65770547 -2.06575251]
 ...
 [ 0.10107596  0.64854965  0.91066577  0.76800134  0.82822074 -0.24202015]
 [-0.06285005 -0.26748093  0.2986592  -0.19193621 -1.29894885 -1.78663 ]
 [-0.0597074  0.64854965  0.2986592  -0.21477209 -0.12253545  0.97828494]]
```

FIGURE 3.2: Standardisation (réduction-centrage)

La figure 3.3 présente l'affichage du data frame.

```
print(Z[0][0])
NYData=pandas.DataFrame(Z,index=X.index,columns=X.columns)
print(NYData)
print(NYData.iloc[0,0])
```

```
-0.06679132499977304
      PRICE  BEDS  BATH  PROPERTYSQFT  LATITUDE  LONGITUDE
0 -0.066791 -0.725496 -0.313347 -0.355109  0.618636 -0.189506
1  1.952439  1.564580  4.582705  6.348261  0.671938 -0.244598
2 -0.067362  0.190534 -0.313347 -0.099762 -1.657705 -2.065753
3 -0.069343 -0.267481 -0.925354 -0.751623  0.620119 -0.190602
4  0.500390  1.564580  0.298659  4.949044  0.680547 -0.150332
..      ...      ...      ...      ...      ...      ...
495  0.101076  0.648550  0.910666  0.768001  0.828221 -0.242020
496 -0.062850 -0.267481  0.298659 -0.191936 -1.298949 -1.786630
497  0.101076  0.648550  0.910666  0.768001  0.828221 -0.242020
498 -0.062850 -0.267481  0.298659 -0.191936 -1.298949 -1.786630
499 -0.059707  0.648550  0.298659 -0.214772 -0.122535  0.978285
```

```
[500 rows x 6 columns]
-0.06679132499977304
```

FIGURE 3.3: Data frame

- La moyenne des variables avant la standardisation fournit une indication de la tendance centrale des données brutes, tandis que la moyenne après standardisation indique comment les données sont centrées autour de zéro après avoir éliminé les différences d'échelle et de dispersion.
- L'écart-type des variables avant la standardisation mesure la dispersion des valeurs autour de la moyenne et fournit des informations sur la variabilité des données brutes. Une fois les données standardisées, l'écart-type des variables standardisées sera égal à 1, car cela permet de mettre les variables à la même échelle.
- La matrice de corrélation après standardisation est une représentation tabulaire des corrélations entre chaque paire de variables une fois que les données ont été standardisées. Chaque élément de la matrice correspond au coefficient de corrélation de Pearson entre deux variables. Cette corrélation mesure la force et la direction de la relation linéaire entre les variables, tout en éliminant l'effet de l'échelle et de la distribution des données originales grâce à la standardisation.

D'après la figure 3.4 :

- Les variables fortement corrélées positivement sont : "Beds" et "Price", "Bath" et "Beds", "Bath" et "PropertySqft".
- Les variables "Longitude" et "Beds", "Latitude" et "Beds" sont faiblement corrélées.

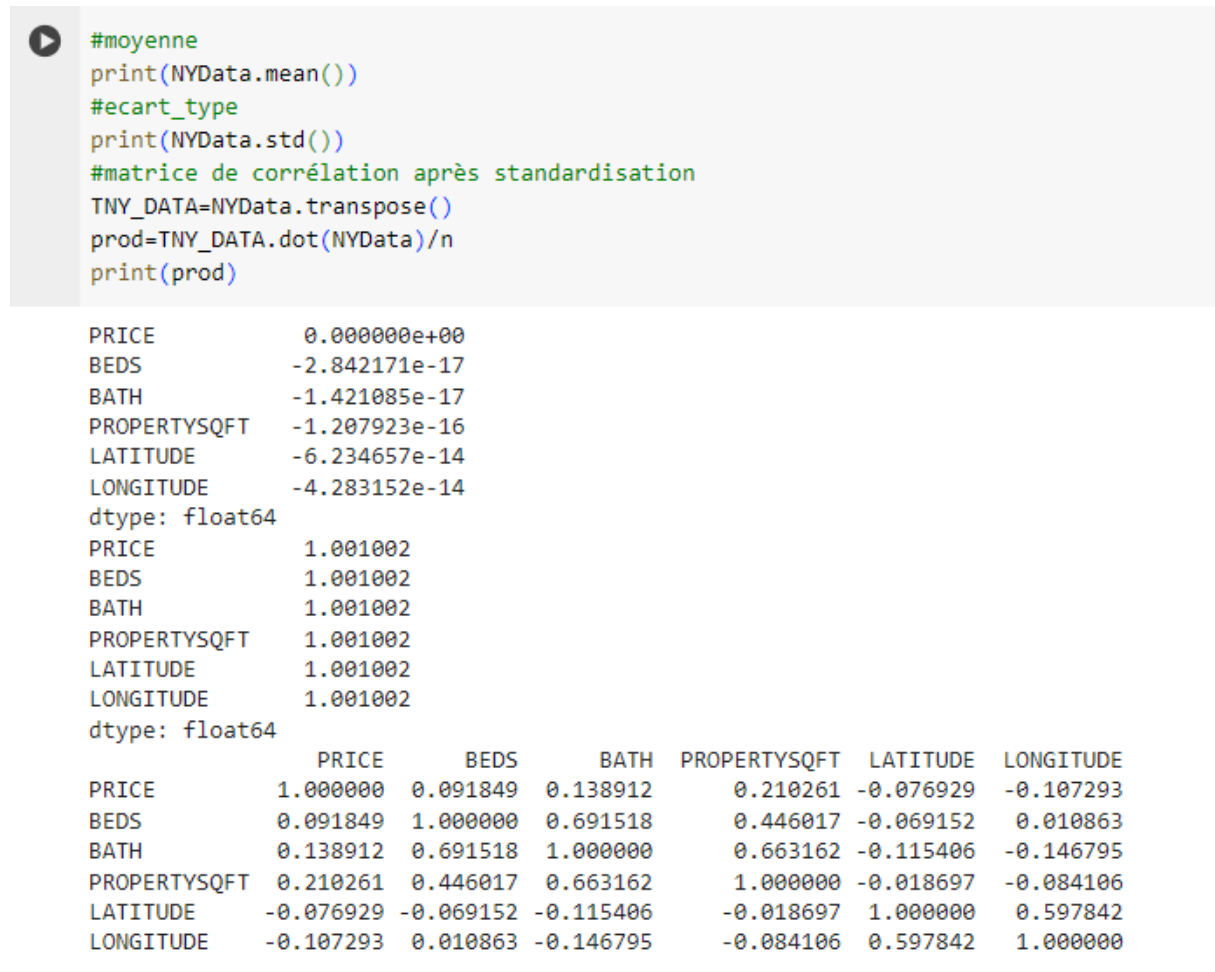


FIGURE 3.4: Moyenne, écart-type et matrice de corrélation après standardisation

La figure 3.5 présente l’affichage des valeurs propres en ordre décroissant.

Les valeurs propres caractérisent les transformations linéaires dans les espaces vectoriels. On a deux valeurs propres qui sont supérieur à 1 : $\lambda_0 = 2.31$ et $\lambda_1 = 1.56$.


```
eigenvalues, eigenvectors = np.linalg.eig(Coor)

diagonal_matrix = np.diag(eigenvalues)

#affichage des valeurs propres en ordre décroissant
print(diagonal_matrix)
```

```
[[2.31472577 0.          0.          0.          0.          0.          ]
 [0.          1.56158028 0.          0.          0.          0.          ]
 [0.          0.          0.9557293  0.          0.          0.          ]
 [0.          0.          0.          0.5641011  0.          0.          ]
 [0.          0.          0.          0.          0.22663931 0.          ]
 [0.          0.          0.          0.          0.          0.37722424]]
```

FIGURE 3.5: Affichage des valeurs propres en ordre décroissant

La figure 3.6 présente la proportion de variance expliquée par chaque composante principale (= axe factoriel). Cette proportion est souvent exprimée en pourcentage.

```
# Calcul du pourcentage de variance expliquée par chacun des axes factoriels
tot = sum(eigenvalues)
var_exp = [(i / tot)*100 for i in sorted(eigenvalues, reverse=True)]
cum_var_exp = np.cumsum(var_exp)
print(pandas.DataFrame({'valprop':eigenvalues,'inertie':var_exp,'inertiecum':cum_var_exp}))
```

	valprop	inertie	inertiecum
0	2.314726	38.578763	38.578763
1	1.561580	26.026338	64.605101
2	0.955729	15.928822	80.533922
3	0.564101	9.401685	89.935608
4	0.226639	6.287071	96.222678
5	0.377224	3.777322	100.000000

FIGURE 3.6: Pourcentage de variance expliquée par chacun des axes factoriels

3.2 Les représentations graphiques de l'étude ACP

La figure 3.7 présente le diagramme des valeurs propres qui est un outil utile pour comprendre la structure de variance dans nos données et pour guider la sélection du nombre approprié de composantes à inclure dans notre analyse en composantes principales. Une valeur propre élevée indique que l'axe explique une grande partie de la variance dans les données. On a deux valeurs propres qui sont supérieures à 1 : 0=2.31 et 1=1.56. On va choisir deux axes factoriels dans la suite de notre analyse !

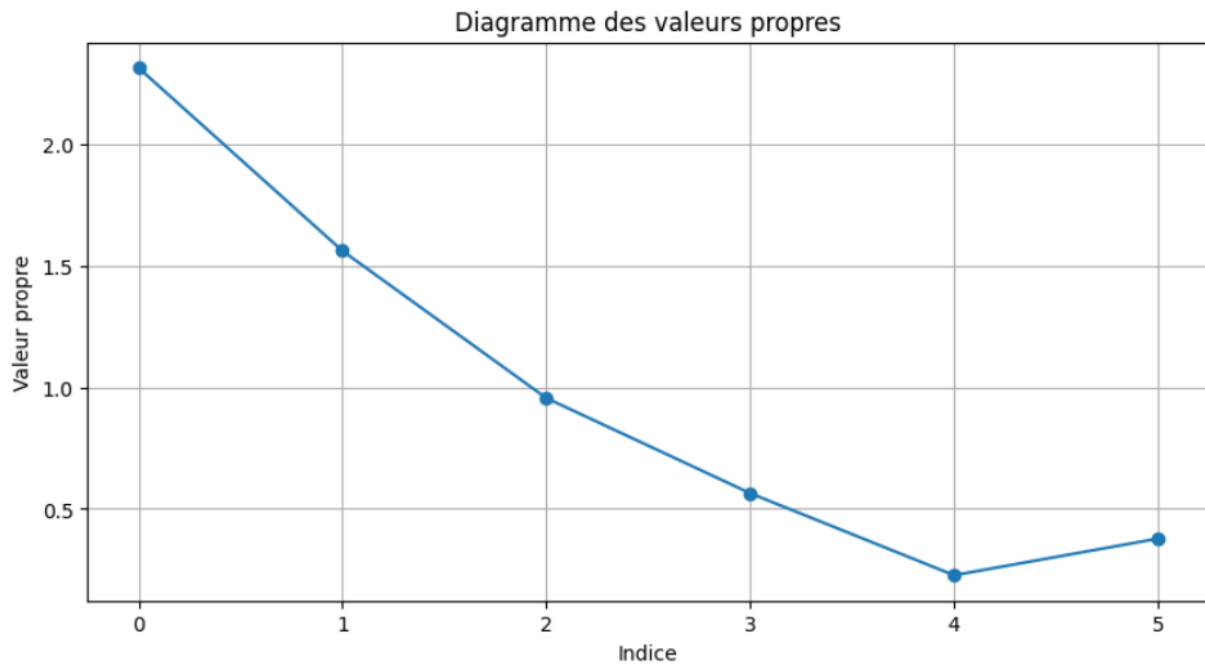


FIGURE 3.7: Diagramme des valeurs propres

La figure 3.8 présente le diagramme des pourcentages cumulés est un outil important pour évaluer la performance de notre analyse en composantes principales et pour prendre des décisions éclairées sur le nombre optimal de composantes à inclure dans notre modèle.

Dans notre cas, le diagramme indique que les pourcentages les plus élevés ($>90\%$) sont pour les deux premières valeurs propres. Donc, ce sont ces valeurs propres qui vont générer les composantes principales qui gardent le plus des informations de départ : perte très faible !

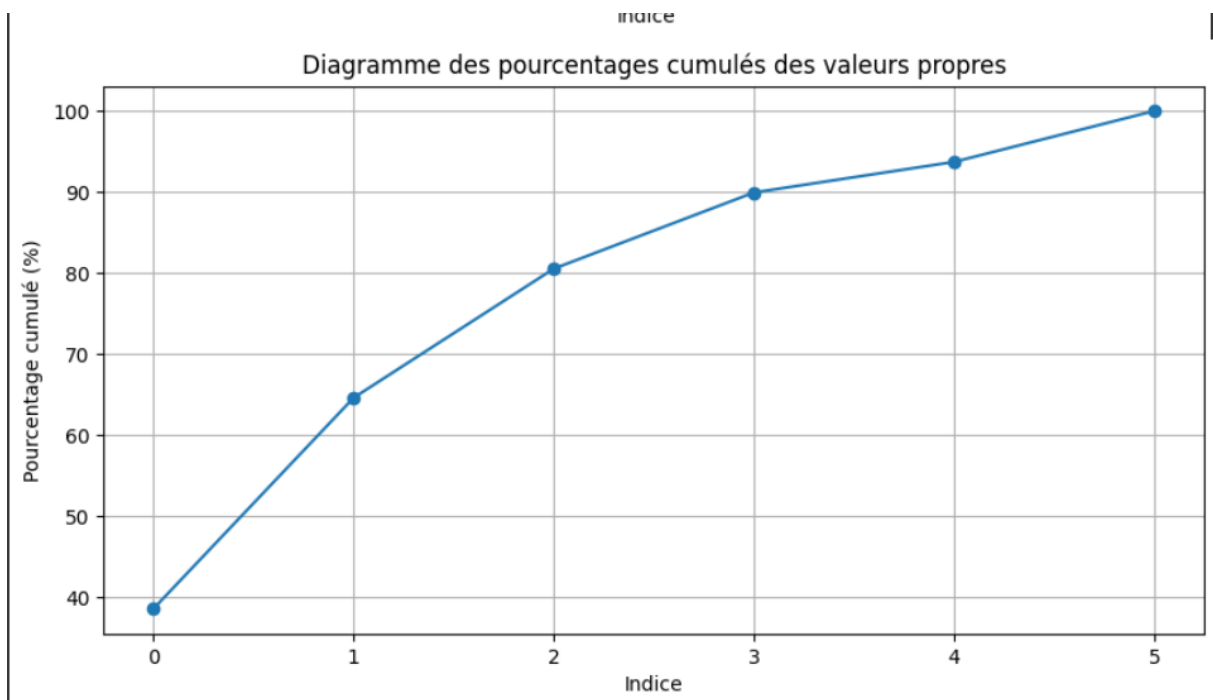


FIGURE 3.8: Diagramme des pourcentages cumulés

La position des individus dans la figure 3.9 de projection peut également nous aider à interpréter les composantes principales. Par exemple, si les individus avec des valeurs élevées sur une variable spécifique sont regroupés dans une certaine direction dans la figure, cela peut indiquer comment cette variable contribue à la structure des composantes principales. Dans notre cas, le maximum de valeurs atteint sur le premier axe est : 12.5.

Et celui pour le deuxième axe est : 3.8.

La projection des individus sur les deux premiers axes factoriels est une technique essentielle pour explorer et comprendre la structure des données dans l'analyse en composantes principales. Elle permet une visualisation efficace des relations entre les individus et des patterns présents dans les données, ce qui facilite l'interprétation et l'analyse des résultats.

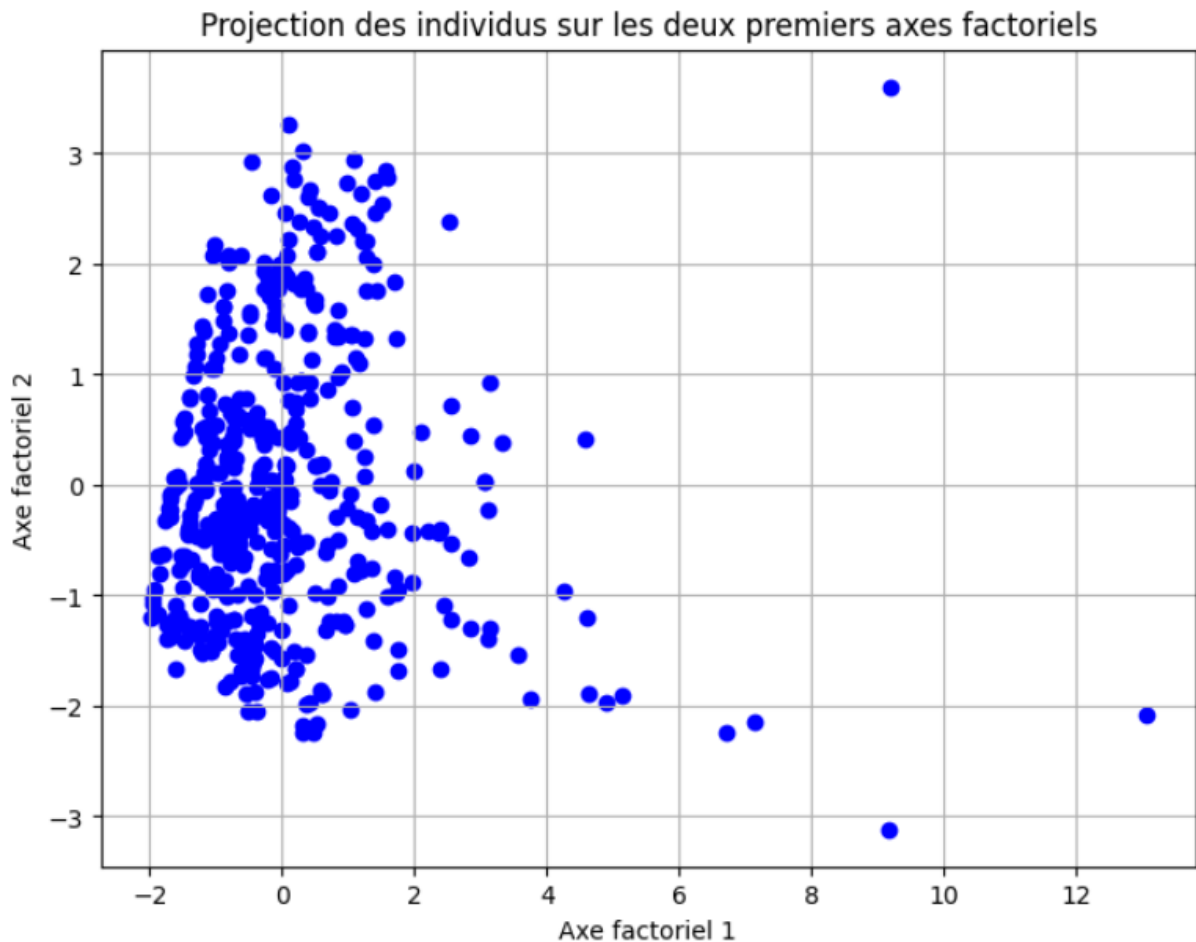


FIGURE 3.9: Projection des individus sur les deux premiers axes factoriels

La projection des variables sur les deux premiers axes factoriels dans la figure 3.10 permet de visualiser comment les variables originales contribuent à ces deux axes principaux, qui représentent la plus grande part de la variance des données.

Selon notre cas, les 4 variables : PRICE, PROPERTYSQFT, BATH et BEDS ont fortement contribué à l'axe 1. Les autres 2 variables LONGITUDE et LATITUDE ont fortement contribué à l'axe 2. Ceci explique de plus le fait que les variables PRICE, PROPERTYSQFT, BATH, BEDS sont fortement proportionnelles. Donc les individus (houses dans notre cas) ayant un prix élevé, sont ceux qui ont la surface la plus grande, le nombre de chambres et de salles d'eau le plus élevé aussi.

De même, les 2 autres variables restantes sont fortement liées géographiquement et mathématiquement tout en exprimant l'une avec l'autre.

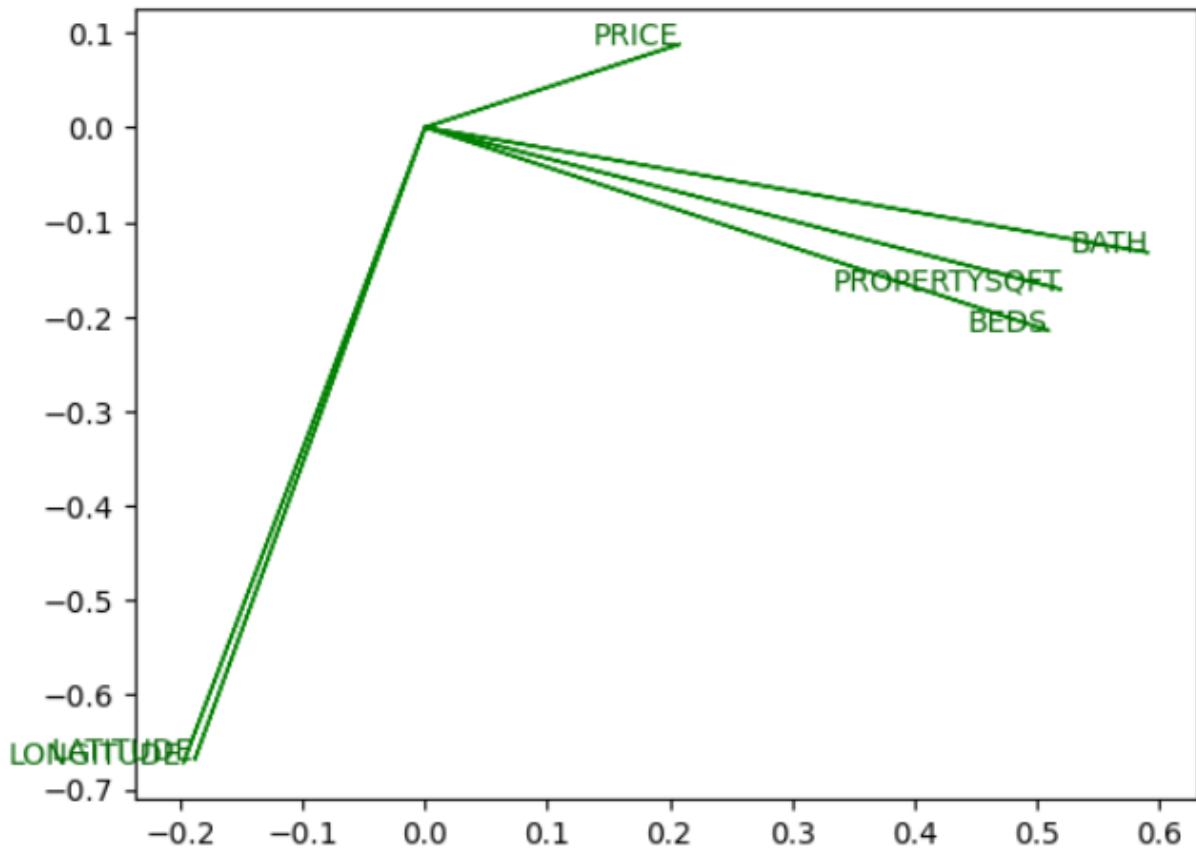


FIGURE 3.10: La contribution des variables aux deux axes principaux

En conclusion, l'Analyse en Composantes Principales (ACP) a été un outil crucial pour notre projet. Elle nous a permis de réduire la dimensionnalité des données, de visualiser des tendances et des regroupements, d'identifier les variables importantes, et de mieux comprendre la structure de nos données. En simplifiant notre analyse, l'ACP a renforcé nos conclusions et enrichi notre compréhension du sujet étudié.

Classification

4.1 Classification Ascendante Hiérarchique (CAH)

La Classification Ascendante Hiérarchique (CAH) est une méthode d'analyse statistique utilisée pour regrouper des individus ou des variables en fonction de similarités ou de dissimilarités mesurées entre eux. La CAH est une technique non supervisée, ce qui signifie qu'elle ne nécessite pas de variables cibles prédéfinies.

Les buts de la CAH sont clairement définis :

1. Regroupement des individus similaires : La méthode CAH permet de regrouper les individus qui partagent des caractéristiques similaires ou des profils similaires dans des clusters ou des groupes. Cela peut être utile pour segmenter une population en sous-groupes homogènes à des fins d'analyse ultérieure ou de prise de décision.

2. Identification de structures sous-jacentes : En regroupant les individus de manière hiérarchique, la méthode CAH peut révéler des structures sous-jacentes dans les données, telles que des relations de similarité ou de dissimilarité entre les individus. Cela peut aider à mieux comprendre la nature des données et à identifier des patterns ou des tendances importantes.

3. Visualisation des regroupements : La méthode CAH peut également être utilisée pour créer des dendrogrammes, qui sont des représentations graphiques des regroupements hiérarchiques. Les dendrogrammes permettent de visualiser la structure des regroupements et facilitent l'interprétation des résultats de l'analyse.

4. Segmentation de marché : En marketing, la méthode CAH est souvent utilisée pour segmenter les clients en groupes ayant des comportements d'achat similaires. Cela permet aux entreprises de mieux comprendre les besoins et les préférences des différents segments de leur marché cible et d'adapter leurs stratégies de marketing en conséquence.

5. Analyse de données biologiques ou génétiques : Dans les domaines de la biologie et de la génétique, la méthode CAH est utilisée pour regrouper des échantillons ou des gènes similaires en fonction de leurs profils d'expression ou de leurs séquences génétiques. Cela peut aider à identifier des sous-populations de patients ou à comprendre les relations phylogénétiques entre

différentes espèces.

Le dendrogramme est un outil puissant pour visualiser et interpréter les résultats de classifications hiérarchiques, ce qui permet de mieux comprendre la structure et les patterns présents dans les données. Il est largement utilisé dans de nombreux domaines pour l'analyse exploratoire des données, la segmentation de populations et la prise de décision.

la figure 4.1 présente le dendrogramme qui est une représentation arborescente des regroupements hiérarchiques des individus. Chaque nœud du dendrogramme représente un regroupement à un niveau spécifique de la hiérarchie, et les branches connectant les nœuds indiquent la similarité entre ces regroupements. Plus les branches sont longues, moins les regroupements sont similaires.

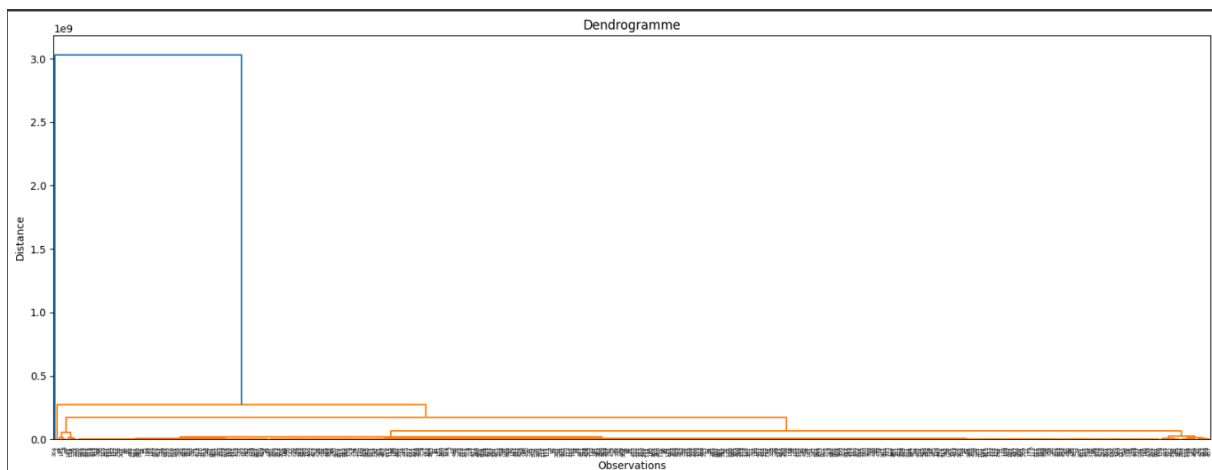


FIGURE 4.1: Le dendrogramme

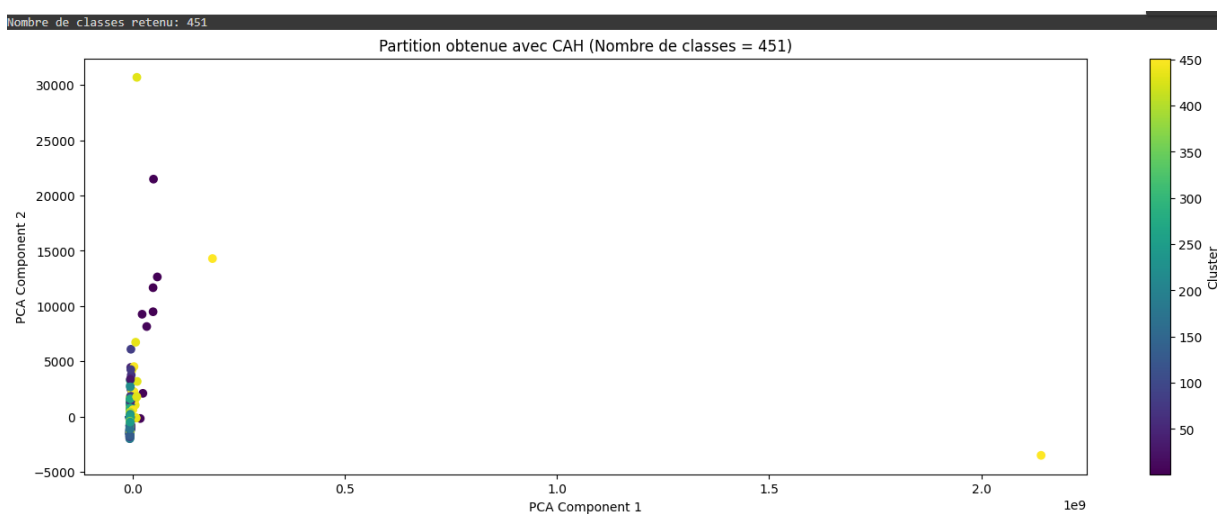


FIGURE 4.2: La partition obtenue avec CAH

la figure 4.2 présente la partition obtenue avec CAH : On a obtenu avec la méthode de CAH, 451 clusters.

4.2 La méthode k-means

Définition : La méthode de k-means est un algorithme de clustering largement utilisé dans le domaine de l'apprentissage automatique et de l'analyse de données pour partitionner un ensemble de données en un certain nombre de clusters. L'objectif de l'algorithme est de partitionner les données en clusters de telle sorte que les points au sein de chaque cluster soient similaires entre eux et différents des points des autres clusters.

Voici comment fonctionne l'algorithme de k-means :

1. ****Initialisation**** : Tout d'abord, le nombre de clusters k est spécifié. Les centres initiaux des clusters, appelés "centroïdes", sont généralement choisis de manière aléatoire à partir des données.
2. ****Affectation des points aux clusters**** : Chaque point de données est attribué au cluster dont le centroïde est le plus proche. La distance utilisée pour mesurer la proximité est généralement la distance euclidienne, bien que d'autres mesures puissent également être utilisées.
3. ****Mise à jour des centroïdes**** : Une fois que tous les points ont été attribués à des clusters, les centroïdes des clusters sont recalculés en prenant la moyenne de tous les points attribués à chaque cluster. Cela déplace le centroïde vers la "moyenne" des points du cluster.
4. ****Répéter les étapes 2 et 3**** : Les étapes d'affectation des points aux clusters et de mise à jour des centroïdes sont répétées jusqu'à ce qu'un critère d'arrêt soit atteint. Ce critère peut être un nombre fixe d'itérations, la convergence des centroïdes (c'est-à-dire lorsque les centroïdes ne changent plus significativement entre les itérations), ou d'autres critères spécifiques.
5. ****Finalisation**** : Une fois que l'algorithme a convergé, les points de données sont regroupés dans les clusters finaux, et les centroïdes représentent les centres de chaque cluster.

La méthode de k-means est utilisée dans de nombreux domaines, y compris la segmentation de marché, la reconnaissance de formes, l'analyse de données géospatiales et la compression d'images, entre autres. Elle est appréciée pour sa simplicité, sa rapidité d'exécution et sa capacité à gérer de grands ensembles de données. Cependant, il est important de noter que le choix initial des centroïdes peut influencer les résultats de l'algorithme, et que l'algorithme peut converger vers un optimum local plutôt que global.

La figure 4.3 représente d'une manière claire la corrélation entre nos variables :

* les variables fortement coorelées positivement sont : "Beds" et "PropertySqft", "Price" et "PropertySqft", "Bath" et "Beds" , "Bath" et "PropertySqft" .

* Les variables "Longitude" , "Beds", "Latitude" et "Price" sont faiblement coorelées.

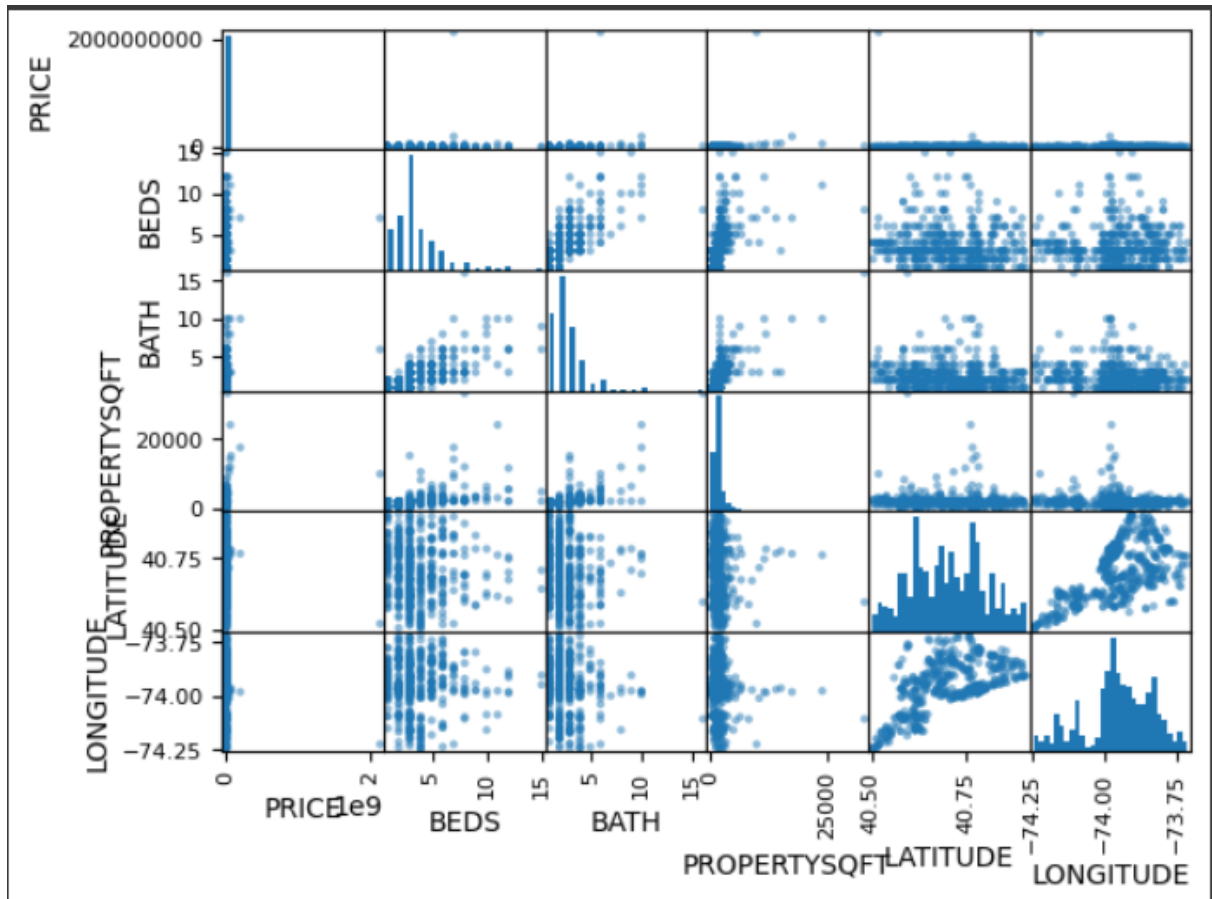


FIGURE 4.3: Corrélation entre les variables

Les figure 4.4 et 4.5 représentent la répartition des individus sur 4 clusters.

A titre d'exemple, les individus d'indices 0, 3 et 499 appartiennent au cluster n°1. Les individus d'indices 1, 2 et 4 appartiennent au cluster n°2.

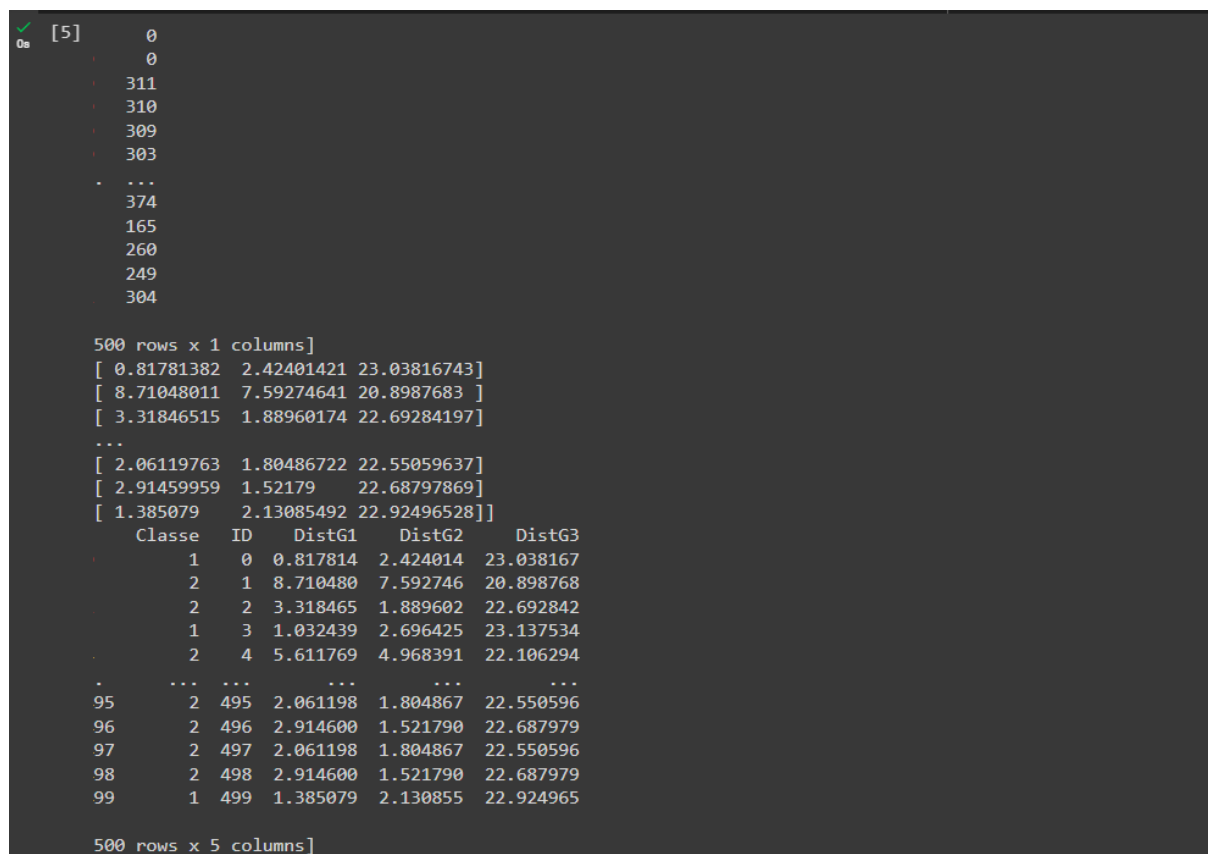
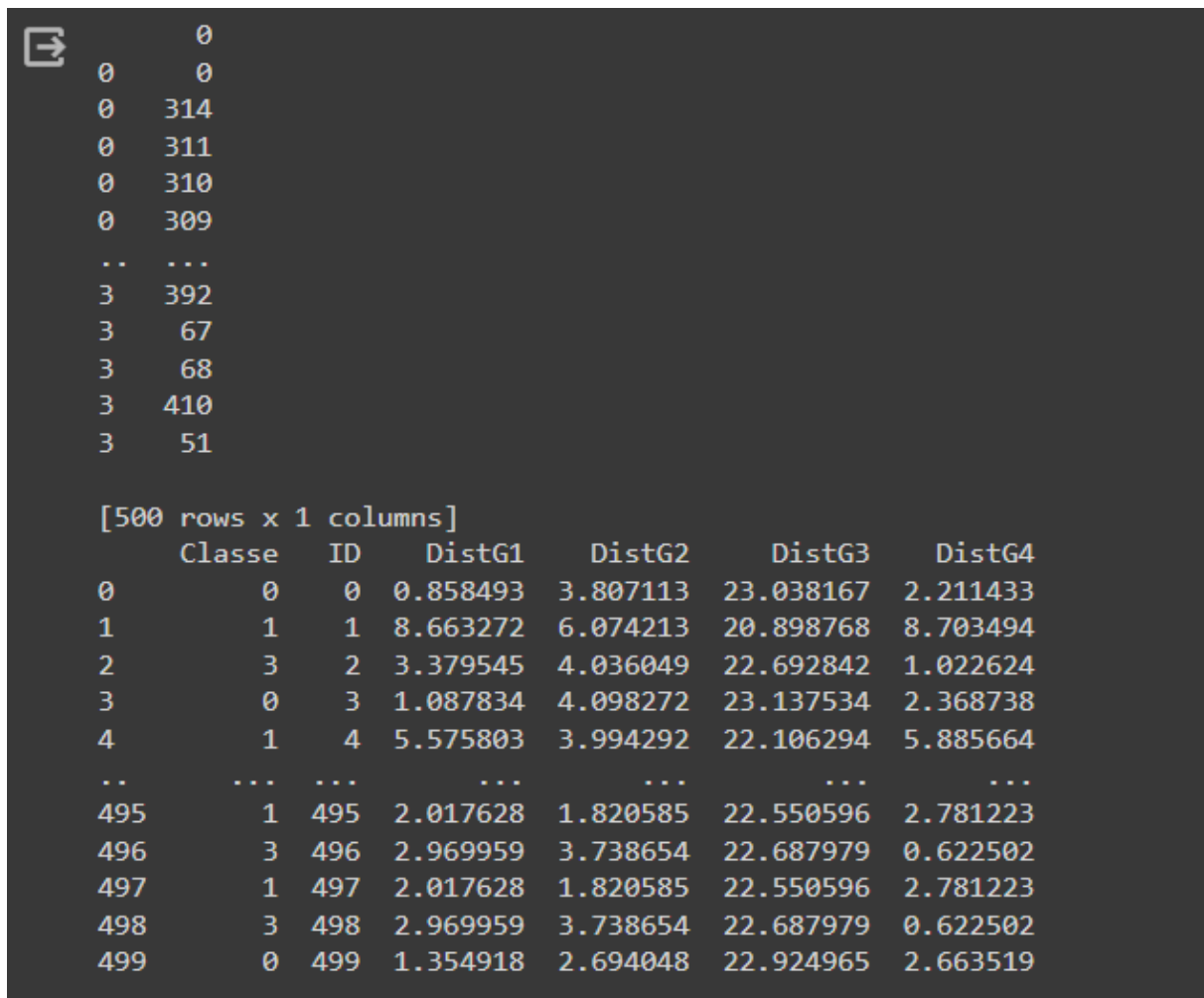


FIGURE 4.4: Répartition des individus sur 4 clusters



```

0
0 0
0 314
0 311
0 310
0 309
.. ...
3 392
3 67
3 68
3 410
3 51

[500 rows x 1 columns]

```

	Classe	ID	DistG1	DistG2	DistG3	DistG4
0	0	0	0.858493	3.807113	23.038167	2.211433
1	1	1	8.663272	6.074213	20.898768	8.703494
2	3	2	3.379545	4.036049	22.692842	1.022624
3	0	3	1.087834	4.098272	23.137534	2.368738
4	1	4	5.575803	3.994292	22.106294	5.885664
..
495	1	495	2.017628	1.820585	22.550596	2.781223
496	3	496	2.969959	3.738654	22.687979	0.622502
497	1	497	2.017628	1.820585	22.550596	2.781223
498	3	498	2.969959	3.738654	22.687979	0.622502
499	0	499	1.354918	2.694048	22.924965	2.663519

FIGURE 4.5: Répartition des individus sur 4 clusters(suite)

La figure 4.6 représente la méthode du coude qui est une technique utilisée pour déterminer le nombre optimal de clusters dans un algorithme de clustering, comme k-means. L'objectif est de trouver le point où l'ajout de clusters supplémentaires n'apporte plus une amélioration significative à la qualité de la partition des données. Voici comment fonctionne la méthode du coude :

1. ****Exécution de l'algorithme de clustering pour différents nombres de clusters**** : Vous exécutez l'algorithme de clustering (par exemple, k-means) sur vos données pour un certain nombre de valeurs de k (le nombre de clusters). Vous choisissez généralement une plage de valeurs de k, par exemple de 1 à 10.
2. ****Calcul de la mesure de qualité du clustering**** : Pour chaque valeur de k, vous calculez une mesure de qualité du clustering, telle que la somme des carrés des distances des points à leurs centres de cluster (SSD en anglais). Cette mesure évalue à quel point les points sont proches de

leur centre de cluster respectif.

3. **Représentation graphique des résultats** : Vous tracez un graphique où l'axe des x représente le nombre de clusters k et l'axe des y représente la mesure de qualité du clustering. Cela vous donne une courbe qui montre comment la qualité du clustering évolue en fonction du nombre de clusters.

4. **Identification du "coude" dans le graphique** : Vous examinez le graphique pour identifier le point où l'ajout de clusters supplémentaires ne conduit plus à une diminution significative de la mesure de qualité du clustering. Ce point est appelé le "coude" du graphique. Au-delà de ce point, l'amélioration de la qualité du clustering devient marginale.

5. **Choix du nombre optimal de clusters** : Le nombre optimal de clusters est généralement choisi comme la valeur de k correspondant au coude dans le graphique. Cependant, il peut parfois être nécessaire de prendre en compte d'autres facteurs, tels que le contexte de l'application ou les exigences spécifiques du problème.

C'est la méthode la plus simple mais efficace pour déterminer le nombre optimal de clusters dans un algorithme de clustering. Elle permet de trouver un compromis entre la complexité du modèle (nombre de clusters) et sa capacité à expliquer la structure des données.

Dans notre cas, le point où l'ajout de clusters supplémentaires ne conduit plus à une diminution significative de la mesure de qualité du clustering est $= 2$.

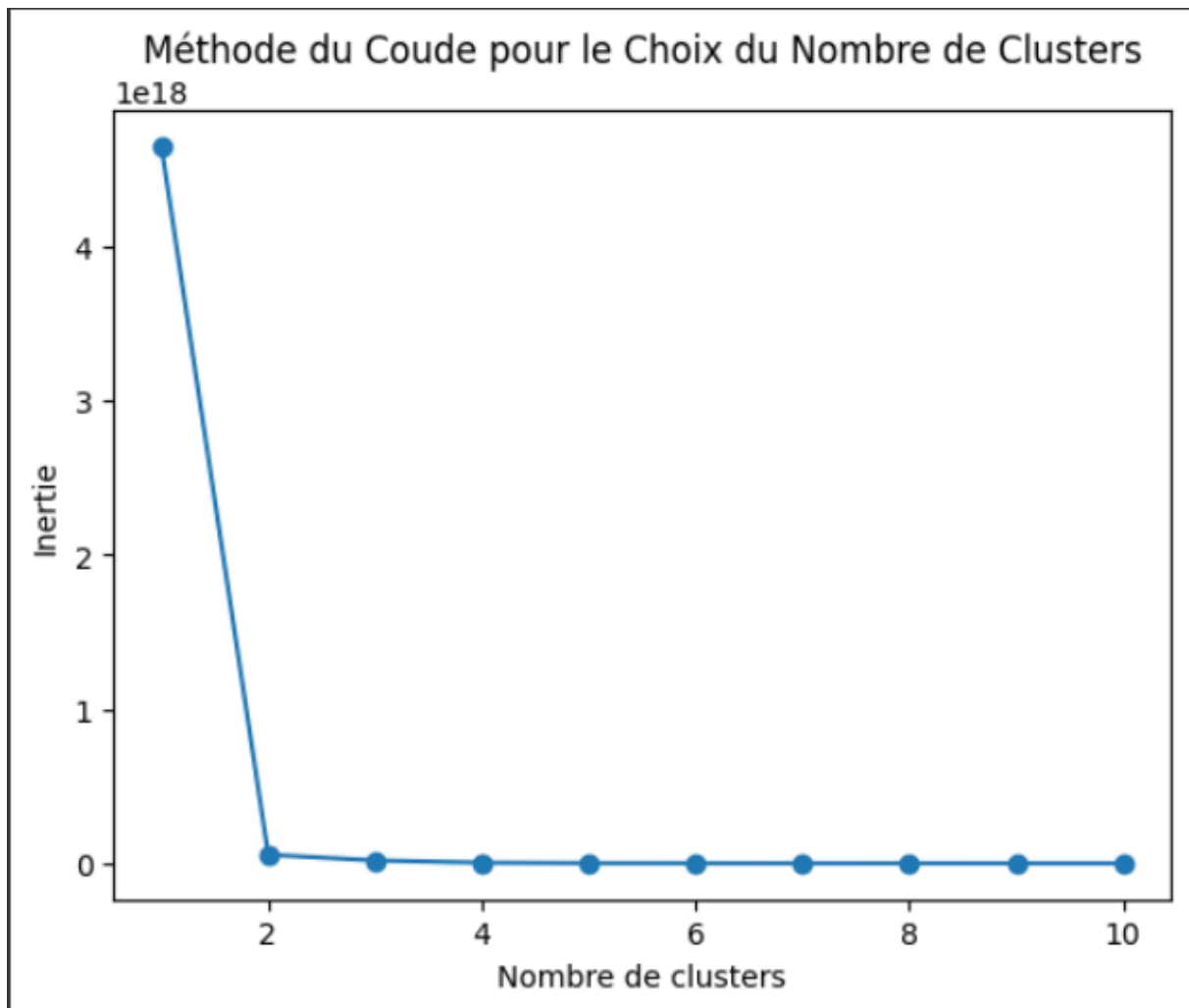


FIGURE 4.6: Méthode du coude pour le choix du nombre de clusters

La figure 4.7 représente les moyennes de caractéristiques de chaque cluster sont des valeurs qui représentent le centre de chaque cluster dans un algorithme de clustering tel que k-means. Chaque caractéristique est représentée par une valeur moyenne calculée à partir des points de données qui appartiennent à ce cluster spécifique. Ces moyennes peuvent fournir des informations importantes sur les caractéristiques distinctives de chaque cluster. Voici quelques-unes des façons dont les moyennes de caractéristiques peuvent être utiles :

1. ****Interprétation des clusters**** : Les moyennes de caractéristiques permettent d'interpréter les clusters en identifiant les valeurs moyennes des différentes variables pour chaque cluster. Par exemple, dans une analyse de segmentation de marché, vous pouvez utiliser les moyennes de caractéristiques pour comprendre les caractéristiques démographiques ou comportementales distinctives de chaque segment de clientèle.
2. ****Comparaison entre les clusters**** : En comparant les moyennes de caractéristiques entre

les clusters, vous pouvez identifier les différences significatives et les similitudes entre eux. Cela peut vous aider à comprendre les profils distincts des groupes de données regroupés dans chaque cluster.

3. ****Validation des clusters**** : Les moyennes de caractéristiques peuvent être utilisées pour valider la qualité des clusters en examinant à quel point les valeurs moyennes sont similaires à l'intérieur de chaque cluster et différentes entre les clusters. Des clusters bien définis auront des moyennes de caractéristiques cohérentes et distinctes.

La variable la plus discriminante dans les 2 clusters est PRICE !

Moyennes des caractéristiques de chaque cluster :						
cluster	PRICE	BEDS	BATH	PROPERTYSQFT	LATITUDE	LONGITUDE
0	1.023357e+06	2.905775	1.905775	1716.088093	40.741916	-73.896499
1	5.254058e+06	4.876471	3.664706	3253.207270	40.624700	-74.058098
2	2.147484e+09	7.000000	6.000000	10000.000000	40.518484	-74.224418

Différences significatives entre les clusters :						
cluster	PRICE	BEDS	BATH	PROPERTYSQFT	LATITUDE	LONGITUDE
1	4.230701e+06	1.970696	1.758931	1537.119176	-0.117215	-0.161598
2	2.142230e+09	2.123529	2.335294	6746.792730	-0.106216	-0.166321

Caractéristiques les plus discriminantes pour chaque cluster :						
Cluster 1:						
Caractéristique la plus discriminante: PRICE, Différence: 4230700.97677454						
Cluster 2:						
Caractéristique la plus discriminante: PRICE, Différence: 2142229589.1235294						

FIGURE 4.7: Les moyennes de caractéristiques de chaque cluster

La figure ci-dessous représente le clustering de nos données (nos individus) sous forme de 3 clusters.

On peut remarquer qu'il existe un cluster formé d'un seul individu !

Distribution dispersée !

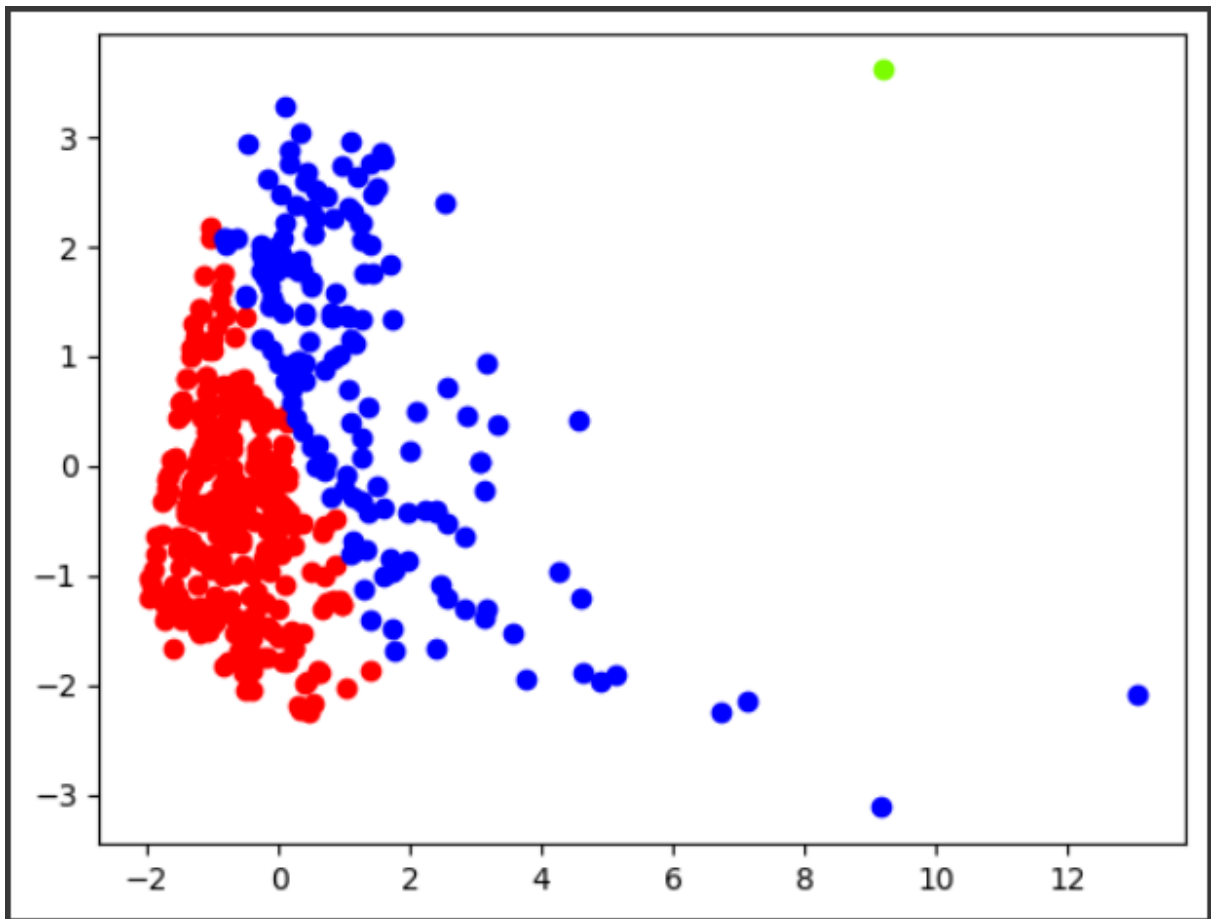


FIGURE 4.8: Le clustering de données

4.3 K-means après élimination des individus

Après avoir éliminé les maisons qui causaient une dispersion excessive de la distribution, nous avons réappliqué la méthode K-means.

La figure ci-dessous représente la Méthode du coude pour le choix du nombre de clusters .

Selon la méthode de coude, on a $K=5$.

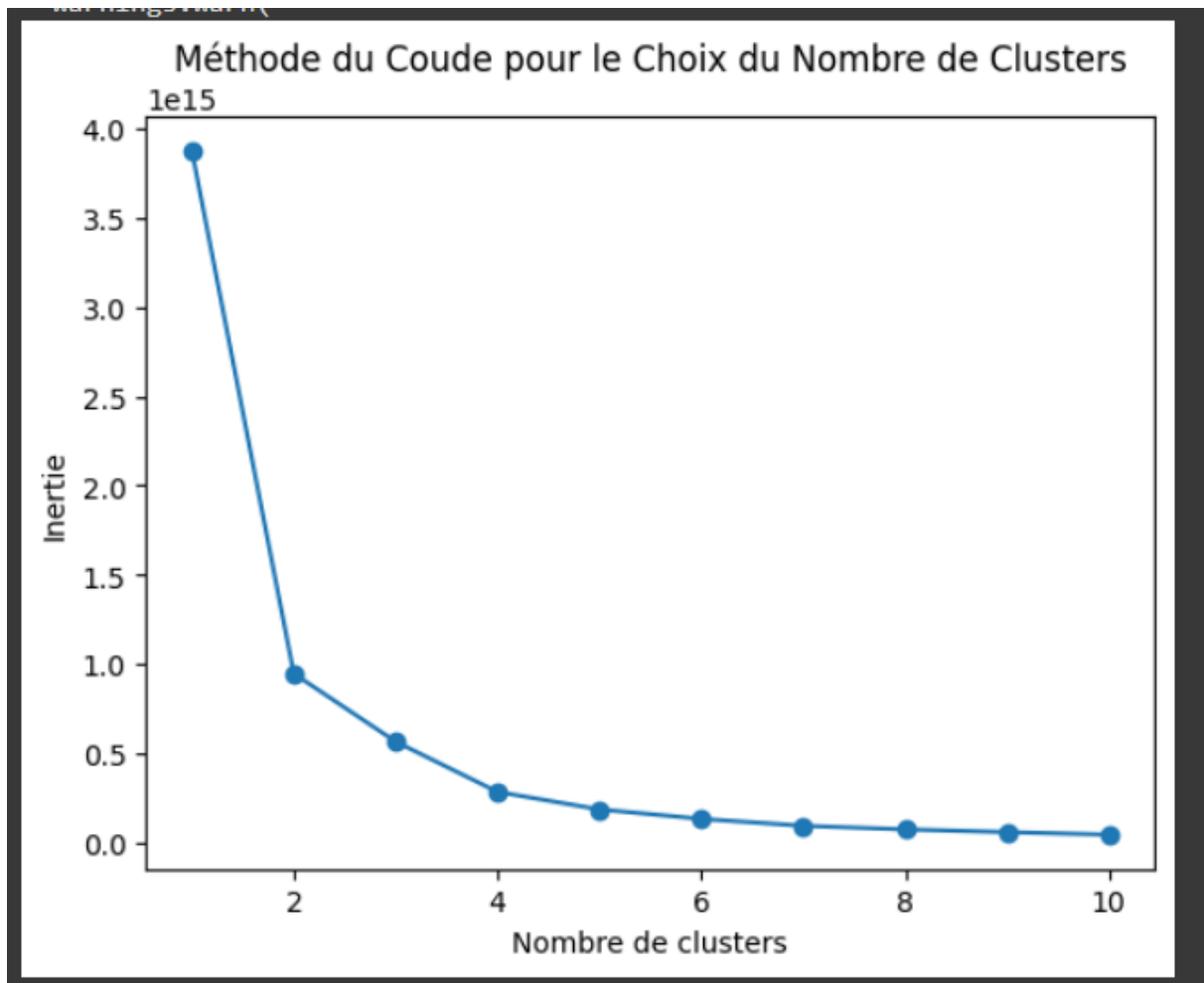


FIGURE 4.9: Méthode du coude pour le choix du nombre de clusters après modification

La figure 4.10 représente les nouvelles moyennes de caractéristiques de chaque cluster

Moyennes des caractéristiques de chaque cluster :						
cluster	PRICE	BEDS	BATH	PROPERTYSQFT	LATITUDE	LONGITUDE
0	9.451380e+05	2.987138	1.938907	1745.318941	40.747108	-73.889544
1	5.735776e+06	7.214286	5.000000	3854.980537	40.701013	-73.959481
2	8.285313e+05	3.161290	2.451613	1780.847725	40.586848	-74.100602
Différences significatives entre les clusters :						
cluster	PRICE	BEDS	BATH	PROPERTYSQFT	LATITUDE	LONGITUDE
1	4.790638e+06	4.227147	3.061093	2109.661596	-0.046094	-0.069938
2	-4.907245e+06	-4.052995	-2.548387	-2074.132812	-0.114166	-0.141120
Caractéristiques les plus discriminantes pour chaque cluster :						
Cluster 1:						
Caractéristique la plus discriminante: PRICE, Différence: 4790637.865698208						
Cluster 2:						
Caractéristique la plus discriminante: LATITUDE, Différence: -0.11416573732719115						

FIGURE 4.10: Les moyennes de caractéristiques de chaque cluster après modification

La figure 4.11 représente le clustering de nos données (nos individus) sous forme de 3 clusters.

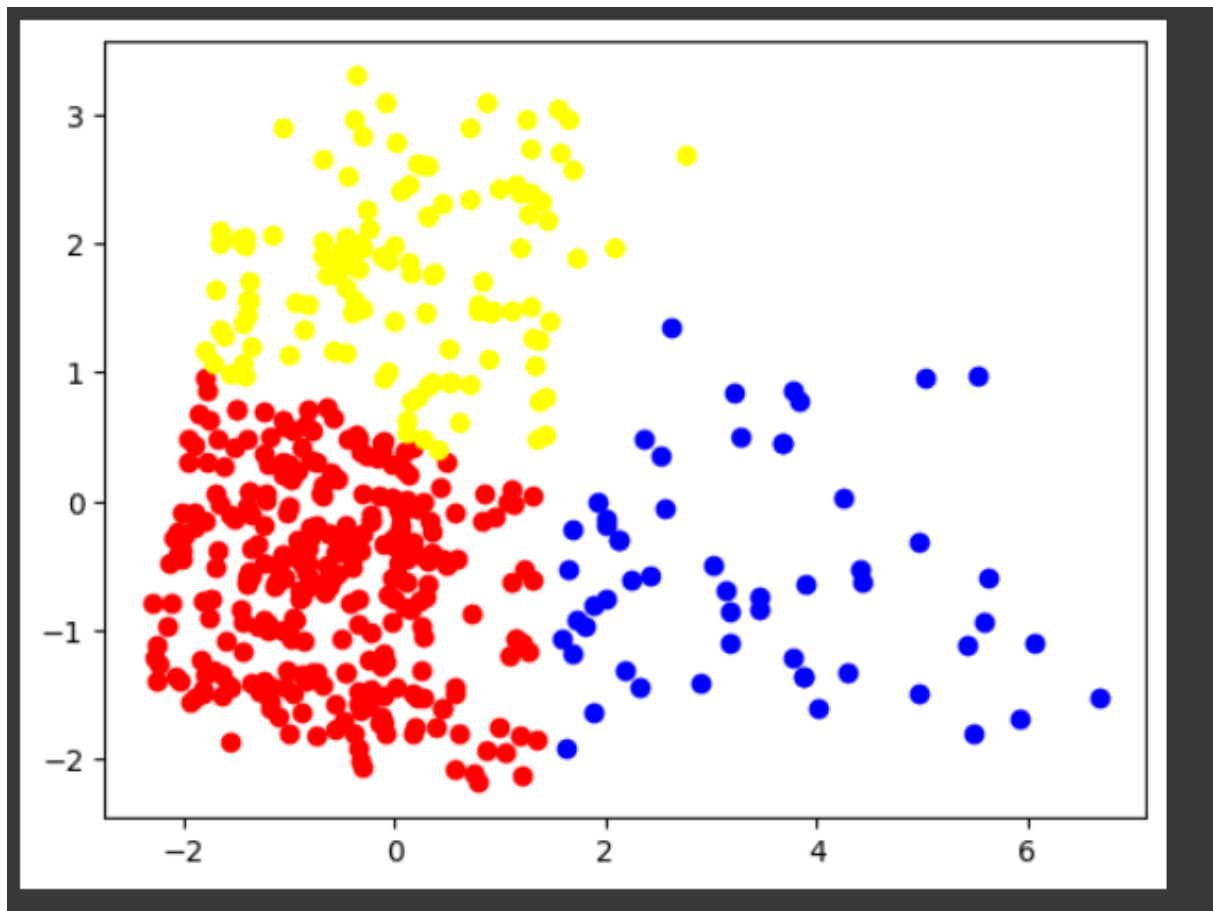


FIGURE 4.11: Le clustering de données après modification

Après élimination de l'individu qui a contribué à la dispersion de nos clusters, on obtient 3 clusters bien répartis entre eux et homogènes !

4.4 Comparaisons entre CAH et K-means

L'Analyse Hiérarchique des Clusters (CAH) et l'algorithme K-means sont deux méthodes de clustering largement utilisées dans l'analyse de données. Voici une comparaison entre les deux :

CAH : La CAH est une méthode de clustering agglomérative, ce qui signifie qu'elle commence par considérer chaque observation comme un cluster individuel et fusionne progressivement les

clusters les plus similaires jusqu'à ce qu'il ne reste qu'un seul cluster contenant toutes les observations.

K-means : K-means est une méthode de clustering partitionnelle qui divise les données en K clusters prédéfinis. Il commence par initialiser aléatoirement les centres de cluster, puis attribue chaque observation au cluster dont le centre est le plus proche. Il réajuste ensuite les centres de cluster en calculant la moyenne des observations appartenant à chaque cluster, et ce processus est répété jusqu'à ce que les centres convergent. La méthode k-means utilisée nécessite normalement moins de calculs et est adaptée à de très grand ensemble de données.

Conclusion Générale

Après avoir examiné les données comprenant 500 individus représentant des propriétés à travers New York City, il est possible de tirer plusieurs conclusions importantes.

Premièrement, en ce qui concerne les variables quantitatives, il est clair que le prix des maisons varie en fonction de plusieurs facteurs, notamment le nombre de chambres, le nombre de salles de bain, la surface de la propriété, ainsi que la latitude et la longitude, ce qui indique une corrélation potentielle entre l'emplacement géographique et le prix des propriétés.

Deuxièmement, les variables qualitatives telles que le type de propriété et la ville de la propriété offrent également des informations précieuses. Le type de propriété peut influencer non seulement le prix, mais aussi d'autres caractéristiques telles que la taille et les équipements disponibles. De même, la ville de la propriété peut jouer un rôle crucial dans la détermination de la valeur, en raison des différences de quartiers, de commodités locales et d'autres facteurs environnementaux.

En conclusion, ce projet offre une opportunité passionnante d'explorer les relations complexes entre les caractéristiques des propriétés et leur valeur à travers New York City. En utilisant des techniques d'analyse de données appropriées, il est possible d'identifier les principaux déterminants du prix des propriétés et de développer des modèles prédictifs précis pour aider les acheteurs, les vendeurs et les investisseurs immobiliers à prendre des décisions éclairées.