

UNIVERSITÉ DU QUÉBEC À CHICOUTIMI

RAPPORT TP-1

PRÉSENTÉ À

MONSIEUR JULIEN MAITRE

COURS

SUJET SPECIAL - FONDAMENTAUX DE L'APPRENTISSAGE AUTOMATIQUE

(8INF950)

PAR

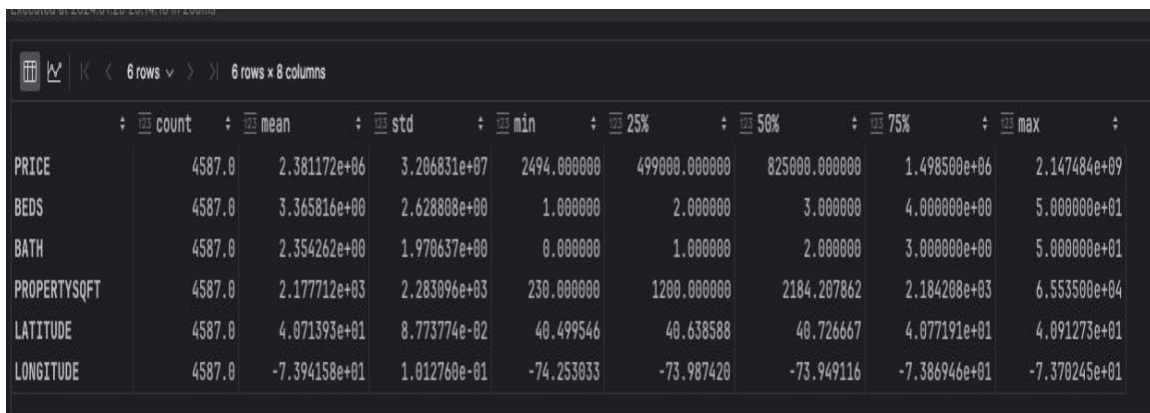
KOFFI DELADEM MOISE ETOU (ETOK01059800)

JEU DE DONNEE: NEW YORK HOUSING MARKET

01 FEVRIER 2024

## I- DESCRIPTION DU JEUX DONNÉES

Le jeu de données que nous allons explorer et traiter est un ensemble de données provenant de la plateforme Kaggle, qui contient des informations immobilières relatives aux prix des maisons à New York. Notre jeu de données comprend 17 colonnes et 4801 observations. Après première analyses de notre jeu de données nous avons 214 lignes en double que nous avons supprimés. Il reste donc 5487 lignes. Nous n'avons pas de valeur nulle et manquante dans nos données. La représentation graphique grâce au diagramme boxplot et au tableau statistique révèlent plusieurs caractéristiques des maisons à New York. On observe que le prix médian des maisons est de 825 000 \$ et que la moyenne est significativement plus élevée. On en déduit donc la présence de valeurs extrêmement élevées comme en témoigne une valeur maximale énorme de près de 2,15 milliards. Une observation visible sur la boxplot des prix à New YORK confirme cette analyse. Le nombre moyen de chambres est d'environ 3,37 avec une médiane similaire à 3, mais une valeur maximale de 50 chambres, ce qui peut également indiquer des valeurs aberrantes. Hypothèse toujours confirmé par la boxplot. De même, le nombre moyen de salles de bain est d'environ 2,35 avec une médiane de 2 et une valeur maximale de 50 salles de bain, suggérant également des valeurs aberrantes. La superficie médiane des propriétés est d'environ 2 184,21 pieds carrés, avec une moyenne similaire, mais une valeur maximale de 65 535 pieds carrés, qui semble être une valeur aberrante.



	count	mean	std	min	25%	50%	75%	max
PRICE	4587.0	2.381172e+06	3.206831e+07	2494.000000	499000.000000	825000.000000	1.498500e+06	2.147484e+09
BEDS	4587.0	3.365816e+00	2.628008e+00	1.000000	2.000000	3.000000	4.000000e+00	5.000000e+01
BATH	4587.0	2.354262e+00	1.970637e+00	0.000000	1.000000	2.000000	3.000000e+00	5.000000e+01
PROPERTYSQFT	4587.0	2.177712e+03	2.283096e+03	230.000000	1200.000000	2184.207862	2.184208e+03	6.553500e+04
LATITUDE	4587.0	4.071393e+01	8.773774e-02	40.499546	40.638588	40.726667	4.077191e+01	4.091273e+01
LONGITUDE	4587.0	-7.394158e+01	1.012760e-01	-74.253033	-73.987420	-73.949116	-7.386946e+01	-7.370245e+01

*Figure1 : Tableau statistique des caractéristiques des maisons à New York*

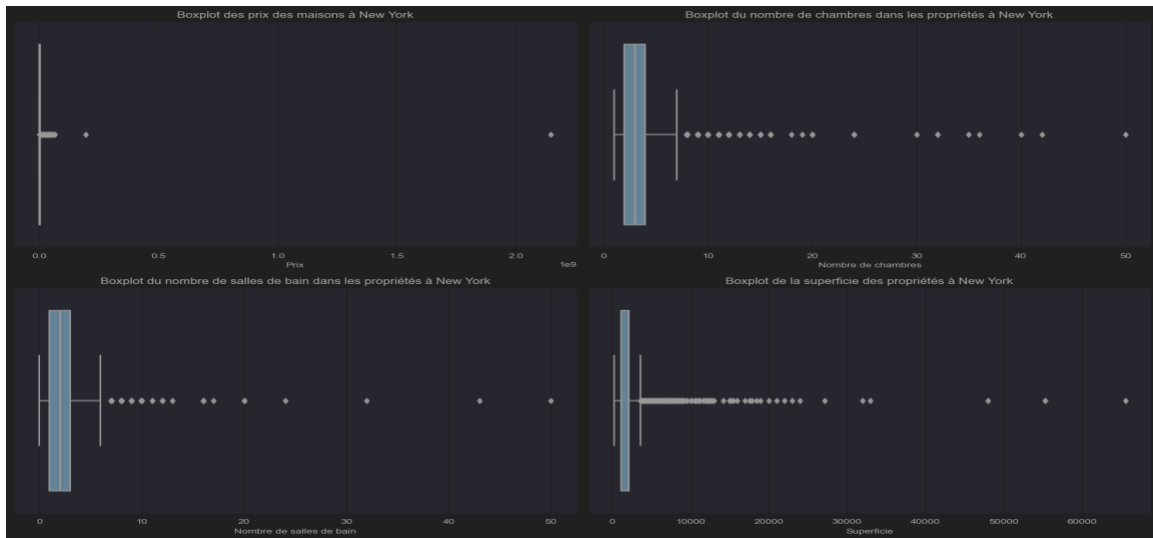


Figure2 : Diagramme de boxplot du prix, du nombre de chambres, de salles de bain et de superficie des maisons à New York

Un nettoyage de valeur aberrante a été effectué afin d’assurer la cohérence des données. Il reste 3495 observations.

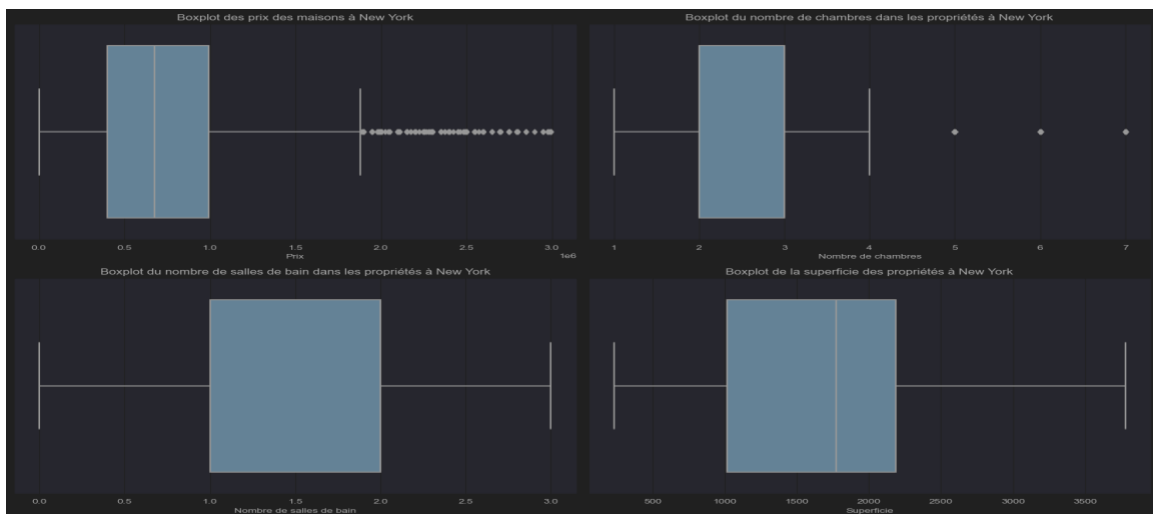


Figure3 : Diagramme de boxplot du prix, du nombre de chambres, de salles de bain et de superficie des maisons à New York après nettoyage des valeurs aberrantes

En affichant la distribution des données des colonnes BEDS et BATH, cela nous permet de visualiser rapidement la répartition des valeurs dans ces colonnes et de repérer des tendances ou

des caractéristiques intéressantes dans les données. Ainsi nous déduisons que la majeure partie des maisons de nos jeux de donnée est de 3. De même la majeure partie des maisons on 2 Salle de bain.

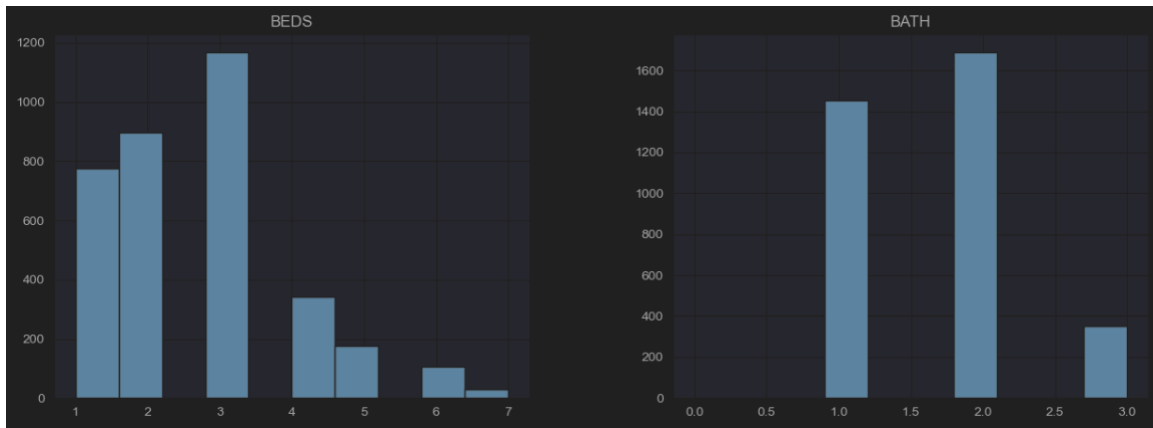


Figure4 : Histogramme de distribution des données des colonnes BEDS et BATH

Le diagramme ci-dessous représente la répartition des types de propriétés dans notre ensemble de données. Nous avons 13 Types de propriété et en examinant le diagramme, nous pouvons observer que le type de propriété le plus courant dans notre ensemble de données est : « Co-op for sale » avec 36,5% suivie par « House for sale » 21,5%, « condo for sale » 19,6% et « Multi-family home for sale » 10,3%.

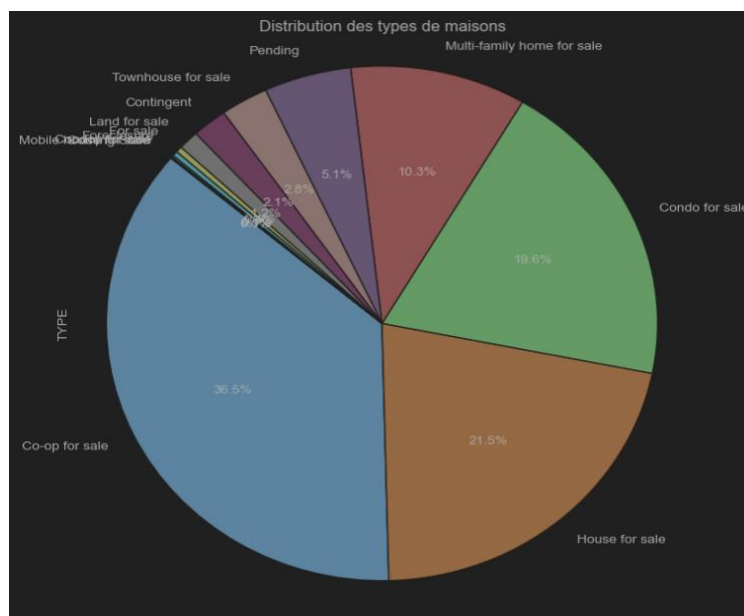
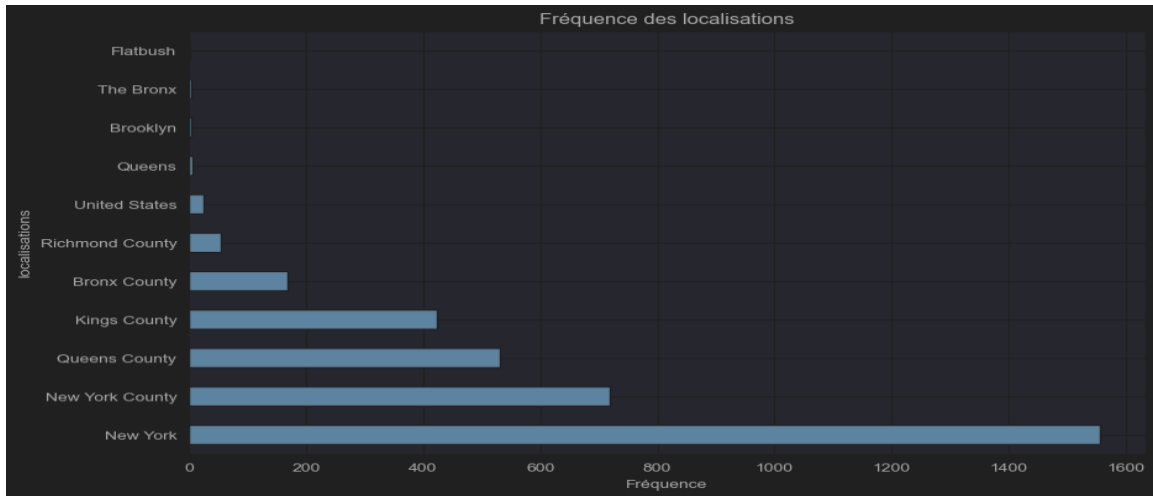


Figure4 : Répartition des types de propriété à New york

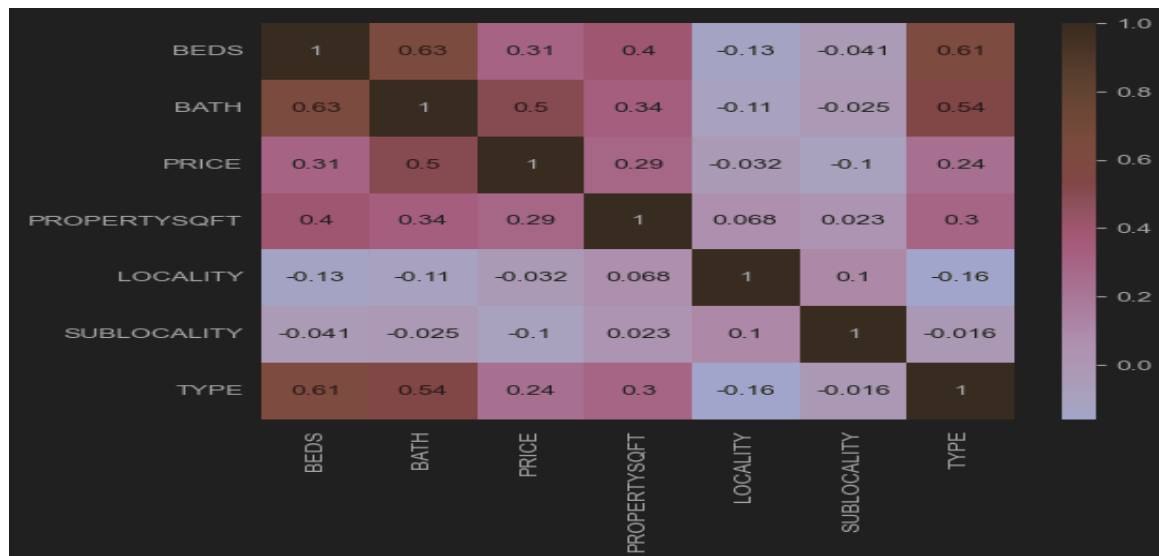
Le diagramme en barres de la figure 5 représente la fréquence des 11 localités présentes dans le jeu de données. Nous pouvons ainsi remarquer que la plus grande localité dans laquelle nous avons des détails immobiliers est « New York », suivie de « New York County » puis « Queens County » et « Kings County ».



*Figure5 : Répartition des propriétés en fonction de la localisation à New york*

La matrice de corrélation de la figure6 nous fournit des informations sur les relations linéaires entre le prix des maisons (**PRICE**), le nombre de chambres (**BEDS**), le nombre de salles de bains (**BATH**), la superficie de la propriété (**PROPERTYSQFT**). Ainsi d'après les observations faites la matrice de corrélation indique qu'il y a une corrélation positive forte (0.63) entre le nombre de chambres (**BEDS**) et le nombre de salles de bain (**BATH**). On pourrait donc dire que les maisons avec un plus grand nombre de chambres ont tendance à avoir un plus grand nombre de salles de bain. Il y'a aussi une corrélation positive modérée entre le nombre de chambres (**BEDS**) et la superficie de la propriété (**PROPERTYSQFT**) et le nombre de salles de bain (**BATH**) avec la superficie de la propriété (**PROPERTYSQFT**) respectivement (0,442687) et (0,500568) Cela signifie que les maisons avec un plus grand nombre de chambres ont tendance à avoir une superficie de propriété plus grande et que les maisons avec un plus grand nombre de salles de bain ont tendance à avoir une superficie de propriété plus grande. On remarque aussi que Les coefficients

de corrélation entre le prix (PRICE) et une corrélation positive modérer entre le nombre de salle de bain de la propriété (**BATH**) et le nombre de Chambre) respectivement (0,5) et (0,31).



*Figure6 : Matrice de corrélation des variable PRICE, BATH, BED, PROPERTYSQFT*

## II- ETAPE TRAITEMENT DU JEUX DE DONNEES

### a) Nettoyage des données

- **Valeur manquante**: Comme mentionné dans la description des données notre jeux de donnée ne contient aucune valeur manquante. Aucun traitement n'a donc été effectué.
- **Suppression de doublons** : Les doublons dans nos jeux de données correspondent à des entrées répétées de la même information d'une maison. Ils ont été supprimés afin d'éviter toute erreur potentielle et de garantir l'intégrité dans notre analyse.
- **Conversion du type** : Dans notre jeu de données, la colonne 'BATH' correspond au nombre de salles de bain dans une maison. Initialement de type 'float', nous avons remarqué la présence de nombres décimaux dans cette colonne. Étant donné que le nombre de salles de bain ne peut pas être décimal, nous avons converti la colonne en type 'int' pour garantir la cohérence de nos données.
- **Suppression des valeur Aberrantes** : Notre jeu de données présente plusieurs valeurs aberrantes, telles que des nombres élevés de chambres, de salles de bain et le prix des maisons. Les analyses approfondies pour comprendre les raisons de ces données ont révélé qu'il s'agit une incohérence

anormale des données. Afin de remédier à cette situation, nous avons choisi de supprimer ces données. Tenter de corriger une donnée pourrait altérer la relation de cette variable avec une autre colonne, notamment étant donné que notre objectif principal est d'analyser les dimensions entre nos données.

### **b) Normalisation des variables**

- **La normalisation standard :** Afin de ramener les données à la même échelle nous avons appliqué dessus la formule de la normalisation standard

### **c) Encoder des variables**

- **Encodage en entiers:** Nous avons encodé les variables TYPE, LOCALITY et SUBLOCALITY qui sont qualitatives, et qui peuvent avoir un impact sur le prix d'une maison. Cette démarche vise à mieux comprendre les relations présentes dans nos jeux de données.

## **4) Séparation des données**

- **Séparation des données:** les données ont été séparées pour avoir une matrice de variables explicatives (X\_train) et une matrice de variable cibles (Y\_train).

## **5) Traitement des données**

**Implémentation PCA - Méthode 1:** Dans un premier temps, nous avons réalisé l'analyse en composantes principales (PCA) en respectant les étapes suivantes :

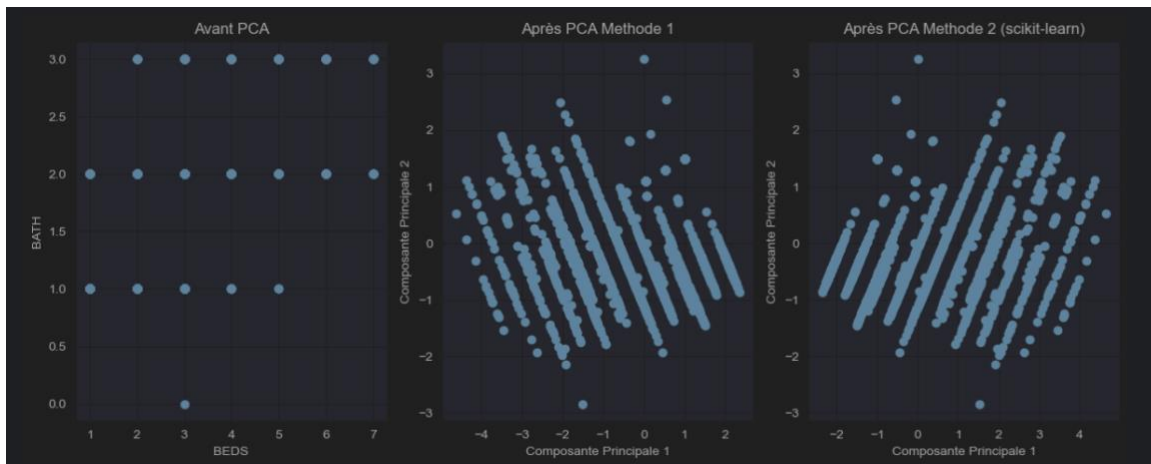
- Normalisation des données comme décrit précédemment.
- Calcul de la matrice de covariance.
- Calcul des valeurs propres et vecteurs propres de la matrice de covariance.
- Tri des valeurs propres en ordre décroissant et obtention des indices de tri
- Sélection des N premiers vecteurs propres pour la réduction de dimensionnalité (dans mon cas j'ai N= choisi 2).
- Transformation des valeurs standardisées en utilisant les vecteurs propres sélectionnés.
- Graphe de visualisation

**Implémentation PCA - Méthode 2 :** Dans un second temps, nous avons effectué l'analyse en composantes principales sur les mêmes données en utilisant la bibliothèque scikit-learn

- Standardisation avec la fonction `StandardScaler()`;
- Définition du nombre d'axe de composant principal et création du model PCA avec fonction `PCA()` qui prend en paramètre le nombre de composant dans mon cas 2
- Entraînement et transformation des valeurs standardisées avec la fonction `fit_transform()`
- Graphe de visualisation

L'implémentation de c'est deux approches nous a permis de réduire la dimensionnalité de nos données d'identifier les composantes principales de notre jeu de donnée et de comparer les résultats des deux méthodes, nous pouvons évaluer l'efficacité de notre implémentation par rapport à l'approche standard de scikit-learn

### III- Illustration avant et après PCA



**Figure 7 : Illustration de la comparaison des méthodes des données avant PCA et après PCA avec la méthode 1 et 2**

À partir de l'illustration ci-dessus, nous observons initialement (sans l'utilisation de la PCA) qu'il est difficile d'observé toutes les données en plus. Nous remarquons aussi une forte colinéarité entre les variables explicatives BATH et BEDS. Après l'application de l'algorithme PCA, nous constatons une réduction de la dimensionnalité et une réorganisation des variances. Il est aussi



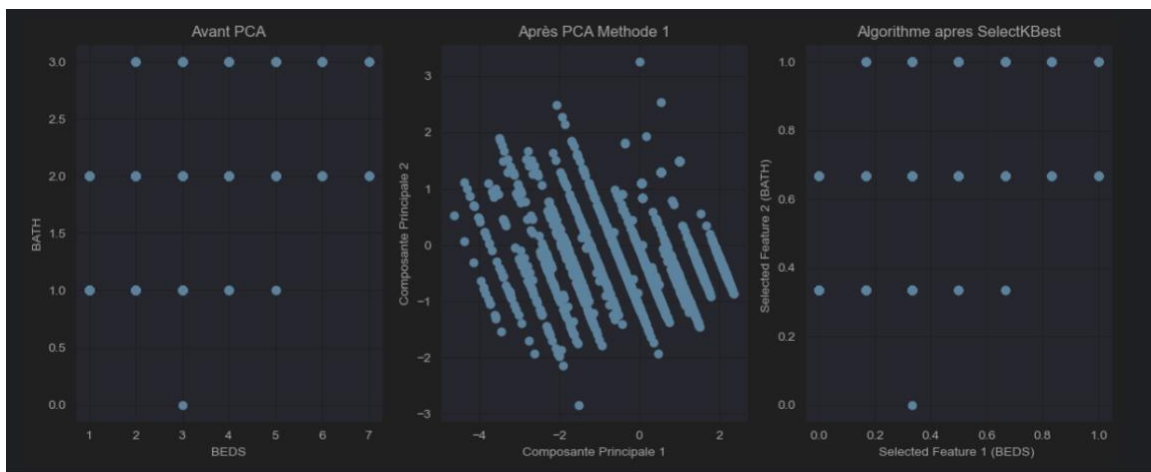
observé la suppression de colinéarité précédemment observer dans la première composante principale et la seconde.

#### **IV- Analyse et comparaisons de mon implémentation de PCA avec celle de la librairie scikit-learn**

L'analyse en composantes principales (PCA) nous a permis de réduire la dimensionnalité de nos données et d'identifier les composantes principales qui expliquent la variance maximale dans notre ensemble de données. Les résultats obtenus des deux méthodes sont approximativement identiques. Cependant, Comme nous pouvons le voir sur la figure7 nous remarquons que la réduction de dimension avec les deux méthodes n'a pas effectuée à la même échelle. La présentation des données entre mon implémentation et celle de la librairie scikit-learn sont symétrique.

Niveau implémentation il est plus simple et plus rapide d'implémenter le PCA avec la Library scikit-learn que de coder soit même les méthodes de calcul. En plus avec Scikit-learn intègre d'autre outils une intégré transparente avec d'autres outils et algorithmes de Machine Learning

#### **V- Comparaison l'algorithme PCA et d'algorithme SelectKBest**



*Figure 8 : Illustration avant PCA, après PCA et après SelectKBest*

L'illustration ci-dessus (Figure8) met en évidence une différence importante entre l'algorithme PCA et SelectKBest. Contrairement au PCA, l'algorithme SelectKBest n'a pas transformé nos données

dans un nouvelle espace. Il s'est simplement contenté de choisir les meilleures caractéristiques (BATH et BEDS) afin de réduire la dimensionnalité tout en gardant les informations importantes de notre jeu de donnée.

## **VI- Conclusion**

Ce rapport d'analyse de données immobilières provenant de la plateforme Kaggle m'a permis d'appliquer et de comprendre les techniques de traitement des données ainsi que les bases du fonctionnement de l'Analyse en Composantes Principales (PCA). J'ai eu à implémenter cette dernière et à la comparer avec celle implémentée dans Scikit-learn. En parallèle, j'ai exploré le SelectKBest, une technique de sélection des caractéristiques. Contrairement au PCA, SelectKBest réduit également la dimensionnalité, et conserve les informations sans transformer les données. Je trouve particulièrement intéressant cet algorithme parce qu'il pourrait être combiné à d'autres algorithmes efficaces et qui prennent beaucoup de temps de traitement pour être plus performants sur un jeu de données volumineux. Cependant, un aspect moins apprécié de l'expérience que j'ai noté est l'inexactitude dans la prise de décision lors du traitement des données. En outre, j'ai appris reconnaître l'importance des prétraitements appropriés des données pour garantir des analyses précises. Dans l'ensemble, cette expérience m'a permis de développer mes compétences en analyse, en prétraitement, et en traitement de donnée tout en m'offrant un aperçu des techniques avancées de réduction de la dimensionnalité.

### ***Reference du jeu de données:***

<https://www.kaggle.com/datasets/nelgiriyeewithana/new-york-housing-market/data>