*Université Abdelmalek Essaadi Faculté des Sciences et techniques de Tanger Département Génie Informatique*



Rapport sur Atelier 2 :

« Clustering »

Réalisé par :

***LASIRI Nouhaila***

***BENSLIMAN Hamza***

Encadré par :

***Pr . ELAACHAk LOTFI***

*Cycle Ingénieur: LSI S4*

*Machine Learning*

*Année : 2019/2020*

Objectif :

L’objectif principal de cet atelier est de pratiquer les concepts du Clustering, en traitant les données d’une Data Sets.

Outils:

Visual Studio Code, Anaconda, Python, Pandas, Sklearn, matplotlib.

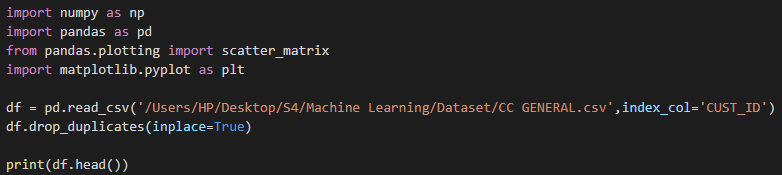
Data Sets:

Credit Card Data Set: <https://www.kaggle.com/arjunbhasin2013/ccdata>

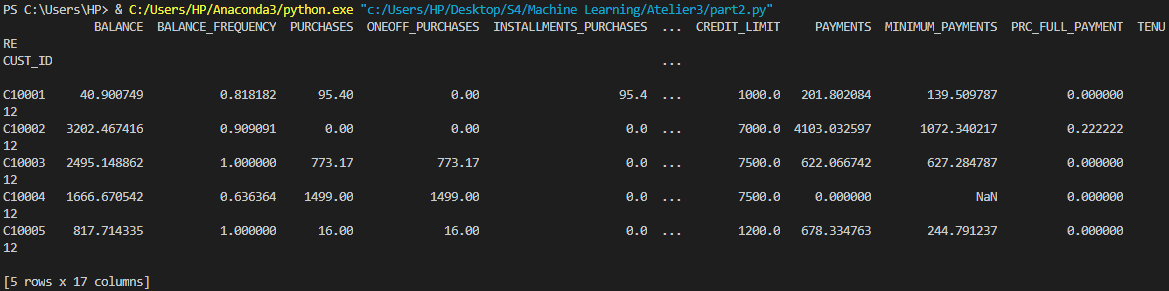
Partie 1 (Data Visualisation):

1. Exploration des données :

*Code Source :*

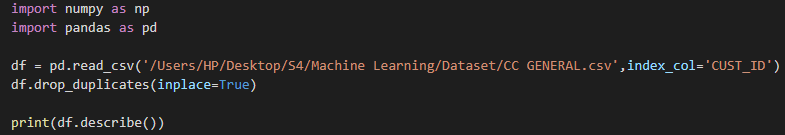


*Résultats :*

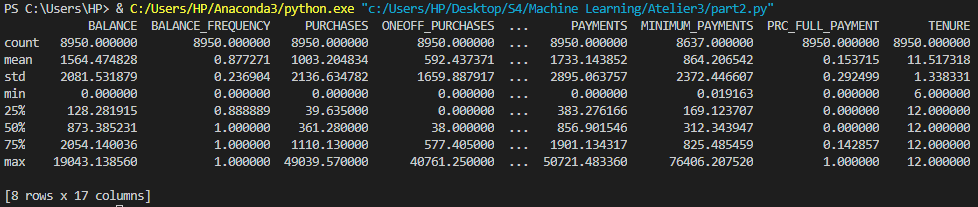


1. Le résumé statistique :

*Code Source :* Avec la fonction describe()

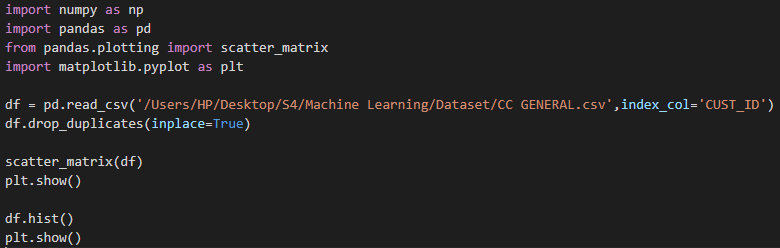


*Résultats :*

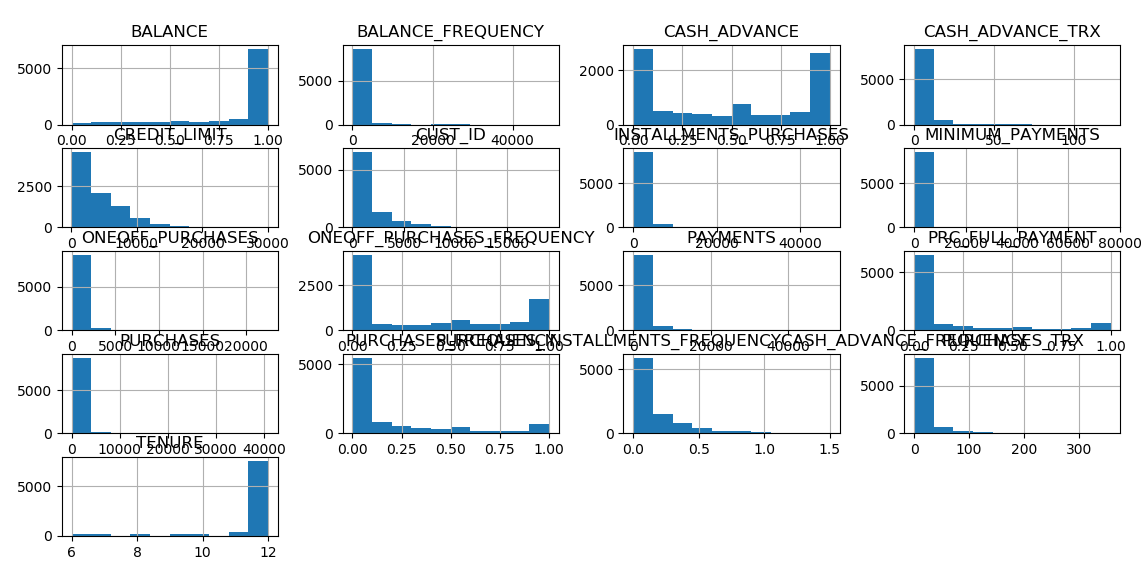


1. Les nuages des points du dataset :

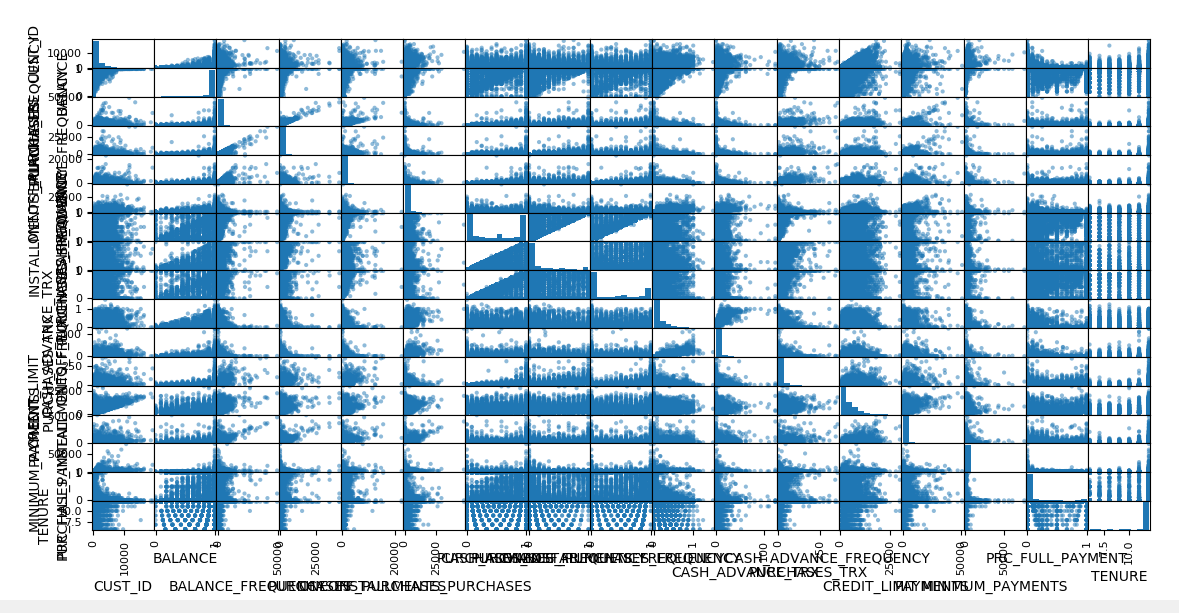
*Code Source :*



* Histogramme :



* Scatter plot matrix :

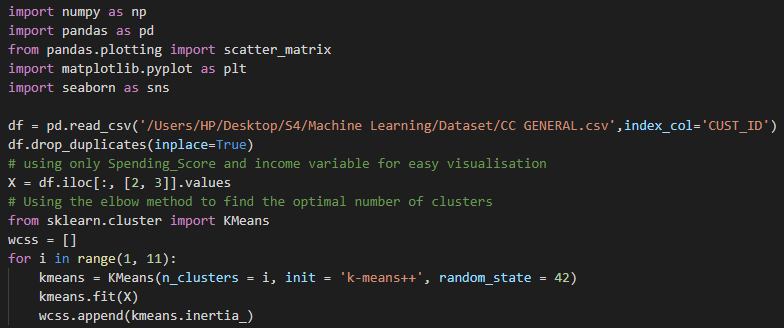


Partie 2 (Clustering):

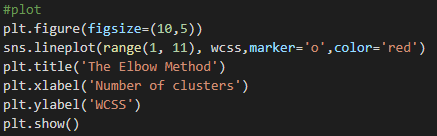
1 & 2. Construire le modèle de clustering avec sklearn &Kmeans + Définir le K nécessaire avec Elbow :

*Code Source :*

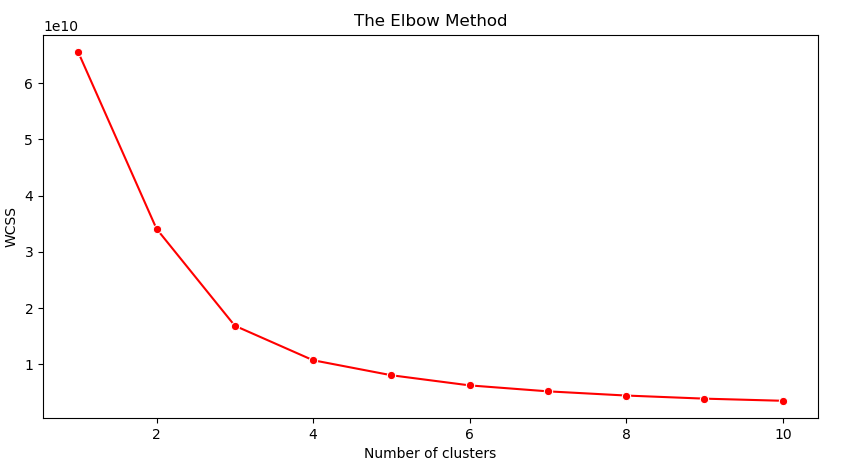
Kmeans et la methode Elbow pour choisir K :



La courbe de la methode Elbow :



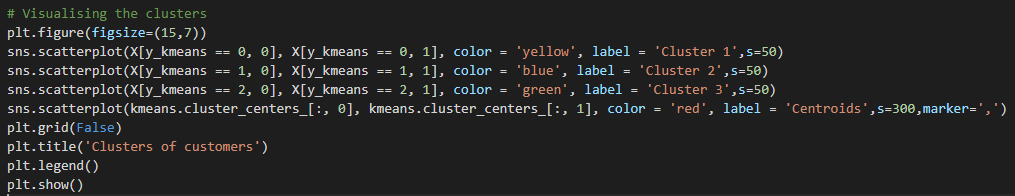
*Résultats :*



 On peut voir qu’après 3, il n’ya pas de diminution significative du WCSS, alors 3 est la meilleure K pour notre dataset.

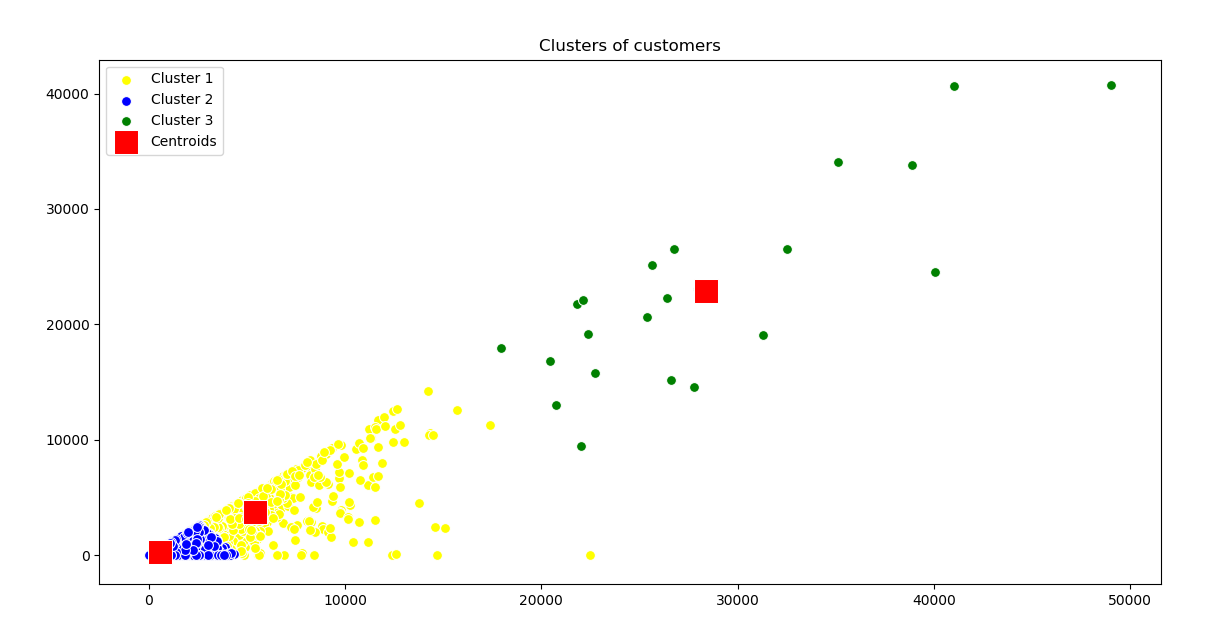
1. Représentation des clusters obtenus :

*Code Source :*



On a pris juste 3 Clusters comme on a trouvé avec Elbow.

*Résultats :*



1. Interprétation des résultats :

D’après ce graphe des résultats en fonction du BALANCE\_FREQUENCY et PURCHACES, on voit que l’algorithme K-means a mis les trois centroids dans le centre de chaque groupe coloré du dataset. Et c’est logique, car les gens riches (avec un grand BALANCE\_FRENQUENCY) dépense plus de solde dans ces achats (PURCHACES).