

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE MINISTÈRE DE L’ENSEIGNEMENT SUPÉRIEUR ET DE RECHERCHE SCIENTIQUE

UNIVERSITÉD’ALGER1

FACULTÉ DES SCIENCES

DEPARTEMENT MATHEMATIQUE ET INFORMATIQUE

# Projet du Module Data Mining

Domaine: INFORMATIQUE

Spécialité: ISIL

Thème

Classification automatique(Clustering)

Fait par:

MR AREZKI Mohemed Nadji

MR Benlalam Mohamed Rachid

MR Benazouz Ouilem

# Partie 1 :

1. **Classification hiérarchique Ascendante (CHA) :**

* **Application de l’algorithme de regroupement hiérarchique a main :**

Voire le fichier joint nommé « CHA.pdf ».

* **Représentation de l’ensemble des points :**

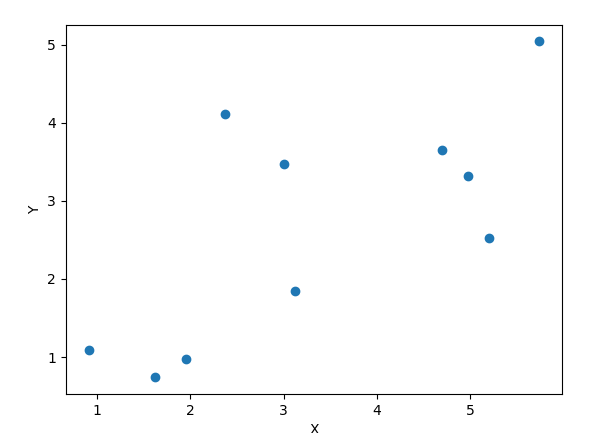


Figure 1 : Représentation de l’ensemble des points

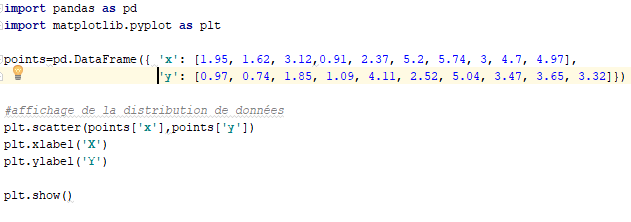


Figure 2 : Code python de l’ensemble des point

* **Dendrogramme et l’hiérarchie :**

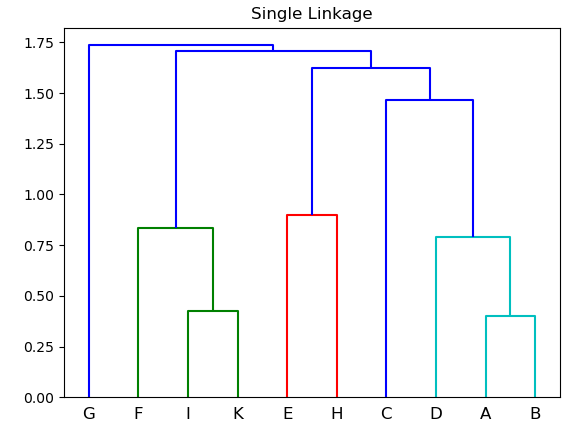


Figure 3 : Dendrogramme Single Linkage

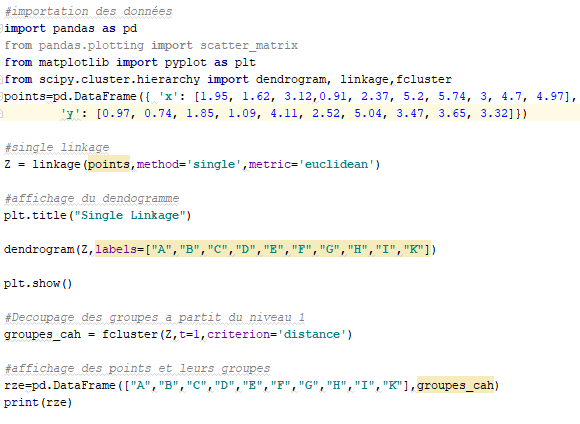


Figure 4 : Code python Dendrogramme Single Linkage

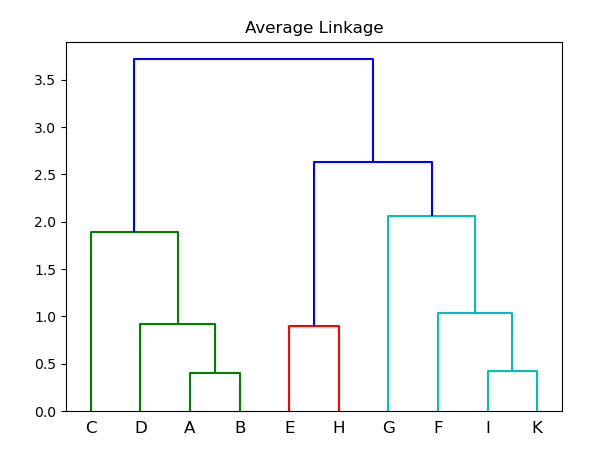


Figure 5 : Dendrogramme Average Linkage

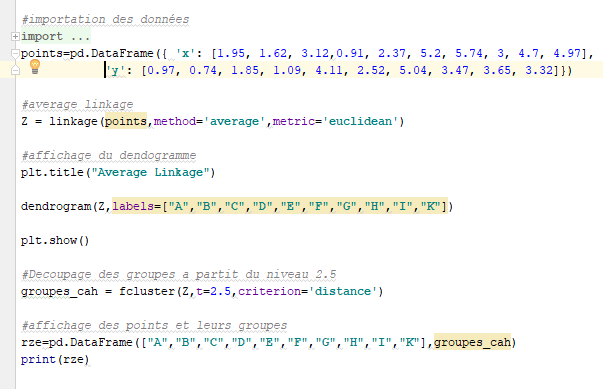


Figure 6 : Code python Dendrogramme Average Linkage

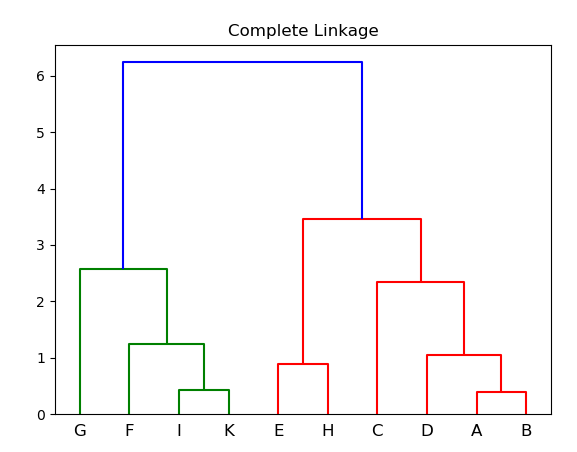


Figure 7 : Dendrogramme Complete Linkage

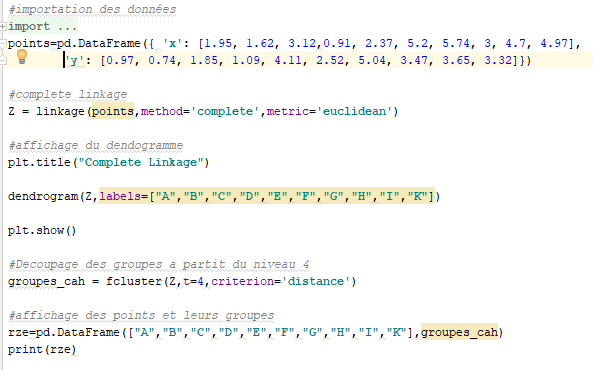


Figure 8 : Code python Dendrogramme Complete Linkage

* M**eilleur partitionnement :**

-Le meilleur partitionnements pour le Single Linkage est un partitionnement de niveau 1 avec le résultat de 5 groupe comme suit [A ,B,D] , [E,H] , [F,I,K] ,[C] , [G]

-Le meilleur partitionnements pour le Average Linkage est un partitionnement de niveau 2.5 avec le résultat de 3 groupes comme suit [A,B,C,D] , [E,H] ,[F,I,G,K]

Le meilleur partitionnement pour le Complete Linkage est un partitionnement de niveau 4 avec le résultat de 2 groupes comme suit [A ,B ,C ,D,E,H] , [F,I,G,K]

* **Vérification des résultats :**

L’application de l’algorithme CHA sur le papier et le code on donnée le même résultat.

1. **Kmeans :**

* **Application de l’algorithme Kmeans a main :**

Voire le fichier joint nommé « Kmeans.pdf ».

* **Utilisation de la bibliothèque Scikit-learn :**

-Calcule de la matrice de la distance avec métrique « Manthatan » :

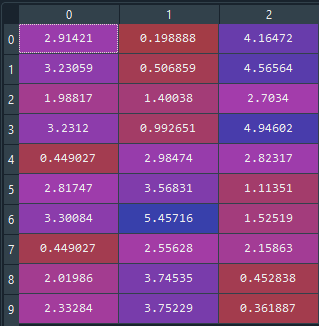


Figure 8 : Matrice des distances Kmeans

-**Déduction du meilleur choix pour le paramètre k :**

-**Méthode Silhouette :**

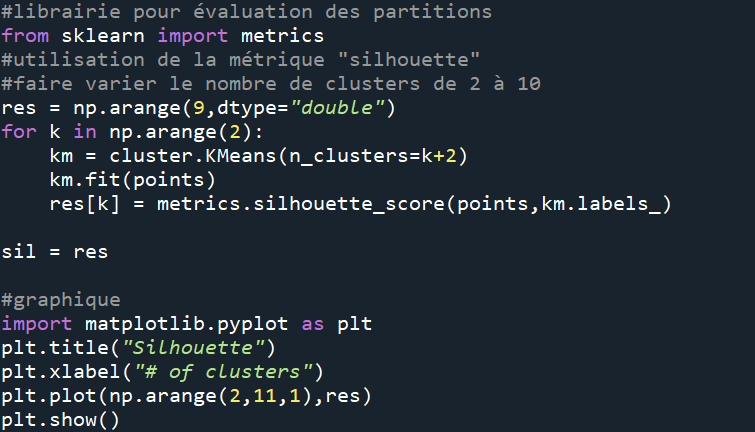


Figure 9 : Code python méthode Silhoutte

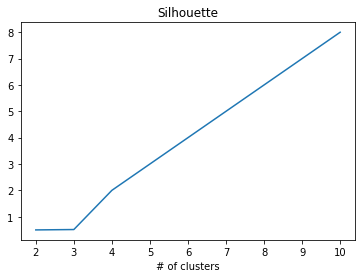


Figure 10 : Graphe Silhouette

**-méthode ELBOW :**

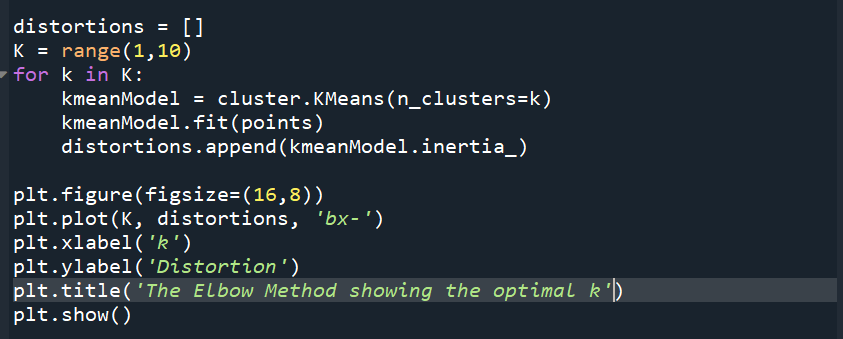


Figure 11 : Code python méthode ELBOW

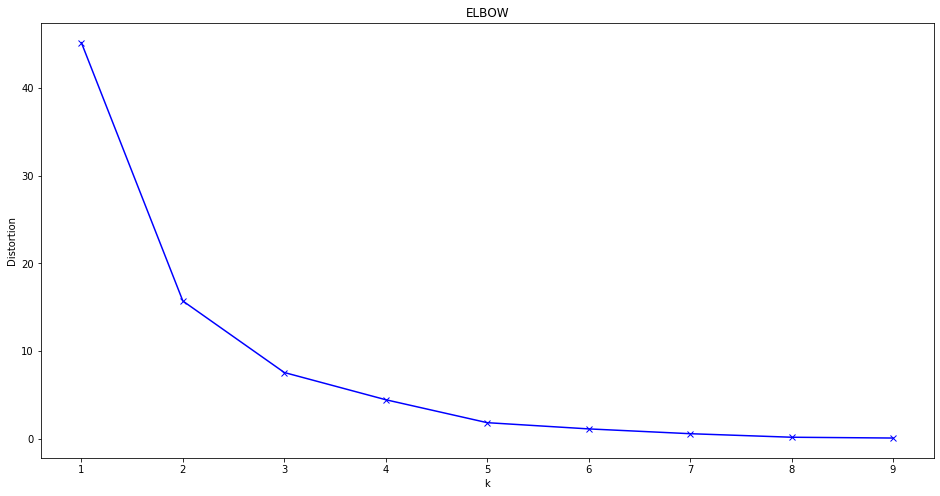


Figure 12 : Graphe ELBOW

En utilisant la méthode ELBOW on déduit que le meilleur paramètre K est 5.

En utilisant la méthode Silhouette on déduit que le meilleur paramètre K est 4.

* **Vérification des résultats**

L’application de l’algorithme Kmeans sur le papier et le code on donnée le même résultat.

# Partie 2 :

1. **Jeu de donnée :**

Nous avons effectué une mise à jour sur le dataset que vous nous avez envoyé a partir du site officiel du ministère de la santé algérienne « <http://www.sante.gov.dz/> » de la date du 29/06/2020.

1. **Utilisation de la bibliothèque Scikit-learn :**

-Lecture du dataset :

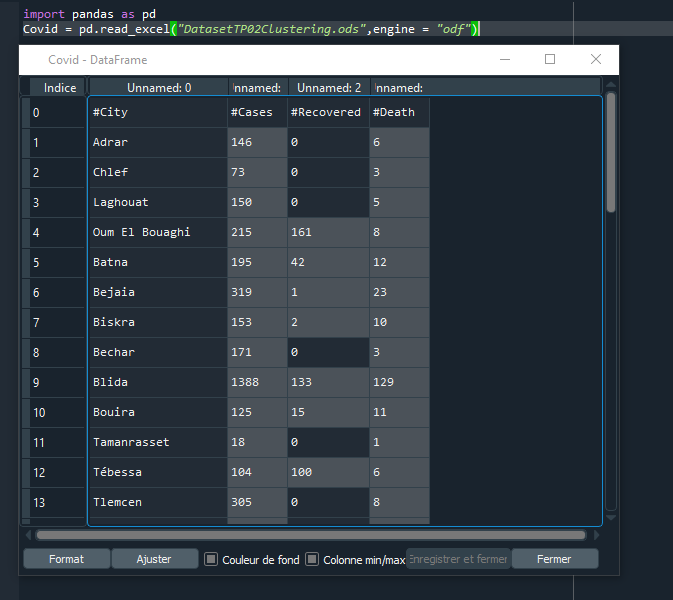


Figure 13 : Lecture du dataset

**-Distribution de données CHA :**

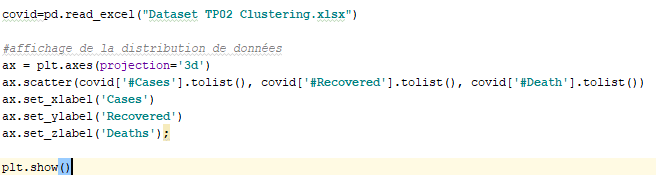


Figure 14 : Code python distribution CHA

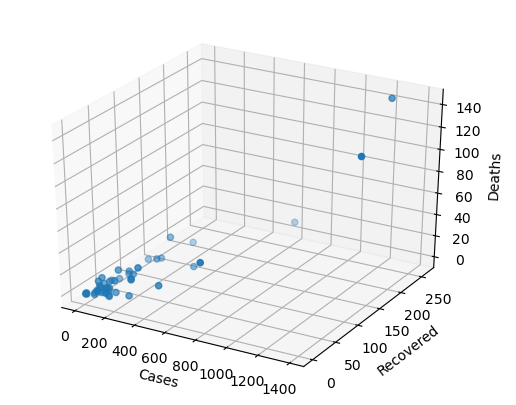


Figure 14 : Graphe de distribution CHA

-Dendrogramme :

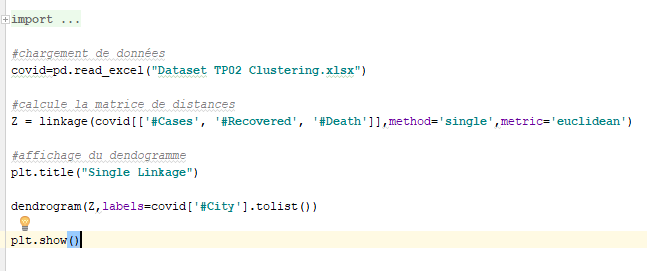


Figure 15 : Code python Single Linkage

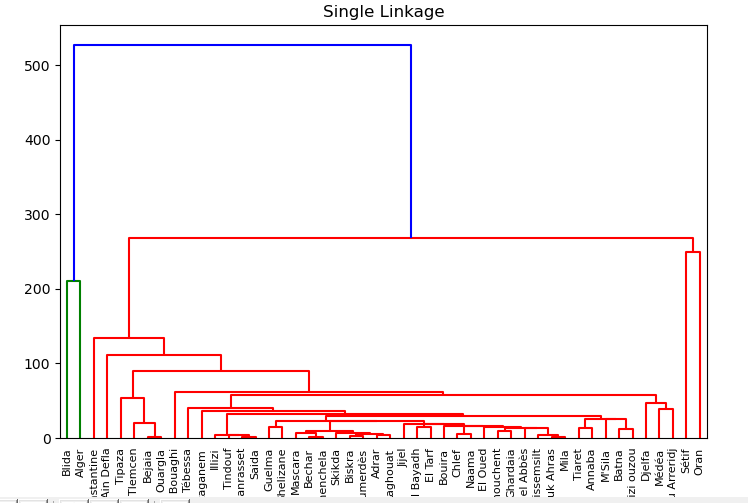
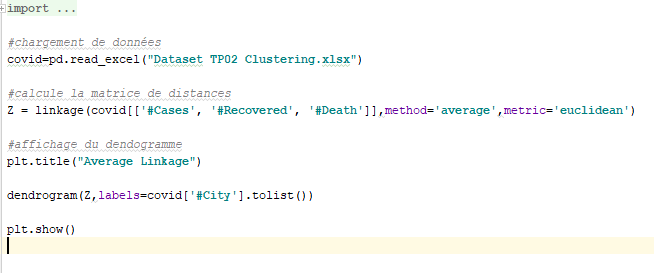


Figure 16 : Dendrogramme Single Linkage

Figure 18 : Code python Average Linkage

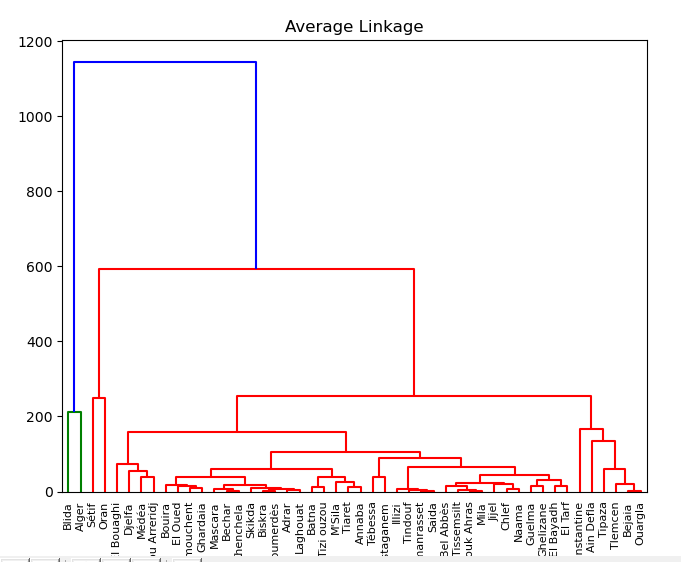


Figure 17 : Dendrogramme Average Linkage

Figure 18 : Code python Complete Linkage

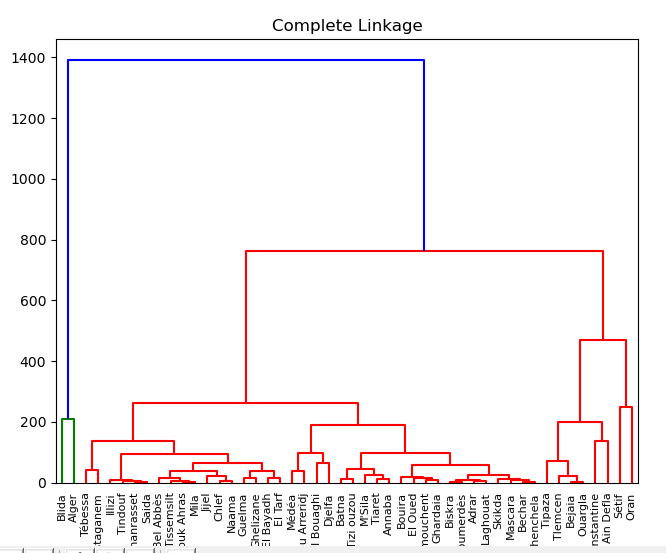


Figure 19 : Dendrogramme Complete Linkage

**-Meilleur Coupe :**

Dans le cas de Single Linkage la meilleure coupe est 260.

Dans le cas d’average Linkage la meilleure coupe est 300.

Dans le cas de Complete linkage la meilleure coupe est 500.

-**Application du Kmeans :** 

Figure 20 : Code python matrice de distance

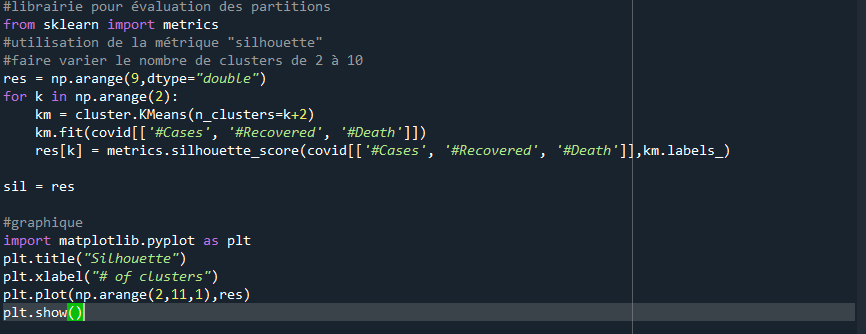


Figure 21 : Code python Silhouette

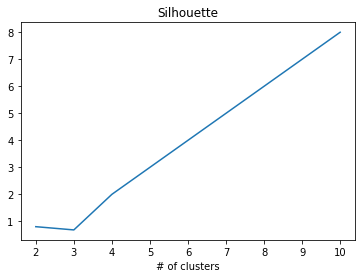


Figure 22 : Graphe Silhouette

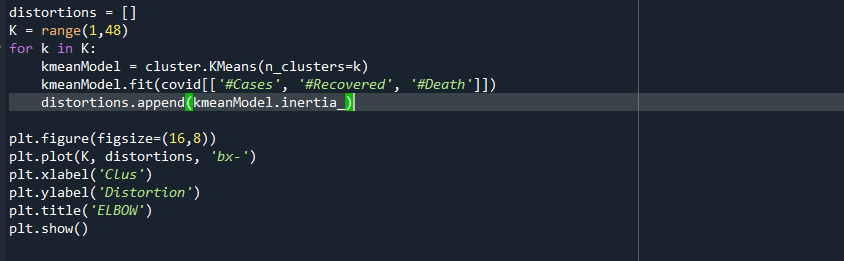


Figure 23 : Code python Silhoutte

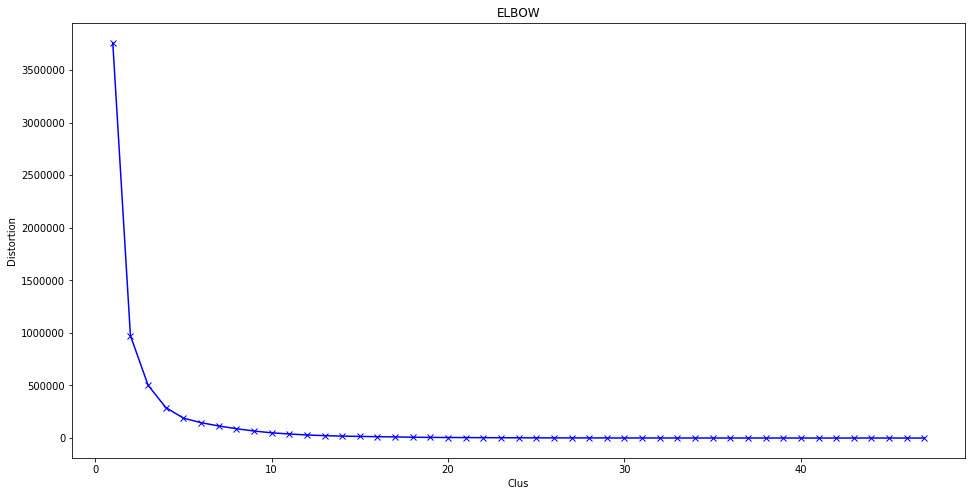


Figure 24 : Graphe ELBOW

**-Interface Graphique :**

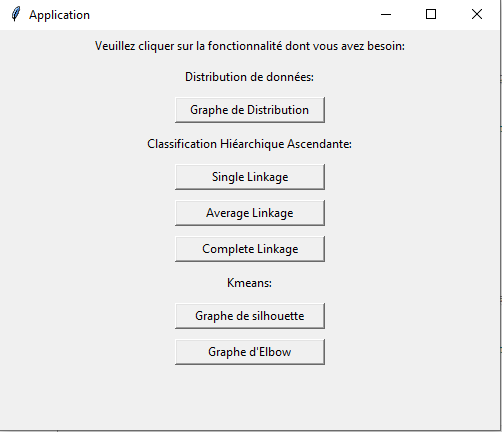


Figure 25 : Interface graphique sou python

**Labialisation du regroupement :**

Groupe 1 labialisé « Critique »  :

Blida , Alger

Groupe 2 labialisé « Modéré  »   :

Oran, Sétif , Ain Defla , Constantine , Ouargla , Bejaia , Telecmann

Groupe 3  labialisé « Contrôlé  »   : le reste des willayas