**Title:** K-Means clustering using python

**Aim:** To implement K-Means clustering using sci-kit learn library.

**Tools:** Anaconda, Python 3.7, Jupiter Notebook

**Theory:**

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. K-Means Clustering groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process. It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The k-means clustering algorithm mainly performs two tasks:

* Determines the best value for K center points or centroids by an iterative process.
* Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Algorithm:

* Select the number K to decide the number of clusters.
* Select random K points or centroids. (It can be other from the input dataset).
* Assign each data point to their closest centroid, which will form the predefined K clusters.
* Calculate the variance and place a new centroid of each cluster.
* Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
* If any reassignment occurs, then go to step-4 else go to FINISH.
* The model is ready.

The performance of the K-means clustering algorithm depends upon highly efficient clusters that it forms. There are different ways to find the optimal number of clusters, elbow method is one of them.

The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:



To find the optimal value of clusters, the elbow method follows the below steps:

* It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).
* For each value of K, calculates the WCSS value.
* Plots a curve between calculated WCSS values and the number of clusters K.
* The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

**Assignment:**

1. The dataset contains the records of different countries that includes different numbers related to child mortality, exports, health, imports, income, etc. Let the United Nations want to design development plan for the different groups of countries. The groups to be formed based on the economic wellbeing of a country. Write a program in python to train a model using K-Means clustering to group the countries based on the dataset available.

**Conclusion:**

The elbow methods shows that K=3 is the optimum value of K. The output image shows the three different clusters, of countries based on the data, with different colors.