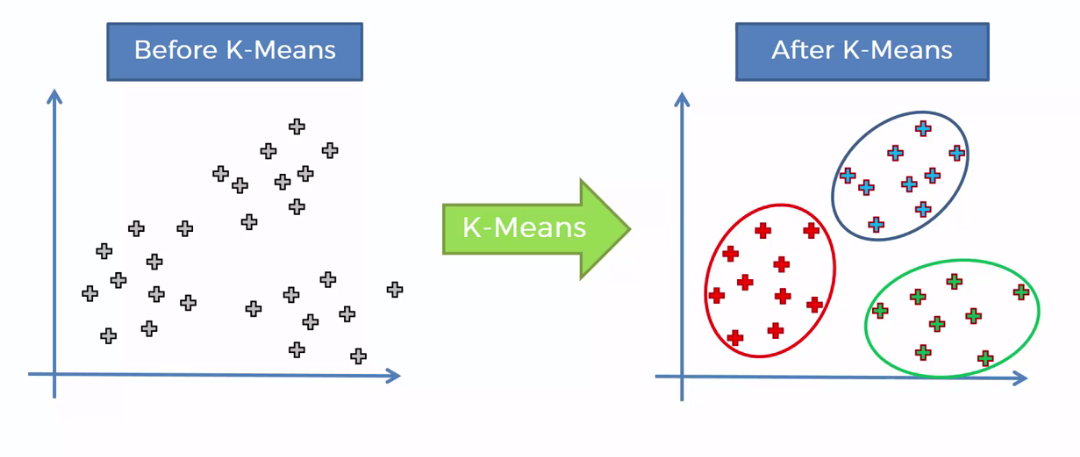
**NAME OF EXPERIMENT**: Implementation of K-means algorithm for clustering in Python.

**INTRODUCTION**: The k-means algorithm is a popular unsupervised machine learning technique used for clustering data points into groups. The "k" in k-means refers to the number of clusters that the algorithm is instructed to form. It is a simple and most popular for cluster analysis. It aims to partition ‘n’ observations into ‘k’ clusters. Each cluster in the k-means clustering algorithm id represented by a centroid point. Centroid point is the average of all the points in the set.

The main idea of the k-means algorithm is to find k-centroid points and every point in the dataset will belong either of k-sets having minimum Euclidean distance.



K-Means Algorithm:

Input:

k = the number of clusters,

D = dataset containing n objects

Output:

A set of k clusters

Method:

1. Randomly select ‘k’ from D as the initial cluster center.
2. Calculate the distance between each datapoint and cluster centers.

Euclidean distance =

1. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
2. Update the cluster means (i.e., calculate the mean value of the objects for each cluster)

MANUAL CALCULATION:

Cluster the following instance of given data with the help of k-means algorithm.

|  |  |  |
| --- | --- | --- |
| S.N. | X | Y |
| 1 | 185 | 72 |
| 2 | 170 | 56 |
| 3 | 168 | 60 |
| 4 | 179 | 68 |

Take K = 2

🡪 Solution,

Step1: Assume two points (k1, k2) randomly as cluster center.

K1 = (185, 72) and K2 = (170, 56)

Euclidean distance for Row 3 (K1, 3)

K1= = = 20.8

Euclidean distance for row 3 (K2, 3) =

= = 4.47

Here, K2 < K1

So the data point 3 belongs to K2.

The resulting cluster is;

K1 = {Row 1}

K2 = {Row2, Row3}

Calculate new dataset for new centroid,

K1 = (185, 72) and K2 = (170+168/2, 56+60/2) = (169, 58)

Step2:

Euclidean distance for Row 4

K1= = = 7.21

K2= = = 14.14

Here, K1< K2

So data point 4 belongs to K1.

The resulting Cluster is;

K1 = {Row 1, Row 4}, K2 = {Row2, Row3}

Calculate new dataset for new centroid,

K1 =(185+179/2, 72+68/2) =(182, 70) and K2 = (169, 58)

The cluster of data points obtained in second iteration is same as the third .So terminate the third one.

Therefore, final cluster = K1 {(185, 72), (179, 68)} and K2 {(170, 56), (168, 60)}.

ADVANTAGES OF K-MEANS CLUSTERING ALGORITHM:

1. It is very simple to implement.
2. It is scalable to a huge dataset and faster to large datasets.
3. It adapts the new examples very frequents.

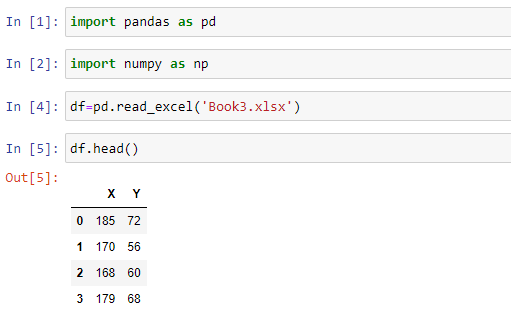
DISADVANTAGES OF K-MEANS CLUSTERINGALGORITHM:

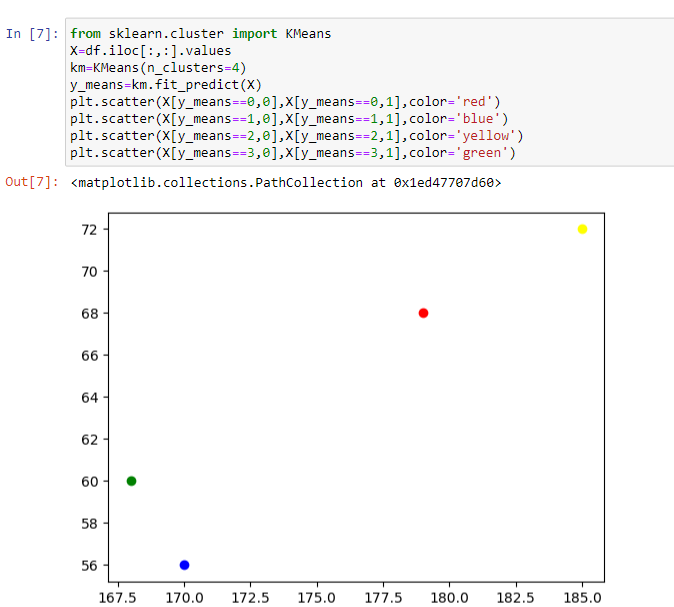
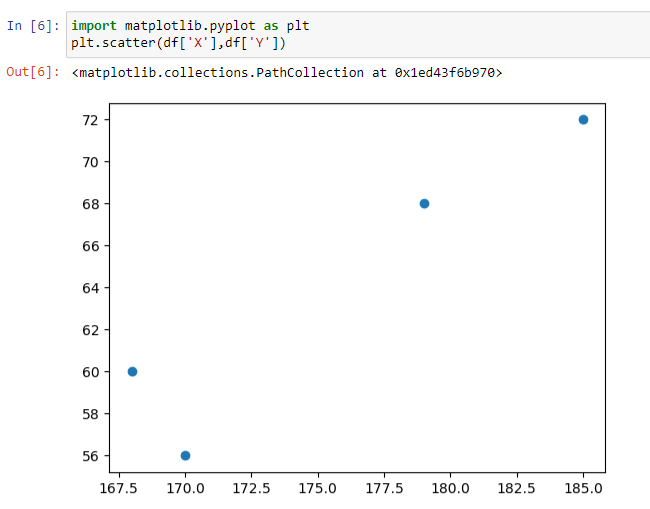
1. Dependent on initial values.
2. It is sensitive to the outliers.
3. Choosing the k values manually is tough job.
4. As the number of dimensions increases its scalability decreases.

PROGRAM IMPLEMENTATION IN PYTHON:

Requirement = Anaconda Navigator

Source Code:





**CONCLUSION:**

Hence, Successful Implementation of k-means clusteringalgorithm using Python.