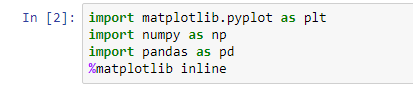
**CS412 – HW1: K-means**

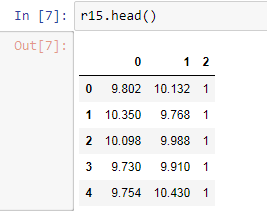
Import the essential libraries:

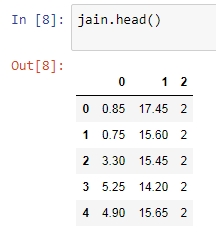


Load the datasets R15 and Jain:

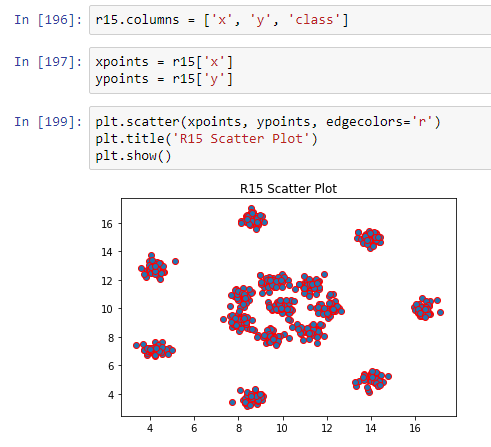


Results:

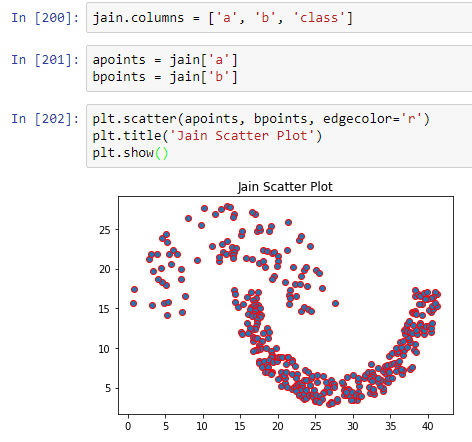




Plot the R15.txt data set:

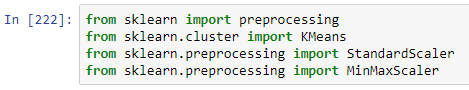


Plot the Jain.txt data set:



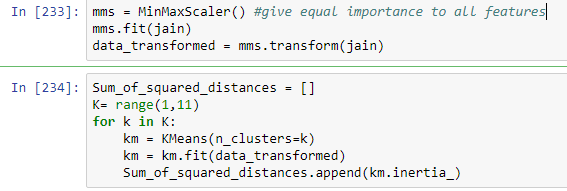
**Applying the K Means Algorithm:**

First, import the essential libraries:



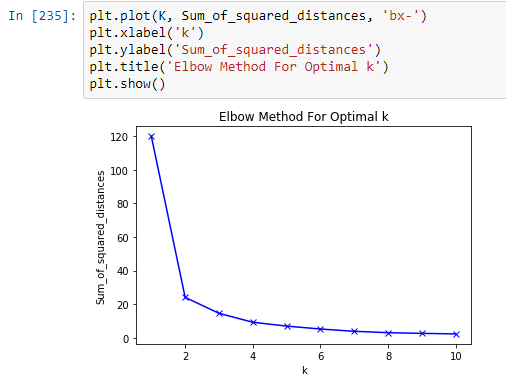
We also need to find the optimal k (cluster number) to succesfully implement the k-means algorithm in data sets. To find the optimal k:

For each value of k, initialize k-means and also use inertia\_ for calculating the SUM OF SQUARED DISTANCES of samples to the centroids of clusters.



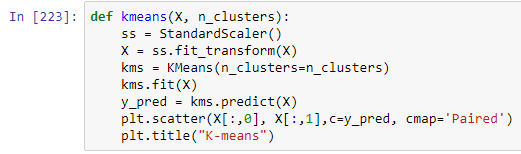
**As number of clusters: k increases, sum of squared distance closes to 0.**

Use **elbow method** to find the optimal k: Ploting the sum of squared distances for each k in the specified range:

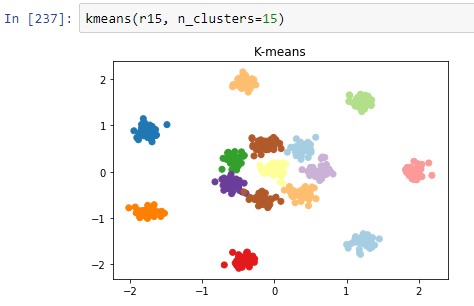


From the elbow method plot we can easily see that the **optimal number of clusters for the data set Jain is 2.**

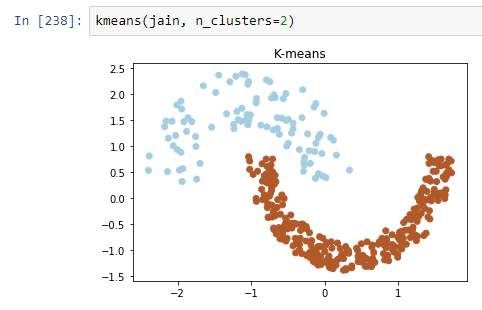
Now implement a k-means function to use with data sets and proper k:



Clustering result for R15.txt with **optimal cluster number as 15**:

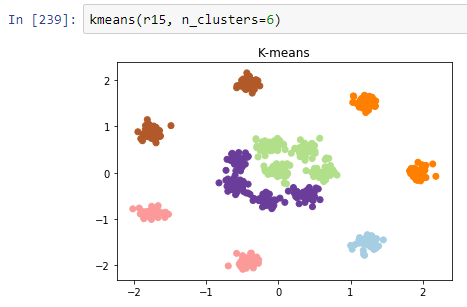


Clustering result for Jain.txt with **optimal cluster number as 2**:



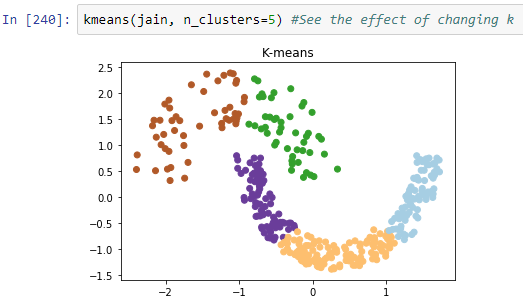
As i explained earlier, to see the effect of changing the k (cluster number) with respect to SSE:

Decrease the number of clusters for R15.txt to 6 and the result will be:



As k decreased, it increased the sum of squared distances between centroids of clusters and samples and we now have a non-optimal result with increased SSE.

Similarly, increase the number of clusters for Jain.txt to 5 and the result will be:



As k increased, it decreased the sum of squared distances between centroids of clusters and samples but we still have a non-optimal result because partitioning the optimal clustering did not add anything to the analysis, it only increased the cluster number. It can be easily seen when compared the first k-means and the second k-means results of Jain.txt. The reason is that the optimal cluster number for Jain must be 2.