ASSIGNMENT – 11

TOPIC: K-MEANS

Q.1. To study and prepare notes for k-means++ algorithm.

Q.2. Find out other methods we can use to choose the number of clusters for the k-means.

Q.3. Try to use k-means algorithm with different combinations of the columns for the datasets.

Ans. 1) Drawback of standard K-means algorithm:

One disadvantage of the K-means algorithm is that it is sensitive to the initialization of the centroids or the mean points. So, if a centroid is initialized to be a "far-off" point, it might just end up with no points associated with it and at the same time more than one clusters might end up linked with a single centroid. Similarly, more than one centroids might be initialized into the same cluster resulting in poor clustering. For example, consider the images shown below.

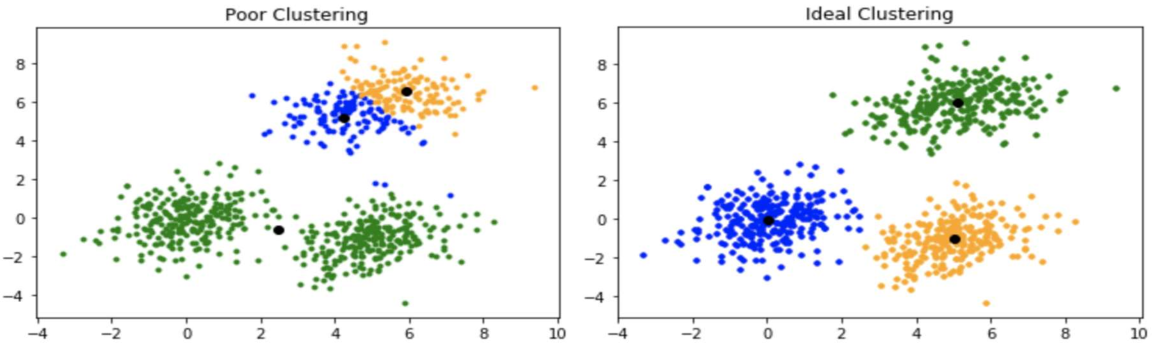


Fig (a) **Fig (b)**

Fig (a): Shows a poor initialisation of centroids resulted in poor clustering.

Fig (b): This is how the clustering should have been and ideal clustering:

K-mean++:

To overcome the above-mentioned drawback we use K-means++. This algorithm ensures a smarter initialization of the centroids and improves the quality of the clustering. Apart from initialization, the rest of the algorithm is the same as the standard K-means algorithm. That is K-means++ is the standard K-means algorithm coupled with a smarter initialization of the centroids.

Intuition:

By following the above procedure for initialization, we pick up centroids which are far away from one another. This increases the chances of initially picking up centroids that lie in different clusters. Also, since centroids are picked up from the data points, each centroid has some data points associated with it at the end.

Note: Although the initialization in K-means++ is computationally more expensive than the standard K-means algorithm, the run-time for convergence to optimum is drastically reduced for K-means++. This is because the centroids that are initially chosen are likely to lie in different clusters already.

Ans. 2) Different methods we can use to choose the number of clusters for the k-means:

Elbow Method:

Elbow method gives us an idea on what a good *k* number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. We pick *k* at the spot where SSE starts to flatten out and forming an elbow.

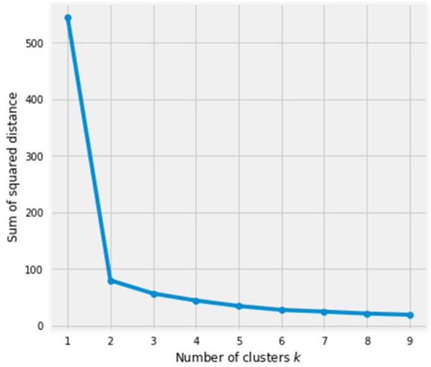


Fig (a) – Elbow at 2

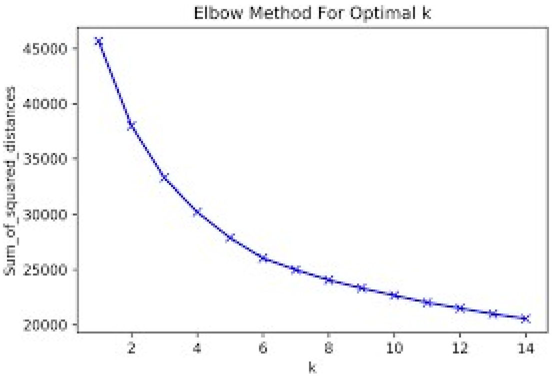


Fig (b) – No elbow

The graph, Fig (a) above shows that k=2 is not a bad choice. Sometimes it's still hard to figure out a good number of clusters to use because the curve is monotonically decreasing and may not show any elbow (as in case of Fig b) or has an obvious point where the curve starts flattening out.

Silhouette Analysis:

Silhouette analysis can be used to determine the degree of separation between clusters. For each sample:

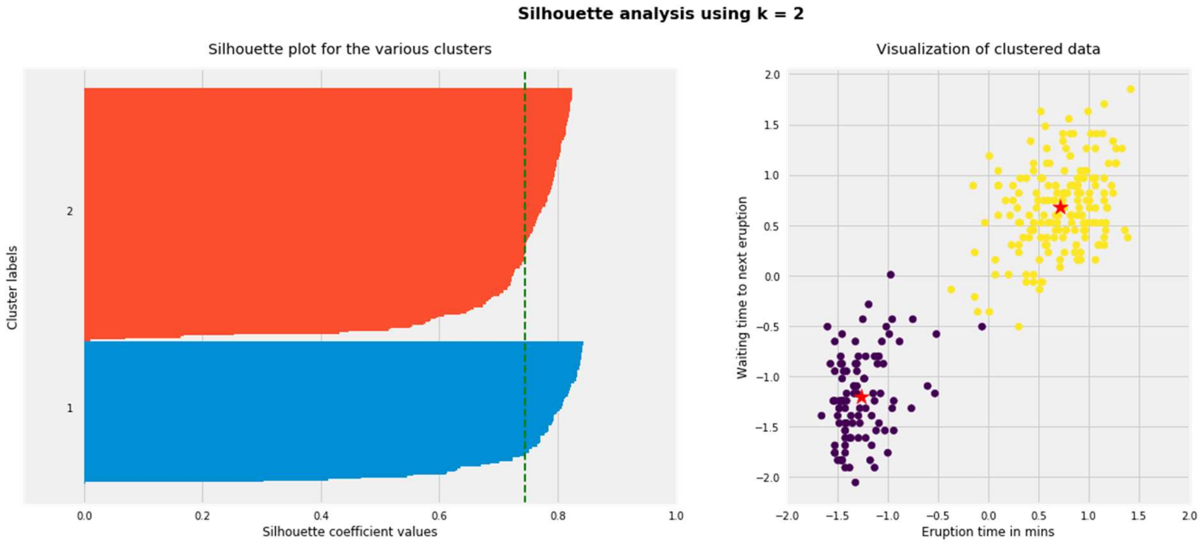
- Compute the average distance from all data points in the same cluster (ai).
- Compute the average distance from all data points in the closest cluster (bi).
- Compute the coefficient:

$$\frac{b^i - a^i}{max(a^i, b^i)}$$

The coefficient can take values in the interval [-1, 1].

- If it is 0 –> the sample is very close to the neighboring clusters.
- It it is 1 –> the sample is far away from the neighboring clusters.
- It it is -1 –> the sample is assigned to the wrong clusters.

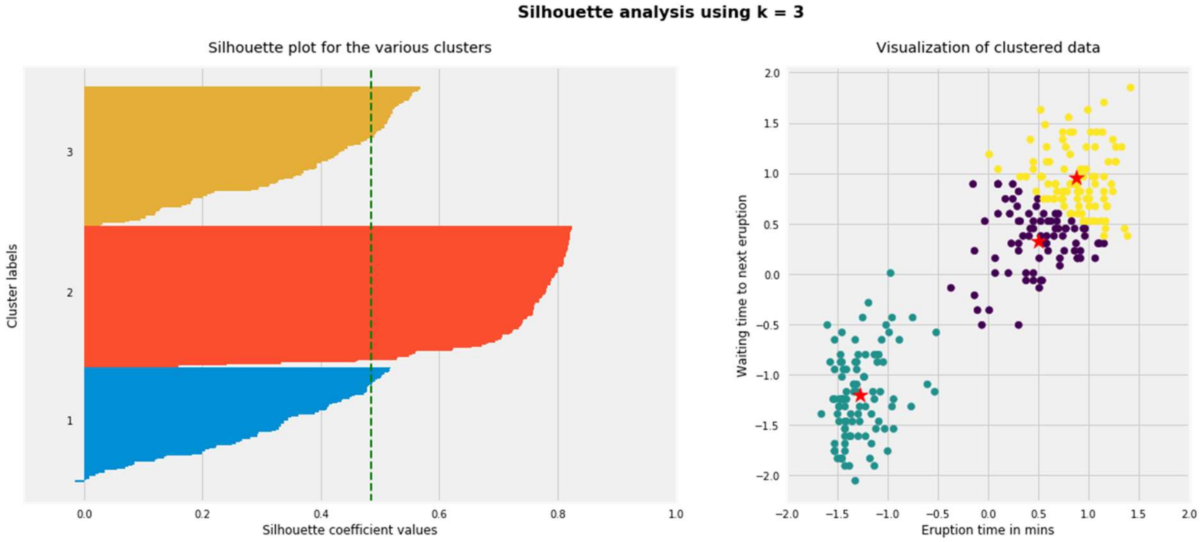
Therefore, we want the coefficients to be as big as possible and close to 1 to have a good clusters.



Silhouette analysis using k = 2

Silhouette plot for the various clusters

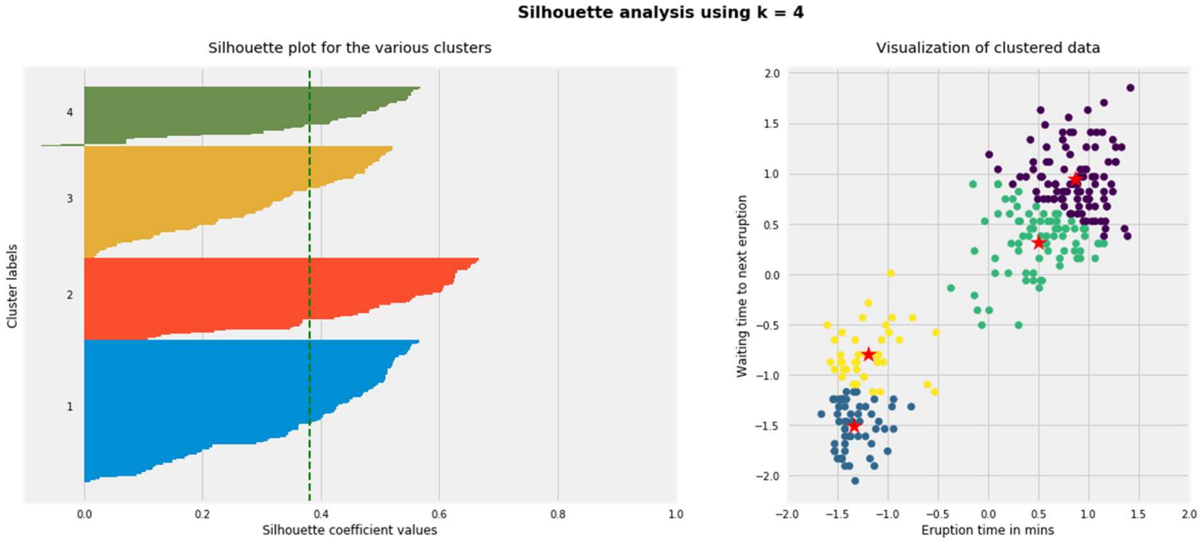
Visualization of clustered data



Silhouette analysis using k = 3

Silhouette plot for the various clusters

Visualization of clustered data



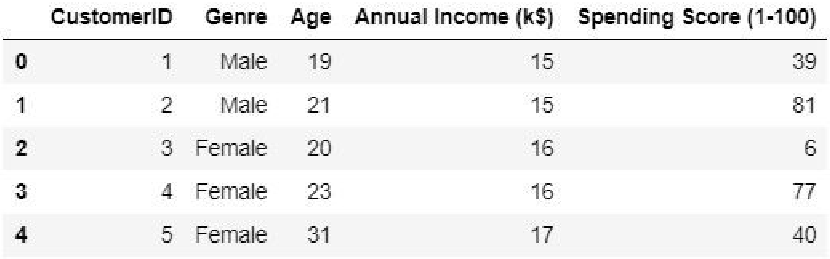
Silhouette analysis using k = 4

Silhouette plot for the various clusters

Visualization of clustered data

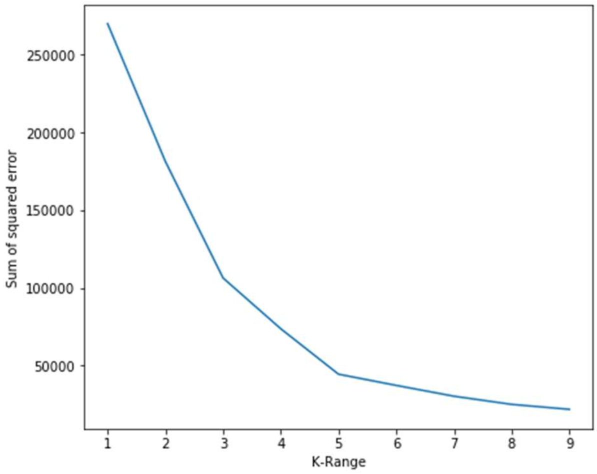
As the above plots show, n_clusters=2 has the best average silhouette score of around 0.75 and all clusters being above the average shows that it is actually a good choice. Also, the thickness of the silhouette plot gives an indication of how big each cluster is. The plot shows that cluster 1 has almost double the samples than cluster 2. However, as we increased n_clusters to 3 and 4, the average silhouette score decreased dramatically to around 0.48 and 0.39 respectively. Moreover, the thickness of silhouette plot started showing wide fluctuations. The bottom line is: Good n_clusters will have a well above 0.5 silhouette average score as well as all of the clusters have higher than the average score.

Ans. 3) Given dataset: (Represnted only 5 rows)

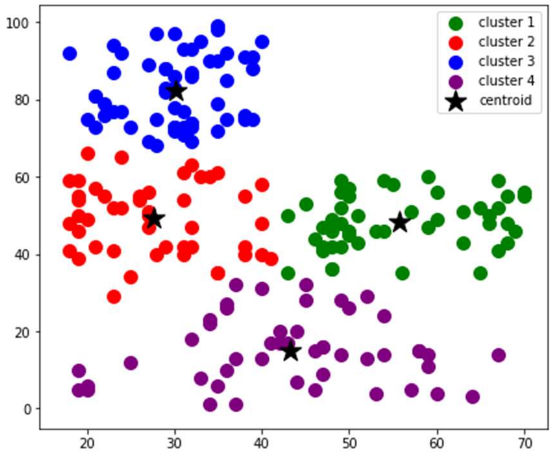


	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Case 1: Considering columns Annual Income (k$) and Spending Score (1-100)



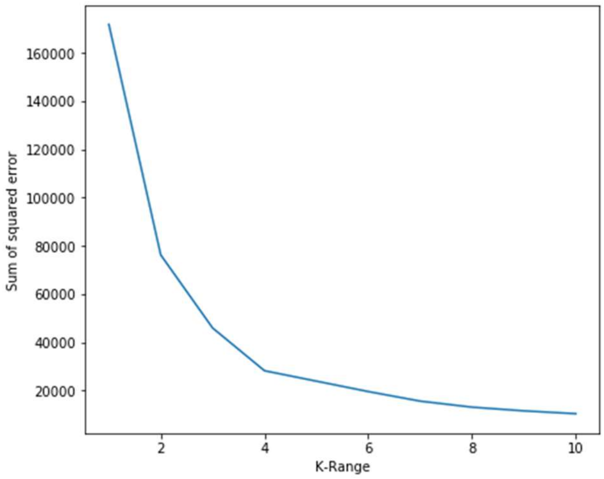
Fig(a)



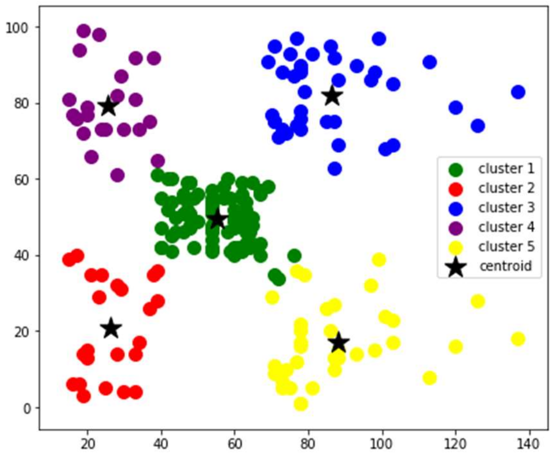
Fig(b)

Here, in Fig (a), we see that after using elbow method k = 5 is the best value for number of clusters and then using n_cluster = 5 we plot a graph with 5 clusters with their respective centroid as shown in the Fig (b)

Case 2: Considering columns Age and Spending Score (1-100)



Fig(c)



Fig(d)

Here, in Fig (c), we see that after using elbow method k = 4 is the best value for number of clusters and then using n_cluster = 4 we plot a graph with 4 clusters with their respective centroid as shown in the Fig (d)