**Subjective Question**

1. Compare and contrast K-means Clustering and Hierarchical Clustering.

* K-means clustering uses pre-specified no. of clusters when compared to hierarchical clustering where k value is not pre-determined; we can cut the dendrogram and take number of k.
* Hierarchical clustering is a time consuming process when compared to K-means as each data points are considered with every other cluster.
* The results produced by running the algorithm multiple times might differ in k-means. While results are reproducible in Hierarchical clustering.

2. Briefly explain the steps of the K-means clustering algorithm.

K-means clustering tries to group similar kinds of items in the form of clusters. It finds the similarity between the data points and groups them into the clusters. K-means clustering algorithm works in three steps..

1. Select the k values.
2. Randomly pick points in space and call them cluster centers.
3. Assign each observation in the data set to cluster that minimizes Euclidian distance.
4. Recompute the centre of each cluster by taking mean of each observation in each cluster.
5. Repeat steps 2 & 3 until either the centers don’t change or your maximum number of iterations has been reached.

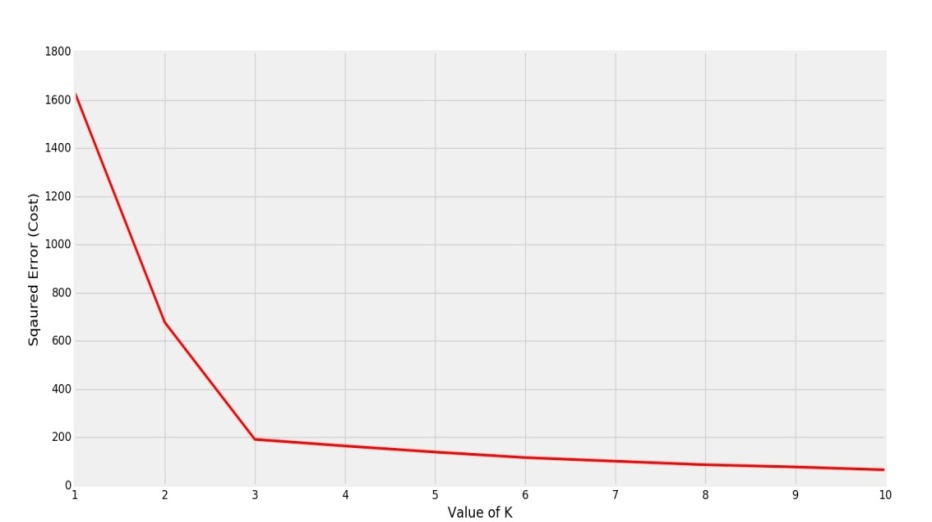
3**.** How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Statistical Analyses

Value of K can be chosen using elbow curve and silhouette score.

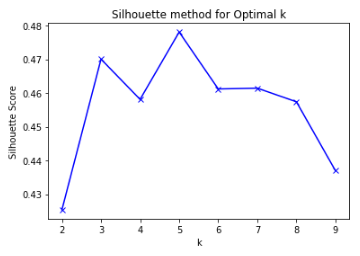
* The elbow curve method uses the sum of square distances method to find the most optimal clusters at the point where there is not much change in the inertia, i.e. where distinct elbows are being formed.

The basic idea behind this method is that it plots the various values of cost with changing *k*. As the value of *K* increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point.



Clearly the elbow is forming at K=3. So the optimal value will be 3 for performing K-Means.

* The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.



k=5 should be chosen for the number of clusters.

Business Aspect

This is mostly based for situations where the business understanding of the problem dictates as to how many clusters need to be created.

4. Explain the necessity for scaling/standardization before performing Clustering.

Standardization is an important step of Data preprocessingasit controls the variability of the dataset, it converts dataset to a comparable scale before applying clustering which generate good quality clusters and improve the accuracy of clustering algorithms, check out the link below to view its effects on k-means analysis.

5. Explain the different linkages used in Hierarchical Clustering.

* Single linkage- here distance between two clusters are defined as the shortest distance between points in the two clusters.
* Complete linkage **-**here distance between two clusters are defined as the maximum distance between any two points in the cluster.
* Average linkage- here distance between two clusters are defined as the average distance between every point of one cluster to every other point of the cluster.