Assignment: Part II

Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly(what EDA you performed, which type of Clustering produced a better result and so on).

Answer:

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer:

a)

Description	K- Means Clustering	Hierarchial Clustering
Definition	K-means clustering works by grouping objects with similar features into groups, such that there are a few disctinct groups that can be clearly observed. Objects belonging to each cluster show distinct properties to that cluster only. There can be some overlap which can be resolved through cluster mapping.	Heirarchical clustering is method that works by grouping similar looking objects into a simiar cluster till all the objects are combined and form on cluster. The clusters are formed in tree like structures. This approach is also referred to as the "Bottom-up" approach. Another way the clustering works is by taking everything as one cluster and continuing to split it into individual clusters till all objects are individually represented. This approach is known as the "Top-down" approach
Cluster Calculation	We need to calculate and decide the number of clusters before hand. This is done by performing silhouette analysis and elbow analysis. Based on these analyses we identify the number of clusters that will give us the discernable clusters that we can use to group our data in.	In Heirarchial clustering we use a graphical representation called the Dendogram. Using the dendogram graph we determine the number of clusters as follows: 1. Determine the largest vertical distance that doesn't intersect any of the other clusters. 2. Draw a horizontal line at both extremities 3. The optimal number of clusters is equal to the number of vertical lines going through the horizontal line
Scalability	K-Means is highly scalable and works well with large data sets or big-data.	This is not very scalable as it becomes difficult to identify the number of clusters.
Efficiency	K-Means is a nonlinear time process and hence is more efficient compared to hierarchial clustering.	Heirarchial clustering is a linear time algorithm, therefore it can be slightly slow depending on the size of the data. This happens due to a large number of merging and splitting operations that take place.
Pros	• Relatively simple implentation • Scalable for large datasets • Converges well • Efficient compared to other clustering algorithms	• Easy to implement • No need to specify the number of clusters • Dendogram provides good visual understanding of the data and

		clusters
	• Easy to Interpret • Flexible to changes	• Works well with identifying outliers
Cons	• No-optimal cluster size. • Needs to know number of clusters/K-value before hand • Works only with numerical data • Doesnt work well with outliers • Suffers from curse of dimensionality	• Doesnt work well with large datasets • Time complexity for large or medium datasets means it takes a lot of time and is slower compared to K-means

b) Briefly explain the steps of the K-means clustering algorithm.

Answer: The K-Means Algorithm works as follows:
- **Step 1: Initialization of Clusters –** Randomly choose k datapoints as centroids
- **Step 2: Cluster Assignment –** Assign the neighboring datapoints to the closest cluster using Euclidean distance between them.
- **Step 3: Move the Centroid –** In the newly formed clusters calculate the new centroid by taking a mean of all the clusters
- **Step 4: Break when centroids stop moving –** Repeat step 2 & 3 till the centroids stop moving. Once they stop moving the algorithm has converged.

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Answer: The initial value of K is chosen at random for calculating the clusters initially. This is usually done statistically using either silhoutte analysis or elbow analysis.

Statistical Aspect: For choosing k using a statistical aspect we can use the elbow score and silhoutte analysis methods to pick an optimal value of K. This can be done by taking the point where there is a sudden drop in the curve in both the cases. Of course, taking a value of 2 as is not advised but might make sense. This can be verified considering the business aspect.

Business Aspect: The business aspect of choosing a K value could be driven by what goal the business is trying to achieve. The understanding of the dataset as well as the business problem can help aid in identifying a value of k.

d) Explain the necessity for scaling/standardisation before performing Clustering.

Answer: Scaling allows for features with different scales and units to come to a common base. Since k-means clustering focuses more on the direction of the distance between points, if one of the features has a higher scale and dimensionality, the entire result will be skewed towards that feature. Scaling will equalise the dimensions of all the features and help increase the performance of our model.

e) Explain the different linkages used in Hierarchical Clustering.

Answer: There are 3 kinds of linkages in Hierarchial Clustering, namely
- **Single Linkage:** Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters

- **Complete Linkage:** Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters
- **Average Linkage:** Here, the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.