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: The “Clustering of Countries” assignment was performed with the objective to identify top 5 countries that are in need of immediate aid.
We identified the countries through the following analysis steps:

1. **Step 1 – Data Inspection & Analysis:**

   We began by understanding the buiness problem provided. After which we proceeded towards importing and analysing the data. Our analysis consisted of the following steps:

   ➔ Importing the data set & checking that it has been correctly imported using head()

   ➔ Checking basic parameters using describe(), shape, info()

   ➔ Checking for null and duplicate values

2. **Step 2 – Data Preperation & EDA**

   We then proceeded to prepare the data for EDA and Clustering. This was done as follows

   ➔ We first converted the *exports, health and imports* columns to their correct numerical values. These were given as % of gdpp. We are working with gdp values in the other columns, therefore these were correctly converted.

   ➔ We then indexed the data based on the *country* column since it is a non-numerical column and would later cause type mismatch issues.

   ➔ Performed outlier analysis on 0.01, 0.25, 0.5, 0.75, 0.90, 0.95 ,0.99 percentile values of the dataset to identify any extreme outliers. Plotted distplots, box plots & heatmaps to understand visually the outliers.

   ➔ We decided not to remove the outliers but to make them equal to the 99% value on the higer side and 1% value on the lower side. Removing the outliers would have resulted in data loss and essentially dropping countries that might actully need aid.

   ➔ We again plotted the distplot, box plot and heatmap. Visibly after the outlier treatment there was change in the values on the heatmap as well as the shape of the graphs.

   ➔ We then performed *Hopkins Analysis* on our dataset. Hopkins analysis allows us to identify the clustering tendency of our dataset.

   ➔ We then performed scaling of our dataset to prevent any highly scaled features to skew the data incorrectly in the favour of that feature.

- ➔ After scaling our dataset we decided to perform Elbow and Silhoutte analysis to decide the number of clusters we should take. Both these analysis resulted with a cluster number of 3.

3. **Step 3 – K-Means Clustering & Visulaization**

   - ➔ Using 3 as our final number of clusters, from the analysis carried out above, we proceeded to perform K-means clustering.
   - ➔ We set the number of clusters to 3 and performed k-means clustering on our scaled dataset. This resulted in a number of cluster ids for every country. The resultant cluster ids were added to a new copy of the dataset without outliers, under the label *cluster_id* .
   - ➔ Then, we grouped the data using the cluster ids and plotted scatter plots for the 3 features, *child_mort, gdpp & income*, between them to see how the clusters lined up.
   - ➔ It was observed that cluster 2 consistently had lower income & gdpp countries. The child mortality rate was also high in these countries.
   - ➔ To confirm our analysis so far, we performed some profiling and found that our observations were correct. We identified 5 countries with the lowest income & gdpp but highest child_mortality rate.

4. **Step 4 – Hierarchial Clustering**

   - ➔ We performed single linkage & complete linkage clustering on the sacaled dataset to identify the number of clusters.
   - ➔ Using complete linkage and the cluster value of 2, we created labels for the linkage based clusters and again plotted scatter plots for the same.
   - ➔ We observed that the countries in cluster 0 for linkage had similar behaviour as countries in cluster 2 for k-means.
   - ➔ We profiled the data and found that the top 5 countries in cluster 0 were the same as that in K-means.

5. **Step 5 – Conclusion**

   - ➔ We concluded that The HELP International NGO should focus on the 5 countries outlined below for immediate aid: Burindi, Liberia, Congo, Dem. Rep., Niger and Sierra Leone.

Question 2: Clustering

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

Answer:

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

	• Easy to Interpret • Flexible to changes	clusters • 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.