Aravapalli Pushkara

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)

Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Steps followed:

1. First the data was cleaned for finding missing values, duplicates – the data did not have any missing values. Typing mistakes or duplicate values.
2. An outlier analysis was done and it resulted in gdpp, income and inflation to have high outliers
3. But, the variable inflation might be deter mental for finding the countries with that were economically backward and needed the funds.
4. Next the variable gdpp was visualized through pair plots
5. From the correlation heat map – we understood that there is significantly high correlation between the following
    a. gdpp and income
    b. imports and exports
    c. total_fer and child_mort
6. To handel the above scaling of the data was performed, if feature scaling was not performed then the values of units could get overrated or underrated.
7. Next, we anlayze for PCA (Principle Component Analysis), applying it on country_norm gives a considerable reduction in dimensionality.
8. Next we plot the cumulative plt to understand the number of variables causing high amount of variance.
9. We then perform dimensionality reduction using incremental PCA.
10. We then check of possible correlations after the PCA which we see have now become closer to ‘0’.
11. We then transpose the data and further analyse.
12. Now, to understand the clustering tendency we use Hopkins statistic which turned out to be greater than 0.7 which means that the data has a good tendency to form clusters.
13. To understand how many clusters need to be taken into consideration, we perform the Silhouette analysis where we found from the sum of squared graph and elbow graph that k=4 can be good enough for further analysis.
14. Scatterplots with PC components and cluster IDs is formed.
15. We get the list of counties that require the fund and which are in direst of needs.
16. Hierarchal clustering (Unsupervised) is now done to group together unlabelled data with similar characteristics.
17. The dendrogram Is cut at K=4
18. The aforementioned clusters are now analysed and subplots created.

19. We ignore the cluster 3 with only one country, from this we understand that cluster 0 could be one detrimental factor and hence bin the cluster.

As per Hierarchical Clustering the countries which require aid are:

- Burundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone
- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Togo

There are similar countries by both K-means and Hierarchical Clustering - Hence concluding that the following are the countries which are in direst need of aid by considering socio – economic factor into consideration:

- Burundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone
- Madagascar
- Mozambique
- Central African Republic
- Malawi

Question 2: Clustering

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

1. K-means clustering can handle large amount of data unlike the Hierarchical Clustering.
2. Time complexity of K-means is linear while Hierarchical Clustering is quadratic.
3. K-means algorithms are parameterized by k values, i.e the number of clusters we would want to create. It starts by creating k number of centroids. However, in Hierarchical Clustering we start with data points having its own clusters.
4. K-means, cluster choice is picked randomly and different results are obtained upon various iterations. Hierarchical Clustering - builds clusters incrementally.
5. K-means needs prior knowledge of K (number of chunks we would want our data to be divided) whereas Hierarchical Clustering is about analysing the dendrograms.

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

1. Kmeans algorithm is a recursive algorithm which partitions the df into K number of pre-defined varied and non-overlapping subgroups or clusters.
2. Here, each data point belongs to only a single cluster.
3. This method makes the intra-cluster data points as same as possible along with keeping the clusters as different or far off as possible.
4. It also assigns data points to certain clusters in a way that the sum of the squared distance between the data points and that of the cluster's centroid (the arithmetic average/mean of all the data points that belong to that cluster) is at the least.
5. The least variation we have within the clusters, the more the homogeneity or the similar the data points are within the same cluster.
6. The way k-means algorithm works is as follows by the steps below:

    1. Specify number of clusters K.
    2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
    3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
    4. Compute the sum of the squared distance between data points and all centroids.
    5. Assign each data point to the closest cluster (centroid).
    6. Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.
    7. The approach kmeans follows to solve the problem is called Expectation-Maximization. The E-step is assigning the data points to the closest cluster.

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

1. Determining the ideal number of groups in an informational index is an essential issue in apportioning bunching, for example, k-implies bunching, which requires the client to indicate the quantity of bunches k to be created.
2. Unfortunately, there is no conclusive response to this action. The ideal number of groups is by one way or another abstract and relies upon the strategy utilized for estimating similitudes and the outliers utilized for analysis.
3. A straightforward and well known arrangement comprises of reviewing the dendrogram delivered utilizing various levelled bunching to check whether it proposes a specific number of groups.

4. These strategies incorporate direct techniques and measurable testing strategies.
5. Statistical testing strategies comprises of looking at results against invalid observations.
    1. **Statistical context:** K-means gives an iterative method to give random initializations at the start of the algorithm. A different approach to initializations can lead to give different number of clusters. It is recommended to run the k-means algorithm with different initializations of the centroids and get to a result which gives the least sum of squared distances.
    2. **Business context**: The number of clusters can be considered by the business problem for example for an ecommerce industry, we can cluster the high value customers and low value customers so that they can be mapped to respective ecommerce campaigns to gain return. For every new customer that gets added this benchmarking can help account retain profitable customers. In this context we can assume k=3 for forming different clusters.

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

1. Before we start with the clustering analysis the clusters tend to change. If we have more equal scales then the percentage of variables significantly contribute towards more of defining the clusters.
2. Especially, if they variables are incomparable units such as age and height then we must standardize the variables, however even if the variables are comparable it is a good practice to scale the data.
3. If the scaling is not performed then there will be uneven variance and the weightage will be put on smaller clusters, and hence clusters would separate along variables that have greater variance.
4. K-means is sensitive to order of order of elements in the df, hence the analysis can be performed multiple times to randomize the elements.
5. Hence standardization helps variables with large clusters and prevent them from dominating the way clusters are defined in the process.

e) Explain the different linkages used in Hierarchical Clustering.

1. This method treats every data point as a single cluster and recursively merges all the clusters until all the data points are merged into one single cluster.
2. This is often represented as a dendrogram.
3. **For complete linkage**-for every set of cluster the smaller one is considered and merged along the diameter.
4. **For single linkage**-for every set of two clusters, one of the closest members with smaller distance is merged i.e. minimum pairwise distance.
5. Single linkage does not always give clear dendrograms and often complex, for visualization purpose a complete linkage gives a better picture to choose the value of K, however choosing the number of clusters depends on the business problem.
6. There are different methods to it,
    a. **Divisive method:** All the data point observations begin with a single cluster, where it is a top – down approach. The tree is split as we move down the hierarchy.

b. **Agglomerative method:** At first each data point is considered to be a single cluster. After every iteration similar clusters merge with other clusters until k clusters are formed.
c. **Single linkage:** This is based on grouping of the clusters in bottom up approach, combining two clusters which contain closest pair of elements not belonging to the same cluster as each other.
d. **Complete Linkage:** It is one of the methods of Agglomerative Hierarchical Clustering where all the clusters are eventually merged to form one whole cluster.
e. **Average linkage:** The distance between any two clusters is taken to be equal to the avg distance from any element of one cluster to another.