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:

Problem Statement: HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. After the recent funding programs, they have been able to raise around 10 million dollars. Now the CEO of the NGO needs to decide how to use this money strategically and effectively on backward countries.

Our main task is to cluster the countries by the factors mentioned above and then present solution and recommendations to the CEO

Let’s divide the entire process of clustering into steps:

Step 1: Reading and understanding data:

- In this step, we read the data and check the statistics of the data like info, describe and dtype methods using pandas library to get an understanding on the data.

Step 2: Exploratory Data analysis:

a) Data cleaning:

- In this step, I’ve converted the imports, exports and health columns to absolute values using GDP column as they were given as percentage of GDP per capita
- I’ve checked for null values and noticed that there are no null values in the data

b) Univariate Analysis:

- I have plotted distribution plot using sea born library for all the numerical columns in the data to understand how the data is varying in different ranges.

- Child mortality (Death of children under 5 years of age per 1000 live births) seems to be below 50 in most of the countries, only few countries child mortality is above 100.

c) Bivariate Analysis:

- I've have used three types of plots in Bivariate analysis: Pair plot, bar plot, heat map
- From these above plots, top 5 countries having lowest GDP are Burundi, Liberia, Congo, Dem. Rep., Niger and Sierra Leone.
- Top 5 countries having lowest net income per person are Congo, Dem. Rep., Liberia, Burundi Niger and Central African Republic
- Top 5 countries with highest child mortality are Haiti, Sierra Leone, Chad, Central African Republic and Mali.
- Top 5 countries with lowest life expectancy are Haiti, Lesotho, Central African Republic, Zambia and Malawi.
- From the heat map, I can infer that there is a high correlation between income and GDP, so, if average net income per person increases the GDP of the country will also increase.

Step 3: Outlier Treatment:

- To check the outliers, I've plotted boxplot or whisker plot and used hard capping to eliminate the outliers.
- Removing the lower range outliers for countries with low child mortality, low inflation and low total fertility.
- Removing the upper range outliers for countries with high income, GDP, imports, exports, life expectancy and health.
- Here, we are using soft capping method for removing these outliers with quantiles of 0.05 for lower range and 0.95 for higher range

Step 4: Scaling data:

- I've used standard scaler to normalize the data with mean 0 and standard deviation 1.
- After scaling we have converted back to data frame to prepare the data for clustering

- After converting to data frame I’ve checked for cluster tendency using Hopkin’s statistics and found that the value is promising to form clusters.

Step 5: Creating k-means clustering algorithm:

- To find the optimum value of number of clusters k, I’ve used Sum of squared distances vs k-value plot which is known as elbow curve method, silhouette analysis and found that optimum value is k = 3
- Assigned labels to main data and used scatter and box plot to check the clusters formed with gdpp, income and child_mort columns.

Step 6: Creating Hierarchical algorithm:

- I’ve created dendrogram using single and complete linkage and found that complete linkage seems to be promising in clusters formation and used cut tree method to form 3 clusters.
- Assigned labels of clusters formed to main data and used scatter, boxplot to check the clusters formed with gdpp, income and child_mort columns.

Step 7: Reporting 5 or more backward countries:

- Finally filtered the data by both the clustering labels and sorted the values with gdpp, income and child_mort columns and noticed that the top 10 backward countries are Burundi, Liberia, Congo, Dem. Rep., Niger, Sierra Leone, Madagascar, Mozambique, Central African Republic, Malawi, Eritrea.

Question 2: Clustering

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

Answer:

K-means Clustering:

In K-means clustering, we initially assign the random points as centroids and we will calculate the Euclidean distance from each point to these centers and assign labels based on the lowest distance to the cluster centers. We will calculate the centroids again using the labels assigned. We will again calculate the Euclidean distance from each point to these new centroids and update the labels and this steps repeats until there is no change in the centroids. In, k-means clustering we need to find the optimum value of k in order to form the clusters. We use elbow curve method or Silhouette analysis to find the optimum k-value. Also, the final clusters formed are based on the initial centroids, so the initial centroids should be randomly taken until we achieve at a decent cluster formation or we can use k-means ++ to initialize the centroids.

Hierarchical Clustering:

Hierarchical Clustering produces a hierarchy that reduces these steps and we can choose any cluster depending upon our business understanding or through some intuition. By just looking at the dendrogram we can see where to cut the line and take the number of clusters as per our own preference.

Comparison:

K-Means Clustering:

Category - Centroid based, partition-based

Method to find the optimal number of clusters - The Elbow method using SSD, Silhouette Analysis

Directional approach- Not any, the only centroid is considered to form clusters

Python Library- sklearn.cluster.KMeans

Hierarchical Clustering:

Category- Hierarchical, Agglomerative or Divisive

Method to find the optimal number of clusters - Dendrogram

Directional approach - Top-down, bottom-up

Python Library - sklearn.cluster.AgglomerativeClusterin

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

Answer:

The steps involved in k-means clustering are:

Let X = {x1,x2,x3,……..,xn} be the set of data points and C = {c1,c2,…….,Cc} be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using:

Where, 'ci' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3.

Finally if there is no change in the cluster centers, we will stop and label the data based the clusters formed.

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:

There are mainly 2 methods used to determine the value of ‘k’ for K-means clustering:

The “Elbow” Method: The Elbow method looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn’t improve much better the total WSS (within-cluster sum of square).The total WSS measures the compactness of the clustering and we want it to be as small as possible.

The Silhouette Method: It determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

The above 2 methods helps us in determining the optimum number of clusters we can form. However, we need to take the Business requirements as well. We need to discuss with the Business team and based on their requirements we need to form the clusters.

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

Answer:

Scaling is necessary before performing clustering incase if the variables have different units. If we have mixed numerical data, where each attribute is something entirely different and has different units attached then these values aren't really comparable anyway and the weights may vary. Standardizing them is a best-practice to give equal weight to them. K-means minimizes the error function using a gradient-based optimization algorithm. Normalizing the data improves convergence of such algorithms. The idea is that if different components of data (features) have different scales, then derivatives tend to align along directions with higher variance, which leads to poorer/slower convergence.

e) Explain the different linkages used in Hierarchical Clustering

Answer:

There are 3 different types of linkages in Hierarchical clustering:

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.

We have to decide what type of linkage should be used by looking at the data. One convenient way to decide is to look at how the dendrogram looks. Usually, a single linkage-type will produce dendrograms which are not structured properly, whereas complete or average linkage will produce clusters which have a proper tree-like structure.