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.

Answer 1:

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.

It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making

this decision are mostly related to choosing the countries that are in the direst need of aid.

My Approach:

I have done the following steps for my assignment.

Data Transformation :

-In our dataset, the 'imports', 'exports' and 'health' variables seem to be in percentage of GDP per capita, and this can sometimes give an incorrect insight in our EDA.

-For example, the health spending of 'United states' is 17.9 and that of 'Sierra Leone' is '13.1',both of which are very close to each other in health spending in terms of their % of GDP per capita, but these figures do not actually tell us the real story of how rich and poor are 'USA' and 'Sierra Leone' is.

-So the best way to tackle it is to convert the % values to ABSOLUTE values.

Outlier Treatment:

-There seems to be outliers in every single variable.

-This is a very delicate situation in terms of Business problem statement & Clustering analysis.

-If we apply outlier treatment by CAPPING, this will change the ranking of a few countries with respect to requirements of Financial Aid, also we will still have some outlier present after Capping, so it's not a wise decision in this business scenario.

-If we apply outlier treatment by deletion based on IQR values, this will remove a few countries from the list that would have really deserved the Financial Aid.

-If we do not apply Outlier treatment, it can impact the clustering model, as the presence of Outlier can change the CENTROID (K-Means) of the cluster.

-After considering all these scenarios, I've decided to go with the model which suits the Business Problem better. That is, not to treat the Outlier, and check the model with different K values to see which one gives a better business outcome.

EDA:

-Most of the data points are 'NOT Normally' distributed.

-Their variance as well as their range are different.

-All the above points indicate the need of standardising the data before we build the model.

-Since we need to compute the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range.

-Thus, scaling down of all attributes to the same normal scale is important here.

Hopkins Statistics(Cluster Tendency):

-Hopkins Statistic over .70 is a good score that indicates that the data is good for cluster analysis.

-A Hopkins Statistic value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

-For the given dataset, average Hopkins Statistics value is more than .90 which reflects the dataset is highly clustered.

Scaling:

-I used the Standardisation method for scaling the data.

CLuster Modelling:

-Using Hierarchical Clustering to identify optimal cluster value.

-Using Silhouette and Elbow method to validate optimal cluster value.

-Use the K-Means cluster method to build the final Cluster Model.

Interpreting Cluster:

-Identify appropriate clusters for financial aid using cluster mean method.

-Analyze the final cluster statistics against other clusters.

-Decision making on the final list based on the descriptive statistics of the final cluster.

Final Countries selection:

-Choose the top 10 countries from the final cluster based on higher child mortality, lower gdpp and lower income.

Question 2: Clustering

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

Answer :

K-Means Clustering	**Hierarchical Clustering**
We need to have the desired number of clusters ahead of time.	We can decide the number of clusters after completion of plotting the dendrogram by cutting the dendrogram at different heights.
It is a collection of data points in one cluster which are similar between them and not similar data points that belong to another cluster.	Clusters have tree-like structures and most similar clusters are first combined which continues until we reach a single branch.
Works very well in large datasets.	Works well in small dataset and not well with large dataset.
The main drawback of k-Means is it doesn't evaluate outliers properly.	Outliers are properly explained in hierarchical clustering.
K-means is only used for numerical.	Hierarchical clustering is used when we have a variety of data as it doesn't require to calculate any distance.

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

Answer :

Step 1: Randomly select K points as initial centroids.

Step 2: All the data points closest to the centroid will create a cluster center according to Euclidean distance function.

Step 3: Once we assign all the points to each of k clusters, we need to update the cluster centers or centroid of that cluster created.

Step 4: Repeat 2,3 steps until cluster centers reach convergence.

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: 'K' value is chosen randomly in K-Means clustering based on statistical aspect. From business aspect, we need to first understand the dataset and based on that we decide number of 'k'. for example, we have a dataset of variables like 'pen', 'pencil', 'books', 'notebooks', 'mobiles', 'charger', 'laptop'. Now if we want to have k values based on statistical aspect, we can use silhouette score to determine that but based on business aspect, after viewing the dataset we can easily make cluster = 2, one in electronics category and another non-electronics.

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

Answer: It is definitely a good idea to do scaling/standardisation because our variables may have units at different scale and as our method stresses

more on calculation of direction of space or distance, so if we have one variable with high scale units then while calculating for k-Means or hierarchical it will create a big difference as the clusters will tend to move with the variables having greater values or variances. By applying standardisation/scaling will increase the performance of our model.

e) Explain the different linkages used in Hierarchical Clustering.

Answer: Linkage is a technique used in Agglomerative Clustering. Linkage helps us to merge two data points into one using the below linkage technique.

Single linkage: The distance between two clusters is calculated by the minimum distance between two points from each cluster.

Complete linkage: The distance between two clusters is calculated by the maximum distance between two points from each cluster.

Average linkage: The distance between two clusters is the average distance between every point of one cluster to the other every point of another cluster.