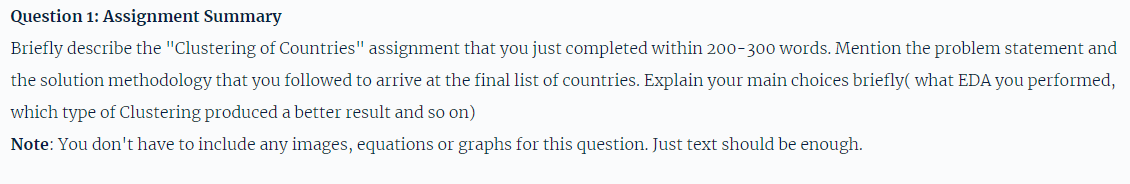
Assignment: Part II



Answer:

Our main objective for this assignment is to find the countries that are in direst need of aid. Our job is to find those countries using socio-economic and heath factors which will show overall development of the country.

So, in order to do so first I analyzed the dataset – It has total 167 countries with no missing values, In our dataset (ngo\_df), the 'imports', 'exports' and 'health' variable are in percentage of GDP per capita, and this can sometimes give an incorrect insight in our EDA and hence converted them to their Absolute values. After that using box plot we see that outliers are present for all variables .If we apply capping (based on IQR treatment) now ranking of countries would change with respect to the business problem i.e. of providing the country that actually require 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. As the end goal is to achieve the business goal for now Outlier treatment is avoided and model with different K-values will be checked giving better business outcome. We then move on with EDA and can see high correlation b/w imports and exports also gdp and health; total\_fer & child\_mort also have correlation. Next 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 .We begin our analysis now with Hierarchical clustering and can see that Complete Linkage gives a better cluster formation(k=3). We could also see that Cluster 0(under-developed cluster) comprises of ~89% of overall data, and has ~148 observations in comparison to 167 total observations; this seems to be a problem. This means that Hierarchical clustering is not giving us a good result as 89% of the data points are segmented into that cluster. We also saw that increasing the cluster number is not solving this problem. We will perform K-Means Clustering and check how that turns out to be. Using Silhouette and Elbow method to validate the optimal cluster value (k=3). K-Means, we could see that using 3 Clusters provided a better output in terms of a balanced cluster size. We will go ahead with MEDIAN and not MEAN, as there seems to be a greater variability in the income and gdpp values). So the approach will be to filter out all countries from our target cluster with income & gdpp less than the Median of 1860 & 932 respectively and Child Mortality >= the Median of 90. --- We will perform the filtering in the order GDPP --> INCOME --> CHILD MORTALITY --- This is due to reason that, we need to identify the countries with lowest GDPP & INCOME first and then with maximum CHILD MORTALITY. This is based on the understanding that, countries with highest child mortality and having higher gdpp & income will not have any impact on the child mortality rate even after the financial aid.So we will consider the 'K-Means with 3 Clusters' as our FINAL MODEL. We then use median values of the target cluster to determine the top 10 countries from the final cluster based on higher child mortality and lower income.

.

**Question 2: Clustering**

1. Compare and contrast K-means Clustering and Hierarchical Clustering.
2. Briefly explain the steps of the K-means clustering algorithm.
3. How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.
4. Explain the necessity for scaling/standardization before performing Clustering.
5. Explain the different linkages used in Hierarchical Clustering.

Answer:

a)

|  |  |
| --- | --- |
| **K-Means Clustering** | **Hierarchical Clustering** |
| We need to have desired number of clusters ahead of time. | We can decide the number of clusters after completion of plotting 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 belongs to another cluster. | Clusters have tree like structures and most similar clusters are first combine which continues until we reach a single branch. |
| Works very good in large dataset | Works well in small dataset and not good with large dataset |
| The main drawback of k-Means is it doesn’t evaluate  properly in presence of outliers. | Outliers are properly explained in hierarchical clustering |
| K-means only used for numerical. | Hierarchical clustering is used when we have variety of data as it doesn’t require to calculate any distance. |

b)

Step 1: Randomly select K points as initial centroids.

Step 2: All the data points closer to the centroid will create 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)

‘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)

It is definitely a good idea to do scaling/standardization 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 standardization/scaling will increase the performance of our model.

e)

Linkage is a technique used in Agglomerative Clustering.

Linkage helps us to merge two data points into one using 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 every point of other cluster.

**Ward linkage**: The distance between clusters is calculated by the sum of squared differences with all clusters.