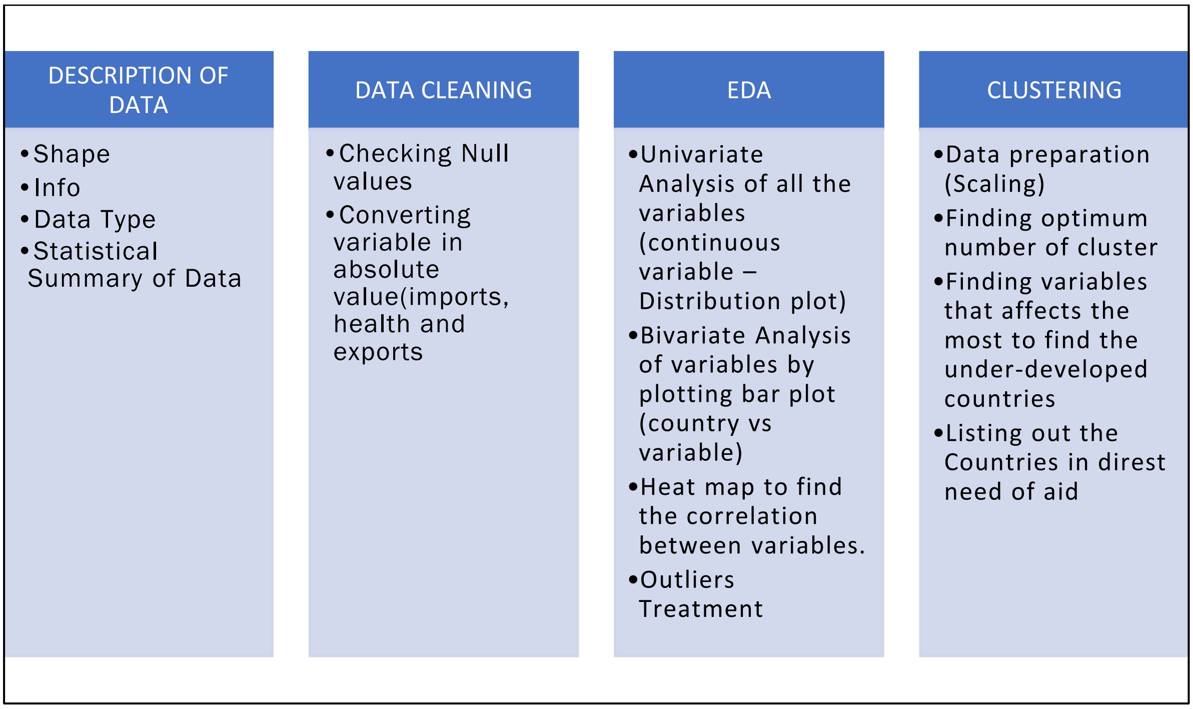
Question 1: Assignment Summary

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)

As a data analyst I need to suggest the CEO of the NGO (HELP International) that which all countries are in direst need of aid since, CEO has to make the decision that how he can use the money strategically and effectively.

For the above problem statement, I have concluded by following the approach and methodology as mentioned below:



DESCRIPTION OF DATA

- Shape
- Info
- Data Type
- Statistical Summary of Data

DATA CLEANING

- Checking Null values
- Converting variable in absolute value(imports, health and exports

EDA

- Univariate Analysis of all the variables (continuous variable – Distribution plot)
- Bivariate Analysis of variables by plotting bar plot (country vs variable)
- Heat map to find the correlation between variables.
- Outliers Treatment

CLUSTERING

- Data preparation (Scaling)
- Finding optimum number of cluster
- Finding variables that affects the most to find the under-developed countries
- Listing out the Countries in direst need of aid

Firstly, I have described the data – number of columns, number of rows, data type of variables and statistical summary of the data. After looking the data set, I have cleaned the data as required by converting variables (imports, health, exports) in absolute values.

Followed by this Exploratory data analysis is carried out by performing univariate analysis of all the variables using distribution plot, Bivariate analysis using Bar plot between country and all other variables and heat map is plotted to find the correlation between the variables, and lastly box plot of all the variables are plotted to find the outliers, and treatment for the same is performed by Capping the upper or lower range outliers as per the business requirement.

Finally, clustering is performed using K-mean clustering and Hierarchical clustering method, and since, countries are not properly assigned using Hierarchical clustering method therefore, final result is extracted by K-mean clustering method. When clustering is performed using Hierarchical method, only one country was assigned under one of the 3 clusters. This is the reason that why I have chosen the K-mean method over Hierarchical clustering method.

Conclusion and Recommendations

- The CEO of the NGO majorly should focus on the countries with low gdpp. Since most of the other factors such as Imports, Exports, health and income is highly correlated with the gdpp factor.
- NGO should also focus on the countries with high child mortality. Child mortality is highly positively correlated with total fertility and negatively correlated with life expectancy factor therefore, no need to look after the other two variables.
- Since countries with less imports and exports will not have source of income from this major sector which helps in improving economy of the country therefore, countries with less Exports and Imports of goods and services per capita should also be considered.
- All other factors are highly correlated with child mortality and gdpp therefore, CEO should focus on the countries with high child mortality and low gdpp only.
- 5 countries which are in direst need of aid are - "Burundi", "Liberia", "Congo, Dem.Rep", "Niger", "Sierra Leone".

Question 2: Clustering

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

There are more than 100 clustering algorithms known but very few of them are used thoroughly. Two of them are centroid models and connectivity models. K-mean clustering algorithm falls under centroid model whereas hierarchical clustering algorithm falls under connectivity model.

K-mean Clustering Algorithm	**Hierarchical Clustering Algorithm**
Number of clusters should be predefined	No need to predefine the number of clusters
Number of clusters can be determined using elbow curve or silhouette score	Number of clusters is derived using dendrogram
Centroid value is used to assign the data points to the clusters	Two type of approach can be used - 1) Top to Bottom (Divisive Clustering), 2) Bottom to Top (Agglomerative Clustering).
Can handle big dataset	Cannot handle big dataset.
Time complexity is linear therefore, less expensive.	Time complexity is quadratic (Computationally Intensive) therefore, more expensive.
Every time we run the code; we might get different result because k-mean clustering algorithm is centroid model based.	This is not the case with Hierarchical clustering.

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

K-mean clustering algorithm follows centroid model. Steps in k-mean algorithm are mentioned below:

- Firstly, initiate the algorithm by giving number of clusters (k) and random k number of centroid points (uk).(Figure1)

- In second step, data points are assigned to the centroid by calculating the distance between the data point (xi) and centroid (uk) {d(xi, uk)}. Data point will get assigned to the nearest centroid. (Figure 2)
- In the third step, mean of all the data points assigned to centroid is calculated. That is algorithm will calculate mean of each cluster formed and then set that mean as the centroid point. (Figure 3)
- Optimization – second step and third step will be repeated until the position of centroid get fixed. (Figure 2, Figure 3)

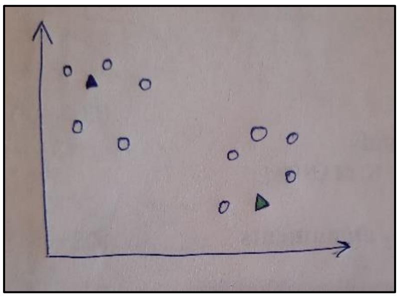


Figure 1 Initialize two centroids

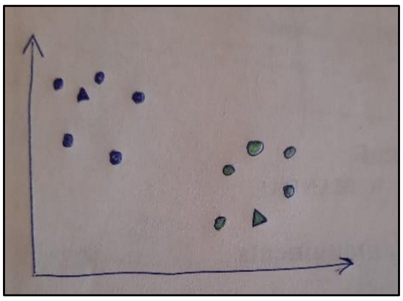


Figure 2 Assigned Data points (calculating distance)

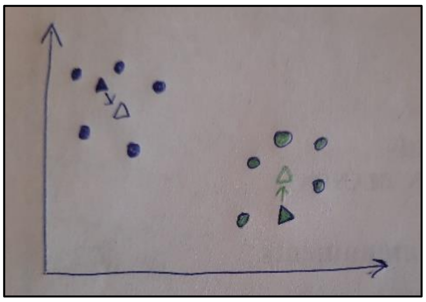


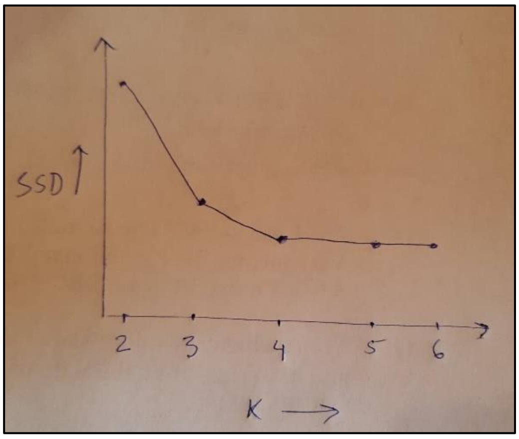
Figure 3 Location of centroid changed

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

There are two methods by which we can chose the value of k in k-mean clustering.

1) Elbow curve
2) Silhouette score

Firstly, we will calculate the SSDs (sum of squared distance) value within clusters for different values of k. In second step we will plot the curve between k (number of clusters) and SSD values. We will almost get a plot as shown below known as elbow curve.



We will finalize the value of k by observing the elbow plot. The point where we are getting sudden change in SSD value will be finalized as value of k. For example, in above plot we are getting break point at k = 3, therefore we will consider number of clusters as 3.

However, in many cases it is difficult to observe the k by just plotting elbow curve. So, many a times we use Silhouette score to finalize the k value.

Silhouette score measures that how a data point is similar to its own cluster as compared to the other clusters. Therefore, Silhouette curve is a metric which can be used to validate the decision made by using elbow curve, and not just for making decision.

$$\text{silhouette score} = \frac{p - q}{max(p, q)}$$

p = average distance from the nearest neighbour cluster(separation)

q = average distance from its own cluster (Cohesion)

- silhouette score value lies between -1 to 1
- Silhouette score close to 1 signifies that data points is very similar to other data points in the cluster, i.e we have chosen right number of clusters.
- Silhouette score close to -1 signifies that data points is very different to other data points in the cluster, i.e we have created too few or too many clusters.

After clustering the data, it is very important to evaluate the clusters according to business need. Our goal is not just about creating the clusters from the data set but it is more about creating accurate and meaningful clusters that will help us in observing inferences about the business problem.

For example, in consumer segmentation business problem we can form the clusters and identify the consumer segments by using variable present in the dataset (demographic, behavioural, etc.) and finally, we can do cluster profiling to understand the consumers and to describe them according to the factors used in clustering.

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

Variable used are often of different scale. Scaling of a variable has a large impact on the cluster formed. We are using Euclidean distance method in the clustering process; therefore, it is must to bring all the variables on the same scale. This can be achieved by using Standardisation.

For example, data of retail purchases of store that sells pen and laptop. So, definitely number of pens sold will be much higher than the number of laptops. Therefore, when we try to cluster this data then number of pens sold will influence the data and the number of laptops sold will have very little effect. This will not be correct from the business perspective, since value of the laptop is much higher than the value of the pen. From this example it is much clear that it is must to perform standardisation before clustering process.

e) Explain the different linkages used in Hierarchical Clustering.

There are three methods of linkage as mentioned below:

1) Single Linkage
2) Complete Linkage

3) Average Linkage

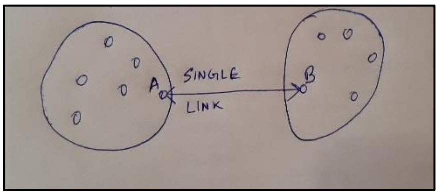


Figure 1 Single Linkage (minimum distance)

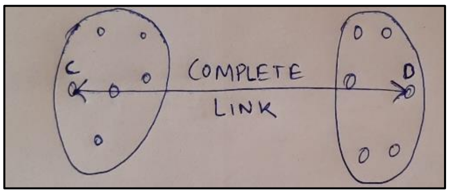


Figure 1Complete Linkage (maximum distance)

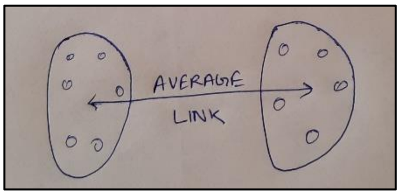
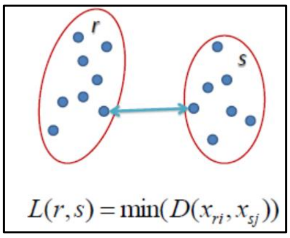


Figure 3 Average Linkage (average distance)

Single Linkage

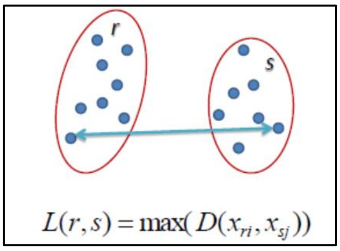
Single linkage used in hierarchical clustering is defined as the minimum distance between two points in each cluster. For example, in Figure 1, distance between the cluster is equal to AB (minimum distance between the cluster)



$$L(r,s)=\min(D(x_{ri},x_{sj}))$$

Complete Linkage

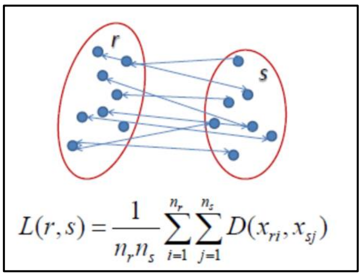
Complete linkage used in hierarchical clustering is defined as the maximum distance between two points in each cluster. For example, in Figure 2, distance between the cluster is equal to CD (maximum distance between the cluster)



$$L(r,s)=\max(D(x_{ri},x_{sj}))$$

Average Linkage

Average linkage used in hierarchical clustering use all the data points in the cluster to find the distance and is defined as the average distance between all the points in one cluster to all the points in the other cluster.



$$L(r,s)=\frac{1}{n_r n_s}\sum_{i=1}^{n_r}\sum_{j=1}^{n_s}D(x_{ri},x_{sj})$$

In all the above methods distance signifies the similarity between the clusters.