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)

The below are the steps used to perform analysis on the Countries dataset provided.

Problem Statement:

To identify the top 5 countries which are in direst need of aid so that the HELP International (which is an international humanitarian NGO that is committed to fighting poverty) can provide the people of backward countries with basic amenities and relief during the time of disasters and natural calamities by providing those funds.

Data Understanding

- Imported the dataset and analysed the columns.
- Checked if there were any missing values present in the dataset. There were no missing values present(i.e. EDA Analysis)
- Converted exports, health and imports from percentage value of GDP per capita to their actual value by multiplying it with 'gdpp' and dividing by 100.
- Next step, I checked for outliers in the dataset. Identified the outliers by using Skewness which explains the extent to which the data is normally distributed. Ideally, the skewness value should be between -1 and +1, and any major deviation from this range indicates the presence of extreme values.
- Capped the outliers for the below columns by capping outliers to 1% and 98% respectively:-
  - income
  - child_mort
  - exports
  - imports
  - gdpp
  - inflation

Modelling

- Rescaled the columns using Standard Scalar.
- Used Hopkins Check to evaluate if the data is feasible for clustering or not (i.e. checking cluster tendency).
- The dataset obtained a Hopkins Check value of 87.4 % (Higher values indicates it has a high tendency form clusters)

K-Means Clustering

- Used Elbow curve/SSD method and Silhouette analysis to find the optimal number of clusters (i.e. value of k)
- Selected the value of k as 3 by visualizing the plots from the above analysis.
- Used KMeans to obtain the clusters and assigned the cluster_id to the original dataset.
- Visualized the clusters using box-plot and scatter plots and concluded that ***Cluster 0*** is the one which is in dire need of aid, since it had the lowest gdpp,lowest income and high child mortality rate
- Fetched the top 5 countries from this cluster by sorting gdpp in ascending order, income in ascending order and child_mort in descending order.

Hierarchical Clustering

- Plotted the dendograms using single, average and complete linkages.
- Based on the complete linkage,I cut the dendogram in 3 clusters/segments.
- Assigned the cluster labels to the dataset
- Visualized the clusters using box-plot and scatter plots and concluded that ***Cluster 0*** is the one which is in dire need of aid, since they have lowest gdpp,lowest income and highest child mortality rate
- Fetched the top 5 countries from this cluster by sorting gdpp in ascending order, income in ascending order and child_mort in descending order.

Conclusion

- K means and Hierarchical Clustering resulted in the same list of top 5 countries
- Below is the final list of the top 5 countries

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id	cluster_labels
88	Liberia	89.30	62.457000	38.5860	302.80200	742.24	5.47	60.8	5.02	331.62	0	0
26	Burundi	93.60	22.243716	26.7960	104.90964	764.00	12.30	57.7	6.26	331.62	0	0
37	Congo, Dem. Rep.	116.00	137.274000	26.4194	165.66400	742.24	20.80	57.5	6.54	334.00	0	0
112	Niger	123.00	77.256000	17.9568	170.86800	814.00	2.55	58.8	7.49	348.00	0	0
132	Sierra Leone	145.16	67.032000	52.2690	137.65500	1220.00	17.20	55.0	5.20	399.00	0	0

Question 2: Clustering

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

Difference between K-Means and Hierarchical Clustering is highlighted below:-

Properties	K –Means	Hierarchical Clustering
Definition	K Means Clustering generates a specific number of disjoint, flat (non-hierarchical) Clusters.	Hierarchical Clustering method construct a hierarchy of Clustering, not just a single partition of objects.
Performance	The performance of K- mean algorithm is better than Hierarchical Clustering Algorithm.	Hierarchical Clustering Algorithm performance is less as compare to K- mean algorithm.
Cluster	K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into	You can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram
Data Set	K Means clustering can handle big data because the time complexity of K Means is linear i.e. O(n)	Hierarchical clustering can't handle big data since time complexity is quadratic i.e. O(n2).
Sensitive To Noise	K-Means is very sensitive to noise (i.e. outliers) in the dataset.	It is less sensitive to noise in the dataset
Execution Time	K -mean algorithm also increases its time of execution.	Hierarchical algorithm its performance is better.

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

K-Means algorithm is the process of dividing the N data points into K groups or clusters. Here the steps of the algorithm are:
1. Start by choosing K random points the initial cluster centres.
2. Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance.
3. For each cluster, compute the new cluster centre which will be the mean of all cluster members.
4. Now re-assign all the data points to the different clusters by taking into account the new cluster centres.
5. Keep iterating through the step 3 & 4 until there are no further changes possible.
At this point, you arrive at the optimal clusters.

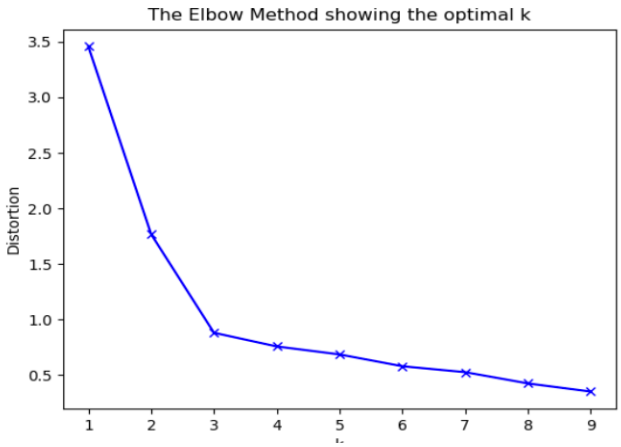
c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

There are 2 ways of selecting the optimal value of k in K-means clustering

a) **Elbow curve/SSD(Sum of squared distances)**
It is a plot where the x-axis will represent the number of clusters and the y-axis will be an evaluation metric

Steps involved:-

- Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters or each k, calculate the total within-cluster sum of square (wss).
- Plot the curve of wss according to the number of clusters k.
- The location of a bend (knee) in the plot is generally considered as an indicator of the
- Appropriate number of clusters.



b) **Silhouette analysis**
Steps involved:-

- Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
- For each k, calculate the average silhouette of observations (avg.sil).

- Plot the curve of avg.sil according to the number of clusters k.
- The location of the maximum is considered as the appropriate number of clusters.

From the business standpoint, it refers to how many groups/clusters does the business wants to classify the segments into so that specific/customised marketing campaigns can be created based on the clusters formed.

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

Let's consider the below scenario.
Consider your data has an age variable which tells about the age of a person in years and an income variable which tells the monthly income of the person in rupees:

ID	Age	Income(rupees)
1	25	80,000
2	30	100,000
3	40	90,000
4	30	50,000
5	40	110,000

Here the Age of the person ranges from 25 to 40 whereas the income variable ranges from 80,000 to 110,000.

Let's now try to find the similarity between observation 1 and 2 using Euclidean distance which is given by:-

$$D = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}$$

Euclidean Distance = [(100000–80000)^2 + (30–25)^2]^(1/2)=**20000.000625**

- It can be noted here that the high magnitude of income affected the distance between the two points. This will impact the performance of all distance based model as it will give higher weightage to variables which have higher magnitude (income in this case).
- We do not want our algorithm to be affected by the magnitude of these variables.
- The algorithm should not be biased towards variables with higher magnitude.
- To overcome this problem, we can bring down all the variables to the same scale.

One of the most common technique to do so is normalization where we calculate the mean and standard deviation of the variable. Then for each observation, we subtract the mean and then divide by the standard deviation of that variable:

$$z = \frac{x - \mu}{\sigma}$$

Apart from normalization, another method is Min-Max Scaling which is given by the following formula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

After applying normalization, the data looks like:-

ID	Age	Income(rupees)
1	-1.192	-0.260
2	-0.447	0.608
3	1.043	0.173
4	-0.447	-1.563
5	1.043	1.042

Let's again calculate the Euclidean distance between observation 1 and 2:

Euclidean Distance = [(0.608+0.260)^2 + (-0.447+1.192)^2]^(1/2) **= 1.1438**

We can clearly see that the distance is not biased towards the income variable. It is now giving similar weightage to both the variables. Hence, it is always advisable to bring all the features to the same scale for applying distance based algorithms like K-Means.

e) Explain the different linkages used in Hierarchical Clustering.

There are 3 types of linkages used in Hierarchical clustering namely single-linkage, complete-linkage, and average-linkage.

Single-Linkage

- Single-linkage (nearest neighbour) is the shortest distance between a pair of observations in two clusters.
- It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can appear spread-out.

Complete-Linkage

- Complete-linkage (farthest neighbour) is where distance is measured between the farthest pair of observations in two clusters.
- This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together.

Average-Linkage

- Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance.
- Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering.

Note: Complete and Average linkage methods give a well separated dendogram whereas single linkage gives us dendogram which are not well supported. We generally want well separated clusters.