Question 1: Assignment Summary

Clustering of Countries

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

Solution Methodology:

Steps followed :

1. **Data Reading and understanding**
2. **Data Cleaning**
3. **Data Visualization:**
    - **Univariate Analysis**
    - **Bivariate analysis**
4. **Outliers Treatment**
5. **Hopkins Statistics**
6. **Data Preparation**
7. **Finding the optimal k :**
    - **Elbow curve &**
    - **Silhouette Analysis Techniques**
8. **K-mean clustering**
9. **Hierarchical clustering**
10. **Conclusion**

1. **Data Reading and understanding:** Imported the dataset and unstood the data . Read the datatypes well. Also read the shape and the statistical point of view checked the features Converted the Three features export,import and health to the actual values .

2. **Data cleaning:** This process performed the cleaning of the data. First checked any null values in any of the columns then after checking that found none and further proceeded with the duplicate rows . The dataset was very clean with no null values and duplicates rows.

3. **Data viualization:**

Univariate Analysis: When performed the displot for all the features

- There is some variation in the distribution of the income dataset .
- The child mortality also shows variation in the distribution of the dataset .

- The gdpp also shows the data distributed with some variation .
- We will takes these features for the data profiling for helping us know the countries in need of aid.

The variation in the data distribution is what need when we do clustering to cluster the countries

- Performed bar plot on the gdpp for all the countries and found the country Luxembourg having the highest Gdpp .

Bivariate Analysis:

We did a scatter plot for gdpp vs income

- We can notice from the plot that countries with the low income have low gdpp.
- We can focus on these countries as they indicate the countries are poor.
- The countries with the low gdpp will definitely be in need of the aid.
- Countries with the high income rate shows good gdpp thus we can include them in the rich countries as of now.

We did a scatter plot for gdpp vs health

- We can notice from the plot that countries with the low gdpp are poor in the health also. We can focus on these countries.
- The countries with the low health will definitely be in need of the aid.
- Countries with the high gdpp rate shows good health thus we can include them in the rich countries as of now.

4. Outlier treatment:

- Since my data is less I will not go for removing the data from the dataset.
- Outlier capping is one option
- but we will not treat for all those features those who have outliers below the range as we may loose the pattern of the data and may not get the desired output(except child mort) .
- we can however treat the child mortality outliers in the lower range but as we see the feature shows outliers in the higher range we will not touch it as it shows that the countries with high child mortality rate is in need of the aid so we will not cap or we may loose those countries .

5. Hopkins Statistics

- Every time we run the above cell we get different hopkins value but if we observe we see its above 90 that means our dataset has a very good tendency to form clusters

6. Data Preparation: After treating the outliers we did the scaling which is the most important step before clustering

7. Finding the optimal k:

- **Elbow curve**
- **Silhouette Analysis Techniques**

The elbow curve and the Silhouette Analysis help in finding the optimal k for clustering. it is good to proceed with either 3 clusters. As 4 clusters might confuse with the behaviors.

8. **K-mean clustering**:When cluster with the k=3 we got three clusters in which the cluster 2 showing the behavior of the countries which are in the dire need of the aid.Country In cluster 2 are exhibiting the following behavior :

- **High child mortality**
- **Low gdpp**
- **Low income**

Country showing low gdpp and low income indicates that the country is not developed and are poor hence they are in need of the aid .

Countries from k-mean are :

Haiti
2.Sierra Leone
3.Chad
4.Central African Republic
5.Mali
6.Nigeria
7.Niger
8.Angola
9.Congo, Dem. Rep.
10.Burkina Faso

9 . **Hierarchical clustering:**

The countries in the cluster 0 shows the behavior of high child mortality, low income and low gdpp .So we can go ahead with the countries in this cluster but there was one more cluster 2 which had only one country in it Nigeria which is definitely one of the country exhibiting the behavior of low income ,low gdpp and high child mortality.But after clustering and sorting with the features high child mortality and low income and low gdpp we get countries similar to the K-mean clustering.

1.Haiti
2.Sierra Leone
3.Chad
4.Central African Republic
5.Mali
6.Niger
7.Angola
8.Congo, Dem. Rep.
9.Burkina Faso
10. Guinea Bissau

10 **Conclusion:** After analyzing the cluster through k-mean and hierarchical clustering **the k-mean cluster provided better clarity on the countries exhibiting similar behaviors like high child mortality, low gdpp and low income** .

Also if notice countries from k-mean and the hierarchical clustering provides similar countries. But Nigeria was one of the country present in the other cluster by the hierarchical cluster but we will still consider it because when noticed it in details we saw Nigeria do have high child mortality with low gdpp and low income

compare to other highly developed countries .And low income and low gdpp indicates the country is poor country still developing and thus needs the aid.
The final countries are been considered from k-mean as the neat behavior is been seen in the k-mean clustering .(However the hierarchical clustering too displays the similar countries)

.

Final Countries:

- **Haiti**
- **Niger**
- **Sierra Leone**
- **Angola**
- **Chad**
- **Congo, Dem. Rep**
- **Central African Republic**
- **Nigeria**
- **Mali**
- **Burkina Faso**

Question 2: Clustering

a) **Compare and contrast K-means Clustering and Hierarchical Clustering.**
Ans:

Difference between K Means and Hierarchical clustering

- Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. $O(n^2)$.
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
- K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram.

	K-Means Clustering	Hierarchical Clustering
Category	*Centroid based, partition-based*	*Hierarchical, Agglomerative*
Method to find the optimal number of	*The Elbow method, Silhouette analysis*	*Dendrogram*

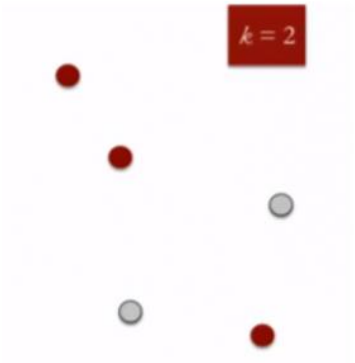
clusters		
Directional approach	*Not any, the only centroid is considered to form clusters*	*Top-down, bottom-up*
Python Library	*sklearn – KMeans*	*sklearn-AgglomerativeClustering*

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

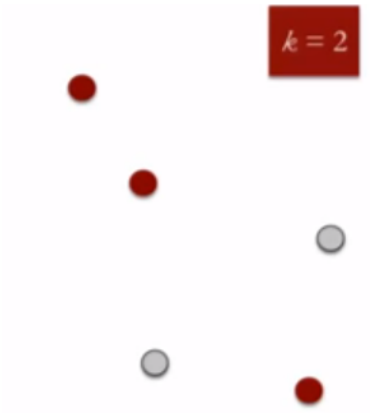
Ans:

K means is an iterative clustering algorithm that aims to find local maxima in each iteration. This algorithm works in these 5 steps :

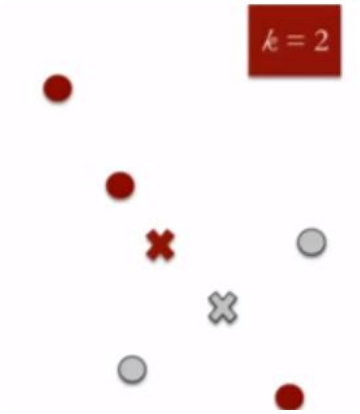
1. Specify the desired number of clusters K : Let us choose k=2 for these 5 data points in 2-D space.



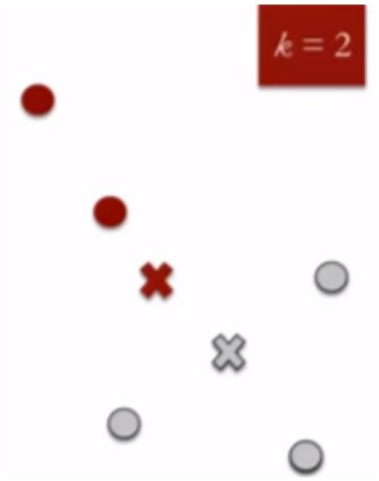
2. Randomly assign each data point to a cluster : Let's assign three points in cluster 1 shown using red color and two points in cluster 2 shown using grey color.



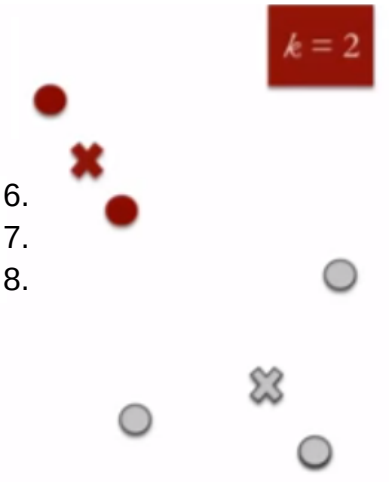
3. Compute cluster centroids : The centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.



4. Re-assign each point to the closest cluster centroid : Note that only the data point at the bottom is assigned to the red cluster even though its closer to the centroid of grey cluster. Thus, we assign that data point into grey cluster



5. Re-compute cluster centroids : Now, re-computing the centroids for both the clusters.



6.
7.
8.

Repeat steps 4 and 5 until no improvements are possible : Similarly, we'll repeat the 4th and 5th steps until we'll reach global optima. When there will be no further switching of data points between two clusters for two successive repeats.

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

Ans:

Determining the optimal number of clusters in a data set is a fundamental issue in partitioning clustering, such as k-means clustering, which requires the user to specify the number of clusters k to be generated.

We use Direct methods consists of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named **elbow and silhouette methods**, respectively.

Statistical method:

Elbow method

Recall that, the basic idea behind partitioning methods, such as k-means clustering, is to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total WSS measures the compactness of the clustering and we want it to be as small as possible.

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.

The optimal number of clusters can be defined as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.

2. For each k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

   Note that, the elbow method is sometimes ambiguous. An alternative is the average silhouette method which can be also used with any clustering approach.

Silhouette Analysis

It measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering. Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k.

The algorithm is similar to the elbow method and can be computed as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the average silhouette of observations (avg.sil).
3. Plot the curve of avg.sil according to the number of clusters k.
4. The location of the maximum is considered as the appropriate number of clusters.

Business aspect

Determining the k from the above techniques definitely helps in the good clustering for our analyzing purpose but not always the k defined by the elbow curve or the Silhouette method doesn't have to be accurate.

Sometimes keeping in mind the business aspects also brings lot of change in the cluster .For eg if the elbow curve suggests a good curve at k=5 or say k=6 and may be even the silhouette method also some gives a good score for k=4 or 5 or 6 . The decision should be based on the purpose of the analysis. Sometimes forming many clusters doesn't make any sense instead making many clusters may confuse the main objective of the analysis and thus end up with the data which might not make any sense .

So keeping the business point of view and also the purpose on what data is been analyzed the k value should be decided.

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

Ans:

Standardisation of data, that is, converting them into z-scores with mean 0 and standard deviation 1,is important for 2 reasons in K-Means algorithm:

• 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 helps in this process.

• The different attributes will have the measures in different units. Thus, standardisation helps in making the attributes **unit-free and uniform**.

e) Explain the different linkages used in Hierarchical Clustering.

Ans:

We took the minimum of all the pairwise distances between the data points as the representative of the distance between 2 clusters. This measure of distance is called **single linkage**. Apart from using the minimum, we can use other methods to compute the distance between the clusters. Let's consider the common types of linkages:-

1. **Single Linkage:** Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters
2. **Complete Linkage:** Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters
3. **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.