Question 1: Assignment Summary

Problem statement:

During the recent funding programs, we have been able to raise around $ 10 million. As an analyst, we have to come up with the countries list that are in the direst need of aid.

Analysis Flow/Methodology

1:-**Data collection and cleaning**

- Import thedata
- Identifying the data quality issuesandcleanthedata

2:- **EDA**

- Converting exports, imports and health spending percentages to absolute values

Visualizing the data

- Visualizing few original data variables to look for any patternorcorrelation.

Outlier Treatment

- For all the columns, we will not cap or drop the outliers at the lower range but we will treat the upper range outliers and this is because, we may end up losing some countries are actually in the need of AID.
- For child_mort, we will not cap or drop the outliers at the upper range but we will treat the lower range outliers.

  i. 1-99: Soft range
  ii. 5-95: mid-range
  iii. 25-75: hard range

3:- **Hopkins Statistics**

- Tocheck if data has tendencytoformclusters

4:- **Scaling the data**

- Standardizing all the continuous variables using the standard scalar

5:- **K means clustering**

- Identify the 'k' by silhouette analysis and sum of squared distances graph.
- Visualizing the clusters with various variables
- Analyzing the clusters
- Identifying the countries which requires aid.

6:-**Hierarchical Clustering**

- Identify the 'n' via dendrogram.
- Visualizing the clusters with various variables
- Analyzing the clusters
- Identifying the countries which requires aid.

7:-**Decision Making**

- Identifying the countries which requires aid by analyzing both K-means and Hierarchical Clustering results.

Question 2: Clustering

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

- Cluster analysis or simply k means clustering is the process of partitioning a set of data objects into subsets. Each subset is a cluster such that the similarity within the cluster is greater and the similarity between the clusters is less.

- In Hierarchical methods, we create hierarchical decomposition of the given set of data. We create hierarchical decomposition in two ways such as from bottom to the top or top to down. On the basis how we create hierarchical decomposition we divide this method into two approaches one is agglomerative approach and other is the divisive approach.

- K-means works well with larger data as hierarchical approach takes huge amount of memory for computation and is very tedious .Hierarchical is a linear method so once a data point is added to cluster it cannot be undone.
- The Challenge with K-mean is to select K , i.e , the number of centroids to start with .

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

Assignment step

- We start by choosing the k initial centroids (selection of k is random , generally between 3- 15)
- Now we calculate distance of each points from the chosen centres and allocate the point to the centre with least distance (euclidean distance is used to calculate distance)

Optimization step

- Recompute the centre for each cluster formed . Now we again go back to assignment step and assign the points to nearest cluster centre

We keep iterating these 2 steps (assignment and optimisation) till the centroids no longer update . At this point optimal clusters have been created

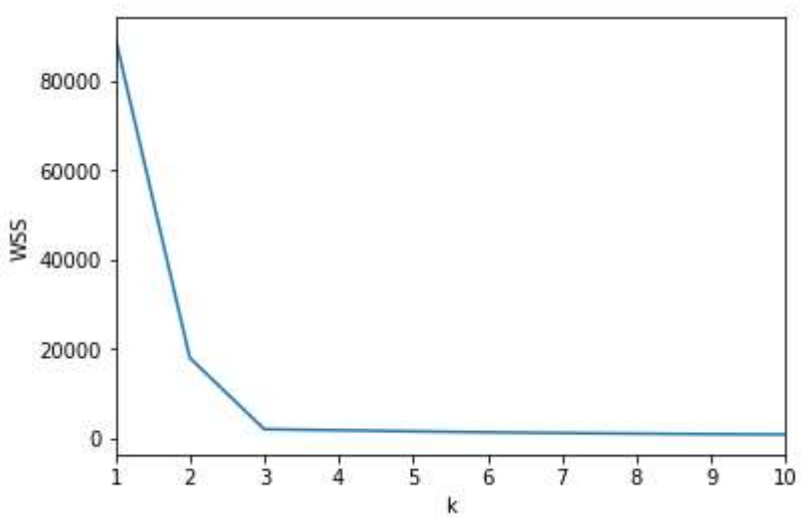
c) How is the value of 'k 'chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

The basic idea behind this method is that it plots the various values of cost with changing *k*. As the value of *K* increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the **elbow point**.

There are two ways to determine k :-The Elbow Method

I. The Silhouette Method

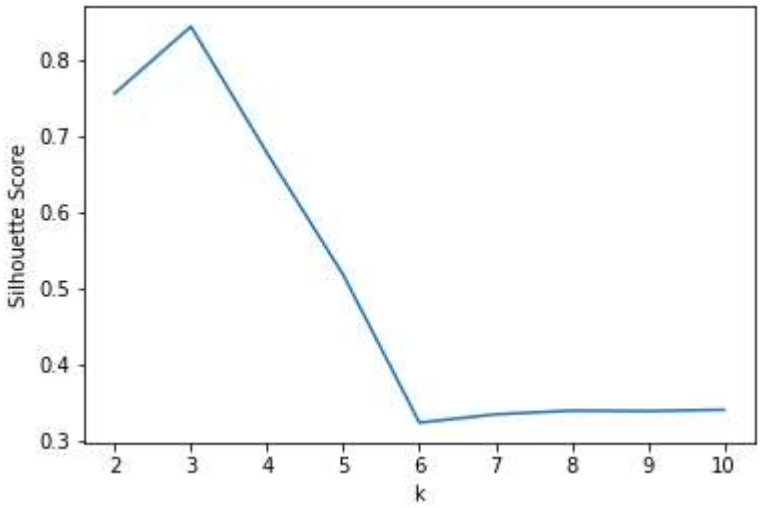
Elbow Method



plot looks like an arm with a clear elbow at k = 3 after plotting the SSD graph .

Silhouette Method

high Silhouette Score is desirable. The Silhouette Score reaches its ***global maximum at the optimal k***. This should ideally appear as a peak in the Silhouette Value-versus-k plot



There is a clear peak at k = 3. Hence, it is optimal.

Impact :-
K-means clustering is an unsupervised algorithm which you can use to organise large amounts of retail data to generate competitive insights about your business. There are many use cases which can help you implement this practice in your business and compete strategically in the retail market

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

Most of the times, your dataset will contain features highly varying in magnitudes, units and range.
But since, most of the machine learning algorithms use eucledian distance between two data points in their computations , we need to take care of it , as features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes .
To supress this effect, we need to bring all features to the same level of magnitudes. This can be acheived by scaling.

e) Explain the different linkages used in Hierarchical Clustering.

There are three types of linkage in hierarchical clustering :

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