Question 1: Assignment Summary

Problem Statement:

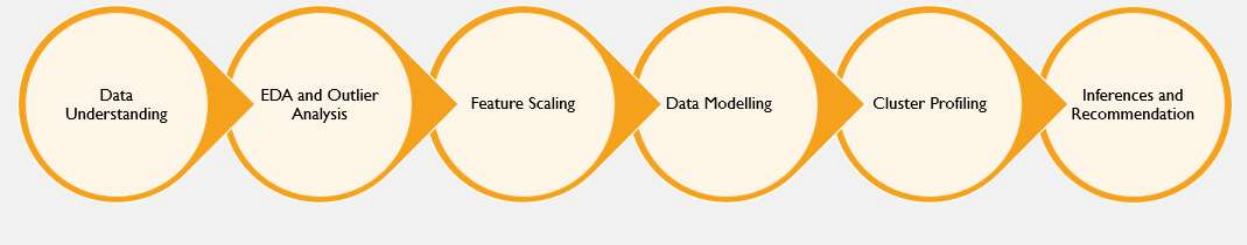
An NGO named HELP is committed to help the under developed countries by providing them financial AID and wants us to analyze the situation of various countries based on it socio – economic and health factors. We need to use the unsupervised learning technique ie clustering model (K means and Hierarchical) and identify the countries which need the aid.

Model Objective: Cluster the countries by the factors mentioned and recommend it to the CEO the top five countries which needs help. I have used columns gdpp, income and child mortality to analyse the clusters.

Business understanding:

- 1. Gross Domestic Product of a country indicates the total value of production of goods and services of that country and thus indicates that the above five countries have the least production happening in their country. This in turn would lead to the need of requiring external support from other countries. When the financial help is provided to the countries based on the low GDP indicator, such countries should be able to produce more goods and services, which in turn helps in the countries overall development.
- 2. Low Income is a sign of the inability to purchase commodities or services which can keep a family comfortable or in dire need of resources in case of very low income families. As such, foreign help would provide the required help to such families at a lower cost.
- 3. Child mortality rate can be the third indicator as it is not directly linked to the resources available, as it is affected to other reasons like incomplete care of the mother during pregnancy, irresponsibility of the health care workers, etc. Thus, child mortality rate can be considered as the third factor in this problem, keeping in mind the malnutrition and other financial problems the country is facing.

Solution Methodology:



- Firstly, I have imported the data and have analyzed it and have converted the columns exports, health and imports to its actual numerical values. Then have checked for missing values and have done detailed exploratory analysis like plotting a distplot for each column (univariate analysis) and pair plot(bivariate analysis). The analysis shows that the features child_mortality , gdpp , income, total_fer ,life_expec shows are widely distributed.

- Then have performed the outlier analysis and have capped all the outliers as the dataset is small and dropping the outliers will lose the information. After which feature scaling has been performed using the Standard scaler method.
- Under data modeling first I have performed the Hopkins stats method to check if the data set is suitable for clustering after which using elbow curve and silhouette score have identified that optimal value for k(num of clusters) is 3.
- Have performed the k means, hierarchical clustering and cluster profiling to identify the countries which have less gdpp, less income and high high mortality rate as these three indicates under developed countries and the top 5 countries of these clusters have been identified. Both the solutions have given the same top 5 countries and in hierarchal I have used the complete linkage method as this method has given me more clearer insights that the single linkage method.
- The top five countries from the under developed country cluster(least gdpp then least income and high mortality rate) have been identified as the countries which are in dire need of AID recommended to the CEO of the organization.

Subjective questions:

Question 2: Clustering

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

K-means Clustering	Hierarchical Clustering
K means is an iterative clustering algorithm that aims to find local maxima in each iteration.	**This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left.**
This algorithm can be used to handle big data	**This algorithm cant be used to handle big data**
Time complexity of K Means is linear i.e. O(n)	Time complexity of **hierarchical clustering is quadratic i.e. O(n2).**
We need to initialize the k value upfront	We need not define the size of cluster .i.e. the k value. This builds clusters incrementally, producing a dendogram.
K – means is a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. It is a division of object is in exactly one cluster, not several.	In Hierarchical clustering, clusters have a tree like structure or a parent child relationship. Here, the two most similar clusters are combined together and continue to combine until all objects are in the same cluster.
K - means will often give unintuitive results.	Hierarchical clustering can be more computationally expensive.

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

K-Means Algorithm:

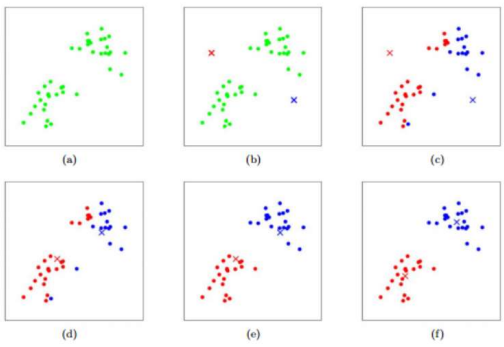
k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) .A cluster refers to a collection of data points aggregated together because of certain similarities.

There are two major steps in this algorithm:

1. Assignment steps
2. Optimization step

Algorithm of k-Means:

1. Initially specify the number of clusters(k= no of clusters) based on the elbow curve and the silhouette score.
2. Initialize the centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids ie assignment of data points to the clusters isn't changing.
4. Compute the sum of the squared distance between the data points and all the centroids.
5. Assign each data point to the closest cluster(centroid).
6. Compute the centroids for the clusters by taking the average of all the data points that belong to the cluster.
7. These above steps are repeated until there is no change in the centroid.
8. Finally, the clusters and its centroid is obtained.
   The below figure shows each step of K Means Clustering algorithm.



K Means Clustering:

1. Maximizes the tightness/Closeness between the clusers(less intra cluster distance)
2. Maximizes the distance between the clusters (more inter cluster distance)

Summary:

Initialize the cluster centers

Assignment: datapoints - > cluster centers (based on the distance between each individual point and the center)

Optimization: recompute Cluster center

The above two steps are repeated until no further optimization can be done.

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

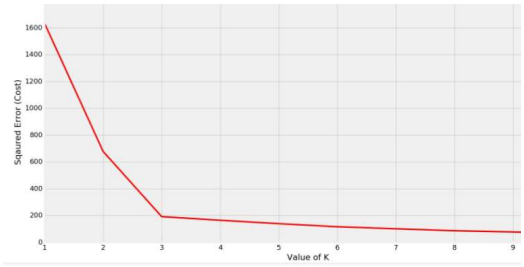
While clustering our data set using algorithms such as k means we need to define the k value at the beginning. So determining the optimal number of clusters in a data set is a fundamental issue.In general, there is no one particular method to determine the exact value of K, but an accurate values of K can be estimated through the following methods.

1. Statistical methods:

Elbow method: The basic idea in 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.

How does it work:

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.



5. At k=3 you get optimal cluster so we can chose k(no of clusters)value as 3

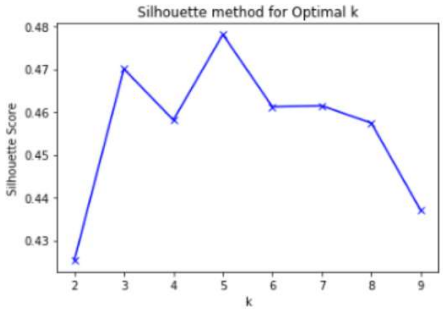
Silhouette method Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k (Kaufman and Rousseeuw 1990).

This is a better measure to decide the number of clusters to be formulated from the data. It is calculated for each instance.

Silhouette Coefficient = (x-y)/ max(x,y)

where, **y** is the mean intra cluster distance: mean distance to the other instances in the same cluster. **x** depicts mean nearest cluster distance i.e. mean distance to the instances of the next closest cluster.

The coefficient varies between -1 and 1. A value close to 1 implies that the instance is close to its cluster is a part of the right cluster. Whereas, a value close to -1 means that the value is assigned to the wrong cluster.



The above graph shows k=3 or k=5 may be the optimal number. Now we can check with the business which might be more meaningful and make the decision.

Gap statistic method

The *gap statistic* has been published by R. Tibshirani, G. Walther, and T. Hastie (Standford University, 2001). The approach can be applied to any clustering method.

The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e, that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points.

A number of other techniques exist for validating K, including cross validation, information criteria, the information-theoretic jump method, the silhouette method and the G-means algorithm. In addition, monitoring the distribution of data points across groups provided insight into how the algorithm is splitting the data for each K.

2. Business Aspect : A person with business domain expertise can have a look at the data and our statistical analysis to find the right K and define the K based on their business requirement. Like the business can check and tell it would be meaningful to cluster only 3 groups instead of 5 because the information that the three groups gives is enough for them to build their strategies and making the cluster size 5 might increase their complications So we can go ahead and build our model using k=3.

So when we combines both our statically analysis and the business expertise decisions we can find an optimal value for k and build the model. For the business to explain our analysis we can use the

hierarchical model(in which we dint define k) so that they can get clear understanding on how the clusters are formed and their k values.

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

Standardization which is known as feature scaling is a process in which we rescale the features or variables of our dataset so that they all are in the same scale. This is done as a part of preprocessing and data preparation for data modeling step. This is usually done when our data set have variables which are of different units (e.g. meters, tones, inches etc.) or when the scales of each of our variables are very different from one another (e.g., 0-1 vs 0-1000). This is very important in cluster analysis because the clusters are defined based on the distance metrics (the distance between points in mathematical space).

For eg Our data has features which means different (i.e. age and weight fields can't be directly compared. We cant say one year is less that 50 kg) or may not have same level of importance in sorting or clustering the records (age might be less important than weight). In such scenario by standardizing the features we can avoid the dependency on the choice of measurement units. Standardizing converts the original measurements into unitless variables.

Standardizing variable tends to make the training process better behaves by improving the numerical condition of the optimization process and ensuring that the default values involved in initialization and termination are appropriate.

e) Explain the different linkages used in Hierarchical Clustering.

Hierarchical Clustering involves either Agglomerative clustering or divisive clustering ie clustering the sub clusters into larger clusters in a bottom up approach or dividing the larger cluster into smaller clusters in a top down approach. To implement this, we need to calculate the distance between two sub clusters or data points. The different approaches to measure this distance between two sub clusters is called different types of linkages.

There are three different types of linkages which are used in hierarchical clustering

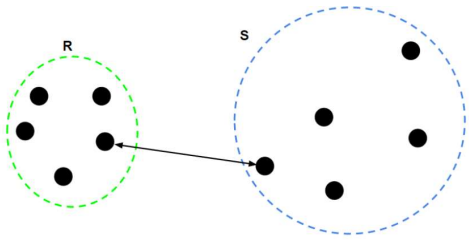
1. Single Linkage
2. Complete linkage
3. Average linkage

Single Linkage:

For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.

Formula:

$$L(R,S) = min(D(i,j)), i \epsilon R, j \epsilon S$$

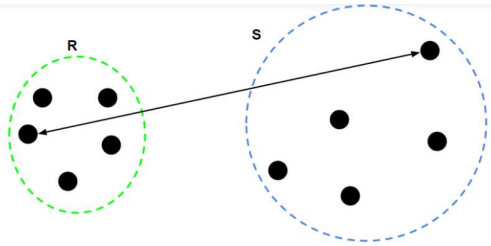


Complete linkage:

For two clusters R and S, the single linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S.

Formula:

$$L(R,S) = max(D(i,j)), i\epsilon R, j\epsilon S$$



Average Linkage:

For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean.

Formula:

$$L(R,S) = \frac{1}{n_R+n_S} \sum_{i=1}^{n_R} \sum_{j=1}^{n_S} D(i,j), i\epsilon R, j\epsilon S$$

nR-> Number of data points of cluster R

nS-> Number of data points of cluster S