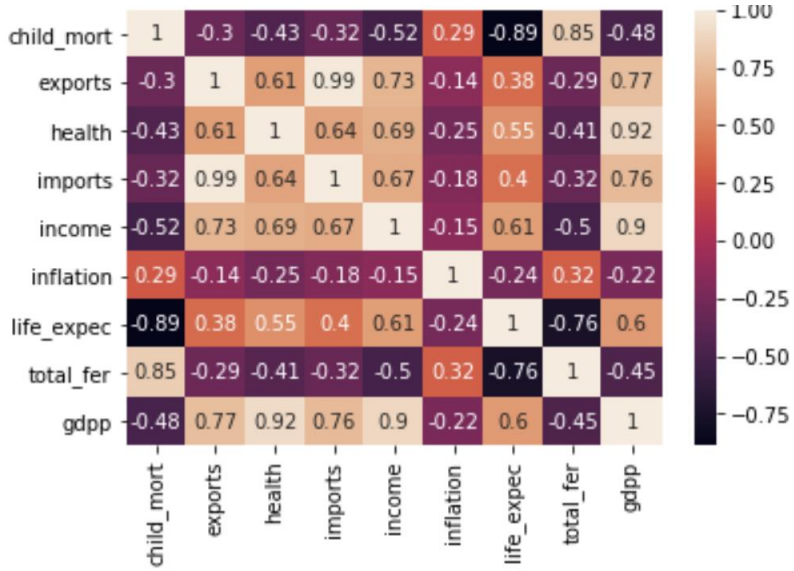
Question 1: Assignment Summary

Problem Statement:

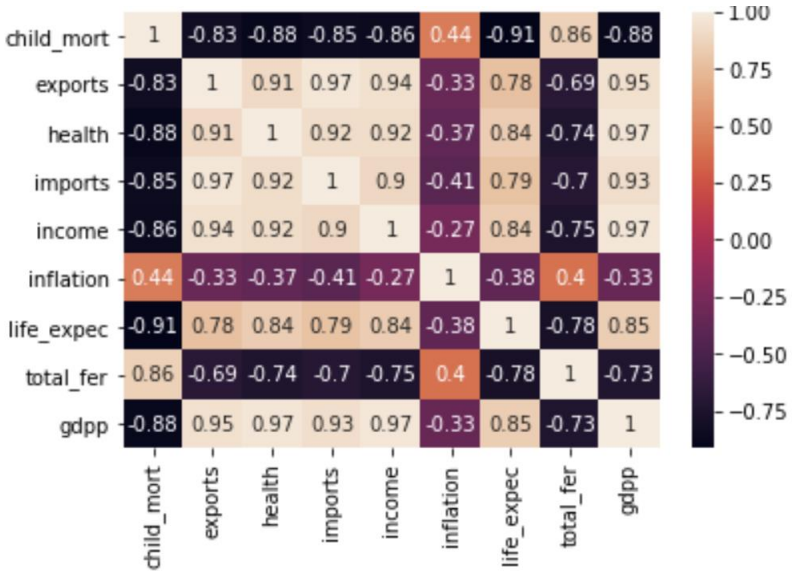
NGO's has raised $ 10 million funds in funding programs to help poor nations.The CEO of the NGO needs to decide how to use this money on countries that are in the direst need of aid.

Visualization of the Countries: Correlation of the variables from the heatmap and pair plot.

1. Some of the variables are having high positive correlation with gdpp vs (income, health, imports) etc.
2. Some of them have high negative correlation like child mortality and life expectancy & gdpp vs health etc.
3. This shows that the dataset is having multicollinearity.



Transformation: Data is skewed towards right. So data is distributed using power transformation.



Model is built by k-means and hierarchical clustering using 3 clusters. Country who needs immediate aid is identified based on key columns with child mortality, low income and low gdpp. Haiti, Sierra Leone, Chad, Central African Republic, Mali are countries who requires immediate support.

Question 2: Clustering

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

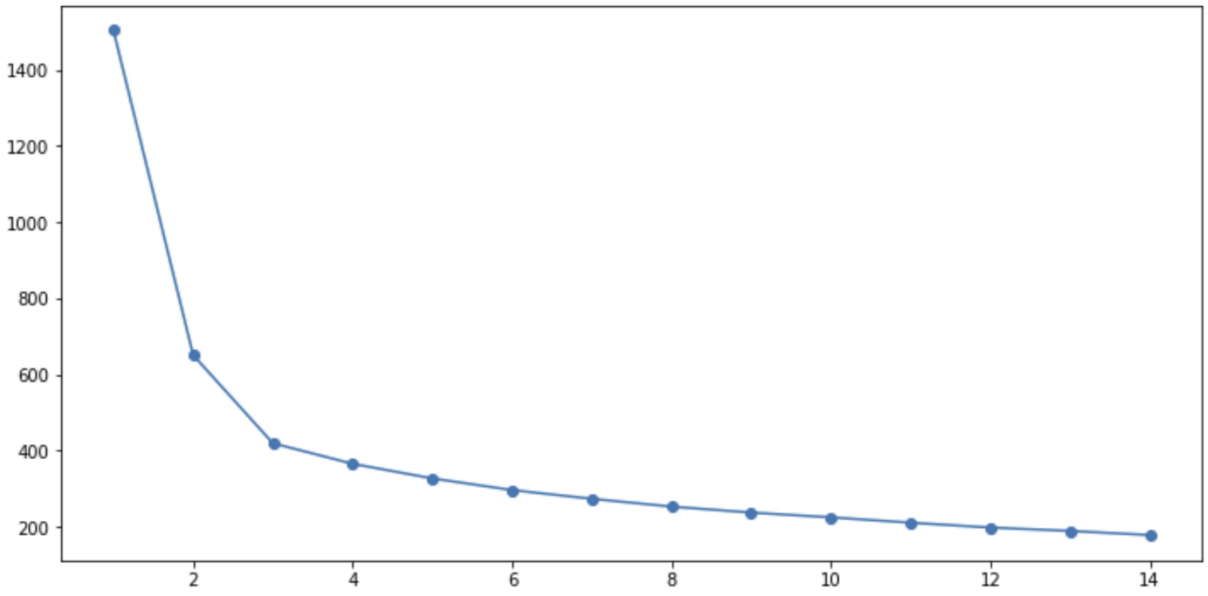
- K- means clustering a simply a division of the set of data objects into non- overlapping subsets (clusters) such that each data object is in exactly one subset). A hierarchical clustering is a set of nested clusters that are arranged as a tree.
- K Means clustering needed advance knowledge of K i.e. no. of clusters one want to divide your data. In hierarchical clustering one can stop at any number of clusters, one find appropriate by interpreting the dendrogram.

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

- Randomly select *'c'* cluster centers.
- Calculate the distance between each data point and cluster centers.
- Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
- Recalculate the new cluster center using:
- Where, *'c_i'* represents the number of data points in *i^{th}* cluster.
- $$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$
- Recalculate the distance between each data point and new obtained cluster centers.
- If no data point was reassigned then stop, otherwise repeat from step 3).

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

Optimum value of the k-value is determined based on elbow method. Based on below graph elbow is forming at K=3 because there is bend at 3. So the optimal value will be 3 for performing K-Means. Later check the distribution of data based on pair plot and box plot.



We are trying to find the optimum value of the k-value based on the business requirements.

So, to achieve this we used silhouette analysis to find the score of range of cluster values from 2 to 16.

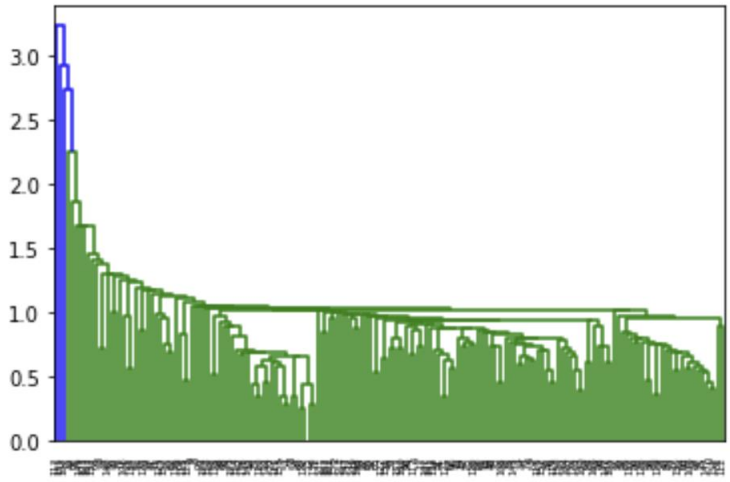
We found below silhouette scores are varied from 0.44351069200040283 to 0.21514790363091

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

When we standardize the data prior to performing cluster analysis, the clusters change. We find that with more equal scales more significantly contributes to defining the clusters. Standardization prevents variables with larger scales from dominating how clusters are define.

e) Explain the different linkages used in Hierarchical Clustering.

Single-Linkage: Single-linkage (nearest neighbor) 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.



Agglomerative Hierarchical Clustering with Complete Linkage: Complete-Linkage: Complete-linkage (farthest neighbor) 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. Along with average-linkage, it is one of the more popular distance metrics.