Clustering Assignment

Gaurav Garg

Problem Statement

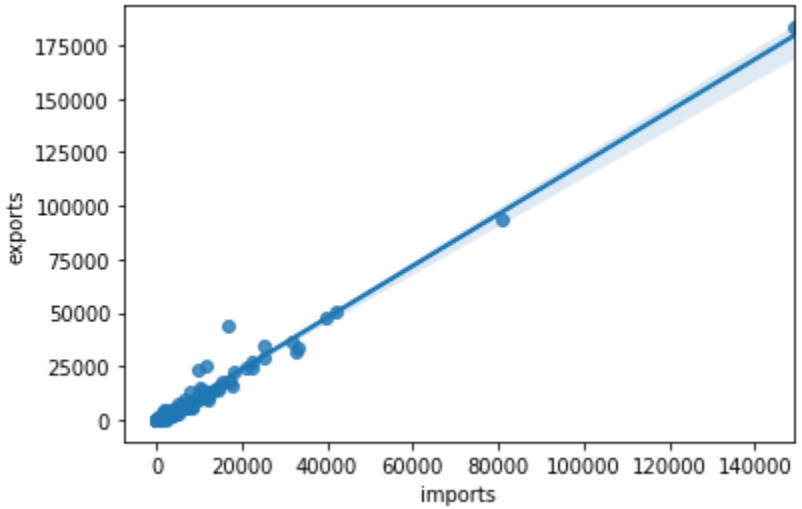
An International humanitarian NGO raised around $10 million in recent funding. The CEO needs to decide how to use this money and which countries to help. From the analysis we need to find out countries which are in direst need of financial aid.

Data Understanding and Analysis

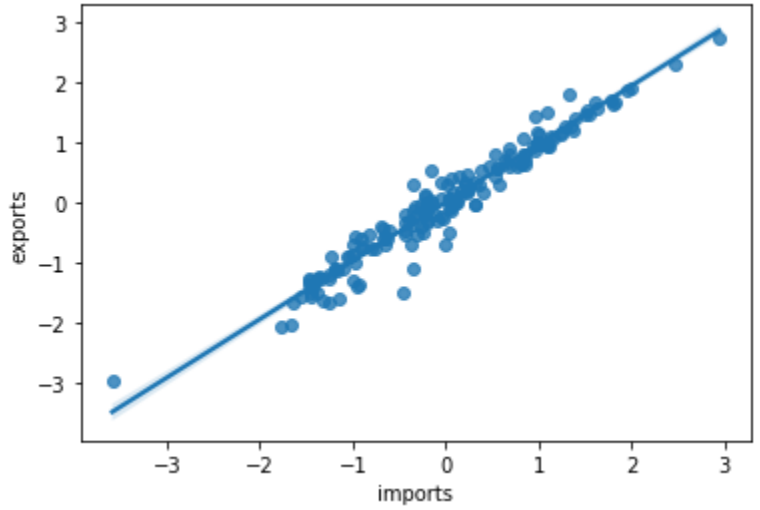
- The data we have, has details of 167 countries across the world. It has details like child mortality, value of total exports, imports, health expenditure, GDP, etc.
- Imports, exports and health are given as percentage of GDP per capita, so first step is to convert them to actual values.
- After plotting seaborn pair plot between these variables we see that their individual distribution plot suggest that in most of the columns like income, imports, exports, gdp the values are left skewed i.e. there are many countries within lower ranges of these columns.
- From the heat map we saw that imports and exports are highly correlated, similarly life_expec and child_mort also have high negative correlation coefficient
- There are no null/NA values in any of the columns

Transformation to log values

- As most of the variable's values are left skewed, if we convert them into log values we would get better dependency between the variables.

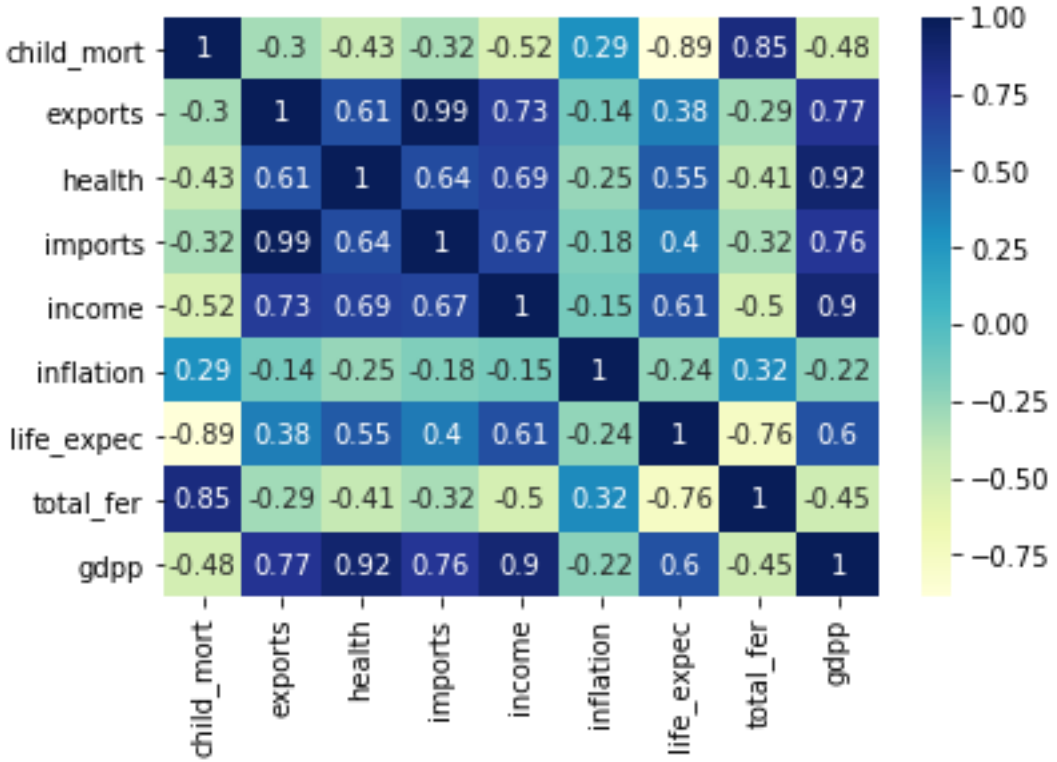


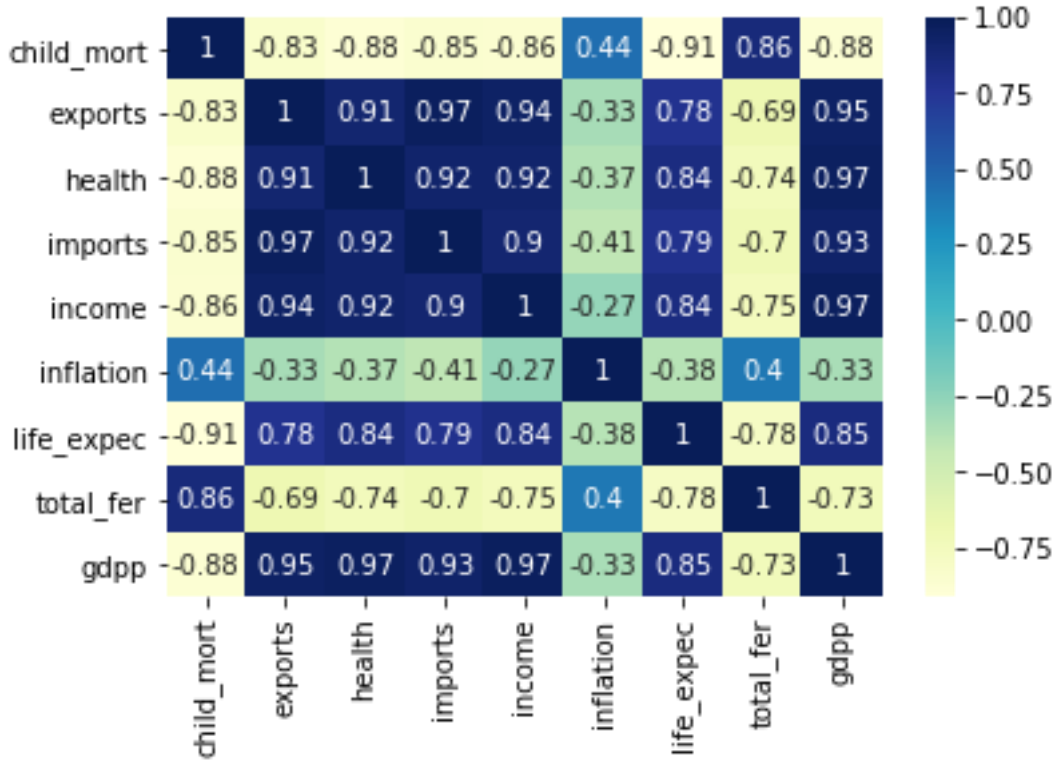
Before PowerTransformation



After PowerTranformation

Heat Map


child_mort 1 -0.3 -0.43 -0.32 -0.52 0.29 -0.89 0.85 -0.48
exports -0.3 1 0.61 0.99 0.73 -0.14 0.38 -0.29 0.77
health -0.43 0.61 1 0.64 0.69 -0.25 0.55 -0.41 0.92
imports -0.32 0.99 0.64 1 0.67 -0.18 0.4 -0.32 0.76
income -0.52 0.73 0.69 0.67 1 -0.15 0.61 -0.5 0.9
inflation 0.29 -0.14 -0.25 -0.18 -0.15 1 -0.24 0.32 -0.22
life_expec -0.89 0.38 0.55 0.4 0.61 -0.24 1 -0.76 0.6
total_fer 0.85 -0.29 -0.41 -0.32 -0.5 0.32 -0.76 1 -0.45
gdpp -0.48 0.77 0.92 0.76 0.9 -0.22 0.6 -0.45 1
child_mort exports health imports income inflation life_expec total_fer gdpp
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75


child_mort 1 -0.83 -0.88 -0.85 -0.86 0.44 -0.91 0.86 -0.88
exports -0.83 1 0.91 0.97 0.94 -0.33 0.78 -0.69 0.95
health -0.88 0.91 1 0.92 0.92 -0.37 0.84 -0.74 0.97
imports -0.85 0.97 0.92 1 0.9 -0.41 0.79 -0.7 0.93
income -0.86 0.94 0.92 0.9 1 -0.27 0.84 -0.75 0.97
inflation 0.44 -0.33 -0.37 -0.41 -0.27 1 -0.38 0.4 -0.33
life_expec -0.91 0.78 0.84 0.79 0.84 -0.38 1 -0.78 0.85
total_fer 0.86 -0.69 -0.74 -0.7 -0.75 0.4 -0.78 1 -0.73
gdpp -0.88 0.95 0.97 0.93 0.97 -0.33 0.85 -0.73 1
child_mort exports health imports income inflation life_expec total_fer gdpp
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75

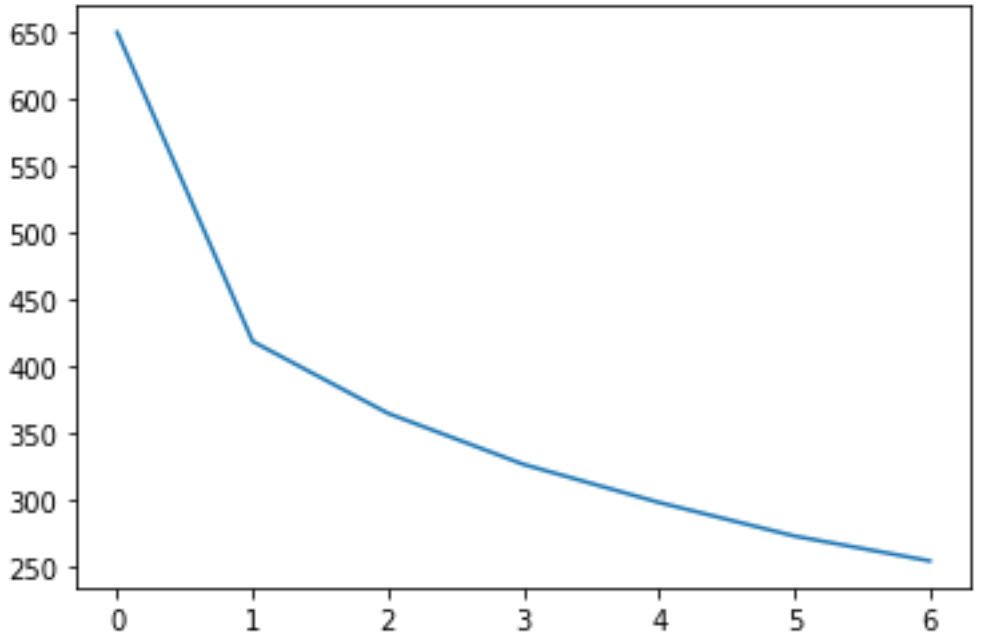
Modeling

- K-Means and Hierarchical clustering both are used for clustering the dataset.
- Almost similar results are obtained from both.
- SSD and Silhouette Analysis are used to find out the optimal value of K.
- Hopkins value obtained is around 0.80-0.85

Sum of Square Differences

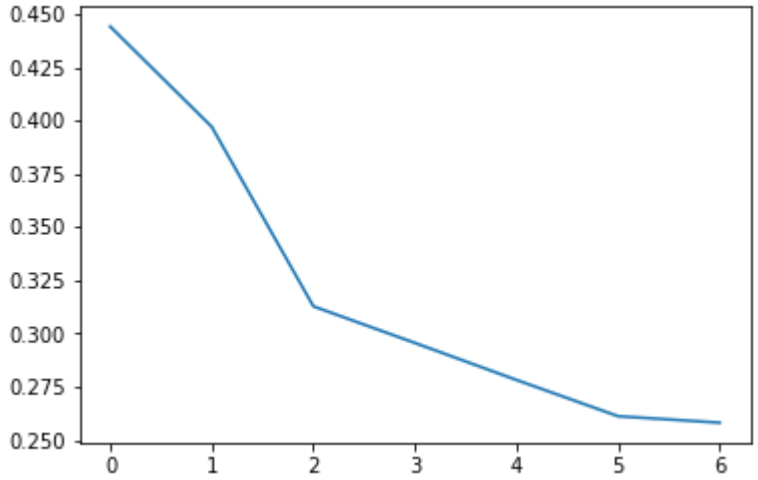
Below Curve is plotted for K=2 to K=8, in the below curve 0 represents 2, 1 represents 3 and so on.

From this curve, we can say that there is noticeably larger drop when value of K moves from 2 to 3 and K decreases less gradually after that.



The Silhouette Analysis

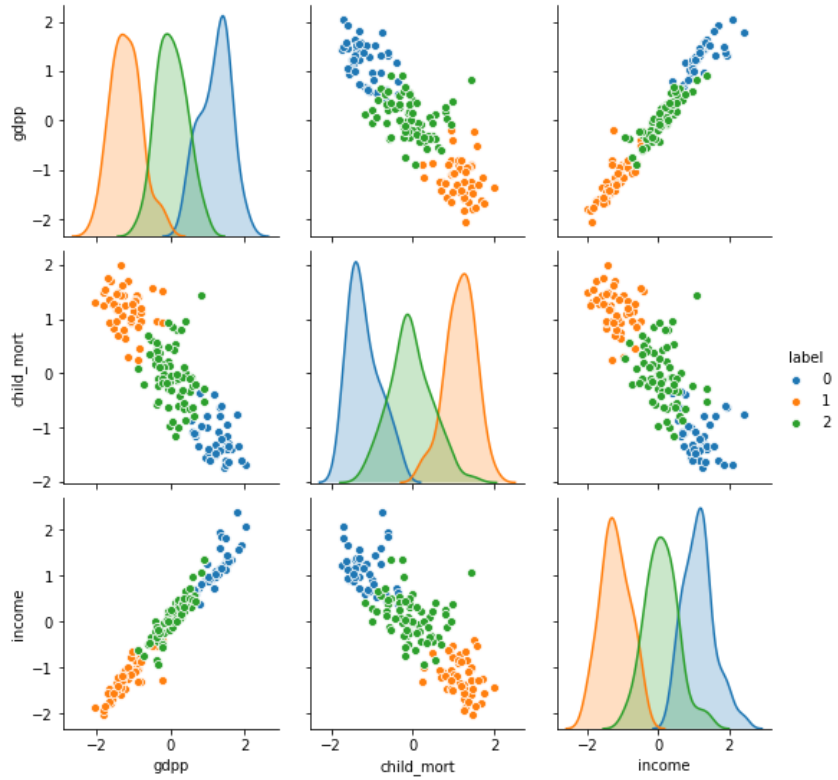
- Silhouette Analysis also supports the idea of K=3 or K=2.
- Silhouette score is large for K=2, but we want a smaller set of countries to be able to predict which are direst need, therefore K=3 is finally chosen.




```
For n_clusters=2, the silhouette score is 0.44420866548727334
For n_clusters=3, the silhouette score is 0.3970904804043125
For n_clusters=4, the silhouette score is 0.3127821211075455
For n_clusters=5, the silhouette score is 0.2955725973648449
For n_clusters=6, the silhouette score is 0.2781808343287792
For n_clusters=7, the silhouette score is 0.2611591377487346
For n_clusters=8, the silhouette score is 0.2581258675958351
```

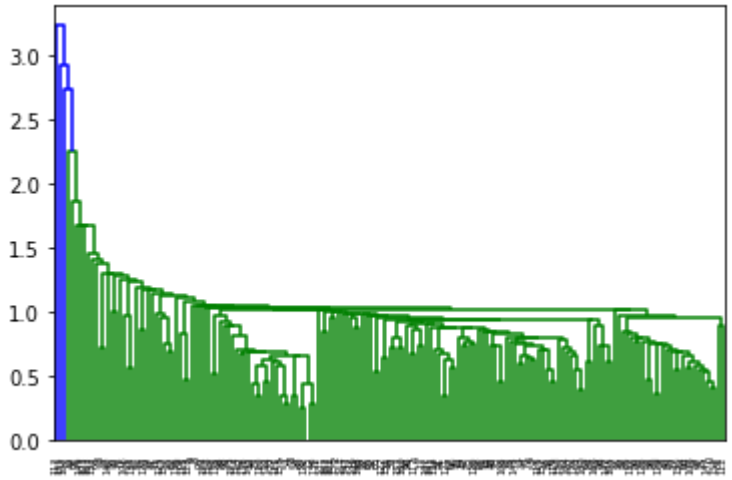
K-Means Clustering

- K-means from sklearn.cluster library is used for K-means algorithm
- From the obtained result from K=3, we see that K-means clustering has perfectly separated out countries based on the three desired variables of our analysis.



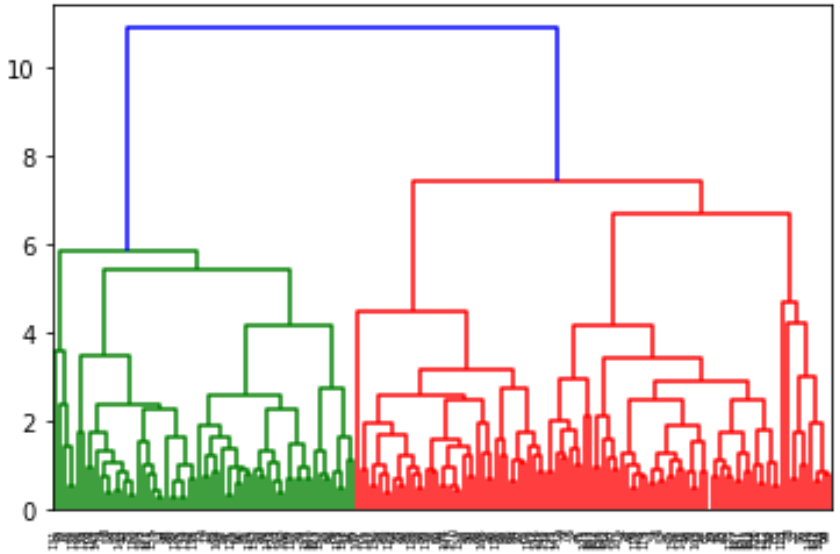
Hierarchical Clustering –single linkage

- Again linkage, dendrogram and cut_tree libraries from scipy.cluster.hierarchy are chosen for the analysis
- Single Linkage could not help much in prediction.
- Below is the obtained dendrogram



Hierarchical Clustering - Complete linkage

- Complete linkage was very helpful in correct prediction.
- Predictions from complete linkage were similar to K-means only difference was complete linkage predicted 17 more countries in the desired cluster (with high child_mort, low income and low gdpp).
- Therefore finally, K-means predictions were chosen for results.



Result – final list of countries

- Below is the final list of countries chosen to be given to CEO to provide financial aid.
- From the 50 list of countries obtained from k-means clustering, top 10 are chosen based on the highest child_mort rate since child mortality is a very serious issue of a country and can be taken as most important.

```
66                      Haiti
132               Sierra Leone
32                        Chad
31     Central African Republic
97                        Mali
113                    Nigeria
112                      Niger
3                       Angola
37             Congo, Dem. Rep.
25                 Burkina Faso
Name: country, dtype: object
```