UNDER DEVELOPED COUNTRIES

in need for Aid

PROBLEM STATEMENT & ANALYSIS APPROACH

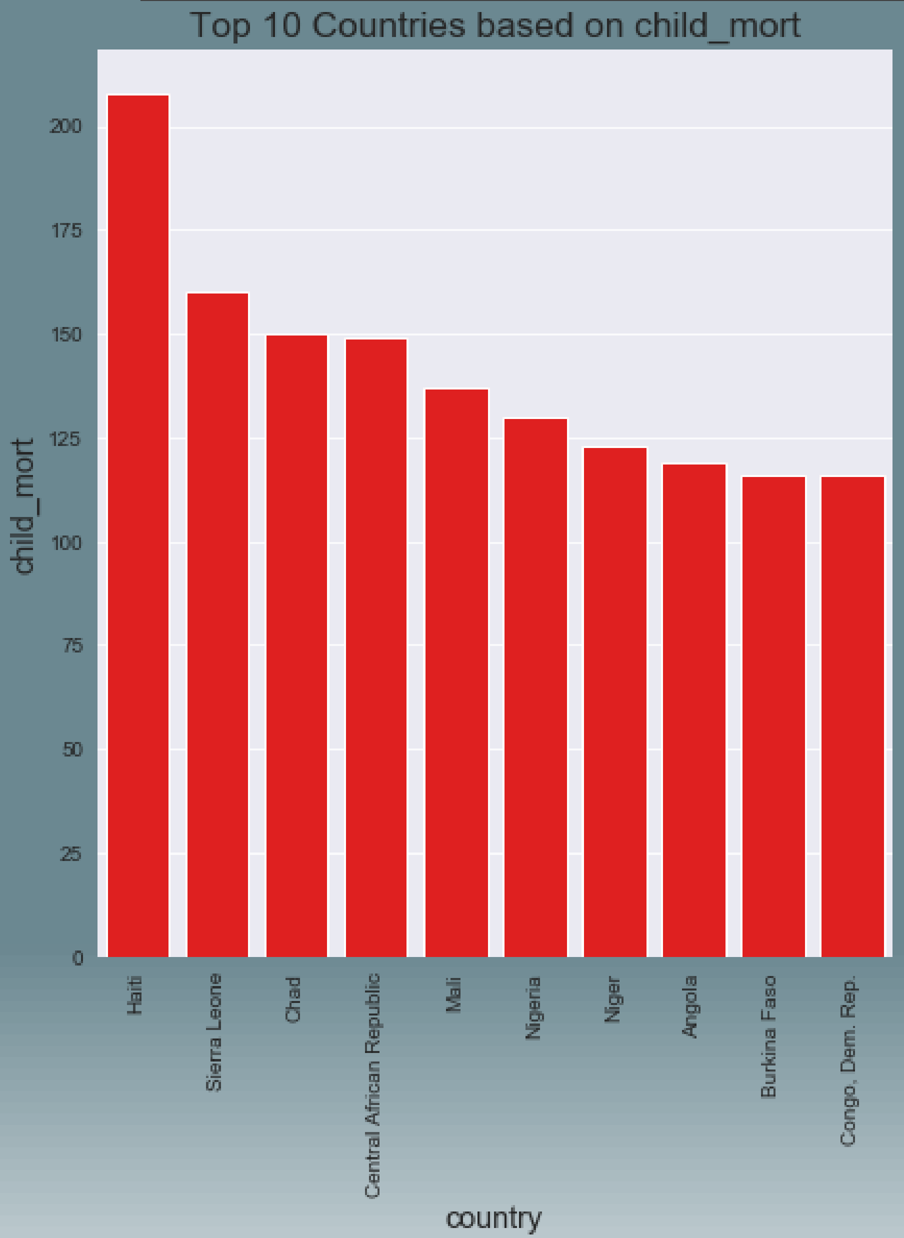
PROBLEM STATEMENT

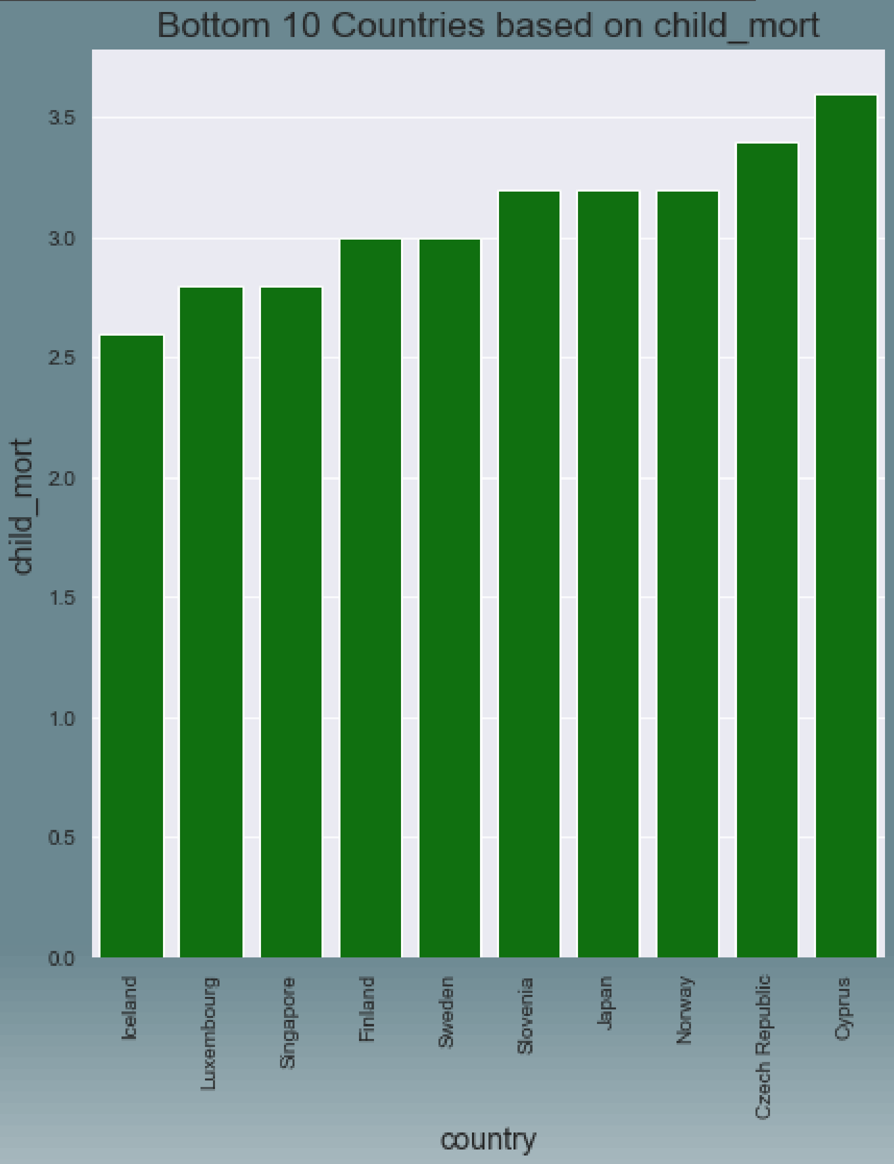
- The CEO of the NGO (HELP International is an international humanitarian NGO) needs to decide how to use the raised $ 10 million strategically and effectively on the countries that are in direst need of aid.
- Choosing those countries is very essential.

ANALYSIS APPROACH

- Understanding and performing Exploratory Data Analysis on the data related to countries around the world based on the factors, such as gdpp, income, child mortality, etc.
- Extract maximum information from all the features and create different clusters for different types of countries, such as Under-Developed, Developing, and Developed Countries.
- Validate the clusters so-found and extract the top-n under-developed countries that require aid.

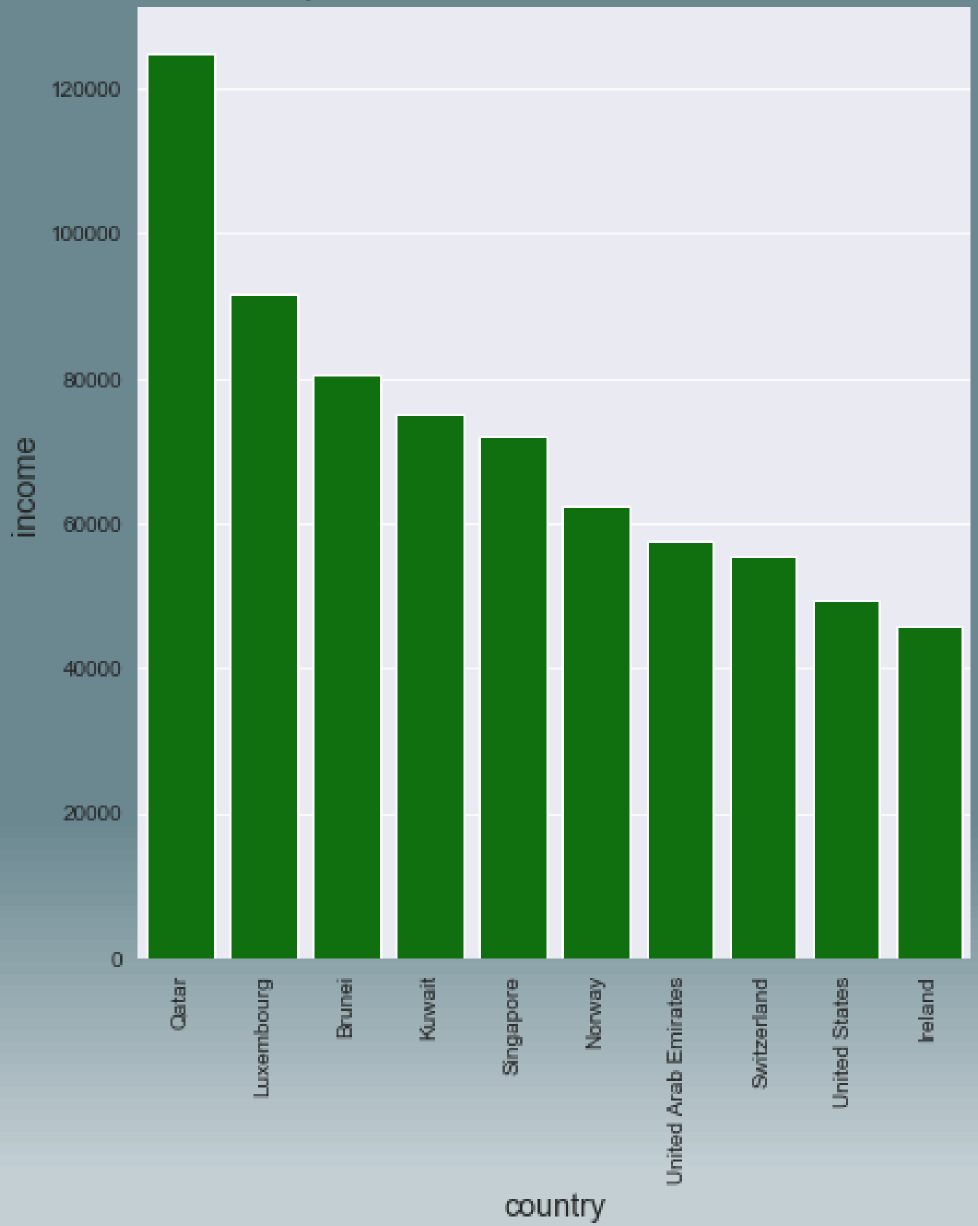
TOP 10 & BOTTOM 10 COUNTRIES BASED ON CHILD MORTALITY



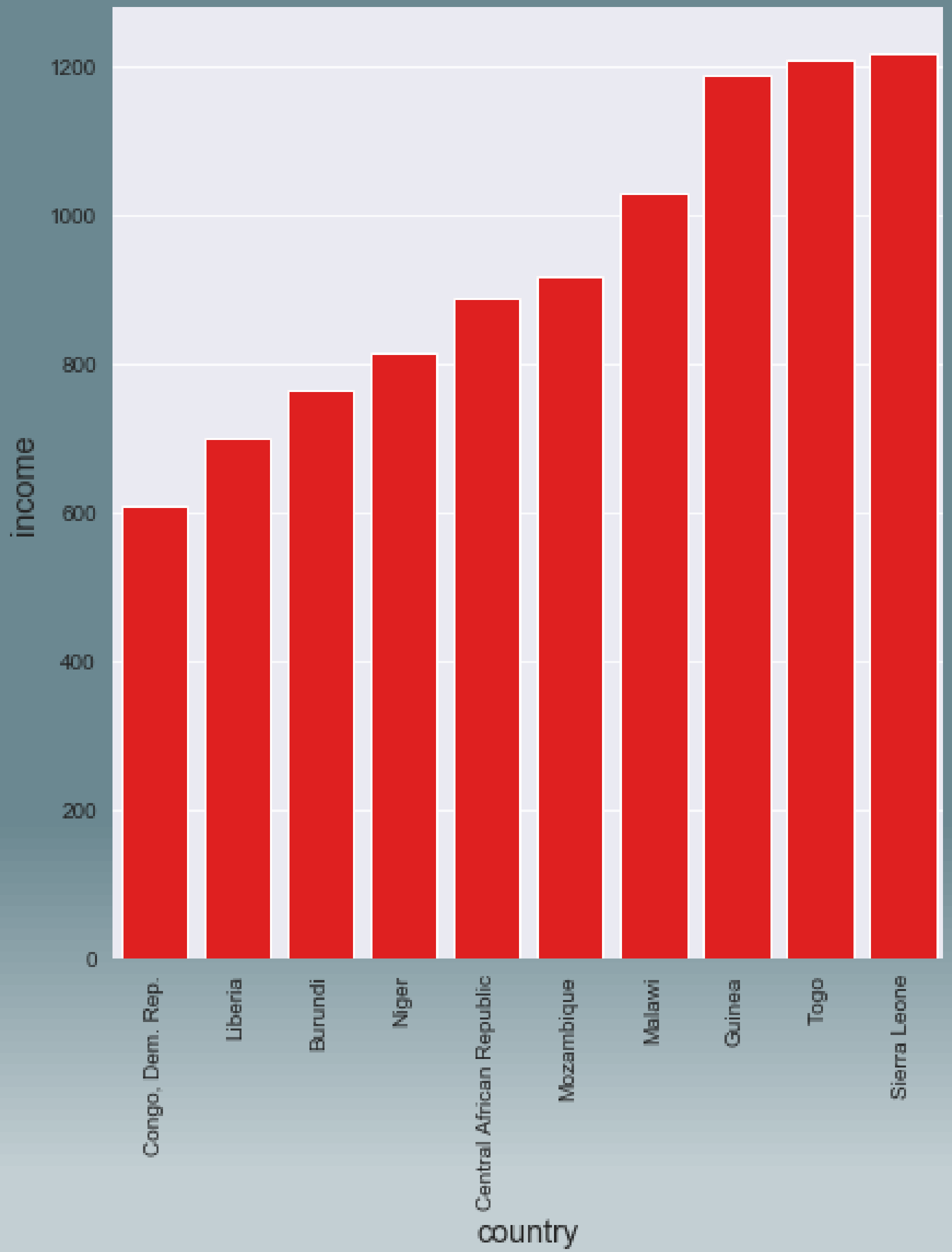


TOP 10 & BOTTOM 10 COUNTRIES BASED ON INCOME

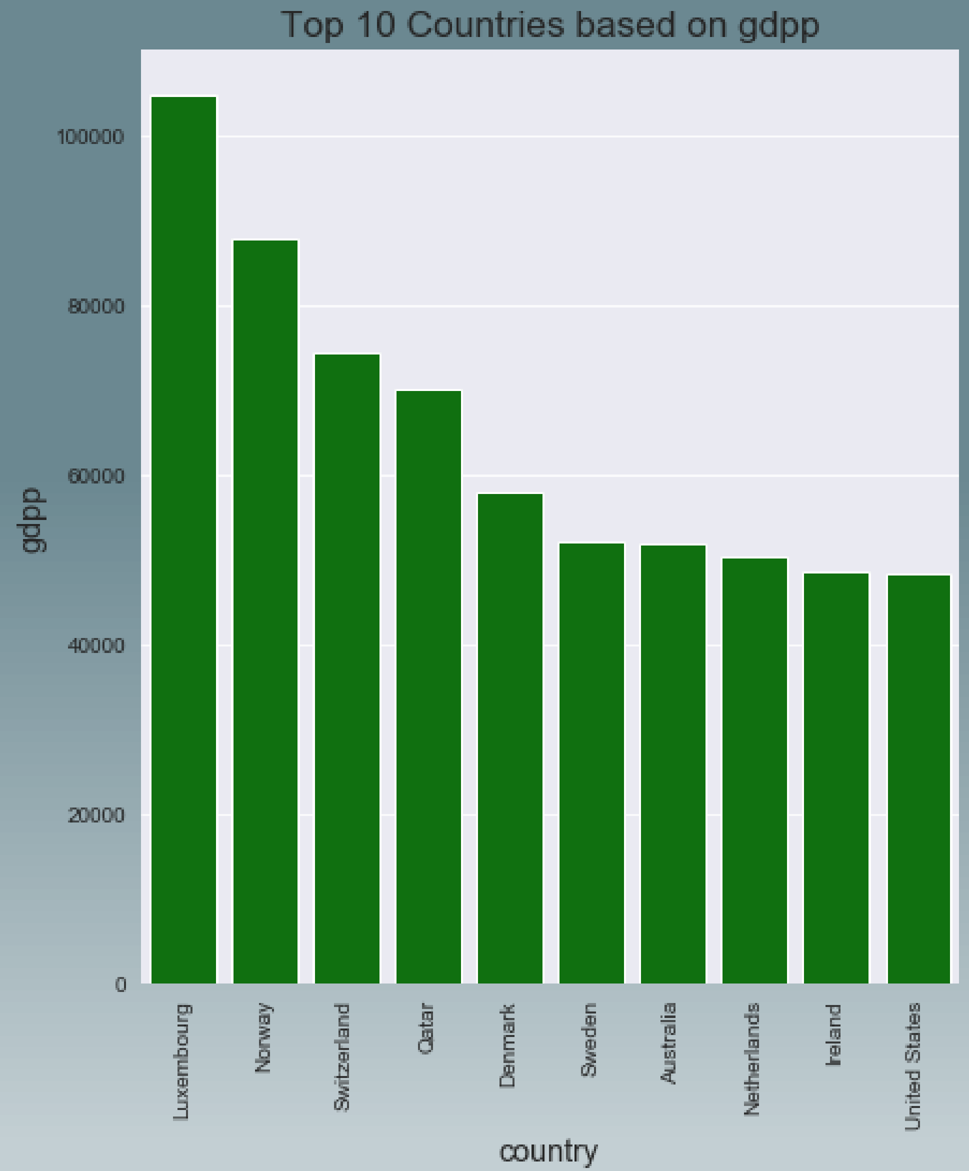
Top 10 Countries based on income

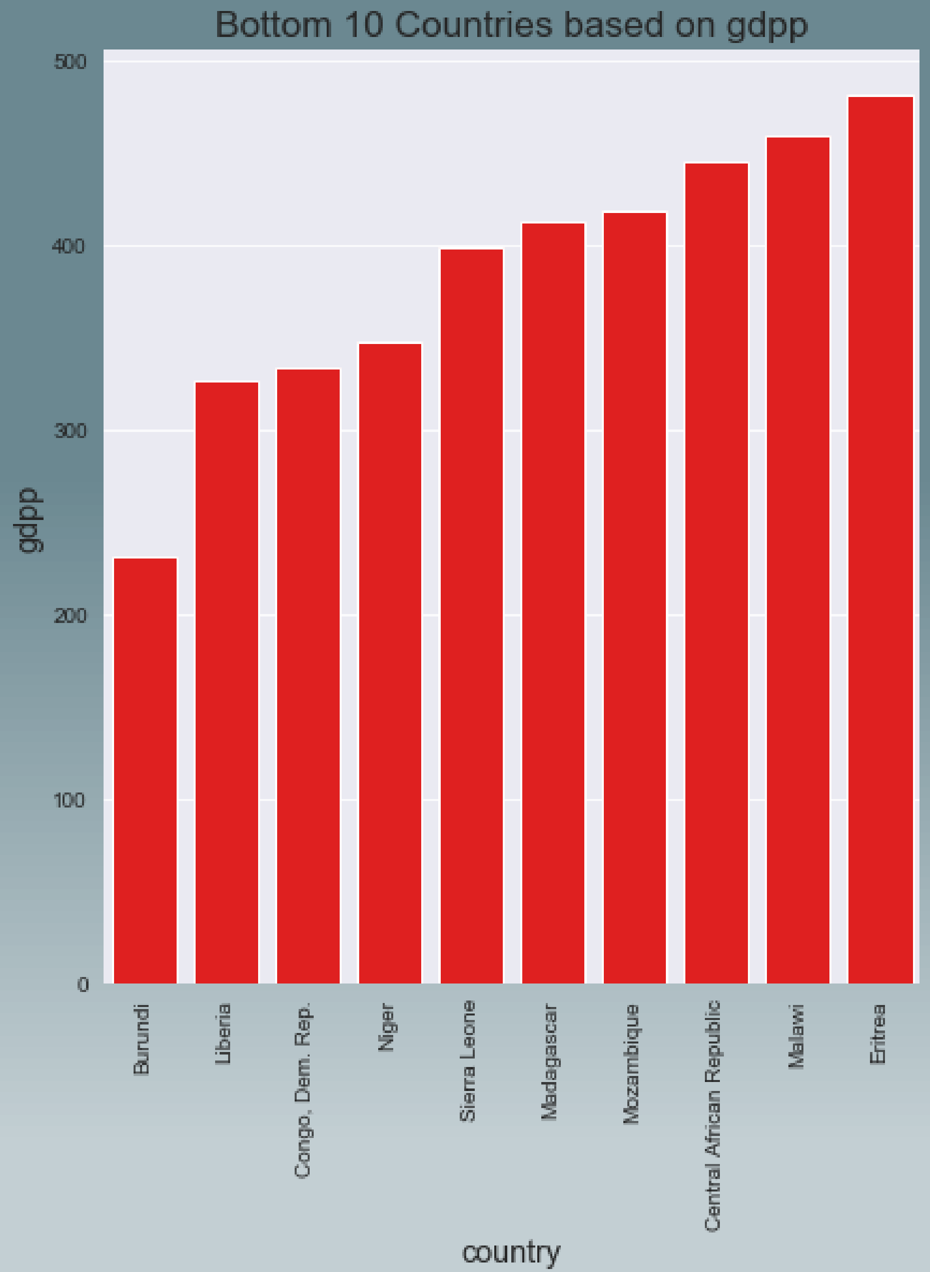


Bottom 10 Countries based on income



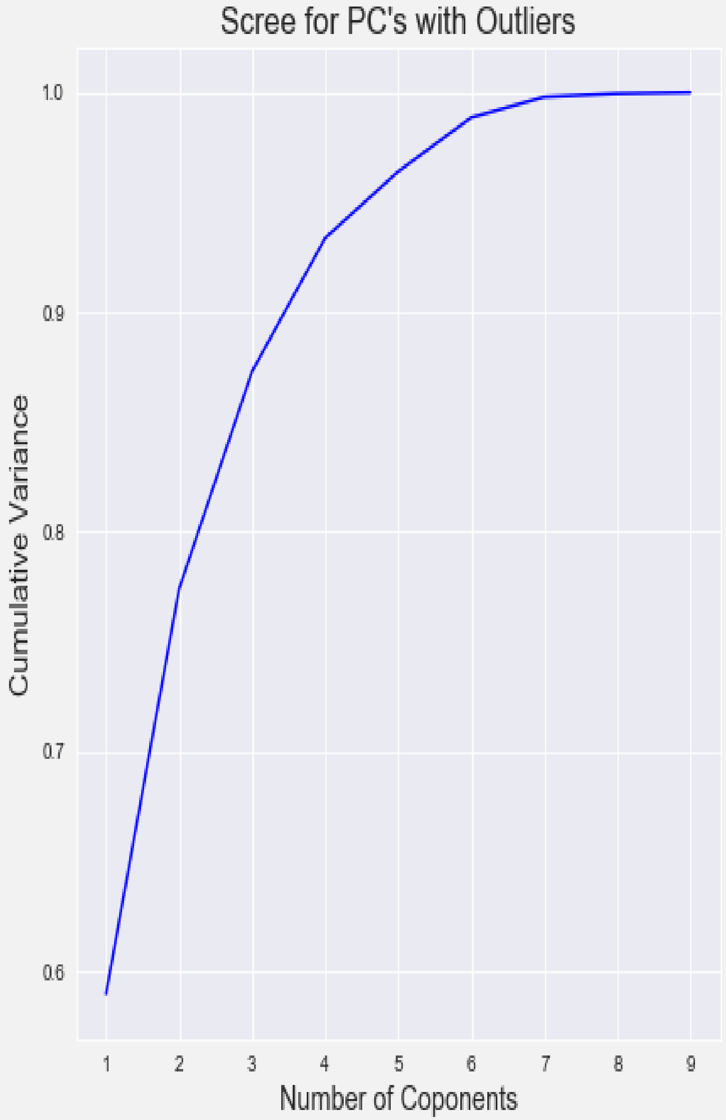
TOP 10 & BOTTOM 10 COUNTRIES BASED ON GDPP

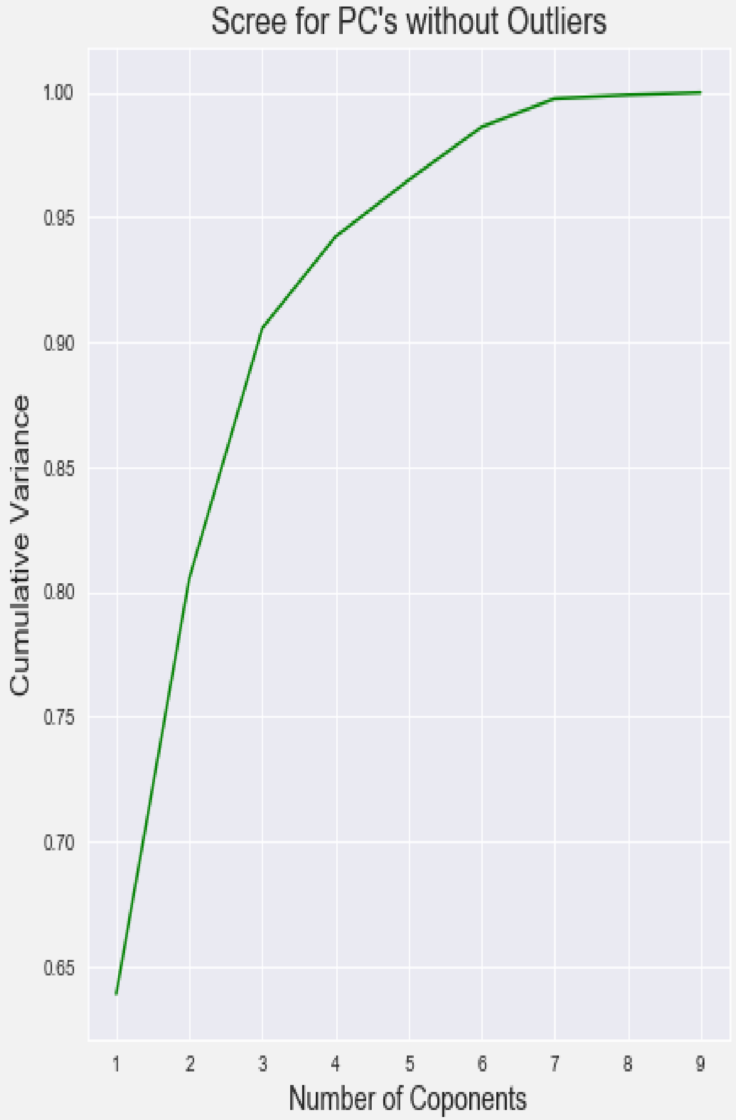




PRINCIPAL COMPONENT ANALYSIS

- PCA applied on two different data sets. One with the outliers and one after treating the outliers.
- The scree plot for both looks quite similar.
- For both, 5 PC's were used, since in both the cases they explain around 96% of the total variance in the data. Post 5 components, the slope in both the cases starts leveling off.

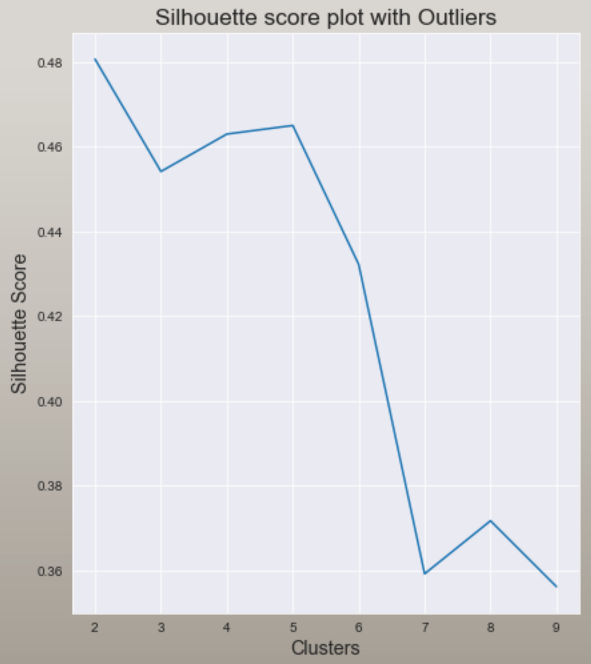

Scree for PC's with Outliers

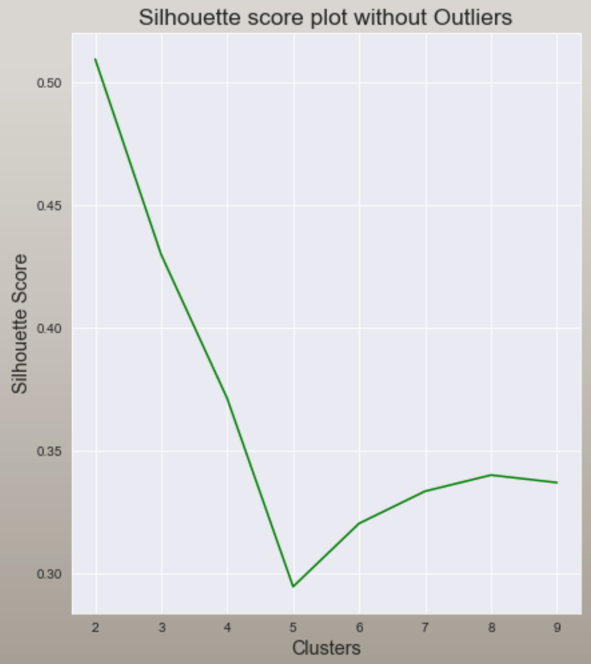

Scree for PC's without Outliers

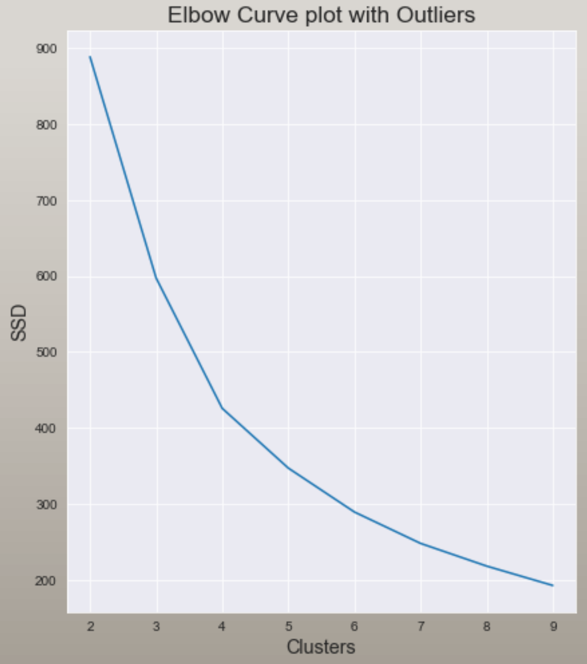
CLUSTERING

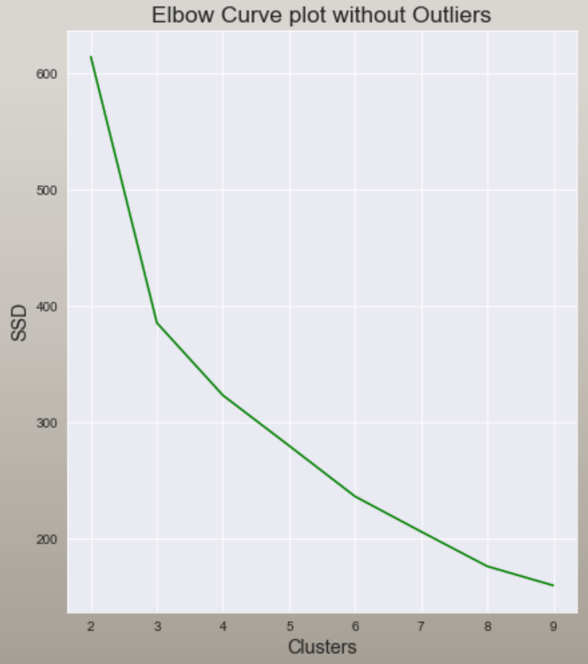
- Two different types of clustering were performed on the PCA data sets (one with outliers and one without the outliers)
- **1st one is K Means clustering**
  - To find the optimal number of cluster's for both the data sets (with and without outliers), the below two methods were chosen:
    - Silhouette Score
    - Elbow Curve Plot
- **2nd is Hierarchical Clustering**
  - Single Linkage
  - Complete Linkage

K-MEANS CLUSTERING


Silhouette score plot with Outliers


Silhouette score plot without Outliers


Elbow Curve plot with Outliers


Elbow Curve plot without Outliers

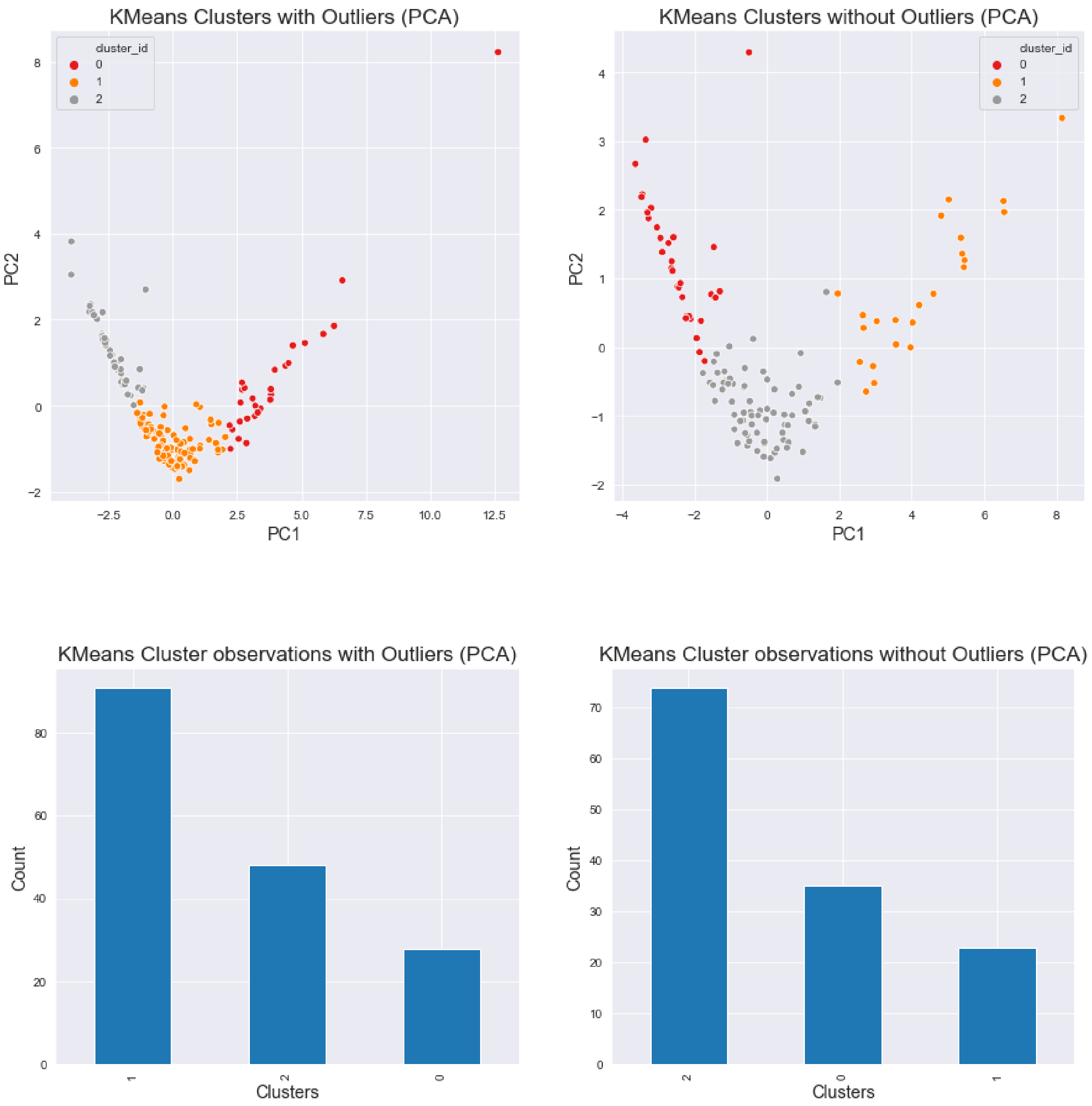
Based on the silhouette score:

- For data set with outliers, number of clusters can be 2, 3, and 4.
- For data set without outliers, number of clusters can also be 2,3, and 4, since they have good score.

From Elbow curve also, the potential number of clusters for both the data frames can be 2,3 and 4, but let's go with 3 for the following reasons:

- Considering 2 will just split the data into half, which will not be very useful.
- Considering 4 clusters will also not make much sense since the number of features are very less.
- Choosing 3 would help from the business perspective as 3 clusters could mean poor, average and good performing countries. (A potential cluster labeling.)

CLUSTERS OBTAINED FROM K-MEANS PERFORMED ON PCA'S

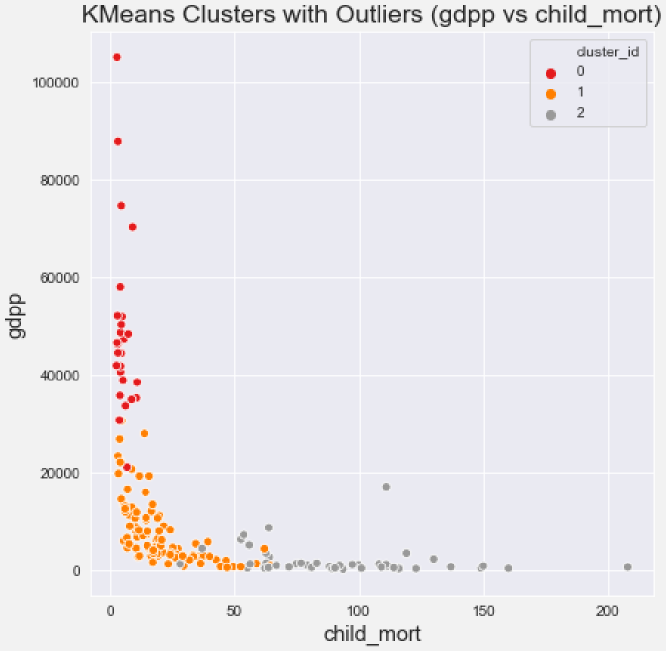

KMeans Clusters with Outliers (PCA)

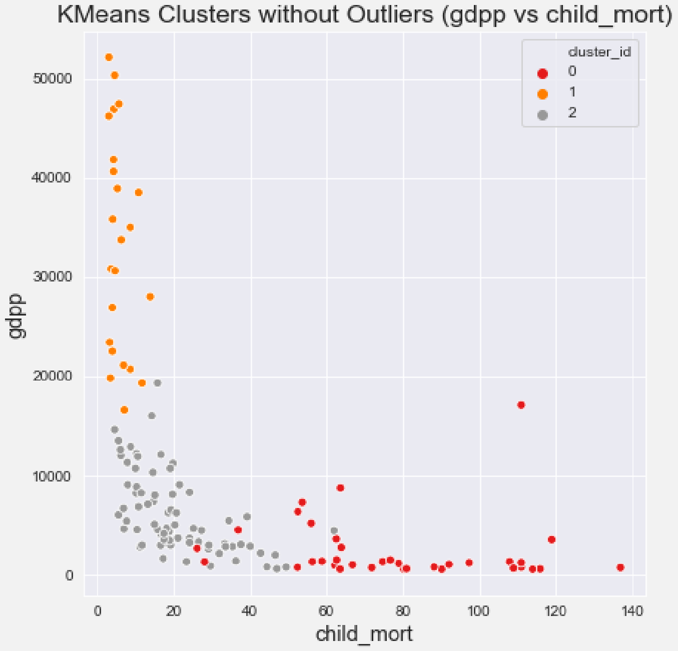
KMeans Clusters without Outliers (PCA)

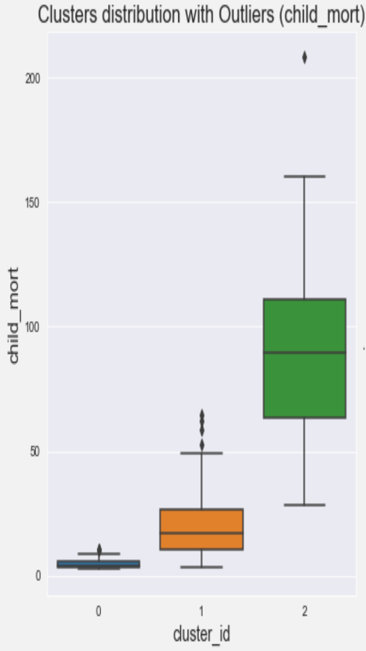
KMeans Cluster observations with Outliers (PCA)

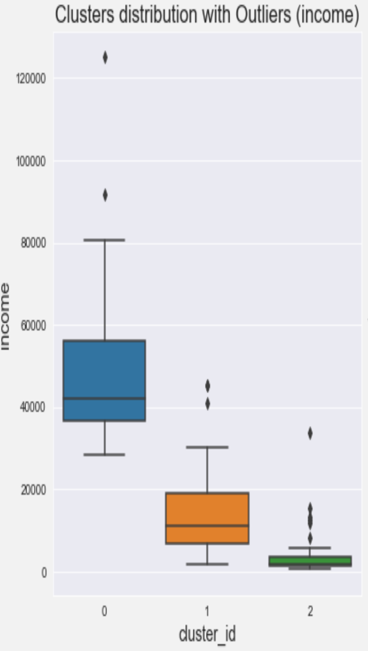
KMeans Cluster observations without Outliers (PCA)

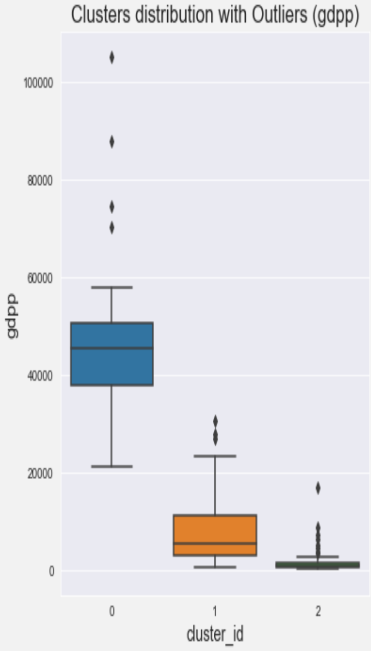
The Clusters so-created have good proportion of distribution.


KMeans Clusters with Outliers (gdpp vs child_mort)


KMeans Clusters without Outliers (gdpp vs child_mort)


Clusters distribution with Outliers (child_mort)


Clusters distribution with Outliers (income)

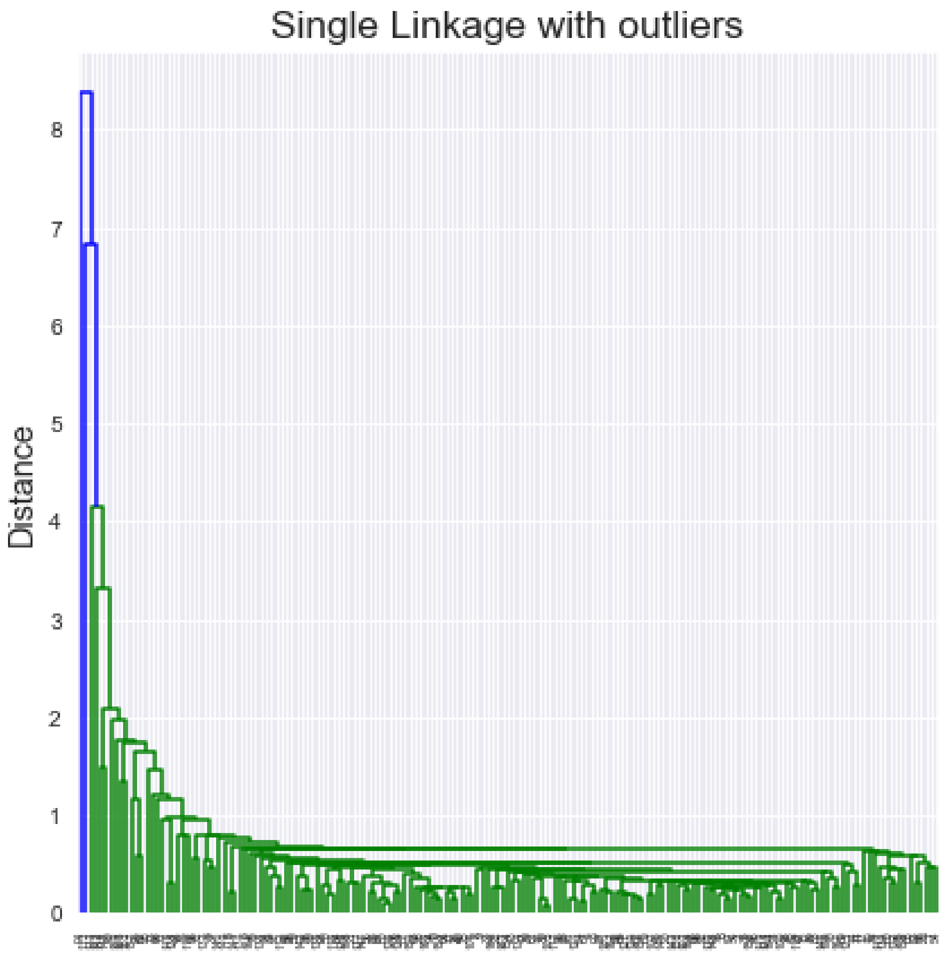

Clusters distribution with Outliers (gdpp)

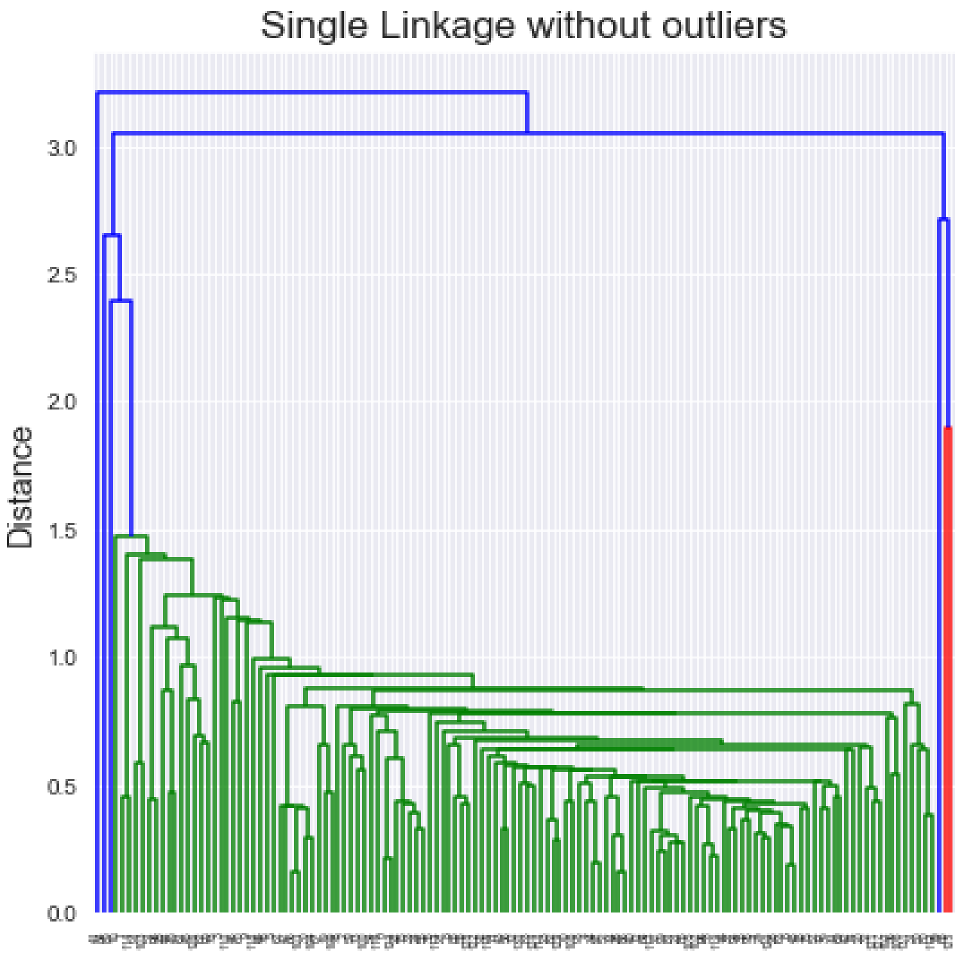
CLUSTER PROFILING (K-MEANS)

- Based on the above clustering for data frame with and without outliers, the clusters can be named as follows:
  - Developed Countries
  - Developing Countries
  - Under-developed Countries.
- The clusters for data set with outliers depicts the under developed ones to be shown by cluster no. 2, since
  - GDPP is very low
  - Child Mortality is very high

Inferences

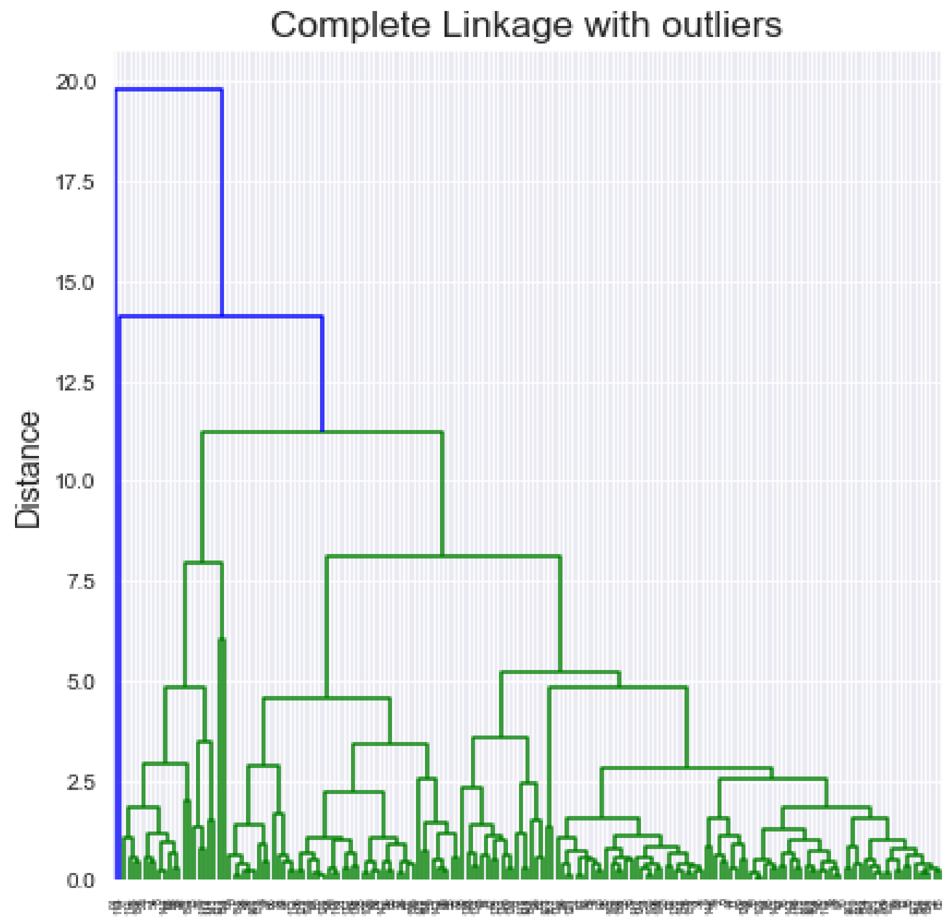
- **Cluster 0 (Developed)** : High gdpp, High income, Low child_mort
- **Cluster 1 (Developing):** Medium gdpp, Medium income, Medium child_mort
- **Cluster 2 (Under-developed):** Low gdpp, Low income, High child_mort

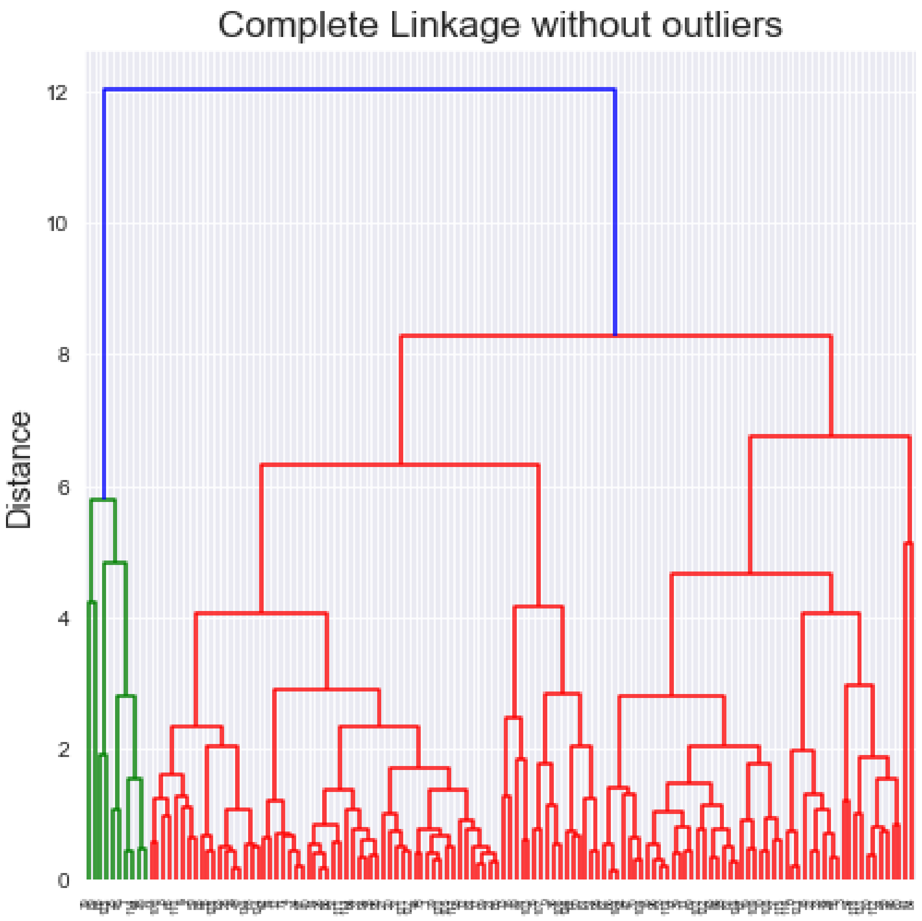




HIERARCHICAL CLUSTERING – SINGLE LINKAGE

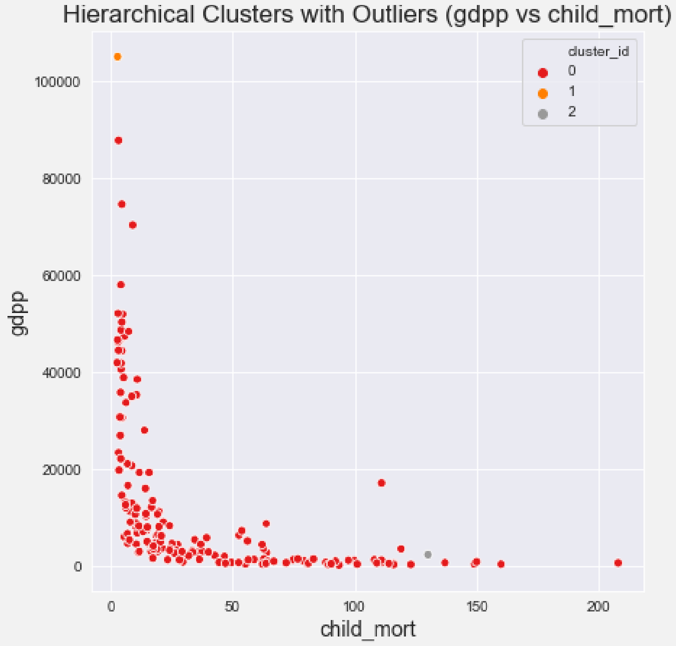
Single linkage didn't produce a good enough result for analyzing the clusters.



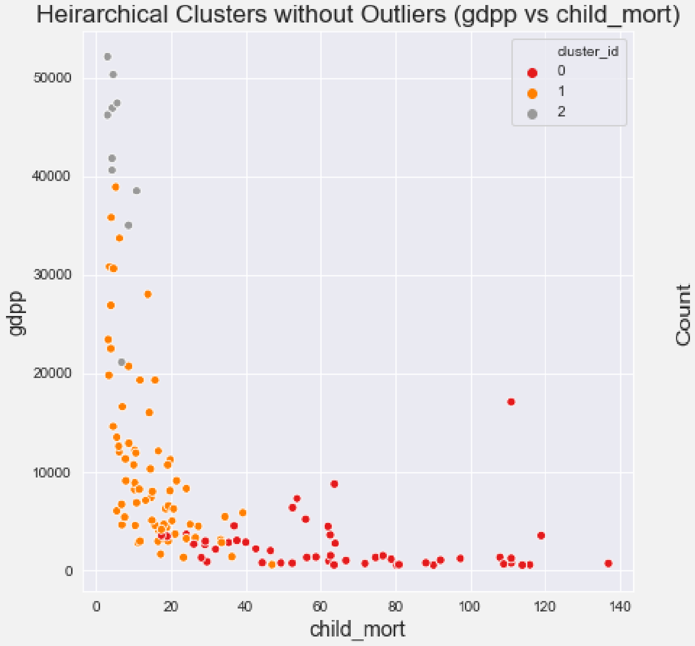


HIERARCHICAL CLUSTERING – COMPLETE LINKAGE

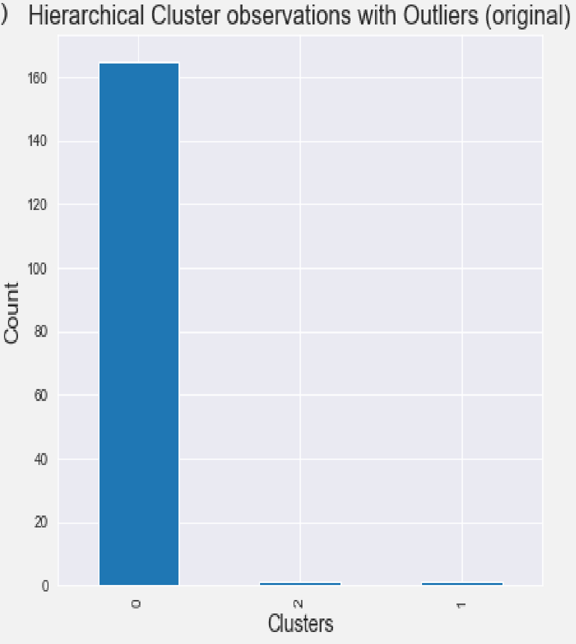
Complete linkage provides better results, when compared to single linkage
hence took 3 clusters as per the business sense.



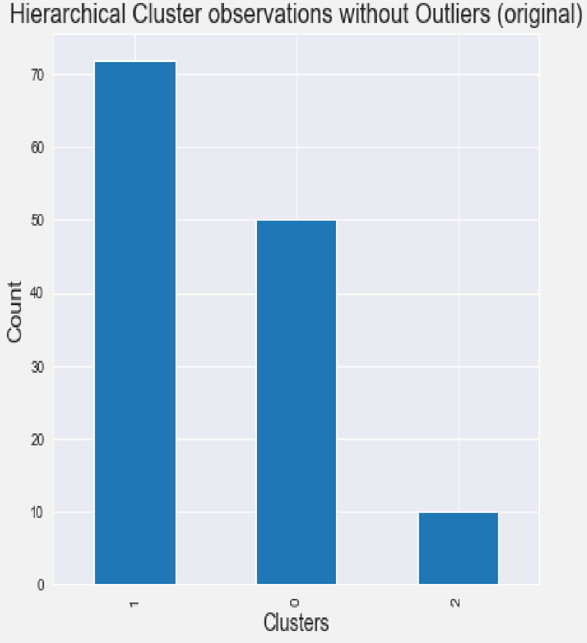
Hierarchical Clusters with Outliers (gdpp vs child_mort)



Heirarchical Clusters without Outliers (gdpp vs child_mort)



Hierarchical Cluster observations with Outliers (original)



Hierarchical Cluster observations without Outliers (original)

CLUSTER PROFILING (HIERARCHICAL-COMPLETE LINKAGE)

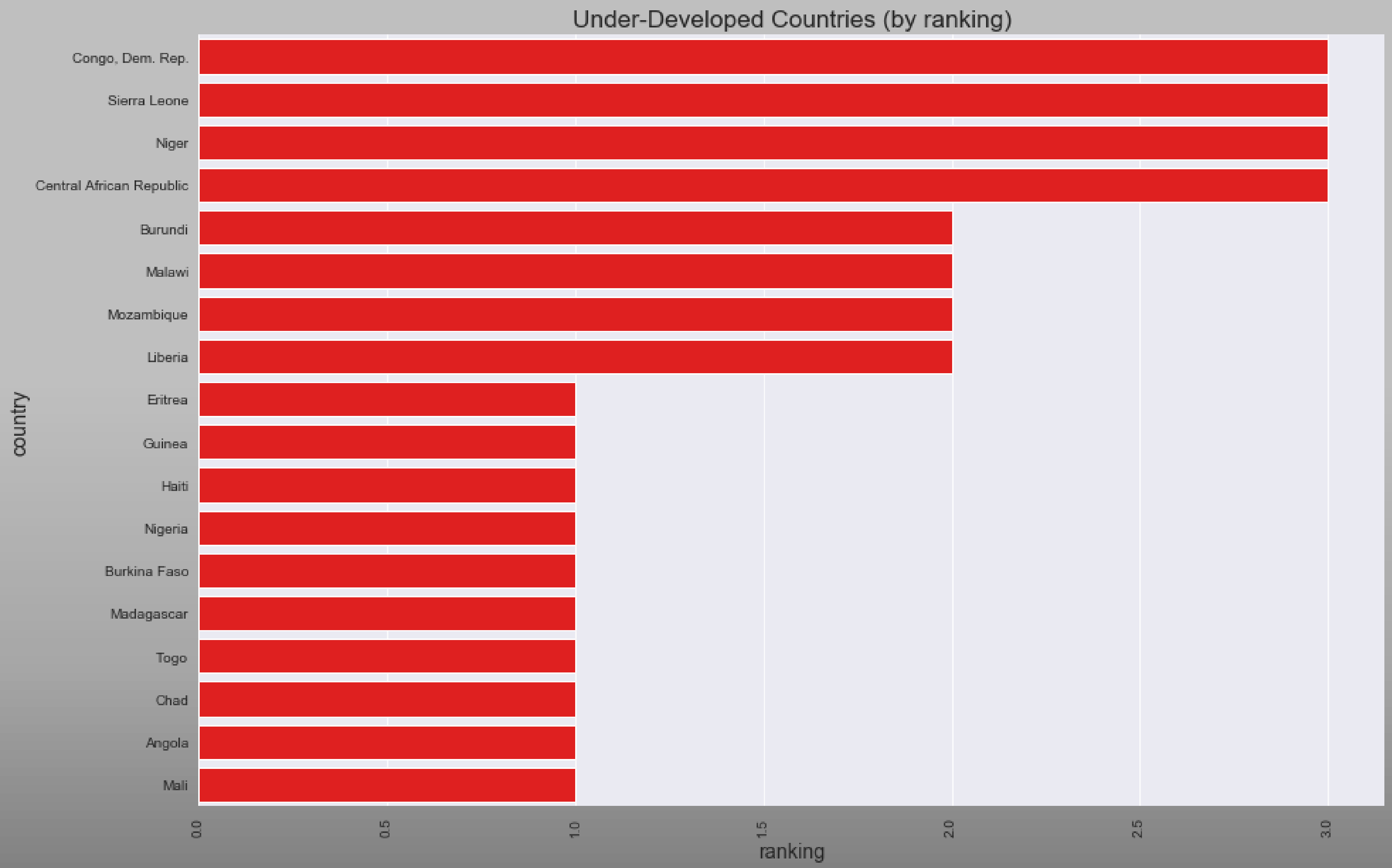
It can be observed (data set with outliers):

- Clusters created using hierarchical clustering (complete) doesn't cluster the countries in good proportion
- There is a lot of imbalance seen here.
- Almost all the countries are grouped into cluster number 0.

RECOMMENDATIONS

- KMeans provides better results when compared to Hierarchical (Complete Linkage) clustering.
- Removing outliers trimmed off few low performing countries, which were potentially the ones to be receiving aid the most.
- To go ahead with K-Means and with the data set containing the outliers. Since the data is less and each data represents a single countries data, removing them might remove few countries that could be in the direst need of aid.

UNDER DEVELOPED COUNTRIES

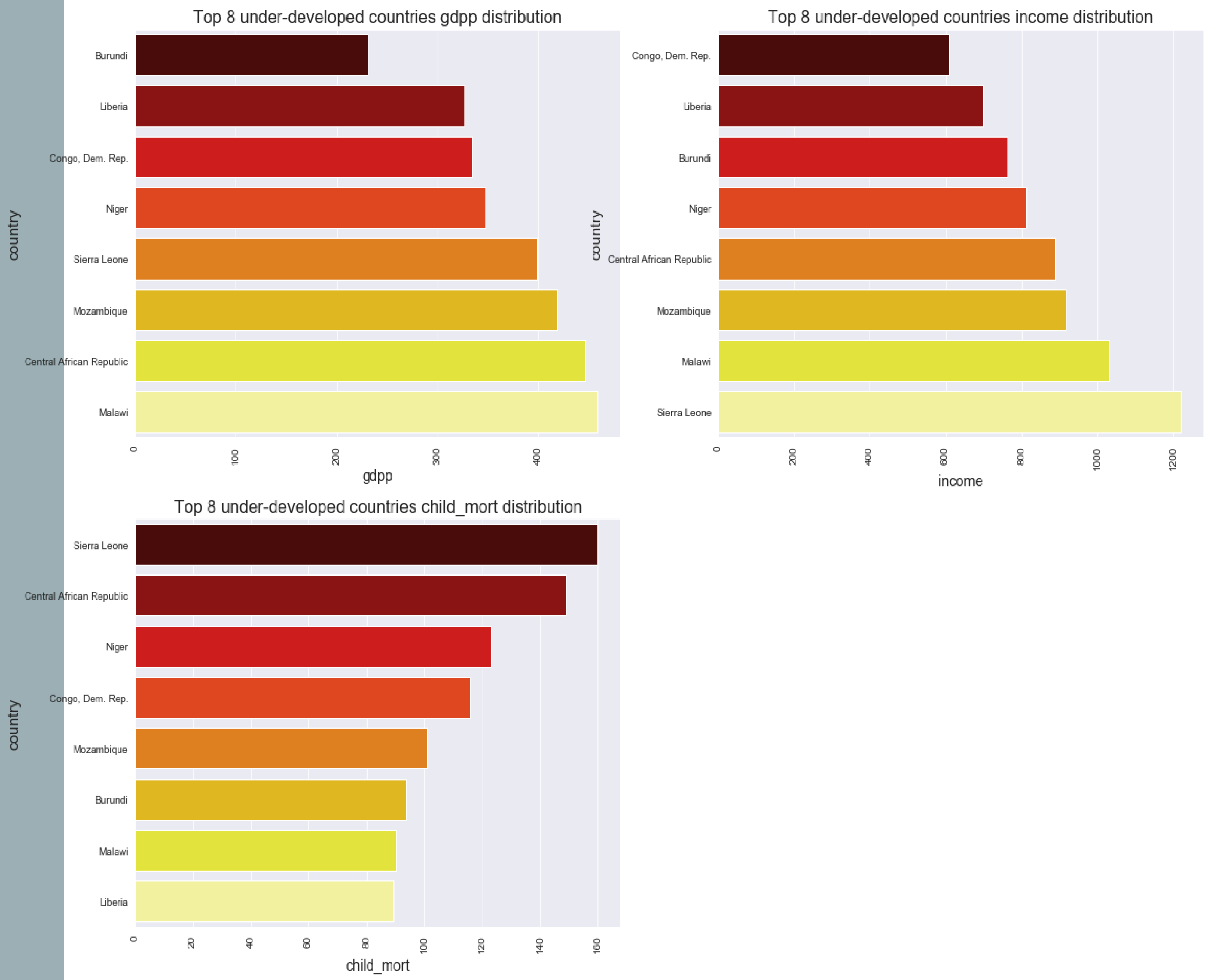


There are 18 under developed countries ordered by rankings. The top most countries are the ones which require immediate attention.

TOP 8 UNDER DEVELOPED COUNTRIES

Final List of countries that require dire aid:

1. Sierra Leone
2. Congo, Dem. Rep.
3. Central African Republic
4. Niger
5. Liberia
6. Burundi
7. Malawi
8. Mozambique


Top 8 under-developed countries gdpp distribution
country
Burundi
Liberia
Congo, Dem. Rep.
Niger
Sierra Leone
Mozambique
Central African Republic
Malawi
gdpp
Top 8 under-developed countries income distribution
country
Congo, Dem. Rep.
Liberia
Burundi
Niger
Central African Republic
Mozambique
Malawi
Sierra Leone
income
Top 8 under-developed countries child_mort distribution
country
Sierra Leone
Central African Republic
Niger
Congo, Dem. Rep.
Mozambique
Burundi
Malawi
Liberia
child_mort