Clustering Case Study

HELP International, is an international humanitarian NGP that is committed to fighting poverty and providing the backward countries with basic amenities and relief during the time of disasters and natural calamities.

After recent funding, CEO of the NGO needs to decide what all countries need to focus on for next campaign.

Prepared by
Vaibhav Swarnkar

Data Manipulations

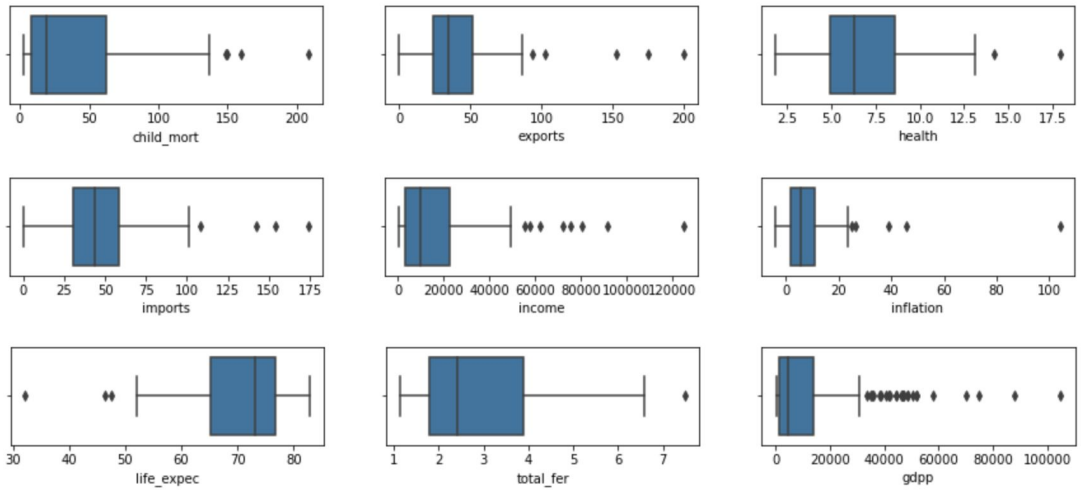
- From the dataset, all of the numerical columns were considered to create the model.
- Inspected dataset using functions - describe, info, head, columns
- Dataset has no missing values
- Handling outliers - Next slide has the charts of columns with outliers
    - Outliers were detected looking by plotting boxplots on each numerical columns

Data Manipulation

2.1.1 Visualizing outliers

Box plt and distribution chart help us visualize the outliers for each column.

```
plot_box(data, number_columns)
```

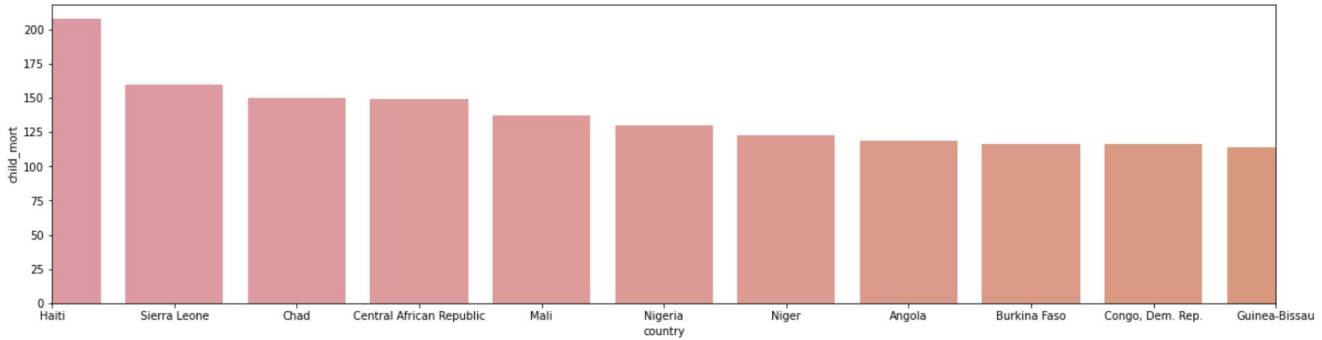


Data Manipulation

- Outliers were treated by capping the values that helps in not eliminating the rows from dataframe.
- All of the columns except **child_mort** and **inflation** will be capped at higher range.

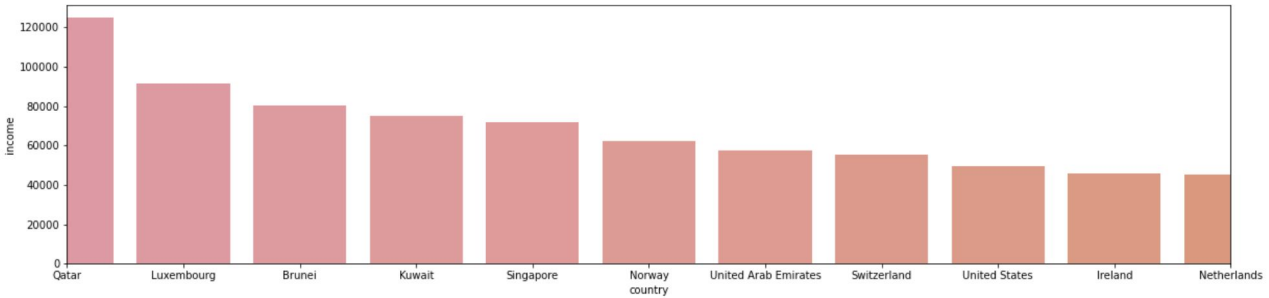
Data Analysis

When compared child mortality rate in countries, Haiti turned out to be with the highest child mortality rate.



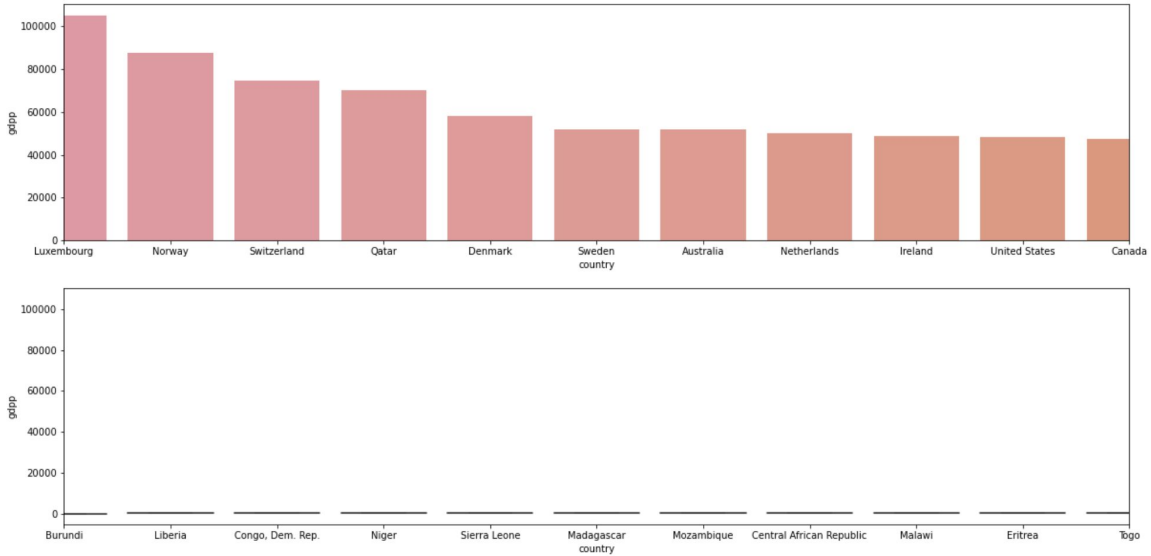
Data Analysis

When compared income in countries, Qatar has the highest income range.



Data Analysis

GDP per capita turned out to be highest in Luxembourg and lowest in Burundi



Modeling

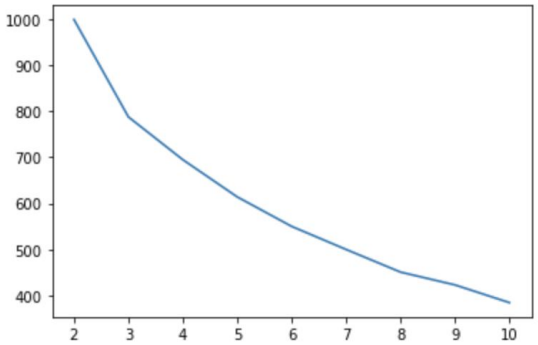
Two modeling techniques were used

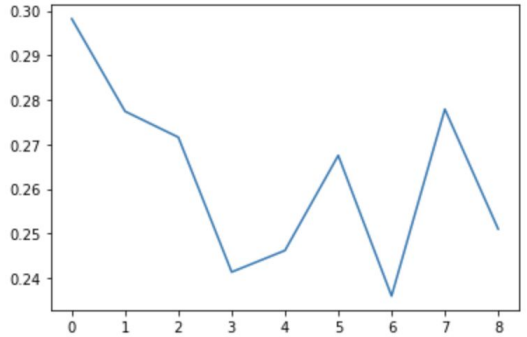
1. KMeans
2. Hierarchical Clustering

Optimal Number of clusters -

After performing Elbow Curve and Silhouette Analysis, it was concluded to use 3 clusters.

Elbow Curve

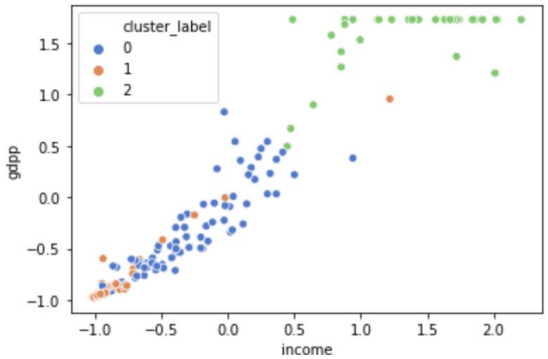




Silhouette Analysis

Visualizing the clusters

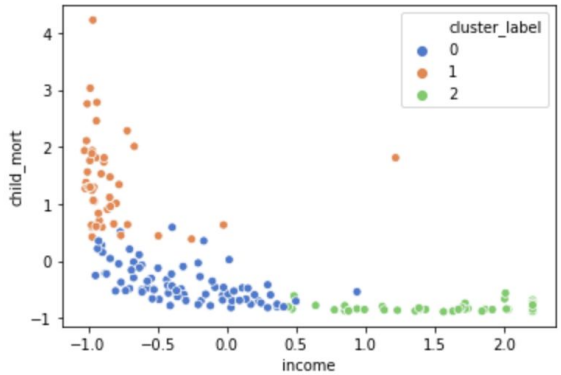
```
: sns.scatterplot(x='income', y='gdpp', hue='cluster_label', data=data_kmeans_scaled, palette='muted')
  plt.show()
```



Income and GDPP

Visualizing Cluster

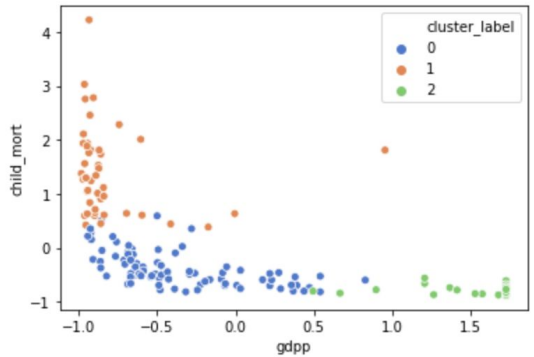
```
sns.scatterplot(x='income', y='child_mort', hue='cluster_label', data=data_kmeans_scaled, palette='muted')
plt.show()
```



Income and Child Mortality

Visualizing Cluster

```
sns.scatterplot(x='gdpp', y='child_mort', hue='cluster_label', data=data_kmeans_scaled, palette='muted')
plt.show()
```



GDPP and Child Mortality

Top Results of KMeans

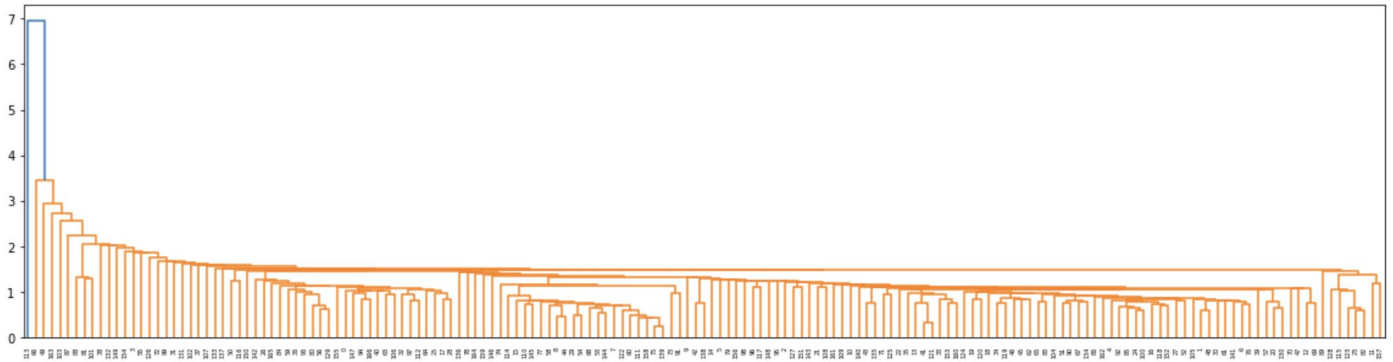
After performing profiling and and considering the cluster 1 as we can see that the countries in cluster 1 have low income and gdpp with high child mortality rate and below are the list of the countries.

1. Burundi
2. Liberia
3. Congo, Dem Rep
4. Niger
5. Sierra Leone

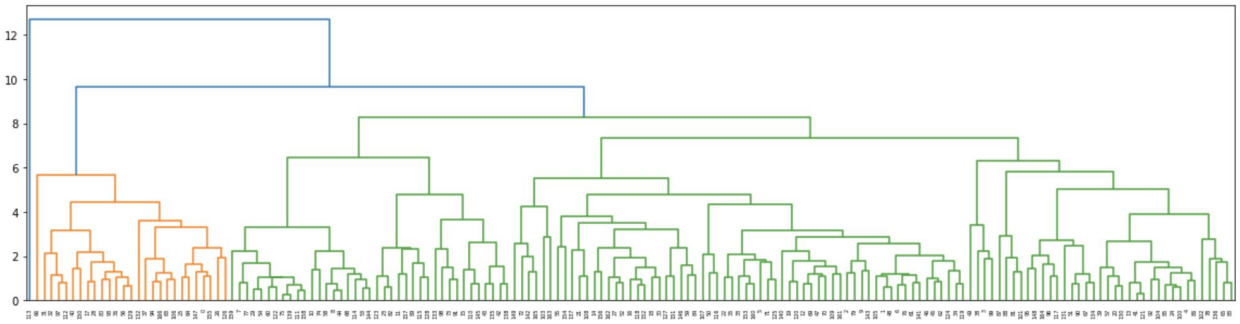
Top Results of Hierarchical

In Hierarchical Clustering, using both techniques Single and Complete Linkage gives only one country **Haiti** in result that needs dire aid.

Single Linkage



Complete Linkage



Conclusion

To conclude the study and outcome, the suggested clustering to go with would be KMeans as it is giving top five countries that needs attention and is much clear through profiling as well.

Final list of countries to be considered

1. Burundi
2. Liberia
3. Congo, Dem Rep
4. Niger
5. Sierra Leone