Clustering Assignment

BY Karamveer Kath

Problem Statement

- HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities
- After the recent funding programmes, they have been able to raise around $ 10 million.
- Now the CEO of the NGO needs to decide how to use this money strategically and effectively.
- As a data analyst your job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

Problem Approach- Steps

- Start off with the necessary data inspection and EDA tasks suitable for this dataset - data cleaning, univariate analysis, bivariate analysis etc.
- **Outlier Analysis**
- Try both K-means and Hierarchical clustering(both single and complete linkage) on this dataset to create the clusters
- Analyse the clusters and identify the ones which are in dire need of aid
- need to perform visualisations on the clusters that have been formed
- Both K-means and Hierarchical may give different results because of previous analysis
- report back at least 5 countries which are in direst need of aid from the analysis work that you perform.

Data Inspection & Cleaning

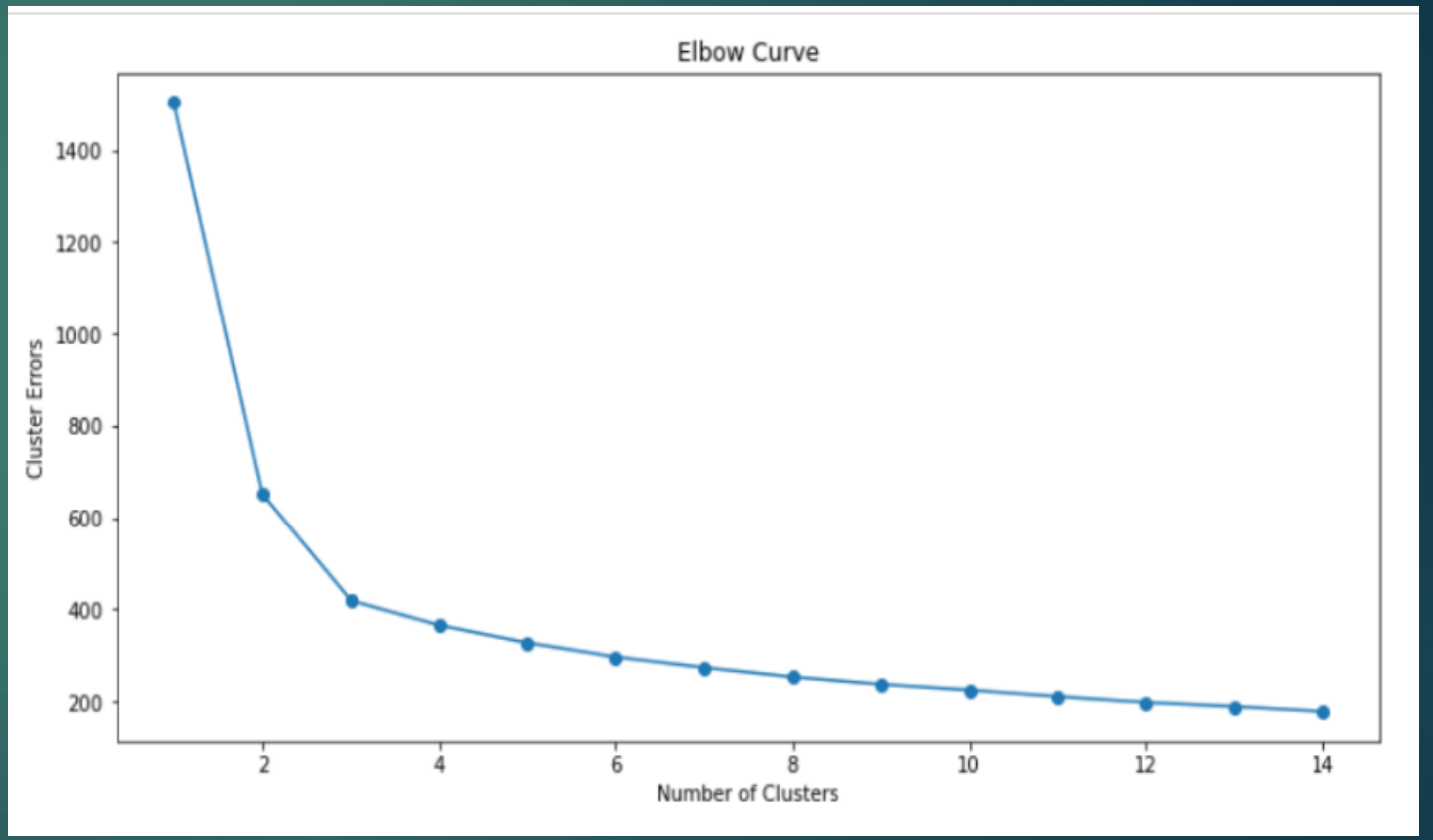
- No missing values in our data
- The Outliers in our data are completely acceptable from Business perspective, as we'll have poor countries and highly developed countries as well.
- To get rid of skewness and make our data normal, we transformed the variables using SKLearns power
- Converted exports, imports and health columns absolute values from percentages
- As we have one row for each country, removing outliers will cause data loss which is not a feasible solution

Finding the Optimal Clusters

- To find the Optimal number of Clusters, I have

used Elbow Curve method

- From the curve, we could see the elbow at number of clusters are 3
- So, I have decided to take Optimal Number of clusters for modelling as 3



Modelling - KMeans

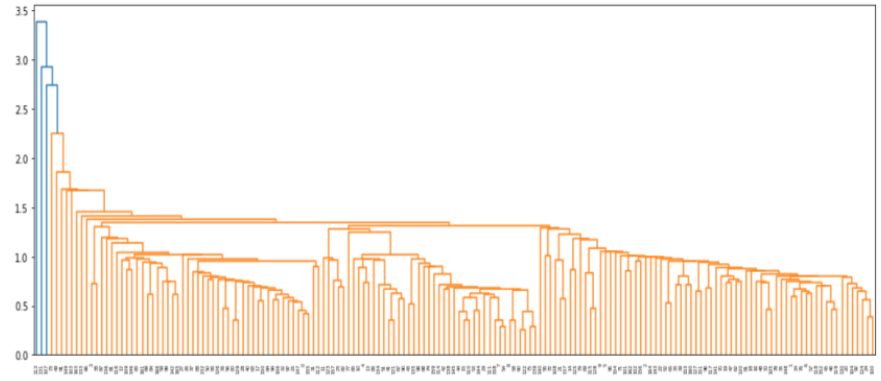
- Using Number of clusters as 3, and init method as “K-means ++” , we build the model using fit
  method
- Predicted the clusters using Predict method of Kmeans

Modelling Hierarchical

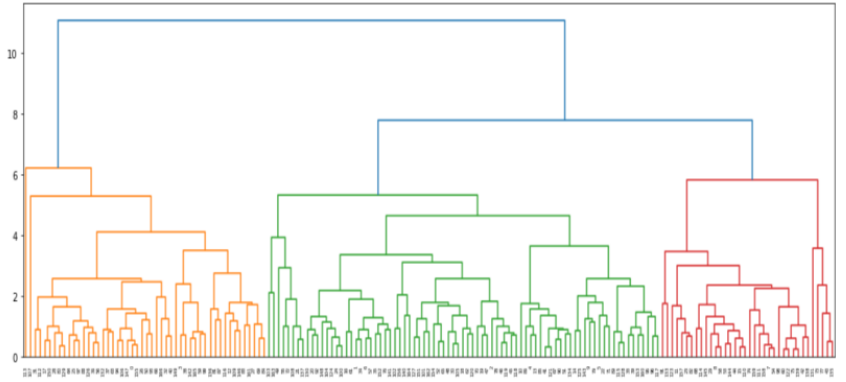
- Built the model using “Euclidean distance” as metric and linkage type as “Single”
- Plotted the Dendrogram for single linkage, we won’t be able to observe good clusters in single
linkage
- Built the model using Complete Linkage, we could clearly observe 3 clusters formed.
- Used Cut_tree with n_clusters = 3 to get the labels of the clusters formed

Modelling Hierarchical

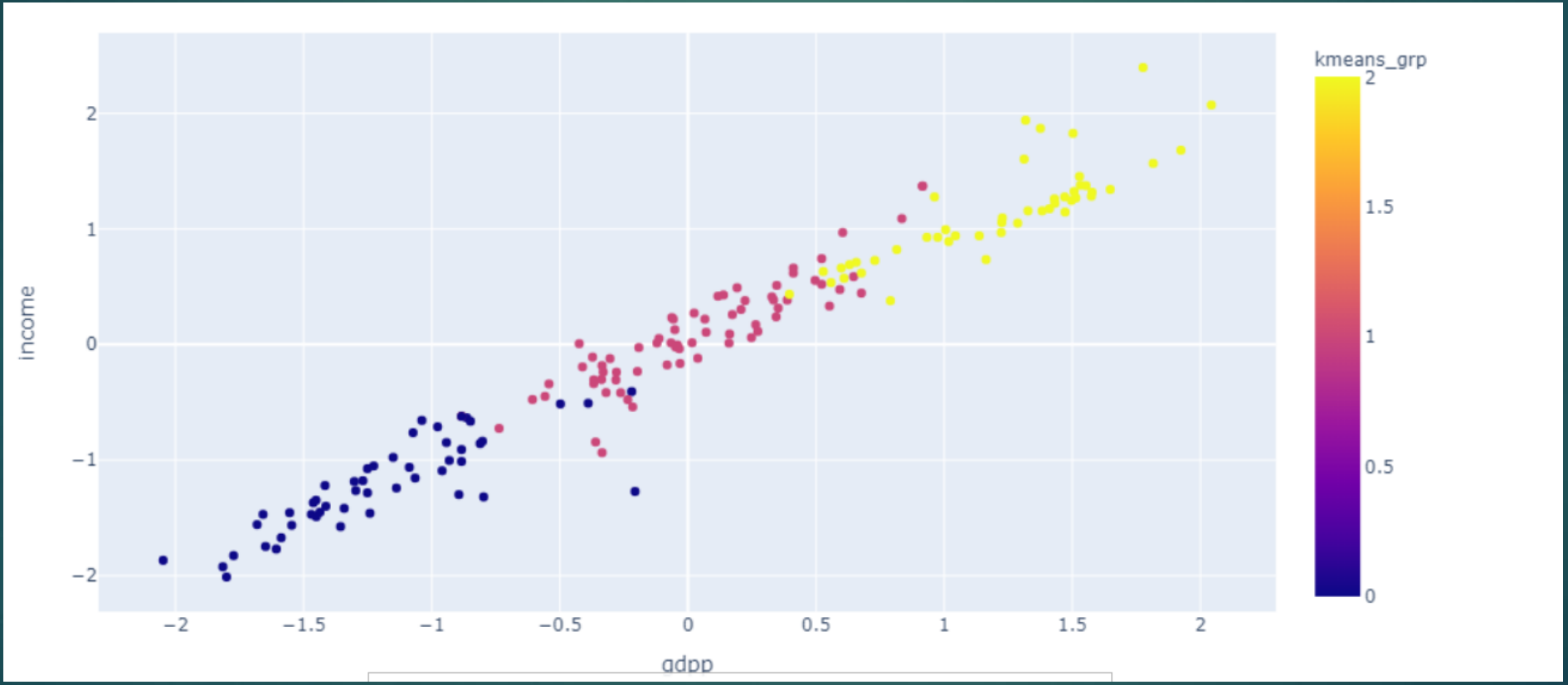
SINGLE LINKAGE



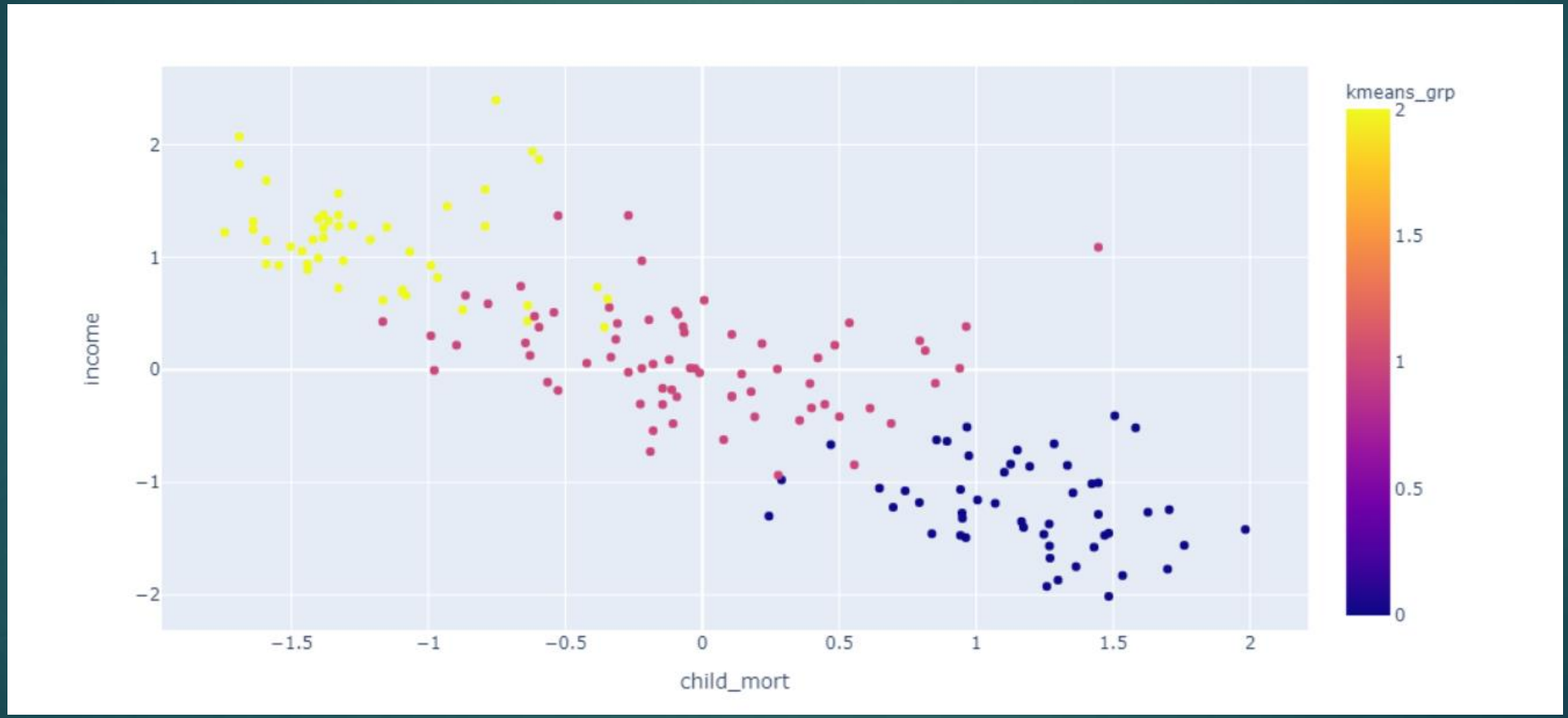
COMPLETE LINKAGE



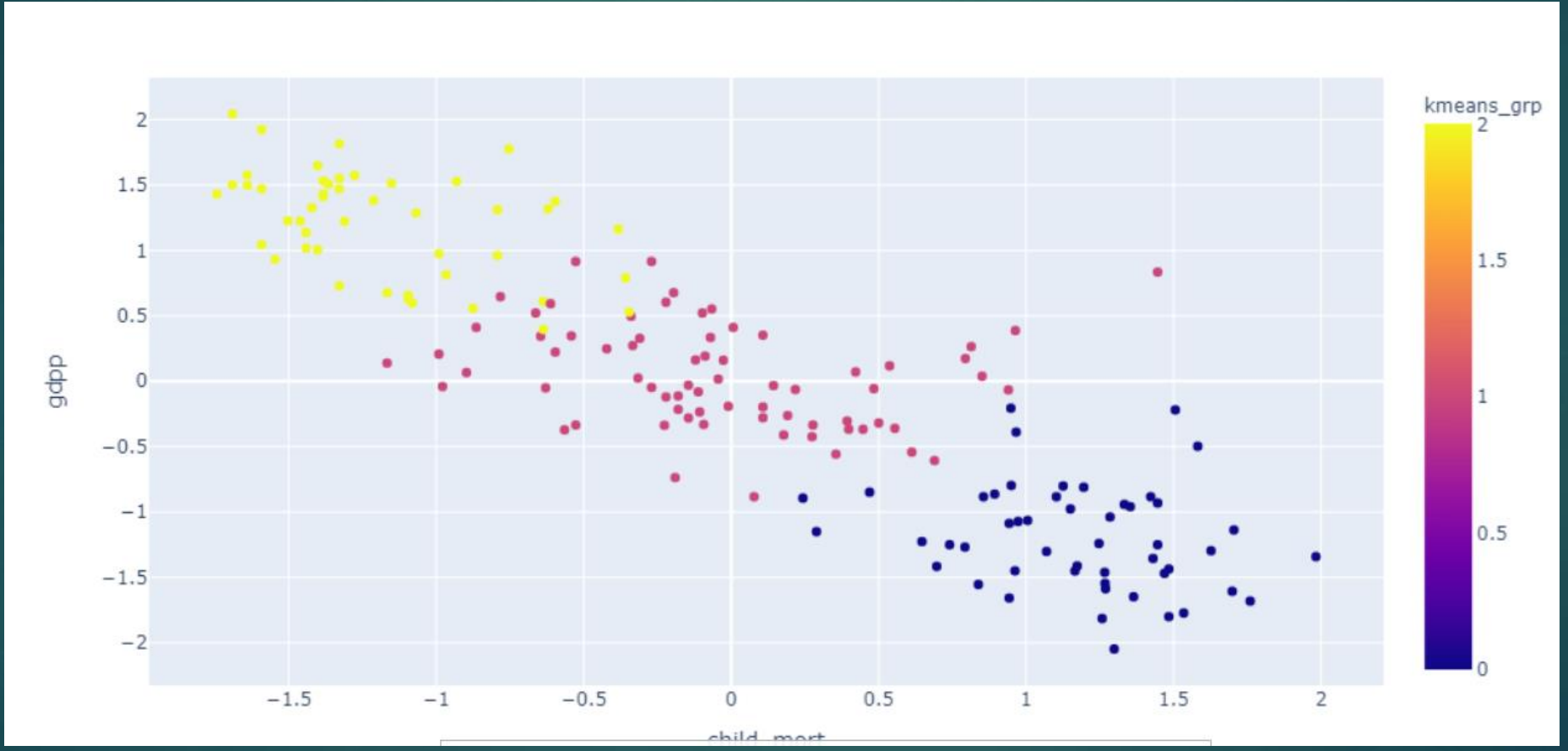
Visualisations – gdpp vs income



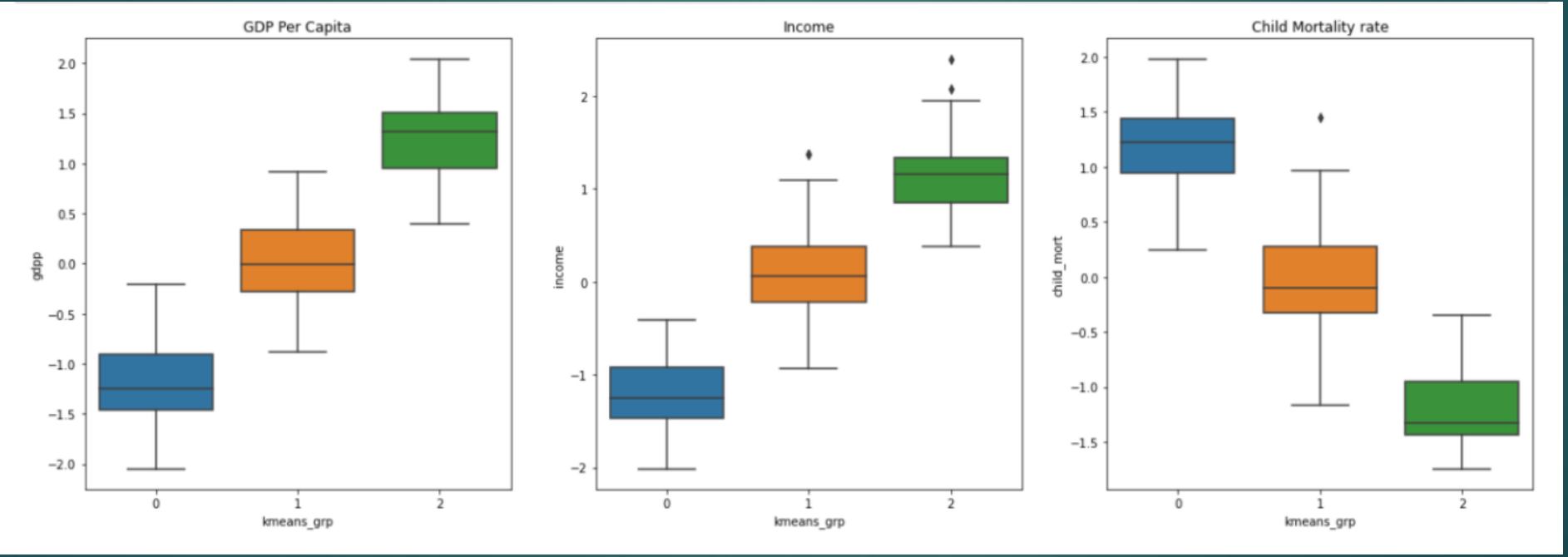
Visualisations – child_mort vs income



Visualisations – child_mort vs gdpp



Boxplots of Clusters formed



Interpretation

- Basis the box plots and scatter plots, we can see the cluster formations clearly
- From these we label the clusters formed as :
  - Cluster - 0 : High Child Mortality, Low Income and Low GDP
  - Cluster - 1 : Average Child Mortality, Average Income and Average GDP
  - Cluster - 2 : Low Child Mortality, High Income and High GDP
- So, we can see that Cluster 0 has High Child mortality rate, low income and low GDP, which contains the poor countries.
- We have a total of 50 poor countries

Result

- From the list of poor countries we obtained, sorted the list on income, gdpp, child_mortality rate.

Top 5 countries which are in direct need :
Congo, Dem Rep
Liberia
Burundi
Niger
Central African Republic

- According to the Business requirements we may modify the list obtained.