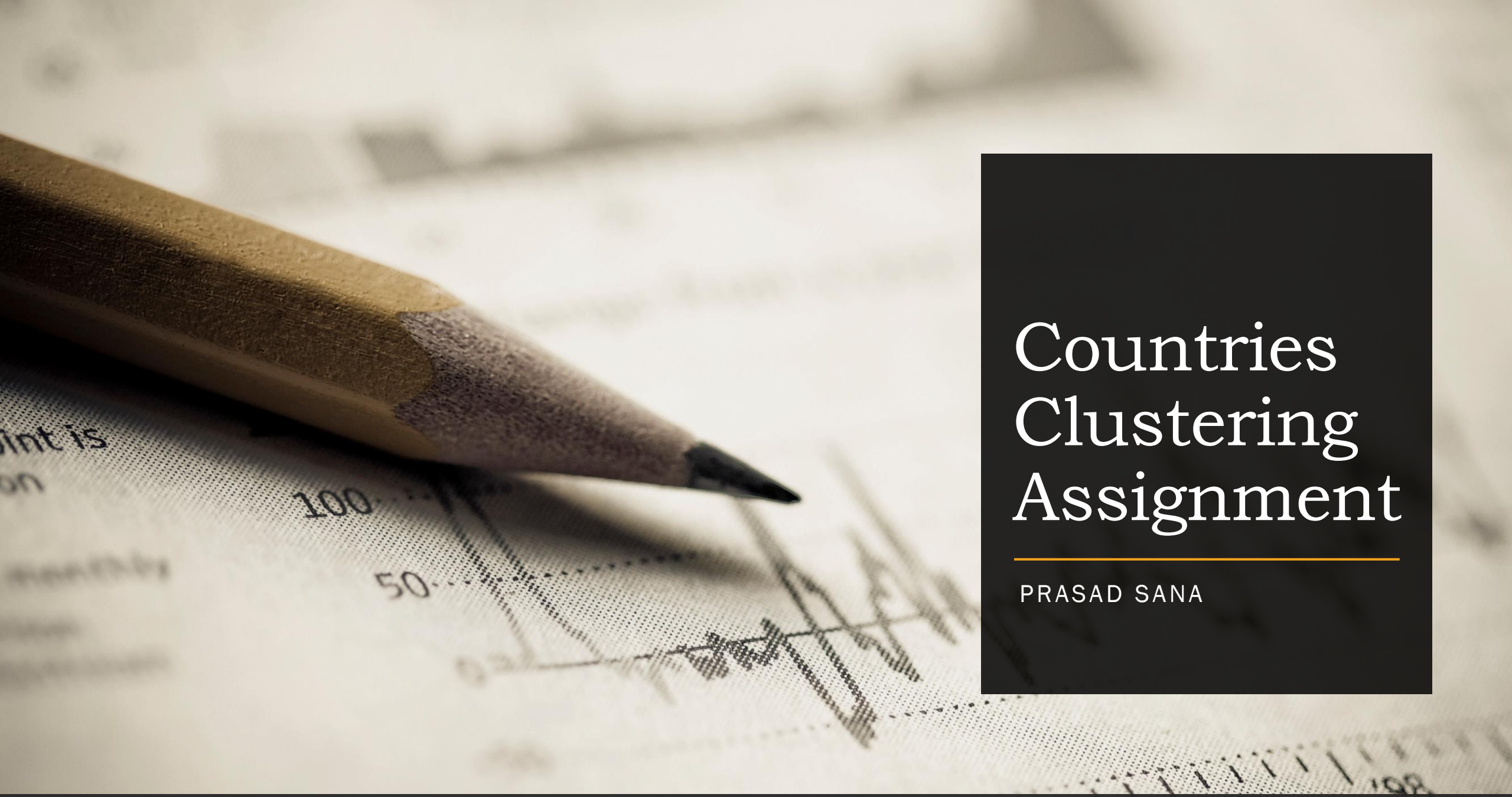


# Countries Clustering Assignment

PRASAD SANA

# 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.
- The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.
- As a Data Scientist, **we need to find the countries in direst need and help CEO of HELP International** in using the fund money to reach right countries

# Problem Approach

- As we have the Data of countries like child mortality rate, GDP Per Capita, Income etc. , we can use **Clustering** to segregate the countries into different groups
- Steps :
  - Data Inspection – Missing Values if any, EDA
  - Outlier Analysis
  - Data Pre-processing
  - Finding Optimal number of Clusters
  - Modelling
    - KMeans Clustering
    - Hierarchical Clustering – Single and Complete Linkages
  - Listing down top 5 countries in need

# Data Inspection and Outlier Analysis

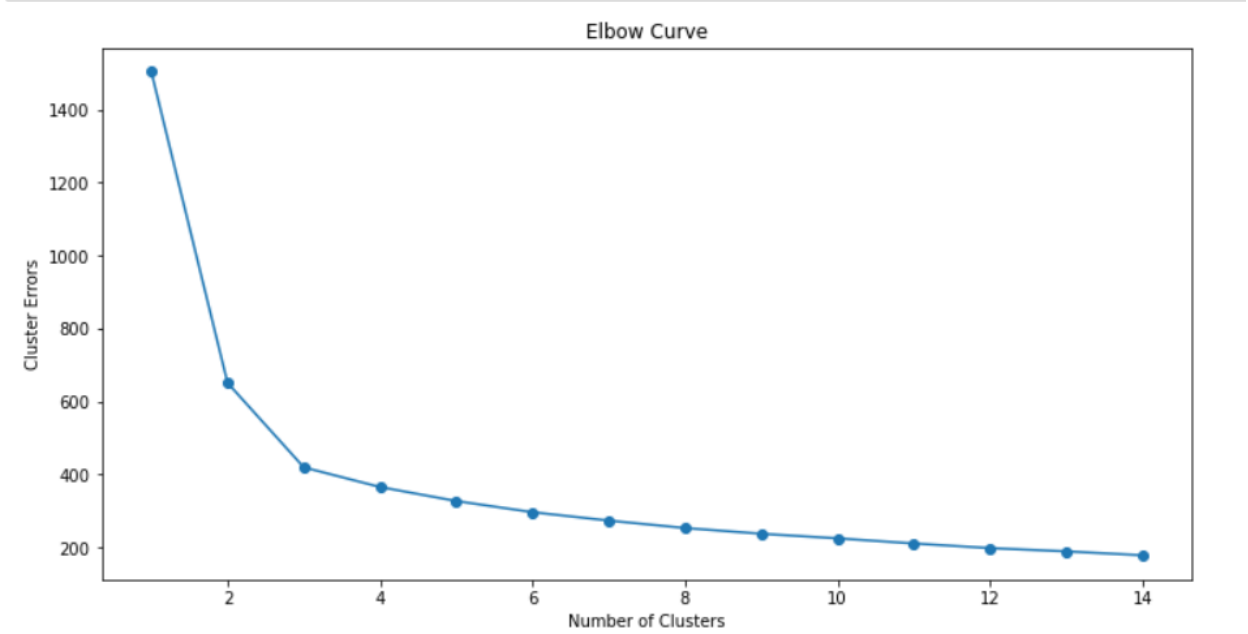
- We do not have any missing values in our data
- Converted exports, imports and health columns absolute values from percentages
- After plotting all the distplots, we observed that almost all of the features are right skewed(except for life span which is left skewed), which is a sign of outliers
- The Outliers in our data are completely acceptable from Business perspective, as we'll have poor countries and highly developed countries as well.
- As we have one row for each country, removing outliers will cause data loss which is not a feasible solution. Also, Capping the data may lead to bad clustering as we are changing the data itself.
- To get rid of skewness and make our data normal, we transformed the variables using SKLearns power transformation.

# Data Pre-processing

- As we already saw that we have skewed data, clustering will result in bad clusters.
- Also, the transformation we did helped in making the data close to Normal distribution, as well as helped us in skipping the Scaling.
- We don’t have to scale the data explicitly

# Finding Optimal number of Clusters

- To find the Optimal number of Clusters, we used Elbow Curve method
- From the curve, we could see the elbow at number of clusters = 3
- So, we decided to take Optimal Number of clusters for modelling as 3



# Modelling - KMeans

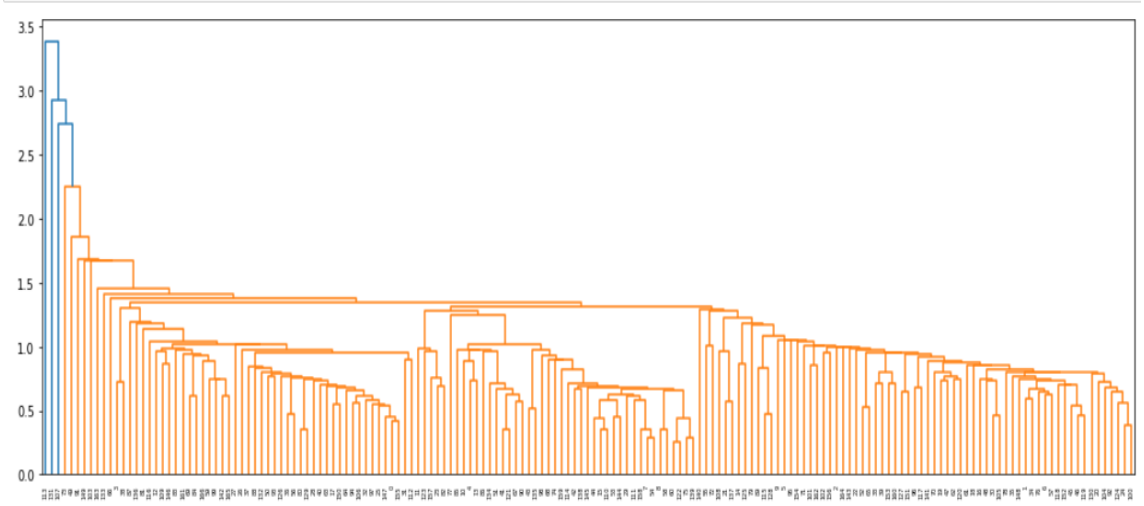
- Using Number of clusters as 3, and init method as “K-means ++” , we build the model using fit method
- Predicted the clusters using Predict method of Kmeans

# Modelling – Hierarchical

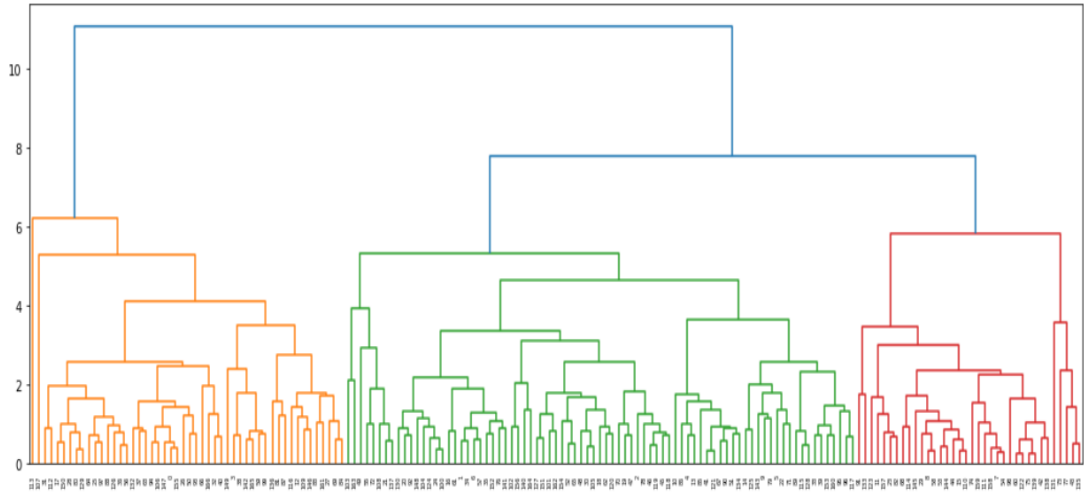
- Built the model using “Euclidean distance” as metric and linkage type as “Single”
- Plotted the Dendrogram for single linkage, we won’t be able to observe good clusters in single linkage
- Built the model using Complete Linkage, we could clearly observe 3 clusters formed.
- Used **Cut_tree** with n_clusters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

# Modelling - Hierarchical

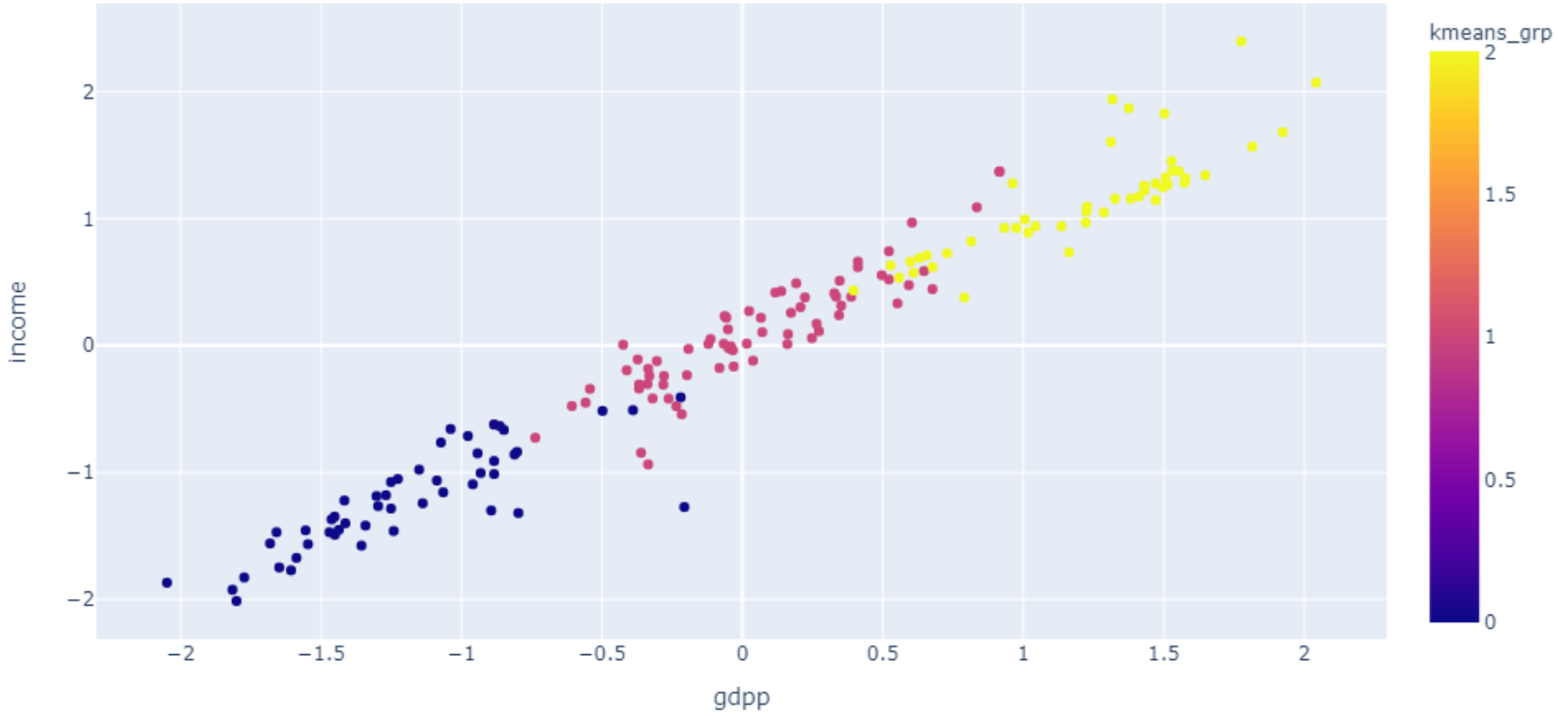
SINGLE LINKAGE



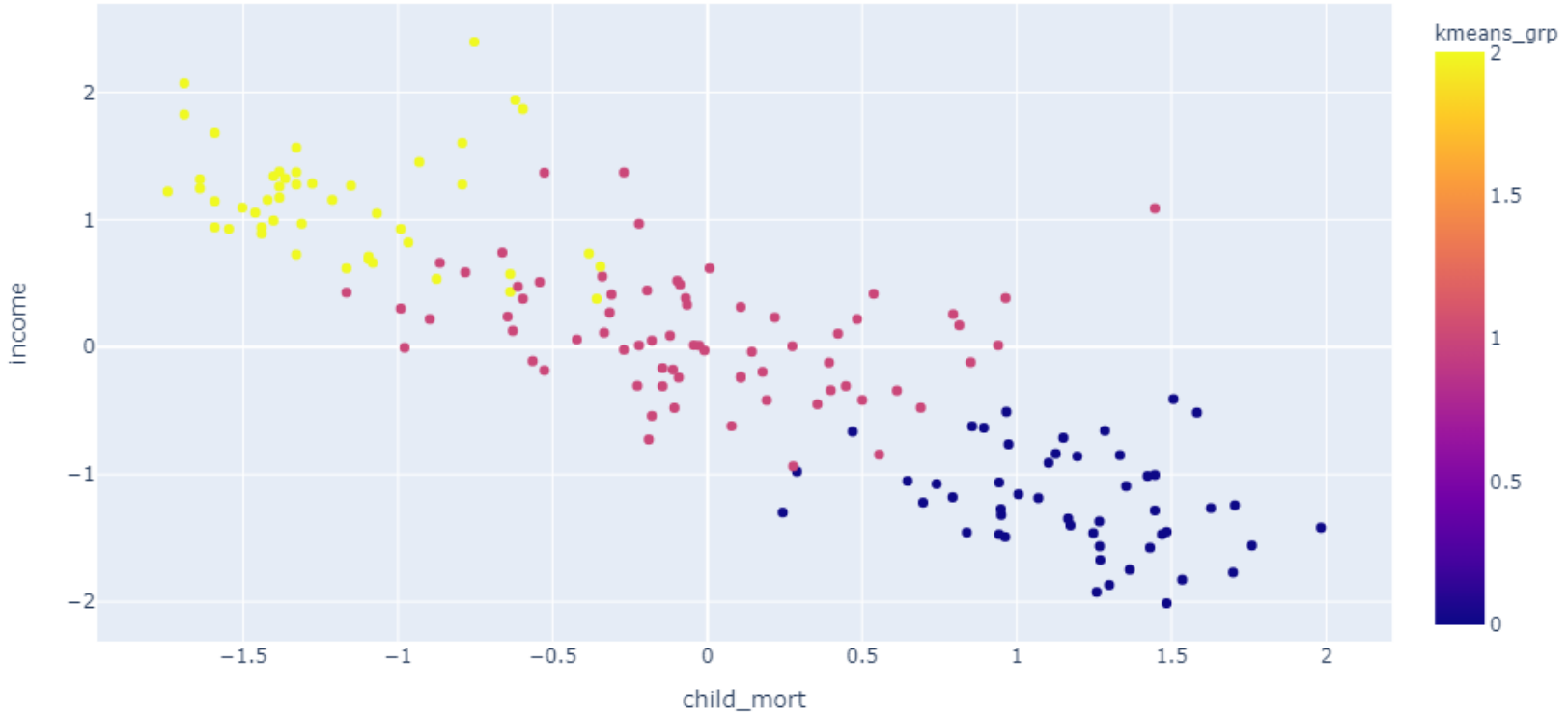

COMPLETE LINKAGE



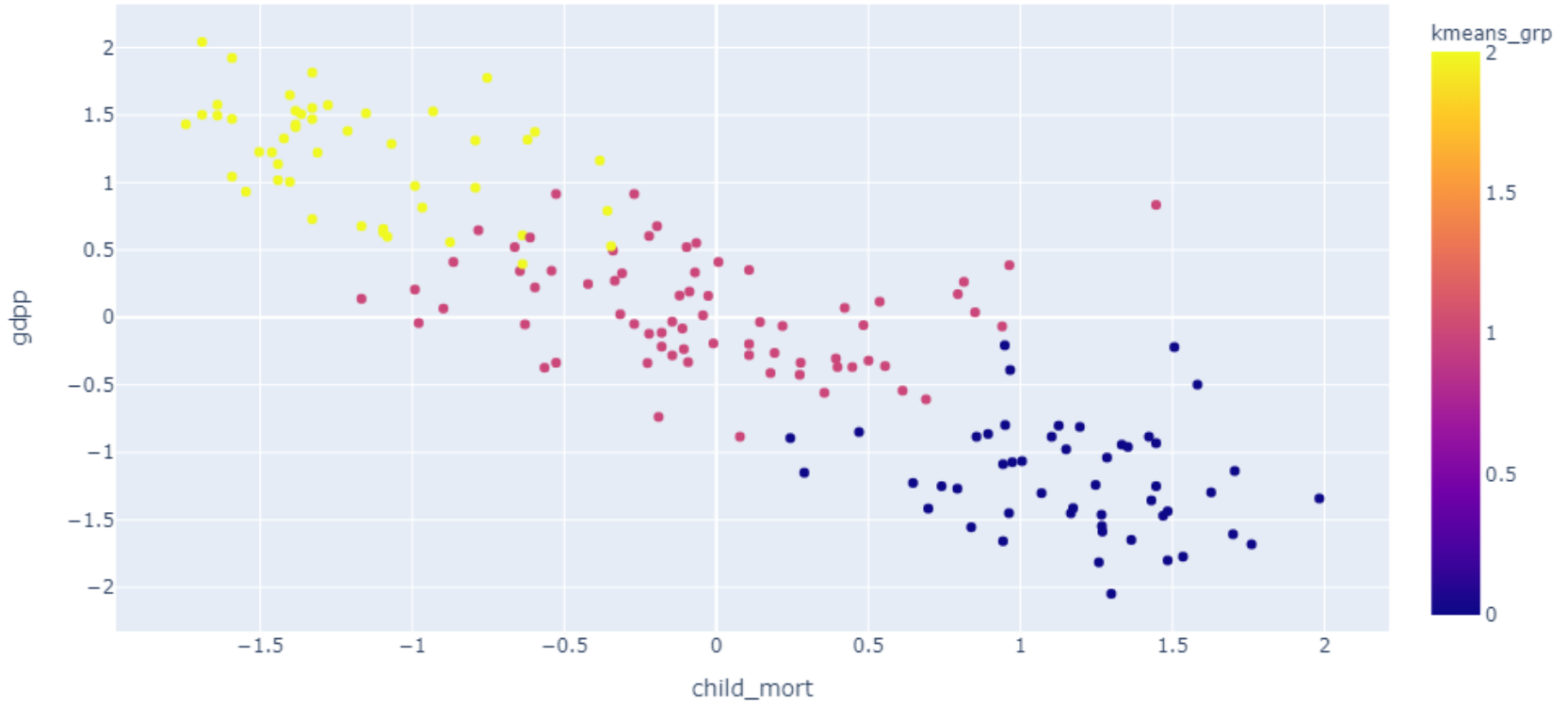
# Visualisations – gdpp vs income



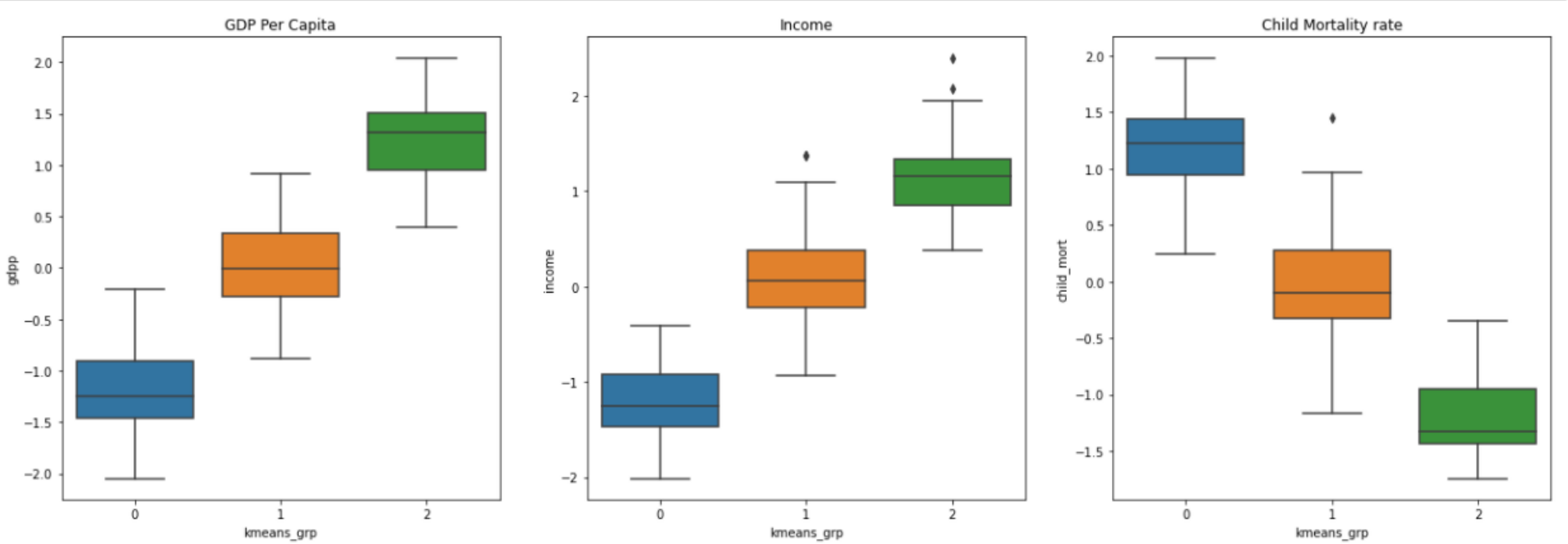
# Visualisations – child_mort vs income



# Visualisations – child_mort vs gdpp



# Boxplots of Clusters formed



# Interpretation

- From the box plots and scatter plots, we could 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 conclude that Cluster 0 has High Child mortality rate, low income and low GDP, which is 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 direst need :
  - *Congo, Dem Rep*
  - *Liberia*
  - *Burundi*
  - *Niger*
  - *Central African Republic*
- Based on Business needs, we can change the sort order to get different list