# Clustering Assignment

Ayush Chauhan

# 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 and Data Scaling if necessary
  - Finding Optimal number of Clusters
  - Modelling
    - KMeans Clustering
      - Use Hopkins Method to check if the dataset is good enough for a cluster analysis
      - Use Silhouette and Elbow method to validate the optimal cluster values.
  - Hierarchical Clustering – Single and Complete Linkages
  - Listing down top 10 countries in need

# DATA INSPECTION AND OUTLIER ANALYSIS

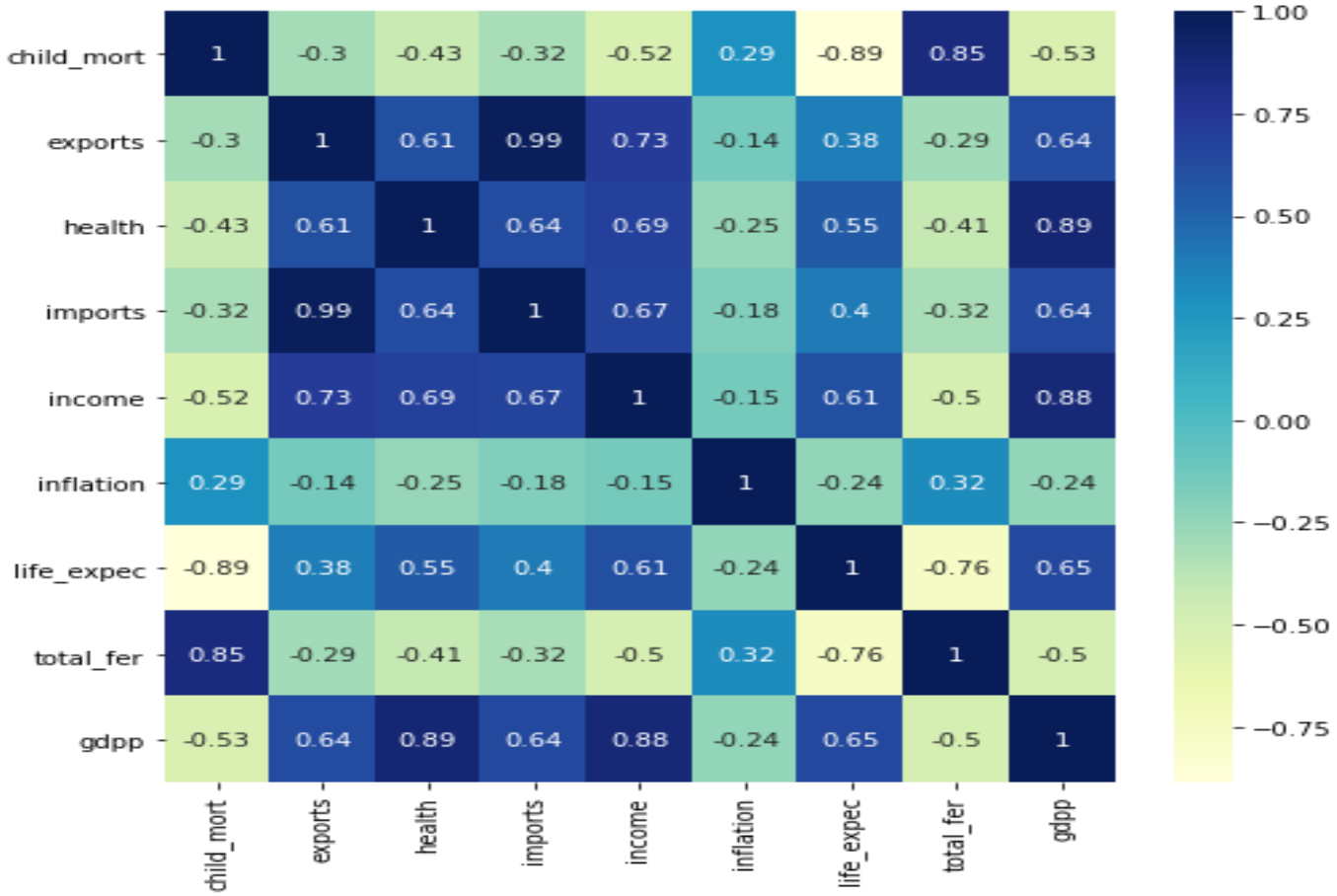
- We do not have any missing values in our data
- Converted exports, imports and health columns absolute values from percentages
- The Outliers in our data are completely acceptable from Business perspective, as we'll have poor countries and highly developed countries as well.
- As we have one row for 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.
- Their range are also differnt All the above points indicates the need of standardising the data before we build the model.
- Since we need to compute the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range. Thus, scaling down of all attributes to the same normal scale is important here.

# DATA PRE-PROCESSING

- used Standardisation method for scaling the data

- Hopkins Statistics
  - Hopkins Statistic over .70 is a good score that indicated that the data is good for cluster analysis.
  - A 'Hopkins Statistic' value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

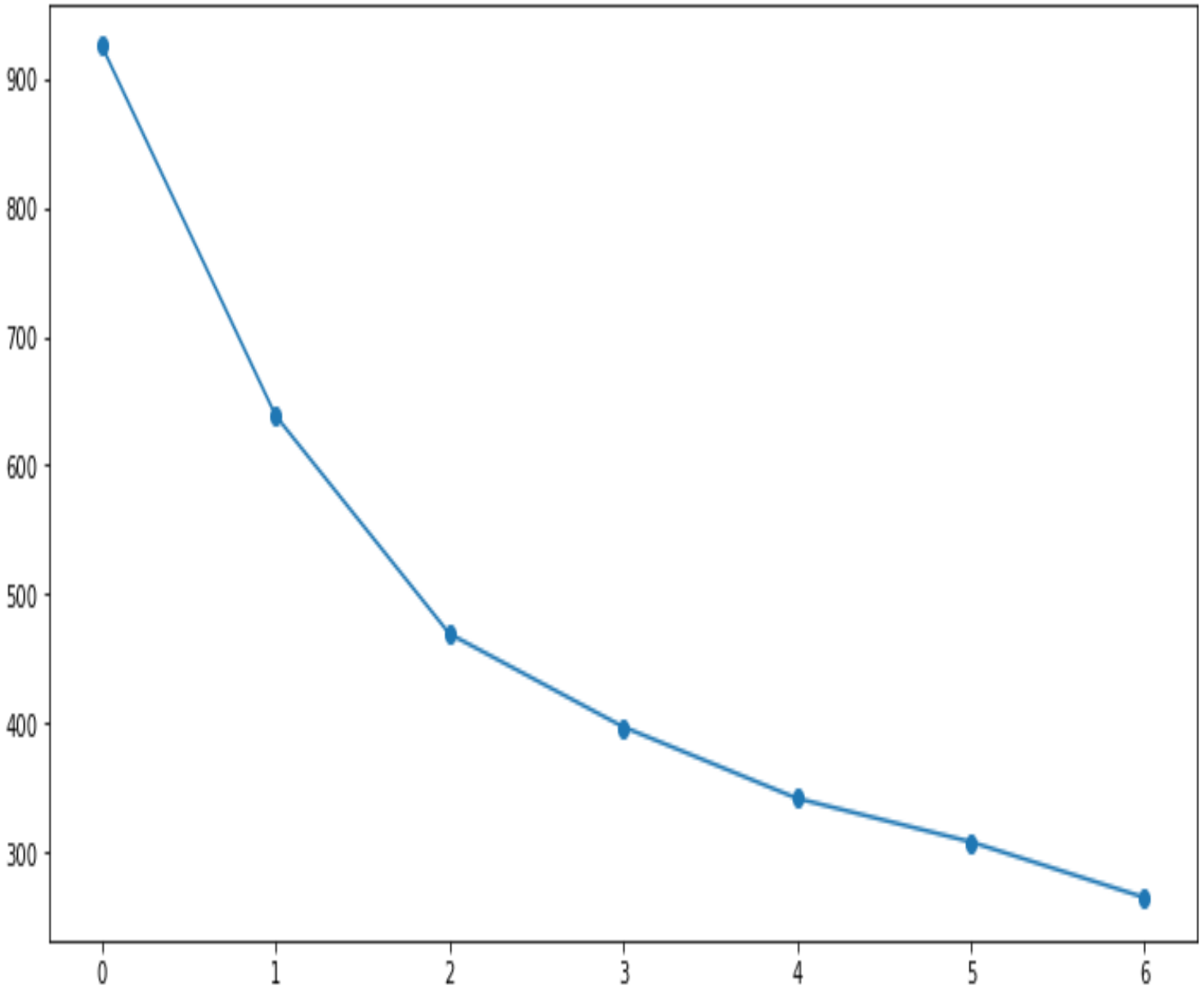
# CORRELATION IN THE DATA:




- After data cleaning , we removed outlier from gdpp column because the country with high gdpp would not require any aid as there are already doing good.
- We did standardized scaling to standardize all parameters on cleaned, outlier removed data.
- Looking at the heatmap, we see that few variables like (total fertility, child mortality) , (income , gdpp) and (imports and exports) have high correlation.

# 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

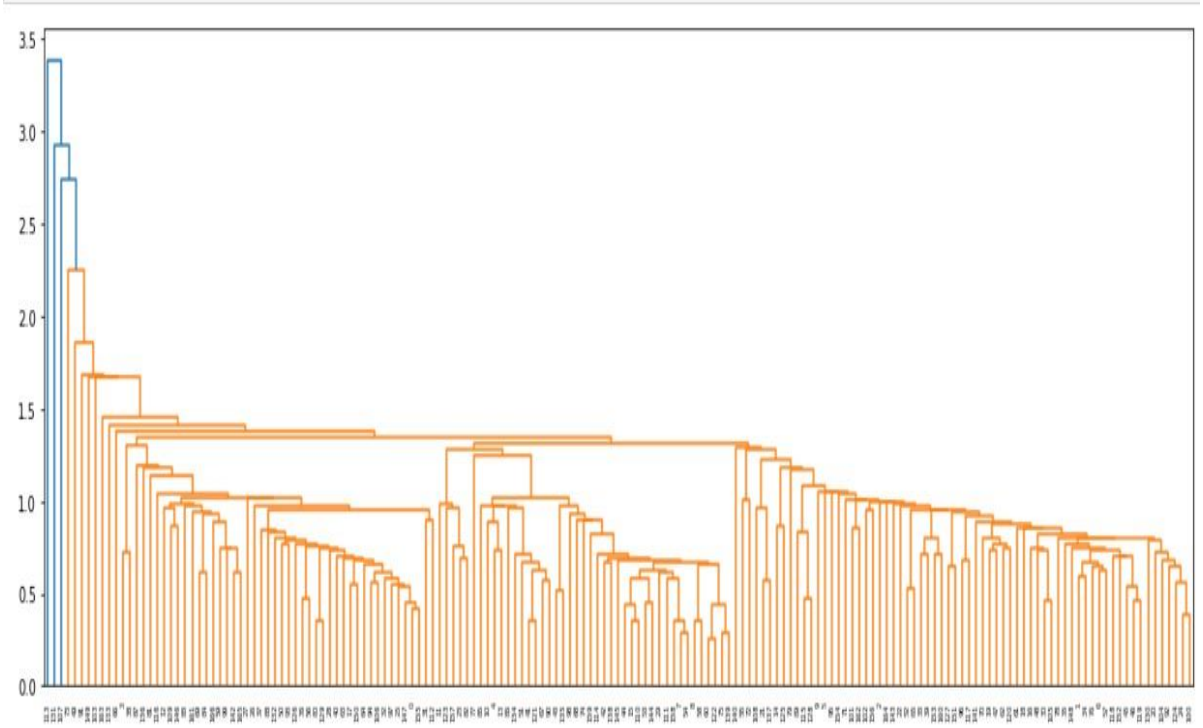
- Using Number of clusters as 3, and init method as “K-means ++” , we build the model using fit method
- Predicted the clusters using Predict method of Kmeans

# MODELLING – HIERARCHICAL

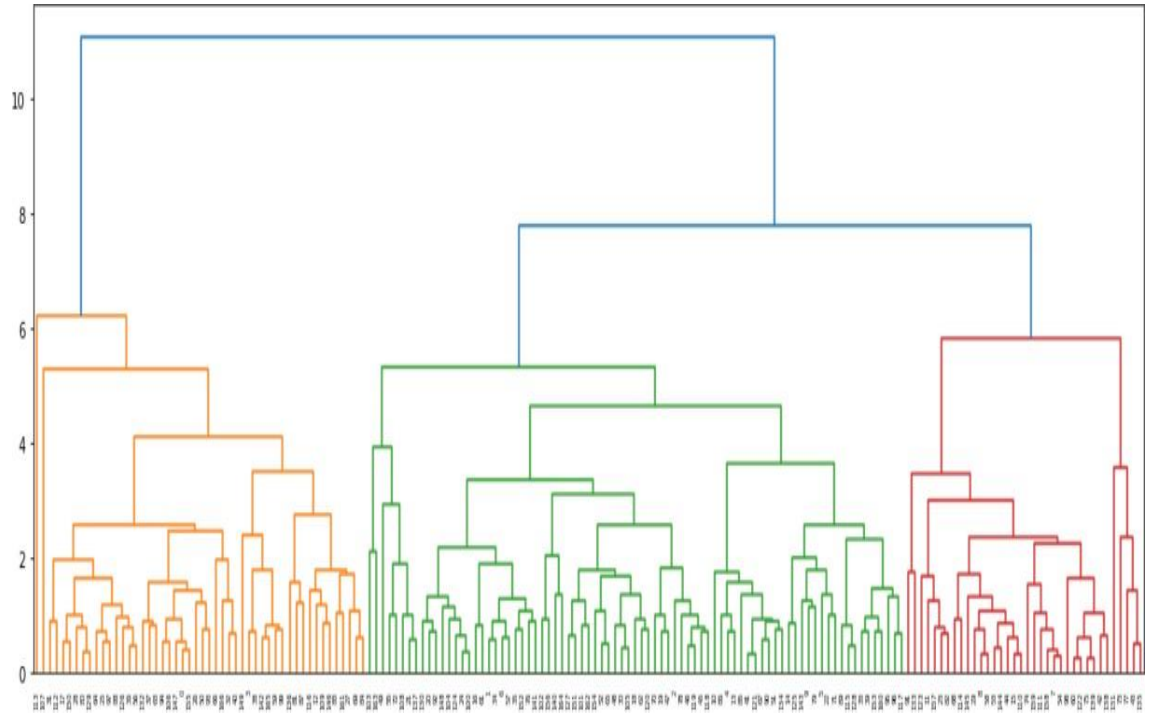
- Built the model using “Euclidean distance” as metric and linkage type as “Single”
- Plotted the Dendrogram for single linkage, we won’t be able to observe good clusters in single linkage
- Built the model using Complete Linkage, we could clearly observe 3 clusters formed.
- Used Cut_tree with n_clusters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

# MODELLING - HIERARCHICAL

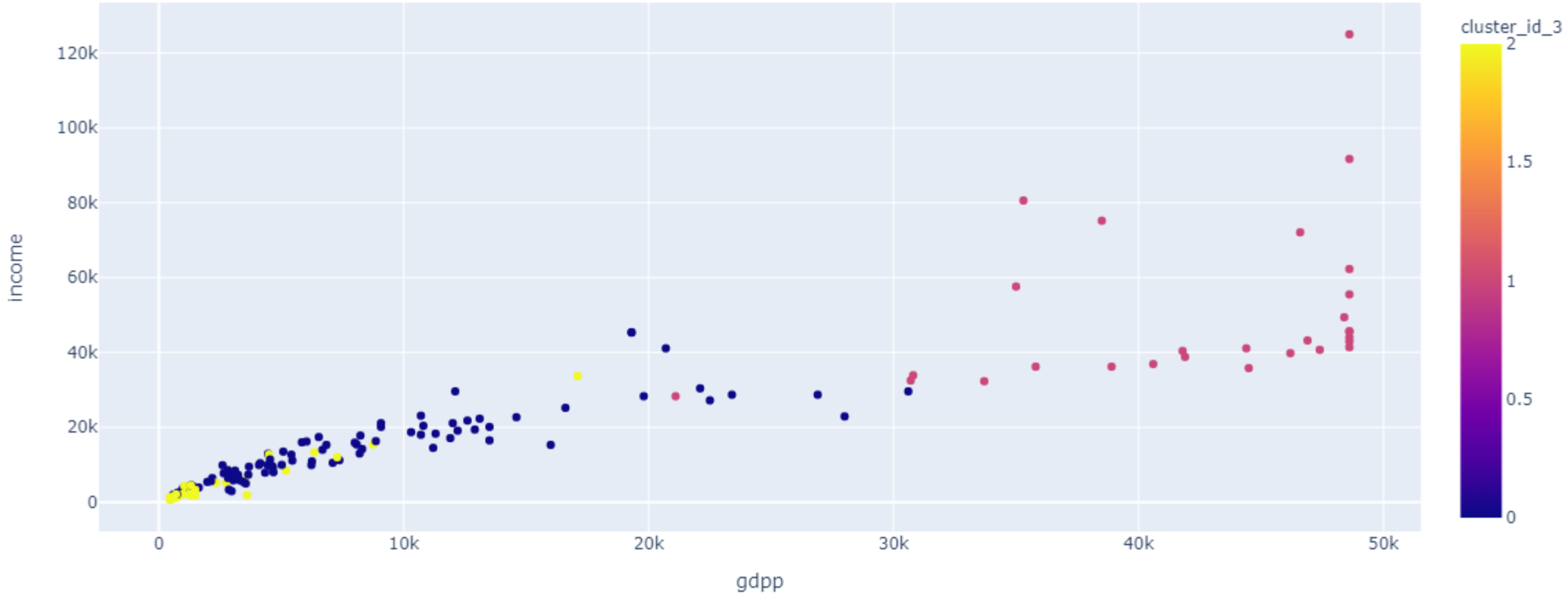
SINGLE LINKAGE



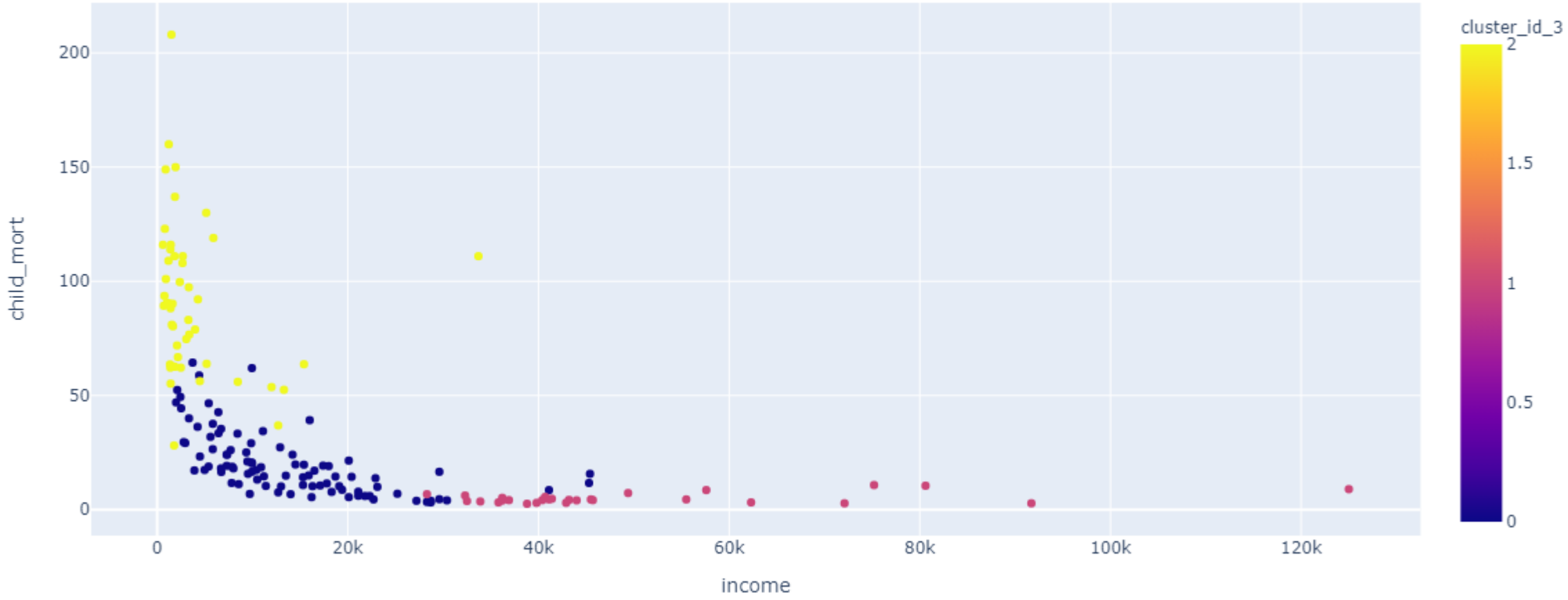

COMPLETE LINKAGE



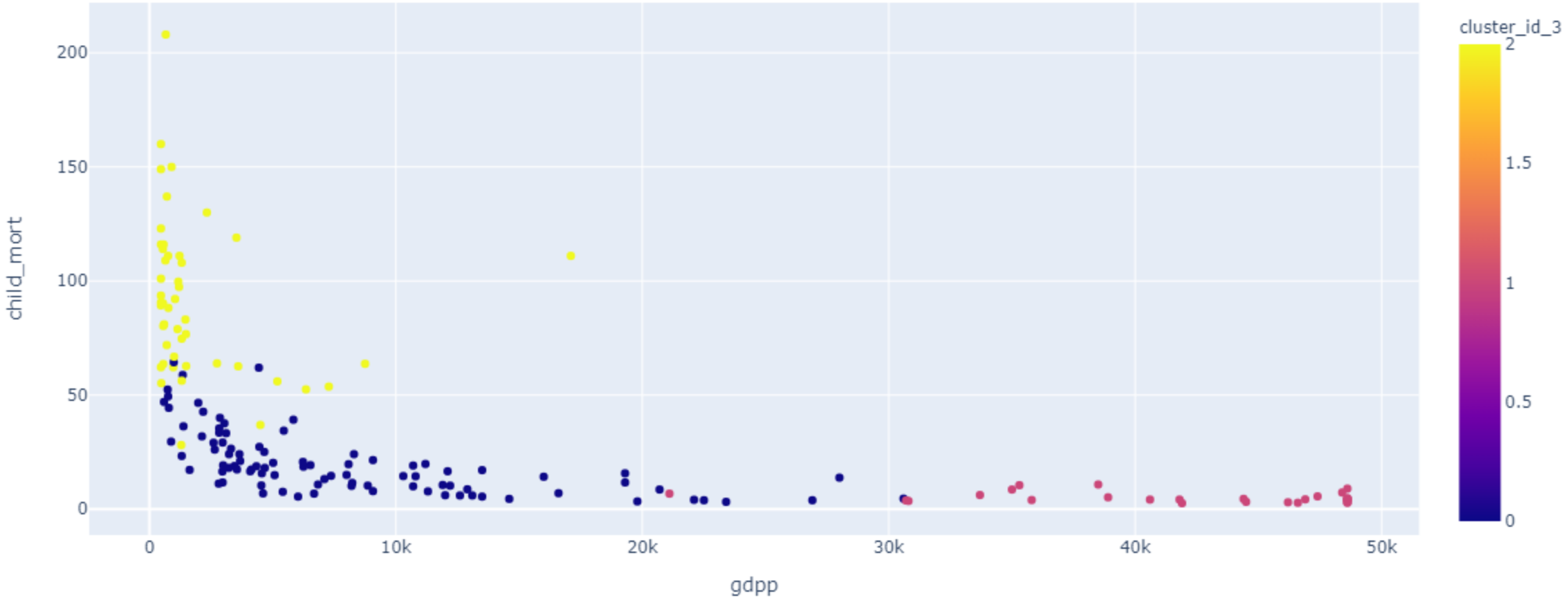
# VISUALISATIONS – GDPP VS INCOME



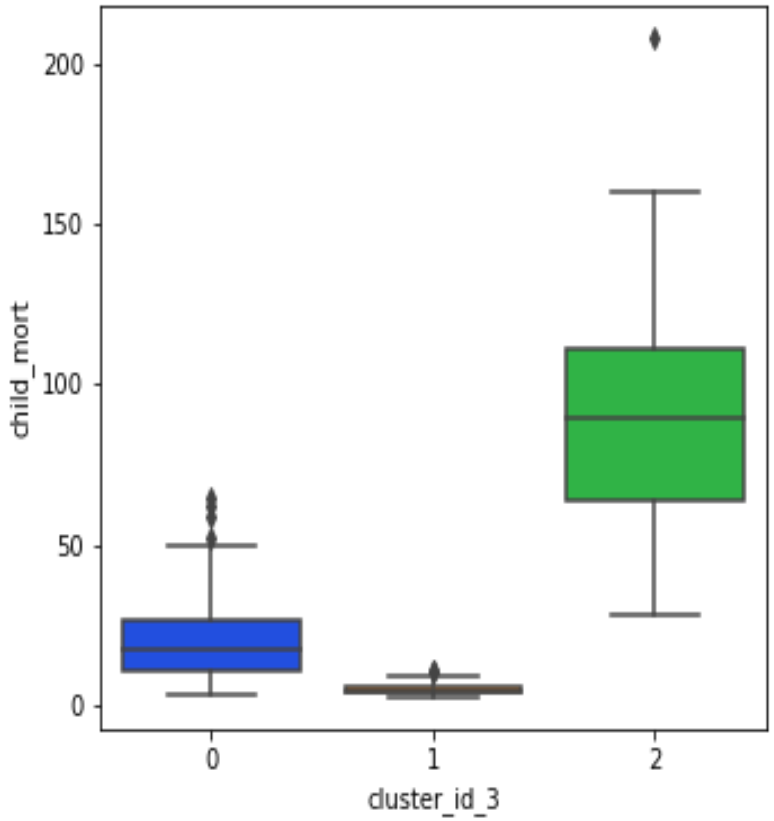
# VISUALISATIONS – CHILD_MORT VS INCOME

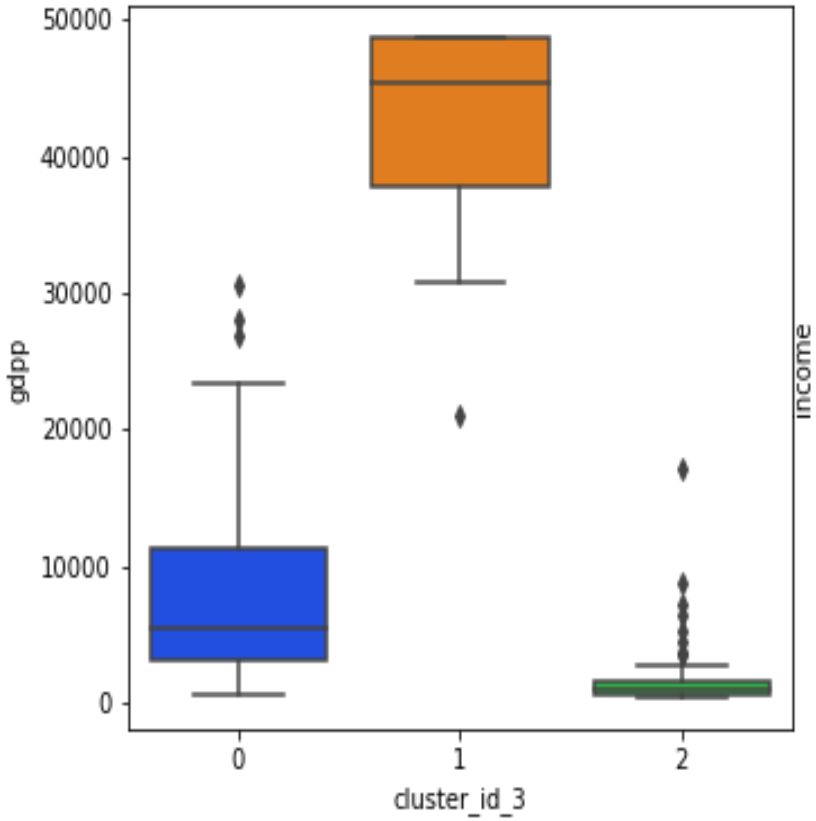


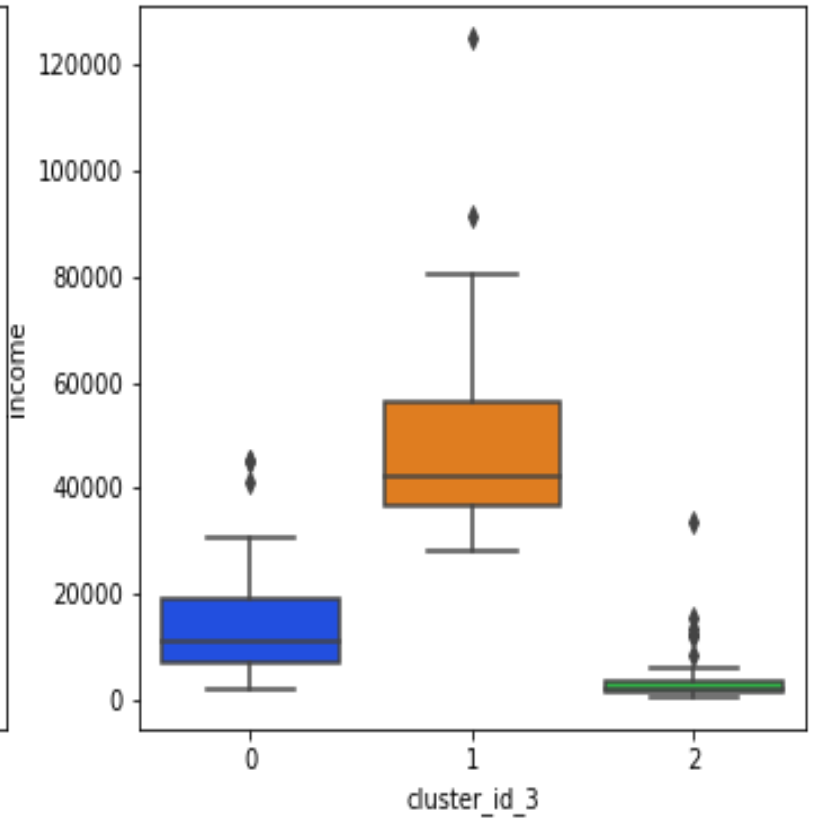
# VISUALISATIONS – CHILD_MORT VS GDPP



# BOXPLOTS OF CLUSTERS FORMED



# DECISION MAKING ON THE FINAL APPROACH

Looking at the descriptive statistics of our Under Developed Countries cluster, we could notice that some of the countries have really high income and gdpp values (looking at the max values of income & gdpp). From a business problem prespective, we would want our numbers of child mortality, income and gdpp somewhere around the MEDIAN level (We decided to go with MEDIAN and not MEAN, as there seems to be a greater variablity in the income and gdpp values). So the approach we are going to take is to filter out all countries from our original list with income & gdpp less than the Median of 1860 & 932 respectively and Child Mortality >= the Median of 90.

We will perform the filtering in the order GDPP --> INCOME --> CHILD MORTALITY

This is due to reason that, we need to identify the countries with lowest GDPP & INCOME first and then with maximum CHILD MORTALITY. This is based on the understanding that, countries with highest child mortality and having higher gdpp & income will not have any impact on the child mortality rate even after the financial aid.

# 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 48 poor countries

# RESULT

- From the list of poor countries we obtained, sorted the list on income, gdpp, child_mortality rate.
- Top 10 countries which are in direst need :

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

Based on Business needs, we can change the sort order to get different list