Clustering Assignment

By Aakash Gadge

Problem Statement

- In this assignment, we have been tasked with recommending countries to the NGO which could largely benefit from the financial aid.
- For this purpose, we have provided with historical data of countries on various factors like its GDP, Income per Capita, expenditure on exports, imports and health, and also some important parameters like child mortality rate, average life expectancy and fertility rate.
- We need to provide recommendation about countries which tend to perform weakly on 3 target parameters i.e. gdpp, income and child_mort

Methodology

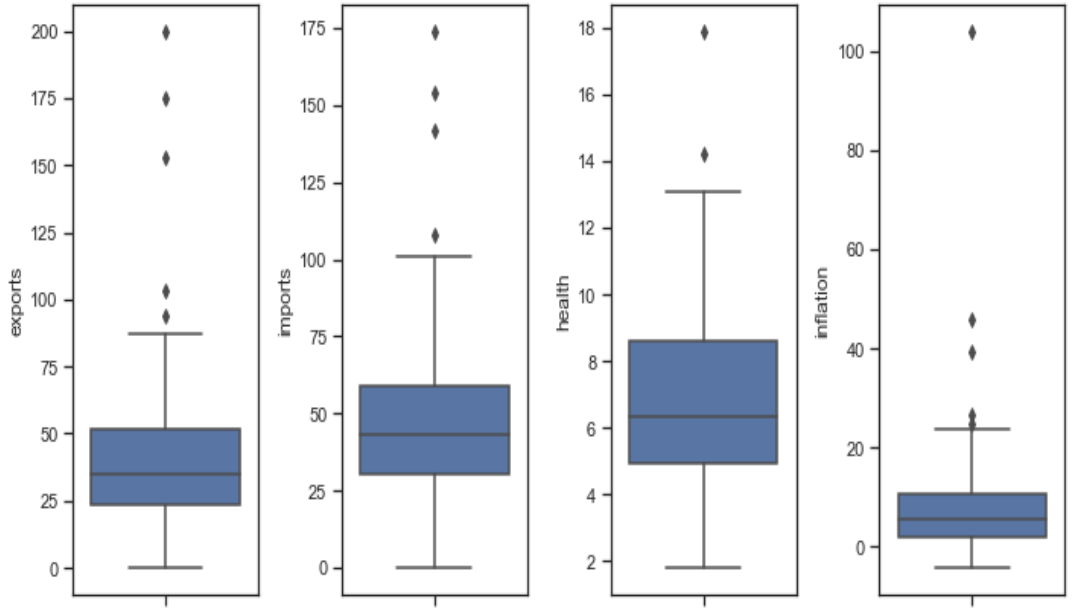
- We will approach this problem by creating clusters having similar trends and based on this classification find the countries which require immediate attention.
- To form clusters, we will be using both K-Means and Hierarchical Clustering and arrive at a solution best represented by both.
- We will determine the optimal number of clusters for K-Means using Silhouette Score
- We shall then compare the clusters formed by both the methods and check properties of each cluster from others.
- We shall inspect which cluster is our target group and then make recommendation from the selected cluster.

Steps

- Data Importing
- Exploratory Data Analysis
  - Uni-variant Analysis
  - Outlier Analysis
  - Bi-variant Analysis
  - Collinearity
- Data Preparation
- K-Means Clustering
- Hierarchical Clustering
- Cluster Analysis
- Final Recommendation

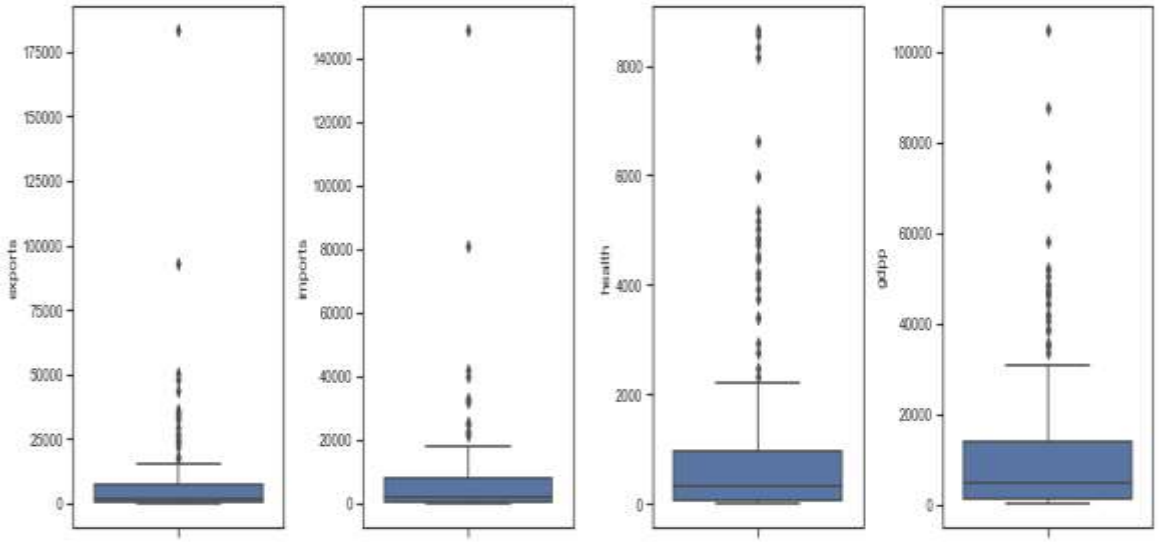
Uni-variant Analysis

- Here we are visualizing the boxplots of percentage columns to check if any outliers are present.
- Since these columns are directly dependent on the gdpp column, we shall convert these columns to absolute values and then check if any outliers are present.



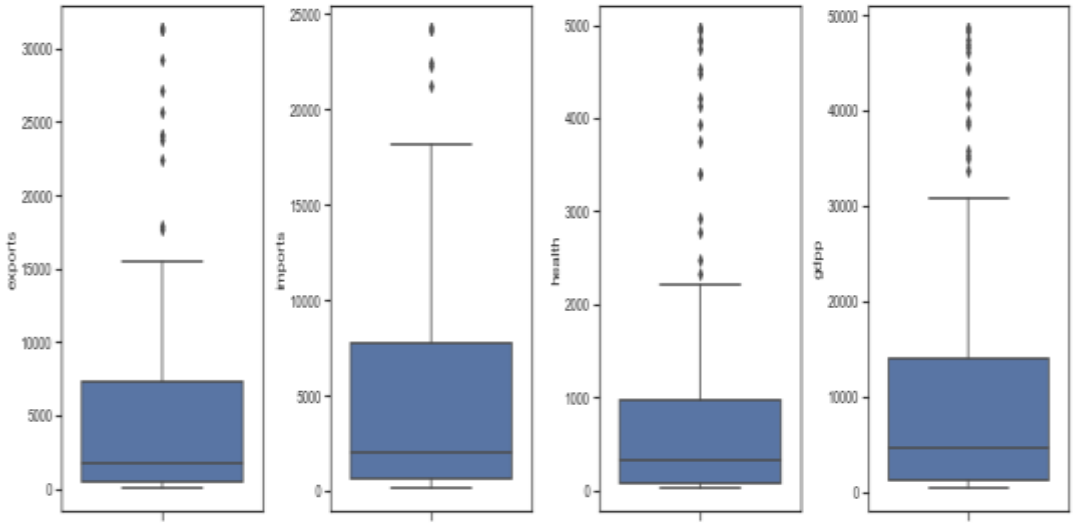
Conversion to Absolute Values

- We have converted the percentage column (exports, imports, health) to absolute values.
- Here we can clearly see the presence of outliers namely in gdpp column as well.
- Since the number of outliers is high, we won't be removing them but instead capping them.
- This way we retain information and cluster distinction is evident.



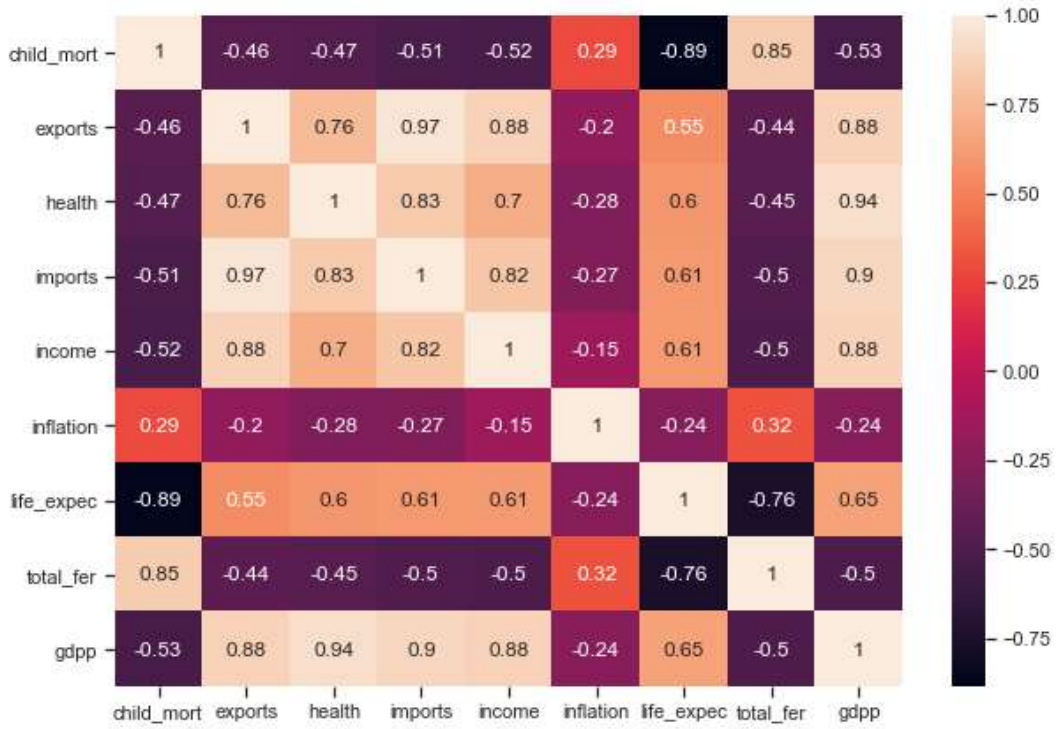
Outlier Treatment

- In order to treat outliers, we will be capping the upper limit as well as the lower limit.
- We will do this by imputing any value which is greater than 95th percentile for the given variable to be equal to the 95th percentile value.
- Similarly, we will be imputing any value which is lower than 5th percentile for the given variable to be equal to the 5th percentile.

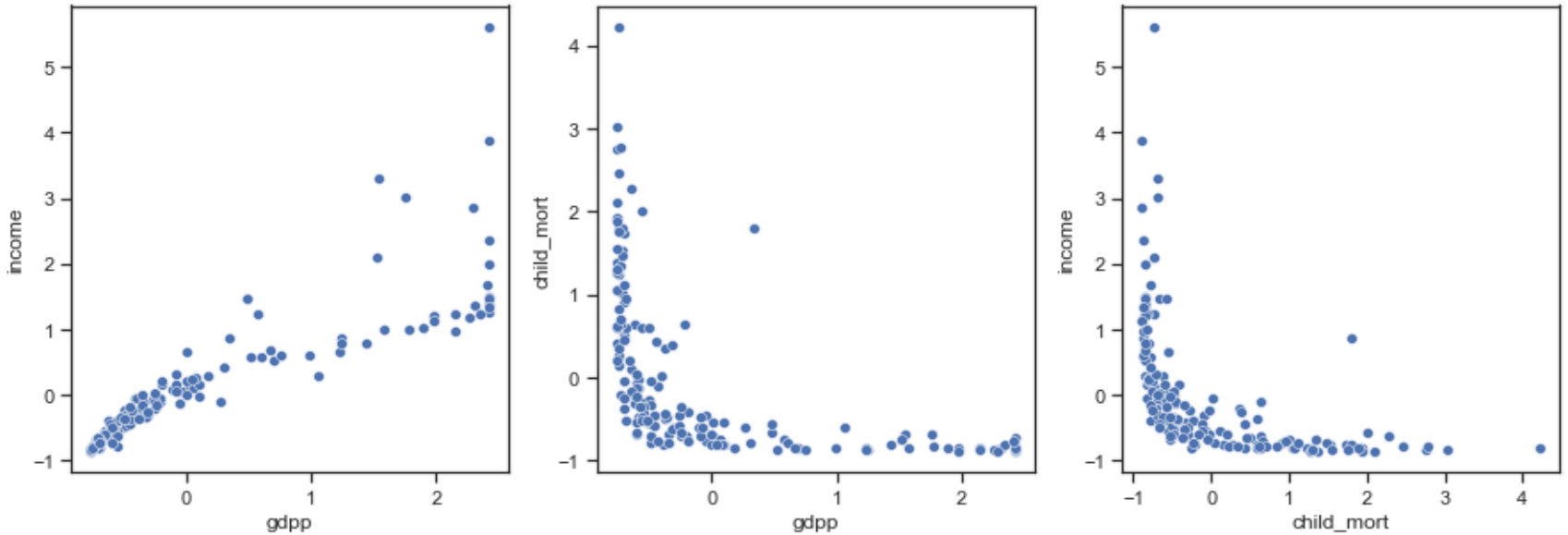


Variable Collinearity

- Here we can visualize the collinearity present amongst the variables in the dataset provided.
- Highly collinear variable (health, imports, exports, income) with “gdpp” is because of the dependency.
- “child_mort” has negative co-efficient with “gdpp”
- And so does, “total_fert” and inflation.



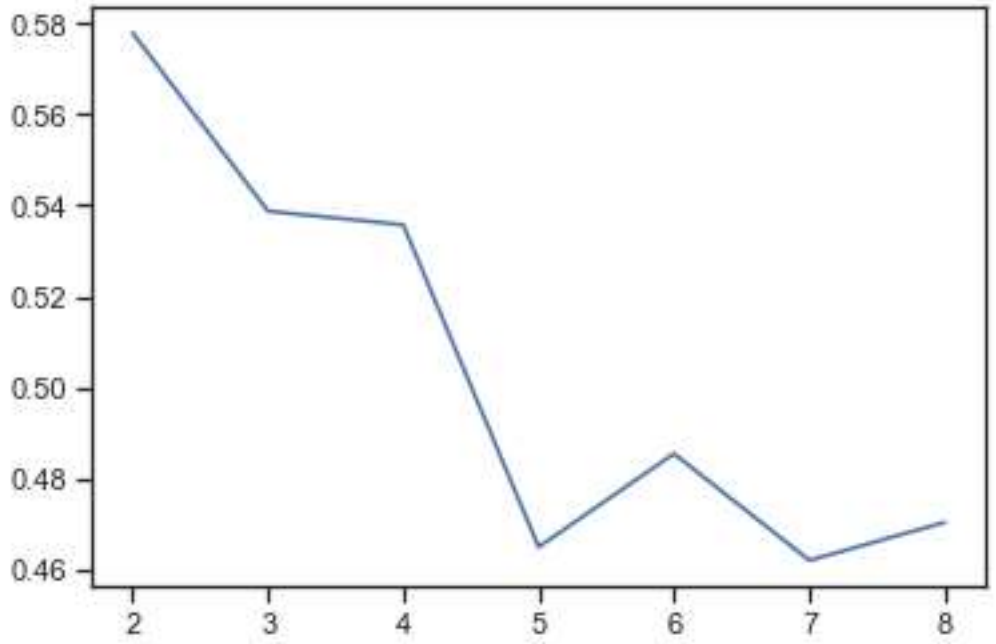
Bi-Variant Analysis



- We can observe that “gdpp” and “income” tend to follow a positive linear trend.
- As the gdpp decreases, the child_mort rate tends to increase.
- Similarly, as the income decreases, child_mort rate tends to increase.

Silhouette's Score

- To obtain the optimal value of k for kmeans clustering, we have tried to plot the silhouette score for k-clusters in this graph.
- We can observe that k=2 has the highest score of around 0.58 but having just 2 clusters isn't quite beneficial for our business problem.
- Hence, we choose k=3 as has the next highest score of 0.54.
- We will performing our further analysis with k=3.

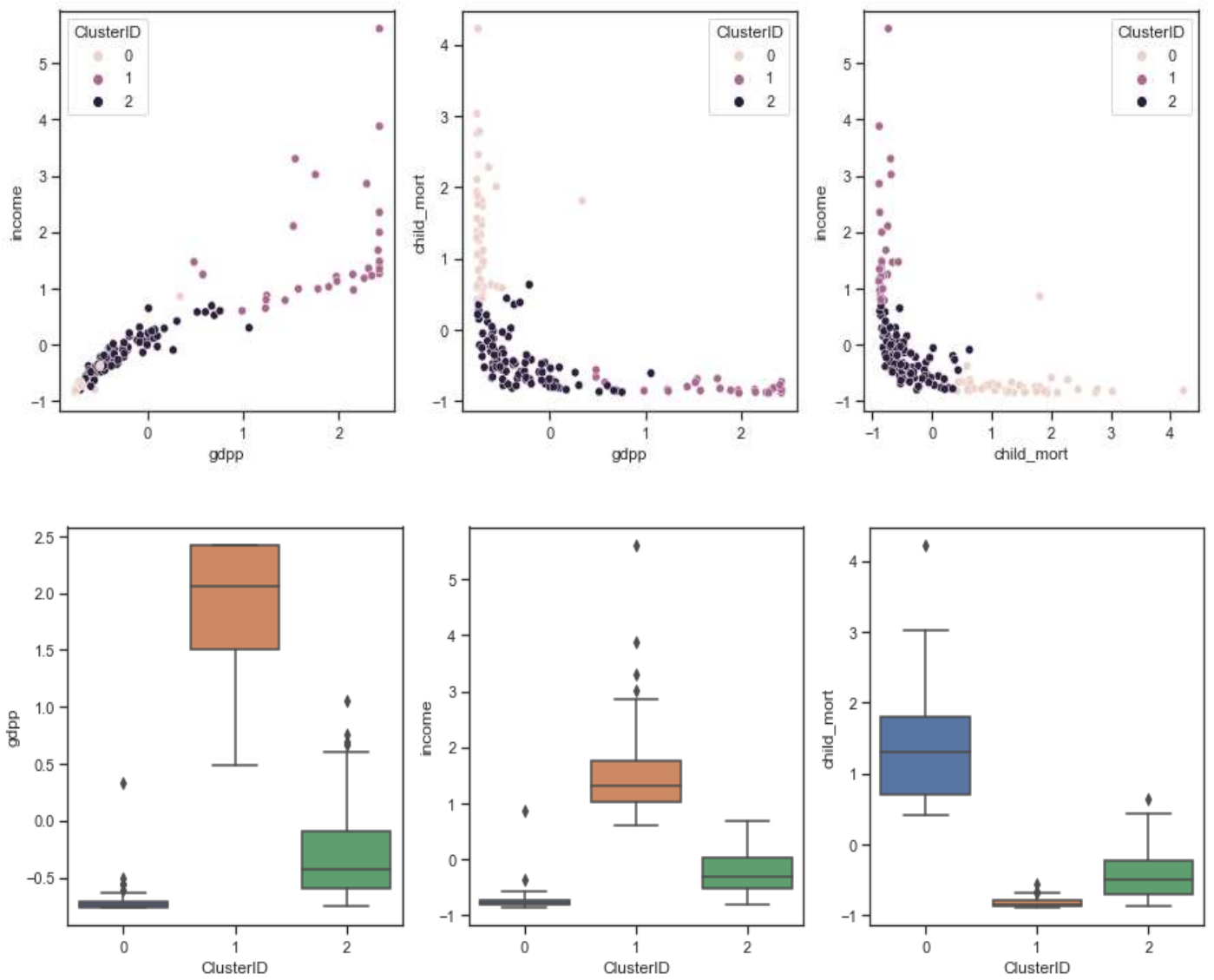


K-Means Clustering Analysis

- Having performed our K-Means clustering by choosing k=3, we have obtained the clusters, and can be visualized here.
- We can see that cluster 1 has high mean gdpp while cluster 0 has the lowest mean gdpp.
- Similarly, mean of income of clusters is as follows:

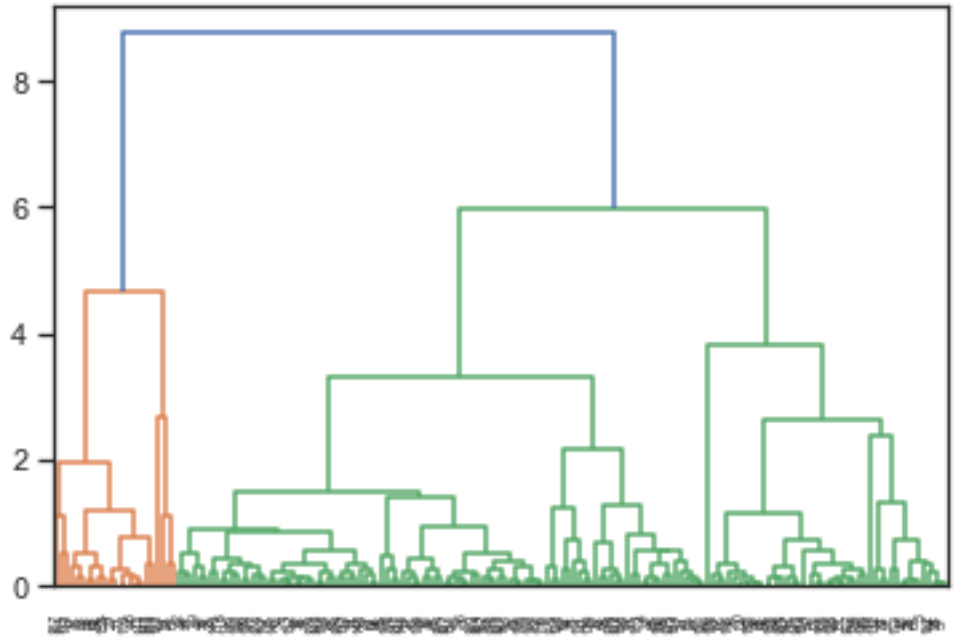
  cluster 1 > cluster 2 > cluster 0
- For gdpp column we can observe that mean of child_mort for the clusters is as follows:

  Cluster 0 > cluster 2 > cluster 1



Hierarchical Clustering

- We will also try to form clusters using hierarchical clustering, as this gives us better output as compared to Kmeans clustering.
- You can visualize the dendrogram formed using the complete method.
- At height of 5, we can obtain 3 clusters which seems to be optimal number of clusters.

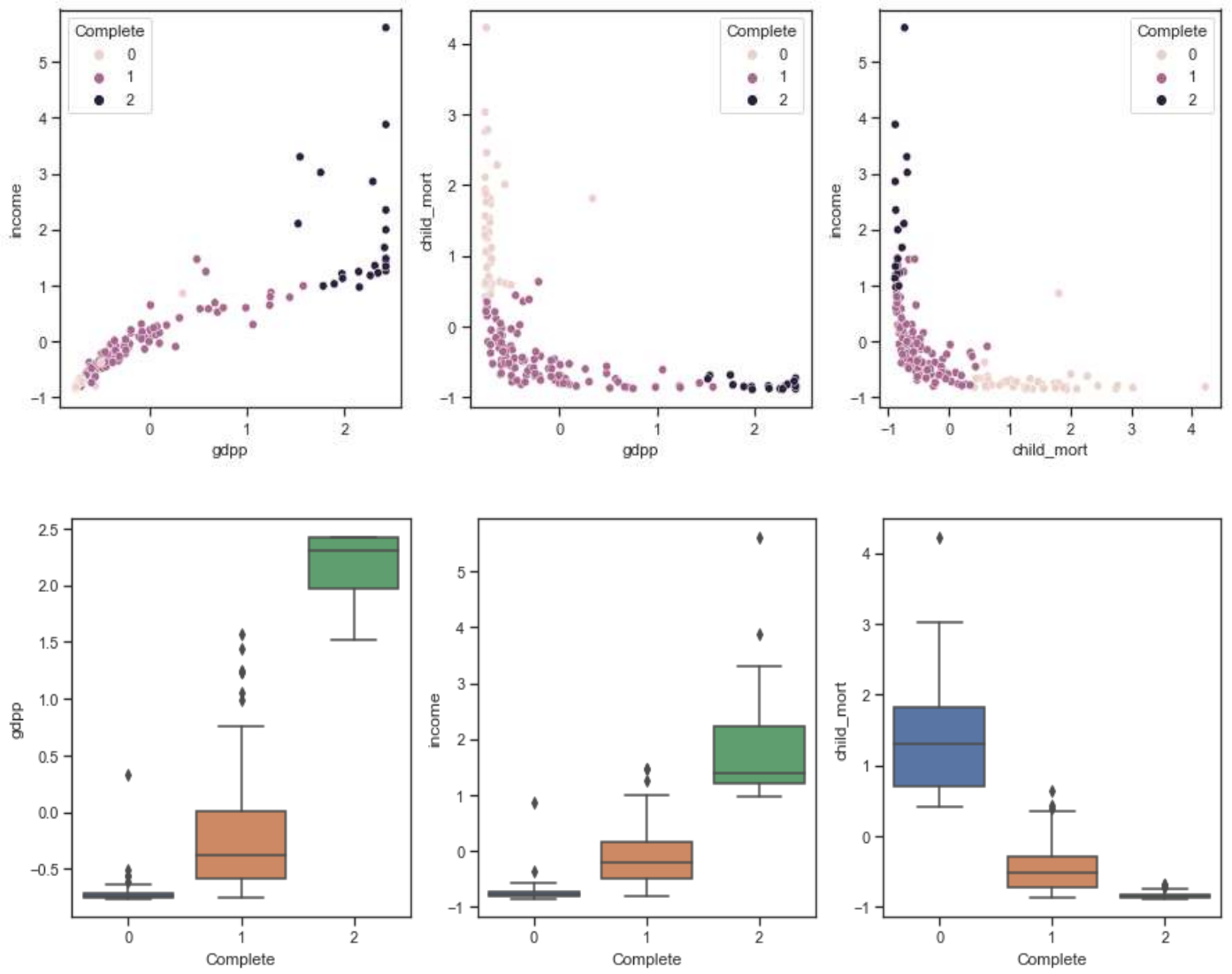


Hierarchical Clustering Analysis

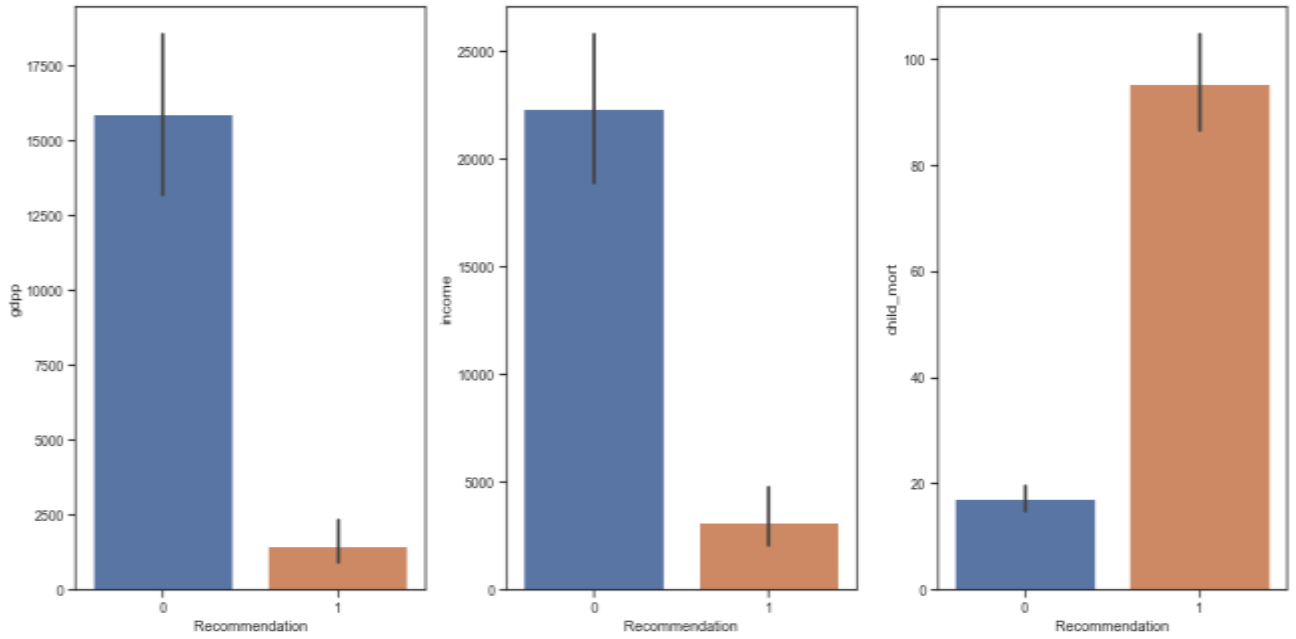
- Having performed our Hierarchical clustering by choosing cut_tree at 3, we have obtained the clusters, and can be visualized here.
- We can see that cluster 2 has high mean gdpp while cluster 0 has the lowest mean gdpp.
- Similarly, mean of income of clusters is as follows:

  cluster 2 > cluster 1 > cluster 0
- For gdpp column we can observe that mean of child_mort for the clusters is as follows:

  Cluster 0 > cluster 1 > cluster 2



Recommendation Analysis



- Based on the clusters formed by hierarchical clustering using complete method, we have selected cluster 0 as our target group and created a new column Recommendation which is assigned value 1 for cluster 0 and 0 for all the other clusters.
- Here we can visualize the difference between the gdpp, income and child_mort for the recommended datapoints, which further supports our hypothesis.

Final Recommendation

- Based on our analysis, we can recommend countries which would largely benefit from the financial aid, which are sorted by highest child mortality rate, lowest income per capita and lowest gdpp.

- The output of top10 such countries is as follows:
  - 'Haiti'
  - 'Sierra Leone'
  - 'Chad'
  - 'Central African Republic'
  - 'Mali'
  - 'Nigeria'
  - 'Niger'
  - 'Angola'
  - 'Congo, Dem. Rep.'
  - 'Burkina Faso'

Conclusion

- With the final recommendation of top 10 countries being made, we can conclude this assignment.
- To summarize,
  - We started by loading the dataset and checking its features
  - Looked at the variable to understand the business problem
  - Checked for presence of any outliers.
  - Performed univariant and bivariant analysis.
  - Checked for collinearity between variables.
  - Scaled the data for modelling.
  - Found the optimal number of k using Silhouette's score analysis.
  - Performed K-Means clustering.
  - Performed Hierarchical clustering.
  - Using the final clusters obtained, we made the final recommendations.