

COUNTRIES IN NEED!

Problem Statement:

- Find top 5 Countries which are in need of immediate AID of money.
- Use the factors such as **GDP per Capita**, **Income per person** and **Child Mortality** to form the clusters and find the countries.
- Perform **PCA** and **Clustering Algorithms** to obtain the desired objective.

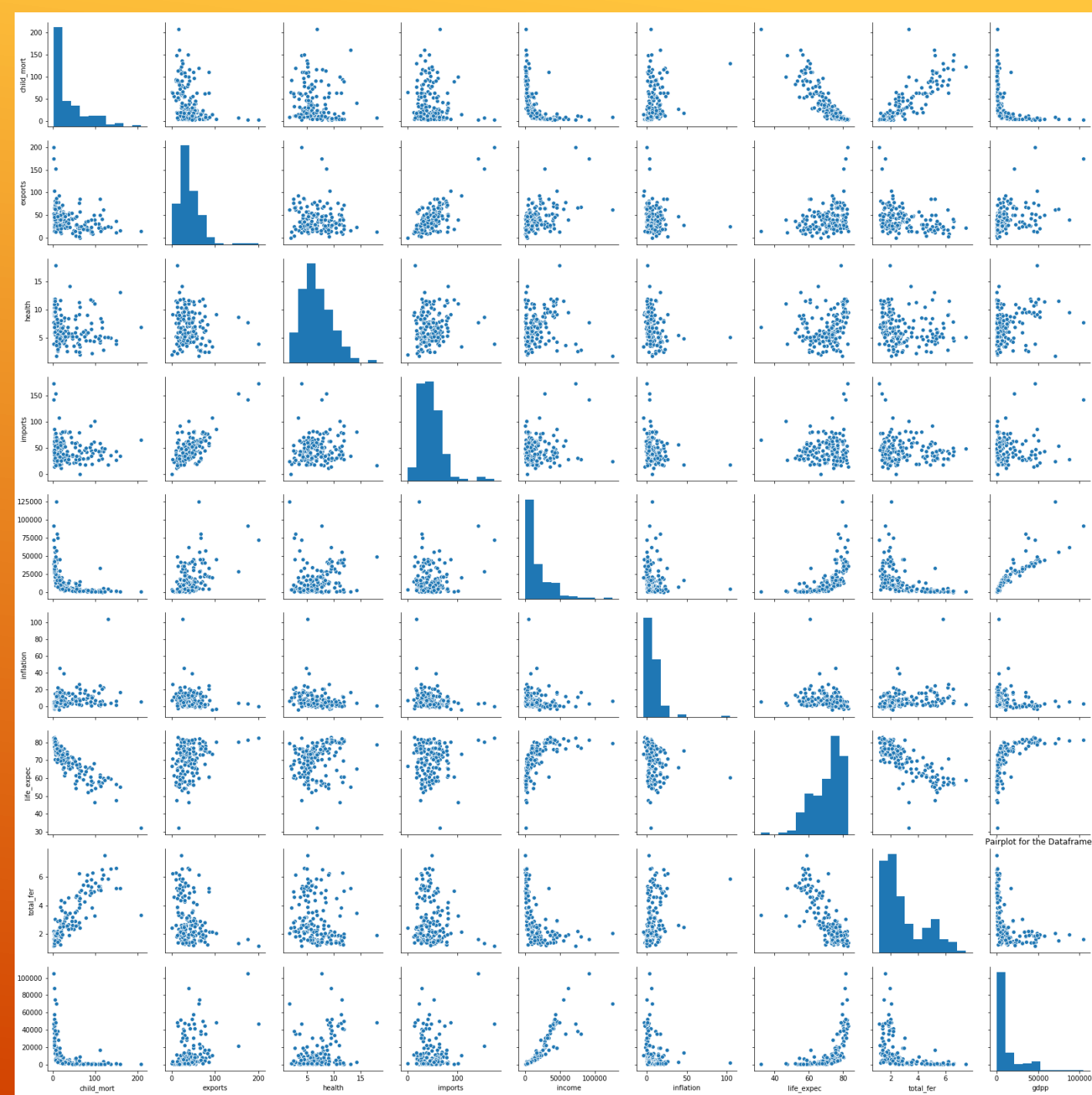
- ▶ The data set has a shape of 165 rows and 10 columns, out of which 9 are Numerical Columns!
- ▶ The pairplot and Outliers gave us some usefull insights and helped in scaling the dataframe to a better statistic.
- ▶ The Hopkins test performed in order to verify that the dataset is suitable for Clustering or not gave a score of 0.91, which indicates that clustering algorithm technqiues can be applied over the data set.
- ▶ The High Correlated pairs are :
 - ▶ Child_Mort is correlated with Total_fer with value of 0.85
 - ▶ Exports is correlated with Imports with value of 0.74
 - ▶ Income is correlated with GDPP with value of 0.9
 - ▶ Life_Expec is correlated with Income with value of 0.61
 - ▶ Life_Expec is correlated with GDPP with value of 0.6

KNOWING ABOUT DATA!

PAIR PLOT FOR THE DATA FRAME!

Inferences:

1. We observe Linear Relationship between most of the variables which displays the MULTICOLLINEARITY and which we can deal by using PCA.
2. We can see that we have outliers in our dataframe, maybe around 5%, and treating them won't hurt our data variance, hence we will go onto do that.
3. Finding the countries which are in desperate need of AID wouldn't be easy, so we need the clustering algorithms for solving the problem statement.





Highly correlated variables/pairs are:

1. Exports – Imports = 0.74
2. Income – Life_Expect = 0.61
3. Income – GDPP = 0.9
4. Child_mort – Total_fer = 0.85

To overcome such high positive correlation we will use the technique of **Principal Component Analysis**.

But before we do that, we will just try to perform the Kmeans on non PCA treated dataframe!

Let's remove the OUTLIERS!

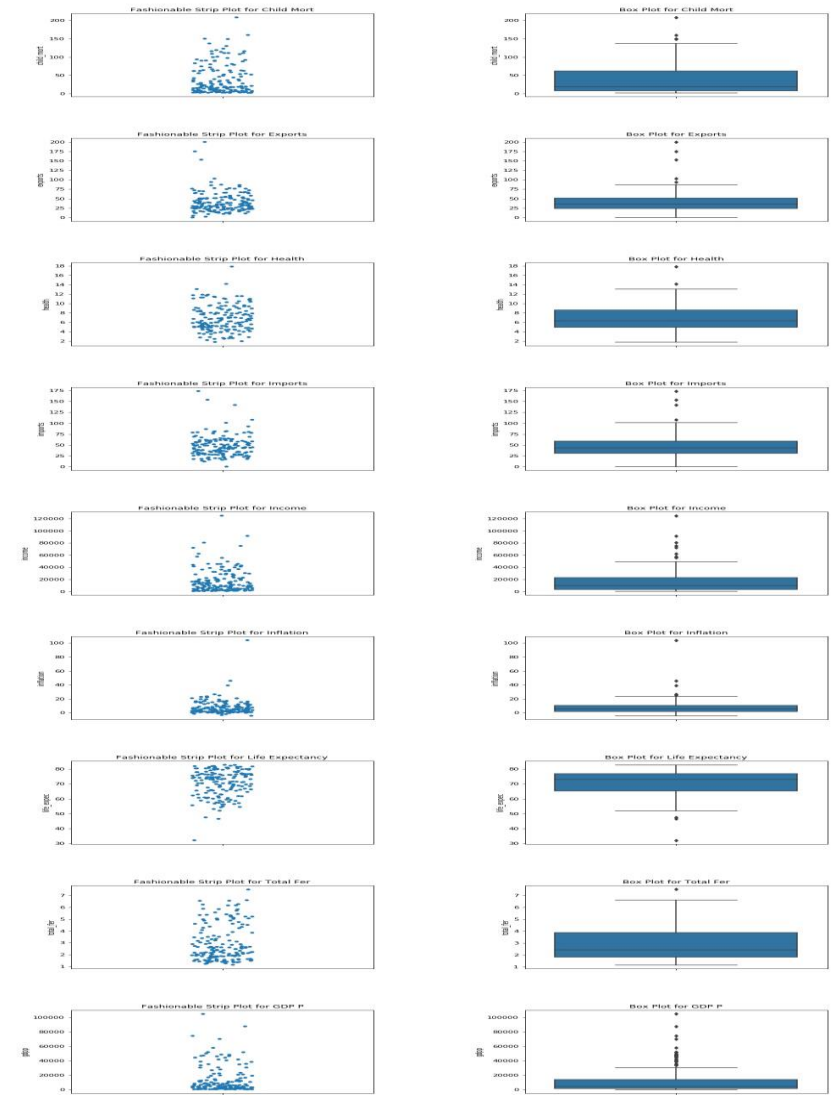
CORRELATION!

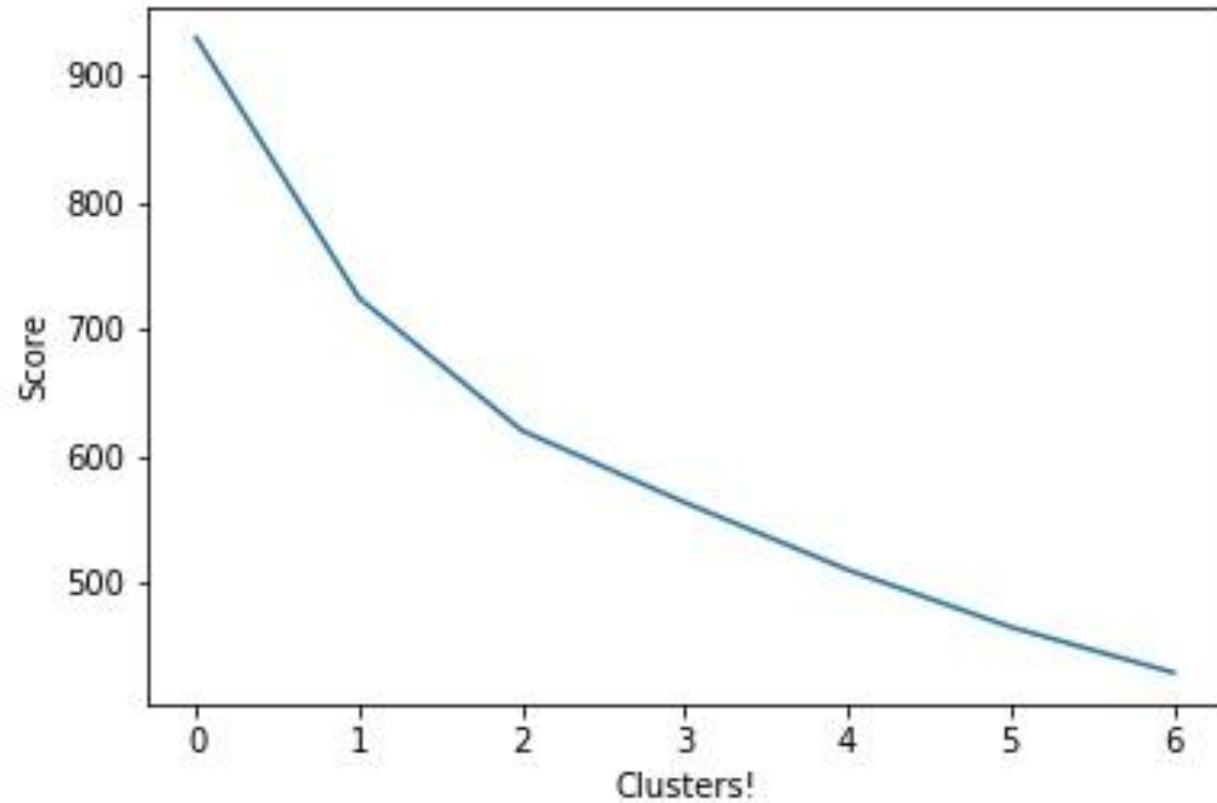
LOOKING AT THE BOX PLOT FOR OUTLIERS!

The order in which the Outliers were visualized are:

1. Strip Plot – Box Plot for Child Mortality.
2. Strip Plot – Box Plot for Exports.
3. Strip Plot – Box Plot for Health.
4. Strip Plot – Box Plot for Imports.
5. Strip Plot – Box Plot for Income.
6. Strip Plot – Box Plot for Inflation.
7. Strip Plot – Box Plot for Life Expectancy.
8. Strip Plot – Box Plot for Total Fer.
9. Strip Plot – Box Plot for GDP per Capita

We used the STATISTICAL approach for treating outliers, which says that only keep those values which have absolute ZSCORE value less than 3 on the COLUMNS AXIS.





Elbow Curve Inferences:

1. The Significant drop can be observed at cluster 3, and the business logic also says that we have 3 class of Population => Lower Class/Middle Class and Upper Class.
2. The Silhouette score also says that for Cluster $K=3$, score is 0.27 after which there is no increase in the score.

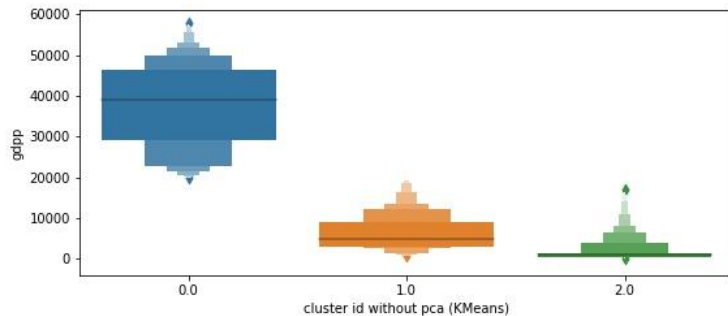
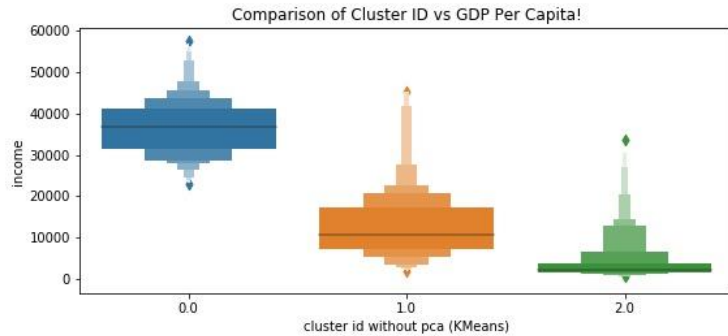
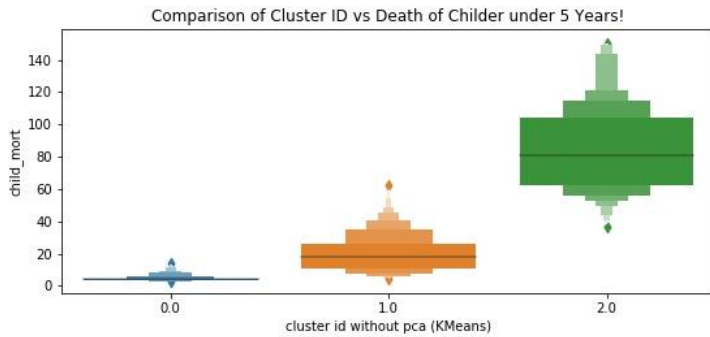
OPTIMUM NUMBER OF CLUSTER SELECTION!

JUST FOR FUN!

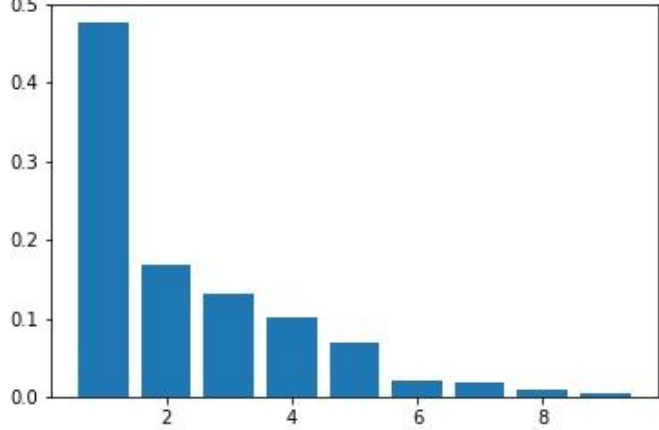
The advantages of this FUN Exercise are:

1. We get to see if the **Outlier Treatment** has benefitted our DATAFRAME! And it has!
2. Next we get to learn/implement the KMEANS for exercise purpose.
3. Identify the clusters which we can get after PCA treatment.
4. See the respective Values for each variable.

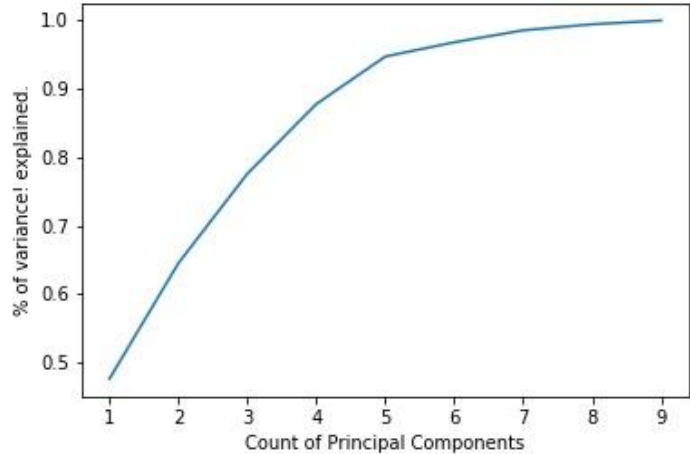
Now let's begin the PCA Components Selection and Cluster Selection!



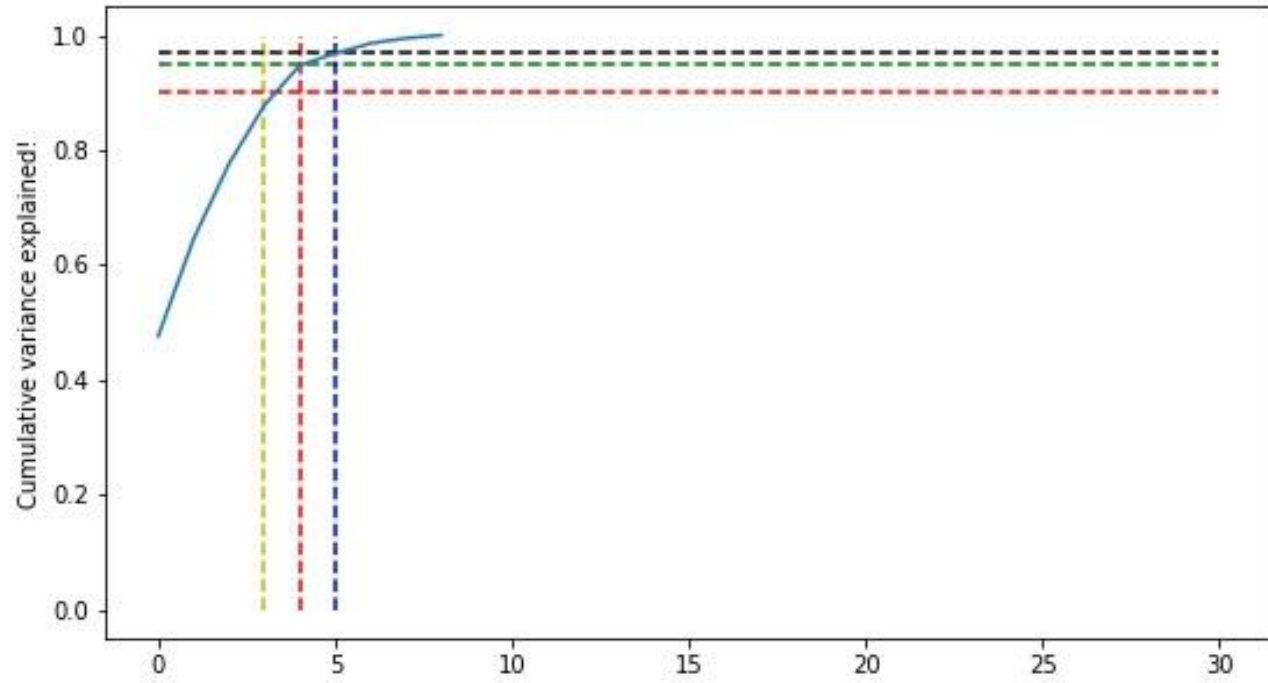
Explanation of Variance by the formed Principal Components!



Scree plot



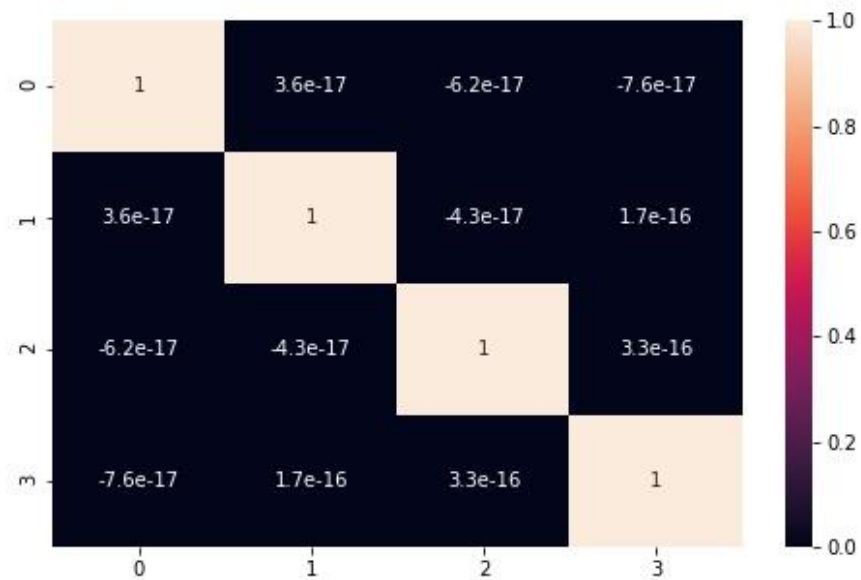
Variance Plot for different PC's to choose the optimum number of Principal Components!



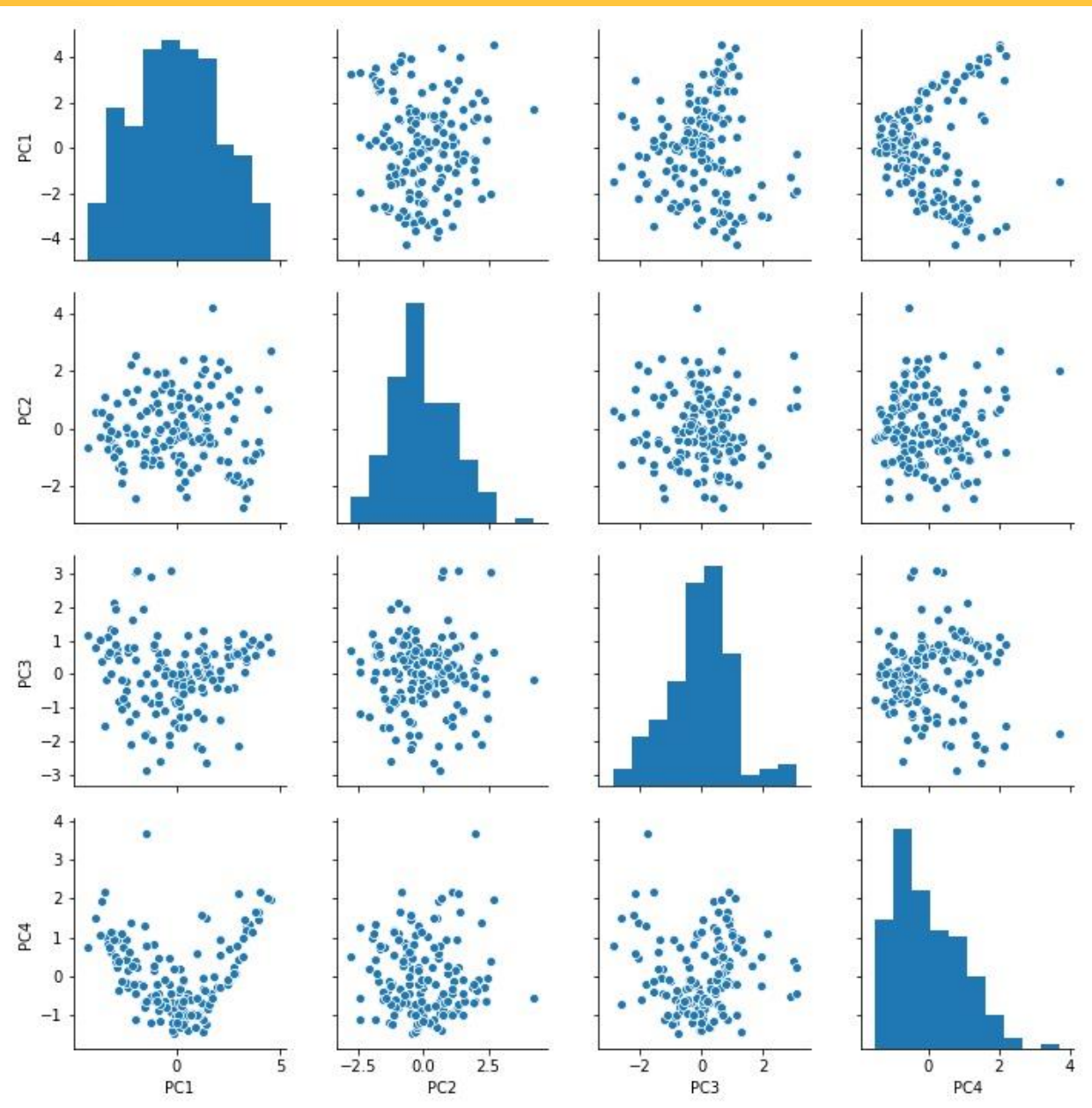
We choose **4 Principal Components!** Because the Maximum variance of 95% is explained by 4 Principal Components.

Moving on in the problem statement, we will use **4 PC's** for our analysis of clustering.

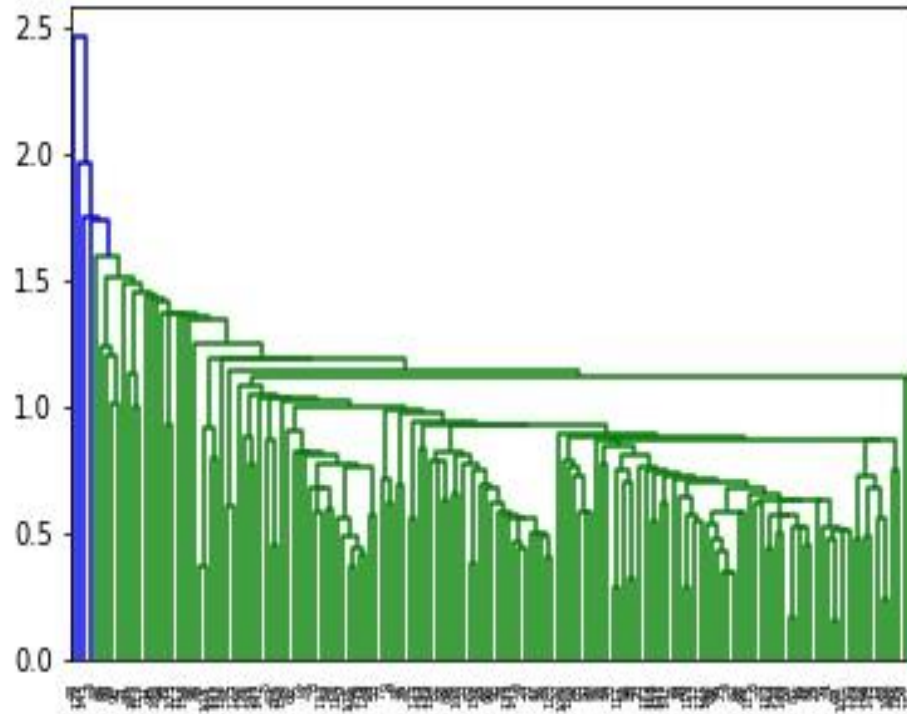
PCA APPLICATION



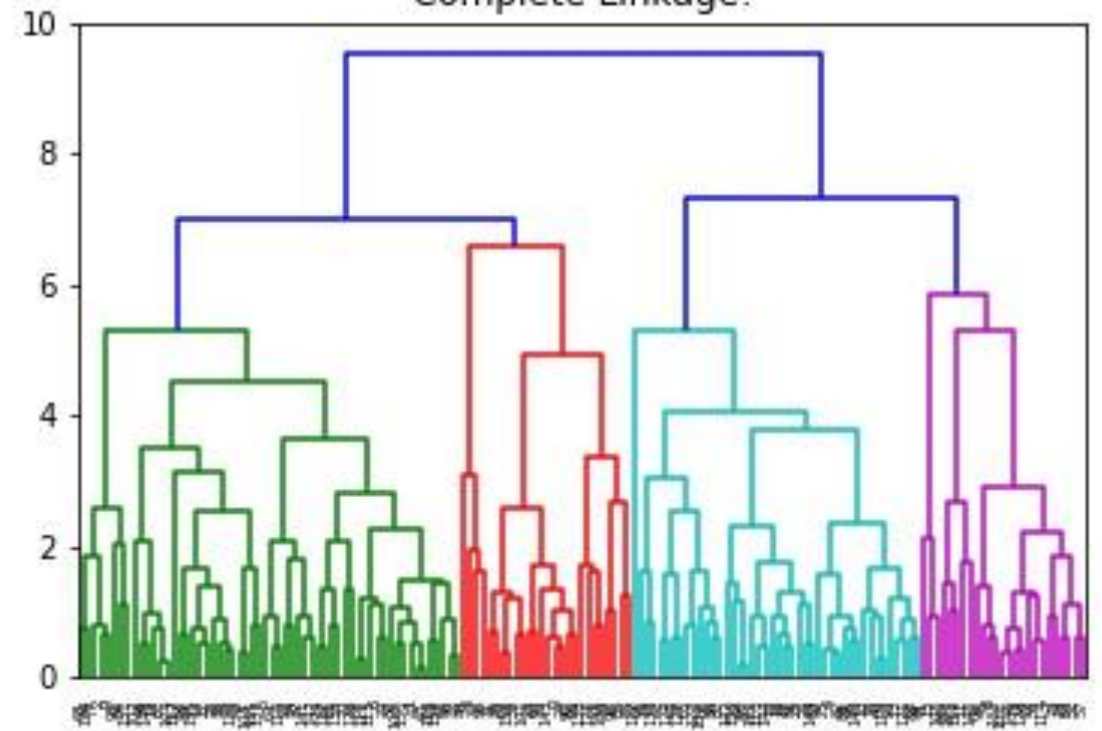
CORRELATION AND PAIR PLOT OF SELECTED PC'S



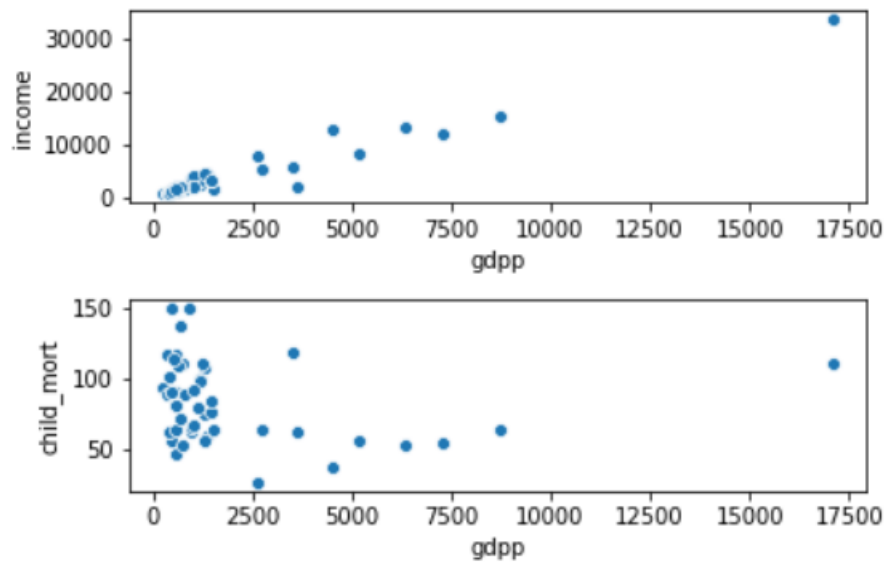
Single Linkage!



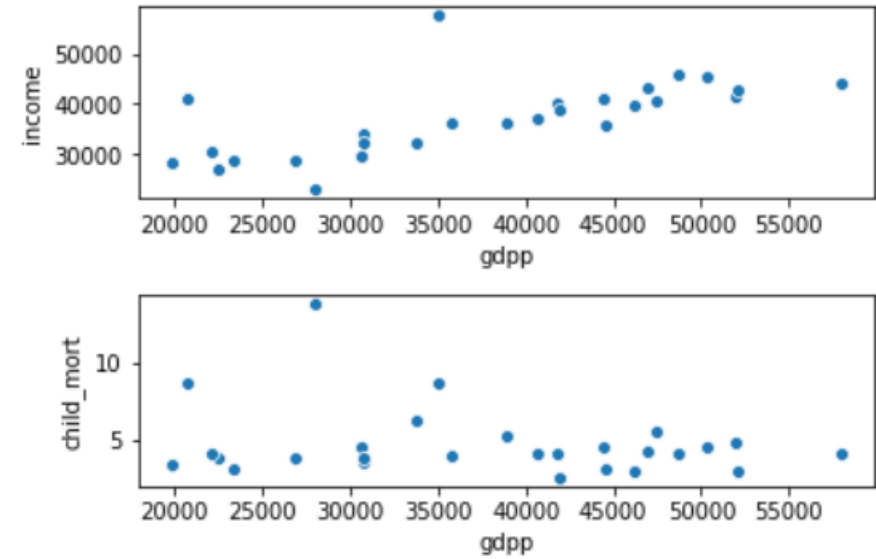
Complete Linkage!



HIERARCHICAL CLUSTERING DENDROGRAMS
VISUALIZATION!

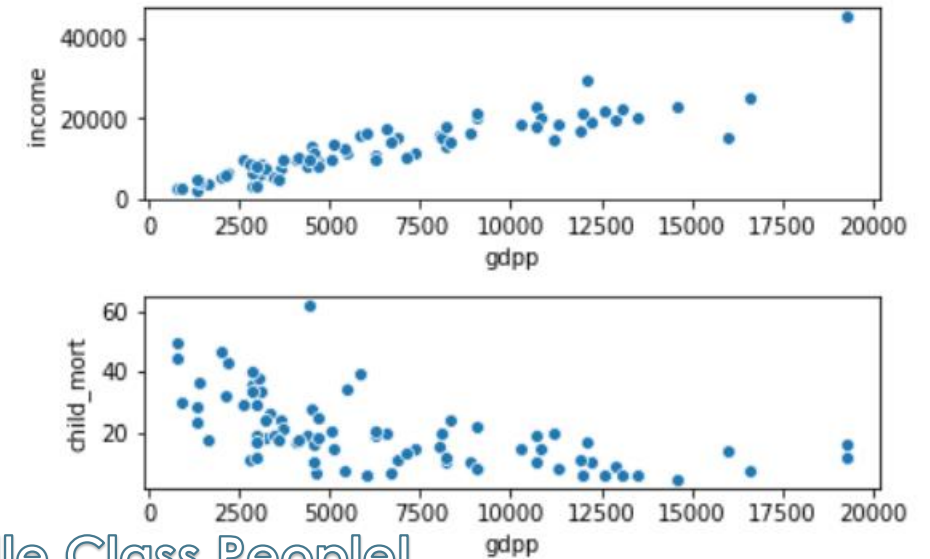


Cluster 0= > Lower Class People!

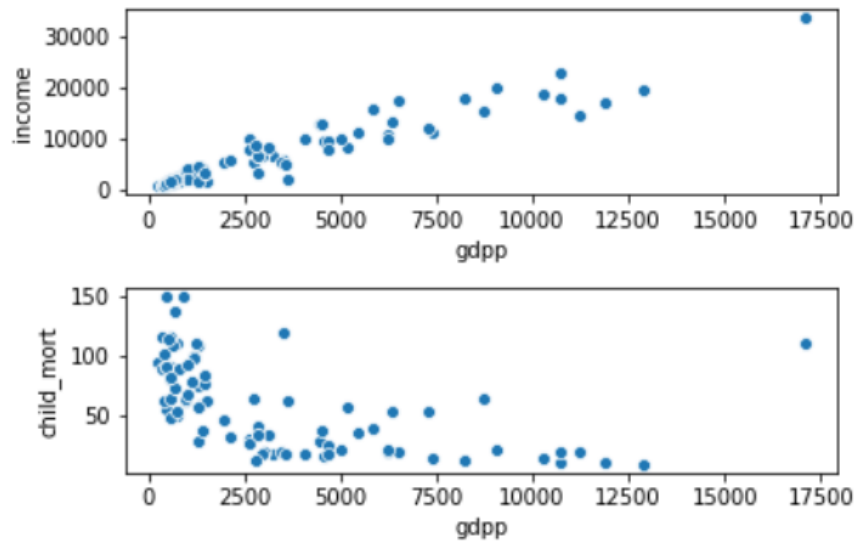


Cluster 1= > Upper Class People!

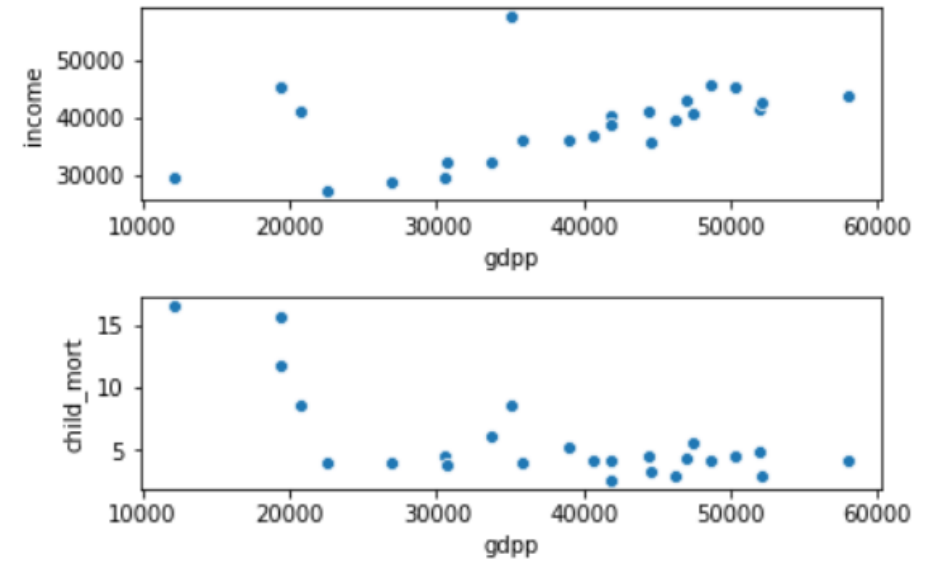
K MEANS VISUALIZATION FOR
CHILD MORTALITY, INCOME
AND GDP PER CAPITA!



Cluster 2= > Middle Class People!

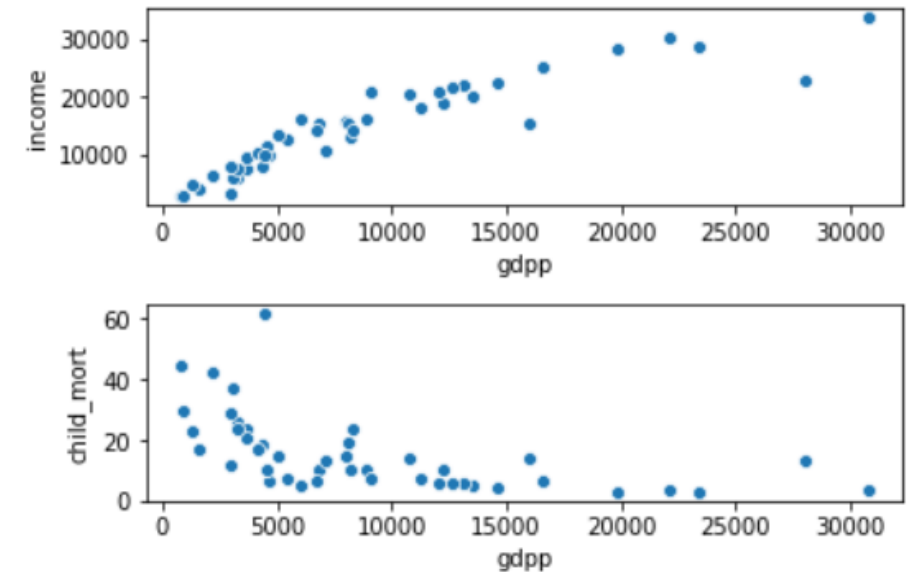


Cluster 0= > Lower Class People!



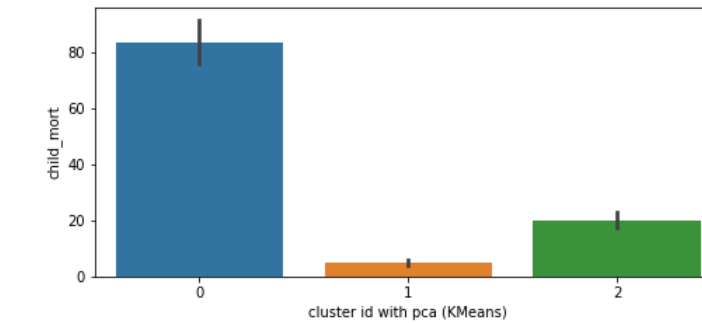
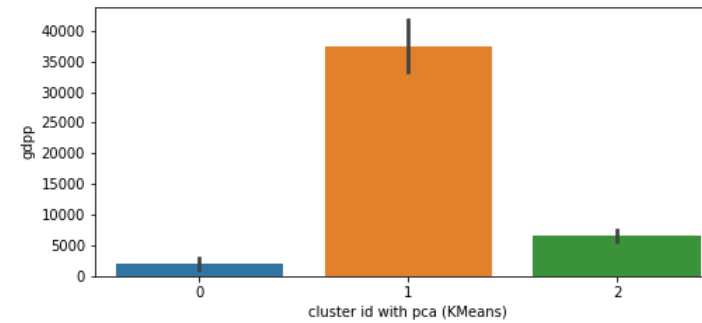
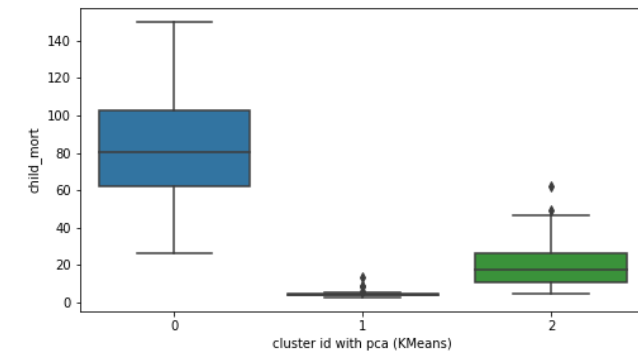
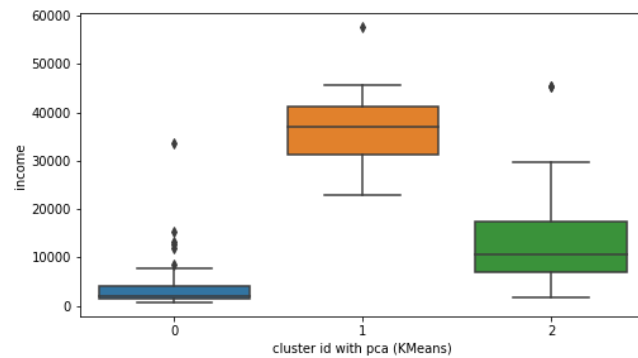
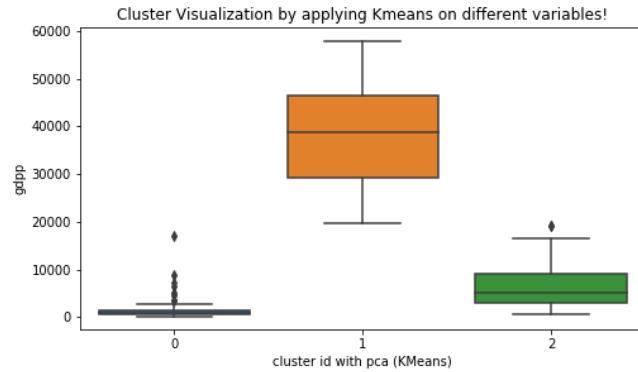
Cluster 1= > Upper Class People!

HIERARCHICAL CLUSTERING VISUALIZATION FOR CHILD MORTALITY, INCOME AND GDP PER CAPITA!

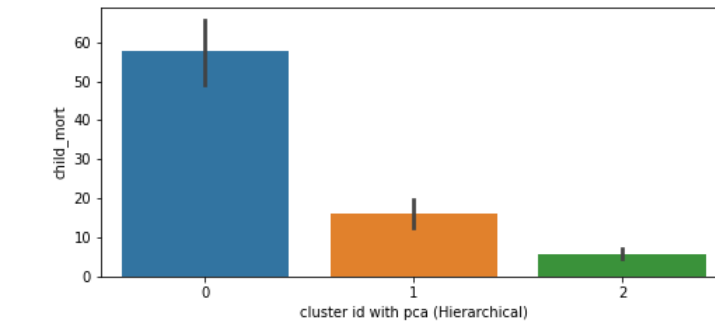
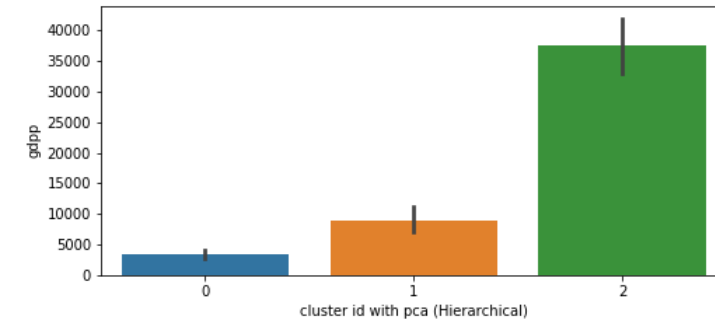
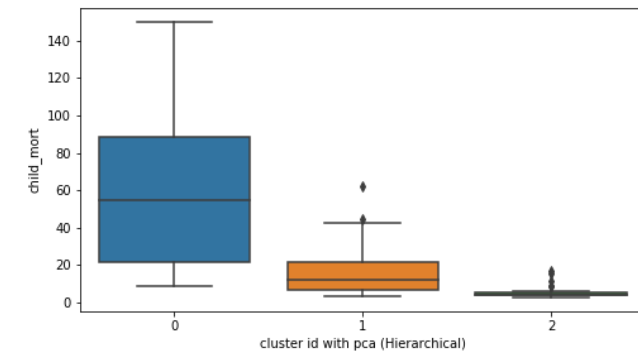
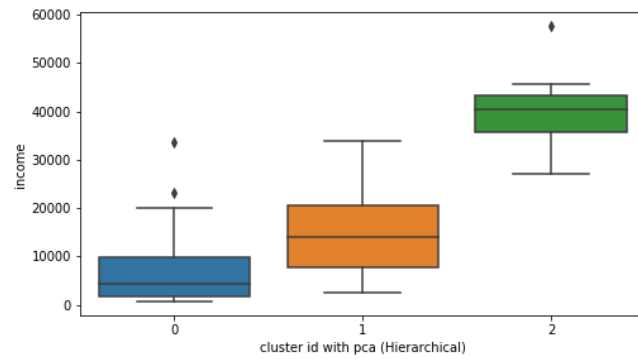
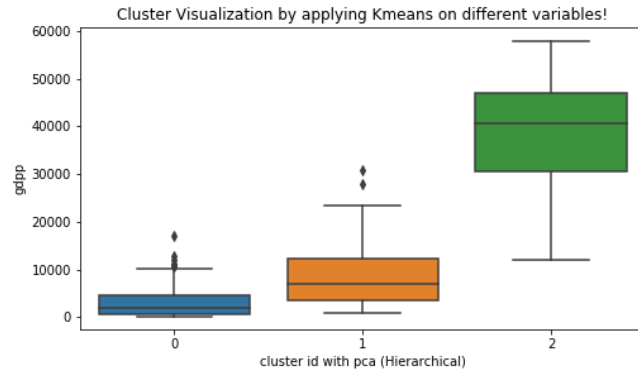


Cluster 2= > Middle Class People!

KMEANS PLOT FOR CLUSTERS!



HIERARCHICAL CLUSTERING PLOT FOR CLUSTERS!



OUR SELECTION:

1. The cluster on which we need to focus is Cluster 0.
2. This cluster contains **High Child Mortality ratio** and comparatively **Low Income and Low GDP per capita** from all the clusters.

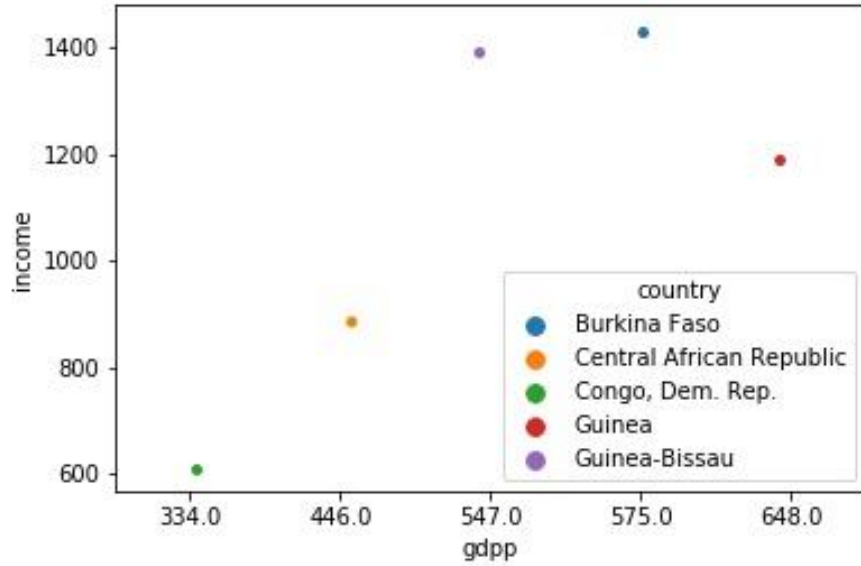
We are going to choose the countries based on the mean values for the selected clusters.

Approach:

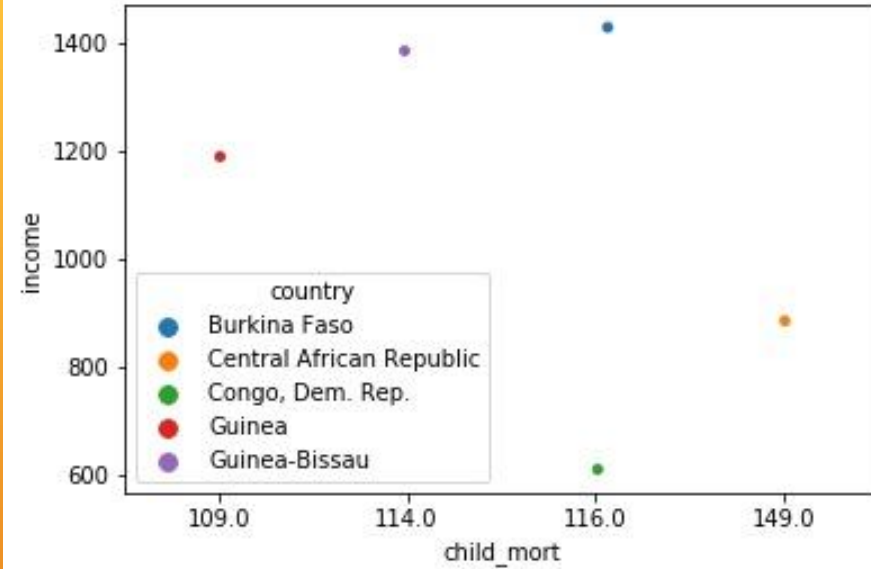
- We take the mean of all the variables(Child Mortality/Income/GDPP) and create a criteria.
- Criteria 1 : The DataFrame should be selected such that the Income values and GDPP values should be lower than the mean values of respective 2 variables.
- Criteria 2 : The dataframe selected from criteria 1 should be used in **AND** condition where the Child Mortality value should be greater than the mean value of the Child Mortality value.

The advantage is that **we will get all the countries where CHILD MORTALITY is HIGH and INCOME/GDPP is low.**

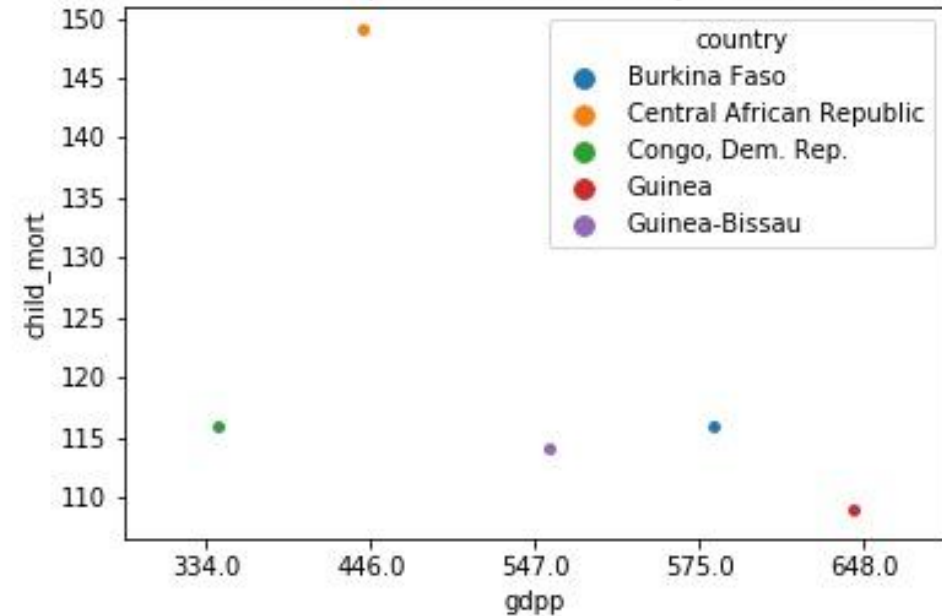
GDP Per Capita vs Income! in KMEANS



Income vs Child Mortality! in KMEANS

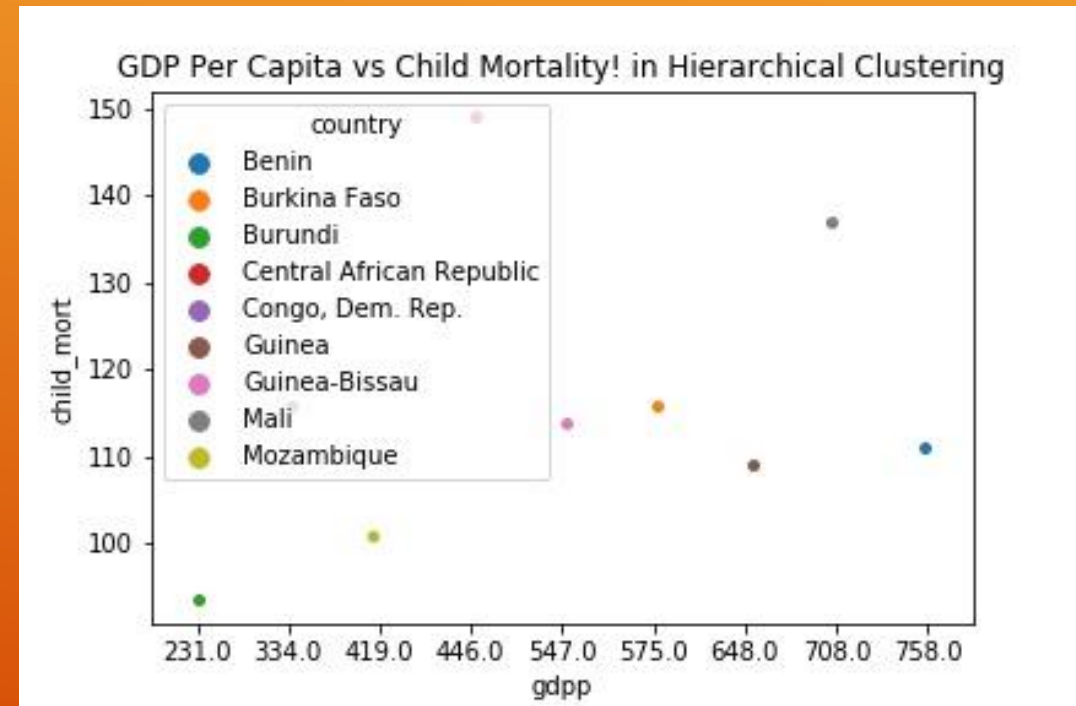
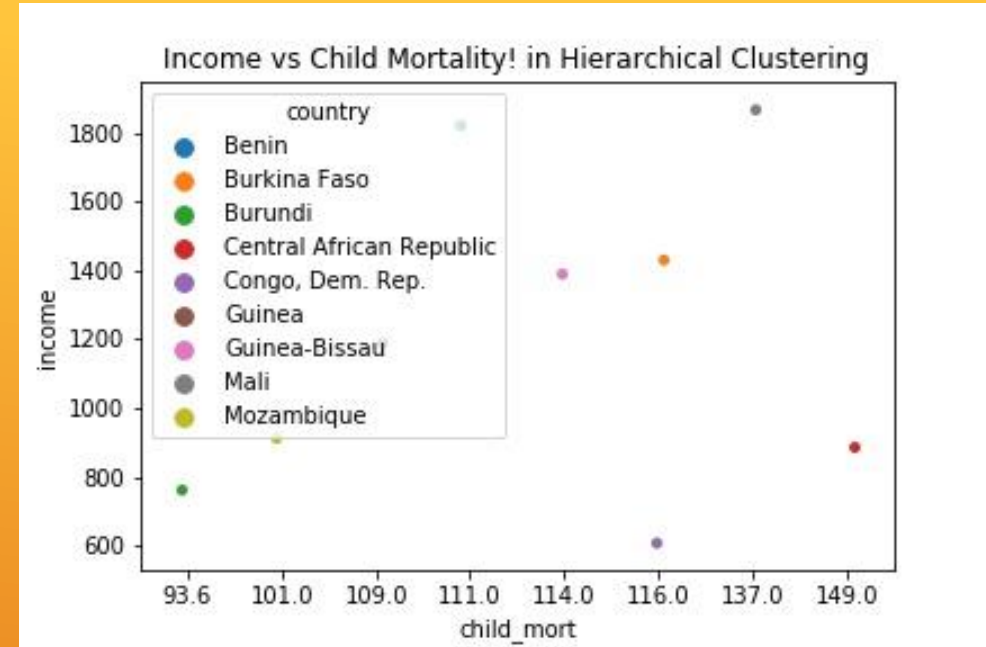
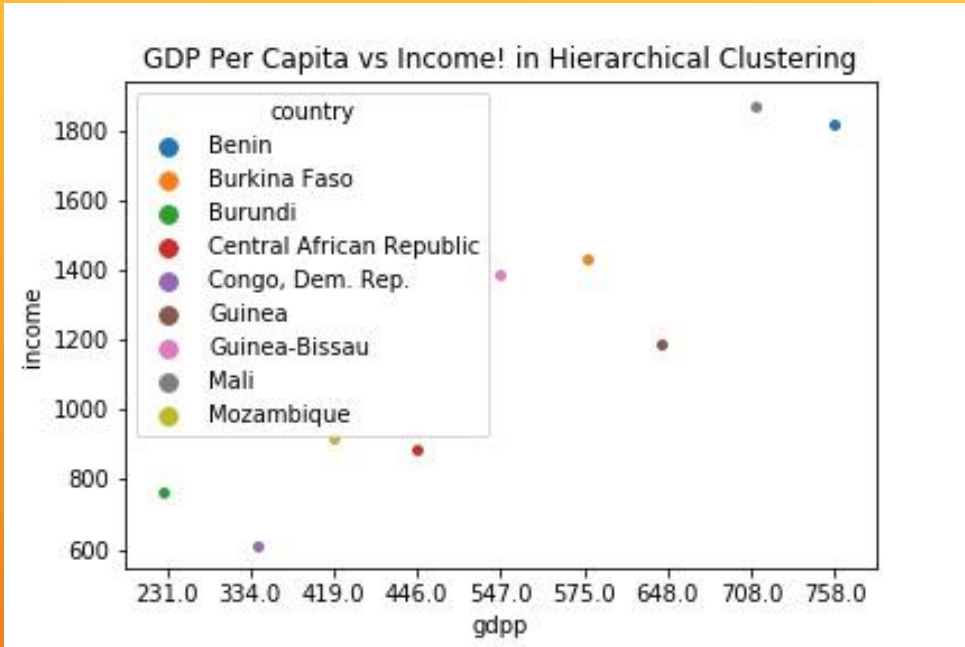


GDP Per Capita vs Child Mortality! in KMEANS



KMEANS SELECTED COUNTRIES

HIERARCHICAL CLUSTERED SELECTED COUNTRIES



► **So the TOP 5 Countries which are similar in both the models are, but we will choose the KMEANS Model for our analysis:**

- Burkina Faso
- Central African Republic
- Congo, Dem. Rep.
- Guinea
- Guinea-Bissau

Remedies which we can provide:

1. Provide proper institutions governed by honest and educated bodies.
2. Interact with the population and provide more employment for these under developed countries.
3. Share the world knowledge and make them aware about the current world scenarios.
4. Help the poor by upskilling them which brings them above the Labour status.
5. Provide medical embodiments which are always up for helping in disastrous situations.

FINAL CONCLUSION