Assignment: Clustering and PCA

SOCIO-ECONOMIC CLASSIFICAITON

Devesh Singh
Date 19/08/2019

Introduction

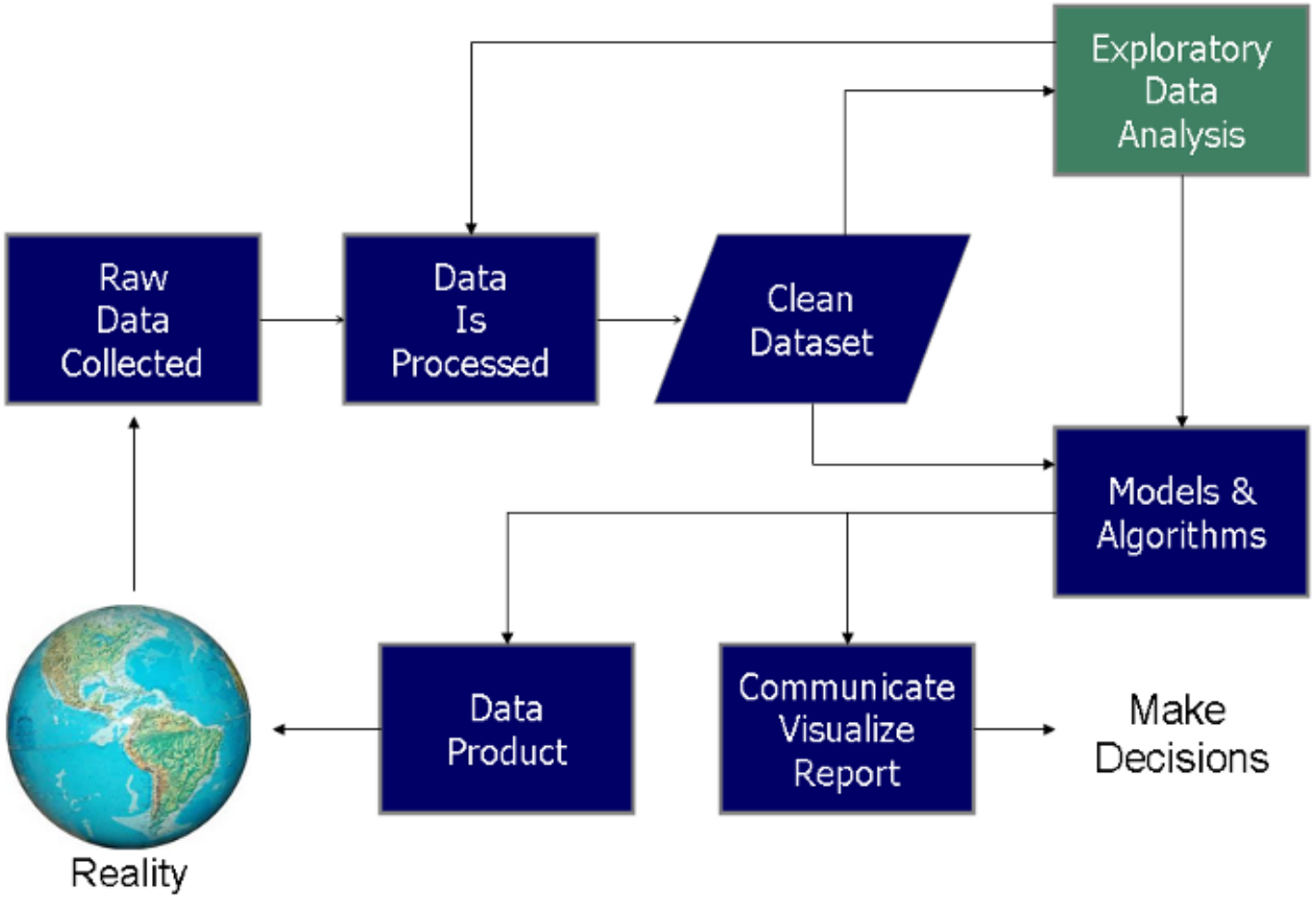
Data

The datasets containing those socio-economic factors and the corresponding data dictionary are provided.

Problem Statement

- To categorise the countries using some socio-economic and health factors that determine the overall development of the country.
- Suggest the countries which HELP NGO needs to focus for there improvement.
- Recommendation

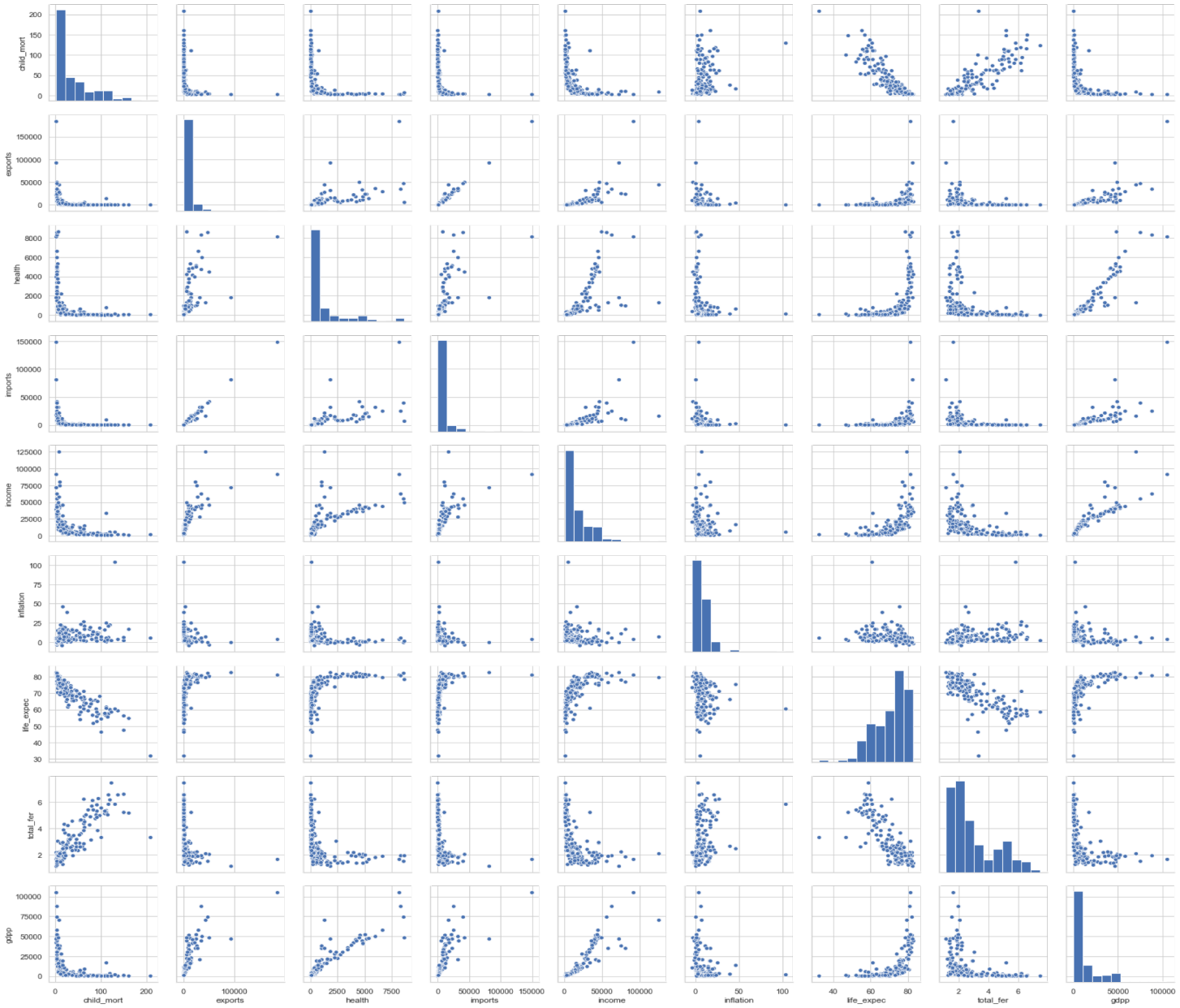
Data Science Process



Flow Chart

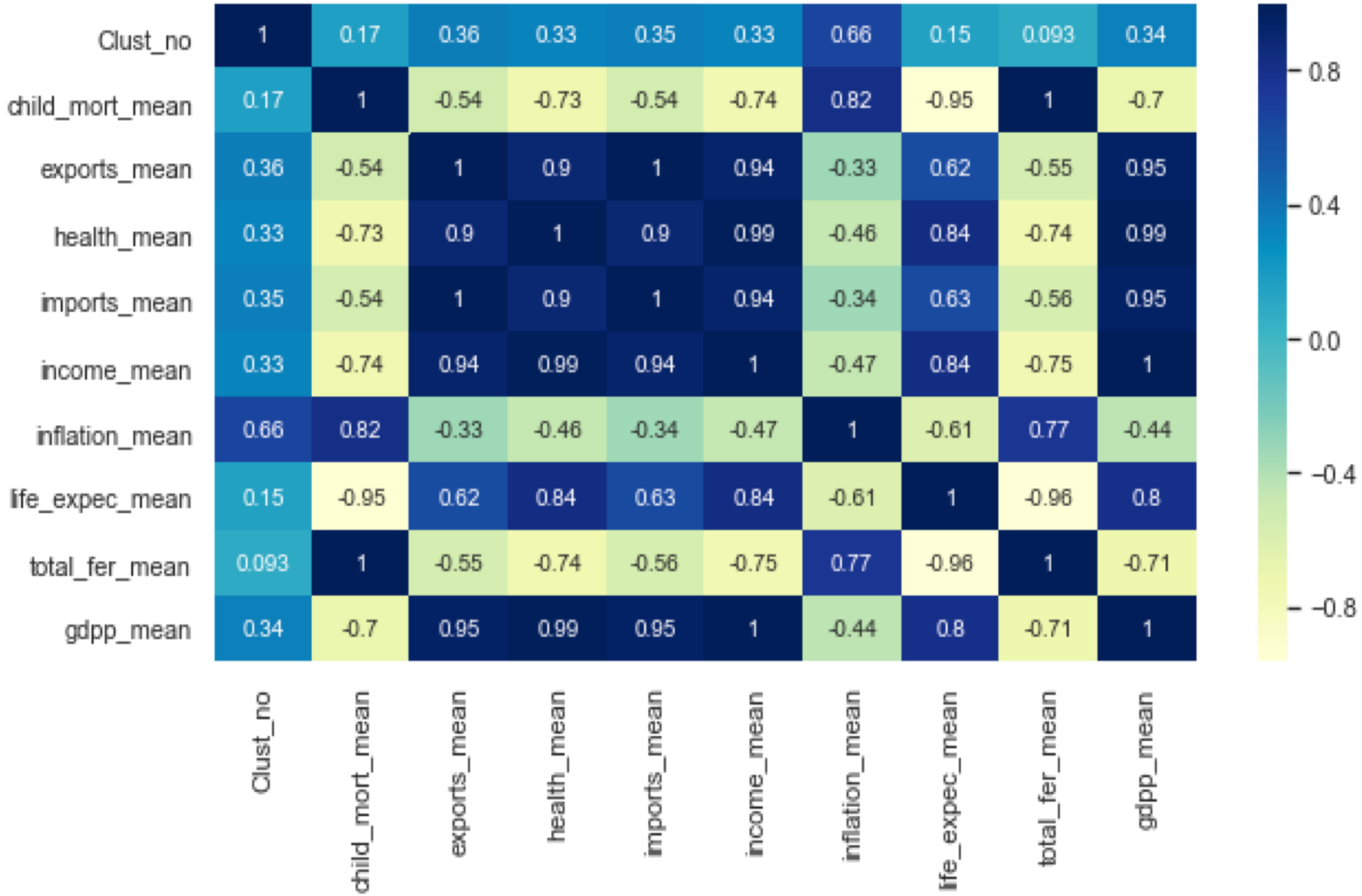
- Raw Data was Imported and stored Data frame for processing
- After Data Processing, the data was cleaned for Exploratory Data Analysis
- Principle component algorithm applied to generate reduced dimension data .
- Clustering technique both k-mean and Hierarchical has been applied to form cluster of data.
- Applying some Visualization technique to tap the problem statement.

EDA



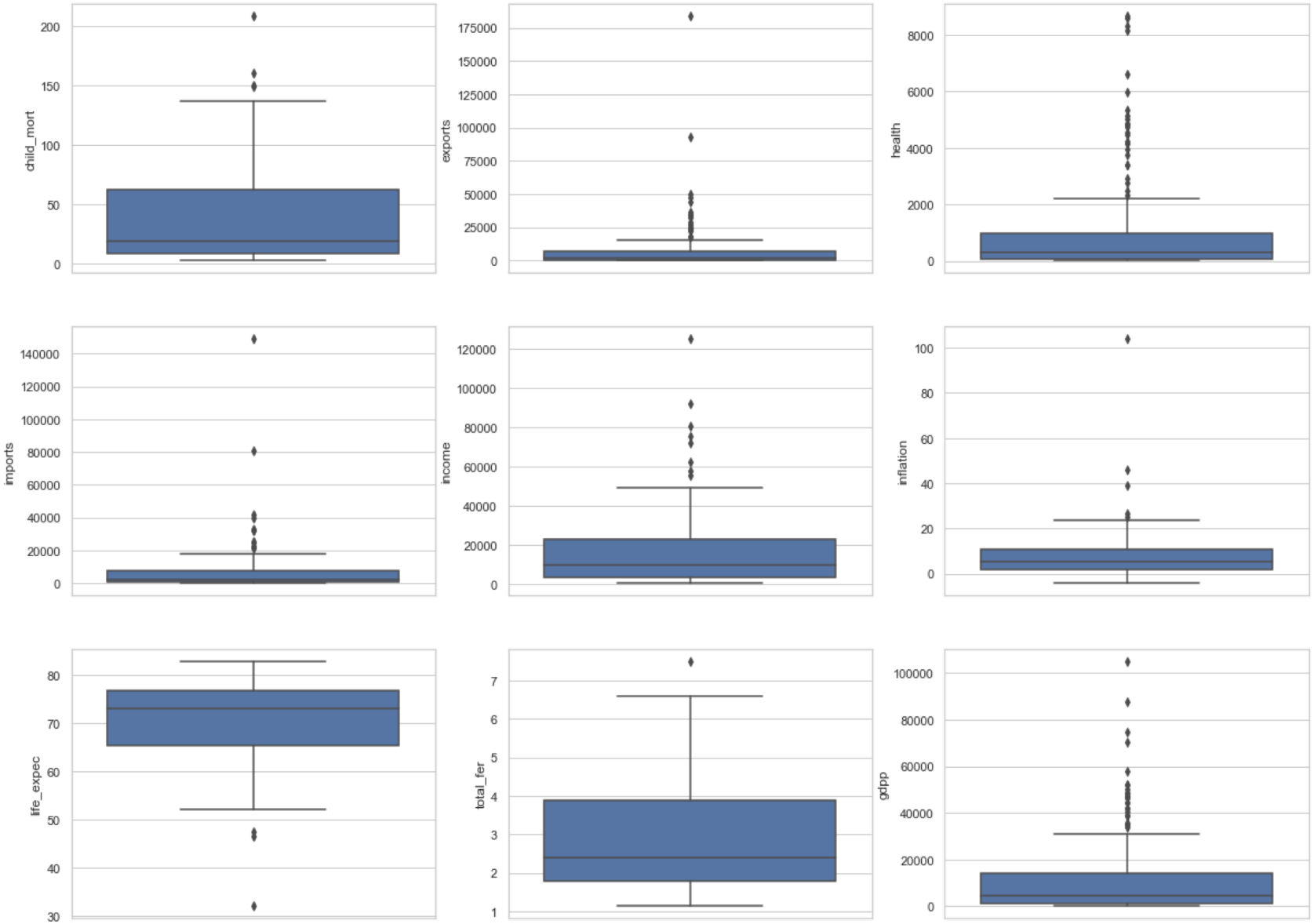
- Pair plot on data shows how data is distributed with each variable .
- Some of graph shows that Data is skewed f.

EDA …



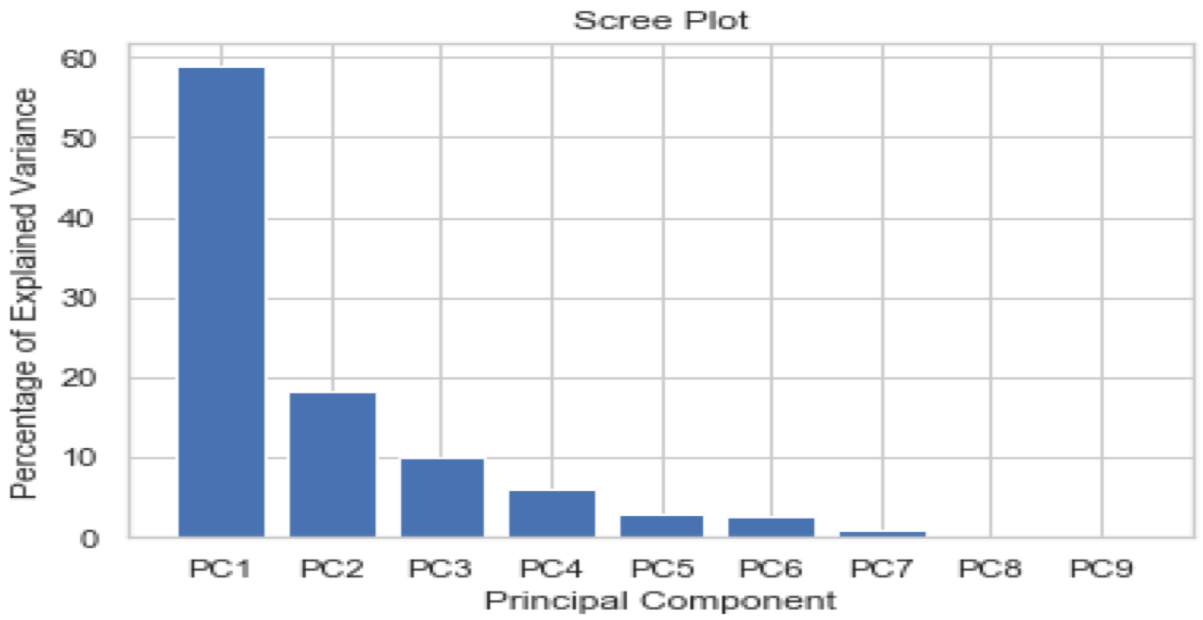
- Heat map on DF, shows that data is very highly correlated around quite a lot featured .
- Hence this become good candidate for PCA
- We can see there is high correlation between some variables, we will use PCA to solve this issue.

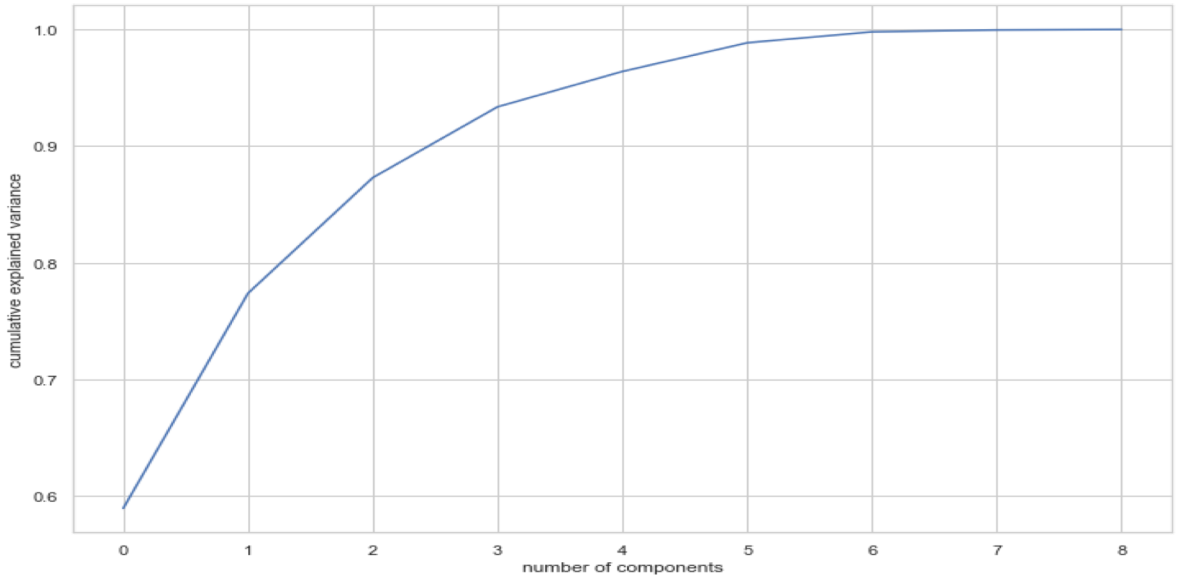
Outlier Treatment



- Any outlier is very sensitive for analysis the any data
- After plotting the box plot for each variable , few outlier to be seen
- As we can see there are a number of outliers in the data.
- Keeping in mind we need to identify backward countries based on socio economic and health factors.
- We will cap the outliers to values accordingly for analysis.

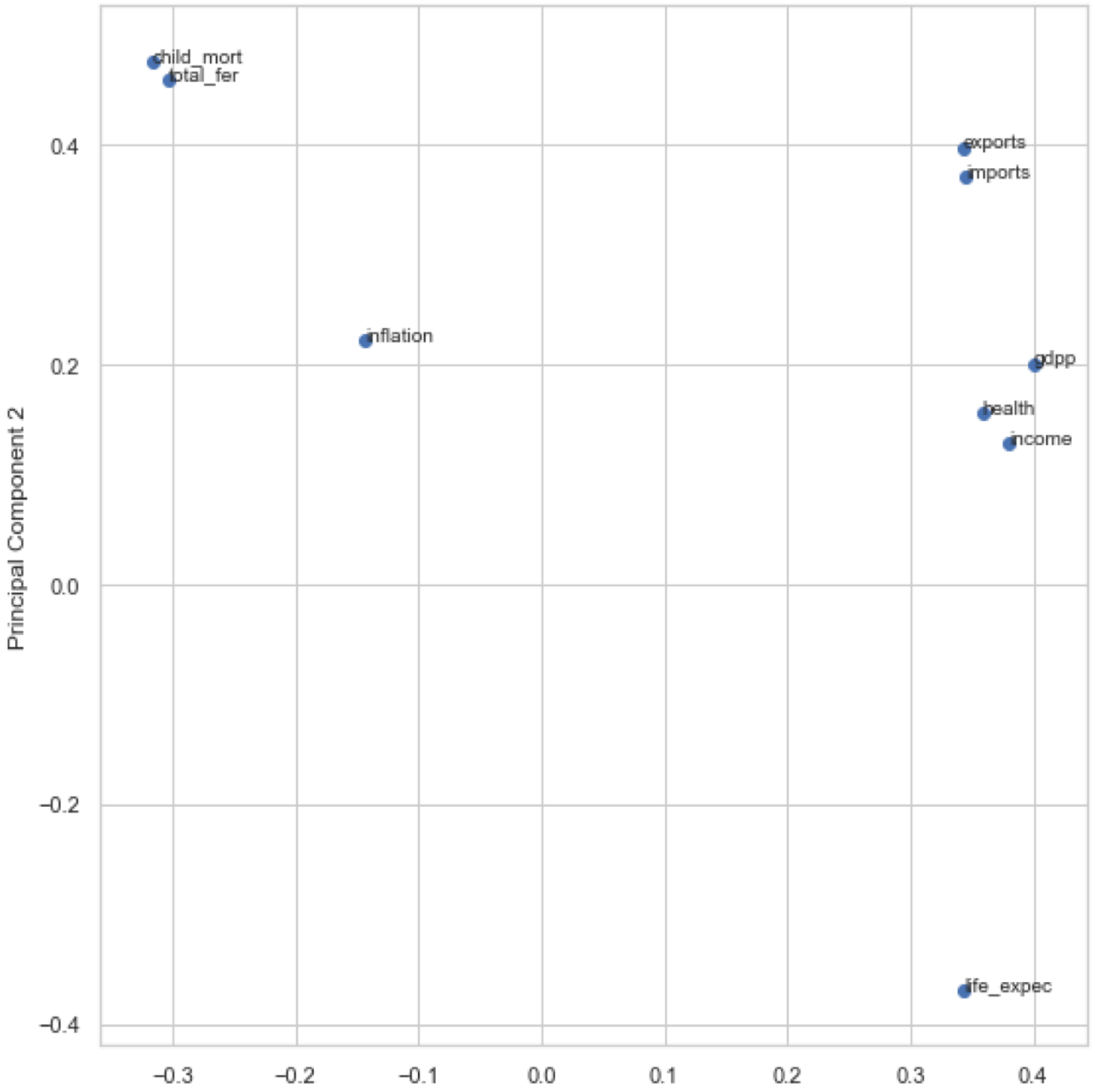
PCA



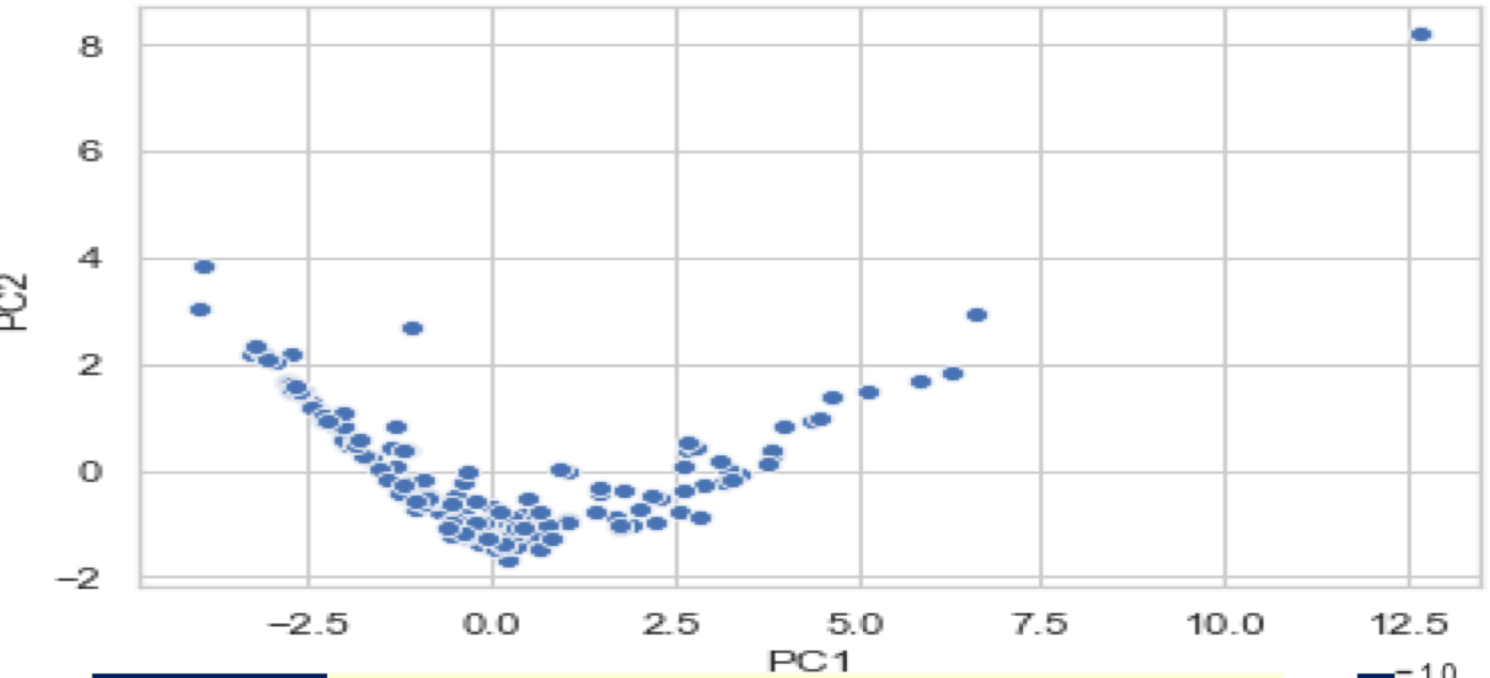


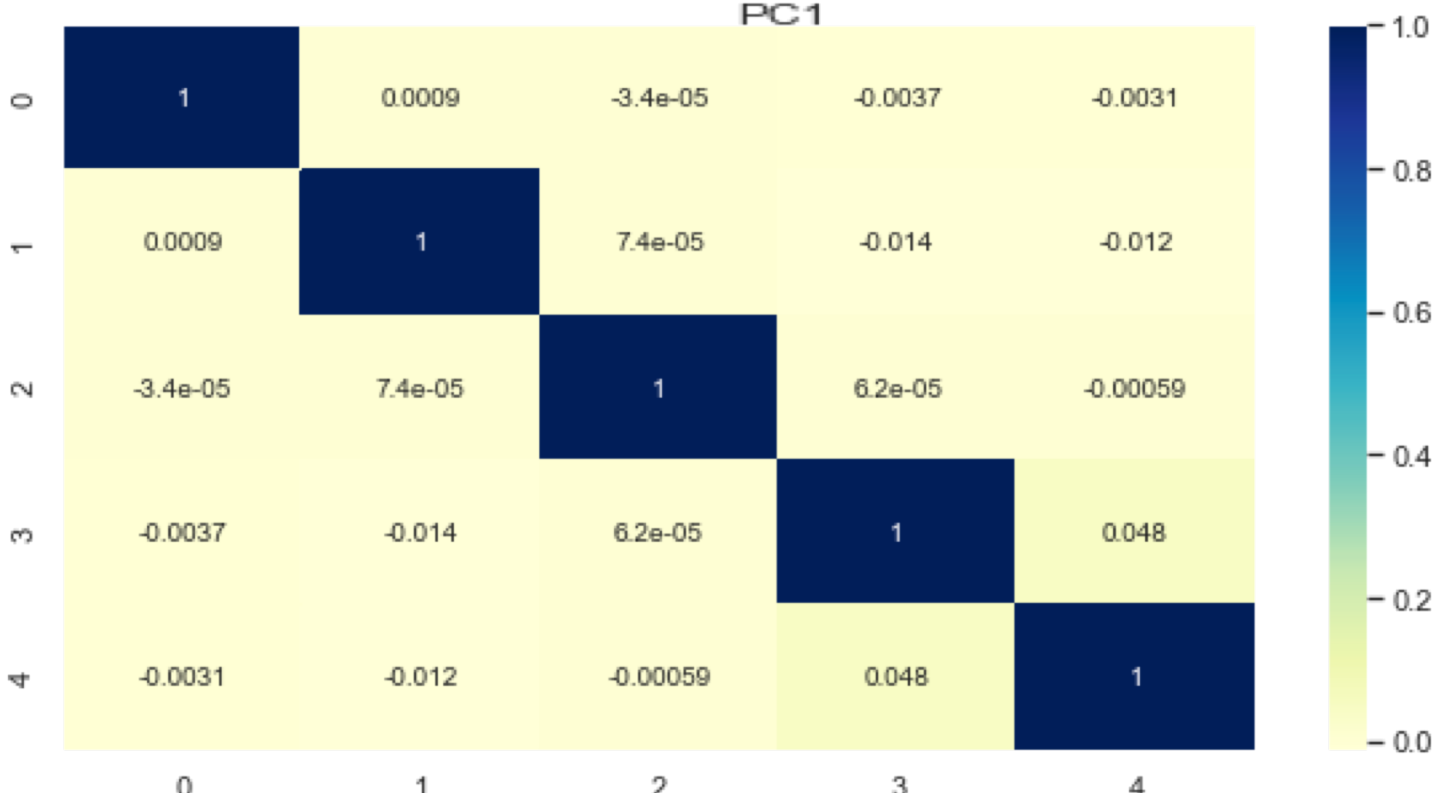
- After Data profiling , its suggested to apply PCA on cleaned Data frame
- Step 1: Scale the data with StandardScaler method on data frame to control the variability.
- Importing the PCA module to apply PCA method on scaled data .
- Next step is to select number of principle component which can be selected using Scree plot
- Scree plot suggest that Around 95% of the information is being explained by 5 components,
- Finally we use

PCA...



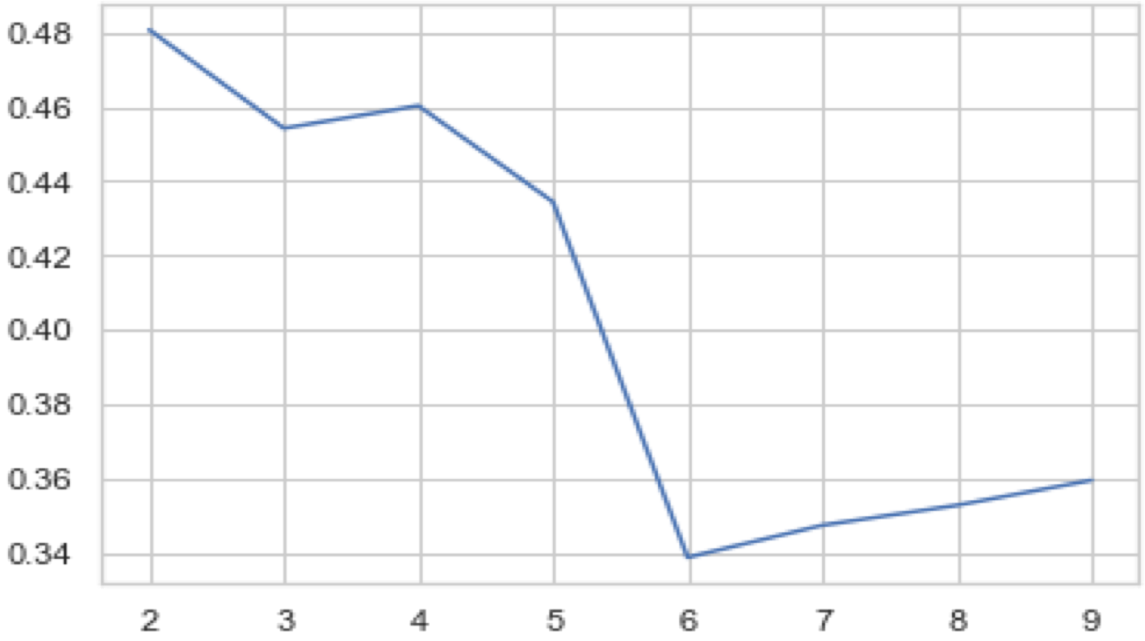
- Understanding how the original variables are loaded on the principal components
- Plotting them on scatter plot to visualise how these features are loaded

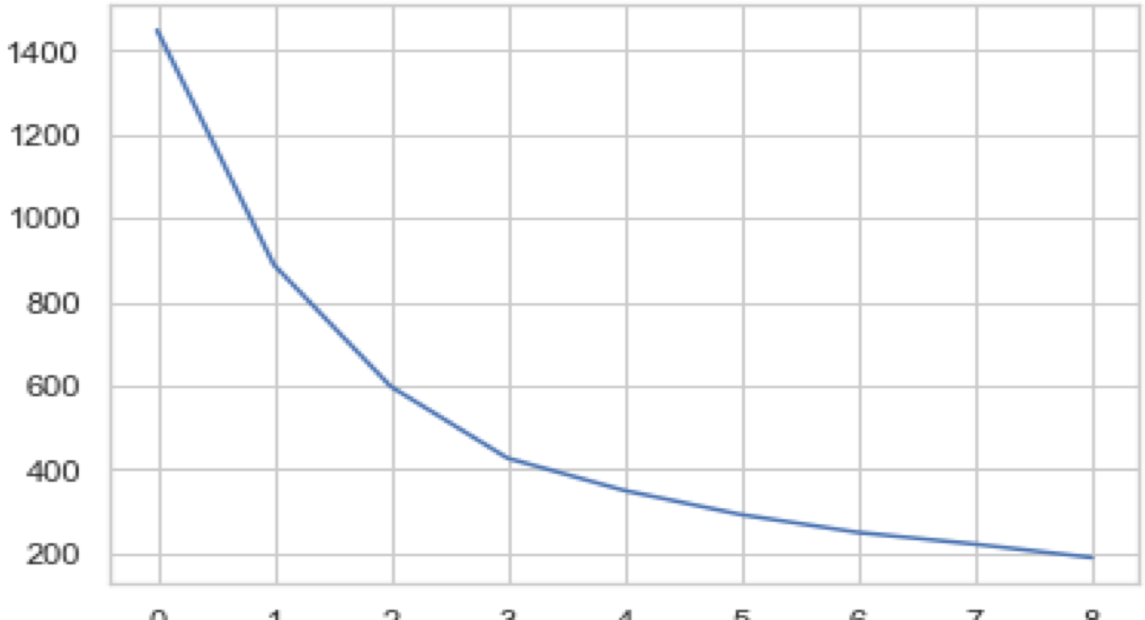




- Visualising the points on the PCs.
- one of the prime advantages of PCA is that you can visualise high dimensional data d
- Creating correlation matrix for the principal components to visualize any correlation
- plotting the correlation matrix and we see the correlation has been minimised
- Data is now ready to be used for clustering.

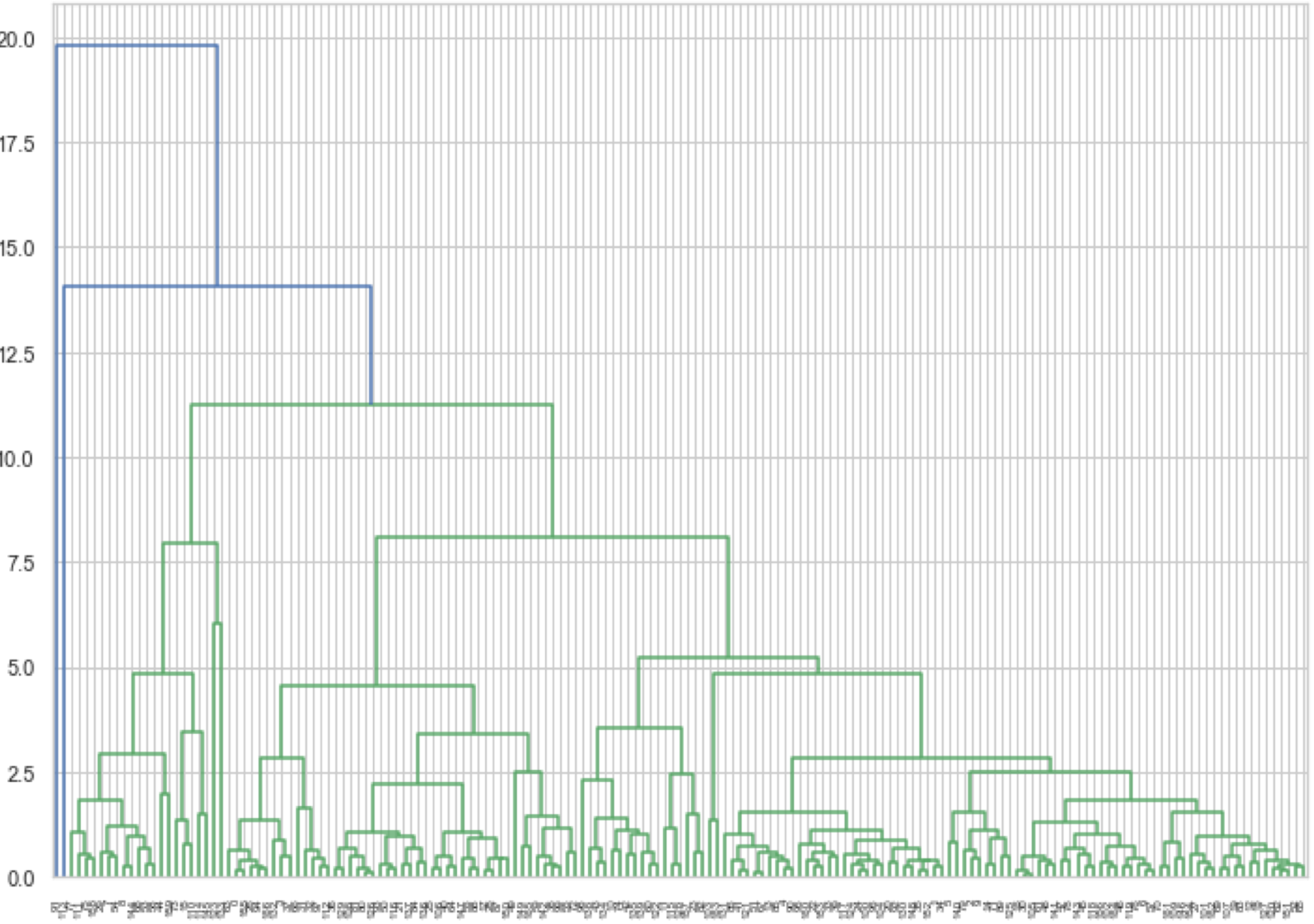
Applying clustering on Reduced dimensional data





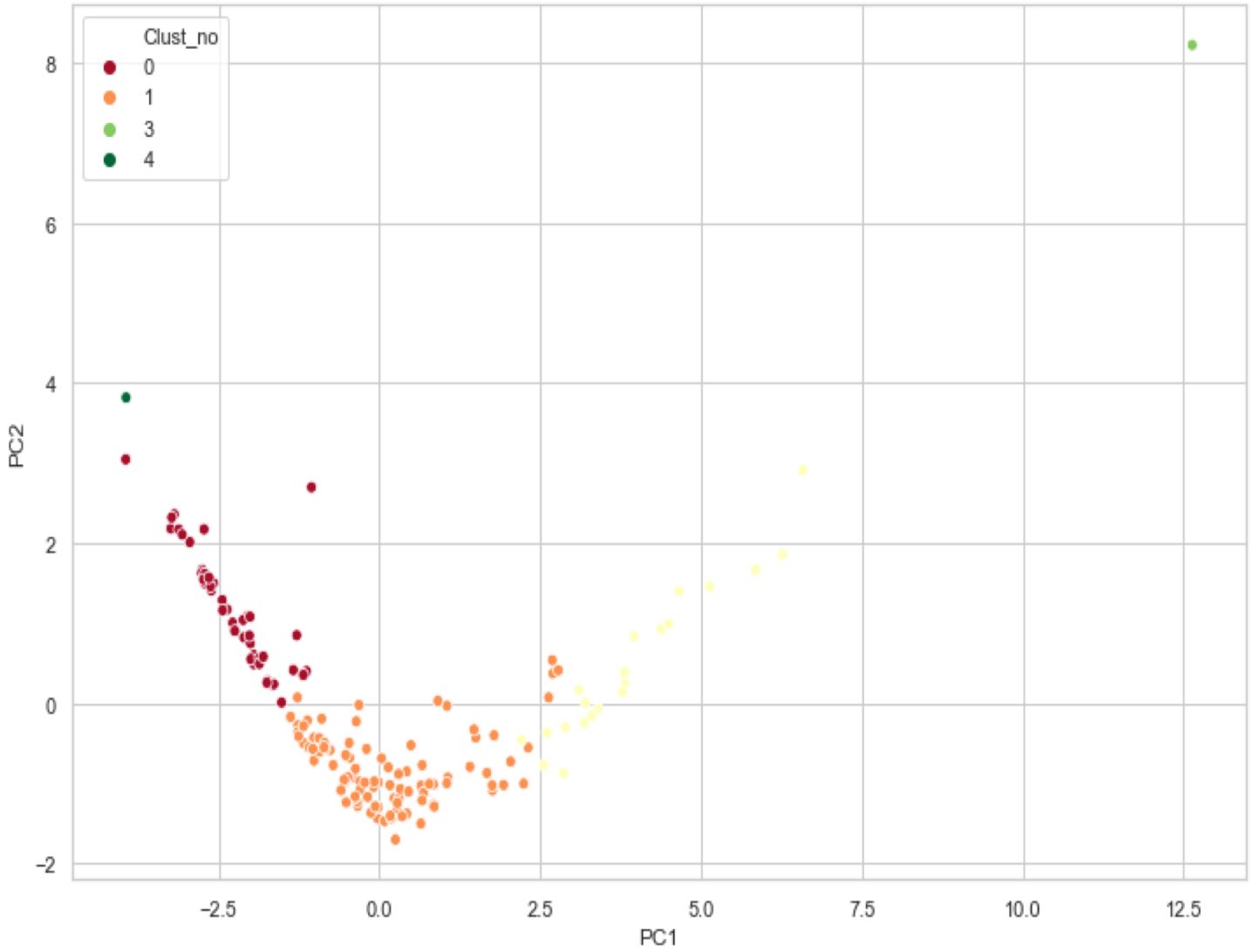
- K-mean clustering and Hierarchical clustering is applied to see which one best cluster the data
- As a pre-request steps We check The Hopkins statistic to observe cluster tendency
- Result shows that value is between {0.7, ..., 0.99}, which has high tendency to cluster.
- Main question is to decide number of cluster , which will be answered using various visualisation technique like silhouette score, elbow curve , dendrogram (see image respectively)
- We Plot all of them and it suggest the 4 or 5is the optimal no of cluster to be used.

Clustering …

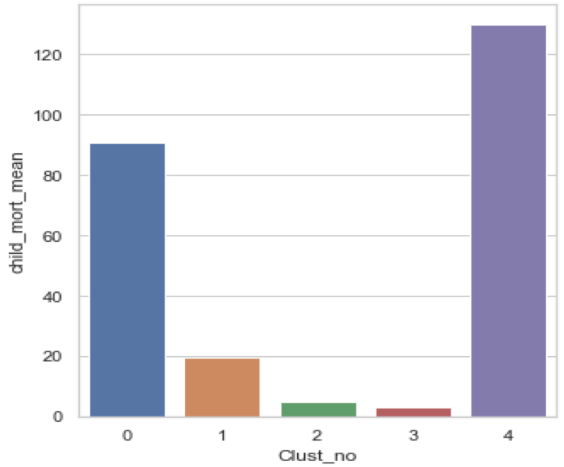


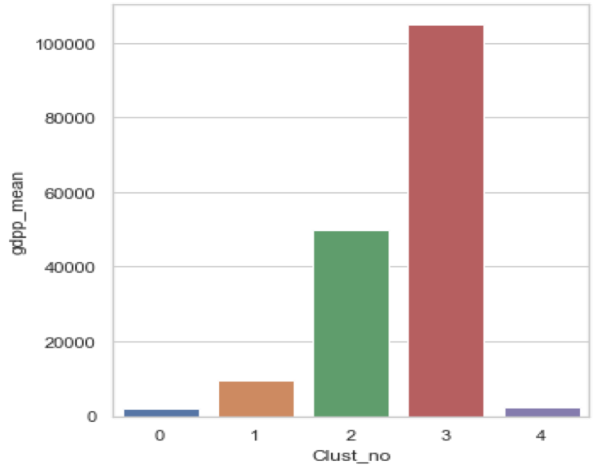
- Dendrogram depict as complete-link (or complete linkage) hierarchical clustering
- We cut it on 5 level to get clustering.
- Merge this data to original data to see the cluster number with each row of data.

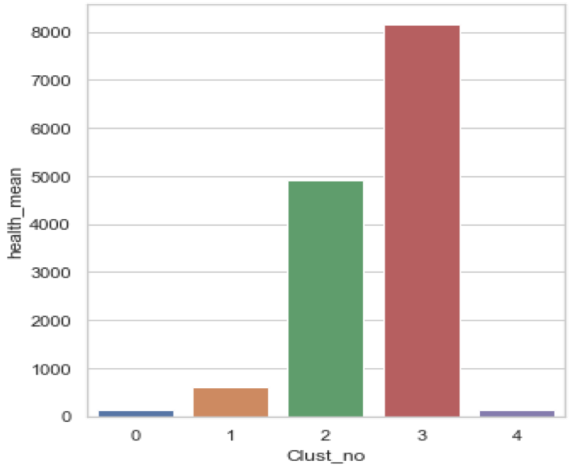
Key Insight from Clustered data

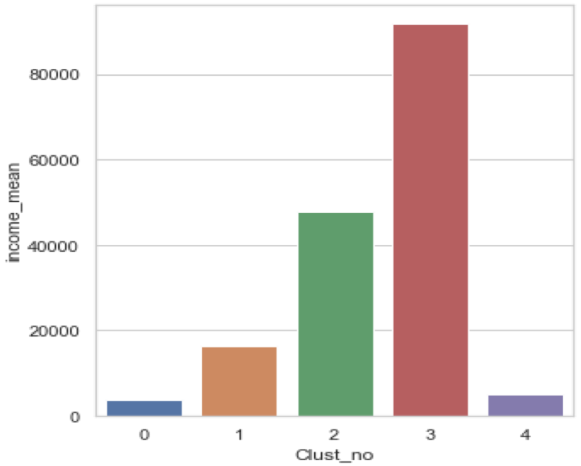


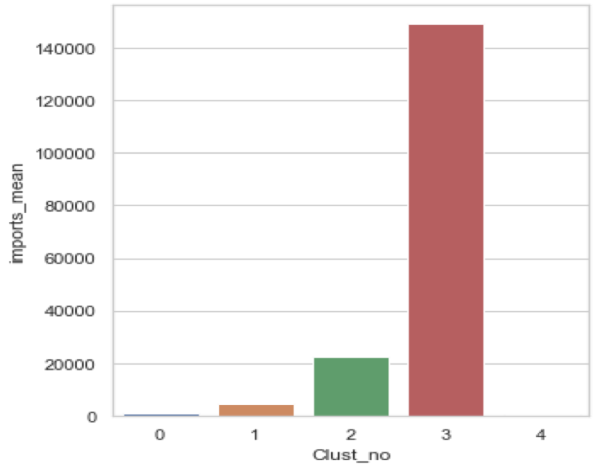
- As shown on scattered plot , each countries has been clustered from 0 to 4 (total 5 cluster)
- There large number of countries which has been part of cluster 0,1,2.
- Cluster 4 and 5 seems to have fewer observation
- Cluster data is now ready to be analysed by transforming as per business need.

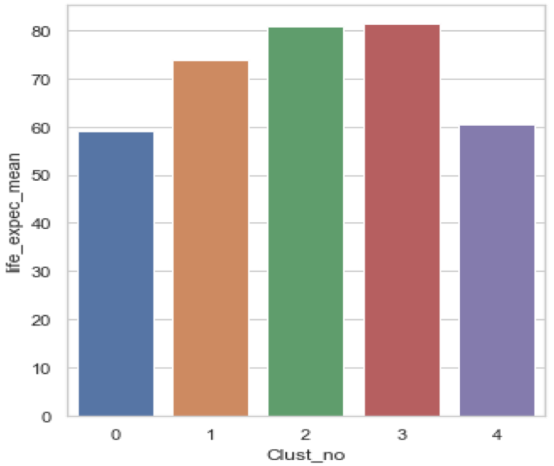


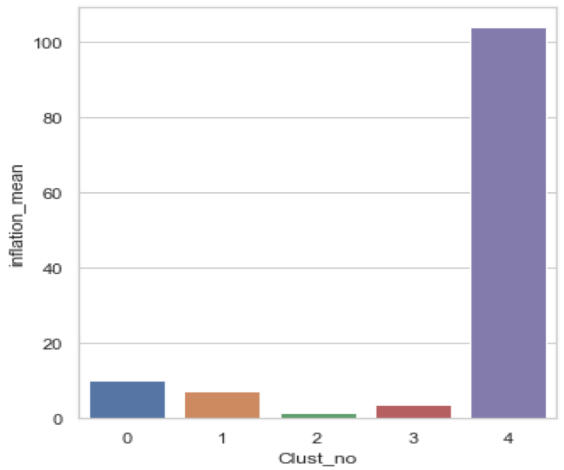


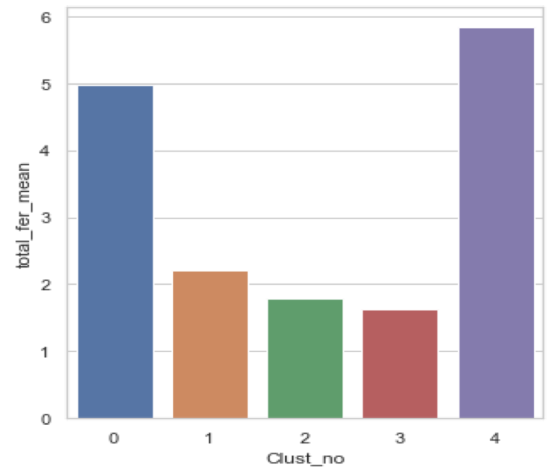


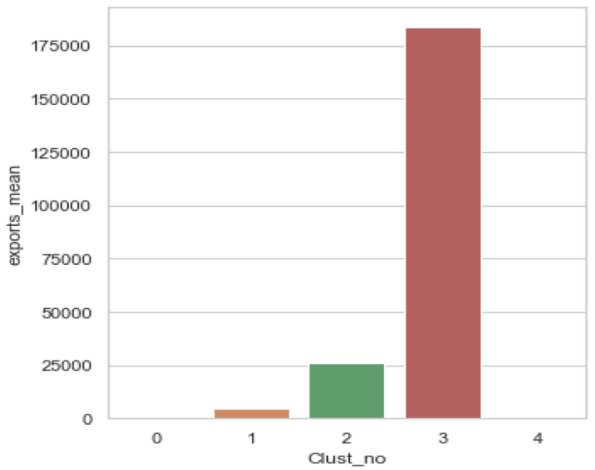












For all cluster data , the mean value of each variable has been calculated and its plotted againt respective cluster number
The bar plot is shown and observation is now drawn below.

The gdpp and income is very low for cluster 0 and 1 countries
Where as child mortality is very high in cluster 4 countries which is very scary

Conclusion

Cluster 0 contains countries that are in direct need of financial aid, since:

a) it has disproportionately high child mortality rate, total_fer & inflation;
b) and it has lowest gdpp, income & life_expectancy

- Cluster 1 countries seems to be developed countries with low gdpp , less health expenditure but avg export , income group
- List of countries under cluster 3 & 4 seem to have insufficient data to draw any insight , hence shouldn’t be considered.
- Nigeria is only country who falls under cluster 4 have very high child mortality rate which seems to be scary

Recommendation

Below countries are most backward countries and they are in direct need of financial aid

Afghanistan', 'Angola', 'Benin', 'Botswana', 'Burkina Faso', 'Burundi', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo, Dem. Rep.', 'Congo, Rep.', "Cote d'Ivoire", 'Equatorial Guinea', 'Eritrea', 'Gabon', 'Gambia', 'Ghana', 'Guinea', 'Guinea-Bissau', 'Haiti', 'Iraq', 'Kenya', 'Kiribati', 'Lao', 'Lesotho', 'Liberia', 'Madagascar', 'Malawi', 'Mali', 'Mauritania', 'Mozambique', 'Namibia', 'Niger', 'Pakistan', 'Rwanda', 'Senegal', 'Sierra Leone', 'Solomon Islands', 'South Africa', 'Sudan', 'Tanzania', 'Timor-Leste', 'Togo', 'Uganda', 'Yemen', 'Zambia’