upGrad

#LifeKoKaroLift

Clustering and PCA ->

Assignment

Prepared by:
Abhimanyu Dasarwar

upGrad

Clustering & PCA: Finding Top 5 Countries in dire need of aid from HELP International NGO

Part I : Data browsing and cleaning

Part II : Outliers Treatment

Part III: Principal Component Analysis

Part IV: Clustering

Part V: Clustering Analysis

Assignment – Problem Statement

upGrad

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.
- The significant issues is which countries should be provided aid?

Approach followed:

- First step is to analyze the given data and perform data cleaning activities.
- Missing values identification.
- Treatment of Outliers
- Principal Component Analysis
- Clustering
- Clustering Analysis

Data browsing and cleaning, Outliers Treatment

upGrad

Data Browsing and routine checkup:

1. Importing dataset to python notebook
2. Browsing data using DF.head()
3. Finding out Number of rows and columns using DF.shape command
4. Applying DF.describe() to identify numerical variables and St. Deviation and percentile values

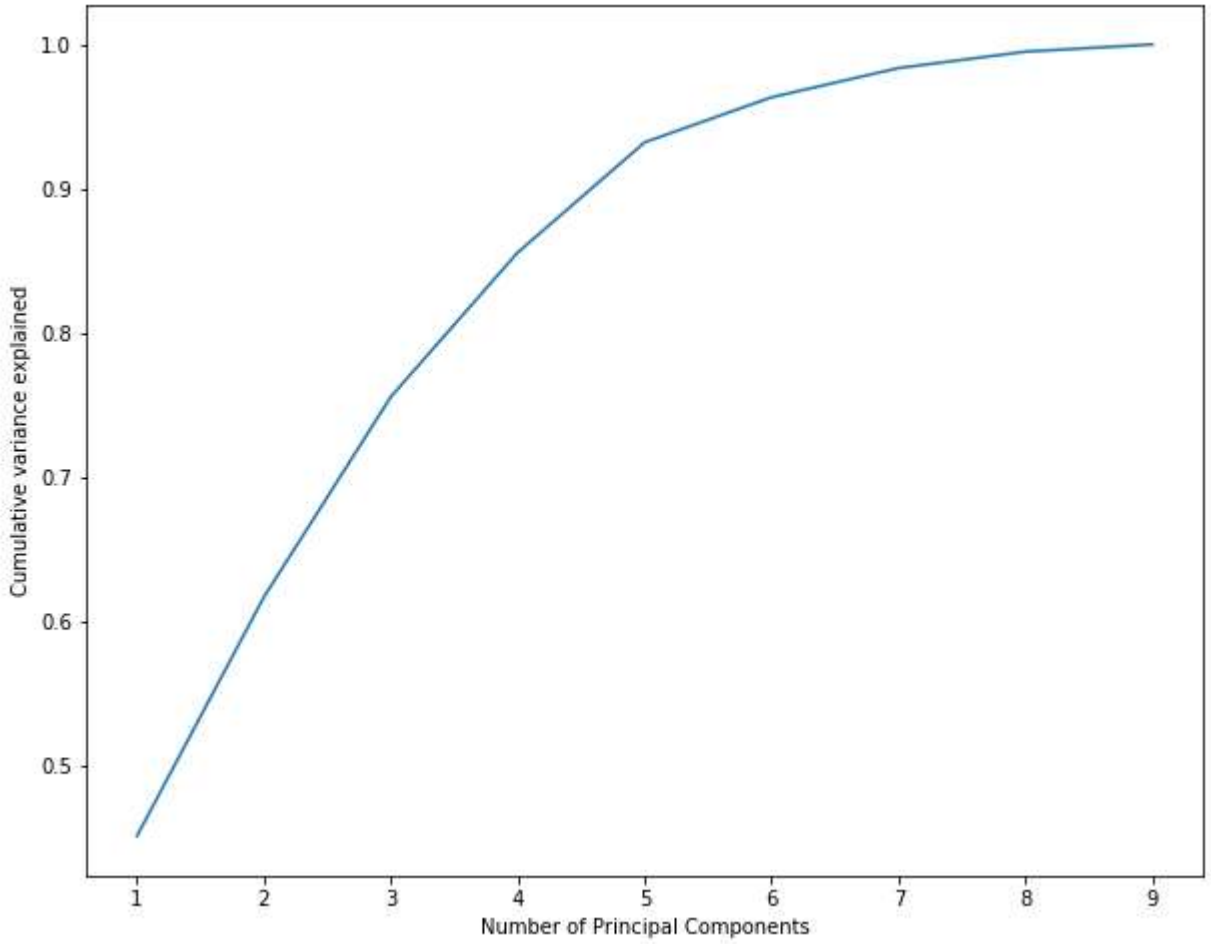
Identification and Treatment of Outliers:

1. Outliers are identified using Boxplot
2. “GDPP” and “Income” features had number of outliers
3. From GDPP, outliers above 80% are removed
4. From Income, outliers above 90% are removed.
5. Scaled the data.
6. Scaled data is used for further analysis.

Principal Component Analysis

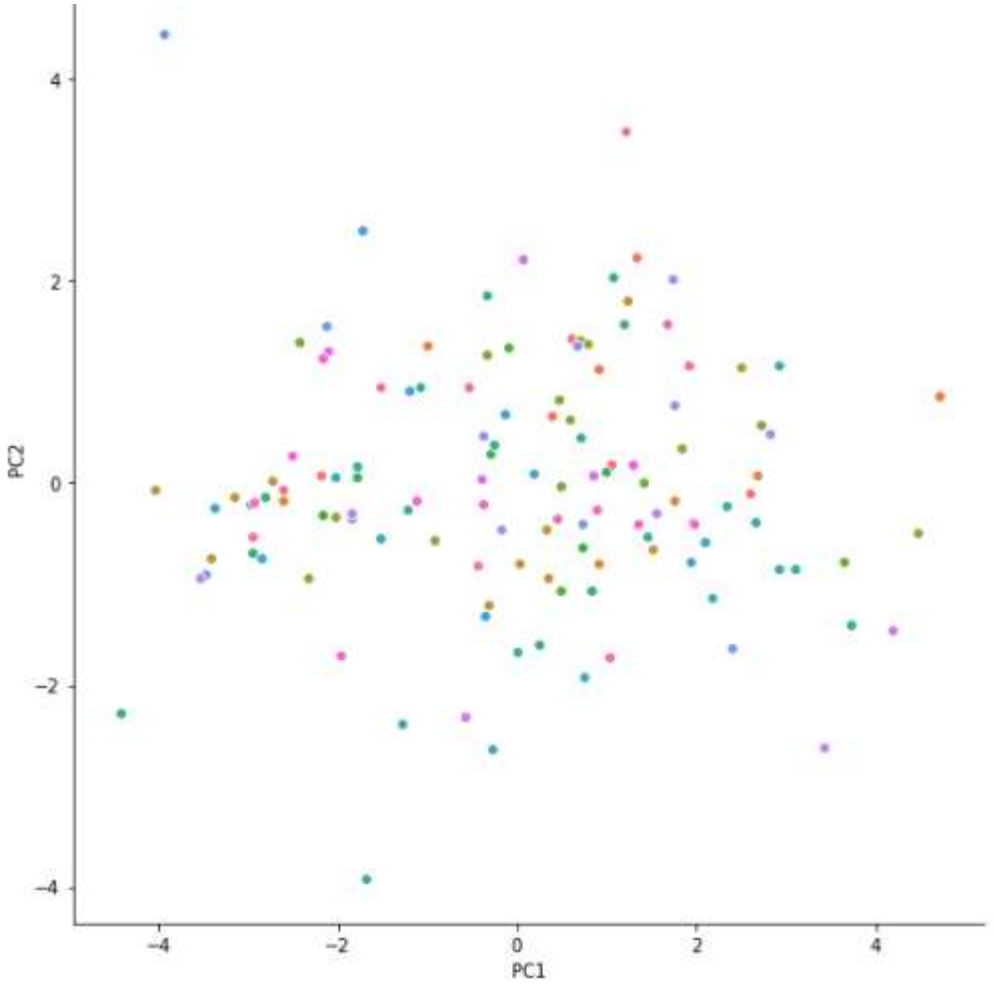
upGrad

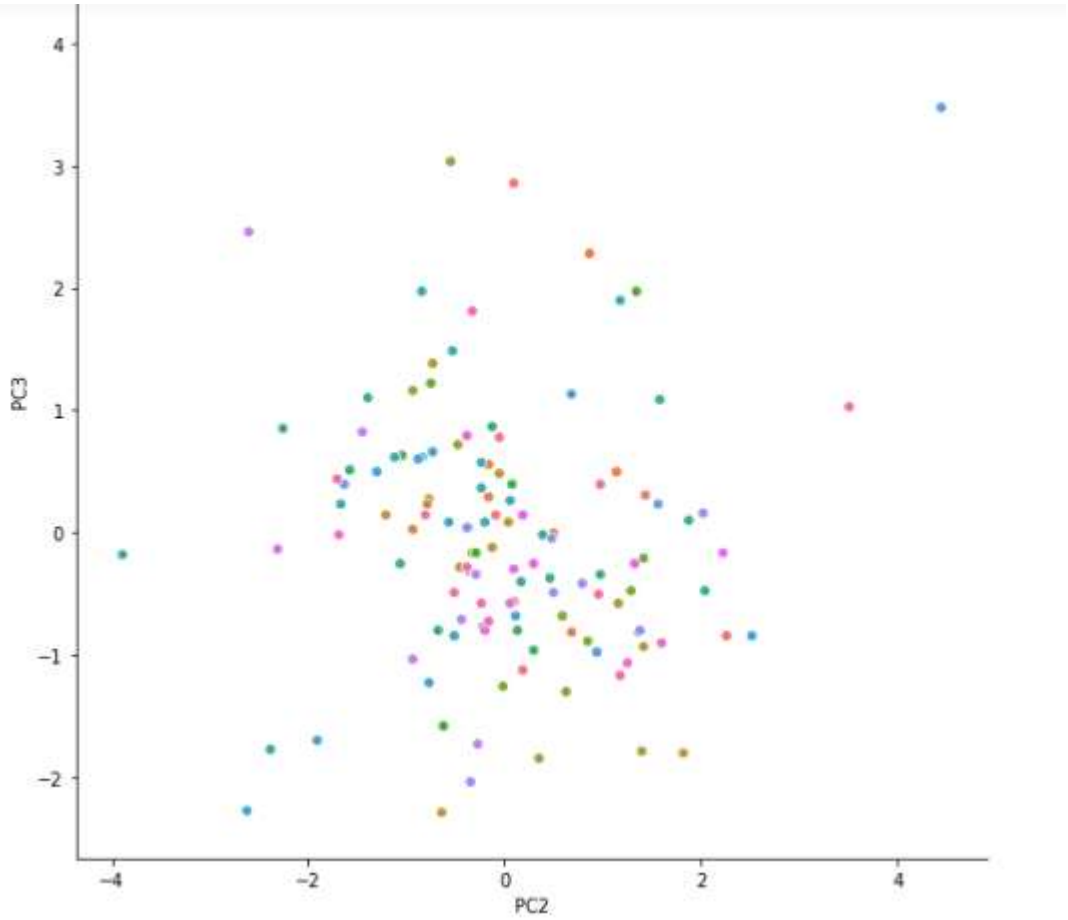
Performed the PCA on Scaled data and obtained the Variance Ration of each feature from the dataset.
Obtained the below scree plot which helps in deciding the number of principal components.



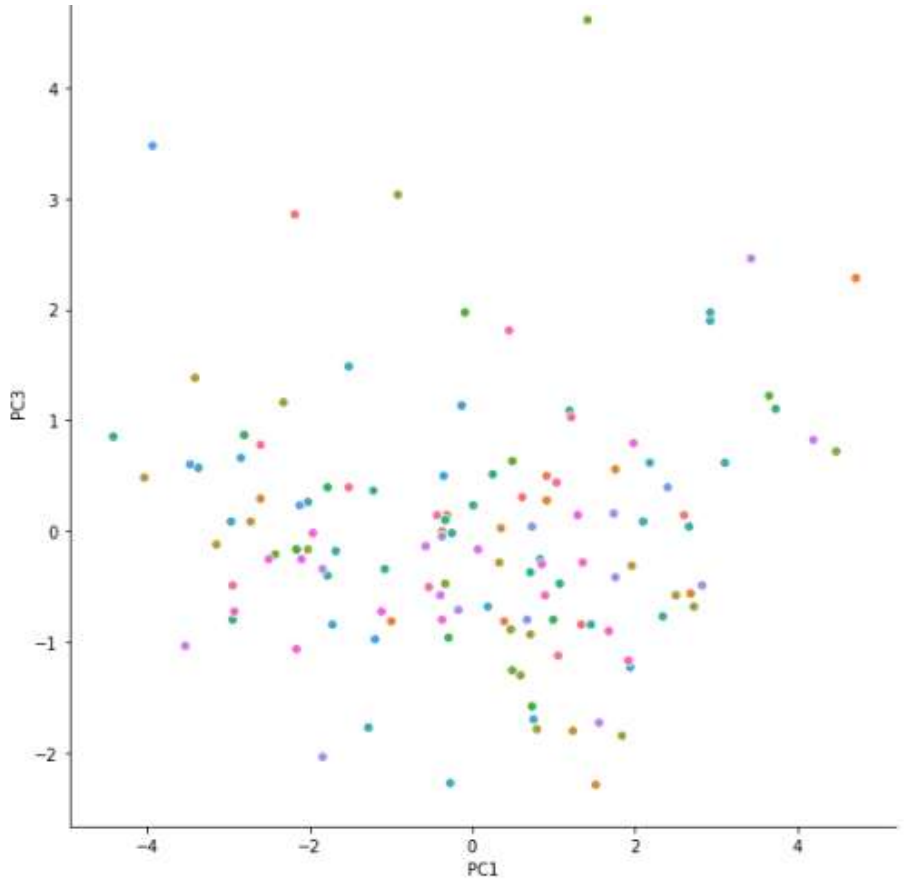
From the plot, it's visible that 5 Components explain upto 95% of the variance in the data.
Hence K=5

- Re-run the PCA with k=5 and combine the country names with the data obtained from PCA.
- Below are scatter plots which confirm that we can form different clusters of this data.





- The “Hopkins Score” of the data varies between 0.70 to 0.79.
- This is considered as a good Hopkins score for clustering.
- Hence we can proceed with forming the clusters of the obtained data.

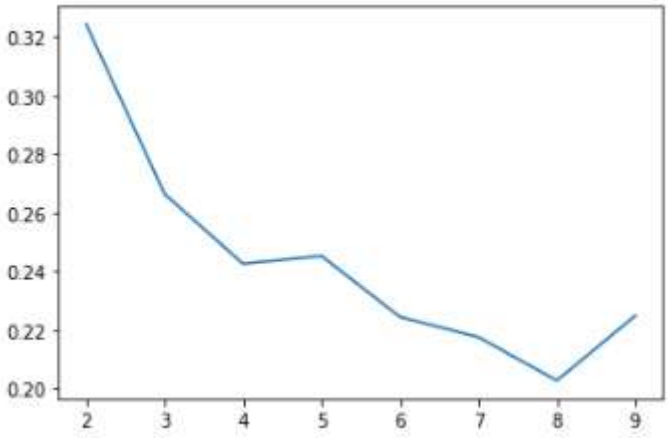


Clustering

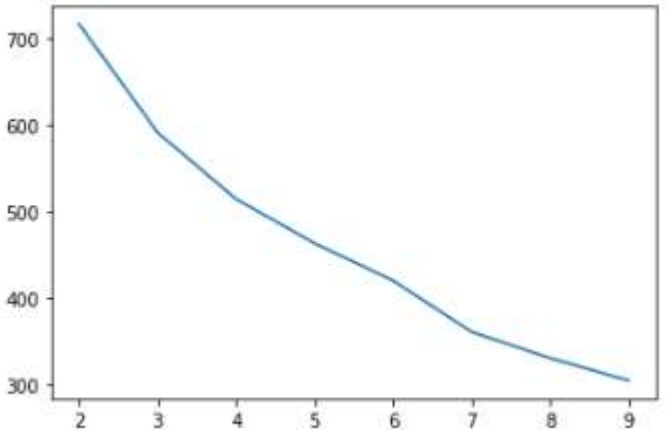
upGrad

KMeans Clustering: Performed the KMeans Clustering on the data.
To determine the number of clusters, we have used “Silhouette Score Plot” and “Elbow Curve”

Silhouette Score Plot

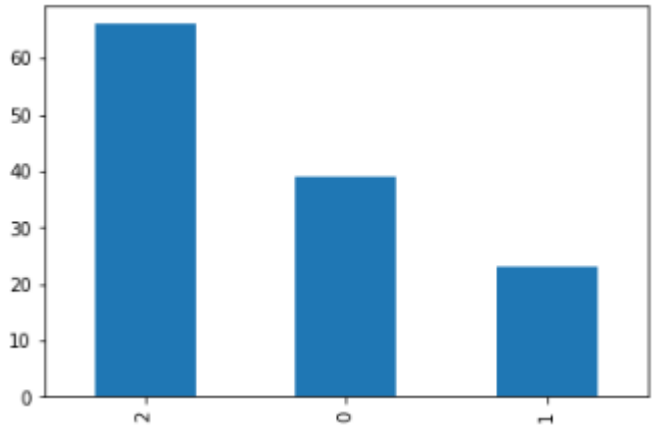


Elbow Curve

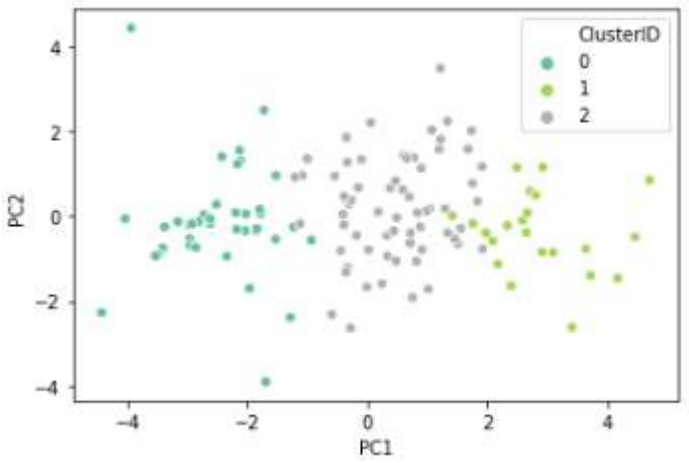


- In Silhouette score plot, average max value is at 3
- In Elbow Curve, we have a elbow at 3.
- Taking into consideration above observations, the number of clusters are taken as 3

Clustering



- The bar graph shows the number of record from each cluster.



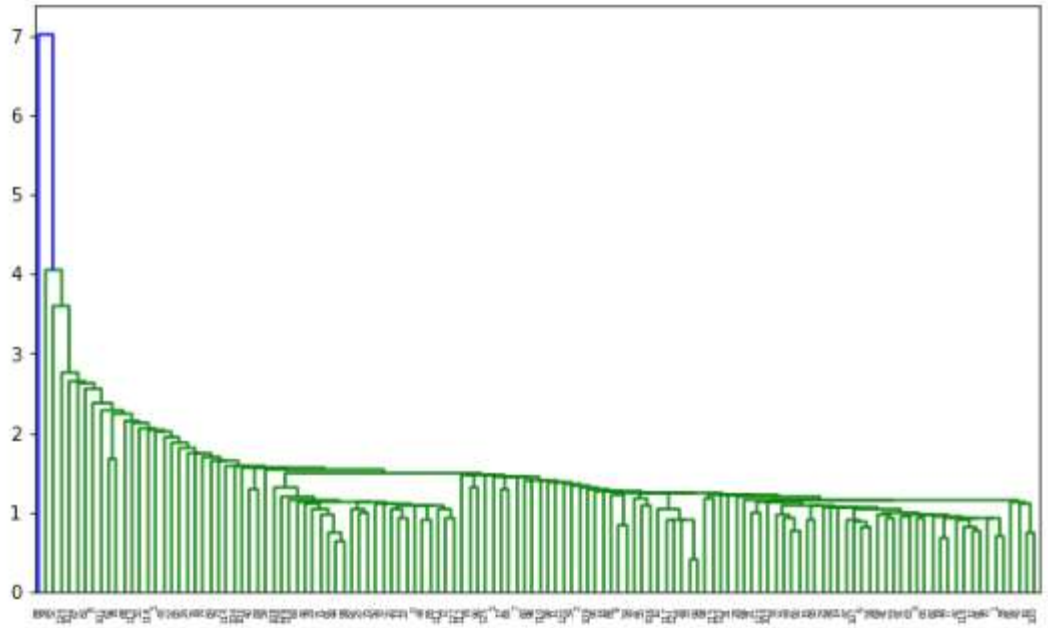
- The scatter graph shows the distribution of clusters.
- Here we can easily get clear idea of cluster distribution.

Clustering

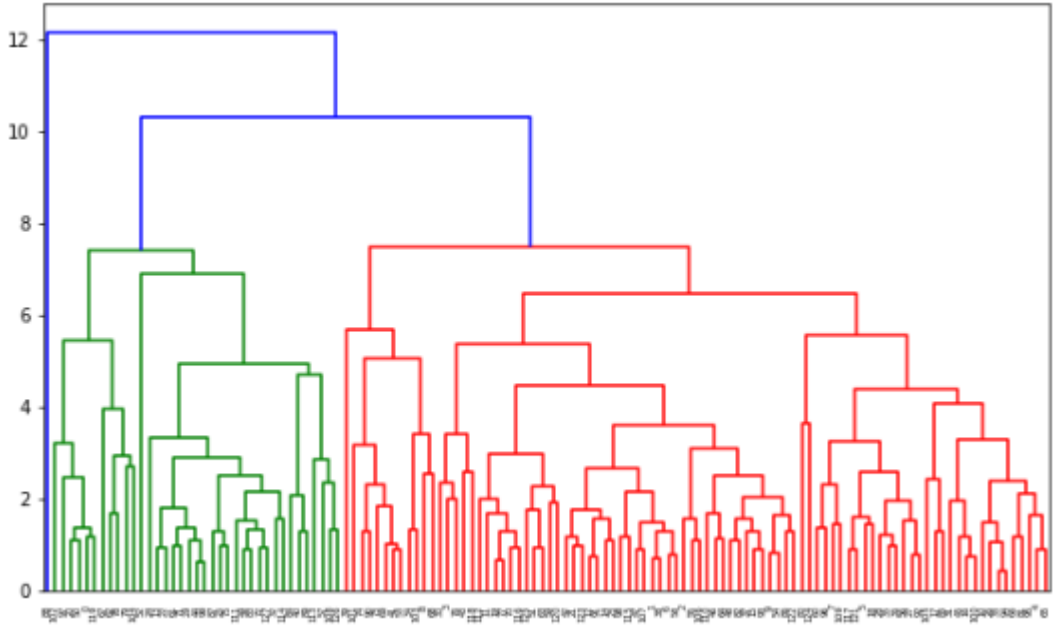
upGrad

Hierarchical Clustering: Formed the Clusters using Hierarchical Clustering also.
Used Single and Complete linkage to find the optimal number of clusters.

Single Linkage



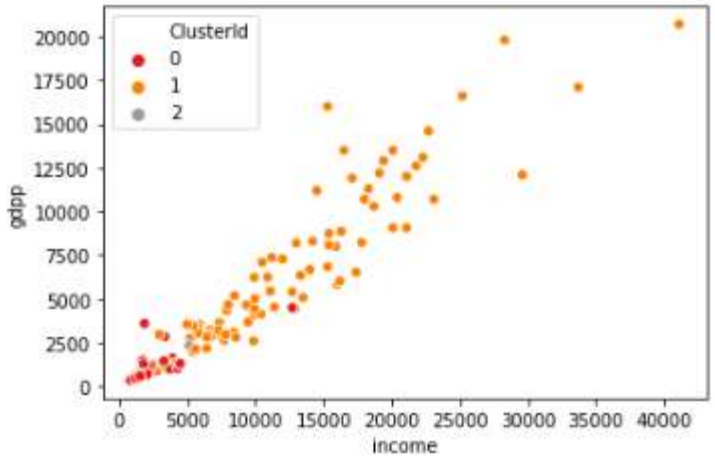
Complete Linkage



From Complete linkage, we can see that number of clusters can be 2,3,5.
Hence, opting for n=3

Clustering

upGrad

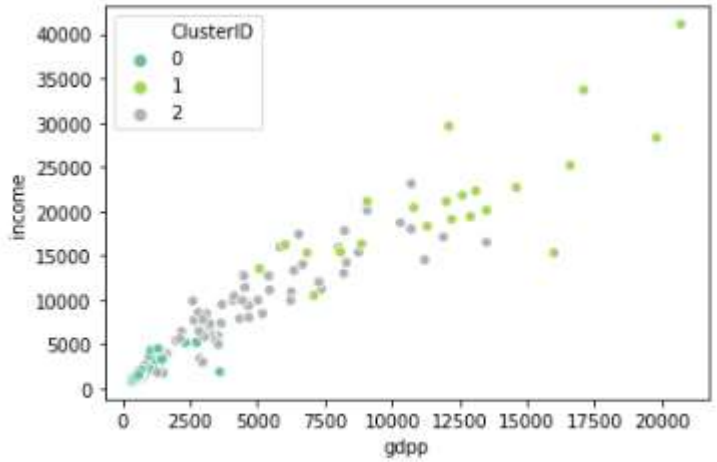


- The scatter lot shows the distribution of clusters, using n=3
- Unlike KMeans, we are not able to visualize 3 clusters as only one record is in 3rd cluster.
- Hence opting for Kmeans for further evaluations.

Clustering Analysis

upGrad

- Using Kmeans, we formed three clusters of the data.
- After observing the data obtained and clusters formed, we can conclude that
  - Cluster 0 is Under Developed Countries.
    - ✓ Low GDPP
    - ✓ Low Income
    - ✓ High Child Mortality Rate
  - Cluster 1 us Developed Countries.
    - ✓ High GDPP
    - ✓ High Income
    - ✓ Low Child Mortality Rate
  - Cluster 2 is Developing Countries.

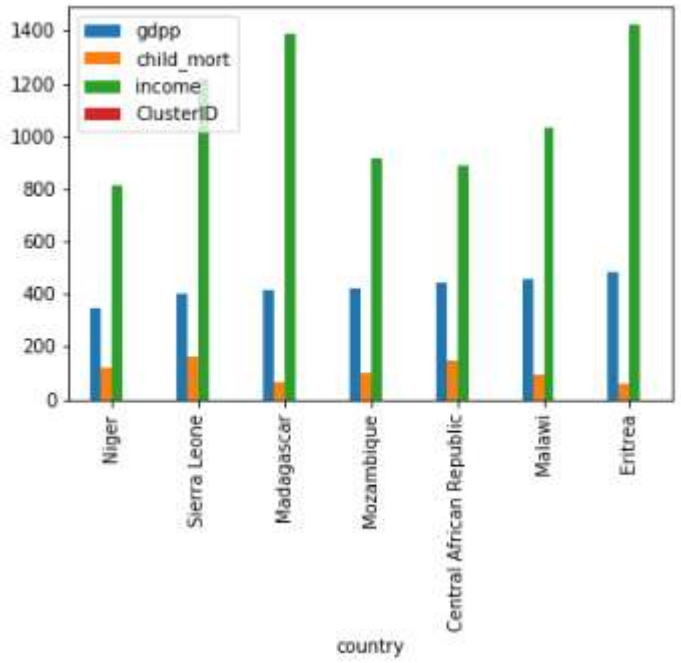


- Using filters, created new dataset for Cluster 0 for more evaluations

Cluster 0 Analysis

upGrad

- Cluster 0 is made up of Under Developed Countries.
- To find out the countries those are in need of help, sorted the data
- Below are the countries those have very low GDPP, low income and high child mortality rate.
  - Niger
  - Sierra Leone
  - Madagascar
  - Mozambique
  - Central African Republic



Conclusion and Key Takeaways

upGrad

From the PCA and Clustering of the data provided, narrowed down the top 5 countries those are in dire need of aid from HEALTH

- Using PCA, found that 95% of variance in data is explained by 5 Principal components.
- Clustered the data in 3 different clusters and observed that :
  - Cluster 0 is for Under Developed Countries.
  - Cluster 1 is made up of Developed Countries as their GDPP and Income is high, while child mortality rate is very low.
  - Cluster 2 is for Developing Countries.
- Top 5 Countries having low GDPP, low Income, High Child Mortality :
  - Niger
  - Sierra Leone
  - Madagascar
  - Mozambique
  - Central African Republic