Clustering and PCA Assignment
Help International Study

Anshul Srivastava

Abstract

Objective

- To categorise the countries using some socio-economic and health factors that determine the overall development of the country.
- To suggest the countries which the CEO needs to focus on the most.
- Help International may choose to utilize this Knowledge while making this decision are mostly related to choosing the countries that are in the direst need of aid.

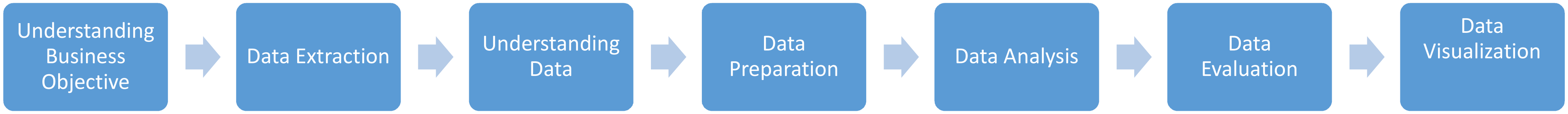
Analysis

- EDA
- -Standardisation
- -PCA
- -Removing Multi Collinearity
- Outlier Treatment
- -Clustering
- -Barplots for Decision taking

Source

- UpGrad's Dataset

Problem Solving Methodology



Understanding Business Objective → Data Extraction → Understanding Data → Data Preparation → Data Analysis → Data Evaluation → Data Visualization

Problem Solving Approach

Data Cleaning Steps

- All NaN Values were taken care
- All columns with same column values were removed
- Removing all the columns with >90% null values
- Basic conversion to matching Data type

Outlier Analysis

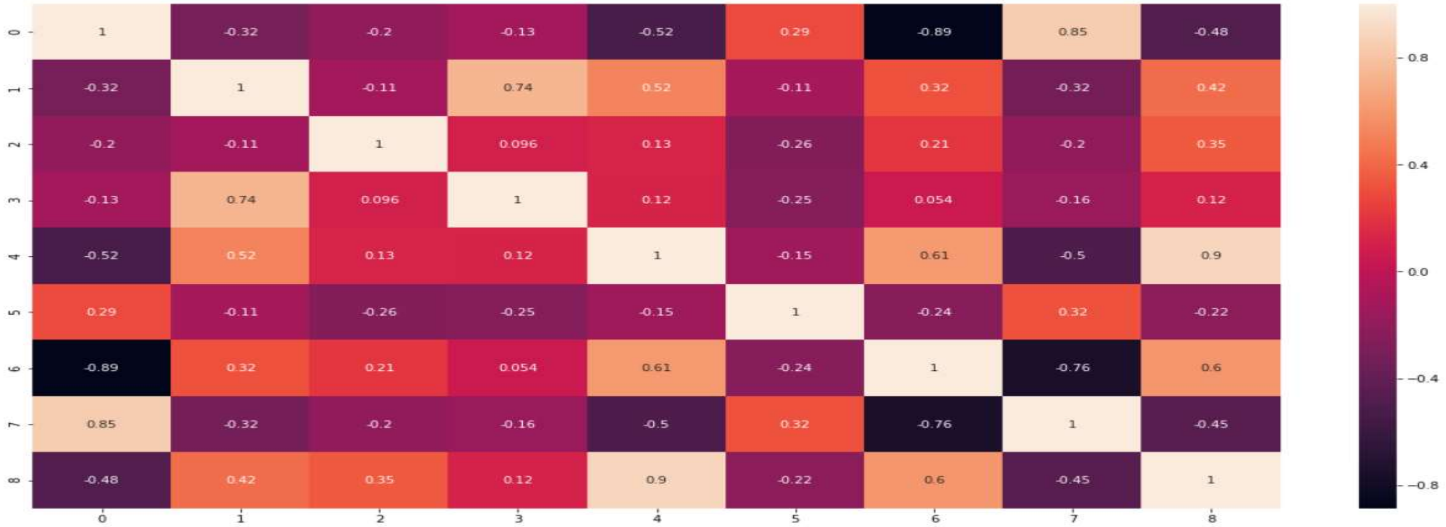
- Analysing relevant columns by plotting them on Boxplot charts

Clustering Analysis

- Analysing 4 or more clusters to make inferences of the dataset

Observations

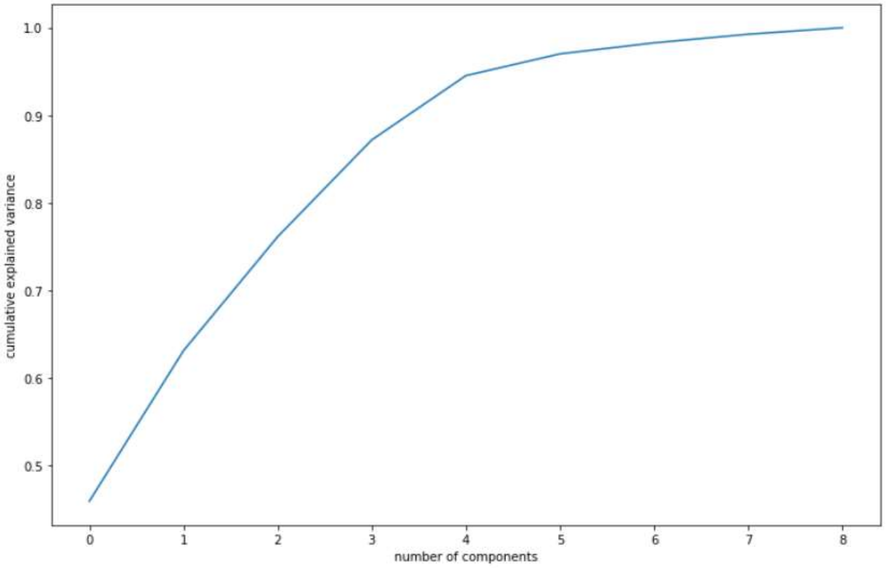
- Outliers majorily in gdpp, income , exports and child mortality.
- Rather than doing Outlier treatment earlier , do it on PCA dataset when multicollinearity is removed.

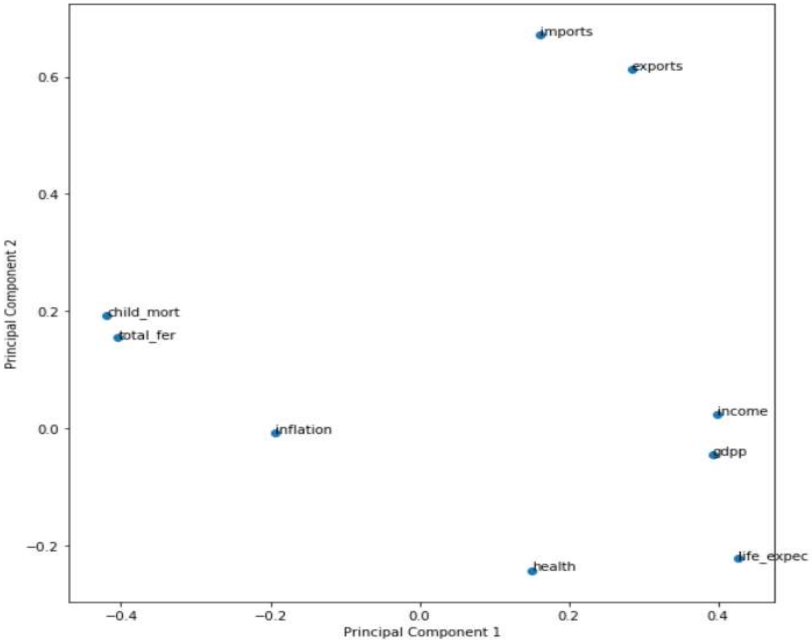


	0	1	2	3	4	5	6	7	8
0	1	-0.32	-0.2	-0.13	-0.52	0.29	-0.89	0.85	-0.48
1	-0.32	1	-0.11	0.74	0.52	-0.11	0.32	-0.32	0.42
2	-0.2	-0.11	1	0.096	0.13	-0.26	0.21	-0.2	0.35
3	-0.13	0.74	0.096	1	0.12	-0.25	0.054	-0.16	0.12
4	-0.52	0.52	0.13	0.12	1	-0.15	0.61	-0.5	0.9
5	0.29	-0.11	-0.26	-0.25	-0.15	1	-0.24	0.32	-0.22
6	-0.89	0.32	0.21	0.054	0.61	-0.24	1	-0.76	0.6
7	0.85	-0.32	-0.2	-0.16	-0.5	0.32	-0.76	1	-0.45
8	-0.48	0.42	0.35	0.12	0.9	-0.22	0.6	-0.45	1

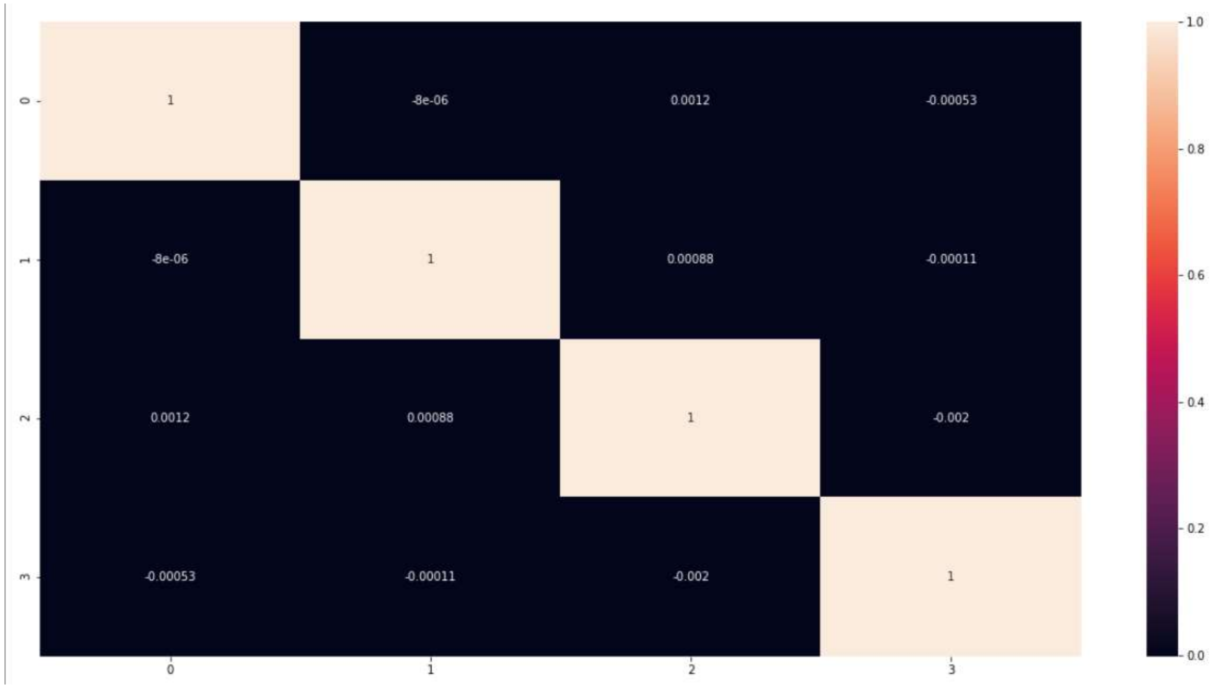
PCA

- PCA is done in order to remove multiple variables which are multi-collinearity.
- Various principal features are created just to explain variance in the dataset.
- Also we can see through the Scree Plot that 4 features are enough to explain 95% of the variance in the dataset rather than 9 variables in the original dataset.





PCA Dataset



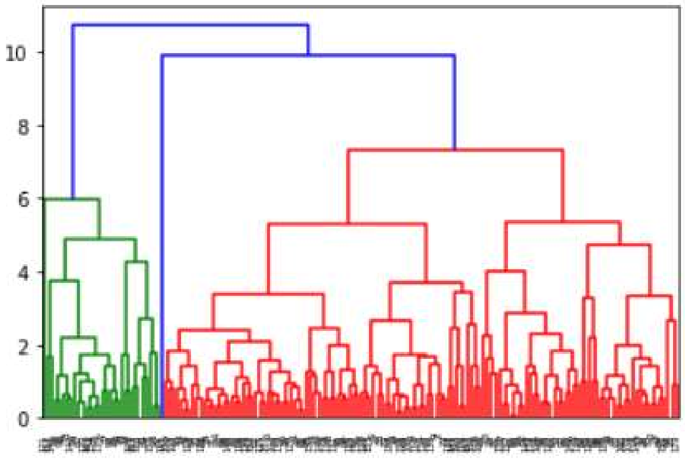
Indeed - there is no correlation between any two components! Good job, PCA!¶

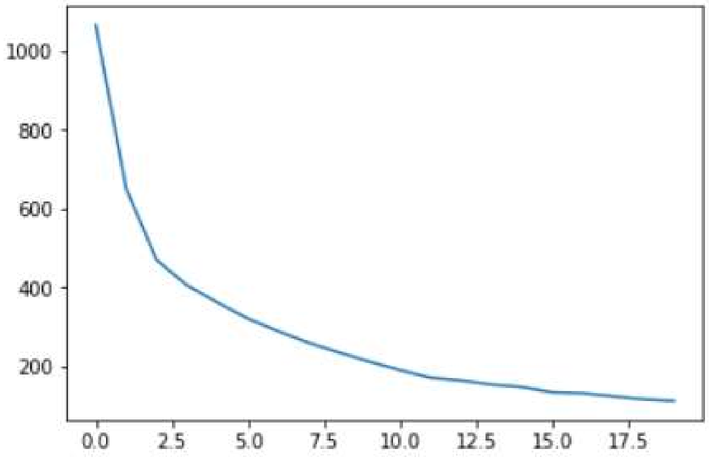
We effectively have removed multicollinearity from our situation, and our models will be much more stable

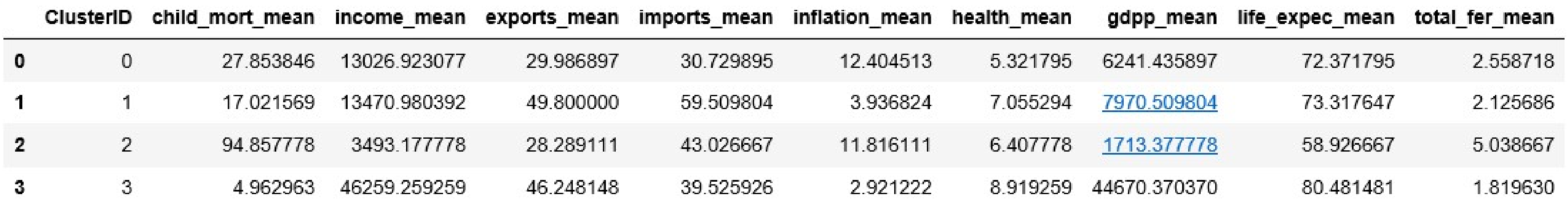
Clustering

- Hopkins Test was done on the PCA Dataset to find out whether clusters can be made out of the data or not.
- Combination of both Hierarchical Clustering and K-means Clustering is done to find out optimal number of clusters.
- Hierarchical clustering gives us a picture as to how many clusters are to be made and then the same number of clusters are made using k-means.
- Clusters are again verified using Elbow Curve and Sum of Squared Errors.

Aggregate Dataset after Clustering

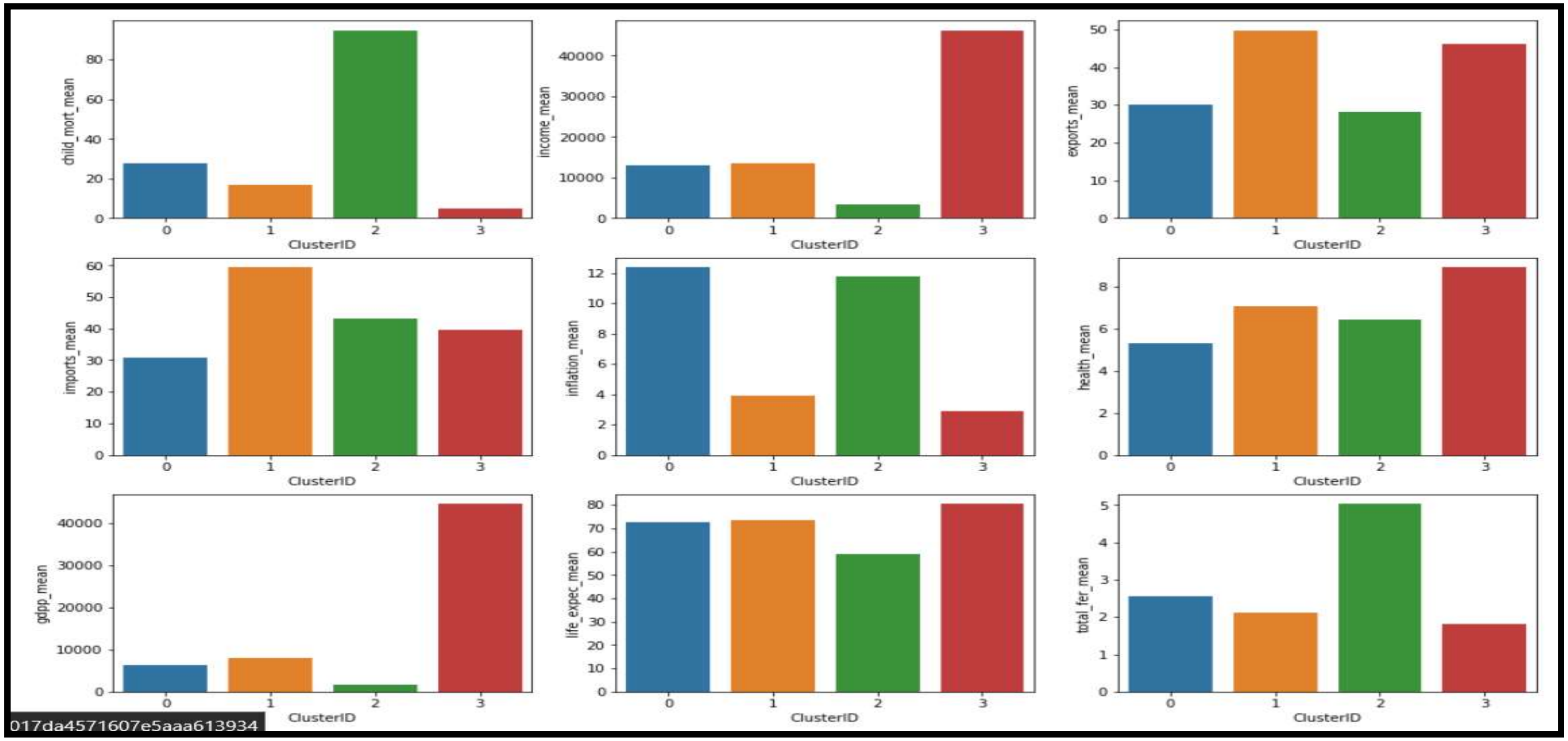






	ClusterID	child_mort_mean	income_mean	exports_mean	imports_mean	inflation_mean	health_mean	gdpp_mean	life_expec_mean	total_fer_mean
0	0	27.853846	13026.923077	29.986897	30.729895	12.404513	5.321795	6241.435897	72.371795	2.558718
1	1	17.021569	13470.980392	49.800000	59.509804	3.936824	7.055294	7970.509804	73.317647	2.125686
2	2	94.857778	3493.177778	28.289111	43.026667	11.816111	6.407778	1713.377778	58.926667	5.038667
3	3	4.962963	46259.259259	46.248148	39.525926	2.921222	8.919259	44670.370370	80.481481	1.819630

Visualizing Clusters and its Properties



Conclusion

- Cluster 1 (0) – Socio-economic factors like child mortality, health and fertility are quite average over here but the inflation is too high that means it is experiencing rapid growth.
- Cluster 2 (1) – Exports and imports are very high in this cluster whereas the economic growth of this country is very low.
- Cluster 3 (2) – Child morality is very high in this cluster and also inflation is too high whereas we can also see that average number of children per woman is quite large compared to other clusters.
- Cluster 4 (3) – Income of people in this cluster is high and they are good with external relations and exports power is also good whereas growth has declined since people countries have already reached saturation stage but the health conditions are proper and gdp is also high.

Recommendation

- Cluster 1 (0) – Since inflation is too high over the countries in this cluster should focus on health facilities and better medical care.
- Cluster 2 (1) – The annual growth rate is too low over here this may be because of the fertility rate or child morality being low which should be focused more in this cluster countries.
- Cluster 3 (2) – Child mortality is quite high as well as fertility rate also shoots which indirectly questions about health care facilities being poor for new-born or children and even old people.
- Cluster 4 (3) – A maximum percentage of health facilities account for gdpp which mean this cluster countries have good health facilities but the growth is stagnant and also imports have to be given extra push for better externa relations.

Thank You!