HELP INTERNATIONAL (CLUSTERING AND PCA)

R.MADHAVAN

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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

- After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

- This is where i come in as a data analyst. My job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then I need to suggest the countries which the CEO needs to focus on the most

DATA AND LIBRARIES USED

DATA USED:

‘Country_data.csv’ which contains the following info

Column Name	Description
country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services. Given as %age of the Total GDP
health	Total health spending as %age of Total GDP
imports	Imports of goods and services. Given as %age of the Total GDP
Income	Net income per person
Inflation	The measurement of the annual growth rate of the Total GDP
life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

This project was entirely executed on python

Libraries Used :

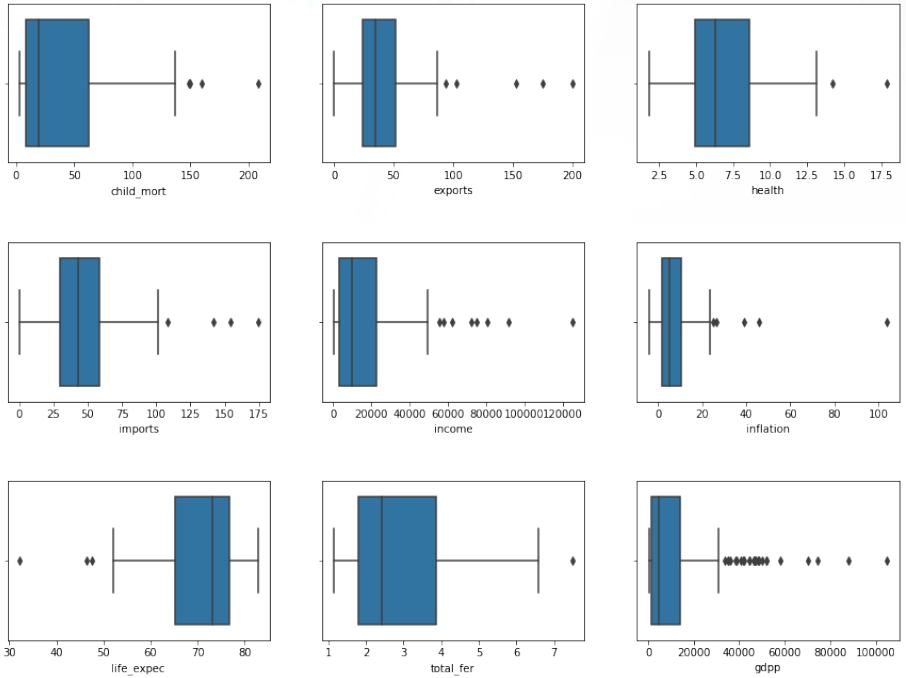
a) Numpy
b) Pandas
c) Scipy
d) Sklearn
e) Pyclustertend
f) Copy

The process

- Outlier Analysis and Treatment
- Dimensionality Reduction (PCA)
- K- means clustering
- Hierarchical clustering
- K-means vs Hierarchical clustering
- Cluster Analysis and Cluster validation
- REPORTING TOP 6 COUNTRIES IN DIRE NEED OF FINANCIAL AID

Outlier analysis

Spread of the features related to Help NGO



Countries removed after outlier analysis

```
'Australia',
'Austria',
'Bahrain',
'Belgium',
'Brunei',
'Canada',
'Cyprus',
'Denmark',
'Equatorial Guinea
'Finland',
'France',
'Germany',
'Iceland',
'Ireland',
'Italy',
'Japan',
'Kuwait',
'Luxembourg',
'Netherlands',
'New Zealand',
'Norway',
'Oman',
'Qatar',
'Saudi Arabia',
'Singapore',
'South Korea',
'Spain',
'Sweden',
'Switzerland',
'United Arab Emirates'
'United Kingdom',
'United States'}
```

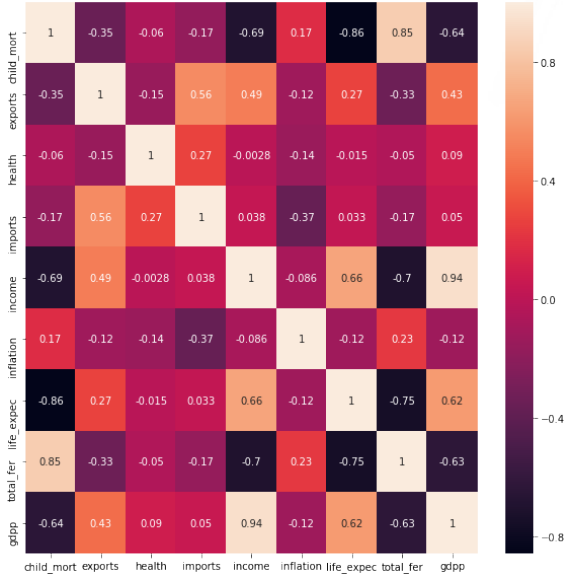
Bahamas
Czech Republic
Greece
Israel
Malta
Portugal
Slovenia

We can verify subjectively that all these countries are Developed Economies

Dimensionality Reduction (Pca)

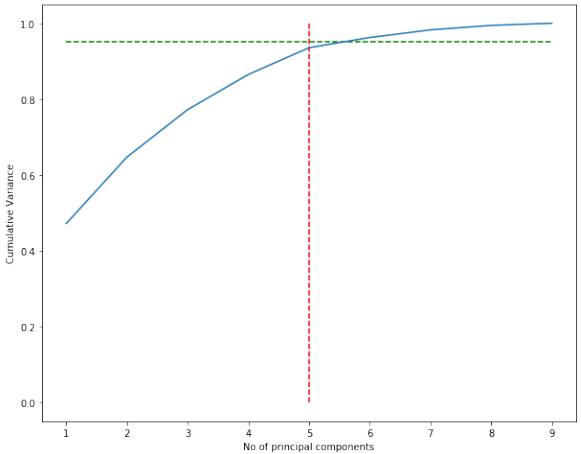
CO-REALTION BETWEEN FEATURES BEFORE PCA

Co-relation matrix of features in HELP NGO

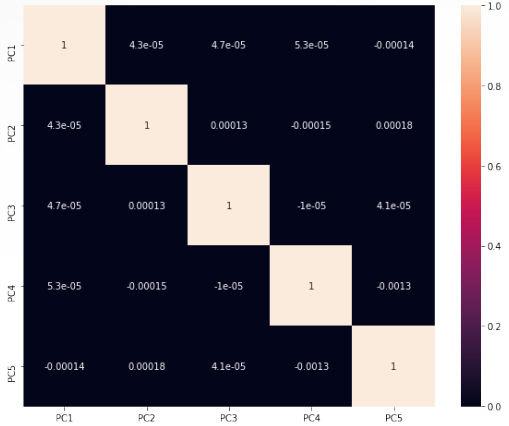


CHOOSING 5 PRINCIPAL COMPONENTS CAPTURING ABOUT 95% OF VARIANCE

Scree-plot to determine number of principal components to retain



Heatmap showing un co-related principal components after dimensionality reduction



HEATMAP SHOWING ZERO CO-RELATION BETWEEN PRINCIPAL COMPONENTS

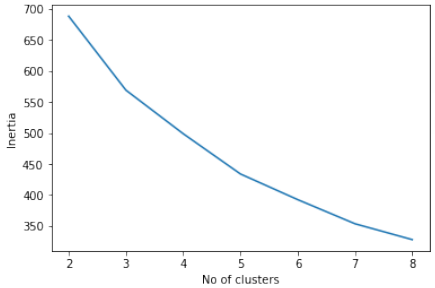
K-Means Clustering

HOPKINS STATISTIC

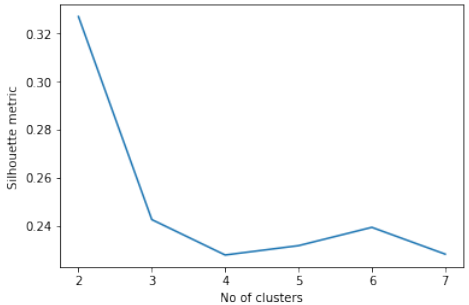
```
my_hopkins(pca_df)
```

```
0.7660381292670496
```

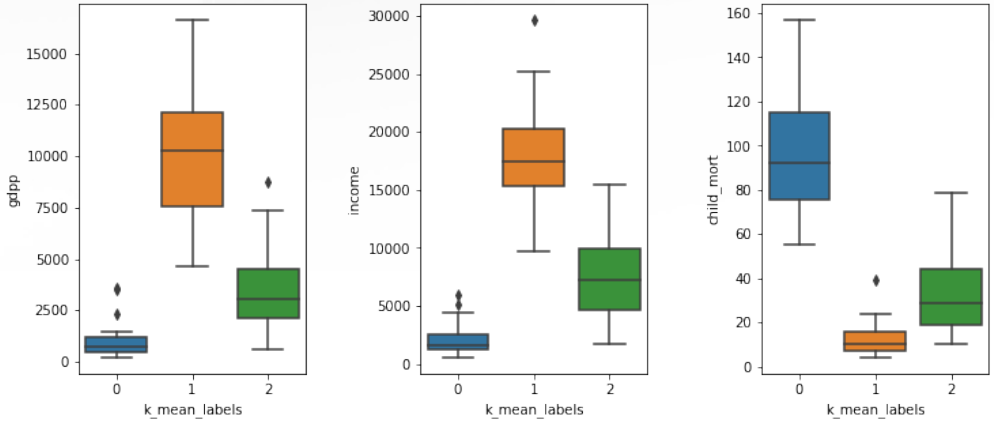
Elbow-curve method for finding the optimal k for k-means



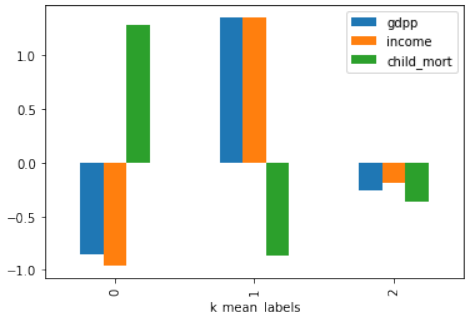
Silhouette method for finding the optimal k for k-means



Boxplots to check inter cluster heteroscedasticity and intra cluster homoscedasticity using K means clustering

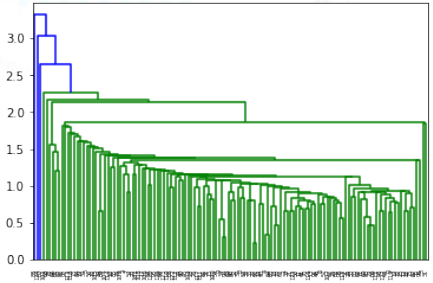


Mean of gdpp,income and child mortality across various cluster labels for K Means

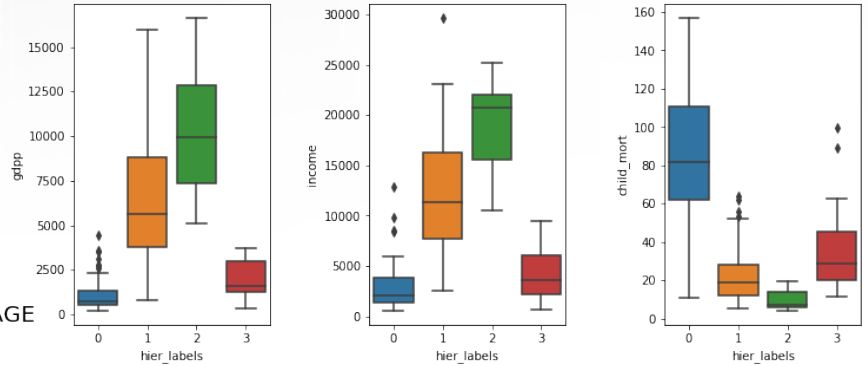


Hierarchical Clustering

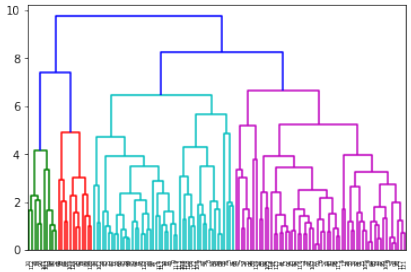
DENDROGRAM FOR REPRESENTING HIERARCHICAL CLUSTERING USING SINGLE LINKAGE



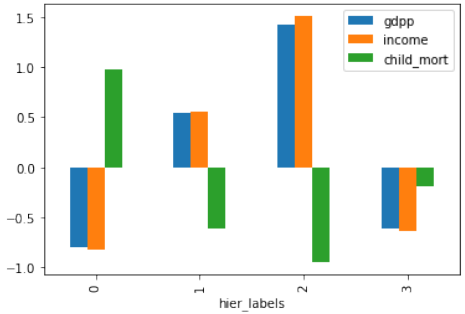
Boxplots to check inter cluster heteroscedasticity and intra cluster homoscedasticity using Hierarchical clustering



DENDROGRAM FOR REPRESENTING HIERARCHICAL CLUSTERING USING COMPLETE LINKAGE

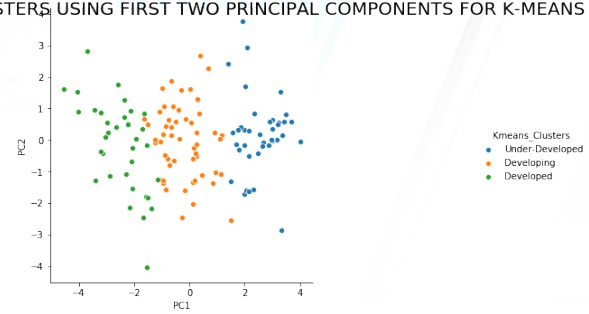


Mean of gdpp,income and child mortality across various cluster labels for Hierarchical

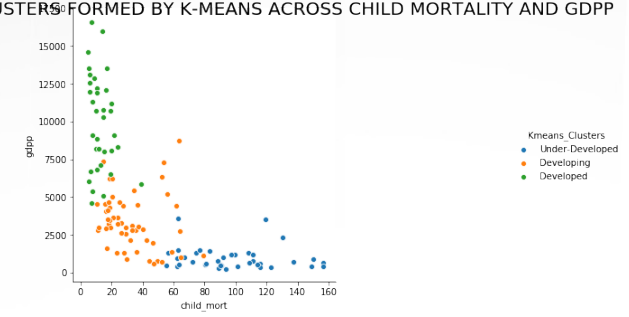


K-means vs Hierarchical clustering, Cluster Analysis

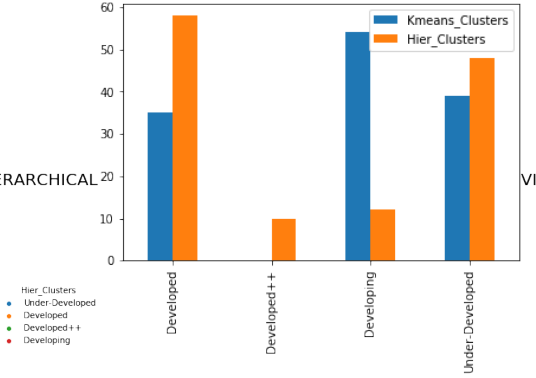
VISUALIZING CLUSTERS USING FIRST TWO PRINCIPAL COMPONENTS FOR K-MEANS



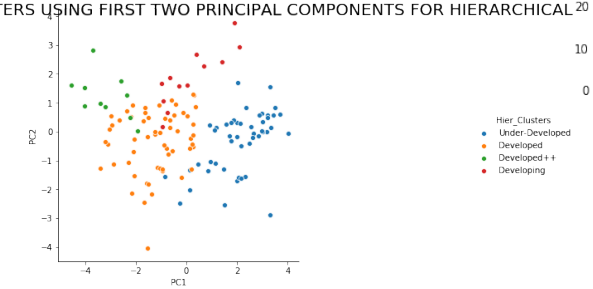
VISUALIZING CLUSTERS FORMED BY K-MEANS ACROSS CHILD MORTALITY AND GDPP



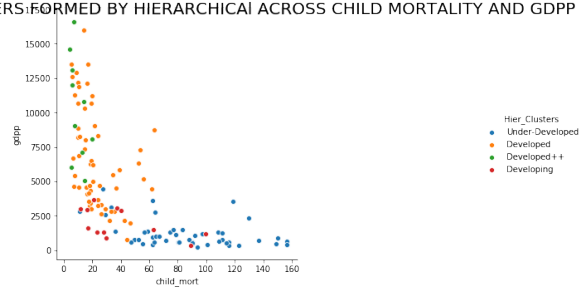
NO OF COUNTRIES UNDER EACH CLUSTER FOR KMEANS AND HIERARCHICAL



VISUALIZING CLUSTERS USING FIRST TWO PRINCIPAL COMPONENTS FOR HIERARCHICAL



VISUALIZING CLUSTERS FORMED BY HIERARCHICAl ACROSS CHILD MORTALITY AND GDPP



REPORTING OF COUNTRIES

The countries in dire need of aid are:

1) Central African Republic
2) Congo
3) Guinea-Bissau
4) Niger
5) Sierra Leone

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Central African Republic	149.0	11.8	3.98	26.5	888	2.01	47.5	5.21	446
Congo, Dem. Rep.	116.0	41.1	7.91	49.6	609	20.80	57.5	6.54	334
Guinea-Bissau	114.0	14.9	8.50	35.2	1390	2.97	55.6	5.05	547
Niger	123.0	22.2	5.16	49.1	814	2.55	58.8	7.49	348
Sierra Leone	156.6	16.8	13.10	34.5	1220	17.20	55.0	5.20	399