Clustering & PCA

ASSIGNMENT

Name: Ashish Gaurav

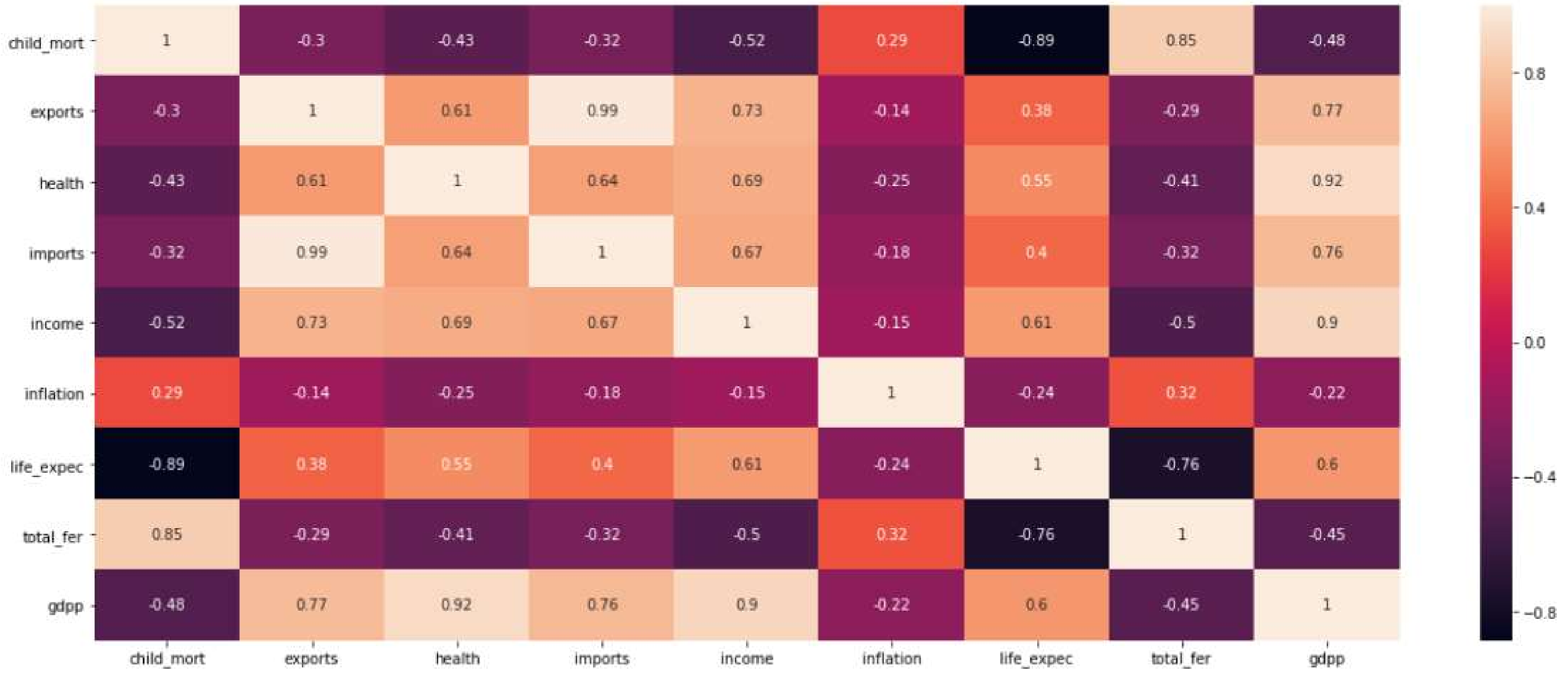
Application ID UpGrad: APFE18805048

<Abstract - Objective and Problem Statement>

- HELP is an international humanitarian NGO that fights poverty and provides basic amenities to the people of backward communities.
- HELP has managed to raise 10 million dollars for the unfortunate.
- But it needs to decide the countries that are in their worst situation right now. The money will be spent for relief to people in those countries.
- Thus we have to identify groups of countries with similar socio-economic conditions.

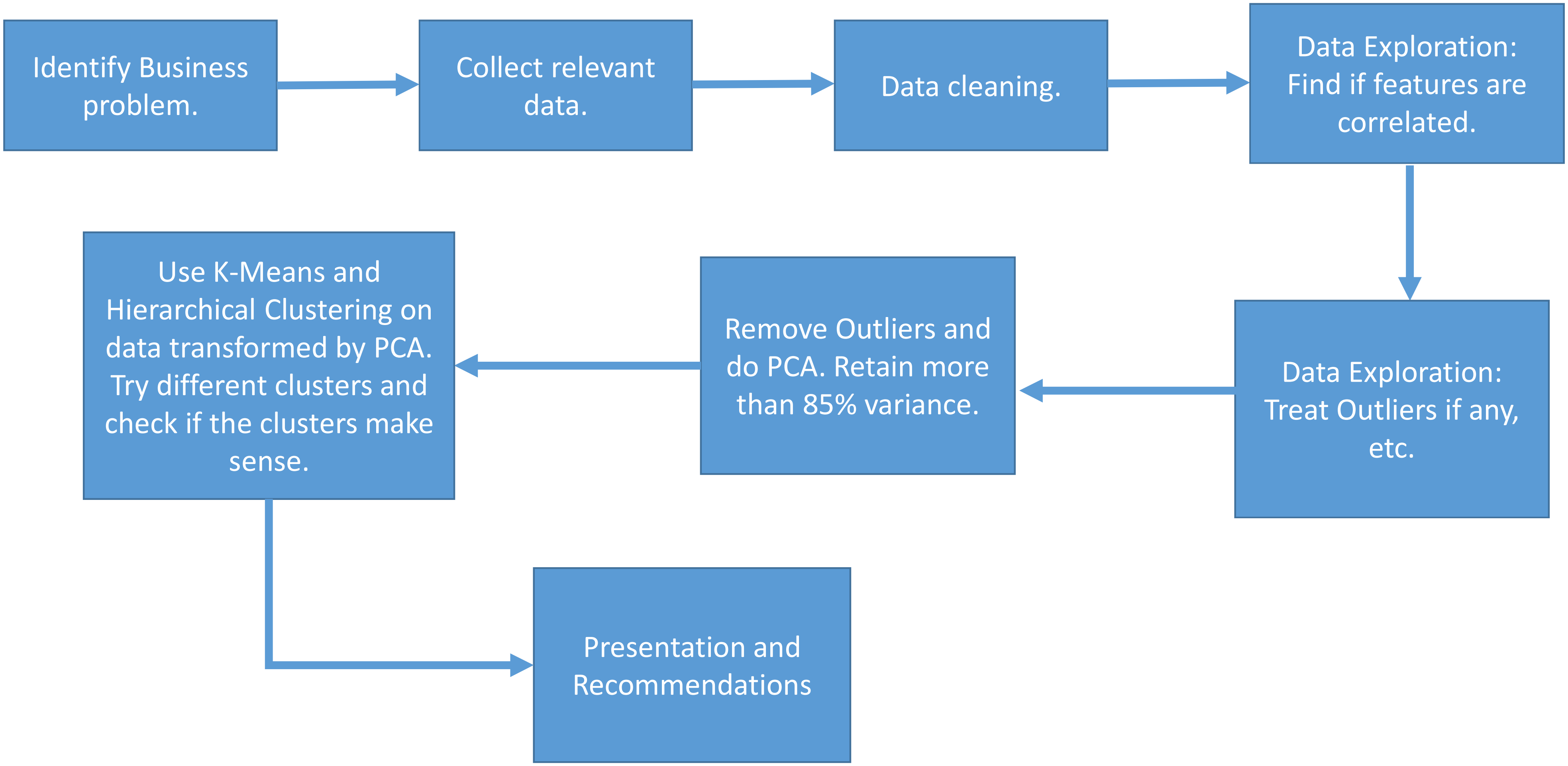
<Some Preliminary Observations>

- Features are highly correlated. For example- income and gdpp, life_expec and income, total_fertility and child_mort, life_expec and child_mort, etc.
- PCA can find new features that are uncorrelated, as well as do dimensionality reduction.



	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52	0.73	0.69	0.67	1	-0.15	0.61	-0.5	0.9
inflation	0.29	-0.14	-0.25	-0.18	-0.15	1	-0.24	0.32	-0.22
life_expec	-0.89	0.38	0.55	0.4	0.61	-0.24	1	-0.76	0.6
total_fer	0.85	-0.29	-0.41	-0.32	-0.5	0.32	-0.76	1	-0.45
gdpp	-0.48	0.77	0.92	0.76	0.9	-0.22	0.6	-0.45	1

<Approach followed>



Identify Business problem.

Collect relevant data.

Data cleaning.

Data Exploration: Find if features are correlated.

Data Exploration: Treat Outliers if any, etc.

Remove Outliers and do PCA. Retain more than 85% variance.

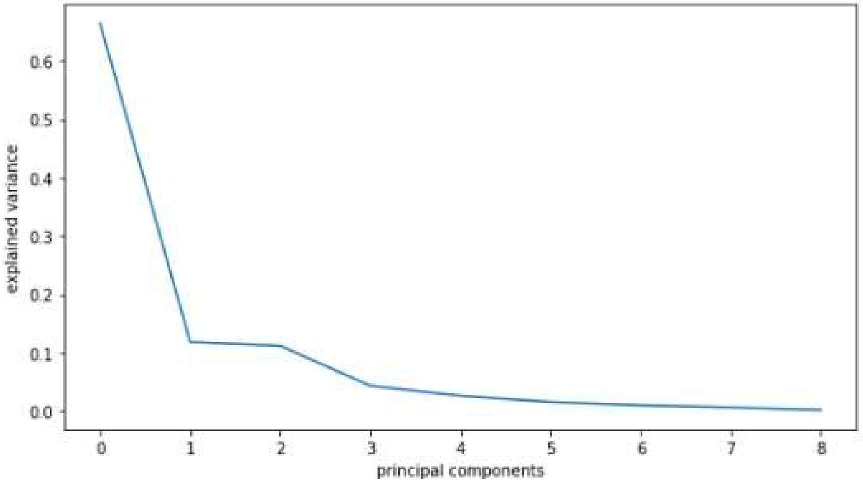
Use K-Means and Hierarchical Clustering on data transformed by PCA. Try different clusters and check if the clusters make sense.

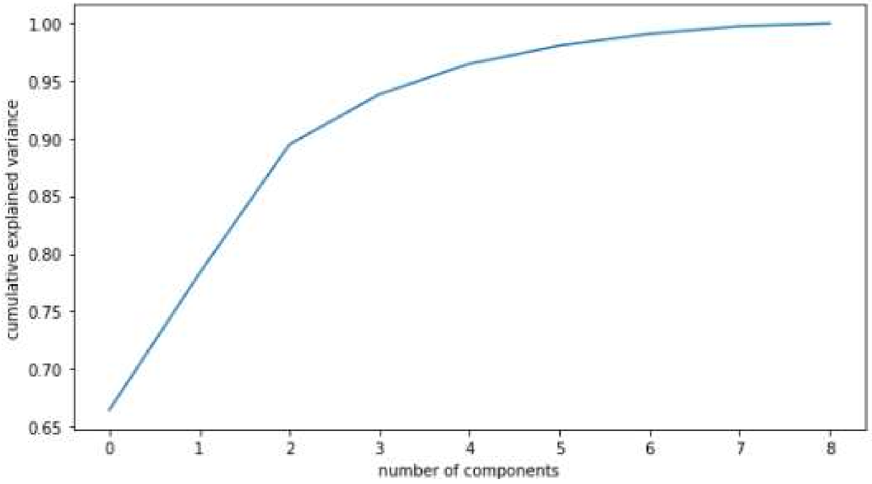
Presentation and Recommendations

<1. PCA-Scree Plot>

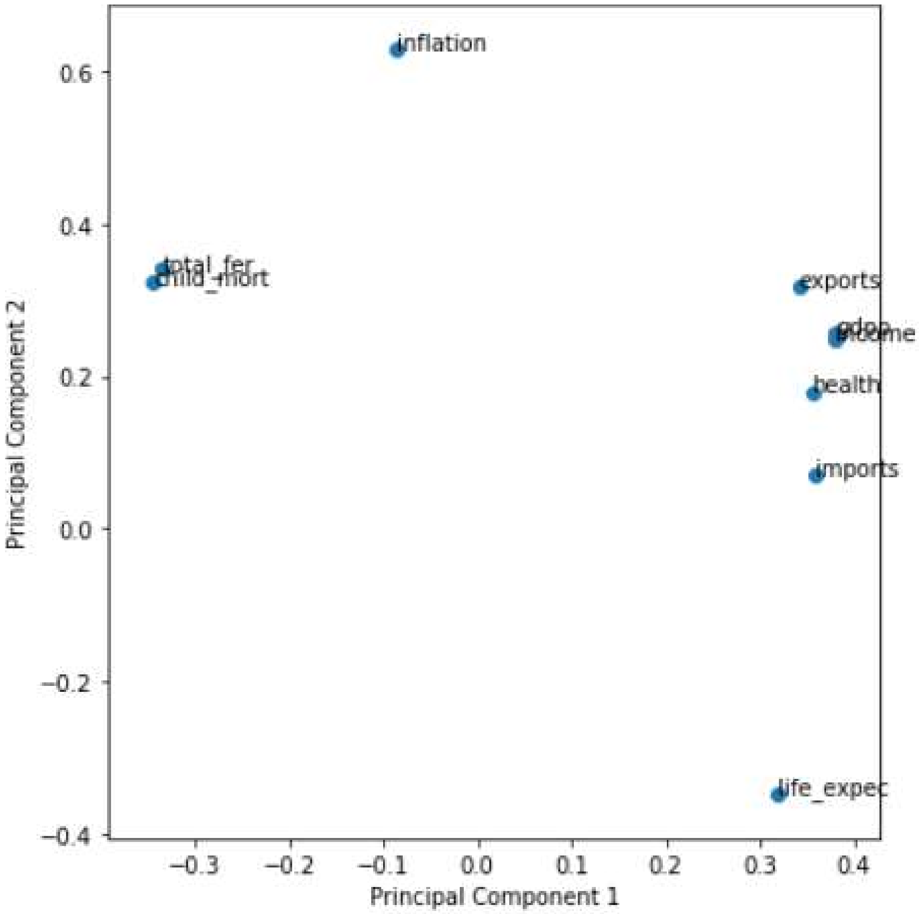
- As we had highly correlated features, we used PCA to find new components that are uncorrelated as well as they retain the variance of the original data.
- Left graph-explained variance ratio. Right graph-Cumulative explained variance.
- Two methods to select number of Principal Components: (https://www.r-bloggers.com/pca-and-k-means-clustering-of-delta-aircraft/)

  1. Pick the number of components which explain 85% or greater of the variation.
  2. Use the 'elbow' method of the scree plot (on left, which has explained variance on y-axis).
- As elbow is not very clear on the left plot, we'll use the cumulative explained variance plot on the right. Thus, we can stat by selecting first 3 principal components as they explain 89.5% of the variance.



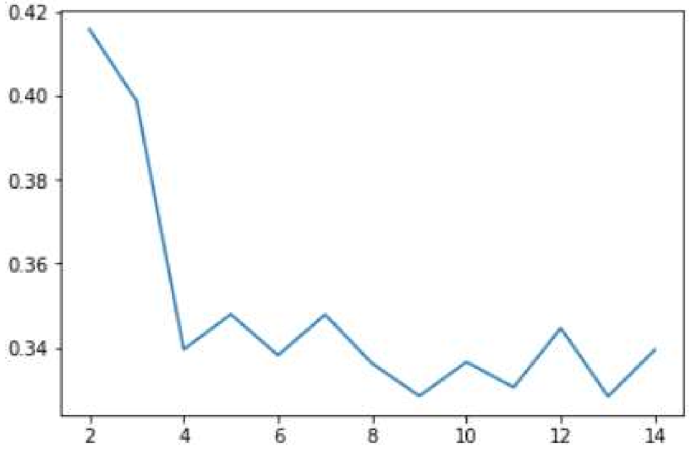


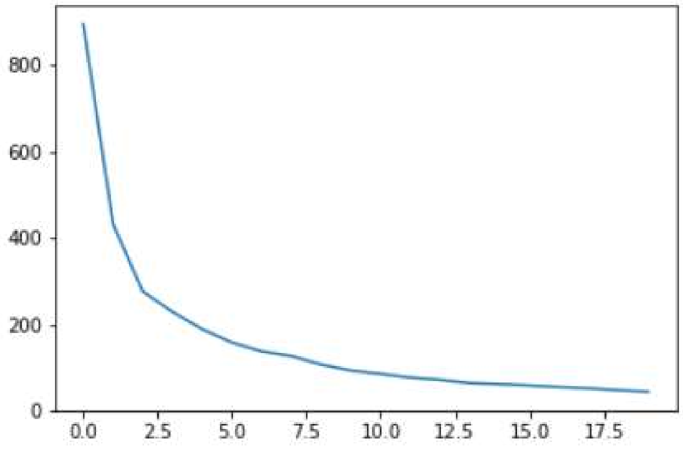
<1. PCA-What does the principal component capture.>



- PCA finds new components or features that are linear combinations of original features.
- On first principal component, imports, health, exports, gdpp, life_expec, income seem to be on right side, and inflation, total_fer and child_mort seem to be on the left side.
- Similarly, for second principal component, import, health, exports, inflation are high and, life_expec is on the lower side.
- The reason that some features appear together is because these features are correlated and vary together.

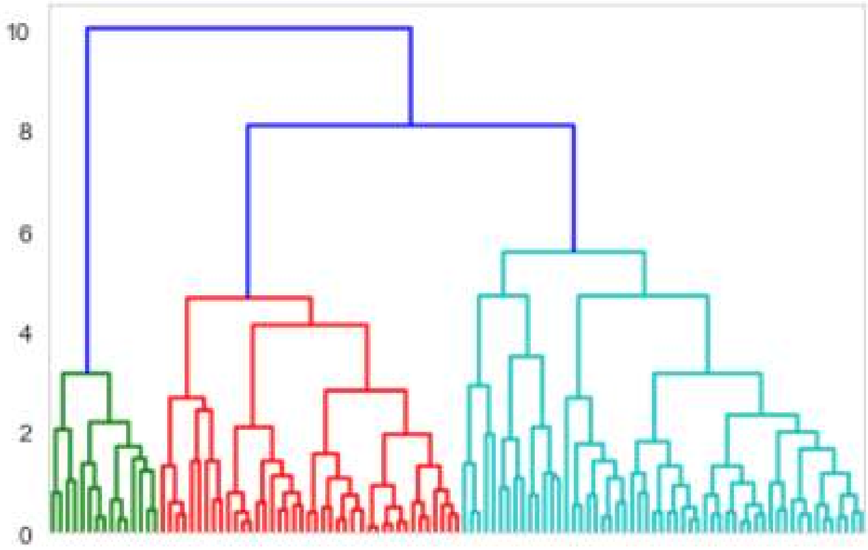
<2. Clustering, a. K-Means Clustering>

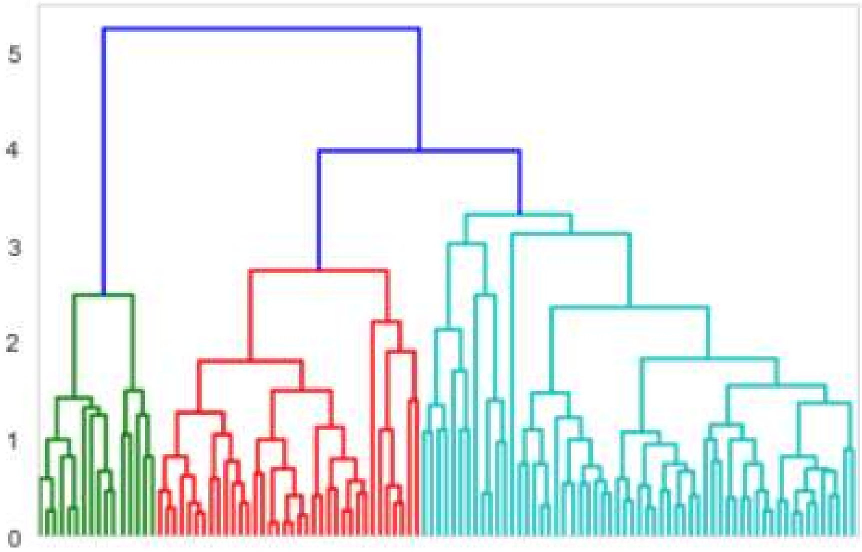




- We have two methods for clustering-a. K-Means Clustering, b. Hierarchical Clustering.
- We need to select the number of clusters when applying K-Means clustering but not in Hierarchical Clustering.
- Now, we apply clustering algorithm on transformed data produced by PCA.
- The top plot is something called silhouette score. And the bottom plot contains plot of within sum of squared distances.
- How do we know that the number of clusters that we choose make sense? These two plots will help us choose the number of clusters.
- X-axis is the number of clusters and Y-axis is the score.
- Silhouette score must be high and within sum of squared distances plot must be low.
- Number of cluster is selected using the same elbow method described in Page 5.
- Elbow in ssd curve seems to be at 3. Let's try 3 and 4 clusters first and see if clusters make sense.
- It is to be noted that these only help us decide the number of clusters. We have to check the data ourselves to identify whether the cluster makes sense or not.

<2. Clustering, b. Hierarchical Clustering>





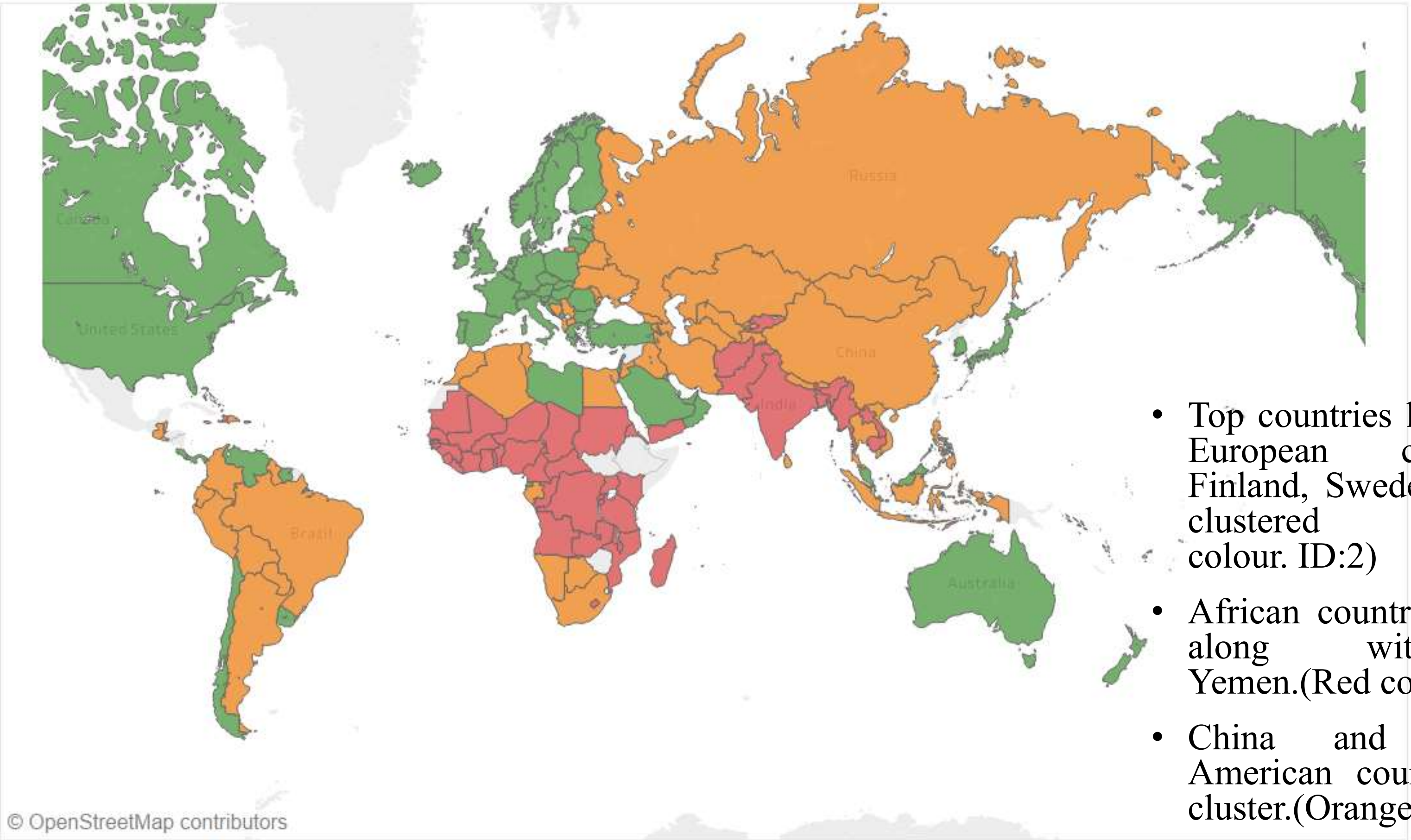
- ‘Average’ and ‘Complete’ linkage method was used.
- ‘Complete’ linkage method seemed to show slightly better results, as the clusters are formed at a slightly lower height. Although, both methods show very similar results. Dendogram can be seen on left.
- ‘Average’ linakage method Dendogram can be seen on right.

<2. Clustering, c. Analysis of Clusters.>

- 2 Clustering Algorithms were used-K-Means and Bottom-Up Hierarchical.
- 3 and 4 clusters were formed using each method.
- The outlier data was then assigned cluster labels using the Euclidean distance from the clusters in case of Hierarchical clustering. And, in case of K-Means clustering the .predict() function of scikit learn was used.
- Original and Outlier data were merged along with their corresponding Cluster labels.
- Original Feature mean and medians were identified for different clusters.
- It was observed that, in this case, Hierarchical Clustering with 3 clusters produced simple, clear and distinct results.
- Results of Hierarchical Clustering with 3 clusters can be seen in the next page.

<2. Clustering, c. Analysis of Clusters.>

Hierarchical, 3 Clusters



© OpenStreetMap contributors

- Top countries like US, Canada, European countries like Finland, Sweden, Germany are clustered together.(Green colour. ID:2)
- African countries are clustered along with Pakistan, Yemen.(Red colour. ID:0)
- China and other South American countries form one cluster.(Orange colour. ID:1)

< 2. Clustering, c. Analysis of Clusters.>

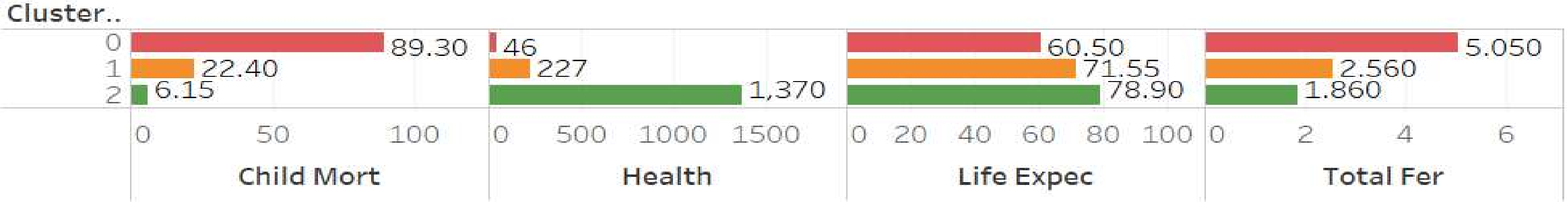
- Characteristics of Clusters:

1. ClusterID 0: This cluster has highest child mortality, lowest income and gdpp, highest inflation, lowest life expectancy and highest fertility rate. Note that instead of having high fertility rate, most of the children die because of diseases. This cluster contains African countries, Lao, Afghanistan and Pakistan.

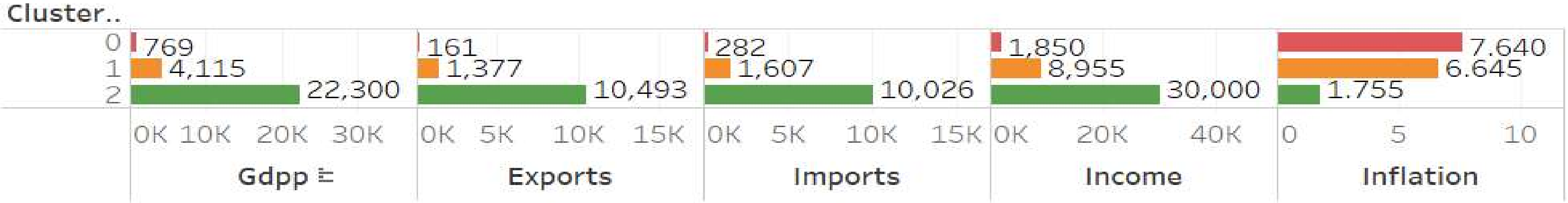
2. ClusterID 1: This cluster contains countries which are in between better doing and worst countries, i.e., Cluster 1 and Cluster 2. Medium Child Mortality, Medium Income, High Inflation, Good Life Expectancy, Apt Fertility rate, Medium GDPP. Some countries belonging to this country are: China, Russia, Brazil, etc.

3. ClusterID 2: This cluster contains first world countries. Low child mortality, High Health and Life Expectancy, Good Exports, High GDPP and Income, etc. US, Nordic Countries, European Countries, Australia, Gulf Countries, Japan, etc. belong to this Cluster.

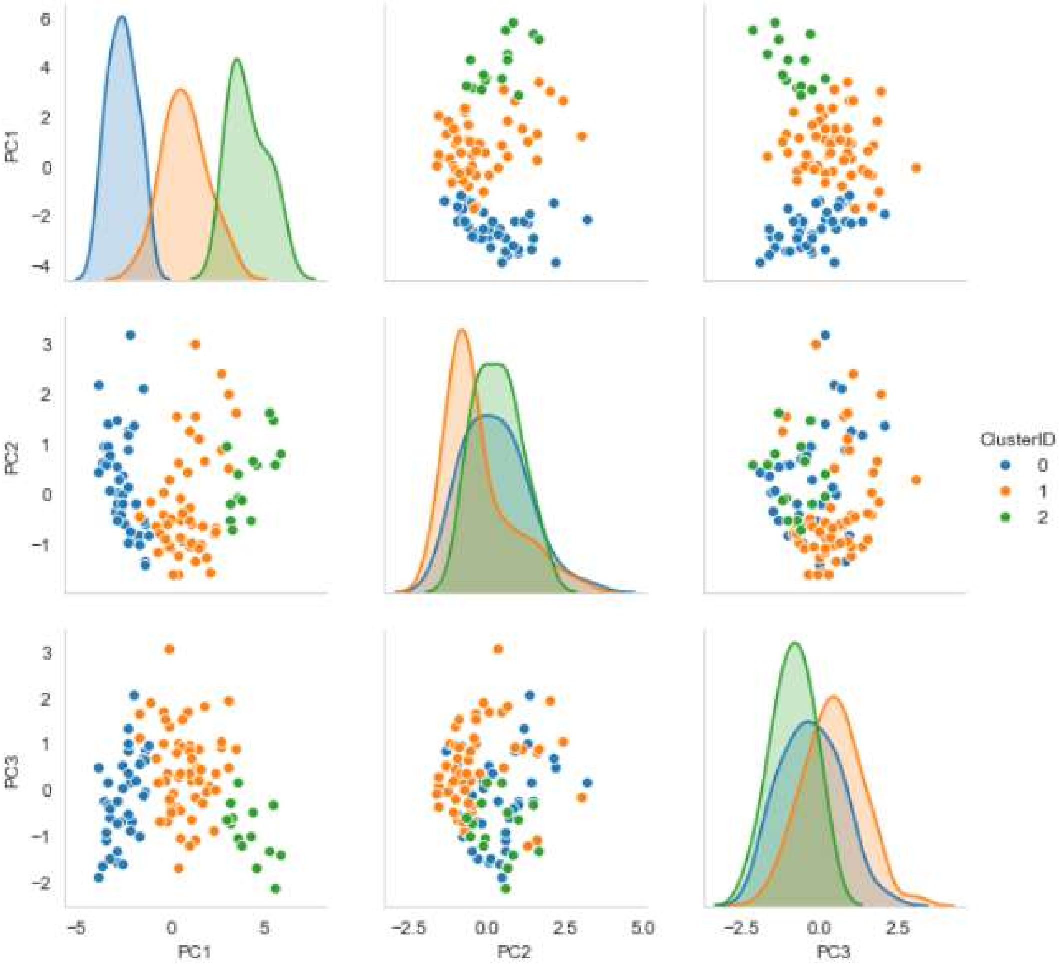
Stats - Health



Stats - Economic



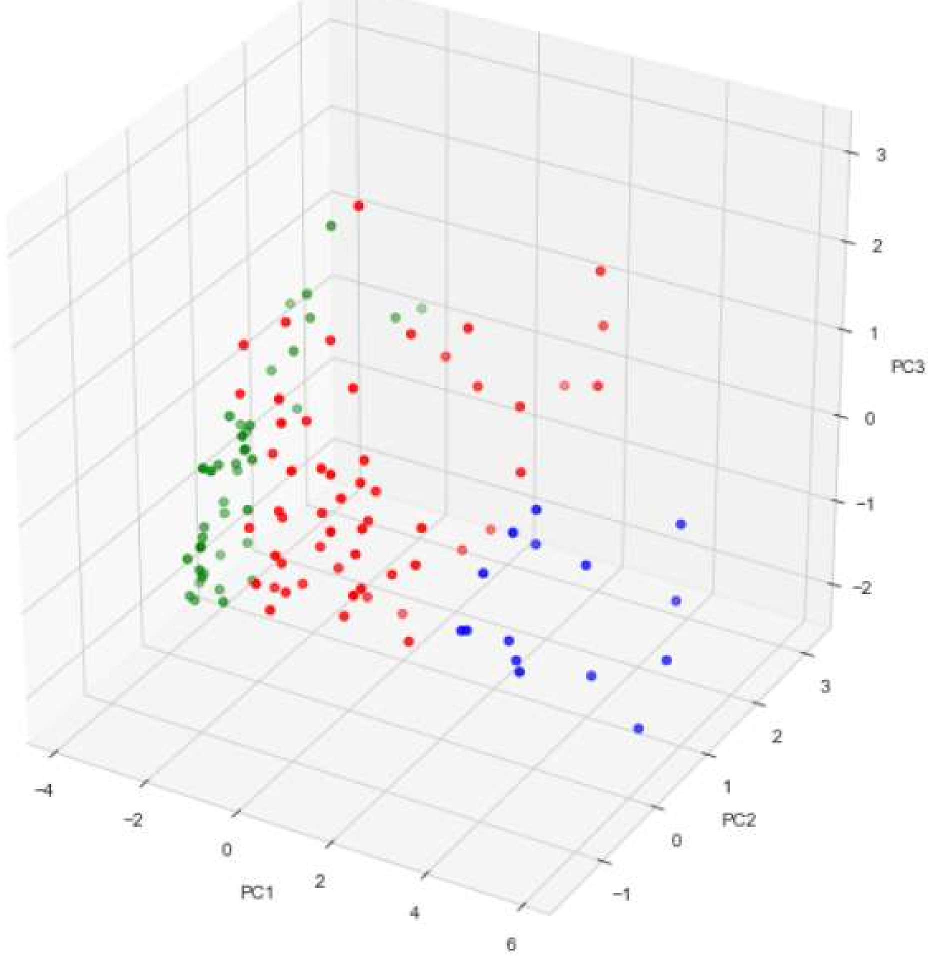
< 2. Clustering, d. Inference.>



- How the data points of principal components are spread.
- What clusters are those data points attached to.
- Hierarchical, 3-Clusters

< 2. Clustering, d. Inference.>

Some intuition in 3D.



<3. Results, a. Clustering on all countries>

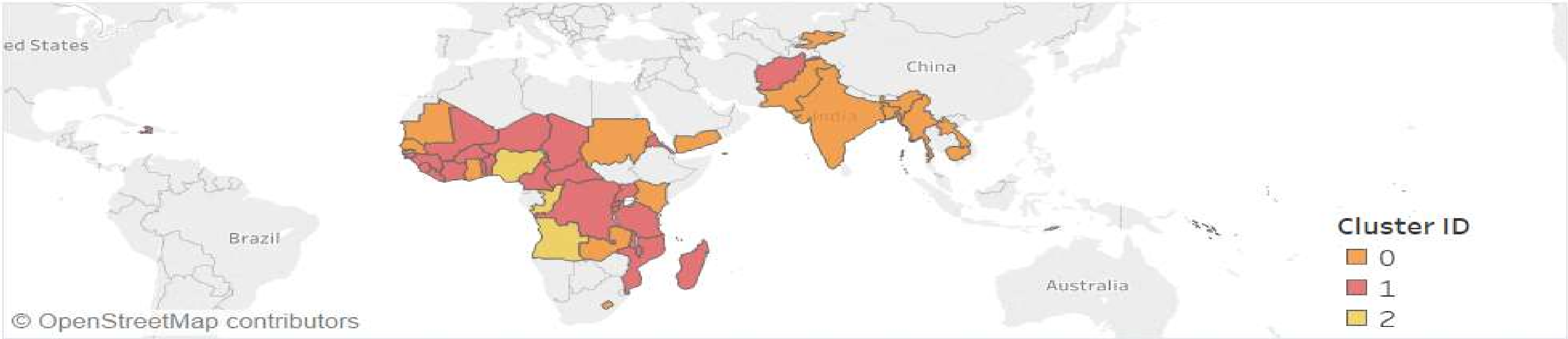
- 'Afghanistan', 'Angola', 'Bangladesh', 'Benin', 'Burkina Faso', 'Burundi', 'Cambodia', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo, Dem. Rep.', 'Congo, Rep.', "Cote d'Ivoire", 'Eritrea', 'Gambia', 'Ghana', 'Guinea', 'Guinea-Bissau', 'Haiti', 'India', 'Kenya', 'Kiribati', 'Kyrgyz Republic', 'Lao', 'Lesotho', 'Liberia', 'Madagascar', 'Malawi', 'Mali', 'Mauritania', 'Mozambique', 'Myanmar', 'Niger', 'Nigeria', 'Pakistan', 'Rwanda', 'Senegal', 'Sierra Leone', 'Solomon Islands', 'Sudan', 'Tanzania', 'Timor-Leste', 'Togo', 'Uganda', 'Yemen', 'Zambia'– 47 countries.
- The above countries belong to cluster 0. These are mostly African countries except Pakistan, Afghanistan and Lao.
- Mostly they are African countries. This is the only place where disease still wipes a lot of population. This is why Bill & Melinda Gates foundation has Africa as a primary target for health and education facilities.
- According to their findings improving health conditions will not increase population as people will choose to have smaller families. And it can also be seen that Cluster 0, i.e. African countries has highest fertility rate.
- https://www.youtube.com/watch?v=obRG-2jurz0 – watch this video.

<3. Results, b. Further clustering on the identified 48 countries >

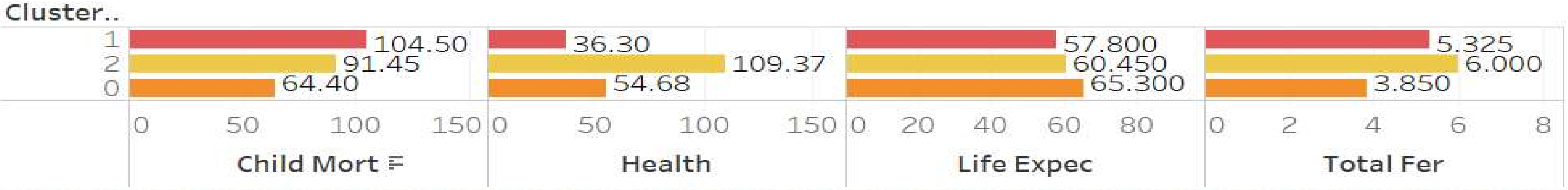
- Further clustering of the identified 47 countries belonging to Cluster0 was done.
- 3 clusters were selected.
- Yellow cluster(ClusterID: 2) and Red cluster(ClusterID: 1) are basically two extremes. Red has poor health and economic conditions and Yellow cluster has better health and economic conditions.
- The Orange(ClusterID: 0) cluster are sort of in between them.
- The order of importance given to clusters must be red(immediate attention), orange, yellow (least attention.)
- As red cluster requires the most attention, we have further shortlisted 26 countries from 48 identified countries earlier.
- Statistics associated with clusters can be seen in the next page. Note that median values are plotted in graph and NOT mean.

<3. Results, b. Further clustering on the identified 48 countries>

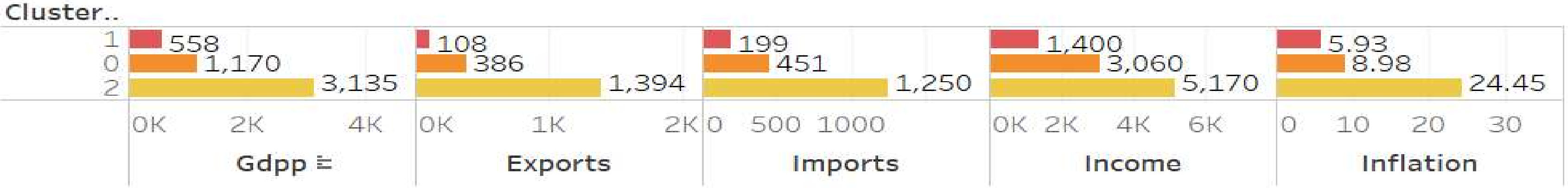
Cluster ID 0, 3 Clusters



Stats - Health, Clus ID 0, 3 Clusters



Stats - Economic, Clus ID 0, 3 Clus



<3. Results, b. Further clustering on the identified 48 countries>

- 26 countries belonging to the red cluster that need immediate attention:
- 'Afghanistan', 'Benin', 'Burkina Faso', 'Burundi', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo, Dem. Rep.', "Cote d'Ivoire", 'Eritrea', 'Gambia', 'Guinea', 'Guinea-Bissau', 'Haiti', 'Liberia', 'Madagascar', 'Malawi', 'Mali', 'Mozambique', 'Niger', 'Rwanda', 'Sierra Leone', 'Tanzania', 'Togo', 'Uganda'