CLUSTERING OF COUNTRIES ASSIGNMENT

Categorising countries based on socio-economic and health factors and choosing the countries for monetary aid

Vishak Nair

17th August2019

STEPS FOLLOWED FOR THE PCA AND CLUSTERING

- Understanding the Data
- Data Preparation and Exploration
- Principal Component Analysis on the dataset
- Outlier analysis after PCA
- K-Means Clustering
- Cluster Analysis
- Hierarchical Clustering

➢The stats and the shape of the dataset where well observed initially before any action.
➢ Null values, Outliers where checked initially and the amount of correlation and taken a call for data preparation.
➢ Standard Scaling was done inorder to bring all the features in comparable scale.
➢ PCA was run and found the optimal Principal components
➢ K-Means algorithm was run to find out distinct clusters.
➢ Analysis on different clusters to find out distinct countries that stand out since they are socio-economically backward
➢Hierarchical Clustering inorder to validate or as a different approach.

DATASETS USED

- Country-data Dataset -> Contains the list of countries and their corresponding various socio economic factors

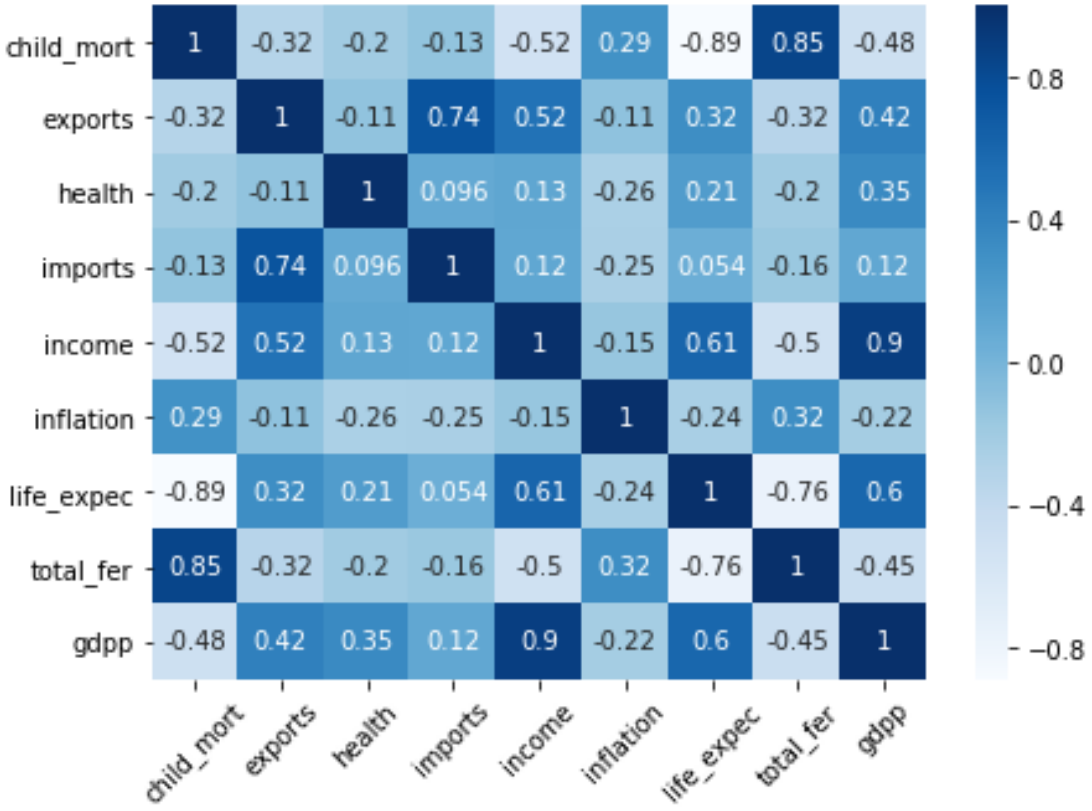
Problem Statement

- The $ 10 million dilemma – Which all countries might need a share?
- Categorise all the countries around the globe into distinct categories based on socio-economic or health factors that determine the overall development of the country.
- The factors that can be considered are Child mortality ratio, exports, imports, health indicators, inflation, life expectancy, total fertility and the most important GDPP.

The Approach

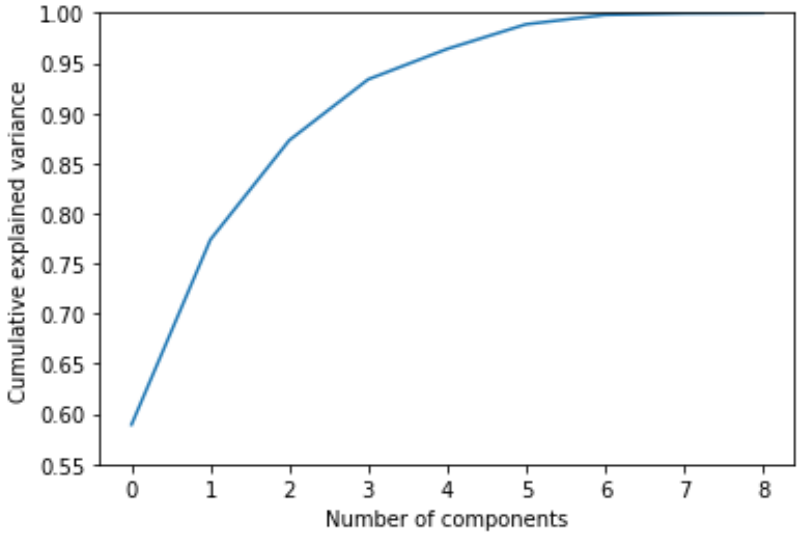
- We started with **understanding the dataset** and tried to get a look and feel of the data. This includes seeing for null values, outlier detection and basic correlation plots as well.
- **Data Preparation** stage includes converting the features such as Exports, Imports and Health which are given as % of GDP, to normal ranges.
- **Standard Scaling** of all the numeric features is other step which we did as part of Data Preparation to get all the features in common range.
- We went ahead with **PCA**. Found the different PCA components and the variance explained by each component. Have used the help of **screeplot** for deciding how many clusters to be selected and how much variance to be explained by them.
- Features were converted to Principal Components(PC) and **outlier analysis** were done on different of these PC's and decision was made whether to retain them or not.
- Checking the dataset is clusterable with **Hopkins approach.**
- Approaches of **Silhouette score** and **Elbow curve** to find out the right number of clusters.
- **K-Means clustering** finally to get the cluster labels and to categorise the clusters.
- **Cluster Analysis** was done on final clusters to identify which are developed nations and which are underperforming or socio-economically backward countries.
- Further use of **Hierarchical Clustering** for validation or as an different approach to K-Means.

Correlation between the Features



- Initial check of the correlation matrix with the above heat map displays that there is high correlation of 0.85 between child mortality and total fertility and value of 0.9 between income and Per Capita GDP which is obvious one.
- Also we can see high negative correlation between life expectancy and child mortality as well life expectancy and total fertility which are also obvious understandings.

Screeplot



➢ From the Screeplot we can find that around 96% of the variance is explained by 5 components.

```
#Explained variance ratio
pca.explained_variance_ratio_.round(2)
```

```
array([0.59, 0.18, 0.1 , 0.06, 0.03, 0.02, 0.01, 0.  , 0.  ])
```

➢ The above shows the variance explained by each of the principal components. The same as been plotted as cumulative sum, using the screeplot.

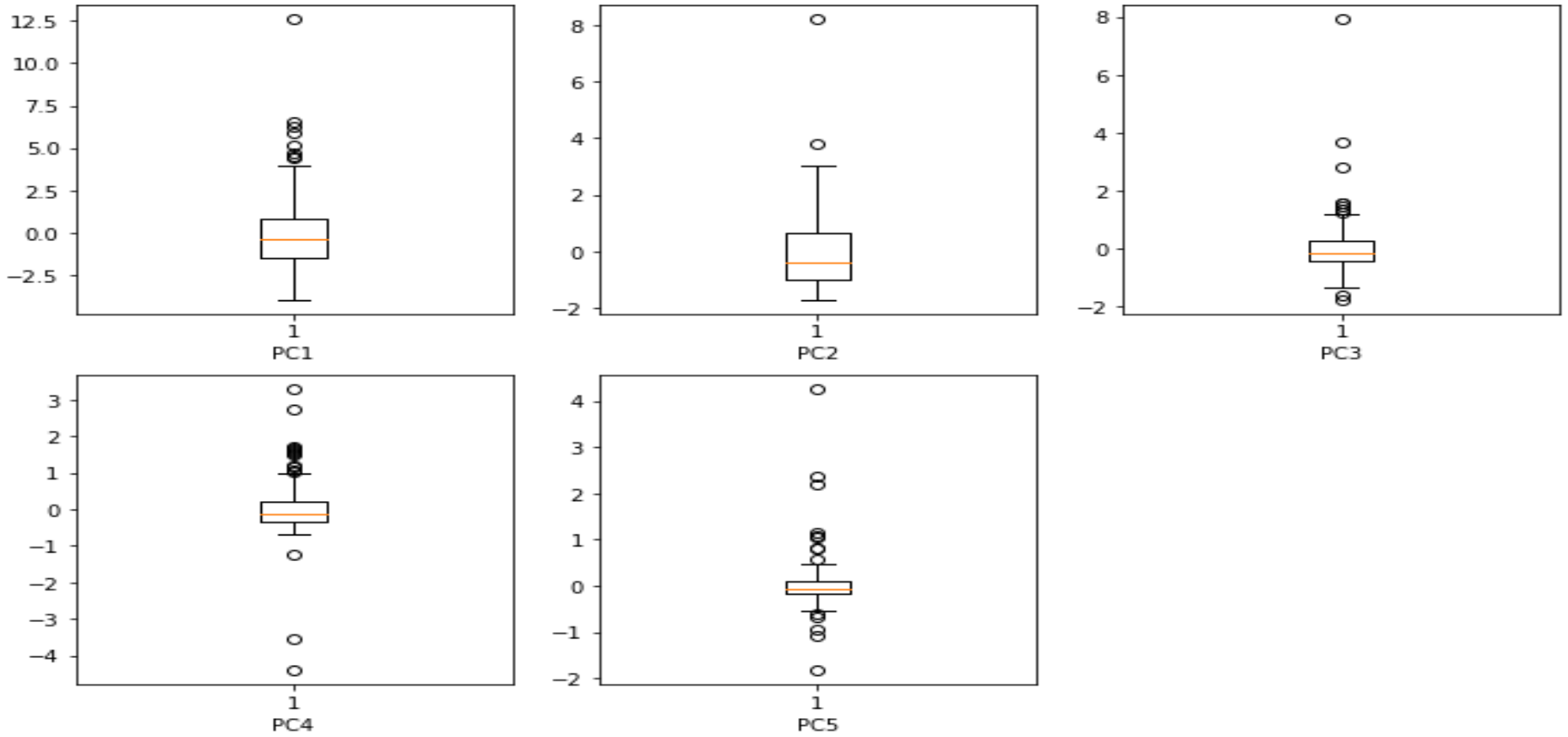
➢ Hence we will proceed with 5 principal components and go and fit transform our country data.

➢ Hence we have essentially done a dimensionality reduction on country dataset with 9 original features to a dataset with 5 principal components.

➢ Principal components thus formed would be independent to each other at the same time captures maximum variance.

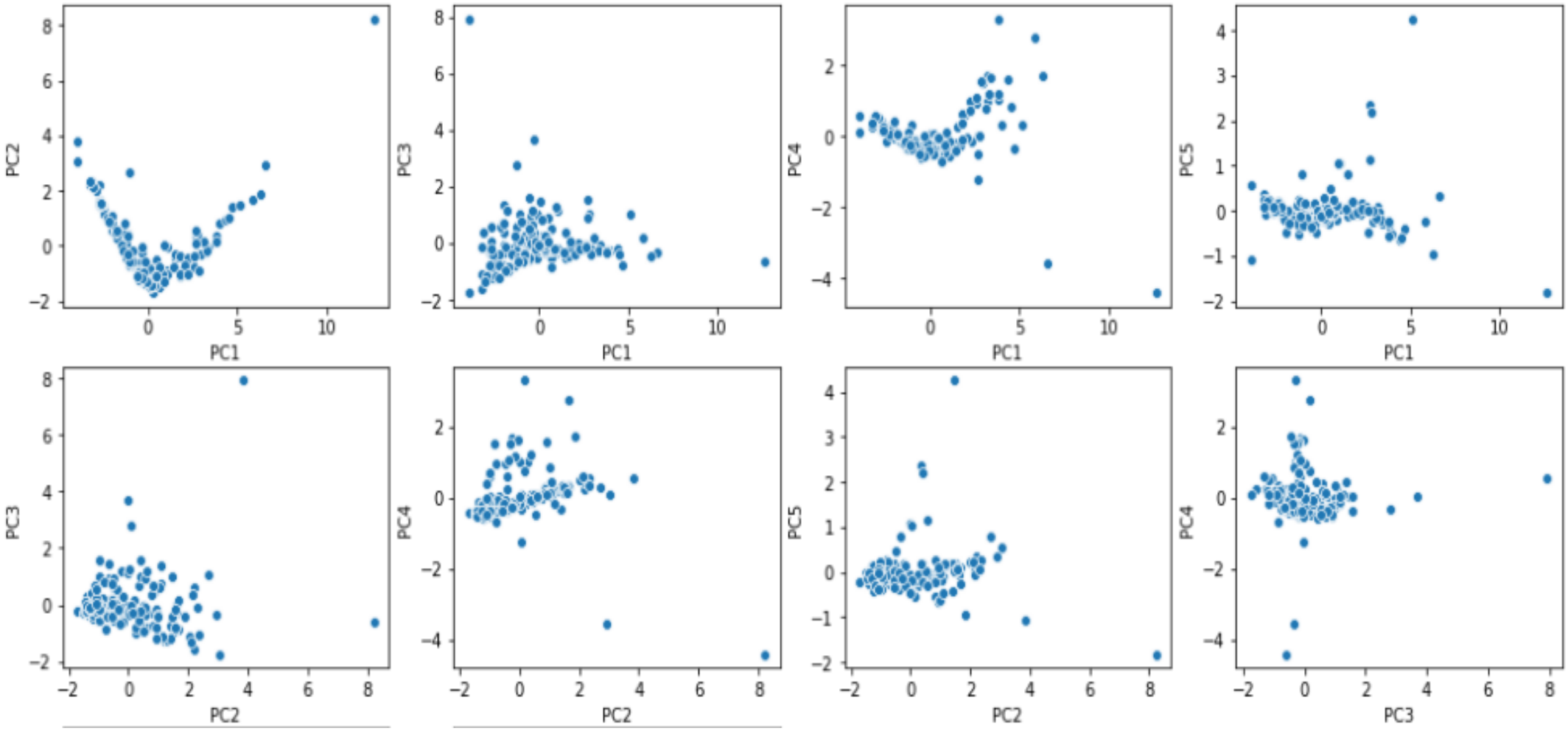
.

Outlier Analysis on Principal Components



- From the above boxplots, for different Principal Components we can see that, there are outliers present in each of the PC's. PC2 seems to be the one least affected by outliers while coming to components 4 and 5 there are many outliers present. We can remove these datapoints before proceeding to clustering, but since we know we have only around 167 datapoints with us, dropping these many points will result in loosing in many information.
- Also these outlier countries would be the countries with the least GDP or high rate of child mortality which are direct indicators of their socio-economical status. Hence we will take it forward for clustering and observe these countries in cluster.

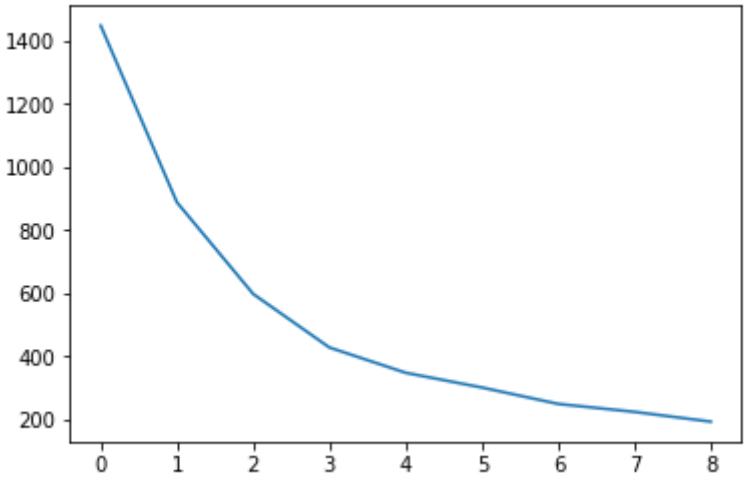
Visualising the datapoints on Principal Components



Selection of Number of Clusters

Elbow Curve

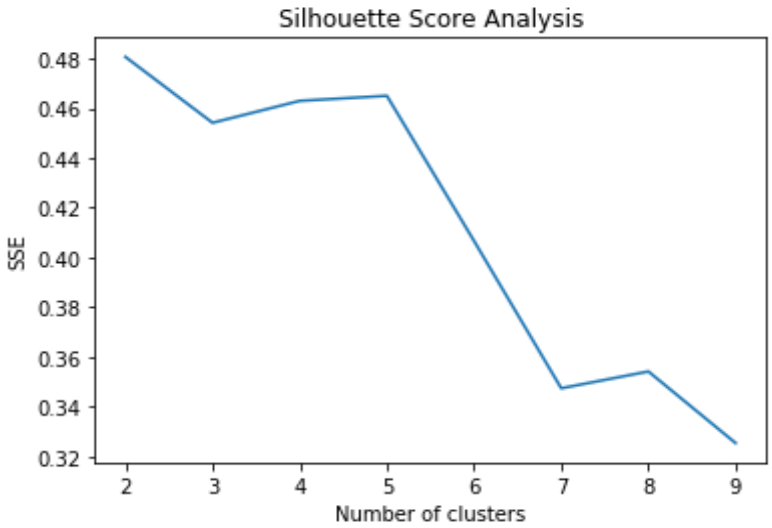
- From the Elbow Curve above we can see that, we can't decide on a clear elbow from the above plot since the plot looks like a fairly smooth curve. But with a common intuition we can say that the number of clusters can be 4 or 5.
- We can check again the above notion, with the help of other important approach of `Silhouette Score`.



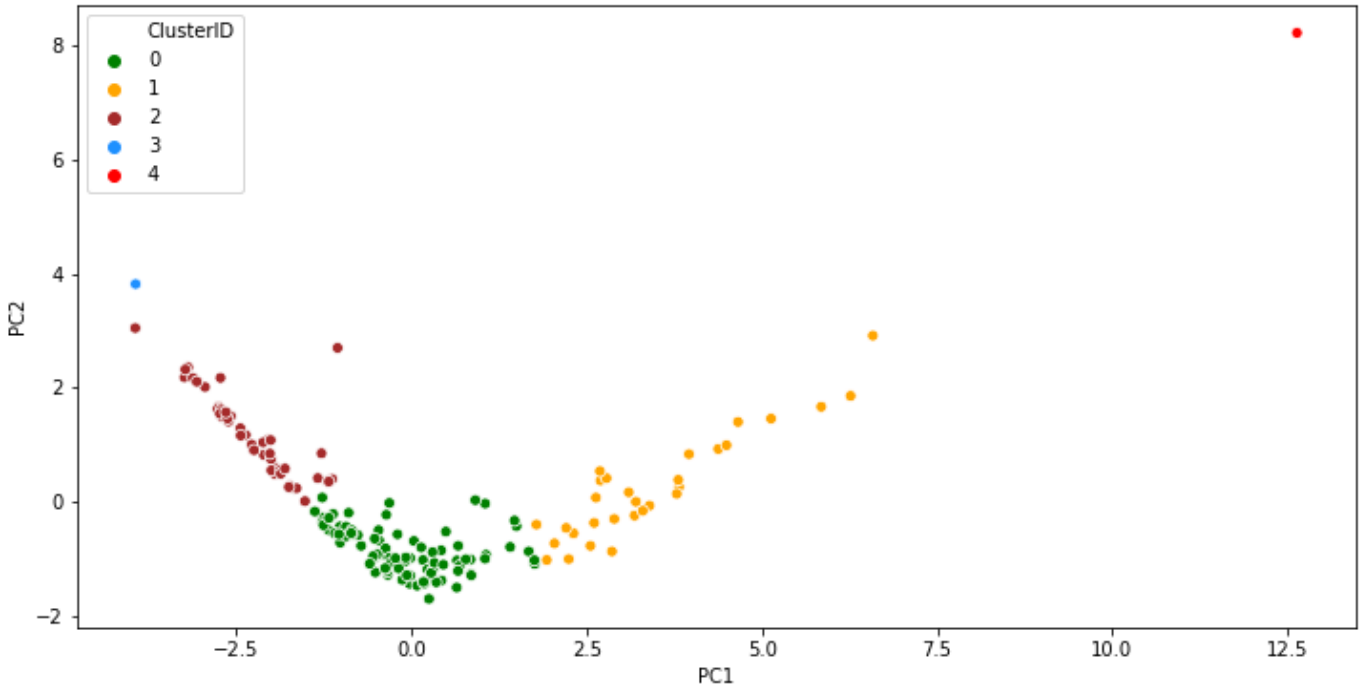
Silhouette Analysis

- From the Silhouette score plot we can see that, we are getting good scores for cluster numbers = 2,3,4 and 5. We can see a dip in the score going from 2 to 3 and an increase further going on to 5 clusters.
- We have found and proceeded with 5 PC's as well. Hence we will go with the initial intuition of 5 clusters(Note: All the cluster size of 2,3,4,5 are giving comparative good Silhouette Scores).

.



Cluster Analysis



- The above plot shows the different clusters over the Principal components of PC1 and PC2.
- We can observe that Cluster 4(one in red) and Cluster 3(one in blue) are kind of standout or in other terms Outlier Clusters.
- Below snippet shows 'NIGERIA' is the one in Cluster 3 and 'LUXEMBOURG' is in Cluster 4.
- Just observing the original features for each of this countries, we can come to know why only these 2 countries are placed in different clusters and also at opposite ends.

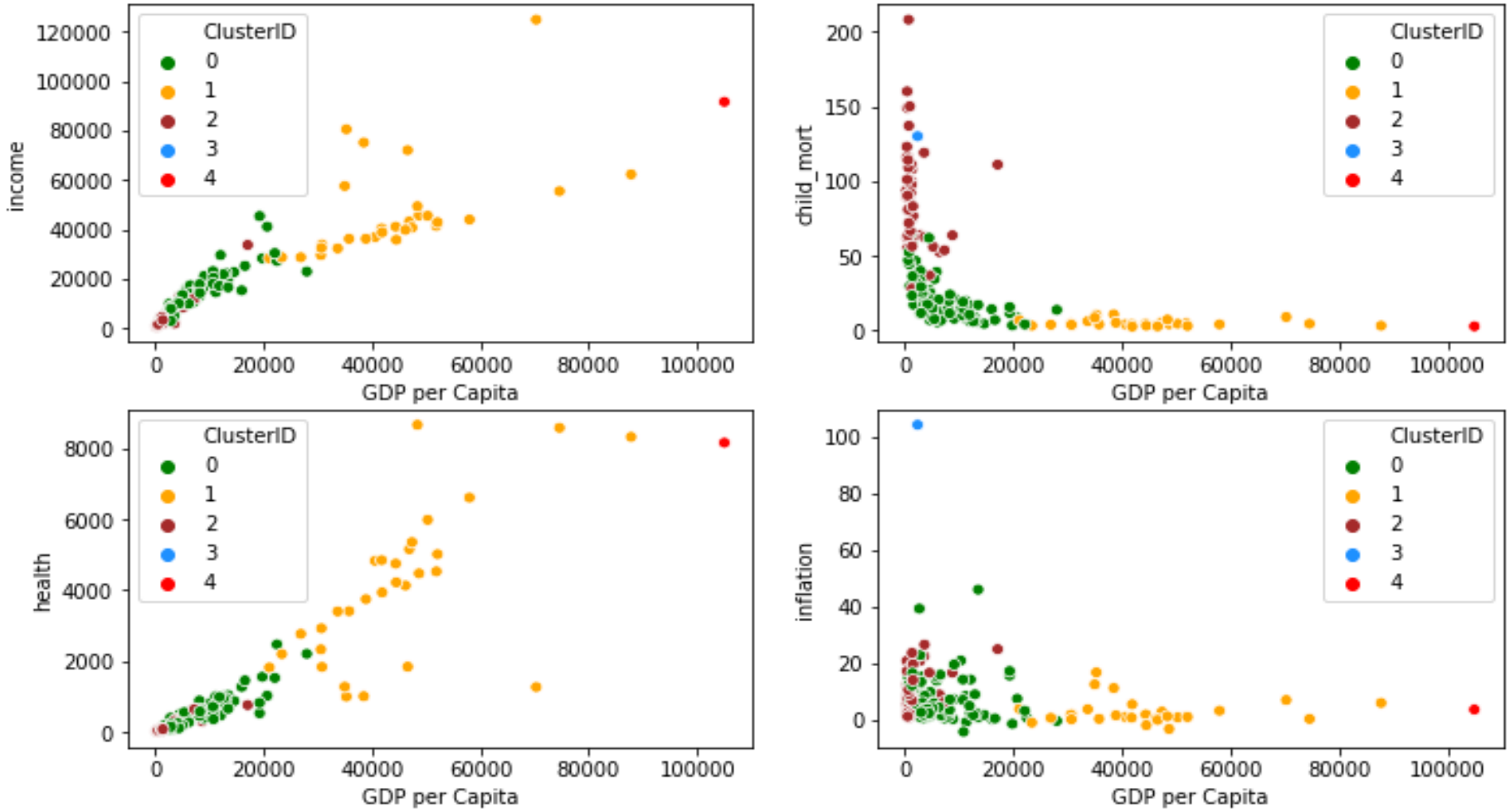
```
country_df_final.loc[country_df_final['ClusterID']==3]
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
113	Nigeria	130.0	589.49	118.131	405.42	5150	104.0	60.5	5.84	2330	3

```
country_df_final.loc[country_df_final['ClusterID']==4]
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
91	Luxembourg	2.8	183750.0	8158.5	149100.0	91700	3.62	81.3	1.63	105000	4

GDP per Capita vs Features such as Income, Child Mortality, Health indicator and Inflation.



➢ We can see that, Cluster 1 and Cluster 4 are the best performing or in other terms developed countries. These countries are having good GDP per capita as well as good health indicators as well as good income. We could find that the health indicator as well as income are kind of having linear relationship with the GDP per capita with these clusters at the higher end. We are getting a fairly intuitive idea of which clusters we need to concentrate(Clusters 0,2 and 3).

So which Clusters to look into?

- So we have got an initial intuition of which Clusters to concentrate into based on the visualisations of GDPP and other features as before. (Clusters 0,2 and 3).
- Now we will see for what stats can offer us for each clusters and take it forward.

	Cluster	child_mort_mean	exports_mean	health_mean	imports_mean	income_mean	inflation_mean	life_expec_mean	total_fer_mean	gdpp_mean
0	0	20.918182	3366.778392	482.897845	3461.435467	13455.568182	7.295045	73.222727	2.242159	7332.636364
1	1	5.006667	23900.726667	4010.316333	20228.370000	46676.666667	2.741567	80.480000	1.791667	44103.333333
2	2	90.793617	885.224660	114.751355	835.999170	3870.702128	9.951809	59.212766	4.974043	1900.255319
3	3	130.000000	589.490000	118.131000	405.420000	5150.000000	104.000000	60.500000	5.840000	2330.000000
4	4	2.800000	183750.000000	8158.500000	149100.000000	91700.000000	3.620000	81.300000	1.630000	105000.000000

- For Cluster 3 which contains only Nigeria, we have seen that Child Mortality rate is very high. The GDP per capita as well as income are very low for the same.
- Now out of the rest 2 clusters i.e. Cluster 0 and Cluster 2, we can analyse the mean for various features. Child mortality rate is pretty high for the countries in Cluster 2. The health indicator is at a pretty low value for this Cluster countries. At the same time GDP per capita and Income are at low side compared to other Clusters.
- We can hence add Nigeria to our list of countries and further drill down Cluster 2 for the worst performing countries in the Cluster.

Cluster 2 WORST performing rankings - GDPP, Child Mortality and Income as features

Based on GDPP

	country	gdpp	child_mort	income
132	Sierra Leone	399	160.0	1220
112	Niger	348	123.0	814
37	Congo, Dem. Rep.	334	116.0	609
88	Liberia	327	89.3	700
26	Burundi	231	93.6	764

Based on Income

	country	gdpp	child_mort	income
31	Central African Republic	446	149.0	888
112	Niger	348	123.0	814
26	Burundi	231	93.6	764
88	Liberia	327	89.3	700
37	Congo, Dem. Rep.	334	116.0	609

Based on Child Mortality

	country	gdpp	child_mort	income
97	Mali	708	137.0	1870
31	Central African Republic	446	149.0	888
32	Chad	897	150.0	1930
132	Sierra Leone	399	160.0	1220
66	Haiti	662	208.0	1500

- We can see that if we sort the cluster 2 countries based on the GDPP, we can see the 5 economically backward countries based on GDPP are the countries - `**Sierra Leone, Niger, Democratic Republic of Congo, Liberia, Burundi**`. GDPP of these countries are way less than the average GDPP of this cluster itself as well as the income is also way less than the mean par of `3870` for this cluster.
- We now sorted the countries based on the income and the lowest that can be ranked on the income are `**Central African Republic, Niger, Burundi, Liberia, Democratic Republic of Congo**`, out of which 4 countries are the one listed with lowest GDPP as well.
- Now we will see countries which come under the highest child mortality ratio. Even though these countries are having comparatively better income, their GDPP is pretty low as well as child mortality is very high. Countries here are `**Mali, Central African Republic, Chad, Sierra Leone, Haiti**`

Cluster 0 WORST performing rankings - GDPP, Child Mortality and Income as features

Why to go for Cluster 0 as well?

➢ Since from all the visual plots for the data points till now, we saw that the countries are closely cluttered. We can analyse the worst performers for Cluster 2 as well, inorder to cement our selection of countries and to find out any other countries that might need an aid, in case we have missed out.

Based on GDPP

	country	gdpp	child_mort	income
83	Kyrgyz Republic	880	29.6	2790
27	Cambodia	786	44.4	2520
12	Bangladesh	758	49.4	2440
146	Tajikistan	738	52.4	2110
109	Nepal	592	47.0	1990

Based on Income

	country	gdpp	child_mort	income
83	Kyrgyz Republic	880	29.6	2790
27	Cambodia	786	44.4	2520
12	Bangladesh	758	49.4	2440
146	Tajikistan	738	52.4	2110
109	Nepal	592	47.0	1990

Based on Child Mortality

	country	gdpp	child_mort	income
12	Bangladesh	758	49.4	2440
146	Tajikistan	738	52.4	2110
69	India	1350	58.8	4410
154	Turkmenistan	4440	62.0	9940
107	Myanmar	988	64.4	3720

➢ If you compare these above countries of Cluster 0 with Cluster2, all countries here are having less child mortality rate compared to cluster 2 countries which we identified earlier. Income is comparatively little better as well, while GDPP is less, in which we can see `**Nepal**` being one of the country with both GDPP and income as very less. In fact, Nepal can be seen values lesser than even Chad or Mali of Cluster 2 .

➢ The feature of Child mortality ratio is a way far below par compared to countries of Cluster 2 which is a good indication that these countries are better performers.

➢ **Tajikistan** is other country which we can add in our list of countries because of its lower income slab.

FINAL LIST OF COUNTRIES and Supporting Statements

❑ List of countries hence selected for the aid are the below listed Countries –

- NIGERIA
-SIERRA LEONE
-NIGER
-DEMOCRATIC REPUBLIC OF CONGO
-LIBERIA
-BURUNDI
-CENTRAL AFRICAN REPUBLIC
-MALI
-CHAD
-HAITI
-NEPAL
-TAJIKISTAN

❑ Points validating the selection -

- Out of all the listed countries we selected, except countries of Nigeria, Tajikistan and Haiti, all countries are the ones listed in LDC(Least Developed Countries) which are the countries according to the United Nations, that exhibit the lowest indicators of socio-economic development, with the lowest Human Development Index ratings of all countries in the world.
- Even though Nigeria is not in the list of the LDC, and is more developed than the majority of African countries, GDP of the country is too low to make it to the list of the developed ones safely, and the industrialization is very far behind most of the countries that made it to this rating. The major problems in Nigeria are poor health care, infant mortality, corruption, and high illiteracy rates, among many others.
- Coming to Haiti, hobbled by foreign interventions, political instability, and natural disasters, the former French colony has long suffered from underdevelopment. Haiti is the poorest country in the Western Hemisphere.
- While for Tajikistan, Deficiencies in the legal framework and the judicial system, as well as weak public administration and an undeveloped financial sector, as made it one of the poorest countries.

THANK YOU