HELP International NGO Fund using PCA

Submitted by: Shadab Hussain

About:

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 project that included a lot of awareness drives and 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.

Objective:

Group the countries given in the data based on different features so that we can identify the development related to each country and identify those countries which needs to be focused most so that funds can be granted to them for their development.

Steps to be followed:

- Explore the data, check for data accuracy and take necessary steps for data processing
- Check for the correlation between the features if any and normalize the data to bring them on same scale
- Using PCA reduce data dimension and remove multicollinearity
- Cluster data on feature created by PCA using k-Means and Hierarchical
- Identify the cluster which needs the focus of NGO
- List down those identified countries

1. Explore the Data

We have 10 variables and 167 data points in the **Country-data.csv** file.

List of features:

```
['country', 'child_mort', 'exports', 'health', 'imports', 'income', 'inflation', 'life_expec', 'total_fer', 'gdpp']
```

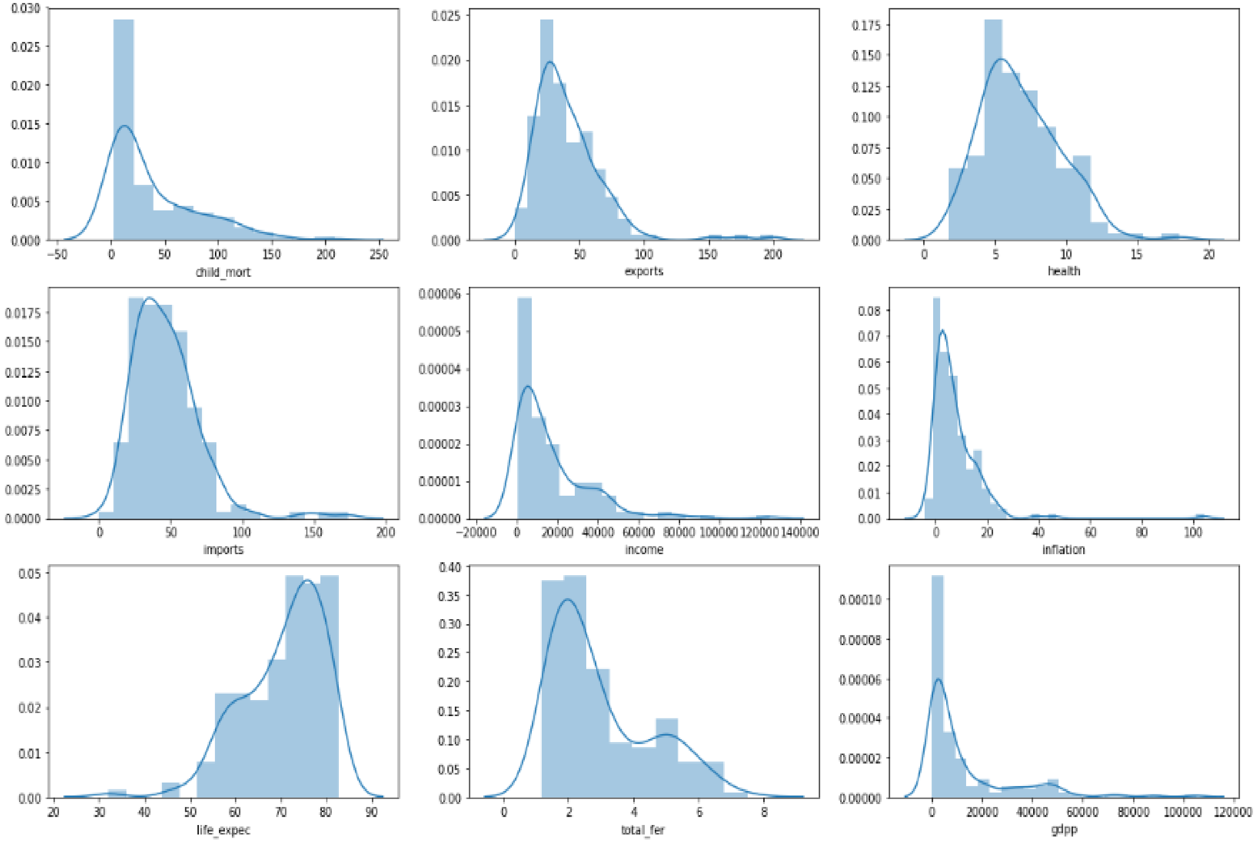
Data Description:

(There are no null values in the given data)

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.

1.1 Checking for Data Distribution

Data Distribution for each of the given features:



Looking for data patterns:

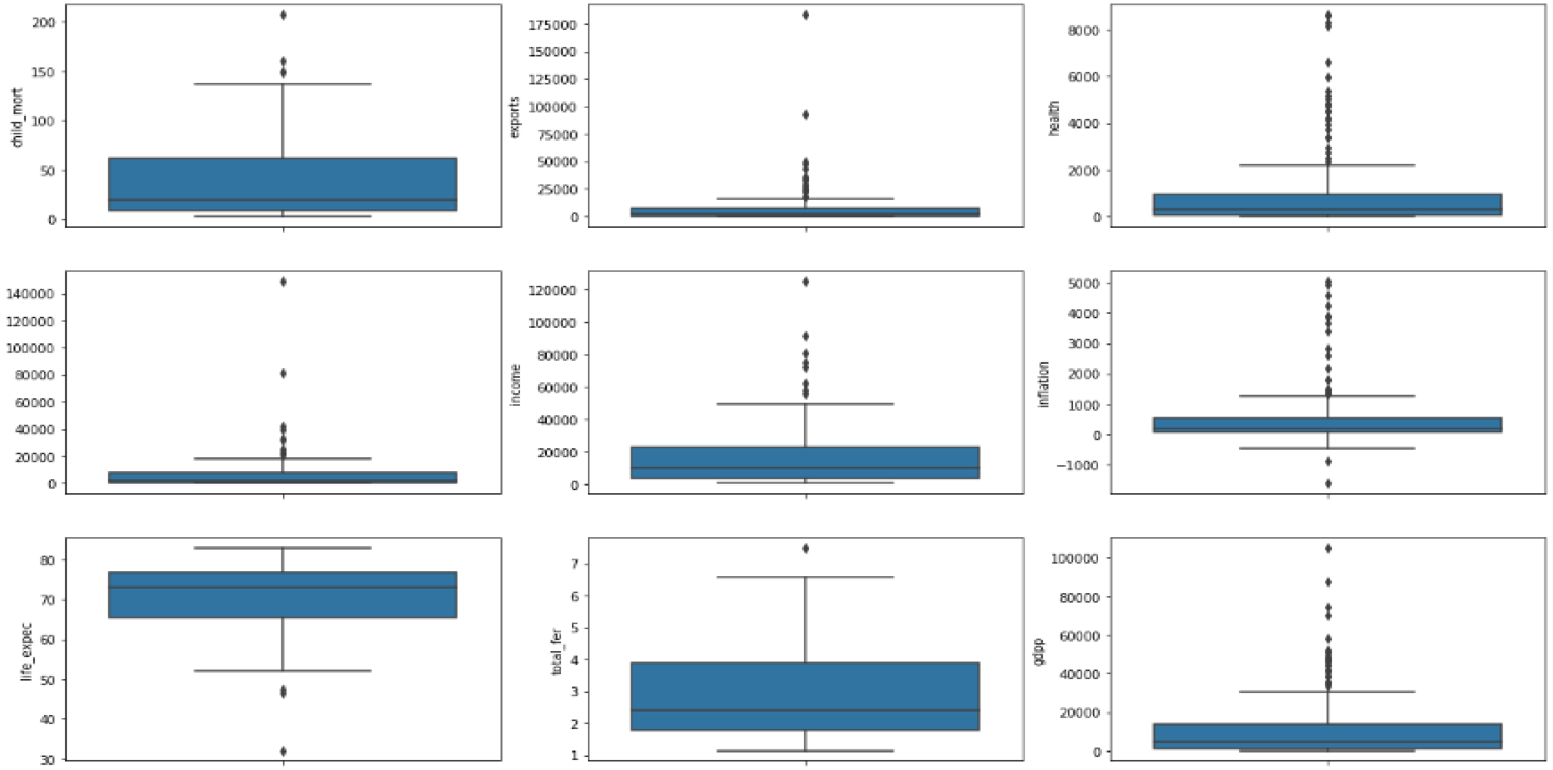
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
std	40.328931	27.412010	2.746837	24.209589	19278.067698	10.570704	8.893172	1.513848	18328.704809
min	2.600000	0.109000	1.810000	0.065900	609.000000	-4.210000	32.100000	1.150000	231.000000
25%	8.250000	23.800000	4.920000	30.200000	3355.000000	1.810000	65.300000	1.795000	1330.000000
50%	19.300000	35.000000	6.320000	43.300000	9960.000000	5.390000	73.100000	2.410000	4660.000000
75%	62.100000	51.350000	8.600000	58.750000	22800.000000	10.750000	76.800000	3.880000	14050.000000
90%	100.220000	70.800000	10.940000	75.420000	41220.000000	16.640000	80.400000	5.322000	41840.000000
95%	116.000000	80.570000	11.570000	81.140000	48290.000000	20.870000	81.400000	5.861000	48610.000000
99%	153.400000	160.480000	13.474000	146.080000	84374.000000	41.478000	82.370000	6.563600	79088.000000
max	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

1.2 Data Accuracy, Conversion and Outlier Detection

Converting columns like exports, health, imports and inflation in terms of their absolute value

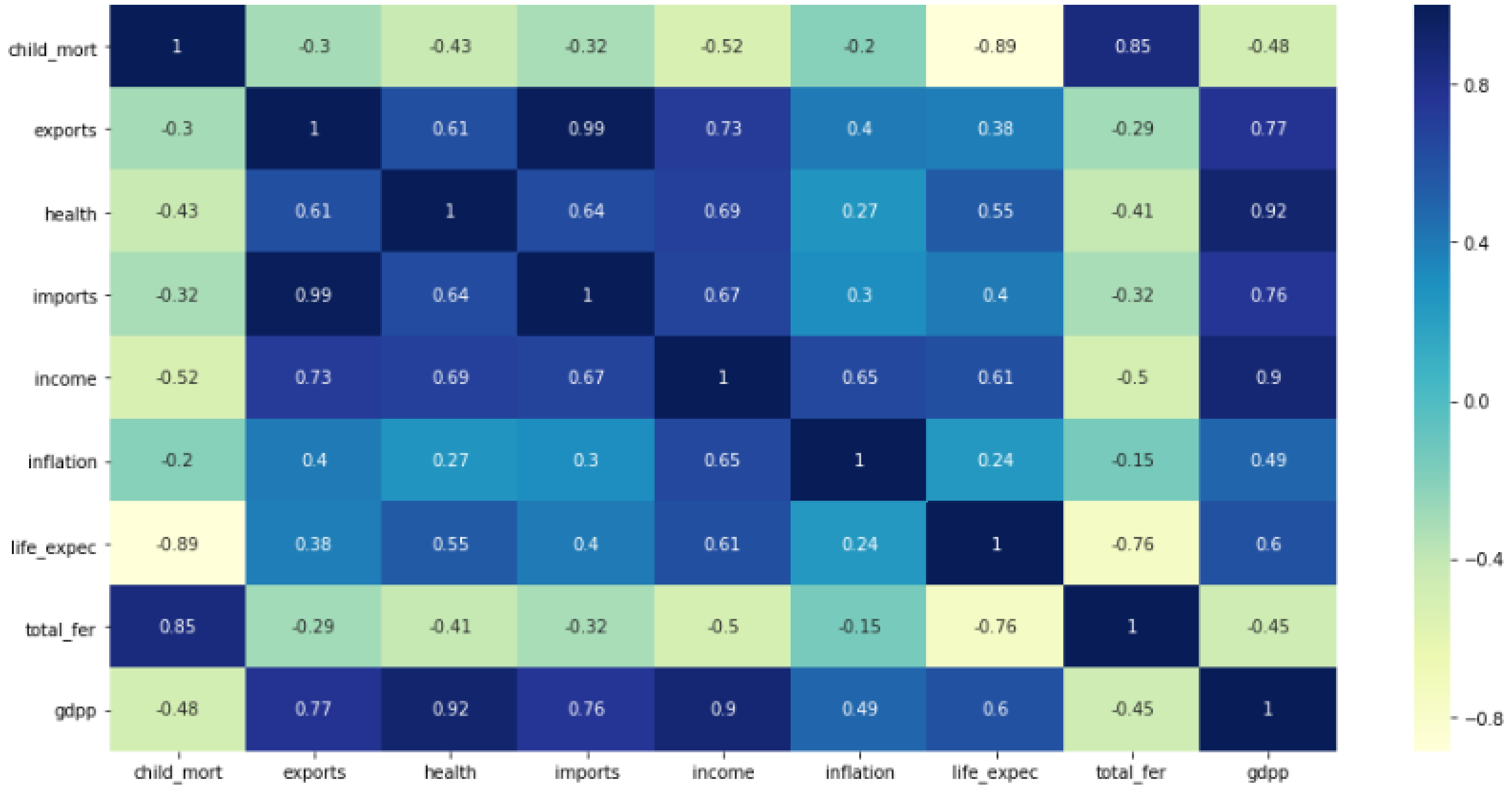
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	55.30	41.9174	248.297	1610	47.700292	56.2	5.82	553
1	Albania	16.6	1145.20	267.8950	1987.740	9930	175.749833	76.3	1.65	4090
2	Algeria	27.3	1712.64	185.9820	1400.440	12900	618.484065	76.5	2.89	4460
3	Angola	119.0	2199.19	100.6050	1514.370	5900	646.013072	60.1	6.16	3530
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100	173.186120	76.8	2.13	12200

Detecting outliers:



2. Correlation and Normalization

Correlation between the features:
Visualizing feature correlation using Heatmap.
We would have to reduce/remove this collinearity using PCA.

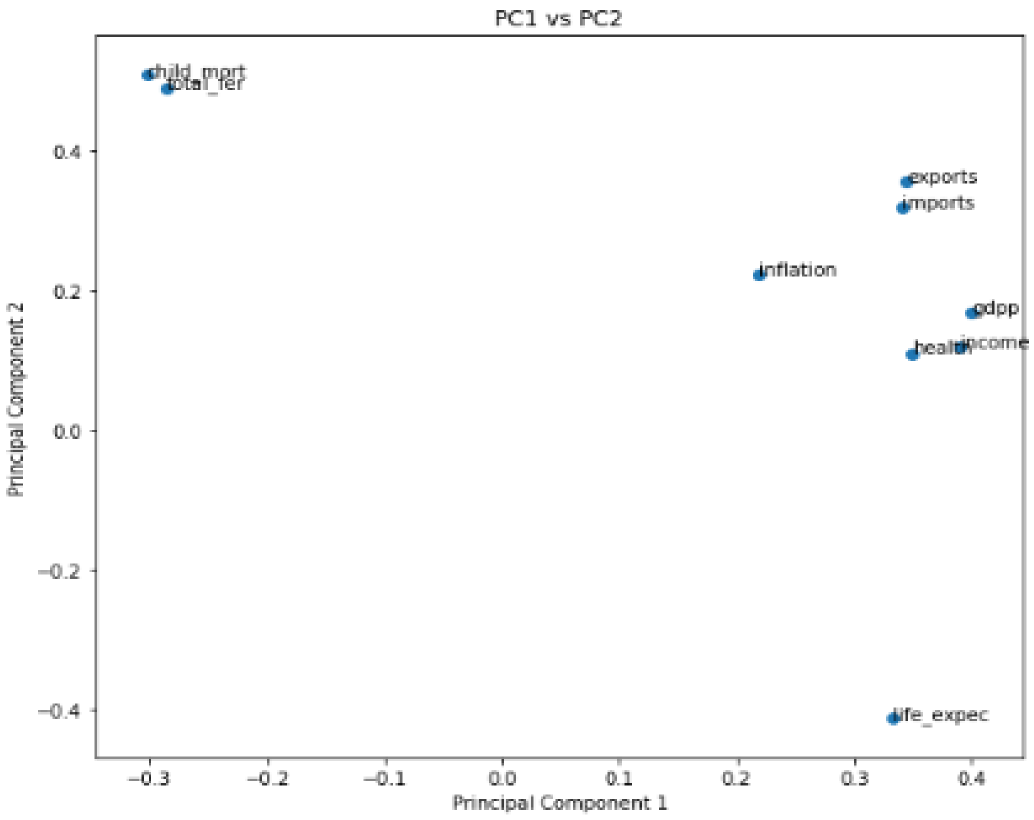


Normalizing the data to bring all the features on same scale

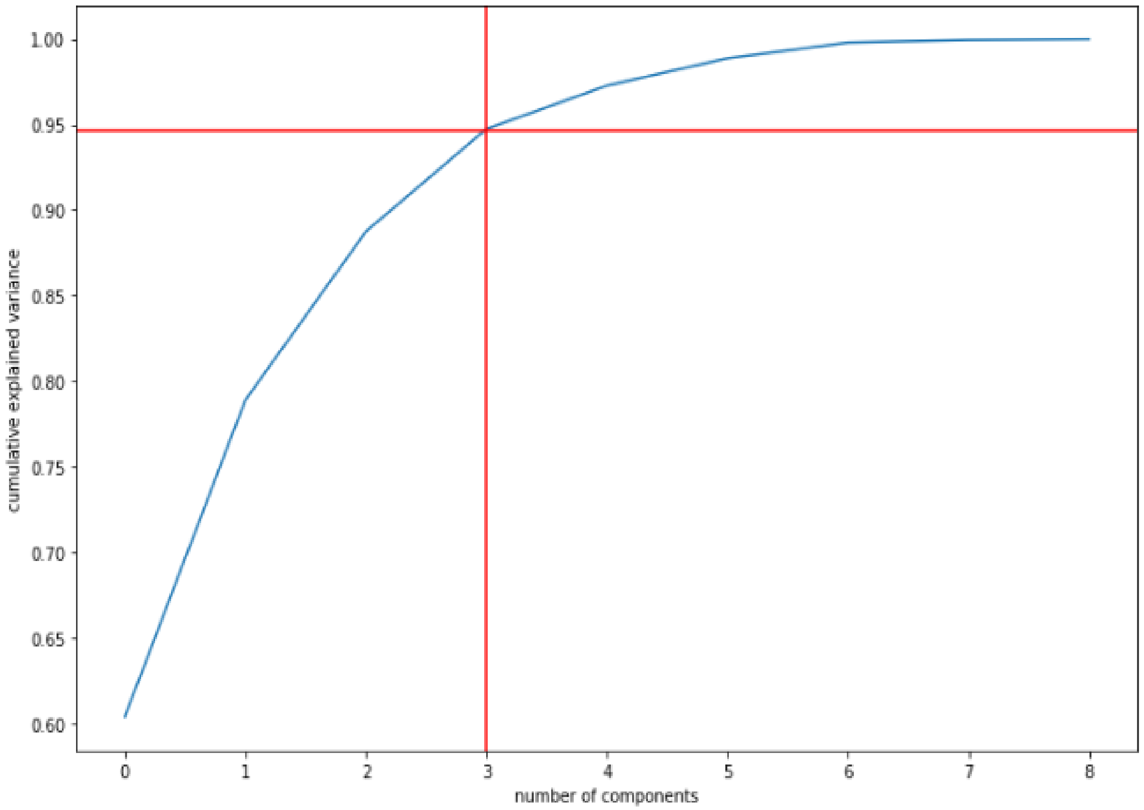
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	1.291532	-0.411011	-0.565040	-0.432276	-0.808245	-0.470611	-1.619092	1.902882	-0.679180
1	-0.538949	-0.350191	-0.439218	-0.313677	-0.375369	-0.340132	0.647866	-0.859973	-0.485623
2	-0.272833	-0.318526	-0.484826	-0.353720	-0.220844	0.111006	0.670423	-0.038404	-0.465376
3	2.007808	-0.291375	-0.532363	-0.345953	-0.585043	0.139058	-1.179234	2.128151	-0.516268
4	-0.695634	-0.104331	-0.178771	0.040735	0.101732	-0.342744	0.704258	-0.541946	-0.041817

3. Principal Component Analysis:

Visualizing features using first two components, i.e., PC1 and PC2
From here we can observe, we can either create 3 or 5 clusters.

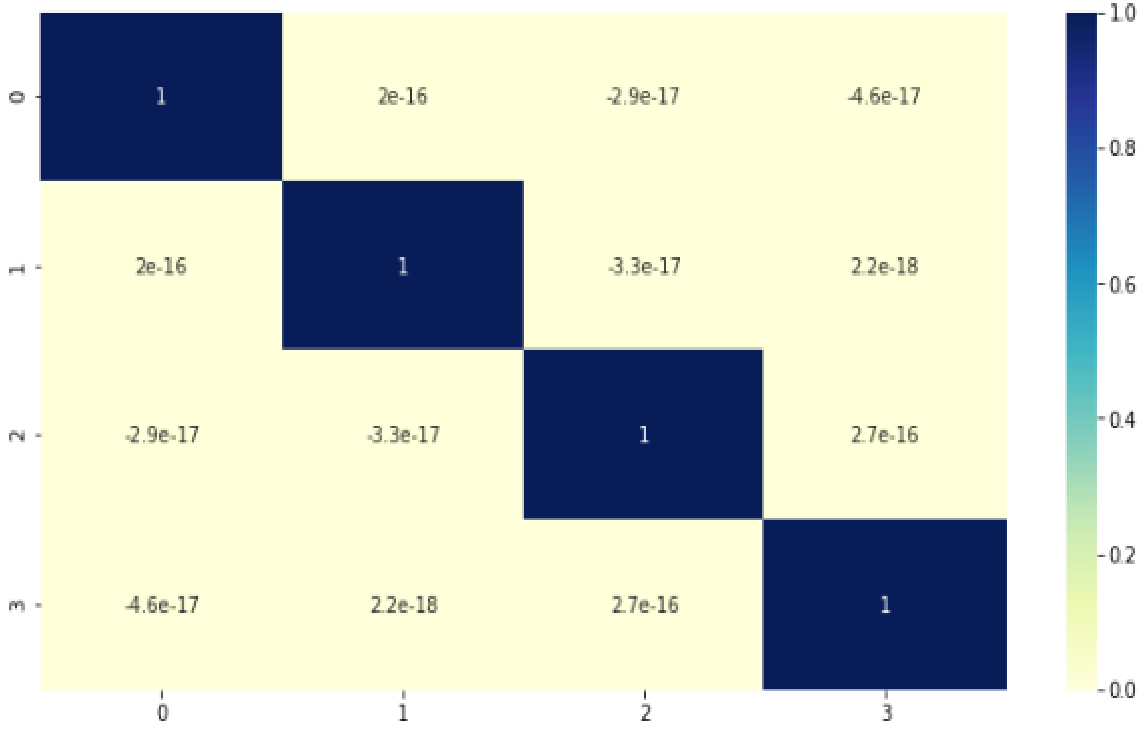


Approx. 94%+ data variance can be covered using 4 principal components

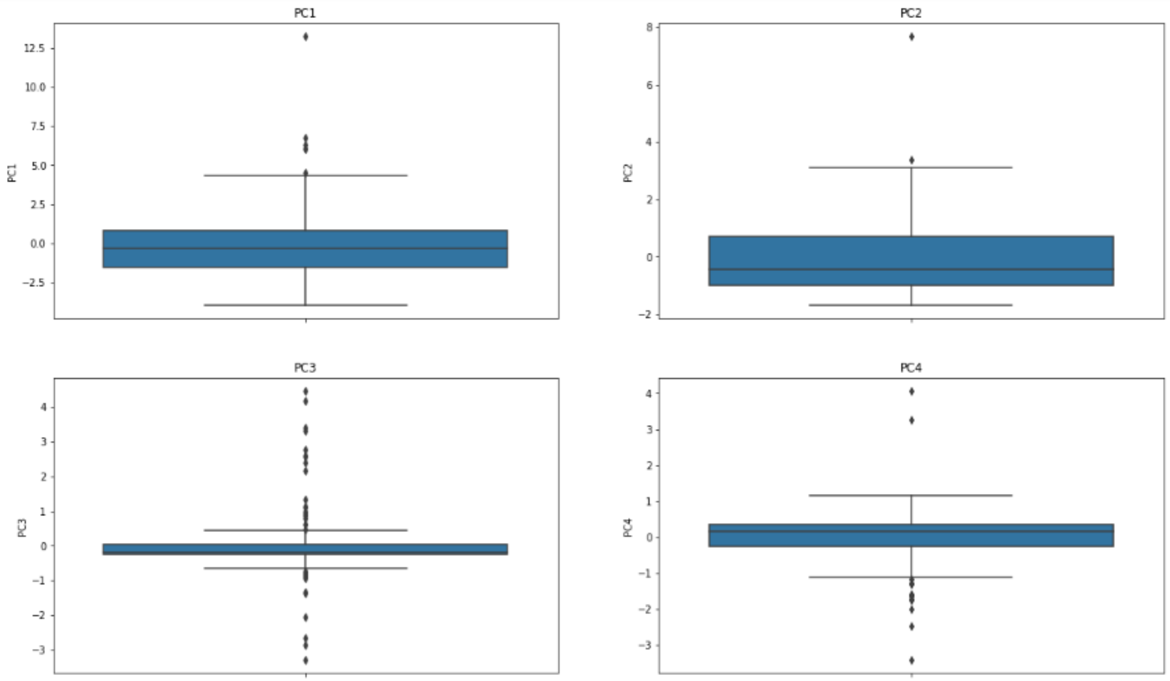


Transformed data and Outlier Detection

Transformed the original data, and after PCA, data is of just 4-dimensions.
max corr: 2.706341341108554e-16
min corr: -4.606488068144064e-17
Below is the Heatmap of features created using PCA:



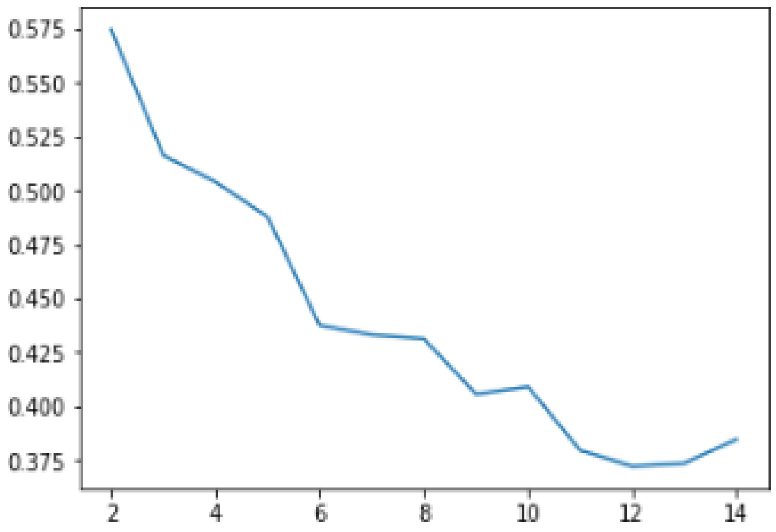
Detecting outliers after data transformation
We can observe we have still few outliers in our data, which we will discard for now and later we will predict clusters for them after building clustering model.



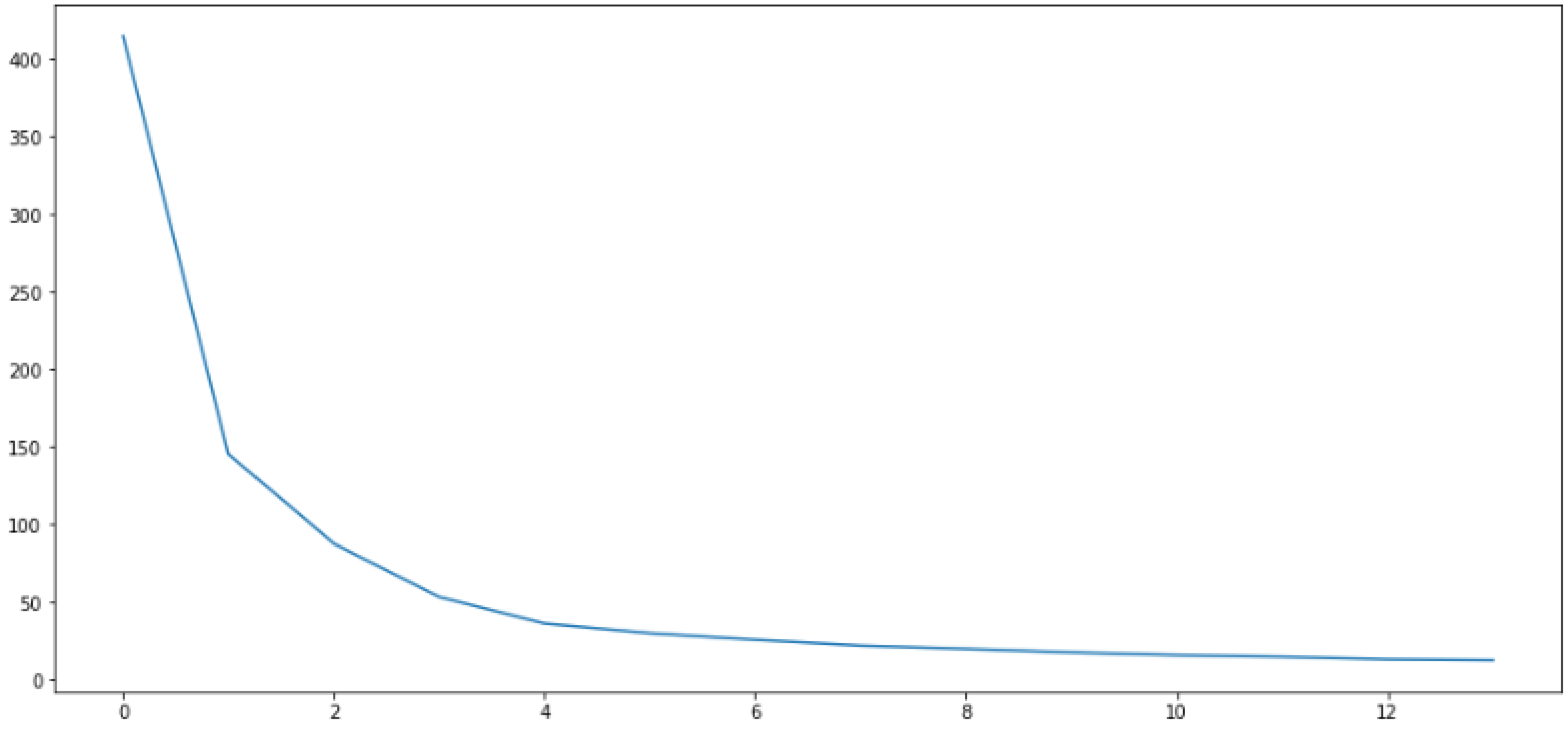
4. Clustering

We got Hopkins statistics more then 80% which indicates data is good fit for clustering.

From Silhoutte Score graph and Elbow curve we can observe best k value lies between 3 to 5



Silhoutte Score Graph



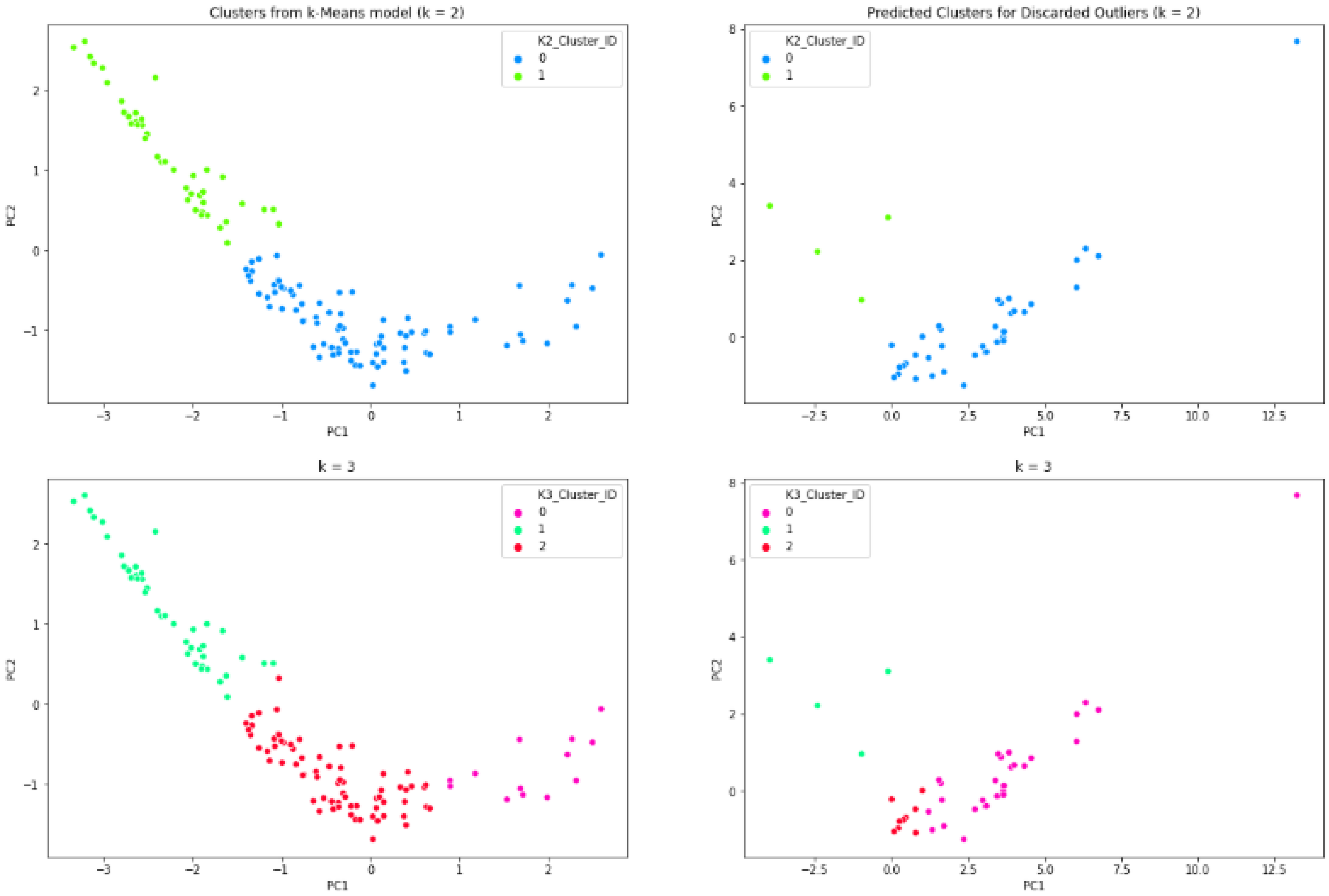
Elbow Curve

k-Means Cluster Visualization

Visualizing cluster on data after discarding outlier and data having discarded ouliers with PC1 on x-axis and PC2 on y-axis

Left graphs are of data after discarding outliers, and right graphs are of discarded outliers data.

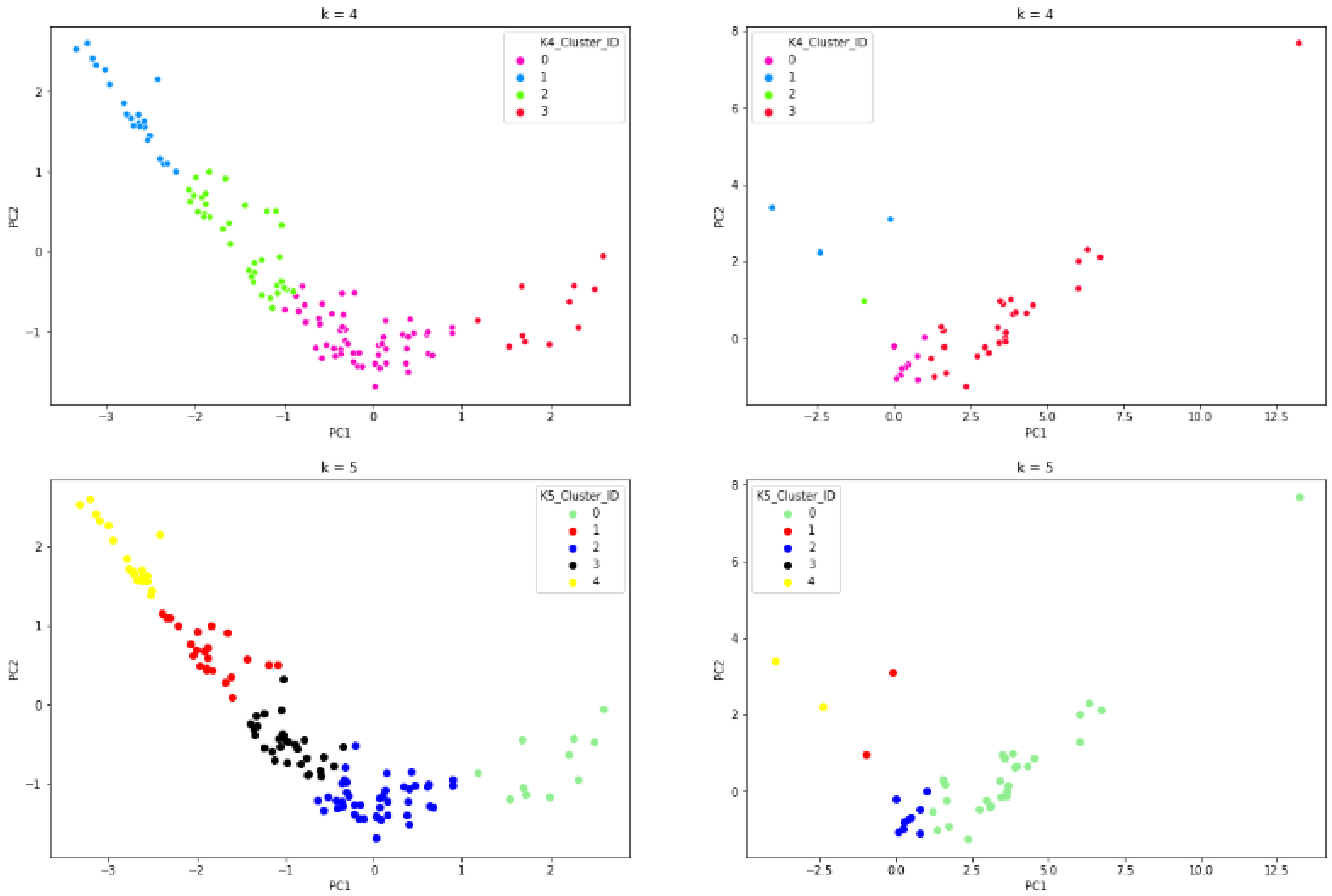
First row is for k=2
Second row is for k=3



k-Means Cluster Visualization (continued…)

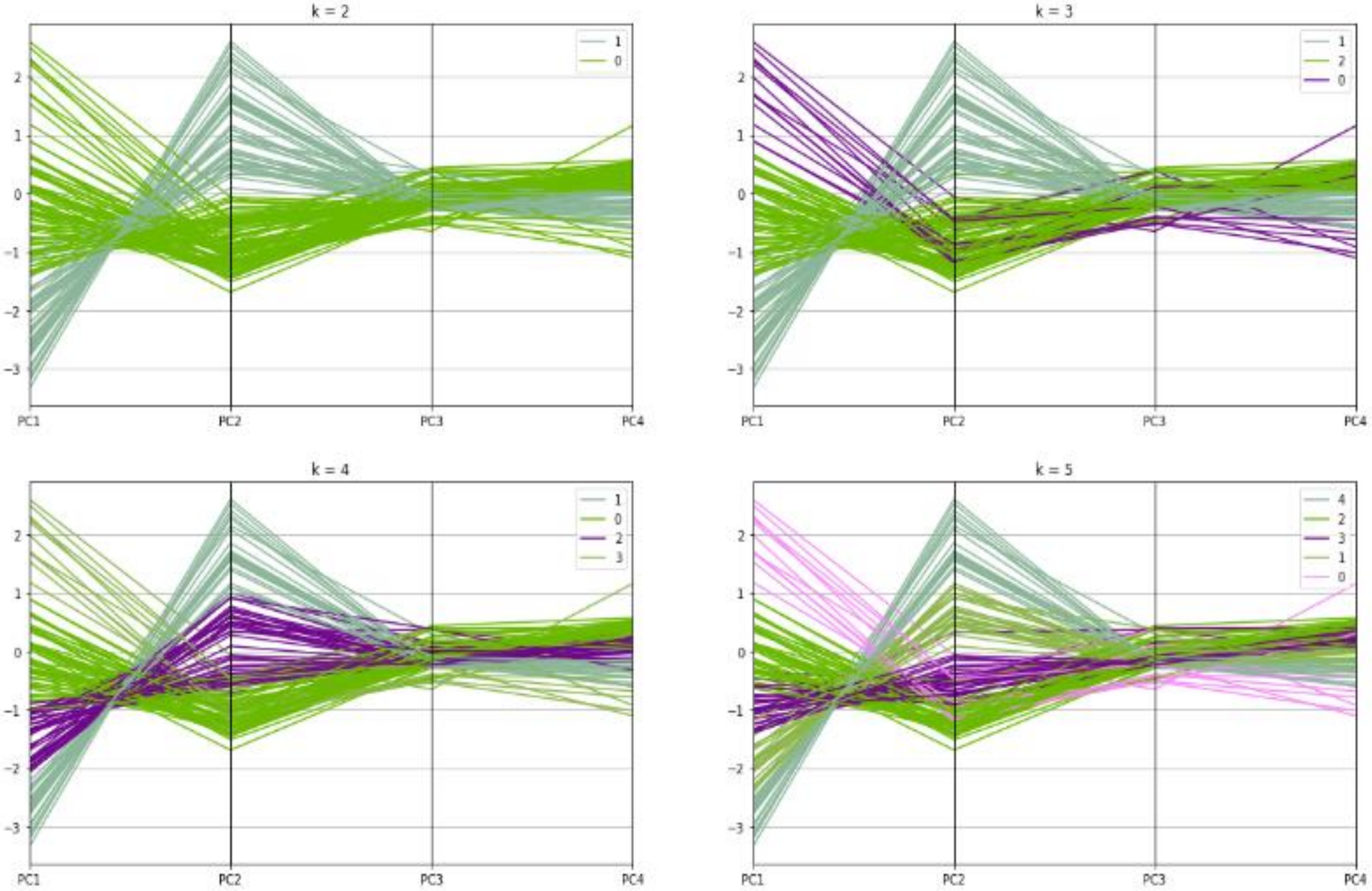
Here,
First row is for k=4
Second row is for k=5

We can observe from the plots, k=5 is best suited for analysis



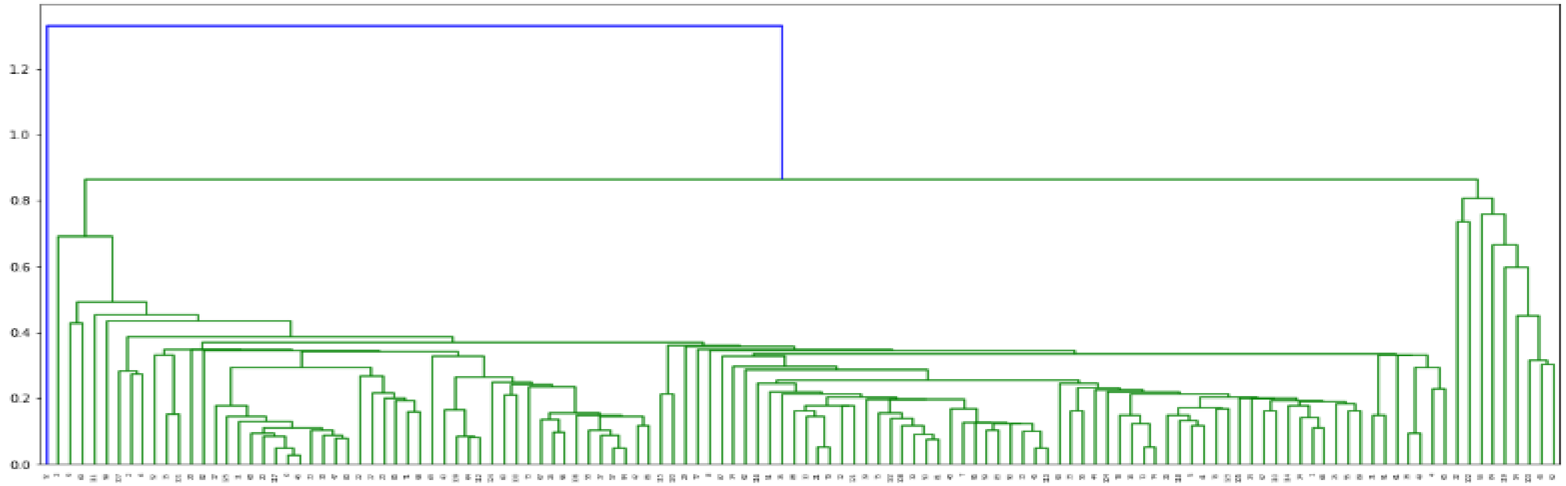
k-Means Cluster Visualization (continued…)

Parallel-Coordinates Visualization:
This gives us understanding of data points are getting clustered for different number of clusters.

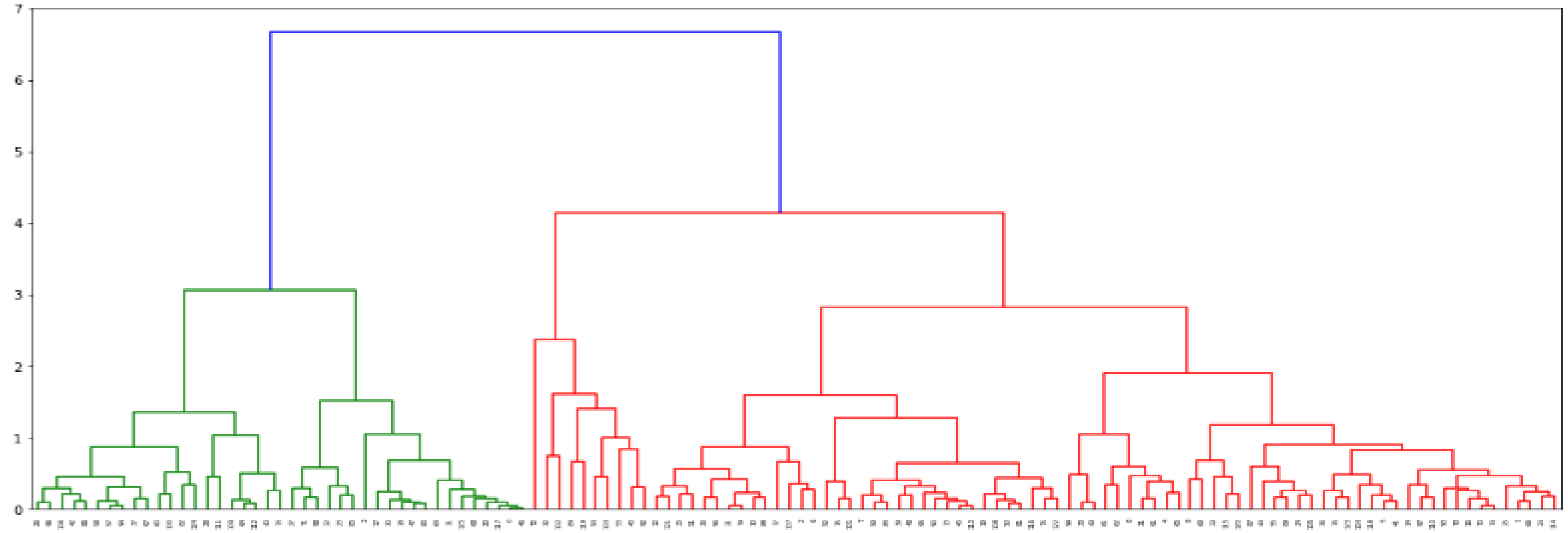


Hierarchical Cluster Visualization

This is the dendogram for method=simple

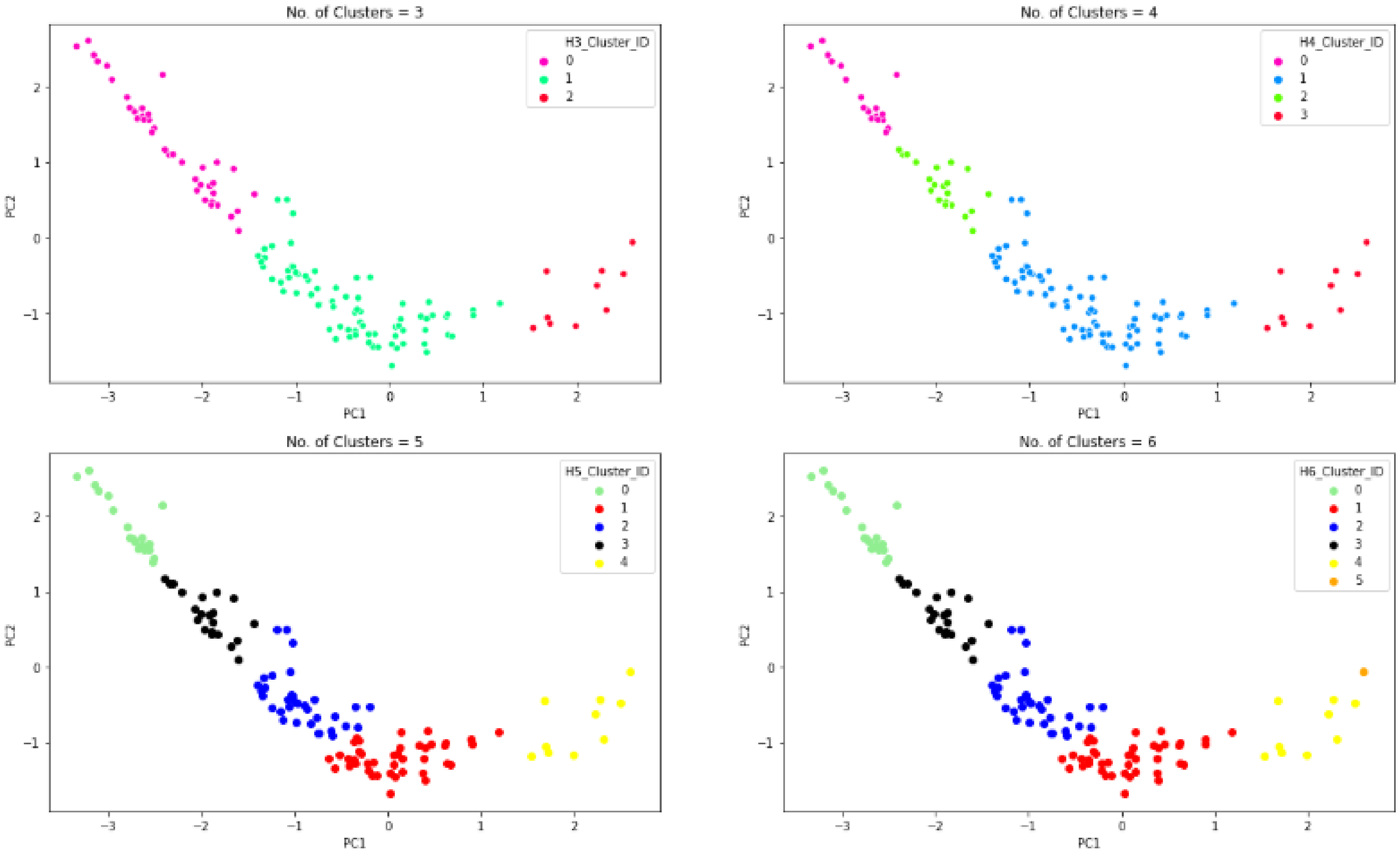


This is the dendogram for method=complete
From here we can observe, 5 clusters are best to group the given data



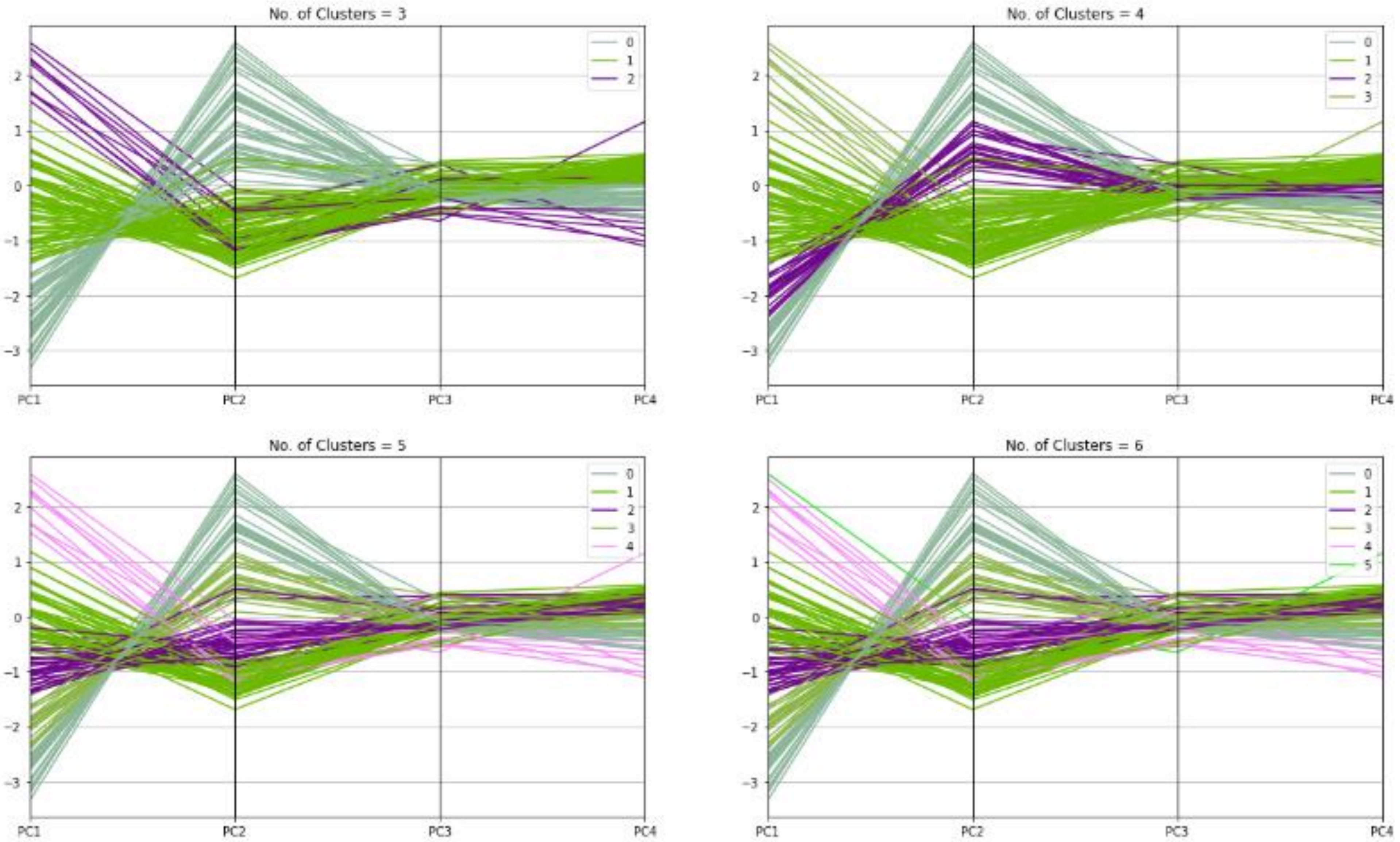
Hierarchical Cluster Visualization (continued…)

Visualization of hierarchical clustering model with PC1 and PC2 with different number of clusters on data after discarding outliers



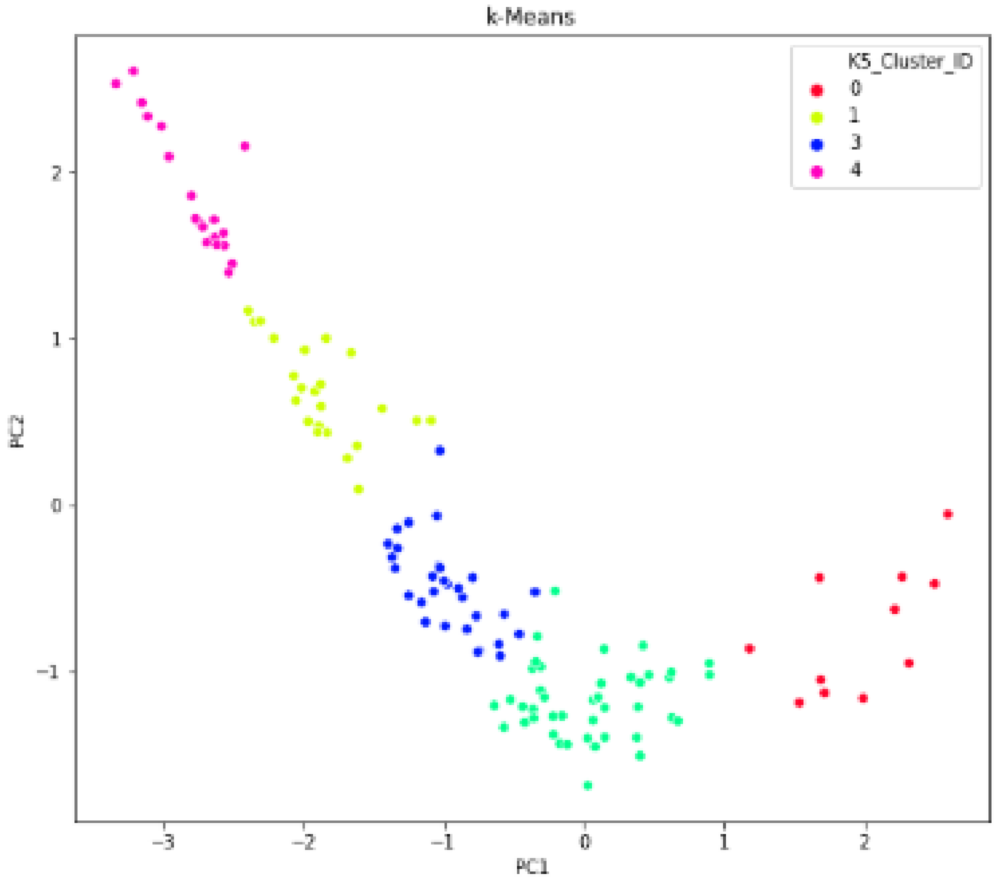
Hierarchical Cluster Visualization (continued…)

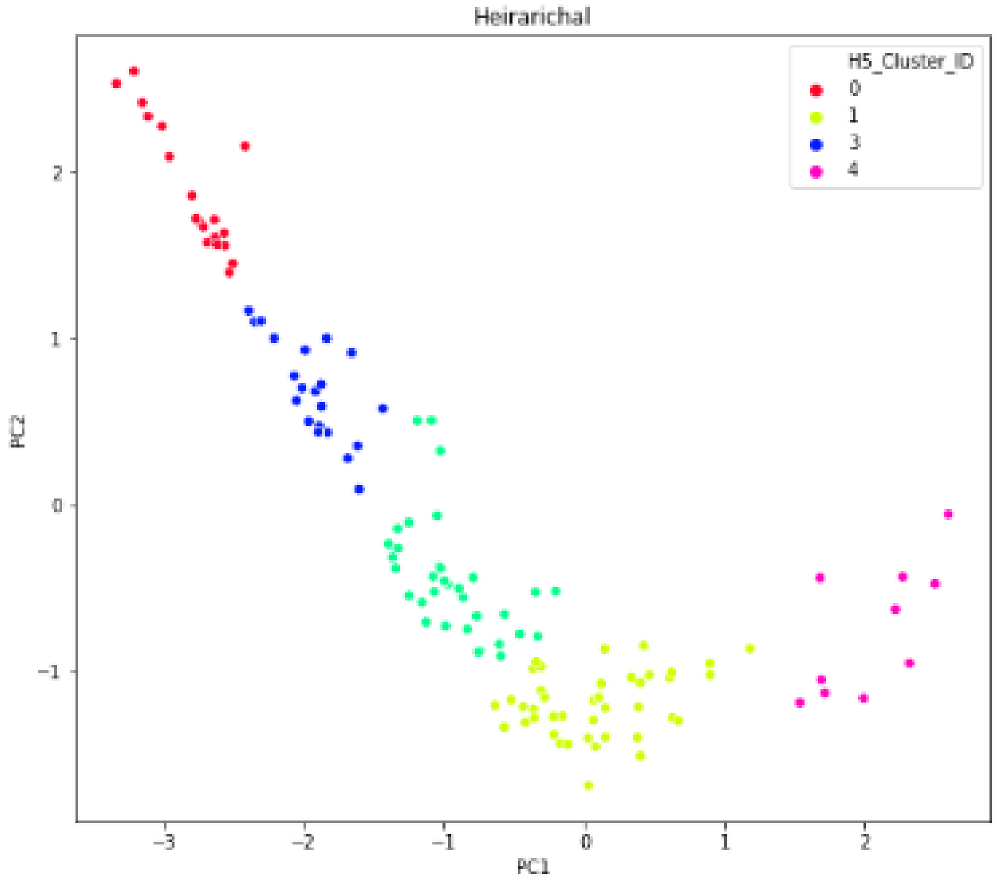
Parallel-Coordinates Visualization:
This gives us understanding of data points are getting clustered for different number of clusters.



Cluster Visualization (continued…)

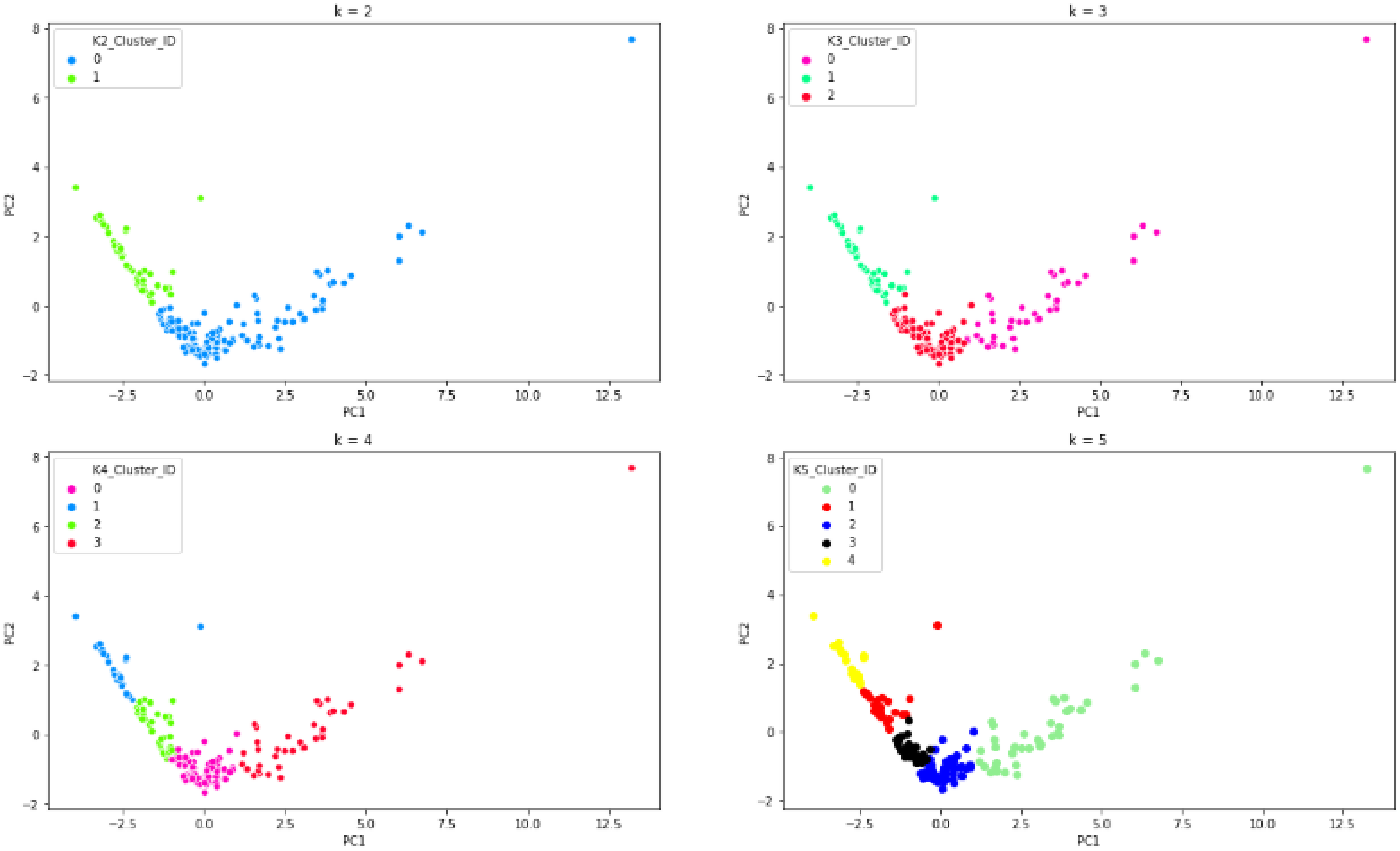
Visualization of clusters for both k-means and hierarchical, both looks almost similar





Cluster Visualization (continued…)

Visualization of k-means clustering model with PC1 and PC2 with different number of clusters on total data including discarded outliers



Cluster Visualization (continued…)

We picked k-means for k=5 and merged the data with features, plotted bar-plots of mean of these variables on y-axis and cluster on x-axis after grouping them by cluster id

From these plots we can conclude:

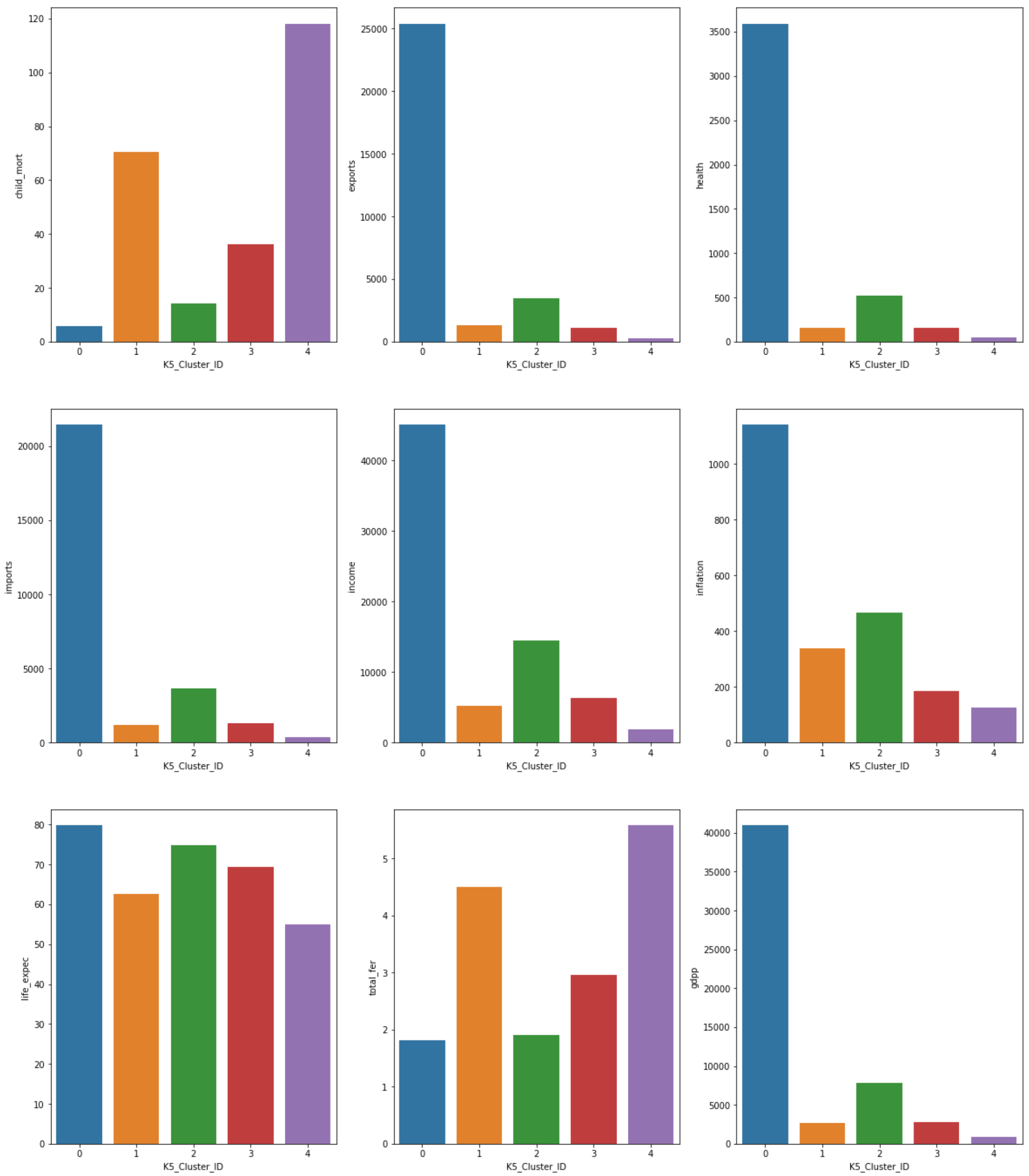
•Child mortality is very high in **Cluster 4** in comparison to other clusters.

•Exports, Health, Income, Imports and, Inflation are very high in **Cluster 0** in comparison to other clusters as well in **Cluster 4** its very low.

•Life expectancy is almost good in all clusters but little less in **Cluster 4.**

•Total Fertility Rate is very high in **Cluster 5**

•GDPP is significantly very low in **Cluster 4** and very high in **Cluster 0**.



Now, let's see how principal components are associated with original features and clusters:

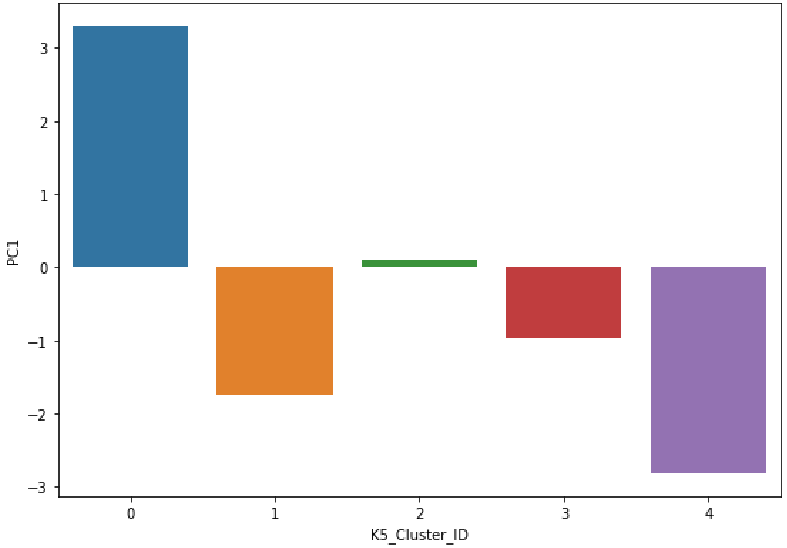
From this we can see:

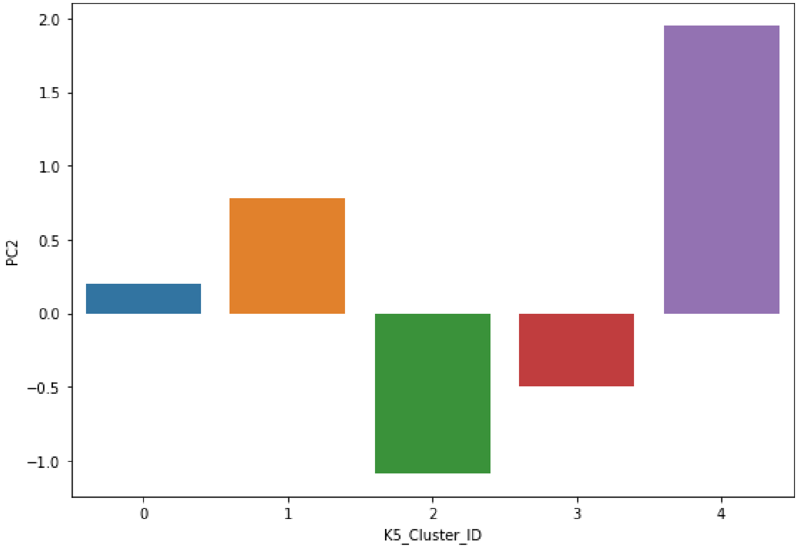
1.PC1 is negatively associated with child_mort and the same we can observe from PC1 plot. It is highly correlated with Cluster 4.

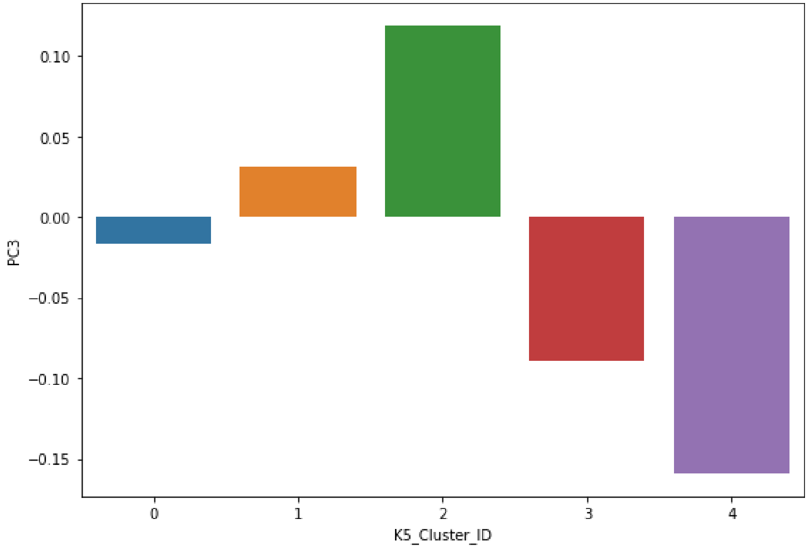
2.Life Expectancy is very low in Cluster 2 and very high in Cluster 4.

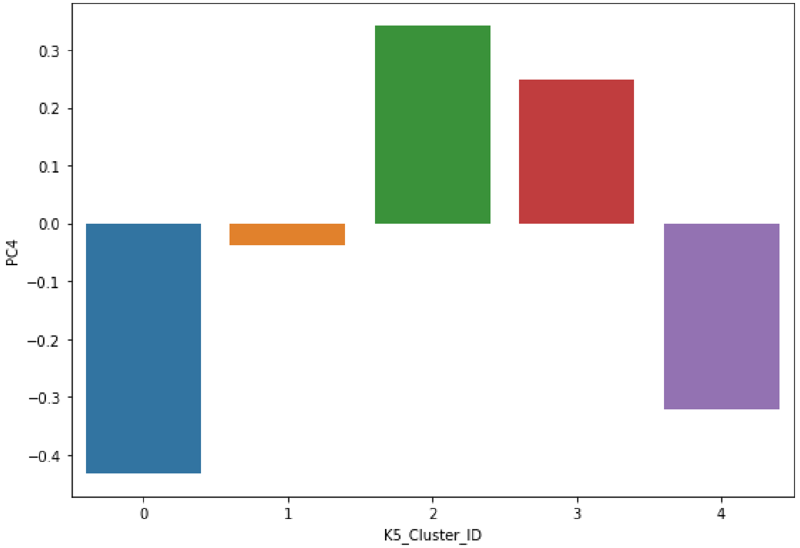
3.PC3 is highly associated with Inflation Rate , which shows the chnage in GDP from last year to current year. We can observe in Cluster 4 it is negatively associated.

4.PC4 is highly negatively associated with health, then imports and then gdpp, effect of which we can observe in Cluster 4 and Cluster 0.









5. Conclusion:

We can conclude, `Cluster 4` is the group of countries which are having high child mortality and low gdpp.
Child_mort, gdpp, health, inflation are major features affecting clustering.
Therefore, these countries should be funded.

Below is the list of those countries:
'Afghanistan', 'Angola', 'Benin', 'Burkina Faso', 'Burundi', 'Cameroon', 'Central African Republic', 'Chad', 'Congo, Dem. Rep.', "Cote d'Ivoire", 'Guinea', 'Guinea-Bissau', 'Haiti', 'Lesotho', 'Malawi', 'Mali', 'Mozambique', 'Niger', 'Nigeria', 'Sierra Leone', 'Uganda', 'Zambia'

Thank You ☺