Clustering & PCA Assignment

Submitted by: Vivek Raj

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.

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

OBJECTIVES

1. Perform PCA on the dataset and obtain the new dataset with the Principal Components. Choose the appropriate number of components k. You need to perform your clustering activity on this new dataset, i.e. the PCA modified dataset with the k components.

2. **Outlier Analysis**: You must perform the Outlier Analysis on the dataset, before or after performing PCA, as per your choice. However, you do have the flexibility of not removing the outliers if it suits the business needs or a lot of countries are getting removed. Hence, all you need to do is find the outliers in the dataset, and then choose whether to keep them or remove them depending on the result you get.

3. Try both K-means and Hierarchical clustering(both single and complete linkage) on this dataset to create the clusters. Analyse the clusters and identify the ones which are in dire need of aid.

4. Also, you need to perform visualisations on the clusters that have been formed. You should also do the same visualisation using any two of the original variables (like gdpp, child_mort, etc.) on the X-Y axes as well. You can also choose other types of plots like boxplots, etc.

5. The final list of countries depends on the number of components that you choose and the number of clusters that you finally form.

Importing the dataset and Pandas Data Frame

```
# Importing the dataset as pandas dataframe. We will observe first 5 rows of the dataframe to get a glimpse of dataset.
df = pd.read_csv("Country-data.csv")
df.head()
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

1. The data frame has 167 rows and 10 columns. Converted “Export”, “Health” & “Imports” column to absolute values.
2. The Dataset has no null/NAN values in any of the columns hence null value imputation or removal is not required.
3. However, there are outliers in the dataset as the standard deviation is very high and there is huge gap in median, 75th percentile and max value.

Findings of Exploratory Data Analysis of Data Frame

1. High Child Mortality rate prevails in countries with very low gdpp, low income and low health, it has negative correlation with life expectancy, positive correlation with total fertility, and not related with the inflation.

2. Export, health, Imports, Income are highly positive correlated with GDPP as usual.

3. There is a huge gap in exports, imports, health in developed vs under-developed countries.

4. Life expectancy is high in most of the countries.

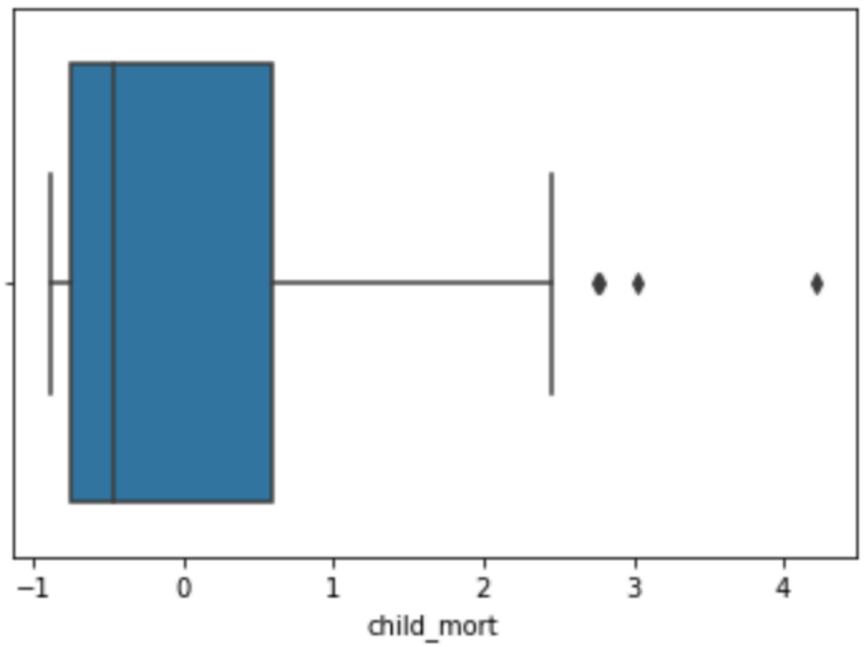
Rescaling of the data frame using Standard Scaler

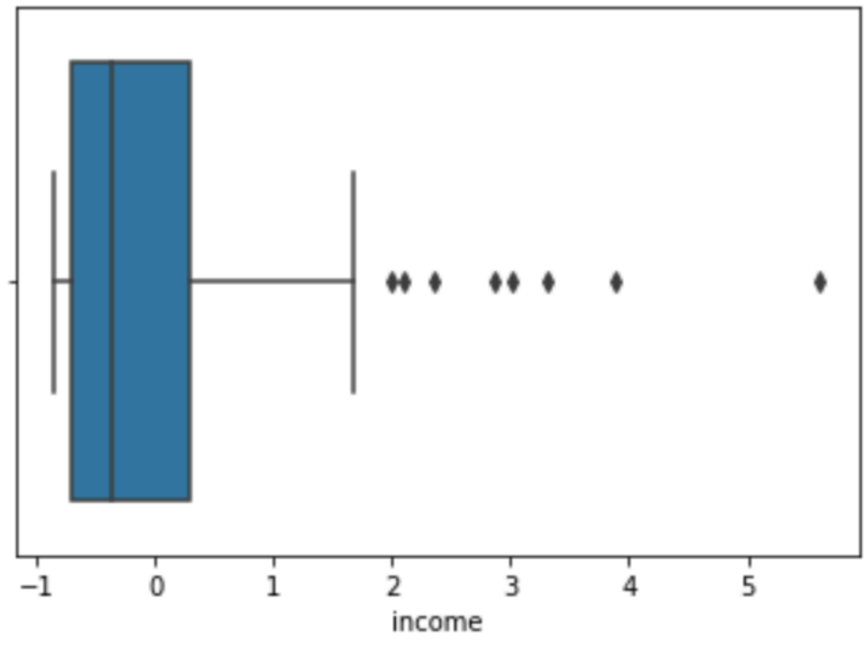
```
scaler = StandardScaler()
scaled = scaler.fit_transform(num_df)
```

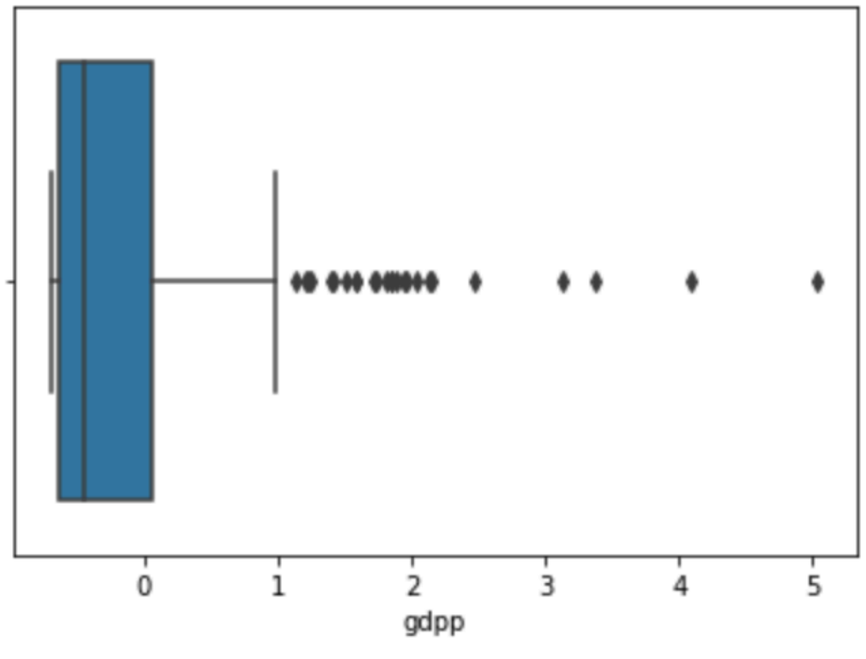
```
scaled_num_df = pd.DataFrame(scaled)
scaled_num_df.columns = num_df.columns
scaled_num_df.head()
```

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	1.291532	-0.411011	-0.565040	-0.432276	-0.808245	0.157336	-1.619092	1.902882	-0.679180
1	-0.538949	-0.350191	-0.439218	-0.313677	-0.375369	-0.312347	0.647866	-0.859973	-0.485623
2	-0.272833	-0.318526	-0.484826	-0.353720	-0.220844	0.789274	0.670423	-0.038404	-0.465376
3	2.007808	-0.291375	-0.532363	-0.345953	-0.585043	1.387054	-1.179234	2.128151	-0.516268
4	-0.695634	-0.104331	-0.178771	0.040735	0.101732	-0.601749	0.704258	-0.541946	-0.041817

Outlier Analysis in Child_Mort, Income & gdpp Columns using Box Plot







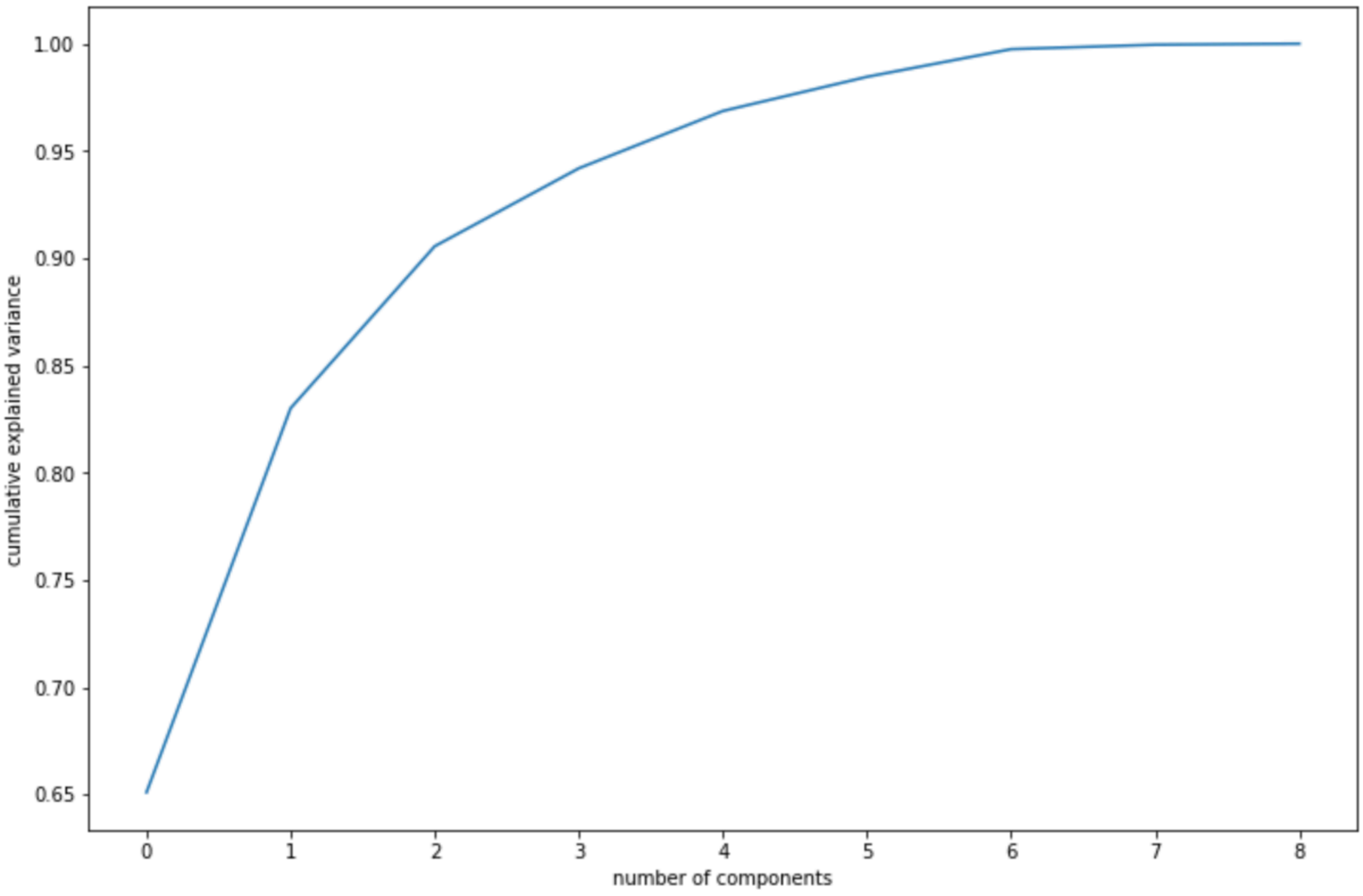
The data set contains outliers which we have removed using Inter Quartile Range method. After Outlier removal we have lost 4 rows which is very small hence we will go with outlier removal before PCA analysis and clustering.

PCA Analysis - Scree Plot

Principal component analysis (**PCA**) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called **principal components**.

The **scree plot** is used to determine the number of factors to retain in principal components to keep in a principal component analysis (PCA).

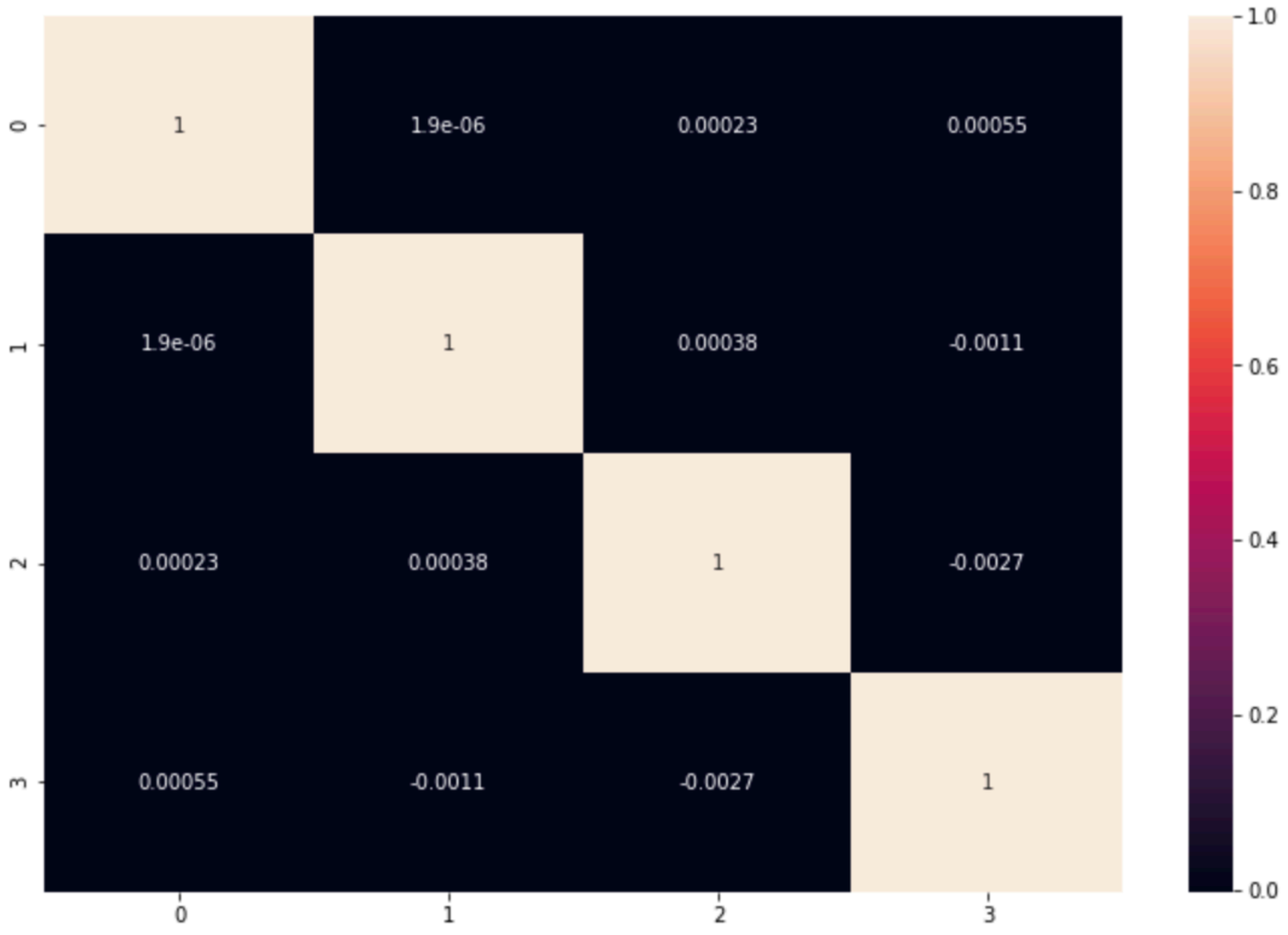
Around 96% of the information is being explained by 4 components.



Data Frame with 4 PCA Components

	Feature	PC1	PC2	PC3	PC4
0	child_mort	-0.405471	0.454189	-0.084379	0.194838
1	exports	0.190279	0.214574	0.053759	0.124178
2	health	0.381437	0.459480	-0.088565	-0.125078
3	imports	0.203451	0.203730	-0.022106	0.086685
4	income	0.348594	0.223513	0.242375	0.238737
5	inflation	-0.121641	-0.021234	0.944603	0.102190
6	life_expec	0.429326	-0.299784	0.118092	-0.606193
7	total_fer	-0.386057	0.448679	0.118621	-0.698833
8	gdpp	0.377923	0.391726	0.051183	0.013544

Correlation Matrix Heat map of 4 PCA Components



We see that correlations are indeed very close to 0

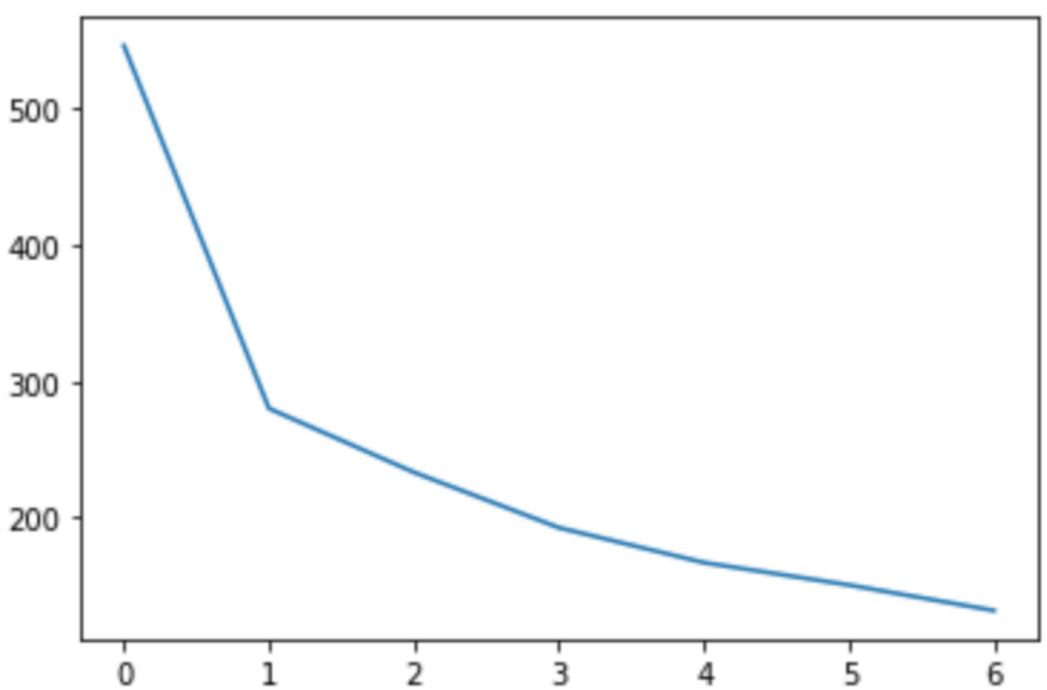
Clustering - K-Means Clustering

Hopkins Statistics calculated for 4 Principal Components is 0.84 showing very good tendency of clustering.

Finding Optimal No. Of Clusters - Elbow Curve (SSD Method)

```
ssd = []
range_n_clusters = [2,3,4,5,6,7,8]
for num_cluster in range_n_clusters:
    kmeans = KMeans(n_clusters=num_cluster, max_iter=50)
    kmeans.fit(pca_scaled_num_df)
    ssd.append(kmeans.inertia_)
# Plot the SSDs for each n_cluster
plt.plot(ssd)
```

```
[<matplotlib.lines.Line2D at 0x1a27733080>]
```



Finding Optimal No. Of Clusters - Silhouette Score

2. Silhouette Score

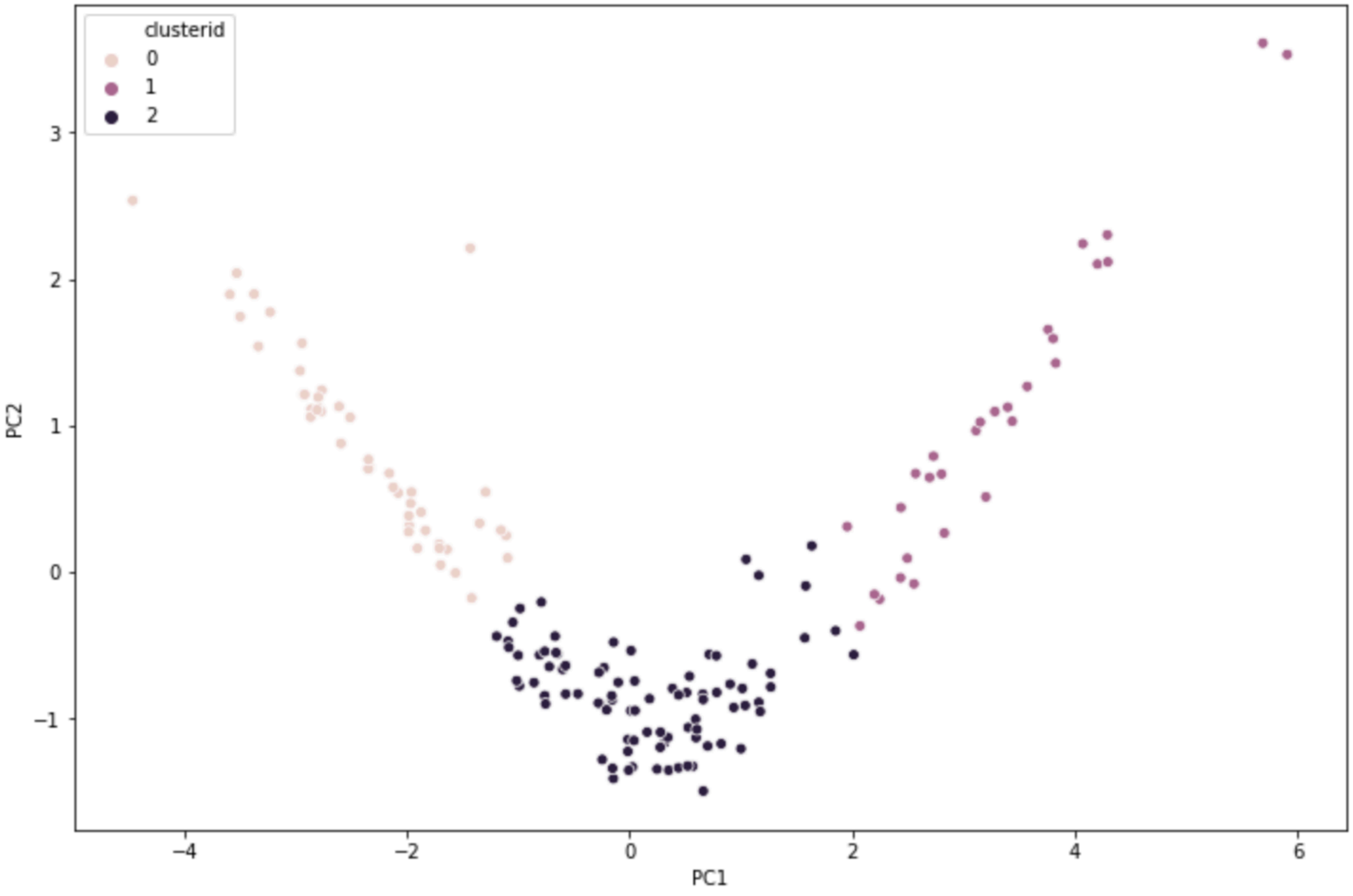
```
range_n_clusters = [2,3,4,5,6,7,8]
for num_cluster in range_n_clusters:
    kmeans = KMeans(n_clusters=num_cluster, max_iter=50)
    kmeans.fit(pca_scaled_num_df)
    cluster_labels = kmeans.labels_
    silhouette_avg = silhouette_score(pca_scaled_num_df, cluster_labels)

    print("For n_Cluster = {0}, the silhouette score is {1}".format(num_cluster, silhouette_avg))
```

```
For n_Cluster = 2, the silhouette score is 0.5017774588536934
For n_Cluster = 3, the silhouette score is 0.5638733473215646
For n_Cluster = 4, the silhouette score is 0.5054535245513277
For n_Cluster = 5, the silhouette score is 0.43374542777913483
For n_Cluster = 6, the silhouette score is 0.3157168829042766
For n_Cluster = 7, the silhouette score is 0.3418757194558599
For n_Cluster = 8, the silhouette score is 0.3500836483939067
```

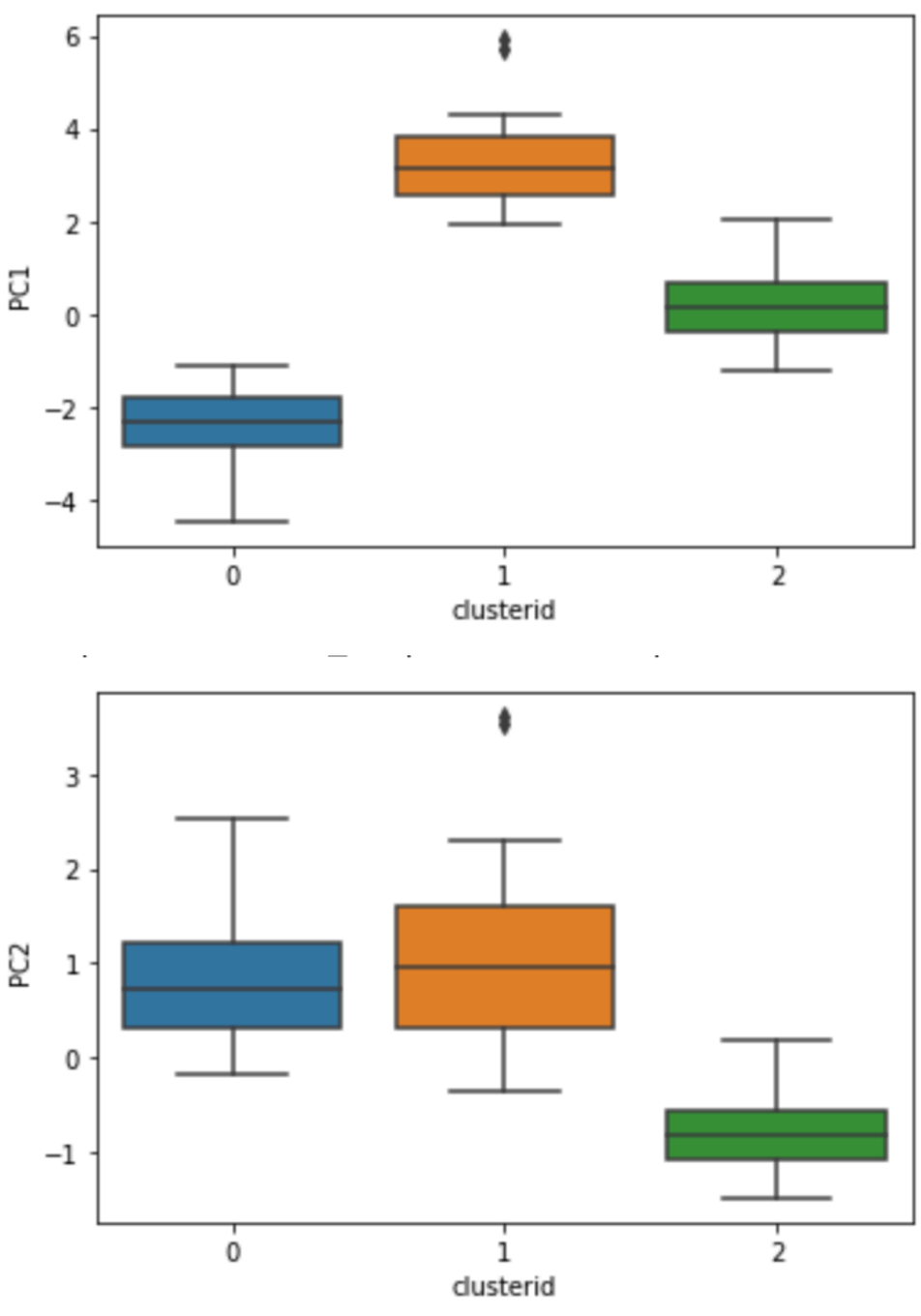
From the above silhouette score, num_cluster = 3 will be suitable for this dataset.

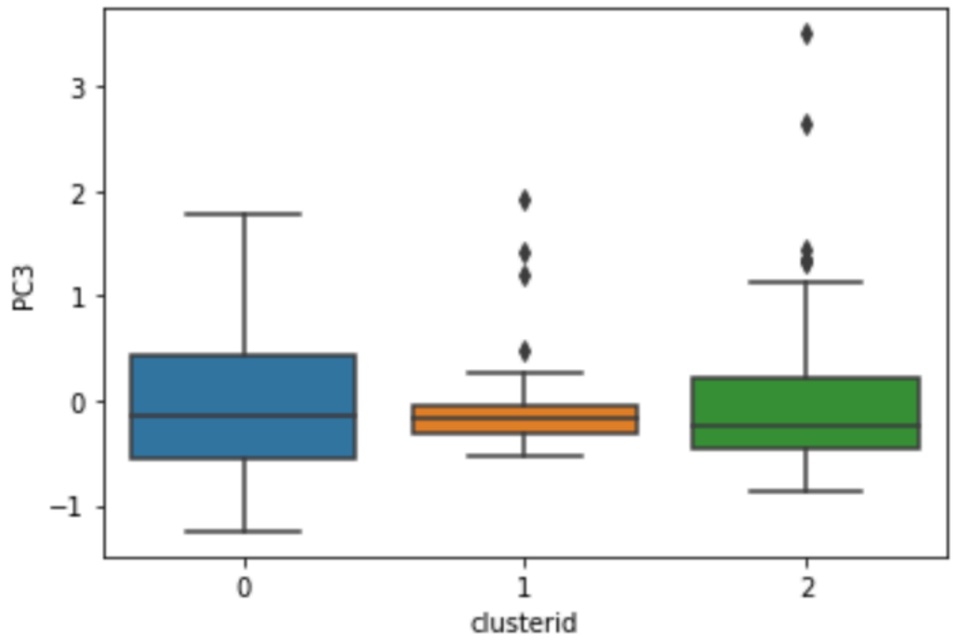
Visualising the Points on Principal Components

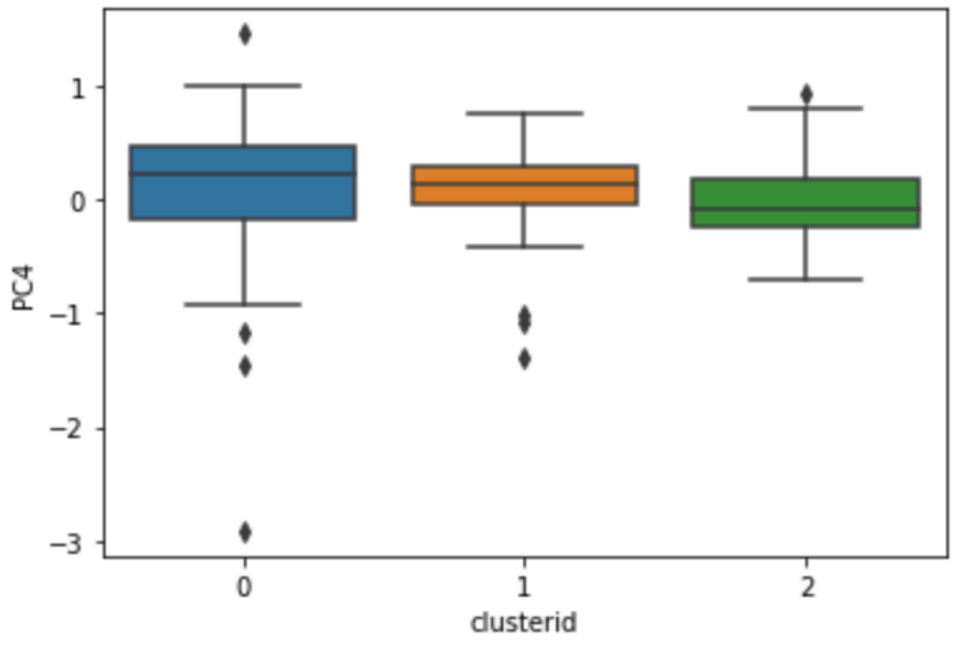


Visualising the Clusters on Principal Components

PC1 & PC2 consists of datapoint with good variations in clusters 0,1, & 2. In PC3 & PC4 however the data in cluster 0,1, & 2 have similar median values but cluster 0 and 2 have different min, max and quartile values.

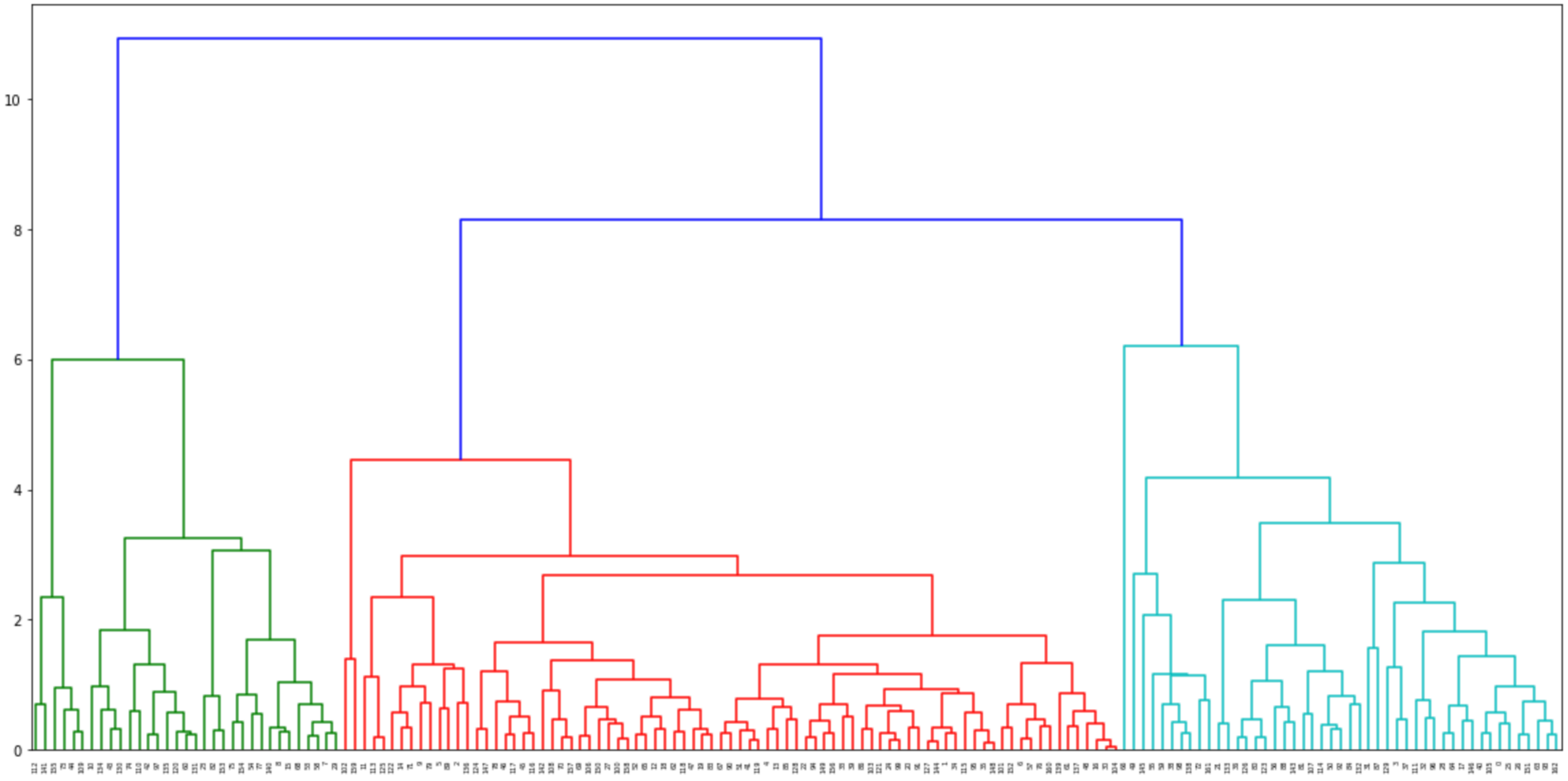






Hierarchical Clustering

The Dendrogram shown below is using complete linkage method and it clearly shows formation of 3 distinct clusters to explain the data points in the dataset.



Countries in Various Clusters

```
cl1_count_kmean = df2[df2.clusterid == 0].country
cl1_count_kmean.head()
```

```
0      Afghanistan
3           Angola
17           Benin
21        Botswana
25    Burkina Faso
Name: country, dtype: object
```

```
cl2_count_kmean = df2[df2.clusterid == 1].country
cl2_count_kmean.head()
```

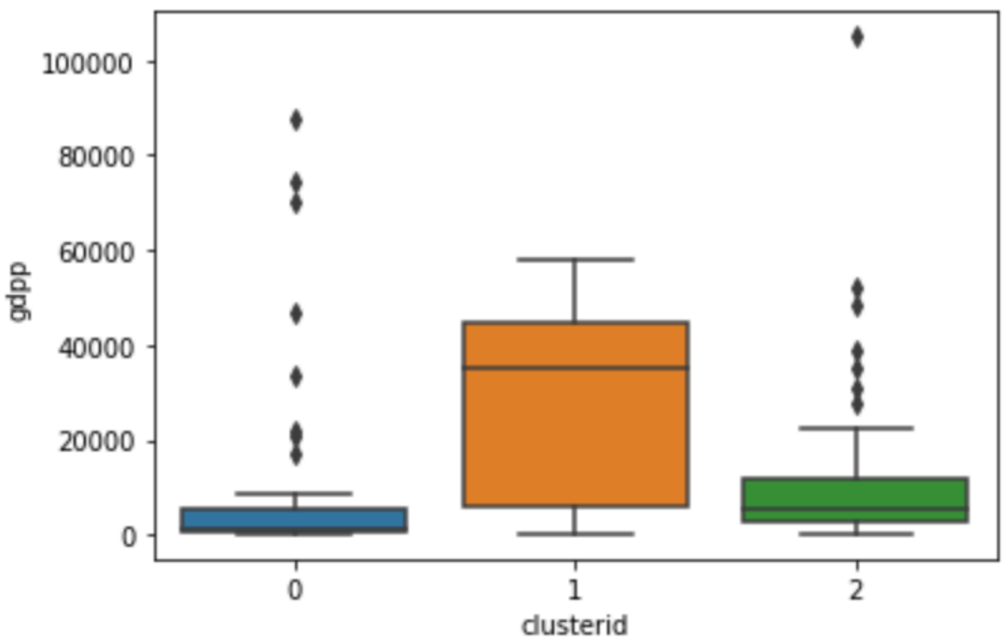
```
7     Australia
8       Austria
15      Belgium
23       Brunei
29       Canada
Name: country, dtype: object
```

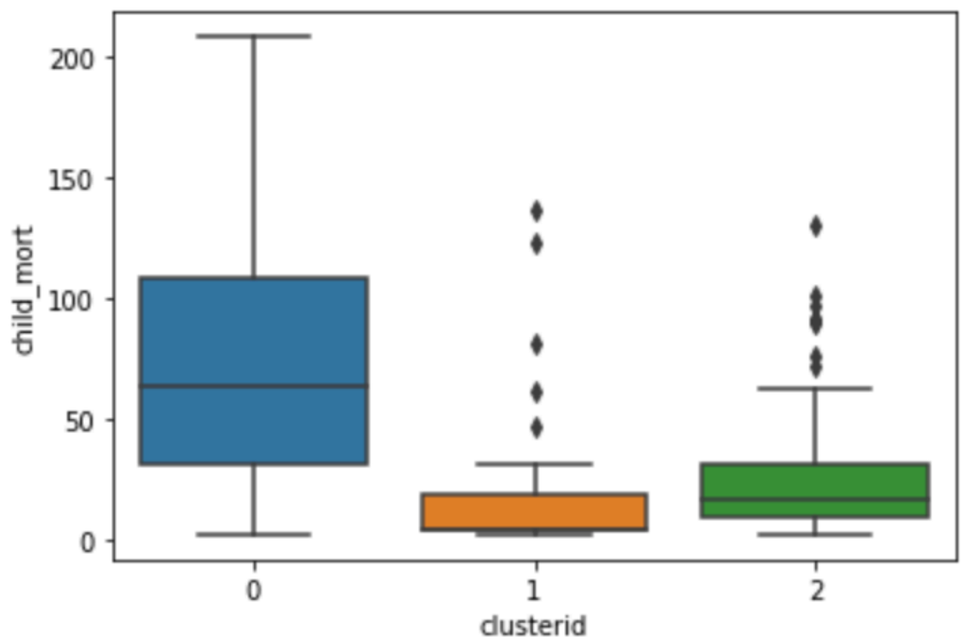
```
cl3_count_kmean = df2[df2.clusterid == 2].country
cl3_count_kmean.head()
```

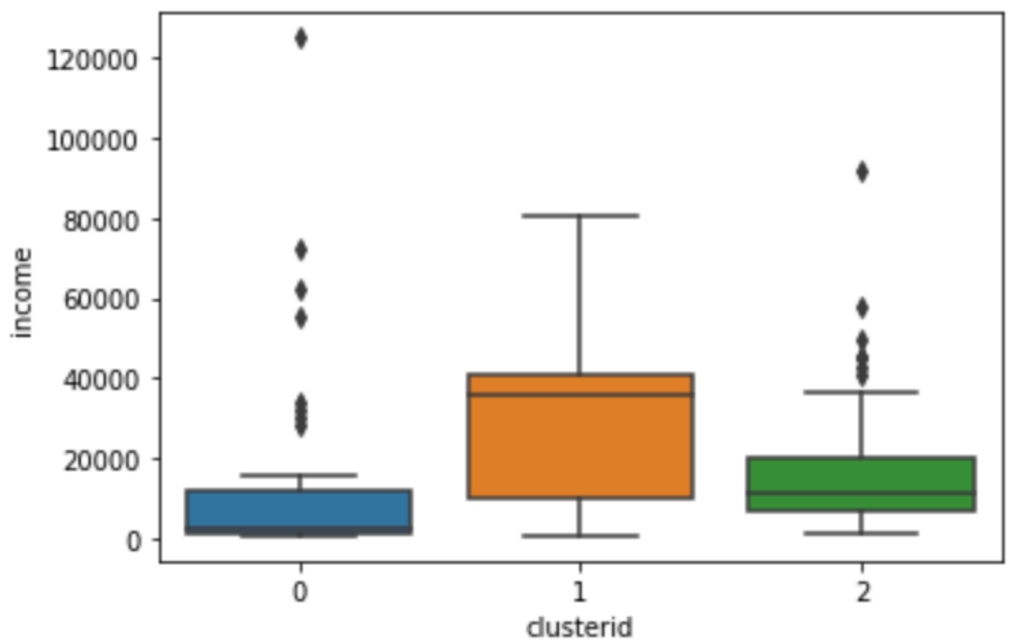
```
1                 Albania
2                 Algeria
4    Antigua and Barbuda
5               Argentina
6                 Armenia
Name: country, dtype: object
```

Comparing Child_mort, gdpp & Income across Clusters Using Box Plots

Hence in all three parameters `gdpp`, `child_mort`, `income`, countries falling under cluster 1 with clusterid = 0 derived from K-Means clustering method are under the category of most under-developed countries and need more focus in terms of financial Aids.







Binning & List of Under Developed Countries with Urgent Need for Financial Aids

Binning

```
fin=df2[df2['child_mort']>= 63.9]
fin=fin[fin['income']<= 1220]
fin1=fin[fin['gdpp']<= 2690]
fin1.head()
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	clusterid	cluster_label
26	Burundi	93.6	20.6052	26.7960	90.552	764	12.30	57.7	6.26	231	0	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888	2.01	47.5	5.21	446	0	0
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334	0	0
63	Guinea	109.0	196.3440	31.9464	279.936	1190	16.10	58.0	5.34	648	0	0
88	Liberia	89.3	62.4570	38.5860	302.802	700	5.47	60.8	5.02	327	0	0

Based on Binning as well as clustering it is clear that countries mentioned above with clusterid/cluster_label = 0 are among the most under-developed countries and need immediate attention in priority in terms of financial Aids.

The End.