# PCA Assignment & Clustering Assignment

SUEL AHMED

# Problem Statement

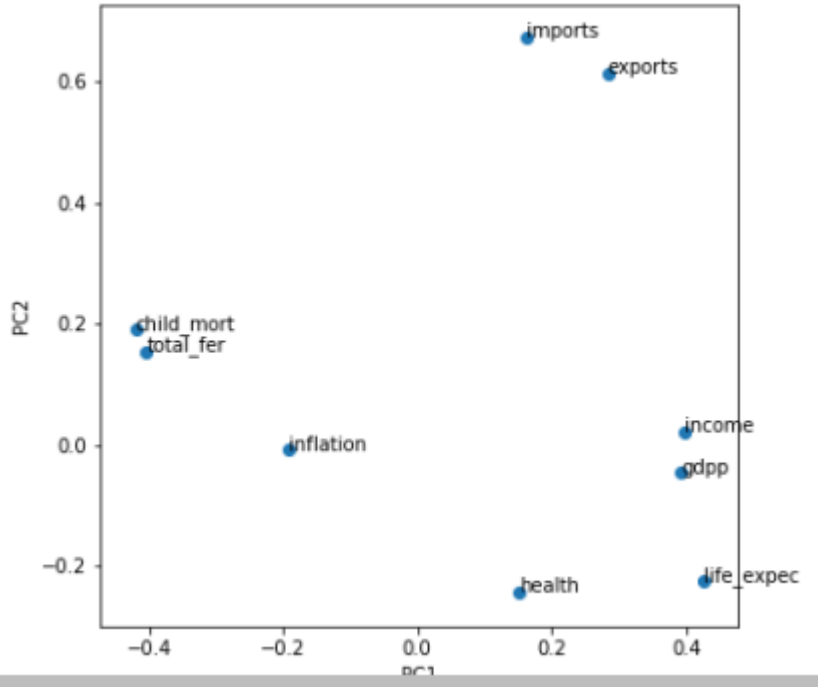
- HELP International is an international humanitarian NGO that is looking to provide aid to the countries in need.

- They have raised US $ 10 million and their CEO is looking to invest this money strategically and effectively.

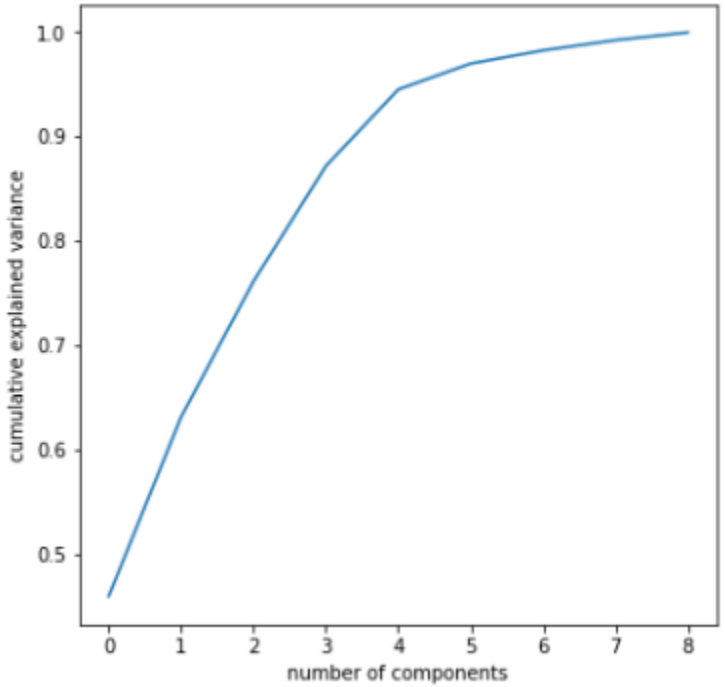
# METHODLOGY & ANALYSIS

- Read the data
- Check the data for null values
- Check the data for outliers
- Apply standard scaling to normalize the data
- Check for correlation, if strong positive or negative correlation are present then remove those variables
- But we used PCA to reduce dimentionality and multi-collinearity

- Draw a plot between PC1 and PC2. Some pattern in data is visible now



- Check for optimal number of components from scree plot. In this case, the number was decided to be 5 as it describe 95% (approx.) variance in data.



- Use incremental PCA with 5 components
- Use transformation (PC) on dataset
- Check for correlation again, No strong correlation is present at the moment. WE ARE SUCCESFULL IN REDUCING THE DIMENSIONS OF DATASET

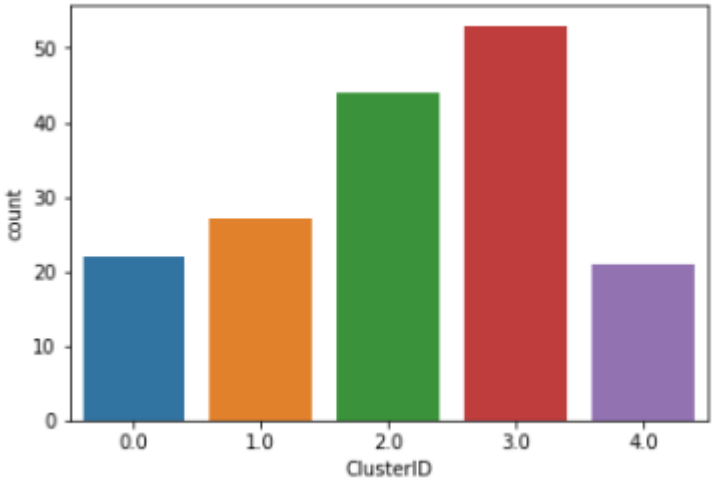
- Perform outlier analysis, the following formula was used to discard outliers

```
plt.boxplot(df_train_pca_1.PC1)
Q1 = df_train_pca_1.PC1.quantile(0.25)
Q3 =df_train_pca_1.PC1.quantile(0.75)
IQR = Q3 - Q1
df_train_pca_1 = df_train_pca_1[(df_train_pca_1.PC1 >= Q1 - 1.5*IQR) & (df_train_pca_1.PC1 <=Q3 + 1.5*IQR)]
```

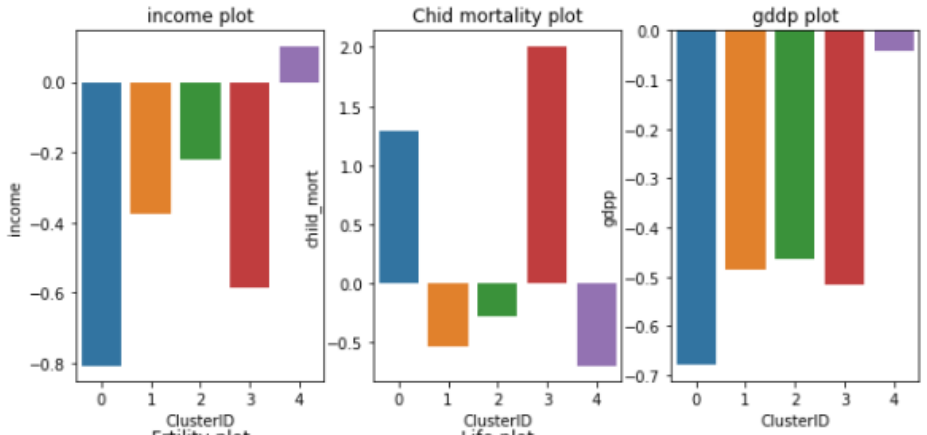
- Run a K mean algorithm & Heirarchial clustering
- Perform a hopkin and sillhoutte analysis on dataset to check whether is suitable for clustering.
- Also check for sum of squared distance.
- Perform the same steps for separate dataset with outlier points and at the end merge the data (outlier and non outlier data).
- Euclidean distance was used to manually reassign outlier point clusters to original data point cluster.
- Visualize the data

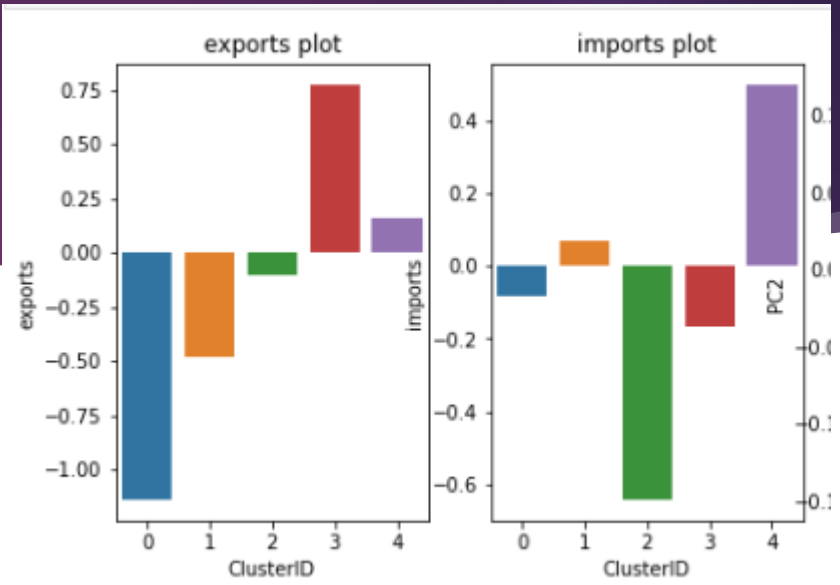
# Analysis with K means

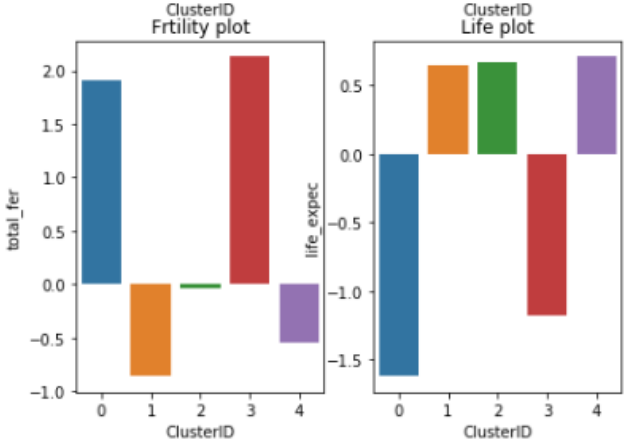
- Cluster formed in k means and their analysis
- Cluster group 3 was more frequent than other groups

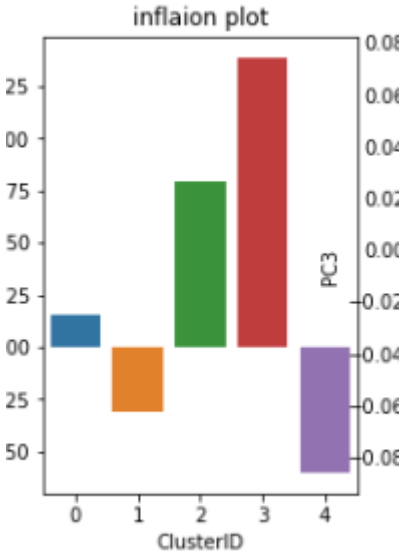


- Cluster group 0 and 3 stands out from rest of clusters because they have negative clusters.
- Whereas cluster 4 seems most suitable as it has high income plot, low child mortality rate, higher gdp and life expentency than others.
- Cluster 4 countries have low inflation and high export as compared to others
- Please refer to attached visualization on next slide.


income plot
income
ClusterID
Chid mortality plot
child_mort
ClusterID
gddp plot
gdpp
ClusterID

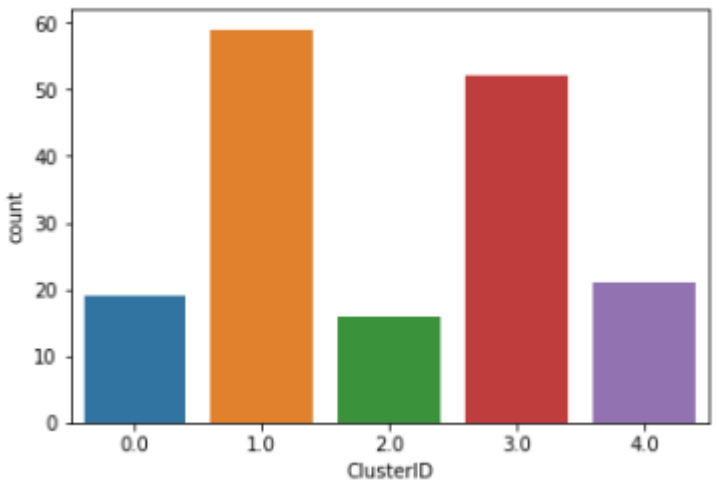

exports plot
exports
ClusterID
imports plot
imports
PC2
ClusterID

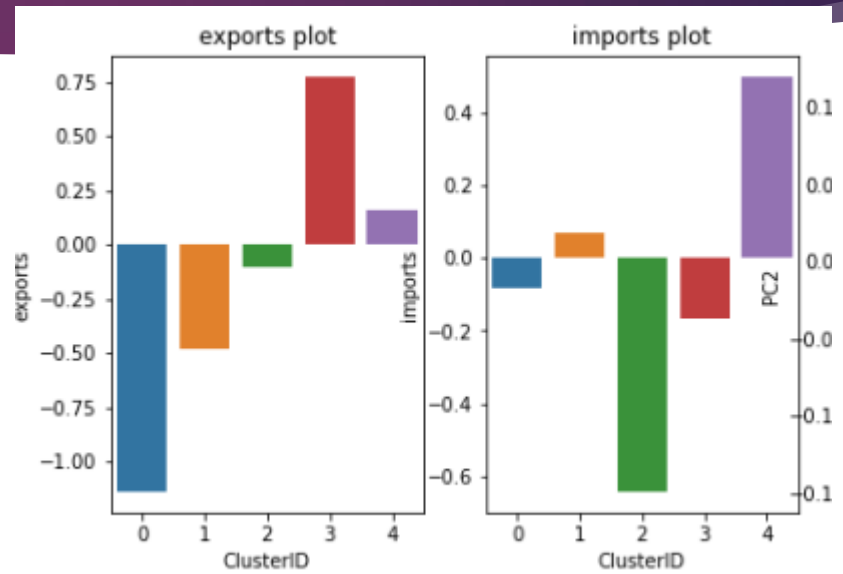

Frtility plot
total_fer
ClusterID
Life plot
life_expec
ClusterID

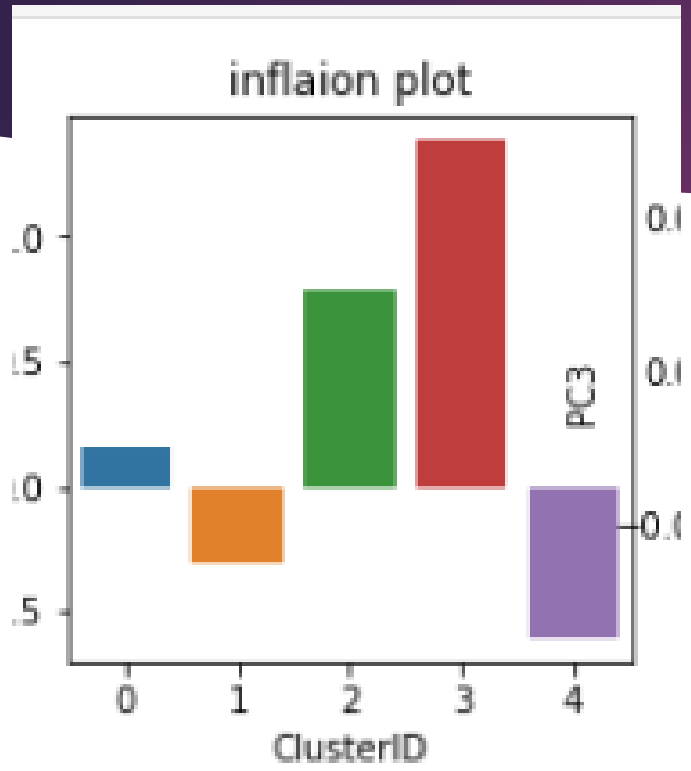

inflaion plot
PC3
ClusterID

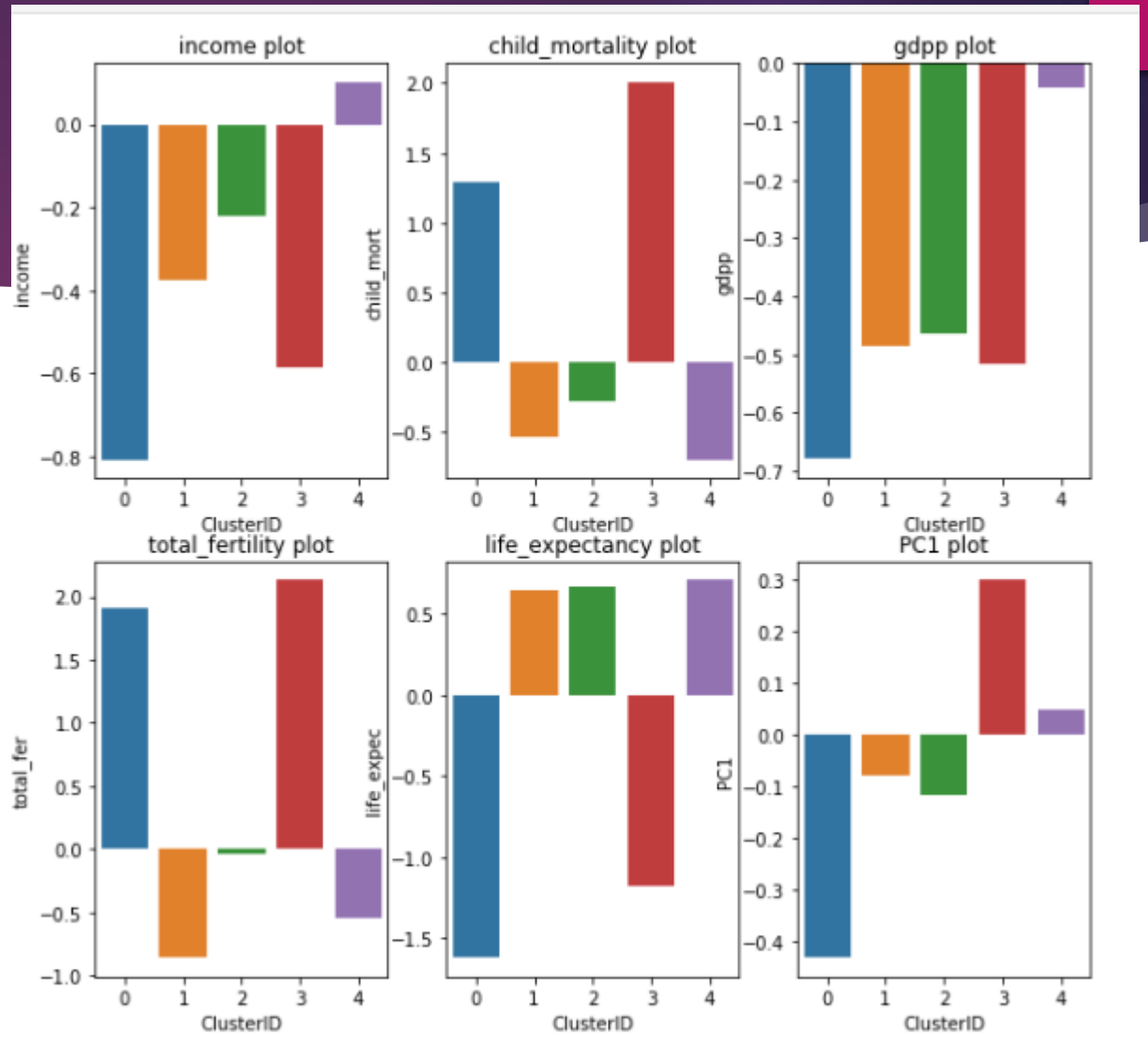
# Analysis with Heirarchial

- With this method too, the results are more or less the same.
- Clusters 1 & 3 were more frequent than others.
- Cluster 4 countries showed more or less positive trend with higher income & gdp, lower child mortality rates, higher life expectancy but low fertility.
- Cluster 4 countries had higher imports whereas cluster 3 countries had higher exports.
- Cluster 4 countries had lowest inflation whereas cluster 3 countries had highest inflation.
- Please refer to attached visualizations


count
ClusterID
0.0
1.0
2.0
3.0
4.0


exports plot
imports plot
exports
imports
PC2
ClusterID
ClusterID


inflaion plot
PC3
ClusterID


income plot
income
child_mortality plot
child_mort
gdpp plot
gdpp
total_fertility plot
total_fer
life_expectancy plot
life_expec
PC1 plot
PC1
ClusterID

# Recommendations

- If the CEO wants, to invest he/she should invest in countries in cluster 4.
- The reason is that the development record of countries in cluster 4 is good.
- They have good life expectancy, low inflation , high export import deficit and most importantly higher income.
- CEO can also invest some money in countries in cluster 3, as they have thriving exports that is a good sign for future potential. But on other areas cluster 3 severely lags behind.
- The ROI can be expected to be high with countries in cluster 4 rather that countries in other clusters.