# Cluster & PCA Assignment

Name:

1. **Pratik Kumar**

# Abstract

An international NGO looking forward to fighting to the poverty and providing the people of backward countries which basic of amenities and relief in the disasters.

**Problem statement** is that NGO want to know how to decide to use their money to help poor and needy peoples.

Solution approach :

We need to choose counties which has very need of aid.

We have a dataset accompanied with some information about countries. We will divide data as per common behaviour of countries to find very low performer countries on the base of given features on dataset.

We will look in some features that is property of low performer countries like **child mortality , health, GDP etc**.

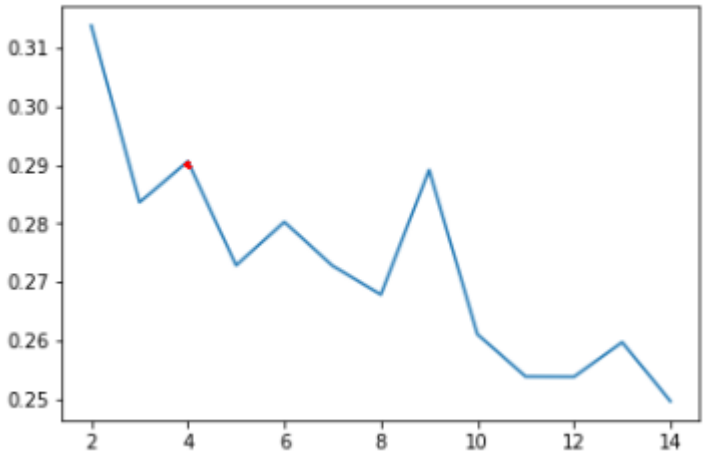
# Problem solving methodology

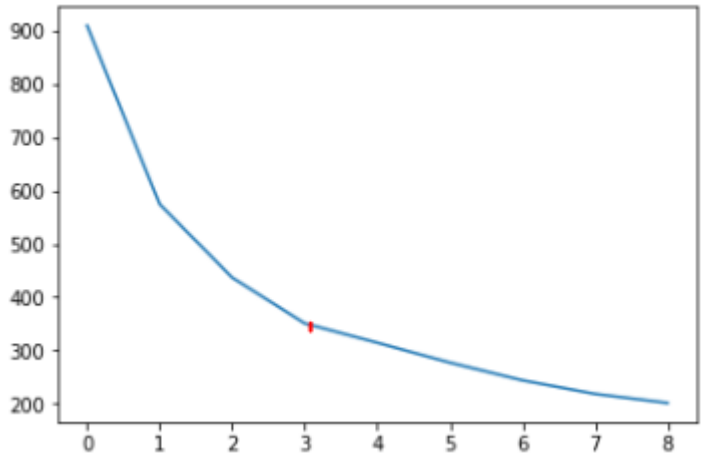
In this case study, We will check missing values and outliers in dataset and handle them as per scenario.

Use **PCA** to decrease features or eliminate multicollinearity without loss information of datasets.

```
from sklearn.decomposition import PCA
pca = PCA(svd_solver='randomized', random_state=42)
```

Using **KMeans** method to divide same properties dataset in clusters. Will use k = 3 as per evidence of elbow and silhouette score.

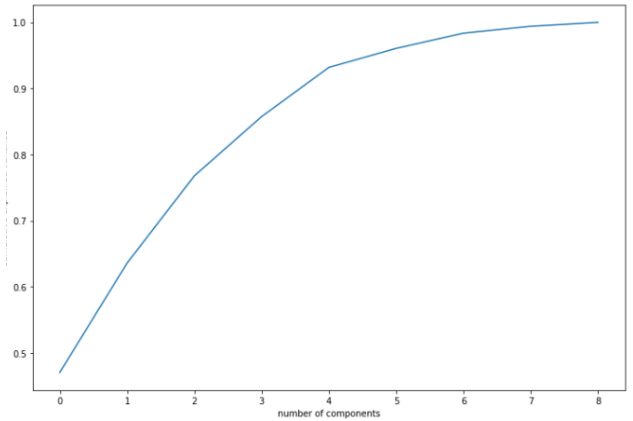


# Principal Component Analysis

Apply PCA  method on dataset and find out relevant PCA components to apply upon cluster Analyses to find desired data.

```
from sklearn.decomposition import PCA
pca = PCA(svd_solver='randomized', random_state=42)
```




After find that four components covering max area of data

So we picked best components of very less equal to zero

Variance on features.

Make final dataframe “mydf”  to process further work. Now

We had featured out data set ready to apply KMean on it.

```
## create dataFrame from the PCA's
mydf = pd.DataFrame(df_train_pca, columns=['A','B','C','D'])
#mydf = pd.DataFrame({'A':df_train_pca[0],'B':df_train_pca[1],
mydf.head()
```

| | A | B | C | D |
|---|---|---|---|---|
| 0 | -0.855482 | -0.227921 | -0.335437 | -1.263436 |
| 1 | 0.047495 | -1.244787 | 1.759583 | -0.410392 |
| 2 | 3.169743 | 0.433401 | 2.578709 | 1.848593 |
| 3 | -2.054186 | 0.551182 | 0.002359 | 0.030439 |
| 4 | -0.766793 | -2.997358 | 0.673137 | 0.208372 |

# Clustering

We did very first pre-processing raw data, we are good so-far.

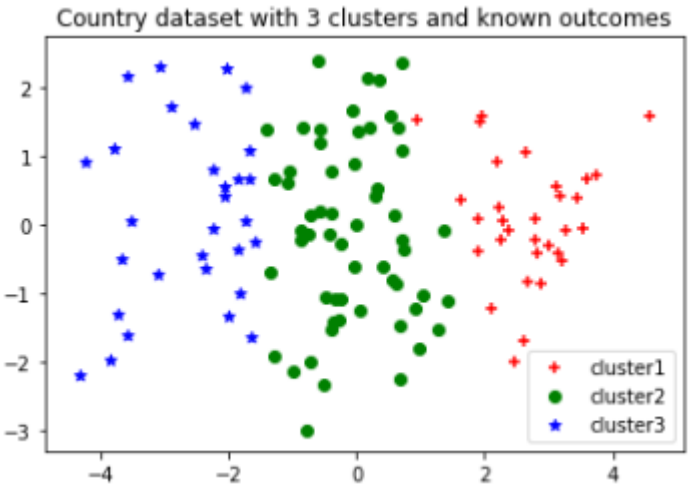
We are going to use **KMeans** & **Hierarchical** Clustering methods.

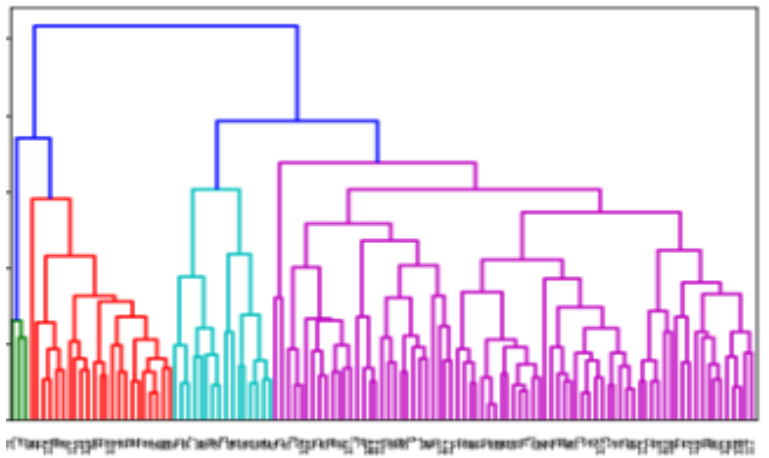
Using Hopkins method checked tendency of data. We good get 72% result. Initialization of k = 4 randomly.

Find cluster centroids using silhouette score and elbow curve methods. We got data in three clusters.

```
3 my_cluster1 = KMeans(n_clusters=3, max_iter=50, random_state=50)
4 my_cluster1.fit(mydf_mod1)
```

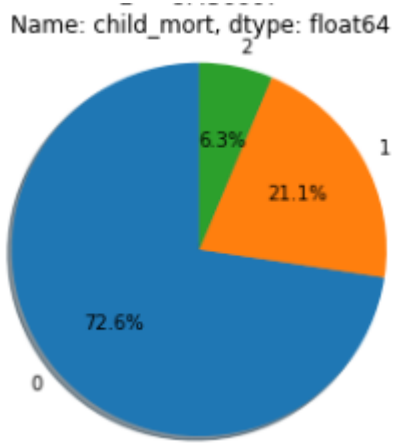
```
h_cluster = linkage(n_df, method = "complete", metric='euclidean')
dendrogram(h_cluster)
plt.show()
```

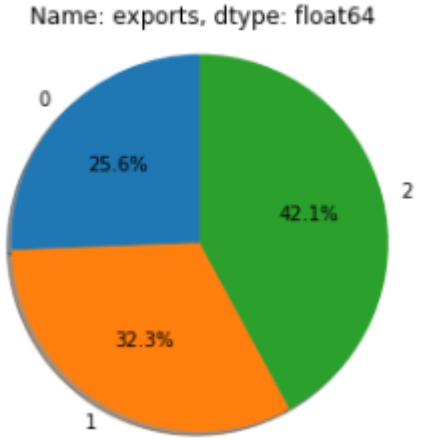


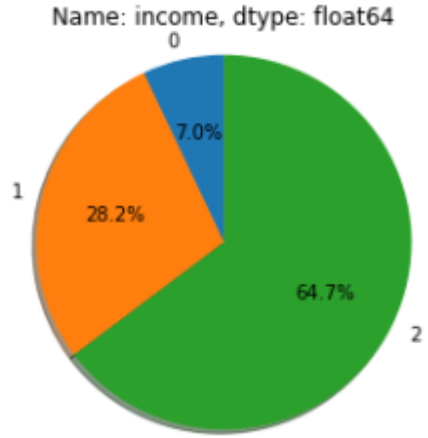



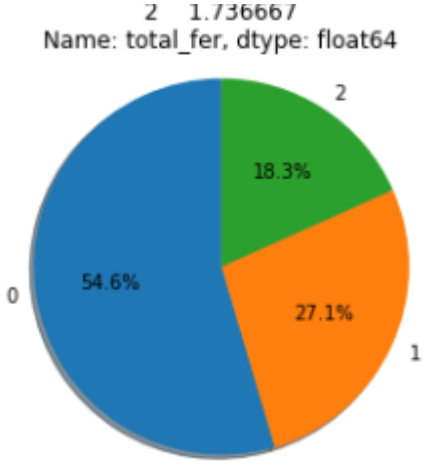
# KMeans Algorithm Results

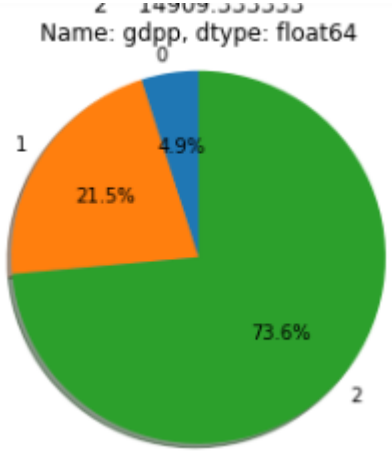
- Cluster 0's country need aid to survive. Cluster 0 look like group of under developed countries.

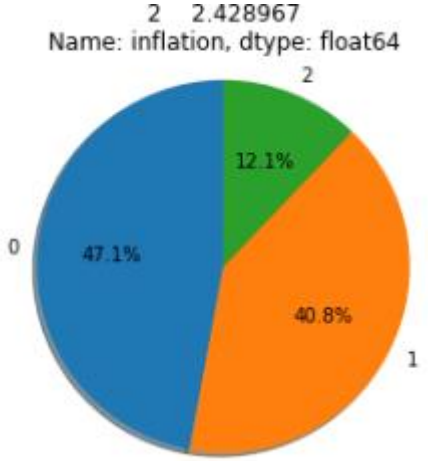


# Result Summary

- As per the Chart in pervious slid. We can observe that cluster id 0 need more aid to survive. So we can recommend cluster 0’s group courtiers to NGO.
- We can see that in cluster 0.

Child mortality = 72.6 %

Export = 25.6 %

Income = 7 % only

Inflation = 47.1 %

**All above evidence proof that group 0 is final selection of countries.**

# Countries Name

These are countries which need aid to survive.
All are under developed countries.

| | |
|---|---|
| 1 | Angola |
| 2 | Benin |
| 3 | Burkina Faso |
| 4 | Cameroon |
| 5 | Chad |
| 6 | Comoros |
| 7 | Congo, Dem. Rep. |
| 8 | Congo, Rep. |
| 9 | Cote d'Ivoire |
| 10 | Gambia |
| 11 | Ghana |
| 12 | Guinea |
| 13 | Guinea-Bissau |
| 14 | Haiti |
| 15 | Kenya |
| 16 | Madagascar |
| 17 | Malawi |
| 18 | Mali |
| 19 | Mauritania |
| 20 | Mozambique |
| 21 | Niger |
| 22 | Senegal |
| 23 | Tanzania |
| 24 | Togo |
| 25 | Uganda |
| 26 | Zambia |

# Thank You