a) Compare and contrast K-means Clustering and Hierarchical Clustering.

Clustering is an unsupervised learning, where you are not interested in prediction because you do not have a target or outcome variable.

K means is one of the unsupervised learning techniques. The main idea is to find the k centres one center for each of the clusters. In simple terms, the algorithm needs to find data points whose values are **similar** to each other and therefore these points would then belong to the same cluster. The method in which any clustering algorithm goes about doing that is through the method of finding something called a “**distance measure**”. The distance measure that is used in K-means clustering is called the **Euclidean Distance** measure. If there are 2 points X and Y having n dimensions

$$X=(X_1,X_2,X_3,\dots X_n)$$

$$Y=(Y_1,Y_2,Y_3,\dots.Y_n)$$

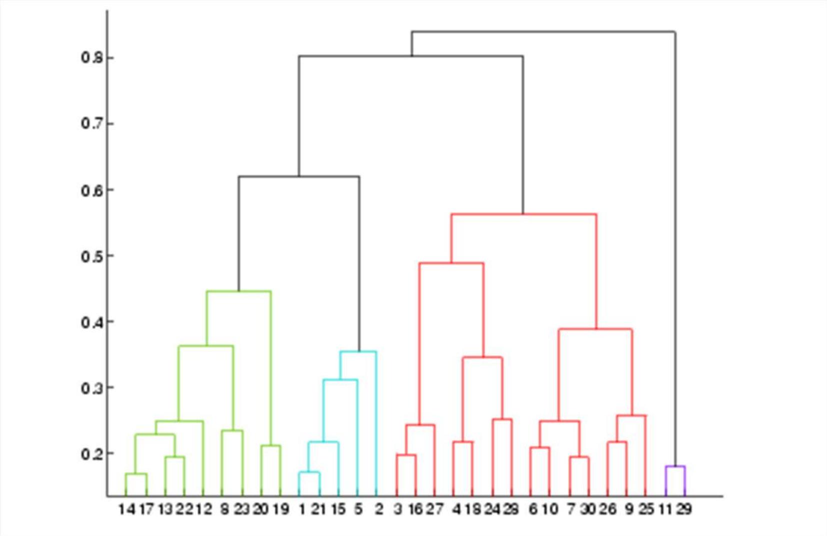
Then the **Euclidean Distance D** is given as

$$D=\sqrt{\boxed{(X_1-Y_1)_2+(X_2-Y_2)_2+\dots(X_n-Y_n)_2}}$$

Essentially, the observations which are closer or more similar to each other would have a low Euclidean distance and the observations which are farther or less similar to each other would have a higher Euclidean distance.

Another algorithm to achieve unsupervised clustering. This is called **Hierarchical Clustering**. Here, instead of pre-defining the number of clusters, we first have to visually describe the similarity or dissimilarity between the different data points and then decide the appropriate number of clusters on the basis of these similarities or dissimilarities. The hierarchical clustering algorithm does not have this restriction.

The output of the hierarchical clustering algorithm is quite different from the K-mean algorithm as well. It results in an inverted tree-shaped structure, called the dendrogram. An example of a dendrogram is shown below.



Dendrogram

b) Briefly explain the steps of the K-means clustering algorithm.

First step is to randomly pick clusters centres' and labelling them which represents cluster centres'.

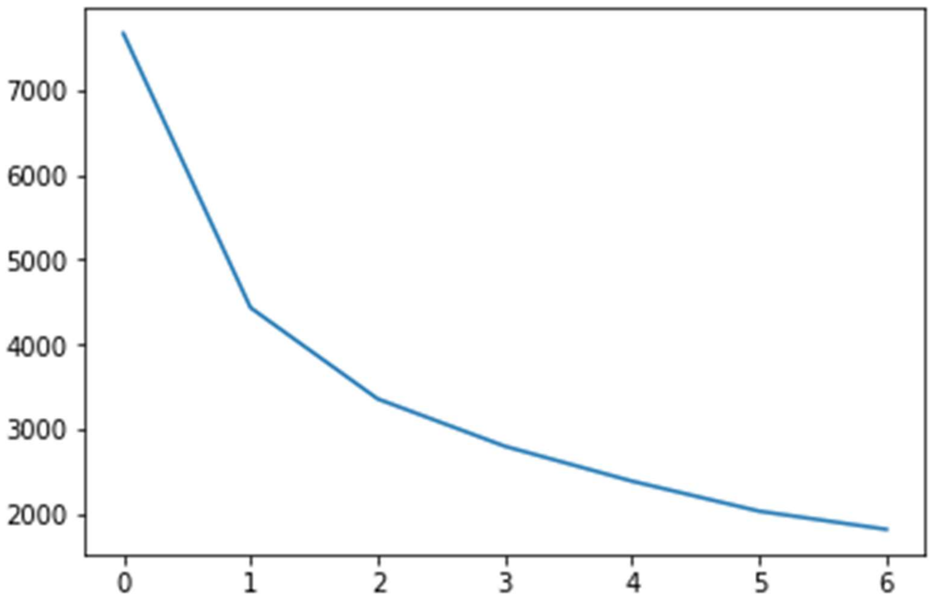
Second step is to assign closest observations closest to the cluster center which will be based on minimum distance.

Third step is to update the cluster center based on data points assigned to them this will find new cluster centers

Last step is to repeat the steps till optimum clusters are reached.

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

There are various techniques to be followed in order to determine the value of k in k means algorithm. The most popular method is the elbow method.



Clearly the elbows are forming at 2 so the optimal value can be taken as 2 to perform k means. The diagram plots various values as k changes as the value of k increases the elements present in the cluster becomes less. That implies element's closer to the cluster centroid. So we need to find that point were the distortion declines this is the elbow point.

d) Explain the necessity for scaling/standardisation before performing Clustering.

The idea behind standardisation is to reduce variance in the data which leads to poorer convergence. Hence standardisation data is the pre processing step to perform cluster analysis. Standardisation helps to put all the variables in the same scale If there are variables with larger weight more importance is provided to that data thus making a bias model.
Hence Scaling/Standardisation is important pre-processing step for clustering analysis.

e) Explain the different linkages used in Hierarchical Clustering.

The different types of linkages.

Single Linkage: Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters

Complete Linkage: Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters

Average Linkage: Here, the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.

a) **Give at least three applications of using PCA.**

PCA is the dimension reduction technique used for model building It can be applied in image recognition, banking sector, bioinformatics. Etc.

b) **Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.**

PCA is basically dimensionality reduction technique there are many ways to attain this technique one is Feature elimination were the variables which are considered worst predictors are removed from the variable though some information is lost in this process another is feature extraction we extract information from the present variables. This involves cmbining the variable in such as way that most information is retained. This is done by transforming the present variables. Here, all the variables are not corelated hence the multicollinearity problem is solved, less independent variables are present in the dataset. PCA “Variance” means “Summative variance” or “Overall Variability.” The original variables after performing PCA are replaced by new variables called as principal components and they have variances also known as eigen values. Hence covariance matrix are computed.

c) **State at least three shortcomings of using Principal Component Analysis.**

One need to perform standardisation before PCA.
Variables sometimes becomes less interpretable.
Sometimes loss of information

Question 1: Assignment Summary

Our job here was to categorise the countries sing some socio-economic and health factors that determine the overall development of the country. Then to suggest the countries which the CEO needs to focus on the most. The datasets contained those socio-economic factors .

Apart from the country variable there were in all 9 dependent variables in the data set .

PCA was performed after scaling the variables to reduce the dimensions. The scree plot was plotted which showed 4 components can be used to explain the data more than 95%.

Hence 4 principal components were kept. The correlation plot was obtained to confirm if the variables shows multicollinearity. Which was not present.

The important variable were:

PC1:Life_expectancy

PC2:Imports

PC3:Inflation

PC4:gdpp

Were kept in the dataset.

K means algorithm was performed the elbow curve gave the optimal value for clustering as 4. Hence final model was performed with 4 clusters.

The silhouette analysis score were ranging from 28%to 35%.

Outlier analysis was performed and post that box plot was plotted to find the clusters distribution.

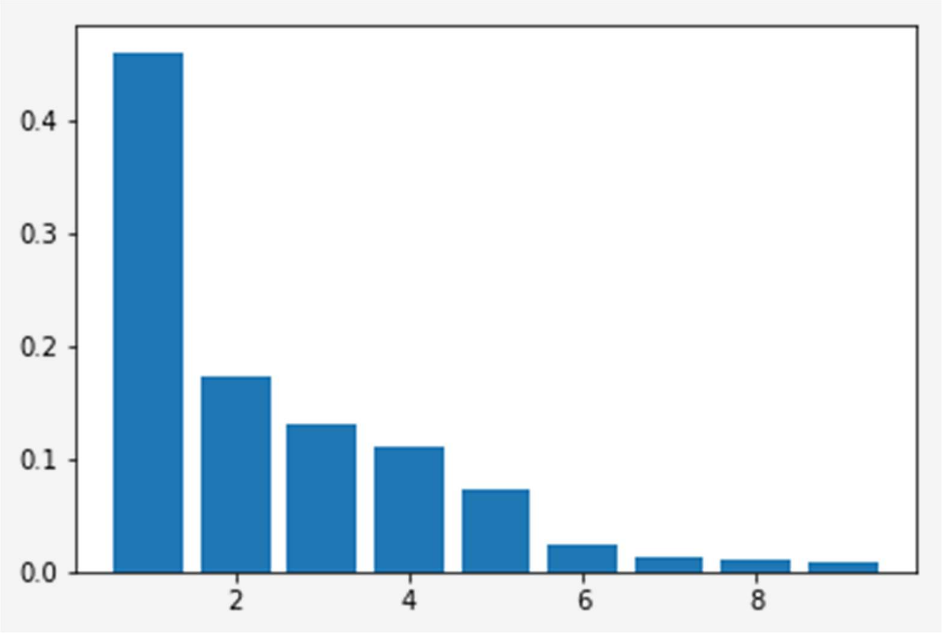
Similarly hierarchical clustering was also performed.

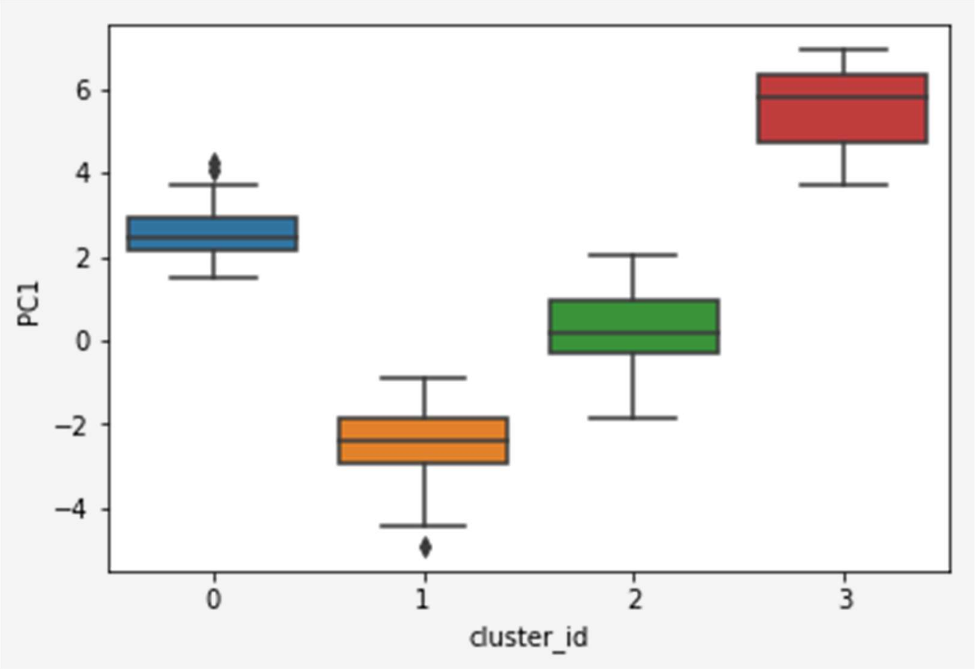
After the analysis we found that country like “Nigeria” is in desperate need for help.

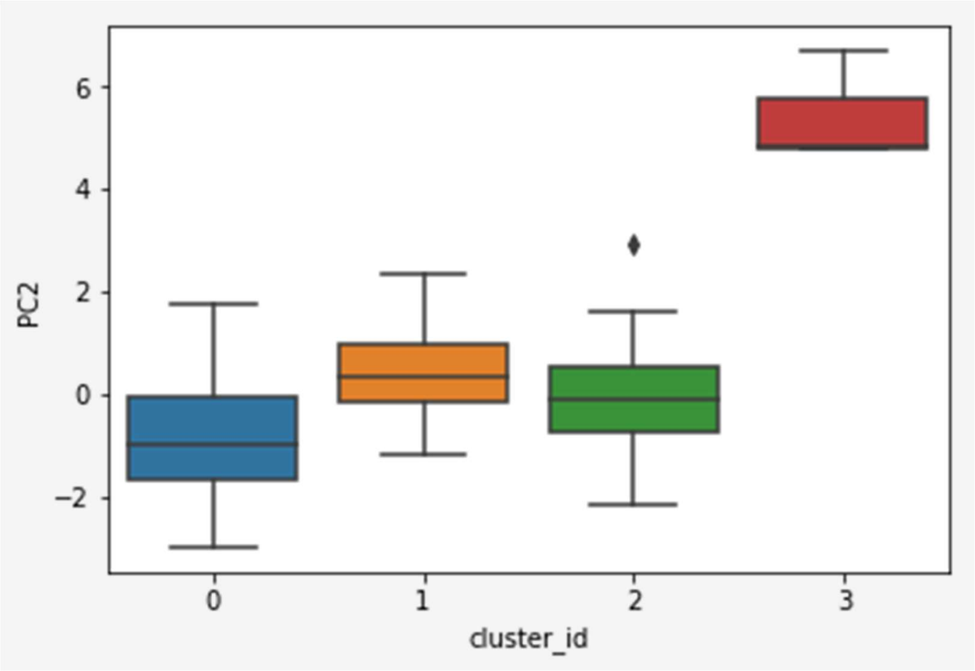
And country like “Luxemborg” are good.

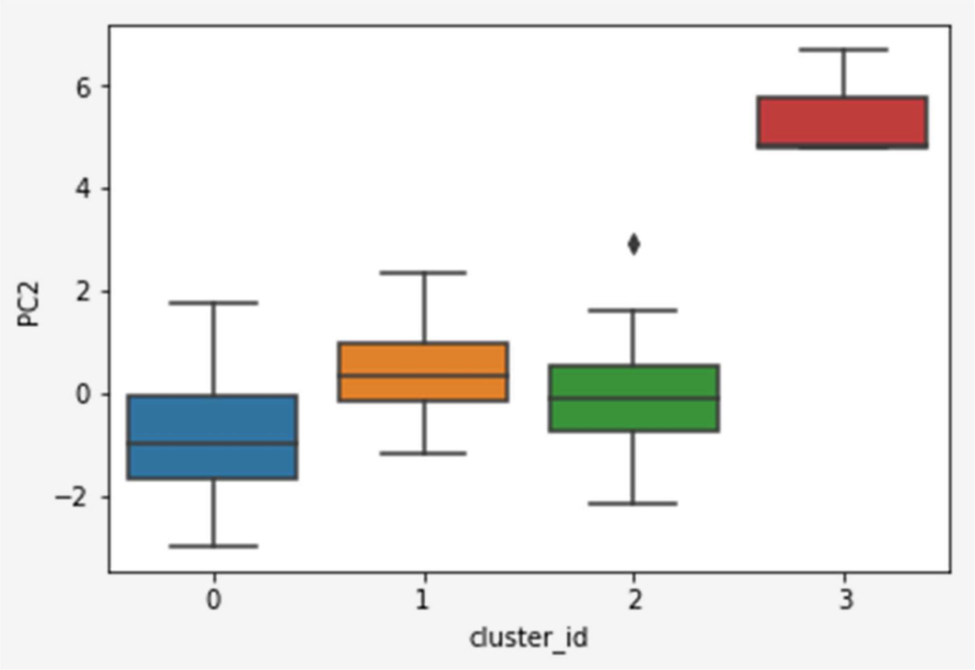
Plots from the data:

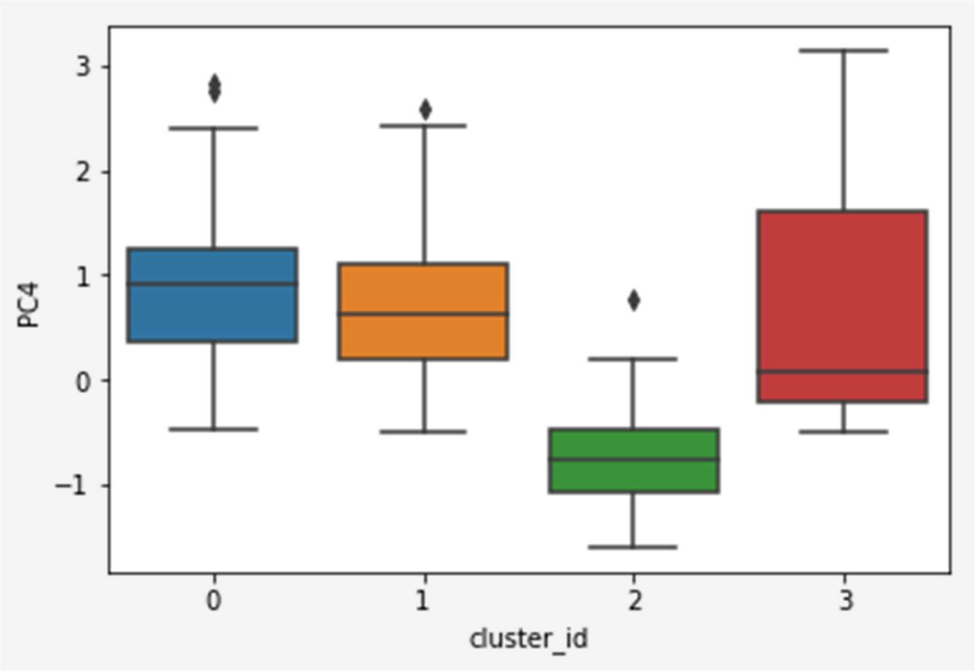
Explained variance ratio plot

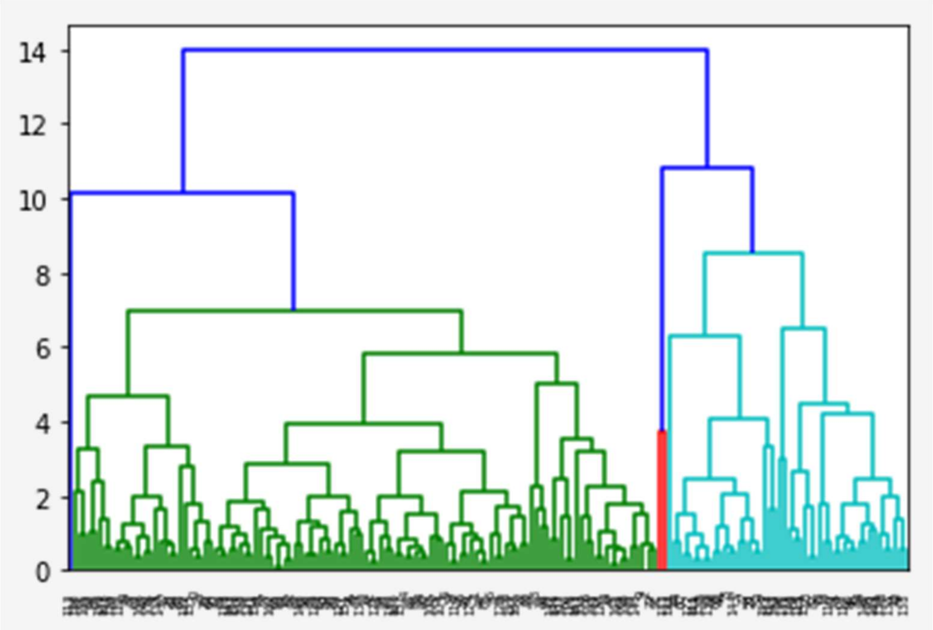





PC2
6
4
2
0
−2
0
1
2
3
cluster_id


PC2
6
4
2
0
−2
0
1
2
3
cluster_id


3
2
1
0
−1
PC4
0
1
2
3
cluster_id


14
12
10
8
6
4
2
0