

Pattern Recognition and Machine Learning

Minor Project

Report

Project-5

Team Members	Roll Number
1. Kalbhavi Vadhi Raj	B21EE030
2. Harsh Vardhan Singh	B21CS032
3. Aneesh Atul Borkar	B21EE079

Problem Statement:-

HELP International has been able to raise around \$ 10 million. The CEO of this NGO wants to donate this money to a country that needs this money.

We are given data from 167 countries and based on this data, we have to decide which country is in immediate need of money.

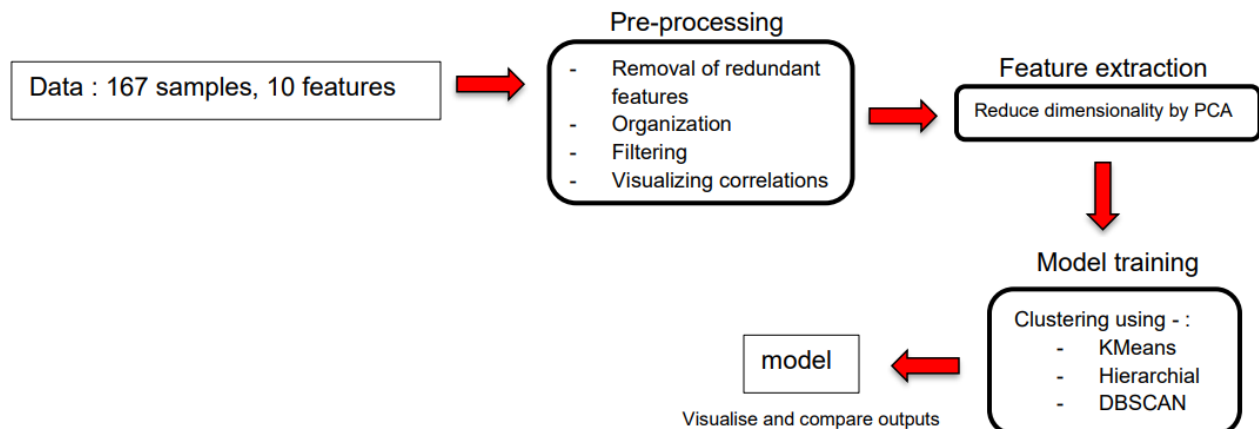


Figure 1: Machine Learning Pipeline

Exploratory Data Analysis:-

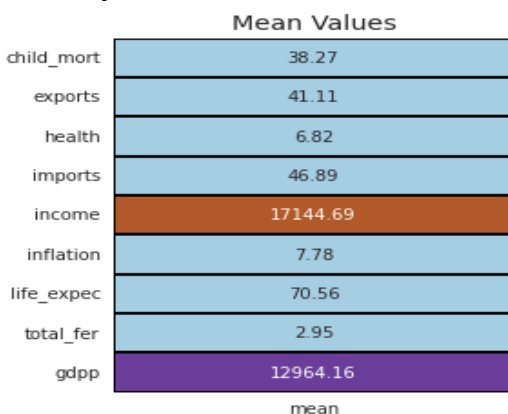
Given Data has 167 rows each representing a single country.
It has 10 columns:-

Column Name	Description
country	It has the names of countries for each row.
child_mort	It contains the deaths of children under 5 years of age per 1000 live births for each country.
Export	It gives each country's export in terms of a percentage of its GDP.
Health	It gives the amount spent on health as a percentage of its GDP.
Imports	It gives the imports of each country in terms of a percentage of its 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 newborn child would live.
total_fer	The number of children that would be born to each woman.
gdpp	The GDP per capita of each country.

Health Factors	child_mort, health, life_expec, total_fer
Trade Factors	imports, exports
Finance Factors	income, inflation, GDP

Using Pandas Describe Function to make a heat map of mean values

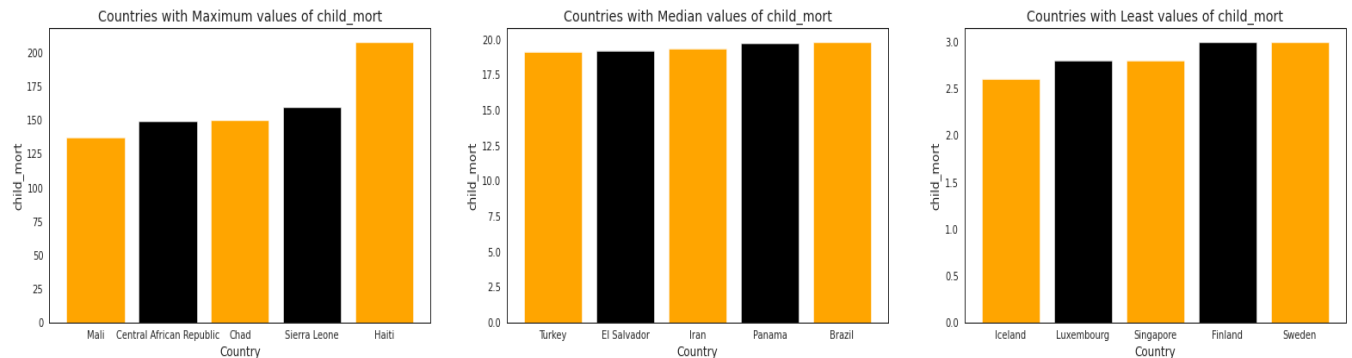
Heat Map:-



Based on Each feature making a Bar graph for Max, Mid and Min countries.

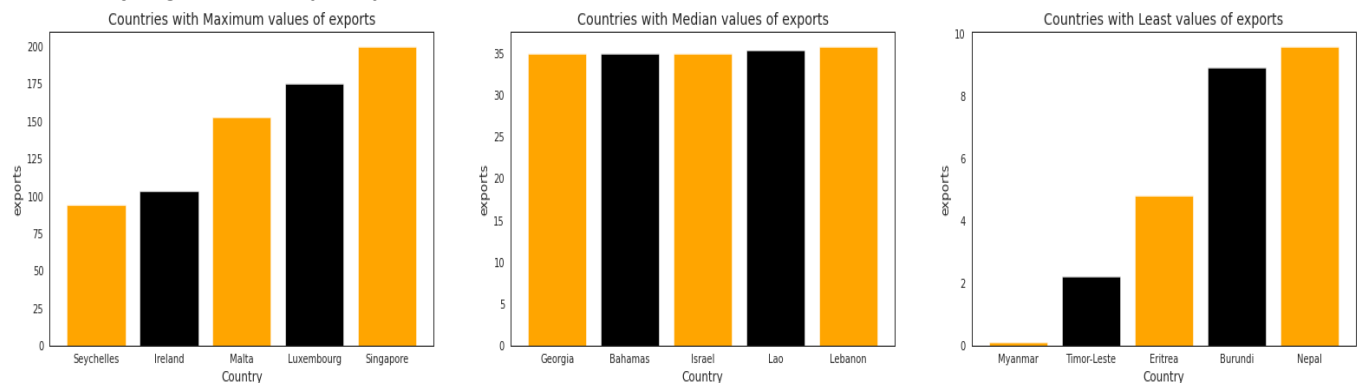
Child Mortality:-(Health Factor)

Countries with the least child mortality are developed and the ones with the maximum need help



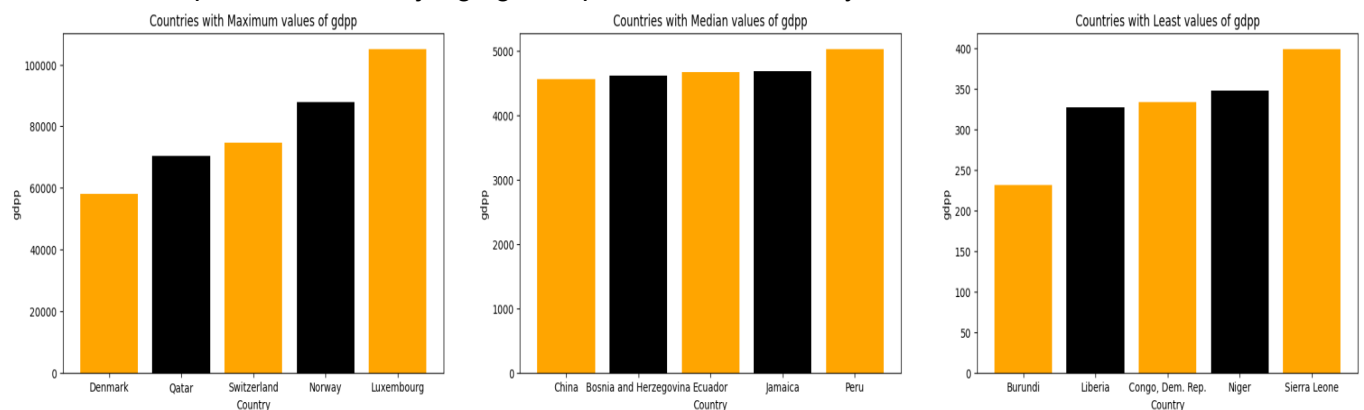
Exports:-(Trade Factor)

We can't judge a country only based on Exports.



GDPP:-(Finance Factor)

GDPP is an important factor for judging how poor or rich a country is.

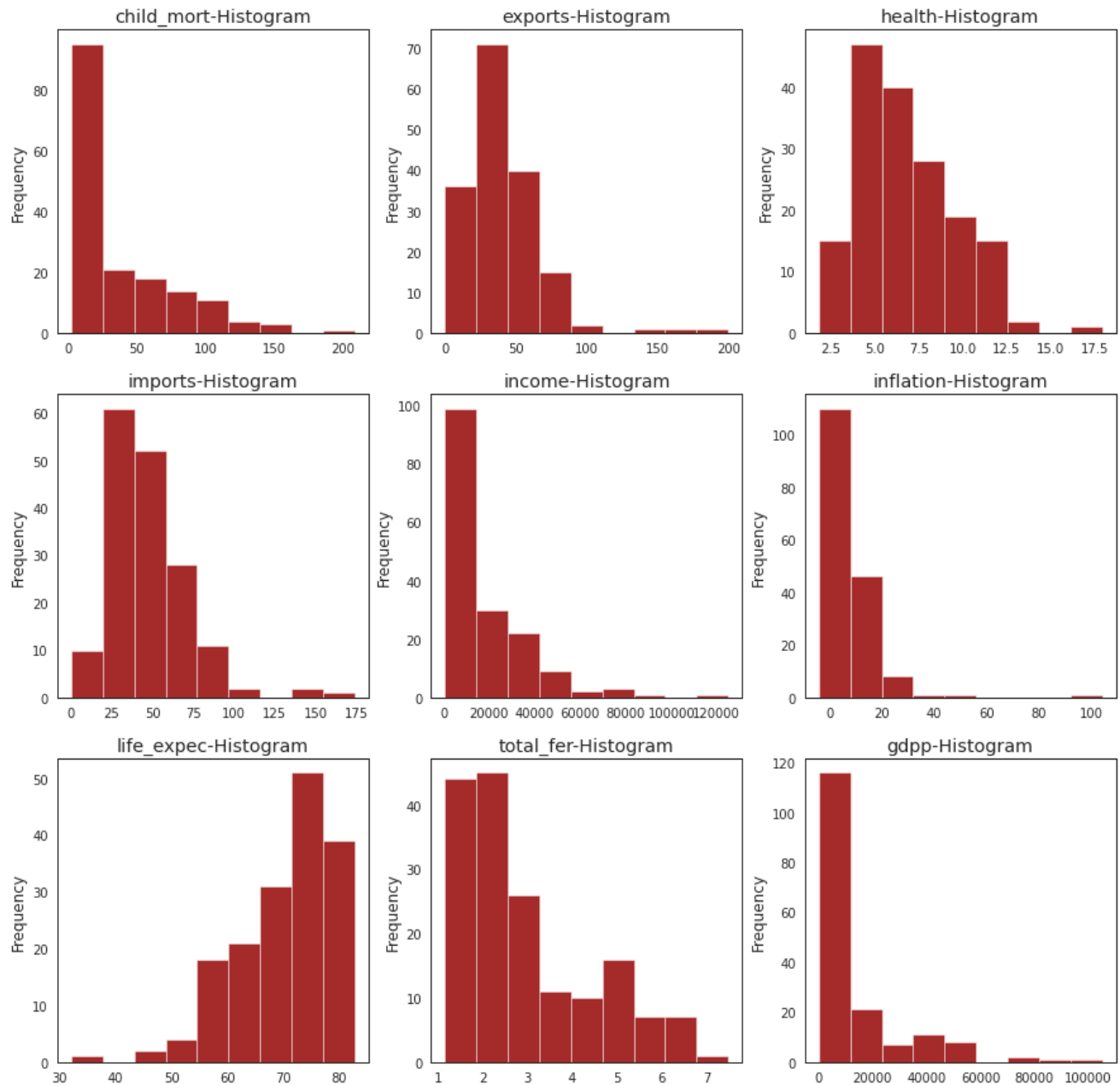


All these features are essential for judging the condition of a country but none of the features can alone define a country.

Hence we are using clustering to group similar countries based on these features.

Histogram of data concerning each feature

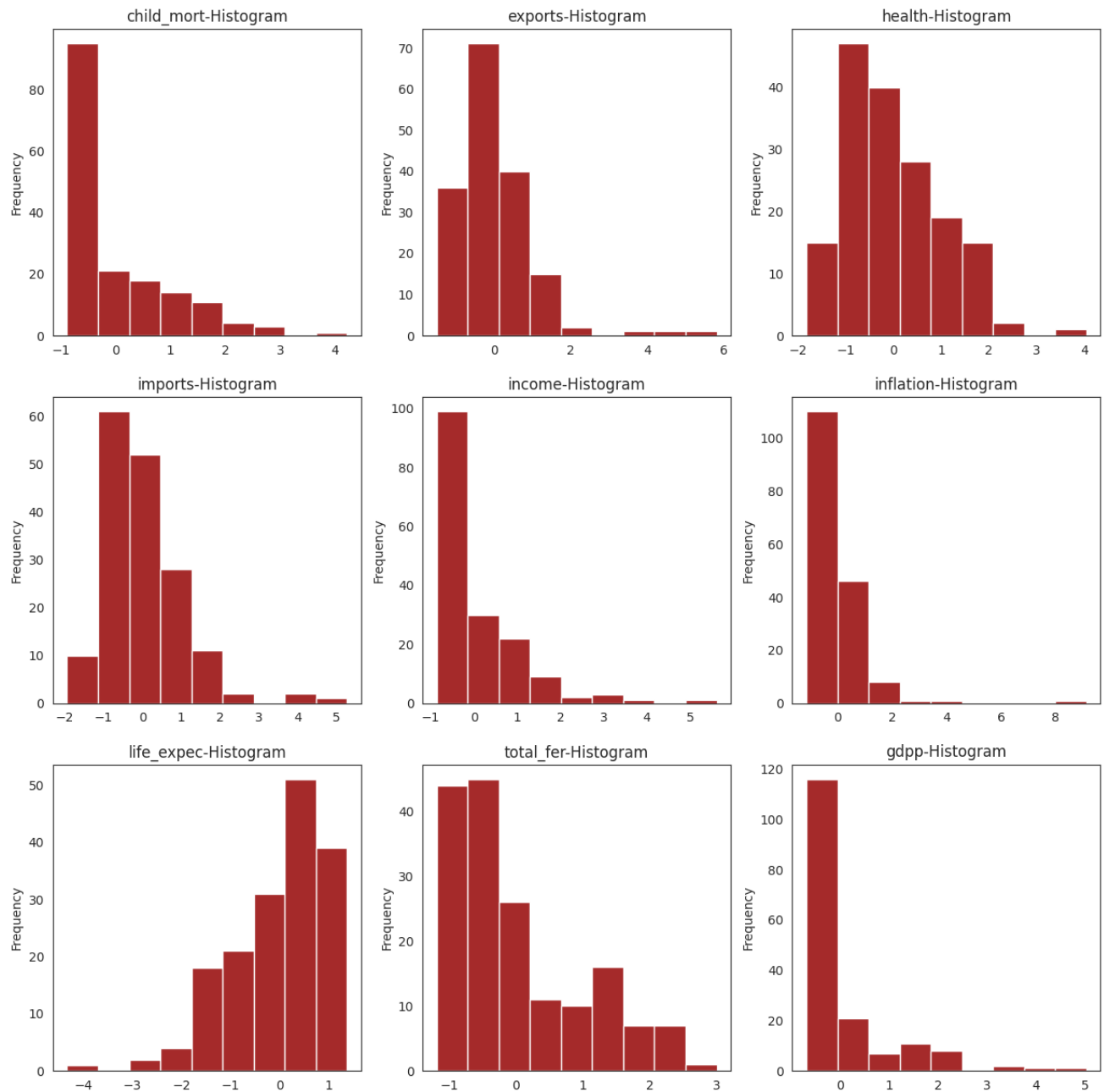
From this, we can understand the distribution of each feature and which has to be standardised or normalised.



As I intend to use PCA on this data for Feature selection, I standardise the data i.e. make the features zero-centered with variance=1.

Preprocessing:-

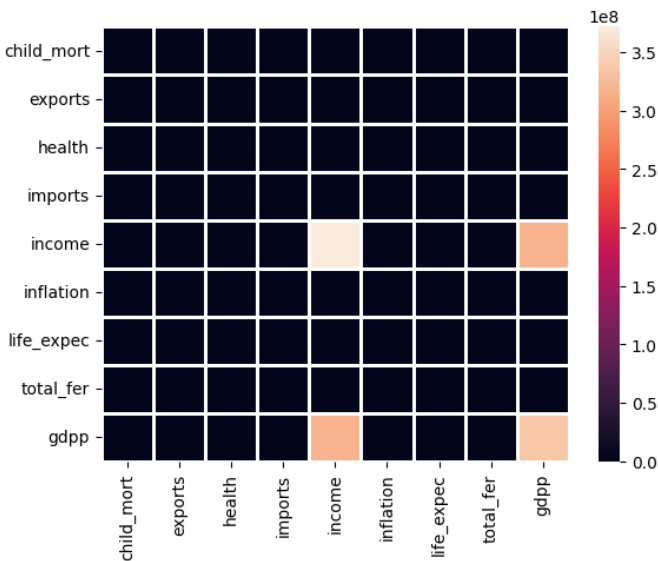
Histograms of each feature after standardisation.



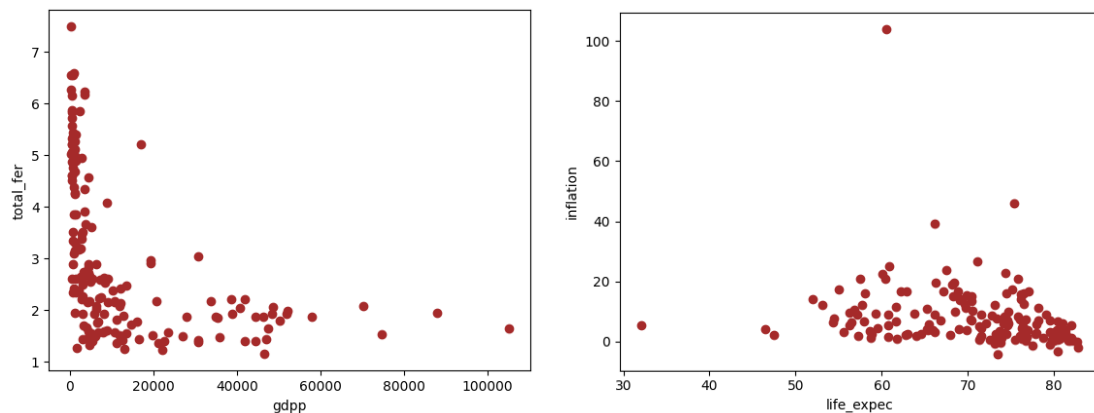
We can see that now the data is centered at zero.

Trying to Visualise the Outliers Based on features with the highest Correlation:-

Heat map for Covariance:-



Some of the plots:-



We can now clearly see the outliers in the data.

Some of these outliers are:

positive outliers:- ones like which high GDP, hence more on the developed countries' side.

Negative outliers:- Having higher values in negative features like child_mort. These outliers are the ones that need help.

Feature Selection:-

Using PCA to Select features:-

After Applying PCA to the data I took all the features which accounted for nearly 90% of the total variance.

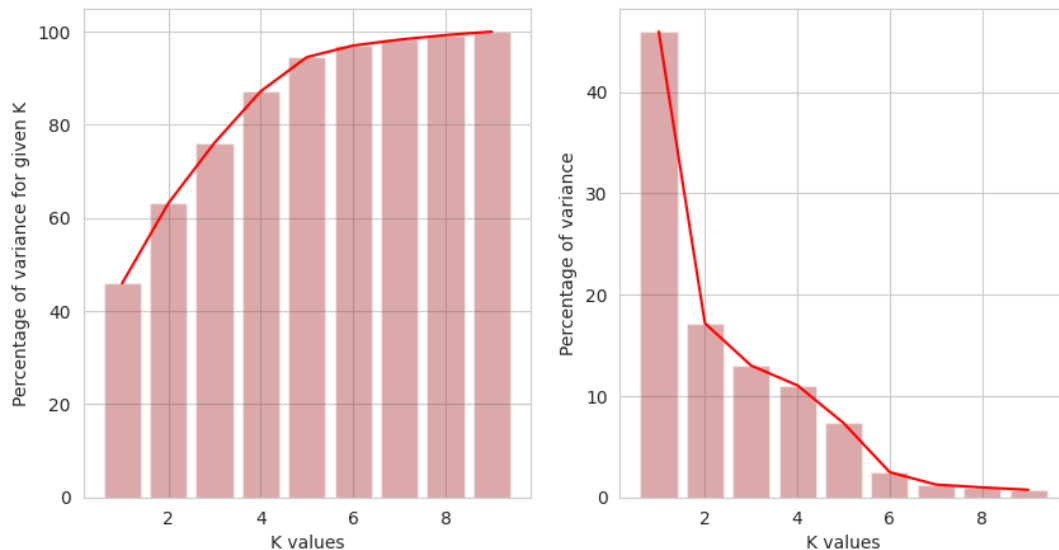
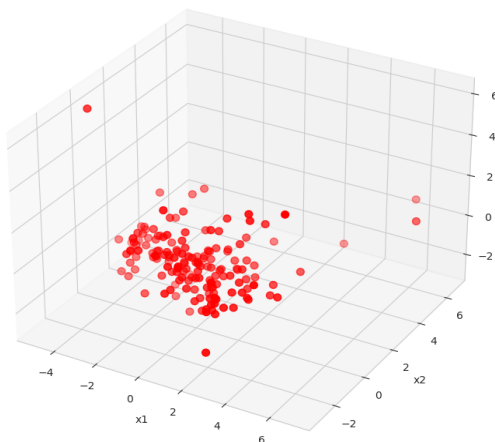


Figure:- (left) Percentage of variance for k features selected, (right) Percentage of variance for a single feature

So now have I reduced the number of features from 9 to 5.

Note:- PCA doesn't help in clustering better, but what it does is just decrease the training time and also makes it easier to visualise the data easily.

3D Scatter after PCA:-



Model Building:-

K-means clustering:-

On original Data

Using Elbow Method and silhouette score method to find the optimal number of Clusters

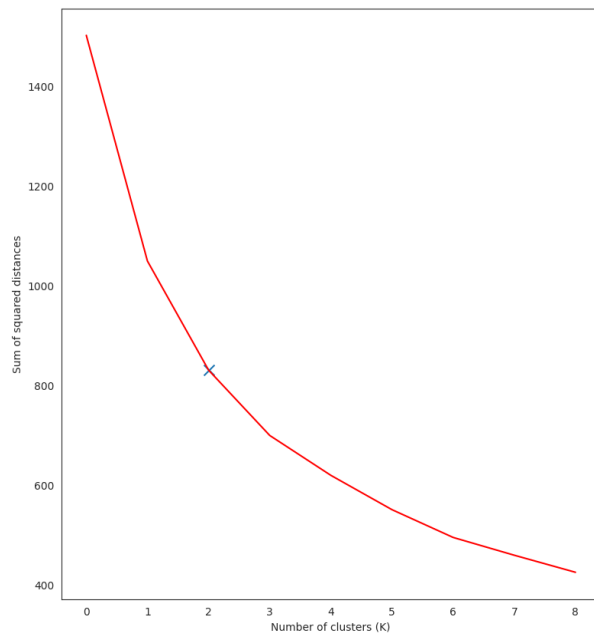


Figure:- Elbow Method

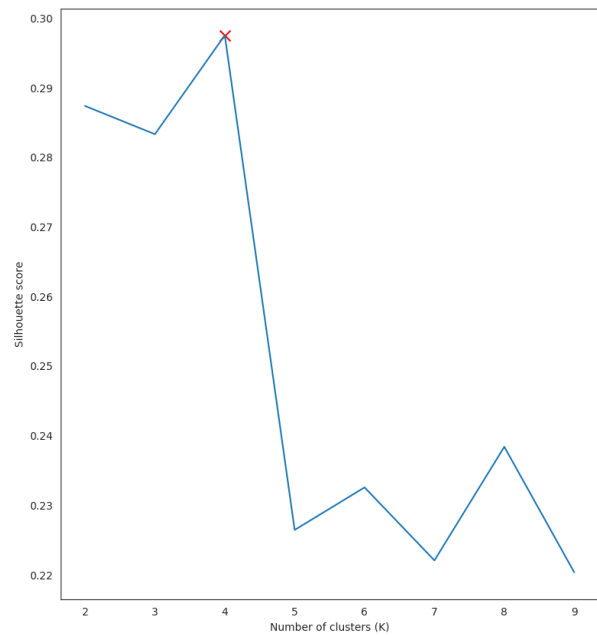
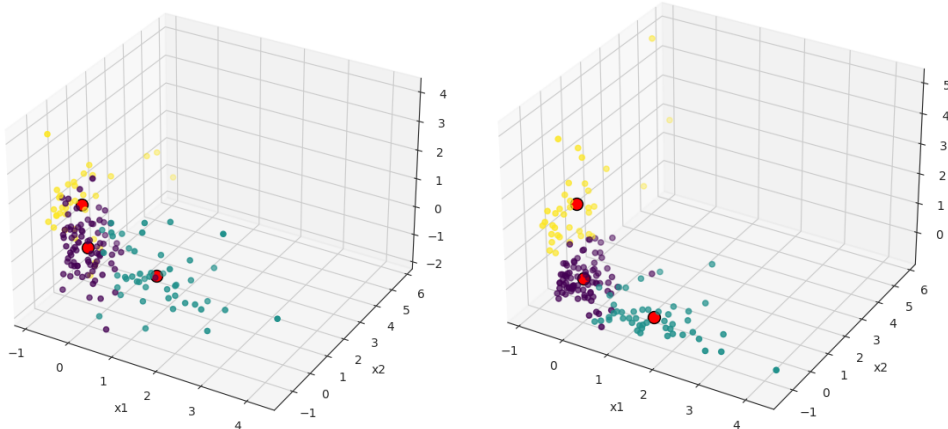


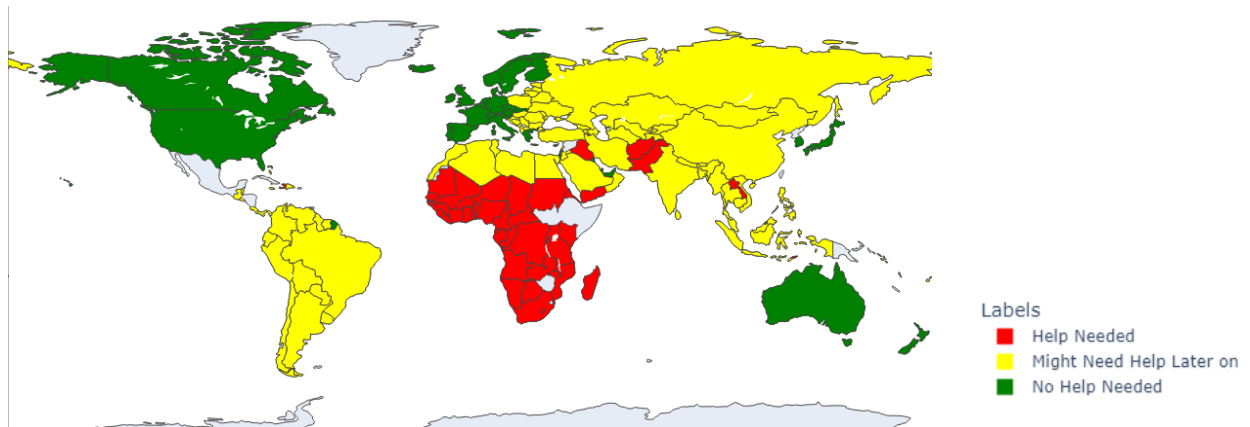
Figure:- Silhouette score maximisation

Based on the output of each method I am taking the Average value i.e. $k=3$

Visualising Clusters formed by K-means clustering:-



Final Output of the K-means on the original Data:-



K-means clustering:-

On PCA Data:-

Using Elbow Method and silhouette score method to find the optimal number of Clusters

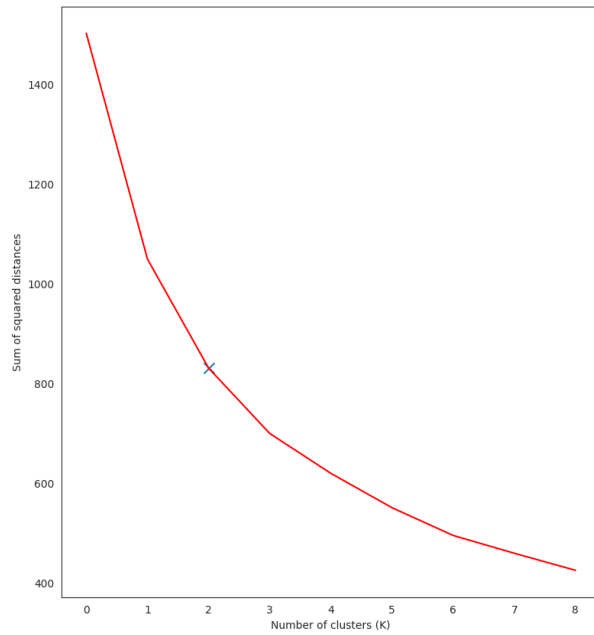


Figure:- Elbow Method

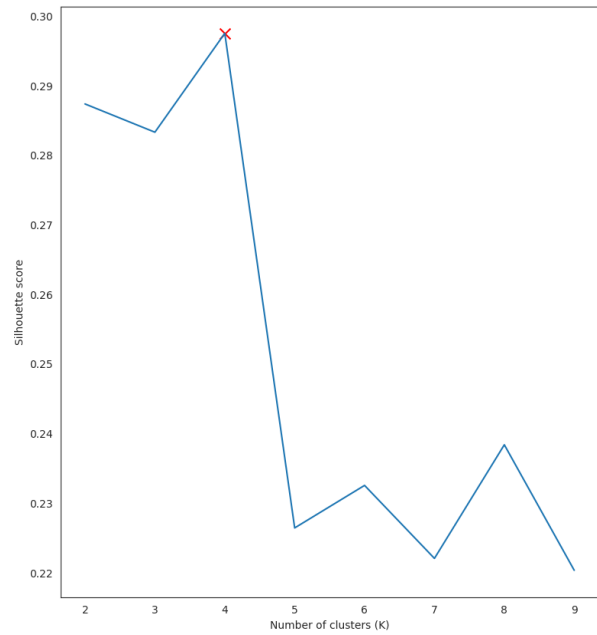
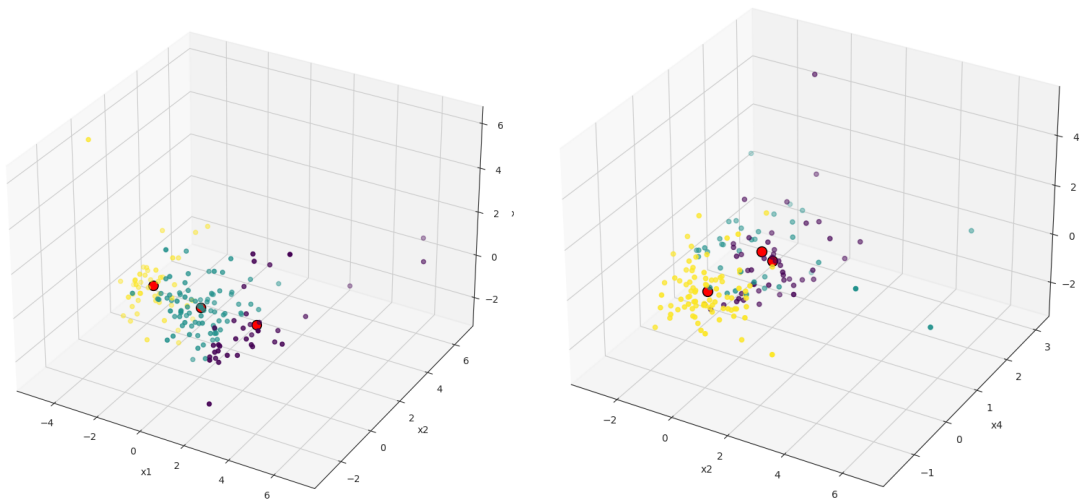
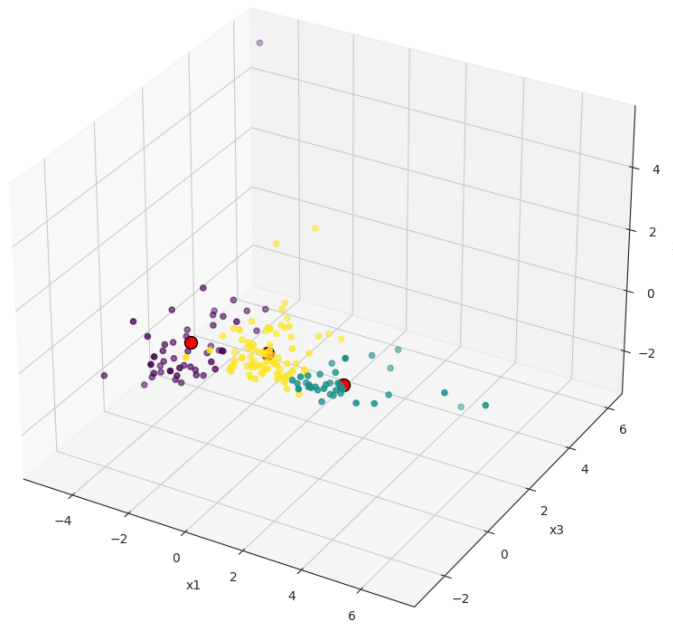


Figure:- Silhouette score method

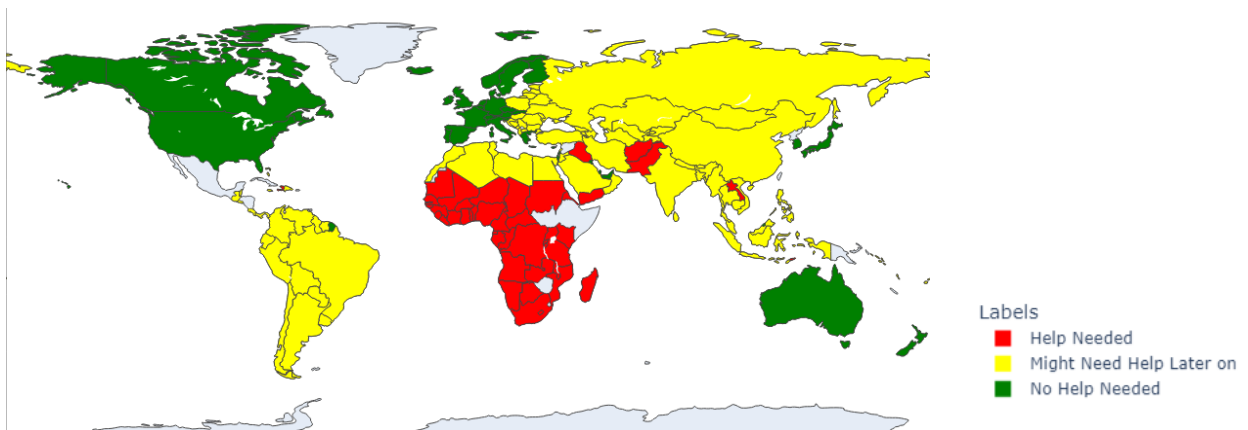
Based on the output of each method I am taking the Average value i.e. $k=3$

Visualising Clusters formed by K-means clustering:-





Final Output of the K-means on the PCA Data:-



DBscan:-

DBscan is a density-based clustering algorithm.

Hyperparameters:-

- **minPts** : It is the minimum number of data points that need to be present in the area of a point to be considered as a core point.
- **Epsilon(Eps)**: It is the radius of the area of a centre point.

On Original Data:-

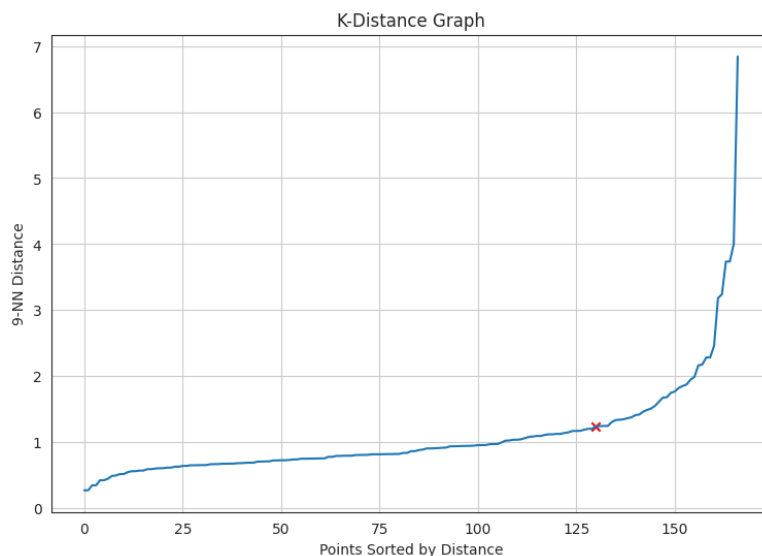
Deciding minPts:-

Generally, we take $\text{minPts} \geq n_{\text{features}} + 1$ or $\text{minPts} = 2 * n_{\text{features}}$ for noisy and small datasets.

Here we have $n_{\text{features}} = 9$ hence taking $\text{minPts} = 10$

Now using KNN I am finding the value of eps

Graph for KNN (k=9):-

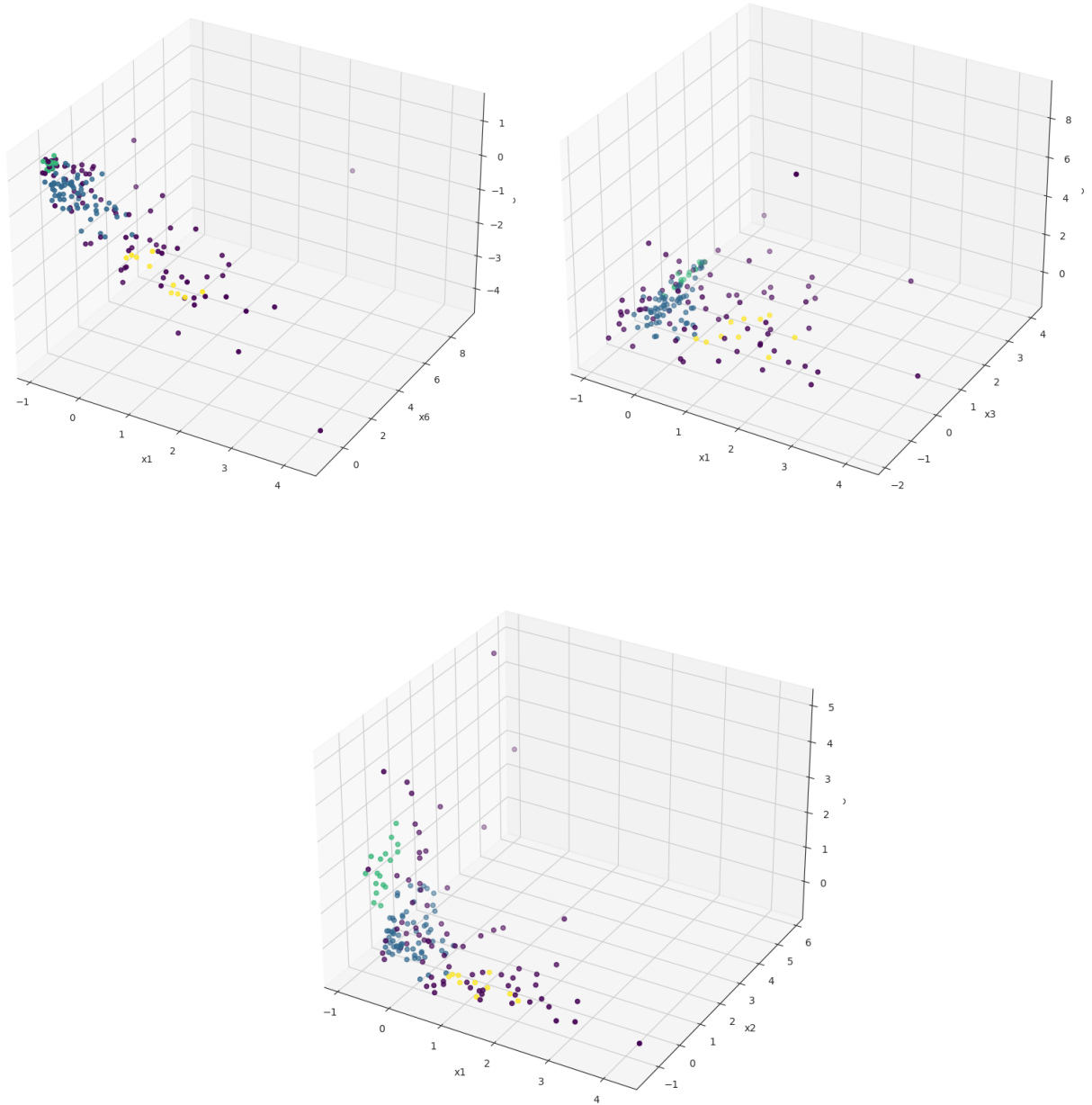


We get $\text{eps} = 1.24$

Observation:- Distance at which there is an abrupt change in the value of 10-NN distance that is taken as the value of the eps.

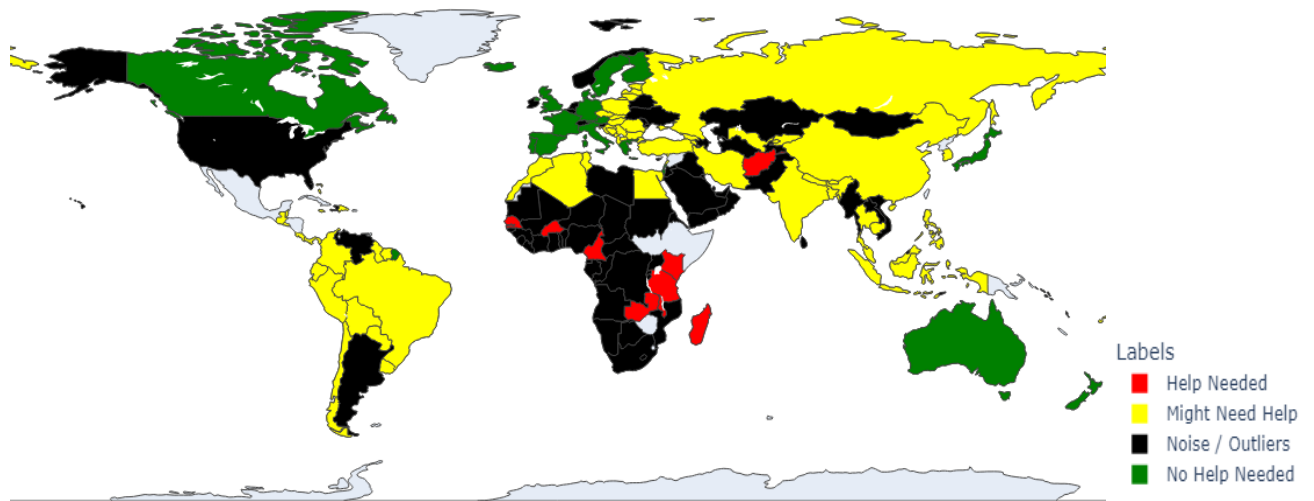
Applying DBSCAN(minPts=10,eps=1.24)

3D scatter plots of the cluster formed by DBSCAN:-



Here we are getting 3 clusters with noise.

Final Output using DBSCAN on Original Data:-



Observations:- With DBSCAN we were able to remove both the positive and negative outliers and tell the result on the remaining data.

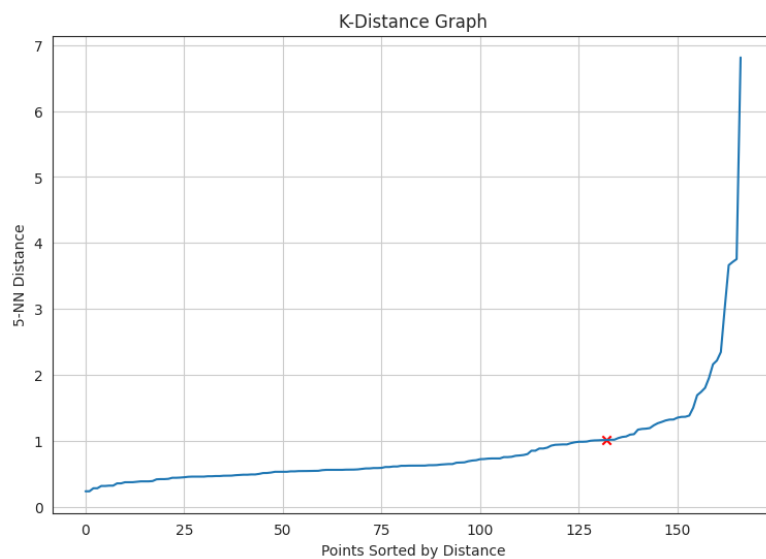
On PCA Data:-

Deciding the value of eps and minPts:-

Here $n_{\text{features}}=5$ hence $\text{minPts}=6$

Using KNN to find the eps value.

Graph of KNN (k=5):-



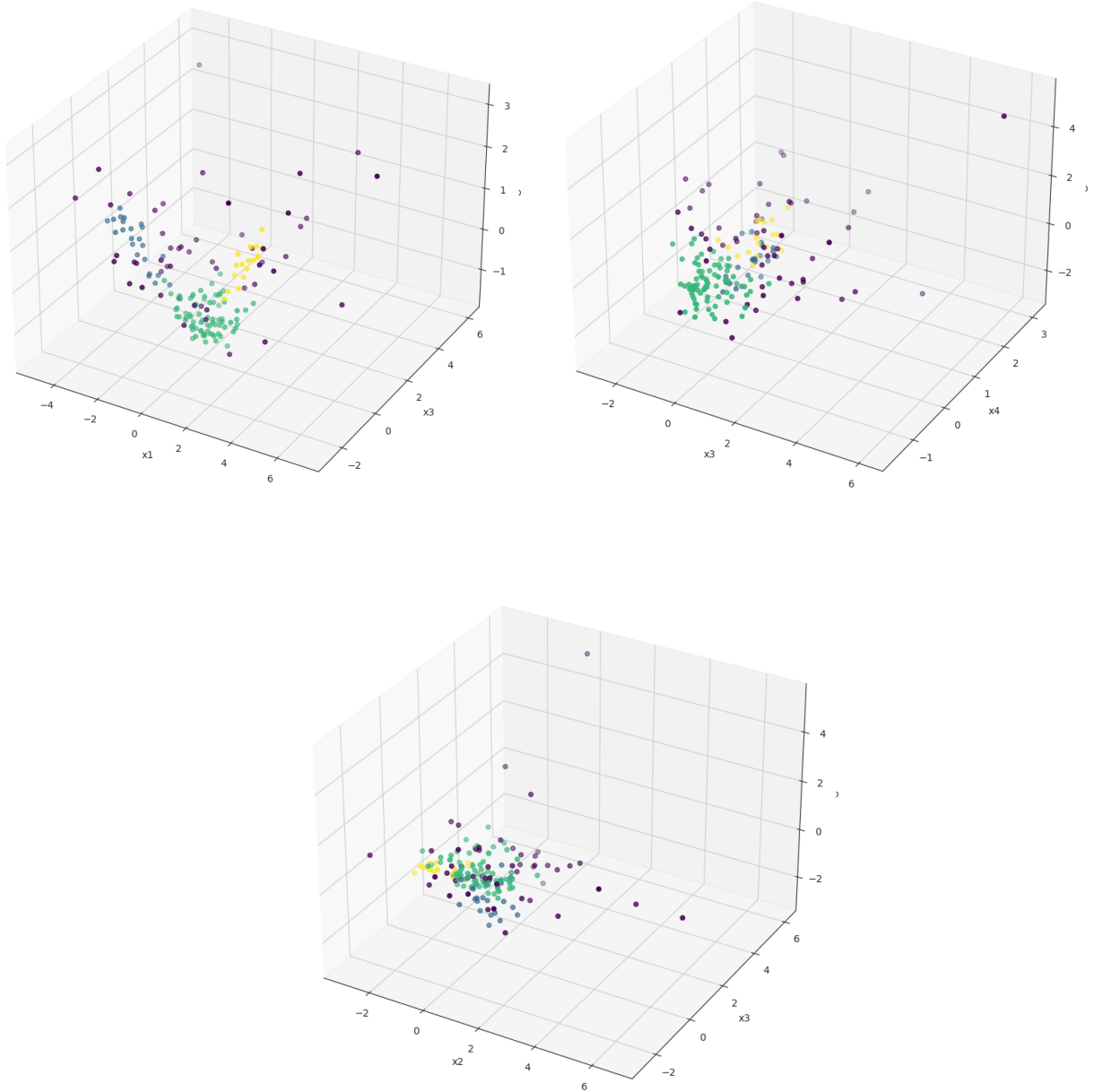
Here we get $\text{eps}=1.013$

Observation:- Distance at which there is an abrupt change in the value of 10-NN distance that is taken as the value of the eps.

Here we got a different value of eps when compared to the Original Data.

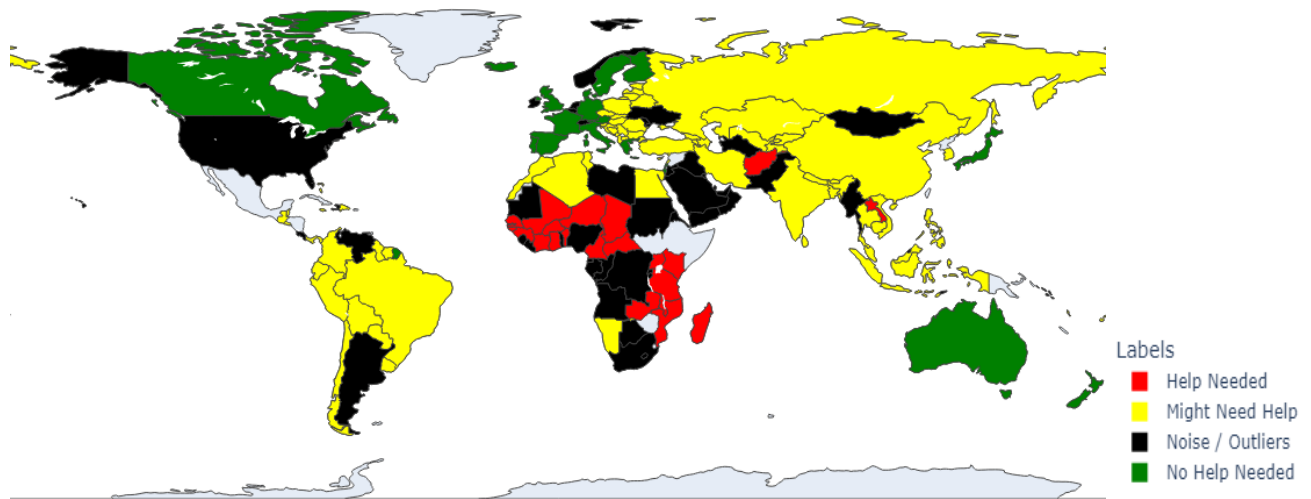
Applying DBSCAN(minPts=6,eps=1.013)

3D scatter plots of the cluster formed by DBSCAN:-



Here also we get 3 clusters and noise.

Final Output using DBSCAN on PCA Data:-



Observations:- With DBSCAN we were able to remove both the positive and negative outliers and tell the result on the remaining data. However, there is a slight difference between the PCA and original data outputs.

Hierarchical Clustering:-

Hierarchical Clustering is a distanced based algorithm that is used for unsupervised learning problems.

It develops the hierarchy of clusters in the form of a tree i.e known as the **dendrogram**

On the Original Data:-

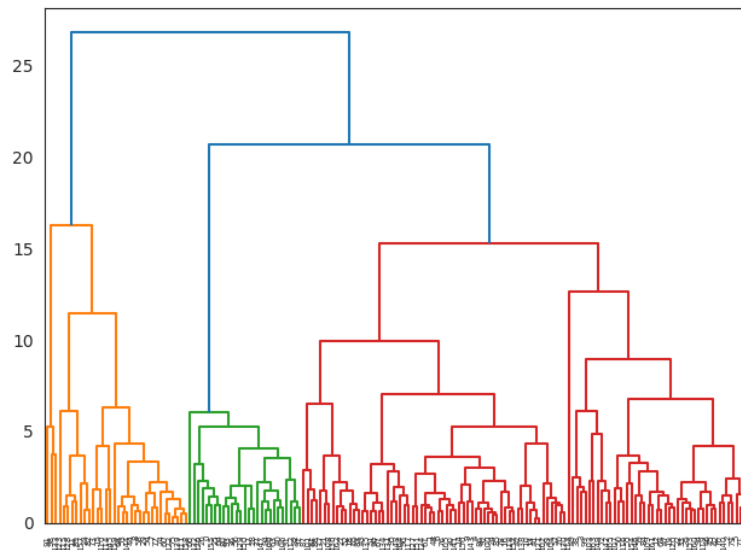
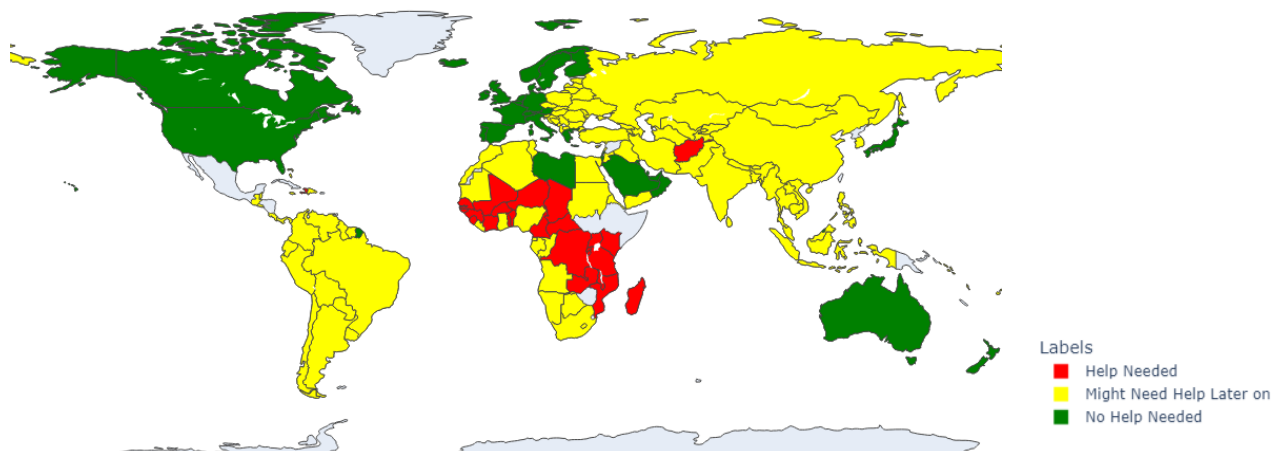


Figure:-Dendrogram

Final Output of Hierarchical clustering on the original data:-



On PCA Data:-

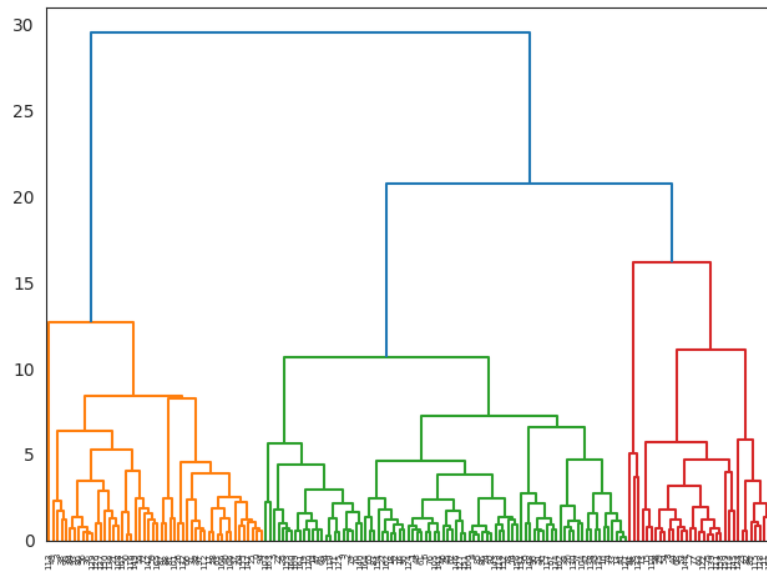
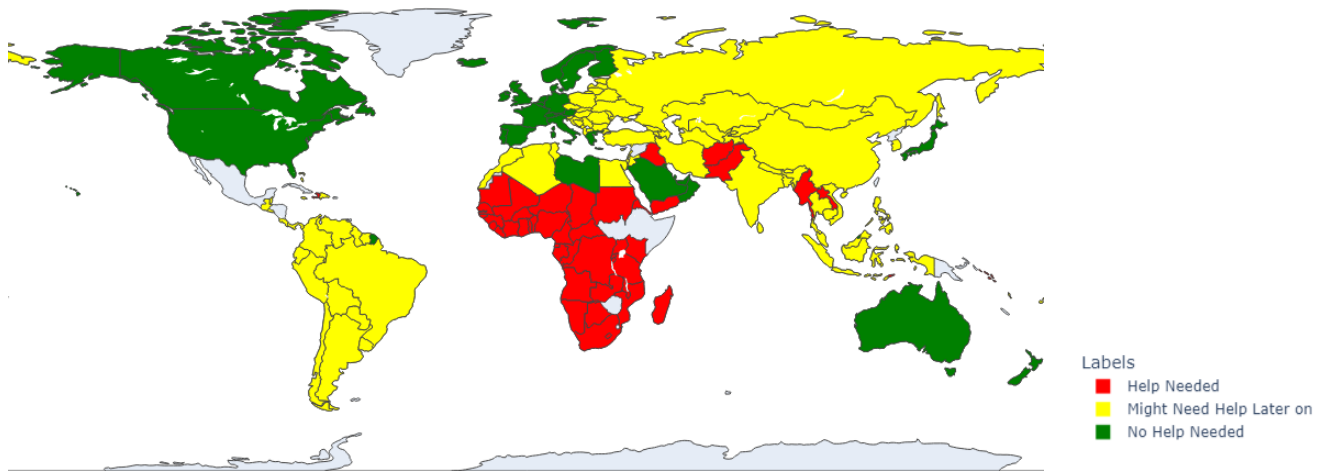


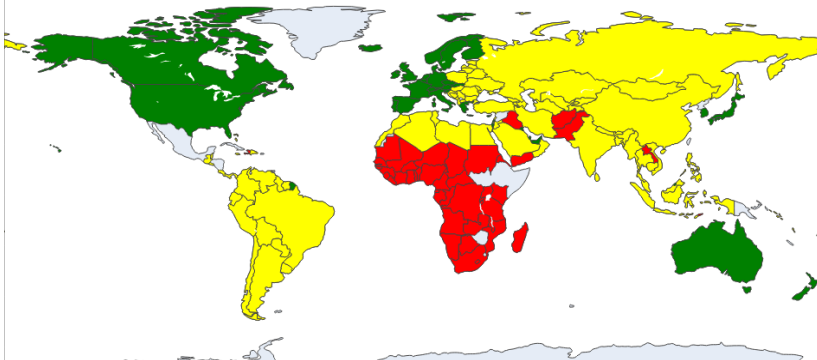
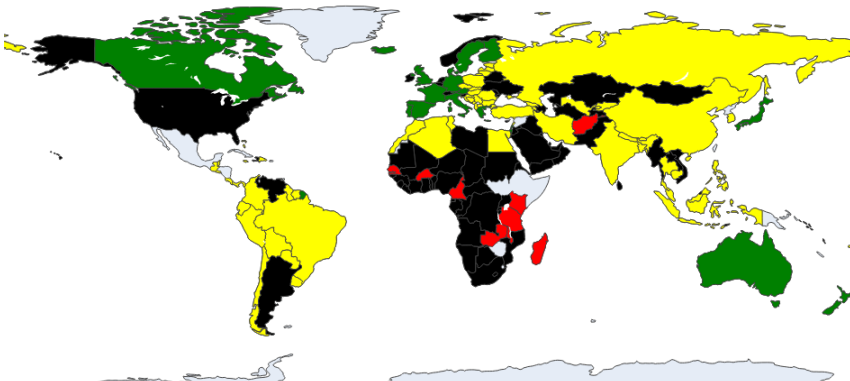
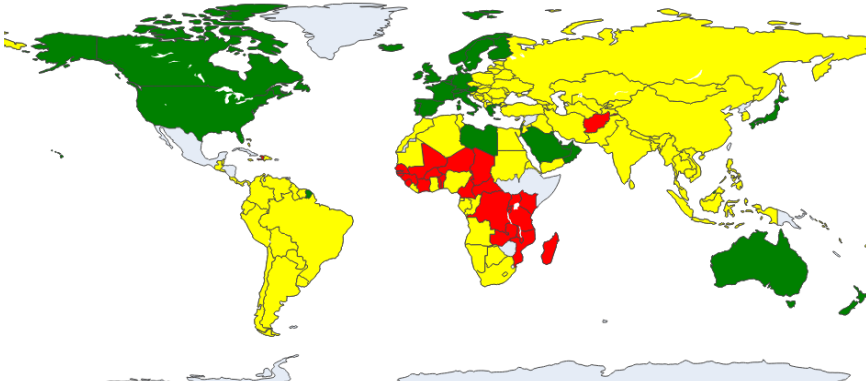
Figure:-Dendrogram

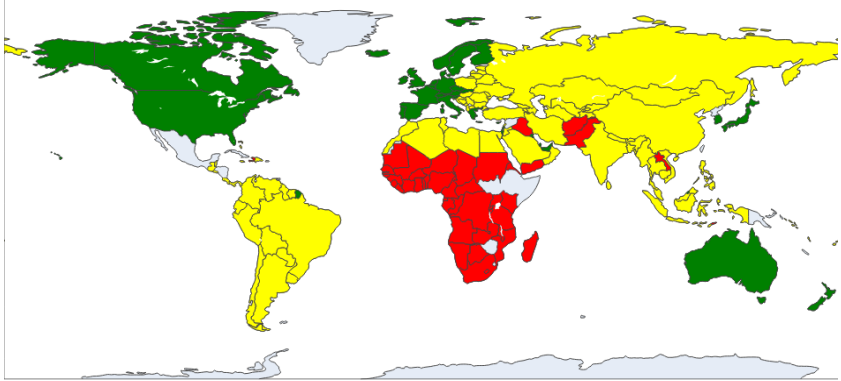
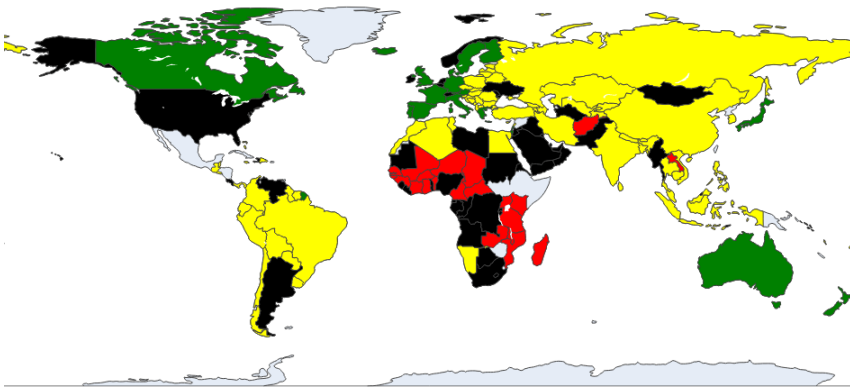
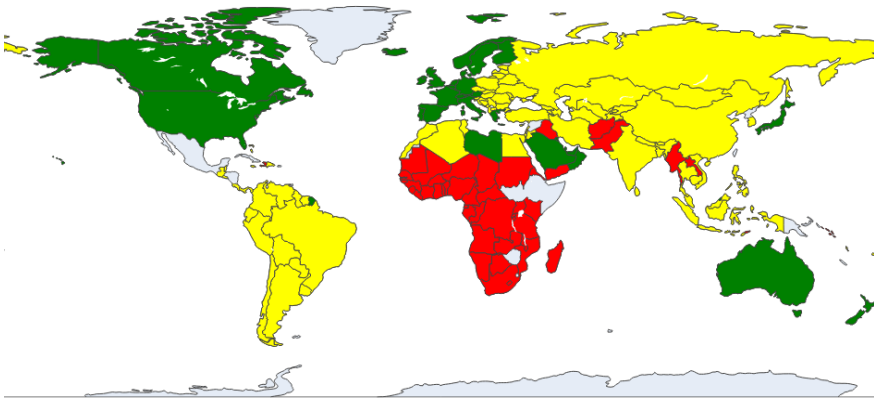
Final Output of Hierarchical clustering on PCA data:-



Conclusion:-

On Original Data

Model	Final Output
K-means	
DBscan	
Hierarchical	

Model	Final Output
K-means	
DBscan	
Hierarchical	

Observation:-

There is no difference between Original and PCA data in the case of K-Means clustering.

However in the case of DBscan and Hierarchical, There is a considerable difference between the output on PCA and Original Data.