PRML MINOR PROJECT

Country Data

Honey Solanki (B21EE068)
Jatin Kumar (B21EE027)
Lokesh Chaudhari (B21CS041)

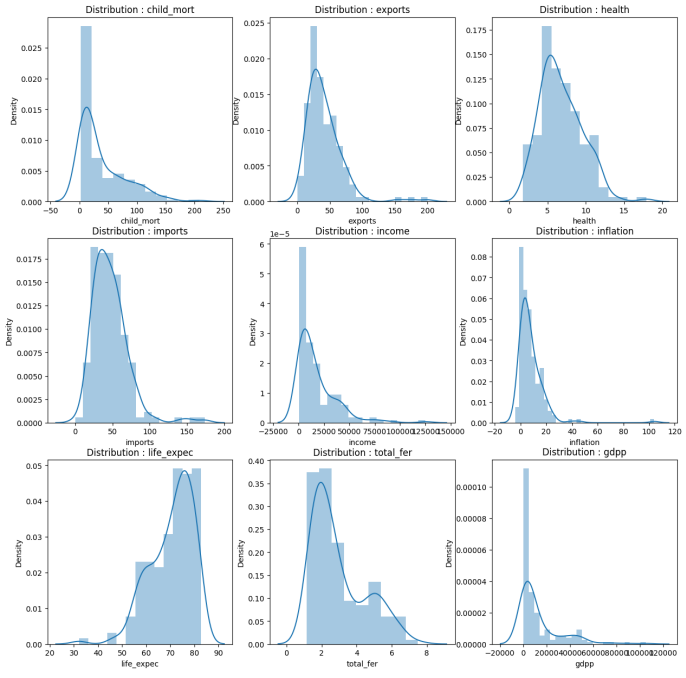
Data Preprocessing

- ➔ Imported necessary libraries namely numpy, pandas, seaborn, matplotlib, sklearn etc.
- ➔ Read the dataset using read_csv
- ➔ Gather required information of the dataset like shape, columns, etc.
- ➔ Checked if any null value is present in the dataset and found out no null values are are present

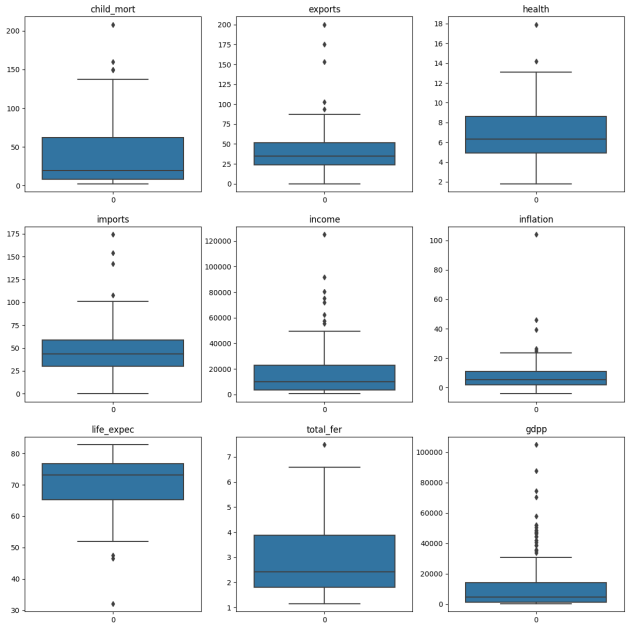
EDA

- ➔ For this dataset, as the number of features are less, we manually check the dataset.
- ➔ Except country, all the features are numerical features with their element data type being either float or integer

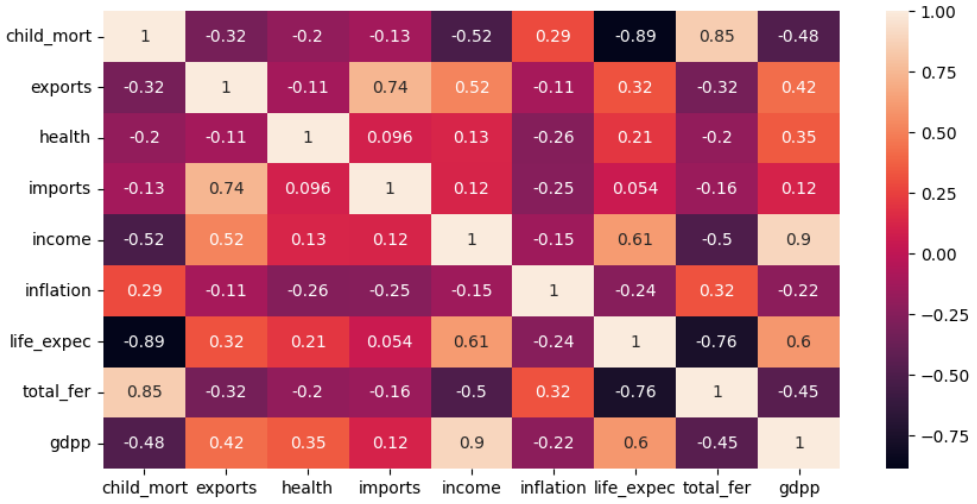
Distribution of features :



Boxplot :



Heatmap : Plotting heatmap to gain information about correlation between features



We can clearly see that some features are essentially from the same category and they have the same reaction to other features of different categories.

➔ The 3 categories of the features are :

- health : child_mort, health, life_expec, total_fer
- trade : imports, exports
- finance : income, inflation, gdpp

➔ Hence, we will dissolve these features into these categories and normalize them!

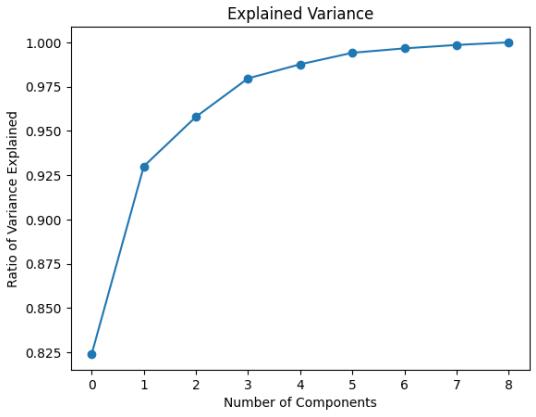
New Dataframe from feature combination :

	Country	Health	Trade	Finance
0	Afghanistan	0.625740	0.139614	0.079820
1	Albania	0.127451	0.199901	0.088756
2	Algeria	0.182485	0.186622	0.212808
3	Angola	0.661381	0.283058	0.236946
4	Antigua and Barbuda	0.116409	0.275189	0.145043

Principal Component Analysis (PCA) :

➔ It is a dimension reduction method that is preferably used for an Unsupervised Learning Problem.
➔ We copied the original dataset to a new dataset namely pca_df.
➔ We standardized the health column and normalized all the remaining columns. Also before standardization and normalization we dropped the country column.

Variance explained vs No. of components :



➔ Typically a number of components with more than 95% of ratio of variance are selected.
➔ In this case, we select the n=2 as the steps generated have significant variances and thus the other features get dominated by their variances.

PCA Data :

	0	1	2
0	0.220482	0.640048	0.086112
1	-0.080903	-0.172000	-0.208695
2	-0.961283	-0.125199	-0.126719
3	-1.505914	0.472144	0.284978
4	-0.264724	-0.237308	-0.059150

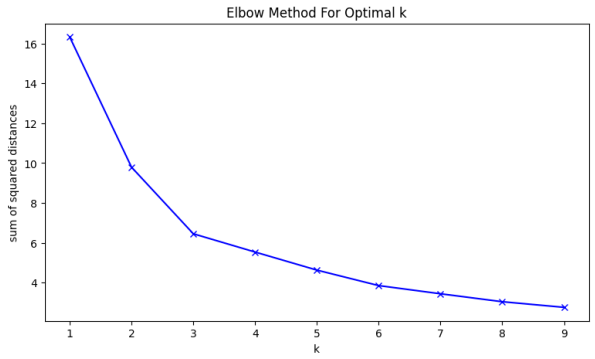
And hence we reduced dimensions!

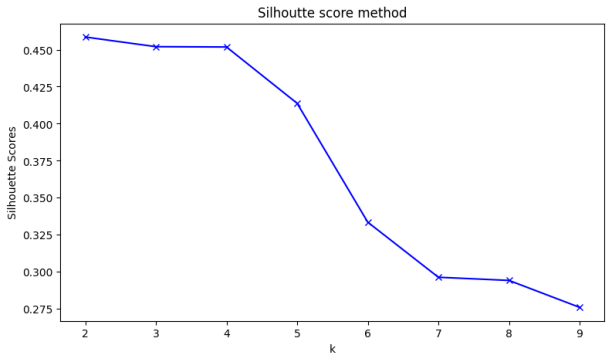
K-Means Clustering :

We performed **K-Means clustering simultaneously on two datasets,** first obtained by feature combination and another by PCA.

Normal Data (Feature Combination) :

We used the elbow method and silhouette score method to obtain an optimal value of k (number of clusters).

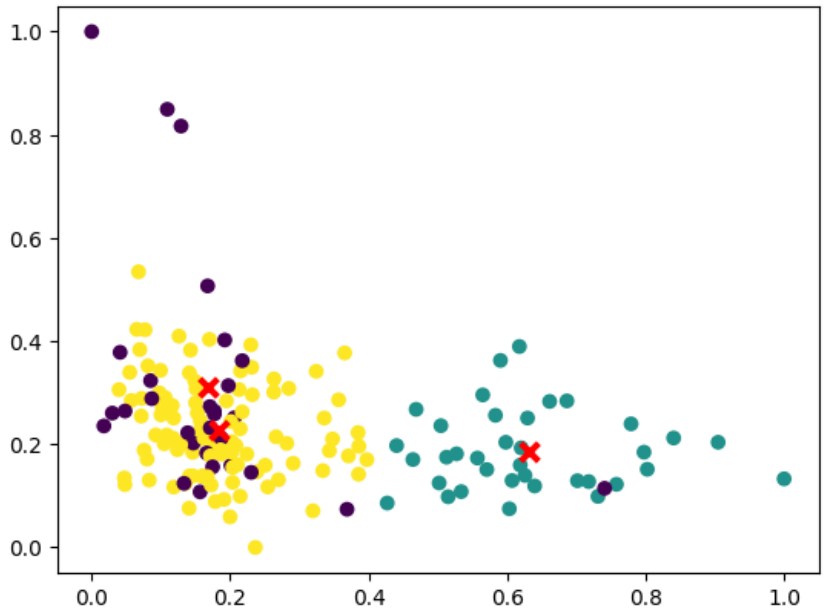




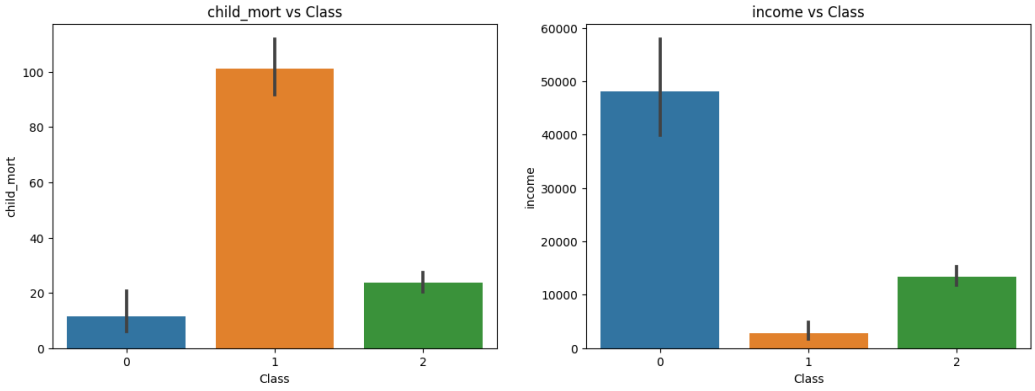
From the results of the above 2 methods, we select :

➔ **k : Clusters = 3**

Assigned clusters with their centroids



- ➔ Now we have got the clusters but we don't know which value corresponds to what!
- ➔ Hence, we draw a barplot of income & child_mort w.r.t labeled clusters.
- ➔ As we know, low income and high child mortality is a sign of an economically backward nation.

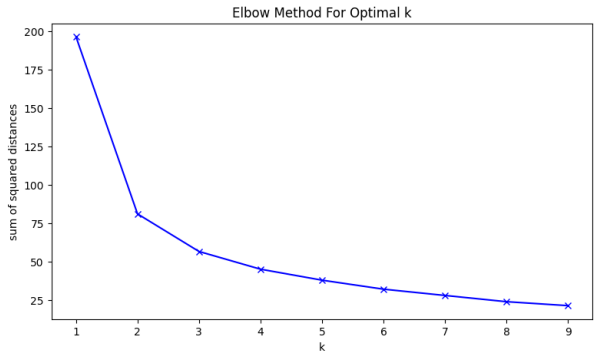


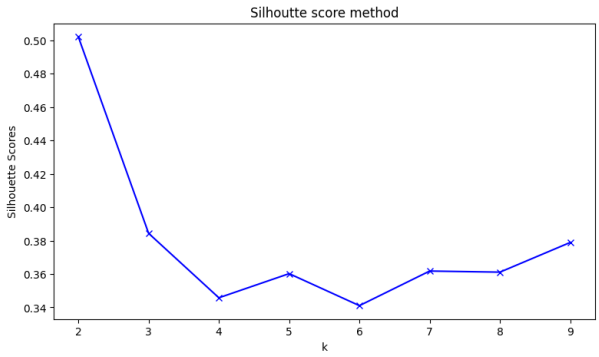
From above plot we can see that

- ➔ **0 :** No Help Needed **(No)**
- ➔ **1 :** Help Needed **(Yes)**
- ➔ **2 :** Might Need Help **(Might)**

PCA Data :

We used the elbow method and silhouette score method to obtain an optimal value of k (number of clusters).

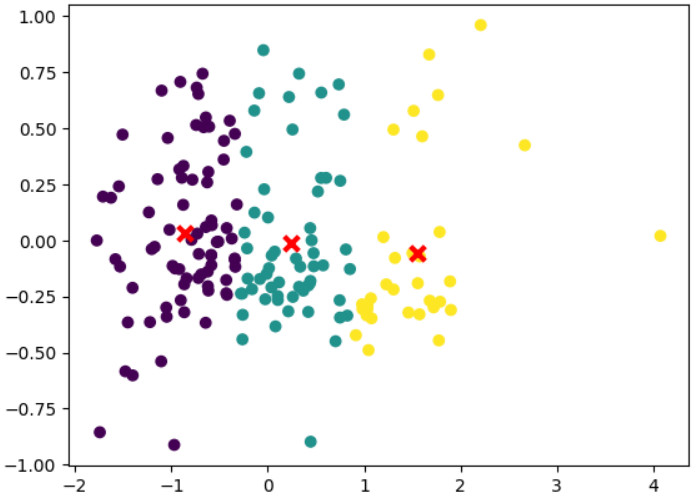




From the results of the above 2 methods, we select :

➔ **k : Clusters = 3**

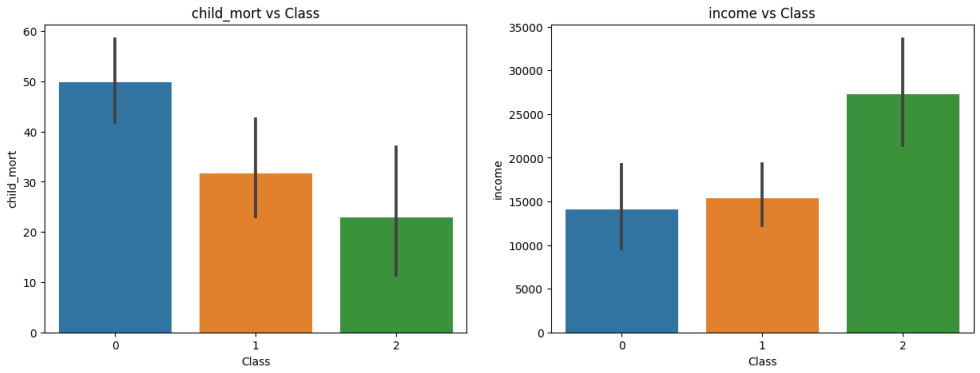
Assigned clusters with their centroids



➔ Now we have got the clusters but we don't know which value corresponds to what!

➔ Hence, we draw a barplot of income & child_mort w.r.t labeled clusters.

➔ As we know, low income and high child mortality is a sign of an economically backward nation.



From above plot we can see that

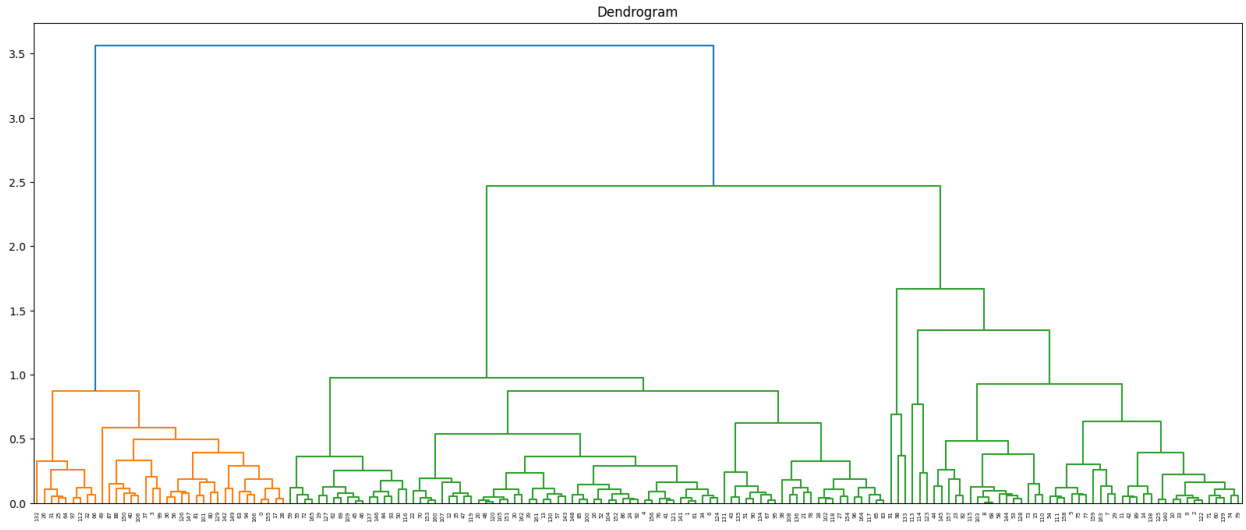
- ➔ **0 :** Help Needed **(Yes)**
- ➔ **1 :** Might Need Help **(Might)**
- ➔ **2 :** No Help Needed **(No)**

Hierarchical Clustering :

We performed **Hierarchical clustering simultaneously on two datasets,** first obtained by feature combination and another by PCA.

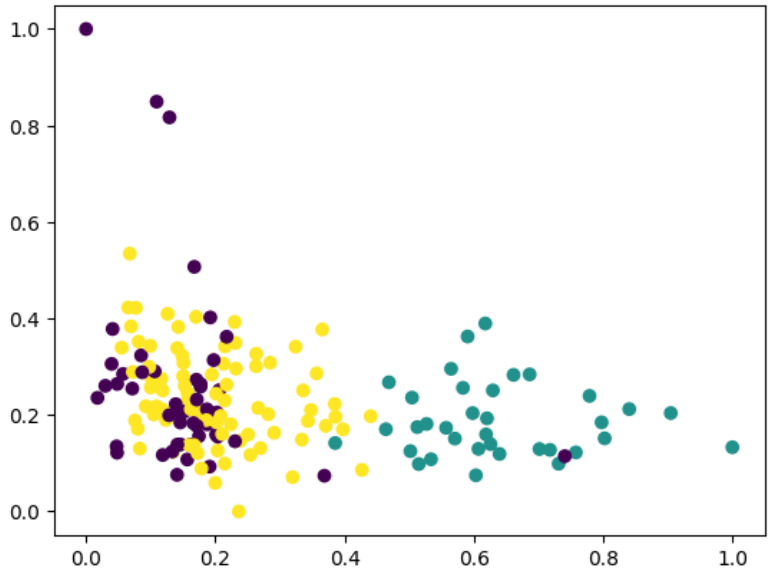
Normal Data (Feature Combination) :

Dendrogram

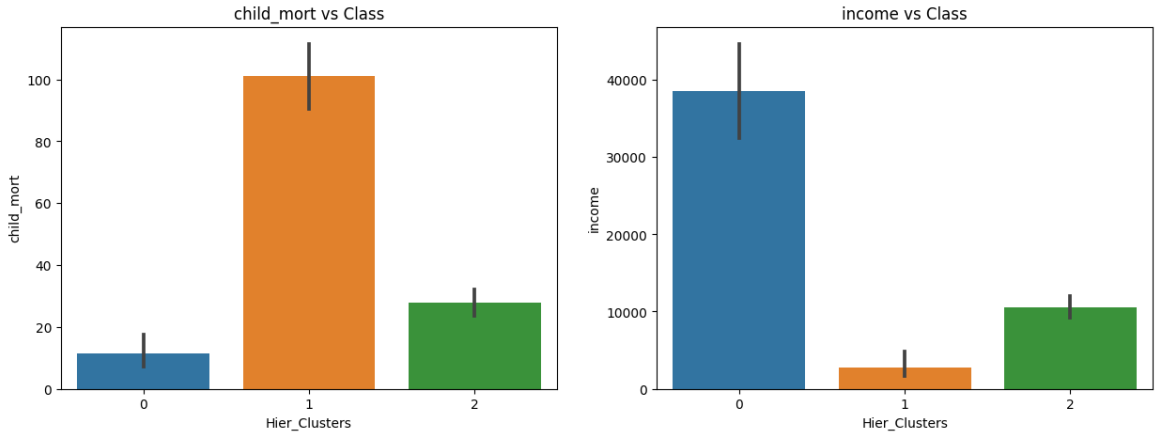


We can see that it has 3 branches hence the **number of clusters to be formed are 3.**

Scatterplot with assigned clusters



- ➔ Now we have got the clusters but we don't know which value corresponds to what!
- ➔ Hence, we draw a barplot of income & child_mort w.r.t labeled clusters.
- ➔ As we know, low income and high child mortality is a sign of an economically backward nation.

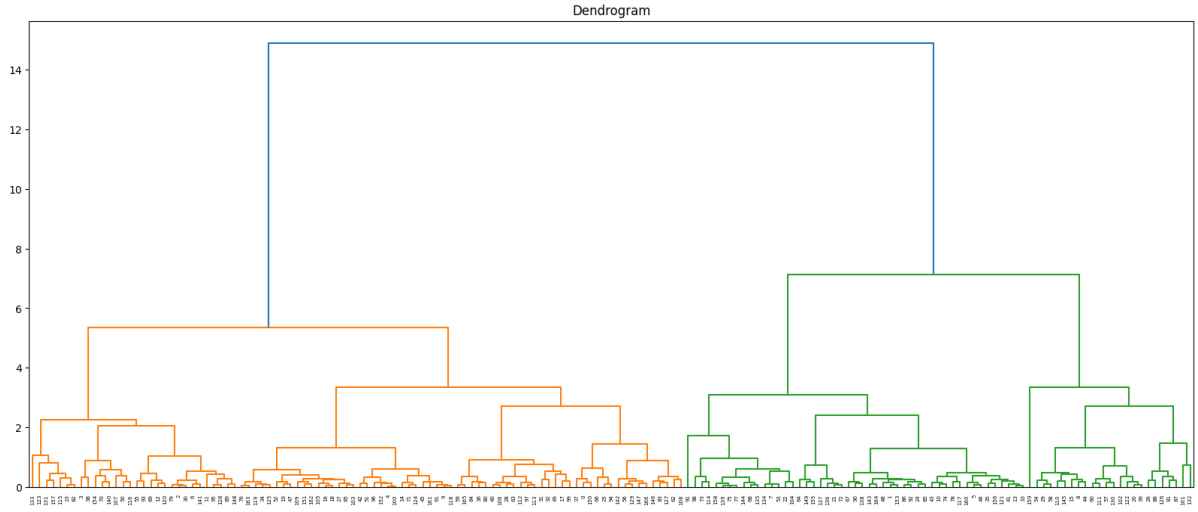


From above plot we can see that

- ➔ **0 :** No Help Needed **(No)**
- ➔ **1 :** Help Needed **(Yes)**
- ➔ **2 :** Might Need Help **(Might)**

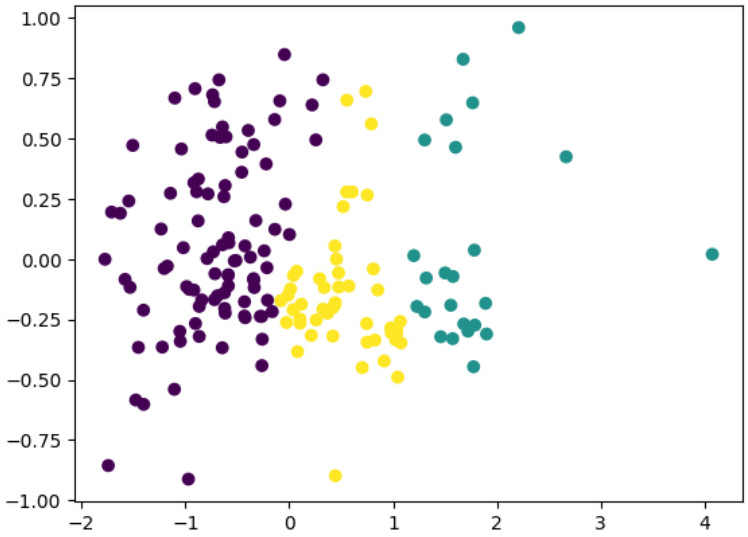
PCA Data :

Dendrogram

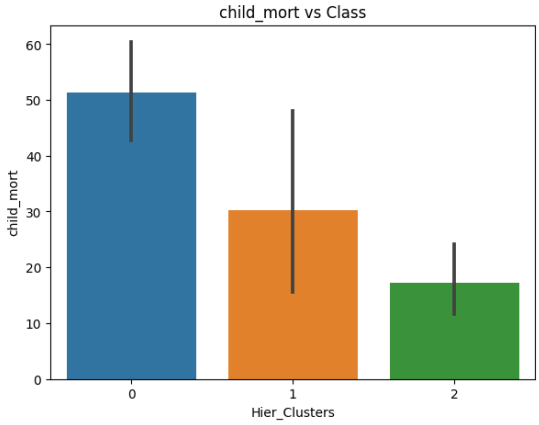


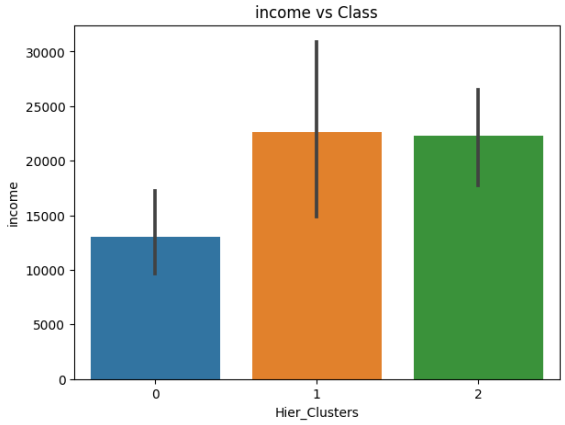
We can see that it has 3 branches hence the number of clusters to be formed are 3.

Scatterplot with assigned clusters



- ➔ Now we have got the clusters but we don't know which value corresponds to what!
- ➔ Hence, we draw a barplot of income & child_mort w.r.t labeled clusters.
- ➔ As we know, low income and high child mortality is a sign of an economically backward nation.





From above plot we can see that

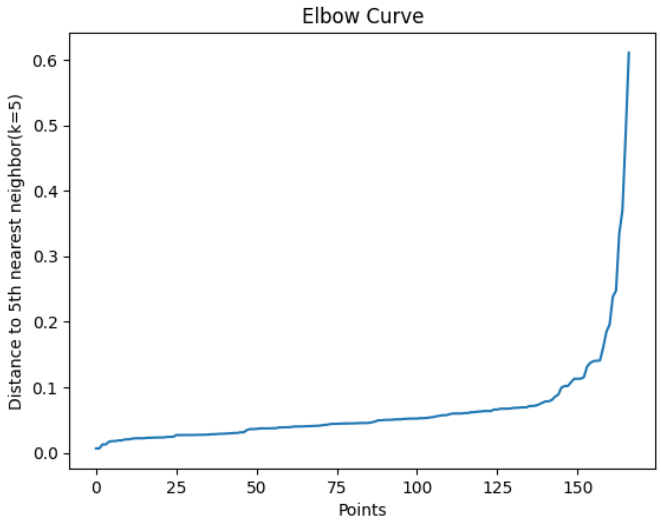
- ➔ **0 :** Help Needed **(Yes)**
- ➔ **1 :** Might Need Help **(Might)**
- ➔ **2 :** No Help Needed **(No)**

DBSCAN :

We performed **DBSCAN simultaneously on two datasets,** first obtained by feature combination and another by PCA.

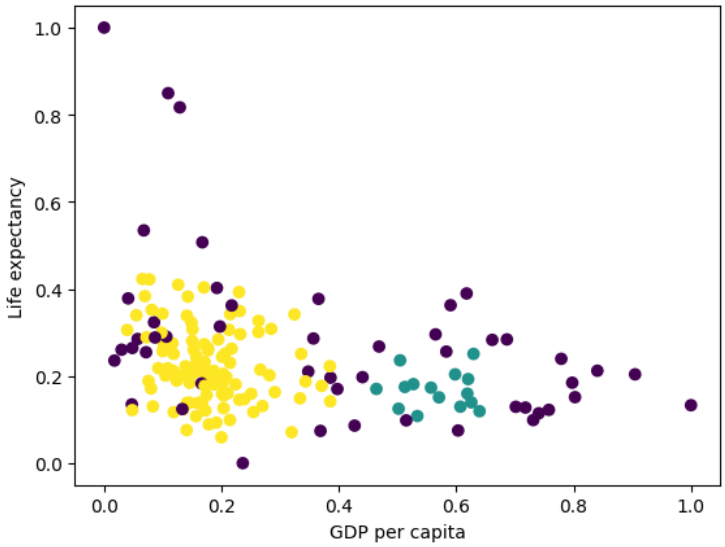
Normal Data (Feature Combination) :

We used the elbow curve to obtain an optimal value for epsilon and minimum samples.

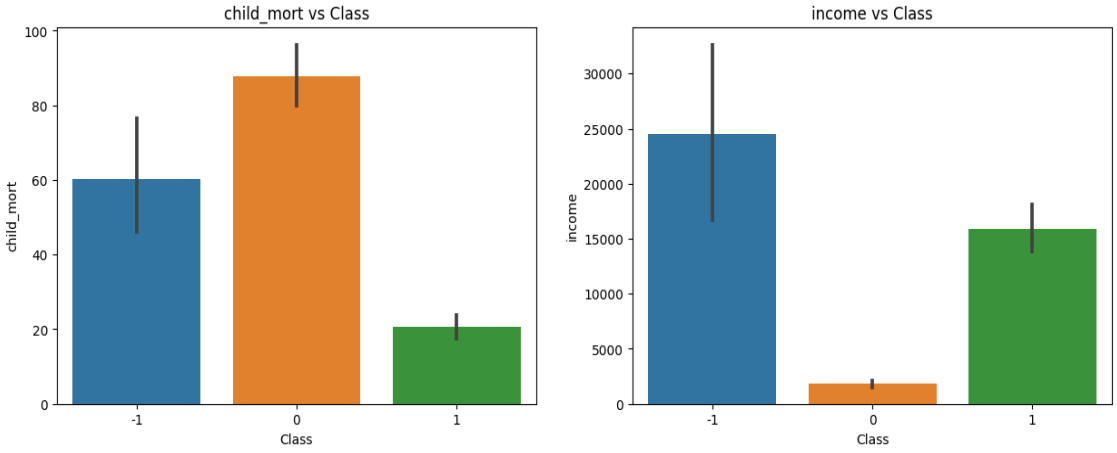


It is clear that the best value for **epsilon is 0.08**, and the required minimal number of observations is 5.

Scatterplot with assigned clusters



- ➔ Now we have got the clusters but we don't know which value corresponds to what!
- ➔ Hence, we draw a barplot of income & child_mort w.r.t labeled clusters.
- ➔ As we know, low income and high child mortality is a sign of an economically backward nation.

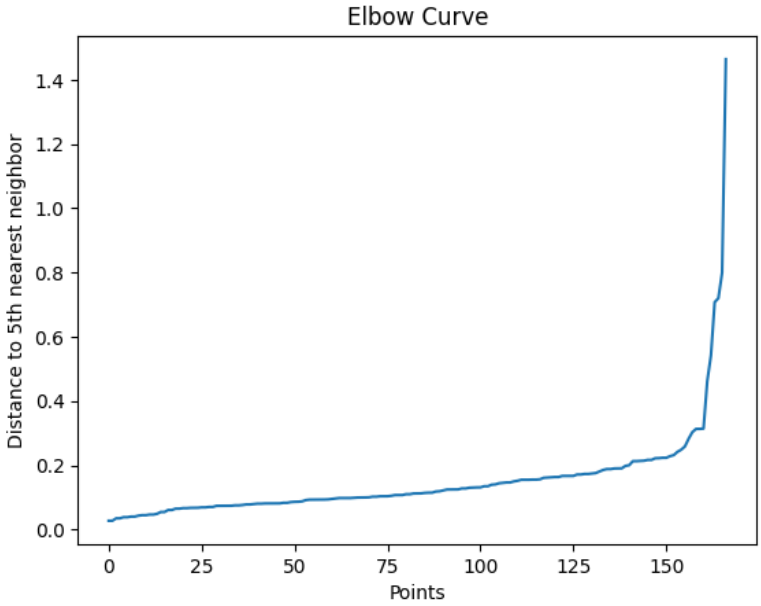


From above plot we can see that

- ➔ **-1 :** Noise **(Noise)**
- ➔ **0 :** Help Needed **(Yes)**
- ➔ **1 :** No Help Needed **(No)**

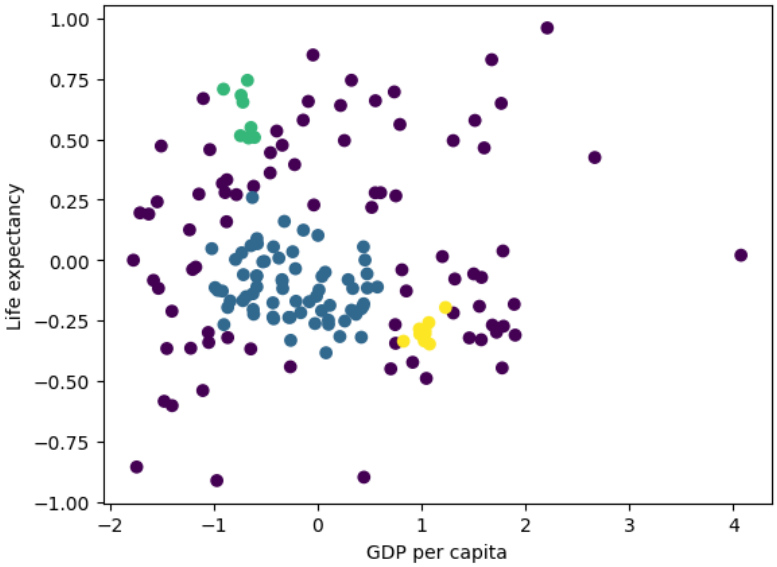
PCA Data :

We used the elbow curve to obtain an optimal value for epsilon and minimum samples.

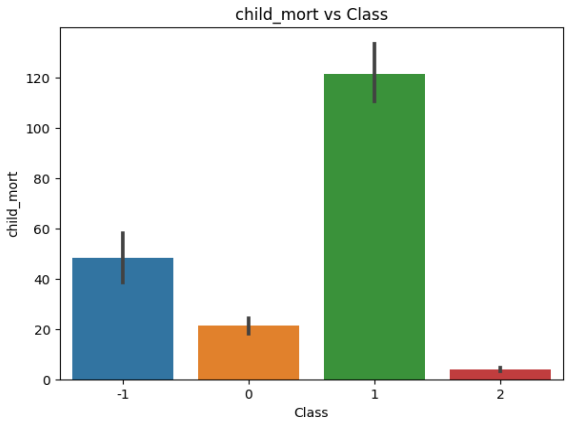


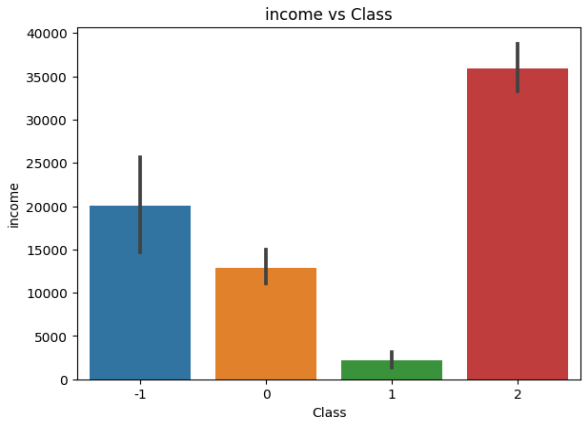
It is clear that the best value for **epsilon is 0.2**, and the required minimal number of observations is 6.

Scatterplot with assigned clusters



- ➔ Now we have got the clusters but we don't know which value corresponds to what!
- ➔ Hence, we draw a barplot of income & child_mort w.r.t labeled clusters.
- ➔ As we know, low income and high child mortality is a sign of an economically backward nation.





From above plot we can see that

- ➔ **-1 :** Noise **(Noise)**
- ➔ **0 :** Might Need Help **(Might)**
- ➔ **1 :** Help Needed **(Yes)**
- ➔ **2 :** No Help Needed **(No)**

Silhouette Score :

K-Means Clustering

- ➔ **Normal Data : 0.45197076714375445**
- ➔ **PCA Data : 0.39202292064405597**

Hierarchical Clustering

- ➔ **Normal data : 0.37795779229139226**
- ➔ **PCA data : 0.4046355541546756**

DBSCAN

- ➔ **Normal data : 0.24109785763130062**
- ➔ **PCA data : -0.07**