PROJECT

DATA MINING

Table of contents

List of Figures

List of Tables

PROBLEM 1 - CLUSTERING

Problem Statement:

The dataset given is about the health and economic conditions in different States of a country. The Group States based on how similar their situation is, so as to provide these groups to the government so that appropriate measures can be taken to escalate their Health and Economic conditions.

Q1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, etc.)

Sample of the Dataset:

	States	Health_indeces1	Health_indices2	Per_capita_income	GDP
0	Bachevo	417	66	564	1823
1	Balgarchevo	1485	646	2710	73662
2	Belasitsa	654	299	1104	27318
3	Belo_Pole	192	25	573	250
4	Beslen	43	8	528	22

Table 1. Sample of the States Dataset.

Size of the Dataset:

- There are 5 features (columns) with 297 observations (rows) in the data frame.

Data Types of Variables in the Dataset:

Feature	Data_Type
States	object
Health_indeces1	int64
Health_indices2	int64
Per_capita_income	int64
GDP	int64

Table 2. Data Types of All Features in the States Dataset.

- All features (variables) in the above dataset are numeric and continuous except states feature.

Basic Information of Dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 297 entries, 0 to 296
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   States             297 non-null    object
 1   Health_indeces1    297 non-null    int64 
 2   Health_indices2    297 non-null    int64 
 3   Per_capita_income  297 non-null    int64 
 4   GDP                297 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 11.7+ KB
```

Data Description:

	Health_indeces1	Health_indices2	Per_capita_income	GDP
count	297.00	297.00	297.00	297.00
mean	2630.15	693.63	2156.92	174601.12
std	2038.51	468.94	1491.85	167167.99
min	-10.00	0.00	500.00	22.00
25%	641.00	175.00	751.00	8721.00
50%	2451.00	810.00	1865.00	137173.00
75%	4094.00	1073.00	3137.00	313092.00
max	10219.00	1508.00	7049.00	728575.00

Table 3. Summary of States Dataset.

Exploratory Data Analysis

Let us Check for duplicate observations

- There are no duplicate observations in the given dataset.

Let us Check for null values

Feature	Count_of_Null_Values
States	0
Health_indeces1	0
Health_indices2	0
Per_capita_income	0
GDP	0

Table 4. Null Values in the States Dataset.

- There are no null values in the given dataset.

Let us check the outliers in the Dataset

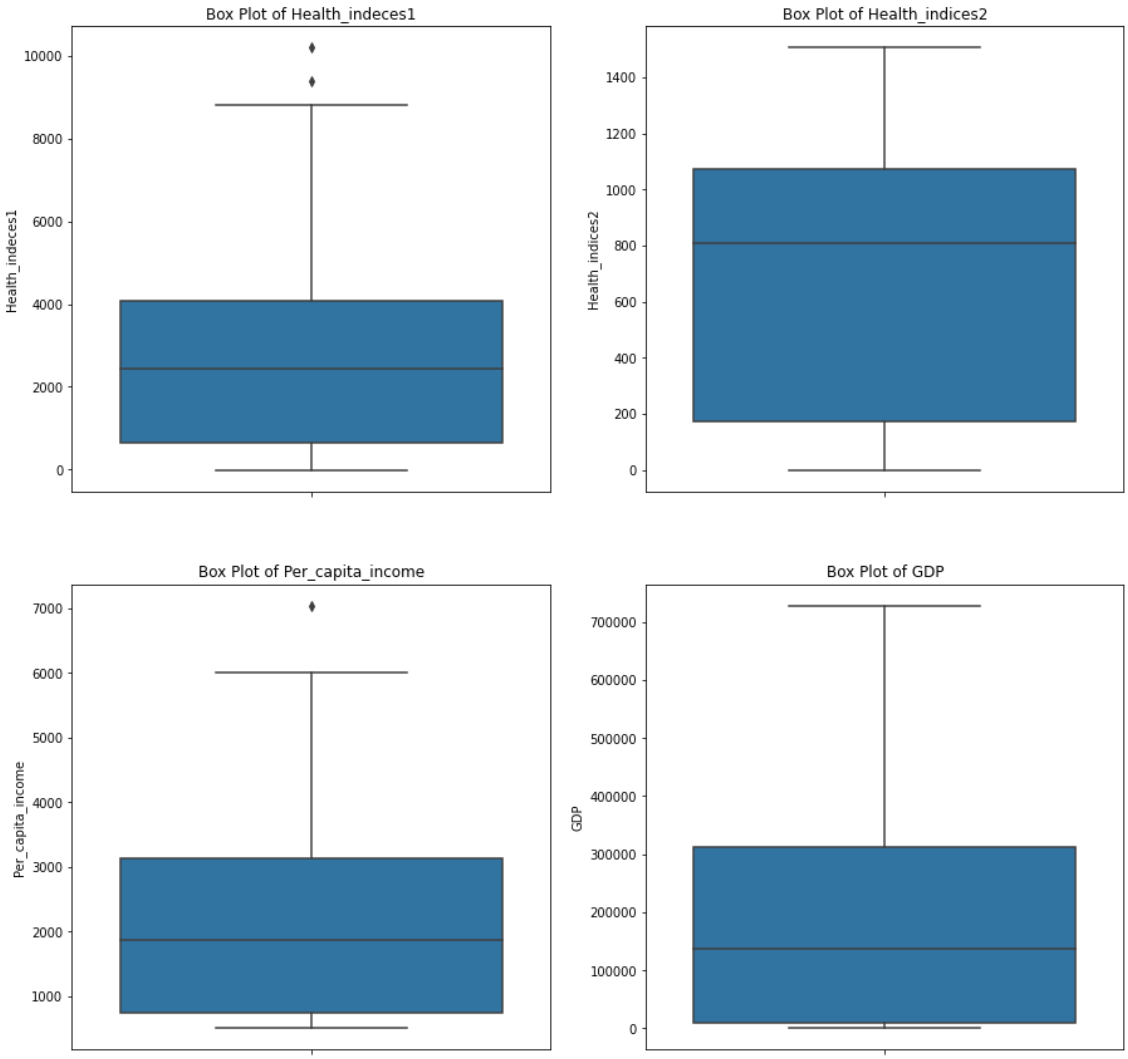


Figure 1. Box Plots of Numerical Features in States Dataset.

Feature	No. of Outliers	Percentage of Outliers
Health_indeces1	2	0.7
Per_capita_income	1	0.3
GDP	0	0.0
Health_indices2	0	0.0

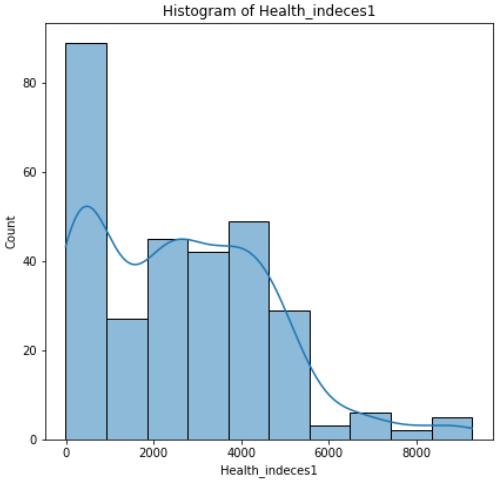
Table 5. Number of Outliers and Percentage of Outliers in States Dataset.

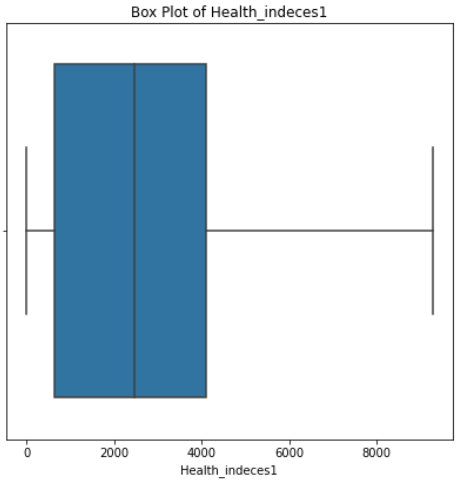
Insights:

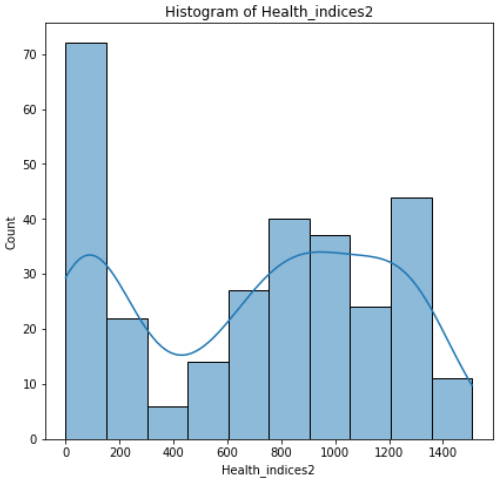
- There are outliers in Health Indices1 and Per capita income features.
- The outliers in Health Indices1 and Per capita income features are 0.7% and 0.3% respectively.

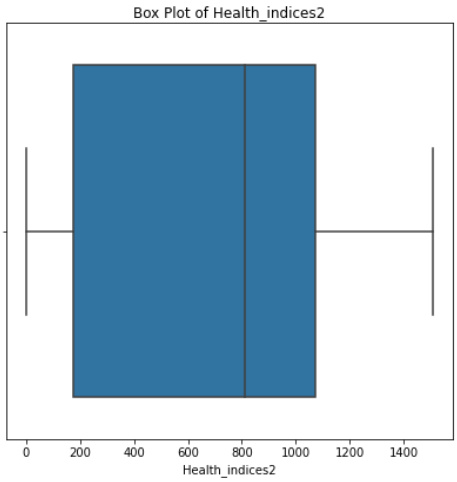
- Outliers are treated by capping and flooring method.

Univariate Analysis – Distribution Plots









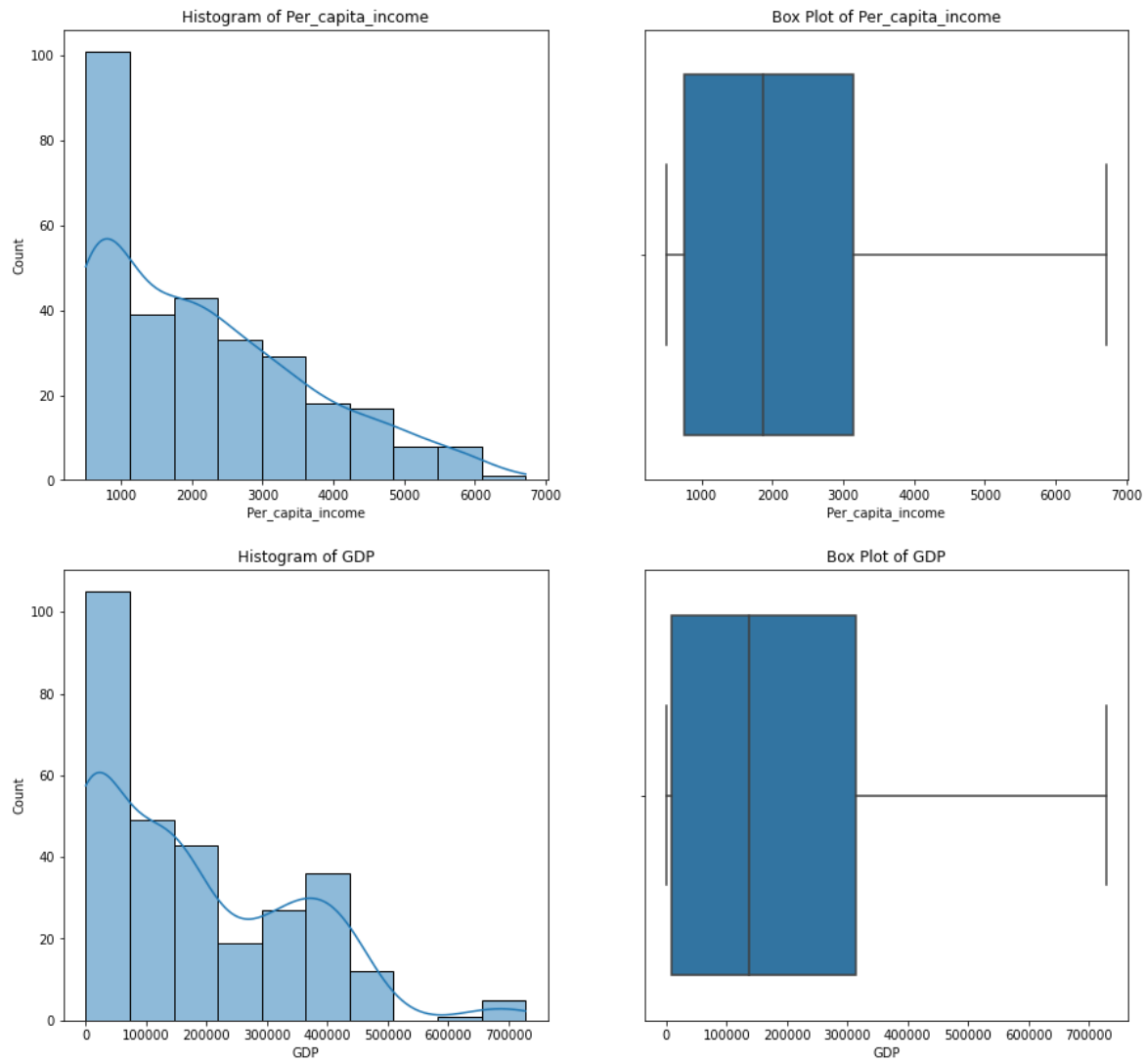


Figure 2. Histograms and Box Plots for Numerical Features in States Dataset.

Skewness:

It is a measure of lack of symmetry in a distribution.

Feature	skewness
Health_indeces1	0.67
Health_indices2	-0.17
Per_capita_income	0.81
GDP	0.83

Table 6. Skewness of Numeric Features in States Dataset.

Kurtosis: It is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution

Feature	kurtosis
Health_indeces1	0.22
Health_indices2	-1.40
Per_capita_income	-0.19
GDP	0.06

Table 7. Kurtosis of Numeric Features in States Dataset.

Insights:

From above plots and tables, we can conclude below points,

1. Except Health Indeces2 feature, all other features are right skewed distributions (Positively skewed).
2. Health Indeces1 and GDP features have positive kurtosis.
3. Health Indeces2 and Per capita income features have negative kurtosis.

Bivariate Analysis – Between Numeric Continuous Variables

Pair Plot:

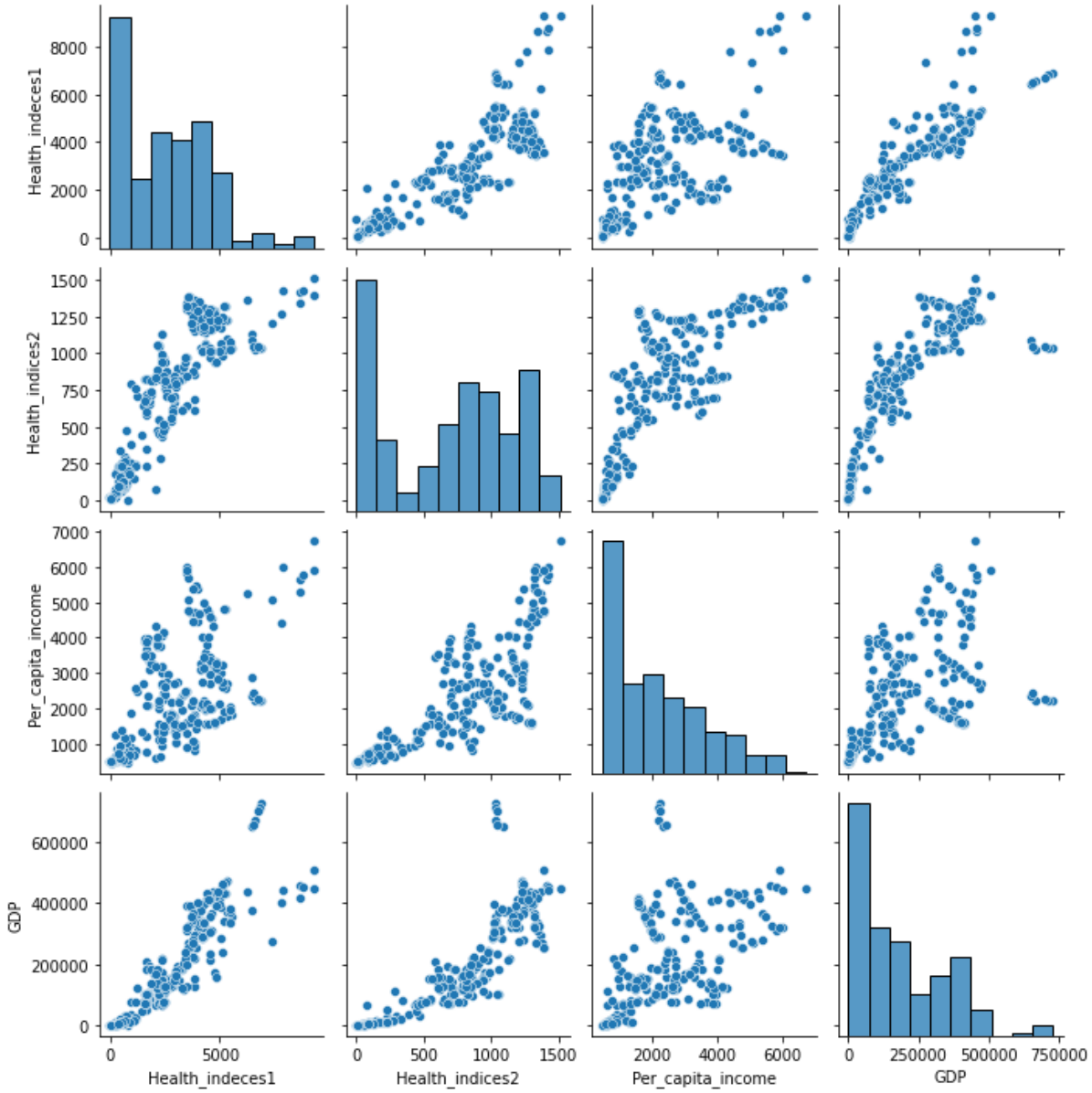


Figure 3. Pair Plot for Numeric Continuous Features in States Dataset.

Heat Map:

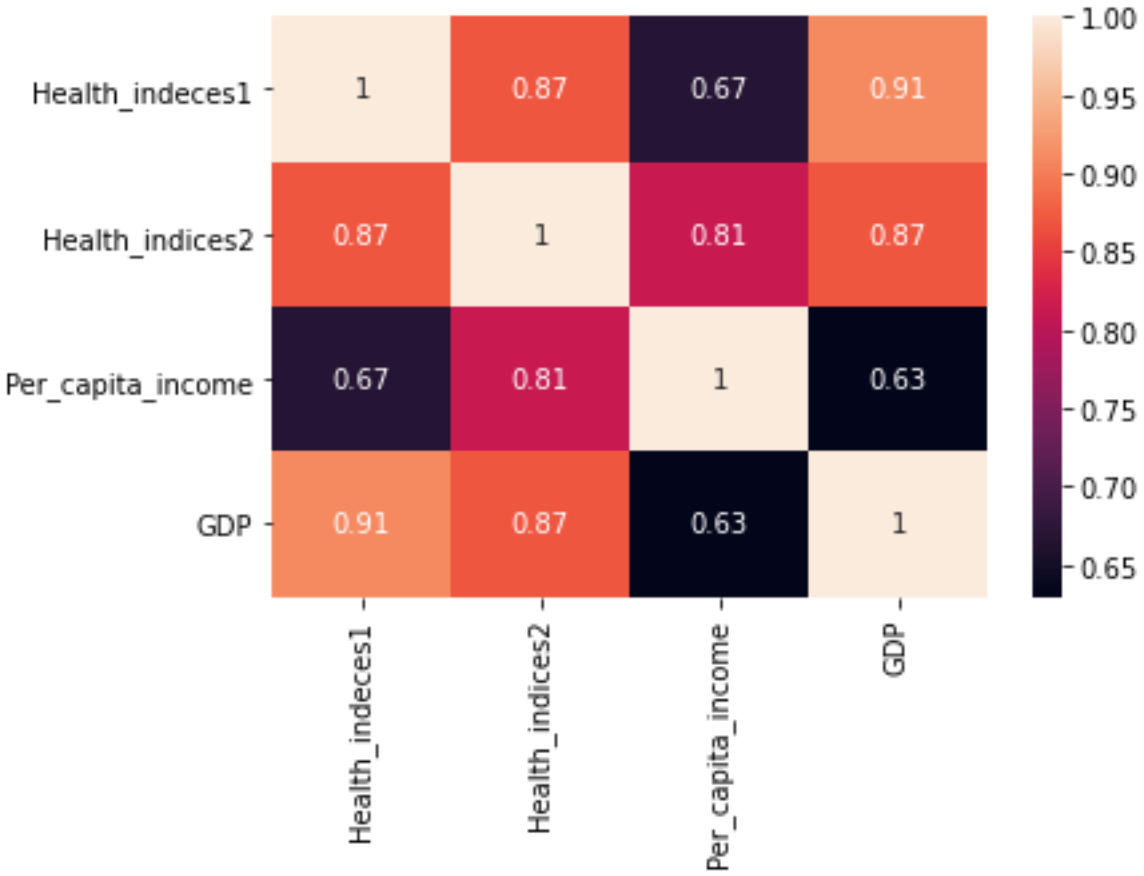


Figure 4. Heat Map for Numeric Continuous Features in States Dataset.

Insights:

From above Pair-Plot and Heatmap, we can conclude below points,

1. Few features have strong correlation between them like Health indeces1 & GDP (0.91), Health indeces2 & GDP (0.87).
2. Few features have moderate correlation between them like Health indeces1 & Per capita income (0.67), GDP & Per capita income (0.63).

Q1.2. Do you think scaling is necessary for clustering in this case? Justify

- Generally, Scaling improves the performance of all distance-based models because if we don't scale the data, it gives higher weightage to features which have higher magnitude. Hence, it is always advisable to bring all the features to the same scale before proceeding to distance-based algorithms like Agglomerative clustering and K-Means Clustering.
- In this dataset, the magnitudes of the statistical parameters like Mean, Standard Deviation, Variance, Minimum and Maximum are significantly different for all features (Refer below

table). **Hence, scaling is required to bring all the features into a common scale before proceeding to clustering.**

- We can use z-score method to scale the data i.e., finding z-score value for each and every observation in the dataset by using following formula.

$$Z\ Score = \frac{(x - \mu)}{Sigma}$$

Where, x = Value of the observation

μ = Mean

Sigma = Standard Deviation

Feature	mean	std	min	max	variance
Health_indeces1	2626.5	2025.9	-10.0	9273.5	4.104160e+06
Health_indices2	693.6	468.9	0.0	1508.0	2.199088e+05
Per_capita_income	2155.8	1488.3	500.0	6716.0	2.214995e+06
GDP	174601.1	167168.0	22.0	728575.0	2.794514e+10

Table 8. Mean, Standard Deviation and Variance of All Numeric Features.

Q1.3. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.

Z-Score Method has been applied to scale the data and the sample of the scaled dataset is shown below.

Sample of the Scaled Dataset:

	Health_indeces1	Health_indices2	Per_capita_income	GDP
0	-1.092498	-1.340654	-1.071354	-1.035304
1	-0.564428	-0.101746	0.373007	-0.604838
2	-0.975314	-0.842955	-0.707908	-0.882536
3	-1.203748	-1.428232	-1.065297	-1.044730
4	-1.277421	-1.464545	-1.095584	-1.046096

Table 9. Sample of the Scaled States Dataset.

Hierarchical Clustering:

- This method is based on hierarchy representation of clusters where parent cluster is connected to further to child clusters.
- A cluster represents collection of similar data points.
- The agglomerative clustering is the most popular and common hierarchical clustering.
- This method starts by considering each data point as a single cluster. In the next step the single clusters are merged into a big cluster based on the similarity between them.
- The procedure is repeated until all the datapoints are merged into one big cluster. The procedure can be represented as hierarchy/tree of clusters.

Dendrogram:

- A dendrogram is a pictorial representation to visualize hierarchical clustering.
- It is mainly used to show the outcome of hierarchical clustering in the form of a tree like diagram that records the sequences of merges and splits.

Linkage:

- Linkage process merges two clusters into one cluster based on the distance or similarity between them.
- The similarity between two clusters is very important parameter for merging and dividing of cluster. Ward's method or minimum variance method is used to calculate similarity between two clusters.

Ward´s linkage:

- The concept is much similar to analysis of variance (ANOVA). The linkage function specifying the distance between two clusters is computed as the increase in the "error sum of squares" (ESS) after combining two clusters into a single cluster.
- Ward´s method selects the successive clustering steps so as to minimize the increase in ESS at each step.

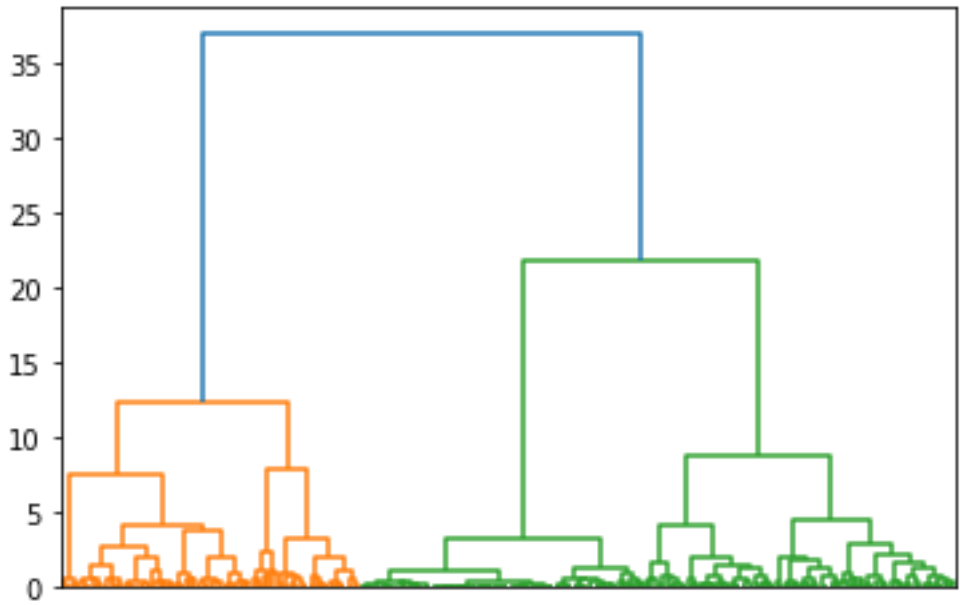


Figure 5. Dendrogram without Truncated.

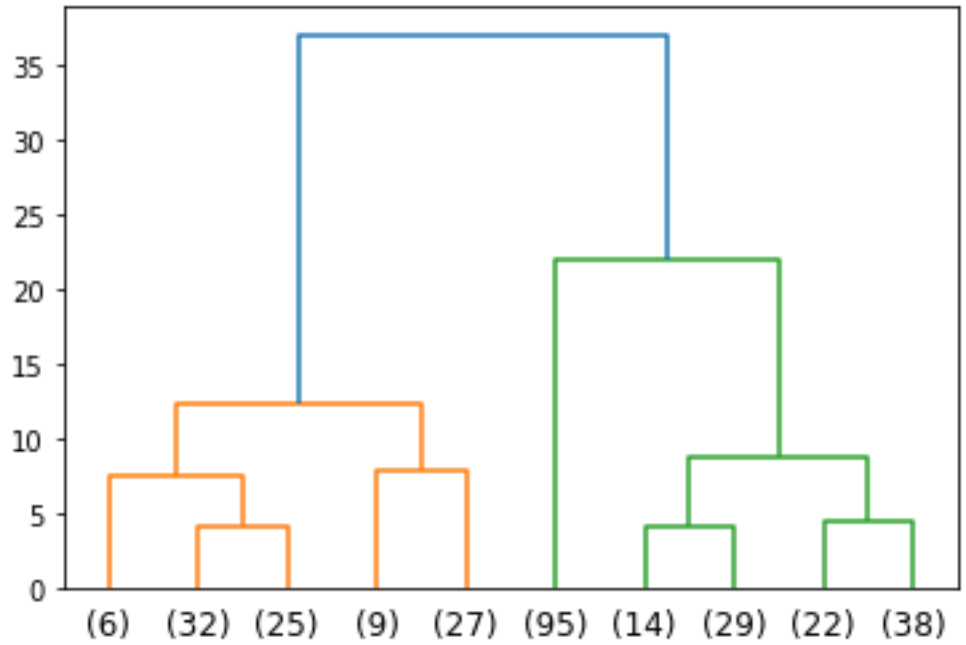


Figure 6. Truncated Dendrogram Representing Last 10 Clusters Formed.

Selecting the Optimum Number of Clusters:

- From above Truncated Dendrogram, it can be noticed that the distance or increase in within sum squares (WSS) is large (length of blue line) to merge last two clusters into single final cluster.

- Hence, we can select the optimum number of clusters are two. But according the business, making two clusters will not add any additional benefit over without clustering. So, it is not correct.
- **The next optimum number of clusters selected based on distance (length of green line) or increase in within sum squares (WSS) are three.**

Hierarchical Clustering Labels:

The following is an array of cluster numbers (labels) for all the observations in the dataset.

```
array([2, 3, 2, 2, 2, 2, 2, 1, 2, 3, 2, 2, 2, 2, 2, 2, 1, 2, 2, 1, 3, 2,
       2, 2, 3, 2, 3, 3, 3, 2, 3, 2, 2, 1, 2, 2, 1, 3, 2, 3, 2, 2, 3, 3,
       3, 2, 2, 1, 2, 2, 2, 3, 2, 1, 2, 3, 3, 2, 2, 3, 3, 2, 2, 1, 3, 2,
       3, 2, 2, 2, 2, 3, 2, 2, 2, 2, 3, 3, 2, 2, 2, 2, 2, 2, 3, 3, 2, 2,
       2, 1, 2, 2, 2, 3, 2, 2, 3, 2, 2, 3, 2, 2, 1, 2, 3, 2, 3, 3, 2, 2,
       2, 1, 3, 3, 2, 2, 2, 3, 2, 2, 3, 2, 2, 2, 1, 2, 3, 2, 2, 3, 2, 2,
       2, 3, 2, 2, 2, 2, 2, 3, 2, 3, 2, 2, 2, 3, 1, 2, 3, 2, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3], dtype=int32)
```

Sample of Clustered Dataset:

	States	Health_indeces1	Health_indices2	Per_capita_income	GDP	Hclusters
0	Bachevo	417.0	66.0	564.0	1823.0	2
1	Balgarchevo	1485.0	646.0	2710.0	73662.0	3
2	Belasitsa	654.0	299.0	1104.0	27318.0	2
3	Belo_Pole	192.0	25.0	573.0	250.0	2
4	Beslen	43.0	8.0	528.0	22.0	2

Table 10. Sample of the Hierarchical Clustered Dataset.

Customer Segmentation:

Hclusters	Health_indeces1	Health_indices2	Per_capita_income	GDP
1	4912.7	1201.6	3371.8	377132.5
2	401.1	104.5	680.7	5388.8
3	2481.8	748.7	2347.6	136004.7

Table 11. Centroids of the Hierarchical Clusters.

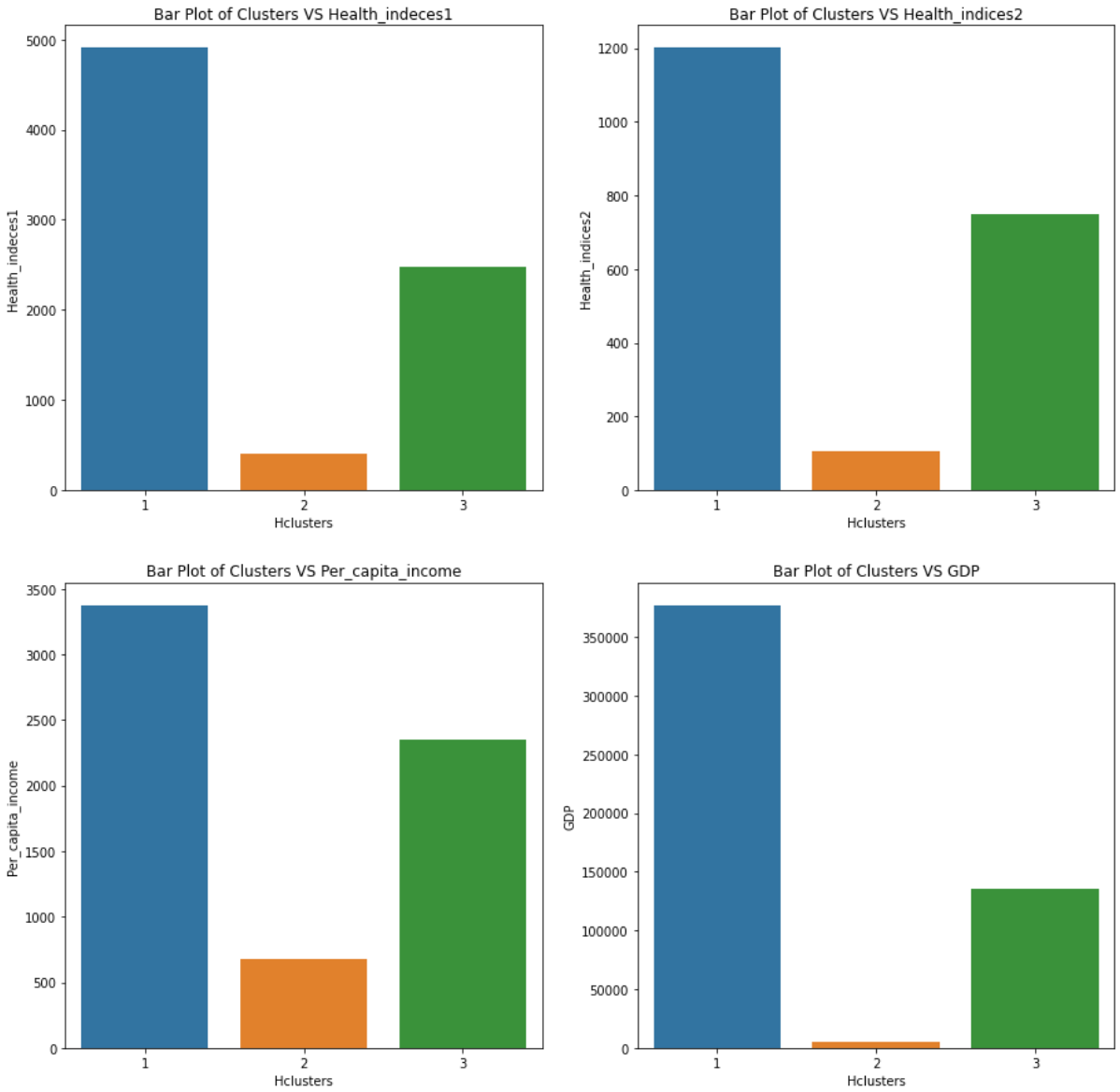


Figure 7. Means of Different Features vs Hierarchical Clusters.

The coordinates of each cluster's centroid are shown in table 11 so that means of each feature in different clusters can be compared. From above table and bar plots, we can write below conclusions.

- Means of all features decreases in the order of cluster1, cluster3, cluster2.
- The states in cluster 1 have high health indices, high Per capita income and high GDP.
- The states in cluster 2 have low health indices, low Per capita income and low GDP.
- The states in cluster 3 have moderate health indices, moderate Per capita income and moderate GDP.

Visualization of Hierarchical Clusters:

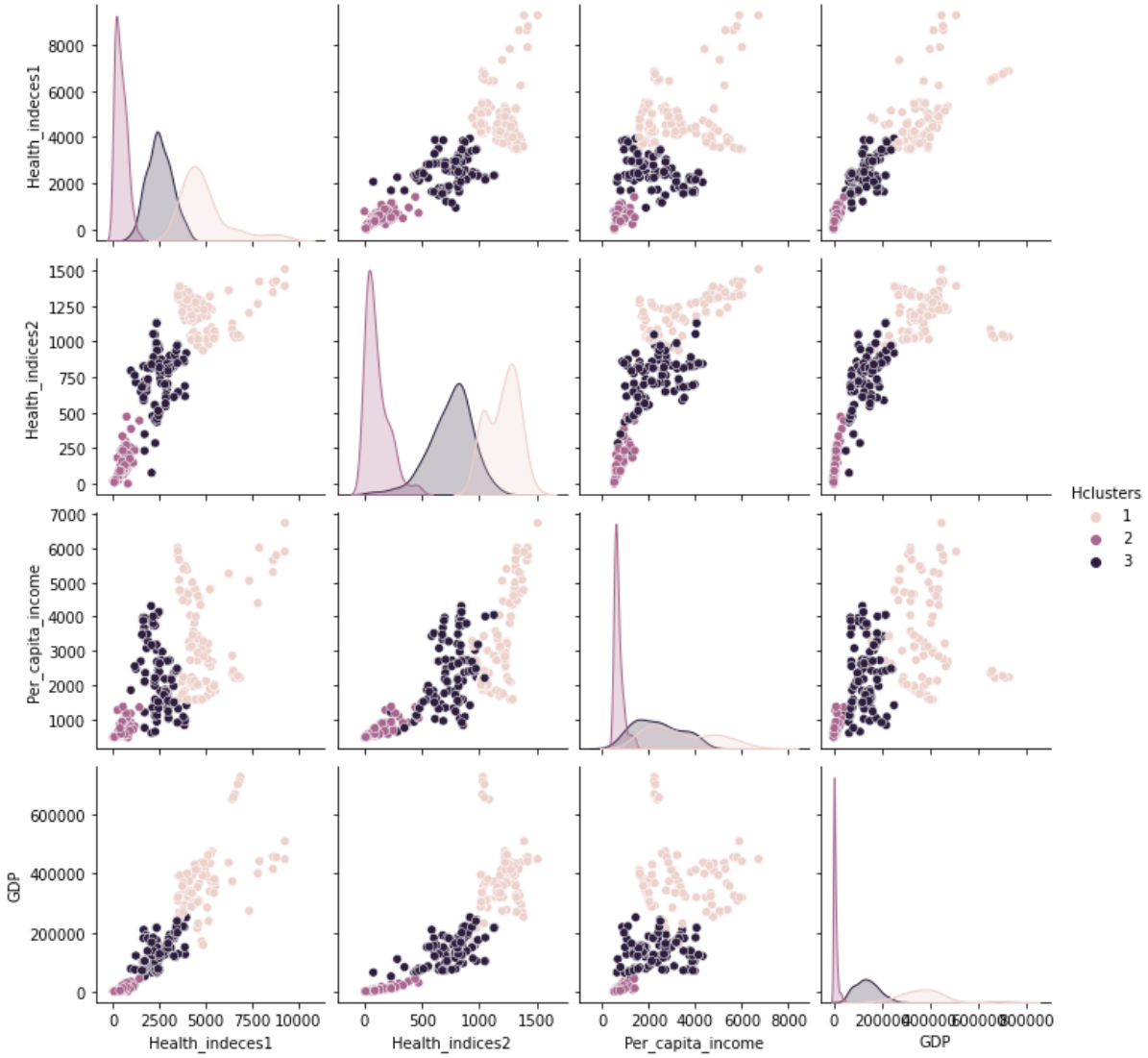


Figure 8. Pair Plot of Numeric Features with Hierarchical Clusters.

The above pair plot indicates that all customers are properly segregated into three clusters based on their similarities.

Q1.4. Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and find the silhouette score.

K-means Clustering:

- K-means clustering is an unsupervised learning algorithm whose goal is to find similar groups or assign the data points to clusters on the basis of their similarity.

- It means the points in same cluster are similar to each other and the points in different clusters are dissimilar with each other.
- In this method, initially we need to choose the number of clusters before applying the model and run the model for different number of clusters then optimum number of clusters can be found by plotting elbow curve for within sum squares (WSS).

Optimum No. of Clusters by Elbow Plot Method:

- This method is based on plotting the values of within sum squares (WSS) against different no. of clusters (k). As the no. of clusters increases, WSS decreases.
- The decrease in WSS is significant upto certain no. of clusters, after which there is no significant decrease in WSS. This no. of clusters at which decrease in WSS is not significant is known as optimum no. of clusters.
- The optimum no. of clusters can be identified from the WSS plot at the elbow point.

	Number_of_Clusters	WSS
0	1	1188.00
1	2	469.38
2	3	258.45
3	4	181.74
4	5	147.73
5	6	116.60
6	7	90.01
7	8	78.99
8	9	70.09
9	10	63.15

Table 12. WSS for Different No. of Clusters.

Elbow Plot:

- In this problem, Within Sum Squares (WSS) are calculated for different no. of clusters and tabulated above and Elbow plot is drawn by taking no. of clusters (k) on x-axis and WSS values on y-axis.
- From Elbow plot, we can notice that elbow exist **at cluster number three.** Hence, we can decide that the **optimum no. of clusters is three** by Elbow plot method.

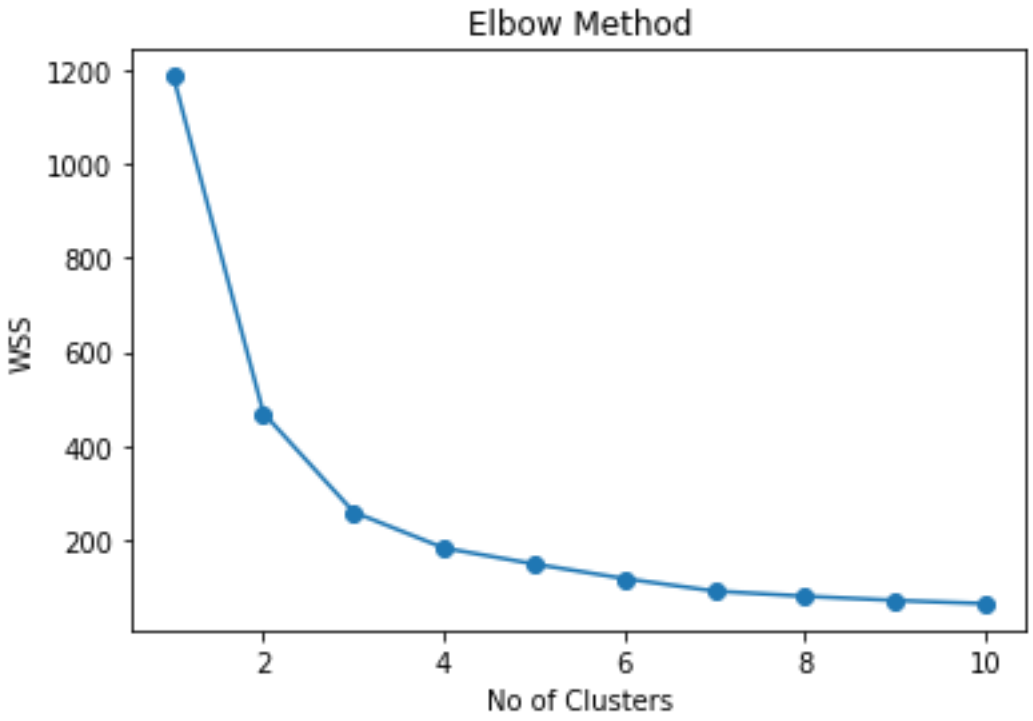


Figure 9. No. of Clusters vs WSS Plot (Elbow Plot).

Optimum No. of Clusters by Silhouette Score Method:

	Number_of_Clusters	Silhouette_Score
0	2	0.53
1	3	0.53
2	4	0.55
3	5	0.52
4	6	0.53
5	7	0.56
6	8	0.53
7	9	0.51
8	10	0.49

Table 13. Silhouette Scores for Different No. of Clusters.

In this problem, Silhouette Scores are calculated for different no. of clusters and tabulated above and Silhouette Score Plot is drawn by taking no. of clusters (k) on x-axis and Silhouette Score values on y-axis.

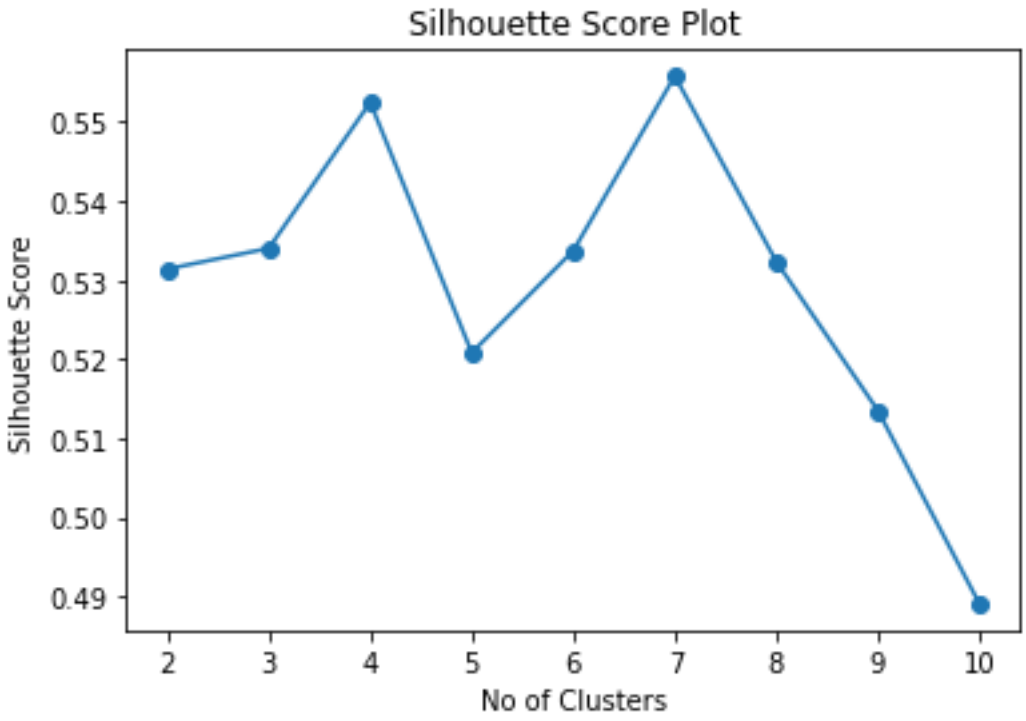


Figure 10. No. of Clusters vs Silhouette Scores Plot.

- From above plot, we can notice that maximum Silhouette Score exist at four clusters (0.55) and seven clusters (0.56). But we have got optimum number of clusters according to WSS plot as three. Hence, it is better to select optimum number of clusters is three because for three clusters we have got reasonably good Silhouette score (0.53).

K-Means Clustering Labels:

The following is an array of cluster numbers (labels) for all the observations in the dataset.

```
array([0, 1, 0, 0, 0, 0, 0, 2, 0, 1, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 1, 0,
       0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 2, 0, 0, 2, 1, 0, 1, 0, 0, 1, 1,
       1, 0, 0, 1, 0, 0, 0, 1, 0, 2, 0, 1, 1, 0, 0, 1, 1, 0, 0, 2, 1, 0,
       1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 2, 0, 1, 0, 1, 1, 0, 0,
       0, 2, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 2, 0, 1, 0, 0, 1, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 2, 0, 1, 0, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1])
```

Sill Width of Samples:

- Maximum Sill width: 0.882
- Minimum Sill width: -0.081
- Average Sil width: 0.534
- Number of data points having negative sill width are four only.

From above data, we can notice that sill width for all the observations is ranging from -0.081 to 0.882. Only four data points have been wrongly mapped to clusters. Hence, it can be concluded that the dataset has been clustered into three groups properly.

Sample of Clustered Dataset:

	States	Health_indeces1	Health_indices2	Per_capita_income	GDP	kmclusters
0	Bachevo	417.0	66.0	564.0	1823.0	0
1	Balgarchevo	1485.0	646.0	2710.0	73662.0	1
2	Belasitsa	654.0	299.0	1104.0	27318.0	0
3	Belo_Pole	192.0	25.0	573.0	250.0	0
4	Beslen	43.0	8.0	528.0	22.0	0

Table 14. Sample of the K-Means Clustered Dataset.

Customer Segmentation:

kmclusters	Health_indeces1	Health_indices2	Per_capita_income	GDP
0	499.2	116.4	693.8	9428.1
1	2597.1	783.0	2464.1	141264.1
2	4919.6	1212.3	3382.3	385648.6

Table 15. Centroids of the K-Means Clusters.

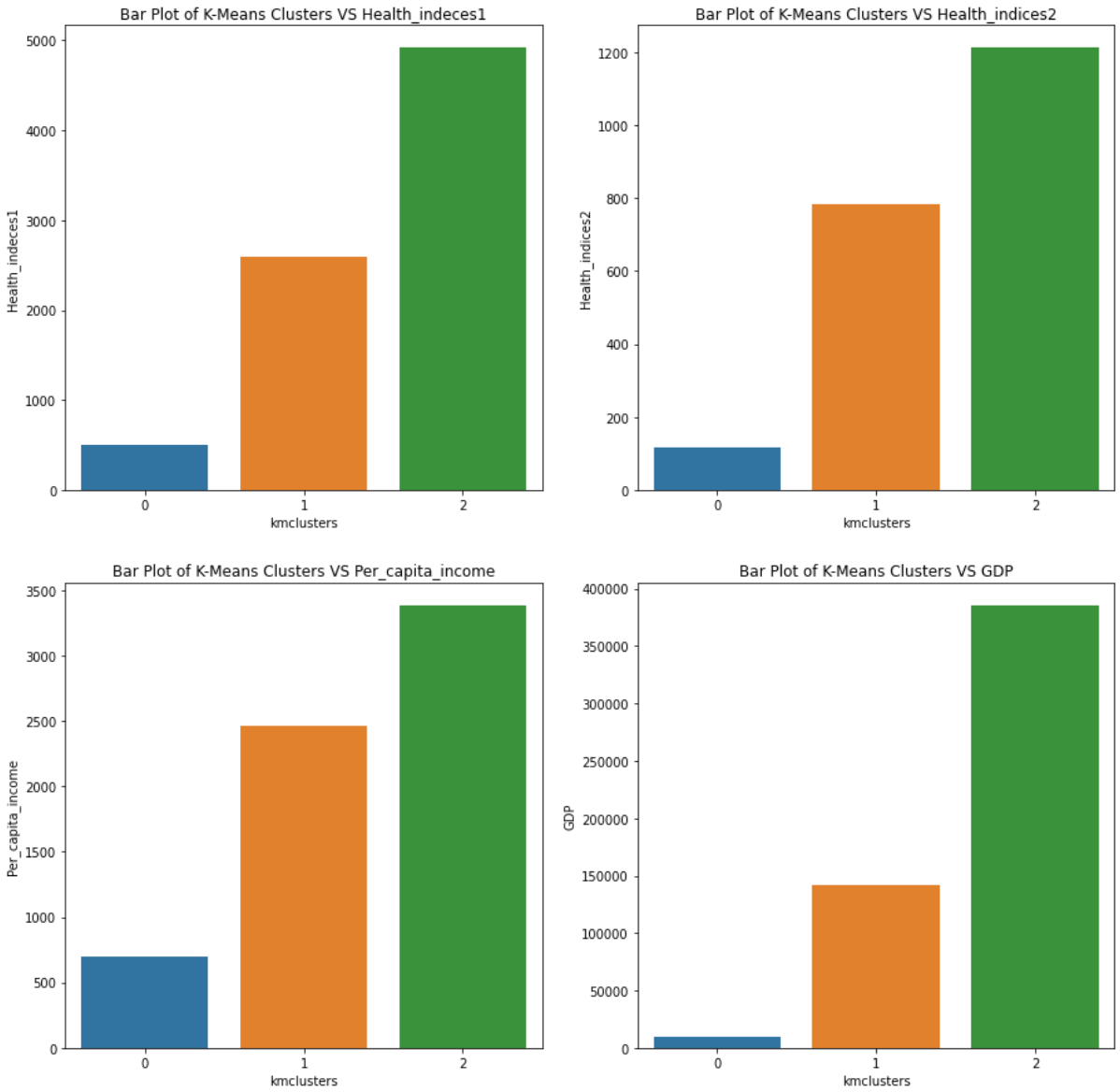


Figure 11. Means of Different Features vs K-Means Clusters.

The coordinates of each cluster's centroid are shown in table 15 so that means of each feature in different clusters can be compared. From above table and bar plots, we can write below conclusions.

- Means of all features increases in the order of cluster 0, cluster 1 and cluster 2.
- The states in cluster 2 have high health indices, high Per capita income and high GDP.
- The states in cluster 0 have low health indices, low Per capita income and low GDP.
- The states in cluster 1 have moderate health indices, moderate Per capita income and moderate GDP.

Visualization of K-Means Clusters:

The below pair plot indicates that all customers are properly segregated into three clusters based on their similarities.

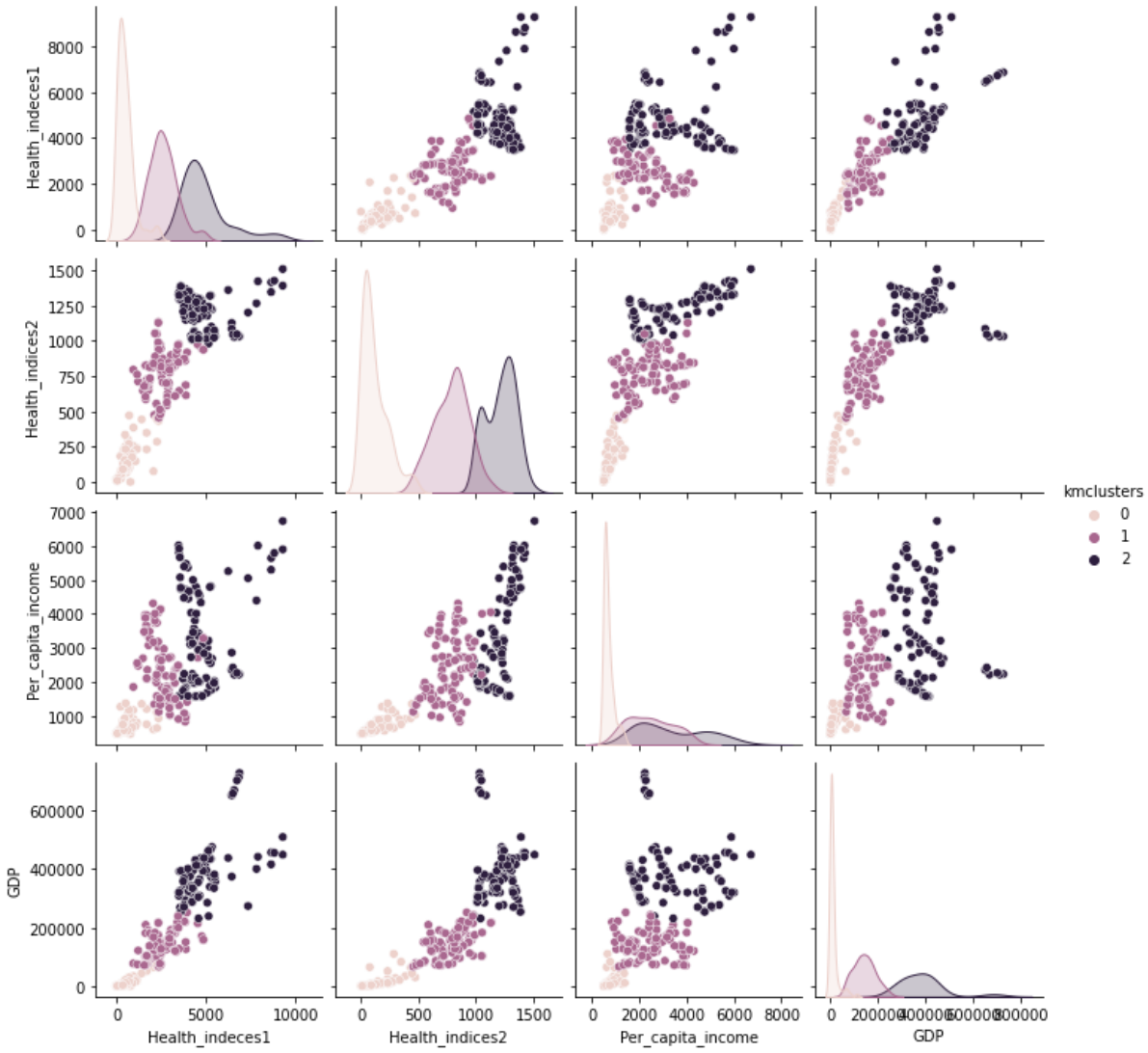


Figure 12. Pair Plot of Numeric Features with K-Means Clusters.

Distribution States among Kmeans clusters:

Kmeans Cluster	Number of States
0	101
1	101
2	95

States in Kmeans cluster 0:

```
array(['Bachevo', 'Belasitsa', 'Belo_Pole', 'Beslen', 'Bogolin',
       'Bogoroditsa', 'Budiltsi', 'Churicheni', 'Churilovo', 'Debochitsa',
       'Dobarsko', 'Dobri_Laki', 'Dolen', 'Drakata', 'Drangovo', 'Garmen',
       'Gega', 'Godeshevo', 'Gyurgevo', 'Ilinden', 'Kamena', 'Klyuch',
       'Kochan', 'Kolarovo', 'Krandzhilitsa', 'Krastiltsi', 'Kulata',
       'Leshko', 'Logodazh', 'Moshtanets', 'Nikudin', 'Osina', 'Padesh',
       'Palat', 'Pletena', 'Polenitsa', 'Rupite', 'Satovcha', 'Struma',
       'Strumeshnitsa', 'Sushitsa', 'Vaklinovo', 'Valkosel', 'Vishlene',
       'Valkovo', 'Yavornitsa', 'Zanoga', 'Blagoevgrad', 'Zelenodol',
       'ZoycheneBallela', 'Ballerin', 'Ballinamallard', 'Ballintoy',
       'Balloo', 'Ballybogy', 'Ballycarry', 'Ballycastle', 'Ballyclare',
       'Ballyeaston', 'Ballygawley', 'Ballygowan', 'Ballyhalbert',
       'Ballyhornan', 'Ballykelly', 'Ballykinler', 'Ballylinney',
       'Ballymacmaine', 'Ballymagorry', 'Ballymartin', 'Ballymena',
       'Ballynahinch', 'Ballyrobert', 'Ballyronan', 'Ballyrory',
       'Ballyvoy', 'Ballywalter', 'Balnamore', 'Banbridge', 'Bangor',
       'Belcoo', 'Belfast', 'Bellaghy', 'Bellarena', 'Belleeks',
       'Benburb', 'Beragh', 'Bessbrook', 'Blackskull', 'Blackwatertown',
       'Blaney', 'Bleary', 'Boho', 'Brackaville', 'Bready',
       'Brookeborough', 'Broughshane', 'Bryansford', 'Buckna',
       'BushmillsCaledon', 'Campsie', 'Drumbeg'], dtype=object)
```

States in Kmeans cluster 1:

```
array(['Balgarchevo', 'Cherniche', 'Gabrovo', 'Gorna_Breznitsa',
       'Ivanovo', 'Kalimantsi', 'Krupnik', 'Lebnitsa', 'Mendovo',
       'Mihnevo', 'Mikrevo', 'Obidim', 'Petrelik', 'Pravo_Bardo',
       'Ribnik', 'Slashten', 'Starchevo', 'Suhostrel', 'Tuhovishta',
       'Volno', 'Zheleznitsa', 'Zhizhevo', 'Ballycassidy', 'Ballylesson',
       'Ballymacnab', 'Ballymoney', 'Ballynure', 'Ballyrashane',
       'Ballyskeagh', 'Ballystrudder', 'Banagher', 'Bannfoot', 'Belleek',
       'Bendooragh', 'Brockagh', 'Broomhill', 'Burnfoot', 'Camlough',
       'Kilbride', 'Cullyhanna', 'Desertmartin', 'Downhill',
       'Downpatrick', 'Draperstown', 'Drinns_Bay', 'Dromara', 'Dromintee',
       'Dromore', 'Drumaness', 'Drumbo', 'Drumlaghy', 'Drumlough',
       'Drummullan', 'Drumnacanvy', 'Drumnakilly', 'Drumquin',
       'Drumraighland', 'Drumsurn', 'Dunadry', 'Dundonald', 'Dundrod',
       'Dundrum', 'Dungannon', 'Dungiven', 'Dunloy', 'Dunnamanagh',
       'Dunmurry', 'Dunnamore', 'Dunnaval', 'DunseverickGalbally',
       'Gamblestown', 'Garrison', 'Garvagh', 'Garvaghey', 'Garvetagh',
       'Gawley', 'GibsonHill', 'Gilford', 'Gillygooly', 'Glack', 'Glebe',
       'Glenarm', 'Glenavy', 'Glengormley', 'Glenmornan', 'Glenoe',
       'Glenone', 'Glynn', 'Gortaclare', 'Gortin', 'Gortnahey',
       'Goshedan', 'Gracehill', 'Grange_Corner', 'Granville',
       'Greencastle', 'Greenisland', 'Greyabbey', 'Greysteel', 'Groggan'],
      dtype=object)
```

States in Kmeans cluster 2:

```
array(['Buchino', 'Dolene', 'Fargovo', 'Kolibite', 'Kribul', 'Polena',
       'Strumyani', 'Ballygalley', 'Ballymaguigan', 'Ballyscullion',
       'Bellanaleck', 'Burren', 'Capecastle', 'Cappagh', 'Cargan',
       'Carnalbanagh', 'Carncastle', 'Carnlough', 'Carnteel',
       'Carrickaness', 'Carrickfergus', 'Carrickmore', 'Carrowclare',
       'Carrowdore', 'Carrybridge', 'Carryduff', 'Castlecaulfield',
       'Castledawson', 'Castlederg', 'Castlerock', 'Castlewellan',
       'Charlemont', 'Clabby', 'Clady', 'Cladymore', 'Clanabogan',
       'Claudy', 'Clogh', 'Clogher', 'Cloghy', 'Clonmore', 'Clonoe',
       'Clough', 'Cloughmills', 'Coagh', 'Coalisland', 'Cogry',
       'Coleraine', 'Collegeland', 'Comber', 'Conlig', 'Cookstown',
       'Corbet', 'Corkey', 'Corrinshego', 'Craigarogan', 'Craigavon',
       'Cranagh', 'Cranford', 'Crawfordsburn', 'Creagh', 'Creggan',
       'Crossgar', 'Crossmaglen', 'Crumlin', 'Cullaville', 'Cullybackey',
       'Culmore', 'Culnady', 'Curran', 'Cushendall', 'CushendunDarkley',
       'Derry_Derrycrin', 'Derrygonnelly', 'Derryhale', 'Derrykeighan',
       'Derrylin', 'Derrymacash', 'Derrymore', 'Derrynaflaw',
       'Derrynoose', 'Derrytrasna', 'Derryvore', 'Dervock', 'Doagh',
       'Dollingstown', 'Donagh', 'Donaghadee', 'Donaghcloney', 'Donaghey',
       'Donaghmore', 'Donegore', 'Dooish', 'Dorsey', 'DouglasBridge'],
      dtype=object)
```

Q1.5. Describe cluster profiles for the clusters defined. Recommend different priority-based actions that need to be taken for different clusters on the bases of their vulnerability situations according to their Economic and Health Conditions.

Cluster Profiles:

- A cluster represents collection of similar data points. In clustering process, data points are segregated into different groups (clusters) based on their similarities.
- Our objective is to minimize sum squares within clusters and maximize sum squares between clusters to ensure proper clustering.

Mapping of K-Means Clusters with Hierarchical Clusters:

Hclusters	Health_indeces1	Health_indices2	Per_capita_income	GDP
1	4912.7	1201.6	3371.8	377132.5
2	401.1	104.5	680.7	5388.8
3	2481.8	748.7	2347.6	136004.7

kmclusters	Health_indeces1	Health_indices2	Per_capita_income	GDP
0	499.2	116.4	693.8	9428.1
1	2597.1	783.0	2464.1	141264.1
2	4919.6	1212.3	3382.3	385648.6

Table 16. Comparison of Centroids of Hierarchical & K-Means Clusters.

By comparing means of different features in Hierarchical Clustering & K-Means Clustering, we can notice below key points.

- Cluster 1 in Hierarchical Clustering (high health indices, high Per capita income and high GDP) is equivalent to Cluster 2 in K-Means Clustering.
- Cluster 2 in Hierarchical Clustering (low health indices, low Per capita income and low GDP) is equivalent to Cluster 0 in K-Means Clustering.
- Cluster 3 in Hierarchical Clustering (moderate health indices, moderate Per capita income and moderate GDP) is equivalent to Cluster 1 in K-Means Clustering.

Even we can cross check above conclusions by observing labelled (both Hierarchical & K-Means) dataset.

	States	Health_indeces1	Health_indices2	Per_capita_income	GDP	Hclusters	kmclusters
0	Bachevo	417.0	66.0	564.0	1823.0	2	0
1	Balgarchevo	1485.0	646.0	2710.0	73662.0	3	1
2	Belasitsa	654.0	299.0	1104.0	27318.0	2	0
3	Belo_Pole	192.0	25.0	573.0	250.0	2	0
4	Beslen	43.0	8.0	528.0	22.0	2	0
5	Bogolin	69.0	14.0	527.0	73.0	2	0
6	Bogoroditsa	307.0	69.0	707.0	1724.0	2	0
7	Buchino	9273.5	1508.0	6716.0	449003.0	1	2
8	Budiltsi	744.0	115.0	809.0	7497.0	2	0
9	Cherniche	2975.0	857.0	1600.0	153299.0	3	1

Table 17. Sample of Dataset with Hierarchical & K-Means Clustering labels.

Priority-based actions:

1. States in **Kmeans cluster 2** have **high health indices, high Per capita income and high GDP**. Hence, we can notice that these sates may be considered as **developed states. No immediate action is required** by the government to improve health indices, per capita income and GDP but government should strictly **keep implementing the strategies which**

are being already executed in healthcare and financial departments (Equivalent to Cluster 1 in Hierarchical Clustering).

2. States in **Kmeans cluster 1** have **moderate health indices, moderate Per capita income and moderate GDP**. Hence, we can notice that these sates may be considered as **developing states. Few actions are required by the government but not immediately. Based on the budget availability,** government should **introduce new strategies** to improve health indices, per capita income and GDP and also government should strictly **keep implementing the strategies which are being already executed** in healthcare and financial departments (Equivalent to Cluster 3 in Hierarchical Clustering).
3. States in **Kmeans cluster 0** have **low health indices, low Per capita income and low GDP**. Hence, we can notice that these sates may be considered as **under developed states. Immediate actions are required** by the government to develop the states in health care and financial sectors. Government should **introduce new strategies** to improve health indices, per capita income and GDP and also government should **review the strategies which are being already executed** in healthcare and financial departments and those **strategies** have to **reformed or discontinued** based on in depth analysis (Equivalent to Cluster 2 in Hierarchical Clustering).
4. Government may look into implementing below strategies to increase health indices.
    a. Developing infra infrastructure.
    b. Providing health policies at free of cost or at low cost.
    c. Recruiting a greater number of health care professionals.
    d. Educating the people about importance of health and role of food hobbies to stay healthy.
5. Government may look into implementing below strategies to increase per capita income and GDP.
    a. Increasing minimum support price for the crops to increase the income of farmer families.
    b. Increasing expenditure and investment in infrastructure.
    c. Attracting large companies to establish their business in under developed states to create a greater number of jobs.

PROBLEM 2: CART – RF – ANN

Problem Statement:

Mortality Outcomes for Females Suffering Myocardial Infarction. The mifem data frame has 1295 rows and 10 columns. This is a Dataset of females having coronary heart disease (CHD). You have to predict with the given information whether the female is dead or alive so as to discover important factors that should be considered crucial in the treatment of the disease. Use CART, RF & ANN, and compare the models' performances in train and test sets.

Q2.1. Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it?

Sample of the Dataset:

	outcome	age	yronset	premi	smstat	diabetes	highbp	hichol	angina	stroke
0	live	63	85	n	x	n	y	y	n	n
1	live	55	85	n	c	n	y	y	n	n
2	live	68	85	y	nk	nk	y	y	y	n
3	live	64	85	n	x	n	y	n	y	n
4	dead	67	85	n	nk	nk	nk	y	nk	nk

Table 18. Sample of CHD Dataset.

Data Types of the Features:

Feature	outcome	age	yronset	premi	smstat	diabetes	highbp	hichol	angina	stroke
Data Type	object	int64	int64	object	object	object	object	object	object	object

Table 19. Data Types of the Features in CHD Dataset.

Insights:

- There are 1295 observations (rows) and 10 features (columns) in the dataset.
- All features in the dataset of object type except age and yronset features
- There are 75 duplicate records (rows) in the dataset which have to be removed before proceeding to next step.
- The size of the dataset after removing duplicate records is (1220 records, 10 features).

Basic Information of the Dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1295 entries, 0 to 1294
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   outcome   1295 non-null   object
 1   age       1295 non-null   int64
 2   yronset   1295 non-null   int64
 3   premi     1295 non-null   object
 4   smstat    1295 non-null   object
 5   diabetes  1295 non-null   object
 6   highbp    1295 non-null   object
 7   hichol    1295 non-null   object
 8   angina    1295 non-null   object
 9   stroke    1295 non-null   object
dtypes: int64(2), object(8)
memory usage: 101.3+ KB
```

Table 20. Basic Information of the CHD Dataset.

Summary of the Dataset:

	outcome	age	yronset	premi	smstat	diabetes	highbp	hichol	angina	stroke
count	1295	1295.000000	1295.000000	1295	1295	1295	1295	1295	1295	1295
unique	2	NaN	NaN	3	4	3	3	2	3	3
top	live	NaN	NaN	n	n	n	y	n	n	n
freq	974	NaN	NaN	928	522	978	813	655	724	1063
mean	NaN	60.922008	88.785328	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	NaN	7.042327	2.553647	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	NaN	35.000000	85.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	NaN	57.000000	87.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	NaN	63.000000	89.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	NaN	66.000000	91.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	NaN	69.000000	93.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 21. Summary of the CHD Dataset.

Checking for Missing Values:

Feature	outcome	age	yronset	premi	smstat	diabetes	highbp	hichol	angina	stroke
Count of Null Values	0	0	0	0	0	0	0	0	0	0

Table 22. Count of Null Values in Each Feature of the CHD Dataset.

- There are no null values in the given dataset

Checking for Anomalies

Getting Unique Counts of All Categorical Variables:

```
Value Counts of  outcome :
live    903
dead    317
Name: outcome, dtype: int64
-------------------------
Value Counts of  premi :
n     864
y     302
nk     54
Name: premi, dtype: int64
-------------------------
Value Counts of  smstat :
n     488
c     361
x     270
nk    101
Name: smstat, dtype: int64
-------------------------
Value Counts of  diabetes :
n     912
y     241
nk     67
Name: diabetes, dtype: int64
```

```
Value Counts of  highbp :
y     764
n     382
nk     74
Name: highbp, dtype: int64
-------------------------
Value Counts of  hichol :
y    614
n    606
Name: hichol, dtype: int64
-------------------------
Value Counts of  angina :
n     664
y     459
nk     97
Name: angina, dtype: int64
-------------------------
Value Counts of  stroke :
n     993
y     150
nk     77
Name: stroke, dtype: int64
```

Table 23. Unique Counts of All Categorical Variables

Getting Unique Entries of All Variables:

```
Feature: outcome
['live' 'dead']
-----------------------------------------------
Feature: age
[63 55 68 64 67 66 46 59 56 60 61 69 52 62 54 57 49 53 35 65 58 45 51 43
 48 50 47 38 41 44 40 42 39 36]
-----------------------------------------------
Feature: yronset
[85 86 87 88 89 90 91 92 93]
-----------------------------------------------
Feature: premi
['n' 'y' 'nk']
```

```
Feature: smstat
['x' 'c' 'nk' 'n']
-------------------------------------------------
Feature: diabetes
['n' 'nk' 'y']
-------------------------------------------------
Feature: highbp
['y' 'nk' 'n']
-------------------------------------------------
Feature: hichol
['y' 'n']
-------------------------------------------------
Feature: angina
['n' 'y' 'nk']
-------------------------------------------------
Feature: stroke
['n' 'nk' 'y']
```

Table 24. Unique Entries of All Variables

Insights:

- By observing above table 23 and table 24, we can conclude that there are no anomalies in the dataset.

Checking for Outliers:

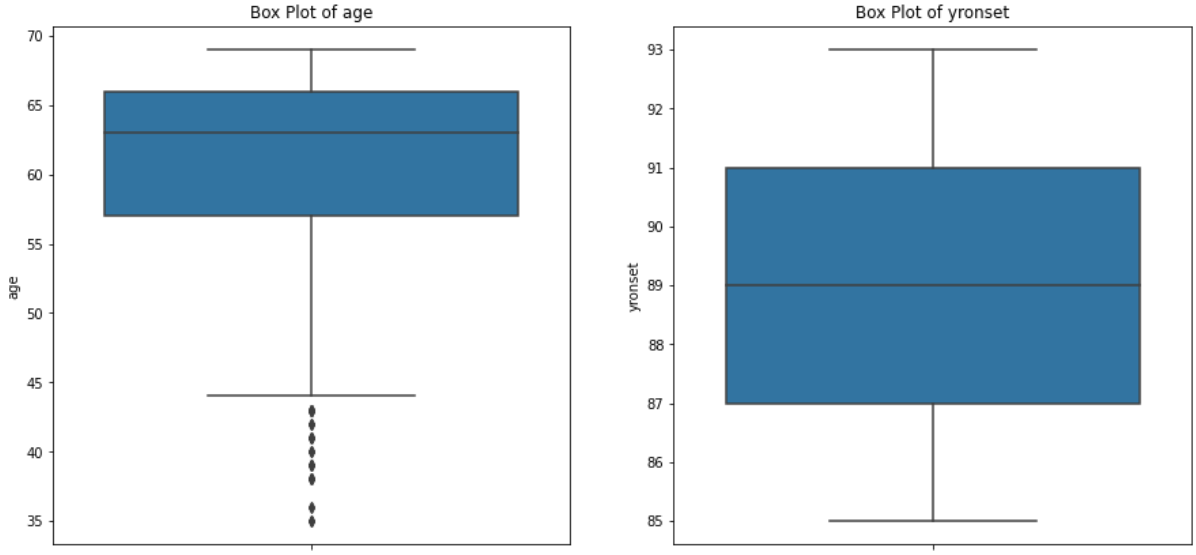


Figure 13. Box Plot of Numeric Features in the CHD Dataset

Feature	No. of Outliers	Percentage of Outliers
age	37	3.0
yronset	0	0.0

Table 25. No. of Outliers and Percentage of Outliers in the CHD Dataset.

Insights:

- There are few outliers (only 3%) in the age feature of the dataset.

Univariate Analysis

Histogram and Box Plots for Numeric Features:

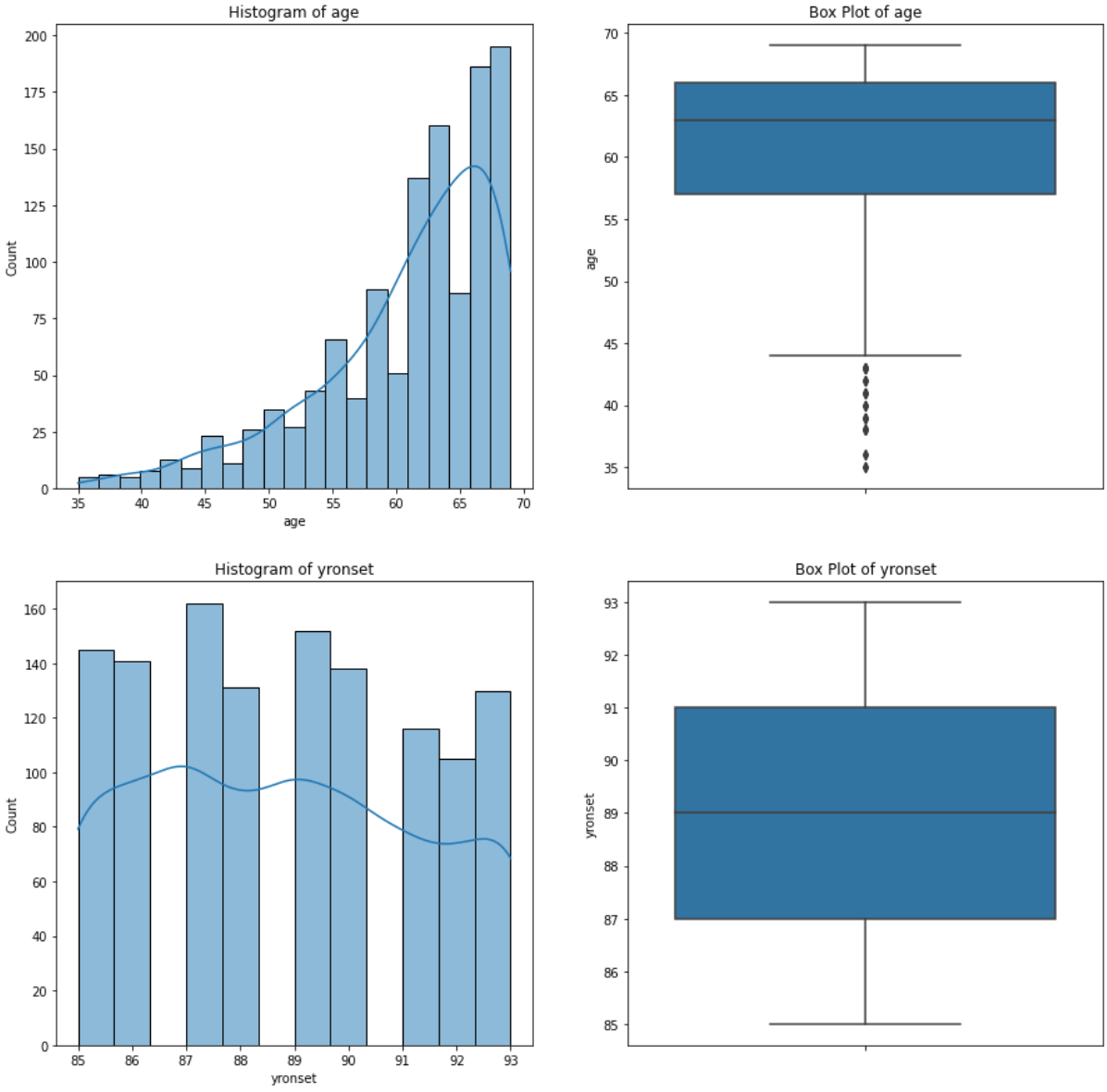


Figure 14. Histogram and Box Plots for Numeric Features in Dataset.

Skewness:

It is a measure of lack of symmetry in a distribution.

Feature	Skewness
age	-1.16
yronset	0.13

Table 26. Skewness of Numeric Features in CHD Dataset.

Kurtosis: It is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution

Feature	Kurtosis
age	0.95
yronset	-1.16

Table 27. Kurtosis of Numeric Features in CHD Dataset.

Insights:

From above plots and tables, we can conclude below points,

1. Age feature has left skewed distribution (Negatively skewed). Yronset has right skewed distribution (Positively skewed).
2. Age feature has positive kurtosis and yronset has negative kurtosis.

Count Plot of Outcome:

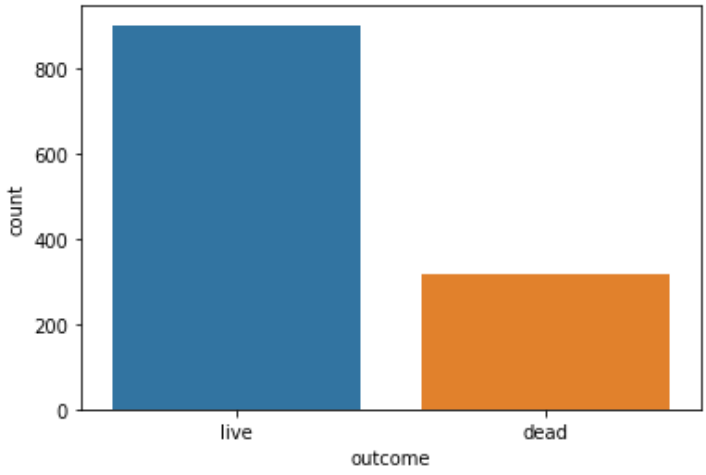


Figure 15. Count Plot of Outcome

```
live    74.0
dead    26.0
Name: outcome, dtype: float64
```

- Count plot for Outcome (Target) is drawn to check the balance of classes in target feature.
- In this dataset, we have enough number entries for both (live and dead) classes.

Bivariate Analysis

Pair Plot for Numeric Features:

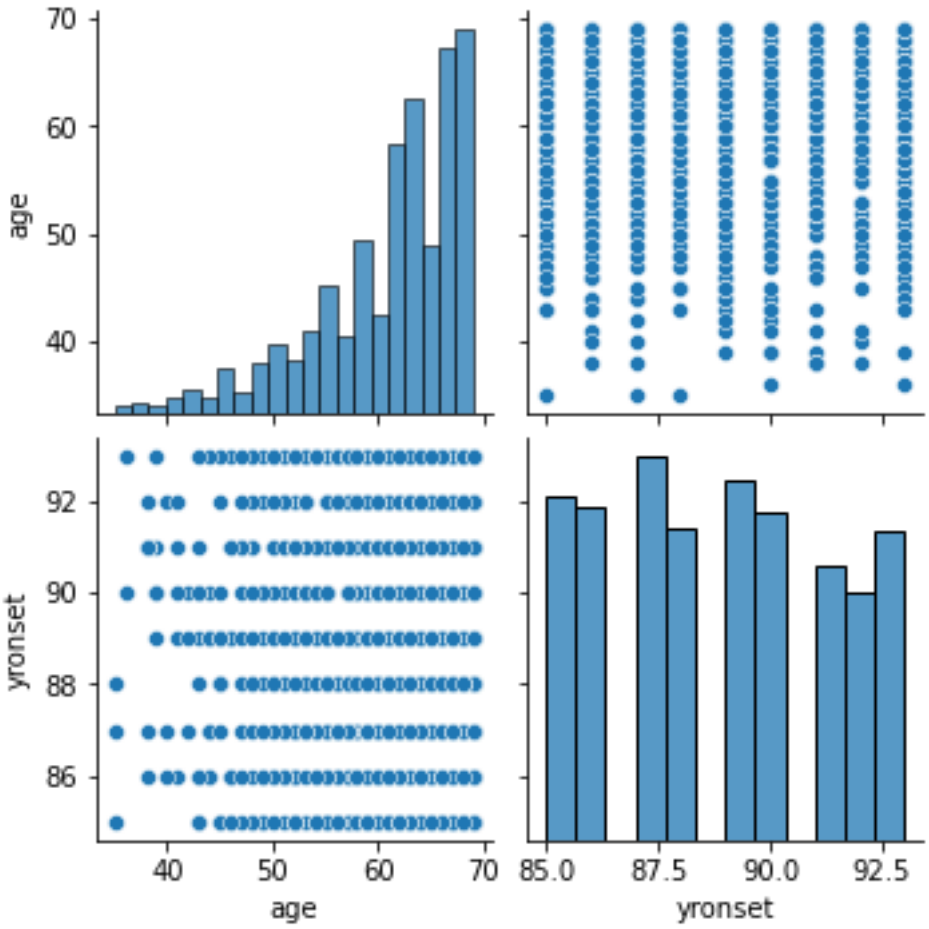


Figure 16. Pair Plot for Numeric Features in CHD Dataset.

Heat Map for Numeric Features:

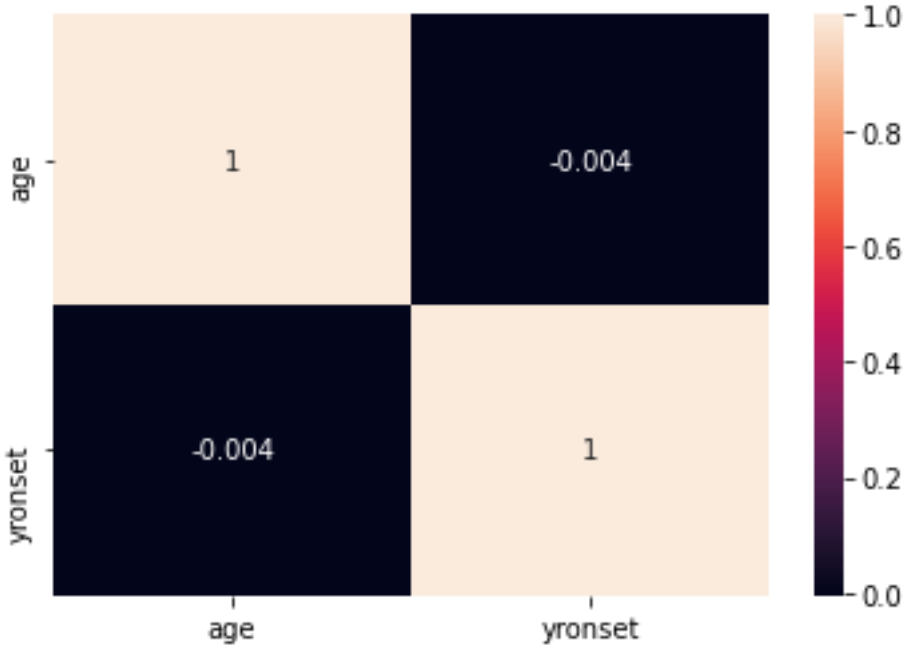
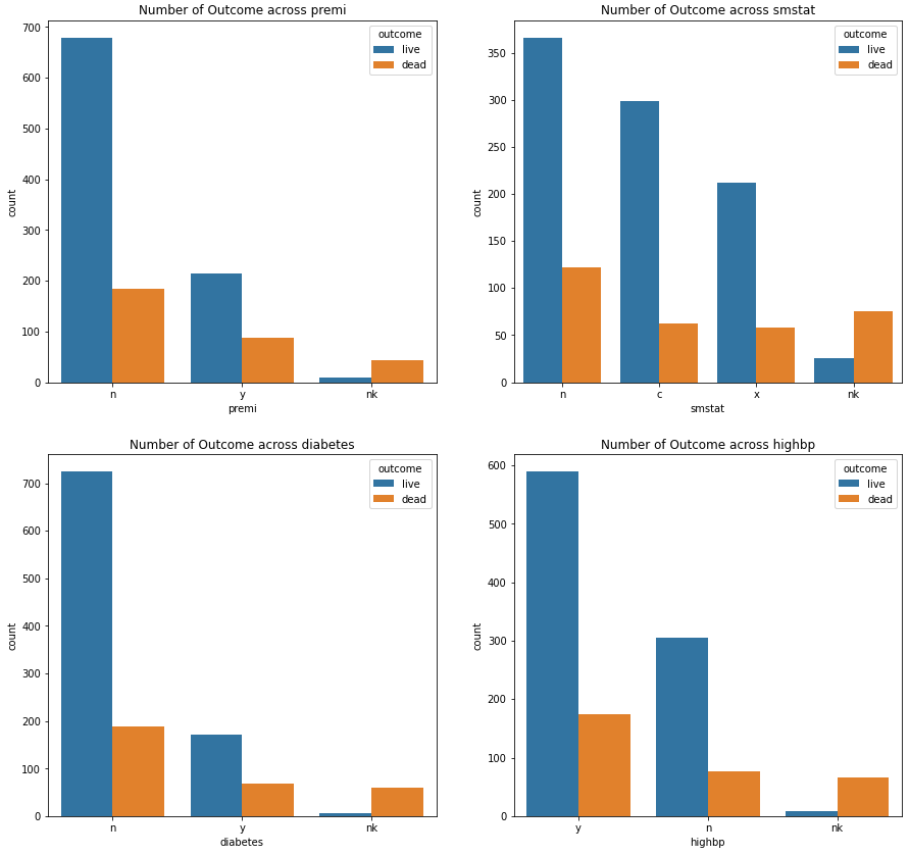


Figure 17. Heat Map for Numeric Features in CHD Dataset.

Insights:

- There is no correlation between the variables.

Count Plots of Outcome with Categorical Features:



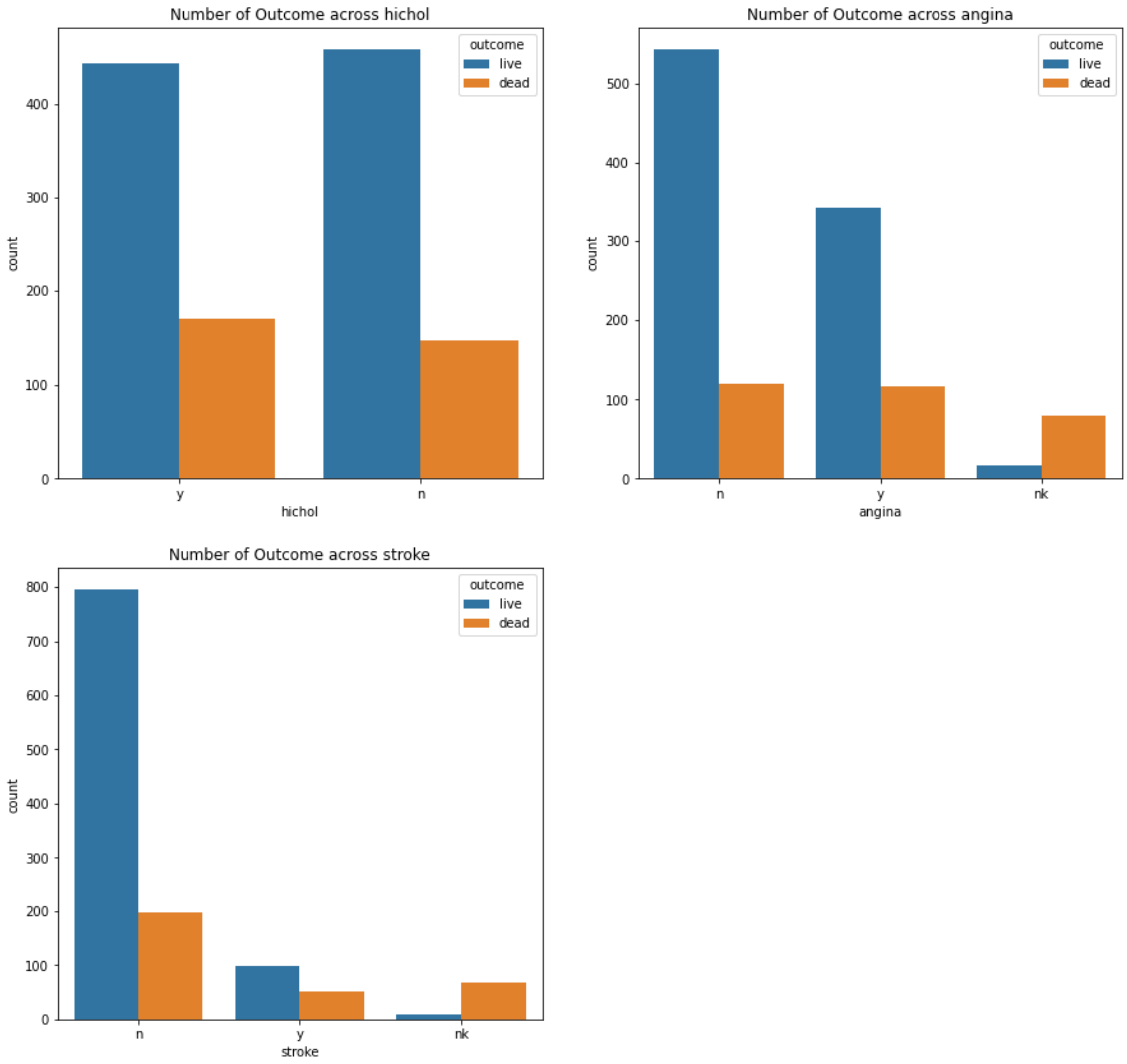


Figure 18. Count Plots of Outcome with Categorical Features.

Insights:

1. There is maximum no. of women with no previous myocardial infarction event and minimum no. of women with previous myocardial infarction event not known. In previous myocardial infarction not known category, there are more no. of dead than live.
2. There is more no. of women with non-smoke category and a smaller number of women in smoking status not known category. In smoking status not known category, there are more no. of dead than live.
3. There is more no. of women with non-diabetic category and a smaller number of women in diabetes not known category. In diabetes status not known category, there are more no. of dead than live.

4. There is more no. of women with non-high BP category and a smaller number of women in high BP not known category. In high BP status not known category, there are more no. of dead than live.
5. There is more no. of women with high cholesterol category and a smaller number of women in no high cholesterol category.
6. There is more no. of women with non-stroke category and a smaller number of women in stroke not known category. In stroke not known category, there are more no. of dead than live.

Bar Plots of Continuous Features with Outcome:

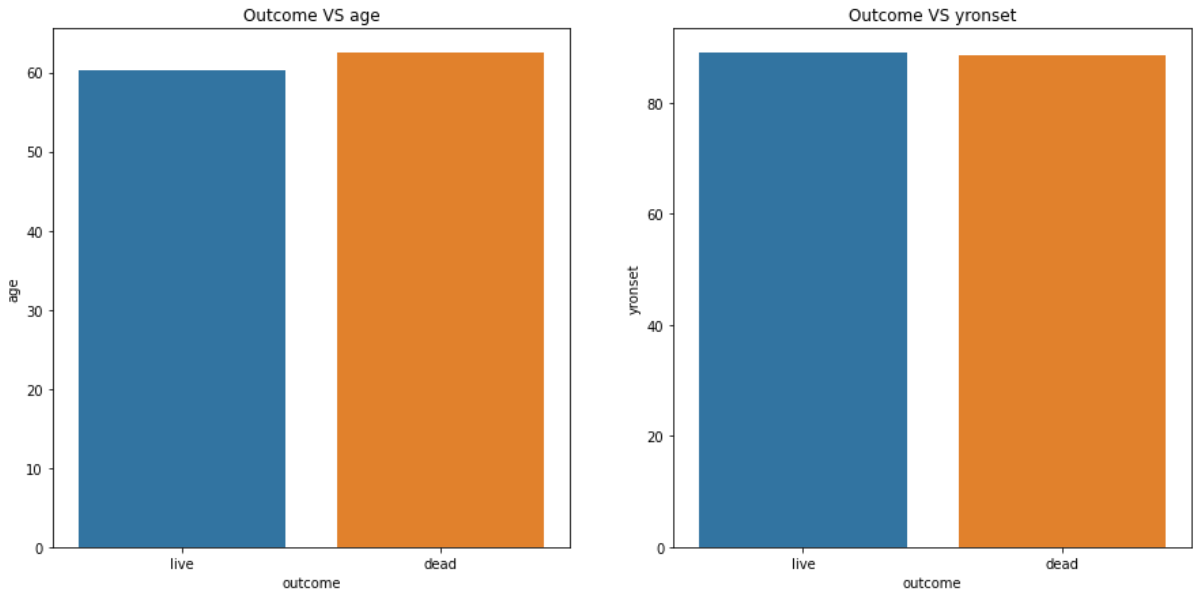


Figure 19. Bar Plots of Continuous Features with Outcome.

Insights:

1. Mean age of women who are dead is almost equal to as that of who are live.
2. Mean year of onset of women who are dead is almost equal to as that of who are live.

Q2.2. Encode the data (having string values) for Modelling. Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network.

Converting Categorical Features into Codes:

```
Feature: premi
['n' 'y' 'nk']
[0 2 1]
----------------
Feature: smstat
['x' 'c' 'nk' 'n']
[3 0 2 1]
----------------
Feature: diabetes
['n' 'nk' 'y']
[0 1 2]
----------------
Feature: highbp
['y' 'nk' 'n']
[2 1 0]
```

```
Feature: hichol
['y' 'n']
[1 0]
----------------
Feature: angina
['n' 'y' 'nk']
[0 2 1]
----------------
Feature: stroke
['n' 'nk' 'y']
[0 1 2]
```

Table 28. Integer Codes of Categorical Features in CHD Dataset.

Data Types of Encoded Dataset:

Feature	Data Type
outcome	int32
age	int64
yronset	int64
premi	int8
smstat	int8
diabetes	int8
highbp	int8
hichol	int8
angina	int8
stroke	int8

Table 29. Data Types of Encoded CHD Dataset.

Sample of Encoded Dataset:

	outcome	age	yronset	premi	smstat	diabetes	highbp	hichol	angina	stroke
0	0	63	85	0	3	0	2	1	0	0
1	0	55	85	0	0	0	2	1	0	0
2	0	68	85	2	2	1	2	1	2	0
3	0	64	85	0	3	0	2	0	2	0
4	1	67	85	0	2	1	1	1	1	1

Table 30. Sample Encoded CHD Dataset.

Splitting the Dataset into Independent and Dependent Features:

Independent Features:

	age	yronset	premi	smstat	diabetes	highbp	hichol	angina	stroke
0	63	85	0	3	0	2	1	0	0
1	55	85	0	0	0	2	1	0	0
2	68	85	2	2	1	2	1	2	0
3	64	85	0	3	0	2	0	2	0
4	67	85	0	2	1	1	1	1	1

Dependent (Target) Feature:

```
0    0
1    0
2    0
3    0
4    1
5    0
6    0
7    1
8    1
9    1
Name: outcome, dtype: int32
```

Proportion of 1s (YES) and 0s (NO):

```
0    74.0
1    26.0
Name: outcome, dtype: float64
```

- There is no issue of class imbalance here as we have reasonable proportions in both the classes.

Splitting the Data into Train and Test Sets:

- Both independent and target datasets have been divided into train and test sets.
- No. of observations in test set is selected as 0.33 times of total data points.
- Then no. of observations in train set will be 0.66 times of total data points.

Checking the Training and Test Data:

```
Size of xtrain: (817, 9)
Size of xtest: (403, 9)
Size of ytrain: (817,)
Size of ytest: (403,)
```

	age	yronset	premi	smstat	diabetes	highbp	hichol	angina	stroke
780	48	89	1	2	1	1	1	1	1
886	57	90	0	1	0	0	0	0	0
73	65	85	0	1	2	2	1	0	0
117	62	85	0	0	0	2	0	0	0
1092	69	92	0	1	0	2	1	1	0

	age	yronset	premi	smstat	diabetes	highbp	hichol	angina	stroke
1059	60	92	0	0	0	0	0	0	0
553	67	89	2	0	0	0	1	0	0
594	61	90	2	0	0	2	1	1	0
267	55	86	2	3	2	2	0	2	2
81	60	85	0	1	0	0	1	2	0

Table 31. Samples of Dependent Train and Test Datasets.

```
780     1
886     0
73      0
117     0
1092    1
Name: outcome, dtype: int32
```

```
1059    0
553     1
594     0
267     0
81      0
Name: outcome, dtype: int32
```

Table 32. Sample of Target Train and Test Datasets.

Distribution of Target Class in Train and Test sets:

```
0    74.1
1    25.9
Name: outcome, dtype: float64
```

```
0    73.9
1    26.1
Name: outcome, dtype: float64
```

Table 32. Unique Counts of Target Class in Train and Test Datasets.

- From above table, we can notice that target class (0s and 1s) is almost uniformly distributed between train and test datasets.

Building Decision Tree Classifier Model (CART)

Below Hyper Parameters have been selected in CART model to optimize by using GridSearchCV.

- Maximum Depth: [5,6,7,8] – It is the depth upto which decision tree is allowed to grow. It is selected based the depth upto which tree has grown uniformly without tuning hyper parameters.
- Minimum samples leaf: [10,15,20] – It is the minimum samples required in leaf node to validate the node as leaf node. It is selected approximately 1-2% of observations in train dataset.
- Minimum samples split: [30,45,60] – It is the minimum samples required in a node to split it. It is selected approximately three times of minimum samples leaf.

Best Parameters:

Below are the best parameters obtained in CART model by using GridSearchCV

- Maximum Depth: 5, Minimum samples leaf: 20, Minimum samples split: 60.

Features Importance in CART model:

Feature	stroke	age	yronset	smstat	premi	hichol	angina	diabetes	highbp
Importance	0.73	0.08	0.05	0.04	0.04	0.04	0.02	0.0	0.0

Table 33. Features Importance in CART model.

From above table, we can notice that stroke is the most important feature in classifying or predicting the class of target variable.

Building Random Forest Model

Below Hyper Parameters have been selected in Random Forest model to optimize by using GridSearchCV.

- Number of Estimators: [51,101,151] – It is the number of trees considered in a Random Forest model. It is selected by default as 101 and added additionally 51 and 151 to optimize.

- Maximum Features: [2,3,4] – It is the number of features considered in each split. It is selected approximately square root of number of features.
- Maximum Depth: [5,6,7,8] – It is the depth upto which decision tree is allowed to grow. It is selected based the depth upto which tree has grown uniformly without tuning hyper parameters.
- Minimum samples leaf: [10,15,20] – It is the minimum samples required in leaf node to validate the node as leaf node. It is selected approximately 1-2% of observations in train dataset.
- Minimum samples split: [30,45,60] – It is the minimum samples required in a node to split it. It is selected approximately three times of minimum samples leaf.

Best Parameters:

Below are the best parameters obtained in Random Forest by using GridSearchCV

- Number of Estimators: 51, Maximum Features: 3, Maximum Depth: 6, Minimum samples leaf: 15, Minimum samples split: 30.

Features Importance in Random Forest model:

Feature	Importance
stroke	0.281
angina	0.166
age	0.140
smstat	0.115
diabetes	0.110
yronset	0.088
premi	0.040
highbp	0.039
hichol	0.020

Table 34. Features Importance in Random Forest model

From above table, we can notice that stroke is the most important feature in classifying or predicting the class of target variable.

Building Artificial Neural Network Model

Below Hyper Parameters have been selected in Artificial Neural Network model to optimize by using GridSearchCV.

- Hidden Layer Sizes: [50,100,150] – Selected default neurons as 100 and added additionally 50 and 150 to optimize.
- Tolerance: [0.01,0.001,0.0001] – Selected default tolerance as 0.001 and added additionally 0.01 and 0.0001 to optimize.
- Maximum Iterations: [500,1000,1500] – Selected default maximum iterations as 1000 and added additionally 500 and 1500 to optimize.

Best Parameters:

Below are the best parameters obtained in Random Forest by using GridSearchCV

- Hidden Layer Sizes: 100, Tolerance: 0.001, Maximum Iterations: 500

Q2.3. Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve, and get ROC_AUC score for each model.

Decision Tree Classifier Model (CART) Evaluation

Model Evaluation Based on Train Set:

Confusion Matrix:

	Predicted 0	Predicted 1
Actual 0	596	9
Actual 1	165	47

Table 35. Confusion Matrix for Train Dataset in CART model.

Classification Report:

	precision	recall	f1-score	support
0	0.78	0.99	0.87	605.00
1	0.84	0.22	0.35	212.00
accuracy	0.79	0.79	0.79	0.79
macro avg	0.81	0.60	0.61	817.00
weighted avg	0.80	0.79	0.74	817.00

Table 36. Classification Report for Train Dataset in CART model.

ROC Curve:

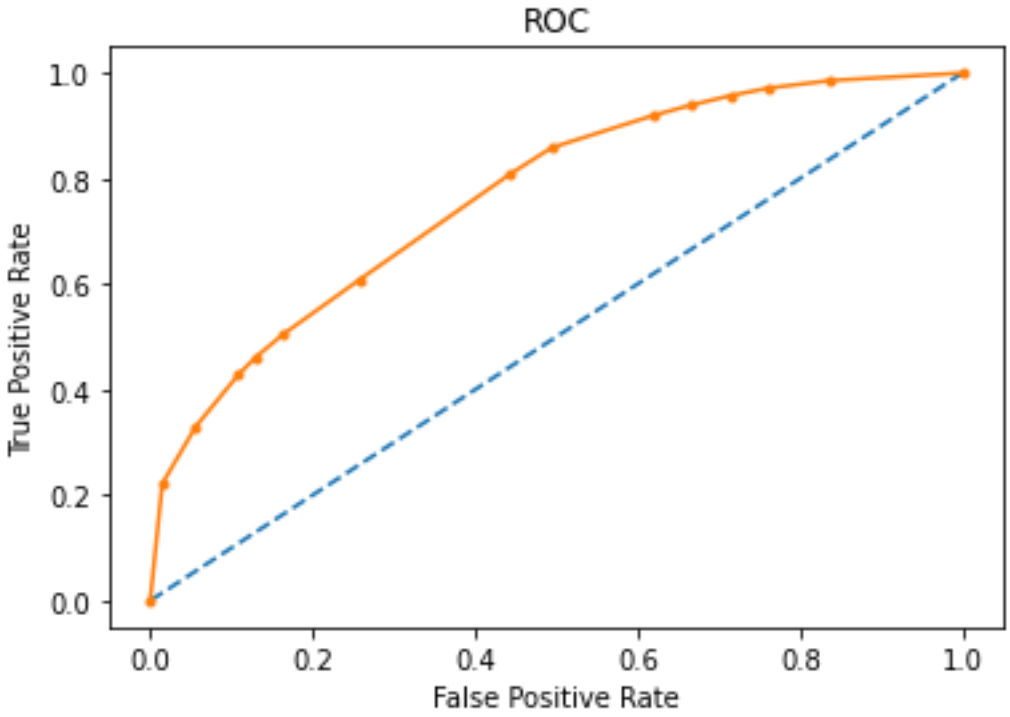


Figure 21. ROC Curve for Train Dataset in CART model.

Accuracy:

Accuracy of the model is 0.79

ROC AUC Score:

ROC AUC Score of the model is 0.77

Model Evaluation Based on Test Set:

Confusion Matrix:

	Predicted 0	Predicted 1
Actual 0	298	0
Actual 1	84	21

Table 37. Confusion Matrix for Test Dataset in CART model.

Accuracy:

Accuracy of the model is 0.79

ROC AUC Score:

ROC AUC Score of the model is 0.65

Classification Report:

	precision	recall	f1-score	support
0	0.78	1.00	0.88	298.00
1	1.00	0.20	0.33	105.00
accuracy	0.79	0.79	0.79	0.79
macro avg	0.89	0.60	0.60	403.00
weighted avg	0.84	0.79	0.73	403.00

Table 38. Classification Report for Test Dataset in CART model.

ROC Curve:

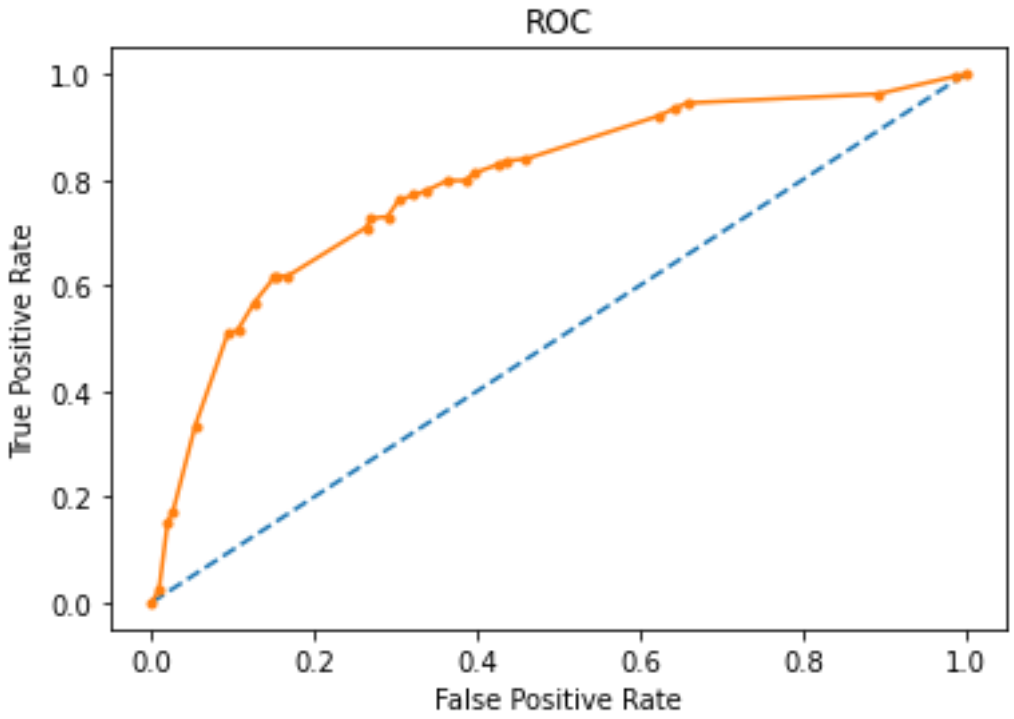


Figure 22. ROC Curve for Test Dataset in CART model.

Conclusions:

Data	**Decision Tree Model**				
	Precision for Class 1	**Recall for Class 1**	**F1 Score for Class 1**	**AUC Score for Class 1**	**Accuracy**
Train Dataset	0.84	0.22	0.35	0.77	0.79

Test Dataset	1	0.2	0.33	0.65	0.79

Table 39. Performance Metrics of CART model

From above table, we can conclude below points.

1. The performance metrics like precision, recall, F1 score, AUC Score and Accuracy for test dataset are approaching train dataset. Hence, there is no over fitting in the model.
2. Accuracy of the model on test dataset (0.79) is more than 0.75. Hence, the model is considered as good model and can be used for predictions.
3. ROC AUC Score of the model on test dataset (0.79) is more than 0.75. Hence, the model is considered as good model and can be used for predictions.
4. Recall on test dataset is 0.2. This is very low. We should work on this metric to improve it by consulting with business.
5. Precision on test dataset is 1. This is good enough to use the model for predictions.
6. F1 score on test dataset is 0.33. This is not sufficient enough to use the model for predictions. We should consult with the business to check correctness of the data and to improve the F1 score.

Random Forest Model Evaluation

Model Evaluation Based on Train Set:

Confusion Matrix:

	Predicted 0	Predicted 1
Actual 0	600	5
Actual 1	164	48

Table 40. Confusion Matrix for Train Dataset in Random Forest model.

Classification Report:

	precision	recall	f1-score	support
0	0.79	0.99	0.88	605.00
1	0.91	0.23	0.36	212.00
accuracy	0.79	0.79	0.79	0.79
macro avg	0.85	0.61	0.62	817.00
weighted avg	0.82	0.79	0.74	817.00

Table 41. Classification Report for Train Dataset in Random Forest model.

Accuracy:

Accuracy of the model is 0.79

ROC AUC Score:

ROC AUC Score of the model is 0.82

ROC Curve:

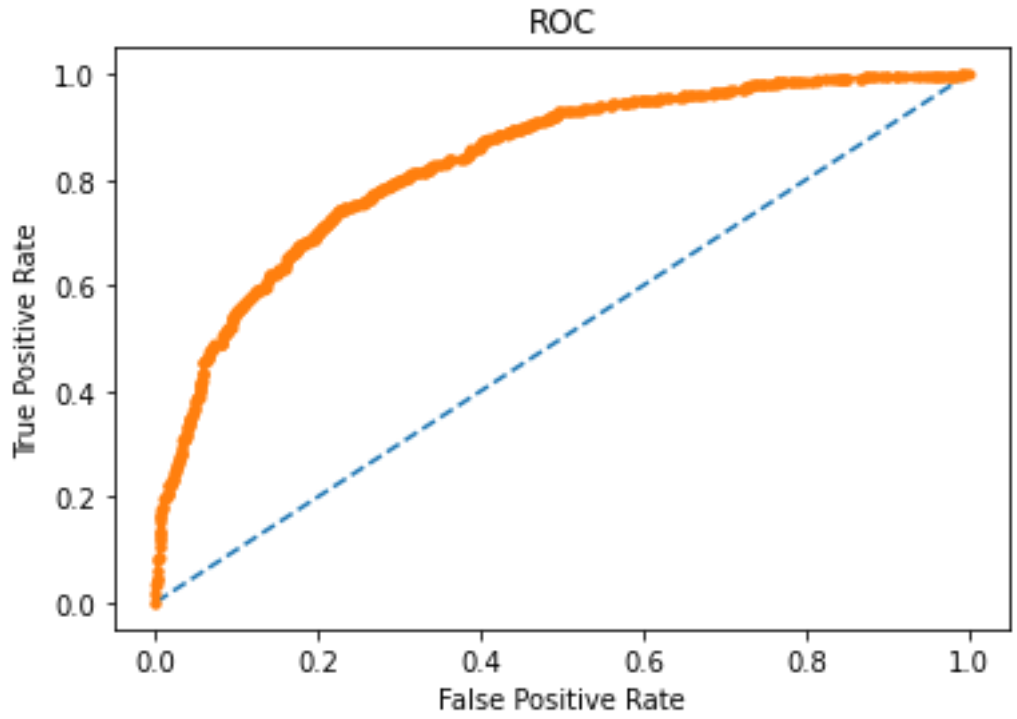


Figure 23. ROC Curve for Train Dataset in Random Forest model.

Model Evaluation Based on Test Set:

Confusion Matrix:

	Predicted 0	Predicted 1
Actual 0	298	0
Actual 1	82	23

Table 42. Confusion Matrix for Test Dataset in Random Forest model.

Accuracy:

Accuracy of the model is 0.8

ROC AUC Score:

ROC AUC Score of the model is 0.67

Classification Report:

	precision	recall	f1-score	support
0	0.78	1.00	0.88	298.0
1	1.00	0.22	0.36	105.0
accuracy	0.80	0.80	0.80	0.8
macro avg	0.89	0.61	0.62	403.0
weighted avg	0.84	0.80	0.74	403.0

Table 43. Classification Report for Test Dataset in Random Forest model.

ROC Curve:

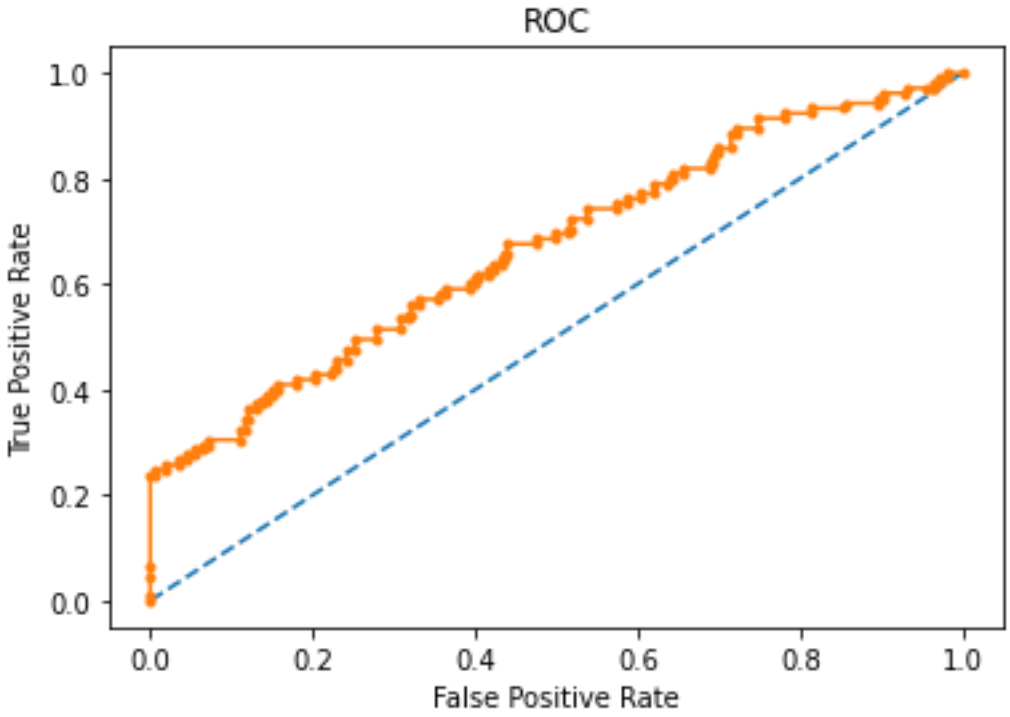


Figure 24. ROC Curve for Test Dataset in Random Forest model.

Conclusions:

Data	**Random Forest Model**				
	Precision for Class 1	**Recall for Class 1**	**F1 Score for Class 1**	**AUC for Class 1**	**Accuracy**
Train Dataset	0.91	0.23	0.36	0.82	0.79

Test Dataset	1	0.22	0.36	0.67	0.8

Table 44. Performance Metrics of Random Forest model

From above table, we can conclude below points.

1. The performance metrics like precision, recall, F1 score, AUC Score and Accuracy for test dataset are approaching train dataset. Hence, there is no over fitting in the model.
2. Accuracy of the model on test dataset (0.8) is more than 0.75. Hence, the model is considered as good model and can be used for predictions.
3. ROC AUC Score of the model on test dataset (0.67) is little low. We should discuss with the business to improve this metric.
4. Recall on test dataset is 0.22. This is little low. We should work on this metric to improve it by consulting with business.
5. Precision on test dataset is 1. This is very high to use the model for predictions.
6. F1 score on test dataset is 0.36. We should discuss with the business to improve this metric.

Artificial Neural Network (ANN) Model Evaluation

Model Evaluation Based on Train Set:

Confusion Matrix:

	Predicted 0	Predicted 1
Actual 0	591	14
Actual 1	142	70

Table 45. Confusion Matrix for Train Dataset in Artificial Neural Network model.

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.98	0.88	605.00
1	0.83	0.33	0.47	212.00
accuracy	0.81	0.81	0.81	0.81
macro avg	0.82	0.65	0.68	817.00
weighted avg	0.81	0.81	0.78	817.00

Table 46. Classification Report for Train Dataset in Artificial Neural Network model.

Accuracy:

Accuracy of the model is 0.81

ROC AUC Score:

ROC AUC Score of the model is 0.76

ROC Curve:

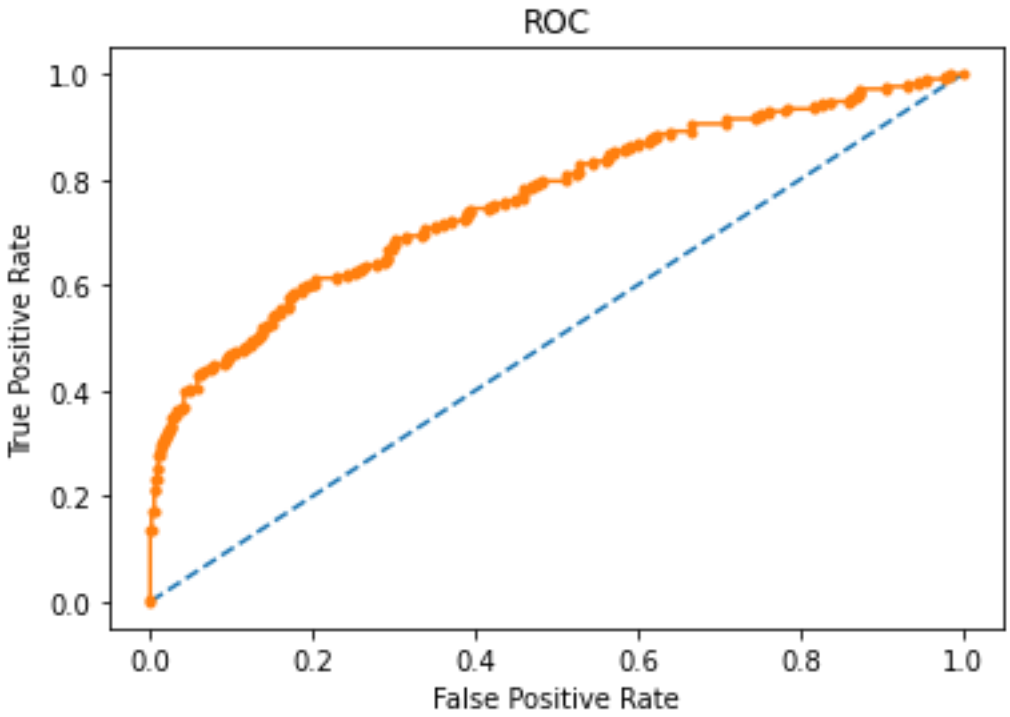


Figure 25. ROC Curve for Train Dataset in Artificial Neural Network model.

Model Evaluation Based on Test Set:

Confusion Matrix:

	Predicted 0	Predicted 1
Actual 0	282	16
Actual 1	77	28

Table 47. Confusion Matrix for Test Dataset in Artificial Neural Network model.

Accuracy:

Accuracy of the model is 0.77

ROC AUC Score:

ROC AUC Score of the model is 0.67

Classification Report:

	precision	recall	f1-score	support
0	0.79	0.95	0.86	298.00
1	0.64	0.27	0.38	105.00
accuracy	0.77	0.77	0.77	0.77
macro avg	0.71	0.61	0.62	403.00
weighted avg	0.75	0.77	0.73	403.00

Table 48. Classification Report for Test Dataset in Artificial Neural Network model.

ROC Curve:

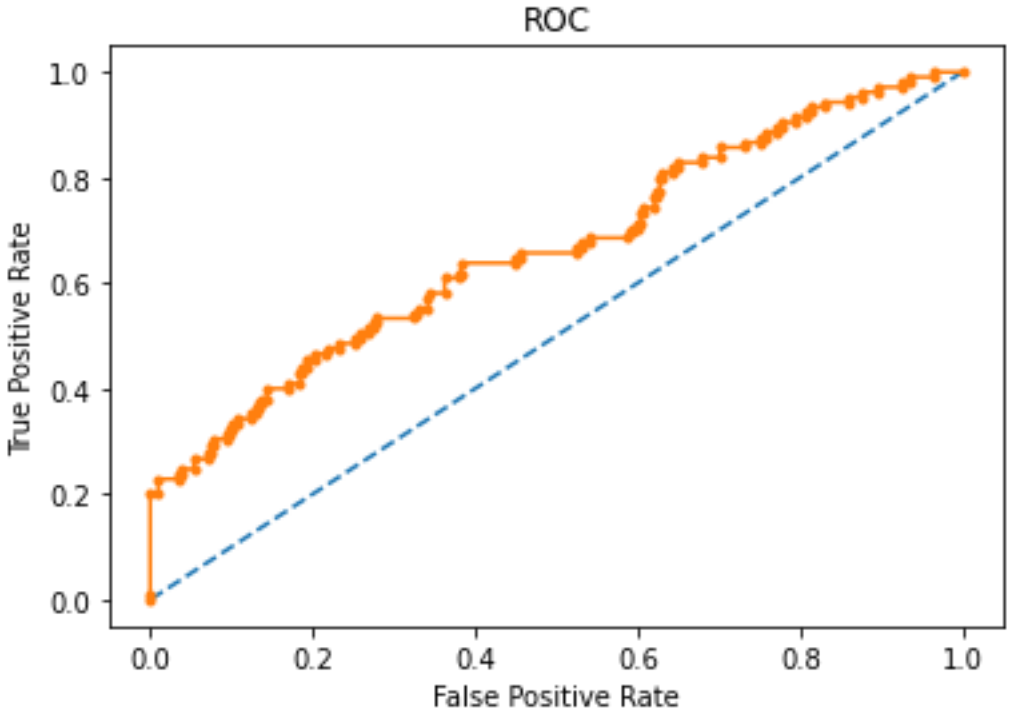


Figure 26. ROC Curve for Test Dataset in Artificial Neural Network model.

Conclusions:

Data	**Artificial Neural Network Model**				
	Precision for Class 1	**Recall for Class 1**	**F1 Score for Class 1**	**AUC for Class 1**	**Accuracy**
Train Dataset	0.83	0.33	0.47	0.76	0.81

Test Dataset	0.64	0.27	0.38	0.67	0.77

Table 49. Performance Metrics of Random Forest model

From above table, we can conclude below points.

1. The performance metrics like precision, recall, F1 score, AUC Score and Accuracy for test dataset are approaching train dataset. Hence, there is no over fitting in the model.
2. Accuracy of the model on test dataset (0.77) is more than 0.75. Hence, the model is considered as good model and can be used for predictions.
3. ROC AUC Score of the model on test dataset (0.81) is more than 0.75. Hence, the model is considered as good model and can be used for predictions.
4. Recall on test dataset is 0.27. This is little low. We should work on this metric to improve it by consulting with business.
5. Precision on test dataset is 0.64. This is sufficient enough to use the model for predictions.
6. F1 score on test dataset is 0.38. This is little low. We should work on this metric to improve it by consulting with business.

Importance of Performance Metrics:

- ✓ F1 score is a harmonic mean of recall and precision and it is more important than recall and precision. Because F1 score takes care of both of them. If either recall or precision decreases, F1 score automatically decreases drastically. Hence, F1 score is most important metric in evaluating model performance and deciding validity of the model.
- ✓ In this model, **Recall is more important than precision**. If the prediction is live but in actual if it is dead (False Negatives), it creates problem and it is not acceptable. To avoid this problem, we should reduce no. of false negatives by reducing the threshold value less than 0.5.
- ✓ If the prediction is dead but in actual if it is live (False Positives), this scenario is not a problem.

Q2.4. Final Model: Compare all the models and write an inference which model is best/optimized.

	Accuracy	AUC	Precision	Recall	F1 Score
CART Train	0.79	0.77	0.84	0.22	0.35
Random Forest Train	0.79	0.82	0.91	0.23	0.36
Neural Network Train	0.81	0.76	0.83	0.33	0.47
CART Test	0.79	0.65	1.00	0.20	0.33
Random Forest Test	0.80	0.67	1.00	0.22	0.36
Neural Network Test	0.77	0.67	0.64	0.27	0.38

Table 50. Comparison of Performance Metrics of all three models.

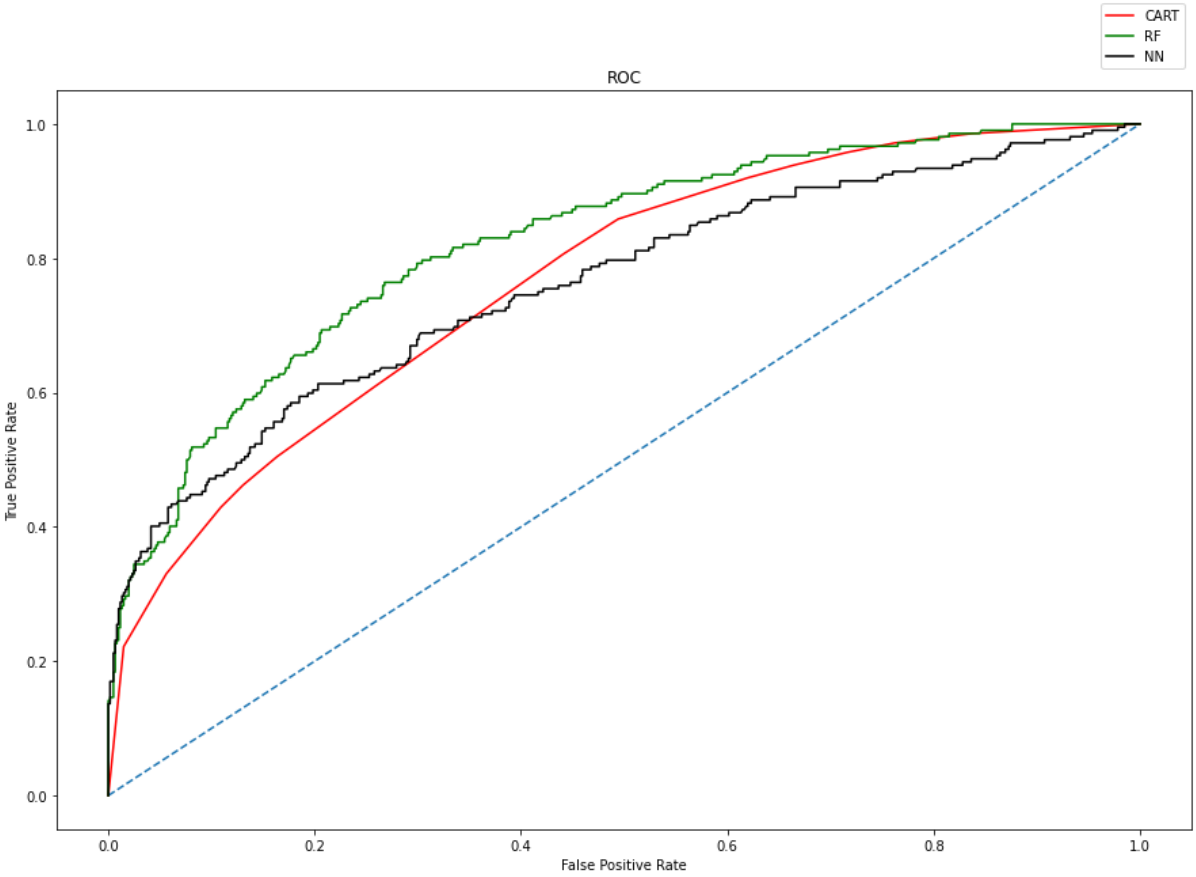


Figure 27. Comparison of ROC Curves for all three models.

Inferences:

From above table and plot, we can derive below inferences.

1. For all the models, the performance metrics of train dataset are approaching to the test dataset. Hence, there is no over fitting in any one of these models.
2. The performance metrics of test datasets for all the models is almost equal and they are not significant enough to use for predictions. **Anyway, Artificial Neural Network**

model is providing little better results (Recall and F1 Score) over CART and Random Forest models. Hence, Artificial Neural Network model is optimized among these three.

Q2.5. Inference: Basis on these predictions, what are the insights and recommendations?

Insights and Recommendations to business:

1. For all the models, the performance metrics of train dataset are approaching to the test dataset. Hence, there is **no over fitting** in any one of these models.
2. The performance metrics of test datasets for all the models is almost equal and but they are not significantly large enough to use for predictions. Hence, we should not use these models for predictions. We need to discuss with the business regarding the correctness of the data and threshold limit to improve recall and f1 score for all the models.
3. **Anyway, Artificial Neural Network model is providing little better results (Recall and F1 Score) over CART and Random Forest models. Hence, Artificial Neural Network model is optimized among these three**.
4. Threshold value can be decreased below 0.5 in all models to decrease false negatives and improve recall of the model.
5. Women who have previous myocardial infarction are more likely to die. In previous myocardial infarction **not known category**, there are more no. of dead than live. This category needs to be investigated further to get accurate effect.
6. Smoking is not influencing much on death of women. But in smoking status not known category, there are more no. of dead than live. This category needs to be investigated further to get accurate effect smoking on death.
7. Women with diabetes are more likely to die. But in diabetes status not known category, there are more no. of dead than live. This category needs to be investigated further to get accurate effect smoking on death.
8. High BP is not influencing much on death of women. In high BP status not known category, there are more no. of dead than live. This category needs to be investigated further to get accurate effect High BP on death.
9. Women with stroke history are more likely to die. In stroke not known category, there are more no. of dead than live. This category needs to be investigated further to get accurate effect stroke on death.