PROJECT

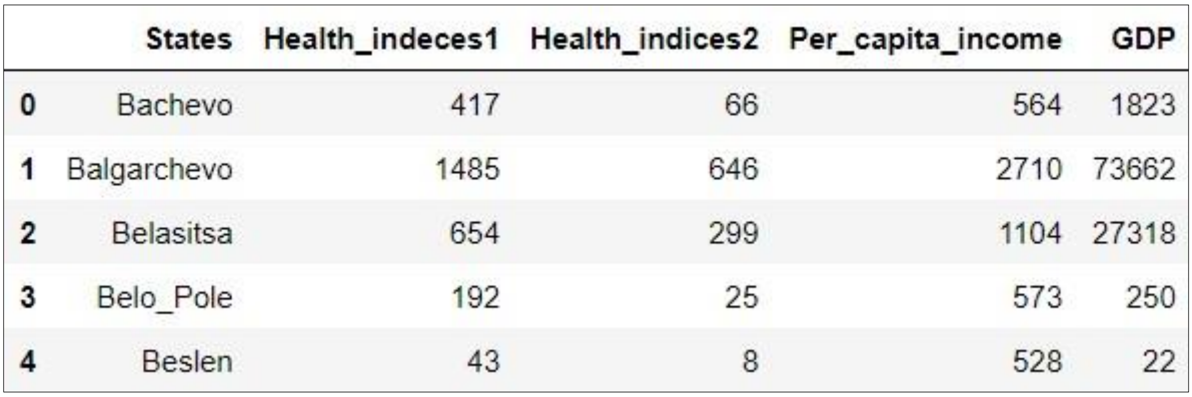
DATA MINING

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:



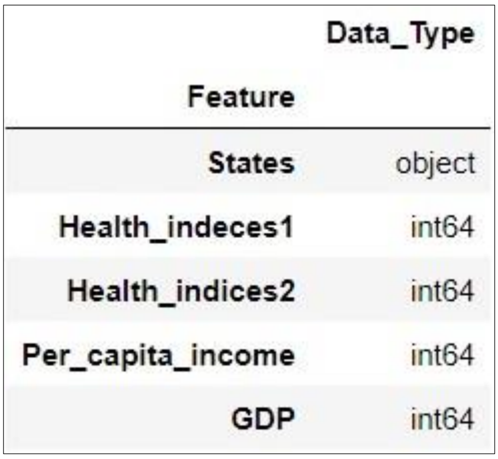
	States	Health_indeces1	Health_indices2	Per_capita_income	GDP
0	Bachevo	417	66	564	1823
1	Balgarchevo	1485	646	2710	73662
2	Belasitsa	654	299	1104	27318
3	Belo_Pole	192	25	573	250
4	Beslen	43	8	528	22

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:

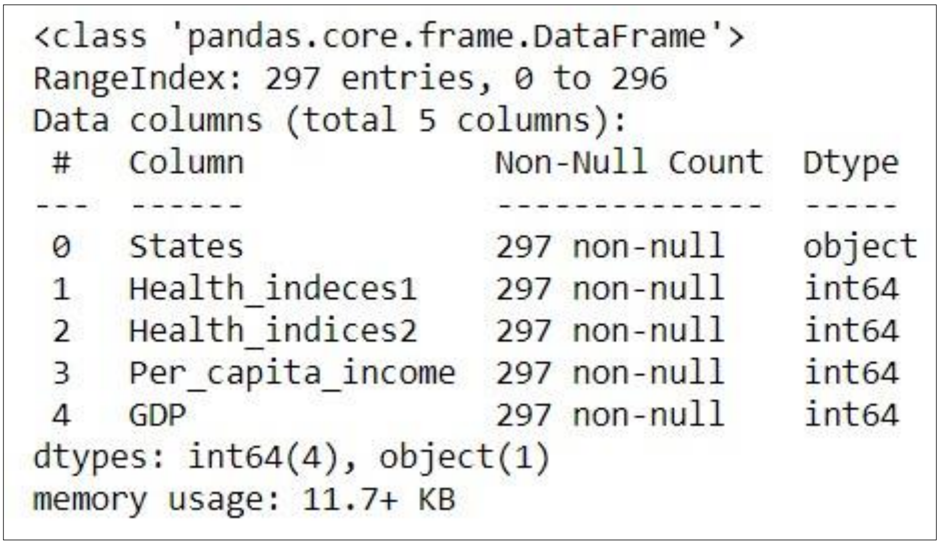


	Data_Type
Feature	
States	object
Health_indeces1	int64
Health_indices2	int64
Per_capita_income	int64
GDP	int64

Data Types of All Features in the States Dataset.

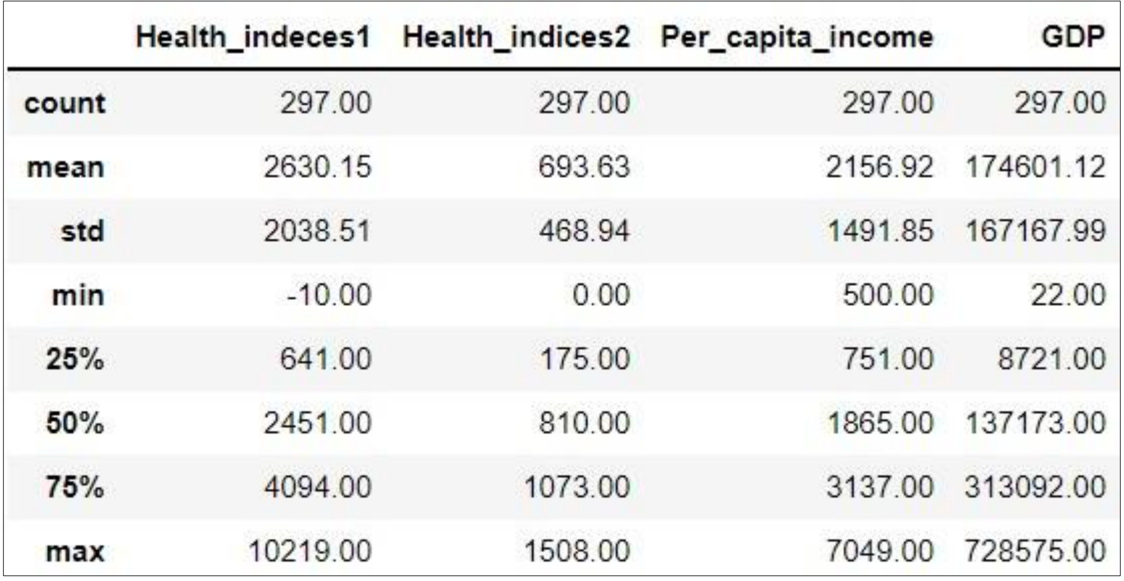
- All features (variables) in the above dataset are numeric and continuous except states feature.

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

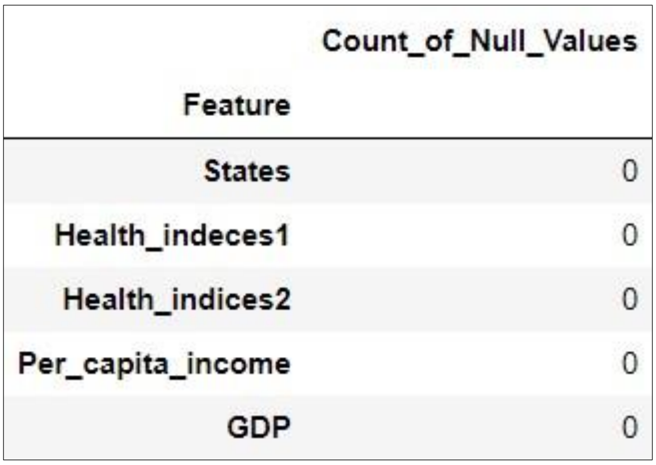
3. Summary of States Dataset.

Exploratory Data Analysis

Let us Check for duplicate observations

- There are no duplicate observations in the given dataset.

Let us Check for null values

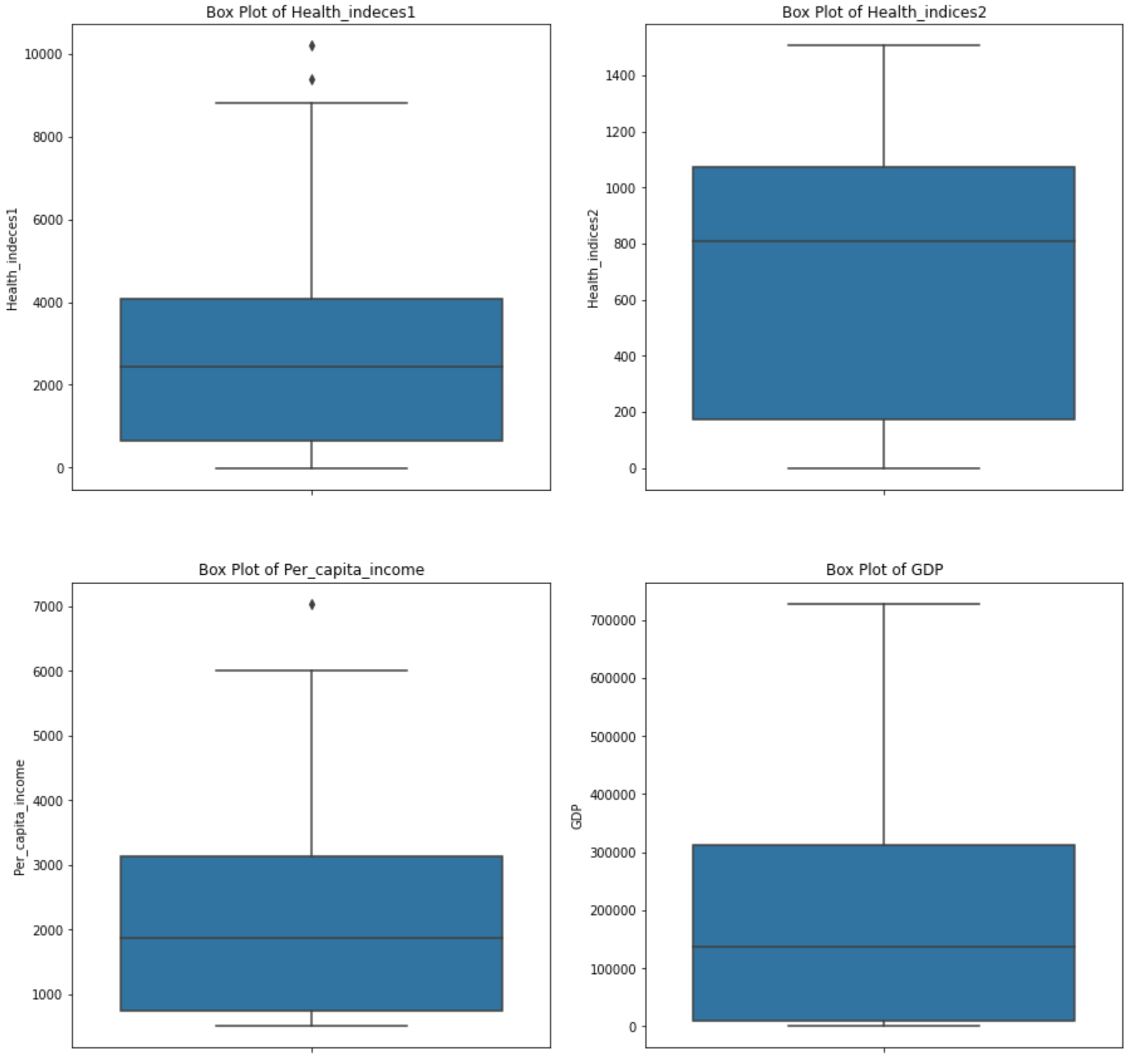


	Count_of_Null_Values
Feature	
States	0
Health_indeces1	0
Health_indices2	0
Per_capita_income	0
GDP	0

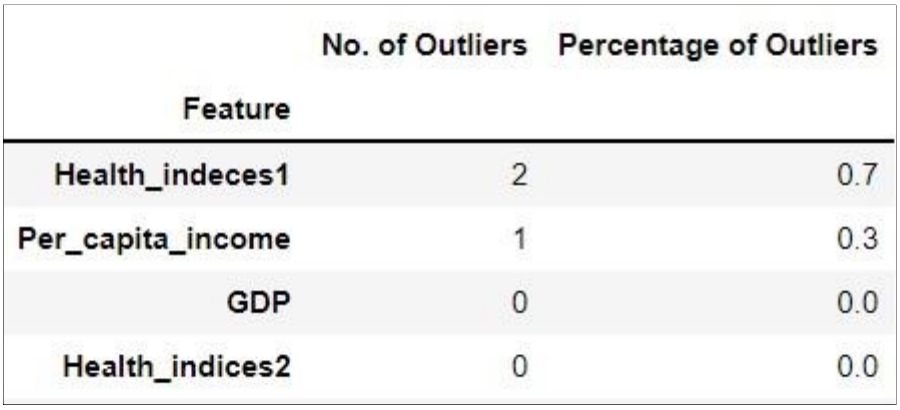
Null Values in the States Dataset.

- There are no null values in the given dataset.

Let us check the outliers in the Dataset



Box Plots of Numerical Features in States Dataset.



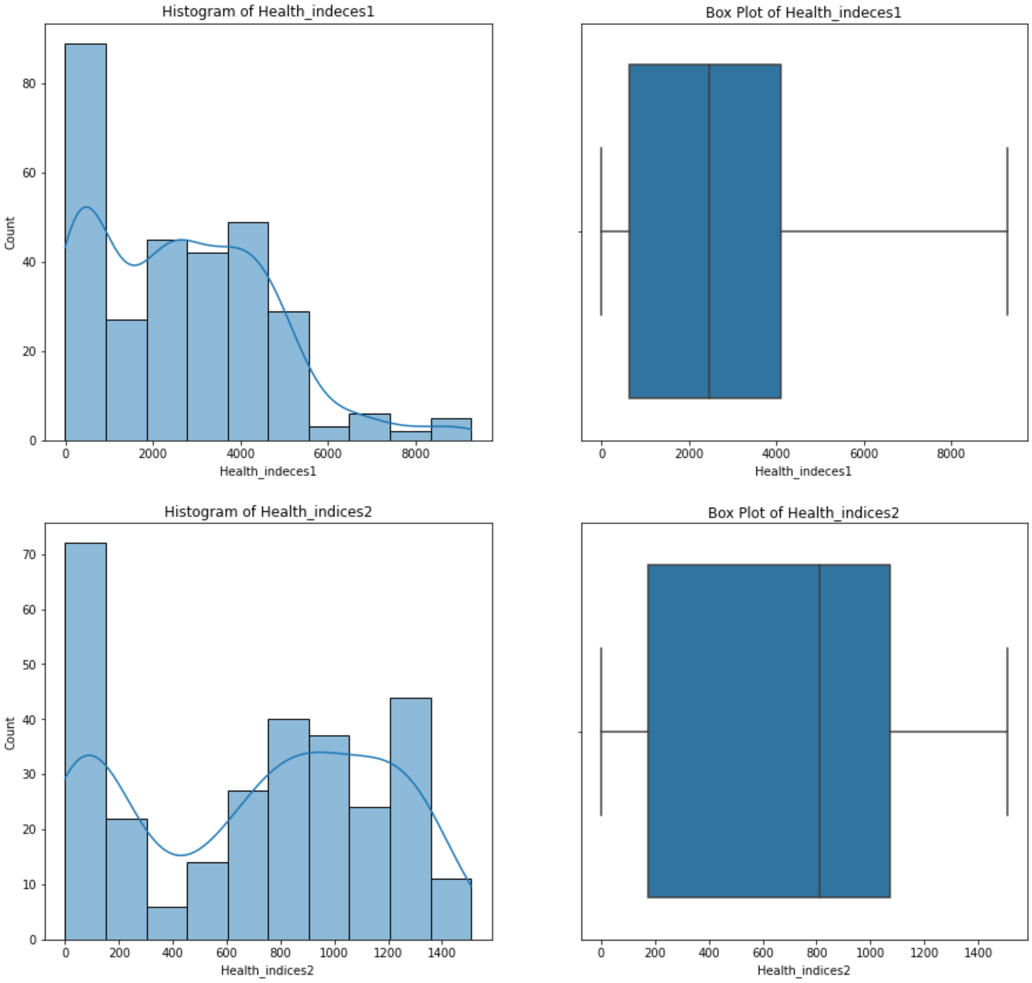
	No. of Outliers	Percentage of Outliers
Feature		
Health_indeces1	2	0.7
Per_capita_income	1	0.3
GDP	0	0.0
Health_indices2	0	0.0

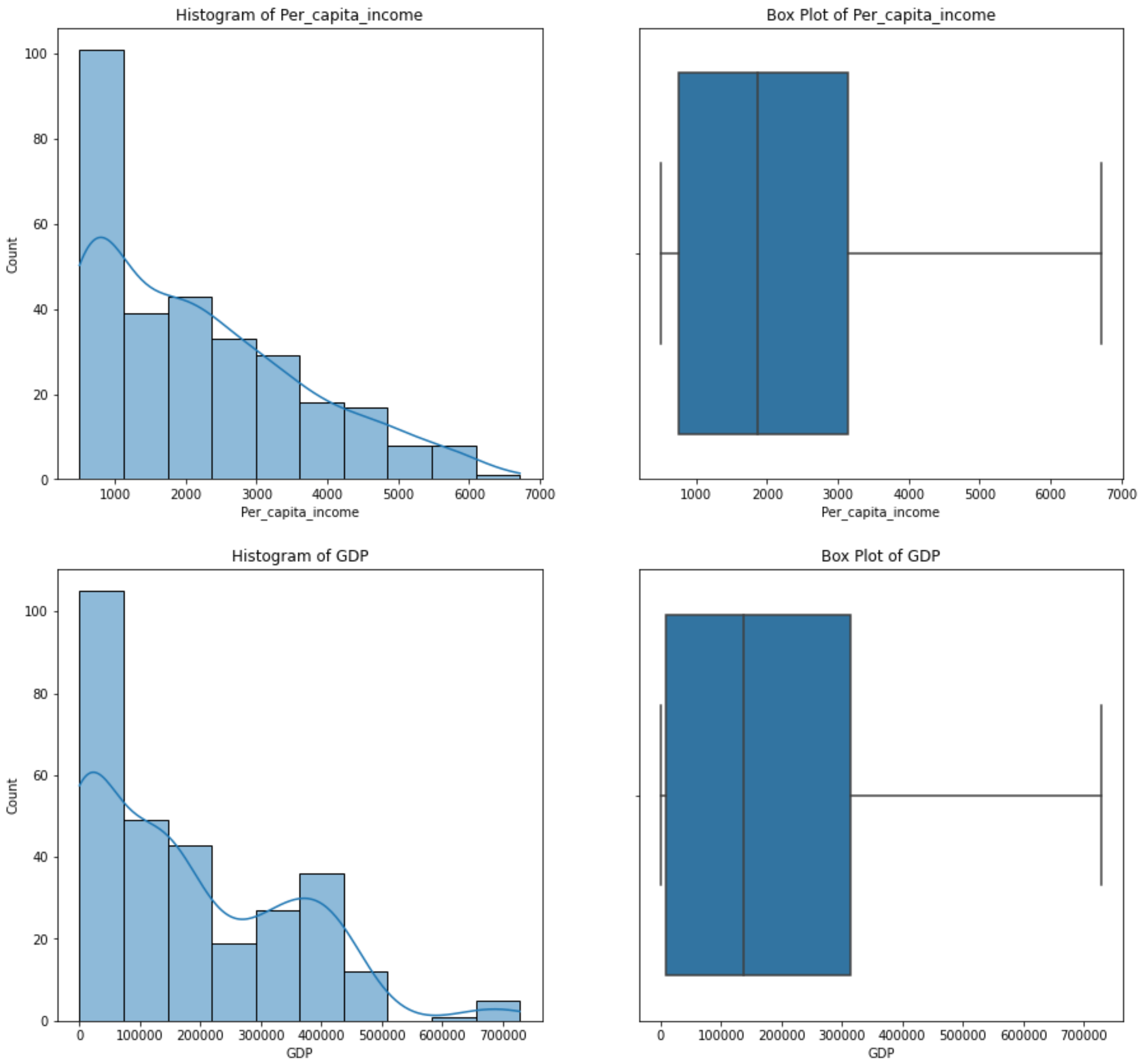
Number of Outliers and Percentage of Outliers in States Dataset.

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.

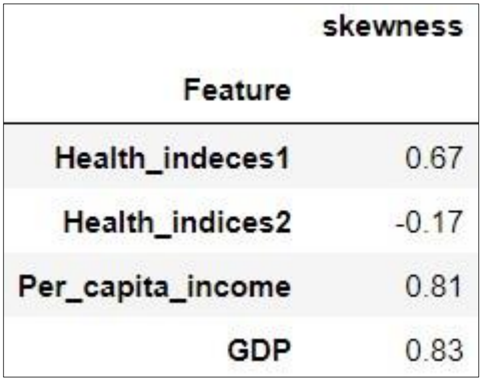
Univariate Analysis – Distribution Plots





Histograms and Box Plots for Numerical Features in States Dataset. Skewness:

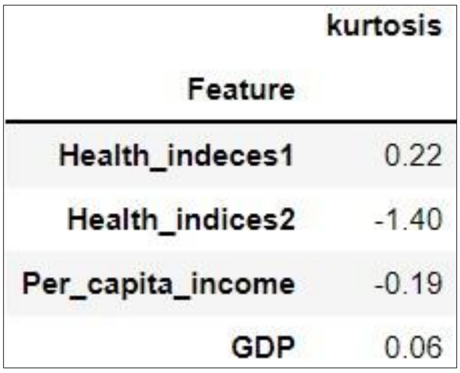
It is a measure of lack of symmetry in a distribution.



	skewness
Feature	
Health_indeces1	0.67
Health_indices2	-0.17
Per_capita_income	0.81
GDP	0.83

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



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

Kurtosis of Numeric Features in States Dataset.

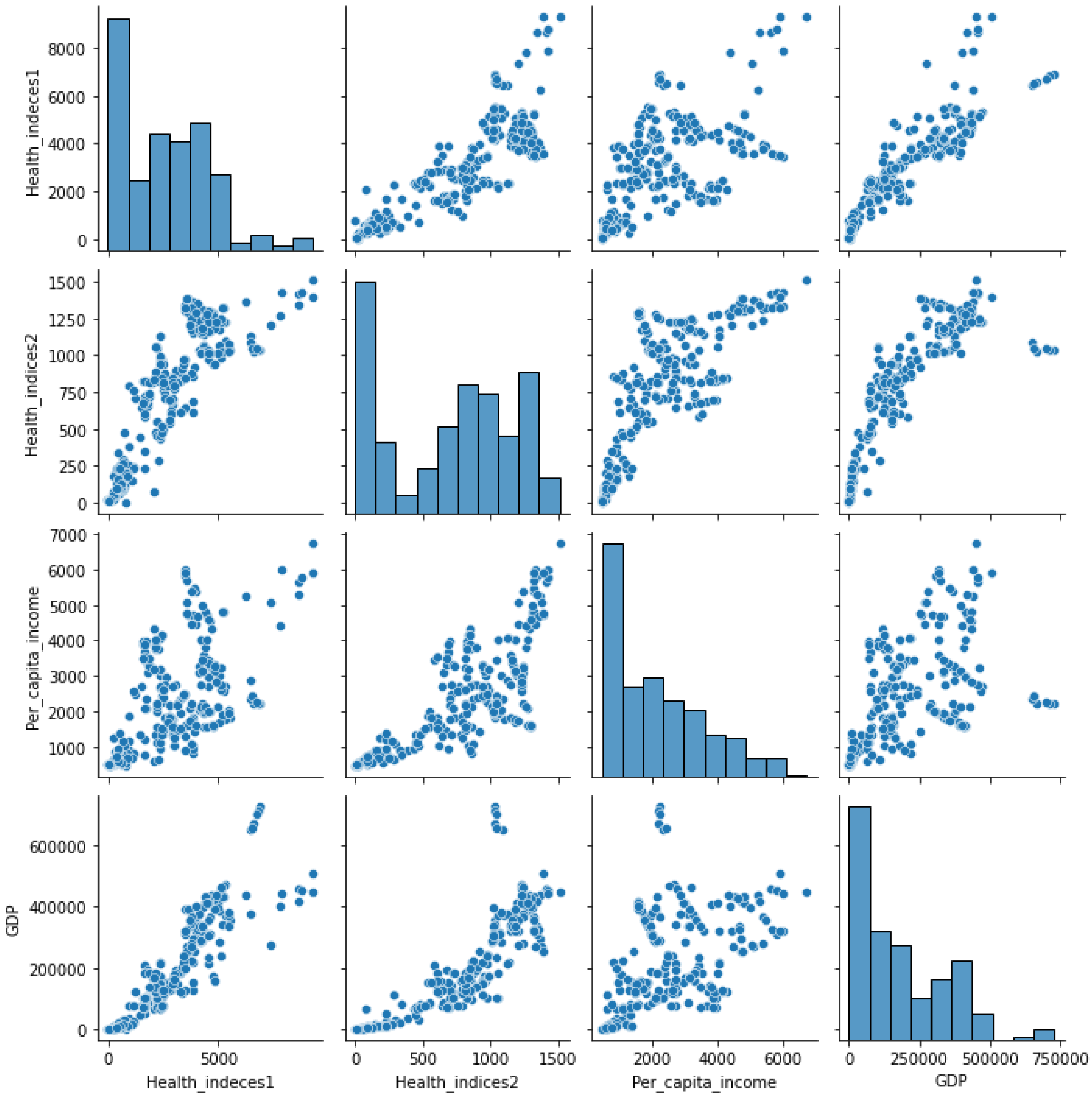
Insights:

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.

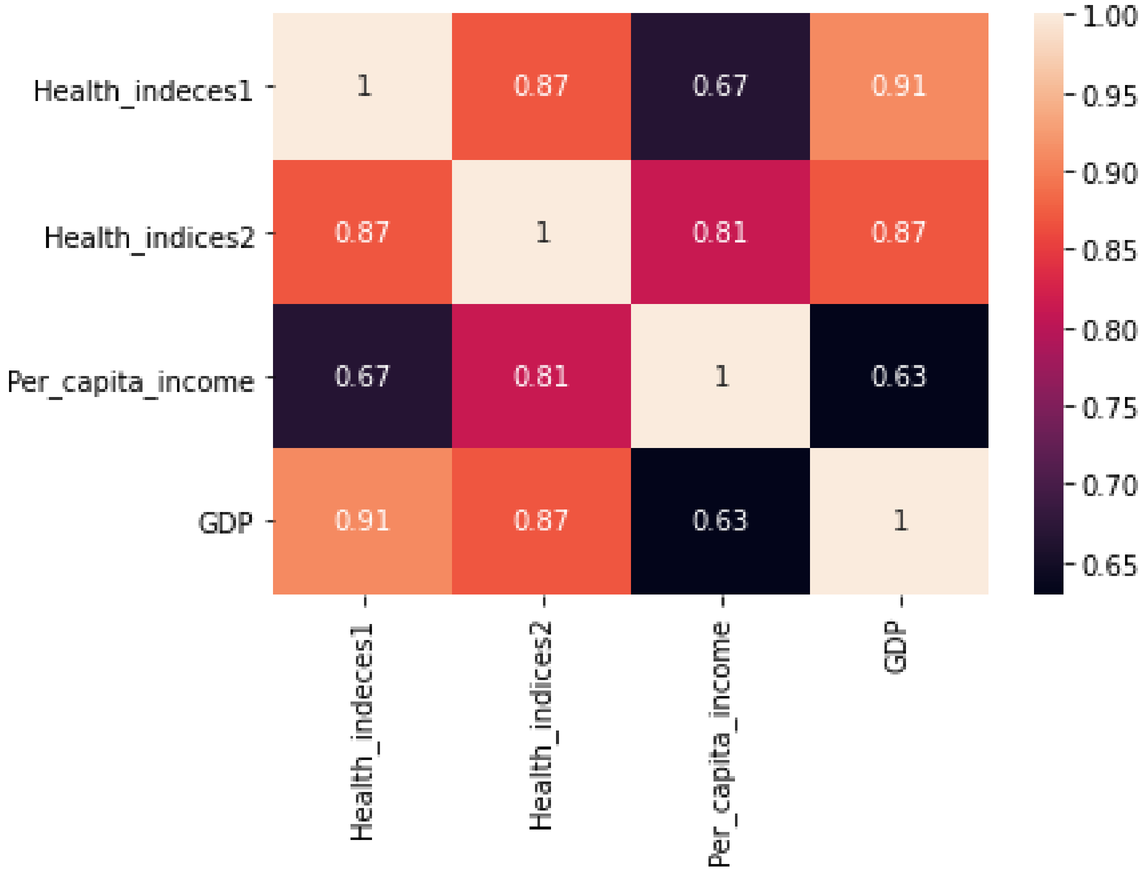
Bivariate Analysis – Between Numeric Continuous Variables

Pair Plot:



Pair Plot for Numeric Continuous Features in States Dataset.

Heat Map:



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

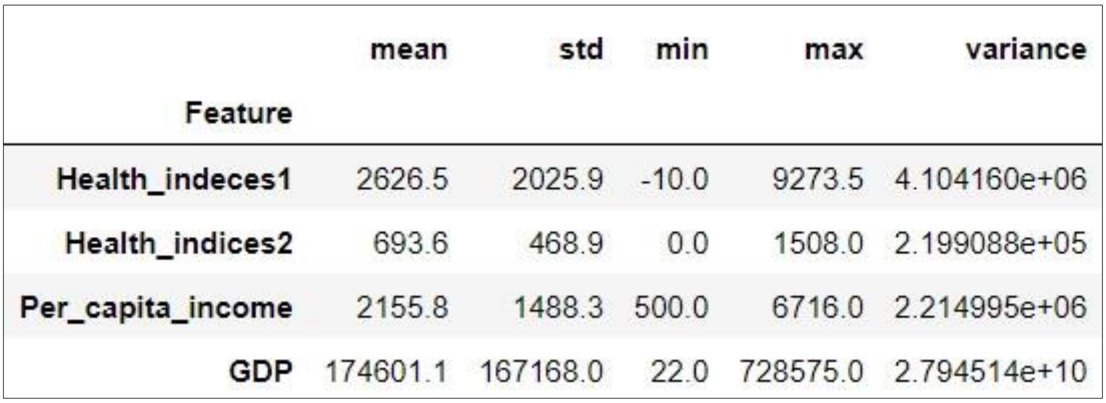
- Generally, Scaling improves the performance of all distance-based models because if we don't scale the data, it gives higher weightage to features which have higher magnitude. Hence, it is always advisable to bring all the features to the same scale before proceeding to distance-based algorithms like 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



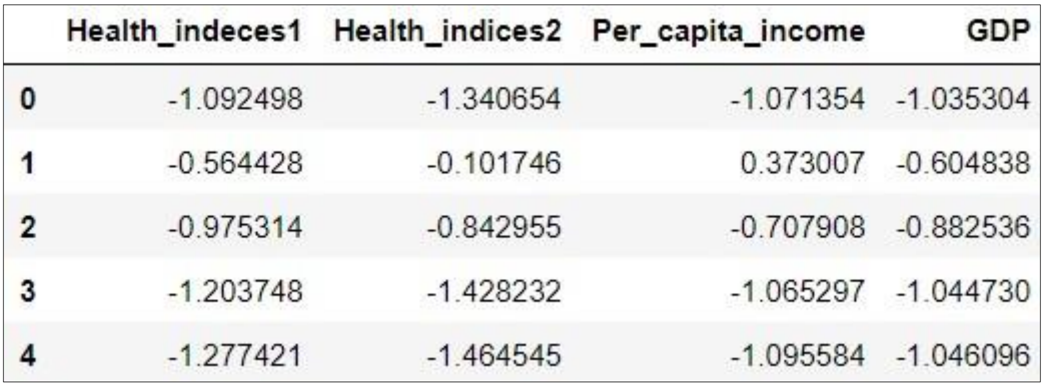
	mean	std	min	max	variance
Feature					
Health_indeces1	2626.5	2025.9	-10.0	9273.5	4.104160e+06
Health_indices2	693.6	468.9	0.0	1508.0	2.199088e+05
Per_capita_income	2155.8	1488.3	500.0	6716.0	2.214995e+06
GDP	174601.1	167168.0	22.0	728575.0	2.794514e+10

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

Sample of the Scaled States Dataset.

Hierarchical Clustering:

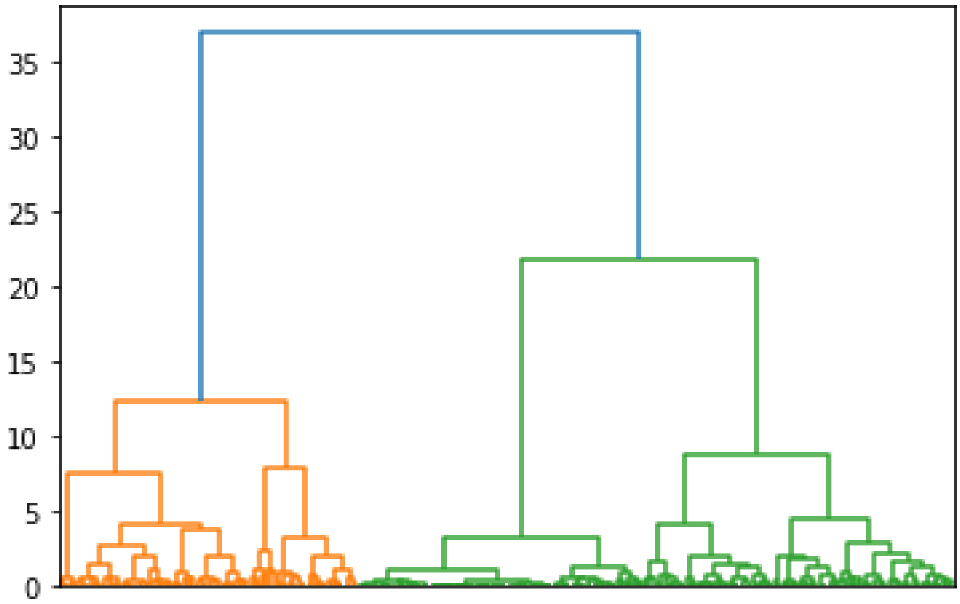
- This method is based on hierarchy representation of clusters where parent cluster is connected to further to child clusters.
- A cluster represents collection of similar data points.
- The agglomerative clustering is the most popular and common hierarchical clustering.
- This method starts by considering 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:

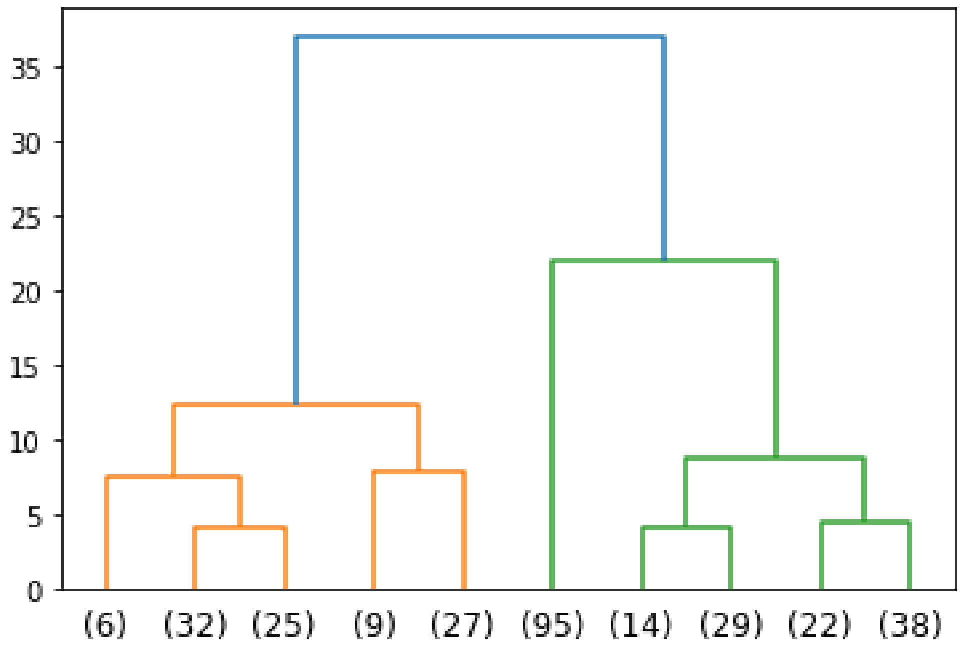
- A dendrogram is a pictorial representation to visualize hierarchical clustering.
- It is mainly used to show the outcome of hierarchical clustering in the form of a tree like diagram that records the sequences of merges and splits.

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.



Dendrogram without Truncated.



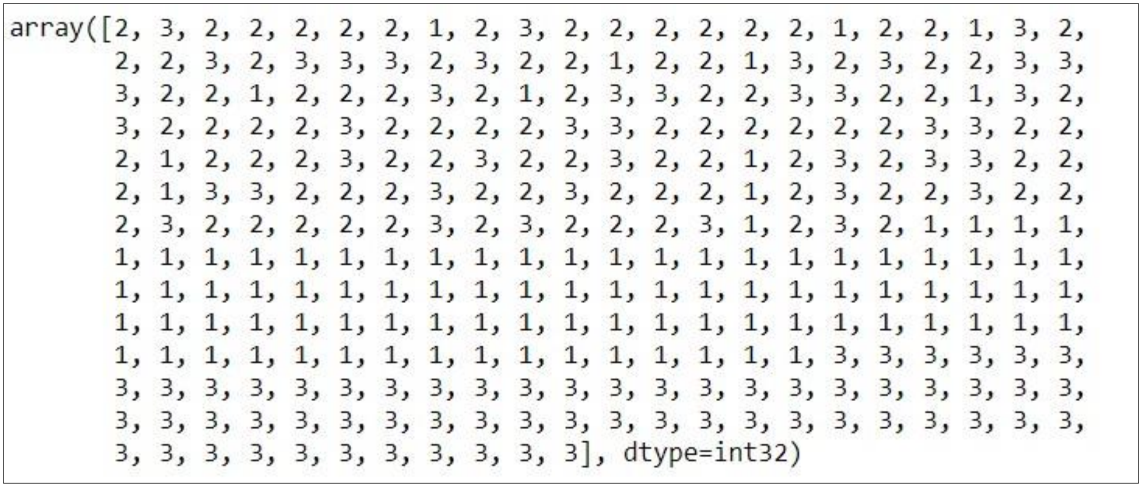
Truncated Dendrogram Representing Last 10 Clusters Formed.

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.**

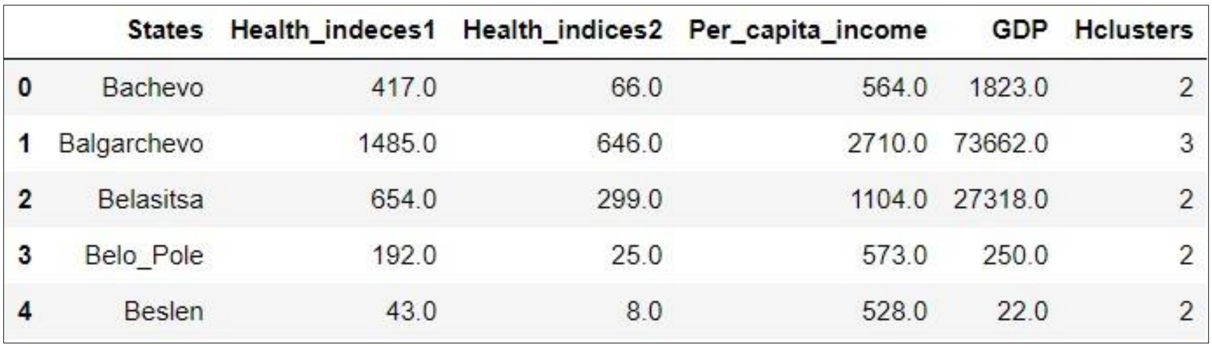
Hierarchical Clustering Labels:

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, 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], dtype=int32)
```

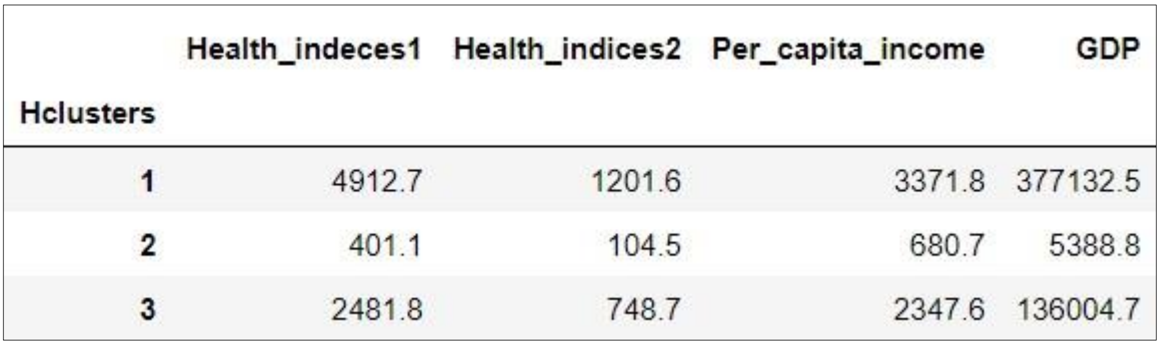
Sample of Clustered Dataset:



	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

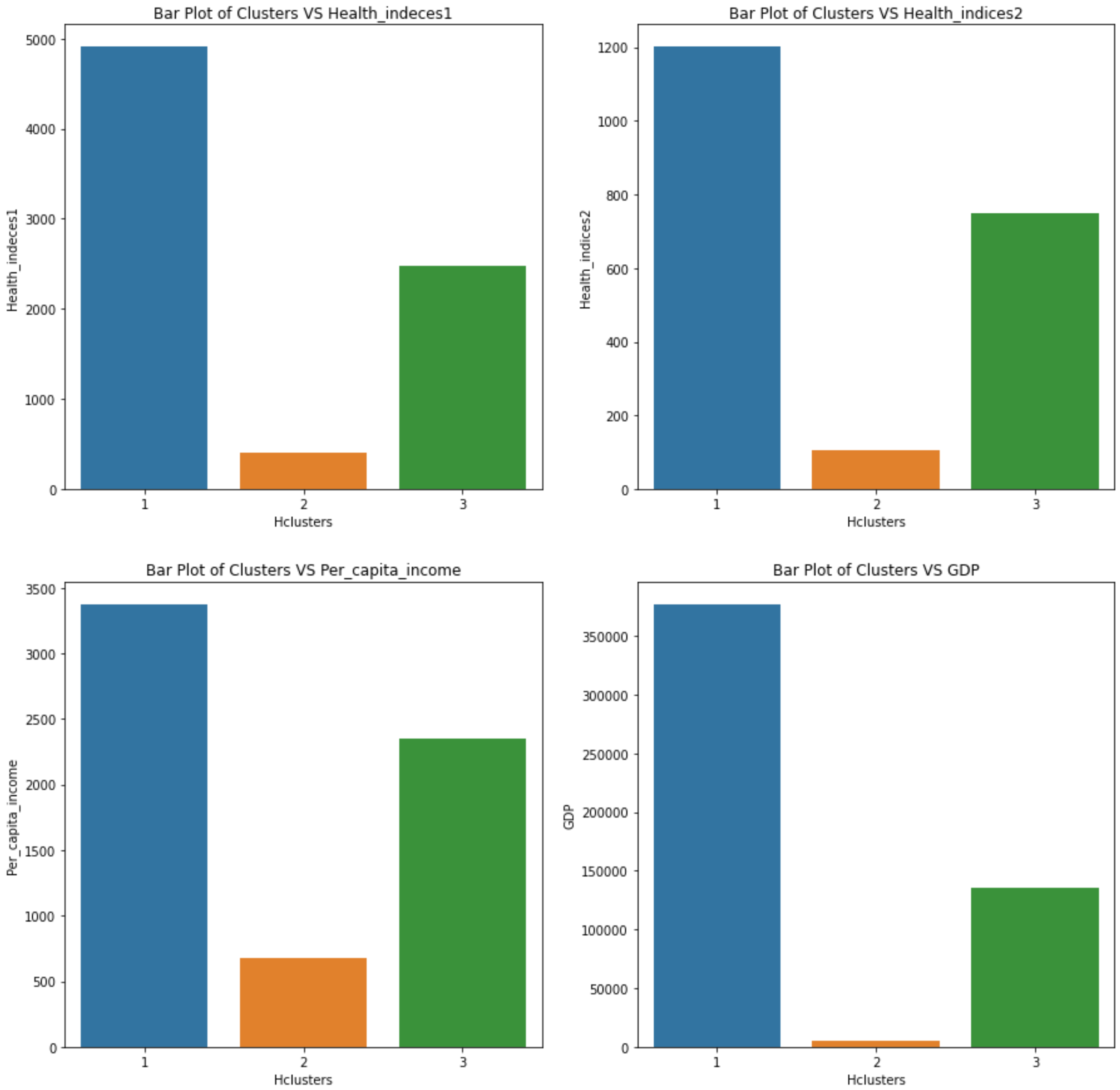
Sample of the Hierarchical Clustered Dataset.

Customer Segmentation:



	Health_indeces1	Health_indices2	Per_capita_income	GDP
Hclusters				
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

Centroids of the Hierarchical Clusters.

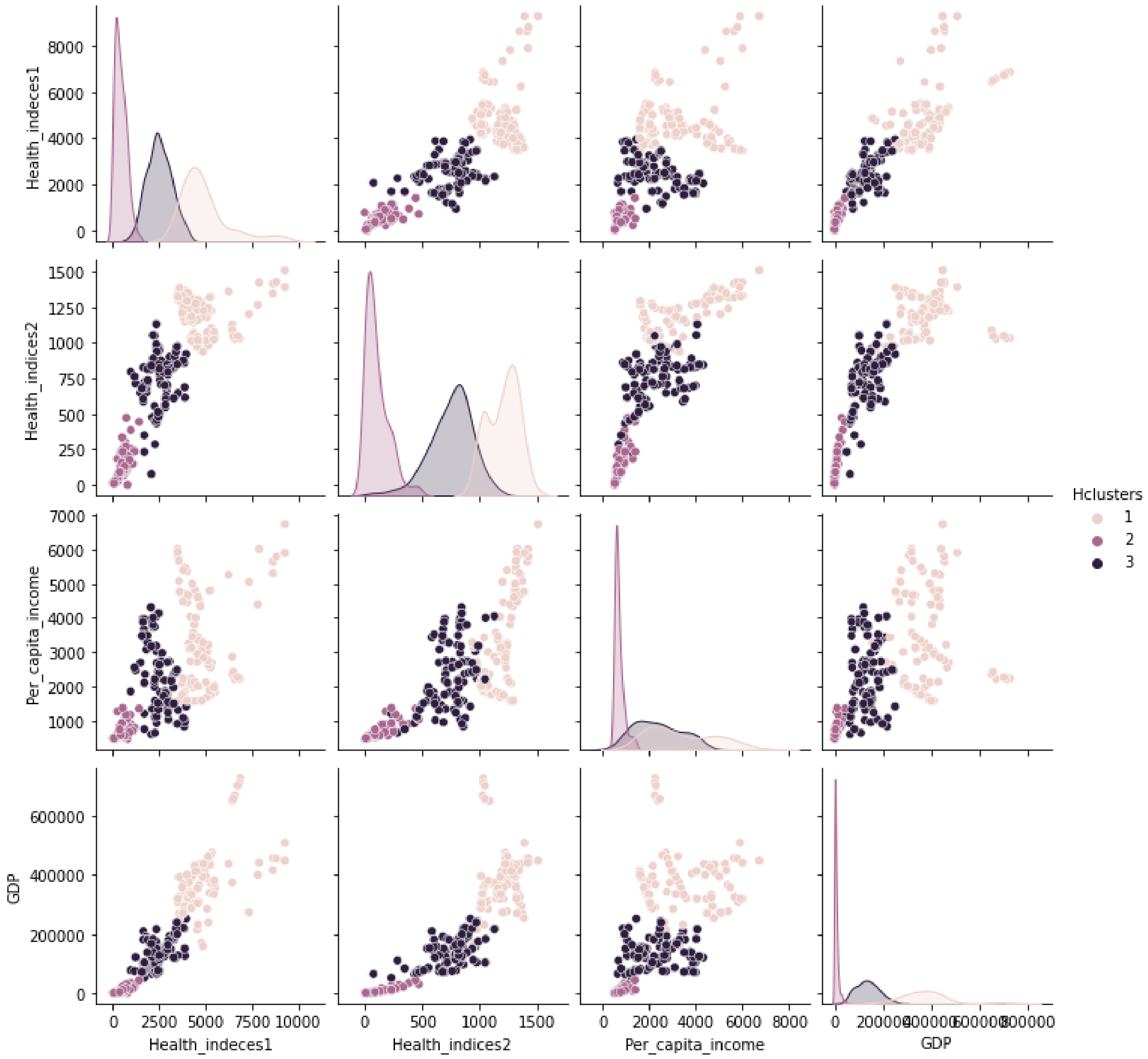


Means of Different Features vs Hierarchical Clusters.

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:



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.

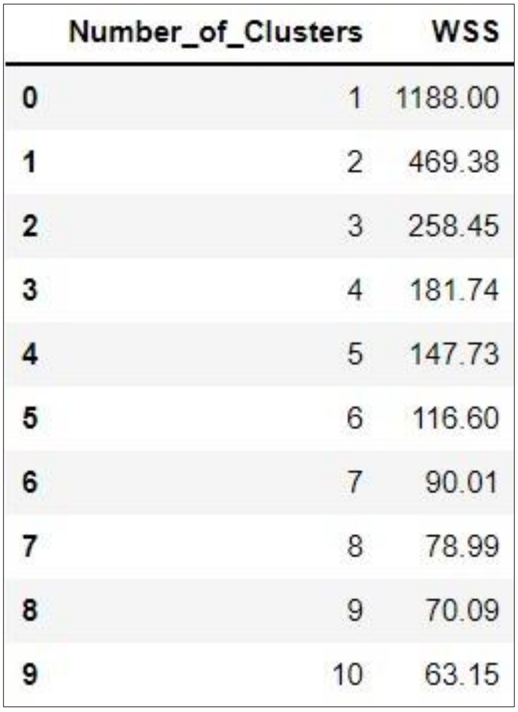
Q1.4. Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and find the silhouette score.

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

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.

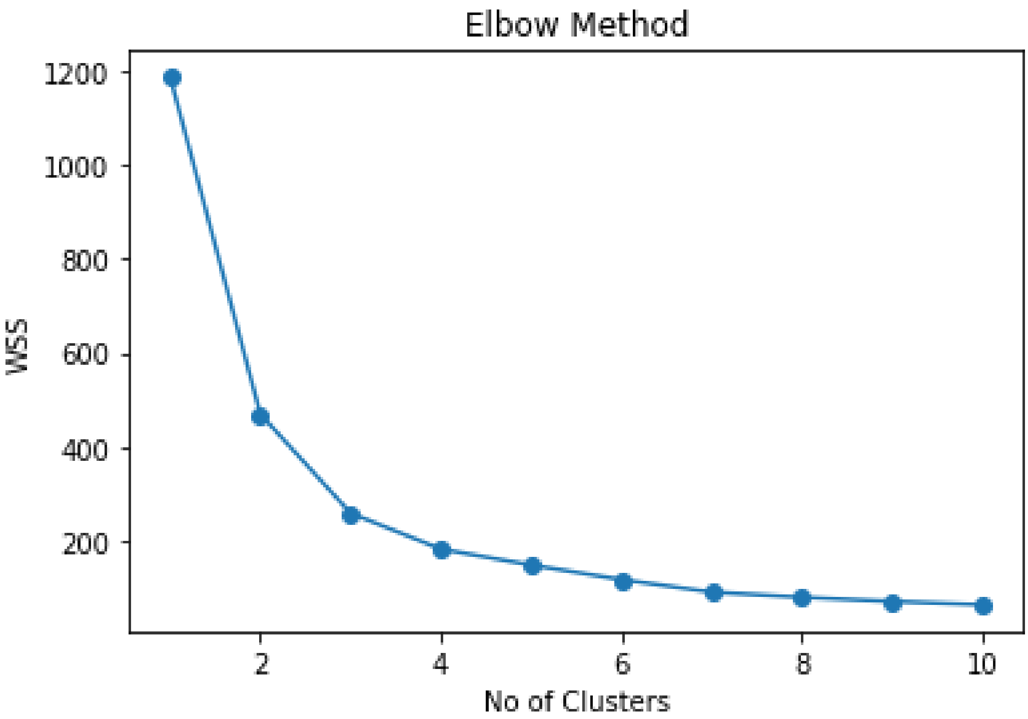
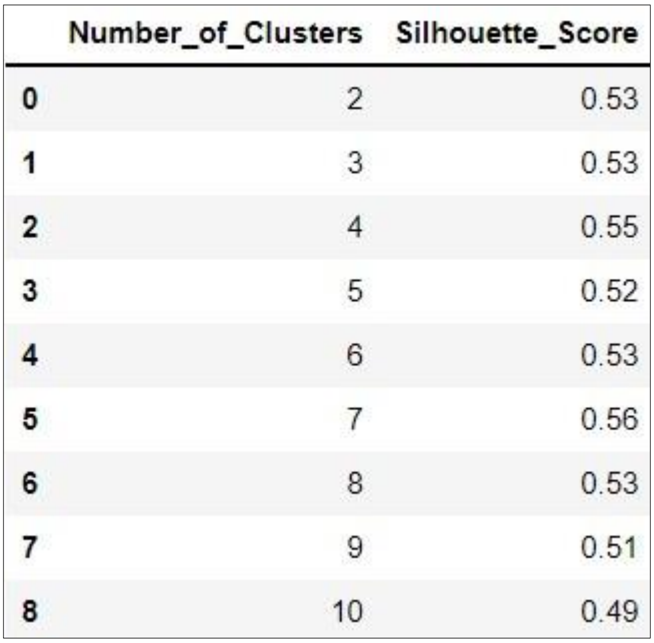


Figure 9. No. of Clusters vs WSS Plot (Elbow Plot).

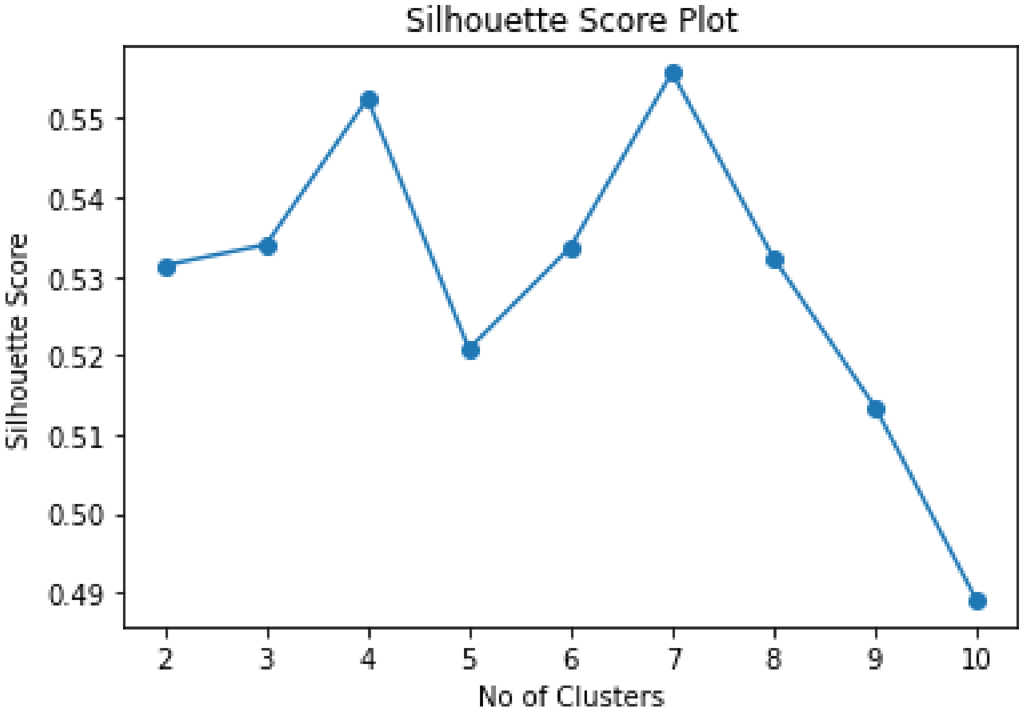
Optimum No. of Clusters by Silhouette Score Method:



	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

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.

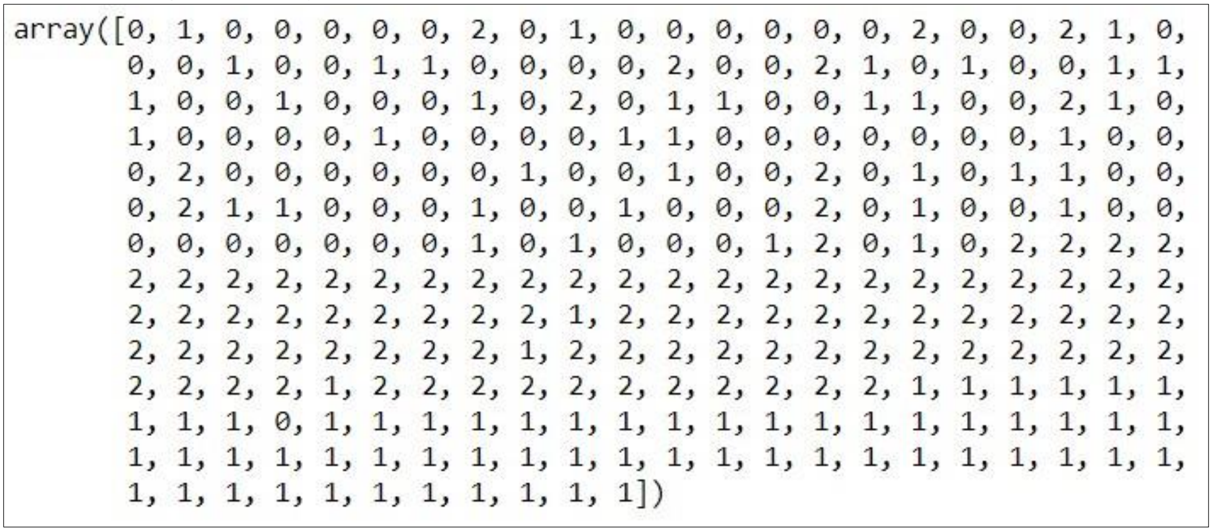


No. of Clusters vs Silhouette Scores Plot.

- 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])
```

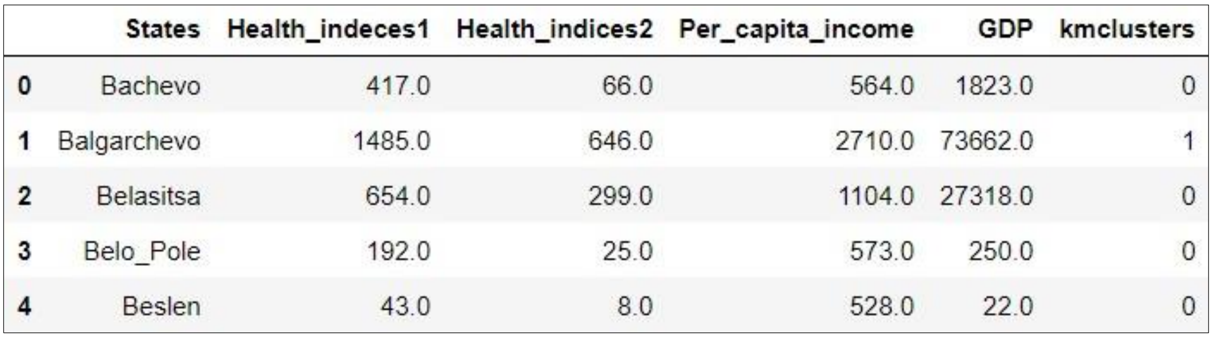
Sill Width of Samples:

- 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.

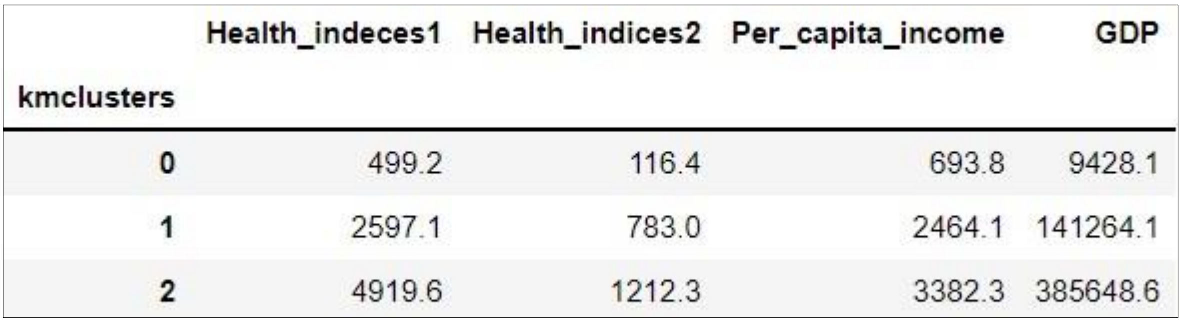
Sample of Clustered Dataset:



	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

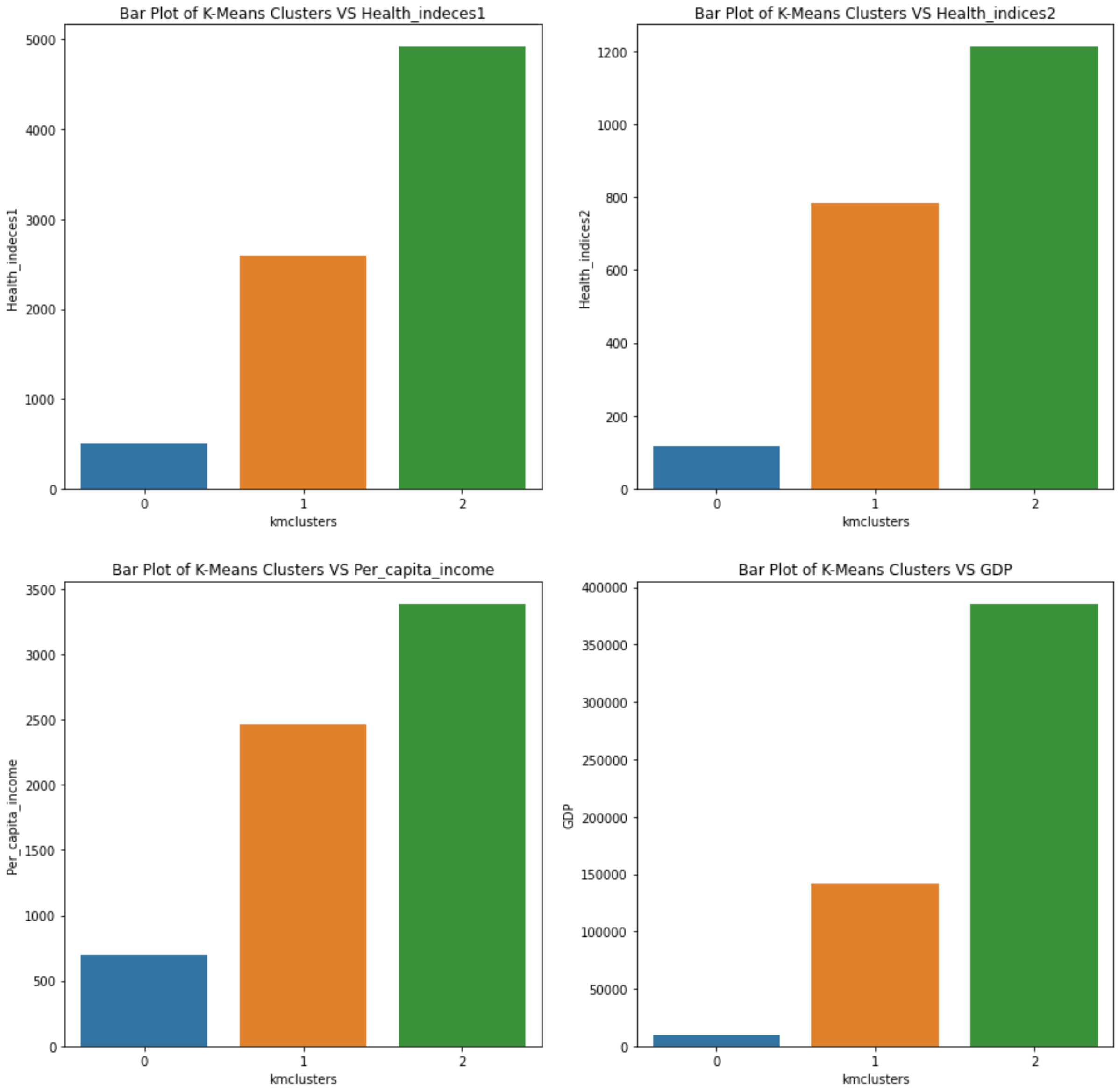
Sample of the K-Means Clustered Dataset.

Customer Segmentation:



	Health_indeces1	Health_indices2	Per_capita_income	GDP
kmclusters				
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

. Centroids of the K-Means Clusters.



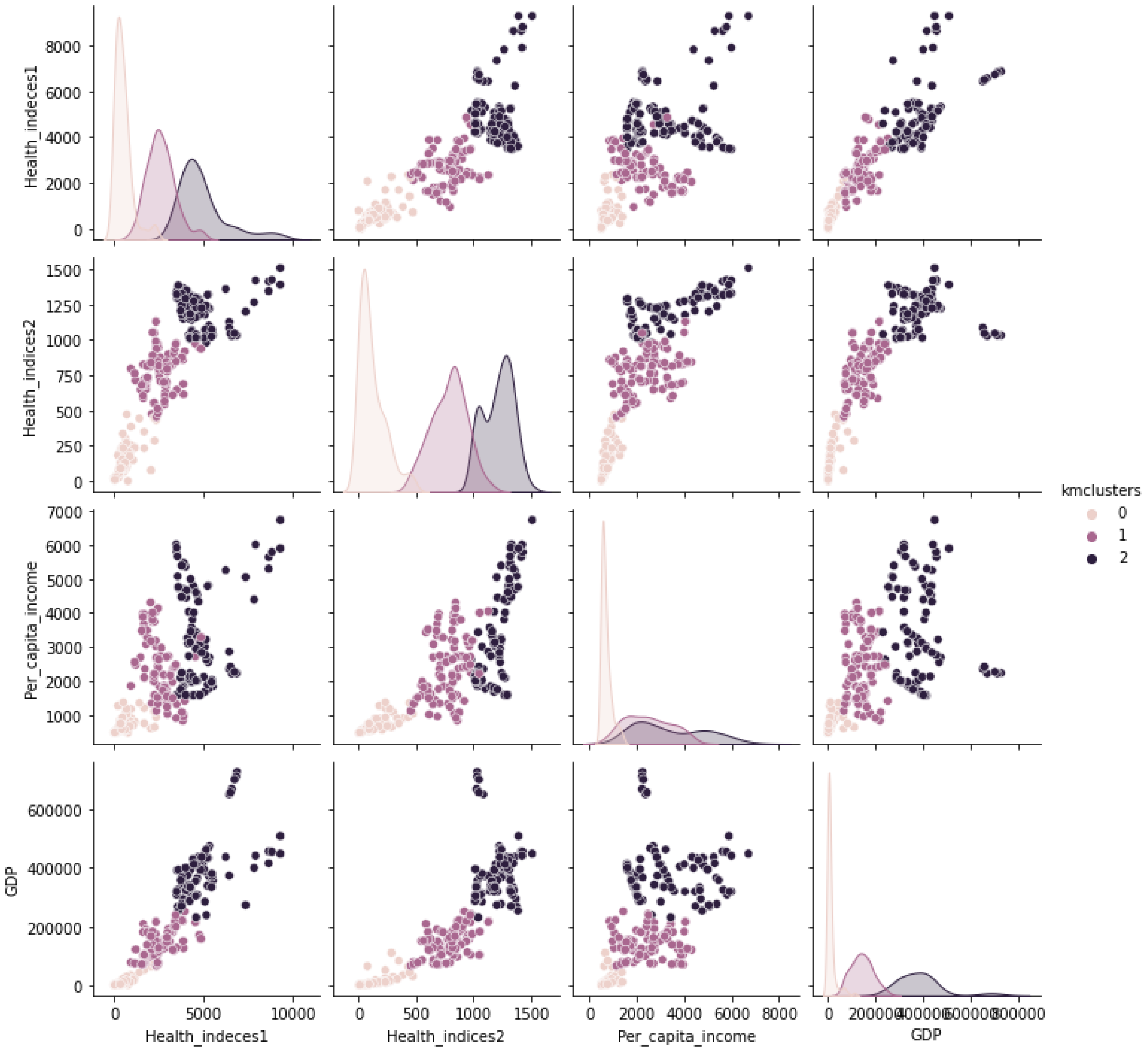
Means of Different Features vs K-Means Clusters.

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.

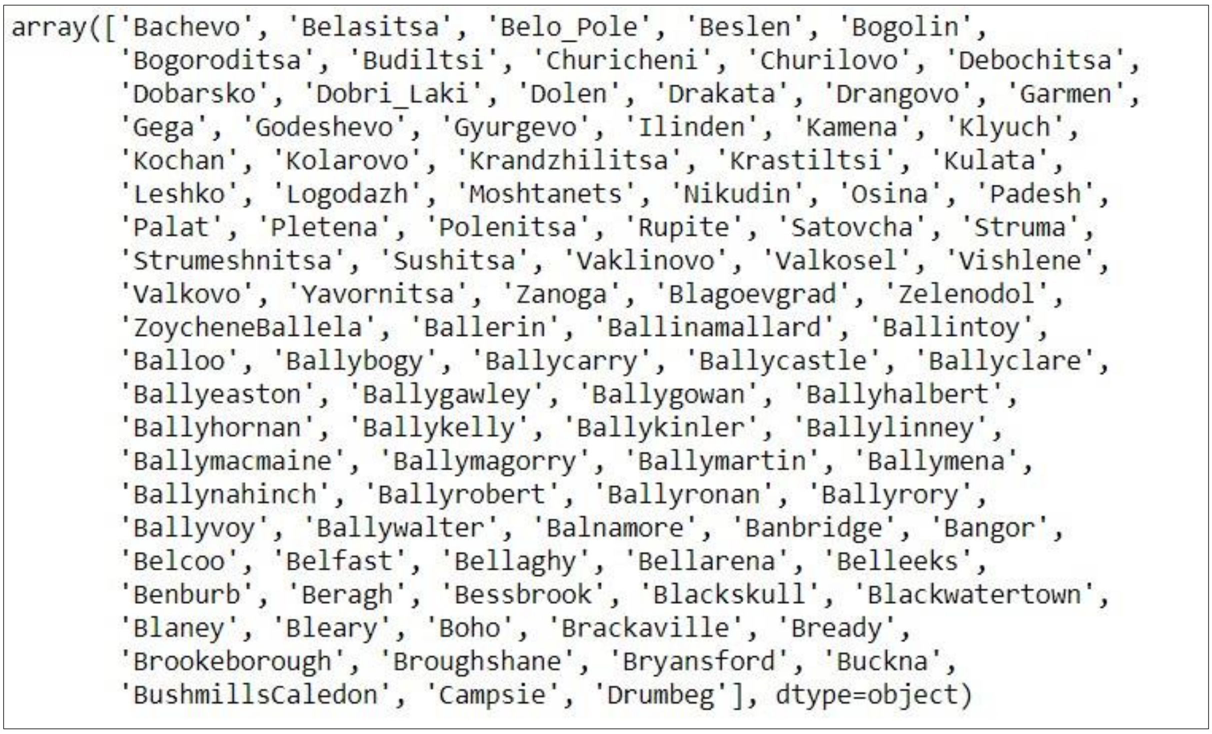


Pair Plot of Numeric Features with K-Means Clusters.

Distribution States among Kmeans clusters:

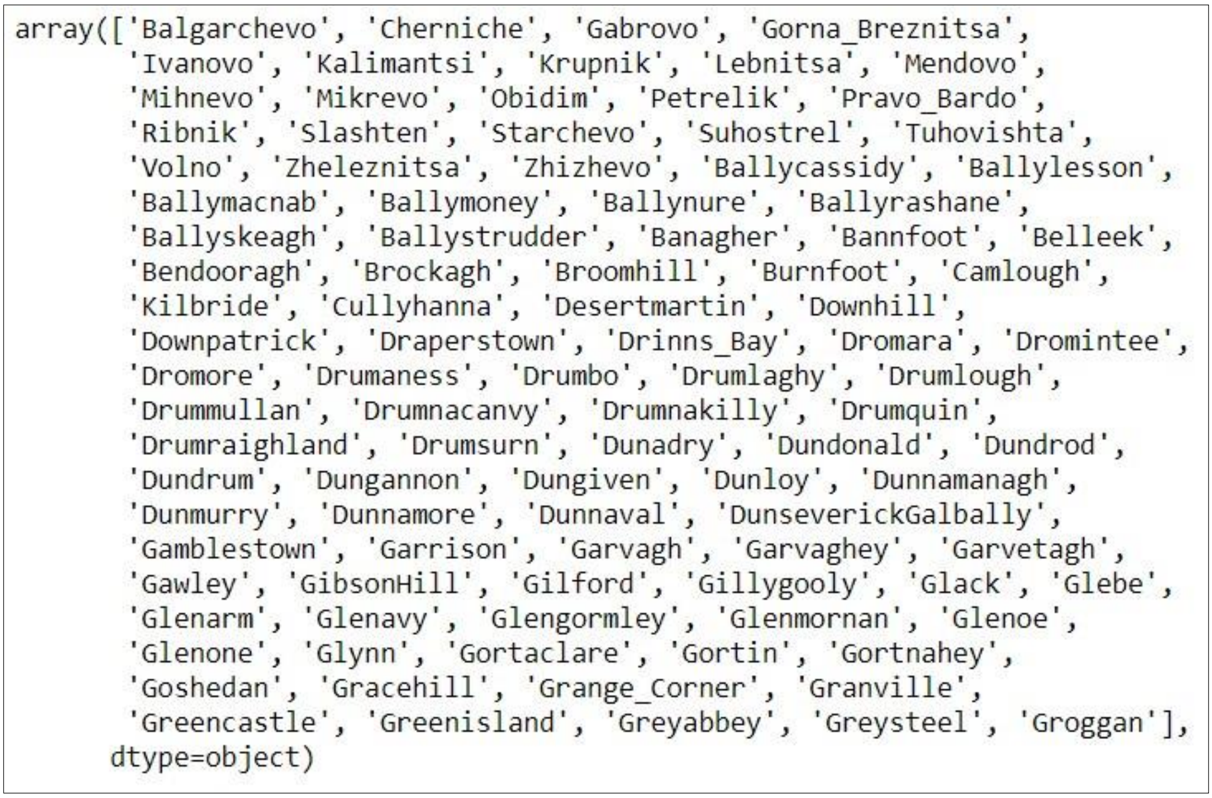
Kmeans Cluster	Number of States
0	101
1	101
2	95

States in Kmeans cluster 0:



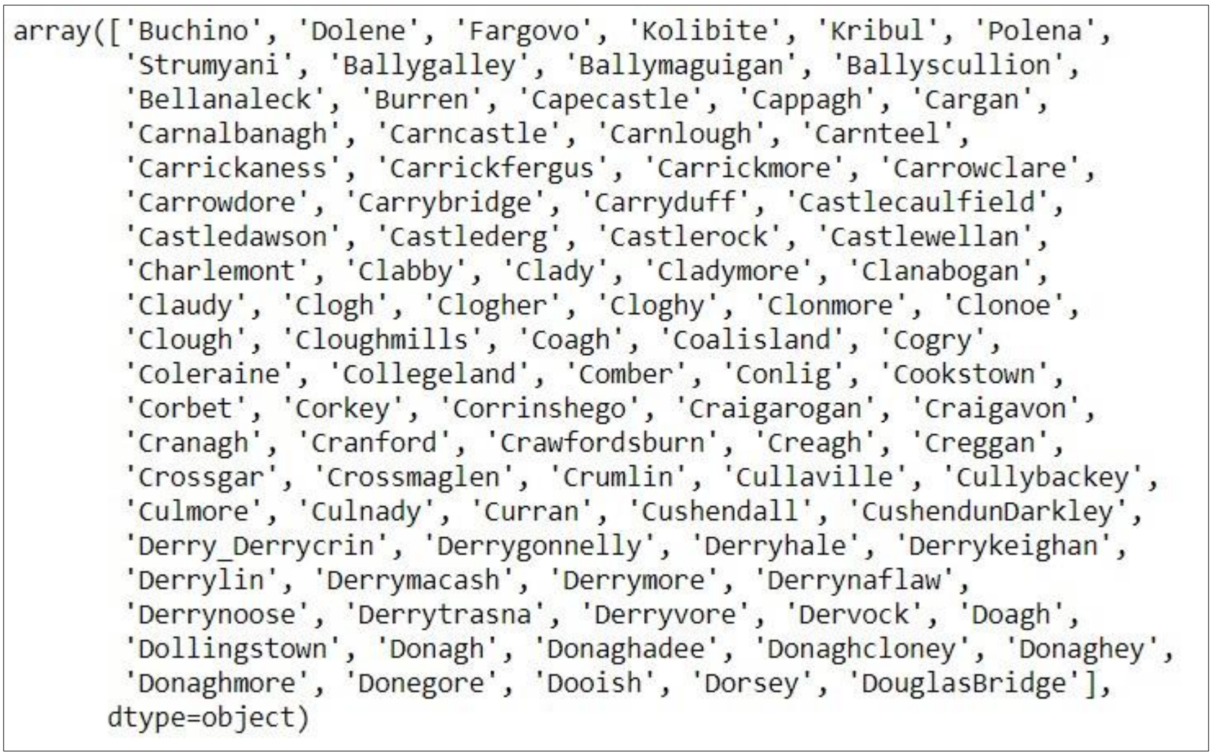
```
array(['Bachevo', 'Belasitsa', 'Belo_Pole', 'Beslen', 'Bogolin',
       'Bogoroditsa', 'Budiltsi', 'Churicheni', 'Churilovo', 'Debochitsa',
       'Dobarsko', 'Dobri_Laki', 'Dolen', 'Drakata', 'Drangovo', 'Garmen',
       'Gega', 'Godeshevo', 'Gyurgevo', 'Ilinden', 'Kamena', 'Klyuch',
       'Kochan', '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:



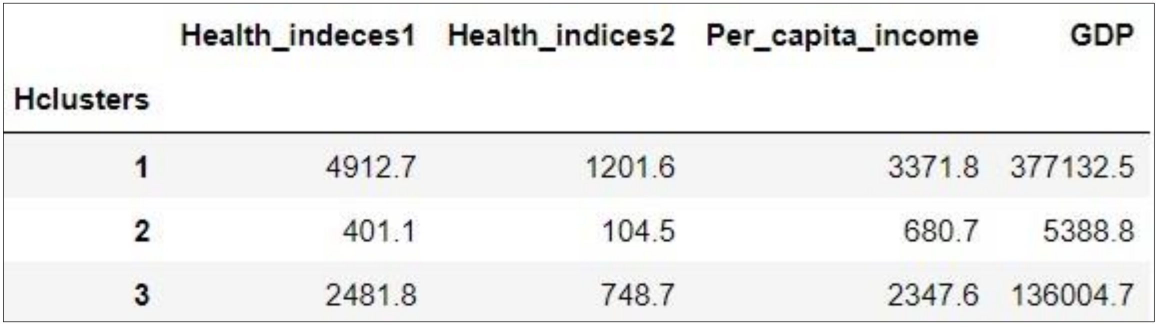
```
array(['Buchino', 'Dolene', 'Fargovo', 'Kolibite', 'Kribul', 'Polena',
       'Strumyani', 'Ballygalley', 'Ballymaguigan', 'Ballyscullion',
       'Bellanaleck', 'Burren', 'Capecastle', 'Cappagh', 'Cargan',
       'Carnalbanagh', 'Carncastle', 'Carnlough', 'Carnteel',
       'Carrickaness', 'Carrickfergus', '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.

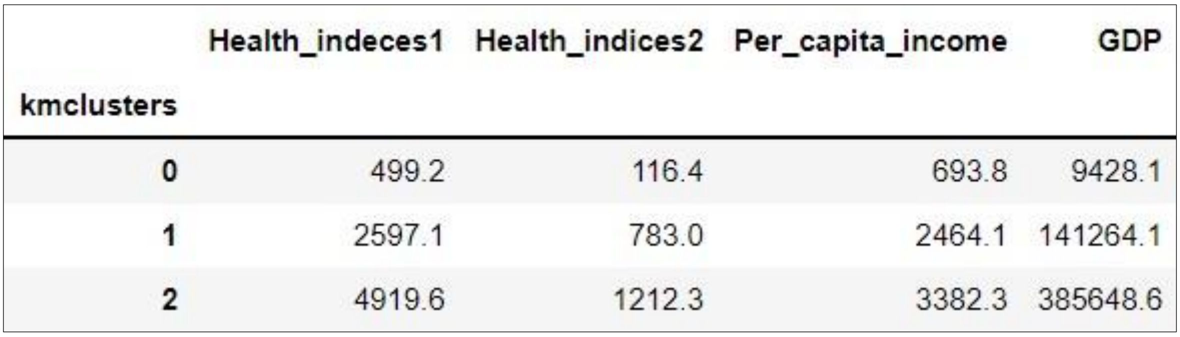
Cluster Profiles:

- 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:



	Health_indeces1	Health_indices2	Per_capita_income	GDP
Hclusters				
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



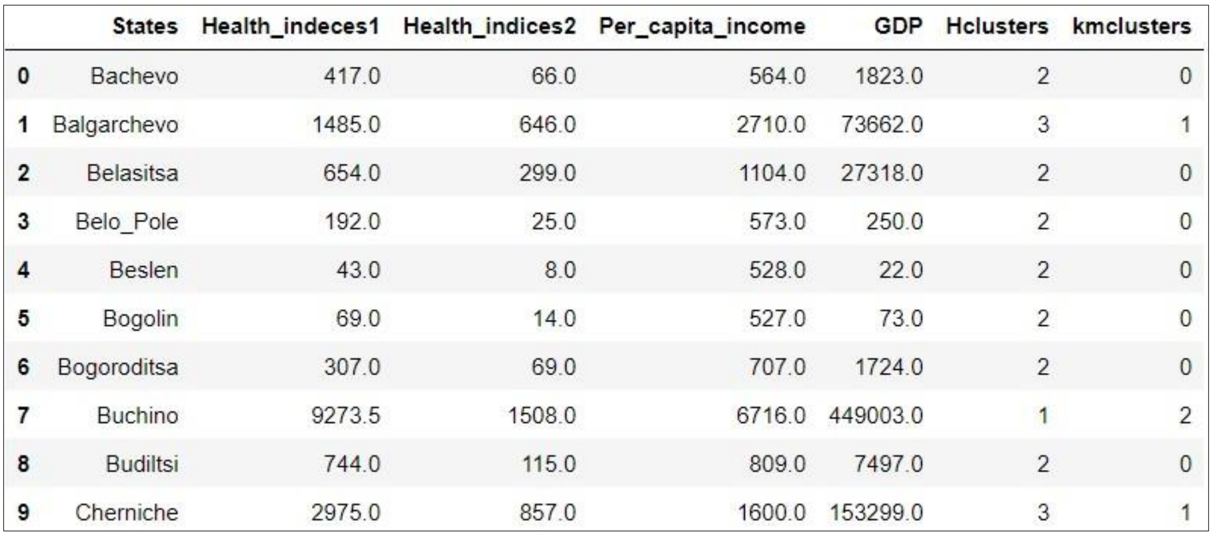
	Health_indeces1	Health_indices2	Per_capita_income	GDP
kmclusters				
0	499.2	116.4	693.8	9428.1
1	2597.1	783.0	2464.1	141264.1
2	4919.6	1212.3	3382.3	385648.6

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 & KMeans) 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

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.