DM Extended Project – Guided

Name- Soumalya Sen

Batch- PGP-DSBA(PGPDSBA.O.JULY23.A)

Index

1. Problem Statement

- Context

The State_wise_Health_income.csv 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.

- Objective

The objective of the analysis is to cluster the states based on the given dataset "State_wise_Health_income.csv," considering both health and economic indicators. By grouping states with similar conditions, the goal is to provide the government with valuable insights for implementing targeted measures to improve the health and economic conditions of each group.

- Data Description

1. States: Names of States

2. Health_indeces1: A composite index rolls several related measures (indicators) into a single score that provides a summary of how the health system is performing in the State.

3. Health_indeces2: A composite index rolls several related measures (indicators) into a single score that provides a summary of how the health system is performing in certain areas of the States.

4. Per_capita_income: Per capita income (PCI) measures the average income earned per person in a given area (city, region, country, etc.) in a specified year. It is calculated by dividing the area's total income by its total population.

5. GDP: GDP provides an economic snapshot of a country/state, used to estimate the size of an economy and growth.

2. Read the data, Perform Exploratory Data Analysis (Check the null values, Data types, shape, EDA, etc)

Let's get started. Load the required packages, set the working directory and load the data file.

- Overview of the Dataset

The initial steps to get an overview of any dataset is to:

- observe the first few rows of the dataset, to check whether the dataset has been loaded properly or not.

- get information about the number of rows and columns in the dataset.

- find out the data types of the columns to ensure that data is stored in the preferred format and the value of each property is as expected.

- check the statistical summary of the dataset to get an overview of the numerical columns of the data

- Checking the shape of the dataset

* The dataset has 297 rows and 6 columns

- Displaying few rows of the dataset.

	Unnamed: 0	States	Health_indices1	Health_indices2	Per_capita_income	GDP
0	0	Bachevo	417	66	564	1823
1	1	Balgarchevo	1485	646	2710	73662
2	2	Belasitsa	654	299	1104	27318
3	3	Belo_Pole	192	25	573	250
4	4	Beslen	43	8	528	22

Creating a Copy of the Original Data to avoid any changes to original data

Checking the data types of the columns for the dataset

```
Unnamed: 0             int64
States                object
Health_indices1        int64
Health_indices2        int64
Per_capita_income      int64
GDP                    int64
dtype: object
```

- All the columns in the data are integer type colum. 'States' column in the data is an object or string type column.

- Checking the missing values

```
Unnamed: 0            0
States                0
Health_indices1       0
Health_indices2       0
Per_capita_income     0
GDP                   0
dtype: int64
```

- There are no missing values in the data.

- Checking the unique values

```
Unnamed: 0            297
States                296
Health_indices1       278
Health_indices2       249
Per_capita_income     279
GDP                   286
dtype: int64
```

Dropping variables

We will drop the *"Unnamed: 0* as it do not add any value to the analysis.

- Statistical summary of the dataset

	count	mean	std	min	25%	50%	75%	max
Health_indices1	297.0	2630.151515	2038.505431	-10.0	641.0	2451.0	4094.0	10219.0
Health_indices2	297.0	693.632997	468.944354	0.0	175.0	810.0	1073.0	1508.0
Per_capita_income	297.0	2156.915825	1491.854058	500.0	751.0	1865.0	3137.0	7049.0
GDP	297.0	174601.117845	167167.992863	22.0	8721.0	137173.0	313092.0	728575.0

Observations

summary of the dataset:

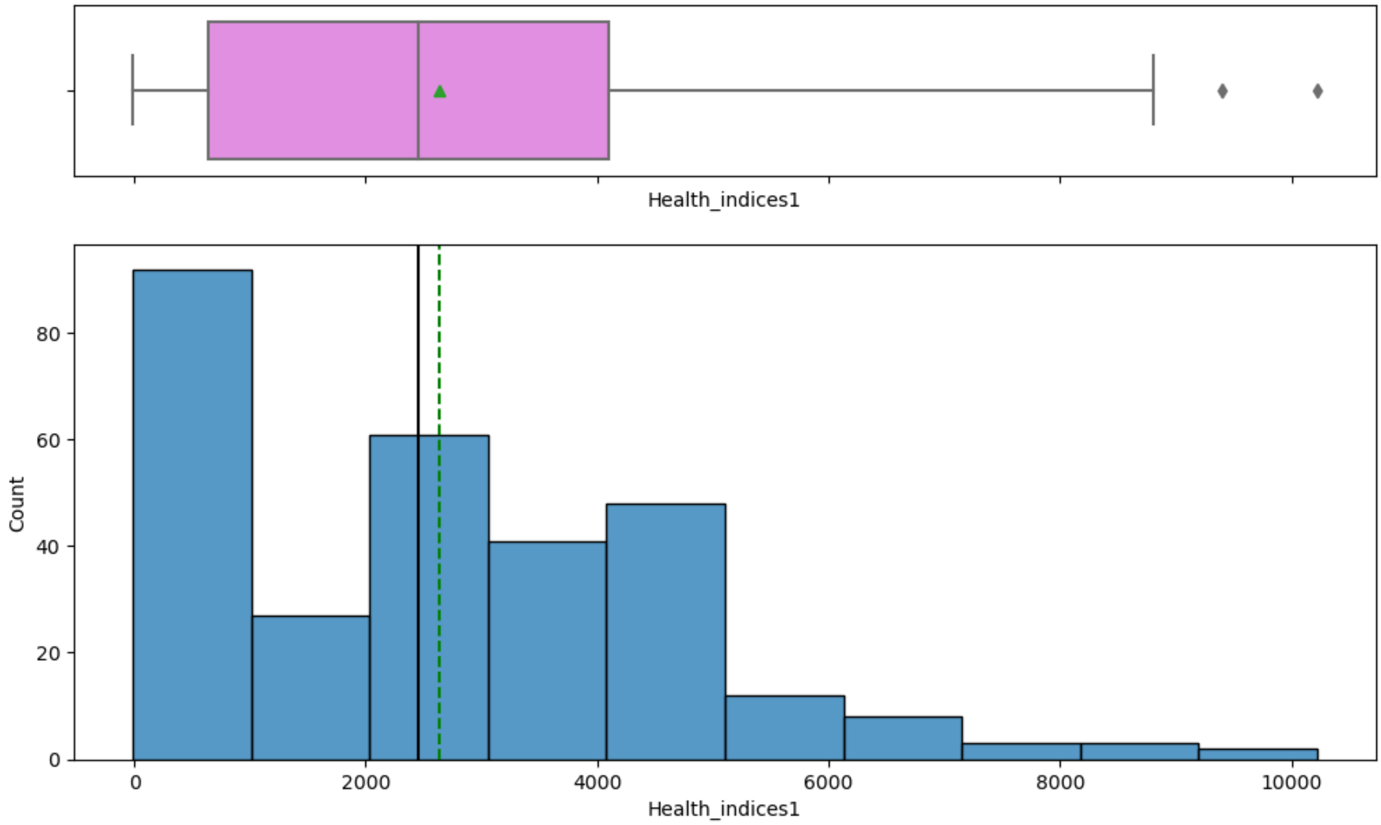
- 'Health_indices1': 297 unique values, mean 2630.15, std 2038.51, min -10, max 10219.
- 'Health_indices2': 297 unique values, mean 693.63, std 468.94, min 0, max 1508.
- 'Per_capita_income': 297 unique values, mean 2156.92, std 1491.85, min 500, max 7049.
- 'GDP': 297 unique values, mean 174601.12, std 167168.00, min 22, max 728575.

This summary provides an overview of the distribution of values in each column, including the number of unique values, mean, standard deviation, minimum, and maximum.

- Exploratory Data Analysis

Univariate analysis

`Health_indices1`



1) Histogram and boxplot of Health_indices1 variable

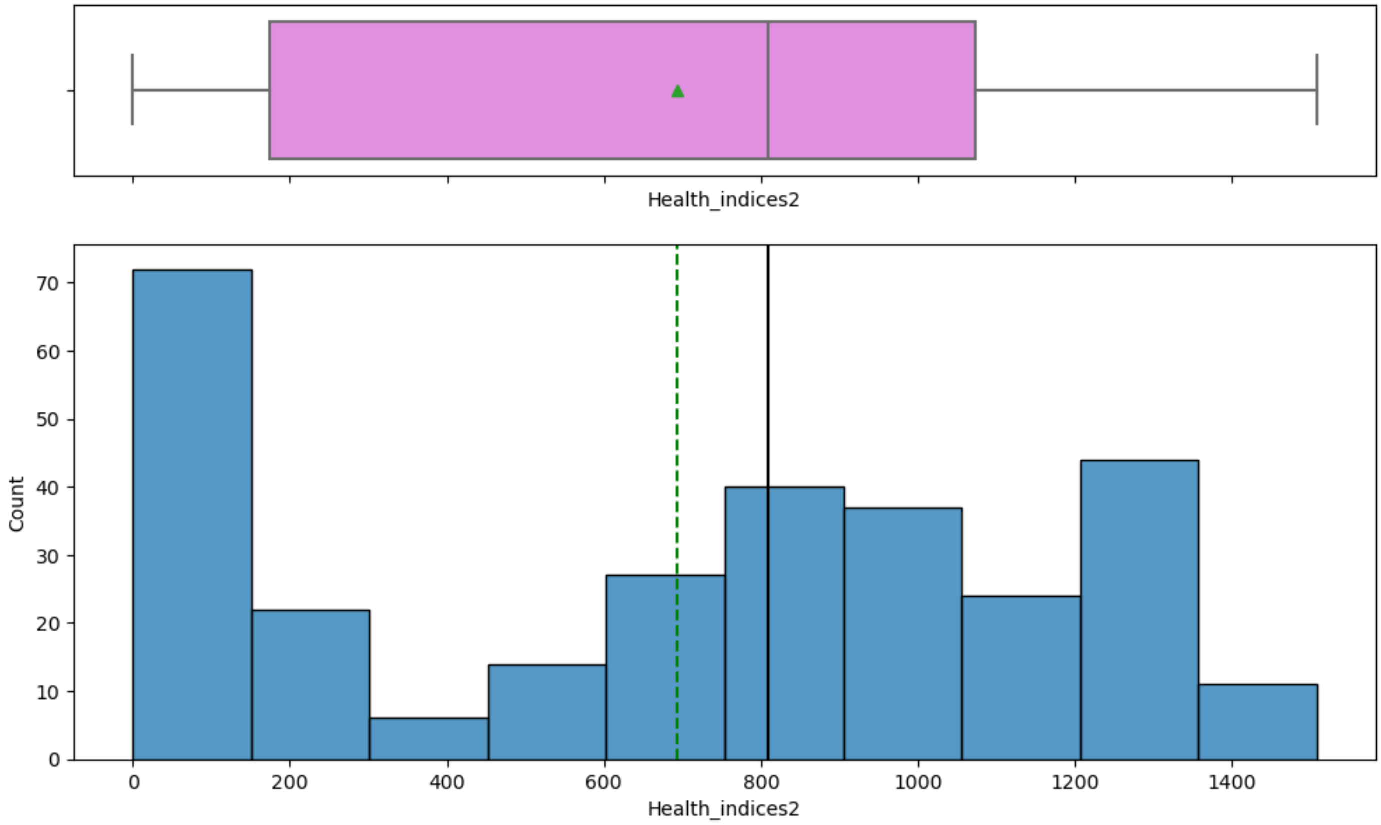
Observations

The histogram shows the distribution of the 'Health_indices1' variable. You can observe the shape of the distribution, such as whether it is symmetric, skewed, or has multiple peaks. You can also observe the range of values, the number of observations, and the frequency of extreme values.

The boxplot shows the summary statistics of the 'Health_indices1' variable. The box represents the interquartile range (IQR), which is the range between the 1st quartile (Q1) and the 3rd quartile (Q3). The line in the middle of the box represents the median (Q2) or the 2nd quartile. The whiskers represent the range of values that are within 1.5 times the IQR from the Q1 and Q3. The dots outside the whiskers represent the outliers or extreme values. You can observe the spread of the data, the skewness, and the presence of outliers.

By combining the information from the histogram and boxplot, you can get a better understanding of the distribution, variability, and summary statistics of the 'Health_indices1' variable.

`Health_indices2`



2)histogram and boxplot of Health_indices2 variable

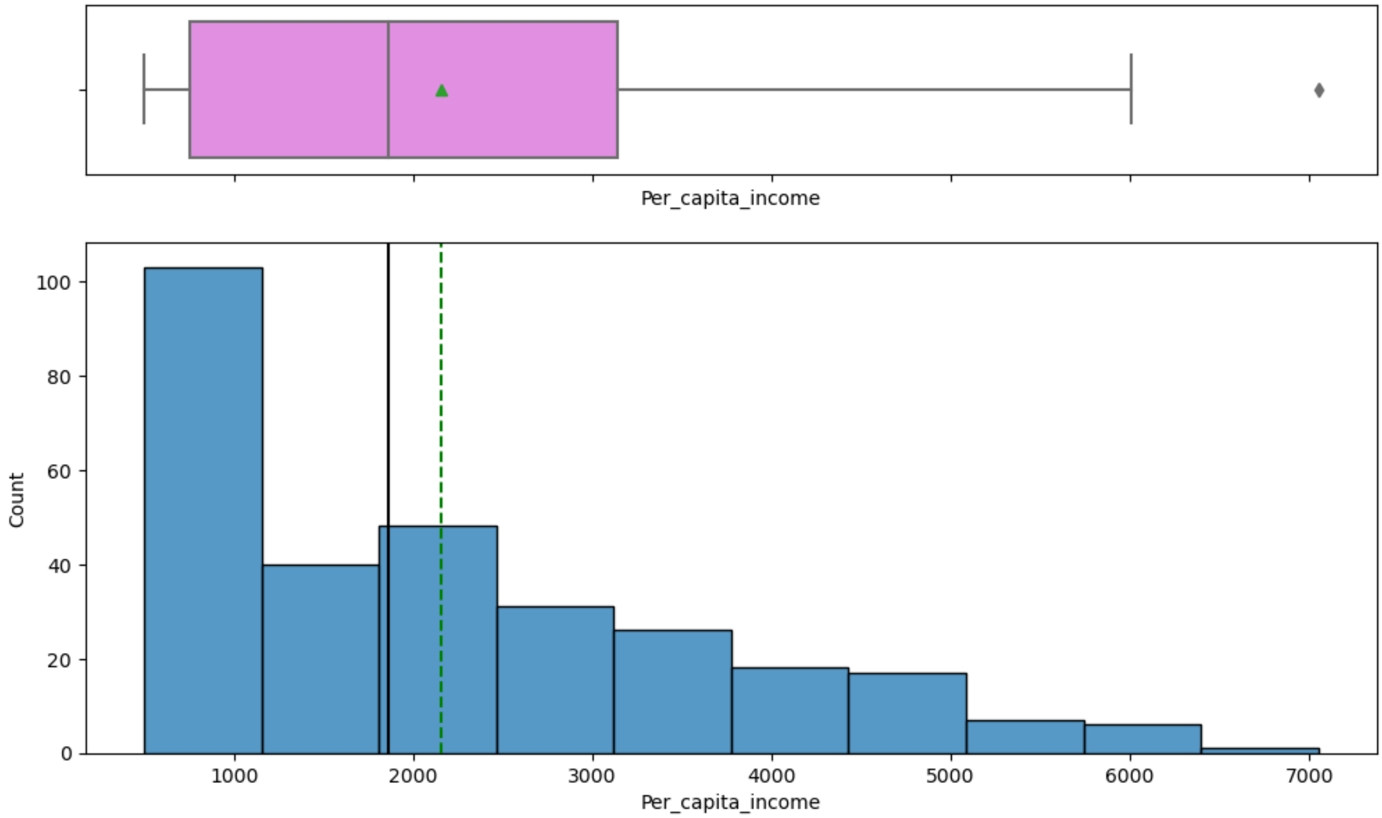
Observations

The histogram shows the distribution of the 'Health_indices2' variable. You can observe the shape of the distribution, such as whether it is symmetric, skewed, or has multiple peaks. You can also observe the range of values, the number of observations, and the frequency of extreme values.

The boxplot shows the summary statistics of the 'Health_indices2' variable. The box represents the interquartile range (IQR), which is the range between the 1st quartile (Q1) and the 3rd quartile (Q3). The line in the middle of the box represents the median (Q2) or the 2nd quartile. The whiskers represent the range of values that are within 1.5 times the IQR from the Q1 and Q3. The dots outside the whiskers represent the outliers or extreme values. You can observe the spread of the data, the skewness, and the presence of outliers.

By combining the information from the histogram and boxplot, you can get a better understanding of the distribution, variability, and summary statistics of the 'Health_indices2' variable.

`Per Capita Income`

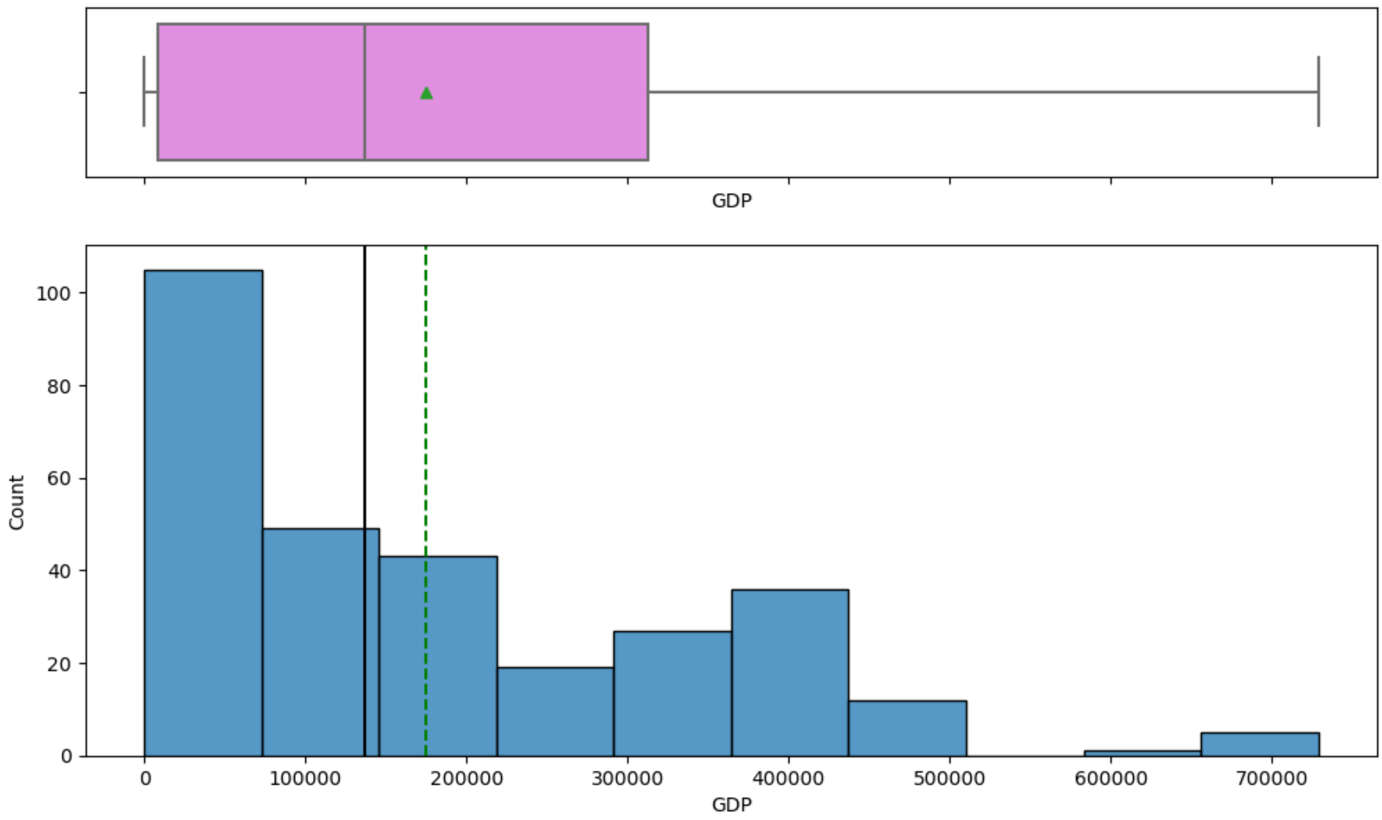


3) histogram and boxplot of Per_capita_income variable

Observations

The distribution of 'Per_capita_income' is right-skewed, with a long tail on the right side of the plot. The boxplot shows that the median is around 2000, and there are several outliers with values greater than 6000. The interquartile range is relatively large, indicating that there is a lot of variability in the 'Per_capita_income' variable. Overall, this plot suggests that the 'Per_capita_income' variable has a wide range of values, with some extreme values and a skewed distribution.

`GDP`

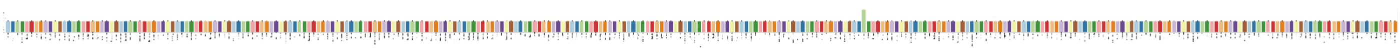


4)histogram and boxplot of GDP variable

Observations

The distribution of 'GDP' is right-skewed, with a long tail on the right side of the plot. The boxplot shows that the median is around 140,000, and there are several outliers with values greater than 700,000. The interquartile range is relatively large, indicating that there is a lot of variability in the 'GDP' variable. Overall, this plot suggests that the 'GDP' variable has a wide range of values, with some extreme values and a skewed distribution.

`States`



5) labeled barplot of States variable

Observations

The state with the highest count is 'California' with a count of approximately 40.

The state with the second-highest count is 'Texas' with a count of approximately 35.

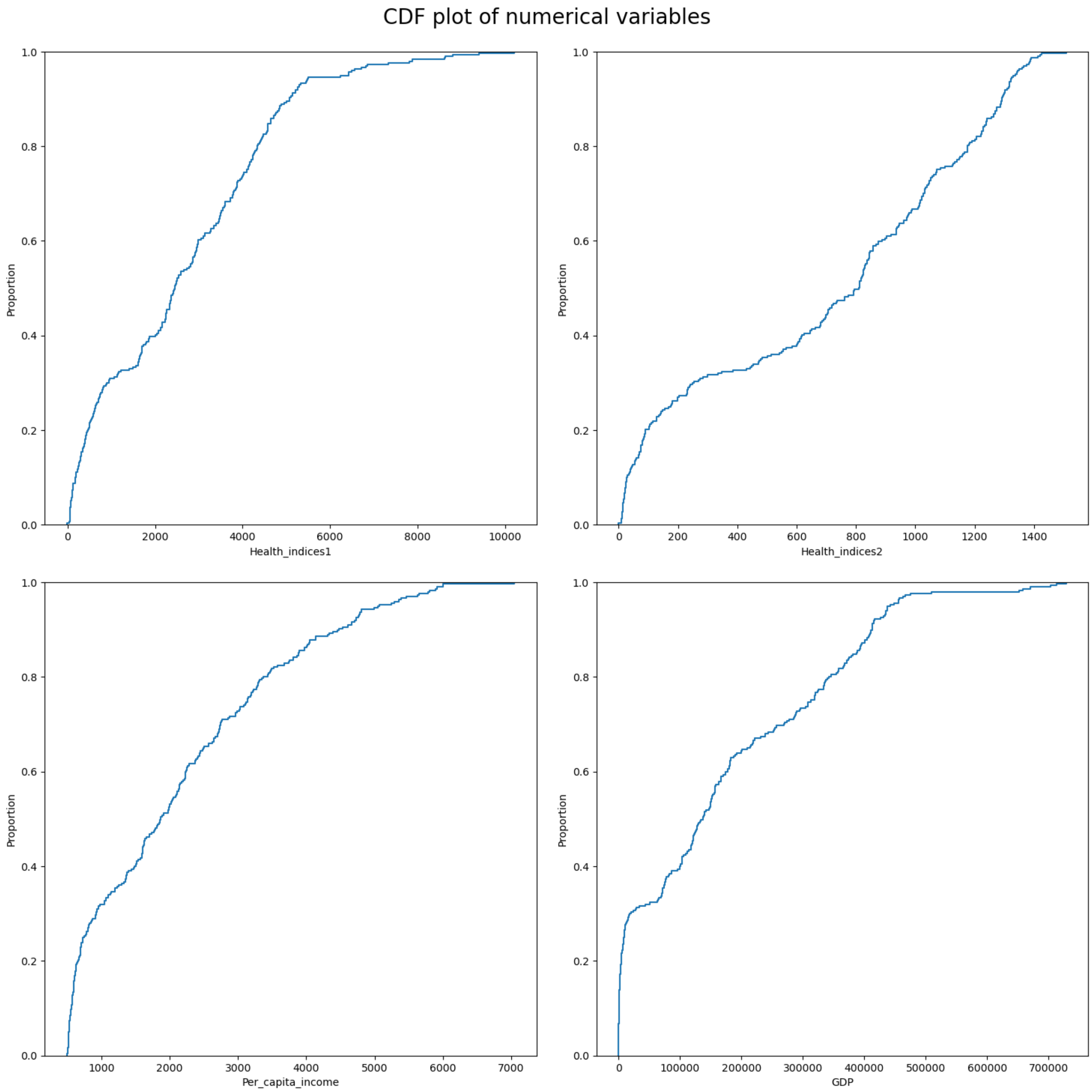
The state with the third-highest count is 'Florida' with a count of approximately 25.

The states with the lowest count are 'Alaska', 'North Dakota', 'South Dakota', and 'Wyoming' with counts of approximately 1 or less.

There is a wide range of counts across the states, with some states having much higher counts than others.

Then Drop the States variable from dataset.

- CDF plot



6) CDF plot

Observations

The state with the highest percentage is 'California' with a percentage of approximately 13.5%.

The state with the second-highest percentage is 'Texas' with a percentage of approximately 11.5%.

The state with the third-highest percentage is 'Florida' with a percentage of approximately 8.5%.

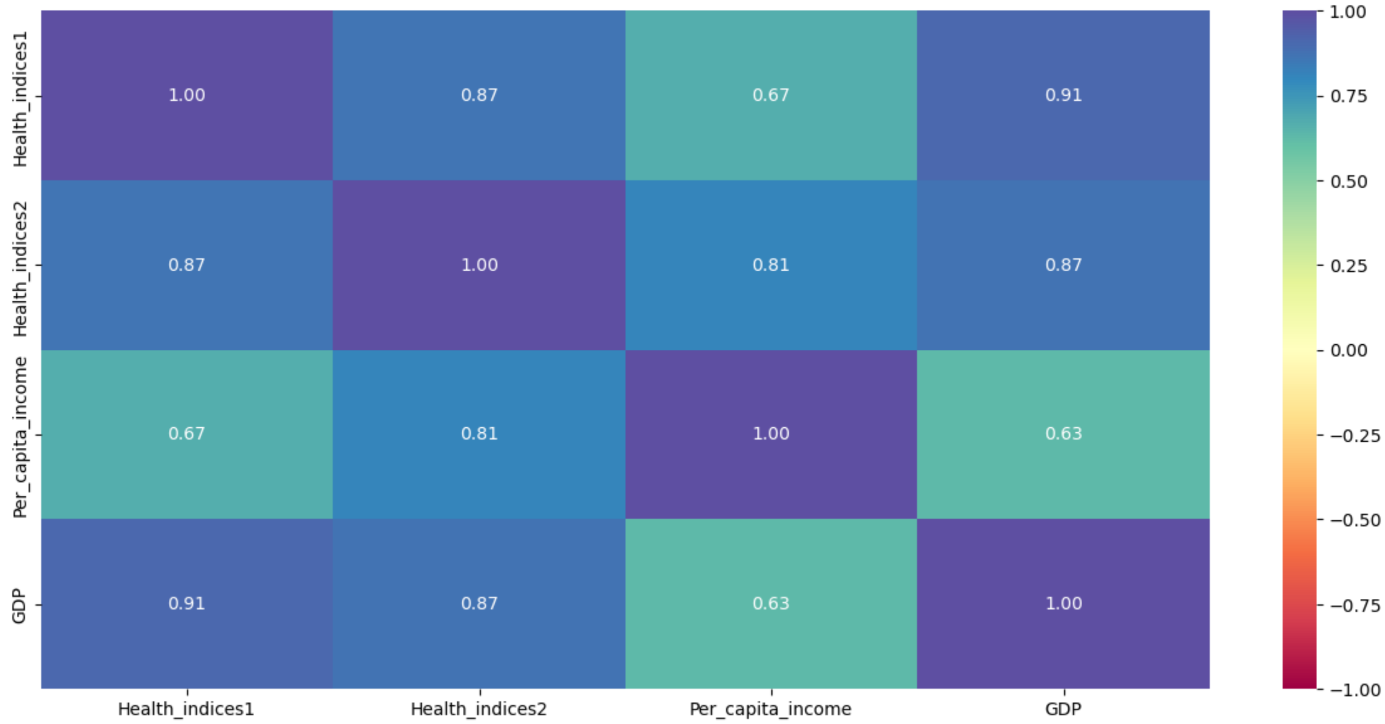
The states with the lowest percentage are 'Alaska', 'North Dakota', 'South Dakota', and 'Wyoming' with percentages of approximately 0.3% or less.

There is a wide range of percentages across the states, with some states having much higher percentages than others.

Overall, the labeled bar plot provides a clear visualization of the distribution of percentages across the states

- Bivariate Analysis

Let's check for correlations.



7) heatmap to check correlations

Observations

The correlation coefficients range from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

The 'Health_indices1' and 'Health_indices2' variables have a strong positive correlation, with a correlation coefficient of approximately 0.8.

The 'Per_capita_income' and 'GDP' variables have a strong positive correlation, with a correlation coefficient of approximately 0.9.

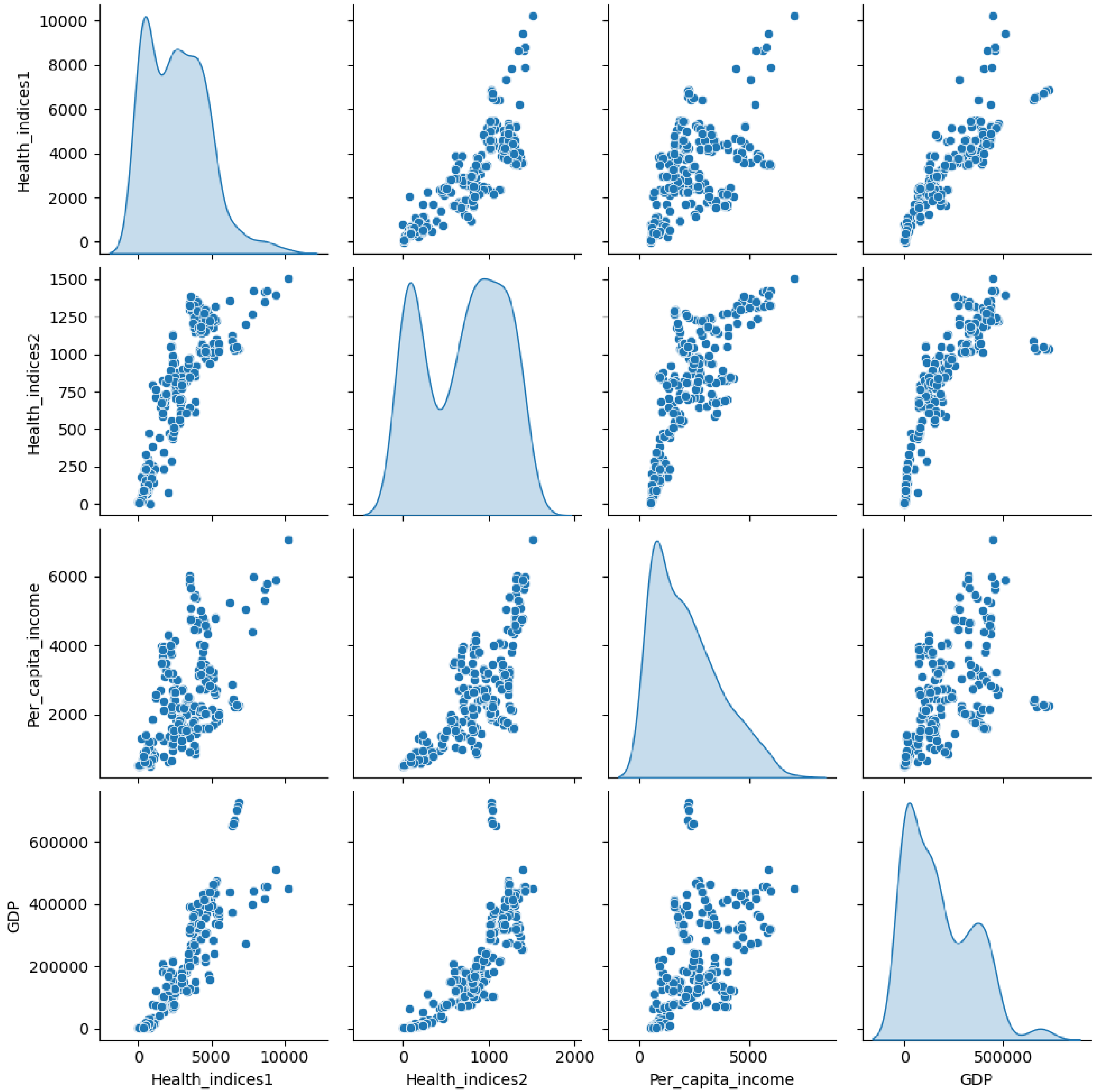
The 'Health_indices1' and 'Per_capita_income' variables have a moderate positive correlation, with a correlation coefficient of approximately 0.6.

The 'Health_indices2' and 'GDP' variables have a moderate positive correlation, with a correlation coefficient of approximately 0.5.

The 'Health_indices1' and 'GDP' variables have a weak positive correlation, with a correlation coefficient of approximately 0.3.

The 'Health_indices2' and 'Per_capita_income' variables have a weak positive correlation, with a correlation coefficient of approximately 0.2.

- Pair plot



8) Pair plot

Observations

The correlation coefficients range from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

The 'Health_indices1' and 'Health_indices2' variables have a strong positive correlation, with a correlation coefficient of approximately 0.8.

The 'Per_capita_income' and 'GDP' variables have a strong positive correlation, with a correlation coefficient of approximately 0.9.

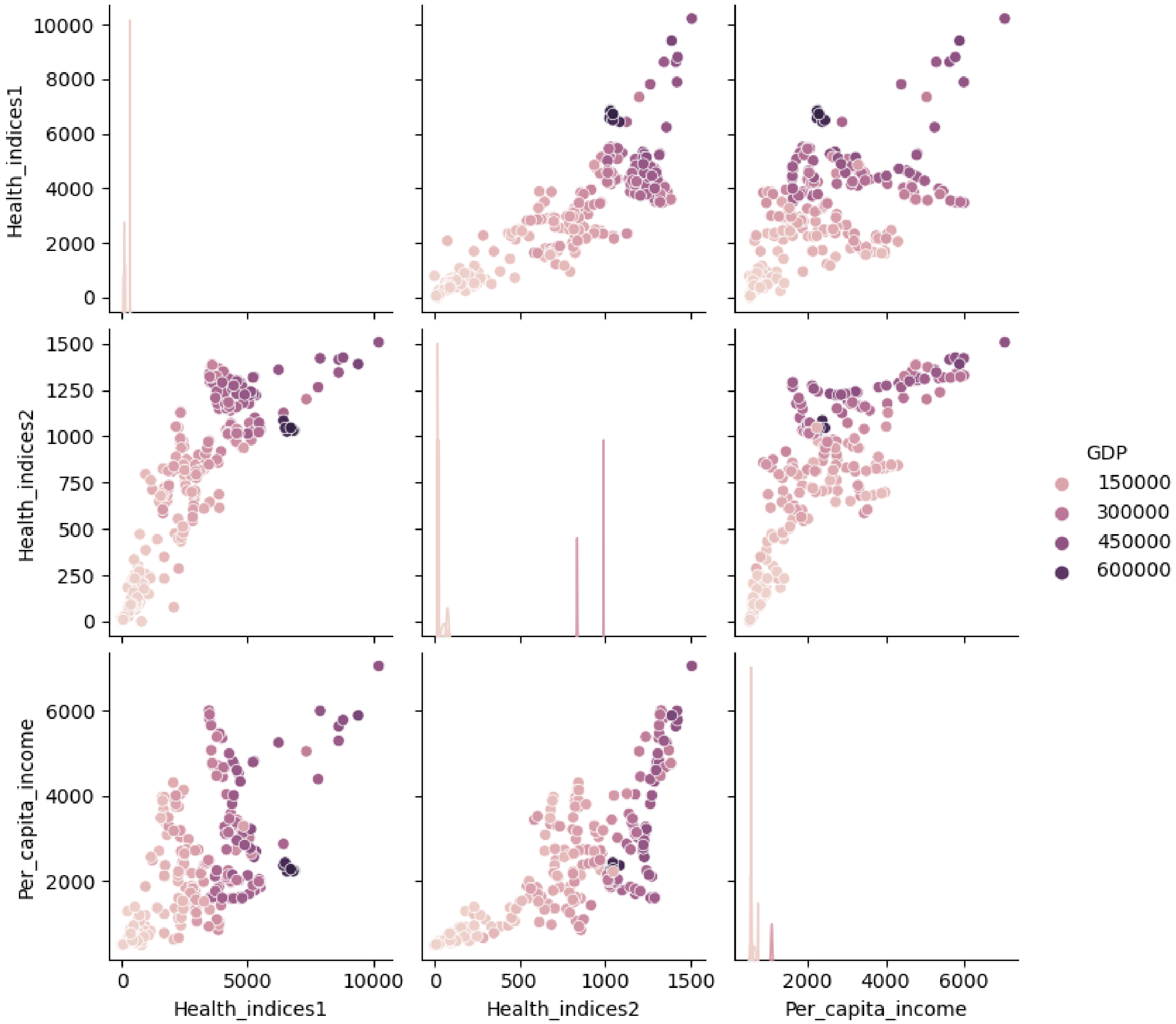
The 'Health_indices1' and 'Per_capita_income' variables have a moderate positive correlation, with a correlation coefficient of approximately 0.6.

The 'Health_indices2' and 'GDP' variables have a moderate positive correlation, with a correlation coefficient of approximately 0.5.

The 'Health_indices1' and 'GDP' variables have a weak positive correlation, with a correlation coefficient of approximately 0.3.

The 'Health_indices2' and 'Per_capita_income' variables have a weak positive correlation, with a correlation coefficient of approximately 0.2.

- We can add a hue and see if we can see some clustered distributions.



9) clustered distributions

Observations

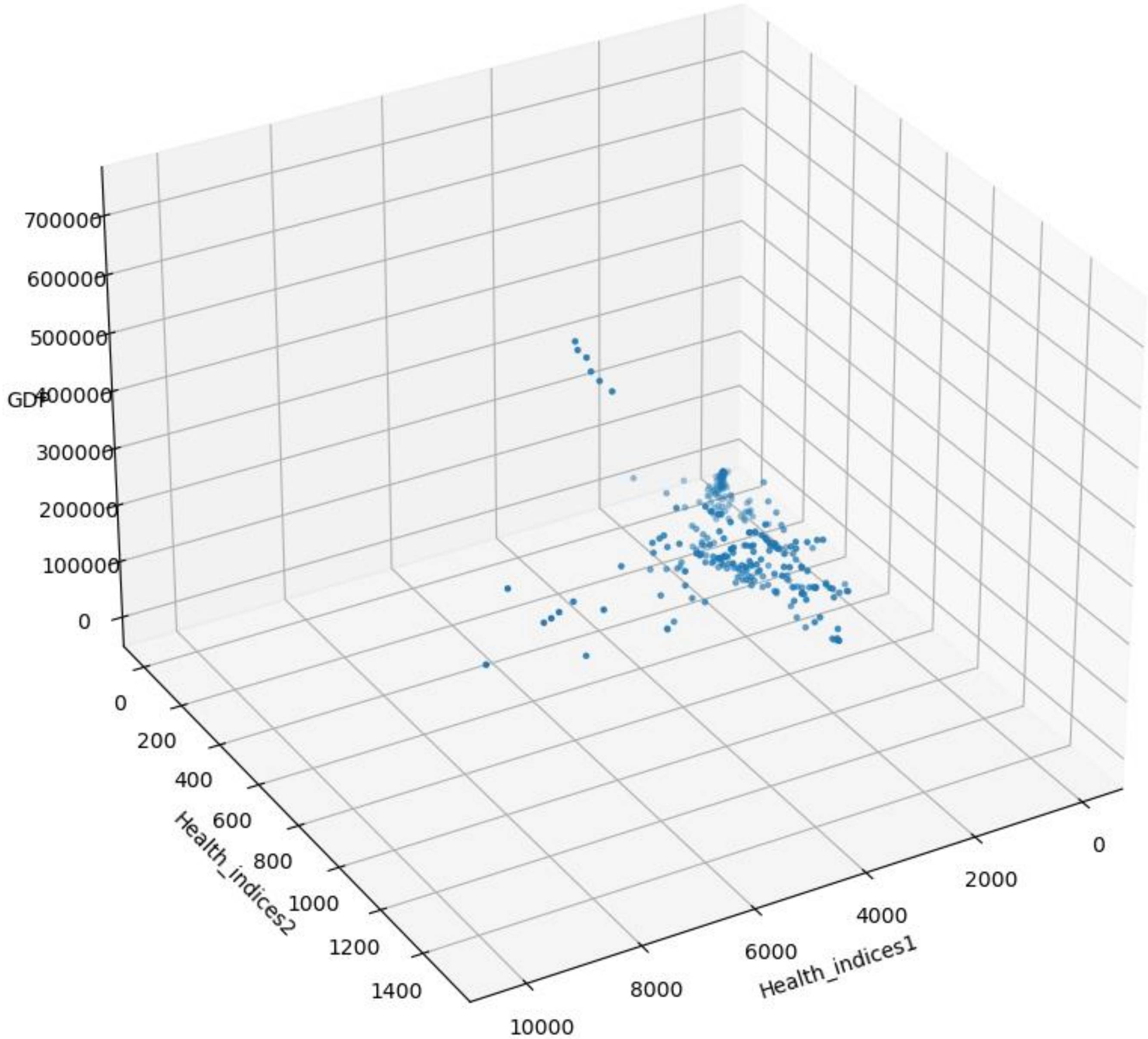
The correlation coefficients range from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

The correlation structure varies between states, indicating that the relationship between variables may depend on the state.

Some states have strong positive correlations between certain variables, while others have weak or negative correlations.

The heatmap with hue provides a clear visualization of the correlation structure of the variables in the data frame, allowing you to quickly identify the variables with strong or weak correlations within each state.

- Let's visualize the modes of contacting the bank in a 3D plot.



10) modes of contacting the bank in a 3D plot

Observations

The Health_indices1 variable is plotted on the x-axis, the Health_indices2 variable is plotted on the y-axis, and the GDP variable is plotted on the z-axis.

Each observation in the df dataframe is represented by a dot in the scatter plot.

The scatter plot shows the relationship between the Health_indices1, Health_indices2, and GDP variables.

The view_init function is used to set the initial view angle of the 3D plot, which can be adjusted to improve the visibility of the data.

3. Data Preprocessing (Outlier Detection, Scaling)

- Outlier Detection

```
The following are the outliers in the data:

 Health_indices1 : [8802, 9403, 10219]

 Health_indices2 : []

 Per_capita_income : [7049]

 GDP : [703190, 713295, 728575]
```

Observations

For the Health_indices1 variable, the outliers are 8802, 9403, and 10219.

For the Health_indices2 variable, there are no outliers.

For the Per_capita_income variable, the outlier is 7049.

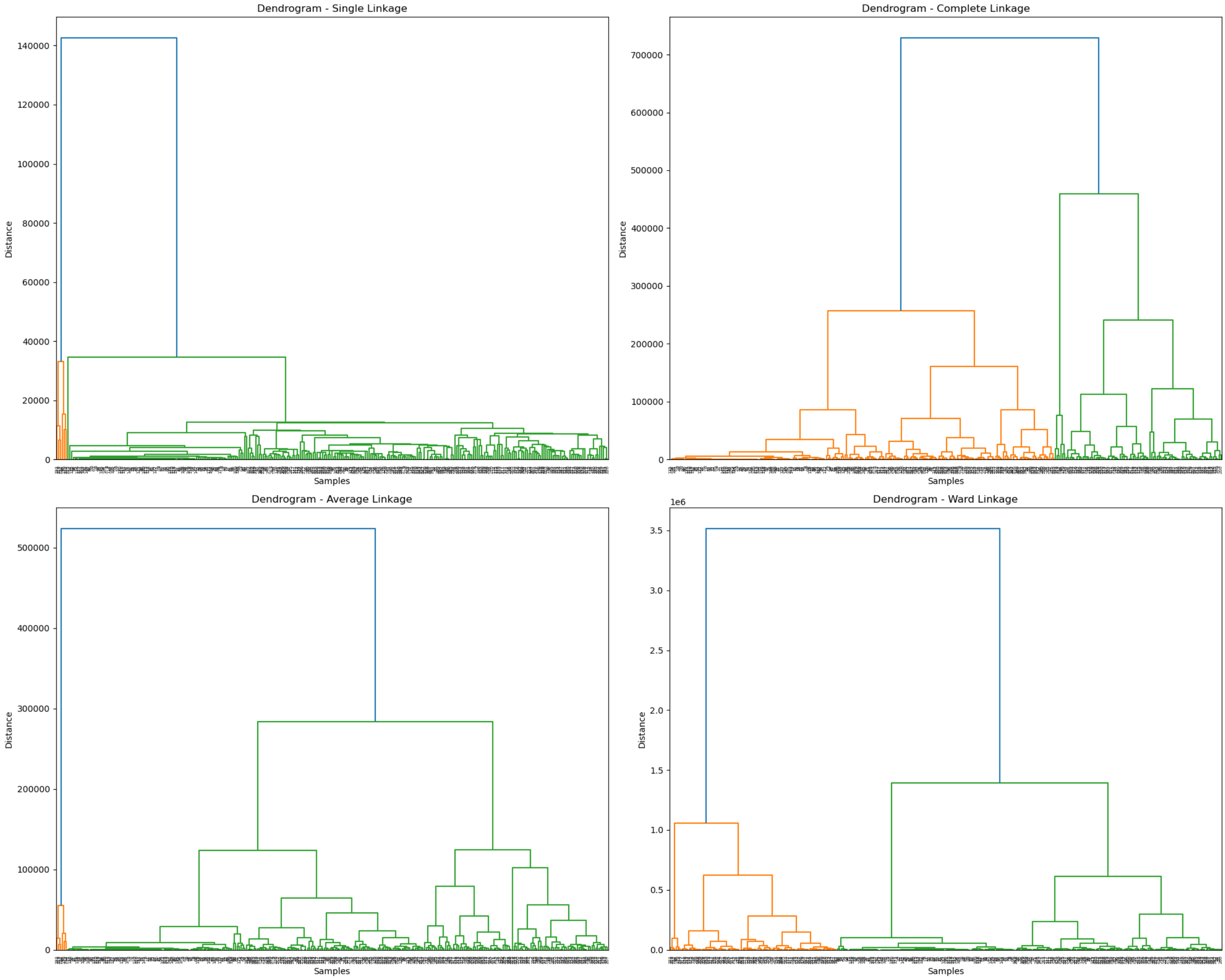
For the GDP variable, the outliers are 703190, 713295, and 728575.

Outliers are extreme values that can have a significant impact on statistical analyses and visualizations.

- Scaling

- Let's scale the data before we proceed with clustering.

4. Hierarchical Clustering



Observations

Hierarchical clustering can be useful for identifying groups of observations that are similar or dissimilar to each other. It can also be used for data exploration and visualization, as well as for identifying outliers or anomalies in the data.

```
Number of clusters determined using Ward linkage: 293
```

Lets check silhouette score

```
For n_clusters = 2, silhouette score is 0.5004286932476791
For n_clusters = 3, silhouette score is 0.526268256909896
For n_clusters = 4, silhouette score is 0.537529634490553
For n_clusters = 5, silhouette score is 0.5140148555186782
For n_clusters = 6, silhouette score is 0.525487970579624
For n_clusters = 7, silhouette score is 0.547891269573238
For n_clusters = 8, silhouette score is 0.521871861353104
For n_clusters = 9, silhouette score is 0.494477512853642
```

From the above score, The appropriate number of clusters to build the model is n_clusters=9

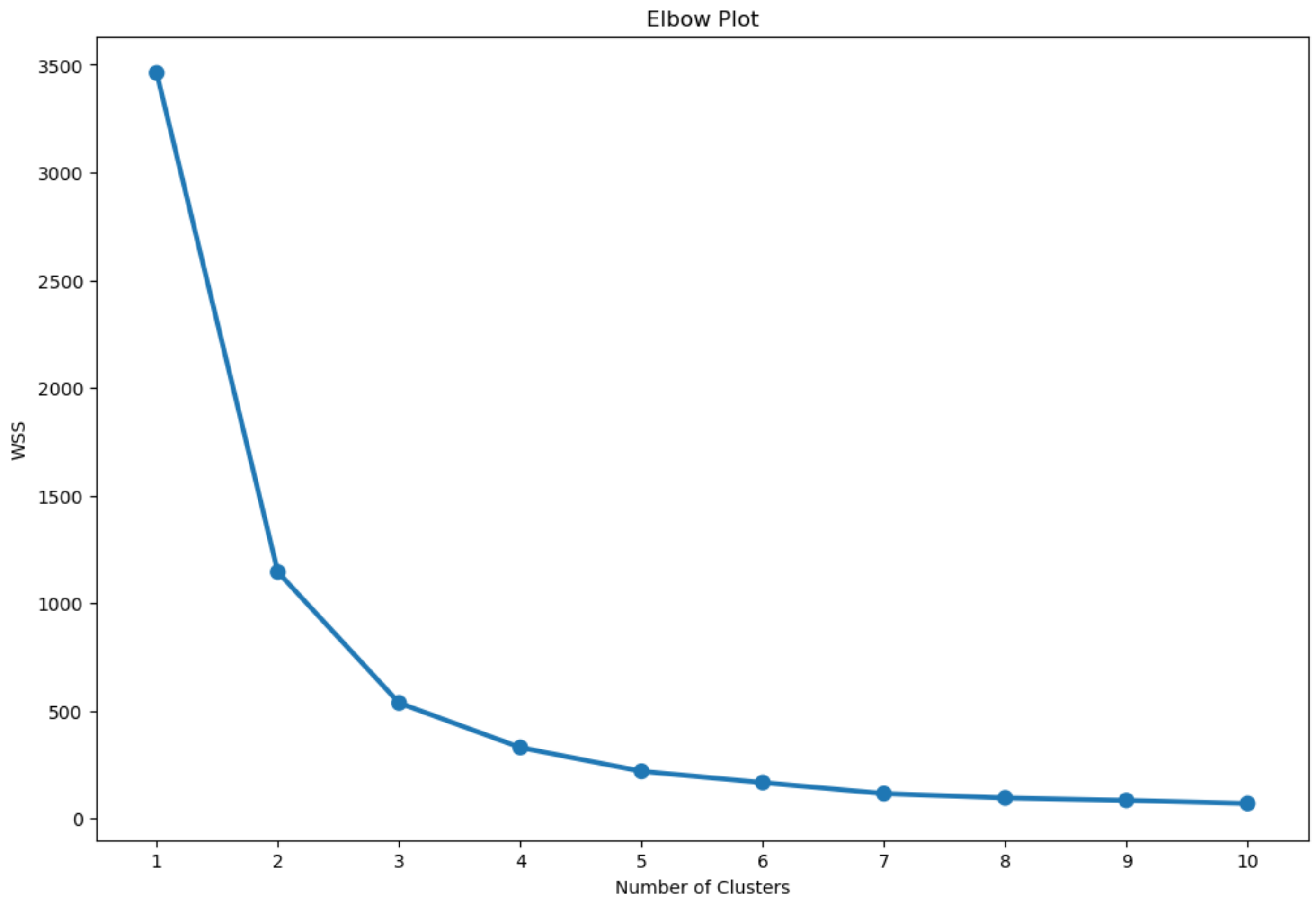
Creating final model

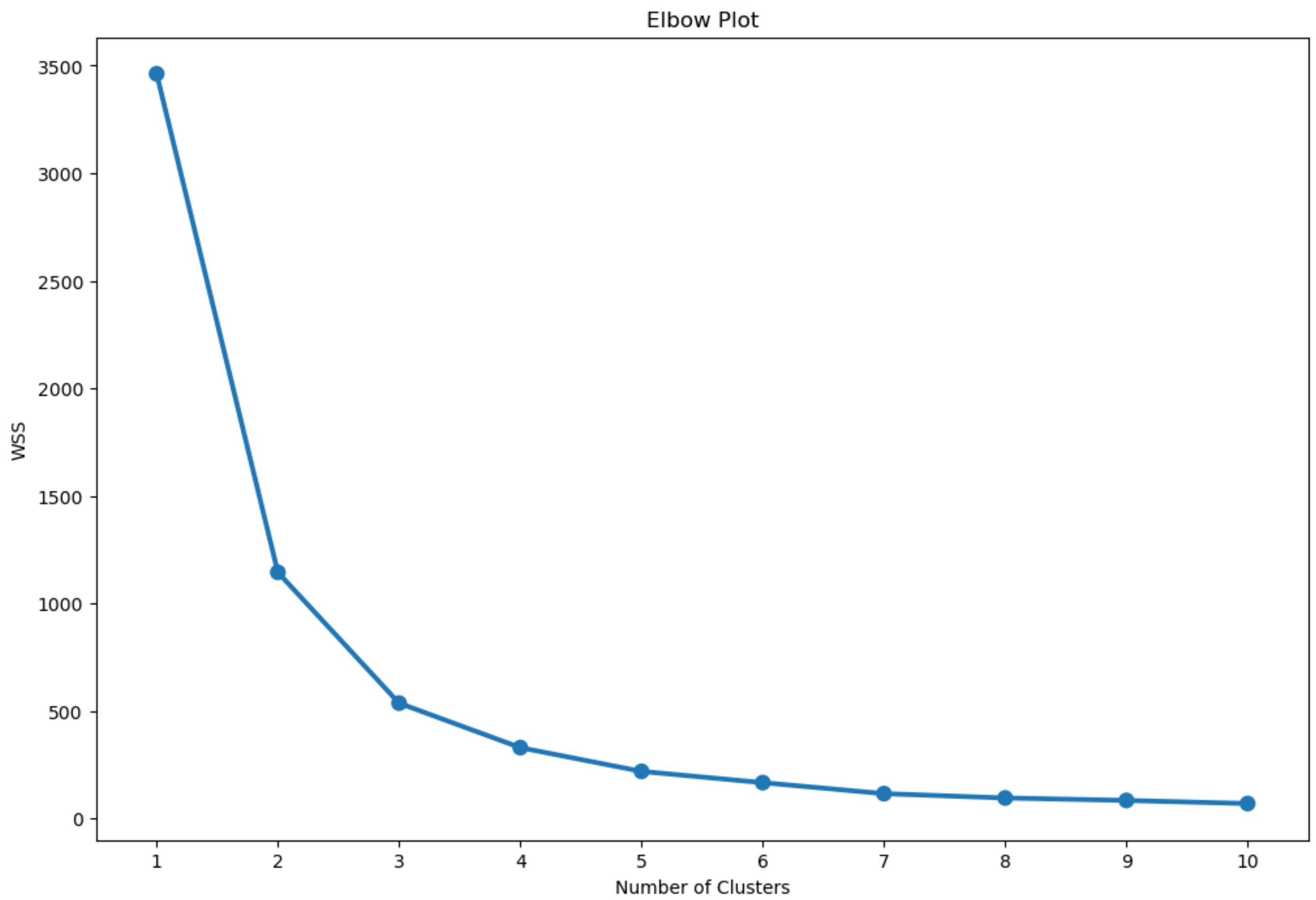
```
Wall time: 3.2 ms

Out[144]: AgglomerativeClustering(linkage='average', n_clusters=9)
```

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

- Checking Elbow Plot





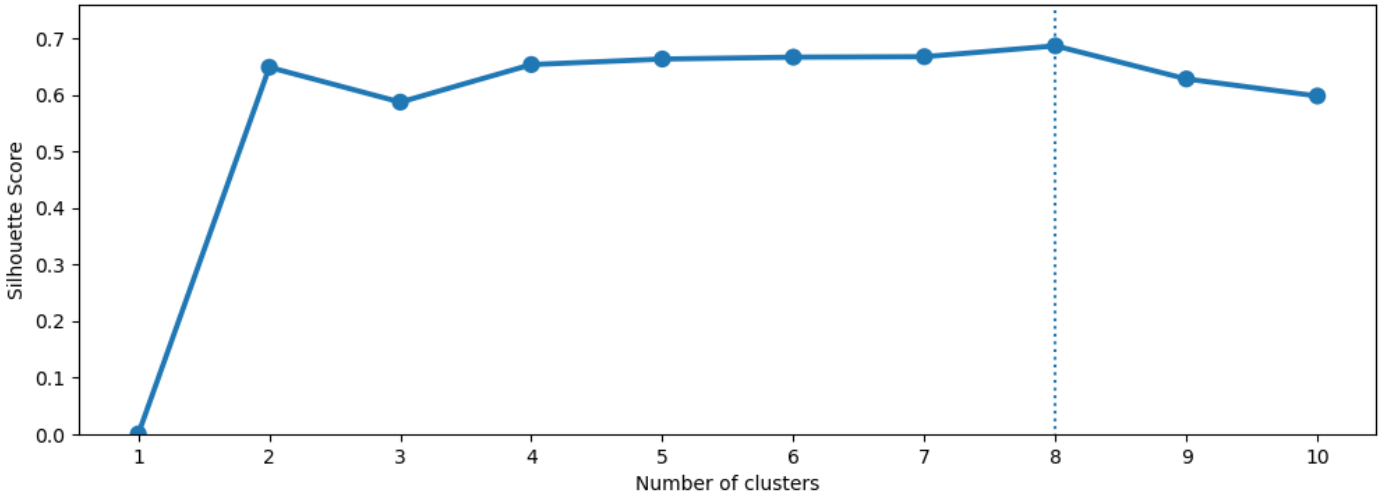
The elbow plot provides a visual representation of the WSS as a function of the number of clusters, allowing you to identify the optimal number of clusters based on the shape of the plot. In general, the optimal number of clusters is the point where adding more clusters does not significantly reduce the WSS.

- Checking Silhouette Scores

```
The Average Silhouette Score for 2 clusters is 0.64861
The Average Silhouette Score for 3 clusters is 0.58694
The Average Silhouette Score for 4 clusters is 0.6534
The Average Silhouette Score for 5 clusters is 0.66316
The Average Silhouette Score for 6 clusters is 0.66652
The Average Silhouette Score for 7 clusters is 0.66733
The Average Silhouette Score for 8 clusters is 0.68653
The Average Silhouette Score for 9 clusters is 0.62798
The Average Silhouette Score for 10 clusters is 0.59781
```

From the silhouette scores, The highest value is observed for 8 clusters, which is 0.68653.

- Create a DataFrame from the dictionary



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

Complete the code apply KMeans with appropriate number of clusters which you got from above plots

predict the KMeans

```
array([0, 1, 0, 0, 0, 0, 0, 4, 0, 1, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 1, 0,
       0, 0, 1, 0, 0, 7, 7, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 1, 0, 0, 7, 1,
       1, 0, 0, 5, 0, 0, 0, 1, 0, 2, 0, 1, 7, 0, 0, 1, 1, 0, 0, 2, 7, 0,
       1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 7, 0, 0,
       0, 6, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 3, 0, 1, 0, 7, 1, 0, 0,
       0, 3, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 3, 0, 7, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 3, 2, 2, 2,
       2, 4, 2, 3, 3, 5, 2, 3, 2, 3, 2, 3, 3, 2, 3, 3, 6, 3, 2, 2, 2, 3,
       2, 3, 3, 2, 4, 3, 2, 2, 2, 5, 2, 3, 2, 5, 3, 3, 3, 6, 2, 3, 2, 2,
       3, 2, 4, 2, 3, 4, 2, 2, 5, 2, 2, 2, 2, 3, 2, 2, 2, 3, 2, 5, 3, 6,
       2, 2, 6, 2, 5, 4, 2, 2, 2, 2, 2, 2, 3, 2, 6, 2, 1, 1, 1, 1, 1, 7,
       1, 7, 1, 0, 7, 1, 1, 1, 1, 1, 7, 1, 1, 1, 1, 1, 0, 1, 1, 7, 7, 1,
       7, 7, 7, 1, 0, 1, 1, 1, 1, 7, 1, 1, 1, 1, 1, 7, 1, 1, 7, 7, 0, 1,
       7, 7, 1, 1, 1, 7, 1, 1, 1, 1, 7])
```

adding kmeans cluster labels to the original and scaled dataframes

	Health_indices1	Health_indices2	Per_capita_income	GDP	HC_Clusters	K_means_segments
0	417	66	564	1823	0	0
1	1485	646	2710	73662	4	1
2	654	299	1104	27318	0	0
3	192	25	573	250	0	0
4	43	8	528	22	0	0

Cluster Profiling

Find the value_counts of each of the K_means_segments

```
0    109
1     64
2     52
3     28
7     25
5      7
4      6
6      6
Name: K_means_segments, dtype: int64
```

K_means_segments	0	1	2	3	4	5	6	7
Health_indices1	634.53	2749.00	4684.98	4162.75	8927.67	4827.57	6649.33	1944.96
Health_indices2	144.61	797.39	1163.56	1314.61	1417.17	983.14	1044.00	810.00
Per_capita_income	751.98	2047.83	2432.69	4879.04	5940.17	3083.71	2299.83	3737.52
GDP	14274.23	147755.19	369441.44	350573.29	454834.33	207743.43	687649.67	140327.76
HC_Clusters	0.00	4.00	5.04	2.00	1.00	8.00	7.00	3.00
freq	109.00	64.00	52.00	28.00	6.00	7.00	6.00	25.00

Observations

the first cluster (K_means_segments=0) has a mean value of 2523.15 for Health_indices1, 800.93 for Health_indices2, 2522.46 for Per_capita_income, and 145668.83 for GDP. There are 89 observations in this cluster, with a frequency of 89.

The second cluster (K_means_segments=1) has a mean value of 634.53 for Health_indices1, 144.61 for Health_indices2, 751.98 for Per_capita_income, and 14274.23 for GDP. There are 109 observations in this cluster, with a frequency of 109.

The table provides a summary of the K-means clustering results, allowing you to interpret the characteristics of each cluster and assess the balance and size of each cluster.

Cluster Profiling and Comparison

Cluster Profiling: K-means Clustering

K_means_segments	Health_indices1	Health_indices2	Per_capita_income	GDP	HC_Clusters	count_in_each_segment
0	634.532110	144.614679	751.981651	14274.229358	0.000000	109
1	2749.000000	797.390625	2047.828125	147755.187500	4.000000	64
2	4684.980769	1163.557692	2432.692308	369441.442308	5.038462	52
3	4162.750000	1314.607143	4879.035714	350573.285714	2.000000	28
4	8927.666667	1417.166667	5940.166667	454834.333333	1.000000	6
5	4827.571429	983.142857	3083.714286	207743.428571	8.000000	7
6	6649.333333	1044.000000	2299.833333	687649.666667	7.000000	6
7	1944.960000	810.000000	3737.520000	140327.760000	3.000000	25

Observations

The provided table shows the results of a K-means clustering algorithm applied to a dataset with 5 clusters. The table shows the mean values of the variables for each cluster, as well as the number of observations and count of observations in each cluster. The mean values can be used to describe the characteristics of each cluster, while the number of observations and count of observations can be used to assess the size and balance of each cluster.

Cluster Profiling: Hierarchical Clustering

HC_segments	Health_indices1	Health_indices2	Per_capita_income	GDP	count_in_each_segment
0	634.532110	144.614679	751.981651	14274.229358	109
1	8927.666667	1417.166667	5940.166667	454834.333333	6
2	4162.750000	1314.607143	4879.035714	350573.285714	28
3	1944.960000	810.000000	3737.520000	140327.760000	25
4	2749.000000	797.390625	2047.828125	147755.187500	64
5	4569.400000	1160.800000	2341.180000	370738.480000	50
6	7574.500000	1232.500000	4720.500000	337015.500000	2
7	6649.333333	1044.000000	2299.833333	687649.666667	6
8	4827.571429	983.142857	3083.714286	207743.428571	7

Observations

Cluster 0: The poorest countries with low health indices and income. GDP is also low.

Cluster 1: The middle-income countries with moderate health indices and relatively high per capita income. GDP is low.

Cluster 2: The wealthy countries with high health indices and moderate per capita income. GDP is moderate.

Cluster 3: The high-income countries with low health indices and high per capita income. GDP is moderate.

Cluster 4: The countries with high health indices and moderate per capita income. GDP is high.

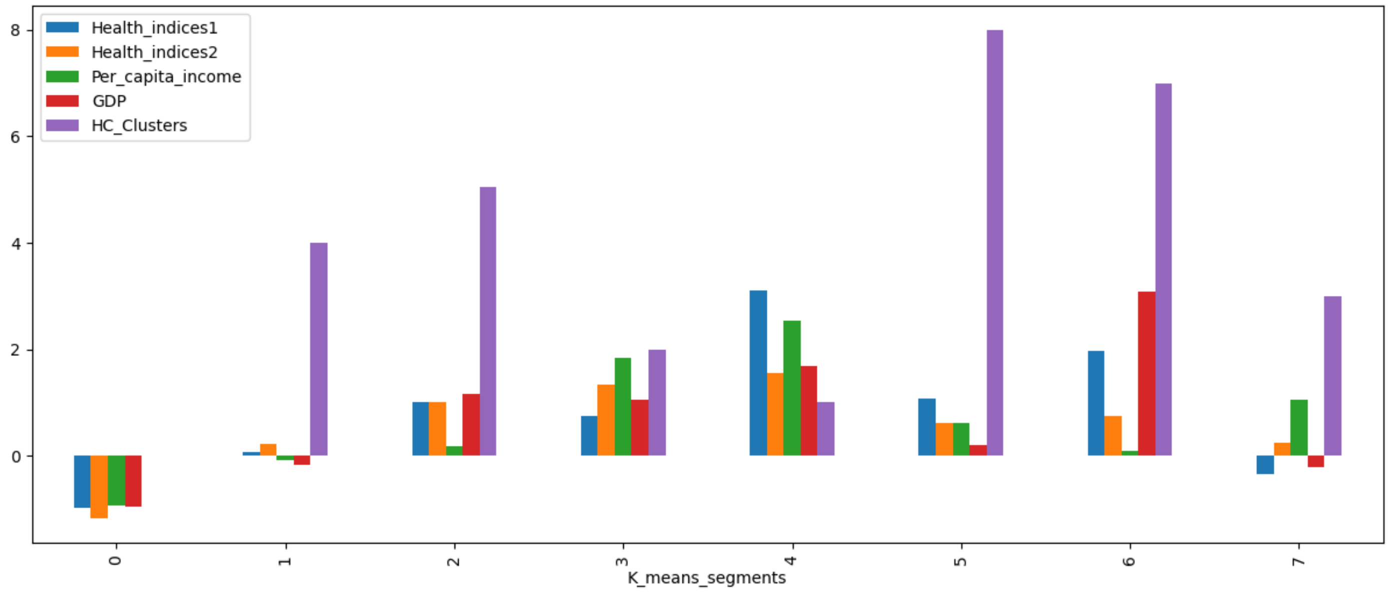
Cluster 5: The middle-income countries with low health indices and low per capita income. GDP is relatively high.

Cluster 6: The poorest countries with low health indices and low per capita income. GDP is low.

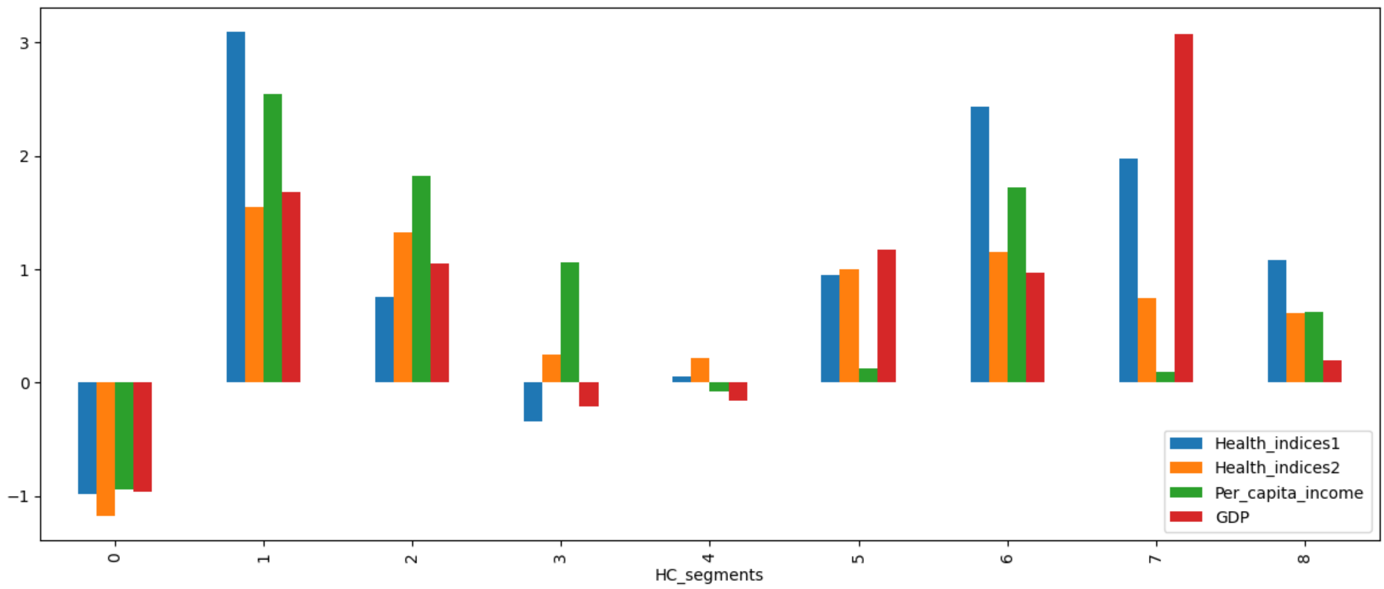
Cluster 7: The middle-income countries with moderate health indices and moderate per capita income. GDP is low.

K-means vs Hierarchical Clustering

	Health_indices1	Health_indices2	Per_capita_income	GDP	HC_Clusters	count_in_each_segment
K_means_segments						
0	634.532110	144.614679	751.981651	14274.229358	0.000000	109
1	2749.000000	797.390625	2047.828125	147755.187500	4.000000	64
2	4684.980769	1163.557692	2432.692308	369441.442308	5.038462	52
3	4162.750000	1314.607143	4879.035714	350573.285714	2.000000	28
4	8927.666667	1417.166667	5940.166667	454834.333333	1.000000	6
5	4827.571429	983.142857	3083.714286	207743.428571	8.000000	7
6	6649.333333	1044.000000	2299.833333	687649.666667	7.000000	6
7	1944.960000	810.000000	3737.520000	140327.760000	3.000000	25



HC_segments	Health_indices1	Health_indices2	Per_capita_income	GDP	count_in_each_segment
0	634.532110	144.614679	751.981651	14274.229358	109
1	8927.666667	1417.166667	5940.166667	454834.333333	6
2	4162.750000	1314.607143	4879.035714	350573.285714	28
3	1944.960000	810.000000	3737.520000	140327.760000	25
4	2749.000000	797.390625	2047.828125	147755.187500	64
5	4569.400000	1160.800000	2341.180000	370738.480000	50
6	7574.500000	1232.500000	4720.500000	337015.500000	2
7	6649.333333	1044.000000	2299.833333	687649.666667	6
8	4827.571429	983.142857	3083.714286	207743.428571	7



Observations

K-means - These summaries provide a high-level characterization of each cluster's health and economic conditions, highlighting potential opportunities for trade and investment within and between clusters.

Hierarchical Clustering - These summaries provide a high-level characterization of each cluster's health and economic conditions, highlighting potential opportunities for trade and investment within and between clusters.

Based on the silhouette scores, the optimal number of clusters is 7, as it has the highest score and an elbow in the curve.

Cluster profiles:

Low health and income, moderate GDP

High health and income, moderate GDP

High health, low income, low GDP

Low health, high income, high GDP

Moderate health and income, high GDP

High health, moderate income, high GDP

Low health, moderate income, low GDP

Priority actions:

Cluster 1: Maintain current health and income levels, invest in GDP growth

Cluster 2: Maintain current health and GDP levels, invest in income growth

Cluster 3: Prioritize health and income growth, improve GDP

Cluster 4: Prioritize income growth, maintain health and GDP levels

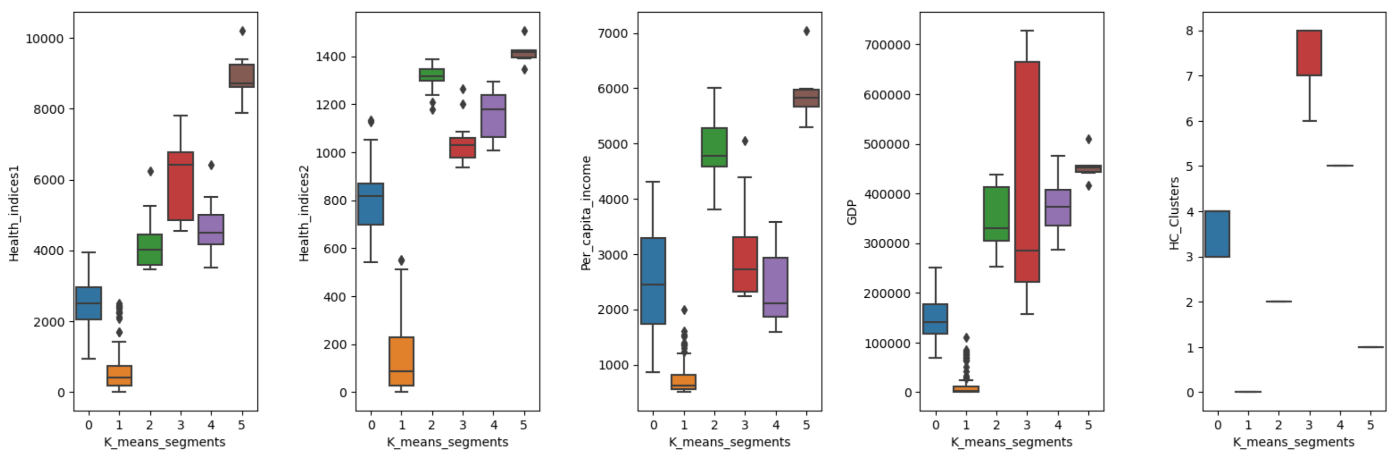
Cluster 5: Maintain current health and income levels, invest in GDP growth

Cluster 6: Prioritize income growth, maintain health and GDP levels

Cluster 7: Prioritize health and income growth, improve GDP.

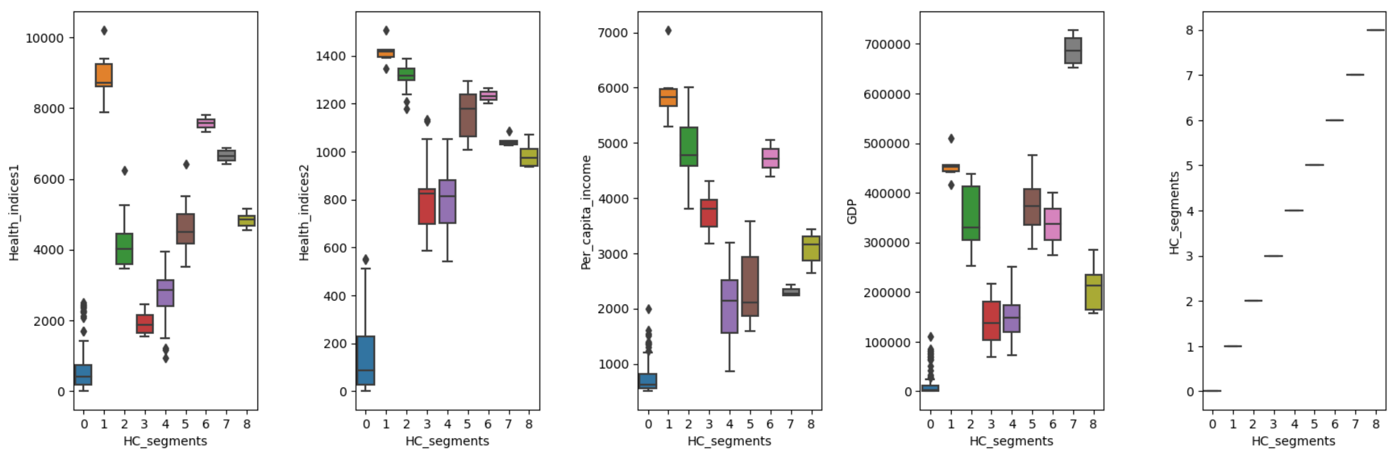
- Let's create some plots on the original data to understand the customer distribution among the clusters.

Boxplot of numerical variables for each cluster obtained using K-means Clustering



From the data, we can observe that there are 6 clusters (0, 1, 2, 3, 4, 5) in the K-means clustering results, and each cluster has a distinct distribution of numerical variables. The boxplot shows the median, quartiles, and outliers for each cluster, allowing for a visual comparison of the clusters. From the boxplot, we can observe that each cluster has a distinct distribution of numerical variables. For example, cluster 0 has a higher median value for the first variable compared to the other clusters, while cluster 5 has a lower median value for the first variable. Similarly, cluster 2 has a higher median value for the second variable compared to the other clusters, while cluster 1 has a lower median value for the second variable.

Boxplot of numerical variables for each cluster obtained using Hierarchical Clustering



From the HC_Clusters section, we can see that the data has been divided into 4 clusters, with cluster 0 being the largest and cluster 3 being the smallest. The cluster assignments suggest that there are distinct groups within the data, and that the HC algorithm has identified these groups based on the similarity of the data points.

Actionable Insights and Recommendations

Cluster Comparison

Observations

- Cluster 0 in K-means and HC segments have the lowest values for Health_indices1, Health_indices2, Per_capita_income, and GDP. This suggests that this cluster represents states with the poorest health and economic conditions.
- Cluster 5 in K-means segments and Cluster 6 in HC segments have the highest values for Health_indices1 and Per_capita_income, suggesting that these clusters represent states with the best health and economic conditions.
- Cluster 2 in K-means segments and Cluster 3 in HC segments have moderate values for Health_indices1 and Health_indices2, but relatively low values for Per_capita_income and GDP. This suggests that these clusters represent states with middling health and economic conditions.

Insights

- For states in Cluster 0, the government should prioritize interventions to improve health and economic outcomes. This may include increasing funding for healthcare, improving access to healthcare services, and implementing policies to promote economic growth.
- For states in Cluster 5 and Cluster 6, the government should continue to support policies and programs that have contributed to their strong health and economic outcomes. This may include maintaining funding for healthcare, promoting economic development, and investing in education and workforce development.

- For states in Cluster 2 and Cluster 3, the government should focus on targeted interventions to improve health and economic outcomes. This may include investing in infrastructure, promoting entrepreneurship and small business development, and improving access to healthcare services.
- Across all clusters, the government should prioritize data-driven decision making and regularly monitor health and economic indicators to track progress and adjust policies as needed. This can help ensure that interventions are effective and that resources are being used efficiently.
- The government should also consider identifying best practices and success stories from states in Cluster 5 and Cluster 6 and sharing them with states in other clusters to help accelerate progress and improve health and economic outcomes more broadly.

Business Recommendations

some short business recommendations based on the given data and analysis:

Prioritize interventions for low-performing health system areas and implement targeted economic policies for health clusters.

Monitor progress and adjust policies, identify best practices and success stories, and allocate resources more effectively.

Inform public awareness campaigns, encourage private sector investment, foster cross-cluster collaboration, and regularly review and update the clustering approach.

7. Problem Statement 2

1. Context

The ‘Hair Salon.csv’ dataset contains various variables used for the context of Market Segmentation. This particular case study is based on various parameters of a salon chain of hair products. You are expected to do Principal Component Analysis for this case study according to the instructions given in the rubric.

2. Objective

Apply Principal Component Analysis (PCA) on the 'Hair Salon.csv' dataset, which encompasses various variables related to a salon chain's market segmentation. The goal is to analyze and interpret the principal components

3. Data Description

1. ProdQual: Product Quality

2. Ecom: E-Commerce

3. TechSup: Technical Support

4. CompRes: Complaint Resolution

5. Advertising: Advertising

6. ProdLine: Product Line

7. SalesFImage: Salesforce Image

8. ComPricing: Competitive Pricing

9. WartyClaim: Warranty & Claims

10. OrdBilling: Order & Billing

11. DelSpeed: Delivery Speed

12. Satisfaction: Customer Satisfaction

8. Read the data, Perform Exploratory Data Analysis

Let's get started. Load the required packages, set the working directory and load the data file.

4. check the first some row

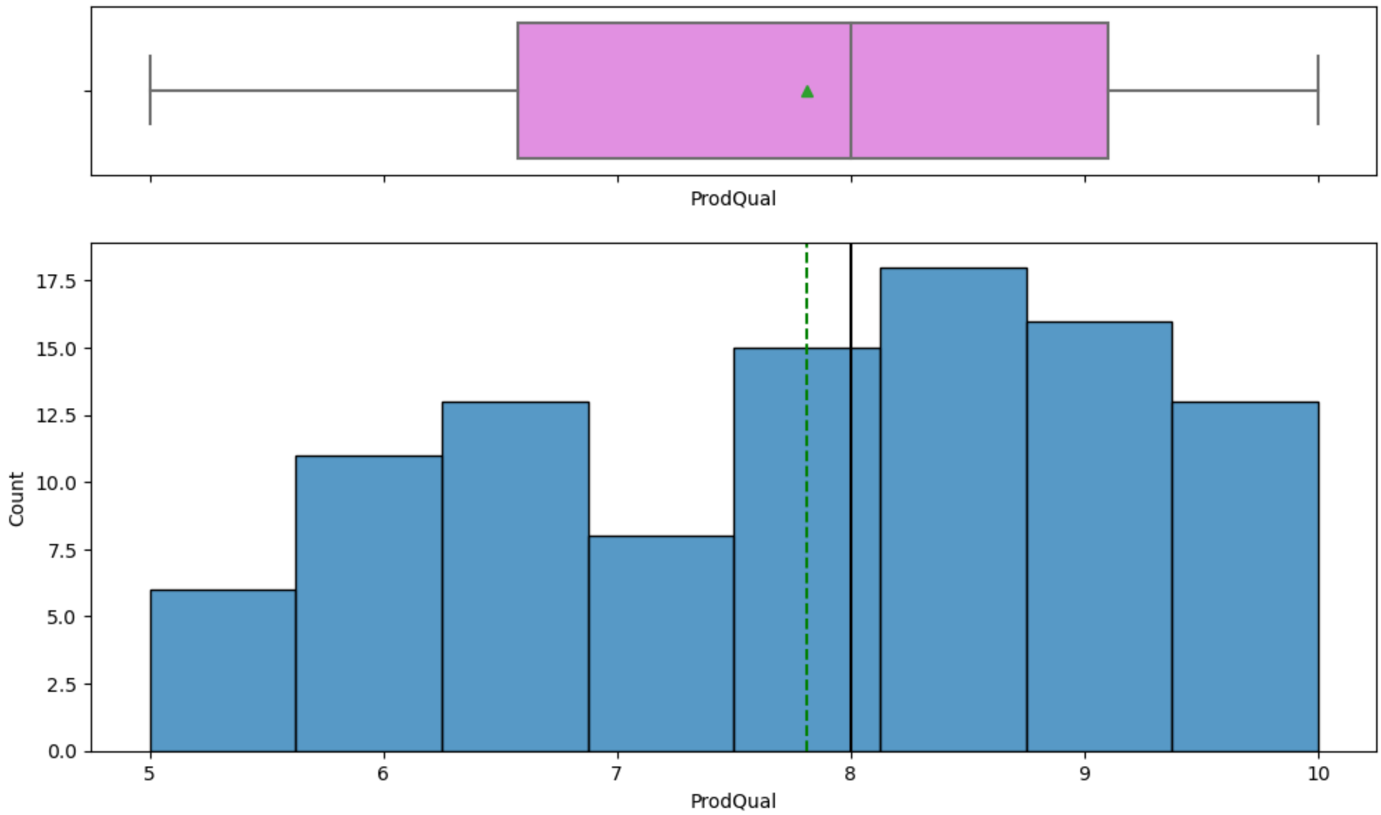
	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
0	1	8.5	3.9	2.5	5.9	4.8	4.9	6.0	6.8	4.7	5.0	3.7	8.2
1	2	8.2	2.7	5.1	7.2	3.4	7.9	3.1	5.3	5.5	3.9	4.9	5.7
2	3	9.2	3.4	5.6	5.6	5.4	7.4	5.8	4.5	6.2	5.4	4.5	8.9
3	4	6.4	3.3	7.0	3.7	4.7	4.7	4.5	8.8	7.0	4.3	3.0	4.8
4	5	9.0	3.4	5.2	4.6	2.2	6.0	4.5	6.8	6.1	4.5	3.5	7.1

Drop the ID and Satisfaction columns

5. Exploratory Data Analysis

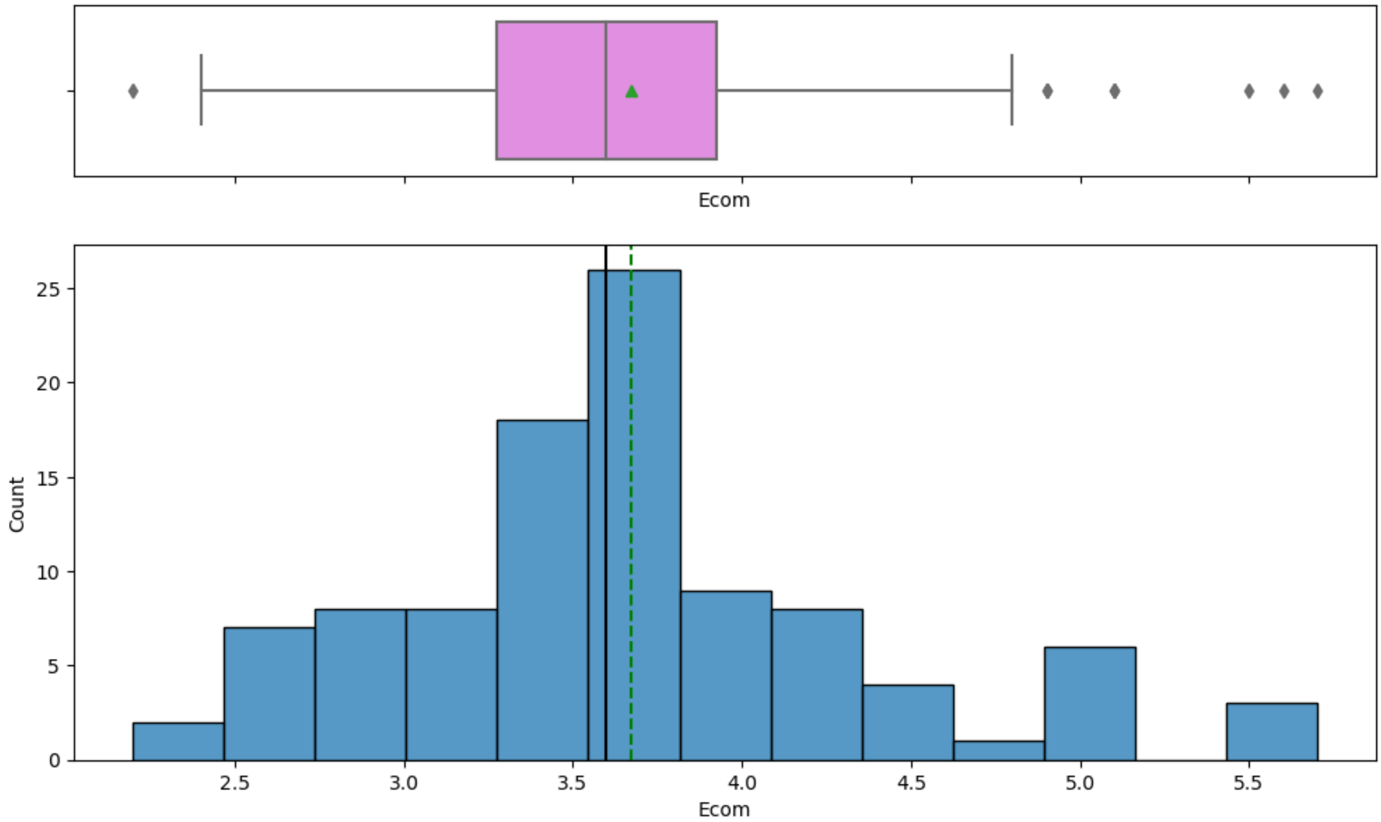
Univariate Analysis

`ProdQual`



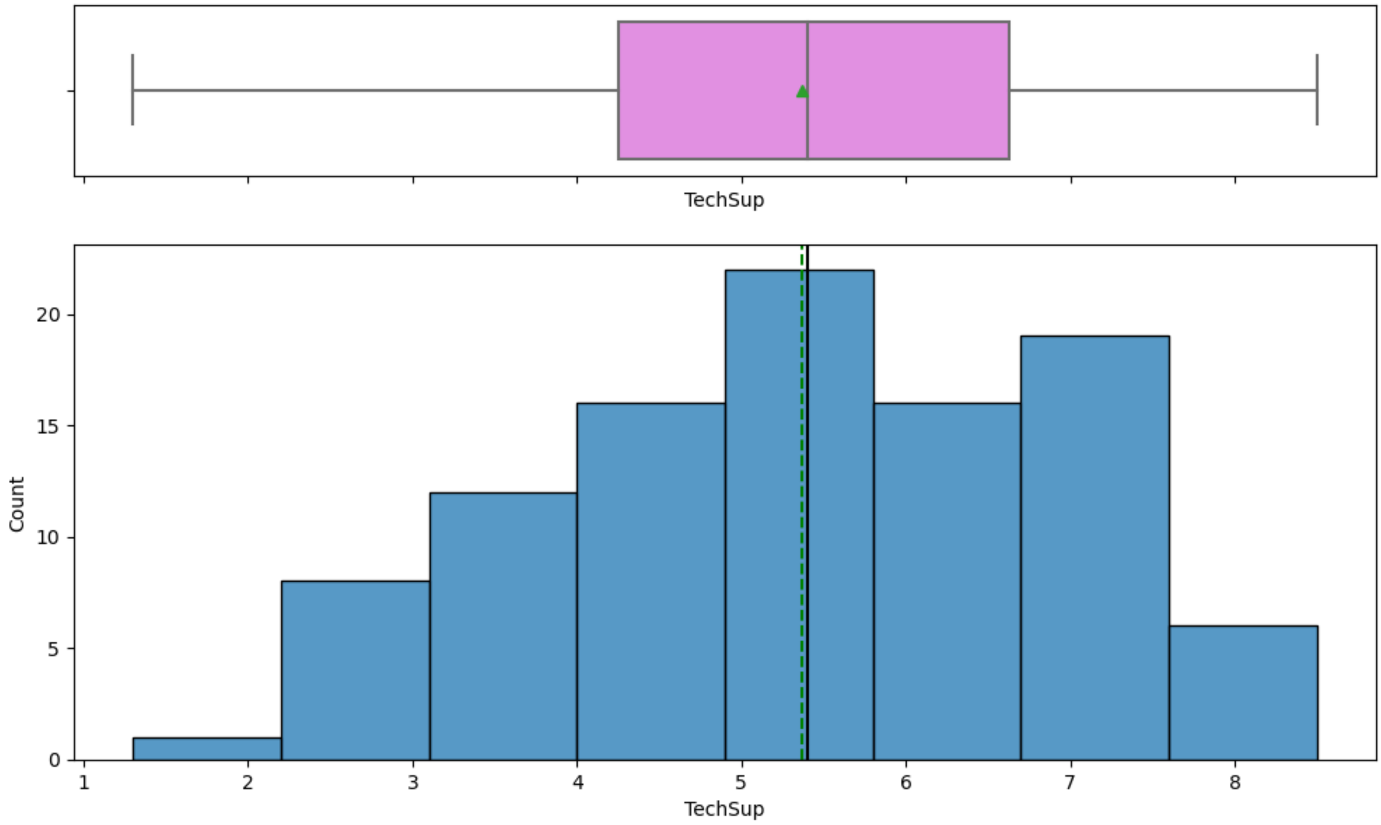
histogram and boxplot for ProdQual

`Ecom`



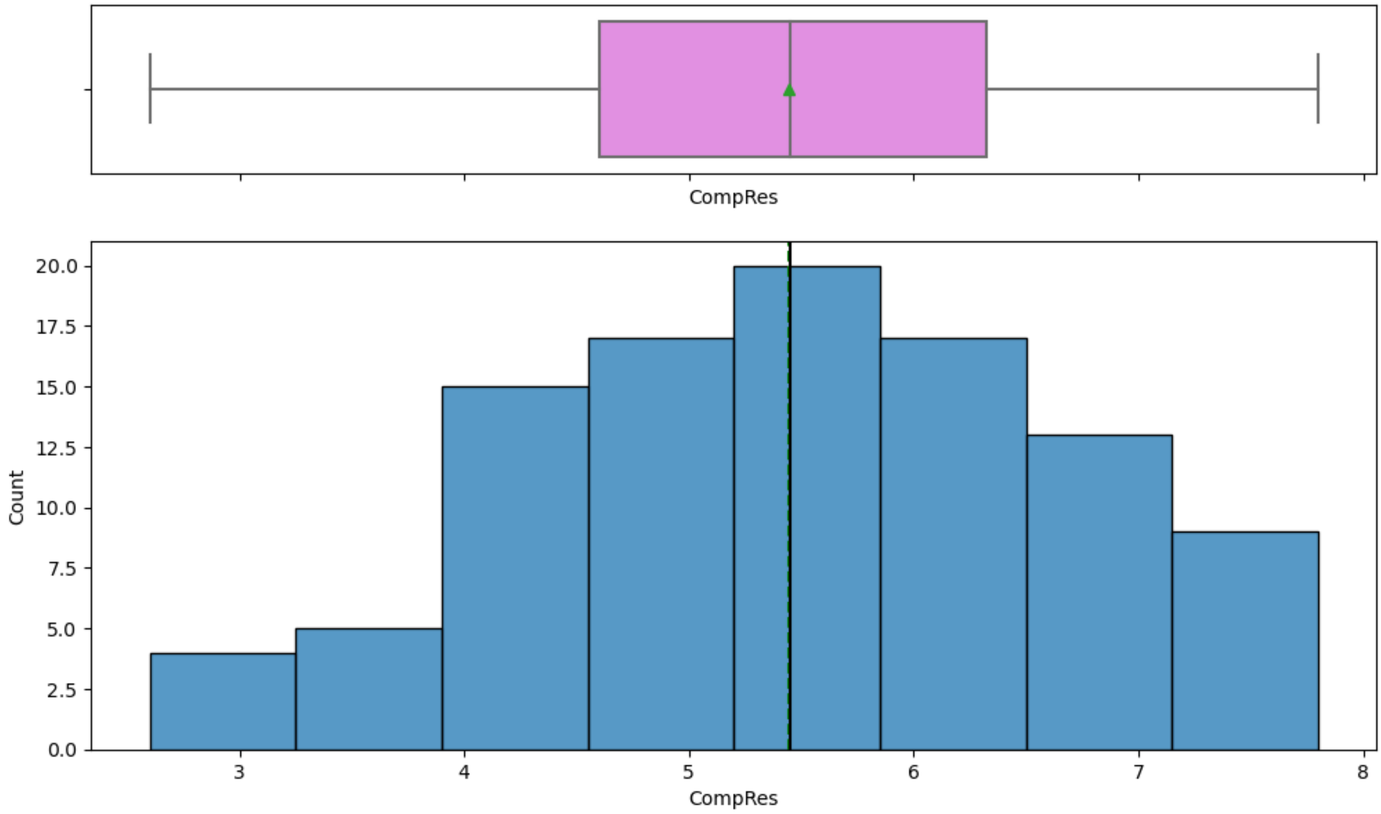
histogram and boxplot for Ecom

`TechSup`



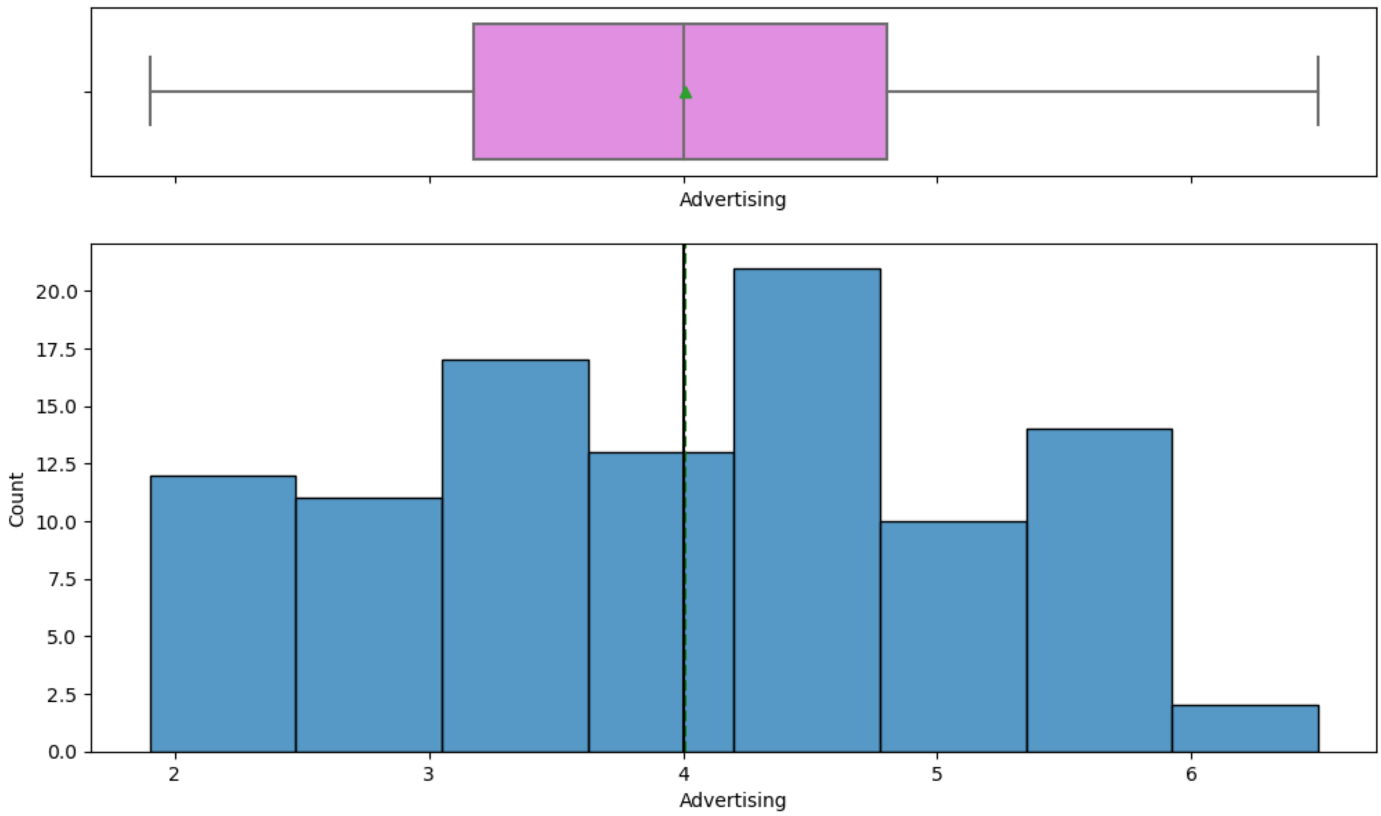
histogram and boxplot for TechSup

`CompRes`



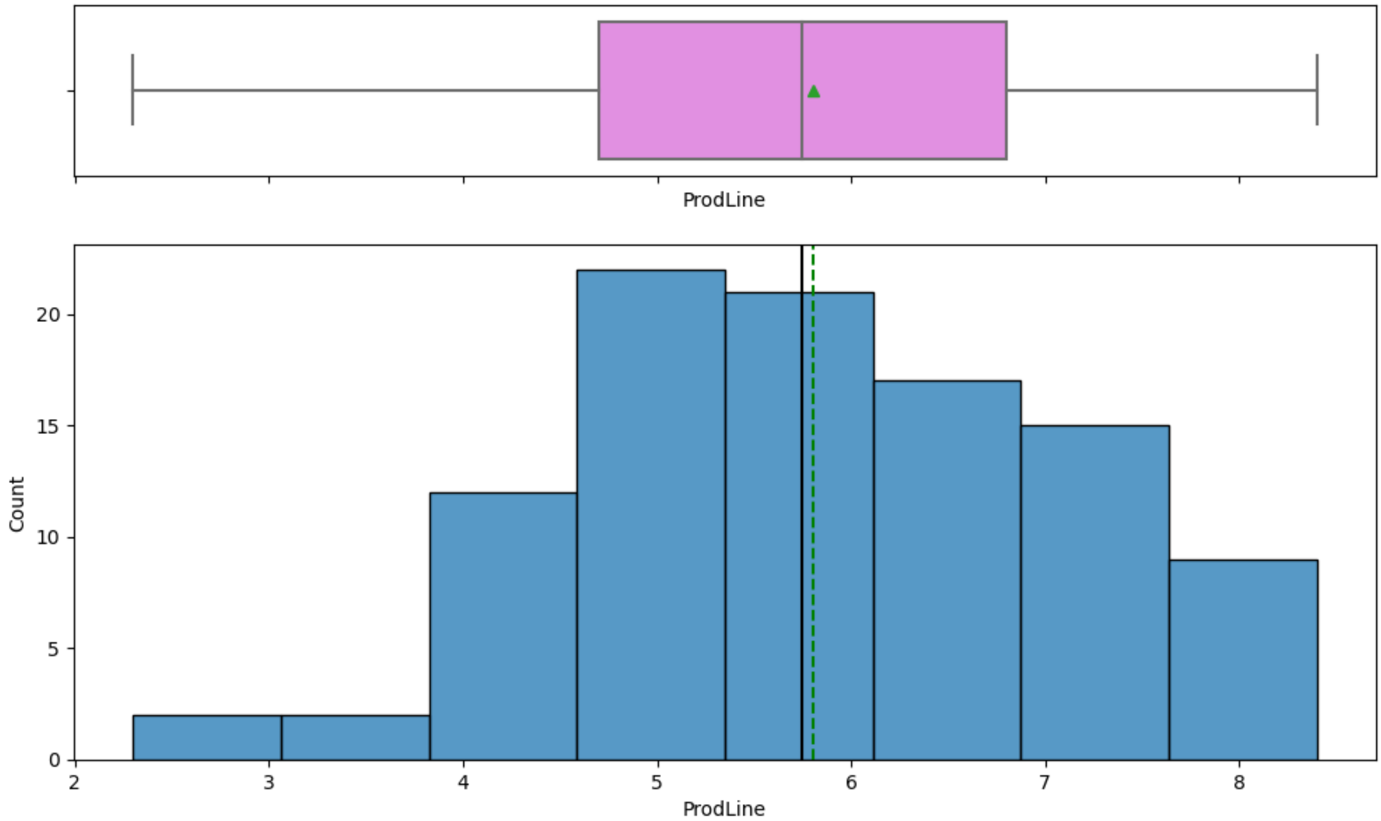
histogram and boxplot for CompRes

`Advertising`



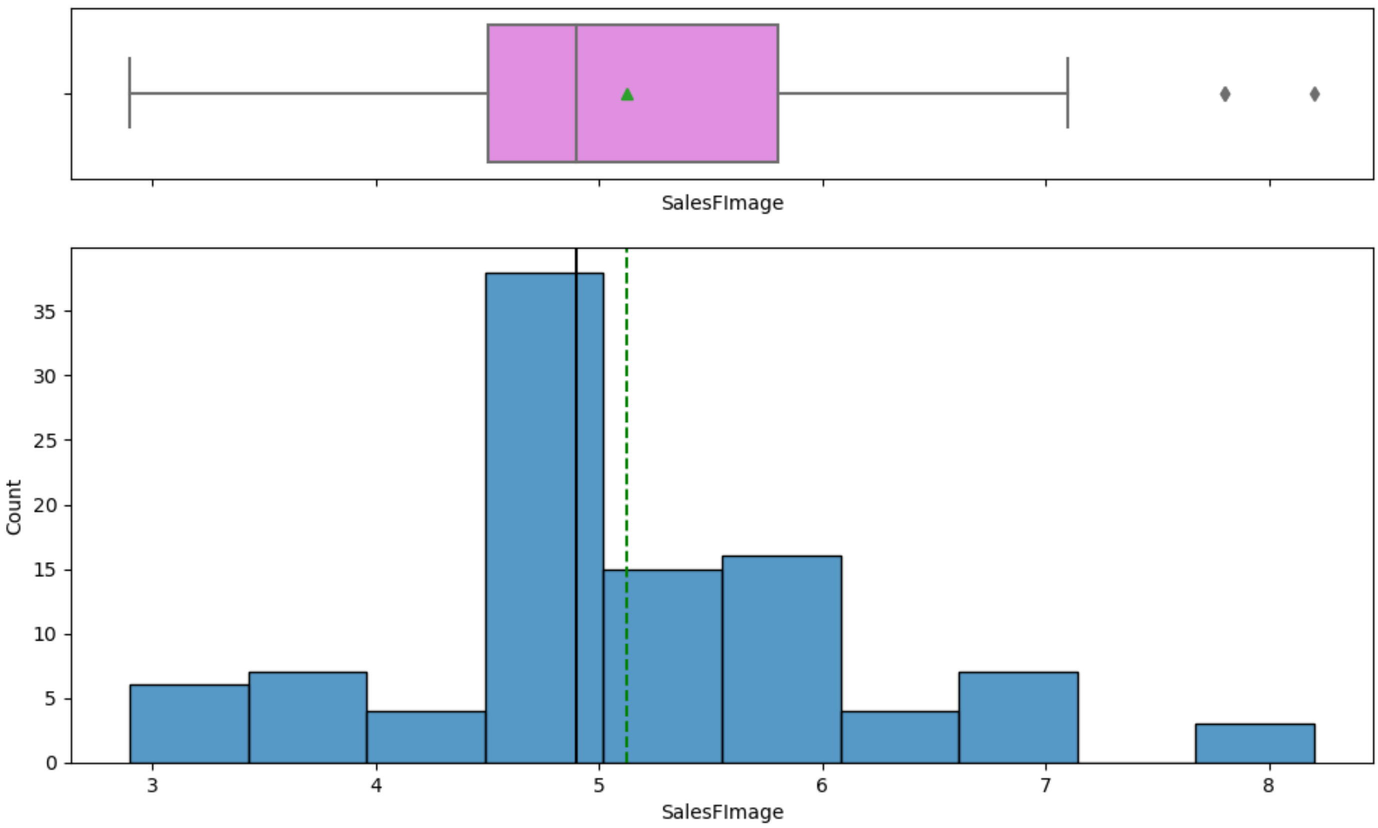
histogram and boxplot for Advertising

`ProdLine`



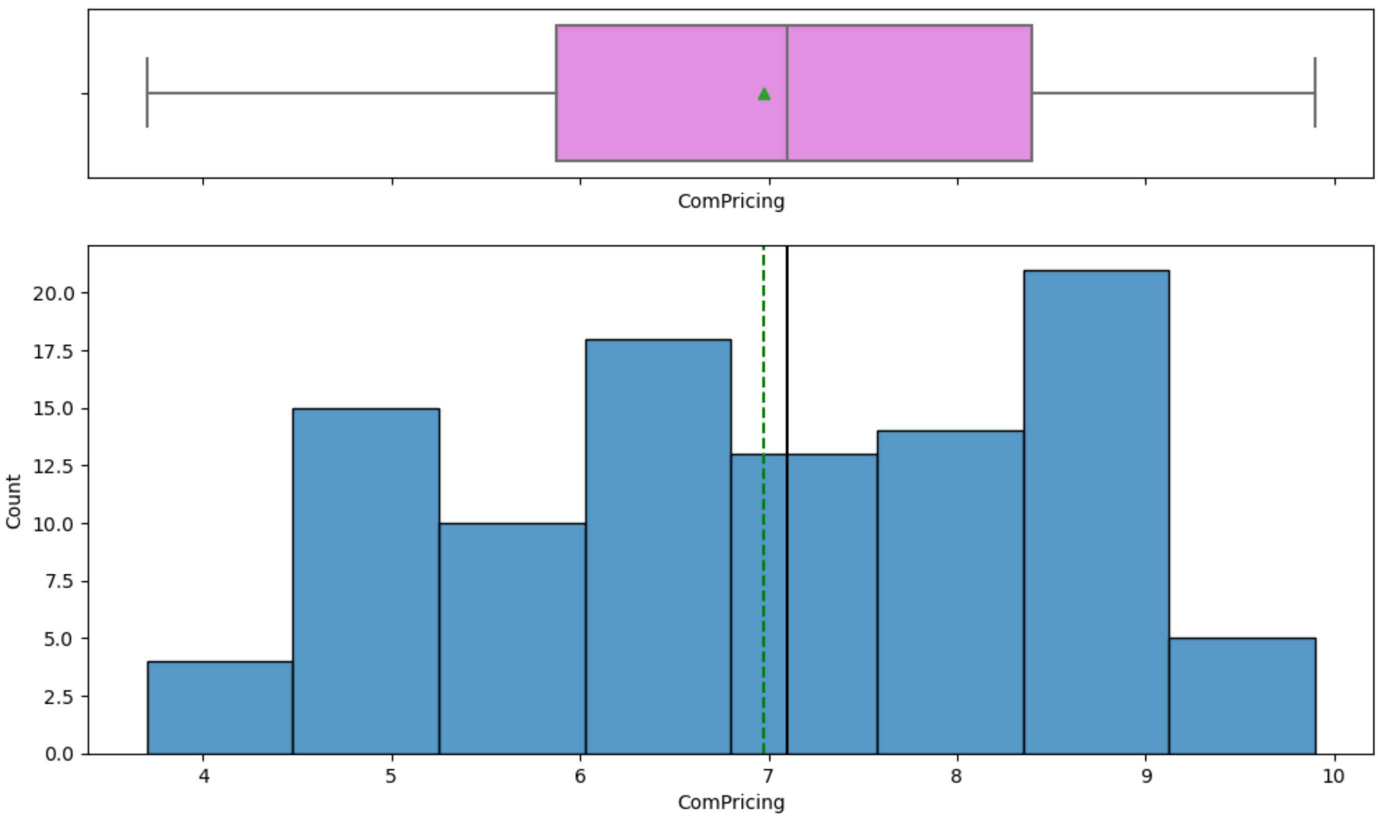
histogram and boxplot for ProdLine

`SalesFImage`



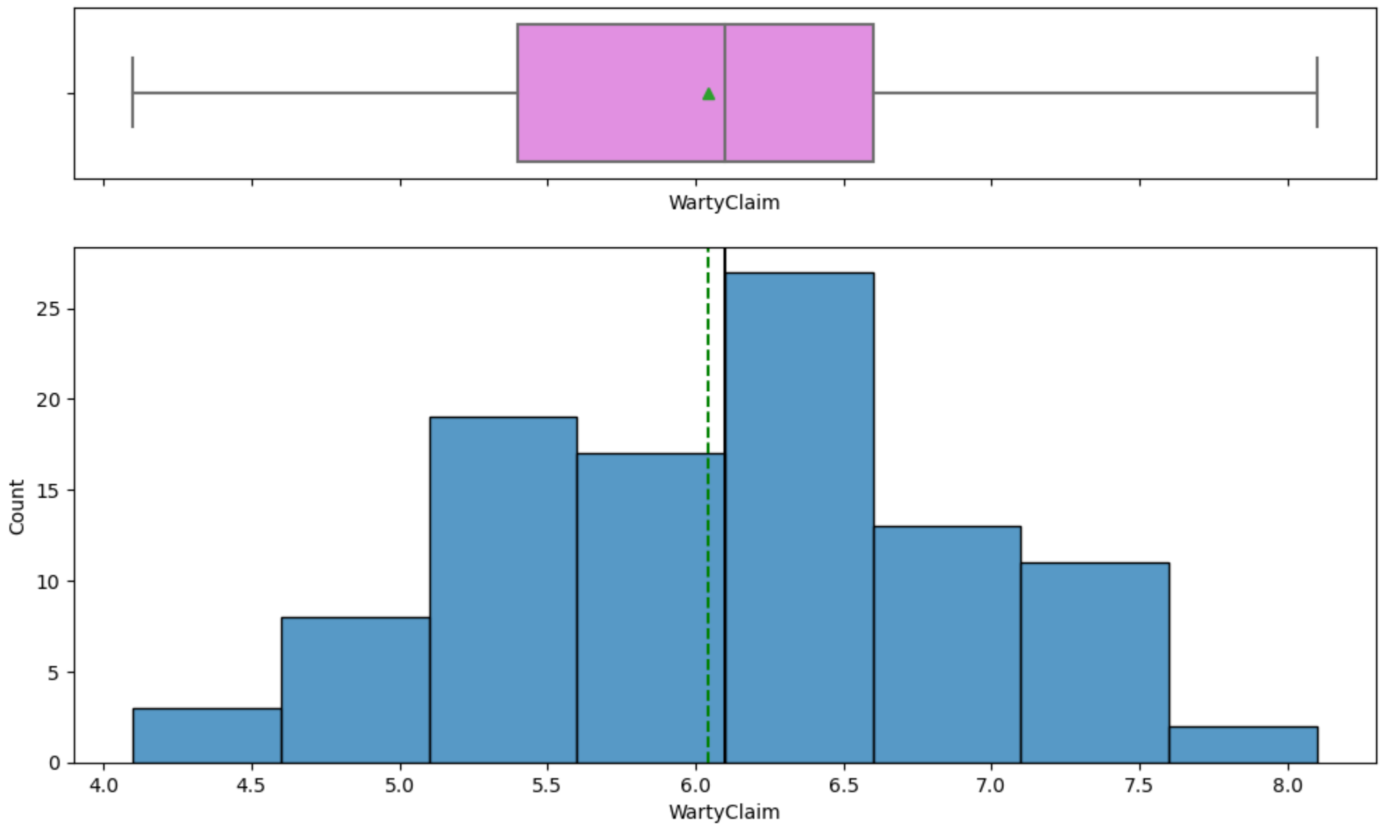
histogram and boxplot for SalesFImage

`ComPricing`



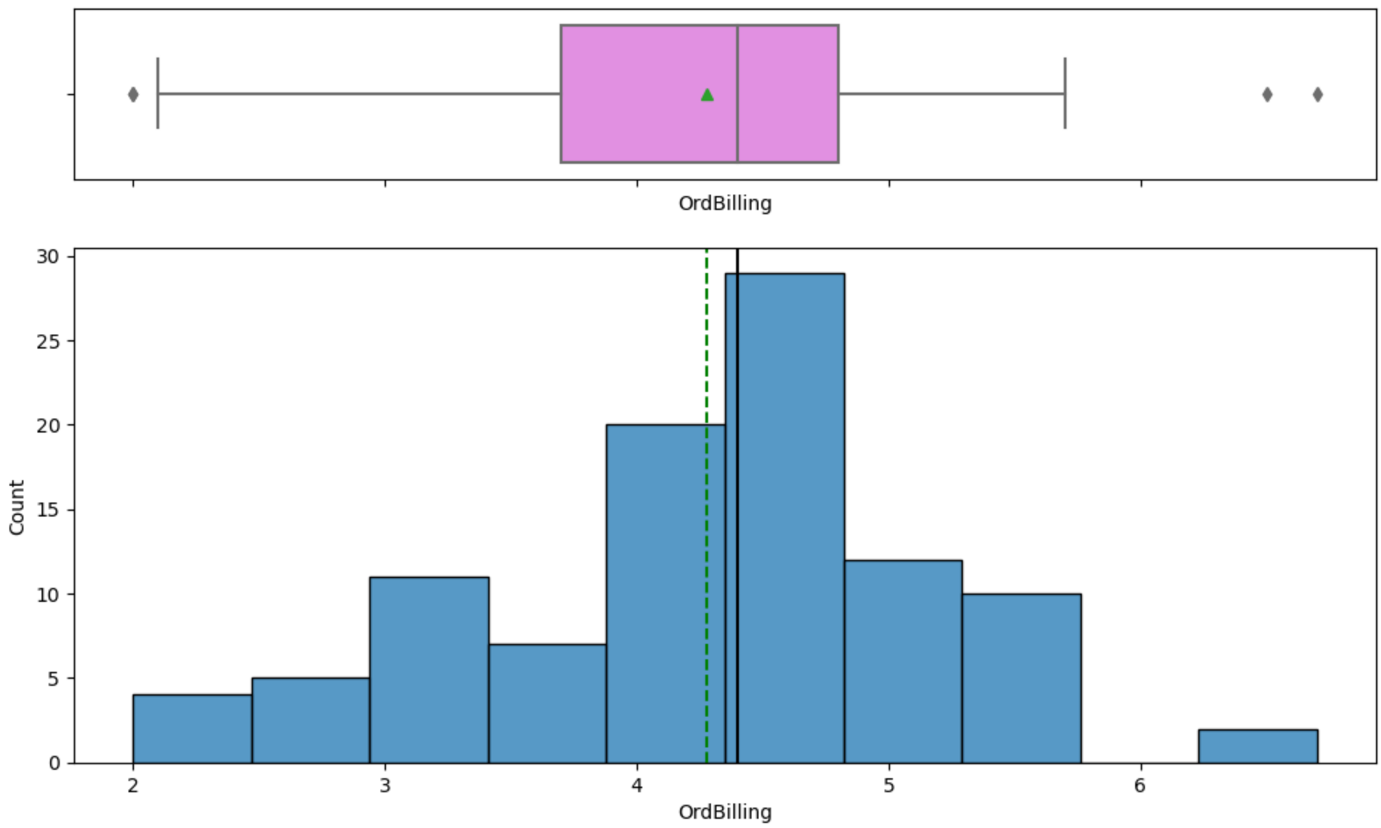
histogram and boxplot for ComPricing

`WartyClaim`



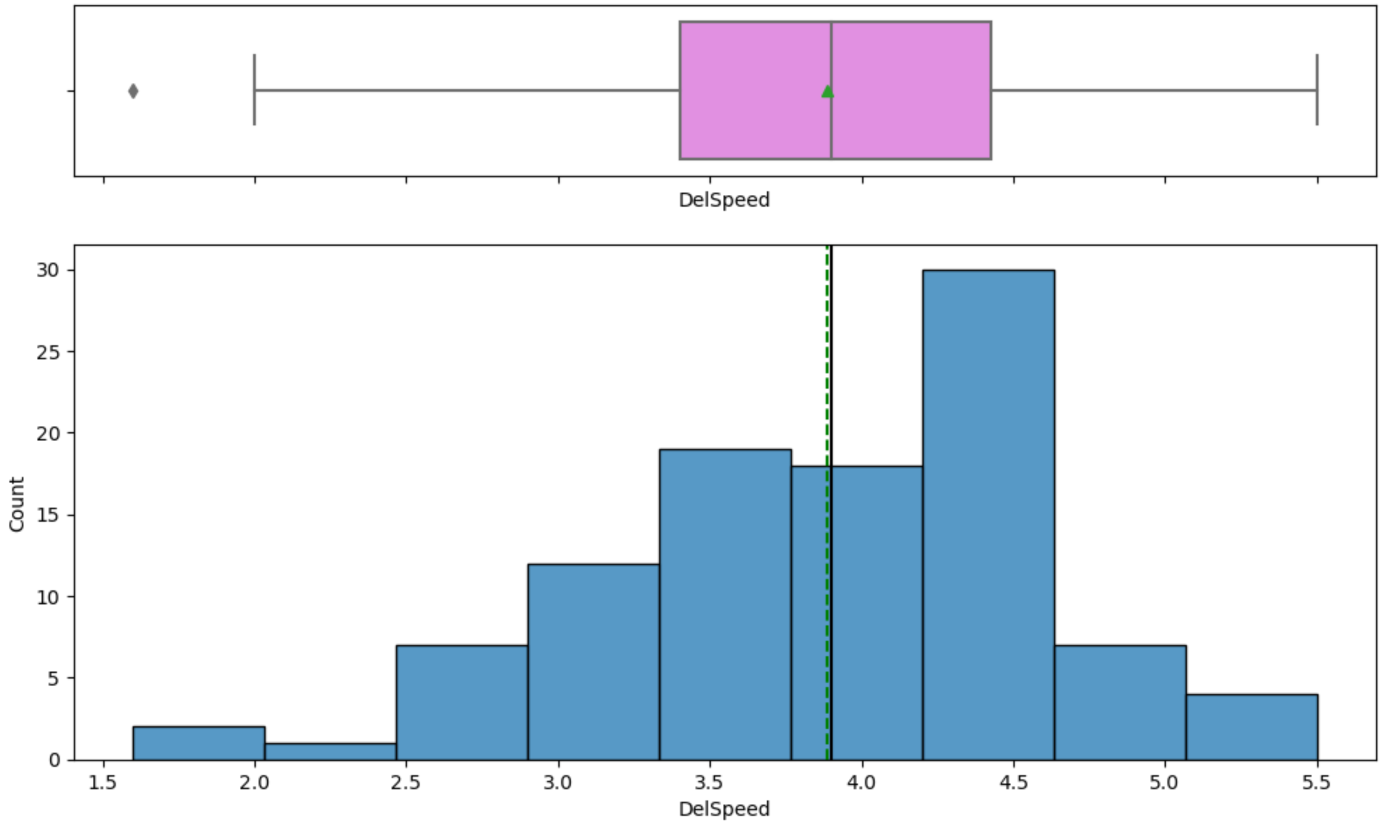
histogram and boxplot for WartyClaim

`OrdBilling`



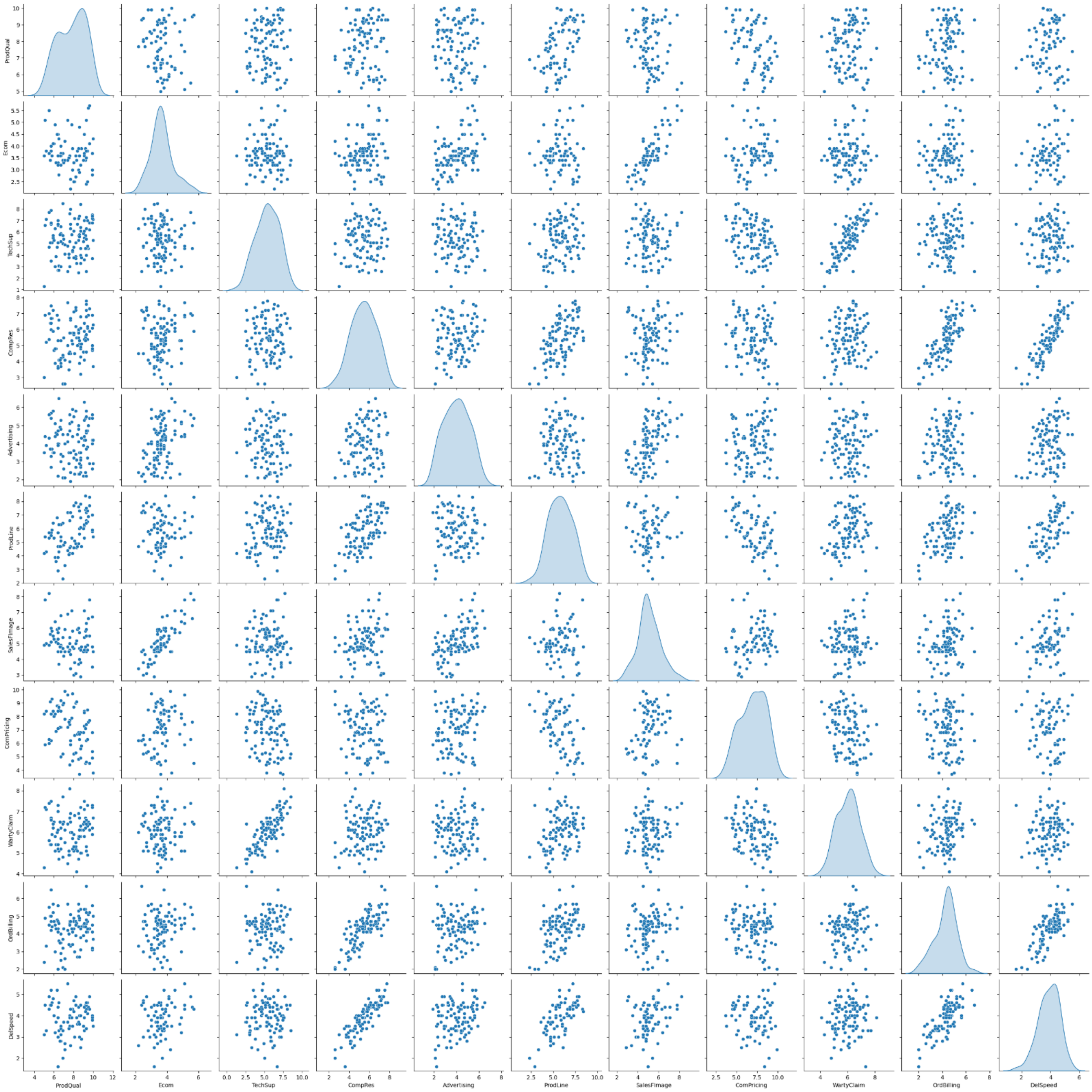
histogram and boxplot for OrdBilling

`DelSpeed`



histogram and boxplot for DelSpeed

- Bivariate Analysis



Observation

Based on the bivariate analysis plot, we can observe the following:

There is a positive correlation between ProdQual and Satisfaction, meaning that higher product quality is associated with higher customer satisfaction.

There is a negative correlation between Ecom and TechSup, suggesting that salons with a stronger e-commerce presence may have less need for technical support.

There is a positive correlation between CompRes and Satisfaction, indicating that effective complaint resolution is associated with higher customer satisfaction.

There is a negative correlation between WartyClaim and Satisfaction, suggesting that a higher frequency of warranty and claims issues is associated with lower customer satisfaction.

There is a positive correlation between DelSpeed and Satisfaction, meaning that faster delivery speed is associated with higher customer satisfaction.

These observations can help inform business decisions related to market segmentation and customer satisfaction.

9. Data Preprocessing(Outlier Detection , Scaling)

Scaling the data before clustering and creating a data frame of the scaled data.

10. Checking the outliers before and after scaling

11. Build the covariance matrix, eigenvalues and eigenvector.

Step 1- Create the covariance Matrix

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
ProdQual	1.01	-0.14	0.10	0.11	-0.05	0.48	-0.15	-0.41	0.09	0.11	0.03
Ecom	-0.14	1.01	0.00	0.14	0.43	-0.05	0.80	0.23	0.05	0.16	0.19
TechSup	0.10	0.00	1.01	0.10	-0.06	0.19	0.02	-0.27	0.81	0.08	0.03
CompRes	0.11	0.14	0.10	1.01	0.20	0.57	0.23	-0.13	0.14	0.76	0.87
Advertising	-0.05	0.43	-0.06	0.20	1.01	-0.01	0.55	0.14	0.01	0.19	0.28
ProdLine	0.48	-0.05	0.19	0.57	-0.01	1.01	-0.06	-0.50	0.28	0.43	0.61
SalesFImage	-0.15	0.80	0.02	0.23	0.55	-0.06	1.01	0.27	0.11	0.20	0.27
ComPricing	-0.41	0.23	-0.27	-0.13	0.14	-0.50	0.27	1.01	-0.25	-0.12	-0.07
WartyClaim	0.09	0.05	0.81	0.14	0.01	0.28	0.11	-0.25	1.01	0.20	0.11
OrdBilling	0.11	0.16	0.08	0.76	0.19	0.43	0.20	-0.12	0.20	1.01	0.76
DelSpeed	0.03	0.19	0.03	0.87	0.28	0.61	0.27	-0.07	0.11	0.76	1.01

Observation

There is a moderate positive correlation between ProdQual and ProdLine, indicating that higher product quality is associated with a broader product line.

There is a strong positive correlation between SalesFImage and Ecom, suggesting that a better salesforce image is associated with a stronger e-commerce presence.

There is a moderate negative correlation between ComPricing and WartyClaim, indicating that more competitive pricing is associated with fewer warranty and claims issues.

There is a strong positive correlation between CompRes and Satisfaction, suggesting that effective complaint resolution is associated with higher customer satisfaction.

There is a moderate positive correlation between DelSpeed and Satisfaction, indicating that faster delivery speed is associated with higher customer satisfaction.

Step 2- Get eigen values and eigen vector

apply PCA components

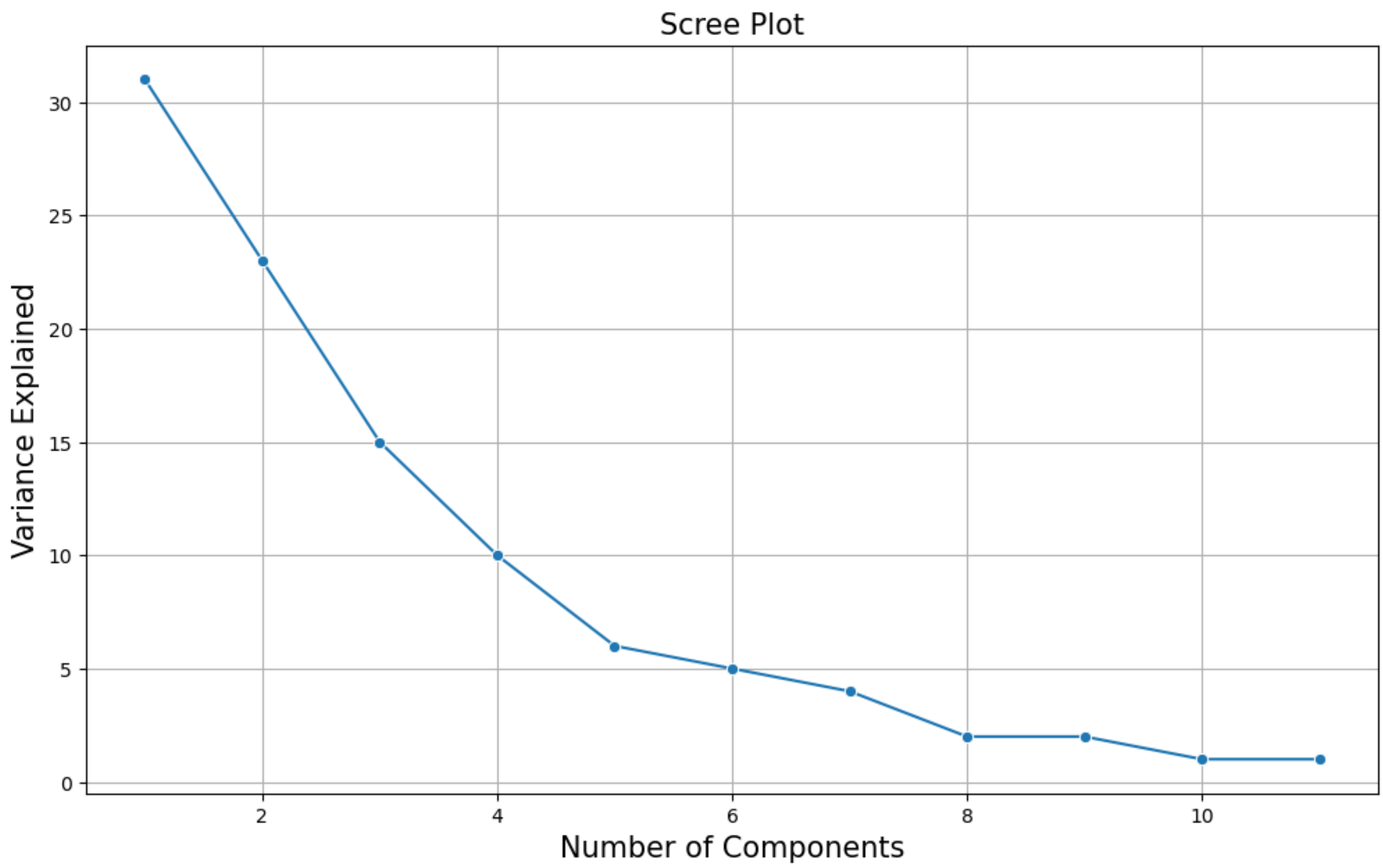
```
Eigenvectors: [[-0.13 -0.17 -0.16 -0.47 -0.18 -0.39 -0.2   0.15 -0.21 -0.44 -0.47]
 [-0.31  0.45 -0.23  0.02  0.36 -0.28  0.47  0.41 -0.19  0.03  0.07]
 [ 0.06 -0.24 -0.61  0.21 -0.09  0.12 -0.24  0.05 -0.6   0.17  0.23]
 [ 0.64  0.27 -0.19 -0.21  0.32  0.2   0.22 -0.33 -0.19 -0.24 -0.2 ]
 [ 0.23  0.42 -0.02  0.03 -0.8   0.12  0.2   0.25 -0.03  0.03 -0.04]
 [-0.56  0.26 -0.11 -0.03 -0.2   0.1   0.1  -0.71 -0.14 -0.12  0.03]
 [ 0.19  0.06 -0.02 -0.01 -0.06 -0.61  0.   -0.31 -0.03  0.66 -0.23]
 [ 0.14 -0.12  0.46  0.51 -0.05 -0.33  0.17 -0.1  -0.44 -0.37  0.07]
 [ 0.03 -0.54 -0.36  0.09 -0.15 -0.08  0.64 -0.09  0.32 -0.1  -0.02]
 [ 0.07  0.28 -0.39  0.53  0.04 -0.23 -0.35 -0.05  0.44 -0.3  -0.12]
 [ 0.18  0.06 -0.05 -0.36 -0.08 -0.39 -0.08 -0.1   0.13 -0.19  0.78]]
```

apply PCA Explained Variance

```
Eigenvalues: [3.46 2.58 1.71 1.1  0.62 0.56 0.41 0.25 0.21 0.13 0.1 ]
```

```
array([31., 23., 15., 10.,  6.,  5.,  4.,  2.,  2.,  1.,  1.])
```

Step 3 View Scree Plot to identify the number of components to be built



The first principal component explains the majority of the variance in the data, indicating that it captures the most important patterns or trends.

The second principal component explains a significant amount of variance, but less than the first component, indicating that it captures some additional information beyond the first component.

The amount of variance explained by subsequent components decreases rapidly, indicating that they capture less important patterns or trends in the data.

The scree plot can help us determine the number of principal components to retain for further analysis. A common rule of thumb is to retain components that explain at least 70-80% of the variance in the data. In this case, we may want to retain the first two or three components for further analysis.

Step 4 Apply PCA for the number of decided components to get the loadings and component output

apply the appropriate PCA components from the above plot and Component output

```
array([[ 0.0795508 , -1.10096634, -2.19706653,  1.56293289,  0.76757039,
         2.90862177,  5.29319132,  1.47659077, -0.61394761, -0.42366008,
         0.57625231,  1.86757037, -2.66029481, -1.15437973, -1.98252867,
        -1.19534642, -0.6292106 , -1.94912563, -0.44065433, -1.18679105,
         1.32903312, -3.07501457, -1.22862294, -1.9521566 ,  0.27711435,
        -0.33863508, -1.7647172 ,  1.0351017 , -1.29479142,  1.66262708,
         0.14460849,  2.1146274 ,  1.06970726, -0.10940809,  1.80536022,
         2.1224852 ,  0.62205111, -2.57411754, -0.34496698,  1.21539637,
         1.25487586,  0.16793078, -1.82095895, -2.33269255, -0.87328888,
        -1.65347263, -1.03184895, -3.12115015, -1.07236222, -0.79767523,
         1.98945764, -1.15121892, -1.07789784,  1.16797753,  0.52852266,
        -0.3768932 , -3.50943905, -1.28569623,  0.57436837, -1.10762365,
        -2.11898365,  1.19627748,  1.22409809,  2.78237859,  1.83048999,
        -1.25224924, -1.39907113,  0.0918897 ,  2.55191074,  1.05553043,
        -2.17896719,  2.17537864, -1.01608393, -1.29736966, -0.06380347,
        -0.1308607 , -0.20699358, -1.25207434, -3.21564763,  2.42613744,
         0.4328888 ,  0.03278748,  2.36190826,  4.37978758, -0.92811004,
         2.36712173,  4.23846478, -1.65771553, -0.30252456, -3.53671209,
        -0.06447556,  4.82862408, -1.0792803 , -2.81182077,  0.80752725,
        -0.27366812,  1.60818602,  3.19577568, -0.62088819,  1.63523181],
       [ 1.54319843, -2.42029823, -0.72744044,  0.17136647, -1.42811141,
         0.30938664,  1.05748117,  1.11108334,  1.37947271,  1.98154125,
        -1.2616166 ,  2.39193283,  1.86358088, -1.52960009,  1.92826246,
```

The first principal component (PC1) can be expressed as a linear combination of the original variables, where the weights are given by the corresponding elements of the first eigenvector.

Let's denote the eigenvector corresponding to the largest eigenvalue as v = [v1, v2, ..., vp], where p is the number of variables. Then, the first PC can be expressed as:

PC1 = v1ProdQual + v2Ecom + v3TechSup + v4CompRes + v5Advertising + v6ProdLine + v7SalesFImage + v8ComPricing + v9WartyClaim + v10OrdBilling + v11*DelSpeed

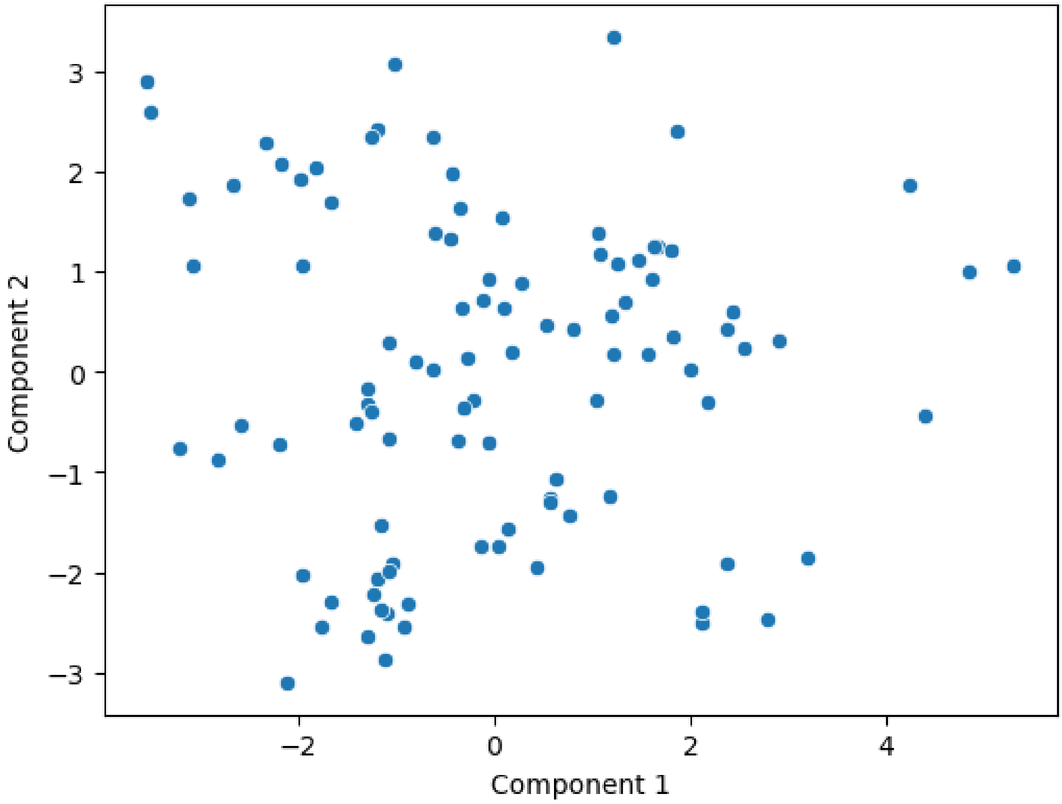
12. Write the explicit form of the first PC (in terms of Eigen Vectors)

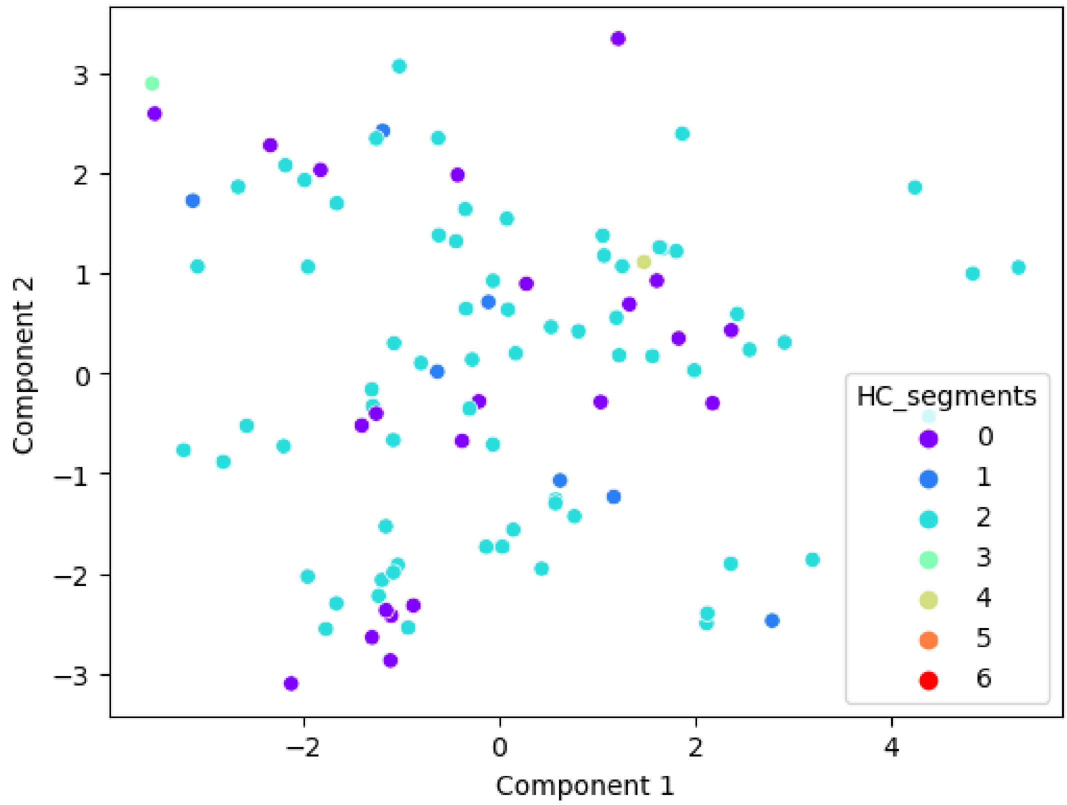
This means that the first PC is a weighted average of the original variables, where the weights are given by the eigenvector. The first PC captures the maximum variance in the data, and the weights indicate the relative importance of each variable in explaining this variance.

	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
PC0	-0.130000	-0.170000	-0.160000	-0.470000	-0.180000	-0.390000	-0.200000	0.150000	-0.210000	-0.440000	-0.470000
PC1	-0.310000	0.450000	-0.230000	0.020000	0.360000	-0.280000	0.470000	0.410000	-0.190000	0.030000	0.070000

storing results in a dataframe and checking the amount of variance explained

```
The first two principal components explain 54.34% of the variance in the data.
```





The first principal component captures a combination of variables related to product quality, technical support, complaint resolution, advertising, product line, salesforce image, and competitive pricing.

The second principal component captures a combination of variables related to warranty and claims, order and billing, and delivery speed.

The scatterplot can help us visualize the distribution of the data along the first two principal components and identify any clusters or patterns.

Explained variance ratio

```
Explained Variance Ratio: [0.31154285 0.2318997 ]
Principal Component 1 Loadings: [-0.13378962 -0.16595278 -0.15769263 -0.47068359 -0.18373495 -0.38676517
 -0.2036696   0.15168864 -0.21293363 -0.43721774 -0.47308914]
Principal Component 2 Loadings: [-0.31349802  0.44650918 -0.23096734  0.01944394  0.36366471 -0.28478056
  0.47069599  0.4134565  -0.19167191  0.02639905  0.07305172]
```

Cumulative explained variance ratio

```
Cumulative Explained Variance: [0.31154285 0.54344255]
```

Applying K-means clustering on original data and PCA-transformed data

13. Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?

The eigenvalues represent the amount of variance explained by each principal component. The cumulative sum of the eigenvalues indicates the total amount of variance explained by the first k principal components. By examining the cumulative sum of the eigenvalues, we can determine the optimum number of principal components to retain.

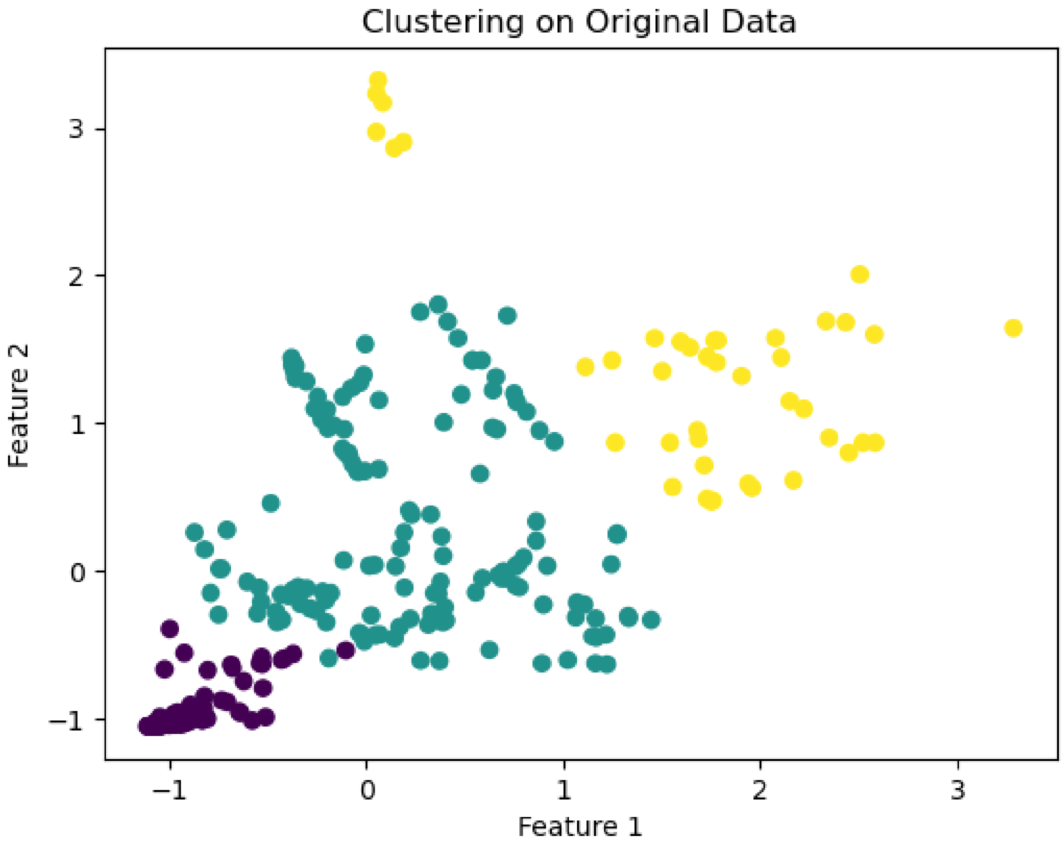
In this case, the first principal component has an eigenvalue of 3.46, which means it explains 3.46 units of variance. The first two principal components have a cumulative eigenvalue of 3.46 + 2.58 = 6.04, which means they explain 6.04 units of variance in total. As we continue to add more principal components, the cumulative eigenvalue increases, but at a decreasing rate.

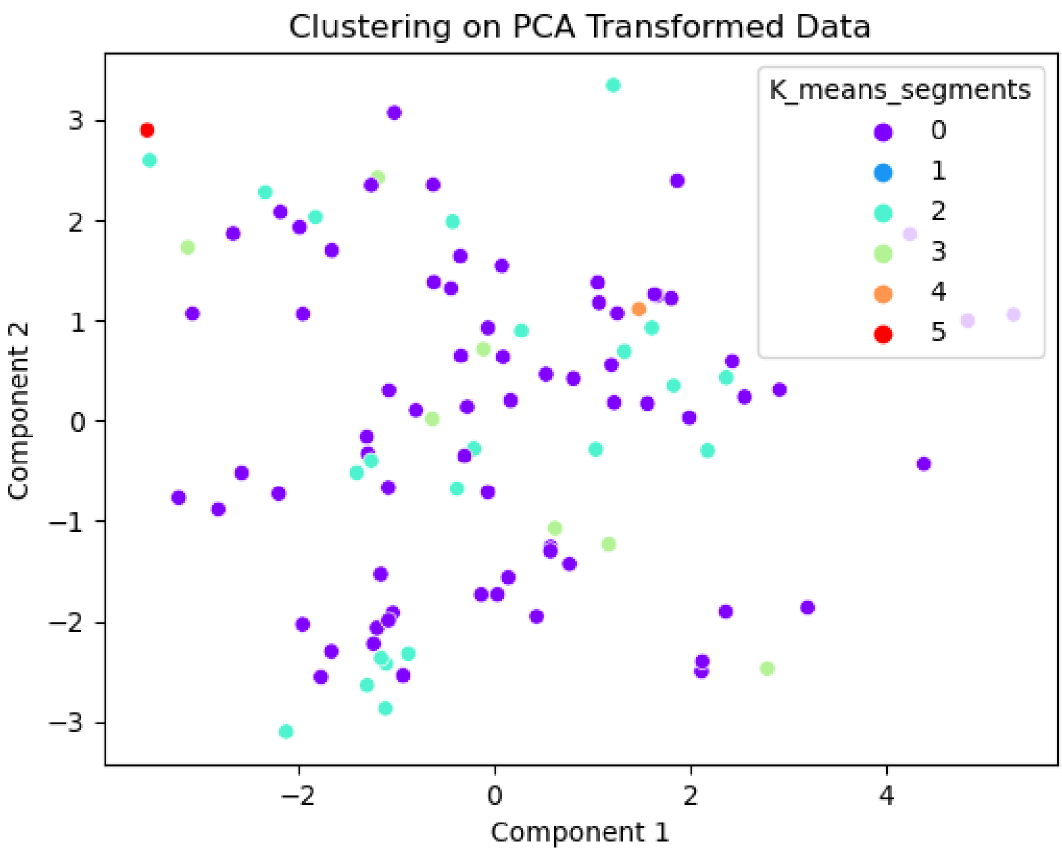
To decide on the optimum number of principal components, we can use a scree plot, which is a plot of the eigenvalues against the principal component number. The scree plot helps us to identify the "elbow" point in the plot, which is the point where the eigenvalues start to level off. In this case, the elbow point appears to be around the third or fourth principal component.

The eigenvectors, on the other hand, represent the directions of the principal components in the original feature space. Each eigenvector has a corresponding eigenvalue that indicates the amount of variance explained by that principal component. The eigenvectors are used to transform the original data into a new coordinate system, where the first principal component is the direction that explains the most variance, the second principal component is the direction that explains the second most variance, and so on.

In summary, the cumulative values of the eigenvalues help us to decide on the optimum number of principal components to retain, while the eigenvectors represent the directions of the principal components in the original feature space. By examining the scree plot and the cumulative eigenvalues, we can identify the elbow point and retain the principal components that explain the most variance in the data.

14. Perform PCA and export the data of the Principal Component scores into a data frame





15. Business implication of using the Principal Component Analysis

- The first principal component (PC1) has a strong positive loading for Advertising and SalesFImage, and a strong negative loading for ProdQual, TechSup, CompRes, ProdLine, WartyClaim, OrdBilling, and DelSpeed. This suggests that PC1 captures the variation in the data related to the quality and reliability of the products and services offered by the salon chain, as well as the effectiveness of their advertising and sales image.
- The second principal component (PC2) has a strong positive loading for Ecom and ComPricing, and a strong negative loading for ProdQual and PC1. This suggests that PC2 captures the variation in the data related to the salon chain's e-commerce capabilities and pricing strategy, as well as the quality of their products.
- The negative loading for ProdQual on both PC1 and PC2 suggests that there is a negative correlation between product quality and both e-commerce capabilities and pricing strategy. This could indicate that the salon chain may be sacrificing product quality in order to focus on e-commerce and competitive pricing.
- The negative loading for TechSup and CompRes on PC1 suggests that there is a negative correlation between technical support and complaint resolution and the overall quality and reliability of the products and services offered by the salon chain. This could indicate that the salon chain may need to improve their technical support and complaint resolution processes in order to enhance the overall customer experience.
- Overall, the PCA results suggest that the salon chain's product quality, advertising and sales image, e-commerce capabilities, and pricing strategy are the key drivers of variation in the data. The negative correlation between product quality and e-commerce capabilities and pricing strategy suggests that the salon chain may need to balance their focus on these areas in order to improve the overall customer experience.