**DATA MINING EXTENDED ASSIGNMENT**

**BY**

**SUDEEP KUMAR DAS**

**PGPDSBA.O. SEP22.B**

**TABLE OF CONTENT**

**LIST OF FIGURES**

**LIST OF TABLES**

**Problem A:**

**Problem Statement \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 05**

**Question 1: Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 05**

**Question 2: Scale the variables and write the inference for using the type of scaling function for this case study. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 14**

**Question 4: Check the dataset for outliers before and after scaling. Draw your inferences from this exercise. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 15**

**Question 5: Build the covariance matrix, eigenvalues and eigenvector. \_\_\_ 17**

**Question 6: Write the explicit form of the first PC. \_\_\_\_\_ 19**

**Question 7: 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? Perform PCA and export the data of the Principal Component scores into a data frame. \_\_\_\_\_\_\_\_ 19**

**Question 8: Mention the business implication of using the Principal Component Analysis for this case study. \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20**

**Problem B:**

**Problem Statement \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 21**

**Question 1: Clustering: Read the data and do exploratory data analysis. Describe the data briefly. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 22**

**Question 2: Do you think scaling is necessary for clustering in this case? Justify. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 26**

**Question 3: Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them. \_\_ 26**

**Question 4: Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and find the silhouette score. \_\_\_\_\_\_\_\_\_ 27**

**Question 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. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 30**

**LIST OF FIGURES**

**Figure 1: Univariate Analysis of Product Quality \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 08**

**Figure 2: Univariate Analysis of Ecom Variable \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 08**

**Figure 3: Univariate Analysis of Technical Support \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 09**

**Figure 4: Univariate analysis of Complain Resolution \_\_\_\_\_\_\_\_\_\_\_\_ 09**

**Figure 5: Univariate Analysis of Advertising \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 10**

**Figure 6: Univariate Analysis of Product Line \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 10**

**Figure 7: Univariate Analysis of Sales Force Image \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 11**

**Figure 8: Univariate Analysis of Competitive Pricing \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 11**

**Figure 9: Univariate Analysis of Warranty & Claim \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 12**

**Figure 10: Univariate Analysis of Order & Billing \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 12**

**Figure 11: Univariate Analysis of Delivery Speed \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 13**

**Figure 12: Multivariate Analysis of the variables \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 13**

**Figure 13: Outliers before Scaling \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 15**

**Figure 14: Outliers After Scaling \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 16**

**Figure 15: Heatmap of 4 PC \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20**

**Figure 16: Info of the dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_23**

**Figure 17: Distribution of Health Indices 1 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 24**

**Figure 18: Distribution of Health Indices 2 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 24**

**Figure 19: Distribution Of Per Capita Income \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 25**

**Figure 20: Distribution of GDP \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 25**

**Figure 21: Hierarchical Clustering using Dendrogram \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 27**

**Figure 22: Elbow Plot \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 28**

**Figure 23: Cluster Profiling \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 30**

**LIST OF TABLES**

**Table 1: Head and Tail of the Dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 05**

**Table 2: Info of the dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 06**

**Table 3: Descriptions of the dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 06**

**Table 4: Null Values check of the dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 07**

**Table 5: Corelation Matrix \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 14**

**Table 6: Output after Z-score Scaling \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 15**

**Table 7: Covariance Matrix \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 17**

**Table 8: Eigen Values \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 18**

**Table 9: Eigen Vectors \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 18**

**Table 10: First few rows of reduced dimension \_\_\_\_\_\_\_\_\_\_ 19**

**Table 11: Principal Component Score \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20**

**Table 12: First and Last Few rows of the Dataset \_\_\_\_\_\_\_\_\_ 22**

**Table 13: Description of the Dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 23**

**Table 14: First Few rows of the Scaled Data \_\_\_\_\_\_\_\_\_\_\_\_\_ 26**

**Table 15: Clustered Dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 27**

**Table 16: K-means clustering for 3 and 4 clusters \_\_\_\_\_\_\_\_\_ 29**

**Table 17: Distribution of data in 3 Clusters & 4 Clusters \_\_\_\_ 29**

**Table 18: Distribution of data using Hierarchical & K-Means Clustering \_\_ 29**

Problem A:

Problem Statement:

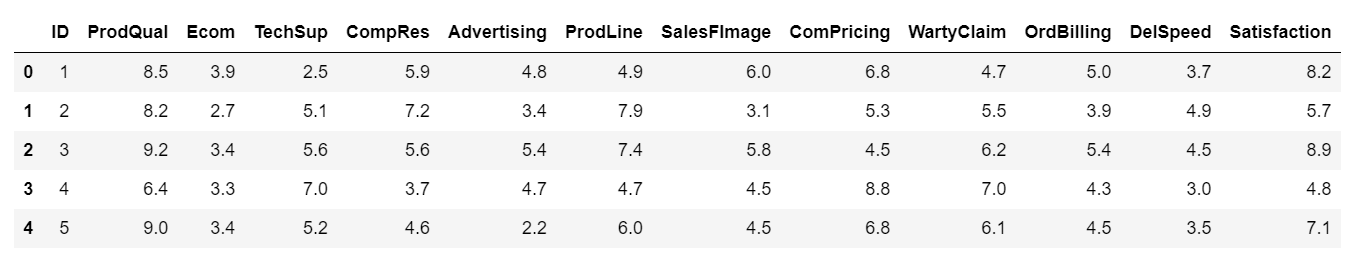
The [‘Hair Salon.csv’](https://olympus.mygreatlearning.com/courses/84255/files/7192422/download?wrap=1) dataset contains various variables used for the context of Market Segmentation. This particular case study is based on various parameters of a salon chain of hair products. You are expected to do Principal Component Analysis for this case study according to the instructions given in the rubric.

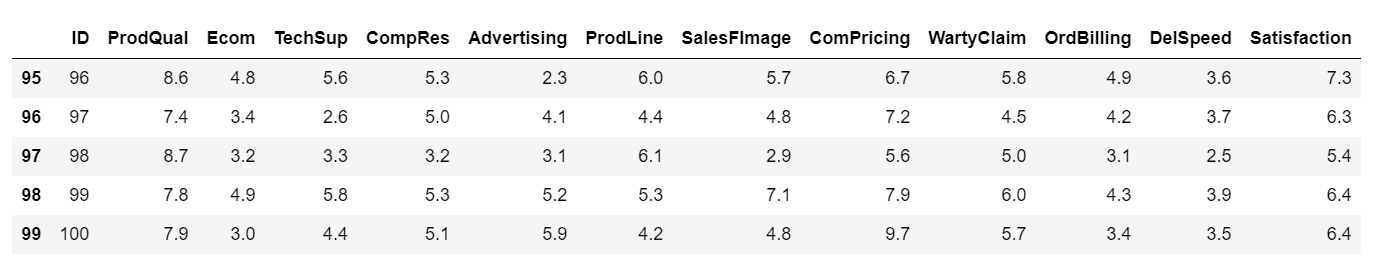
**Question 1:**

**Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented.**

**Answer 1:**

The dataset Hair Salon.csv contains various variables used for the context of Market Segmentation. We will first import the necessary libraries and then read the data set. The first and last few rows of the dataset are as follows:

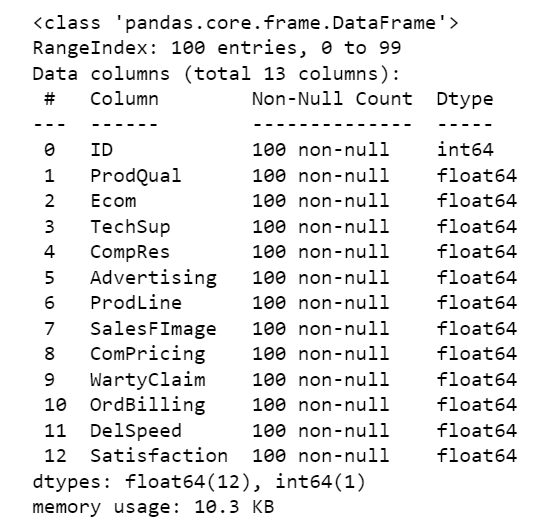




***Table 1: Head and Tail of the Dataset***

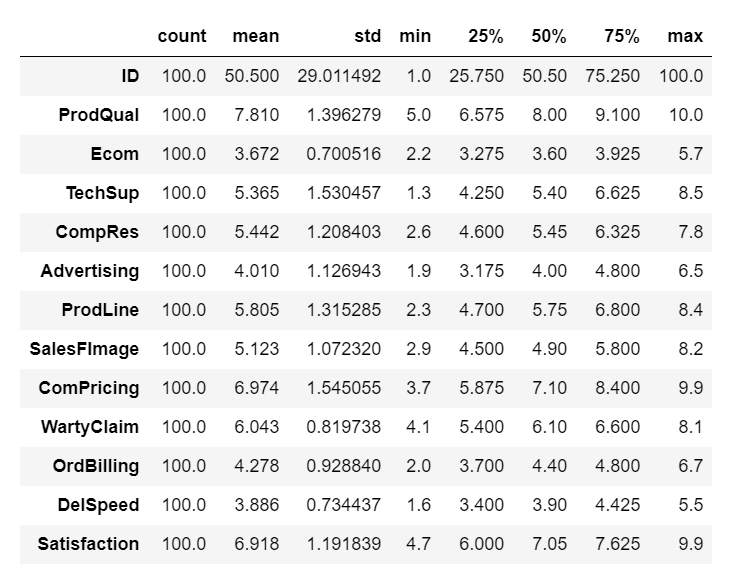
There are 100 rows and 13 columns in the dataset.

The info of the dataset is as follows:



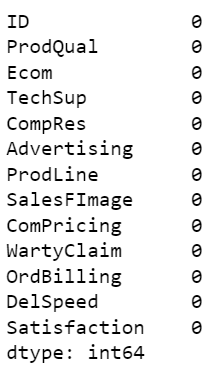
***Table 2: Info of the dataset***

Now let’s look at the basic descriptions of the dataset:



***Table 3: Descriptions of the dataset***

Null Values Check:

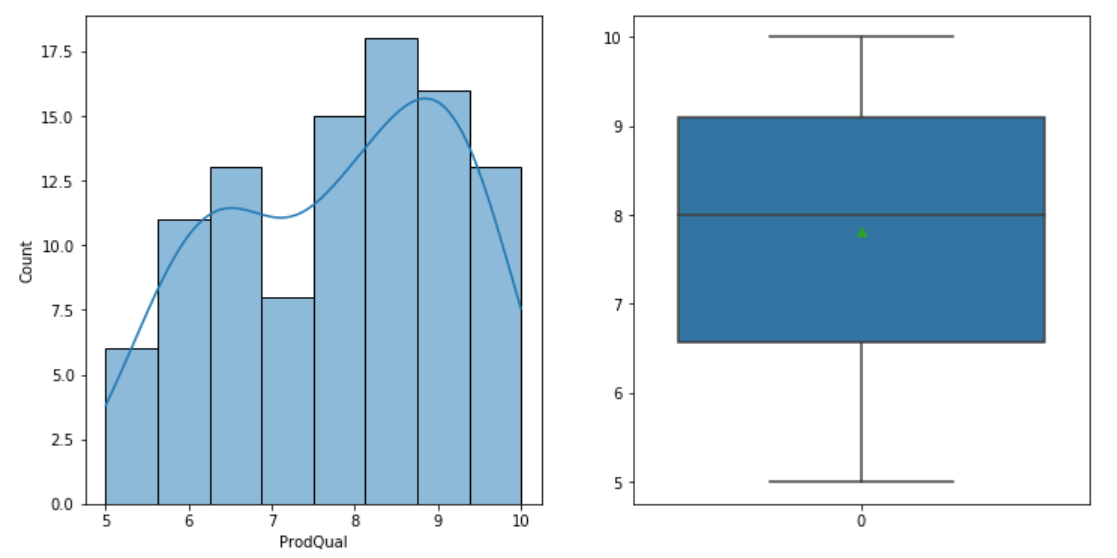


***Table 4: Null Values check of the dataset***

Observations:

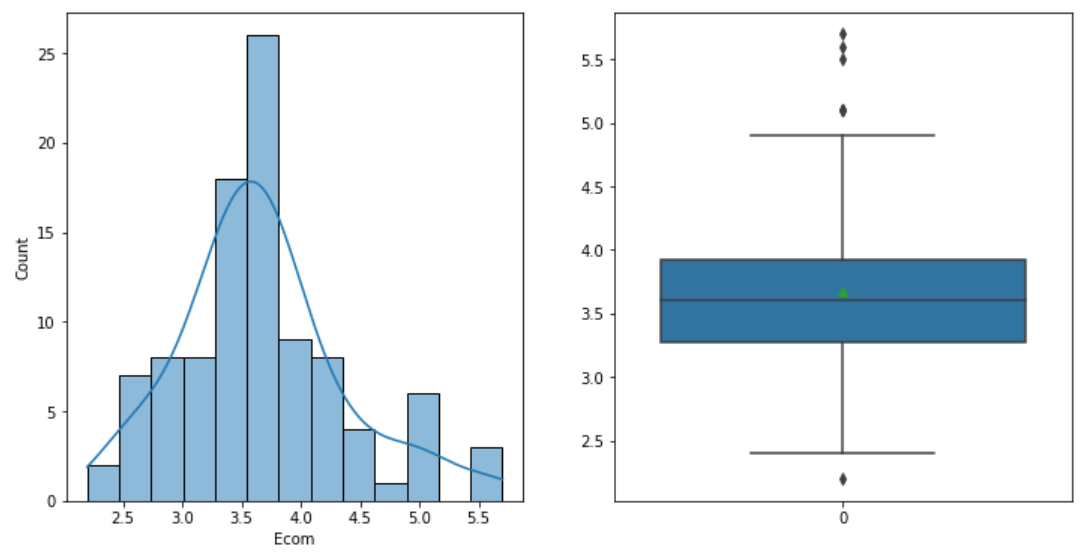
* The dataset consists of variables used for the context of market segmentation.
* There are 100 rows and 13 columns in the dataset.
* There are no missing values in the dataset.
* The dataset contains a variable named ‘Satisfaction’ which is the Target variable.
* The mean product quality is 7.8 with minimum being 5 and maximum being 10.
* All of the variables are continuous in nature and have float data type except ‘id’ variable which is int data type.

Exploratory Data Analysis:



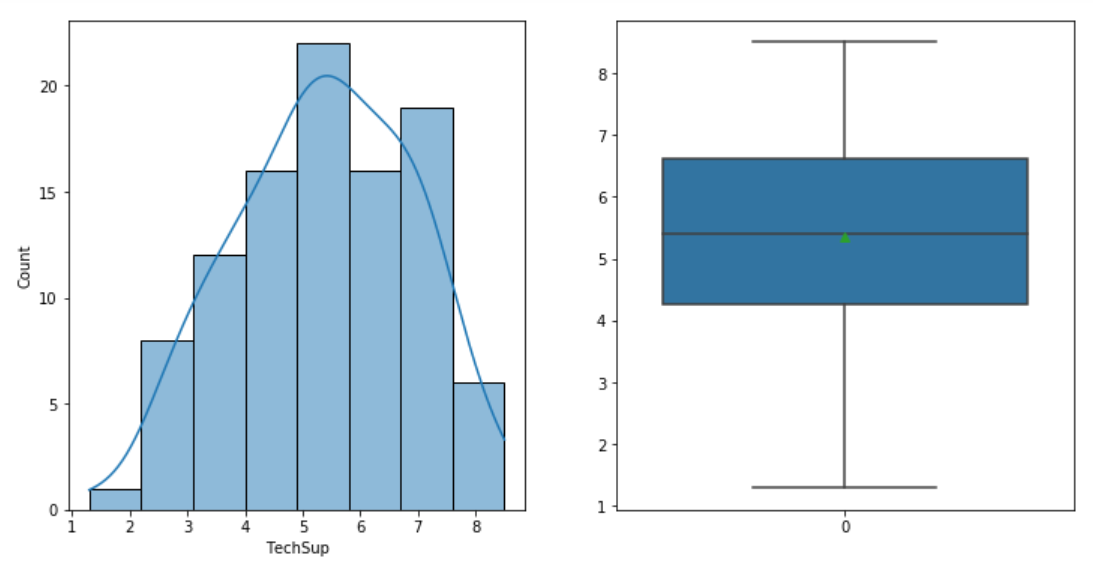
***Figure 1: Univariate Analysis of Product Quality***

The Product Quality is not normally distributed. The mean of the variable is 7.8, minimum being 5 and maximum is 10. There are no outliers present.



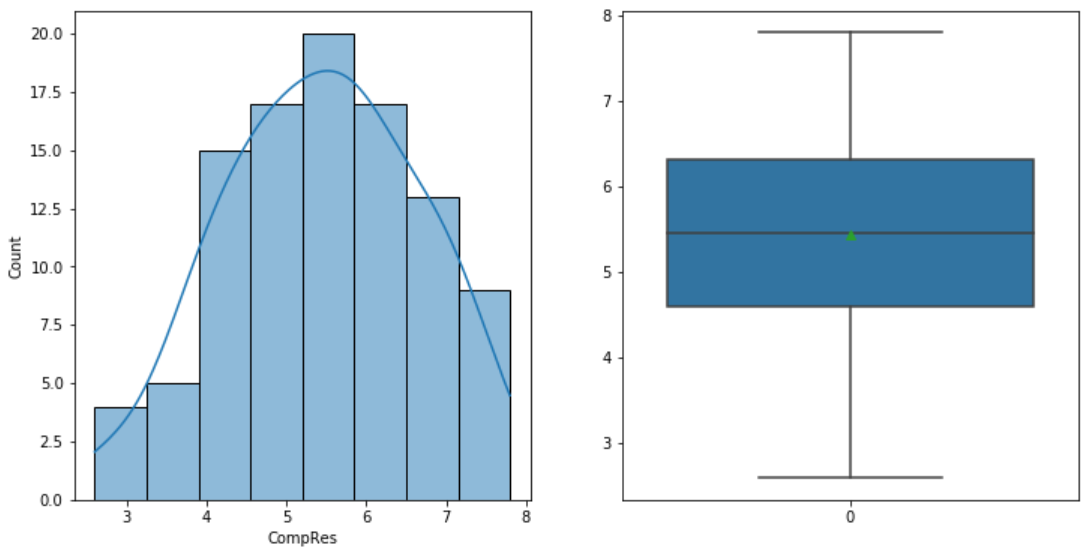
***Figure 2: Univariate Analysis of Ecom Variable***

The Ecom Variable seems to be having a normal distribution. There are some outliers present in the dataset which we will treat at a later stage. The mean is 3.67, minimum is 2.2 and maximum is 5.7.



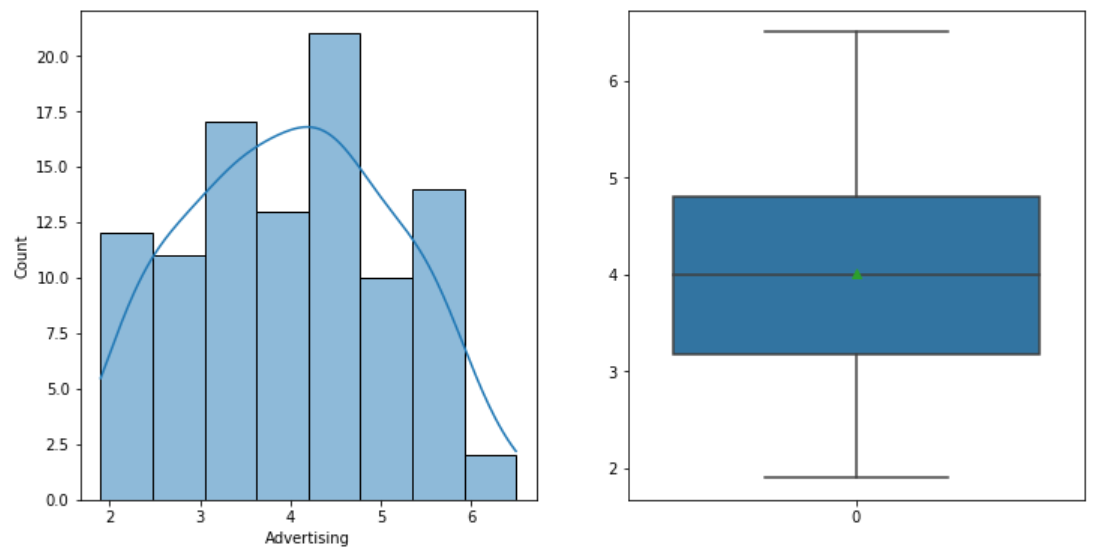
***Figure 3: Univariate Analysis of Technical Support***

The variable Technical Support is following a normal distribution. There are no outlier in the dataset and the mean is 5.36, minimum value is 1.3 and maximum value is 8.8.



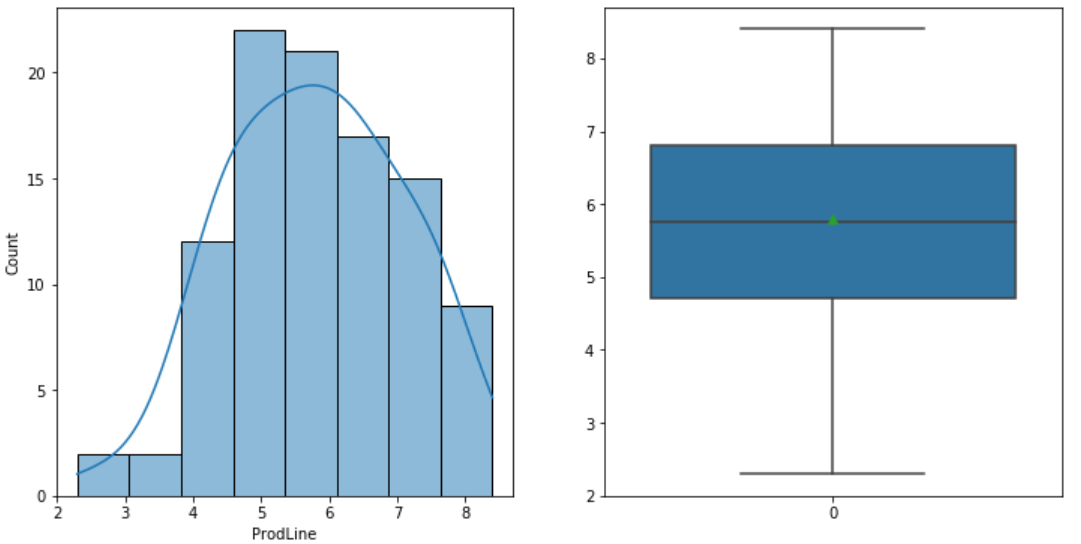
***Figure 4: Univariate analysis of Complain Resolution***

The variable Complain Resolution is more or less normally distributed. The mean is 5.4 with the minimum value being 2.6 and maximum value is 7.8.



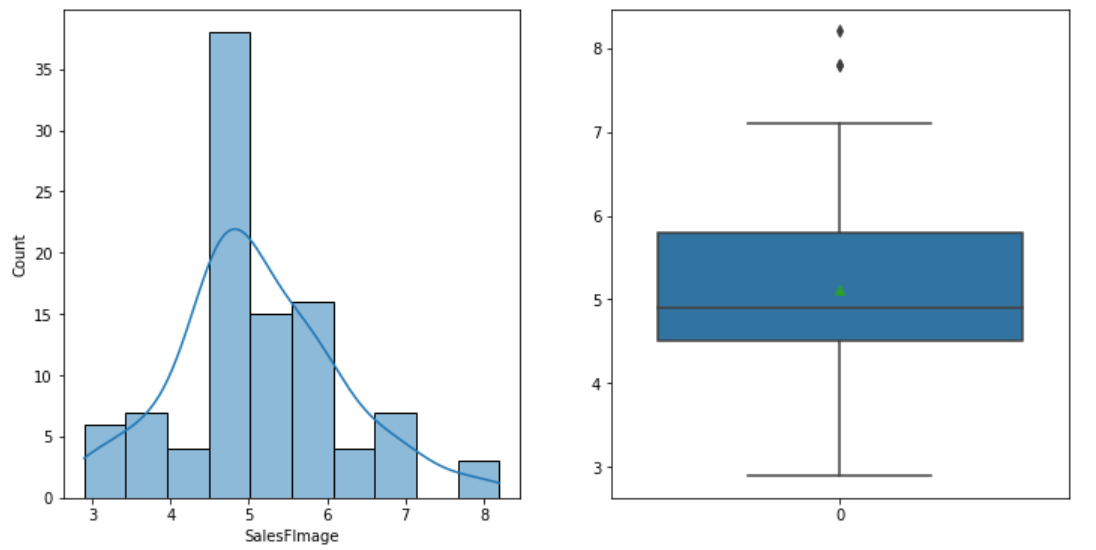
***Figure 5: Univariate Analysis of Advertising***

The variable Advertising is a little normally distributed. There are no outliers in this variable. The mean is 4 with maximum being 6.5 and minimum being 1.9.



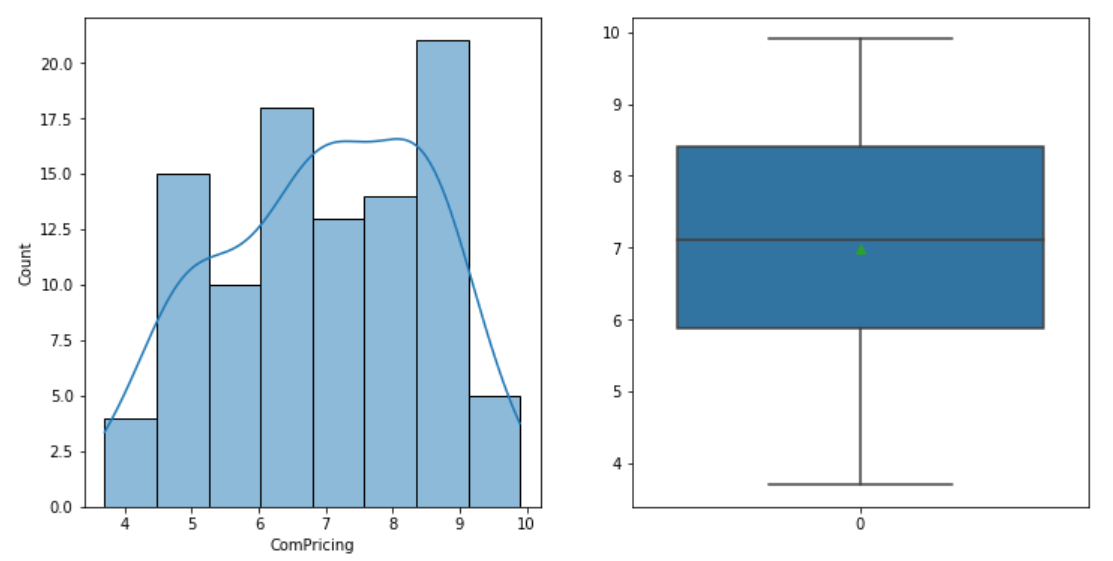
***Figure 6: Univariate Analysis of Product Line***

The variable product line has no outliers present. It is normally distributed with mean being 5.8, minimum being 2.3 and maximum being 8.4.



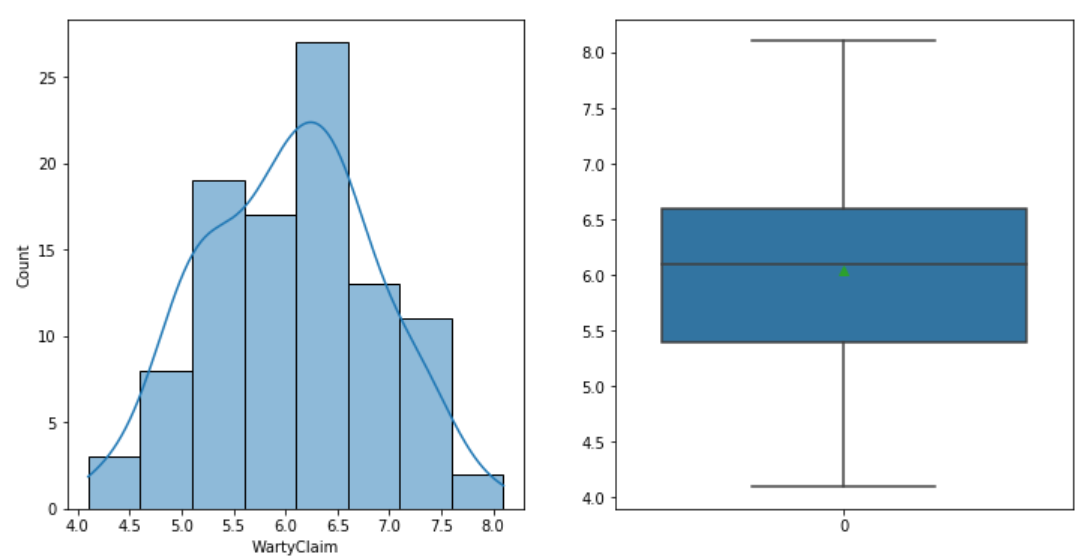
***Figure 7: Univariate Analysis of Sales Force Image***

The Variable Sales Force image is not normally distributed. There are outliers present in the variable. The mean is 5.1 minimum is 2.9 and the maximum 8.2.



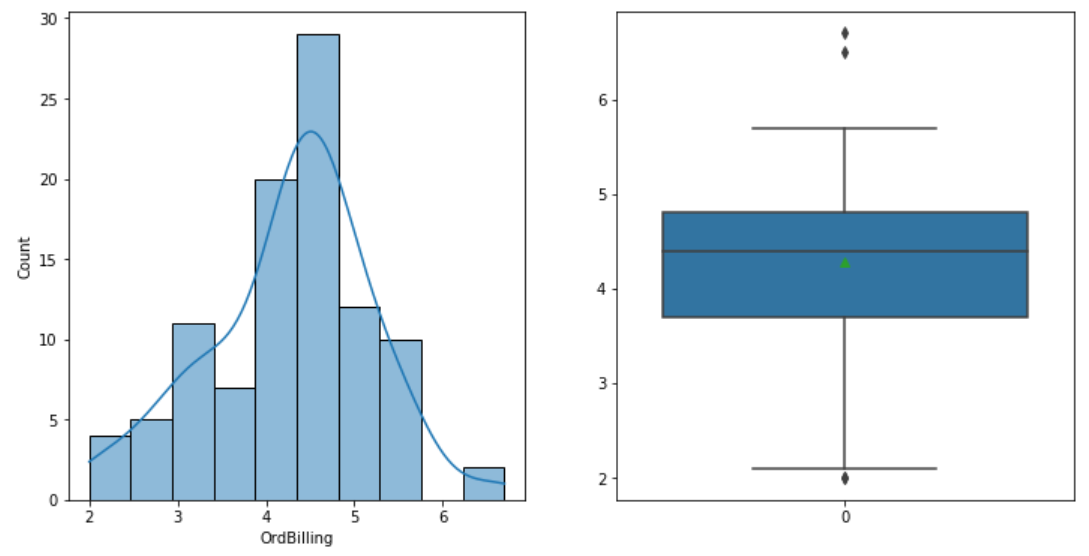
***Figure 8: Univariate Analysis of Competitive Pricing***

The variable Competitive Pricing is not normally distributed. There are no outliers present in the variable. The mean value is 6.9 the minimum value is 3.7 and the maximum value is 9.9.



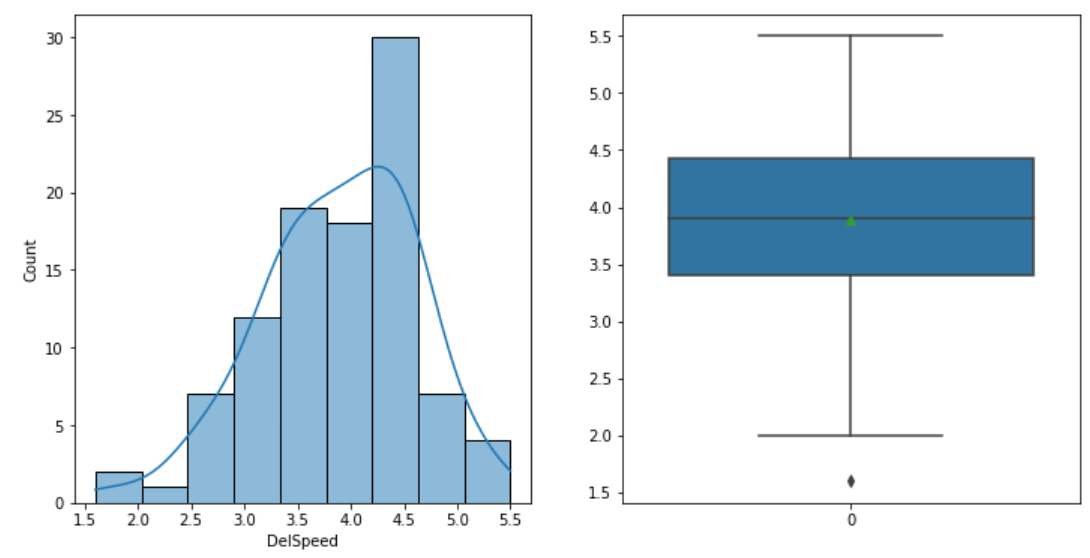
***Figure 9: Univariate Analysis of Warranty & Claim***

The variable Warranty & Claim is normally distributed. There are no outliers present in the variable. The mean is 6, the minimum value is 4.1 and the maximum value is 8.1.



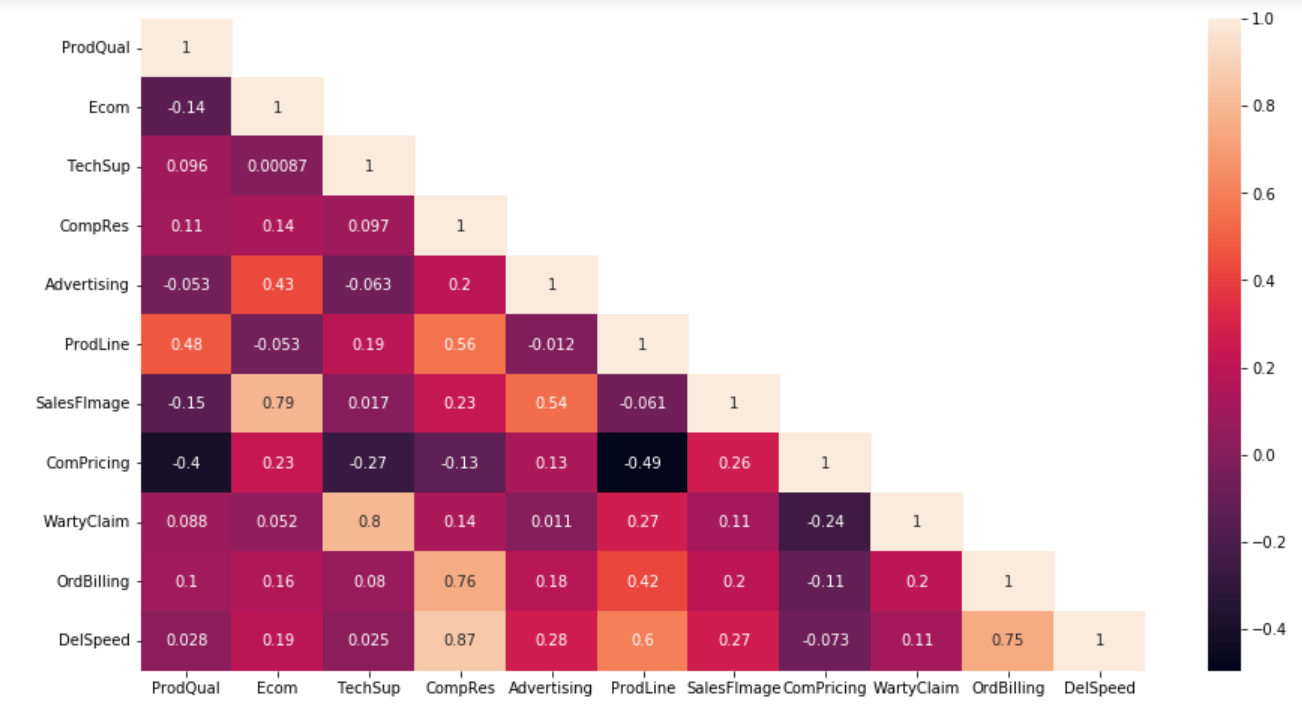
***Figure 10: Univariate Analysis of Order & Billing***

The variable Order and Billing is approximately normally distributed. There are some outliers present in the variable. The mean value is 4.2, the minimum value is 2, the maximum value is 6.7.



***Figure 11: Univariate Analysis of Delivery Speed***

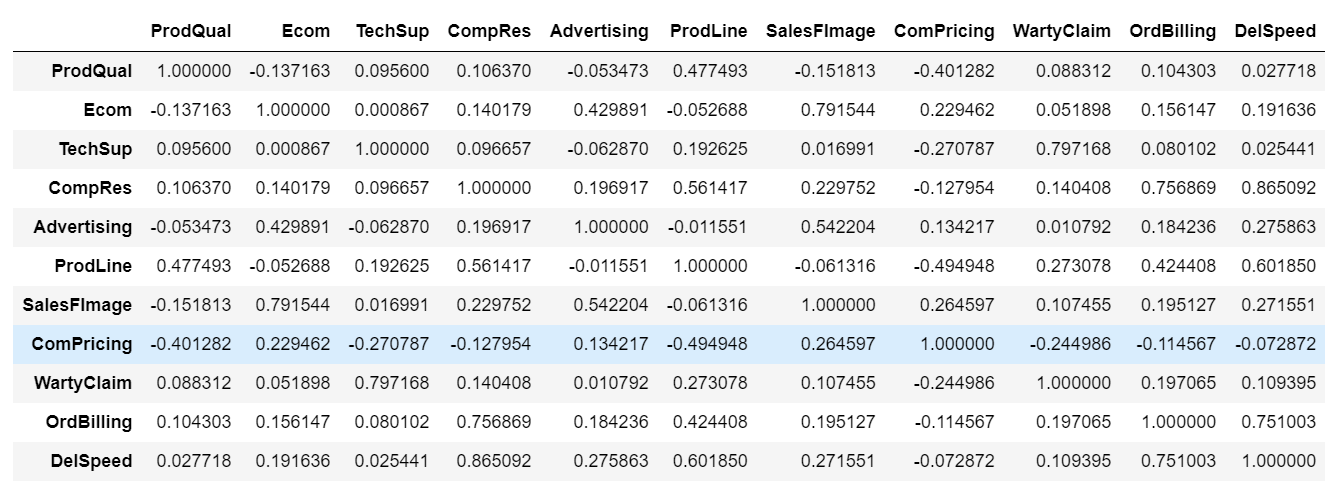
The Variable Delivery Speed seems to be normally distributed. There is one outlier present in the variable. The mean value is 3.8, the minimum value is 1.6, the maximum value is 5.5.



***Figure 12: Multivariate Analysis of the variables***

The multivariate analysis shows the corelation between the variables. The lighter colours depict high corelation. The darker colour depicts low corelation between variables.

The corelation matrix is as follows:



***Table 5: Corelation Matrix***

Observations:

Some key observations that can be drawn from the exploratory data analysis are as follows:

* The dataset contains 100 rows and 13 columns.
* Most of the independent variables are normally distributed.
* There are no missing or duplicate data in the dataset.
* 4 of the independent variables i.e., Ecom, SalesFImage, OrdBilling & DelSpeed contains outliers.
* In the corelation matrix we can observe that SalesFImage & Ecom, WartyClaim & TechSupport, OrdBilling & CompRes, DelSpeed & CompRes, DelSpeed & OrdBilling have high Corelation.

**Question 2:**

**Scale the variables and write the inference for using the type of scaling function for this case study.**

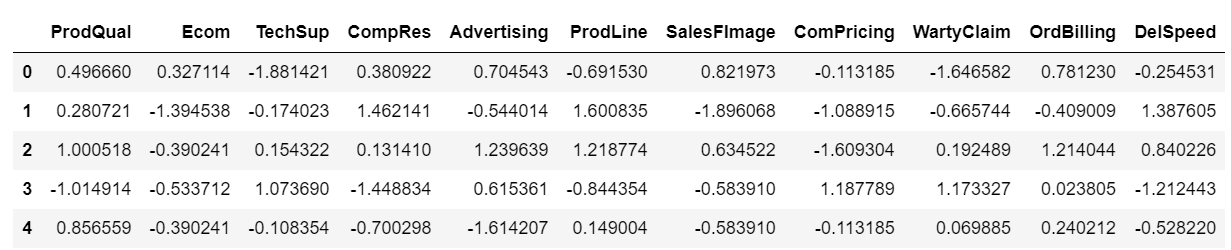
**Answer 2:**

Z-score is a variation of scaling that represents the number of standard deviations away from the mean. We would use z-score to ensure our feature distributions have mean = 0 and std = 1. The formula for calculating the z-score of a point, *x*, is as follows:

x′=(x−μ)/σ

It’s useful when there are a few outliers, but not so extreme that you need clipping. In our dataset we have very few outliers se we can conclude that zscore is the best scaling technique for our dataset.

After importing the necessary library, we will perform the scaling. The output of zscore scaling is as follows:



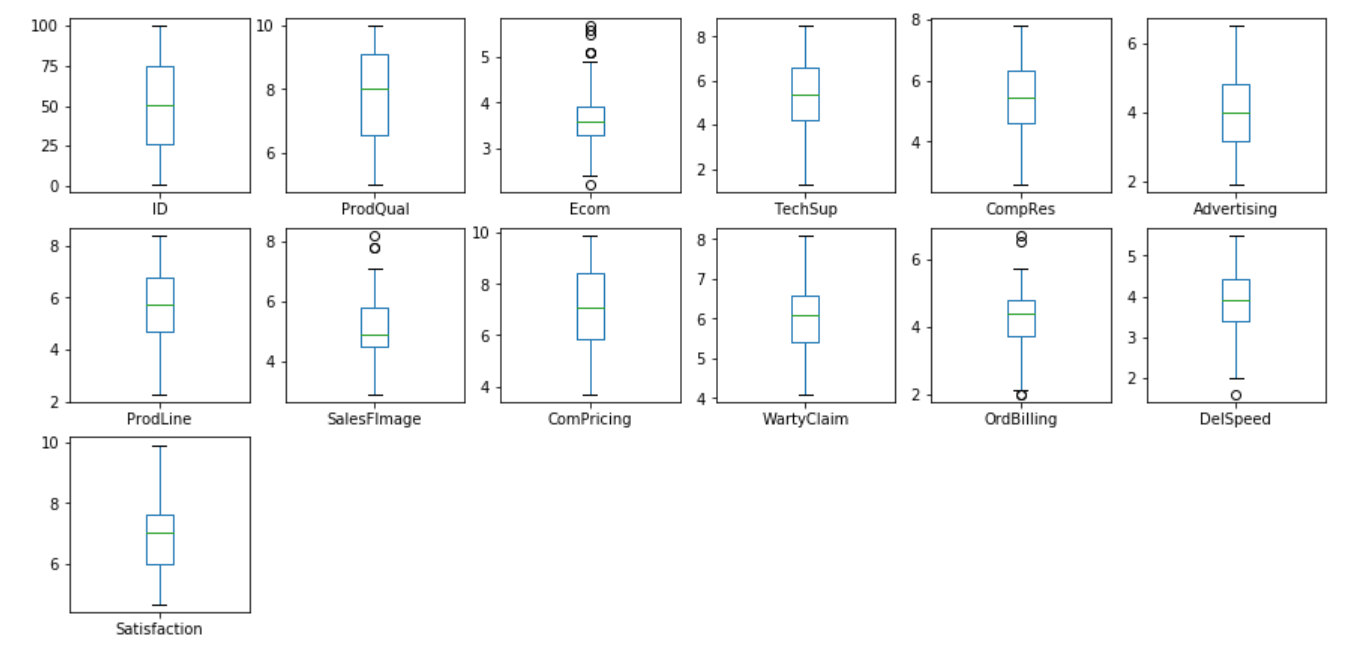
***Table 6: Output after zscore Scaling***

**Question 4:**

**Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.**

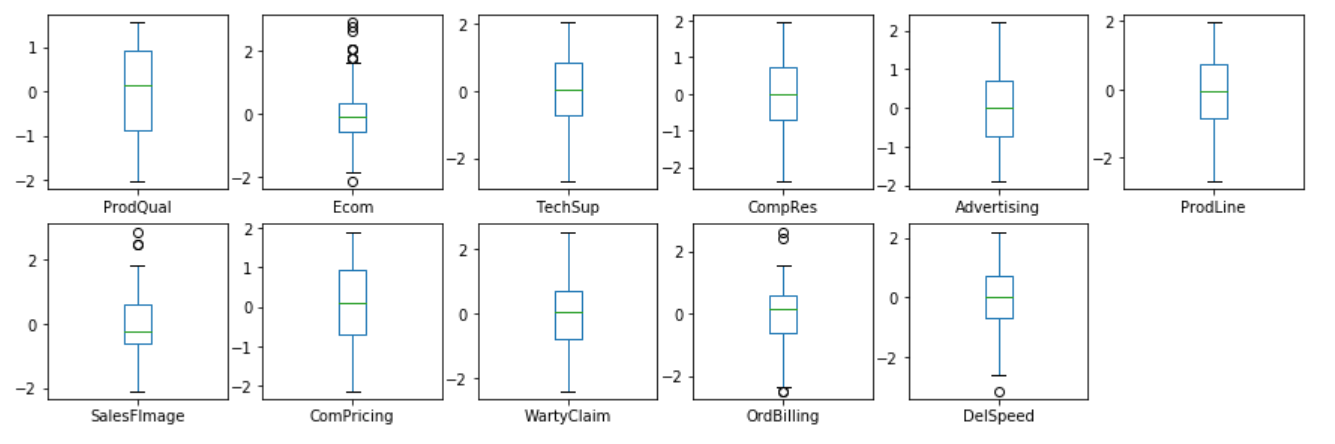
**Answer:**

The outliers before scaling of the dataset is as follows:



***Figure 13: Outliers before Scaling***

The outliers after scaling are as follows:



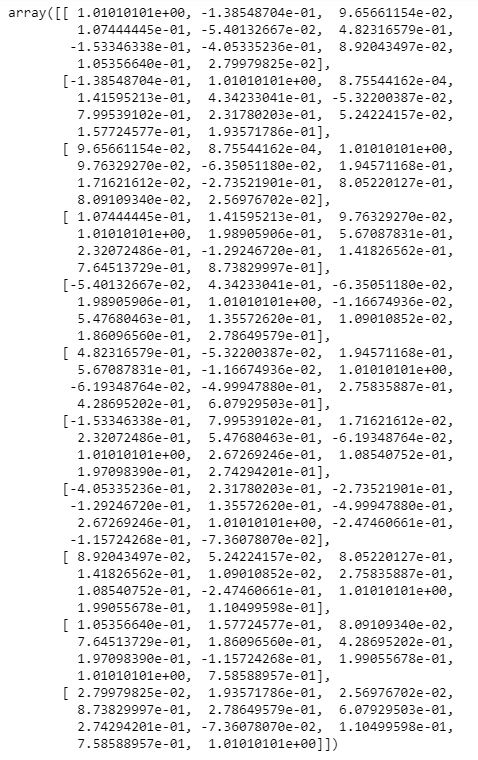
***Figure 14: Outliers After Scaling***

**Question 5:**

**Build the covariance matrix, eigenvalues and eigenvector.**

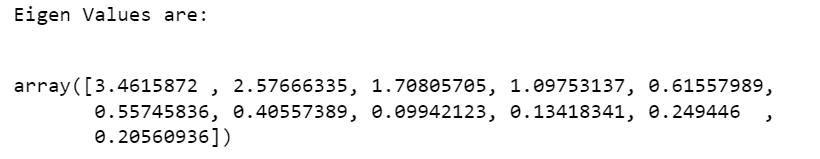
**Answer:**

The covariance matrix is as follows:



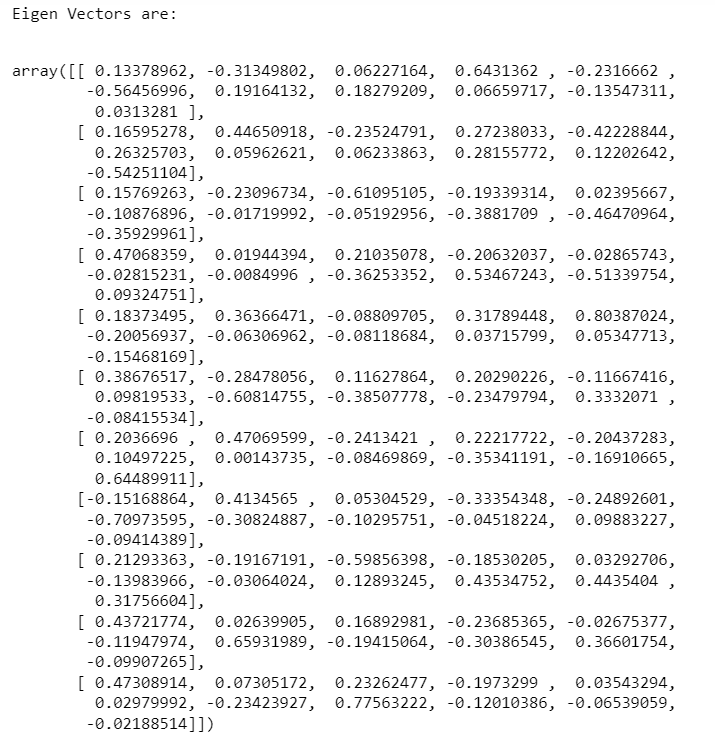
***Table 7: Covariance Matrix***

The Eigen Values are as follows:



***Table 8: Eigen Values***

The Eigen Vectors are as follows:



***Table 9: Eigen Vectors***

**Question 6:**

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

**Answer:**

The explicit form of the first PC

3.4615872\* ProdQual + 2.57666335\*Ecom + 1.70805705\*TechSup + 1.09753137\*CompRes + 0.61557989\*Advertising + 0.55745836\*ProdLine + 0.40557389\*SalesFImage + 0.249446\*ComPricing + 0.20560936\*WartyClaim + 0.13418341\*OrdBilling + 0.09942123\*DelSpeed.

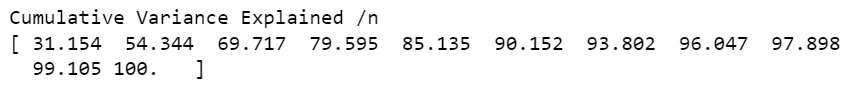
**Question 7:**

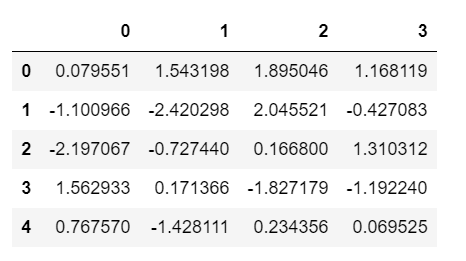
**Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate? Perform PCA and export the data of the Principal Component scores into a data frame.**

**Answer:**

The cumulative value up to fourth Principal Component is 79.6. General rule of thumb is to choose first k PC’s such that the first k PC’s explain 70-90% of the total variance. Hence from the cumulative values of eigen values, help in selecting the required no. of PC’s. In this case first Fourth PC’s have been selected capturing 79.6% of variation and thereby reducing our dimension from 18 to 4.

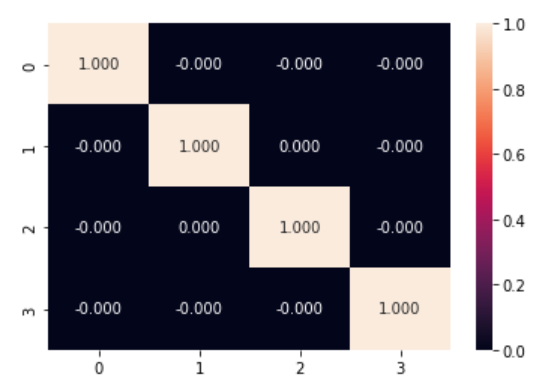
Below is the figure of the same.





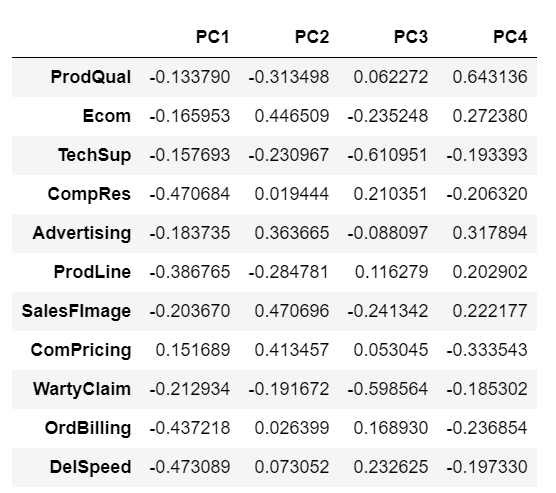
***Table 10: First few rows of reduced dimension***

The Heatmap is as follows:



***Figure 15: Heatmap of 4 PC***

The data of the principal component score is as follows:



***Table 11: Principal Component Score***

**Question 8:**

**Mention the business implication of using the Principal Component Analysis for this case study.**

**Answer:**

Principal Component Analysis (PCA) is an unsupervised learning algorithm technique used to examine the interrelations among a set of variables. PCA forms the basis of multivariate data analysis based on projection methods. The most important use of PCA is to represent a multivariate data table as smaller set of variables (summary indices) in order to observe trends, jumps, clusters and outliers. It finds a sequence of linear combinations of variables. PCA also serves as a tool for better data visualization of high dimensional data. We can create a heat map to show the correlation between each component. It is often used to help in dealing with multi- collinearity before a model is developed.

PCA reduces the number of variables that are correlated to each other into fewer independent variables without losing the essence of these variables. It provides an overview of linear relationships between inputs and variables. In this case there were 18 variables at the beginning which were to be analysed to find the relation with the target variable. But with the use of PCA we were able to reduce the dimensionality to just 4 variables which was able to establish a relationship with the target variables.

**Problem B:**

**Problem Statement**:

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.

The Data Dictionary is as follows:

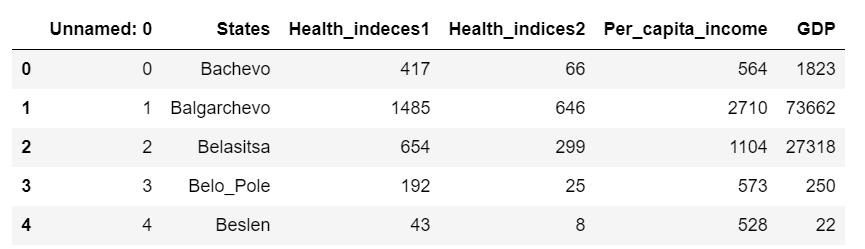
1. States: names of States  
   2. Health\_indeces1: A composite index rolls several related measures (indicators) into a single score that provides a summary of how the health system is performing in the State.  
   3. Health\_indeces2: A composite index rolls several related measures (indicators) into 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 rate.

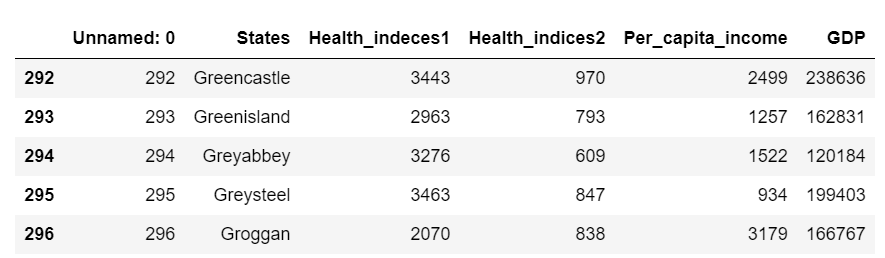
**Question 1:**

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

**Answer:**

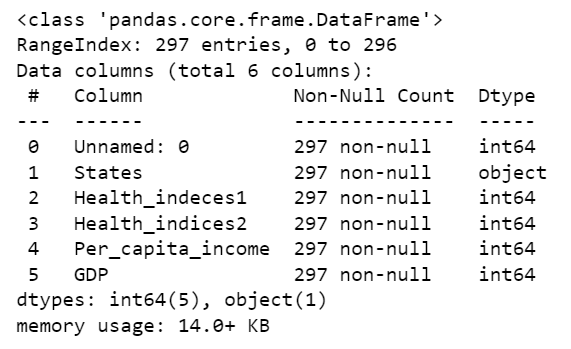
After importing the necessary libraries, we will import the dataset and perform the EDA. The First and Last few rows of the dataset is as follows:





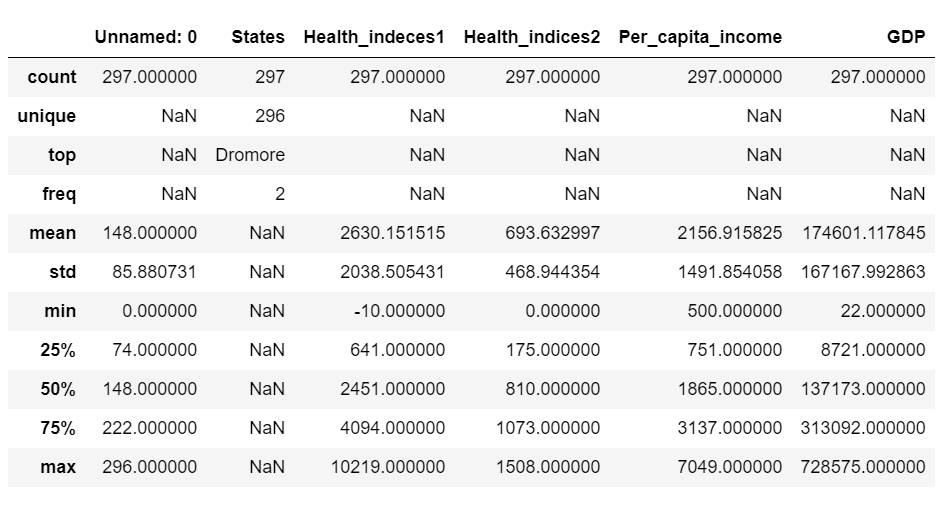
***Table 12: First and Last Few rows of the Dataset***

There are 297 rows and 6 Columns of the dataset. The info of the dataset is as follows:



***Figure 16: info of the dataset***

The basic Description of the dataset is as follows:

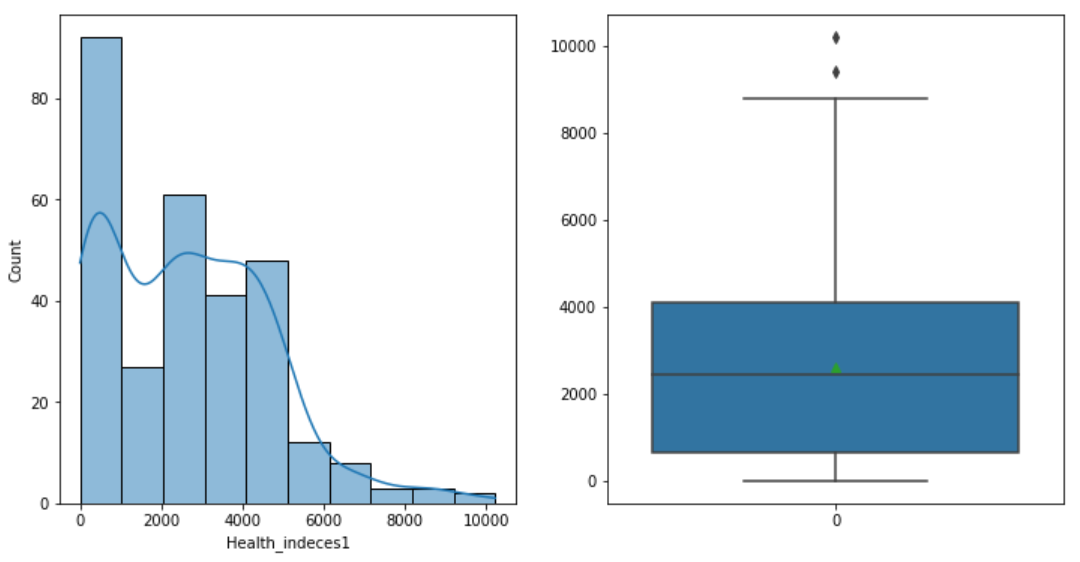


***Table 13: Description of the Dataset***

Observations:

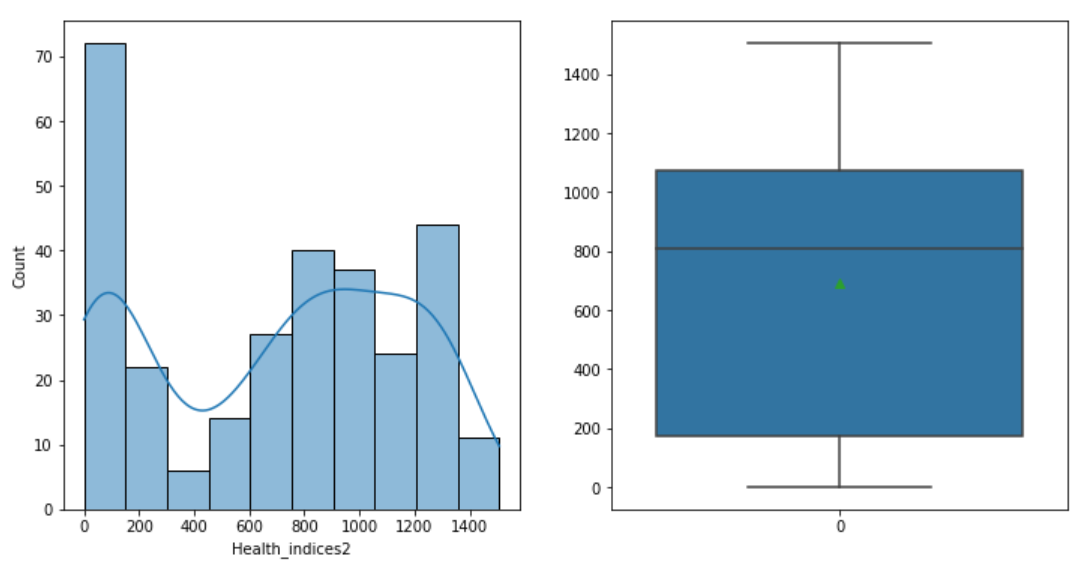
* The dataset contains data about the Health and economic conditions in different States of a country.
* There are 297 rows and 6 columns in the dataset.
* There are no null or missing values or duplicate values in the dataset.
* The dataset contains data of Health and economic conditions of 296 states.
* Health Indices 1 has a mean value of 2630 and Health Indices 2 has a mean value of 693.
* Per capita income has a mean value of 2156.9 with minimum per capita income at 500 and maximum at 7049.
* Mean GDP is at 174601 with Minimum GDP of a state is at 22 and the maximum at 728575.

Exploratory Data Analysis.



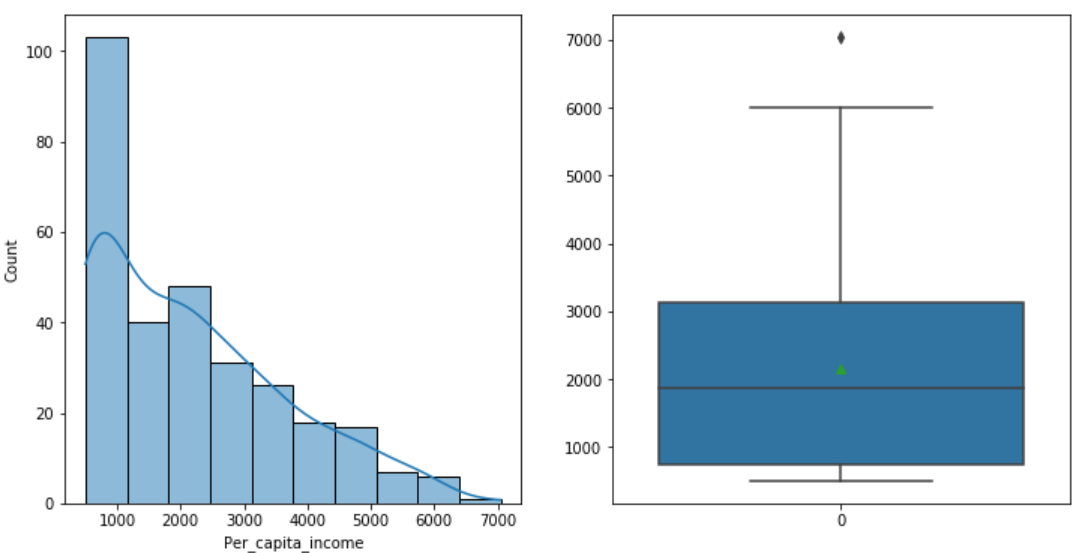
***Figure 17: Distribution of Health Indices 1***

The data is left skewed. The mean is at 2630. We also observe that there are outliers in the dataset.



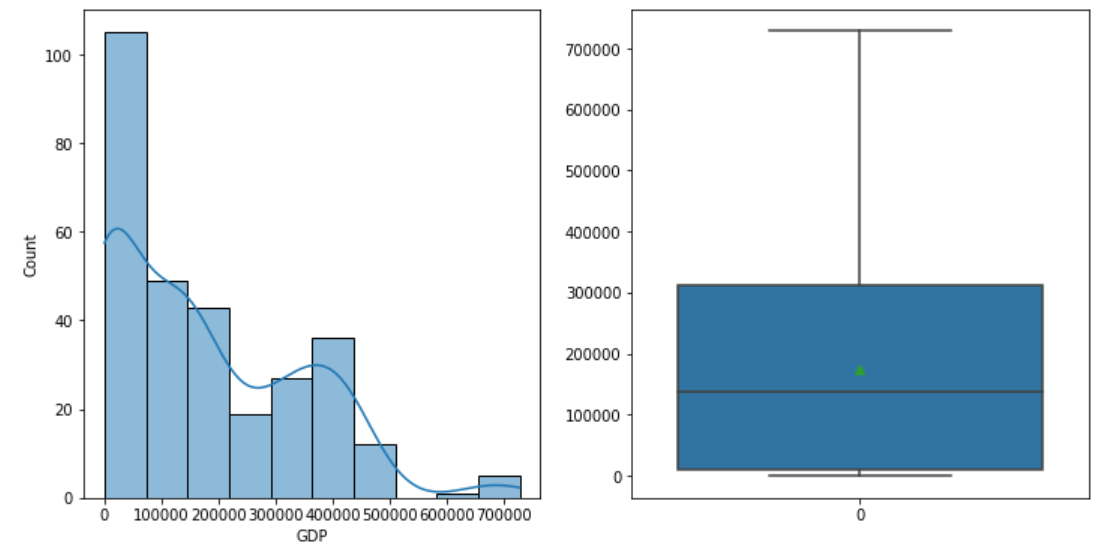
***Figure 18: Distribution of Health Indices 2***

We observe that the variable is not normally distributes. The mean of the health indices 2 is 693. There are no outliers in the dataset.



***Figure 19: Distribution Of Per Capita Income***

We observe that the variable is left skewed and not normally distributed. The mean is 2156.9 and there are outliers present.

** ***Figure 20: Distribution of GDP***

The data is not normally distributed. It is left skewed. The mean value is 174601 and there are no outliers present in the data.

**Question 2:**

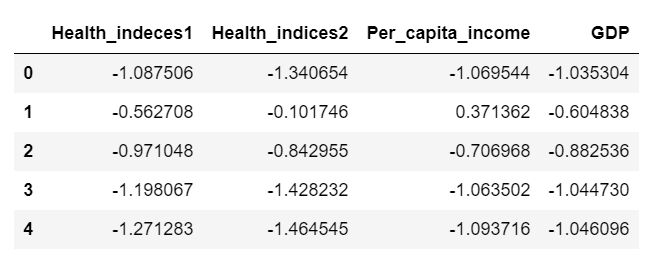
**Do you think scaling is necessary for clustering in this case? Justify.**

**Answer:**

Scaling is a step of Data Pre-Processing that is applied to independent variables or features of data. It helps to normalize the data within a particular range. Scaling the data can help to balance the impact of all variables on the distance calculation and can help to improve the performance of the algorithm. Scaling means that we’re transforming your data so that it fits within a specific scale, like 0-100 or 0-1. We want to scale data when we’re using methods based on measures of how far apart data points.

The aim of the study of this dataset is to form clusters of 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. So, we need to do scaling as the dataset contains data points which vary from one another very much. Hence, we need to perform scaling to bring the data points withing one range and perform out clustering techniques.

The scaled data is as follows:



***Table 14: First Few rows of the Scaled Data***

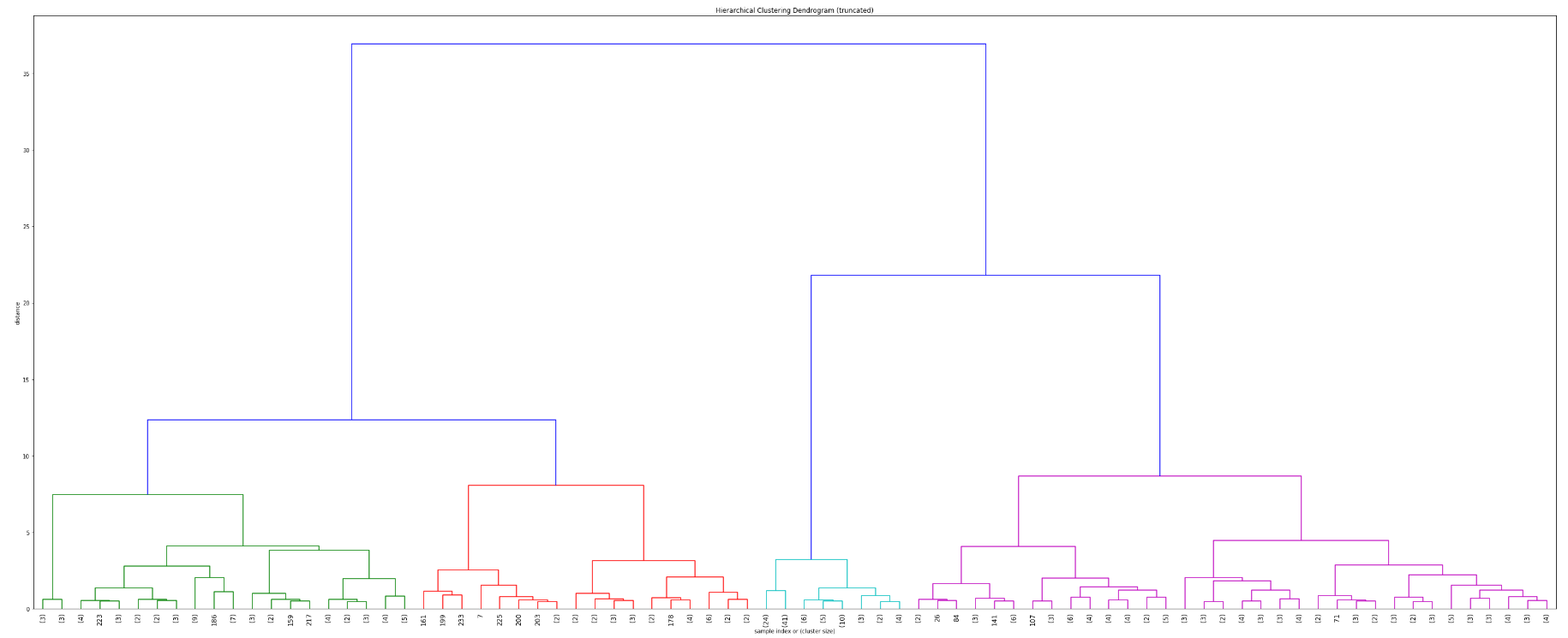
**Question 3:**

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

**Answer:**

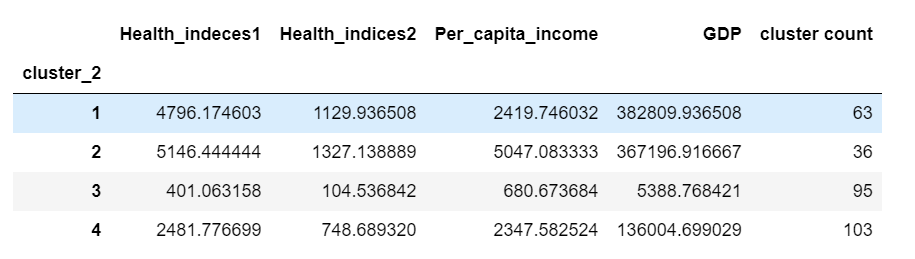
Hierarchical clustering is a method of cluster analysis in data mining that creates a hierarchical representation of the clusters in a dataset. The method starts by treating each data point as a separate cluster and then iteratively combines the closest clusters until a stopping criterion is reached.

The Hierarchical clustering is as follows:



***Figure 21: Hierarchical Clustering using Dendrogram***

The Dendrogram gives us the clusters in which the data is divided. We will choose 4 clusters as our optimum clusters. The following table shows the clustered data.



***Table 15: Clustered Dataset***

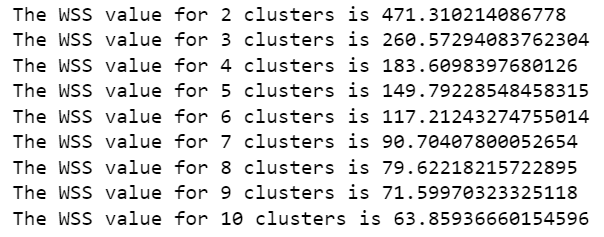
**Question 4:**

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

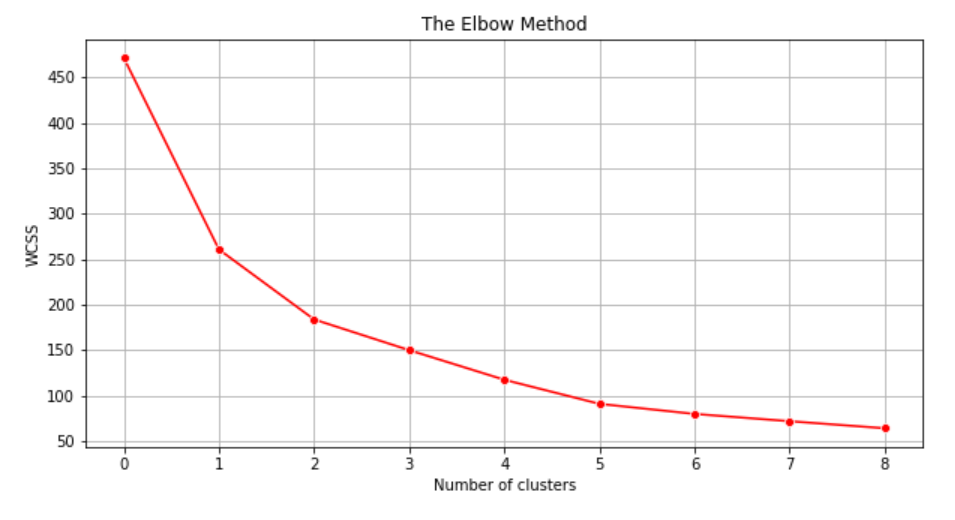
**Answer:**

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

First we will find out the WSS(within sum of squares) for our dataset which will help in determining the optimum number of clusters. The WSS for a range of 2 to 11 is as follows.



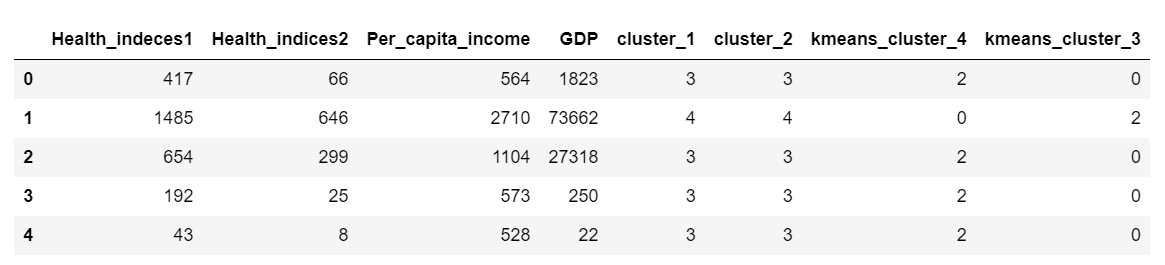
We will look at the elbow plot to determine the optimum number of clusters for our dataset. The elbow plot is as follows.



***Figure 22: Elbow Plot***

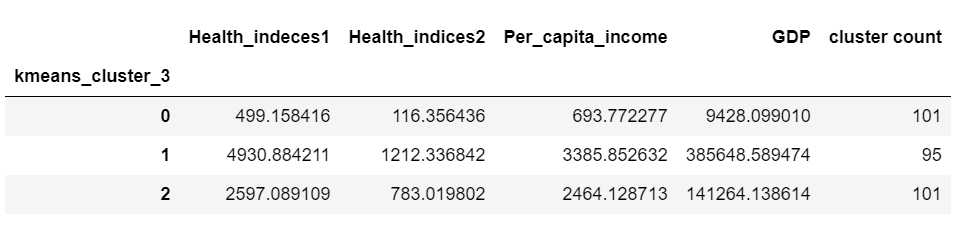
The elbow plot hints that the optimum number of clusters should be 3 or 4. So, we will perform K-means for 3 and 4 clusters.

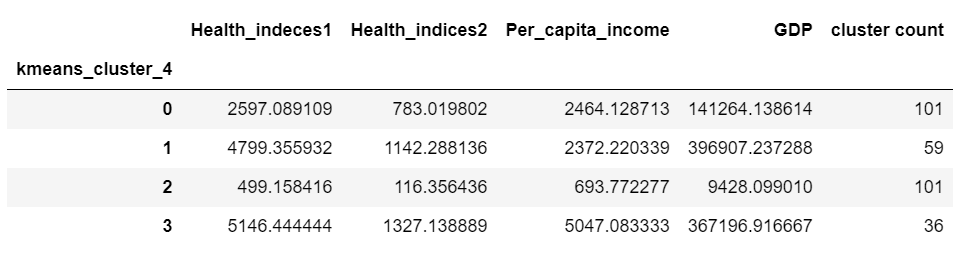
On performing the K-Means clustering on our scaled dataset we get the following output for 3 clusters and 4 clusters.



***Table 16: K-means clustering for 3 and 4 clusters***

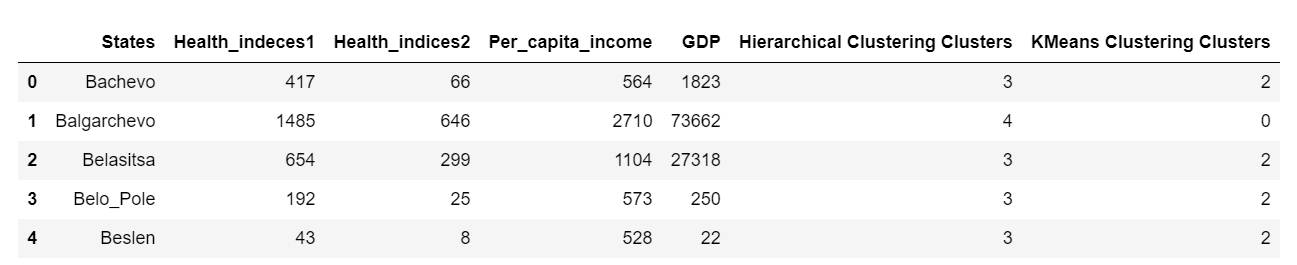
The distribution of the data in cluster 3 and cluster4 is as follows.





***Table 17: Distribution of data in 3 Clusters & 4 Clusters***

The final Distribution of data using Hierarchical & K-Means Clustering is as follows.



***Table 18: Distribution of data using Hierarchical & K-Means Clustering***

The silhouette score of the K-means clustering with 4 clusters is as follows.



**Question 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.**

**Answer:**

***Figure 23: Cluster Profiling***

* 34% of the states belong to cluster 0. They consist of 101 states.
* 20% of the states belong to cluster 1. They consist of 59 states.
* 34% of the states belong to cluster 2. They consist of 101 states.
* 12% of the states belong to cluster 3. They consist of 36 states.