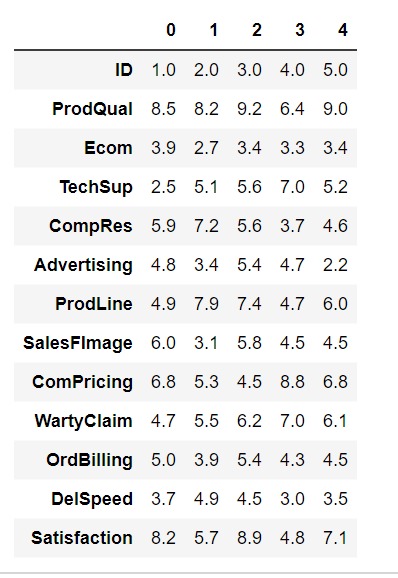
**DATA MINING PROJECT BUSINESS REPORT**

**PART 1: PRINCIPAL COMPONENT ANALYSIS (PCA)**

**Problem Statement:** The [‘Hair Salon.csv’](https://olympus.mygreatlearning.com/courses/84255/files/7192422/download?wrap=1)[View in a new window](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. **Kindly refer to the**[**PCA\_Data\_Dictionary.jpg**](https://olympus.mygreatlearning.com/courses/84255/files/7192423/download?wrap=1)**[View in a new window](https://olympus.mygreatlearning.com/courses/84255/files/7192423/download?wrap=1) file for the Data Dictionary of the Dataset.**  
**Note: This particular dataset contains the target variable satisfaction as well. Please drop this variable before doing Principal Component Analysis.**

**Described all the heads of data.**

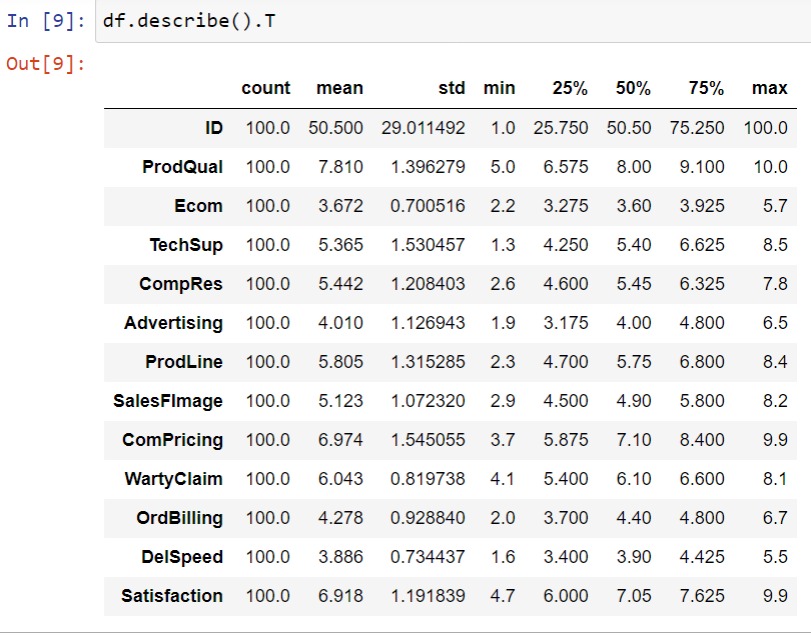
****

**Shape of the data:**

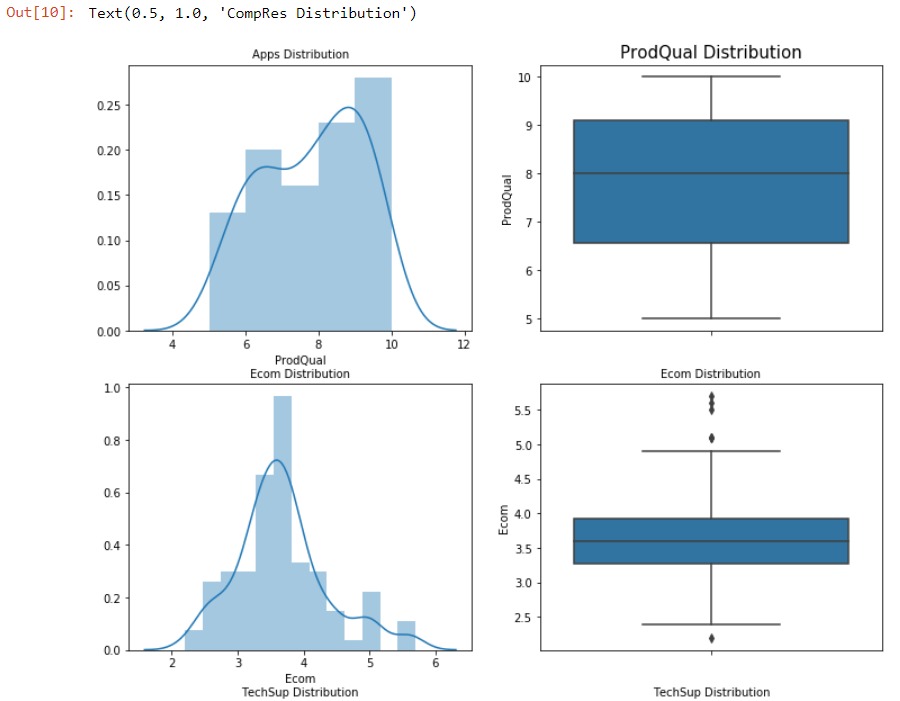
(100, 13)

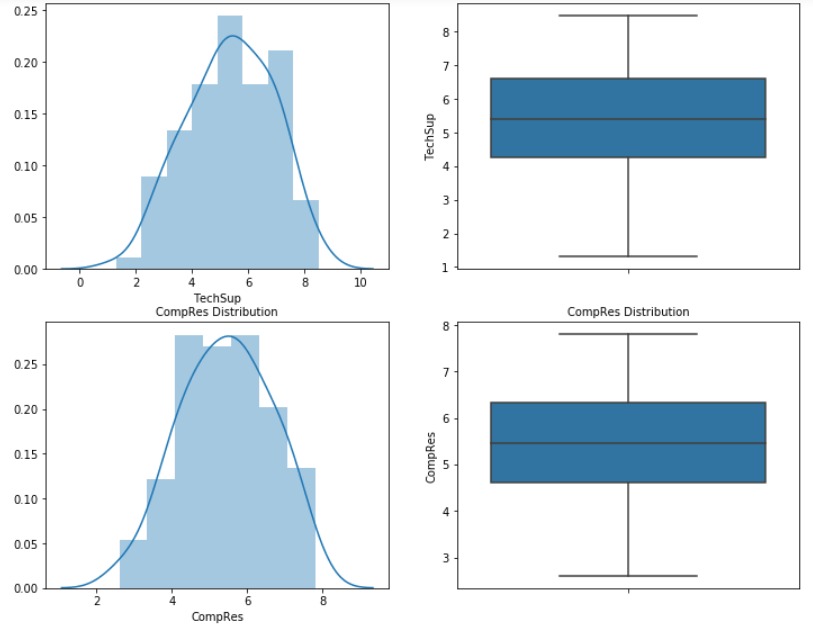
There are 0 null values and 0 duplicated values.

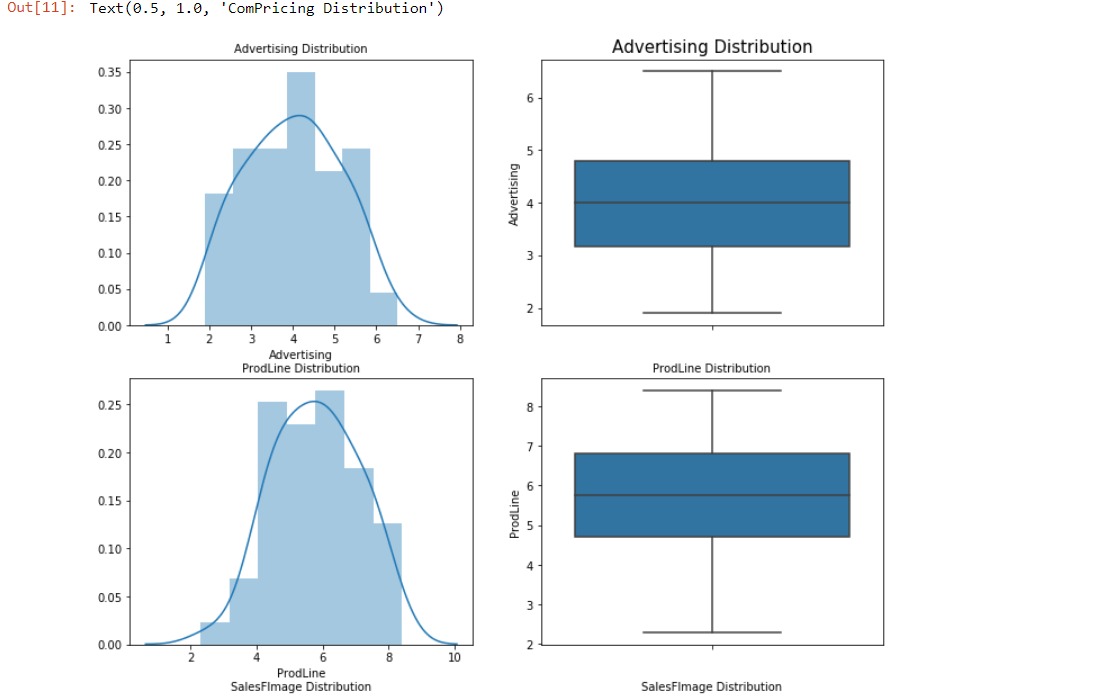
Described all the values.

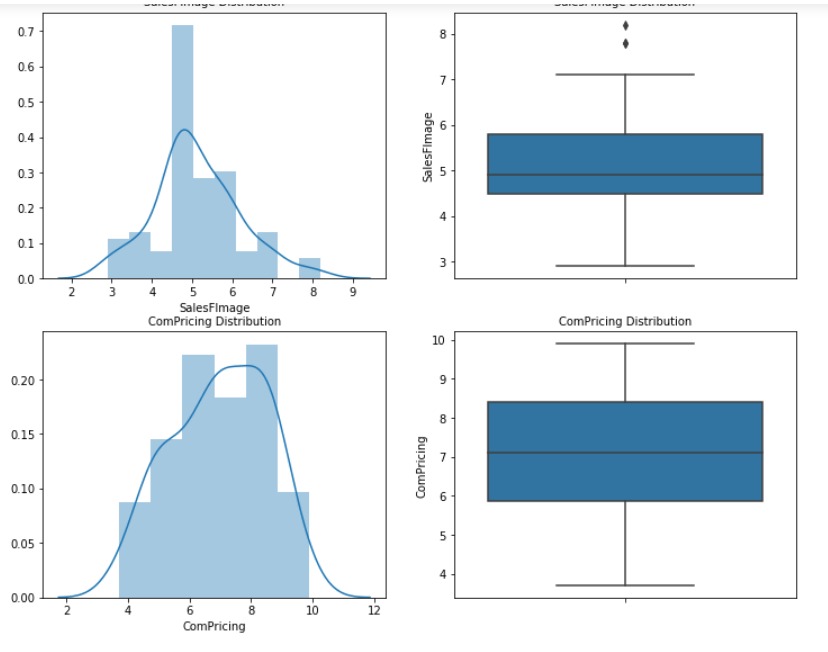


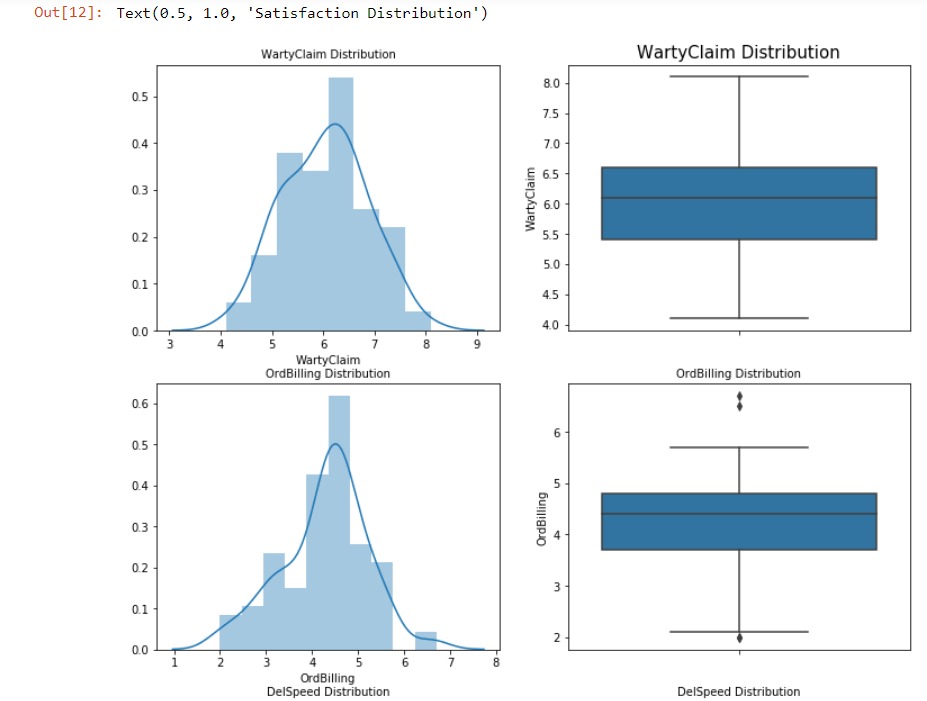
**1)Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented.**

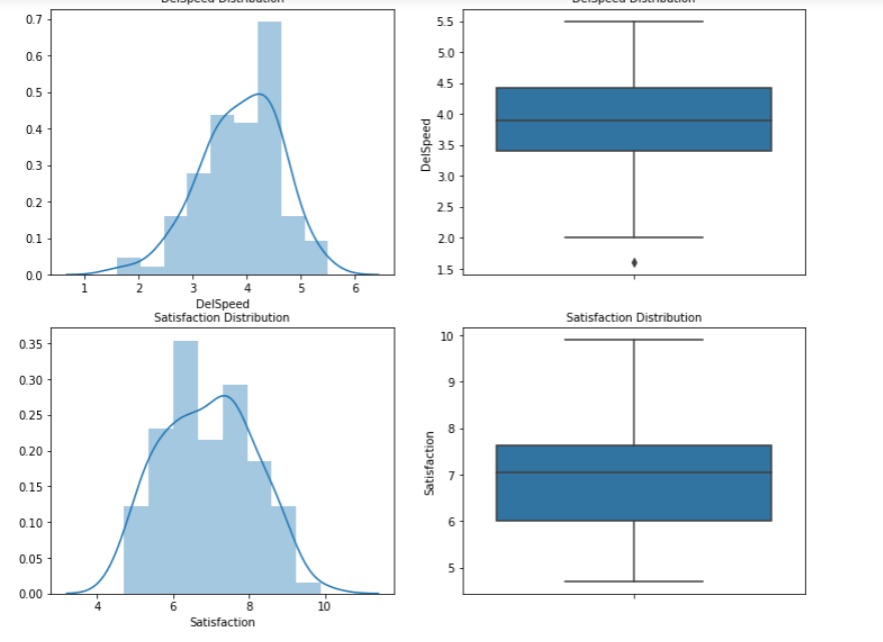
****

****

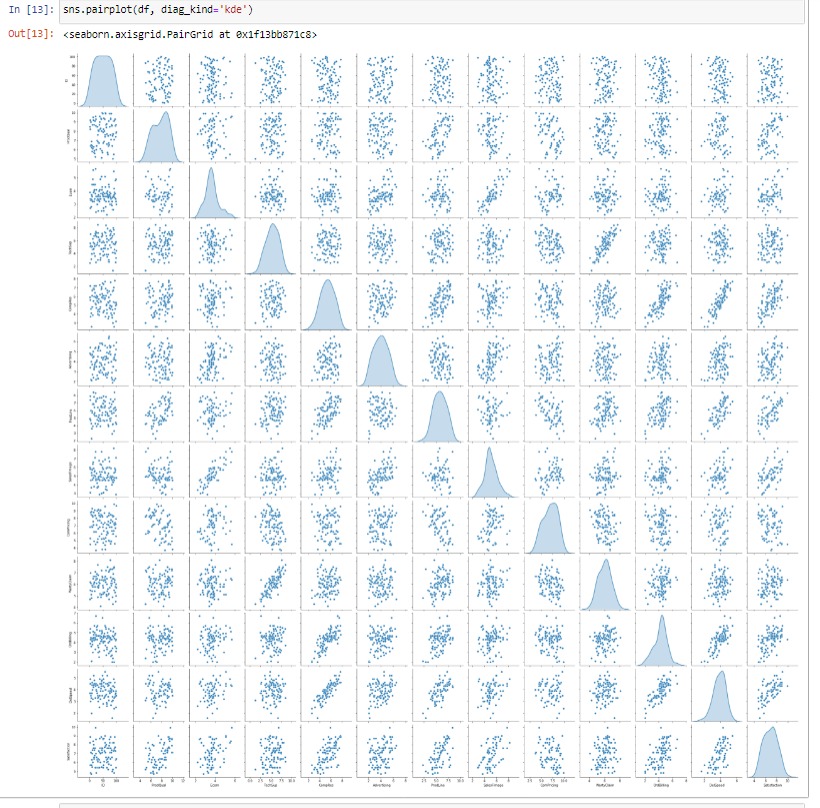
****

****

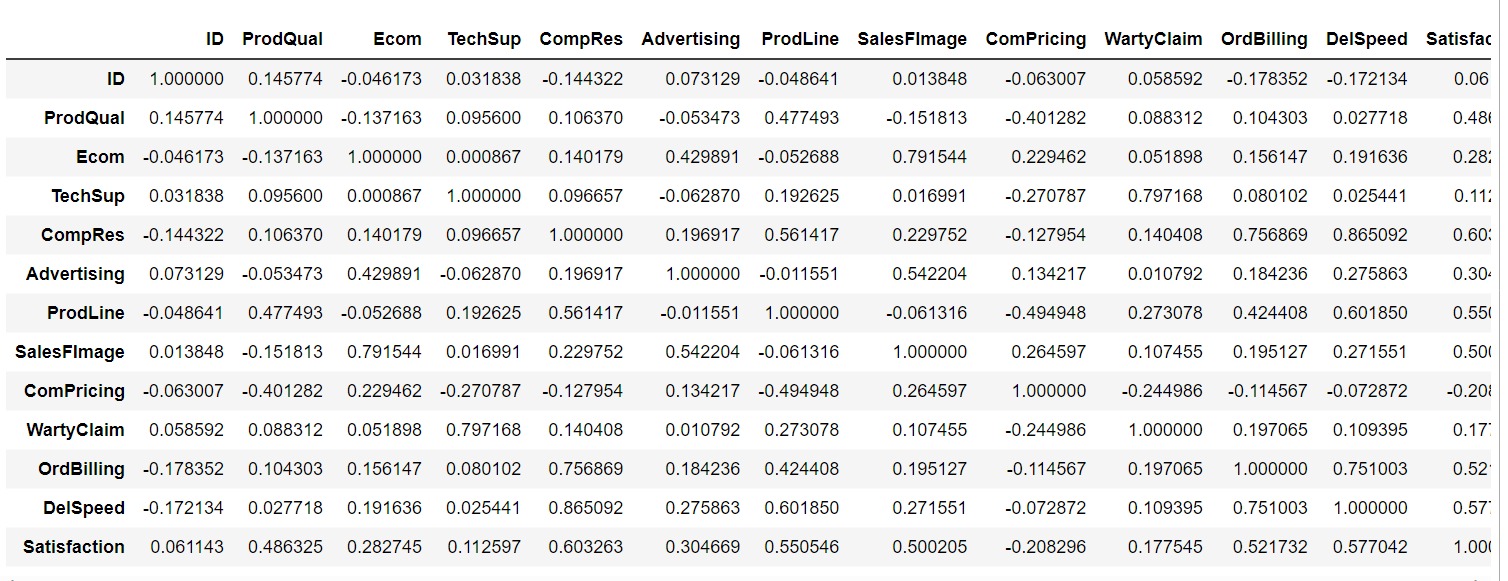
****



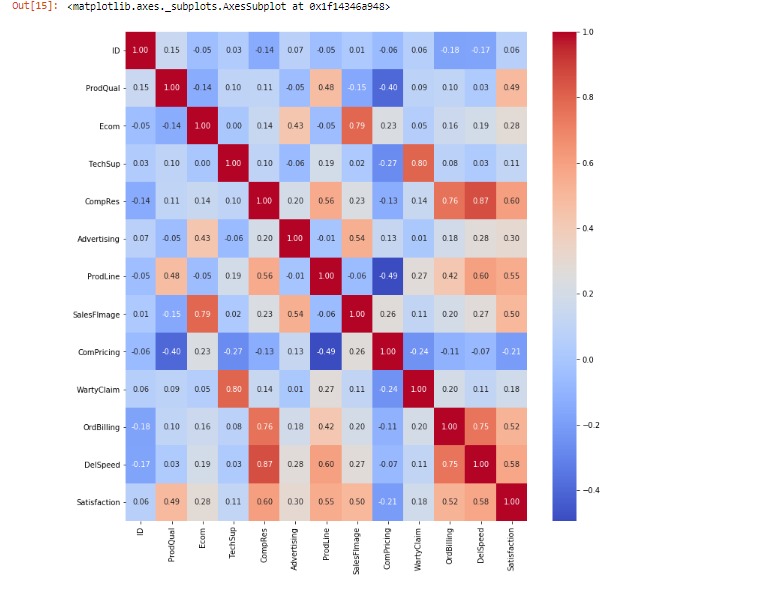
 Shown all the outliers of the data respectively.



We can detect that in the above pairplot.



Correlations of the data.



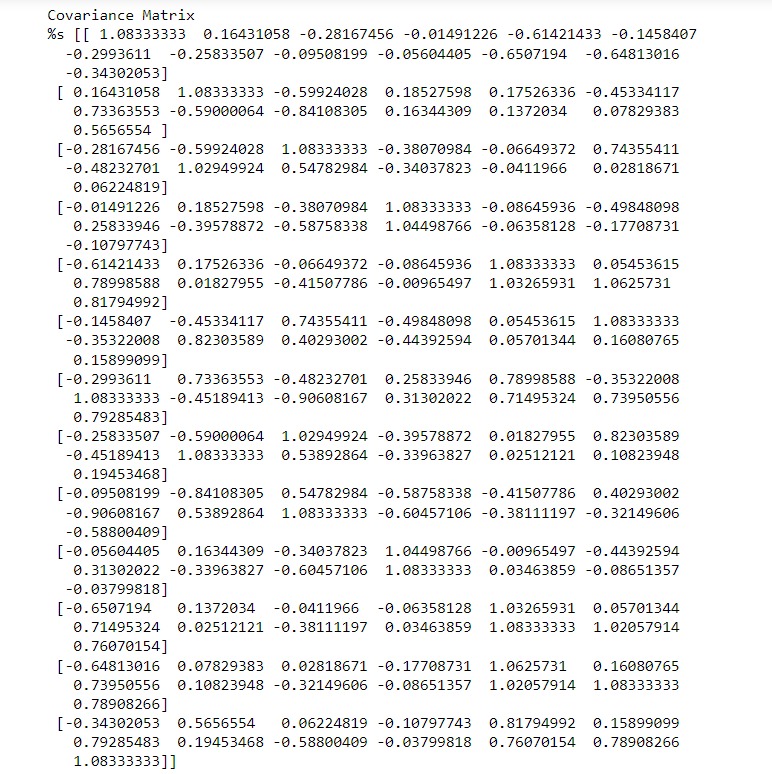
We can see the perfect heatmap of the hair salon.

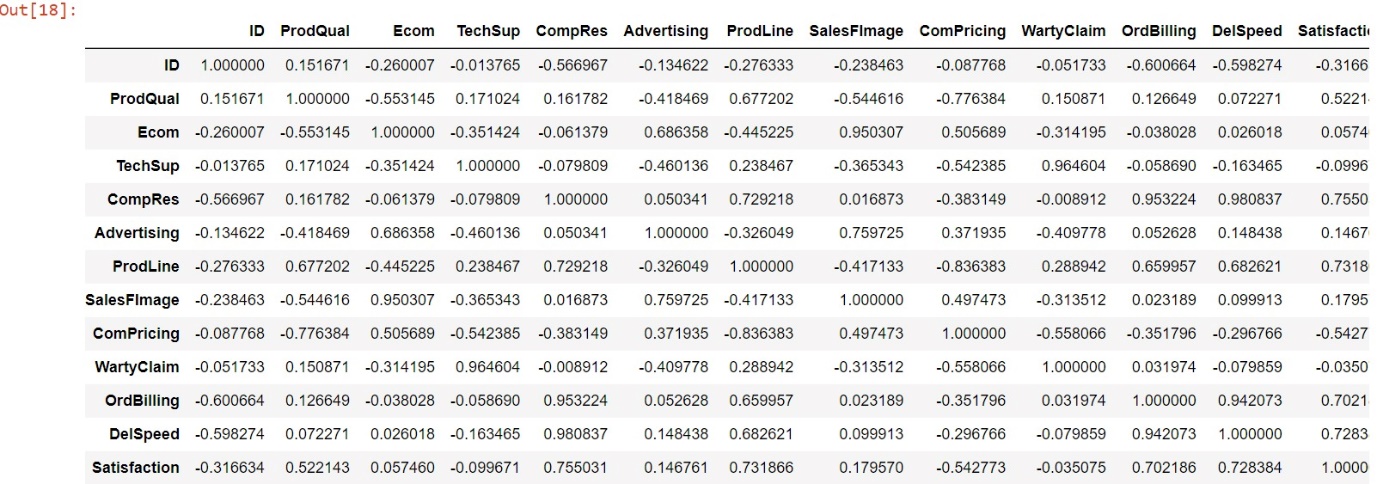
## 2)Scale the variables and write the inference for using the type of scaling function for this case study.

## 

Using zscore we can find the perfect scaling functions.

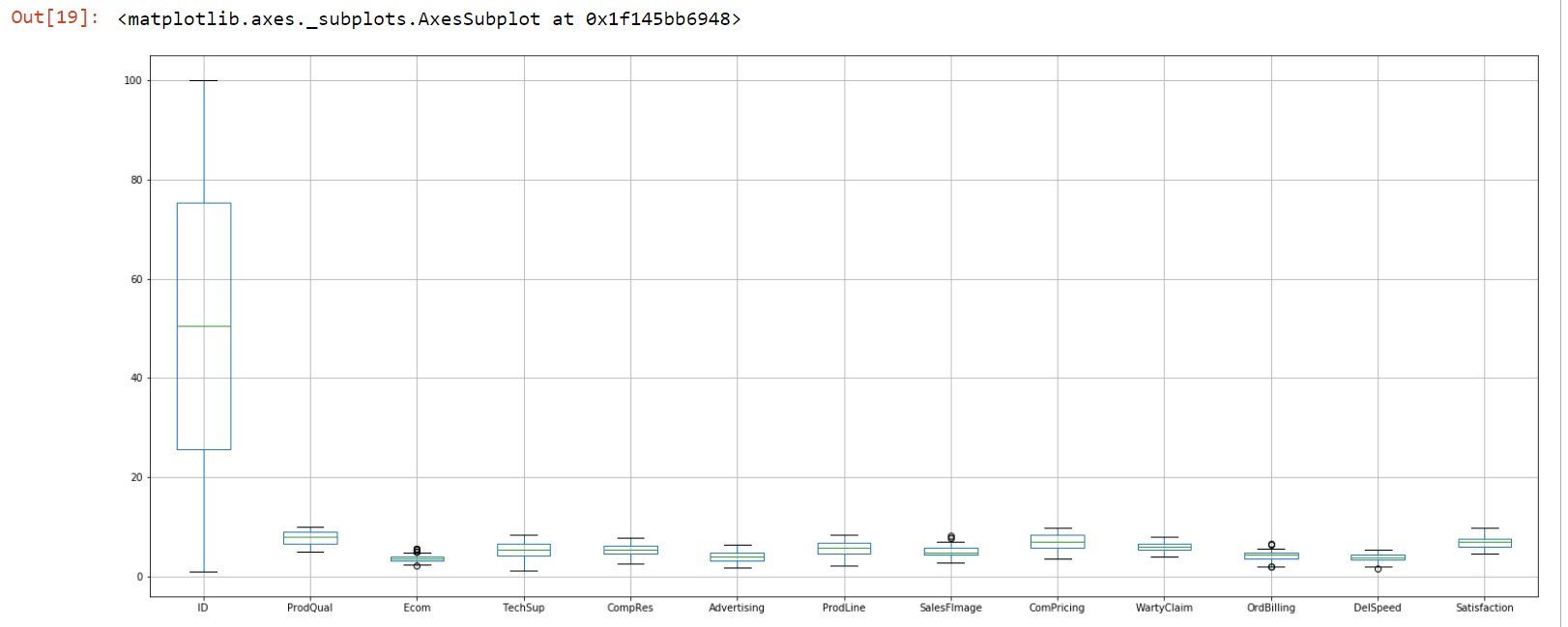
## 3)Comment on the comparison between covariance and the correlation matrix after scaling.

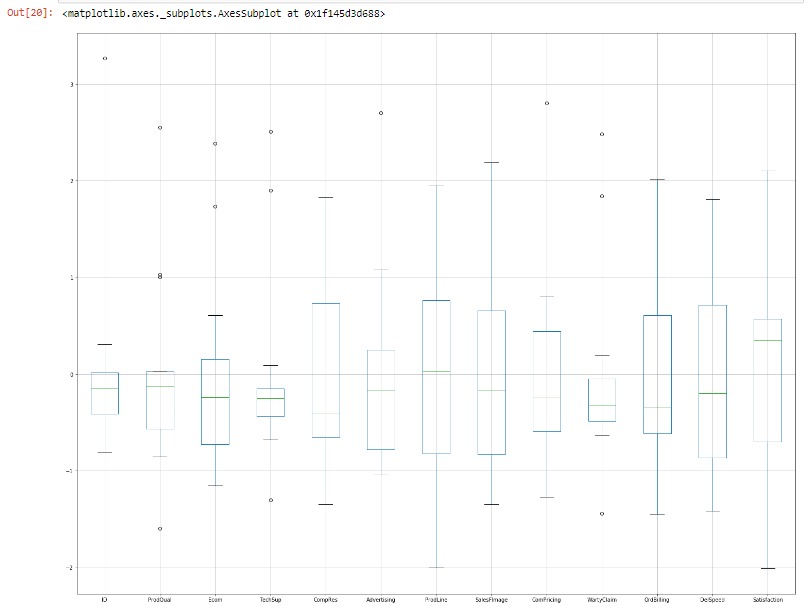


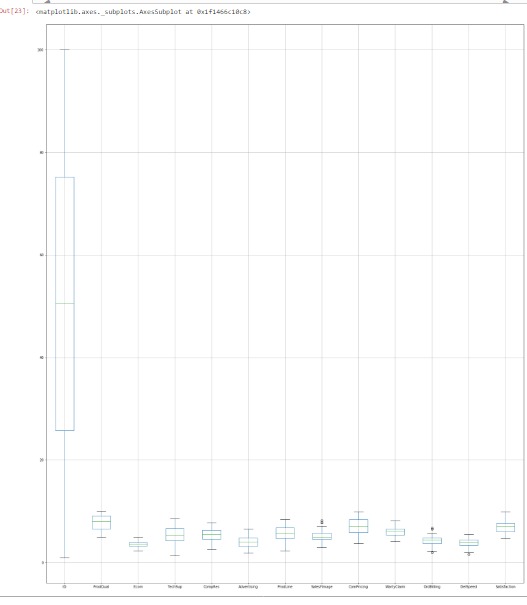


Both correlation and covariance measures are also unaffected by the change in location. However, when it comes to making a choice between covariance vs correlation to measure relationship between variables, correlation is preferred over covariance because it does not get affected by the change in scale.

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

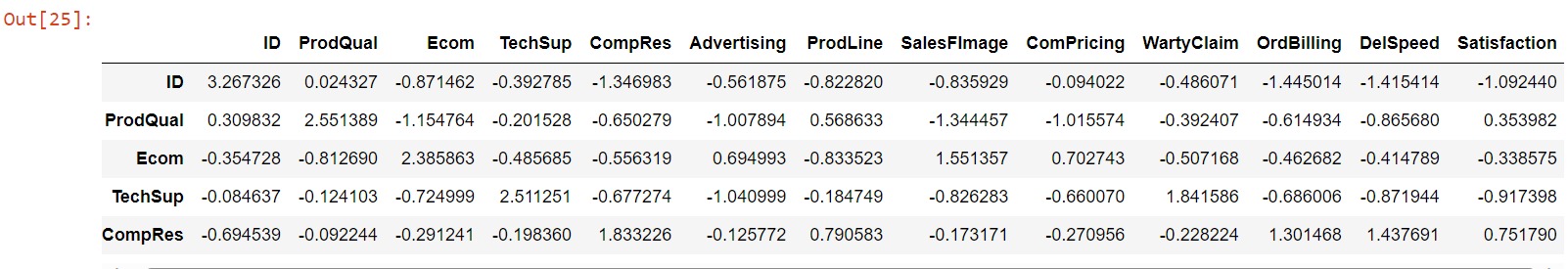


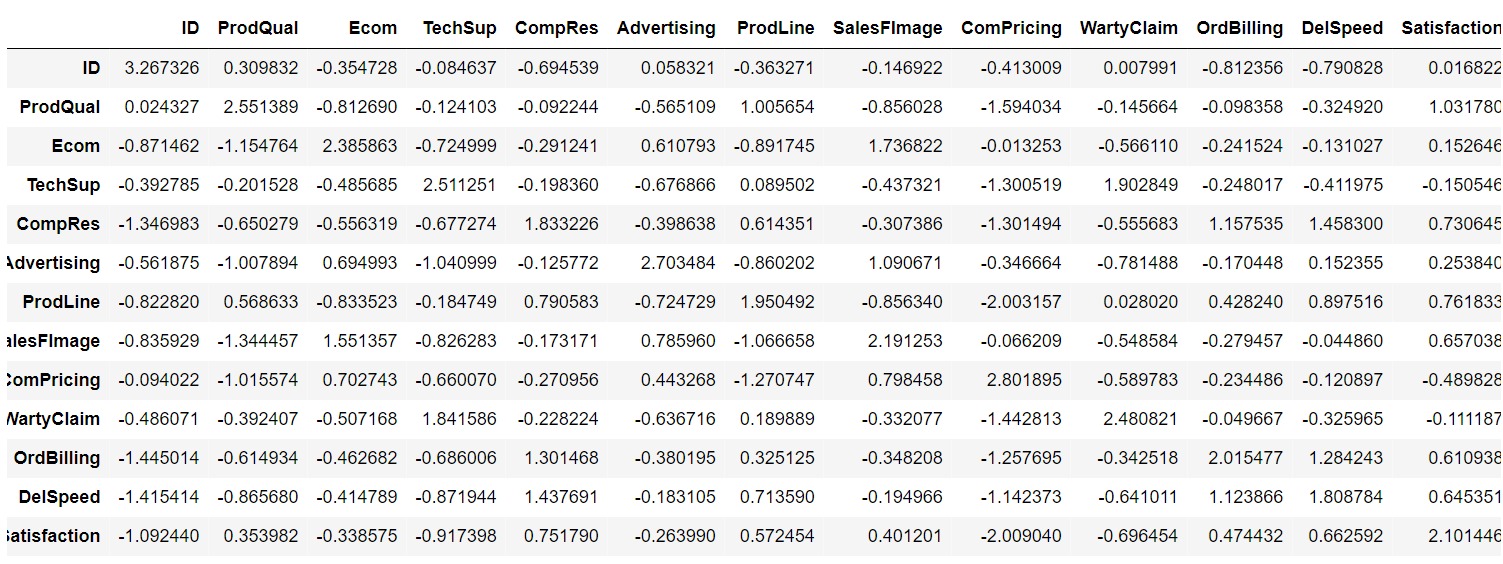


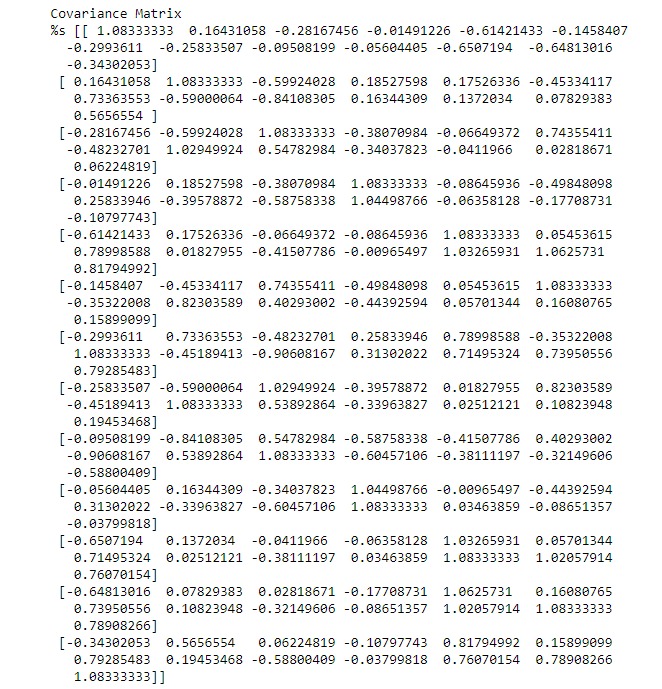


Checked all the dataset for outliers.

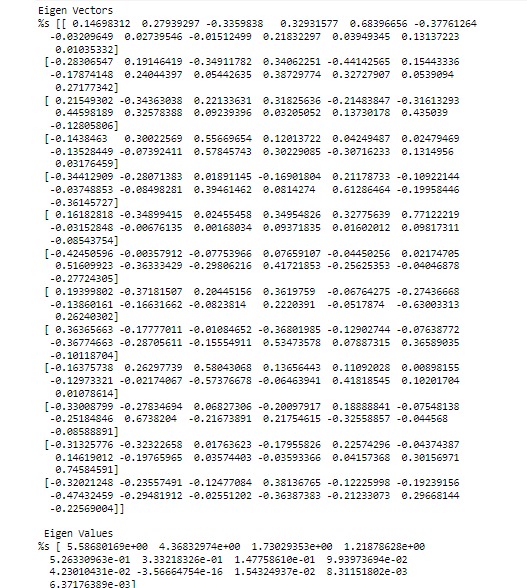
## 5)Build the covariance matrix, eigenvalues, and eigenvector.



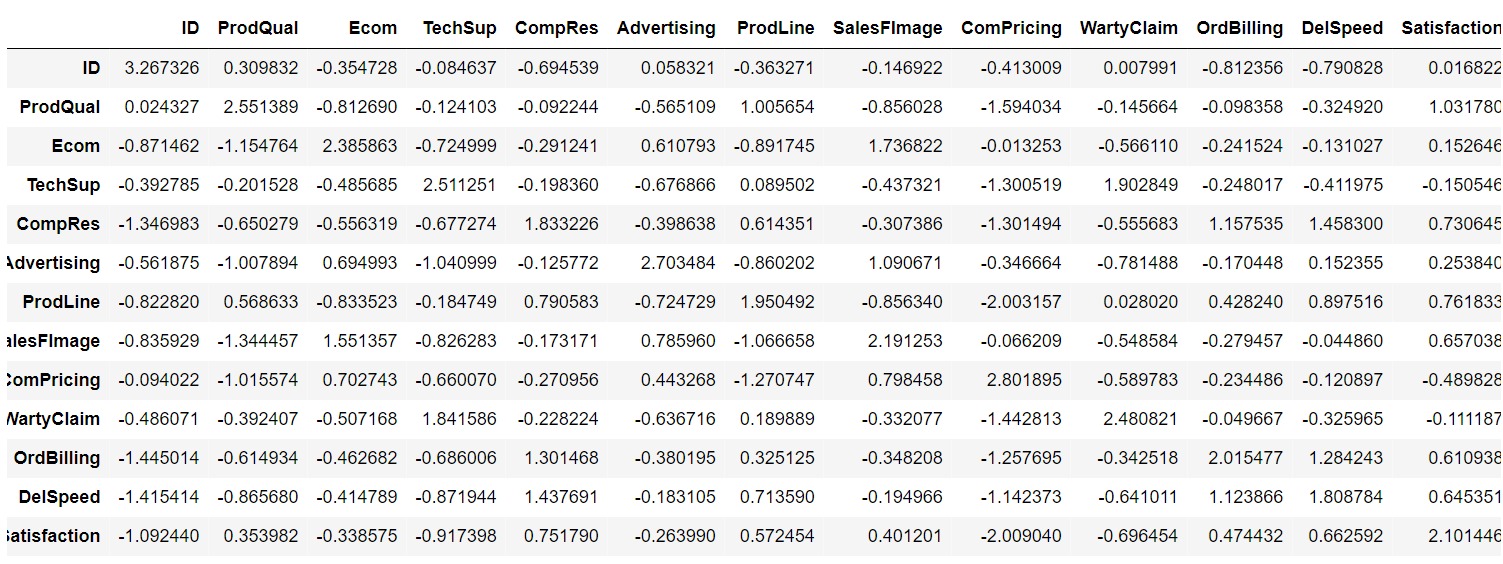




We can see all covariance matrix in the above figure.

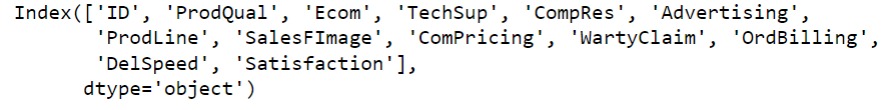


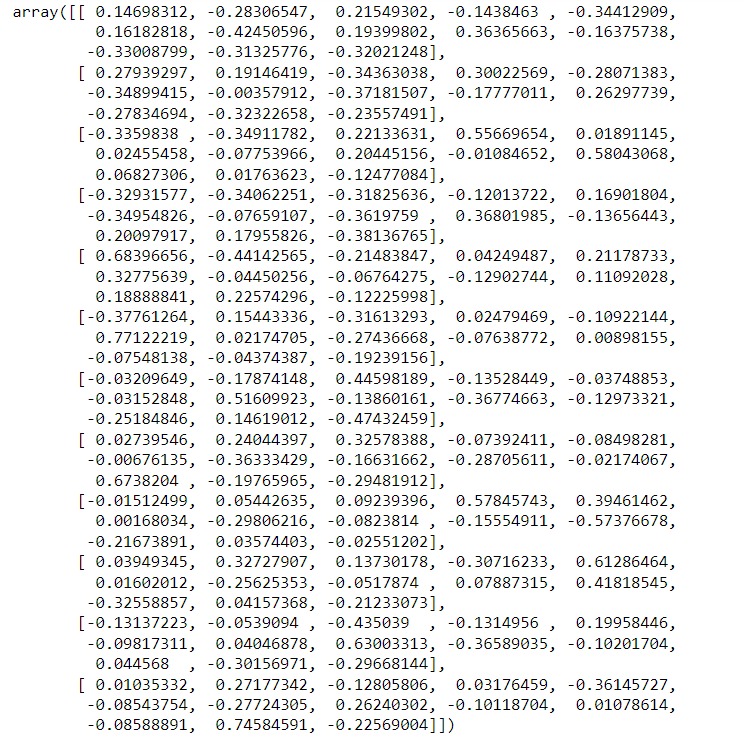
Found all the eigen vectors and eigen values respectively.

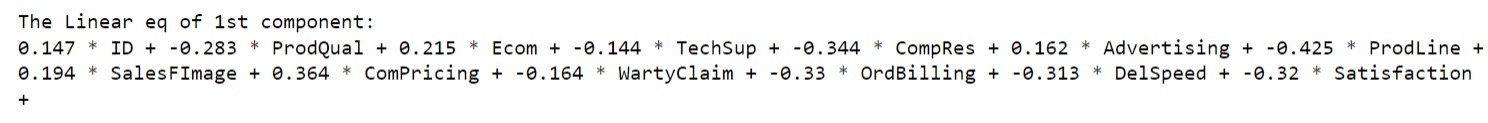


## 7)Write the explicit form of the first PC (in terms of Eigen Vectors).

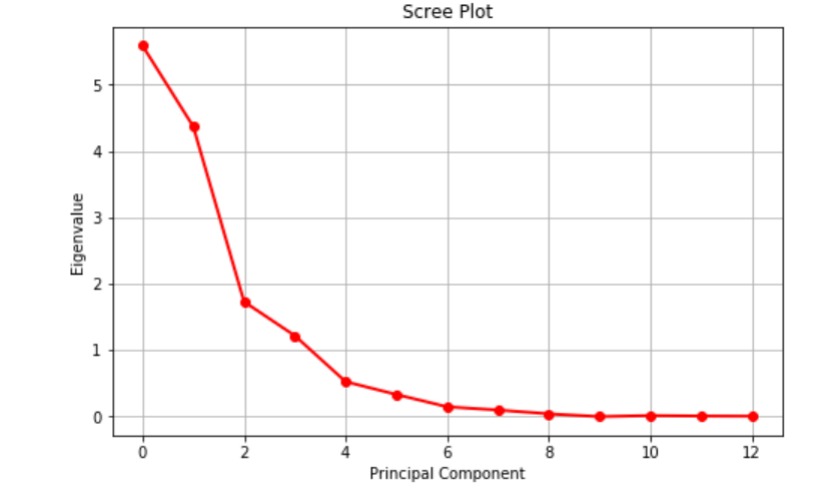
## 



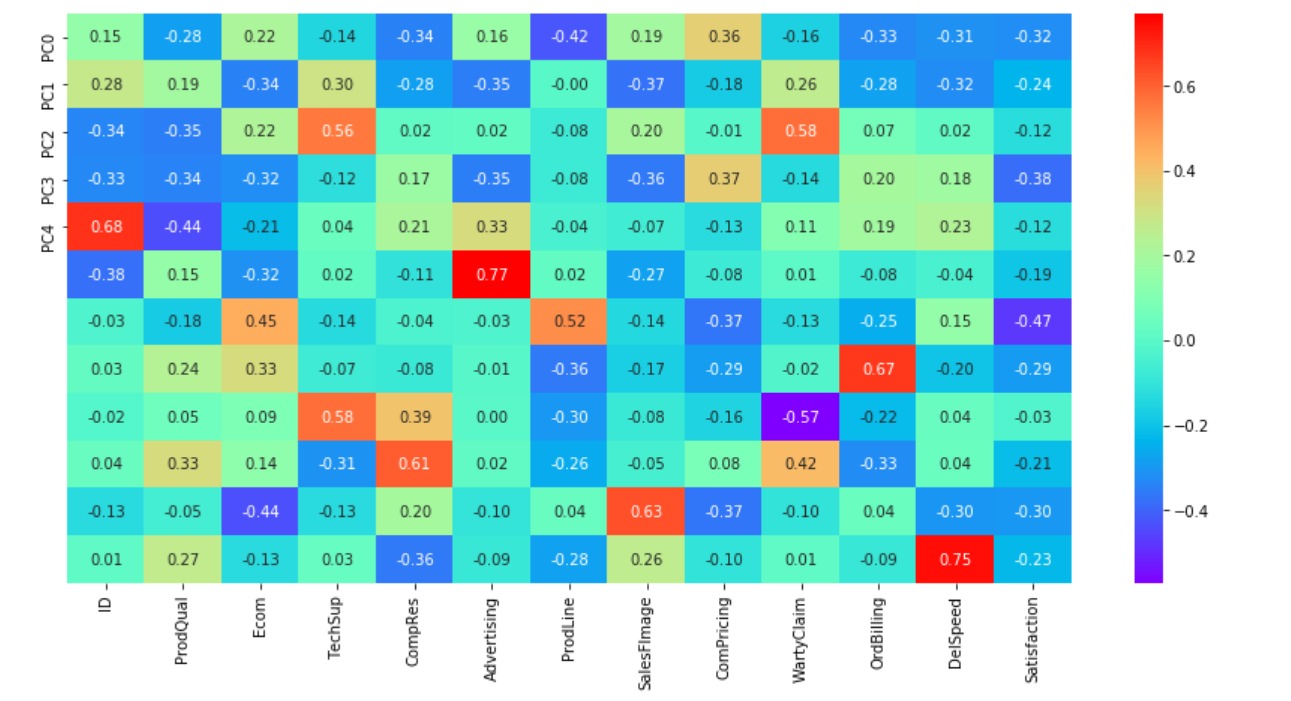




## 8)Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?







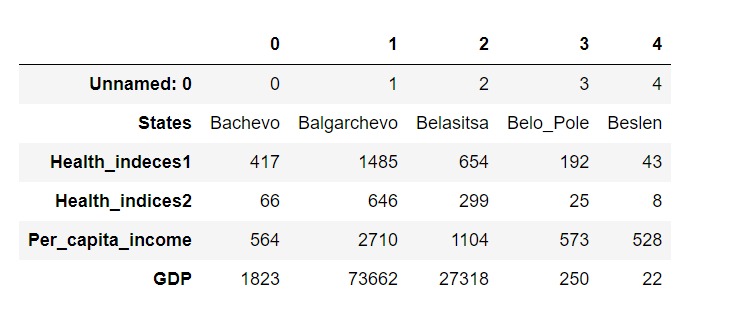
## 8)Mention the business implication of using the Principal Component Analysis for this case study. [Hint: Write Interpretations of the Principal Components Obtained]

#### This business case study is about hair saloon dataset which contain the names of various colleges, which has various details of saloon. To understand more about the dataset we perform univariate analysis and multivariate analysis which gives us the understanding about the variables. From analysis we can understand the distribution of the dataset, skew, and patterns in the dataset. From multivariate analysis we can understand the correlation of variables. Inference of multivariate analysis shows we can understand multiple variables highly correlated with each otsher. The scaling helps the dataset to standardize the variable in one scale. Outliers are imputed using IQR values once the values are imputed we can perform PCA. The principal component analysis is used reduce the multicollinearity between the variables. Depending on the variance of the dataset we can reduce the PCA components. The PCA components for this business case is 5 where we could understand the maximum variance of the dataset. Using the components we can now understand the reduced multicollinearity in the dataset

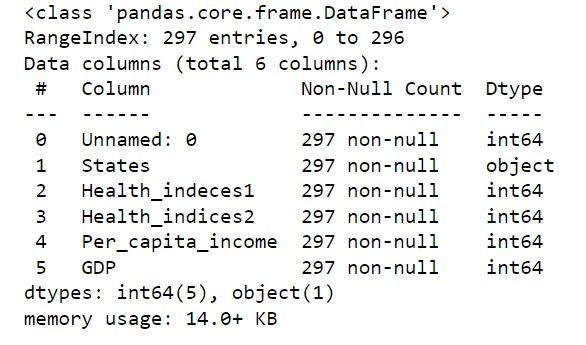
**PART 2: CLUSTERING**

[The](https://olympus.mygreatlearning.com/courses/84255/files/7192425/download?wrap=1)[View in a new window](https://olympus.mygreatlearning.com/courses/84255/files/7192425/download?wrap=1)[State\_wise\_Health\_income.csv](https://olympus.mygreatlearning.com/courses/84255/files/7192425/download?wrap=1)[View in a new window](https://olympus.mygreatlearning.com/courses/84255/files/7192425/download?wrap=1) 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.  
2.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, etc, etc)  
2.2. Do you think scaling is necessary for clustering in this case? Justify  
2.3. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.  
2.4. Apply K-Means clustering on scaled data and determine optimum clusters. Apply the elbow curve and find the silhouette score.  
2.5. 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.  
**Data Dictionary for State\_wise\_Health\_income Dataset:**  
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.

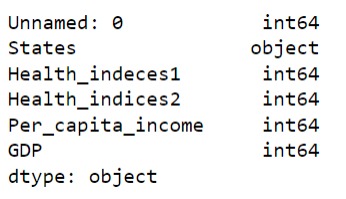
**1)Clustering: Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, etc)**



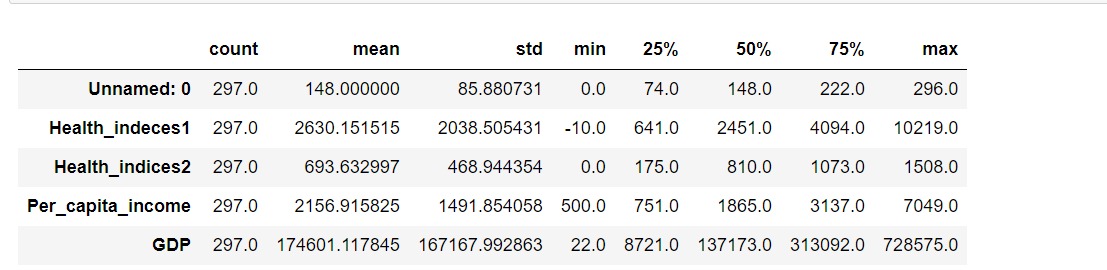
Defined all heads in the dataset.



Information is shown respectively.

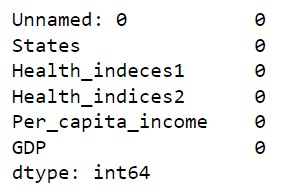


Datatypes shown in the above table.



Describes all the values

There are 0 duplicated values.



Null values are defined.

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

### In most cases yes. But the answer is mainly based on the similarity/dissimilarity function you used in k-means. If the similarity measurement will not be influenced by the scale of your attributes, it is not necessary to do the scaling job.