Revanth Janapriyan

 PGP-DSBA

**Advanced statistics**

Contents

[***PROBLEM 1*** 3](#_Toc115293294)

[executive summary 3](#_Toc115293295)

[Intoduction 3](#_Toc115293296)

[Data description 3](#_Toc115293297)

[**1.1 PERFORM EXPLORATORY DATA ANALYSIS [BOTH UNIVARIATE AND MULTIVARIATE ANALYSIS TO BE PERFORMED]. THE INFERENCES DRAWN FROM THIS SHOULD BE PROPERLY DOCUMENTED.** 4](#_Toc115293298)

[Data sample 4](#_Toc115293299)

[EXPLORATORY DATA ANALYSIS 4](#_Toc115293300)

[**1.2 SCALE THE VARIABLES AND WRITE THE INFERENCE FOR USING THE TYPE OF SCALING FUNCTION FOR THIS CASE STUDY.** 12](#_Toc115293301)

[**1.3 COMMENT ON THE COMPARISON BETWEEN COVARIANCE AND THE CORRELATION MATRIX AFTER SCALING** 14](#_Toc115293302)

[**1.4 CHECK THE DATASET FOR OUTLIERS BEFORE AND AFTER SCALING. DRAW YOUR INFERENCES FROM THIS EXERCISE.** 15](#_Toc115293303)

[**1.5 BUILD THE COVARIANCE MATRIX, EIGENVALUES AND EIGENVECTOR.** 17](#_Toc115293304)

[**1.6 Write the explicit form of the first PC (in terms of Eigen Vectors)** 18](#_Toc115293305)

[**1.7 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](#_Toc115293306)

[**1.8 MENTION THE BUSINESS IMPLICATION OF USING THE PRINCIPAL COMPONENT ANALYSIS FOR THIS CASE STUDY** 20](#_Toc115293307)

[***PROBLEM 2*** 21](#_Toc115293308)

[executive summary 21](#_Toc115293309)

[Intoduction 21](#_Toc115293310)

[Data description 21](#_Toc115293311)

[Data SAMPLE 22](#_Toc115293312)

[EXPLORATORY DATA ANALYSIS 22](#_Toc115293313)

[**2.1 STATE THE NULL AND ALTERNATE HYPOTHESIS FOR CONDUCTING ONE-WAY ANOVA FOR BOTH THE VARIABLES ‘MANUFACTURER’ AND ‘TECHNICIAN' INDIVIDUALLY.** 23](#_Toc115293314)

[**2.2 PERFORM ONE-WAY ANOVA FOR VARIABLE ‘MANUFACTURER’ WITH RESPECT TO THE VARIABLE ‘SERVICE TIME’. STATE WHETHER THE NULL HYPOTHESIS IS ACCEPTED OR REJECTED BASED ON THE ANOVA RESULTS.** 24](#_Toc115293315)

[**2.3** **PERFORM ONE-WAY ANOVA FOR VARIABLE ‘TECHNICIAN’ WITH RESPECT TO THE VARIABLE ‘SERVICE TIME’. STATE WHETHER THE NULL HYPOTHESIS IS ACCEPTED OR REJECTED BASED ON THE ANOVA RESULTS.** 25](#_Toc115293316)

[**2.4** **ANALYSE THE EFFECTS OF ONE VARIABLE ON ANOTHER WITH THE HELP OF AN INTERACTION PLOT. WHAT IS AN INTERACTION BETWEEN TWO TREATMENTS?** 27](#_Toc115293317)

[**2.5 PERFORM A TWO-WAY ANOVA BASED ON THE VARIABLES ‘MANUFACTURER’ & ‘TECHNICIAN’ WITH RESPECT TO THE VARIABLE ‘SERVICE TIME’ AND STATE YOUR RESULTS.** 28](#_Toc115293318)

[**2.6 MENTION THE BUSINESS IMPLICATIONS OF PERFORMING ANOVA FOR THIS PARTICULAR CASE STUDY.** 28](#_Toc115293319)

# PROBLEM 1

## **executive summary**

The Hair Salon 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 following rubric.

## **Intoduction**

In this project, we have taken up various steps to determine the best components which are contributing in marketing of various hair products of the client. Exploration of the dataset have been done which includes looking at the shape of the dataset, checking for missing values, checking for outliers and imputing them, univariate analysis, bivariate analysis and graphical representations. Apart from data exploration, methods such as data scaling, comparing covariance and correlation, etc have been carried out in order to find out the best principal components of the given problem.

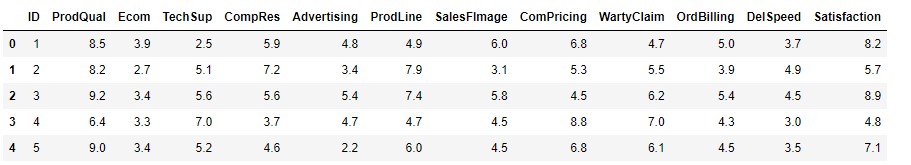
## **Data description**

* **ProdQual:** Contain information about product quality.
* **Ecom:** Contains information about E-commerce.
* **TechSup:** Contains information about Tech Support.
* **CompRes:** Contains information about Complaint Resolution.
* **Advertising:** Advertising
* **ProdLine:** Contains information about Product Line.
* **SaleFImage:** Contains information about Salesforce Image.
* **CompPricing:** Contains information about Competitive Pricing.
* **WartyClaim:** Warranty & Claims.
* **OrdBilling:** Contains information about Order & Billing.
* **DelSpeed:** Delivery speed.
* **Satisfaction:**  Customer Satisfaction.

### 1.1 PERFORM EXPLORATORY DATA ANALYSIS [BOTH UNIVARIATE AND MULTIVARIATE ANALYSIS TO BE PERFORMED]. THE INFERENCES DRAWN FROM THIS SHOULD BE PROPERLY DOCUMENTED.

## **Data sample**

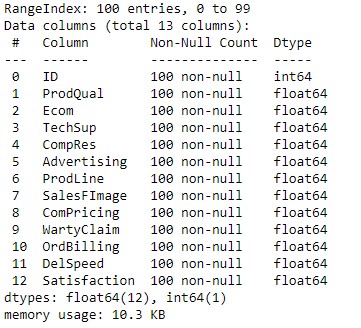
The below table shows the variables and their first five values.



**Table1.1**

There are 13 features available in the given dataset.

## **EXPLORATORY DATA ANALYSIS**

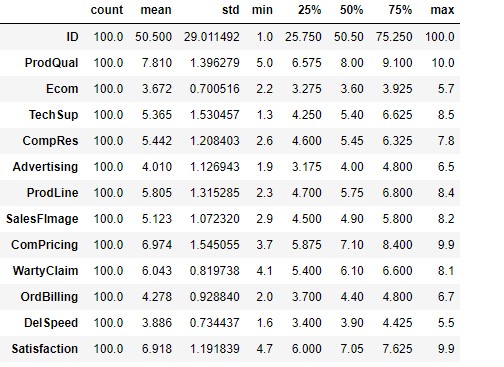


**Table 1.2**

The table 1.2, shows the features of the dataset along with its data type and null value counts of each variable. There are total of 12 float variables and 1 integer variable.

Descriptive Analysis

Now, let us see the descriptive statistics of the given dataset using the below table.



**Table 1.3**

The table 1.3 provides various information about mean, median, standard deviation, minimum and maximum value, first quartile and third quartile of each feature of the dataset.

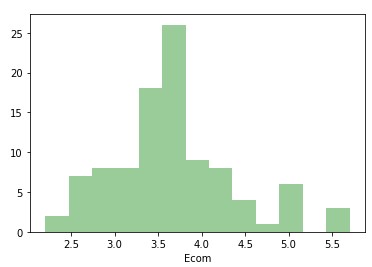
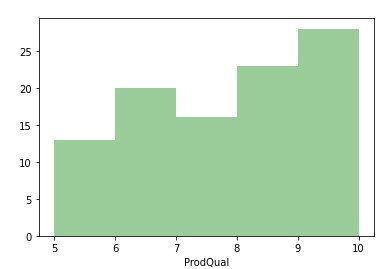
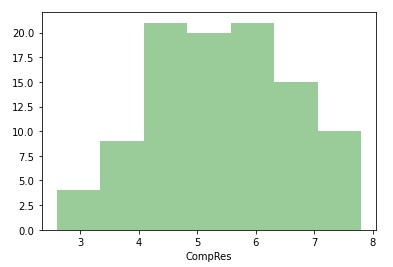
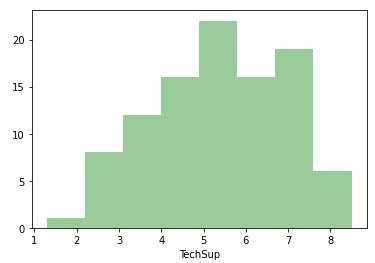
There are no null values present in the variables of the given dataset (Refer table1.4)



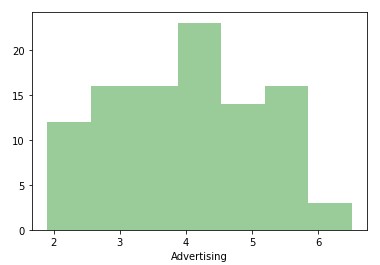
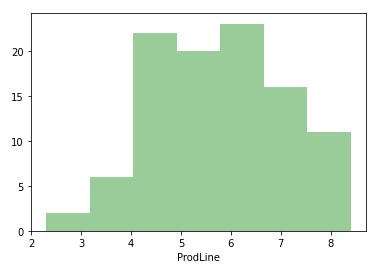
**Table 1.4**

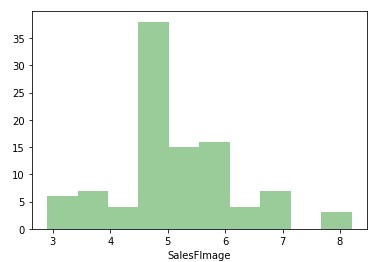
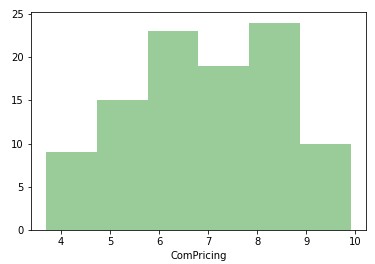
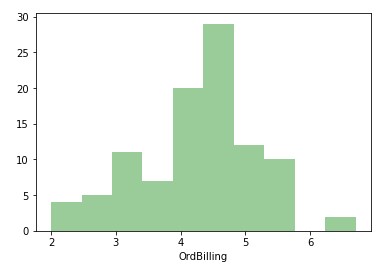
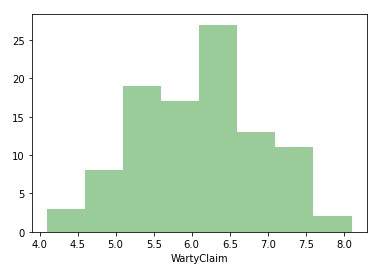
UNIVARIATE Analysis

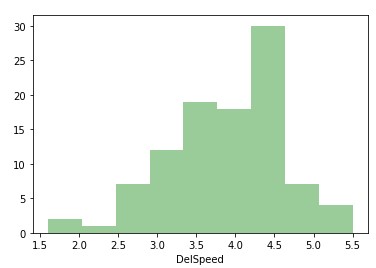
Univariate Analysis for each variable present in the dataset has been visualized below.



**Fig 1.1**

****

**   
**

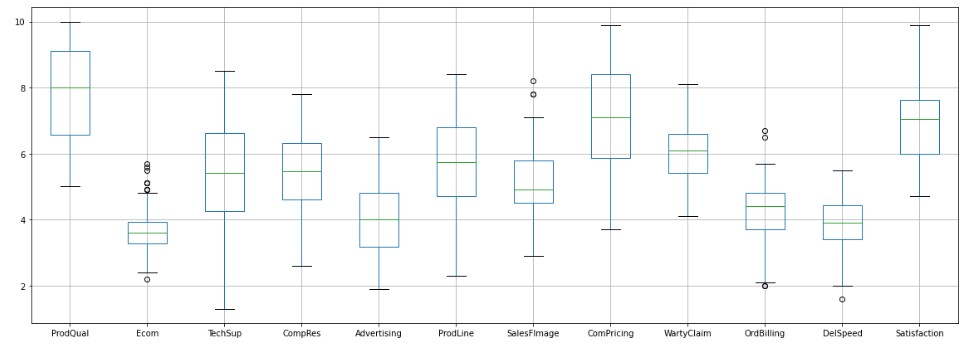


**Fig 1.2**

* From the fig 1.2, it can be concluded that none of the variable present in the dataset is normally distributed.
* Tech support, Delivery speed and Product Quality are Left skewed.
* Rest of the variables are distributed in different manners such that they are neither right or left skewed.

OUTLIERS

The outliers present in each variable has also been visualised using the boxplot below.

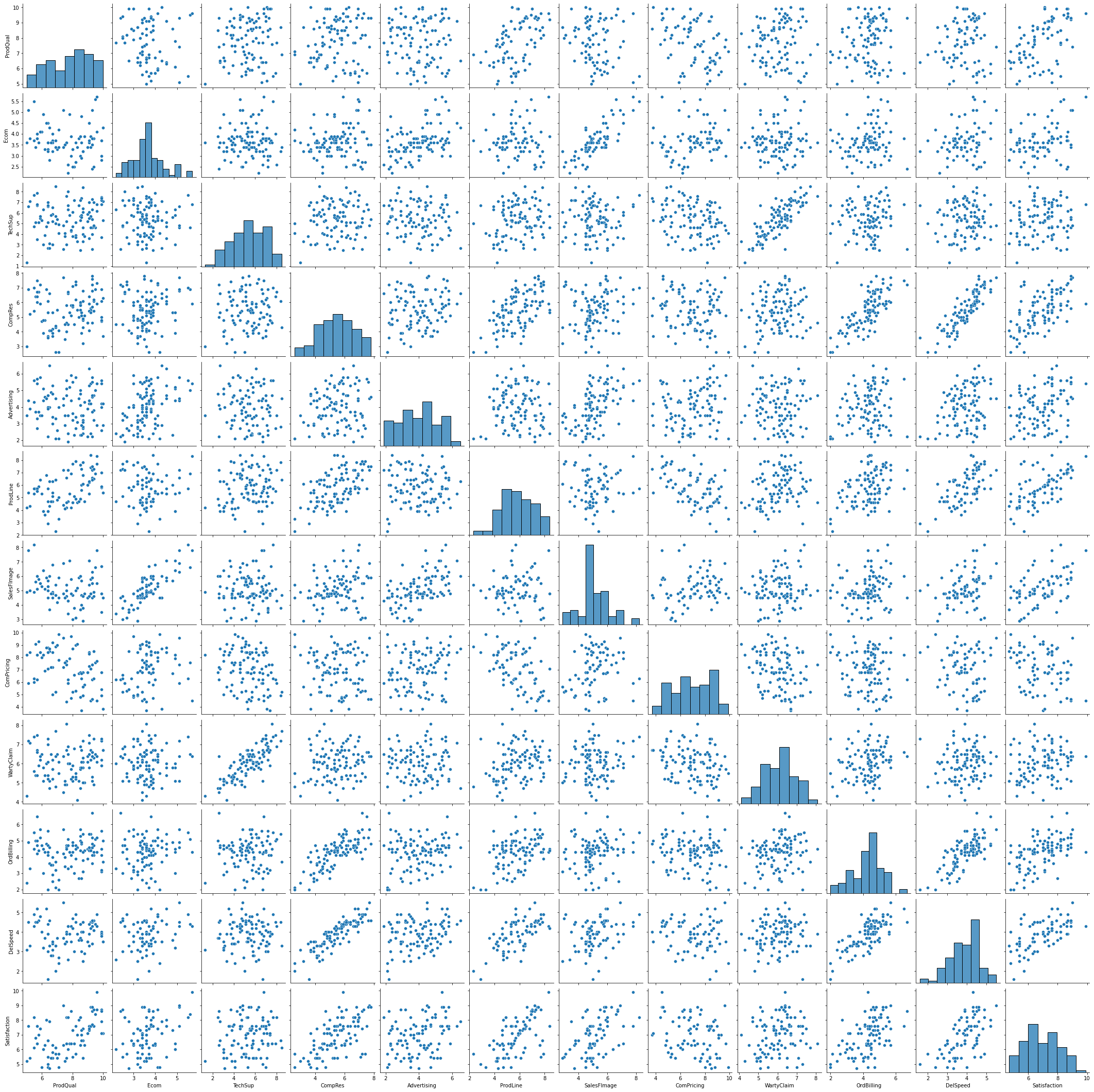


**Fig 1.3**

Above Fig 1.3, shows that the variables Ecom, SalesFImage, OrdBilling and Delspeed has outliers.

PAIRPLOT

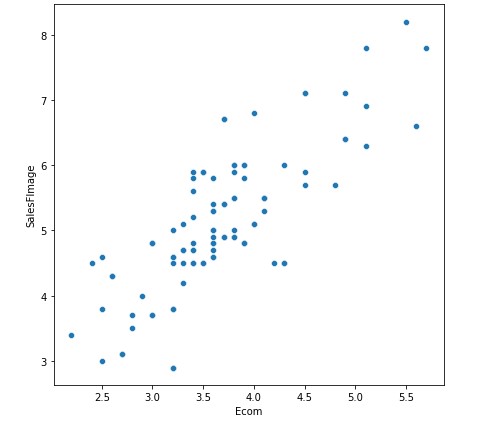
Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

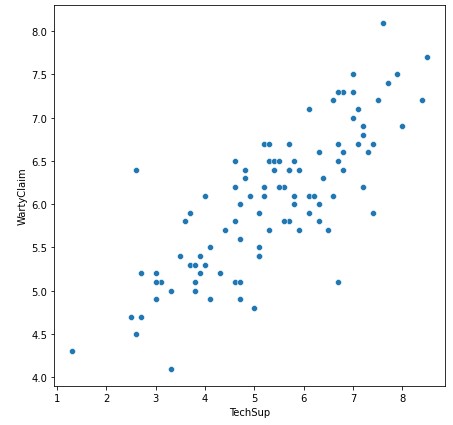


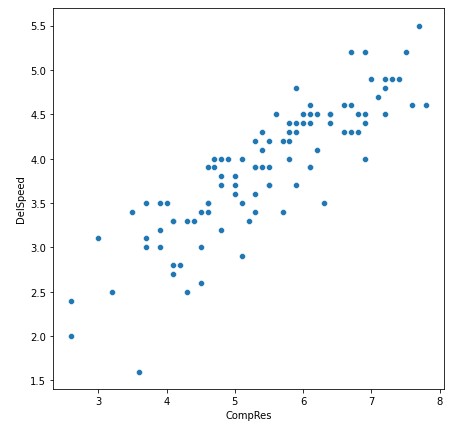
**Fig 1.4**

CoRRELATION

As seen in the pairplot previous, only feel variables are correlated with each other. Let us zoom in a bit and see which variables are correlated using below figure.





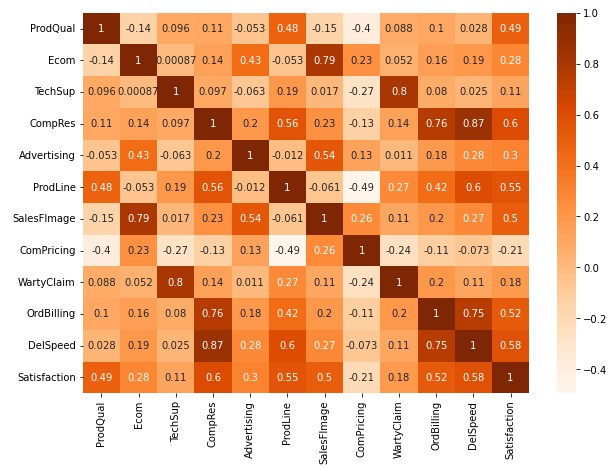
****

**Fig 1.5**

The Fig 1.5, visualize the positive correlation between the variables SalesFImage and Ecom, WartyClaim and TechSup, Delspeed and CompRes.

HEATMAP

The below figure shows the correlation of each variable in the form of heatmap.

****

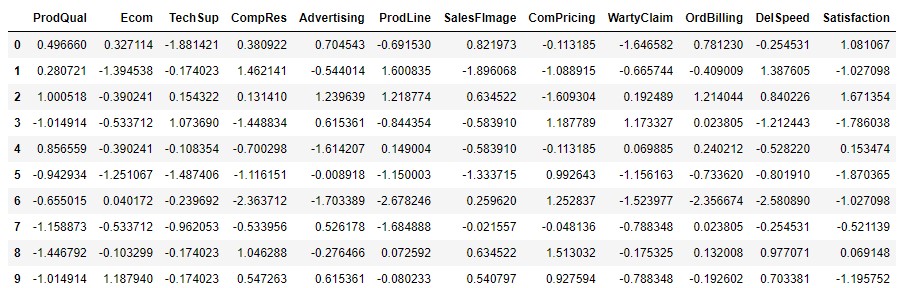
**Fig 1.6**

As we saw earlier in Fig 1.5, the heatmap also adds up that there is a positive correlation between only six features present in the dataset. All other features don’t have a good correlation between them.

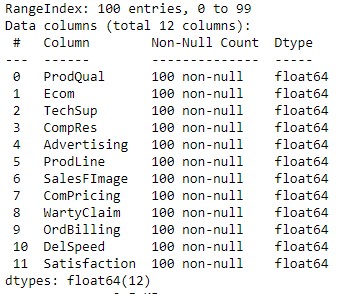
### 1.2 SCALE THE VARIABLES AND WRITE THE INFERENCE FOR USING THE TYPE OF SCALING FUNCTION FOR THIS CASE STUDY.

Scaling is a technique to make them closer to each other or in simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower. Scaling data makes it easy for a model to learn and understand the problem.

Below tables table 1.5, table 1.6 shows the scaled dataset and the datatypes of each scaled feature in the dataset. StandardScaler is used to scale data here.



**Table 1.5**



**Table 1.6**

Table 1.6 shows that the datatypes of all the features in dataset is changed to float datatype after scaling the dataset.

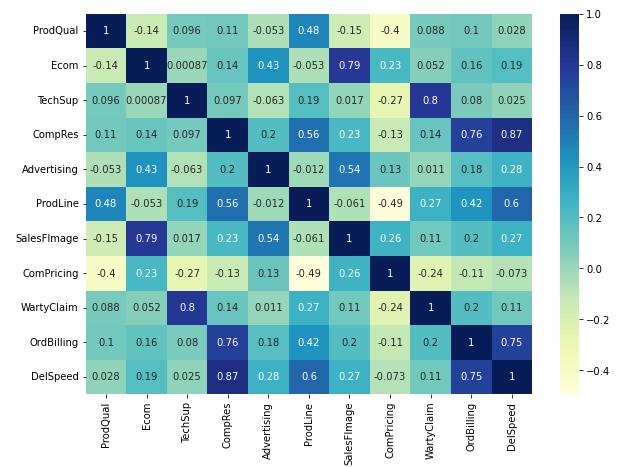
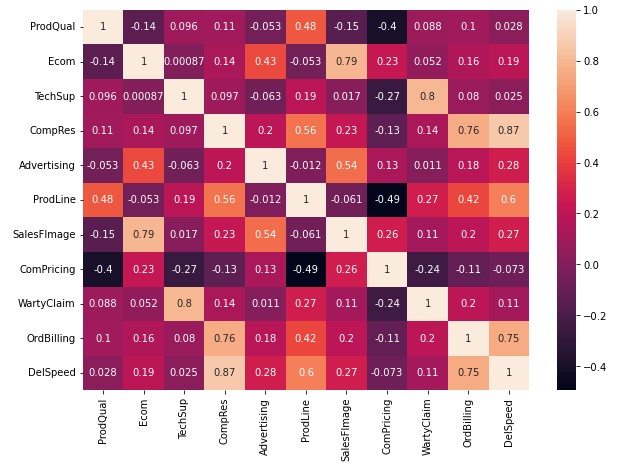


**Table 1.7**

Table above shows the descriptive statistics of the scaled data. We can see that the standard deviation among all the features is same. Thus, all the datapoints are normalized using the Scaling technique.

### 1.3 COMMENT ON THE COMPARISON BETWEEN COVARIANCE AND THE CORRELATION MATRIX AFTER SCALING

Covariance and correlation are two terms that are opposed and are both used in statistics. Covariance shows you how the two variables differ, whereas correlation shows you how the two variables are related. Keeping the definition, let us look at the heatmaps of covariance matrix and correlation matrix to see if there is any difference.

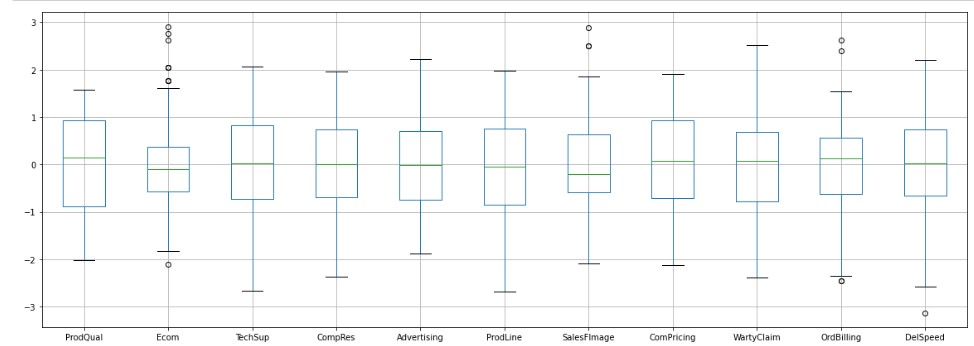


**Fig 1.7**

The fig 1.7, shows the heatmap of covariance and correlation matrix. Comparing both heatmaps of covariance and correlation matrix there no visible difference can be inferred.

### 1.4 CHECK THE DATASET FOR OUTLIERS BEFORE AND AFTER SCALING. DRAW YOUR INFERENCES FROM THIS EXERCISE.

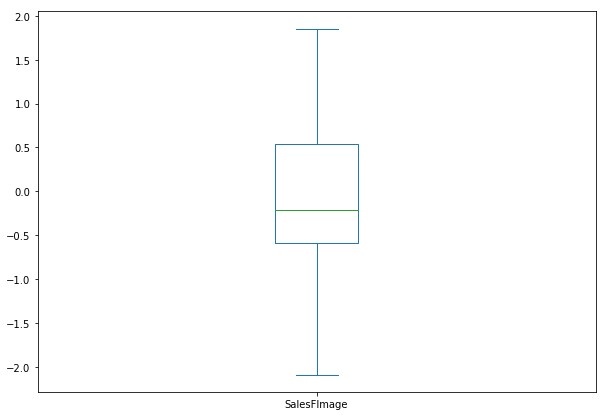
As, we know that only the Ecom, SalesFImage, OrdBilling and DelSpeed has outliers from the previous data exploration done using unscaled data. Let us visualize the same for the scaled data and see if there is any outliers present.

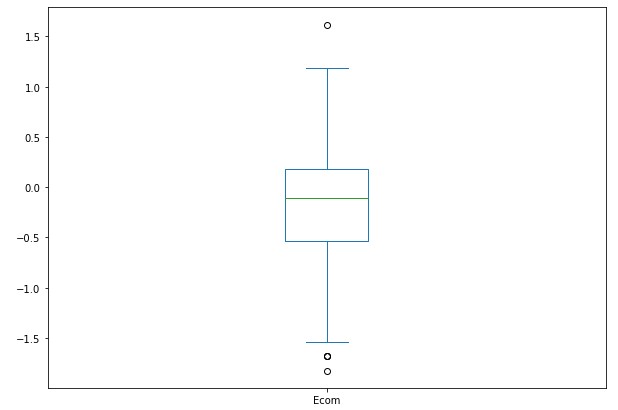


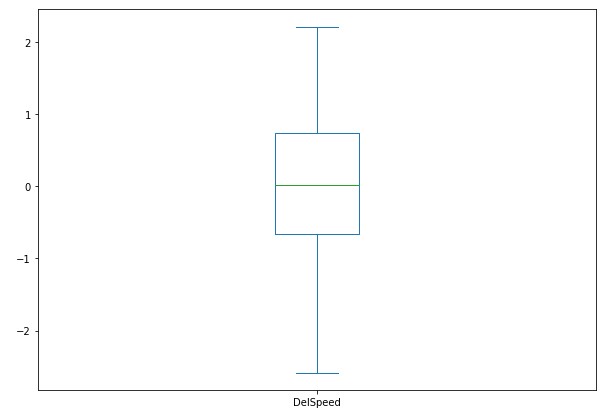
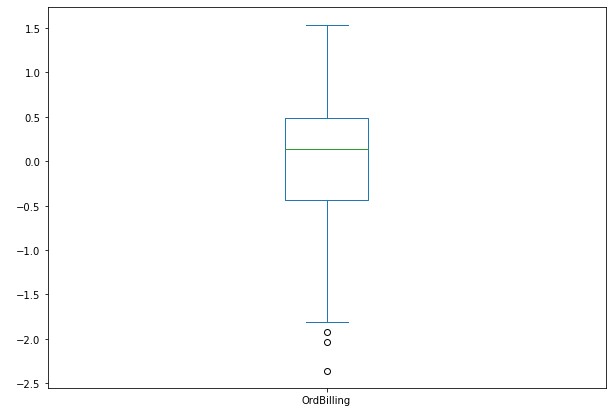
**Fig 1.8**

Similar to the unscaled dataset, Ecom, SalesFImage, OrdBilling and DelSpeed are the features with outliers in the scaled dataset too. (Refer fig1.8)

Necessary steps have been taking to treat the outliers such as imputing with upper value and lower value. The treated features are again visualized and it can be seen that the outliers present in the features Ecom, SalesFImage, OrdBilling and DelSpeed are treated. (Refer fig 1.9)



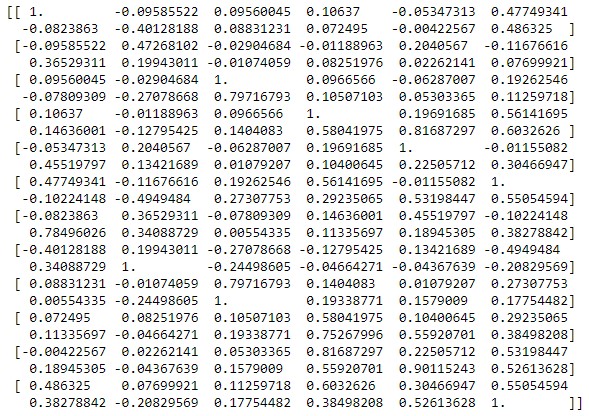


  
**Fig 1.9**

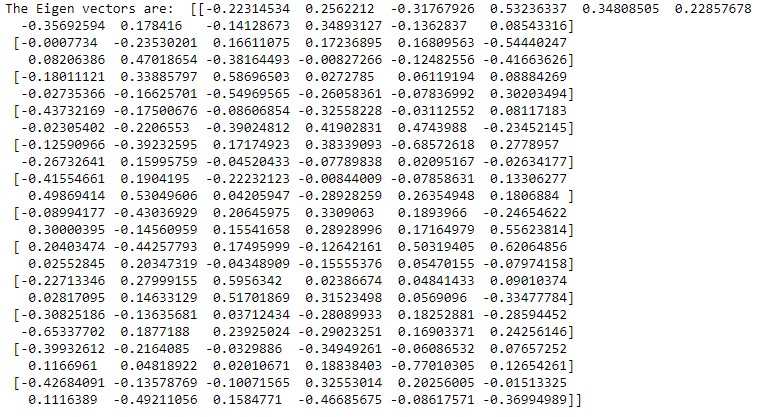
### 1.5 BUILD THE COVARIANCE MATRIX, EIGENVALUES AND EIGENVECTOR.

The covariance matrix, eigen values and eigen vectors are build for the scaled dataset.

* The covariance matrix of the scaled dataset is given below.



* The eigen values of the scaled dataset are 3.03969016, 2.16167412, 1.50614006, 1.03871336, 0.63648233, 0.42590603, 0.39890628, 0.10400615, 0.13491754, 0.25156462, 0.213473.
* The Eigen vectors are given below.

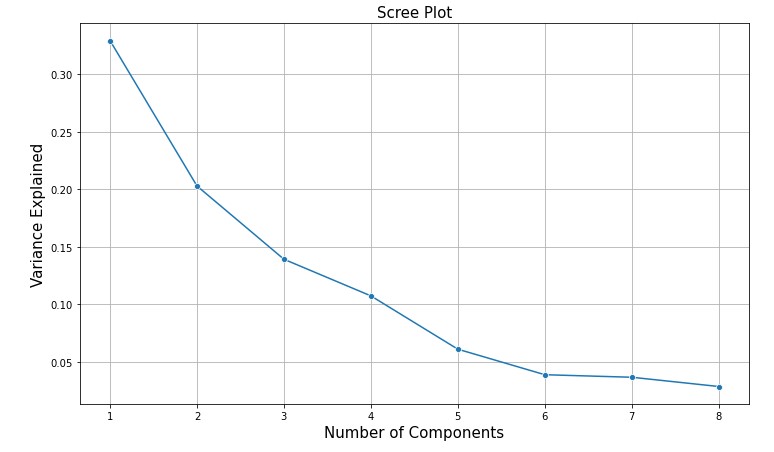


### 1.6 Write the explicit form of the first PC (in terms of Eigen Vectors)

* Before we proceed further, we need to check the significance of correlation and adequacy of the features using bartlett sphericity test and kmo.
* Barlett sphericity test is carried out assuming the null hypothesis as all the features are significantly correlated and the alternative hypothesis as none of them are correlated. From the result of the barlett sphericity test, null hypothesis is rejected as the p-value is less than 0.5.
* Similarly, kmo is done. Since MSA is close to 0.7, we can expect a considerable reduction in the dimension and extraction of meaningful components to some extent.
* The pca components are found and using which explained variance and ratio of explained variance of the principal components.
* The explained variance of principal components is 3.0703941, 2.18350922, 1.5213536, 1.04920541, 0.64291145, 0.43020811, 0.40293564, 0.25410568.
* The ratio of explained variance of principal components is 0.30668398, 0.21809816, 0.15195925, 0.10479908, 0.06421672, 0.04297101, 0.04024692, 0.02538115.
* From the explained variance and ratio of the explained variance it is clearly evident that first 5 components explains the variance better.

### 1.7 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.

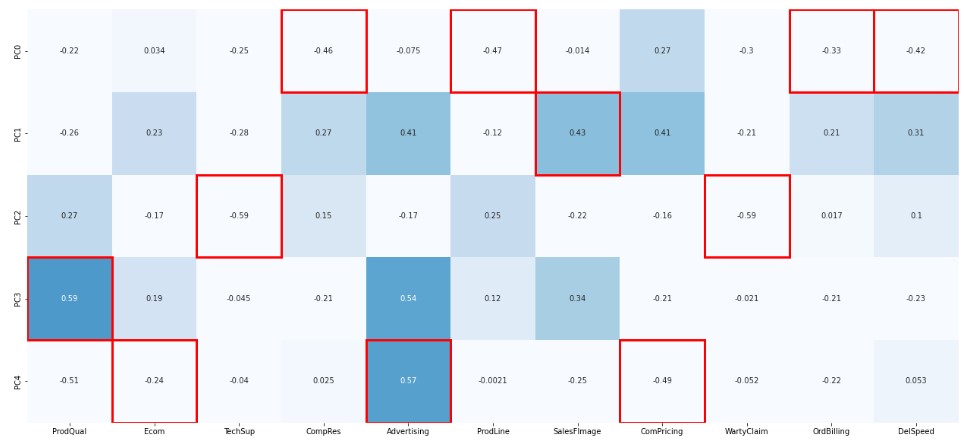
* The Cumulative values of eigenvalues shows that almost 80% of the variance is explained by the first 5 principal components.
* The rest of the eigenvalues accounts for the variance which is lesser than what these 5 principal components accounts for.
* Thus, it can play a major part in deciding 5 principal components as optimum number of principal components for the given analysis. The scree plot below in the fig 1.10 shows the pca.



**Fig 1.10**

The Eigenvector indicate the direction across which the data is spread most.

Upon performing pca, the principal components have been identified and are visualized.



**Fig 1.11**

The figure above shows the 5 principal components which explained maximum variance across each feature.

### 1.8 MENTION THE BUSINESS IMPLICATION OF USING THE PRINCIPAL COMPONENT ANALYSIS FOR THIS CASE STUDY

The pca has fetched us various conclusion such as

* The Ecom increase with the Sales Force Image increase, which clearly shows that more transactions can be made with better Sales force Image.
* Similarly, the warranty claims, tech support, complaint resolution and delivery speed are good among the all the aspects considered. So, it is time to concentrate on other features which can propel the business further in the right path.
* Advertising, Product Quality are some of the important features which has to be given more importance in the future.

# PROBLEM 2

## **executive summary**

The staff of a service centre for electrical appliances include three technicians who specialize in repairing three widely used electrical appliances by three different manufacturers. It was desired to study the effects of Technician and Manufacturer on the service time. Each technician was randomly assigned five repair jobs on each manufacturer's appliance and the time to complete each job (in minutes) was recorded.

## **Intoduction**

In this problem, the dataset is given is about technicians of a service centre who are specialized in repairing three widely used electrical appliances by three different manufacturers. The dataset has been analysed using descriptive analysis and made ready for working on Anova. Both One way Anova and two way Anova are used to find solution for this problem.

## **Data description**

The given dataset contains the following columns and their descriptions are provided below:

1. **Technician** – Contains technicians group information.
2. **Manufacturer** – Contains information about the different manufacturers.
3. **Job** – Contains information about types of jobs.
4. **Service\_Time** – Contains service time taken for each job by each technician.

## **Data SAMPLE**

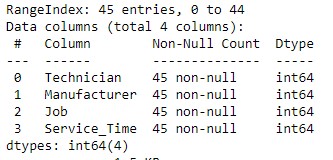
The sample of the given dataset is shown in the table below.



**Table 2.1**

## **EXPLORATORY DATA ANALYSIS**

Let us check for missing values and data types of the features in the dataset before proceeding further.

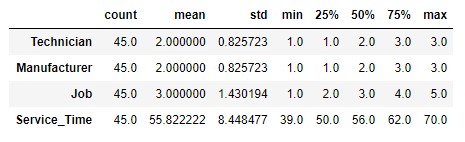


**Table 2.2**

The dataset contains four features and forty-five non-null values in each feature. All the features present in the dataset set has integer as datatype. Also, upon checking the unique counts of each feature, it is found that each feature – Technician, Manufacturer and Job has three different category and there are 15 entries for those three categories under each feature.

Descriptive Analysis

Now, let us see the descriptive statistics of the given dataset using the below table.

****

**Table 2.3**

From the table 2.3, we can see that the average time taken for a single service is 55.8 minutes and the longest is 70 minutes.

### 2.1 STATE THE NULL AND ALTERNATE HYPOTHESIS FOR CONDUCTING ONE-WAY ANOVA FOR BOTH THE VARIABLES ‘MANUFACTURER’ AND ‘TECHNICIAN' INDIVIDUALLY.

1. The null and alternative hypothesis for conducting one-way anova for the variable ‘Manufacturer’ is given below:

* **Null Hypothesis:** The mean service time taken for each manufacturer is same.
* **Alternative Hypothesis:** The mean service time taken for at least one manufacturer is different.

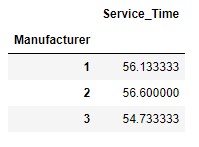
1. The null and alternative hypothesis for conducting one-way anova for the variable ‘Technician’ is given below:

* **Null Hypothesis:** The mean service time taken for all three technician group is same.
* **Alternative Hypothesis:** The mean service time taken for at least one technician group is different.

### 2.2 PERFORM ONE-WAY ANOVA FOR VARIABLE ‘MANUFACTURER’ WITH RESPECT TO THE VARIABLE ‘SERVICE TIME’. STATE WHETHER THE NULL HYPOTHESIS IS ACCEPTED OR REJECTED BASED ON THE ANOVA RESULTS.

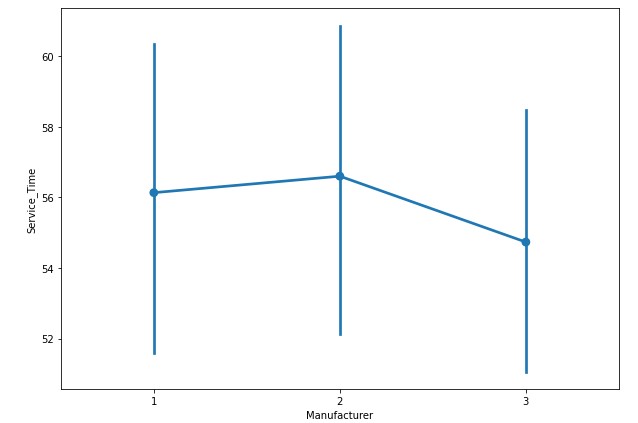
One-way anova for variable ‘Manufacturer’ with respect to the variable ‘Service Time’ has been performed and the results are given below.

The mean of the groups of variable ‘Manufacturer’ with respect to ‘Service Time’ are show in the below table.



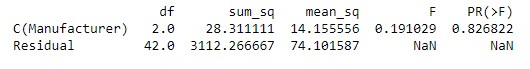
**Table 2.4**

The mean of all the three groups is almost equal expect the third group slightly coming short. The figure 2.1 below also visualizes the mean of the three groups.



**Fig 2.1**

Now let us have a look at the p-value using the table 2.5 below.



**Table 2.5**

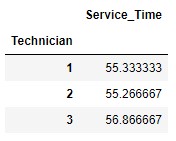
The table 2.5, shows that the p-value is significantly greater than 0.5.

Since, the mean value of the groups are almost same and the p-value is greater than 0.5, the null hypothesis cannot be rejected and it remains true here.

### PERFORM ONE-WAY ANOVA FOR VARIABLE ‘TECHNICIAN’ WITH RESPECT TO THE VARIABLE ‘SERVICE TIME’. STATE WHETHER THE NULL HYPOTHESIS IS ACCEPTED OR REJECTED BASED ON THE ANOVA RESULTS.

One-way anova for variable ‘Technician’ with respect to the variable ‘Service Time’ has been performed and the results are given below.

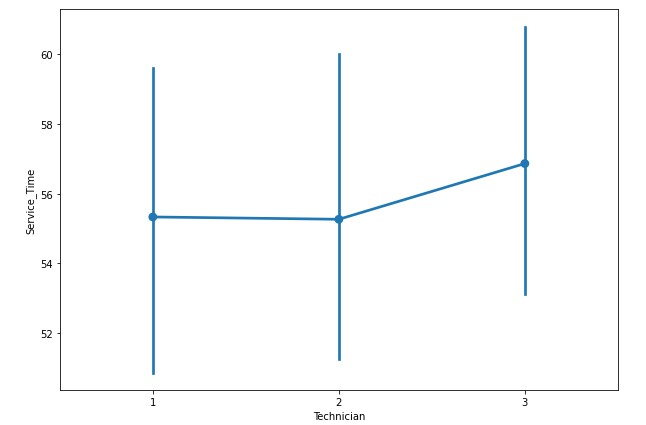
The mean of the groups of variable ‘Technician’ with respect to ‘Service Time’ are show in the below table.



**Table 2.6**

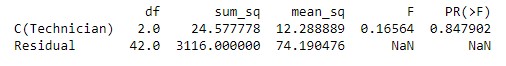
The mean of all the three groups is almost equal expect the third group slightly coming high.

Let us visualize the mean using the pointplot and see how it is spread. The figure below also helps us understand visually that all the three means are closer.



**Fig 2.2**

Now let us have a look at the p-value using the table 2.7 below.

****

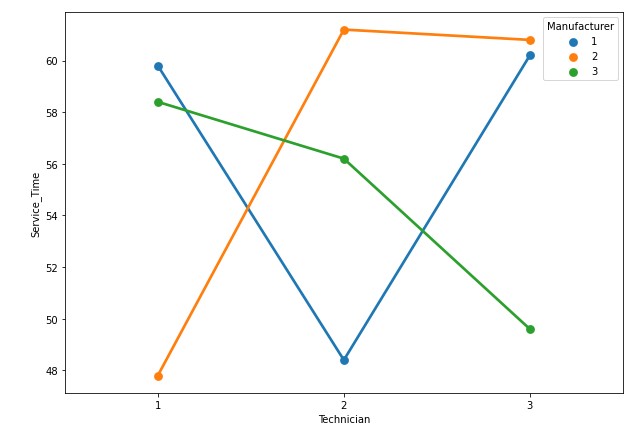
**Table 2.7**

The p-value shown in the table 2.7 is higher than the significance level 0.5.

From, the one-way anova performed above, the null hypothesis remains true as the p-value is greater than 0.5 and also the mean of all the three technician groups is closer to each other.

### ANALYSE THE EFFECTS OF ONE VARIABLE ON ANOTHER WITH THE HELP OF AN INTERACTION PLOT. WHAT IS AN INTERACTION BETWEEN TWO TREATMENTS?

The effect of one variable on another is visualized in the figure 2.3 using the interaction plot.



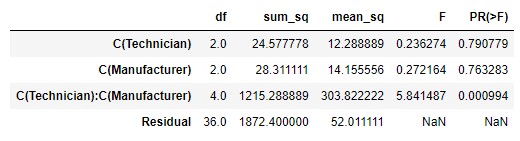
**Fig 2.3**

Figure 2.3 shows the interaction among each variable and their effect on each other. It can be clearly seen that the service time for each technician group varies with respect to the manufacturer.

* Technician group 1 finishes the work quicker than others when it is manufacturer 2 and has a high service time for manufacturer 1.
* Likewise, technician group 2 has lesser service time for manufacturer 1 and vice versa for manufacturer 2.
* Technician group 3 completes the service quicker when it is manufacturer 3 when compared to other two manufacturer.

### 2.5 PERFORM A TWO-WAY ANOVA BASED ON THE VARIABLES ‘MANUFACTURER’ & ‘TECHNICIAN’ WITH RESPECT TO THE VARIABLE ‘SERVICE TIME’ AND STATE YOUR RESULTS.

Two-way anova has been performed based on the variables ‘Manufacturer’ and ‘Technician’ with respect to ‘Service Time’. The result of the two-anova performed is displayed in the table below.



**Table 2.8**

From the table above, it can be concluded that when the variables technician and manufacturer are considered separately the variance is not that much. But when both variables are considered as a combination, the p-value is lesser than 0.5 which explains that there is variance among them.

### 2.6 MENTION THE BUSINESS IMPLICATIONS OF PERFORMING ANOVA FOR THIS PARTICULAR CASE STUDY.

Upon performing the one-way and two-anova, variation business implications can be made:

* Since, certain technicians have shorter service time with respect to certain manufacturer the service centre can make sure that the exact technicians are assigned with the manufacturer the perform well. This will help in achieving faster service.
* Cross-training session for every technician can be organised by the service centre collaborating with each manufacturer. This will ensure that all the 45 technicians are trained to service all the appliances manufactured by all three manufacturers, which in turn helps in achieving a balanced work distribution among the group.
* Since the dataset provided contains less data about the service time, technician and manufacturer, a detailed study can be done in order to analysis why a particular group takes more time while servicing a specific manufacturer’s appliance. This will helps us figure out the challenges faced by each technician while servicing different manufacturer’s appliance and provided necessary support/training to overcome the challenges.