A Project Report on PCA & ANOVA

By

Soumya Probhat Roy

Table Of Content

ANOVA

1.1 Problem Statement

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. The data for this particular experiment is thus attached.

Questions:

1) State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables 'Manufacturer' and 'Technician individually. – 3 points

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. - 3 points

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. - 3 points

4) Analyse the effects of one variable on another with the help of an interaction plot. What is an interaction between two treatments?
[hint: use the 'pointplot' function from the 'seaborn' graphical subroutine in Python] - 4 points

5) Perform a two-way ANOVA based on the variables 'Manufacturer' & 'Technician' with respect to the variable 'Service Time' and state your results. - 5 points

6) Mention the business implications of performing ANOVA for this particular case study. - 5 Points

Null & Alternate Hypothesis

Null and Alternate hypothesis with respect to Manufacturer

H0 = There is no significance difference in service time means due to Manufacturer

H1 = There is significance difference in service time means due to Manufacturer

Null and Alternate hypothesis with respect to Technician

H0 = There is no significance difference in service time means due to Technician

H1 = There is significance difference in service time means due to Technician

One Way Anova (Manufacturer Vs Service Time)

```
                    df     sum_sq   mean_sq         F    PR(>F)
C(Service_Time)   24.0  14.333333  0.597222  0.762411  0.739253
Residual          20.0  15.666667  0.783333       NaN       NaN
```

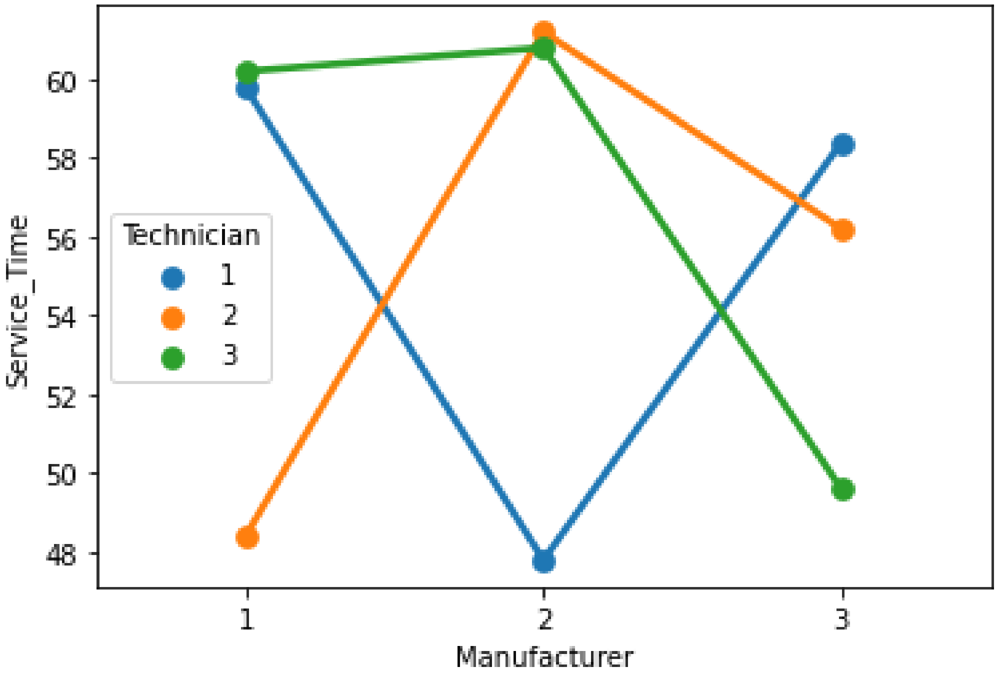
Because the P value is higher than 0.05, the null hypothesis may be ruled out with a 95% confidence level. Regarding the manufacturer, there is a significant difference in service time methods.

One Way Anova (Technician Vs Service Time)

```
                    df     sum_sq   mean_sq         F    PR(>F)
C(Service_Time)   24.0  14.166667  0.590278  0.745614  0.755742
Residual          20.0  15.833333  0.791667       NaN       NaN
```

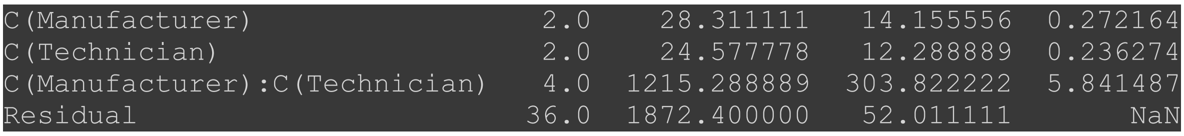
So the P value is higher than 0.05, the null hypothesis may be ruled out with a 95% confidence level. There is a significant difference in the technicians' service time means.

Effect of One Variable Over Other

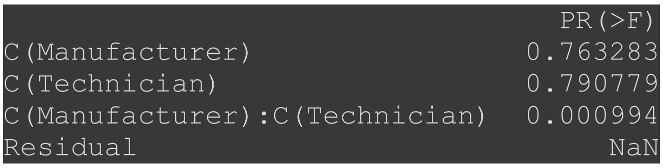


When the lines are parallel then there will be interaction. The above plot shows that there are no interaction between the two variables.

Two Way Anova (Manufacturer, Technician Vs Service Time)



```
C(Manufacturer)                  2.0    28.311111   14.155556  0.272164
C(Technician)                    2.0    24.577778   12.288889  0.236274
C(Manufacturer):C(Technician)    4.0  1215.288889  303.822222  5.841487
Residual                        36.0  1872.400000   52.011111       NaN
```



```
                                 PR(>F)
C(Manufacturer)                0.763283
C(Technician)                  0.790779
C(Manufacturer):C(Technician)  0.000994
Residual                            NaN
```

Because the combination's P value is less than 0.05, we might have to accept the null hypothesis involving two variables. Accordingly, there is no appreciable change in the service time means for the Technicians and Manufacturer collectively.

Business Implementation

The investigation reveals that the impact of the manufacturer and technicians on the mean service time is minimal. But for a specific manufacturer, there can be some professionals who are more effective and require less maintenance. The two-way Anova explains this. In order to have an optimised mean service time, the appropriate technicians must be assigned to the appropriate manufacturer.

2 PCA

2.1 Problem Statement

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 following rubric.

Note: This particular dataset contains the target variable satisfaction as well. Please do drop this variable before doing Principal Component Analysis.

Questions:

1) Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented. – **5 points**

2) Scale the variables and write the inference for using the type of scaling function for this case study. - **3 points**

3) Comment on the comparison between covariance and the correlation matrix after scaling. - **2 points**

4) Check the dataset for outliers before and after scaling. Draw your inferences from this exercise. - **3 points**

5) Build the covariance matrix, eigenvalues and eigenvector. - **4 points**

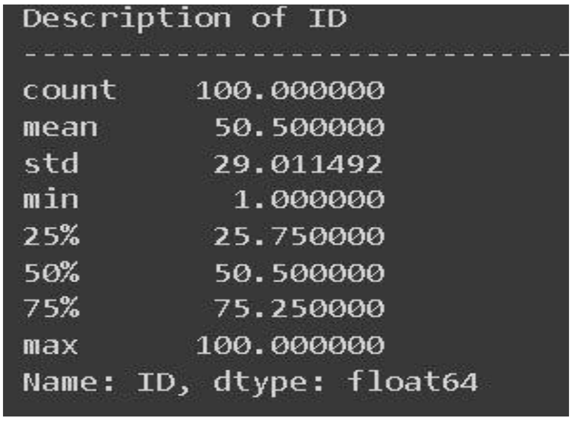
6) Write the explicit form of the first PC (in terms of Eigen Vectors) – **5 points**

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. – **10 points**

8) Mention the business implication of using the Principal Component Analysis for this case study. – **5 points**

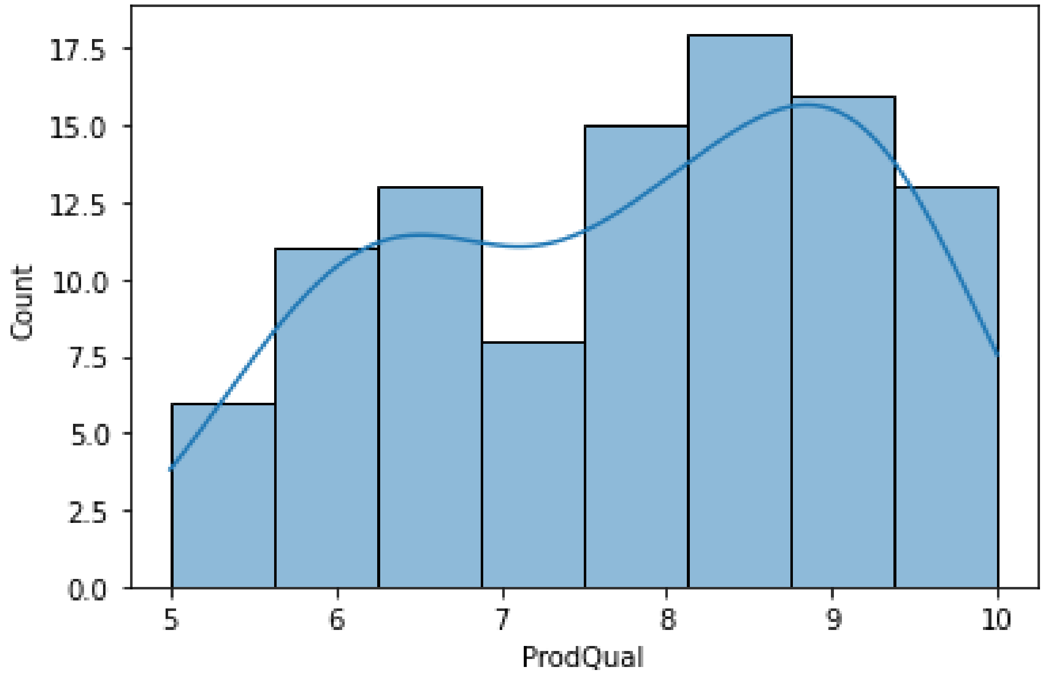
2.2 Explanatory Data Analysis

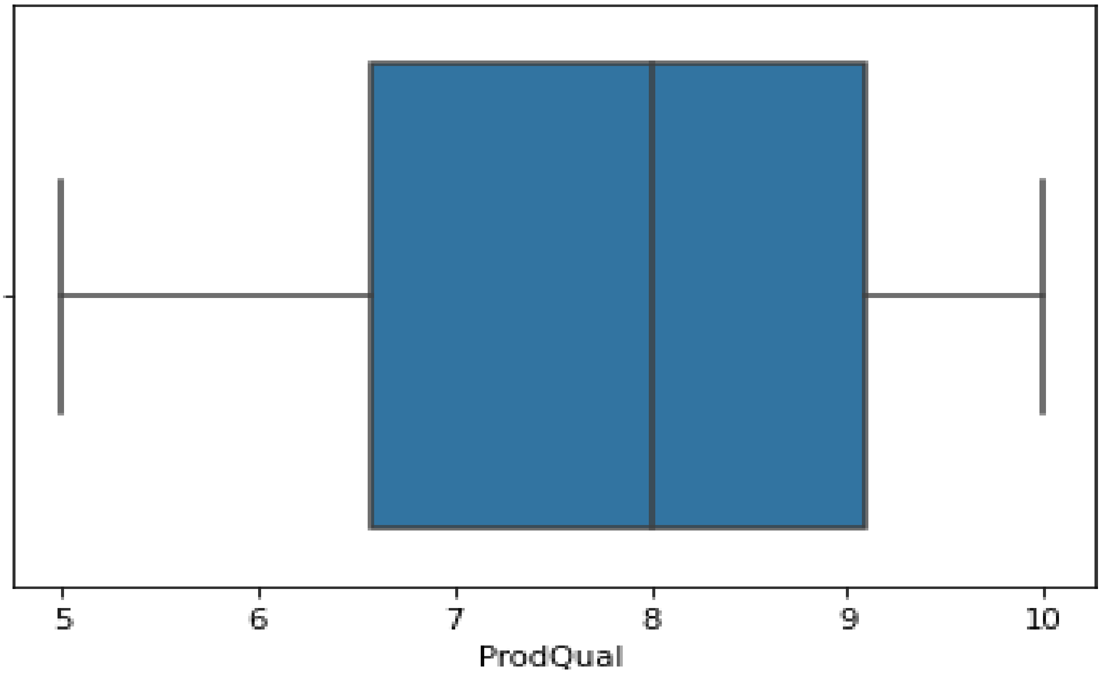
- There are 100 rows and 13 columns
- There is target variable satisfaction and ID column which are dropped for the analysis purpose,
- All data types are float data type
- There are no duplicates present in the dataset
- There are no null values present in the data set.
- Basic statistics of the dataset is as follows



```
Description of ID
-----------------------------
count    100.000000
mean      50.500000
std       29.011492
min        1.000000
25%       25.750000
50%       50.500000
75%       75.250000
max      100.000000
Name: ID, dtype: float64
```

Univariate Analysis

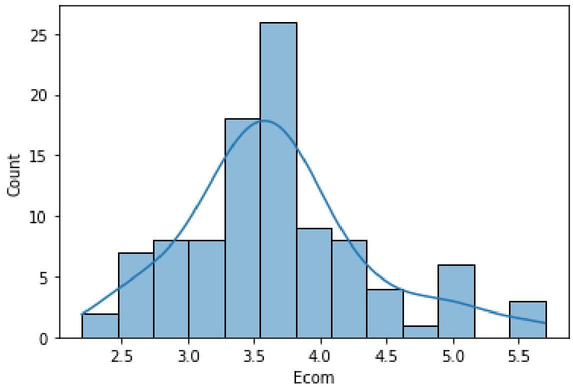


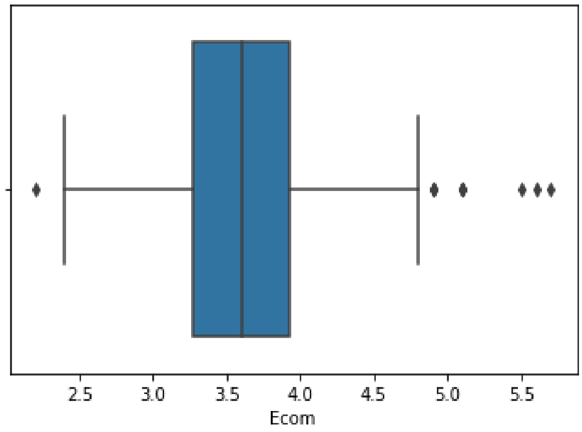


With the univariate analysis we shall understand the distribution of data & find out the pattern.

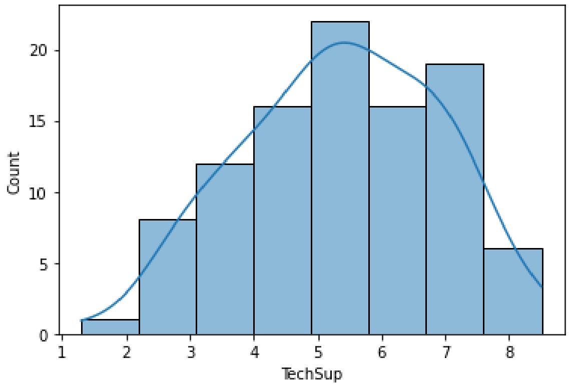
The ‘product quality’ ranges between 6.5 and 9.2. There are no outliers present in the data. There are more than one peak of the data.

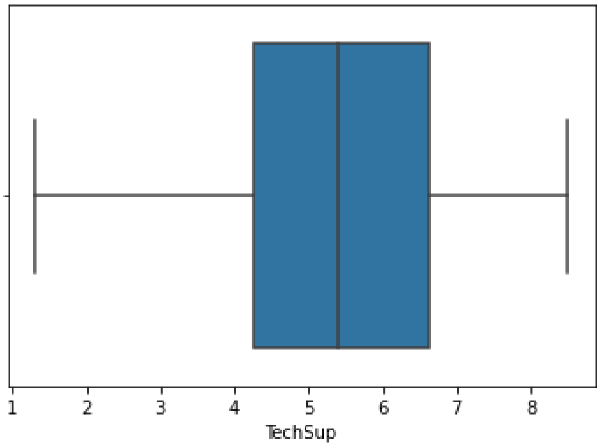
The data from E commerce is normally distributed with the data present between 3.25 & 4.0 there are also few outliers present in the variable.



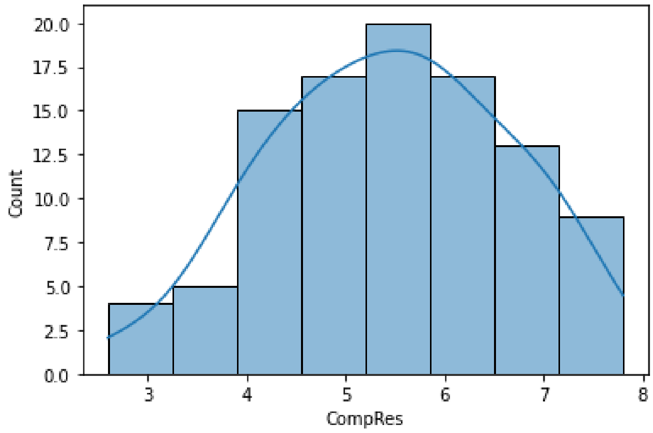


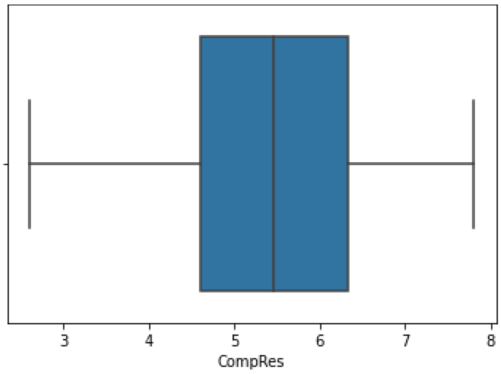
The data for tech support is normally distributed and data ranges between 4 and 7.



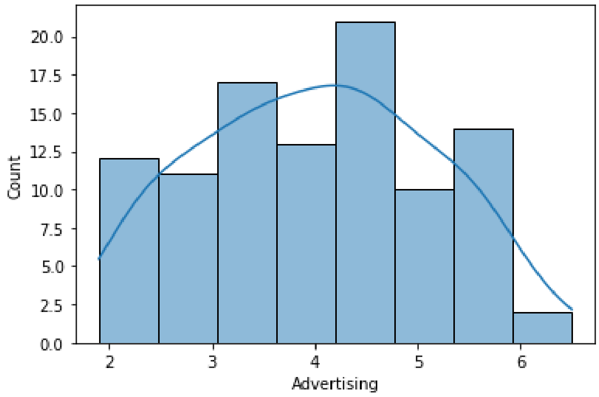


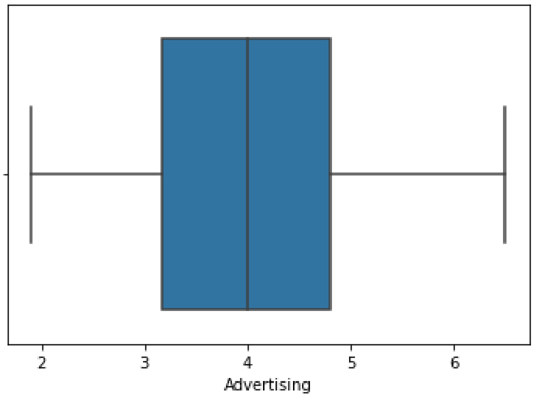
Compliant Resolution data is normally distributed with data ranges between 4.5 to 5



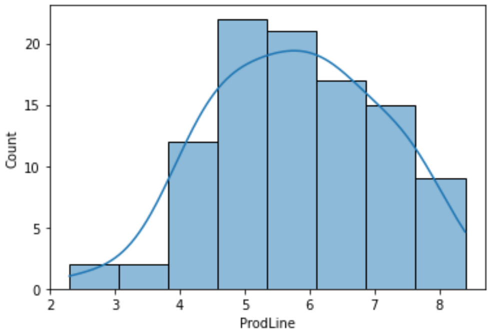


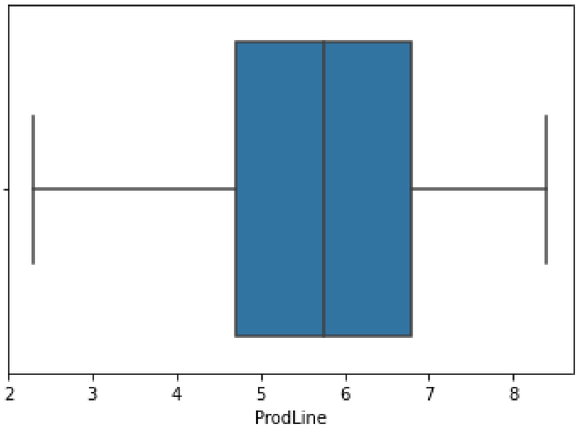
The data for advertising is normally distributed with data ranging between 3 & 5



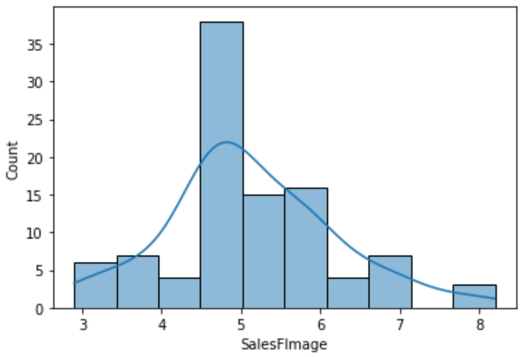


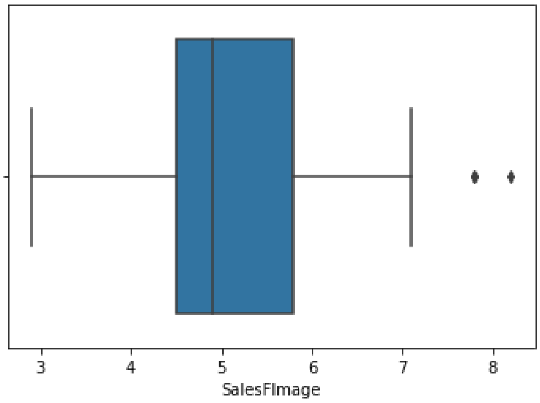
The data for product line is normally distributed and data ranging between 4.54 to 7



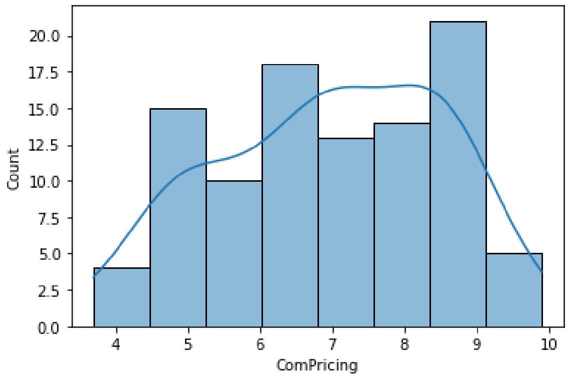


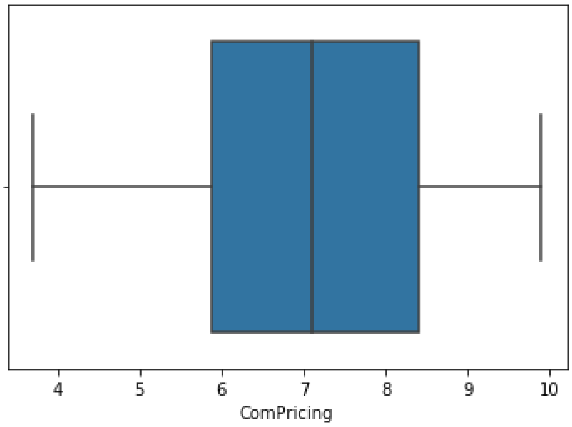
The data for salesforce image is normally distributed with data ranges between 4.5 and 6 and data have median slightly lies toward left.



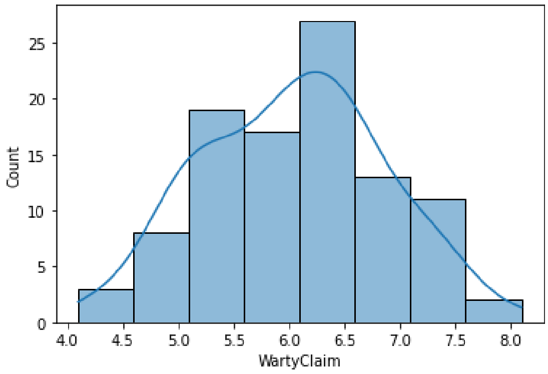


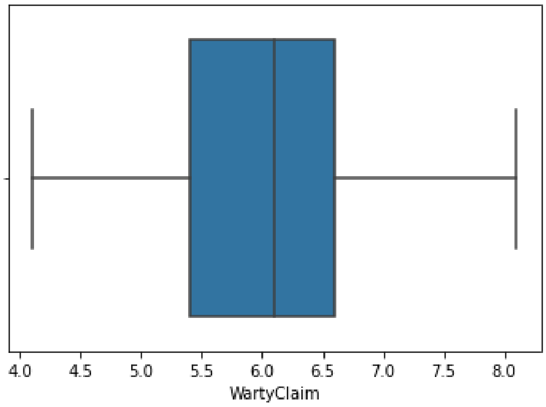
The data for Comp Pricing is distributed with data ranging 5.8 & 8.5



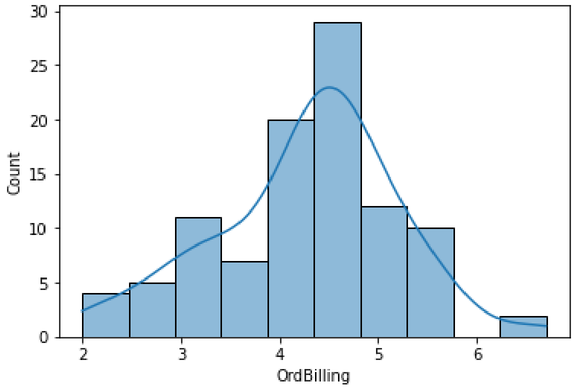


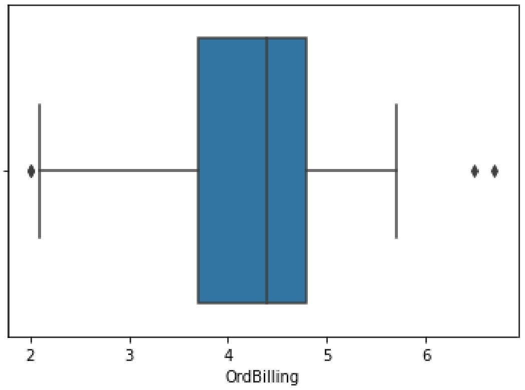
The data of warranty claim is distributed with date ranging from 5.3 & 6.6



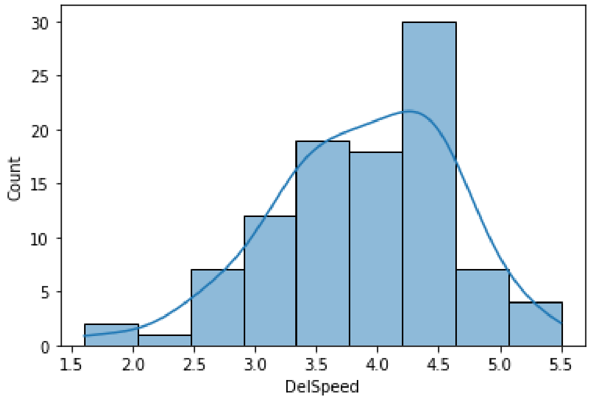


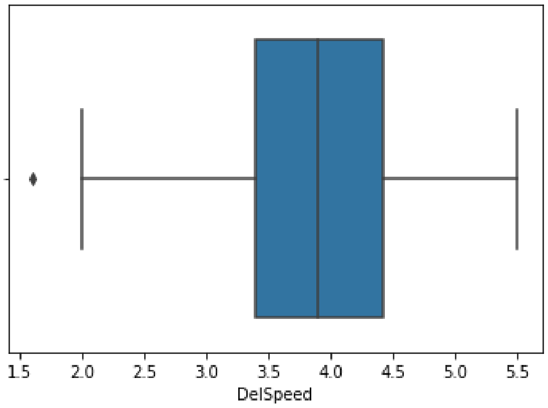
The data is normally distributed with single peak and data ranges between 3.8, 5. There are certain outliers



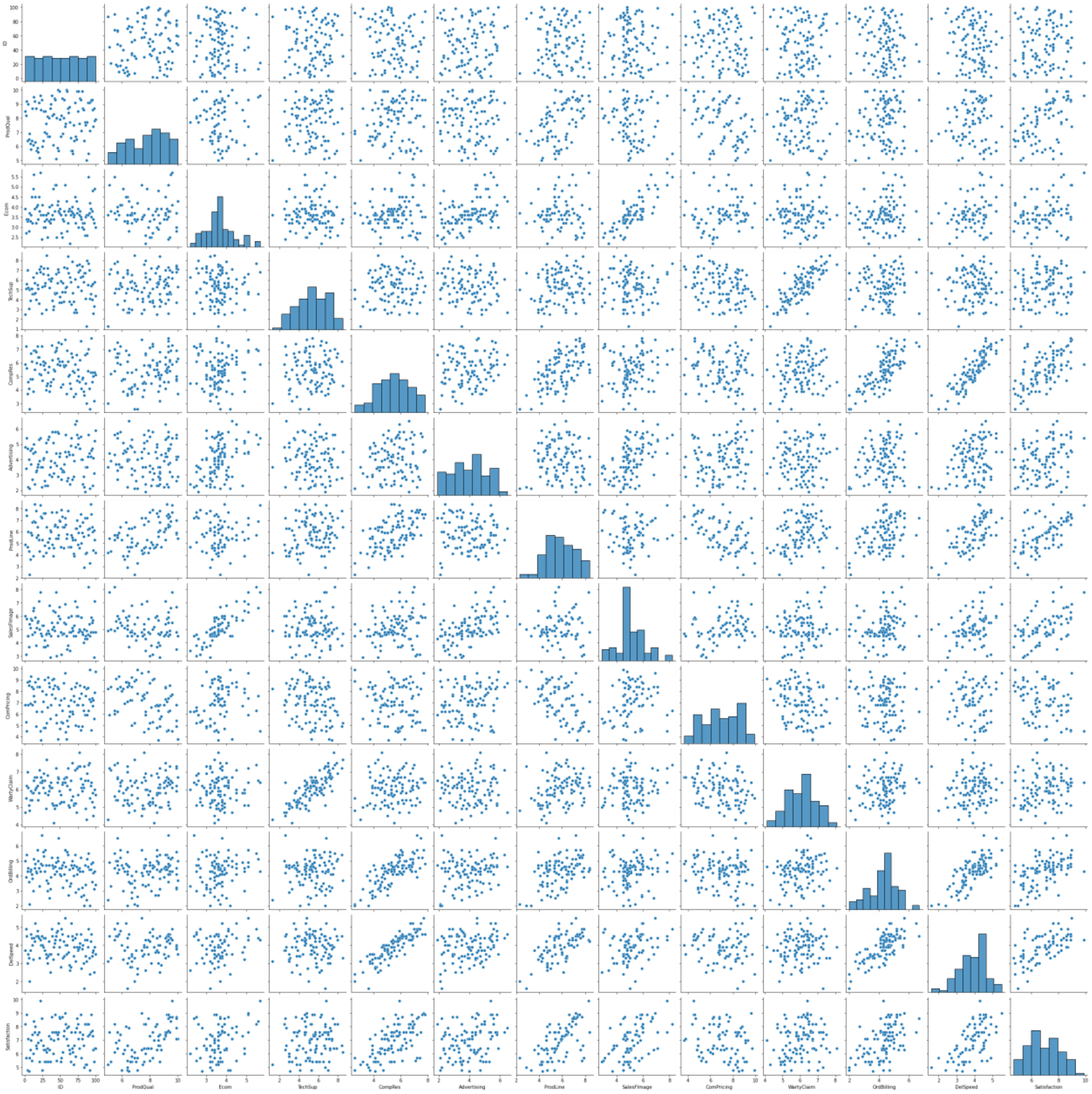


The data distributed between 3.4 and 4.5 with very few outliers





Multi variate Analysis



The Heat map shows the corelation between the data

The following are the variable of the strong corelation

Salesforce Image and ecom is 0.79

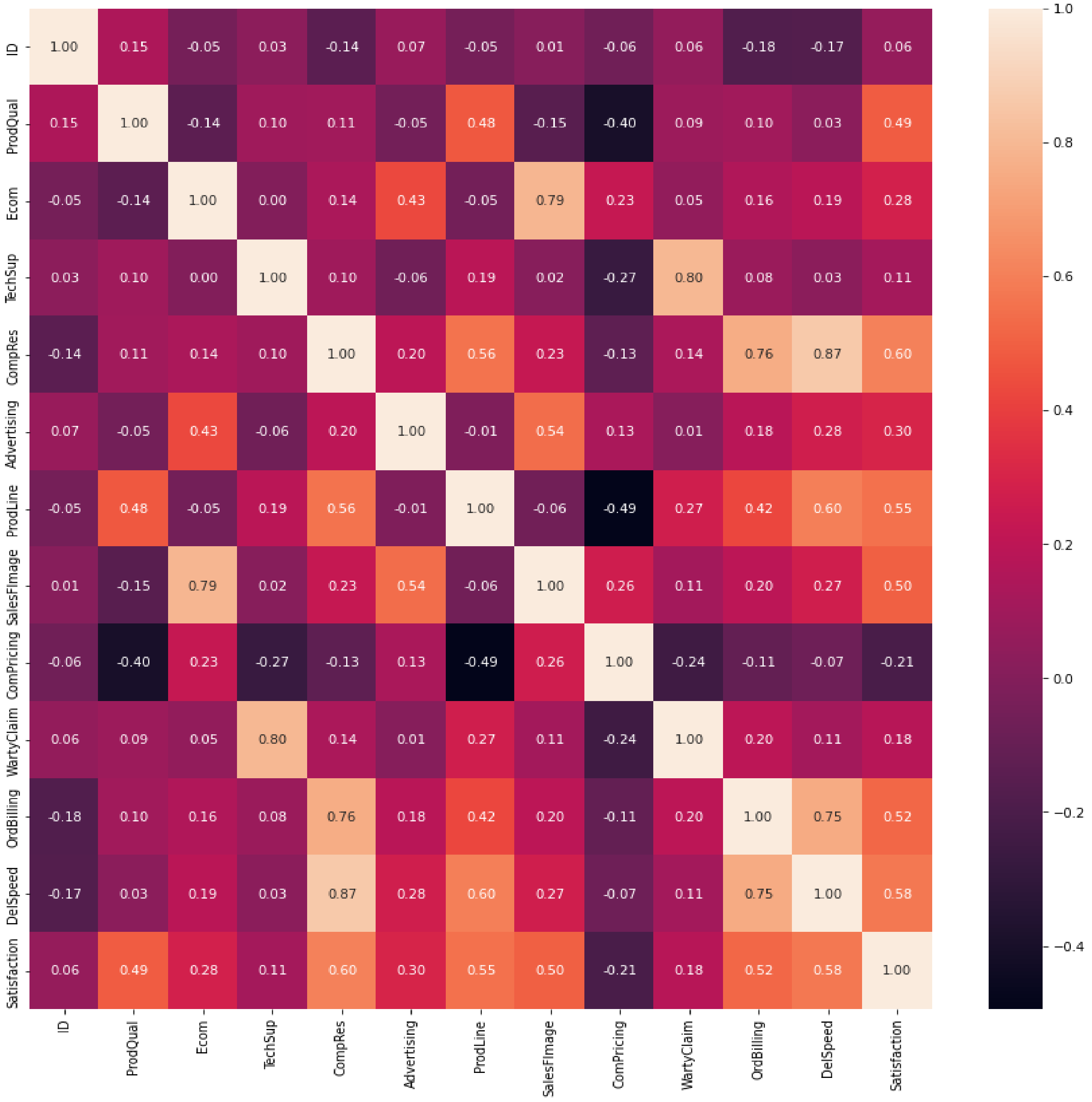
Correlatiab between wartyclaim and techsetup Is 0.80

Compres and DelSpeed is 0.87

DelSpeed and Prodline is 0.602

Ordbling and CompRes is 0.76

OrdBling and DelSpeed is 0.75



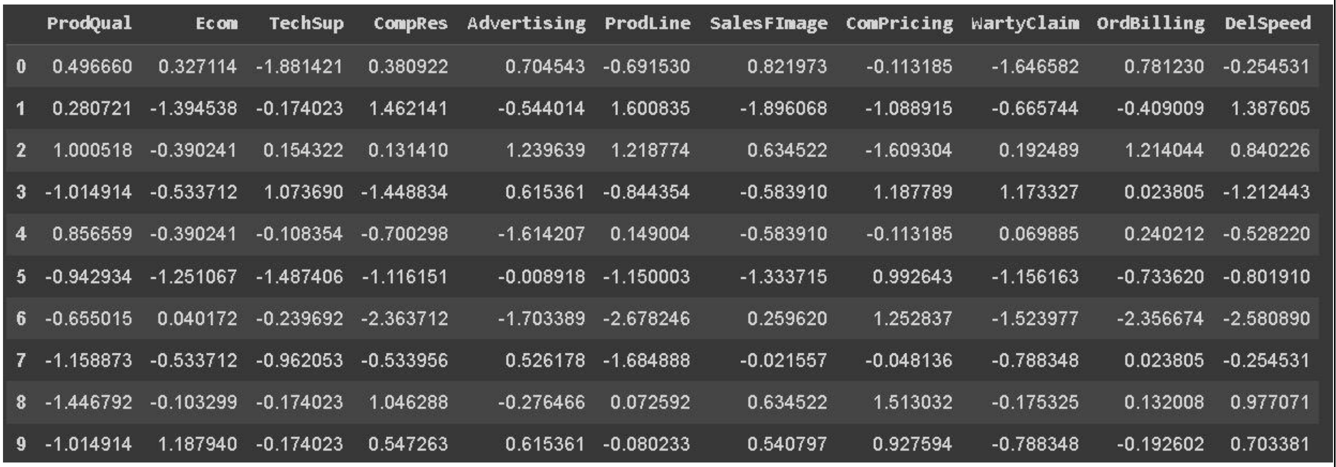
Also, there is a negative correaltion between ProdLine and ComPricing (-0.49), ProdQual and ComPricing (-0.40) Etc

2.3 Scaling of Data

Before scaling of the data, ID and satisfaction columns are dropped as ID is has just unique ID and satisfaction is target variable

The Z-Score scaling is used to scale the data. The scaling is required for the data as each variable has the data skewed between 2-3 data points (for example order billing data lies between 4 & 5whereas product quality lies between 7 and 9)

This will help in standardization of data



	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
0	0.496660	0.327114	-1.881421	0.380922	0.704543	-0.691530	0.821973	-0.113185	-1.646582	0.781230	-0.254531
1	0.280721	-1.394538	-0.174023	1.462141	-0.544014	1.600835	-1.896068	-1.088915	-0.665744	-0.409009	1.387605
2	1.000518	-0.390241	0.154322	0.131410	1.239639	1.218774	0.634522	-1.609304	0.192489	1.214044	0.840226
3	-1.014914	-0.533712	1.073690	-1.448834	0.615361	-0.844354	-0.583910	1.187789	1.173327	0.023805	-1.212443
4	0.856559	-0.390241	-0.108354	-0.700298	-1.614207	0.149004	-0.583910	-0.113185	0.069885	0.240212	-0.528220
5	-0.942934	-1.251067	-1.487406	-1.116151	-0.008918	-1.150003	-1.333715	0.992643	-1.156163	-0.733620	-0.801910
6	-0.655015	0.040172	-0.239692	-2.363712	-1.703389	-2.678246	0.259620	1.252837	-1.523977	-2.356674	-2.580890
7	-1.158873	-0.533712	-0.962053	-0.533956	0.526178	-1.684888	-0.021557	-0.048136	-0.788348	0.023805	-0.254531
8	-1.446792	-0.103299	-0.174023	1.046288	-0.276466	0.072592	0.634522	1.513032	-0.175325	0.132008	0.977071
9	-1.014914	1.187940	-0.174023	0.547263	0.615361	-0.080233	0.540797	0.927594	-0.788348	-0.192602	0.703381

2.4 Covariance & Corelation Matrix

Covariance and correlation helps in measuring the relationship and dependency between two variables in the dataset Scaling will not impact the covariance or correlation as scaling is used only for the standardization of the data. Covariance helps find whether two variables are directly proportional (positive) or inversely proportional (negative). in the data set that is covariance helps to find linear relationship between the variable. Covariance matrix of the data is shown as below, and it clearly shows the positive and negative linear relationship between the data

Correlation shows the how much is both the variables are correlated. The correlation matrix between and after the scaling is same.

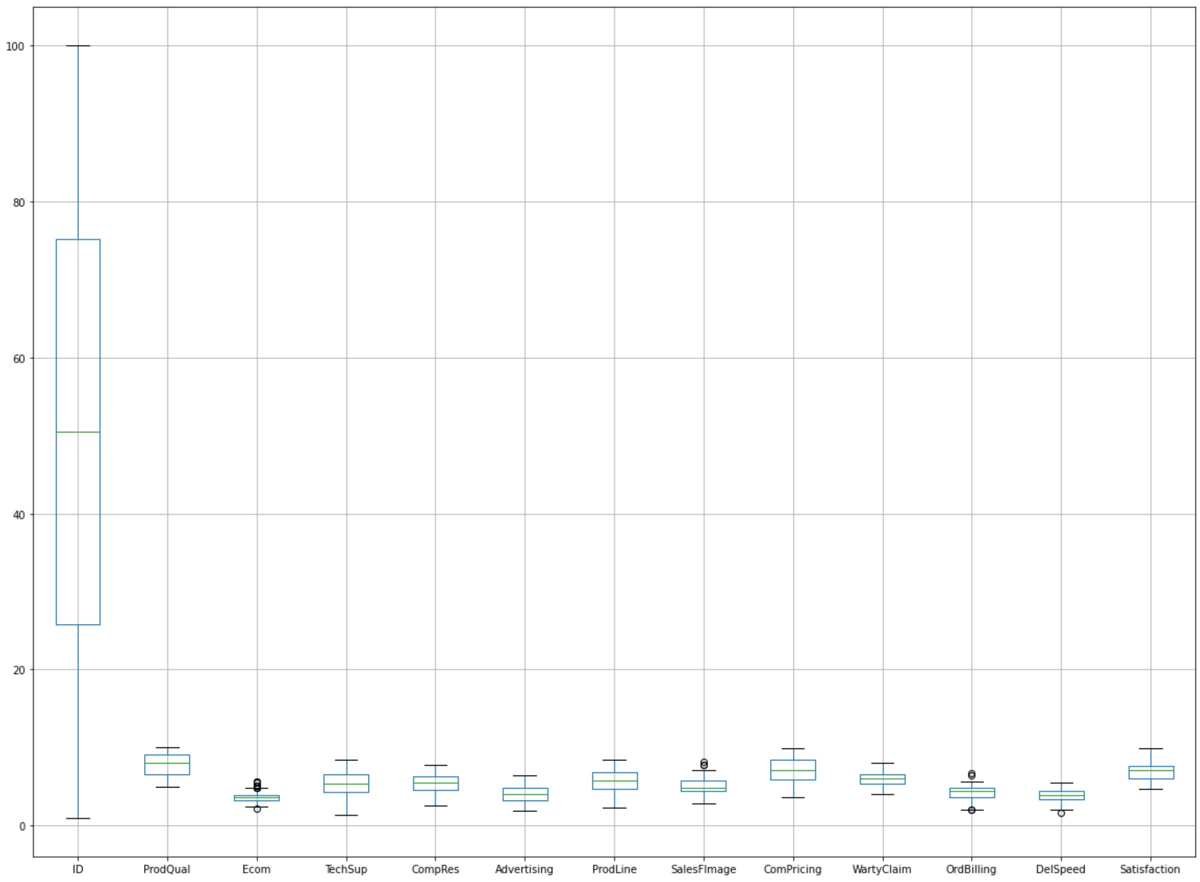
The below matrix shows the correlation between the variables and how strong the relationship between the variable

```
             ProdQual      Ecom   TechSup   CompRes  Advertising  ProdLine  \
ProdQual     1.000000 -0.137163  0.095600  0.106370    -0.053473  0.477493
Ecom        -0.137163  1.000000  0.000867  0.140179     0.429891 -0.052688
TechSup      0.095600  0.000867  1.000000  0.096657    -0.062870  0.192625
CompRes      0.106370  0.140179  0.096657  1.000000     0.196917  0.561417
Advertising -0.053473  0.429891 -0.062870  0.196917     1.000000 -0.011551
```

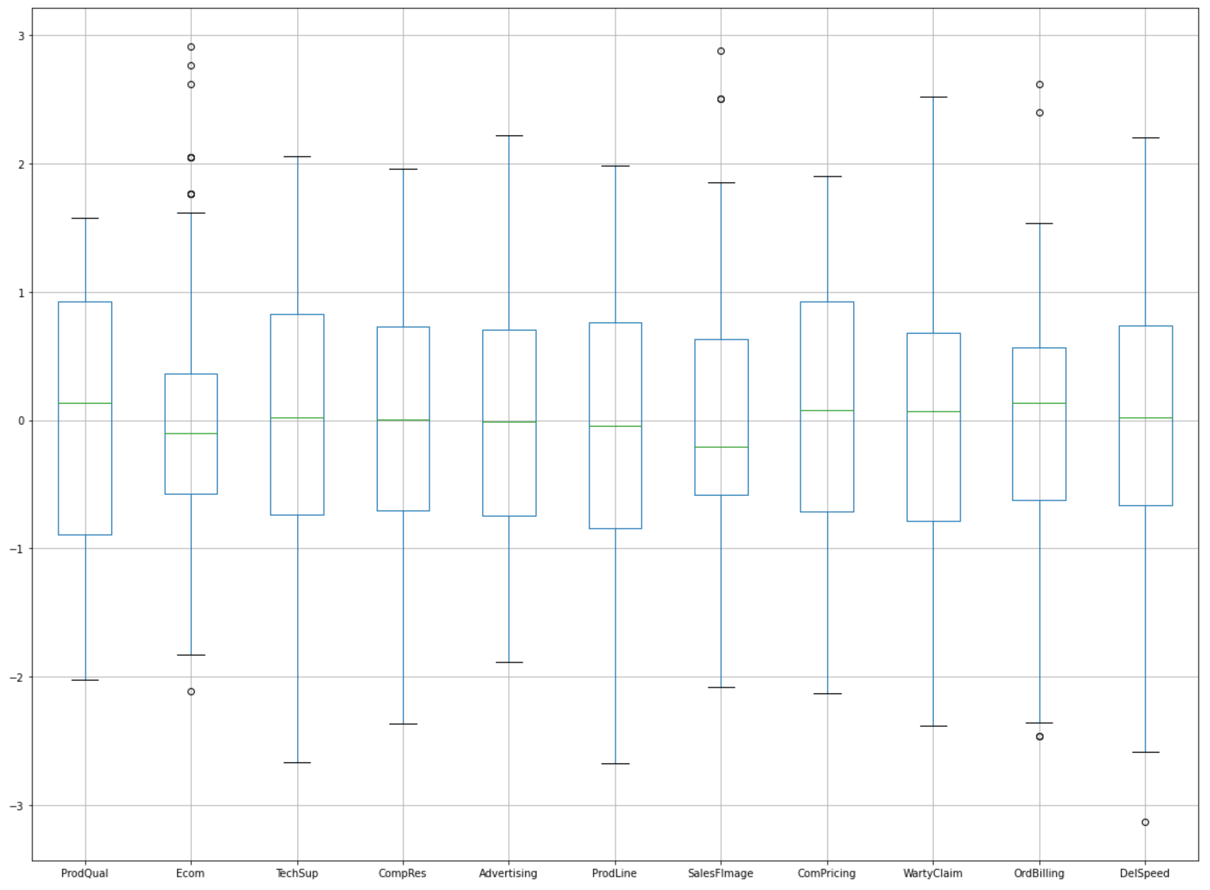
```
ProdLine     0.477493 -0.052688  0.192625  0.561417    -0.011551  1.000000
SalesFImage -0.151813  0.791544  0.016991  0.229752     0.542204 -0.061316
ComPricing  -0.401282  0.229462 -0.270787 -0.127954     0.134217 -0.494948
WartyClaim   0.088312  0.051898  0.797168  0.140408     0.010792  0.273078
OrdBilling   0.104303  0.156147  0.080102  0.756869     0.184236  0.424408
DelSpeed     0.027718  0.191636  0.025441  0.865092     0.275863  0.601850

             SalesFImage  ComPricing  WartyClaim  OrdBilling  DelSpeed
ProdQual       -0.151813   -0.401282    0.088312    0.104303  0.027718
Ecom            0.791544    0.229462    0.051898    0.156147  0.191636
TechSup         0.016991   -0.270787    0.797168    0.080102  0.025441
CompRes         0.229752   -0.127954    0.140408    0.756869  0.865092
Advertising     0.542204    0.134217    0.010792    0.184236  0.275863
ProdLine       -0.061316   -0.494948    0.273078    0.424408  0.601850
SalesFImage     1.000000    0.264597    0.107455    0.195127  0.271551
ComPricing      0.264597    1.000000   -0.244986   -0.114567 -0.072872
WartyClaim      0.107455   -0.244986    1.000000    0.197065  0.109395
OrdBilling      0.195127   -0.114567    0.197065    1.000000  0.751003
DelSpeed        0.271551   -0.072872    0.109395    0.751003  1.000000
```

2.5 Outliers before and after scaling

Outliers before Scaling:



Outliers after Scaling :



There are no changes to outliers before and after scaling, means scaling will not impact outliers.

Ecommerce, salesforce image, order billing and delivery speed have few outliers. Though the number of outliers is less it will be better to remove the outliers in order to have a better models.

2.6 Covariance Matrix, Eigenvalue, Eigenvectors

Covariance Matrix

```
%s [[ 1.01010101e+00 -1.38548704e-01  9.65661154e-02  1.07444445e-01
  -5.40132667e-02  4.82316579e-01 -1.53346338e-01 -4.05335236e-01
   8.92043497e-02  1.05356640e-01  2.79979825e-02]
 [-1.38548704e-01  1.01010101e+00  8.75544162e-04  1.41595213e-01
   4.34233041e-01 -5.32200387e-02  7.99539102e-01  2.31780203e-01
   5.24224157e-02  1.57724577e-01  1.93571786e-01]
 [ 9.65661154e-02  8.75544162e-04  1.01010101e+00  9.76329270e-02
  -6.35051180e-02  1.94571168e-01  1.71621612e-02 -2.73521901e-01
   8.05220127e-01  8.09109340e-02  2.56976702e-02]
 [ 1.07444445e-01  1.41595213e-01  9.76329270e-02  1.01010101e+00
   1.98905906e-01  5.67087831e-01  2.32072486e-01 -1.29246720e-01
   1.41826562e-01  7.64513729e-01  8.73829997e-01]
 [-5.40132667e-02  4.34233041e-01 -6.35051180e-02  1.98905906e-01
   1.01010101e+00 -1.16674936e-02  5.47680463e-01  1.35572620e-01
   1.09010852e-02  1.86096560e-01  2.78649579e-01]
 [ 4.82316579e-01 -5.32200387e-02  1.94571168e-01  5.67087831e-01
  -1.16674936e-02  1.01010101e+00 -6.19348764e-02 -4.99947880e-01
   2.75835887e-01  4.28695202e-01  6.07929503e-01]
 [-1.53346338e-01  7.99539102e-01  1.71621612e-02  2.32072486e-01
   5.47680463e-01 -6.19348764e-02  1.01010101e+00  2.67269246e-01
   1.08540752e-01  1.97098390e-01  2.74294201e-01]
 [-4.05335236e-01  2.31780203e-01 -2.73521901e-01 -1.29246720e-01
   1.35572620e-01 -4.99947880e-01  2.67269246e-01  1.01010101e+00
  -2.47460661e-01 -1.15724268e-01 -7.36078070e-02]
 [ 8.92043497e-02  5.24224157e-02  8.05220127e-01  1.41826562e-01
   1.09010852e-02  2.75835887e-01  1.08540752e-01 -2.47460661e-01
   1.01010101e+00  1.99055678e-01  1.10499598e-01]
 [ 1.05356640e-01  1.57724577e-01  8.09109340e-02  7.64513729e-01
   1.86096560e-01  4.28695202e-01  1.97098390e-01 -1.15724268e-01
   1.99055678e-01  1.01010101e+00  7.58588957e-01]
 [ 2.79979825e-02  1.93571786e-01  2.56976702e-02  8.73829997e-01
   2.78649579e-01  6.07929503e-01  2.74294201e-01 -7.36078070e-02
   1.10499598e-01  7.58588957e-01  1.01010101e+00]]
```

Eigen Values

```
 %s [3.4615872  2.57666335 1.70805705 1.09753137 0.61557989 0.55745836
 0.40557389 0.09942123 0.13418341 0.249446   0.20560936]
```

Eigen Vectors

```
 %s [[ 0.13378962 -0.31349802  0.06227164  0.6431362  -0.2316662  -0.56456996
   0.19164132  0.18279209  0.06659717 -0.13547311  0.0313281 ]
 [ 0.16595278  0.44650918 -0.23524791  0.27238033 -0.42228844  0.26325703
   0.05962621  0.06233863  0.28155772  0.12202642 -0.54251104]
 [ 0.15769263 -0.23096734 -0.61095105 -0.19339314  0.02395667 -0.10876896
  -0.01719992 -0.05192956 -0.3881709  -0.46470964 -0.35929961]
 [ 0.47068359  0.01944394  0.21035078 -0.20632037 -0.02865743 -0.02815231
  -0.0084996  -0.36253352  0.53467243 -0.51339754  0.09324751]
 [ 0.18373495  0.36366471 -0.08809705  0.31789448  0.80387024 -0.20056937
  -0.06306962 -0.08118684  0.03715799  0.05347713 -0.15468169]
 [ 0.38676517 -0.28478056  0.11627864  0.20290226 -0.11667416  0.09819533
  -0.60814755 -0.38507778 -0.23479794  0.3332071  -0.08415534]
 [ 0.2036696   0.47069599 -0.2413421   0.22217722 -0.20437283  0.10497225
   0.00143735 -0.08469869 -0.35341191 -0.16910665  0.64489911]
 [-0.15168864  0.4134565   0.05304529 -0.33354348 -0.24892601 -0.70973595
  -0.30824887 -0.10295751 -0.04518224  0.09883227 -0.09414389]
 [ 0.21293363 -0.19167191 -0.59856398 -0.18530205  0.03292706 -0.13983966
  -0.03064024  0.12893245  0.43534752  0.4435404   0.31756604]
 [ 0.43721774  0.02639905  0.16892981 -0.23685365 -0.02675377 -0.11947974
   0.65931989 -0.19415064 -0.30386545  0.36601754 -0.09907265]
 [ 0.47308914  0.07305172  0.23262477 -0.1973299   0.03543294  0.02979992
  -0.23423927  0.77563222 -0.12010386 -0.06539059 -0.02188514]]
```

2.7 First PC

The Linear eq of 1st component:

-0.134 * ProdQual + -0.166 * Ecom + -0.158 * TechSup + -0.471 * CompRes + -0.184 * Advertising + -0.387 * ProdLine + -0.204 * SalesFImage + 0.152 * ComPricing + -0.213 * WartyClaim + -0.437 *OrdBilling + -0.473 * DelSpeed +

2.8 PCA & PCA Score

Cumulative Variance Explained [31.1542848   54.34425491 69.71676832 79.59455066 85.1347697

90.15189496 93.80205993 96.04707397 97.89755822 99.10520892

100.

From the above array, we can clearly find that total sums to 100.

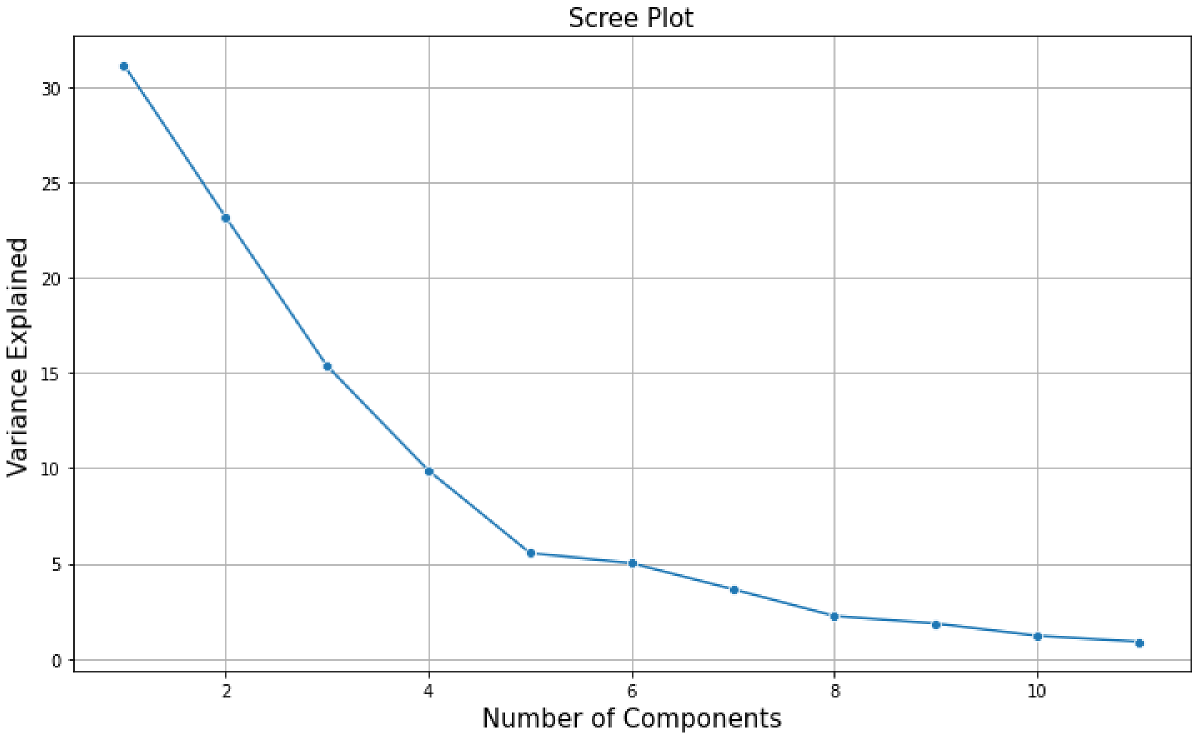
1. Check for cumulative variance upto 90%. Check the corresponding associated

2. The incremental value between the components should not be less than 5%

3. We select 5 as the principal components as after 6 the incremental value between the components is more than 5%

4. So we select 5 principal components for this analysis

The scree plot is shown below as well:



Thus PCA is performed and exported into data frame and multicollinearity is reduced. With this wecan run various models which will help in getting better efficiency scores in the models.

Consider 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?

Cumulative Variance Explained [31.1542848 54.34425491 69.71676832 79.59455066 85.1347697

90.15189496 93.80205993 96.04707397 97.89755822 99.10520892

100.

As a general rule 80-20 is taken, for choosing the number of principal components which are chosen from the cumulative variance explained. Here, we see that 81% is achieved after the 6th Eigen value, hence 6 principal components have been chosen. The Eigenvectors determine the directions of the new attribute space, and the eigenvalues determine their magnitude. As can be seen in the PCA, the components of the eigen vectors determine the PCs.

2.9 Business Implementation

The case study is centred on the numerous items used in hair salons for market segmentation. We were able to understand the relationships between the variables and the distribution of data for each variable from the univariate and multivariate analyses. Instead of analysing all 11 variables in the next step of the study, PCA enables us to eliminate multicollinearity and analyse the 5 factors.

These five elements can be combined to create a common portfolio that can be used to create segmentation strategies.