2021

Advanced Statistics Project

Prajoth

Great learning

Table of Contents

List of Tables:

List of Figures:

Problem 1:Analysis of variance (ANOVA)

The staff of a service center for electrical appliances includes 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

Data Description:

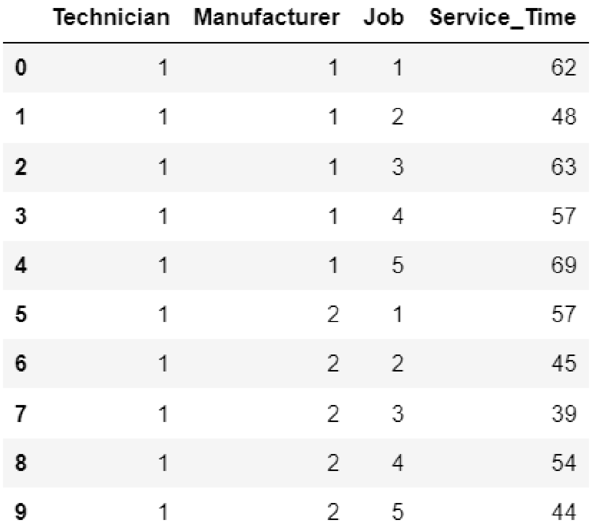
Technician - Three technicians who specialize in repairing.

Manufacturer- Three widely used electrical appliances by three different manufacturers.

Job- Each technician was randomly assigned five repair jobs on each manufacturer's appliance.

Service Time- Time to complete each job (in minutes) was recorded.

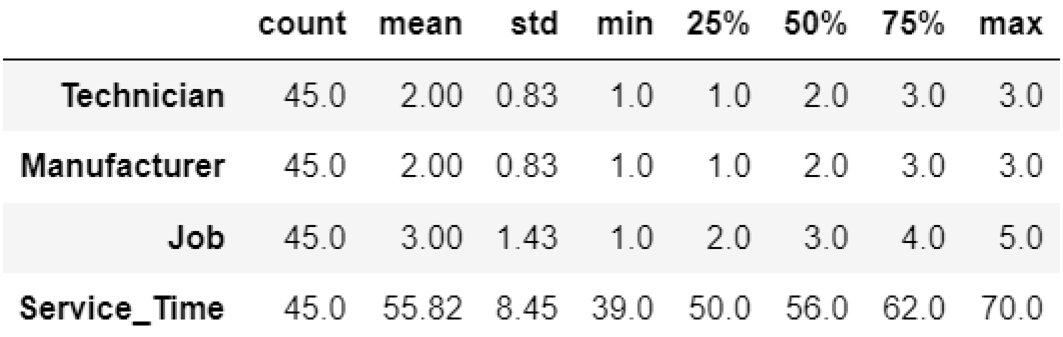
Reading the Dataset:



	Technician	Manufacturer	Job	Service_Time
0	1	1	1	62
1	1	1	2	48
2	1	1	3	63
3	1	1	4	57
4	1	1	5	69
5	1	2	1	57
6	1	2	2	45
7	1	2	3	39
8	1	2	4	54
9	1	2	5	44

Table 1: Reading the data

Summary of Dataset:



	count	mean	std	min	25%	50%	75%	max
Technician	45.0	2.00	0.83	1.0	1.0	2.0	3.0	3.0
Manufacturer	45.0	2.00	0.83	1.0	1.0	2.0	3.0	3.0
Job	45.0	3.00	1.43	1.0	2.0	3.0	4.0	5.0
Service_Time	45.0	55.82	8.45	39.0	50.0	56.0	62.0	70.0

Table 2: Summary of data

Observation:

- Dataset consist 4 variables all are integer.
- Dataset has 45 rows and 4 columns.
- There is no null values and duplicates in dataset.

1.1 State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables ‘Manufacturer’ and ‘Technician individually

Null and the Alternate hypothesis for Technician:

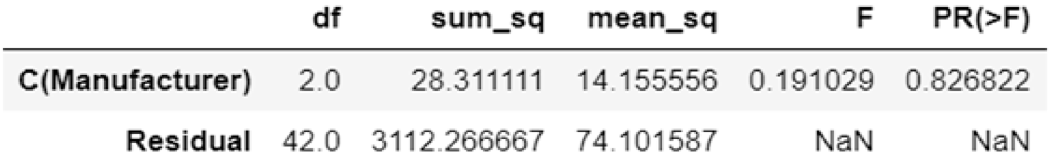
- H0 : Mean of Service Time for all Technician is same
- Ha: For at least one level of Technician mean is different

Null and the Alternate hypothesis for Manufacturer:

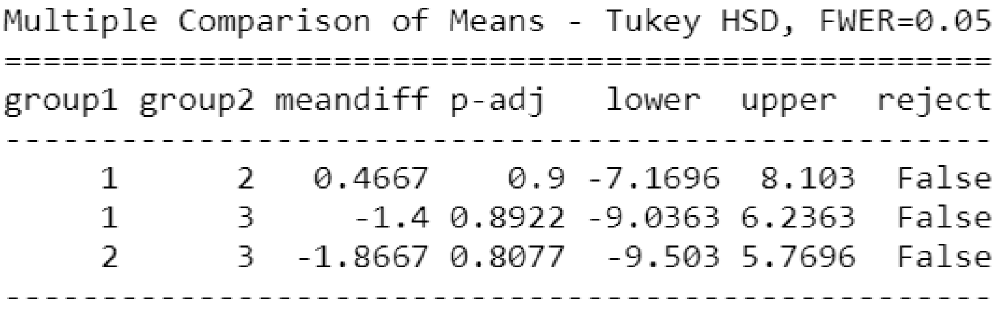
- H0 : Mean of Service Time for all Manufacturer is same
- Ha: For at least one level Manufacturer of mean is different

Let’s assume significance level=0.05

1.2 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.



	df	sum_sq	mean_sq	F	PR(>F)
C(Manufacturer)	2.0	28.311111	14.155556	0.191029	0.826822
Residual	42.0	3112.266667	74.101587	NaN	NaN



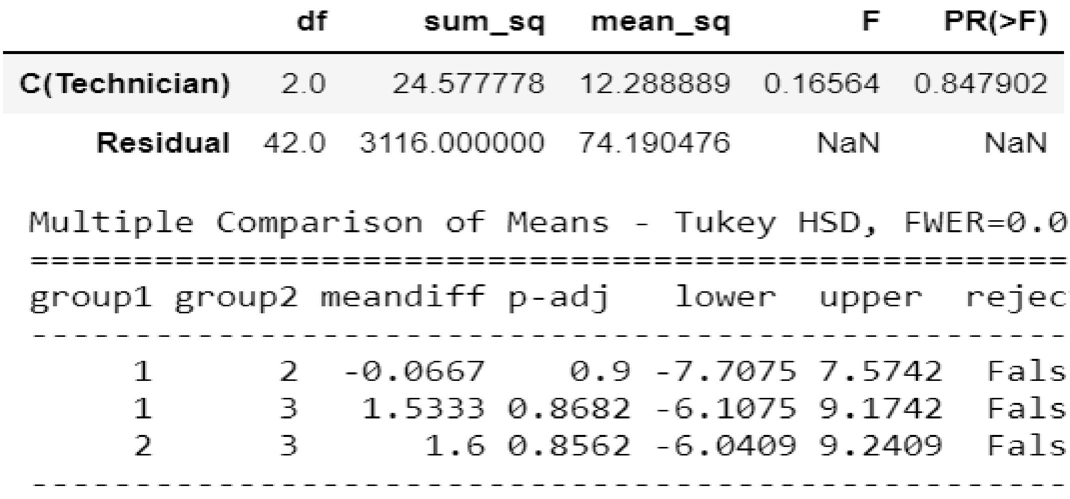
```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj   lower  upper  reject
----------------------------------------------------
     1      2   0.4667    0.9 -7.1696  8.103  False
     1      3     -1.4 0.8922 -9.0363 6.2363  False
     2      3  -1.8667 0.8077  -9.503 5.7696  False
----------------------------------------------------
```

Table 3: 1-Way Anova with Manufacturer and Service Time

Observation:

- Since the p value is greater than the significance level(0.05), we failed to reject the null hypothesis and states that there is a no difference in the mean of Service Time based on Technician
- All the combinations mean difference in Manufacturer is greater than 0.05 ie significance level hence we failed to reject the Null hypothesis.

1.3 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.



	df	sum_sq	mean_sq	F	PR(>F)
C(Technician)	2.0	24.577778	12.288889	0.16564	0.847902
Residual	42.0	3116.000000	74.190476	NaN	NaN

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj   lower  upper  reject
-----------------------------------------------------
     1      2  -0.0667    0.9 -7.7075 7.5742  False
     1      3   1.5333 0.8682 -6.1075 9.1742  False
     2      3      1.6 0.8562 -6.0409 9.2409  False
-----------------------------------------------------
```

Table 4:1-Way Anova with Technician and Service Time

Observation:

- Since the p value is greater than the significance level(0.05), we failed to reject the null hypothesis and states that there is a no difference in the mean of Service Time based on Technician.
- All the combinations mean difference in Technician is greater than 0.05 i.e. significance level hence we failed to reject the Null hypothesis

1.4) Analyse the effects of one variable on another with the help of an interaction plot. What is an interaction between two treatments?

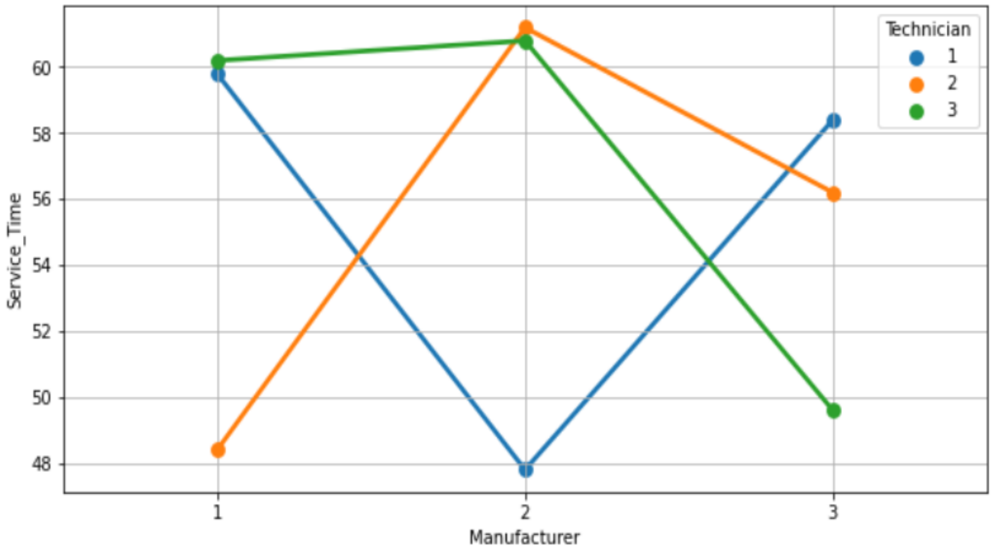


Figure 1: Interaction Plot

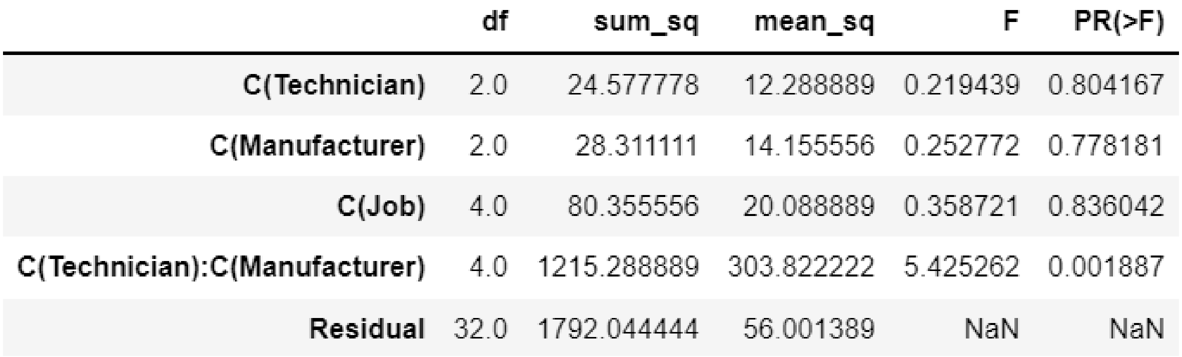
Observation:

- There is good interaction between manufacturer and technician based on service time.
- Technician 1 has low service time with manufacturer 2 when compare to all technicians and has high service time with manufacturer 3 when compared to all technicians.
- Technician 2 with manufacturer 2 has high service time than all technicians and low service time with manufacture 1.
- Technician 3 as high service time with manufacturer 1 & 2 and low with manufacturer 3.

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

- There is good interaction between manufaturer and technician based on service time so we need to perform two-way ANOVA for better prediction.

C(Technician):C(Manufacturer)-Interaction Effect term



	df	sum_sq	mean_sq	F	PR(>F)
C(Technician)	2.0	24.577778	12.288889	0.219439	0.804167
C(Manufacturer)	2.0	28.311111	14.155556	0.252772	0.778181
C(Job)	4.0	80.355556	20.088889	0.358721	0.836042
C(Technician):C(Manufacturer)	4.0	1215.288889	303.822222	5.425262	0.001887
Residual	32.0	1792.044444	56.001389	NaN	NaN

Table 5: 2-way Anova with all Variables

Observation:

- Since the p value is greater than the significance level(0.05), we failed to reject the null hypothesis and states that there is a no difference in the mean of Service Time based on Technician, Manufacturer and jobs.

we see that the p-value of the interaction effect term of 'Technician' and 'Manufacturer' suggests that the Null Hypothesis is rejected in this case. So there is a difference in mean while interaction of both Technician and Manufacturer. Interaction term(C(Technician):C(Manufacturer)) has highest F-Value when compared to Technician and Manufacturer.

1.6) Mention the business implications of performing ANOVA for this particular case study.

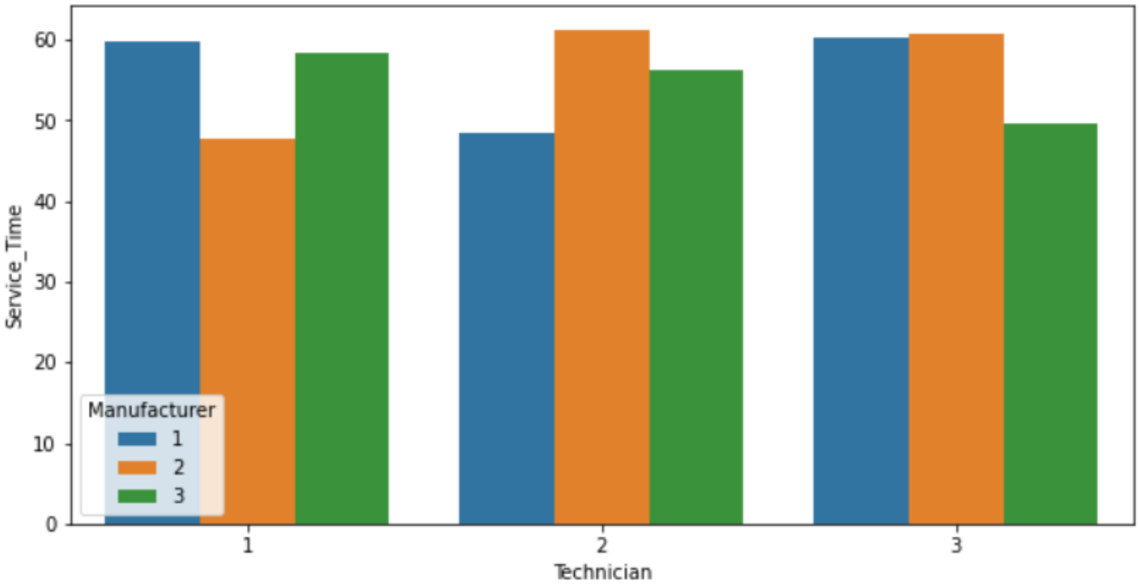


Figure 2: Bar Plot with all variables

Observation:

- Technician 1 has low service time with manufacturer 2 when compare to all technicians and has high service time with manufacturer 3 when compared to all technicians.
- Technician 2 with manufacturer 2 has high service time than all technicians and low service time with manufacture 1.
- Technician 3 as high service time with manufacturer 1 & 2 and low with manufacturer 3.

Problem 2:Principal Component Analysis (PCA)

Problem Statement:

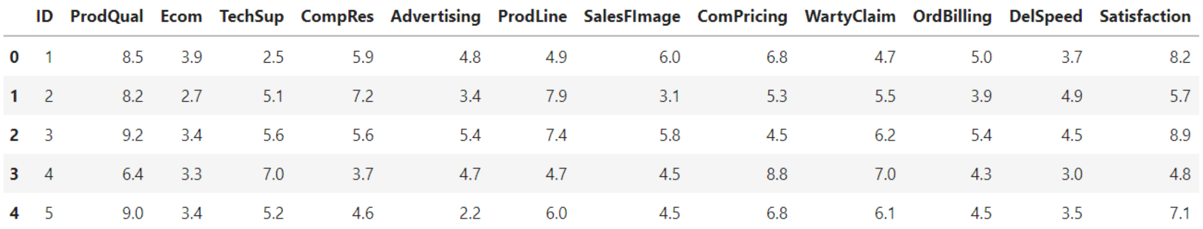
The 'Hair Salon.csv' dataset contains various variables used for the context of Market Segmentation. This particular case study is based on various parameters of a salon chain of hair products. You are expected to do Principal Component Analysis for this case study according to the instructions given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

Data Description:

Variable	Expansion
Id	Customer Id
ProdQual	Product Quality
E-Com	E-Commerce
TechSup	Tech Support
ComRes	Complaint Resolution

Advertising	Advertising
ProdLine	Product Line
SalesFImage	Sales Face Image
ComPricing	Competitive Pricing
WartyClaim	Warrenty & Claims
OrdBilling	Order & Billing
DelSpeed	Delivery Speed
Satisfaction	Satisfaction

Reading the Dataset:



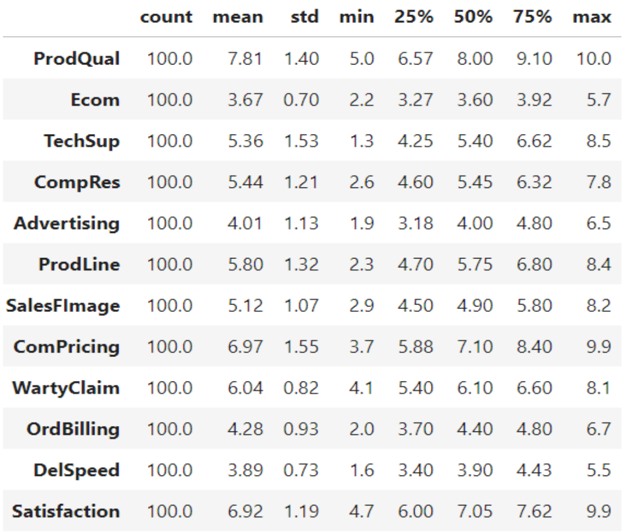
	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
0	1	8.5	3.9	2.5	5.9	4.8	4.9	6.0	6.8	4.7	5.0	3.7	8.2
1	2	8.2	2.7	5.1	7.2	3.4	7.9	3.1	5.3	5.5	3.9	4.9	5.7
2	3	9.2	3.4	5.6	5.6	5.4	7.4	5.8	4.5	6.2	5.4	4.5	8.9
3	4	6.4	3.3	7.0	3.7	4.7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

Table 6: Reading the data

Observation:

- ➢ Dataset has 100 rows and 13 columns.
- ➢ Dataset consists of 12 float data type and one integer type.
- ➢ Satisfaction is target or dependent variable in dataset.

Summary of Dataset:



	count	mean	std	min	25%	50%	75%	max
ProdQual	100.0	7.81	1.40	5.0	6.57	8.00	9.10	10.0
Ecom	100.0	3.67	0.70	2.2	3.27	3.60	3.92	5.7
TechSup	100.0	5.36	1.53	1.3	4.25	5.40	6.62	8.5
CompRes	100.0	5.44	1.21	2.6	4.60	5.45	6.32	7.8
Advertising	100.0	4.01	1.13	1.9	3.18	4.00	4.80	6.5
ProdLine	100.0	5.80	1.32	2.3	4.70	5.75	6.80	8.4
SalesFImage	100.0	5.12	1.07	2.9	4.50	4.90	5.80	8.2
ComPricing	100.0	6.97	1.55	3.7	5.88	7.10	8.40	9.9
WartyClaim	100.0	6.04	0.82	4.1	5.40	6.10	6.60	8.1
OrdBilling	100.0	4.28	0.93	2.0	3.70	4.40	4.80	6.7
DelSpeed	100.0	3.89	0.73	1.6	3.40	3.90	4.43	5.5
Satisfaction	100.0	6.92	1.19	4.7	6.00	7.05	7.62	9.9

Table 7: Summary of Data

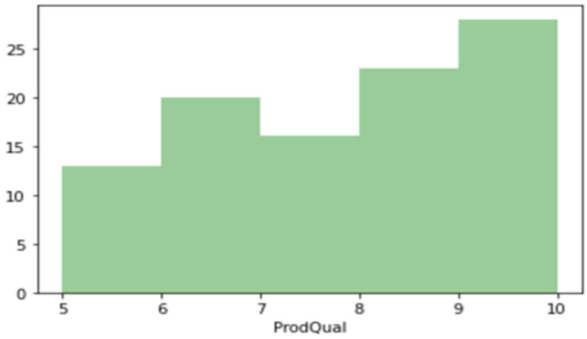
Observation:

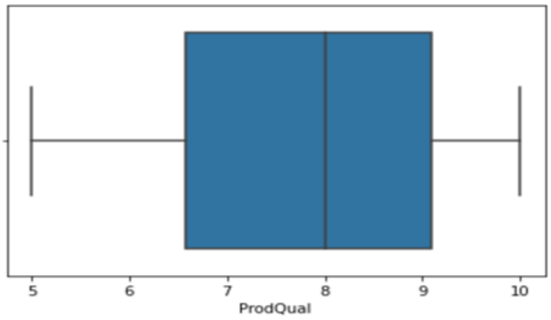
- ➢ There is no null values and duplicates in dataset.

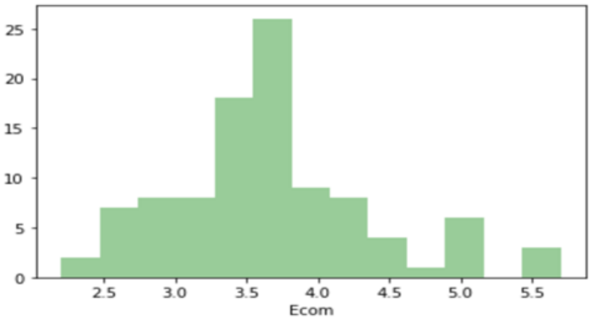
- Means and Median of all features is almost same.

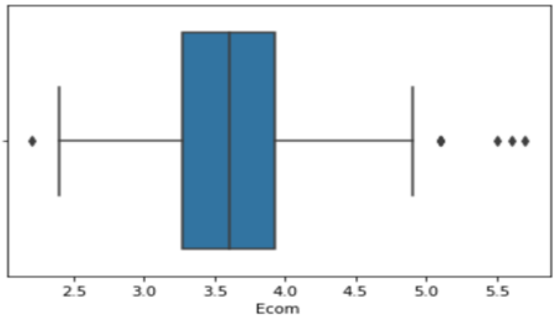
2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented

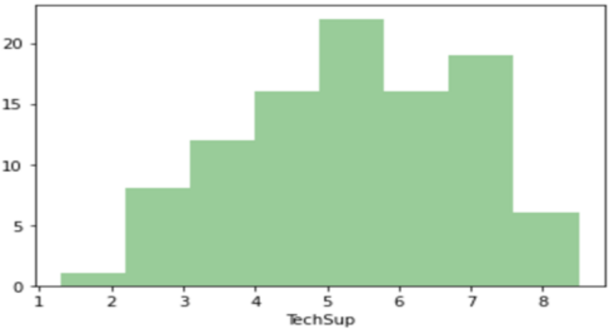
Univariate Analysis:

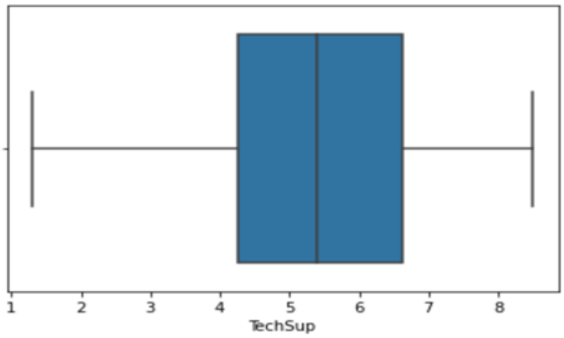


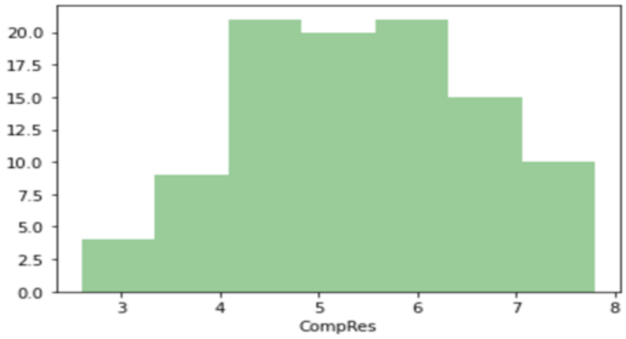


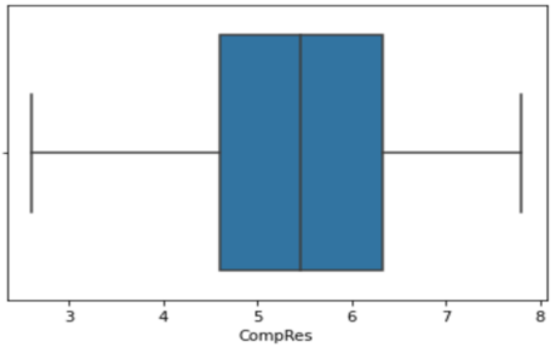


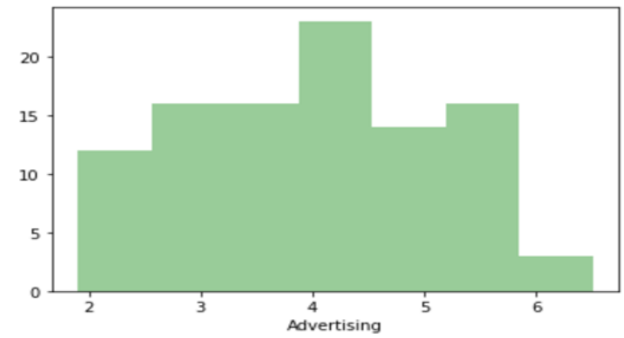


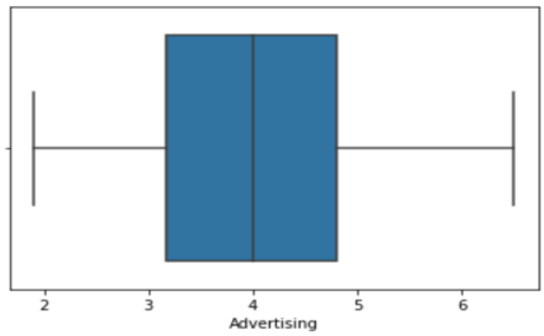


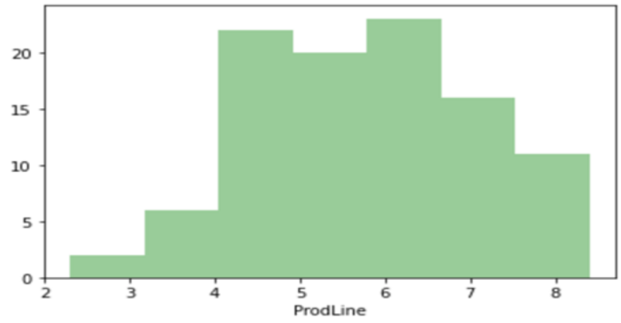


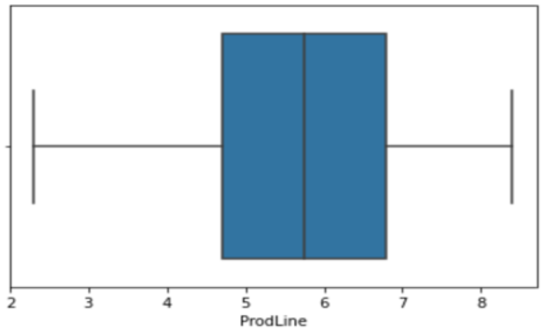


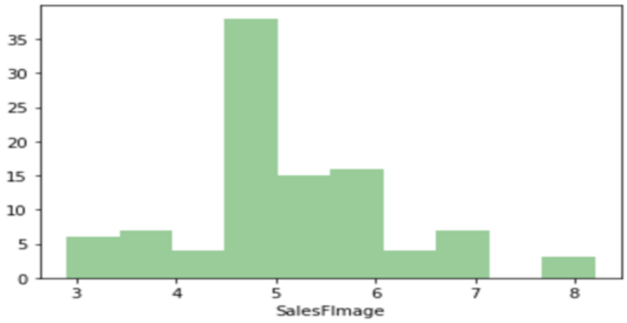


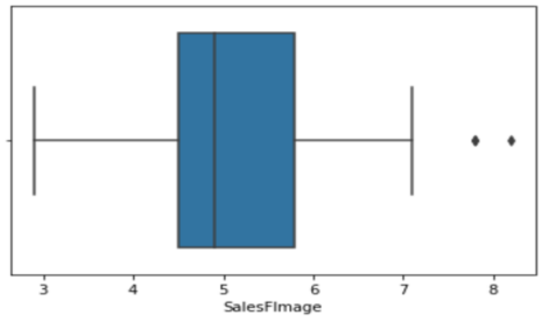

0
5
10
15
20
2
3
4
5
6
Advertising

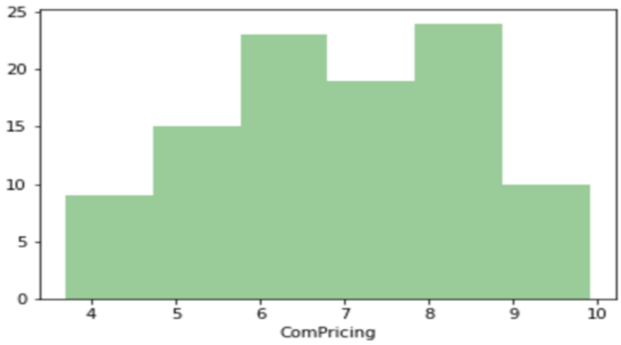

2
3
4
5
6
Advertising

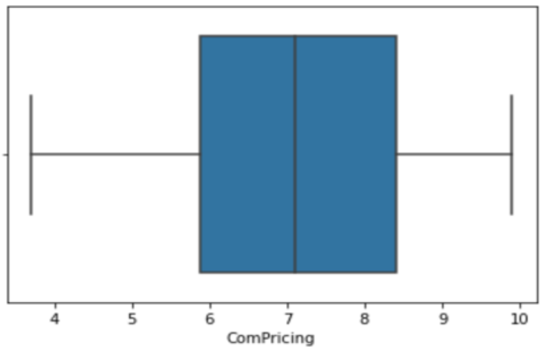

0
5
10
15
20
2
3
4
5
6
7
8
ProdLine


2
3
4
5
6
7
8
ProdLine


0
5
10
15
20
25
30
35
3
4
5
6
7
8
SalesFImage


3
4
5
6
7
8
SalesFImage


0
5
10
15
20
25
4
5
6
7
8
9
10
ComPricing


4
5
6
7
8
9
10
ComPricing

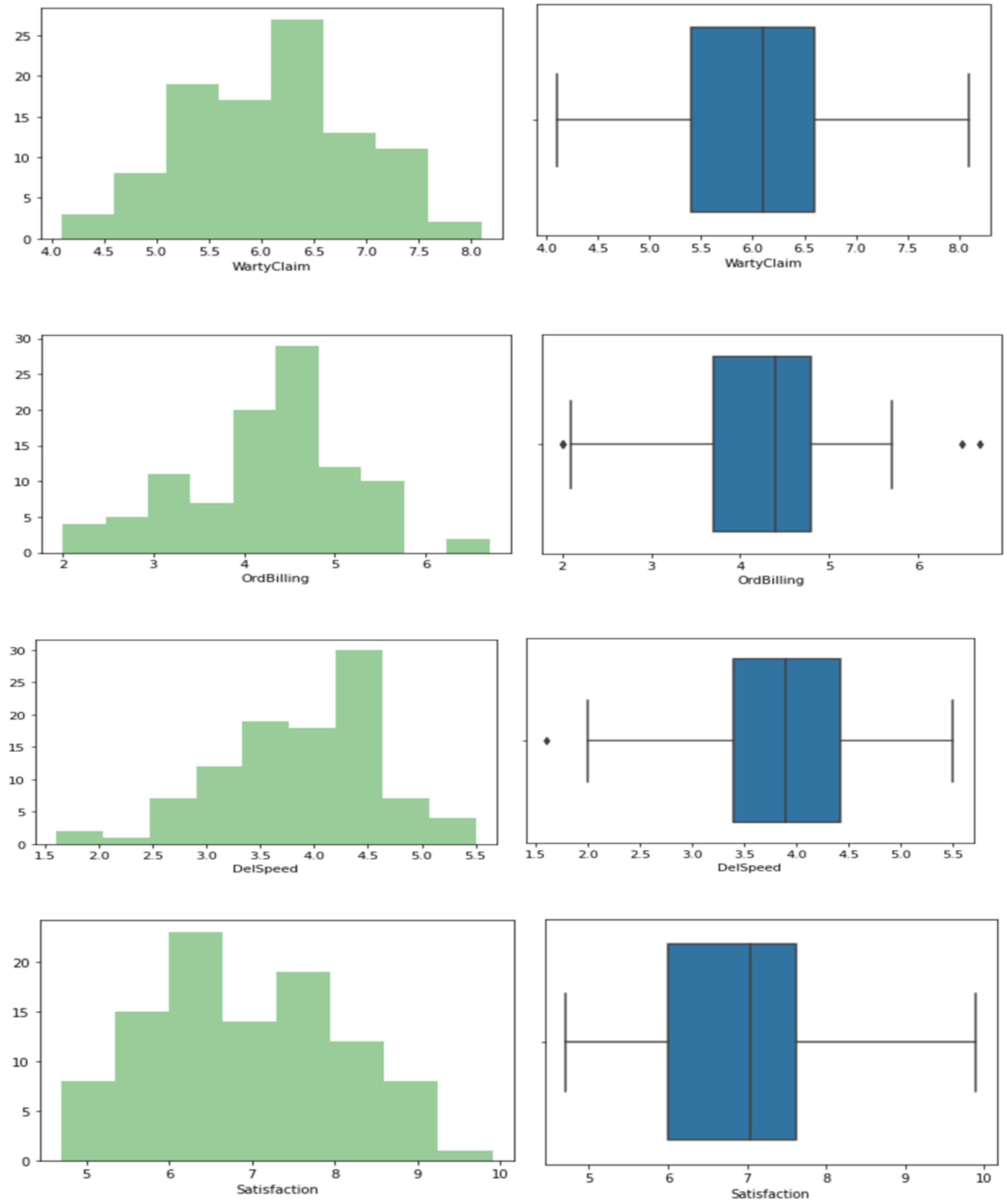


Figure 3: Dist and Box Plot for all Variables

Observation:

- There are 12 features in dataset all are float data type.
- Satisfaction is target or dependent variable in dataset.
- Means and medians of all features is almost same so they are normally distributed.
- E-com, Sales Image , Order & Billing, Delivery Speed has outliers in data.
- Satisfaction from customer’s min is 4.7 and max is 9.9.

Multivariate Analysis:

Pair Plot with Contineous Variable:

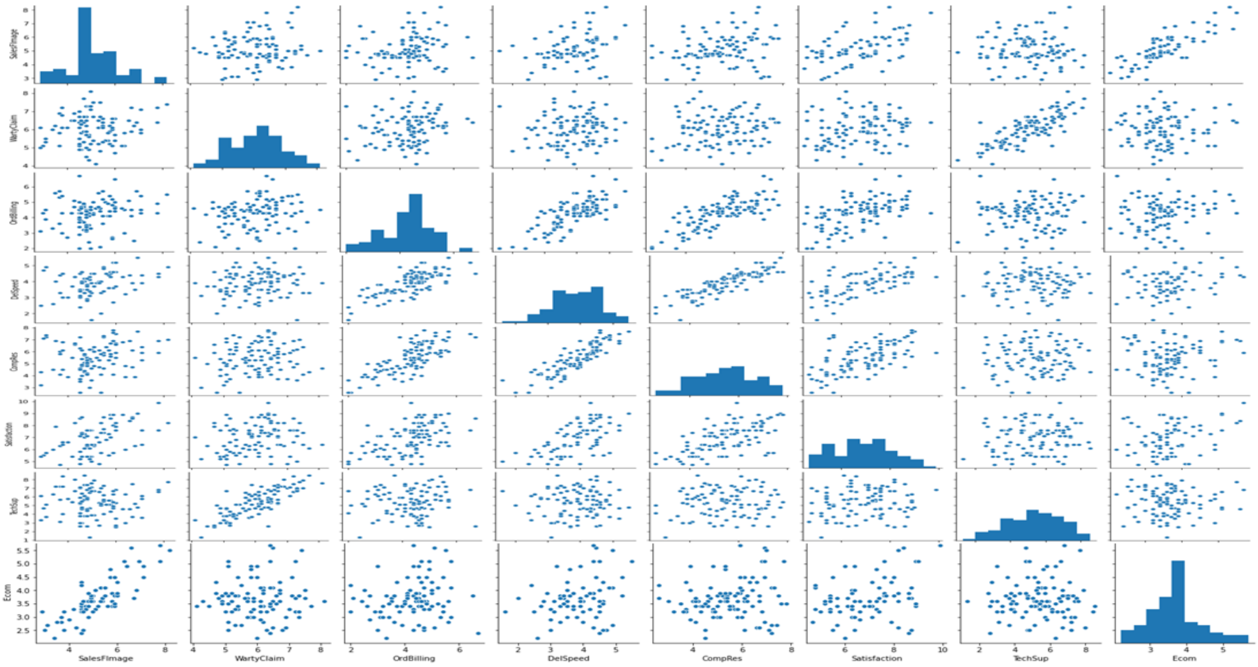


Figure 4: Pair Plot with Continuous Variable

Heatmap:

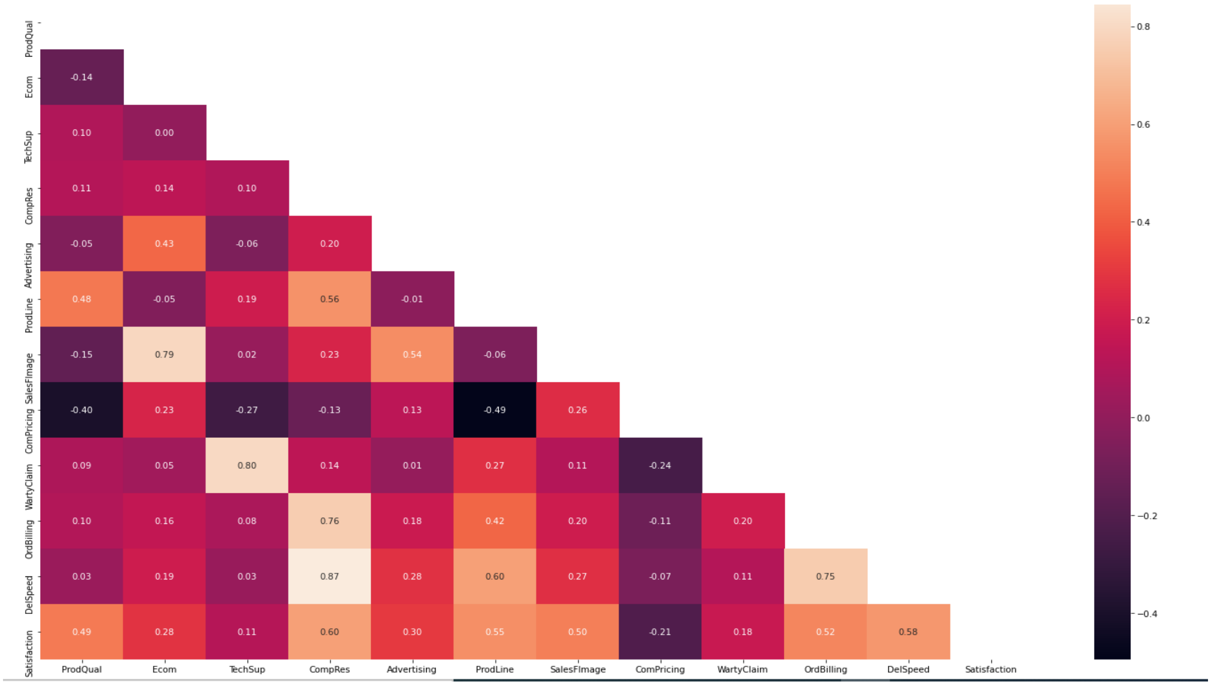


Figure 5:Heat Map

Observation:

- Complaint resolution and delivery speed **(0.87),** Warranty claim and Tech Support **(0.80)** , E-com and Sales Image **(0.79)** which has **positive correlation** in descending order.
- Competitive pricing and Tech support **(-0.27),**Competitive pricing and Product Quality **(-0.40)**, Competitive pricing and product line **(-0.49)** which has top 3 **negative correlation**.

Outlier Treatment:

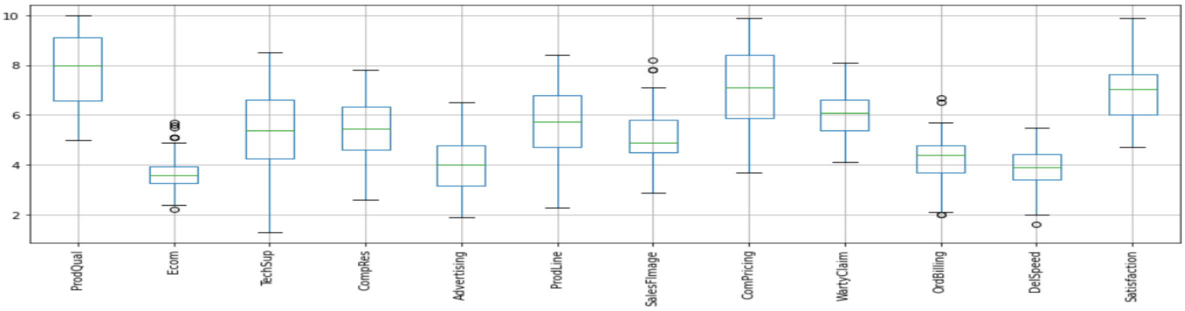


Figure 6: Before treating Outliers

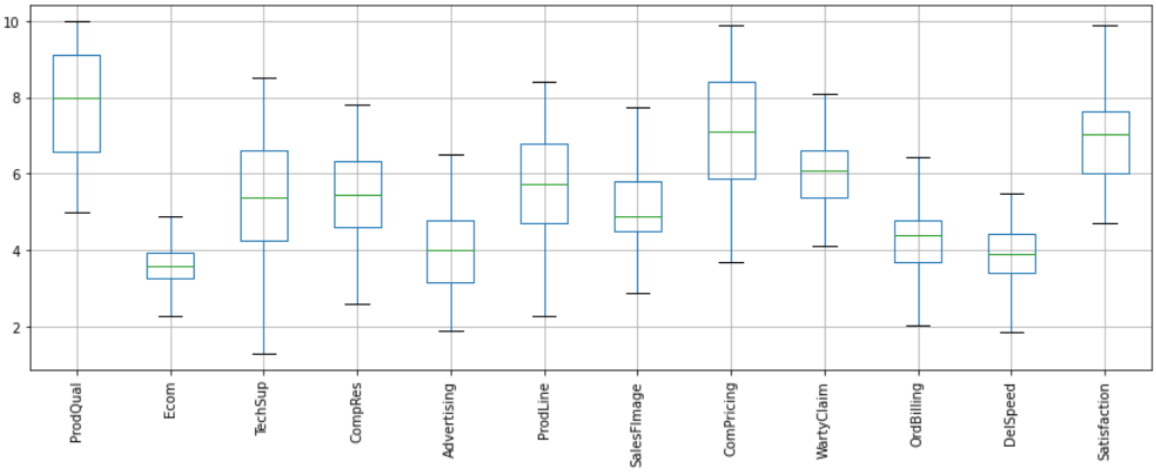
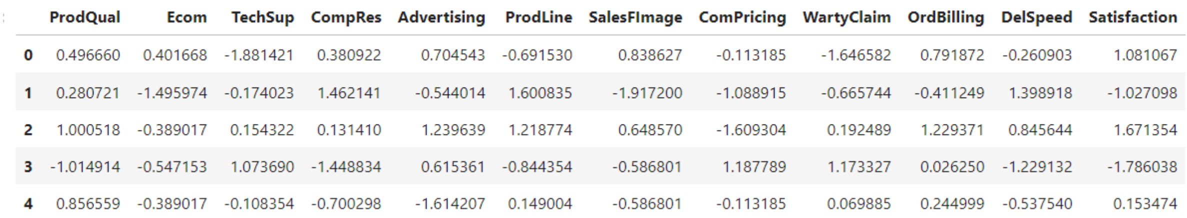


Figure 7: After treating Outliers

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



	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
0	0.496660	0.401668	-1.881421	0.380922	0.704543	-0.691530	0.838627	-0.113185	-1.646582	0.791872	-0.260903	1.081067
1	0.280721	-1.495974	-0.174023	1.462141	-0.544014	1.600835	-1.917200	-1.088915	-0.665744	-0.411249	1.398918	-1.027098
2	1.000518	-0.389017	0.154322	0.131410	1.239639	1.218774	0.648570	-1.609304	0.192489	1.229371	0.845644	1.671354
3	-1.014914	-0.547153	1.073690	-1.448834	0.615361	-0.844354	-0.586801	1.187789	1.173327	0.026250	-1.229132	-1.786038
4	0.856559	-0.389017	-0.108354	-0.700298	-1.614207	0.149004	-0.586801	-0.113185	0.069885	0.244999	-0.537540	0.153474

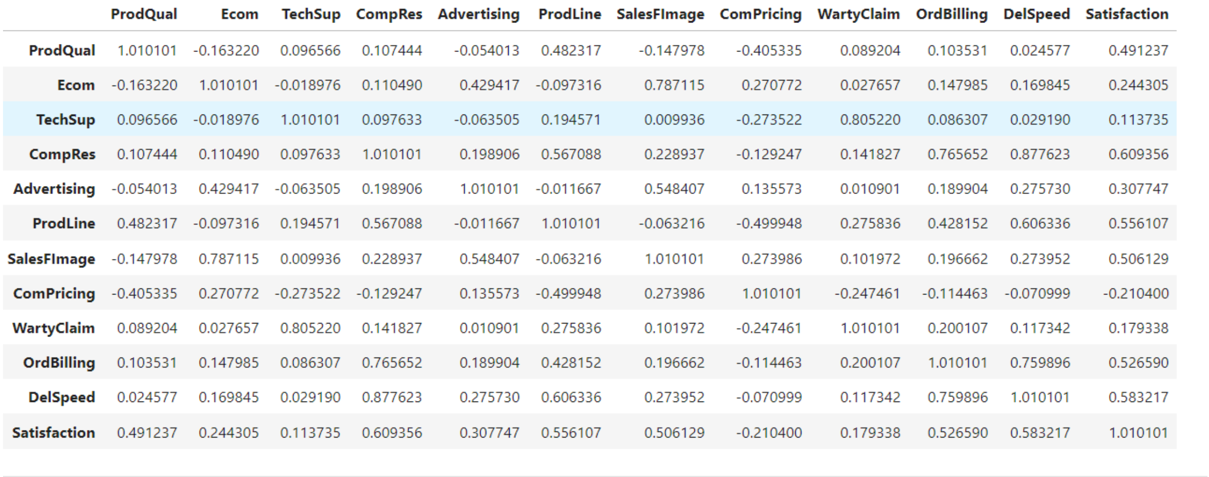
Table 8: Scaling the data

Observation:

- Normalisation is used to transform all variables in the data to a same range. It doesn't solve the problem caused by outliers.

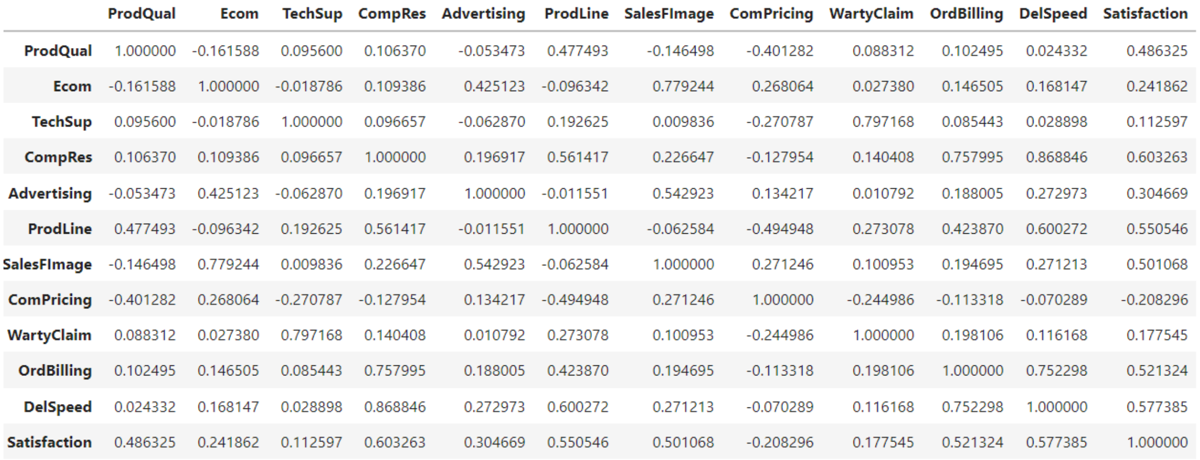
- ➢ So, as you can see the after normalisation also, the outliers remains outliers. Only the range is changed.
- ➢ Sacling should be done to center data before performing PCA since the transformation relies on the data being around the origin.
- ➢ Some data might already follow a standard normal distribution with mean zero and standard deviation of one and so would not have to be scaled before PCA.So PCA will not give accurate results.

2.3 Comment on the comparison between covariance and the correlation matrix after scaling



	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.010101	-0.163220	0.096566	0.107444	-0.054013	0.482317	-0.147978	-0.405335	0.089204	0.103531	0.024577	0.491237
Ecom	-0.163220	1.010101	-0.018976	0.110490	0.429417	-0.097316	0.787115	0.270772	0.027657	0.147985	0.169845	0.244305
TechSup	0.096566	-0.018976	1.010101	0.097633	-0.063505	0.194571	0.009936	-0.273522	0.805220	0.086307	0.029190	0.113735
CompRes	0.107444	0.110490	0.097633	1.010101	0.198906	0.567088	0.228937	-0.129247	0.141827	0.765652	0.877623	0.609356
Advertising	-0.054013	0.429417	-0.063505	0.198906	1.010101	-0.011667	0.548407	0.135573	0.010901	0.189904	0.275730	0.307747
ProdLine	0.482317	-0.097316	0.194571	0.567088	-0.011667	1.010101	-0.063216	-0.499948	0.275836	0.428152	0.606336	0.556107
SalesFImage	-0.147978	0.787115	0.009936	0.228937	0.548407	-0.063216	1.010101	0.273986	0.101972	0.196662	0.273952	0.506129
ComPricing	-0.405335	0.270772	-0.273522	-0.129247	0.135573	-0.499948	0.273986	1.010101	-0.247461	-0.114463	-0.070999	-0.210400
WartyClaim	0.089204	0.027657	0.805220	0.141827	0.010901	0.275836	0.101972	-0.247461	1.010101	0.200107	0.117342	0.179338
OrdBilling	0.103531	0.147985	0.086307	0.765652	0.189904	0.428152	0.196662	-0.114463	0.200107	1.010101	0.759896	0.526590
DelSpeed	0.024577	0.169845	0.029190	0.877623	0.275730	0.606336	0.273952	-0.070999	0.117342	0.759896	1.010101	0.583217
Satisfaction	0.491237	0.244305	0.113735	0.609356	0.307747	0.556107	0.506129	-0.210400	0.179338	0.526590	0.583217	1.010101

Table 9: covariance matrix after scaling



	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ProdQual	1.000000	-0.161588	0.095600	0.106370	-0.053473	0.477493	-0.146498	-0.401282	0.088312	0.102495	0.024332	0.486325
Ecom	-0.161588	1.000000	-0.018786	0.109386	0.425123	-0.096342	0.779244	0.268064	0.027380	0.146505	0.168147	0.241862
TechSup	0.095600	-0.018786	1.000000	0.096657	-0.062870	0.192625	0.009836	-0.270787	0.797168	0.085443	0.028898	0.112597
CompRes	0.106370	0.109386	0.096657	1.000000	0.196917	0.561417	0.226647	-0.127954	0.140408	0.757995	0.868846	0.603263
Advertising	-0.053473	0.425123	-0.062870	0.196917	1.000000	-0.011551	0.542923	0.134217	0.010792	0.188005	0.272973	0.304669
ProdLine	0.477493	-0.096342	0.192625	0.561417	-0.011551	1.000000	-0.062584	-0.494948	0.273078	0.423870	0.600272	0.550546
SalesFImage	-0.146498	0.779244	0.009836	0.226647	0.542923	-0.062584	1.000000	0.271246	0.100953	0.194695	0.271213	0.501068
ComPricing	-0.401282	0.268064	-0.270787	-0.127954	0.134217	-0.494948	0.271246	1.000000	-0.244986	-0.113318	-0.070289	-0.208296
WartyClaim	0.088312	0.027380	0.797168	0.140408	0.010792	0.273078	0.100953	-0.244986	1.000000	0.198106	0.116168	0.177545
OrdBilling	0.102495	0.146505	0.085443	0.757995	0.188005	0.423870	0.194695	-0.113318	0.198106	1.000000	0.752298	0.521324
DelSpeed	0.024332	0.168147	0.028898	0.868846	0.272973	0.600272	0.271213	-0.070289	0.116168	0.752298	1.000000	0.577385
Satisfaction	0.486325	0.241862	0.112597	0.603263	0.304669	0.550546	0.501068	-0.208296	0.177545	0.521324	0.577385	1.000000

Table 10: correlation matrix after scaling

Observation:

- ➢ Covariance and Correlation measures aid in establishing this. Covariance brings about the variation across variables. We use covariance to measure how much two variables change with each other. We use correlation to determine how strongly linked two variables are to each other.
- ➢ If you scale (z-score) your data then your covariance matrix is a correlation matrix.

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

Before Scaling:

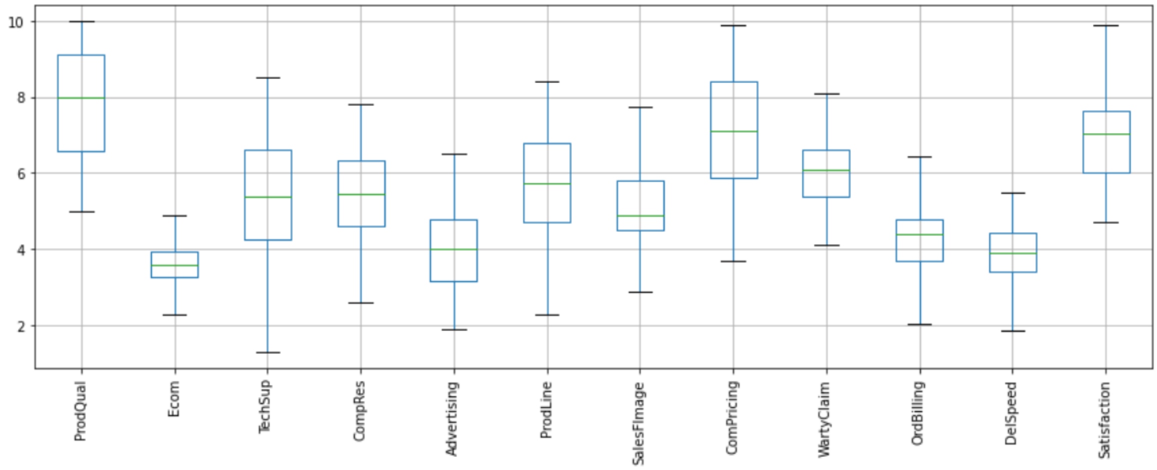


Figure 8: Before Scaling

After Scaling:

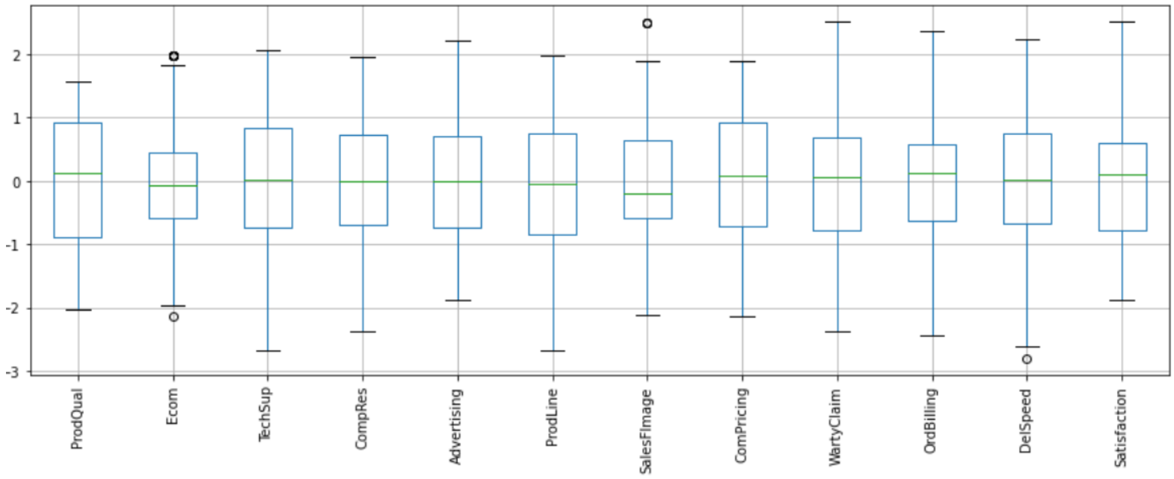


Figure 9: After Scaling

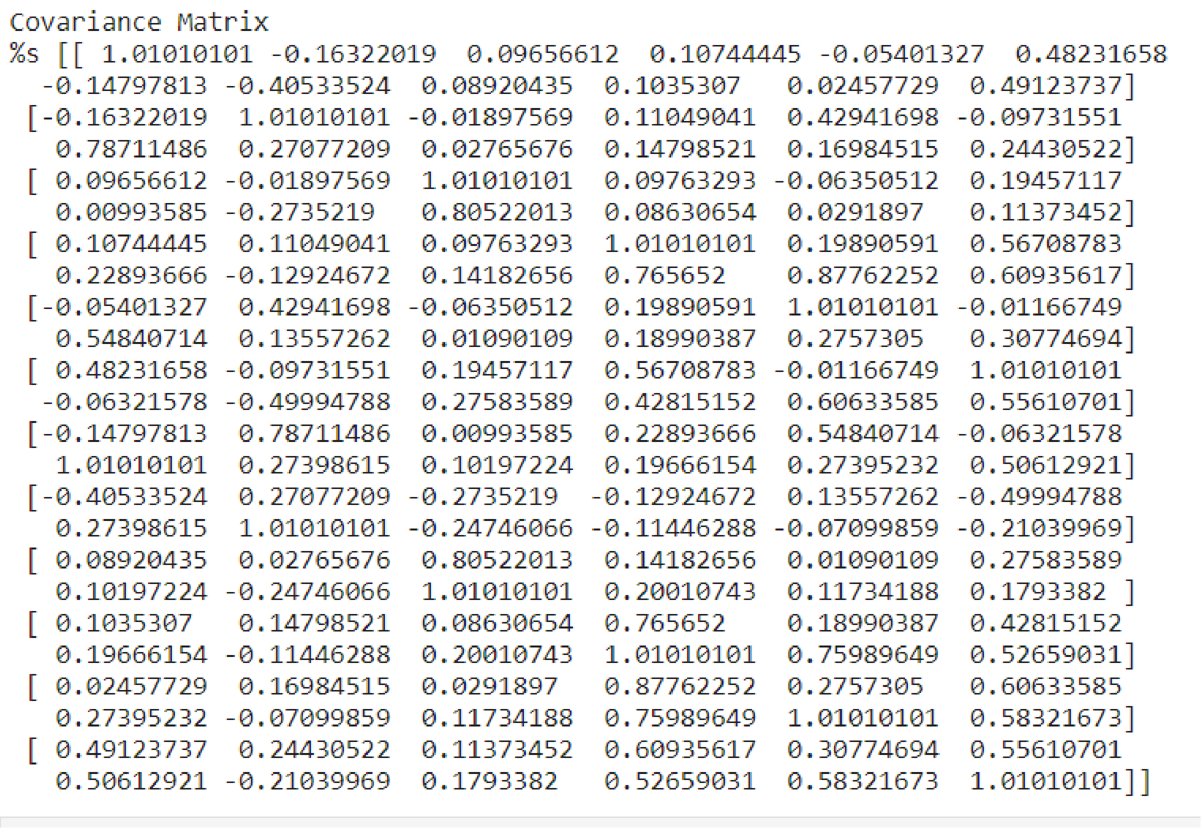
Obsevation:

- Normalization is used to transform all variables in the data to a same range. It doesn't solve the problem caused by outliers. So, as you can see the after normalization also, the outliers remain outliers. Only the range is changed.
- Removal of outliers creates a normal distribution in some of my variables, and makes transformations for the other variables more effective. Therefore, it seems that removal of outliers before transformation is the better option.
- Compare the respective medians of each box plot. If the median line of a box plot lies outside of the box of a comparison box plot, then there is likely to be a difference between the two groups.

- Scaling should be done to center data before performing PCA since the transformation relies on the data being around the origin. Some data might already follow a standard normal distribution with mean zero and standard deviation of one and so would not have to be scaled before PCA. So PCA will not give accurate results.

2.5 Build the covariance matrix, eigenvalues and eigenvector?

covariance matrix:

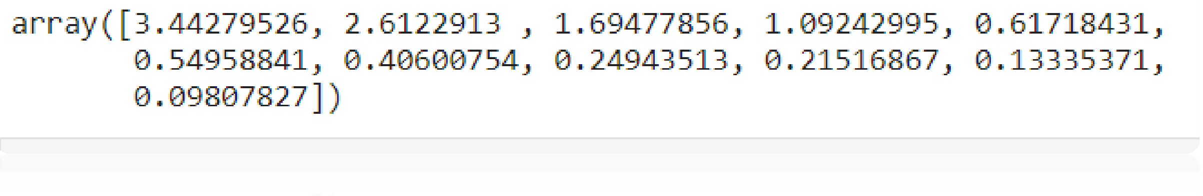


```
Covariance Matrix
%s [[ 1.01010101 -0.16322019  0.09656612  0.10744445 -0.05401327  0.48231658
  -0.14797813 -0.40533524  0.08920435  0.1035307   0.02457729  0.49123737]
 [-0.16322019  1.01010101 -0.01897569  0.11049041  0.42941698 -0.09731551
   0.78711486  0.27077209  0.02765676  0.14798521  0.16984515  0.24430522]
 [ 0.09656612 -0.01897569  1.01010101  0.09763293 -0.06350512  0.19457117
   0.00993585 -0.2735219   0.80522013  0.08630654  0.0291897   0.11373452]
 [ 0.10744445  0.11049041  0.09763293  1.01010101  0.19890591  0.56708783
   0.22893666 -0.12924672  0.14182656  0.765652    0.87762252  0.60935617]
 [-0.05401327  0.42941698 -0.06350512  0.19890591  1.01010101 -0.01166749
   0.54840714  0.13557262  0.01090109  0.18990387  0.2757305   0.30774694]
 [ 0.48231658 -0.09731551  0.19457117  0.56708783 -0.01166749  1.01010101
  -0.06321578 -0.49994788  0.27583589  0.42815152  0.60633585  0.55610701]
 [-0.14797813  0.78711486  0.00993585  0.22893666  0.54840714 -0.06321578
   1.01010101  0.27398615  0.10197224  0.19666154  0.27395232  0.50612921]
 [-0.40533524  0.27077209 -0.2735219  -0.12924672  0.13557262 -0.49994788
   0.27398615  1.01010101 -0.24746066 -0.11446288 -0.07099859 -0.21039969]
 [ 0.08920435  0.02765676  0.80522013  0.14182656  0.01090109  0.27583589
   0.10197224 -0.24746066  1.01010101  0.20010743  0.11734188  0.1793382 ]
 [ 0.1035307   0.14798521  0.08630654  0.765652    0.18990387  0.42815152
   0.19666154 -0.11446288  0.20010743  1.01010101  0.75989649  0.52659031]
 [ 0.02457729  0.16984515  0.0291897   0.87762252  0.2757305   0.60633585
   0.27395232 -0.07099859  0.11734188  0.75989649  1.01010101  0.58321673]
 [ 0.49123737  0.24430522  0.11373452  0.60935617  0.30774694  0.55610701
   0.50612921 -0.21039969  0.1793382   0.52659031  0.58321673  1.01010101]]
```

Principle Component Analysis(PCA):

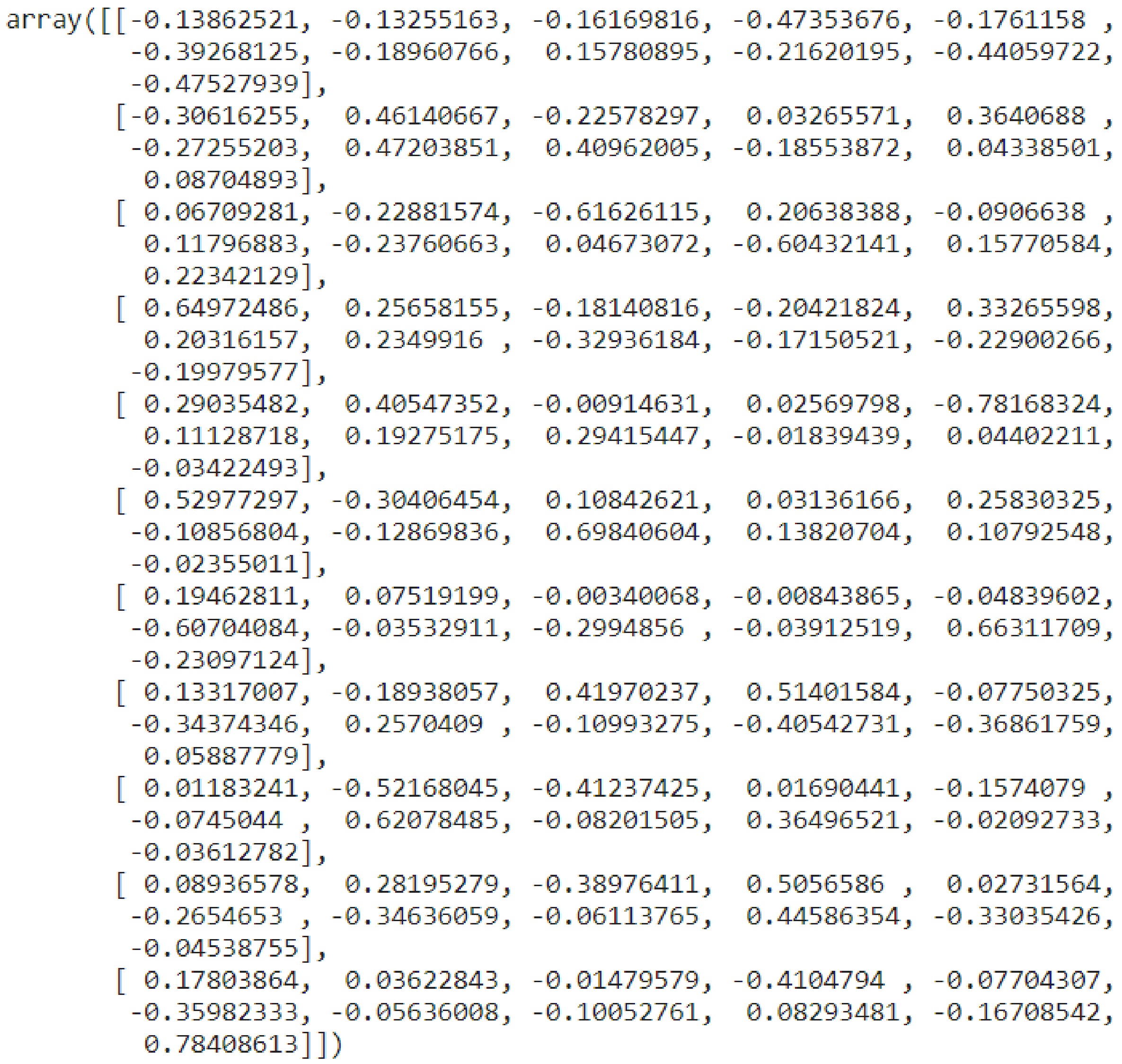
- Principal Components represent the directions of the data that specify a maximum amount of variance i.e. the lines that capture most information present in the data.
- Principal Components as new axes that provide the best angle to visualize and evaluate the data, in order that the difference between the observations is best visible.

Eigen Values:



```
array([3.44279526, 2.6122913 , 1.69477856, 1.09242995, 0.61718431,
       0.54958841, 0.40600754, 0.24943513, 0.21516867, 0.13335371,
       0.09807827])
```

Eigen Vector:



```
array([[-0.13862521, -0.13255163, -0.16169816, -0.47353676, -0.1761158 ,
        -0.39268125, -0.18960766,  0.15780895, -0.21620195, -0.44059722,
        -0.47527939],
       [-0.30616255,  0.46140667, -0.22578297,  0.03265571,  0.3640688 ,
        -0.27255203,  0.47203851,  0.40962005, -0.18553872,  0.04338501,
         0.08704893],
       [ 0.06709281, -0.22881574, -0.61626115,  0.20638388, -0.0906638 ,
         0.11796883, -0.23760663,  0.04673072, -0.60432141,  0.15770584,
         0.22342129],
       [ 0.64972486,  0.25658155, -0.18140816, -0.20421824,  0.33265598,
         0.20316157,  0.2349916 , -0.32936184, -0.17150521, -0.22900266,
        -0.19979577],
       [ 0.29035482,  0.40547352, -0.00914631,  0.02569798, -0.78168324,
         0.11128718,  0.19275175,  0.29415447, -0.01839439,  0.04402211,
        -0.03422493],
       [ 0.52977297, -0.30406454,  0.10842621,  0.03136166,  0.25830325,
        -0.10856804, -0.12869836,  0.69840604,  0.13820704,  0.10792548,
        -0.02355011],
       [ 0.19462811,  0.07519199, -0.00340068, -0.00843865, -0.04839602,
        -0.60704084, -0.03532911, -0.2994856 , -0.03912519,  0.66311709,
        -0.23097124],
       [ 0.13317007, -0.18938057,  0.41970237,  0.51401584, -0.07750325,
        -0.34374346,  0.2570409 , -0.10993275, -0.40542731, -0.36861759,
         0.05887779],
       [ 0.01183241, -0.52168045, -0.41237425,  0.01690441, -0.1574079 ,
        -0.0745044 ,  0.62078485, -0.08201505,  0.36496521, -0.02092733,
        -0.03612782],
       [ 0.08936578,  0.28195279, -0.38976411,  0.5056586 ,  0.02731564,
        -0.2654653 , -0.34636059, -0.06113765,  0.44586354, -0.33035426,
        -0.04538755],
       [ 0.17803864,  0.03622843, -0.01479579, -0.4104794 , -0.07704307,
        -0.35982333, -0.05636008, -0.10052761,  0.08293481, -0.16708542,
         0.78408613]])
```

Bartletts Test of Sphericity:

- Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population:

Null and Alternative Hypothesis:

- H0: All variables in the data are uncorrelated.
- Ha: At least one pair of variables in the data are correlated.

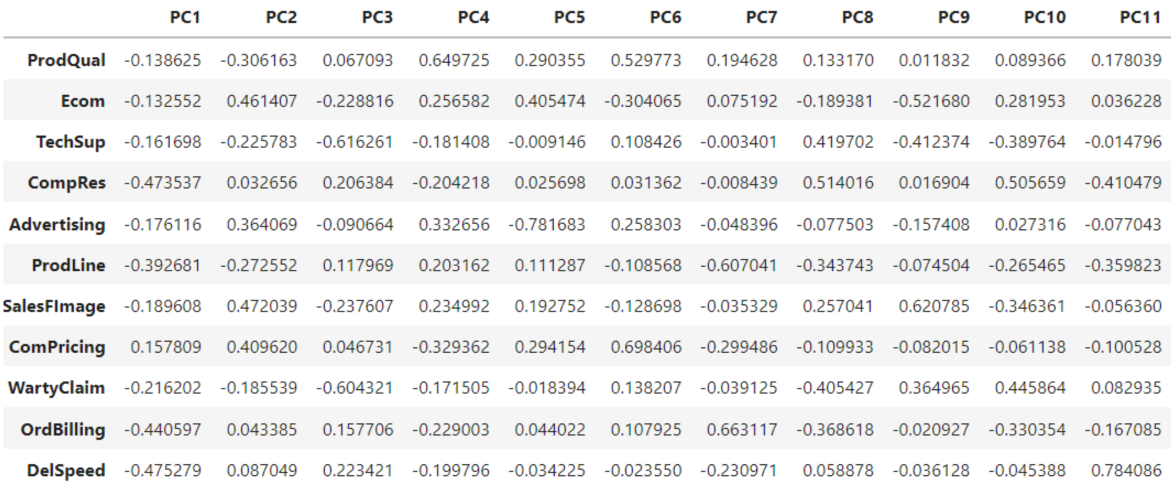
KMO Test:

- The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

Observation:

- The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.
- Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

Create a dataframe containing the loadings or coefficients of all PCs:



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
ProdQual	-0.138625	-0.306163	0.067093	0.649725	0.290355	0.529773	0.194628	0.133170	0.011832	0.089366	0.178039
Ecom	-0.132552	0.461407	-0.228816	0.256582	0.405474	-0.304065	0.075192	-0.189381	-0.521680	0.281953	0.036228
TechSup	-0.161698	-0.225783	-0.616261	-0.181408	-0.009146	0.108426	-0.003401	0.419702	-0.412374	-0.389764	-0.014796
CompRes	-0.473537	0.032656	0.206384	-0.204218	0.025698	0.031362	-0.008439	0.514016	0.016904	0.505659	-0.410479
Advertising	-0.176116	0.364069	-0.090664	0.332656	-0.781683	0.258303	-0.048396	-0.077503	-0.157408	0.027316	-0.077043
ProdLine	-0.392681	-0.272552	0.117969	0.203162	0.111287	-0.108568	-0.607041	-0.343743	-0.074504	-0.265465	-0.359823
SalesFImage	-0.189608	0.472039	-0.237607	0.234992	0.192752	-0.128698	-0.035329	0.257041	0.620785	-0.346361	-0.056360
ComPricing	0.157809	0.409620	0.046731	-0.329362	0.294154	0.698406	-0.299486	-0.109933	-0.082015	-0.061138	-0.100528
WartyClaim	-0.216202	-0.185539	-0.604321	-0.171505	-0.018394	0.138207	-0.039125	-0.405427	0.364965	0.445864	0.082935
OrdBilling	-0.440597	0.043385	0.157706	-0.229003	0.044022	0.107925	0.663117	-0.368618	-0.020927	-0.330354	-0.167085
DelSpeed	-0.475279	0.087049	0.223421	-0.199796	-0.034225	-0.023550	-0.230971	0.058878	-0.036128	-0.045388	0.784086

Table 11: Data frame containing coefficients of all PCs

Scree Plot to decide on the Number of PC'S:

- A Scree plot always displays the Eigen values in a downward curve, ordering the eigenvalues from largest to smallest. According to the Scree test, the ”elbow” of the graph where the Eigenvalues seem to level off is found and factors or components to the left of this point should be retained as significant.
- The Scree Plot is used to determine the number of Principal Components to keep in a Principal Component Analysis (PCA).

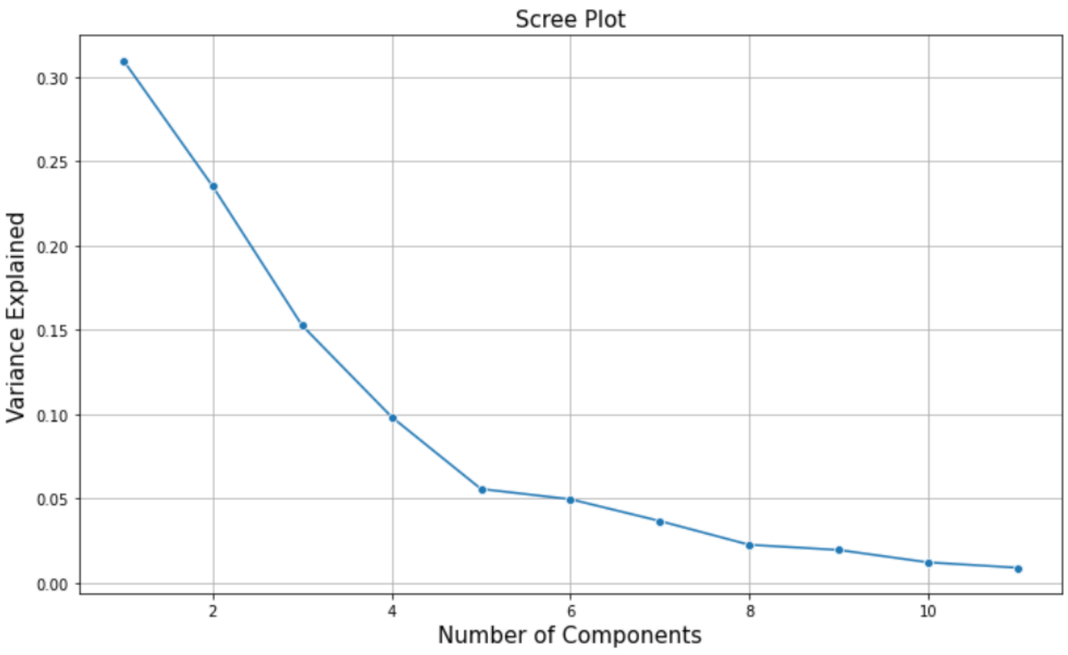
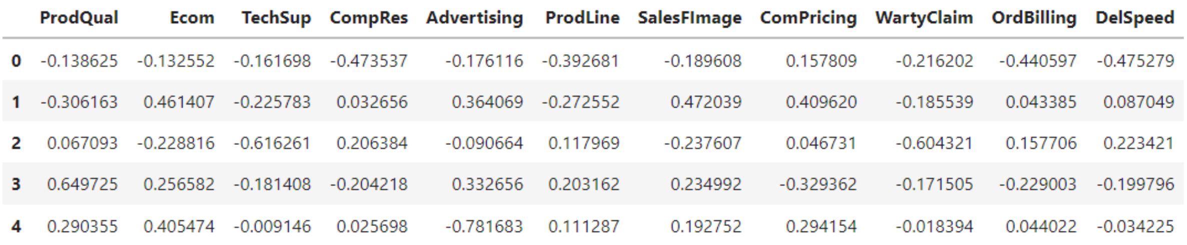


Figure 10:Scree Plot

Observation:

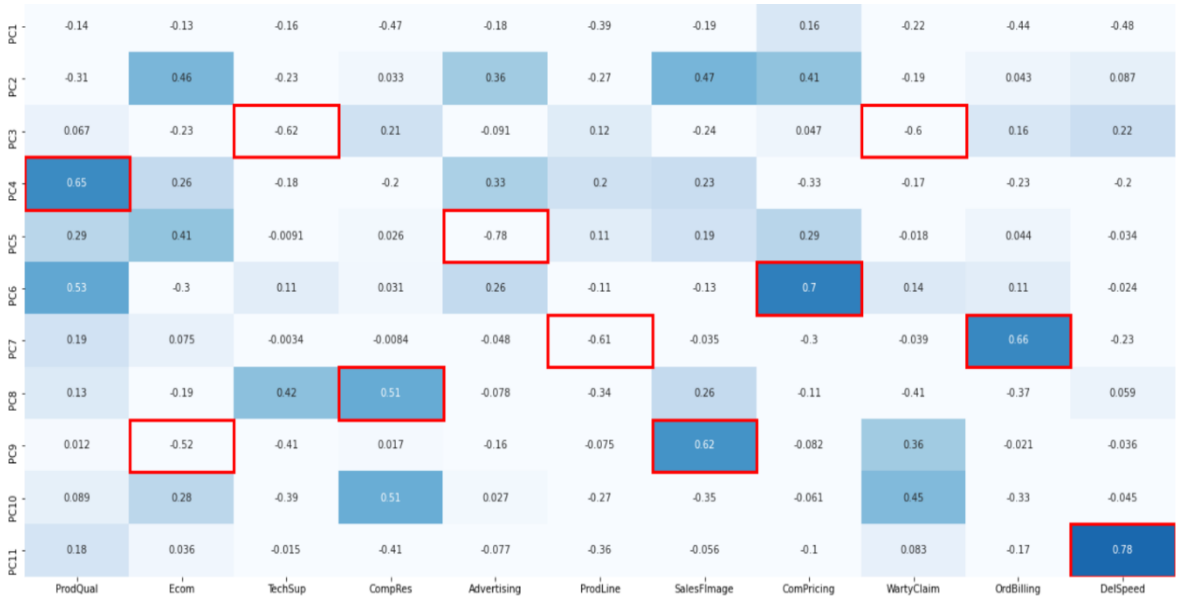
- we can see that it is basically an elbow-shaped graph. Where in for our dataset up until 5 Principal Components there is a significant variation or change in the variance explained. So this by looking at the Scree Plot we can consider the number of Principal Components to be 5 for our dataset.



	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
0	-0.138625	-0.132552	-0.161698	-0.473537	-0.176116	-0.392681	-0.189608	0.157809	-0.216202	-0.440597	-0.475279
1	-0.306163	0.461407	-0.225783	0.032656	0.364069	-0.272552	0.472039	0.409620	-0.185539	0.043385	0.087049
2	0.067093	-0.228816	-0.616261	0.206384	-0.090664	0.117969	-0.237607	0.046731	-0.604321	0.157706	0.223421
3	0.649725	0.256582	-0.181408	-0.204218	0.332656	0.203162	0.234992	-0.329362	-0.171505	-0.229003	-0.199796
4	0.290355	0.405474	-0.009146	0.025698	-0.781683	0.111287	0.192752	0.294154	-0.018394	0.044022	-0.034225

Table 12: Extracted Data using PCA

- Let's identify which features have maximum loading across the components.
- We will first plot the component loading on a heatmap.
- For each feature, we find the maximum loading value across the components and mark the same with help of rectangular box.
- Features marked with rectangular red box are the one having maximum loading on the respective component. We consider these marked features to decide the context that the component represents.



	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed
PC1	-0.14	-0.13	-0.16	-0.47	-0.18	-0.39	-0.19	0.16	-0.22	-0.44	-0.48
PC2	-0.31	0.46	-0.23	0.033	0.36	-0.27	0.47	0.41	-0.19	0.043	0.087
PC3	0.067	-0.23	-0.62	0.21	-0.091	0.12	-0.24	0.047	-0.6	0.16	0.22
PC4	0.65	0.26	-0.18	-0.2	0.33	0.2	0.23	-0.33	-0.17	-0.23	-0.2
PC5	0.29	0.41	-0.0091	0.026	-0.78	0.11	0.19	0.29	-0.018	0.044	-0.034
PC6	0.53	-0.3	0.11	0.031	0.26	-0.11	-0.13	0.7	0.14	0.11	-0.024
PC7	0.19	0.075	-0.0034	-0.0084	-0.048	-0.61	-0.035	-0.3	-0.039	0.66	-0.23
PC8	0.13	-0.19	0.42	0.51	-0.078	-0.34	0.26	-0.11	-0.41	-0.37	0.059
PC9	0.012	-0.52	-0.41	0.017	-0.16	-0.075	0.62	-0.082	0.36	-0.021	-0.036
PC10	0.089	0.28	-0.39	0.51	0.027	-0.27	-0.35	-0.061	0.45	-0.33	-0.045
PC11	0.18	0.036	-0.015	-0.41	-0.077	-0.36	-0.056	-0.1	0.083	-0.17	0.78

Figure 11: Heat Map with all Components

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

PC1=-0.13_ProdQual-0.17_Ecom-0.16_TechSup-0.47_CompRes-0.18_Advertising-0.39_ProdLine-0.2_SalesFImage+0.15_ComPricing-0.21_WartyClaim-0.44_OrdBilling-0.47*DelSpeed.

Observation:

- It shows how many times component 1(PC0) related to each features in dataset
- Eg: It relates PC0 is -0.13 times of Prod Qual.

2.7 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

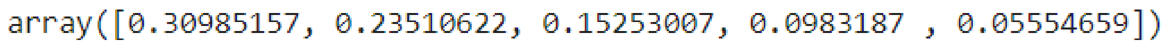
Cumulative values of the eigenvalues:

```
Cumulative Variance Explained: [ 30.98515733  54.49577899  69.74878605  79.58065556  85.13531437
  90.0816101   93.73567797  95.98059413  97.91711216  99.11729559
 100.        ]
```

PCA for Optimal Numbers:

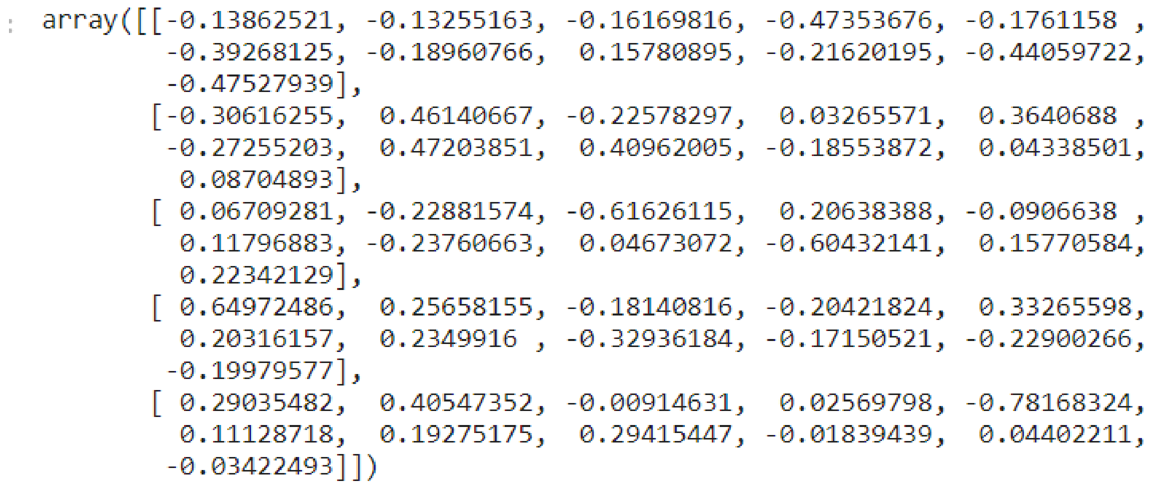
- Apply PCA for the number of decided components to get the loadings and component output
- Using scikit learn PCA here. It does all the above steps and maps data to PCA dimensions in one shot
- NOTE -we are generating only 5 PCA dimensions (dimensionality reduction from 11 to 5)

Eigen Values for Optimal Values:



```
array([0.30985157, 0.23510622, 0.15253007, 0.0983187 , 0.05554659])
```

Eigen Vector for Optimal Values:



```
: array([[-0.13862521, -0.13255163, -0.16169816, -0.47353676, -0.1761158 ,
          -0.39268125, -0.18960766,  0.15780895, -0.21620195, -0.44059722,
          -0.47527939],
         [-0.30616255,  0.46140667, -0.22578297,  0.03265571,  0.3640688 ,
          -0.27255203,  0.47203851,  0.40962005, -0.18553872,  0.04338501,
           0.08704893],
         [ 0.06709281, -0.22881574, -0.61626115,  0.20638388, -0.0906638 ,
           0.11796883, -0.23760663,  0.04673072, -0.60432141,  0.15770584,
           0.22342129],
         [ 0.64972486,  0.25658155, -0.18140816, -0.20421824,  0.33265598,
           0.20316157,  0.2349916 , -0.32936184, -0.17150521, -0.22900266,
          -0.19979577],
         [ 0.29035482,  0.40547352, -0.00914631,  0.02569798, -0.78168324,
           0.11128718,  0.19275175,  0.29415447, -0.01839439,  0.04402211,
          -0.03422493]])
```

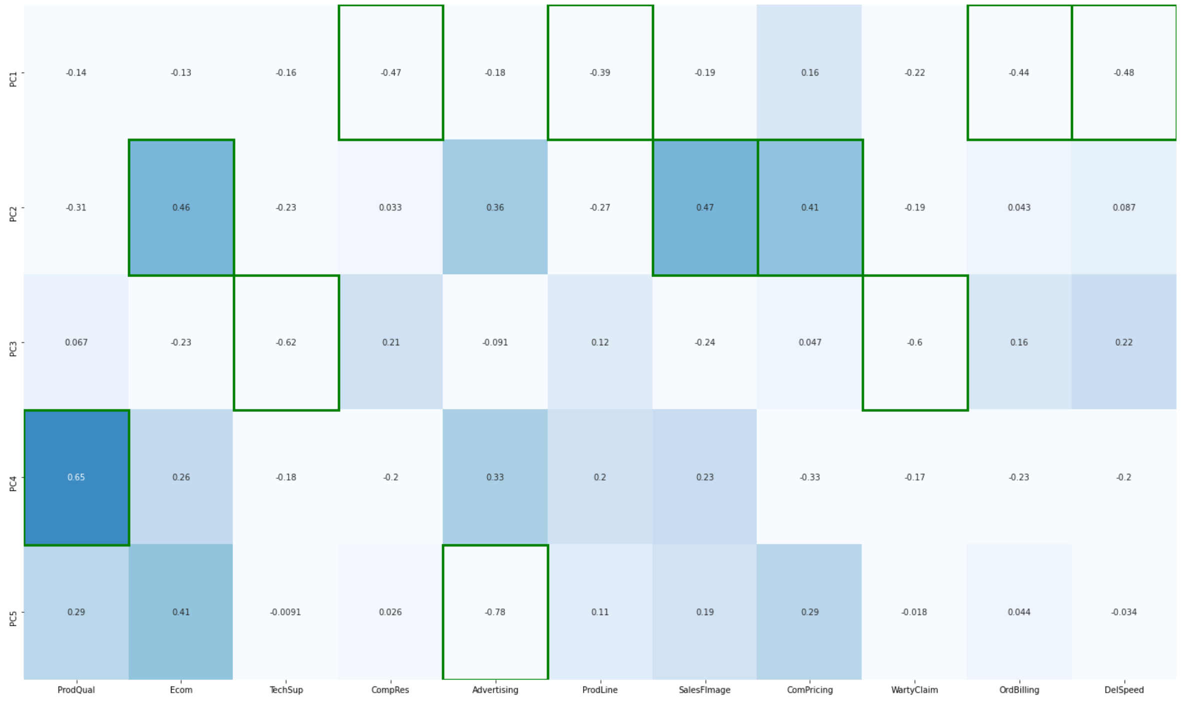
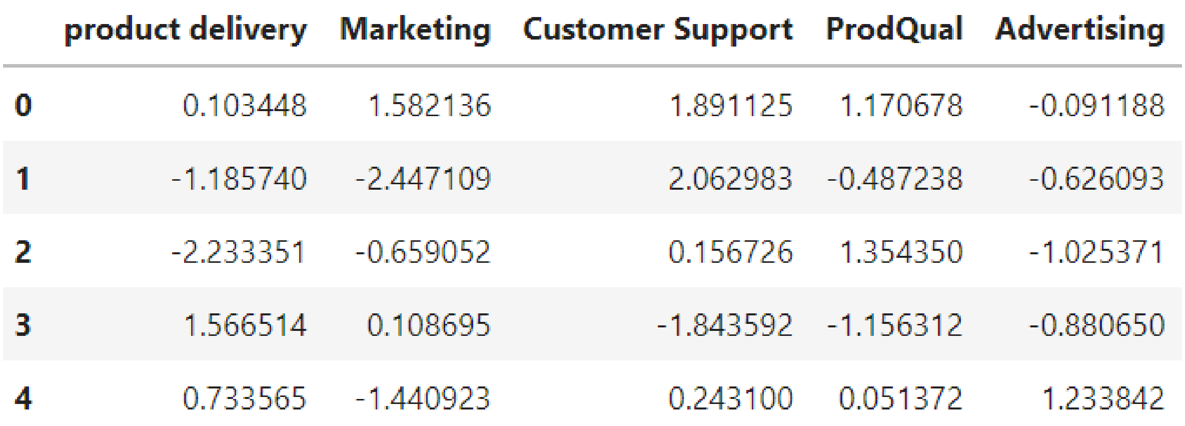


Figure 12: Heat Map with optimal components

Observation:

- PC1 is the linear combination of feature Comp Res, Product Line, Order & Billing and Del Speed.
- PC2 is the linear combination of feature E-com, SalesFImage, Com Pricing.
- PC3 is the linear combination of feature Tech Sup and Warrenty Claim.
- PC4 has Prod Qual.
- PC5 has Advertising.

New DataFrame using Principal Component scores:



	product delivery	Marketing	Customer Support	ProdQual	Advertising
0	0.103448	1.582136	1.891125	1.170678	-0.091188
1	-1.185740	-2.447109	2.062983	-0.487238	-0.626093
2	-2.233351	-0.659052	0.156726	1.354350	-1.025371
3	1.566514	0.108695	-1.843592	-1.156312	-0.880650
4	0.733565	-1.440923	0.243100	0.051372	1.233842

Table 13: Principal Component scores with reduced features

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

- Large datasets are increasingly common and are often difficult to interpret. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss.
- It does so by creating new uncorrelated variables that successively maximize variance. Finding such new variables, the principal components, reduces to solving an eigenvalue /eigenvector problem, and the new variables are defined by the dataset at hand, not a priority, hence making PCA an adaptive data analysis technique.
- Customer need to engage with advertisement and promotional offers , so it will increase the revenue. we need to deep dive in to customer satisfaction who has given low rating and check which product they purchased. we should call the customer and get feedback from them. According to feedback from customers we need to analyse the product to retain the customer.
- Low rating in product quality we need to coordinate with quality team and shuld take a call to increase the quality of product.