This Business Report shall provide detailed explanation of how we approached each problem given in the assignment. It shall also provide relative resolution and explanation with regards to the problems

Advanced Statistics – Project

Business Report | July 2021

Thakur Arun Singh

Contents

Problem A1: 2

Problem A1.1 2

Problem A1.2 3

Problem A1.3 3

Problem A1.4 4

Problem A1.5 5

Problem A1.6 6

Problem 2: 6

Problem 2.1 6

Problem 2.2 12

Problem 2.3 13

Problem 2.4 14

Problem 2.5 15

Problem 2.6 18

Problem 2.7 19

Problem 2.8 20

Problem A1:

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. The data for this particular experiment is thus attached.

Problem A1.1

State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables 'Manufacturer' and 'Technician individually?

Resolution:

- First we import all the necessary libraries in Python, and then import the data file which is 'Service (2)'. Once we import the file we confirm whether the data has been uploaded correctly or not using 'head' function. Using this function we can view the data and all the columns and headers whether they are aligning correctly or not.
- Then using the 'shape' function we can understand how many row and columns are there in our data set
- To check the data type of all the columns and also to check the null values, 'info' function. Has been used.
- To see the detail description of the data such as, Count, Mean, Median, Min, Max, Standard Deviations etc,

	count	mean	std	min	25%	50%	75%	max
Technician	45	2	0.825723	1	1	2	3	3
Manufacturer	45	2	0.825723	1	1	2	3	3
Job	45	3	1.430194	1	2	3	4	5
Service Time	45	55.822222	8.448477	39	50	56	62	70

- Using the 'isnull' function, one can understand if there are any null values in the data set. And we do not have any null values in the existing data set.
- Using the 'dups' function we check for the duplicates and there were no duplicate values.

FORMING HYPOTHESIS FOR ONE WAY ANOVA

Variable Technician

Null hypothesis states that there will be no effect of Technician at any job on its service time.

Alternate hypothesis states that there will be an effect of Technician in at least one job on its service time.

$H0$: μT1 = μT2 = μT3

$H1$: μT1 ≠ μT2 = μT3 or μT1 = μT2 ≠ μT3 or μT1 ≠ μT2 ≠ μT3

Variable Manufacturer

Null hypothesis states that there will be no effect of Manufacturer at any job on its service time.

Alternate hypothesis states that there will be an effect of Manufacturer in at least one job on its service time.

$H0$: µM1 = µM2 = µM3

$H1$: µM1 ≠ µM2 = µM3 or µM1 = µM2 ≠ µM3 or µM1 ≠ µM2 ≠ µM3

Problem A1.2

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.

Resolution:

One-way ANOVA for variable 'Manufacturer' with respect to the variable 'Service Time'

Level of Significance i.e, α = 0.05

```
In [7]: formula = ols('Service_Time ~ Manufacturer', data = df).fit()
        aov_table = sm.stats.anova_lm(formula, type=1)
        print(aov_table)
```

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

Interpretation: The P-value obtained from the ANOVA analysis for Manufacturer is statistically significant as (P > 0.05), we fail to reject the Null Hypothesis. This means that there is no effect of Manufacturer on the Service Time.

Problem A1.3

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.

Resolution:

One-way ANOVA for variable 'Technician' with respect to the variable 'Service Time'

Level of Significance i.e, α = 0.05

```
In [8]: formula1 = ols('Service_Time ~ Technician', data = df).fit()
aov_table1 = sm.stats.anova_lm(formula1, type =1)
print(aov_table1)
```

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

Interpretation: The P-value obtained from the ANOVA analysis for Technician is statistically significant as (P > 0.05), we fail to reject the Null Hypothesis. This means that there is no effect of Technician on the Service Time.

Problem A1.4

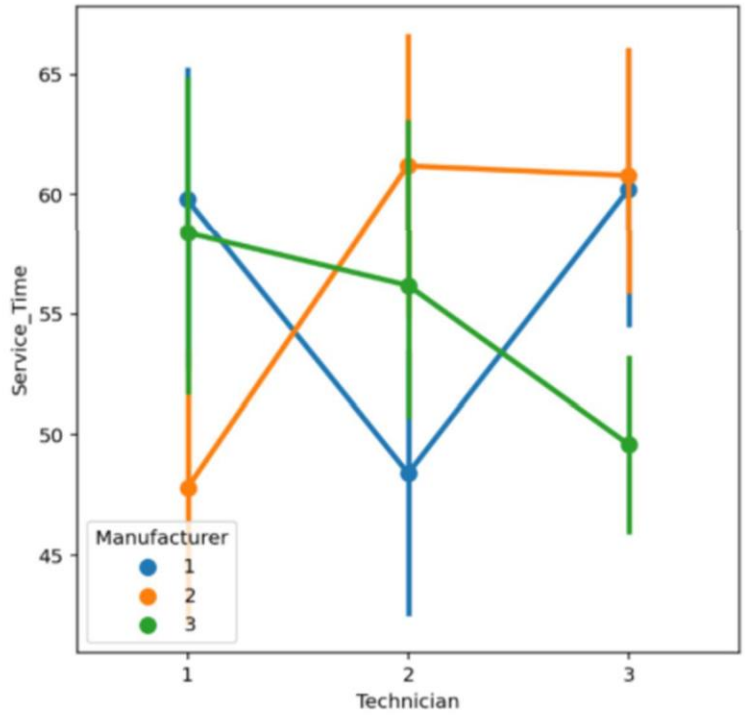
Analyze the effects of one variable on another with the help of an interaction plot. What is an interaction between two treatments?
[hint: use the 'pointplot' function from the 'seaborn' graphical subroutine in Python]

Resolution:

Interaction between the variables

The Interaction between the variables can be seen using the "pointplot" in the Seaborn library.

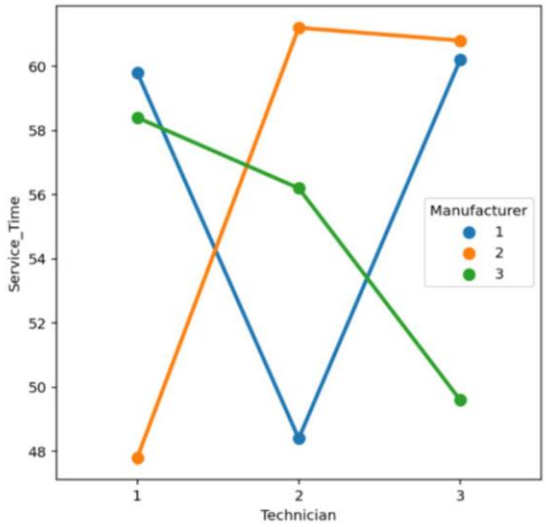
```
In [9]: plt.figure(figsize = (6,6))
sns.pointplot(x= 'Technician' , y = 'Service_Time', hue='Manufacturer', data = df);
```



Interpretation: We can see through the plot that there is significant interaction between the variables, as they cross over overlap each other at various intervals, if there was no interactionthe lines would appear parallel to each other

Interaction plot after removing the confidence intervals

```
In [10]: plt.figure(figsize = (6,6))
         sns.pointplot(x= 'Technician' , y = 'Service_Time', hue='Manufacturer', data = df , ci=None);
```



Interpretation: Interaction effect occurs when the effect of one variable is dependent on the value of another variable. It indicates that a third variable could influence the relationship between an independent and a dependent variable in this case 'Service Time' is the outcome, which is influenced according to the 'Manufacturer' and 'Technician' i.e., indicating a relationship between the variables 'Manufacturer' and 'Technician' at some level.

Problem A1.5

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

Resolution:

Two way ANOVA analysis on the variables w.r.t Service Time

Level of Significance i.e, α = 0.05

```
In [11]: formula2 = ols('Service_Time ~ Manufacturer + Technician + Manufacturer*Technician', data = df).fit()
         aov_tbl = sm.stats.anova_lm(formula2 , type = 2)
         print(aov_tbl)
```

	df	sum_sq	mean_sq	F	PR(>F)
Manufacturer	2.0	28.311111	14.155556	0.272164	0.763283
Technician	2.0	24.577778	12.288889	0.236274	0.790779
Manufacturer:Technician	4.0	1215.288889	303.822222	5.841487	0.000994
Residual	36.0	1872.400000	52.011111	NaN	NaN

Interpretation: The p-value for the interaction between Manufacturer: Technician is 0.000994, i.e., less the level of significance, which indicates that the relationship between Manufacturer and Service Time depends on the value of Technician. Because the interaction effect between Manufacturer and Technician is statistically significant, we cannot interpret the main effects without considering the interaction effect.

Since the P value of the variables Manufacturer and Technician in Two-way Anova are greater than the Significance Level, we can say that they don't individually affect the Service Time.

The Interaction of the same variables significantly affects the Service Time, since the P value of their Interaction is less than the significance level.

Problem A1.6

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

Resolution:

The ANOVA test allows a comparison of two or more groups at once to determine if a relationship exists between them, also to determine variability between the samples and also within them

In this case study, ANOVA would help to study the effects of a Technician or a Manufacturer working on various jobs, on their Service Time

1. The lesser the Service Time, the more is the customer satisfaction.
2. Greater customer satisfaction leads to customer loyalty and increase in business.
3. It would also help the business to rectify if there are any particular issues which are causing higher service times.
4. To check for any inefficiency on the part of either the technician or the manufacturer.
5. To pin point if a particular electric appliance is requiring more amount of time to be serviced, also the frequency of the service etc.

Identifying and rectifying these issues would certainly help the business a great deal.

Problem 2:

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.

Problem 2.1

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

Resolution:

- First we import all the necessary libraries in Python, and then import the data file which is 'Service(2)'. Once we import the file we confirm whether the data has been uploaded correctly or not using 'head' function. Using this function we can view the data and all the columns and headers whether they are aligning correctly or not.
- Then using the 'shape' function we can understand how many row and columns are there in our data set.
- To check the data type of all the columns and also to check the null values, 'info' function. Has been used.
- To see the detail description of the data such as, Count, Mean, Median, Min, Max, Standard Deviations etc,

:

	count	mean	std	min	25%	50%	75%	max
ID	100.0	50.500	29.011492	1.0	25.750	50.50	75.250	100.0
ProdQual	100.0	7.810	1.396279	5.0	6.575	8.00	9.100	10.0
Ecom	100.0	3.672	0.700516	2.2	3.275	3.60	3.925	5.7
TechSup	100.0	5.365	1.530457	1.3	4.250	5.40	6.625	8.5
CompRes	100.0	5.442	1.208403	2.6	4.600	5.45	6.325	7.8
Advertising	100.0	4.010	1.126943	1.9	3.175	4.00	4.800	6.5
ProdLine	100.0	5.805	1.315285	2.3	4.700	5.75	6.800	8.4
SalesFImage	100.0	5.123	1.072320	2.9	4.500	4.90	5.800	8.2
ComPricing	100.0	6.974	1.545055	3.7	5.875	7.10	8.400	9.9
WartyClaim	100.0	6.043	0.819738	4.1	5.400	6.10	6.600	8.1
OrdBilling	100.0	4.278	0.928840	2.0	3.700	4.40	4.800	6.7
DelSpeed	100.0	3.886	0.734437	1.6	3.400	3.90	4.425	5.5
Satisfaction	100.0	6.918	1.191839	4.7	6.000	7.05	7.625	9.9

- Using the 'isnull' function, one can understand if there are any null values in the data set. And we do not have any null values in the existing data set.
- Using the 'dups' function we check for the duplicates and there were no duplicate values.
- We have both categorical and continuous data. For categorical data, we have we have cut, color and clarity. For continuous data, we have carat, depth, table and price.
- We also identified the unique values in categorical data.

After reviewing the data thoroughly, and based on the above analysis we can say that, we have seven variables, Mean and Median values are almost equal, and Standard deviation for 'Spending' is higher than other variables. There are no duplicates in the data set.

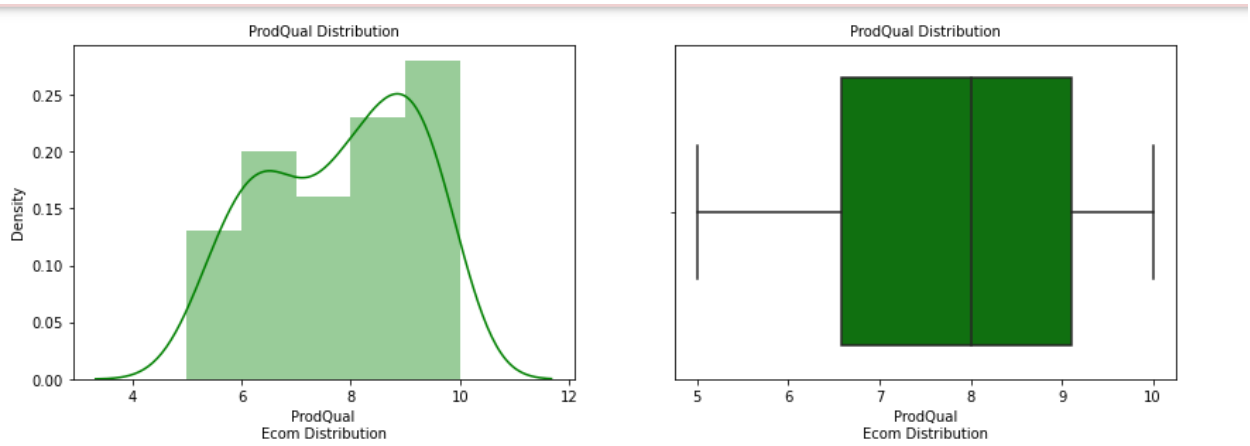
	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
0	1	8.5	3.9	2.5	5.9	4.8	4.9	6.0	6.8	4.7	5.0	3.7	8.2
1	2	8.2	2.7	5.1	7.2	3.4	7.9	3.1	5.3	5.5	3.9	4.9	5.7
2	3	9.2	3.4	5.6	5.6	5.4	7.4	5.8	4.5	6.2	5.4	4.5	8.9
3	4	6.4	3.3	7.0	3.7	4.7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

	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
95	96	8.6	4.8	5.6	5.3	2.3	6.0	5.7	6.7	5.8	4.9	3.6	7.3
96	97	7.4	3.4	2.6	5.0	4.1	4.4	4.8	7.2	4.5	4.2	3.7	6.3
97	98	8.7	3.2	3.3	3.2	3.1	6.1	2.9	5.6	5.0	3.1	2.5	5.4
98	99	7.8	4.9	5.8	5.3	5.2	5.3	7.1	7.9	6.0	4.3	3.9	6.4
99	100	7.9	3.0	4.4	5.1	5.9	4.2	4.8	9.7	5.7	3.4	3.5	6.4

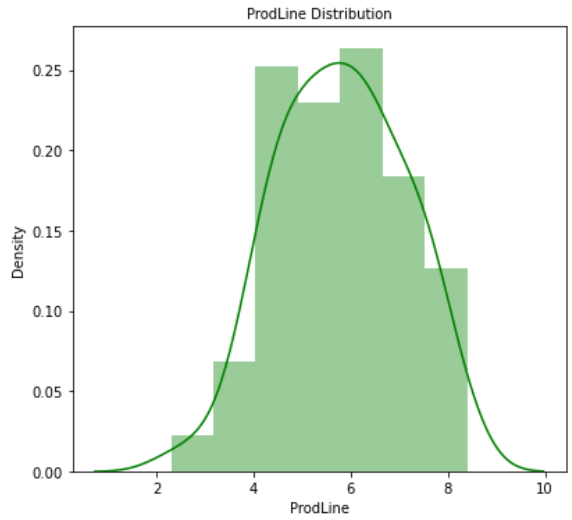
```
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   ID            100 non-null    int64
 1   ProdQual      100 non-null    float64
 2   Ecom          100 non-null    float64
 3   TechSup       100 non-null    float64
 4   CompRes       100 non-null    float64
 5   Advertising   100 non-null    float64
 6   ProdLine      100 non-null    float64
 7   SalesFImage   100 non-null    float64
 8   ComPricing    100 non-null    float64
 9   WartyClaim    100 non-null    float64
 10  OrdBilling    100 non-null    float64
 11  DelSpeed      100 non-null    float64
 12  Satisfaction  100 non-null    float64
dtypes: float64(12), int64(1)
memory usage: 10.3 KB
```

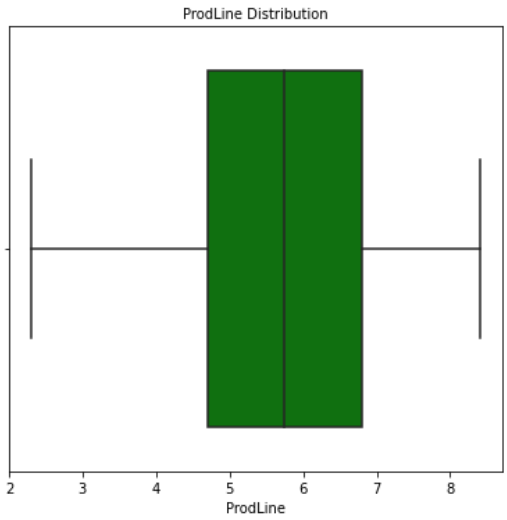
Exploratory data analysis

Univariate and multivariate analysis

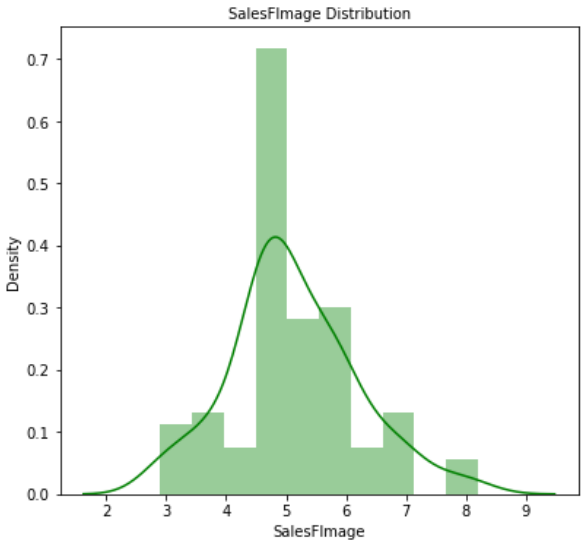


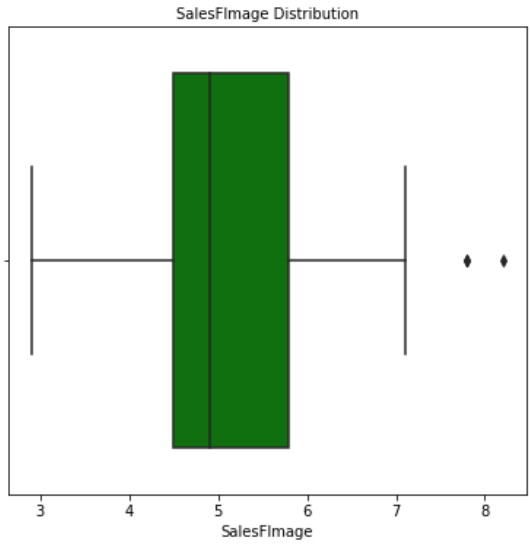
The distribution of data seems to be positively skewed as there are multiple peak points in the distribution there could be multimode and the box plot of ProdQual seems to have no outliers. In the range of 5 to 10, where majority of data resides.



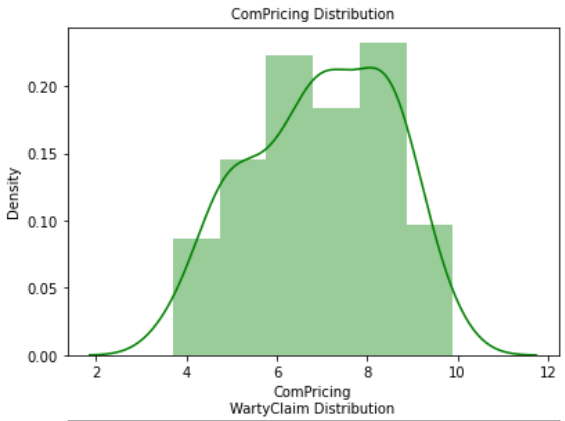


The above dist plot shows the normal distribution of data from 3 – 8. Boxplot shows that there are no outliers.

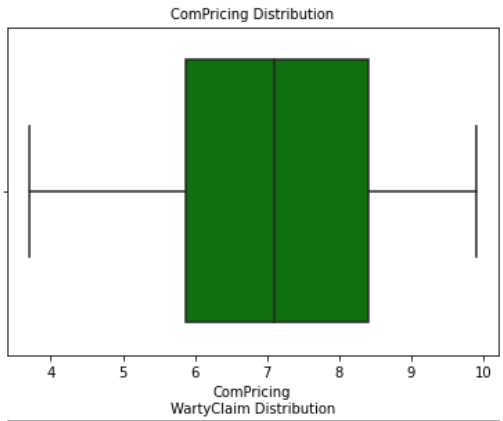




The above dist plot shows the distribution of data from 3 – 7 and is positively skewed. Boxplot shows that there are a few outliers.

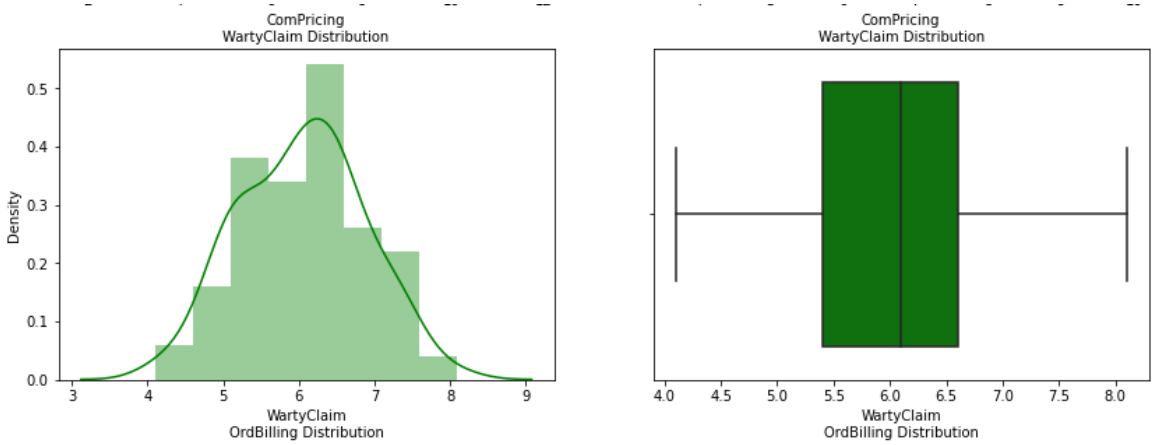


WartyClaim Distribution

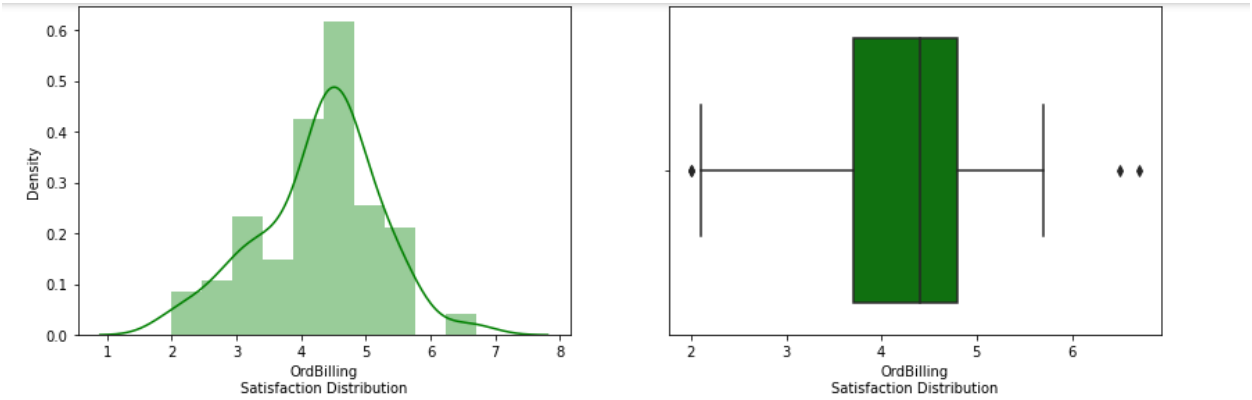


WartyClaim Distribution

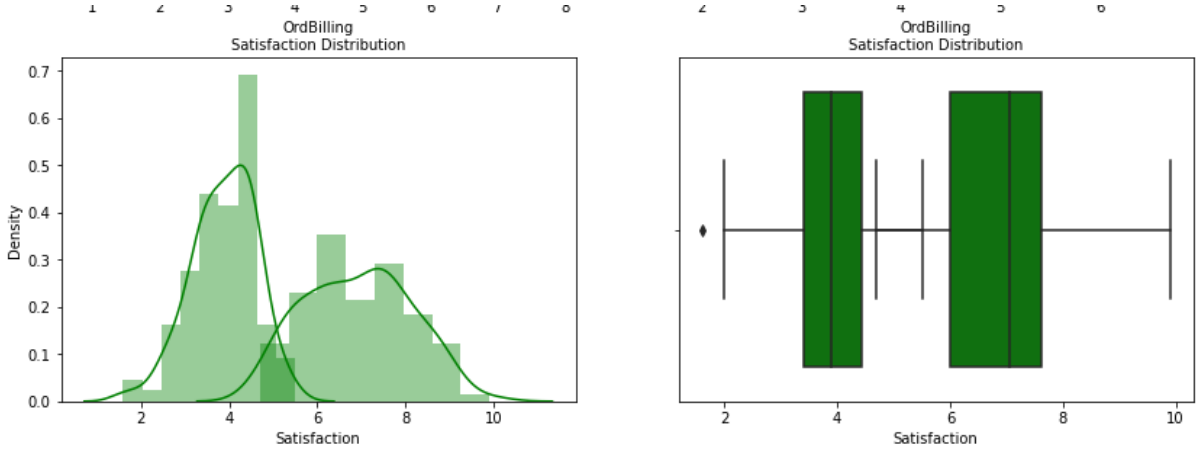
The above dist plot shows the distribution of data from 4 – 10 and is positively skewed. Boxplot shows that there are no outliers.



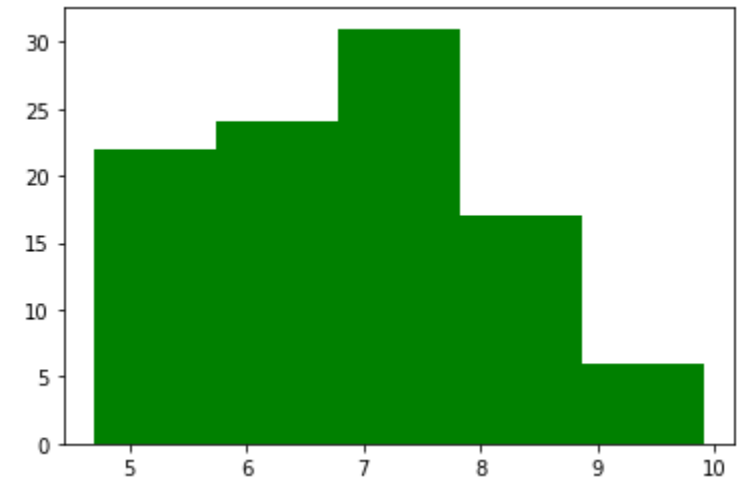
The above dist plot shows the distribution of data from 3 – 8 and is positively skewed. Boxplot shows that there are no outliers. The distribution is too much positively skewed.

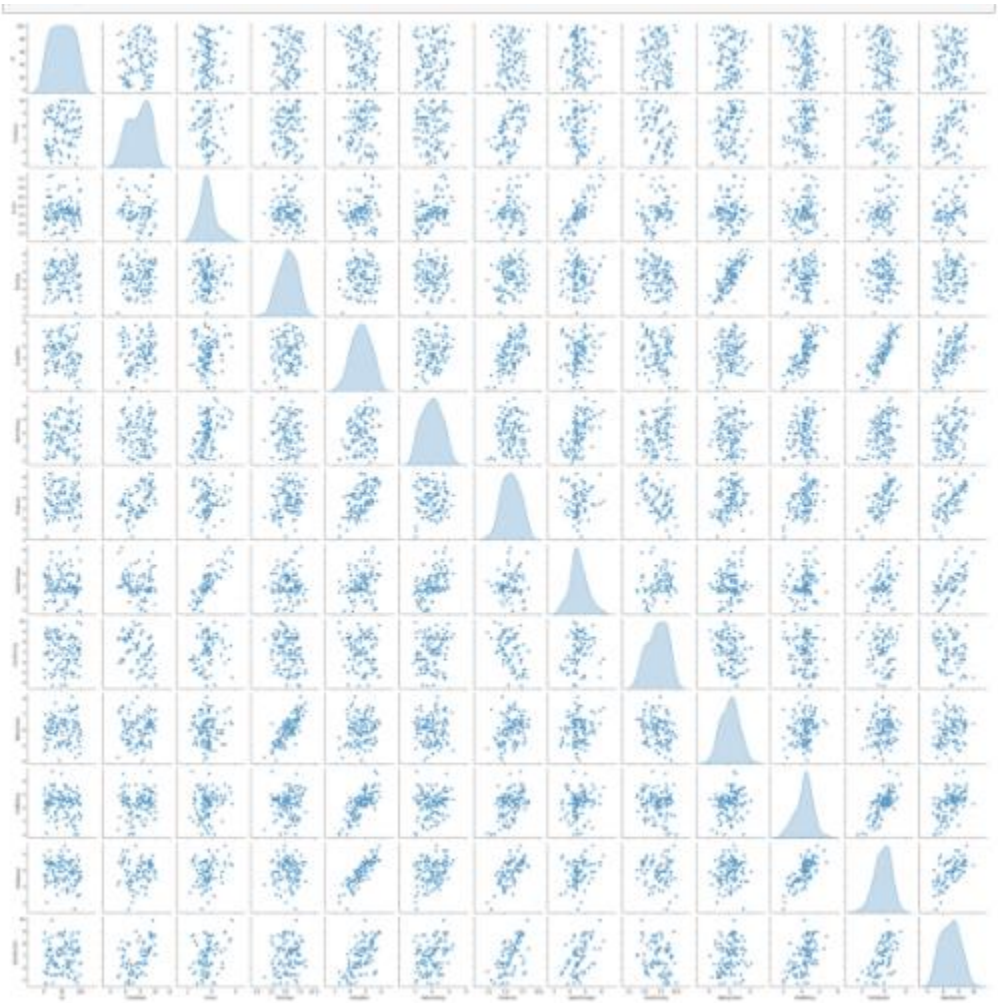


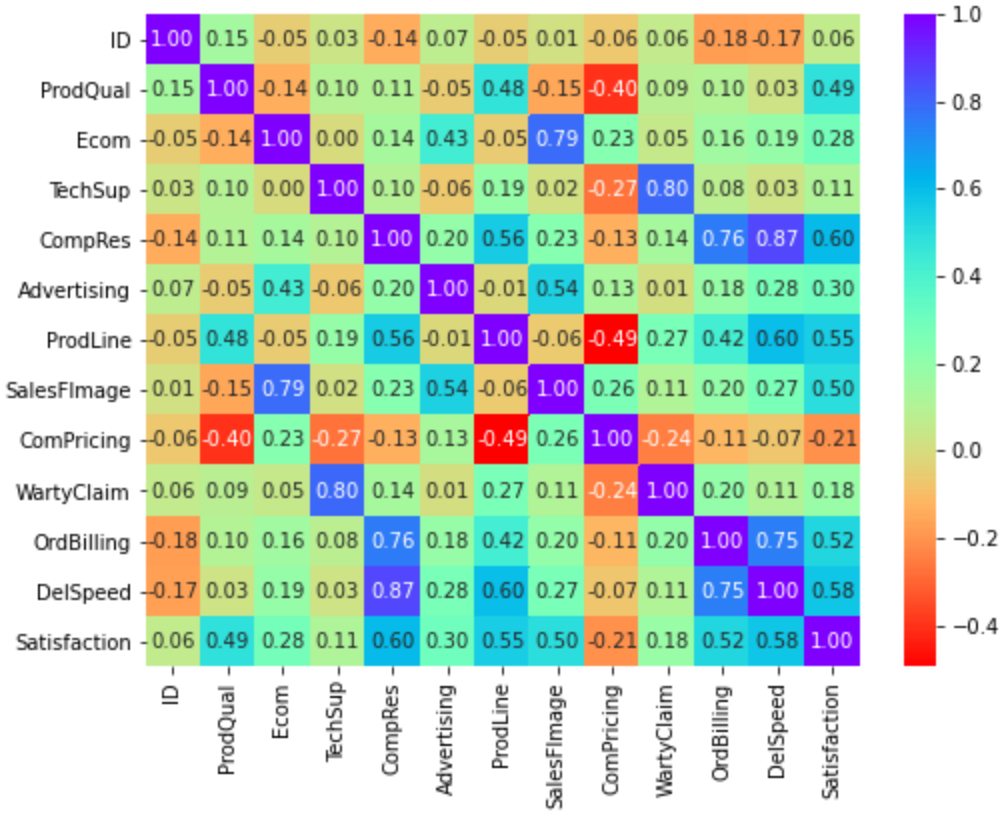
The above dist plot shows the distribution of data from 2 – 5 and is positively skewed. Boxplot shows that there are a few outliers.



The above dist plot shows the distribution of data from 2 – 9 and is positively skewed. Boxplot shows that there are no outliers.







Problem 2.2

Is scaling necessary for PCA in this case? Give justification and perform scaling.

Resolution:

PCA is impacted by scaling. If variables in the data set have large differences in their variances, then all variables need to be scaled. Otherwise the variables(s) with larges variance will have disproportionately more influence on the construction of PCs

Use StandardScaler to standardize the dataset's features onto unit scale (mean = 0 and variance = 1) which is a requirement for the optimal performance of many machine learning algorithms. The standard score of a sample x is calculated as:

$$Z = \frac{(X - X)}{S}$$

where u is the mean of the samples and s is the standard deviation of the sample. Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the dataset. Mean and standard deviation are then stored to be used on later data using transform.

We used zscore from scipy stats to standardise the data. Below is a transposed dataframe head

]:

	0	1	2	3	4
ID	-1.714816	-1.680173	-1.645531	-1.610888	-1.576245
ProdQual	0.496660	0.280721	1.000518	-1.014914	0.856559
Ecom	0.327114	-1.394538	-0.390241	-0.533712	-0.390241
TechSup	-1.881421	-0.174023	0.154322	1.073690	-0.108354
CompRes	0.380922	1.462141	0.131410	-1.448834	-0.700298
Advertising	0.704543	-0.544014	1.239639	0.615361	-1.614207
ProdLine	-0.691530	1.600835	1.218774	-0.844354	0.149004
SalesFImage	0.821973	-1.896068	0.634522	-0.583910	-0.583910
ComPricing	-0.113185	-1.088915	-1.609304	1.187789	-0.113185
WartyClaim	-1.646582	-0.665744	0.192489	1.173327	0.069885
OrdBilling	0.781230	-0.409009	1.214044	0.023805	0.240212
DelSpeed	-0.254531	1.387605	0.840226	-1.212443	-0.528220
Satisfaction	1.081067	-1.027098	1.671354	-1.786038	0.153474

As we can see from the above df, data after scaling will transform every value in such a waythat the mean will be 0 and standard deviation will be 1.

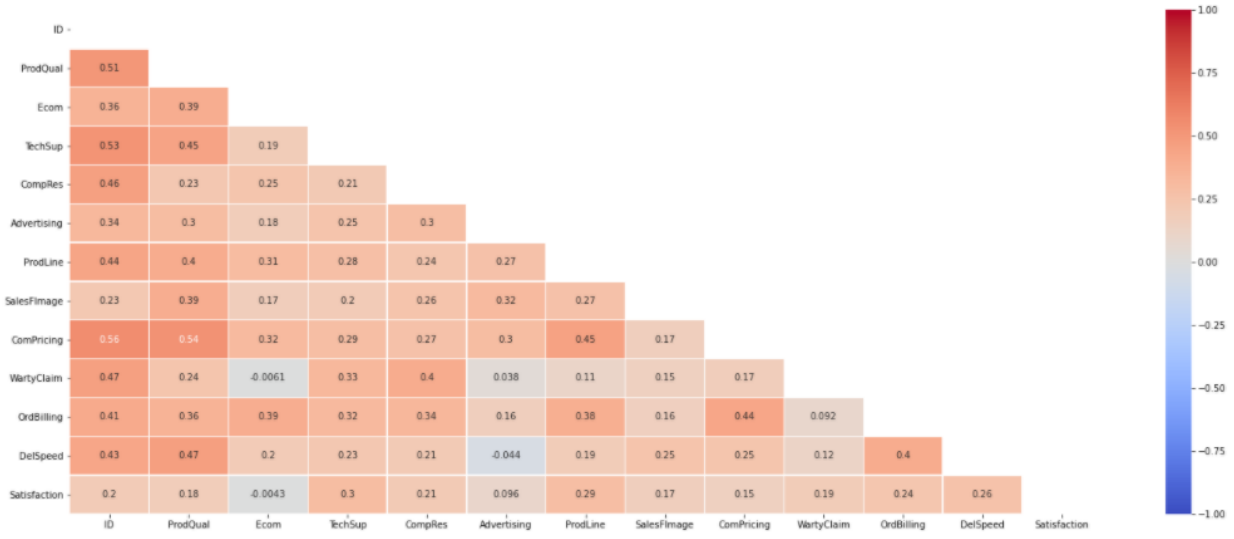
Problem 2.3

Comment on the comparison between the covariance and the correlation matrices from this data.

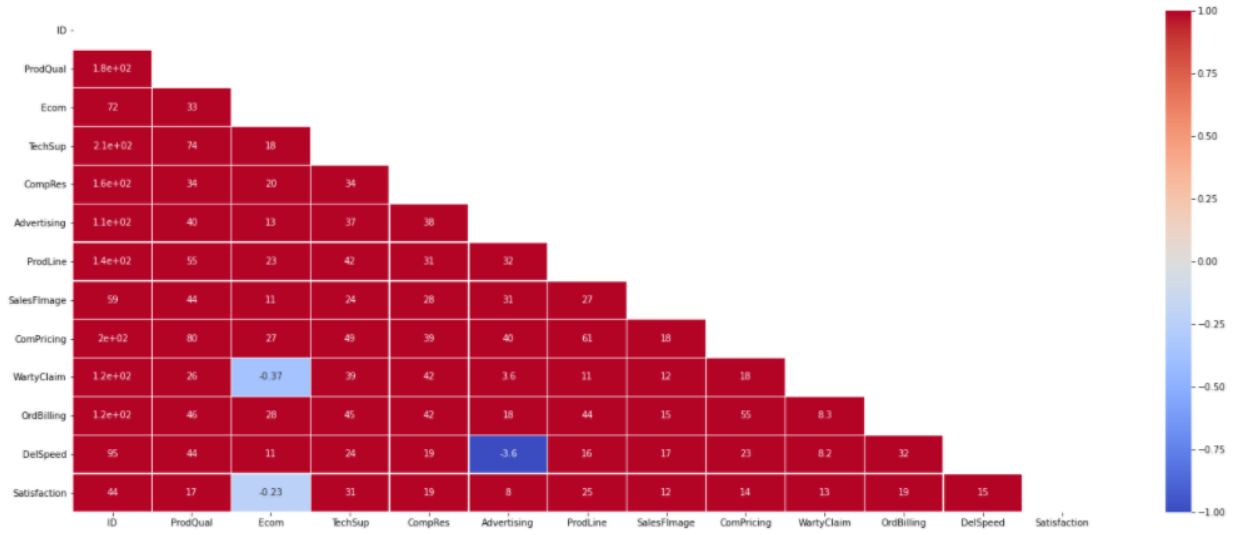
Resolution:

From the below images, it is clear is that the correlation and the covariance matrices are same afterz-score (mean =0 an¢I sd=1) sca1‹ng is performed.

Correlation Heat Map



Covariance Heat Map



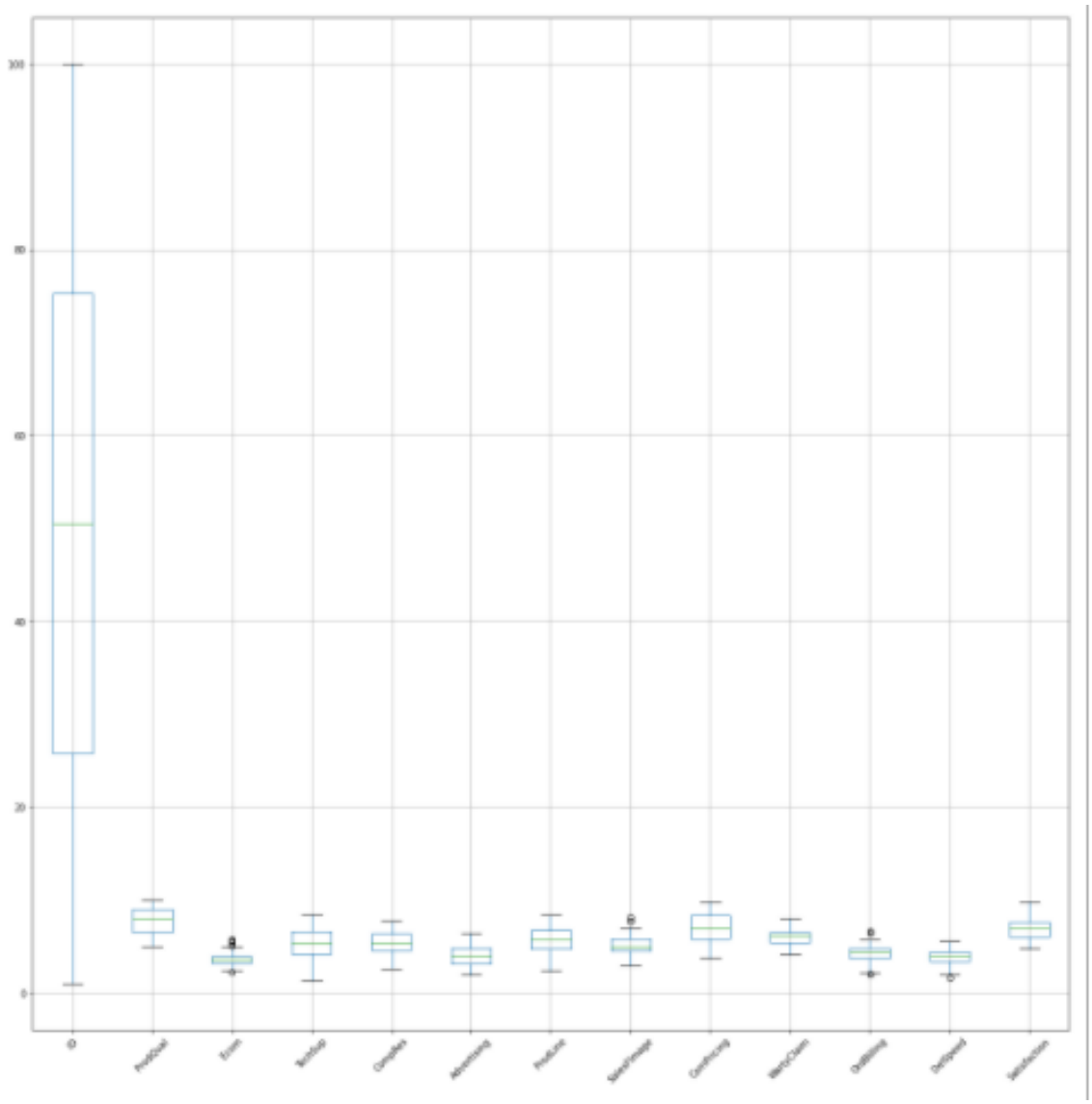
Problem 2.4

Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]

Resolution:

Scaling on data does not have any impact on outliers. The primary purpose of scaling is to make sure all the variables are on same scale, but that does not have any effect on the existing outliers.

Similar to – Univariate analysis boxplots, we can see that same outliers exist even after scaling the data.



From the above box plot we can infer that there are certain outliers before scaling.

After scaling, the above outliers are no longer outliers.

Problem 2.5

Build the covariance matrix, eigenvalues and eigenvectors.

Resolution:

Below is Covariance Matrix

Covariance Matrix
%s [[8.41666667e+02 5.90505051e+00 -9.38383838e-01 1.41363636e+00
-5.05959596e+00 2.39090909e+00 -1.85606061e+00 4.30808081e-01
-2.82424242e+00 1.39343434e+00 -4.80606061e+00 -3.66767677e+00
2.11414141e+00]
[5.90505051e+00 1.94959596e+00 -1.34161616e-01 2.04292929e-01
1.79474747e-01 -8.41414141e-02 8.76919192e-01 -2.27303030e-01
-8.65696970e-01 1.01080808e-01 1.35272727e-01 2.84242424e-02

```
   8.09313131e-01]
 [-9.38383838e-01 -1.34161616e-01  4.90723232e-01  9.29292929e-04
   1.18662626e-01  3.39373737e-01 -4.85454545e-02  5.94589899e-01
   2.48355556e-01  2.98020202e-02  1.01600000e-01  9.85939394e-02
   2.36064646e-01]
 [ 1.41363636e+00  2.04292929e-01  9.29292929e-04  2.34229798e+00
   1.78757576e-01 -1.08434343e-01  3.87752525e-01  2.78838384e-02
  -6.40313131e-01  1.00010606e+00  1.13868687e-01  2.85959596e-02
   2.05383838e-01]
 [-5.05959596e+00  1.79474747e-01  1.18662626e-01  1.78757576e-01
   1.46023838e+00  2.68161616e-01  8.92313131e-01  2.97711111e-01
  -2.38896970e-01  1.39084848e-01  8.49519192e-01  7.67765657e-01
   8.68832323e-01]
 [ 2.39090909e+00 -8.41414141e-02  3.39373737e-01 -1.08434343e-01
   2.68161616e-01  1.27000000e+00 -1.71212121e-02  6.55222222e-01
   2.33696970e-01  9.96969697e-03  1.92848485e-01  2.28323232e-01
   4.09212121e-01]
 [-1.85606061e+00  8.76919192e-01 -4.85454545e-02  3.87752525e-01
   8.92313131e-01 -1.71212121e-02  1.72997475e+00 -8.64797980e-02
  -1.00582828e+00  2.94429293e-01  5.18494949e-01  5.81383838e-01
   8.63040404e-01]
 [ 4.30808081e-01 -2.27303030e-01  5.94589899e-01  2.78838384e-02
   2.97711111e-01  6.55222222e-01 -8.64797980e-02  1.14986970e+00
   4.38381818e-01  9.44555556e-02  1.94349495e-01  2.13860606e-01
   6.39278788e-01]
 [-2.82424242e+00 -8.65696970e-01  2.48355556e-01 -6.40313131e-01
  -2.38896970e-01  2.33696970e-01 -1.00582828e+00  4.38381818e-01
   2.38719596e+00 -3.10284848e-01 -1.64416162e-01 -8.26909091e-02
  -3.83567677e-01]
 [ 1.39343434e+00  1.01080808e-01  2.98020202e-02  1.00010606e+00
   1.39084848e-01  9.96969697e-03  2.94429293e-01  9.44555556e-02
  -3.10284848e-01  6.71970707e-01  1.50046465e-01  6.58606061e-02
   1.73460606e-01]
 [-4.80606061e+00  1.35272727e-01  1.01600000e-01  1.13868687e-01
   8.49519192e-01  1.92848485e-01  5.18494949e-01  1.94349495e-01
  -1.64416162e-01  1.50046465e-01  8.62743434e-01  5.12315152e-01
   5.77571717e-01]
 [-3.66767677e+00  2.84242424e-02  9.85939394e-02  2.85959596e-02
   7.67765657e-01  2.28323232e-01  5.81383838e-01  2.13860606e-01
  -8.26909091e-02  6.58606061e-02  5.12315152e-01  5.39397980e-01
   5.05103030e-01]
 [ 2.11414141e+00  8.09313131e-01  2.36064646e-01  2.05383838e-01
   8.68832323e-01  4.09212121e-01  8.63040404e-01  6.39278788e-01
  -3.83567677e-01  1.73460606e-01  5.77571717e-01  5.05103030e-01
   1.42048081e+00]]
```

Below are Eigan Values and Eigen Vectors

Eigen Vectors
%s [[-9.99912384e-01 4.67506304e-04 -5.65878055e-03 -1.81980517e-04
-1.13459689e-02 1.32040445e-03 1.18508452e-03 1.65805267e-04
-2.49380566e-03 1.39599530e-03 -1.62487445e-03 4.60277483e-04
1.55546927e-04]
[-7.03193857e-03 -3.59759307e-01 1.85156279e-01 3.39959708e-01
5.96444194e-01 4.09015425e-01 -1.65646704e-01 -1.98916911e-01
-1.97345518e-01 2.26566299e-01 1.12985476e-01 1.38986625e-01
9.08941074e-02]
[1.11704017e-03 -9.65088475e-03 -2.30059336e-01 -1.17021022e-01
1.85639575e-01 -1.30762390e-01 2.64171262e-01 1.34885319e-01
-4.68722757e-01 4.16771358e-01 -1.58427432e-01 -6.06862090e-01
8.34433987e-02]
[-1.68710374e-03 -2.77977655e-01 2.75675872e-01 -7.93599159e-01
6.16672057e-02 1.77594428e-01 -8.78304959e-02 -5.44262681e-02
1.54563583e-01 1.86982642e-01 -3.25808942e-01 7.49744023e-02
-1.68727351e-02]
[6.02576484e-03 -3.47140278e-01 -3.16511010e-01 2.14560507e-02
-4.31722661e-01 1.00755869e-01 -4.93788671e-02 -2.48450750e-01
2.12321621e-01 5.03733662e-01 3.92118749e-01 -6.36147305e-02
-2.47958598e-01]
[-2.83918258e-03 -4.95618777e-02 -3.85693159e-01 -9.70866331e-02
2.96482960e-01 -4.16530351e-01 -7.51016700e-01 4.04898018e-02
6.76289320e-02 -3.61832287e-02 -2.03358293e-02 -5.14789298e-02
-5.77825258e-02]
[2.20686821e-03 -4.78070858e-01 2.25388663e-02 1.66799848e-01
-1.79001426e-01 8.52498399e-02 -6.91400344e-02 7.82955236e-01
-1.04656308e-01 -6.13440371e-02 -1.33787758e-01 8.16083764e-02
-2.16170714e-01]
[-5.07993493e-04 -4.43466354e-02 -4.37929648e-01 -2.10955144e-01
3.16497386e-01 -1.99141283e-01 4.46357167e-01 9.80048429e-02
-7.18508478e-02 6.35717460e-02 1.43579372e-01 6.14008906e-01
-7.90902547e-02]
[3.36785909e-03 4.19232153e-01 -4.52741681e-01 -1.31054055e-01
1.77428558e-02 7.32042946e-01 -1.59560945e-01 1.87893695e-01
6.72800925e-04 -4.24334925e-02 -3.31731814e-02 -1.68075294e-02
-4.39708759e-02]
[-1.65798442e-03 -1.58570410e-01 8.70724717e-02 -3.69807918e-01
1.19861875e-02 6.28513036e-02 -3.38391143e-02 7.41157429e-02
-2.94309088e-01 -4.11652393e-01 7.16636000e-01 -1.94547888e-01
1.10857618e-01]
[5.71996513e-03 -2.32462117e-01 -2.16317074e-01 2.92690374e-03

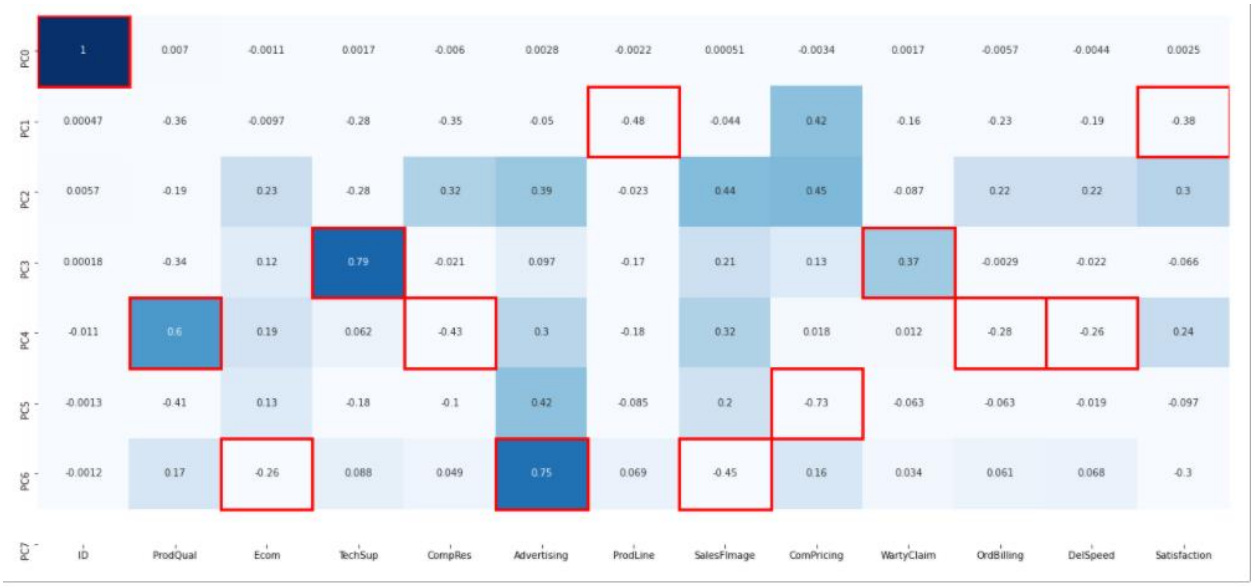
-2.75858049e-01 6.29372543e-02 -6.07876406e-02 -4.42718776e-01
-5.70681661e-01 -3.41862380e-01 -3.45814061e-01 1.31098183e-01
-2.00129073e-01]
[4.36676088e-03 -1.93670591e-01 -2.15326291e-01 2.20523161e-02
-2.55001201e-01 1.93089082e-02 -6.81997910e-02 3.99650215e-02
2.90177449e-03 4.55121101e-02 -8.10200709e-02 1.51589714e-01
9.01304938e-01]
[-2.51090346e-03 -3.83770866e-01 -2.99694350e-01 6.55396706e-02
2.40375199e-01 9.65447911e-02 2.99380298e-01 -1.22026289e-01
4.89268537e-01 -4.32370294e-01 -1.49922752e-01 -3.71642870e-01
7.63729154e-03]]
Eigen Values
%s [8.41813864e+02 5.41445001e+00 3.41610042e+00 2.50518082e+00
1.48883147e+00 1.20292011e+00 6.65562632e-01 5.65091638e-01
3.01560595e-01 2.54668356e-01 1.61475362e-01 8.26320424e-02
6.88176085e-02]

Problem 2.6

Write down the explicit form of the first PC (in terms of the eigenvectors)

Resolution:

The equation for first PC would be cross multiplication of variables and Eigen vectors of 0 indexes. Eigen vectors derived from sklearn, Numpy and Statsmodels might differ a bit numerically or on sign prespective.



	PC1
ID	0.05
ProdQual	-0.16
Ecom	-0.17
TechSup	-0.12
CompRes	-0.42
Advertising	-0.18
ProdLine	-0.35
SalesFImage	-0.22
ComPricing	0.13
WartyClaim	-0.17
OrdBilling	-0.39
DelSpeed	-0.42
Satisfaction	-0.41

PC1 = (0.05 x ID)+(-0.16 x ProdQual)+(-0.17 x Ecom)+(-0.12 x TechSup)+(-0.42 x CompRes)+(-0.18 x Advertising)+(-0.35 x ProdLine)+(-0.22 x SalesFImage)+(0.13 x ComPricing)+(-0.17 x WartyClaim)+(-0.39 x OrdBilling)+(-0.42 x DelSpeed)+(-0.41 x Satisfaction)

Problem 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

Resolution:

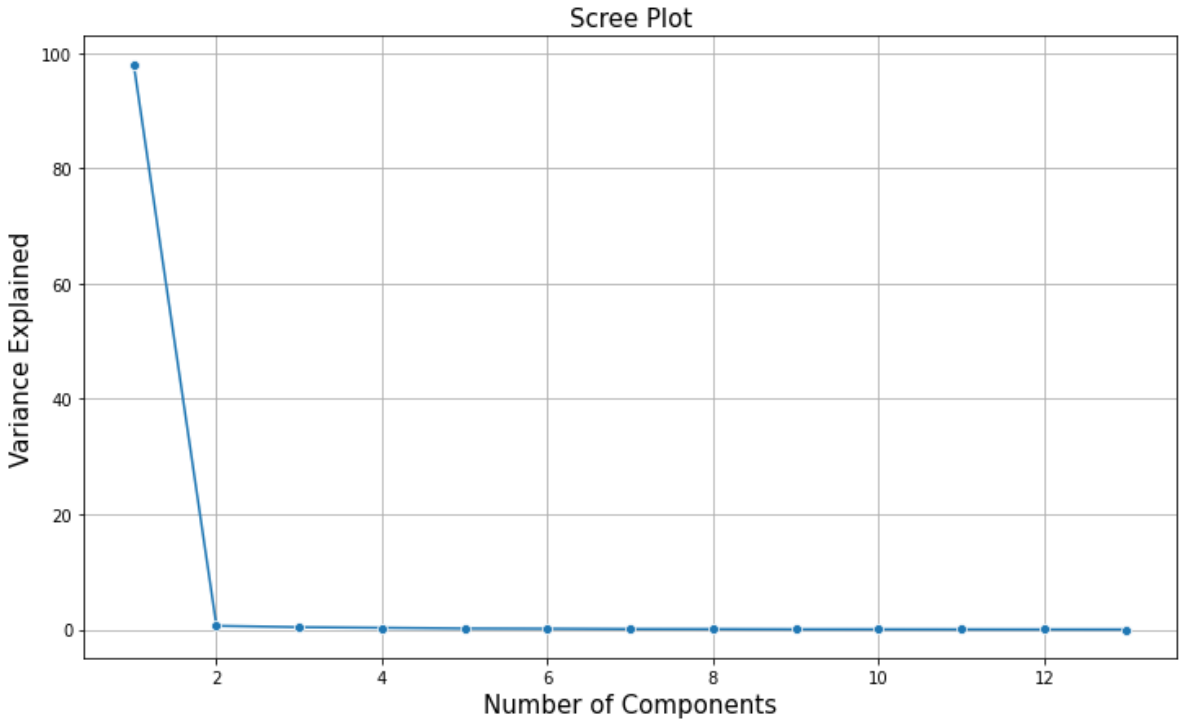
Explained variance by each principal component:

[98.12 98.75 99.15 99.44 99.62 99.76 99.83 99.9 99.93 99.96 99.98 99.99 100.]

The first PC explain 98.12% of variance in the data set followed by 98.75%of variance explained by PC2 and so on.

Cumulative Explained Variance:

[98.12 98.75 99.15 99.44 99.62 99.76 99.83 99.9 99.93 99.96 99.98 99.99 100]



The cumulative percentage gives the percentage of variance accounted for the number of components. For instance the cumulative percentage of the second component is sum of the percentage of variance for the first and the second component. It helps in deciding the number of components by selecting the components which explains the higher variance.

Problem 2.8

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

Resolution:

Interpretation of the principal component is based on which variables are the most strongest correlated with each other.

Example Component Summary:

PC1 = (0.05 x ID)+(-0.16 x ProdQual)+(-0.17 x Ecom)+(-0.12 x TechSup)+(-0.42 x CompRes)+(-0.18 x Advertising)+(-0.35 x ProdLine)+(-0.22 x SalesFImage)+(0.13 x ComPricing)+(-0.17 x WartyClaim)+(-0.39 x OrdBilling)+(-0.42 x DelSpeed)+(-0.41 x Satisfaction)

Here all the values are scaled (normalized) variables are used to construct PCs.

Similarly the other PCs can also be expressed in terms of scaled variables.

For business implications, the following can be a way to explain the Principal component analysis:

- The first principal component picks up about 98% of the variability in the data. That implies picking up considerable amount of variation in the data. So, in business we need to look at the ease of doing things wells.
- The explanation of each component along with their weights is also one of the ways to look at it.

- Here we are considering only 5 Principal Components for this sample interpretation.

	PC1	PC2	PC3	PC4	PC5
ID	0.05	0.05	-0.23	-0.50	-0.78
ProdQual	-0.16	0.32	0.00	-0.52	0.30
Ecom	-0.17	-0.44	-0.25	-0.09	0.30
TechSup	-0.12	0.24	-0.57	0.29	-0.01
CompRes	-0.42	-0.00	0.21	0.17	-0.21
Advertising	-0.18	-0.35	-0.13	-0.21	-0.11
ProdLine	-0.35	0.30	0.10	-0.09	0.10
SalesFImage	-0.22	-0.46	-0.27	-0.13	0.15
ComPricing	0.13	-0.42	0.07	0.17	-0.19
WartyClaim	-0.17	0.21	-0.57	0.28	-0.07
OrdBilling	-0.39	-0.01	0.18	0.22	-0.17
DelSpeed	-0.42	-0.06	0.24	0.18	-0.20
Satisfaction	-0.41	-0.02	-0.02	-0.31	0.10

Decision regarding which correlation value is high may vary from one case to the other. In the above example we have taken a considerable magnitude irrespective of the sign.

Principal component analysis is a very versatile technique and its applications in the number of situations.

The End

Thakur Arun Singh

*************************^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^*************************