PROJECT ON ADVANCED STATISTIC S

BUSINESS REPORT

Sankesh A Nagrare
Sankeshnagrare92@gmail.com
PGP – Data Science & Business Analytics

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Case Study 1 - Service Data Analysis Using ANOVA

Overview:

The staff of a service center 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.

Summary:

This business report provides detailed explanation on the approach to each problem definition, solution to those the problems provides some key insights/recommendations to the business.

Q1.1)State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables 'Manufacturer' and 'Technician individually

1)The Hypothesis of One-Way ANOVA for 'Manufacturer' with respect to Service Time

H_0: Service time is not depend on Manufacturer
H_A: Service time is depend on Manufacturer

2)The Hypothesis of One-Way ANOVA for Technician with respect ServiceTime

H_0: Service time is not dependent on Technician.

H_A: Service time is dependent on Technician.

Where,
H_0 = Null Hypothesis
H_A = Alternate Hypothesis

Also, it is given that the dataset qualifies all the assumptions for ANOVA.

- Each group sample is drawn from a normally distributed population
- All populations have a common variance
- All samples are drawn independently of each other
- Within each sample, the observations are sampled randomly and independently of each other
- Factor effects are additive

Q1.2) 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.

we perform One-Way ANOVA

The Hypothesis of One-Way ANOVA for 'Manufacturer' with respect to 'Service Time'
H_0: Service time is not depend on Manufacturer.
H_A: Service time is depend on Manufacturer .

Below is the result from python code:

	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

*Table 1: ANOVA results for variable '*Manufacturer' *with respect to variable* 'Service Time'

From above,
we can say that the corresponding p-value is great than alpha (0.05). Thus, we Fail to reject the Null Hypothesis and accept the alternate hypothesis. And state that Servicve Time is not depend on the manufacturer.

.

Q1.3) 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.

we perform One-Way ANOVA

1) The Hypothesis of One-Way ANOVA for 'Technician' with respect 'ServiceTime'

H_0: Service time is not dependent on Technician.

H_A: Service time is dependent on Technician.

Below is the result from python code:

	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

Table 2: ANOVA results for variable Technician' *with respect to variable* 'ServiceTime'

From above,
we can say that the corresponding p-value is great than alpha (0.05). Thus, we Fail to rejectthe Null Hypothesis and accept the alternate hypothesis. And state that Servicve Time is not depend on the Technician.

Q1.4) 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]

1)Point Plot of Manufacturer' vs 'Service Time'

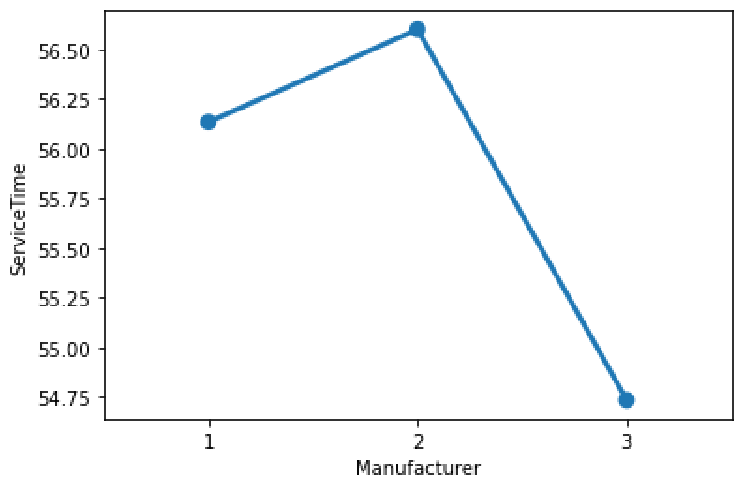

Figure 1: Plot of Manufacturer vs Service Time

From the above graph, we can say that:

- As we can say that as compare to Manufacturer1 taking service time is less and Manufacturers 2 is taking more time for service least service time taken by the manufacturer 3 as compare to both manufacturer.
- There is a moderate difference in service time of individuals.

2)Point Plot of Technician vs 'Service Time'

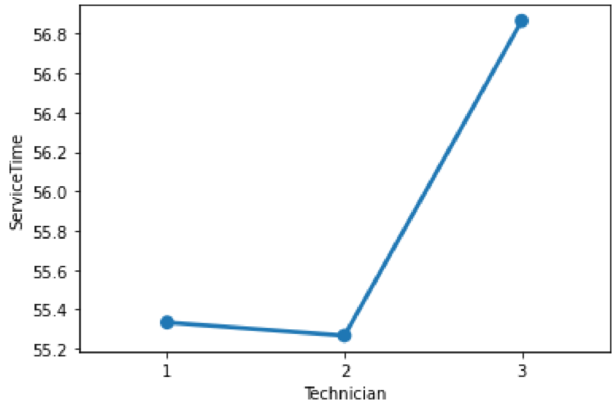

Figure 2: Plot of Technician vs Service Time

From the above graph, we can say that:

- As we can say that as compare to Technician 3 is taking more time to service. And the least time taken by the Technician2 we can analyses from the above figure2

The interaction between two treatments (*Manufacturer* & Technician in this case) with respect to continuous measure (*Service Time* variable in this case) is said to exist, if, the response of continuous measure to one categorical variable not depends on another categorical variable.

- Interaction effects represent the combined effects of factors on the independent measure.
- When an interaction effect is present, the impact of one factor depends on the level of the other factor.

Interaction plot shows level of interaction by the number of intersection points.

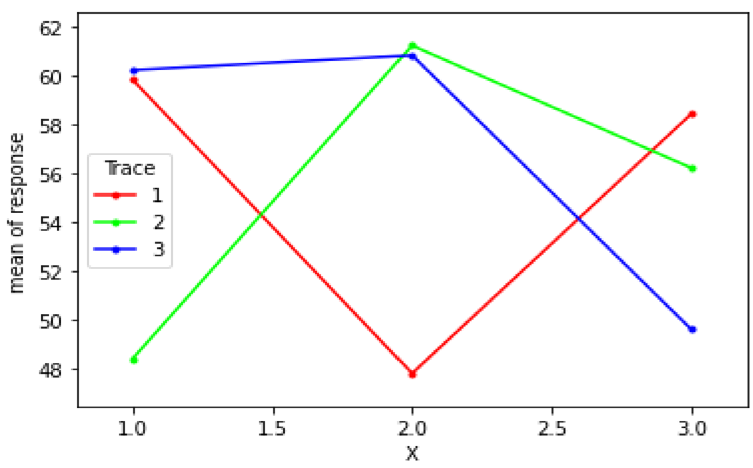

Figure 3: Interaction Plot between Manufacturer & Technician *variables*

- More the number of intersection points in the graph, higher the interaction level between concerned variables and vice-versa.

From the graph above, we can say that:

- There are More than five intersection points in the graph which shows there is a good level of interaction between *Manufacturer* & Technician variable.

Q1.5) Perform a two-way ANOVA based on the variables 'Manufacturer' & 'Technician' with respect to the variable 'Service Time' and state your results.

A two-way ANOVA with interaction tests three null hypotheses at the same time:

Null Hypothesis (H_0):

- Service time is not dependent on Manufacturer.
- Service time is not dependent on Technician.
- There is no interaction effect between Manufacturer and Technician on Service

Time salary.Alternate Hypothesis (H_A):

- Service time is dependent on Manufacturer.
- Service time is dependent on Technician.
- There is interaction effect between Manufacturer and Technician service time.

Below is the result from python code:

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

Table 4: ANOVA results of variables 'Manufacturer'*&* 'Technician' *with respect to variable* 'Service Time'*along with their interaction*

Since, One-Way ANOVA has already been performed individually for Manufacturer & Technician variable above, we are concerned only with the results of third hypothesis here (interaction test between Manufacturer & Technician with respect to Service time)

And, since, the p-value of the interaction effect term of Manufacturer and Technician is less than 0.05 the Null Hypothesis is rejected in this case and we can accept the alternate hypothesis.

Therefore,

- Service time is dependent on Manufacturer.
- Service time is dependent on Technician.
- There is no interaction effect between Manufacturer and Technician service time.

Q1.6) Explain the business implications of performing ANOVA for this particular case study.

By using ANOVA for the dataset, we came to know that –

- As we can see that service time is not depend on the Manufacturers.
- Service time is not depend on the Technician.
- There is interaction effect between Manufacturer and Technician service time

Case Study 2- Hair Salon (EDA & PCA)

Overview

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.

Q2.1) Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented.

- Data has 13 columns and 100 rows
- The entire dataset is of float data type. However, column 'ID' is Integer datatype.
- No duplicate records.
- No null values.
- We have to investigate further for outliers.

Univariate Analysis

1) Boxplot for outlier identification

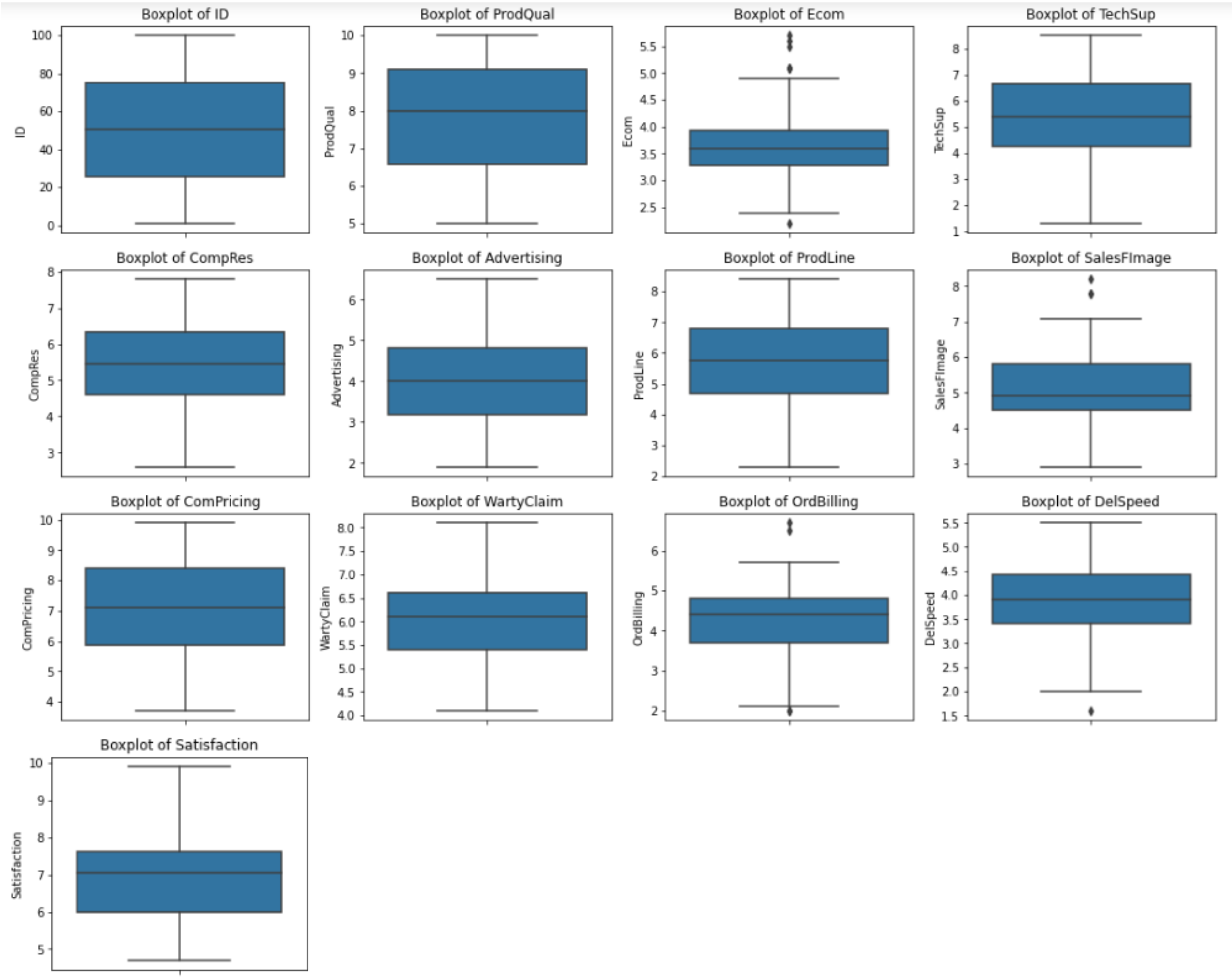

Figure 4: Boxplot for Outlier Identification

From the graph above, we can clearly say that, there are a lot of outliers in the dataset. Hence, we are going to replace them with either the maximum or the minimum value based on which side of the boxplot they lie.

2) Distplot for studying variable distribution

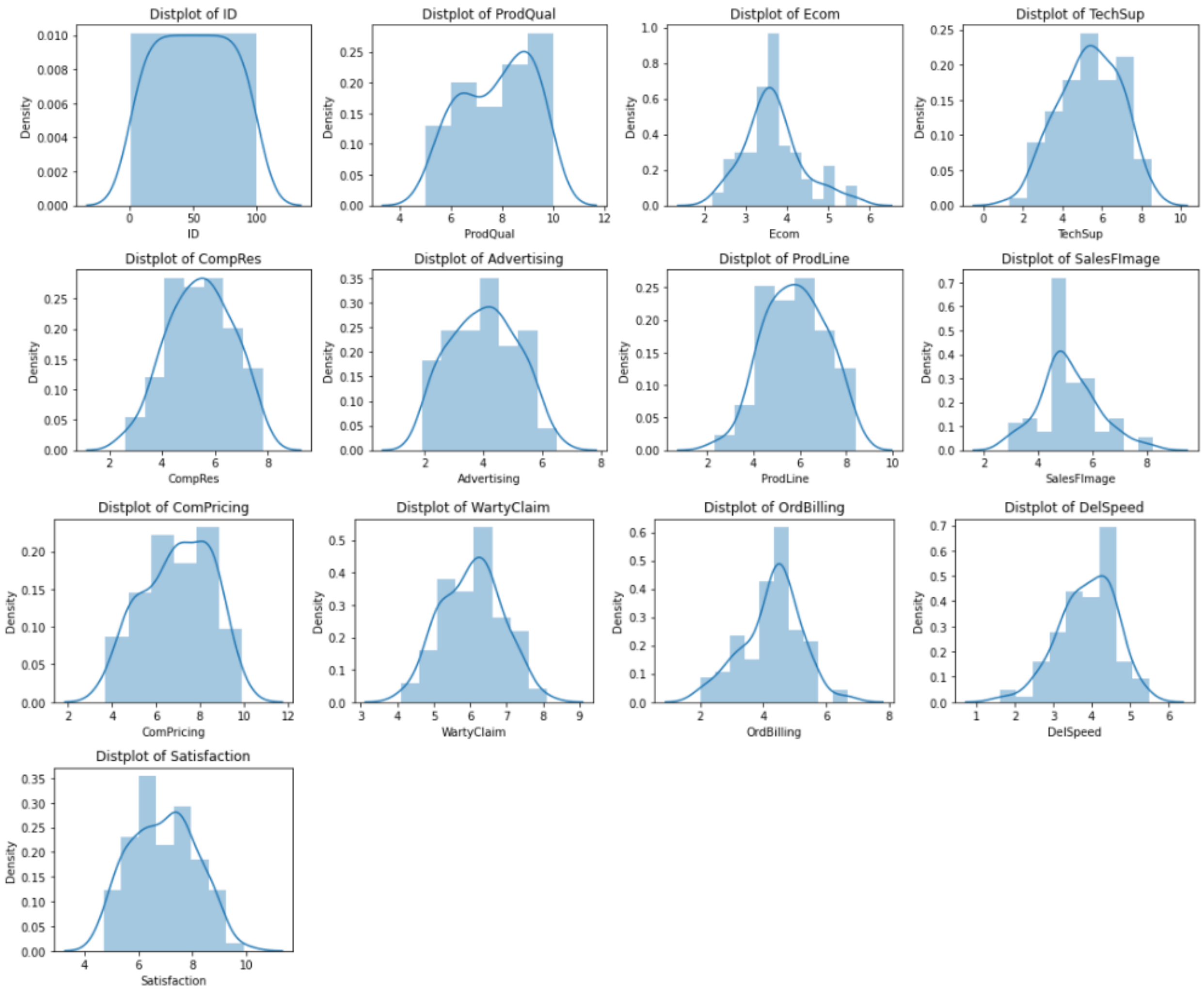

Figure 5: Distplot for studying variable distribution.

Observations:

- The variables Prodqual have a slightly left-skewed distribution.
- As we can see that all distribution have a normal distribution.

Multivariate Analysis

1) Heatmap to study correlation between variables

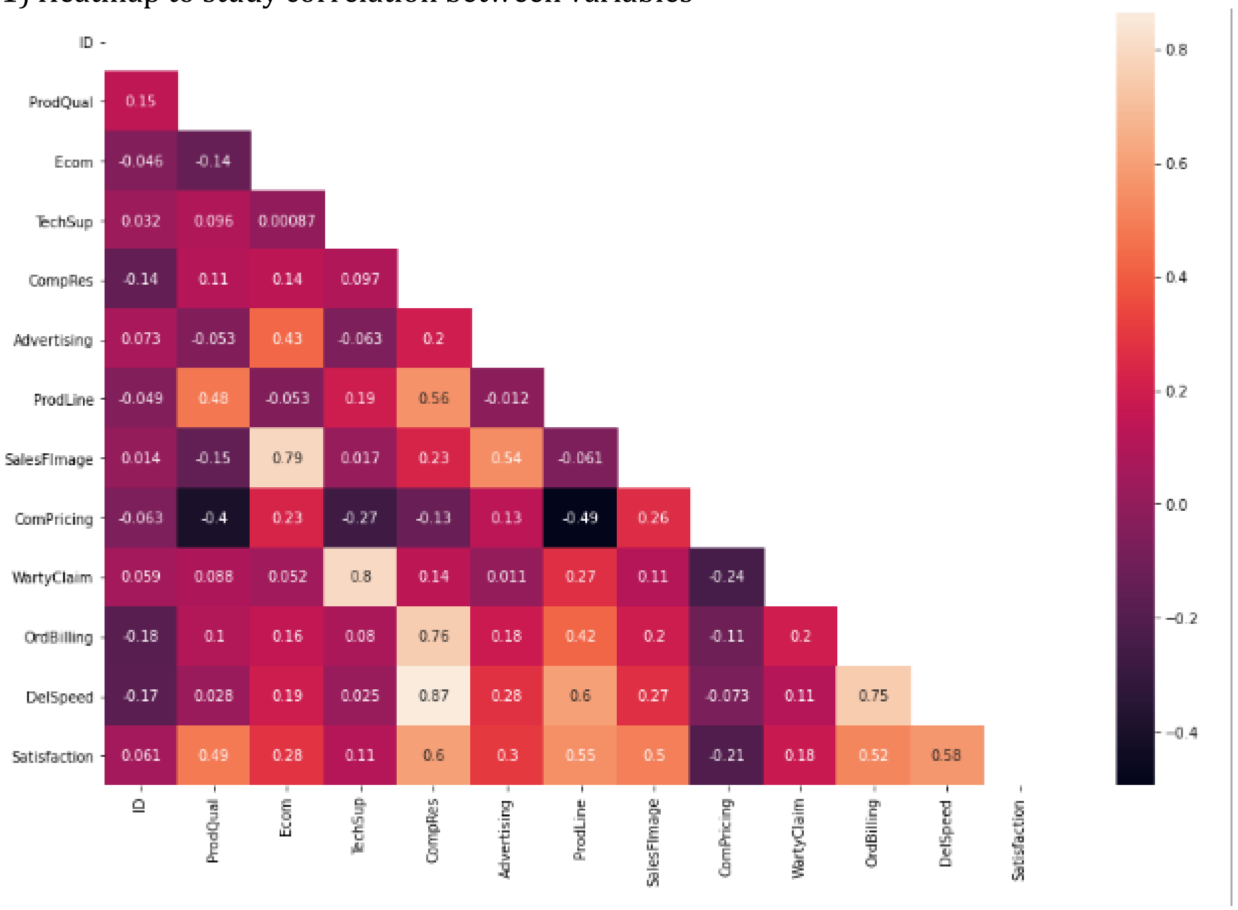

Figure 6: Correlation Heatmap

Observations:

Variable ProdQual with Techsup, Techsup with Compres, Compres with Delspeed have a strong correlation with one another.

Q2.2) Scale the variables and write the inference for using the type of scaling function for this case study.

Yes, it is necessary to normalize data before performing PCA in this case because there are somevariables which are in integer.

- The PCA calculates a new projection of the dataset and the new axis is based on the standard deviation of our variables. So, a variable with a high standard deviation will have a higher weight for the calculation of axis than a variable with a low standard deviation. If we normalize your data, all variables will have the same standard deviation, thus all variables will have the same weight and our PCA calculates relevant axis.

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

Table 5: Sample of data after scaling

Q2.3) Comment on the comparison between covariance and the correlation matrix after scaling.

Correlation Matrix:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369491	0.(
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758	0.337583	0.
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469	0.308274	0.:
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828	0.491135	-0.:
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862	0.524749	-0.:

Table 6: Correlation Matrix of the scaled dataset

Covariance Matrix

	ID	ProdQual	Ecom	TechSup	CompRes	Advertising	ProdLine	SalesFImage	ComPricing	WartyClaim	OrdBilling	DelSpeed	Satisfaction
ID	1.000000	0.145774	-0.046173	0.031838	-0.144322	0.073129	-0.048641	0.013848	-0.063007	0.058592	-0.178352	-0.172134	0.061143
ProdQual	0.145774	1.000000	-0.137163	0.095600	0.106370	-0.053473	0.477493	-0.151813	-0.401282	0.088312	0.104303	0.027718	0.486325
Ecom	-0.046173	-0.137163	1.000000	0.000867	0.140179	0.429891	-0.052688	0.791544	0.229462	0.051898	0.156147	0.191636	0.282745
TechSup	0.031838	0.095600	0.000867	1.000000	0.096657	-0.062870	0.192625	0.016991	-0.270787	0.797168	0.080102	0.025441	0.112597
CompRes	-0.144322	0.106370	0.140179	0.096657	1.000000	0.196917	0.561417	0.229752	-0.127954	0.140408	0.756869	0.865092	0.603263

Table 7: Covariance Matrix of the scaled dataset

From the above two tables we can say that there is a slight or negligible difference in their values. This is because the dataset considered to calculate covariance & correlation is already scaled. If the dataset wouldn't have been scaled the covariance matrix would have differed a lot.

Both Correlation and Covariance are very closely related to each other and yet they differ a lot. When it comes to choosing between Covariance vs Correlation, the latter stands to be the first choice as it remains unaffected by the change in dimensions, location, and scale, and can also be used to make a comparison between two pairs of variables.

Q2.4) Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.?

Outliers in the dataset before scaling:

Figure 4 shows the outliers before scaling

Outliers in the dataset after scaling:

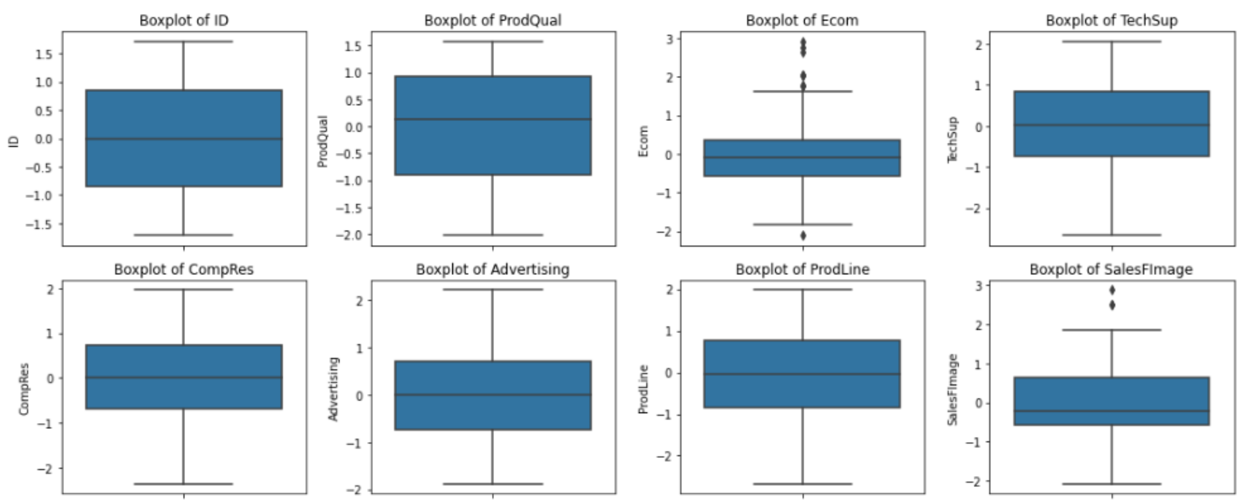

Figure 6: Outliers after scaling the dataset

From the above two graphs, we can say that:

- There is no difference in outliers of the dataset before & after scaling. Scaling of data just transforms all variables in the dataset to a same range and it has no effect whatsoever on the presence of outliers in the dataset or treats them in any way.

Q2.5) Extract the eigenvalues and eigenvectors.
(Using Sklearn PCA print both)

Eigen vectors:
Below is the output of python code:

```
array([[ 0.04659362, -0.15527921, -0.16649956, -0.12357735, -0.42372779,
        -0.17945503, -0.35204872, -0.21735578,  0.13280731, -0.17304207,
        -0.38957753, -0.42388058, -0.41081139],
       [-0.04983935, -0.31871546,  0.43719256, -0.24176757,  0.00333435,
         0.35166727, -0.2983661 ,  0.45952847,  0.4200957 , -0.20560727,
         0.01190881,  0.05703423,  0.01812171],
       [-0.23101384,  0.00350694, -0.24857742, -0.56970117,  0.21422033,
        -0.13320679,  0.10241352, -0.26689457,  0.06735008, -0.56554168,
         0.18313799,  0.23758551, -0.02355175],
       [ 0.49737305,  0.52458938,  0.08627072, -0.29368824, -0.17225577,
         0.20753035,  0.09192432,  0.12748761, -0.16813296, -0.28042302,
        -0.22041485, -0.18178685,  0.31265255],
       [-0.77967247,  0.30438026,  0.30202119, -0.01164363, -0.20503738,
        -0.11194799,  0.09999132,  0.15296919, -0.19120973, -0.0709626 ,
        -0.17007695, -0.19698481,  0.10399316],
       [-0.11581052, -0.26205579, -0.09075031, -0.05256564, -0.05876359,
         0.69277812,  0.06253771, -0.10886837, -0.58106764, -0.04736066,
        -0.07248005,  0.04588085, -0.24600185],
       [-0.2434289 ,  0.40025604, -0.4135531 ,  0.12809222, -0.03792792,
         0.51300709, -0.16638963, -0.17542557,  0.4860585 ,  0.12431805,
         0.0718449 , -0.05915848,  0.09828528],
       [ 0.00670863,  0.12242084,  0.0069788 , -0.01430496, -0.00299504,
        -0.07491646, -0.63459988,  0.02358494, -0.34443129, -0.04038476,
         0.62756051, -0.23611464,  0.0772913 ],
       [-0.10531516, -0.29399737, -0.51442042,  0.10903036,  0.13815804,
        -0.09001346, -0.22501788,  0.32985481, -0.15530003, -0.12226945,
        -0.3456201 , -0.01042558,  0.52981233],
       [-0.03330521, -0.18687338, -0.23131479, -0.53917654, -0.44253679,
        -0.03673186,  0.22952031,  0.17964334,  0.02644731,  0.49812059,
         0.24437143, -0.07837693,  0.15141982],

       [-0.03642261,  0.2077462 ,  0.02350541, -0.42537132,  0.58002662,
        -0.02395951, -0.25381921,  0.05235779, -0.08811167,  0.45618859,
        -0.32499983, -0.06159741, -0.21109323],
       [ 0.00261414,  0.22775251, -0.02731935, -0.01723457, -0.37927468,
        -0.0970485 , -0.34729945,  0.0716641 , -0.10633714,  0.08229667,
        -0.15708004,  0.78375834, -0.10500781],
       [ 0.02144534,  0.21495202, -0.34920787,  0.11007793,  0.05228422,
        -0.05105266,  0.18824266,  0.6622454 , -0.0100872 , -0.16428112,
         0.15271613, -0.05566047, -0.53518017]])
```

Table 8: Eigen vectors

Eigen Values:

Below is the output of python code:

```
array([4.09031163, 2.58246009, 1.74408311, 1.38449513, 0.84530284,
       0.63737703, 0.55277881, 0.40687424, 0.32136047, 0.23782883,
       0.14469341, 0.10013442, 0.08361313])
```

Table 8: Eigen values

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
ID	0.046594	-0.049839	-0.231014	0.497373	-0.779672	-0.115811	-0.243429	0.006709	-0.105315	-0.033305	-0.036423	0.002614	0.021445
ProdQual	-0.155279	-0.318715	0.003507	0.524589	0.304380	-0.262056	0.400256	0.122421	-0.293997	-0.186873	0.207746	0.227753	0.214952
Ecom	-0.166500	0.437193	-0.248577	0.086271	0.302021	-0.090750	-0.413553	0.006979	-0.514420	-0.231315	0.023505	-0.027319	-0.349208
TechSup	-0.123577	-0.241768	-0.569701	-0.293688	-0.011644	-0.052566	0.128092	-0.014305	0.109030	-0.539177	-0.425371	-0.017235	0.110078
CompRes	-0.423728	0.003334	0.214220	-0.172256	-0.205037	-0.058764	-0.037928	-0.002995	0.138158	-0.442537	0.580027	-0.379275	0.052284
Advertising	-0.179455	0.351667	-0.133207	0.207530	-0.111948	0.692778	0.513007	-0.074916	-0.090013	-0.036732	-0.023960	-0.097049	-0.051053
ProdLine	-0.352049	-0.298366	0.102414	0.091924	0.099991	0.062538	-0.166390	-0.634600	-0.225018	0.229520	-0.253819	-0.347299	0.188243
SalesFImage	-0.217356	0.459528	-0.266895	0.127488	0.152969	-0.108868	-0.175426	0.023585	0.329855	0.179643	0.052358	0.071664	0.662245
ComPricing	0.132807	0.420096	0.067350	-0.168133	-0.191210	-0.581068	0.486058	-0.344431	-0.155300	0.026447	-0.088112	-0.106337	-0.010087
WartyClaim	-0.173042	-0.205607	-0.565542	-0.280423	-0.070963	-0.047361	0.124318	-0.040385	-0.122269	0.498121	0.456189	0.082297	-0.164281
OrdBilling	-0.389578	0.011909	0.183138	-0.220415	-0.170077	-0.072480	0.071845	0.627561	-0.345620	0.244371	-0.325000	-0.157080	0.152716
DelSpeed	-0.423881	0.057034	0.237586	-0.181787	-0.196985	0.045881	-0.059158	-0.236115	-0.010426	-0.078377	-0.061597	0.783758	-0.055660
Satisfaction	-0.410811	0.018122	-0.023552	0.312653	0.103993	-0.246002	0.098285	0.077291	0.529812	0.151420	-0.211093	-0.105008	-0.535180

Q2.6) Write the explicit form of the first PC (in terms of Eigen Vectors)?

The explicit form of the first PC ($a_1x_1 + a_2x_2 + \ldots + a_nx_n$)

Below is the output from python code

```
(0.05 * -1.71)+ (-0.16 * 0.5)+ (-0.17 * 0.33)+ (-0.12 * -1.88)+ (-0.42 * 0.38)+ (-0.18 * 0.7)+ (-0.35 * -0.69)+ (-0.22 * 0.82)+
(0.13 * -0.11)+ (-0.17 * -1.65)+ (-0.39 * 0.78)+ (-0.42 * -0.25)+ (-0.41 * 1.08)+
```

Table 9: Sample of Dataframe of Explicit of the first PC

Q2.7) 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?

Below is the output of python code:

```
array([0.31149296, 0.50815723, 0.64097587, 0.7464105 , 0.81078356,
       0.85932227, 0.9014185 , 0.93240354, 0.95687638, 0.97498796,
       0.98600692, 0.99363254, 1.        ])
```

Table 9: Sample of Dataframe of Explicit of the first PC

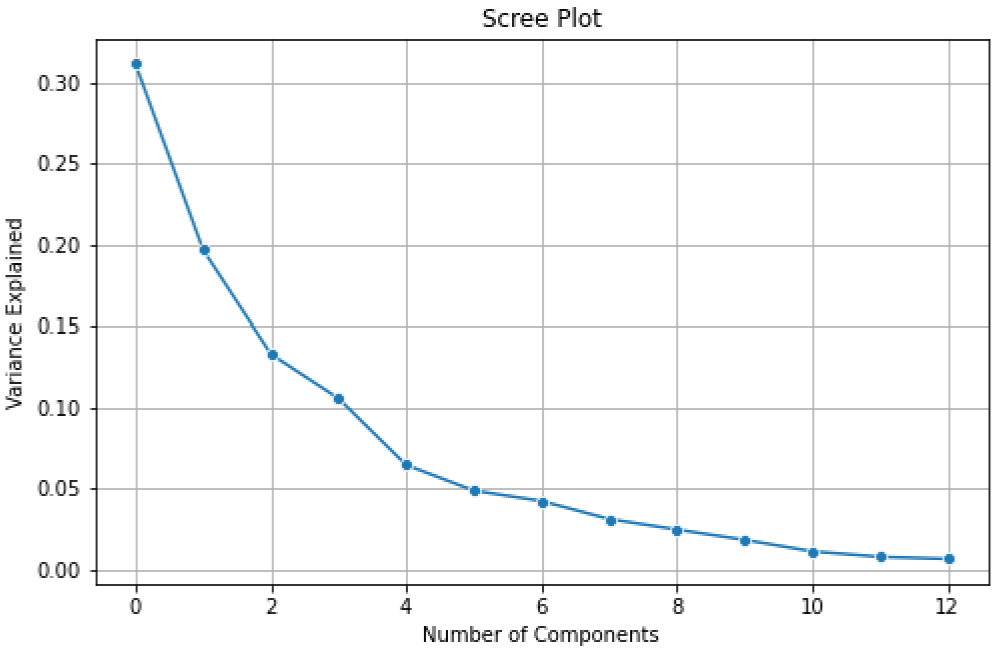

Figure 8: Scree Plot

Cumulative values of eigen values help us decide the optimum number of principal components by considering the cumulative explained variance ratio with a certain confidence interval.

In this case, we take confidence level as 85%, hence, we use 7 principal components.

The Eigenvectors indicate the direction of the principal components (new axes)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
ID	0.046594	-0.049839	-0.231014	0.497373	-0.779672	-0.115811	-0.243429
ProdQual	-0.155279	-0.318715	0.003507	0.524589	0.304380	-0.262056	0.400256
Ecom	-0.166500	0.437193	-0.248577	0.086271	0.302021	-0.090750	-0.413553
TechSup	-0.123577	-0.241768	-0.569701	-0.293688	-0.011644	-0.052566	0.128092
CompRes	-0.423728	0.003334	0.214220	-0.172256	-0.205037	-0.058764	-0.037928
Advertising	-0.179455	0.351667	-0.133207	0.207530	-0.111948	0.692778	0.513007
ProdLine	-0.352049	-0.298366	0.102414	0.091924	0.099991	0.062538	-0.166390
SalesFImage	-0.217356	0.459528	-0.266895	0.127488	0.152969	-0.108868	-0.175426
ComPricing	0.132807	0.420096	0.067350	-0.168133	-0.191210	-0.581068	0.486058
WartyClaim	-0.173042	-0.205607	-0.565542	-0.280423	-0.070963	-0.047361	0.124318
OrdBilling	-0.389578	0.011909	0.183138	-0.220415	-0.170077	-0.072480	0.071845
DelSpeed	-0.423881	0.057034	0.237586	-0.181787	-0.196985	0.045881	-0.059158
Satisfaction	-0.410811	0.018122	-0.023552	0.312653	0.103993	-0.246002	0.098285

Q2.8) Mention the business implication of using the Principal Component Analysis for this case study.?

- Principal Component Analysis in this case study reduced the dimensionality of the dataset from 13 to 7 as it gives us better perspective and less complexity.
- It helps in minimizing redundant data and helps in refining useful data as when we use process-intensive algorithms (like many supervised algorithms) on the data so we need to get rid of redundancy.
- PCA gave us linearly independent and different combinations of features which we can further to describe our data differently as it gives a whole new perspective.

7 Principal Components are enough to perform further analysis (as they cover 90% of variance of the dataset).