ADVANCED STATISTIC BUSINESS REPORT

-----------------\*\*\*\*\*\*------------------

**BY:- ANIKET HIRGUDE**

***INDEX***

***CONTENT PAGE NO.***

***PROBLEM 1:***

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

***2) Perform one-way ANOVA for variable ‘Manufacturer’ with respect to the variable ‘Service Time’. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results. - 3 points***

***3) Perform one-way ANOVA for variable ‘Technician’ with respect to the variable ‘Service Time’. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results. - 3 points***

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

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

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

***INDEX***

***CONTENT PAGE NO.***

***PROBLEM 2:***

***1) Perform Exploratory Data Analysis [both univariate and multivariate analysis to be***

***performed]. The inferences drawn from this should be properly documented. – 5 points***

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

***study. - 3 points***

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

***points***

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

***exercise. - 3 points***

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

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

***7) Discuss the cumulative values of the eigenvalues. How does it help you to decide on the***

***optimum number of principal components? What do the eigenvectors indicate? Perform PCA***

***and export the data of the Principal Component scores into a data frame. – 10 points***

***8) Mention the business implication of using the Principal Component Analysis for this case***

***study. – 5 points***

***\*\*\****

**PROBLEM 1:**

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.

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

***:--***

One-way ANOVA for education Null Hypothesis H0: The mean salary is the same across all.

Alternate Hypothesis H1: The mean Service time is different in at least one category of Manufacturer.

One-way ANOVA for occupation Null Hypothesis H0: The mean Service time is the same across all. Alternate Hypothesis H1: The mean Service time is different in at least one category of Technician.

# 1.2 Perform one-way ANOVA for 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

Since the p-value = 0.826822 is less than the significance level (alpha = 0.05), we can reject the null hypothesis and conclude that there is a significant difference in the mean salaries for at least one category of Manufacturer.

# 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

# :--

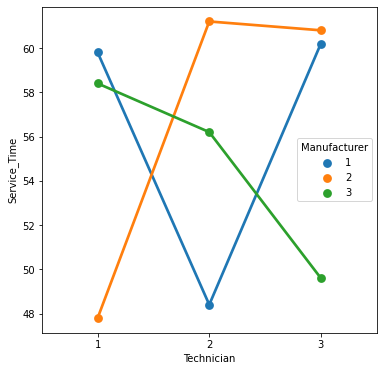
df sum\_sq mean\_sq F PR(>F)

C(Technician) 2.0 24.577778 12.288889 0.16564 0.847902

# Residual 42.0 3116.000000 74.190476 NaN NaN

Since the p-value = 0.847902 is greater than the significance level (alpha = 0.05), we fail to reject the null hypothesis (i.e., we accept H0) and conclude that there is no significant difference in the mean salaries across the categories of Technician.

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



The interaction plot shows that there is significant amount of interaction between the categorical variables, Technician & Manufacturer

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

# :--

# H0 (Null Hypothesis): The effect of the independent variable ‘Manufacturer’ on the mean ‘SERVICE TIME’ does not depend on the effect of the other independent variable ‘Occupation’ (i.e., there is no interaction effect between the 2 independent variables, Technician, Manufacturer

# H1 (Alternative Hypothesis): There is an interaction effect between the independent variable ‘Manufacturer’ and the independent variable ‘Technician’ on the mean salary

# dfsum\_sqmean\_sqFPR(>F)C(Manufacturer)2.028.31111114.1555560.2721640.763283C(Technician)2.024.57777812.2888890.2362740.790779C(Manufacturer):C(Technician)4.01215.288889303.8222225.8414870.000994Residual36.01872.40000052.011111NaNNaN

# From the table, we see that there is a significant amount of interaction between the variables, Technician & Manufacturer. As p value = 0.000994 is lesser than the significance level (alpha = 0.05), we reject the null hypothesis. Thus, we see that there is an interaction effect between Technician & Manufacturer on the mean service time

# 1.6 Explain the business implications of performing ANOVA for this particular case study

CONCLUSION The above TWO way ANOVA test we clearly see that p-value is greater than significance level(alpha = 0.05), we accept null hypothesis.

(there is no interaction effect between the 2 independent variables, Manufacturer & Technician)

\*\*

PROBLEM 2:

Problem Statement: The ‘Hair Salon.csv’ dataset contains various variables used for the

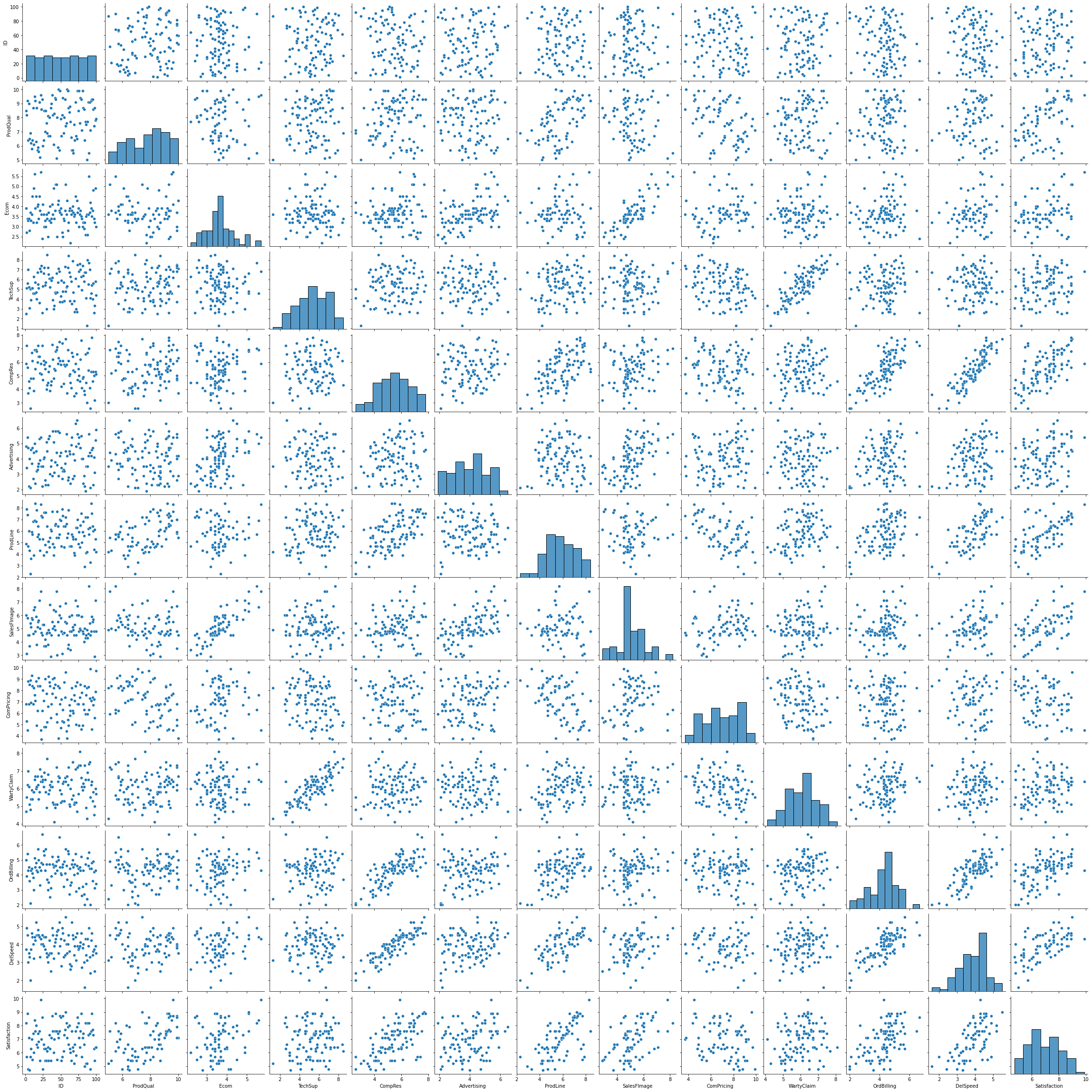
context of Market Segmentation. This particular case study is based on various parameters of a

salon chain of hair products. You are expected to do Principal Component Analysis for this case

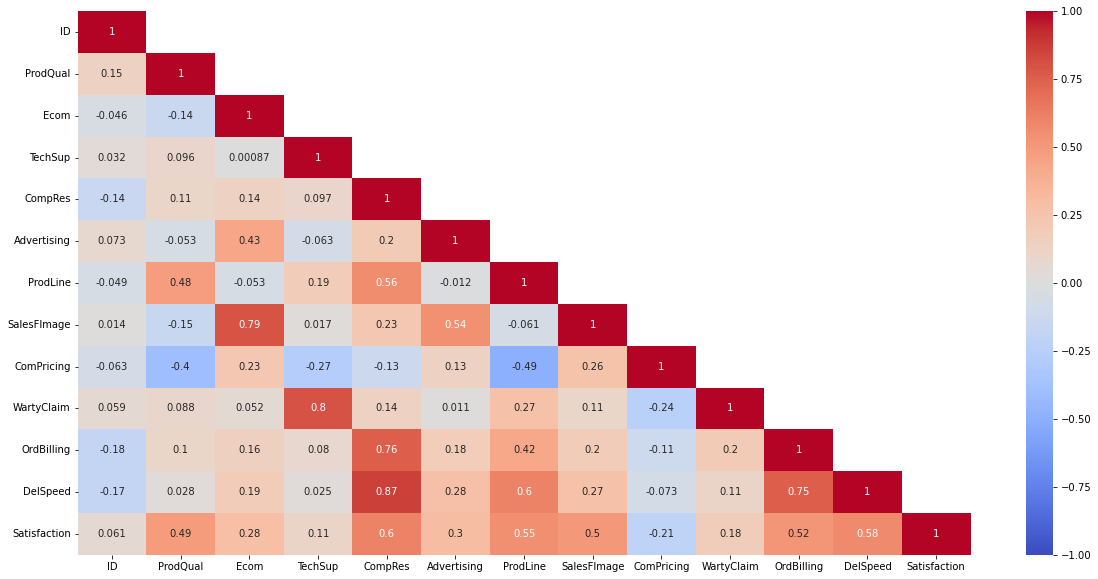
study according to the instructions given in the following rubric.

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

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



# Multivariate Analysis



Few pairs have very high co-relation

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

|  | **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 |

In [ ]:



​

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

Covariance Matrix

%s [[ 1.01010101e+00 1.47246863e-01 -4.66398292e-02 3.21596548e-02

-1.45780253e-01 7.38679326e-02 -4.91322907e-02 1.39879563e-02

-6.36432983e-02 5.91842408e-02 -1.80153686e-01 -1.73872596e-01

6.17605408e-02]

[ 1.47246863e-01 1.01010101e+00 -1.38548704e-01 9.65661154e-02

1.07444445e-01 -5.40132667e-02 4.82316579e-01 -1.53346338e-01

-4.05335236e-01 8.92043497e-02 1.05356640e-01 2.79979825e-02

4.91237372e-01]

[-4.66398292e-02 -1.38548704e-01 1.01010101e+00 8.75544162e-04

1.41595213e-01 4.34233041e-01 -5.32200387e-02 7.99539102e-01

2.31780203e-01 5.24224157e-02 1.57724577e-01 1.93571786e-01

2.85601025e-01]

[ 3.21596548e-02 9.65661154e-02 8.75544162e-04 1.01010101e+00

9.76329270e-02 -6.35051180e-02 1.94571168e-01 1.71621612e-02

-2.73521901e-01 8.05220127e-01 8.09109340e-02 2.56976702e-02

1.13734524e-01]

[-1.45780253e-01 1.07444445e-01 1.41595213e-01 9.76329270e-02

1.01010101e+00 1.98905906e-01 5.67087831e-01 2.32072486e-01

-1.29246720e-01 1.41826562e-01 7.64513729e-01 8.73829997e-01

6.09356166e-01]

[ 7.38679326e-02 -5.40132667e-02 4.34233041e-01 -6.35051180e-02

1.98905906e-01 1.01010101e+00 -1.16674936e-02 5.47680463e-01

1.35572620e-01 1.09010852e-02 1.86096560e-01 2.78649579e-01

3.07746944e-01]

[-4.91322907e-02 4.82316579e-01 -5.32200387e-02 1.94571168e-01

5.67087831e-01 -1.16674936e-02 1.01010101e+00 -6.19348764e-02

-4.99947880e-01 2.75835887e-01 4.28695202e-01 6.07929503e-01

5.56107006e-01]

[ 1.39879563e-02 -1.53346338e-01 7.99539102e-01 1.71621612e-02

2.32072486e-01 5.47680463e-01 -6.19348764e-02 1.01010101e+00

2.67269246e-01 1.08540752e-01 1.97098390e-01 2.74294201e-01

5.05257885e-01]

[-6.36432983e-02 -4.05335236e-01 2.31780203e-01 -2.73521901e-01

-1.29246720e-01 1.35572620e-01 -4.99947880e-01 2.67269246e-01

1.01010101e+00 -2.47460661e-01 -1.15724268e-01 -7.36078070e-02

-2.10399686e-01]

[ 5.91842408e-02 8.92043497e-02 5.24224157e-02 8.05220127e-01

1.41826562e-01 1.09010852e-02 2.75835887e-01 1.08540752e-01

-2.47460661e-01 1.01010101e+00 1.99055678e-01 1.10499598e-01

1.79338201e-01]

[-1.80153686e-01 1.05356640e-01 1.57724577e-01 8.09109340e-02

7.64513729e-01 1.86096560e-01 4.28695202e-01 1.97098390e-01

-1.15724268e-01 1.99055678e-01 1.01010101e+00 7.58588957e-01

5.27001932e-01]

[-1.73872596e-01 2.79979825e-02 1.93571786e-01 2.56976702e-02

8.73829997e-01 2.78649579e-01 6.07929503e-01 2.74294201e-01

-7.36078070e-02 1.10499598e-01 7.58588957e-01 1.01010101e+00

5.82870984e-01]

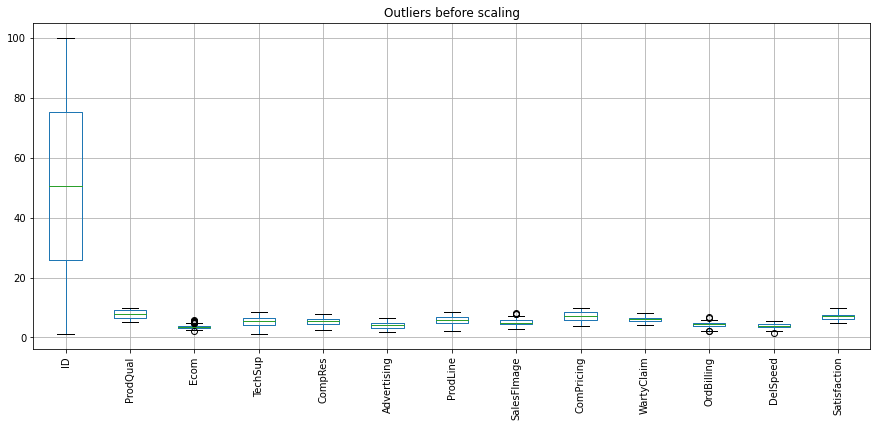
[ 6.17605408e-02 4.91237372e-01 2.85601025e-01 1.13734524e-01

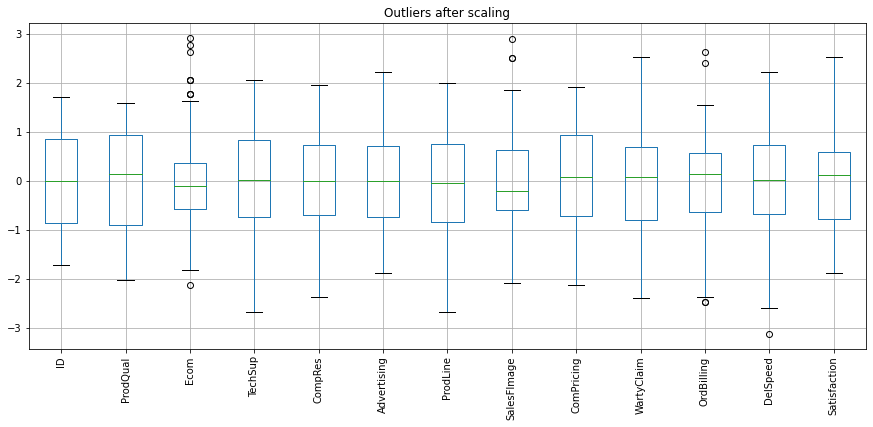
6.09356166e-01 3.07746944e-01 5.56107006e-01 5.05257885e-01

-2.10399686e-01 1.79338201e-01 5.27001932e-01 5.82870984e-01

1.01010101e+00]]

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





***The above scaling has not removed any outliers. An outlier treatment needs to be done subsequently. But since we aren’t asked to do so we won’t remove outliers.***

***The standard deviation of all variables has come to 1***

# 2.5 Build the covariance matrix, eigenvalues and eigenvector.

Eigen Values

%s [4.09031163 2.58246009 1.74408311 1.38449513 0.84530284 0.63737703

0.55277881 0.40687424 0.32136047 0.23782883 0.14469341 0.08361313

0.10013442]

Eigen Vectors

%s [[ 0.04659362 0.04983935 -0.23101384 -0.49737305 -0.77967247 -0.11581052

0.2434289 0.00670863 0.10531516 -0.03330521 0.03642261 0.02144534

-0.00261414]

[-0.15527921 0.31871546 0.00350694 -0.52458938 0.30438026 -0.26205579

-0.40025604 0.12242084 0.29399737 -0.18687338 -0.2077462 0.21495202

-0.22775251]

[-0.16649956 -0.43719256 -0.24857742 -0.08627072 0.30202119 -0.09075031

0.4135531 0.0069788 0.51442042 -0.23131479 -0.02350541 -0.34920787

0.02731935]

[-0.12357735 0.24176757 -0.56970117 0.29368824 -0.01164363 -0.05256564

-0.12809222 -0.01430496 -0.10903036 -0.53917654 0.42537132 0.11007793

0.01723457]

[-0.42372779 -0.00333435 0.21422033 0.17225577 -0.20503738 -0.05876359

0.03792792 -0.00299504 -0.13815804 -0.44253679 -0.58002662 0.05228422

0.37927468]

[-0.17945503 -0.35166727 -0.13320679 -0.20753035 -0.11194799 0.69277812

-0.51300709 -0.07491646 0.09001346 -0.03673186 0.02395951 -0.05105266

0.0970485 ]

[-0.35204872 0.2983661 0.10241352 -0.09192432 0.09999132 0.06253771

0.16638963 -0.63459988 0.22501788 0.22952031 0.25381921 0.18824266

0.34729945]

[-0.21735578 -0.45952847 -0.26689457 -0.12748761 0.15296919 -0.10886837

0.17542557 0.02358494 -0.32985481 0.17964334 -0.05235779 0.6622454

-0.0716641 ]

[ 0.13280731 -0.4200957 0.06735008 0.16813296 -0.19120973 -0.58106764

-0.4860585 -0.34443129 0.15530003 0.02644731 0.08811167 -0.0100872

0.10633714]

[-0.17304207 0.20560727 -0.56554168 0.28042302 -0.0709626 -0.04736066

-0.12431805 -0.04038476 0.12226945 0.49812059 -0.45618859 -0.16428112

-0.08229667]

[-0.38957753 -0.01190881 0.18313799 0.22041485 -0.17007695 -0.07248005

-0.0718449 0.62756051 0.3456201 0.24437143 0.32499983 0.15271613

0.15708004]

[-0.42388058 -0.05703423 0.23758551 0.18178685 -0.19698481 0.04588085

0.05915848 -0.23611464 0.01042558 -0.07837693 0.06159741 -0.05566047

-0.78375834]

[-0.41081139 -0.01812171 -0.02355175 -0.31265255 0.10399316 -0.24600185

-0.09828528 0.0772913 -0.52981233 0.15141982 0.21109323 -0.53518017

0.10500781]]

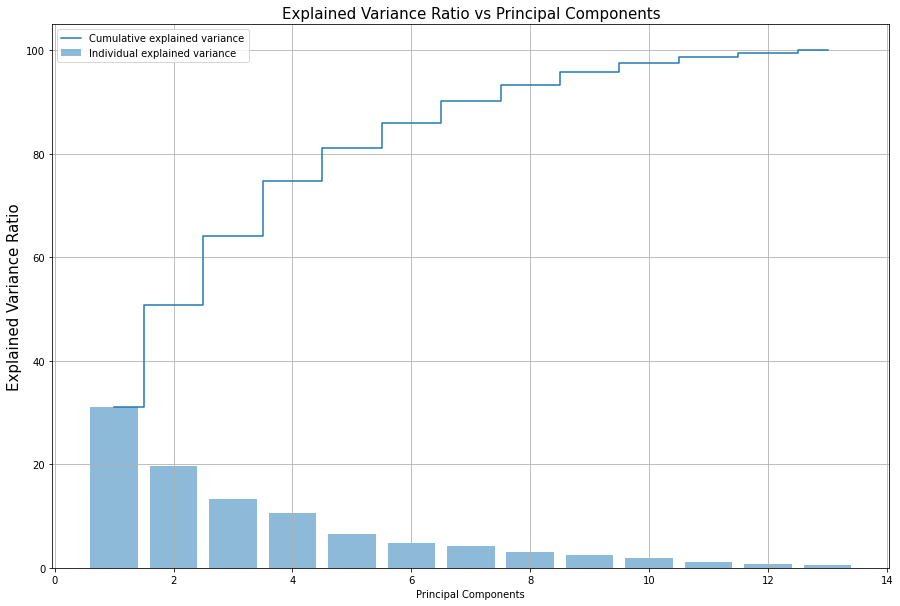
# 2.6 Write the explicit form of the first PC (in terms of Eigen Vectors)

# 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 Variance Explained [ 31.14929626 50.81572309 64.09758676 74.64104968 81.07835592

85.93222712 90.14185037 93.24035419 95.68763779 97.49879577

98.60069174 99.36325388 100. ]



***Based on above cumulative values and explained variance ratio we clearly see that the first 7 PC’s (Principal components) capture 90% of the variances of the data. Hence, we can decide that the number of principal components for our dataset is 7.***

***Each eigen vector represents a direction of variance. The eigen vector corresponding to the largest eigen value will give the direction of pg. 31 maximum variance. This is the first principal component. Then, the eigen vector corresponding to the 2nd largest eigen value will give the direction of the second largest variance. This is the second principal component. And, so on.***

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

***This business case study is about Hair Salon dataset which contain the names of various Market Segmentation., which has various details of various parameters of salon chain of hair products.. To understand more about the dataset, we perform univariate analysis and multivariate analysis which gives us the understanding about the variables. From analysis we can understand the distribution of the dataset, skew, and patterns in the dataset. From multivariate analysis we can understand the correlation of variables. The scaling helps the dataset to standardize the variables in one scale. The principal component analysis is used reduce the multicollinearity between the variables. Depending on the variance of the dataset we can reduce the components. The PCA components for this business case is 7 where we could capture the maximum variance of the dataset. Using the components, multicollinearity of variables in the dataset has been reduced which further will help in building machine learning models.***

***\*\*\*END\*\*\****