Project: Advanced Statistics

HARESH P TAYADE

PGP-DSBA ONLINE

SEPT'2021

Date: 12/12/2021

Project: Advanced Statistics

Contents

ANNOVA

(Problem 1A & 1B)

1. State the null and the alternate hypothesis for conducting one way annova for both education and occupation individually
2. Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.
3. Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.
4. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. **(Non-Graded)**
5. What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the 'pointplot' function from the 'seaborn' function]
6. Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?
7. Explain the business implications of performing ANOVA for this particular case study.

(Problem 2)

1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?
2. Is scaling necessary for PCA in this case? Give justification and perform scaling.
3. Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data]
4. Check the dataset for outliers before and after scaling. What insight do you derive here?
5. Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both]
6. Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features
7. Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]
8. Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?
9. Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]

List of Figures:

List of Tables:

Executive Summary

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csv] are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

[Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.]

Introduction

The purpose of this whole exercise is to explore the dataset.Do the exploratory data analysis.as the dataset is about the salary of 40 individuals from different educational and occupation.

1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

Answer:

For Education

Ho- null hypothesis-the mean salary of educational qualification is same at three level (high school graduate,bachelor,doctorate).

H1- Alternate hypothesis-the mean salary is different in one of the educational qualification

For Occupation

Ho - Null Hypothesis - the mean salary of occupation is same at all 4 levels (administrative,clerical,sales,professional).

H1 - Alternate Hypothesis - the mean salary of occupation is different at atleast 1 levels from the given 4 .

Where significance difference=0.05

if p value is lesser then 0.05 we reject the null hypothesis

if p value is greater or equal to 0.05 we accept the null hypothesis.

1.2 Perform one-way ANOVA for Education with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

Answer:

TABLE 1: LIST OF EDUCATION

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

```
# As we can see above the p value is less than (0.05),we can say that we reject the null Hypothesis Ho.
```

1.3 Perform one-way ANOVA for variable Occupation with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

Answer:

TABLE 2: LIST OF OCCUPATION

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	0.884144	0.458508
Residual	36.0	1.528092e+11	4.244701e+09	NaN	NaN

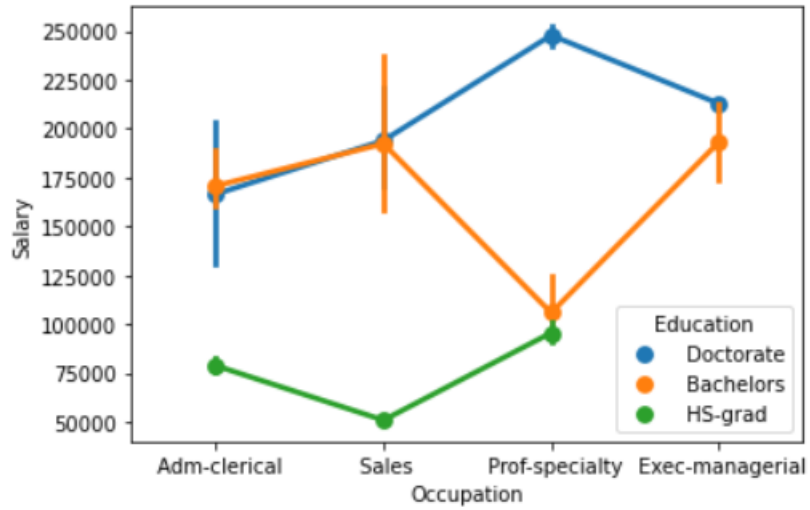
```
# In this as we can see that the p value is greater than (0.05), we can say that we cannot reject the null hypothesis Ho.
```

1.5 What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.

Answer:

FIGURE 1: INTERACTION BETWEEN EDUCATION AND OCCUPATION

```
<AxesSubplot:xlabel='Occupation', ylabel='Salary'>
```



From above plot we can make the Interaction .

1. Adm-clerical has good interaction with Doctorate and Bachelors
2. Bachelors and Doctorates is good with Sales.
3. Exec-managerial has no interaction with Education.
4. As compare to salary Sales having low salary as compare to Doctorate and bachelor.
5. Exec-managerial foes have high secondary grade.
6. As compared to bachelors and HS grad having good Interaction in prof-speciality.

1.6 Perform a two-way ANOVA based on the Education and Occupation (along with their interaction Education*Occupation) with the variable 'Salary'. State the null and alternative hypotheses and state your results. How will you interpret this result?

Answer:

The null and alternate hypothesis for 2 way anova test for Occupation and Education are

Null hypothesis - H0 -the mean salary for Occupation and education are same

Alternate Hypothesis - H1 - the mean salary is different for atleast 1 of the (education or Occupation)

for p value = 0.05

if p value is less than 0.05 we need to reject the null hypothesis.

if p value is more than 0.05 we need to accept the null hypothesis.

TABLE 3: TWO WAY ANNOVA

Two way anova

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	31.257677	1.981539e-08
C(Occupation)	3.0	5.519946e+09	1.839982e+09	1.120080	3.545825e-01
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

In this two way anova test we can see the interaction between two treatment.

so need to do again two way anova with :

TABLE 4 : TWO WAY ANNOVA

Two way anova:

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

as from above both the table we can see changes in p value of education and occupation.

as we can see that the p value of "Education' is reject because the null hypothesis is rejected in this case.

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

Answer:

Anova stands for “analysis of variance” and it is used for statistics for hypothesis testing to see interaction between independent variable and dependent variable.

Anova is used when two or more group are compared.

As we can see from above table, education salary increase, and the doctorate salary is high than bachelors and hs-grad. But sometime doctorate is not getting high salary than hs-grad and bachelors. So that doctorate can be prefer for all or not or they are not qualified for other jobs. As for occupation is getting lesser salary than education , we can note in graph that high salary is offered to degree holder than doctorates for few occupations.As per years of experience , specialisation ,or domain knowledge as per that it can change the salary cycle the data for years of experience ,specialization,or domain ,job profile ,knowledge is missing .so as per that we can say the data missing

ANOVA test indicates that the education level coupled with occupation has significant influence over salary that alone occupation type with comparison to Educational background.

(Problem 2)

Executive Summary

The dataset Education - Post 12th Standard.csv contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx.

- *Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?*
- *Is scaling necessary for PCA in this case? Give justification and perform scaling.*
- *Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].*
- *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]*
- *Extract the eigenvalues and eigenvectors.[Using Sklearn PCA Print Both]*
- *Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features*
- *Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]*
- *Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?*
- *Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [**Hint:** Write Interpretations of the Principal Components Obtained]*

2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

Answer:

First we should know that what answers are we need to find. Do the data is correct or not, we need to check the data , do we have any missing value, is there any duplicate values ,or any outliers . we have to perform as annova test (one –way, two-way),EDA ,PCA.

EDA

Univariate analysis:

Univariate is to analysis single variable in this we need to find the patterns of data.

The describe function shows all the numeric, from this we can plot a graph as well as we can check the outliers.

TABLE 5: STATISTICAL DESCRIPTION OF DATA SET

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.380952	165.105360	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.171171	5221.768440	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.463320	17.177710	10.0	53.0	65.0	78.0	118.0

Observations:

From above table we can see that:

1) The data consists of 777 rows and 18 coloumns

2) As we can see that it has 777 different universities with 18 variables

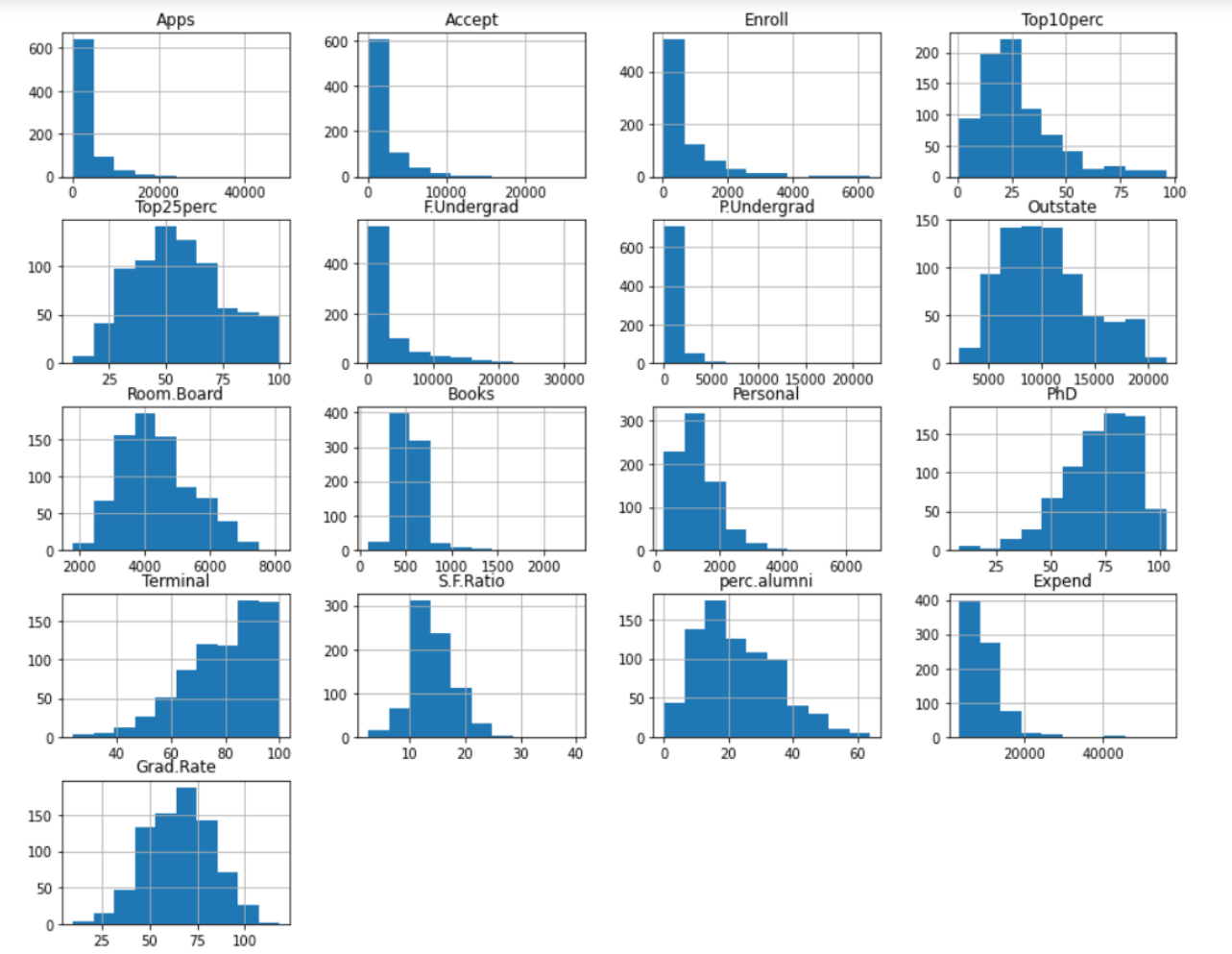
3) we have different index as follows:
'Apps,Accept,Enroll,Top10perc,Top25perc,F.undergrad,P.undergrad,Outstate,Room.Board,Books,Personal,Phd,Terminal,S.f.ratio,Perc.alumni,Expend,Grad.rate'.

4) From above table we can see that perc.alumni has minimum 0.0 value and the maximum value is 64.0 in the dataset.

5) P.Undergrad has a minimum value 1 and the maximum value is 21836.0

6) Top10perc has a minimum value 1 and the maximum value is 96.0

FIGURE 2.1.1:HIST PLOT TO CHECK DISTRIBUTION AND DENSITY OF VARIABLES



Hist plot for check Distribution and density of each variable.

Observations:

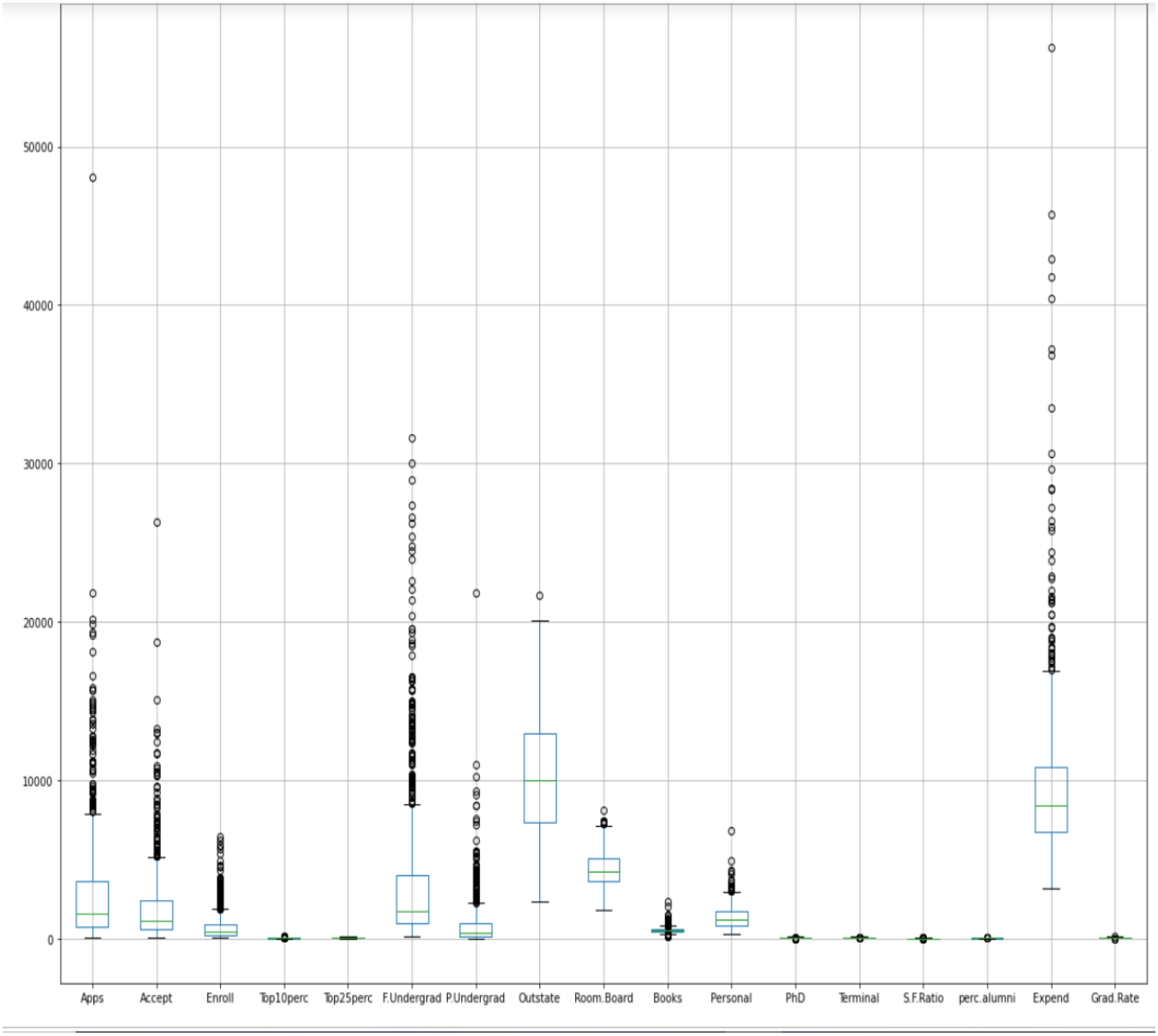
As from above table we can see that

'Phd' & 'Terminal' has right skewed data

Data is normally distributed in 'grad.rate','top25perc'. 'Apps','Accept','Enroll','Top10perc',F.undergrad,'P.undergrad','outstate','roomboard','books','personal','s.f.ratio','perc.alumini','expend'.

BOX PLOT for outliers in each variable or univariate analysis of all variable.

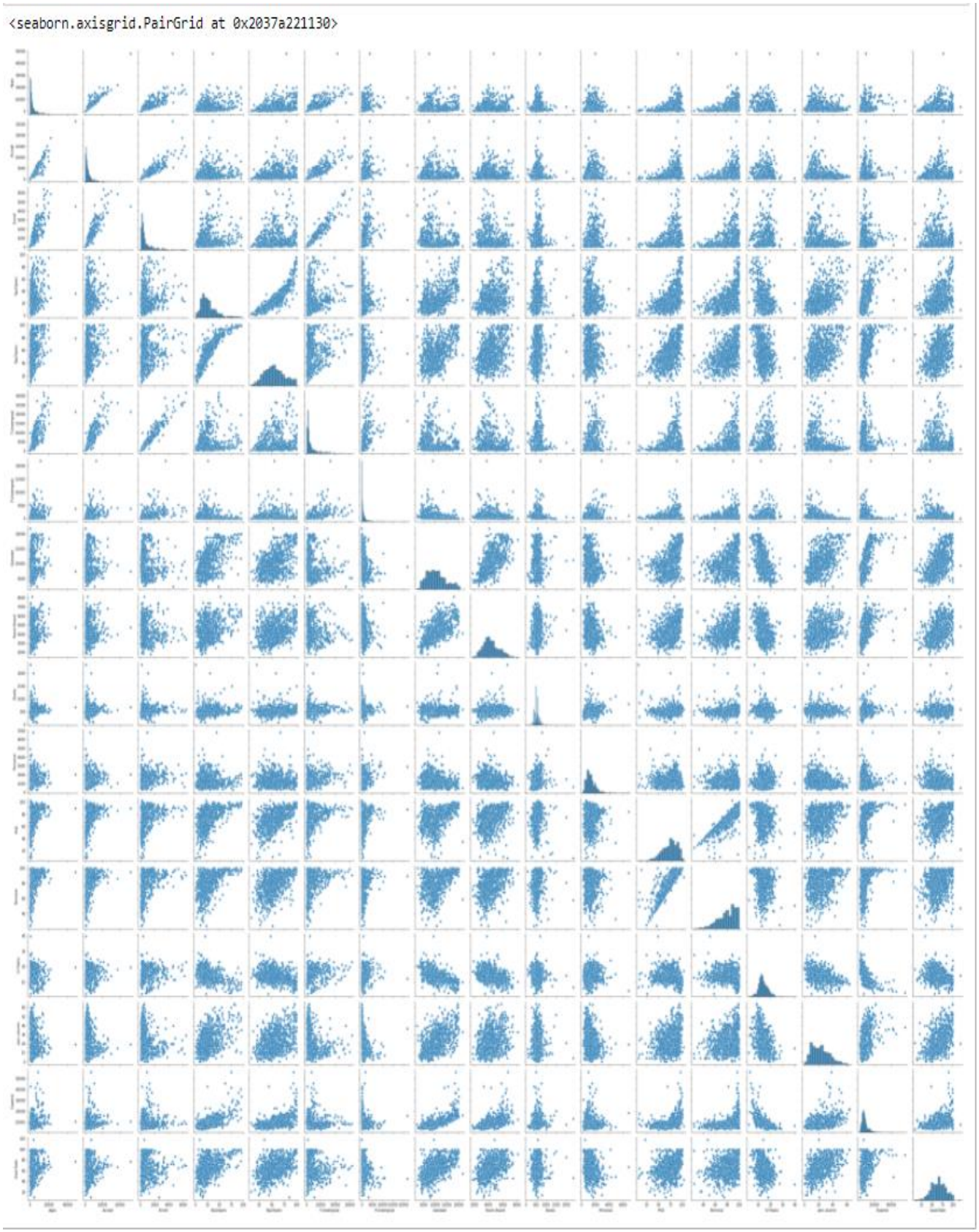
FIGURE 2.1.2: BOX PLOT FOR OUTLIERS



Multivariate analysis:

Pairplot to see relationship of all variables among each other.

FIGURE 2.1.3: PAIR PLOT FOR RELATION OF VARIABLE



Observation:

Few pair have very high correlation

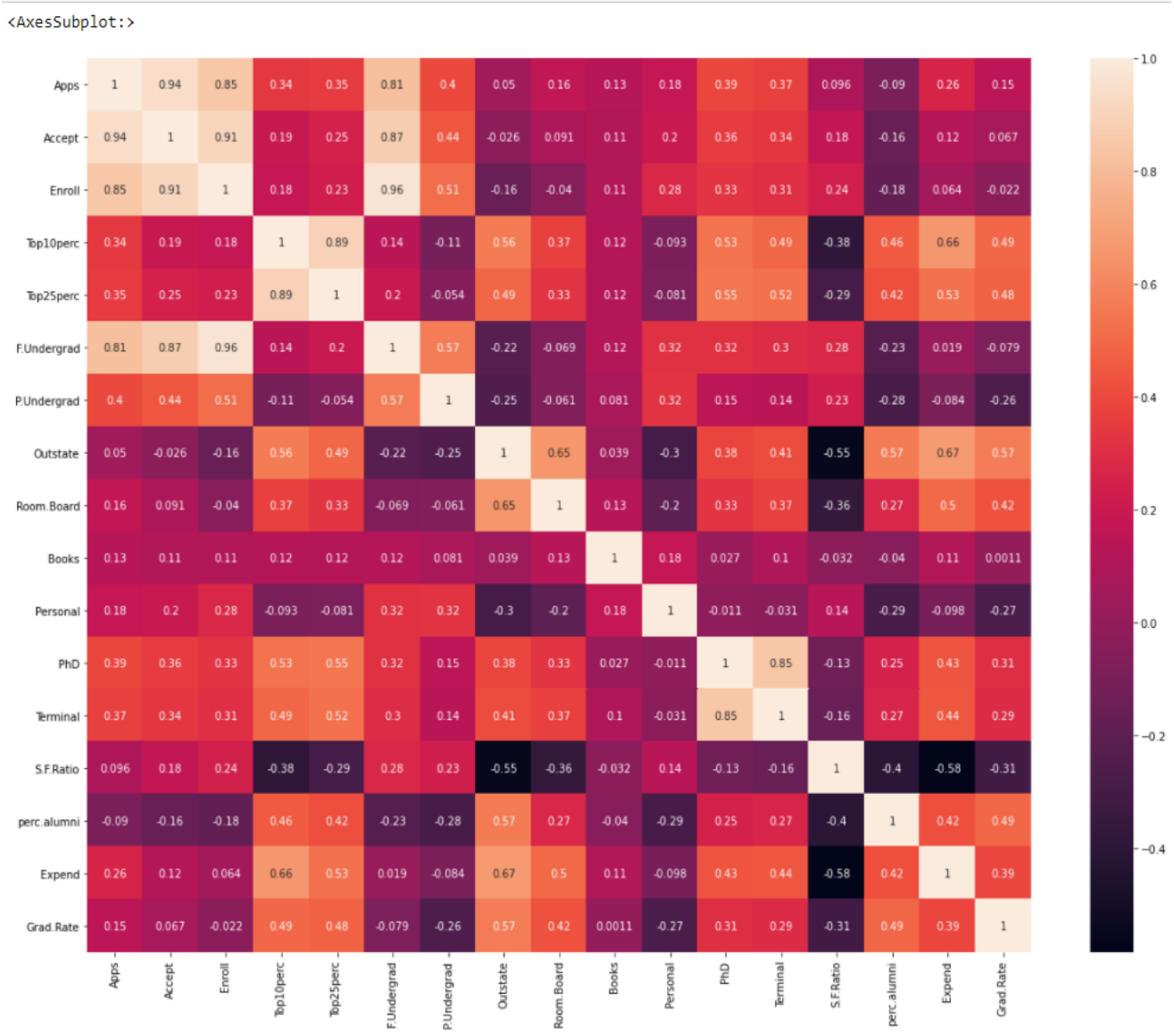
Application and acceptance

pHd and terminal

Below heatmap exhibits multicollinearity issue as significant number of high co-relation variables pairs/features.

FIGURE 2.1.4:HEAT MAP FOR COLLINEARITY

Heatmap to check the collinearity of original data



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

Answer:

yes it is necessary to normalize data before performing PCA.Scaling of data can be done using Z-score method .

z=value - mean/standard deviation

$$z = \frac{value - mean}{standard\ deviation}$$

As by the formula for z-score.

Figure 2.2.1: Histogram Plot before scaling

Histogram Plot before scaling

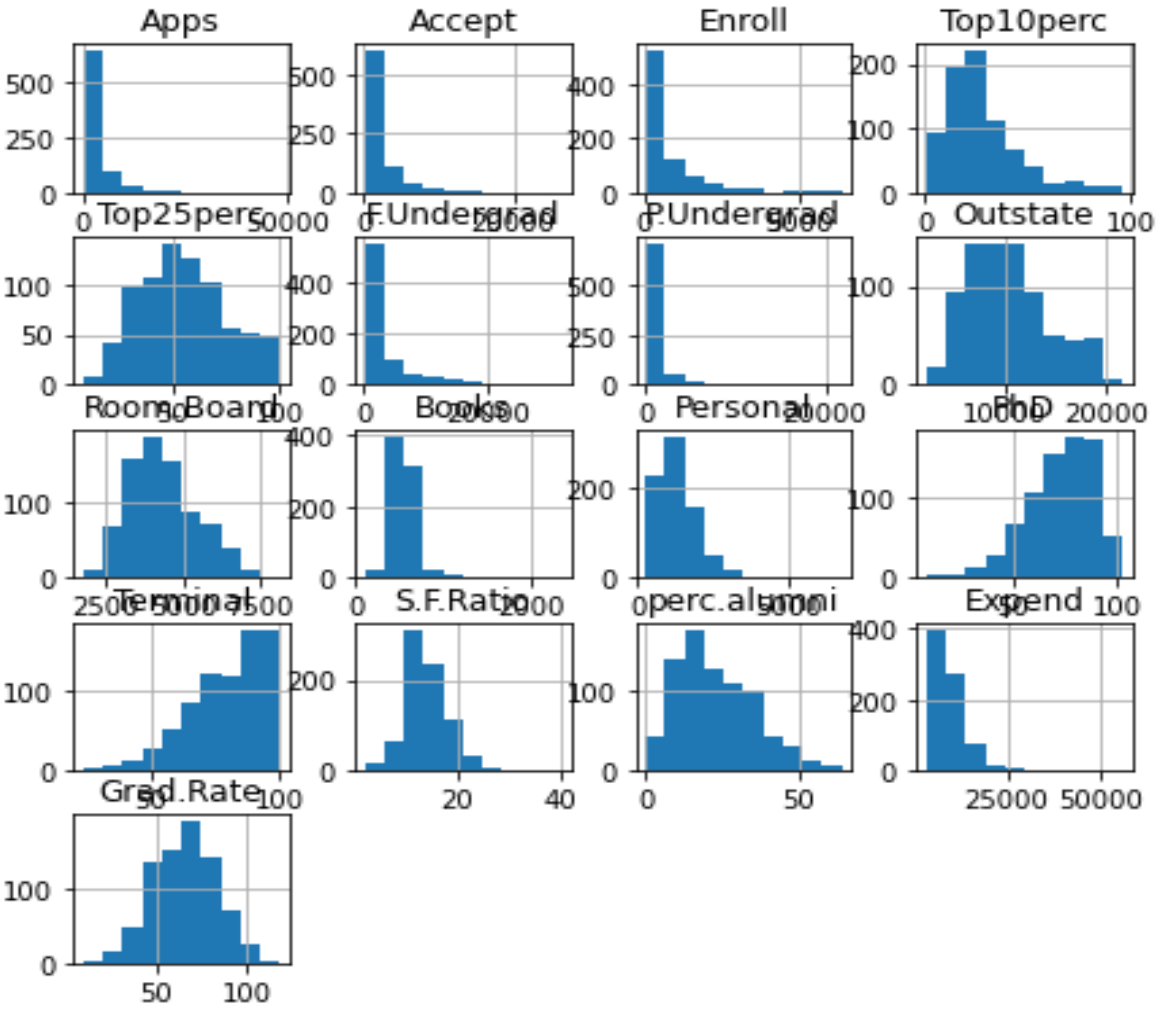


Figure 2.2.2: Histogram plot after scaling

Histogram plot after scaling

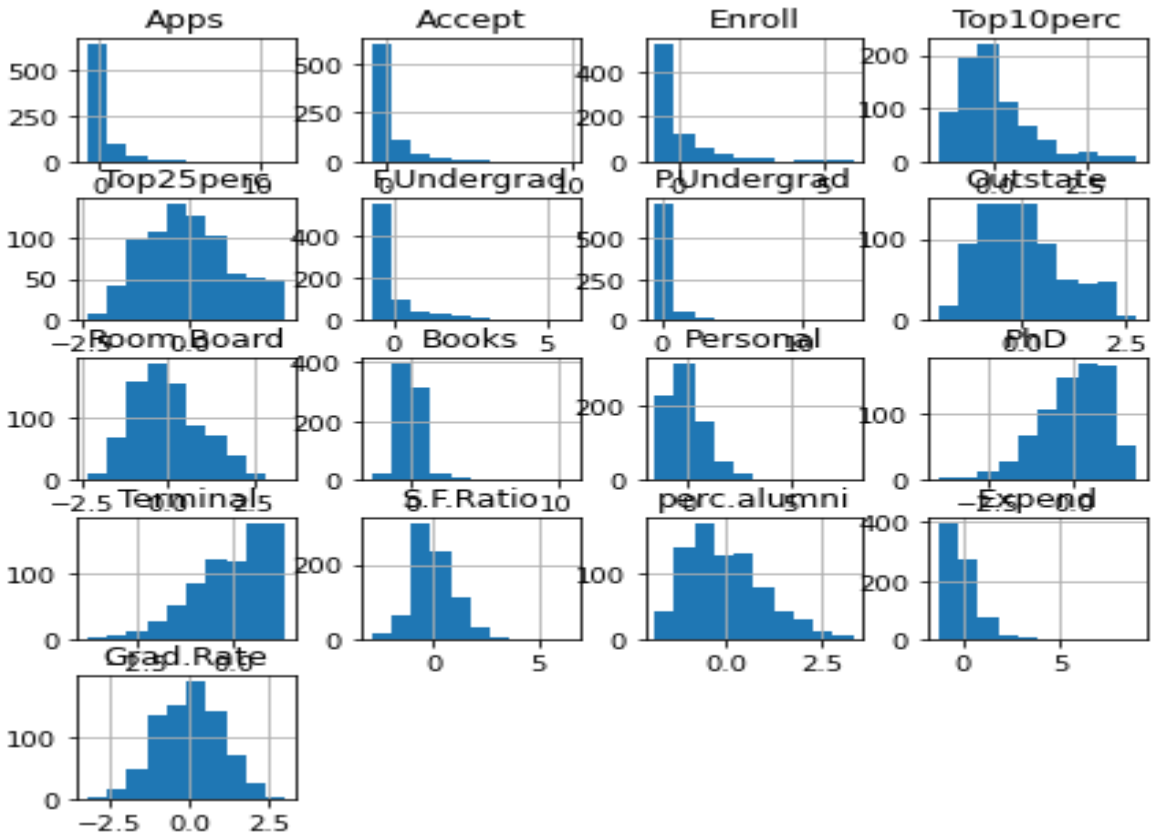


TABLE 6:FOR SCALING OF DATA SET

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Rat
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.163028	-0.115729	1.01377
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.675646	-3.378176	-0.47770
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1.204845	-0.931341	-0.30074
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.185206	1.175657	-1.61527
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0.204672	-0.523535	-0.55354

As we have done the scaling as in the before graph we have done the after Scaling.

2.3 Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data]

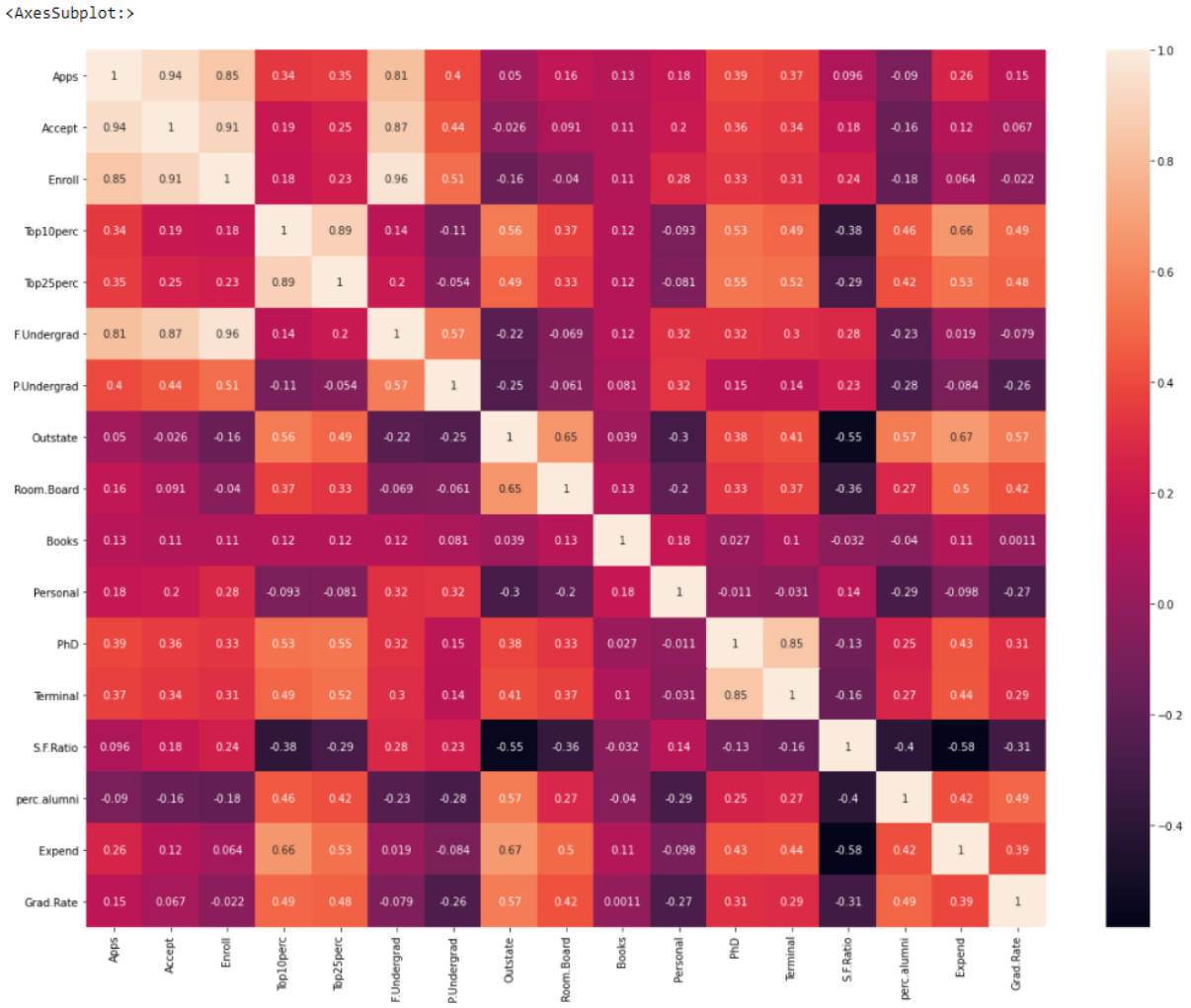
Answer:

 # covariance indicates the direction of linear relationship between variables

 # correlation measures both strength and direction of linear relationship between two variables

 # both covariance and correlation measure relationship.

FIGURE 2.3.1: CORRELATION OF DATA SET



 # From this above table

#Low correlation

 # apps & s.f.ratio has low correlation

 # S.F ratio and terminal has low correlation

 # Expend and S.F.ratio has low correlation

 # High correlation

 # enroll with f.undergrad

 # enroll with accept

 # apps with accept & accept with apps.

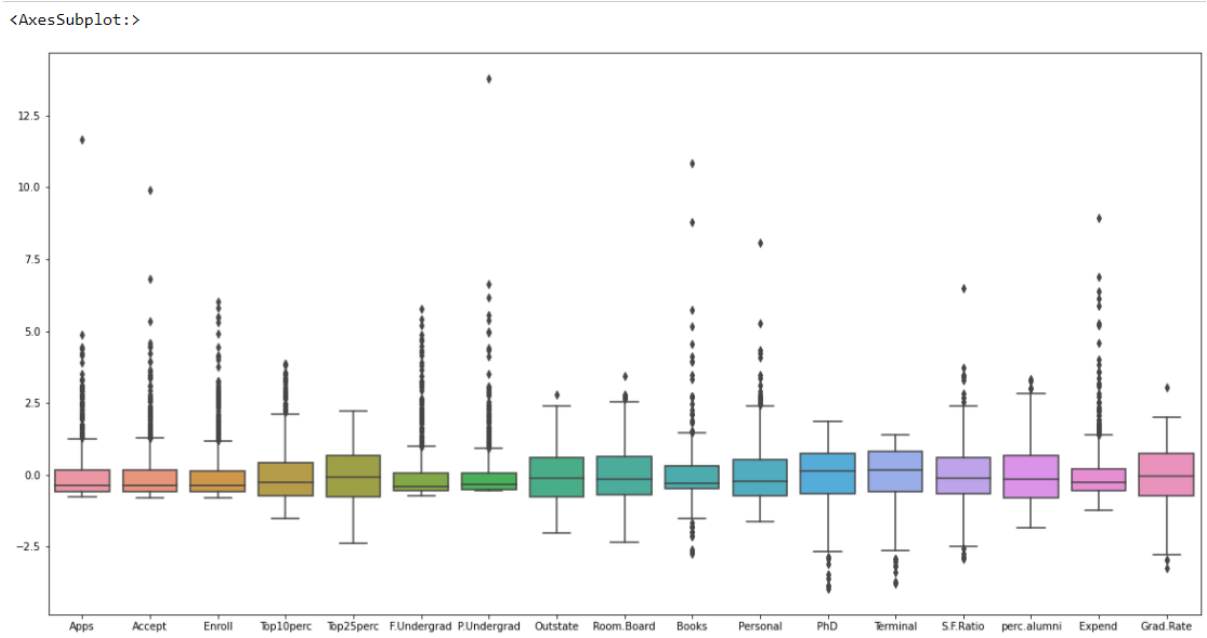
2.4 Check the dataset for outliers before and after scaling. What insight do you derive here?

Answer:

While performing univariate analysis we have plotted boxplot for all variable for checking outliers presence.

After scaling no much difference in terms of outliers reduction.

FIGURE 2.4.1-BOX PLOT FOR SCALES DATA ALL VARIABLE.



As in above we can see that there is no such difference.However the outliers, have an influence when computing the empirical mean and standard deviation .

TABLE 7 – SCALED DATA SUMMARY

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	6.355797e-17	1.000644	-0.755134	-0.575441	-0.373254	0.160912	11.658671
Accept	777.0	6.774575e-17	1.000644	-0.794764	-0.577581	-0.371011	0.165417	9.924816
Enroll	777.0	-5.249269e-17	1.000644	-0.802273	-0.579351	-0.372584	0.131413	6.043678
Top10perc	777.0	-2.753232e-17	1.000644	-1.506526	-0.712380	-0.258583	0.422113	3.882319
Top25perc	777.0	-1.546739e-16	1.000644	-2.364419	-0.747607	-0.090777	0.667104	2.233391
F.Undergrad	777.0	-1.661405e-16	1.000644	-0.734617	-0.558643	-0.411138	0.062941	5.764674
P.Undergrad	777.0	-3.029180e-17	1.000644	-0.561502	-0.499719	-0.330144	0.073418	13.789921
Outstate	777.0	6.515595e-17	1.000644	-2.014878	-0.776203	-0.112095	0.617927	2.800531
Room.Board	777.0	3.570717e-16	1.000644	-2.351778	-0.693917	-0.143730	0.631824	3.436593
Books	777.0	-2.192583e-16	1.000644	-2.747779	-0.481099	-0.299280	0.306784	10.852297
Personal	777.0	4.765243e-17	1.000644	-1.611860	-0.725120	-0.207855	0.531095	8.068387
PhD	777.0	5.954768e-17	1.000644	-3.962596	-0.653295	0.143389	0.756222	1.859323
Terminal	777.0	-4.481615e-16	1.000644	-3.785982	-0.591502	0.156142	0.835818	1.379560
S.F.Ratio	777.0	-2.057556e-17	1.000644	-2.929799	-0.654660	-0.123794	0.609307	6.499390
perc.alumni	777.0	-6.022638e-17	1.000644	-1.836580	-0.786824	-0.140820	0.666685	3.331452
Expend	777.0	1.213101e-16	1.000644	-1.240641	-0.557483	-0.245893	0.224174	8.924721
Grad.Rate	777.0	3.886495e-16	1.000644	-3.230876	-0.726019	-0.026990	0.730293	3.060392

So even if there is an outliers in the data , they will not be affected by Standardization.

2.5 Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both]

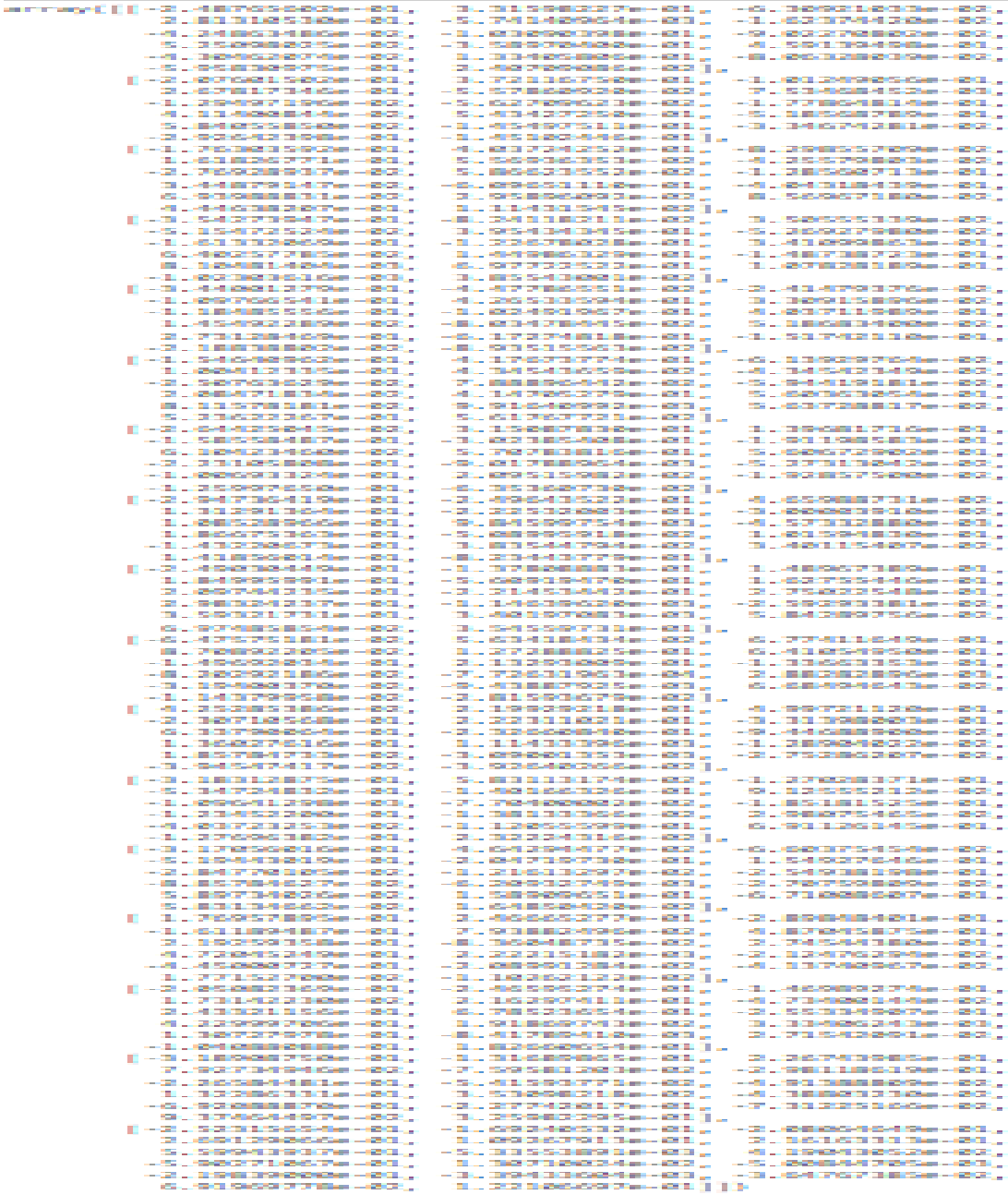
Answer:

We need to get eigen value and eigen vector .

TABLE 8- FOR EIGEN VALUE.

```
array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,
       0.84849117, 0.6057878 , 0.58787222, 0.53061262, 0.4043029 ,
       0.02302787, 0.03672545, 0.31344588, 0.08802464, 0.1439785 ,
       0.16779415, 0.22061096])
```

TABLE 9: FOR EIGEN VECTOR



As the above table (Eigen vector) we can see clear in Jyupter file.

2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features.

Answer:

in PCA given the mnean center dataset x with n sample and p variables

the first principal component PC1 isgiven by the linear combination of the original variables X_1, X_2, …,

X_pPC_1 = w_{17}X_1 + w_{16}X_2 + … + w_{1p}X_p

The first principal component PC1 represents the component that retains the maximum variance ofthe data.

w1 corresponds to aneigenvector of the covariance matrix

∑ = (1/N-1)X

2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

Answer:

2.8 Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?

Answer:

2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]

Answer:

The business implication of using PCA

#PCA is used in EDA and for making predictive models can only be done on continuous variable. #PCA used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain #lower-dimensional data while preserving as much of the data's variation as possible. # In this case we can reduce dimensions from 17 to 9that explains over 110%variances. #The first principal component can equivalently be defined as a direction that maximizes the variance of the projected data #The ith principal component can be taken as a direction (i.e. at 110 degrees to one another.) #to the first-1principal components that maximizes the variance of the projected data.

_ _

_ _ _ _ _ _ _ _ _ _ _ _*OVER*_ _ _ _ _ _ _ _ _ _ _ _