Business Report

On

Education - Post 12th Standard & Salary dataset

Neha Mishra

PGP_DSBA Online

Dec 22021

Date:03/04/2022

Table of contents

Problem 2-Education Dataset:-

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 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.

The dataset Education - Post 12th Standard.csv contains information about names of various universities and colleges, numbers of applications received, number of applications accepted, number od new students enrolled, percentage of new students from top 10% of higher secondary class, percentage of new students from top 25% of higher secondary class, number of full-time undergraduate students, number of part-time undergraduate students, number of students for whom the particular college or university is out-of-state tuition, cost of room and board, estimated book costs for a student, estimated personal spending for student, percentage of facilities with Ph.Ds.', percentage of facilities with terminal degree, student /faculty ratio, percentage of alumni who donate, the instructional expenditure per student and graduate rate.

Data Description

Education	: Doctorate, Bachelors and HS-grad
Occupation	: Adm-clerical, Sales, Prof-specialty and Exec-managerial
Salary	: Continuous from 50103.0 to 260151.0

Sample of Dataset

	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769

Table no 1 : Sample Dataset

Exploratory Data Analysis

Let us check the types of variables in the data frame.

Education : object
Occupation : object
Salary : int64

Check for missing values in the dataset:

	Column	Non-Null Count	Count	Dtype
0	Education	non-null	40	object
1	Occupation	non-null	40	object
2	Salary	non-null	40	int64

Table no. 2 : Missing value check

1.1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for Education

$H0$: The means of 'Education' variable with respect to each Salary is equal.

$H1$: At least one of the means of 'Education' variable with respect to each Salary is unequal.

1.1.2 State the null and the alternate hypothesis for conducting one-way ANOVA for Occupation

$H0$: The means of 'Occupation' variable with respect to each Salary is equal.

$H1$: At least one of the means of 'Occupation' variable with respect to each Salary is unequal.

1.2 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

	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

Table no.3 : One way Anova test for Education

So, here in this case the p_value(1.257709e-08) is less than 0.05. We have enough proof to reject null hypothesis.
Conclude that At least one of the means of 'Education' variable with respect to each Salary is unequal

1.3 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

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

Table no.4: One way Anova test for Occupation

So, here in this case the p_value (0.458508) is greater than 0.05. We have enough proof to accept null hypothesis.
Conclude that The means of 'Occupation' variable with respect to each Salary is equal.

1.4 If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result.

The Null Hypothesis is rejected in Scenario Education .We will check class mean are different due to difference in the group means by using Tukeyhsd( ) function in Python.

group1	group2	p-adj	reject
Bachelors	Doctorate	0.0146	True
Bachelors	HS-grad	0.001	True
Doctorate	HS-grad	0.001	True

Table no.5 : Tukeyhsd result

1.5 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.

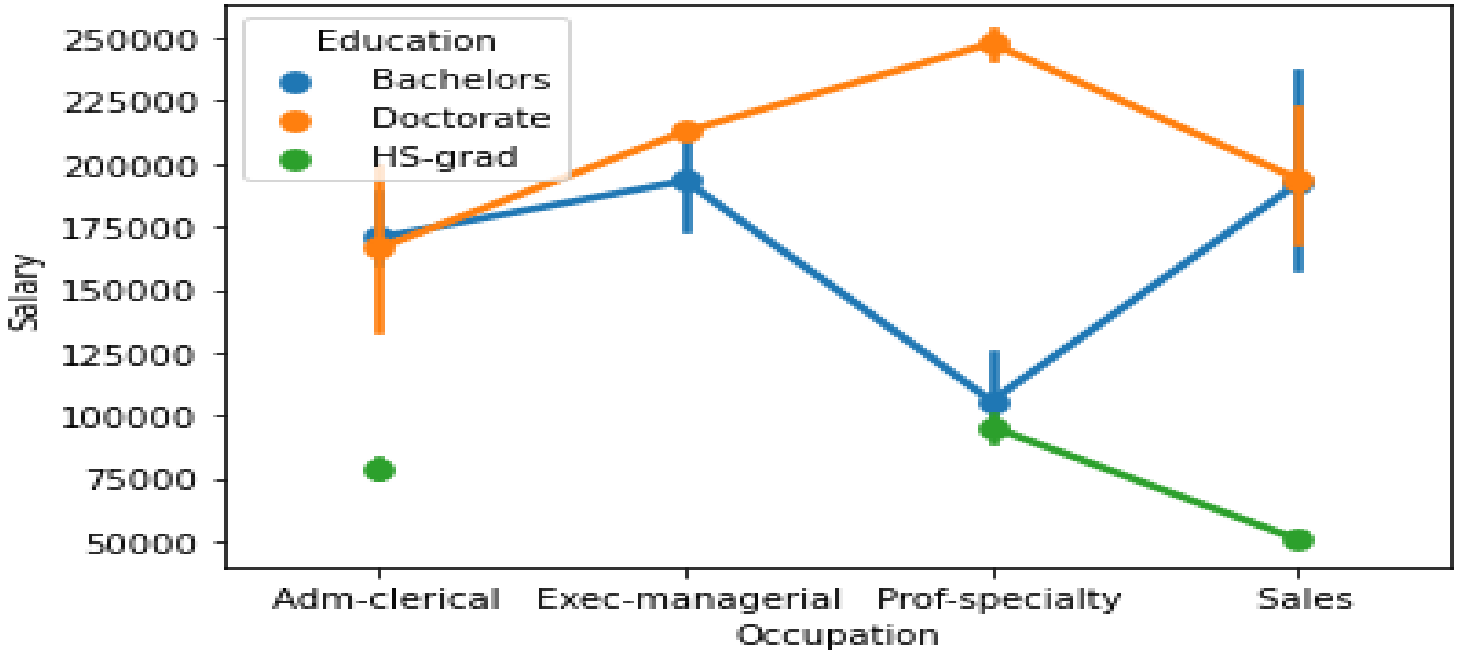

Fig 1: Point plot

From the above point plot, there seems to be an interaction between the two Education & Occupation variables.

Doctorate & Bachelors from Education variables corresponds with Adm-Clerical & Sales from salary variable.

It also seems that no interaction between HS-grad with occupation.

1.6 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?

$H0$: There is no interaction between factors of Occupation and Education.

$H1$: there is an interaction between factors of Occupation and Education.

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	5.277862	4.993238e-03
C(Education)	2.0	9.695663e+10	4.847831e+10	68.176603	1.090908e-11
C(Occupation): C(Education)	6.0	3.523330e+10	5.872217e+09	8.258287	2.913740e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Table no.6 : Two way Anova test between Occupation and Education

Since, the p_value (4.993238e-03) is less than 0.05. We have enough proof to reject null hypothesis.
Conclude that there is an interaction between factors of Occupation and Education.

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

By performing One way test between Education and Occupation with respect to Salary We found that different in the means of Education with Salary variable and constant means of Occupation with respect to salary
And by Two-way Anova test between Education and Occupation we found that there are some sort of interaction between factor of Education and Occupation.

Exploratory Data Analysis

Let us check the types of variables and missing value in the data frame.

	Column	Non-Null Count	Dtype
0	Names	777 non-null	object
1	Apps	777 non-null	int64
2	Accept	777 non-null	int64
3	Enroll	777 non-null	int64
4	Top10perc	777 non-null	int64
5	Top25perc	777 non-null	int64
6	F.Undergrad	777 non-null	int64
7	P.Undergrad	777 non-null	int64
8	Outstate	777 non-null	int64
9	Room.Board	777 non-null	int64
10	Books	777 non-null	int64
11	Personal	777 non-null	int64
12	PhD	777 non-null	int64
13	Terminal	777 non-null	int64
14	S.F.Ratio	777 non-null	float64
15	perc.alumni	777 non-null	int64
16	Expend	777 non-null	int64
17	Grad.Rate	777 non-null	int64

Table no.6: data type

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

Univariate Analysis

Univariate analysis will help us to understand the distribution along the individual columns.

1. Apps

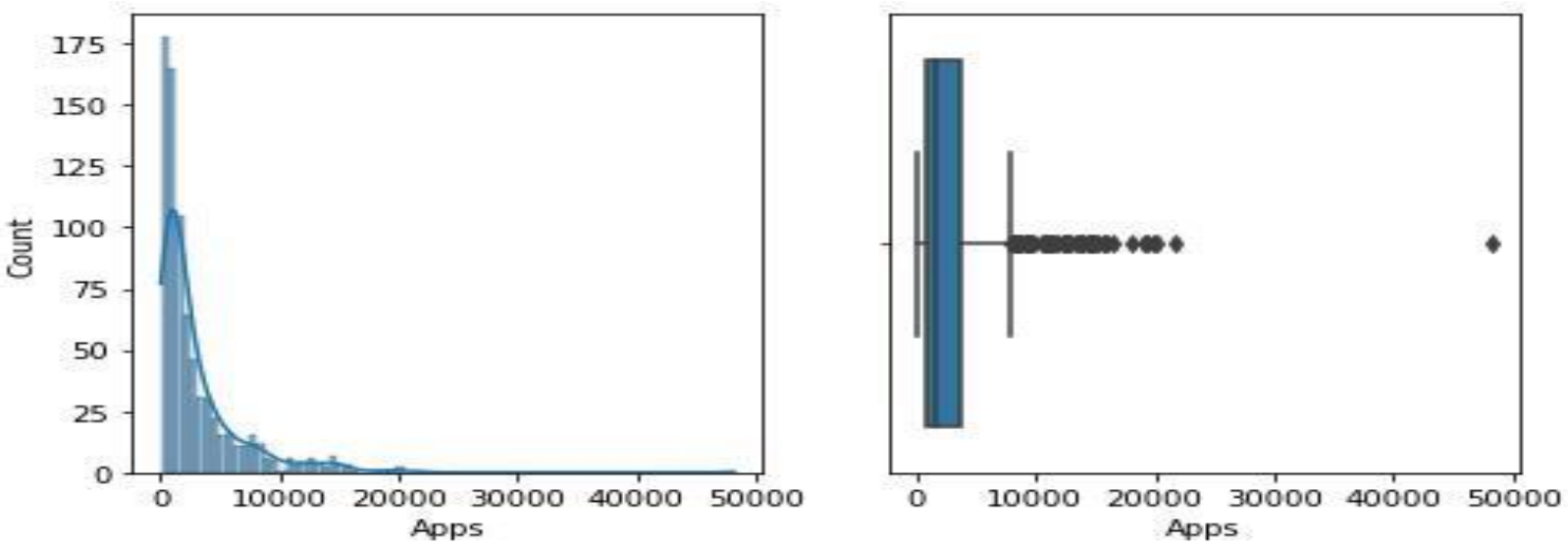

Fig. no.1: histogram and boxplot for Apps

From above histogram and boxplot, the data point are right skewed and having outliers in the column.

2. Accept

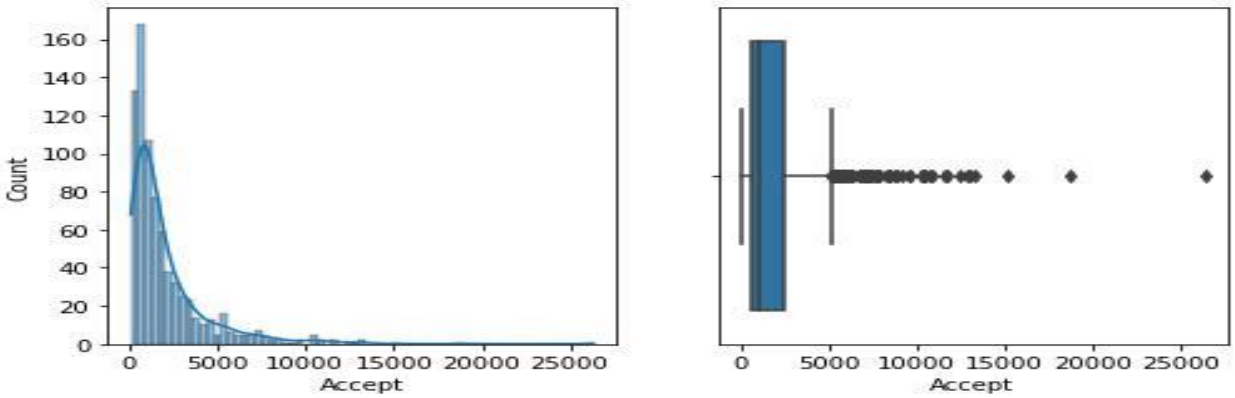

Fig. no.2 : histogram and boxplot for Accept

From above histogram and boxplot, the data point are right skewed and having outliers in the column.

3. Enroll

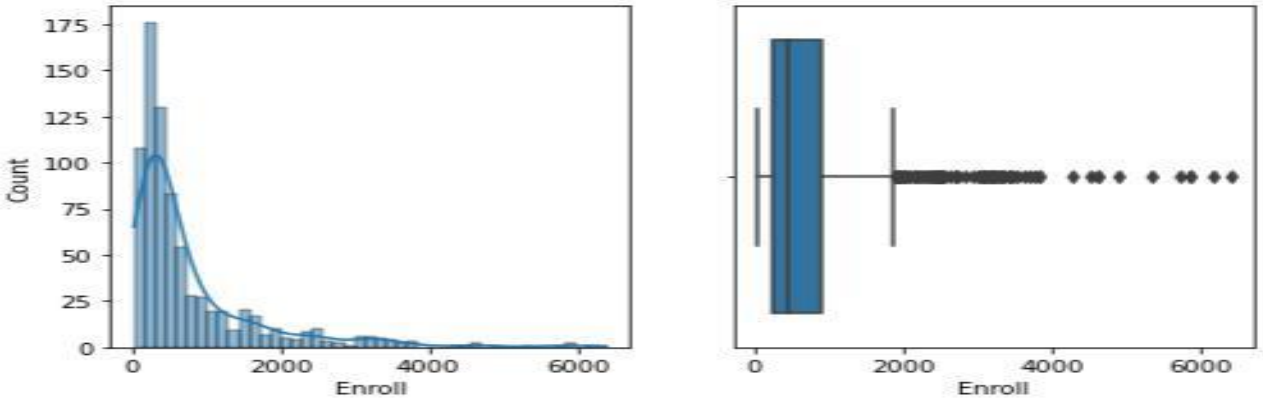

Fig. no.3: histogram and boxplot for Enroll

From above histogram and boxplot, the data point are right skewed and having outliers in the column.

4. Top10perc

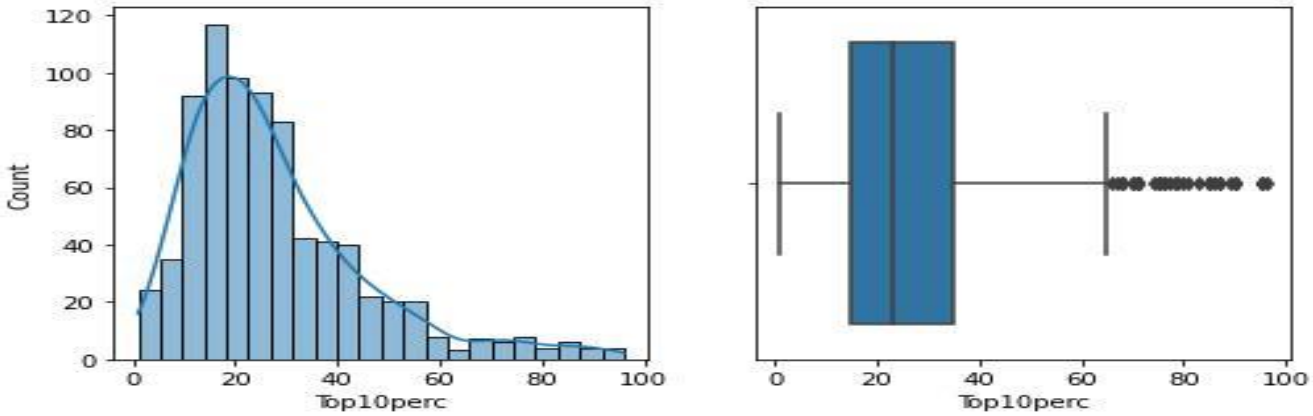

Fig. no.4: histogram and boxplot for Top10perc

From above histogram and boxplot, the data point are right skewed and having outliers in the column.

5. Top25perc

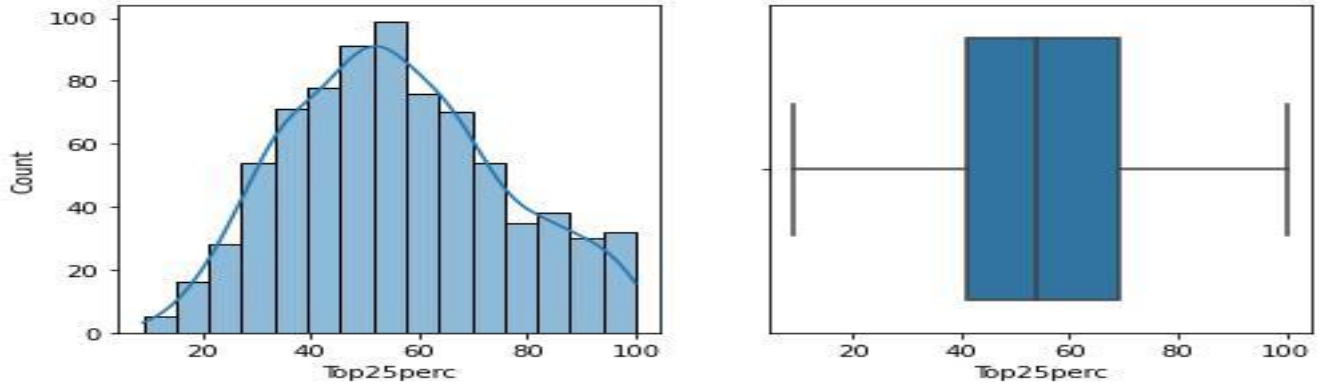

Fig. no.5 histogram and boxplot for Top25perc

From above histogram and boxplot, the data point are normally distributed and having no outliers in the column.

6. F. Undergrad

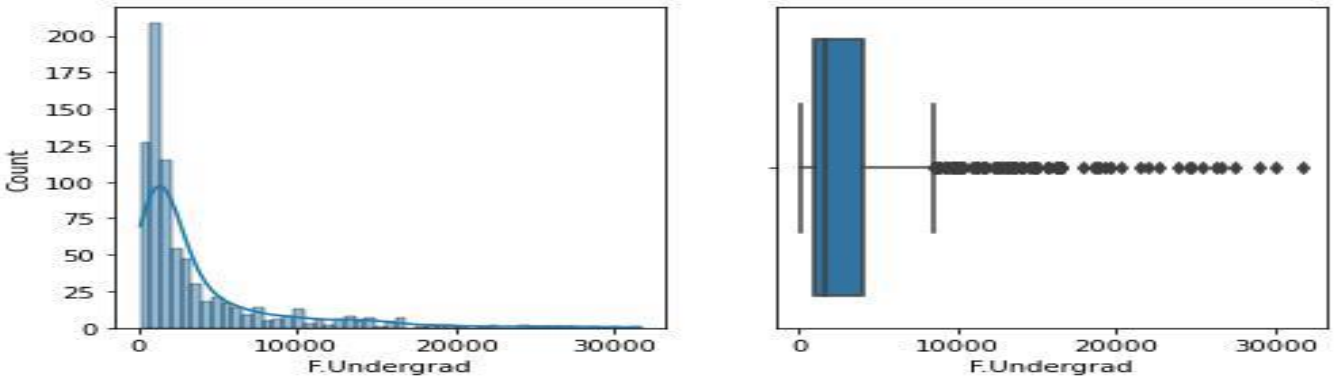

Fig. no.6 : histogram and boxplot for F.Undergrad

From above histogram and boxplot, the data point are right skewed and having outliers in the column.

7. P.Undergrad

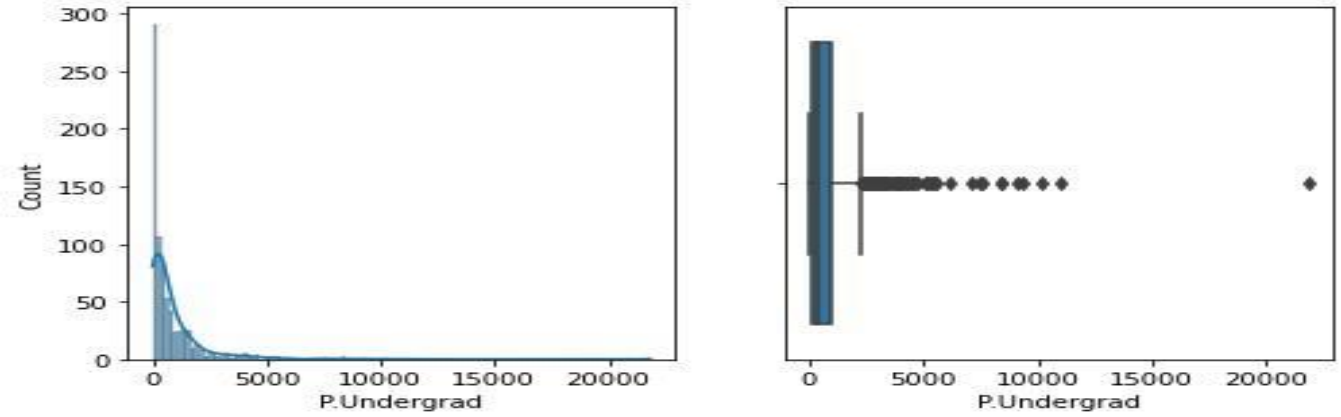

Fig. no.7: histogram and boxplot for P.Undergrad

From above histogram and boxplot, the data point are right skewed and having outliers in the column.

8. Outstate

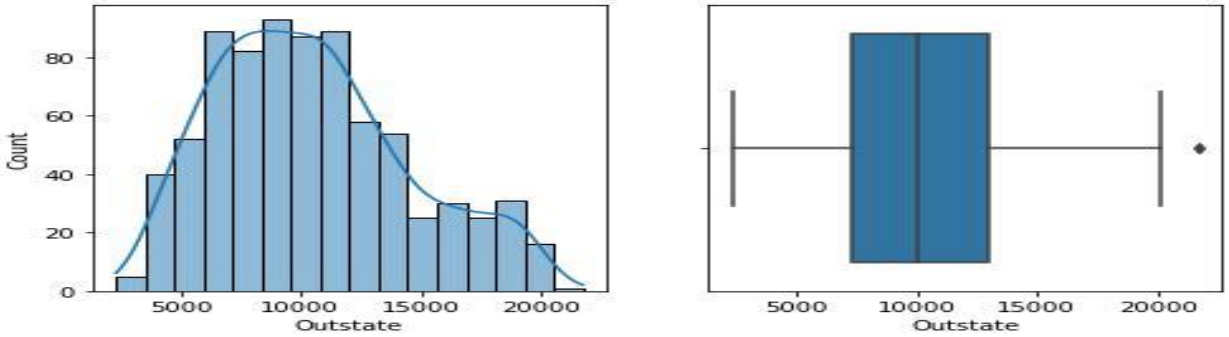

Fig. no.8: histogram and boxplot for Outstate

From above histogram and boxplot, the data point are normally distributed and having outliers in the column.

9. Room.Board

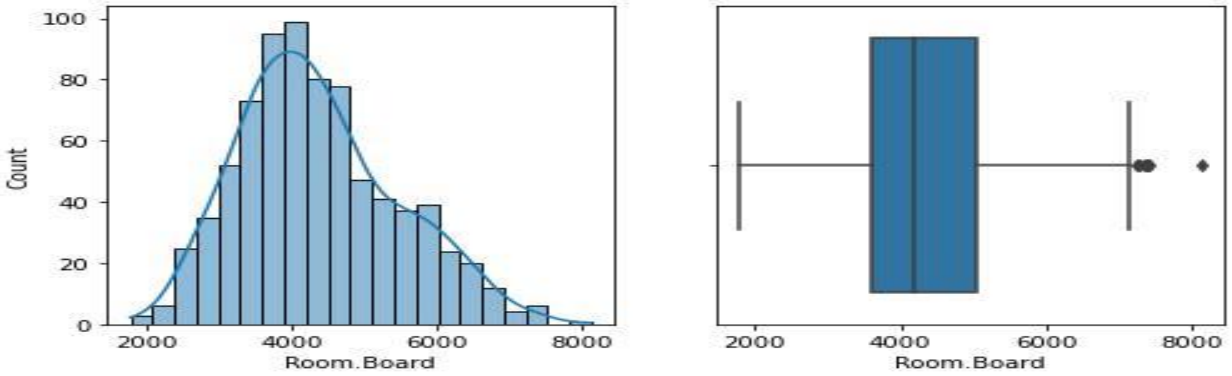

Fig. no.9: histogram and boxplot for Room.Board

From above histogram and boxplot, the data point are normally distributed and having outliers in the column.

10. Books

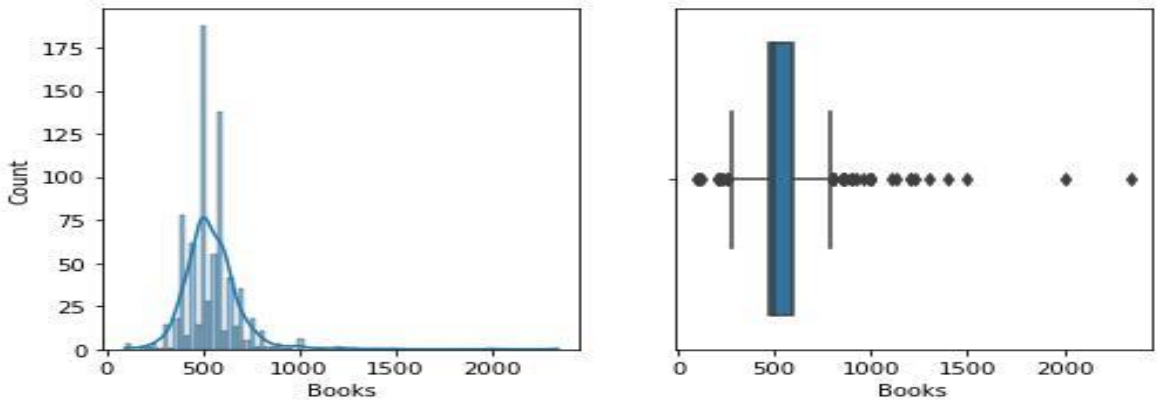

Fig. no.10 : histogram and boxplot for Books

From above histogram and boxplot, the data point are normally distributed and having outliers in the column.

11.Personal

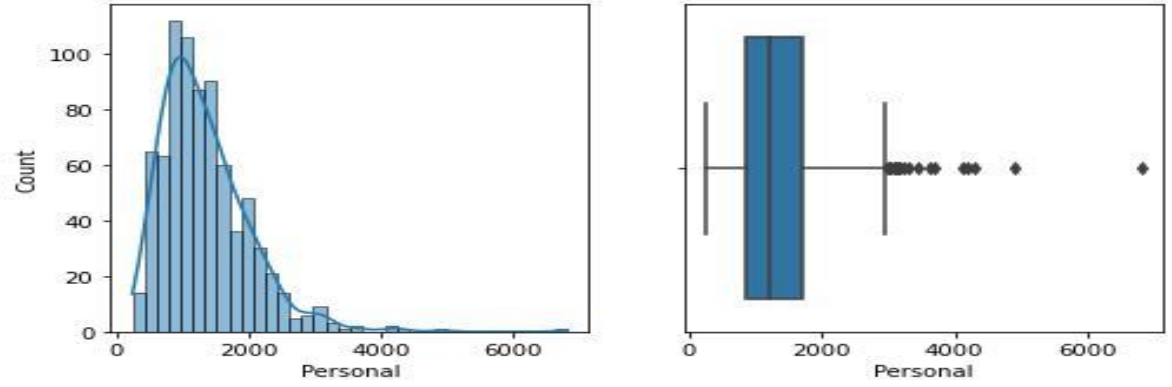

Fig. no.11 : histogram and boxplot for Personal

From above histogram and boxplot, the data point are right skewed and having outliers in the column.

12. PhD

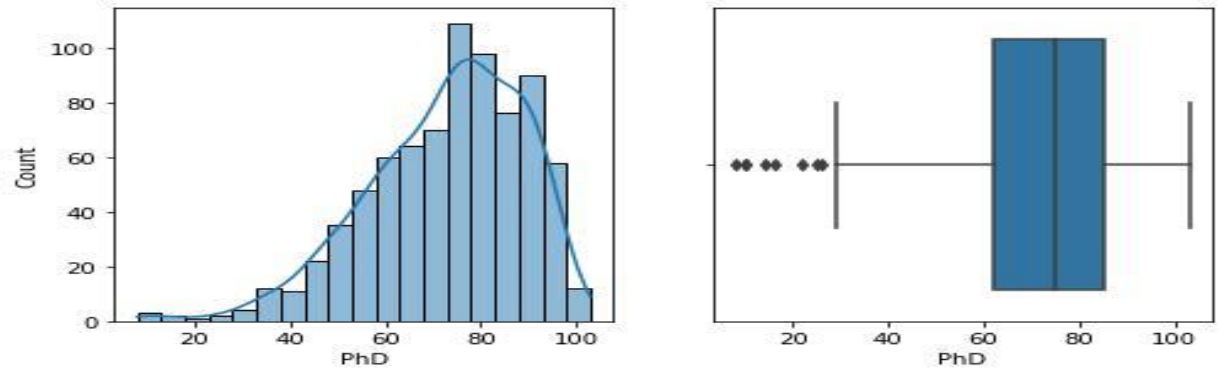

Fig. no.12 : histogram and boxplot for PhD

From above histogram and boxplot, the data point are left skewed and having outliers in the column.

13. Terminal

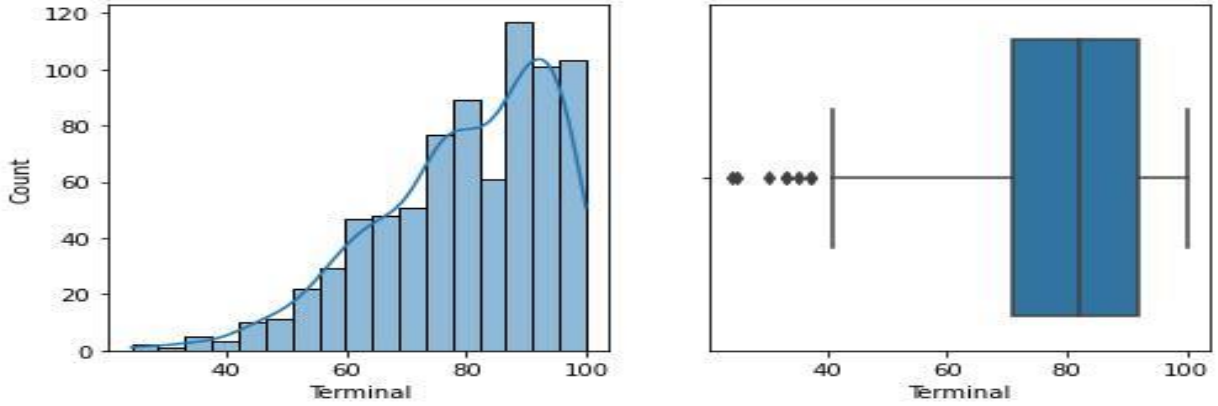

Fig. no.13 : histogram and boxplot for Terminal

From above histogram and boxplot, the data point are left skewed and having outliers in the column.

14. S.F.Ratio

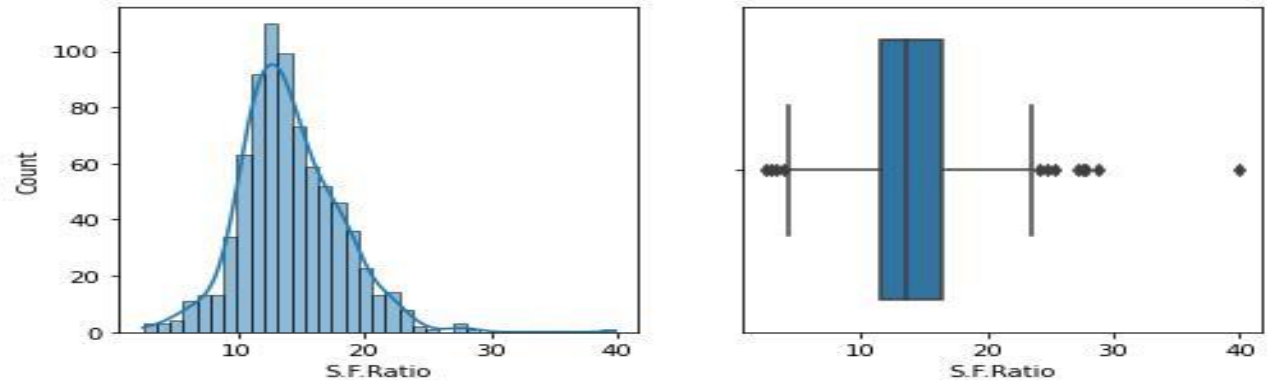

Fig. no.14: histogram and boxplot for S.F.Ratio

From above histogram and boxplot, the data point are left normally distributed and having outliers in the column.

15. perc. alumni

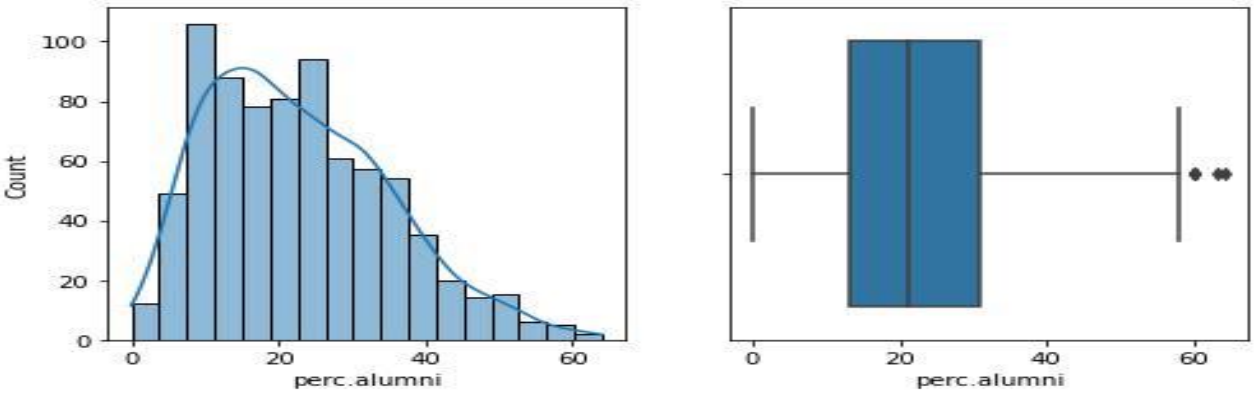

Fig. no.15: histogram and boxplot for perc.alumni

From above histogram and boxplot, the data point are right skewed and having outliers in the column.

16. Expend

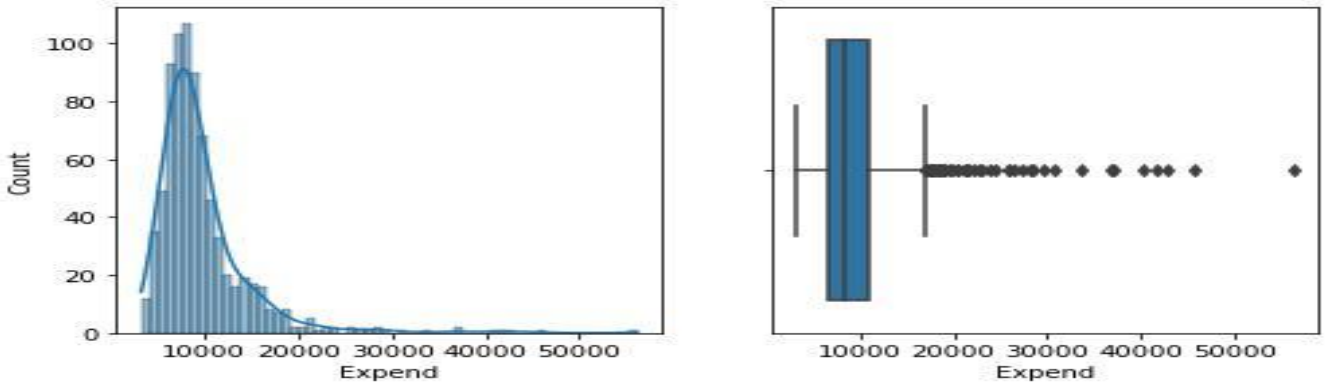

Fig. no.16: histogram and boxplot for Expend

From above histogram and boxplot, the data point are right skewed and having outliers in the column.

17. Grad. Rate

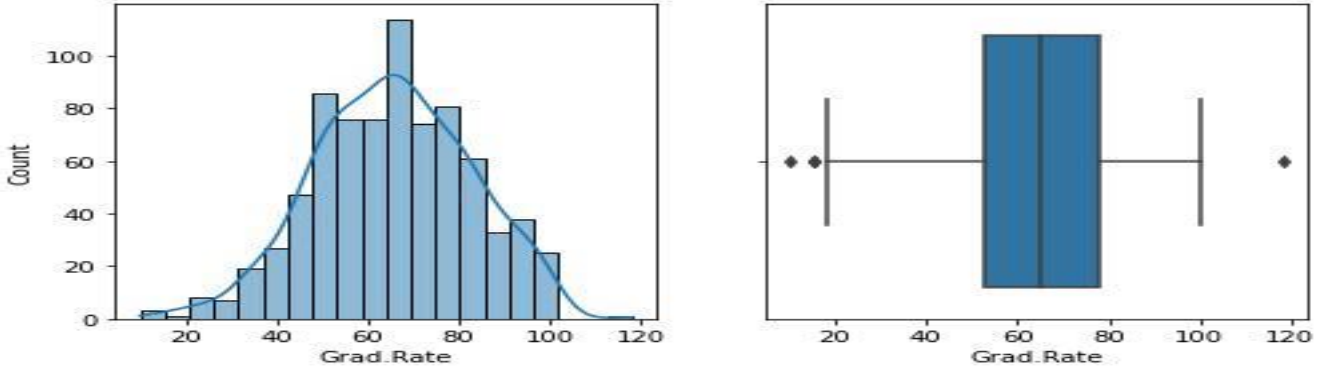

Fig. no.17: histogram and boxplot for Grad. Rate

From above histogram and boxplot, the data point are normally distributed and having outliers in the column.

Multivariate Analysis

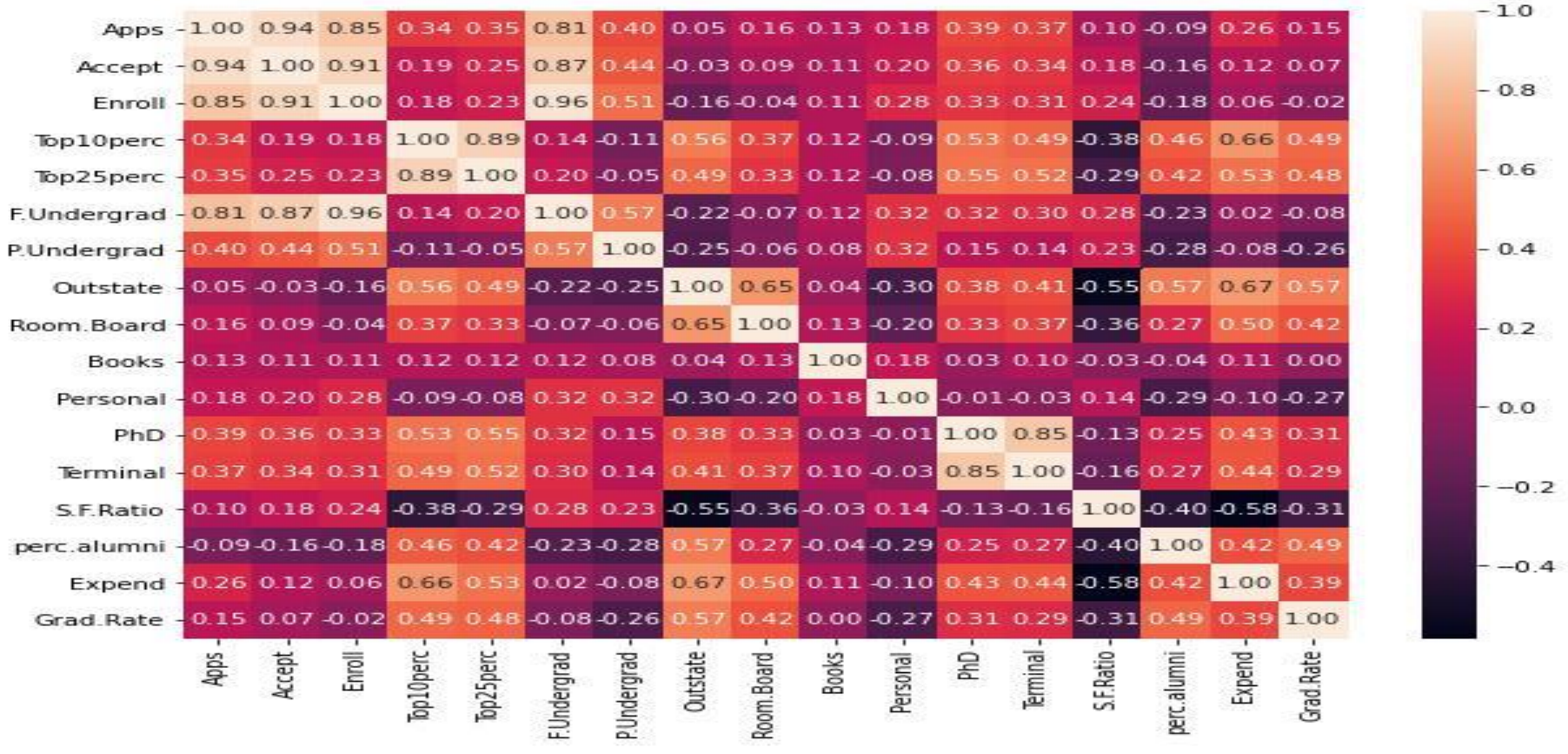

Fig.no 18: heat map for dataset

From the heat map, we can say that applications has high correlation with the app. Accepted, Enrolled and full-time graduate.

Student to faculty ratio is negatively correlated with the top 10perc, top 25perc, outstate and room. board

PhD and Terminal has good correlation between Top10perc and Top25perc

Grad.rate have nice of relation in outstate, room. Board, perc. alumni and expand.

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

Since the data has been scale in different manner for different column, we have to scale data. For particular this dataset we will go for z-score scaling. It will help to better model building and analysis.

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

Covariance matrix

[[1.00128866 0.94466636 0.84791332 0.33927032 0.35209304 0.81554018 0.3987775 0.05022367 0.16515151 0.13272942 0.17896117 0.39120081
0.36996762 0.09575627 -0.09034216 0.2599265 0.14694372]
[0.94466636 1.00128866 0.91281145 0.19269493 0.24779465 0.87534985 0.44183938 -0.02578774 0.09101577 0.11367165 0.20124767 0.35621633
0.3380184 0.17645611 -0.16019604 0.12487773 0.06739929]
[0.84791332 0.91281145 1.00128866 0.18152715 0.2270373 0.96588274 0.51372977 -0.1556777 -0.04028353 0.11285614 0.28129148 0.33189629
0.30867133 0.23757707 -0.18102711 0.06425192 -0.02236983]
[0.33927032 0.19269493 0.18152715 1.00128866 0.89314445 0.1414708 -0.10549205 0.5630552 0.37195909 0.1190116 -0.09343665 0.53251337
0.49176793 -0.38537048 0.45607223 0.6617651 0.49562711]
[0.35209304 0.24779465 0.2270373 0.89314445 1.00128866 0.19970167 -0.05364569 0.49002449 0.33191707 0.115676 -0.08091441 0.54656564
0.52542506 -0.29500852 0.41840277 0.52812713 0.47789622]
[0.81554018 0.87534985 0.96588274 0.1414708 0.19970167 1.00128866 0.57124738 -0.21602002 -0.06897917 0.11569867 0.31760831 0.3187472
0.30040557 0.28006379 -0.22975792 0.01867565 -0.07887464]
[0.3987775 0.44183938 0.51372977 -0.10549205 -0.05364569 0.57124738 1.00128866 -0.25383901 -0.06140453 0.08130416 0.32029384 0.14930637
0.14208644 0.23283016 -0.28115421 -0.08367612 -0.25733218]
[0.05022367 -0.02578774 -0.1556777 0.5630552 0.49002449 -0.21602002 -0.25383901 1.00128866 0.65509951 0.03890494 -0.29947232 0.38347594
0.40850895 -0.55553625 0.56699214 0.6736456 0.57202613]
[0.16515151 0.09101577 -0.04028353 0.37195909 0.33191707 -0.06897917 -0.06140453 0.65509951 1.00128866 0.12812787 -0.19968518 0.32962651
0.3750222 -0.36309504 0.27271444 0.50238599 0.42548915]
[0.13272942 0.11367165 0.11285614 0.1190116 0.115676 0.11569867 0.08130416 0.03890494 0.12812787 1.00128866 0.17952581 0.0269404
0.10008351 -0.03197042 -0.04025955 0.11255393 0.00106226]
[0.17896117 0.20124767 0.28129148 -0.09343665 -0.08091441 0.31760831 0.32029384 -0.29947232 -0.19968518 0.17952581 1.00128866 -0.01094989 -
0.03065256 0.13652054 -0.2863366 -0.09801804 -0.26969106]

[0.39120081 0.35621633 0.33189629 0.53251337 0.54656564 0.3187472 0.14930637 0.38347594 0.32962651 0.0269404 -0.01094989 1.00128866 0.85068186 -0.13069832 0.24932955 0.43331936 0.30543094]
[0.36996762 0.3380184 0.30867133 0.49176793 0.52542506 0.30040557 0.14208644 0.40850895 0.3750222 0.10008351 -0.03065256 0.85068186 1.00128866 -0.16031027 0.26747453 0.43936469 0.28990033]
[0.09575627 0.17645611 0.23757707 -0.38537048 -0.29500852 0.28006379 0.23283016 -0.55553625 -0.36309504 -0.03197042 0.13652054 -0.13069832 -0.16031027 1.00128866 -0.4034484 -0.5845844 -0.30710565]
[-0.09034216 -0.16019604 -0.18102711 0.45607223 0.41840277 -0.22975792 -0.28115421 0.56699214 0.27271444 -0.04025955 -0.2863366 0.24932955 0.26747453 -0.4034484 1.00128866 0.41825001 0.49153016]
[0.2599265 0.12487773 0.06425192 0.6617651 0.52812713 0.01867565 -0.08367612 0.6736456 0.50238599 0.11255393 -0.09801804 0.43331936 0.43936469 -0.5845844 0.41825001 1.00128866 0.39084571]
[0.14694372 0.06739929 -0.02236983 0.49562711 0.47789622 -0.07887464 -0.25733218 0.57202613 0.42548915 0.00106226 -0.26969106 0.30543094 0.28990033 -0.30710565 0.49153016 0.39084571 1.00128866]]

Table no 7: Covariance Matrix

Both covariance and the correlation matrices provide the relation between the two variable It scale the data in the same units and it is simple pair plot in the numerical form.

From the covariance matrix and correlation matrix table we get the same value numerically

Formula :

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
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	0.095633	-0.090226	0.259592	0.146755
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	0.176229	-0.159990	0.124717	0.067313
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	0.237271	-0.180794	0.064169	-0.022341
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	-0.384875	0.455485	0.660913	0.494989
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	-0.294629	0.417864	0.527447	0.477281
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.318337	0.300019	0.279703	-0.229462	0.018652	-0.078773
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.081200	0.319882	0.149114	0.141904	0.232531	-0.280792	-0.083568	-0.257001
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.038855	-0.299087	0.382982	0.407983	-0.554821	0.566262	0.672779	0.571290
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.127963	-0.199428	0.329202	0.374540	-0.362628	0.272363	0.501739	0.424942
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.000000	0.179295	0.026906	0.099955	-0.031929	-0.040208	0.112409	0.001061
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.000000	-0.010936	-0.030613	0.136345	-0.285968	-0.097892	-0.269344
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1.000000	0.849587	-0.130530	0.249009	0.432762	0.305038
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.099955	-0.030613	0.849587	1.000000	-0.160104	0.267130	0.438799	0.289527
S.F.Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628	-0.031929	0.136345	-0.130530	-0.160104	1.000000	-0.402929	-0.583832	-0.306710
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009	0.267130	-0.402929	1.000000	0.417712	0.490898
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739	0.112409	-0.097892	0.432762	0.438799	-0.583832	0.417712	1.000000	0.390343
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.571290	0.424942	0.001061	-0.269344	0.305038	0.289527	-0.306710	0.490898	0.390343	1.000000

Table no 8: Correlation Matrix

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

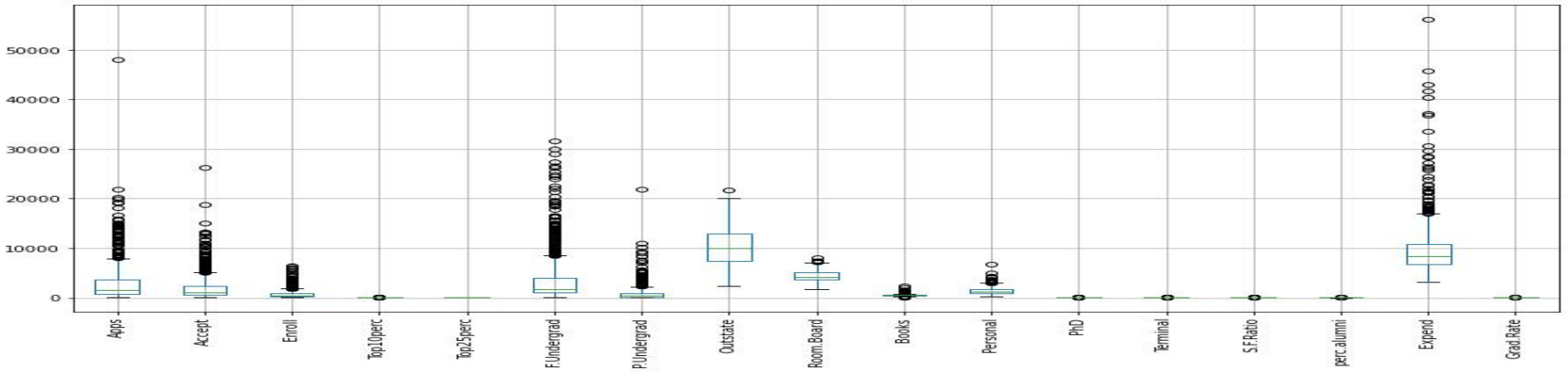

Fig no 19 : boxplot Before scale

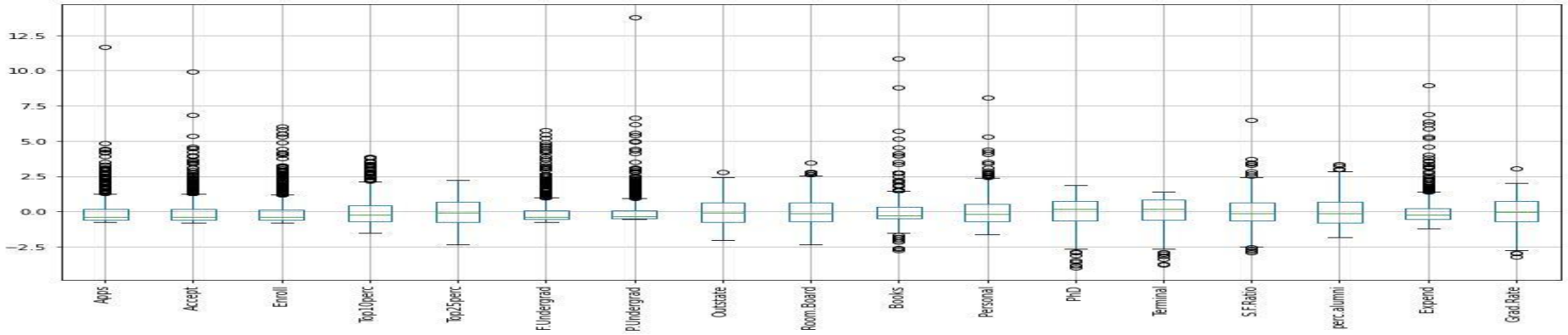

Fig no 20: Boxplot After scale

From the above figure, it is clearly seen that after scaling the boxplot is looking clear. But having outliers in the both plots because it only scale data in one unit not remove outliers.

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

```
array([[ 0.0929684 , 0.06592707, 0.03166929, 0.33452816, 0.36427546, 0.01149875, -0.04622402, 0.37830181,
0.29777508, 0.0401252 , -0.10944135, 0.31241468, 0.31563577, -0.238936 , 0.28566756, 0.24699418, 0.31117828],
[ 0.32104652, 0.3319699 , 0.35033549, 0.06754279, 0.13142781, 0.32452837, 0.20972858, -0.20665209, -0.07383062,
0.13395669, 0.29386253, 0.30728219, 0.28923834, 0.27738993, -0.26160327, -0.02920582, -0.12680266],
[ 0.06660652, 0.07883241, 0.01381154, -0.32328505, -0.41399578, 0.02193807, 0.1038968 , 0.2446006 , 0.65435548,
0.06932308, 0.02906196, 0.01051885, 0.07534077, -0.19726187, -0.35032544, 0.14253472, -0.14343185],
[-0.0129432 , -0.03420729, -0.01122623, 0.21141739, 0.19443136, -0.01744947, -0.02676078, 0.02000679, -0.07736692,
0.29066279, 0.60616613, -0.21336602, -0.22034156, -0.50833033, -0.05443422, 0.17754101, -0.26409767],
[ 0.24674827, 0.22877472, 0.19114851, 0.07741651, 0.11797053, 0.15029942, 0.06989958, 0.04640648, 0.20589468,
0.05440207, -0.01326297, -0.44256735, -0.48498252, 0.128203 , -0.08731305, -0.0657708 , 0.54379453],
[ 0.00650339, 0.02487384, 0.03094669, -0.32096376, -0.37686172, 0.01698553, 0.00474311, 0.05172404, -0.0558972 ,
0.08091173, 0.52880595, 0.07682269, 0.10434108, 0.06989583, 0.56468516, -0.05988252, 0.34192987], [-0.2400326 , -
0.27288698, -0.26523924, 0.09974811, 0.18713202, -0.21021832, -0.08937877, -0.05471221, 0.31550426, 0.50411092,
0.19661216, 0.04689621, 0.06899073, 0.48926256, -0.1645982 , -0.1378192 , 0.11257333],
[ 0.13180129, 0.12917757, 0.12023338, -0.02170441, -0.01531995, 0.10021068, 0.0537958 , 0.02283038, 0.11989134,
0.48393685, -0.33142515, -0.18837226, -0.09780965, 0.16012666, 0.52854784, 0.05801938, -0.47485834],
[ 0.01773119, 0.0195307 , 0.00760842, -0.13379303, -0.18486533, 0.01175077, -0.04181848, -0.17673816, -0.33073591,
0.6116081 , -0.31893212, 0.0983305 , 0.12229609, -0.36247736, -0.21127982, 0.02433122, 0.35725708],
[ 0.03400089, 0.06133927, 0.00639234, 0.00700079, -0.12749803, -0.02574847, -0.13366948, 0.75697305, -0.42976697,
0.10870293, 0.03384734, -0.02010444, -0.06046494, 0.31211523, -0.21023249, 0.20131984, -0.05772333], [-0.03784437,
-0.00907891, 0.01188448, -0.08516089, 0.13131824, 0.01666731, 0.08576745, 0.0571668 , -0.05833278, - 0.0378997 ,
0.01497856, -0.70146927, 0.67932403, -0.01040467, -0.03469458, -0.04429777, 0.03267103], [-0.22445438, -
0.17576693, -0.04255122, -0.14064341, 0.18030241, 0.06116207, 0.8706629 , 0.21583357, -0.09487917, 0.05038066, -
0.03722249, 0.09315117, -0.11394885, -0.06491485, 0.02561499, -0.12534642, 0.03283629],
[-0.11116423, -0.15058739, -0.01589244, 0.29511711, -0.29567976, 0.05081468, 0.24738219, -0.27224079, -0.01923638, -
0.04789243, -0.00703647, -0.10037346, 0.01716347, 0.20799672, 0.01145122, 0.76406889, 0.10963549],
[ 0.00914552, -0.00281244, -0.02985794, -0.69375696, 0.511526 , 0.01136756, -0.14019587, -0.09997777, -0.02476745, -
0.02826868, 0.0047994 , 0.0434578 , -0.08863951, 0.07241455, -0.00883861, 0.45737401, 0.02419997],
[ 0.5569348 , 0.26755069, -0.49315933, 0.00901027, -0.00267915, -0.55093616, 0.23891433, -0.07289772, -0.06810512, -
0.01670698, 0.02161283, -0.00454371, 0.02707478, 0.02278083, 0.01480452, 0.04412393, -0.01406045],
[ 0.59269472, -0.70710813, -0.13493584, -0.02159112, -0.01682207, 0.35140689, -0.03885668, 0.04554559, -0.01161555,
-0.00902118, -0.0047103 , -0.00548016, 0.00898686, -0.01073745, -0.00137953, -0.04641723, -0.00788293], [-
0.14346037, 0.32336365, -0.69930928, 0.0310452 , -0.00890279, 0.61814444, -0.04589764, 0.00367566, -0.00556533, -
0.00416292, 0.00630645, -0.00804 , -0.0048878 , -0.00662521, 0.01383698, -0.01193343, 0.00123347]])
```

Table no 9: Eigen vectors

Here are eigen vectors for the scaled data.

Eigen Values

Eigen values for scaled data are as follows:

```
array([4.75579369, 2.3800885 , 0.88497491, 0.81453646, 0.72423975, 0.52688069, 0.47958062,
0.41127635, 0.36620193, 0.23942458, 0.12943793, 0.09751277, 0.08189987, 0.06059116, 0.03582106,
0.01435481, 0.00793972])
```

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

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	0.09	0.32	0.07	-0.01	0.25	0.01	-0.24	0.13	0.02	0.03	-0.04	-0.22	-0.11	0.01	0.56	0.59	-0.14
Accept	0.07	0.33	0.08	-0.03	0.23	0.02	-0.27	0.13	0.02	0.06	-0.01	-0.18	-0.15	-0.00	0.27	-0.71	0.32
Enroll	0.03	0.35	0.01	-0.01	0.19	0.03	-0.27	0.12	0.01	0.01	0.01	-0.04	-0.02	-0.03	-0.49	-0.13	-0.70
Top10perc	0.33	0.07	-0.32	0.21	0.08	-0.32	0.10	-0.02	-0.13	0.01	-0.09	-0.14	0.30	-0.69	0.01	-0.02	0.03
Top25perc	0.36	0.13	-0.41	0.19	0.12	-0.38	0.19	-0.02	-0.18	-0.13	0.13	0.18	-0.30	0.51	-0.00	-0.02	-0.01
F.Undergrad	0.01	0.32	0.02	-0.02	0.15	0.02	-0.21	0.10	0.01	-0.03	0.02	0.06	0.05	0.01	-0.55	0.35	0.62
P.Undergrad	-0.05	0.21	0.10	-0.03	0.07	0.00	-0.09	0.05	-0.04	-0.13	0.09	0.87	0.25	-0.14	0.24	-0.04	-0.05
Outstate	0.38	-0.21	0.24	0.02	0.05	0.05	-0.05	0.02	-0.18	0.76	0.06	0.22	-0.27	-0.10	-0.07	0.05	0.00
Room.Board	0.30	-0.07	0.65	-0.08	0.21	-0.06	0.32	0.12	-0.33	-0.43	-0.06	-0.09	-0.02	-0.02	-0.07	-0.01	-0.01
Books	0.04	0.13	0.07	0.29	0.05	0.08	0.50	0.48	0.61	0.11	-0.04	0.05	-0.05	-0.03	-0.02	-0.01	-0.00
Personal	-0.11	0.29	0.03	0.61	-0.01	0.53	0.20	-0.33	-0.32	0.03	0.01	-0.04	-0.01	0.00	0.02	-0.00	0.01
PhD	0.31	0.31	0.01	-0.21	-0.44	0.08	0.05	-0.19	0.10	-0.02	-0.70	0.09	-0.10	0.04	-0.00	-0.01	-0.01
Terminal	0.32	0.29	0.08	-0.22	-0.48	0.10	0.07	-0.10	0.12	-0.06	0.68	-0.11	0.02	-0.09	0.03	0.01	-0.00
S.F.Ratio	-0.24	0.28	-0.20	-0.51	0.13	0.07	0.49	0.16	-0.36	0.31	-0.01	-0.06	0.21	0.07	0.02	-0.01	-0.01
perc.alumni	0.29	-0.26	-0.35	-0.05	-0.09	0.56	-0.16	0.53	-0.21	-0.21	-0.03	0.03	0.01	-0.01	0.01	-0.00	0.01
Expend	0.25	-0.03	0.14	0.18	-0.07	-0.06	-0.14	0.06	0.02	0.20	-0.04	-0.13	0.76	0.46	0.04	-0.05	-0.01
Grad.Rate	0.31	-0.13	-0.14	-0.26	0.54	0.34	0.11	-0.47	0.36	-0.06	0.03	0.03	0.11	0.02	-0.01	-0.01	0.00

Table no 10: Principal Component

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]

Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0.09	0.07	0.03	0.33	0.36	0.01	-0.05	0.38	0.30	0.04	-0.11	0.31	0.32	-0.24	0.29	0.25	0.31

The linear equation for 1st component :

PC1=0.09*Apps+0.07*Accept+0.03*Enroll+0.33*Top10perc+0.01*F.Undergrad
-0.05*P.Undergrad-0.38*Outstate+0.30*Room.Board+0.04*Books-0.11*Personal
+0.31*PhD+0.32*Terminal-0.24*S.F.Ratio+0.29*perc.alumni+0.25*Expend+0.31*Grad.Rate

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?

```
array([0.39596786, 0.59413427, 0.66781737, 0.73563576, 0.79593603, 0.83980417, 0.8797341 , 0.91397701,
0.94446702, 0.96440153, 0.97517855, 0.98329747, 0.99011646, 0.99516129, 0.99814376, 0.99933894, 1. ])
```

As the cumulative value are in ascending order, we can say that upto 80% values we will get 5 PC's.

From scree plot ,after 5th PC the graph is looking very flat

So we will go with 5 PC's to define final correlation

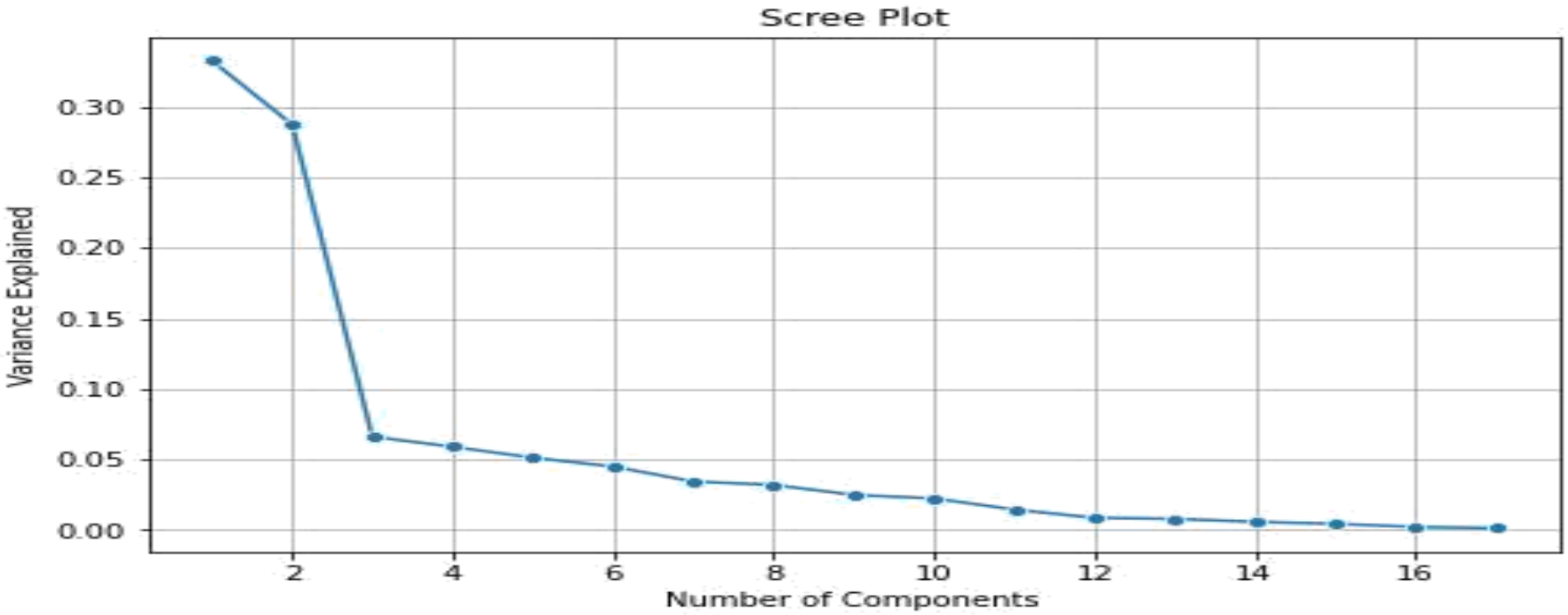

Fig no 21: Scree Plot

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]

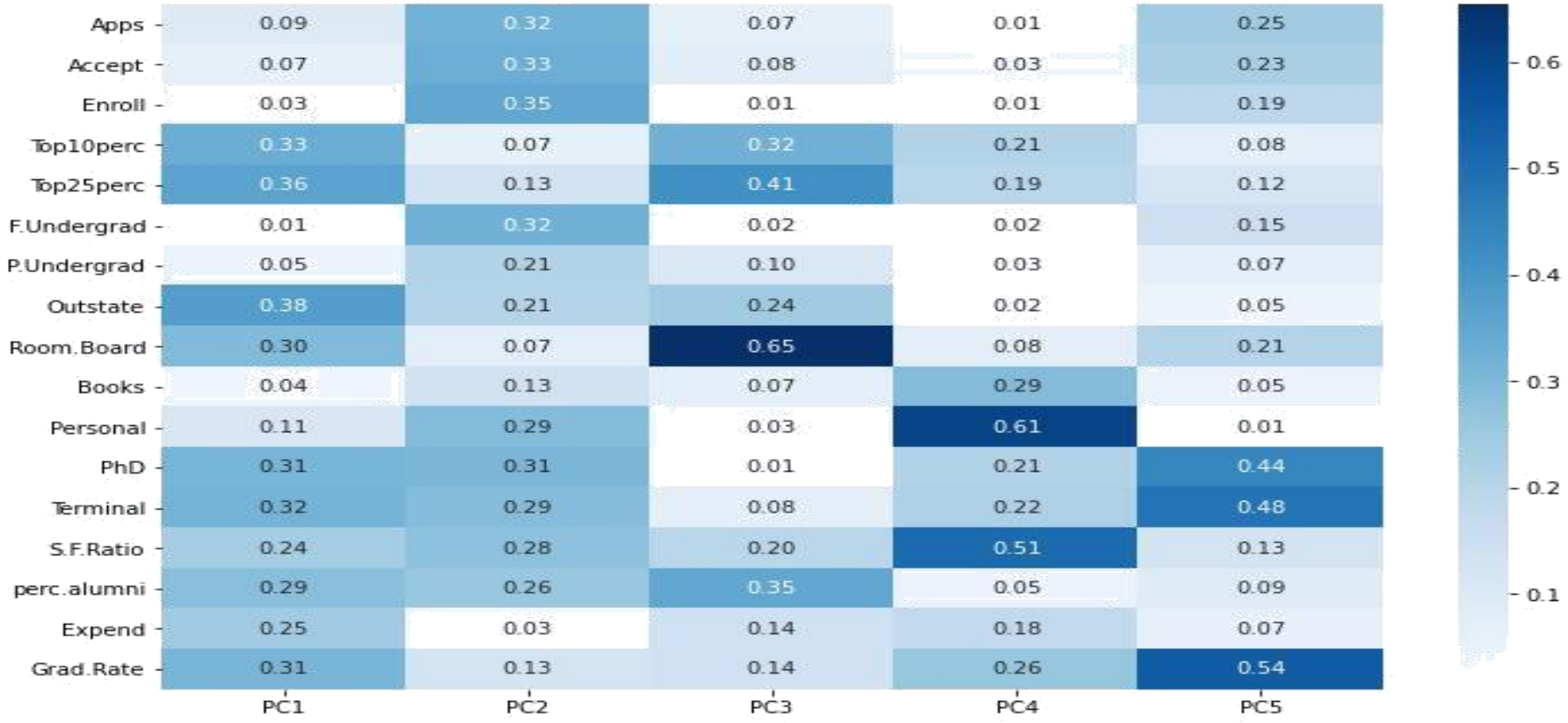

Fig. no 22: Heat map for 5 PC's

Now from the above heat map we can conclude that, we reduce 17 numerical variable to 5 principal components. We check the skewness of the data, remove outliers.

Univariate, multivariate and heat map among the different variable told us the various stories about the datapoint.

We also studied the impact of scaling on the datapoint of the variables.

So, this study will help us to build robust machine learning model.

The End……!