Project
Advanced Statistics

Megha R

PGP-DSBA Online March' 22

Date: 26/06/2022

Table of Contents

Contents

1 - Salary Data Analysis.....4
Problem 1.1.....4
Problem 1.2.....5
Problem 1.3.....5
Problem 1.4.....5
Problem 1.5.....6
Problem 1.6.....8
Problem 1.7.....9

2 - Salary Data Analysis.....11
Problem 2.1.....11
Problem 2.2.....41
Problem 2.3.....42
Problem 2.4.....43
Problem 2.5.....46
Problem 2.6.....44
Problem 2.7.....47
Problem 2.6.....44
Problem 2.7.....47
Problem 2.8.....48
Problem 2.9.....48

Problem 1:

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.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using one-way ANOVA and other parameters. The data consists of different Salary with educational qualifications and occupation. This project should help the student in exploring the summary statistics & one-way ANOVA.

Sample of the 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
5	Doctorate	Sales	219420
6	Doctorate	Sales	237920
7	Doctorate	Sales	160540
8	Doctorate	Sales	180934
9	Doctorate	Prof-specialty	248156

Exploratory Data Analysis

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

```
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Education   40 non-null     object
 1   Occupation  40 non-null     object
 2   Salary      40 non-null     int64
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

There are 40 entries of which 1 is integer data type and the rest 2 are of float.

Check for missing values in the dataset:

```
Education     0
Occupation    0
Salary        0
dtype: int64
```

From the above results we can see that there is no missing value present in the dataset.

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

Solution:

One way ANOVA (Education)

Null Hypothesis H0: The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, HS-Grad).

Alternate Hypothesis H1: The mean salary is different in at least one category of education.

One way ANOVA (Occupation)

Null Hypothesis H0: The mean salary is the same across all the 4 categories of occupation(Prof-Specialty, Sales, Adm-clerical, Exec-Managerial).

Alternate Hypothesis H1: The mean salary is different in at least one category of occupation.

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.

Solution:

	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

The above is the ANOVA table for Education variable.

Since the p value = 1.257709e-08 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 education.

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.

Solution:

	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

The above is the ANOVA table for Occupation variable.

Since the p value = 0.458508 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 4 categories of occupation.

1.4 If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result.

Solution:

Using, the Tukey Honest Significant Difference test, we get the following table for the category education:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Bachelors	Doctorate	43274.0667	0.0146	7541.1439	79006.9894	True
Bachelors	HS-grad	-90114.1556	0.001	-132035.1958	-48193.1153	True
Doctorate	HS-grad	-133388.2222	0.001	-174815.0876	-91961.3569	True

The table shows that since the p- values (p-adj in the table) are lesser than the significance level for all the three categories of education, this implies that the mean salaries across all categories of education are different.

Using, the Tukey Honest Significant Difference test, we get the following table for the category occupation:

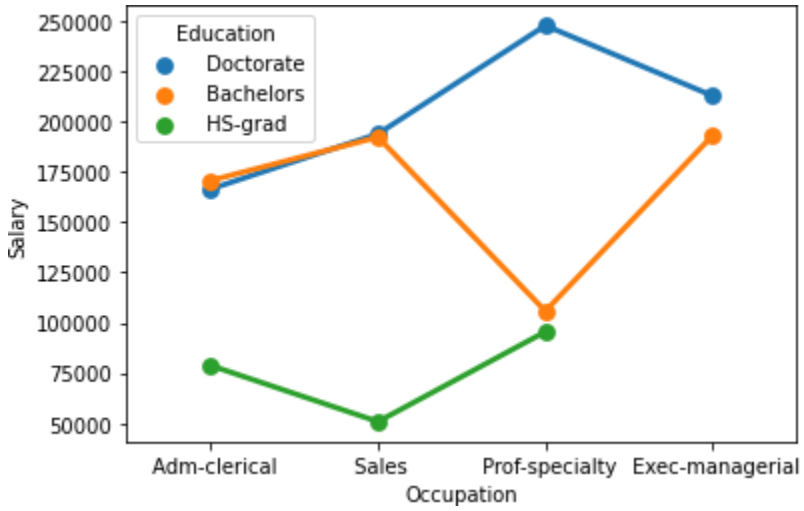
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Adm-clerical	Exec-managerial	55693.3	0.4146	-40415.1459	151801.7459	False
Adm-clerical	Prof-specialty	27528.8538	0.7252	-46277.4011	101335.1088	False
Adm-clerical	Sales	16180.1167	0.9	-58951.3115	91311.5449	False
Exec-managerial	Prof-specialty	-28164.4462	0.8263	-120502.4542	64173.5618	False
Exec-managerial	Sales	-39513.1833	0.6507	-132913.8041	53887.4374	False
Prof-specialty	Sales	-11348.7372	0.9	-81592.6398	58895.1655	False

For the category occupation, the Tukey Honest Significant Difference test has further confirmed that the mean salaries across all occupation classes are significantly same (as seen in solution 1.3). The table above confirms the same, wherein we see that all p-values are greater than 0.05.

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.

Solution:



The interaction plot shows that there is significant amount of interaction between the

categorical variables, Education and Occupation.

The following are some of the observations from the interaction plot:

- People with HS-grad education do not reach the position of Exec-managerial and they hold only Adm-clerk, Sales and Prof-Specialty occupations.
- People with education as Bachelors or Doctorate and occupation as Adm-clerical and Sales almost earn the same salaries (salaries ranging from 170000 – 190000).
- People with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupations as Adm-clerical and Sales.
- People with education as Bachelors and occupation Sales earn higher than people with education as Bachelors and occupation Prof-Specialty whereas people with education as Doctorate and occupation Sales earn lesser than people with Doctorate and occupation Prof-Specialty. We see a reversal in this part of the plot.
- Similarly, people with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupation Exec-Managerial whereas people with education as Doctorate and occupation as Prof-Specialty earn higher than people with education as Doctorate and occupation Exec-Managerial. There is a reversal in this part of the plot too.
- Salespeople with Bachelors or Doctorate education earn the same salaries and earn higher than people with education as HS-grad.
- Adm clerical people with education as HS-grad earn the lowest salaries when compared to people with education as Bachelors or Doctorate.
- Prof-Specialty people with education as Doctorate earn maximum salaries and people with education as HS-Grad earn the minimum.
- People with education as HS -Grad earn the minimum salaries.
- There are no people with education as HS -grad who hold Exec-managerial occupation.
- People with education as Bachelors and occupation, Sales and Exec-Managerial earn the same salaries.

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?

Solution:

H0: The effect of the independent variable 'education' on the mean 'salary' does not depend on the effect of the other independent variable 'occupation' (i. e. there is no interaction effect between the 2 independent variables, education and occupation).

H1: There is an interaction effect between the independent variable 'education' and the independent variable 'occupation' on the mean Salary.

By performing two way ANOVA, we get the following table:

	df	sum_sq	mean_sq	F	\
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	
Residual	29.0	2.062102e+10	7.110697e+08	NaN	

	PR(>F)
C(Education)	5.466264e-12
C(Occupation)	7.211580e-02
C(Education):C(Occupation)	2.232500e-05
Residual	NaN

From the table, we see that there is a significant amount of interaction between the variables, Education and Occupation.

As p value = 2.232500e-05 is lesser than the significance level (alpha = 0.05), we reject the null hypothesis and this implies that there is significant amount of interaction between the variables, Education and Occupation. Hence, at least one of the means of the salary variable with respect to each education category and occupation is unequal.

Thus, we can say that there is an interaction effect between education and occupation on the mean salary.

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

Solution:

From the interaction plot above, there is significant amount of interaction between the categorical variables, Education and Occupation.

The following are some of the observations from the interaction plot:

- People with HS-grad education do not reach the position of Exec-managerial and they hold only Adm-clerk, Sales and Prof-Specialty occupations.
- People with education as Bachelors or Doctorate and occupation as Adm-clerical and Sales almost earn the same salaries (salaries ranging from 170000 – 190000).
- People with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupations as Adm-clerical and Sales.
- People with education as Bachelors and occupation Sales earn higher than people with education as Bachelors and occupation Prof-Specialty whereas people with education as Doctorate and occupation Sales earn lesser than people with Doctorate and occupation Prof-Specialty. We see a reversal in this part of the plot.
- Similarly, people with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupation Exec-Managerial whereas people with education as Doctorate and occupation as Prof-Specialty earn higher than people with education as Doctorate and occupation Exec-Managerial. There is a reversal in this part of the plot too.
- Salespeople with Bachelors or Doctorate education earn the same salaries and earn higher than people with education as HS-grad.
- Adm clerical people with education as HS-grad earn the lowest salaries when compared to people with education as Bachelors or Doctorate.
- Prof-Specialty people with education as Doctorate earn maximum salaries and people with education as HS-Grad earn the minimum.
- People with education as HS -Grad earn the minimum salaries.
- There are no people with education as HS -grad who hold Exec-managerial occupation.
- People with education as Bachelors and occupation, Sales and Exec-Managerial earn the same salaries.

From the ANOVA method and the interaction plot, we see that education combined with occupation results in higher and better salaries among the people. It is clearly seen that people with education as Doctorate draw the maximum salaries and people with education HS-grad earn the least. Thus, we can conclude that Salary is dependent on educational qualifications and occupation.

Problem 2:

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.

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

Solution:

The following are some observations after initial exploration of the data:

- The dataset consists of 777 rows and 18 columns.
- The 'Names' field is an object data type, and all the remaining 17 fields are numeric fields.
- The field 'S.F. Ratio' is a float data type and the remaining 16 numeric fields are integer data type.
- There are no missing values in the data.
- Also, there are no bad data which is seen from the output of the 'info' command.
- There are no duplicate records in the data set.
- There are lots of outliers in the data as evident from the boxplots seen in Figure 1. (No outlier treatment is being done as per the instructions)

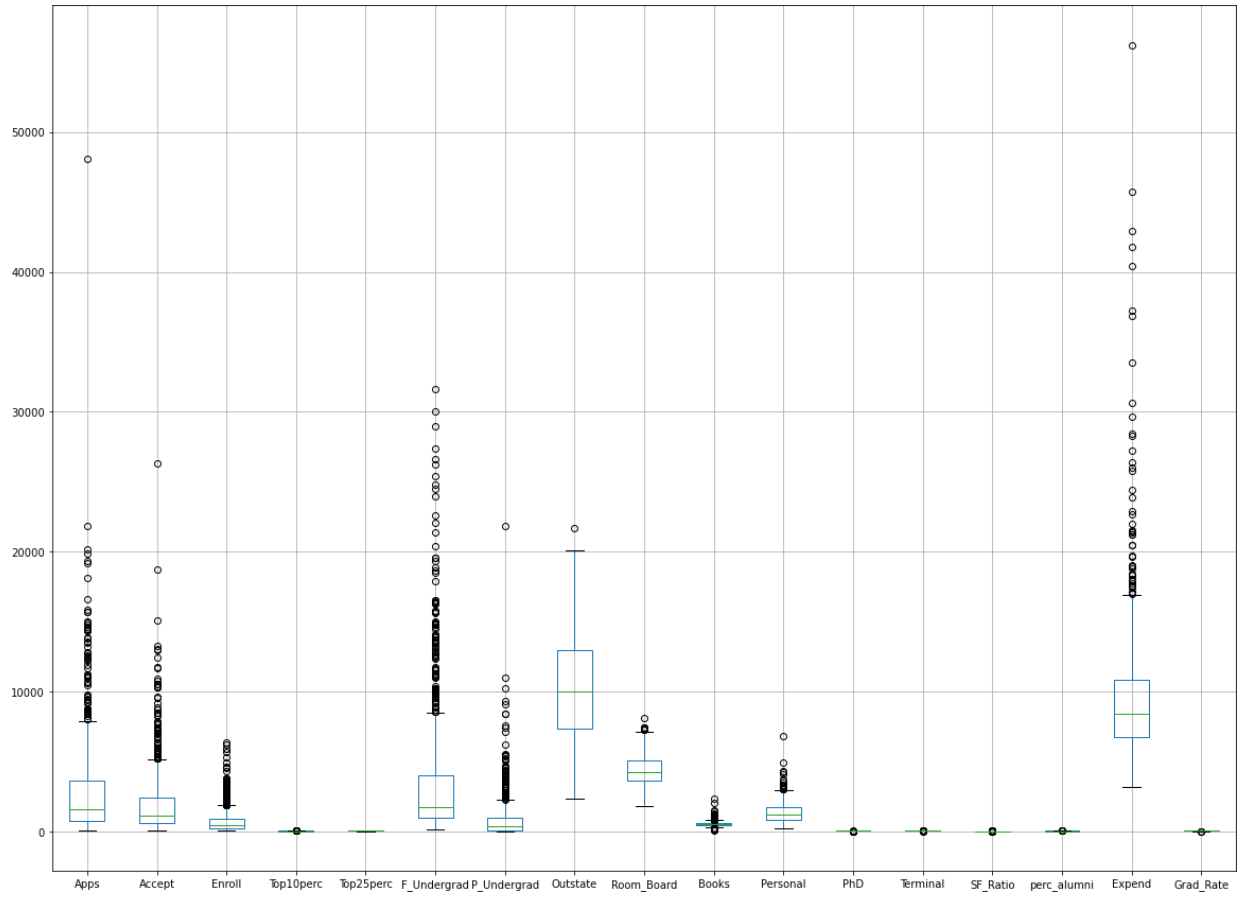


Figure 1- Boxplots for all the attributes in the data set

Treating Anomalies

- From the description of the data, we see us that there are a few anomalies in the data. The graduation rate Grad.Rate has a data with value 118 and percentage of faculty with PhDs have a value 103 which are anomalies as the upper limit can be only 100, being percentages. These two values have been imputed with the median.

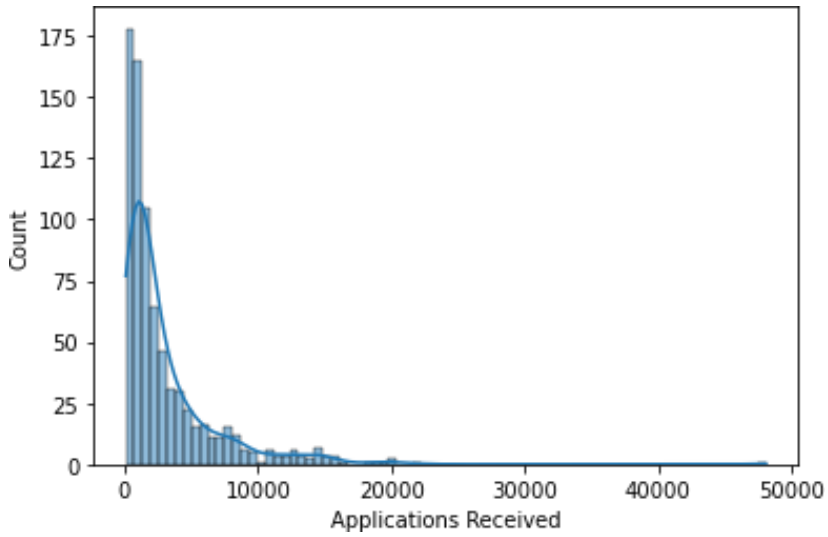
Data Visualization- Univariate Analysis

Insights from the Data

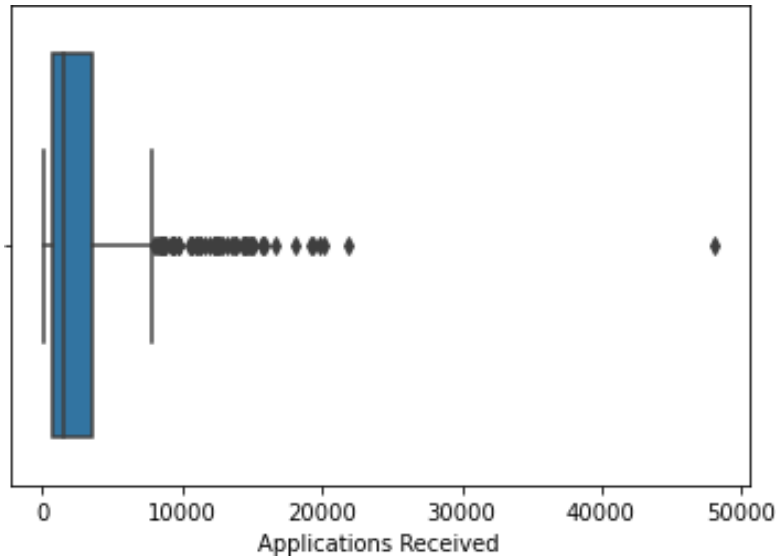
Apps: Number of applications a college/university receives

- The number of applications received by a college ranges from 81 to 48094.
- It is seen that Rutgers at New Brunswick receives the maximum number of applications (48094) and Christendom College receives the least (81).

- The mean number of applications is 3001 and the median is 1558.
- The mean being higher than the median indicates that the distribution is right skewed as seen in the below figure.



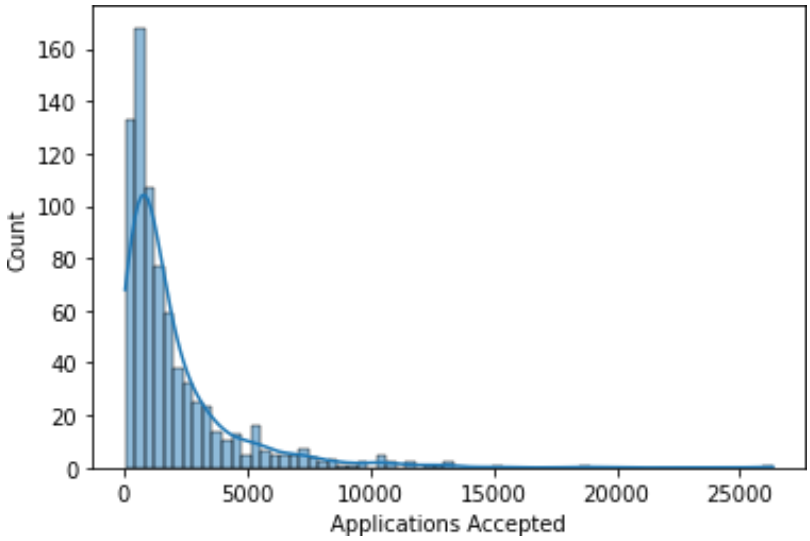
The below figure shows the presence of outliers in this attribute.



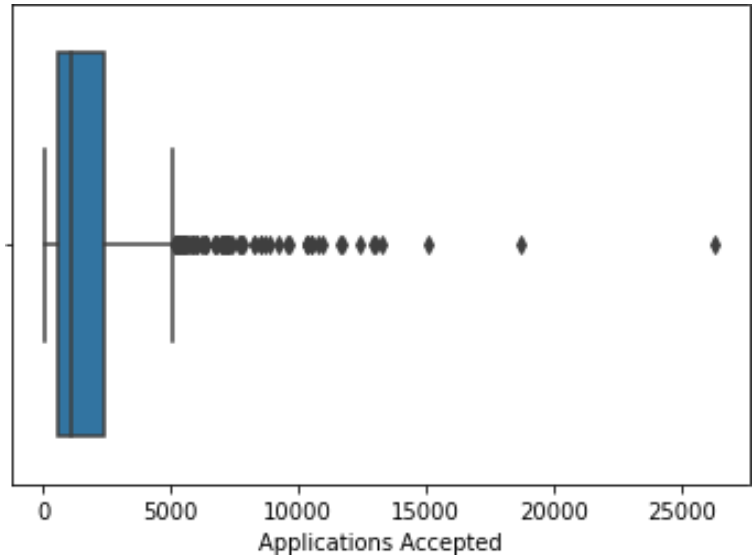
Accept: Number of applications a college/university accepts

- The number of applications accepted by a college ranges from 72 to 26330.
- It is seen that Rutgers at New Brunswick accepts the maximum number of applications
- (26330) and Christendom College accepts the least (79).
- The mean number of applications accepted is 2018 and the median is 1110.

- The mean being higher than the median indicates that the distribution is right skewed as seen in the below figure.

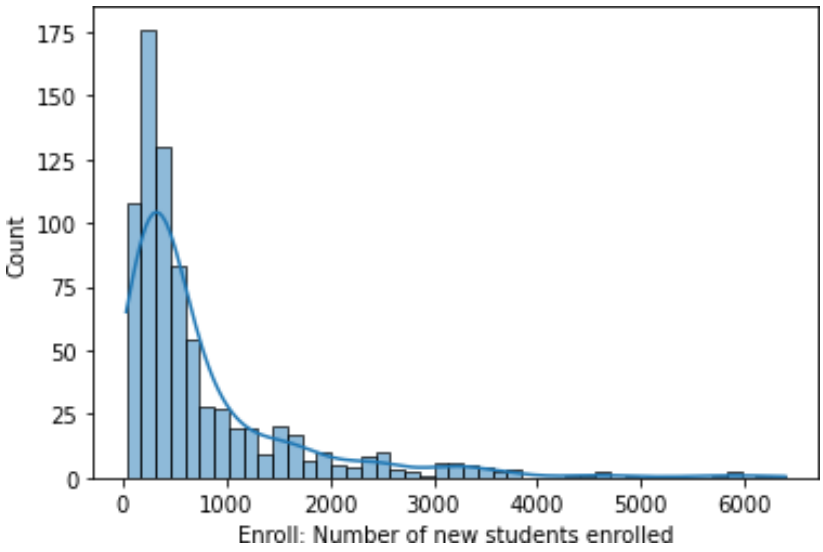


The below figure shows the presence of outliers in this attribute.

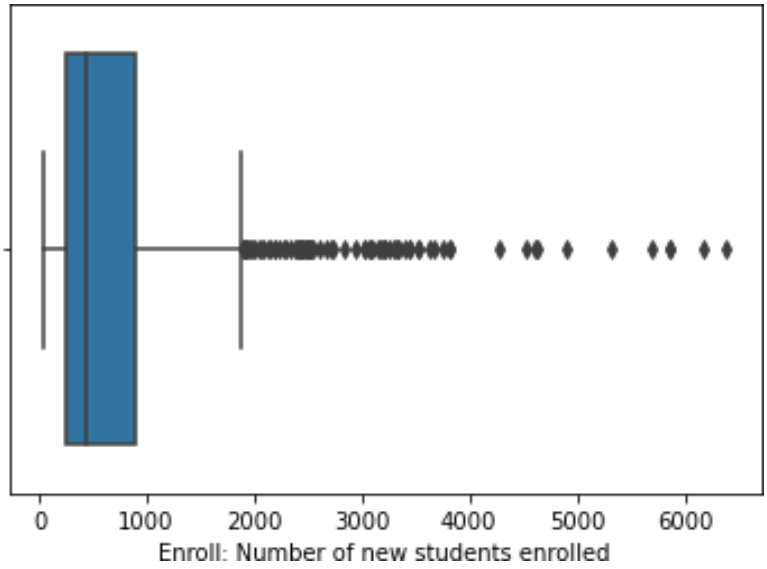


Enroll: Number of students who enroll

- The number of students who enroll ranges from 35 to 6392.
- Texas A&M Univ. at College Station has maximum enrollment with 6392 students and Capitol College has the least with 35 students.
- The mean number of students enrolled is 779.97 and the median is 434. The mean being higher than the median indicates that the distribution is right skewed as seen in the below figure.



The below boxplot shows the presence of outliers in this attribute.

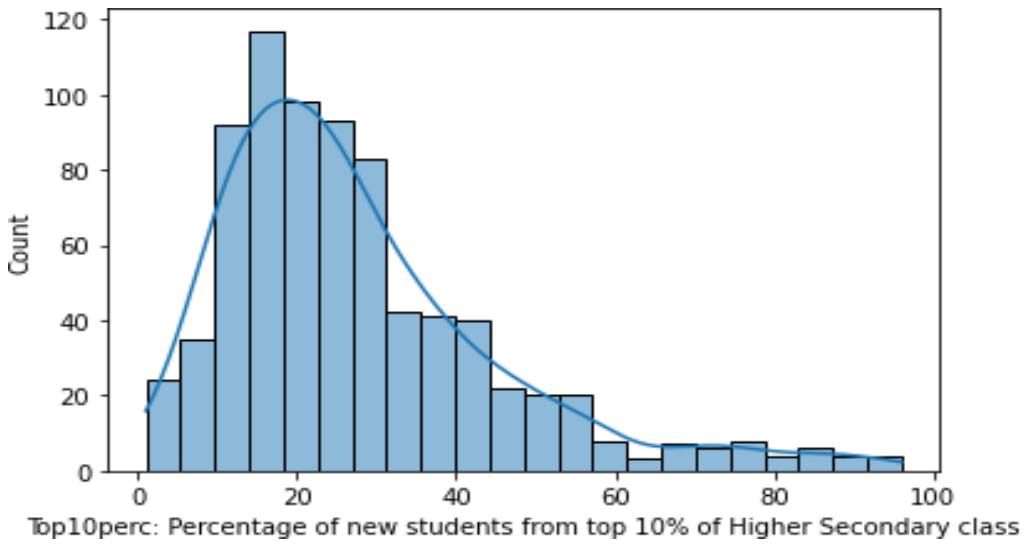


Top10perc : Percentage of new students from top 10% of Higher Secondary class

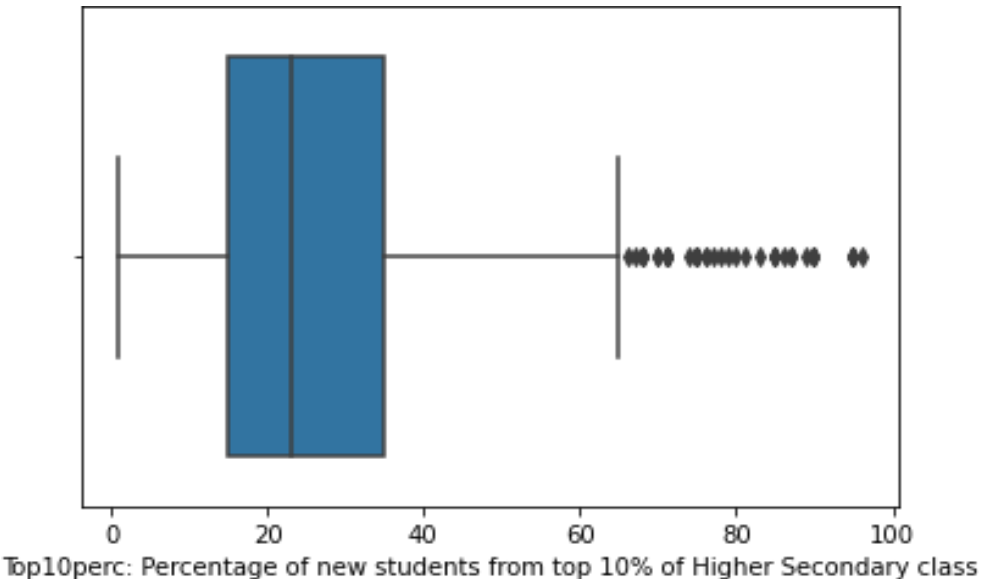
- The percentage of new students from top 10% of Higher Secondary class ranges from 1 to 96.
- Massachusetts Institute of Technology has maximum percentage of new students from top 10% of Higher Secondary class with 96% and three colleges as seen in the below table have the least with 1%.

```
University wise listing   Top10perc
-----------------------------------------------
Names
Massachusetts Institute of Technology    96
Harvey Mudd College                      95
University of California at Berkeley     95
Yale University                          95
Princeton University                     90
                                         ..
Morris College                            2
Virginia State University                 2
Fayetteville State University             1
North Adams State College                 1
Center for Creative Studies               1
Name: Top10perc, Length: 777, dtype: int64
```

- The mean percentage of new students from top 10% of Higher Secondary class is 27% and the median is 23%.
- The mean being higher than the median indicates that the distribution is right skewed as seen in the below figure.



The below boxplot shows the presence of outliers in this attribute.



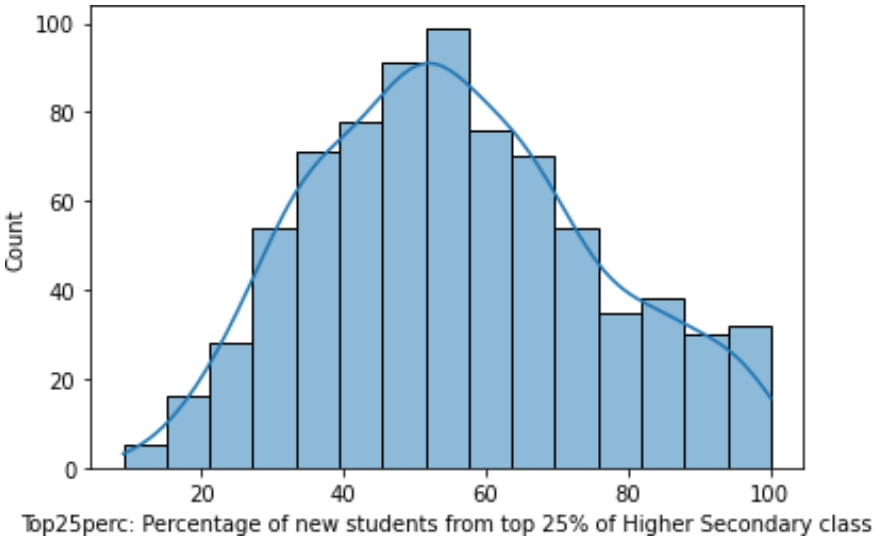
Top25perc: Percentage of new students from top 25% of Higher Secondary class

- The percentage of new students from top 25% of Higher Secondary class ranges from 9 to 100.
- There are 7 universities/colleges which have the maximum percentage (100%) of new students from top 25% of Higher Secondary class and Huron University has the least with 9% as can be seen below.

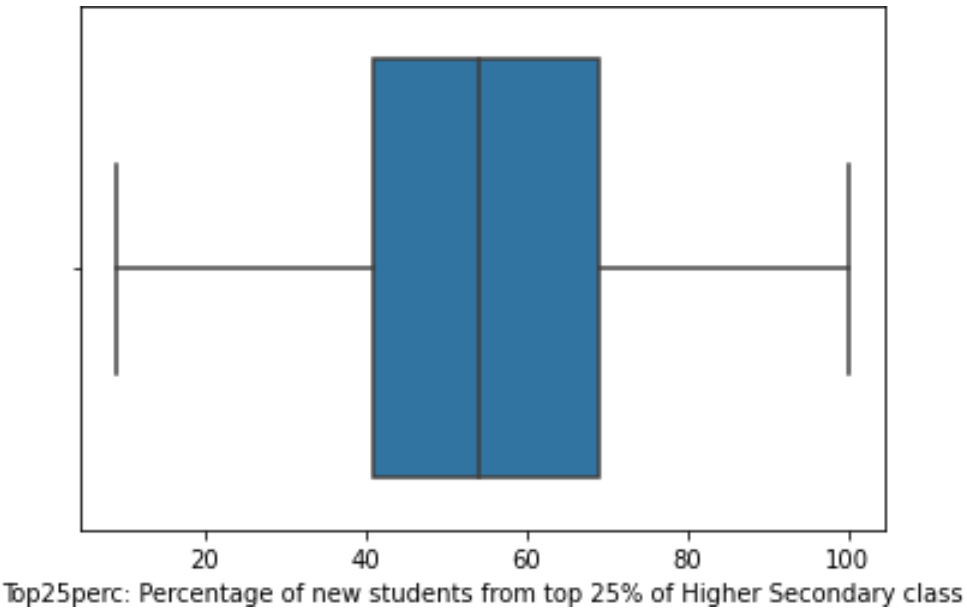
```
Names
University of California at Irvine      100
Bowdoin College                         100
Harvard University                      100
Harvey Mudd College                     100
University of Pennsylvania              100
SUNY at Buffalo                         100
University of California at Berkeley    100
Yale University                          99
Georgia Institute of Technology          99
Niagara University                       99
Name: Top25perc, dtype: int64
```

```
Names
North Adams State College          19
Dominican College of Blauvelt      19
Mesa State College                 18
St. Paul's College                 17
Fayetteville State University      16
Franklin Pierce College            14
Morris College                     13
Johnson State College              13
Mayville State University          12
Huron University                    9
Name: Top25perc, dtype: int64
```

- The mean percentage of new students from top 25% of Higher Secondary class is 55.7966% and the median is 54%.
- The mean is very close to the median indicating that the distribution is almost normal as seen in the below figure.

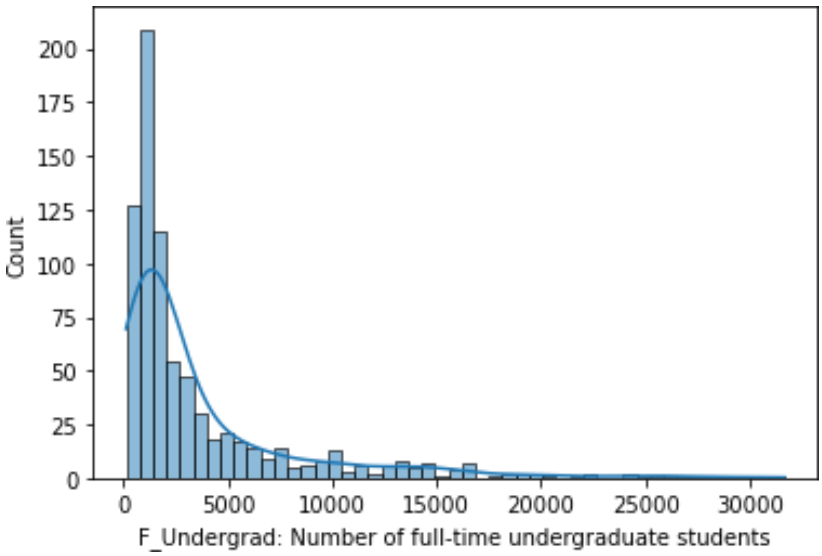


The attribute has no outliers as seen in the below boxplot.

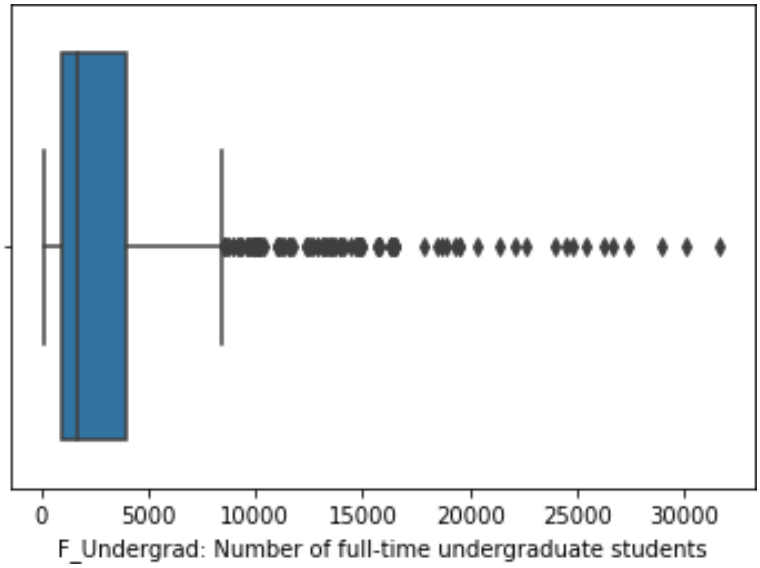


F_Undergrad: Number of full-time undergraduate students

- The number of full time under graduate students ranges from 139 to 31643.
- Texas A&M University at College Station has maximum number of full time undergraduate students with 31643 students and Christendom College has the least with 139 students.
- The mean number of full-time undergraduate students is 3699 and the median is 1707.
- The mean being higher than the median indicates that the distribution is right skewed as seen in the below figure.



The below boxplot shows the presence of outliers in this attribute.

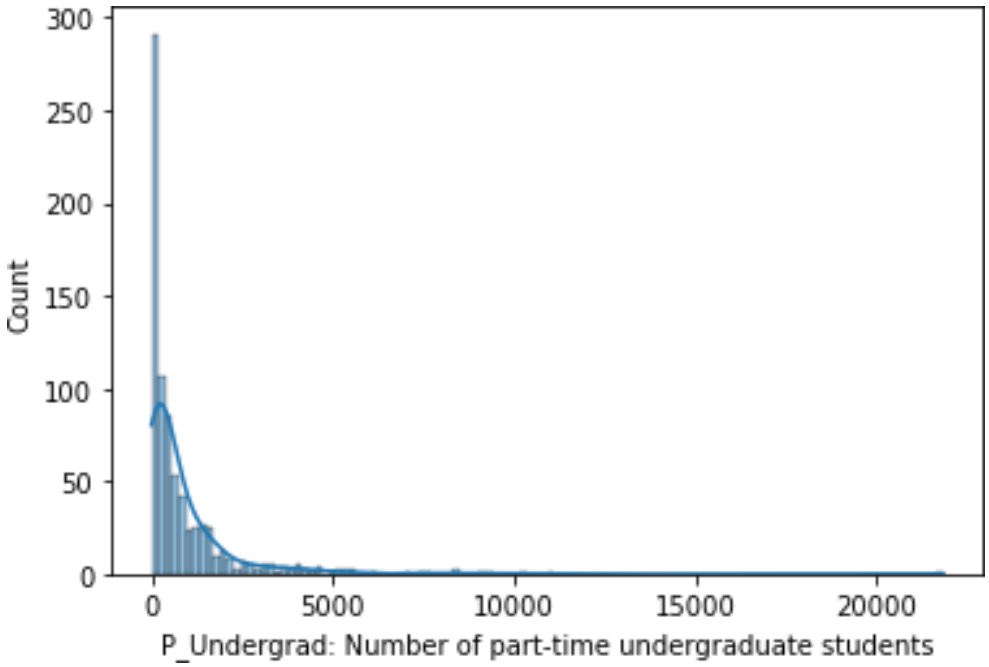


P_Undergrad: Number of part-time undergraduate students

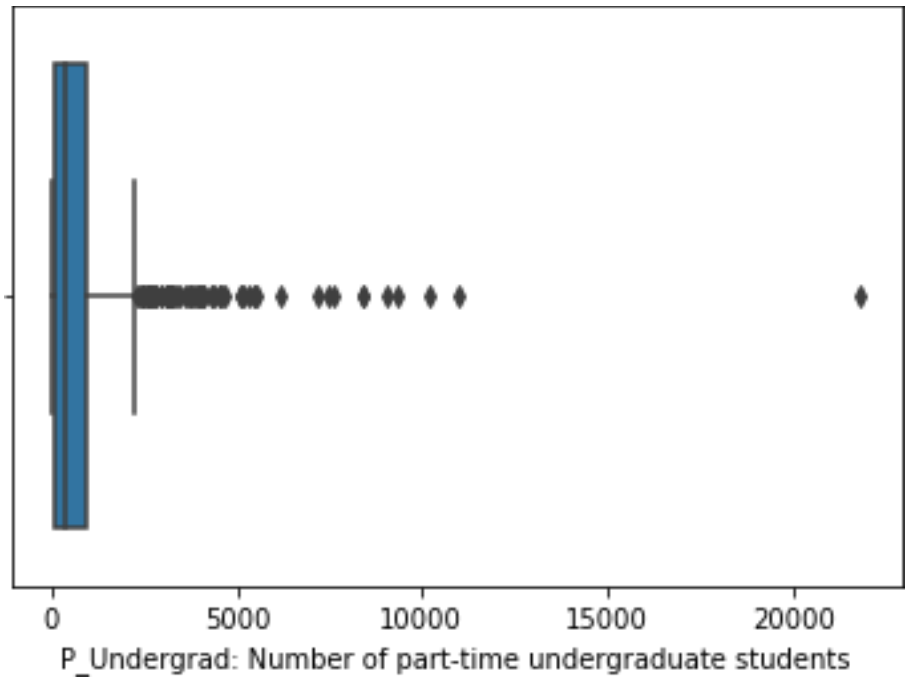
- The Number of part-time undergraduate students ranges from 1 to 21836.
- University of Minnesota Twin cities has the maximum number of part time undergraduate students with 21836 students and the below 4 colleges have the least with 1 student.

```
Names
University of Minnesota Twin Cities    21836
University of South Florida            10962
Northeastern University                10221
Florida International University        9310
Georgia State University                9054
                                        ...
Bennington College                         2
Claremont McKenna College                  1
Hampden - Sydney College                   1
Kenyon College                             1
College of Wooster                         1
Name: P_Undergrad, Length: 777, dtype: int64
```

- The mean number of part time undergraduate students is 855.30 and the median is 353.
- The mean being higher than the median indicates that the distribution is right skewed as seen in the below figure.

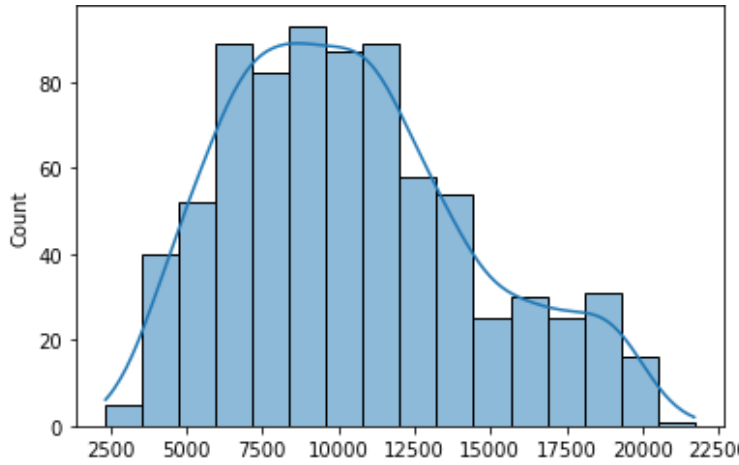


The below boxplot shows the outliers in the data**.**



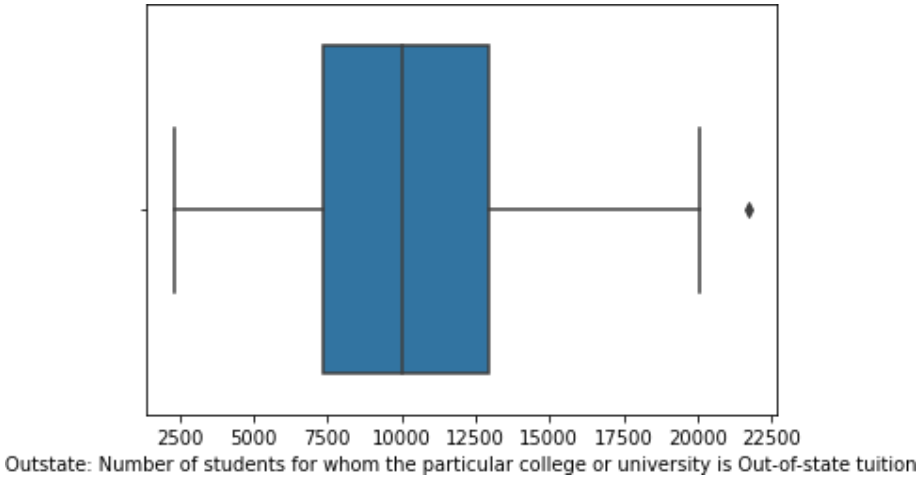
Outstate: Number of students for whom the particular college or university is Out-of-state tuition

- The number of students for whom the particular college or university is Out-of-state tuition ranges from 2340 to 21700.
- Bennington College has the maximum number of students for whom the particular college or university is Out-of-state tuition with 21700 students and Brigham Young University at Provo has the least with 2340 students.
- The mean number of students for whom the particular college or university is Out-of-state tuition is 10440.67 and the median is 9990.
- The mean is quite close to the median indicating that the distribution is almost normal as seen in the below figure.



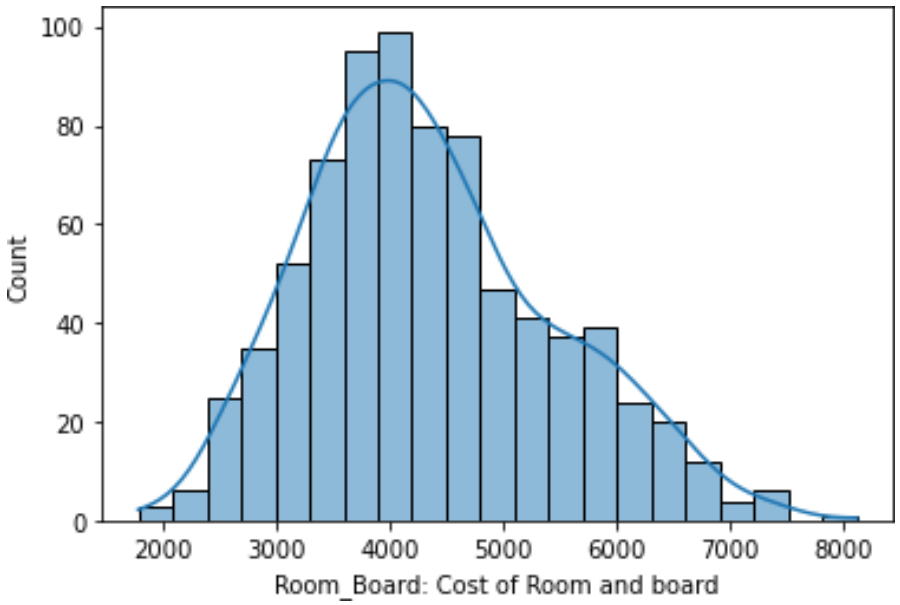
.

This is also evident from the below boxplot which has only one outlier.

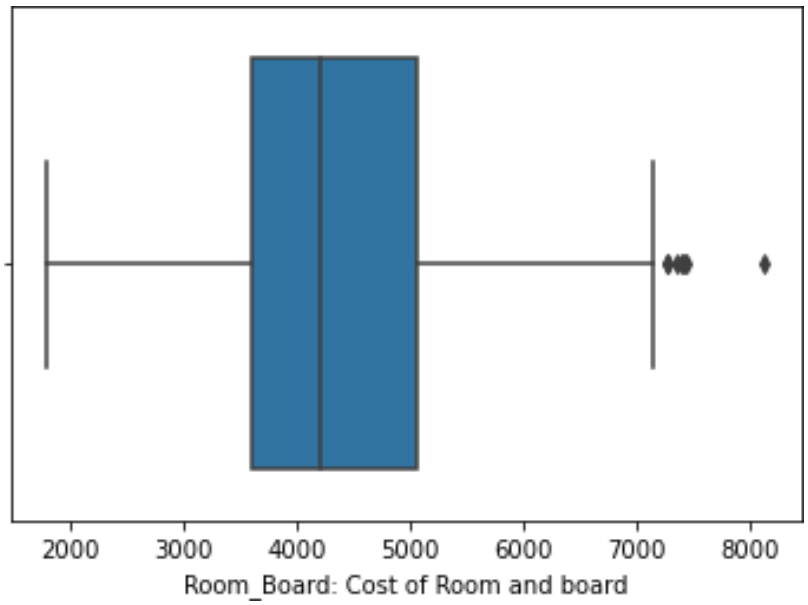


Room_Board: Cost of Room and board

- The cost of room and board ranges from $1780 to $8124.
- Barnard College has the maximum cost of room and board ($ 8124) and North Carolina A & T State University has the least cost of room and board ($1780).
- The mean cost of room and board is $4357.53 and the median is $ 4200.
- The mean is quite close to the median indicating that the distribution is almost normal wi th a shorter right tail as seen in the below figure.

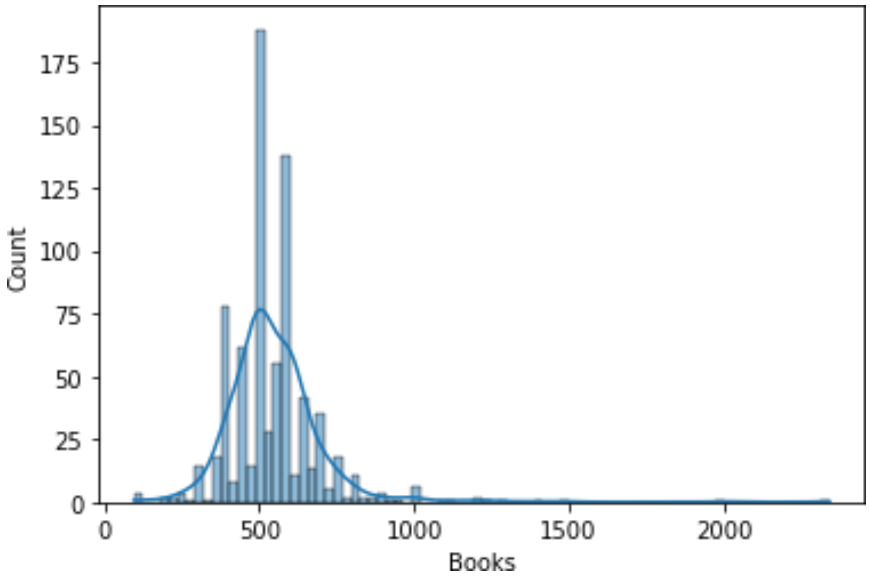


The boxplot below shows the presence of a few outliers.

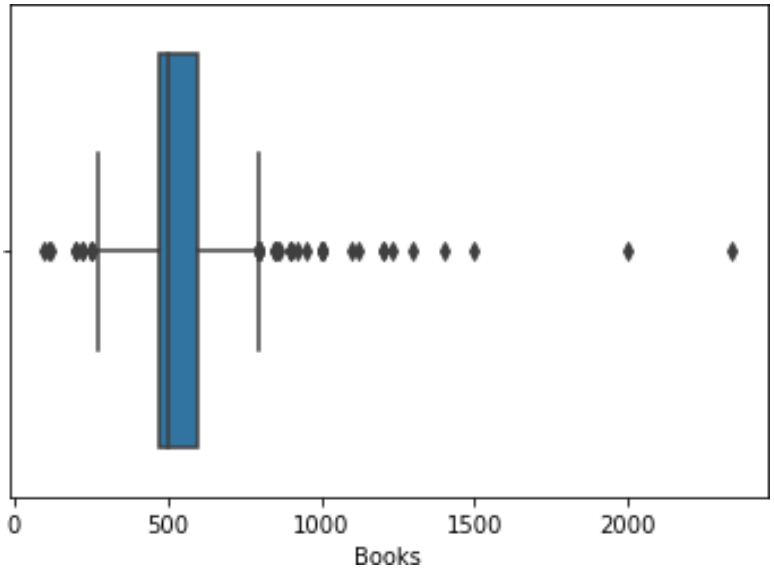


Books: Estimated book costs for a student

- The estimated book costs for a student ranges from $96 to $2340.
- Center for Creative Studies has the maximum estimated book costs ($ 2340) for a student and Appalachian State University has the least book costs for a student ($ 96).
- The mean estimated book cost is $549.38 and the median is $500.
- The mean is greater than the median indicating that the distribution is right skewed as see n in the below figure.

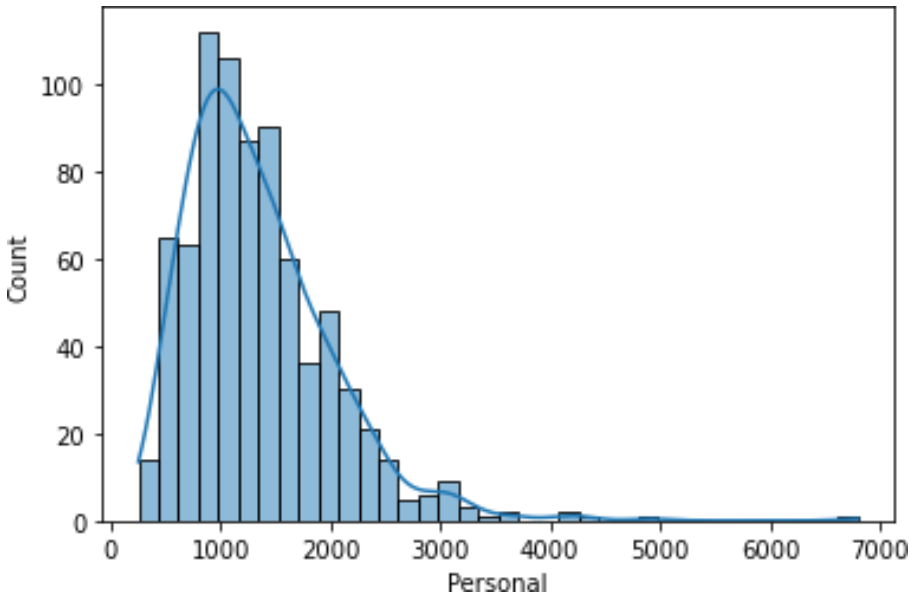


The below box plot shows the presence of outliers in the data.

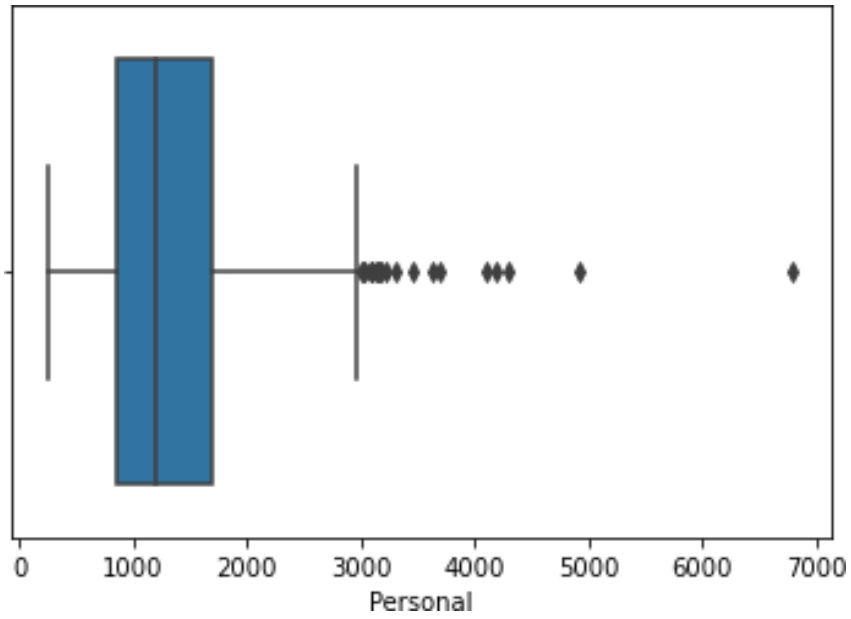


Personal: Estimated personal spending for a student

- The estimated personal spending for a student ranges from $250 to $6800.
- Saint Louis University has the maximum estimated personal spending for a student ($6800) and Benedictine College has the least ($ 250)
- The mean estimated personal spending for a student is $ 1340.64 and the median is $1200
- The mean is greater than the median indicating that the distribution is right skewed as seen in the below figure.



The boxplot below shows the outliers in the data.

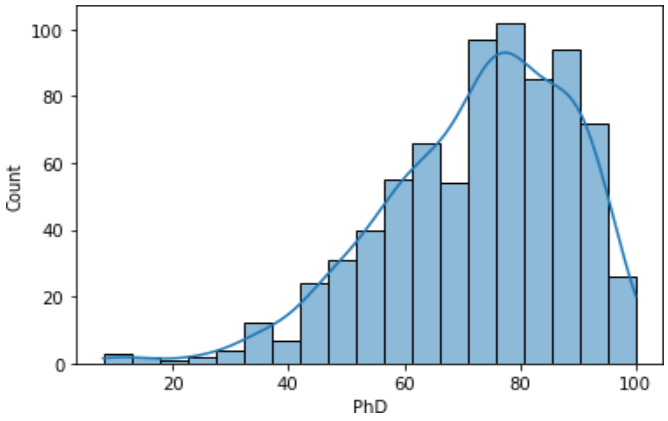


PhD: Percentage of faculties with Ph.D.'s

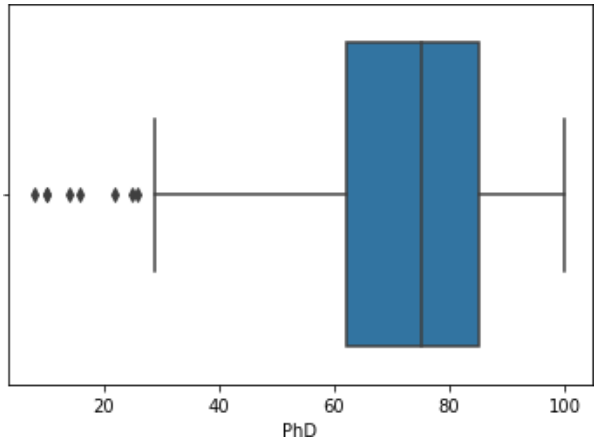
- The percentage of faculties with Ph.D.'s ranges from 8% to 100%.
- Three colleges have the maximum percentage of faculties with Ph.D.'s (100%) and Center for Creative Studies has the least (8%).

```
University wise listing   PhD
-------------------------------------------------
Names
Harvey Mudd College                          100
Pitzer College                               100
Bryn Mawr College                            100
New Mexico Institute of Mining and Tech.      99
Claremont McKenna College                     99
                                             ...
University of the Arts                        16
Savannah Coll. of Art and Design              14
Capitol College                               10
Wentworth Institute of Technology             10
Center for Creative Studies                    8
Name: PhD, Length: 777, dtype: int64
```

The mean percentage of faculties with Ph.D.'s is 72.62% and the median is 75%. The mean is lesser than the median indicating that the distribution is left skewed as seen in the below figure.



The boxplot below shows the presence of outliers.



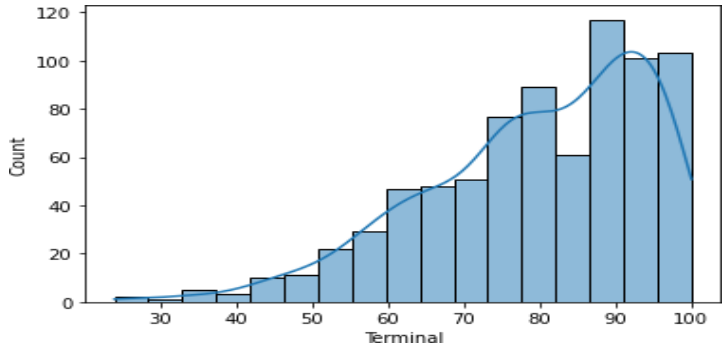
Terminal: Percentage of faculties with terminal degree

- The percentage of faculties with terminal degree ranges from 24% to 100%.
- The below universities/colleges have the maximum percentage of faculties with terminal Degree (100%) and Salem-Teikyo University has the least (24%).

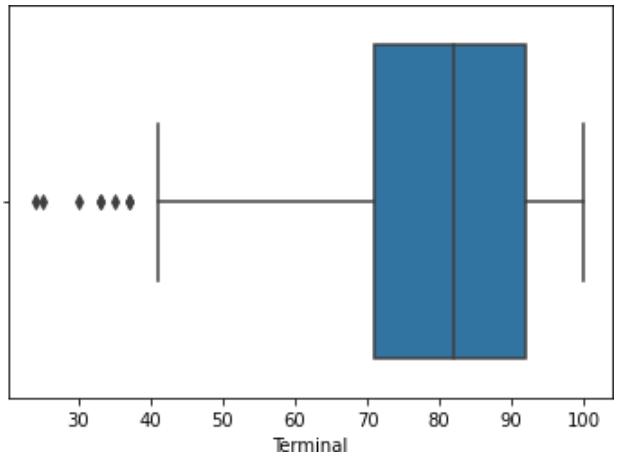
```
University wise listing   Terminal
--------------------------------------------------
Names
Scripps College                                100
Brown University                               100
University of Texas at San Antonio             100
New Mexico Institute of Mining and Tech.       100
Niagara University                             100
                                               ...
MidAmerica Nazarene College                     33
West Liberty State College                      33
Adelphi University                              30
Goldey Beacom College                           25
Salem-Teikyo University                         24
Name: Terminal, Length: 777, dtype: int64
```

The mean percentage of faculties with terminal degree is 79.70 % and the median is 92%.

The mean is lesser than the median indicating that the distribution is left skewed as seen in the below figure.



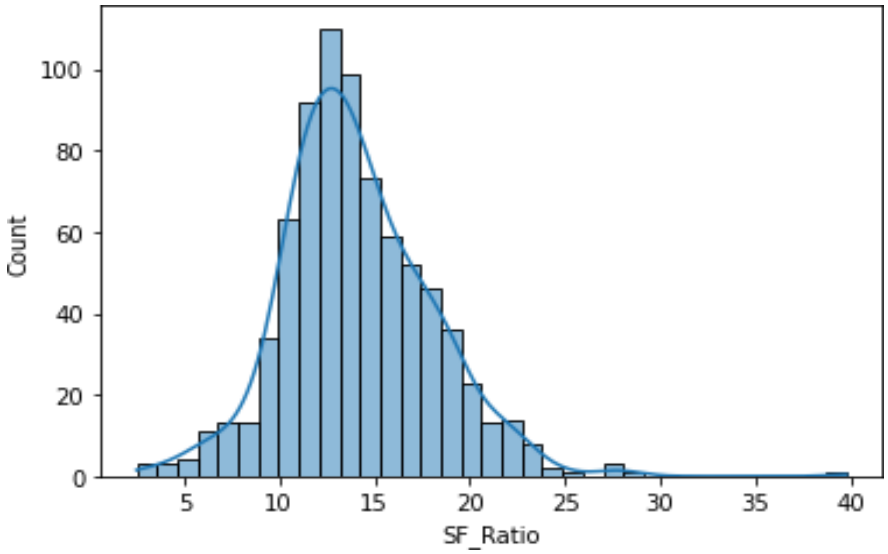
The below boxplot shows the presence of outliers.



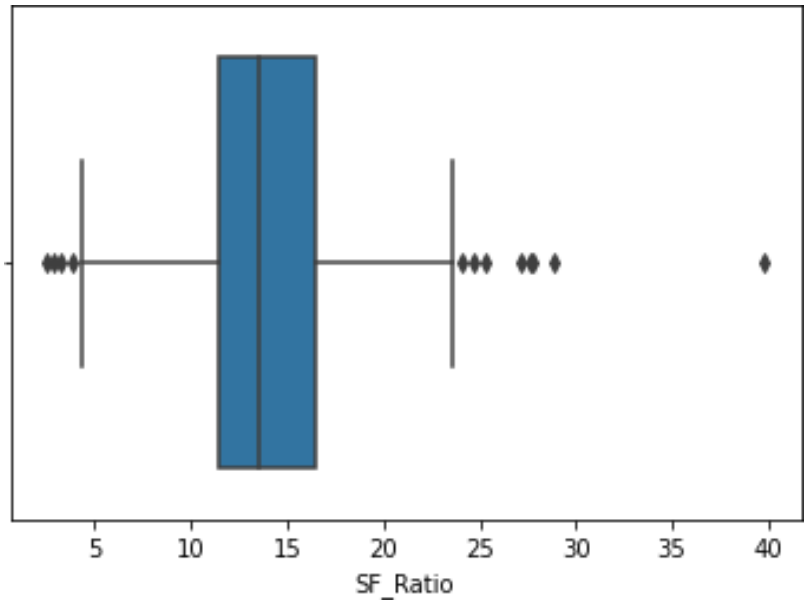
S.F. Ratio: Student/faculty ratio

- The student faculty ratio ranges from 2.5 to 39.8.
- Indiana Wesleyan University has the maximum student faculty ratio (39.8) and University of Charleston has the least (2.5).

- The mean student faculty ratio is 14.09 and the median is 13.6.
- The mean is quite close to the median indicating that the distribution is almost normal with short right tail as seen below.



The below boxplot shows the presence of outliers.



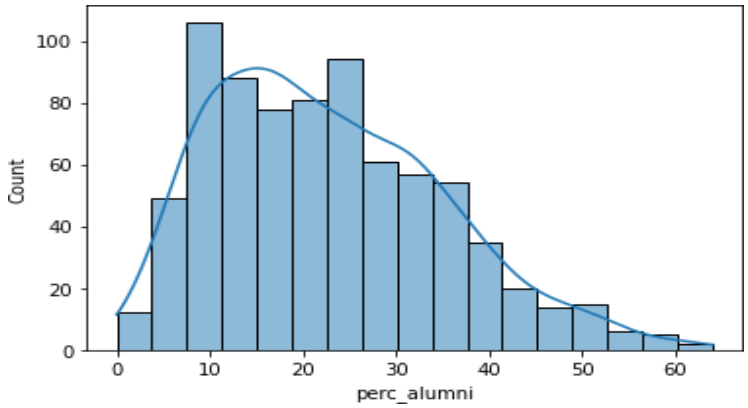
perc.alumni: Percentage of alumni who donate

- The percentage of alumni who donate ranges from 0% to 64%
- Williams College has the maximum percentage of alumni who donate (64%) and 2

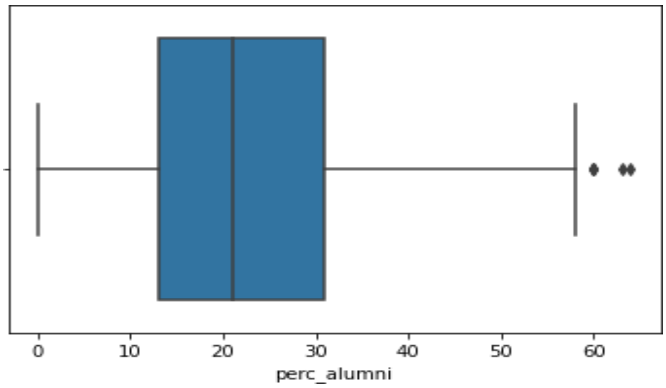
colleges as given in the below table have the least (0%).

```
University wise listing   perc_alumni
-------------------------------------------
Names
Williams College                        64
Amherst College                         63
Centre College                          60
Carleton College                        60
Hamilton College                        60
                                        ..
Columbia College MO                      2
Prairie View A. and M. University        1
University of Wisconsin at Green Bay     1
Central Washington University            0
University of Southern Colorado          0
Name: perc_alumni, Length: 777, dtype: int64
```

The mean percentage of alumni who donate is 22.74% and the median is 21% . The mean is quite close to the median indicating that the distribution is almost normal with a short right tail as seen below



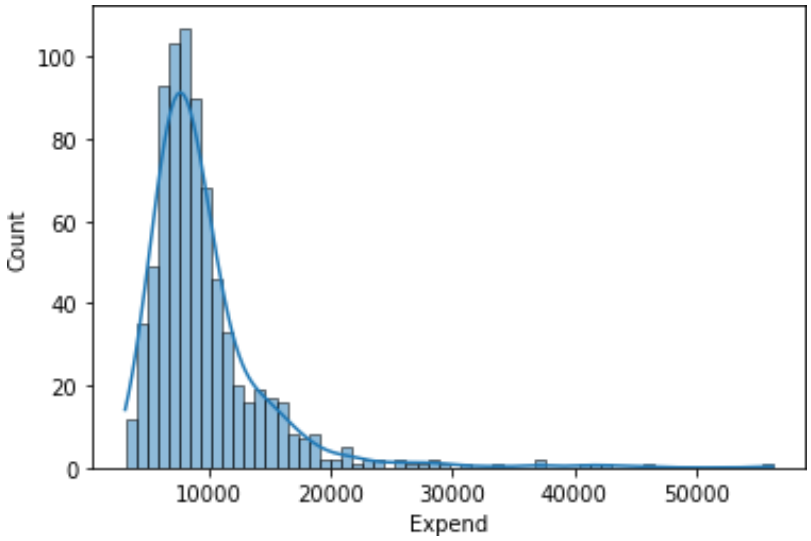
The below plot shows the outliers.



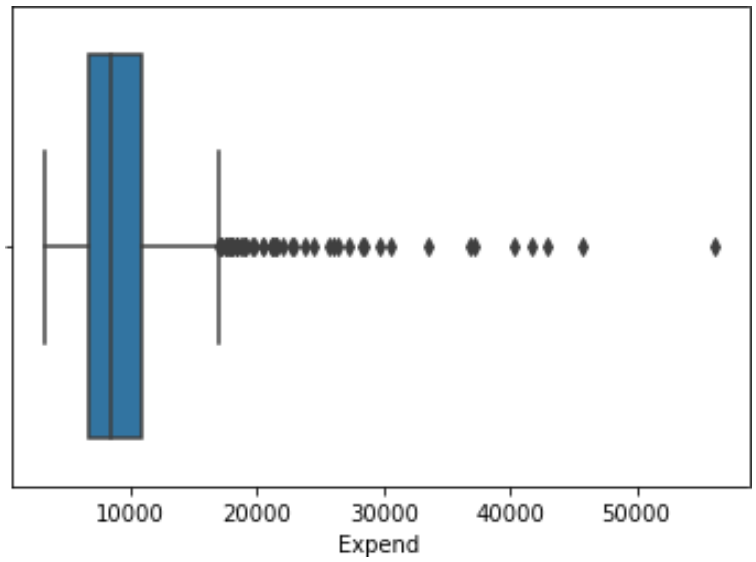
Expend: The Instructional expenditure per student

- The instructional expenditure per student ranges from $3186 to $56233.
- Johns Hopkins University has the maximum instructional expenditure per student ($56233) and Jamestown College has the least ($3186).

- The mean instructional expenditure per student is $9660.17 and the median is $ 8377.
- The mean is greater than the median indicating that the distribution is right skewed as seen in the below figure.



The below figure shows the presence of outliers.



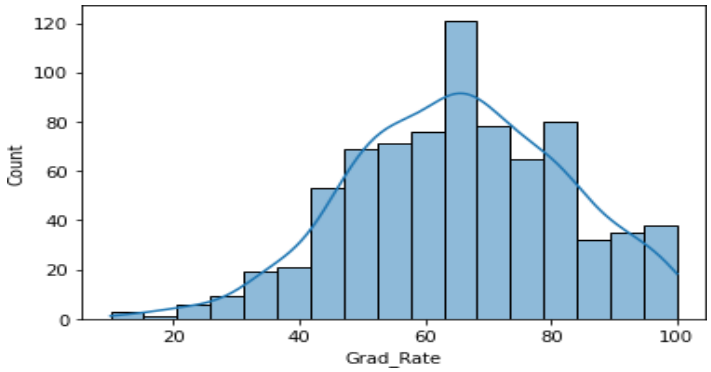
Grad.Rate: Graduation rate

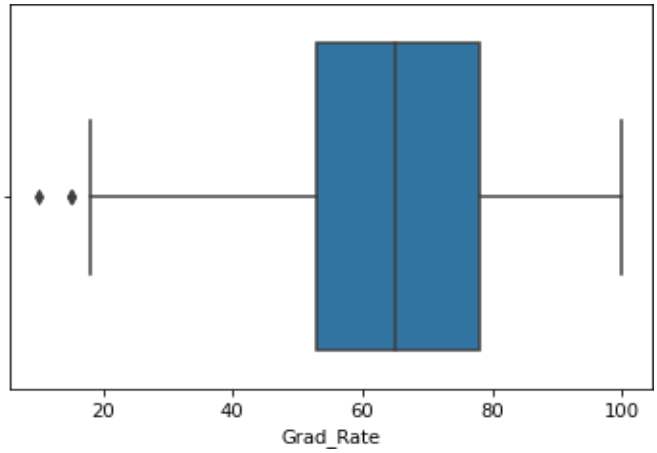
- **The graduation rate ranges from 10% to 100%.**
- The below table gives the universities with the maximum graduation rate (100%) and Tex as Southern University has the least (10%).

```
University wise listing   Grad_Rate
-----------------------------------------
Names
University of Richmond             100
Harvard University                 100
College of Mount St. Joseph        100
Harvey Mudd College                100
Missouri Southern State College    100
                                   ...
Claflin College                     21
Brewton-Parker College              18
Montreat-Anderson College           15
Alaska Pacific University           15
Texas Southern University           10
Name: Grad_Rate, Length: 777, dtype: int64
```

- The mean graduation rate is 65.40% and the median is 65%.
- The mean is almost equal to the median indicating that the distribution is normal as seen

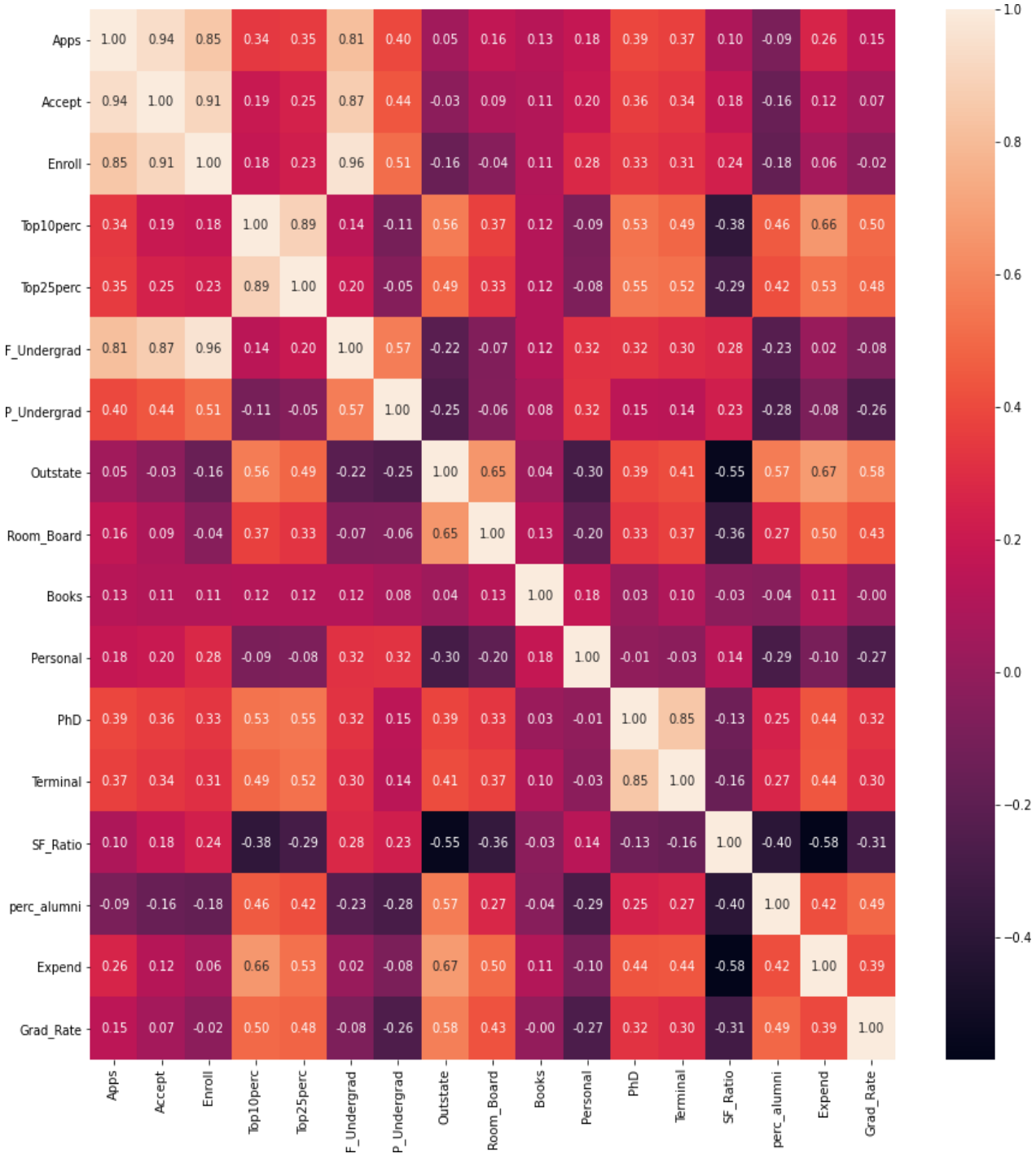
in the below figure. The outliers are seen in the boxplot.



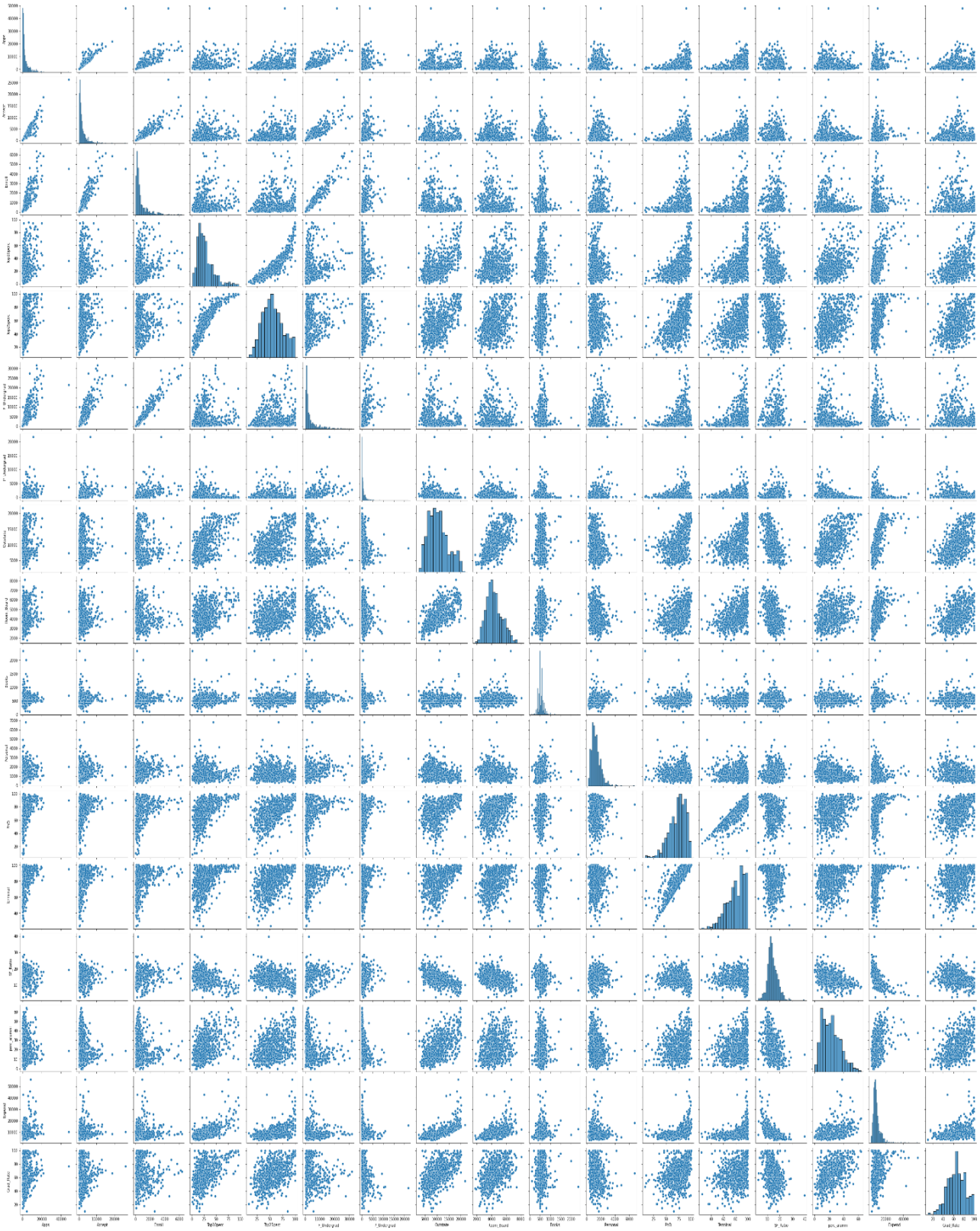


Multivariate Analysis- Bivariate Analysis

The below heat map gives the correlation between the 17 variables.

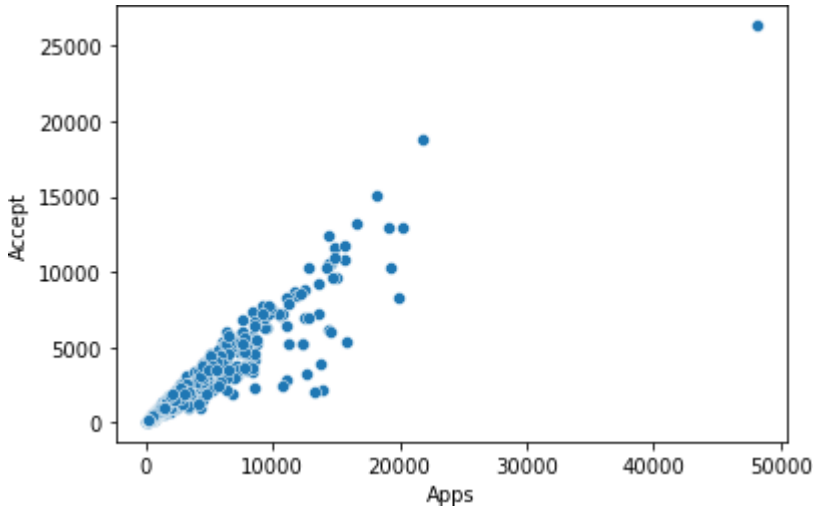


The below plot gives the pairwise scatterplot between the variables.

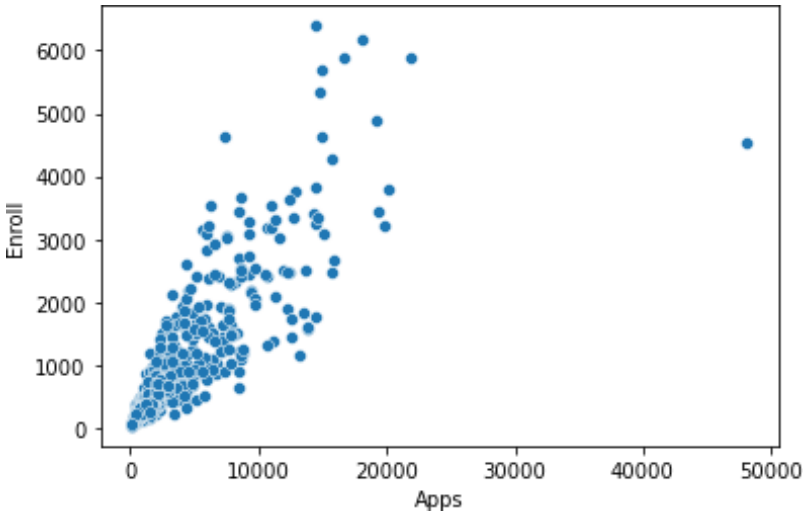


Pairplot of the 17 variables

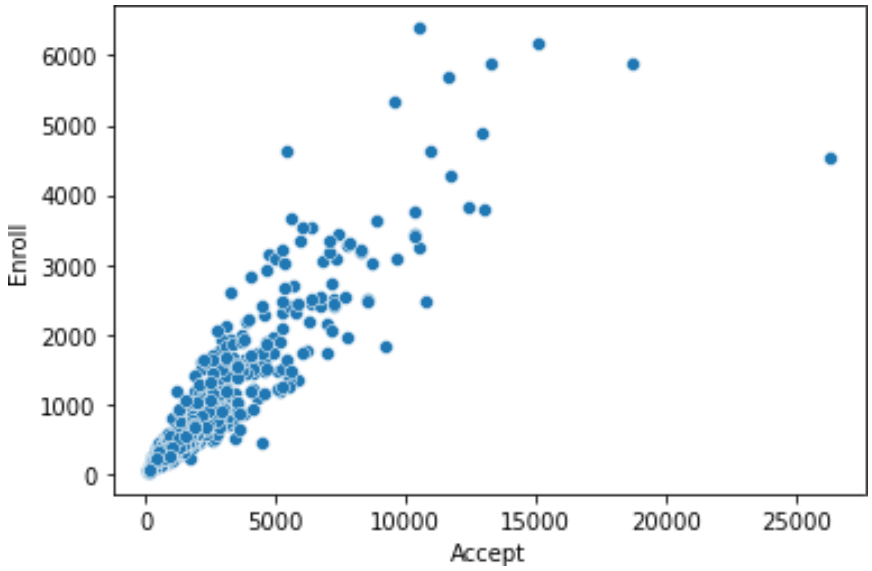
The significant plots out of these have been presented below**.**



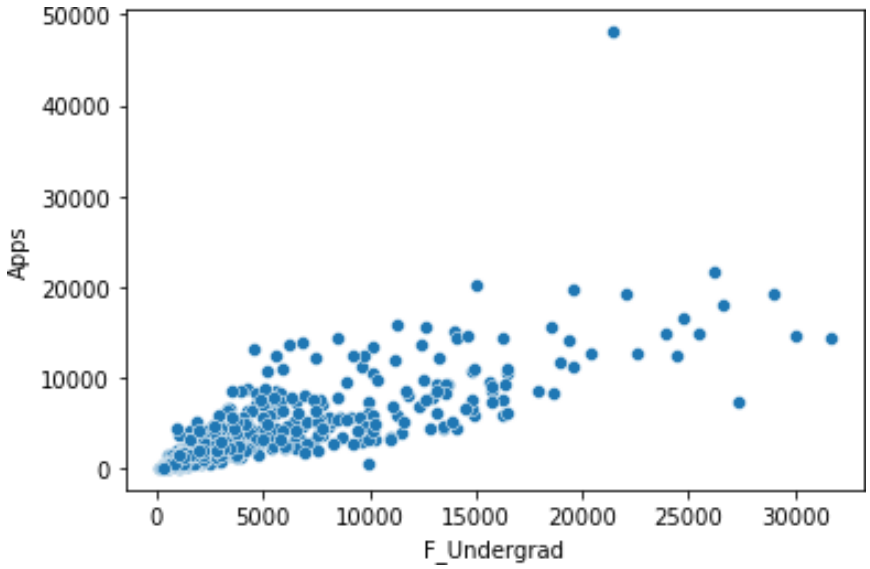
- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.94) between the number of applications received by a university and the number of applications accepted.



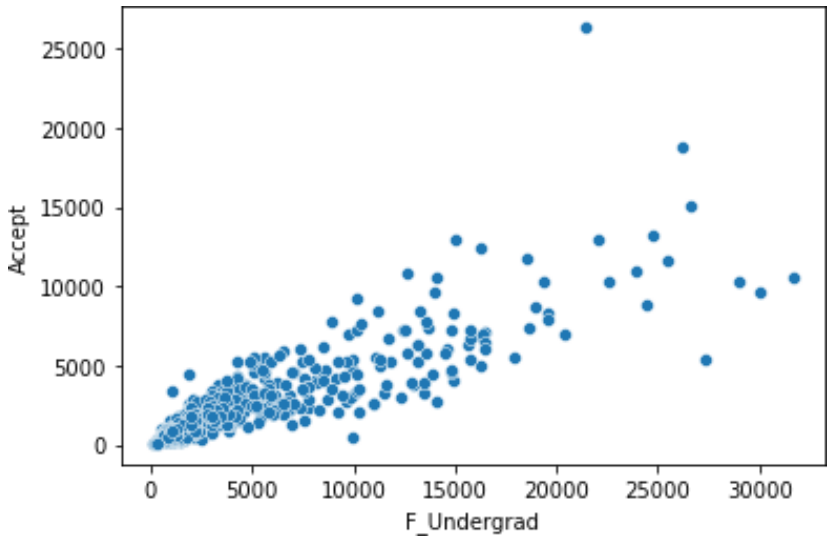
- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.85) between the number of applications received by a university and the number of students enrolled.



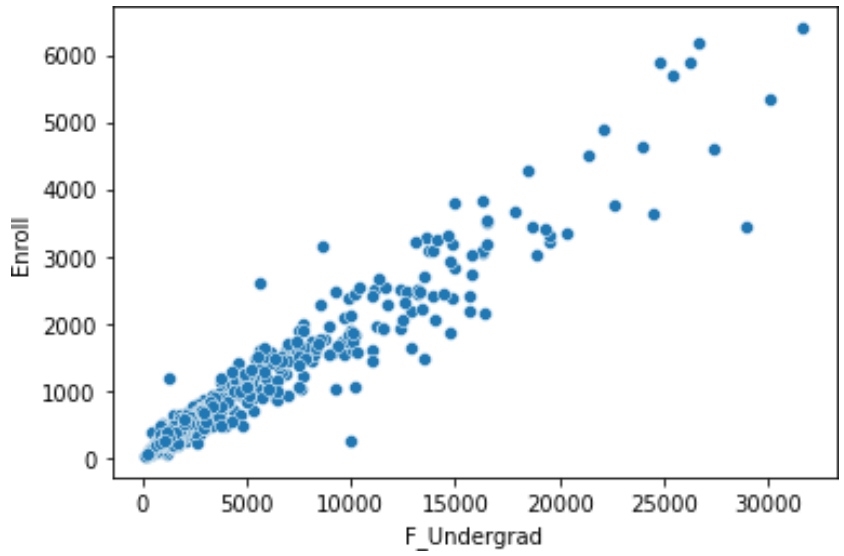
- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.91) between the number of applications accepted by a university and the number of students enrolled.



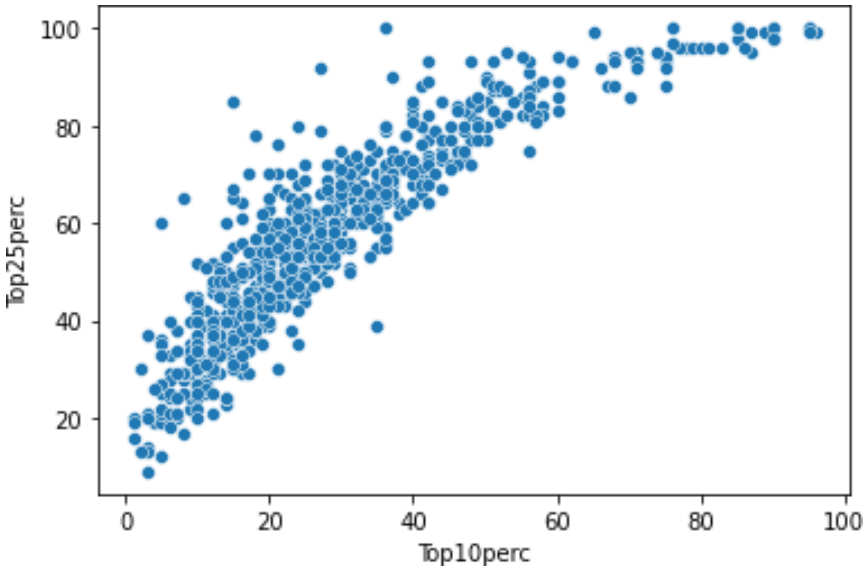
- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.81) between the number of full time undergraduate students and the number of applications received by a university.



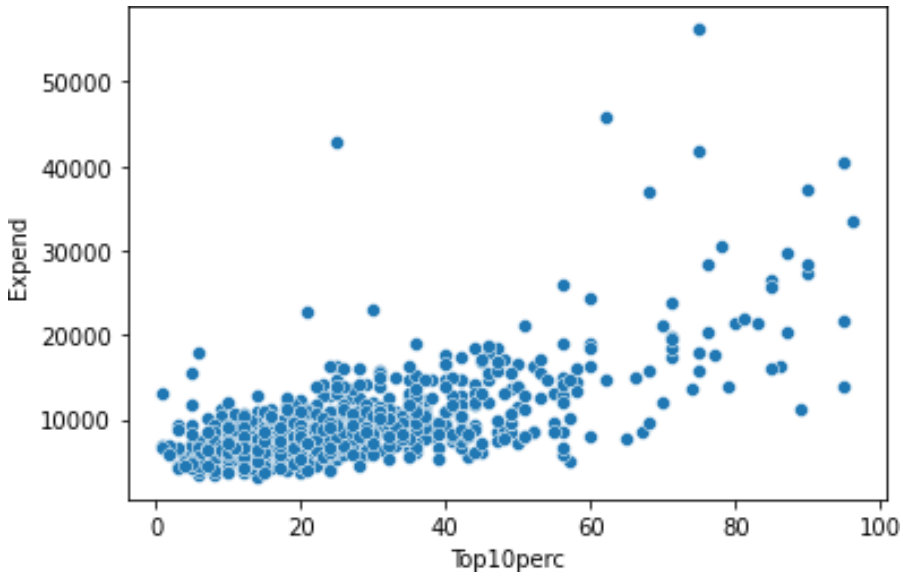
- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.87) between the number of full time undergraduate students and the number of applications accepted by a university.



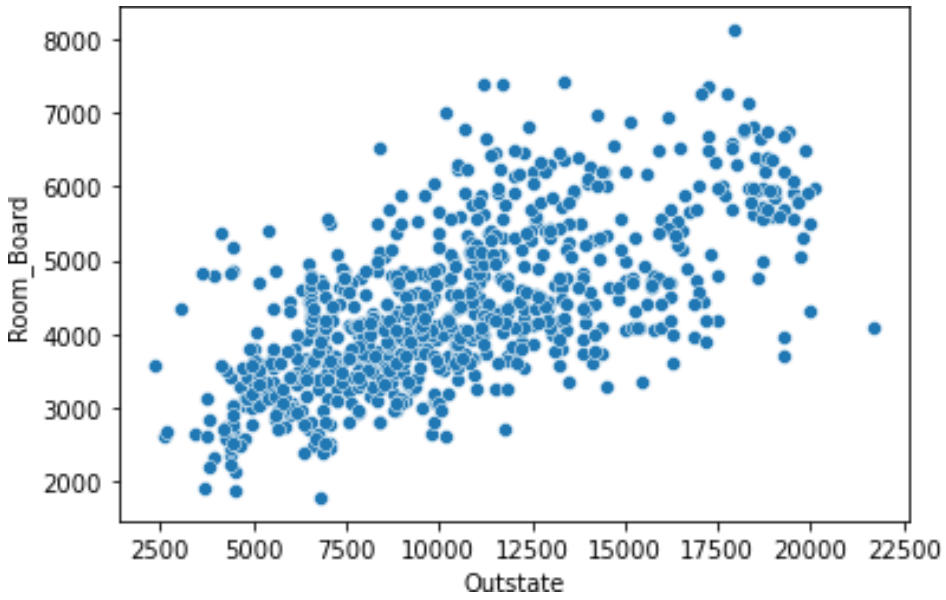
- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.96) between the number of full time undergraduate students and the number of students enrolled in a university.



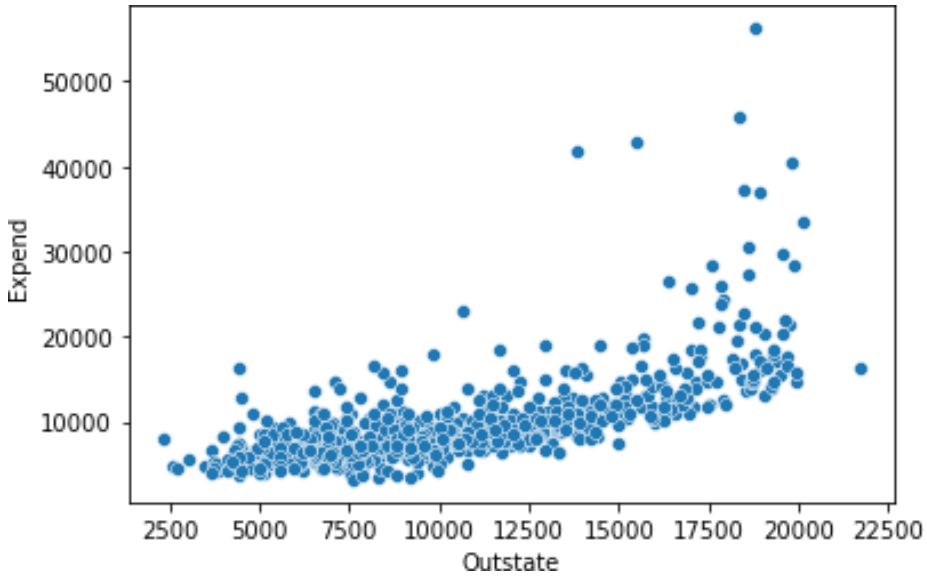
- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.89) between the percentage of new students from top 10% of Higher Secondary class and percentage of new students from top 25% of Higher Secondary class.



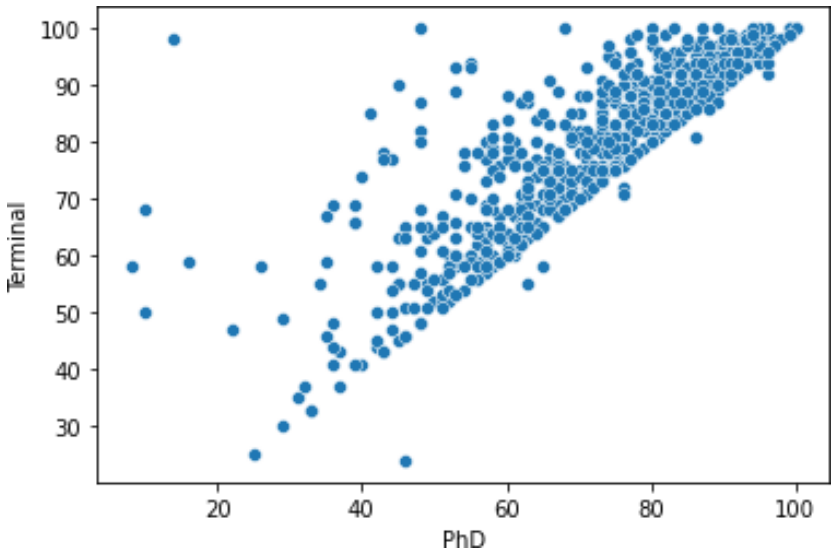
- From the heat map and the above scatter plot, we find a moderately positive correlation (correlation coefficient = 0.66) between the percentage of new students from top 10% of Higher Secondary class and the instructional expenditure per student.



- From the heat map and the above scatter plot, we find a moderately positive correlation (correlation coefficient = 0.65) between the cost of room and board and the number of students for whom the particular college or university is Out-of-state tuition



- From the heat map and the above scatter plot, we find a moderately positive correlation (correlation coefficient = 0.67) between the number of students for whom the particular college or university is Out-of-state tuition and the instructional expenditure per student.



- From the heat map and the above scatter plot, we find a very strong positive correlation (correlation coefficient = 0.85) between the percentage of faculties with Ph.D.'s and percentage of faculties with terminal degree.
- We can calculate the acceptance ratio or acceptance rate of a university/college by dividing the number of accepted applications divided by the total number of applications received.
- It is seen that the there are 6 universities/ colleges with 100% acceptance rates and Princeton University has the least acceptance rate with 15.45% as seen from the below tables.

	Names	Acceptance Ratio
729	Wayne State College	100.00
192	Emporia State University	100.00
355	Mayville State University	100.00
535	Southwest Baptist University	100.00
697	University of Wisconsin-Superior	100.00
368	MidAmerica Nazarene College	100.00
25	Arkansas Tech University	99.71
538	Southwestern Adventist College	99.07
452	Pikeville College	99.01
391	Mount Marty College	98.92

	Names	Acceptance Ratio
763	Williams College	29.74
144	Columbia University	28.57
174	Duke University	28.23
158	Dartmouth College	26.47
221	Georgetown University	25.92
70	Brown University	25.73
16	Amherst College	23.06
775	Yale University	22.91
250	Harvard University	15.61
459	Princeton University	15.45

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

Solution:

Yes. Scaling is necessary for PCA in this case. Since, the sample variances of the original variables show differences by large order of magnitude, variables need to be normalized (as seen in the below table).

```
The variances of the numeric variables are given below

Apps           1.497846e+07
Accept         6.007960e+06
Enroll         8.633684e+05
Top10perc      3.111825e+02
Top25perc      3.922292e+02
F_Undergrad    2.352658e+07
P_Undergrad    2.317799e+06
Outstate       1.618466e+07
Room_Board     1.202743e+06
Books          2.725978e+04
Personal       4.584258e+05
PhD            2.654282e+02
Terminal       2.167478e+02
SF_Ratio       1.566853e+01
perc_alumni    1.535567e+02
Expend         2.726687e+07
Grad_Rate      2.915125e+02
dtype: float64
```

Scaling ensures that the attribute means are all 0 and variances 1. A snapshot of the scaled data is seen in the below table.

	Apps	Accept	Enroll	Top10perc	Top25perc	F_Undergrad	P_Undergrad	Outstate	Room_Board	Books	Personal	PhD	Terminal	SF_
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.161177	-0.115729	1.0
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.679375	-3.378176	-0.4
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1.205308	-0.931341	-0.3
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.190052	1.175657	-1.6
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0.207340	-0.523535	-0.5
5	-0.624307	-0.628611	-0.669812	0.592287	0.313426	-0.623421	-0.535212	0.760947	-0.932970	-0.299280	-0.983753	-0.345435	-0.455567	-1.1
6	-0.684808	-0.685356	-0.729043	-0.598931	-0.545505	-0.677472	-0.410988	0.708713	1.243144	-0.299280	0.235515	1.067213	0.903786	-0.6
7	-0.285088	-0.121984	-0.313353	0.535563	0.616579	-0.434450	-0.541127	0.852479	0.427443	-0.602312	-0.725120	1.005793	1.379560	-0.0
8	-0.507700	-0.481644	-0.595505	0.138490	0.363952	-0.562562	-0.361036	1.282036	0.038754	-1.511408	-1.242385	0.391599	0.292077	-0.7
9	-0.625600	-0.620854	-0.654735	-0.372032	-0.596031	-0.598459	-0.510893	0.006798	-0.891911	0.670422	0.678885	-2.003761	-2.630532	-0.6

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

Solution:

For a scaled data, the covariance matrix and the correlation matrix are the same.

In this case too, we find that the covariance matrix reduces to the correlation matrix after scaling.

A snapshot of the two matrices are given below.

Correlation matrix

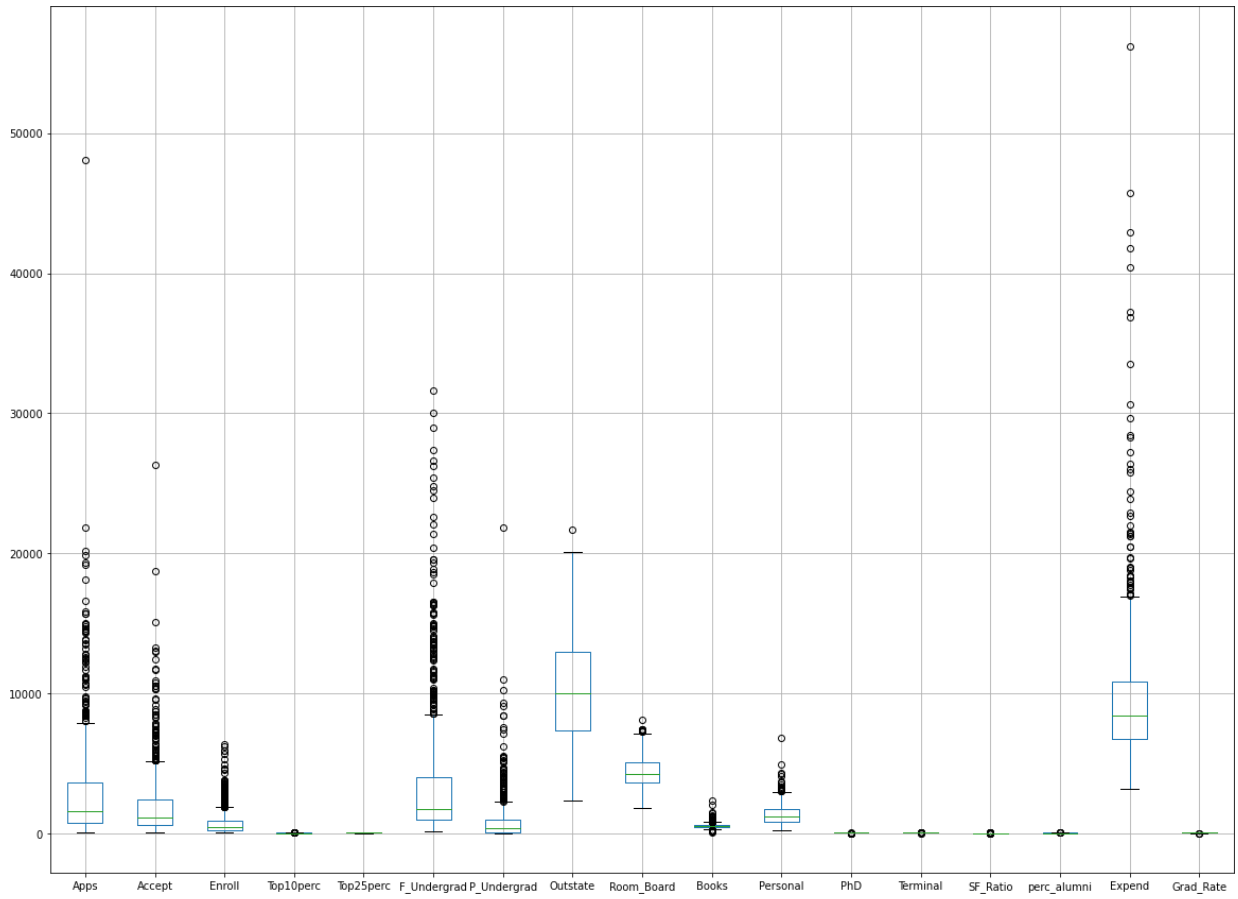
```
                Apps    Accept    Enroll  Top10perc  Top25perc  F_Undergrad
Apps        1.000000  0.943451  0.846822   0.338834   0.351640     0.814491
```

Covariance Matrix

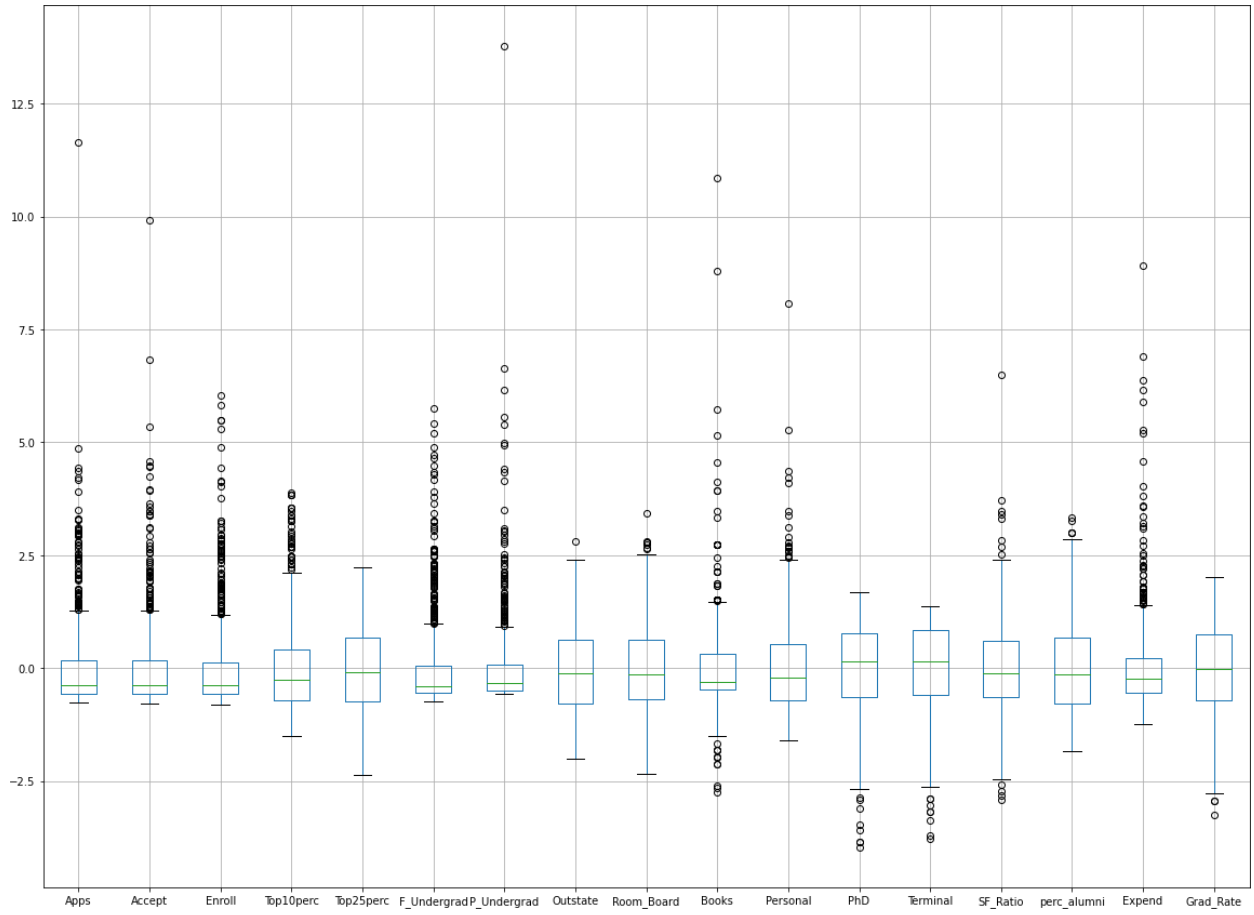
```
Covariance Matrix
%s [[ 1.00128866e+00  9.44666359e-01  8.47913316e-01  3.39270321e-01
   3.52093041e-01  8.15540181e-01  3.98777500e-01  5.02236717e-02
```

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

Solution:



Boxplot before scaling of data



Boxplot after scaling of data

The above two plots shows the boxplots before and after scaling. From, the 2 boxplots, we see the presence of outliers in the data. Normalisation of data does not remove the outliers, but only the range of the data changes.

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

Solution:

PCA has been performed and the principal component scores have been loaded into a data frame. The below gives the screenshot of the PC data frame.(Please refer Python file)

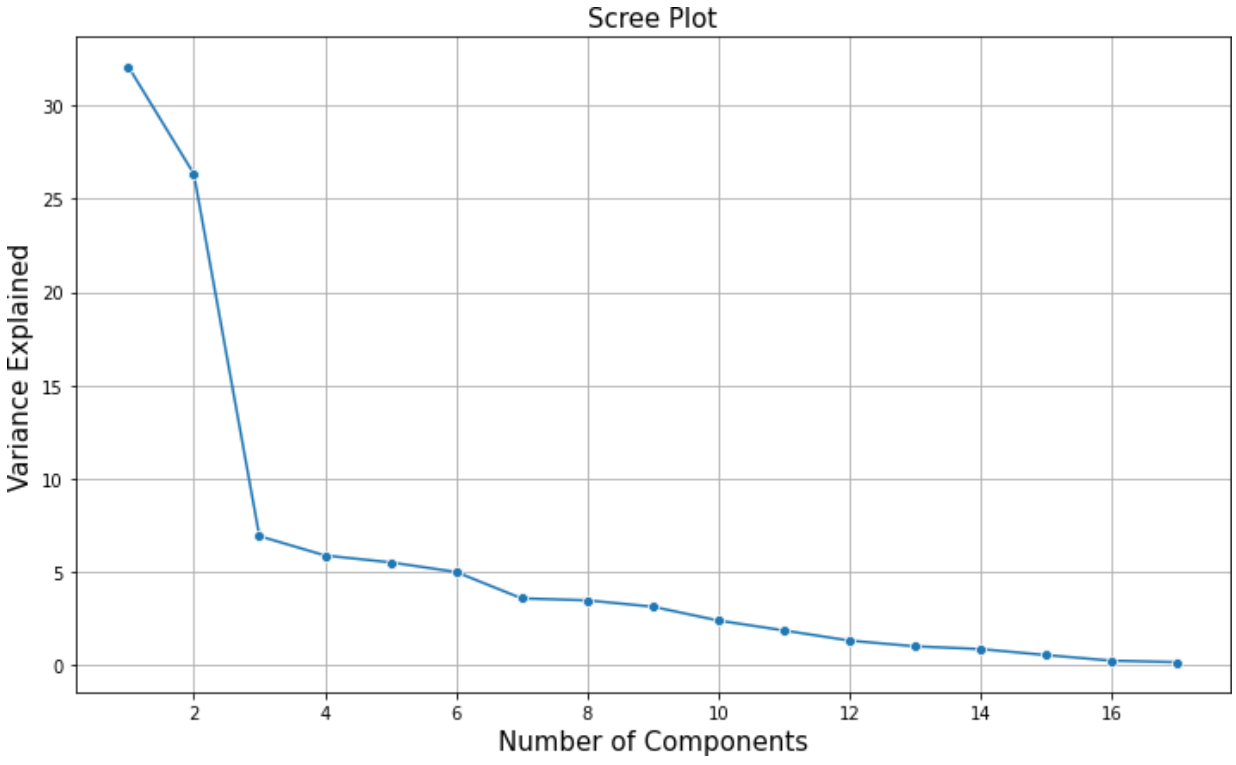
	Apps	Accept	Enroll	Top10perc	Top25perc	F_Undergrad	P_Undergrad	Outstate	Room_Board	Books	Personal	PhD	Terminal
0	0.247533	0.206300	0.175139	0.353991	0.343702	0.153528	0.025786	0.294966	0.248897	0.064283	-0.042598	0.319579	0.316776
1	0.332427	0.372875	0.404251	-0.081508	-0.043914	0.418089	0.315119	-0.248748	-0.136928	0.056607	0.219634	0.058854	0.047001
2	-0.059678	-0.097759	-0.081358	0.034279	-0.025509	-0.060563	0.137967	0.048149	0.152144	0.679694	0.495300	-0.131903	-0.070972
3	0.285097	0.271788	0.163457	-0.055180	-0.115855	0.101292	-0.158951	0.136572	0.191576	0.071380	-0.249033	-0.529195	-0.518145
4	0.000170	0.050617	-0.058798	-0.394019	-0.423913	-0.045408	0.306058	0.220140	0.556677	-0.131995	-0.217224	0.150814	0.214616
5	-0.012281	0.011277	-0.040120	-0.054194	0.030869	-0.041727	-0.193339	-0.026773	0.167220	0.640262	-0.337698	0.083092	0.149282

Terminal	SF_Ratio	perc_alumni	Expend	Grad_Rate
0.316776	-0.177164	0.205418	0.318606	0.255627
0.047001	0.246077	-0.246024	-0.130727	-0.168686
-0.070972	-0.291240	-0.147090	0.227919	-0.205241
-0.518145	-0.168468	0.016115	0.086067	0.243113
0.214616	-0.077059	-0.215815	0.074993	-0.115718
0.149282	0.485844	-0.047150	-0.297458	0.215657

6 Principal components have been generated with a cumulative variance of 81%. We have chosen 6 PCs since as a general rule 80-20 is taken. The below gives the cumulative variance explained.

```
Cumulative Variance Explained [ 32.0988744   58.44246116  65.34259565  71.20525863  76.69617559
  81.67414057  85.23346016  88.68686349  91.80133169  94.17208083
  96.01178946  97.30603027  98.29167258  99.13175814  99.64913029
  99.86482309 100.        ]
```

The scree plot is shown below as well.



2.5 Extract the eigenvalues and eigenvectors.

Solution:

The Eigen values and the Eigen vectors are given below:

```
Eigen Values
%s [5.46384062 4.4841809  1.17453449 0.99793705 0.93465879 0.84734458
0.60586408 0.58783511 0.53014189 0.40354672 0.02300969 0.03671503
0.3131535  0.08806661 0.14299858 0.16777512 0.22030447]
```

```
Eigen Vectors
 %s [[-2.47532537e-01  3.32426877e-01  5.96777039e-02 -2.85096690e-01
   1.70019622e-04  1.22811932e-02  3.06596836e-02  1.03577277e-01
   8.93297613e-02 -5.07359465e-02  3.59723152e-01 -4.59049537e-01
   4.31741045e-02 -1.32822763e-01 -6.60495278e-02 -5.97045093e-01
  -2.33859080e-02]
 [-2.06299756e-01  3.72875058e-01  9.77593828e-02 -2.71787742e-01
   5.06165527e-02 -1.12771266e-02  2.81197218e-03  5.55282328e-02
   1.76710689e-01 -4.05496164e-02 -5.44054193e-01  5.17547918e-01
  -5.86528967e-02  1.44994783e-01 -2.62518280e-02 -2.92920428e-01
   1.46702734e-01]
 [-1.75138842e-01  4.04250640e-01  8.13575684e-02 -1.63457087e-01
  -5.87975580e-02  4.01196268e-02  2.35400786e-02 -5.83157504e-02
   1.28035155e-01 -3.14617978e-02  6.09048127e-01  4.05328352e-01
  -6.83645378e-02 -2.99576263e-02  7.39350561e-02  4.46935450e-01
  -1.21051996e-02]
 [-3.53990557e-01 -8.15077765e-02 -3.42794420e-02  5.51796334e-02
  -3.94019023e-01  5.41942245e-02  1.64625571e-01  1.29536151e-01
  -3.38331896e-01 -6.40272957e-02 -1.44813035e-01  1.47935152e-01
  -8.77395752e-03 -6.98036916e-01  1.07536920e-01  1.60339719e-03
  -3.78844097e-02]
 [-3.43702467e-01 -4.39140881e-02  2.55085723e-02  1.15855059e-01
  -4.23913172e-01 -3.08687410e-02  1.25835762e-01  1.08116507e-01
  -4.01041645e-01 -1.43050699e-02  8.04965700e-02 -5.19055777e-02
  -2.74016941e-01  6.17396656e-01 -1.49296761e-01 -2.48611440e-02
   9.11239455e-02]
[-1.53527590e-01  4.18089297e-01  6.05628562e-02 -1.01292070e-01
 -4.54078528e-02  4.17274288e-02  2.40116146e-02 -7.82383577e-02
  5.92590638e-02 -1.74906259e-02 -4.14017608e-01 -5.60675304e-01
 -7.99532742e-02 -1.00315012e-02  4.28876301e-02  5.25319912e-01
 -5.76487187e-02]
[-2.57859287e-02  3.15119134e-01 -1.37966832e-01  1.58951251e-01
  3.06058094e-01  1.93339422e-01 -2.51111263e-02 -5.70135018e-01
 -5.61028863e-01  2.23974904e-01  8.72484281e-03  5.31278335e-02
  1.02020732e-01 -2.09983126e-02 -1.94333337e-02 -1.26509734e-01
  6.28499474e-02]
[-2.94965994e-01 -2.48748173e-01 -4.81491197e-02 -1.36572002e-01
  2.20139878e-01  2.67734158e-02 -1.12135600e-01 -1.46220489e-02
  1.20304645e-03 -1.81786941e-01  5.14125438e-02 -1.02095362e-01
  1.46199343e-01 -3.95392860e-02  2.58416552e-02  1.45891405e-01
  8.22908398e-01]
[-2.48896885e-01 -1.36927922e-01 -1.52143504e-01 -1.91576166e-01
  5.56677349e-01 -1.67220434e-01 -2.18058092e-01  2.12995569e-01
 -2.80843531e-01 -2.93448662e-01  1.23870180e-03  2.56620894e-02
 -3.60912134e-01 -3.69908819e-03  5.72153462e-02  7.18354751e-02
 -3.52608505e-01]
[-6.42827785e-02  5.66066526e-02 -6.79693620e-01 -7.13795514e-02
 -1.31994951e-01 -6.40262169e-01  1.50371205e-01 -2.07821399e-01
  1.35859726e-01  8.50316393e-02  7.40409608e-04 -2.85111489e-03
  3.34809127e-02  9.57613442e-03  6.82096145e-02 -9.82435927e-03
  2.76901401e-02]
```

```
[ 4.25981738e-02  2.19634151e-01 -4.95300078e-01  2.49032745e-01
 -2.17223922e-01  3.37697849e-01 -6.37299867e-01  2.08241997e-01
  8.38764902e-02 -1.41756567e-01 -1.20672459e-03  1.31210253e-02
 -2.07268141e-02  2.65877604e-03 -3.01808078e-02 -4.00538504e-02
  4.01144680e-02]
[-3.19579232e-01  5.88540448e-02  1.31903121e-01  5.29195380e-01
  1.50814264e-01 -8.30924584e-02  2.57764237e-03  7.72742010e-02
  1.86611542e-01  1.18174458e-01  1.40741118e-02 -2.98639738e-02
  3.98157243e-02  1.10964608e-01  6.96287228e-01 -1.11116039e-01
 -1.97930769e-02]
[-3.16776477e-01  4.70008357e-02  7.09715596e-02  5.18144779e-01
  2.14615639e-01 -1.49282189e-01  4.17554494e-02  1.40630479e-02
  2.59934804e-01  7.70099176e-02  6.18241305e-03  2.70252109e-02
 -6.25480083e-02 -1.57564764e-01 -6.69882232e-01  3.94111128e-02
 -1.86203119e-02]
[ 1.77164491e-01  2.46076655e-01  2.91240091e-01  1.68468205e-01
 -7.70591919e-02 -4.85844391e-01 -2.13980087e-01  7.54935950e-02
 -2.79116404e-01 -4.73970264e-01 -2.22862123e-03  2.11914505e-02
  4.42559093e-01  2.10406761e-02 -4.10690876e-02  1.66448115e-02
  1.06883947e-02]
[-2.05418014e-01 -2.46023579e-01  1.47090056e-01 -1.61146847e-02
 -2.15815334e-01  4.71502314e-02 -2.21096515e-01 -6.89202681e-01
  2.46797747e-01 -4.23595852e-01 -1.88728215e-02 -3.92402220e-03
 -1.33965467e-01  8.21266099e-03  3.13661076e-02 -1.02793833e-01
 -1.80051409e-01]

  [-3.18605544e-01 -1.30727369e-01 -2.27918829e-01 -8.60666647e-02
    7.49927681e-02  2.97457585e-01  2.24278705e-01  6.26002869e-02
    5.30219126e-02 -1.39085118e-01 -3.56405319e-02  4.42782385e-02
    6.89350125e-01  2.28296993e-01 -7.37876749e-02  9.04328465e-02
   -3.27702760e-01]
  [-2.55626868e-01 -1.68686285e-01  2.05241119e-01 -2.43113496e-01
   -1.15718370e-01 -2.15656623e-01 -5.64575953e-01 -1.48592841e-02
   -4.94830868e-02  5.91786591e-01 -1.45932934e-02  7.66081425e-03
    2.24433233e-01  4.01950518e-03 -4.72561238e-02  6.58020893e-02
   -1.27378733e-01]]
```

2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).

Solution:

The first PC is given by

0.25*Apps + 0.21*Accept + 0.18*Enroll + 0.36*Top10perc + 0.34*Top25perc +

0.15*F_Undergrad + 0.03*P_Undergrad + 0.29*Outstate + 0.25*Room_Board+

0.06*Books - 0.04*Personal + 0.32*PhD + 0.32*Terminal - 0.18*SF_Ratio +

0.21*perc_alumni + 0.32* Expend + 0.26*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?

Solution:

The below gives the cumulative variance explained.

```
Cumulative Variance Explained [ 32.0988744   58.44246116  65.34259565  71.20525863  76.69617559
  81.67414057  85.23346016  88.68686349  91.80133169  94.17208083
  96.01178946  97.30603027  98.29167258  99.13175814  99.64913029
  99.86482309 100.        ]
```

As a general rule 80-20 is taken, for choosing the number of principal components which are chosen from the cumulative variance explained. Here, we see that 81% is achieved after the 6^{th} Eigen value, hence 6 principal components have been chosen.

The Eigenvectors determine the directions of the new attribute space, and the eigenvalues determine their magnitude. As can be seen in the PCA, the components of the eigen vectors determine the PCs.

2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis?

Solution:

Business Implications:

In this case study, the data set consisted of a large number of attributes/variables (17- excluding the Names field). Performing PCA reduced the dimensionality of this large data set by transforming the set of attributes into a smaller one that still contained most of the information in the large set.

Smaller data sets are easier to analyse and faster for ML algorithms without extraneous attributes to process. This has been simple to achieve in this data set as the data set contained a large number of correlated variables and PCA is a powerful tool which reduces this multicollinearity. Thus, in this case study, PCA has reduced the number of attributes of the data set, at the same time has retained as much information is possible.

In this data set, using the information on the Eigen values, Eigen Vectors and Cumulative

Variance Explained, the 6 PCS out of the 17 have been identified. Since, choosing 6 PCS has captured 81% of the variance and information in the original data set. As a general rule 80-20 is taken, for choosing the number of principal components which are chosen from the cumulative variance explained. Here, we see that 81% is achieved after the 6th Eigen value, hence 6 principal components have been chosen.

The below gives the screenshot of the 6 PCs data frame. (Please refer Python file)

	Apps	Accept	Enroll	Top10perc	Top25perc	F_Undergrad	P_Undergrad	Outstate	Room_Board	Books	Personal	PhD	Terminal
0	0.247533	0.206300	0.175139	0.353991	0.343702	0.153528	0.025786	0.294966	0.248897	0.064283	-0.042598	0.319579	0.316776
1	0.332427	0.372875	0.404251	-0.081508	-0.043914	0.418089	0.315119	-0.248748	-0.136928	0.056607	0.219634	0.058854	0.047001
2	-0.059678	-0.097759	-0.081358	0.034279	-0.025509	-0.060563	0.137967	0.048149	0.152144	0.679694	0.495300	-0.131903	-0.070972
3	0.285097	0.271788	0.163457	-0.055180	-0.115855	0.101292	-0.158951	0.136572	0.191576	0.071380	-0.249033	-0.529195	-0.518145
4	0.000170	0.050617	-0.058798	-0.394019	-0.423913	-0.045408	0.306058	0.220140	0.556677	-0.131995	-0.217224	0.150814	0.214616
5	-0.012281	0.011277	-0.040120	-0.054194	0.030869	-0.041727	-0.193339	-0.026773	0.167220	0.640262	-0.337698	0.083092	0.149282

Terminal	SF_Ratio	perc_alumni	Expend	Grad_Rate
0.316776	-0.177164	0.205418	0.318606	0.255627
0.047001	0.246077	-0.246024	-0.130727	-0.168686
-0.070972	-0.291240	-0.147090	0.227919	-0.205241
-0.518145	-0.168468	0.016115	0.086067	0.243113
0.214616	-0.077059	-0.215815	0.074993	-0.115718
0.149282	0.485844	-0.047150	-0.297458	0.215657

The explicit form of the first PC is given by

0.25*Apps + 0.21*Accept + 0.18*Enroll + 0.36*Top10perc + 0.34*Top25perc + 0.15*F_Undergrad + 0.03*P_Undergrad + 0.29*Outstate + 0.25*Room_Board+ 0.06*Books - 0.04*Personal + 0.32*PhD + 0.32*Terminal - 0.18*SF_Ratio + 0.21*perc_alumni + 0.32* Expend + 0.26*Grad_Rate

The above coefficients, 0.25, 0.21…. indicate the weights associated with each of these 17 attributes which make up the first PC. Similarly, the other coefficients or the weights can be got from the PC data frame shown above.

Further Analysis:

Using a component loading on a heat map, the features that have maximum loading across the components can be identified.

For each feature, the maximum loading value across the components can be found and the same can be marked with the help of rectangular box as seen in the below plot.

Features marked with rectangular red box are the ones having the maximum loading on the respective component. These marked features are used to decide the context that the

component represents. Using the components additional rules can be derived and analyzed. Unsupervised learning like clustering can further be applied on the data to segment the universities/colleges based on the components created and further analyzed.

Component Loading on a Heat Map