By: Vaishnavi Karelia

Advanced Statistics

PGPDSBA Online Feb_D 2021

Project: Advanced Statistics

Index

Problem 1A:

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

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

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

Formulate the hypothesis of the One-way ANOVA of 'Education' variable with the 'Salary' variable.

H_0 : There is no effect of education on Salary.

H_1 : At least one of the Education level, mean Salary is different from others.

H_0 : μS = μE = μO = μP

H_1 : Atleast one of the mean is different from others

Formulate the hypothesis of the One-way ANOVA of 'Occupation' variable with the 'Salary' variable.

H_0 : There is no effect of Occupation on Salary.

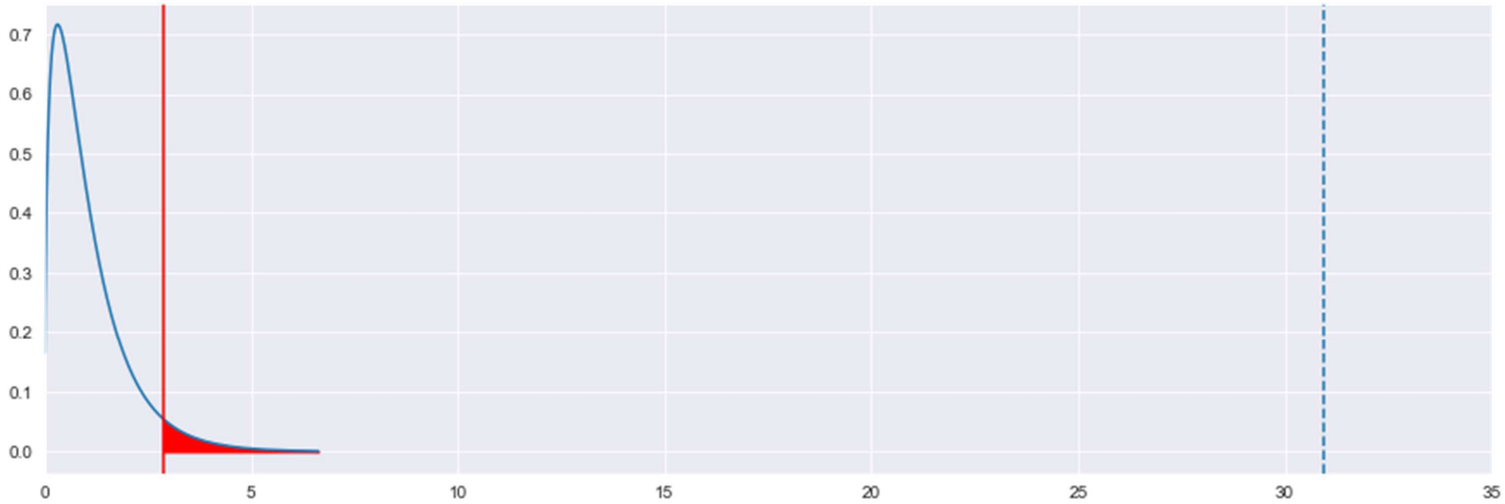
H_1 : At least one of the Occupation level, mean Salary is different from others.

H_0 : μS = μO = μE

H_1 : Atleast one of the mean is different from others

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.

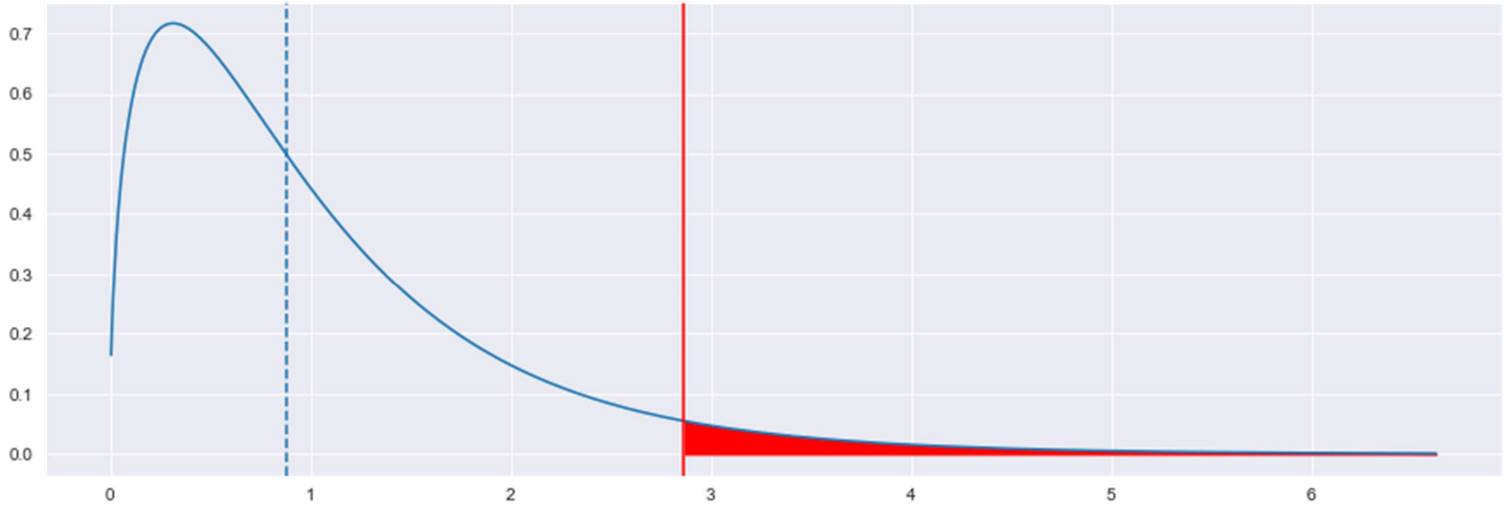
	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0000	102,695,466,735.8860	51,347,733,367.9430	30.9563	0.0000
Residual	37.0000	61,372,559,274.4889	1,658,717,818.2294	nan	nan



Conclusion: We can see that the p value is less than 5% (which was our αα) hence we reject the null hypothesis. This goes to indicate that the salary is dependent on at least one of education level. Mean Salary for at least one of the education level is different.

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.

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0000	11,258,782,926.4660	3,752,927,642.1553	0.8841	0.4585
Residual	36.0000	152,809,243,083.9090	4,244,701,196.7752	nan	nan



Conclusion : We can see that the p value is more than 5% (which was our α) hence we do not have sufficient evidence to reject the null hypothesis. This goes to indicate that the salary is dependent on 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.

```
           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
-----------------------------------------------------------------------------
```

Interpretation

We can see that the p value is less than 5% (which was our α) hence we reject the null hypothesis. This also indicates that mean of the each education category is equal.

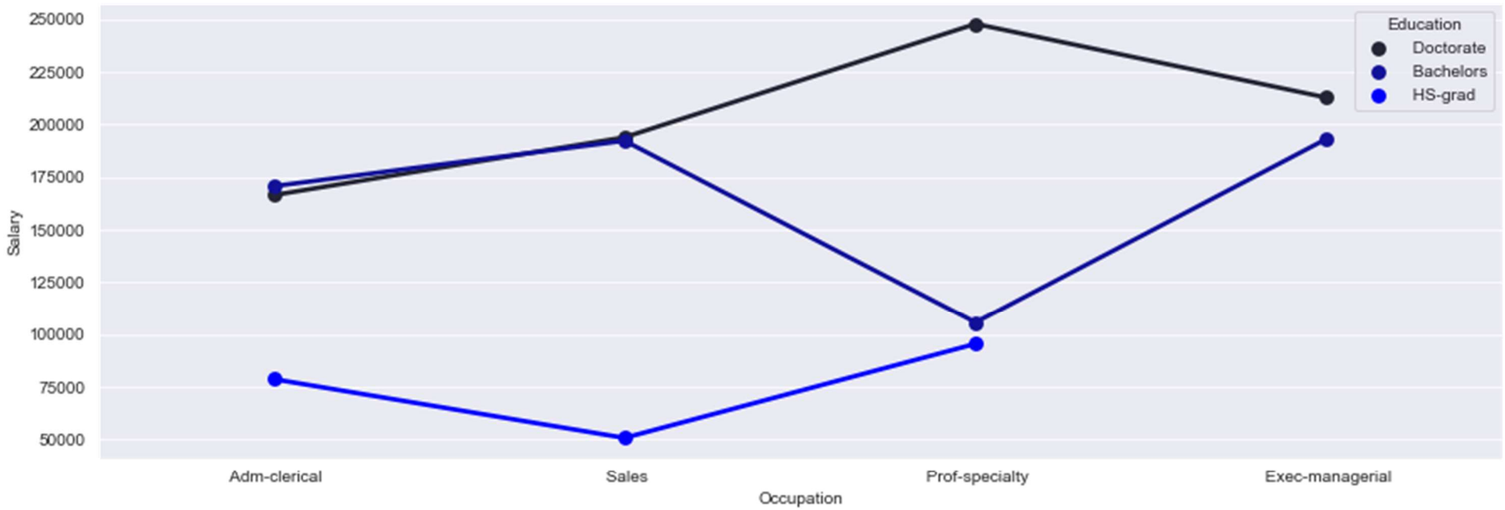
```
               Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================================================
     group1           group2        meandiff  p-adj     lower        upper    reject
---------------------------------------------------------------------------------
    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
---------------------------------------------------------------------------------
```

Interpretation

We can see that the p value is more than 5% (which was our α) hence we do not have sufficient evidence to reject the null hypothesis. This also indicates that mean of the each ocupation category is not equal.

Problem 1B:

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.



Analyse

From the interaction plot we can see very significant evidence of interaction between Doctorate and Bachelors in Adm-clerical and Sales Occupation. Also, there is weak evidence of interaction between Bachelors and HS-grad in Prof-speciality occupation.

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?

H_0 : The mean Salary for each combination of Education and Occupation category is equal. There is no interaction between Occupation and Education category.

H_1 : The mean Salary for atleast one combination of Education and Occupation is different from the others.**OR** there is interaction

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0000	11,258,782,926.4660	3,752,927,642.1553	5.2779	0.0050
C(Education)	2.0000	96,956,629,862.7808	48,478,314,931.3904	68.1766	0.0000
C(Education):C(Occupation)	6.0000	35,639,495,779.3358	5,939,915,963.2226	8.3535	0.0000
Residual	29.0000	20,621,020,503.0333	711,069,672.5184	nan	nan

Interpretation

We can see that the p value is less than 5% (which was our α) hence we reject the null hypothesis. This also indicates that the mean Salary for atleast one combination of Education and Occupation is different from the others.**OR** there is interaction

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

The main hypotheses is that the Salary variable is depended upon the Education of the employee and Occupation of the employee.

It is observed that the Salary is significantly impacted by Occupation and Education along with their interaction effect. Also, each of the occupational category is significantly different from each other whereas the education category is equal.

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?

Univariate Analysis

Categorical Columns

```
Univariate Analysis for column: Names
--------------------------------------------------

count                777
unique               777
top       Marist College
freq                   1
Name: Names, dtype: object
--------------------------------------------------
```

Since the categorical column has 777 unique values which is equal to total no of rows and each variable with 1 frequency the univariate analysis won't make much sense.

Univariate Analysis

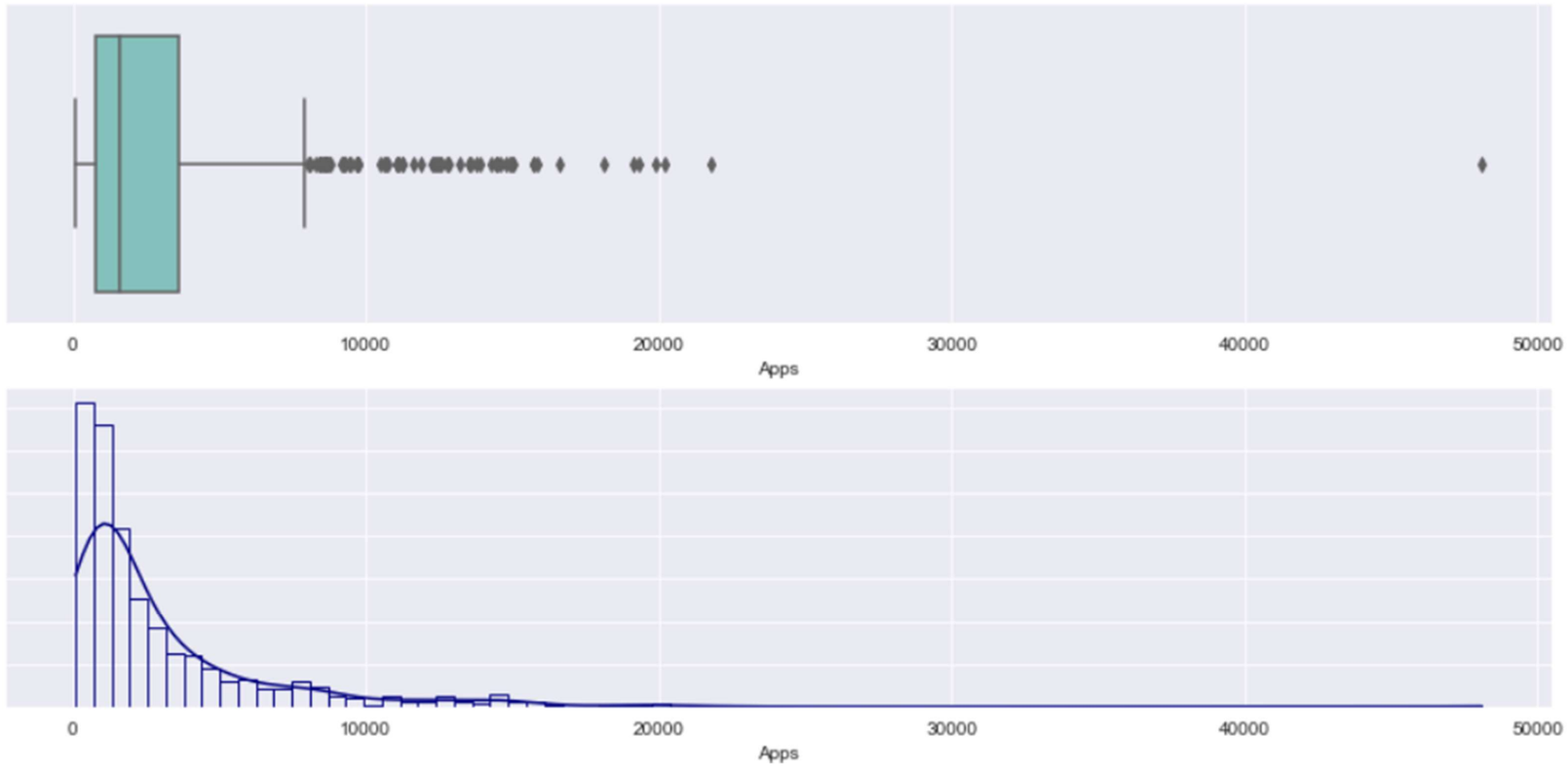
Numerical Columns

```
Univariate Analysis for column: Apps
--------------------------------------------------

count      777.0000
mean     3,001.6384
std      3,870.2015
min         81.0000
25%        776.0000
50%      1,558.0000
75%      3,624.0000
max     48,094.0000
Name: Apps, dtype: float64
--------------------------------------------------

Apps variable has outliers


Skewnes of the column is = -0.16655717024327263
        Hence, the column is negatively skewed or left skewed and is not normally distributed.
```

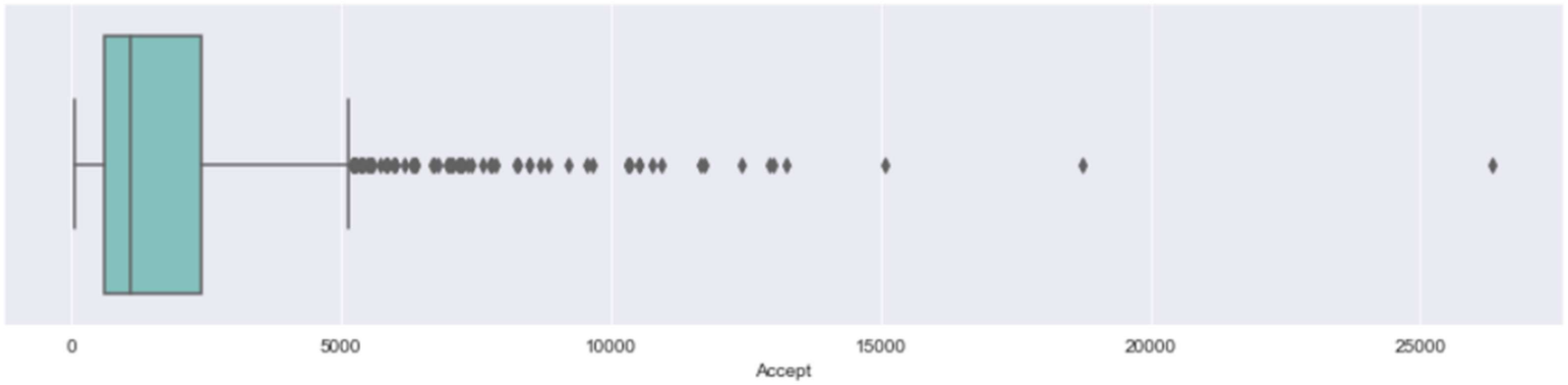


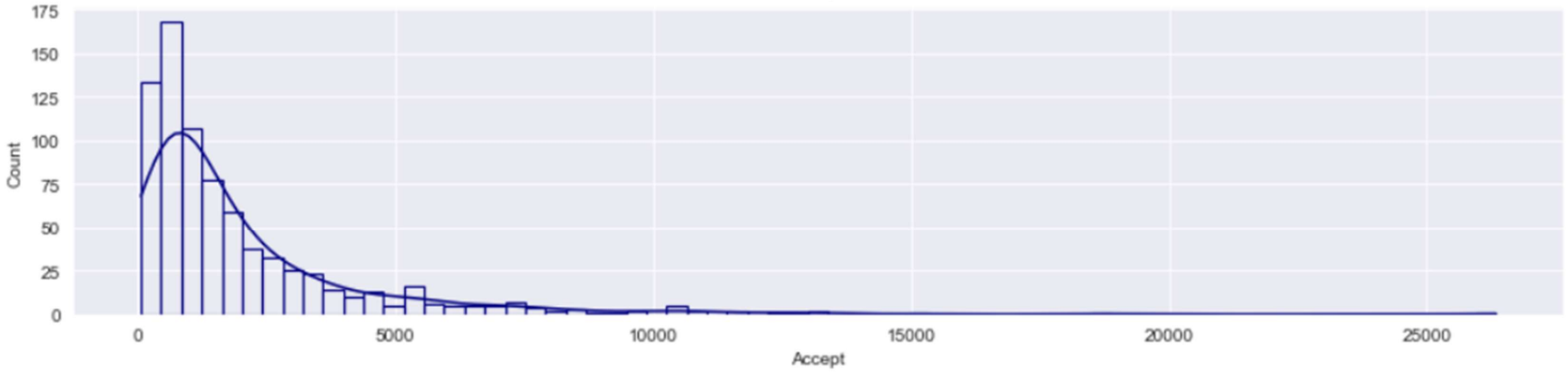
```
Univariate Analysis for column: Accept
--------------------------------------------------

count      777.0000
mean     2,018.8044
std      2,451.1140
min         72.0000
25%        604.0000
50%      1,110.0000
75%      2,424.0000
max     26,330.0000
Name: Accept, dtype: float64
--------------------------------------------------

Accept variable has outliers


Skewnes of the column is = 3.7165574035202718
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```



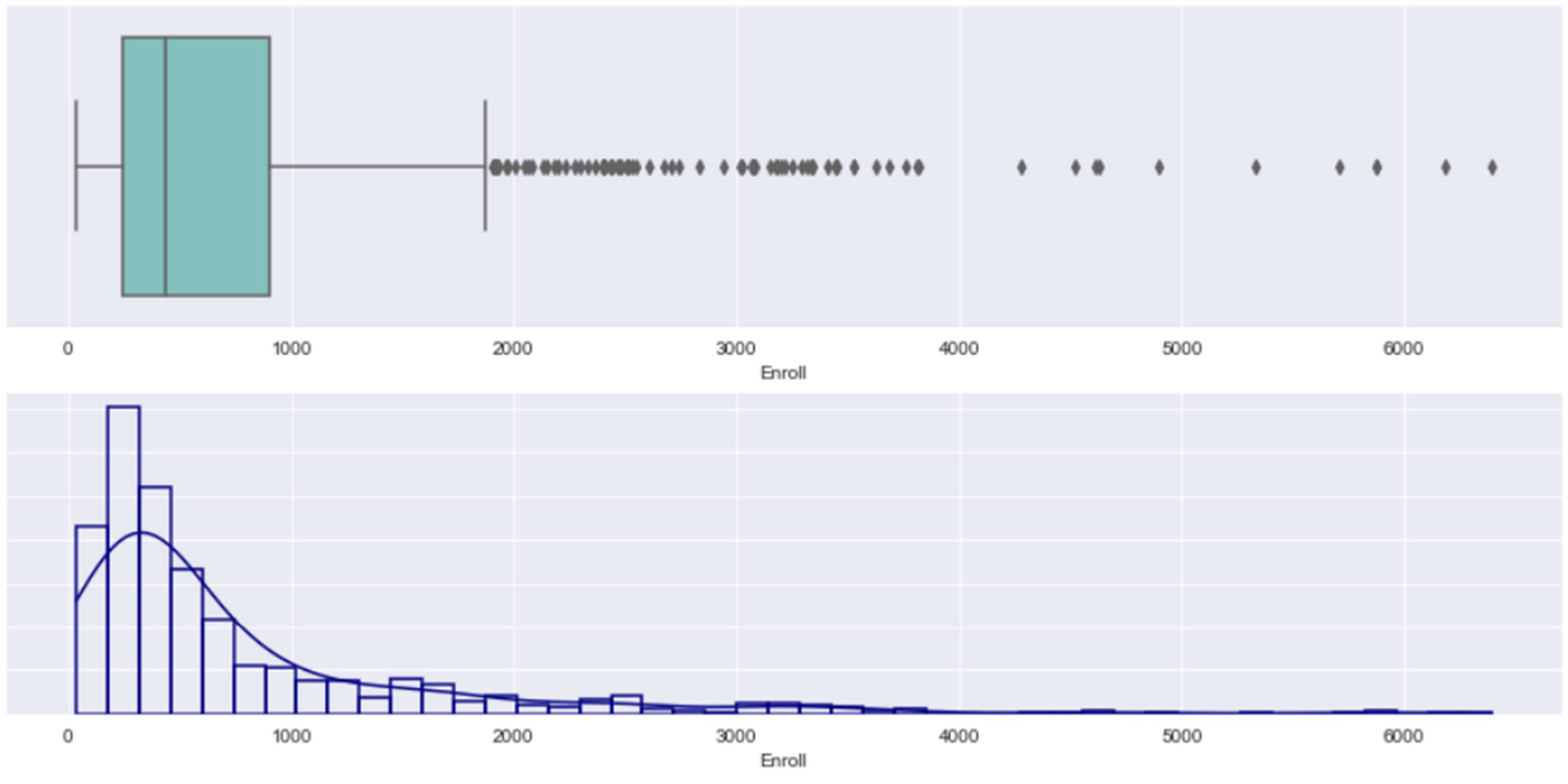


```
Univariate Analysis for column: Enroll
--------------------------------------------------

count     777.0000
mean      779.9730
std       929.1762
min        35.0000
25%       242.0000
50%       434.0000
75%       902.0000
max     6,392.0000
Name: Enroll, dtype: float64
--------------------------------------------------

Enroll variable has outliers


Skewnes of the column is = 3.4111258724395235
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```

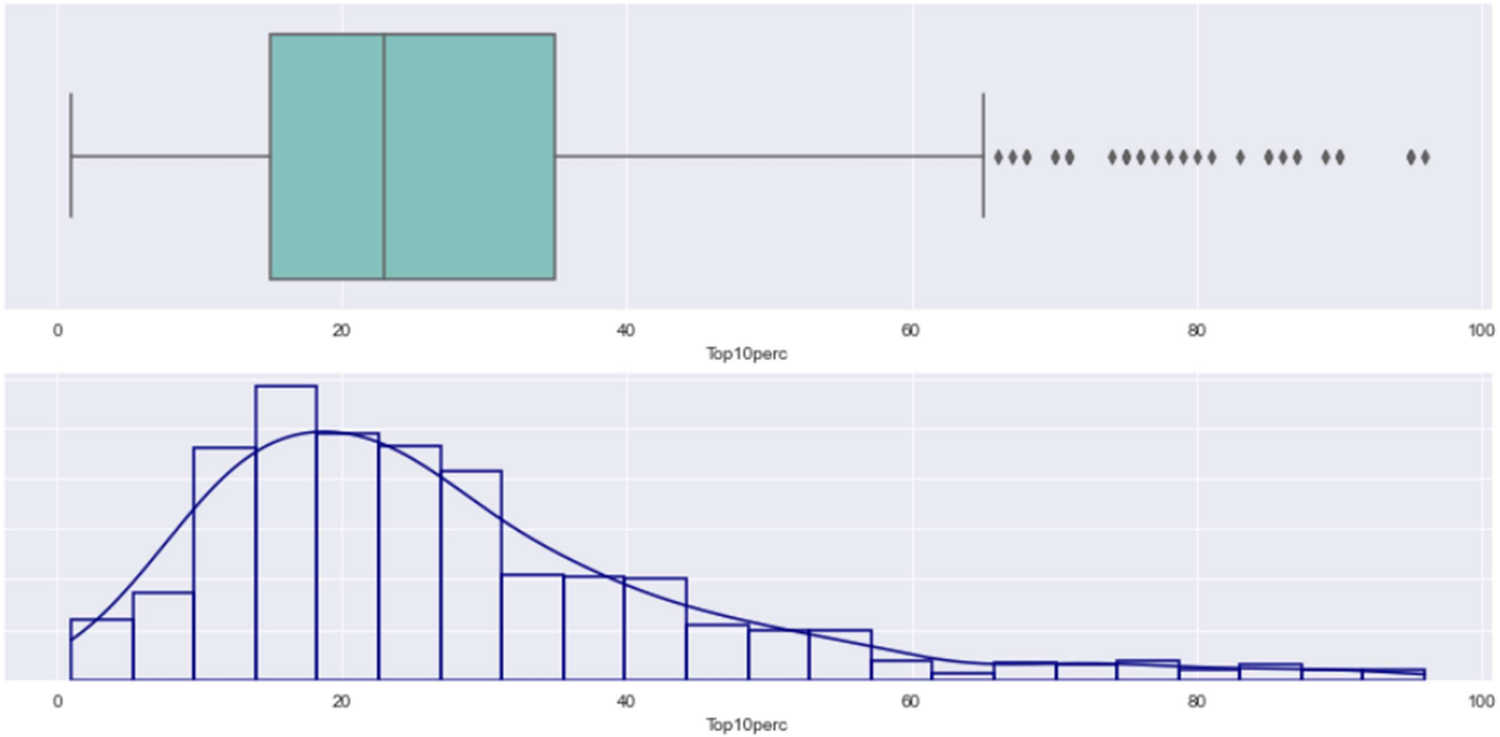


```
Univariate Analysis for column: Top10perc
--------------------------------------------------

count   777.0000
mean     27.5586
std      17.6404
min       1.0000
25%      15.0000
50%      23.0000
75%      35.0000
max      96.0000
Name: Top10perc, dtype: float64
--------------------------------------------------

Top10perc variable has outliers


Skewnes of the column is = 2.6852679191653412
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```

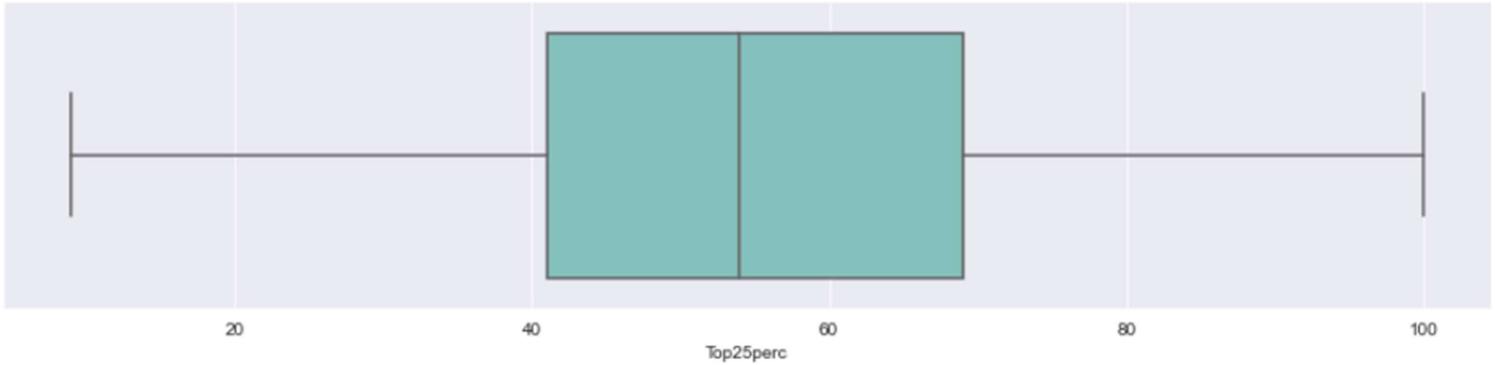


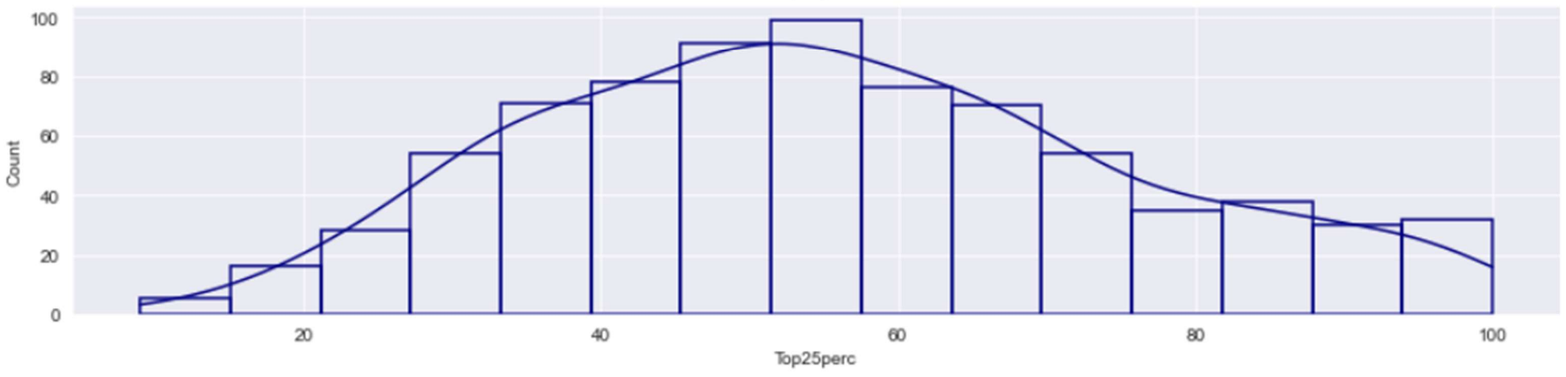
```
Univariate Analysis for column: Top25perc
--------------------------------------------------

count   777.0000
mean     55.7967
std      19.8048
min       9.0000
25%      41.0000
50%      54.0000
75%      69.0000
max     100.0000
Name: Top25perc, dtype: float64
--------------------------------------------------

Top25perc variable does not have any outliers


Skewnes of the column is = 1.410487098842332
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```



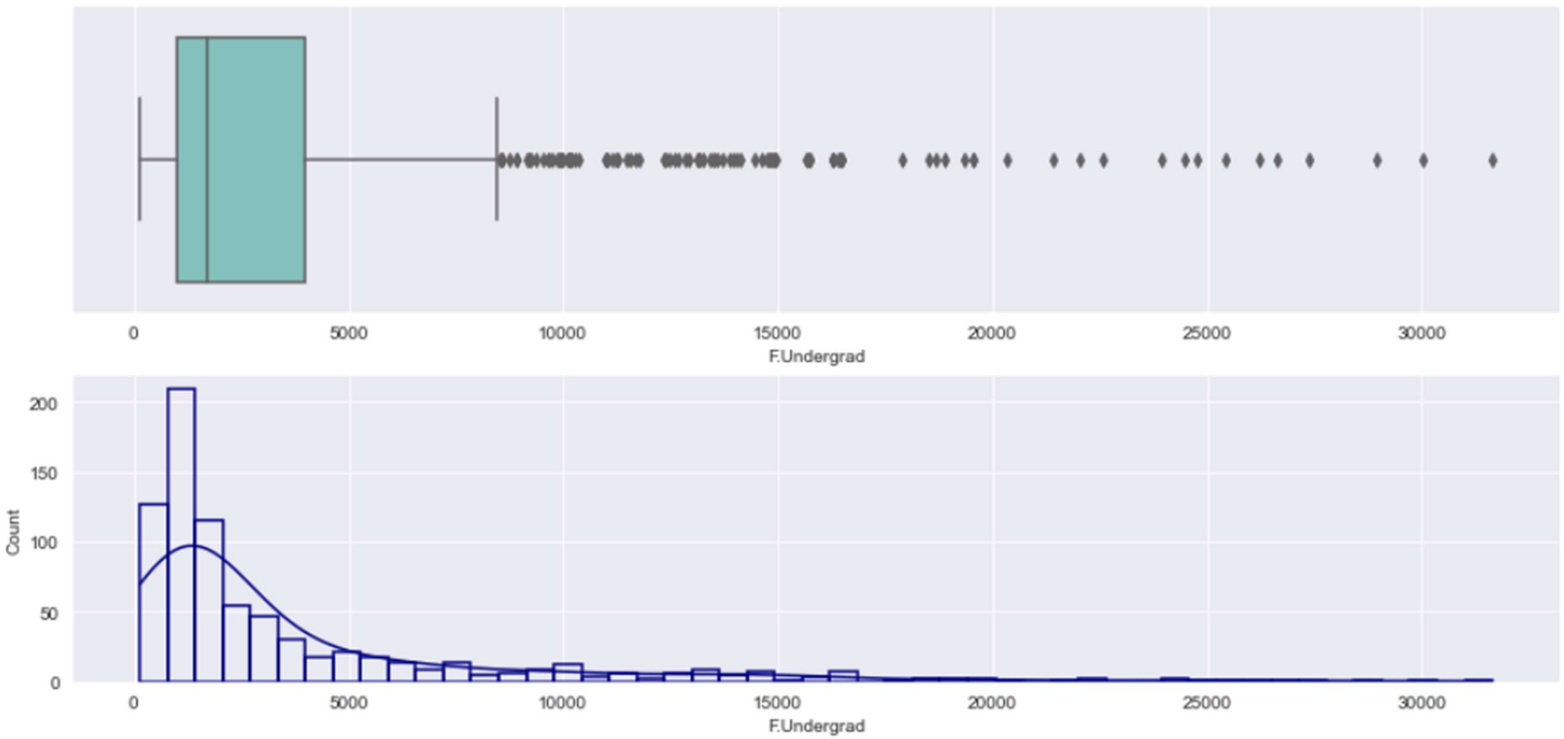


```
Univariate Analysis for column: F.Undergrad
--------------------------------------------------

count      777.0000
mean     3,699.9073
std      4,850.4205
min        139.0000
25%        992.0000
50%      1,707.0000
75%      4,005.0000
max     31,643.0000
Name: F.Undergrad, dtype: float64
--------------------------------------------------

F.Undergrad variable has outliers


Skewnes of the column is = 0.2588394269741162
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```

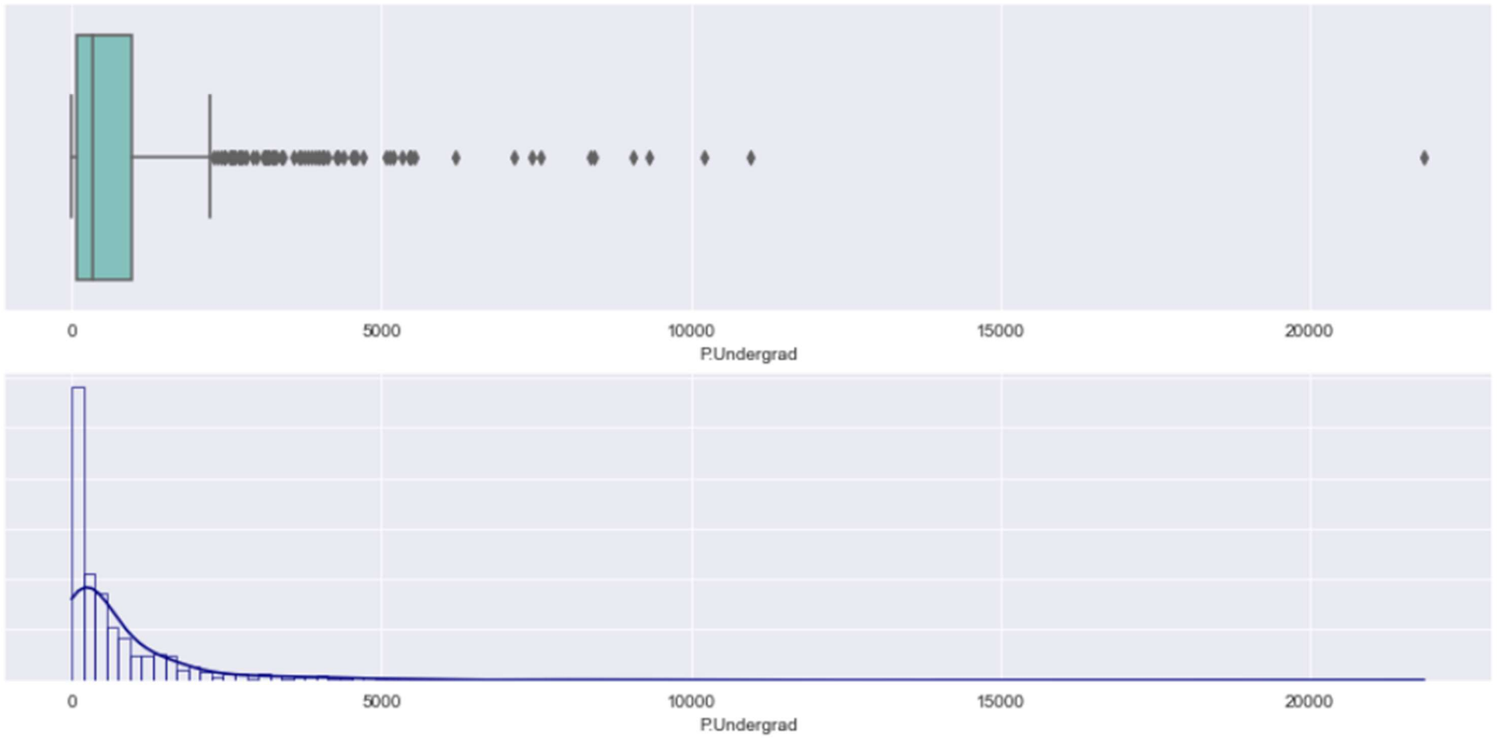


```
Univariate Analysis for column: P.Undergrad
--------------------------------------------------

count      777.0000
mean       855.2986
std      1,522.4319
min          1.0000
25%         95.0000
50%        353.0000
75%        967.0000
max     21,836.0000
Name: P.Undergrad, dtype: float64
--------------------------------------------------

P.Undergrad variable has outliers


Skewnes of the column is = 2.6054157486361564
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```

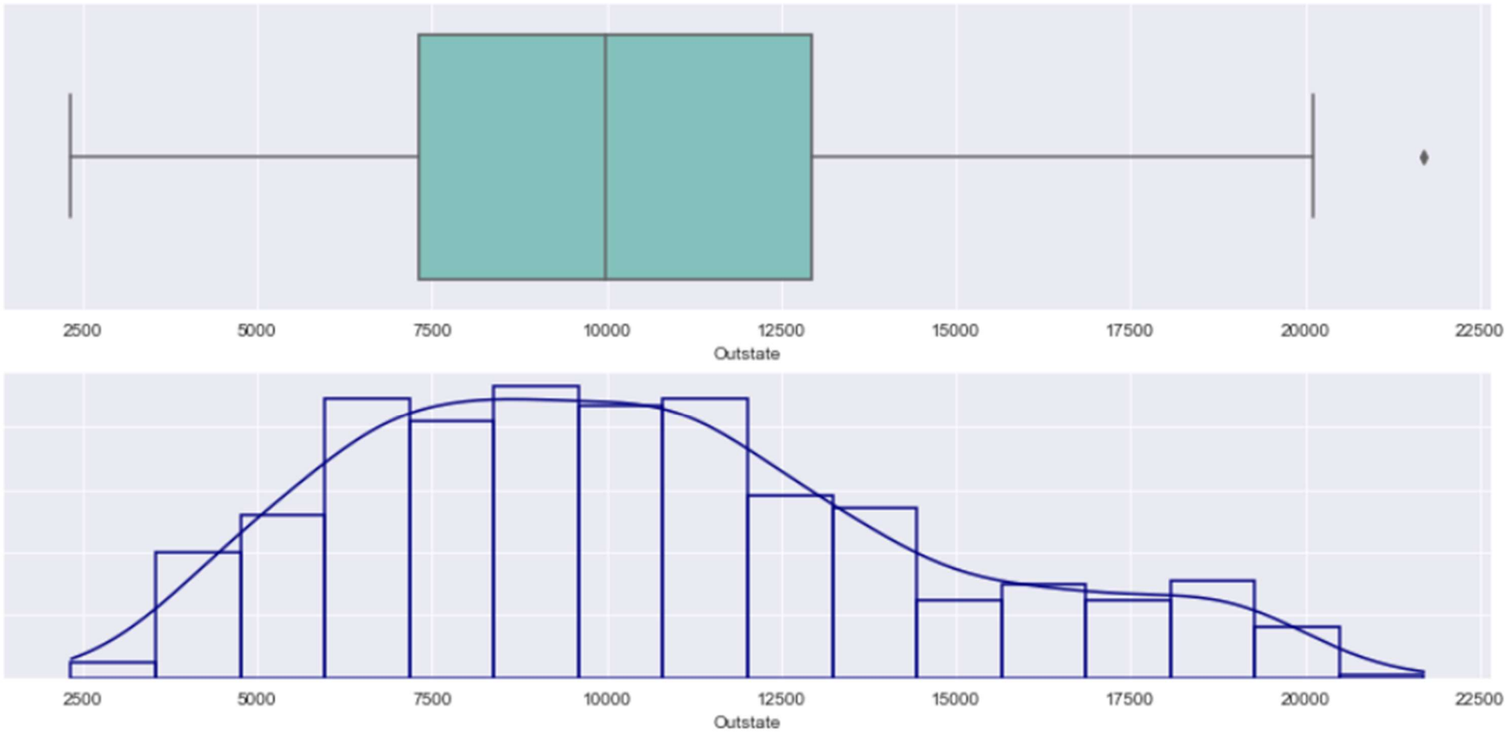


```
Univariate Analysis for column: Outstate
--------------------------------------------------

count      777.0000
mean    10,440.6692
std      4,023.0165
min      2,340.0000
25%      7,320.0000
50%      9,990.0000
75%     12,925.0000
max     21,700.0000
Name: Outstate, dtype: float64
--------------------------------------------------

Outstate variable does not have any outliers


Skewnes of the column is = 5.681358169711681
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```

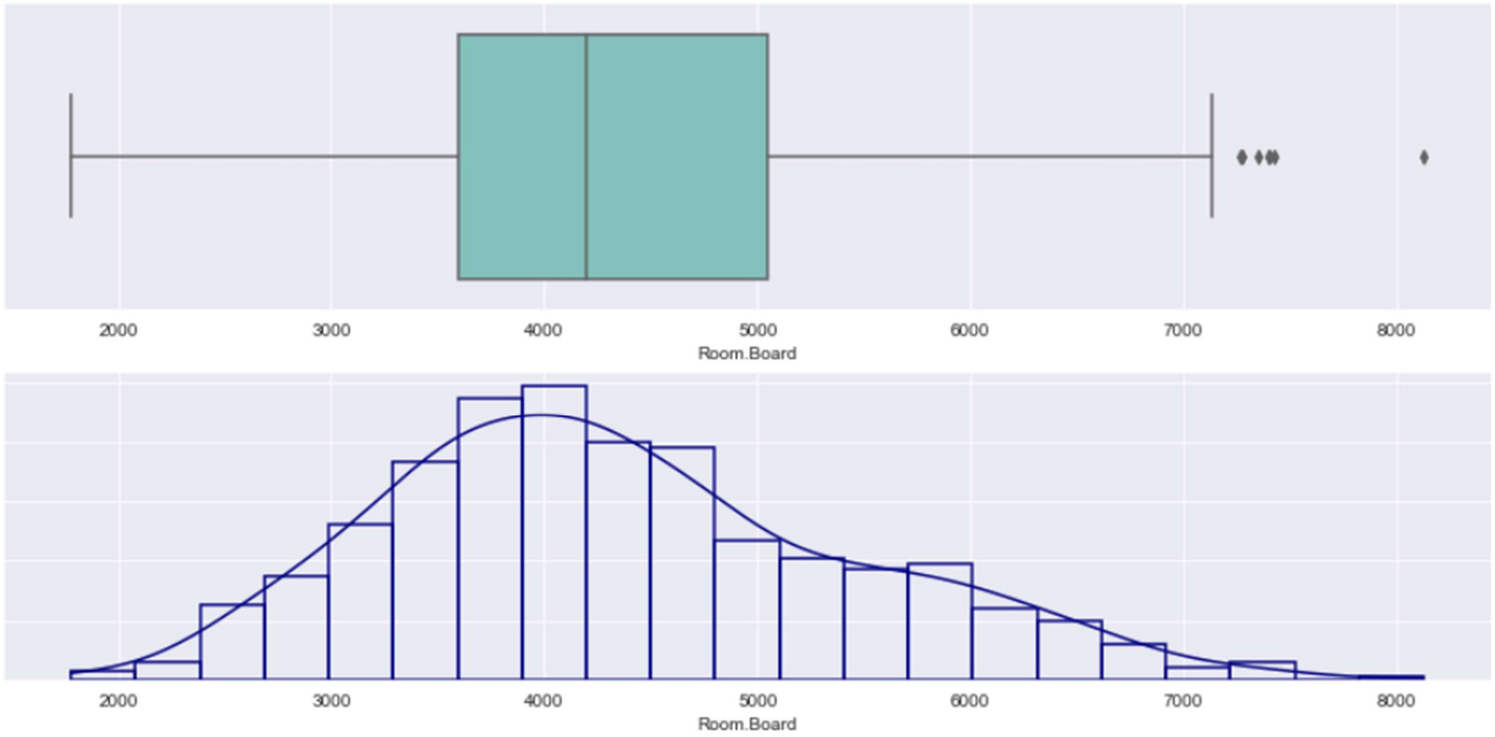


```
Univariate Analysis for column: Room.Board
--------------------------------------------------

count     777.0000
mean    4,357.5264
std     1,096.6964
min     1,780.0000
25%     3,597.0000
50%     4,200.0000
75%     5,050.0000
max     8,124.0000
Name: Room.Board, dtype: float64
--------------------------------------------------

Room.Board variable has outliers


Skewnes of the column is = 0.508294284359404
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```

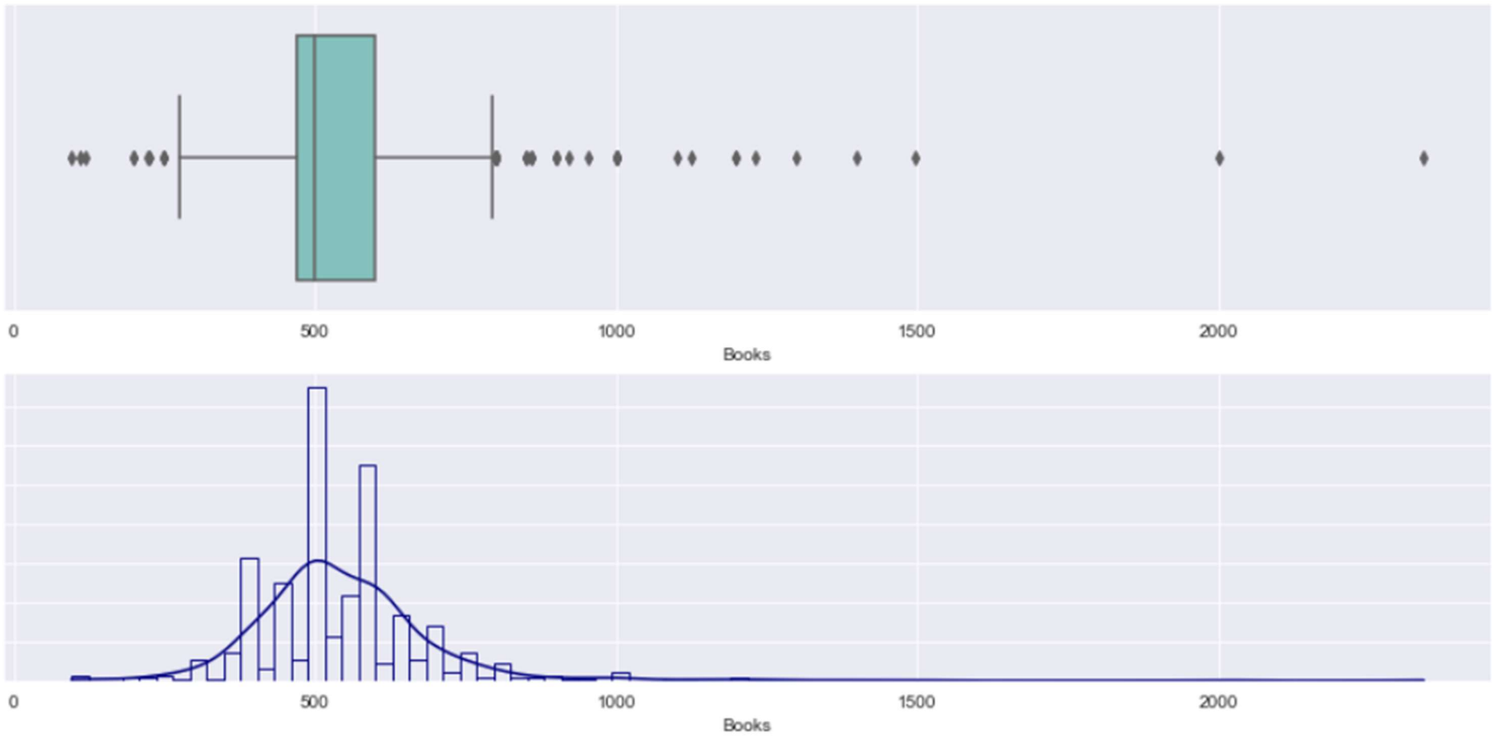


```
Univariate Analysis for column: Books
--------------------------------------------------

count     777.0000
mean      549.3810
std       165.1054
min        96.0000
25%       470.0000
50%       500.0000
75%       600.0000
max     2,340.0000
Name: Books, dtype: float64
--------------------------------------------------

Books variable has outliers


Skewnes of the column is = 0.4764335489968277
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```

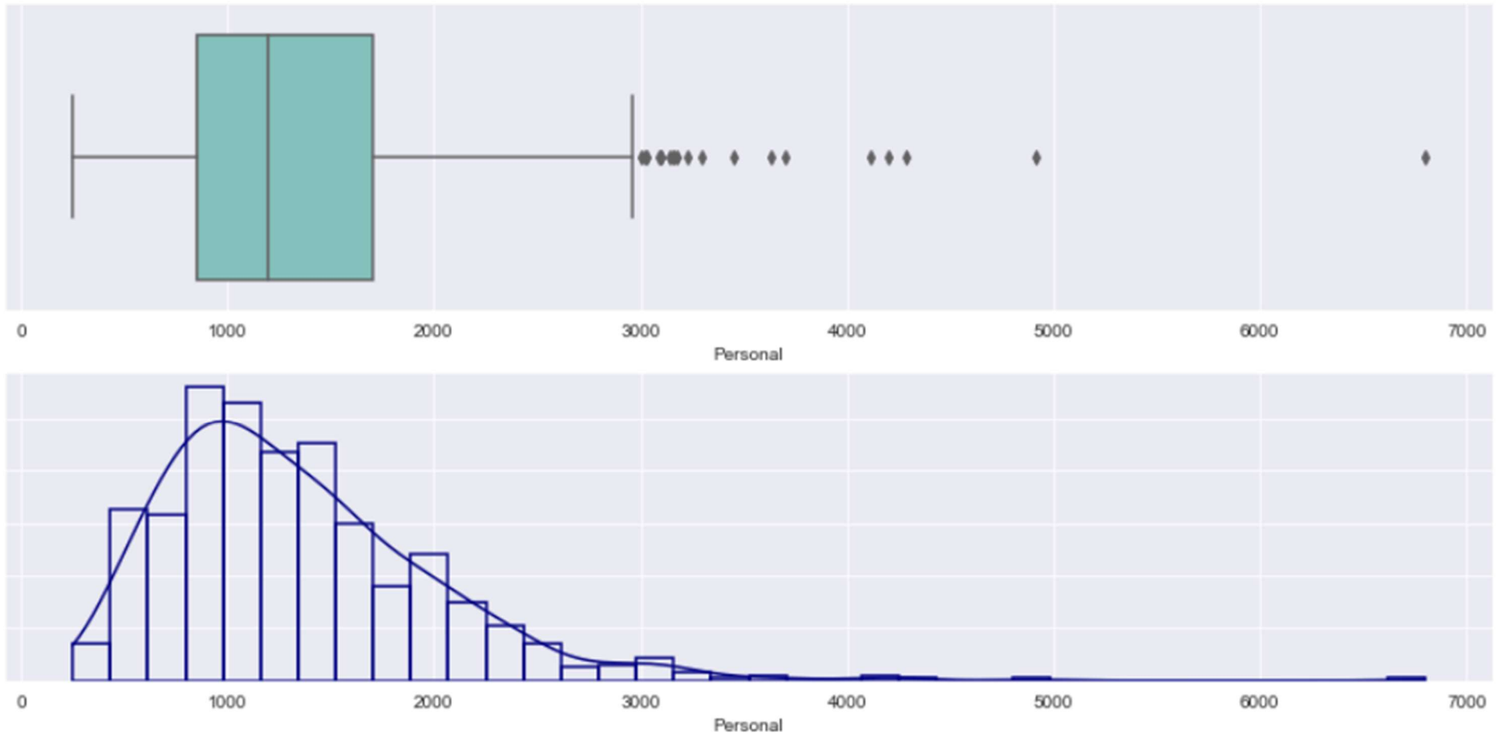


```
Univariate Analysis for column: Personal
--------------------------------------------------

count     777.0000
mean    1,340.6422
std       677.0715
min       250.0000
25%       850.0000
50%     1,200.0000
75%     1,700.0000
max     6,800.0000
Name: Personal, dtype: float64
--------------------------------------------------

Personal variable has outliers


Skewnes of the column is = 3.478293278376379
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```

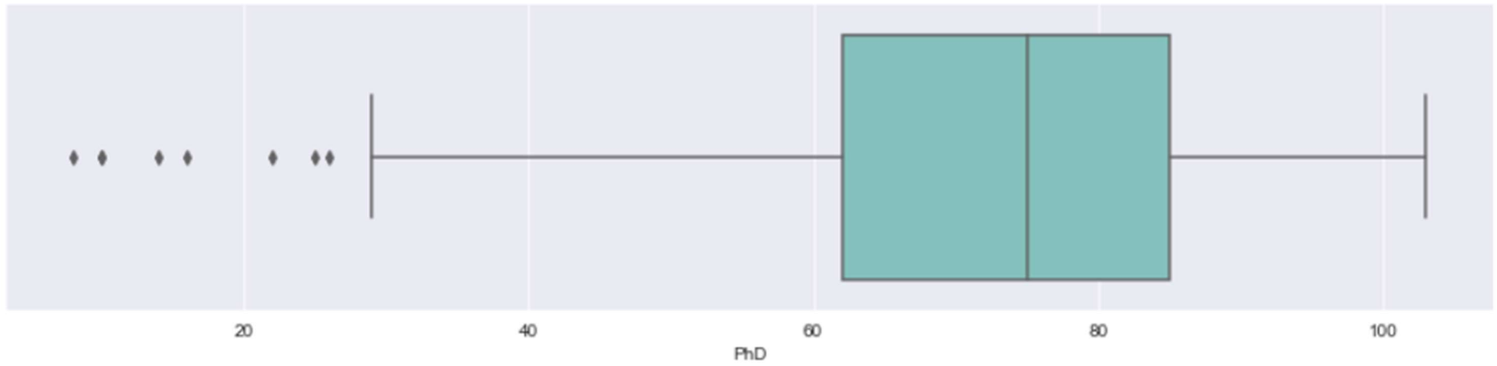


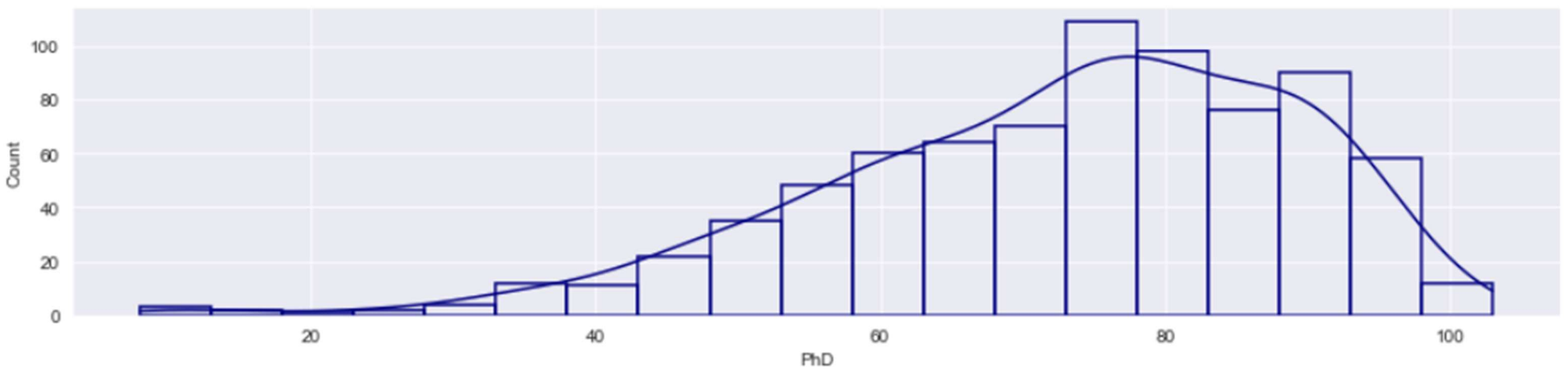
```
Univariate Analysis for column: PhD
--------------------------------------------------

count   777.0000
mean     72.6602
std      16.3282
min       8.0000
25%      62.0000
50%      75.0000
75%      85.0000
max     103.0000
Name: PhD, dtype: float64
--------------------------------------------------

PhD variable has outliers


Skewnes of the column is = 1.7391308384291781
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```



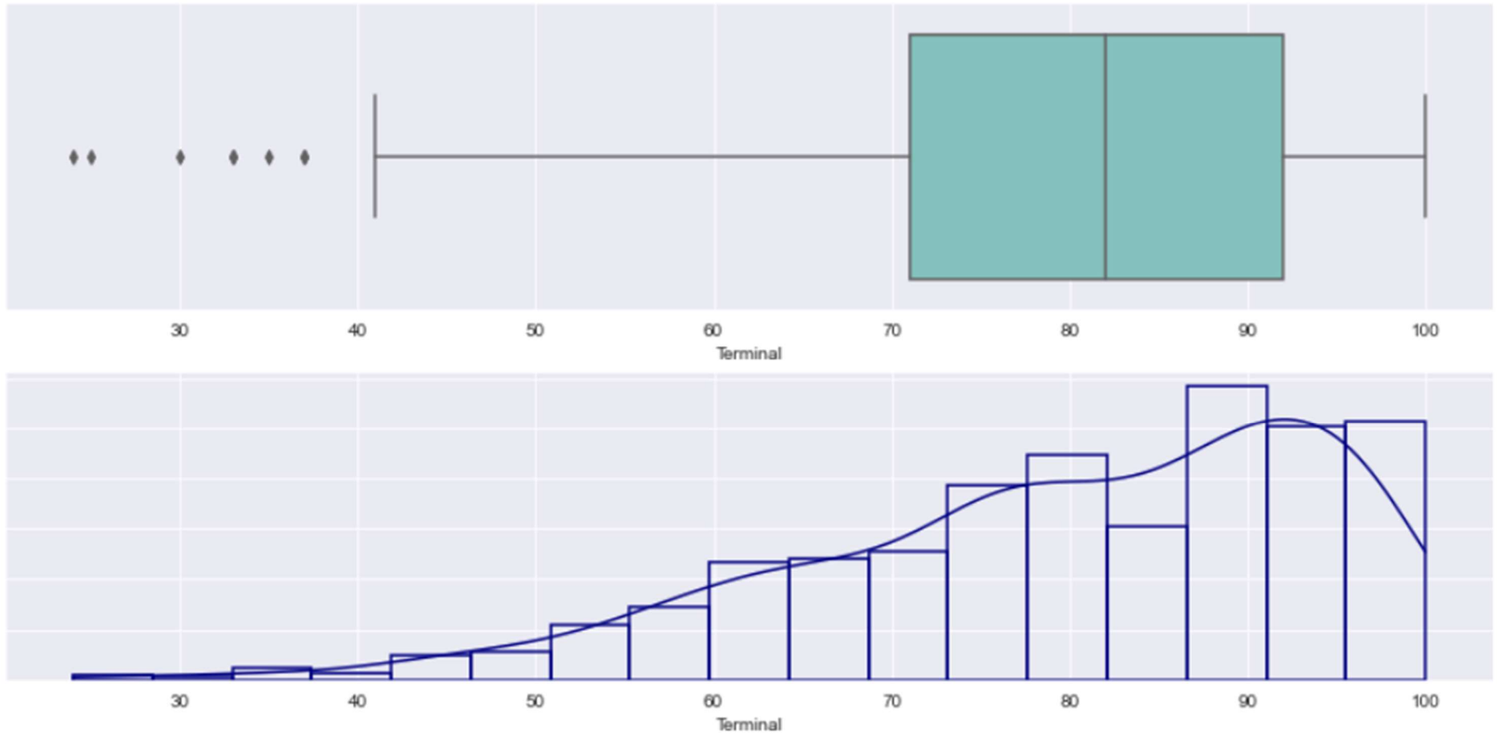


```
Univariate Analysis for column: Terminal
--------------------------------------------------

count   777.0000
mean     79.7027
std      14.7224
min      24.0000
25%      71.0000
50%      82.0000
75%      92.0000
max     100.0000
Name: Terminal, dtype: float64
--------------------------------------------------

Terminal variable has outliers


Skewnes of the column is = -0.7666863621506335
        Hence, the column is negatively skewed or left skewed and is not normally distributed.
```

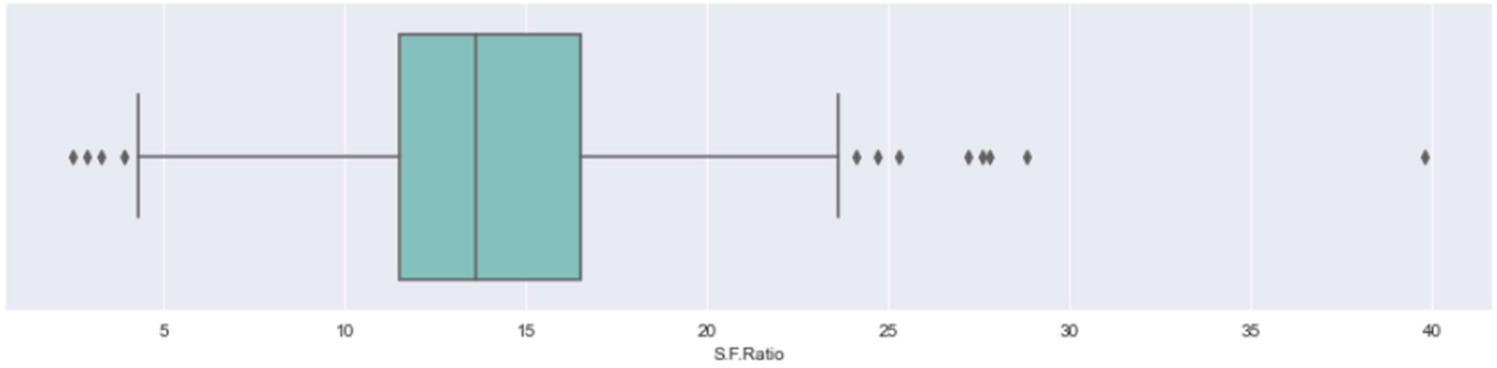


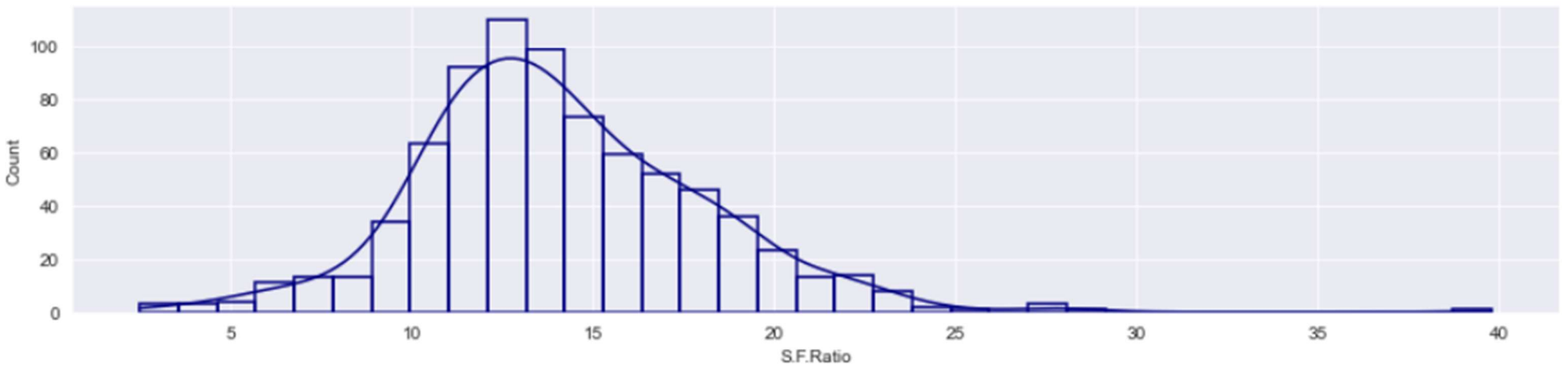
```
Univariate Analysis for column: S.F.Ratio
--------------------------------------------------

count   777.0000
mean     14.0897
std       3.9583
min       2.5000
25%      11.5000
50%      13.6000
75%      16.5000
max      39.8000
Name: S.F.Ratio, dtype: float64
--------------------------------------------------

S.F.Ratio variable has outliers


Skewnes of the column is = -0.8149651536781263
        Hence, the column is negatively skewed or left skewed and is not normally distributed.
```



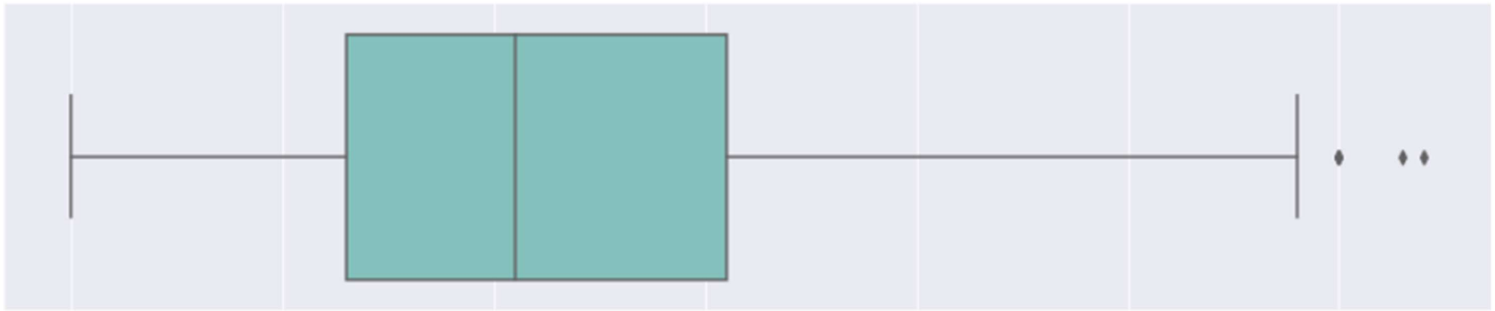


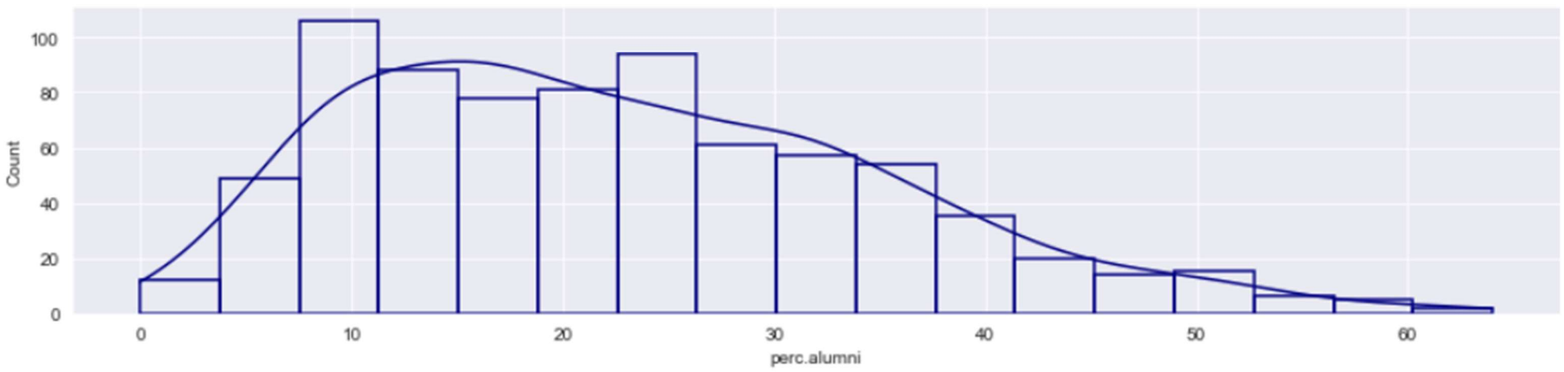
```
Univariate Analysis for column: perc.alumni
--------------------------------------------------

count   777.0000
mean     22.7439
std      12.3918
min       0.0000
25%      13.0000
50%      21.0000
75%      31.0000
max      64.0000
Name: perc.alumni, dtype: float64
--------------------------------------------------

perc.alumni variable has outliers


Skewnes of the column is = 0.6661461873546756
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```



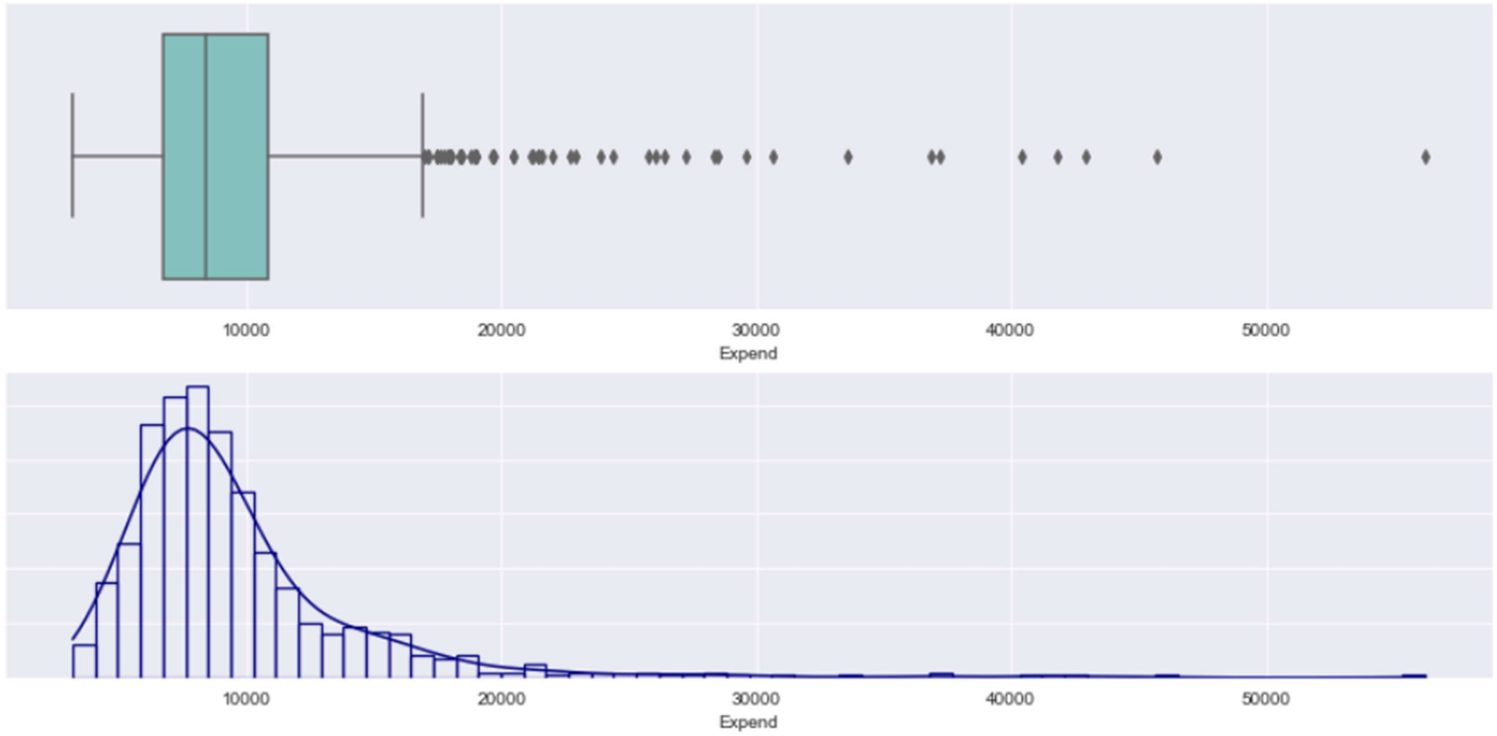


```
Univariate Analysis for column: Expend
--------------------------------------------------

count      777.0000
mean     9,660.1712
std      5,221.7684
min      3,186.0000
25%      6,751.0000
50%      8,377.0000
75%     10,830.0000
max     56,233.0000
Name: Expend, dtype: float64
--------------------------------------------------

Expend variable has outliers


Skewnes of the column is = 0.6057189848601131
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```

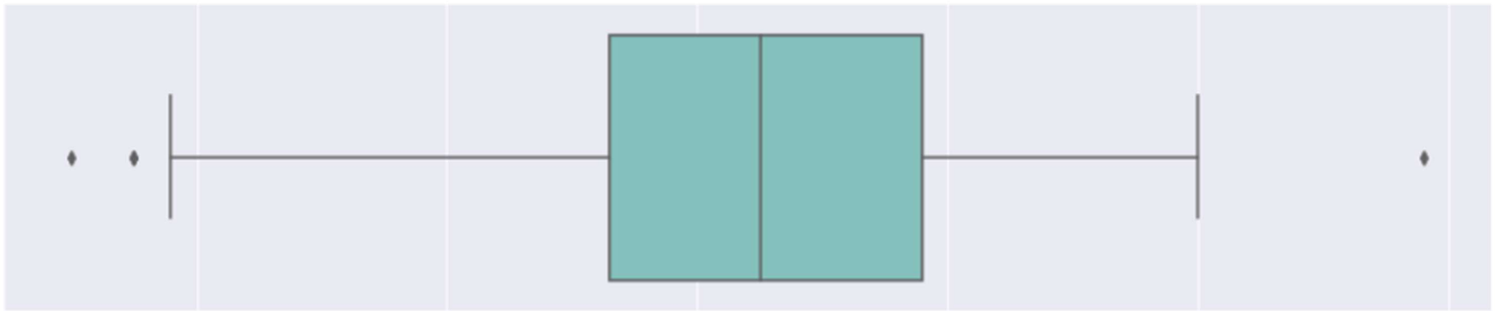


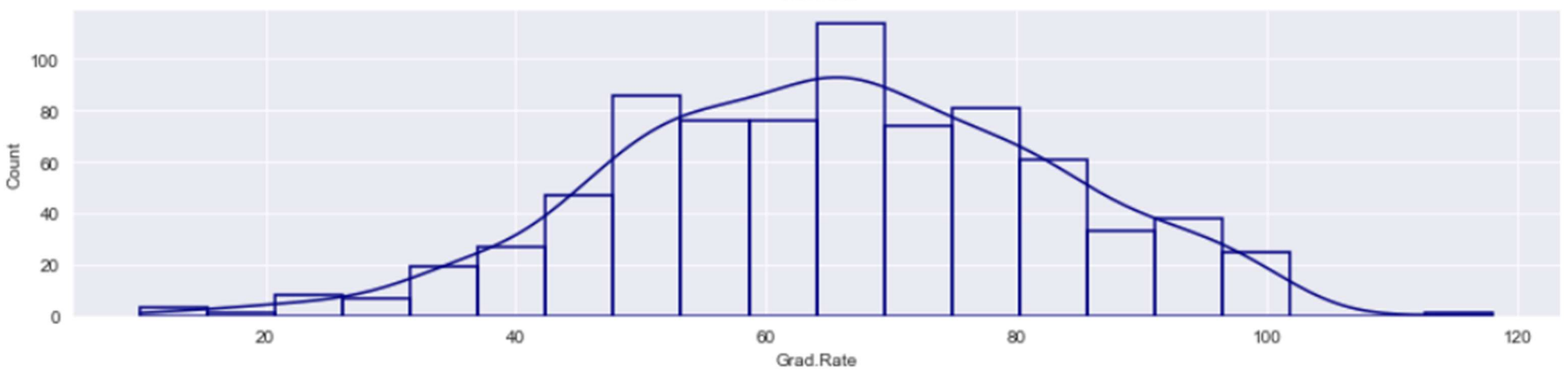
```
Univariate Analysis for column: Grad.Rate
--------------------------------------------------

count   777.0000
mean     65.4633
std      17.1777
min      10.0000
25%      53.0000
50%      65.0000
75%      78.0000
max     118.0000
Name: Grad.Rate, dtype: float64
--------------------------------------------------

Grad.Rate variable has outliers


Skewnes of the column is = 3.4526399033472197
        Hence, the column is positively skewed or right skewed and is not normally distributed.
```





Analysis

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data.

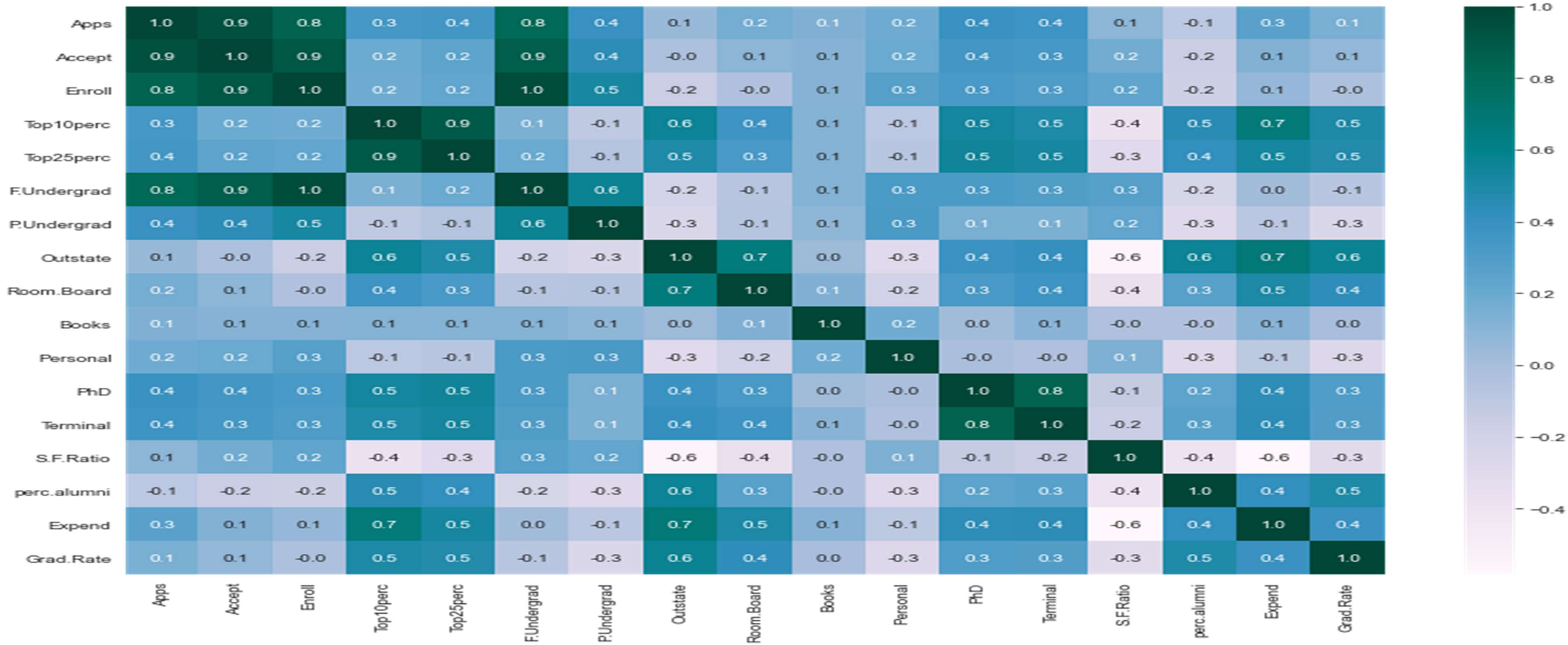
> All the numeric variables are skewed and hence, are not normally distributed.

While calculating the Z-score we re-scale and center the data and look for data points which are too far from zero. These data points which are way too far from zero will be treated as the outliers. In most of the cases a threshold of 3 or -3 is used i.e if the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers.

> Based on z-scores, all the variables have outliers except Top25perc and Outstate.

Bivariate Analysis

Numerical against Numerical



Analysis

> Apps, Accept and Enroll has a high and positive correlation with F.Undergrad.

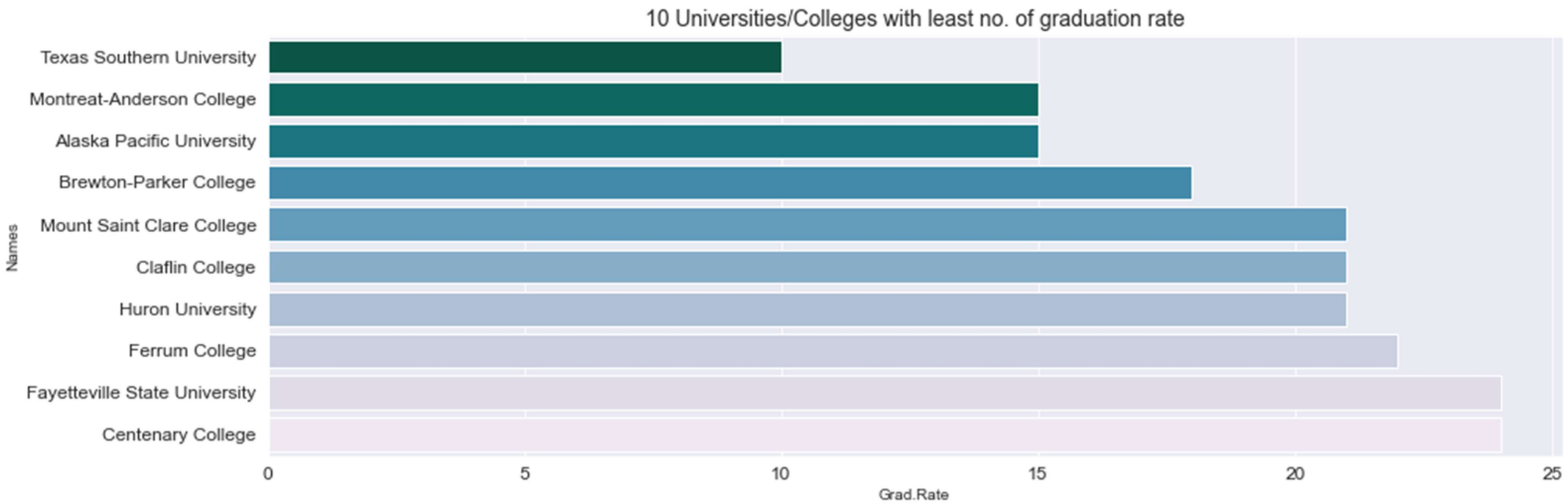
> Outstate has a high correlation with Expend, Room Board and Top10perc. Top10perc is also highly correlated with Top25perc.

> Outstate and S.F.Ratio (student faculty ratio) are negatively correlated. S.F Ratio is also negatively correlated with Top10perc, Room Board and Expend.

Bivariate Analysis

Numerical against Categorical

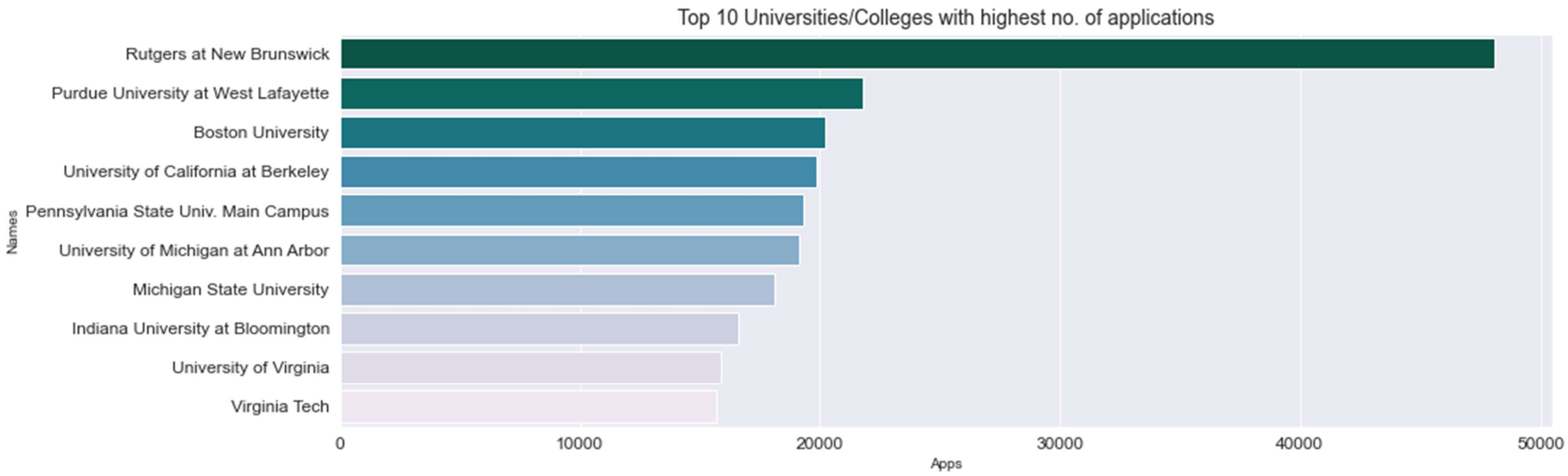
10 Universities or Colleges with least number of Graduation rate



Analysis

Texas Southern University has the least number of graduation rate followed by Montreat-Anderson College and Alaska Pacific University.

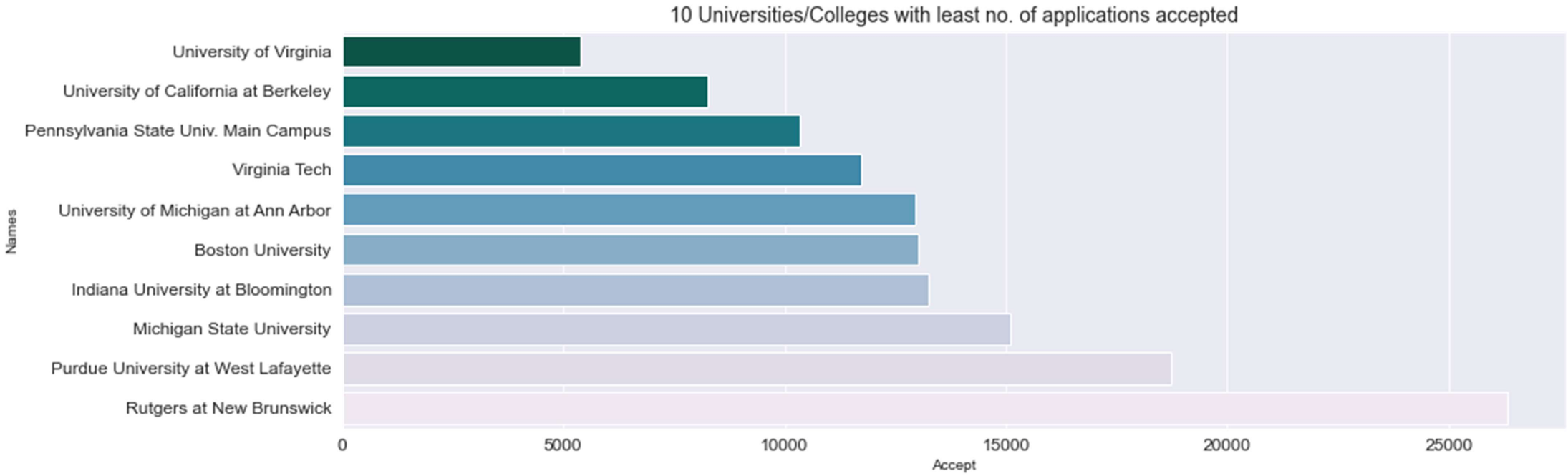
Top 10 Universities or Colleges with highest applications received



Analysis

`Rutgers at New Brunswick` is the leading university with the highest of 48094 number of applications received. The second leading university `Purdue University at West Lafayette` has 21804 number of applications which is approx half of Rutgers at New Brunswick

10 Universities/Colleges with least no. of applications accepted

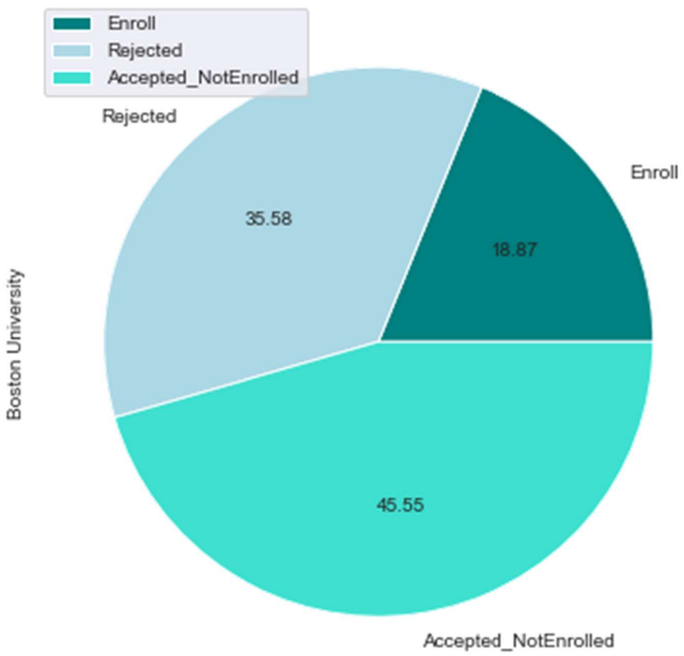


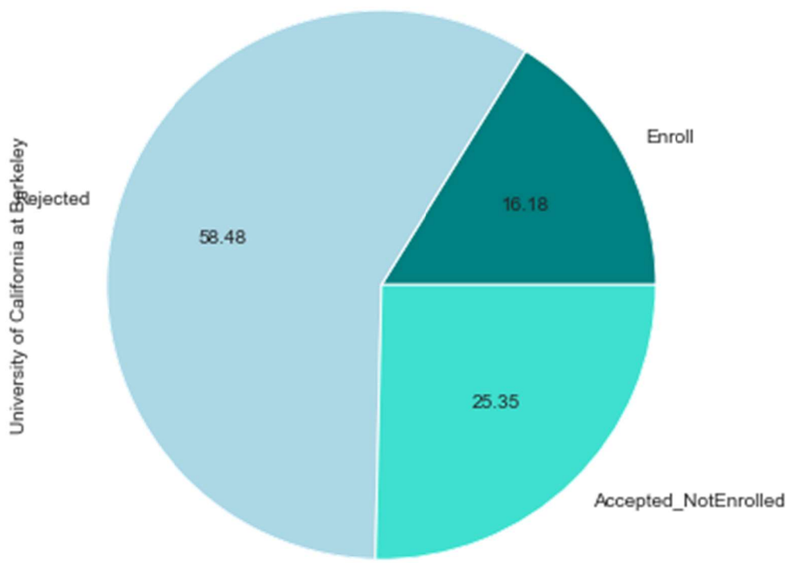
Analysis

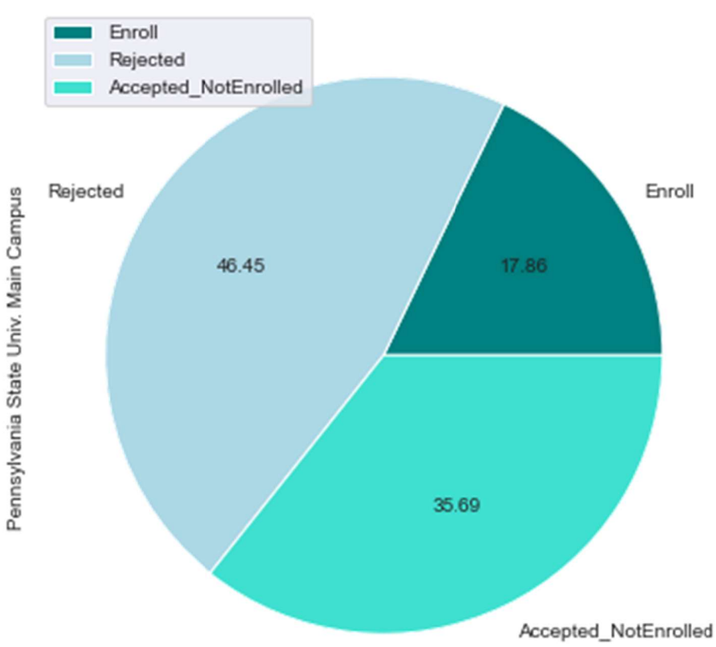
`University of Virginia` has received 15849 applications outof which only 5384 were accepted i.e. only 33.97% of the total applications received was accepted.

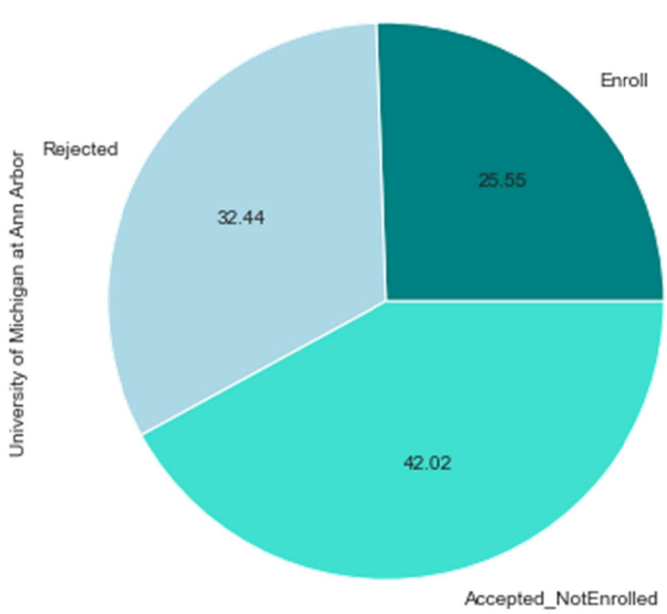
Multivariate Analysis

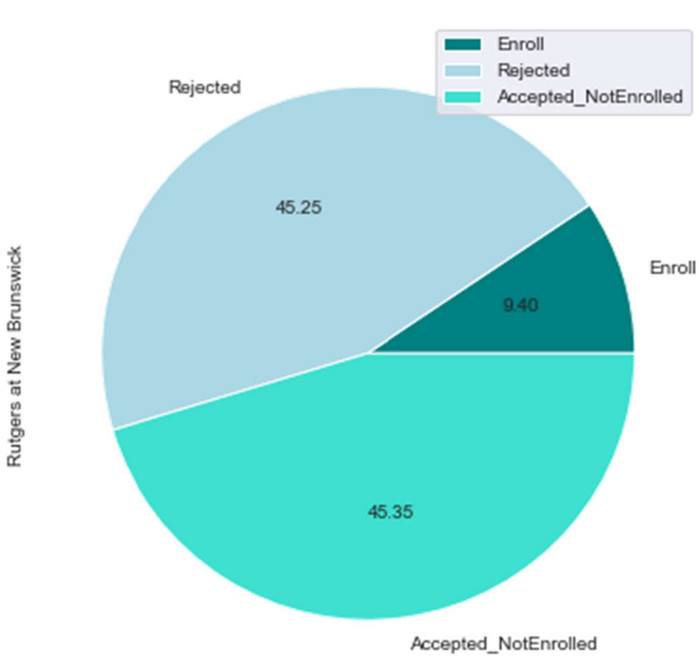
Top universities with highest no. of applications Rejected, Enrolled and Accepted but not enrolled

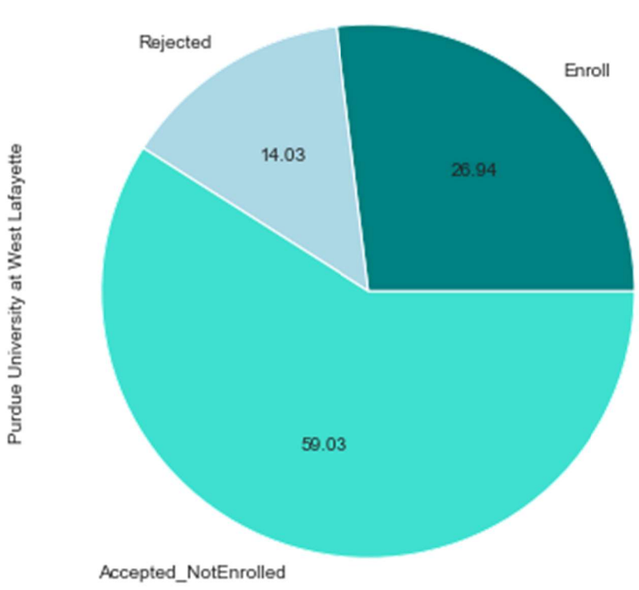

Enroll
Rejected
Accepted_NotEnrolled
Boston University
Rejected
Enroll
35.58
18.87
45.55
Accepted_NotEnrolled


University of California at Berkeley
Rejected
Enroll
58.48
16.18
25.35
Accepted_NotEnrolled


Enroll
Rejected
Accepted_NotEnrolled
Pennsylvania State Univ. Main Campus
Rejected
Enroll
46.45
17.86
35.69
Accepted_NotEnrolled


University of Michigan at Ann Arbor
Rejected
Enroll
32.44
25.55
42.02
Accepted_NotEnrolled


Enroll
Rejected
Accepted_NotEnrolled
Rutgers at New Brunswick
Rejected
Enroll
45.25
9.40
45.35
Accepted_NotEnrolled


Purdue University at West Lafayette
Rejected
Enroll
14.03
26.94
59.03
Accepted_NotEnrolled

Analyses

Rutgers at New Brunswick - Out of total applications received: only 9.4% students were enrolled, 45.25% were rejected and 45.35% were accecpted but the students did not enrolled.

Purdue University at West Lafayette - Out of total applications received: only 26.94% students were enrolled, 14.03% were rejected and 59.03% were accecpted but the students did not enrolled.

Pennsylvania State Univ. Main Campus - Out of total applications received: only 17.86% students were enrolled, 46.45% were rejected and 35.69% were accecpted but the students did not enrolled.

University of Michigan at Ann Arbor - Out of total applications received: only 25.55% students were enrolled, 32.44% were rejected and 42.02% were accecpted but the students did not enrolled.

Boston University - Out of total applications received: only 18.87% students were enrolled, 35.58% were rejected and 45.55% were accecpted but the students did not enrolled.

University of California at Berkeley - Out of total applications received: only 16.18% students were enrolled, 58.48% were rejected and 25.35% were accecpted but the students did not enrolled.

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

The goal of scaling is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.

PCA is necessary to normalize the data. The PCA calculates a new projection of the data set. If one component (e.g. meters) varies less than another does (e.g. kg) because of their respective scales (meters vs. kilos), PCA might determine that the direction of maximal variance more closely corresponds with the 'kilos' axis, if those features are not scaled. Thus leading to incorrect results.

In this case, perc.alumni, Terminal, Top10perc and Top25perc are in percetage. S.F Ratio is the student/faculty ratio and most the columns are the total number of sudents for a particular column. This alone concludes that the variables of the data are not of same scale. Hence, scaling is necessary for PCA in this case.

Scaling

	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	-0.3469	-0.3212	-0.0635	-0.2586	-0.1918	-0.1681	-0.2092	-0.7464	-0.9649	-0.6023	1.2700	-0.1630	-0.1157	1.0138	-0.8676	-0.5019	-0.3183
1	-0.2109	-0.0387	-0.2886	-0.6557	-1.3539	-0.2098	0.2443	0.4575	1.9092	1.2159	0.2355	-2.6756	-3.3782	-0.4777	-0.5446	0.1661	-0.5513
2	-0.4069	-0.3763	-0.4781	-0.3153	-0.2929	-0.5496	-0.4971	0.2013	-0.5543	-0.9053	-0.2596	-1.2048	-0.9313	-0.3007	0.5859	-0.1773	-0.6678
3	-0.6683	-0.6817	-0.6924	1.8402	1.6776	-0.6581	-0.5208	0.6266	0.9968	-0.6023	-0.6882	1.1852	1.1757	-1.6153	1.1512	1.7929	-0.3765
4	-0.7262	-0.7646	-0.7807	-0.6557	-0.5960	-0.7119	0.0090	-0.7165	-0.2167	1.5189	0.2355	0.2047	-0.5235	-0.5535	-1.6751	0.2418	-2.9396

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

Correlation Matrix

	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	1.0000	0.9435	0.8468	0.3388	0.3516	0.8145	0.3983	0.0502	0.1649	0.1326	0.1787	0.3907	0.3695	0.0956	-0.0902	0.2596	0.1468
Accept	0.9435	1.0000	0.9116	0.1924	0.2475	0.8742	0.4413	-0.0258	0.0909	0.1135	0.2010	0.3558	0.3376	0.1762	-0.1600	0.1247	0.0673
Enroll	0.8468	0.9116	1.0000	0.1813	0.2267	0.9646	0.5131	-0.1555	-0.0402	0.1127	0.2809	0.3315	0.3083	0.2373	-0.1808	0.0642	-0.0223
Top10perc	0.3388	0.1924	0.1813	1.0000	0.8920	0.1413	-0.1054	0.5623	0.3715	0.1189	-0.0933	0.5318	0.4911	-0.3849	0.4555	0.6609	0.4950
Top25perc	0.3516	0.2475	0.2267	0.8920	1.0000	0.1994	-0.0536	0.4894	0.3315	0.1155	-0.0808	0.5459	0.5247	-0.2946	0.4179	0.5274	0.4773
F.Undergrad	0.8145	0.8742	0.9646	0.1413	0.1994	1.0000	0.5705	-0.2157	-0.0689	0.1155	0.3172	0.3183	0.3000	0.2797	-0.2295	0.0187	-0.0788
P.Undergrad	0.3983	0.4413	0.5131	-0.1054	-0.0536	0.5705	1.0000	-0.2535	-0.0613	0.0812	0.3199	0.1491	0.1419	0.2325	-0.2808	-0.0836	-0.2570
Outstate	0.0502	-0.0258	-0.1555	0.5623	0.4894	-0.2157	-0.2535	1.0000	0.6543	0.0389	-0.2991	0.3830	0.4080	-0.5548	0.5663	0.6728	0.5713
Room.Board	0.1649	0.0909	-0.0402	0.3715	0.3315	-0.0689	-0.0613	0.6543	1.0000	0.1280	-0.1994	0.3292	0.3745	-0.3626	0.2724	0.5017	0.4249
Books	0.1326	0.1135	0.1127	0.1189	0.1155	0.1155	0.0812	0.0389	0.1280	1.0000	0.1793	0.0269	0.1000	-0.0319	-0.0402	0.1124	0.0011
Personal	0.1787	0.2010	0.2809	-0.0933	-0.0808	0.3172	0.3199	-0.2991	-0.1994	0.1793	1.0000	-0.0109	-0.0306	0.1363	-0.2860	-0.0979	-0.2693
PhD	0.3907	0.3558	0.3315	0.5318	0.5459	0.3183	0.1491	0.3830	0.3292	0.0269	-0.0109	1.0000	0.8496	-0.1305	0.2490	0.4328	0.3050
Terminal	0.3695	0.3376	0.3083	0.4911	0.5247	0.3000	0.1419	0.4080	0.3745	0.1000	-0.0306	0.8496	1.0000	-0.1601	0.2671	0.4388	0.2895
S.F.Ratio	0.0956	0.1762	0.2373	-0.3849	-0.2946	0.2797	0.2325	-0.5548	-0.3626	-0.0319	0.1363	-0.1305	-0.1601	1.0000	-0.4029	-0.5838	-0.3067
perc.alumni	-0.0902	-0.1600	-0.1808	0.4555	0.4179	-0.2295	-0.2808	0.5663	0.2724	-0.0402	-0.2860	0.2490	0.2671	-0.4029	1.0000	0.4177	0.4909
Expend	0.2596	0.1247	0.0642	0.6609	0.5274	0.0187	-0.0836	0.6728	0.5017	0.1124	-0.0979	0.4328	0.4388	-0.5838	0.4177	1.0000	0.3903
Grad.Rate	0.1468	0.0673	-0.0223	0.4950	0.4773	-0.0788	-0.2570	0.5713	0.4249	0.0011	-0.2693	0.3050	0.2895	-0.3067	0.4909	0.3903	1.0000

Covariance Matrix

	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	1.0013	0.9447	0.8479	0.3393	0.3521	0.8155	0.3988	0.0502	0.1652	0.1327	0.1790	0.3912	0.3700	0.0958	-0.0903	0.2599	0.1469
Accept	0.9447	1.0013	0.9128	0.1927	0.2478	0.8753	0.4418	-0.0258	0.0910	0.1137	0.2012	0.3562	0.3380	0.1765	-0.1602	0.1249	0.0674
Enroll	0.8479	0.9128	1.0013	0.1815	0.2270	0.9659	0.5137	-0.1557	-0.0403	0.1129	0.2813	0.3319	0.3087	0.2376	-0.1810	0.0643	-0.0224
Top10perc	0.3393	0.1927	0.1815	1.0013	0.8931	0.1415	-0.1055	0.5631	0.3720	0.1190	-0.0934	0.5325	0.4918	-0.3854	0.4561	0.6618	0.4956
Top25perc	0.3521	0.2478	0.2270	0.8931	1.0013	0.1997	-0.0536	0.4900	0.3319	0.1157	-0.0809	0.5466	0.5254	-0.2950	0.4184	0.5281	0.4779
F.Undergrad	0.8155	0.8753	0.9659	0.1415	0.1997	1.0013	0.5712	-0.2160	-0.0690	0.1157	0.3176	0.3187	0.3004	0.2801	-0.2298	0.0187	-0.0789
P.Undergrad	0.3988	0.4418	0.5137	-0.1055	-0.0536	0.5712	1.0013	-0.2538	-0.0614	0.0813	0.3203	0.1493	0.1421	0.2328	-0.2812	-0.0837	-0.2573
Outstate	0.0502	-0.0258	-0.1557	0.5631	0.4900	-0.2160	-0.2538	1.0013	0.6551	0.0389	-0.2995	0.3835	0.4085	-0.5555	0.5670	0.6736	0.5720
Room.Board	0.1652	0.0910	-0.0403	0.3720	0.3319	-0.0690	-0.0614	0.6551	1.0013	0.1281	-0.1997	0.3296	0.3750	-0.3631	0.2727	0.5024	0.4255
Books	0.1327	0.1137	0.1129	0.1190	0.1157	0.1157	0.0813	0.0389	0.1281	1.0013	0.1795	0.0269	0.1001	-0.0320	-0.0403	0.1126	0.0011
Personal	0.1790	0.2012	0.2813	-0.0934	-0.0809	0.3176	0.3203	-0.2995	-0.1997	0.1795	1.0013	-0.0109	-0.0307	0.1365	-0.2863	-0.0980	-0.2697
PhD	0.3912	0.3562	0.3319	0.5325	0.5466	0.3187	0.1493	0.3835	0.3296	0.0269	-0.0109	1.0013	0.8507	-0.1307	0.2493	0.4333	0.3054
Terminal	0.3700	0.3380	0.3087	0.4918	0.5254	0.3004	0.1421	0.4085	0.3750	0.1001	-0.0307	0.8507	1.0013	-0.1603	0.2675	0.4394	0.2899
S.F.Ratio	0.0958	0.1765	0.2376	-0.3854	-0.2950	0.2801	0.2328	-0.5555	-0.3631	-0.0320	0.1365	-0.1307	-0.1603	1.0013	-0.4034	-0.5846	-0.3071
perc.alumni	-0.0903	-0.1602	-0.1810	0.4561	0.4184	-0.2298	-0.2812	0.5670	0.2727	-0.0403	-0.2863	0.2493	0.2675	-0.4034	1.0013	0.4183	0.4915
Expend	0.2599	0.1249	0.0643	0.6618	0.5281	0.0187	-0.0837	0.6736	0.5024	0.1126	-0.0980	0.4333	0.4394	-0.5846	0.4183	1.0013	0.3908
Grad.Rate	0.1469	0.0674	-0.0224	0.4956	0.4779	-0.0789	-0.2573	0.5720	0.4255	0.0011	-0.2697	0.3054	0.2899	-0.3071	0.4915	0.3908	1.0013

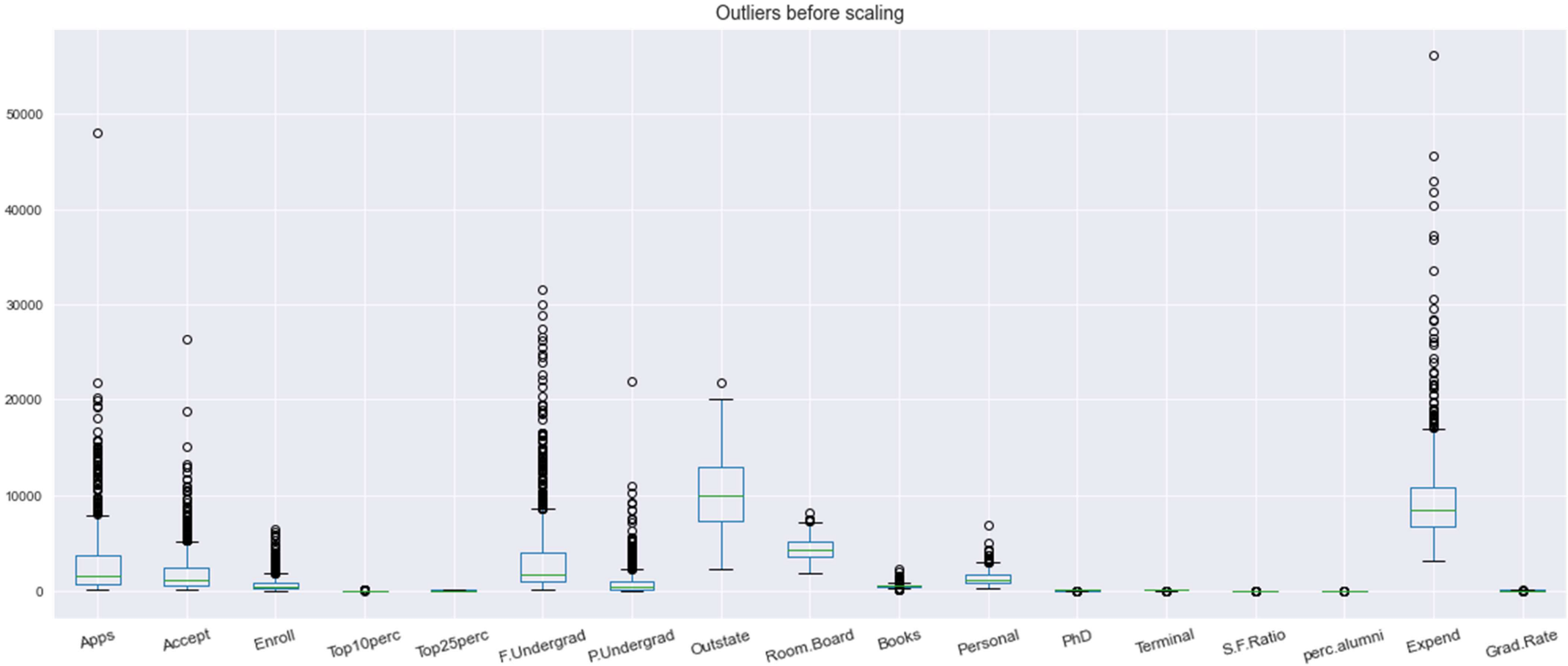
Interpretation

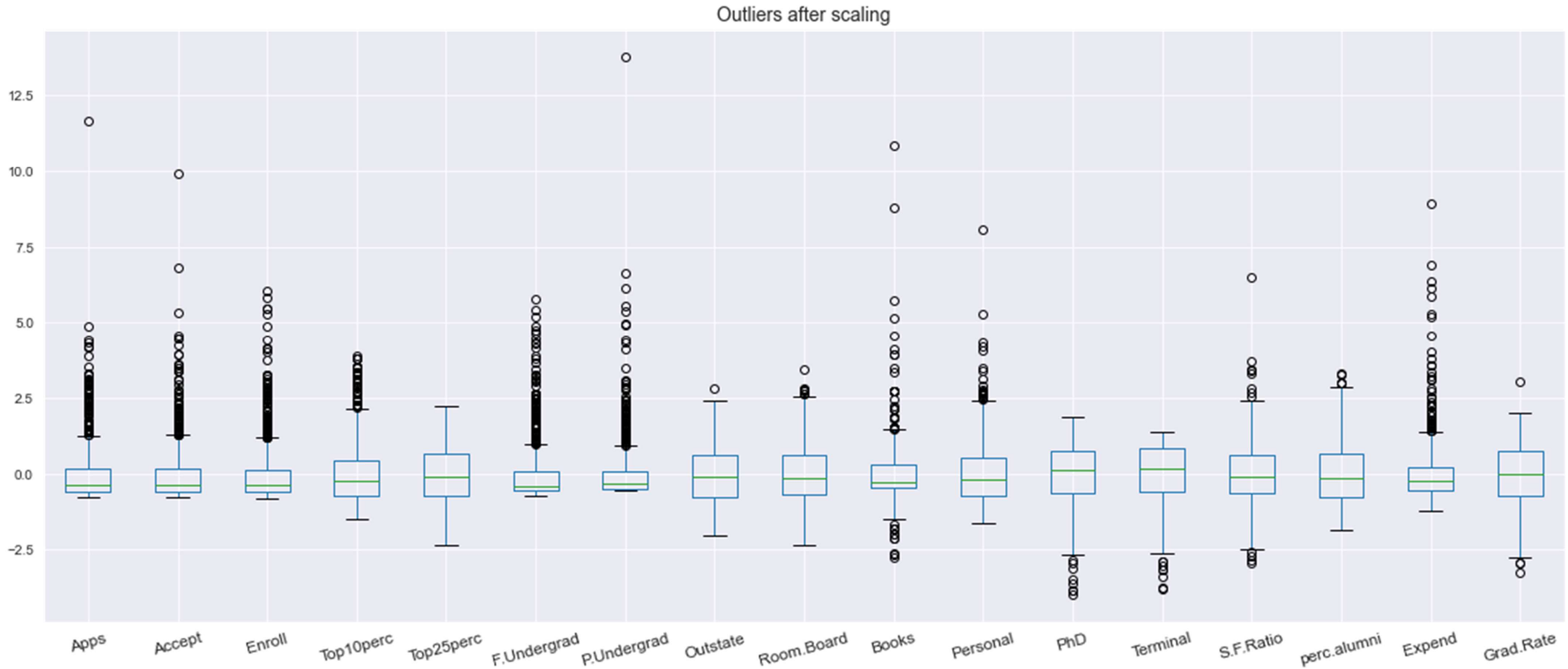
Both Correlation and Covariance are very closely related to each other and yet they differ a lot.

A covariance matrix is used to study the direction of the linear relationship between variables. A correlation matrix is used to study the strength of a relationship between two variables.

Covariance vs Correlation, the latter stands to be the most preferred choice as it remains unaffected by the change in dimensions, location, and scale, and can also be used to make a comparison between two pairs of variables. Since it is limited to a range of -1 to +1, it is useful to draw comparisons between variables across domains. However, an important limitation is that both these concepts measure the only linear relationship.

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


Outliers before scaling
50000
40000
30000
20000
10000
0
Apps
Accept
Enroll
Top10perc
Top25perc
F.Undergrad
P.Undergrad
Outstate
Room.Board
Books
Personal
PhD
Terminal
S.F.Ratio
perc.alumni
Expend
Grad.Rate



Insights

Before scaling the variables were on different scale as we can see in the figure *Outliers before scaling*. The scaling shrinked the range of the feature values and median is now close to 0 as shown in figure *Outliers after scaling*.

2.5 Extract the eigenvalues and eigenvectors. [print both]

```
Eigen Values:
[5.45052162 4.48360686 1.17466761 1.00820573 0.93423123 0.84849117
 0.6057878  0.58787222 0.53061262 0.4043029  0.02302787 0.03672545
 0.31344588 0.08802464 0.1439785  0.16779415 0.22061096]

Eigen Vectors:
[[-2.48765602e-01 -2.07601502e-01 -1.76303592e-01 -3.54273947e-01
  -3.44001279e-01 -1.54640962e-01 -2.64425045e-02 -2.94736419e-01
  -2.49030449e-01 -6.47575181e-02  4.25285386e-02 -3.18312875e-01
```

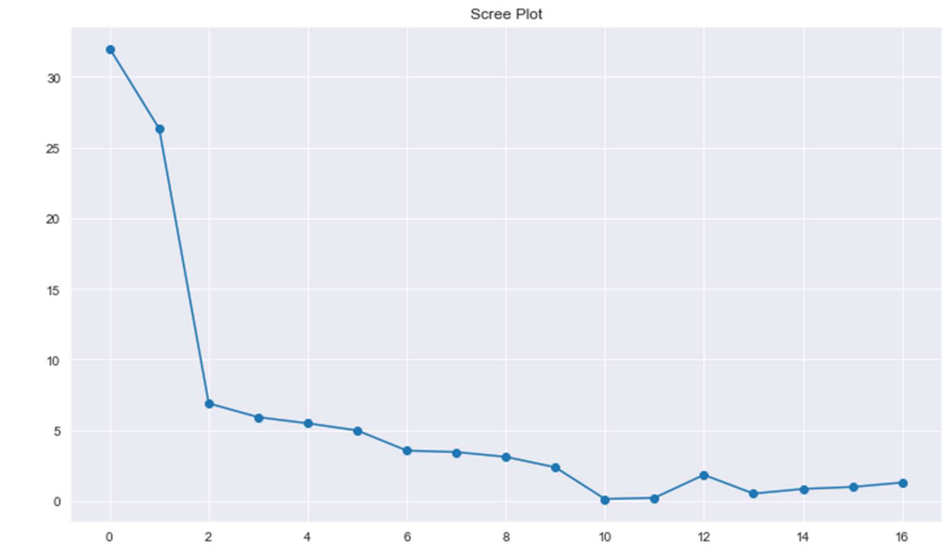
```
 -3.17056016e-01  1.76957895e-01 -2.05082369e-01 -3.18908750e-01
 -2.52315654e-01]
[ 3.31598227e-01  3.72116750e-01  4.03724252e-01 -8.24118211e-02
 -4.47786551e-02  4.17673774e-01  3.15087830e-01 -2.49643522e-01
 -1.37808883e-01  5.63418434e-02  2.19929218e-01  5.83113174e-02
  4.64294477e-02  2.46665277e-01 -2.46595274e-01 -1.31689865e-01
 -1.69240532e-01]
[ 6.30921033e-02  1.01249056e-01  8.29855709e-02 -3.50555339e-02
  2.41479376e-02  6.13929764e-02 -1.39681716e-01 -4.65988731e-02
 -1.48967389e-01 -6.77411649e-01 -4.99721120e-01  1.27028371e-01
  6.60375454e-02  2.89848401e-01  1.46989274e-01 -2.26743985e-01
  2.08064649e-01]
[-2.81310530e-01 -2.67817346e-01 -1.61826771e-01  5.15472524e-02
  1.09766541e-01 -1.00412335e-01  1.58558487e-01 -1.31291364e-01
 -1.84995991e-01 -8.70892205e-02  2.30710568e-01  5.34724832e-01
  5.19443019e-01  1.61189487e-01 -1.73142230e-02 -7.92734946e-02
 -2.69129066e-01]
[ 5.74140964e-03  5.57860920e-02 -5.56936353e-02 -3.95434345e-01
 -4.26533594e-01 -4.34543659e-02  3.02385408e-01  2.22532003e-01
  5.60919470e-01 -1.27288825e-01 -2.22311021e-01  1.40166326e-01
  2.04719730e-01 -7.93882496e-02 -2.16297411e-01  7.59581203e-02
 -1.09267913e-01]
[ 1.62374420e-02 -7.53468452e-03  4.25579803e-02  5.26927980e-02
 -3.30915896e-02  4.34542349e-02  1.91198583e-01  3.00003910e-02
 -1.62755446e-01 -6.41054950e-01  3.31398003e-01 -9.12555212e-02
 -1.54927646e-01 -4.87045875e-01  4.73400144e-02  2.98118619e-01
 -2.16163313e-01]
[ 4.24863486e-02  1.29497196e-02  2.76928937e-02  1.61332069e-01
  1.18485556e-01  2.50763629e-02 -6.10423460e-02 -1.08528966e-01
 -2.09744235e-01  1.49692034e-01 -6.33790064e-01  1.09641298e-03
  2.84770105e-02 -2.19259358e-01 -2.43321156e-01  2.26584481e-01
 -5.59943937e-01]
[ 1.03090398e-01  5.62709623e-02 -5.86623552e-02  1.22678028e-01
  1.02491967e-01 -7.88896442e-02 -5.70783816e-01 -9.84599754e-03
  2.21453442e-01 -2.13293009e-01  2.32660840e-01  7.70400002e-02
  1.21613297e-02  8.36048735e-02 -6.78523654e-01  5.41593771e-02
  5.33553891e-03]
[ 9.02270802e-02  1.77864814e-01  1.28560713e-01 -3.41099863e-01
 -4.03711989e-01  5.94419181e-02 -5.60672902e-01  4.57332880e-03
 -2.75022548e-01  1.33663353e-01  9.44688900e-02  1.85181525e-01
  2.54938198e-01 -2.74544380e-01  2.55334907e-01  4.91388809e-02
 -4.19043052e-02]
[-5.25098025e-02 -4.11400844e-02 -3.44879147e-02 -6.40257785e-02
 -1.45492289e-02 -2.08471834e-02  2.23105808e-01 -1.86675363e-01
 -2.98324237e-01  8.20292186e-02 -1.36027616e-01  1.23452200e-01
  8.85784627e-02 -4.72045249e-01 -4.22999706e-01 -1.32286331e-01
  5.90271067e-01]
[ 3.58970400e-01 -5.43427250e-01  6.09651110e-01 -1.44986329e-01
  8.03478445e-02 -4.14705279e-01  9.01788964e-03  5.08995918e-02
  1.14639620e-03  7.72631963e-04 -1.11433396e-03  1.38133366e-02
  6.20932749e-03 -2.22215182e-03 -1.91869743e-02 -3.53098218e-02
 -1.30710024e-02]
[-4.59139498e-01  5.18568789e-01  4.04318439e-01  1.48738723e-01
 -5.18683400e-02 -5.60363054e-01  5.27313042e-02 -1.01594830e-01
  2.59293381e-02 -2.88282896e-03  1.28904022e-02 -2.98075465e-02
  2.70759809e-02  2.12476294e-02 -3.33406243e-03  4.38803230e-02
  5.00844705e-03]
[ 4.30462074e-02 -5.84055850e-02 -6.93988831e-02 -8.10481404e-03
 -2.73128469e-01 -8.11578181e-02  1.00693324e-01  1.43220673e-01
 -3.59321731e-01  3.19400370e-02 -1.85784733e-02  4.03723253e-02
```

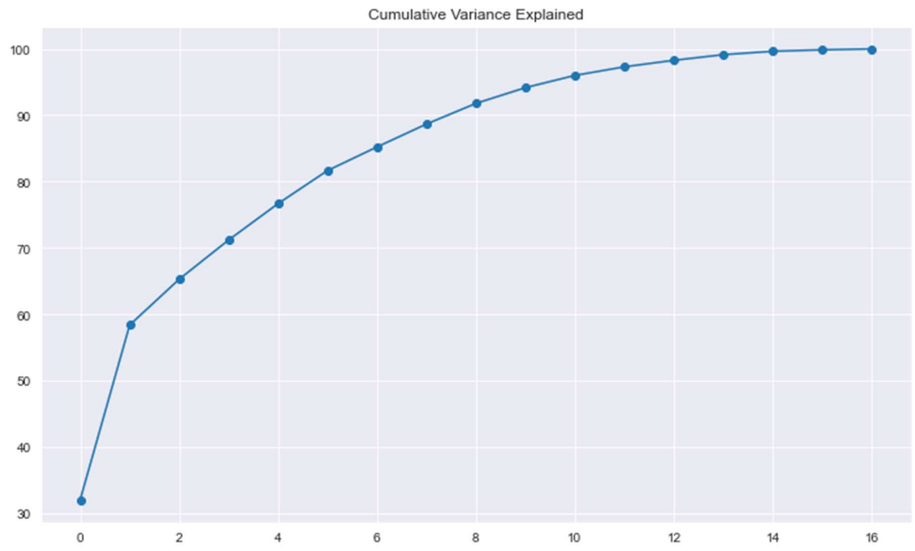
```
 -5.89734026e-02  4.45000727e-01 -1.30727978e-01  6.92088870e-01
  2.19839000e-01]
[-1.33405806e-01  1.45497511e-01 -2.95896092e-02 -6.97722522e-01
  6.17274818e-01 -9.91640992e-03 -2.09515982e-02 -3.83544794e-02
 -3.40197083e-03  9.43887925e-03  3.09001353e-03  1.12055599e-01
 -1.58909651e-01  2.08991284e-02  8.41789410e-03  2.27742017e-01
  3.39433604e-03]
[ 8.06328039e-02  3.34674281e-02 -8.56967180e-02 -1.07828189e-01
  1.51742110e-01 -5.63728817e-02  1.92857500e-02 -3.40115407e-02
 -5.84289756e-02 -6.68494643e-02  2.75286207e-02 -6.91126145e-01
  6.71008607e-01  4.13740967e-02 -2.71542091e-02  7.31225166e-02
  3.64767385e-02]
[-5.95830975e-01 -2.92642398e-01  4.44638207e-01 -1.02303616e-03
 -2.18838802e-02  5.23622267e-01 -1.25997650e-01  1.41856014e-01
  6.97485854e-02 -1.14379958e-02 -3.94547417e-02 -1.27696382e-01
  5.83134662e-02  1.77152700e-02 -1.04088088e-01  9.37464497e-02
  6.91969778e-02]
[ 2.40709086e-02 -1.45102446e-01  1.11431545e-02  3.85543001e-02
 -8.93515563e-02  5.61767721e-02 -6.35360730e-02 -8.23443779e-01
  3.54559731e-01 -2.81593679e-02 -3.92640266e-02  2.32224316e-02
  1.64850420e-02 -1.10262122e-02  1.82660654e-01  3.25982295e-01
  1.22106697e-01]]
```

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

PCA based on variance explained and Cumulative_Variance_Explained

	Eigen_Values	Variance_Explained	Cumulative_Variance_Explained
0	5.4505	32.0206	32.0206
1	4.4836	26.3402	58.3608
2	1.1747	6.9009	65.2618
3	1.0082	5.9230	71.1847
4	0.9342	5.4884	76.6732
5	0.8485	4.9847	81.6579
6	0.6058	3.5589	85.2167
7	0.5879	3.4536	88.6703
8	0.5306	3.1172	91.7876
9	0.4043	2.3752	94.1628
10	0.0230	0.1353	96.0042
11	0.0367	0.2158	97.3002
12	0.3134	1.8414	98.2860
13	0.0880	0.5171	99.1318
14	0.1440	0.8458	99.6490
15	0.1678	0.9858	99.8647
16	0.2206	1.2960	100.0000





Eigen Vectors into a data frame with the original 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	-0.2488	-0.2076	-0.1763	-0.3543	-0.3440	-0.1546	-0.0264	-0.2947	-0.2490	-0.0648	0.0425	-0.3183	-0.3171	0.1770	-0.2051	-0.3189	-0.2523
1	0.3316	0.3721	0.4037	-0.0824	-0.0448	0.4177	0.3151	-0.2496	-0.1378	0.0563	0.2199	0.0583	0.0464	0.2467	-0.2466	-0.1317	-0.1692
2	0.0631	0.1012	0.0830	-0.0351	0.0241	0.0614	-0.1397	-0.0466	-0.1490	-0.6774	-0.4997	0.1270	0.0660	0.2898	0.1470	-0.2267	0.2081
3	-0.2813	-0.2678	-0.1618	0.0515	0.1098	-0.1004	0.1586	-0.1313	-0.1850	-0.0871	0.2307	0.5347	0.5194	0.1612	-0.0173	-0.0793	-0.2691
4	0.0057	0.0558	-0.0557	-0.3954	-0.4265	-0.0435	0.3024	0.2225	0.5609	-0.1273	-0.2223	0.1402	0.2047	-0.0794	-0.2163	0.0760	-0.1093
5	0.0162	-0.0075	0.0426	0.0527	-0.0331	0.0435	0.1912	0.0300	-0.1628	-0.6411	0.3314	-0.0913	-0.1549	-0.4870	0.0473	0.2981	-0.2162
6	0.0425	0.0129	0.0277	0.1613	0.1185	0.0251	-0.0610	-0.1085	-0.2097	0.1497	-0.6338	0.0011	0.0285	-0.2193	-0.2433	0.2266	-0.5599
7	0.1031	0.0563	-0.0587	0.1227	0.1025	-0.0789	-0.5708	-0.0098	0.2215	-0.2133	0.2327	0.0770	0.0122	0.0836	-0.6785	0.0542	0.0053
8	0.0902	0.1779	0.1286	-0.3411	-0.4037	0.0594	-0.5607	0.0046	-0.2750	0.1337	0.0945	0.1852	0.2549	-0.2745	0.2553	0.0491	-0.0419
9	-0.0525	-0.0411	-0.0345	-0.0640	-0.0145	-0.0208	0.2231	-0.1867	-0.2983	0.0820	-0.1360	0.1235	0.0886	-0.4720	-0.4230	-0.1323	0.5903
10	0.3590	-0.5434	0.6097	-0.1450	0.0803	-0.4147	0.0090	0.0509	0.0011	0.0008	-0.0011	0.0138	0.0062	-0.0022	-0.0192	-0.0353	-0.0131
11	-0.4591	0.5186	0.4043	0.1487	-0.0519	-0.5604	0.0527	-0.1016	0.0259	-0.0029	0.0129	-0.0298	0.0271	0.0212	-0.0033	0.0439	0.0050
12	0.0430	-0.0584	-0.0694	-0.0081	-0.2731	-0.0812	0.1007	0.1432	-0.3593	0.0319	-0.0186	0.0404	-0.0590	0.4450	-0.1307	0.6921	0.2198
13	-0.1334	0.1455	-0.0296	-0.6977	0.6173	-0.0099	-0.0210	-0.0384	-0.0034	0.0094	0.0031	0.1121	-0.1589	0.0209	0.0084	0.2277	0.0034
14	0.0806	0.0335	-0.0857	-0.1078	0.1517	-0.0564	0.0193	-0.0340	-0.0584	-0.0668	0.0275	-0.6911	0.6710	0.0414	-0.0272	0.0731	0.0365
15	-0.5958	-0.2926	0.4446	-0.0010	-0.0219	0.5236	-0.1260	0.1419	0.0697	-0.0114	-0.0395	-0.1277	0.0583	0.0177	-0.1041	0.0937	0.0692
16	0.0241	-0.1451	0.0111	0.0386	-0.0894	0.0562	-0.0635	-0.8234	0.3546	-0.0282	-0.0393	0.0232	0.0165	-0.0110	0.1827	0.3260	0.1221

	PCA-1	PCA-2	PCA-3	PCA-4
0	-1.5929	0.7673	-0.1011	-0.9218
1	-2.1924	-0.5788	2.2788	3.5889
2	-1.4310	-1.0928	-0.4381	0.6772
3	2.8556	-2.6306	0.1417	-1.2955
4	-2.2120	0.0216	2.3870	-1.1145

Principal Component (eigenvectors) into a data frame with the original 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
PCA-1	0.2488	0.2076	0.1763	0.3543	0.3440	0.1546	0.0264	0.2947	0.2490	0.0648	-0.0425	0.3183	0.3171	-0.1770	0.2051	0.3189	0.2523
PCA-2	0.3316	0.3721	0.4037	-0.0824	-0.0448	0.4177	0.3151	-0.2496	-0.1378	0.0563	0.2199	0.0583	0.0464	0.2467	-0.2466	-0.1317	-0.1692
PCA-3	-0.0631	-0.1012	-0.0830	0.0351	-0.0242	-0.0614	0.1397	0.0466	0.1490	0.6774	0.4997	-0.1270	-0.0660	-0.2898	-0.1470	0.2267	-0.2081
PCA-4	0.2813	0.2678	0.1618	-0.0516	-0.1098	0.1004	-0.1586	0.1313	0.1850	0.0871	-0.2307	-0.5347	-0.5194	-0.1612	0.0173	0.0793	0.2691

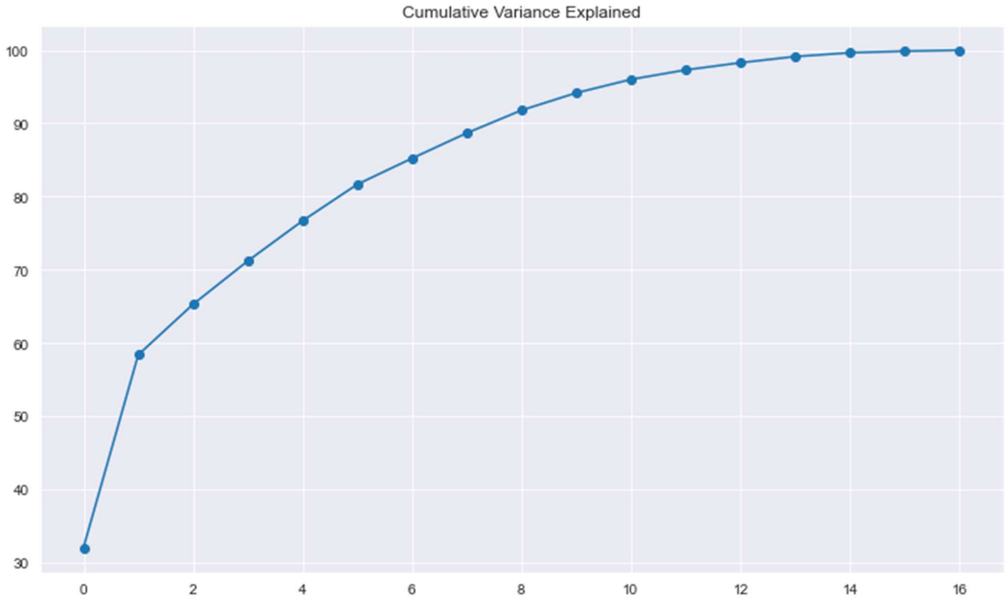
2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).

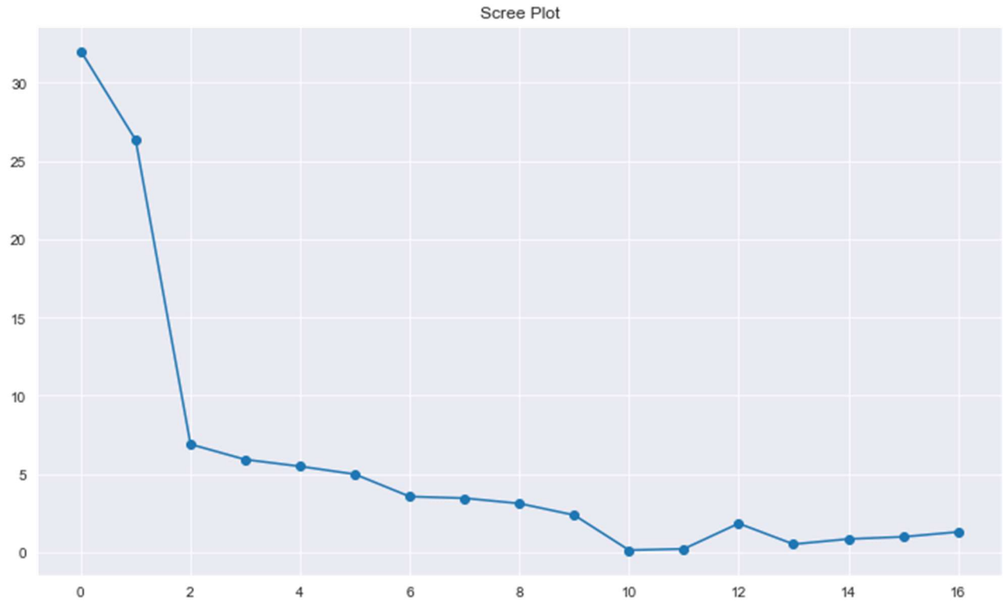
```
Explicit form of the first PC (in terms of the eigenvectors):

0.25 * Apps + 0.21 * Accept + 0.18 * Enroll + 0.35 * 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 * S.F.Ratio + 0.21 * perc.alumni + 0.32 * Expend + 0.25 * 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?

Variance Explained and Cumulative_Variance_Explained - Eigen values





	Eigen_Values	Variance_Explained	Cumulative_Variance_Explained
0	5.4505	32.0206	32.0206
1	4.4836	26.3402	58.3608
2	1.1747	6.9009	65.2618
3	1.0082	5.9230	71.1847
4	0.9342	5.4884	76.6732
5	0.8485	4.9847	81.6579
6	0.6058	3.5589	85.2167
7	0.5879	3.4536	88.6703
8	0.5306	3.1172	91.7876
9	0.4043	2.3752	94.1628
10	0.0230	0.1353	96.0042
11	0.0367	0.2158	97.3002
12	0.3134	1.8414	98.2860
13	0.0880	0.5171	99.1318
14	0.1440	0.8458	99.6490
15	0.1678	0.9858	99.8647
16	0.2206	1.2960	100.0000

The eigenvalue-one criterion.

In principal component analysis, one of the most commonly used criteria for solving the number-of-components problem is the eigenvalue-one criterion, also known as the Kaiser criterion (Kaiser, 1960). With this approach, you retain and interpret any component with an eigenvalue greater than 1.00. On the other hand, a component with an eigenvalue less than 1.00 is accounting for less variance than had been contributed by one variable and are viewed as trivial, and are not retained.

The scree test.

With the scree test (Cattell, 1966), you plot the eigenvalues associated with each component and look for a "break" between the components with relatively large eigenvalues and those with small eigenvalues. The components that appear before the break are assumed to be meaningful and are retained for rotation; those apppearing after the break are assumed to be unimportant and are not retained.

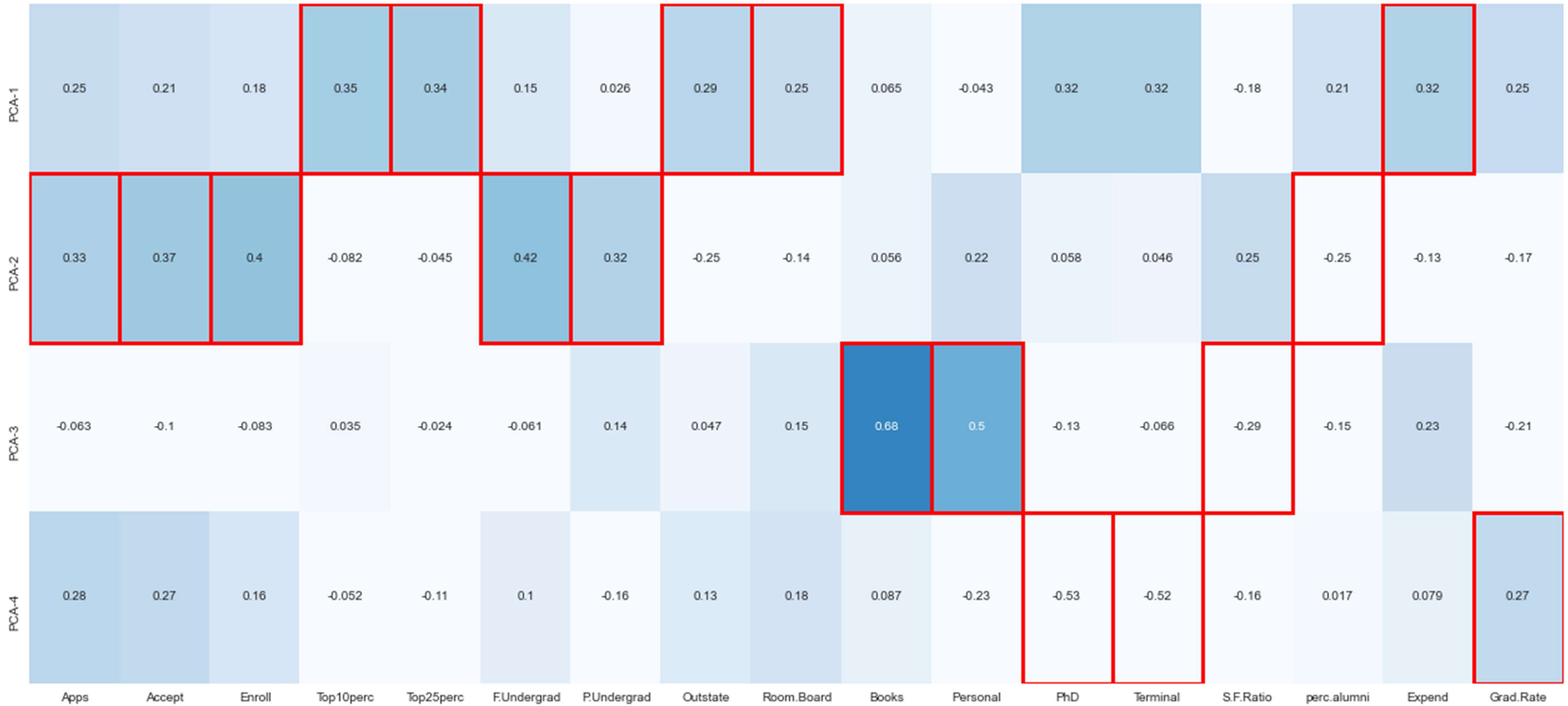
Proportion of variance accounted for

An alternative criterion is to retain enough components so that the cumulative percent of variance accounted for is equal to some minimal value. When researchers use the "cumulative percent of variance accounted for" as the criterion for solving the number-ofcomponents problem, they usually retain enough components so that the cumulative percent of variance accounted for at least 70% (and sometimes 80%).

Based on the above criterion, the first 4 PC have eigenvalue of more than 1 and it accounts for more than approx 70% of the cumulative variance explained (71.19%).

Eigen Vector indicates the direction of the principal components (new axes). It determine the directions of the new feature space. ¶

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]


PCA-1
0.25 0.21 0.18 0.35 0.34 0.15 0.026 0.29 0.25 0.065 -0.043 0.32 0.32 -0.18 0.21 0.32 0.25
PCA-2
0.33 0.37 0.4 -0.082 -0.045 0.42 0.32 -0.25 -0.14 0.056 0.22 0.058 0.046 0.25 -0.25 -0.13 -0.17
PCA-3
-0.063 -0.1 -0.083 0.035 -0.024 -0.061 0.14 0.047 0.15 0.68 0.5 -0.13 -0.066 -0.29 -0.15 0.23 -0.21
PCA-4
0.28 0.27 0.16 -0.052 -0.11 0.1 -0.16 0.13 0.18 0.087 -0.23 -0.53 -0.52 -0.16 0.017 0.079 0.27
Apps
Accept
Enroll
Top10perc
Top25perc
F.Undergrad
P.Undergrad
Outstate
Room.Board
Books
Personal
PhD
Terminal
S.F.Ratio
perc.alumni
Expend
Grad.Rate

Component Summaries

First Principal Component Analysis - PCA1:

The first principal component is a measure of the quality of Top10pec, Top25perc, Expend and to some extend Outstate and Room.Board. They are all positively correlated to PCA1.

Second Principal Component Analysis - PCA2:

The second principal component is a measure of the quality of Apps (applications received), Accept and Enroll as well as F.Undergrad and P.Undergrad. They are all positively correlated to PCA2 except perc.alumni which is negatively correlated to PCA2. The second PCA primarily measures the applications of students received, enrolled and accepted and whether they are F.Undergrad or P.Undergrad.

Third Principal Component Analysis - PCA3:

The third principal component is a measure of the quality of Books, Personal that is positively correlated to PCA3 and S.F Ratio i.e. Student/Faculty Ratio which is negatively correlated to PCA3.

Fourth Principal Component Analysis - PCA4:

The fourth principal component denotes Graduation Rate, Terminal and PhD of the universities or colleges.