# ADVANCED STATISTICS PROJECT
# BUSINESS REPORT

BY

RITUSRI MOHAN

# CONTENTS

# PROBLEM 1

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.

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

One -way ANOVA(Education)

Null Hypothesis:
The mean salary is *same* across all three categories of Education, i.e., Doctorate, Bachelors and HS-Grad.

Alternate Hypothesis:
The mean salary is *different for at least one* of the categories of Education.

One -way ANOVA(Occupation)

Null Hypothesis:
The mean salary is *same* across all four categories of Occupation, i.e., (Prof-Specialty, Sales, Adm-clerical, Exec-Managerial)

Alternate Hypothesis:
The mean salary is *different for at least one* of the categories 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.

| | 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 |

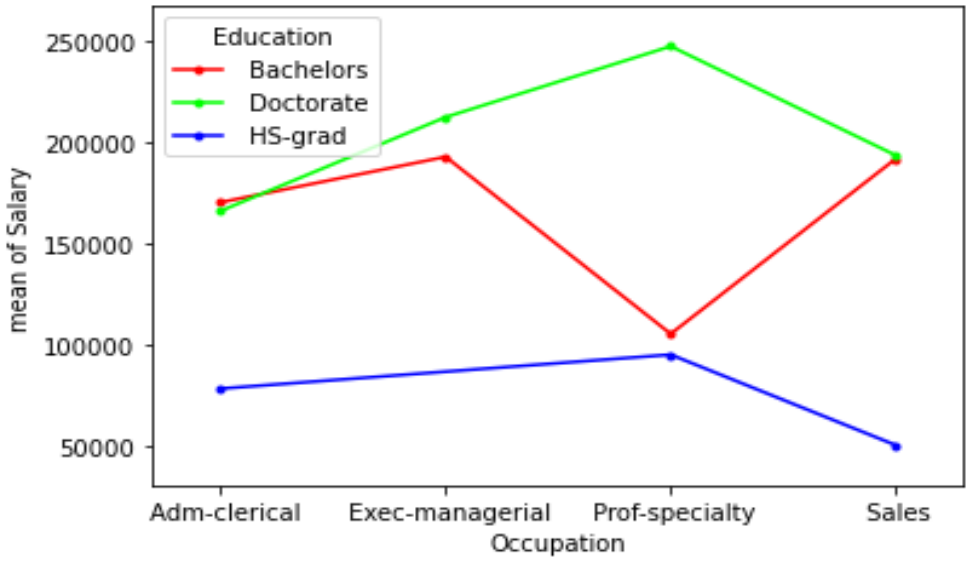
Since the p value, i.e.,1.257709e-08 is *less than* the significance level (*alpha = 0.05*), the *null hypothesis is rejected* and it is concluded that there is a difference in the mean salaries for at least one category of education.

**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.

| | 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 |

Since the p value, i.e., 0.458508 is *greater than* the significance level (*alpha = 0.05*), the *null hypothesis cannot be rejected* and it is concluded that there is no significant difference in the mean salaries across the 4 categories of occupation**.**

**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**.**



- ✓ 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 (ranging from 170000–190000).

- ✓ People with education as Bachelors and occupation as Prof-Specialty earn much lesser than people with education as Bachelors and occupation as Adm-clerical, Exec-managerial 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. There is a reversal in this part of the plot.

- ✓ 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.

- ✓ Sales people with Bachelors or Doctorate education earn the almost same salaries and earn higher than people with education as HS-graduates.

- ✓ 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.

- ✓ People with education as Bachelors and occupation, Sales and Exec-Managerial earn almost 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?

Null Hypothesis:
The effect of the independent variable 'education' on the mean 'salary' is *not dependent* on the effect of the other independent variable 'occupation'.

Alternate Hypothesis:
There *is an interaction effect* between the independent variable 'education' and the independent variable 'occupation' on the mean salary**.**

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

As p value, i.e., 2.232500e-05 is lesser than the significance level (alpha = 0.05), the null hypothesis is rejected.
Therefore, 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.

- ✓ By using ANOVA in this case study, we can understand the interaction and dependency between variables.
- ✓ From the interaction plot, we understood the trend of salaries for different occupations at different levels of education.

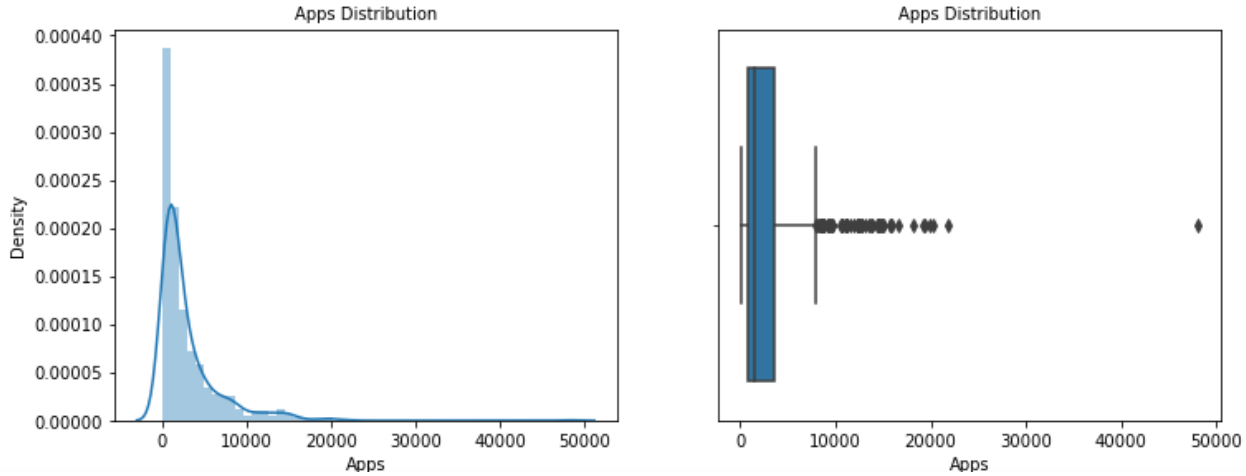
# PROBLEM 2

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?

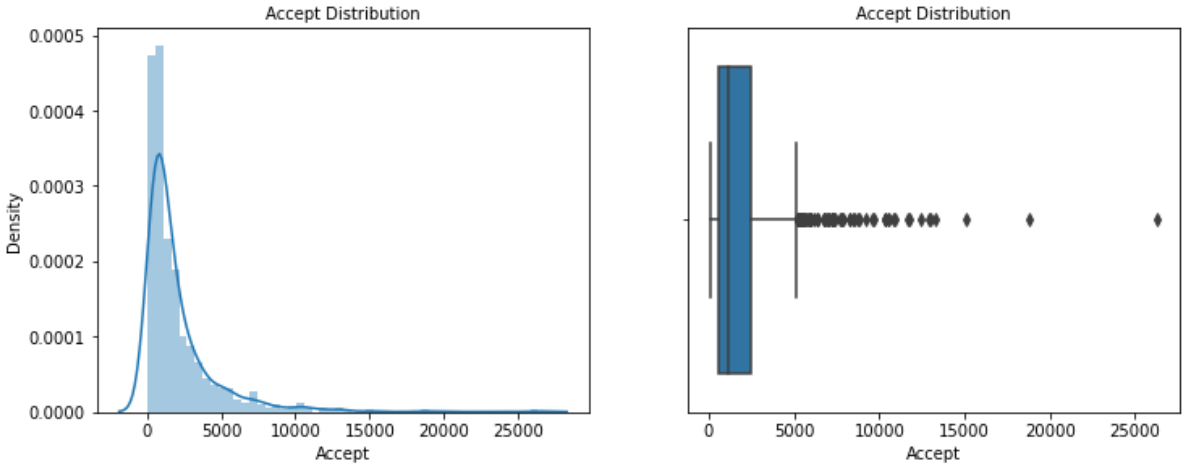
## UNIVARIATE ANALYSIS

### APPS:



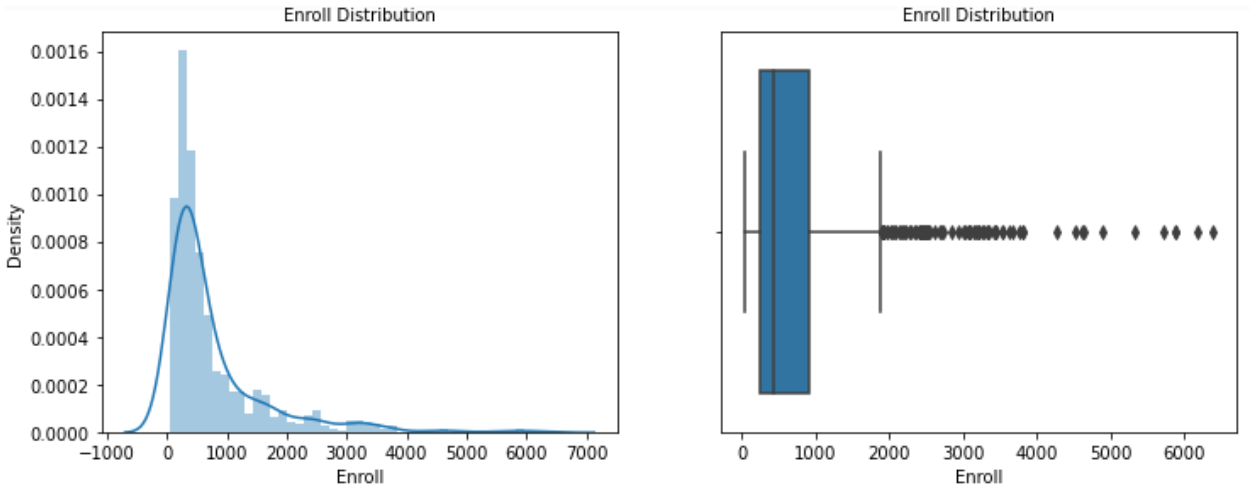

The Box plot of Apps variable shows outliers. The distribution of the data is skewed. The max applications are around 50,000.

### ACCEPT DISTRIBUTION:



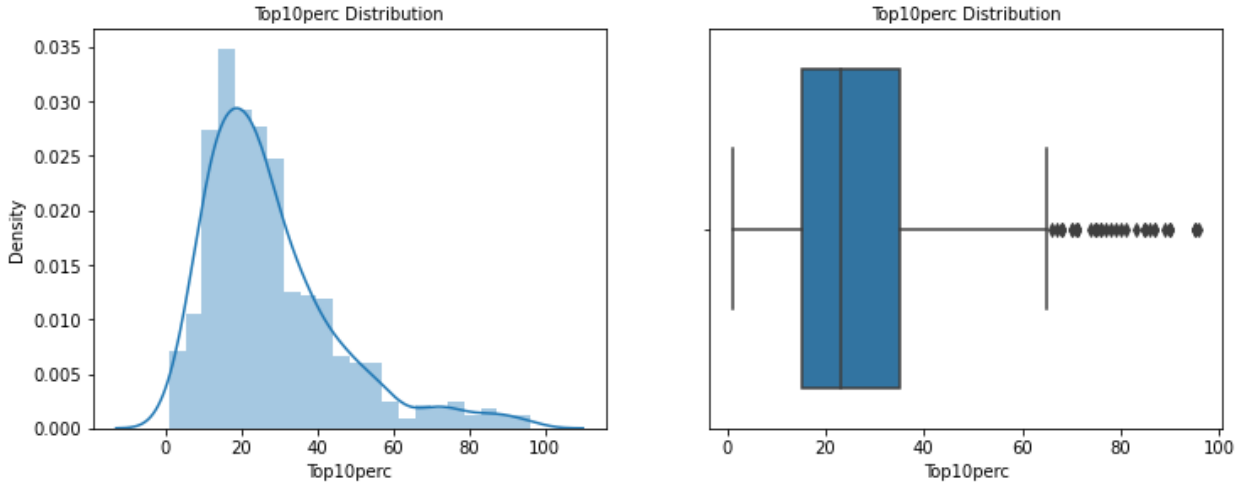
The Accept variable has outliers. The dist plot shows us the majority of applications accepted from each university are in the range - 70 to 1500.The accept variable is be positively skewed.

ENROLL DISTRIBUTION:



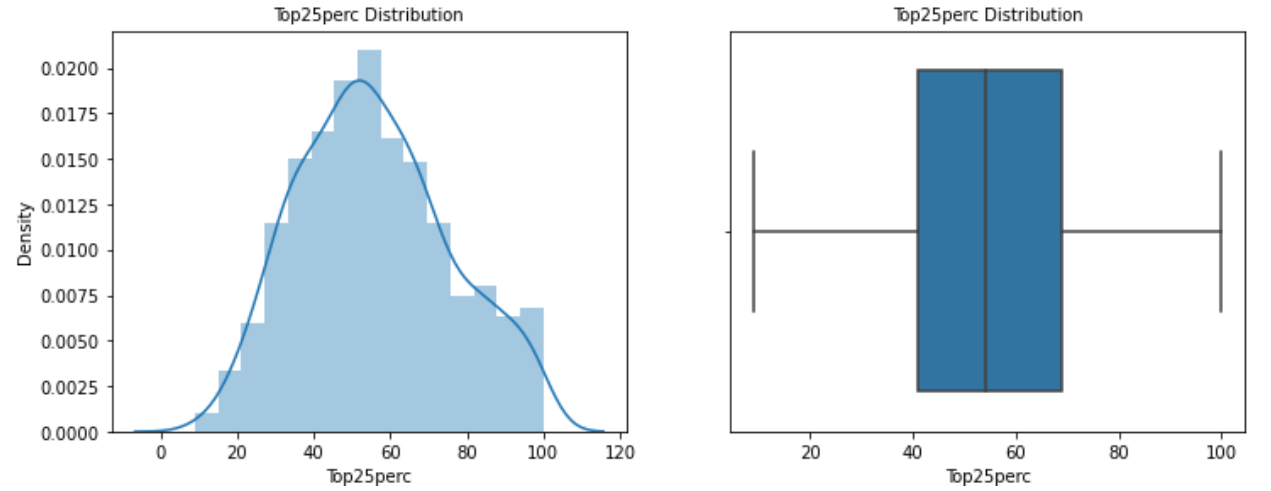

The box plot of the Enroll variable also has outliers. The distribution of the data is positively skewed. From the dist plot we can understand most colleges have enrolled students in the range of 200 to 600 students.

TOP 10 PERC DISTRIBUTION:



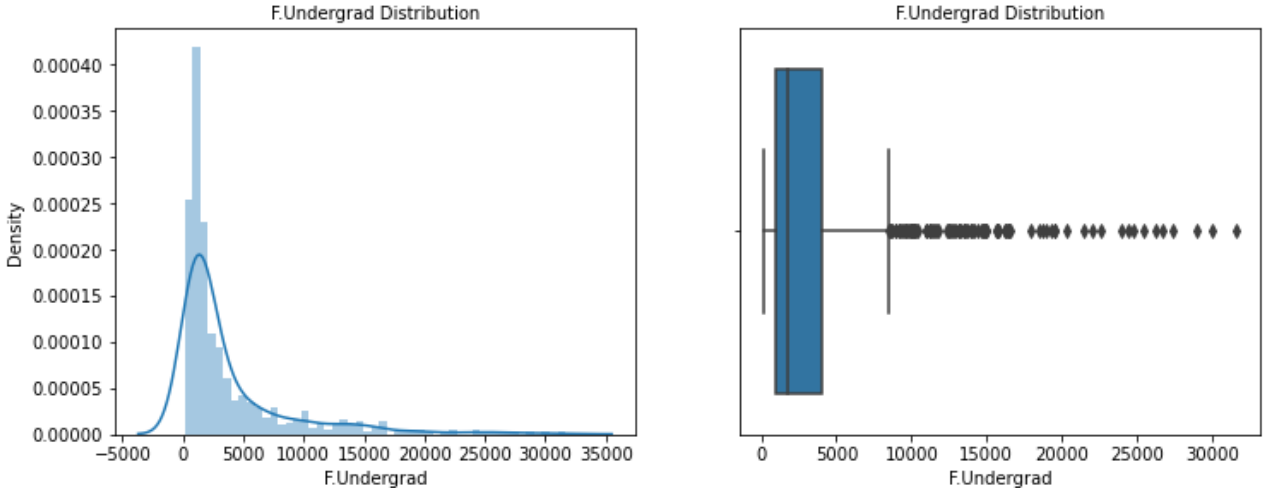

The box plot of the students from top 10 percentage of higher secondary class seems to have outliers. The distribution seems to be positively skewed. There is good amount of intake about 25 to 50 students from top 10 percentage of higher secondary class.

TOP 25 PERC DISTRIBUTION:



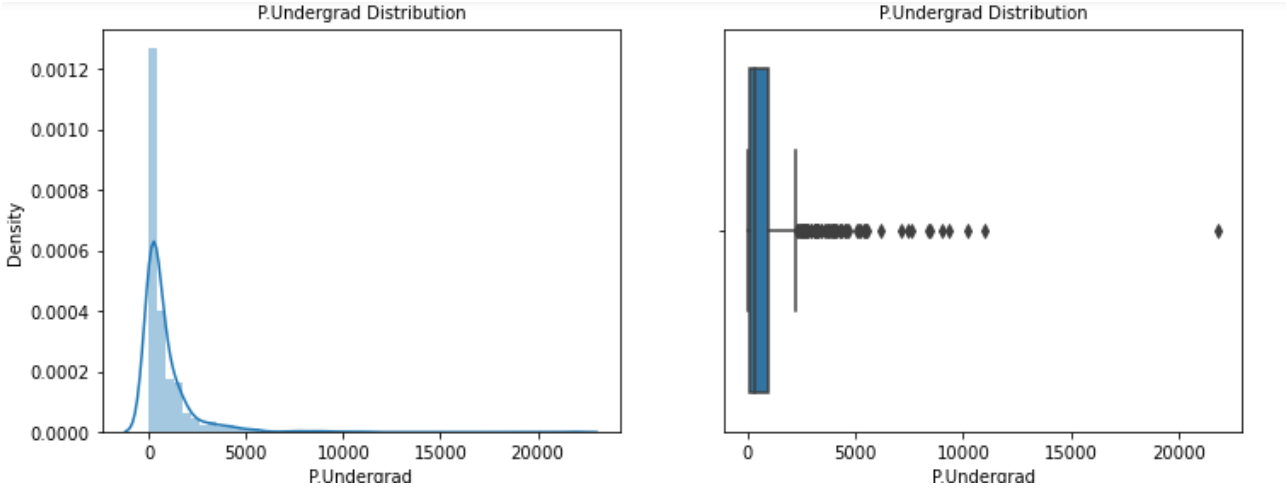

The box plot for the top 25% shows outliers. The distribution is almost normally distributed.

FULL TIME UNDERGRADUATE:



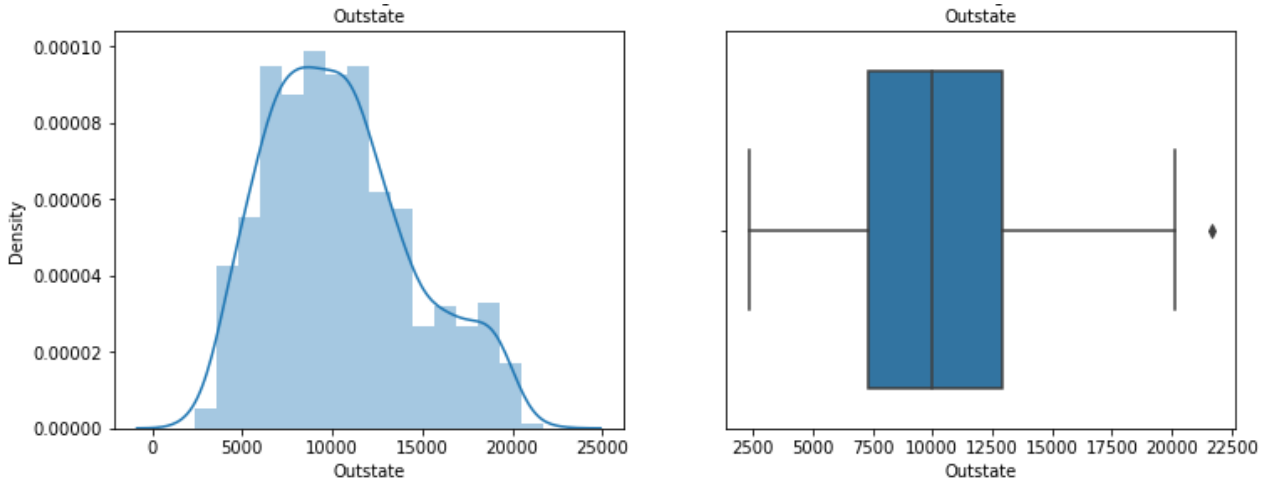

The box plot of the full time graduates has many outliers. The distribution of the data is positively skewed. The full time graduates are around 5000.

PART TIME UNDERGRADUATE:



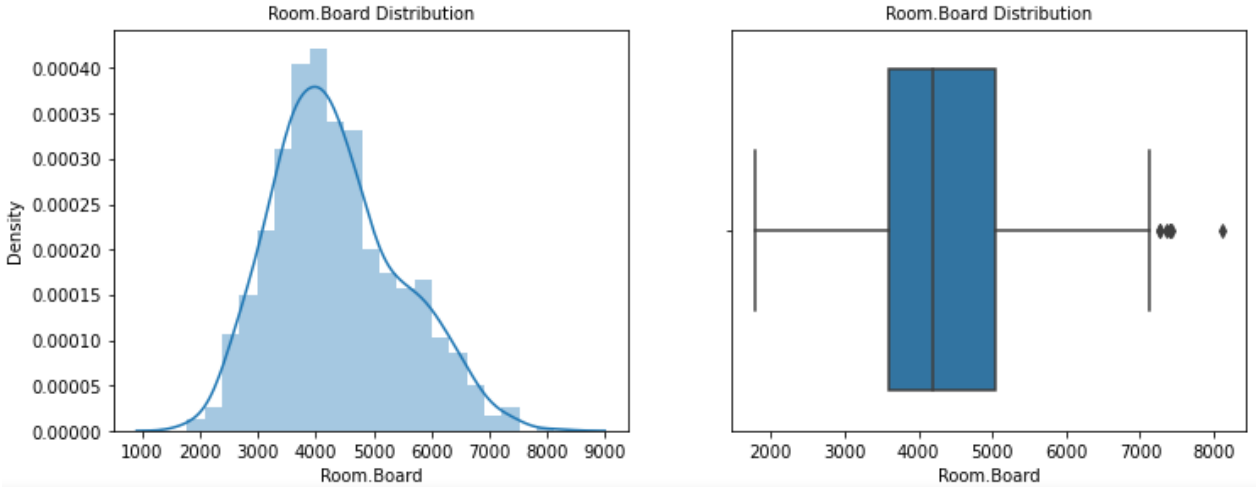

The box plot of the part time graduates has outliers. The distribution of the data is positively skewed. Part-time graduates are around 3000.

OUTSTATE:



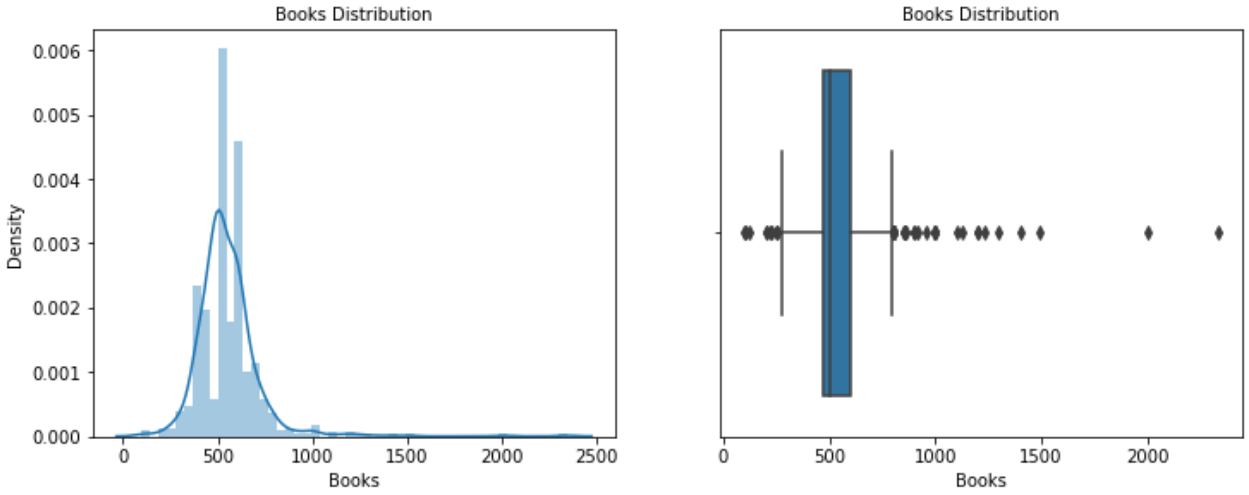

The box plot of outstate has only one outlier and the distribution is almost normally distributed

ROOM BOARD:



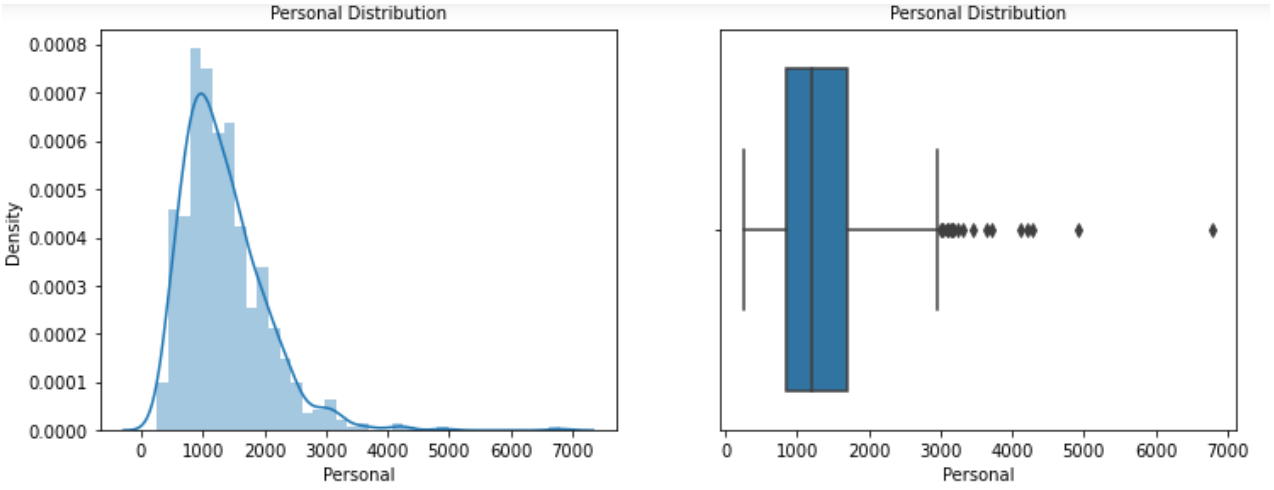

The Room Board has few outliers and the distribution is normally distributed.

BOOKS:



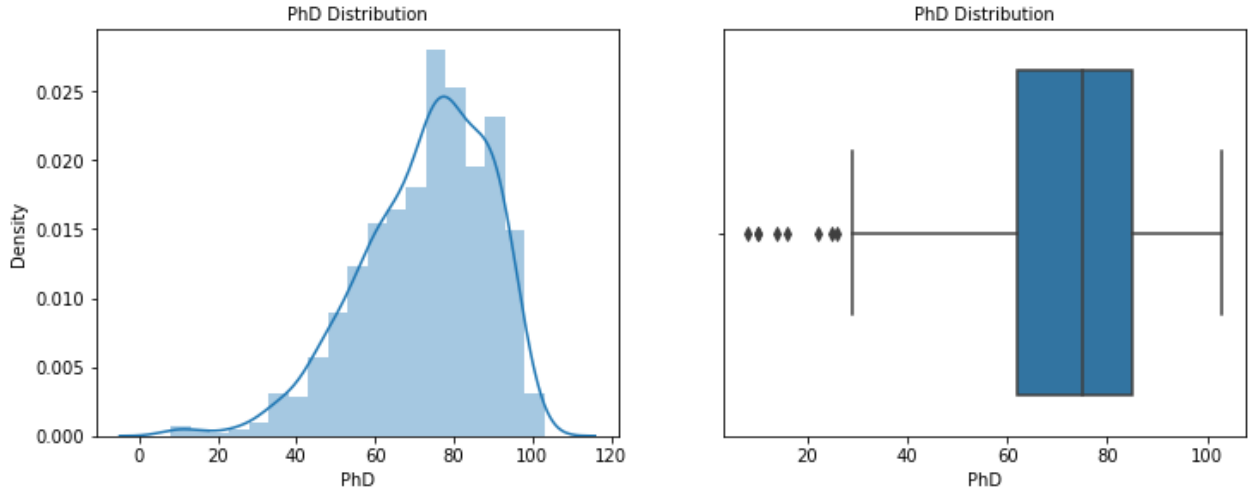

The box plot of books has outlier and the distribution seems to be bimodal. The cost of books per student seems to be in the range of 400 to 1000.

PERSONAL:



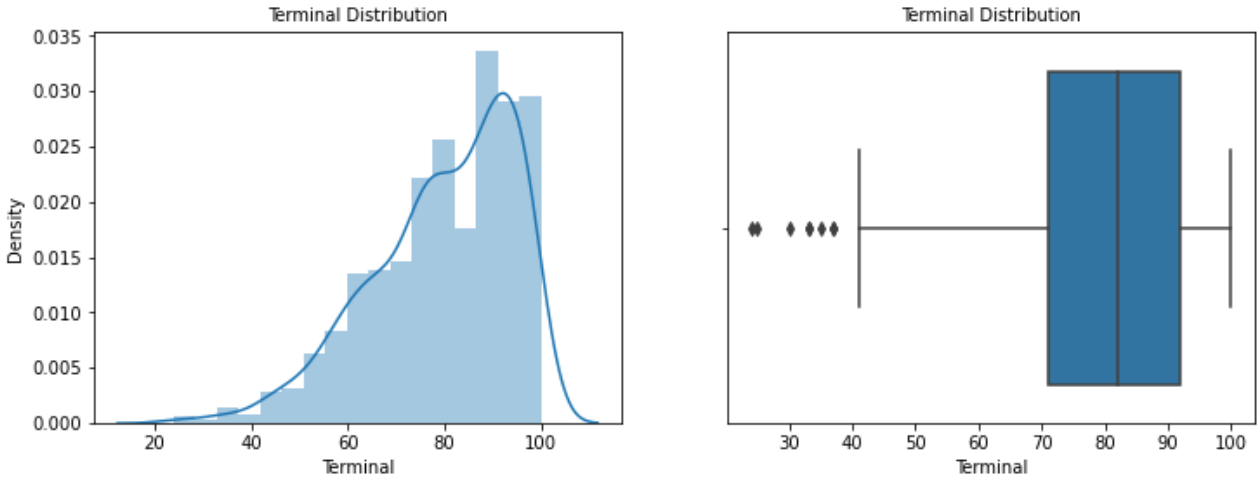

The box plot of personal expense has outliers. Some student's personal expenses are much more than the other students and the distribution seems to be positively skewed.

PHD:



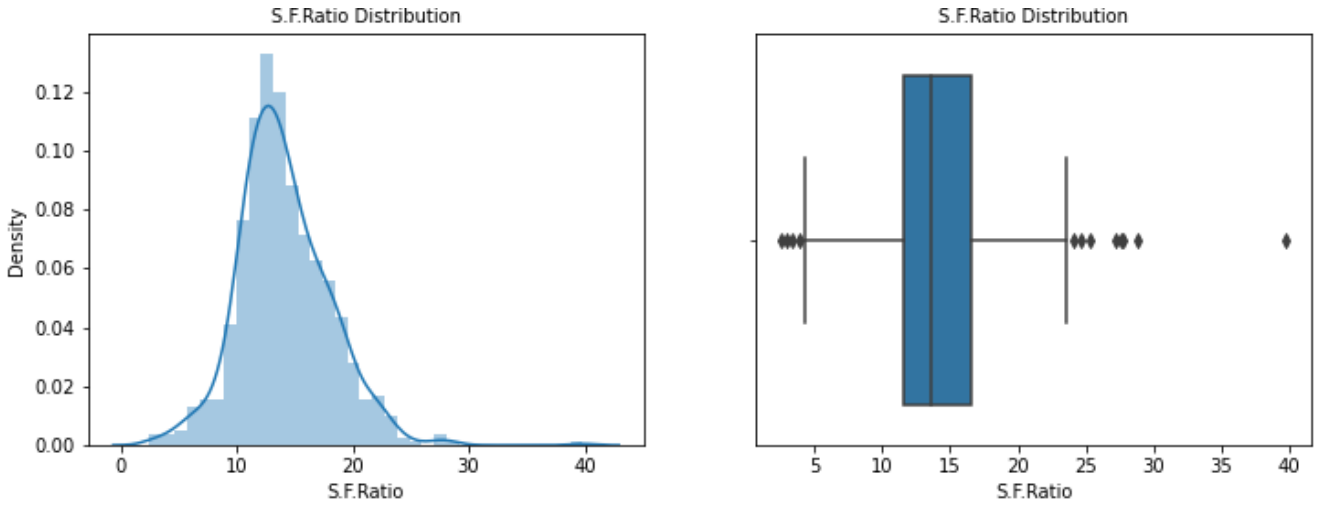

The box plot of PHD has outliers and the distribution is negatively skewed.

TERMINAL:



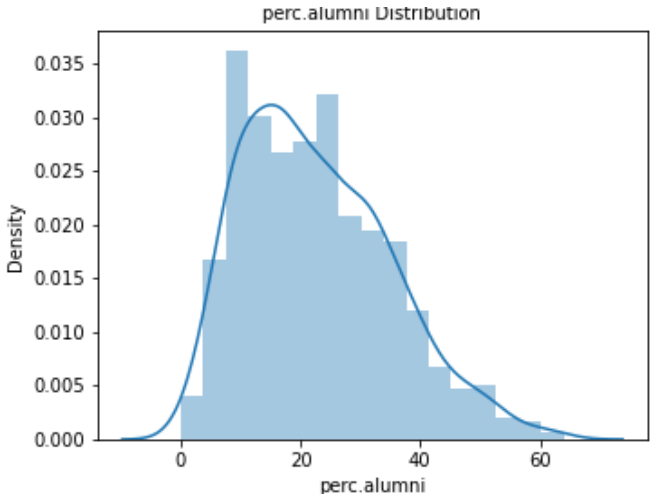

The box plot of terminal seems to have outliers in the dataset and the distribution for the terminal also seems to be negatively skewed.

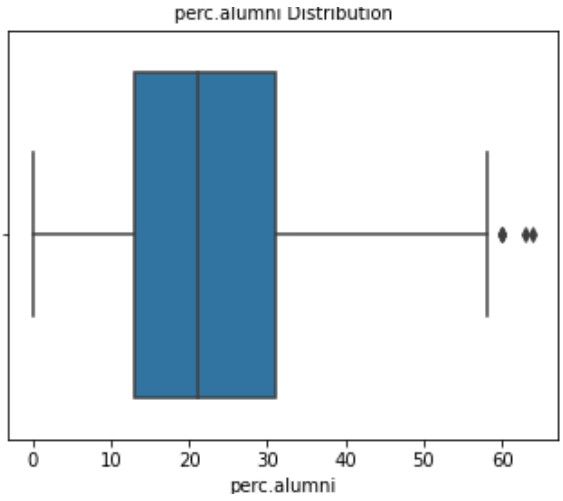
SF RATIO:




The SF ratio variable also has outliers in the dataset. The distribution is almost normally distributed. The student faculty ratio is almost same in all the university and colleges.

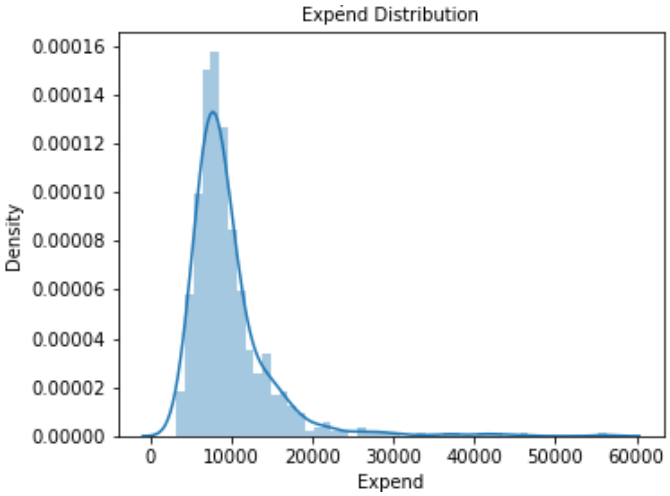
PERC ALUMNI:

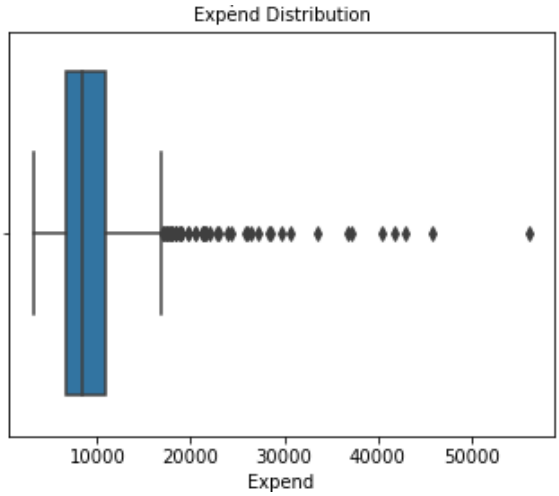



The percentage of alumni box plot seems to have outliers in the dataset and the distribution is almost normally distributed.

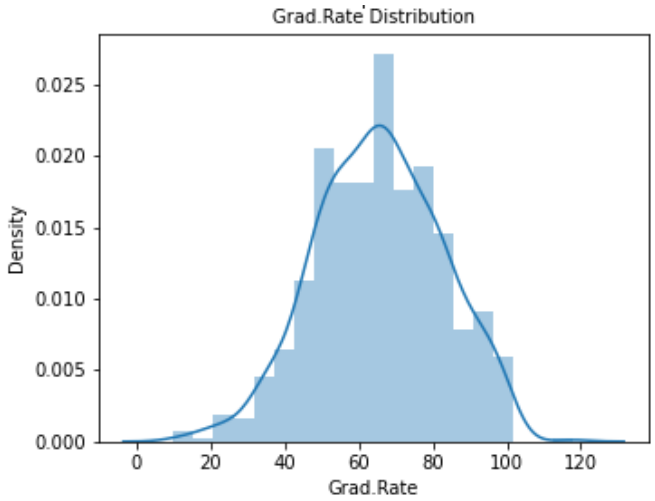
EXPENDITURE:

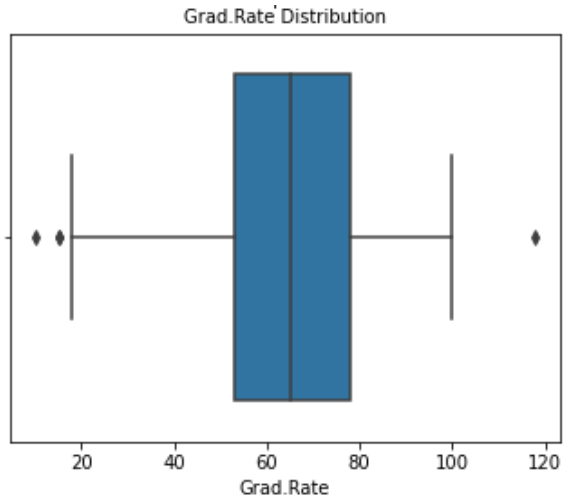



The expenditure variable also has outliers in the dataset and the distribution of the expenditure is positively skewed.

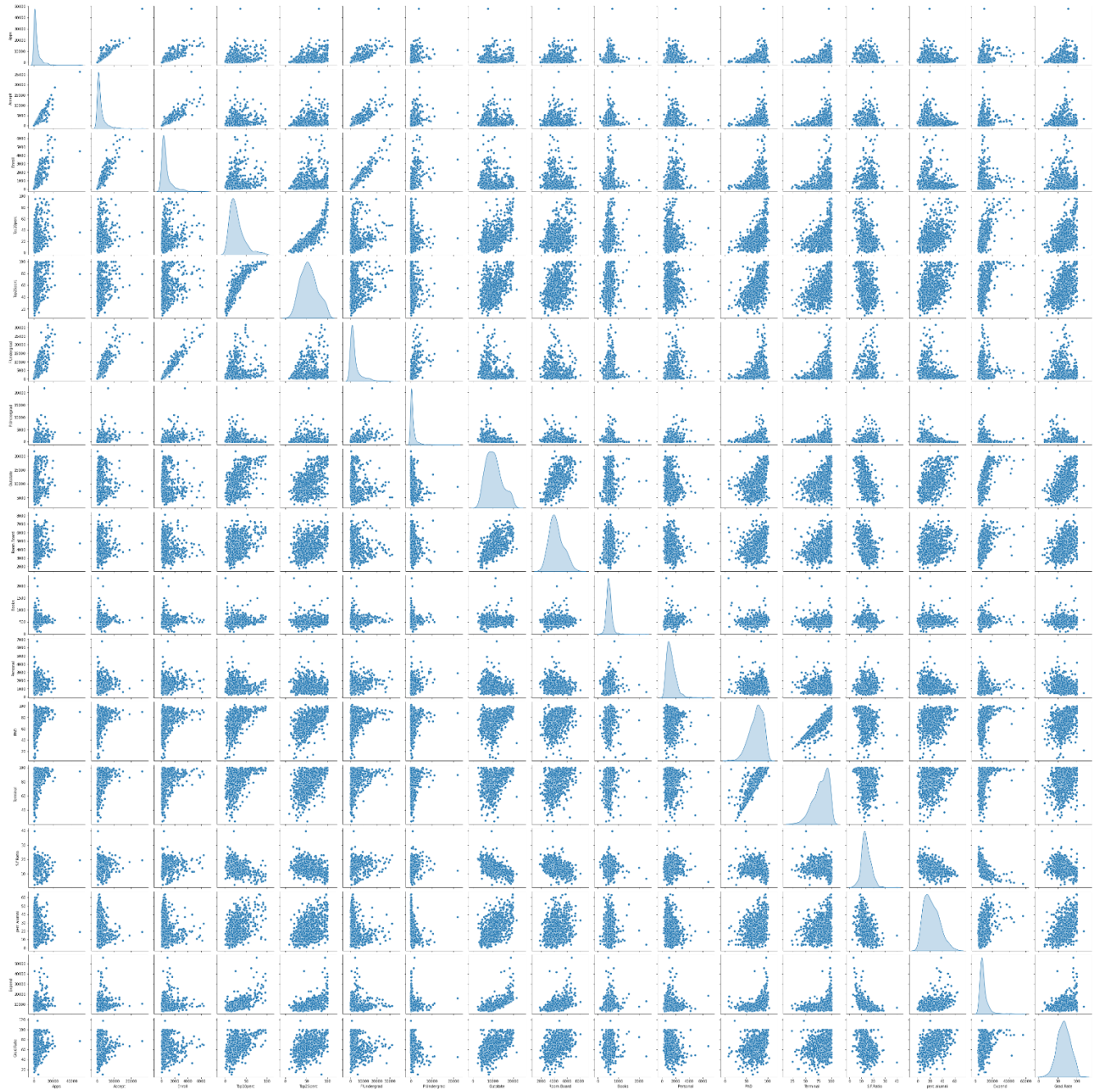
GRAD RATE:



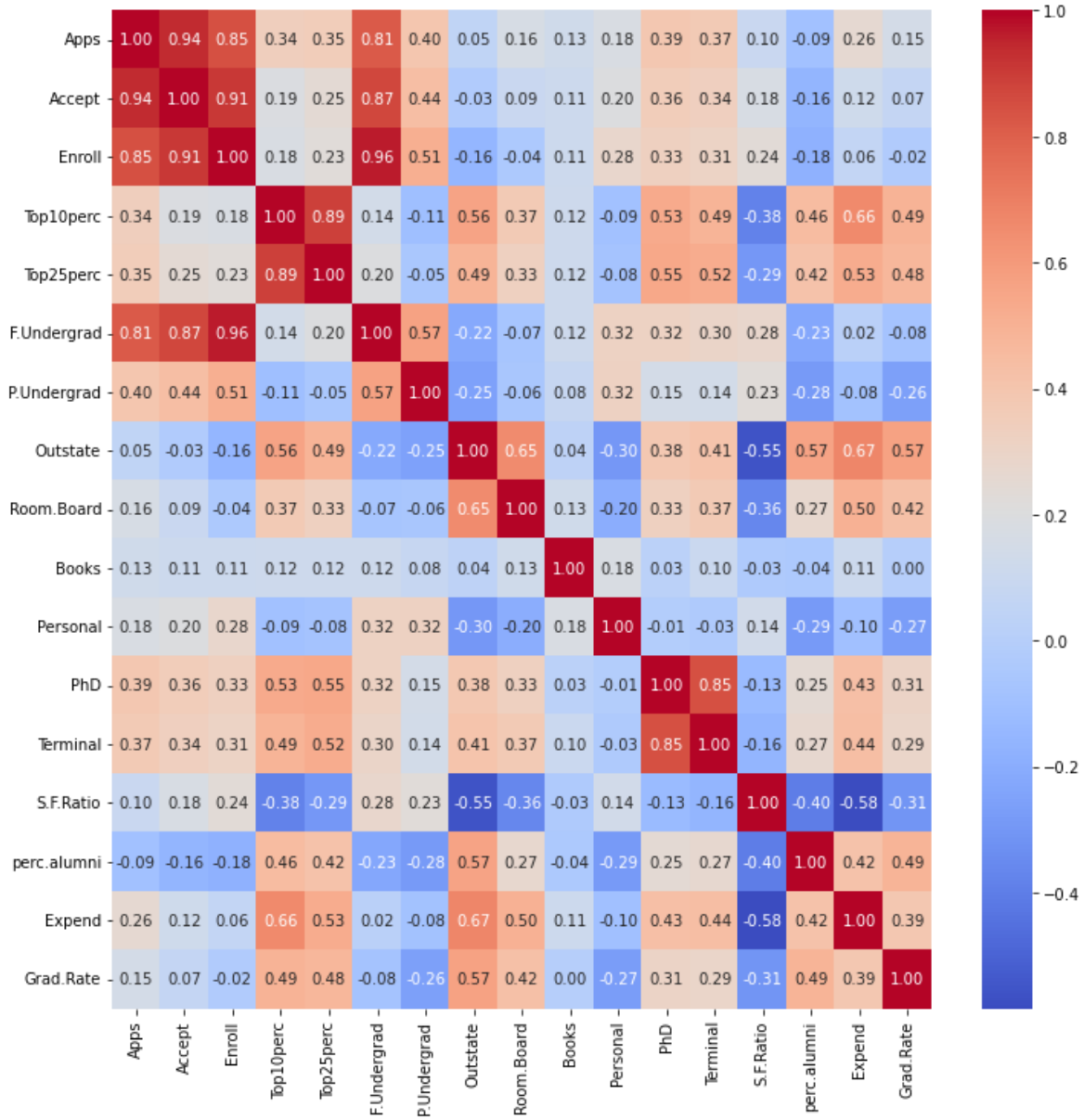

The graduation rate among the students in all the university above 60%. The box plot of the graduation rate has outliers in the dataset and the distribution is normally distributed.

# MULTIVARIATE ANALYSIS



This plot helps us to understand the relationship between all the numerical values in the dataset and establish the trends in the dataset.

HEATMAP




- ✓ From the Heatmap we understand that the application variable is highly positively correlated with application accepted, students enrolled and full time graduates. So this relationship gives the insights onwhen student submits the application it is accepted and the student is enrolled as fulltime graduate.

- ✓ There is a negative correlation between application and percentage of alumni which means that not all students are part of alumni.

- ✓ The applications with top 10, 25 of higher secondary class, room board, books, personal, outstate, PhD, terminal, S.F ratio, expenditure and Graduation ratio are positively correlated.

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

After removing the Names variable from the dataset, it has 18 numerical columns with different scales. Z-score method has been used for this case study. This method tells us how much the standard deviation is away from the mean and in which direction.

In this case study-

- ✓ The application, accepted application, enrolled fulltime graduates, part-time graduates, outstate denote the number of students.
- ✓ The top10 percent and top20 percent students are denoted in percentage.
- ✓ Room board, books, and personal are denoted with money.
- ✓ The PhD, SF ratio, percentage of alumni are denoted with percentage values of different students.
- ✓ The graduation rate is denoted by the percentage value of graduates who get graduated every year.

Since, different types of numerical representation is used in the dataset, scaling is used as PCA requires that all the variables in dataset have similar scale of measurement.

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

The comparison between the covariance and correlation matrix is that they both measure the relationship and the dependency between two variables.

Covariance indicates the direction of the linear relationship between the variables whether it is positive (directly proportional) or negative (inversely proportional).

The Covariance matrix for this case study is as follows-

Covariance Matrix
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.81554018
  0.3987775   0.05022367  0.16515151  0.13272942  0.17896117  0.39120081
  0.36996762  0.09575627 -0.09034216  0.2599265   0.14694372]
 [ 0.94466636  1.00128866  0.91281145  0.19269493  0.24779465  0.87534985
  0.44183938 -0.02578774  0.09101577  0.11367165  0.20124767  0.35621633
  0.3380184   0.17645611 -0.16019604  0.12487773  0.06739929]
 [ 0.84791332  0.91281145  1.00128866  0.18152715  0.2270373   0.96588274
  0.51372977 -0.1556777  -0.04028353  0.11285614  0.28129148  0.33189629
  0.30867133  0.23757707 -0.18102711  0.06425192 -0.02236983]
 [ 0.33927032  0.19269493  0.18152715  1.00128866  0.89314445  0.1414708
 -0.10549205  0.5630552   0.37195909  0.1190116  -0.09343665  0.53251337
  0.49176793 -0.38537048  0.45607223  0.6617651   0.49562711]
 [ 0.35209304  0.24779465  0.2270373   0.89314445  1.00128866  0.19970167
 -0.05364569  0.49002449  0.33191707  0.115676   -0.08091441  0.54656564
  0.52542506 -0.29500852  0.41840277  0.52812713  0.47789622]
 [ 0.81554018  0.87534985  0.96588274  0.1414708   0.19970167  1.00128866
  0.57124738 -0.21602002 -0.06897917  0.11569867  0.31760831  0.3187472
  0.30040557  0.28006379 -0.22975792  0.01867565 -0.07887464]
 [ 0.3987775   0.44183938  0.51372977 -0.10549205 -0.05364569  0.57124738
  1.00128866 -0.25383901 -0.06140453  0.08130416  0.32029384  0.14930637
  0.14208644  0.23283016 -0.28115421 -0.08367612 -0.25733218]
 [ 0.05022367 -0.02578774 -0.1556777   0.5630552   0.49002449 -0.21602002
 -0.25383901  1.00128866  0.65509951  0.03890494 -0.29947232  0.38347594
  0.40850895 -0.55553625  0.56699214  0.6736456   0.57202613]
 [ 0.16515151  0.09101577 -0.04028353  0.37195909  0.33191707 -0.06897917
 -0.06140453  0.65509951  1.00128866  0.12812787 -0.19968518  0.32962651
  0.3750222  -0.36309504  0.27271444  0.50238599  0.42548915]
 [ 0.13272942  0.11367165  0.11285614  0.1190116   0.115676    0.11569867
  0.08130416  0.03890494  0.12812787  1.00128866  0.17952581  0.0269404
  0.10008351 -0.03197042 -0.04025955  0.11255393  0.00106226]
 [ 0.17896117  0.20124767  0.28129148 -0.09343665 -0.08091441  0.31760831
  0.32029384 -0.29947232 -0.19968518  0.17952581  1.00128866 -0.01094989
 -0.03065256  0.13652054 -0.2863366  -0.09801804 -0.26969106]
 [ 0.39120081  0.35621633  0.33189629  0.53251337  0.54656564  0.3187472
  0.14930637  0.38347594  0.32962651  0.0269404  -0.01094989  1.00128866
  0.85068186 -0.13069832  0.24932955  0.43331936  0.30543094]
 [ 0.36996762  0.3380184   0.30867133  0.49176793  0.52542506  0.30040557
  0.14208644  0.40850895  0.3750222   0.10008351 -0.03065256  0.85068186
  1.00128866 -0.16031027  0.26747453  0.43936469  0.28990033]
 [ 0.09575627  0.17645611  0.23757707 -0.38537048 -0.29500852  0.28006379
  0.23283016 -0.55553625 -0.36309504 -0.03197042  0.13652054 -0.13069832
 -0.16031027  1.00128866 -0.4034484  -0.5845844  -0.30710565]
 [-0.09034216 -0.16019604 -0.18102711  0.45607223  0.41840277 -0.22975792

```
 -0.28115421  0.56699214  0.27271444 -0.04025955 -0.2863366   0.24932955
  0.26747453 -0.4034484   1.00128866  0.41825001  0.49153016]
[ 0.2599265   0.12487773  0.06425192  0.6617651   0.52812713  0.01867565
 -0.08367612  0.6736456   0.50238599  0.11255393 -0.09801804  0.43331936
  0.43936469 -0.5845844   0.41825001  1.00128866  0.39084571]
[ 0.14694372  0.06739929 -0.02236983  0.49562711  0.47789622 -0.07887464
 -0.25733218  0.57202613  0.42548915  0.00106226 -0.26969106  0.30543094
  0.28990033 -0.30710565  0.49153016  0.39084571  1.00128866]]
```

Correlation measures the strength (positively correlated or negatively correlated) and the direction of the linear relationship between two variables.

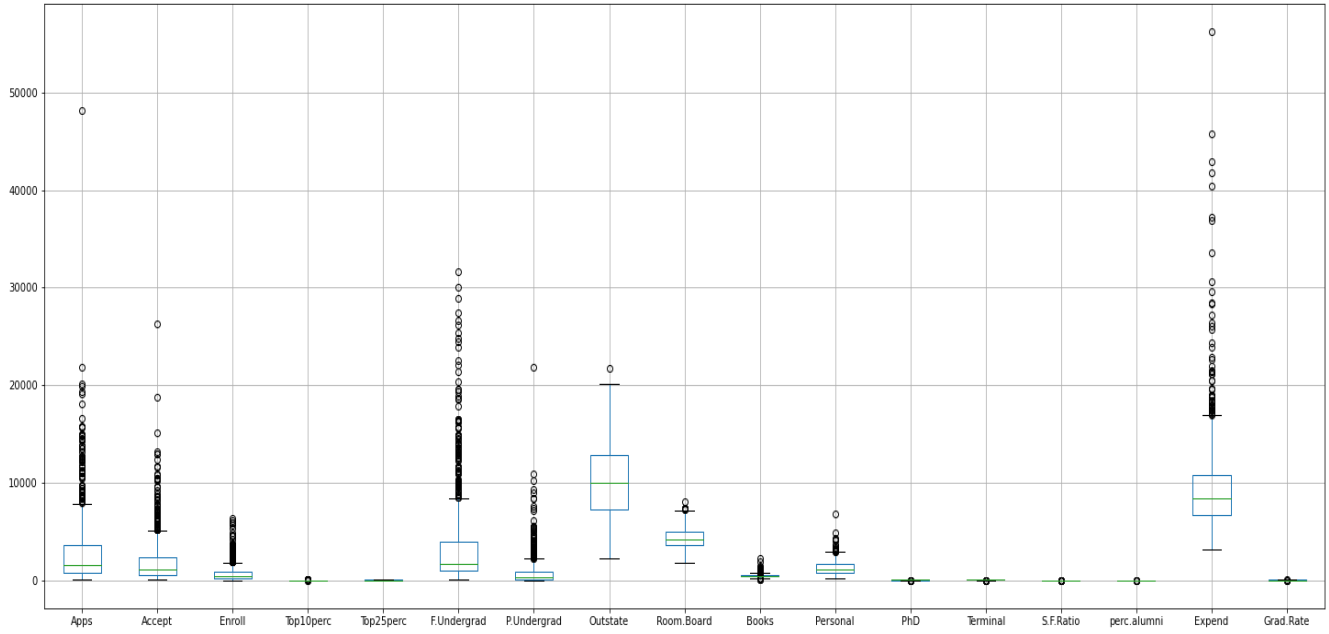
Application, Acceptance, Enrollment and Fulltime graduates are highly positively correlated. Top 10 percentage and Top 25 percentage are positively correlated.

Correlation Matrix (***refer to the Jupyter file to read the values of the matrix***) for this case study is-

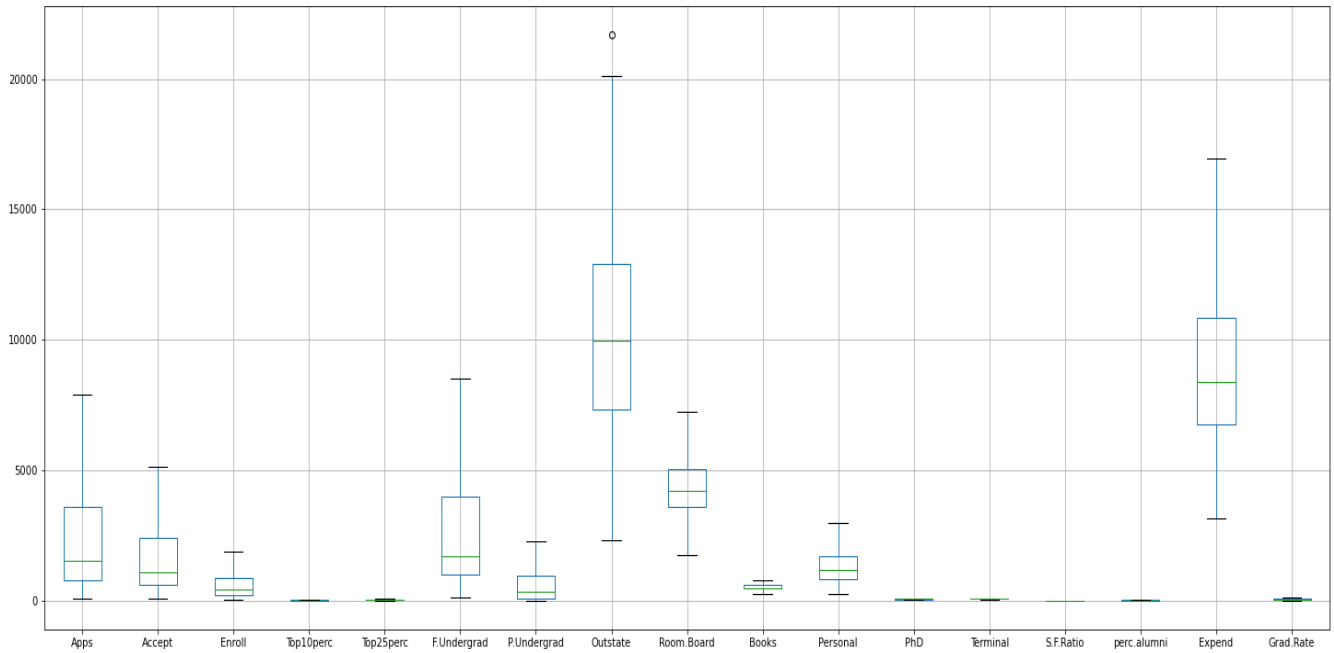
| | Apps | Accept | Enroll | Top10perc | Top25perc | F.Undergrad | P.Undergrad | Outstate | Room.Board | Books | Personal | PhD | Terminal | S.F.Ratio | perc.alumni | Expend | Grad.Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Apps | 1.000000 | 0.943451 | 0.846822 | 0.338834 | 0.351640 | 0.814491 | 0.398264 | 0.050159 | 0.164939 | 0.132559 | 0.178731 | 0.390697 | 0.369491 | 0.095633 | -0.090226 | 0.259592 | 0.146755 |
| Accept | 0.943451 | 1.000000 | 0.911637 | 0.192447 | 0.247476 | 0.874223 | 0.441271 | -0.025755 | 0.090899 | 0.113525 | 0.200989 | 0.355758 | 0.337583 | 0.176229 | -0.159990 | 0.124717 | 0.067313 |
| Enroll | 0.846822 | 0.911637 | 1.000000 | 0.181294 | 0.226745 | 0.964640 | 0.513069 | -0.155477 | -0.040232 | 0.112711 | 0.280929 | 0.331469 | 0.308274 | 0.237271 | -0.180794 | 0.064169 | -0.022341 |
| Top10perc | 0.338834 | 0.192447 | 0.181294 | 1.000000 | 0.891995 | 0.141289 | -0.105356 | 0.562331 | 0.371480 | 0.118858 | -0.093316 | 0.531828 | 0.491135 | -0.384875 | 0.455485 | 0.660913 | 0.494989 |
| Top25perc | 0.351640 | 0.247476 | 0.226745 | 0.891995 | 1.000000 | 0.199445 | -0.053577 | 0.489394 | 0.331490 | 0.115527 | -0.080810 | 0.545862 | 0.524749 | -0.294629 | 0.417864 | 0.527447 | 0.477281 |
| F.Undergrad | 0.814491 | 0.874223 | 0.964640 | 0.141289 | 0.199445 | 1.000000 | 0.570512 | -0.215742 | -0.068890 | 0.115550 | 0.317200 | 0.318337 | 0.300019 | 0.279703 | -0.229462 | 0.018652 | -0.078773 |
| P.Undergrad | 0.398264 | 0.441271 | 0.513069 | -0.105356 | -0.053577 | 0.570512 | 1.000000 | -0.253512 | -0.061326 | 0.081200 | 0.319882 | 0.149114 | 0.141904 | 0.232531 | -0.280792 | -0.083568 | -0.257001 |
| Outstate | 0.050159 | -0.025755 | -0.155477 | 0.562331 | 0.489394 | -0.215742 | -0.253512 | 1.000000 | 0.654256 | 0.038855 | -0.299087 | 0.382962 | 0.407983 | -0.554821 | 0.566262 | 0.672779 | 0.571290 |
| Room.Board | 0.164939 | 0.090899 | -0.040232 | 0.371480 | 0.331490 | -0.068890 | -0.061326 | 0.654256 | 1.000000 | 0.127963 | -0.199428 | 0.329202 | 0.374540 | -0.362628 | 0.272363 | 0.501739 | 0.424942 |
| Books | 0.132559 | 0.113525 | 0.112711 | 0.118858 | 0.115527 | 0.115550 | 0.081200 | 0.038855 | 0.127963 | 1.000000 | 0.179295 | 0.026906 | 0.099955 | -0.031929 | -0.040208 | 0.112409 | 0.001061 |
| Personal | 0.178731 | 0.200989 | 0.280929 | -0.093316 | -0.080810 | 0.317200 | 0.319882 | -0.299087 | -0.199428 | 0.179295 | 1.000000 | -0.010936 | -0.030613 | 0.136345 | -0.285968 | -0.097892 | -0.269344 |
| PhD | 0.390697 | 0.355758 | 0.331469 | 0.531828 | 0.545862 | 0.318337 | 0.149114 | 0.382962 | 0.329202 | 0.026906 | -0.010936 | 1.000000 | 0.849587 | -0.130530 | 0.249009 | 0.432762 | 0.305038 |
| Terminal | 0.369491 | 0.337583 | 0.308274 | 0.491135 | 0.524749 | 0.300019 | 0.141904 | 0.407983 | 0.374540 | 0.099955 | -0.030613 | 0.849587 | 1.000000 | -0.180194 | 0.267130 | 0.430799 | 0.289527 |
| S.F.Ratio | 0.095633 | 0.176229 | 0.237271 | -0.384875 | -0.294629 | 0.279703 | 0.232531 | -0.554821 | -0.362628 | -0.031929 | 0.136345 | -0.130530 | -0.180194 | 1.000000 | -0.402929 | -0.583832 | -0.306710 |
| perc.alumni | -0.090226 | -0.159990 | -0.180794 | 0.455485 | 0.417864 | -0.229462 | -0.280792 | 0.566262 | 0.272363 | -0.040208 | -0.285968 | 0.249009 | 0.267130 | -0.402929 | 1.000000 | 0.417712 | 0.490898 |
| Expend | 0.259592 | 0.124717 | 0.064169 | 0.660913 | 0.527447 | 0.018652 | -0.083568 | 0.672779 | 0.501739 | 0.112409 | -0.097892 | 0.432762 | 0.430799 | -0.583832 | 0.417712 | 1.000000 | 0.390343 |
| Grad.Rate | 0.146755 | 0.067313 | -0.022341 | 0.494989 | 0.477281 | -0.078773 | -0.257001 | 0.571290 | 0.424942 | 0.001061 | -0.269344 | 0.305038 | 0.289527 | -0.306710 | 0.490898 | 0.390343 | 1.000000 |

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

DATASET BEFORE SCALING




DATASET AFTER SCALING



INSIGHTS

The outliers are still present in dataset because scaling does not remove outliers scaling scales the values on a Z score distribution.

In this case study, outliers are imputed using IQR (***refer to the Jupyter file to see the code to remove outliers***).

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

Eigen Vectors

```
%s [[-2.48765602e-01  3.31598227e-01 -6.30921033e-02  2.81310530e-01
  -5.74140964e-03 -1.62374420e-02 -4.24863486e-02 -1.03090398e-01
  -9.02270802e-02  5.25098025e-02 -3.58970400e-01  4.59139498e-01
  -4.30462074e-02  1.33405806e-01 -8.06328039e-02 -5.95830975e-01
   2.40709086e-02]
 [-2.07601502e-01  3.72116750e-01 -1.01249056e-01  2.67817346e-01
  -5.57860920e-02  7.53468452e-03 -1.29497196e-02 -5.62709623e-02
  -1.77864814e-01  4.11400844e-02  5.43427250e-01 -5.18568789e-01
   5.84055850e-02 -1.45497511e-01 -3.34674281e-02 -2.92642398e-01
  -1.45102446e-01]
 [-1.76303592e-01  4.03724252e-01 -8.29855709e-02  1.61826771e-01
   5.56936353e-02 -4.25579803e-02 -2.76928937e-02  5.86623552e-02
  -1.28560713e-01  3.44879147e-02 -6.09651110e-01 -4.04318439e-01
   6.93988831e-02  2.95896092e-02  8.56967180e-02  4.44638207e-01
   1.11431545e-02]
 [-3.54273947e-01 -8.24118211e-02  3.50555339e-02 -5.15472524e-02
   3.95434345e-01 -5.26927980e-02 -1.61332069e-01 -1.22678028e-01
   3.41099863e-01  6.40257785e-02  1.44986329e-01 -1.48738723e-01
   8.10481404e-03  6.97722522e-01  1.07828189e-01 -1.02303616e-03
   3.85543001e-02]
 [-3.44001279e-01 -4.47786551e-02 -2.41479376e-02 -1.09766541e-01
   4.26533594e-01  3.30915896e-02 -1.18485556e-01 -1.02491967e-01
   4.03711989e-01  1.45492289e-02 -8.03478445e-02  5.18683400e-02
   2.73128469e-01 -6.17274818e-01 -1.51742110e-01 -2.18838802e-02
  -8.93515563e-02]
 [-1.54640962e-01  4.17673774e-01 -6.13929764e-02  1.00412335e-01
   4.34543659e-02 -4.34542349e-02 -2.50763629e-02  7.88896442e-02
  -5.94419181e-02  2.08471834e-02  4.14705279e-01  5.60363054e-01
   8.11578181e-02  9.91640992e-03  5.63728817e-02  5.23622267e-01
   5.61767721e-02]
 [-2.64425045e-02  3.15087830e-01  1.39681716e-01 -1.58558487e-01
  -3.02385408e-01 -1.91198583e-01  6.10423460e-02  5.70783816e-01
   5.60672902e-01 -2.23105808e-01 -9.01788964e-03 -5.27313042e-02
  -1.00693324e-01  2.09515982e-02 -1.92857500e-02 -1.25997650e-01
  -6.35360730e-02]
```

[-2.94736419e-01 -2.49643522e-01  4.65988731e-02  1.31291364e-01
 -2.22532003e-01 -3.00003910e-02  1.08528966e-01  9.84599754e-03
 -4.57332880e-03  1.86675363e-01 -5.08995918e-02  1.01594830e-01
 -1.43220673e-01  3.83544794e-02  3.40115407e-02  1.41856014e-01
 -8.23443779e-01]
[-2.49030449e-01 -1.37808883e-01  1.48967389e-01  1.84995991e-01
 -5.60919470e-01  1.62755446e-01  2.09744235e-01 -2.21453442e-01
  2.75022548e-01  2.98324237e-01 -1.14639620e-03 -2.59293381e-02
  3.59321731e-01  3.40197083e-03  5.84289756e-02  6.97485854e-02
  3.54559731e-01]
[-6.47575181e-02  5.63418434e-02  6.77411649e-01  8.70892205e-02
  1.27288825e-01  6.41054950e-01 -1.49692034e-01  2.13293009e-01
 -1.33663353e-01 -8.20292186e-02 -7.72631963e-04  2.88282896e-03
 -3.19400370e-02 -9.43887925e-03  6.68494643e-02 -1.14379958e-02
 -2.81593679e-02]
[ 4.25285386e-02  2.19929218e-01  4.99721120e-01 -2.30710568e-01
  2.22311021e-01 -3.31398003e-01  6.33790064e-01 -2.32660840e-01
 -9.44688900e-02  1.36027616e-01  1.11433396e-03 -1.28904022e-02
  1.85784733e-02 -3.09001353e-03 -2.75286207e-02 -3.94547417e-02
 -3.92640266e-02]
[-3.18312875e-01  5.83113174e-02 -1.27028371e-01 -5.34724832e-01
 -1.40166326e-01  9.12555212e-02 -1.09641298e-03 -7.70400002e-02
 -1.85181525e-01 -1.23452200e-01 -1.38133366e-02  2.98075465e-02
 -4.03723253e-02 -1.12055599e-01  6.91126145e-01 -1.27696382e-01
  2.32224316e-02]
[-3.17056016e-01  4.64294477e-02 -6.60375454e-02 -5.19443019e-01
 -2.04719730e-01  1.54927646e-01 -2.84770105e-02 -1.21613297e-02
 -2.54938198e-01 -8.85784627e-02 -6.20932749e-03 -2.70759809e-02
  5.89734026e-02  1.58909651e-01 -6.71008607e-01  5.83134662e-02
  1.64850420e-02]
[ 1.76957895e-01  2.46665277e-01 -2.89848401e-01 -1.61189487e-01
  7.93882496e-02  4.87045875e-01  2.19259358e-01 -8.36048735e-02
  2.74544380e-01  4.72045249e-01  2.22215182e-03 -2.12476294e-02
 -4.45000727e-01 -2.08991284e-02 -4.13740967e-02  1.77152700e-02
 -1.10262122e-02]
[-2.05082369e-01 -2.46595274e-01 -1.46989274e-01  1.73142230e-02
  2.16297411e-01 -4.73400144e-02  2.43321156e-01  6.78523654e-01
 -2.55334907e-01  4.22999706e-01  1.91869743e-02  3.33406243e-03
  1.30727978e-01 -8.41789410e-03  2.71542091e-02 -1.04088088e-01
  1.82660654e-01]
[-3.18908750e-01 -1.31689865e-01  2.26743985e-01  7.92734946e-02
 -7.59581203e-02 -2.98118619e-01 -2.26584481e-01 -5.41593771e-02
 -4.91388809e-02  1.32286331e-01  3.53098218e-02 -4.38803230e-02
 -6.92088870e-01 -2.27742017e-01 -7.31225166e-02  9.37464497e-02
  3.25982295e-01]
[-2.52315654e-01 -1.69240532e-01 -2.08064649e-01  2.69129066e-01

1.09267913e-01  2.16163313e-01  5.59943937e-01 -5.33553891e-03
4.19043052e-02 -5.90271067e-01  1.30710024e-02 -5.00844705e-03
-2.19839000e-01 -3.39433604e-03 -3.64767385e-02  6.91969778e-02
1.22106697e-01]]

Eigen Values

%s [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]

**2.6** Perform PCA and export the data of the 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.262230 | 0.230624 | 0.189345 | 0.338871 | 0.334694 | 0.163364 | 0.022537 | 0.283419 | 0.244165 | 0.096734 | -0.035198 | 0.326418 | 0.323126 | -0.163118 | 0.188578 | 0.328949 | 0.238796 |
| 1 | 0.314086 | 0.344579 | 0.382776 | -0.099381 | -0.059567 | 0.398603 | 0.357544 | -0.251926 | -0.131948 | 0.093949 | 0.232450 | 0.055092 | 0.042982 | 0.259837 | -0.257128 | -0.160080 | -0.167560 |
| 2 | -0.080996 | -0.107629 | -0.085522 | 0.078731 | 0.050686 | -0.073699 | -0.040299 | -0.014779 | 0.021247 | 0.697109 | 0.530972 | -0.081157 | -0.059012 | -0.274183 | -0.103779 | 0.184231 | -0.246356 |
| 3 | 0.096893 | 0.118256 | 0.009421 | -0.369062 | -0.416766 | 0.014042 | 0.225334 | 0.263127 | 0.580772 | -0.036305 | -0.115165 | -0.147493 | -0.089232 | -0.259459 | -0.223892 | 0.213763 | -0.036091 |
| 4 | 0.219854 | 0.189592 | 0.162328 | 0.157346 | 0.144606 | 0.102739 | -0.095765 | 0.037403 | -0.069435 | 0.035393 | -0.000373 | -0.550741 | -0.590385 | -0.142072 | 0.128301 | -0.022538 | 0.356800 |
| 5 | 0.002202 | -0.016510 | -0.068076 | -0.088870 | -0.027635 | -0.051646 | -0.024533 | -0.020424 | 0.237310 | 0.638604 | -0.381497 | 0.003317 | 0.035382 | 0.468715 | 0.012561 | -0.231553 | 0.313588 |
| 6 | -0.028340 | -0.012919 | -0.015204 | -0.257599 | -0.239198 | -0.031141 | -0.009996 | 0.094239 | 0.094533 | -0.111137 | 0.639264 | 0.089264 | 0.091784 | 0.152058 | 0.391649 | -0.150372 | 0.468649 |
| 7 | -0.089954 | -0.137615 | -0.144235 | 0.289416 | 0.345527 | -0.108767 | 0.123887 | 0.010994 | 0.389817 | -0.239880 | 0.277411 | -0.034219 | -0.090252 | 0.242697 | -0.565908 | -0.118831 | 0.180745 |
| 8 | -0.130530 | -0.142229 | -0.050846 | 0.122497 | 0.193949 | -0.001451 | 0.634804 | 0.008703 | 0.220477 | -0.021035 | -0.017332 | -0.166554 | -0.112653 | 0.153950 | 0.539234 | -0.024219 | -0.316940 |
| 9 | -0.156399 | -0.149176 | -0.064872 | -0.035823 | 0.006476 | -0.000151 | 0.546414 | -0.232378 | -0.254709 | 0.091182 | -0.127801 | 0.106897 | 0.085925 | -0.470803 | -0.147213 | 0.080508 | 0.488149 |
| 10 | -0.086434 | -0.042769 | -0.043825 | 0.001604 | -0.101816 | -0.034874 | 0.252354 | 0.593994 | -0.475559 | 0.043590 | 0.015264 | -0.039163 | -0.084277 | 0.362556 | -0.174291 | 0.392739 | 0.087687 |
| 11 | -0.089573 | -0.158544 | 0.035076 | 0.039689 | -0.146262 | 0.132962 | -0.049774 | -0.559506 | 0.106827 | -0.051378 | -0.009535 | 0.073842 | -0.166316 | 0.240169 | 0.048034 | 0.690530 | 0.158937 |
| 12 | -0.089362 | -0.044502 | 0.062227 | -0.069789 | 0.096536 | 0.087911 | -0.044833 | -0.069298 | -0.017254 | -0.035593 | 0.011776 | -0.702378 | 0.661894 | 0.048570 | -0.035677 | 0.128891 | 0.063599 |
| 13 | -0.549345 | -0.291579 | 0.416942 | -0.008978 | 0.010930 | 0.570753 | -0.146416 | 0.211125 | 0.101164 | 0.028590 | -0.033797 | 0.064235 | -0.098780 | -0.062084 | -0.028021 | -0.128701 | 0.007188 |
| 14 | 0.005605 | 0.014452 | -0.049919 | -0.723634 | 0.655440 | 0.025204 | -0.039705 | -0.001885 | -0.028220 | -0.008088 | 0.001425 | 0.083125 | -0.113346 | 0.003863 | -0.007290 | 0.145276 | -0.003250 |
| 15 | 0.599192 | -0.661475 | -0.233289 | -0.022080 | -0.032318 | 0.367638 | -0.026257 | 0.081262 | -0.026715 | -0.010481 | -0.004534 | -0.012454 | 0.017926 | -0.018328 | 0.000099 | -0.056008 | -0.014807 |
| 16 | -0.182177 | 0.391054 | -0.716684 | 0.056206 | -0.019673 | 0.542765 | -0.029503 | -0.001053 | -0.009853 | -0.004362 | 0.010872 | -0.013315 | -0.007381 | -0.008857 | 0.024056 | -0.010555 | 0.002513 |

(***refer to the Jupyter file to read the value of the data frame clearly***)

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

The Linear equation of 1st principal component:

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?

```
array([ 32.0206282 ,  58.36084263,  65.26175919,  71.18474841,
        76.67315352,  81.65785448,  85.21672597,  88.67034731,
        91.78758099,  94.16277251,  96.00419883,  97.30024023,
        98.28599436,  99.13183669,  99.64896227,  99.86471628,
       100.        ])
```

- ✓ The Eigen values sum up to 100.
- ✓ To decide on the optimum number of principal components, the incremental value between the components should not be less than 5%.
- ✓ Based on this, the optimum number of components, for this case study is taken as 5.

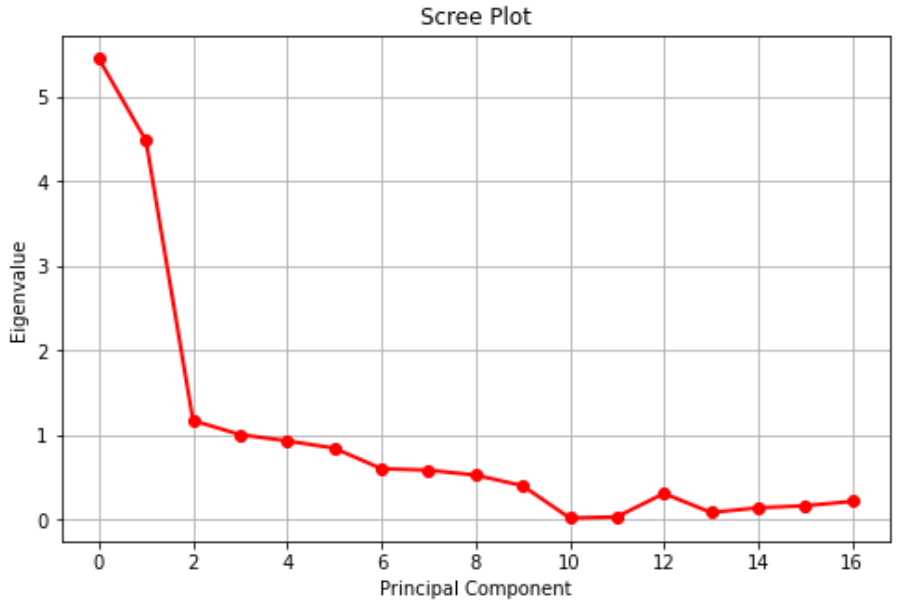
```
array([[ 0.2487656 ,  0.2076015 ,  0.17630359,  0.35427395,  0.34400128,
         0.15464096,  0.0264425 ,  0.29473642,  0.24903045,  0.06475752,
        -0.04252854,  0.31831287,  0.31705602, -0.17695789,  0.20508237,
         0.31890875,  0.25231565],
       [ 0.33159823,  0.37211675,  0.40372425, -0.08241182, -0.04477866,
         0.41767377,  0.31508783, -0.24964352, -0.13780888,  0.05634184,
         0.21992922,  0.05831132,  0.04642945,  0.24666528, -0.24659527,
        -0.13168986, -0.16924053],
       [-0.06309212, -0.10124904, -0.08298556,  0.03505554, -0.02414794,
        -0.06139299,  0.13968172,  0.04659887,  0.14896739,  0.67741165,
         0.49972112, -0.12702837, -0.06603754, -0.2898484 , -0.14698927,
         0.22674399, -0.20806465],
       [ 0.28131053,  0.26781734,  0.16182677, -0.05154725, -0.10976654,
         0.10041234, -0.15855849,  0.13129136,  0.18499599,  0.08708922,
        -0.23071057, -0.53472483, -0.51944302, -0.16118949,  0.01731422,
```

0.07927349,  0.26912907],
[ 0.0057414 ,  0.0557861 , -0.05569362, -0.39543434, -0.4265336 ,
 -0.04345438,  0.30238541,  0.222532  ,  0.56091947, -0.12728883,
 -0.22231102,  0.14016632,  0.20471973, -0.07938825, -0.21629741,
  0.07595812, -0.10926791]])

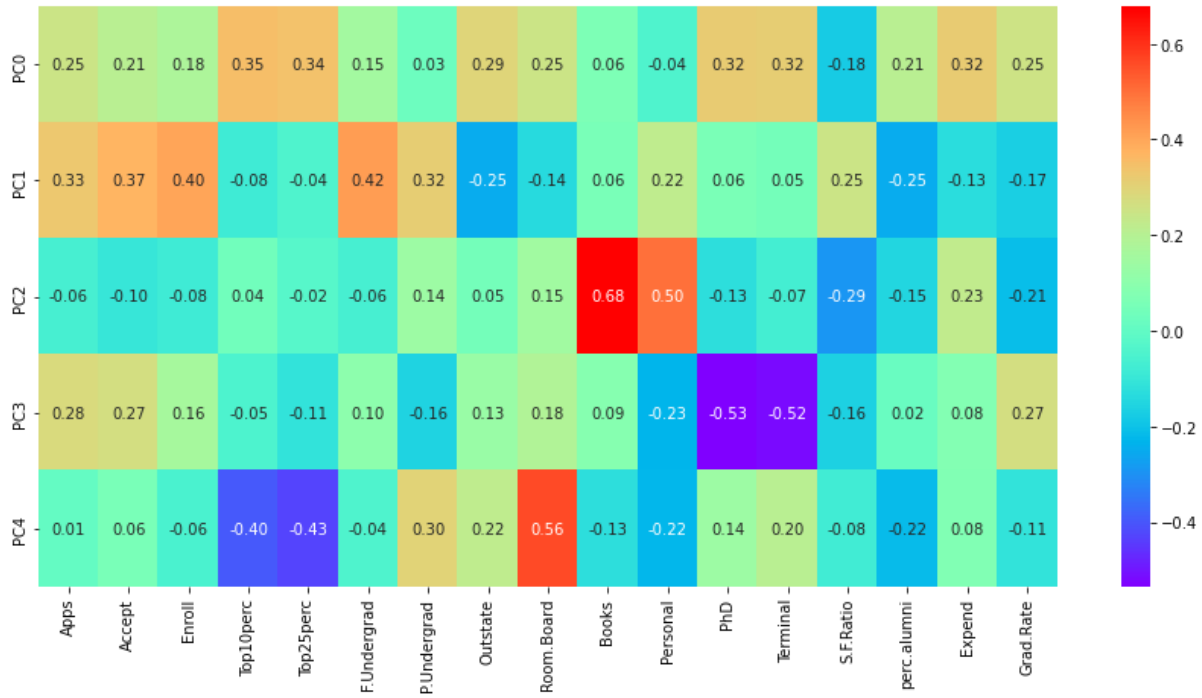
- ✓ The first component explains 32.02% variance in dataset.
- ✓ The first two components explain 58.36% variance in dataset.
- ✓ The first three components explain 65.26% variance in dataset.
- ✓ The first four components explain 71.18% variance in dataset.
- ✓ The first five components explain 76.67% variance in dataset.

The number of Eigen vectors for this case study is five. From these eigen vectors we can understand which variable has more weightage and influences the dataset in the principal components.

SCREE PLOT



HEATMAP




**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]

- ✓ To understand more about the dataset, we perform univariate analysis and multivariate analysis which gives us the understanding about the variables.
- ✓ From analysis the distribution of the dataset, skew, and patterns in the dataset is understood
- ✓ Multivariate analysis shows that multiple variables are highly correlated with each other.
- ✓ The scaling helps the dataset to bring the variables to one scale. Outliers are imputed using IQR values and then PCA is performed.
- ✓ The principal component analysis is used reduce the multi collinearity between the variables.
- ✓ The PCA components for this business case is 5 where maximum variance of the dataset is understood.
- ✓ Using the components, the reduced multi collinearity is understood in the dataset.

***THE END***