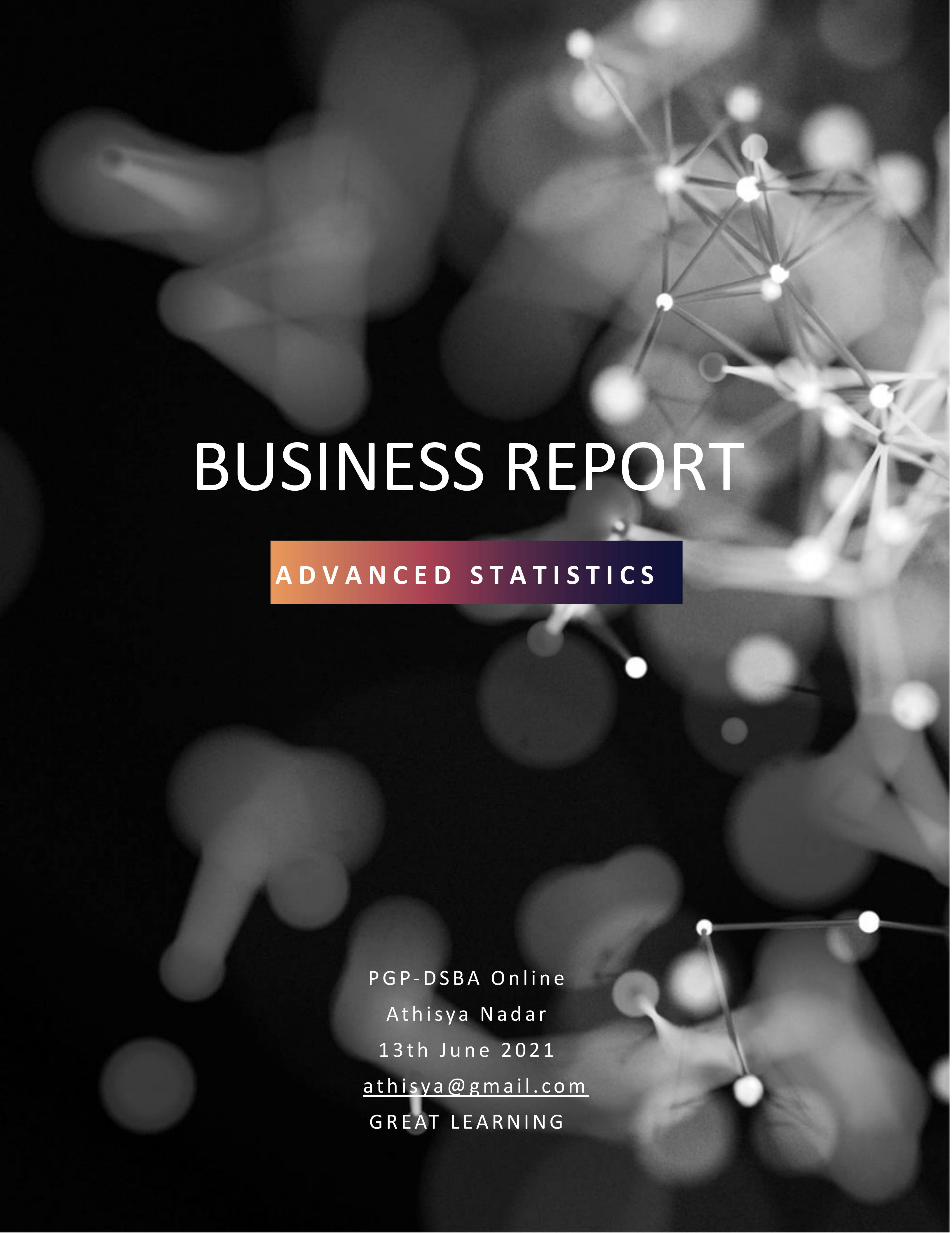


BUSINESS REPORT

ADVANCED STATISTICS

PGP-DSBA Online
Athisya Nadar
13th June 2021
athisya@gmail.com
GREAT LEARNING

Table of Contents

Problem 1A:

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals Salary Data 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. **State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.**
2. **Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.**
3. **Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.**
4. **If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. (Non-Graded)**

Problem 1B:

1. **What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the 'pointplot' function from the 'seaborn' function]**
2. **Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?**
3. **Explain the business implications of performing ANOVA for this particular case study.**

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

ANOVA stands for analysis of Analysis of Variance. We are going to look at the assumptions before proceeding with ANOVA:

- The samples drawn from different populations are independent and random.
- The response variables of all the populations are normally distributed.
- The variances of all the populations are equal.

We can use the Shapiro-Wilk Test to check normality. If the p-value ≤ 0.05, then we reject the null hypothesis i.e. we assume the distribution of our variable is not normal/gaussian. If the p-value > 0.05, then we fail to reject the null hypothesis i.e. we assume the distribution of our variable is normal/gaussian.

Anderson-Darling Normality Test is another general normality tests designed to determine if the data comes from a specified distribution, in our case, the normal distribution. It gives a range of critical values, at which the null hypothesis can be failed to rejected if the calculated statistic is less than the critical value.

The Levene test is a test of variance which tests the null hypothesis that all input samples are from populations with equal variances.

Since we are assuming that the data follows normal assumption, we can formulate the null hypothesis for conducting one-way ANOVA for both Education and Occupation individually as stated below:

Hypotheses of One-Way ANOVA(Education)

Null Hypothesis ($H0$): The mean salary earned is same across different education level

Alternate Hypothesis (HA): The mean salary is different in at least one education level

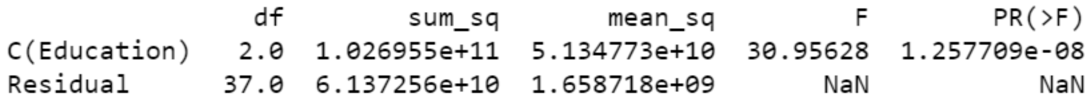
Hypotheses of One-Way ANOVA(Occupation)

Null Hypothesis($H0$): The mean salary earned is same across different occupation

Alternate Hypothesis(HA): The mean salary is different in at least one 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.

After performing the 1 way ANOVA in python for education with respect to the variable salary we have got the following result:

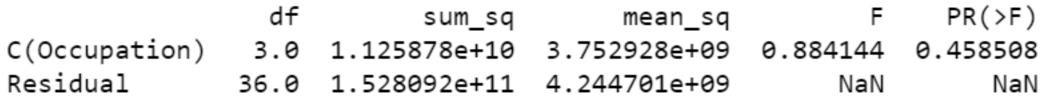


```
                 df        sum_sq       mean_sq         F        PR(>F)
C(Education)    2.0  1.026955e+11  5.134773e+10  30.95628  1.257709e-08
Residual       37.0  6.137256e+10  1.658718e+09       NaN           NaN
```

From this table we can clearly see that p-value<0.05, hence we reject the Null Hypothesis and accept the alternate hypothesis. Thus, we can say that the mean salary is different in at least one education level.

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.

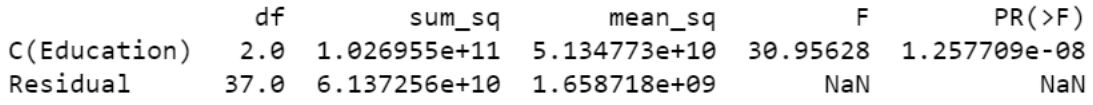
After performing 1-way ANOVA in Python for variable occupation with respect to the variable salary we get the following table:



```
                  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
```

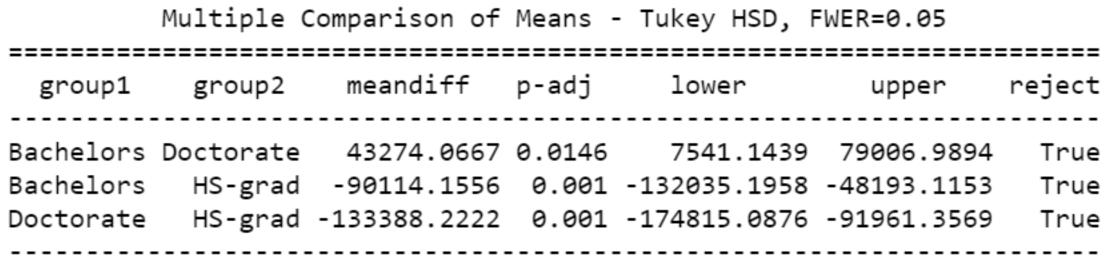
From this table we can conclude that p-value >0.05, hence we fail to reject the Null Hypothesis. Thus, we can say, the mean salary earned is same across different occupation.

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.



```
                 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 null hypothesis is rejected after performing the 1 way ANOVA in python for education with respect to the variable salary, we use The Tukey HSD Test to find out which class means are significantly different. What his test does is compare the differences between means of values rather than comparing pairs of values.

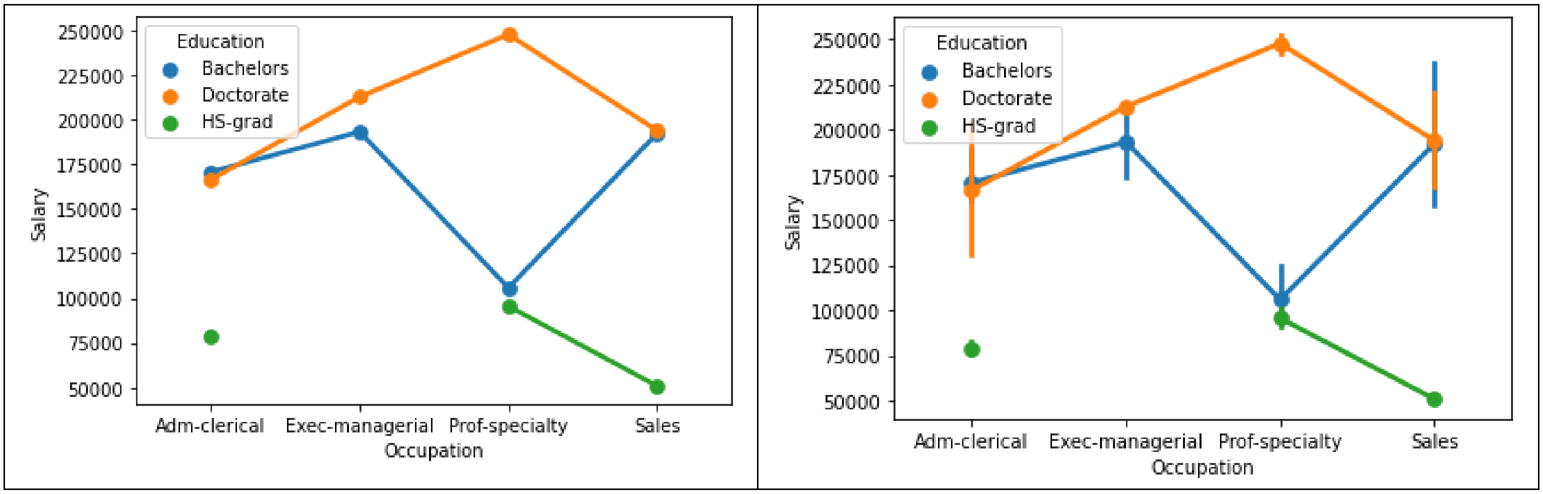


```
          Multiple Comparison of Means - Tukey HSD, FWER=0.05
=======================================================================
  group1    group2    meandiff   p-adj     lower        upper    reject
-----------------------------------------------------------------------
Bachelors Doctorate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

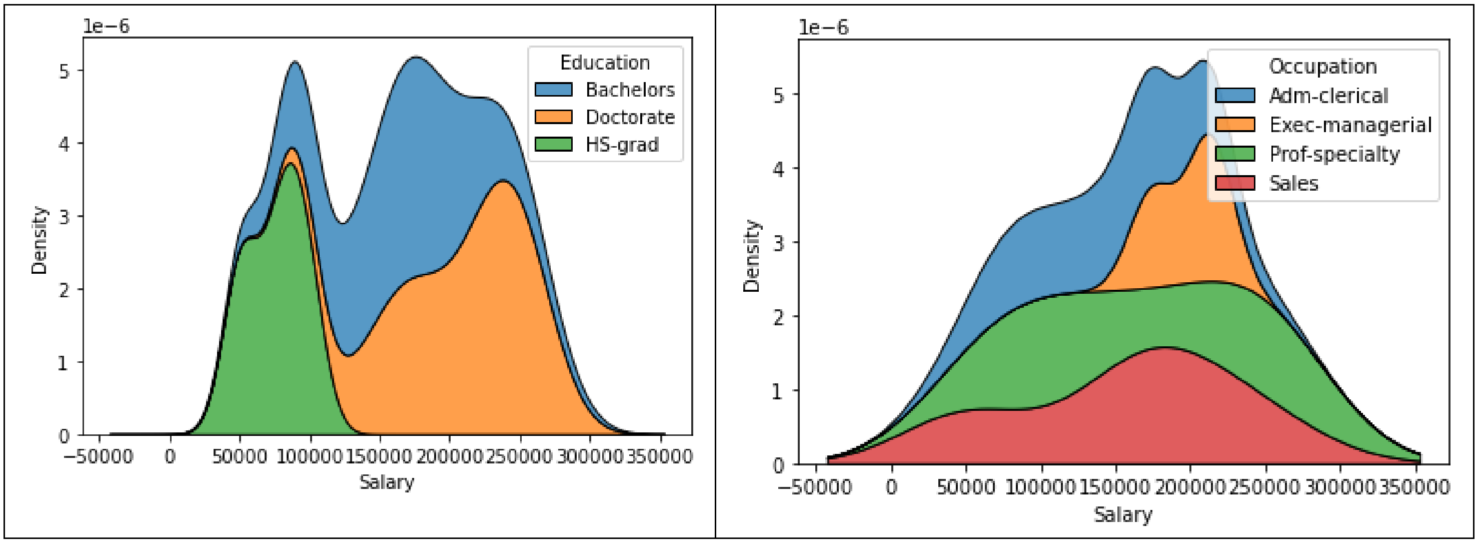
F-Distribution or F-Statistic is the ratio of MSB to MSW. It gives the degree of how relatively greater the difference is 'between group means' (MSB) compared to 'within group variance' (MSW). If the ratio is greater than expected will mean that not all the group means are the same and at least one mean is substantially different.

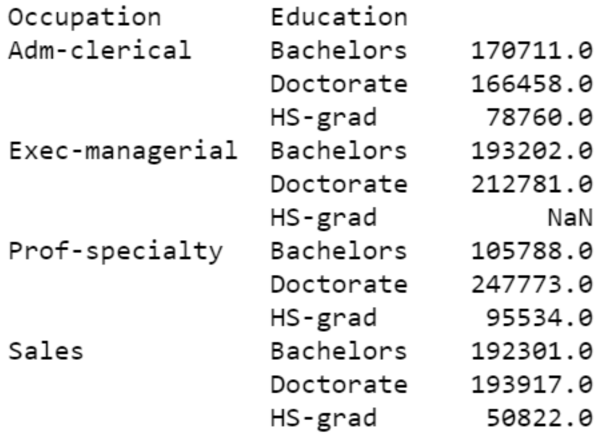
Since p-adj is less than 0.05 in all 3 scenarios, the means are different across all three education levels. Thus, mean salary(Doctorate) not equal to mean salary(Bachelors) not equal to mean salary (HSgrad)

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.



When we try to analyze the effect of one variable on another, i.e effect of Education on Occupation, we can clearly see that the salary earned by Bachelor’s degree holder is greater than doctorate among Adm-clerical and among Sales Occupation salary earned by Bachelors and Doctorates is same. None of the Exec-managerial occupation has anybody with HS grad education level.





```
Occupation        Education
Adm-clerical      Bachelors     170711.0
                  Doctorate     166458.0
                  HS-grad        78760.0
Exec-managerial   Bachelors     193202.0
                  Doctorate     212781.0
                  HS-grad            NaN
Prof-specialty    Bachelors     105788.0
                  Doctorate     247773.0
                  HS-grad        95534.0
Sales             Bachelors     192301.0
                  Doctorate     193917.0
                  HS-grad        50822.0
```

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?

The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors). The primary purpose of a two-way ANOVA to understand if there is an interaction between the two independent variables on the dependent variable.

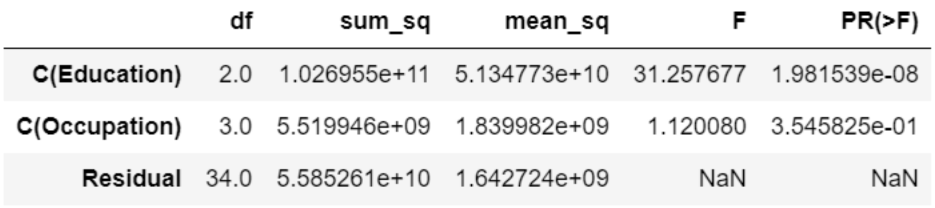
Assumptions of two-way ANOVA

- Dependent variable should be measured at the continuous level.

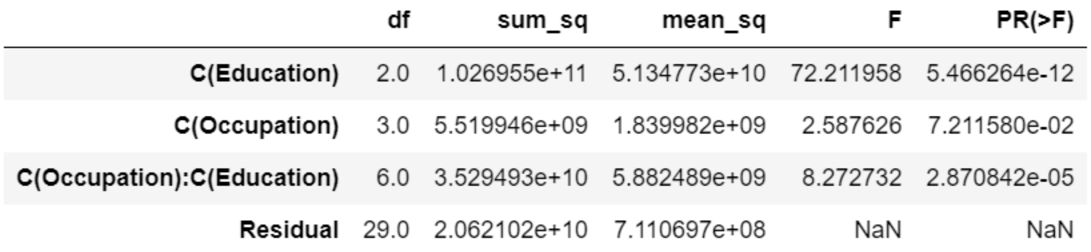
- Two independent variables should each consist of two or more categorical, independent groups.

- There should be no significant outliers.

- Dependent variable should be approximately normally distributed for each combination of the groups of the two independent variables



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

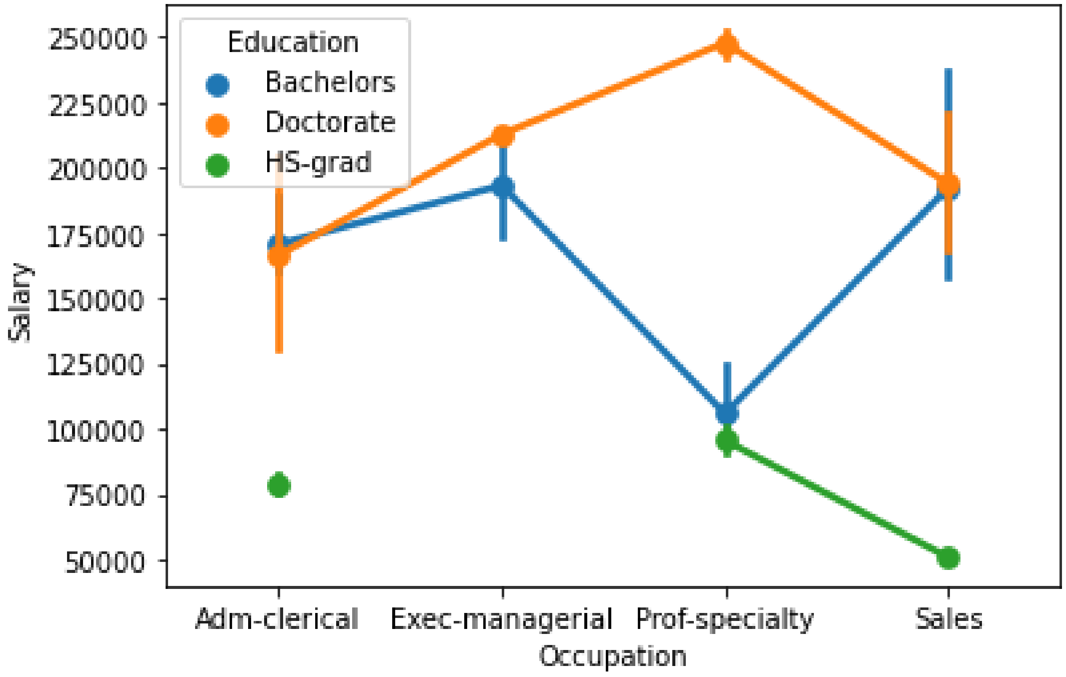


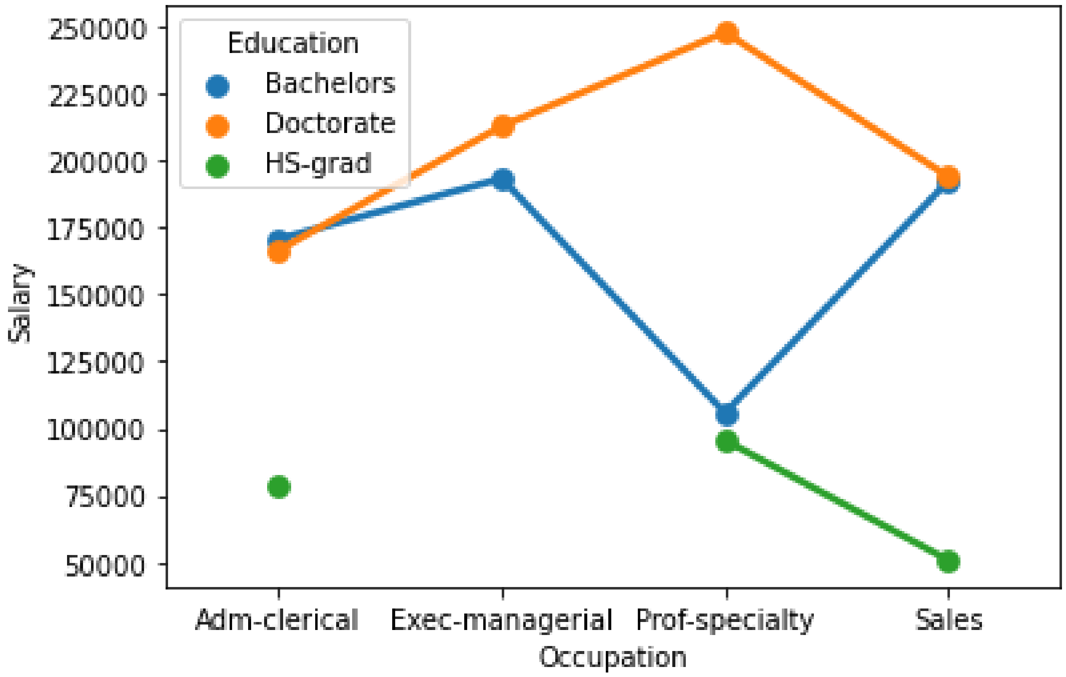
	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Occupation):C(Education)	6.0	3.529493e+10	5.882489e+09	8.272732	2.870842e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

The p-value for Education is 5.466264e-12, which indicates that the levels of Education are associated with different Salary.

The p-value for Occupation is 7.211580e-02, which indicates that the levels of Occupation are not associated with different Salary.

The p-value for the interaction between Education* Occupation is 2.232500e-05, which indicates that the relationship between Occupation and Salary depends on the value of Education.





When we try to analyze the effect of one variable on another, i.e effect of Education on Occupation, we can clearly see that the salary earned by Bachelor's degree holder is greater than doctorate among Adm-clerical and among Sales Occupation salary earned by Bachelors and Doctorates is same. None of the Exec-managerial occupation has anybody with HS grad education level.

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

Performing ANOVA in a business context, helps to budget and manage revenue. This analysis can be used to forecast future performance. Effective variance analysis can help a company spot trends, issues, opportunities and threats to short-term or long-term success.

Variance analysis helps maintain control over a project's expenses by monitoring planned versus actual costs. This analysis will help the business to manage human resource in a more resourceful way. Making optimum use of the talent pool across different education level and Occupation can be a challenge sometimes, this variance study can help to shed some light on how we can improve the approach towards addressing these issues.

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.

- **Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?**
- **Is scaling necessary for PCA in this case? Give justification and perform scaling.**
- **Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].**
- **Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]**

- **Extract the eigenvalues and eigenvectors.[print both]**
- **Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features**
- **Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]**
- **Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?**
- **Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]**

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

Exploratory data analysis helps to deal with bad values, anomalies, removing duplicates, missing values, outlier treatment, normalizing and scaling, encoding categorical variable, Univariate, bivariate and multivariate analysis.

Removing duplicates helps to decrease computation time, after checking the dataset for duplicates we do not find any duplicate values.

Missing values affect mean, mode, median. Models do not work if there are missing values in the dataset, after checking for missing values we could not find any missing value in the dataset.

Treating **outliers** is crucial to get the best fit for models, so we have replaced outliers with upper range and lower range to minimize the effect of outliers. This deals with data within the column.

Normalizing and scaling helps to scale the data between -3 or 3 or between 0 to 1. There are different types of scaler: Min/max scaler is a subset of standard scaler which scales the data between 0 and 1. Z-score scaling scales the data such that the mean is 0 and standard deviation is 1.

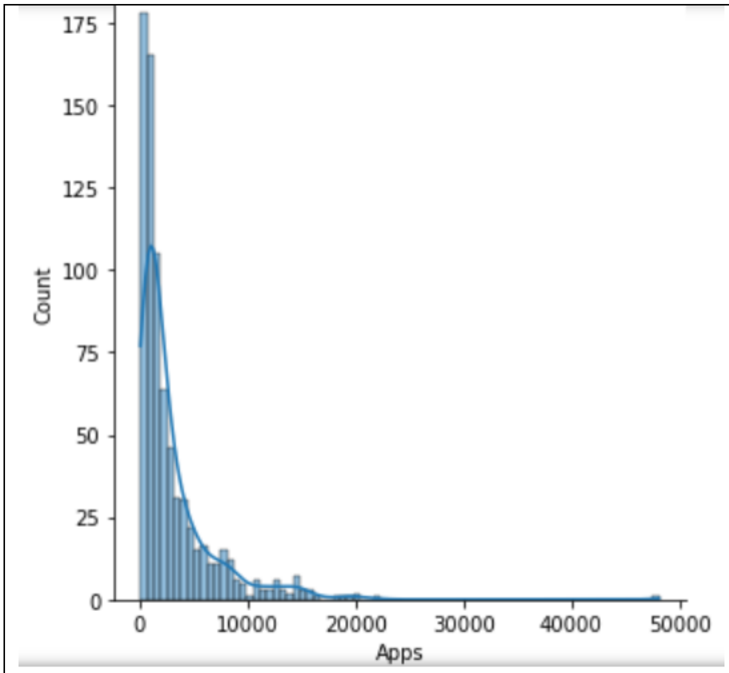
Inspecting the distributions of the features is a very important step, If the features are not normally distributed, have a high Skewness then using the data as is may produce misleading results. When the distribution of the continuous data is non-normal, **transformations** of data are applied to make the data as "normal" as possible and, thus, increase the validity of the associated statistical analyses. Inspect the skewness of the feature, apply Log, Square root or Cube root for removing Right-skewness. Square, Cube or exponential to remove the left skewness.

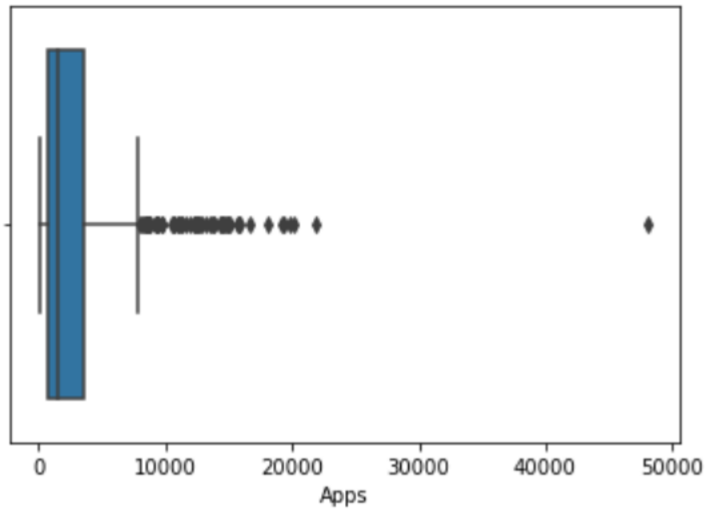
Encoding allows us to change categorical data to numerical continuous data. Three most commonly used Encoding techniques - 1. One-Hot Encoding : In this technique, for each category of a feature, we create a new column (sometimes called a dummy variable) with binary encoding (0 or 1) to denote whether a particular row belongs to this category. Label Encoding: In label encoding, we replace the categorical value with a numeric value between 0 and the number of classes minus 1. Ordinal Encoding An Ordinal Encoder is used to encode categorical features into an ordinal numerical value.

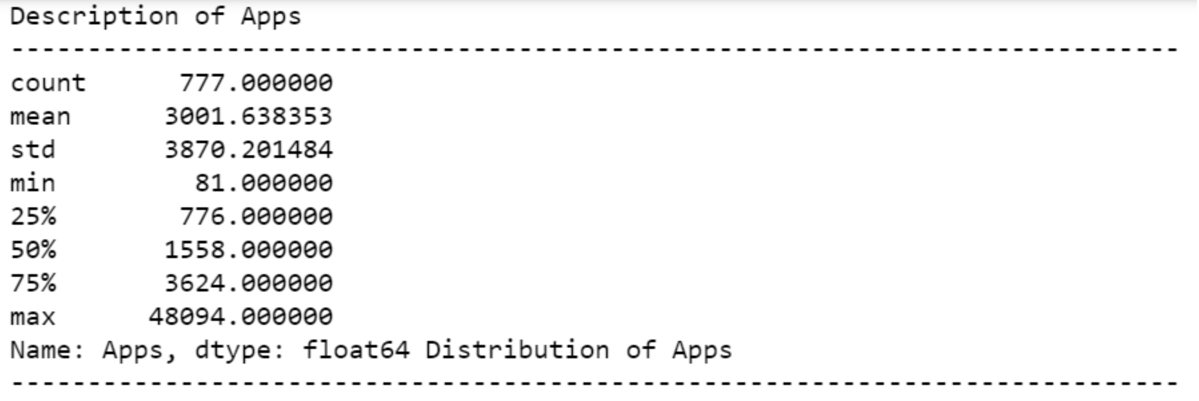
Univariate Analysis

The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately. It is possible for two kinds of variables- Categorical and Numerical.

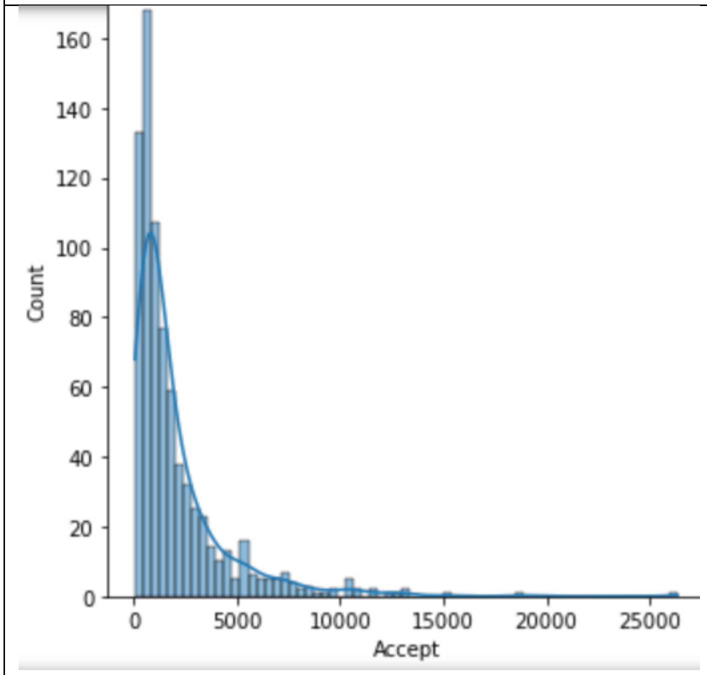
Some patterns that can be easily identified with univariate analysis are Central Tendency (mean, mode and median), Dispersion (range, variance), Quartiles (interquartile range), and Standard deviation.

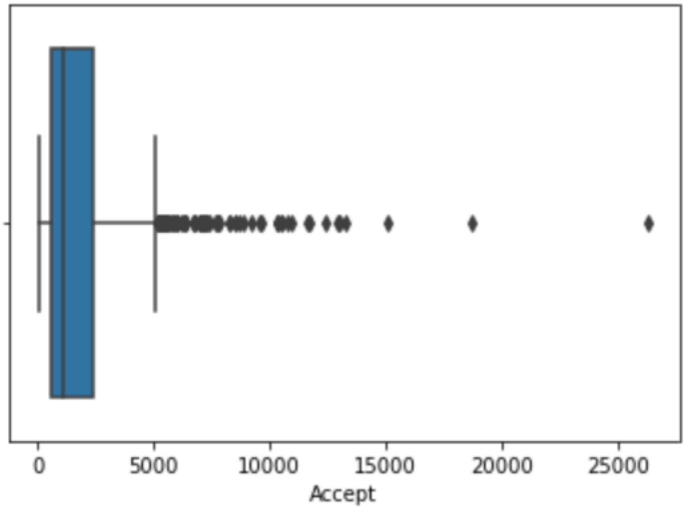


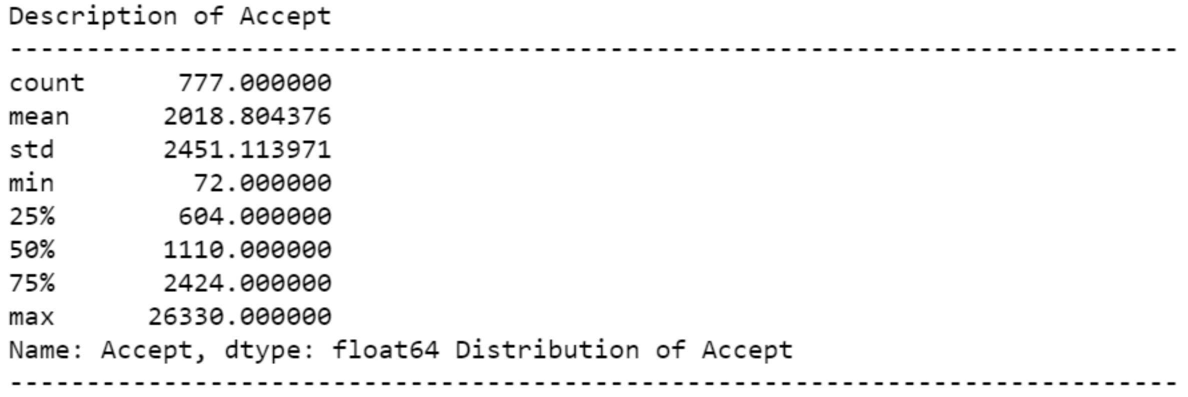




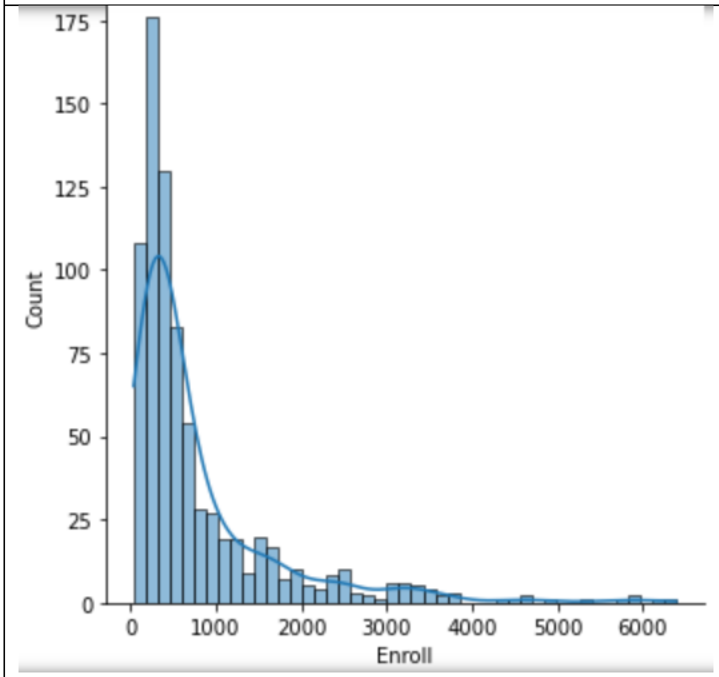
```
Description of Apps
---------------------------------------------------------------------------
count      777.000000
mean      3001.638353
std       3870.201484
min         81.000000
25%        776.000000
50%       1558.000000
75%       3624.000000
max      48094.000000
Name: Apps, dtype: float64 Distribution of Apps
---------------------------------------------------------------------------
```

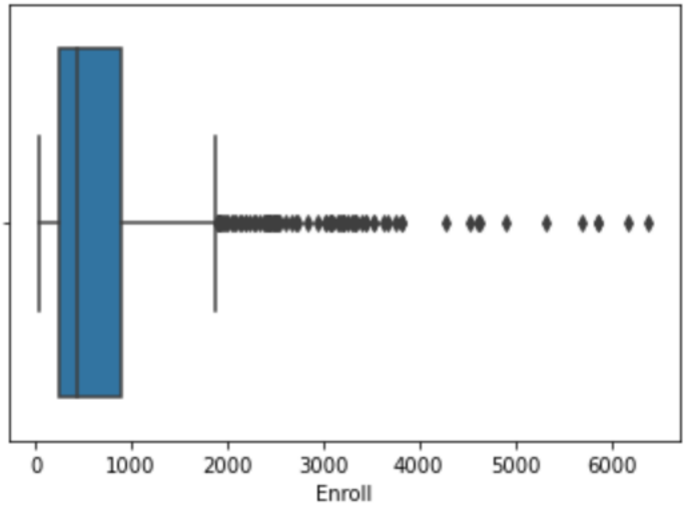


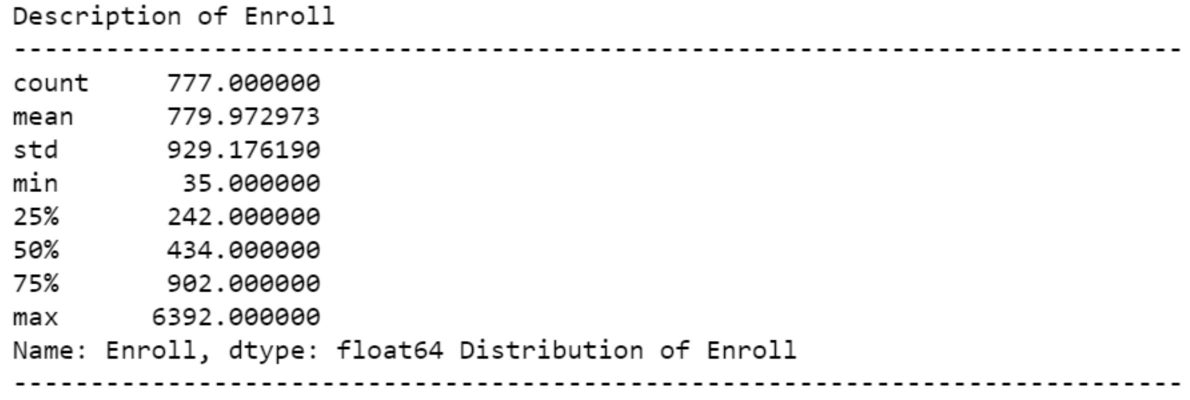




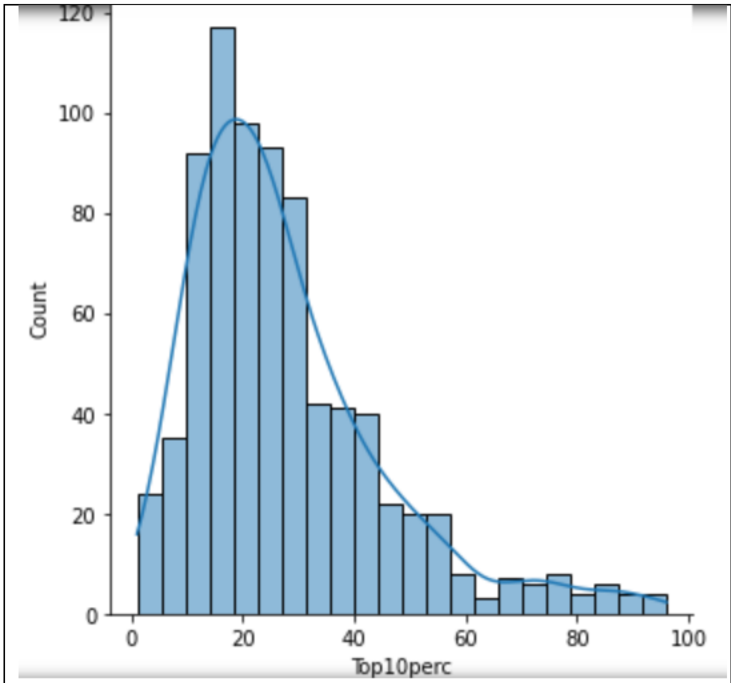
```
Description of Accept
----------------------------------------------------------------------------
count      777.000000
mean      2018.804376
std       2451.113971
min         72.000000
25%        604.000000
50%       1110.000000
75%       2424.000000
max      26330.000000
Name: Accept, dtype: float64 Distribution of Accept
----------------------------------------------------------------------------
```

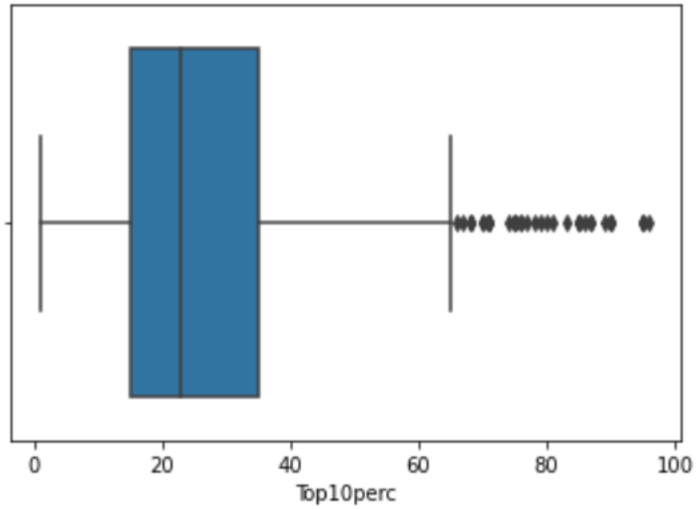


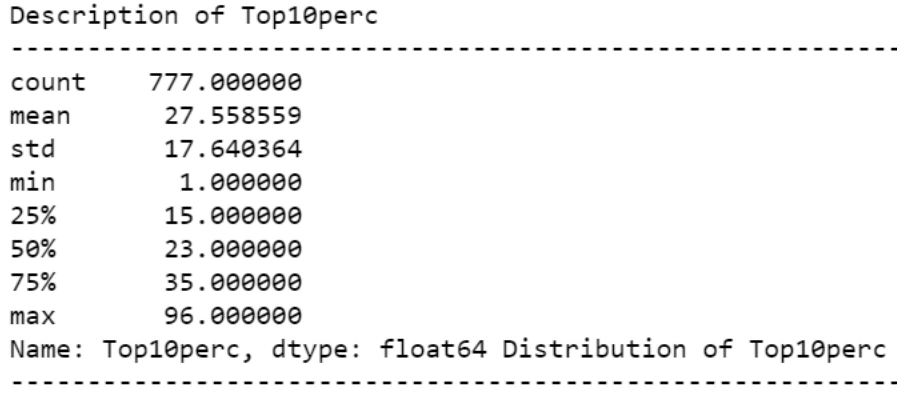




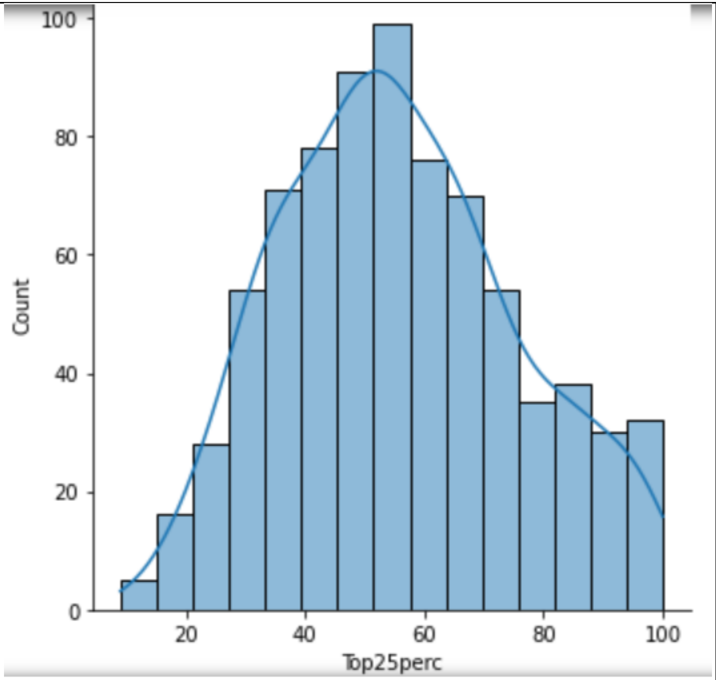
```
Description of Enroll
----------------------------------------------------------------------------
count      777.000000
mean       779.972973
std        929.176190
min         35.000000
25%        242.000000
50%        434.000000
75%        902.000000
max       6392.000000
Name: Enroll, dtype: float64 Distribution of Enroll
----------------------------------------------------------------------------
```

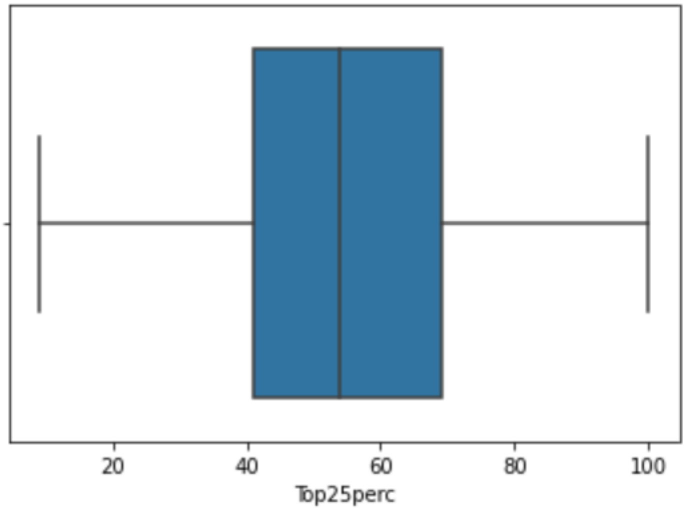


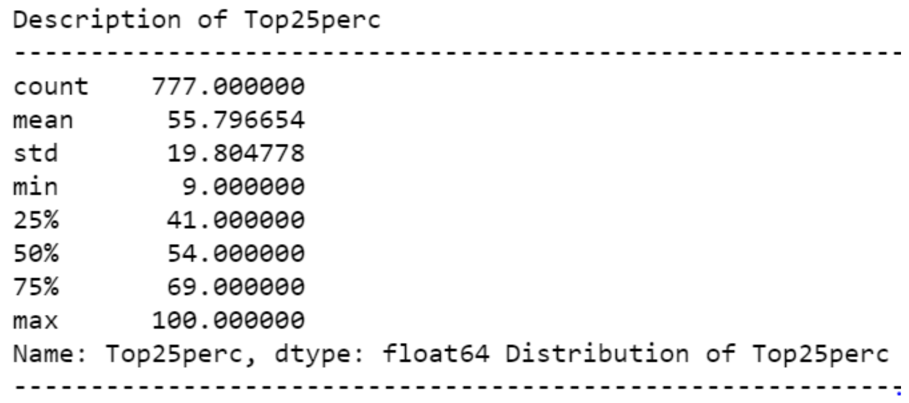




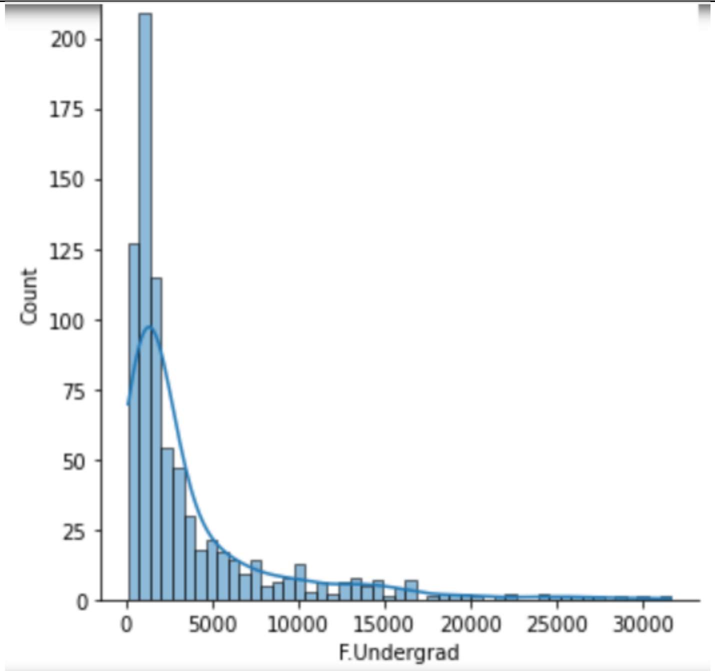
```
Description of Top10perc
-----------------------------------------------------------
count    777.000000
mean      27.558559
std       17.640364
min        1.000000
25%       15.000000
50%       23.000000
75%       35.000000
max       96.000000
Name: Top10perc, dtype: float64 Distribution of Top10perc
-----------------------------------------------------------
```

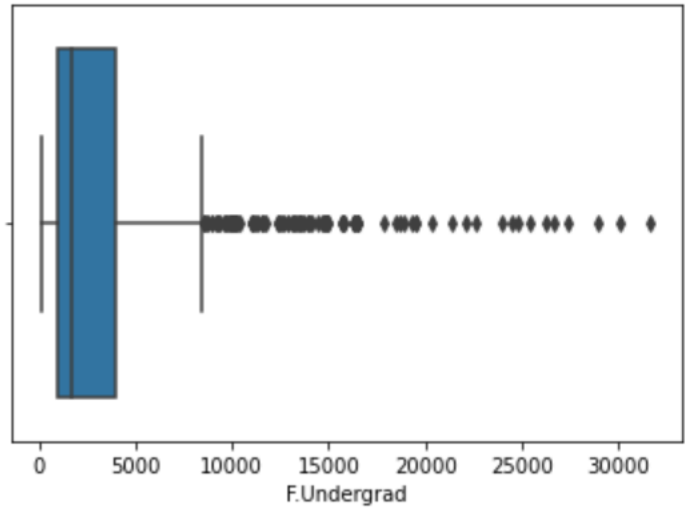


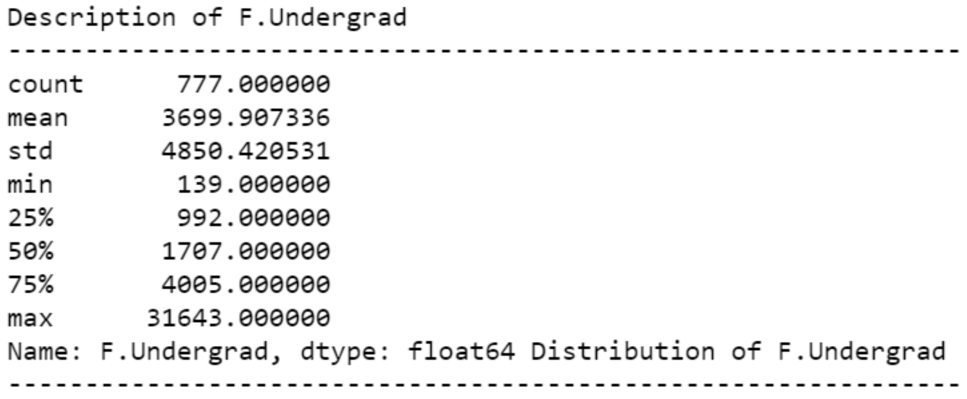




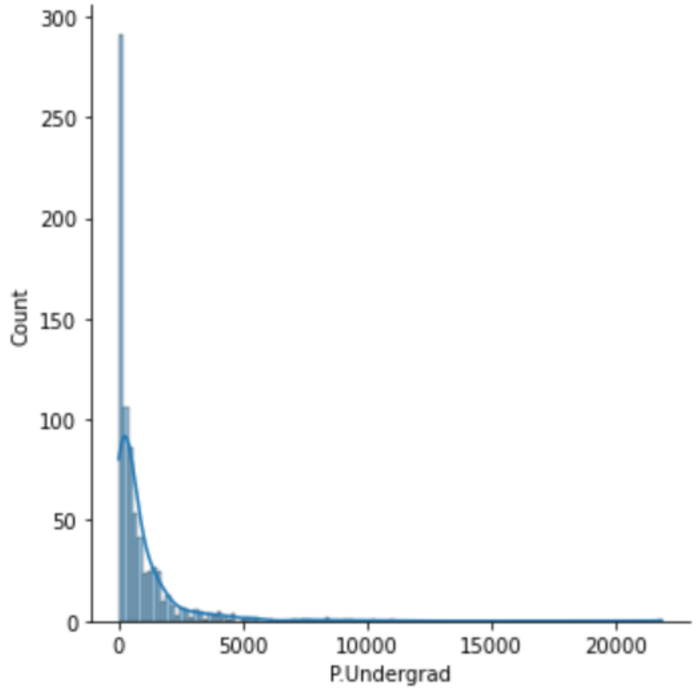
```
Description of Top25perc
----------------------------------------------------------
count    777.000000
mean      55.796654
std       19.804778
min        9.000000
25%       41.000000
50%       54.000000
75%       69.000000
max      100.000000
Name: Top25perc, dtype: float64 Distribution of Top25perc
----------------------------------------------------------
```

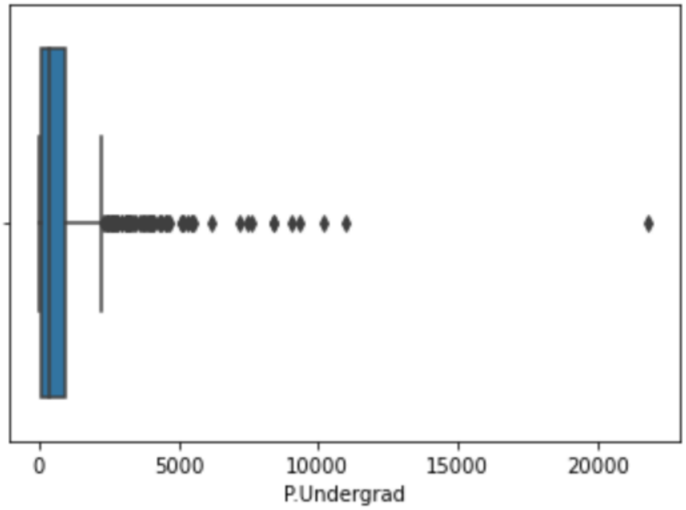


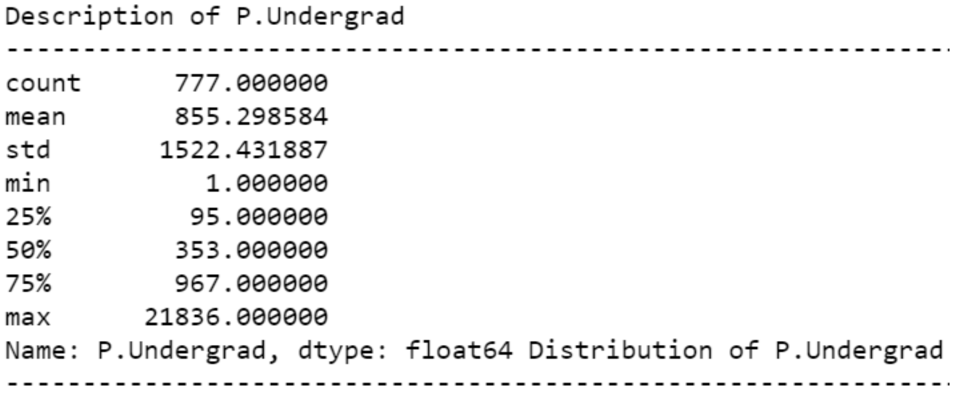




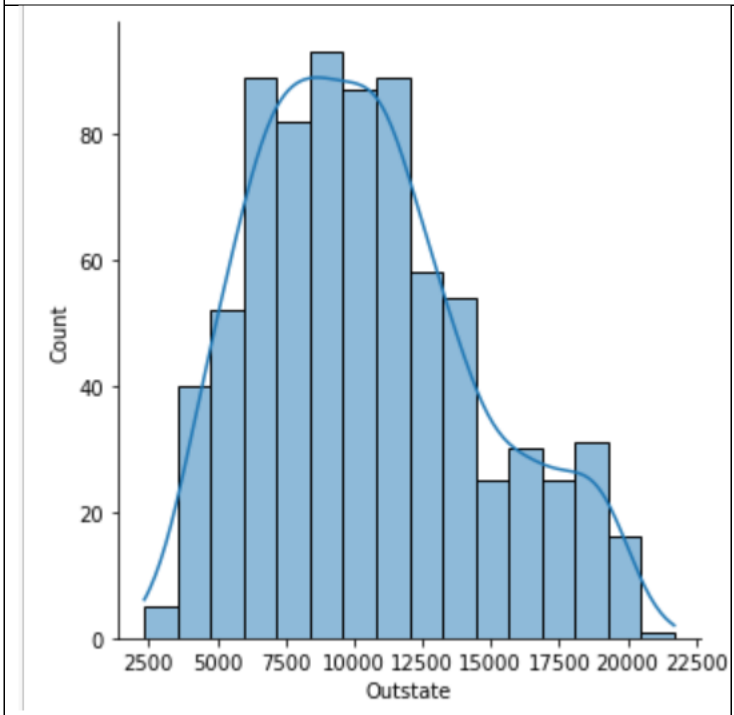
```
Description of F.Undergrad
--------------------------------------------------------------
count      777.000000
mean      3699.907336
std       4850.420531
min        139.000000
25%        992.000000
50%       1707.000000
75%       4005.000000
max      31643.000000
Name: F.Undergrad, dtype: float64 Distribution of F.Undergrad
--------------------------------------------------------------
```

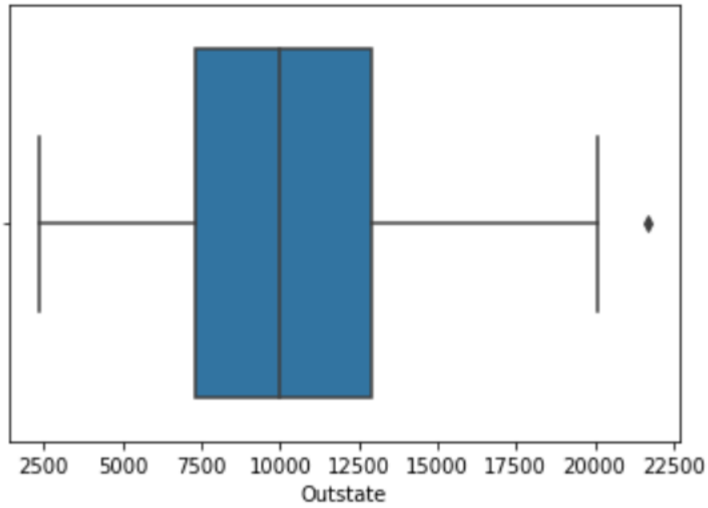


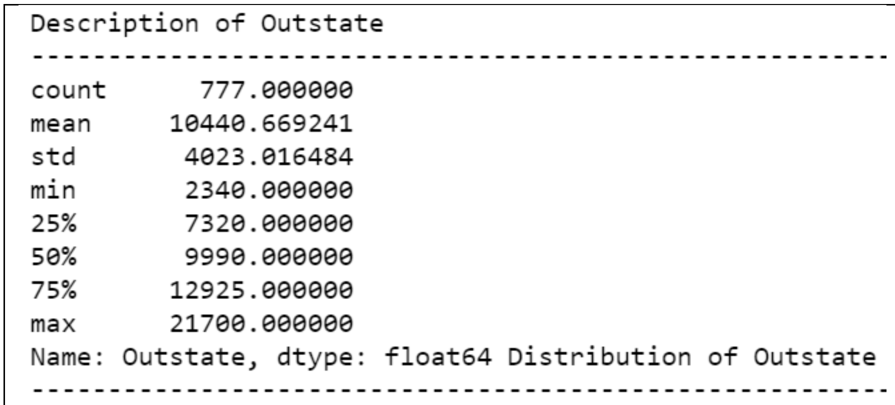




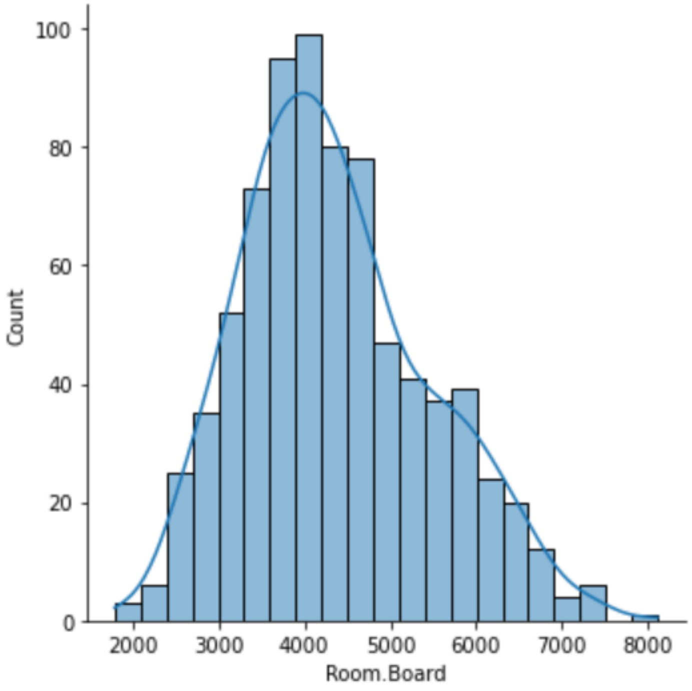
```
Description of P.Undergrad
--------------------------------------------------------------
count       777.000000
mean        855.298584
std        1522.431887
min           1.000000
25%          95.000000
50%         353.000000
75%         967.000000
max       21836.000000
Name: P.Undergrad, dtype: float64 Distribution of P.Undergrad
--------------------------------------------------------------
```

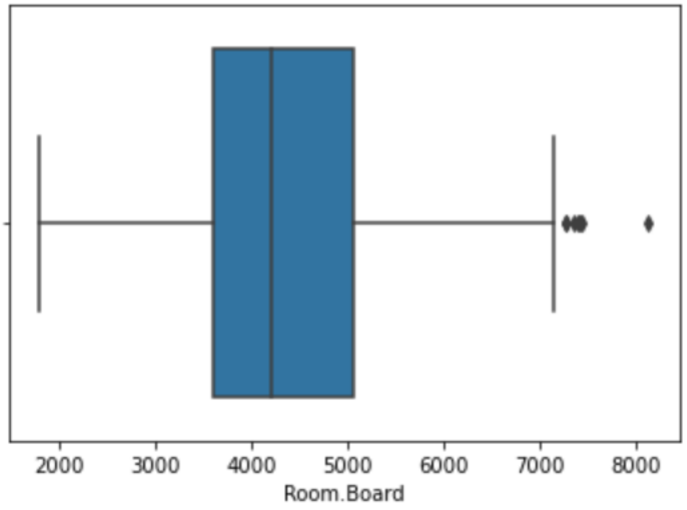


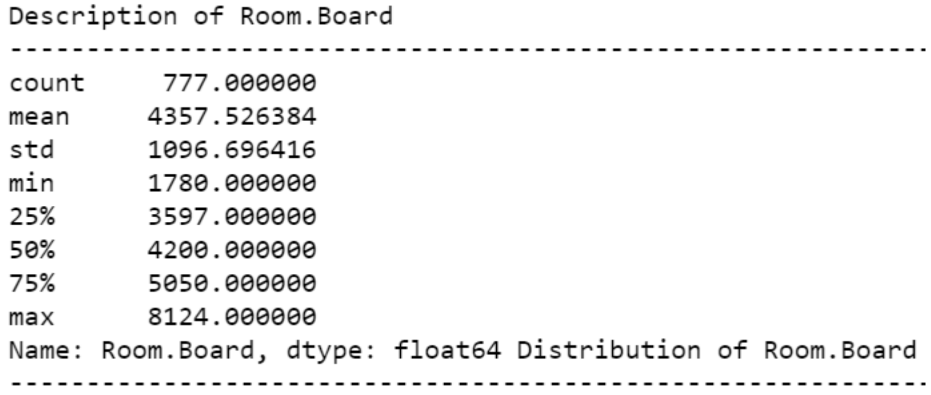




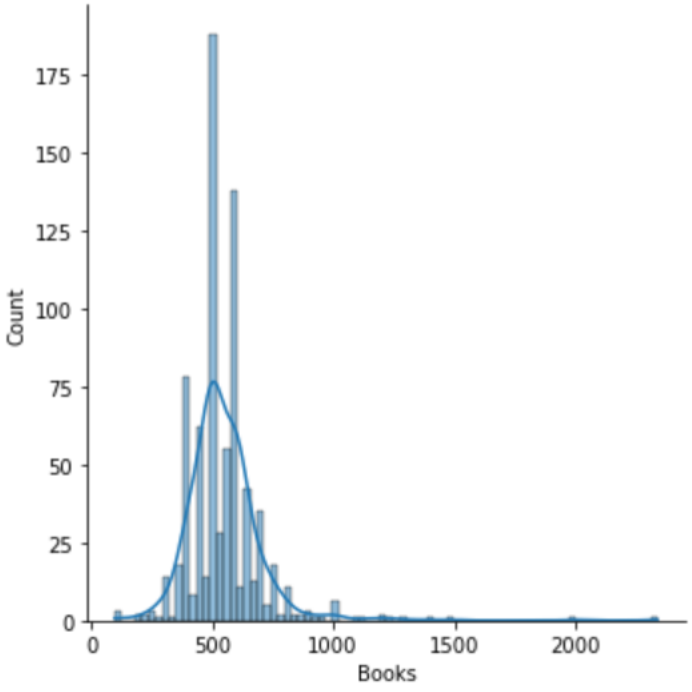
```
Description of Outstate
--------------------------------------------------------
count      777.000000
mean     10440.669241
std       4023.016484
min       2340.000000
25%       7320.000000
50%       9990.000000
75%      12925.000000
max      21700.000000
Name: Outstate, dtype: float64 Distribution of Outstate
--------------------------------------------------------
```

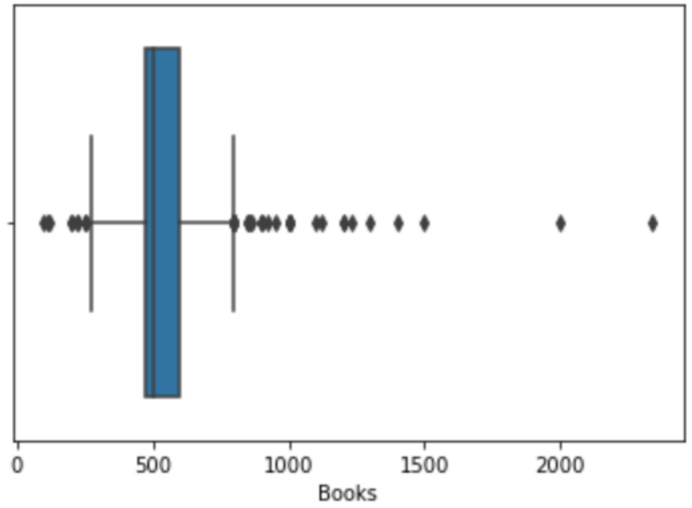




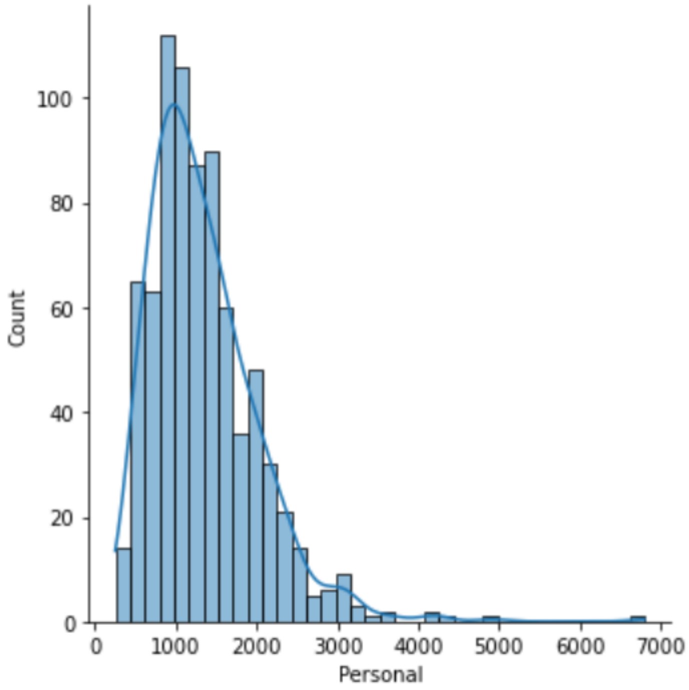


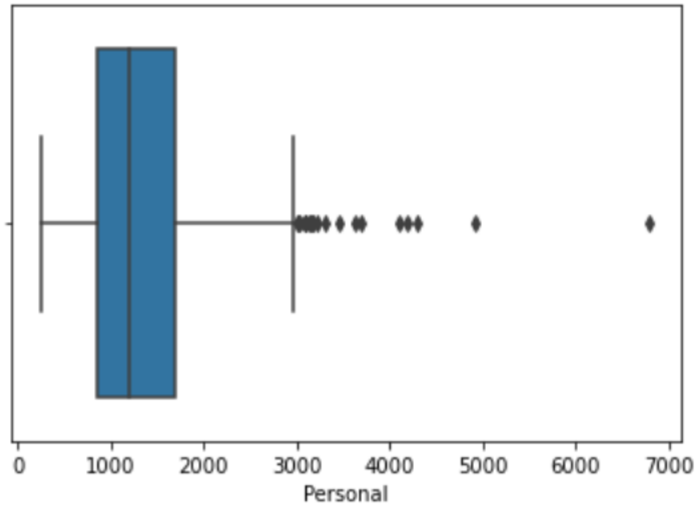
```
Description of Room.Board
------------------------------------------------------------
count      777.000000
mean      4357.526384
std       1096.696416
min       1780.000000
25%       3597.000000
50%       4200.000000
75%       5050.000000
max       8124.000000
Name: Room.Board, dtype: float64 Distribution of Room.Board
------------------------------------------------------------
```

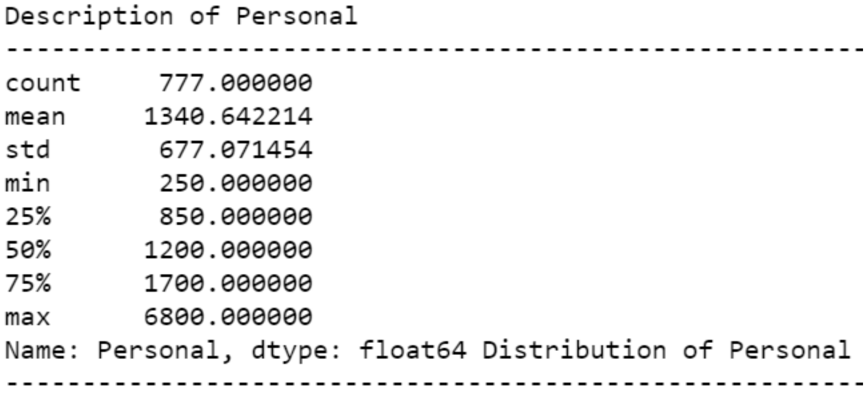




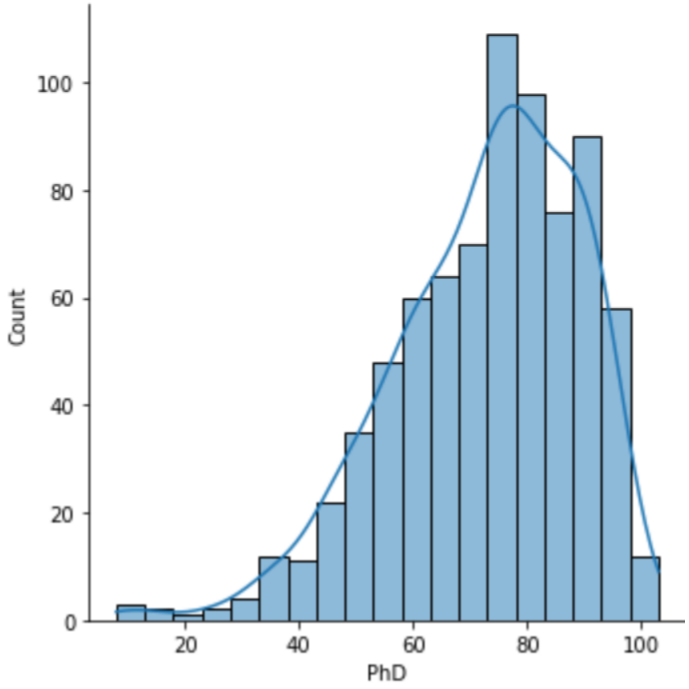
```
Description of Books
--------------------------------------------------
count      777.000000
mean       549.380952
std        165.105360
min         96.000000
25%        470.000000
50%        500.000000
75%        600.000000
max       2340.000000
Name: Books, dtype: float64 Distribution of Books
--------------------------------------------------
```

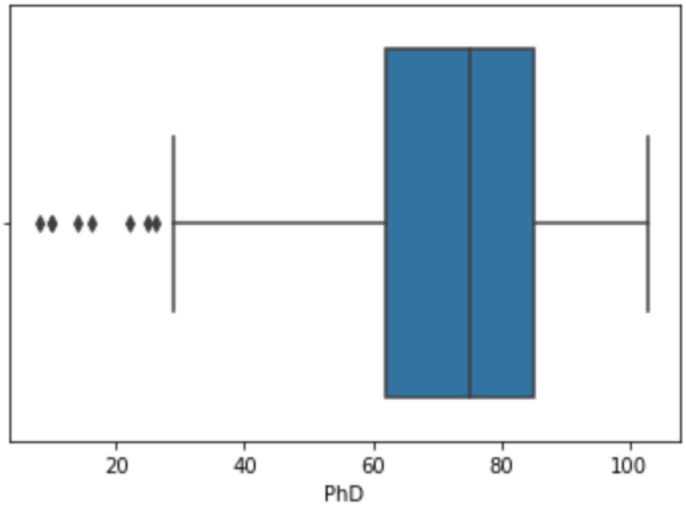


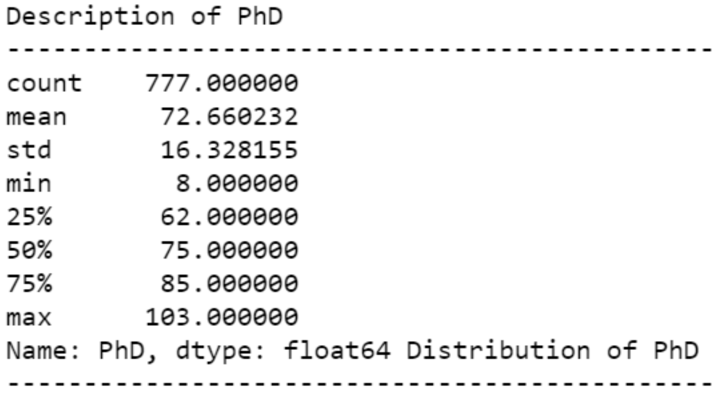




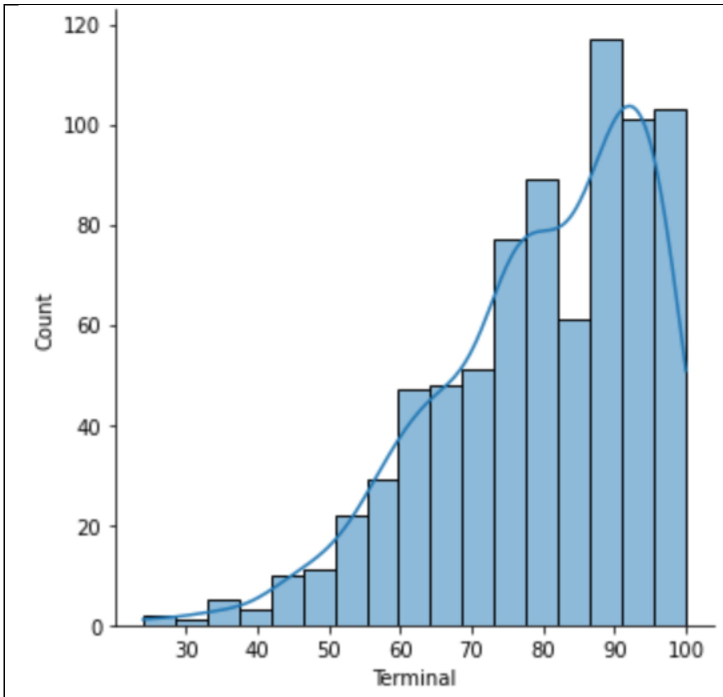
```
Description of Personal
--------------------------------------------------------
count      777.000000
mean      1340.642214
std        677.071454
min        250.000000
25%        850.000000
50%       1200.000000
75%       1700.000000
max       6800.000000
Name: Personal, dtype: float64 Distribution of Personal
--------------------------------------------------------
```

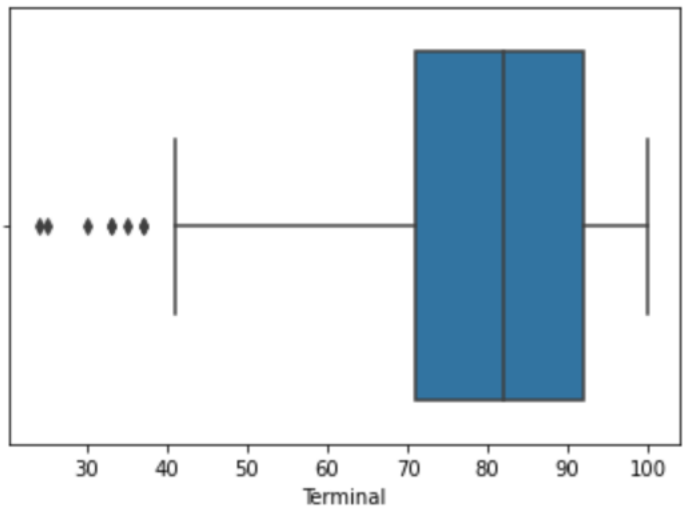


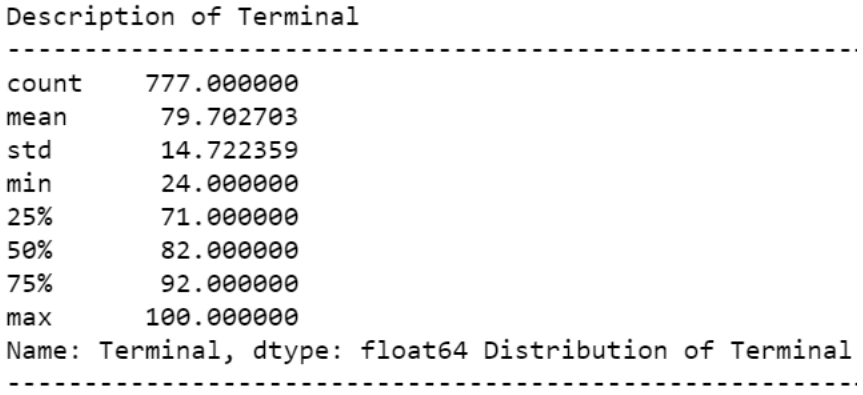




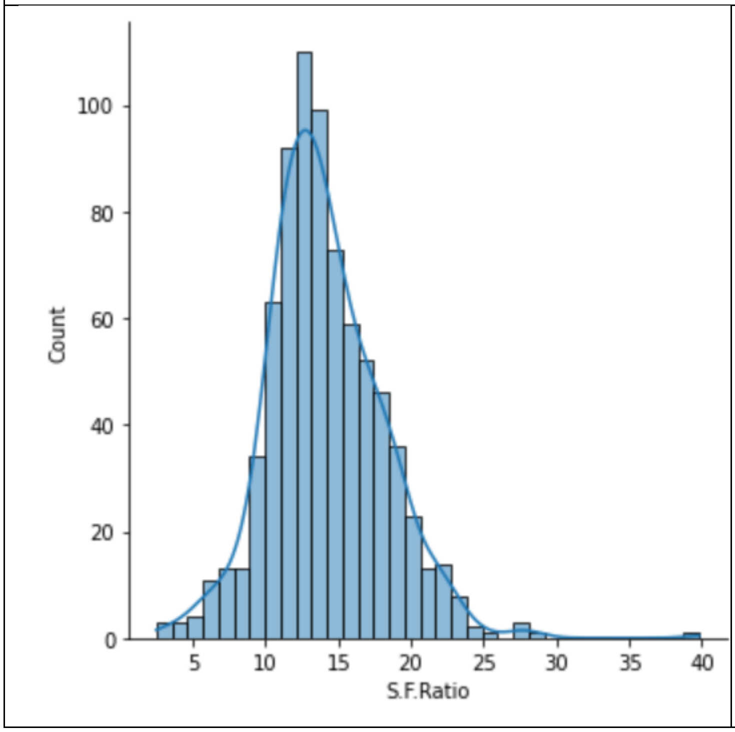
```
Description of PhD
----------------------------------------------
count     777.000000
mean       72.660232
std        16.328155
min         8.000000
25%        62.000000
50%        75.000000
75%        85.000000
max       103.000000
Name: PhD, dtype: float64 Distribution of PhD
----------------------------------------------
```

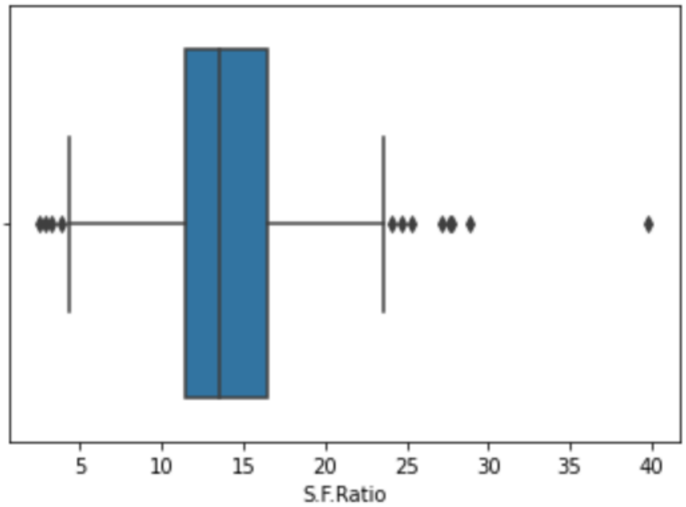


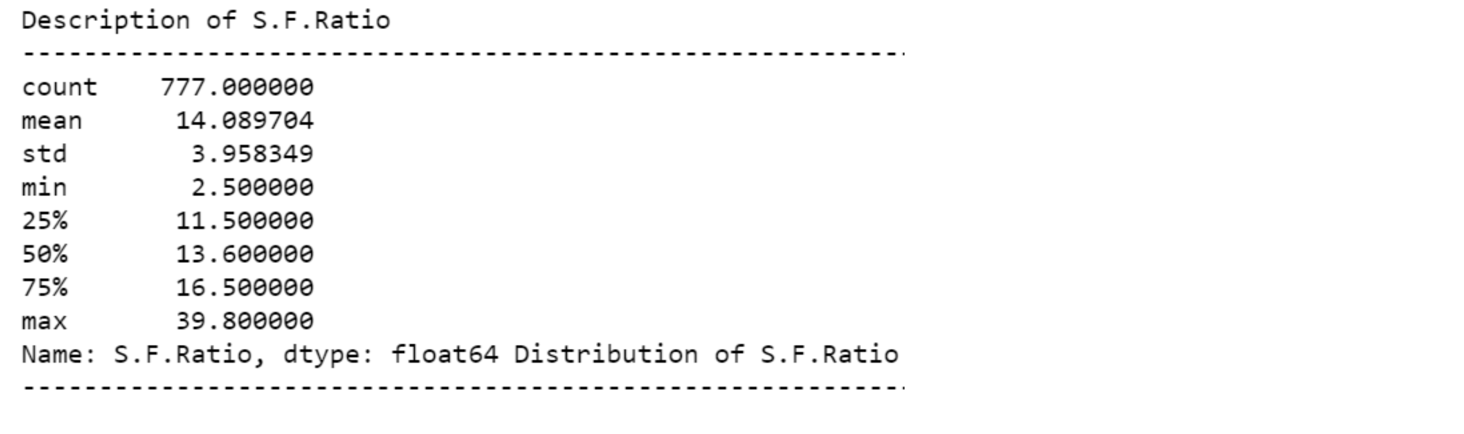




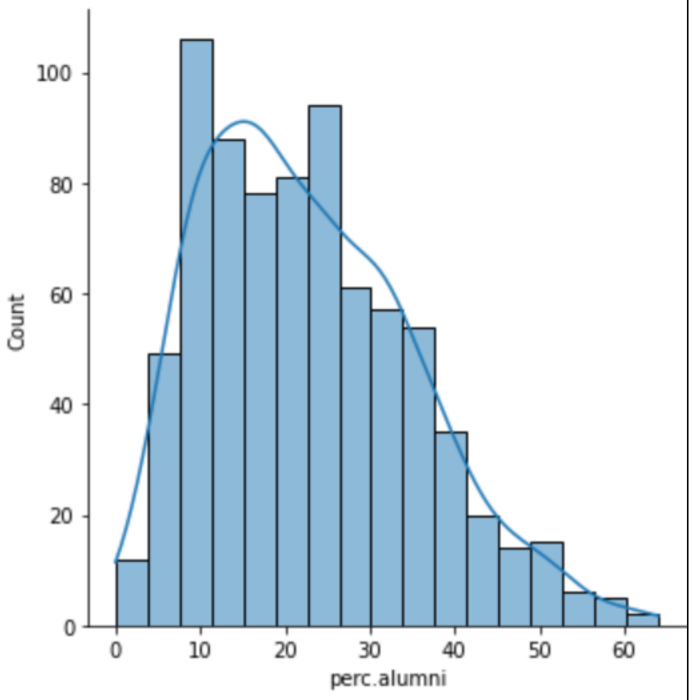
```
Description of Terminal
-------------------------------------------------------
count    777.000000
mean      79.702703
std       14.722359
min       24.000000
25%       71.000000
50%       82.000000
75%       92.000000
max      100.000000
Name: Terminal, dtype: float64 Distribution of Terminal
-------------------------------------------------------
```

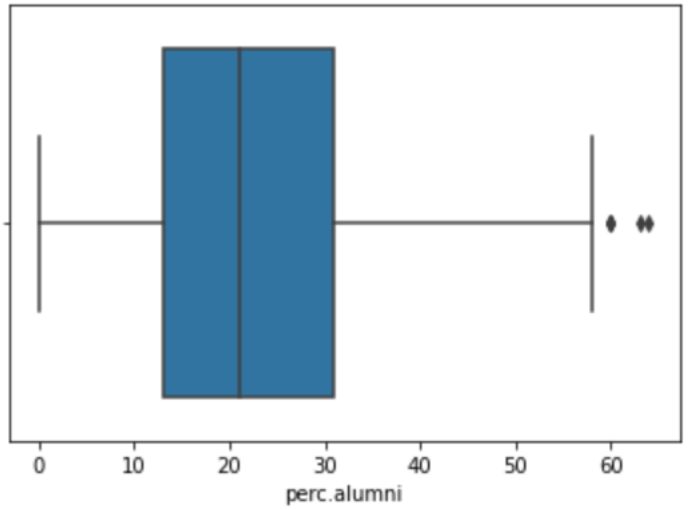


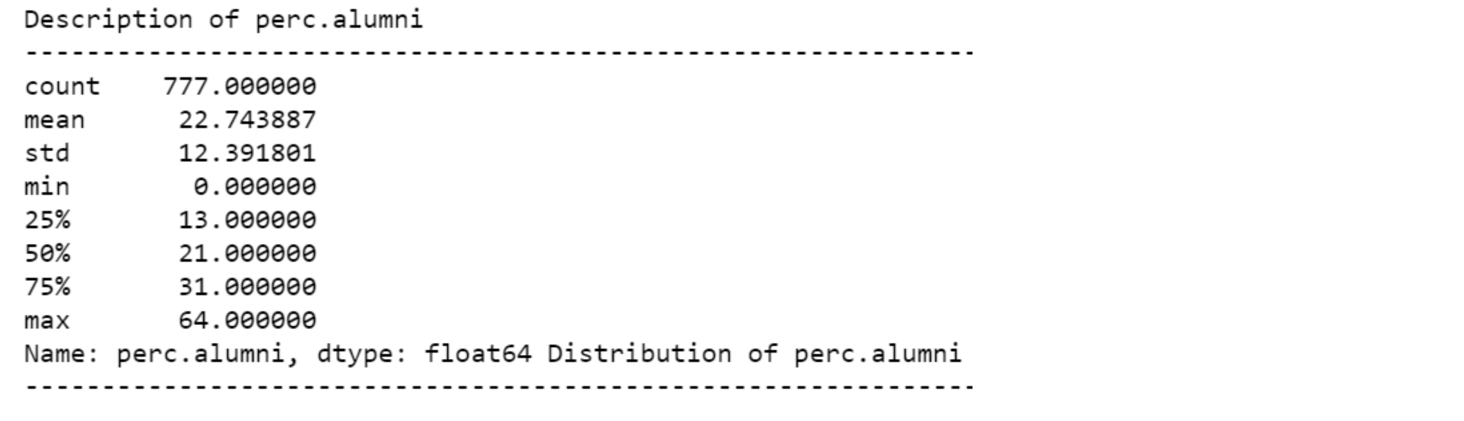




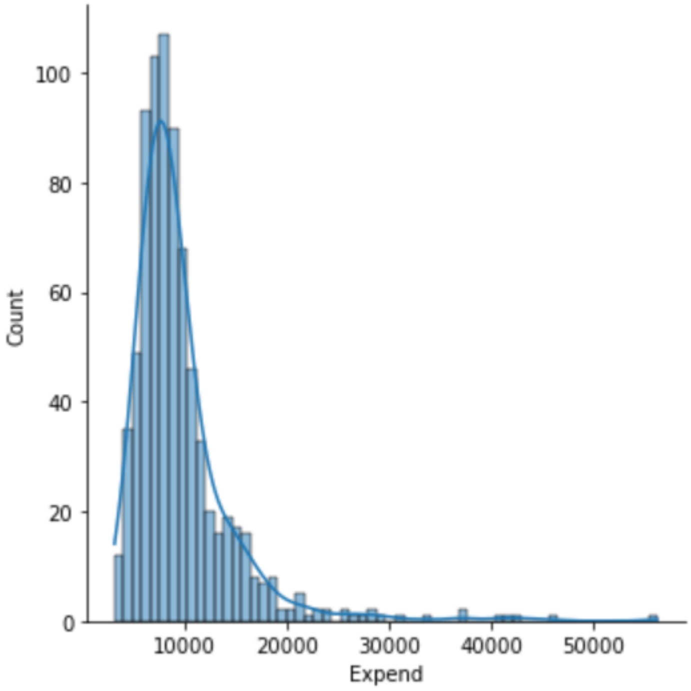
```
Description of S.F.Ratio
-----------------------------------------------------------
count    777.000000
mean      14.089704
std        3.958349
min        2.500000
25%       11.500000
50%       13.600000
75%       16.500000
max       39.800000
Name: S.F.Ratio, dtype: float64 Distribution of S.F.Ratio
-----------------------------------------------------------
```

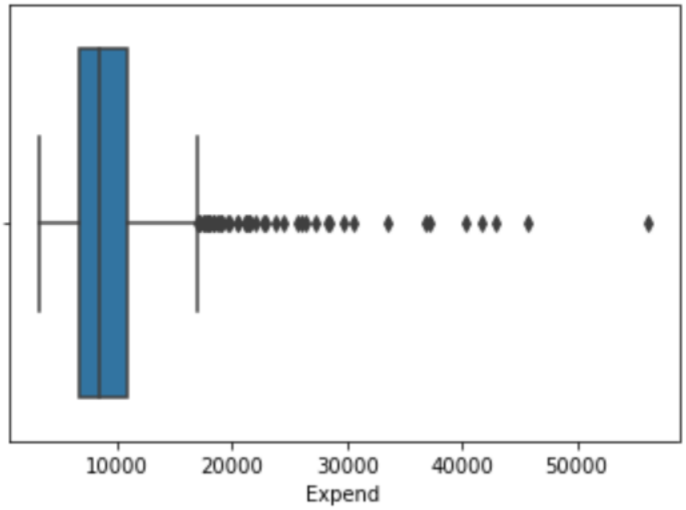


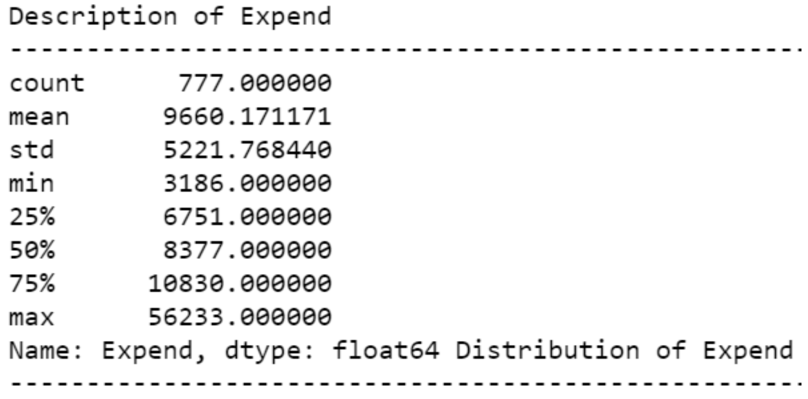




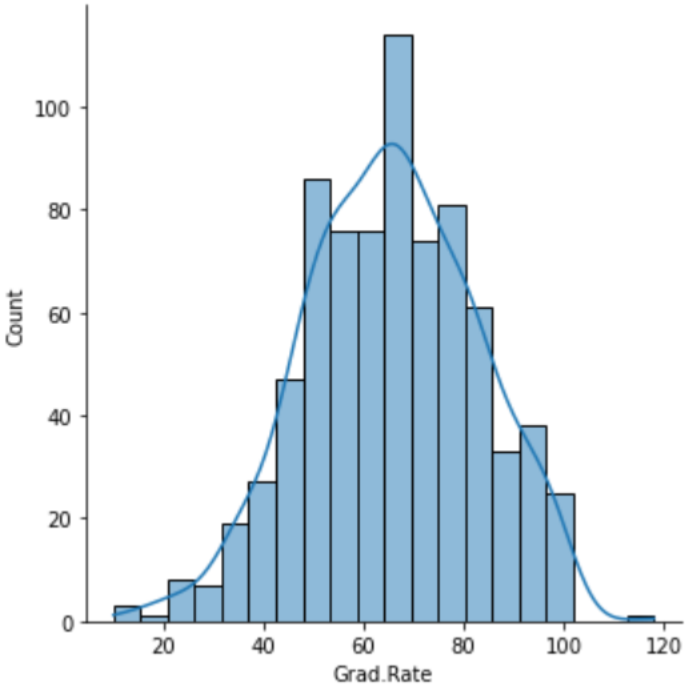
```
Description of perc.alumni
--------------------------------------------------------------
count    777.000000
mean      22.743887
std       12.391801
min        0.000000
25%       13.000000
50%       21.000000
75%       31.000000
max       64.000000
Name: perc.alumni, dtype: float64 Distribution of perc.alumni
--------------------------------------------------------------
```

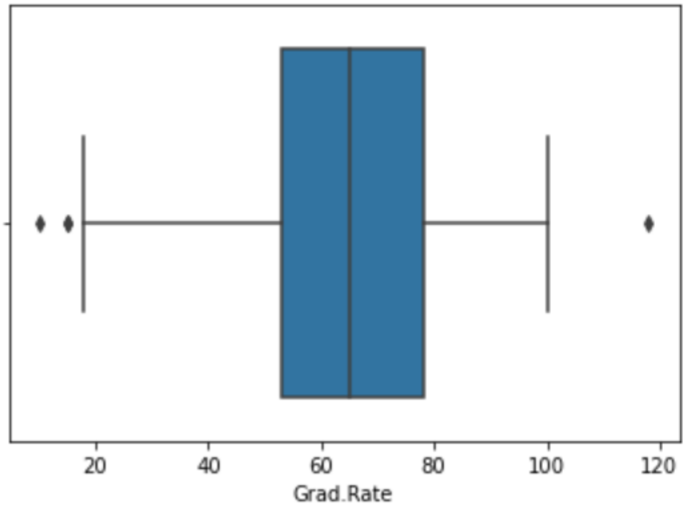


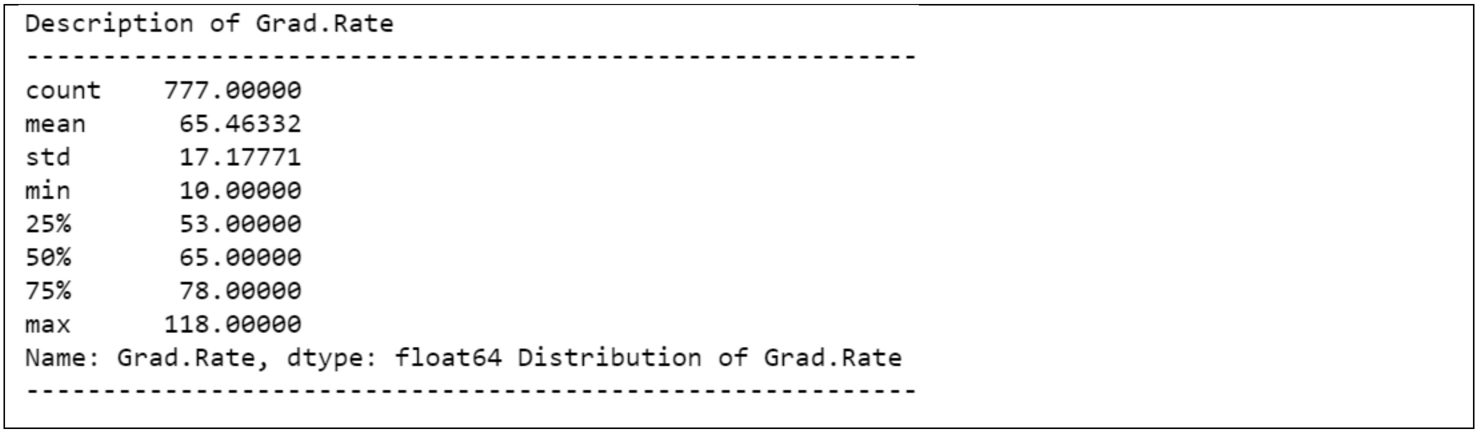




```
Description of Expend
----------------------------------------------------
count      777.000000
mean      9660.171171
std       5221.768440
min       3186.000000
25%       6751.000000
50%       8377.000000
75%      10830.000000
max      56233.000000
Name: Expend, dtype: float64 Distribution of Expend
----------------------------------------------------
```







```
Description of Grad.Rate
-----------------------------------------------------------
count    777.00000
mean      65.46332
std       17.17771
min       10.00000
25%       53.00000
50%       65.00000
75%       78.00000
max      118.00000
Name: Grad.Rate, dtype: float64 Distribution of Grad.Rate
-----------------------------------------------------------
```

After performing univariate analysis, we can see that the data in the Application column is Right skewed with average of 3001.638353 applications per college. Out of the 777 colleges, there is 1 college with 48094 applications.

The data in the Accepted column is also right skewed with average of 2018.804376 applications being accepted, however only average of 779.972973 enrolled. On an average only 27.558559 are from top 10% of Higher Secondary class and 55.796654 are from top 25% of Higher Secondary School. There are 3699.907336 full time under graduate students and 855.298584 part time students on an average.

We can also see that most students prefer to study out of state with an average of 10440.669241 students enrolling in colleges out of their state. Across colleges average cost of Room Board is 4357.526384.Cost of books is estimated to be 549.380952 on an average. Average personal spending amounts to 1340.642214. Looking at the Faculty, 72.660232 percent are PHDs while 79.702703 percent have terminal degree on an average.

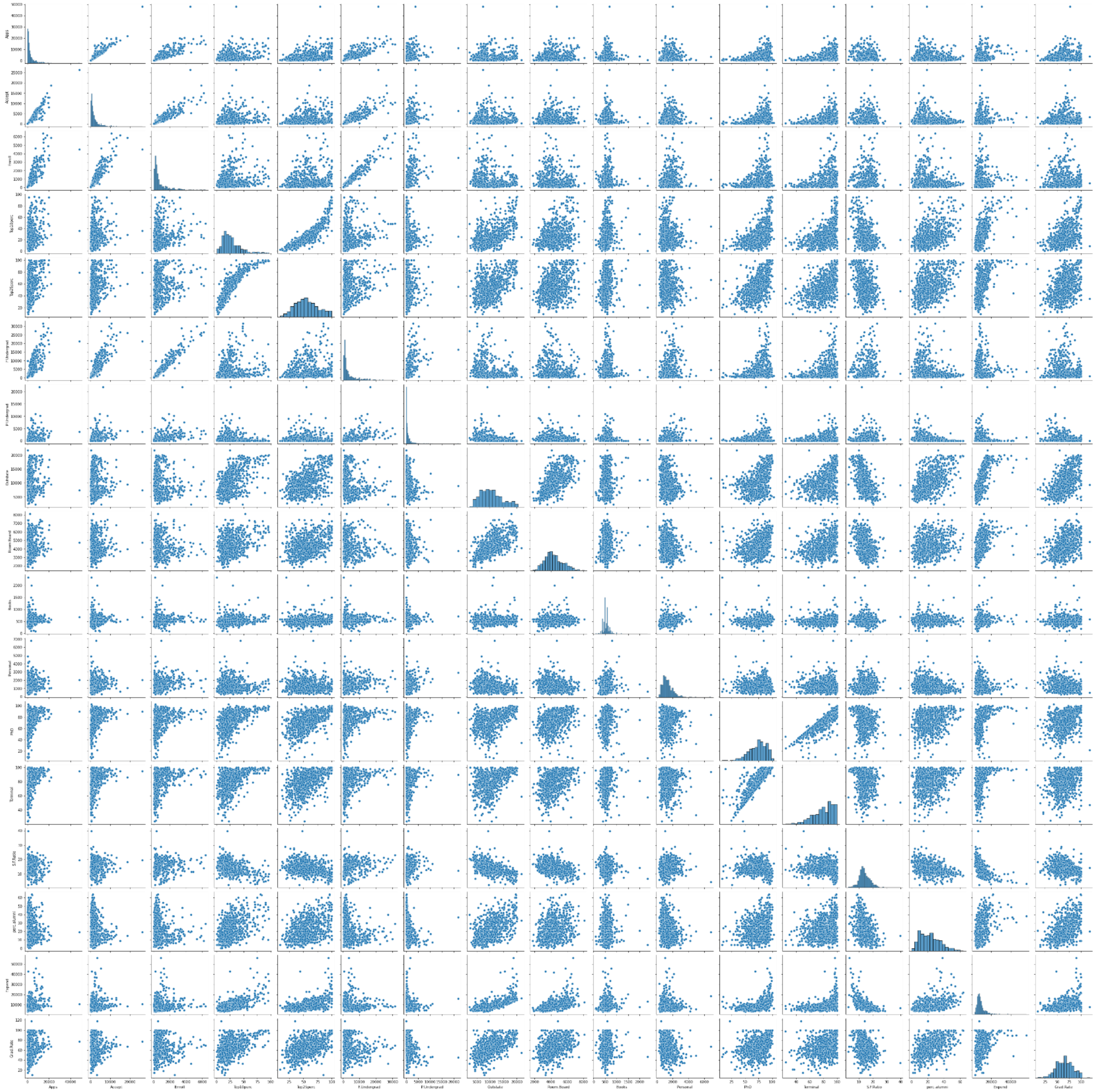
On an average, 14.089704 is the student to Faculty ratio and 22.743887 percent of alumni who donate. 9660.171171 is the average instruction expenditure per student. Only 65.46332 percent actually graduate.

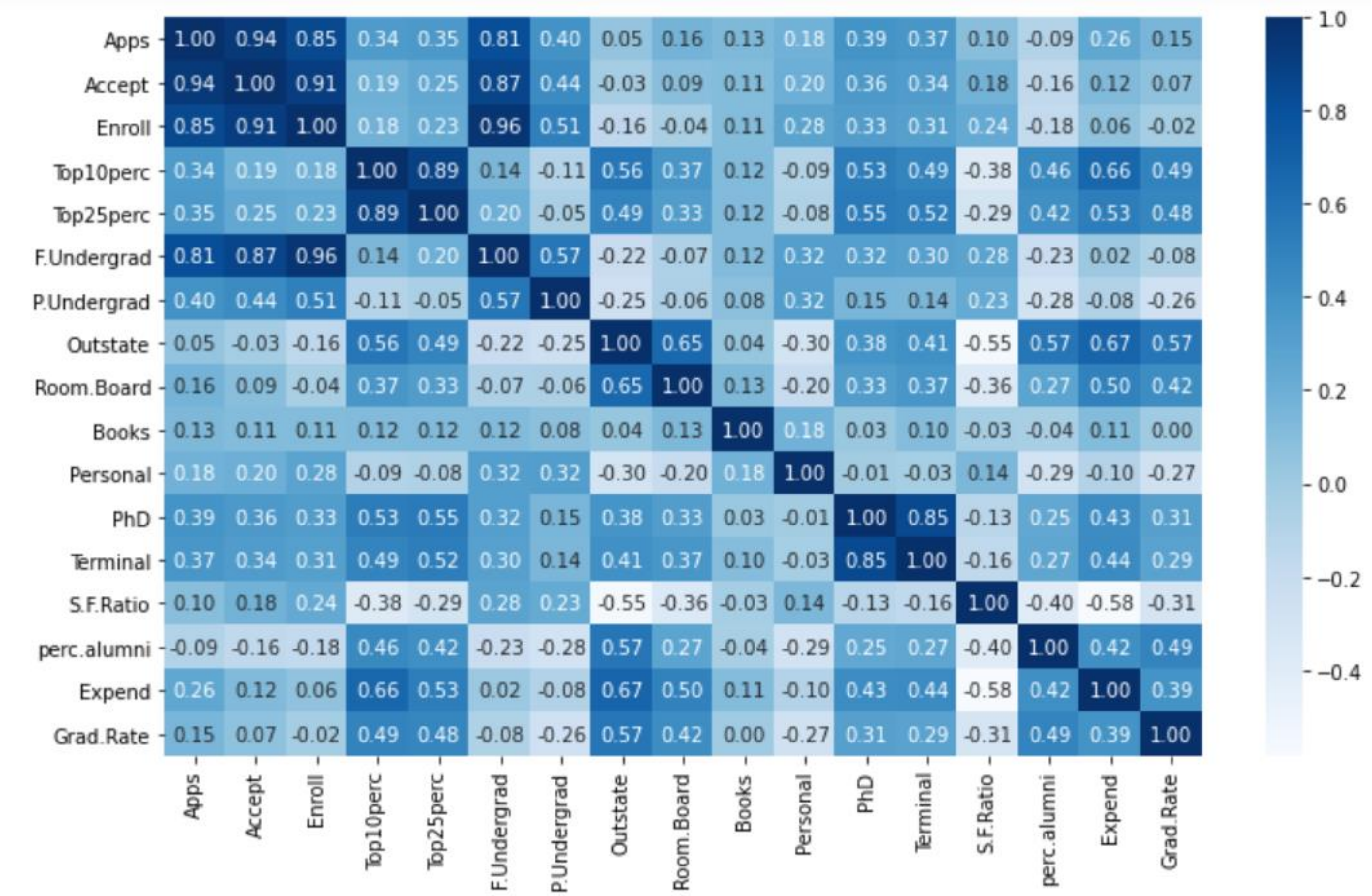
Bivariate Analysis

Bi means two and variate means variable, so here there are two variables. The analysis is related to cause and the relationship between the two variables. The following pairplot and the heatmap displays the correlation among the variables.

Correlation matrices are an essential tool of exploratory data analysis. Correlation heatmaps contain the same information in a visually appealing way. they show in a glance which variables are correlated, to what degree, in which direction, and alerts us to potential multicollinearity problems.

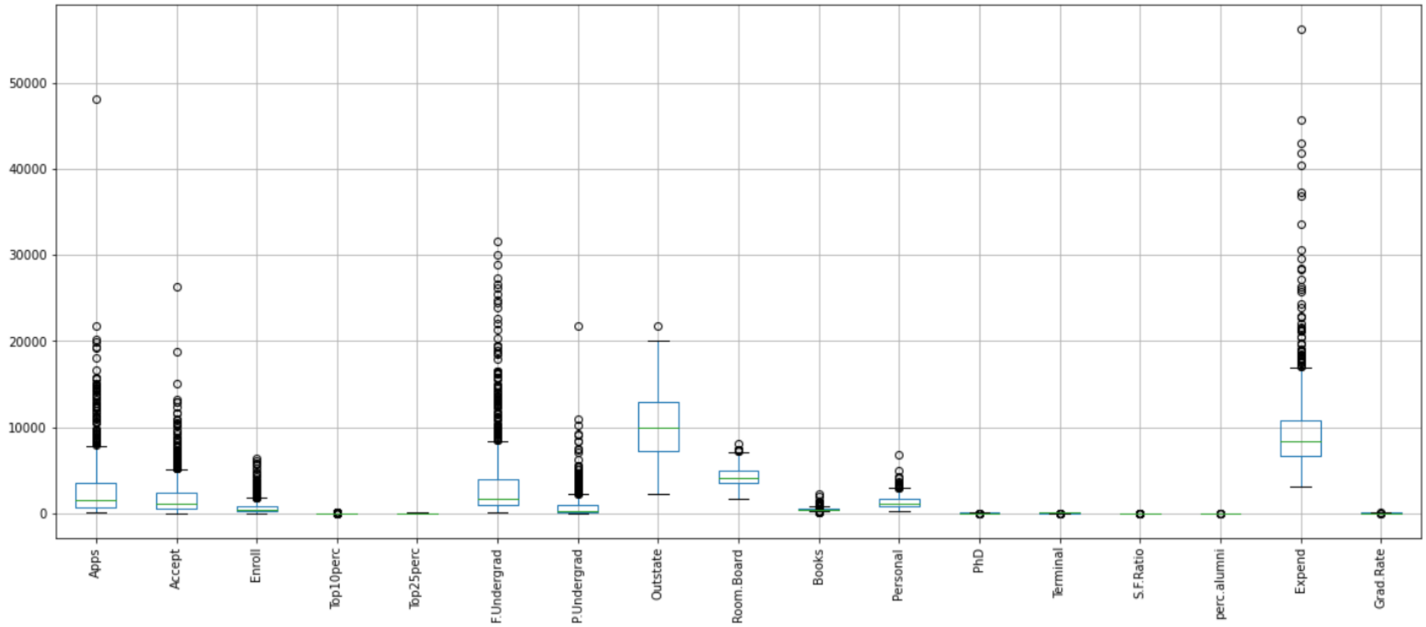
Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables. The close to 1 the correlation is the more positively correlated they are; that is as one increases so does the other and the closer to 1 the stronger this relationship is.


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	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.00	0.94	0.85	0.34	0.35	0.81	0.40	0.05	0.16	0.13	0.18	0.39	0.37	0.10	-0.09	0.26	0.15
Accept	0.94	1.00	0.91	0.19	0.25	0.87	0.44	-0.03	0.09	0.11	0.20	0.36	0.34	0.18	-0.16	0.12	0.07
Enroll	0.85	0.91	1.00	0.18	0.23	0.96	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	-0.18	0.06	-0.02
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.37	0.12	-0.09	0.53	0.49	-0.38	0.46	0.66	0.49
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33	0.12	-0.08	0.55	0.52	-0.29	0.42	0.53	0.48
F.Undergrad	0.81	0.87	0.96	0.14	0.20	1.00	0.57	-0.22	-0.07	0.12	0.32	0.32	0.30	0.28	-0.23	0.02	-0.08
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	-0.06	0.08	0.32	0.15	0.14	0.23	-0.28	-0.08	-0.26
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.65	0.04	-0.30	0.38	0.41	-0.55	0.57	0.67	0.57
Room.Board	0.16	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.65	1.00	0.13	-0.20	0.33	0.37	-0.36	0.27	0.50	0.42
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.08	0.04	0.13	1.00	0.18	0.03	0.10	-0.03	-0.04	0.11	0.00
Personal	0.18	0.20	0.28	-0.09	-0.08	0.32	0.32	-0.30	-0.20	0.18	1.00	-0.01	-0.03	0.14	-0.29	-0.10	-0.27
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.03	-0.01	1.00	0.85	-0.13	0.25	0.43	0.31
Terminal	0.37	0.34	0.31	0.49	0.52	0.30	0.14	0.41	0.37	0.10	-0.03	0.85	1.00	-0.16	0.27	0.44	0.29
S.F.Ratio	0.10	0.18	0.24	-0.38	-0.29	0.28	0.23	-0.55	-0.36	-0.03	0.14	-0.13	-0.16	1.00	-0.40	-0.58	-0.31
perc.alumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.27	-0.04	-0.29	0.25	0.27	-0.40	1.00	0.42	0.49
Expend	0.26	0.12	0.06	0.66	0.53	0.02	-0.08	0.67	0.50	0.11	-0.10	0.43	0.44	-0.58	0.42	1.00	0.39
Grad.Rate	0.15	0.07	-0.02	0.49	0.48	-0.08	-0.26	0.57	0.42	0.00	-0.27	0.31	0.29	-0.31	0.49	0.39	1.00

Except the column showing the percentage of students from top 25 percent of higher secondary school, all the other columns have outliers.



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

Scaling (normalization) is a technique often applied as part of data preparation. 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.

When the feature data value is large, scaling can be used to obtain better Convergence in lesser number of iterations and speed up the overall iterative process.

Types of Scaling –
The two most commonly used scaling techniques are Z-score and Min-Max Scaling.

1. Z-Score Scaling – A method in which all the values are converted to z-scores. Z-score removes the mean i.e. brings the mean to zero and scales the data to unit variance.

Z-scores are calculated using the following formula: z = (x-mean(x))/stdev(x)

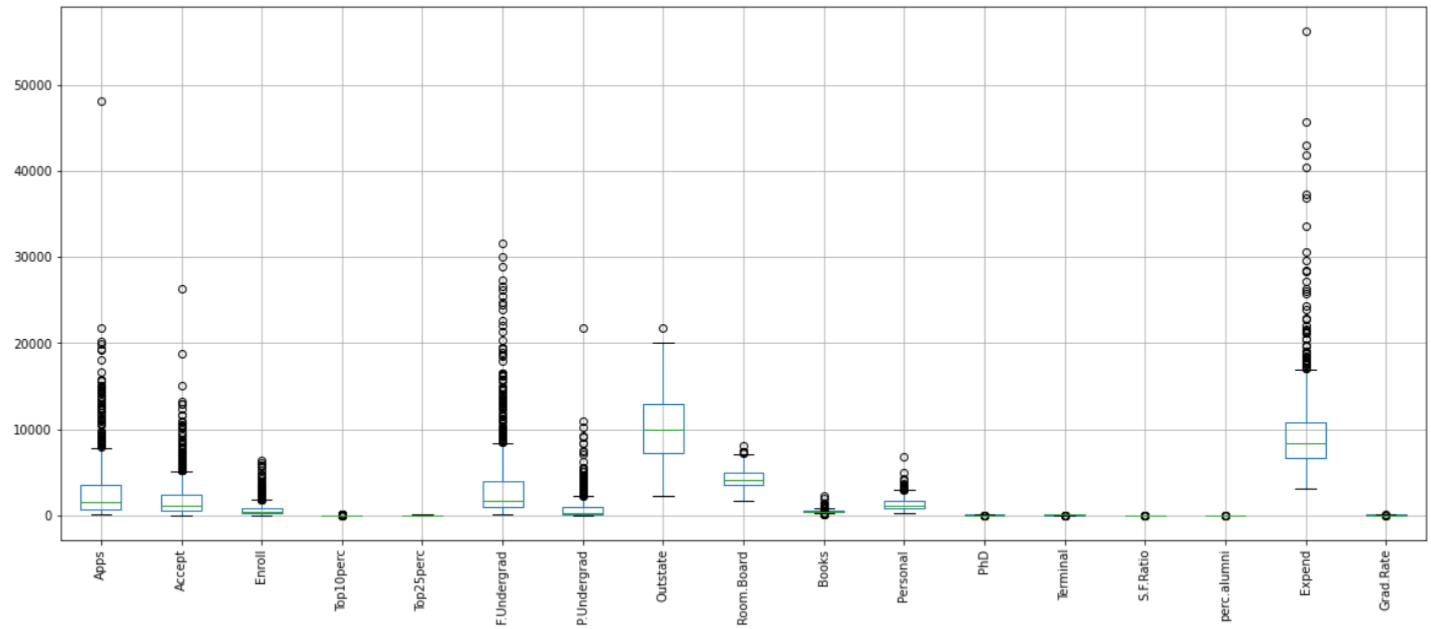
One thing to note at this point is, the expression of z-score makes use of mean and std. deviation of the distribution, which are both affected by the presence of outlier values. This leads to an imbalanced feature scales in presence of significant outlier values. Thus being aware of the outliers and performing appropriate treatment in accordance with the business becomes very crucial.

2. Min-Max Scaling: The MinMax method linearly rescales every feature to the [0,1] interval. The presence of this bounded range - in contrast to z-score scaling - is that we will end up with smaller standard deviations, which can suppress the effect of outliers.

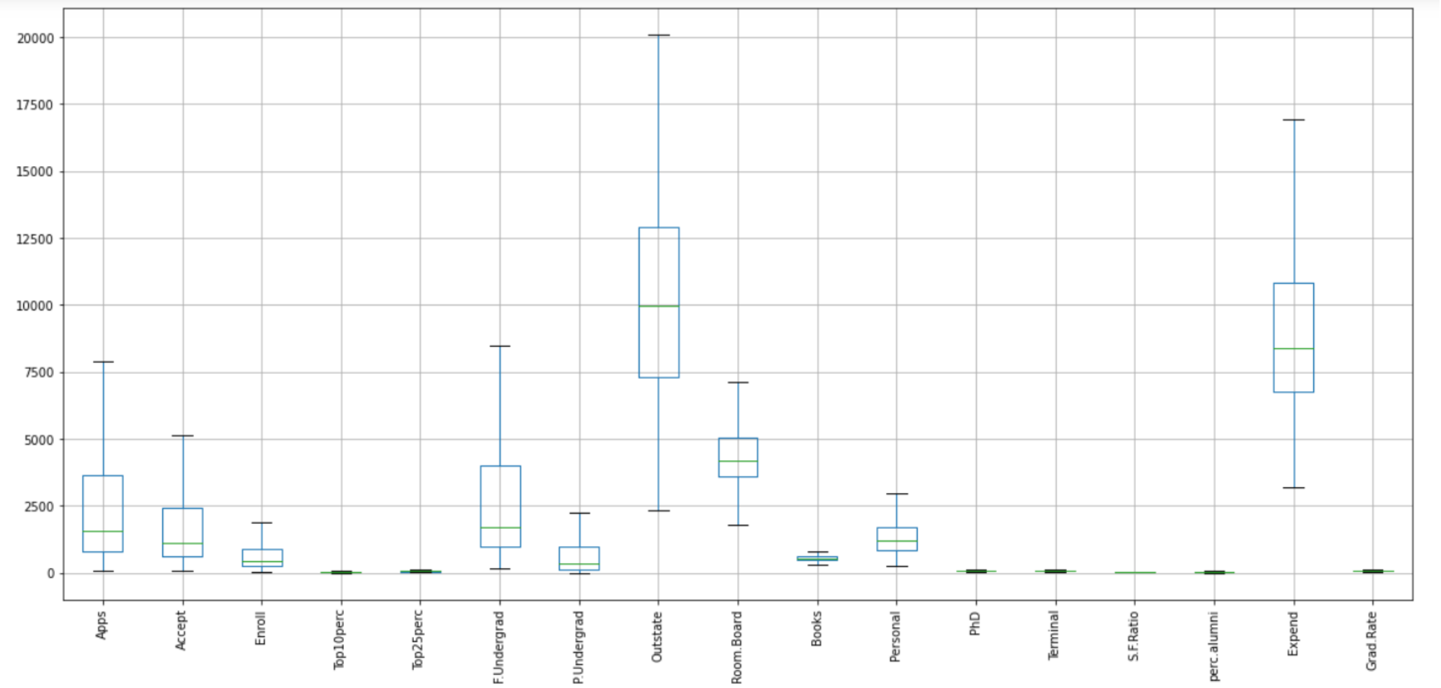
When it comes to choosing the appropriate scaling method for the dataset, there are no hard coded rules instead; it depends on the case study in hand.

1. Min-Max is good to use when you know that the distribution of your data does not follow a Gaussian distribution. In addition, this can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.
2. Z-Score on the other hand, can be helpful in cases where the data follows a Normal (or near normal) distribution or the algorithm demands a Normal distribution of the features.

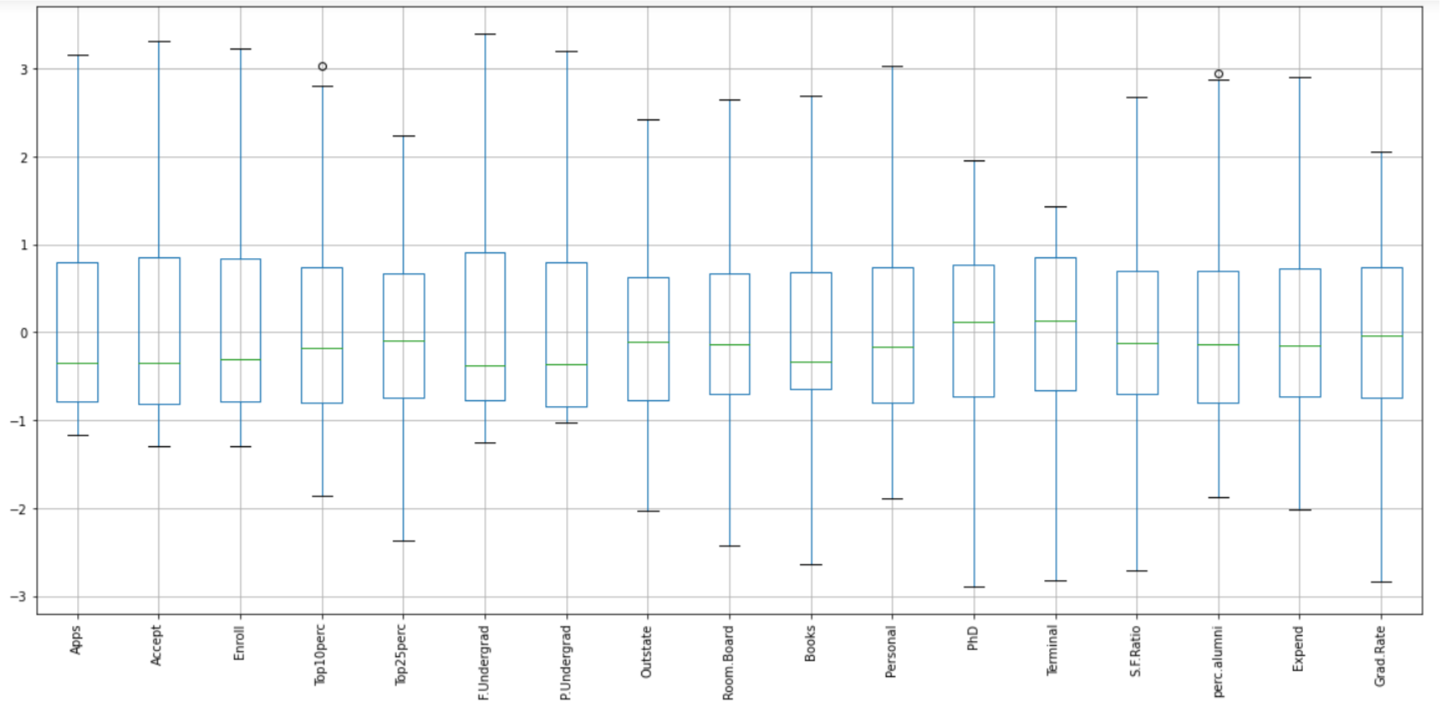
The below boxplot shows the outliers present in the data:



First we are going to treat the outlier, so what we have done is we have replaced the outliers with upper bound and lower bound values to minimize the effect of outliers. After treating the outliers we get the following boxplot:



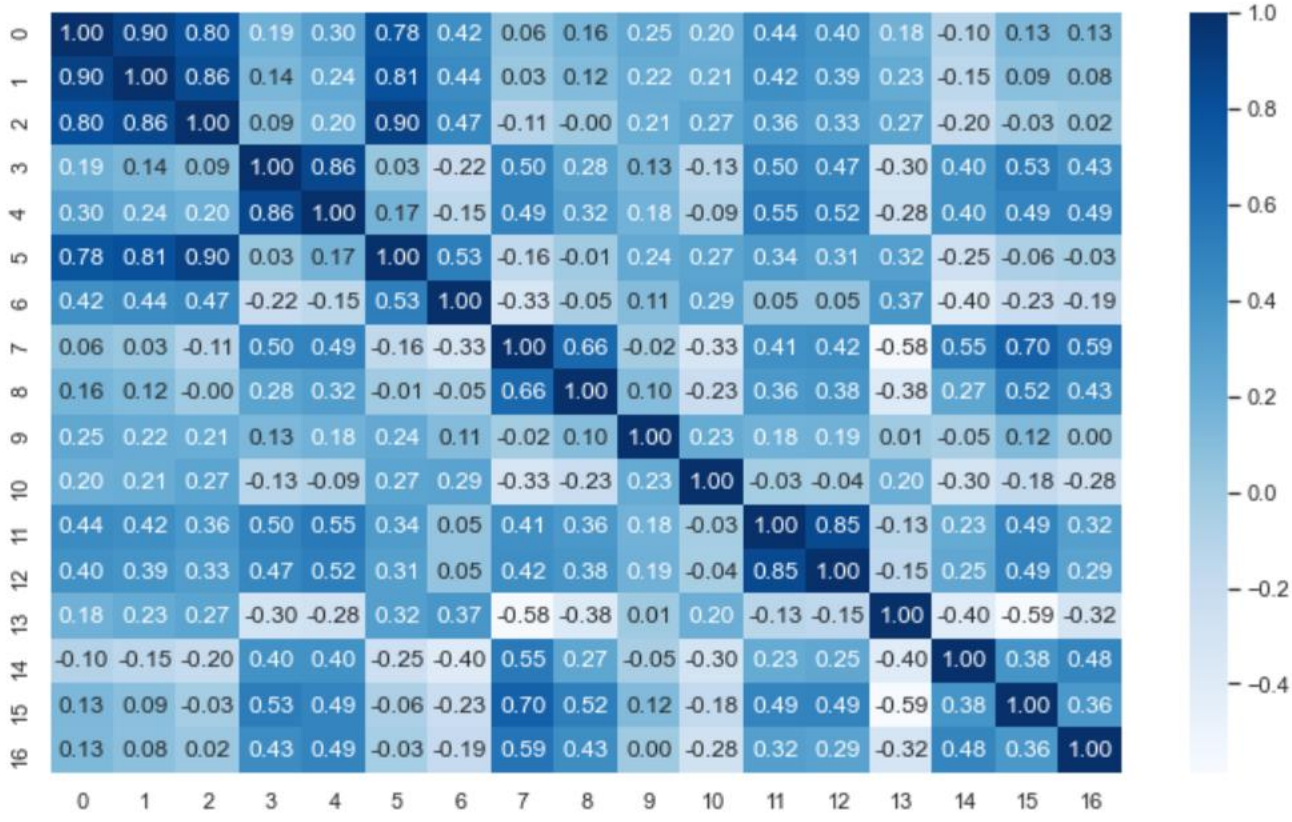
Now we are going to scale the data using the zscore scaling, since our data follows normal distribution. The following boxplot captures the data after scaling:



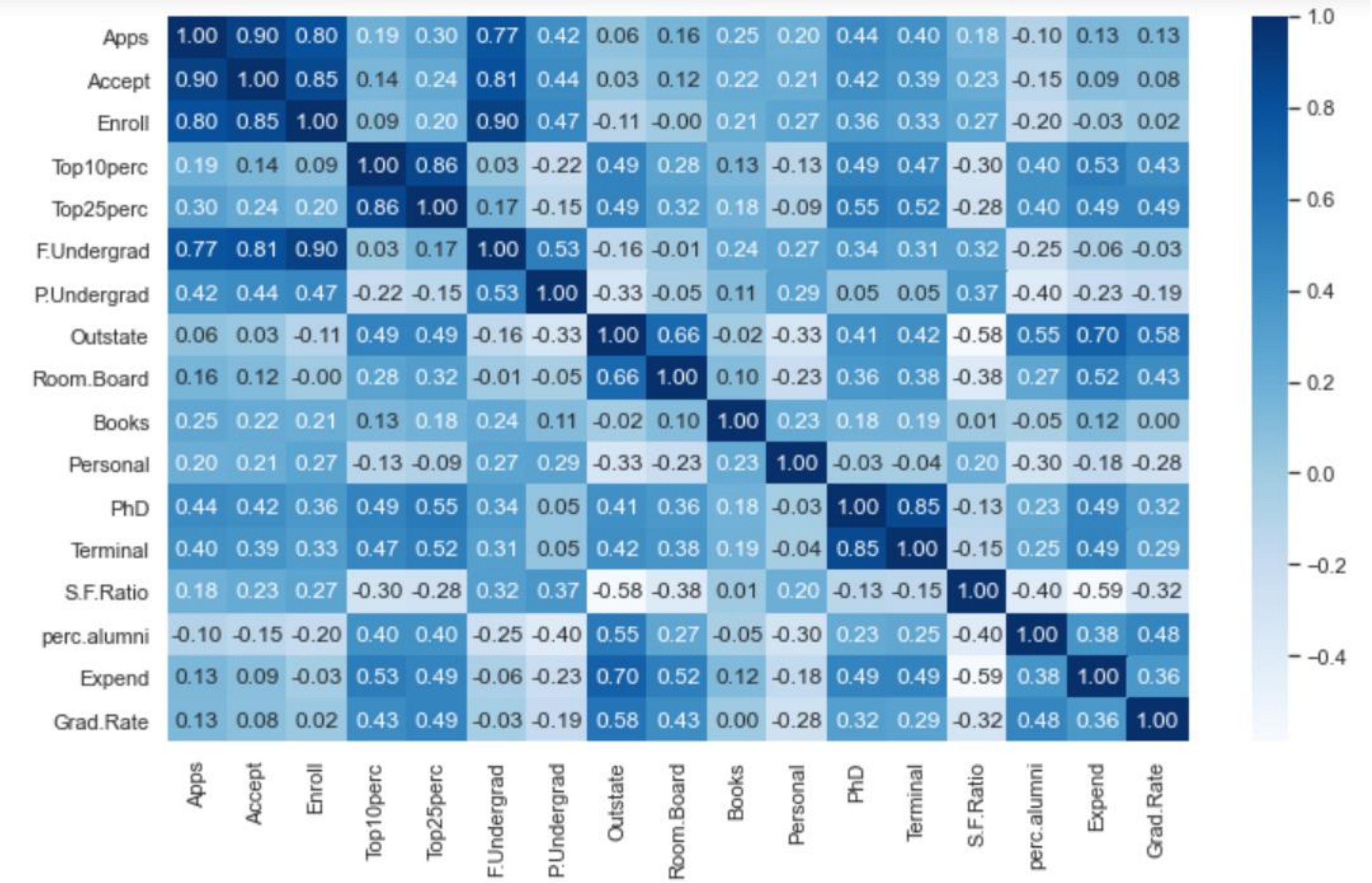
2.3 Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data]

Covariance is a measure to indicate the extent to which two random variables change in tandem. Covariance can vary between -∞ and +∞. Covariance is zero in case of independent variables (if one variable moves and the other doesn't) because then the variables do not necessarily move together.

Correlation is a measure used to represent how strongly two random variables are related to each other. Correlation ranges between -1 and +1. completely independent variables have a zero correlation.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	1.00	0.90	0.80	0.19	0.30	0.78	0.42	0.06	0.16	0.25	0.20	0.44	0.40	0.18	-0.10	0.13	0.13
1	0.90	1.00	0.86	0.14	0.24	0.81	0.44	0.03	0.12	0.22	0.21	0.42	0.39	0.23	-0.15	0.09	0.08
2	0.80	0.86	1.00	0.09	0.20	0.90	0.47	-0.11	-0.00	0.21	0.27	0.36	0.33	0.27	-0.20	-0.03	0.02
3	0.19	0.14	0.09	1.00	0.86	0.03	-0.22	0.50	0.28	0.13	-0.13	0.50	0.47	-0.30	0.40	0.53	0.43
4	0.30	0.24	0.20	0.86	1.00	0.17	-0.15	0.49	0.32	0.18	-0.09	0.55	0.52	-0.28	0.40	0.49	0.49
5	0.78	0.81	0.90	0.03	0.17	1.00	0.53	-0.16	-0.01	0.24	0.27	0.34	0.31	0.32	-0.25	-0.06	-0.03
6	0.42	0.44	0.47	-0.22	-0.15	0.53	1.00	-0.33	-0.05	0.11	0.29	0.05	0.05	0.37	-0.40	-0.23	-0.19
7	0.06	0.03	-0.11	0.50	0.49	-0.16	-0.33	1.00	0.66	-0.02	-0.33	0.41	0.42	-0.58	0.55	0.70	0.59
8	0.16	0.12	-0.00	0.28	0.32	-0.01	-0.05	0.66	1.00	0.10	-0.23	0.36	0.38	-0.38	0.27	0.52	0.43
9	0.25	0.22	0.21	0.13	0.18	0.24	0.11	-0.02	0.10	1.00	0.23	0.18	0.19	0.01	-0.05	0.12	0.00
10	0.20	0.21	0.27	-0.13	-0.09	0.27	0.29	-0.33	-0.23	0.23	1.00	-0.03	-0.04	0.20	-0.30	-0.18	-0.28
11	0.44	0.42	0.36	0.50	0.55	0.34	0.05	0.41	0.36	0.18	-0.03	1.00	0.85	-0.13	0.23	0.49	0.32
12	0.40	0.39	0.33	0.47	0.52	0.31	0.05	0.42	0.38	0.19	-0.04	0.85	1.00	-0.15	0.25	0.49	0.29
13	0.18	0.23	0.27	-0.30	-0.28	0.32	0.37	-0.58	-0.38	0.01	0.20	-0.13	-0.15	1.00	-0.40	-0.59	-0.32
14	-0.10	-0.15	-0.20	0.40	0.40	-0.25	-0.40	0.55	0.27	-0.05	-0.30	0.23	0.25	-0.40	1.00	0.38	0.48
15	0.13	0.09	-0.03	0.53	0.49	-0.06	-0.23	0.70	0.52	0.12	-0.18	0.49	0.49	-0.59	0.38	1.00	0.36
16	0.13	0.08	0.02	0.43	0.49	-0.03	-0.19	0.59	0.43	0.00	-0.28	0.32	0.29	-0.32	0.48	0.36	1.00

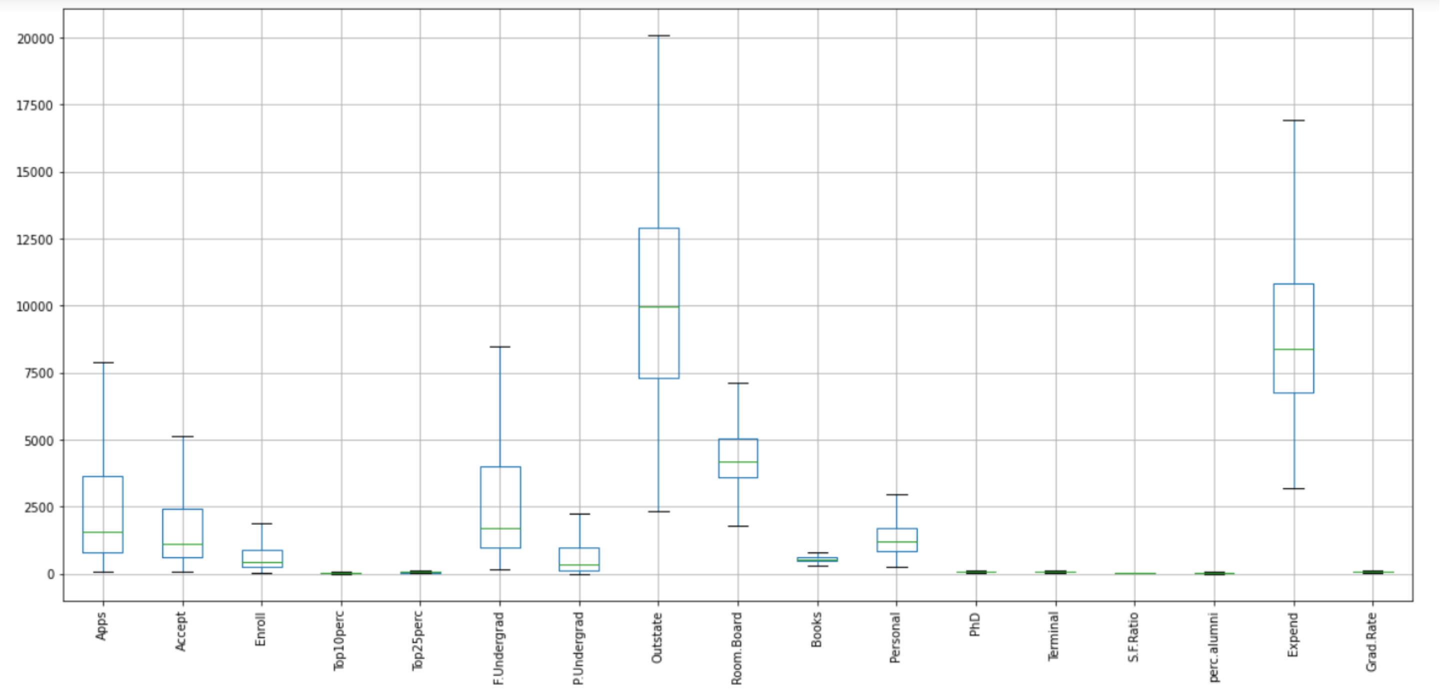


	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.00	0.90	0.80	0.19	0.30	0.77	0.42	0.06	0.16	0.25	0.20	0.44	0.40	0.18	-0.10	0.13	0.13
Accept	0.90	1.00	0.85	0.14	0.24	0.81	0.44	0.03	0.12	0.22	0.21	0.42	0.39	0.23	-0.15	0.09	0.08
Enroll	0.80	0.85	1.00	0.09	0.20	0.90	0.47	-0.11	-0.00	0.21	0.27	0.36	0.33	0.27	-0.20	-0.03	0.02
Top10perc	0.19	0.14	0.09	1.00	0.86	0.03	-0.22	0.49	0.28	0.13	-0.13	0.49	0.47	-0.30	0.40	0.53	0.43
Top25perc	0.30	0.24	0.20	0.86	1.00	0.17	-0.15	0.49	0.32	0.18	-0.09	0.55	0.52	-0.28	0.40	0.49	0.49
F.Undergrad	0.77	0.81	0.90	0.03	0.17	1.00	0.53	-0.16	-0.01	0.24	0.27	0.34	0.31	0.32	-0.25	-0.06	-0.03
P.Undergrad	0.42	0.44	0.47	-0.22	-0.15	0.53	1.00	-0.33	-0.05	0.11	0.29	0.05	0.05	0.37	-0.40	-0.23	-0.19
Outstate	0.06	0.03	-0.11	0.49	0.49	-0.16	-0.33	1.00	0.66	-0.02	-0.33	0.41	0.42	-0.58	0.55	0.70	0.58
Room.Board	0.16	0.12	-0.00	0.28	0.32	-0.01	-0.05	0.66	1.00	0.10	-0.23	0.36	0.38	-0.38	0.27	0.52	0.43
Books	0.25	0.22	0.21	0.13	0.18	0.24	0.11	-0.02	0.10	1.00	0.23	0.18	0.19	0.01	-0.05	0.12	0.00
Personal	0.20	0.21	0.27	-0.13	-0.09	0.27	0.29	-0.33	-0.23	0.23	1.00	-0.03	-0.04	0.20	-0.30	-0.18	-0.28
PhD	0.44	0.42	0.36	0.49	0.55	0.34	0.05	0.41	0.36	0.18	-0.03	1.00	0.85	-0.13	0.23	0.49	0.32
Terminal	0.40	0.39	0.33	0.47	0.52	0.31	0.05	0.42	0.38	0.19	-0.04	0.85	1.00	-0.15	0.25	0.49	0.29
S.F.Ratio	0.18	0.23	0.27	-0.30	-0.28	0.32	0.37	-0.58	-0.38	0.01	0.20	-0.13	-0.15	1.00	-0.40	-0.59	-0.32
perc.alumni	-0.10	-0.15	-0.20	0.40	0.40	-0.25	-0.40	0.55	0.27	-0.05	-0.30	0.23	0.25	-0.40	1.00	0.38	0.48
Expend	0.13	0.09	-0.03	0.53	0.49	-0.06	-0.23	0.70	0.52	0.12	-0.18	0.49	0.49	-0.59	0.38	1.00	0.36
Grad.Rate	0.13	0.08	0.02	0.43	0.49	-0.03	-0.19	0.58	0.43	0.00	-0.28	0.32	0.29	-0.32	0.48	0.36	1.00

Since we have scaled (z-score) data then the covariance matrix is a correlation matrix. ... Scaling to unit variance is scaling the original data to the standard deviation. Thus whether we scale to before or after the covariance matrix, the end result is still the same pattern of variation.

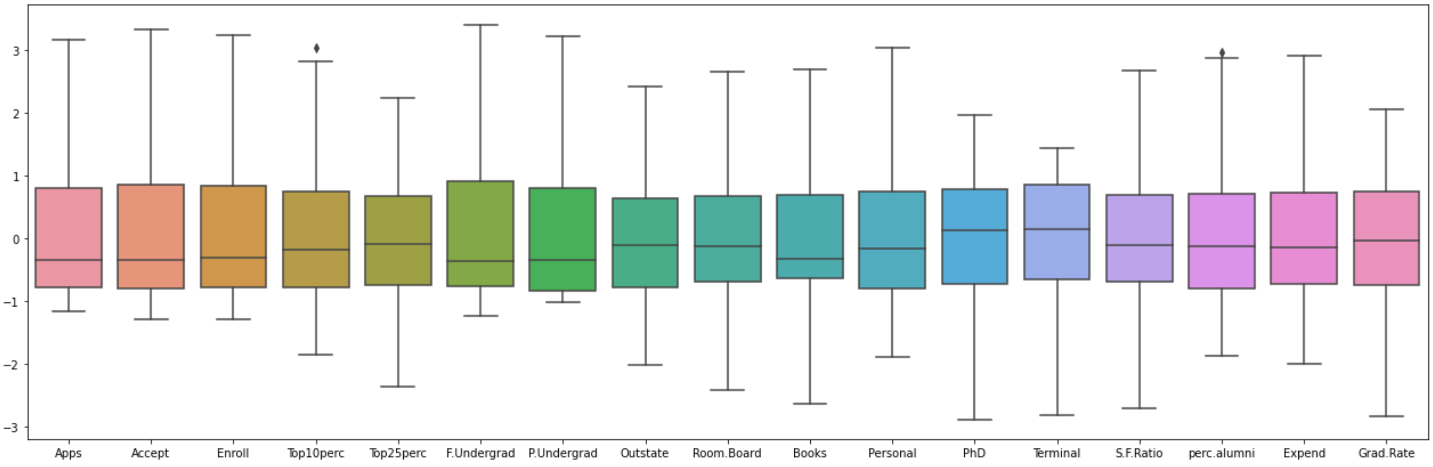
2.4 Check the dataset for outliers before and after scaling. What insight do you derive here?

Except the column showing the percentage of students from top 25 percent of higher secondary school, all the other columns had outliers. so what we have done is we have replaced the outliers with upper bound and lower bound values to minimize the effect of outliers. After treating the outliers we get the following boxplot:



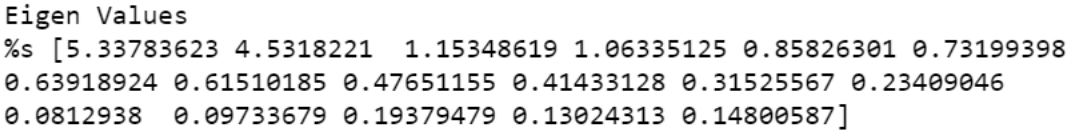
After z-score scaling, we can see that Z-score scaling removes the mean i.e. brings the mean to zero and scales the data to unit variance. The whole point of normalization is to change your observations so that they can be described as a normal distribution.

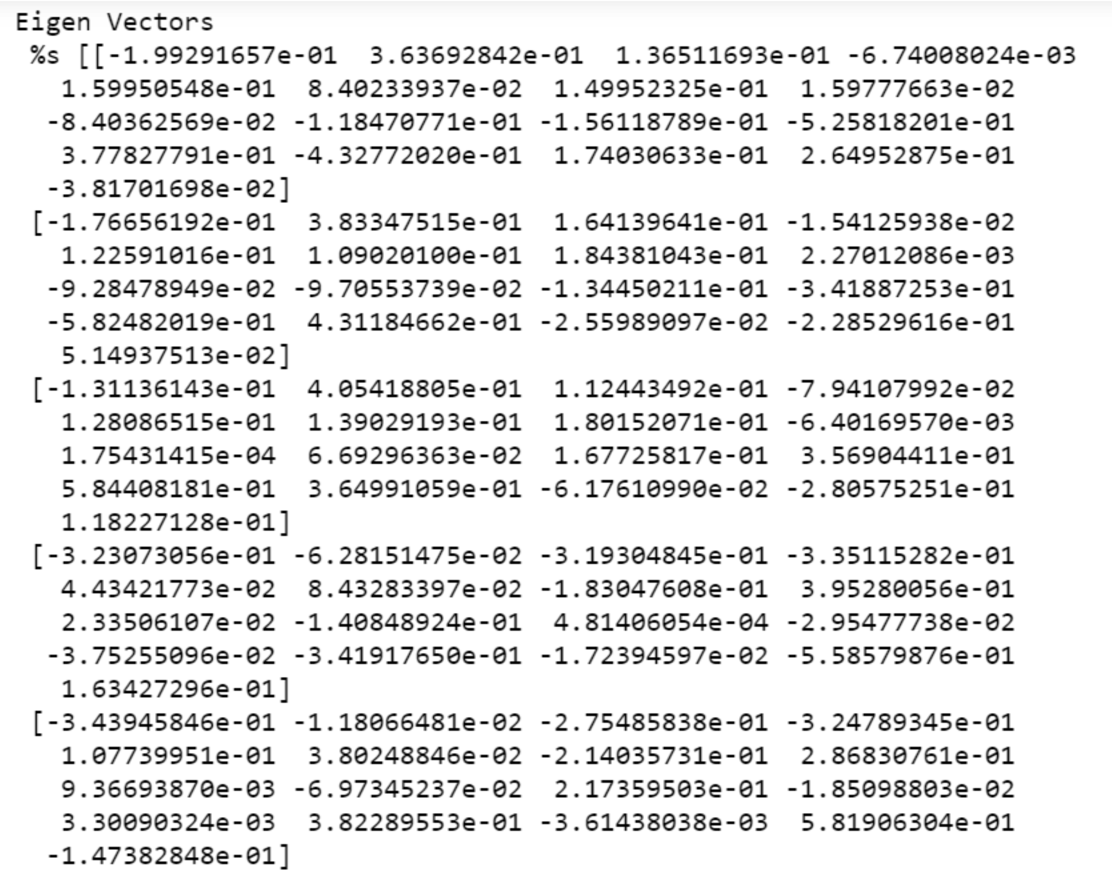
Normal distribution or Gaussian distribution is a bell curve, which is a specific statistical distribution where a roughly equal observations fall above and below the mean, the mean and the median are the same, and there are more observations closer to the mean.



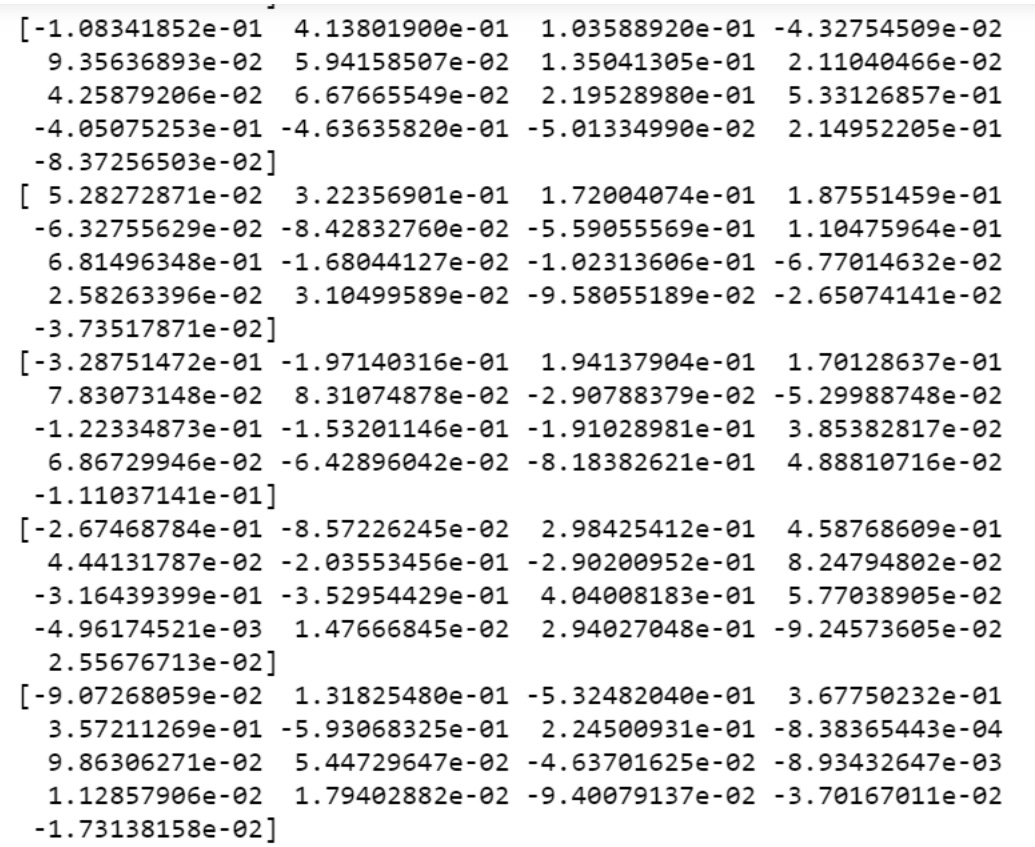
2.5 Extract the eigenvalues and eigenvectors. [print both]

After performing the decomposition of Covariance Matrix of the scaled data, we can extract the Eigen values and the Eigen Vectors

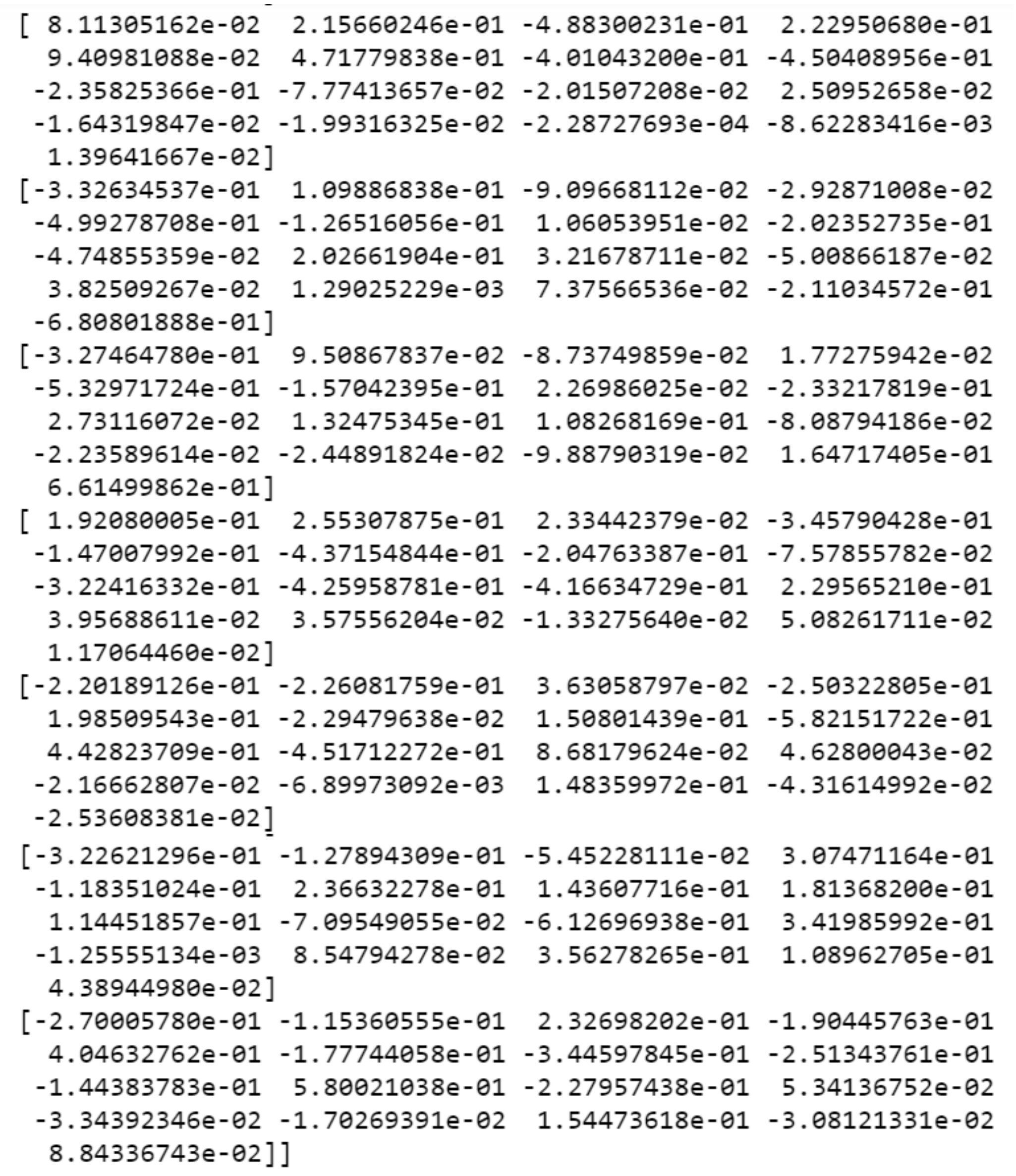

```
Eigen Values
%s [5.33783623 4.5318221  1.15348619 1.06335125 0.85826301 0.73199398
0.63918924 0.61510185 0.47651155 0.41433128 0.31525567 0.23409046
0.0812938  0.09733679 0.19379479 0.13024313 0.14800587]
```



```
Eigen Vectors
 %s [[-1.99291657e-01  3.63692842e-01  1.36511693e-01 -6.74008024e-03
   1.59950548e-01  8.40233937e-02  1.49952325e-01  1.59777663e-02
  -8.40362569e-02 -1.18470771e-01 -1.56118789e-01 -5.25818201e-01
   3.77827791e-01 -4.32772020e-01  1.74030633e-01  2.64952875e-01
  -3.81701698e-02]
 [-1.76656192e-01  3.83347515e-01  1.64139641e-01 -1.54125938e-02
   1.22591016e-01  1.09020100e-01  1.84381043e-01  2.27012086e-03
  -9.28478949e-02 -9.70553739e-02 -1.34450211e-01 -3.41887253e-01
  -5.82482019e-01  4.31184662e-01 -2.55989097e-02 -2.28529616e-01
   5.14937513e-02]
 [-1.31136143e-01  4.05418805e-01  1.12443492e-01 -7.94107992e-02
   1.28086515e-01  1.39029193e-01  1.80152071e-01 -6.40169570e-03
   1.75431415e-04  6.69296363e-02  1.67725817e-01  3.56904411e-01
   5.84408181e-01  3.64991059e-01 -6.17610990e-02 -2.80575251e-01
   1.18227128e-01]
 [-3.23073056e-01 -6.28151475e-02 -3.19304845e-01 -3.35115282e-01
   4.43421773e-02  8.43283397e-02 -1.83047608e-01  3.95280056e-01
   2.33506107e-02 -1.40848924e-01  4.81406054e-04 -2.95477738e-02
  -3.75255096e-02 -3.41917650e-01 -1.72394597e-02 -5.58579876e-01
   1.63427296e-01]
 [-3.43945846e-01 -1.18066481e-02 -2.75485838e-01 -3.24789345e-01
   1.07739951e-01  3.80248846e-02 -2.14035731e-01  2.86830761e-01
   9.36693870e-03 -6.97345237e-02  2.17359503e-01 -1.85098803e-02
   3.30090324e-03  3.82289553e-01 -3.61438038e-03  5.81906304e-01
  -1.47382848e-01]
```



```
[-1.08341852e-01  4.13801900e-01  1.03588920e-01 -4.32754509e-02
  9.35636893e-02  5.94158507e-02  1.35041305e-01  2.11040466e-02
  4.25879206e-02  6.67665549e-02  2.19528980e-01  5.33126857e-01
 -4.05075253e-01 -4.63635820e-01 -5.01334990e-02  2.14952205e-01
 -8.37256503e-02]
[ 5.28272871e-02  3.22356901e-01  1.72004074e-01  1.87551459e-01
 -6.32755629e-02 -8.42832760e-02 -5.59055569e-01  1.10475964e-01
  6.81496348e-01 -1.68044127e-02 -1.02313606e-01 -6.77014632e-02
  2.58263396e-02  3.10499589e-02 -9.58055189e-02 -2.65074141e-02
 -3.73517871e-02]
[-3.28751472e-01 -1.97140316e-01  1.94137904e-01  1.70128637e-01
  7.83073148e-02  8.31074878e-02 -2.90788379e-02 -5.29988748e-02
 -1.22334873e-01 -1.53201146e-01 -1.91028981e-01  3.85382817e-02
  6.86729946e-02 -6.42896042e-02 -8.18382621e-01  4.88810716e-02
 -1.11037141e-01]
[-2.67468784e-01 -8.57226245e-02  2.98425412e-01  4.58768609e-01
  4.44131787e-02 -2.03553456e-01 -2.90200952e-01  8.24794802e-02
 -3.16439399e-01 -3.52954429e-01  4.04008183e-01  5.77038905e-02
 -4.96174521e-03  1.47666845e-02  2.94027048e-01 -9.24573605e-02
  2.55676713e-02]
[-9.07268059e-02  1.31825480e-01 -5.32482040e-01  3.67750232e-01
  3.57211269e-01 -5.93068325e-01  2.24500931e-01 -8.38365443e-04
  9.86306271e-02  5.44729647e-02 -4.63701625e-02 -8.93432647e-03
  1.12857906e-02  1.79402882e-02 -9.40079137e-02 -3.70167011e-02
 -1.73138158e-02]
```



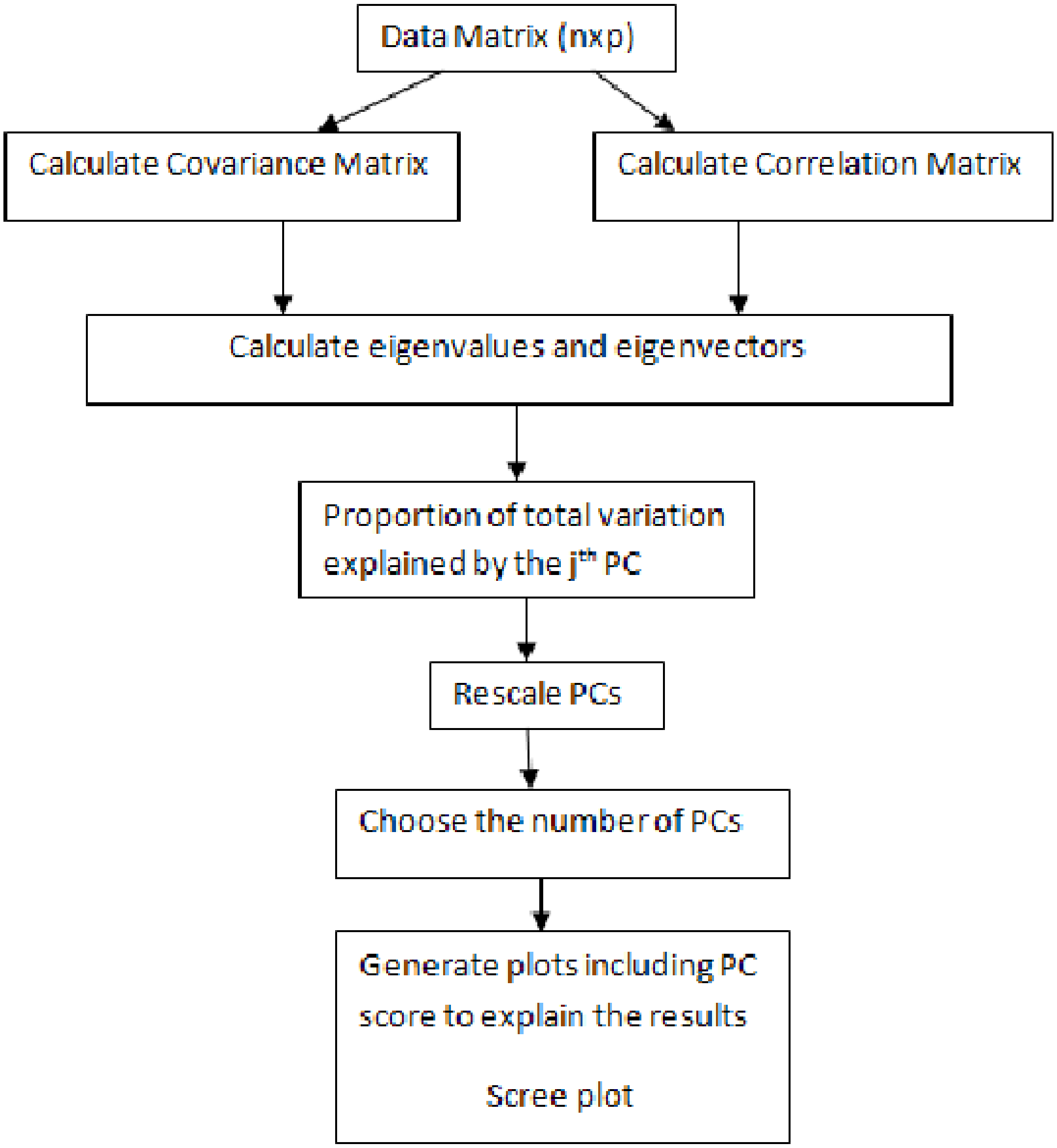
```
 [ 8.11305162e-02  2.15660246e-01 -4.88300231e-01  2.22950680e-01
   9.40981088e-02  4.71779838e-01 -4.01043200e-01 -4.50408956e-01
  -2.35825366e-01 -7.77413657e-02 -2.01507208e-02  2.50952658e-02
  -1.64319847e-02 -1.99316325e-02 -2.28727693e-04 -8.62283416e-03
   1.39641667e-02]
 [-3.32634537e-01  1.09886838e-01 -9.09668112e-02 -2.92871008e-02
  -4.99278708e-01 -1.26516056e-01  1.06053951e-02 -2.02352735e-01
  -4.74855359e-02  2.02661904e-01  3.21678711e-02 -5.00866187e-02
   3.82509267e-02  1.29025229e-03  7.37566536e-02 -2.11034572e-01
  -6.80801888e-01]
 [-3.27464780e-01  9.50867837e-02 -8.73749859e-02  1.77275942e-02
  -5.32971724e-01 -1.57042395e-01  2.26986025e-02 -2.33217819e-01
   2.73116072e-02  1.32475345e-01  1.08268169e-01 -8.08794186e-02
  -2.23589614e-02 -2.44891824e-02 -9.88790319e-02  1.64717405e-01
   6.61499862e-01]
 [ 1.92080005e-01  2.55307875e-01  2.33442379e-02 -3.45790428e-01
  -1.47007992e-01 -4.37154844e-01 -2.04763387e-01 -7.57855782e-02
  -3.22416332e-01 -4.25958781e-01 -4.16634729e-01  2.29565210e-01
   3.95688611e-02  3.57556204e-02 -1.33275640e-02  5.08261711e-02
   1.17064460e-02]
 [-2.20189126e-01 -2.26081759e-01  3.63058797e-02 -2.50322805e-01
   1.98509543e-01 -2.29479638e-02  1.50801439e-01 -5.82151722e-01
   4.42823709e-01 -4.51712272e-01  8.68179624e-02  4.62800043e-02
  -2.16662807e-02 -6.89973092e-03  1.48359972e-01 -4.31614992e-02
  -2.53608381e-02]
 [-3.22621296e-01 -1.27894309e-01 -5.45228111e-02  3.07471164e-01
  -1.18351024e-01  2.36632278e-01  1.43607716e-01  1.81368200e-01
   1.14451857e-01 -7.09549055e-02 -6.12696938e-01  3.41985992e-01
  -1.25555134e-03  8.54794278e-02  3.56278265e-01  1.08962705e-01
   4.38944980e-02]
 [-2.70005780e-01 -1.15360555e-01  2.32698202e-01 -1.90445763e-01
   4.04632762e-01 -1.77744058e-01 -3.44597845e-01 -2.51343761e-01
  -1.44383783e-01  5.80021038e-01 -2.27957438e-01  5.34136752e-02
  -3.34392346e-02 -1.70269391e-02  1.54473618e-01 -3.08121331e-02
   8.84336743e-02]]
```

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

Multivariate Analysis

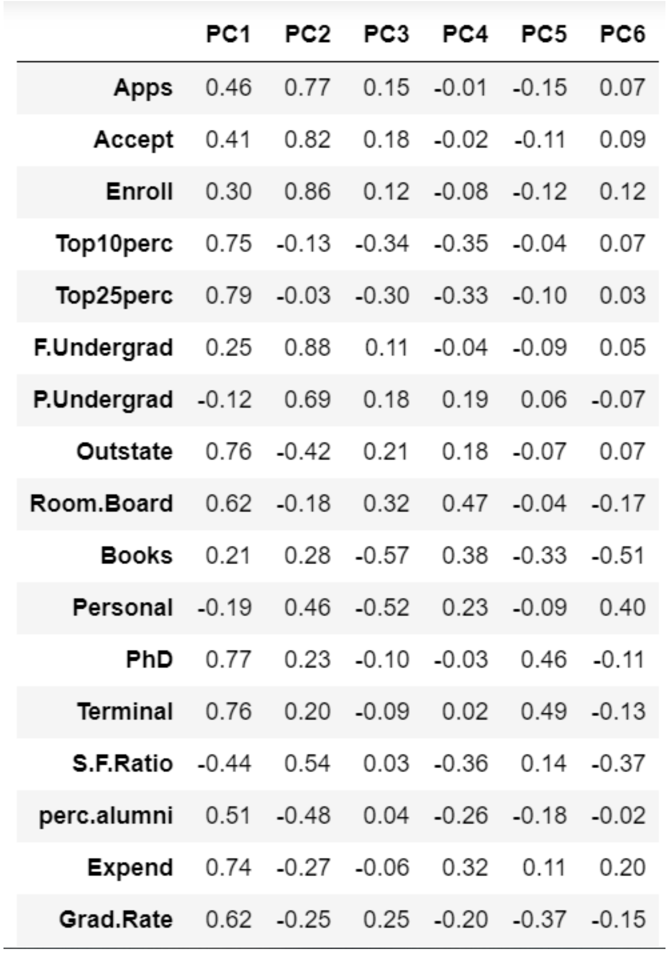
Multivariate analysis is required when more than two variables have to be analyzed simultaneously. It is a tremendously hard task for the human brain to visualize a relationship among 4 variables in a graph and thus multivariate analysis is used to study more complex sets of data. Types of Multivariate Analysis include Cluster Analysis, Factor Analysis, Multiple Regression Analysis, Principal Component Analysis, etc. More than 20 different ways to perform multivariate analysis exist and which one to choose depends upon the type of data and the end goal to achieve.

The steps involved in principal component analysis can be summarized as follows:

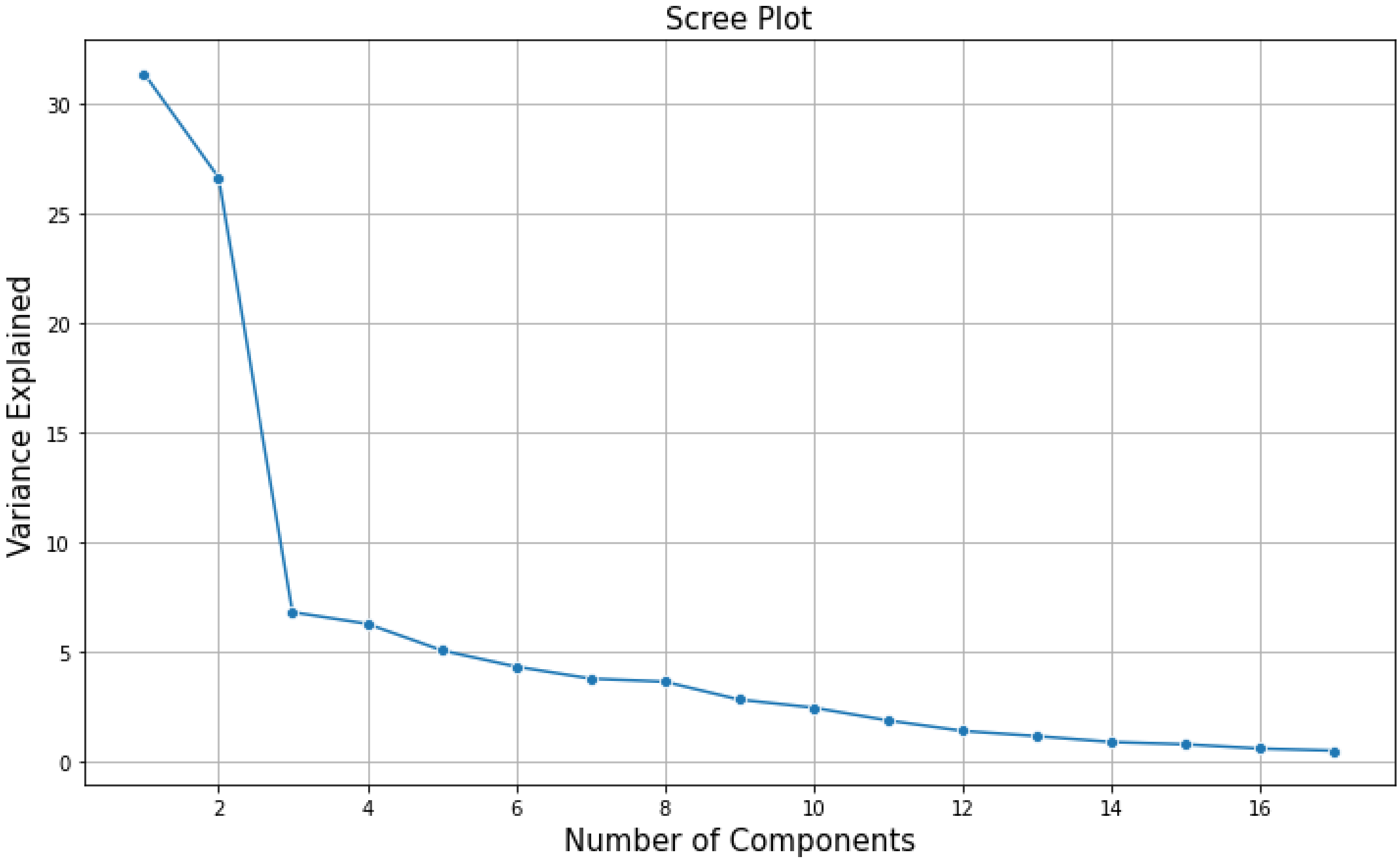


Data Matrix (nxp)
Calculate Covariance Matrix
Calculate Correlation Matrix
Calculate eigenvalues and eigenvectors
Proportion of total variation explained by the jth PC
Rescale PCs
Choose the number of PCs
Generate plots including PC score to explain the results
Scree plot

Principal components are linear combinations of the original variables. Each PC is a linear combination of all variables, or scaled variables, as the case may be. It is possible that some of the coefficients are very small numbers or close to 0. We present the linear combinations that make up the first 6 PC's.



	PC1	PC2	PC3	PC4	PC5	PC6
Apps	0.46	0.77	0.15	-0.01	-0.15	0.07
Accept	0.41	0.82	0.18	-0.02	-0.11	0.09
Enroll	0.30	0.86	0.12	-0.08	-0.12	0.12
Top10perc	0.75	-0.13	-0.34	-0.35	-0.04	0.07
Top25perc	0.79	-0.03	-0.30	-0.33	-0.10	0.03
F.Undergrad	0.25	0.88	0.11	-0.04	-0.09	0.05
P.Undergrad	-0.12	0.69	0.18	0.19	0.06	-0.07
Outstate	0.76	-0.42	0.21	0.18	-0.07	0.07
Room.Board	0.62	-0.18	0.32	0.47	-0.04	-0.17
Books	0.21	0.28	-0.57	0.38	-0.33	-0.51
Personal	-0.19	0.46	-0.52	0.23	-0.09	0.40
PhD	0.77	0.23	-0.10	-0.03	0.46	-0.11
Terminal	0.76	0.20	-0.09	0.02	0.49	-0.13
S.F.Ratio	-0.44	0.54	0.03	-0.36	0.14	-0.37
perc.alumni	0.51	-0.48	0.04	-0.26	-0.18	-0.02
Expend	0.74	-0.27	-0.06	0.32	0.11	0.20
Grad.Rate	0.62	-0.25	0.25	-0.20	-0.37	-0.15



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]

Following in the linear equation for the first component:

PC1 = a1x1 + a2x2 + a3x3 + a4x4 + a5x5 + a6x6 + a7x7 + a8x8 + a9x9 + a10x10 + a11x11 + a12x12 + a13x13 + a14x14 + a15x15 + a16x16 + a17x17

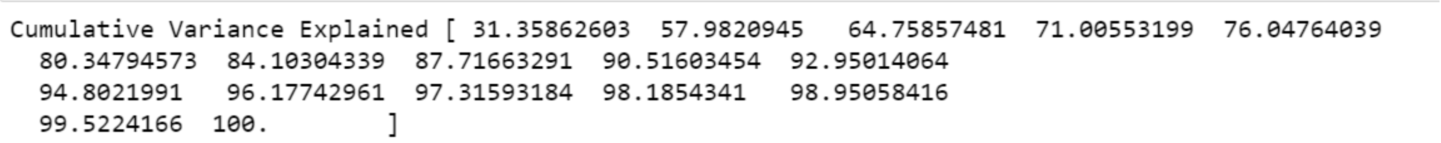
PC1 = 0.20 * **Apps** + 0.18 * **Accept** + 0.13 * **Enroll** + 0.32 * **Top10perc** + 0.34 * **Top25perc** + 0.11 * **Undergrad** + -0.05 * **P.Undergrad** + 0.33 * **Outstate** + 0.27* **Room.Board** + 0.09* **Books +** -0.08* **Personal +** 0.33* **PhD +** 0.33* **Terminal +** -0.19* **S.F.Ratio +** 0.22* **perc.alumni +** 0.32 * **Expend +** 0.27* **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?

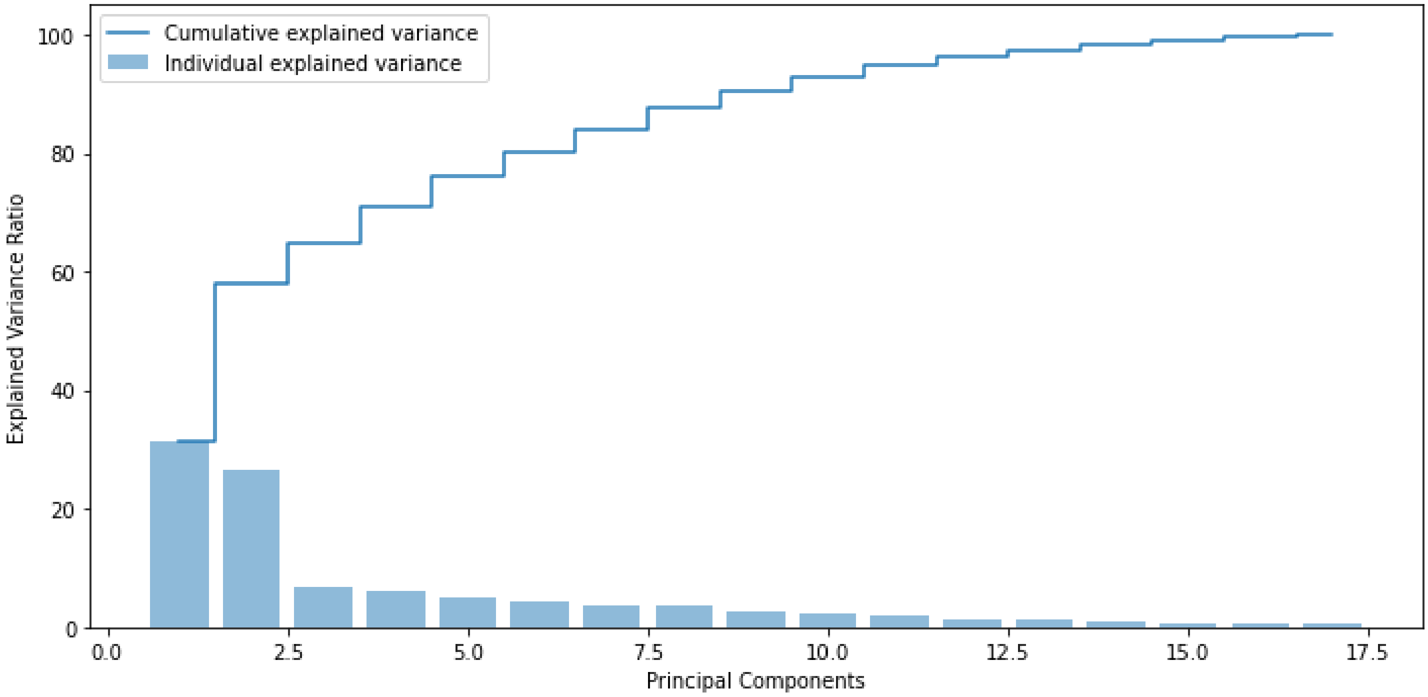
All principal components are extracted at one go and then optimum number of components decided.
There are 17 observed variables and hence 17 PCs are generated. The principal components are constructed in decreasing order of magnitude of their standard deviations, which is equivalent to decreasing order of magnitude of their variances.

The proportion of variance of a principal component is obtained by dividing the variance of the component (obtained by squaring the standard deviation), by total variance. The cumulative proportion upto the k-th principal component is the sum of the proportions of variances upto the$k$-th component.

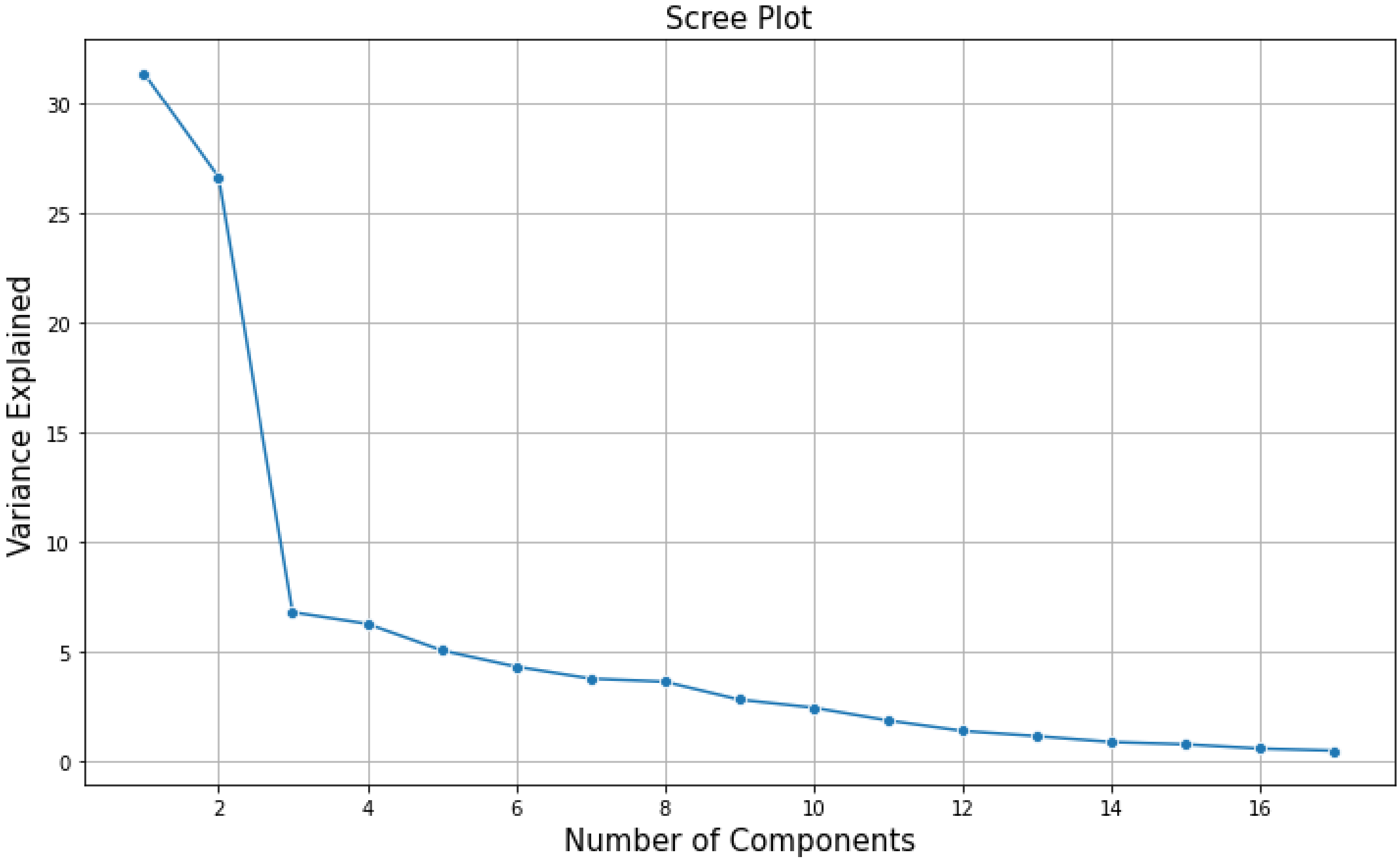
If k = 6 , cumulative proportion is 80.35%. Although there are 17 observed variables, the first 6 principal components can explain more than 80% of the total variation. Hence it is sufficient to use the first 6 PCs instead of the original 17 variables, thereby reducing the dimensions by half



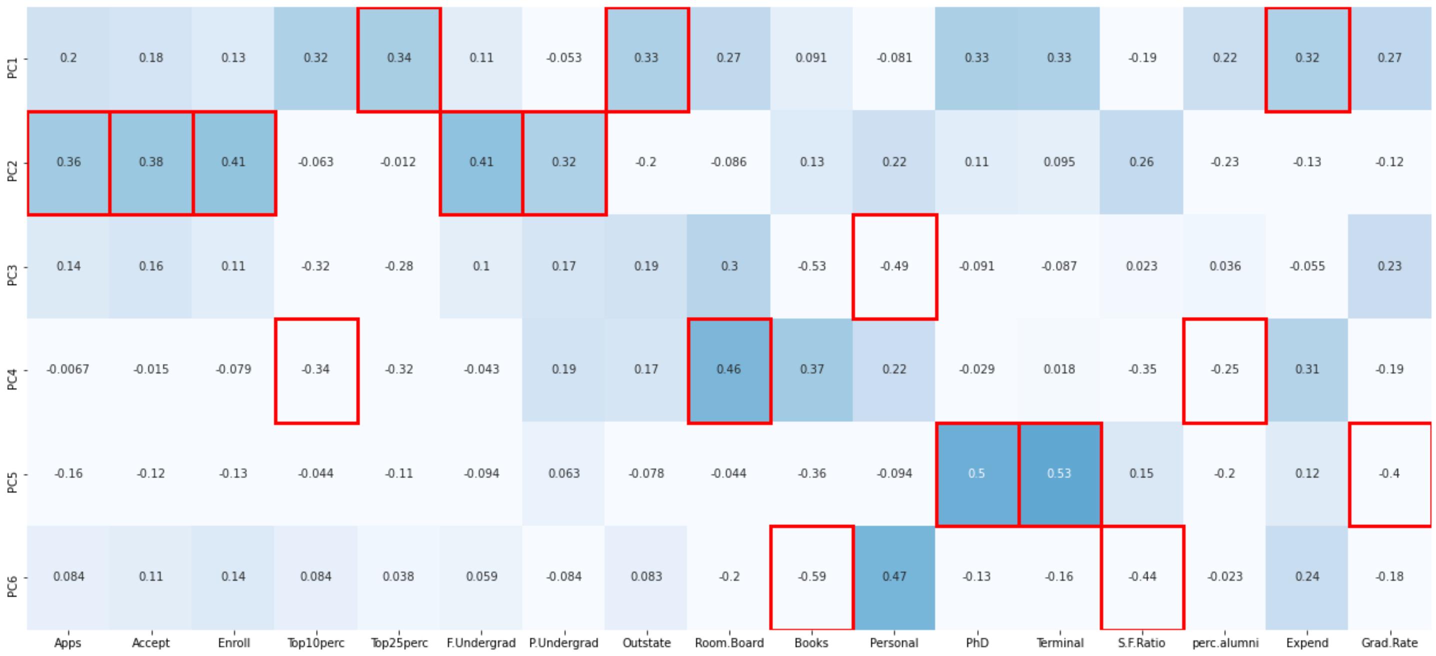
```
Cumulative Variance Explained [ 31.35862603  57.9820945   64.75857481  71.00553199  76.04764039
  80.34794573  84.10304339  87.71663291  90.51603454  92.95014064
  94.8021991   96.17742961  97.31593184  98.1854341   98.95058416
  99.5224166  100.        ]
```



We can also use the scree plot to decide the number of components, based on the analysis we have decided to used first 6 PCs. Principal components are linear combinations of the original variables. Each PC is a linear combination of all variables, or scaled variables.



To check that the PCs are orthogonal, correlation matrix is computed.



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
PC1	0.2	0.18	0.13	0.32	0.34	0.11	-0.053	0.33	0.27	0.091	-0.081	0.33	0.33	-0.19	0.22	0.32	0.27
PC2	0.36	0.38	0.41	-0.063	-0.012	0.41	0.32	-0.2	-0.086	0.13	0.22	0.11	0.095	0.26	-0.23	-0.13	-0.12
PC3	0.14	0.16	0.11	-0.32	-0.28	0.1	0.17	0.19	0.3	-0.53	-0.49	-0.091	-0.087	0.023	0.036	-0.055	0.23
PC4	-0.0067	-0.015	-0.079	-0.34	-0.32	-0.043	0.19	0.17	0.46	0.37	0.22	-0.029	0.018	-0.35	-0.25	0.31	-0.19
PC5	-0.16	-0.12	-0.13	-0.044	-0.11	-0.094	0.063	-0.078	-0.044	-0.36	-0.094	0.5	0.53	0.15	-0.2	0.12	-0.4
PC6	0.084	0.11	0.14	0.084	0.038	0.059	-0.084	0.083	-0.2	-0.59	0.47	-0.13	-0.16	-0.44	-0.023	0.24	-0.18

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]

Principal component analysis is technique used for dimension reduction or representation in lower dimension space with larger variability capture.

PCA also helps with the signal to noise ratio. Greater the SNR (signal to Noise ratio) better the model.

Signal to Noise Ratio = Variance in Noise / Variance in Signal

PCA also helps to address the problem of multicollinearity.

When Should I Do PCA?

- Do you want to reduce the number of variables, but aren’t able to identify variables to completely remove from consideration?
- Do you want to ensure your variables are independent of one another?
- Are you comfortable making your interdependent variable interpretable?
- How can we reduce the dimensions without losing the information content present in the variables? Whenever we remove any of the features, we are losing the signal or the information available in the data

In this case study, since we were presented with data for 777 Colleges across USA there were too many features influencing the information of various colleges. This data can be used by colleges to provide better infrastructure and understand the needs of students. Depending on the business requirement, after reducing the data to 6 principal components we can build models to help reduce cost of tuition and other expenditure, rate the standard of quality of education.

Principal component analysis is usually used as an intermediary step to further analysis. Outcome of PCA may be used in regression, such as principal component regression or partial least squares regression. It can also be used in extraction of latent factors, which often provides important insights into various business applications, such as customer behavior, ecommerce etc. This is known as Factor Analysis.

Problem 1A

- one-way ANOVA (Education)
- one-way ANOVA(Occupation)

Problem 1B

- two-way ANOVA (Education and Occupation)

Problem 2

- Exploratory Data Analysis
- Principal Component Analysis

THANK YOU