

SALARY AND EDUCATION

DATA

ANALYSIS

SIDDHANT SINGH

PGP

-

DSBA Online

AUGUST,22

Date: 14/08/2022

**Problem1...........................................................................................**

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

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.

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.

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.

1.5 What is the interaction between the two treatments? Analyse the effects of one variable on the other (Education and Occupation) with the help of an interaction plot. *Page 4* 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?

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

**Problem2...........................................................................................**

2.1 Perform Exploratory Data Analysis (both univariate and multivariate analysis to be

performed). What insight do you draw from the EDA?

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

2.3 Comment on the comparison between the covariance and the correlation matrices from this data (on scaled data).

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

*Page 10* 2.5 Extract the eigenvalues and eigenvectors.

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

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

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?

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).

# **Problem 1A**

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [[SalaryData.csv](https://olympus.mygreatlearning.com/courses/78048/files/6264520/download?verifier=5ppokrH0tIDVqC7v7msZifopeRroBmzqjCylk7Z0&wrap=1)] 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

**For Salary and Education**

H0 : Salary is dependent on educational qualification.

H1 : Salary is not dependent on educational qualification.

Confidence level (α) = 0.05

**For Salary and Occupation**

HO : Salary is dependent on occupation

H1 : Salary is not dependent on educational qualification

Confidence level (α) = 0.05

# 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

**Checking the Normality Condition of ANOVA**

The Shapiro- Wilk test tests the Null Hypothesis that the data was drawn from a normal distribution

**Shapiro Result(statistic=0.8952829837799072, pvalue=0.0675690770149231)**

**Shapiro Result(statistic=0.9607304334640503, pvalue=0.7050924301147461)**

**Shapiro Result(statistic=0.885286271572113, pvalue=0.1783432960510254)**

since p- value for all 3 continuous variables which are grouped by their respective educational qualifications is greater than 0.05, we fail to reject the Null Hypothesis. Therefore we can say that they are Normally distributed.

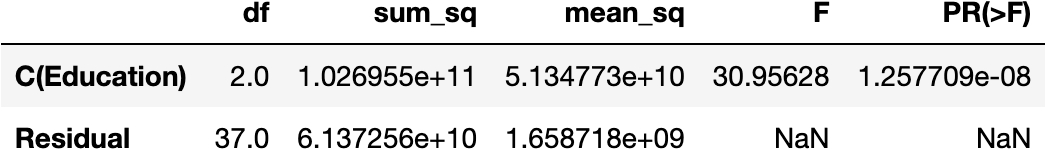
**Checking variance equality condition of ANOVA**

The Levene test tests the Null Hypothesis that all input samples are from populations with equal variances

**Levene Result(statistic=1.8800921605836554, pvalue=0.16686425699301183)**

since p- value is greater than 0.05 we failed to reject the null hypothesis. Therefore, we can say that the input samples are from populations with equal variances

**Analyses of Variance Test**



***since p- value is less than the significance level, we can reject the Null Hypothesis and state that Salary is not dependent on educational qualification***

# 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

**Checking the Normality Condition of ANOVA**

The Shapiro- Wilk test tests the Null Hypothesis that the data was drawn from a normal distribution

**Shapiro Result(statistic=0.736305832862854, pvalue=0.001313115586526692)**

**Shapiro Result(statistic=0.8897128701210022, pvalue=0.11683973670005798)**

**Shapiro Result(statistic=0.9189430475234985, pvalue=0.34822404384613037)**

**Shapiro Result(statistic=0.6901877522468567, pvalue=0.007539781276136637)**

since p- value for Occupation - 'Prof-specialty' and 'Exec-managerial' is less than 0.05, we can say that they are not Normally Distributed. But for Occupation - 'Sales' and 'Adm-clerical' p- value is more than 0.05. we can say that they are Normally Distributed. Since 50% of the samples are Normally Distributed, we can proceed with ANOVA

**Checking variance equality condition of ANOVA**

The Levene test tests the Null Hypothesis that all input samples are from populations with equal variances

**Levene Result(statistic=2.4378177404396832, pvalue=0.0803790714975064)**

since p- value is greater than 0.05 we failed to reject the null hypothesis. Therefore, we can say that the input samples are from populations with equal variances

**Analyses of Variance Test**

**Table

Description automatically generated with low confidence**

***since p- value is greater than the significance level, we failed to reject the Null Hypothesis. We can say that salary is dependent on Occupation***

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

**For Salary and Education**

Table

Description automatically generated with low confidence

**For Salary and Occupation**

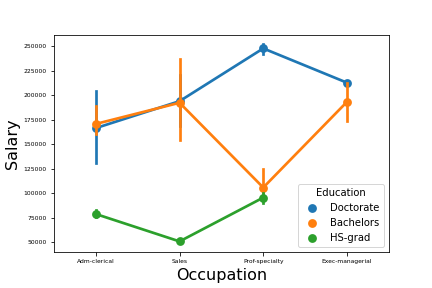
Table

Description automatically generated with low confidence

***mean\_sq difference is more between Educational qualification class this means that Education class means are significantly different.***

# **Problem 1B**

1. What is the interaction between two treatments? Analyse the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the ‘point plot’ function from the ‘seaborn’ function]



Following are the key points observed from the interaction plot:

* Adm-clerical and sales having bachelors and doctorate degree earn similar salary
* Prof-specialty having Doctorate degree earns the maximum
* Prof-specialty having degree Bachelors and HS-grad have similar salary.

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?

HO : The Salary is dependent on Both Education and Occupation

H1 : The Salary is not dependent on Both Education and Occupation

Confidence Level (α): 0.05

**Checking the Normality Condition of ANOVA**

The Shapiro- Wilk test tests the Null Hypothesis that the data was drawn from a normal distribution

**Shapiro Result(statistic=0.8571751117706299, pvalue=0.1796833723783493)**

**Shapiro Result(statistic=0.9059374332427979, pvalue=0.4435662627220154)**

**Shapiro Result(statistic=0.9969509243965149, pvalue=0.9896460175514221)**

**Shapiro Result(statistic=0.8238510489463806, pvalue=0.15232744812965393)**

**Shapiro Result(statistic=0.9072796702384949, pvalue=0.4681086838245392)**

**Shapiro Result(statistic=0.7925583720207214, pvalue=0.09685297310352325)**

**Shapiro Result(statistic=0.7352886199951172, pvalue=0.027851471677422523)**

**Shapiro Result(statistic=0.9998774528503418, pvalue=0.9788563251495361)**

**Shapiro Result(statistic=0.7566614151000977, pvalue=0.014755938202142715)**

**Shapiro Result(statistic=0.9523223638534546, pvalue=0.579590380191803)**

For more than 50% of the data collected, p- value is greater than 0.05. Therefore, for more than 50% of the data collected we failed to reject the Null Hypothesis. As a result more than 50% of the sample data is normally distributed and we can proceed with ANOVA test

**Checking variance equality condition of ANOVA**

The Levene test tests the Null Hypothesis that all input samples are from populations with equal variances

**Levene Result(statistic=1.9901869818184943, pvalue=0.07242074266567151)**

since p- value is greater than 0.05, we failed to reject the Null Hypothesis. Therefore we can conclude that all input samples are from populations with equal variance

**Analyses of Variance Test**

**Table

Description automatically generated**

***since p- value is less than 0.05, we failed to reject the Null Hypothesis. We can state that Salary is dependent on both Education and Occupation***

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

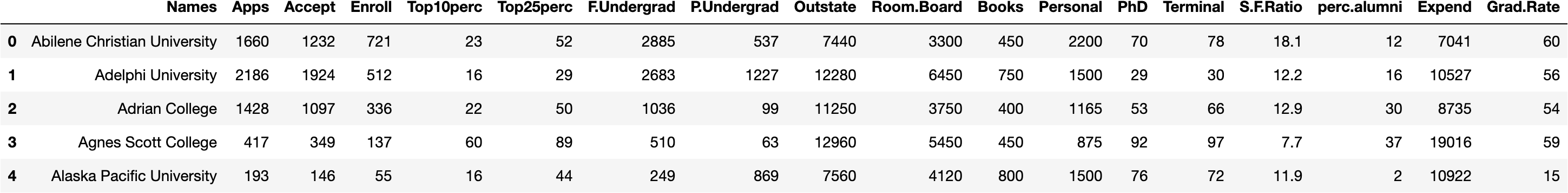
By performing ANOVA, we can conclude that salary is dependent on occupation but it is not dependent on education qualification. When we consider both the class i.e. education and occupation we can say that salary is moderately dependent on both of these classes. Adm-Clerical and Sales professionals with Bachelors and Doctorate degrees earn almost similar salary packages, where HS-grad has low salary packages in every Occupation as compare to Bachelors and Doctorate.

# **Problem 2A**

The dataset [Education - Post 12th Standard.csv](https://olympus.mygreatlearning.com/courses/78048/files/5473308/download?verifier=L3qWQDr2maTP7u4LKvMz7hEFleIBNEnen8zTSyC8&wrap=1) 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](https://olympus.mygreatlearning.com/courses/78048/files/5473307/download?verifier=JWXFQwEXkgMdjJvY5jv37PYJmMZofHwnZKdREdBA&wrap=1).

After loading the dataset, we perform Initial Diagnosis on the loaded dataset.

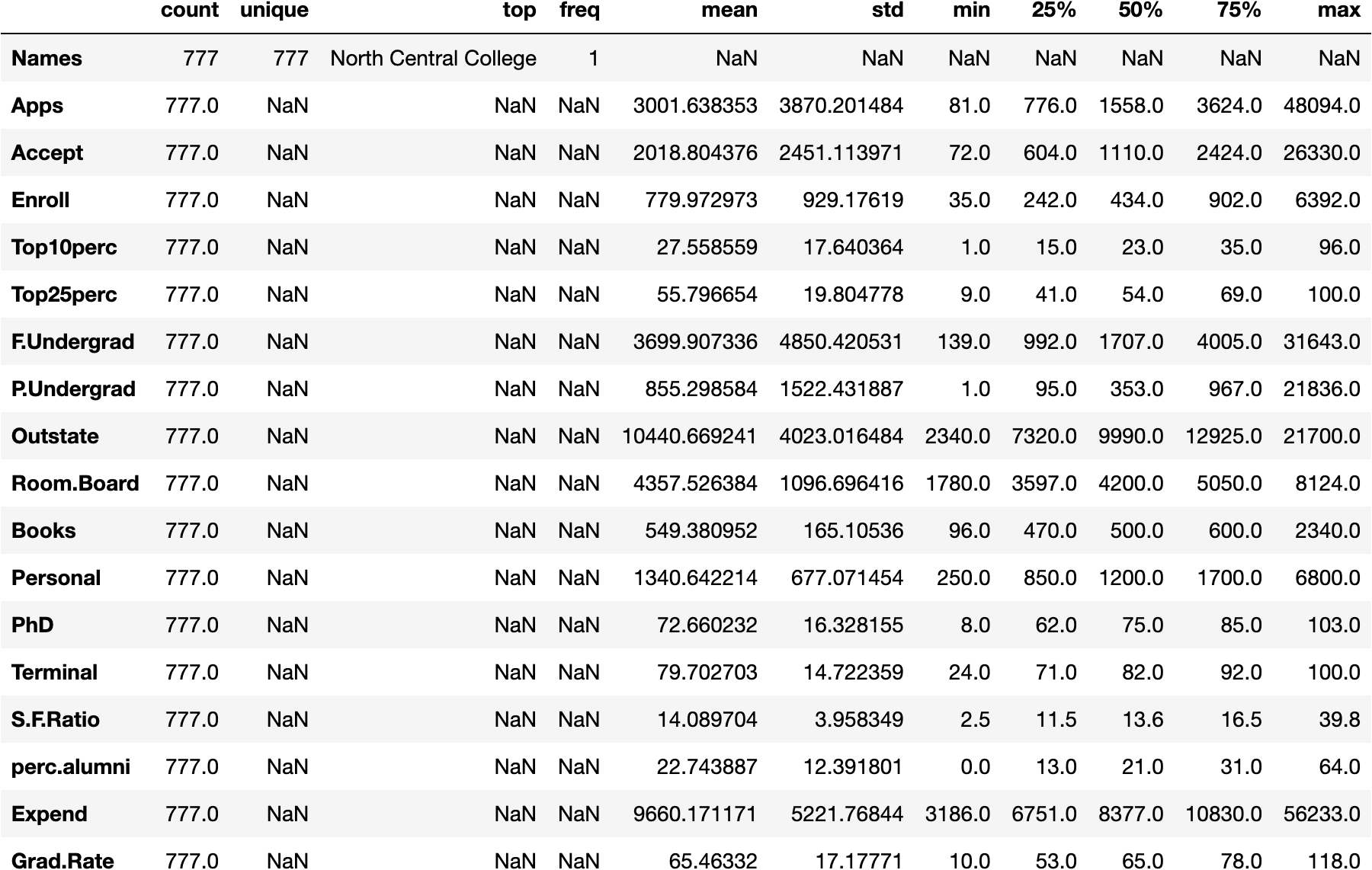
The top 5 rows of the data set are:



The bottom 5 rows of the data set are:



On checking the summary of the data we get the following result:

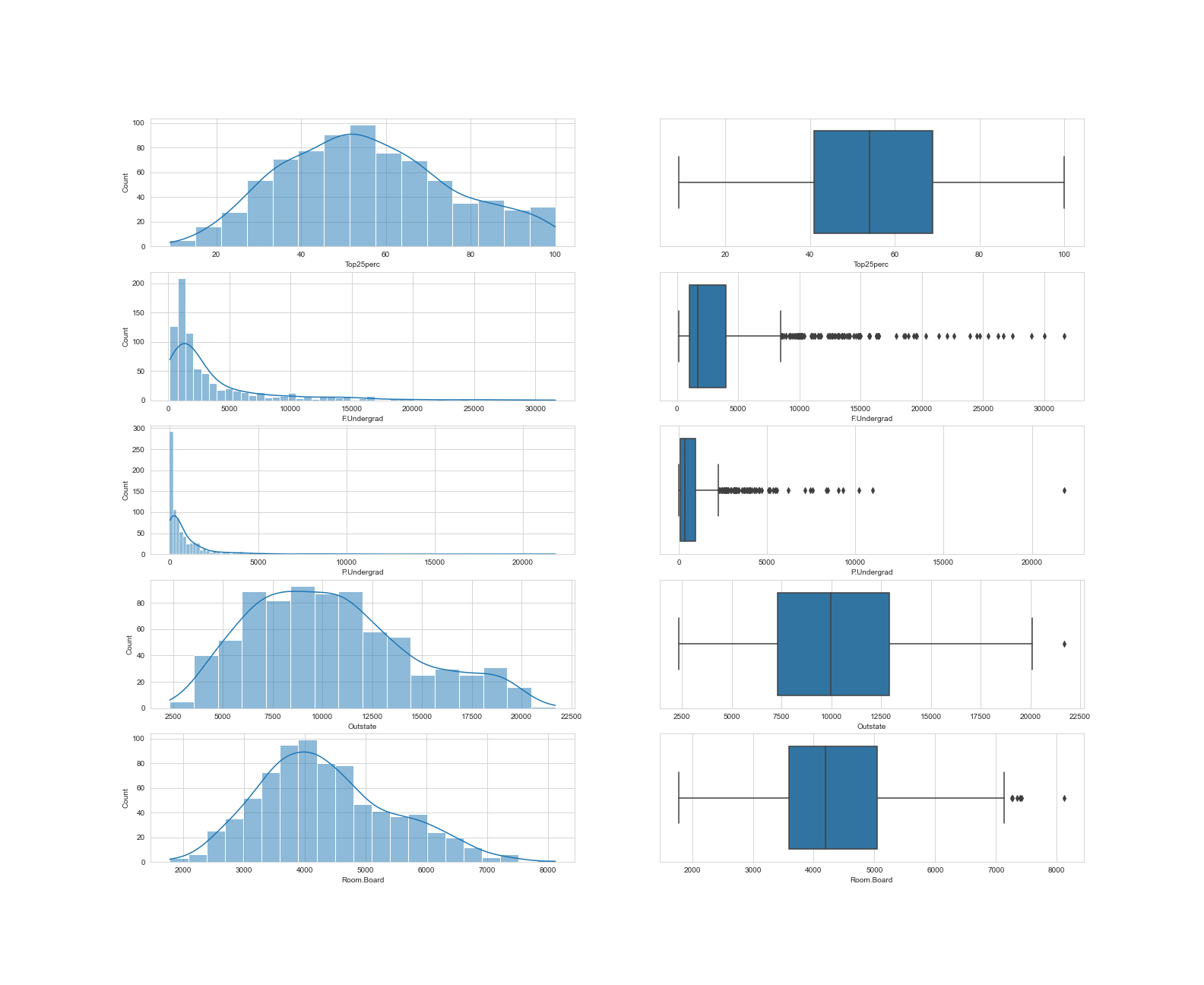


***After performing the initial diagnosis on the dataset we find that the data has 777 rows and 18 columns in total. The given data has 16 integer, 1 float and 1 object type data. There are no duplicated row in the dataset. There are no null values in the dataset as well.***

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

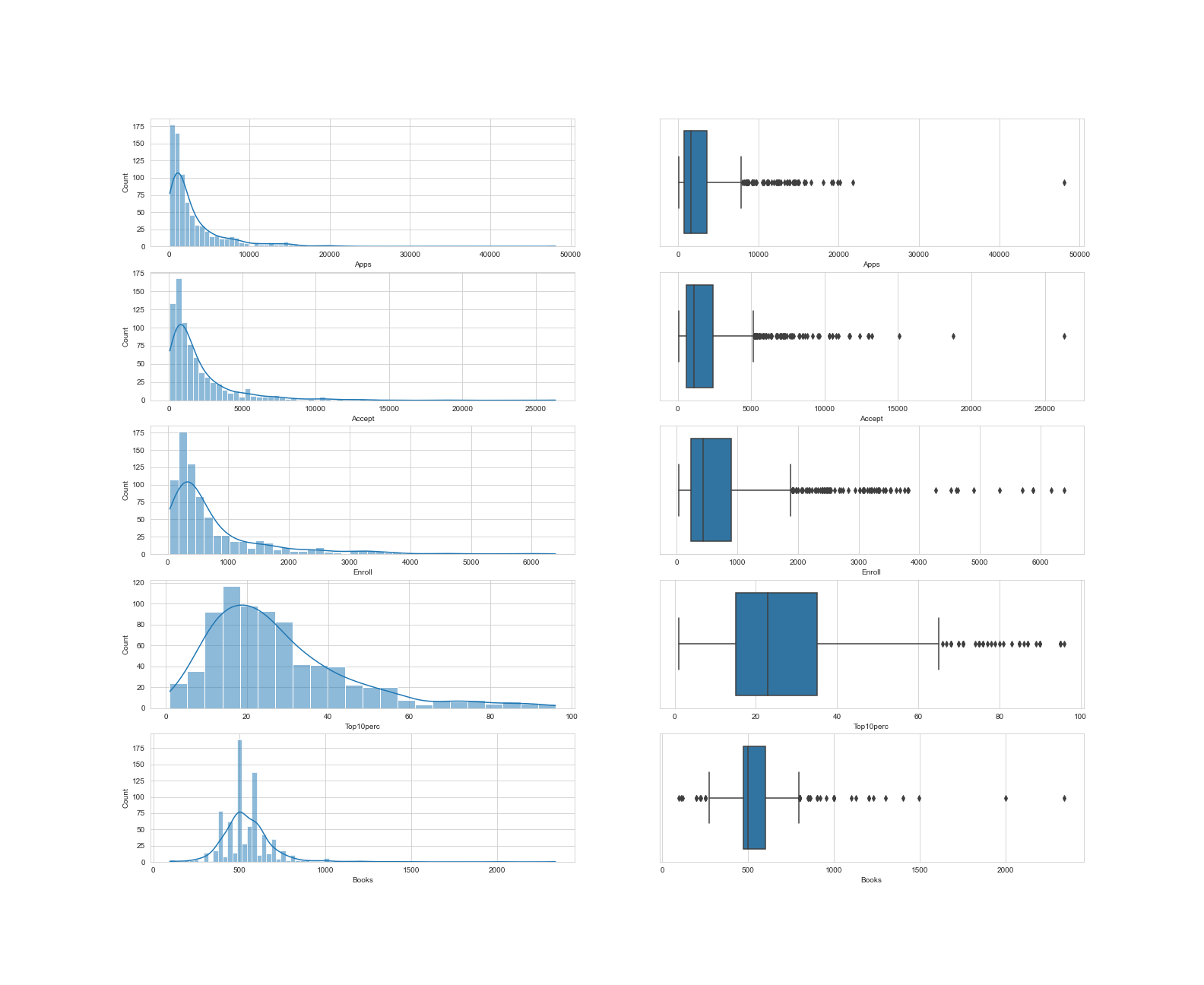
**Univariate Analysis**

We perform univariate analysis to describe the data and find any pattern within the data.

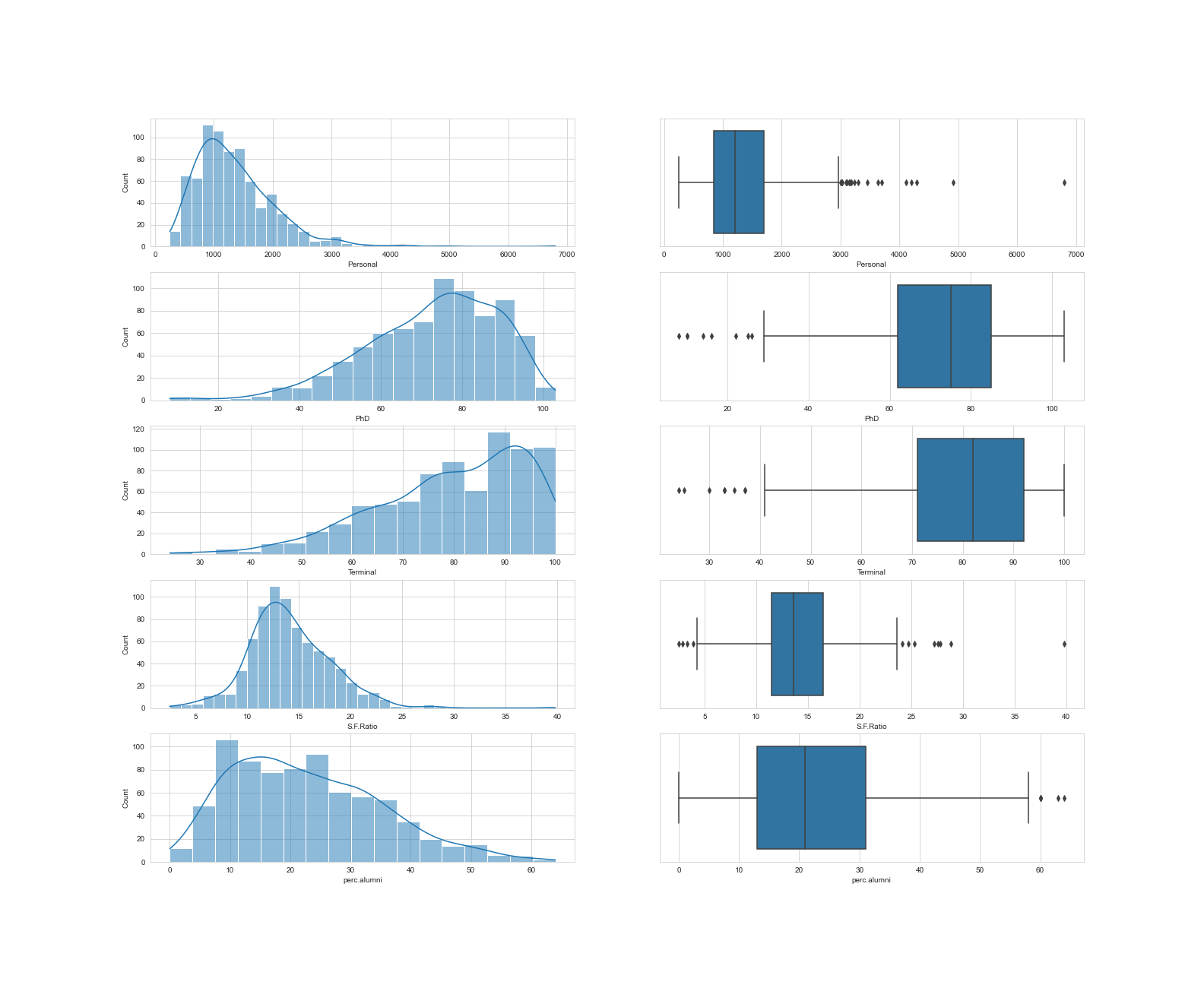


In the above graph we perform univariate analysis on 5 numeric variables:

1. **Top25Perc** 
   1. It is almost normally distributed
   2. It has no outliers
   3. Majority of the students are from top 25% of higher secondary class.
2. **F.Undergrad**
   1. It is positively skewed
   2. It has outliers
   3. full time graduates studying in all the university are in the range of about 2000 to 4000
3. **P.Undergrad**
   1. It is positively skewed
   2. It has outliers
4. **Outstate**
   1. It is almost normally distributed
   2. It has only one outlier
5. **Room.Board**
   1. It is normally distributed
   2. It has outliers

****In the above graph we perform univariate analysis on another 5 numeric variables:

1. **Apps**
   1. The Distribution of the data is skewed to the right
   2. It has outliers
   3. Application count goes till 50K due to the presence of an outlier
2. **Accept**
   1. The Distribution of the data is skewed to the right
   2. It has outliers
   3. accepted application goes till 25k
3. **Enroll**
   1. The Distribution of the data is skewed is skewed to the right
   2. It has outliers
   3. Enrolled application count goes till 6k
4. **Top10Perc**
   1. The Distribution of the data is almost skewed to the right
   2. It has outliers
   3. Intake about 10 to 40 students from top 10 percentage of higher secondary class.
5. **Books**
   1. The distribution seems to be bimodal in nature
   2. It has outliers
   3. The cost of books per student seems to be in the range of 100 to 700.

****

In the above graph we perform univariate analysis on another 5 numeric variables:

1. **Personal**
   1. The Distribution is almost normally distributed with a little skewness on the right
   2. It has outliers
   3. Some student’s personal expense are way bigger than the rest of the students
2. **PhD**
   1. The distribution is almost skewed to the left
   2. It has Outliers
3. **Terminal**
   1. The distribution is skewed to the left
   2. It has outliers
4. **S.F. Ratio**
   1. The distribution is almost normal
   2. It has outliers
   3. student to faculty ratio is in range of 5 to 25 with 12-16 have the highest in colleges
5. **Perc.alumni**
   1. The distribution is almost normal
   2. It has outliers
   3. 12-30% of alumni have highest donation

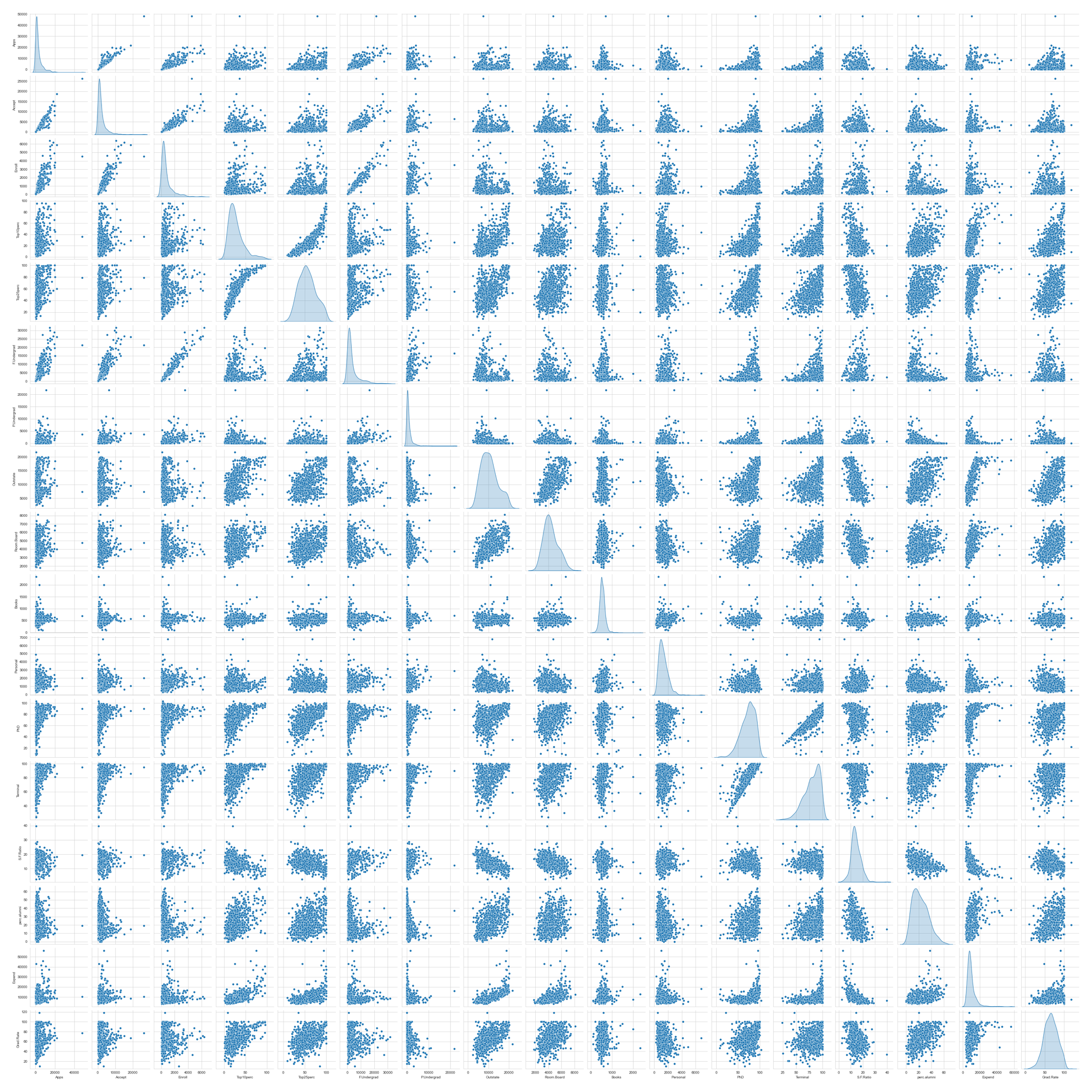
**Chart, histogram

Description automatically generated**

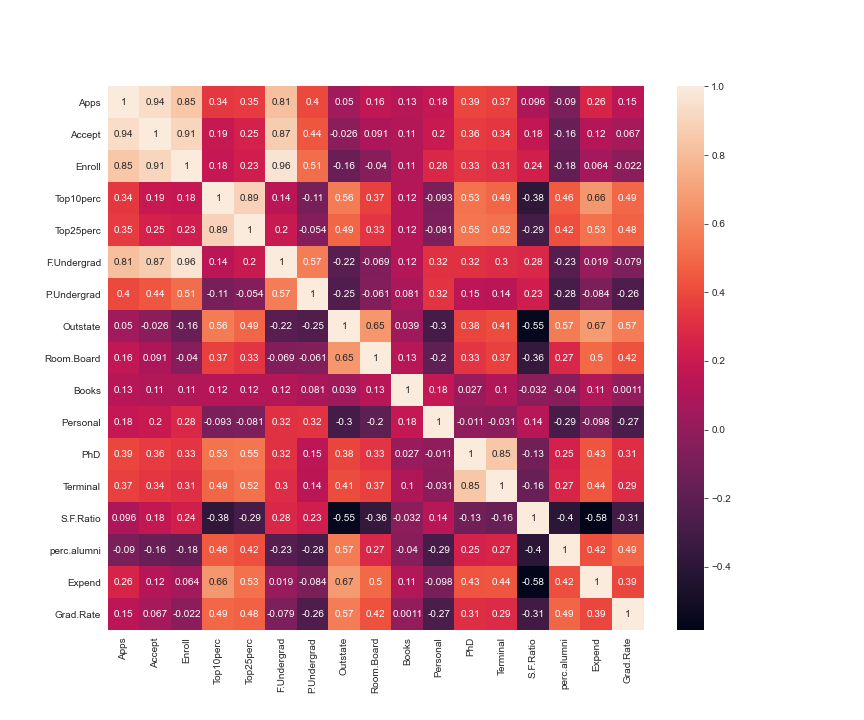
In the above graph we perform univariate analysis on last 2 numeric variables:

1. **Expend**
   1. The distribution is skewed to the right
   2. It has outliers
2. **Grad.Rate**
   1. The distribution is normal
   2. It has outliers (single outlier)
   3. GradRate of most of the colleges goes from 20 to 100% with one outlier.

**Multivariate Analysis**



***The above plot is a pair plot. A pair plot is a pairwise relationships in a dataset. The pair plot function creates a grid of Axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis across a single column.***



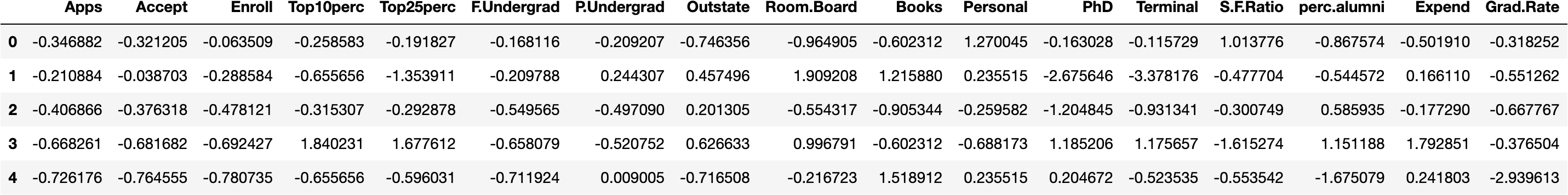
***The above plot is a Heatmap. A heat map helps you visualize density. In the above Heatmap we can observe the following:***

1. The application variable shows high correlation with application accepted.
2. The application variable shows high correlation with enrolled students. The application variable shows high correlation with full time undergraduates. the application variable is positively correlated with variables like application accepted, students enrolled and full time graduates. So this suggests that when student submits the application it is accepted and the student is enrolled as fulltime graduate.
3. The application variable shows negative correlation with percentage of alumni who donate.
4. We observe fairly high level of correlation between percentage of alumni who donate and outstate which suggests that more than half of the alumni who donate belongs to outstate.

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

It is important to perform scaling in PCA. We need scaling because data set might have data with different weights and to compare them to give correct results they should be on the same scale. In order to bring them in the same scale we perform scaling. Feature scaling (also known as data normalization) is the method used to standardize the range of features of data.

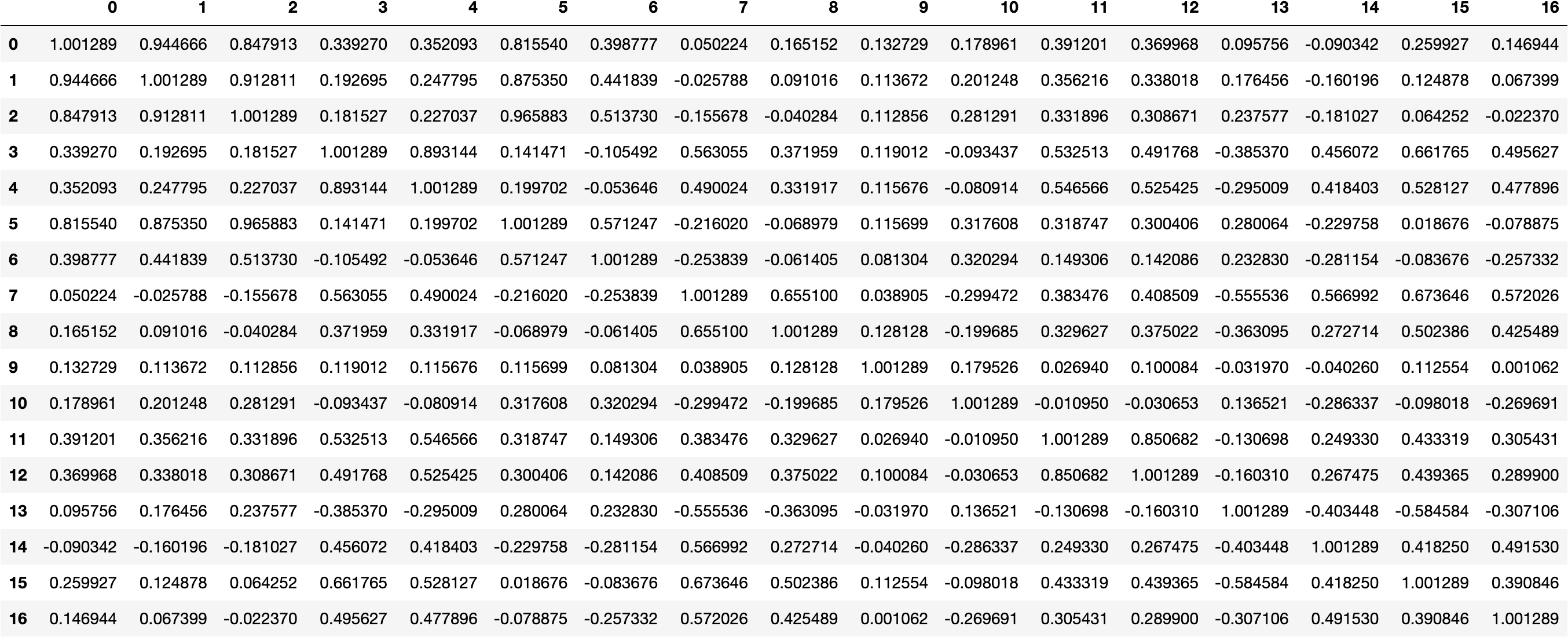
We get the following output, after we perform scaling using Z score.



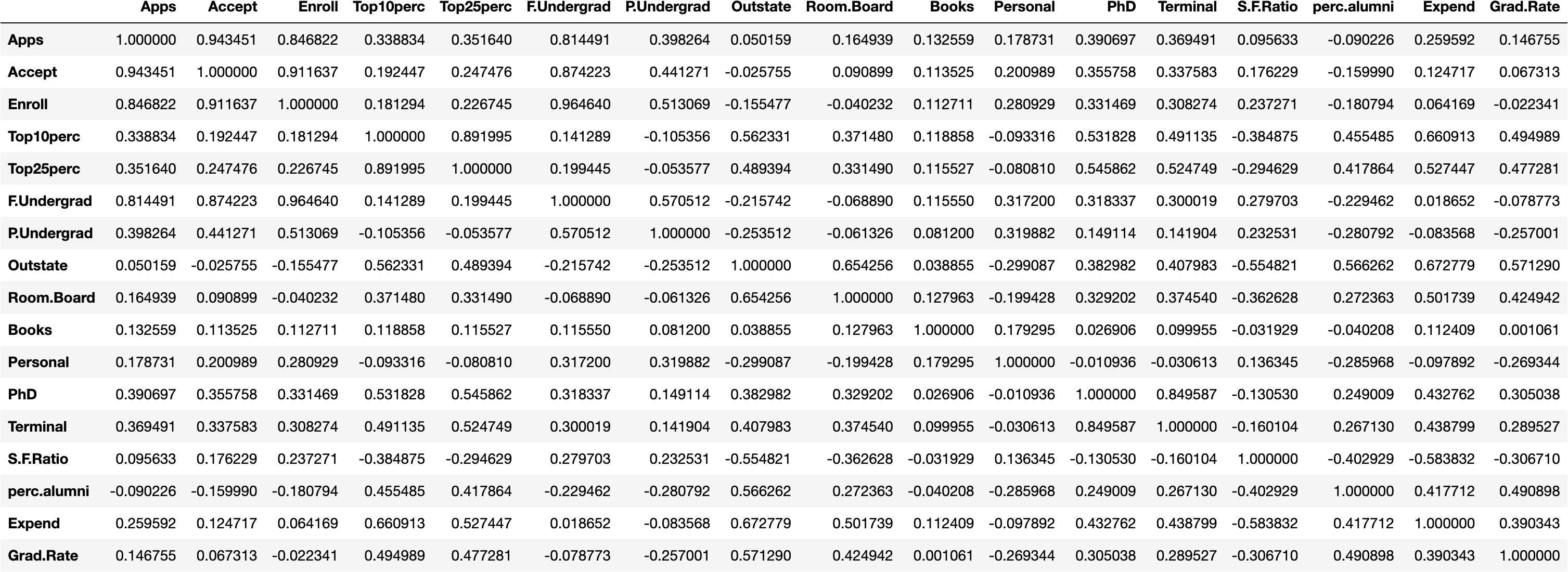
Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].

Covariance is a quantitative measure of the degree to which the deviation of one variable from its mean is related to the deviation of another variable from its mean. Whereas, Correlation tells us both the strength and the direction of this relationship. Correlation is best used for multiple variables that express a linear relationship with one another. When we assume a correlation between two variables, we are essentially deducing that a change in one variable impacts a change in another variable. Correlation is a scaled version of covariance.

**Covariance Matrix**

****

**Correlation Matrix**

****

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]

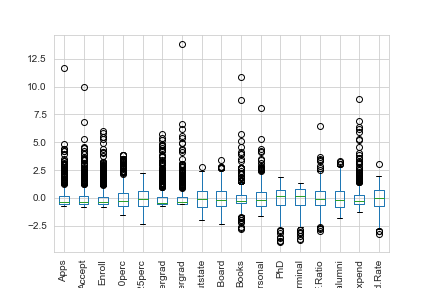
***After plotting boxplot for data set after and before scaling we observe that the standard deviation remains same for all the variables and that the outliers are still present. Therefore we can say that the variables after scaling have been reduced to same weight.***

**Before Scaling**

**Chart

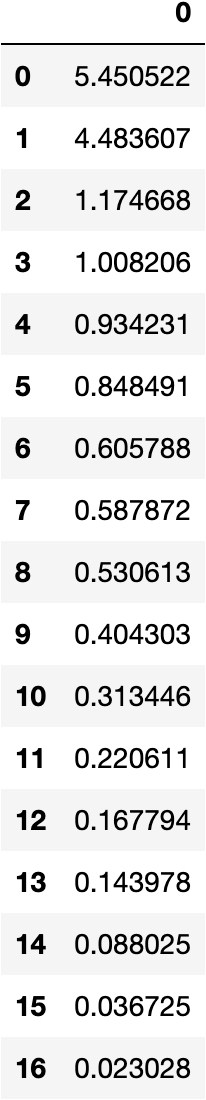
Description automatically generated**

**After Scaling**

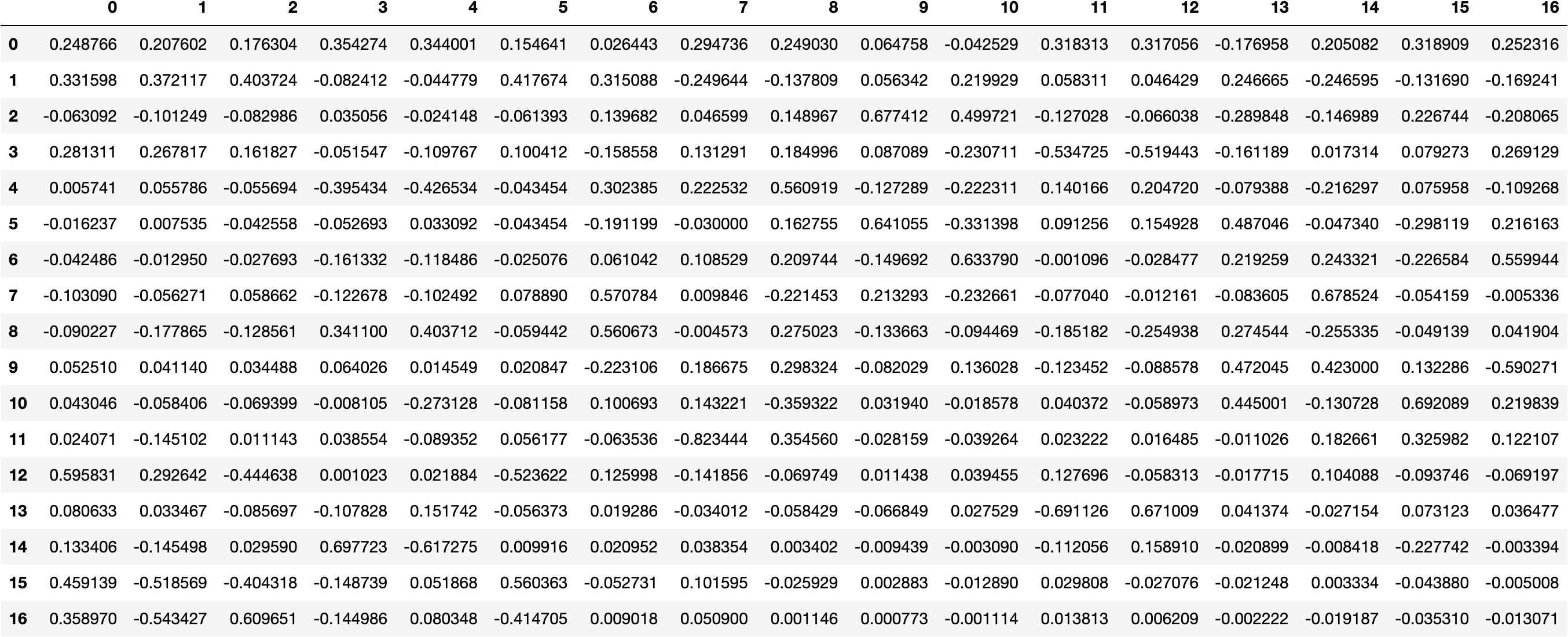
****

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

**Eigen Values**

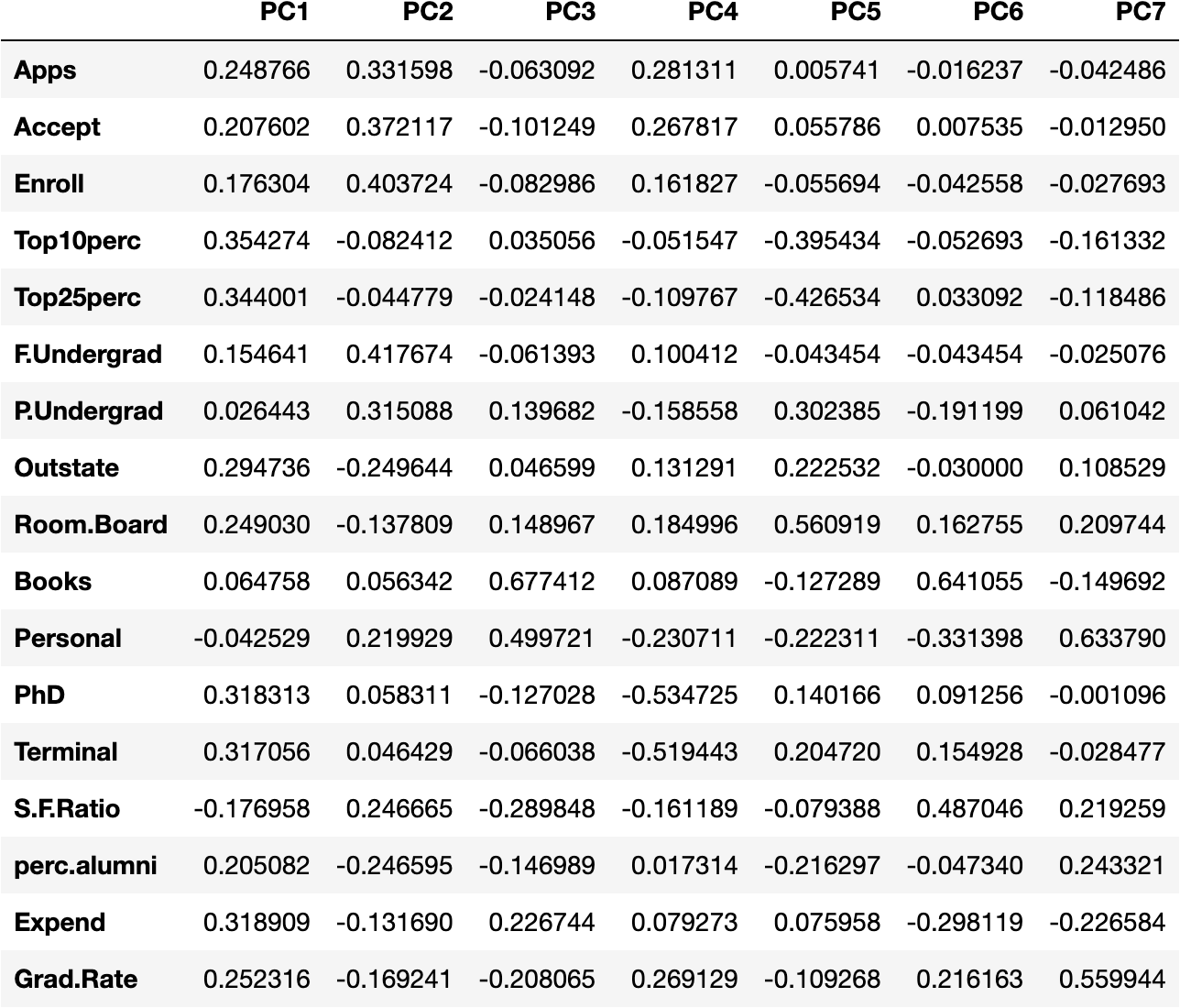
****

**Eigen Vectors**

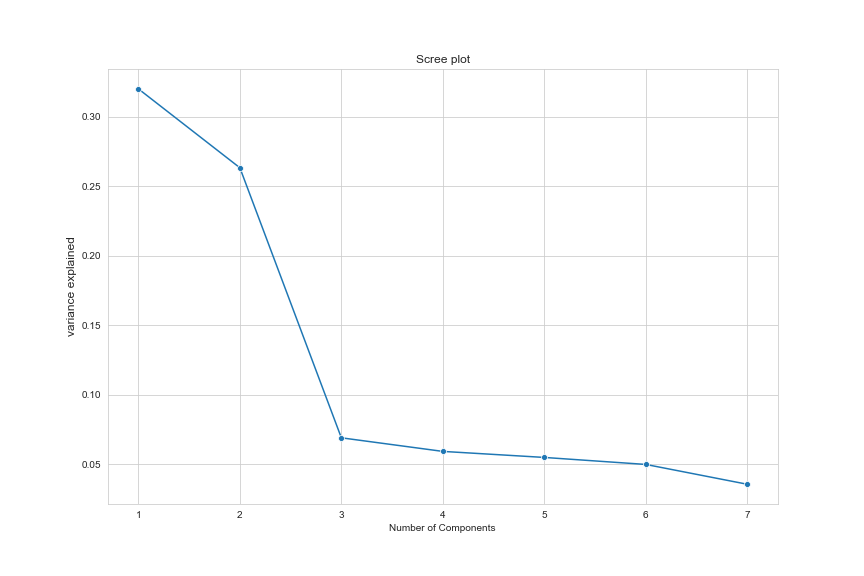
****

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

We apply PCA to all the required features. Then we extract the eigen values and eigen vectors. After that we check the explained variance for each PC. Now, we create a data frame which contains the coefficients of all PC’s.



A scree plot is used to check the cumulative explained variance ratio to find a cut off for selecting the number of PC’s



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 explicit form of the first PC (in terms of the eigenvectors. Using values with two places of decimals only) is given below:

***array([ 0.25, 0.21, 0.18, 0.35, 0.34, 0.15, 0.03, 0.29, 0.25,***

***0.06, -0.04, 0.32, 0.32, -0.18, 0.21, 0.32, 0.25])***

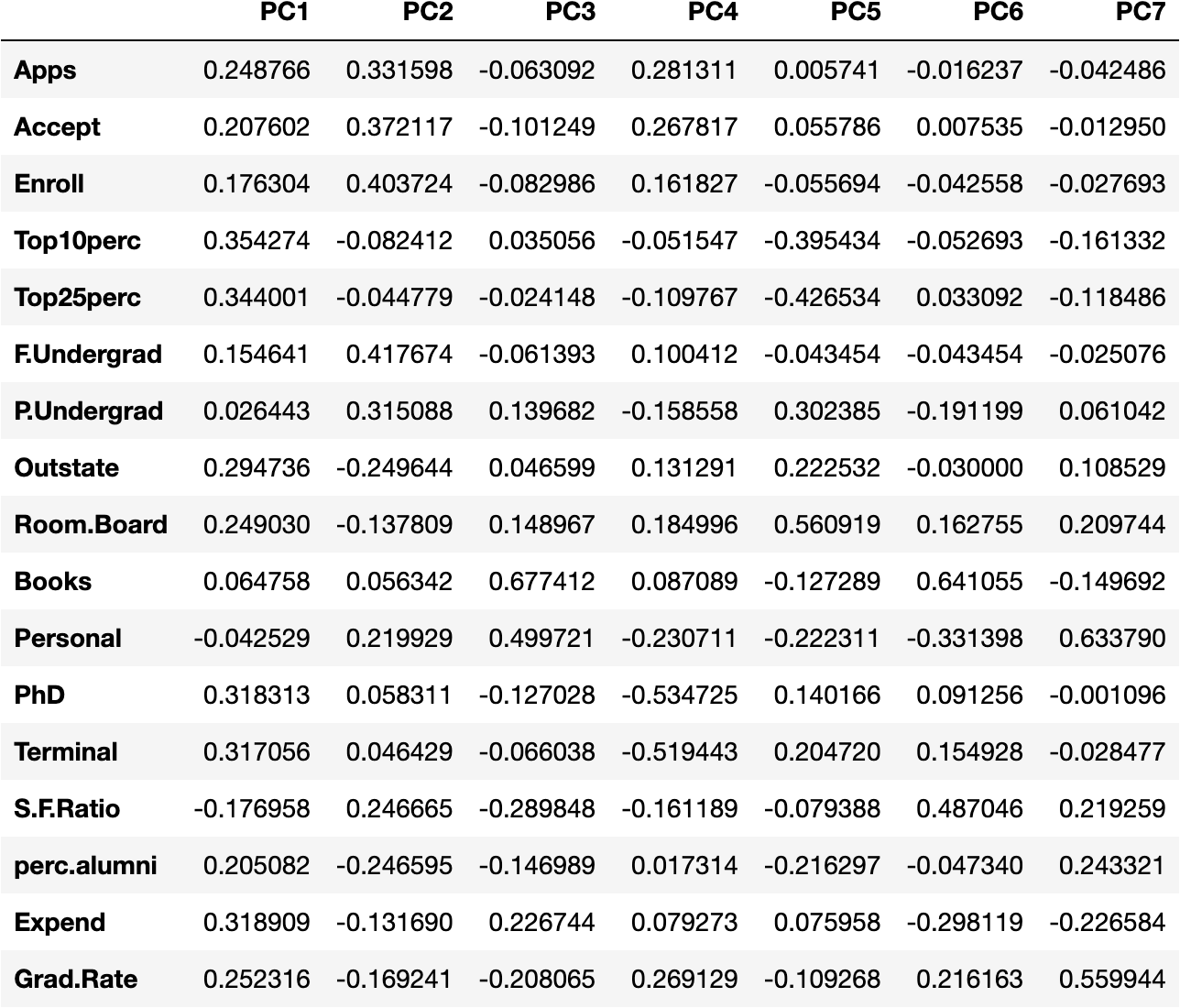
the linear equation of PC in terms of eigenvectors and corresponding features is given below:

***PC1 = .25 \* Apps + 0.22 \* Accept + 0.18 \* Enroll + 0.35 \* Top10perc +0.34 \* Top25perc + 0.15 \* FUndergrad +.03 \* PUndergrad +.29 \* Outstate + 0.25 \* RoomBoard +0.06 \* Books -.04 \* Personal + 0.32 \* PhD +0.32 \* Terminal -0.18 \* SFRatio + 0.21 \* PercAlumni + 0.32 \* Expend +0.25 \* GradRate***

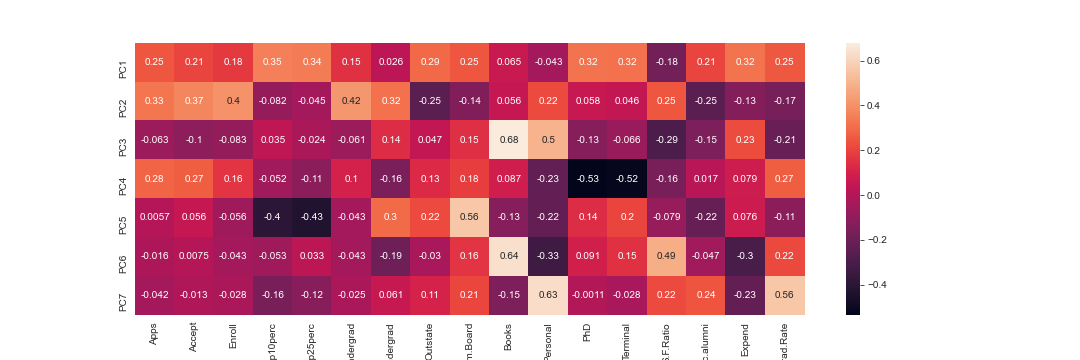
Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?

The Cumulative % gives the percentage of variance accounted for by the n components. We can see that around 7 principal components explained over 85% of the variance. Thus, the optimum number of principal components can be 7. Eigenvectors indicate the direction of the principal components, we can multiply the original data by the eigenvectors to re-orient our data onto the new axes.

***array([0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154, 0.81657854, 0.85216726])***



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]



1. Principal component analysis is a technique for dimension reduction . so it combines input variables in a specific way, to drop the “least important” variables while still retaining the most valuable parts of all of the variables. As an added benefit, each of the “new” variables after PCA are all independent of one another.
2. In this case study, we had 17 numeric variables to be assessed, with PCA we did dimensionality reduction from 17 to 8 (representing more than 90% of the variance). Thus, as far as business implication of using PCA is concerned, in this case, we are reducing a high dimensional space (with 17 variables) and converting it to a lower dimensional space without (theoretically) losing much information.
3. To understand more about the dataset we perform univariate analysis and multivariate analysis which gives us the understanding about the variables.
4. This heatmap and the colour bar basically represent the correlation between the various feature and the principal component. The Principal Component 3 (PC3) shows high correlation with student's expense on Books and Personal category.
5. PC5 is more related to Room.Board and Outstate.
6. PC6 Highlights estimated personal spending for a student and the instructional expenditure per student