**ADVANCED STATISTICS PROJECT**



**ANOVA | EDA | PCA**

**Gaurav Sood**

**PGP-DSBA Online July 2021**

**Date: 17th October, 2021**

Table Of Contents

1. Problem Statement: ............................................................................................................4

1.1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually……………………………………………………………….…4

1.2. 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…………………………………....5

1.3. 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……………………………………5

1.4. 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)…………………………………………...…5

1.5. 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]…………………………………………………………………….6

1.6. 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?...............................................................7

1.7. Explain the business implications of performing ANOVA for this particular case study………8

2. Problem Statement: ……………………………………………………………………………………………………………………9

2.1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?....................................................................9

2.2. Is scaling necessary for PCA in this case? Give justification and perform scaling…………..18

2.3. Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data]………………………………………………………………………………………………………..19

2.4. 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]……………………………………22

2.5. Extract the eigenvalues and eigenvectors.[Using Sklearn PCA Print Both]…………………23

2.6. Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features…………………………………………………………………………27

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]…………………………………………………………………………………28

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?.........................28

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]…………………………………………………………………………………..29

**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.greatlearning.in/courses/53598/files/3876976/download?verifier=3TOGwOxaEzKC2Icqk9KVFb4bGJ8AaMYuGrLYfDDJ&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.]

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

**One-way ANOVA (Education)**

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

Alternate Hypothesis H1: The mean salary is different in at least one level of educational qualification.

**One-way ANOVA (Occupation)**

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

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

Where, Alpha = 0.05

If the p-value is < 0.05, then we reject the null hypothesis.

If the p-value is > or = 0.05, then we fail to reject the null hypothesis.

**1.2. 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.**



Since the p-value (1.257709e-08) is less than the significance level (alpha = 0.05), we can reject the null hypothesis and in turn, conclude that there is indeed a difference in the mean salaries for at least one level of educational qualification.

**1.3. 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.**



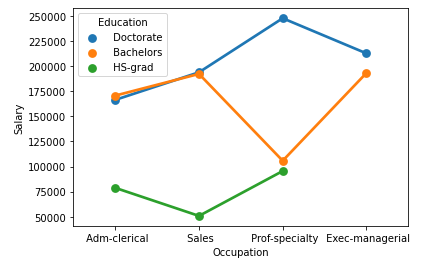
Since the p-value (0.458508) is greater than the significance level (alpha = 0.05), we fail to reject the null hypothesis (thus accepting H0) and conclude that there is no significant difference in the mean salaries across all 4 types of occupation.

**1.4. 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:**

**1.5. 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]**

In order to analyse the effects of one variable on the other, we can make use of an interaction plot.



The interaction plot shows that there is a significant amount of interaction between the categorical variables of Education and Occupation.

Following are some of the observations from the above plot:

* Doctorates generally result in high salaries whereas people with just HS-grad education belong to low income brackets. Meanwhile, people with Bachelors education are in mid-income range.
* People with HS-grad educational level are unable to reach Exec-managerial positions. They are restricted to Adm-clerical, Sales and Prof-specialty occupations and earn less in each of these occupations in comparison to their higher educational level counterparts.
* Strong interaction between people with Bachelors or Doctorate level education and occupation as Adm-clerical and Sales shows them earning similar salaries, possibly hinting at Doctorates being overqualified for such positions.
* People with Bachelor level education and occupation as Prof-specialty are by far the lowest earners within the same educational level, earning barely more than HS-grads in the same occupation.
* On the other hand, people holding Doctorates and occupation Prof-specialty are found to be the highest earners among any occupation across all educational levels.
* People with Doctorates in the occupation Exec-managerial earn slightly more than people with Bachelors in the same occupation.

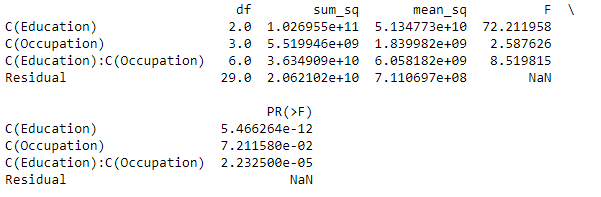
**1.6. Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education\*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?**

**Two-way ANOVA**

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

H1: There is an interaction effect between the independent variable ‘Education’ and the independent variable ‘Occupation’ on the mean Salary.

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



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

As p-value (2.232500e-05) is less than the significance level (alpha = 0.05), we reject the null hypothesis.

Thus, we can say that there is an interaction effect between the two independent variables, Education and Occupation on the mean Salary.

However, the introduction of a few more independent variables could help us further understand to what extent does the interaction between Education and Occupation impact the mean Salary. One such possible independent variable is Work Experience.

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

ANOVA stands for “Analysis of Variance” and is used in statistics when you are testing a hypothesis to understand how different groups respond to each other by making connections between dependent and independent variables.

ANOVA is a statistical test that compares the means of groups in order to determine if there is a difference between them. It is used when more than two groups’ means are compared. For two group means, we can do a simple t-test.

In a business context, ANOVA is used to help manage salary by comparing education to occupation. It can also be used to forecast Salary trends by analysing patterns in data to better understand future hike of Salary.

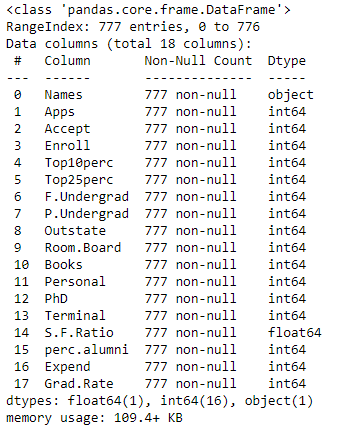
Assuming this report is for an HR Department of an organisation or an HR consulting firm, some key takeaways are as follows:

* An employee or graduate’s salary is heavily dependent upon their level of education as compared to their occupation. Upgrade in the Education level results in a higher salary. On average a Doctorate earns a higher salary a Bachelor or a HS-Grad. Conversely, HS-Grads are by far the lowest earners in the dataset provided.
* Despite occupation’s lesser significance, it does impact salary on some level.
* For some occupations, a higher salary may be awarded to a person holding a Bachelor’s degree in comparison to one with a Doctorate. This indicates the possibility of a Doctorate not being actively sought after for certain roles and instead, being overqualified for the position.
* This brings forth an important shortcoming of the given dataset which reduces the accuracy of tests and analyses performed, namely the absence of a few other essential variables that affect the salary, such as work experience, field of work, specialization in education, etc.

**Problem 2:**

The dataset [Education - Post 12th Standard.csv](https://olympus.greatlearning.in/courses/53598/files/3059973/download?verifier=bI98mccZs5IQWS4cLdXQLlNXyYng6Y0d82IK2Wf3&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.greatlearning.in/courses/53598/files/3059972/download?verifier=vlBaolP8zYRk7aNlGzzJ4syO4W3B0raR562uMvJZ&wrap=1).

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

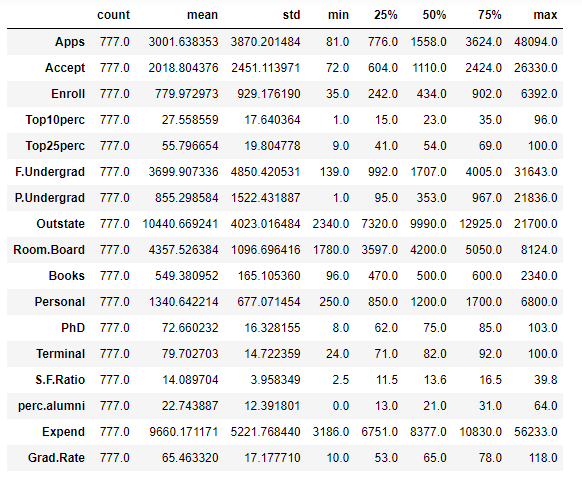
****

* The ‘Education – Post 12th Standard’ dataset has 777 rows with 18 variables in total.
* There are no categorical variables in the dataset.
* Most of the variables are of integer type, except for ‘Names’ which is an object type and ‘S.F. Ratio’ which is of float type.
* There are no missing values in the dataset.
* There are no duplicate data records.

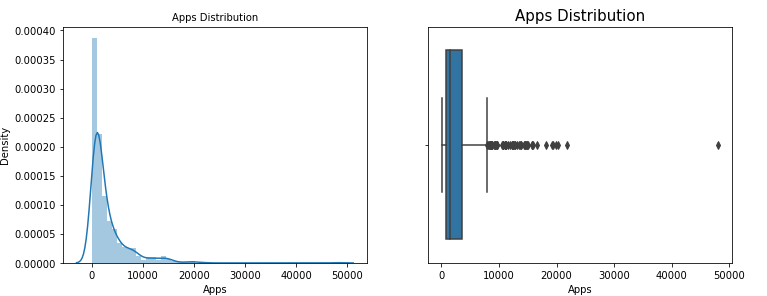
**UNIVARIATE ANALYSIS**

Univariate analysis refers to analysis of a single variable. This helps us to understand the distribution of data in the dataset. With univariate analysis, we can find patterns and summarize the data.

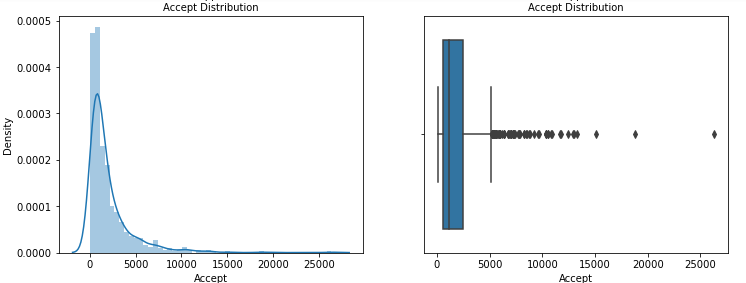
**Statistical Description of Dataset**



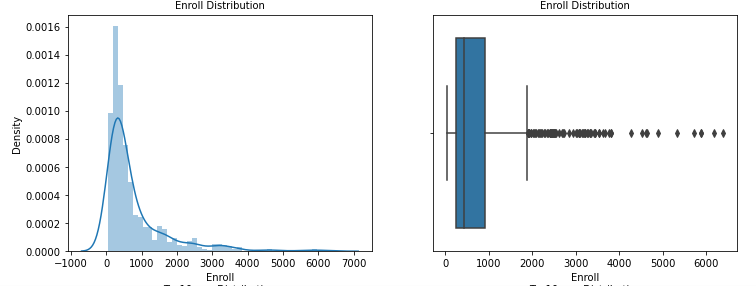
**APPS**



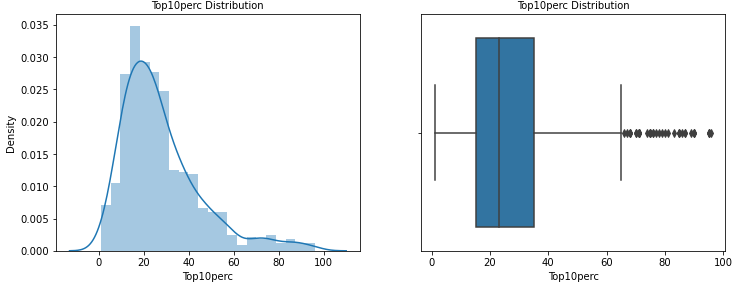
**ACCEPT**

****

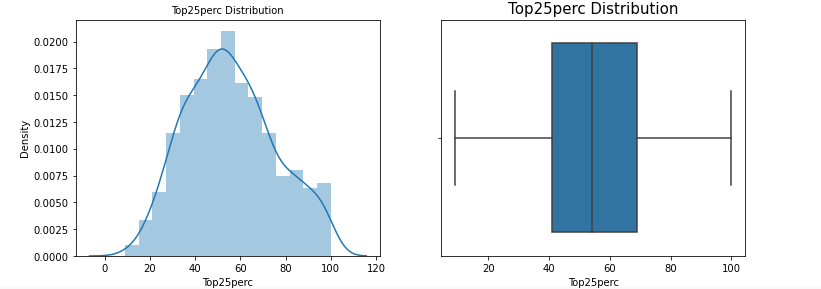
**ENROLL**

****

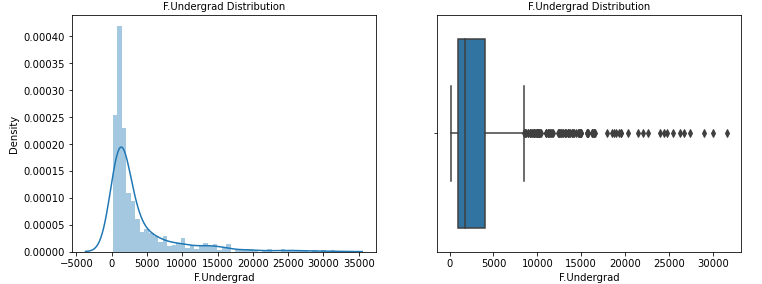
**TOP 10 PERCENT**

****

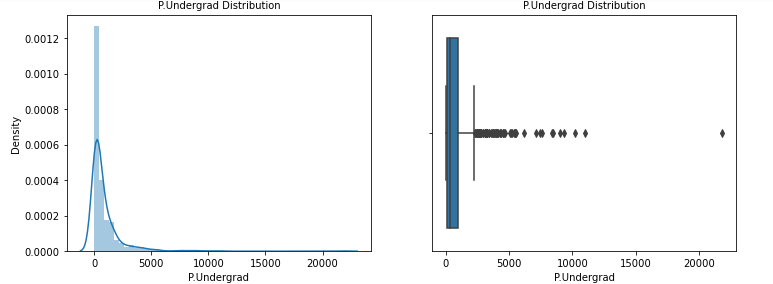
**TOP 25 PERCENT**

****

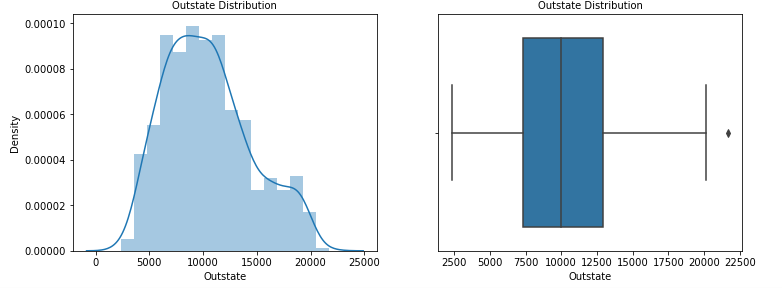
**FULL TIME UNDERGRADUATE**

****

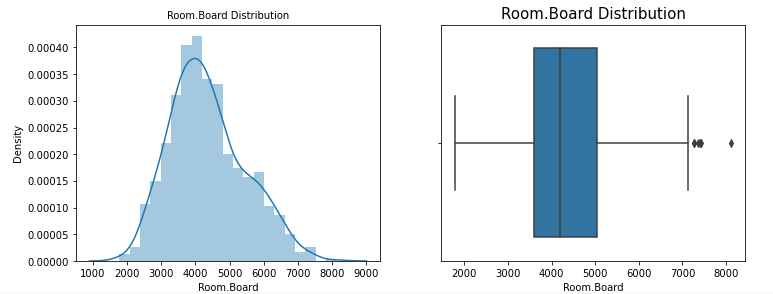
**PART TIME UNDERGRADUATE**

****

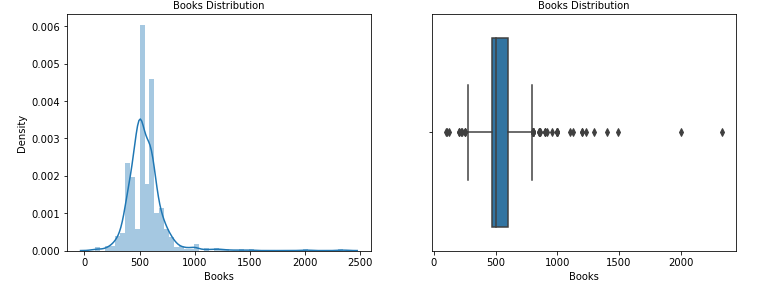
**OUTSTATE**

****

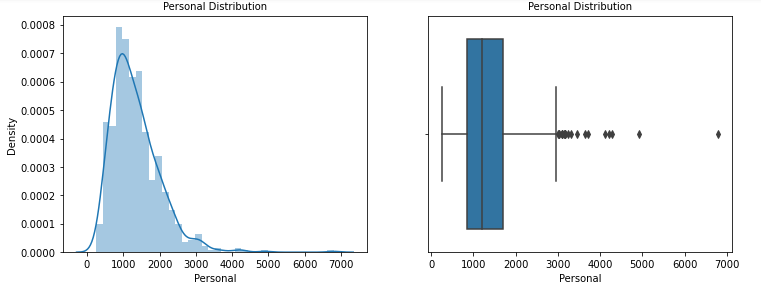
**ROOM BOARD**

****

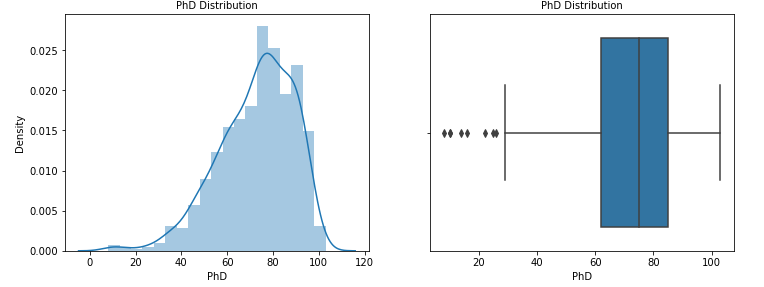
**BOOKS**

****

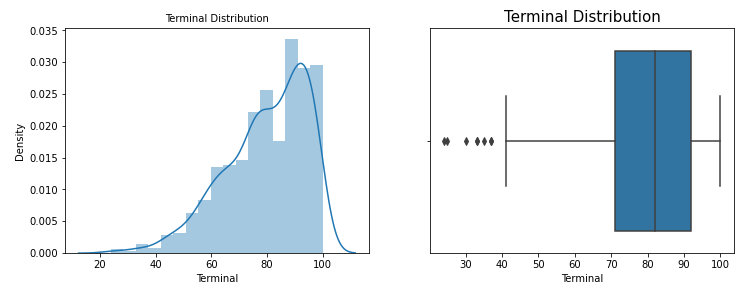
**PERSONAL**

****

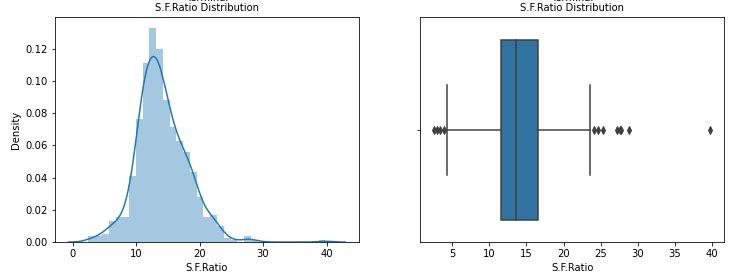
**PHD**

****

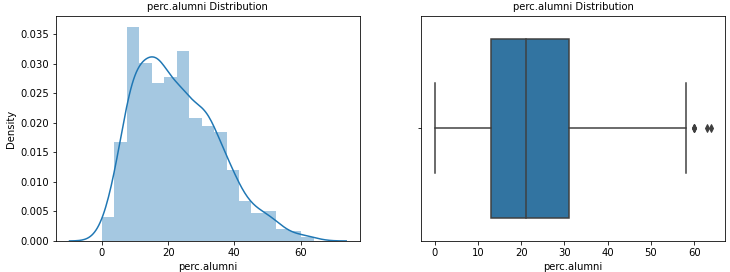
**TERMINAL**

****

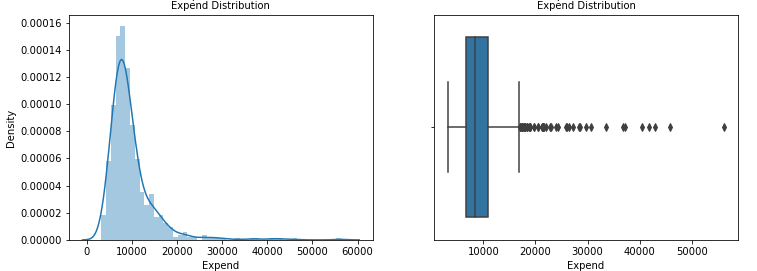
**S.F RATIO**

****

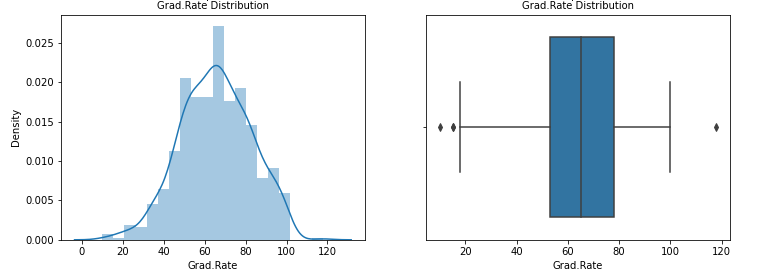
**PERCI ALUMNI**

****

**EXPENDITURE**

****

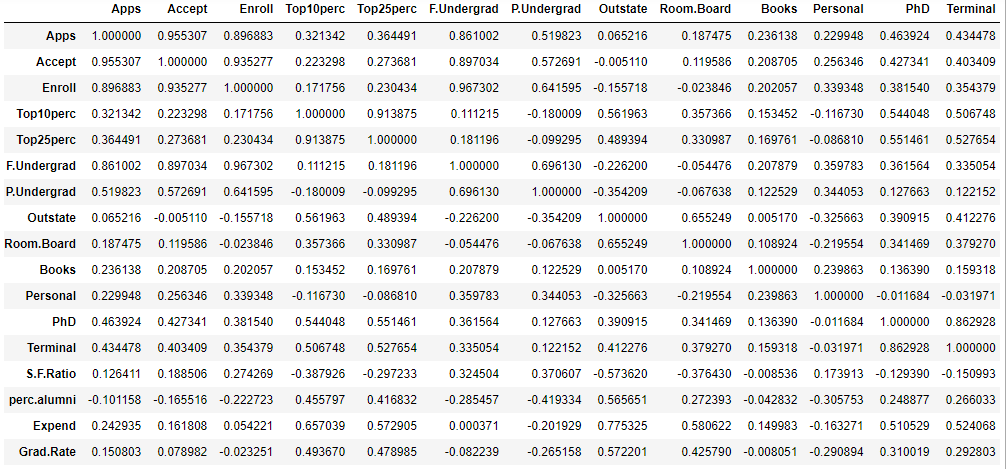
**GRAD RATE**

****

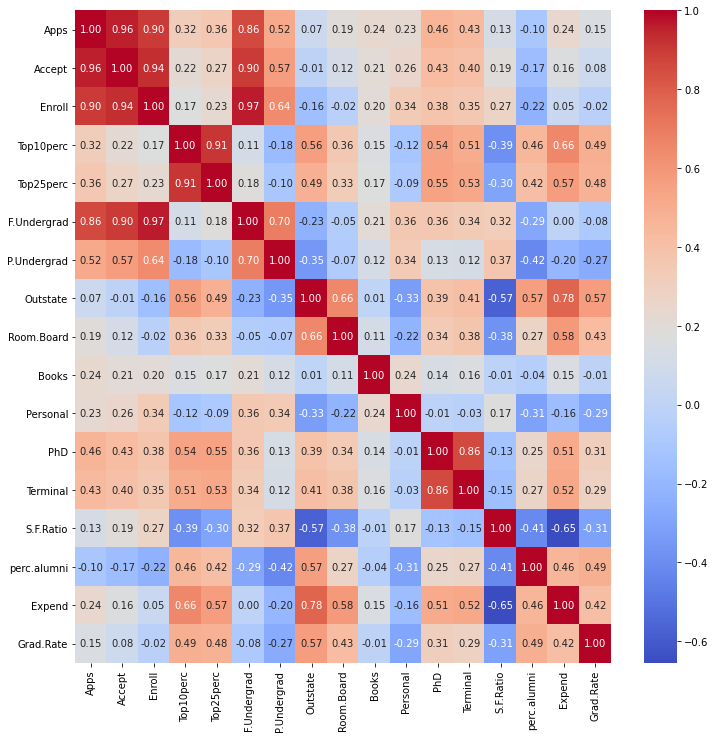
Some valuable inferences from the above graphs are as follows:

* The variables Apps, Accept, Enroll, F Undergrad and P Undergrad are heavily right skewed.
* The variables PhD and Terminal are left skewed.
* Room Board, S.F. Ratio and Grad Rate are about normally distributed.
* The mean percentage of students in a University coming from the best (top 10) schools is 27.5%.
* The mean and median graduation rate is 65%.

**MULTIVARIATE ANALYSIS**

****

**Heatmap showing Correlation Coefficients**

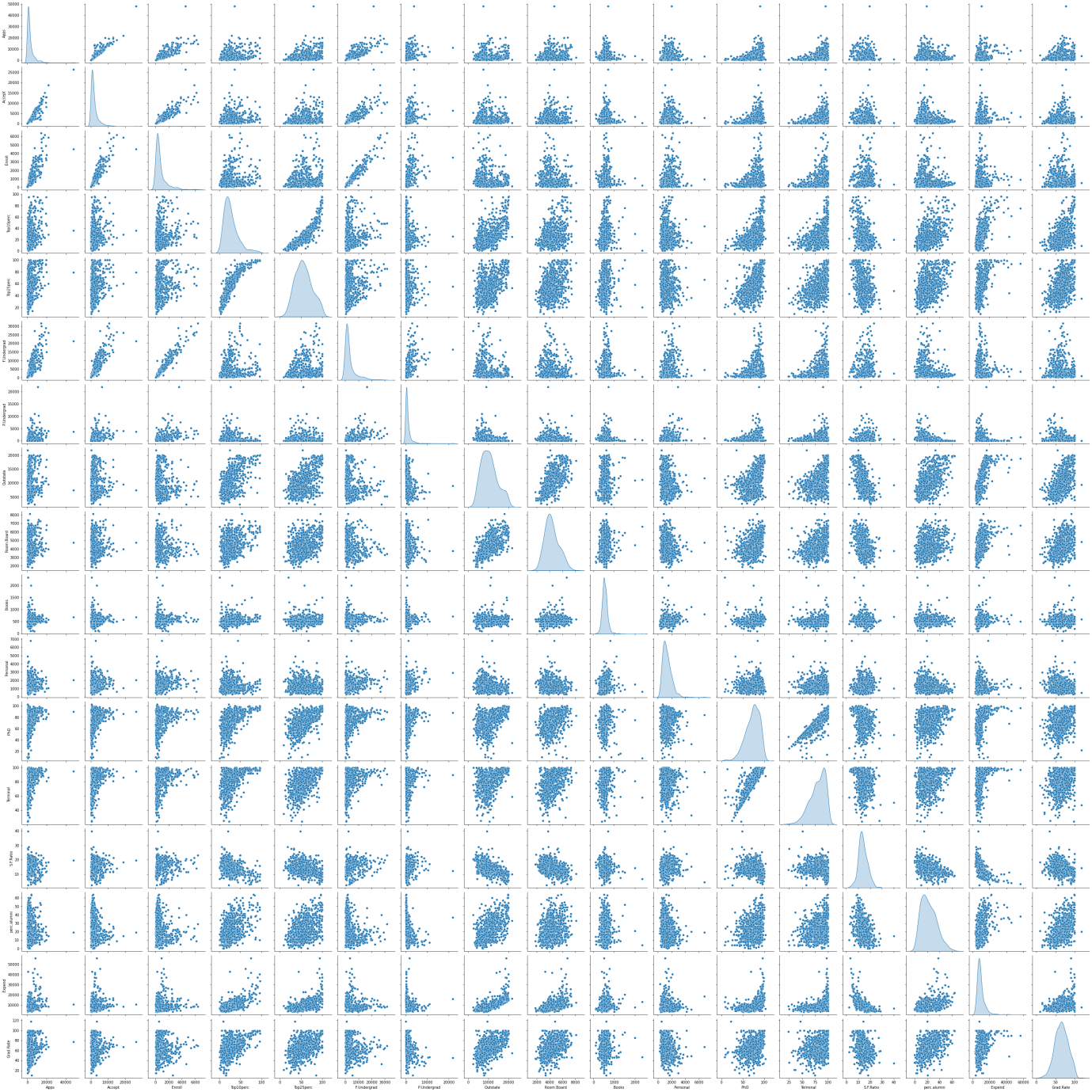
****

A few pairs are found to have high correlation, namely:

* Application and acceptance
* Students from top 10% schools and top 25% schools
* Terminal and PhD qualified faculties
* Enrolment and full-time undergrad students

The heatmap exhibits the problem of multi-collinearity which can be observed with significant number of high correlation pairs of features. This is a problem as it undermines the statistical significance of an independent variable or feature.

**Pair plot to see relationship of variables between each other**

****

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

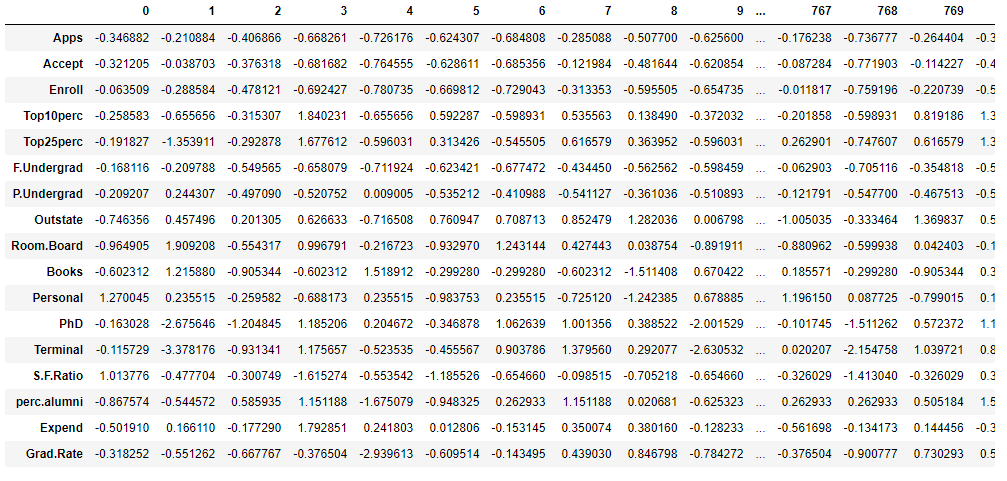
The main objective of scaling or standardization is to normalize the data within a particular range. Since the range of data may vary widely, it becomes a necessary step in data pre-processing.

Before scaling, we drop the ‘Names’ variable which is categorical.

Now, the dataset contains only numerical values. So, we apply z-score method for this case study.

We can also use min-max function to scale the variables.

**Dataset after data has been scaled by applying Z-score**

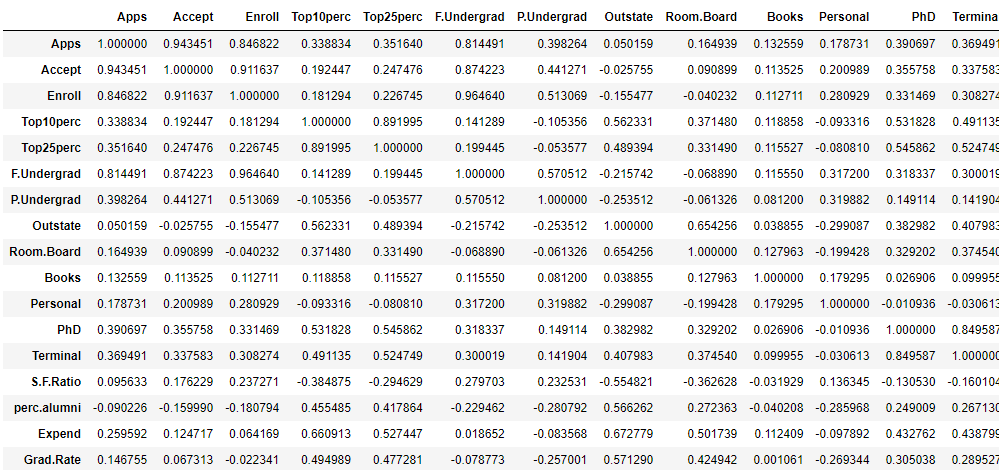
****

Z score tells us how many standard deviation the particular statistic is away from the mean as well as the direction. All variables are scaled using z-score function. Scaling is one of the most important steps to undertake before implementing models.

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

* Both the terms, Correlation and Covariance measure the relationship and dependency between two variables.
* ‘Covariance’ indicates the direction of the linear relationship between variables.
* ‘Correlation’ measures both the strength and direction of the linear relationship between two variables.
* Correlation refers to the scaled form of Covariance.
* The two parameters always have the same sign (positive, negative, or 0). When the sign is positive, the variables are said to be positively correlated; when the sign is negative, the variables are said to be negatively correlated; and when the sign is 0 the variables are said to be uncorrelated.

**Correlation Matrix**

****

**Covariance Matrix**

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]

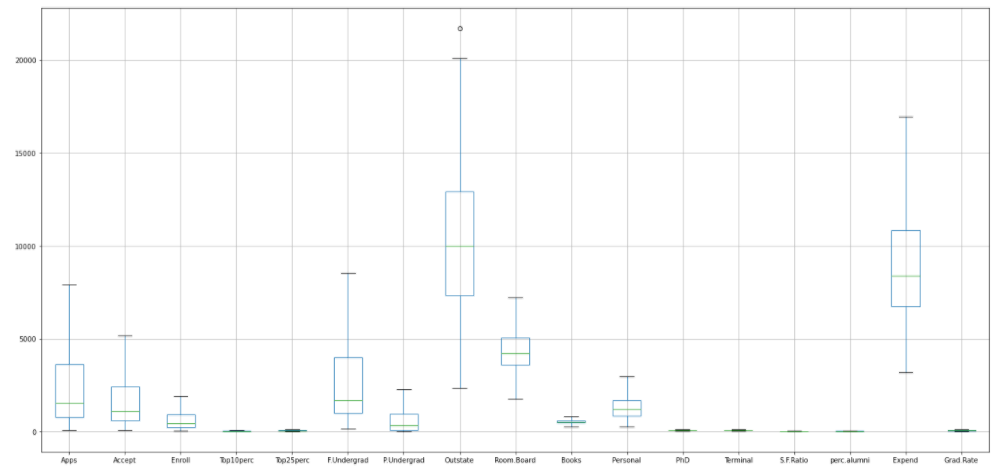
[ 0.14694372 0.06739929 -0.02236983 0.49562711 0.47789622 -0.07887464

-0.25733218 0.57202613 0.42548915 0.00106226 -0.26969106 0.30543094

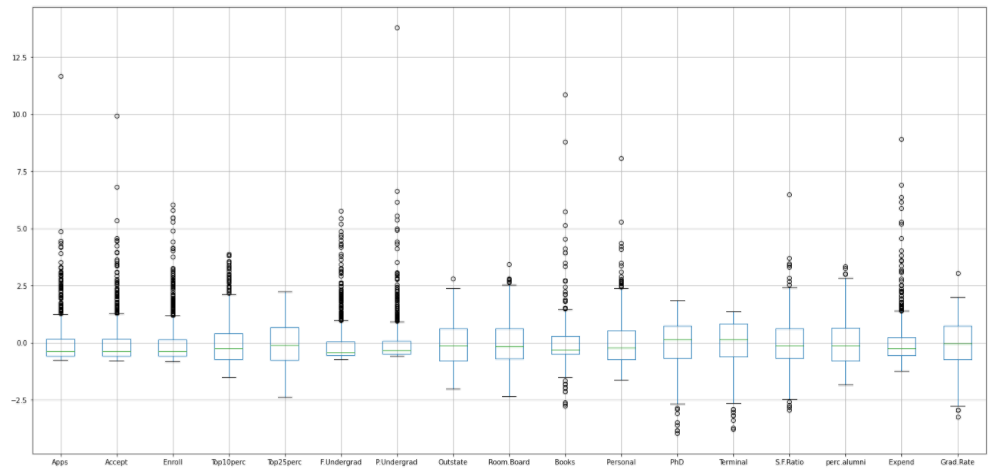
0.28990033 -0.30710565 0.49153016 0.39084571 1.00128866]]

**2.4. 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]**

**Checking the data prior to Scaling**

****

**Checking the data after Scaling**

****

After observing the data before and after scaling, we can see that the outliers are still present in the dataset. We need to take care of outliers separately.

This is because scaling does not remove outliers, rather it just scales the values on a Z-score distribution. We can use any method to remove outliers for further processes.

The scaled data can now be compared easily with each other as all variables are on the same z-score.

The mean of each variable is comparable to 0 and standard deviation is 1.

The range of all variables is standardized and hence, all variables appear unit-less.

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

* Eigenvalues and Eigen vectors are mainly used to capture key information that is stored in a large matrix.
* Eigenvalues are a special set of scalar values that are associated with the set of linear equations in a matrix equation.
* Eigenvectors are non-zero vectors that can be changed at most at its scalar factor after applying linear transformations.
* Eigenvalues and Eigenvectors can improve the efficiency in computationally intensive tasks by reducing the dimensions after making sure that key information is not lost.

**Covariance Matrix**

[[ 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]]

**Eigen Vectors**

[[-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]]

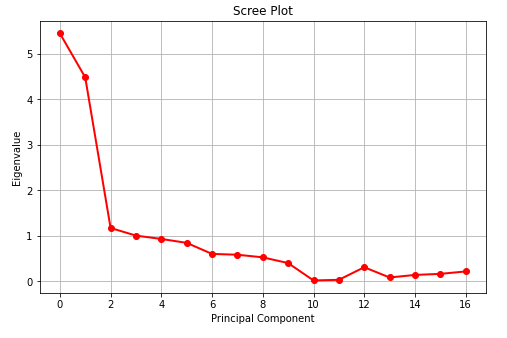
**Eigen Values**

[5.45052162 4.48360686 1.17466761 1.00820573 0.93423123 0.84849117

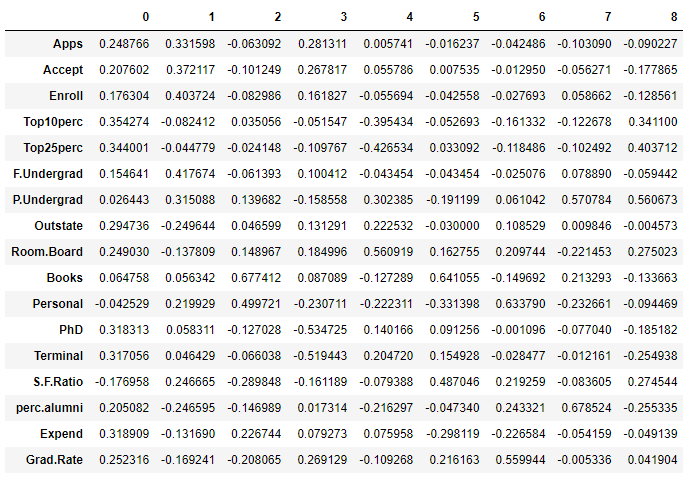
0.6057878 0.58787222 0.53061262 0.4043029 0.02302787 0.03672545

0.31344588 0.08802464 0.1439785 0.16779415 0.22061096]

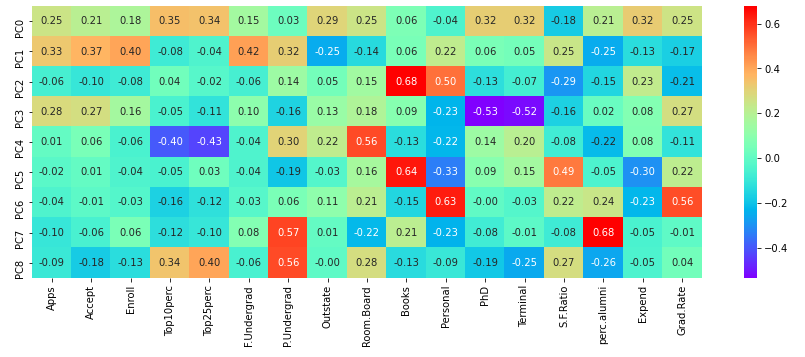
**2.6. Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features**

****

**Principal Score into Data frame**

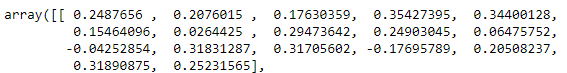
****

**Correlation between components and 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). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]**

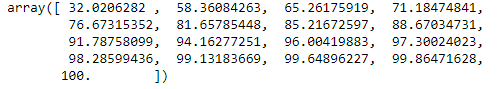
**Explicit form of the First PC:**

****

**The Linear equation of 1st component:**

****

**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?**

****

Adding all the Eigenvalues gives us a sum of 100.

To decide the optimum number of principal components, we check the cumulative variance up to 90%, after which we check the corresponding number associated with 90%.

So, on this basis we decided that the optimum number of principal components to be 9.

The Eigenvectors or Principal Components for this case study are 9, so:

* The first Eigenvalue shows that 32.02% of information is contained by all 17 variables.
* The second Eigenvalue (sum of first and second) shows that 58.36% information is covered.
* And so on and so forth, until over 90% of information is covered by the ninth Eigenvalue.
* Following this, we can overlook the rest as they are less significant in terms of information.

The Eigenvectors (Principal Components) determine the directions of the new feature space, in PCA.

These vectors are essentially multiplication of a vector with a matrix that changes the base of the vector and also its direction in a linear transformation.

**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 (PCA) is a statistical technique used in exploratory data analysis and used for making predictive models. However, the central idea of PCA is to reduce the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the dataset.
* In other words, PCA uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables. It is also a tool to reduce multi-dimensional data into lower dimensions while retaining most of the information.
* PCA can only be done on continuous variables.
* Reduction of dimension requires sacrificing a certain amount of variance. A balance has to be struck so as to achieve a significant reduction in the number of dimensions by sacrificing the least amount of variance.
* For the above case study, the original dataset contained 18 variables. After carefully applying the concepts of PCA we were able to reduce them to just 9 components which will capture 91.78% of the variance in the given dataset.

**THE END**