ADVANCED STATISTICS PROJECT REPORT

ANSWER REPORT

SULOCHANA

**Problem 1:**

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/78180/files/6370335/download?verifier=BM88Y7a8Yw7gDjGGKEGEeuO2Px62yQY8qzZok30z&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.]

**Problem 1A:**

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

**A) Hypothesis for One way ANOVA (Education)**

Null Hypothesis 𝐻0: The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, and HS-Grad).

Alternate Hypothesis 𝐻A: The mean salary is different in at least one category of education.

H0:µD=µB=µH

HA: at least one mean is different from the others

**B) Hypothesis for One way ANOVA (Occupation)**

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

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

H0:µP=µS=µA=µE

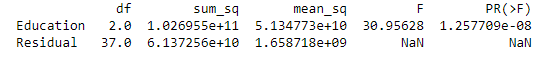
HA: at least one mean is different from the others

**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.**

The level of significance is α = 0.05

If p - value is less than α value, we can reject the null hypothesis. If the p-value is more than level of significance value then we fail to reject null hypothesis.

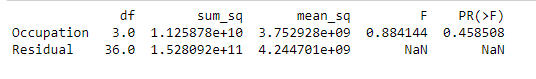
The below is output from the one way ANOVA statistical test.



The above is the ANOVA table for Education variable. Since the p value = **1.257709e-08** is less than the significance level **(alpha = 0.05)**, we can reject the null hypothesis and conclude that there is a significant difference in the mean salaries for at least one category of education.we conclude salary depends on level of education.

**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.**

The below is output from the one way ANOVA statistical test.

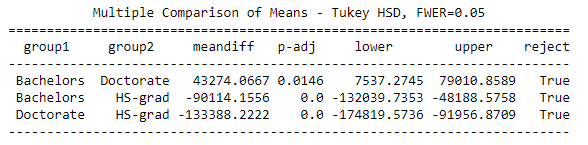


The above is the ANOVA table for Occupation variable. Since the p value = **0.458508** is greater than the significance level **(alpha = 0.05),** we fail to reject the null hypothesis (i.e. we accept H0) and conclude that there is no significant difference in the mean salaries across the 4 categories of occupation.

**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.**

After performing the one way anova for education with respect to the variable salary, we concluded that there is a significant difference in the mean salaries for at least one category of education. Hence we need to find out which educational category mean is different from others mean.

For this, we can perform Tukey HSD test.

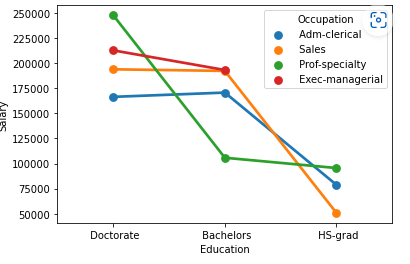


Based on the output it is clear that mean salary is significantly different for each pair of means.

**Problem 1B:**

**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.**

INTERACTION BETWEEN EDUCATION AND OCCUPATION



Interaction plot shows the effect of one variable on the value of another variable. The above graph shows the interaction between the education and occupation. The lines are not parallel in this interaction this means it indicate interaction between occupation and education.

**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?**

**In this case of two way anova education and occupation hypotheses are considered simultaneously.**

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

Alternate Hypothesis 𝐻A: The mean salary is different in at least one category of education.

H0:µD=µB=µH

HA: at least one mean is different from the others

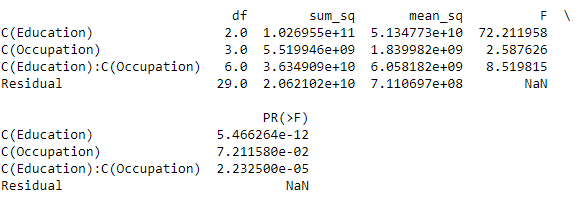
Null Hypothesis 𝐻0: The mean salary is the same across all the 4 categories of occupation (Prof-Specialty, Sales, Adm-clerical, and Exec-Managerial).

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

H0:µP=µS=µA=µE

HA: at least one mean is different from the others

The level of significance is α = 0.05



#### From the table, we see that there is a significant amount of interaction between the variables, Education and Occupation. From the ANOVA method, we see that education combined with occupation results in higher and better salaries among the people.

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

ANOVA test compares more than two means as well as more than two groups at the same time it shows whether the all means are equal or not.

After performing the ANOVA test, the above output indicates the salary only depends on educational qualification but not on the occupation.

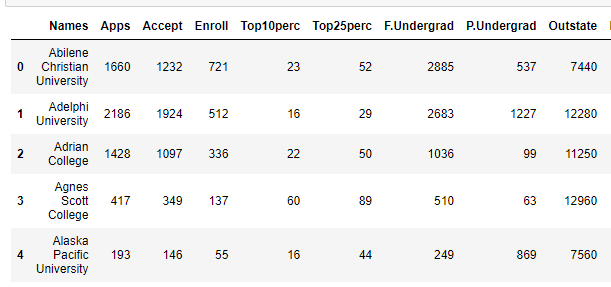
There is a interaction between education and occupation. This means a level of occupation is based on the levels of educational qualification.

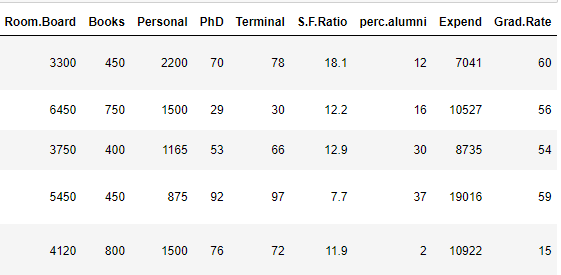
**PROBLEM 2**

The dataset [Education - Post 12th Standard.csv](https://olympus.mygreatlearning.com/courses/78180/files/5510379/download?verifier=V7FXa5qDDiX6kKrYf6QnDsNFLacD0bddD8fB2orc&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/78180/files/5510378/download?verifier=VV7P65zCUC8DHgEGTjlwSDaCH3PcDbJLHkBOqa4S&wrap=1).

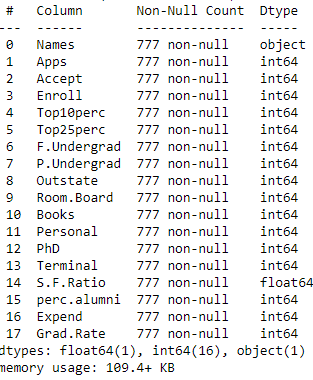
**Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?**

**Data head:**





**Data info:**

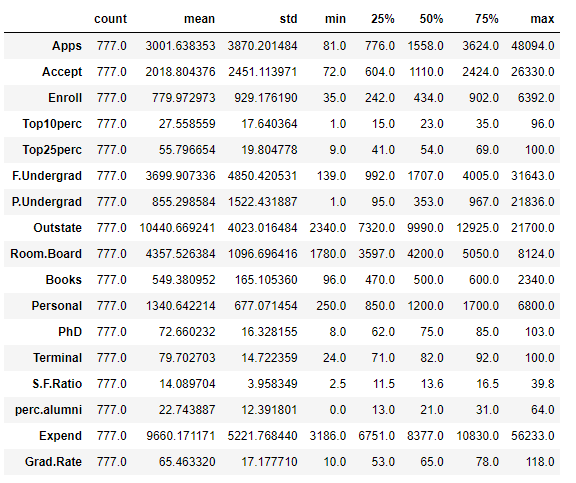


Based on above output there is no null values in data set.

**Data shape:**

(777, 18)

**Data describe:**

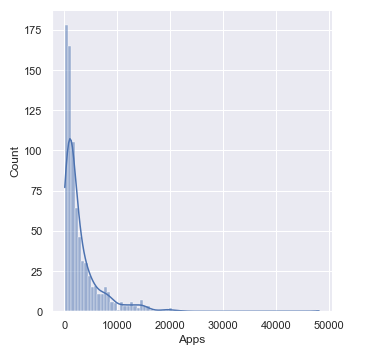


The above table shows 5 point summary of the data. we can see the mean of applications is 3001. So U.S universities received 3001 applications, 2020 applications are accepted and 780 students are enrolled.

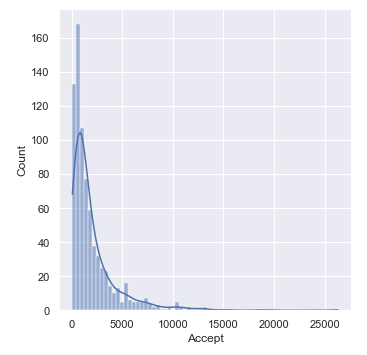
Here we can observe, the average number of full time undergrad students is 3700 and part time undergrad students are 850.

**Univariate Analysis:**

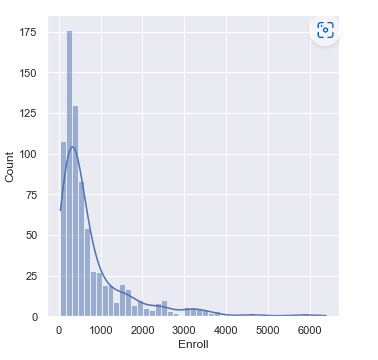
Apps:



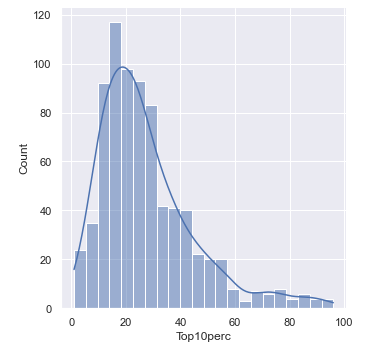
Accept:



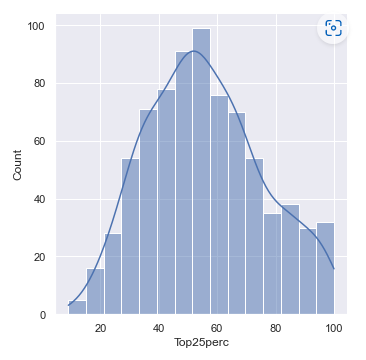
Enroll:



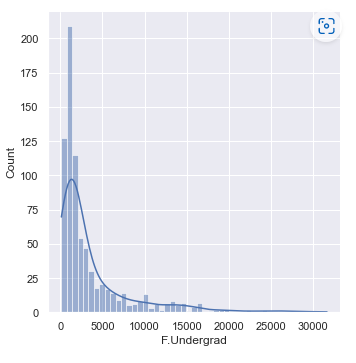
Top10perc:



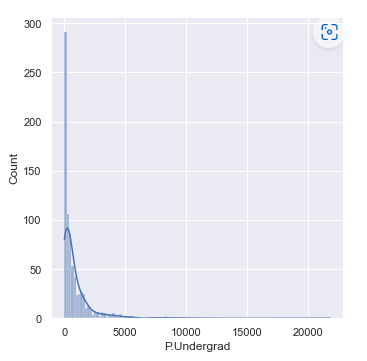
Top25perc:



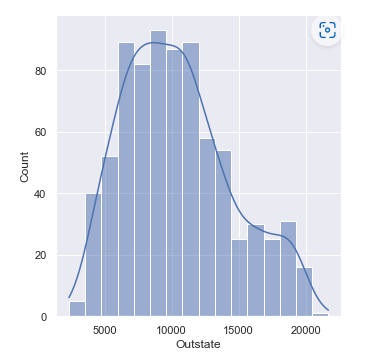
F.undergrad:



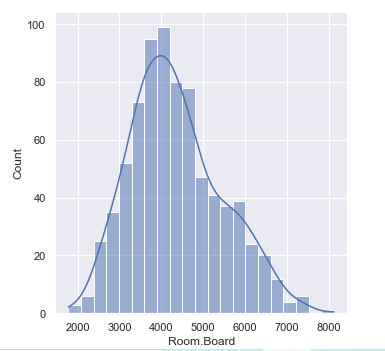
P. undergrad:



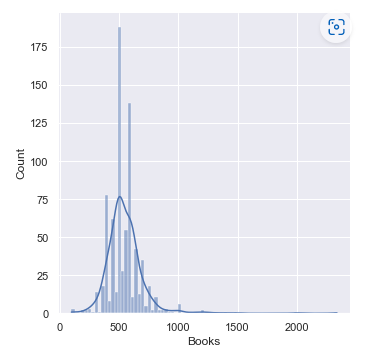
Outstate:



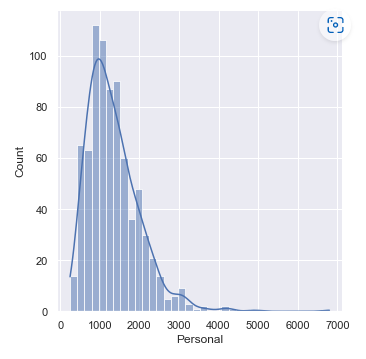
Room..board:



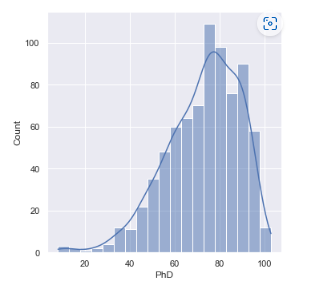
Books:



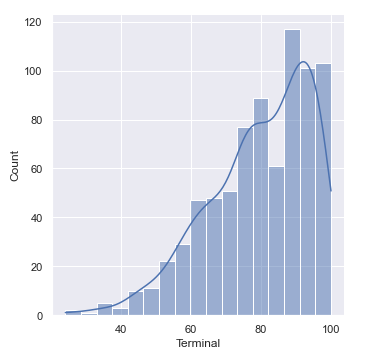
Personal:



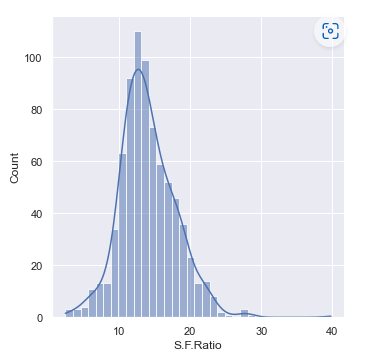
phD



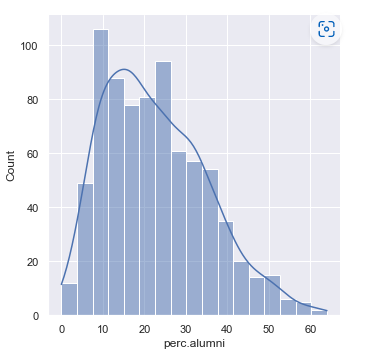
Terminal:



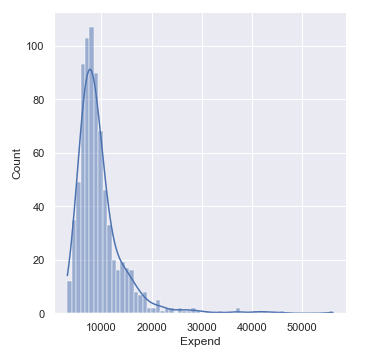
S.F.Ratio:



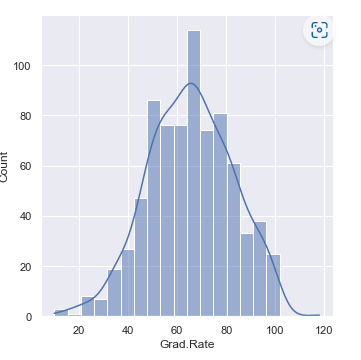
Perc.alumni:



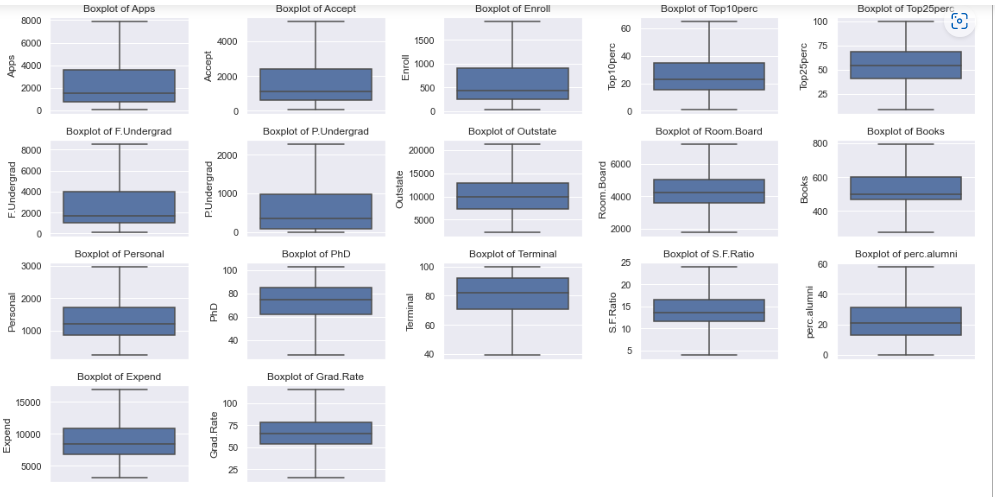
Expend:



Grad Ratio:



Box plots are used to identify the outliers in the data.

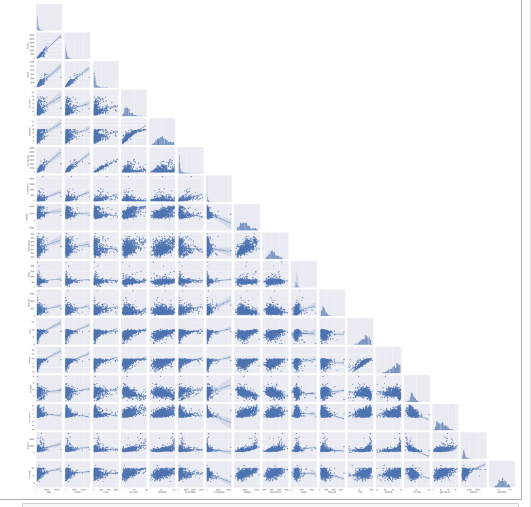


The distplots and boxplots show that apps, accept, enroll, top10perc, F.undergrad, p.undergrad, books, personal, expend variables are highly skewed and all these variables have outliers. Top25perc is the only variable which does not have outlier.

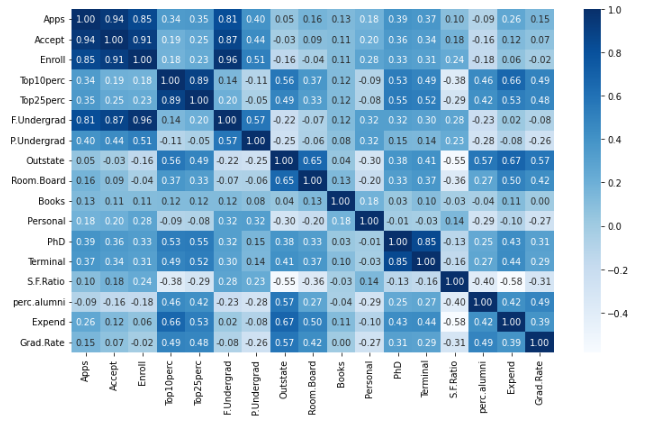
Outstate, room.board, S.F ration, perc.alumni seems to have a moderate right skew.

Phd and terminal variables are moderately left skewed.

Bivariate analysis/ Multivariate analysis:



In the above plot scatter diagrams are plotted for all the numerical columns in the dataset. A scatter plot is a visual representation of the degree of correlation between any two columns.



There is a positive relationship between variables:

Apps and accept

Apps and enroll

Apps and F.undergrad

Accept and enroll

Accept and F.undergrad

Enroll and F.undergrad

Top10perc and top25percent

PhD and Terminal

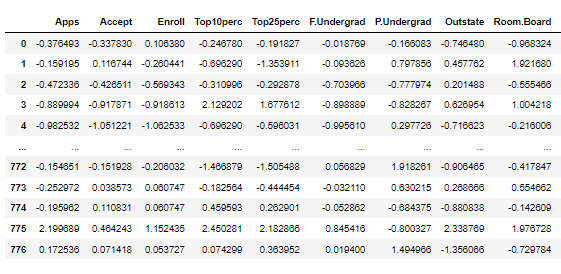
**Is scaling necessary for PCA in this case? Give justification and perform scaling.**

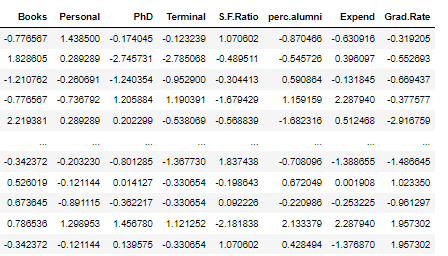
**Yes, it is necessary to normalize data before performing PCA. The pca calculates a new projection of your data set and new axis are based on the standard deviation of variables.**

**The Variables in the data set are different scales i.e. one variable is in thousands and another variable in only two digits. Since the data in these variables are of different scales, it is tough to compare these variables.**

**Feature scaling is the method used to standardize the range of features of data. In this method we convert variables with different scales of measurements into a single scale.**

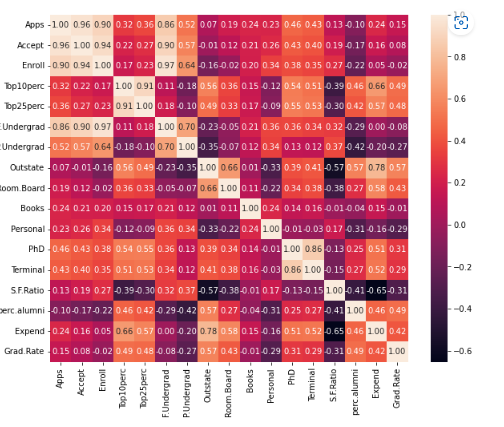
**Standardscaler normalizes the data using the formula (x-mean)/standard deviation.**

****

****

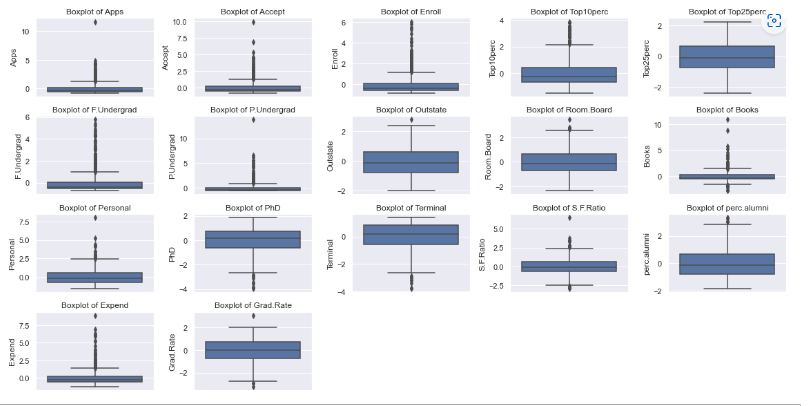
**After scaling if we look at the variables , all has been normalized and scaled in one scale now.**

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

****

From the above images we can say comparison between the covariance and the correlation matrices are same after z score scaling performed.

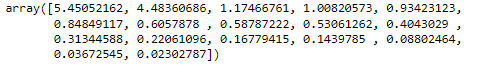
**Check the dataset for outliers before and after scaling. What insight do you derive here?**

****

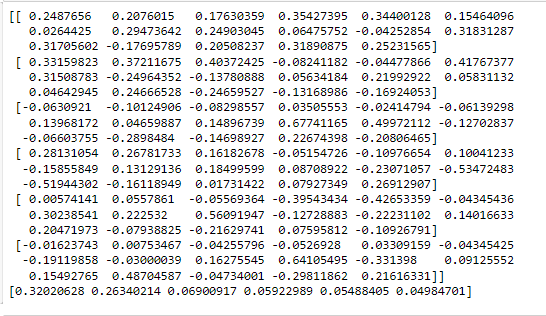
From above boxplots shows outliers in the variables. The scaling of the data does not have any impact on outliers.

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

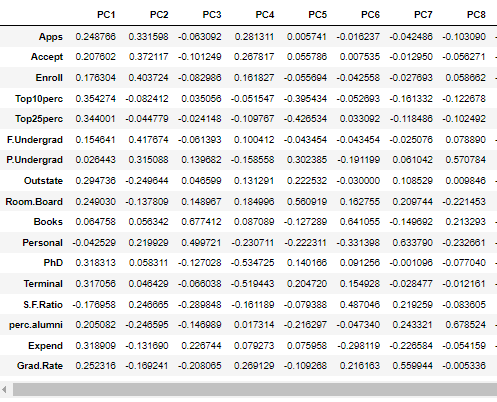
**Eigenvalues extracted by using sklearn**

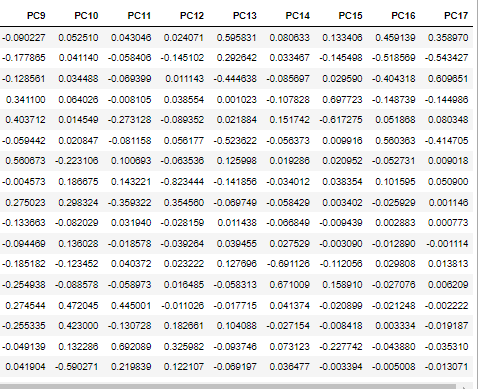
****

**Eigenvectors extracted by using sklearn**

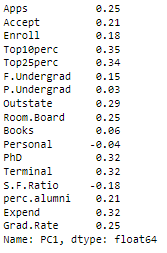
****

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

****

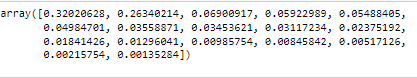
****

**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]**

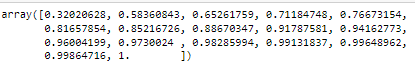
****

PC1: (0.25 x apps)+ (0.21 x accept )+ (0.18 x enroll)+ (0.35 x top10perc)+ (0.34 x top25perc)+ (0.15 x F.undergrad)+ (0.03 x p.undergrad)+ (0.29 x outstate)+ (0.25 x room.board)+ (0.06 x books)+ (0.04 x personel)+ (0.32 x phd)+ (0.32 x terminal)+ (0.18 x S.F ratio)+ (0.21 x perc\_alumni)+ (0.32 x expend)+ (0.25 x Grad\_rate)

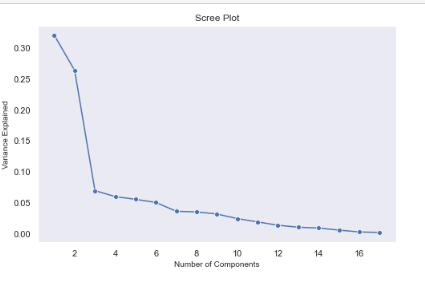
**Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?**

****

The PC1 explained 32% of variance in the data followed by PC2 26% and so on.

****

Above table shows cumulative explained variance.

****

Based on cumulative variance and screen plot we can decide how many PCA components we can choose for further analysis.

**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]**

Interpretation of the principal components is based on which variables are most strongly correlated with each component.

If we discuss about PCA components the first component covers 32% of the data variability. So it covers most of the variables in the data set followed by pca second, third and four components.

**Interpretations of the principal components:**

**PC1**

The first principal component measures apps, top10perc, top25perc, terminal and PhD. These variables are positively correlated and measures high scores.

**PC2**

The second principal component measures accept, enrollment, F.undergrad, P. undergrad, outstate, perc.alomni. It measure high scores in enrollment and F.undergrad.

**PC3**

The third principal component measures books and personal spending of the student.

**PC4**

The fourth principal component measures the cost of boarding and personal spending.