**Business Report on**

***Advanced Statistics***

***Submitted to***



**Great Learning Olympus**

***Prepared by***

**Mrinmoy Majumdar**

**June-D Batch**

**Post Graduate Program in Data Science & Business Analytics**

**From**

****

UT Austin

**September, 2021**

TABLE OF CONTENT

**Problem 1A ………………………………………………………4**

**1.** **S**tate the null and the alternate hypothesis for conducting one-way

ANOVA for both Education and Occupation individually……………6

**2. P**erform a one-way ANOVA on Salary with respect to Education.

State whether the null hypothesis is accepted or rejected based on

the ANOVA results……………………………………………………6

**3. P**erform a one-way ANOVA on Salary with respect to Occupation.

State whether the null hypothesis is accepted or rejected based on

the ANOVA results……………………………………………………6

**4. I**f the null hypothesis is rejected in either (2) or in (3), find out which

class means are significantly different. Interpret the result……………7

**Problem 1B ……………………………………………………… 8**

**1. W**hat 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……………………………………………………..8

**2. P**erform 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? ..........................9

**3. E**xplain the business implications of performing ANOVA for this

particular case study…………………………………………………10

**Problem 2……………………………………………………….11**

**1. P**erform Exploratory Data Analysis [both univariate and multi Variate analysis to be performed]. What insight do you draw from the EDA? …………………………………………………………………………… 14

**2.** **I**s scaling necessary for PCA in this case? Give justification and

perform scaling? ……………………………………………………18

**3. C**omment on the comparison between the covariance and the

correlation matrices from this data [on scaled data]. ……………….20

**4. C**heck 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

**5. E**xtract the eigenvalues and eigenvectors…………………………..22

**6. P**erform PCA and export the data of the Principal Component

(Eigen vectors) into a data frame with the original

features……………………………………………………………..24

**7. W**rite down the explicit form of the first PC (in terms of the

Eigenvectors. Use values with two places of decimals

only)………………………………………………………………28

**8. C**onsider 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…………………………………29

**9. E**xplain the business implication of using the Principal Component

Analysis for this case study. How may PCs help in the

further analysis ….………………………………………………….29

**Problem 1A**

**Problem Statement:**

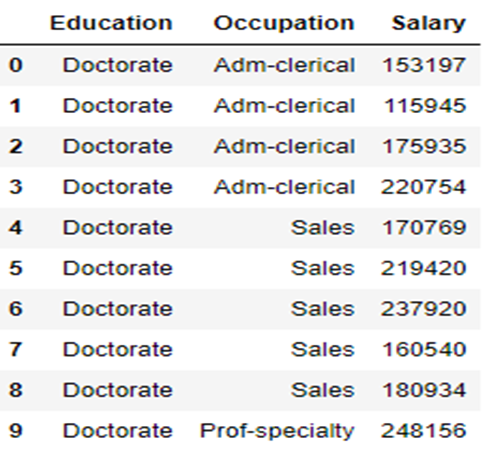
Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csv] are collected and each person’s educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

[Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.]

# Attribute Information:

* Salary: Earnings of 40 different individuals
* Education: Educational qualifications of individuals
* Occupation: Profession of individuals

Displaying Salary Data:

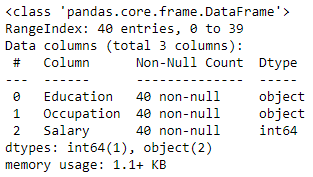
****

**Table 1: Top 10 rows of salary data Frame**

## Basic EDA:

* Checking shape and information of data Frame

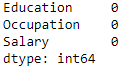
(40,3) – The data set contains 40 observations of data and 3 variables.



**Image 1: Information on salary dataset**

### The data has 40 instances with 3 attributes. 1 integer type and 2 object type (Strings in the column).

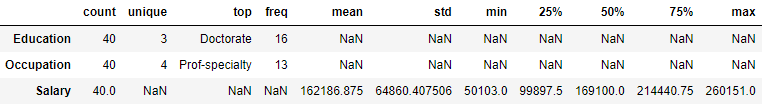
* Check the presence of missing values



**Image 2: Checking null values in data**

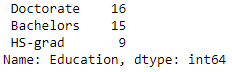
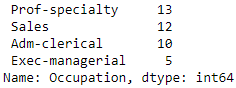
### There are no null values in any of the columns.

* Checking summary of data Frame



**Image 3: Description of salary dataset**

* Checking distinct values of Education and Occupation

**Image 4: Count of education and occupation at different levels**

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

**The hypothesis for conducting one way ANOVA for education:**

## H0: The mean salary of individuals is the same at 3 levels of education.

## Ha: For at least one level of education, mean salary of individuals is different.

**The hypothesis for conducting one way ANOVA for occupation:**

## H0: The mean salary of individuals is the same at 4 levels of occupation.

## Ha: For at least one level of occupation, mean salary of individuals is different.

**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.**

**Now, let us go ahead and perform one way ANOVA with 'Salary' with respect to 'Education'.**



**Image 5: AOV table output of Salary w.r.t Education**

Since the p value is less than the significance level (0.05), we can reject the null hypothesis and states that there is a difference in mean salaries of individuals for at least one level of education.

**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.**

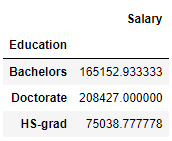
**Let us now perform One Way ANOVA with the variable 'Salary' with respect to 'Occupation'.**



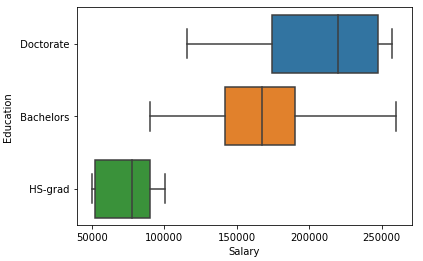
**Image 6: AOV table output of Salary w.r.t Occupation**

**Now, we see that the corresponding p-value is greater than alpha (0.05). Thus, we fail to** reject**the**Null Hypothesis**(**H0**) and states that the mean salary of individuals is same at 4 levels of occupation.**

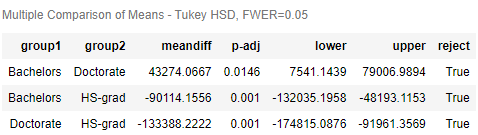
**4. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result.**



**Image 7: Mean salary for different education levels**



**Image 8: Boxplot of Salary vs Education**



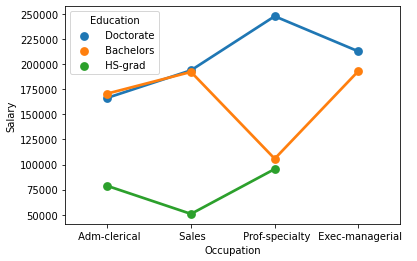
**Image 9: Tukey test output**

From the above results, it is clear that mean salary of individuals with HS-grad education level is significantly different from other levels of education.

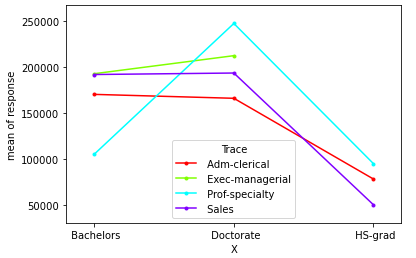
**Problem 1B**

**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.**

**Let us check whether there is any interaction effect between the treatments.**



**Image 10: Point plot of Occupation vs Salary**



**Image 11: Interaction plot of Education vs Salary**

**Still, we can see that there is some sort of interaction between the two treatments.**

**Conclusion –**

* **There is a high level of interaction between Administrative-clerical and Sales occupation individuals with Doctorate and Bachelors education levels signifying similar salary packages.**
* **There is no interaction between Doctorate and HS-graduates independent of occupation.**
* Individuals with Professional or specialty as their occupation and a Doctorate level of education have the maximum salary package and high school graduates with Sales as their occupation have minimum salary package.
* With respect to education level, individuals with high school graduate level of education have the least salary and those with a Doctorate earn the most.
* Among individuals with a high school graduate level of education, the ones working in Professional or specialty earn the most, the ones working in Sales earn the least.
* For individuals with a Bachelors level of education, the ones working in Sales and Executive or managerial occupations earn almost identical salary packages.
* Among those with Doctorate level of education, the individuals with Administrative-Clerical level of occupation earn the least and the ones working in Professional-specialty earn the most.

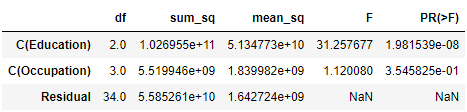
**2. Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education\*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?**

Null and alternative hypothesis for conducting two-way ANOVA based on Salary with respect to both Education and Occupation –

## H0: The mean salary of individuals is the same at all levels of education and occupation.

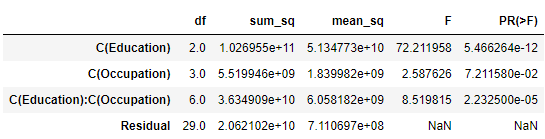
## Ha: For at least one level of education and occupation, mean salary of individuals is different.

**Let us now perform the Two Way ANOVA. We will now analyse the effect of both the treatments on the 'Salary' variable.**



**Image 12: AOV table of Salary w.r.t Education and Occupation without interaction**

**The p-value in both the treatments is less than α (0.05).**



**Image 13: AOV table of Salary w.r.t Education and Occupation with interaction**

**Due to the inclusion of the interaction effect term, we can see a slight change in the p-value of the first two treatments as compared to the Two-Way ANOVA without the interaction effect terms. And we see that the p-value of the interaction effect term of ‘Education' and 'Occupation' suggests that the Null Hypothesis is rejected in this case.**

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

The organisation wants to know whether salary of individuals is dependent on education and occupation levels. ANOVA test helps organisation to conclude that occupation is not a factor that affects salary of individuals rather education contributes to variation in salaries. Education and occupation together has a combined impact on indifferent salaries of individuals. Hence, organisations need to give more priority to individual’s education than occupation while performing their background verification which will ultimately decide their salary packages.

**Problem 2**

**Problem Statement:**

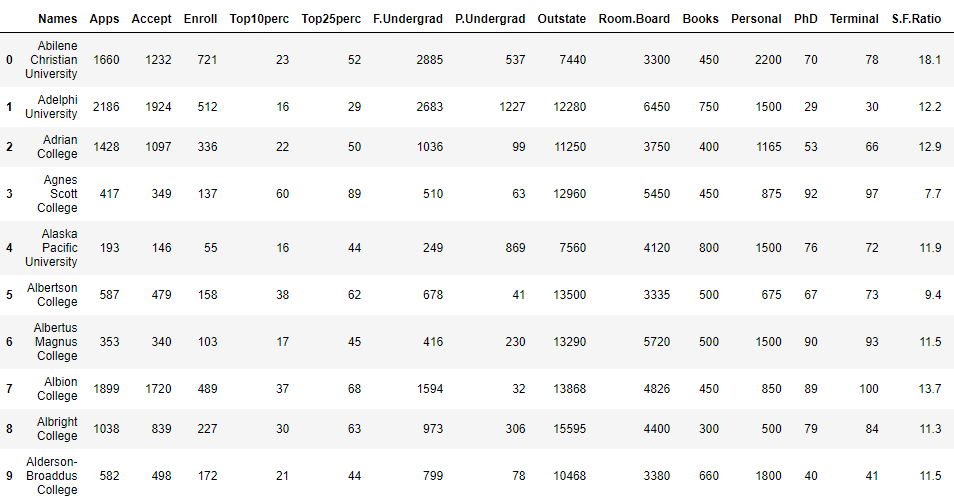
The dataset [Education - Post 12th Standard.csv](https://olympus.greatlearning.in/courses/52662/files/2974420/download?verifier=xpEka1iy4yFudwHr3eTUKDYPk3mSEYotk1MI7yaO&wrap=1) contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according 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/52662/files/2974419/download?verifier=W2bWF0HysNAKVfy3VPb8CyVLZDLsAQwmFJmnJe8X&wrap=1).

Basic Data Exploration

In this step, we will perform the below operations to check what the data set comprises of. We will check the below things:

* head of the dataset
* shape of the dataset
* info of the dataset
* summary of the dataset

Dataset Snapshot –

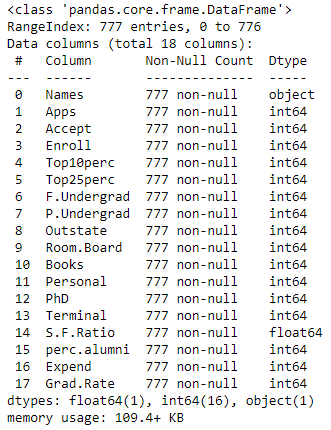


**Image 14: Head of dataset**

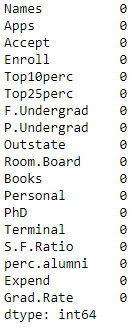
* (777, 18)-The dataset contains 777 observations of data and 18 variables.

### The data has 777 instances with 18 attributes. 16 integer type, 1 float type and 1 object type (Strings in the column).

* There are no null values in the given dataset.

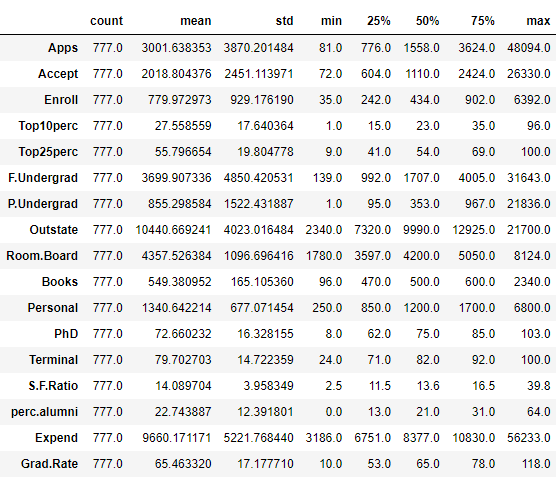


**Image 15: Information on datatypes**



**Image 16: Count of null values in each column of dataset**

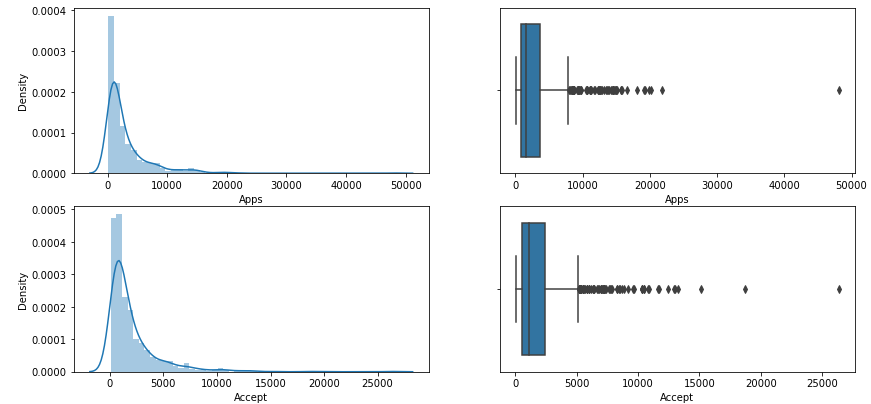
Summary of Dataset

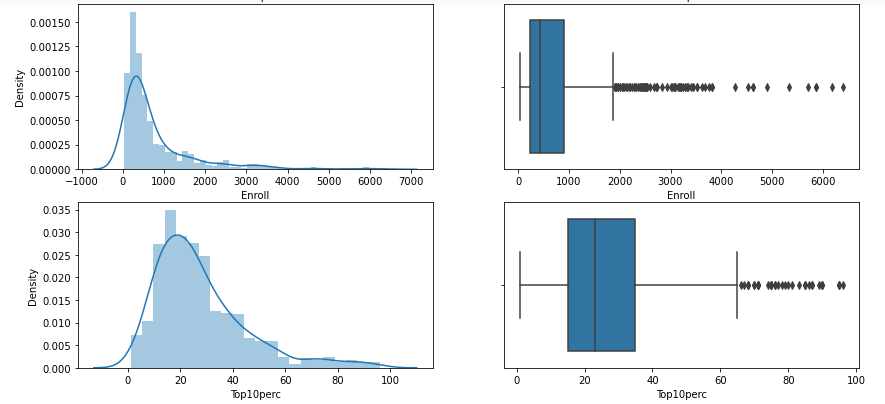


**Image 17: Descriptive summary of dataset**

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

**Univariate Analysis-**



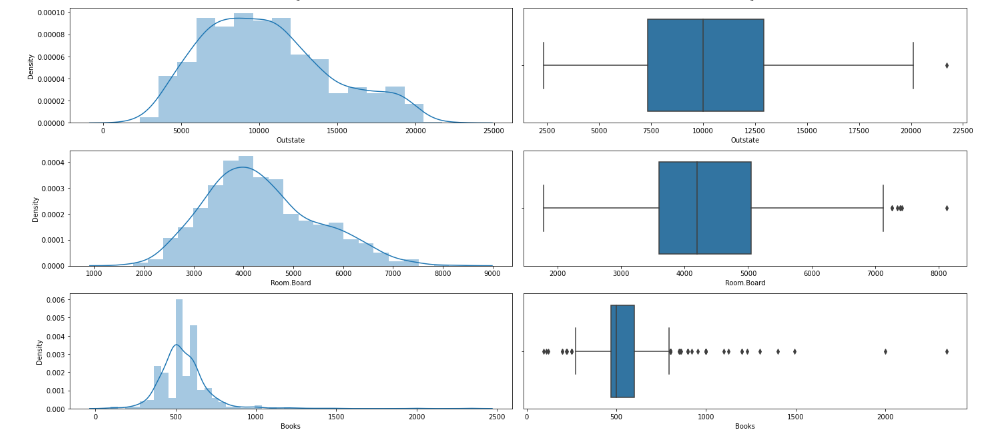
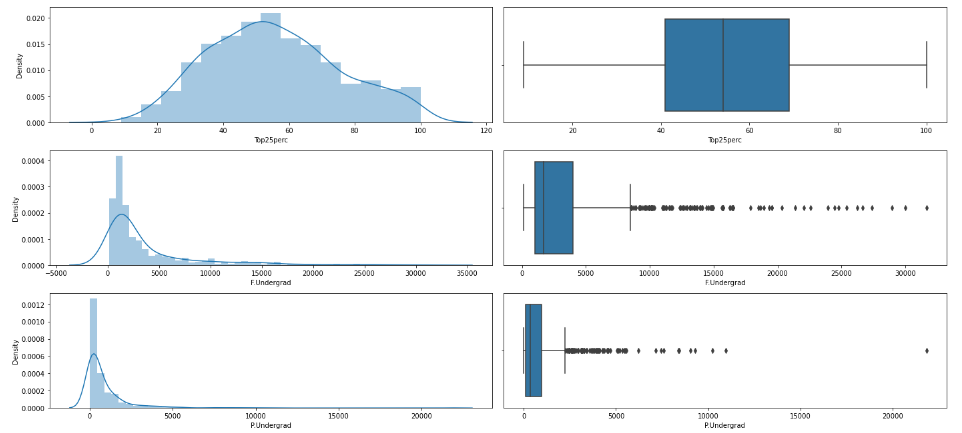


**Image 18: Histogram and boxplots of mentioned columns**

#### 

#### Insights:

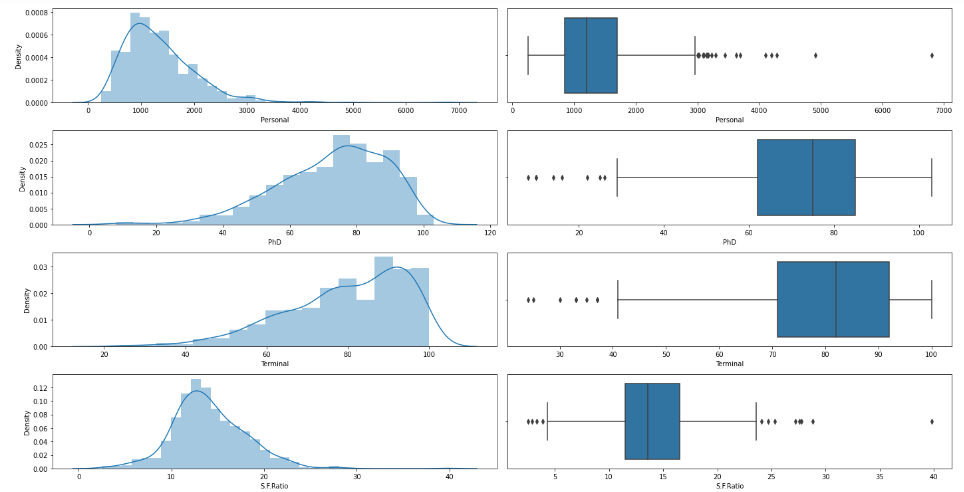
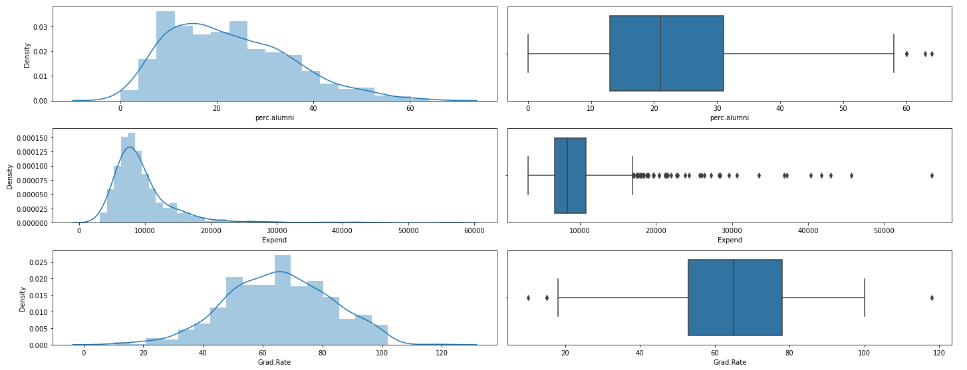
Apps, Accept, Enroll and Top10Perc have outliers present and are right skewed.



**Image 19: Boxplot and histogram of featured columns**

#### Insights:

* Top25Perc has no outliers and is normally distributed.
* P.Undergrad and F.Undergrad have outliers and are highly skewed.
* Outstate variable has outliers in upper values and seems normally distributed.
* Room.Board column is normally distributed and has outliers in upper values.
* Books have outliers in upper and lower values and has high degree of skewness.

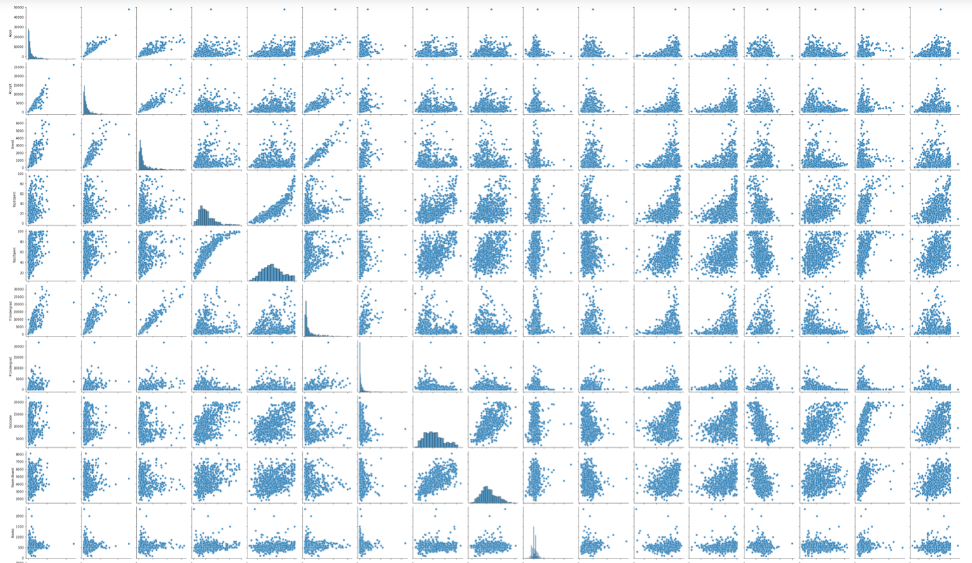
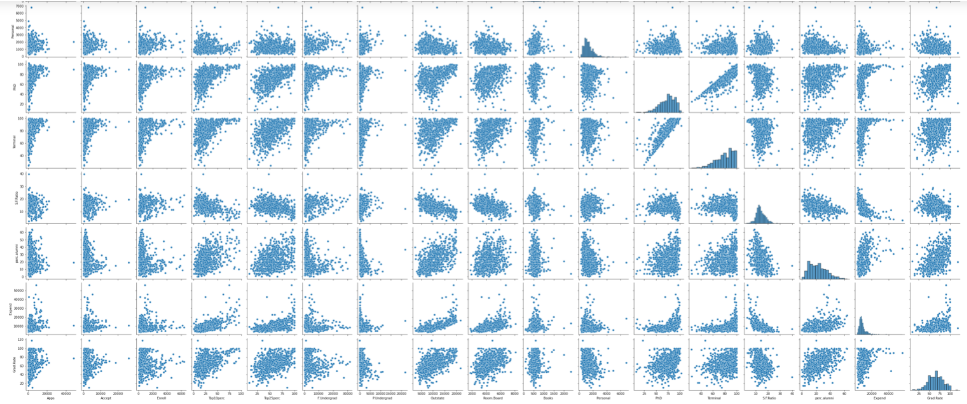
 

**Image 20: Boxplot and histogram of remaining columns**

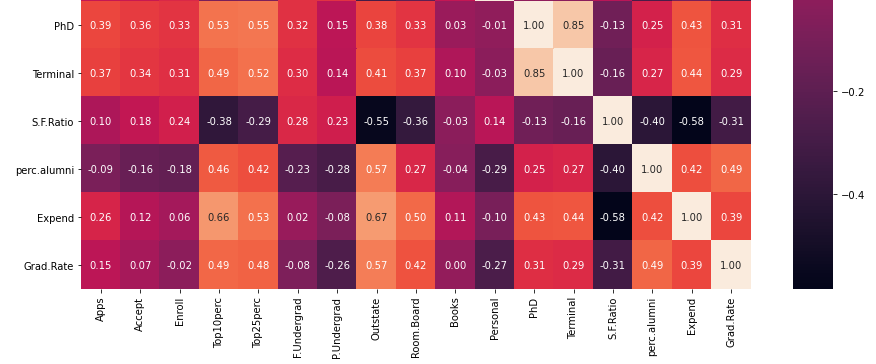
#### Insights:

* Personal have outliers and is right skewed.
* PhD and Terminal is left skewed and has outliers in lower values.
* S.F. ratio has outliers in lower and upper values.
* perc.alumni and expend is right skewed and has outliers in upper values.
* Grad.Rate is left skewed and has outliers in both upper and lower values.

**Multivariate Analysis-**

**Image 21: Pair plot of different dataset columns**

**Image 22: Heat map of dataset**

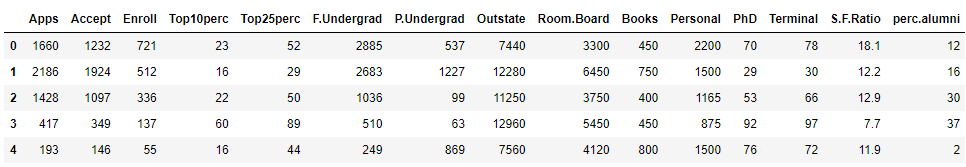
### Insights-

* It is observed that there is high positive correlation between Apps, Accept , Enroll and F. Undergrad
* PhD and Terminal are highly correlated.
* Positive correlation between Apps, Books, Personal, PhD and Terminal.
* perc.alumni, Apps, Accept, Enroll are negatively correlated.

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

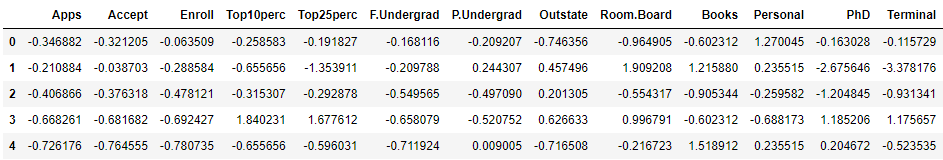
**Scaling is necessary for PCA in this case as dataset has features with different “weights”. In “Distance” based algorithms it is recommended to transform the features so that all features are in same “scale”. It is used in weight based techniques like PCA. When variances in the dataset is widely different and variables are scaled in different units, it is difficult to perform PCA on non-linear data structure. There are many variables at hand and it is tough to focus on particular variable of importance, therefore standardising the data will give mean tending to 0 and standard deviation tending to 1 and analysts can easily compare variables for further analysis.**

**“Names” variable does not add any value to model building hence it is dropped before performing scaling on numerical columns.**

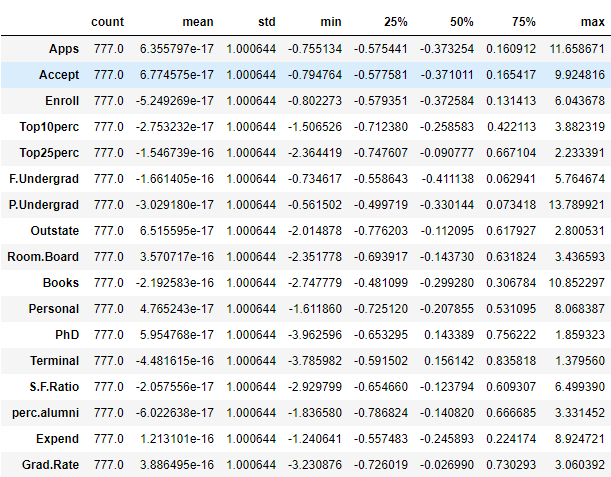


**Image 23: Top 5 records of dataset after dropping “Names” column**

**Z-score technique is performed to scale the data as shown below –**



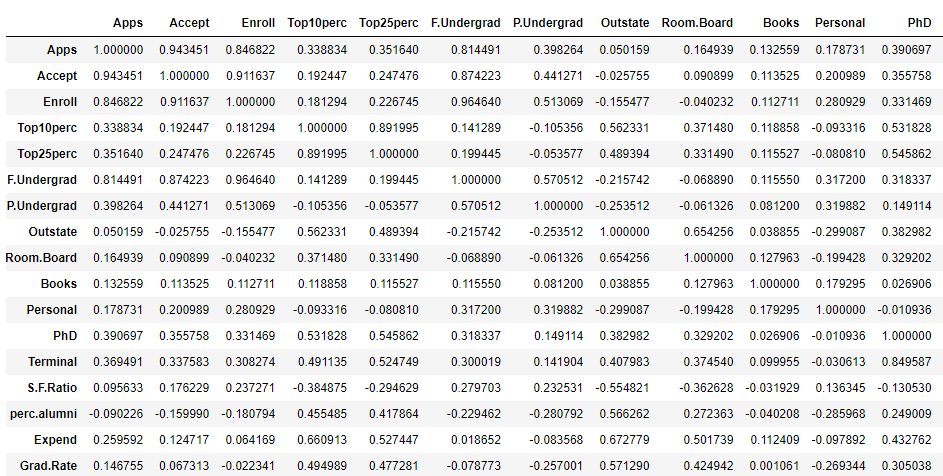
**Image 24: Head of scaled dataset**



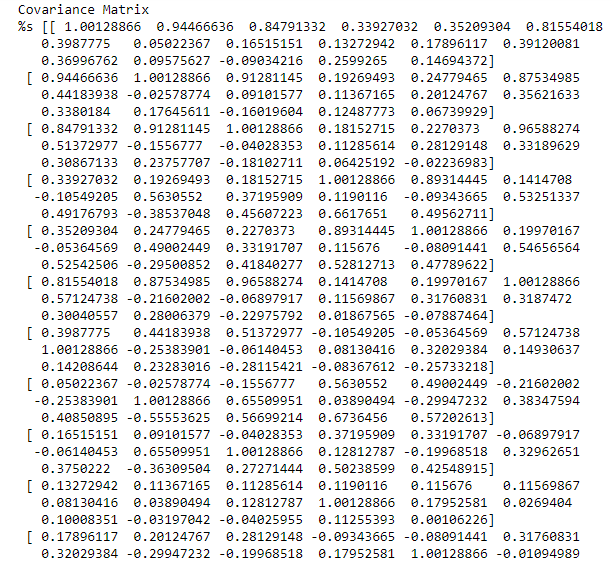
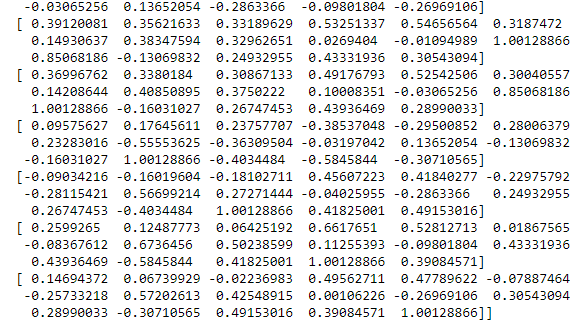
**Image 25: Summary of scaled data**

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

Covariance indicates the direction of linear relationship between variables while correlation measures both the strength and direction of the linear relationship between two variables. Covariance matrix is a mathematical representation of total variance of individual dimension and across dimensions. Correlation between two variables is expressed by either +1 or -1.When one variable increases as the other increases, then correlation is positive. When one decreases as the other decreases, then it is positive. Generating covariance matrix or correlation matrix for all dimensions is one of the four steps while performing PCA.

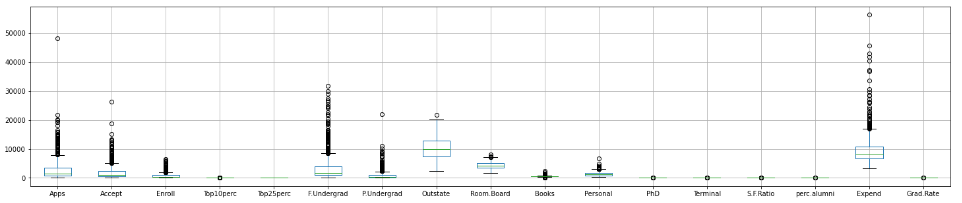


**Image 26: Correlation matrix of scaled data**

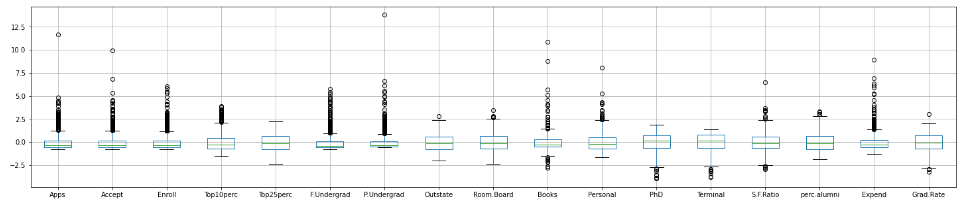
 

**Image 27: Covariance matrix for scaled data**

**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]**



**Image 28: Box plot pre-scaling**

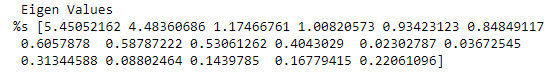


**Image 29: Box plot post-scaling**

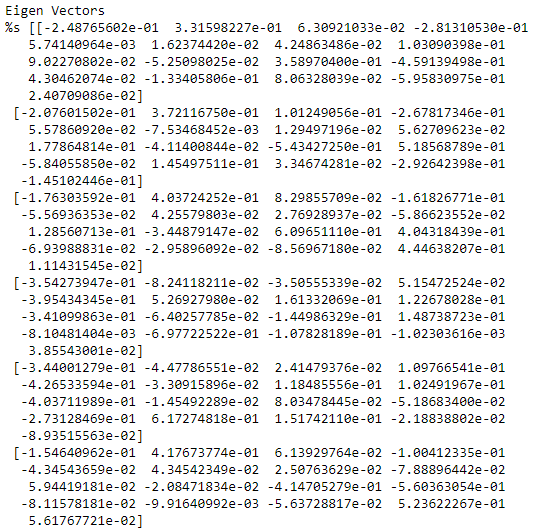
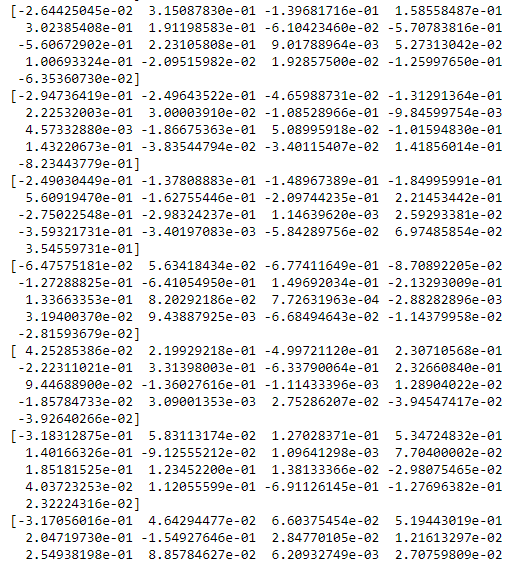
**Insights-**

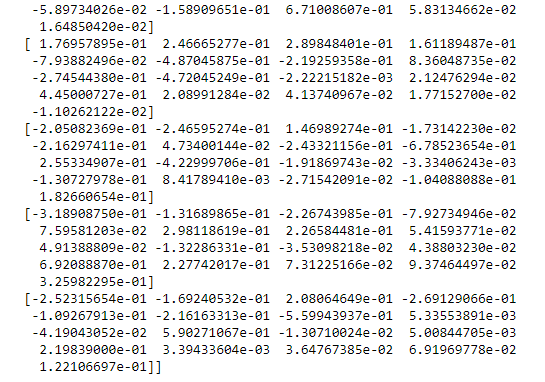
**Prior to scaling, there were lot of variations in the data. Post scaling, median of all the columns are quite close to each other. All variables have same standard deviation and outliers are still present as outliers have not been treated. Scaling brings the data to unit variance and PCA can be applied easily.**

**5. Extract the eigenvalues and eigenvectors.**



**Image 30: Eigen values of scaled data**



**Image 31: Eigen vector of scaled data**

**Eigen values and Eigen vectors were calculated from covariance matrix.**

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

### Statistical tests to be done before PCA-

#### Bartletts Test of Sphericity

Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

* H0: All variables in the data are uncorrelated
* Ha: At least one pair of variables in the data are correlated

If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is small (0.0 in this problem), then we can reject the null hypothesis and agree that there is at least one pair of variables in the data which are correlated hence PCA is recommended.

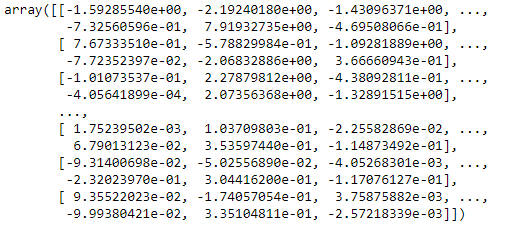
#### KMO Test

The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

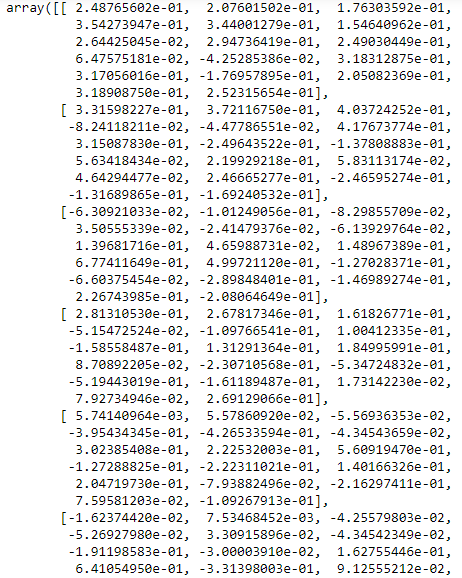
Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components. MSA is 0.813 in this case study.

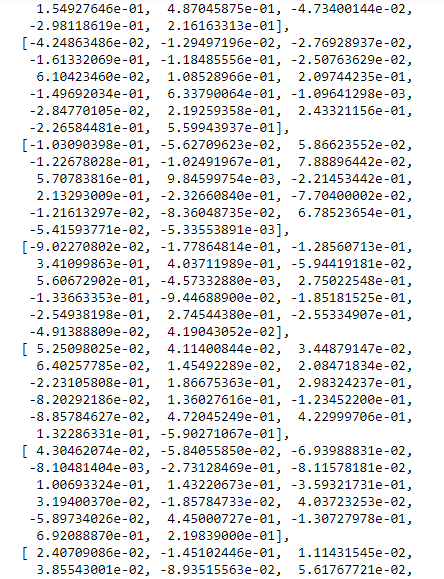
Steps for PCA-

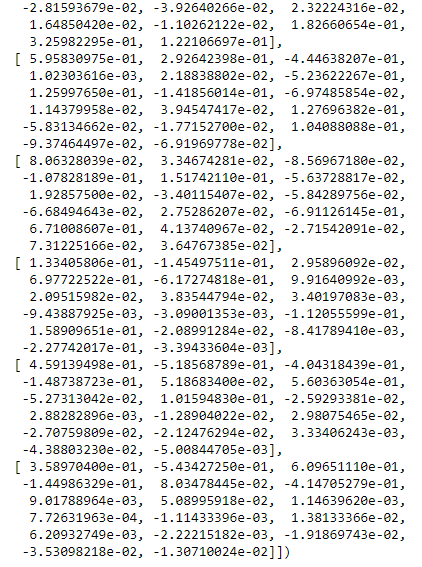
* Begin by standardising the data. Data on all the dimensions are subtracted from their means to shift the data points to the origin i.e. the data is centred on the origins.
* Generate the covariance matrix or correlation matrix for all the dimensions.
* Perform Eigen decomposition, that is, compute Eigen vectors which are principal components and the corresponding Eigen values which are magnitudes of variance captured.
* Sort the Eigen pairs in descending order of Eigen values and select the one with largest value. This is the first principal component that covers the maximum information from original data.



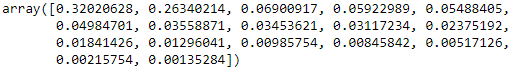
**Image 32: PCA component output**



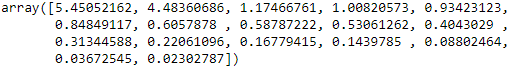
 



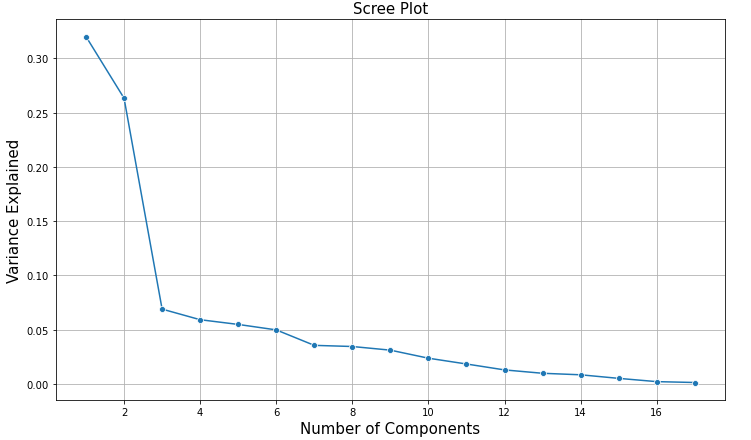
**Image 33: Loading of each feature on the components**



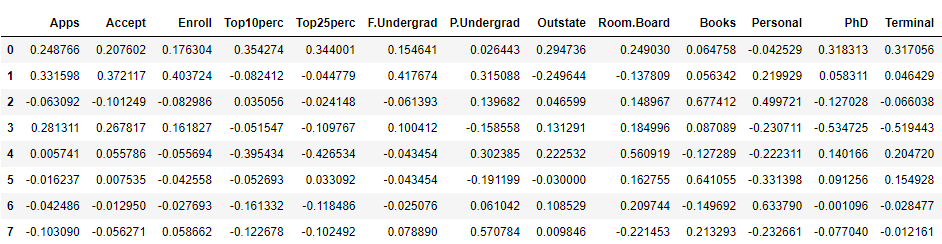
**Image 34: Explained variance of each principal component**



**Image 35: Checking Eigen values in descending order**



**Image 36: Scree Plot to identify number of components**



**Image 37: Eigen vectors exported to data frame with original features**

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

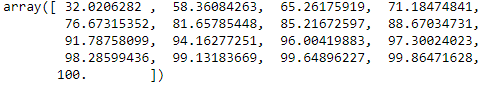


**Image 38: Eigen vector of first PC**

Linear equation of First PC –

0.25 \* Apps + 0.21 \* Accept + 0.18 \* Enroll + 0.35 \* Top10perc + 0.34 \* Top25perc + 0.15 \* F.Undergrad + 0.03 \* P.Undergrad + 0.29 \* Outstate + 0.25 \* Room.Board + 0.06 \* Books + -0.04 \* Personal + 0.32 \* PhD + 0.32 \* Terminal + -0.18 \* S.F.Ratio + 0.21 \* perc.alumni + 0.32 \* Expend + 0.25 \* Grad.Rate

**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?**



**Image 39: Cumulative values of Eigen values**

Cumulative variance up to 90% and incremental value between the components should not be less than 5% are the two factors considered while deciding on optimum number of principal components. Thus, optimum number of components from above image considering two factors is 5.

Eigen vectors indicates the direction of principal component, weights attached to each variable. Eigen vectors signifies individual contribution of each variable to the principal components.

**9. Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis?**

With help of PCA we have been able to reduce 17 numeric features into 5 components which is able to explain 80% of variance in the data.

Reduced components makes it easier to recognize underlying patterns to gain better insights. Using the components additional rules can be derived and analysed.

Principal components can be used to deal with multicollinearity in the data. Before selecting any particular model, it is essential to perform PCA so that variables involved in dataset are independent of one another. PCA also helps to drop “least important” variables which do not play any role in model building. PCA further guides universities to consider background verification of students post 12th education based on components listed below before admitting them to colleges.



**Image 40: Maximum feature loading in rectangular plot**

Interpretations of the Principal Components Obtained

* PC0: Explains Expend**,** Outstate
* PC1: RepresentsApps, Accept**,** Enroll, F. Undergrad, P. Undergrad, perc.alumni
* PC2: ExplainsBooks, Personal, S.F. Ratio
* PC3: Details PhD, Terminal, Grad. Rate
* PC4:HighlightsTop10Perc, Top25Perc**,** Room.Board