**Business Report**

**Project –SMDM**

**Created by Amit Jain**

Table of Contents

[**1.** **SALARY DATA ANALYSIS.** 4](#_Toc101543794)

[**Insight of the data set:** 4](#_Toc101543795)

[**1A-1 : State the null and the alternate hypothesis for conducting one-way ANOVA** 5](#_Toc101543796)

[**1A-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](#_Toc101543797)

[**1A-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.** 6](#_Toc101543798)

[**1A-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)** 7](#_Toc101543799)

[**1B - 1: What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the ‘point plot’ function from the ‘seaborn’ function]** 9](#_Toc101543800)

[**Problem 1B - 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?** 11](#_Toc101543801)

[**Problem 1B - 3: Explain the business implications of performing ANOVA for this particular case study.** 12](#_Toc101543802)

[**Problem 2: College and Student data analysis** 13](#_Toc101543803)

[**2.1 : Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?** 16](#_Toc101543804)

[univariate Analysis 16](#_Toc101543805)

[Multivariate Analysis 19](#_Toc101543806)

[**2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling.** 22](#_Toc101543807)

[**2.3 Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].** 25](#_Toc101543808)

[**2.4: Check the dataset for outliers before and after scaling. What insight do you derive ? [Do not treat Outliers unless specifically asked to do so]** 26](#_Toc101543809)

[**Check for Outliers Before scaling** 26](#_Toc101543810)

[**Check for Outliers After scaling :** 27](#_Toc101543811)

[**2.5 Extract the eigenvalues and eigenvectors.[Using Sklearn PCA Print Both]** 28](#_Toc101543812)

[**2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features** 29](#_Toc101543813)

[**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]** 30](#_Toc101543814)

[**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?** 31](#_Toc101543815)

[**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]** 32](#_Toc101543816)

**List of Figure**

|  |  |  |
| --- | --- | --- |
| **Fig No** | **Description** | **Page No** |
| 1 | Point Plot for Mean Salary for Education Categories | 8 |
| 2 | Count Plot for Mean salary for Education Categories | 8 |
| 3 | Intersection Plot for Education and Occupation categories | 10 |
| 4 | Count Plot for Education and Occupation categories | 10 |
| 5 | Intersection Plot for Education and Occupation categories | 11 |
| 6 | Distribution and Box Plot for all data elements | 17, 18 19 |
| 7 | Heat Map for all data elements | 20 |
| 8 | Box Plot for all data elements | 21 |
| 9 | Box Plot without Scaling after data correction | 26 |
| 10 | Box Plot after Scaling after data correction | 27 |
| 11 | Weightage chart for each PCA component | 31 |

# **SALARY DATA ANALYSIS.**

**Problem Statement:**

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/67673/files/5109354/download?verifier=MumDb4pLo5Q694KNfg1X6tGzT1dhgvywq2HGYyyH&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.

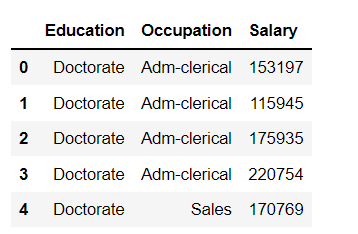
Introduction: This report explains the business requirements and provide the detailed solution based on the data provided for each problem statement. given in the assignment.

**Assumption**:

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

**Step of understanding the data:**

* Import the data: Imported the data using Python notebooks and analyzed the effects of Education and Occupations over salary field.

This is how the data look like:

# **Insight of the data set:**

1. Data Have 3 fields, Education, Occupation and Salary
2. Education have 3 categories: Doctorate, Bachelors and HS-Grad
3. Occupation have 4 Categories: Adm-clerical', ' Sales', ' Prof-specialty', ' Exec-managerial'
4. We need to analyze effects of Education and Occupation Categories over Salary.
5. Shape of the data set is 40 Rows and 3 Columns
6. There are no null values in any Column values
7. Salary ranging between 50103 and 260151 is a good sign, and seems all data is normal in nature, no bad data

# **1A-1 : State the null and the alternate hypothesis for conducting one-way ANOVA**

**Hypothesis for Education Qualification**

Null Hypothesis Ho : Mean Salary is same across all Education Qualification

Alternate Hypothesis Ha: Mean salary is different for at least one level of Education Qualification.

**Hypothesis for Occupation Qualification**

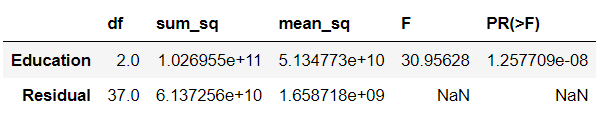
Null Hypothesis Ho : Mean Salary is same accross all Occupation Level

Alternate Hypothesis Ha: Mean salary is different for at least one level of Occupation Level

# **1A-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.**

For performing One way ANOVA Test we have used anova\_lm module from statsmodels.stats.anova python Library

After performing one way ANOVA test, following P value was generated:

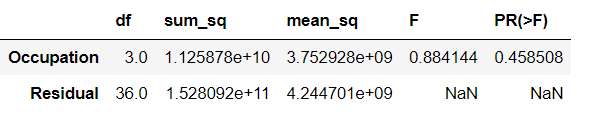


Since the P value is less than alpha value 0.05, so we REJECT the NULL Hypothesis, which says, that mean Salary is same across all Education Qualification level, which means there are differences in mean salary based on Education

# **1A-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.**

For performing One way ANOVA Test we have used anova\_lm module from statsmodels.stats.anova python Library

After performing one way ANOVA test, following P value was generated:



Since the P value is Greater than alpha value 0.05, so we FAIL TO REJECT the NULL Hypothesis, which says, that Mean Salary is same across all Occupation level Which means , Mean salary is same for all Occupation categories

# **1A-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)**

As per results from Problem 1A-2 and 1A-3, that There was significant Difference between

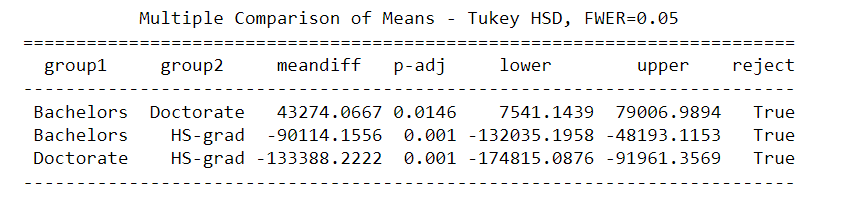
Mean salary for Educational Categories, so Let’s compare the mean for Educations levels.

Education have 3 categories with in it: ' Doctorate', ' Bachelors', ' HS-grad'

For analyzing the mean Salary , based on multiple categories of Education, we would use

MultiComparison module from statsmodels.stats.multicomp Python Library, and after

analysis, Following P values will be driven, for all categories of Education Column:



From the Output of the MultiCompare Test result we observe that the P values of all the

Category of Education Levels are lower than 0.05, So we can say that Mean Salary would be

different for all categories of Education levels.

Now check for Mean Salaries for the Given sample data set for all 3 Education levels:

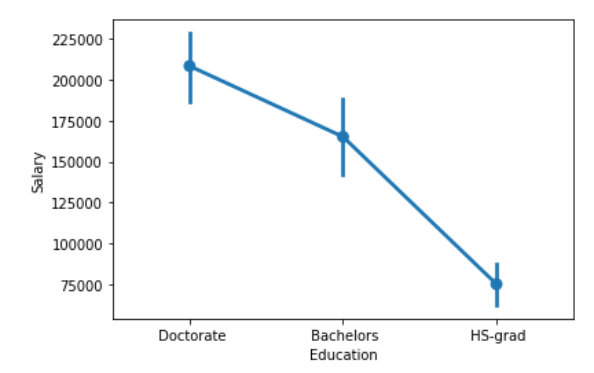
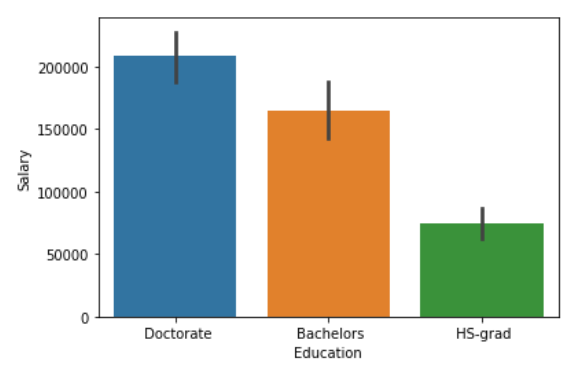


Fig:1 Point Plot Fig:2 Count Plot

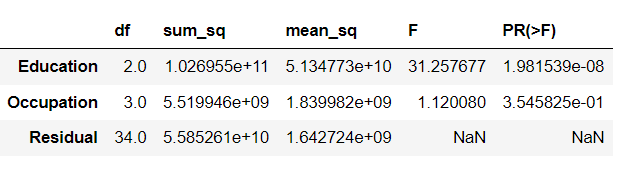
Based on Point Plot and Bar Plot , we can clearly see that mean salary for Doctorate Educational level will be Higher than all and HS-grad Mean salary should be least in all Educational levels.

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

Since this problem statement is seeking comparative analysis of both treatments (Education

and Occupation ) Over salary field, So we would use multi compression and check the P-values

for Combined effect of education and Occupation over salary :

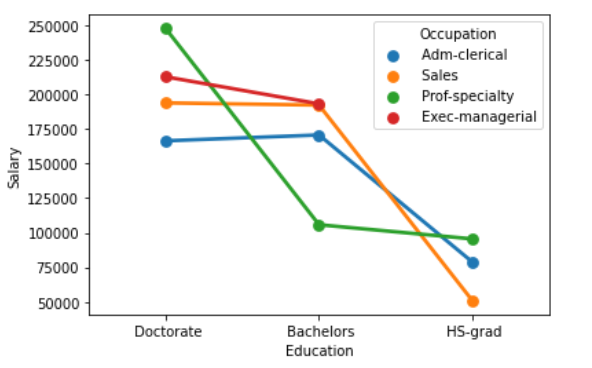


The p-value in the treatments are Less than 𝛼 (0.05), so with 95% confident level we can say that

Mean Salary are not same for Combined effect of Education and Occupation, and there will be

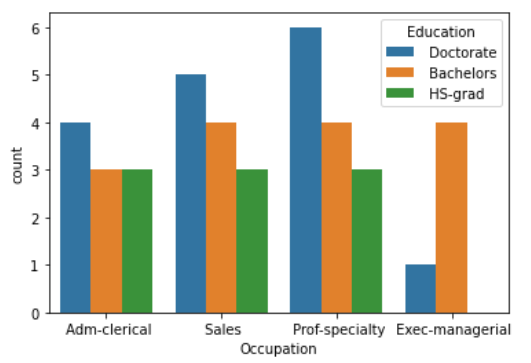
significant differences in Salaries based on combination of Education and Occupation .

Lets Check the interactions using Point Plot method



**Fig 3: Intersection Plot**

**Check for the count plot too for each treatment:**



**Fig 4: Count Plot**

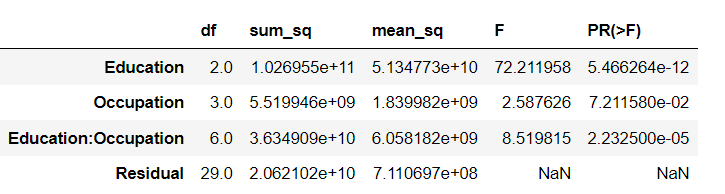
# **Problem 1B - 2: Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education\*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?**

As we confirmed from the interaction plot that there is interactions between education and Occupation levels, So we would conduct a two way anova to check the interaction between Education and occupation.

Null Hypothesis Ho : There is no interaction between Education and Occupation

Alternate Hypothesis Ha : There is an interaction between Education and Occupation

We will use the ANOVA test for Interaction effect of both treatment over Salary . Generate the P value metrics



**We will now check Interaction Plot method**

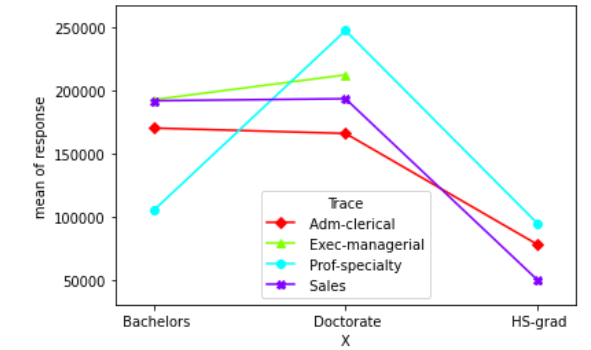


Fig 5: Interaction Plot

We can see that there is some sort of interaction between the two treatments.

Since P value is less than 0.05, So we reject the null hypothesis , which says that There is no interaction between Education and Occupations over Salary and with 95% significance level there is an interaction between Education and Occupation. which means the mean salary is affected by an interaction of Education and Occupation. From the interaction plot, we can say that there is interaction between Education and Occupation levels

# **Problem 1B - 3: Explain the business implications of performing ANOVA for this particular case study.**

By conducting One way ANOVA test : We analyzed Mean salary for any specific Category, So there is no difference in Mean salary for any occupation levels, whereas there are differences in Mean Salary for Educational Levels.

Highest earning Salary have done the Doctorate and lowest earning group is High school graduate.

When we Test for Combined effect of Education and Occupation, we did 2-way ANOVA test, and found that Education and Occupation together are affecting the salary as well and there are differences in mean salary for the combinations of Both categories. and visualized it through intersection plot and observed that there is a interaction between the two categories.

# **Problem 2: College and Student data analysis**

**Problem Statement:**

The dataset Education – Post 12th Standard.csv contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given. The data dictionary of the ‘Education – Post 12th Standard.csv’ can be found in the following file: Data Dictionary.xlsx.

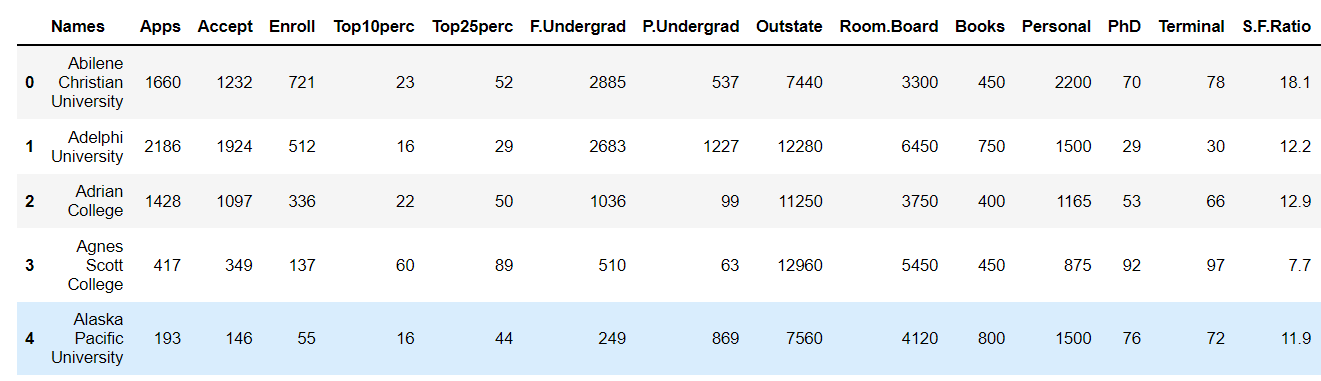
Introduction: This report explains the business requirements and provide the detailed solution based on the data provided for each problem statement. given in the assignment.

**Assumption**:

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

**Step of understanding the data:**

* Import the data: Imported the data using Python notebooks and analyzed the effects of Education and Occupations over salary field.

This is how the data look like:

Insights of the data set:

1. As per the data dictionary given, following are the details of all the Columns in the

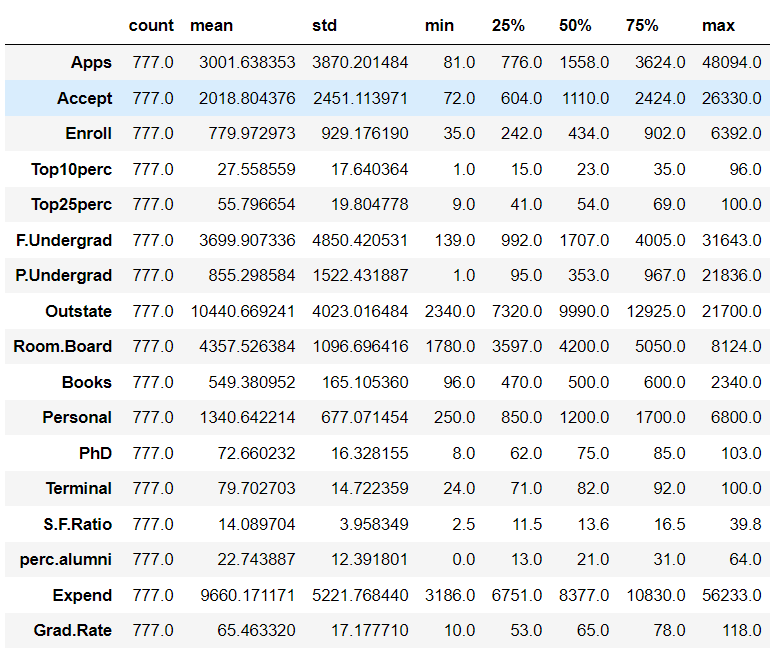
source data set:

* 1. Names: Names of various university and colleges
  2. Apps: Number of applications received
  3. Accept: Number of applications accepted
  4. Enroll: Number of new students enrolled
  5. Top10perc: Percentage of new students from top 10% of Higher Secondary class
  6. Top25perc: Percentage of new students from top 25% of Higher Secondary class
  7. F.Undergrad: Number of full-time undergraduate students
  8. P.Undergrad: Number of part-time undergraduate students
  9. Outstate: Number of students for whom the particular college or university is Out-of-state tuition
  10. Room.Board: Cost of Room and board
  11. Books: Estimated book costs for a student
  12. Personal: Estimated personal spending for a student
  13. PhD: Percentage of faculties with Ph.D.’s
  14. Terminal: Percentage of faculties with terminal degree
  15. S.F.Ratio: Student/faculty ratio
  16. perc.alumni: Percentage of alumni who donate
  17. Expend: The Instructional expenditure per student
  18. Grad.Rate: Graduation rate

1. There are total 17 Numeric fields in the data set
2. Data set have one column “Names” , which is of Object data type and it can be ignored

for the data analysis

1. There are no null values in the data set
2. S.F.Ratio is float64
3. All other columns are in int64 data type,
4. Describe the data set:



1. When Compare to Describe Output with Given data Dictionary Given , Following are

implications:

* 1. "Names" are the Object data type, which can be Dropped for performing EDA
  2. "Apps", "Accepts" , "Enroll" , "F.Undergrad", "P.Undergrad", "Outstate", "Room.Board", "Books", "Personal" seems normal Columns.
  3. "PhD" , "Terminal" and "Grad.Rate" are the percentage values as per data dictionary, but it has maximum values 100 or more than 100, which is not normal, need to change

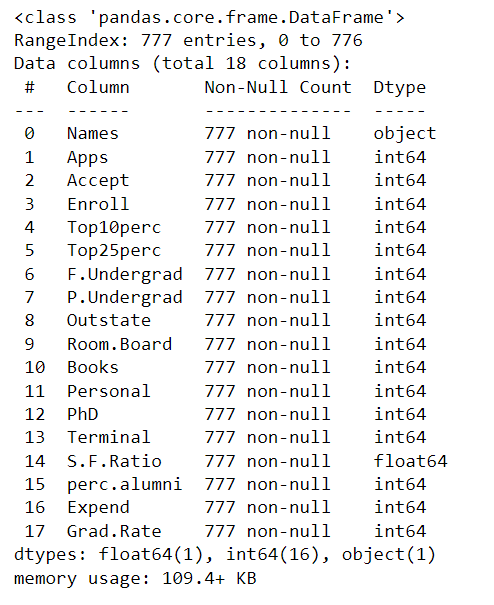
# **2.1 : Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?**

# **univariate Analysis**

in Univariate analysis we need to check Individual columns and its data set distribution , Outliers, missing values , Duplicates, Special characters etc. everything should be tested against each

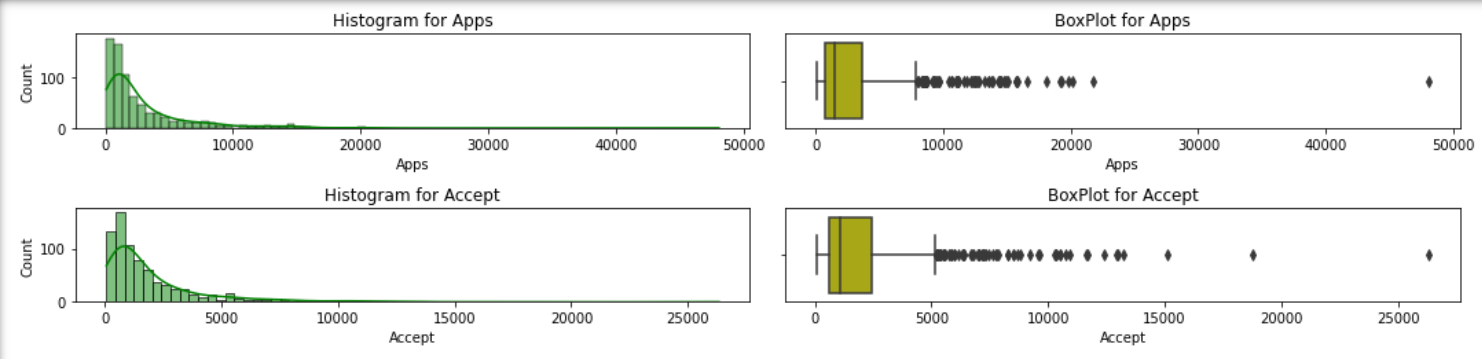
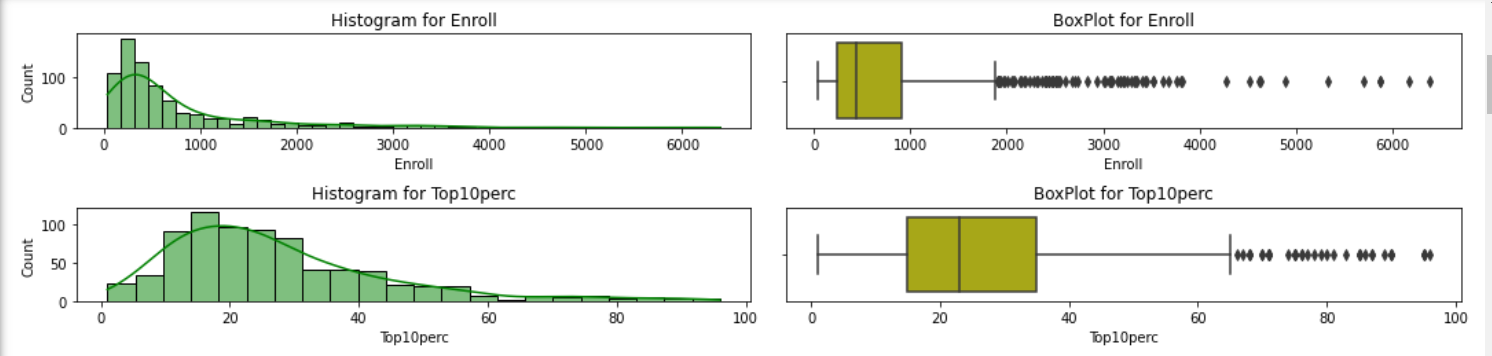
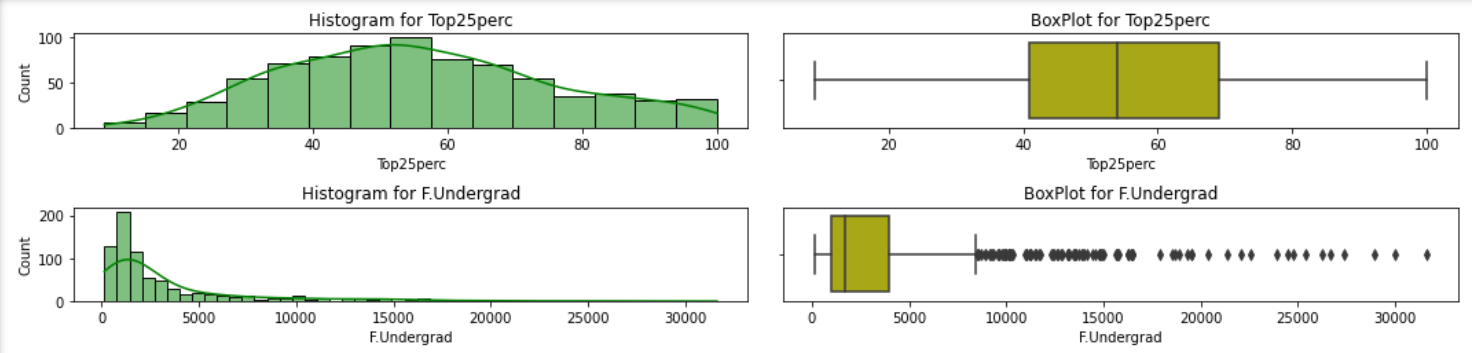
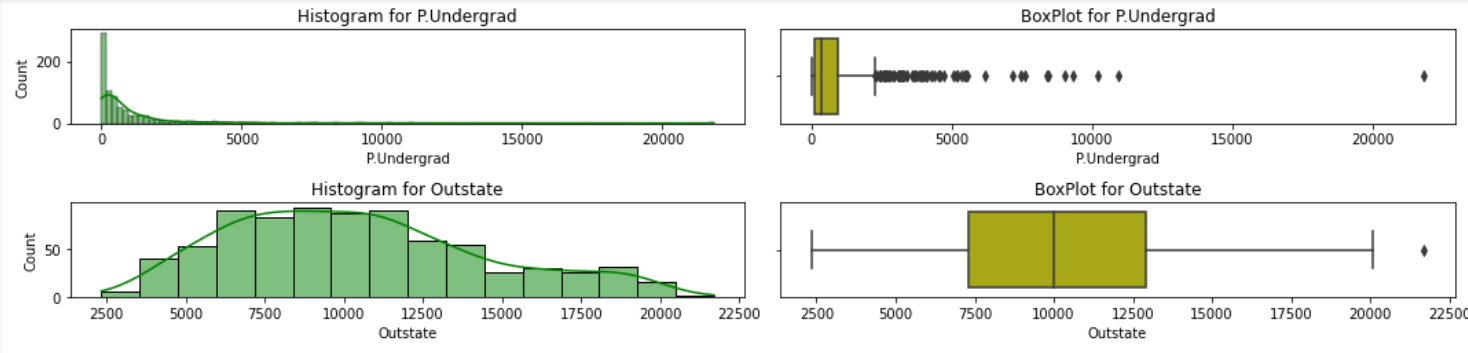
individual column.

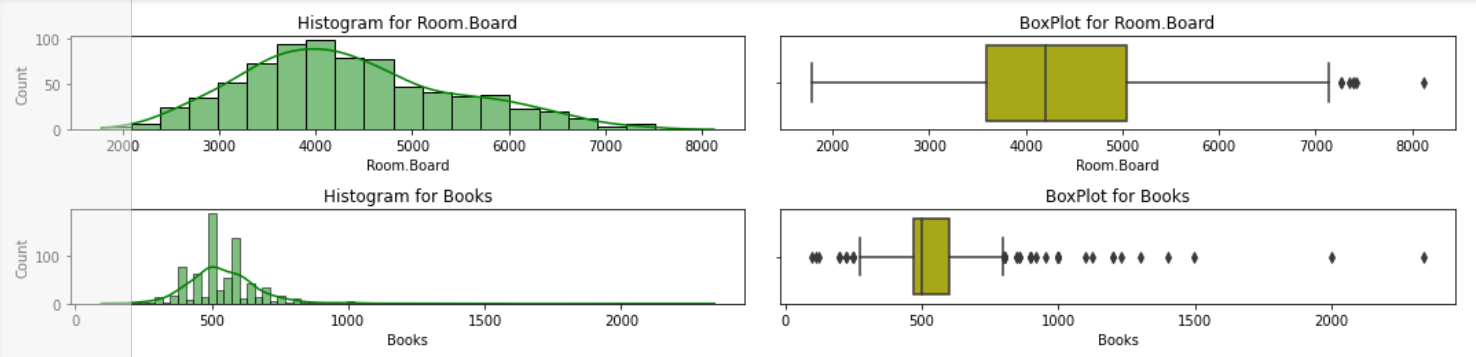
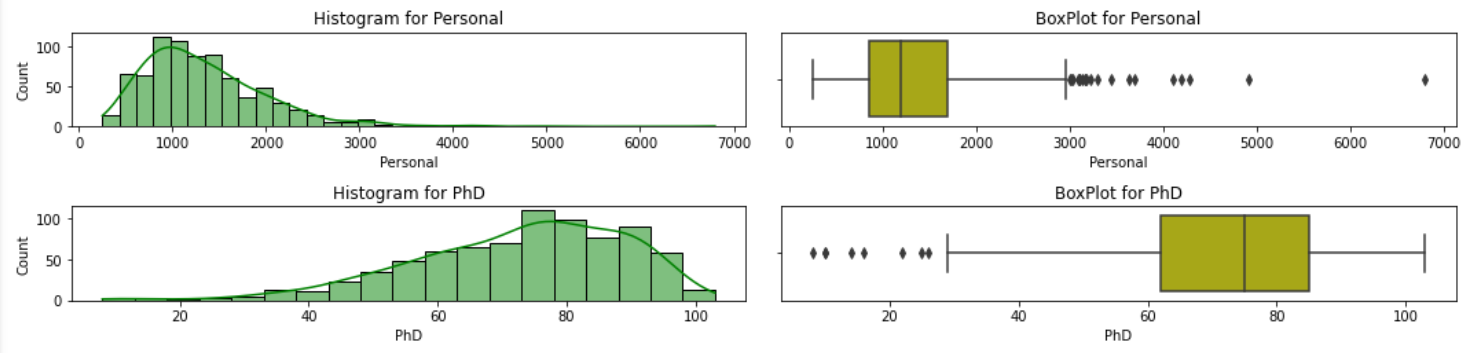
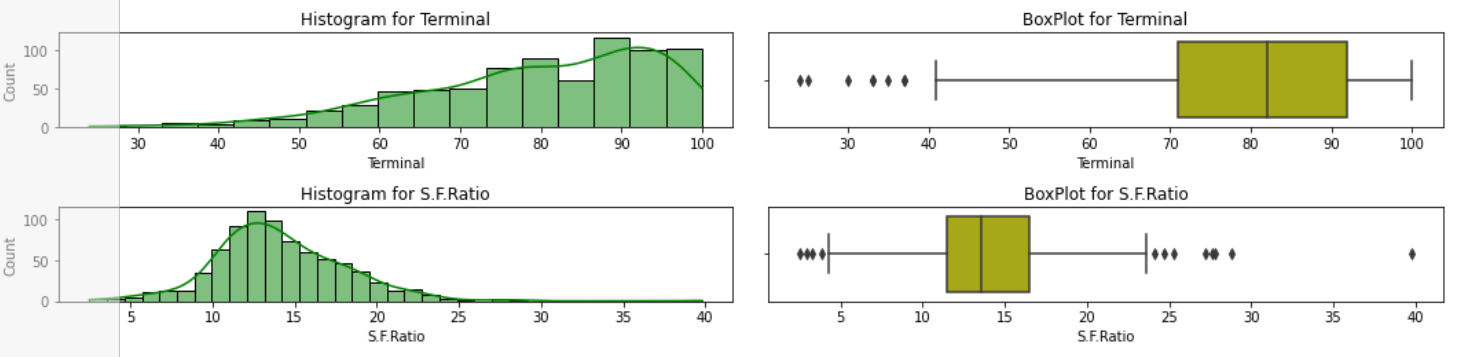
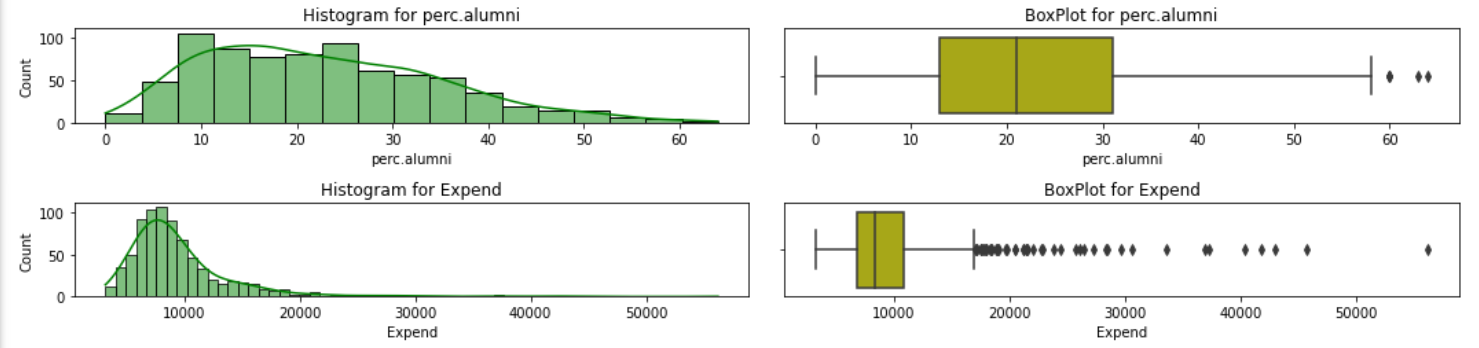
**Check for missing values :**



So No Null values in any of the field.

**Check for Distribution and Outliers in the data set:**





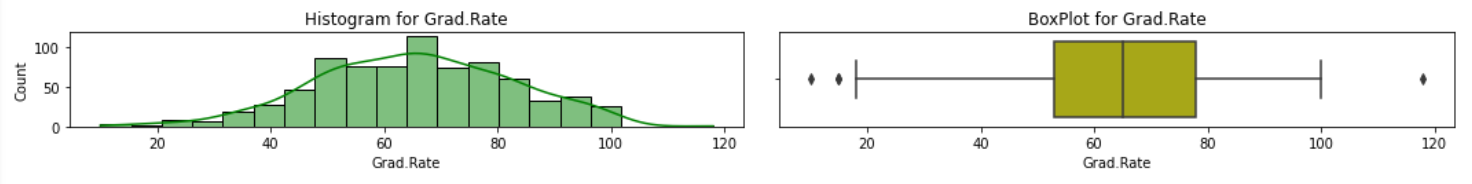
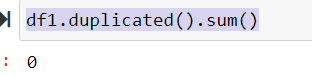


Fig 6: Distribution and Box Plot

Insights: Looking at Box Plot, we can say that : Except Top25perc , remaining all other fields

have outliers in upper values or in lower values.

**Check for Duplicates:**



We don’t have any Duplicate values in the data set.

# **Multivariate Analysis**

In Multivariate analysis , we analyze the combined relations between different fields, and we can consider 2 or more columns and their effects on each other . Directions and strength of the relationships between them. We use Covariance and correlation mechanism for this comparison.

correlations Metrics: We will us heat map for checking the co relation between fields:

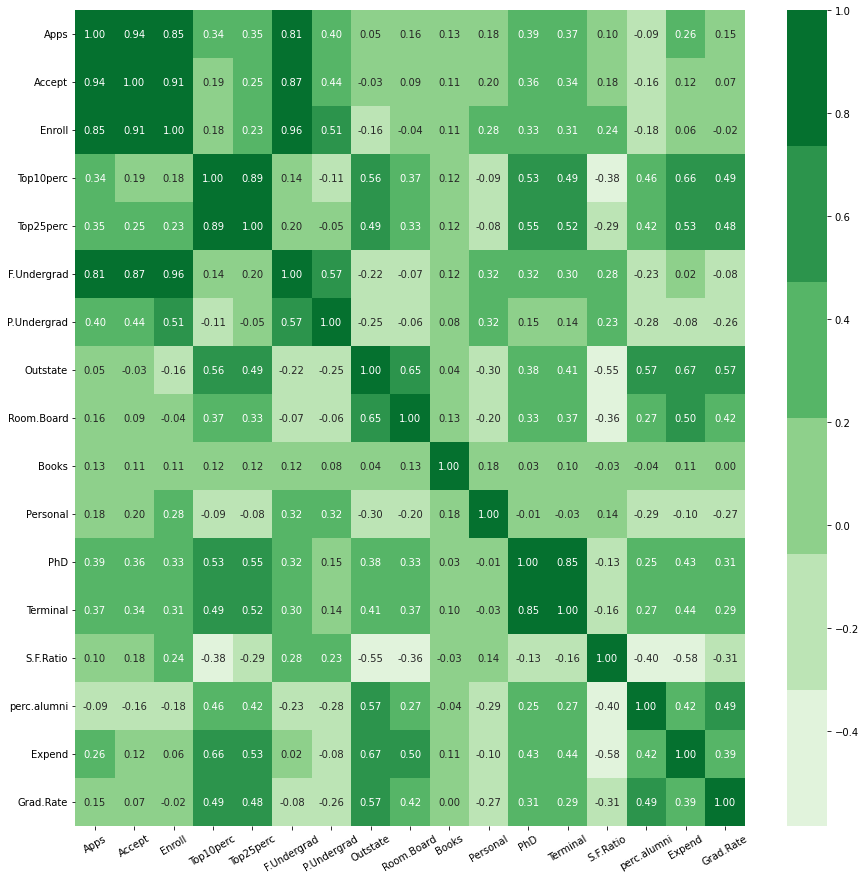


Fig 7 : Heat Map

Check for the Box Plot , with compare to each other , when having single Y – Axis:

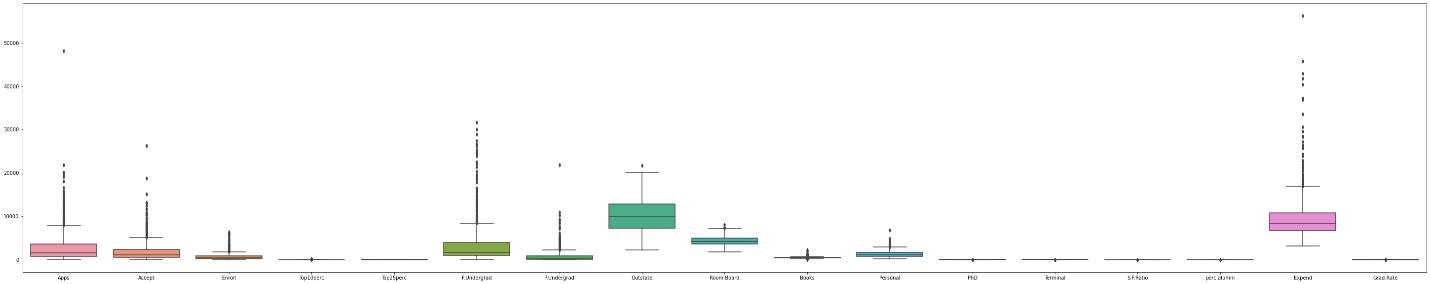


Fig 8: Combined Box Plot

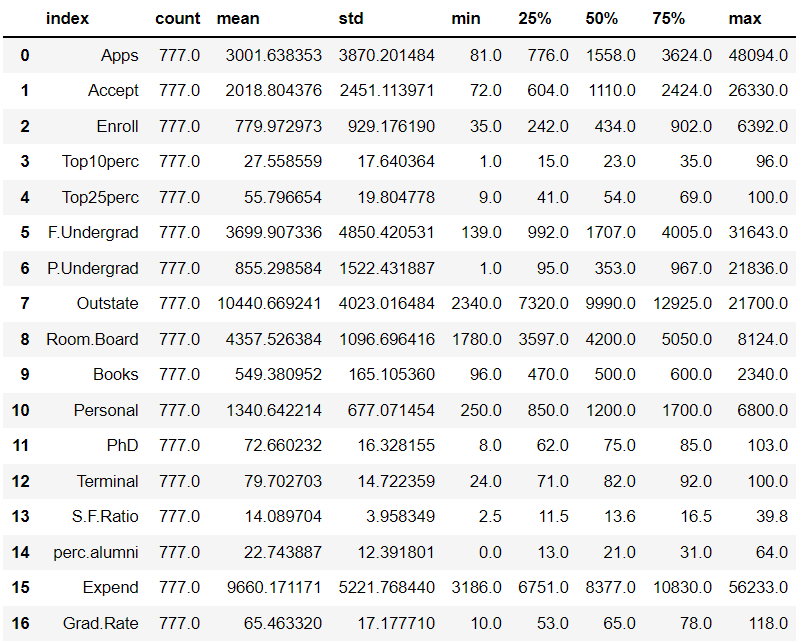
Attached Image for the Box Plot for better visibility :



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

We need to Describe data set, its Minimum and maximum values, Mean, median etc, for

evaluating, whether we need scaling or not



By looking at the min and max values, we can understand, that Minimum values accross all the column ranging between 1 to 3186 and maximum values ranging between 39 and 56233.0,

which means, we have different Scales for all the columns, that’s Why "WE NEED TO SCALE THE DATA BEFORE PROCESSING"

#### **Treat BAD Data**

Based on the data Describe Implications are: When Compare to Describe Output with Given

data Dictionary Given , Following are implications:

1. "Names" are the Object data type, which can be Dropped for perforing EDA

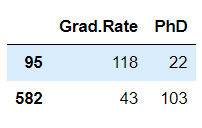
2. "Apps", "Accepts" , "Enroll" , "F.Undergrad", "P.Undergrad", "Outstate", "Room.Board",

"Books", "Personal" seems normal Columns.

3. "PhD" and "Grad.Rate" are the percentage values as per data dictionary, but it has maximum values more than 100, which is not normal, need to change.

Check how many rows , we have with percentage value of "PhD" and "Grad.Rate"

more than 100.



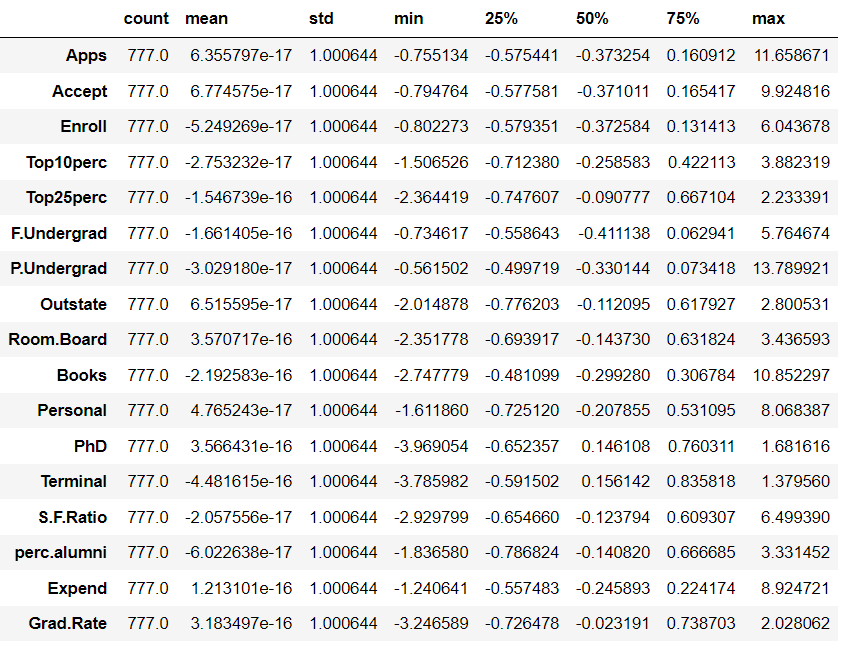
So we have only 2 rows, in row Index 95, we have Grad.Rate more than 100 and in Index = 582 , we have PhD percent more than 100 Lets Treat both them one by one and replace them with Mean value of that Portion

We will replace above values with the MEAN value of that column.

#### **Perform Scaling using z-Score**

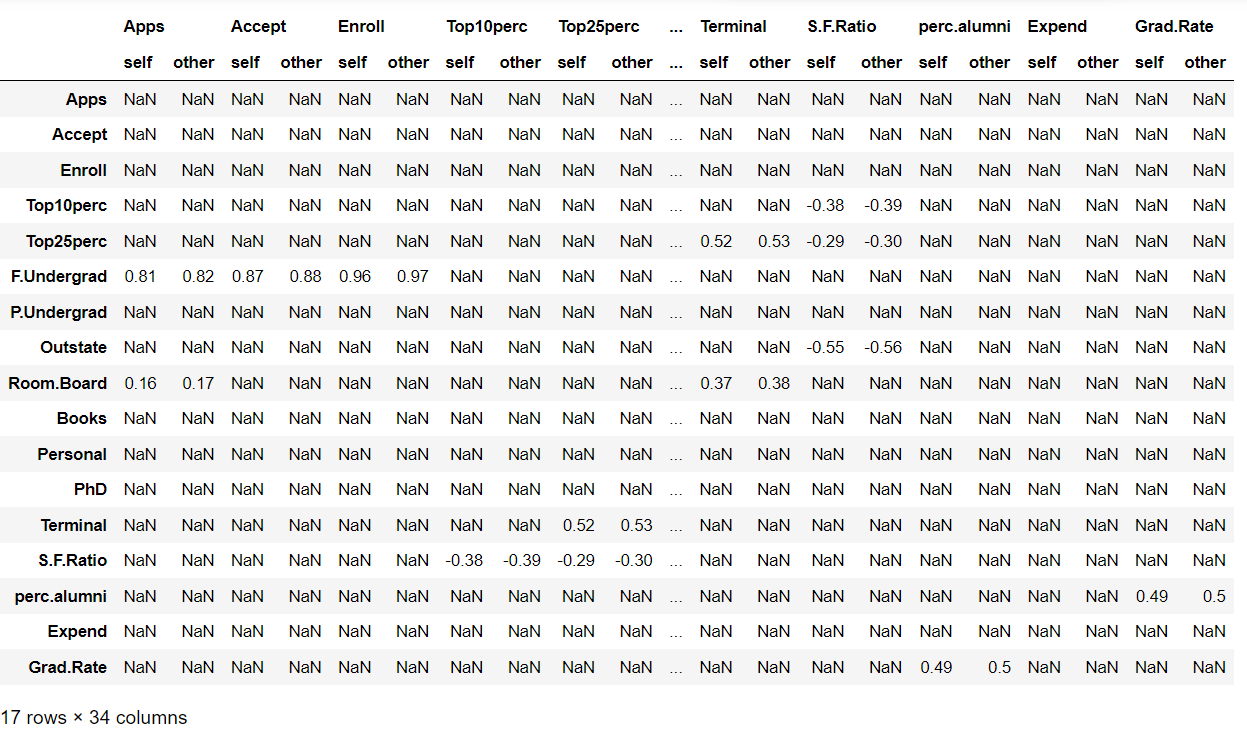
z-Score is the best method for performing scaling , which suits in most of the cases.

After performing z-Score scaling, Describe the data again :



Now we can see that Minimum and maximum values do not have difference of thousands numbers, as we had earlier before scaling as shown in previous Charts.

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

Following are the correlation and covariance comparison charts :

After scaling, when we compare covariance and correlation Metrics, both heat map showing same Graphs. There is very slight difference in the numbers, but those difference numbers are also negligible , if we will round it to ONE decimal point.

CO-Variance Shows the Direction of the relation between 2 variables CO-Relation shows the Direction as well as Strength of the relation between 2 variables

Since both CoVar and CoRel are almost same, we can say that We have good data set after scaling and it is good for processing.

# **2.4: Check the dataset for outliers before and after scaling. What insight do you derive ? [Do not treat Outliers unless specifically asked to do so]**

# **Check for Outliers Before scaling**

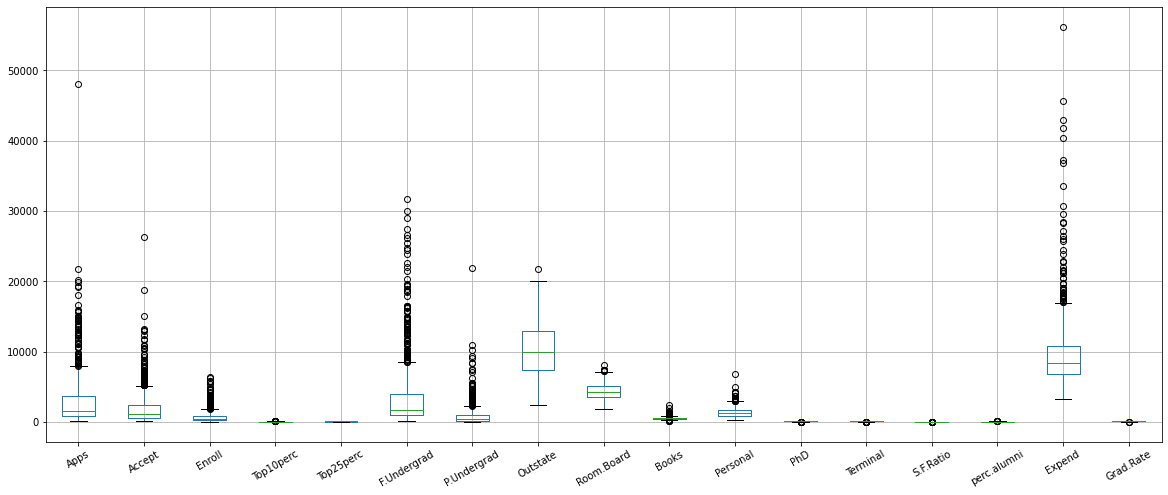


Fig 9 : Box Plot without Scaling after data correction

# **Check for Outliers After scaling :**

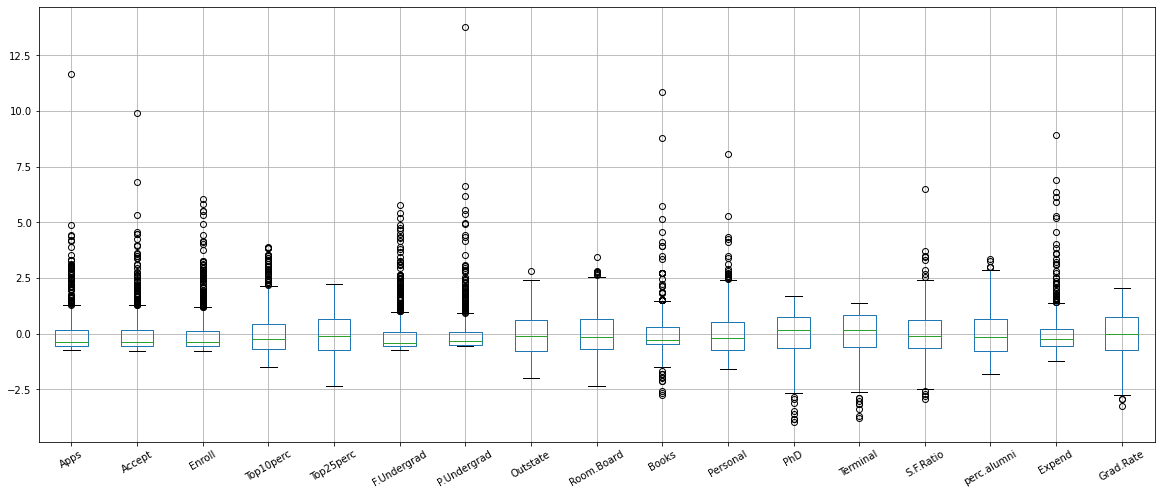


Fig 10: Box Plot after Scaling after data correction

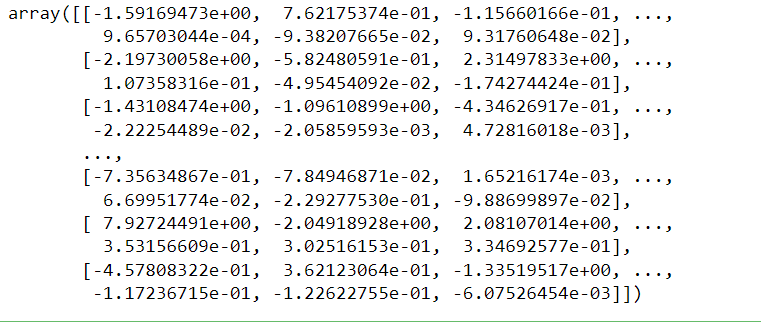
Insights:

1. We can see after scaling Median is closer to 0 for all the Fields.
2. whereas it Median was scattered across all the numbers.
3. We have outliers for about all the fields before and after scaling, as we have not treated the outliers.

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

We have generated the eigenvalues and eigen vectors for all the scaled data set.

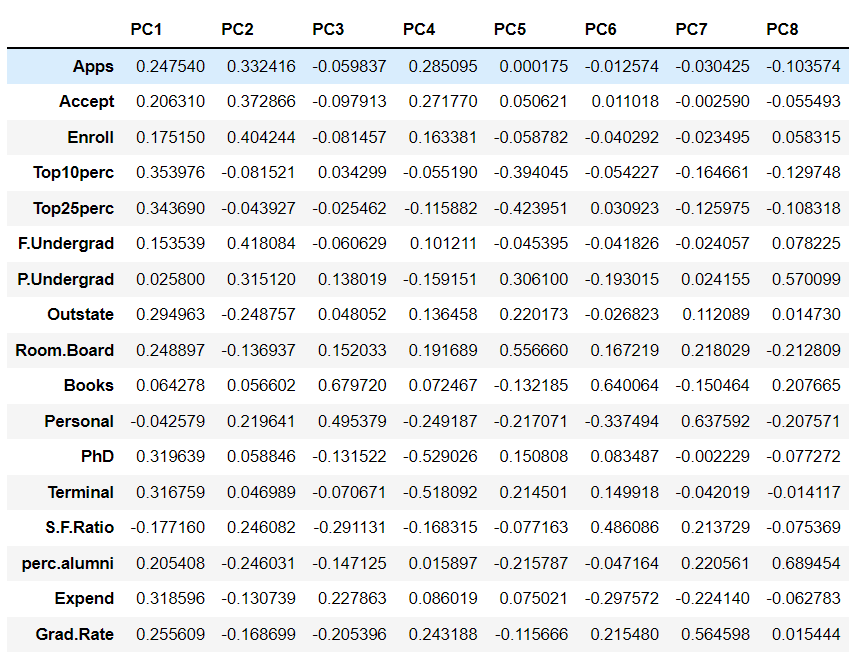
We have used **PCA** Module from **sklearn.decomposition** Python Library for this analysis



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

The PCA analysis was performed and the principal components were exported into

a data fram.



# **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]**

Linear Equation Example:

Y = mX + C

Similarly, we write Equation in terms of the principal components

The explicit form of first PC in terms of eigen vectors up to two places of decimals is obtained as follows:

The linear equation of PC in terms of eigenvectors and corresponding features is shown below

The Linear Equation for the 1st Component will be

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 +0.26 \*

Grad.Rate

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

Cumulative sum is as follows:

array([0.3210007 , 0.584437 , 0.65343689, 0.71205208, 0.76696205,

0.81674035, 0.85233271, 0.88686661, 0.91800986, 0.9417166 ,

0.96011355, 0.97305584, 0.9829117 , 0.99131706, 0.99649123,

0.99864821, 1. ])

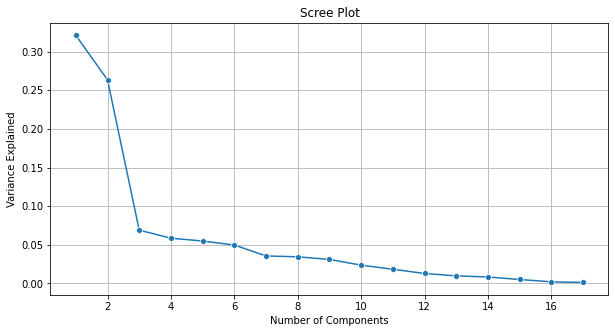


Fig: Weightage chart for each PCA component

We have total 17 columns , but based on cumulative Summation , we can see that initial 7-8 component is contributing approx. 90% of the information, so in our case we will Consider first 8 values of principal components , which will make 88.68% of information. We can take it as optimum number. If we try to take more components it will increase the dimensionality.

# **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]**

We have performed PCA for making all data points come to the vectors and reduce the

dimensionality of the data set. PCA also helps us in understanding the distribution of

every ingredient of the principal components extracted out from the data.

We can further disintegrate each principal component and see that which contents is contributing most in each individual PCA component. By performing the PCA we ensure to eliminate the Noise from data , which comes from the degree of the data points away from normal vectors.

When we performed the PCA, we have got following Cumulative summation of percentage

contribution of each PCA element:

array([**0.3210007 , 0.584437 , 0.65343689, 0.71205208, 0.76696205,**

**0.81674035, 0.85233271, 0.88686661,** 0.91800986, 0.9417166 ,

0.96011355, 0.97305584, 0.9829117 , 0.99131706, 0.99649123,

0.99864821, 1. ])

Out of above data points , we have chosen 8 elements and reduced dimensionality by 9

element ( there were total 17 element). All 8 components are contributing 88.68% of

information, which is enough for data evaluation.

Other Insights are as follows:

1. By using PCA , we could come to know, that 7-8 PCA components are enough for getting more than 85% of the information , Whereas we had 17 Numeric fields to analyze, so we reduce the Dimensionality more than 50%

2. In our analysis we took 8 Components

3. While analyzing individual component, we could come to know, that First Component is covering about 32% of information, and within 1st component itself following fields are playing major role in College "Top10 Percent", "top 25 percent", "Expend", "PhD" and "Terminal" degree,

4. throughout all other components also above listed Fields are playing Vital role in the College information.

5. Since PhD and terminal degree of the teacher are playing good role in each component, we can say that Student prefer to Stick with college with good faculty

6. Though 1st and 2nd Component always have major roles from most of the columns, but when we analyze remaining 6 component out 8 chosen components, we can see that "Books" , "Personal" and "Room Board" also plays major role in information and , might be affecting the reason for choosing the colleges.