Business Report

***Sanjay Srinivasan***

***PGP-DSBA Online***

***JULY’ 21 Batch***

***Date: 09-10-2021***

# *INDEX*

|  |  |  |
| --- | --- | --- |
| ***S. No*** | ***Contents*** | ***Page* No** |
| **1.** | **Problem - 1** | 4 |
|  | Summary | 4 |
|  | Introduction | 4 |
|  | Data Description | 4 |
|  | Sample Dataset | 4 |
|  | Exploratory Data Analysis | 4 |
|  | Checking for missing values in the dataset | 5 |
|  | **Problem 1A** | 5 |
|  | 1) State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually. | 5 |
|  | 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 |
|  | 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 |
|  | 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) | 6 |
|  | **Problem 1B** | 7 |
|  | 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 ‘pointplot’ function from the ‘seaborn’ function] | 7 |
|  | 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? | 7 |
|  | 3) Explain the business implications of performing ANOVA for this particular case study. | 8 |
|  |  |  |
| **2.** | **Problem - 2** | 9 |
|  | Summary | 9 |
|  | Introduction | 9 |
|  | Data Description | 9 |
|  | Sample Dataset | 10 |
|  | Exploratory Data Analysis | 10 |
|  | Checking for missing values in the dataset | 10 |
|  | 2.1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA? | 11 |
|  | 2.2. Is scaling necessary for PCA in this case? Give justification and perform scaling. | 13 |
|  | 2.3. Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data]. | 14 |
|  | 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] | 15 |
|  | 2.5. Extract the eigenvalues and eigenvectors.[Using Sklearn PCA Print Both] | 15 |
|  | 2.6. Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features | 17 |
|  | 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] | 17 |
|  | 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? | 19 |
|  | 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]. | 20 |

# *List Of Tables*

|  |  |  |
| --- | --- | --- |
| ***S.No*** | ***Content*** | ***Page No*** |
| 1.1 | Dataset sample | 4 |
| 1.2 | Datatypes of the variable | 4 |
| 1.3 | Check null values | 5 |
| 1.4 | One-way Anova for Salary vs. Education | 6 |
| 1.5 | One-way Anova for Salary vs. Occupation | 6 |
| 1.2 | Items across regions and channels | 8 |
| 2.1 | Dataset sample | 9 |
| 2.2 | Datatypes of the variable | 10 |
| 2.3 | Check null values | 10 |
| 2.2.1 | Data Before scaling | 13 |
| 2.2.2 | Data After scaling | 13 |
| 2.3.1 | Covariance matrix | 14 |
| 2.3.2 | Correlation matrix | 14 |
| 2.6 | Eigen vector as Dataframe | 17 |
| 2.7.1 | Eigen vector dataframe first row | 17 |
| 2.7.2 | Explicit form of first PC | 18 |
| 2.7.3 | PC values | 18 |
| 2.9.1 | Sample data of new dataframe after PCA | 20 |
| 2.9.2 | Data description of new dataframe after PCA | 21 |

***List Of Figures***

|  |  |  |
| --- | --- | --- |
| ***S.No*** | ***Content*** | ***Page No*** |
| 1.1.1 | Hypothesis for Salary vs. Education | 5 |
| 1.1.2 | Hypothesis for Salary vs. Occupation | 5 |
| 1.4 | Pointplot for Salary vs. Education | 6 |
| 1.5 | Pointplot for Education vs. Salary vs. Occupation | 7 |
| 1.6.1 | Hypothesis for two way anova | 7 |
| 1.6.2 | probability of interaction between Occupation and Education | 7 |
| 2.1.1 | Univariate analysis distplot and boxplot | 11 |
| 2.1.2 | Multivariate analysis of pairplot | 12 |
| 2.1.3 | Multivariate analysis heatmap | 13 |
| 2.4.1 | Boxplot before scaling | 15 |
| 2.4.2 | Boxplot after scaling | 15 |
| 2.5.1 | Output for probability of Bartletts Test | 15 |
| 2.5.2 | Output for probability of KMO Test | 15 |
| 2.5.3 | Output for PCA component | 16 |
| 2.5.4 | Output for Eigen vector | 16 |
| 2.5.5 | Output for Eigen value | 16 |
| 2.5.6 | Scree Plot for Eigen value | 17 |
| 2.7.1 | Heatmap for PC values | 19 |
| 2.8.1 | Cumulative Eigen value | 19 |
| 2.8.2 | Scree Plot for Cumulative Eigen value | 20 |

# Problem - 1

*Summary*

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals 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*

The purpose of this exercise is to explore the dataset across the mean difference of two or more independent variable. The exploratory data analysis of the dataset as the salary details across the individual with different educational and occupation. From this three different education level has different occupation level has the salary difference. This assignment helps in exploring the dataset with anova test.

*Data Description*

1. Education: Individual with three different level of education.

2. Occupation: Individual with four different level of occupation.

3. Salary: Salary details for every individual.

*Sample of the dataset:*

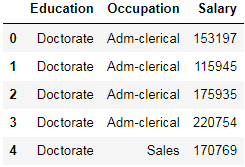


Table 1.1 Dataset Sample

Dataset has 3 variables with 2 different types of the variable. Each variable has different level of education and occupation. Based on the characteristic salary of each individual on each individual is defined.

*Exploratory Data Analysis*

*Let us check the types of variables in the data frame.*

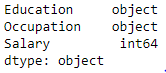


Table- 1.2. Datatypes of the variable

There are total 40 rows and 3 columns in the dataset. Out of 3, 2 columns (Education and Occupation) are of object type and rest is of either integer data type.

*Check for missing values in the dataset:*

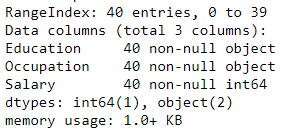


Table- 1.3. Check null values

### Problem 1A:

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

For Education:

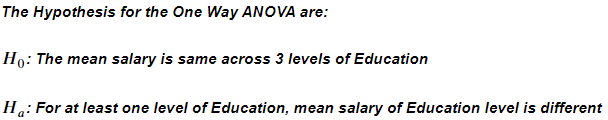


Fig – 1.1.1 Hypothesis for Salary vs. Education

For Occupation:

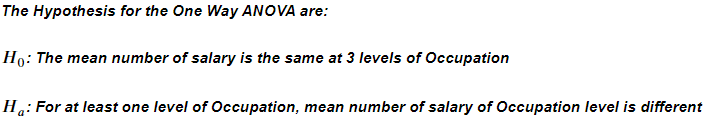


Fig – 1.1.2 Hypothesis for Salary vs. Occupation

From this, Null hypothesis and the Alternative hypothesis are framed for the Salary against Education and Salary against Occupation.

### 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.

The one-way anova for salary with Education is framed.

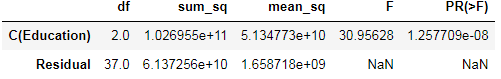


Table – 1.4 One-way Anova for Salary vs. Education

Now, we see that the corresponding p-value (1.257709e-08) is less than alpha (0.05). Thus, we 𝐣𝐞𝐜𝐭 the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 ( 𝐻0).

### 1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

The one-way anova for salary with Occupation is framed.

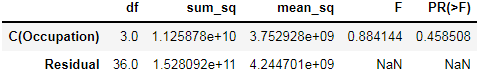


Table – 1.5 One-way Anova for Salary vs. Occupation

Now, we see that the corresponding p-value (0.458508) is greater than alpha (0.05). Thus, we 𝐟𝐚𝐢𝐥 𝐭𝐨 𝐫𝐞𝐣𝐞𝐜𝐭 the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 ( 0 ).

### 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)

From the one-way anova result.

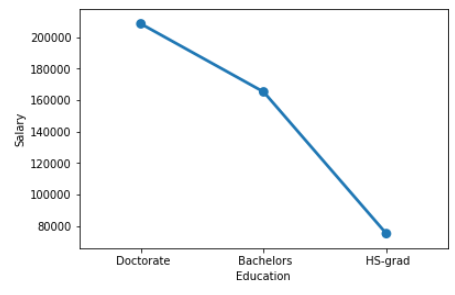


Fig – 1.4 Pointplot for Salary vs. Education

The null hypothesis is rejected for Salary Vs. Education. From this above Pointplot, the Mean are significantly different for all classes.

### Problem 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 ‘pointplot’ function from the ‘seaborn’ function]

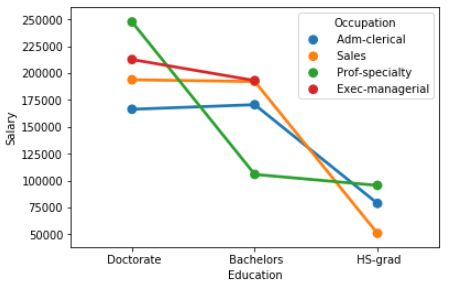


Fig – 1.5 Pointplot for Education vs. Salary vs. Occupation

Inference from the above pointplot, Doctorate education in the prof-speciality has the highest salary of 250000, whereas HS-grad education with the sales occupation has the least salary of around 50000.Bachelors education with the occupation (Sales and Exec-managerial) has the same salary of 200000.

**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?**

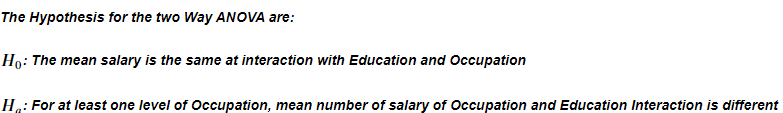


Fig – 1.6.1 Hypothesis for two way anova

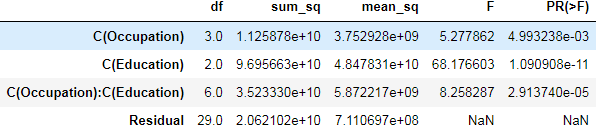


Fig – 1.6.2 probability of interaction between Occupation and Education

The p-Value (2.913740e-05) is less than the alpha (significant value - 0.05). We failed to reject the null hull hypothesis. The mean salary of interacted Education and occupation is same.

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

From the ANOVA method and the interaction plot, we see that education combined with occupation results in higher and better salaries among the people. It is clearly seen that people with education as Doctorate draw the maximum salaries and people with education with HS-grad earn the least. Thus, we can conclude that Salary is dependent on educational qualifications and occupation.

# Problem – 2

*Summary*

The data is gathered about the students, joined at different College/University after completing the12th grade. This dataset consist of data of 777 students who have enrolled for the different university. In this problem statement, we will explore the dataset and perform Principle Component Analysis (PCA).

*Introduction*

The purpose of this exercise is to perform PCA by reducing the dimensionality without losing the data. The Principle Component Analysis of this dataset will reduce the dimensionality by reducing the variable of the dataset, while preserving as much information as possible. This dataset consist of 777 rows and 18 columns, by reducing dimensionality, for eg.(from 3d to 2d) without losing of data.

*Data Description*

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

*Sample of the dataset:*

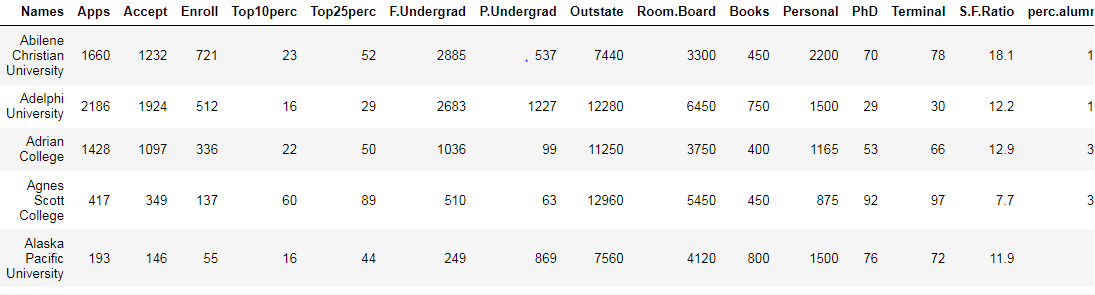


Table 2.1 Dataset Sample

Dataset has 18 variables with student details. Based on the Student details, who have enrolled in different colleges is defined.

*Exploratory Data Analysis*

*Let us check the types of variables in the data frame.*

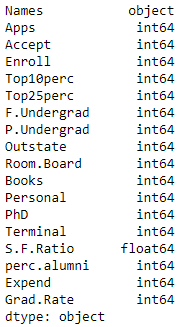


Table 2.2 Datatypes of the variable

There are total 777 rows and 18 columns in the dataset. Out of 18, 1 column is of object type, 1 column is of float (Decimal value) type and rest 16 are of integer data type.

*Check for missing values in the dataset:*

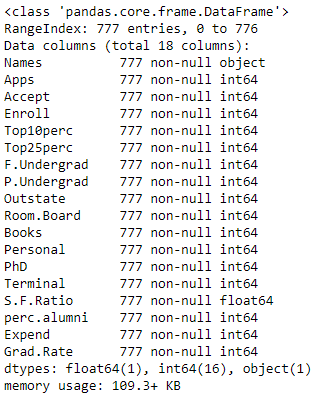


Table 2.3 Check null values

From this, it is clear that there are no null values present in the dataset.

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

#### Univariate Analysis:

Univariate analysis is the simplest form of analysing data. Analysing each variable in detail.

Insights :

Both Enroll and Accept rate have outliers in upper values.

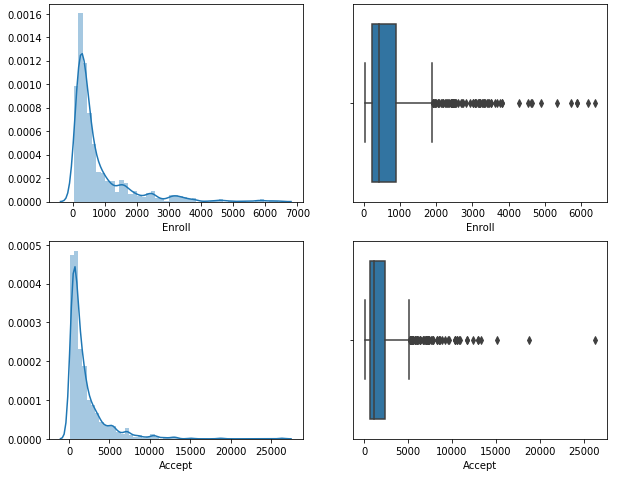


Fig – 2.1.1 Univariate analysis distplot and boxplot

**Multivariate Analysis:**

Analysing the data with two variables.

Insights :

There is a strong correlation observed between few fields. 'Apps' is highly correlated to 'Accept' and 'Enroll'

Also, 'Apps' shows high correlation with 'F.undergrad'

Whereas, 'S.F.Ratio' shows least correlation with 'Outstate'

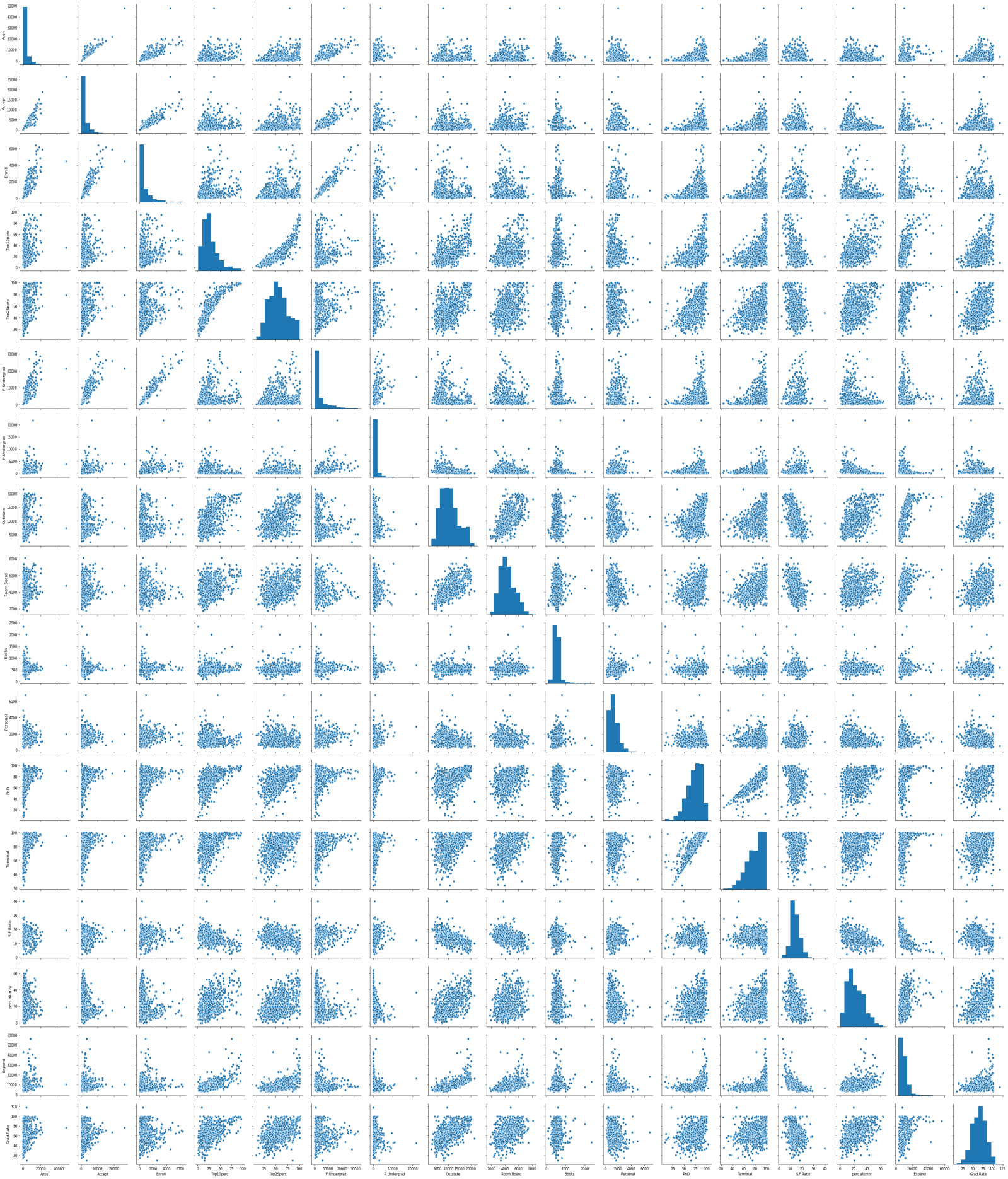


Fig – 2.1.2 Multivariate analysis of pairplot

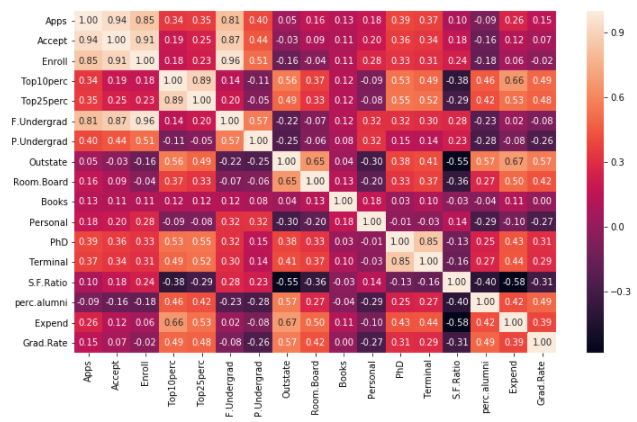


Fig – 2.1.3 Multivariate analysis heatmap

**2.2. 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. ... If you normalize your data, all variables have the same standard deviation, thus all variables have the same weight and your PCA calculates relevant axis.

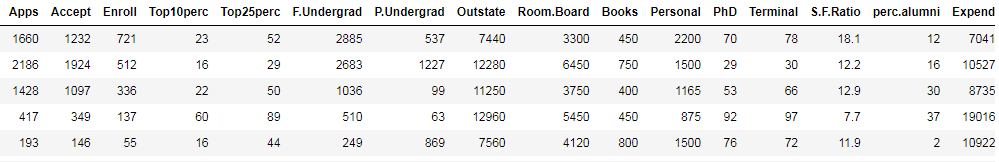


Table 2.2.1 Data Before scaling

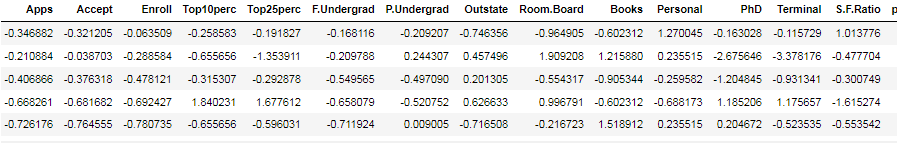


Table 2.2.2 Data after scaling

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

Both covariance and correlation measure the relationship and the dependency between two variables.

|  |  |
| --- | --- |
| Covariance | Correlation |
| 1. Covariance indicates the direction of the linear relationship between variables. | 1. Correlation measures both the strength and direction of the linear relationship between two variables. |
| 2. Covariance values are not standardized. | 2. Correlation values are standardized. |
| 3. Covariance can vary between - infinity to + infinity. | 3. Correlation can vary between - 1 to + 1. |

***Covariance Matrix:***

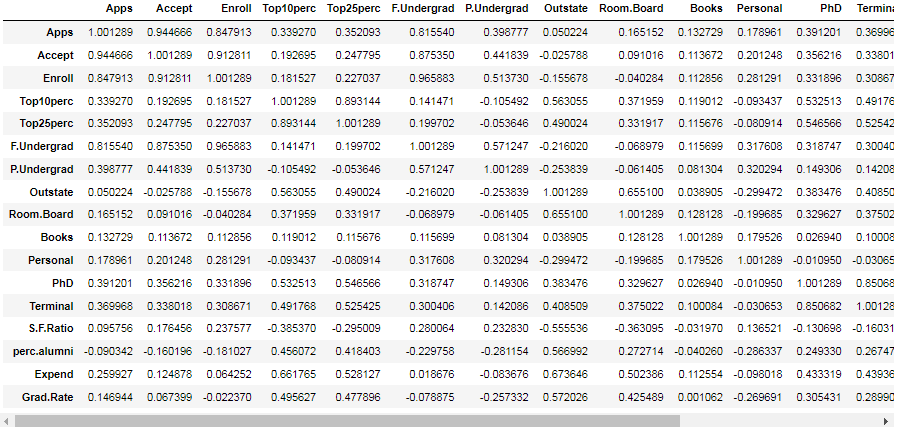


Table 2.3.1 covariance matrix

***Correlation Matrix:***



Table 2.3.2 correlation matrix

**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]**

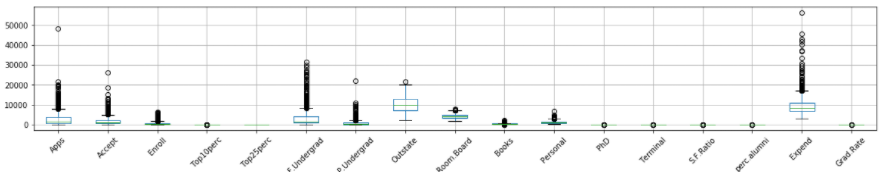


Fig – 2.4.1 Boxplot before scaling

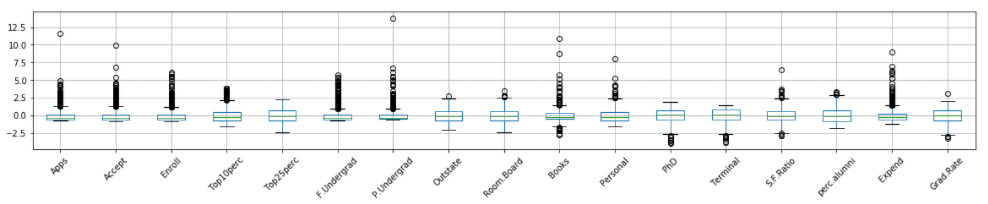


Fig – 2.4.2 Boxplot after scaling

The range in y axis is in 10000 difference in the original dataframe, whereas the range of scaled data where in 2.5 difference. The data are in the different scaling. So, the values in 'Top10perc' and 'Top25perc' box plot is not visible in the original dataframe, whereas, after scaling all the values are converted to the same range.

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

**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 is correlated.

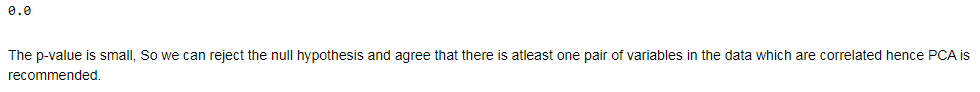


Fig – 2.5.1 Output for probability of Bartletts Test

**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.

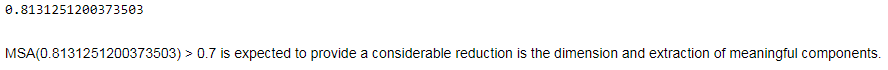


Fig – 2.5.2 Output for probability of KMO Test

From this Bartletts test and KMO test, the data is sufficient and good to perform PCA.

The Component Output is

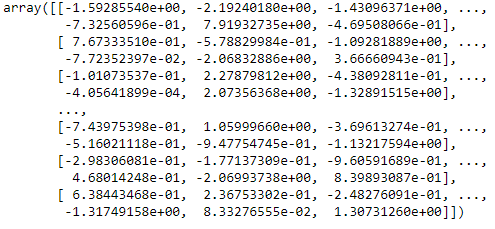


Fig – 2.5.3 Output for PCA component

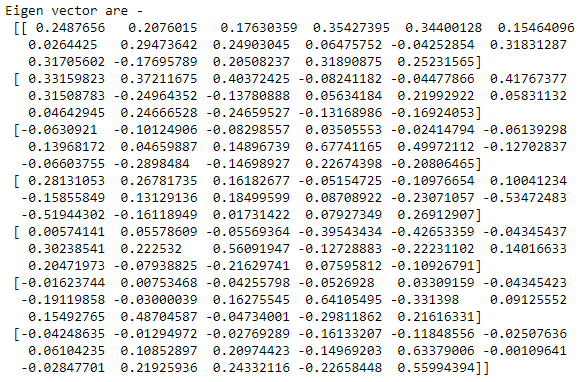


Fig – 2.5.4 Output for Eigen vector



Fig – 2.5.5 Output for Eigen value

Eigen values are plotted as Scree plot. From Screeplot we can infer that around 85% of data lies in the first 8 components of the Eigen values in PCA

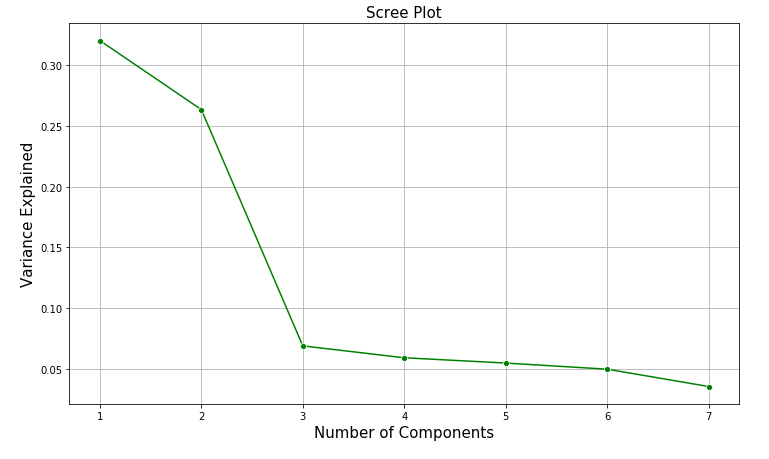


Fig – 2.5.6 Scree Plot for Eigen value

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

To perform PCA, All the Eigen vectors are loaded into a dataframe with 7 rows and 17 columns (excluding of categorical column from the original dataframe)

The Eigen vectors are converted as the dataframe. Sample dataframe image is attached below.

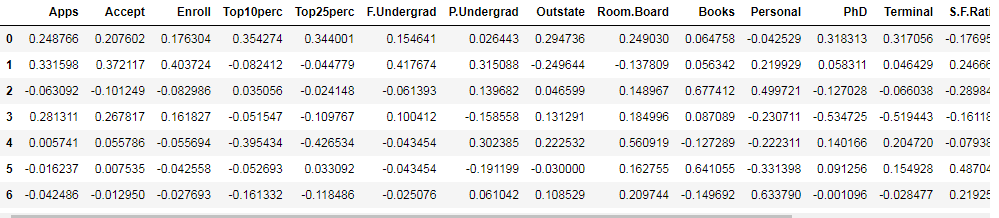


Table 2.6 Eigen vector as Dataframe

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]

First row from the Eigen vector dataframe



Table 2.7.1 Eigen vector dataframe first row

**Linear equation formula:**

PC = m1x1+m2x2+m3x3+....m17x17

Where,

m - Represents Eigen vector

x – Represents the variable

**Explicit Form of first PC:**

PC1= 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

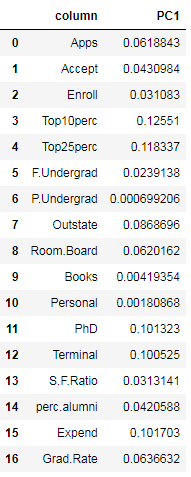


Table 2.7.2 Explicit form of first PC

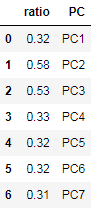


Table 2.7.3 PC values

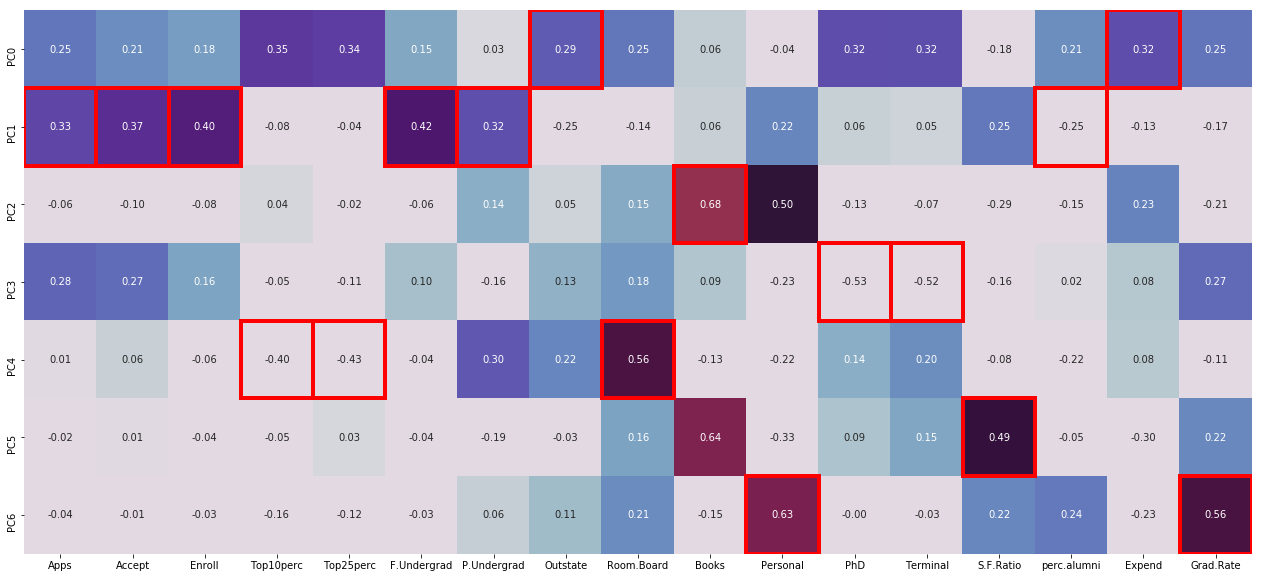


Fig – 2.7.1 Heatmap for PC values

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?

The cumulative value up to seventh Principal Component is 85.21. General rule of thumb is to choose first k PC’s such that the first k PC’s explaining 70-90% of the total variance. Hence from the cumulative values of Eigen values, help in selecting the required no. of PC’s. In this case first seven PC’s have been selected capturing 85.2% of variation and thereby reducing our dimension by half.

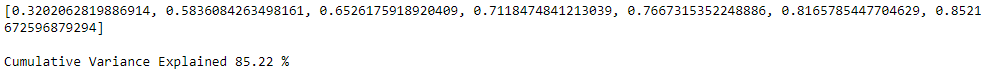


Fig – 2.8.1 Cumulative Eigen value

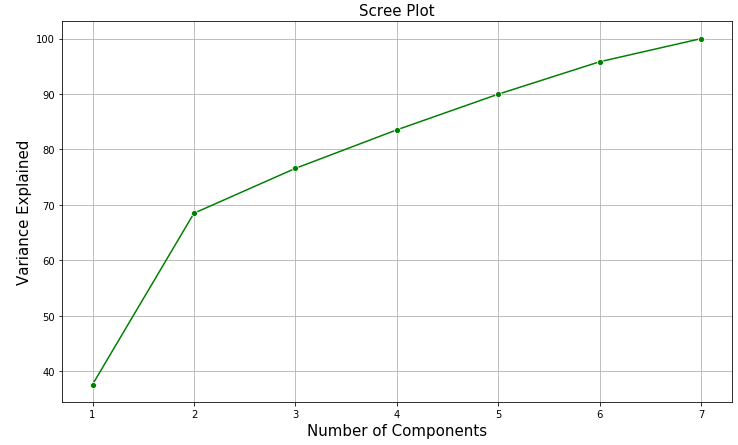


Fig – 2.8.2 Scree Plot for Cumulative Eigen value

The above Scree plot shows, the sum of Eigen values are added and plotted in the graph.

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]

PCA is a “dimensionality reduction” method. It reduces the number of variables that are correlated to each other into fewer independent variables without losing the essence of these variables. It provides an overview of linear relationships between inputs and variables.

***Business Implication:***

From this dataset, 17 continuous variables have been reduced to 7 variables without losing of the data. The output of the PCA from this dataset, after performing the PCA 85% data have been is available in the 7 variable by reducing the dimensionality.

After completing the PCA, 7 PC's are used for the further process.

The sample new dataframe after concatenating the categorical and numerical ( After PCA) dataframe.

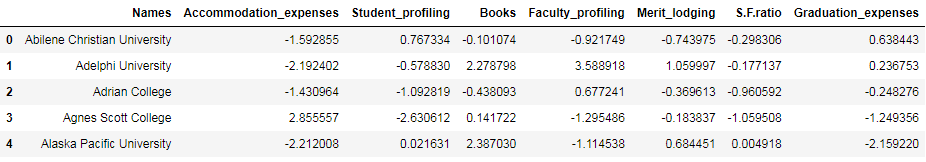


Table 2.9.1 Sample data of new dataframe after PCA

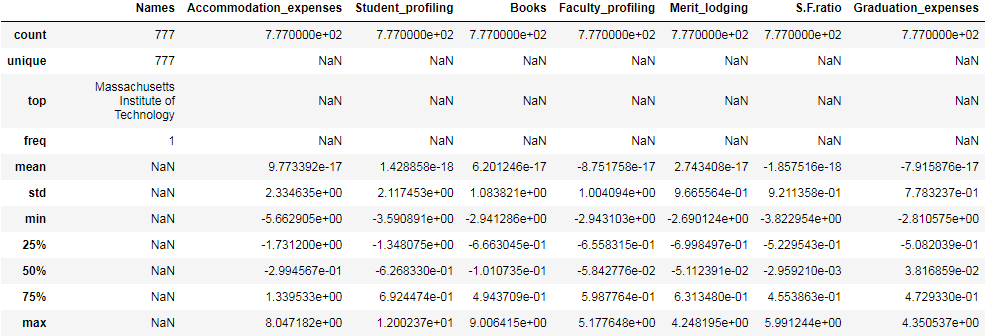


Table 2.9.2 Data description of new dataframe after PCA