**ADVANCE STATISTICS PROJECT**

Text

Description automatically generated

BUSINESS REPORT – By Tribid Maji

**CONTENTS**

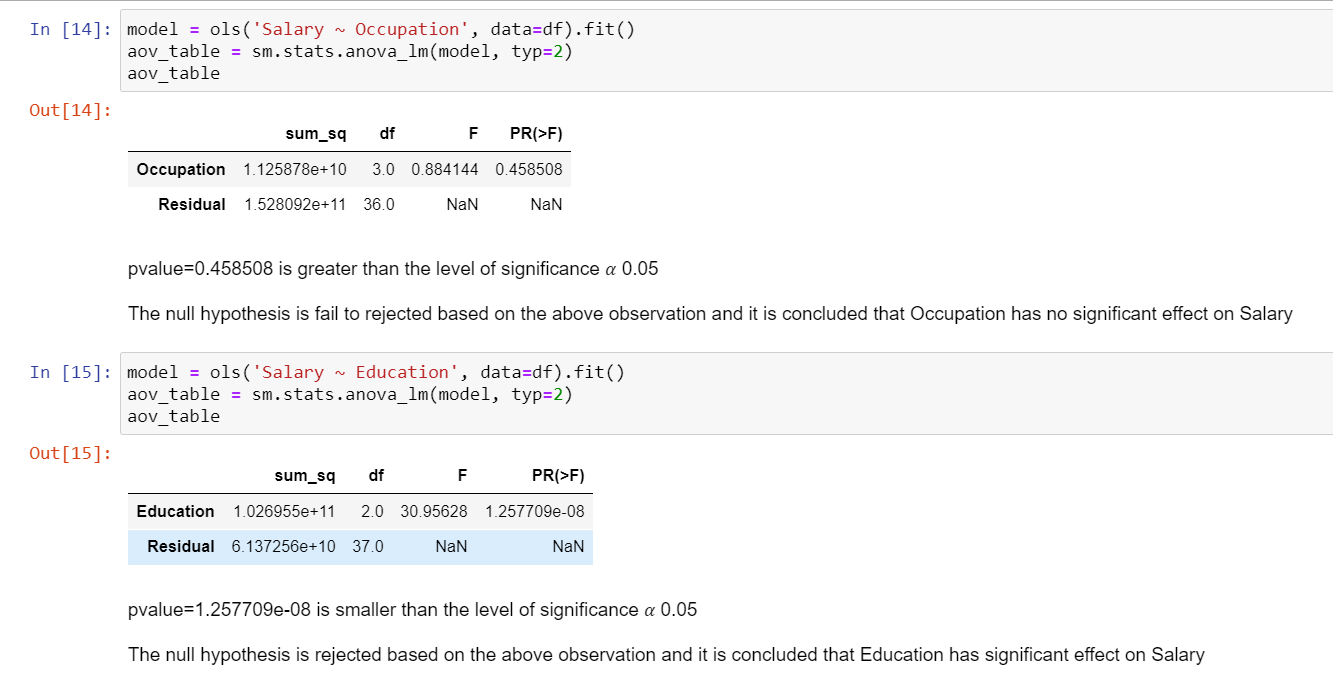
1. **Problem-1**
   1. **State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.**
   2. **Perform one-way ANOVA for Education with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results.**
   3. **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.**
   4. **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.**
   5. **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.**
   6. **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?**
   7. **Explain the business implications of performing ANOVA for this particular case study.**
2. **Problem-2**
   1. **Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?**
   2. **Is scaling necessary for PCA in this case? Give justification and perform scaling.**
   3. **Comment on the comparison between the covariance and the correlation matrices from this data.**
   4. **Check the dataset for outliers before and after scaling. What insight do you derive here?**
   5. **Perform PCA and export the data of the Principal Component scores into a data frame.**
   6. **Extract the eigenvalues, and eigenvectors.**
   7. **Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).**
   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?**
   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]**
3. **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.greatlearning.in/courses/32927/files/2327652/download?verifier=gxg4JnRexuRK727ef3gK5BVK9FiVCYUocjDfcAsM&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.**

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

Answer:

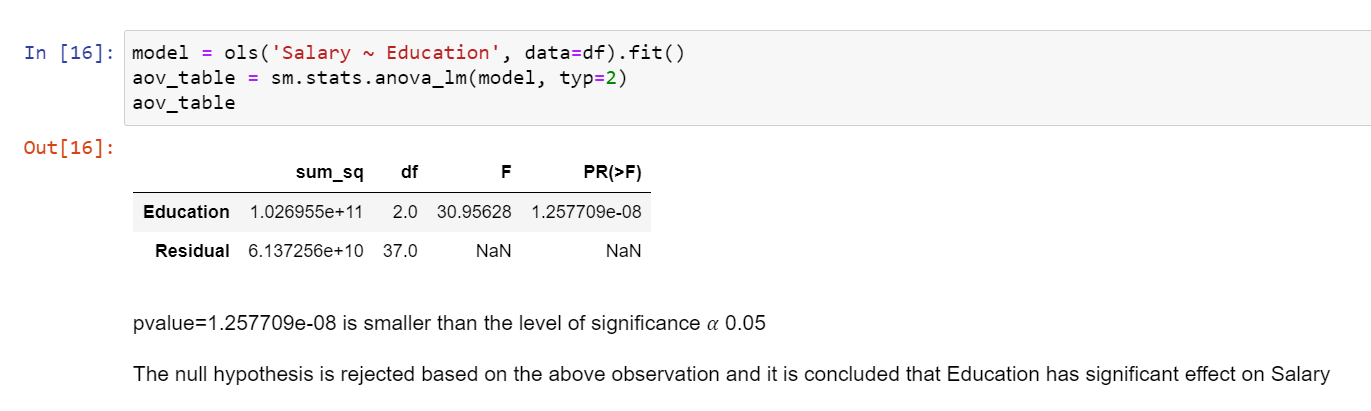
Null hypothesis = H0: Salary is depend on educational qualification and occupation. Alternate hypothesis = H1: Salary is not depend on educational qualification and occupation ;i.e. Salary is dependent on either one of the component(Education or Occupation).



* 1. **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.**

Answer: pvalue=1.257709e-08 is smaller than the level of significance 𝛼 0.05

The null hypothesis is rejected based on the above observation and it is concluded that Education has significant effect on Salary

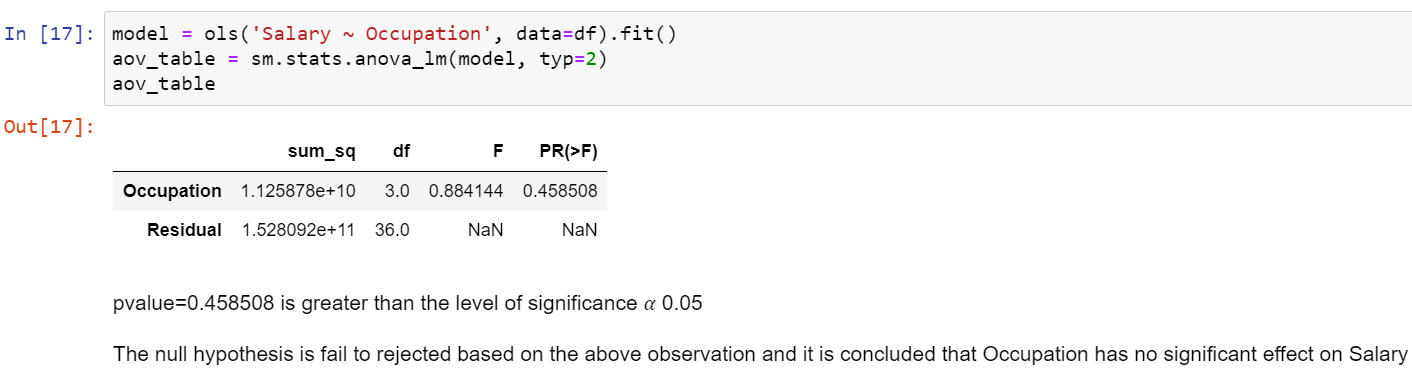


* 1. **Perform one-way ANOVA for variable Occupation with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results.**

Answer: pvalue=0.458508 is greater than the level of significance 𝛼 0.05

The null hypothesis is fail to rejected based on the above observation and it is concluded that Occupation has no significant effect on Salary

​



* 1. **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.**

Answer:

The null hypothesis is rejected based on the above observation and it is concluded that,

Education has significant effect on Salary

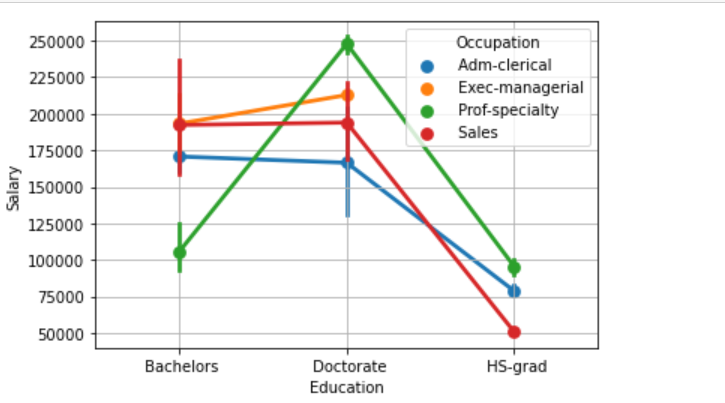
#The null hypothesis is fail to rejected based on the above observation and it is concluded that, Occupation has no significant effect on Salary

* 1. **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.**

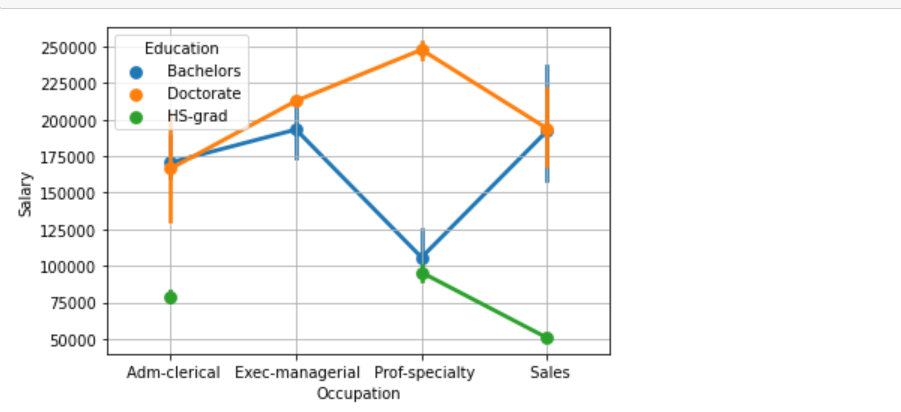
Answer:

As seen from the above two interaction plots, there seems to be very interaction amongst the two categorical variables.

Education:



Occupation:



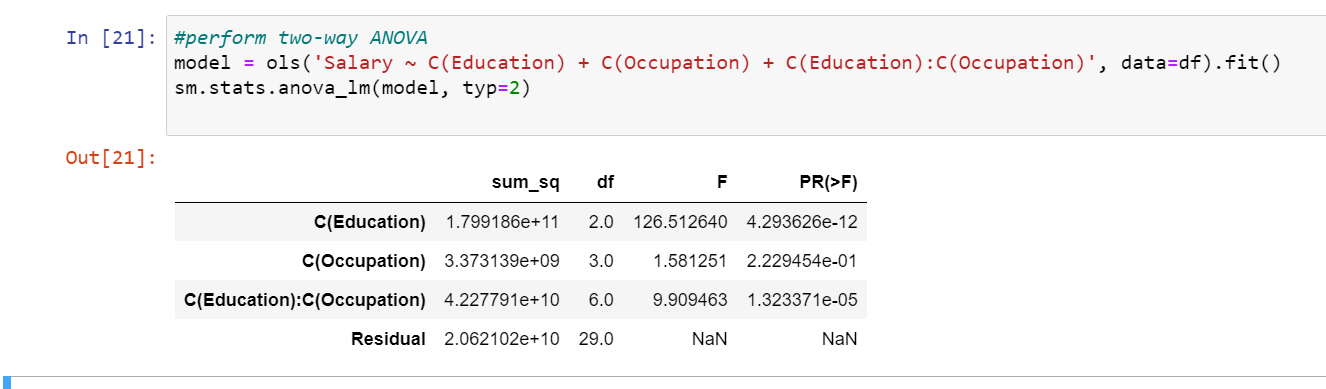
* 1. **Perform a two-way ANOVA based on the Education and Occupation (along with their interaction Education\*Occupation) with the variable ‘Salary’. State the null and alternative hypotheses and state your results. How will you interpret this result?**

Answer:

#As Education and Occupation interaction is 1.323371e-05 which is <0.05 , there seems to be very statistical interaction has significant impact on Salary.

#Since the p-value for Education is 4.293626e-12 lesser than .05, this means the factor have a statistically significant effect on Salary

#Since the p-value for Occupation is 2.229454e-01 greater than .05, this means the factor have a statistically no significant effect on Salary.



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

Answer:

#As Education and Occupation interaction is 1.323371e-05 which is <0.05 , there seems to be very statistical interaction has significant impact on Salary.

#Since the p-value for Education is 4.293626e-12 lesser than .05, this means the factor have a statistically significant effect on Salary

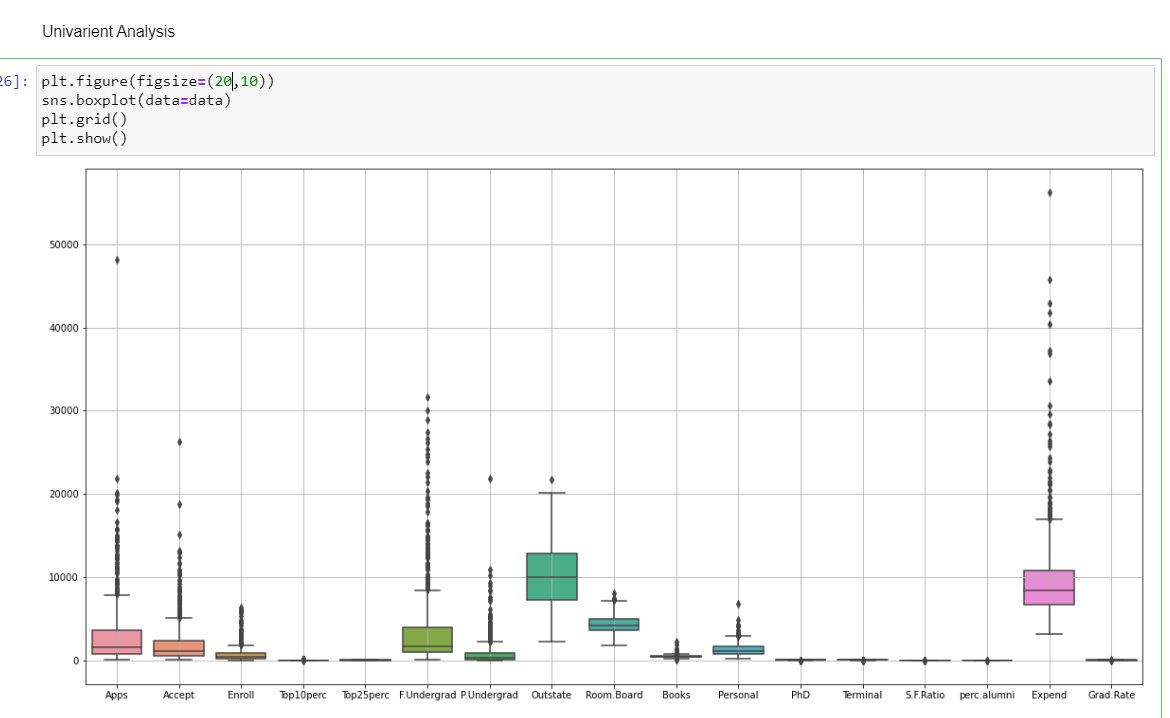
#Since the p-value for Occupation is 2.229454e-01 greater than .05, this means the factor have a statistically no significant effect on Salary.

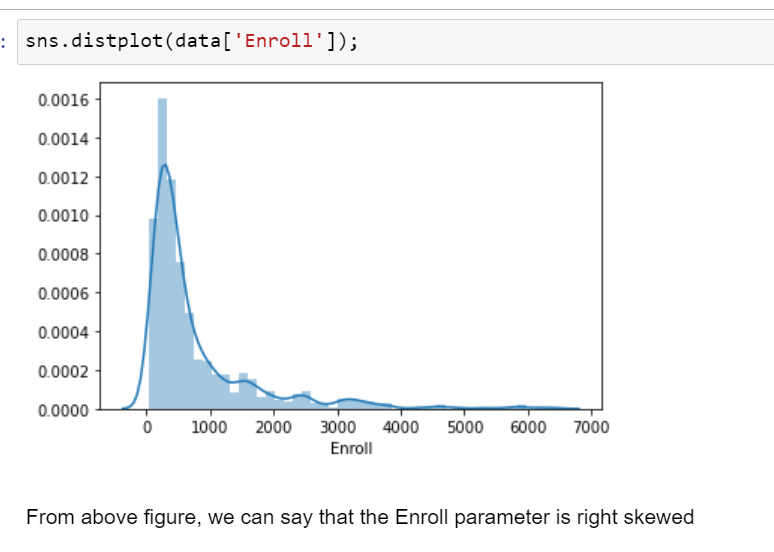
1. **PROBLEM – 2**

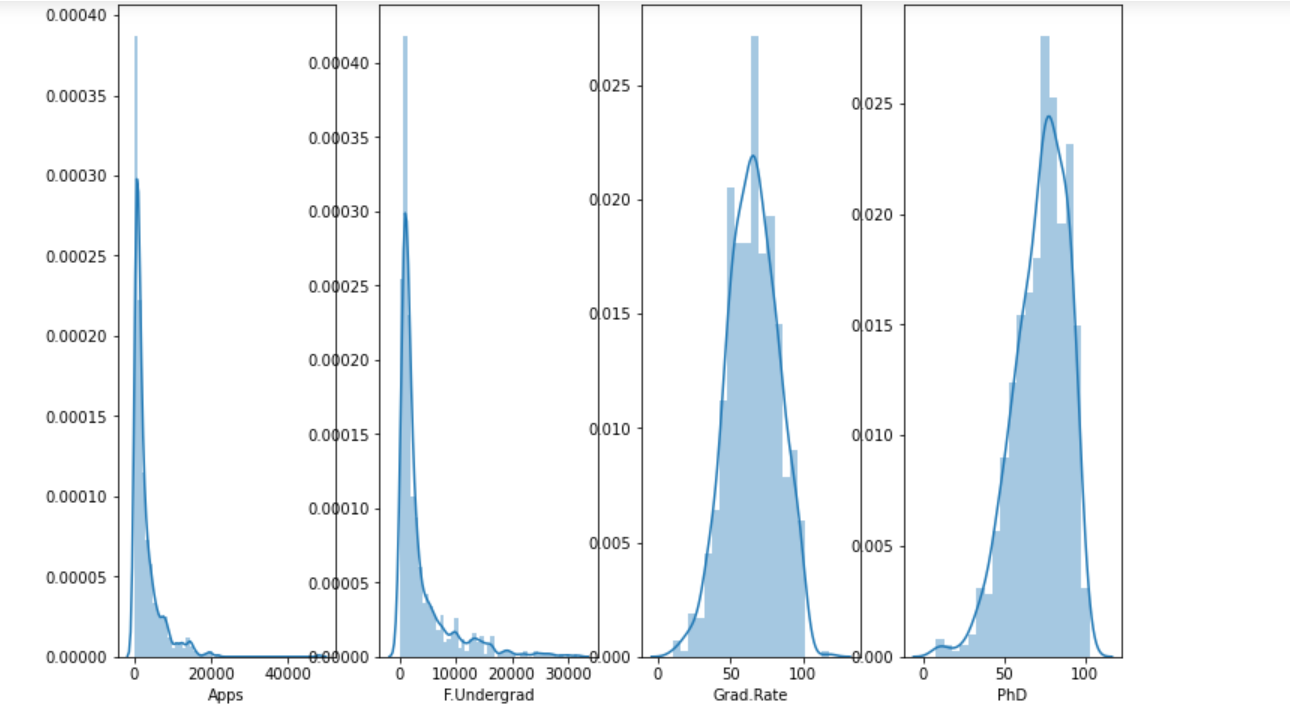
The dataset [Education - Post 12th Standard.csv](https://olympus.greatlearning.in/courses/32927/files/1735740/download?verifier=EEok57ZhKNfq0cklE2ZD935qD1Xamocut6gvOU4Z&wrap=1) contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according 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/32927/files/1735739/download?verifier=6znIBvbflHKvlMmOfntBBxlvaDLUiRcNAew7tUQV&wrap=1).

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

Answer:



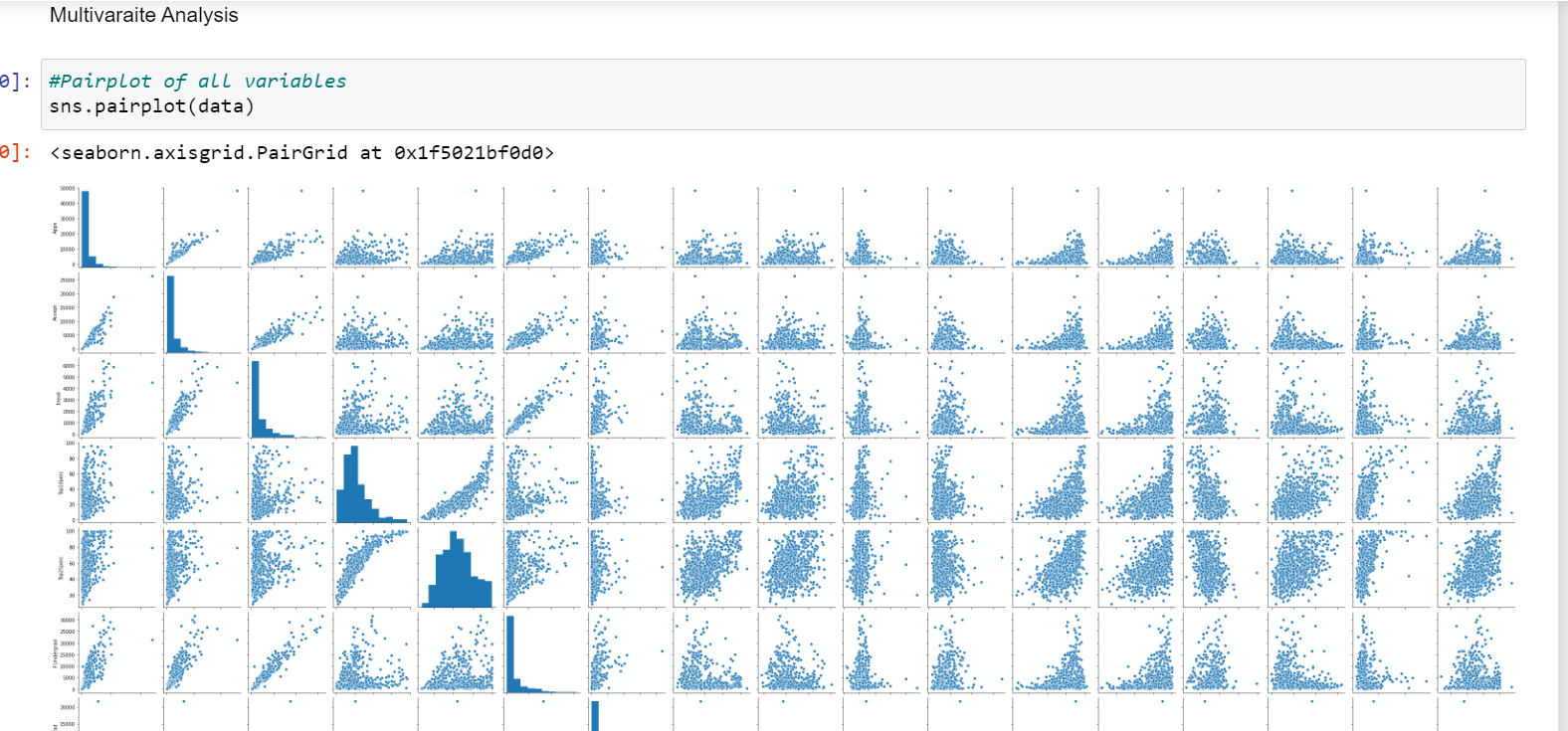


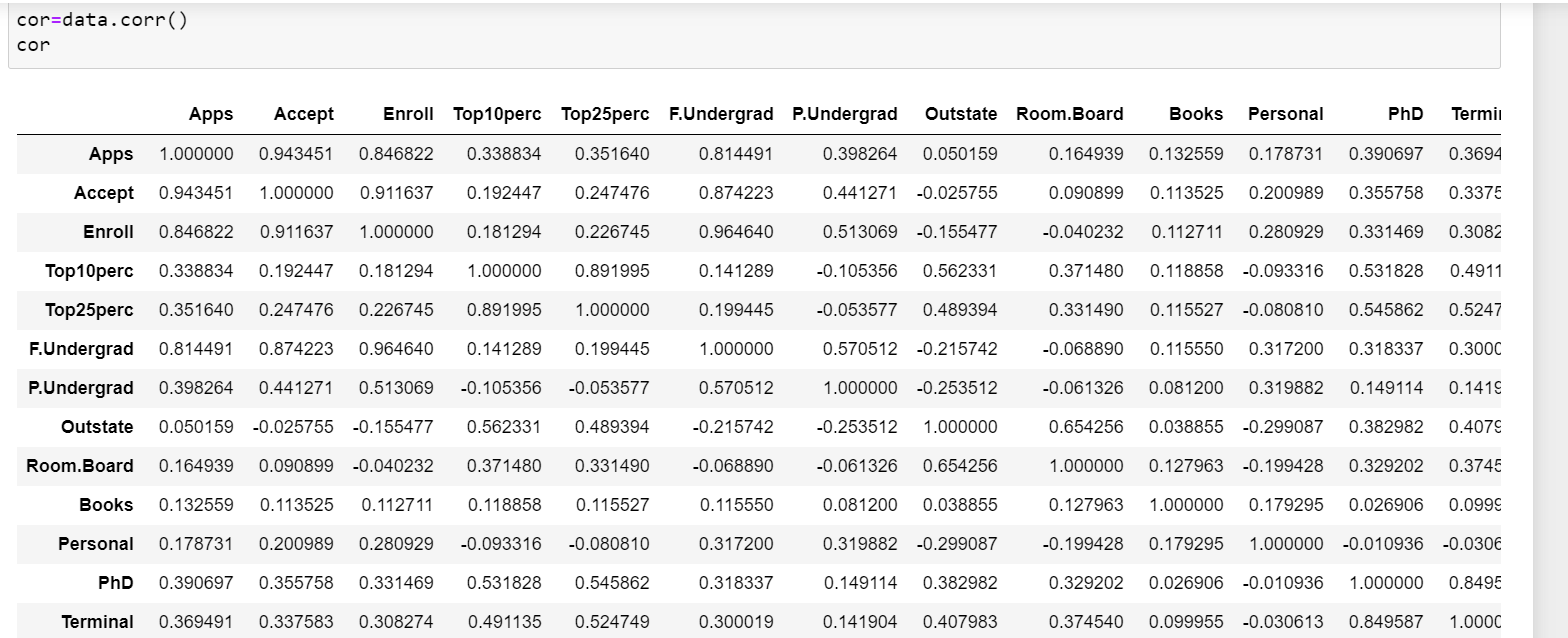


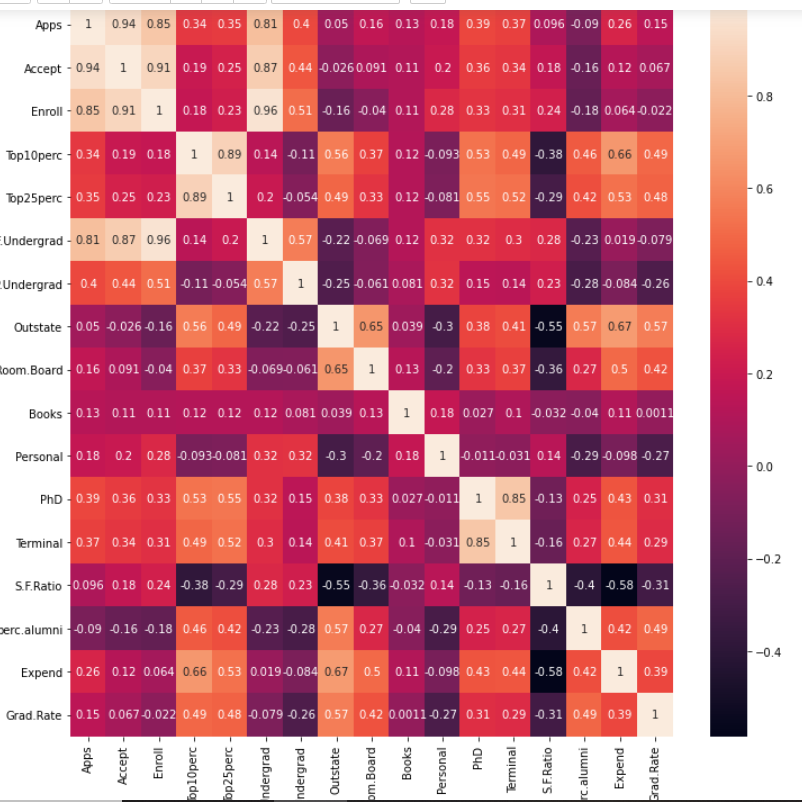
#form the above figure we can conclude that "Applications Received","Full time Under Grad" are right skewed

#form the above figure we can conclude that "PhD" are Left skewed

#from the above figure we can conclude the "Graduation Rate" is Normally distributed







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. The pair plot function in seaborn makes it very easy to generate joint scatter plots for all the columns in the data.

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

Answer:

Often the variables of the data set are of different scales i.e. one variable is in millions and other in only 100.

#in this data set 3 different types of values are there 1)Number(4 to 5 digit) 2)Ratio(2 digit with deciaml) 3)Percentage(2 digit).Since the data in these variables are of different scales, it is tough to compare these variables.

Number: Apps,Accept,Enroll,F.Undergrad,P.Undergrad,Outstate,Room.Board,Books,Personal. Percentage :Top10perc,Top25perc,PhD,Terminal,perc.alumni. Ratio:Grad.Rate,

#Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

In this method, we convert variables with different scales of measurements into a single scale.

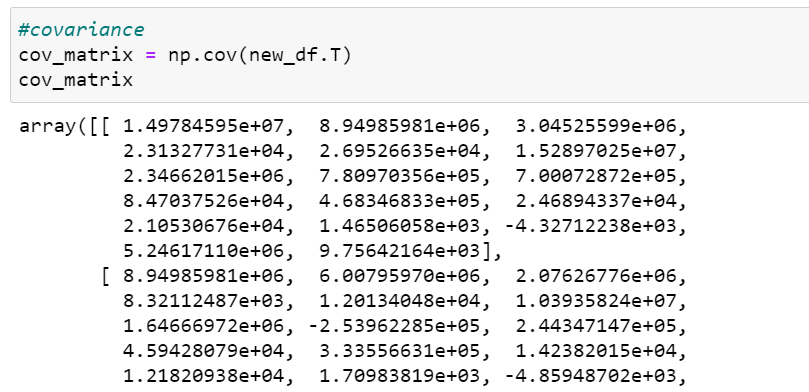
StandardScaler normalizes the data using the formula (x-mean)/standard deviation.

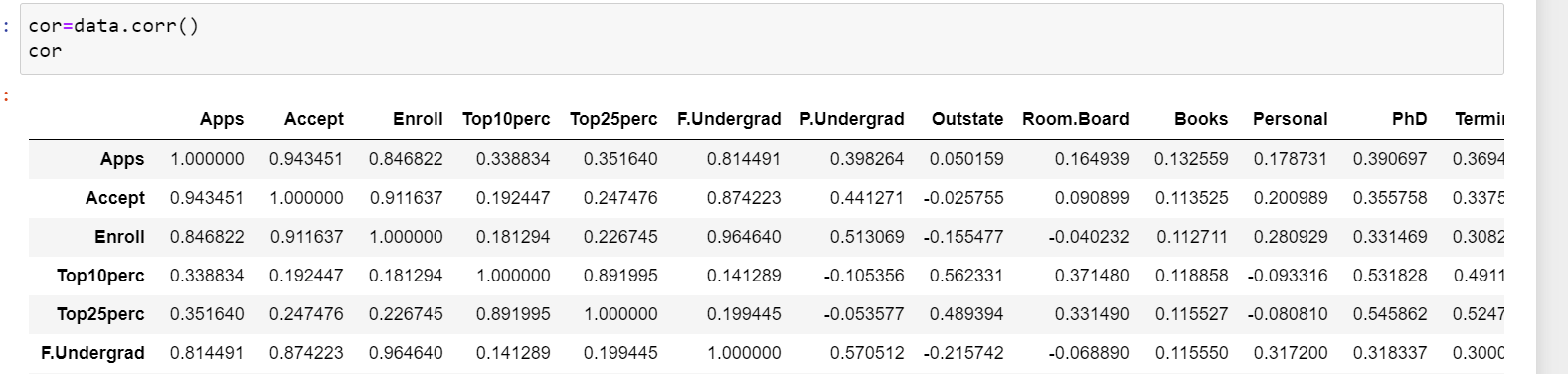
* 1. **Comment on the comparison between the covariance and the correlation matrices from this data.**

Answer:

Covariance and Correlation are two mathematical concepts which are quite commonly used in statistics. Both of these two determine the relationship and measures the dependency between two random variables. Despite, some similarities between these two mathematical terms, they are different from each other. Correlation is when the change in one item may result in the change in the another item. On the other hand, covariance is when two items vary together. Read the given article to know the differences between covariance and correlation.

A measure used to indicate the extent to which two random variables change in tandem is known as covariance. A measure used to represent how strongly two random variables are related known as correlation. Covariance is nothing but a measure of correlation. On the contrary, correlation refers to the scaled form of covariance. The value of correlation takes place between -1 and +1. Conversely, the value of covariance lies between -∞ and +∞. Covariance is affected by the change in scale, i.e. if all the value of one variable is multiplied by a constant and all the value of another variable are multiplied, by a similar or different constant, then the covariance is changed. As against this, correlation is not influenced by the change in scale. Correlation is dimensionless, i.e. it is a unit-free measure of the relationship between variables. Unlike covariance, where the value is obtained by the product of the units of the two variables





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

Answer:

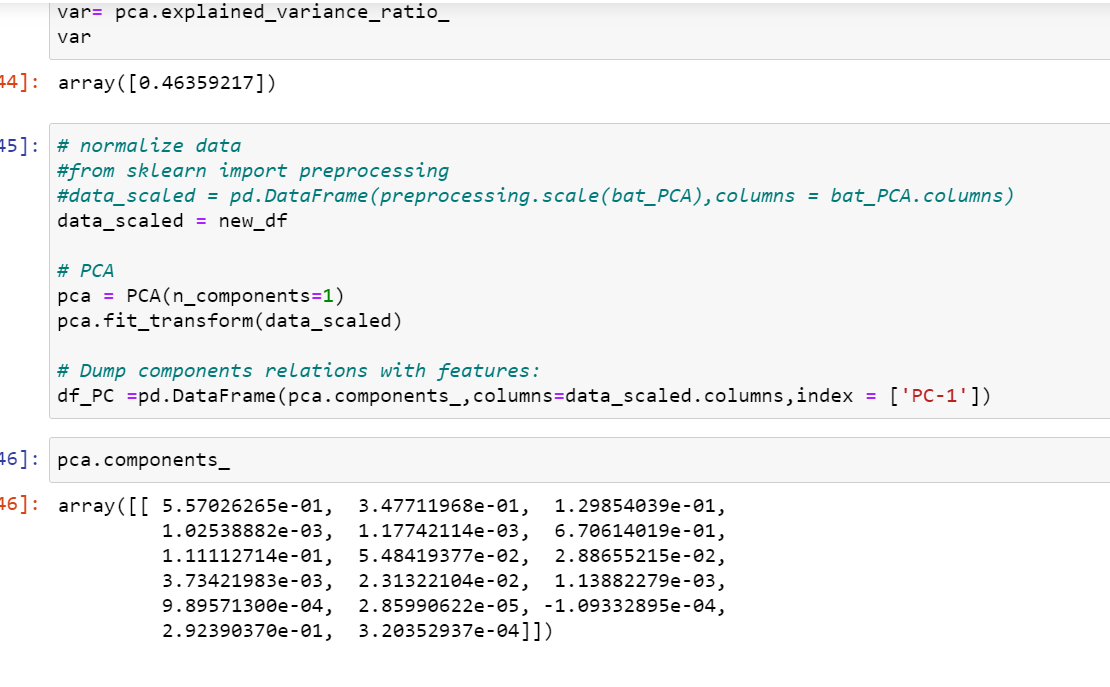
Applying zscore or using StandardScalar give us the same results

It scales the data in such a way that the mean value of the features tends to 0 and the standard deviation tends to 1

Min-Max method ensure that the data scaled to have values in the range 0 to 1

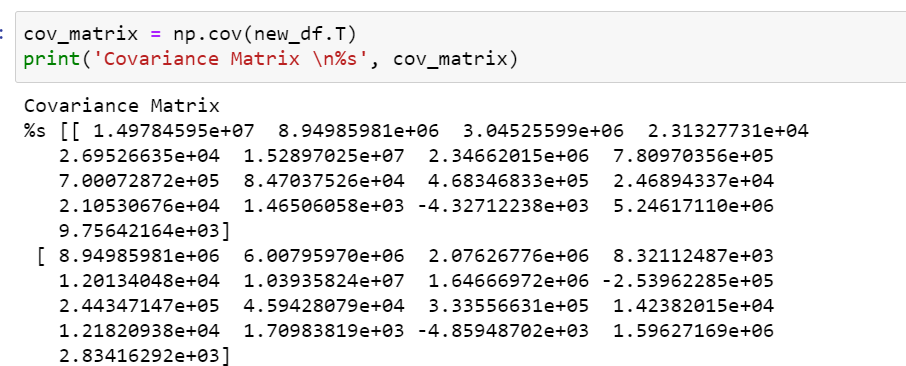
* 1. **Perform PCA and export the data of the Principal Component scores into a data frame.**

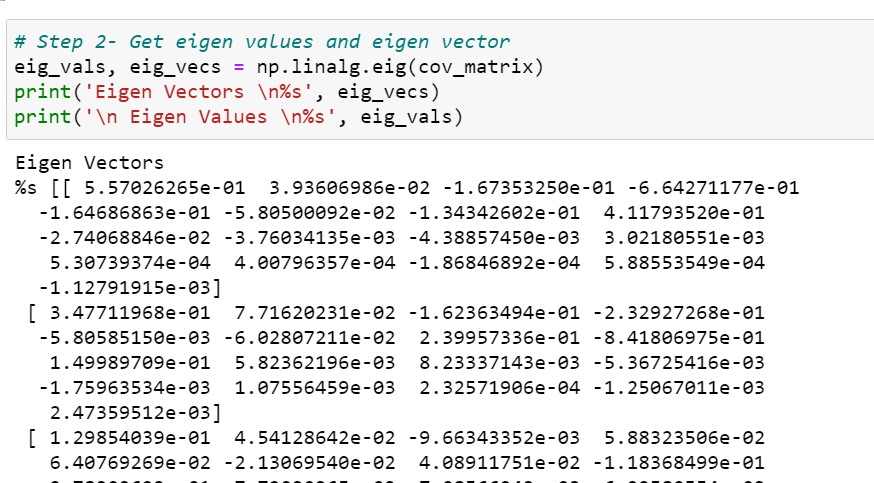
Answer:

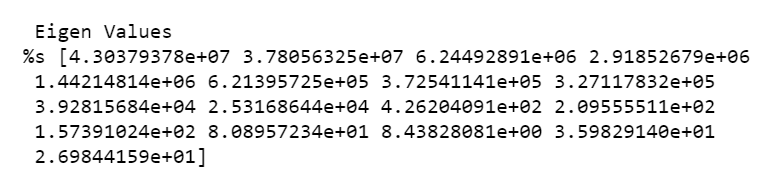


* 1. **Extract the eigenvalues, and eigenvectors.**

Answer:





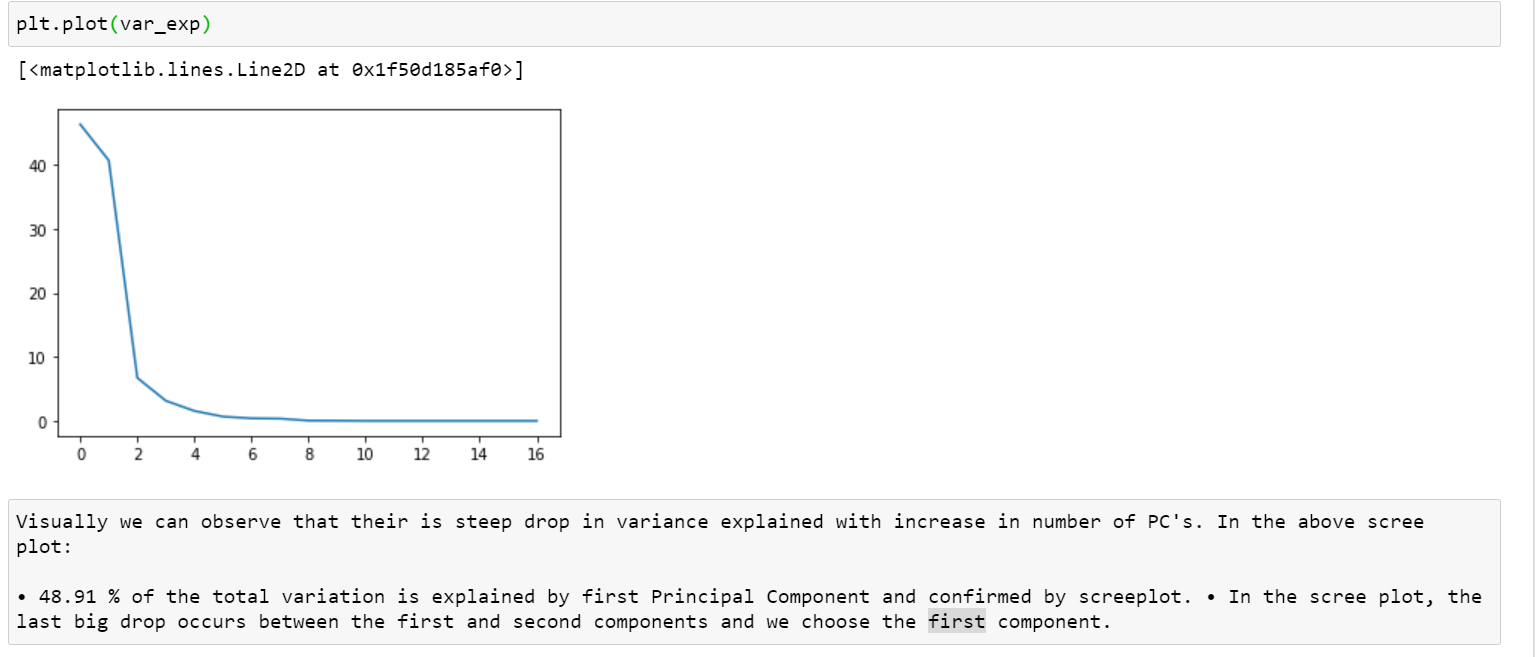


* 1. **Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).**

Answer:

Visually we can observe that their is steep drop in variance explained with increase in number of PC's. In the above scree plot:

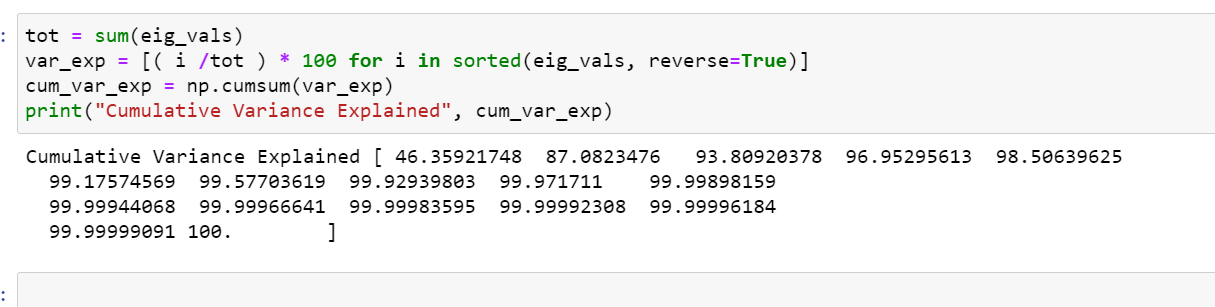
• 48.91 % of the total variation is explained by first Principal Component and confirmed by screeplot. • In the scree plot, the last big drop occurs between the first and second components and we choose the first component.



* 1. **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?**

Answer:

The **eigenvectors** and **eigenvalues** of a covariance (or correlation) matrix represent the “core” of a PCA: The **eigenvectors** (principal components) determine the directions of the new feature space, and the **eigenvalues** determine their magnitude. In other words, the eigenvalues explain the variance of the data along the new feature axes.



* 1. **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]**

Answer:

#in this data set 3 different types of values are there 1)Number(4 to 5 digit) 2)Ratio(2 digit with deciaml) 3)Percentage(2 digit).Since the data in these variables are of different scales, it is tough to compare these variables.

Number: Apps,Accept,Enroll,F.Undergrad,P.Undergrad,Outstate,Room.Board,Books,Personal. Percentage :Top10perc,Top25perc,PhD,Terminal,perc.alumni. Ratio:Grad.Rate,

#Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

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..

# When it comes to choosing between Covariance vs Correlation, the latter stands to be the first choice as it remains unaffected by the change in dimensions, location, and scale, and can also be used to make a comparison between two pairs of variables. Since it is limited to a range of -1 to +1, it is useful to draw comparisons between variables across domains. However, an important limitation is that both these concepts measure the only linear relationship.

# Applying zscore or using StandardScalar give us the same results

It scales the data in such a way that the mean value of the features tends to 0 and the standard deviation tends to 1

Min-Max method ensure that the data scaled to have values in the range 0 to 1