BUSINESS REPORT -ADVANCED STATISTICS

BY

JASPER SHELDON M

Contents

[Problem 1a: 5](#_Toc77457511)

[Executive Summary: 5](#_Toc77457512)

[Introduction: 5](#_Toc77457513)

[Sample data: 5](#_Toc77457514)

[Basic information about the data: 6](#_Toc77457515)

[Groups in Education: 6](#_Toc77457516)

[Groups in Occupation: 6](#_Toc77457517)

[1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually. 7](#_Toc77457518)

[Anova Test: 7](#_Toc77457519)

[Hypothesis for the test: 7](#_Toc77457520)

[*1.*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. 7](#_Toc77457521)

[1.3 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. 7](#_Toc77457522)

[1.4 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. 8](#_Toc77457523)

[Problem 1b: 8](#_Toc77457524)

[1.5 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. 9](#_Toc77457525)

[1.6 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? 10](#_Toc77457526)

[1.7 Explain the business implications of performing ANOVA for this particular case study. 11](#_Toc77457527)

[Problem 2: 12](#_Toc77457528)

[Summary: 12](#_Toc77457529)

[2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA? 12](#_Toc77457530)

[Sample of the data: 12](#_Toc77457531)

[Basic information of the dataset: 13](#_Toc77457532)

[Colleges with least applications: 16](#_Toc77457533)

[2.1 Insights from EDA: 21](#_Toc77457534)

[2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling. 22](#_Toc77457535)

[2.3 Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data]¶ 23](#_Toc77457536)

[2.4 Check the dataset for outliers before and after scaling. What insight do you derive here? 23](#_Toc77457537)

[2.5 Extract the eigenvalues and eigenvectors. 24](#_Toc77457538)

[2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features 27](#_Toc77457539)

[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] 28](#_Toc77457540)

[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? 28](#_Toc77457541)

[2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? 28](#_Toc77457542)

|  |  |  |
| --- | --- | --- |
|  | List of tables |  |
| Index | Name | p.No |
| 1 | basic info about salary data | 4 |
| 2 | Categories in education | 4 |
| 3 | Categories in Occupation | 4 |
| 4 | Result of TukeyHSD on education | 5 |
| 5 | Results of 2 way anova | 8 |
| 6 | Sample of the dataset Colleges | 10 |
| 7 | basic info about College data | 10 |
| 8 | Sample of Scaled data | 19 |
| 9 | PCA components | 23 |
| 10 | Cumulative sum of explained variance | 24 |

|  |  |  |
| --- | --- | --- |
|  | List of Figures |  |
| Index | Name | p.No |
| 1 | Count plot of education on occupation | 6 |
| 2 | Occupation vs Salary | 7 |
| 3 | Scatter plot Education vs Salary | 7 |
| 4 | Box plot of College dataset | 11 |
| 5 | Colleges with the most number of education | 12 |
| 6 | Colleges with the least number of education | 13 |
| 7 | Colleges preferred by top 10 HS school | 13 |
| 8 | Histogram | 14 |
| 9 | Pairplot | 15 |
| 10 | Enrollment vs room cost | 16 |
| 11 | Enrollment vs Book cost | 16 |
| 12 | Enrollment vs Personal cost | 16 |
| 13 | Phd vs Terminal | 17 |
| 14 | Expenditure vs Outstate | 17 |
| 15 | Heatmap | 18 |
| 16 | Covariance and Correlation | 20 |
| 17 | Box plot of Scaled data | 21 |
| 18 | Eigen vectors | 21 |
| 19 | Eigen values | 23 |

# Problem 1a:

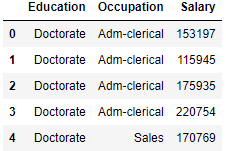
## Executive Summary:

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/43034/files/3076283/download?verifier=S1v4XG3nOkdxC7201g1aIq7s4i4UW9qQPvg6ZqWQ&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.

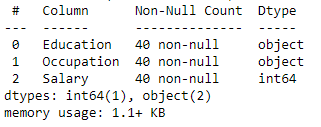
## Introduction:

The intention of this case study is to find whether there is a significant relationship between the categorical variables (Education and Occupation) on the variable Salary. We are going to statistically prove whether a relationship exists with the help of ANOVA. Anova can help us find whether the relationship is significant by analyzing the means with the help of variances. To find the relationship we are going to proceed with one way anova and two way anova to evaluate the interaction between the independent variables.

## Sample data:

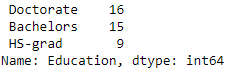


## Basic information about the data:



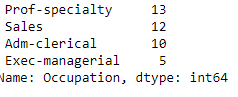
From the table we can see the dataset has 40 entries and there are no missing values present .Two columns are object type and salary is integer type.

## Groups in Education:



From the table we can see the number of people in each category.

## Groups in Occupation:



From the table we can see the number of people in each category in occupation.

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

## Anova Test:

### Hypothesis for the test:

Education:

𝐻0 : The mean salary for the entire category in education is same

𝐻𝑎: The mean salary is different for at least one of the category in education

Occupation:

𝐻0: the mean salaries for all occupation categories are same

𝐻𝑎: The mean salary is different for at least one of the category in occupation

## *1.*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.

After deciding the null and alternative hypothesis we are going to perform the anova test on the education with respect to salary and found the P value to be **1.257709e-08** which is less than the alpha value (0.05) which means we have enough evidence to reject Ho.

Rejecting the Ho, we can conclude there is difference in the mean salary within the different levels of education so education has a significant relationship on salary.

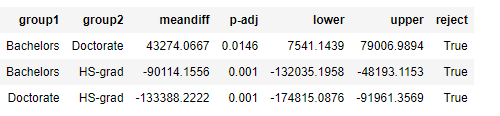
## 1.3 Perform one-way ANOVA for variable Occupation with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

After deciding the null and alternative hypothesis , performing the anova test on the occupation with respect to salary, the P value was found out to be **0.458508** which is not less than the alpha value (0.05) which means we do not have enough evidence to reject Ho so we fail to reject Ho.

We are concluding that there is no significant difference in the mean salary obtained by different categories in the occupation, so Occupation does not have a significant relationship on salary.

## 1.4 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.

From the anova results we rejected null hypothesis in the case of 1.2 that is for education and accepted null hypothesis for occupation to figure out the class means in Education with significant difference we are proceeding with post Hoc test. Post Hoc test does multiple T-tests comparing the means without inflating the type 1 error and keeps it under threshold level (0.05). In post Hoc test we have multiple tests available to check for differences but we have chosen Tukey HSD to check for multiple comparisons of means.

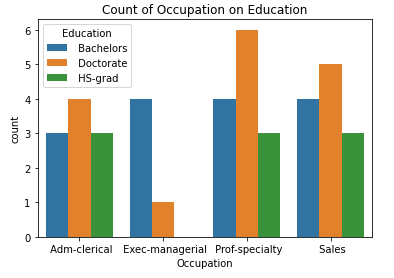


The above is the result of Tukey HSD from the result we can see that all three groups have significant difference in their mean with respect other groups, which means all 3 have different means.

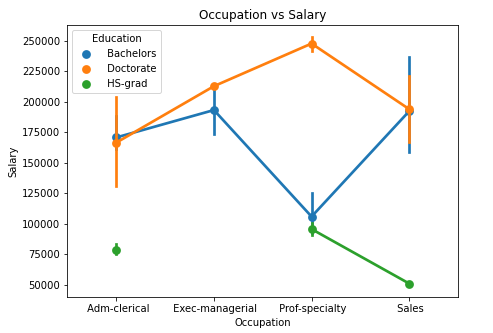
# Problem 1b:

In this section of the case study we are going to estimate the cause and effect of the variables on each other and also the relationship between both the independent variables (Education and occupation).Understanding the relationship also helps us to evaluate the interaction of them with respect to salary.

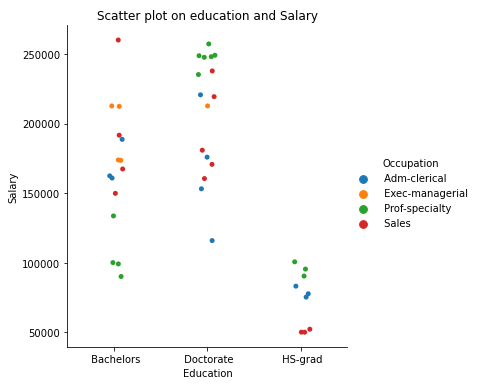
## 1.5 What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.



From the bar graph we can see majority of the people working as professors have done Doctorate, same is applicable for sales also. For Manager Roles most of them are from Bachelor degree .Clerical jobs also has majority of doctorate students.



From the point plot we can see students who have done Honors has the minimum salary across all occupations, Bachelors have the intermediate salary and those who have done Doctorate get the maximum salary in all occupation types. We can clearly see that education plays a vital role in deciding the salary across all occupations. There is a positive linear interaction present, higher the education higher the salary



People who have done Honors and works as professors earn the maximum and those in sales earn the minimum salary where as in the case of Bachelors professors earn the minimum salary and sales people earn the maximum.

People who have done doctorate and works as professors have the maximum salary and clerical jobs have the minimum salary.

## 1.6 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?

The hypotheses are:

𝐻1o: The means are same for all groups in the education category

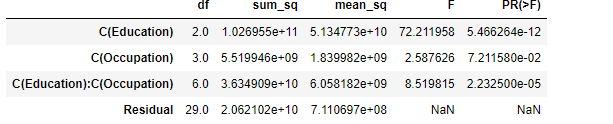
𝐻1a: at least one group has a different mean in the education category

𝐻2o : The means are same for all groups in the occupation category

𝐻2a : at least one group has a different mean in the occupation category

𝐻3o : There is no interaction effect between education and occupation

𝐻3a: There is interaction effect between education and occupation



**Two way anova results**

As we can see the p value is less for education, occupation and the interaction variable so we can reject the null hypothesis in all three cases. As we have enough statistical evidence to reject the null hypothesis we can conclude at least one category in education has different means and at least one category in occupation also has different mean. There is a significant interaction present between education and occupation on the salary earned. We can note in the beginning occupation does not affect the salary significantly but the presence of interaction variable has changed the effect of occupation on salary.

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

Basically we perform ANOVA to find whether there is a significant difference in the mean between the groups. If there is no significant difference it means that particular factor does not affect the dependent variable.

In this case we can see anova results show that there are significant differences in the groups for education and occupation which means education has a significant effect on the salary so is the occupation, that is salary is affected by education and occupation individually in other words we can conclude the type of education and the occupation we do decides the salary we earn. Anova also gives a third result showing there is a significant interaction present between the education and occupation. This means the presence of education influences the effect of occupation on salary. Education and occupation together affects the salary.

From business point of view we can say that education and occupation individually affects the salary we earn also both together has an impact on the salary.

# Problem 2:

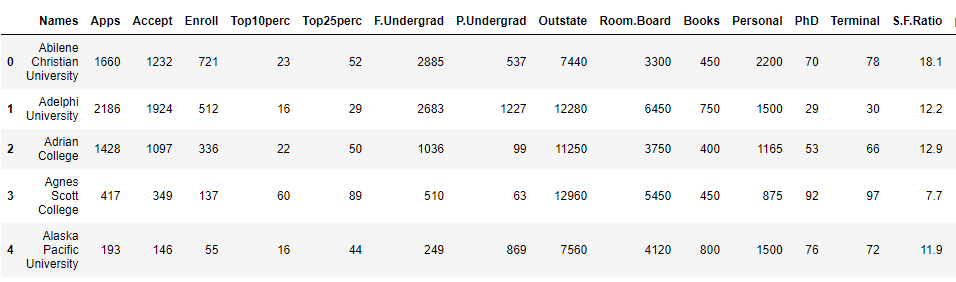
The dataset [Education - Post 12th Standard.csv](https://olympus.greatlearning.in/courses/43034/files/2468434/download?verifier=c5oDcKhvBPnpGEMhoJl2Vjw30PI4Xvk6gEDYZkrX&wrap=1) contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according 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/43034/files/2468433/download?verifier=3VMnRSXXj3dVHLK43VlYQ9SQClrO2J5DoX991v3F&wrap=1).

## Summary:

The dataset contains the list of all the colleges with the number people applying for it. It consists of people being accepted and people joining the college. It also includes the information about faculty and the category of students joining. The cost of books, rooms and institution is also provided we can explore the data to get some insight and try to do PCA if applicable.

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

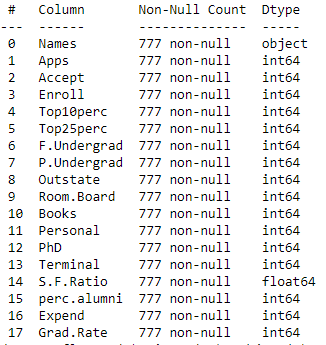
## Sample of the data:



**First 5 rows from the dataset**

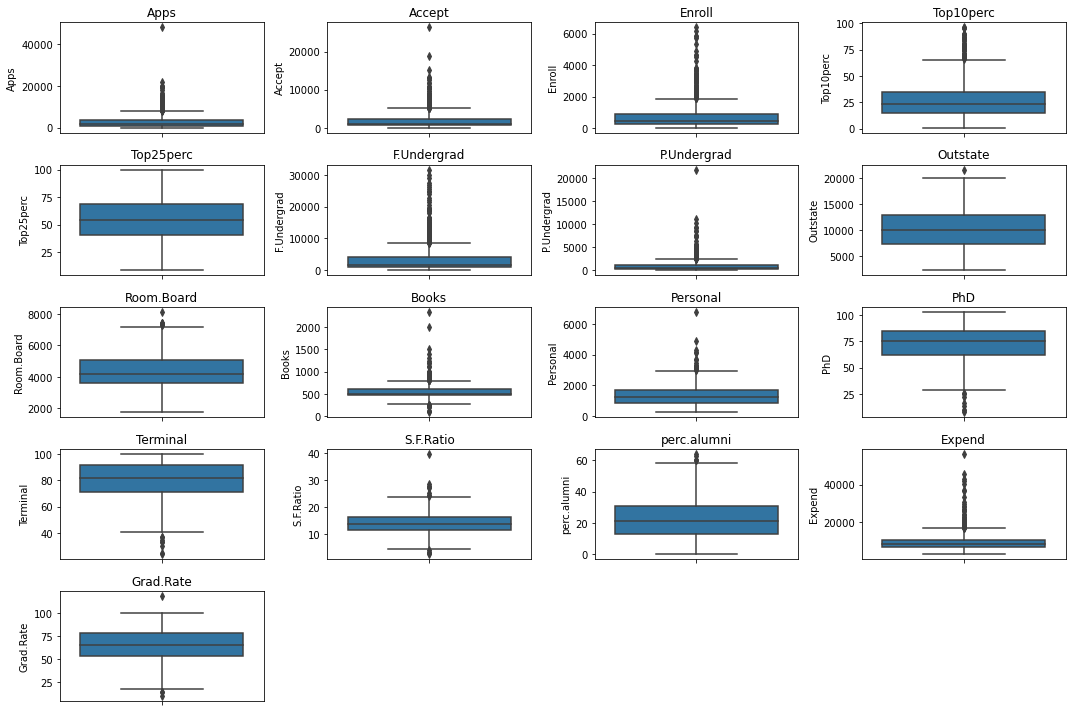
From the sample we can see the name of the college alone is of textual data and other columns are of numerical type.

## Basic information of the dataset:



From the figure we can see we have a total of 17 columns which are numerical type 16 integer types and 1 float type. There are no missing values present in the dataset.

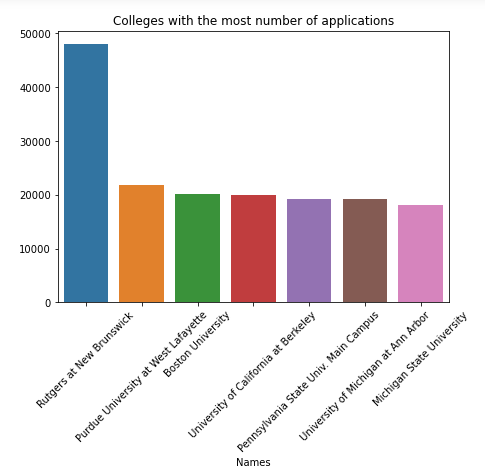
Test for outliers:



**Box plot of all the variables**

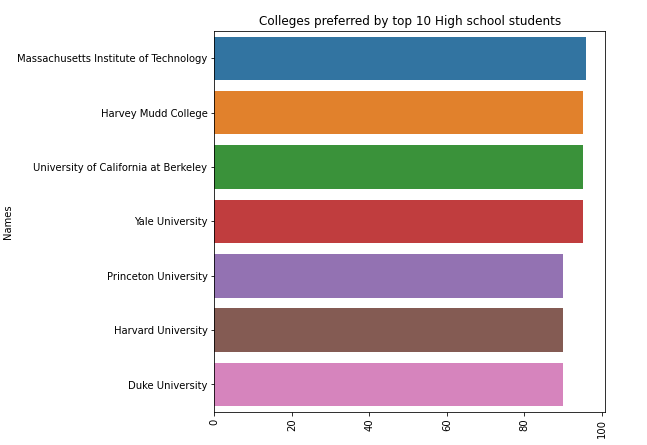
Only top 25percent column is free from outliers rest of the columns have outliers.

**Colleges with the most number of application:**

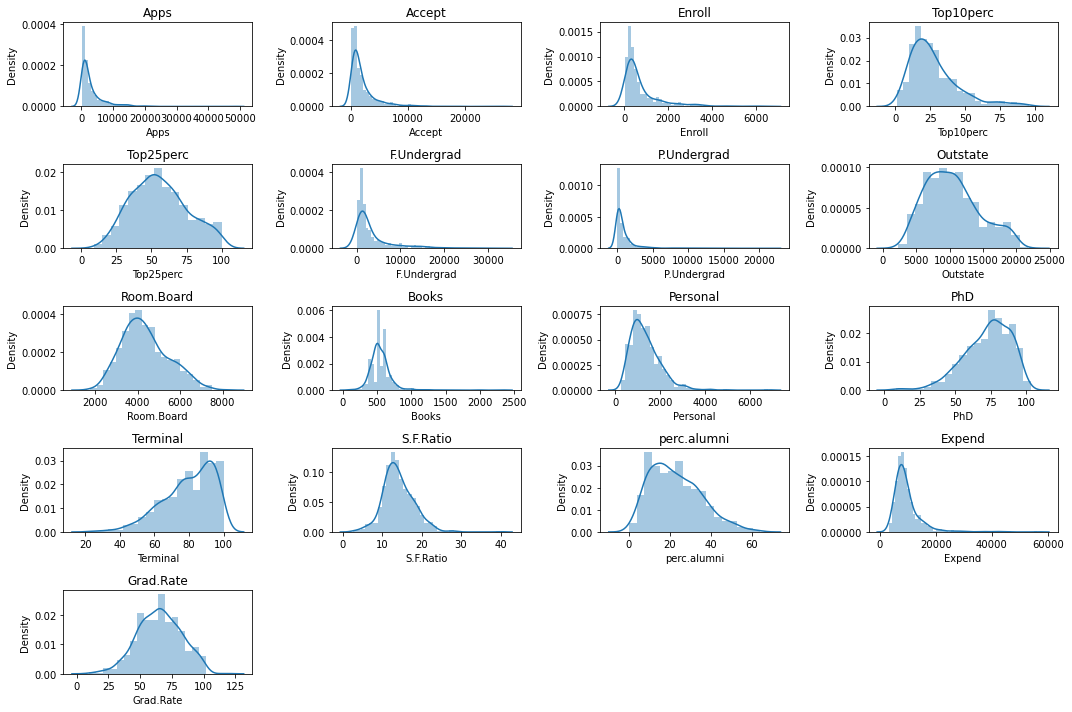


## Colleges with least applications:

***Colleges preferred by students from top 10 higher secondary schools:***



**Univariate Analysis:**



Few distributions have right skewness in them and other half has left skewness and few distributions tend to follow normal distribution.

***Distributions that follow right skewness :***

Apps ,Accept,Enroll,Top10perc,F.undergrad,P.undergrad,Personal,S.F.Ratio,Expend,perc.alumni

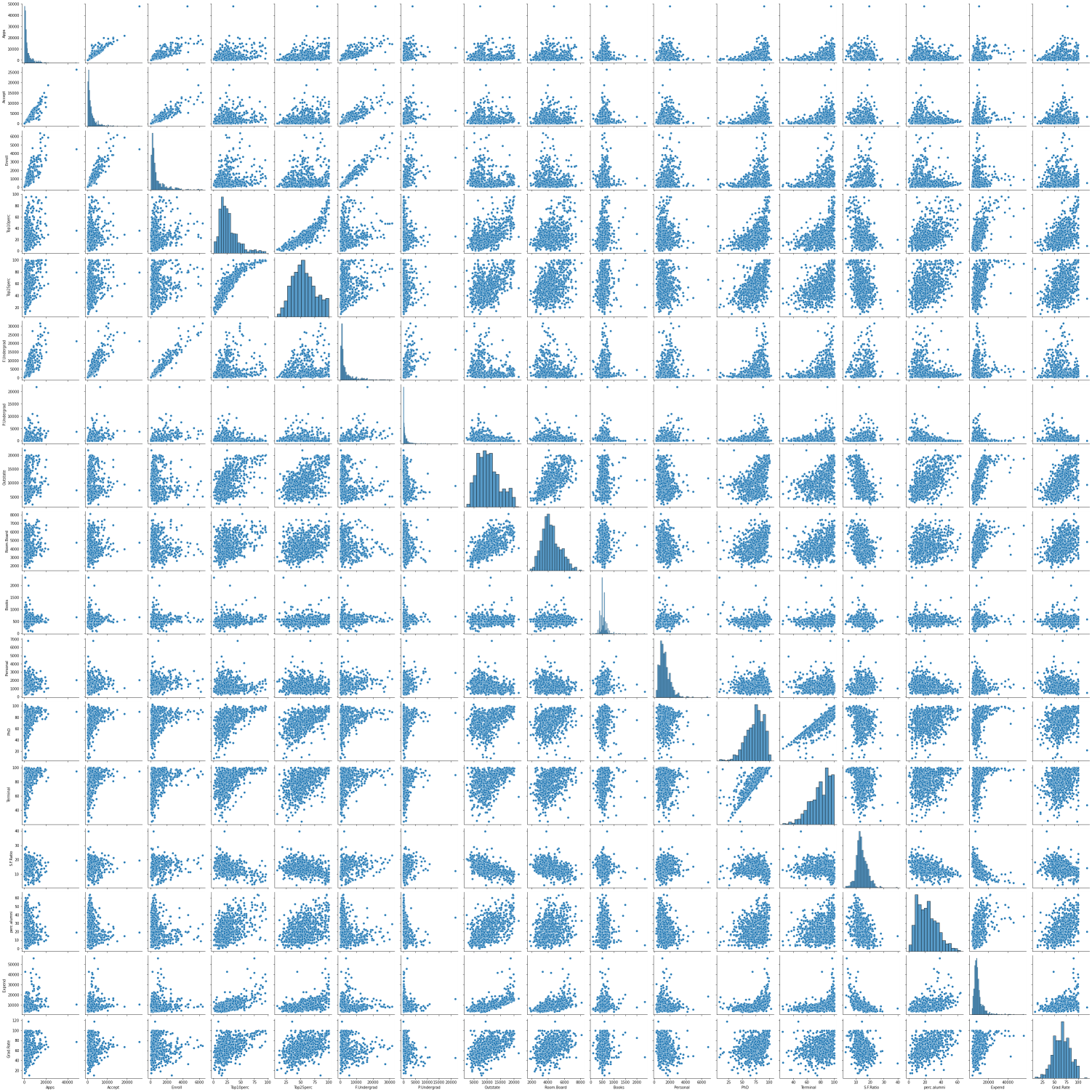
***Distributions that follow left skewness:***

Phd,Terminal

***Distribution that approximately normality:***

Top25Perc,Outstate,Room.Board,Grad.Rate

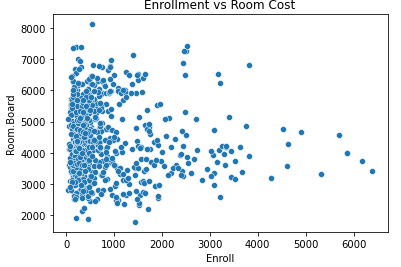
Pair plots:

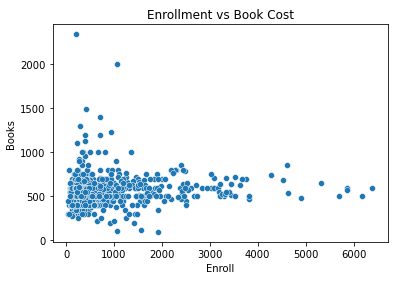


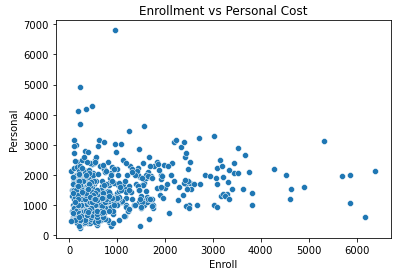
**Pair plots of College dataset**

From the pair plot we can take out the variables with linear relationship and try to plot and see

**Does number of students enrolling a college depends on cost:**

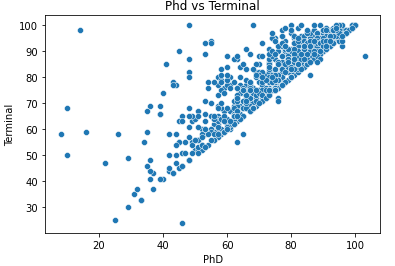






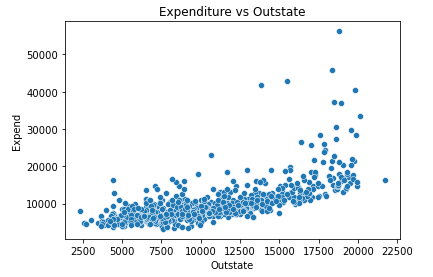
Enrollment does not depend on the room charges but it depends on book charges and personal charges, the lesser the charges more the enrollment.

PhD vs. Terminal:



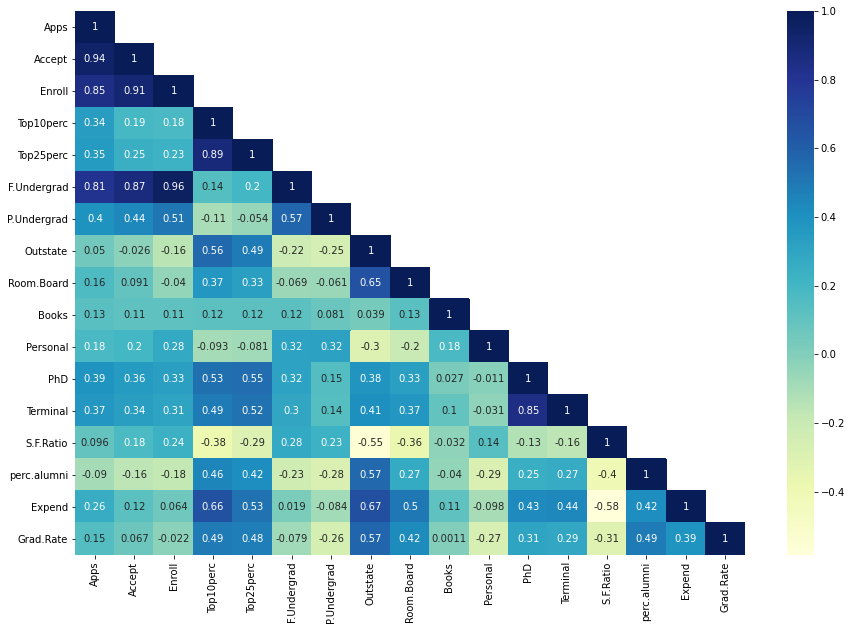
From the plot we can see there is strong liner relationship between PhD and Terminal.

Expenditure vs. Outstate:



From the plot we can see that expenditure is linearly dependent on outstate.

Heat map:



From the plot we can see there is strong correlation present between few variables. There is a strong positive correlation and mild negative correlation present between the variables

## 2.1 Insights from EDA:

Rutgurs at New Brunswick is the college with the most number of applications and Christendom College is the one which receives the least number of applications.

Massachusetts Institute of Technology is the top preferred colleges by the students of top 10 higher secondary school

There are distribution which has left skewness , right skewness and few distribution follow normality

Number of students enrolling for the particular college is dependent on Book cost and personal cost, lesser the cost more the enrollment.

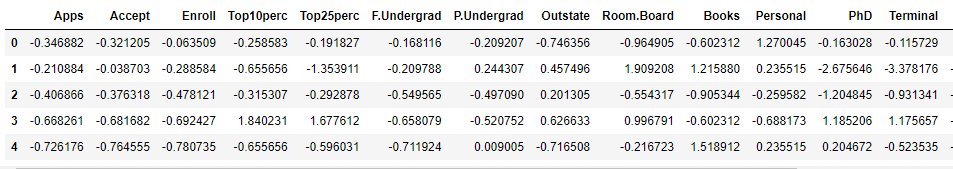
Expenditure increases for colleges if it’s out of state. The further the college higher is the expenditure.

Most of the variables in the dataset are correlated to each other which indicate we can opt for PCA to reduce the noise and increase the information.

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

PCA works by projecting the data into a new mathematical space in order to reduce the unwanted variances and captures most of the variance from the data and tries to represent with a minimum number of dimension. Scaling is necessary for PCA because as we can see there is lot of correlation present between the variables so to reduce this unexplained variance we must centralize these data points and rotate the axes in such a way we have very less correlation. Scaling does the job of centralizing the data points and forming the Eigen vectors will rotate the axes to minimize the variations. Without centralizing the data points creating Eigen vectors won’t be effective and the noise won’t be reduced.

Sample of Scaled data:

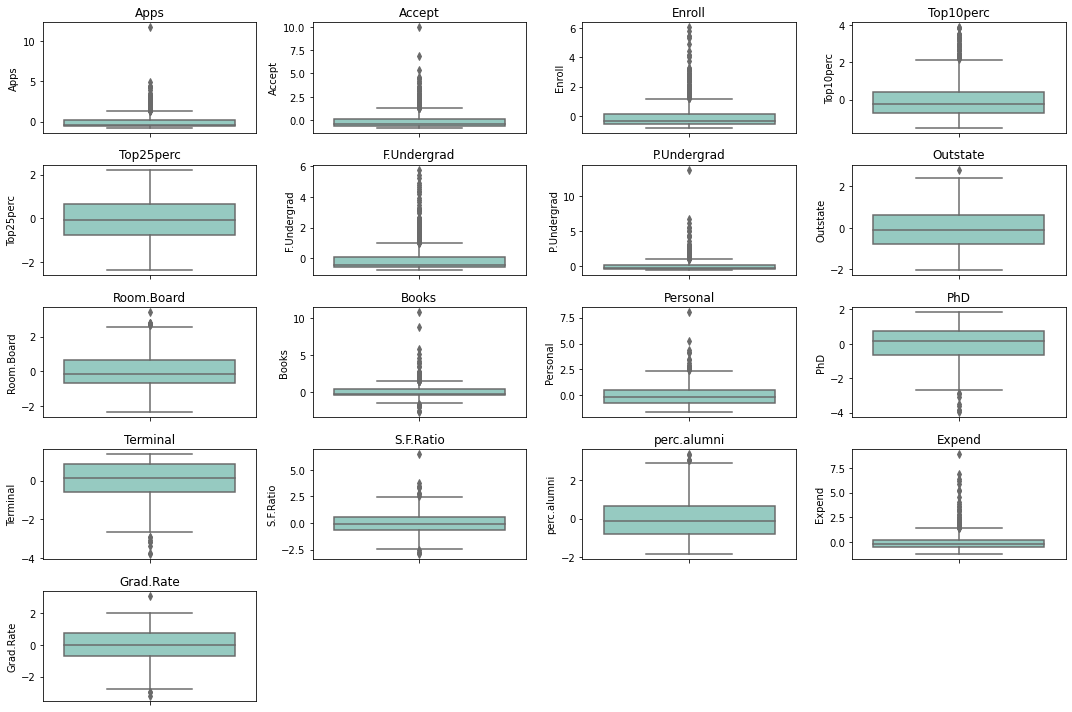


## 2.3 Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data][¶](http://localhost:8888/notebooks/Advanced%20Statistics/Assignement/AS_ASSIGNMENT_JASPER_M.ipynb#2.3-Comment-on-the-comparison-between-the-covariance-and-the-correlation-matrices-from-this-data.[on-scaled-data])

Covariance just shows the strength of the relationship between variables it ranges from -∞ to + ∞whereas the correlation matrix not only tells the strength but also the direction. It ranges from -1 to 1 where as in this case after scaling the covariance matrix and correlation matrix is almost equal the plot looks identical.

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

Outliers are plotted in the section 2.1 before scaling and we could see outliers present in the data set. The below plot consists of box plot of data after scaling.



We can see Scaling does not treat outliers, even after scaling outliers are still present only the range is being changed.

## 2.5 Extract the eigenvalues and eigenvectors.

After performing the scaling on the data set we must find whether there is a significant correlation present in the data set for that we can do Bartlett Sphericity test to prove a correlation exists statistically. For the above test we got P value to be 0 so we can proceed with the next step.

Secondly, after doing Bartlett test we must check for adequacy of the dataset for that we are performing kmo test on the data. We found the result to be 0.81. As this is greater than 0.7 we are going to do pca on the dataset and going to construct Eigen vectors and find the Eigen values.

Eigen vectors:

[ 2.48765602e-01, 2.07601502e-01, 1.76303592e-01,

3.54273947e-01, 3.44001279e-01, 1.54640962e-01,

2.64425045e-02, 2.94736419e-01, 2.49030449e-01,

6.47575181e-02, -4.25285386e-02, 3.18312875e-01,

3.17056016e-01, -1.76957895e-01, 2.05082369e-01,

3.18908750e-01, 2.52315654e-01],

[ 3.31598227e-01, 3.72116750e-01, 4.03724252e-01,

-8.24118211e-02, -4.47786551e-02, 4.17673774e-01,

3.15087830e-01, -2.49643522e-01, -1.37808883e-01,

5.63418434e-02, 2.19929218e-01, 5.83113174e-02,

4.64294477e-02, 2.46665277e-01, -2.46595274e-01,

-1.31689865e-01, -1.69240532e-01],

[-6.30921033e-02, -1.01249056e-01, -8.29855709e-02,

3.50555339e-02, -2.41479376e-02, -6.13929764e-02,

1.39681716e-01, 4.65988731e-02, 1.48967389e-01,

6.77411649e-01, 4.99721120e-01, -1.27028371e-01,

-6.60375454e-02, -2.89848401e-01, -1.46989274e-01,

2.26743985e-01, -2.08064649e-01],

[ 2.81310530e-01, 2.67817346e-01, 1.61826771e-01,

-5.15472524e-02, -1.09766541e-01, 1.00412335e-01,

-1.58558487e-01, 1.31291364e-01, 1.84995991e-01,

8.70892205e-02, -2.30710568e-01, -5.34724832e-01,

-5.19443019e-01, -1.61189487e-01, 1.73142230e-02,

7.92734946e-02, 2.69129066e-01],

[ 5.74140964e-03, 5.57860920e-02, -5.56936353e-02,

-3.95434345e-01, -4.26533594e-01, -4.34543659e-02,

3.02385408e-01, 2.22532003e-01, 5.60919470e-01,

-1.27288825e-01, -2.22311021e-01, 1.40166326e-01,

2.04719730e-01, -7.93882496e-02, -2.16297411e-01,

7.59581203e-02, -1.09267913e-01],

[-1.62374420e-02, 7.53468452e-03, -4.25579803e-02,

-5.26927980e-02, 3.30915896e-02, -4.34542349e-02,

-1.91198583e-01, -3.00003910e-02, 1.62755446e-01,

6.41054950e-01, -3.31398003e-01, 9.12555212e-02,

1.54927646e-01, 4.87045875e-01, -4.73400144e-02,

-2.98118619e-01, 2.16163313e-01],

[-4.24863486e-02, -1.29497196e-02, -2.76928937e-02,

-1.61332069e-01, -1.18485556e-01, -2.50763629e-02,

6.10423460e-02, 1.08528966e-01, 2.09744235e-01,

-1.49692034e-01, 6.33790064e-01, -1.09641298e-03,

-2.84770105e-02, 2.19259358e-01, 2.43321156e-01,

-2.26584481e-01, 5.59943937e-01],

[-1.03090398e-01, -5.62709623e-02, 5.86623552e-02,

-1.22678028e-01, -1.02491967e-01, 7.88896442e-02,

5.70783816e-01, 9.84599754e-03, -2.21453442e-01,

2.13293009e-01, -2.32660840e-01, -7.70400002e-02,

-1.21613297e-02, -8.36048735e-02, 6.78523654e-01,

-5.41593771e-02, -5.33553891e-03],

[-9.02270802e-02, -1.77864814e-01, -1.28560713e-01,

3.41099863e-01, 4.03711989e-01, -5.94419181e-02,

5.60672902e-01, -4.57332880e-03, 2.75022548e-01,

-1.33663353e-01, -9.44688900e-02, -1.85181525e-01,

-2.54938198e-01, 2.74544380e-01, -2.55334907e-01,

-4.91388809e-02, 4.19043052e-02],

[ 5.25098025e-02, 4.11400844e-02, 3.44879147e-02,

6.40257785e-02, 1.45492289e-02, 2.08471834e-02,

-2.23105808e-01, 1.86675363e-01, 2.98324237e-01,

-8.20292186e-02, 1.36027616e-01, -1.23452200e-01,

-8.85784627e-02, 4.72045249e-01, 4.22999706e-01,

1.32286331e-01, -5.90271067e-01],

[ 4.30462074e-02, -5.84055850e-02, -6.93988831e-02,

-8.10481404e-03, -2.73128469e-01, -8.11578181e-02,

1.00693324e-01, 1.43220673e-01, -3.59321731e-01,

3.19400370e-02, -1.85784733e-02, 4.03723253e-02,

-5.89734026e-02, 4.45000727e-01, -1.30727978e-01,

6.92088870e-01, 2.19839000e-01],

[ 2.40709086e-02, -1.45102446e-01, 1.11431545e-02,

3.85543001e-02, -8.93515563e-02, 5.61767721e-02,

-6.35360730e-02, -8.23443779e-01, 3.54559731e-01,

-2.81593679e-02, -3.92640266e-02, 2.32224316e-02,

1.64850420e-02, -1.10262122e-02, 1.82660654e-01,

3.25982295e-01, 1.22106697e-01],

[ 5.95830975e-01, 2.92642398e-01, -4.44638207e-01,

1.02303616e-03, 2.18838802e-02, -5.23622267e-01,

1.25997650e-01, -1.41856014e-01, -6.97485854e-02,

1.14379958e-02, 3.94547417e-02, 1.27696382e-01,

-5.83134662e-02, -1.77152700e-02, 1.04088088e-01,

-9.37464497e-02, -6.91969778e-02],

[ 8.06328039e-02, 3.34674281e-02, -8.56967180e-02,

-1.07828189e-01, 1.51742110e-01, -5.63728817e-02,

1.92857500e-02, -3.40115407e-02, -5.84289756e-02,

-6.68494643e-02, 2.75286207e-02, -6.91126145e-01,

6.71008607e-01, 4.13740967e-02, -2.71542091e-02,

7.31225166e-02, 3.64767385e-02],

[ 1.33405806e-01, -1.45497511e-01, 2.95896092e-02,

6.97722522e-01, -6.17274818e-01, 9.91640992e-03,

2.09515982e-02, 3.83544794e-02, 3.40197083e-03,

-9.43887925e-03, -3.09001353e-03, -1.12055599e-01,

1.58909651e-01, -2.08991284e-02, -8.41789410e-03,

-2.27742017e-01, -3.39433604e-03],

[ 4.59139498e-01, -5.18568789e-01, -4.04318439e-01,

-1.48738723e-01, 5.18683400e-02, 5.60363054e-01,

-5.27313042e-02, 1.01594830e-01, -2.59293381e-02,

2.88282896e-03, -1.28904022e-02, 2.98075465e-02,

-2.70759809e-02, -2.12476294e-02, 3.33406243e-03,

-4.38803230e-02, -5.00844705e-03],

[ 3.58970400e-01, -5.43427250e-01, 6.09651110e-01,

-1.44986329e-01, 8.03478445e-02, -4.14705279e-01,

9.01788964e-03, 5.08995918e-02, 1.14639620e-03,

7.72631963e-04, -1.11433396e-03, 1.38133366e-02,

6.20932749e-03, -2.22215182e-03, -1.91869743e-02,

-3.53098218e-02, -1.30710024e-02]

Eigen values:

array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,

0.84849117, 0.6057878 , 0.58787222, 0.53061262, 0.4043029 ,

0.31344588, 0.22061096, 0.16779415, 0.1439785 , 0.08802464,

0.03672545, 0.02302787])

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | PC1 | | PC2 | | PC3 | | PC4 | | PC5 | | PC6 | | PC7 | | PC8 | | PC9 |
| Apps | | 0.248766 | | 0.331598 | | -0.06309 | | 0.281311 | | 0.005741 | | -0.01624 | | -0.04249 | | -0.10309 | | -0.09023 |
| Accept | | 0.207602 | | 0.372117 | | -0.10125 | | 0.267817 | | 0.055786 | | 0.007535 | | -0.01295 | | -0.05627 | | -0.17786 |
| Enroll | | 0.176304 | | 0.403724 | | -0.08299 | | 0.161827 | | -0.05569 | | -0.04256 | | -0.02769 | | 0.058662 | | -0.12856 |
| Top10perc | | 0.354274 | | -0.08241 | | 0.035056 | | -0.05155 | | -0.39543 | | -0.05269 | | -0.16133 | | -0.12268 | | 0.3411 |
| Top25perc | | 0.344001 | | -0.04478 | | -0.02415 | | -0.10977 | | -0.42653 | | 0.033092 | | -0.11849 | | -0.10249 | | 0.403712 |
| F.Undergrad | | 0.154641 | | 0.417674 | | -0.06139 | | 0.100412 | | -0.04345 | | -0.04345 | | -0.02508 | | 0.07889 | | -0.05944 |
| P.Undergrad | | 0.026443 | | 0.315088 | | 0.139682 | | -0.15856 | | 0.302385 | | -0.1912 | | 0.061042 | | 0.570784 | | 0.560673 |
| Outstate | | 0.294736 | | -0.24964 | | 0.046599 | | 0.131291 | | 0.222532 | | -0.03 | | 0.108529 | | 0.009846 | | -0.00457 |
| Room.Board | | 0.24903 | | -0.13781 | | 0.148967 | | 0.184996 | | 0.560919 | | 0.162755 | | 0.209744 | | -0.22145 | | 0.275023 |
| Books | | 0.064758 | | 0.056342 | | 0.677412 | | 0.087089 | | -0.12729 | | 0.641055 | | -0.14969 | | 0.213293 | | -0.13366 |
| Personal | | -0.04253 | | 0.219929 | | 0.499721 | | -0.23071 | | -0.22231 | | -0.3314 | | 0.63379 | | -0.23266 | | -0.09447 |
| PhD | | 0.318313 | | 0.058311 | | -0.12703 | | -0.53472 | | 0.140166 | | 0.091256 | | -0.0011 | | -0.07704 | | -0.18518 |
| Terminal | | 0.317056 | | 0.046429 | | -0.06604 | | -0.51944 | | 0.20472 | | 0.154928 | | -0.02848 | | -0.01216 | | -0.25494 |
| S.F.Ratio | | -0.17696 | | 0.246665 | | -0.28985 | | -0.16119 | | -0.07939 | | 0.487046 | | 0.219259 | | -0.0836 | | 0.274544 |
| perc.alumni | | 0.205082 | | -0.2466 | | -0.14699 | | 0.017314 | | -0.2163 | | -0.04734 | | 0.243321 | | 0.678524 | | -0.25533 |
| Expend | | 0.318909 | | -0.13169 | | 0.226744 | | 0.079273 | | 0.075958 | | -0.29812 | | -0.22658 | | -0.05416 | | -0.04914 |
| Grad.Rate | | 0.252316 | | -0.16924 | | -0.20806 | | 0.269129 | | -0.10927 | | 0.216163 | | 0.559944 | | -0.00534 | | 0.041904 |
|  | PC10 | | PC11 | | PC12 | | PC13 | | PC14 | | PC15 | | PC16 | | PC17 | |
| Apps | 0.05251 | | 0.043046 | | 0.024071 | | 0.595831 | | 0.080633 | | 0.133406 | | 0.459139 | | 0.35897 | |
| Accept | 0.04114 | | -0.05841 | | -0.1451 | | 0.292642 | | 0.033467 | | -0.1455 | | -0.51857 | | -0.54343 | |
| Enroll | 0.034488 | | -0.0694 | | 0.011143 | | -0.44464 | | -0.0857 | | 0.02959 | | -0.40432 | | 0.609651 | |
| Top10perc | 0.064026 | | -0.0081 | | 0.038554 | | 0.001023 | | -0.10783 | | 0.697723 | | -0.14874 | | -0.14499 | |
| Top25perc | 0.014549 | | -0.27313 | | -0.08935 | | 0.021884 | | 0.151742 | | -0.61727 | | 0.051868 | | 0.080348 | |
| F.Undergrad | 0.020847 | | -0.08116 | | 0.056177 | | -0.52362 | | -0.05637 | | 0.009916 | | 0.560363 | | -0.41471 | |
| P.Undergrad | -0.22311 | | 0.100693 | | -0.06354 | | 0.125998 | | 0.019286 | | 0.020952 | | -0.05273 | | 0.009018 | |
| Outstate | 0.186675 | | 0.143221 | | -0.82344 | | -0.14186 | | -0.03401 | | 0.038354 | | 0.101595 | | 0.0509 | |
| Room.Board | 0.298324 | | -0.35932 | | 0.35456 | | -0.06975 | | -0.05843 | | 0.003402 | | -0.02593 | | 0.001146 | |
| Books | -0.08203 | | 0.03194 | | -0.02816 | | 0.011438 | | -0.06685 | | -0.00944 | | 0.002883 | | 0.000773 | |
| Personal | 0.136028 | | -0.01858 | | -0.03926 | | 0.039455 | | 0.027529 | | -0.00309 | | -0.01289 | | -0.00111 | |
| PhD | -0.12345 | | 0.040372 | | 0.023222 | | 0.127696 | | -0.69113 | | -0.11206 | | 0.029808 | | 0.013813 | |
| Terminal | -0.08858 | | -0.05897 | | 0.016485 | | -0.05831 | | 0.671009 | | 0.15891 | | -0.02708 | | 0.006209 | |
| S.F.Ratio | 0.472045 | | 0.445001 | | -0.01103 | | -0.01772 | | 0.041374 | | -0.0209 | | -0.02125 | | -0.00222 | |
| perc.alumni | 0.423 | | -0.13073 | | 0.182661 | | 0.104088 | | -0.02715 | | -0.00842 | | 0.003334 | | -0.01919 | |
| Expend | 0.132286 | | 0.692089 | | 0.325982 | | -0.09375 | | 0.073123 | | -0.22774 | | -0.04388 | | -0.03531 | |
| Grad.Rate | -0.59027 | | 0.219839 | | 0.122107 | | -0.0692 | | 0.036477 | | -0.00339 | | -0.00501 | | -0.01307 | |

## 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]

E1=0.249\*Apps +0.208\*Accept +0.176\*Enroll +0.354\*Top10perc +0.344\*Top25perc +0.155\*F.Undergrad +0.026\*P.Undergrad +0.295\*Outstate +0.249\*Room.Board +0.065\*Books -0.043\*Personal +0.318\*PhD +0.317\*Terminal -0.177\*S.F.Ratio +0.205\*perc.alumni +0.319\*Expend +0.252\*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?

array([0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154,

0.81657854, 0.85216726, 0.88670347, 0.91787581, 0.94162773,

0.96004199, 0.9730024 , 0.98285994, 0.99131837, 0.99648962,

0.99864716, 1. ])

Cumulative sum of explained variance explains the total covariance captured cumulatively by each vectors. from the output we can see 7 vectors captures 85% of the total variance from the original data and 9 vectors captures a total of 91% of variances. With the help of cumulative sum of variance captured we can decide the number of Eigen vectors needed .If we need 85 % of the variance to be captured we can go ahead with 7 vectors.

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

In this case study the data set has 17 columns with covariance existing between lots of variables this might reduce the signal to noise ratio unwanted covariance in the dataset will impair the performance of the model while generalizing. To alter these effects we need to reduce the covariance between the independent variables also we can reduce the dimensions to increase the S/N ratio.PCA helps in reducing the dimension without losing any information and captures the variances among the variables. A model needs to learn effectively and to do that with the minimalistic data without much of a loss and noise pca comes in handy and forms the Eigen vectors with which we can proceed with the further analysis. From the explained variance we can see that 7 Eigen vectors have captured 85% of the total variance present in the data so we can conclude that 7 vectors is enough to proceed with the analysis

Summarizing the business implication of pca is to represent a huge dataset with a lesser number of properties which helps in the performance of the model prediction also it reduces the noise and bring the information from the dataset.