

Advanced Statistics PROJECT REPORT

AS



March 4, 2022

Brijal Panchal

Contents

[Problem 1 2](#_Toc97304646)

[1.1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually. 3](#_Toc97304647)

[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. 3](#_Toc97304648)

[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. 4](#_Toc97304649)

[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. 4](#_Toc97304650)

[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. 5](#_Toc97304651)

[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? 5](#_Toc97304652)

[1.7. Explain the business implications of performing ANOVA for this particular case study. 6](#_Toc97304653)

[Problem 2 7](#_Toc97304654)

[2.1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA? 8](#_Toc97304655)

[2.2. Is scaling necessary for PCA in this case? Give justification and perform scaling. 12](#_Toc97304656)

[2.3. Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data] 13](#_Toc97304657)

[2.4. Check the dataset for outliers before and after scaling. What insight do you derive here? 14](#_Toc97304658)

[2.5. Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both] 15](#_Toc97304659)

[2.6. Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features 15](#_Toc97304660)

[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](#_Toc97304661)

[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? 18](#_Toc97304662)

[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] 19](#_Toc97304663)

# Problem 1

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

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

Data Discerption

* Education: education type (High school graduate, Bachelor, Doctorate)
* Occupation: occupation type (Administrative and clerical, Sales, Professional, specialty)
* Salary: salary value

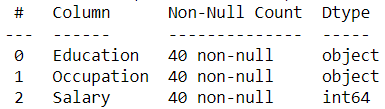
Sample of the dataset:



Dataset has three variables with three different kind of education and four different type of occupation. Based on different combination salary is defined.

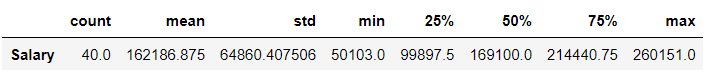
Exploratory Data Analysis

Let us check the types of variables in the data frame and missing values in dataset.



There are total 40 rows and 3 column in the dataset. Out of three columns, two are object and one is integer. There is no missing value present in the dataset.

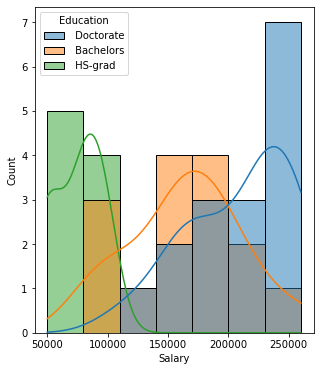
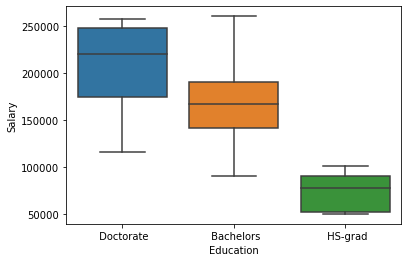
Summary of description



We can see mean and median of salary is almost same around 16K. Minimum salary is 50K while maximum salary is 26K.

* 1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.
* Null and alternate hypothesis for Education:
  + 𝐻0: 𝑚𝑒𝑎𝑛 𝑠𝑎𝑙𝑎𝑟𝑦 𝑜𝑓 𝑎𝑙𝑙 𝐸𝑑𝑢𝑐𝑎𝑡𝑖𝑜𝑛 𝑐𝑎𝑡𝑎𝑔𝑜𝑟𝑦 𝑎𝑟𝑒 𝑠𝑎𝑚𝑒.
  + 𝐻1: 𝑎𝑡𝑙𝑒𝑎𝑠𝑡 𝑜𝑛𝑒 𝑐𝑎𝑡𝑎𝑔𝑜𝑟𝑦 𝑜𝑓 𝐸𝑑𝑢𝑐𝑎𝑡𝑖𝑜𝑛 𝑠𝑎𝑙𝑎𝑟𝑦 𝑚𝑒𝑎𝑛 𝑖𝑠 𝑑𝑖𝑓𝑓𝑟𝑒𝑛𝑡.
* Null and alternate hypothesis for Occupation:
  + 𝐻0: 𝑚𝑒𝑎𝑛 𝑠𝑎𝑙𝑎𝑟𝑦 𝑜𝑓 𝑎𝑙𝑙 Occupation 𝑐𝑎𝑡𝑎𝑔𝑜𝑟𝑦 𝑎𝑟𝑒 𝑠𝑎𝑚𝑒.
  + 𝐻1: 𝑎𝑡𝑙𝑒𝑎𝑠𝑡 𝑜𝑛𝑒 𝑐𝑎𝑡𝑎𝑔𝑜𝑟𝑦 𝑜𝑓 Occupation 𝑠𝑎𝑙𝑎𝑟𝑦 𝑚𝑒𝑎𝑛 𝑖𝑠 𝑑𝑖𝑓𝑓𝑟𝑒𝑛𝑡.
  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.

Visualise Education category dataset

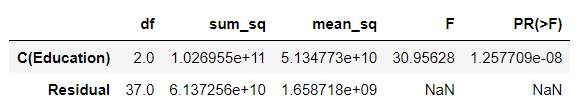
 

From histogram and box-plot of salary over education, we are seeing significant mean difference of salary in all category of education.

Formula for one-way Anova for Education

Formula = 'Salary ~ C(Education)'

One-way Anova table for Education

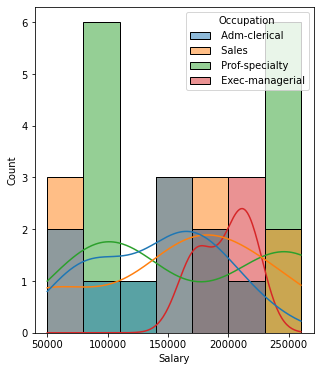
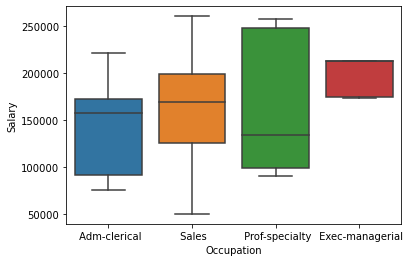


Conclusion

We see that the p-value is less than alpha (0.05). Thus, we 𝐑𝐞𝐣𝐞𝐜𝐭 the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 (𝐻0). This means at least one particular category in the 'Education' variable has different mean of salary as compared to the other categories.

* 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.

Visualise Occupation category dataset

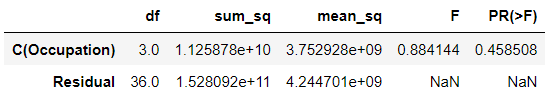
 

From histogram and box-plot of salary over occupation, we are not seeing significant mean difference of salary in all category of occupation.

Formula for one-way Anova for Occupation

Formula = 'Salary ~ C(Occupation)'

One-way Anova table for Occupation

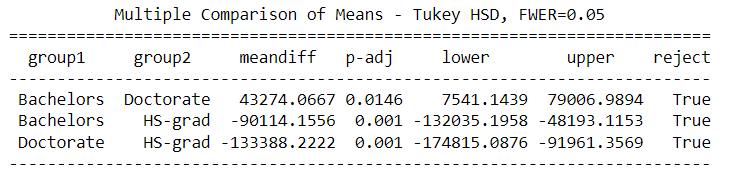


Conclusion

We see that the p-value is greater than alpha (0.05). Thus, we **failed** **to** 𝐑𝐞𝐣𝐞𝐜𝐭 the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 (𝐻0). This means salary of all category in the Occupation variable have same mean of 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.

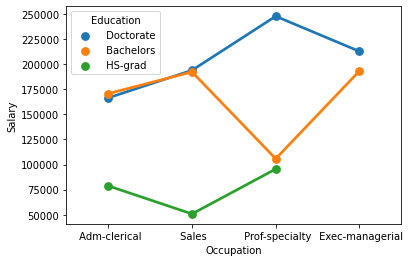
As we can see (1.2) has failed. let’s do Tukeyhsd hypothesis test to find category who's mean are significantly different



As per tukeyhsd table, all combination category of education has different mean.

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

Interaction effects represent the combined effects of factors on the dependent measure.



From interaction plot we can see that there is some sort of interaction of education over occupation.

* Hs-grad on other two category for adm-cierical and sales occupation
* Bachelors on other two category for prof-specialty 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?

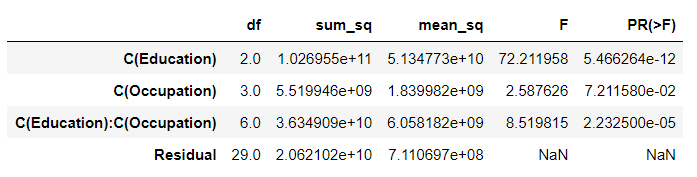
The null and the alternate hypothesis for conducting two-way ANOVA for interaction

* 𝐻0: 𝑇ℎ𝑒𝑟𝑒 𝑖𝑠 𝑛𝑜 𝑖𝑛𝑡𝑒𝑟𝑎𝑐𝑡𝑖𝑜𝑛 𝑏𝑒𝑡𝑤𝑒𝑒𝑛 𝐸𝑑𝑢𝑐𝑎𝑡𝑖𝑜𝑛 𝑎𝑛𝑑 𝑂𝑐𝑐𝑢𝑝𝑎𝑡𝑖𝑜𝑛.
* 𝐻1: 𝑇ℎ𝑒𝑟𝑒 𝑖𝑠 𝑖𝑛𝑡𝑒𝑟𝑎𝑐𝑡𝑖𝑜𝑛 𝑏𝑒𝑡𝑤𝑒𝑒𝑛 𝐸𝑑𝑢𝑐𝑎𝑡𝑖𝑜𝑛 𝑎𝑛𝑑 𝑂𝑐𝑐𝑢𝑝𝑎𝑡𝑖𝑜𝑛.

Formula for two-way Anova test for occupation and education

Formula = 'Salary ~ C(Education)+C(Occupation)+C(Education):C(Occupation)'

Two-way anova table



Conclusion

We see that the p-value is less than alpha (0.05). Thus, we 𝐑𝐞𝐣𝐞𝐜𝐭 the 𝐍𝐮𝐥𝐥 𝐇𝐲𝐩𝐨𝐭𝐡𝐞𝐬𝐢𝐬 (𝐻0). This means there is interaction effect between education and occupation categories.

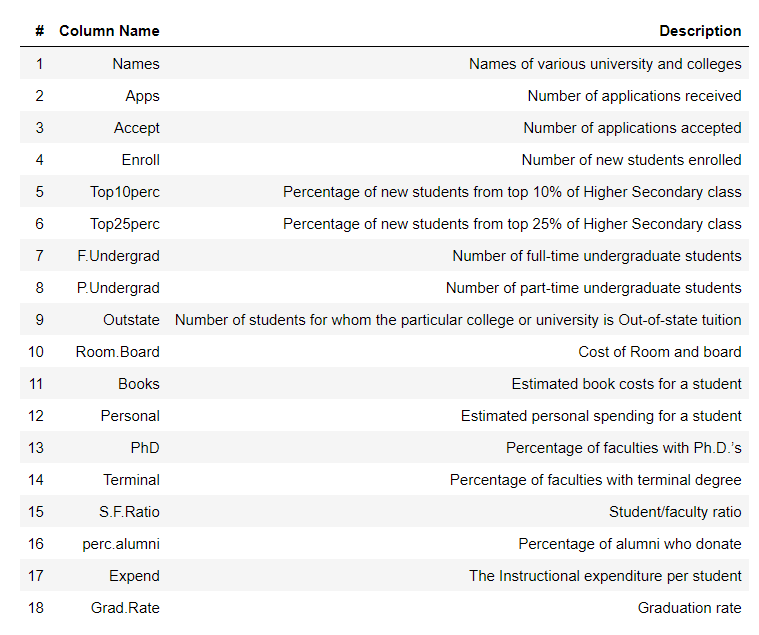
* 1. Explain the business implications of performing ANOVA for this particular case study.

After visualizing and analysing as well as performing Anova hypotheses tests on the dataset of salary over Education and occupation we found bellow insights.

* All type of education have different mean for salary
* Combination of educations and occupations will also give different mean

# Problem 2

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.

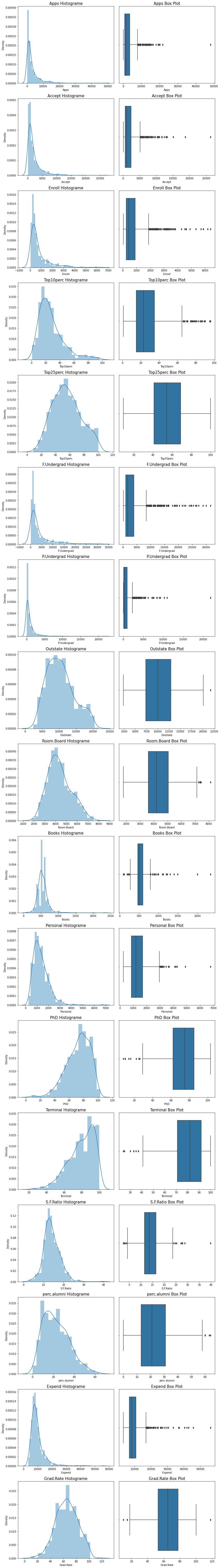
Data Discerption 

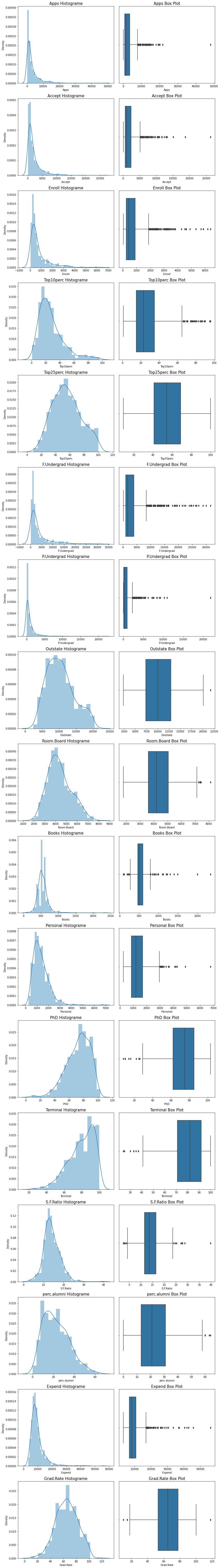
Sample of the dataset:

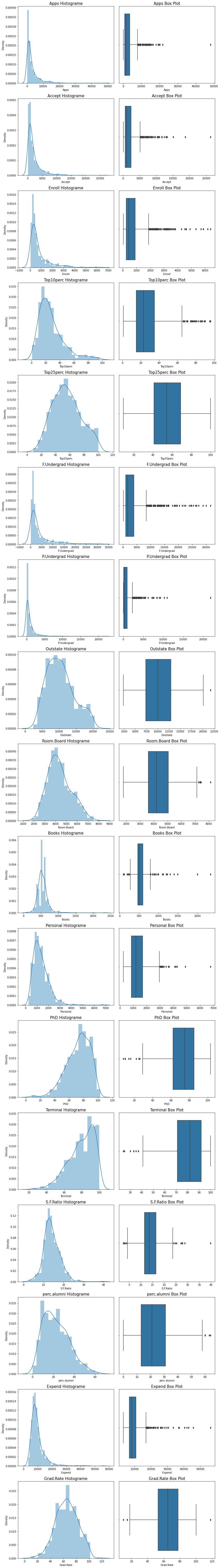


## Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

Univariate analysis



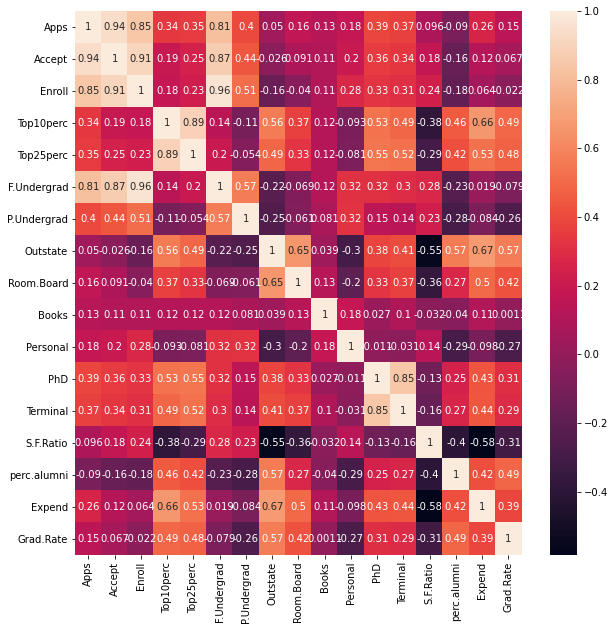




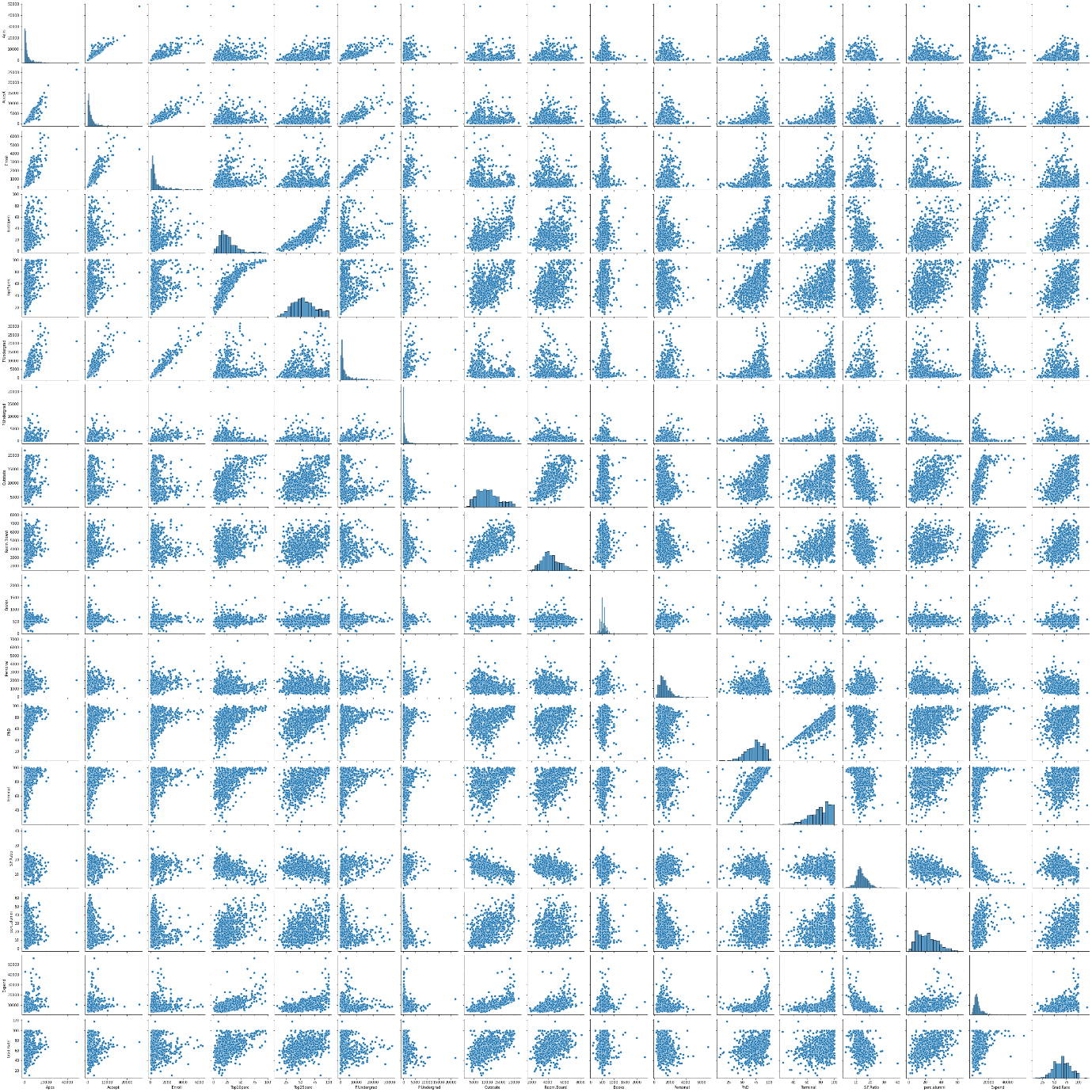
We observe from the plots that the following variables have outliers - Apps, Accept, Enroll, Top10perc, F.Undergrad, P.Undergrad, Outstate, Room.Board, Books, Personal, PhD,Terminal, S.F.Ratio, perc.alumni, Expend, Grad.Rate

Bi- Variate Analysis

* Heat map



* Pair plot



We can see that many columns are co-related to each other or in other words the correlation exists for many columns and the highest is between F.Undergrad and Enroll

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

Descriptive of measure before scaling the dataset



From descriptive measure of statistics, we can see that all categories are at different scale. We need to perform ZScore to make all category at same scale.

Scaling

Dataset after scaling



Descriptive of measure after scaling the dataset



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

Covariance Matrix

[[ 1.00128866 0.94466636 0.84791332 0.33927032 0.35209304 0.81554018

0.3987775 0.05022367 0.16515151 0.13272942 0.17896117 0.39120081

0.36996762 0.09575627 -0.09034216 0.2599265 0.14694372]

[ 0.94466636 1.00128866 0.91281145 0.19269493 0.24779465 0.87534985

0.44183938 -0.02578774 0.09101577 0.11367165 0.20124767 0.35621633

0.3380184 0.17645611 -0.16019604 0.12487773 0.06739929]

[ 0.84791332 0.91281145 1.00128866 0.18152715 0.2270373 0.96588274

0.51372977 -0.1556777 -0.04028353 0.11285614 0.28129148 0.33189629

0.30867133 0.23757707 -0.18102711 0.06425192 -0.02236983]

[ 0.33927032 0.19269493 0.18152715 1.00128866 0.89314445 0.1414708

-0.10549205 0.5630552 0.37195909 0.1190116 -0.09343665 0.53251337

0.49176793 -0.38537048 0.45607223 0.6617651 0.49562711]

[ 0.35209304 0.24779465 0.2270373 0.89314445 1.00128866 0.19970167

-0.05364569 0.49002449 0.33191707 0.115676 -0.08091441 0.54656564

0.52542506 -0.29500852 0.41840277 0.52812713 0.47789622]

[ 0.81554018 0.87534985 0.96588274 0.1414708 0.19970167 1.00128866

0.57124738 -0.21602002 -0.06897917 0.11569867 0.31760831 0.3187472

0.30040557 0.28006379 -0.22975792 0.01867565 -0.07887464]

[ 0.3987775 0.44183938 0.51372977 -0.10549205 -0.05364569 0.57124738

1.00128866 -0.25383901 -0.06140453 0.08130416 0.32029384 0.14930637

0.14208644 0.23283016 -0.28115421 -0.08367612 -0.25733218]

[ 0.05022367 -0.02578774 -0.1556777 0.5630552 0.49002449 -0.21602002

-0.25383901 1.00128866 0.65509951 0.03890494 -0.29947232 0.38347594

0.40850895 -0.55553625 0.56699214 0.6736456 0.57202613]

[ 0.16515151 0.09101577 -0.04028353 0.37195909 0.33191707 -0.06897917

-0.06140453 0.65509951 1.00128866 0.12812787 -0.19968518 0.32962651

0.3750222 -0.36309504 0.27271444 0.50238599 0.42548915]

[ 0.13272942 0.11367165 0.11285614 0.1190116 0.115676 0.11569867

0.08130416 0.03890494 0.12812787 1.00128866 0.17952581 0.0269404

0.10008351 -0.03197042 -0.04025955 0.11255393 0.00106226]

[ 0.17896117 0.20124767 0.28129148 -0.09343665 -0.08091441 0.31760831

0.32029384 -0.29947232 -0.19968518 0.17952581 1.00128866 -0.01094989

-0.03065256 0.13652054 -0.2863366 -0.09801804 -0.26969106]

[ 0.39120081 0.35621633 0.33189629 0.53251337 0.54656564 0.3187472

0.14930637 0.38347594 0.32962651 0.0269404 -0.01094989 1.00128866

0.85068186 -0.13069832 0.24932955 0.43331936 0.30543094]

[ 0.36996762 0.3380184 0.30867133 0.49176793 0.52542506 0.30040557

0.14208644 0.40850895 0.3750222 0.10008351 -0.03065256 0.85068186

1.00128866 -0.16031027 0.26747453 0.43936469 0.28990033]

[ 0.09575627 0.17645611 0.23757707 -0.38537048 -0.29500852 0.28006379

0.23283016 -0.55553625 -0.36309504 -0.03197042 0.13652054 -0.13069832

-0.16031027 1.00128866 -0.4034484 -0.5845844 -0.30710565]

[-0.09034216 -0.16019604 -0.18102711 0.45607223 0.41840277 -0.22975792

-0.28115421 0.56699214 0.27271444 -0.04025955 -0.2863366 0.24932955

0.26747453 -0.4034484 1.00128866 0.41825001 0.49153016]

[ 0.2599265 0.12487773 0.06425192 0.6617651 0.52812713 0.01867565

-0.08367612 0.6736456 0.50238599 0.11255393 -0.09801804 0.43331936

0.43936469 -0.5845844 0.41825001 1.00128866 0.39084571]

[ 0.14694372 0.06739929 -0.02236983 0.49562711 0.47789622 -0.07887464

-0.25733218 0.57202613 0.42548915 0.00106226 -0.26969106 0.30543094

0.28990033 -0.30710565 0.49153016 0.39084571 1.00128866]]

Correlation Matrix



Covariance indicates the direction of the linear relationship between variables. Correlation on the other hand measures both the strength and direction of the linear relationship between two variables. Correlation is a function of the covariance. You can obtain the correlation coefficient of two variables by dividing the covariance of these variables by the product of the standard deviations of the same values.

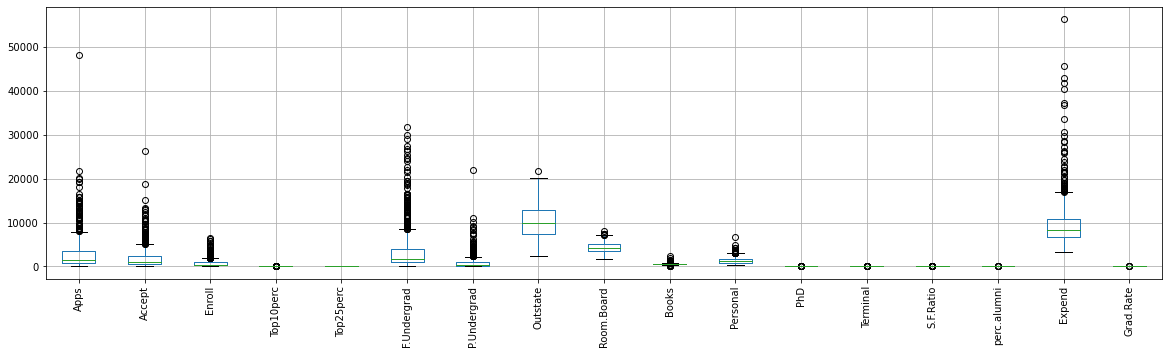
We can state that above three approaches yield the same eigenvectors and eigenvalue pair:

* Eigen decomposition of the covariance matrix after standardizing the data.
* Eigen decomposition of the correlation matrix after standardizing the data.

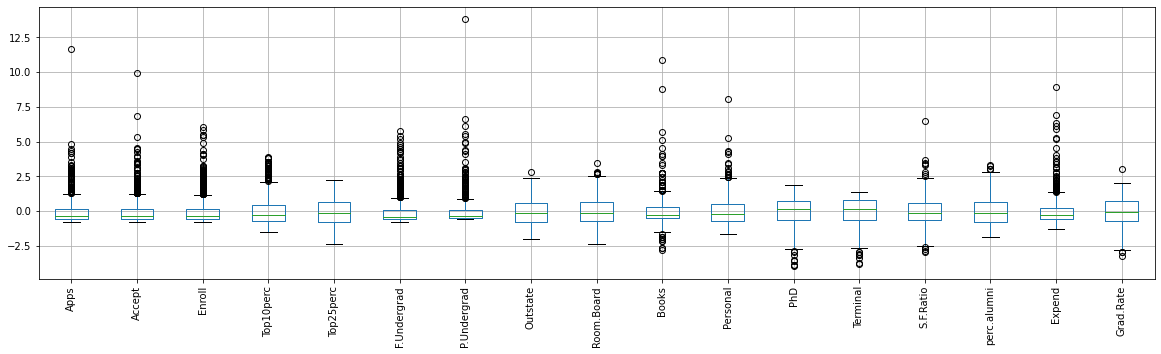
Finally, we can say that after scaling - the covariance and the correlation have the same values

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

Box-plot before scaling



Box-plot after scalling



As we can see box-plot of dataset after scaling and before scaling, mean tends to zero and ranges of all category are also became same.

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

Eigen values

[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]

Eigen vectors

[[-1.59285540e+00 7.67333510e-01 -1.01073537e-01 ... 1.75239502e-03

-9.31400698e-02 9.35522023e-02]

[-2.19240180e+00 -5.78829984e-01 2.27879812e+00 ... 1.03709803e-01

-5.02556890e-02 -1.74057054e-01]

[-1.43096371e+00 -1.09281889e+00 -4.38092811e-01 ... -2.25582869e-02

-4.05268301e-03 3.75875882e-03]

...

[-7.32560596e-01 -7.72352397e-02 -4.05641899e-04 ... 6.79013123e-02

-2.32023970e-01 -9.99380421e-02]

[ 7.91932735e+00 -2.06832886e+00 2.07356368e+00 ... 3.53597440e-01

3.04416200e-01 3.35104811e-01]

[-4.69508066e-01 3.66660943e-01 -1.32891515e+00 ... -1.14873492e-01

-1.17076127e-01 -2.57218339e-03]]

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

Percentage of variance in PCA components

[0.32020628, 0.26340214, 0.06900917, 0.05922989, 0.05488405,

0.04984701, 0.03558871, 0.03453621, 0.03117234, 0.02375192,

0.01841426, 0.01296041, 0.00985754, 0.00845842, 0.00517126,

0.00215754, 0.00135284]

Cumulative Variance Explained

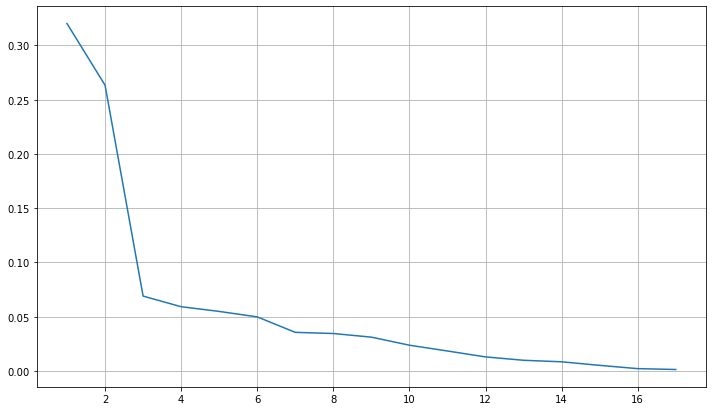
[0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154,

0.81657854, 0.85216726, 0.88670347, 0.91787581, 0.94162773,

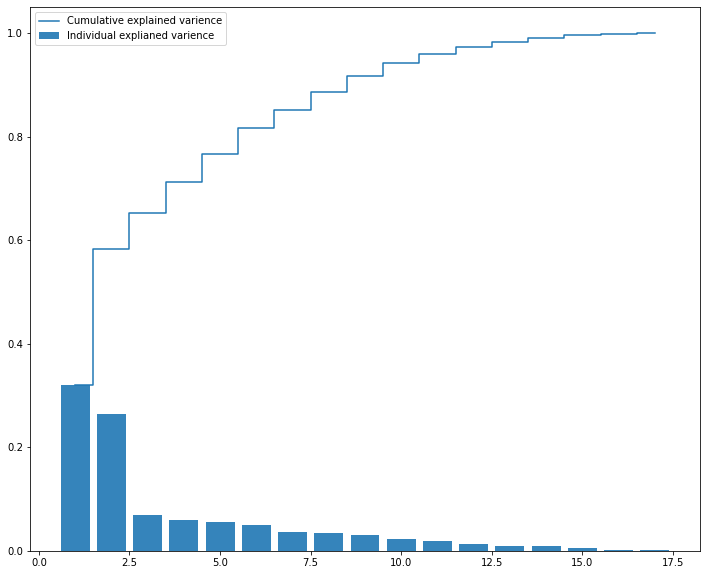
0.96004199, 0.9730024 , 0.98285994, 0.99131837, 0.99648962,

0.99864716, 1. ]

Scree plot



Cumulative explained variance and individual explained variance vs Principal Components



Visually we can observe that their is step drop in variance explained with increase in number of PC's.

We will proceed with 7 components here. But depending on requirement 88% variation or 8 components will also do good

Principal Component (eigenvectors) into a data frame with the original features



## 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]

Loading or weights of each feature corresponding to Eigen vector/component

[[ 0.2487656 , 0.2076015 , 0.17630359, 0.35427395, 0.34400128,

0.15464096, 0.0264425 , 0.29473642, 0.24903045, 0.06475752,

-0.04252854, 0.31831287, 0.31705602, -0.17695789, 0.20508237,

0.31890875, 0.25231565],

[ 0.33159823, 0.37211675, 0.40372425, -0.08241182, -0.04477866,

0.41767377, 0.31508783, -0.24964352, -0.13780888, 0.05634184,

0.21992922, 0.05831132, 0.04642945, 0.24666528, -0.24659527,

-0.13168986, -0.16924053],

[-0.0630921 , -0.10124906, -0.08298557, 0.03505553, -0.02414794,

-0.06139298, 0.13968172, 0.04659887, 0.14896739, 0.67741165,

0.49972112, -0.12702837, -0.06603755, -0.2898484 , -0.14698927,

0.22674398, -0.20806465],

[ 0.28131053, 0.26781735, 0.16182677, -0.05154725, -0.10976654,

0.10041234, -0.15855849, 0.13129136, 0.18499599, 0.08708922,

-0.23071057, -0.53472483, -0.51944302, -0.16118949, 0.01731422,

0.07927349, 0.26912907],

[ 0.00574141, 0.05578609, -0.05569364, -0.39543434, -0.42653359,

-0.04345437, 0.30238541, 0.222532 , 0.56091947, -0.12728883,

-0.22231102, 0.14016633, 0.20471973, -0.07938825, -0.21629741,

0.07595812, -0.10926791],

[-0.01623744, 0.00753468, -0.04255798, -0.0526928 , 0.03309159,

-0.04345423, -0.19119858, -0.03000039, 0.16275545, 0.64105495,

-0.331398 , 0.09125552, 0.15492765, 0.48704587, -0.04734001,

-0.29811862, 0.21616331],

[-0.04248635, -0.01294972, -0.02769289, -0.16133207, -0.11848556,

-0.02507636, 0.06104235, 0.10852897, 0.20974423, -0.14969203,

0.63379006, -0.00109641, -0.02847701, 0.21925936, 0.24332116,

-0.22658448, 0.55994394]]

First row of original scaled data = a



Loadings for Eigen vector/component 0 (PC0) = x

[ 0.2487656 , 0.2076015 , 0.17630359, 0.35427395, 0.34400128,

0.15464096, 0.0264425 , 0.29473642, 0.24903045, 0.06475752,

-0.04252854, 0.31831287, 0.31705602, -0.17695789, 0.20508237,

0.31890875, 0.25231565]

PC0 - New value (PC0 Score for 1st row)

=

= (0.25\*-0.346882) + (0.20\*-0.321205) + (0.18\*-0.063509) +

(0.35\*-0.258583) + (0.34\*-0.191827) + (0.15\*-0.168116) +

(0.03\*-0.209207) + (0.29\*-0.746356) + (0.25\*-0.964905) +

(0.06\*-0.602312) + (-0.04\*1.270045) + (0.32\*-0.163028) +

(0.32\*-0.115729) + (-0.18\*1.013776) + (0.21\*-0.867574) +

(0.32\*-0.50191) + (0.25\*-0.318252)

= -1.5882686299999997

## 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 of variance explained with [n] features

[32. , 58.3, 65.2, 71.1, 76.6, 81.6, 85.2]

Eigen vectors

[[-1.59285540e+00 7.67333510e-01 -1.01073537e-01 ... -7.43975398e-01

-2.98306081e-01 6.38443468e-01]

[-2.19240180e+00 -5.78829984e-01 2.27879812e+00 ... 1.05999660e+00

-1.77137309e-01 2.36753302e-01]

[-1.43096371e+00 -1.09281889e+00 -4.38092811e-01 ... -3.69613274e-01

-9.60591689e-01 -2.48276091e-01]

...

[-7.32560596e-01 -7.72352397e-02 -4.05641899e-04 ... -5.16021118e-01

4.68014248e-01 -1.31749158e+00]

[ 7.91932735e+00 -2.06832886e+00 2.07356368e+00 ... -9.47754745e-01

-2.06993738e+00 8.33276555e-02]

[-4.69508066e-01 3.66660943e-01 -1.32891515e+00 ... -1.13217594e+00

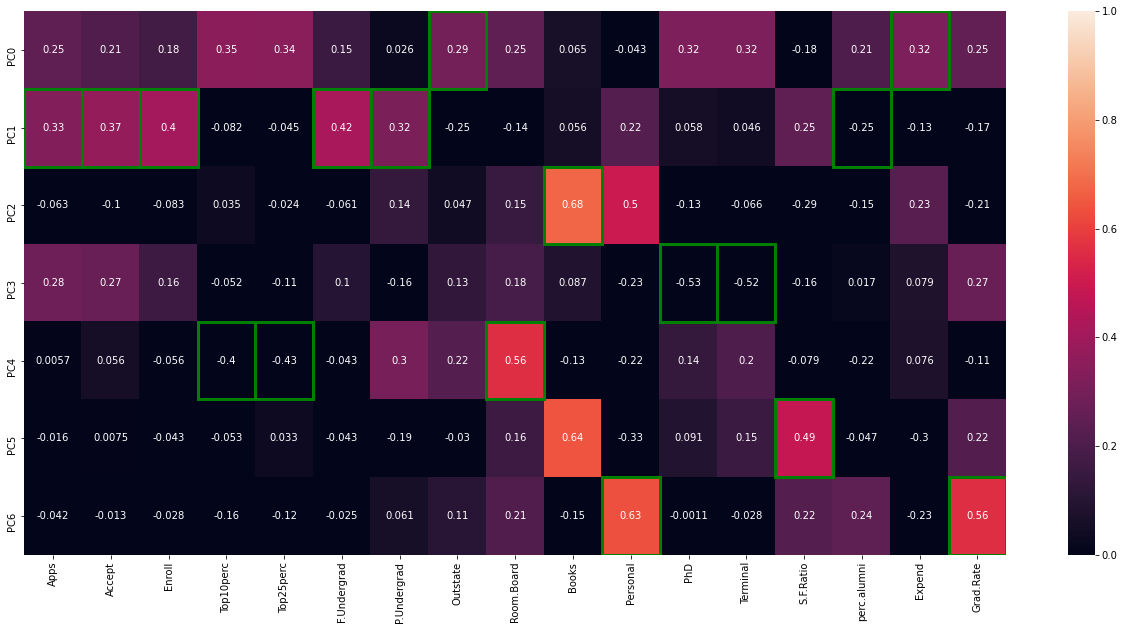
8.39893087e-01 1.30731260e+00]]

The Cumulative % gives the percentage of variance accounted for by the n components. For example, the cumulative percentage for the second component is the sum of the percentage of variance for the first and second components. It helps in deciding the number of components by selecting the components, which explained the high variance.

In the above array, we see that the first feature explains 32% of the variance within our data set while the first two explain 58.3 and so on. If we employ seven features we capture ~ 85% of the variance within the dataset, thus we gain very little by implementing an additional feature.

The eigenvectors (principal components) determine the directions of the new feature space

## 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]



This heatmap and the color bar basically represent the correlation between the various feature and the principal component itself

1. PC0 looks more related to Expenditure and number of student out of state we can rename it with Exp.Outstate
2. PC1 looks more related to apps, accept, enroll, full time and part time undergratuate and percentage of alumani who donated we can rename it with student
3. PC2 looks more related to books we can rename it with books
4. PC3 looks more related to percentage of faculties with PhD and Terminal degrees we can rename it with faculty
5. PC4 looks more related to percentage of students top 10% and 25% of higher secondary class and cost of room/band we can rename it with top.room.band
6. PC5 looks more related to student/faculty ratio we can rename it with S.F.Ratio
7. PC6 looks more related to personal spending for students and graduation rate we can rename it with personal.Grad

New dataset after concatenating PCA and original dataset

