**Project: Advanced Statistics**

Pavan Kalyan B | PGP DSBA Oct\_C’21 | February 2022

**Table of contents**

**I ANOVA**

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

Education and Occupation individually. ---------------------------------------------------------------4

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. --------5

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 -5

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. -----------------------------------------------------------5

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. -6

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? --------7

1.7 Explain the business implications of performing ANOVA for this particular case study. ---8

**II EDA & PCA**

2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA? --------------------------------------------9

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] -------------------------------------------------------------------------------15

2.4 Check the dataset for outliers before and after scaling. What insight do you derive here? ----------------------------------------------------------------------------------------------------------15

2.5 Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both] -----------------17

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] ----------------------------------------------------------19

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 Figures**

Fig 1: 1.5.1 Interaction between Education and Occupation -----------------------------------------------6

Fig 2: 2.1.2 Boxplots for univariate analysis -------------------------------------------------------------------10

Fig 3: 2.1.3 Histograms for univariate analysis ----------------------------------------------------------------11

Fig 4: 2.1.4 pair plot for multivariate analysis -----------------------------------------------------------------12

Fig 5: 2.1.5 heatmap for multivariate analysis ----------------------------------------------------------------13

Fig 6: 2.4.1 boxplots having outliers -----------------------------------------------------------------------------15

Fig 6: 2.4.2 boxplots after removing outliers ------------------------------------------------------------------16

Fig 7: 2.6.1 heatmap for 8 components reduction -----------------------------------------------------------18

Fig 8: 2.8.1 principal component vs explained\_variance\_ratio\_ -----------------------------------------19

**List of Tables**

Table 1: 1.2.1 Anova for Education --------------------------------------------------------------------------------5

Table 2: 1.3.1 Anova for Occupation ------------------------------------------------------------------------------5

Table 3: 1.4.1 For Education ----------------------------------------------------------------------------------------6

Table 4: 1.4.2 For Occupation --------------------------------------------------------------------------------------6

Table 5: 1.6.1 Two-way Anova without interaction effect ---------------------------------------------------7

Table 6: 1.6.2 Two-way Anova with interaction effect -------------------------------------------------------7

Table 7: 2.1.1 Summary of data ------------------------------------------------------------------------------------9

Table 8: 2.2.1 Summary of data before scaling ---------------------------------------------------------------14

Table 8: 2.2.2 Summary of data after scaling -----------------------------------------------------------------14

Table 9: 2.5.1 eigen vectors and eigen values ----------------------------------------------------------------17

Table 10: 2.6.1 eigen vectors into data frame for 17PCS ---------------------------------------------------17

Table 11: 2.6.2 eigen vectors into data frame for 8PCS ----------------------------------------------------18

Table 12: 2.7.1 eigen vectors for first PC -----------------------------------------------------------------------19

Table 13: 2.8.1 eigen vectors for first PC17 to 8pcs coverage --------------------------------------------20

**Problem 1A:**

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [[SalaryData.csv](https://olympus.mygreatlearning.com/courses/63574/files/4516098/download?verifier=TFJD7aifuuAX5w8gABWgnmMQBJaW271XGI9y6RMk&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.

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

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

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

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

**Answer:**

**For Education:**

Null Hypothesis

H0: Mu1 = Mu2 = Mu3

Alternate Hypothesis

Ha: At least one of the means for education is different

**For Occupation:**

Null Hypothesis

H0: Mu1 = Mu2 = Mu3 = Mu4

Alternate Hypothesis

Ha: At least one of the means for occupation is different

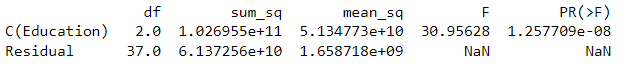
Alpha – 0.05

If Pvalue < 0.05, Then we reject Null hypothesis H0

If Pvalue >= 0.05, Then we fail to reject Null hypothesis H0

* 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.**

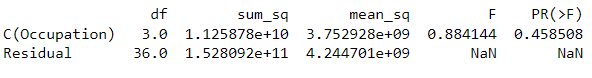
Table 1: 1.2.1 Anova for Education



Since the p-value is less than Alpha, we reject Null Hypothesis (H0) for Education

* 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.**

Table 2: 1.3.1 Anova for Occupation



Since the p-value is greater than Alpha, we fail to reject Null Hypothesis (H0) for Occupation

* 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.**

ANOVA tells us whether our results are significant or not, but it does not tell us about where the results are significant. we must use a Tukey Test to find out where the statistical significance is occurring in our data.

Table 3: 1.4.1 For Education

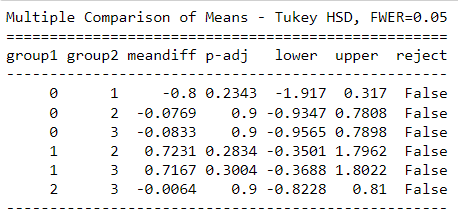
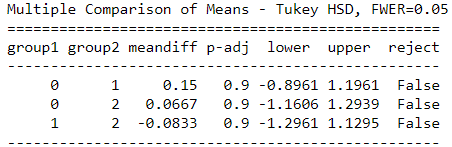
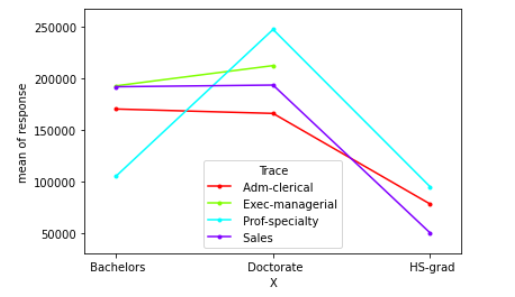


Table 4: 1.4.2 For Occupation



* 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.**

Fig 1: 1.5.1 Interaction between Education and Occupation



Observations:

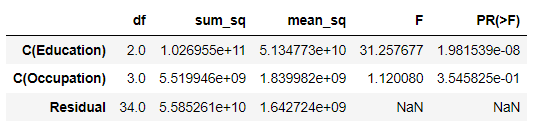
From above plot we can make out that the interaction between people with:

* Bachelors with Adam-clerical, Exec-managerial and sales are fairly good.
* Doctorate with prof-speciality, Adam-clerical, Exec-managerial and sales is absolutely good than the other degrees.
* Hs-grad with Adam-clerical, Exec-managerial and sales are less than other education levels.
* Exec-Managerial job role has no interactions with any other educational background.

With respect to income:

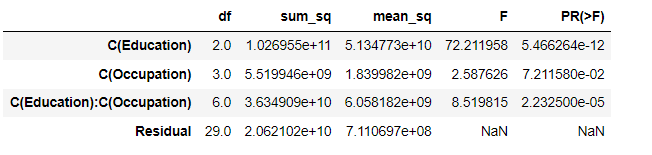
* Doctorate with prof-speciality has the higher income of all
* Doctorate with Exec-managerial has the income higher than the bachelors with Exec-managerial.
* Hs-grad people with all occupation has lesser income
* Bachelors with prof-speciality has lesser income than the doctorate people but higher than the Hs-grad.
* Sales with doctorate and bachelors degree has nearest salary equal
  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?**

Table 5: 1.6.1 Two-way Anova without interaction effect



When we perform Two-way anova without interaction effect, there is some sort of interactions between the education and occupation.

Table 6: 1.6.2 Two-way Anova with interaction effect



Due to the inclusion of the interaction effect term, [ education : occupation ] we can see a slight change in the p-value of the first two treatments as compared to the Two-Way ANOVA without the interaction effect terms.

we see that the p-value of the interaction effect term of 'Education' and 'Occupation' suggests that we reject the Null Hypothesis.

* 1. **Explain the business implications of performing ANOVA for this particular case study.**
* As the Education level increases the salary gets increases.
* Doctorates are suitable for all the roles and they are getting paid highly than anyone.
* We must also note that occupation and education alone impact the salary, there are few more other important variables which can impact salary such as years of experience, specialisation, past experience, extra courses, certifications, etc.
* To decide the salary HR department plays important role by setting up salary based on the experiences.

Anoav compares means, if the means are organized through one factore, it is one way anova. If these means are corresponded to two different treatements. This is called Two-way Anova. With or without interaction. Anova executes the business decision by simply comparing an associated p-value of the F-statistic with a significance level.

Assumptions of Anova:

* Dependent variables should be measured at continuous level.
* Two independent variables should each consist of tow or more independent groups.
* There should be no outliers.
* Dependent variable should be normally distributed.

**==============================================================================**

**Problem 2:**

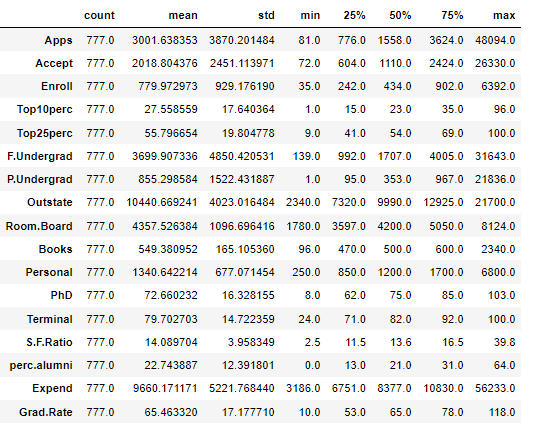
The dataset [Education - Post 12th Standard.csv](https://olympus.mygreatlearning.com/courses/63574/files/3937740/download?verifier=ZuN4KxxuNJ8ywK9WqSC2rDOp9Oj3aL5Io893sFBd&wrap=1) contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according 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.mygreatlearning.com/courses/63574/files/3937739/download?verifier=xvLtBsqcFLpysaaTwZv3S7DddMX0VEe8qI5kEIRF&wrap=1).

* Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?
* Is scaling necessary for PCA in this case? Give justification and perform scaling.
* Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].
* 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]
* Extract the eigenvalues and eigenvectors.[Using Sklearn PCA Print Both]
* Perform PCA and export the data of the 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]
* 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?
* 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]

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

**Univariate Analysis:**

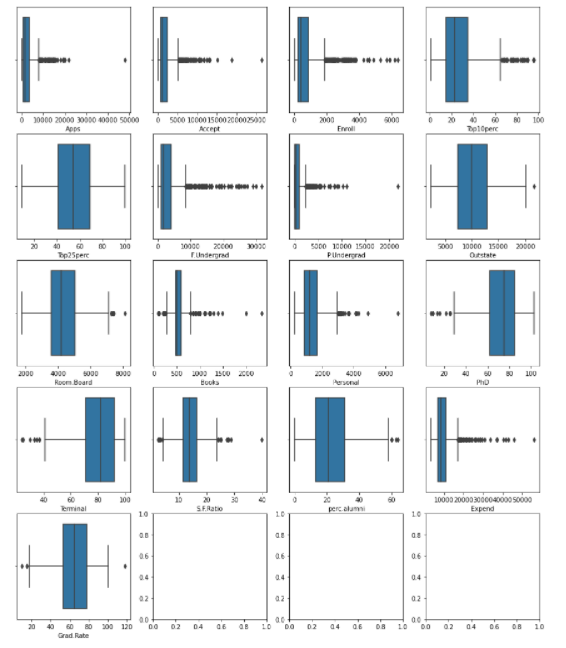
Table 7: 2.1.1 Summary of data



Univariate analysis revers to analysis of single variable. The main purpose is to summarise and find patterns in data. The statistical description of the numeric variable, histogram or distplot etc..

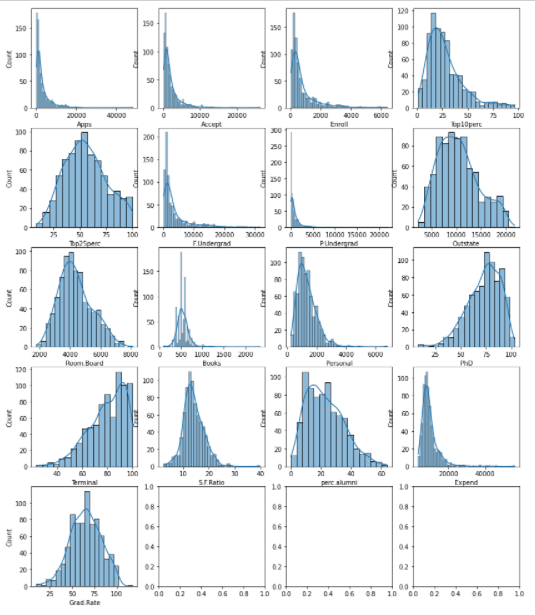
* Data has 777 records for all variables and doesn’t have any null values.
* Apps, Accept, F.undergrad, P.undergrad, Outstate, Personal, Expend – these all variables needs a clean up

Fig 2: 2.1.2 Boxplots for univariate analysis



* From the above plot we can observe that only ‘Top25perc’ doesn’t have any outlier, where other all variables have outliers.
* We have to treat those outliers by using scaling.

Fig 3: 2.1.3 Histograms for univariate analysis

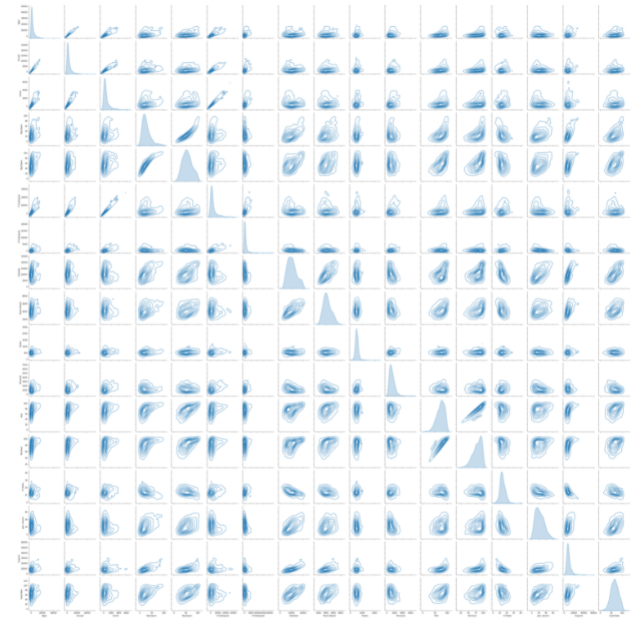


Observations:

* Data distributed on right side: 'PhD', 'Terminal.
* Data distributed on left side: Apps', 'Accept', 'Enroll', 'Top10perc', 'F.Undergrad', 'P.Undergrad', 'Books', 'Personal', 'S.F.Ratio','perc.alumni', 'Expend'.
* Data is symmetric: Top25perc , 'Outstate', ','Grad.Rate', , 'Room.Board'

**Multi-variate Analysis:**

Fig 4: 2.1.4 pair plot for multivariate analysis

****

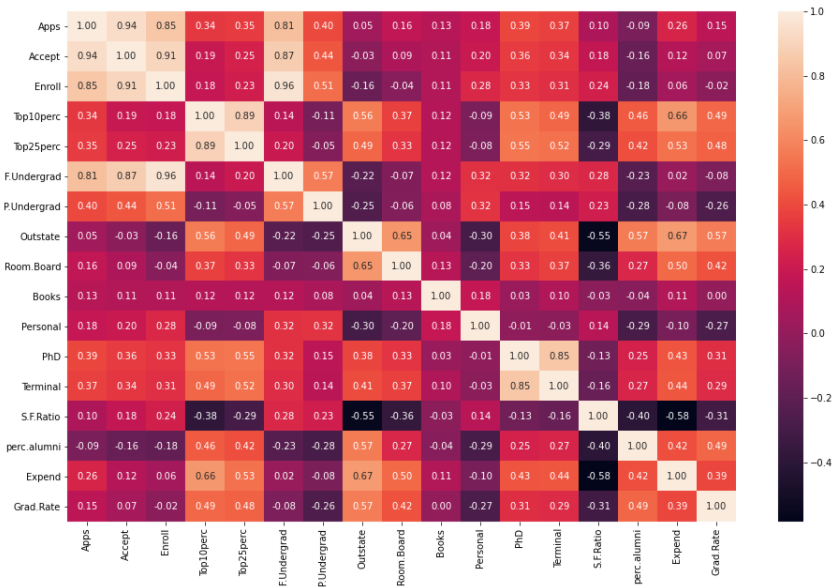
**Observation:**

Some pairs are correlated:

* Apps and Accept are highly correlated
* Accept and Enroll are highly correlated
* Top25perc and Top10perc is correlated
* Accept and F.undergrad is correlated
* Expend and Room.Board are correlated

As many of the variables are correlated, these fall under the multicollinearity and hence we these fall under the curse of dimensionality.

Fig 5: 2.1.5 heatmap for multivariate analysis



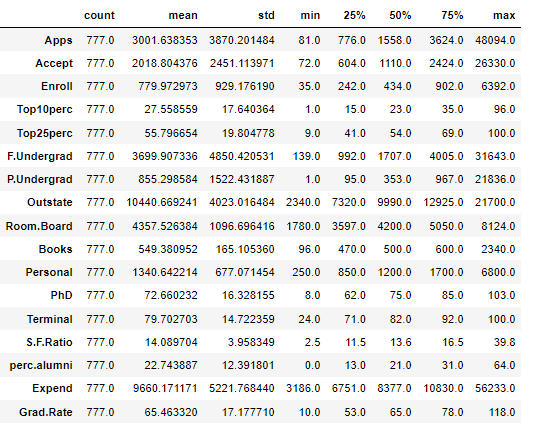
The above heat map shows the high number of correlations and these fall under the category of multicollinearity.

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

* Data set contains 18 attributes and hence we remove the ‘Name’ column as it is unique for all.
* We continue the process with 17 attributes and perform pca
* For 17 attributes we get 17 PC’S
* Scaling is necessary before performing PCA.
* As we perform scaling the data falls into new projection and those range between -3 to 3.
* This normalizing tunes the PCA towards high magnitude features. We can speed up gradient descent and model calculations by scaling.
* Using zscore we scale the data
* Where z = (x-xbar)/(std/n)

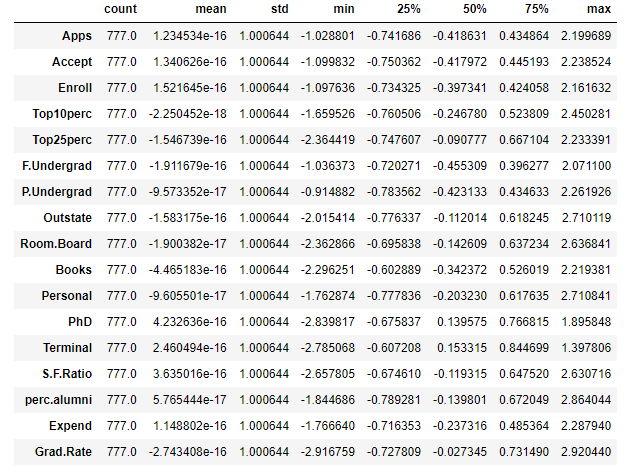
**Before scaling:**

Table 8: 2.2.1 Summary of data before scaling



**After scaling:**

Table 8: 2.2.2 Summary of data after scaling



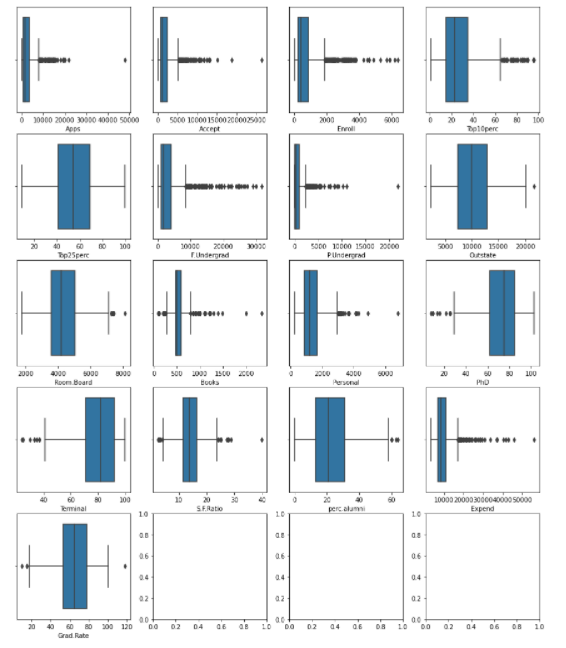
**2.3 Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data]**

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

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

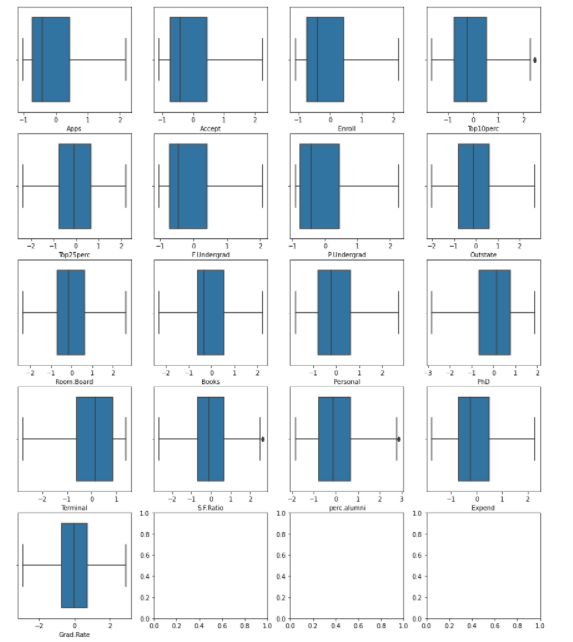
**Before Scaling checking outliers:**

Fig 6: 2.4.1 boxplots having outliers



**After Scaling checking outliers:**

Fig 6: 2.4.2 boxplots after removing outliers

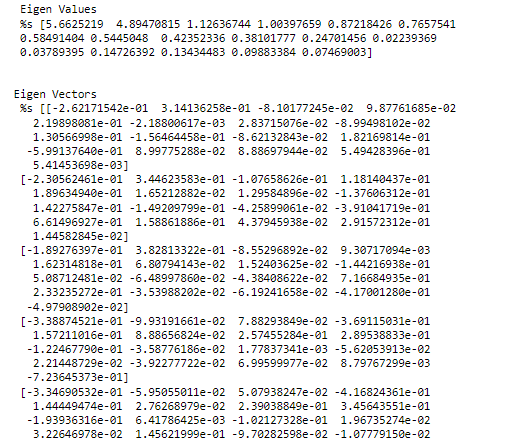
****

While performing Univariate Analysis we have plotted Boxplots for all the variables for checking of Outliers presence. After scaling no much difference in terms of outliers reduction.

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

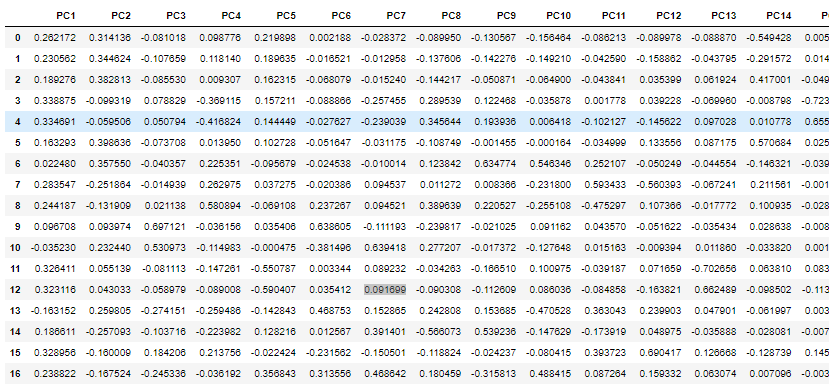
* From the covariance matrix, we derive eigen values and eigen vectors

Table 9: 2.5.1 eigen vectors and eigen values



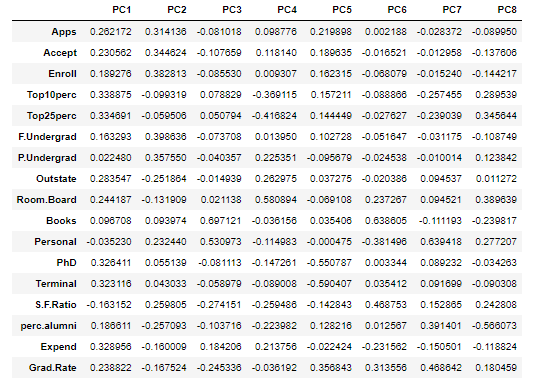
**2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features**

Table 10: 2.6.1 eigen vectors into data frame for 17PCS

****

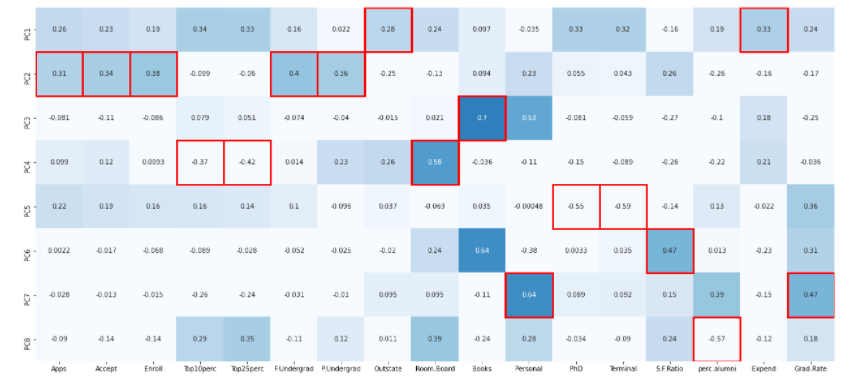
Here we have 17 pcs extracted with 17 attributes

Table 11: 2.6.2 eigen vectors into data frame for 8PCS



Here we have reduced the dimensions or components from 17 to 8 principal components.

Fig 7: 2.6.1 heatmap for 8 components reduction



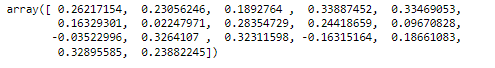
This shows the reduction from 17 pcs to 8 pcs with in the rectangular box which has the highest value.

**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]**

The first principal component PC1 represents the component that has the maximum variance of the data. 33% of data has been covered in pc1.

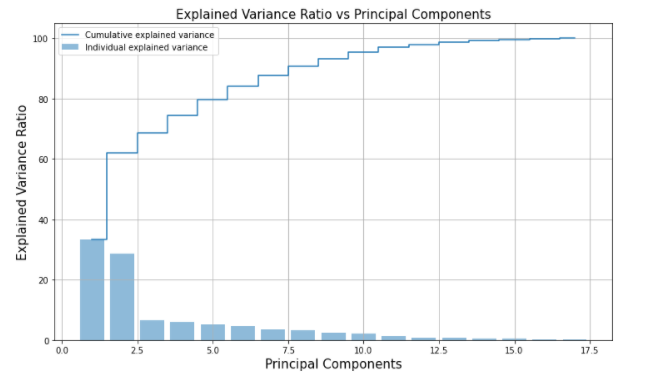
Linear equation of pc1 = a1\*x1 + a2\*x2 + a3\*x3+ … +an\*xn

Table 12: 2.7.1 eigen vectors for first PC



**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?**

Fig 8: 2.8.1 principal component vs explained\_variance\_ratio\_



The plot visually shows how much of variance is allocated to each principal component. For the 1st PC variance 33%, second PC explains 62% and so on.

Table 13: 2.8.1 eigen vectors for first PC17 to 8pcs coverage



By analysing 8 Principle components we covered upto 90% of data instead of the 17 attributes in dataset.

**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]**

The business implication of using the PCA :

* PCA is used for dimensionality reduction
* PCA increases the signal content that is provided to your algorithm to build your model.
* PCA helps to eliminate the unwanted, unnecessary dimensions
* when the two independent variables are strongly interacting with each other, where 'r' value close to 1. then we are providing the same amount of information to our algorithm in two dimensions. these are call redundant.
* when we have too many redundant dimensions, then we are exploring it ourself to curse of dimensionality.

Implications on case study:

* So, in our case study there are 18 attributes which are correlated to each and hence it is making the curse of dimensionality. That is why we are reducing 18 to 8 attributes and hence
* The first component contained 33% of data and second component has 62% and so on, after 90% there was no much variation in the projected data, so we stopped with 8 components.

**==============================================================================**