Advanced Statistics Project
Module 3 – Advanced Statistics

Submitted to

Great Learning

By
……**DHRUV DIWAN**……

Sun group 1- B

Post Graduate Program in Data Science and Business Analytics

Mentor
……Mr. Aniket Goel……..

Date of submission April 2021

Table of Contents

List of Tables and Figures

Problem Statement

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

Theory

ANOVA

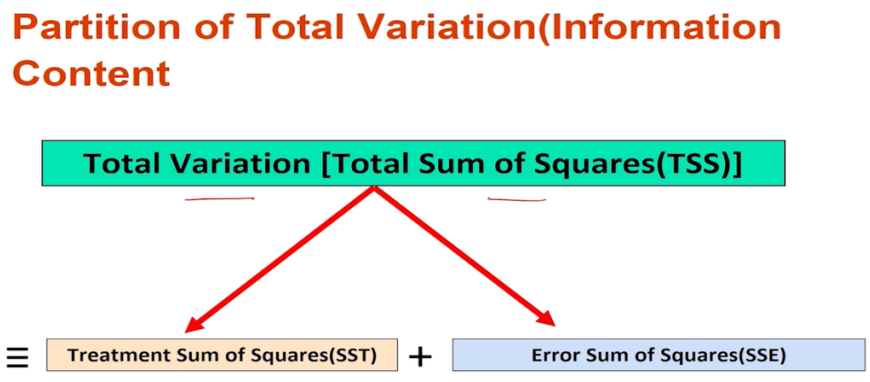
- An experiment has been carried out for the purpose to compare means
- They are having a strong cause and effect relationship
- Randomisation-
- Compare more than 2 Population Means

Assumptions-

- The samples drawn from different populations are independent variables and random
- The response variables are continuous and ideally normally distributed(if not the test doesn't go wrong but may be subject to variation) – Plot a normal Distribution)
- The variance of all the population are equal(only in means not in variance)

ANOVA decomposes in two halves

1. Differences in all the distributions
2. Due to variations in the data



Introduction

The purpose of this whole exercise is to explore the dataset. Do the Hypothesis test analysis. Here our main aim is to understand the relationship between Independent and Dependent variables. The data consists of Salaries of 40 different individuals and their Occupation and Education. This will help us understand and compare the means of different Continuous variables with independent variables.

Describe and Sample Data

	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769

	Salary
count	40.000000
mean	162186.875000
std	64860.407506
min	50103.000000
25%	99897.500000
50%	169100.000000
75%	214440.750000
max	260151.000000

Table 1 – Description(Left) and sample(Right) of the dataset Salary

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Education   40 non-null     object
 1   Occupation  40 non-null     object
 2   Salary      40 non-null     int64
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

Table 2 – Informative Summary of the dataset Salary

Inferences from the above tables -

- Describe and Info process clearly shows that there are 2 object type and 1 integer type which means 2 category or independent factors and 1 continuous variable. Therefore there is no need to convert categorical data. (No EDA Process Required)
- The Standard deviation is 64860.41. So Now we have to check whether this comes from Occupation or education or is there any interaction between the two
- The mean(162186.87) and median(169100) are quite similar and we can assume it to be normal distribution
- OLS is ordinary least squares
- Education and Occupation are object based or independent terms or factors and we need to check whether they are related to a continuous variable integer i.e. Salary
- The shape is (40,3) which means there are 40 entities in each column.
- There are Non-Null Categories in each Columns

```
Prof-specialty     13
Sales              12
Adm-clerical       10
Exec-managerial     5
Name: Occupation, dtype: int64
```

```
Doctorate     16
Bachelors     15
HS-grad        9
Name: Education,
```

Table 3- Value count of each Independent variable

- The above table shows the Categories of each Independent Variables or factors
- Occupation is divided into 4 categories Prof-Speciality, Sales, Adm-Clerical, Exec-managerial while Education is divided into 3 categories Doctorate, Bachelors, HS-Grad
- The values of each person is explained above

Problem 1 A)

Q1.1) State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

Hypothesis test for one way ANOVA

1. **For Occupation with four levels:**

- NULL HYPOTHESIS- H_0: $\mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$

  The mean salary of different group of occupations are same or All Population means are equal

- ALTERNATE HYPOTHESIS- H_1: Means are not all equal.

  At least one of the group of occupation has mean salary different from other or For at least one pair, the populations means are unequal

2. **For education with 3 levels:**

- NULL HYPOTHESIS- H_0: $\mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$

  The mean salary of different group of education levels are same or All Population means are equal or Salary do not differ significantly among the educations.

- ALTERNATE HYPOTHESIS- H_1: Means are not all equal.

  For at least one pair, the populations means are unequal or At least one of the group of education has mean salary different from other or at least one education's salary differ signicantly.

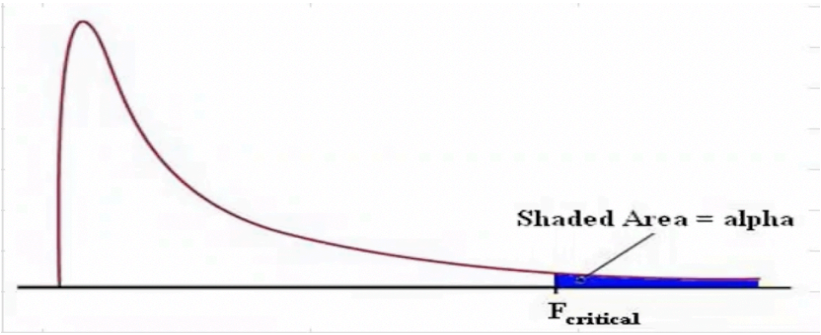
Q1.2). Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

So here we need to decide, is Education significant in relation with a continuous variable or dependent variable Salary.

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

Table 4 – One way ANOVA on Education

- **Degrees of Freedom** = N-1. Here Education has 3 categories so Degrees of freedom is 3-1=2. Total No. of observations consists of 40 total line items where 2 already accounted through education leaving 37 as residual.
- **Sums of Square** – The education categories (Doctorate, bachelors, Hs-Graduate) have 3 different means. Those sample means are compared to the overall means and that is 102695500000. It is also called the between sums of square. Then we have within Sums of Square in the residual output.
- **Mean Sums of Square** – Sums of Square/ Degree of Freedom =102695500000/2 = 51347750000. Same with residual
- **F-Statistic** – It is Education Mean sums of square/ Residual mean sums of square = 30.95682. This tells us that the variance between education Categories is about 30 times within each category of education
- **P-Value -** Since P value is less than Alpha, we reject the null hypothesis and we can clearly say that atleast one of the group of education qualifications has mean salary different from others. Therefore Education qualification(cause) is a differentiator in the salary (Effect)Structure.
- **F-Critical-** There is no output for F-critical in python. We can also calculate F-critical value through Excel Using F.INV.RT .



Note-

- **When the P_Value > Alpha, We Fail to reject the Null Hypothesis**
- **When the P_Value < Alpha, We Reject the Null Hypothesis**
- **If F-Statistic > F-Critical , We Reject the Null Hypothesis**
- **If F-Statistic > F-Critical, We fail to reject the Null Hypthesis**

Inference

Since P value is less than Alpha, we reject the null hypothesis and we can clearly say that at least one of the group of education qualifications has mean salary different from others. **Therefore Education qualification is a differentiator in the salary Structure.**

Q1.3). 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

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	0.884144	0.458508
Residual	36.0	1.528092e+11	4.244701e+09	NaN	NaN

Table 5 - One way ANOVA on Occupation

- **Degrees of Freedom** = N-1. Here Occupation has 4 categories so Degrees of freedom is 4-1=3. Total No. of observations consists of 40 total line items where 3 already accounted through Occupation leaving 36 as residual.
- **Sums of Square** – The Occupation categories (Adm-Clerical, Exec-managerial, Sales, Prof-Speciality) have 3 different means. Those sample means are compared to the overall means and that is 11258780000. It is also called the between sums of square. Then we have within Sums of Square in the residual output.
- **Mean Sums of Square** – Sums of Square/ Degree of Freedom = 11258780000/2 = 5629390000. Same with residual
- **F-Statistic** – It is Occupation Mean sums of square/ Residual mean sums of square = 0.884. This tells us that the variance between Occupation Categories

is about 0.884 times within each category of Occupation which is very less compared to Education as a factor

- **P-Value -** Since the p Value is greater than alpha i.e 0.05 therfore we fail to reject the null hypthesis where we can clearly state that mean salary of different group of occupation are the same. Therefore Occupation is not a differentiator in the salary Structure
- **F-Critical-** There is no output for F-critical in python. We can also calculate F-critical value through Excel Using F.INV.RT .

Note-

- **When the P_Value > Alpha, We Fail to reject the Null Hypothesis**
- **When the P_Value < Alpha, We Reject the Null Hypothesis**
- **If F-Statistic > F-Critical , We Reject the Null Hypothesis**
- **If F-Statistic > F-Critical, We fail to reject the Null Hypthesis**

Inference -

Since the P Value is greater than alpha i.e. 0.05 therefore we fail to reject the null hypothesis where we can clearly state that mean salary of different group of occupation are the same. Therefore Occupation is not a differentiator in the salary Structure.

Q1.4). 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.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Bachelors	Doctorate	43274.0667	0.0146	7541.1439	79006.9894	True
Bachelors	HS-grad	-90114.1556	0.001	-132035.1958	-48193.1153	True
Doctorate	HS-grad	-133388.2222	0.001	-174815.0876	-91961.3569	True

Table 6- Multiple Comparison For Education

Problem 1. B)

Q1.5) What is the interaction between two treatments? Analyse the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.

Two Way ANOVA-

- When multiple treatments which have two independent variables (called factors)
- To check If there is interaction between 2 independent variables and dependent variables

Assumptions

- Dependent variables are measured at a continuous level
- There should be no significant outliers
- Dependent variable should be normally distributed
- Randomisation is done before hand

I have performed a two-way ANOVA to Check the interaction between two treatments. This is required to check the interaction between the two by creating a Plot. Two-way ANOVA without Interaction is required **where we take Education as the first factor as Education has a P-Value less than Alpha.**

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	31.257677	1.981539e-08
C(Occupation)	3.0	5.519946e+09	1.839982e+09	1.120080	3.545825e-01
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

Table 7- Two-way ANOVA

Here if we take education and occupation together, we can see that Education's P value is 0.00000001981539 which is less than alpha while the value of occupation is 0.3545825 more than alpha. Hence even if we take education and occupation together education is alone a significant factor in determining salary of the sample Population.

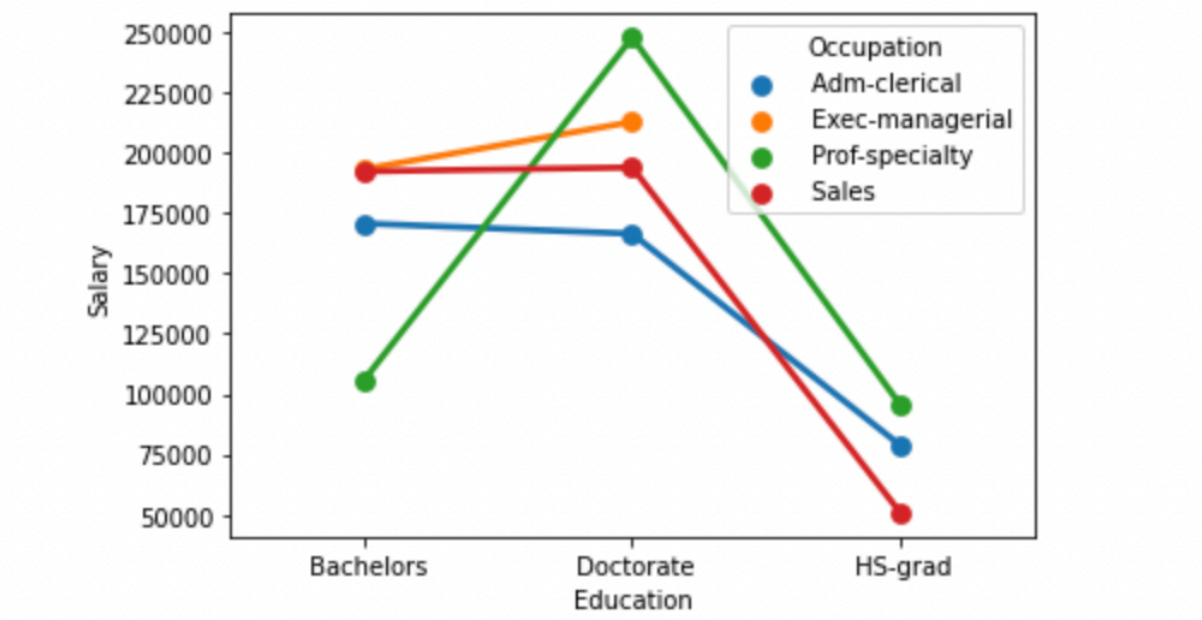


Figure 1 – Interaction Plot with Factor Education and Occupation together

Inference

From the above plot, there is not a significant interaction between the variables. We can clearly see that the salary at High School Graduate is the lowest be it in Occupation which is increasing as they have bachelor's degree while the highest with doctorate Degrees. While according, if we take it through Occupation Perspective, Prof-Specialty has the highest salary while High school students of working in sales have the lowest salary. We can also see that Even having a doctorate degree in the field of sales and Adm-Clerical departments, there is not much of difference with a bachelor's degree.

Q1.6 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?

There are three sets of hypothesis with the two-way ANOVA.

The null hypotheses and Alternate Hypothesis for each of the sets are given below.

1. **For Occupation**

- NULL HYPOTHESIS- H_0: $\mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$

  The mean salary of different group of occupations are same or All Population means are equal

- ALTERNATE HYPOTHESIS- H_1: Means are not all equal.

  At least one of the group of occupation has mean salary different from other or For at least one pair, the populations means are unequal

2. **For Education -**

- NULL HYPOTHESIS- H_0: $\mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$

  The mean salary of different group of education levels are same or All Population means are equal or Salary do not differ significantly among the educations.

- ALTERNATE HYPOTHESIS- H_1: Means are not all equal.

  For at least one pair, the populations means are unequal or At least one of the group of education has mean salary different from other or at least one education's salary differ signicantly.

3. **For Interaction between Education and Salary –**

- NULL HYPOTHESIS-
  There is no interaction between Occupation and Education
- ALTERNATE HYPOTHESIS-
  There is Interaction Between Occupation and Education

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Table 8– Two Way ANOVA with Interaction

- **Degrees of Freedom** = N-1. Here Occupation and Education together have 7 categories so Degrees of freedom is 7-1=6. Total No. of observations consists of 40 total line items where 5 already accounted through Occupation leaving 29 as residual.
- **Sums of Square** – The Occupation and Education categories together have different means. Those sample means are compared to the overall means and that is 36349090000. It is also called the between sums of square. Then we have within Sums of Square in the residual output.
- **Mean Sums of Square** – Sums of Square/ Degree of Freedom = 36349090000/2 = 6058182000.
- **F-Statistic** – It is Occupation and Education Together Mean sums of square/ Residual mean sums of square = 8.519815. This tells us that the variance between Occupation and Education Categories is about 8.5 times.
- **P-Value -** Since the p Value 0.000022325 is Less than alpha i.e 0.05 therefore we reject the null hypothesis where we can clearly state that mean salary of different group of occupation and Education together are Different. Therefore Occupation and Education have some kind of an Interaction. Due to the inclusion of the interaction effect term, 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 and we see that the p-value of the interaction effect term of 'Education' and 'Occupation' suggests that the Null Hypothesis is rejected in this case.

Q1.7). Explain the business implications of performing ANOVA for this particular case study.

- Education alone is a significant variable to salary
- Occupation has no effect on Salary while education and occupation together have an interaction
- Human Resource can use this method to implement Salary Structure
- We can find other factors to improve our output like Experience

EXPLORATORY DATA ANALYSIS AND PRINCIPAL COMPONENT ANALYSIS

Executive Summary

The dataset 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 is also available.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of 777 different colleges with 17 unique features. Analyse the different attributes of the Colleges and reducing the dimensions of the data by scaling and finding Eigen Values and Vectors and performing PCA .

Sample Data and Description

Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Abilene Christian niversity	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
Adelphi niversity	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
Alaska Pacific niversity	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Names	777	777	Western Washington University	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Apps	777	NaN	NaN	NaN	3001.64	3870.2	81	776	1558	3624	48094
Accept	777	NaN	NaN	NaN	2018.8	2451.11	72	604	1110	2424	26330
Enroll	777	NaN	NaN	NaN	779.973	929.176	35	242	434	902	6392
Top10perc	777	NaN	NaN	NaN	27.5586	17.6404	1	15	23	35	96
Top25perc	777	NaN	NaN	NaN	55.7967	19.8048	9	41	54	69	100
F.Undergrad	777	NaN	NaN	NaN	3699.91	4850.42	139	992	1707	4005	31643
P.Undergrad	777	NaN	NaN	NaN	855.299	1522.43	1	95	353	967	21836
Outstate	777	NaN	NaN	NaN	10440.7	4023.02	2340	7320	9990	12925	21700
Room.Board	777	NaN	NaN	NaN	4357.53	1096.7	1780	3597	4200	5050	8124
Books	777	NaN	NaN	NaN	549.381	165.105	96	470	500	600	2340
Personal	777	NaN	NaN	NaN	1340.64	677.071	250	850	1200	1700	6800
PhD	777	NaN	NaN	NaN	72.6602	16.3282	8	62	75	85	103
Terminal	777	NaN	NaN	NaN	79.7027	14.7224	24	71	82	92	100
S.F.Ratio	777	NaN	NaN	NaN	14.0897	3.95835	2.5	11.5	13.6	16.5	39.8
perc.alumni	777	NaN	NaN	NaN	22.7439	12.3918	0	13	21	31	64
Expend	777	NaN	NaN	NaN	9660.17	5221.77	3186	6751	8377	10830	56233
Grad.Rate	777	NaN	NaN	NaN	65.4633	17.1777	10	53	65	78	118

Table 9 – Sample and Description of the dataset

IF we go by the process of EDA

1. Know the Problem Statement or the Business Objective
2. Load and view the given data
3. Check the relevance of the data against the objective or goal to be achieved
    1. Scope of the data
    2. Time relevance of the data
    3. Quantum of data
    4. Features of the data
4. Understand each feature in the data with help of Data Dictionary
5. Know the central tendency and data distribution of each feature

Describe data

- We have 18 features and 777 Colleges
- The range or Min and Max are each taken accurately as per the requirement of the data.
- The Mean and the median are almost equal indicating that most of the numerical items are normally distributed. I
- t may be inferred that some of the variables look highly skewed like Apps, Accept, Enrol, F.Undergrad etc. This is expected as some colleges are expected to have greater number of applications, more Enrolment etc. Further analysis is done during univariate analysis.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 18 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Names        777 non-null    object 
 1   Apps         777 non-null    int64  
 2   Accept       777 non-null    int64  
 3   Enroll       777 non-null    int64  
 4   Top10perc    777 non-null    int64  
 5   Top25perc    777 non-null    int64  
 6   F.Undergrad  777 non-null    int64  
 7   P.Undergrad  777 non-null    int64  
 8   Outstate     777 non-null    int64  
 9   Room.Board   777 non-null    int64  
 10  Books        777 non-null    int64  
 11  Personal     777 non-null    int64  
 12  PhD          777 non-null    int64  
 13  Terminal     777 non-null    int64  
 14  S.F.Ratio    777 non-null    float64
 15  perc.alumni  777 non-null    int64  
 16  Expend       777 non-null    int64  
 17  Grad.Rate    777 non-null    int64  
dtypes: float64(1), int64(16), object(1)
memory usage: 109.4+ KB
```

Table 10 – Summarized information

Inference -

- We have all Non-Null Values so there are no Missing Values
- We have 3 Types of Data Types which Corresponds with the requirement of the dataset 16 Integer, 1 Object and 1 Float.

Data Dictionary

Data Dictionary	Type
1) Names: Names of various university and colleges	(Categorical)
2) Apps: Number of applications received	(Numerical)
3) Accept: Number of applications accepted	(Numerical)
4) Enroll: Number of new students enrolled	(Numerical)
5) Top10perc: Percentage of new students from top 10% of Higher Secondary class	(Numerical)
6) Top25perc: Percentage of new students from top 25% of Higher Secondary class	(Numerical)
7) F.Undergrad: Number of full-time undergraduate students	(Numerical)
8) P.Undergrad: Number of part-time undergraduate students	(Numerical)
9) Outstate: Number of students for whom the particular college or university is Out-of-state tuition	(Numerical)
10) Room.Board: Cost of Room and board	(Numerical)
11) Books: Estimated book costs for a student	(Numerical)
12) Personal: Estimated personal spending for a student	(Numerical)
13) PhD: Percentage of faculties with Ph.D.'s	(Numerical)
14) Terminal: Percentage of faculties with terminal degree	(Numerical)
15) S.F.Ratio: Student/faculty ratio	(Numerical)
16) perc.alumni: Percentage of alumni who donate	(Numerical)
17) Expend: The Instructional expenditure per student	(Numerical)
18) Grad.Rate: Graduation rate	(Numerical)

Table 11- Data Dictionary

Data Pre-processing

Practical data set generally has lot of "noise" and/or" undesired" data points which might impact the outcome, hence pre-processing is an important step As these "noise" elements are so well amalgamated with the complete dataset ,cleansing process is more governed by the data scientist ability. These noise elements are in the form of –

- Bad values
- Anomalies
- Missing values
- Not Useful Data

There are No Anomalies Or bad Values. Also There are no Duplicate Values so there is no need for Data Pre-Processing

Data Visualization

Visualization is a technique for creating diagrams, images or animations to communicate a message
Usage of charts or graphs to visualize huge amounts of complex data is easier than poring over spreadsheets or reports
Data Analysis using Visualization includes:

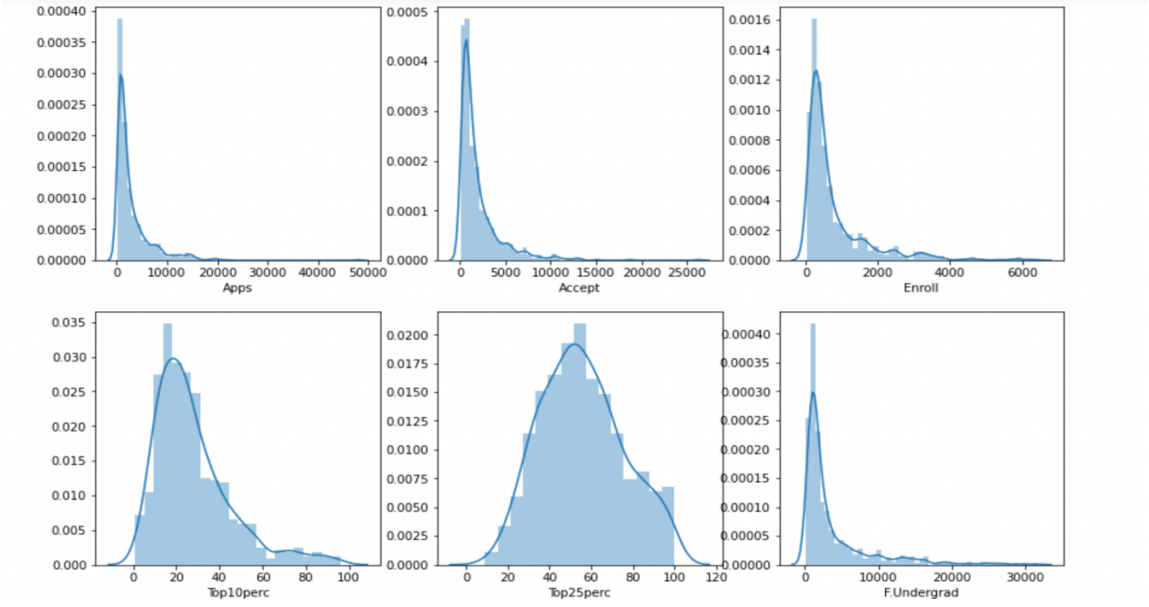
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
  Key for this analysis is generating insights/inferences aligned with the business problem

Q2.1).Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

Univariate Analysis

For Numerical Categories –

For Univariate analysis I have used a distribution Plot as well as check their skewness to understand the data to understand the distribution of the Data.



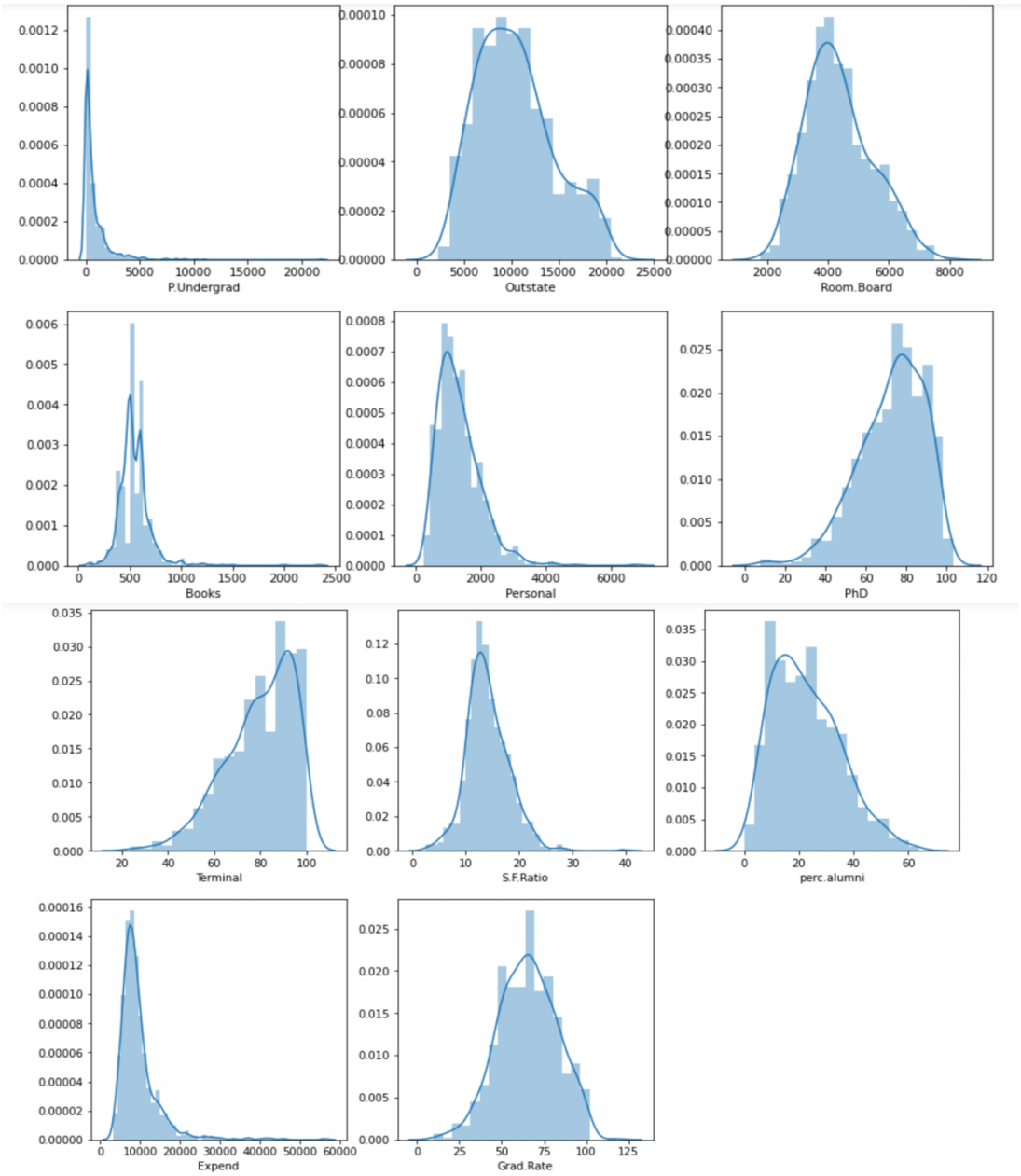


Figure 2 – Distplot for Univariate Analysis

```
P.Undergrad     5.692353
Apps            3.723750
Books           3.485025
Expend          3.459322
Accept          3.417727
Enroll          2.690465
F.Undergrad     2.610458
Personal        1.742497
Top10perc       1.413217
S.F.Ratio       0.667435
perc.alumni     0.606891
Outstate        0.509278
Room.Board      0.477356
Top25perc       0.259340
Grad.Rate      -0.113777
PhD            -0.768170
Terminal       -0.816542
dtype: float64
```

Table 12- Skewness of the data

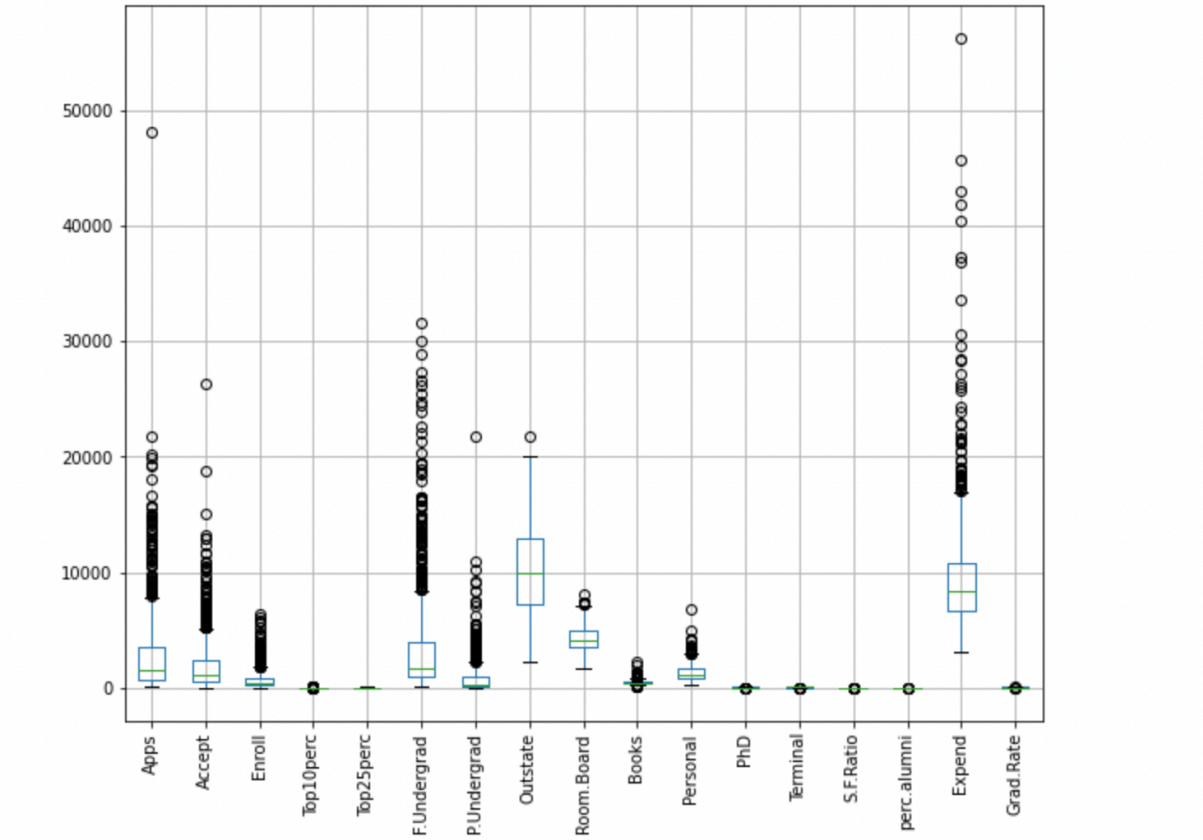


Figure 3- Boxplot Analysis

Inference –

From the Above Skewness data, distribution plots and the description table mentioned in Table 3, This means that most of the colleges are within a range of values corresponding to the individual variables and few colleges are outside this range and nearly symmetrically placed on either side of the range. It is clear that P. Undergrad, Apps, Books, Expend, Accept, Enroll, F. Undergrad, Personal, Top 10 Perc, S.F Ratio, Perc Alumni, Outstate, Room Board as well as Top 25 Perc are Positively or right Skewed as their Mean value is greater than Median. This right skewness may be observed because only a few colleges have higher values for the corresponding variables and most of the other colleges have low values and Their tail falls in the right Side of the distribution plot while Grad.Rate, PHD, Terminal Are negatively skewed or Left Skewed as their mean is less than the Median. This is observed because there are only few colleges which have less percentage of faculties with PhD or Terminal degrees. The Boxplot shows that there are many Outliers in the data.

Multivariate Analysis-

Numerical Data

For Bivariate Analysis, I have used a Correlation matrix as well Scatter plot to understand the relationship between Two variables.

Heat Map or Correlation Matrix-

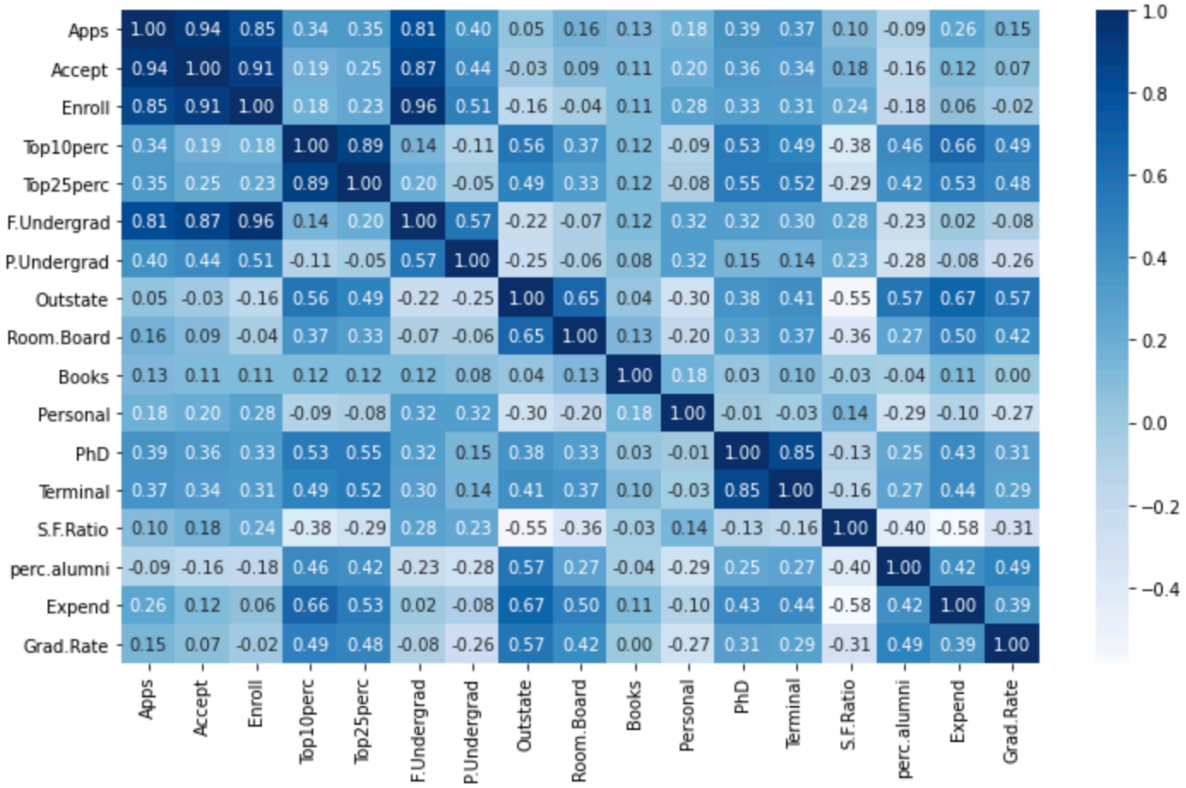
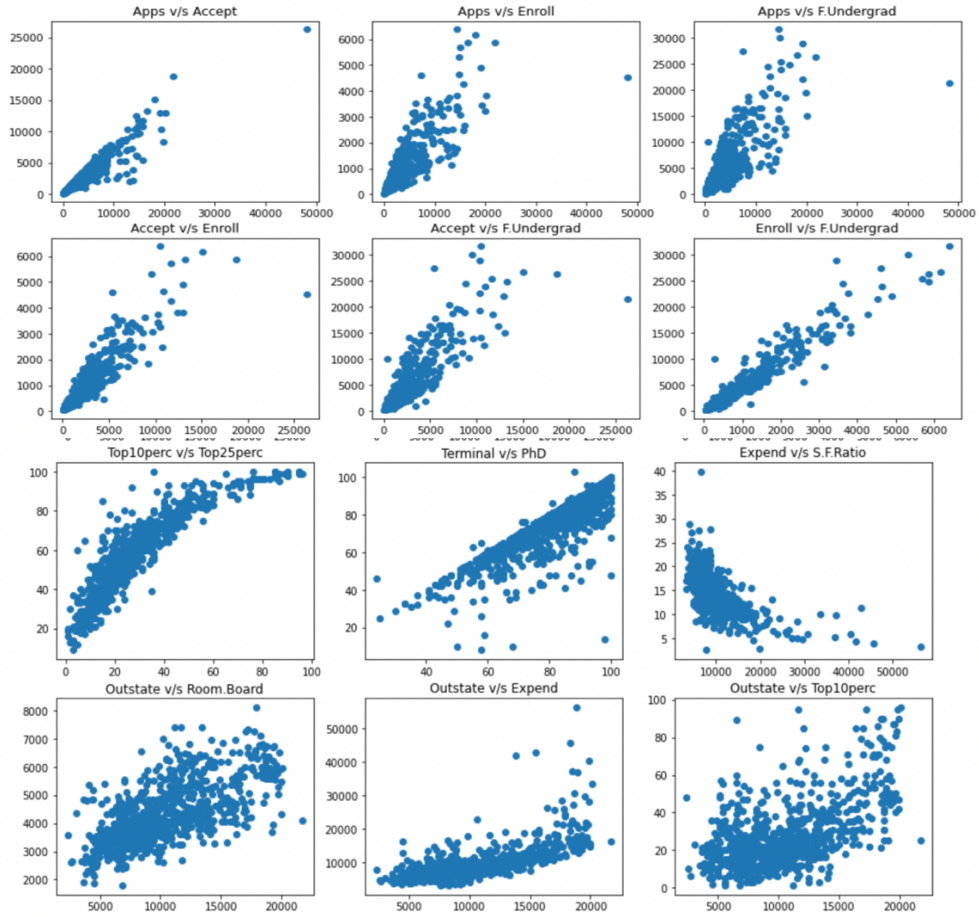


Figure 4 – Correlation Matrix



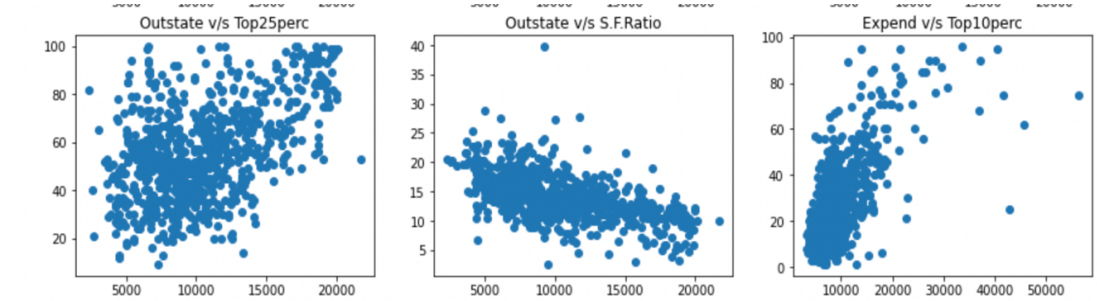
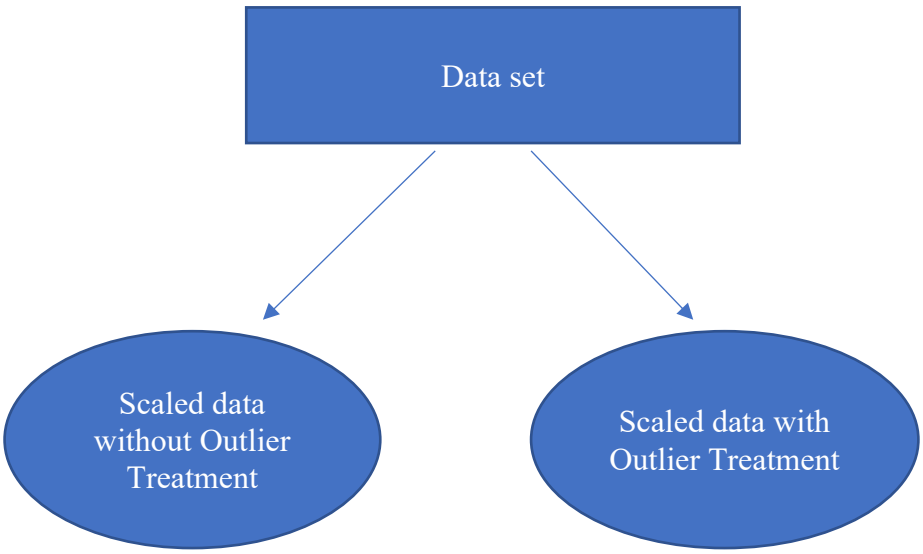


Figure 5- Scatter Plot for Bivariate Analysis

Inferences:

- There is strong positive correlation between 'Apps', 'Accept', 'Enrol' and 'F.Undergrad'. The logic behind the strong positive correlation between 'Apps', 'Accept' and 'Enrol' may be as number of applications increase, this implies more number of acceptance counts and hence more number of enrolments.
- There are outliers in almost all Bivariate analysis mostly in apps vs accept, apps vs enrol, apps vs f.undergrad, accept vs enrol, Accept vs F. undergrad
- There is strong positive correlation between 'Top10perc' and 'Top25perc'. The reason is the students present in the top 10% of higher secondary class are also present in top 25%.
- Also we can observe high positive correlation between 'Terminal' and 'PhD'. This may be because 'Terminal' degree holder is most probably also a PhD holder.
- There is a medium negative correlation observed between 'Expend' and 'S.F.Ratio'; the reason may be because higher Expend ratio means student pays higher instructional expenses but higher student to faculty ratio means more students per faculty. Thus as 'S.F.Ratio' will increase, the expenses shared by the students towards 'Expend' will decrease.
- There is medium positive correlation between 'Outstate' and 'Room.Board' and also between 'Outstate' and 'Expend'. The reason could be higher fees for public universities for out of state students.
- There is medium correlation between 'Outstate' and 'Top10perc', 'Outstate' and 'Top25perc'. The reason could be because the top 10% and top 25% students are distributed throughout the country.
- There is a lot of corelation while PCA will help us to remove the corelations.



Q2.2). Is scaling necessary for PCA in this case? Give justification and perform scaling.

The dataset contains certain variables which are counts like 'Apps' having a mean value of 3001 approx. and variables like 'Expend' which are expressed in currency units having a mean value of 9660 approx. whereas there are certain variables which are ratios and percentages like 'S.F.Ratio' and 'Top25perc' having much lesser magnitude of values. Since we are going to perform PCA, which essentially captures the variance in different directions, if we consider the dataset as it is, it will affect the PCA analysis. With the variables with higher magnitude and hence higher variance dominating the results. Thus to perform a fair and proper PCA analysis it is important to do scaling of the variables. Typically while performing PCA we do mean centering and then scaling by dividing by standard deviation. So we will do z-score scaling.

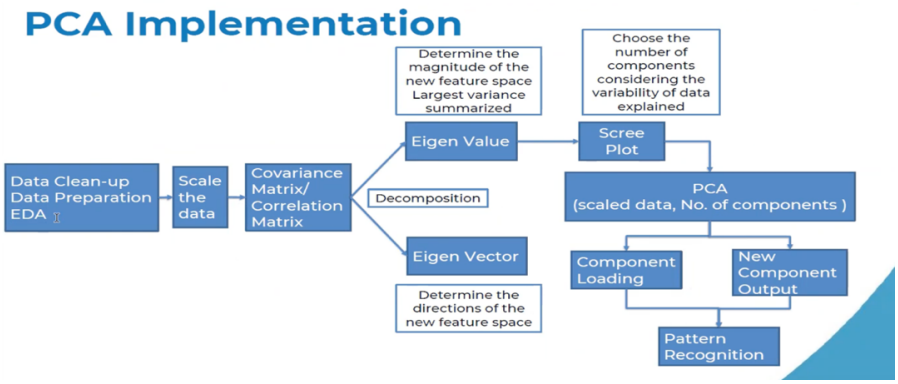
Often the variables of the data set are of different scales i.e. one variable is in millions and other in only 100. Since the data in these variables are of different scales, it is tough to compare these variables. Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data pre-processing 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.We will be doing this only for the numerical variables only.
From the data below, it is clear that

- It may be observed from the description that the mean is nearly 0 and standard deviation is nearly 1, which is the effect of z-score scaling.
- The values of the variables are now comparable and hence will give a better PCA.

	0	1	2	3	4
Apps	-0.346882	-0.210884	-0.406866	-0.668261	-0.726176
Accept	-0.321205	-0.038703	-0.376318	-0.681682	-0.764555
Enroll	-0.063509	-0.288584	-0.478121	-0.692427	-0.780735
Top10perc	-0.258583	-0.655656	-0.315307	1.840231	-0.655656
Top25perc	-0.191827	-1.353911	-0.292878	1.677612	-0.596031
F.Undergrad	-0.168116	-0.209788	-0.549565	-0.658079	-0.711924
P.Undergrad	-0.209207	0.244307	-0.497090	-0.520752	0.009005
Outstate	-0.746356	0.457496	0.201305	0.626633	-0.716508
Room.Board	-0.964905	1.909208	-0.554317	0.996791	-0.216723
Books	-0.602312	1.215880	-0.905344	-0.602312	1.518912
Personal	1.270045	0.235515	-0.259582	-0.688173	0.235515
PhD	-0.163028	-2.675646	-1.204845	1.185206	0.204672
Terminal	-0.115729	-3.378176	-0.931341	1.175657	-0.523535
S.F.Ratio	1.013776	-0.477704	-0.300749	-1.615274	-0.553542
perc.alumni	-0.867574	-0.544572	0.585935	1.151188	-1.675079
Expend	-0.501910	0.166110	-0.177290	1.792851	0.241803
Grad.Rate	-0.318252	-0.551262	-0.667767	-0.376504	-2.939613

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	6.355797e-17	1.000644	-0.755134	-0.575441	-0.373254	0.160912	11.658671
Accept	777.0	6.774575e-17	1.000644	-0.794764	-0.577581	-0.371011	0.165417	9.924816
Enroll	777.0	-5.249269e-17	1.000644	-0.802273	-0.579351	-0.372584	0.131413	6.043678
Top10perc	777.0	-2.753232e-17	1.000644	-1.506526	-0.712380	-0.258583	0.422113	3.882319
Top25perc	777.0	-1.546739e-16	1.000644	-2.364419	-0.747607	-0.090777	0.667104	2.233391
F.Undergrad	777.0	-1.661405e-16	1.000644	-0.734617	-0.558643	-0.411138	0.062941	5.764674
P.Undergrad	777.0	-3.029180e-17	1.000644	-0.561502	-0.499719	-0.330144	0.073418	13.789921
Outstate	777.0	6.515595e-17	1.000644	-2.014878	-0.776203	-0.112095	0.617927	2.800531
Room.Board	777.0	3.570717e-16	1.000644	-2.351778	-0.693917	-0.143730	0.631824	3.436593
Books	777.0	-2.192583e-16	1.000644	-2.747779	-0.481099	-0.299280	0.306784	10.852297
Personal	777.0	4.765243e-17	1.000644	-1.611860	-0.725120	-0.207855	0.531095	8.068387
PhD	777.0	5.954768e-17	1.000644	-3.962596	-0.653295	0.143389	0.756222	1.859323
Terminal	777.0	-4.481615e-16	1.000644	-3.785982	-0.591502	0.156142	0.835818	1.379560
S.F.Ratio	777.0	-2.057556e-17	1.000644	-2.929799	-0.654660	-0.123794	0.609307	6.499390
perc.alumni	777.0	-6.022638e-17	1.000644	-1.836580	-0.786824	-0.140820	0.666685	3.331452
Expend	777.0	1.213101e-16	1.000644	-1.240641	-0.557483	-0.245893	0.224174	8.924721
Grad.Rate	777.0	3.886495e-16	1.000644	-3.230876	-0.726019	-0.026990	0.730293	3.060392

Table 13- Sample(Above) and Description(Below) of Scaled data



PCA Implementation Process

PCA works on only continuous data. As, we have 17 features we have 17 PCA components in the data. After Scaling the data Covariance matrix and correlation matrix are the same but as information is explained in terms of variance we use covariance matrix. We will use these matrix to decompose eigen values and eigen vectors.

Statistical tests to be done before PCA

1st Method

Bartletts Test of Sphericity

Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

- H0: All variables in the data are uncorrelated
- Ha: At least one pair of variables in the data are correlated

If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is small, then we can reject the null hypothesis and agree that there is at least one pair of variables in the data which are correlated hence PCA is recommended.

We get the P-Value as 0 which means we reject the null hypothesis. Therefore there is enough Evidence that there is Correlation in the data

2nd Method

KMO Test

The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

KMO test came out to be 0.813 which is greater than 0.7 therefore PCA is Important for this data set.

Q2.3). Comment on the comparison between the covariance and the correlation matrices from this data.

On a scaled data covariance matrix = Correlation Matrix
Diagonal variables have a 1-unit variable in Correlation matrix

Covariance matrix gives a relationship about the **direction** of the dataset. By direction, it means that if the variables are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable).

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.001289	0.944666	0.847913	0.339270	0.352093	0.815540	0.398777	0.050224	0.165152	0.132729	0.178961	0.391201	0.369968	0.095756	-0.090342	0.259927	0.146944
Accept	0.944666	1.001289	0.912811	0.192695	0.247795	0.875350	0.441839	-0.025788	0.091016	0.113672	0.201248	0.356216	0.338018	0.176456	-0.160196	0.124878	0.067399
Enroll	0.847913	0.912811	1.001289	0.181527	0.227037	0.965883	0.513730	-0.155678	-0.040284	0.112856	0.281291	0.331896	0.308671	0.237577	-0.181027	0.064252	-0.022370
Top10perc	0.339270	0.192695	0.181527	1.001289	0.893144	0.141471	-0.105492	0.563055	0.371959	0.119012	-0.093437	0.532513	0.491768	-0.385370	0.456072	0.661765	0.495627
Top25perc	0.352093	0.247795	0.227037	0.893144	1.001289	0.199702	-0.053646	0.490024	0.331917	0.115676	-0.080914	0.546566	0.525425	-0.295009	0.418403	0.528127	0.477896
F.Undergrad	0.815540	0.875350	0.965883	0.141471	0.199702	1.001289	0.571247	-0.216020	-0.068979	0.115699	0.317608	0.318747	0.300406	0.280064	-0.229758	0.018676	-0.078875
P.Undergrad	0.398777	0.441839	0.513730	-0.105492	-0.053646	0.571247	1.001289	-0.253839	-0.061405	0.081304	0.320294	0.149306	0.142086	0.232830	-0.281154	-0.083676	-0.257332
Outstate	0.050224	-0.025788	-0.155678	0.563055	0.490024	-0.216020	-0.253839	1.001289	0.655100	0.038905	-0.299472	0.383476	0.408509	-0.555536	0.566992	0.673646	0.572026
Room.Board	0.165152	0.091016	-0.040284	0.371959	0.331917	-0.068979	-0.061405	0.655100	1.001289	0.128128	-0.199685	0.329627	0.375022	-0.363095	0.272714	0.502386	0.425489
Books	0.132729	0.113672	0.112856	0.119012	0.115676	0.115699	0.081304	0.038905	0.128128	1.001289	0.179526	0.026940	0.100084	-0.031970	-0.040260	0.112554	0.001062
Personal	0.178961	0.201248	0.281291	-0.093437	-0.080914	0.317608	0.320294	-0.299472	-0.199685	0.179526	1.001289	-0.010950	-0.030653	0.136521	-0.286337	-0.098018	-0.269691
PhD	0.391201	0.356216	0.331896	0.532513	0.546566	0.318747	0.149306	0.383476	0.329627	0.026940	-0.010950	1.001289	0.850682	-0.130698	0.249330	0.433319	0.305431
Terminal	0.369968	0.338018	0.308671	0.491768	0.525425	0.300406	0.142086	0.408509	0.375022	0.100084	-0.030653	0.850682	1.001289	-0.160310	0.267475	0.439365	0.289900
S.F.Ratio	0.095756	0.176456	0.237577	-0.385370	-0.295009	0.280064	0.232830	-0.555536	-0.363095	-0.031970	0.136521	-0.130698	-0.160310	1.001289	-0.403448	-0.584584	-0.307106
perc.alumni	-0.090342	-0.160196	-0.181027	0.456072	0.418403	-0.229758	-0.281154	0.566992	0.272714	-0.040260	-0.286337	0.249330	0.267475	-0.403448	1.001289	0.418250	0.491530
Expend	0.259927	0.124878	0.064252	0.661765	0.528127	0.018676	-0.083676	0.673646	0.502386	0.112554	-0.098018	0.433319	0.439365	-0.584584	0.418250	1.001289	0.390846
Grad.Rate	0.146944	0.067399	-0.022370	0.495627	0.477896	-0.078875	-0.257332	0.572026	0.425489	0.001062	-0.269691	0.305431	0.289900	-0.307106	0.491530	0.390846	1.001289

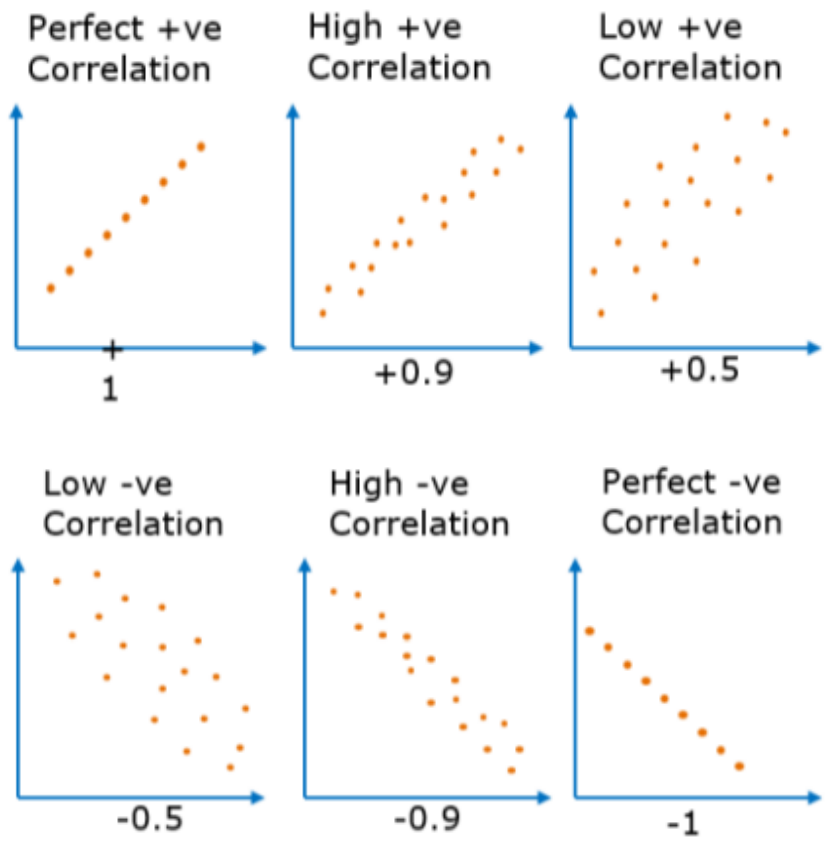
Table 14- Covariance Matrix

Inference-

- Once we standardize the correlation matrix, we get covariance matrix
- Negatives indicate that they tend to be high together or low together
- Variances become 1 and covariance becomes correlations
- Taking into Consideration, Figure 4 – Correlation Matrix. The correlation coefficient is a dimensionless metric and its value ranges from -1 to +1. The closer it is to +1 or -1, the more closely the two variables are related.
- When the correlation coefficient is positive, an increase in one variable also increases the other. When the correlation coefficient is negative, the changes in the two variables are in opposite directions.

If there is no relationship at all between two variables, then the correlation coefficient will certainly be 0. However, if it is 0 then we can only say that there is no linear relationship. There could exist other functional relationships between the variables.

Covariance and correlation are related to each other, in the sense that covariance determines the type of interaction between two variables, while correlation determines the direction as well as the strength of the relationship between two variables.



Q2.4).
Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]

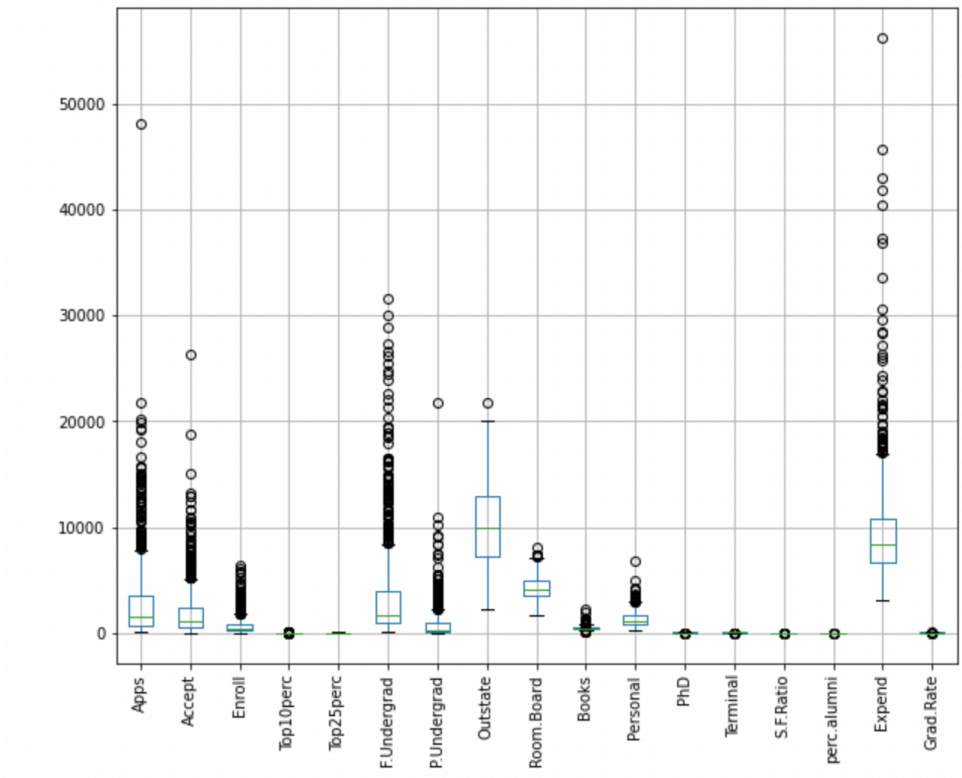


Figure 3(Repeat) - Boxplot without scaling and outlier treatment

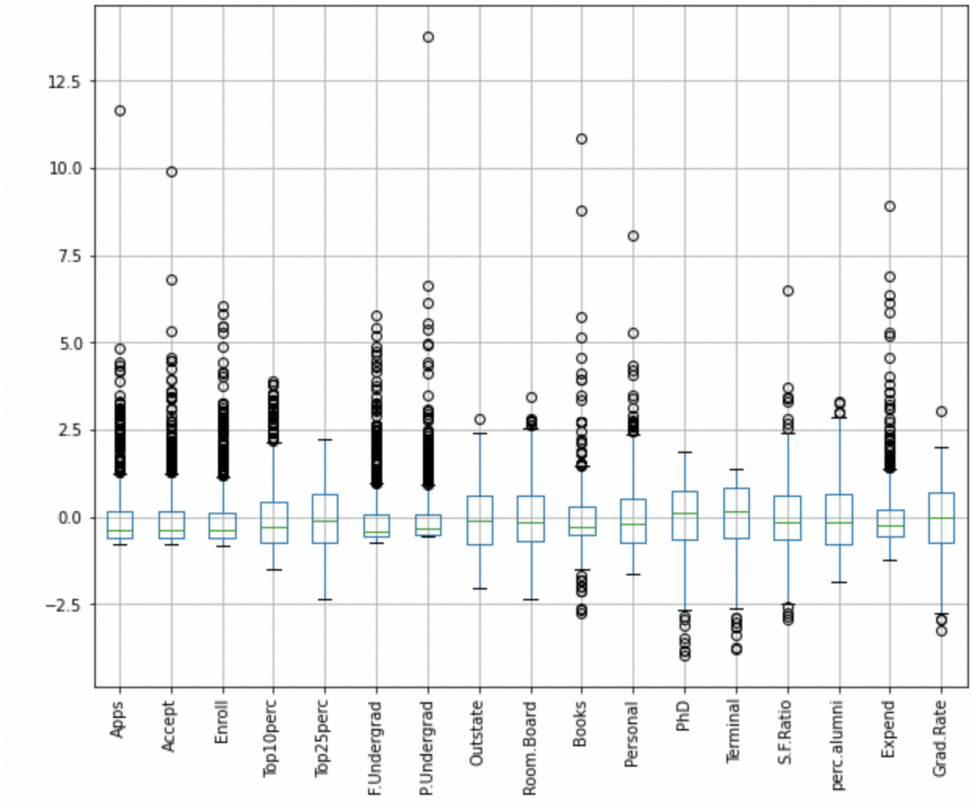


Figure 6- Boxplot with scaling and without outlier treatment

Additional –

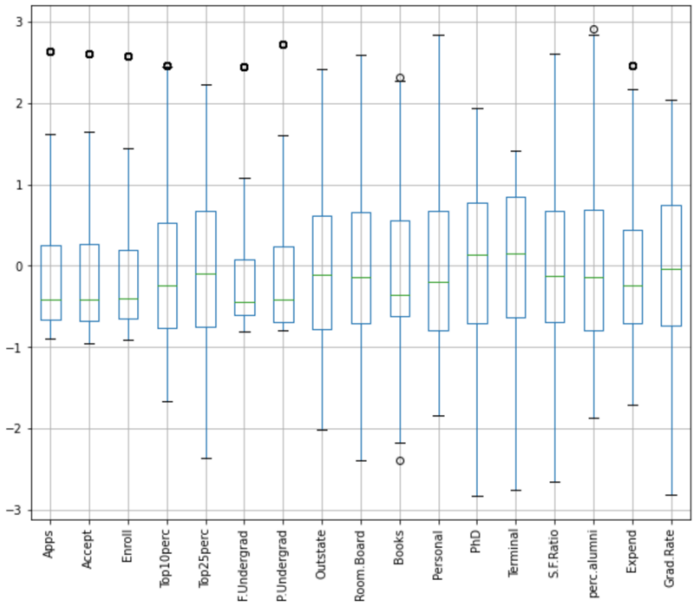


Figure 7-Boxplot with scaling and outlier treatment (Not required) [Treatment of outliers is shown in Jupyter]

Inferences

- Shifts distribution's mean to 0 & unit variance
- There is no predetermined range
- Best to use on data that is approximately normally distributed

Q2.5). Extract the eigenvalues and eigenvectors.[print both]

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.

```
Eigen Values
%s [5.45052162 4.48360686 1.17466761 1.00820573 0.93423123 0.84849117
0.6057878  0.58787222 0.53061262 0.4043029  0.02302787 0.03672545
0.31344588 0.08802464 0.1439785  0.16779415 0.22061096]
```

Table 15- Eigen Values

```
Eigen Vectors
 %s [[-2.48765602e-01  3.31598227e-01  6.30921033e-02 -2.81310530e-01
    5.74140964e-03  1.62374420e-02  4.24863486e-02  1.03090398e-01
    9.02270802e-02 -5.25098025e-02  3.58970400e-01 -4.59139498e-01
    4.30462074e-02 -1.33405806e-01  8.06328039e-02 -5.95830975e-01
    2.40709086e-02]
  [-2.07601502e-01  3.72116750e-01  1.01249056e-01 -2.67817346e-01
    5.57860920e-02 -7.53468452e-03  1.29497196e-02  5.62709623e-02
    1.77864814e-01 -4.11400844e-02 -5.43427250e-01  5.18568789e-01
   -5.84055850e-02  1.45497511e-01  3.34674281e-02 -2.92642398e-01
   -1.45102446e-01]
  [-1.76303592e-01  4.03724252e-01  8.29855709e-02 -1.61826771e-01
   -5.56936353e-02  4.25579803e-02  2.76928937e-02 -5.86623552e-02
    1.28560713e-01 -3.44879147e-02  6.09651110e-01  4.04318439e-01
   -6.93988831e-02 -2.95896092e-02 -8.56967180e-02  4.44638207e-01
    1.11431545e-02]
  [-3.54273947e-01 -8.24118211e-02 -3.50555339e-02  5.15472524e-02
   -3.95434345e-01  5.26927980e-02  1.61332069e-01  1.22678028e-01
   -3.41099863e-01 -6.40257785e-02 -1.44986329e-01  1.48738723e-01
   -8.10481404e-03 -6.97722522e-01 -1.07828189e-01 -1.02303616e-03
    3.85543001e-02]
  [-3.44001279e-01 -4.47786551e-02  2.41479376e-02  1.09766541e-01
   -4.26533594e-01 -3.30915896e-02  1.18485556e-01  1.02491967e-01
   -4.03711989e-01 -1.45492289e-02  8.03478445e-02 -5.18683400e-02
   -2.73128469e-01  6.17274818e-01  1.51742110e-01 -2.18838802e-02
   -8.93515563e-02]
 [-1.54640962e-01  4.17673774e-01  6.13929764e-02 -1.00412335e-01
   -4.34543659e-02  4.34542349e-02  2.50763629e-02 -7.88896442e-02
    5.94419181e-02 -2.08471834e-02 -4.14705279e-01 -5.60363054e-01
   -8.11578181e-02 -9.91640992e-03 -5.63728817e-02  5.23622267e-01
    5.61767721e-02]
  [-2.64425045e-02  3.15087830e-01 -1.39681716e-01  1.58558487e-01
    3.02385408e-01  1.91198583e-01 -6.10423460e-02 -5.70783816e-01
   -5.60672902e-01  2.23105808e-01  9.01788964e-03  5.27313042e-02
    1.00693324e-01 -2.09515982e-02  1.92857500e-02 -1.25997650e-01
   -6.35360730e-02]
  [-2.94736419e-01 -2.49643522e-01 -4.65988731e-02 -1.31291364e-01
    2.22532003e-01  3.00003910e-02 -1.08528966e-01 -9.84599754e-03
    4.57332880e-03 -1.86675363e-01  5.08995918e-02 -1.01594830e-01
    1.43220673e-01 -3.83544794e-02 -3.40115407e-02  1.41856014e-01
   -8.23443779e-01]
  [-2.49030449e-01 -1.37808883e-01 -1.48967389e-01 -1.84995991e-01
    5.60919470e-01 -1.62755446e-01 -2.09744235e-01  2.21453442e-01
   -2.75022548e-01 -2.98324237e-01  1.14639620e-03  2.59293381e-02
   -3.59321731e-01 -3.40197083e-03 -5.84289756e-02  6.97485854e-02
    3.54559731e-01]
  [-6.47575181e-02  5.63418434e-02 -6.77411649e-01 -8.70892205e-02
   -1.27288825e-01 -6.41054950e-01  1.49692034e-01 -2.13293009e-01
    1.33663353e-01  8.20292186e-02  7.72631963e-04 -2.88282896e-03
    3.19400370e-02  9.43887925e-03 -6.68494643e-02 -1.14379958e-02
   -2.81593679e-02]
  [ 4.25285386e-02  2.19929218e-01 -4.99721120e-01  2.30710568e-01
   -2.22311021e-01  3.31398003e-01 -6.33790064e-01  2.32660840e-01
    9.44688900e-02 -1.36027616e-01 -1.11433396e-03  1.28904022e-02
   -1.85784733e-02  3.09001353e-03  2.75286207e-02 -3.94547417e-02
   -3.92640266e-02]
 [-3.18312875e-01  5.83113174e-02  1.27028371e-01  5.34724832e-01
   1.40166326e-01 -9.12555212e-02  1.09641298e-03  7.70400002e-02
   1.85181525e-01  1.23452200e-01  1.38133366e-02 -2.98075465e-02
   4.03723253e-02  1.12055599e-01 -6.91126145e-01 -1.27696382e-01
   2.32224316e-02]
 [-3.17056016e-01  4.64294477e-02  6.60375454e-02  5.19443019e-01
   2.04719730e-01 -1.54927646e-01  2.84770105e-02  1.21613297e-02
   2.54938198e-01  8.85784627e-02  6.20932749e-03  2.70759809e-02
  -5.89734026e-02 -1.58909651e-01  6.71008607e-01  5.83134662e-02
   1.64850420e-02]
 [ 1.76957895e-01  2.46665277e-01  2.89848401e-01  1.61189487e-01
  -7.93882496e-02 -4.87045875e-01 -2.19259358e-01  8.36048735e-02
  -2.74544380e-01 -4.72045249e-01 -2.22215182e-03  2.12476294e-02
   4.45000727e-01  2.08991284e-02  4.13740967e-02  1.77152700e-02
  -1.10262122e-02]
 [-2.05082369e-01 -2.46595274e-01  1.46989274e-01 -1.73142230e-02
  -2.16297411e-01  4.73400144e-02 -2.43321156e-01 -6.78523654e-01
   2.55334907e-01 -4.22999706e-01 -1.91869743e-02 -3.33406243e-03
  -1.30727978e-01  8.41789410e-03 -2.71542091e-02 -1.04088088e-01
   1.82660654e-01]
 [-3.18908750e-01 -1.31689865e-01 -2.26743985e-01 -7.92734946e-02
   7.59581203e-02  2.98118619e-01  2.26584481e-01  5.41593771e-02
   4.91388809e-02 -1.32286331e-01 -3.53098218e-02  4.38803230e-02
   6.92088870e-01  2.27742017e-01  7.31225166e-02  9.37464497e-02
   3.25982295e-01]
 [-2.52315654e-01 -1.69240532e-01  2.08064649e-01 -2.69129066e-01
  -1.09267913e-01 -2.16163313e-01 -5.59943937e-01  5.33553891e-03
  -4.19043052e-02  5.90271067e-01 -1.30710024e-02  5.00844705e-03
   2.19839000e-01  3.39433604e-03  3.64767385e-02  6.91969778e-02
   1.22106697e-01]]
```

Table 16 – Eigen Vectors

Each eigenvector direction is orthogonal to the other eigenvectors. The corresponding coefficients of a particular eigenvector are the loadings corresponding to each of the variables of the original dataset, if the eigenvectors are calculated for a covariance matrix of a standard scaled (z-scaled) data then these coefficients may be considered to be the correlations with the variables of the original dataset.

Q2.6).Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features.

We will use all the 17 components to perform PCA.

```
array([[-1.59285540e+00, -2.19240180e+00, -1.43096371e+00, ...,
        -7.32560596e-01,  7.91932735e+00, -4.69508066e-01],
       [ 7.67333510e-01, -5.78829984e-01, -1.09281889e+00, ...,
        -7.72352397e-02, -2.06832886e+00,  3.66660943e-01],
       [-1.01073537e-01,  2.27879812e+00, -4.38092811e-01, ...,
        -4.05641899e-04,  2.07356368e+00, -1.32891515e+00],
       ...,
       [ 1.75239502e-03,  1.03709803e-01, -2.25582869e-02, ...,
         6.79013123e-02,  3.53597440e-01, -1.14873492e-01],
       [-9.31400698e-02, -5.02556890e-02, -4.05268301e-03, ...,
        -2.32023970e-01,  3.04416200e-01, -1.17076127e-01],
       [ 9.35522023e-02, -1.74057054e-01,  3.75875882e-03, ...,
        -9.99380421e-02,  3.35104811e-01, -2.57218339e-03]])
```

Table 17- Sample of dataset after performing PCA

```
**Eigen Vector corresponding to maximum Eigen value :**
[ 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]
```

Table 18- Loading Principal Components

PCA Components are same as eigen Vectors (take reference from Table 16)

	0	1	2	3	4	5	6	7	8	9
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.103090	-0.090227	0.052510
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950	-0.056271	-0.177865	0.041140
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693	0.058662	-0.128561	0.034488
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678	0.341100	0.064026
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486	-0.102492	0.403712	0.014549
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076	0.078890	-0.059442	0.020847
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042	0.570784	0.560673	-0.223106
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529	0.009846	-0.004573	0.186675
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744	-0.221453	0.275023	0.298324
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692	0.213293	-0.133663	-0.082029
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790	-0.232661	-0.094469	0.136028
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096	-0.077040	-0.185182	-0.123452
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477	-0.012161	-0.254938	-0.088578
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	0.219259	-0.083605	0.274544	0.472045
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321	0.678524	-0.255335	0.423000
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584	-0.054159	-0.049139	0.132286
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944	-0.005336	0.041904	-0.590271

Table 19- Loading all Principal Components with Original features

Q2.7). Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).

Linear Combination of variables
C= W1(Y1) + W2(Y2) + W3(Y3) + W4(Y4)
C – Components
W1,2,3,4 – PCA components loading
Y1,2,3,4 – Features

```
The Linear eq of 1st component:
0.2 * Apps + 0.2 * Accept + 0.2 * Enroll + 0.4 * Top10perc + 0.3 * Top25perc + 0.2 * F.Undergrad + 0.0 * P.Undergrad
+ 0.3 * Outstate + 0.2 * Room.Board + 0.1 * Books + -0.0 * Personal + 0.3 * PhD + 0.3 * Terminal + -0.2 * S.F.Ratio +
0.2 * perc.alumni + 0.3 * Expend + 0.3 * Grad.Rate +
```

Table 20 – Explicit for of the first PC

To gather the explicit form in graphical manner we take 6 Components to gather 83% of the data.

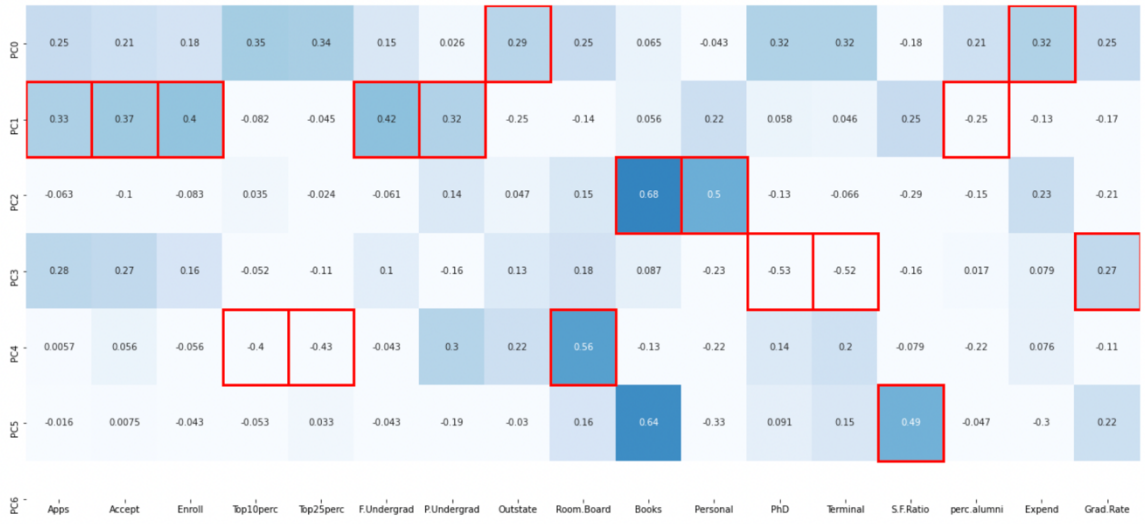


Figure 8 – Heatmap to show Explicit Form

These are the highest magnitude for each column. For E.g., Apps is loaded on PC1. Now As we can see Apps, Accept, Enroll, F.Undergrad, P.Undergrad and Perc. Alumni are maximum loaded on PC1 and are talking about similar business structure. Now we can club them and introduce a new dataset With PC1 or their relation to interpret further results.

Q2.8). 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?

```
Cumulative Variance Explained [ 33.15185743  61.52550945  67.98957029  73.84487717  79.11892366
  83.61602283  87.06508197  90.32266981  92.9263317   95.17182864
  96.61489423  97.47757639  98.27677262  99.00385952  99.44252192
  99.77139178 100.        ]
```

Table 21- Cumulative variance of the Eigen Values

According to eigen values the first and second captures more than half of the variability
They are orthogonal or independent of each other

Total of all the values is 17
And 80% of the information lies in the first 6 eigen values

- The above values are the eigen values of the covariance matrix which show the variance captured by each principal component in decreasing order.

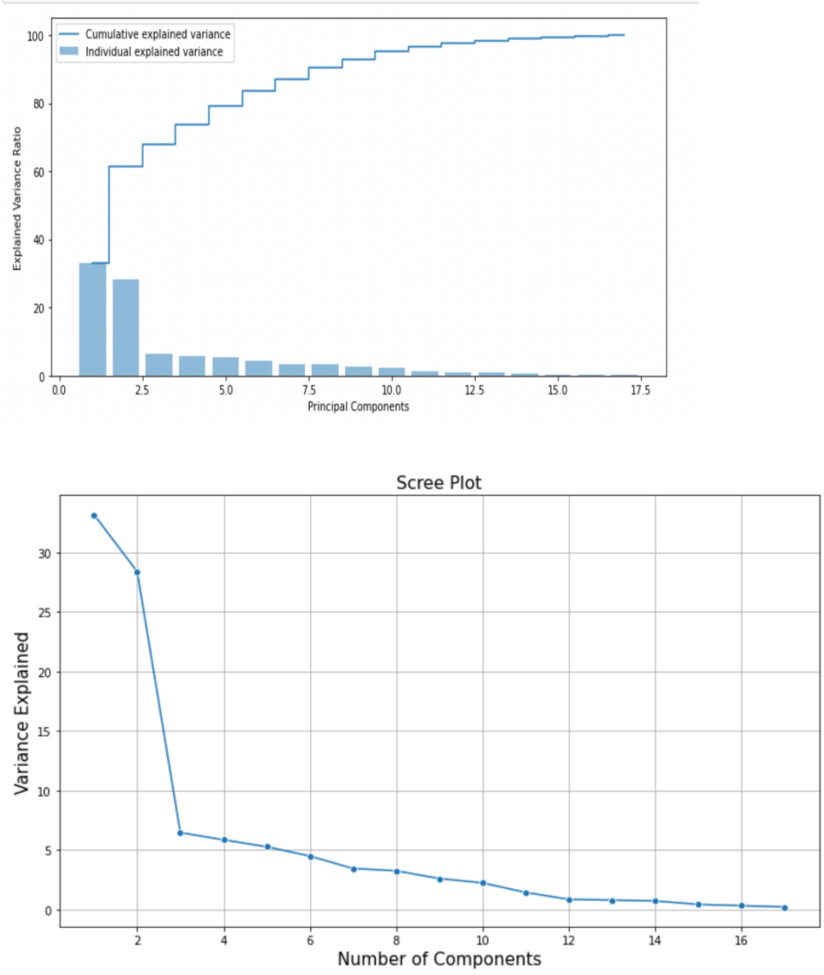


Figure 9- Scree Plot

- Shows the individual explained variances by the principal components.
- We can observe a sudden decrease in slope from the third principal component onwards. This means the maximum variances are captured by the first two principal components. This point is also called inflection point
- The first two principal components capture approximately 62% of the total variance.
- There is sudden drop in variance captured after second principal component.
- 90% of the total variance is captured by first 8 principal components.
- After the first 11 principal components there is less than 1% increase in variance captured consecutively by the remaining principal components.

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]

- I have selected 6 out of the 17 new dimensions now to explain 83.75% of the variance and reduce the dimensionality of the dataset accordingly.
- The new dimension variables are independent of each other, which also helps in certain algorithms.
- The dimensionality reduction as obtained from PCA helps in lesser computing power, i.e. faster processing for further analysis.
- The dimensionality reduction also helps in lesser storage space.The dimensionality reduction also helps in addressing the overfitting issue, which mainly occurs when there are too many variables.
- In our case study, after performing multivariate analysis we have observed that many of the variables are correlated. Thus we don't need all these variables for analysis but we are not sure which variables to drop and which to select, hence we perform PCA, which captures the information (in the form of variance) from all these variables into new dimension variables. Now based on the requirement of information we can select the number of new dimension variables required.
- Range of the values is very high. Therefore it is important to scale the data.