PARAKANDLA. MANISHA

PGP-DSBA ONLINE - GREATLEARNING

PROJECT: ADVANCED STATISTICS

DATE: 17-10-2021

==========

ADVANCED STATISTICS

BUSINESS REPORT

==========

Contents

List of tables:

PROBLEM1:

PROBLEM2:

List of Figures:

Problem1 & Problem2:

Problem1: (Anova)

EXECUTIVE SUMMARY:

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.

INTRODUCTION:

In this problem, we perform **ANOVA** to find the mean salary across all the three categories of Educational Qualification and four categories of occupation. The problem statement here was divided into two parts.

Problem 1A:

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

INTRODUCTION TO ANOVA (Analysis of Variance):

ANOVA is a technique which belongs to the domain called **"Experimental Designs"**. ANOVA helps in establishing an exact way and the Cause- Effect relation between variables. From the statistical inference point of view, ANOVA is an extension of independent t test for testing the equality of two population means. When more than two population means have to be compared, ANOVA technique is used. In ANOVA, the null hypothesis is defined as **Ho** and the Alternate Hypothesis is defined as **H1**.

H0: μ 1 = μ 2 =μ 3 =μ 4 =……=μ k for testing the equality of population means for k populations where μ denotes the mean of the population.

H1: At least one population mean out of K populations are different.

For the given problem, an analysis of salary data has been performed and the results and business insights drawn are listed.

SAMPLE OF A DATASET:

	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

1.1 Table of sample dataset

The dataset consists of 40 observations with three categories of Educational Qualification (Highschool Graduate, Bachelor, Doctorate) and four categories of Occupation (Prof-Specialty, Sales, Adm-clerical, Exec-Managerial) and their respective salaries.

EXPLORATORY DATA ANALYSIS:

The dataset has 40 rows and 3 columns. The given dataset has no missing values & the datatypes of dataset are represented using info function.

```
<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
```

Describing the data:

	Education	Occupation	Salary
count	40	40	40.000000
unique	3	4	NaN
top	Doctorate	Prof-specialty	NaN
freq	16	13	NaN
mean	NaN	NaN	162186.875000
std	NaN	NaN	64860.407506
min	NaN	NaN	50103.000000
25%	NaN	NaN	99897.500000
50%	NaN	NaN	169100.000000
75%	NaN	NaN	214440.750000
max	NaN	NaN	260151.000000

1.2 Table showing Description of data

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

Null and Alternate Hypothesis for one way ANOVA (EDUCATION):

→Null Hypothesis $H0$: The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, HS-Grad).

→Alternate Hypothesis $H1$: The mean salary is different in at least one category of education.

Null and Alternate Hypothesis for one way ANOVA(Occupation):

→Null Hypothesis $H0$: The mean salary is the same across all the 4 categories of occupation (Prof-Specialty, Sales, Adm-clerical, Exec-Managerial).

→Alternate Hypothesis $H1$: The mean salary is different in at least one category of occupation.

1.2 Perform one-way ANOVA for Education with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

Null Hypothesis $H0$: The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, HS-Grad).

Alternate Hypothesis $H1$: The mean salary is different in at least one category of education.

One way ANOVA for Education:

	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

1.3 Table for One way ANOVA (Education)

Since the p value = 1.257709e-08 is less than the significance level, **we reject the null hypothesis** and conclude that there is a significant difference in the mean salaries for at least one category of education.

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.

Null Hypothesis $H0$: The mean salary is the same across all the 4 categories of occupation (Prof-Specialty, Sales, Adm-clerical, Exec-Managerial).

Alternate Hypothesis $H1$: The mean salary is different in at least one category of occupation.

One way ANOVA for Occupation:

	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

1.4 Table for One way ANOVA Occupation

Since the p value = 0.458508 is greater than the significance level, **we fail to reject the null hypothesis** (i.e., **we accept Null Hypothesis H0**) and conclude that there is no significant difference in the mean salaries across the 4 categories of occupation.

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.

To find out which class means are significantly different, the Tukey Honest Significant Difference test is performed. Using, the Tukey Honest Significant Difference test,

we get the following table for the category Education:

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

1.5 Table for Tukey HSD Significance Difference (Education)

This table shows that the p_value (p-adj) is less than the significance level (alpha = 0.05) for all the three categories in Education.

Hence, the mean salary across all 3 Categories of Education is different.

We can visualize the mean difference in Salary for category Education Using **Pointplot.**

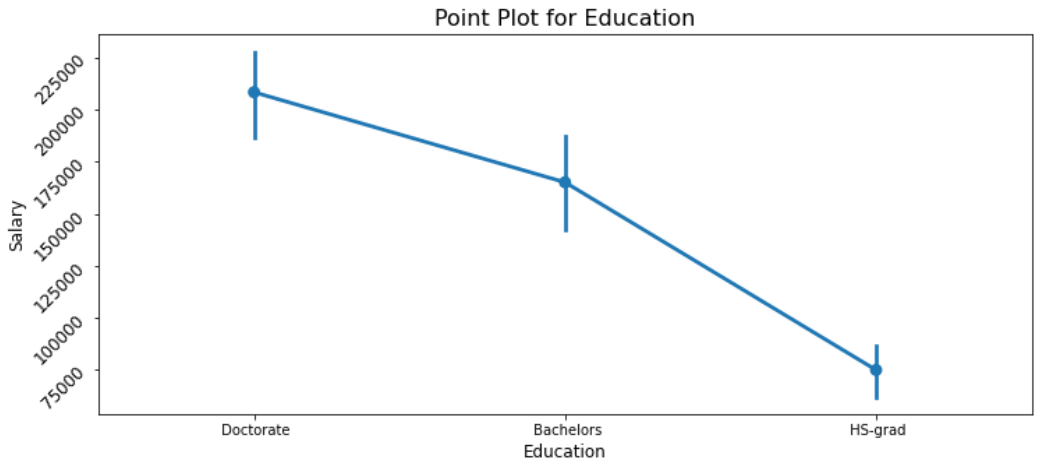

Fig1. Point Plot for Education

Using, the Tukey Honest Significant Difference test, we get the following table for the category Education:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Adm-clerical	Exec-managerial	55693.3	0.4146	-40415.1459	151801.7459	False
Adm-clerical	Prof-specialty	27528.8538	0.7252	-46277.4011	101335.1088	False
Adm-clerical	Sales	16180.1167	0.9	-58951.3115	91311.5449	False
Exec-managerial	Prof-specialty	-28164.4462	0.8263	-120502.4542	64173.5618	False
Exec-managerial	Sales	-39513.1833	0.6507	-132913.8041	53887.4374	False
Prof-specialty	Sales	-11348.7372	0.9	-81592.6398	58895.1655	False

1.6 Table for Tukey HSD Significance Difference (Occupation)

This table shows that the p_value(p-adj) is greater than the significance level (alpha = 0.05) for all the four categories in Occupation.

Hence, Salary across all 4 categories of Occupation is same.

We can visualize the mean difference in Salary for category Occupation Using **Pointplot.**

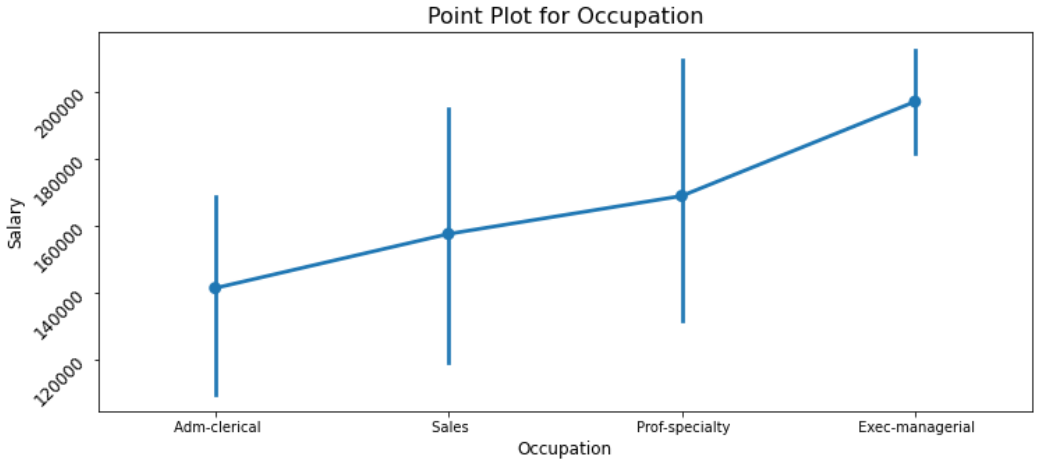

Fig2. Point Plot for Occupation

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.

The interaction plot below shows that, there is a significant amount of interaction between the two categorical variables, Education and Occupation.

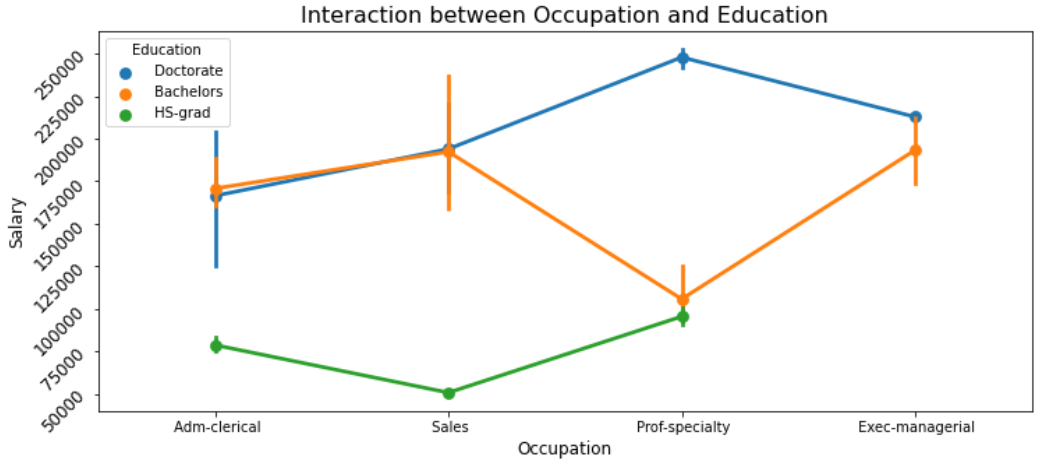

Fig3. Interaction plot for Education-Occupation

Analyzing the effects of one variable on the other from the Interaction plot:

→People with HS-grad education do not reach the position of Exec-managerial and they hold only Adm-clerk, Sales and Prof-Specialty occupations.

→People with education as Bachelors or Doctorate and occupation as Adm-clerical and Sales almost earn the same salaries (salaries ranging from 170000–190000).

→People with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupations as Adm-clerical and Sales.

→ People with education as Bachelors and occupation Sales earn higher than people with education as Bachelors and occupation Prof-Specialty whereas people with education as Doctorate and occupation Sales earn lesser than people with Doctorate and occupation Prof-Specialty. We see a reversal in this part of the plot.

→Similarly, people with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupation Exec-Managerial whereas people with education as Doctorate and occupation as Prof-Specialty earn higher than people with education as Doctorate and occupation Exec-Managerial. There is a reversal in this part of the plot too.

→ Salespeople with Bachelors or Doctorate education earn the same salaries and earn higher than people with education as HS-grad.

→ Adm clerical people with education as HS-grad earn the lowest salaries when compared to people with education as Bachelors or Doctorate.

→ Prof-Specialty people with education as Doctorate earn maximum salaries and people with education as HS-Grad earn the minimum.

→ People with education as HS -Grad earn the minimum salaries.

→ There are no people with education as HS -grad who hold Exec-managerial occupation.

→ People with education as Bachelors and occupation, Sales and Exec-Managerial earn the same salaries.

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?

$H0$: The effect of the independent variable 'education' on the mean 'salary' does not depend on the effect of the other independent variable 'occupation' (i.e., there is no interaction effect between the 2 independent variables, education and occupation).

$H1$: There is an interaction effect between the independent variable 'education' and the independent variable 'occupation' on the mean salary.

Two-way ANOVA based on Education & Occupation:

	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

1.7 Table for Two-way ANOVA based on Education & Occupation

From two-way ANOVA, we can see that there is a significant amount of interaction between the variables, Education and Occupation.

Since, **p value = 2.232500e-05** is **less than** the significance level (**alpha = 0.05**), we reject the null hypothesis.

Hence. We can now see that there is an interaction effect between variables education and occupation on the mean salary.

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

→From the Analysis of Variance ANOVA method, we can observe that combination of Education-Occupation results in higher salaries.

→It is clear that, People with Educational Qualification as Doctorate draw the maximum salaries.

→People with Educational Qualification High School grad (HS Grad) draw the minimum (least) salaries.

→Salary is dependent on Educational Qualifications and Occupation.

Problem2: (Principal Component Analysis)

EXECUTIVE SUMMARY:

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.

DATA DICTIONARY:

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

INTRODUCTION (PCA):

The dataset contains the names of various Universities and Colleges with the number of Applications received, Accepted, Enrolled, Percentage of new students from top 10% of higher secondary class, Percentage of new students from top 25% of higher secondary class, Number of fulltime undergraduates, Number of Parttime Undergraduate students, Number of students for whom the particular college is out of state tuition, cost of room and board, estimated book costs for a student, estimated personal spending for a student, percentage of faculties with PHD, percentage of faculties with terminal degree, student/faculty ratio, percentage of alumni who donate, The instructional expenditure per student, Graduation Rate.

SAMPLE OF A DATASET:

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

2.1 Table for sample of a dataset

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

EXPLORATORY DATA ANALYSIS:

The dataset consists of 777 rows and 18 columns.

There are no duplicate values in the dataset.

The dataset does not have missing values.

```
<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
```

DATA DESCRIPTION:

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.380952	165.105360	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.171171	5221.768440	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.463320	17.177710	10.0	53.0	65.0	78.0	118.0

2.2 Table for description of dataset

UNIVARIATE ANALYSIS:

Univariate analysis is perhaps the simplest form of Statistical Analysis. Like other forms of statistics, it can be inferential or descriptive. The key factor for Univariate Analysis is that only one variable is involved.

It helps us to understand the distribution of data in the dataset. With univariate analysis, we can find patterns and we can summarize the data for

Boxplot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending from the boxes (*whiskers*) indicating variability outside the upper and lower quartiles, hence the terms **box-and-whisker plot** and **box-and-whisker diagram**. Outliers may be plotted as individual points. The spacings between the different parts of the box indicate the degree of spread and skewness in the data, and show Outliers. Box plots can be drawn either horizontally or vertically.

Distplot plots a univariate distribution of observations.

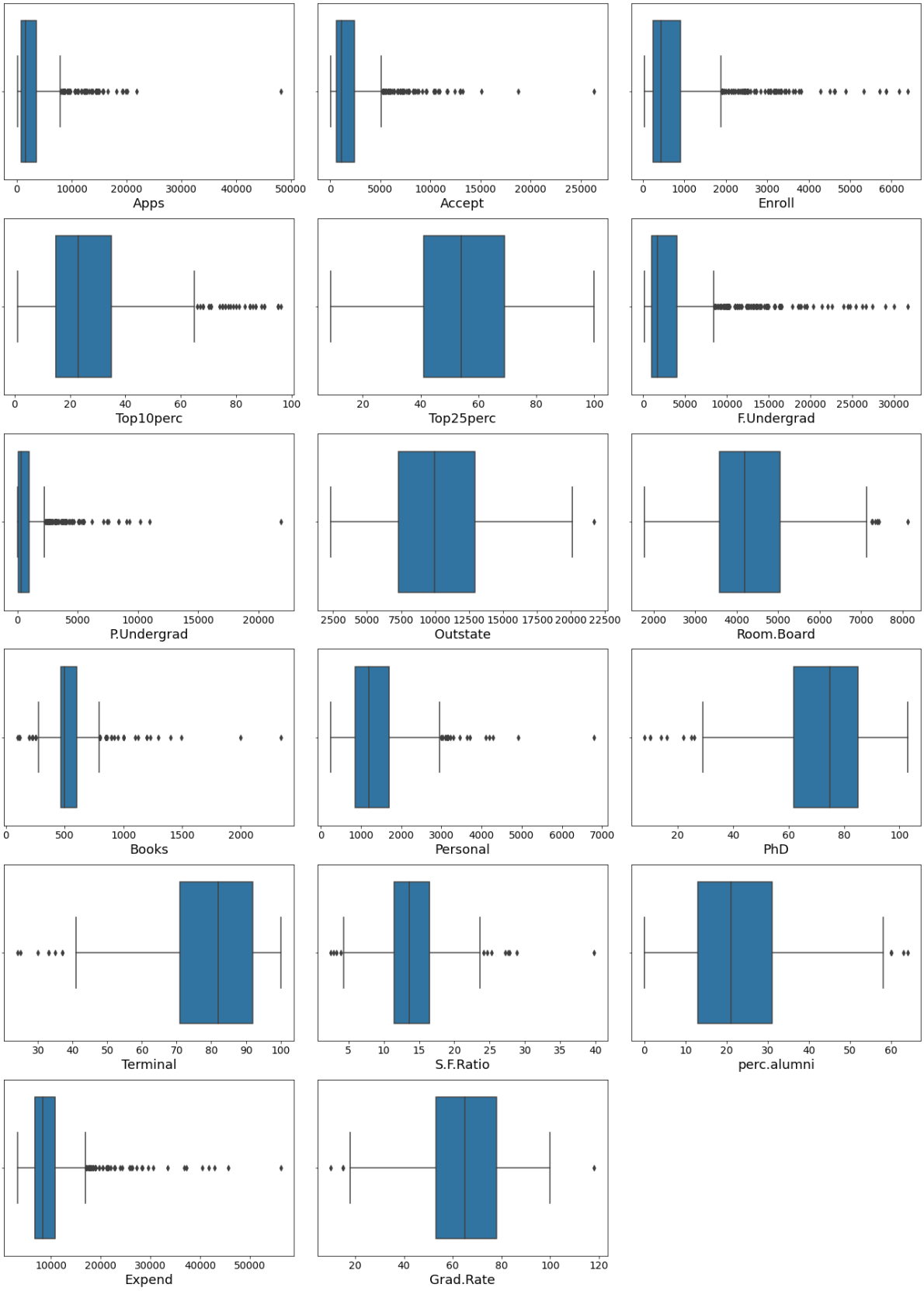

Fig4. Boxplot for Univariate Analysis

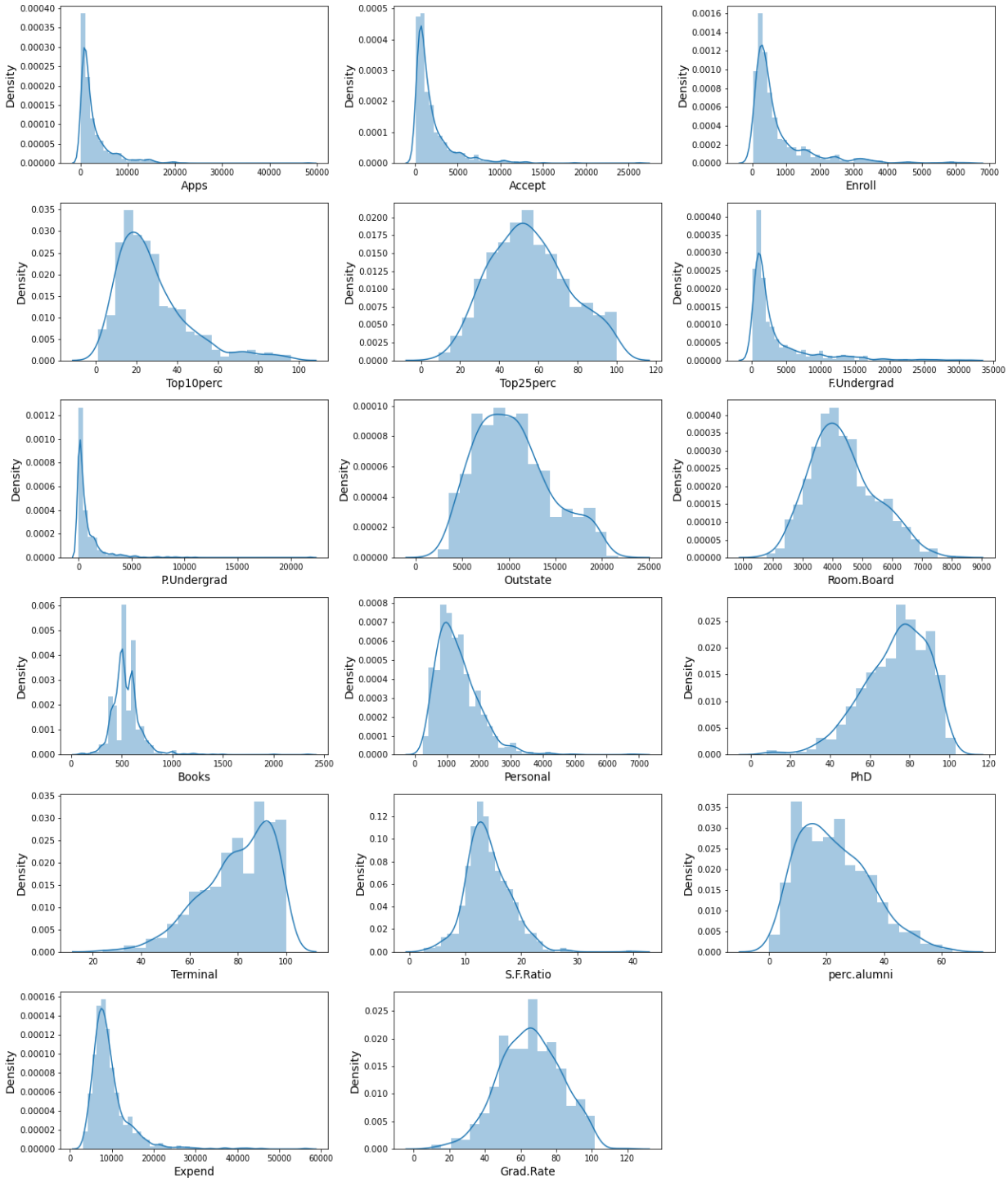

Fig5. Distplot for Univariate Analysis

→The boxplot of **Apps** variable contains outliers, the distribution of the data is skewed we could also understand that each college or university offers application in the range 3000 to 5000. The maximum applications would be around 50,000.

→ The boxplot of **Accept** variable contains outliers. The dist plot shows the majority of applications accepted from each university are in the range from 70 to 1500.The accept variable seems to be positively skewed.

→ The boxplot of **Enroll** variable contains outliers. The distribution of the data is positively skewed. From distplot, we can understand majority of the colleges have enrolled students in the range of 200 to 500 students.

→ The box plot of the students from **top 10 % of higher secondary class** contains outliers. The distribution seems to be positively skewed. There is good amount of intake about 30 to 50 students from top 10 percentage of higher secondary class.

→ The box plot for the **top 25%** has no outliers. The distribution is almost normally distributed. Majority of the students are from top 25% of higher secondary class.

→ The box plot of the **fulltime graduates** has outliers. The distribution of the data is positively skewed. In the range about 3000 to 5000 they are full time graduates studying in all the university.

→ The box plot of the **parttime graduates** has outliers. The distribution of the data is positively skewed. In the range about 1000 to 3000 they are part-time graduates studying in all the university.

→ The box plot of **outstate** has only one outlier. The distribution is almost normally distributed.

→ The **Room board** has few outliers. The distribution is normally distributed.

→ The box plot of **Books** has outliers. The distribution seems to be bimodal. The cost of books per student seems to be in the range of 500 to 100.

→ The box plot of **Personal** expense has outliers. Some student's personal expense are way bigger than the rest of the students. The distribution seems to be positively skewed.

→ The box plot of **PHD** has outliers. The distribution seems to be negatively skewed.

→ The box plot of **Terminal** contains outliers in the dataset. The distribution for the terminal also seems to be negatively skewed.

→ The **SF ratio** variable also has outliers in the dataset. The distribution is almost normally distributed. The student faculty ratio is almost same in all the university and colleges.

→ The **percentage of alumni** box plot contains outliers in the dataset. The distribution is almost normally distributed.

→ The **Expenditure** variable also has outliers in the dataset. The distribution of the expenditure is positively skewed.

→ The **graduation rate** among the students in all the university above 60%. The box plot of the graduation rate has outliers in the dataset. The distribution is normally distributed.

MULTIVARIATE ANALYSIS:

Multivariate analysis is **a Statistical procedure for analysis of data involving more than one type of measurement or observation**. It may also mean solving problems where more than one dependent variable is analyzed simultaneously with other variables.

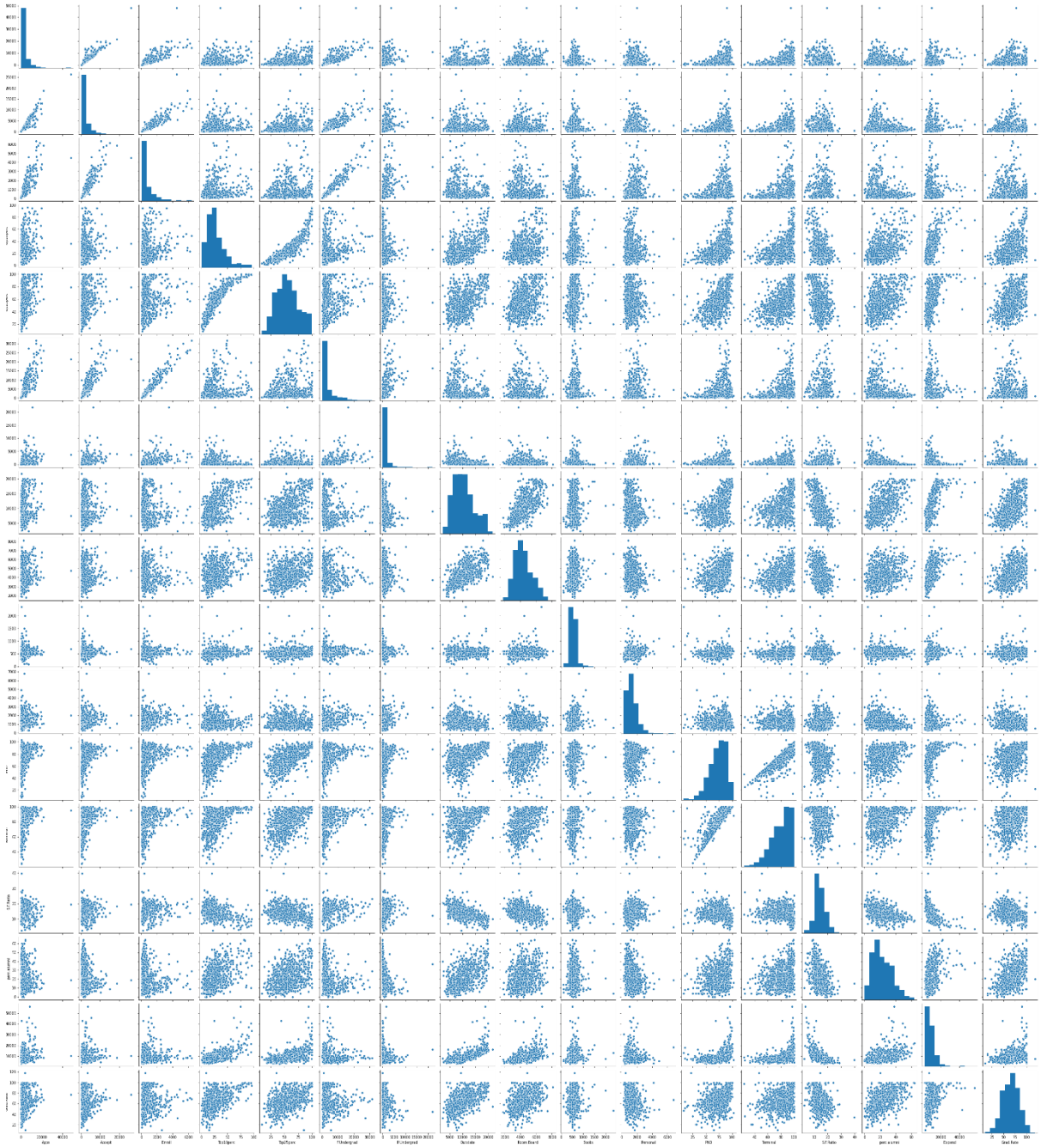

Fig6.Pairplot for Multivariate Analysis

The pair plot helps us to understand the relationship between all the numerical values in the dataset. On comparing all the variables with each other we could understand the patterns or trends in the dataset.

HEATMAP:

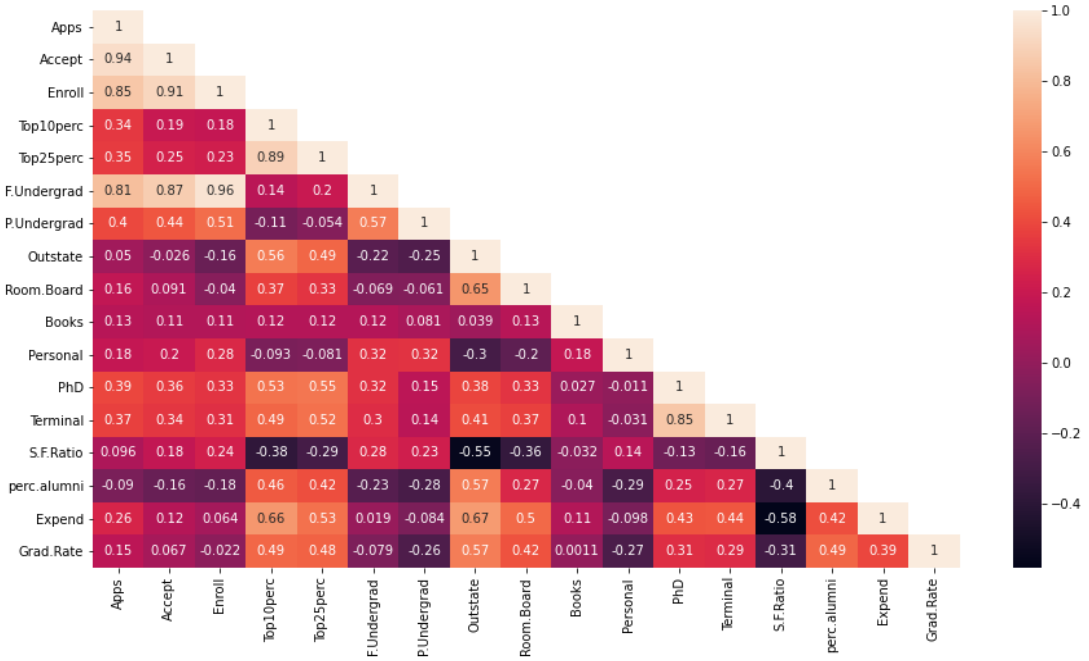

Fig7. Heatmap

Heat map gives us the correlation between two numerical values.

→The variable Apps is highly positively correlated with application accepted, students enrolled and fulltime graduates. The inference we find here is, when student submits the application, it is accepted and student enrolled as fulltime graduate.

→There is a negative correlation between application and perc alumini. It indicates that not all students of their college or university are part of alumini.

→The Application with top 10% & top 25% of higher secondary class, outage, room board, books, personal, PHD, terminal, SF ratio, expenditure and graduation ratio are positively correlated.

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

The names variable is dropped before scaling because it was a categorical variable.

Before Scaling:

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.380952	165.105360	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.171171	5221.768440	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.463320	17.177710	10.0	53.0	65.0	78.0	118.0

2.3 Dataset before scaling

Yes, Scaling is necessary for PCA in this case. Here we can observe all the columns have different scales, some of them have very short range i.e., from 2.5 to 39.8 and some have very high range from 81 to 48094.

Hence, we need to scale the data to bring all of them to same scale.

After Scaling:

	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

2.4 Dataset after scaling

Scaling is one of the most important method to follow before implementing models.

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

The comparison between the covariance and correlation matrix measures the relationship and the dependency between two variables.

Scaling is the representation of the dataset. The numbers will not change. We are bringing the dataset into one unit.

Covariance indicates the direction of the linear relationship between the variables whether it is positive or negative. By direction means it is directly proportional or inversely proportional.

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

Scaled Covariance:

	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

Fig8. Scaled Covariance

Correlation measures the strength (how much?) and the direction of the linear relationship between two variables. Strength is that is that positively correlated or negatively correlated.

$$Correlation = \frac{Cov\ (x,y)}{\sigma x * \sigma y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

Scaled Correlation:

	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	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369491	0.095633	-0.090226	0.259592	0.146755
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758	0.337583	0.176229	-0.159990	0.124717	0.067313
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469	0.308274	0.237271	-0.180794	0.064169	-0.022341
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828	0.491135	-0.384875	0.455485	0.660913	0.494989
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862	0.524749	-0.294629	0.417864	0.527447	0.477281
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.318337	0.300019	0.279703	-0.229462	0.018652	-0.078773
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.081200	0.319882	0.149114	0.141904	0.232531	-0.280792	-0.083568	-0.257001
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.038855	-0.299087	0.382982	0.407983	-0.554821	0.566262	0.672779	0.571290
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.127963	-0.199428	0.329202	0.374540	-0.362628	0.272363	0.501739	0.424942
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.000000	0.179295	0.026906	0.099955	-0.031929	-0.040208	0.112409	0.001061
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.000000	-0.010936	-0.030613	0.136345	-0.285968	-0.097892	-0.269344
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1.000000	0.849587	-0.130530	0.249009	0.432762	0.305038
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.099955	-0.030613	0.849587	1.000000	-0.160104	0.267130	0.438799	0.289527
S.F.Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628	-0.031929	0.136345	-0.130530	-0.160104	1.000000	-0.402929	-0.583832	-0.306710
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009	0.267130	-0.402929	1.000000	0.417712	0.490898
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739	0.112409	-0.097892	0.432762	0.438799	-0.583832	0.417712	1.000000	0.390343
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.571290	0.424942	0.001061	-0.269344	0.305038	0.289527	-0.306710	0.490898	0.390343	1.000000

Fig9.Scaled Correlation

→The correlation matrix which gives the strength and the relationship between the variables.

→For the correlation matrix before scaling and after scaling will remain the same.

→The variables which are highly positively correlated and the variables which are highly negatively correlated. We can observe that the variables which are moderately correlated with each other.

→The variables application, acceptance, enroll and fulltime graduates are highly positively correlated

→The top 10 percentage and top 25 percentage are highly positively correlated

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

Outliers before Scaling:

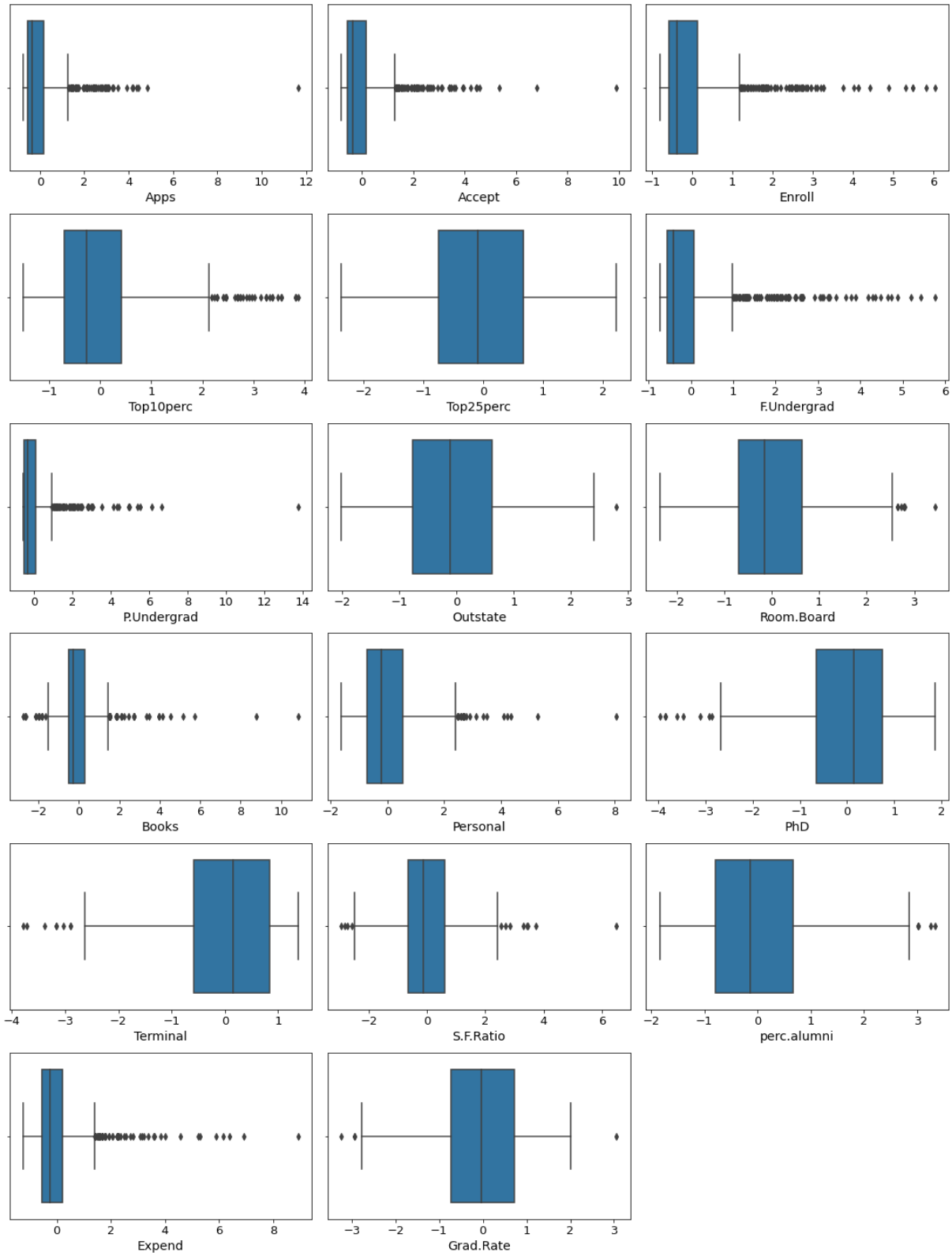

Fig10.Outliers before scaling

Outliers after scaling:

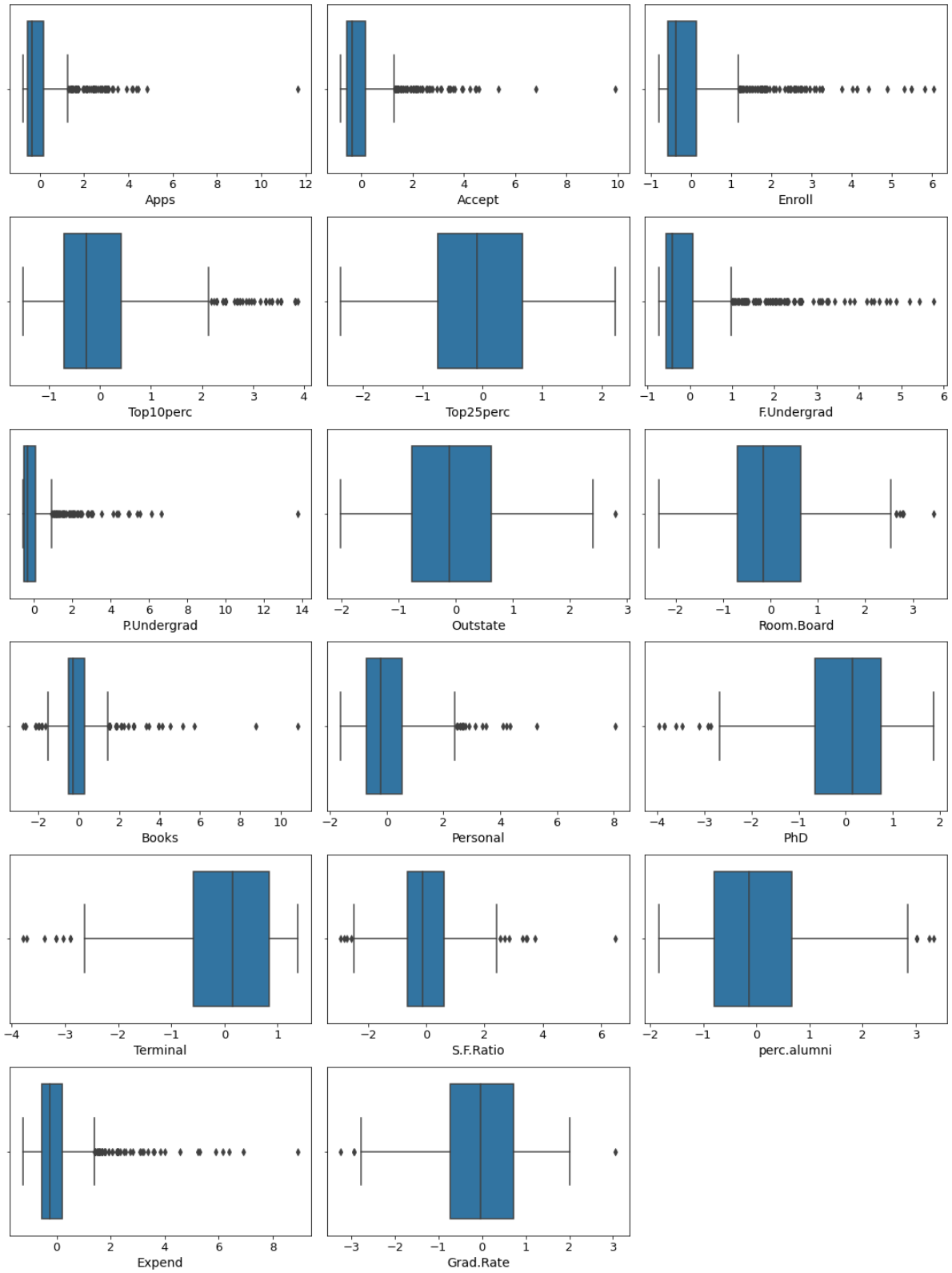

Fig11.Outliers after scaling

Outliers do not change when scaling is performed. while performing scaling, only the scale of data gets change rest everything will be same as before.

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

To Print the Eigen Values and Eigen Vectors we use the Sklearn PCA.

ho: sigma1= sigma2 = sigma3 = sigmak

h1: alteast one pair of sigma is not equal

If the p_value is less than 0.05 can proceed with PCA.

Since kmo_model value 0.8131251200373523 > 0.7. Hence, we can proceed with PCA

Eigen Values:

```
array[[ 2.48765602e-01,  2.07601502e-01,  1.76303592e-01,
         3.54273947e-01,  3.44001279e-01,  1.54640962e-01,
         2.64425045e-02,  2.94736419e-01,  2.49030449e-01,
         6.47575181e-02, -4.25285386e-02,  3.18312875e-01,
         3.17056016e-01, -1.76957895e-01,  2.05082369e-01,
         3.18908750e-01,  2.52315654e-01],
       [ 3.31598227e-01,  3.72116750e-01,  4.03724252e-01,
        -8.24118211e-02, -4.47786551e-02,  4.17673774e-01,
         3.15087830e-01, -2.49643522e-01, -1.37808883e-01,
         5.63418434e-02,  2.19929218e-01,  5.83113174e-02,
         4.64294477e-02,  2.46665277e-01, -2.46595274e-01,
        -1.31689865e-01, -1.69240532e-01],
       [-6.30921033e-02, -1.01249056e-01, -8.29855709e-02,
         3.50555339e-02, -2.41479376e-02, -6.13929764e-02,
         1.39681716e-01,  4.65988731e-02,  1.48967389e-01,
         6.77411649e-01,  4.99721120e-01, -1.27028371e-01,
        -6.60375454e-02, -2.89848401e-01, -1.46989274e-01,
         2.26743985e-01, -2.08064649e-01],
       [ 2.81310530e-01,  2.67817346e-01,  1.61826771e-01,
        -5.15472524e-02, -1.09766541e-01,  1.00412335e-01,
        -1.58558487e-01,  1.31291364e-01,  1.84995991e-01,
         8.70892205e-02, -2.30710568e-01, -5.34724832e-01,
        -5.19443019e-01, -1.61189487e-01,  1.73142230e-02,
         7.92734946e-02,  2.69129066e-01],
       [ 5.74140964e-03,  5.57860920e-02, -5.56936353e-02,
        -3.95434345e-01, -4.26533594e-01, -4.34543659e-02,
         3.02385408e-01,  2.22532003e-01,  5.60919470e-01,
        -1.27288825e-01, -2.22311021e-01,  1.40166326e-01,
         2.04719730e-01, -7.93882496e-02, -2.16297411e-01,
         7.59581203e-02, -1.09267913e-01],
       [-1.62374420e-02,  7.53468452e-03, -4.25579803e-02,
        -5.26927980e-02,  3.30915896e-02, -4.34542349e-02,
        -1.91198583e-01, -3.00003910e-02,  1.62755446e-01,
         6.41054950e-01, -3.31398003e-01,  9.12555212e-02,
         1.54927646e-01,  4.87045875e-01, -4.73400144e-02,
```

```
 -2.98118619e-01,  2.16163313e-01],
[-4.24863486e-02, -1.29497196e-02, -2.76928937e-02,
 -1.61332069e-01, -1.18485556e-01, -2.50763629e-02,
  6.10423460e-02,  1.08528966e-01,  2.09744235e-01,
 -1.49692034e-01,  6.33790064e-01, -1.09641298e-03,
 -2.84770105e-02,  2.19259358e-01,  2.43321156e-01,
 -2.26584481e-01,  5.59943937e-01],
[-1.03090398e-01, -5.62709623e-02,  5.86623552e-02,
 -1.22678028e-01, -1.02491967e-01,  7.88896442e-02,
  5.70783816e-01,  9.84599754e-03, -2.21453442e-01,
  2.13293009e-01, -2.32660840e-01, -7.70400002e-02,
 -1.21613297e-02, -8.36048735e-02,  6.78523654e-01,
 -5.41593771e-02, -5.33553891e-03],
[-9.02270802e-02, -1.77864814e-01, -1.28560713e-01,
  3.41099863e-01,  4.03711989e-01, -5.94419181e-02,
  5.60672902e-01, -4.57332880e-03,  2.75022548e-01,
 -1.33663353e-01, -9.44688900e-02, -1.85181525e-01,
 -2.54938198e-01,  2.74544380e-01, -2.55334907e-01,
 -4.91388809e-02,  4.19043052e-02],
[ 5.25098025e-02,  4.11400844e-02,  3.44879147e-02,
  6.40257785e-02,  1.45492289e-02,  2.08471834e-02,
 -2.23105808e-01,  1.86675363e-01,  2.98324237e-01,
 -8.20292186e-02,  1.36027616e-01, -1.23452200e-01,
 -8.85784627e-02,  4.72045249e-01,  4.22999706e-01,
  1.32286331e-01, -5.90271067e-01],
[ 4.30462074e-02, -5.84055850e-02, -6.93988831e-02,
 -8.10481404e-03, -2.73128469e-01, -8.11578181e-02,
  1.00693324e-01,  1.43220673e-01, -3.59321731e-01,
  3.19400370e-02, -1.85784733e-02,  4.03723253e-02,
 -5.89734026e-02,  4.45000727e-01, -1.30727978e-01,
  6.92088870e-01,  2.19839000e-01],
[ 2.40709086e-02, -1.45102446e-01,  1.11431545e-02,
  3.85543001e-02, -8.93515563e-02,  5.61767721e-02,
 -6.35360730e-02, -8.23443779e-01,  3.54559731e-01,
 -2.81593679e-02, -3.92640266e-02,  2.32224316e-02,
  1.64850420e-02, -1.10262122e-02,  1.82660654e-01,
  3.25982295e-01,  1.22106697e-01],
[ 5.95830975e-01,  2.92642398e-01, -4.44638207e-01,
  1.02303616e-03,  2.18838802e-02, -5.23622267e-01,
  1.25997650e-01, -1.41856014e-01, -6.97485854e-02,
  1.14379958e-02,  3.94547417e-02,  1.27696382e-01,
 -5.83134662e-02, -1.77152700e-02,  1.04088088e-01,
 -9.37464497e-02, -6.91969778e-02],
[ 8.06328039e-02,  3.34674281e-02, -8.56967180e-02,
 -1.07828189e-01,  1.51742110e-01, -5.63728817e-02,
  1.92857500e-02, -3.40115407e-02, -5.84289756e-02,
 -6.68494643e-02,  2.75286207e-02, -6.91126145e-01,
  6.71008607e-01,  4.13740967e-02, -2.71542091e-02,
  7.31225166e-02,  3.64767385e-02],
[ 1.33405806e-01, -1.45497511e-01,  2.95896092e-02,
  6.97722522e-01, -6.17274818e-01,  9.91640992e-03,
```

```
 2.09515982e-02,  3.83544794e-02,  3.40197083e-03,
-9.43887925e-03, -3.09001353e-03, -1.12055599e-01,
 1.58909651e-01, -2.08991284e-02, -8.41789410e-03,
-2.27742017e-01, -3.39433604e-03],
[ 4.59139498e-01, -5.18568789e-01, -4.04318439e-01,
-1.48738723e-01,  5.18683400e-02,  5.60363054e-01,
-5.27313042e-02,  1.01594830e-01, -2.59293381e-02,
 2.88282896e-03, -1.28904022e-02,  2.98075465e-02,
-2.70759809e-02, -2.12476294e-02,  3.33406243e-03,
-4.38803230e-02, -5.00844705e-03],
[ 3.58970400e-01, -5.43427250e-01,  6.09651110e-01,
-1.44986329e-01,  8.03478445e-02, -4.14705279e-01,
 9.01788964e-03,  5.08995918e-02,  1.14639620e-03,
 7.72631963e-04, -1.11433396e-03,  1.38133366e-02,
 6.20932749e-03, -2.22215182e-03, -1.91869743e-02,
-3.53098218e-02, -1.30710024e-02]])
```

Eigen Vectors:

```
array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,
       0.84849117, 0.6057878 , 0.58787222, 0.53061262, 0.4043029 ,
       0.31344588, 0.22061096, 0.16779415, 0.1439785 , 0.08802464,
       0.03672545, 0.02302787])
```

A **scree plot** is a line plot of the eigenvalues of factors or principal components in an analysis. The scree plot is used to determine the number of factors to retain in an exploratory factor analysis or principal components to keep in a principal component analysis.

scree plot always displays the eigenvalues in a downward curve, ordering the eigenvalues from largest to smallest.

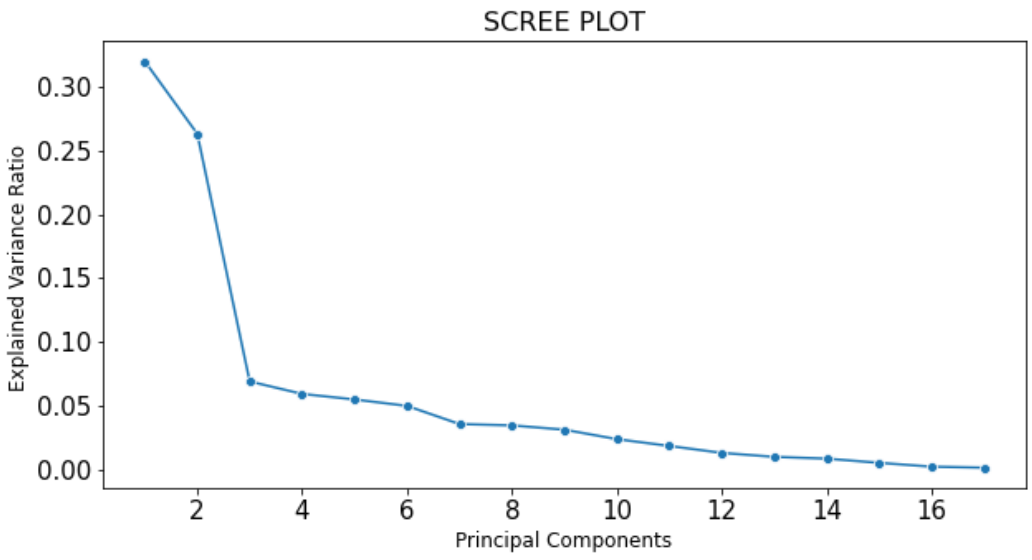

Fig12. Scree plot

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

Here the Eigen vectors are rounded to two decimal places.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	0.25	0.33	-0.06	0.28	0.01	-0.02	-0.04	-0.10	-0.09	0.05	0.04	0.02	0.60	0.08	0.13	0.46	0.36
Accept	0.21	0.37	-0.10	0.27	0.06	0.01	-0.01	-0.06	-0.18	0.04	-0.06	-0.15	0.29	0.03	-0.15	-0.52	-0.54
Enroll	0.18	0.40	-0.08	0.16	-0.06	-0.04	-0.03	0.06	-0.13	0.03	-0.07	0.01	-0.44	-0.09	0.03	-0.40	0.61
Top10perc	0.35	-0.08	0.04	-0.05	-0.40	-0.05	-0.16	-0.12	0.34	0.06	-0.01	0.04	0.00	-0.11	0.70	-0.15	-0.14
Top25perc	0.34	-0.04	-0.02	-0.11	-0.43	0.03	-0.12	-0.10	0.40	0.01	-0.27	-0.09	0.02	0.15	-0.62	0.05	0.08
F.Undergrad	0.15	0.42	-0.06	0.10	-0.04	-0.04	-0.03	0.08	-0.06	0.02	-0.08	0.06	-0.52	-0.06	0.01	0.56	-0.41
P.Undergrad	0.03	0.32	0.14	-0.16	0.30	-0.19	0.06	0.57	0.56	-0.22	0.10	-0.06	0.13	0.02	0.02	-0.05	0.01
Outstate	0.29	-0.25	0.05	0.13	0.22	-0.03	0.11	0.01	-0.00	0.19	0.14	-0.82	-0.14	-0.03	0.04	0.10	0.05
Room.Board	0.25	-0.14	0.15	0.18	0.56	0.16	0.21	-0.22	0.28	0.30	-0.36	0.35	-0.07	-0.06	0.00	-0.03	0.00
Books	0.06	0.06	0.68	0.09	-0.13	0.64	-0.15	0.21	-0.13	-0.08	0.03	-0.03	0.01	-0.07	-0.01	0.00	0.00
Personal	-0.04	0.22	0.50	-0.23	-0.22	-0.33	0.63	-0.23	-0.09	0.14	-0.02	-0.04	0.04	0.03	-0.00	-0.01	-0.00
PhD	0.32	0.06	-0.13	-0.53	0.14	0.09	-0.00	-0.08	-0.19	-0.12	0.04	0.02	0.13	-0.69	-0.11	0.03	0.01
Terminal	0.32	0.05	-0.07	-0.52	0.20	0.15	-0.03	-0.01	-0.25	-0.09	-0.06	0.02	-0.06	0.67	0.16	-0.03	0.01
S.F.Ratio	-0.18	0.25	-0.29	-0.16	-0.08	0.49	0.22	-0.08	0.27	0.47	0.45	-0.01	-0.02	0.04	-0.02	-0.02	-0.00
perc.alumni	0.21	-0.25	-0.15	0.02	-0.22	-0.05	0.24	0.68	-0.26	0.42	-0.13	0.18	0.10	-0.03	-0.01	0.00	-0.02
Expend	0.32	-0.13	0.23	0.08	0.08	-0.30	-0.23	-0.05	-0.05	0.13	0.69	0.33	-0.09	0.07	-0.23	-0.04	-0.04
Grad.Rate	0.25	-0.17	-0.21	0.27	-0.11	0.22	0.56	-0.01	0.04	-0.59	0.22	0.12	-0.07	0.04	-0.00	-0.01	-0.01

2.5 PCA loading Eigen vectors into a dataframe with original features.

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]

Linear Equation of PC1 in terms of Eigen Vectors:

PC1 = 0.25*Apps* + *0.21Accept* + *0.18Enroll* + *0.35Top10perc* + *0.34Top25perc* + *0.15 F.Undergrad* + *Outstate0.29* + *Room.Board0.25* + *Room.Board0.25* + *Books*0.06* + *Personal*-0.04 + PhD*0.32* + *Terminal*0.32* + *S.F.Ratio*-0.18 + perc.alumni*0.21* + *Expend*0.32 + Grad.Rate*0.25

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?

Adding the Eigen values we can get sum of 100.

To decide the optimum number of principal components, we need to

→Check for cumulative variance up to 90%, check the corresponding associated with 90%
→The incremental value between the components should not be less than five percent.

Cumulative values of Eigen Values:

```
array([ 32.0206282 ,  58.36084263,  65.26175919,  71.18474841,
        76.67315352,  81.65785448,  85.21672597,  88.67034731,
        91.78758099,  94.16277251,  96.00419883,  97.30024023,
        98.28599436,  99.13183669,  99.64896227,  99.86471628,
       100.        ])
```

→Based on this we can decide the optimum number of principal components as 5. Because after this, the incremental value between them is less than 5%.

The first component has **32.02%** variance in data

The first two component has **58.36%** variance in data

The first three component has **65.26%** variance in data

The first four component has **71.18%** variance in data

The first five component has **76.67%** variance in data

→ Hence, we select 5 principal components for this case study. They are given below:

```
array([[ 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],
       [-0.06309206, -0.10124911, -0.08298561,  0.03505552, -0.02414793,
        -0.06139292,  0.13968171,  0.04659888,  0.14896739,  0.67741165,
         0.49972112, -0.12702837, -0.06603755, -0.2898484 , -0.14698927,
         0.22674398, -0.20806465],
       [ 0.28131048,  0.26781741,  0.16182683, -0.05154724, -0.10976655,
         0.10041226, -0.15855848,  0.13129135,  0.18499599,  0.08708922,
        -0.23071057, -0.53472483, -0.51944302, -0.16118948,  0.01731422,
         0.0792735 ,  0.26912907],
       [ 0.00574126,  0.05578627, -0.05569348, -0.39543429, -0.42653361,
        -0.04345458,  0.30238543,  0.22253197,  0.56091948, -0.12728883,
        -0.22231102,  0.14016632,  0.20471973, -0.07938824, -0.21629741,
         0.07595813, -0.10926791]])
```

→The Eigen vectors or Principal components for this case study are five, we can understand how much each variable contributes to the principal components.

→ With this Eigen vectors we can understand which variable has more weightage and influences the dataset in the 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]

→This business case study is about education dataset which contain the names of various colleges, which has various details of colleges and university.

→To understand more about the dataset, we performed univariate and multivariate analysis which gives us the understanding about the variables.

→ From univariate analysis we can understand the distribution of the dataset, skew, and patterns in the dataset.

→From multivariate analysis we can understand the correlation of variables.

→ Inference of multivariate analysis shows we can understand multiple variables highly correlated with each other.

→The scaling helps us to standardize the variable in one scale.

→Outliers are imputed using IQR values once the values are imputed, we can perform PCA.

→The principal component analysis (PCA) is used reduce the multicollinearity between the variables.

→Depending on the variance of the dataset we can reduce the PCA components.

→The PCA components for this business case are 5 where we can be able to understand the maximum variance of the dataset.

→Using these components, we can now understand the reduced multicollinearity in the dataset. with this analysis we can perform further analysis and model building.

THE END!!