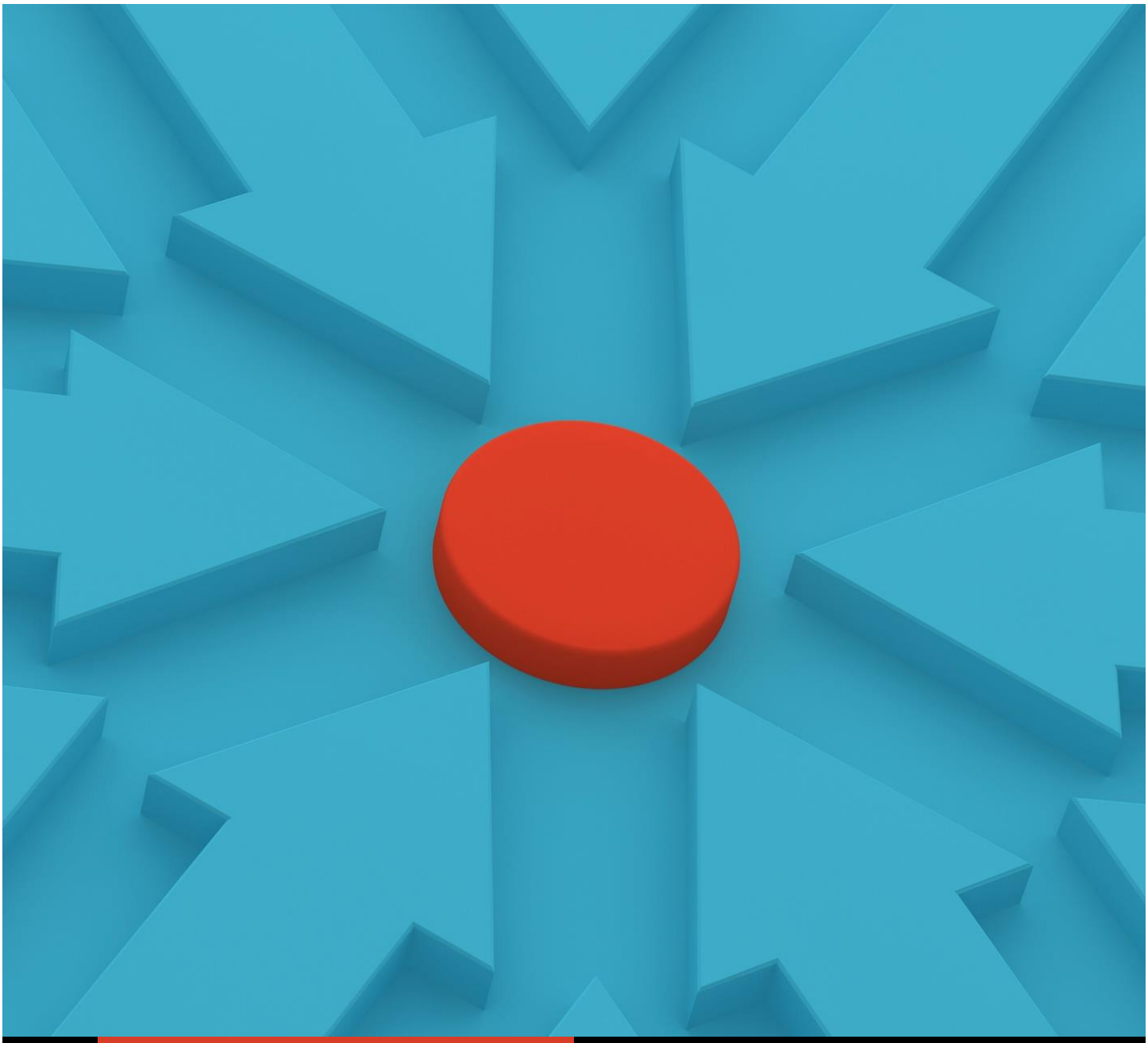


ADVANCED STATISTICS

PROJECT REPORT

SUNDAR RAM S
PGPDSBA
ONLINE DEC_C 2021
03-APR-2022

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF EQUATIONS

TABLE OF TABLES

EXECUTIVE SUMMARY

This report is based on two data sets that contain Salary and Education post 12th standard information. Salary dataset is analyzed to present the understanding, insights and business conclusions through ANOVA while the Education post 12th standard is to present the understanding, insights and business impact using Exploratory Data Analysis and Principal Component Analysis.

INTRODUCTION

This report provides a detailed explanation on approach used, record inferences, insights and provide suitable business solutions. The techniques of Advanced Statistics that employs ANOVA, Exploratory Data Analysis & Principal Component Analysis are leveraged in this exercise.

PROBLEM 1 – SALARY DATA ANALYSIS

Problem Statement

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High School Graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination. [Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.]

Dataset Description

1 Education : Educational Qualification of the person
(High School Graduate, Bachelor, and Doctorate)

2 Occupation : Occupation of the person
(Administrative and Clerical, Sales, Professional/Specialty & Executive/ Managerial)

3 Salary : Salary of the person

Sample of the 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

Table 1.1 – Sample of the Dataset

The data consists of 3 columns having the details - Educational Qualification, Occupation & Salary details of 40 persons.

Exploratory Data Analysis

Variable type

Education : object

Occupation : object

Salary : int64

Variable type is the type of data each column is holding. Here we are able to notice that, Education & Occupation are variables of type object & Salary is of type int64.

Check for missing values in the dataset

Education : 40 non-null

Occupation : 40 non-null

Salary : 40 non-null

It is evident that there are no missing values in the entire dataset

ANOVA

ANOVA is one of the types of Hypothesis testing, which is an extension of t-test for comparison of more than 2 population means. ANOVA is the abbreviated form of **AN**alysis **O**f **VA**riance. ANOVA is further classified in to One Way ANOVA & Two Way ANOVA.

Problem 1A - One Way ANOVA

It is a hypothesis test used to test the equality of 3 or more means simultaneously by analyzing their variances. There will be only 1 independent variable with three or more levels i.e., One Way ANOVA examines the effect of one variable only.

We assume that, (i) Samples are drawn from different populations are independent & random, (ii) The response variable of all populations is continuous and ideally normally distributed, (iii) The variances of all the populations are approximately equal & (iv) Population differs in their means.

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

Null & Alternate Hypothesis based on Educational Qualification

Ho: Mean Salary of individuals with different educational Qualification are equal

Ha: Not all Mean Salary of individuals with different educational Qualification are equal

Null & Alternate Hypothesis based on Occupation

Ho: Mean Salary of individuals with different occupations are equal

Ha: Not all Mean Salary of individuals with different occupations are equal

Q 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.

ANOVA decomposes the total variation in to components of variation. The total sum of squares is equal to the sum of square due to causes.

Total Sum of Squares = Treatment Sum of Squares + Error Sum of Squares

Eq. 1.1 – One Way ANOVA

The variation in sum of squares of the response variable (dependent variable) is caused only by treatment and anything unexplained by the treatment is attributed to error term.

To perform one way ANOVA, we use the pandas to first convert the Education variable type to Categorical type. We then create a formula, import the formula in to a model to calculate ordinary least squares and create the ANOVA table. The necessary algorithms and linear models are available in the statsmodels library of Python which are imported.

formula = 'Salary ~ C(Education)' .....Eq. 1.2 – Formula One Way ANOVA

model = ols(formula, df).fit().....Eq. 1.3 – Model One Way ANOVA

aov_table = anova_lm(model) ..... Eq. 1.4 – ANOVA Table One Way ANOVA

Formula defines the problem statement to analyze Salary as a function of Education. Model defines the algorithm Ordinary Least Squares (ols) that needs to be used on the formula and the dataset (df) on which the algorithm needs to be applied. The anova_lm(model) develops the required ANOVA table based on which we decide to reject or accept the null hypothesis.

	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 1.2 – One Way ANOVA Education w.r.t Salary

Assuming alpha to be 0.05, we can notice that p value (1.257709e-08) is less than Alpha. Hence, ***we reject the null hypothesis*** and conclude that, not all Mean Salary of individuals with different educational Qualification are equal (for at least one pair the mean is not equal).

Q 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.

Similar to Q 1.2, here we would be performing one way ANOVA for Occupation w.r.t Salary. Hence the Occupation variable is first converted in to categorical type, Education in the formula is replaced with occupation and the same subsequent steps are followed.

formula = 'Salary ~ C(Occupation)' .....Eq. 1.5 – Formula One Way ANOVA

ANOVA table for Q1.3 as follows

	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 1.3 – One Way ANOVA Occupation w.r.t Salary

Assuming alpha to be 0.05, we can notice that p value (0.458508) is greater than Alpha. Hence, ***we fail to reject the null hypothesis*** and conclude that, Mean Salary of individuals with different occupation are equal with 65 % confidence level.

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.

From the above solutions it is evident that null hypothesis is rejected in 1.2. On analyzing using a pivot table, we are able to understand that the mean Salary of HS Graduate is significantly different from the mean salary of Doctorate with a value of 1,33,388. Also, from the Pivot table we are able to interpret that the mean salary shows an increasing trend with educational qualification. Higher the educational qualification, higher is the mean salary.

Row Labels	Average of Salary
Bachelors	1,65,152.93
Doctorate	2,08,427.00
HS-grad	75,038.78
Grand Total	**1,62,186.88**

Table 1.4 – Mean salary w.r.t Educational Level

Problem 1B – Two Way ANOVA

Two-way ANOVA is an extension of one way ANOVA. It is a hypothesis test used to compare the mean difference between groups that have been split on two independent variables. It helps us understand if there is an interaction between two independent variables on dependent variable. There will be 2 independent variables with multiple levels.

We assume that, (i) Dependent variable should be measured at continuous level, (ii) Two independent variables should each consist of two or more categorical independent groups, (iii) There should be no significant outliers & (iv) Dependent variables should be approximately normally distributed for each combination of the group of two independent variables.

Q 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.

In the above One Way ANOVA, we were able to interpret that Educational Qualification as the primary factor influencing Salary. However, we still have only 65 % confidence level that mean salary is equal for different occupation. Hence, we also need to find if there is some interaction between these two treatments (Educational Qualification & Occupation).

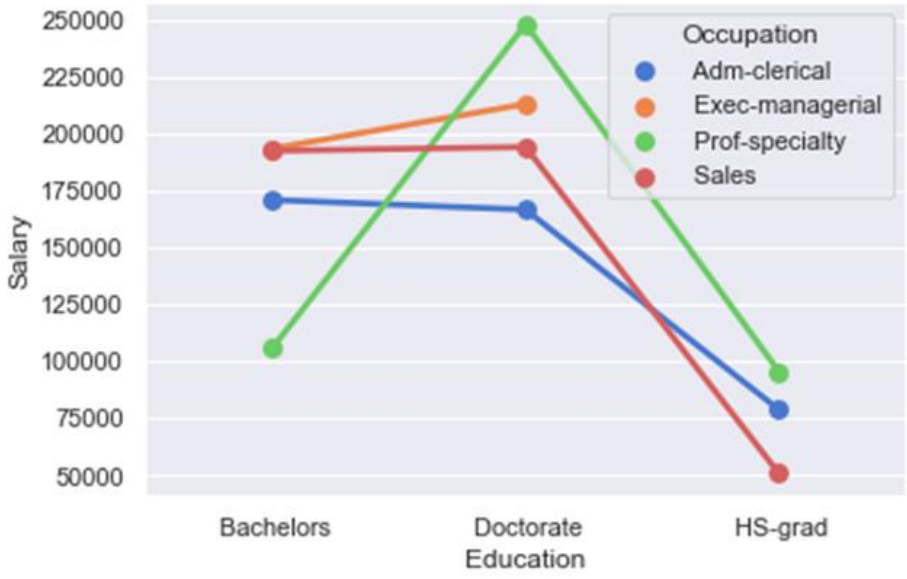


Fig 1.1 – Point Plot - Interaction b/w Educational Qualification & Occupation

From the above point plot, it is evident that occupation interacts with Educational Qualification of the person in determining the salary.

Some of the interpretations are listed below

- To become a high paid Professor with a specialization in particular subject, it is important to complete the Doctorate education in that subject. This also enhances credibility.
- A doctorate person is paid more than a Bachelors person in Managerial, Sales & Professor, while a bachelor is paid more in the Adm Clerical occupation.
- A HS Grad pursuing sales has the least mean salary while HS Grad pursuing his profession in Adm Clerical or Professor is paid higher.
- The most preferred job for Bachelors is Adm Clerical/ Exec Managerial, while it is Professor for Doctorate Holder & HS Grad.

Q 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?

Null & Alternate Hypothesis based on Educational Qualification

α_o: Mean Salary of individuals with different educational Qualification are equal

α_a: Not all Mean Salary of individuals with different educational Qualification are equal

Null & Alternate Hypothesis based on Occupation

β_o: Mean Salary of individuals with different occupations are equal

β_a: Not all Mean Salary of individuals with different occupations are equal

Null & Alternate Hypothesis based on interaction of Education with Occupation

γ_o: There is no interaction between treatments Education and Occupation to influence the mean salary

γ_a: There is interaction between treatments Education and Occupation and influences the mean salary

The steps in Two Way ANOVA are same as that of One-way ANOVA, except the fact that, we would be involving two treatments in the formula and also the interaction between the two treatments. The formula will be designed as per the below equation.

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

Eq. 1.6 – Formula Two Way ANOVA with Interaction

Formula defines the problem statement to analyze Salary as a function of Education, Occupation and an interaction term i.e., influence of education on occupation.

	df	sum_sq	mean_sq	F	\
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	
Residual	29.0	2.062102e+10	7.110697e+08	NaN	

	PR(>F)
C(Education)	5.466264e-12
C(Occupation)	7.211580e-02
C(Education):C(Occupation)	2.232500e-05
Residual	NaN

Table 1.5 – Two-way ANOVA with interaction

Assuming alpha to be 0.05, we can notice the below points.

- For Education treatment, p value (5.466264e-12) is less than 0.05. Hence, we reject the Null hypothesis and conclude that education plays a significant role in determining the salary of a person. From Two Way ANOVA P value, we are further able to capitalize our findings, as p value in two-way ANOVA is further less than that obtained in one way ANOVA.
- For Occupation treatment, p value (7.211580e-02) is greater than 0.05. Hence, we fail to reject the Null hypothesis. This was also noticed in the On way ANOVA worked out for Occupation. However, the p value in two-way ANOVA tends to be closer towards 0.05. This is because of the interaction that Occupation has with the Education.

- For the interaction term, p value (2.232500e-05) is lesser than 0.05. Hence, we reject the null hypothesis and conclude that there is interaction between Occupation and education and this also influences the mean salary of the person.

This statistical procedure of Two-way ANOVA establishes with sufficient proof, the result obtained in the point plot for interaction (Refer Fig 1.1).

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

By performing one way and two-way ANOVA in this case study, we are able to identify the best paying job for a particular profession. This helps businesses in understanding and fixing salaries of persons with different educational qualification. This further aids in cost cutting, where businesses may hire a lower education qualification person in place of a higher educational qualification to perform the same job.

PROBLEM 2 – EDUCATION POST 12TH STANDARD

Problem Statement

The dataset Education - Post 12th Standard.csv contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx.

Dataset Description

Names	:	Names of various university and colleges
Apps	:	Number of applications received
Accept	:	Number of applications accepted
Enroll	:	Number of new students enrolled
Top10perc	:	Percentage of new students from top 10% of Higher Secondary class
Top25perc	:	Percentage of new students from top 25% of Higher Secondary class
F.Undergrad	:	Number of full-time undergraduate students
P.Undergrad	:	Number of part-time undergraduate students

Outstate	:	Number of students for whom the particular college or university is Out-of-state tuition
Room.Board	:	Cost of Room and board
Books	:	Estimated book costs for a student
Personal	:	Estimated personal spending for a student
PhD	:	Percentage of faculties with Ph.D.'s
Terminal	:	Percentage of faculties with terminal degree
S.F.Ratio	:	Student/faculty ratio
perc.alumni	:	Percentage of alumni who donate
Expend	:	The Instructional expenditure per student
Grad.Rate	:	Graduation rate

Sample of the Dataset

	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
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15

Table 2.1 – Sample of the Dataset

The data consists of 777 records with 18 columns having the details of educational institutions post 12th standard.

Q 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 (EDA) is the initial phase of any Data Science project. It is an approach to analyze data using both visual and non-visual techniques. It involves thorough analysis of data to understand the current business situation. EDA involves the below four approaches.

1. Describe Data

2. Data Pre-processing
3. Data Visualization
4. Data Preparation

Describe data

The data consists of 777 records with 18 columns having the details of educational institutions post 12th standard.

Variable Type

Names	:	Object
Apps	:	int64
Accept	:	int64
Enroll	:	int64
Top10perc	:	int64
Top25perc	:	int64
F.Undergrad	:	int64
P.Undergrad	:	int64
Outstate	:	int64
Room.Board	:	int64
Books	:	int64
Personal	:	int64
PhD	:	int64
Terminal	:	int64
S.F.Ratio	:	float64
perc.alumni	:	int64
Expend	:	int64
Grad.Rate	:	int64

There are 16 variables of type int64 and 1 variable each in type float64 & object.

Check for missing values in the Dataset

Names	:	777 non-null
Apps	:	777 non-null
Accept	:	777 non-null
Enroll	:	777 non-null
Top10perc	:	777 non-null
Top25perc	:	777 non-null
F.Undergrad	:	777 non-null
P.Undergrad	:	777 non-null
Outstate	:	777 non-null
Room.Board	:	777 non-null
Books	:	777 non-null
Personal	:	777 non-null
PhD	:	777 non-null
Terminal	:	777 non-null
S.F.Ratio	:	777 non-null
perc.alumni	:	777 non-null
Expend	:	777 non-null
Grad.Rate	:	777 non-null

It is evident that there are no null values in the given data set.

Summary of the Dataset

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.00000
mean	3001.638353	2018.804376	779.972973	27.558559	55.796654	3699.907336	855.298584	10440.669241	4357.526384	549.380952	1340.642214	72.660232	79.702703	14.089704	22.743887	9660.171171	65.46332
std	3870.201484	2451.113971	929.176190	17.640364	19.804778	4850.420531	1522.431887	4023.016484	1096.696416	165.105360	677.071454	16.328155	14.722359	3.958349	12.391801	5221.768440	17.17771
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	2340.000000	1780.000000	96.000000	250.000000	8.000000	24.000000	2.500000	0.000000	3186.000000	10.00000
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	7320.000000	3597.000000	470.000000	850.000000	62.000000	71.000000	11.500000	13.000000	6751.000000	53.00000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	9990.000000	4200.000000	500.000000	1200.000000	75.000000	82.000000	13.600000	21.000000	8377.000000	65.00000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	12925.000000	5050.000000	600.000000	1700.000000	85.000000	92.000000	16.500000	31.000000	10830.000000	78.00000
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000	31643.000000	21836.000000	21700.000000	8124.000000	2340.000000	6800.000000	103.000000	100.000000	39.800000	64.000000	56233.000000	118.00000

Table 2.2 – Summary of the Dataset

Summary of the dataset gives us the data on the 5-point summary of all the variables.

Data Pre-Processing

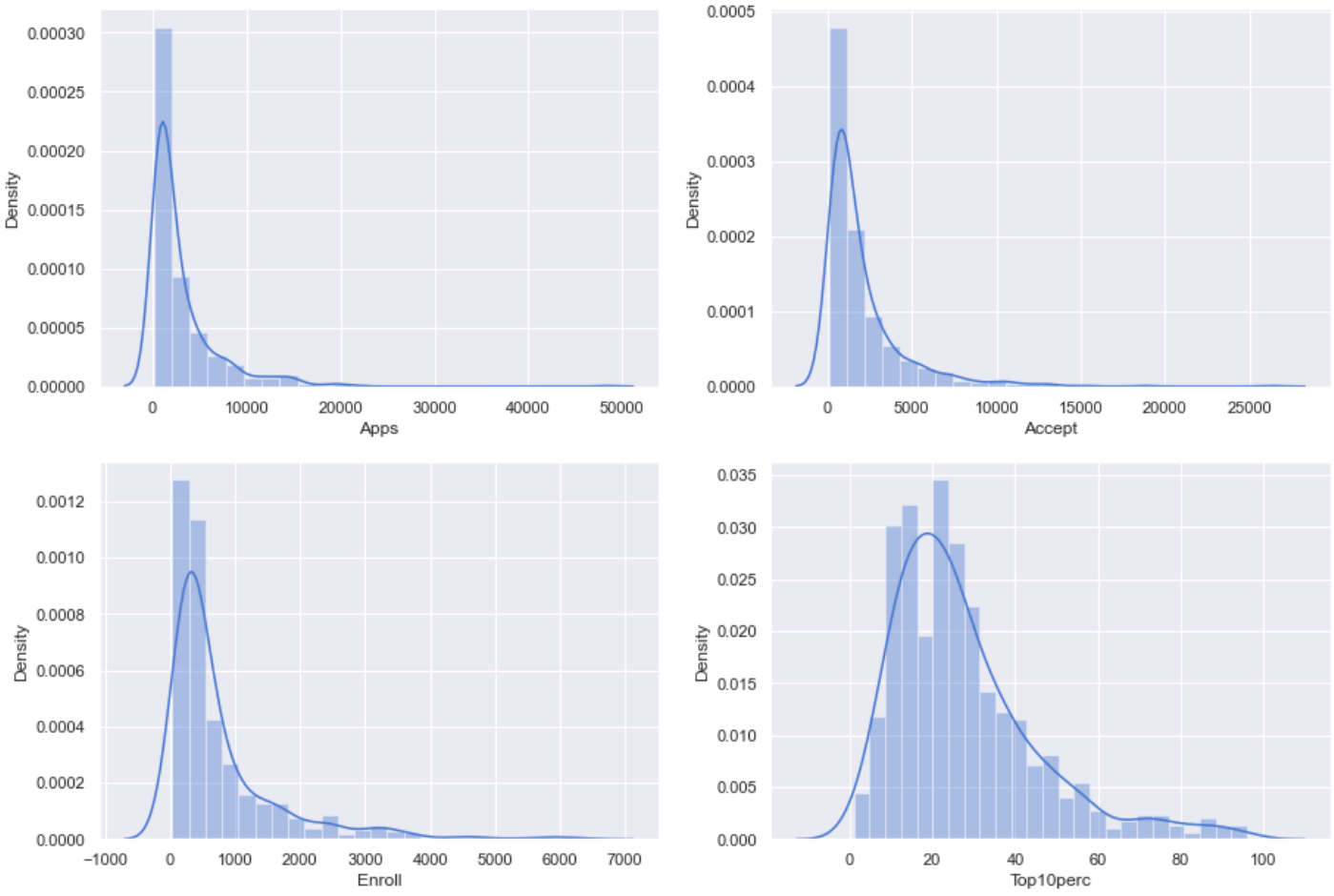
Data Pre-processing is a process of cleansing noise/ unwanted data points that might impact the outcome. Noise can be in the form of bad values, anomalies, missing values, & not useful data. From the summary of the dataset, we can conclude that there no noise in the dataset. It is also noted that there are no duplicate rows in the data set.

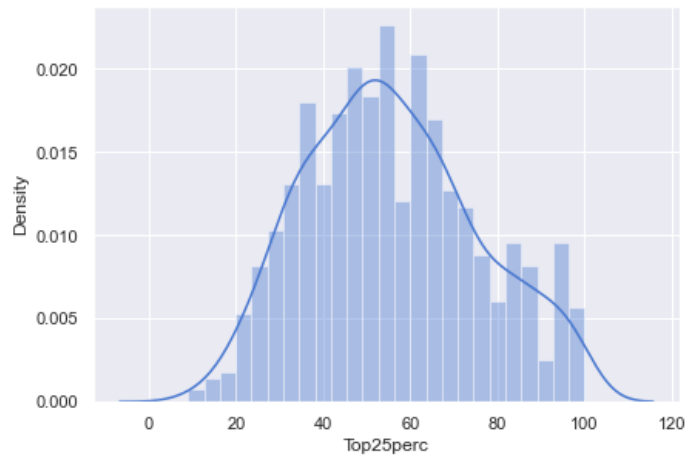
Data Visualization

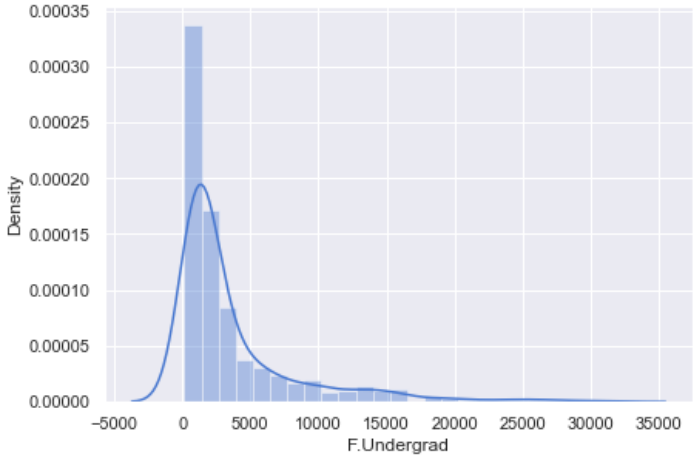
Data Visualization is a technique for creating diagrams, images and charts for the purpose of analysis through visualization. Data Analysis using visualization includes Univariate, Bivariate and Multivariate analysis.

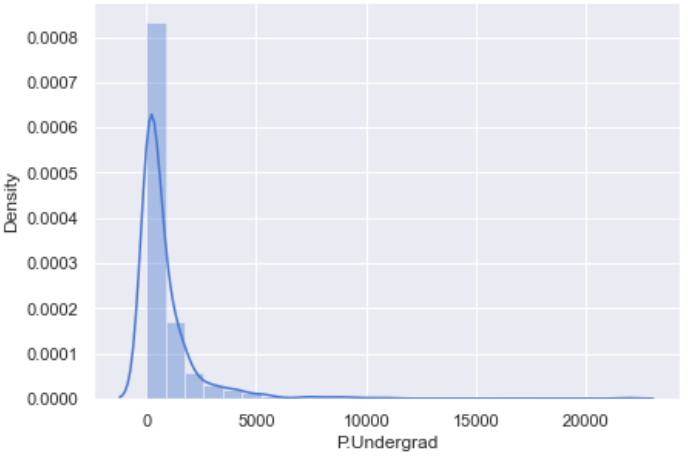
Univariate Analysis:

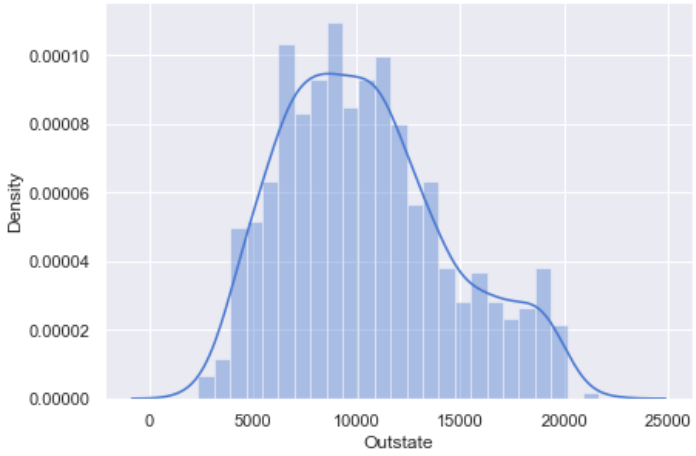
Univariate as the name suggests is the analysis of a single variable. The plot can be histogram/ box plot. Histogram helps us understand the distribution, while box plot helps us understand the five-point summary.

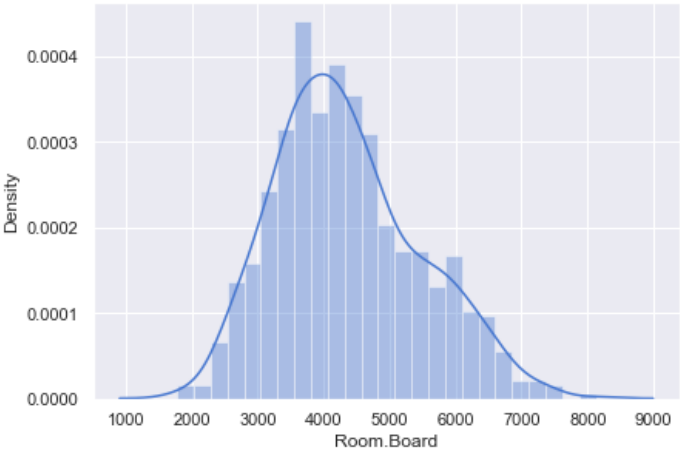


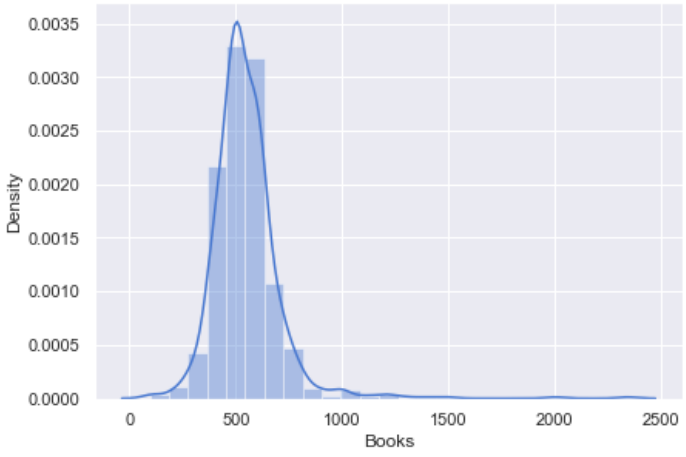

Density
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
Top25perc

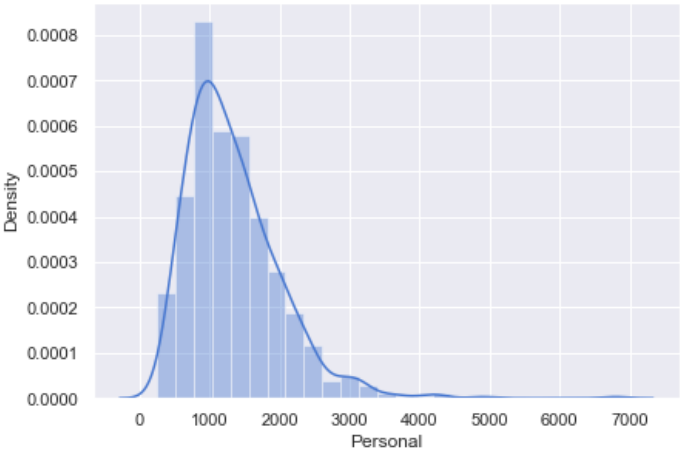

Density
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
−5000
0
5000
10000
15000
20000
25000
30000
35000
F.Undergrad

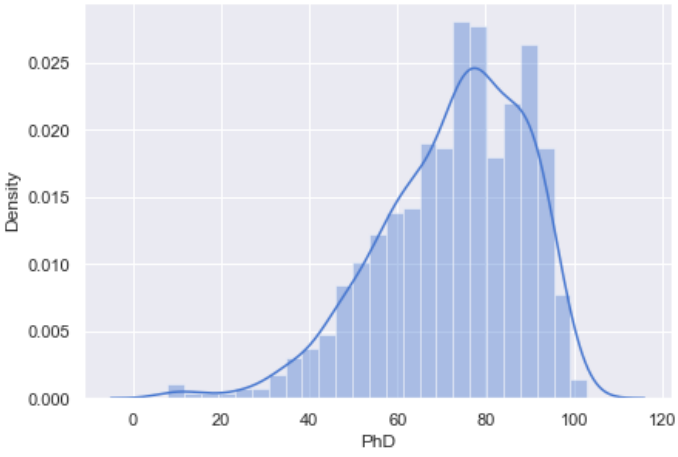

Density
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
5000
10000
15000
20000
P.Undergrad


Density
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
5000
10000
15000
20000
25000
Outstate


Density
0.0004
0.0003
0.0002
0.0001
0.0000
1000
2000
3000
4000
5000
6000
7000
8000
9000
Room.Board


Density
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
500
1000
1500
2000
2500
Books


Density
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
1000
2000
3000
4000
5000
6000
7000
Personal


Density
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
PhD

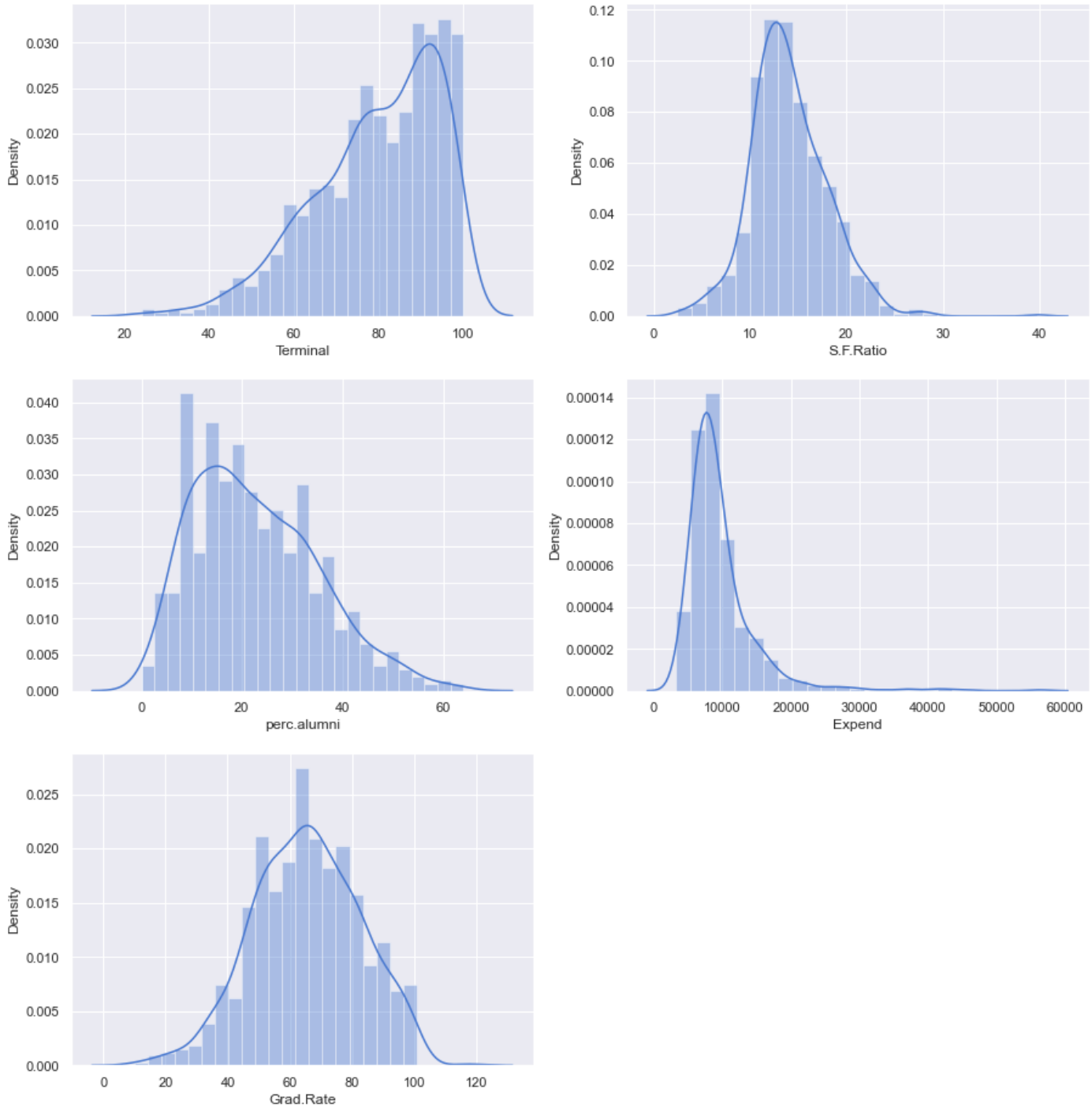
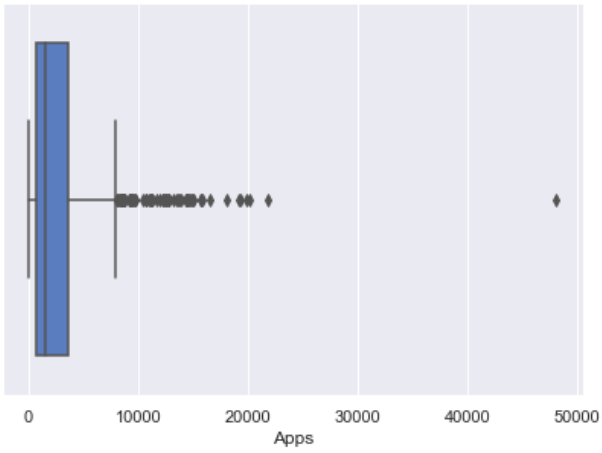
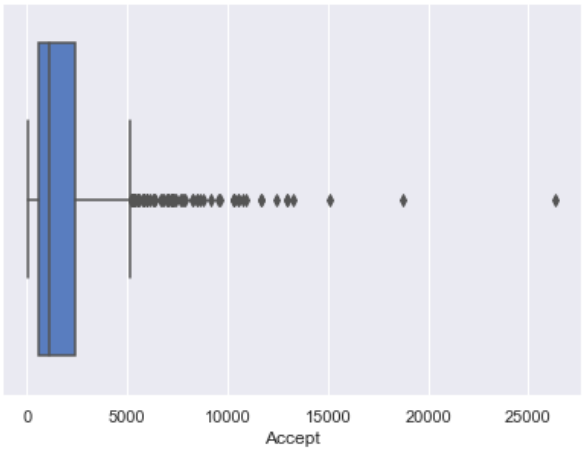
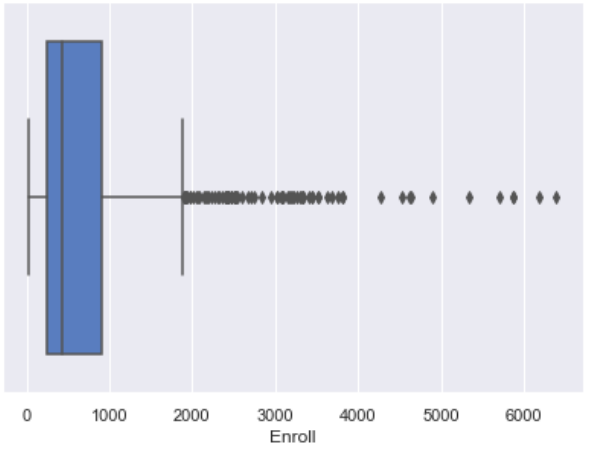


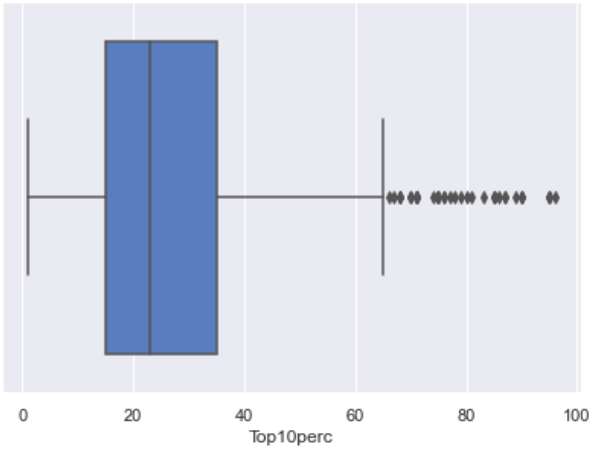
Fig 2.1 – Histogram Plots of all variables – Univariate Analysis

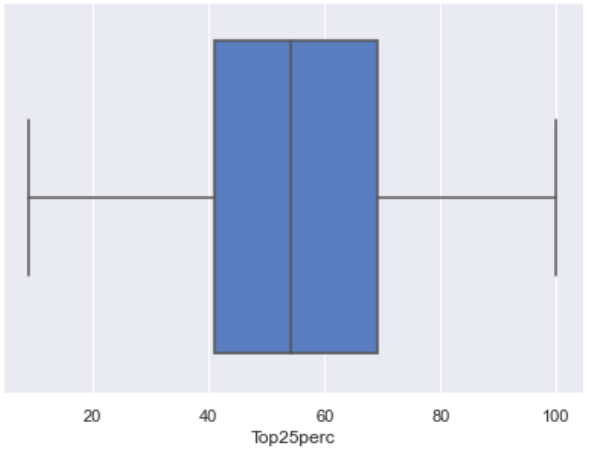
From the above plots, we can see that, most of the variables are skewed. Variables Apps, Accept, Enroll, Top10perc, F.Undergrad, P.Undergrad, Outstate, Room.Board, Books, Personal, S.F.Ratio, perc.alumni, Expend are right skewed while PhD, Terminal & Grad.Rate are left skewed. Top25perc distribution is tending to be normal.

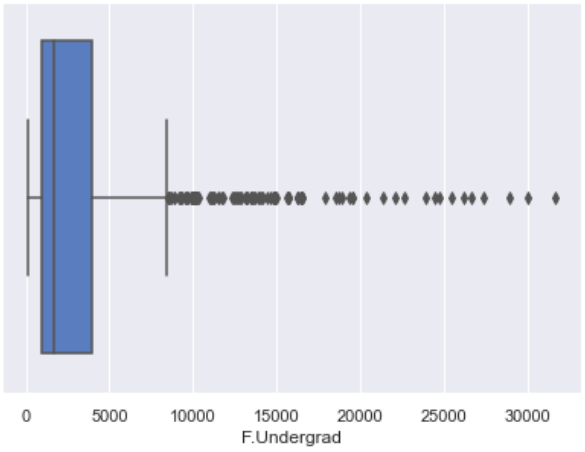

0
10000
20000
30000
40000
50000
Apps

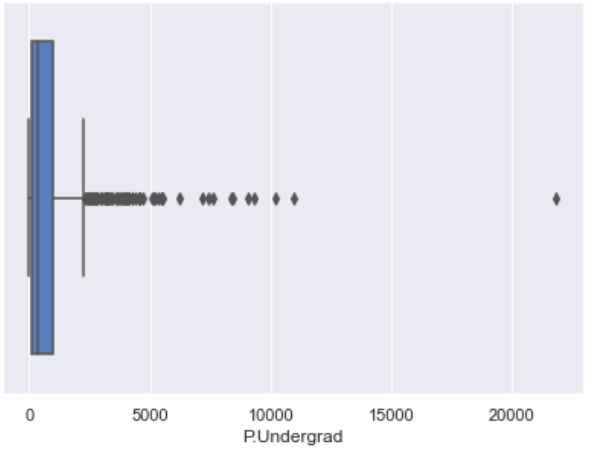

0
5000
10000
15000
20000
25000
Accept

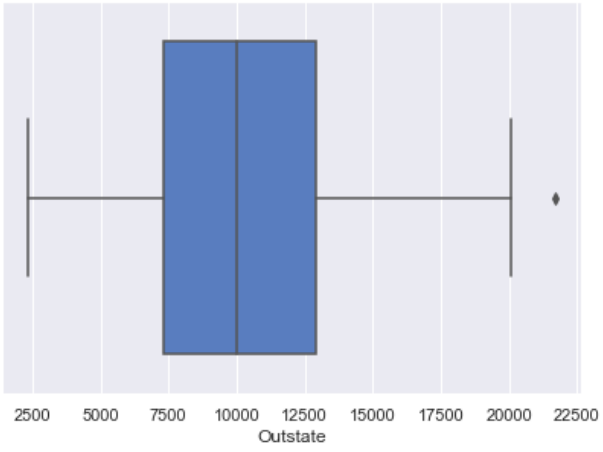

0
1000
2000
3000
4000
5000
6000
Enroll

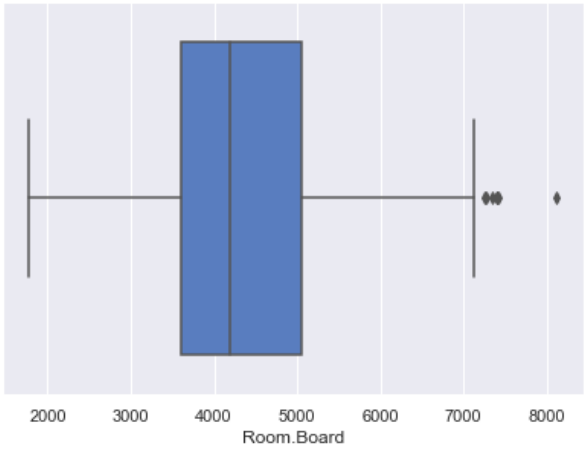

0
20
40
60
80
100
Top10perc

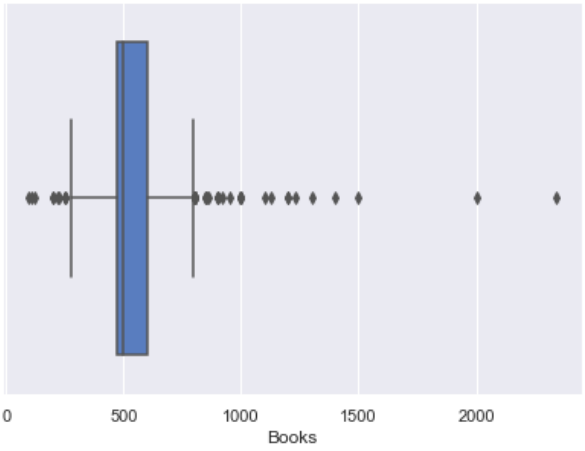

20
40
60
80
100
Top25perc

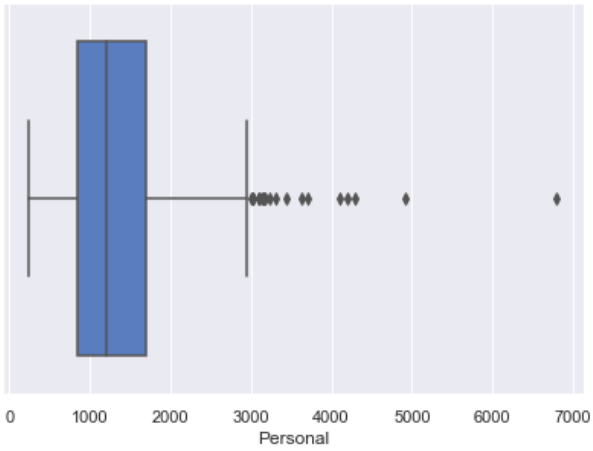

0
5000
10000
15000
20000
25000
30000
F.Undergrad

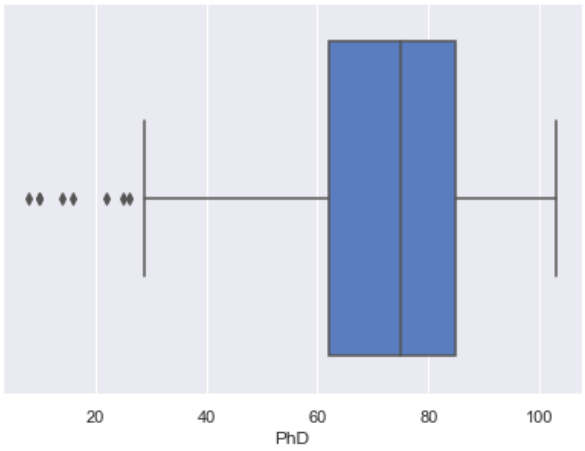

0
5000
10000
15000
20000
P.Undergrad

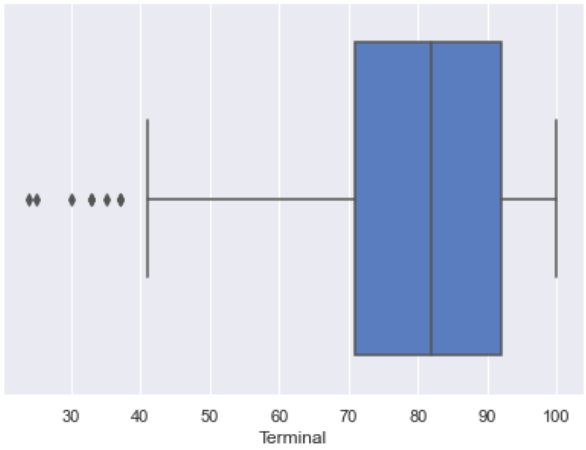

2500
5000
7500
10000
12500
15000
17500
20000
22500
Outstate

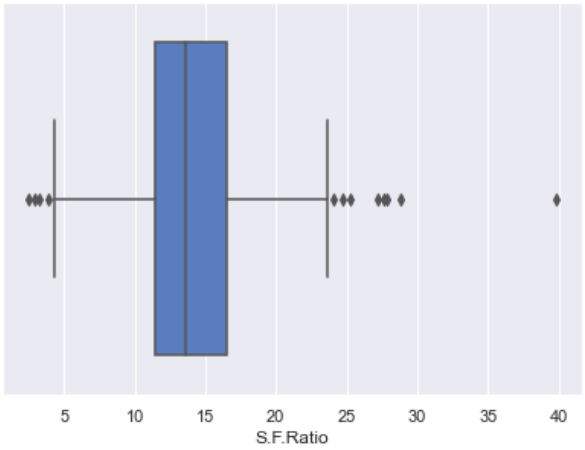

2000
3000
4000
5000
6000
7000
8000
Room.Board

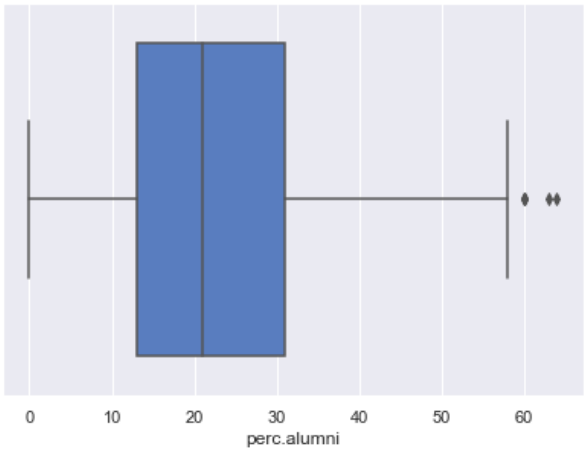

0
500
1000
1500
2000
Books

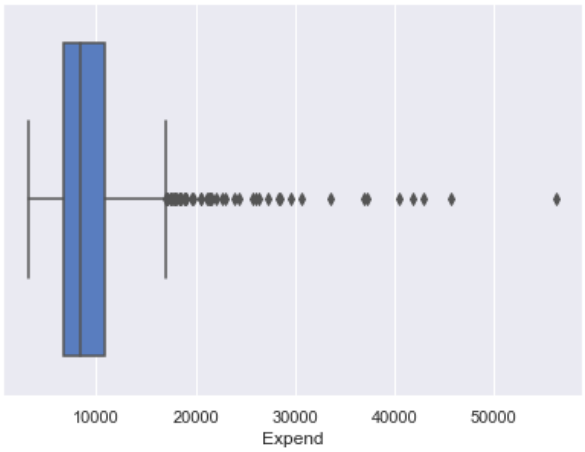

0
1000
2000
3000
4000
5000
6000
7000
Personal


20
40
60
80
100
PhD


30
40
50
60
70
80
90
100
Terminal


5
10
15
20
25
30
35
40
S.F.Ratio


0
10
20
30
40
50
60
perc.alumni


10000
20000
30000
40000
50000
Expend

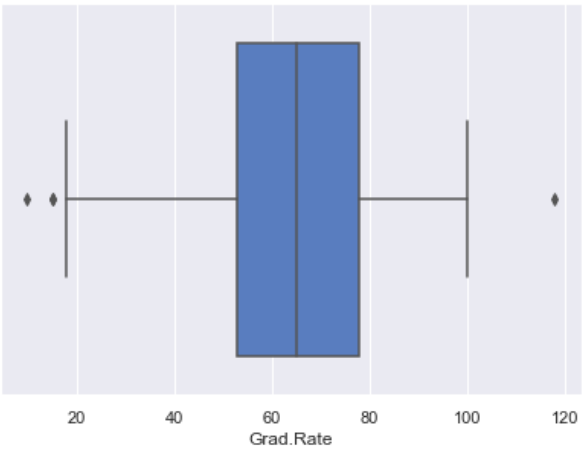


Fig 2.2 – Box Plots of all variables – Univariate Analysis

From the above plot, it is evident that, there are significant number of outliers in almost all the variables except Top25Perc. Top25perc has no outliers while variables Apps, Accept, Enroll, Top10perc, F.Undergrad, P.Undergrad, Outstate, Room.Board, Personal, perc.alumni, Expend have outliers to the right side of the max whisker and PhD, Terminal have outliers to the left of min whisker. Books, S.F.Ratio, & Grad.Rate have outliers in both sides of the box plot.

Multivariate Analysis:

Multivariate Analysis, as the name suggests establishes the relationship of a variable with each of the other variable. This is obtained through a heat map (correlation plot) and pair plot.

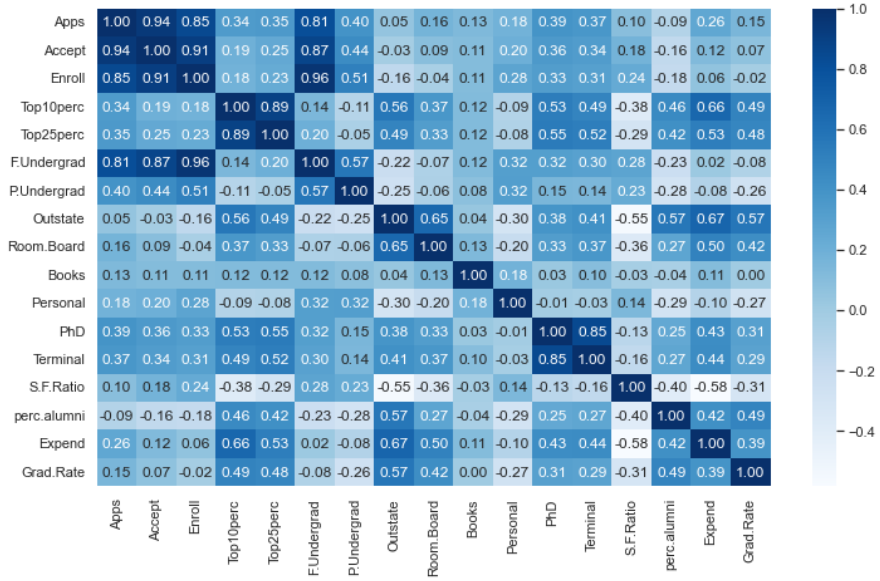


Fig 2.3 – Heat map (Correlation Plot)– Multivariate Analysis

From the above plot, it is evident that, the number of applications that are being received by a college is highly correlated to the number of applications getting shortlisted and number of actual students getting enrolled. It can also be established that, most of the students applying, getting shortlisted and getting enrolled are Ful time graduates.

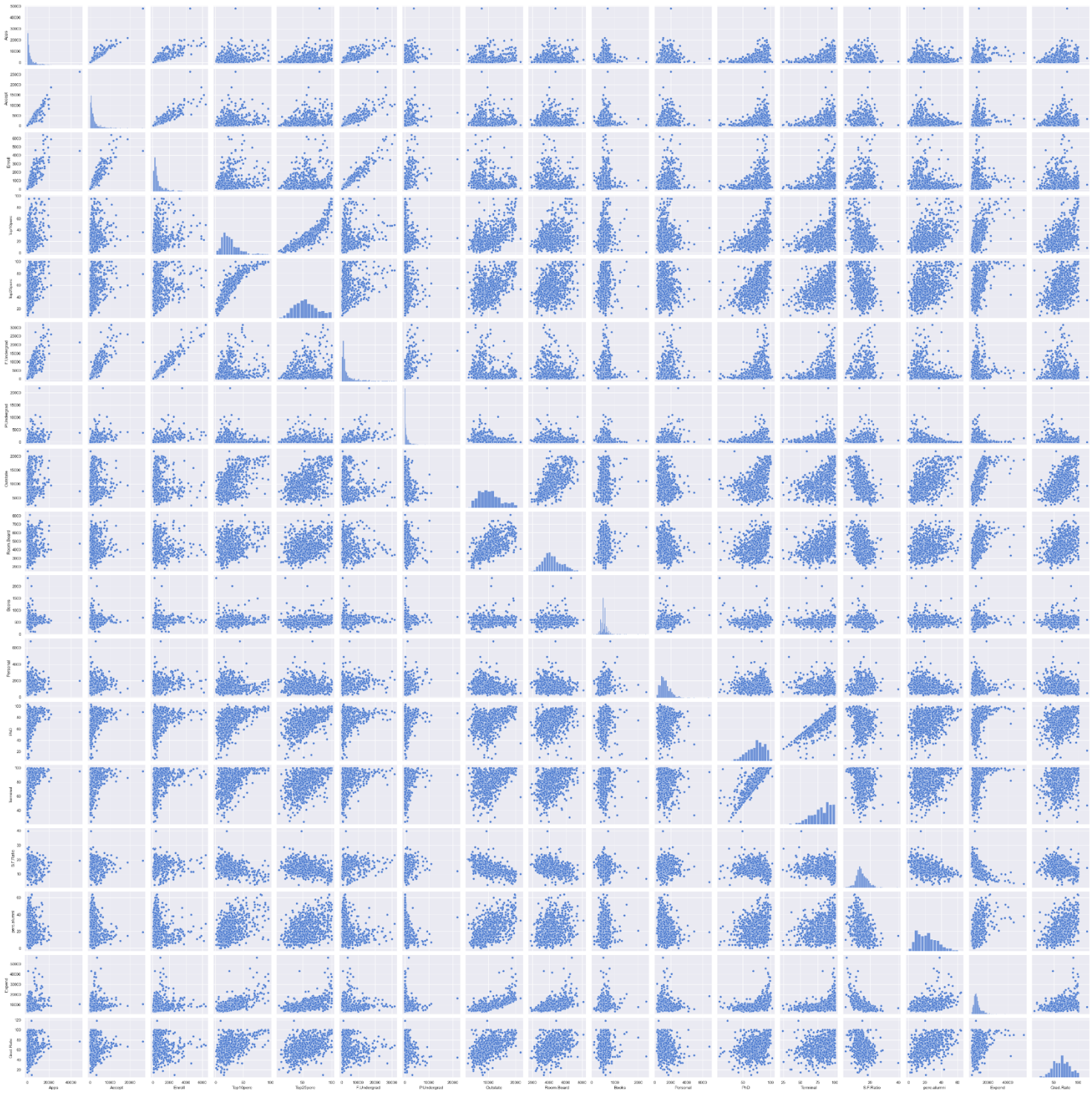


Fig 2.4 – Pair Plot – Multivariate Analysis

From the above pair plot, it can be seen that, the number of applications getting shortlisted and getting enrolled are following a linear trend with the total number of applications received. It can also be noticed that the number of PhD & Terminal education holders are increasing in number w.r.t the applications received. It can also be seen that the number of students enrolling for Part time is significantly low compared to Full time undergrad. Number of students from outstation are moderate irrespective of the number of students applying & enrolling. Cost for hostellers and Cost spent for books have almost remained the same irrespective of the number of students applying and getting enrolled. Cost of boarding has been a constant irrespective of number of students enrolling for both Full time and parttime undergrads.

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

Scaling is necessary for PCA in this case as there are variables with quantitative, cost, fractional (ratio) & percentage type. Scaling helps us bring all these variable types to a common magnitude (scale) i.e., shift the data points to origin, so that the machine learning algorithms becomes stays relevant and unbiased.

Before performing scaling operation, we need to drop the columns that are not relevant for PCA. In this dataset we shall drop the Names column and store it in a new data frame.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
1	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
2	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
3	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
4	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15

Table 2.3 – Sample of the new Dataset with Names column removed

One of the methods through which scaling can be done is through the Z-Score method.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	-0.35	-0.32	-0.06	-0.26	-0.19	-0.17	-0.21	-0.75	-0.96	-0.60	1.27	-0.16	-0.12	1.01	-0.87	-0.50	-0.32
1	-0.21	-0.04	-0.29	-0.66	-1.35	-0.21	0.24	0.46	1.91	1.22	0.24	-2.68	-3.38	-0.48	-0.54	0.17	-0.55
2	-0.41	-0.38	-0.48	-0.32	-0.29	-0.55	-0.50	0.20	-0.55	-0.91	-0.26	-1.20	-0.93	-0.30	0.59	-0.18	-0.67
3	-0.67	-0.68	-0.69	1.84	1.68	-0.66	-0.52	0.63	1.00	-0.60	-0.69	1.19	1.18	-1.62	1.15	1.79	-0.38
4	-0.73	-0.76	-0.78	-0.66	-0.60	-0.71	0.01	-0.72	-0.22	1.52	0.24	0.20	-0.52	-0.55	-1.68	0.24	-2.94

Table 2.4 – Scaled Dataset – Z-score method

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

Correlation is usually performed on the scaled data while covariance is calculated on the unscaled data. Therefore, we will be getting a significant change on performing PCA on unscaled data. However, on calculating covariance and correlation on scaled data, we are able to see that the coefficient values of covariance matrix are on a slightly higher side. The covariance matrix has a value of 1.001289 along the Leading diagonal, while correlation matrix has a value of 1 along the leading diagonal. Therefore this will impact the values of eigen vector and eigen values if PCA is performed. Correlation comes out to be the best option for PCA, as it has better accuracy when compared to correlation and removes the bias that is achieved through scaling.

	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

Table 2.5 – Correlation matrix of Scaled Data

	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 2.6 – Covariance matrix of Scaled Data

Q 2.4 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]

On visualizing with a box plot, we are able to note that scale of the data set has changed (notable at the y axis). There is no significant impact on the outliers as they are not treated. Hence to get a data free from outliers, we need to treat them for which there are several logics based on the dataset we are working with.

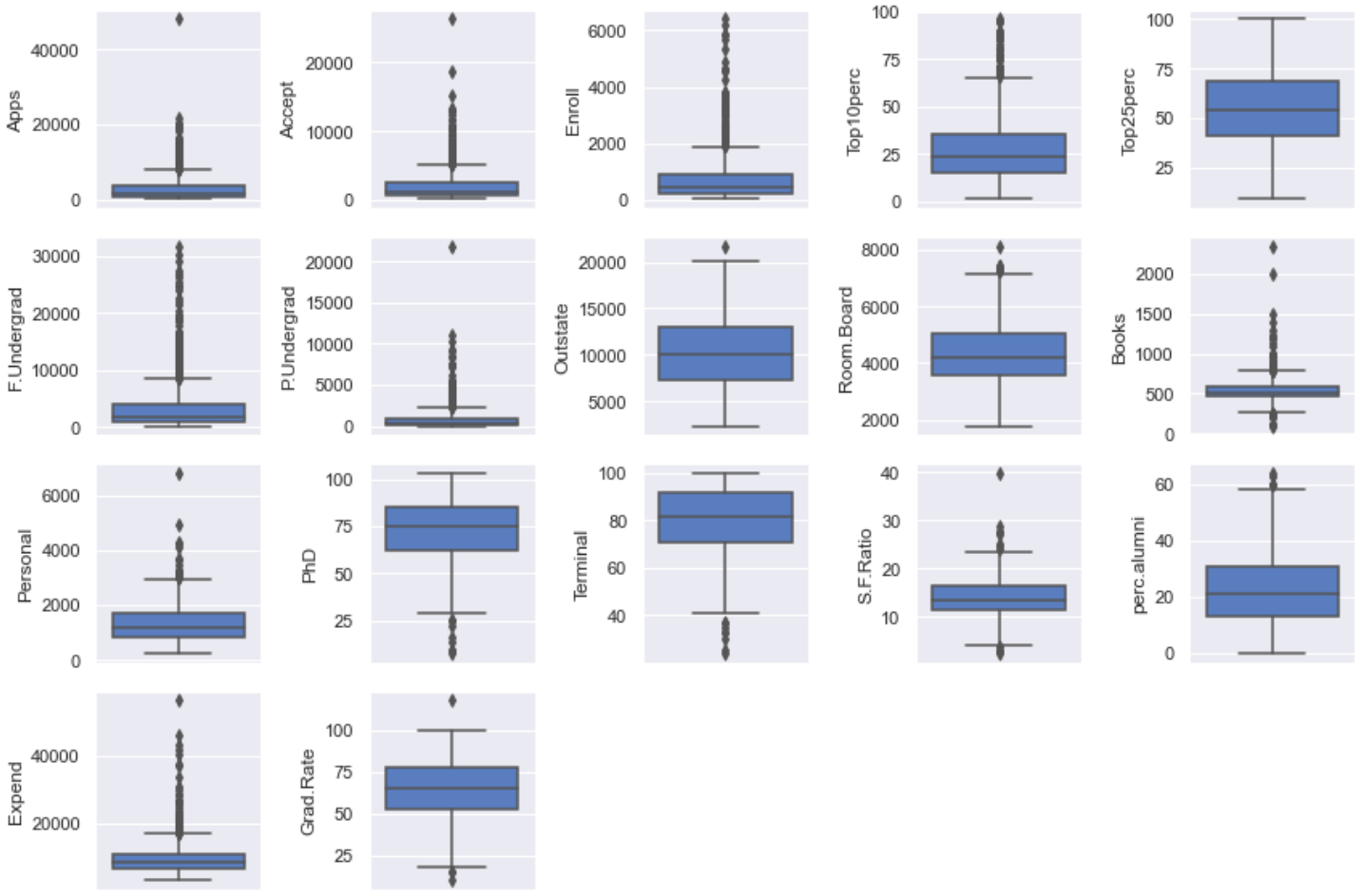


Fig 2.5 – Box plot – Before Scaling

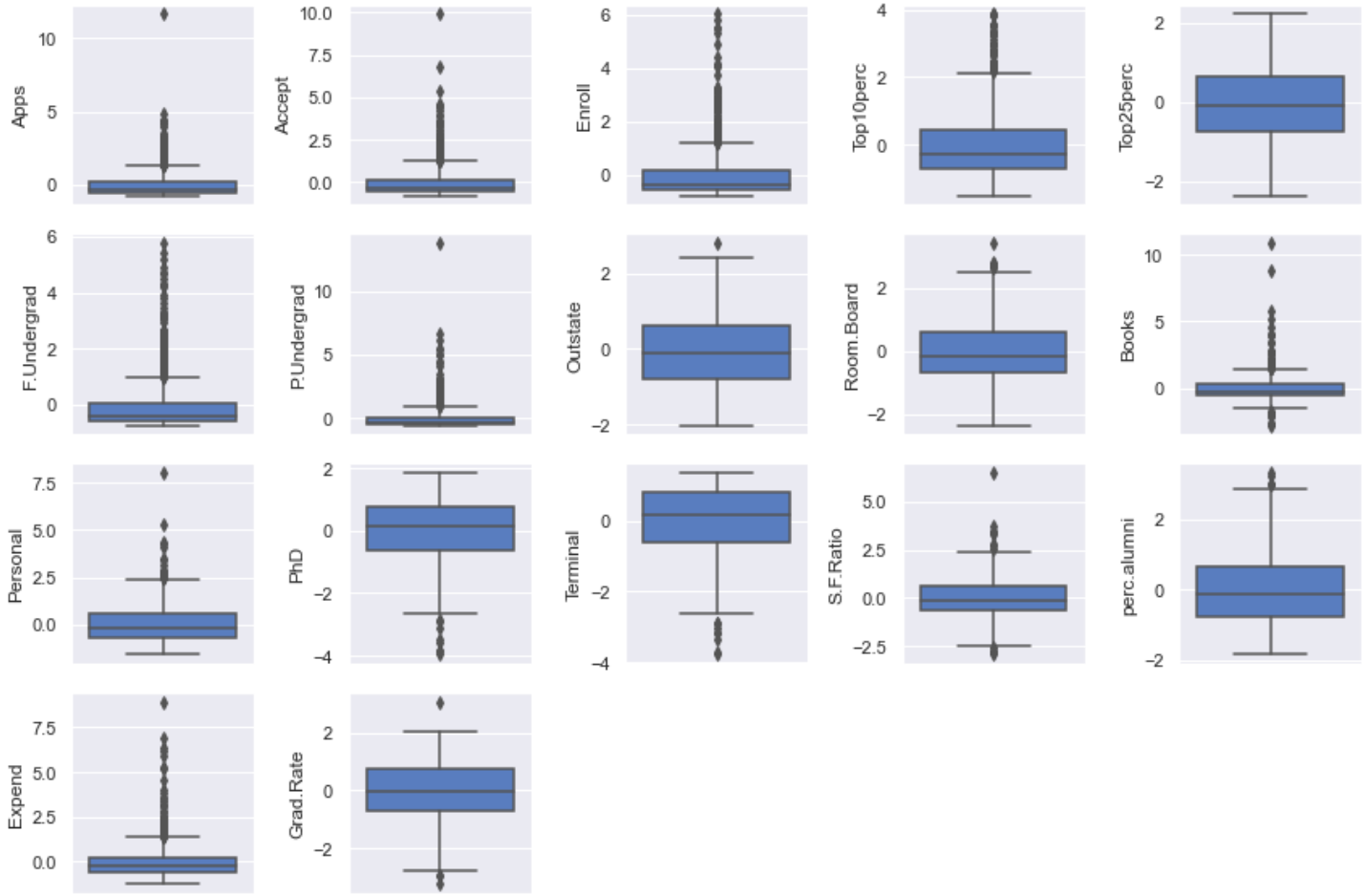


Fig 2.6 – Box plot – After Scaling

Q 2.5 Extract the eigenvalues and eigen vectors. [Using Sklearn PCA Print Both]

Eigen values capture the magnitude of variance in the data while Eigen vector are the coefficients/ loadings of the Principal components.

The Eigen values and Eigen vectors can be extracted after applying the PCA on the dataset. The PCA is imported from the sklearn.decomposition library. The code pca.explained_variance_ gives us the eigen values and pca.components_ gives us the eigen vectors.

Eigen Values array

```
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])
```

Eigen Vector array

```
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]])
```

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

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	0.2488	0.3316	-0.0631	0.2813	0.0057	-0.0162	-0.0425	-0.1031	-0.0902	0.0525	0.0430	0.0241	0.5958	0.0806	0.1334	0.4591	0.3590
Accept	0.2076	0.3721	-0.1012	0.2678	0.0558	0.0075	-0.0129	-0.0563	-0.1779	0.0411	-0.0584	-0.1451	0.2926	0.0335	-0.1455	-0.5186	-0.5434
Enroll	0.1763	0.4037	-0.0830	0.1618	-0.0557	-0.0426	-0.0277	0.0587	-0.1286	0.0345	-0.0694	0.0111	-0.4446	-0.0857	0.0296	-0.4043	0.6097
Top10perc	0.3543	-0.0824	0.0351	-0.0515	-0.3954	-0.0527	-0.1613	-0.1227	0.3411	0.0640	-0.0081	0.0386	0.0010	-0.1078	0.6977	-0.1487	-0.1450
Top25perc	0.3440	-0.0448	-0.0241	-0.1098	-0.4265	0.0331	-0.1185	-0.1025	0.4037	0.0145	-0.2731	-0.0894	0.0219	0.1517	-0.6173	0.0519	0.0803
F.Undergrad	0.1546	0.4177	-0.0614	0.1004	-0.0435	-0.0435	-0.0251	0.0789	-0.0594	0.0208	-0.0812	0.0562	-0.5236	-0.0564	0.0099	0.5604	-0.4147
P.Undergrad	0.0264	0.3151	0.1397	-0.1586	0.3024	-0.1912	0.0610	0.5708	0.5607	-0.2231	0.1007	-0.0635	0.1260	0.0193	0.0210	-0.0527	0.0090
Outstate	0.2947	-0.2496	0.0466	0.1313	0.2225	-0.0300	0.1085	0.0098	-0.0046	0.1867	0.1432	-0.8234	-0.1419	-0.0340	0.0384	0.1016	0.0509
Room.Board	0.2490	-0.1378	0.1490	0.1850	0.5609	0.1628	0.2097	-0.2215	0.2750	0.2983	-0.3593	0.3546	-0.0697	-0.0584	0.0034	-0.0259	0.0011
Books	0.0648	0.0563	0.6774	0.0871	-0.1273	0.6411	-0.1497	0.2133	-0.1337	-0.0820	0.0319	-0.0282	0.0114	-0.0668	-0.0094	0.0029	0.0008
Personal	-0.0425	0.2199	0.4997	-0.2307	-0.2223	-0.3314	0.6338	-0.2327	-0.0945	0.1360	-0.0186	-0.0393	0.0395	0.0275	-0.0031	-0.0129	-0.0011
PhD	0.3183	0.0583	-0.1270	-0.5347	0.1402	0.0913	-0.0011	-0.0770	-0.1852	-0.1235	0.0404	0.0232	0.1277	-0.6911	-0.1121	0.0298	0.0138
Terminal	0.3171	0.0464	-0.0660	-0.5194	0.2047	0.1549	-0.0285	-0.0122	-0.2549	-0.0886	-0.0590	0.0165	-0.0583	0.6710	0.1589	-0.0271	0.0062
S.F.Ratio	-0.1770	0.2467	-0.2898	-0.1612	-0.0794	0.4870	0.2193	-0.0836	0.2745	0.4720	0.4450	-0.0110	-0.0177	0.0414	-0.0209	-0.0212	-0.0022
perc.alumni	0.2051	-0.2466	-0.1470	0.0173	-0.2163	-0.0473	0.2433	0.6785	-0.2553	0.4230	-0.1307	0.1827	0.1041	-0.0272	-0.0084	0.0033	-0.0192
Expend	0.3189	-0.1317	0.2267	0.0793	0.0760	-0.2981	-0.2266	-0.0542	-0.0491	0.1323	0.6921	0.3260	-0.0937	0.0731	-0.2277	-0.0439	-0.0353
Grad.Rate	0.2523	-0.1692	-0.2081	0.2691	-0.1093	0.2162	0.5599	-0.0053	0.0419	-0.5903	0.2198	0.1221	-0.0692	0.0365	-0.0034	-0.0050	-0.0131

Table 2.7 – Data Frame containing Principal Components

The above table shows all the Principal Components in a data frame against all the features.

Q 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]

Any Principal Component is a Linear equation in the form

PC1 = A1X1 + A2X2 + A3X3+…+AnXn ..... Eq. 2.1 – Linear Equation form of PC

where, A1, A2, A3, …, An are the Eigen vectors of the corresponding Feature/ Dimensions and X1, X2, X3, …, Xn are their respective features.

In this case the linear equation of the PC1 is as follows;

PC1 = 0.25 Apps + 0.21 Accept + 0.18 Enroll + 0.35 Top10perc + 0.34 Top25perc + 0.15 F.Undergrad + 0.29 Outstate + 0.25 Room.Board + 0.06 Books – 0.04 Personal + 0.32 PhD + 0.32Terminal – 0.18 S.F.Ratio + 0.21 percalumni + 0.32 expend + 0.25 Grad.Rate ..... Eq. 2.2 – PC1 equation

Q 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?

Cumulative values of the Eigenvalues help us to decide the optimum number of Principal Components. This is because the are a measure of relative variance in the data that is captured in the principal components. By default, the Eigen values when calculated in Python will be given in the descending order of their magnitude. Therefore, the First PC will have the maximum variance captured. Cumulatively calculating the Eigen values helps us identify up to which PC the how much percentage of variance of the dataset has been captured. Most of the times, considering 85% of the variance captured gives us accurate results. This is how we decide upon the optimum number of the principal components. In python we will be able to visualize this using a scree plot and the cumulative percentage can also be calculated for better understanding.

In this case, the cumulative sum of variance in the data captured is as in the below scree plot and their corresponding values as in the below table.

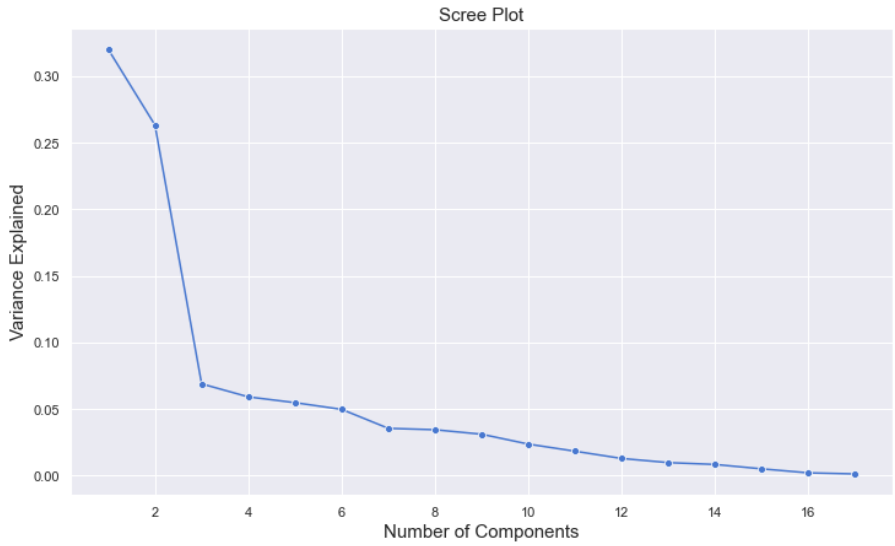


Fig 2.7 – Scree Plot – Capturing Cumulative values Eigen values

Principal Component	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Cumulative Sum of Eigen Values	0.32	0.584	0.653	0.712	0.767	0.817	0.852	0.887	0.918	0.942	0.96	0.973	0.983	0.991	0.996	0.999	1

Table 2.8 – Cumulative values of Eigen values.

From the above plot and values, 85% of the data are captured in the first seven principal components. Hence, it would be optimum to consider up to seven Principal components.

Eigen vectors are the coefficients/ loadings of the principal components that can be seen in the data frame table 2.5 against each Principal Component.

Q 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]

From Q 2.8, we are able to identify that 85% of the data are captured in the first 7 principal components. To proceed further, we must ensure that these 7 PCs are void of correlation and are independent of each other. The same can be witnessed when PCA is performed only for 7 components again and the plot is as below.

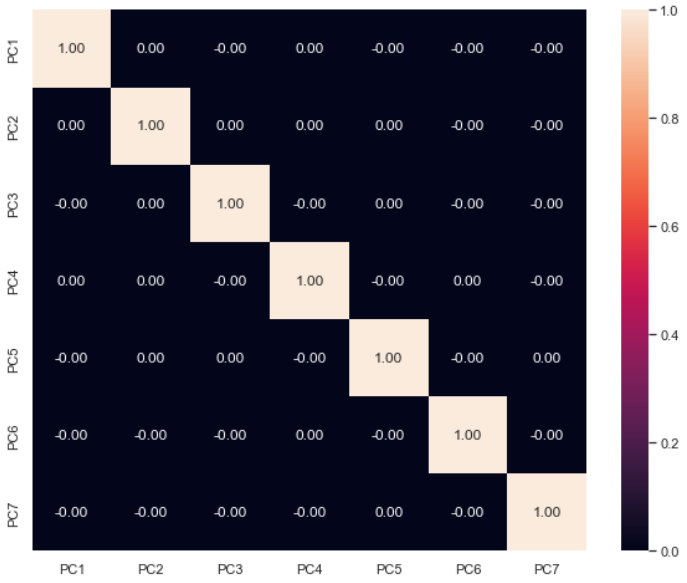


Fig 2.8 – Heat map– Selected PCs Correlation

We are also aware that PC1 captures majority of the data. Hence using the absolute values, we will be able to identify which feature is the prime factor that needs to be targeted to achieve a good business result. The same can be analyzed using a heat map as below.

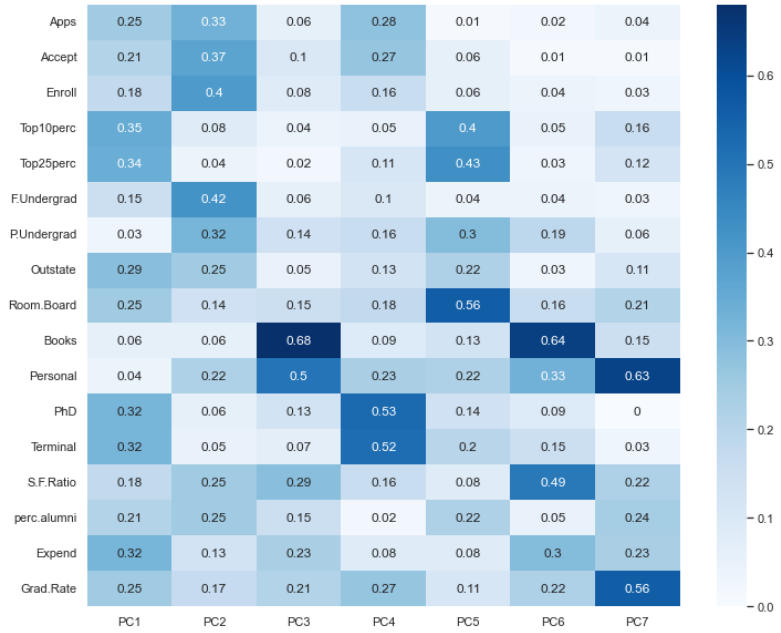


Fig 2.9 – Heat map– Analysis of PCs

From the above heatmap Fig 2.8, we are able to note that the colleges have higher percentage of faculties with PhD/ Terminal degree. It is also evident that the expenditure towards institutional charges is significant. It can also be seen that top 10 and 25 % scorers have joined the colleges after 12th standard. Hence, these insights sets up the target consumers for the business.

We can see that there are also other significant factors in other PCs. These also play a major role. But the PC1 is prioritized as it covers majority of the data.

Through PCA we have reduced the dimensions from 17 to 7, thus making it easier for analysis.

END