ADVANCED STATISTICS

PROJECT REPORT

BY

M.P.KARTHIKEYAN

Contents Problem

Problem 1A

Problem 1B:

Problem 2

- List of Figures

List of Tables

Problem 1A:

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.

Executive summary

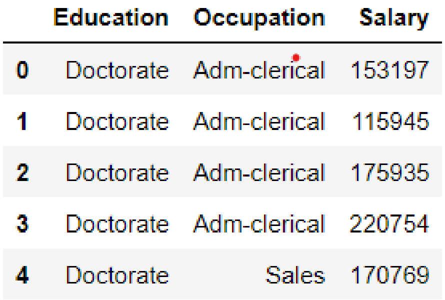
Educational Qualification and Occupational details of 40 salaried individuals are collected to determine it’s impact on the individual’s salary. Educational Qualification has three levels such as High school graduate, Bachelor, and Doctorate. Occupation is at four levels such as Administrative and clerical, Sales, Professional or specialty, and Executive or managerial.

Introduction

The given dataset has details of 40 salaried individuals. Exploratory Data Analysis is done. To determine the dependency of salary on educational and occupational level ANOVA TEST is performed.Both One-way and Two-way ANOVA test was done.

Sample Dataset

Table1.1 Sample 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

Exploratory Data Analysis

Let us check the type of variables

```
Education    object
Occupation   object
Salary       int64
```

The dataset contains 40 rows and 3 columns.Out of 3 columns 2 columns are Object type and 1 column is integer type.

Check for missing values in dataset

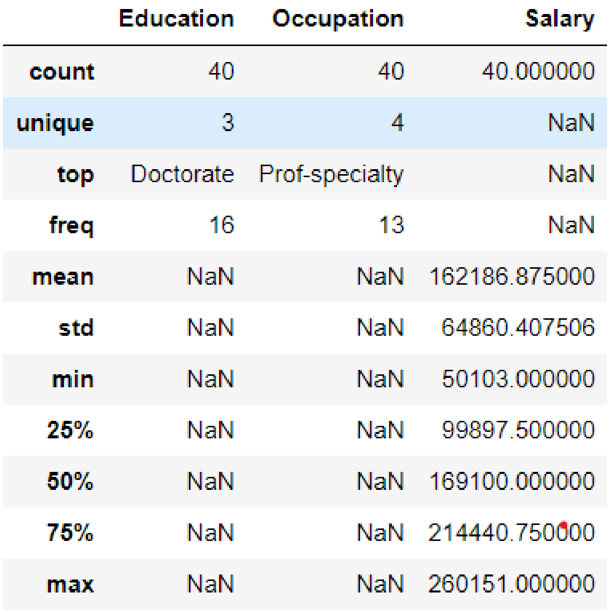
```
Education    40 non-null
Occupation   40 non-null
Salary       40 non-null
```

From the above values it is clear that there are no missing values in dataset.

Descriptive Statistics

Descriptive statistics are used to describe about the variables present in the dataset by giving a short summaries about the sample and the measures of data.
The most recognized types of descriptive statistics are measures of centre**: the mean, median, and mode**, which are used at almost all levels of math and statistics.

Table 1.2-Summary of Dataset



	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

From the above table we found that the **salary range** is found to be between **50103** to **260151.** Out of **40** employees **16** employees have completed **Doctorate** and **13** out of **40** employees are working as **Prof-speciality** making them as most common education level and occupation level respectively of the dataset.

NaN Values are present in some variables as the measures of centre can’t be calculated.

Calculating Salary of different Education & Occupation levels

Education	**Salary**
1.Bachelors	165152.933333
2.Doctorate	208427.000000
3.HS-grad	75038.777778

Occupation	**Salary**
1.Adm-clerical	141424.300000
2.Exec-managerial	197117.600000
3.Prof-specialty	168953.153846
4.Sales	157604.416667

Fig-1.1 Histogram of salary distribution

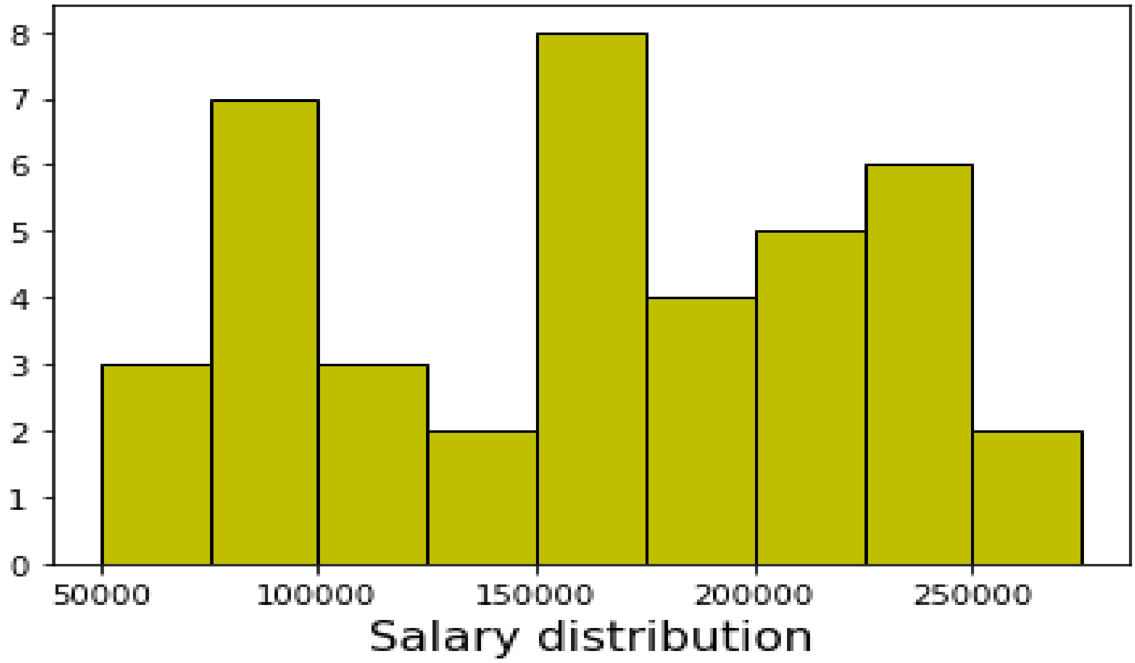
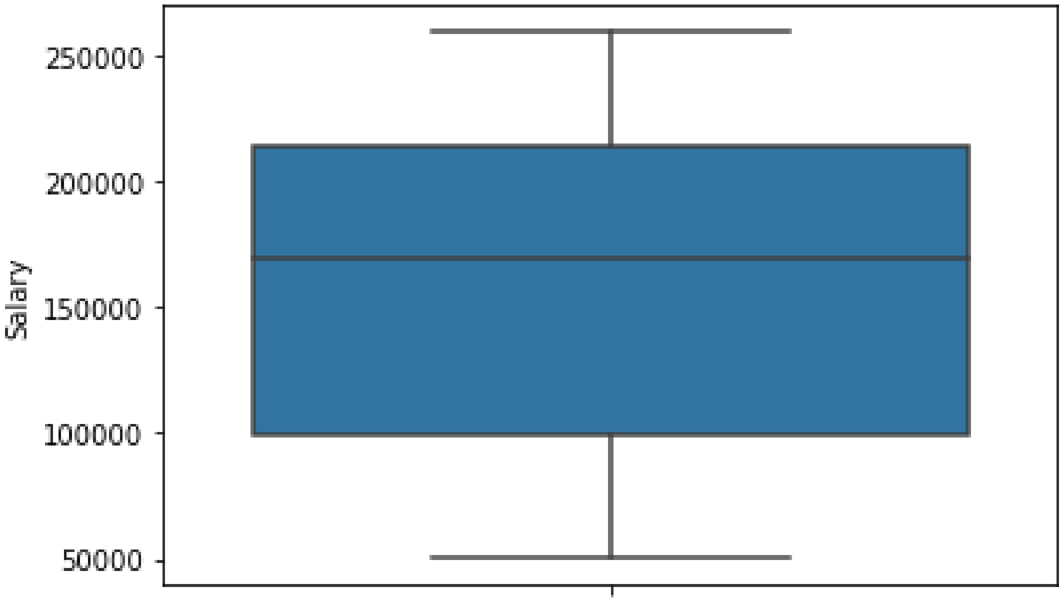


Fig-1.2 Boxplot of Salary Distribution



ANOVA TEST

The ANOVA (Analysis of Variance) technique can be used when it is needed to compare more than two population means.This technique also establish the causation of why the means are behaving in a particular manner.There are two types of ANOVA such as One-Way Anova and Two-way Anova.

Assumptions of ANOVA

The following assumptions are for anova test

1.The samples drawn are independent and random

2.The response variables of population are continuous & normally distributed

3.The variance of all the populations are equal at least approximately

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

Hypothesis of one-way ANOVA for Education

Let **H0** be **Null hypothesis** & **Ha** be **Alternate hypothesis**

$\boldsymbol{H0}$: $\mu 1 = \mu 2 = \mu 3$

$\boldsymbol{Ha}$: At least one Salary level is different from the rest.

Where
$\mu 1$, $\mu 2$, $\mu 3$ represent the population mean salary of 3 different education levels such as Bachelors, Doctorate & HS-grad.

Hypothesis of one-way ANOVA for Occupation

Let **H0** be **Null hypothesis** & **Ha** be **Alternate hypothesis**

$\boldsymbol{H0}$: $\mu 1 = \mu 2 = \mu 3 = \mu 4$

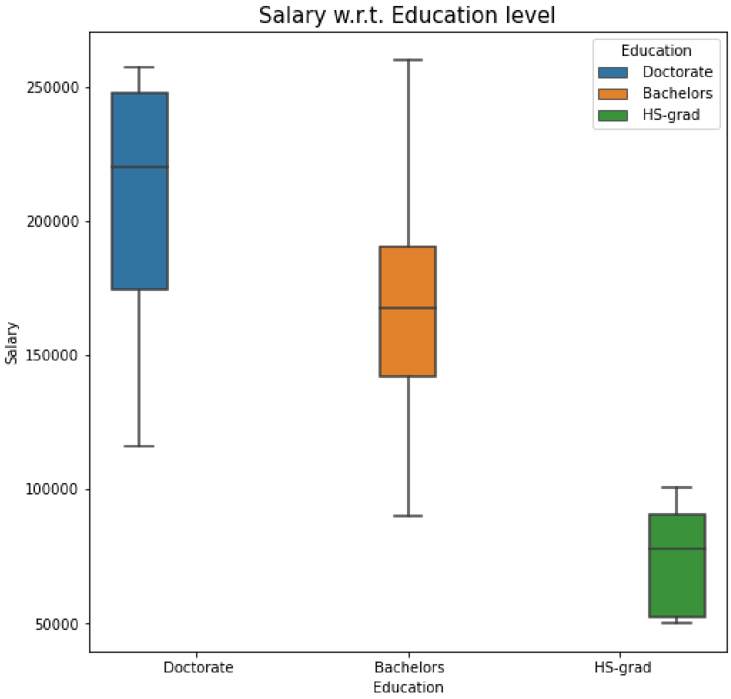
$\boldsymbol{Ha}$: At least one Salary level is different from the rest.

Where
$\mu 1$, $\mu 2$, $\mu 3$, $\mu 4$ represent the population mean salary of 4 different Occupation levels such as Administrative and clerical, Sales, Professional or specialty, and Executive or managerial.

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.

Check for Outliers

Fig-1.3 Boxplot on salary w.r.t Education level



The above plot shows us there is no outliers present in dataset hence ANOVA test can be performed **.**

One-way ANOVA on Salary w.r.t Education

Table 1.3 One way ANOVA on Salary w.r.t Education

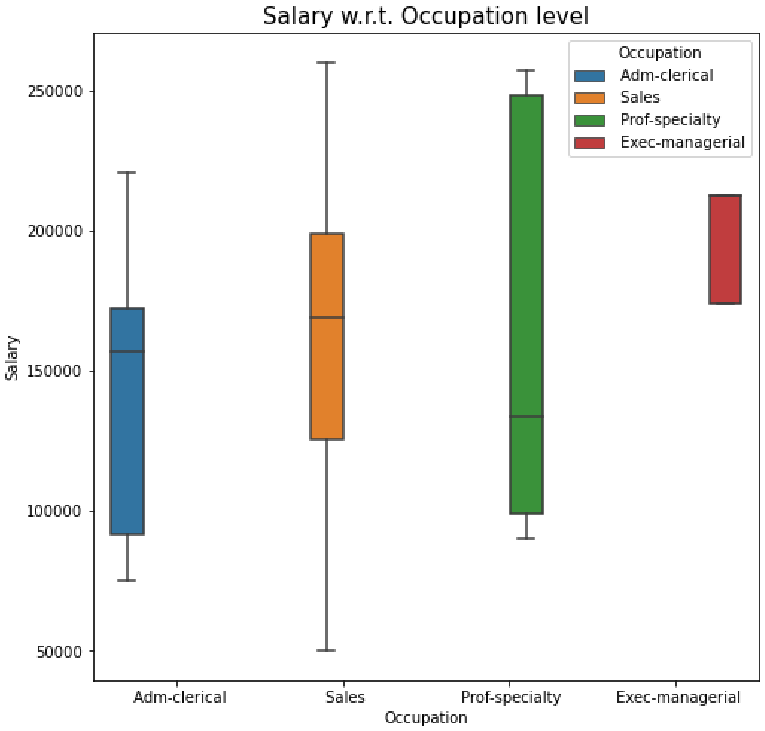
	df	sum_sq	mean_sq	F	PR(>F)
Education	2	1.03e+11	5.13e+10	30.9563	1.26e-08
Residual	37	6.14e+10	1.66e+09	NaN	NaN

From the above ANOVA table since **p value = 1.26e-08** which is less than the significance level (al**pha = 0.05**) we can **reject the null hypothesis** and conclude that atleast one Salary level is different from the rest based on education.

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.

Check for Outliers

Fig-1.4 Boxplot on salary w.r.t Occupation level



The above plot shows us there is no outliers present in dataset hence ANOVA test can be performed.

One-way ANOVA on Salary w.r.t Occupation

Table 1.4 One way ANOVA on Salary w.r.t Occupation

	df	sum_sq	mean_sq	F	PR(>F)
Occupation	3	1.125878e+10	3.752928e+09	0.884144	0.458508
Residual	36	1.528092e+11	4.244701e+09	NaN	NaN

From the above ANOVA table since **p value = 0.458508** which is greater than the significance level (al**pha = 0.05**) we **fail to reject the null hypothesis** and conclude that there is no significant difference in population mean salary of 4 different Occupation levels .

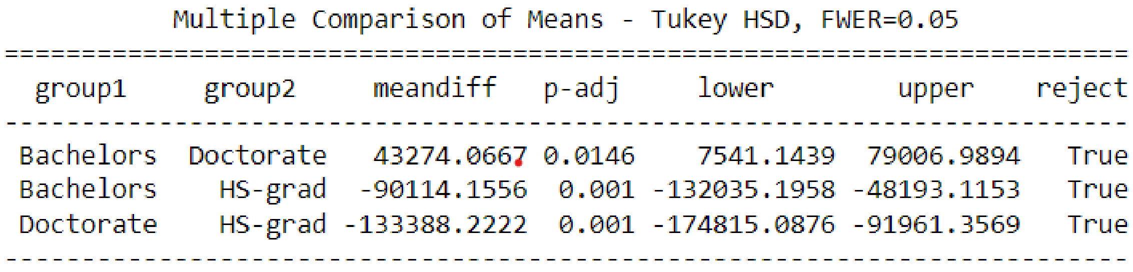
4.If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. (Non-Graded)

The null hypothesis got rejected in (2) one-way ANOVA on Salary with respect to Education.
To find out which class means are significantly different Multiple (pair-wise) comparisons using Tukey's HSD can be performed
the Tukey Honest Significant Difference test,

Hypothesis For Tukey's HSD
Null Hypothesis $H0$: All pairs of group means are equal against

Alternate Hypothesis Ha: At least one groupmean is different from the rest.

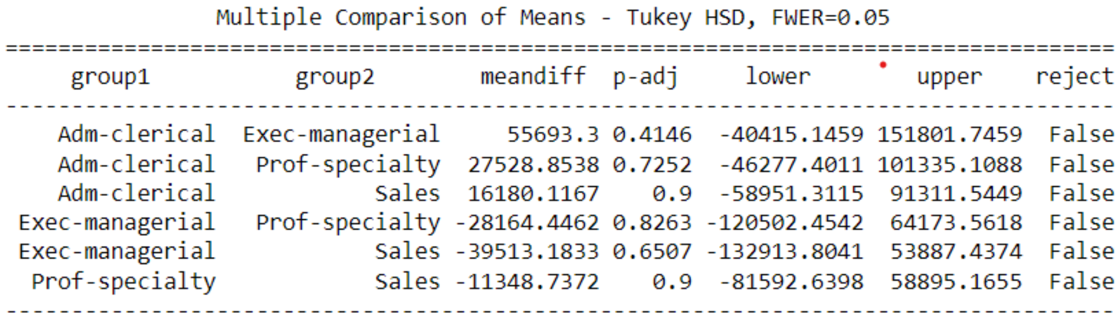
Table-1.5 Multiple comparison of Means of Education-Tukey HSD



```
          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
-------------------------------------------------------------------------
```

From the above table we find p- values(p-adj) are lesser than the significance level(0.05) for all the three categories of education, this implies that the mean salaries across all categories of education are different**.**

Table-1.6 Multiple comparison of Means of Occupation-Tukey HSD



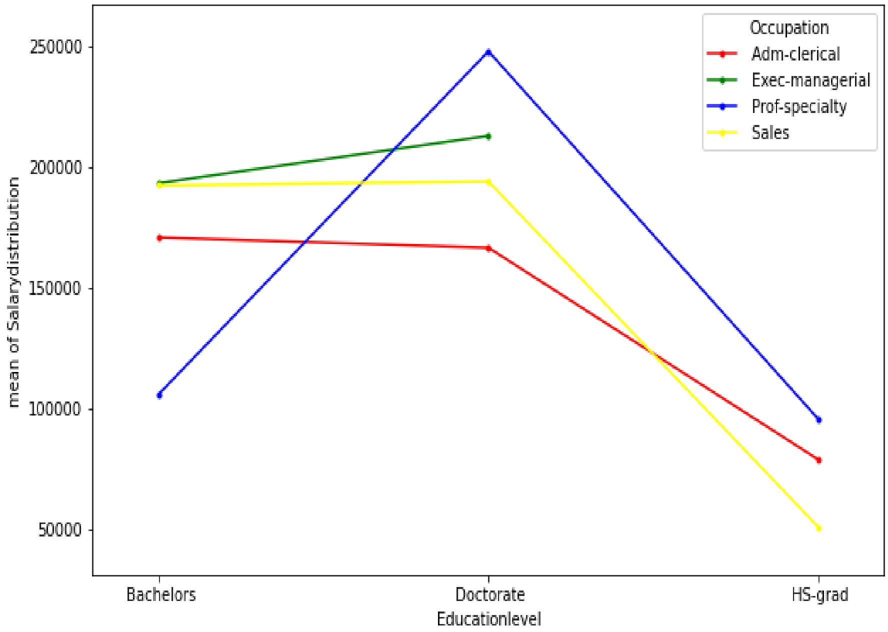
```
                  Multiple Comparison of Means - Tukey HSD, FWER=0.05
==================================================================================
     group1          group2        meandiff  p-adj      lower        upper    reject
----------------------------------------------------------------------------------
   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
----------------------------------------------------------------------------------
```

Here (p-adj > alpha) ,we fail to reject the null hypothesis thus we conclude all pairs of group means are equal .

Problem 1B:

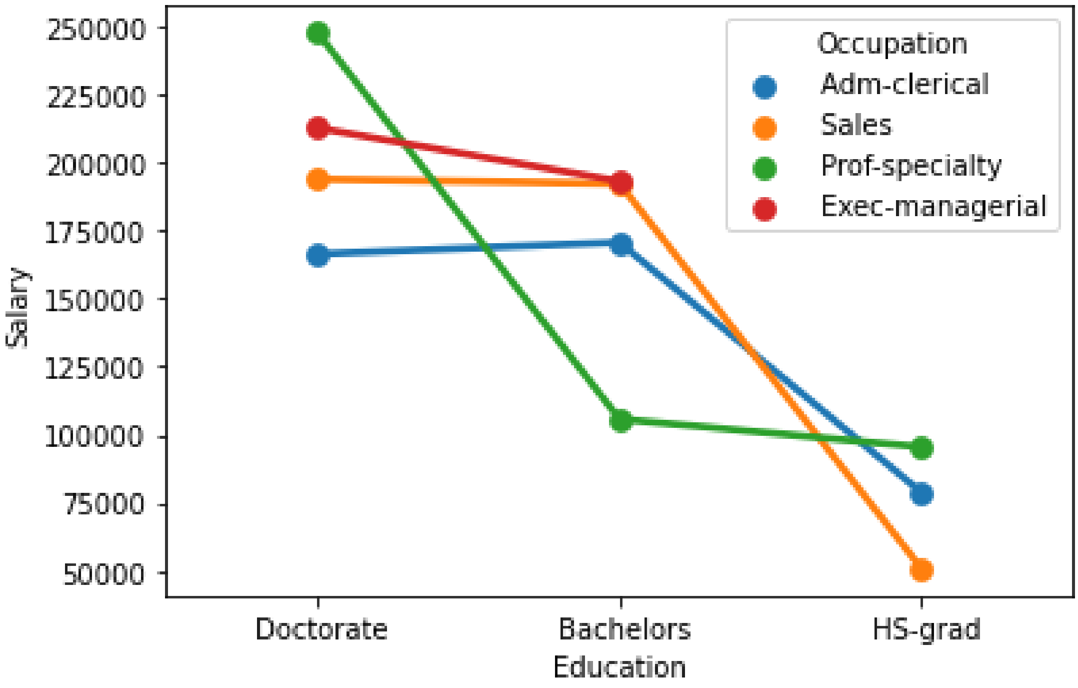
1.What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the ‘pointplot’ function from the ‘seaborn’ function]

Fig-1.5 Interaction Plot



The above interaction plot shows that there is significant amount of interaction between the categorical variables, Education and Occupation

Fig-1.6 Point Plot of interaction



From the above plot we conclude the following observations:

.People with Bachelors or Doctorate as education and Adm-clerical and Sales as occupation almost earn the same salaries .

.People with Hs-grad education and sales as occupation earns less than Adm-clerical with Hs-grad education.

· Exec-managerial position is offered to people with Doctorate and Bachelors Education and not to people with Hs-grad.

· Prof-Specialty people with education as Doctorate earn maximum salaries

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

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

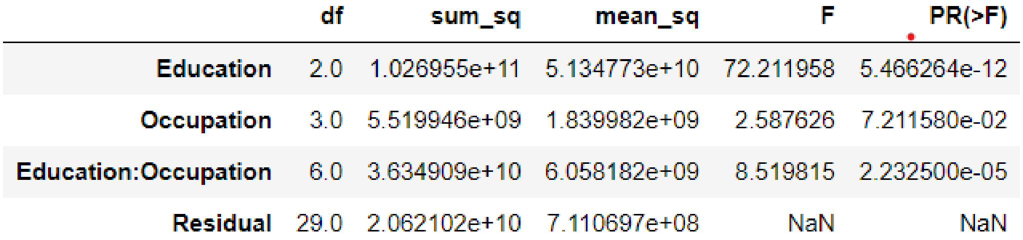
2.Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?

*H*0:There is no interaction effect between the 2 independent variables, education and occupation).

*H*1: There is an interaction effect between the variableS ‘education’ and ‘occupation’ on the mean Salary.

By performing two way ANOVA, we get the following table:

Table 1.7 Two way ANOVA on Salary w.r.t Occupation



	df	sum_sq	mean_sq	F	PR(>F)
Education	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
Occupation	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
Education: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

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

From the table, we see that there is a significant amount of interaction between the variables, Education and Occupation.

Thus, we see that there is an interaction effect between education and occupation on the mean salary.

The education combined with occupation results in higher and better salaries among the people.People with education as Doctorate draw the maximum salaries and people with education HS-grad earn the least. Thus, we can conclude that Salary is dependent on educational qualifications and occupation.

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

ANOVA stands for “analysis of variance” and is used in statistics when you are testing a hypothesis to understand how different groups respond to each other by making connections between independent and dependent variables. ANOVA is a statistical test that compares the means of groups in order to determine if there is a difference between them. Here the given dataset has Educational Qualification and Occupational details of 40 salaried individuals.From the results of ANOVA tests we see that there is an interaction effect between education and occupation on the mean salary.The education combined with occupation results in higher and better salaries among the people.People with education as Doctorate draw the maximum salaries and people with education HS-grad earn the least. Thus, we can conclude that Salary is dependent on educational qualifications and occupation.

Problem 2:

The dataset Education-Post 12 th 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.xslx

Executive summary

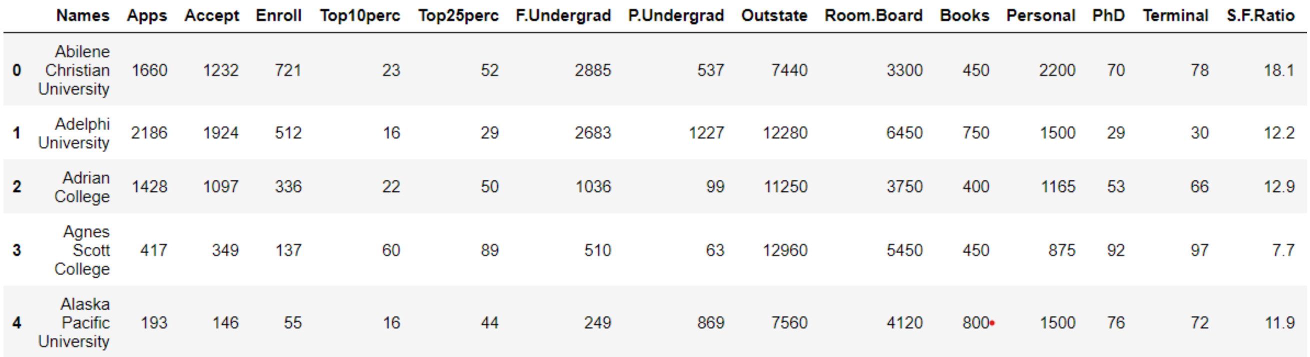
The dataset has information about 777 Colleges/Universities such as the applications received, details about the programmes enrolled by the students,expense for students towards room,board and books.The qualification of Faculties and student/faculty ratio for the institutions ,Graduation rate of institutions are provided.Exploratory Data Analysis and PCA are to be performed on the dataset.

Introduction

The given dataset has datas collected regarding 777 colleges/Universities .EDA and PCA are performed on the dataset.The business implications of the PCA are analysed.

Sample Dataset

Table 2.1 Sample Dataset



	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio
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

Exploratory Data Analysis

Let us check the type of variables

```
 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
```

The dataset contains 777 rows and 18 columns.Out of 18 columns 1 column is Object type, 16 columns are integer type and 1 column is float type.

Check for missing values in dataset

```
Names          777 non-null   object
Apps           777 non-null   int64
Accept         777 non-null   int64
Enroll         777 non-null   int64
Top10perc      777 non-null   int64
Top25perc      777 non-null   int64
F.Undergrad    777 non-null   int64
P.Undergrad    777 non-null   int64
Outstate       777 non-null   int64
Room.Board     777 non-null   int64
Books          777 non-null   int64
Personal       777 non-null   int64
PhD            777 non-null   int64
Terminal       777 non-null   int64
S.F.Ratio      777 non-null   float64
perc.alumni    777 non-null   int64
Expend         777 non-null   int64
Grad.Rate      777 non-null   int64
```

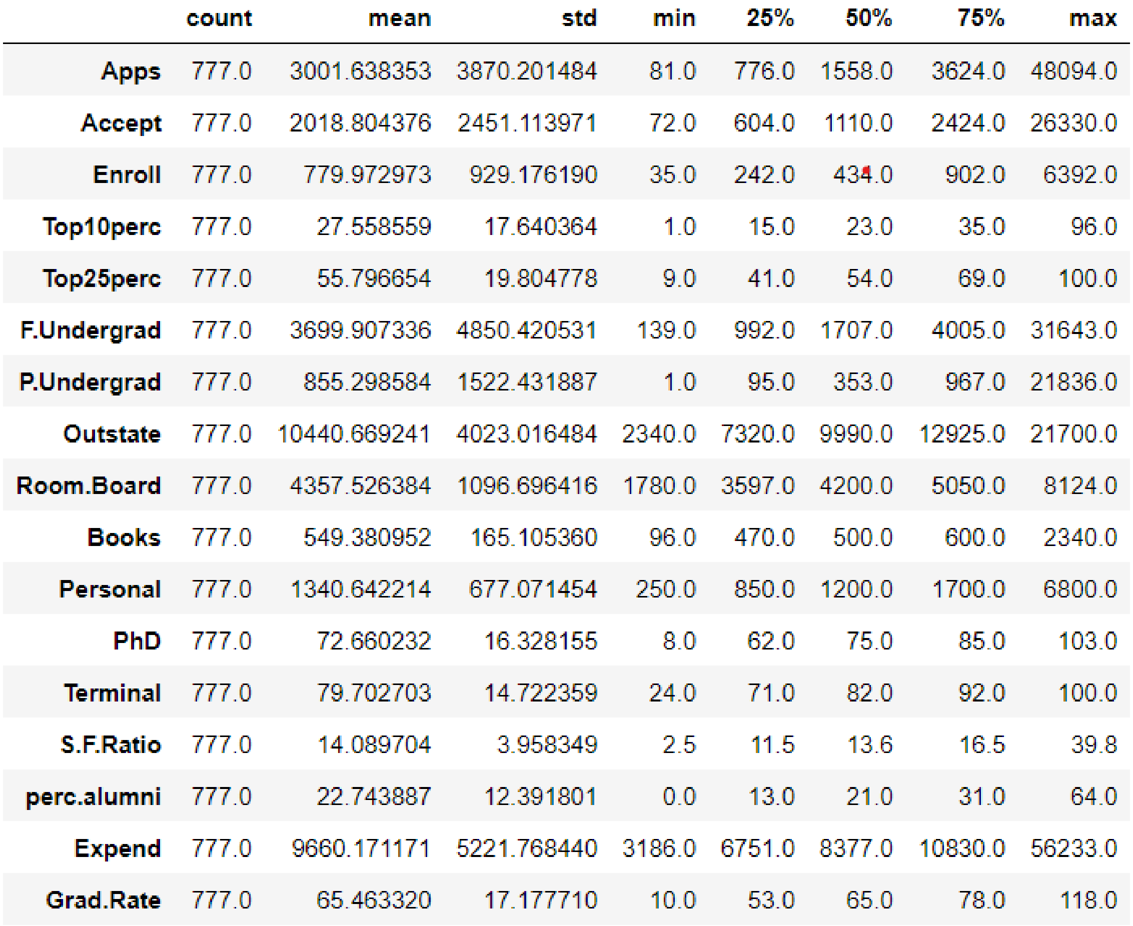
From the above values it is clear that there are no missing values in dataset.

Descriptive Statistics

Descriptive statistics are used to describe about the variables in dataset by giving short summaries about the sample and the measures of data.

The most recognized types of descriptive statistics are measures of centre**: the mean, median, and mode**, which are used at almost all levels of math and statistics.

Table 2.2-Summary of Dataset



	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

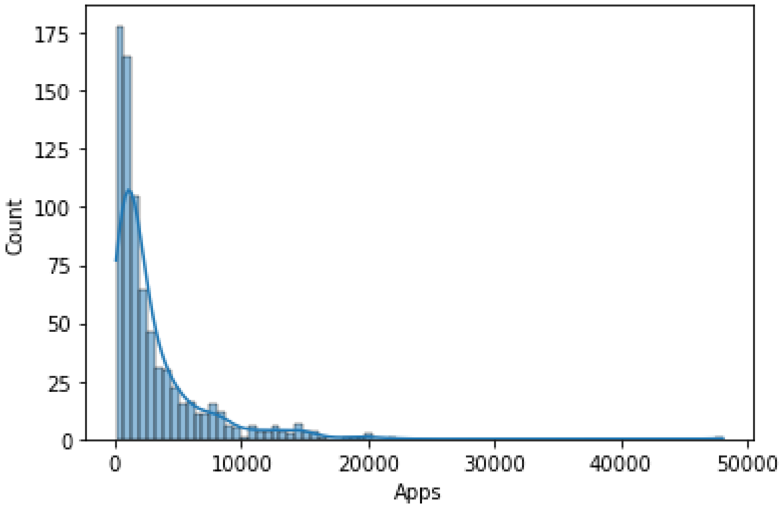
From the above table we can see that on an average most of the colleges receive **3001.63** applications the least number of applications received by a college was 81 and the highest number of applications received by a college was 48094.After selection the average number of students enrolled stands at **779.97**.The enrollment ranges from as low as **35** students to as high as **6392.** The number of students pursuing full time undergraduate course is higher than the number of students pursuing part time undergraduate course.The cost of room and board ranges between **1780** to **8124**. Estimated book costs for a student will be from **96** to **2340** and the average personal expense will be **1340.64**.On an average **72.66%** of faculties have PhD and **79.7%** of faculties have Terminal degree.The average Student/faculty ratio is **14.089.**The average percent of alumni donating to colleges is **22.74%.**The Overall average Graduation rate is **65.46.**

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

Univariate Analysis

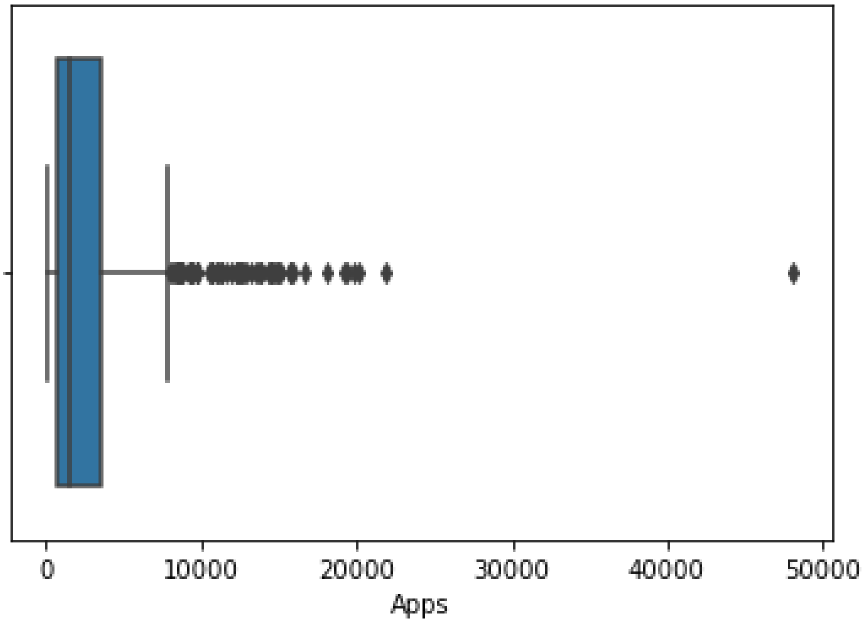
1.Apps

Fig-2.1 Distribution of Apps



The distribution of the data is skewed.we can see that on an average most of the colleges receive **3001.63** applications.The maximum number of applications is around 50000.

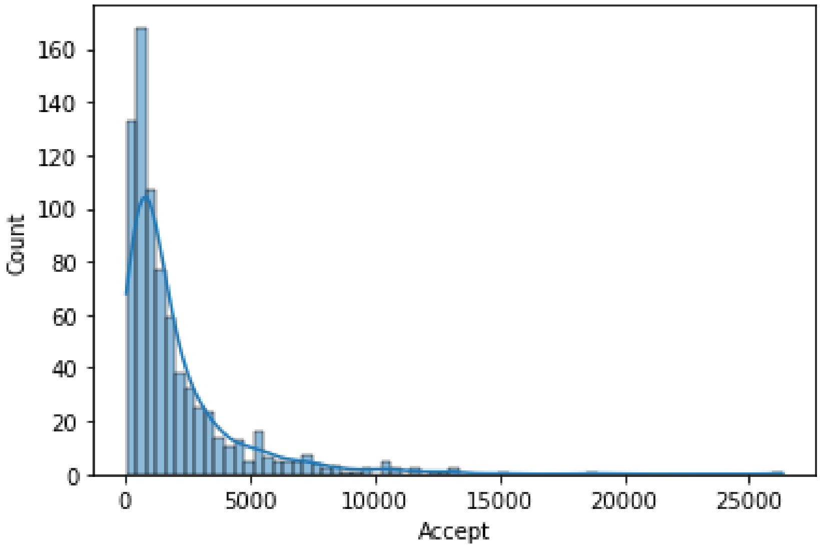
Fig-2.2 Boxplot on Distribution of Apps



From the Boxplot we can see the presence of outliers in the dataset.

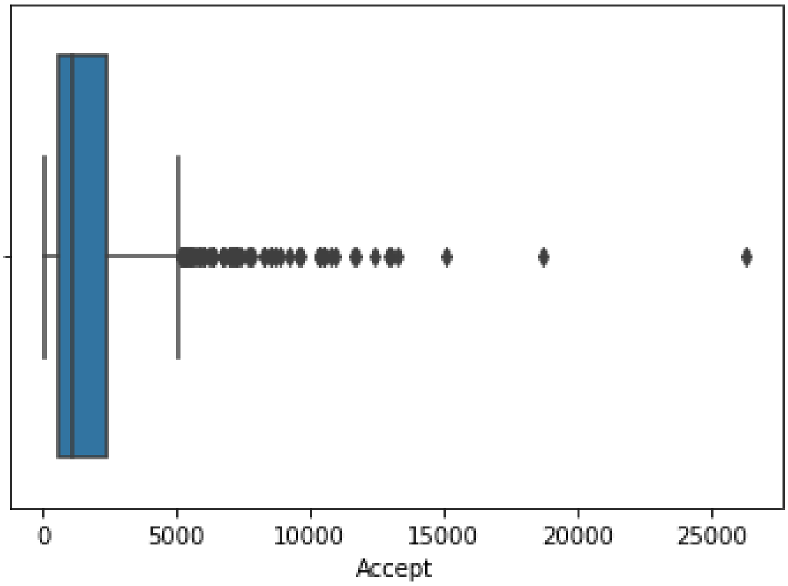
2.Accept

Fig-2.3 Distribution of Accept



The distribution of the data is skewed.we can see that the average no. of applications accepted by the colleges is **2018.80** .The maximum number of applications accepted is around 26000.

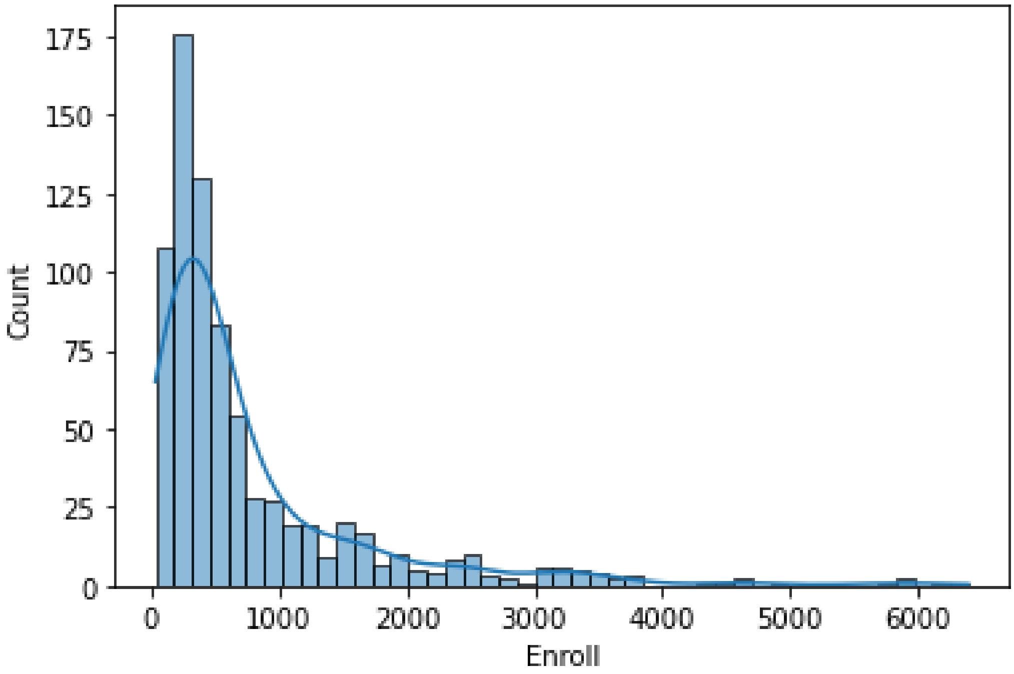
Fig-2.4 Boxplot on Distribution of Accept



From the Boxplot we can see the presence of outliers in the dataset.

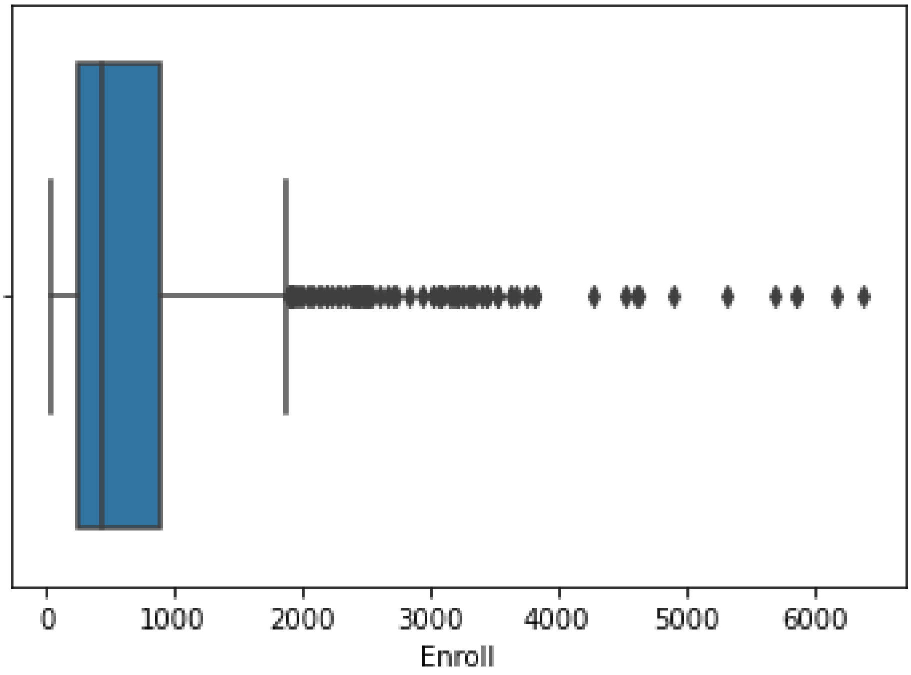
3.Enroll

Fig-2.5 Distribution of Enroll



The distribution of the data is positively skewed. we can see that the average no. of students enrolled for the colleges is 779.97 .The maximum number of enrolment is above 6000.

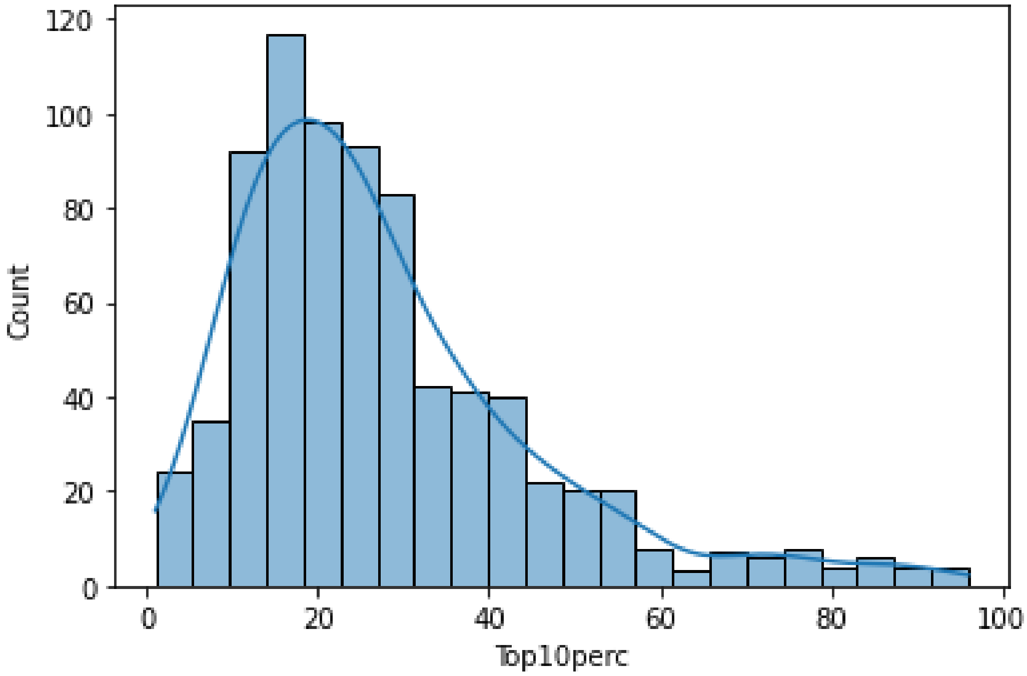
Fig-2.6 Boxplot on Distribution of Enroll



From the Boxplot we can see the presence of outliers in the dataset.

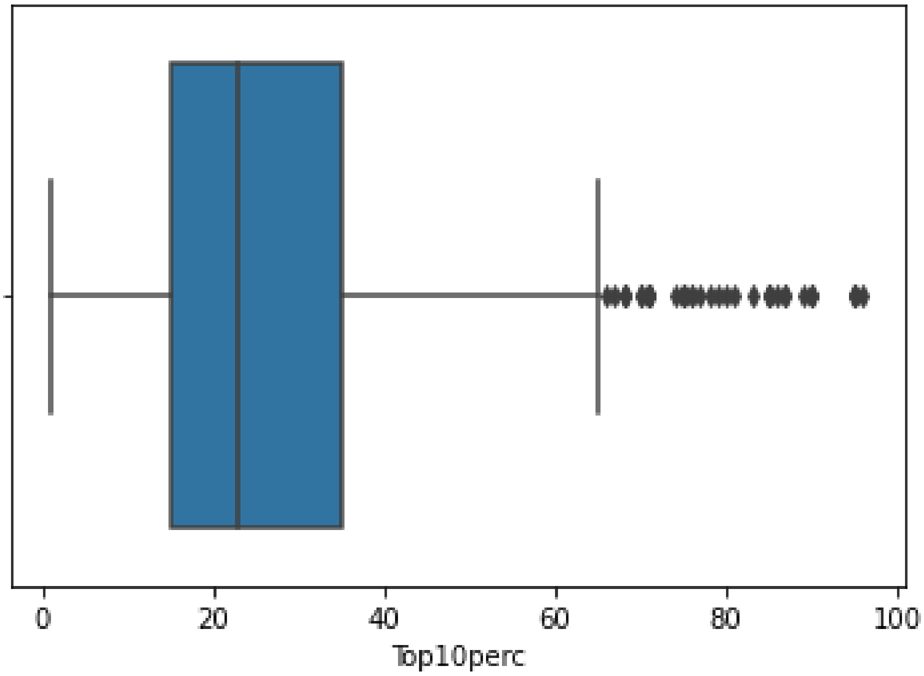
4. Top10perc

Fig-2.7 Distribution of Top10perc



The distribution of the data is positively skewed.The average percent of students from top 10% of Higher Secondary class enrolled for the colleges is 27.55 .The maximum percent of students from top 10% of Higher Secondary class joining a particular institute is around 100.

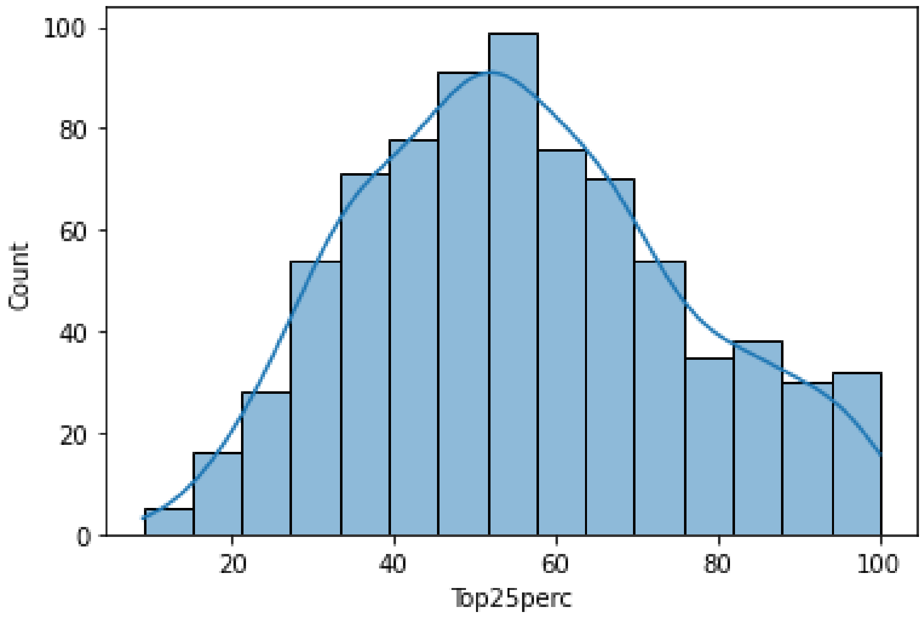
Fig-2.8 Boxplot on Top10perc



From the Boxplot we can see the presence of outliers in the dataset.

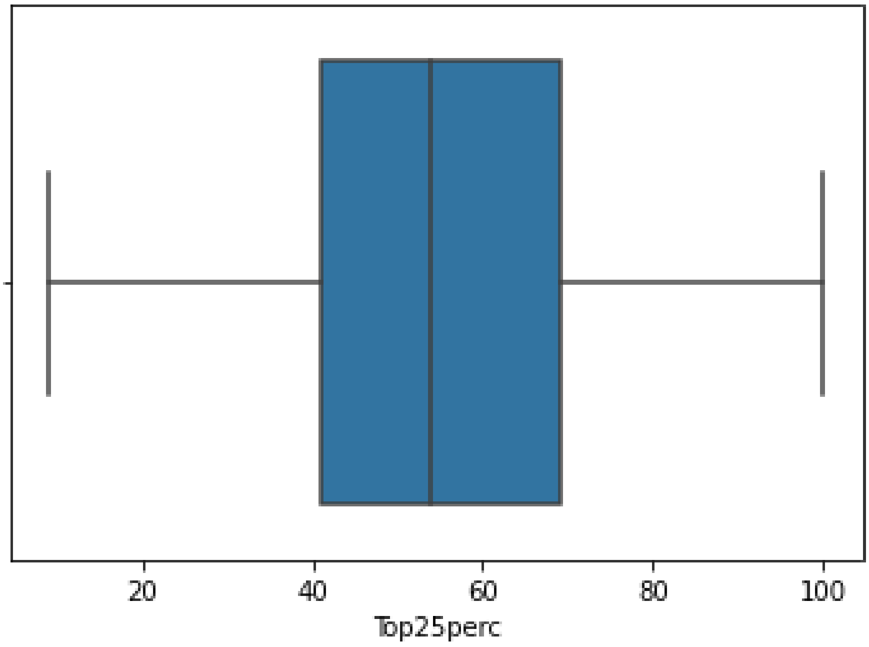
5. Top25perc

Fig-2.9 Distribution of Top25perc



The distribution of the data is almost normal.The average percent of students from top 25% of Higher Secondary class enrolled for the colleges is 55.79 .The maximum percent of students from top 25% of Higher Secondary class joining a particular institute is around 100.

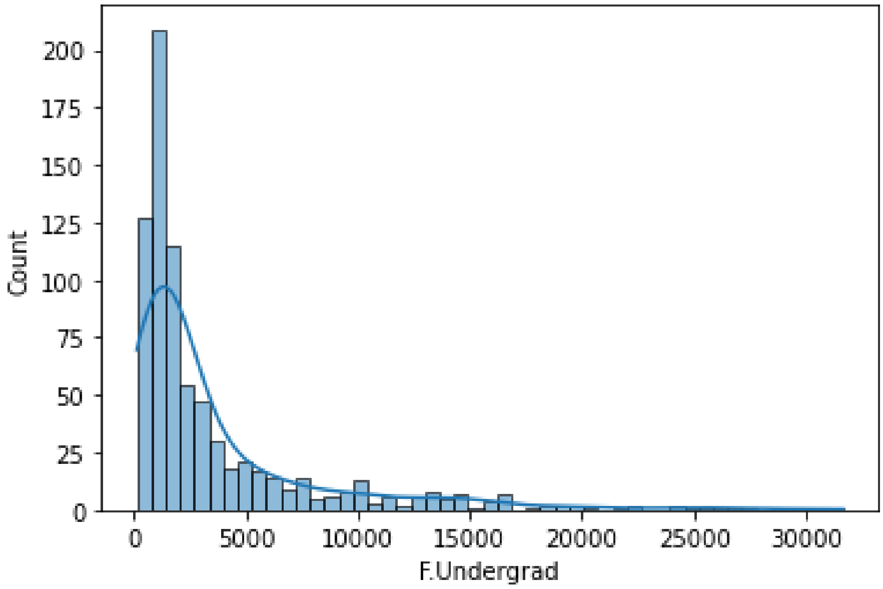
Fig-2.10 Boxplot on Top25perc



From the Boxplot we can see there is no presence of outliers in the dataset.

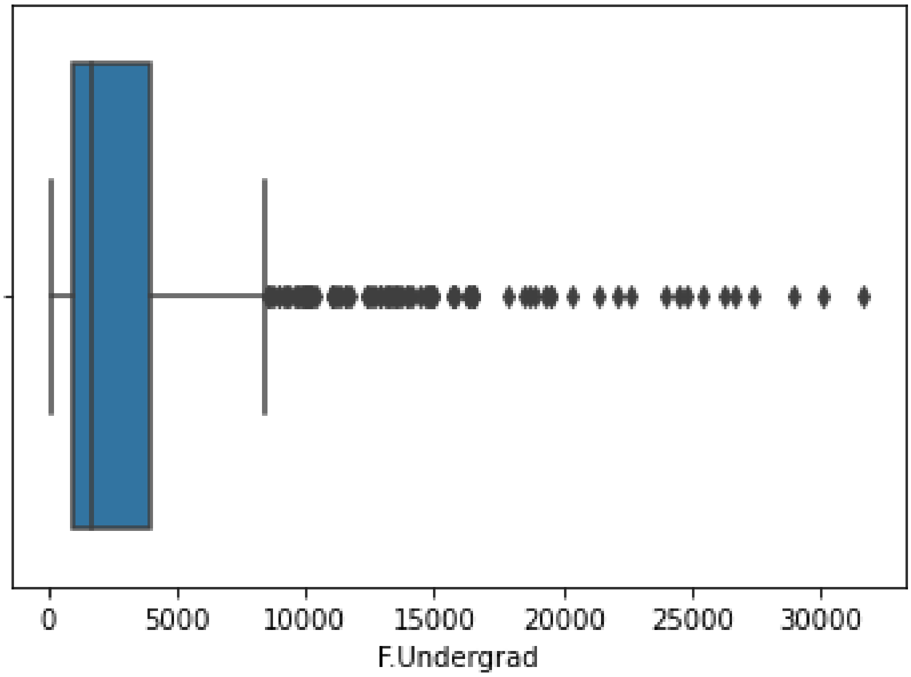
6. F.Undergrad

Fig-2.11 Distribution of   F.Undergrad



The distribution of the data is positive skewed.The average number of students enrolled for full time Undergraduate course is 3699.90 .The maximum number of students enrolled for full time Undergraduate course is 31643.

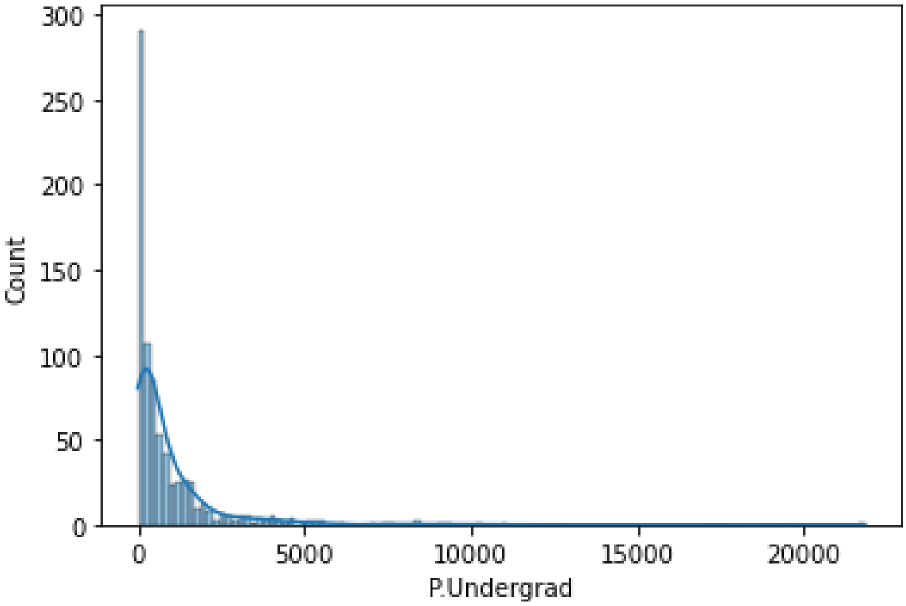
Fig-2.12 Boxplot on F.Undergrad



From the Boxplot we can see there is presence of outliers in the dataset.

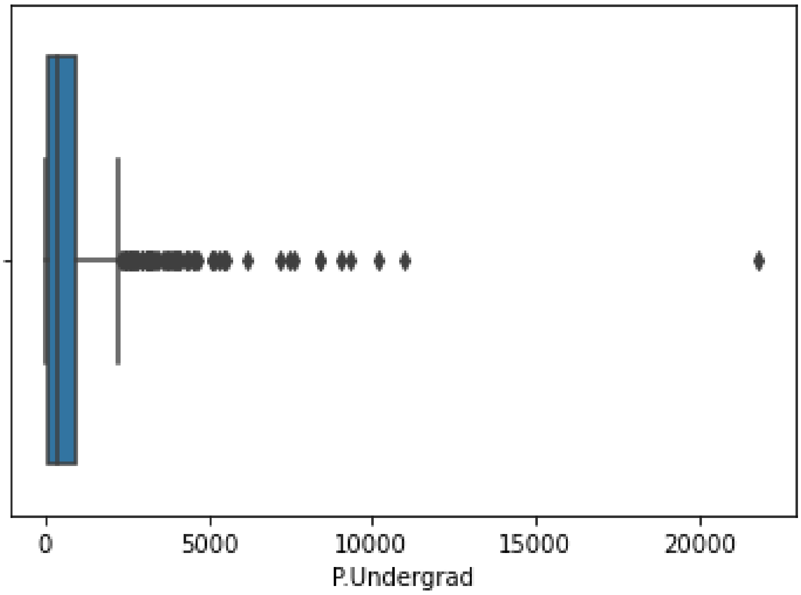
7. P.Undergrad

Fig-2.13 Distribution of   P.Undergrad



The distribution of the data is highly positive skewed.The average number of students enrolled for part time Undergraduate course is 855.29 .The maximum number of students enrolled for part time Undergraduate course is 21836.

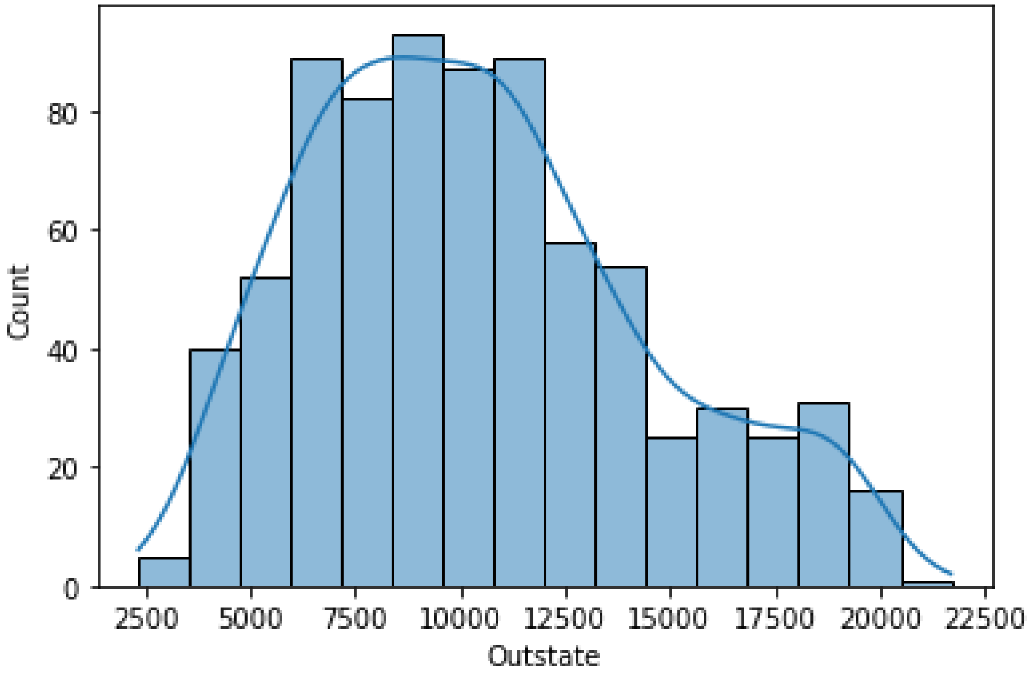
Fig-2.14 Boxplot on P.Undergrad



From the Boxplot we can see there is presence of outliers in the dataset.

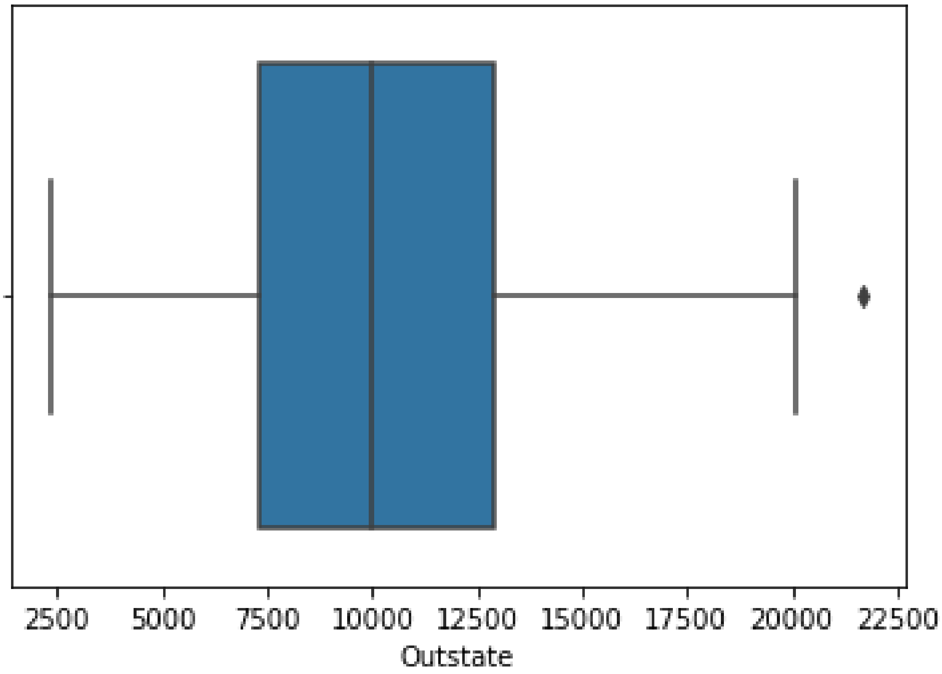
8. Outstate

Fig-2.15 Distribution of Outstate



The distribution of the data is nearly normal.

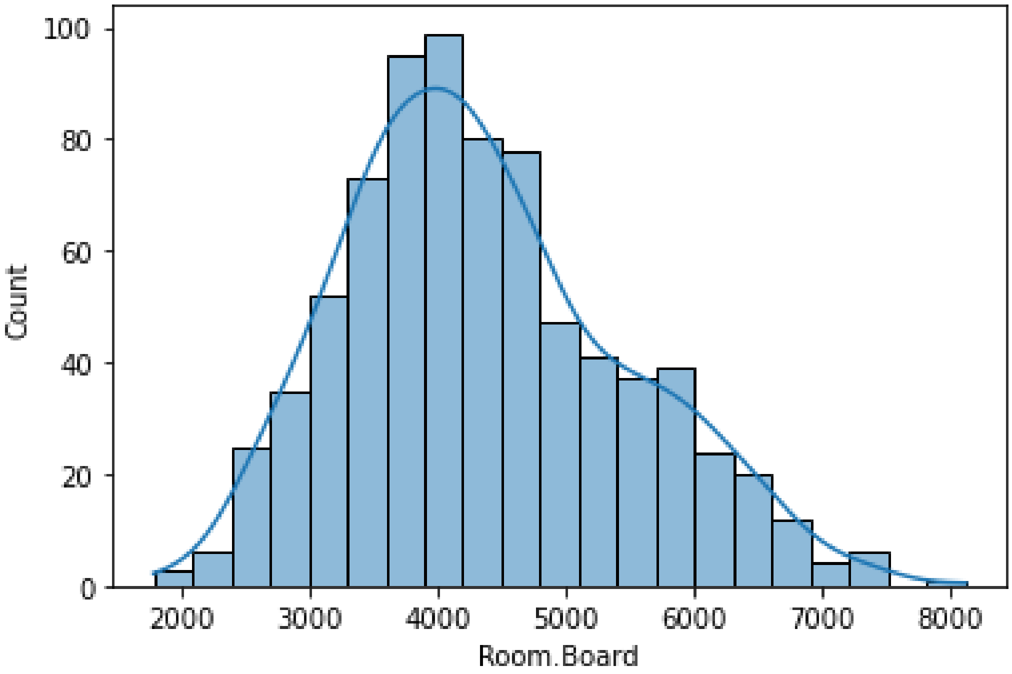
Fig-2.16 Boxplot on Outstate



From the Boxplot we can see there is presence of one outlier in the dataset.

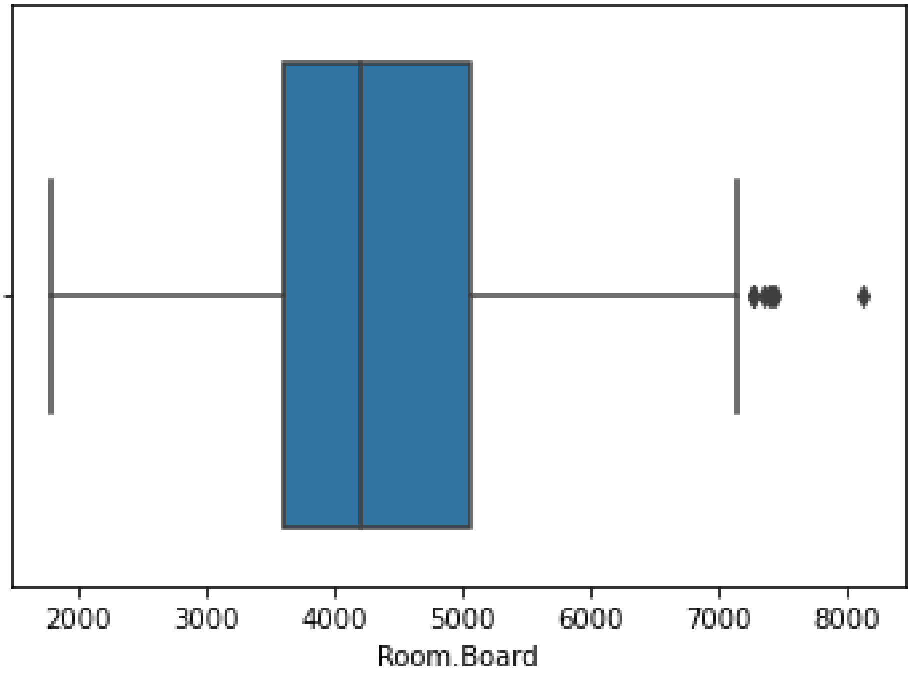
9.Room.Board

Fig-2.17 Distribution of Room.Board



The distribution of the data is nearly normal.

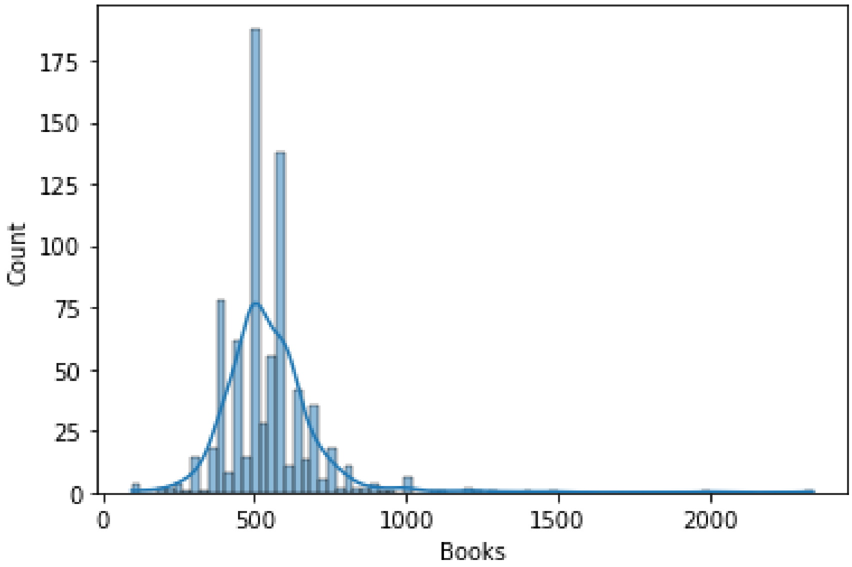
Fig-2.18 Boxplot on Room.Board



From the Boxplot we can see there is presence of outliers in the dataset.

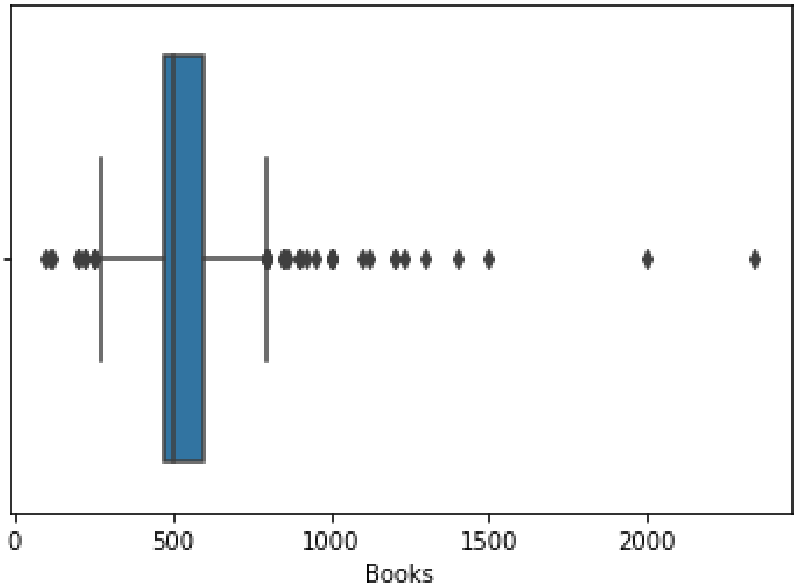
10.Books

Fig-2.19 Distribution of Books



The distribution of the data is positively skewed. Estimated book costs for a student will be from **96** to **2340**

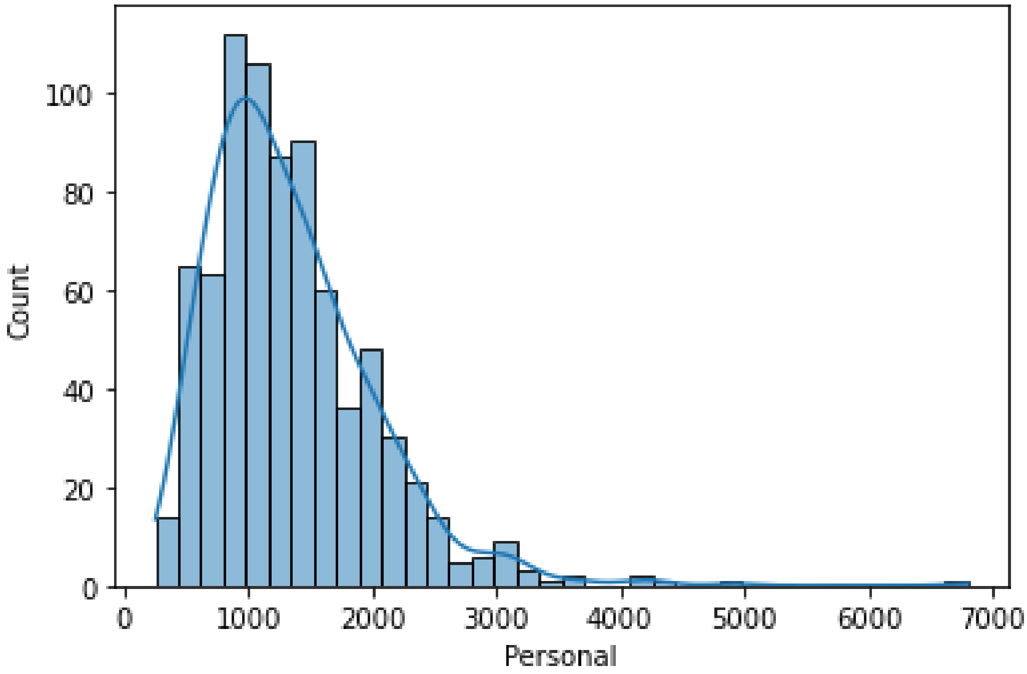
Fig-2.20 Boxplot on Books



From the Boxplot we can see there is presence of outliers in the dataset.

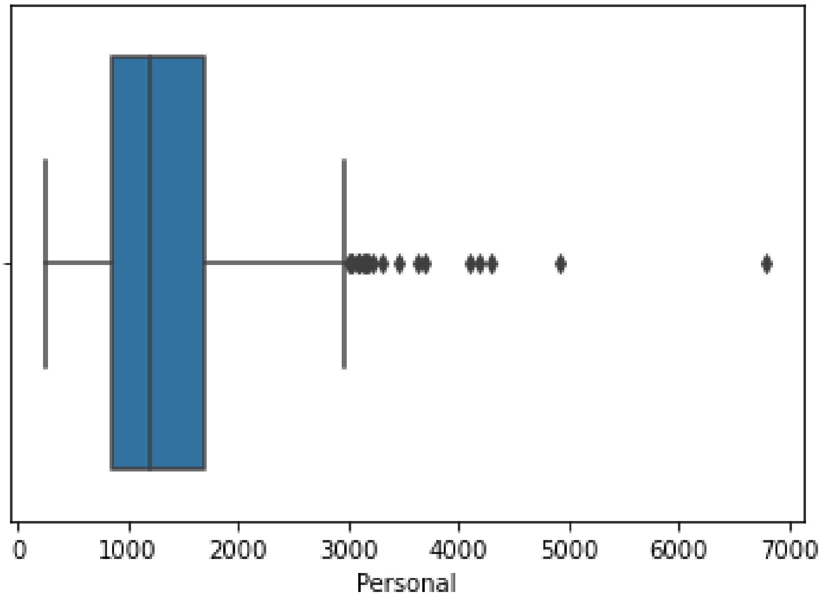
11.Personal

Fig-2.21 Distribution of Personal



The distribution of the data is positively skewed.The maximum personal expense is around 6800.

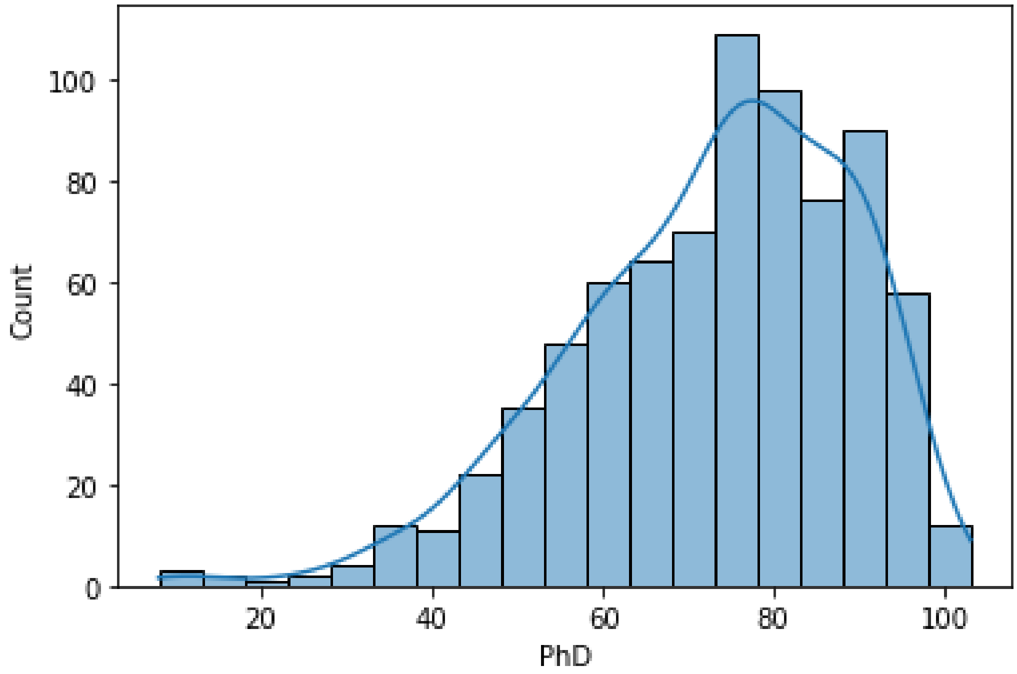
Fig-2.22 Boxplot on Personal



From the Boxplot we can see there is presence of outliers in the dataset.

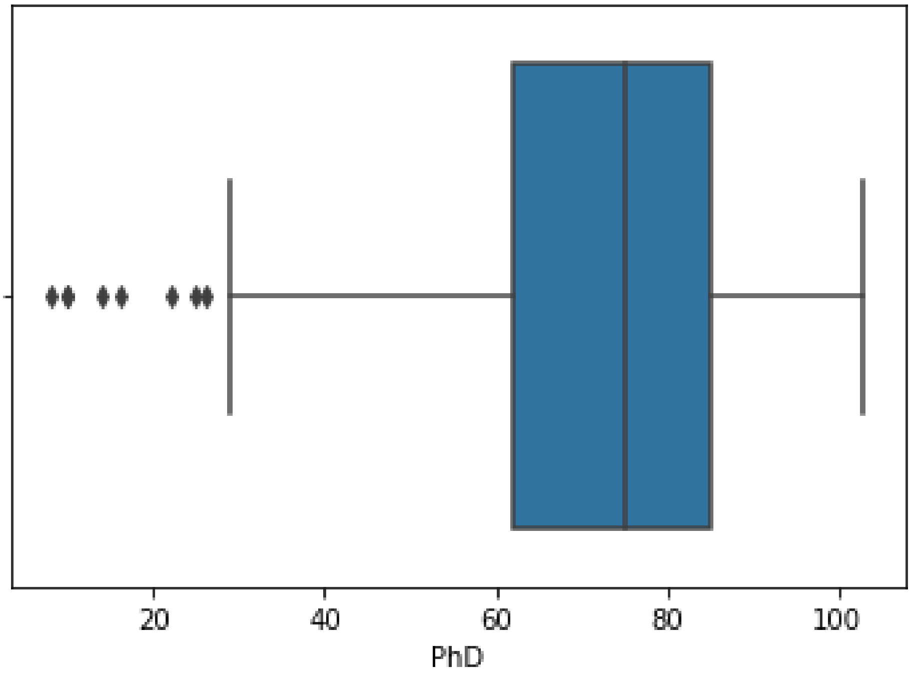
12.PhD

Fig-2.23 Distribution of PhD



The distribution of the data is negatively skewed.

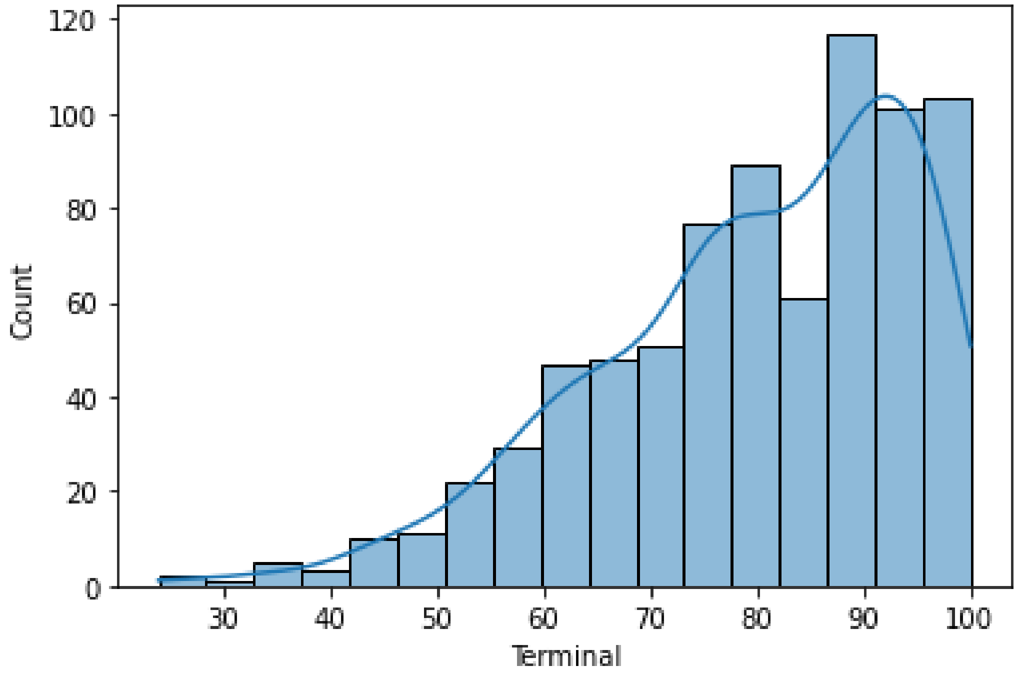
Fig-2.24 Boxplot on PhD



From the Boxplot we can see there is presence of outliers in the dataset.

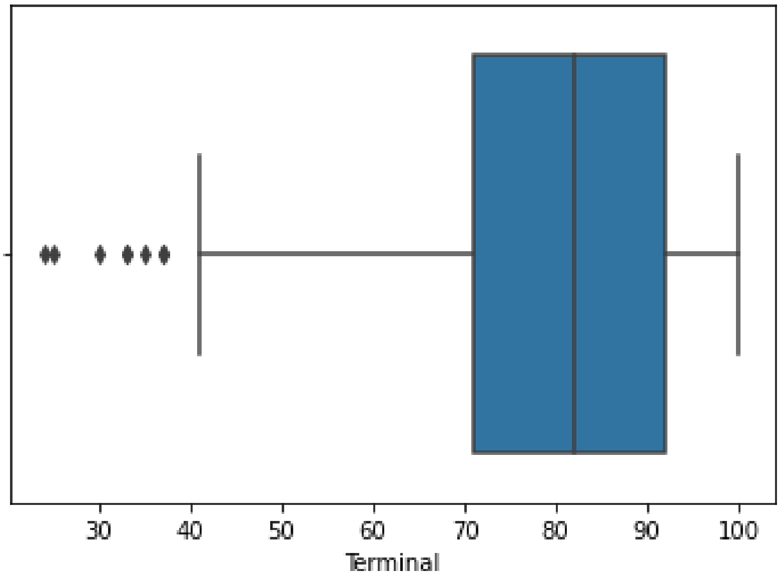
13.Terminal

Fig-2.25 Distribution of Terminal



The distribution of the data is negatively skewed.

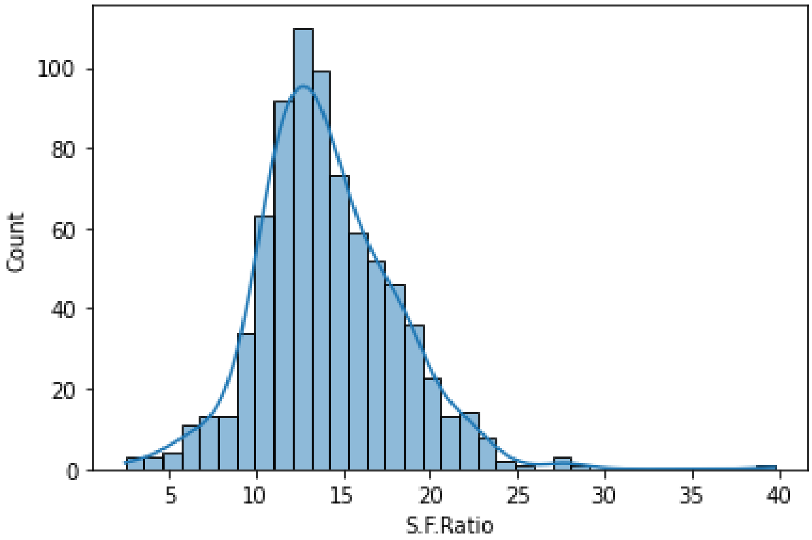
Fig-2.26 Boxplot on Terminal



From the Boxplot we can see there is presence of outliers in the dataset.

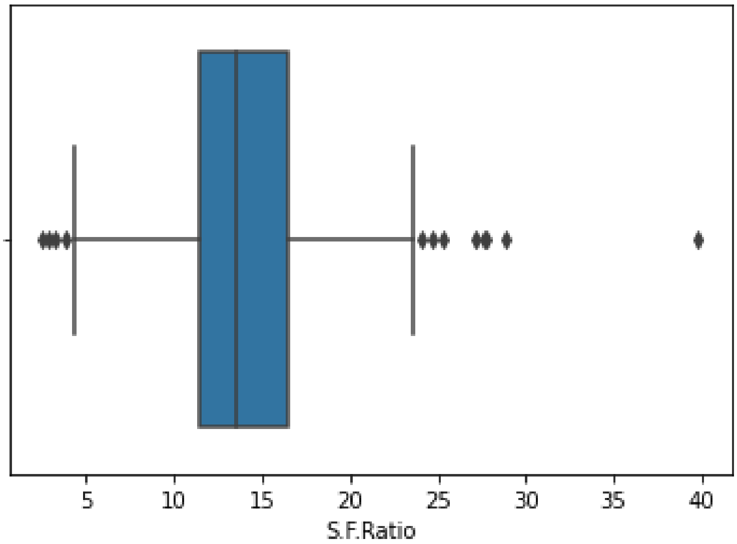
14.S.F.Ratio

Fig-2.27 Distribution of S.F.Ratio



The distribution of the data is almost normal

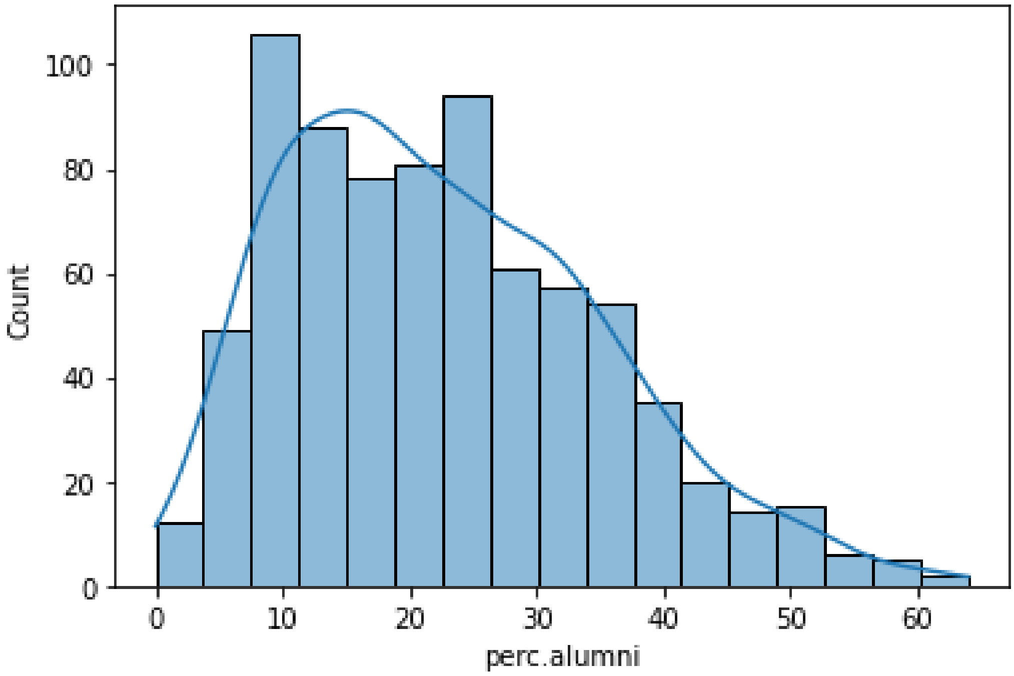
Fig-2.28 Boxplot on S.F Ratio



From the Boxplot we can see there is presence of outliers on both ends in the dataset.

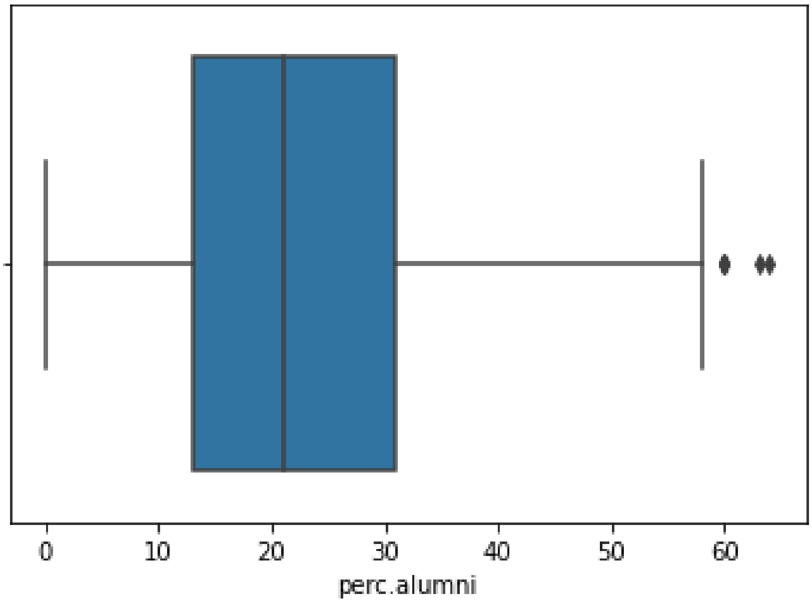
15.Perc.alumni

Fig-2.29 Distribution of Perc.alumni



The distribution of the data is almost normal

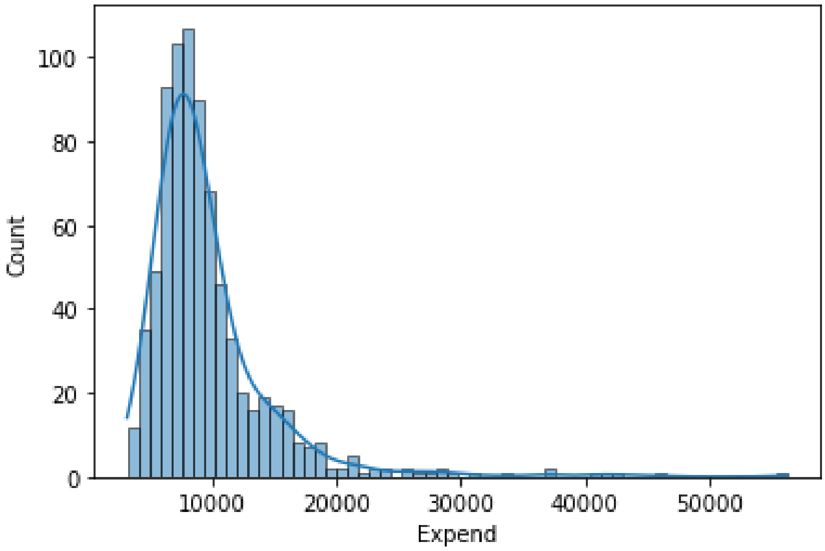
Fig-2.30 Boxplot on Perc.alumni



From the Boxplot we can see there is presence of outliers on both ends in the dataset.

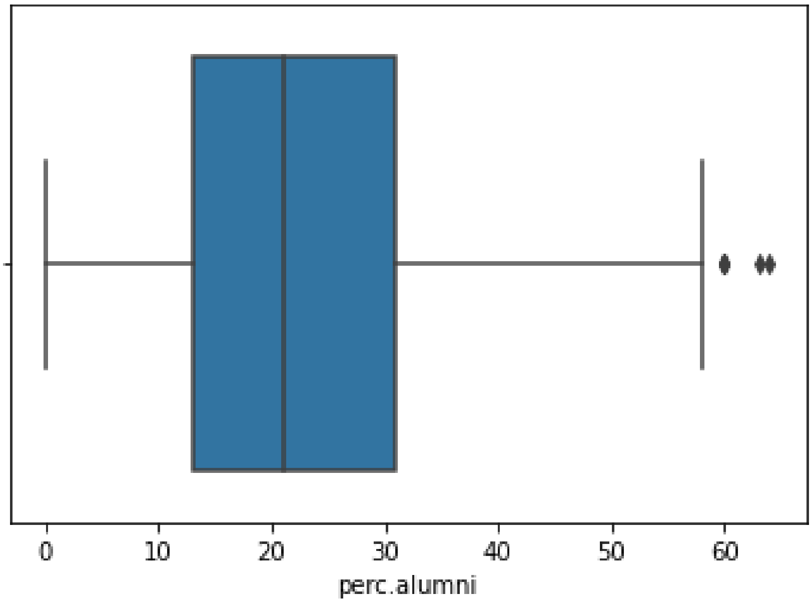
16.Expend

Fig-2.31 Distribution of Expend



The distribution of the data is positively skewed.

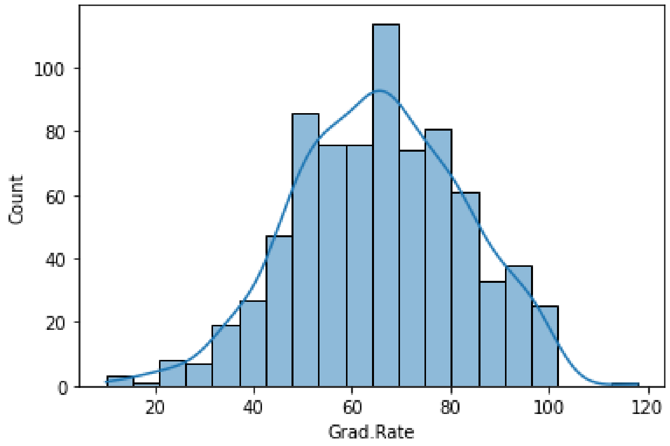
Fig-2.32 Boxplot on Expend



From the Boxplot we can see there is presence of outliers in the dataset.

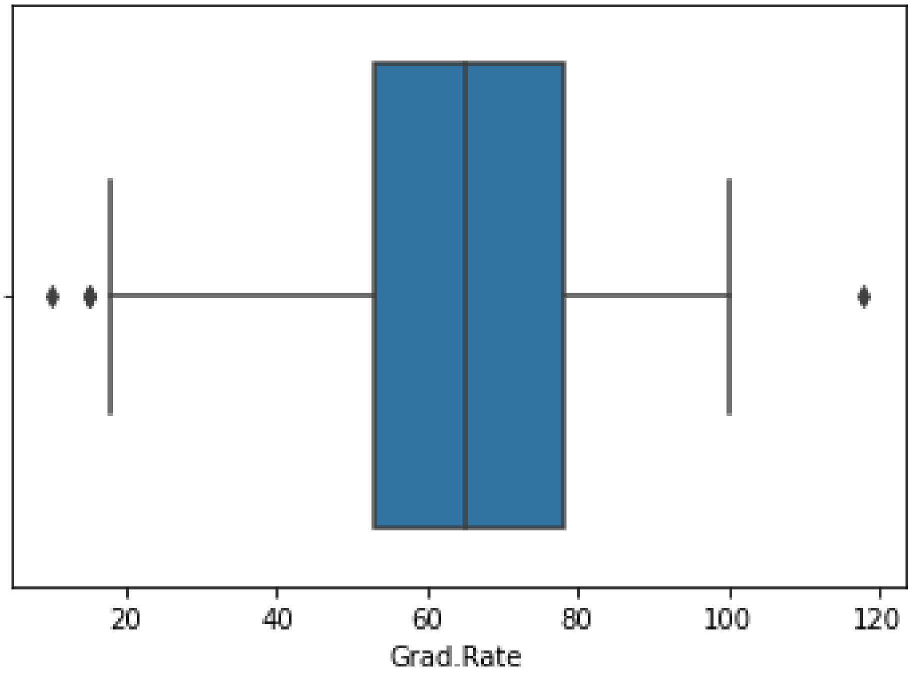
17.Grad.rate

Fig-2.33 Distribution of Grad.rate



The distribution of the data is almost normal.

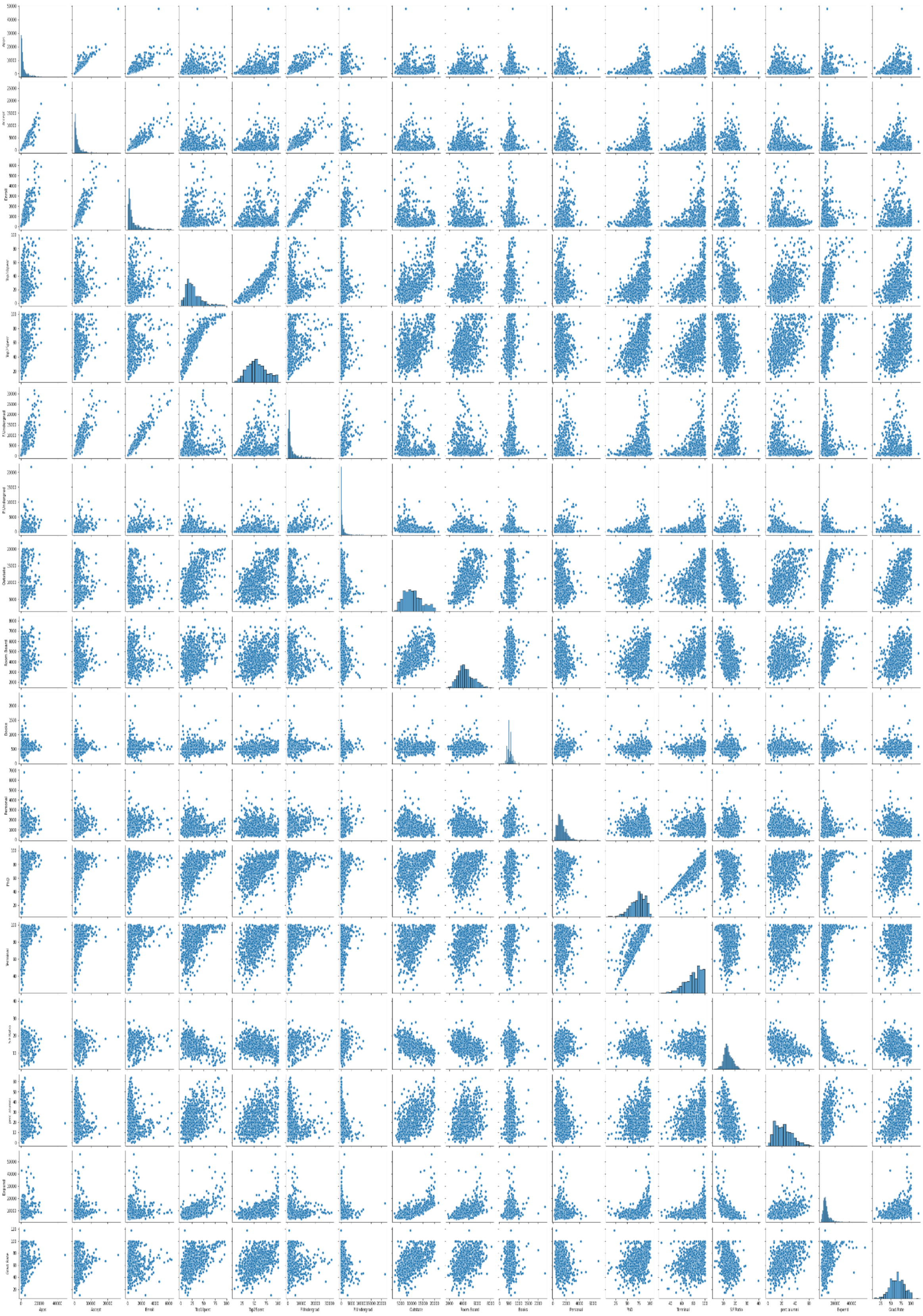
Fig-2.34 Boxplot on Grad.rate



From the Boxplot we can see there is presence of outliers in the dataset.

Bi/Multivariate Analysis

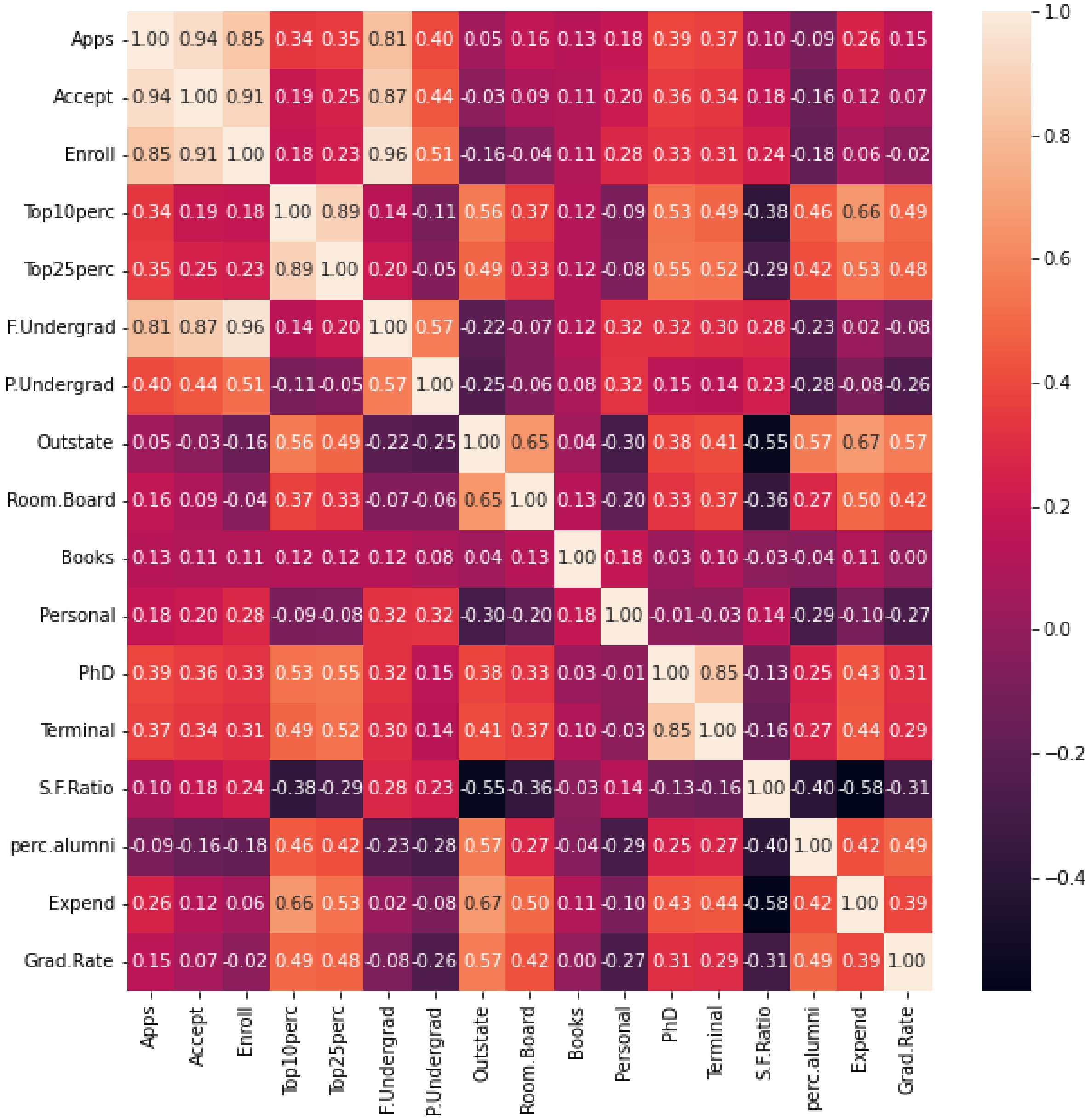
Fig -2.35 Pairplot of Variables



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

Fig -2.36 Heatmap of Variables

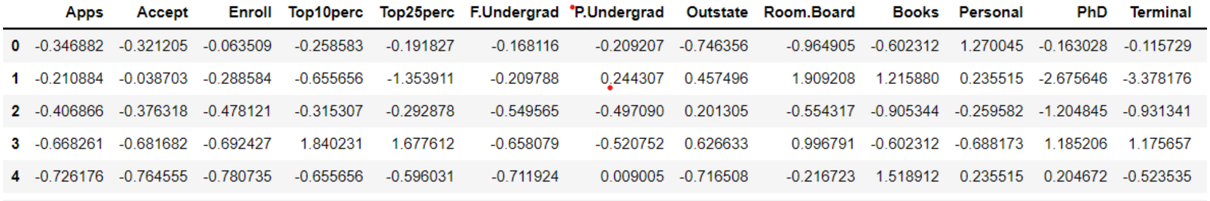


This Heat map gives us the correlation between two numerical values.The highly correlated variables have value around 1.0 we see that the application variable is highly positively correlated with application accepted, students enrolled and full time graduates.From this heatmap insights on application acceptance and the student enrolment as fulltime graduate can be found.High negative correlation is seen between application and percentage of alumni.

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

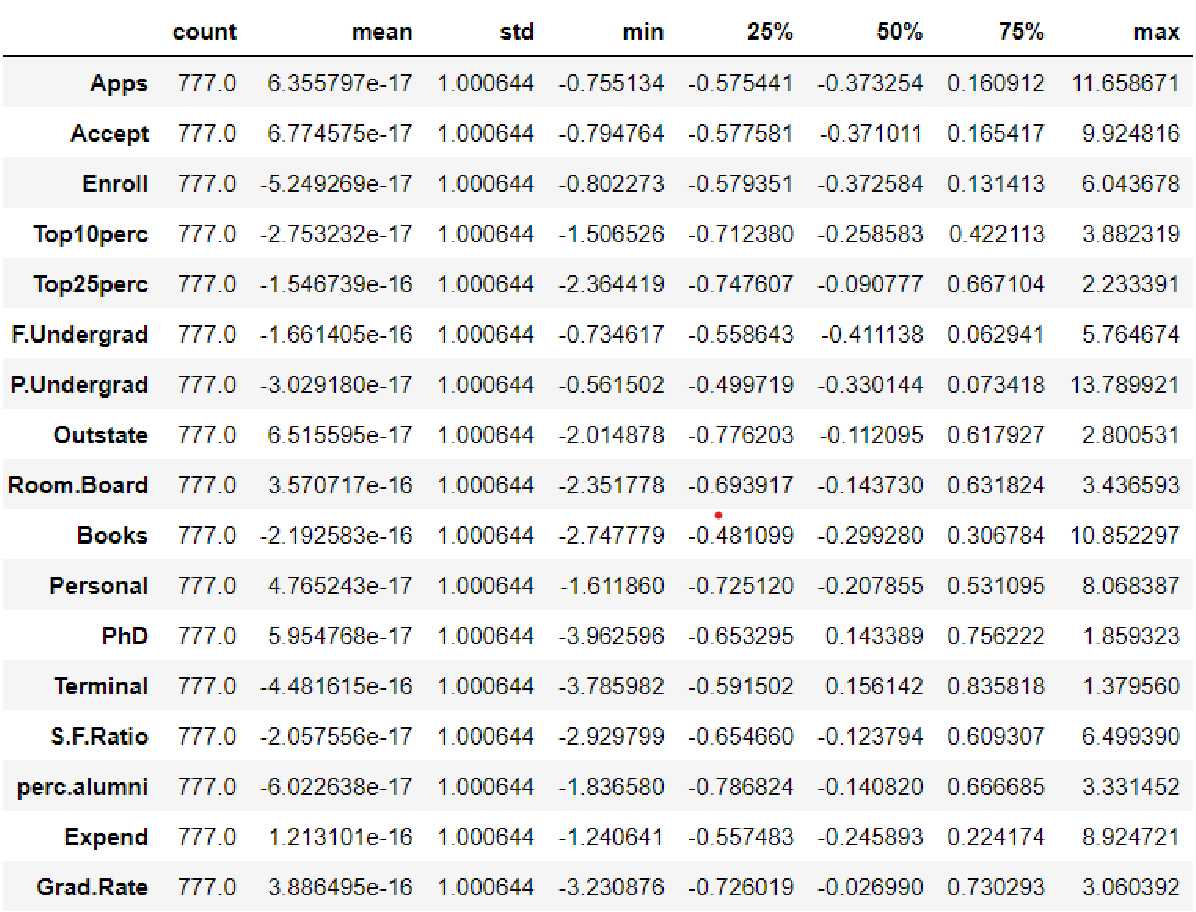
Scaling the target value is a good idea in regression modelling; scaling of the data makes iteasy for a model to learn and understand the problem. Scaling of the data comes under the set of steps of data pre-processing. Here the numeric variables are of different scales which will impact the results of PCA, hence scaling is performed using **z scaling** method .

Table-2.3 Sample Scaled Dataset



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.163028	-0.115729
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.675646	-3.378176
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1.204845	-0.931341
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.185206	1.175657
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0.204672	-0.523535

Table-2.4 Summary Scaled Dataset



	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

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

Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency.

The value of covariance lies in the range of -∞ and +∞.

Correlation is a statistical measure that indicates how strongly two variables are related.

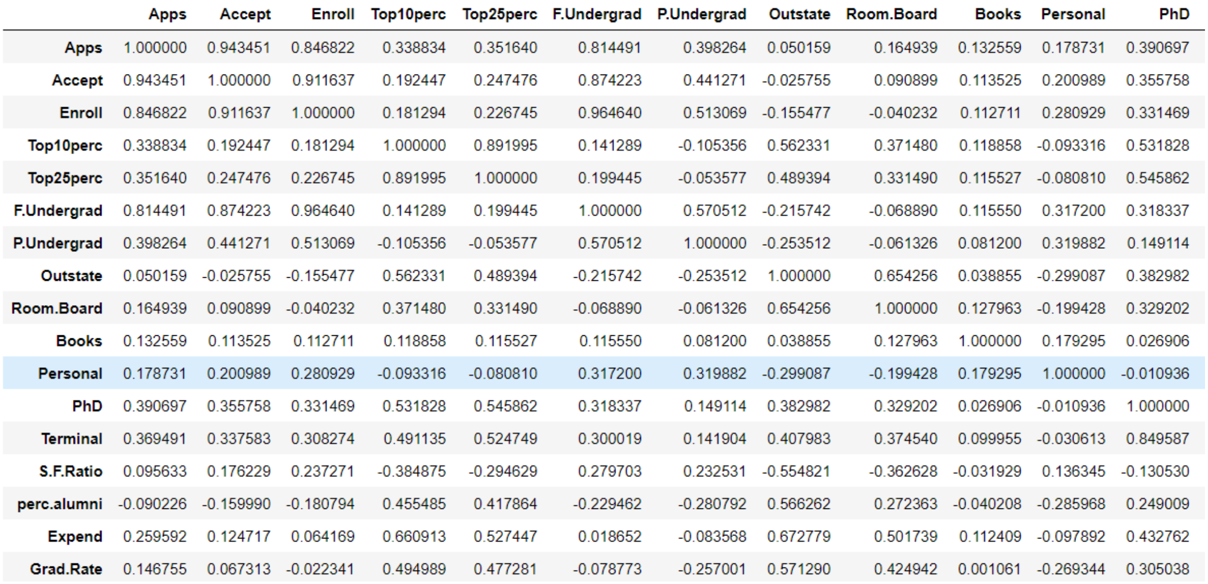
Correlation is limited to values between the range -1 and +1

Covariance Matrix

```
[[ 1.00128866,  0.94466636,  0.84791332,  0.33927032,  0.35209304,
         0.81554018,  0.3987775 ,  0.05022367,  0.16515151,  0.13272942,
         0.17896117,  0.39120081,  0.36996762,  0.09575627, -0.09034216,
         0.2599265 ,  0.14694372],
       [ 0.94466636,  1.00128866,  0.91281145,  0.19269493,  0.24779465,
         0.87534985,  0.44183938, -0.02578774,  0.09101577,  0.11367165,
         0.20124767,  0.35621633,  0.3380184 ,  0.17645611, -0.16019604,
         0.12487773,  0.06739929],
       [ 0.84791332,  0.91281145,  1.00128866,  0.18152715,  0.2270373 ,
         0.96588274,  0.51372977, -0.1556777 , -0.04028353,  0.11285614,
         0.28129148,  0.33189629,  0.30867133,  0.23757707, -0.18102711,
         0.06425192, -0.02236983],
       [ 0.33927032,  0.19269493,  0.18152715,  1.00128866,  0.89314445,
         0.1414708 , -0.10549205,  0.5630552 ,  0.37195909,  0.1190116 ,
        -0.09343665,  0.53251337,  0.49176793, -0.38537048,  0.45607223,
         0.6617651 ,  0.49562711],
       [ 0.35209304,  0.24779465,  0.2270373 ,  0.89314445,  1.00128866,
         0.19970167, -0.05364569,  0.49002449,  0.33191707,  0.115676  ,
        -0.08091441,  0.54656564,  0.52542506, -0.29500852,  0.41840277,
         0.52812713,  0.47789622],
       [ 0.81554018,  0.87534985,  0.96588274,  0.1414708 ,  0.19970167,
         1.00128866,  0.57124738, -0.21602002, -0.06897917,  0.11569867,
         0.31760831,  0.3187472 ,  0.30040557,  0.28006379, -0.22975792,
         0.01867565, -0.07887464],
       [ 0.3987775 ,  0.44183938,  0.51372977, -0.10549205, -0.05364569,
         0.57124738,  1.00128866, -0.25383901, -0.06140453,  0.08130416,
         0.32029384,  0.14930637,  0.14208644,  0.23283016, -0.28115421,
        -0.08367612, -0.25733218],
       [ 0.05022367, -0.02578774, -0.1556777 ,  0.5630552 ,  0.49002449,
        -0.21602002, -0.25383901,  1.00128866,  0.65509951,  0.03890494,
        -0.29947232,  0.38347594,  0.40850895, -0.55553625,  0.56699214,
         0.6736456 ,  0.57202613],
       [ 0.16515151,  0.09101577, -0.04028353,  0.37195909,  0.33191707,
        -0.06897917, -0.06140453,  0.65509951,  1.00128866,  0.12812787,
        -0.19968518,  0.32962651,  0.3750222 , -0.36309504,  0.27271444,
         0.50238599,  0.42548915],
       [ 0.13272942,  0.11367165,  0.11285614,  0.1190116 ,  0.115676  ,
         0.11569867,  0.08130416,  0.03890494,  0.12812787,  1.00128866,
        0.17952581,  0.0269404 ,  0.10008351, -0.03197042, -0.04025955,
        0.11255393,  0.00106226],
       [ 0.17896117,  0.20124767,  0.28129148, -0.09343665, -0.08091441,
        0.31760831,  0.32029384, -0.29947232, -0.19968518,  0.17952581,
        1.00128866, -0.01094989, -0.03065256,  0.13652054, -0.2863366 ,
       -0.09801804, -0.26969106],
       [ 0.39120081,  0.35621633,  0.33189629,  0.53251337,  0.54656564,
        0.3187472 ,  0.14930637,  0.38347594,  0.32962651,  0.0269404 ,
       -0.01094989,  1.00128866,  0.85068186, -0.13069832,  0.24932955,
        0.43331936,  0.30543094],
       [ 0.36996762,  0.3380184 ,  0.30867133,  0.49176793,  0.52542506,
        0.30040557,  0.14208644,  0.40850895,  0.3750222 ,  0.10008351,
       -0.03065256,  0.85068186,  1.00128866, -0.16031027,  0.26747453,
        0.43936469,  0.28990033],
       [ 0.09575627,  0.17645611,  0.23757707, -0.38537048, -0.29500852,
        0.28006379,  0.23283016, -0.55553625, -0.36309504, -0.03197042,
        0.13652054, -0.13069832, -0.16031027,  1.00128866, -0.4034484 ,
       -0.5845844 , -0.30710565],
       [-0.09034216, -0.16019604, -0.18102711,  0.45607223,  0.41840277,
       -0.22975792, -0.28115421,  0.56699214,  0.27271444, -0.04025955,
       -0.2863366 ,  0.24932955,  0.26747453, -0.4034484 ,  1.00128866,
        0.41825001,  0.49153016],
       [ 0.2599265 ,  0.12487773,  0.06425192,  0.6617651 ,  0.52812713,
        0.01867565, -0.08367612,  0.6736456 ,  0.50238599,  0.11255393,
       -0.09801804,  0.43331936,  0.43936469, -0.5845844 ,  0.41825001,
        1.00128866,  0.39084571],
       [ 0.14694372,  0.06739929, -0.02236983,  0.49562711,  0.47789622,
       -0.07887464, -0.25733218,  0.57202613,  0.42548915,  0.00106226,
       -0.26969106,  0.30543094,  0.28990033, -0.30710565,  0.49153016,
        0.39084571,  1.00128866]]
```

Correlation Matrix

Table -2.5 sample covariance matrix



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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

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]

The presence of outliers in dataset can be found by plotting boxplot for the variables.Boxplot for dataset before and after scaling is plotted as follows

Fig-2.37 Boxplot of variables before scaling

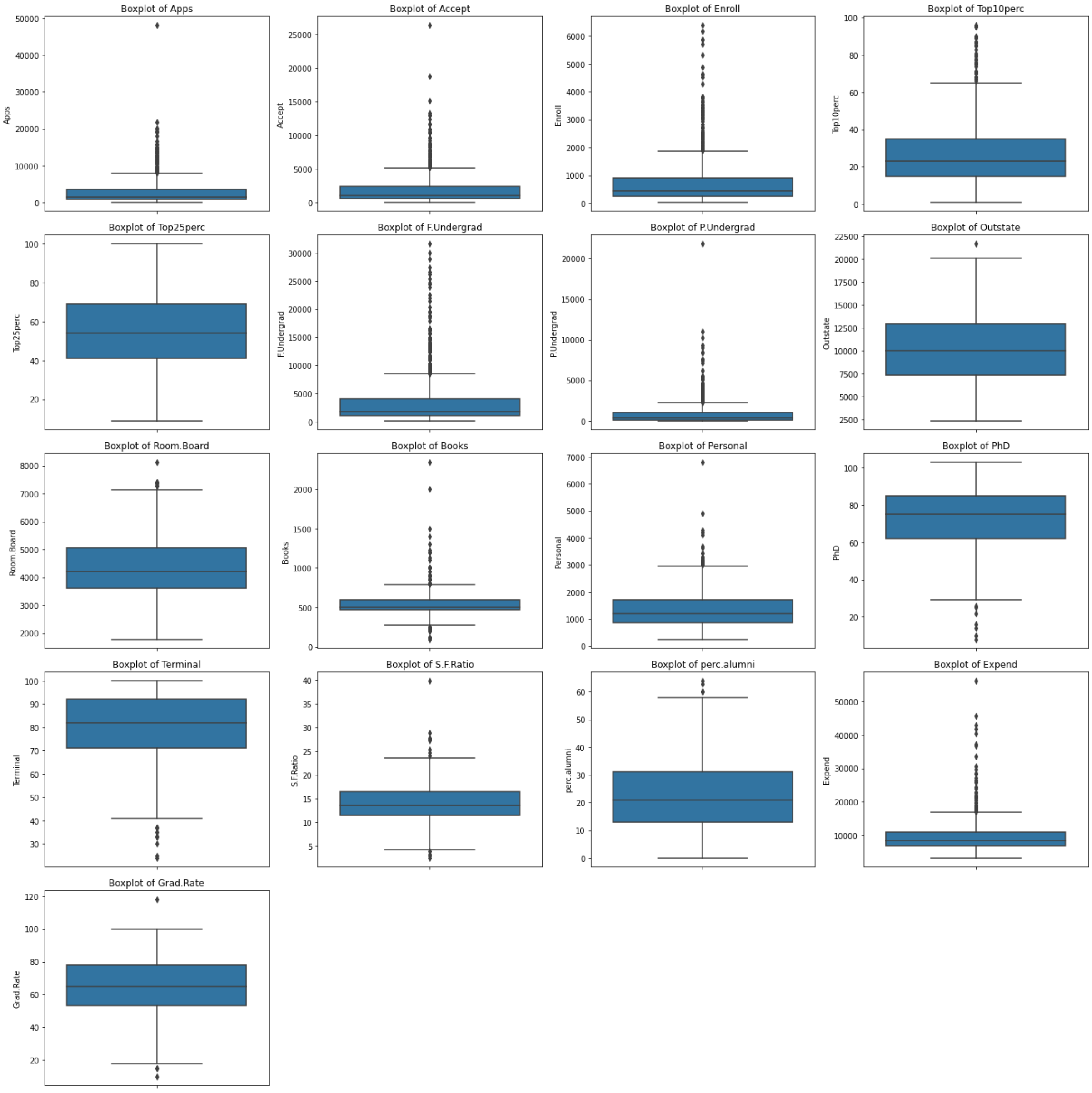
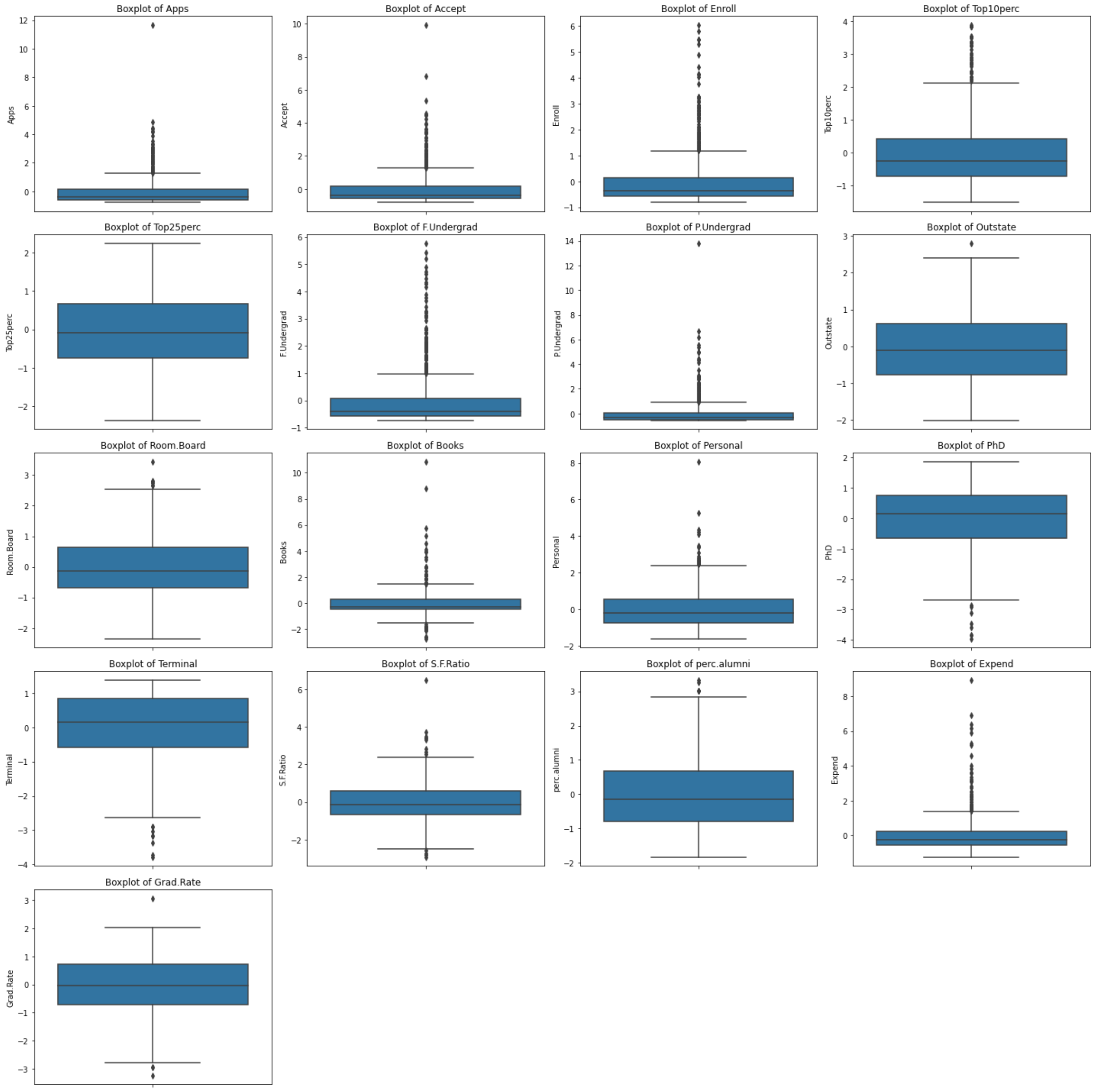


Fig-2.38 Boxplot of variables after scaling



Inference

We observe that the outliers are present in dataset both before and after scaling.The dataset has to be treated to remove the outliers by using appropriate methods such as capping the outlier values to the central measure or any quantile.

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

Eigen Vectors

```
[[ 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]]
```

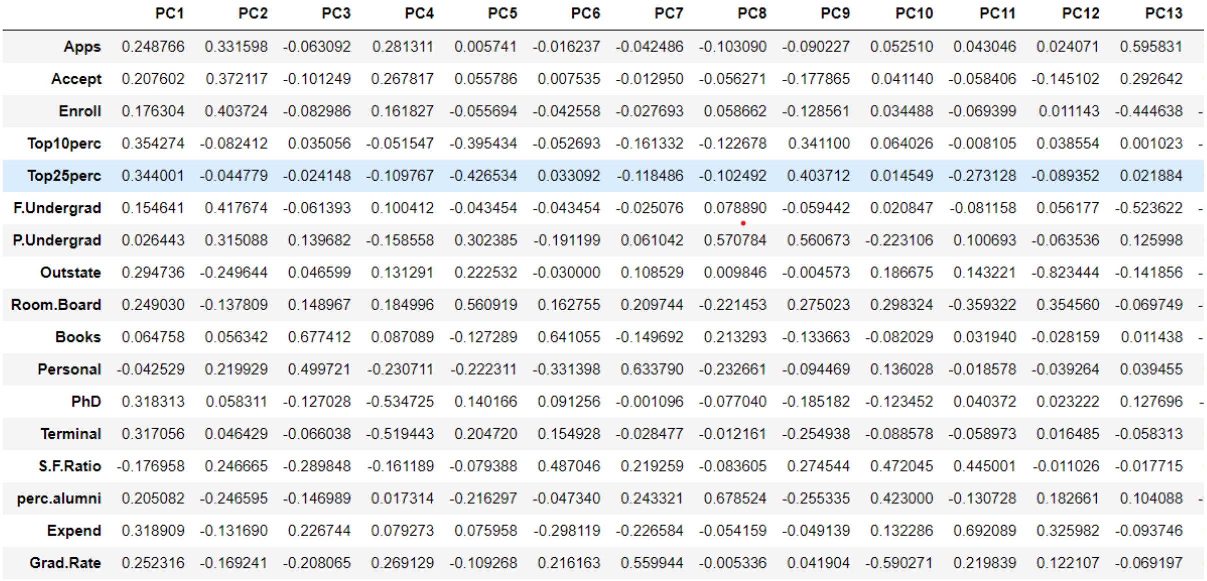
Eigen Values

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

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

The Principal Components with the original features is presented as the following table

Table -2.6 Sample PCA dataset



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.103090	-0.090227	0.052510	0.043046	0.024071	0.595831	
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950	-0.056271	-0.177865	0.041140	-0.058406	-0.145102	0.292642	
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693	0.058662	-0.128561	0.034488	-0.069399	0.011143	-0.444638	-
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678	0.341100	0.064026	-0.008105	0.038554	0.001023	-
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486	-0.102492	0.403712	0.014549	-0.273128	-0.089352	0.021884	
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076	0.078890	-0.059442	0.020847	-0.081158	0.056177	-0.523622	-
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042	0.570784	0.560673	-0.223106	0.100693	-0.063536	0.125998	
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529	0.009846	-0.004573	0.186675	0.143221	-0.823444	-0.141856	-
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744	-0.221453	0.275023	0.298324	-0.359322	0.354560	-0.069749	-
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692	0.213293	-0.133663	-0.082029	0.031940	-0.028159	0.011438	-
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790	-0.232661	-0.094469	0.136028	-0.018578	-0.039264	0.039455	
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096	-0.077040	-0.185182	-0.123452	0.040372	0.023222	0.127696	-
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477	-0.012161	-0.254938	-0.088578	-0.058973	0.016485	-0.058313	
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	0.219259	-0.083605	0.274544	0.472045	0.445001	-0.011026	-0.017715	
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321	0.678524	-0.255335	0.423000	-0.130728	0.182661	0.104088	-
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584	-0.054159	-0.049139	0.132286	0.692089	0.325982	-0.093746	
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944	-0.005336	0.041904	-0.590271	0.219839	0.122107	-0.069197	

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]

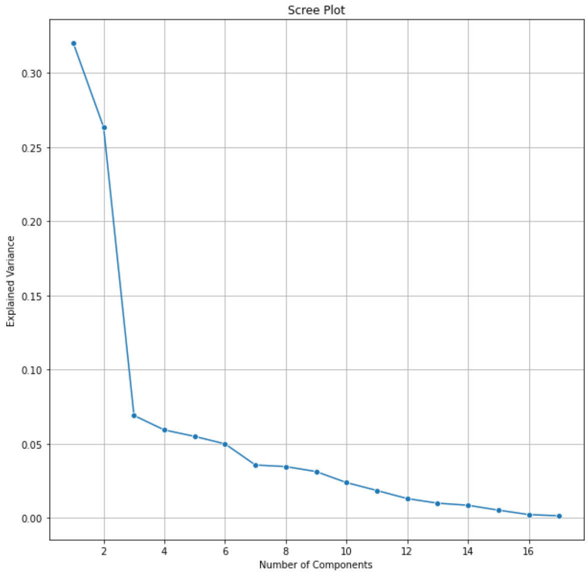
Explicit form of first pc
0.25*Apps + 0.21*Accept + 0.18*Enroll + 0.35*Top10perc + 0.34*Top25perc + 0.15*F.Undergrad + 0.03*P.Undergrad + 0.29*Outstate + 0.25*Room.Board + 0.06*Books + -0.04*Personal + 0.32*PhD + 0.32*Terminal + -0.18*S.F.Ratio + 0.21*perc.alumni + 0.32*Expend + 0.25*Grad.Rate

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

[0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154,
0.81657854, 0.85216726, 0.88670347, 0.91787581, 0.94162773,
0.96004199, 0.9730024 , 0.98285994, 0.99131837, 0.99648962,
0.99864716, 1.

Fig-2.39 Scree Plot



Adding the Eigen values we will get sum of 1
To decide the optimum number of principal components
1. Check for cumulative variance up to 90%, check the corresponding associated with 90%
2. The incremental value between the components should not be less than five percent.

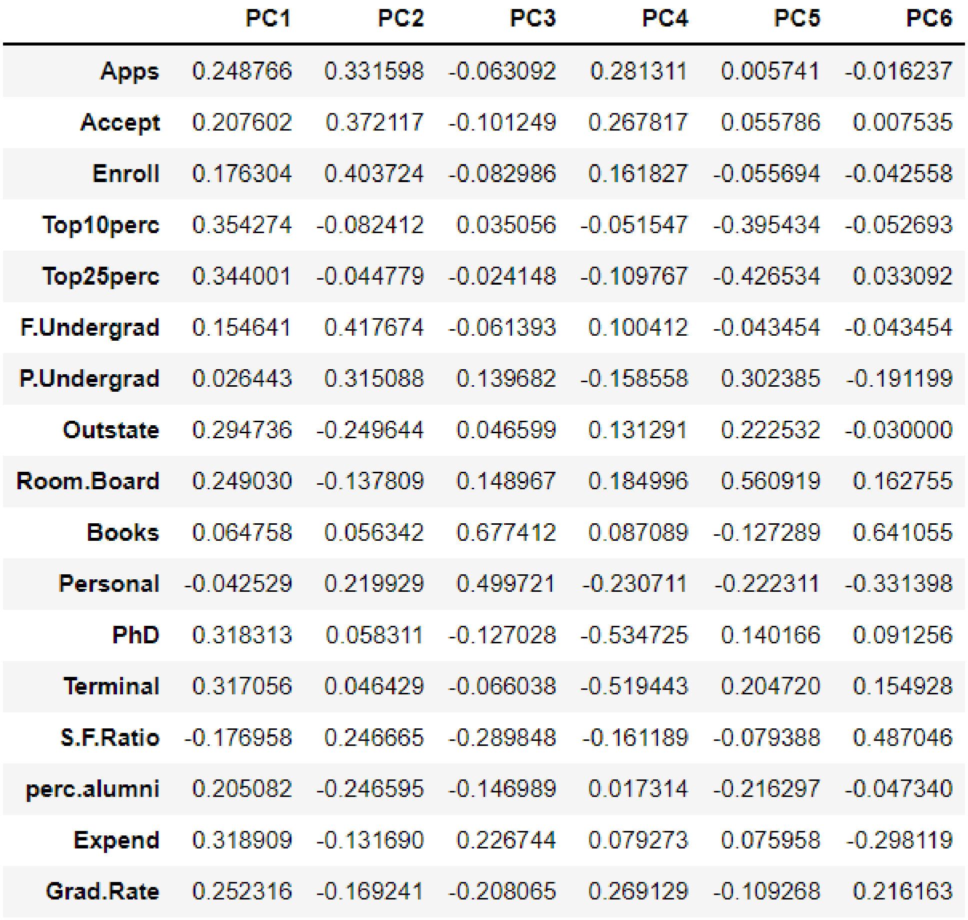
So basis on this we can decide the optimum number of principal components as 6. So, we select 6 principal components for this case study.

The first components explain **32.02%** variance in data
The first two components explains **58.36%** variance in data
The first three components explains **65.26%** variance in data
The first four components explains **71.18%** variance in data
The first five components explains **76.67%** variance in data
The first six components explains **81.66%** variance in data

Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]

The Optimum PCs are decided as 6 for this case study for further analysis.

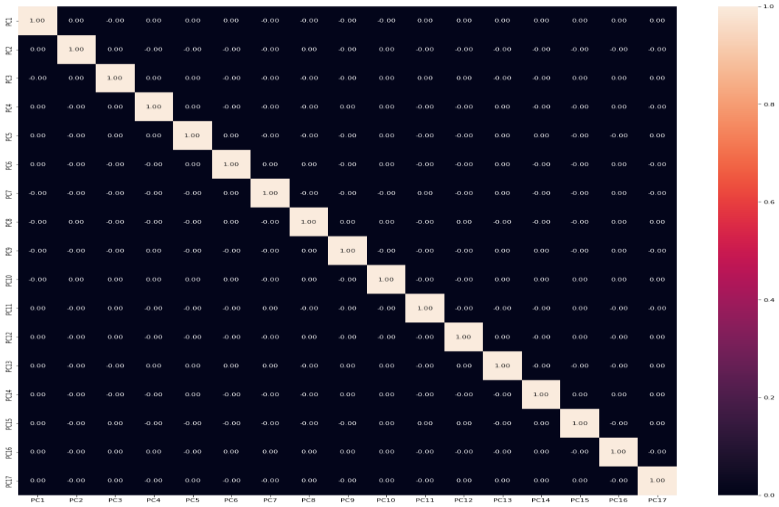
Table -2.7 Dataset of PC



	PC1	PC2	PC3	PC4	PC5	PC6
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163

After PCA the multi collinearity is highly reduced it can be represented by the following heatmap

Fig-2.40 Heat map



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
PC1	1.00	0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PC2	0.00	1.00	0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
PC3	-0.00	0.00	1.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00	0.00	0.00	0.00
PC4	0.00	-0.00	0.00	1.00	0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00
PC5	0.00	-0.00	0.00	0.00	1.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	0.00
PC6	0.00	-0.00	-0.00	-0.00	0.00	1.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00
PC7	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	1.00	-0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00
PC8	-0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	1.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00
PC9	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00	1.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00
PC10	-0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	-0.00	1.00	-0.00	-0.00	0.00	0.00	0.00	0.00	0.00
PC11	0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	1.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00
PC12	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	1.00	0.00	-0.00	0.00	-0.00	0.00
PC13	0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	1.00	-0.00	-0.00	-0.00	0.00
PC14	0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	1.00	0.00	0.00	0.00
PC15	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	-0.00	0.00	1.00	-0.00	0.00
PC16	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00	1.00	-0.00
PC17	0.00	-0.00	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	1.00

The Business implication of using the Principal Component Analysis

The dataset containing information about 777 Colleges/Universities is considered.Exploratory Data Analysis is performed on the daraset.Both Univariate and Bi/Multivariate Analysis are performed on the dataset.The dataset is analysed for the presence of outliers.Principal component analysis (PCA) is a technique for reducing the dimensionality of datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. PCA an adaptive data analysis technique improve the efficiency of machine learning models.Here PCA is done and the optimum no. of PC is considered as 6 which can be used for feeding into machine learning models.