ADVANCE STATISTICS PROJECT

DSBA – 2021 BATCH

SAUMYA RAMANAN

MAY 2021

TABLE OF CONTENTS

Problem 1 3

PROBLEM 1 QUESTION 3

PROBLEM 1.1 SOLUTIONS 3

1.1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually. 3

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. 3

1.3. Perform one-way ANOVA for OCCUPATION with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results. 4

1.4. 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. 5

1.5. 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? 6

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

Problem 2 QUESTION 7

PROBLEM 2 SOLUTION 7

2.1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA? 7

2.2. Is scaling necessary for PCA in this case? Give justification and perform scaling. 17

2.3. Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data]. 17

2.4. Check the dataset for outliers before and after scaling. What insight do you derive here? 21

2.5. Extract the eigenvalues and eigenvectors.[print both] 38

2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features 40

2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). 41

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? ..41

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] 41

REFERENCES 42

PROBLEM 1

PROBLEM 1 QUESTION

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.

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

PROBLEM 1.1 SOLUTIONS

1.1. STATE THE NULL AND THE ALTERNATE HYPOTHESIS FOR CONDUCTING ONE-WAY ANOVA FOR BOTH EDUCATION AND OCCUPATION INDIVIDUALLY.

For Education:

Null Hypothesis: #H0 - The mean of Salary is same for all 3 levels of Treatment (Education).

Alternate: H1 - For atleast one level of education, the mean of Salary is different.

For Occupation:

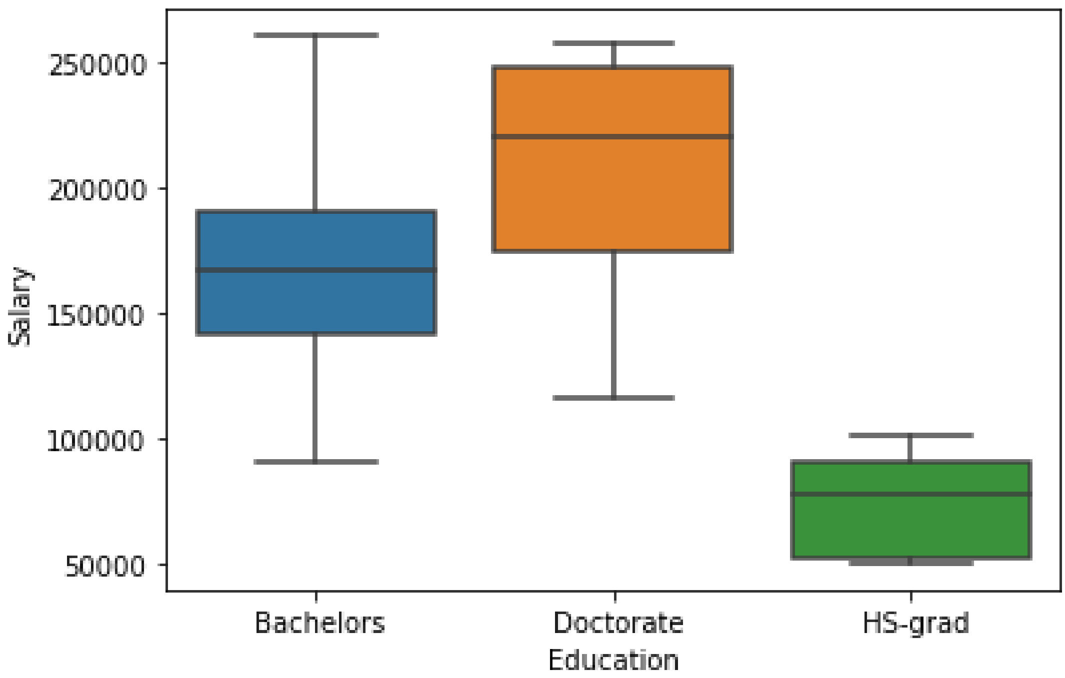
Null: #H0 - The mean of Salary is same for all 4 levels of Treatment (Occupation).

#H1 - For atleast one level of Occupation, the mean of Salary is different.

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.

Data Quality:

It was observed that the data had no nulls or duplicate. The distribution via box plot showed normal distribution with varying means:



Performing 1 way Anova for Test Hypothesis condictions mentioned above, we get the following result:

	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

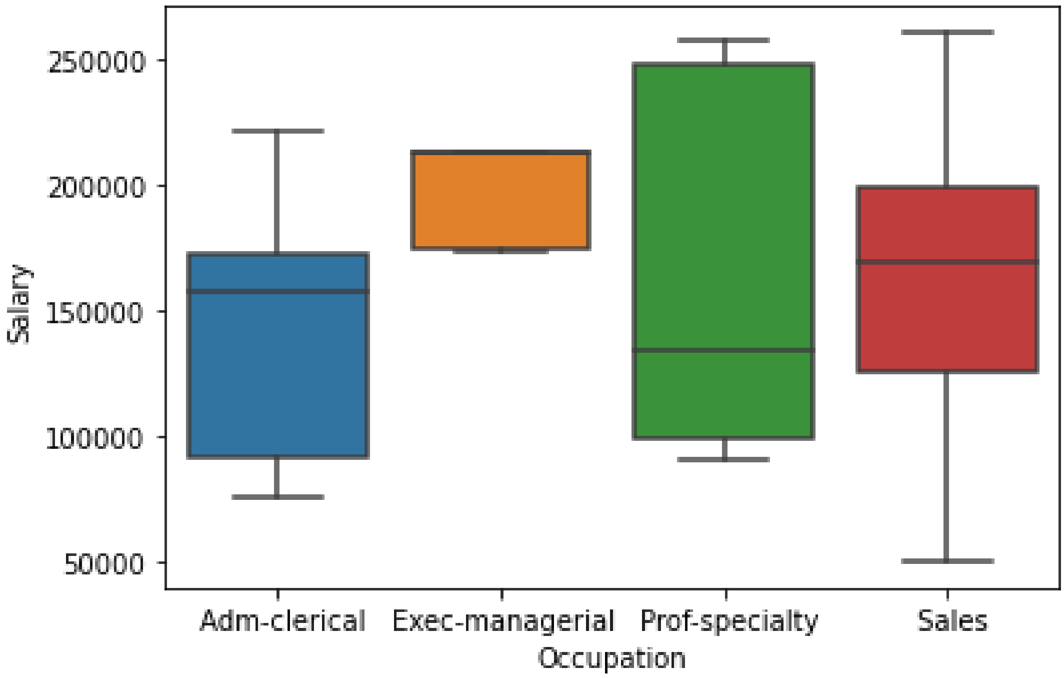
Since the P value is < 0.05 (we assume alpha = 0.05), the null hypothesis is rejected.

Indicating that : For atleast one level of education, the mean of Salary is different.

1.3. PERFORM ONE-WAY ANOVA FOR OCCUPATION WITH RESPECT TO THE VARIABLE 'SALARY'. STATE WHETHER THE NULL HYPOTHESIS IS ACCEPTED OR REJECTED BASED ON THE ANOVA RESULTS.

Data Quality:

It was observed that the data had no nulls or duplicate. The distribution via box plot showed normal distribution with varying means:



Performing 1 way Anova for Test Hypothesis condictions mentioned above, we get the following result:

	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

Since the P value is < 0.05 (we assume alpha = 0.05), we fail to reject the null hypothesis.

Indicating that : The mean of Salary is same for all 4 levels of Treatment (Occupation).

1.4. 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.

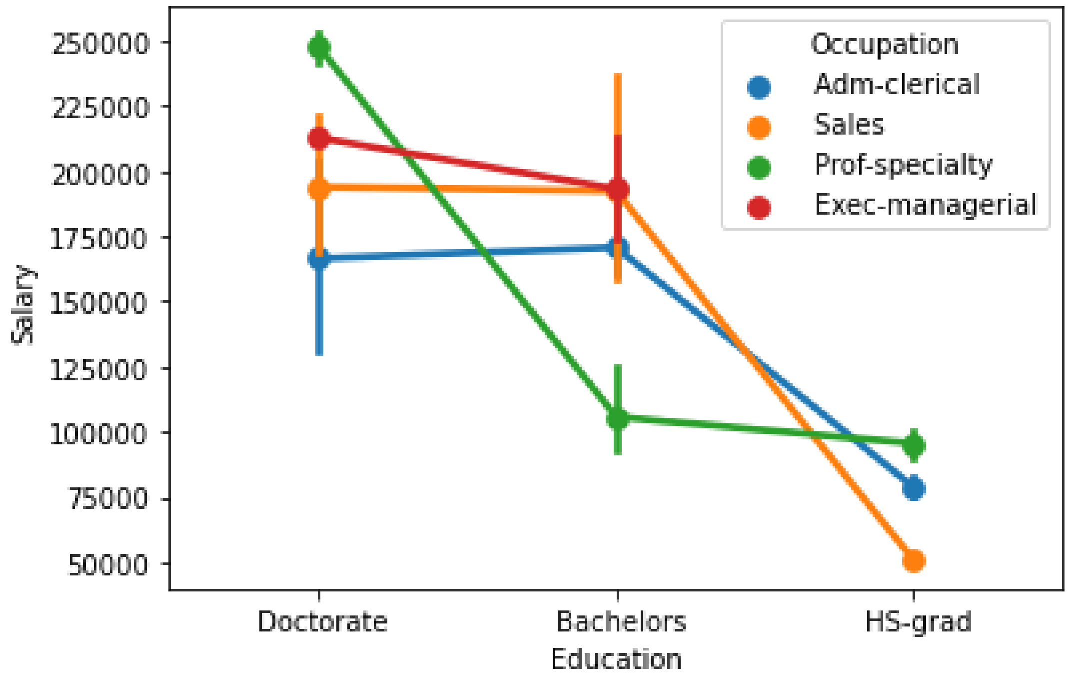
On performing Anova on the interaction of the 2 variables we get the following output:

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	2.284576	9.648715e-02
C(Education)	2.0	9.695663e+10	4.847831e+10	29.510933	3.708479e-08

Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

The p-value in the one of the treatments is less than alpha (0.05)

Let us check whether there is any interaction effect between the treatments.



We observe that there is some interaction between education and occupation (esp in Prof Spl).

1.5. PERFORM A TWO-WAY ANOVA BASED ON THE EDUCATION AND OCCUPATION (ALONG WITH THEIR INTERACTION EDUCATION*OCCUPATION) WITH THE VARIABLE 'SALARY'. STATE THE NULL AND ALTERNATIVE HYPOTHESES AND STATE YOUR RESULTS. HOW WILL YOU INTERPRET THIS RESULT?

H0 – Null Hypothesis: There is no interaction of effect of Education and Occupation on each other .

H1: Alternate Hypothesis: There is an interaction of occupation & educcation

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

C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

The P value 2.23x10^-5 is less than alpha, hence we can conclude there is no interaction between occupation and education that impacts the mean Salary value.

1.6. EXPLAIN THE BUSINESS IMPLICATIONS OF PERFORMING ANOVA FOR THIS PARTICULAR CASE STUDY.

There are 2 treatments in this use case – Occupation and Education. We saw interference between the 2 treatments and we needed to find the impact to Salary due to interactions of these treatment. ANOVA is a great technique to achieve this.

ANOVA can also be used to forecast trends by analyzing patterns in data to better understand the Salary trends (in this case). It's also a widely used statistical technique for comparing the relationship between factors that cause a change in Salary, such as educational and occupational impact on salaries and this helps take measures for the future.

PROBLEM 2 QUESTION

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.

PROBLEM 2 SOLUTION

2.1. PERFORM EXPLORATORY DATA ANALYSIS [BOTH UNIVARIATE AND MULTIVARIATE ANALYSIS TO BE PERFORMED]. WHAT INSIGHT DO YOU DRAW FROM THE EDA?

DATA QUALITY CHECKS:

DATA QUALITY CHECKS WERE DONE FOR NULLS, DUPLICATES, AND DATA WAS FOUND TO BE CLEAN. ALSO EXCEP FOR THE NAMES FIELD THE REST OF FIELDS ARE INT (ONLY S.F.RATIO BEING FLOAT).

UNIVARIATE ANALYSIS:

THE MEAN, MODE AND MEDIAN WAS DONRE FOR ALL INT & FLOAT FIELDS.

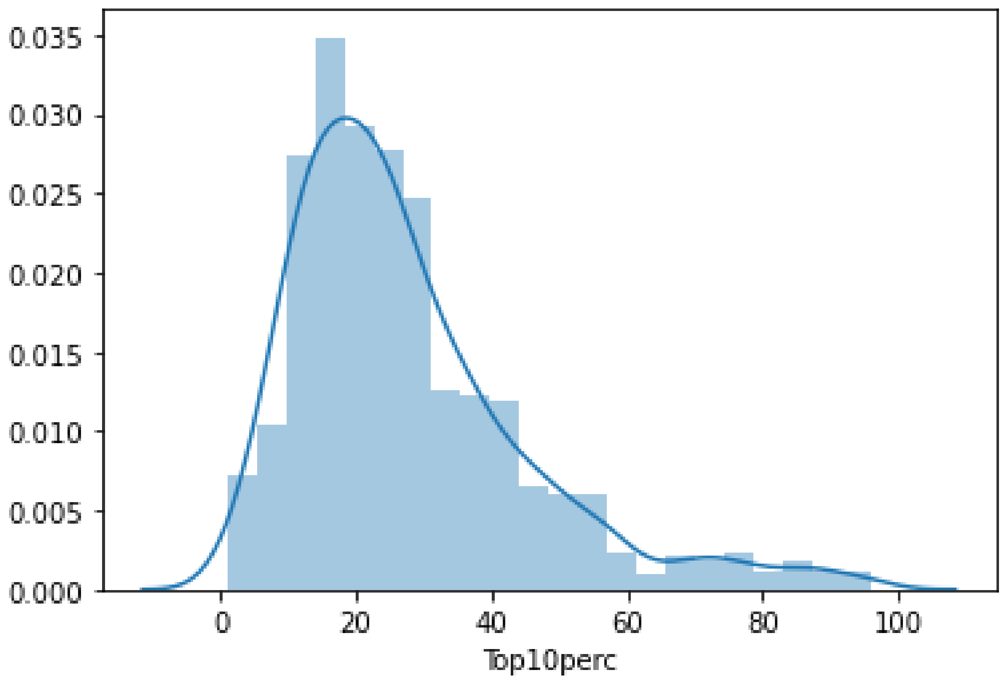
THE DATA WAS FOUND TO BE CLOSE TO NORMAL DISTRIBUTION for INTEGER CONTINUOS VARIABLES Like Top10perc, Top25perc, Perc.Alumni – these indicate the percentage of students applying from top(10,25 %)and also % of Alumni who denote and graduation Rate .

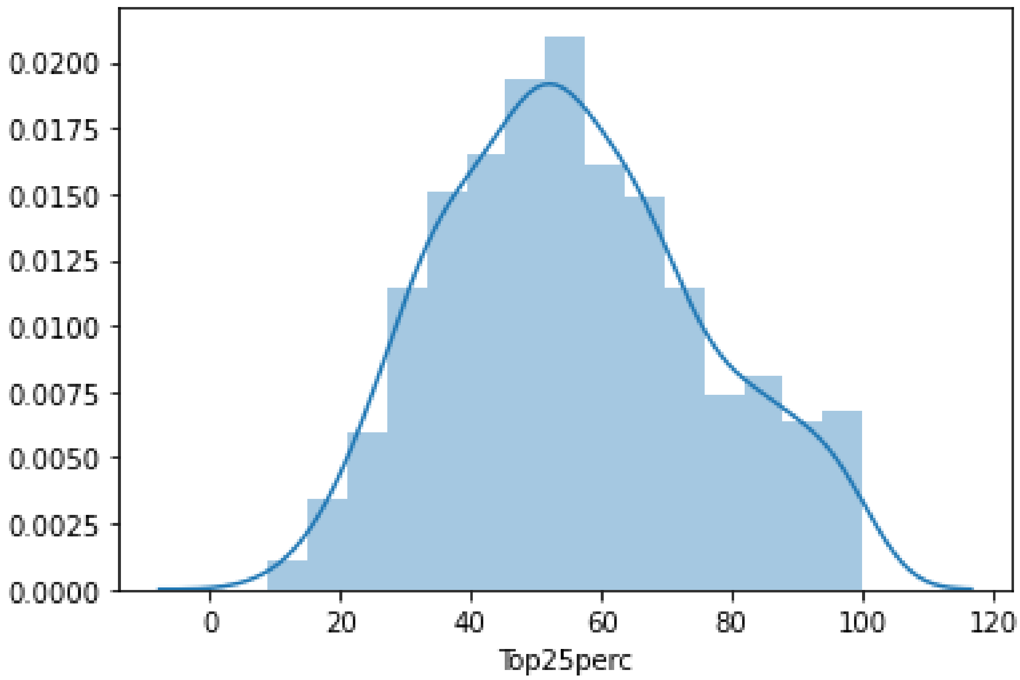
ON expenditure – the Expend, Personal, Books showed normal distribution.

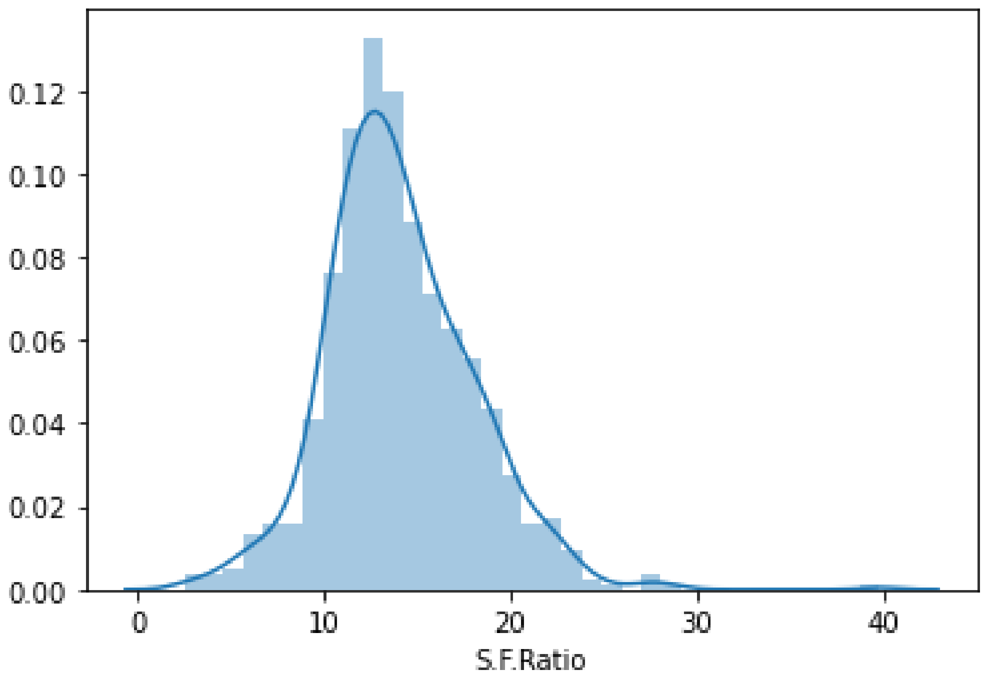
```
count      777.000000
mean      3001.638353
std       3870.201484
min         81.000000
25%        776.000000
50%       1558.000000
75%       3624.000000
max      48094.000000
Name: Apps, dtype: float64
count      777.000000
mean      2018.804376
std       2451.113971
min         72.000000
25%        604.000000
50%       1110.000000
75%       2424.000000
max      26330.000000
Name: Accept, dtype: float64
count     777.000000
mean      779.972973
std       929.176190
min        35.000000
25%       242.000000
50%       434.000000
75%       902.000000
max      6392.000000
Name: Enroll, dtype: float64
count    777.000000
mean      27.558559
std       17.640364
min        1.000000
25%       15.000000
50%       23.000000
75%       35.000000
max       96.000000
Name: Top10perc, dtype: float64
count    777.000000
mean      55.796654
std       19.804778
min        9.000000
```

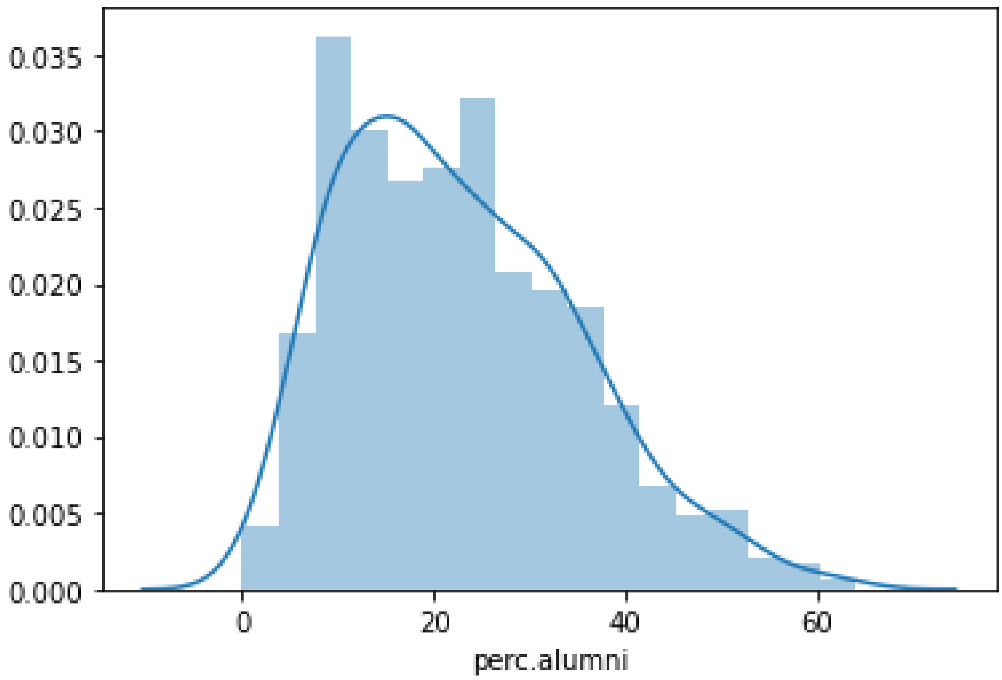
```
25%        41.000000
50%        54.000000
75%        69.000000
max       100.000000
Name: Top25perc, dtype: float64
count      777.000000
mean      3699.907336
std       4850.420531
min        139.000000
25%        992.000000
50%       1707.000000
75%       4005.000000
max      31643.000000
Name: F.Undergrad, dtype: float64
count      777.000000
mean       855.298584
std       1522.431887
min          1.000000
25%         95.000000
50%        353.000000
75%        967.000000
max      21836.000000
Name: P.Undergrad, dtype: float64
count      777.000000
mean     10440.669241
std       4023.016484
min       2340.000000
25%       7320.000000
50%       9990.000000
75%      12925.000000
max      21700.000000
Name: Outstate, dtype: float64
count     777.000000
mean     4357.526384
std      1096.696416
min      1780.000000
25%      3597.000000
50%      4200.000000
75%      5050.000000
max      8124.000000
Name: Room.Board, dtype: float64
count     777.000000
mean      549.380952
std       165.105360
min        96.000000
25%       470.000000
50%       500.000000
75%       600.000000
max      2340.000000
Name: Books, dtype: float64
count     777.000000
mean     1340.642214
std       677.071454
min       250.000000
25%       850.000000
50%      1200.000000
75%      1700.000000
max      6800.000000
```

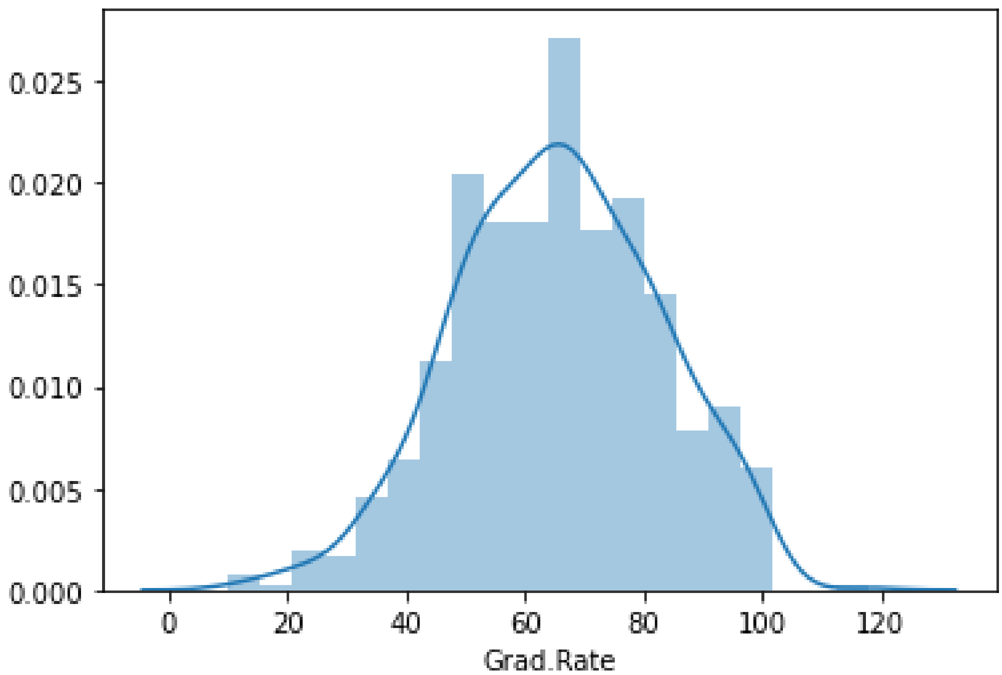
```
Name: Personal, dtype: float64
count    777.000000
mean      72.660232
std       16.328155
min        8.000000
25%       62.000000
50%       75.000000
75%       85.000000
max      103.000000
Name: PhD, dtype: float64
count    777.000000
mean      79.702703
std       14.722359
min       24.000000
25%       71.000000
50%       82.000000
75%       92.000000
max      100.000000
Name: Terminal, dtype: float64
count    777.000000
mean      14.089704
std        3.958349
min        2.500000
25%       11.500000
50%       13.600000
75%       16.500000
max       39.800000
Name: S.F.Ratio, dtype: float64
count    777.000000
mean      22.743887
std       12.391801
min        0.000000
25%       13.000000
50%       21.000000
75%       31.000000
max       64.000000
Name: perc.alumni, dtype: float64
count      777.000000
mean      9660.171171
std       5221.768440
min       3186.000000
25%       6751.000000
50%       8377.000000
75%      10830.000000
max      56233.000000
Name: Expend, dtype: float64
count    777.00000
mean      65.46332
std       17.17771
min       10.00000
25%       53.00000
50%       65.00000
75%       78.00000
max      118.00000
Name: Grad.Rate, dtype: float64
```


0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
Top10perc

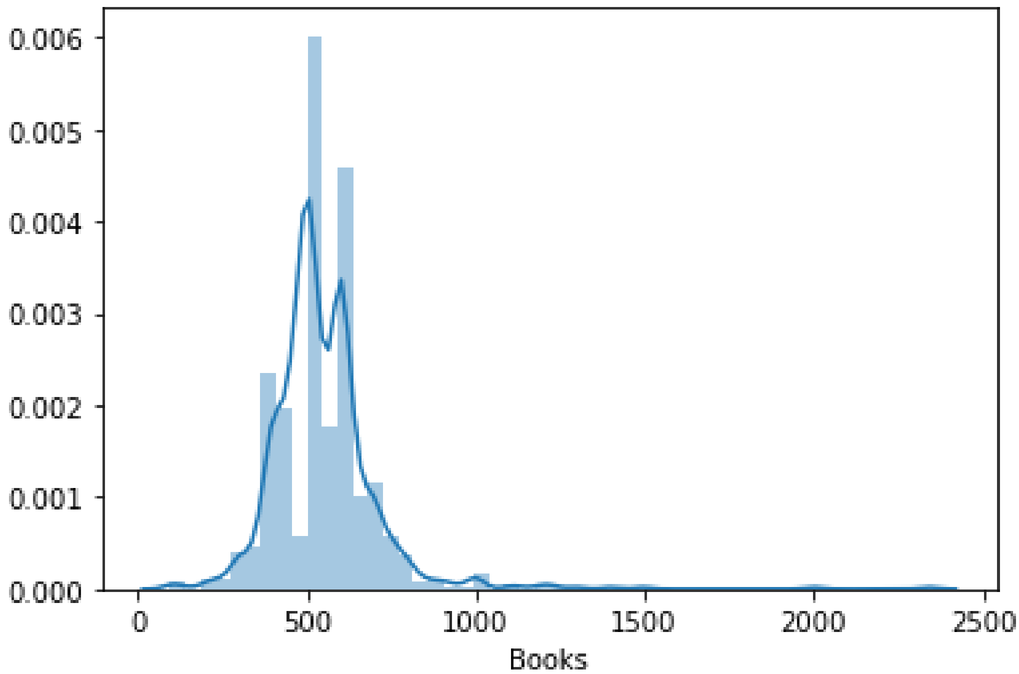

0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
20
40
60
80
100
120
Top25perc


0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
S.F.Ratio

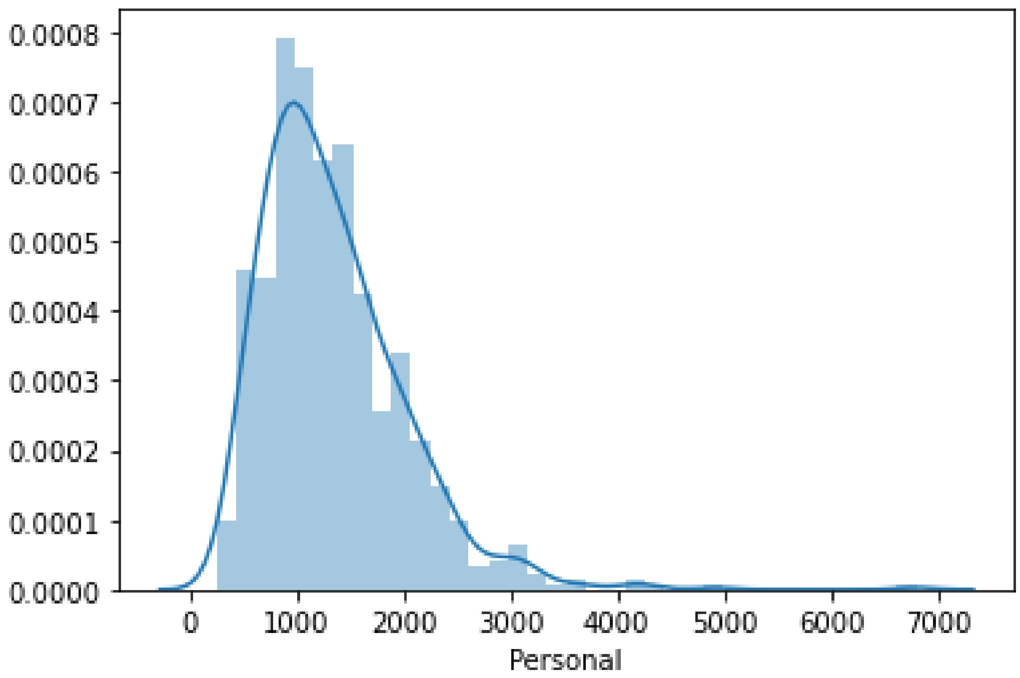

0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
perc.alumni

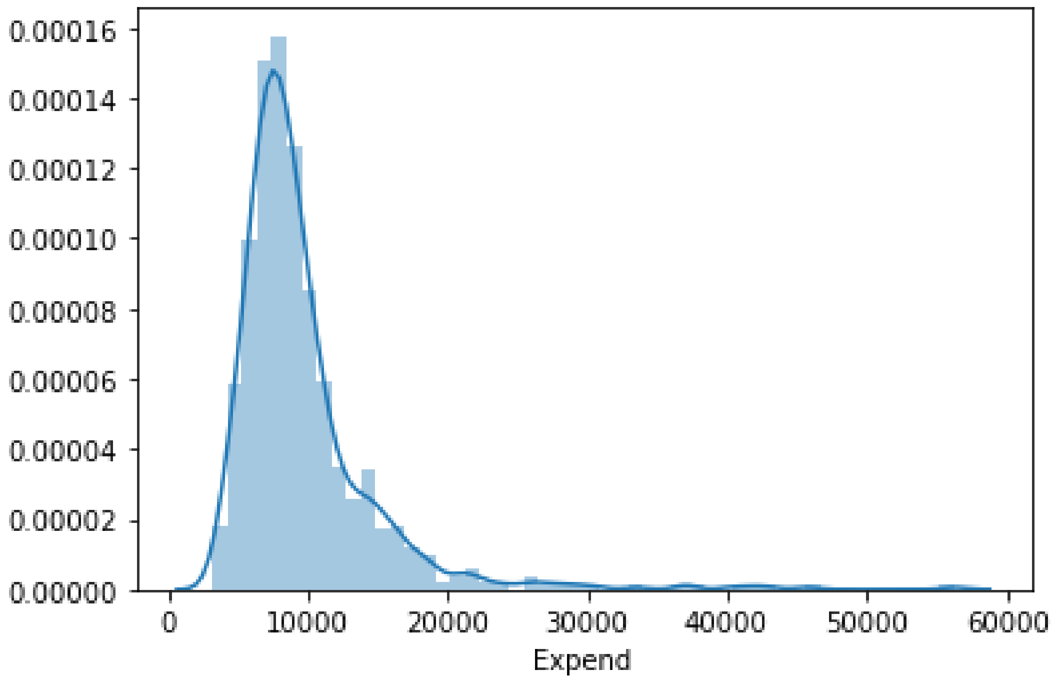


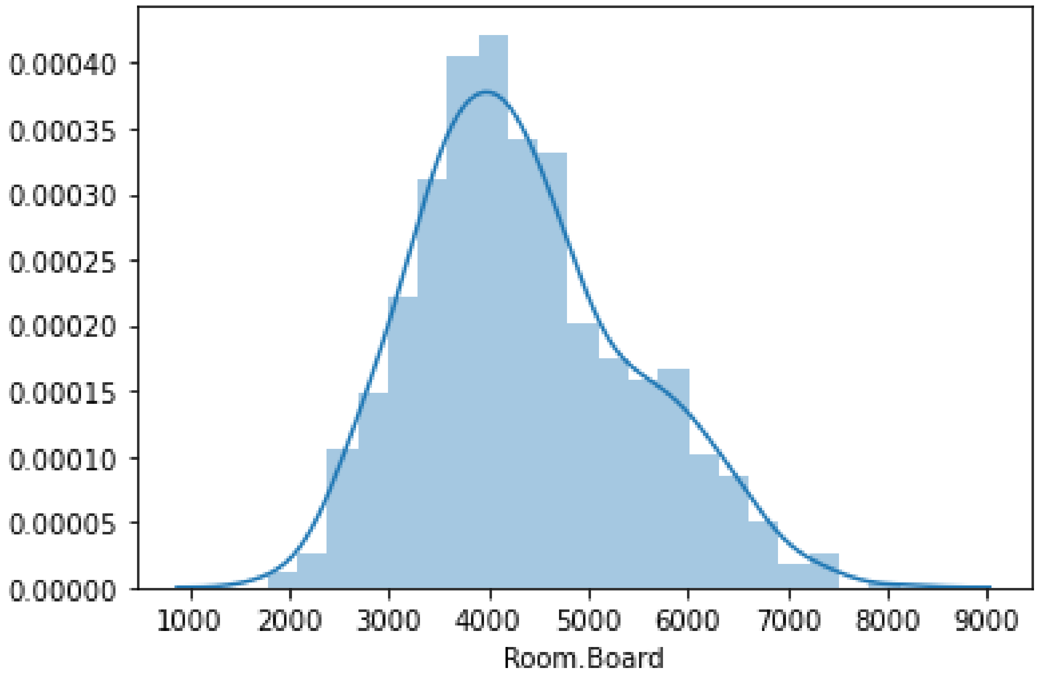
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
Grad.Rate



0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
500
1000
1500
2000
2500
Books


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


0.00016
0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
10000
20000
30000
40000
50000
60000
Expend



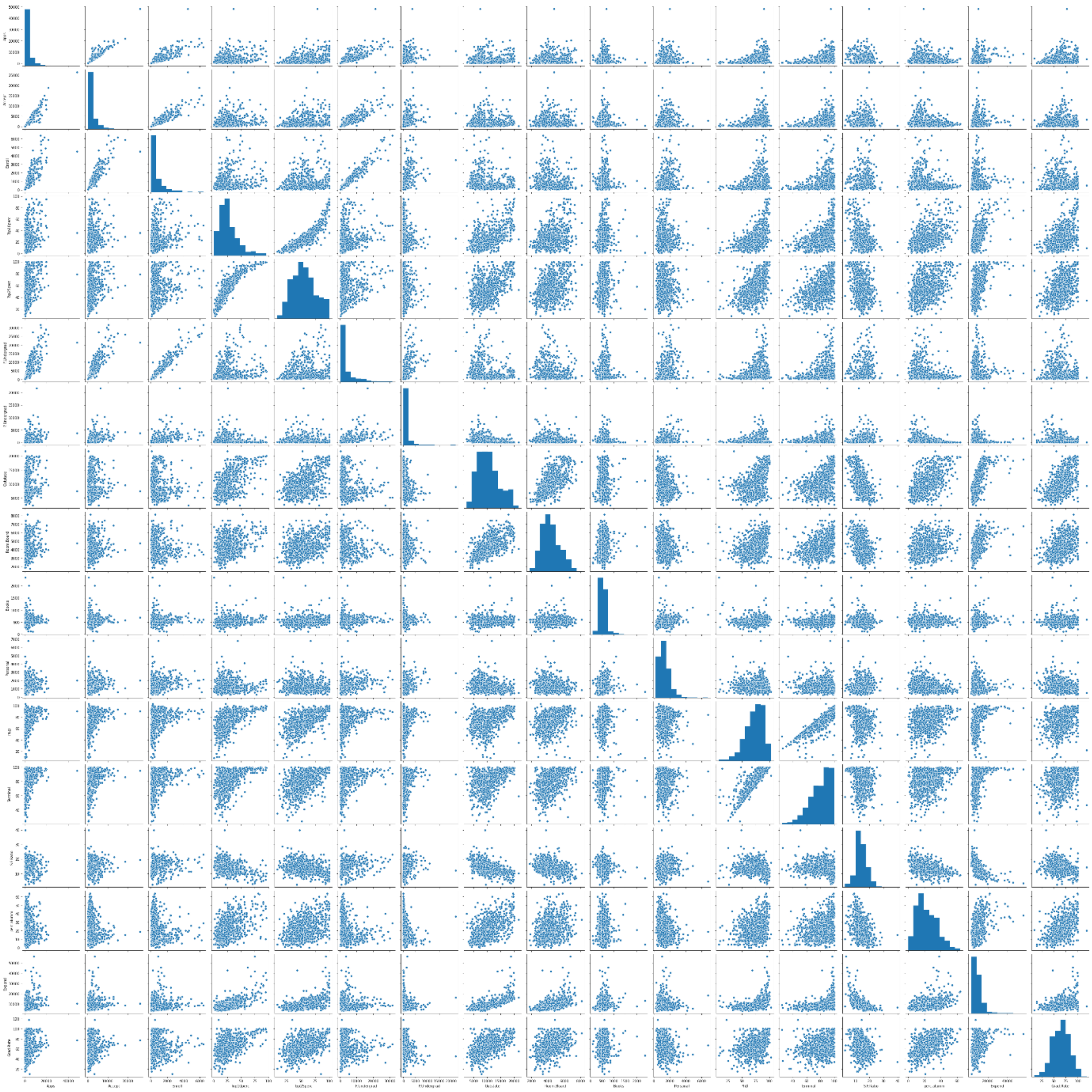
Multivariate Analysis:

A pairplot made for all variable is below:

Following variable pairs – show linear relationship.

Variables	Interpretation
Outstate Vs Expend	An increase in no. of outstation students shows increase in expenditure
Apps Vs Accept	The acceptance number increased with no. of applications
Accept Vs Enroll	The enrollment number increased with no. of acceptance
AppsVs F.undergrat	The application numbers and the numbers of Full time graduates are propotional
Accept Vs F.undergrat	The acceptance rate and the number of Full time graduates are linearly propotional
Enroll Vs F.undergrat	The enrollment rate and the number of Full time graduates are linearly propotional

Top 10% Vs phd	The students from Top 10 pct Higher sec school was propotional to number of PHD teachers (they were probably allocated to such students conciously)
Top 25% Vs phd	The students from Top 25 pct Higher sec school was propotional to number of PHD teachers (they were probably allocated to such students conciously)
Room.board Vs expend	The Room expenses directly correlated to expenses
Terminal Vs Phd	Terminal(last degree) was sharply directly propotional to PhD(indicating PHD was highest degree)



2.2. IS SCALING NECESSARY FOR PCA IN THIS CASE? GIVE JUSTIFICATION AND PERFORM SCALING.

SOLUTION: YES. SCALING IS NECESSARY AS THERE ARE 16 INT VARIABLES – they are at different scales – some represent number of applications, enrollments etc while some others represent the expenditure and few fields on percent of graduates.

Some of the popular Scaling methods that can be used are : Simple Scalar , Min-Max (adjusts for outliers) or z-score.

In our solution we replace all numeric fields with their **z-score.** This is a widely accepted standards for PCA.

It scales the data in such a way that the mean value of the features tends to 0 and the standard deviation tends to 1.

2.3. COMMENT ON THE COMPARISON BETWEEN THE COVARIANCE AND THE CORRELATION MATRICES FROM THIS DATA.[ON SCALED DATA].

Below table shows the covariance and correlation Table for raw-data – with outliers and without 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 1.497846e+07 8.949860e+06 3.045256e+06 23132.773138 26952.663479
1.528970e+07 2.346620e+06 7.809704e+05 7.000729e+05 84703.752639
4.683468e+05 24689.433666 21053.067602 1465.060576 -4327.122381
5.246171e+06 9756.421641

Accept 8.949860e+06 6.007960e+06 2.076268e+06 8321.124872 12013.404757
1.039358e+07 1.646670e+06 -2.539623e+05 2.443471e+05 45942.807867
3.335566e+05 14238.201489 12182.093828 1709.838189 -4859.487022
1.596272e+06 2834.162918

Enroll 3.045256e+06 2.076268e+06 8.633684e+05 2971.583415 4172.592435
4.347530e+06 7.257907e+05 -5.811885e+05 -4.099706e+04 17291.199742
1.767380e+05 5028.961166 4217.086027 872.684773 -2081.693787
3.113454e+05 -356.587977

Top10perc 2.313277e+04 8.321125e+03 2.971583e+03 311.182456 311.630480
1.208911e+04 -2.829475e+03 3.990718e+04 7.186706e+03 346.177405 -
1.114551e+03 153.184870 127.551581 -26.874525 99.567208 6.087931e+04
149.992164

Top25perc 2.695266e+04 1.201340e+04 4.172592e+03 311.630480 392.229216
1.915895e+04 -1.615412e+03 3.899243e+04 7.199904e+03 377.759266 -
1.083605e+03 176.518449 153.002612 -23.097199 102.550946 5.454648e+04
162.371398

F.Undergrad 1.528970e+07 1.039358e+07 4.347530e+06 12089.113681 19158.952782
2.352658e+07 4.212910e+06 -4.209843e+06 -3.664582e+05 92535.764728
1.041709e+06 25211.784197 21424.241746 5370.208581 -13791.929691
4.724040e+05 -6563.307527

P.Undergrad 2.346620e+06 1.646670e+06 7.257907e+05 -2829.474981 -1615.412144
4.212910e+06 2.317799e+06 -1.552704e+06 -1.023919e+05 20410.446674
3.297324e+05 3706.756219 3180.596615 1401.302563 -5297.337090 -
6.643512e+05 -6721.062488

Outstate 7.809704e+05 -2.539623e+05 -5.811885e+05 39907.179832 38992.427500 -
4.209843e+06 -1.552704e+06 1.618466e+07 2.886597e+06 25808.242145 -8.146737e+05
25157.515051 24164.147673 -8835.253539 28229.553066 1.413324e+07
39479.681796

Room.Board 7.000729e+05 2.443471e+05 -4.099706e+04 7186.705605 7199.903568 -
3.664582e+05 -1.023919e+05 2.886597e+06 1.202743e+06 23170.313390 -1.480838e+05
5895.034749 6047.299735 -1574.205914 3701.431379 2.873308e+06
8005.360183

Books 8.470375e+04 4.594281e+04 1.729120e+04 346.177405 377.759266
9.253576e+04 2.041045e+04 2.580824e+04 2.317031e+04 27259.779946
2.004303e+04 72.534242 242.963918 -20.867207 -82.263132
9.691258e+04 3.008837

Personal 4.683468e+05 3.335566e+05 1.767380e+05 -1114.551186 -1083.605065
1.041709e+06 3.297324e+05 -8.146737e+05 -1.480838e+05 20043.025650
4.584258e+05 -120.898783 -305.154186 365.415770 -2399.310824 -
3.460978e+05 -3132.614944

PhD 2.468943e+04 1.423820e+04 5.028961e+03 153.184870 176.518449
2.521178e+04 3.706756e+03 2.515752e+04 5.895035e+03 72.534242 -
1.208988e+02 266.608636 204.231332 -8.436492 50.383230 3.689806e+04
85.557109

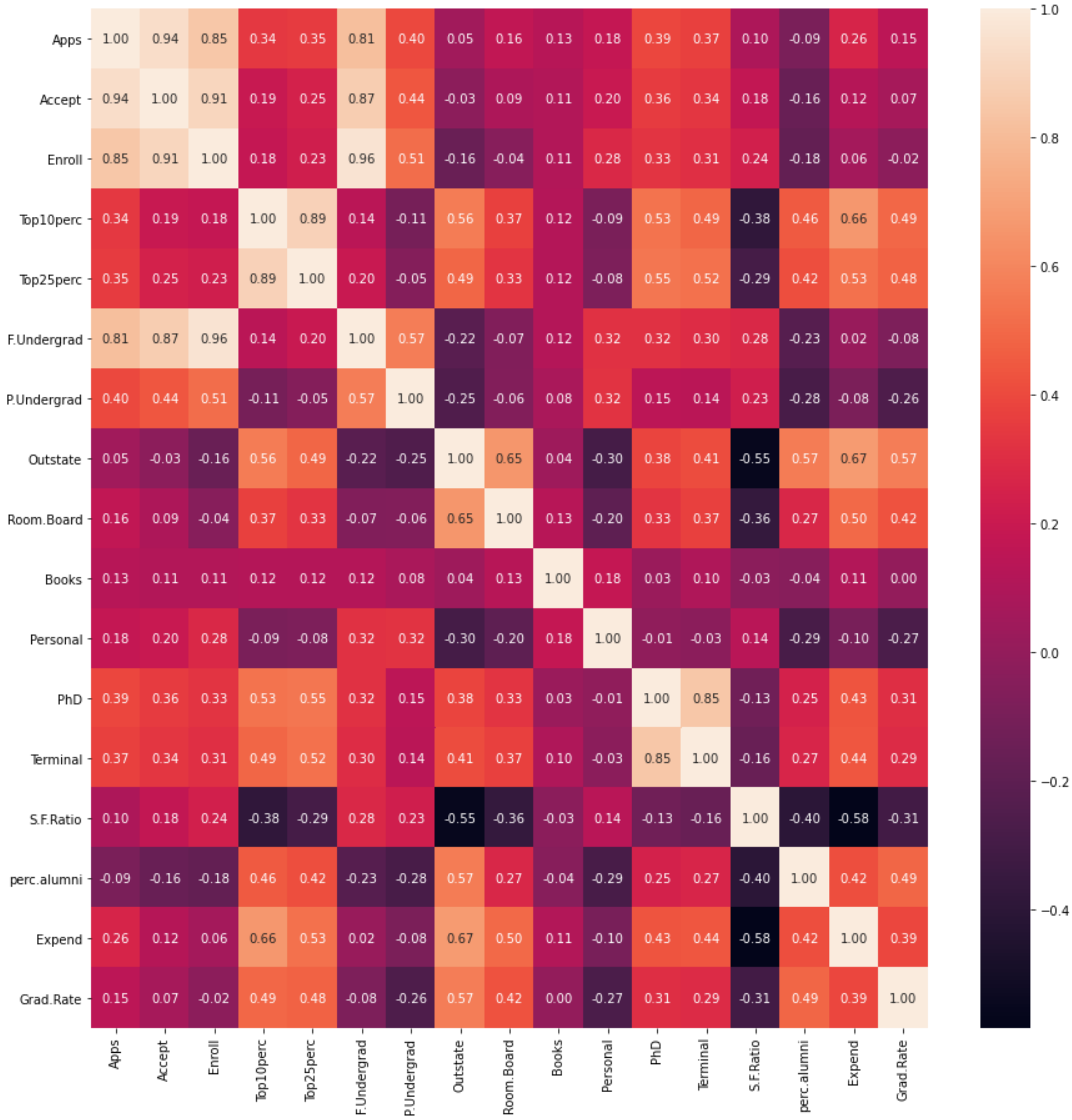
Terminal 2.105307e+04 1.218209e+04 4.217086e+03 127.551581 153.002612
2.142424e+04 3.180597e+03 2.416415e+04 6.047300e+03 242.963918 -
3.051542e+02 204.231332 216.747841 -9.330256 48.734327 3.373346e+04
73.220396

S.F.Ratio 1.465061e+03 1.709838e+03 8.726848e+02 -26.874525 -23.097199
5.370209e+03 1.401303e+03 -8.835254e+03 -1.574206e+03 -20.867207
3.654158e+02 -8.436492 -9.330256 15.668528 -19.764109 -
1.206756e+04 -20.854888

perc.alumni -4.327122e+03 -4.859487e+03 -2.081694e+03 99.567208 102.550946 -
1.379193e+04 -5.297337e+03 2.822955e+04 3.701431e+03 -82.263132 -2.399311e+03
50.383230 48.734327 -19.764109 153.556744 2.702892e+04 104.493815

Expend 5.246171e+06 1.596272e+06 3.113454e+05 60879.310196 54546.483305
4.724040e+05 -6.643512e+05 1.413324e+07 2.873308e+06 96912.580326 -
3.460978e+05 36898.058233 33733.456882 -12067.564601 27028.921473 2.726687e+07
35012.968271

Grad.Rate 9.756422e+03 2.834163e+03 -3.565880e+02 149.992164 162.371398 -
6.563308e+03 -6.721062e+03 3.947968e+04 8.005360e+03 3.008837 -3.132615e+03
85.557109 73.220396 -20.854888 104.493815 3.501297e+04 295.073717



Some of the specific High correlation fields are as below:

Set 1 of highly correlated variables: 'Apps','Accept','Enroll','F.Undergrad', 'P.Undergrad'

Set 2 of highly correlated variables : 'Room.Board','Outstate','Expend', 'Books'

	Top10perc	Top25perc	Grad.Rate	PhD	Terminal
Top10perc	1.000000	0.891995	0.494989	0.531828	0.491135
Top25perc	0.891995	1.000000	0.477281	0.545862	0.524749
Grad.Rate	0.494989	0.477281	1.000000	0.305038	0.289527
PhD	0.531828	0.545862	0.305038	1.000000	0.849587
Terminal	0.491135	0.524749	0.289527	0.849587	1.000000

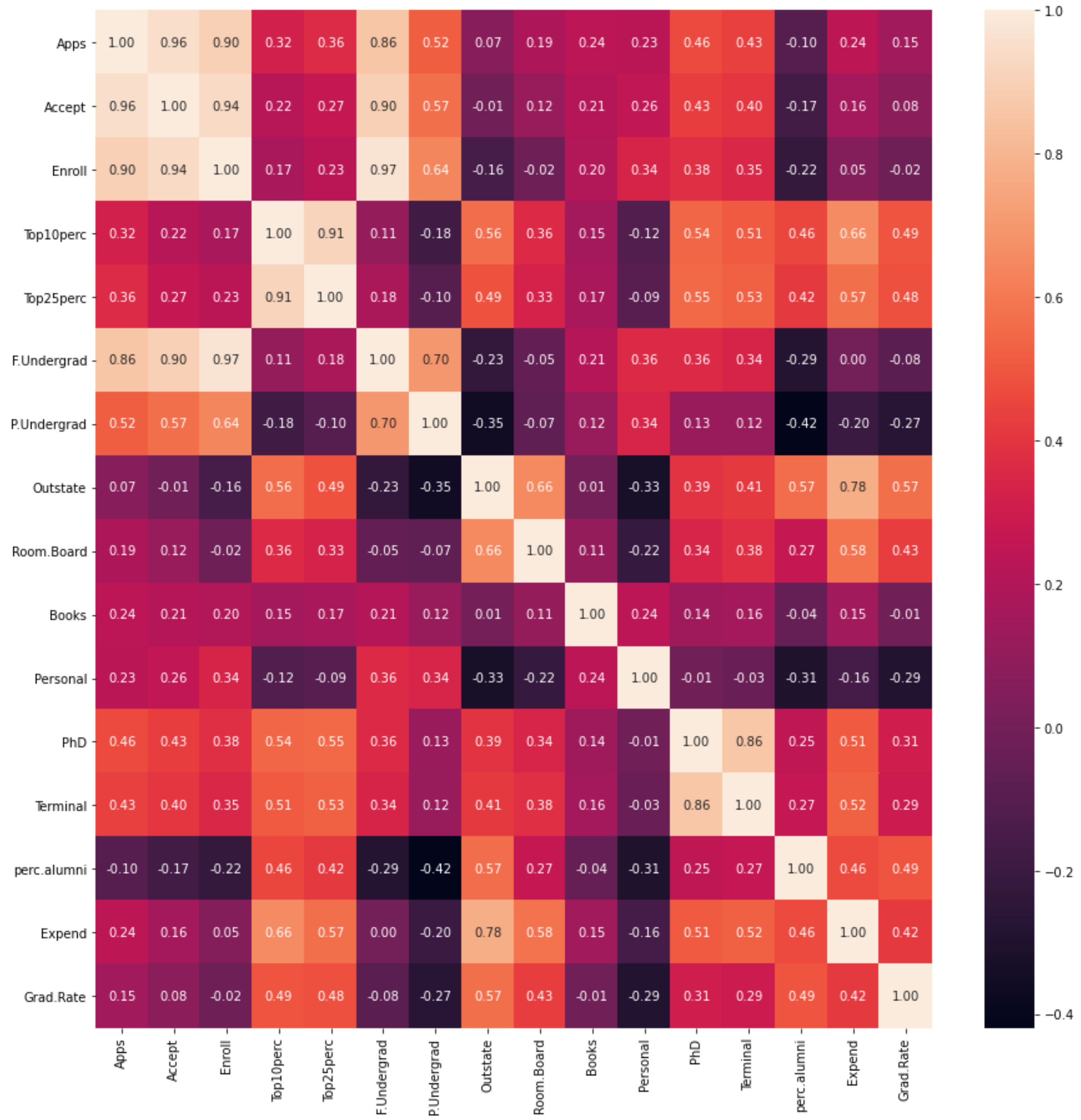
	Room.Board	Outstate	Expend	Books	Personal
Room.Board	1.000000	0.654256	0.501739	0.127963	-0.199428
Outstate	0.654256	1.000000	0.672779	0.038855	-0.299087
Expend	0.501739	0.672779	1.000000	0.112409	-0.097892
Books	0.127963	0.038855	0.112409	1.000000	0.179295
Personal	-0.199428	-0.299087	-0.097892	0.179295	1.000000

AFTER SCALING:

BELOW IS THE COVARIANCE AND CORRELATION AFTER Z-SCORE SCALING AND REMOVING OUTLIERS.

THE **COVARIANCE ON SCALED DATA BECOMES THE CORRELATION.**

THERE IS NO SIGNIFICANT DIFFENECE IN CORRELATION/COVARIANCE AFTER SCALING.



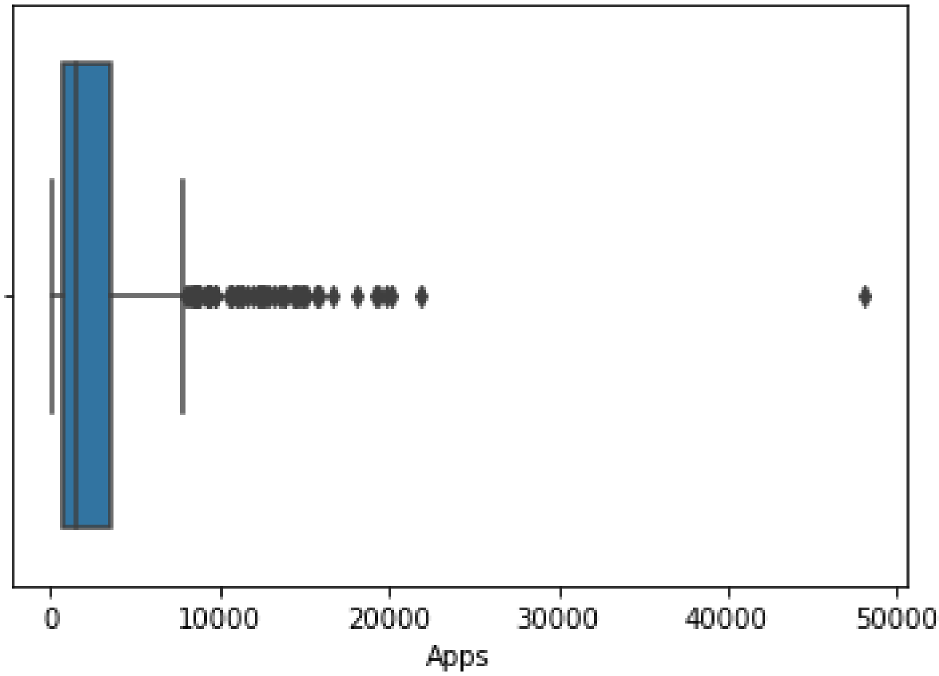
2.4. CHECK THE DATASET FOR OUTLIERS BEFORE AND AFTER SCALING. WHAT INSIGHT DO YOU DERIVE HERE?

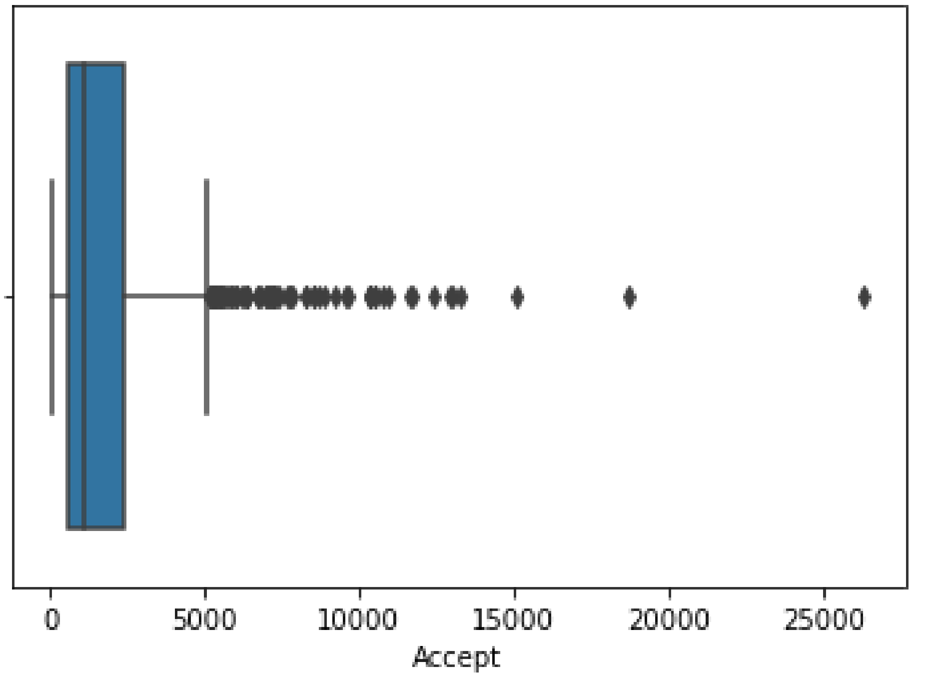
The following is the Boxplot for all integer variables before scaling.

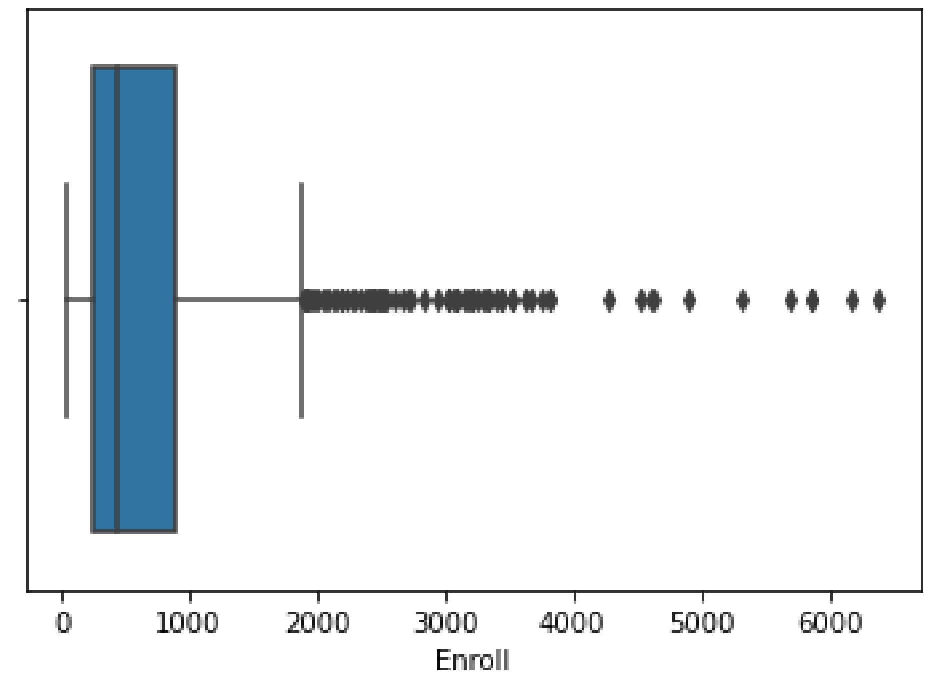
There are several outliers in the data.

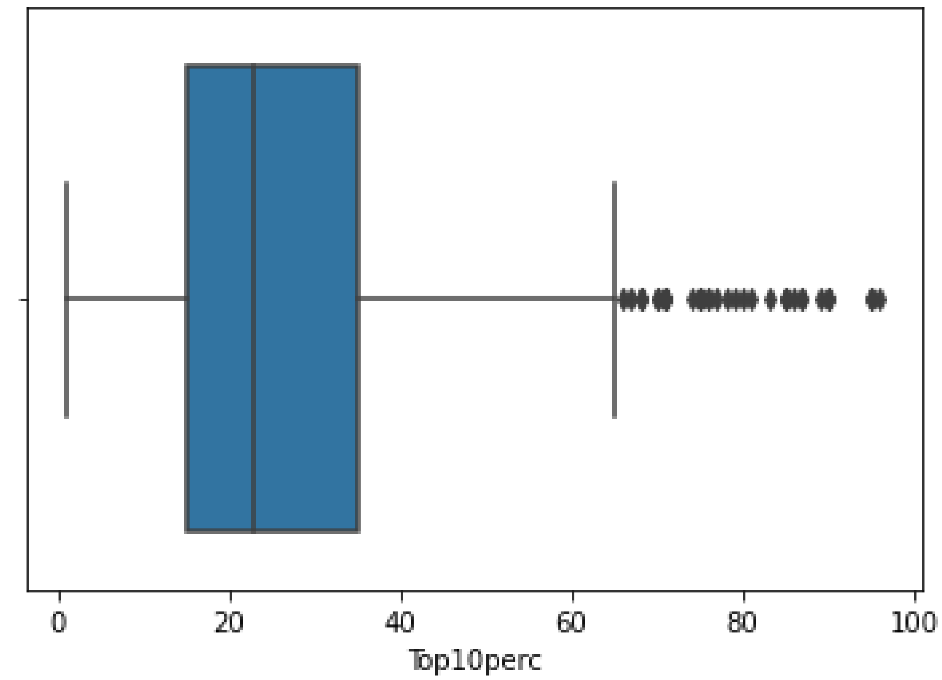
Adjusting for outliers by replacing the Max values with upper whisker (Q3 + 1.5 IQR) and Min value with (Q1 + 1.5 IQR) lower whisker removes outliers.

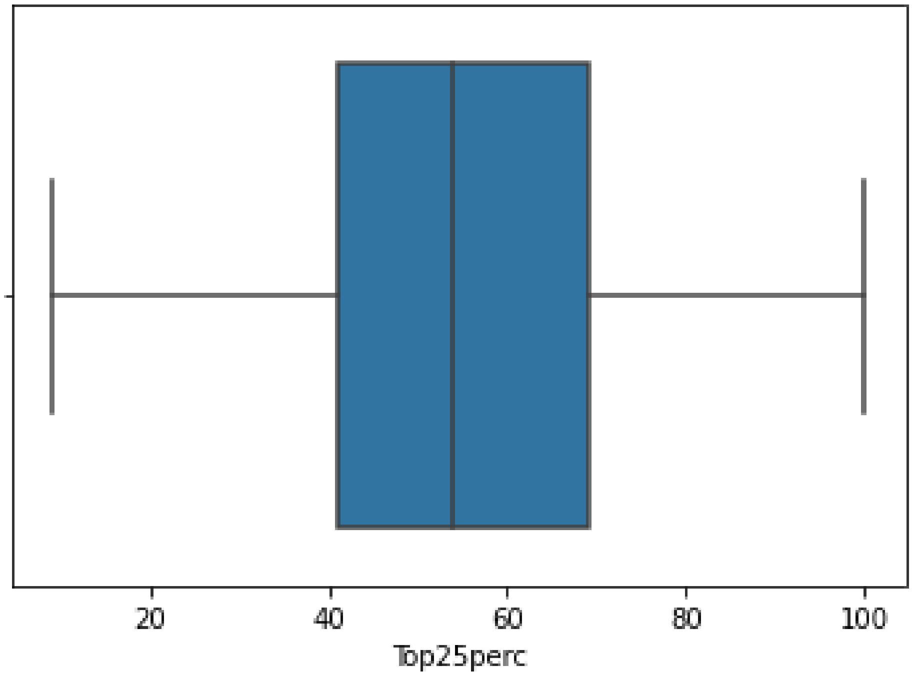
<u>Box-Plot before scaling:</u>

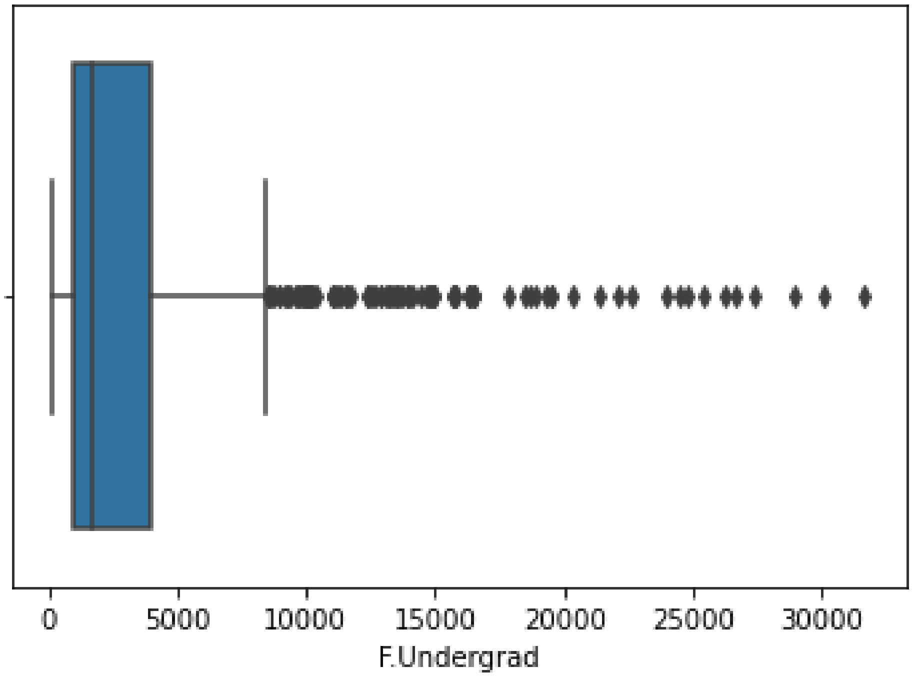

0
10000
20000
30000
40000
50000
Apps

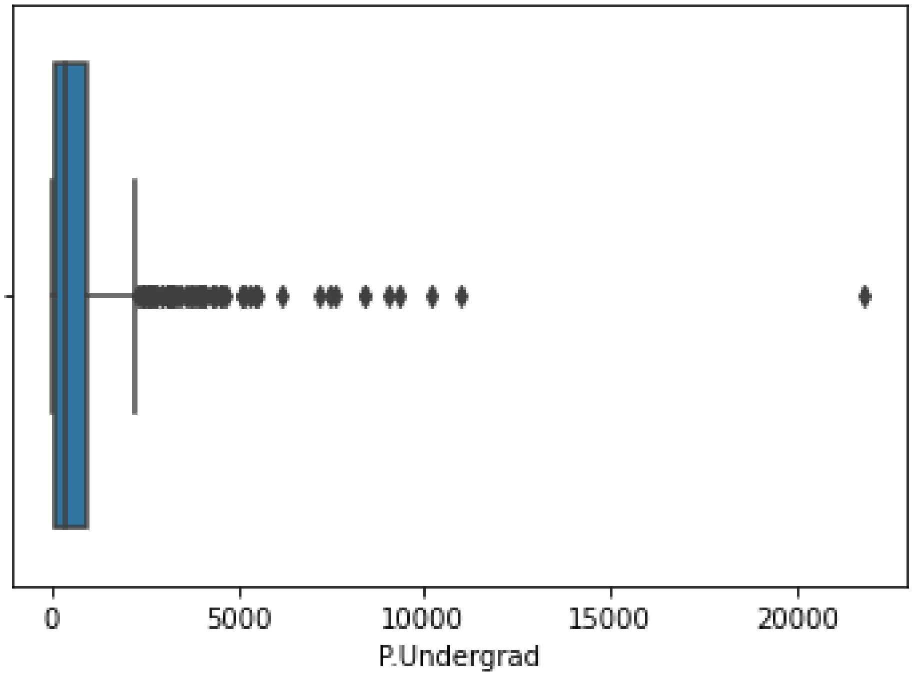

0
5000
10000
15000
20000
25000
Accept

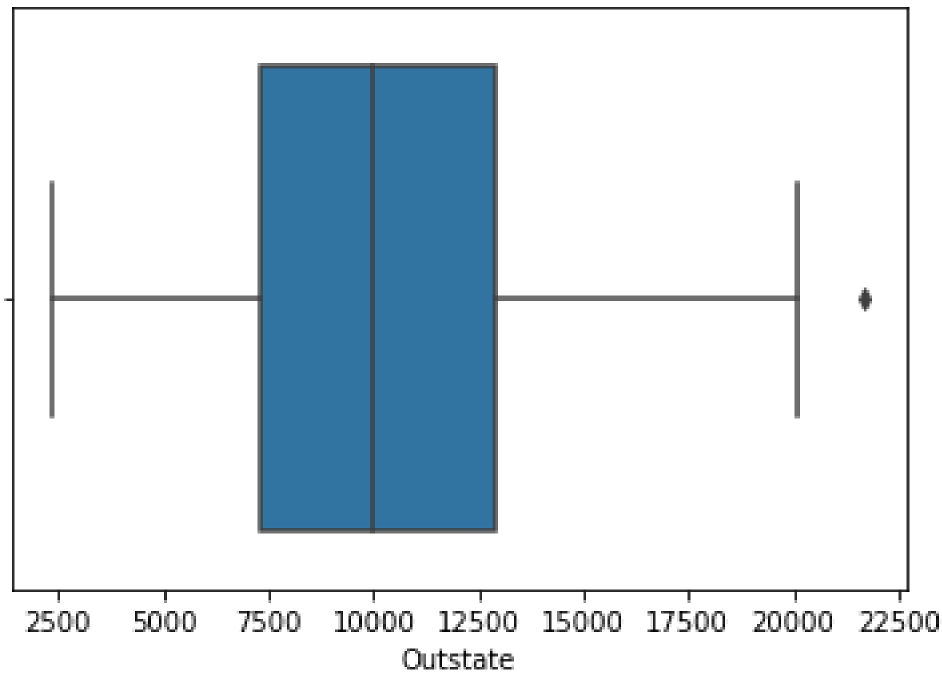

0
1000
2000
3000
4000
5000
6000
Enroll

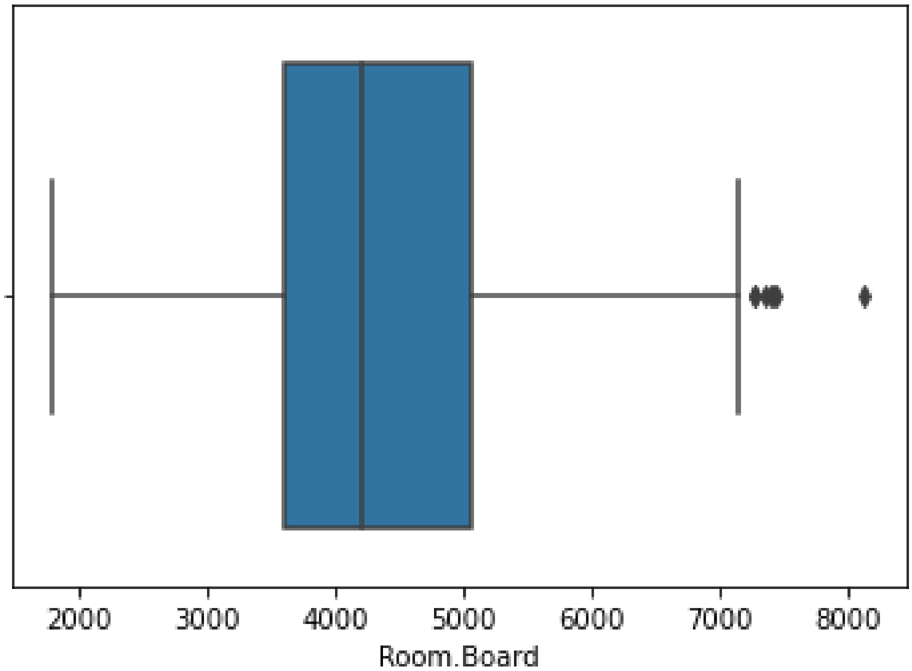

0
20
40
60
80
100
Top10perc

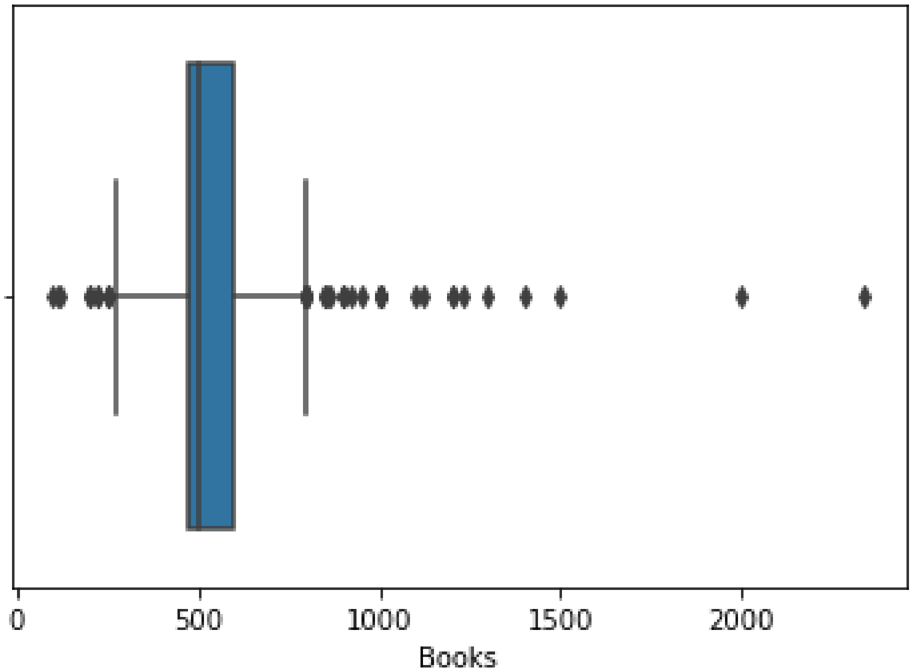

20
40
60
80
100
Top25perc

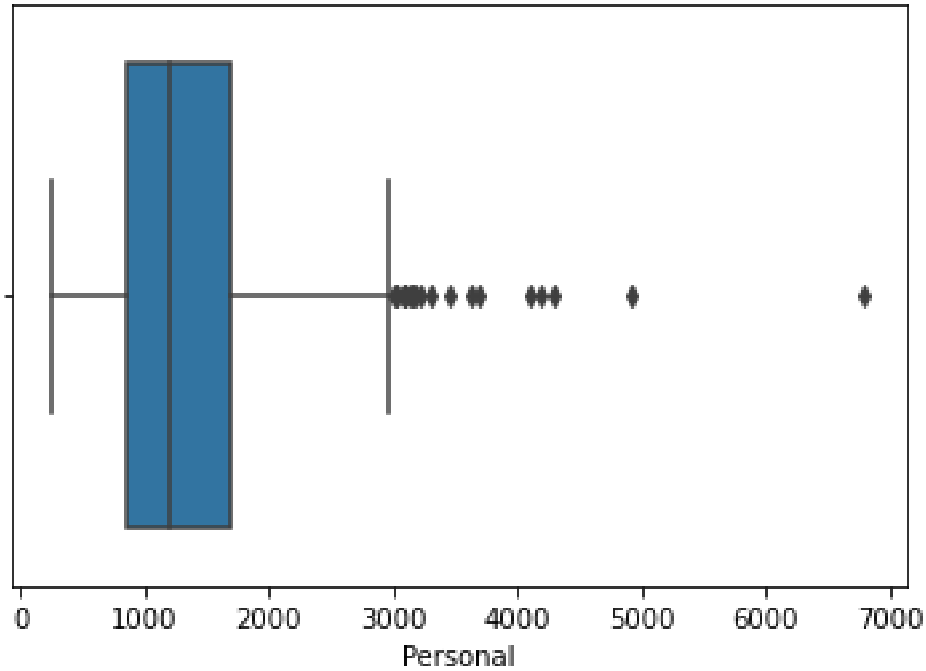

0
5000
10000
15000
20000
25000
30000
F.Undergrad

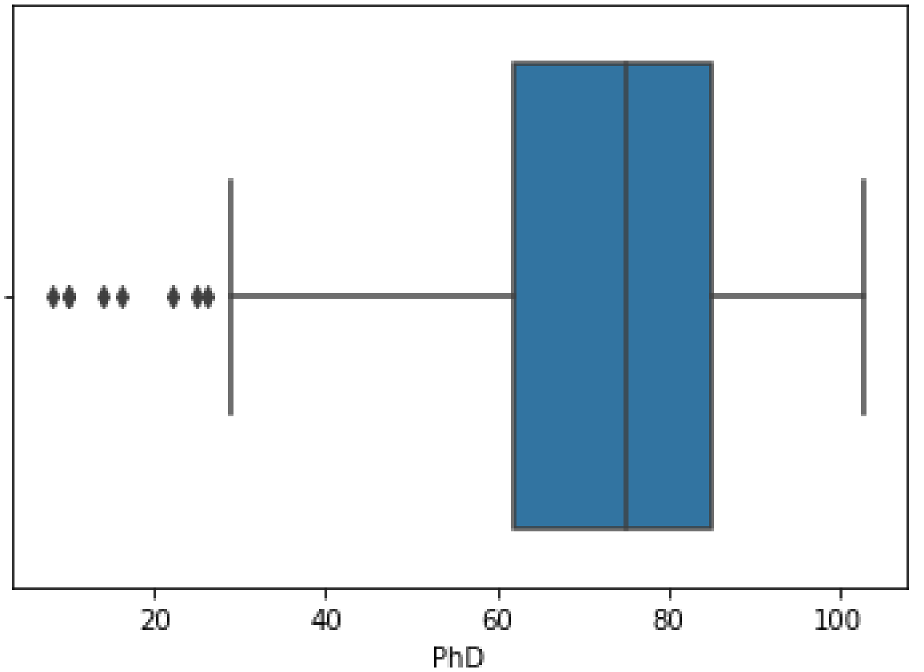

0
5000
10000
15000
20000
P.Undergrad

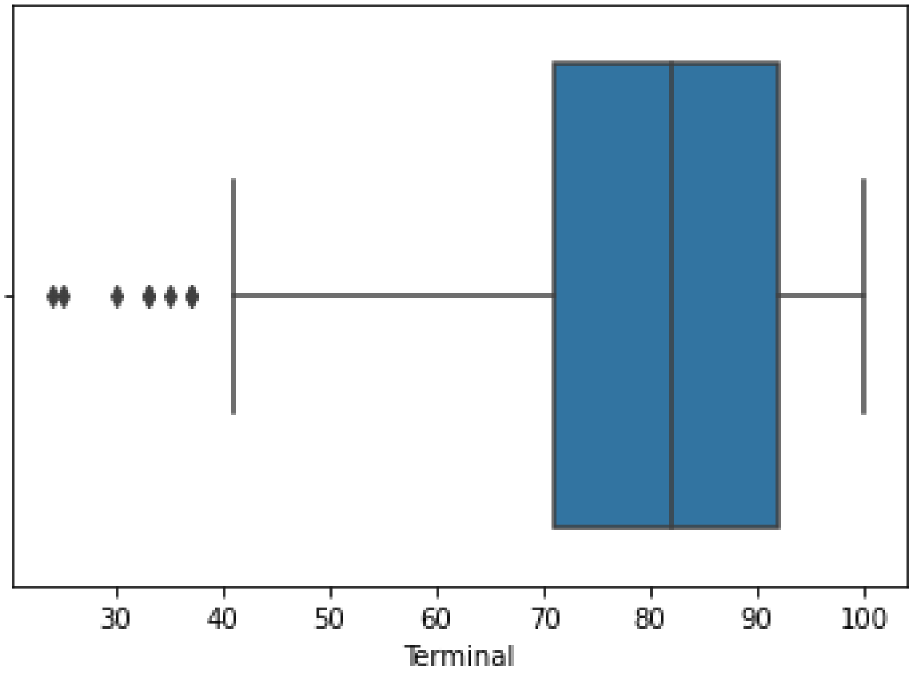

2500
5000
7500
10000
12500
15000
17500
20000
22500
Outstate

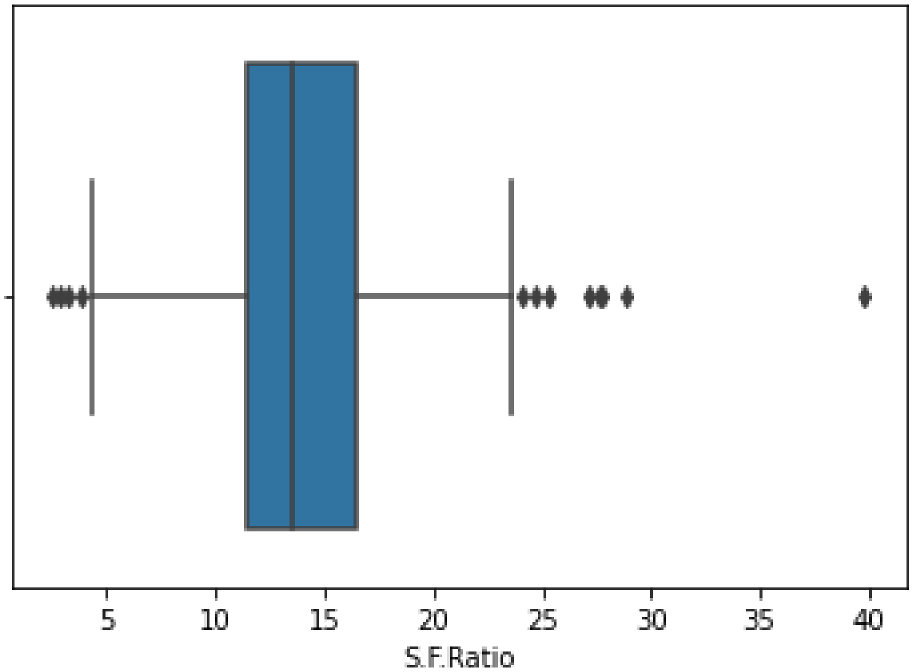

2000
3000
4000
5000
6000
7000
8000
Room.Board

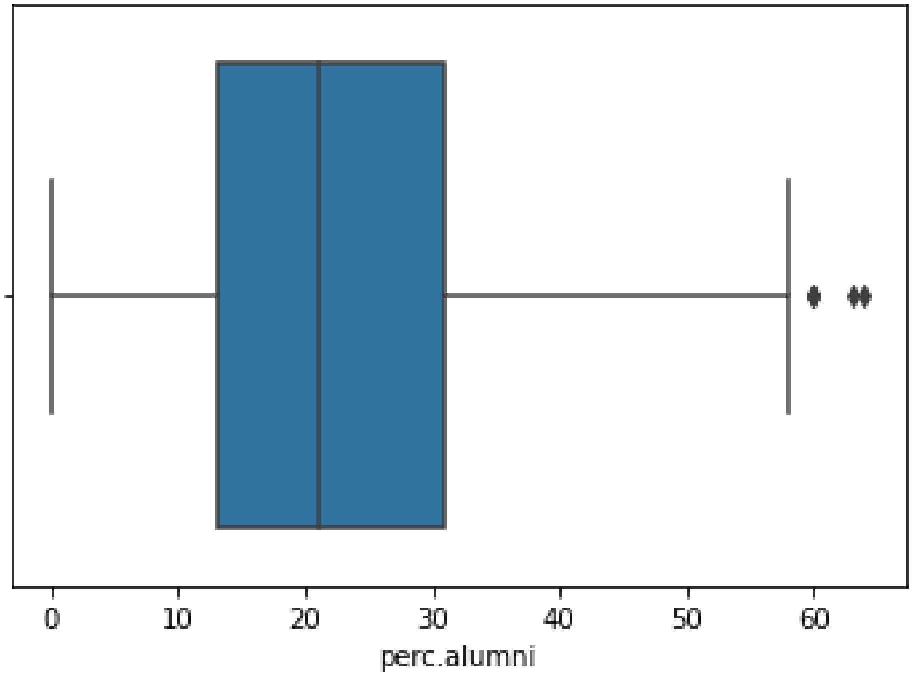

0
500
1000
1500
2000
Books

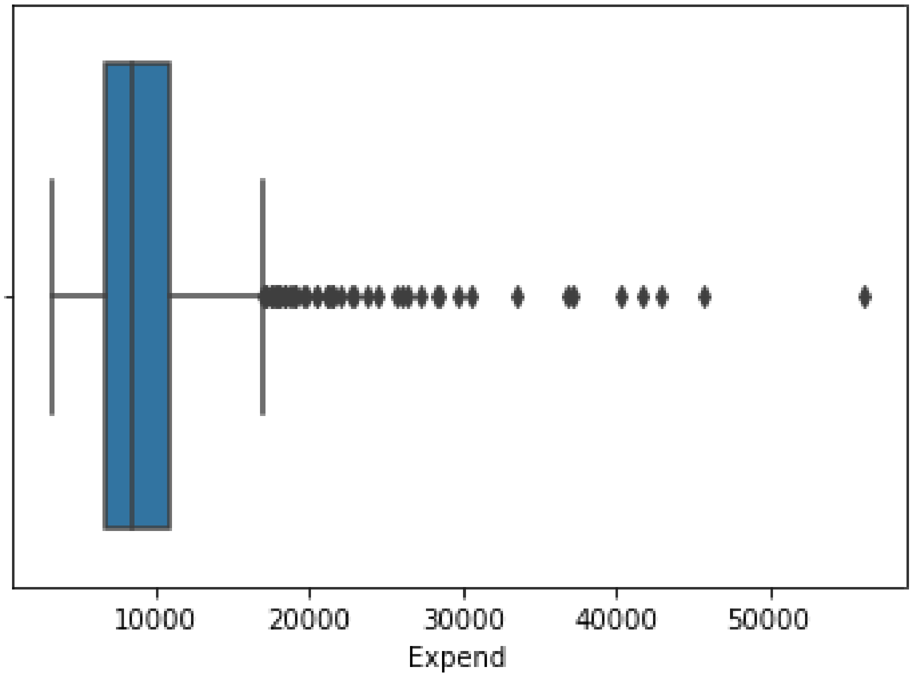

0
1000
2000
3000
4000
5000
6000
7000
Personal

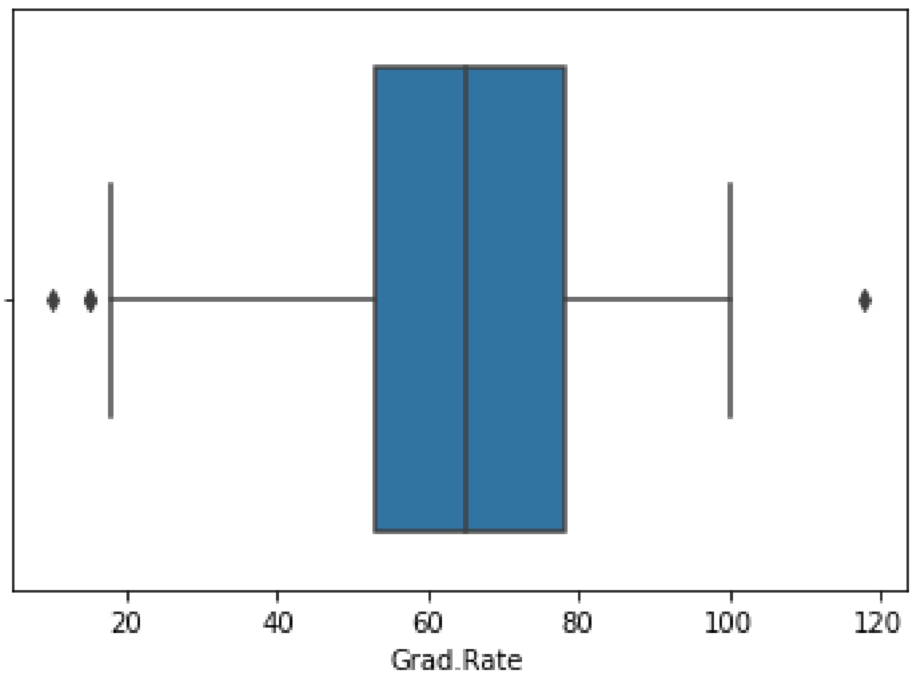

20
40
60
80
100
PhD


30
40
50
60
70
80
90
100
Terminal

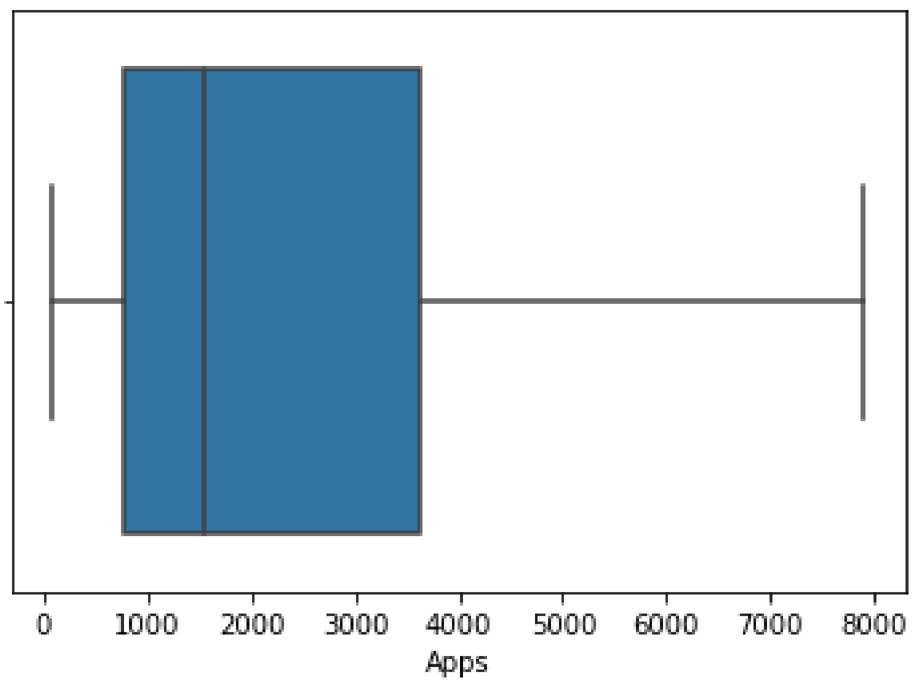

5
10
15
20
25
30
35
40
S.F.Ratio

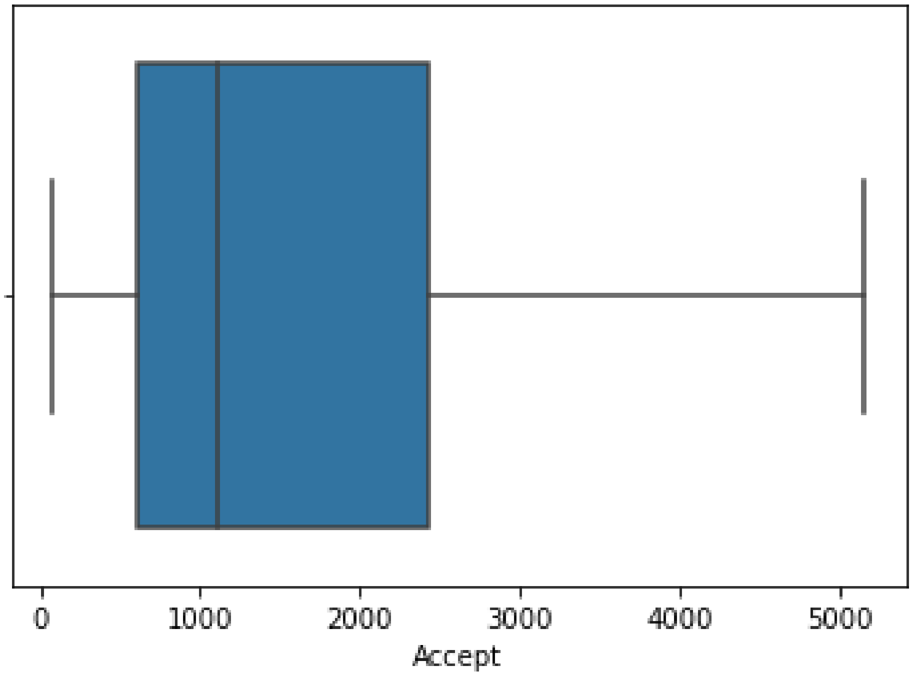

0
10
20
30
40
50
60
perc.alumni

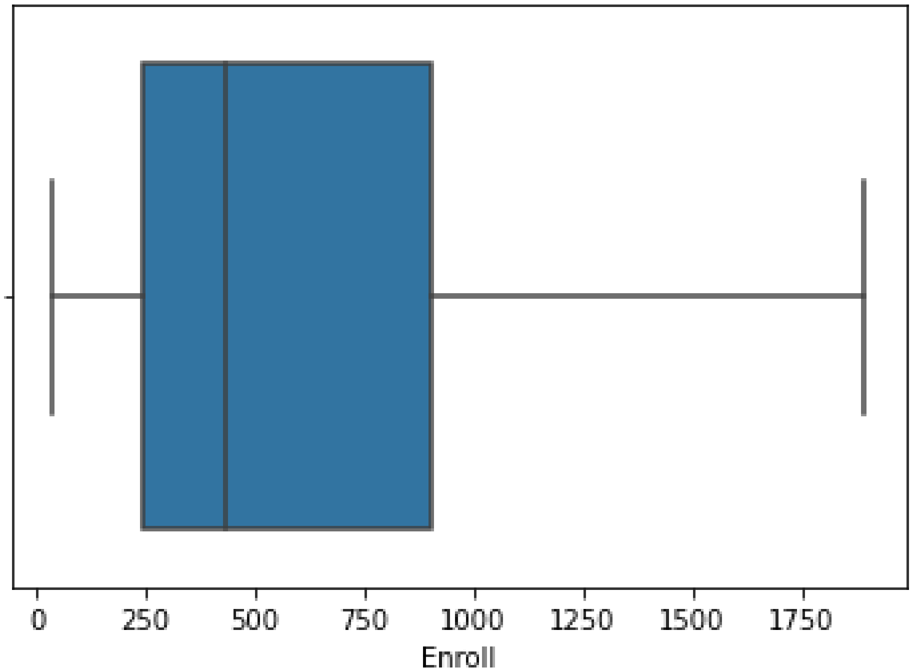

10000
20000
30000
40000
50000
Expend

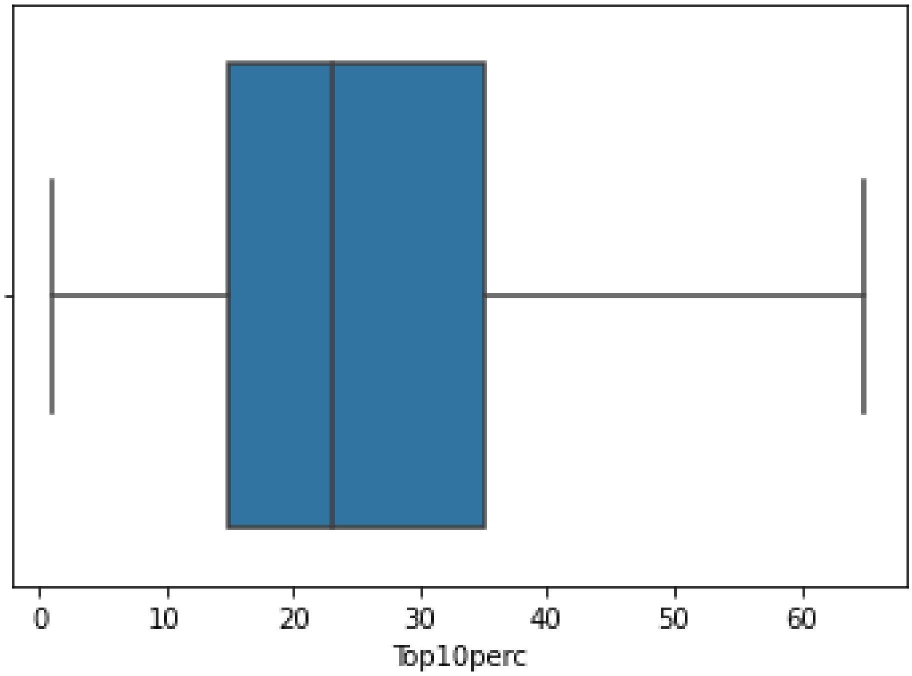


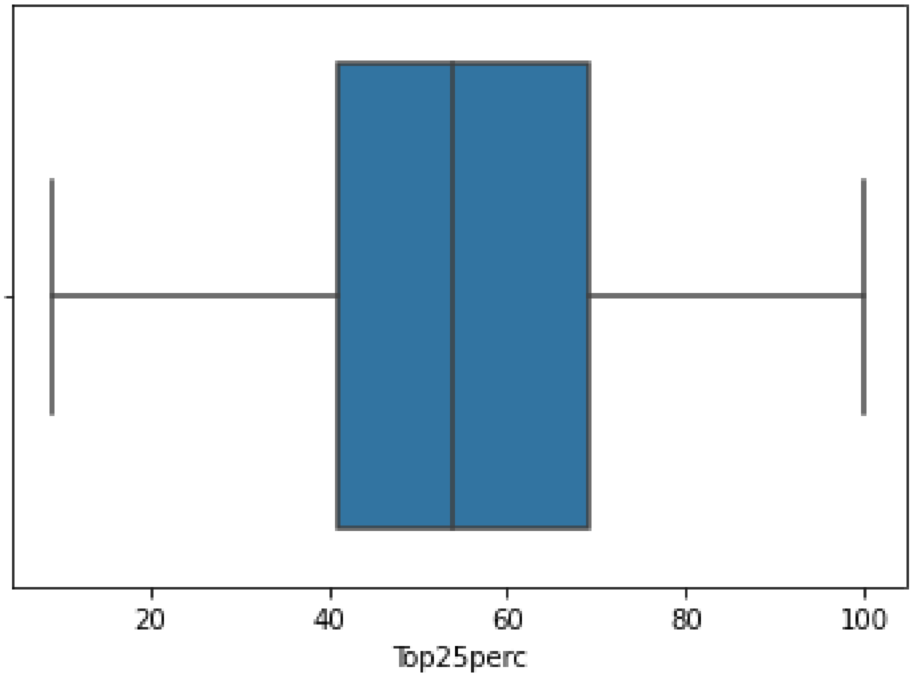
Box-Plots after adjusting for Outliers:

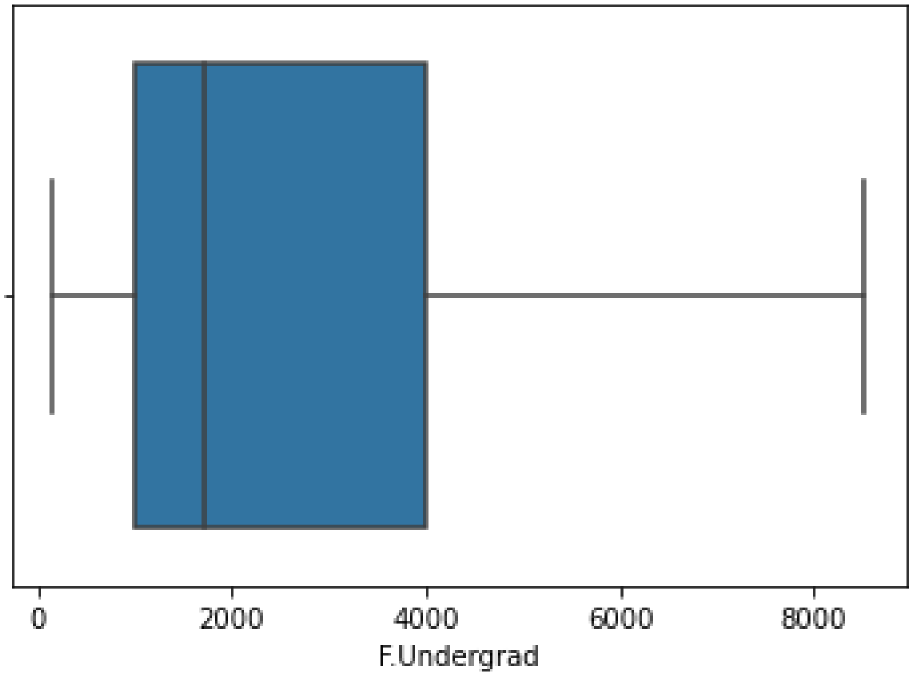


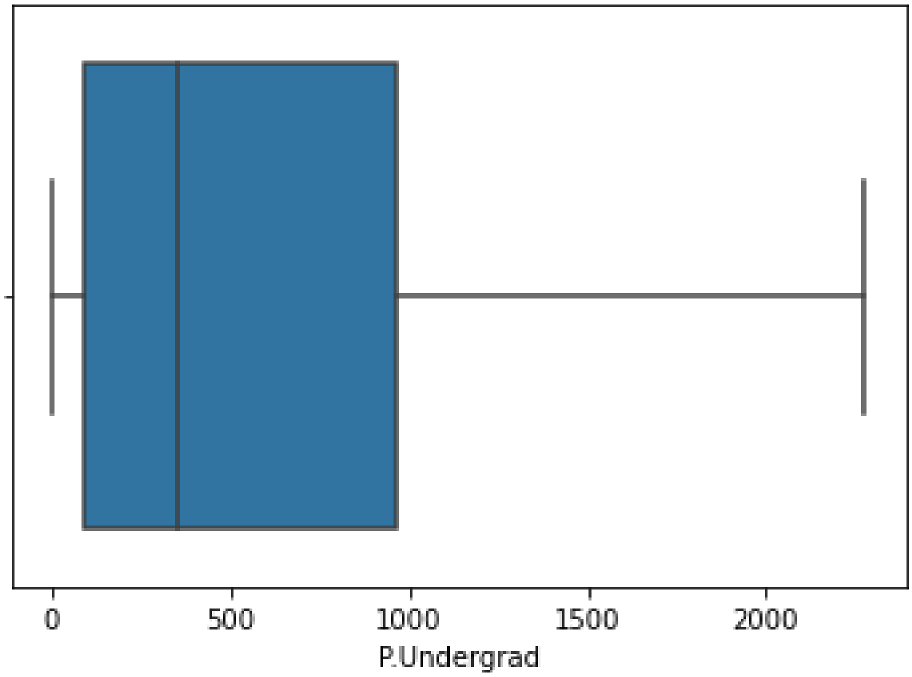

0
1000
2000
3000
4000
5000
Accept

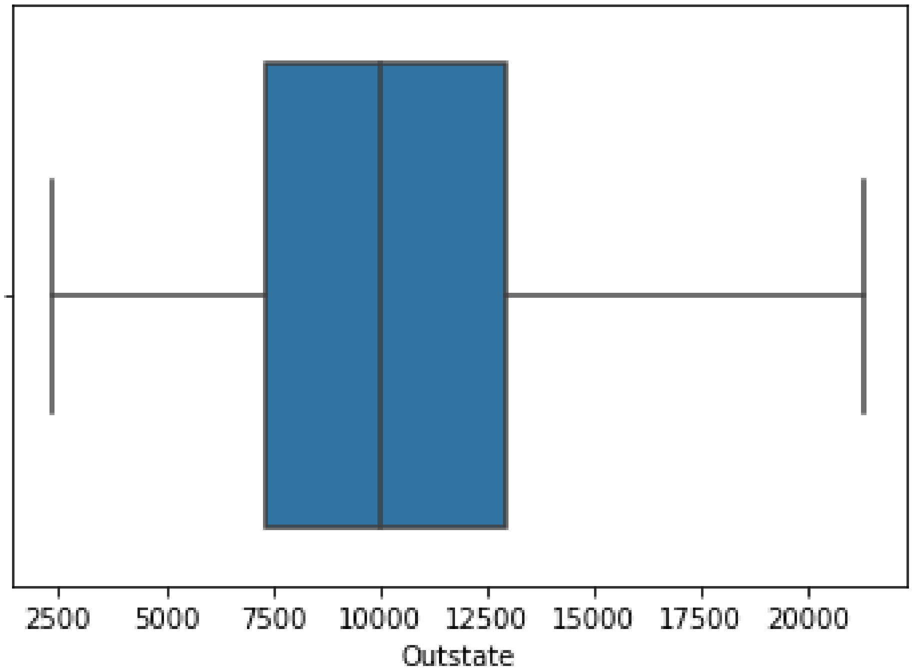

0
250
500
750
1000
1250
1500
1750
Enroll

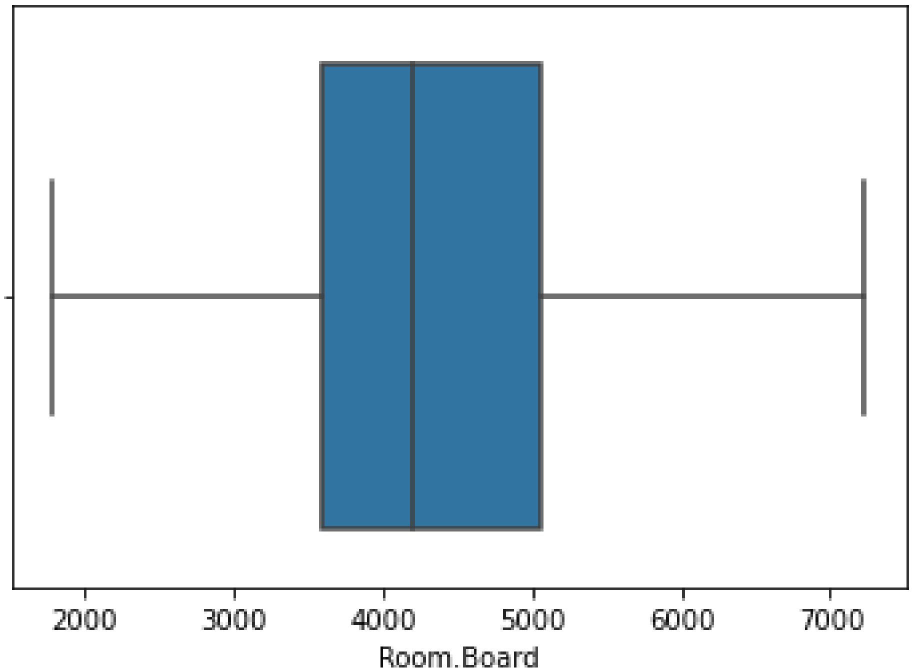

0
10
20
30
40
50
60
Top10perc

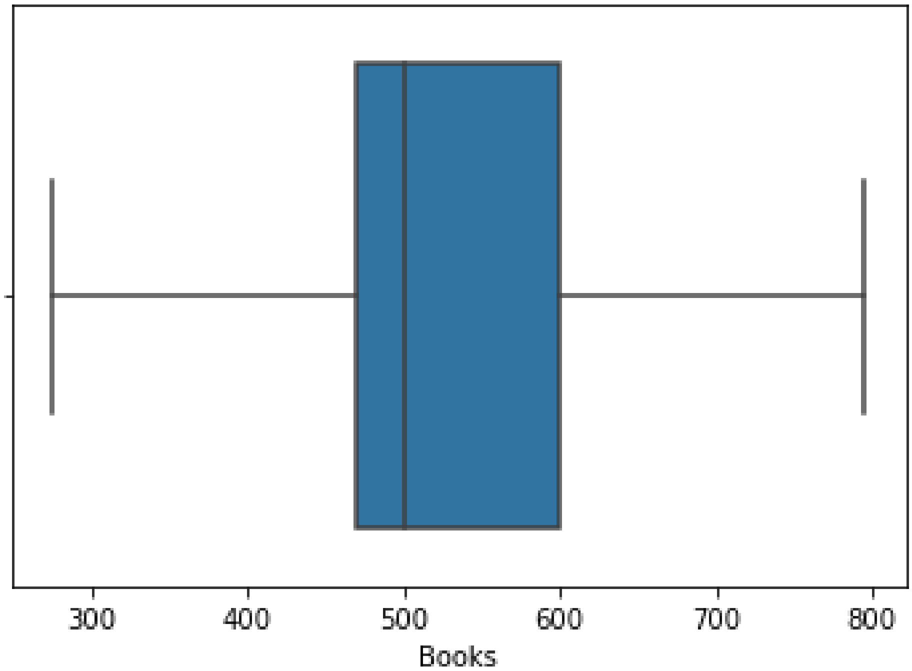

20
40
60
80
100
Top25perc

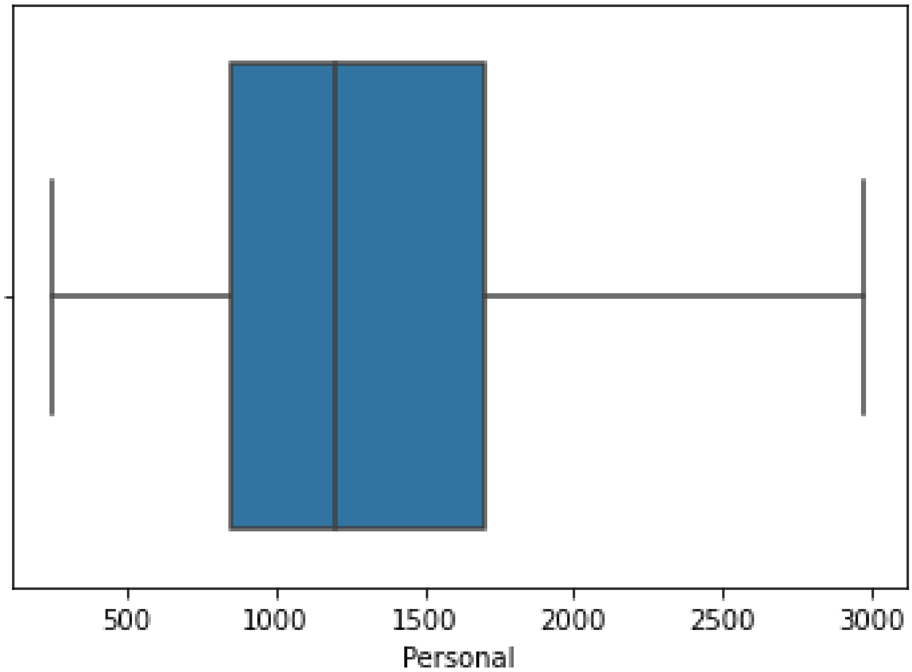

0
2000
4000
6000
8000
F.Undergrad

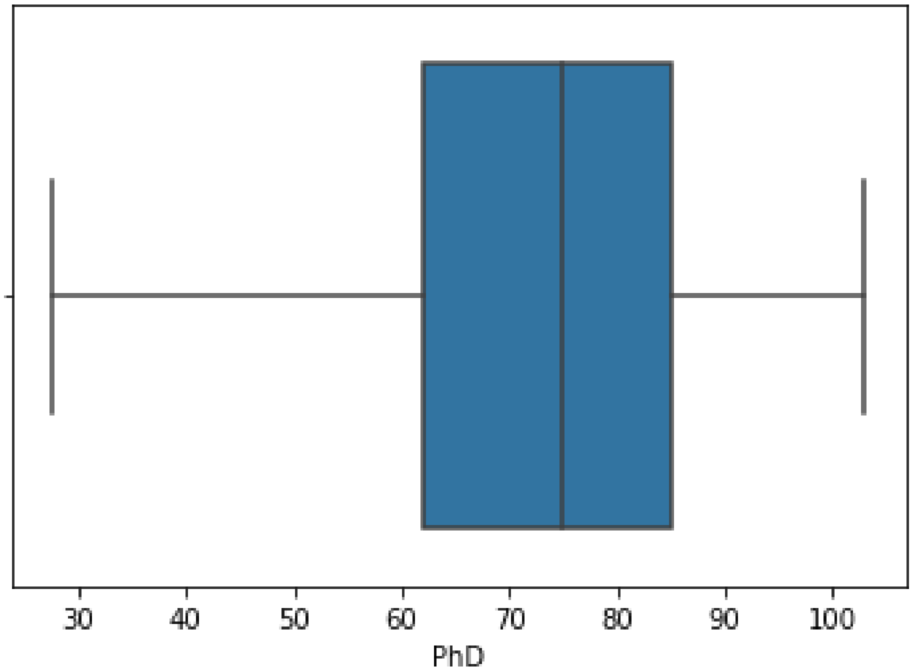

0
500
1000
1500
2000
P.Undergrad

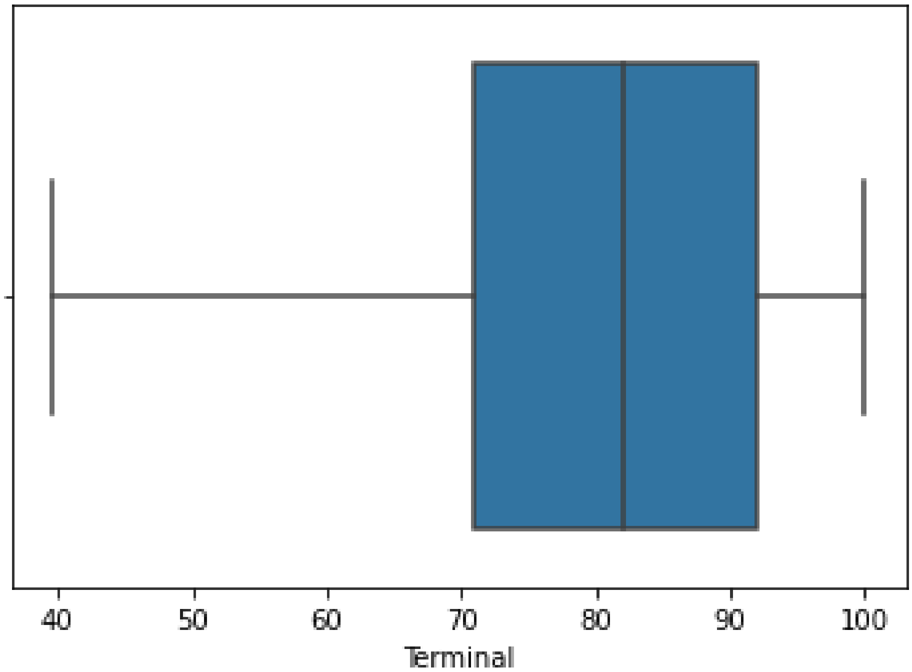

2500
5000
7500
10000
12500
15000
17500
20000
Outstate

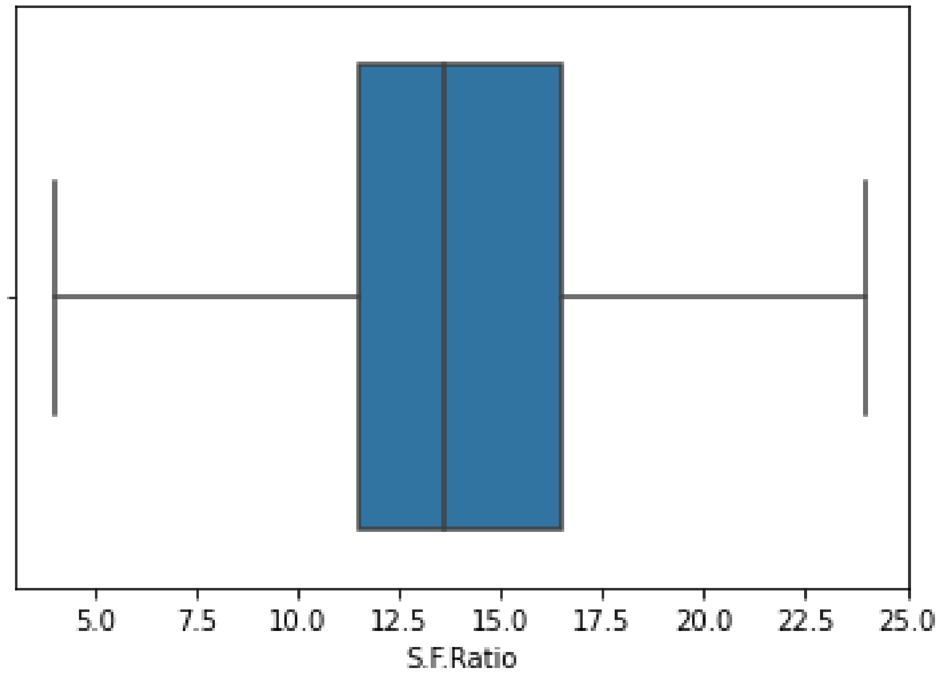

2000
3000
4000
5000
6000
7000
Room.Board

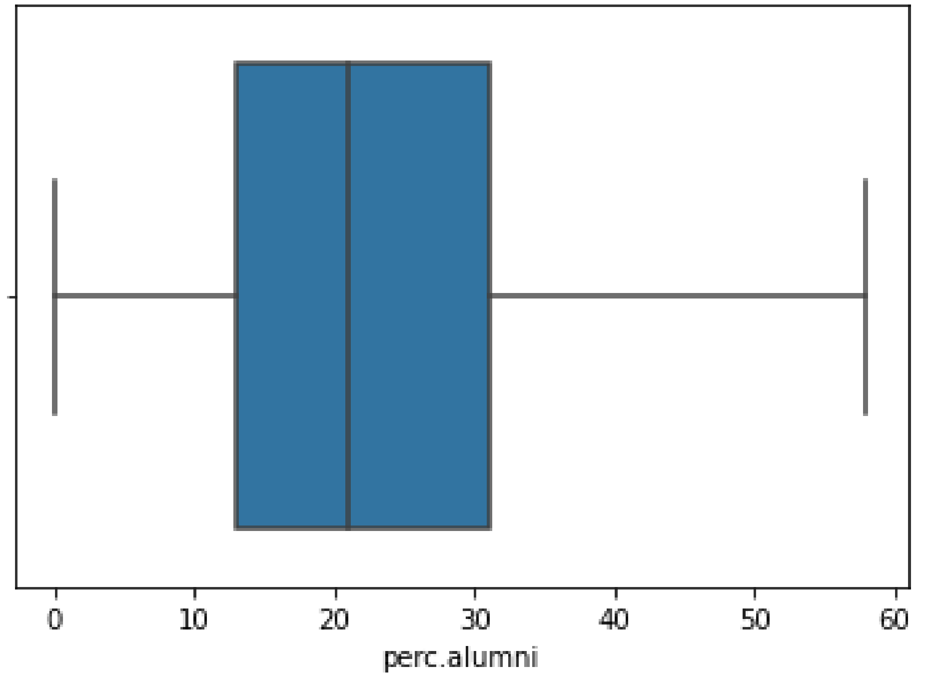

300
400
500
600
700
800
Books

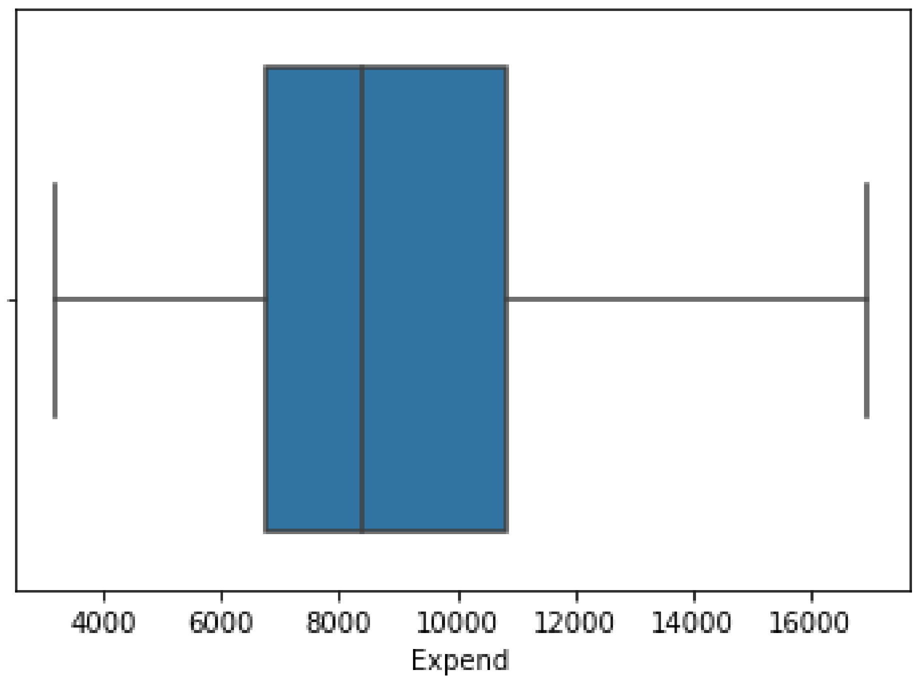

500
1000
1500
2000
2500
3000
Personal

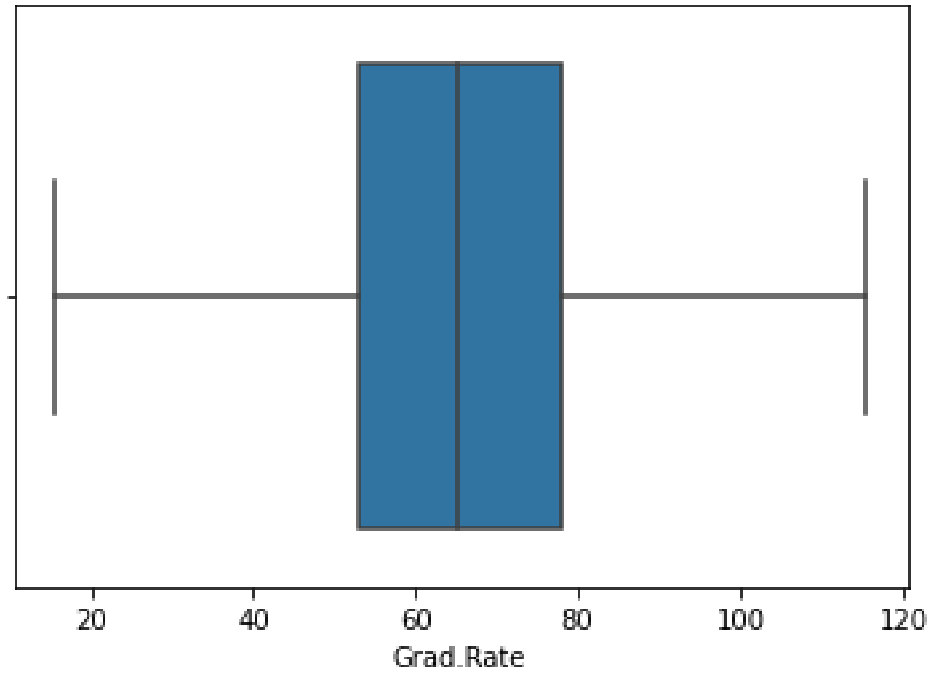

30
40
50
60
70
80
90
100
PhD


40
50
60
70
80
90
100
Terminal


5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
S.F.Ratio


0
10
20
30
40
50
60
perc.alumni





2.5. EXTRACT THE EIGENVALUES AND EIGENVECTORS.[PRINT BOTH]

EIGEN VECTORS:

```
Eigen Vectors
%s [[ 3.17814172e-01  2.61253505e-01  1.25956560e-01 -2.02659037e-02
  -2.13785276e-01  5.09494393e-03 -2.73447826e-03 -1.11637097e-01
  -1.77030063e-01 -1.23575352e-01  1.84110996e-01 -5.99971829e-01
  -4.87133853e-02 -7.06078585e-02  5.55352927e-01 -3.95889770e-03]
 [ 2.92315664e-01  2.99321411e-01  1.60243348e-01 -2.79594866e-02
  -1.78571638e-01  1.07523051e-02 -3.73351890e-02 -1.41061491e-01
  -1.88987598e-01 -8.90716147e-02 -3.93425293e-01  6.60457280e-01
  -1.23962993e-01 -1.96584922e-02  3.07570126e-01 -1.45901270e-02]
 [ 2.58850257e-01  3.46898358e-01  1.13697029e-01  7.87065035e-02
  -1.27108233e-01  4.67087824e-03 -4.00284817e-02 -1.36117130e-01
  -7.73009395e-02 -4.41861897e-02  7.16937431e-01  2.39144525e-01
   3.36346969e-02  5.42114993e-02 -4.14379846e-01  4.82679874e-02]
 [ 3.16599334e-01 -1.78582523e-01 -1.83242981e-01  3.48227421e-01
  -1.04606454e-01 -8.00076650e-02  3.74997172e-01  8.98908162e-02
   1.09456040e-01 -3.67406193e-02 -5.68085940e-02  2.09753184e-02
   1.26218784e-02 -7.57220290e-02  2.32823380e-03  7.23940334e-01]
 [ 3.20861634e-01 -1.40230929e-01 -1.77114455e-01  3.84014882e-01
  -9.79152138e-02 -2.54501770e-02  3.60898100e-01  1.88530825e-01
   2.06643913e-01 -1.48418667e-01  2.01109558e-02  3.36076369e-02
  -1.13022087e-01  1.14797372e-01  1.23989286e-03 -6.55974108e-01]
 [ 2.36348432e-01  3.68172581e-01  9.99029714e-02  5.79713465e-02
  -7.39896854e-02  2.65839910e-02 -1.92282699e-02 -8.98461954e-02
  -7.39932829e-03 -1.16424466e-02 -5.40066143e-01 -3.67027764e-01
   1.09610419e-01  6.43306226e-02 -5.80695926e-01 -2.63077132e-02]
 [ 8.88648643e-02  3.60184404e-01  1.18358177e-01 -1.89371030e-01
   8.30286874e-02 -1.94799589e-02  4.83987906e-02  1.45932136e-01
   7.69196536e-01  3.98309572e-01  2.95389969e-02  2.65058857e-02
  -6.44171208e-02 -3.72659320e-02  1.50230289e-01  3.98989652e-02]
 [ 2.31872820e-01 -3.16439306e-01  7.18312295e-02 -2.64880914e-01
  -8.03766666e-02 -2.11569337e-01 -1.12364200e-02 -1.66920903e-01
  -9.10828853e-02  2.87318733e-01  1.69583120e-03 -7.99534563e-02
  -7.46526619e-01 -1.50028963e-02 -1.91827380e-01  5.46475817e-04]
 [ 2.15986364e-01 -1.87127280e-01  1.00804509e-01 -6.50716853e-01
  -7.81239720e-02 -1.41056084e-01  1.06989319e-01  2.19190118e-01
   1.70557528e-01 -5.73971064e-01  9.88332370e-03  2.72636076e-02
   1.66982702e-01 -2.75940596e-02 -1.02207716e-01  2.79891853e-02]
 [ 1.12196803e-01  7.74326943e-02 -7.30623688e-01 -2.83581077e-01
  -2.15785264e-01  5.47169431e-01 -9.99200107e-02 -4.58210055e-02
   1.09645487e-02  6.15042143e-02  4.68424624e-03  1.13474344e-02
  -5.98171876e-02 -2.95201999e-02 -2.41212595e-02  7.78110527e-03]
 [ 5.85034366e-03  2.54481517e-01 -5.07909665e-01  3.35832619e-02
   4.39203463e-02 -7.20502530e-01 -3.37202094e-01  1.83689918e-01
  -4.73422462e-02 -5.80198275e-02 -1.15214494e-02  3.15110328e-03
  -2.83149719e-02  1.14717763e-02  3.03300820e-02 -1.10547667e-03]
 [ 3.34901874e-01 -2.41464824e-02 -1.38829609e-03  5.40643633e-02
   5.54359149e-01  1.28616688e-01 -1.43927243e-01  1.43032345e-01
  -1.10215433e-01  2.62118958e-03  1.43925051e-02  1.51281130e-02
   8.85807340e-03 -7.05035211e-01 -5.01800528e-02 -8.39148988e-02]
 [ 3.29129870e-01 -3.48931907e-02 -1.01396266e-02 -1.96939301e-02
   5.77514829e-01  1.58196495e-01 -1.70624077e-01  8.90283238e-02
  -6.58000855e-02 -3.80456586e-02  7.88216419e-03 -1.70859686e-02
  -4.54822637e-02  6.80612870e-01  1.00404528e-01  1.13412683e-01]
 [ 1.37165124e-01 -3.06986928e-01  2.27098552e-02  2.36751710e-01
  -1.00310808e-01  9.55228462e-03 -5.62948261e-01 -4.87032307e-01
   4.46868149e-01 -2.44147454e-01 -2.44696068e-02 -1.05464227e-03
```

```
   6.22141835e-02 -4.04534304e-02  2.54921083e-02  7.54978787e-03]
 [ 2.90799895e-01 -2.22651235e-01 -1.00105123e-01 -2.11867513e-01
   2.74998024e-02 -2.41500719e-01  1.78965859e-01 -3.70987912e-01
  -1.13806162e-01  4.67225038e-01  3.55410679e-03  4.08808051e-02
   5.69944557e-01  3.58966824e-02  5.66286984e-02 -1.41230995e-01]
 [ 2.06628476e-01 -2.31197517e-01  1.97097008e-01  7.25989726e-02
  -4.00800954e-01  9.24261073e-02 -4.35183246e-01  6.00960018e-01
  -1.08169618e-01  3.06876947e-01 -2.97216391e-03  1.38133028e-02
   1.77797701e-01  4.50178181e-02 -1.69575329e-02  3.63263587e-03]]
```

EIGEN VALUES:

```
Eigen Values
%s [5.56822194 4.56403917 1.09258336 0.96486806 0.86410796 0.64817632
 0.57464166 0.50191334 0.41905686 0.28891907 0.02242577 0.03802336
 0.1644268  0.13465697 0.09986333 0.0746946 ]
```

2.6 PERFORM PCA AND EXPORT THE DATA OF THE PRINCIPAL COMPONENT (EIGENVECTORS) INTO A DATA FRAME WITH THE ORIGINAL FEATURES

Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad
	Outstate	Room.Board	Books	Personal	PhD	Terminal
	perc.alumni	Expend	Grad.Rate			
0	0.317814	0.292316	0.258850	0.316599	0.320862	0.236348
	0.088865	0.231873	0.215986	0.112197	0.005850	0.334902
	0.329130	0.137165	0.290800	0.206628		
1	0.261254	0.299321	0.346898	-0.178583	-0.140231	0.368173
	0.360184	-0.316439	-0.187127	0.077433	0.254482	-0.024146
	-0.034893	-0.306987	-0.222651	-0.231198		
2	-0.125957	-0.160243	-0.113697	0.183243	0.177114	-0.099903
	-0.118358	-0.071831	-0.100805	0.730624	0.507910	0.001388
	0.010140	-0.022710	0.100105	-0.197097		
3	0.020266	0.027959	-0.078707	-0.348227	-0.384015	-0.057971
	0.189371	0.264881	0.650717	0.283581	-0.033583	-0.054064
	0.019694	-0.236752	0.211868	-0.072599		
4	0.213785	0.178572	0.127108	0.104606	0.097915	0.073990
	-0.083029	0.080377	0.078124	0.215785	-0.043920	-0.554359
	-0.577515	0.100311	-0.027500	0.400801		

2.7 WRITE DOWN THE EXPLICIT FORM OF THE FIRST PC (IN TERMS OF THE EIGENVECTORS. USE VALUES WITH TWO PLACES OF DECIMALS ONLY).

First PC:

```
[[ 0.32  0.26  0.13 -0.02 -0.21  0.01 -0.   -0.11 -0.18 -0.12  0.18 -0.
6
  -0.05 -0.07  0.56 -0.  ]
```

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?

The table below shows cumulative values:

```
Eigen Values
%s [5.56822194 4.56403917 1.09258336 0.96486806 0.86410796 0.64817632
 0.57464166 0.50191334 0.41905686 0.28891907 0.02242577 0.03802336
 0.1644268  0.13465697 0.09986333 0.0746946 ]


Cumulative Variance Explained [ 34.75659769  63.24513048  70.06498798
76.08765221  81.48137628
  85.52726451  89.11415261  92.24707374  94.86280831  96.66622848
  97.69257335  98.53309624  99.15643876  99.6226792   99.86001933
 100.        ]
```

It shows that the cumulative of first 8 PC will account for understanding PCA to the extent of ~92%. Hence we consider only 8 PCs to be fed for ML.

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]

PCA is a widely used multivariate data analysis method. It is particularly useful for data with collinearity and more variables than samples. In this example through multivariate exploration we found correlation between Apps/Accept etc and again for Expend/Food etc..besides other fields.

Based on the original variables, PCA calculates a set of new variables that describes as much as possible of the variance in the data. The new 'variables' are named principal components (PCs). The PCs will be ranked according to how much of the original variance they explain: PC1 will explain the most variance, PC2 the second most and so on.

Calculation of PCs may be done with several methods. Here we use Eigen decomposition on covariance matrix. The number of PCs to include in a given case can be based on a criterion for the explained variance. This is calculated for each PCA. A criterion of >90% is normally the default used in the calculation software. Often only one or a few PCs are needed to sufficiently explain the variance in the data, simplifying significantly the evaluation.

REFERENCES:

https://iwaponline.com/ws/article/19/8/2256/69018/Principal-component-analysis-for-decision-support

https://www.researchoptimus.com/article/what-is-anova.php#:~:text=ANOVA%20is%20used%20in%20a,the%20future%20performance%20of%20sales.