ADVANCE STATISTICS

Submitted By: Jyotiranjan Padhiary

MARCH 6, 2022

Contents

ADVANCE STATISTICS

Problem-1

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals. are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

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

Education

H0: The mean salary of Education is same across all levels

Ha: The mean salary of Education is different for at least one level.

Occupation

H0: The mean salary of occupation is same across all types

Ha: The mean salary of occupation is different for at least one type

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.

	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

Confidence interval 95% considered

Here we have found that P<0.05 so we can reject the null hypothesis.

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

	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

Here we have found that P>0.05 so we can accept the null hypothesis.

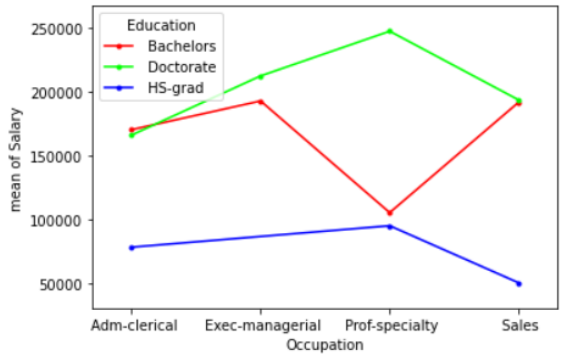
1.4 If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result.

Null hypothesis is rejected in 1.2, and the class means of Education level are significantly different.

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

In the two treatments the mean salary is significantly different in the different levels of Education, but in Occupation it is same. So we would like to see if there is any interaction between this two variable which can be found from interaction plot.

Figure 1 Interaction Plot of Education & Occupation



In *figure 1* we can observe that there is an interaction between Education and Occupation

- For "Doctorate" level of education the salary is highest across all occupation type.
- For "HS-grad" level of education the salary is lowest across all occupation.
- For "Sales" type occupation, only Doctorate or Bachelors are earning more.
- For "Prof-Specialty" type occupation up to Bachelor level education gets the same salary almost.

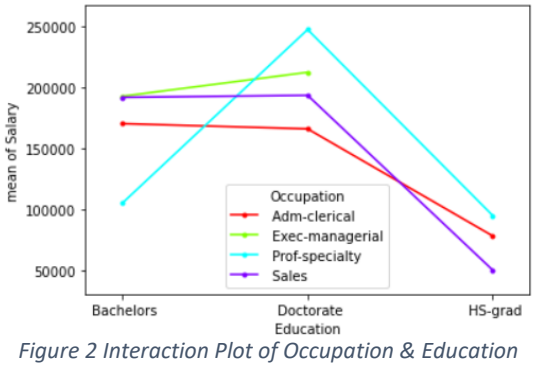


Figure 2 Interaction Plot of Occupation & Education

In *figure 2* there is an interaction observed between Occupation and Education.

- "Exec managerial" type occupations are done by only Bachelors or Doctorate education level people.
- To get better salary in sales one has to be a bachelor level education.
- For a Doctorate level education least salary is offered in Admin-clerical type occupation.
- For a HS-grad best salary is offered in Prof-specialty type occupation.

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

H0: There is no difference in average salary between the different levels of education.

H0: There is no difference in average salary between the different types of occupation.

H0: The effect of education level on average salary does not depend on the effect of occupation type.

Ha: There is a significant difference in the average salary by education level.

Ha: There is a significant difference in average salary by occupation types.

Ha: There is an interaction effect between the education level and occupation type on average salary.

Two Way

	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

P<0.05 for the interaction effect term so we have to reject the H0

Anova Output:

Here we have rejected the null hypothesis and accepted the alternate which means that there is an interaction between Education level and Occupation type on average salary. We can interpret it like, better level of education gives better occupations with better salary.

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

By performing ANOVA, we could now know that whether the different levels of education have some sort of effect on the occupation type and average salary.

So we have found from our analysis that Education level has greater effect on the average salary, compared to occupation type, and if both education and occupation are considered together then they affect the average salary to a great extent.

In layman term we can say that to get better salary one needs to have better education level as well as choose the right occupation type.

For businesses now it is easier to decide on pay cheque by considering that employee's education level and the occupation type he is in.

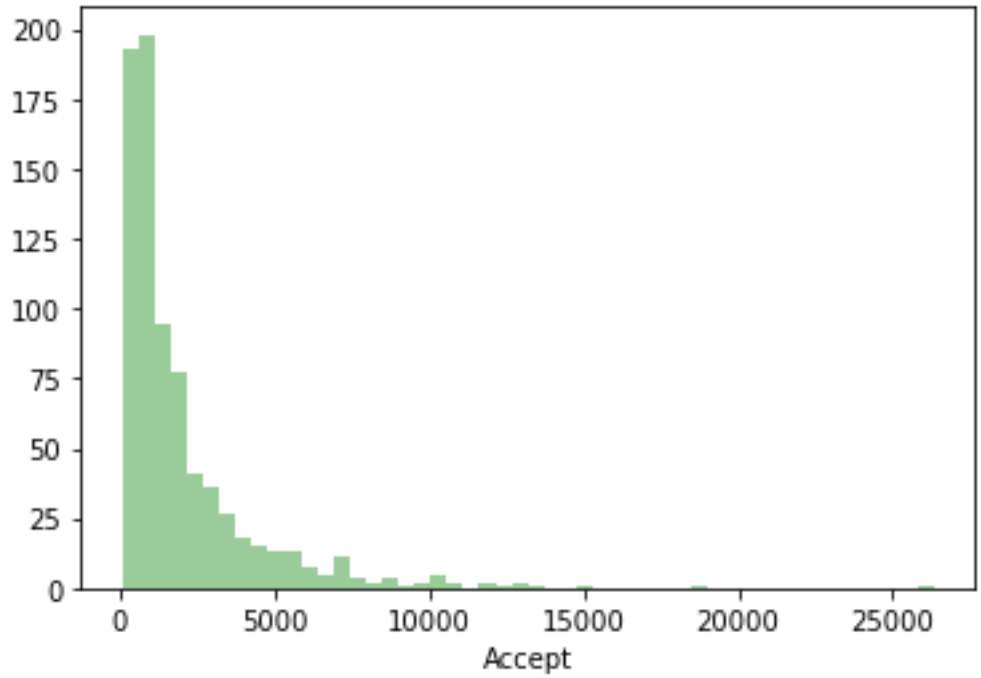
Problem- 2

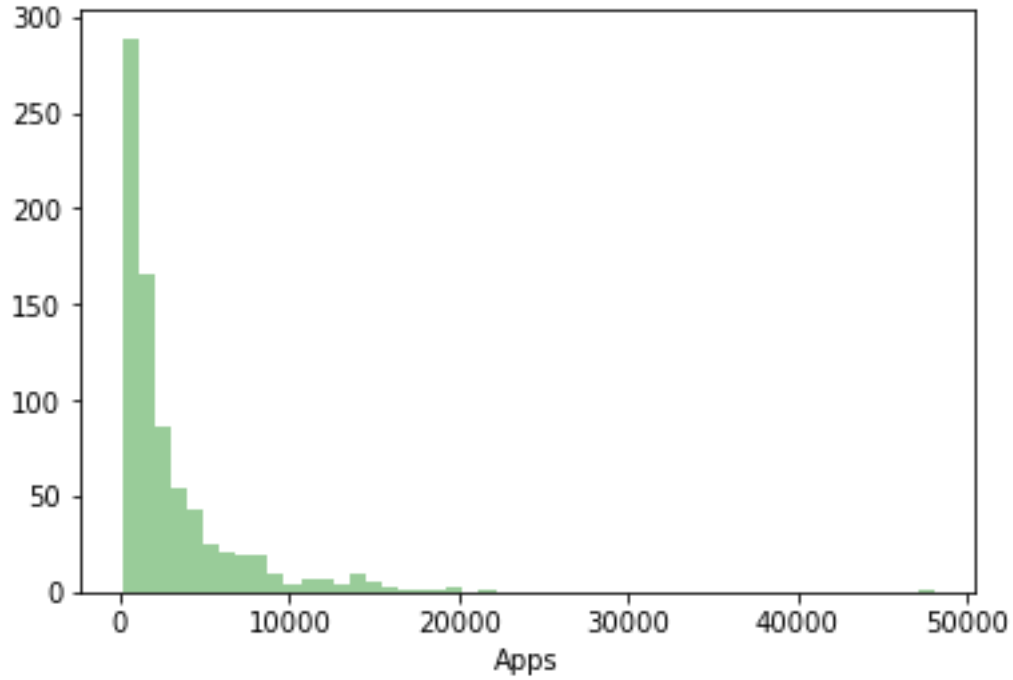
The dataset Education - Post 12th Standard contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data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

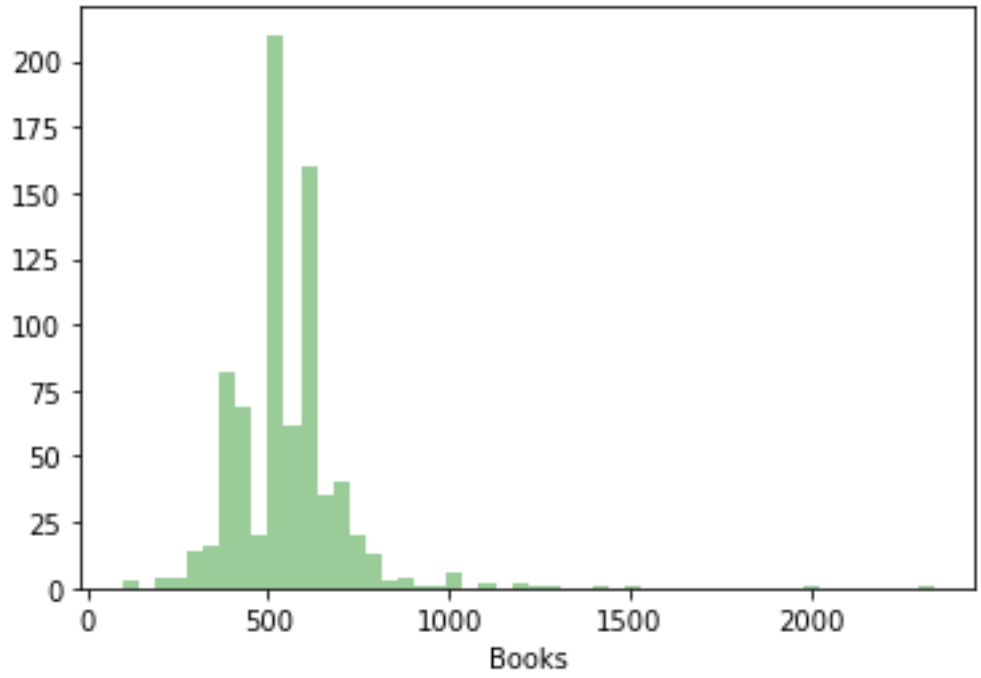
2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

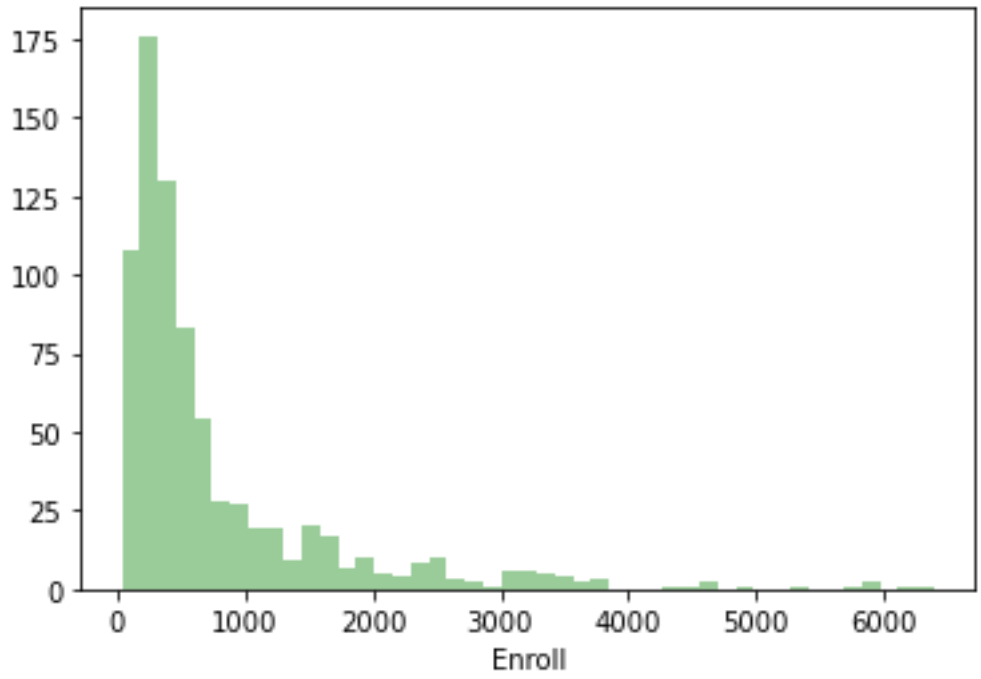
Univariate Analysis

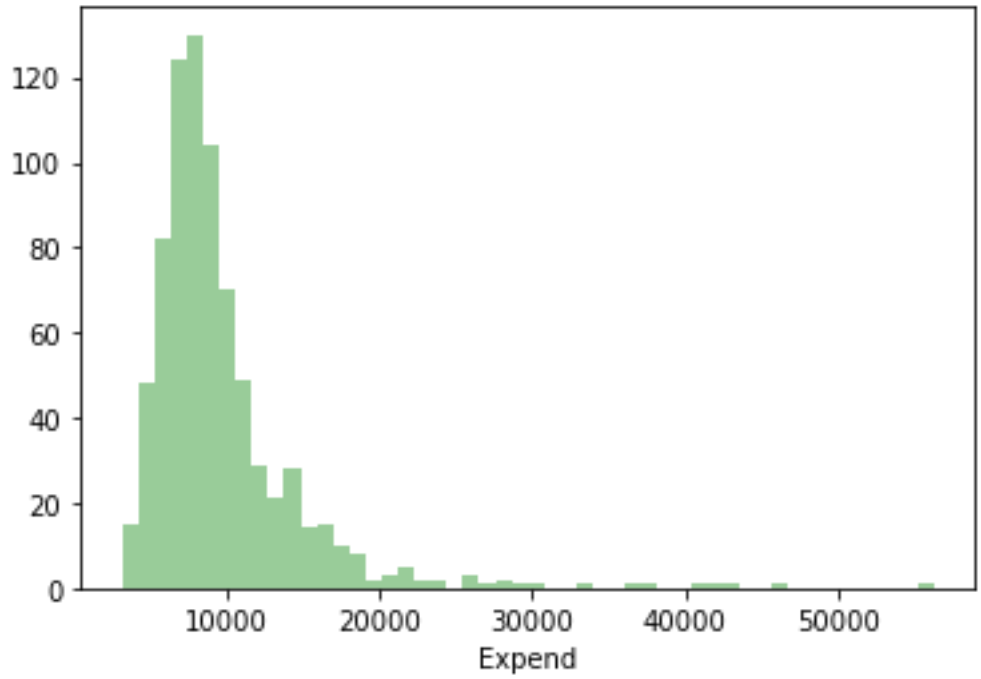
Dist Plot: Below we have plotted the distribution plots of all the 17 numerical variables of the data.

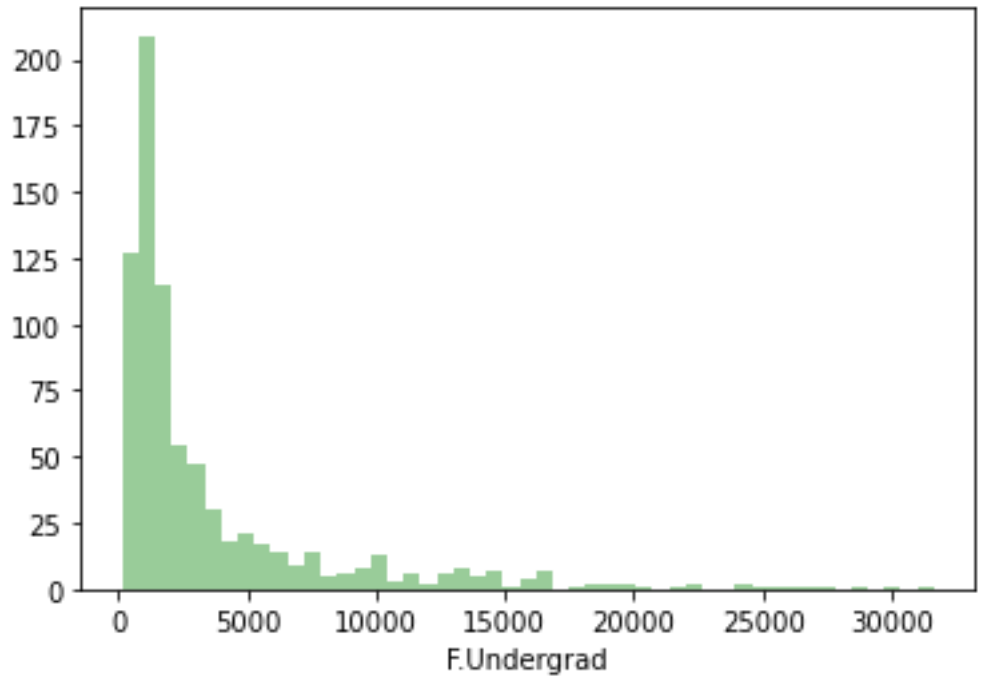

200
175
150
125
100
75
50
25
0
0
5000
10000
15000
20000
25000
Accept

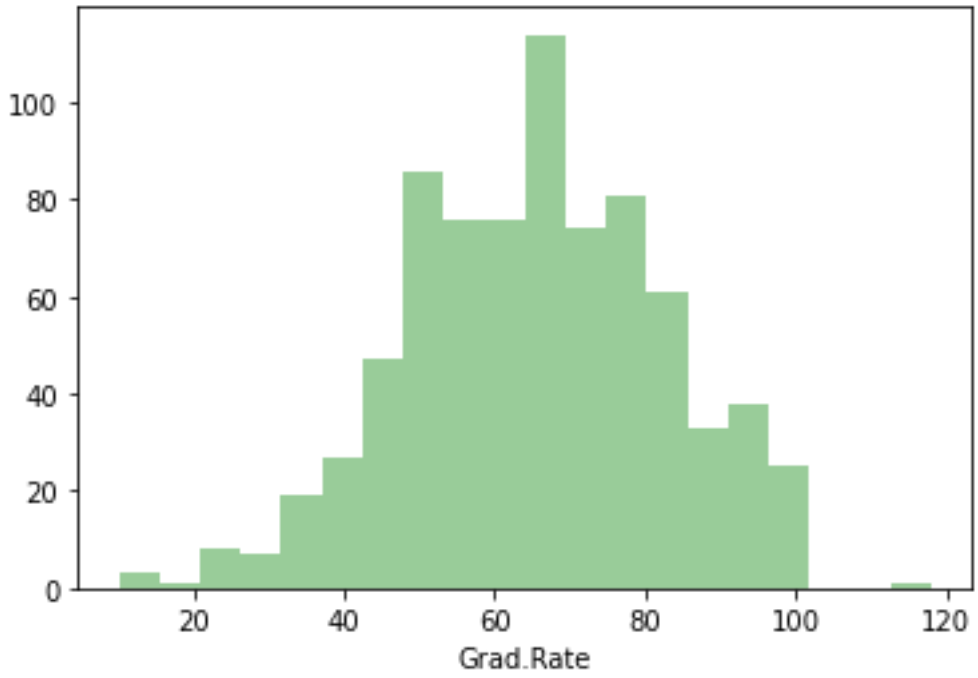

300
250
200
150
100
50
0
0
10000
20000
30000
40000
50000
Apps

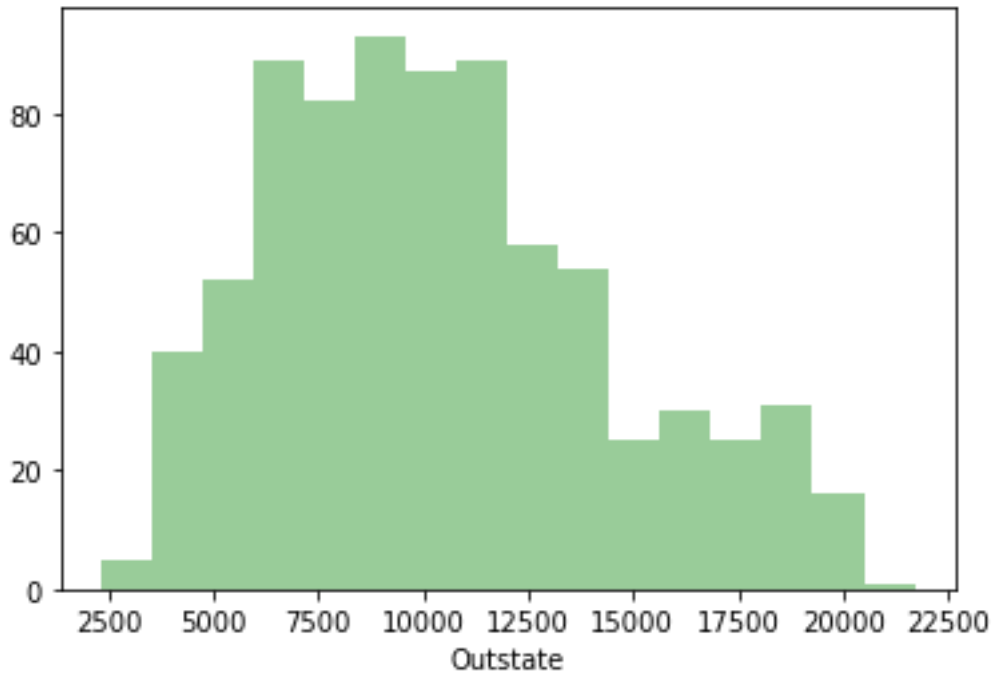

200
175
150
125
100
75
50
25
0
0
500
1000
1500
2000
Books

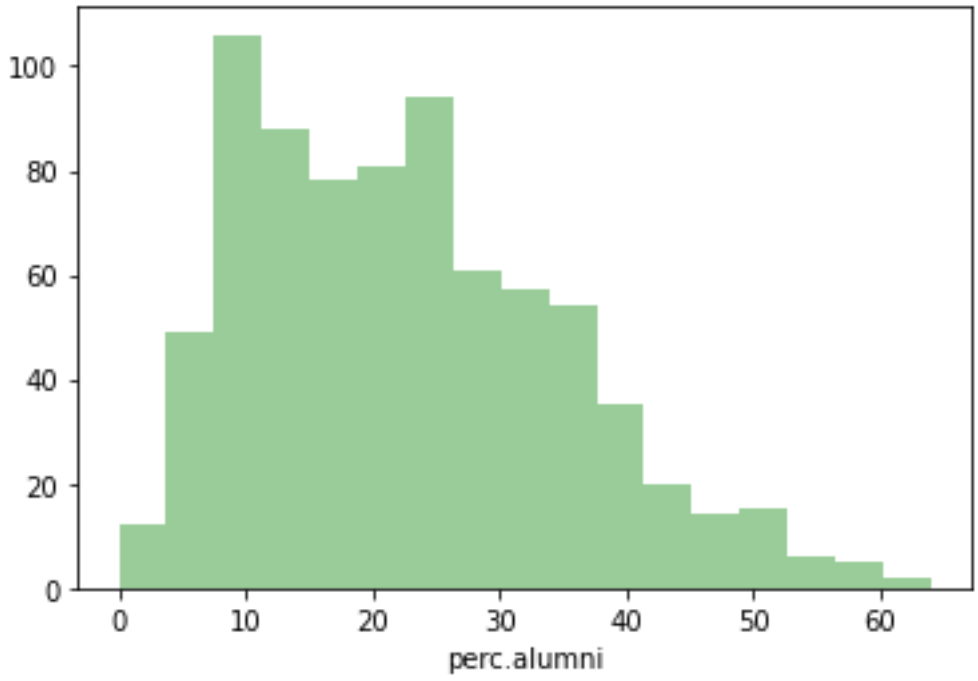

175
150
125
100
75
50
25
0
0
1000
2000
3000
4000
5000
6000
Enroll

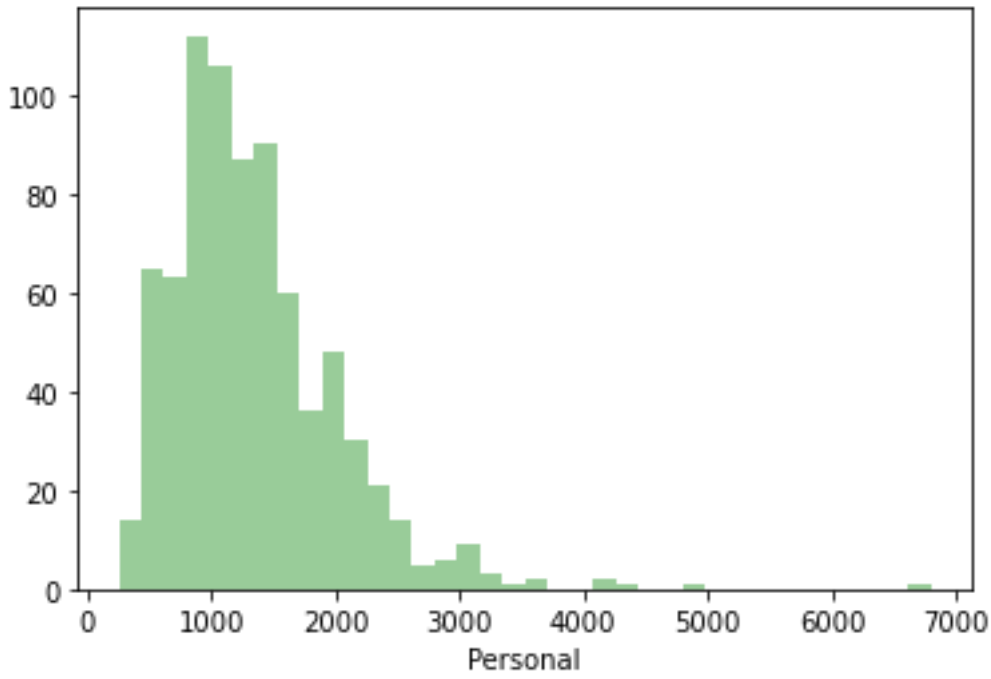

120
100
80
60
40
20
0
10000
20000
30000
40000
50000
Expend

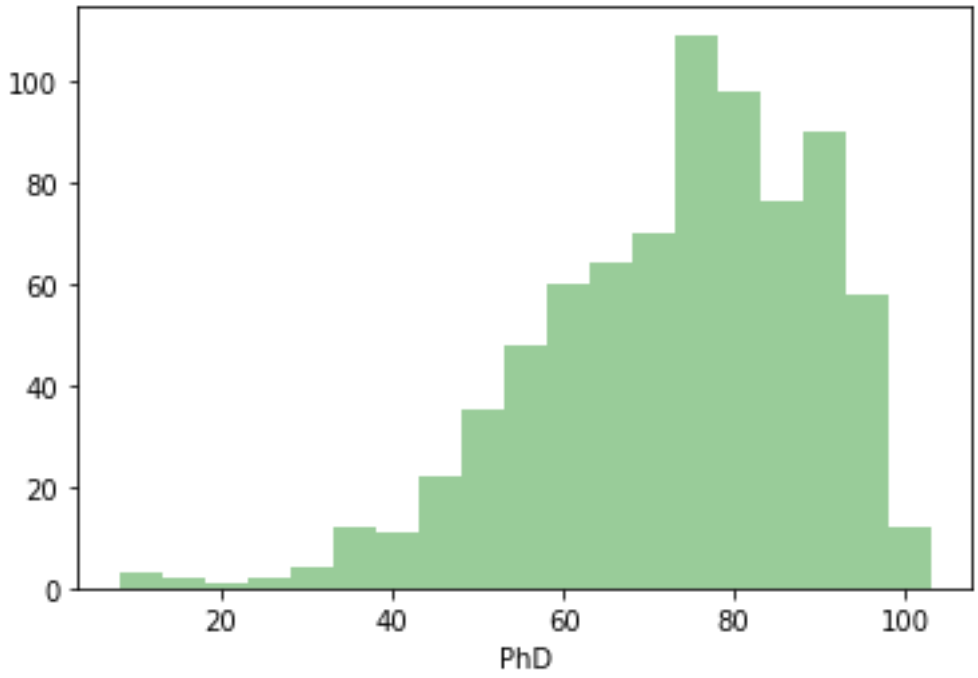

200
175
150
125
100
75
50
25
0
0
5000
10000
15000
20000
25000
30000
F.Undergrad

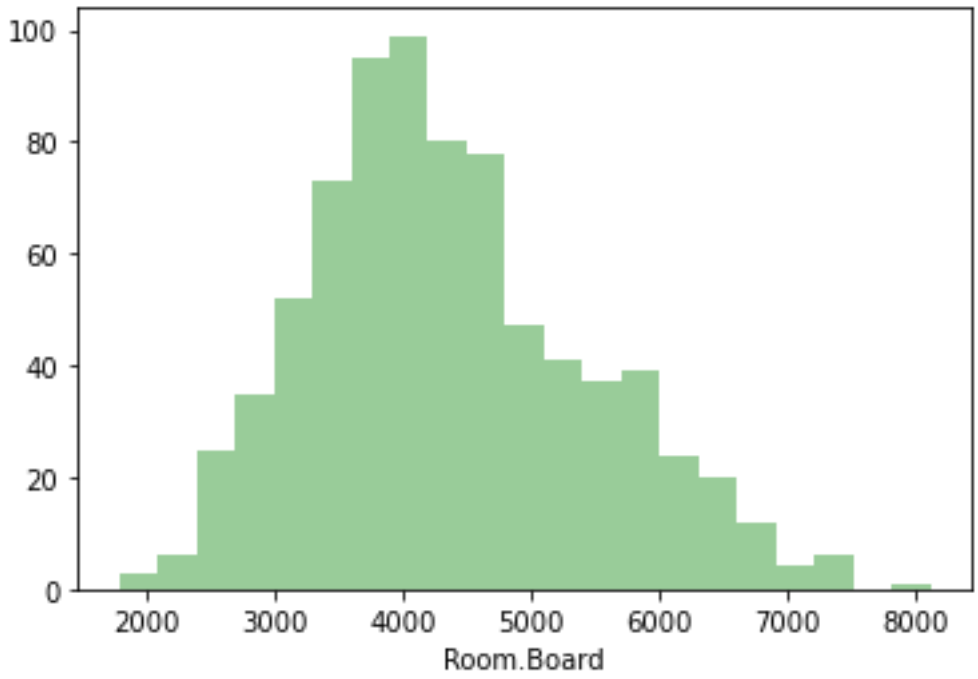

100
80
60
40
20
0
20
40
60
80
100
120
Grad.Rate

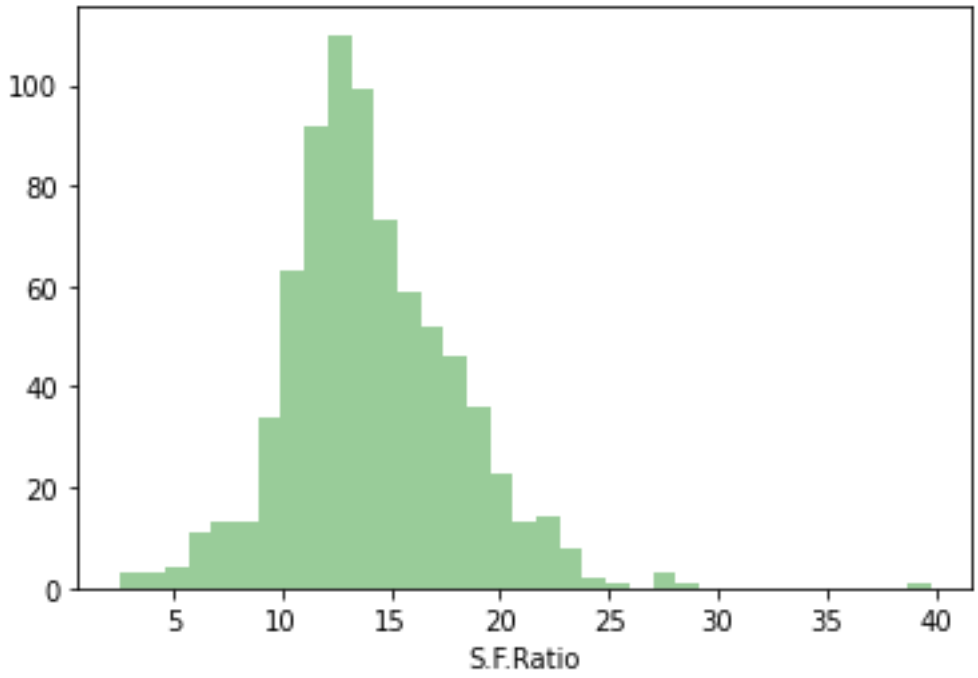

80
60
40
20
0
2500
5000
7500
10000
12500
15000
17500
20000
22500
Outstate

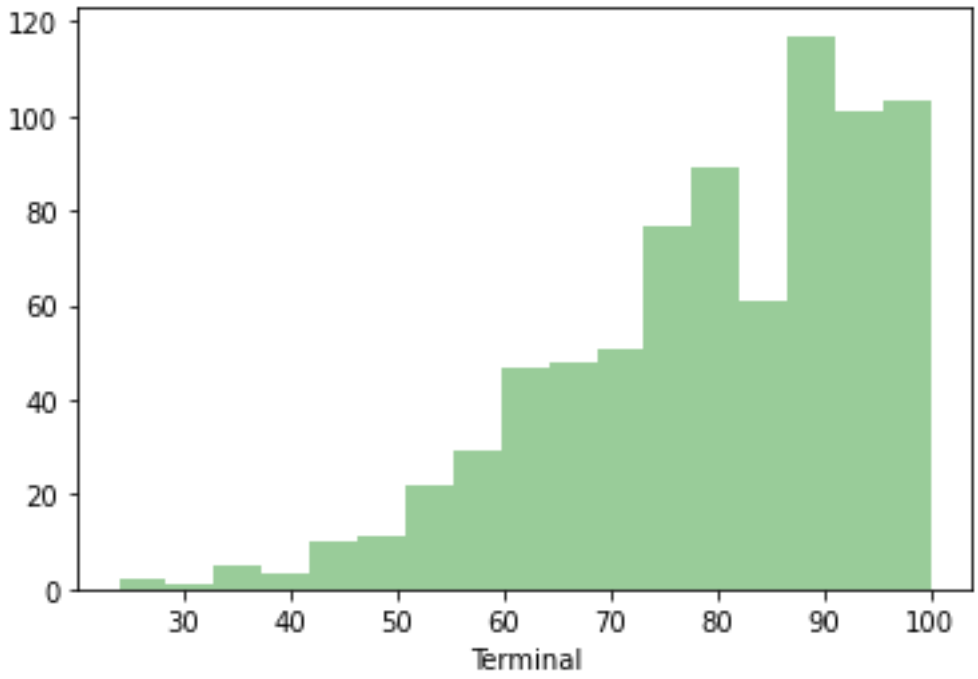

100
80
60
40
20
0
0
10
20
30
40
50
60
perc.alumni

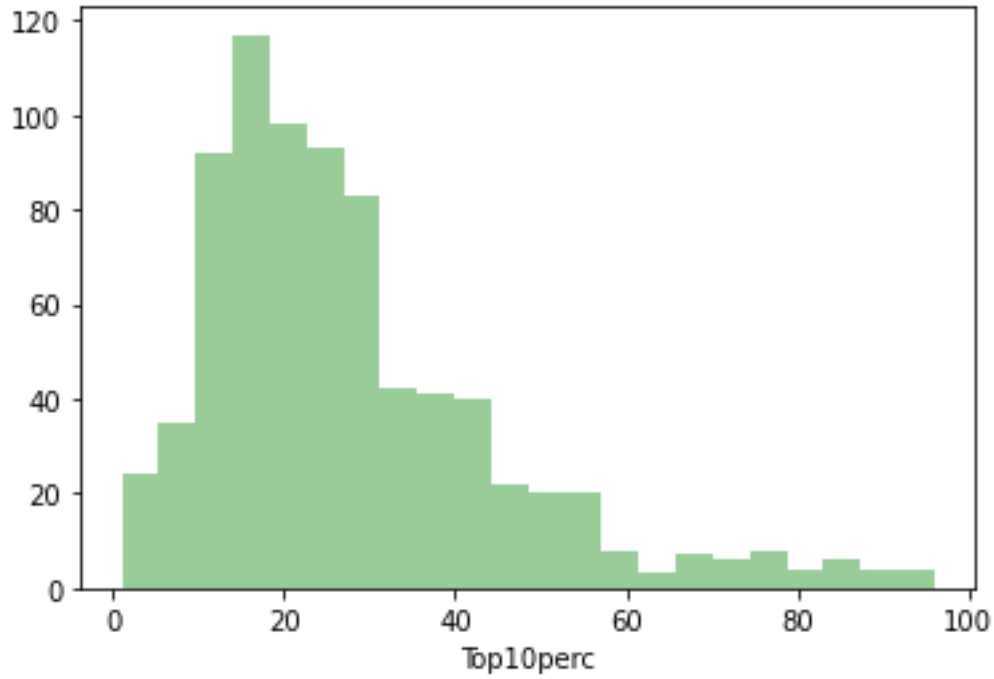

100
80
60
40
20
0
0
1000
2000
3000
4000
5000
6000
7000
Personal

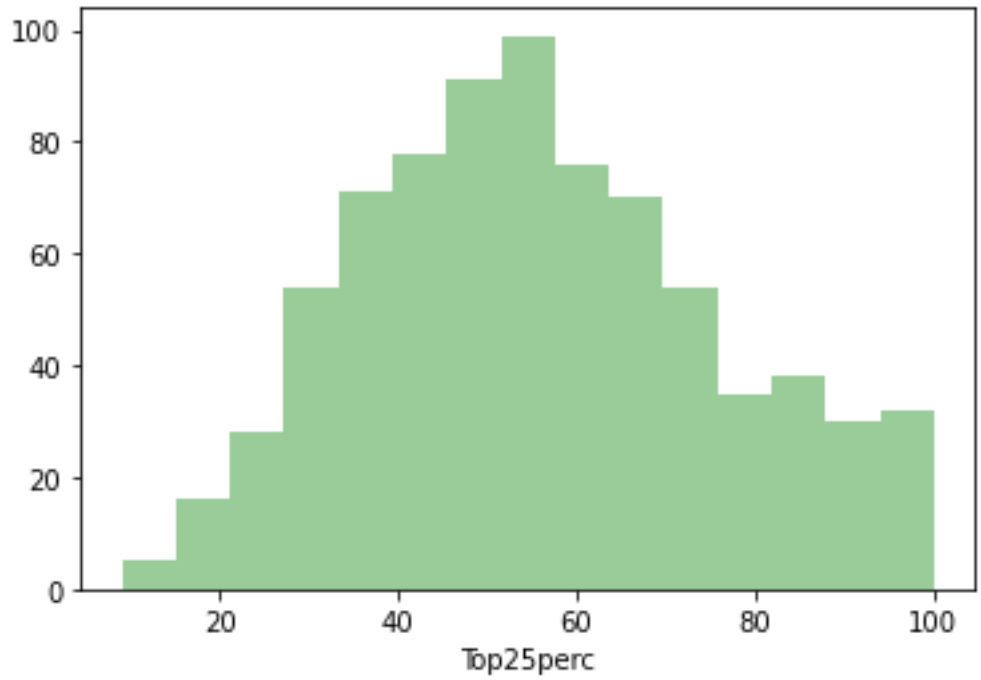

100
80
60
40
20
0
20
40
60
80
100
PhD


100
80
60
40
20
0
2000
3000
4000
5000
6000
7000
8000
Room.Board


100
80
60
40
20
0
5
10
15
20
25
30
35
40
S.F.Ratio


120
100
80
60
40
20
0
30
40
50
60
70
80
90
100
Terminal





Insights: Looking at the distributions we can divide the variable into three categories

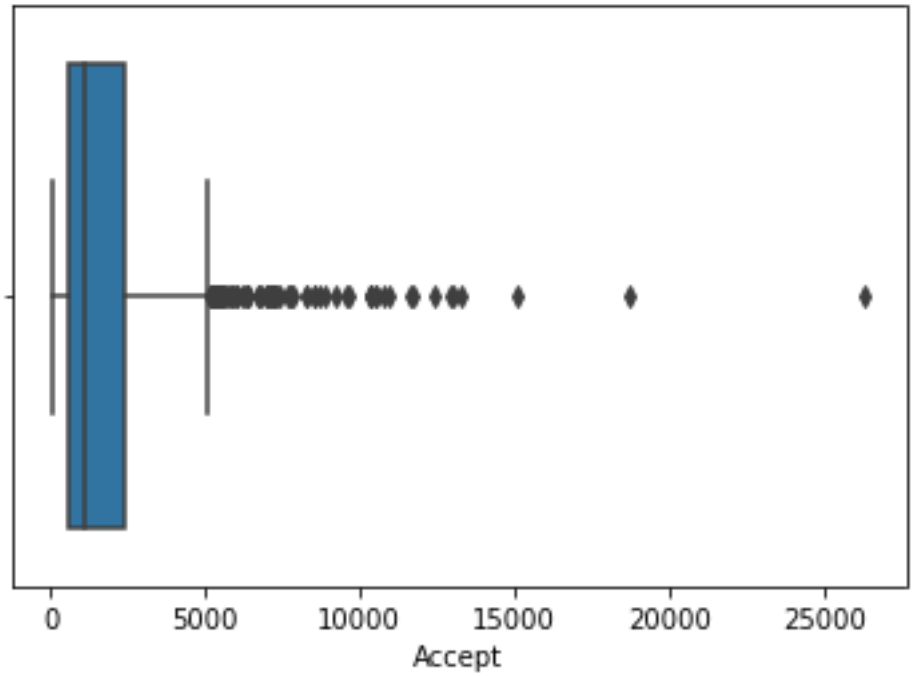
Normal Distribution – Top25perc

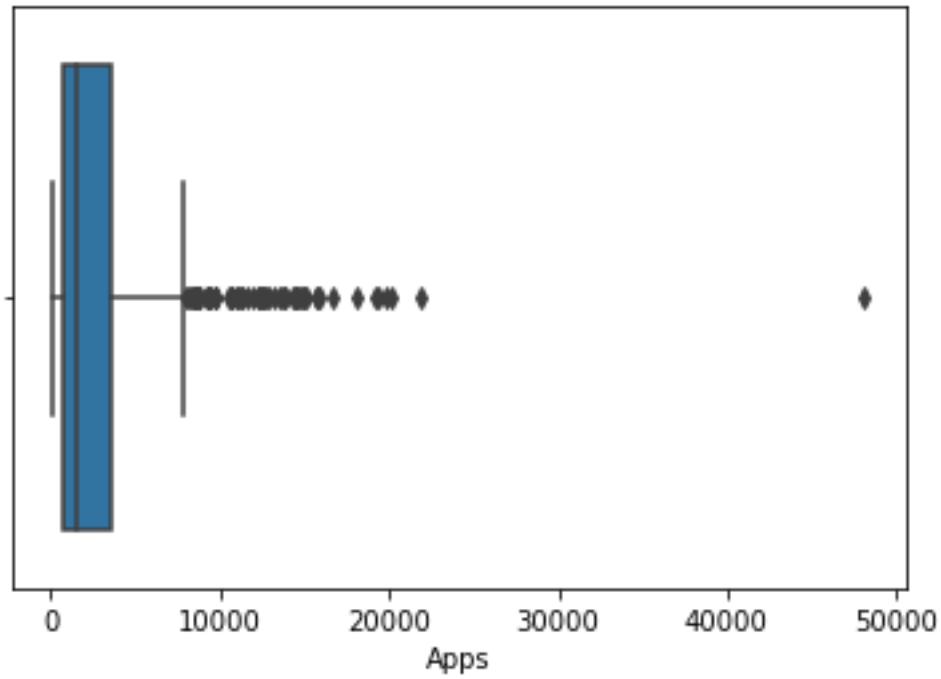
Left Skewed – PhD, Terminal

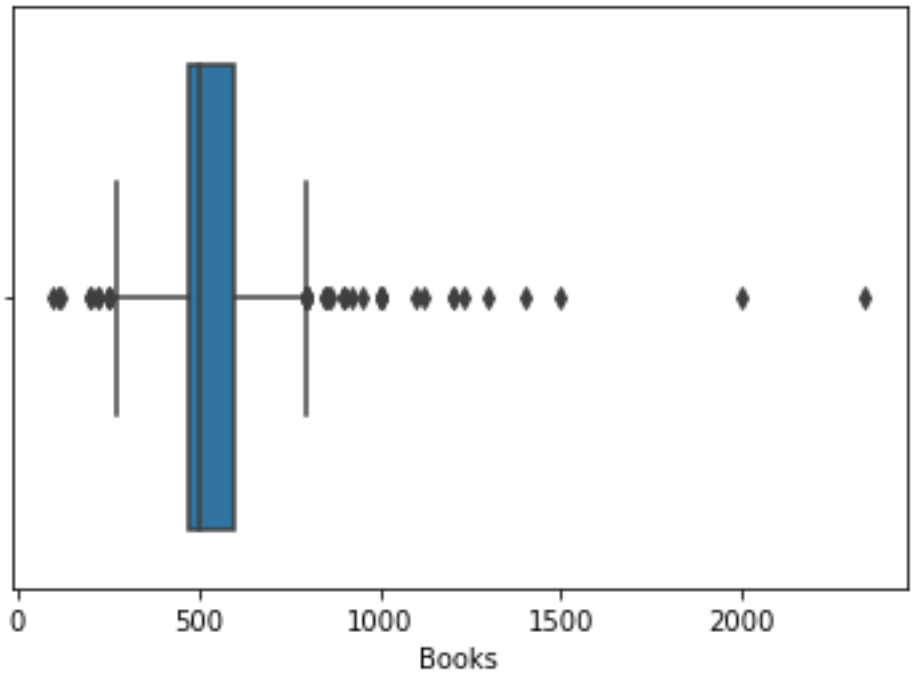
Right Skewed – Accept, Apps, Books, Enroll, Expend, F. Undergrad, Per.alumni, Personal, Top10perc

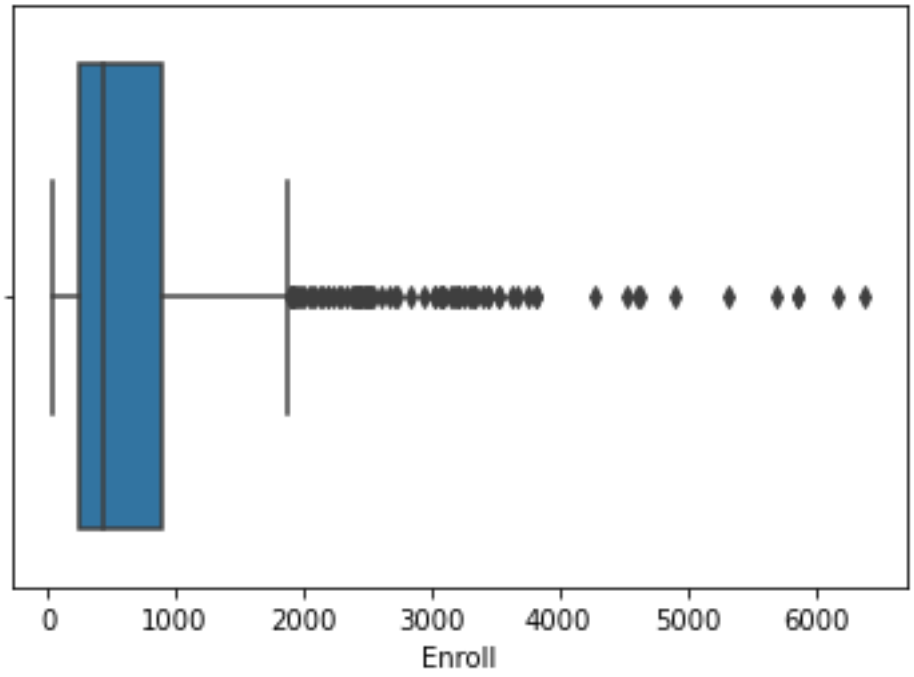
Majority of the data are right skewed.

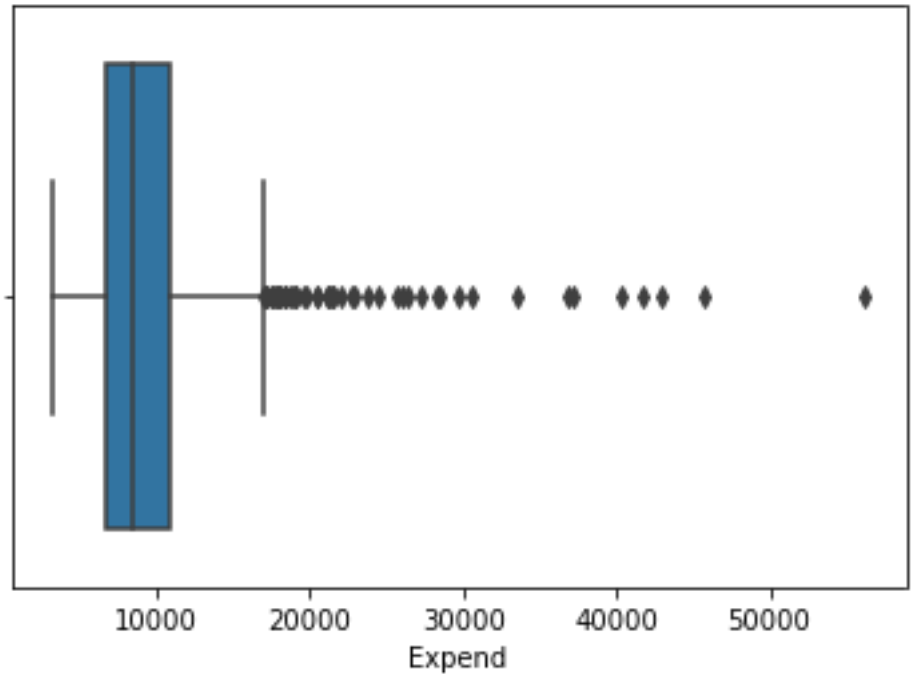
Box Plot

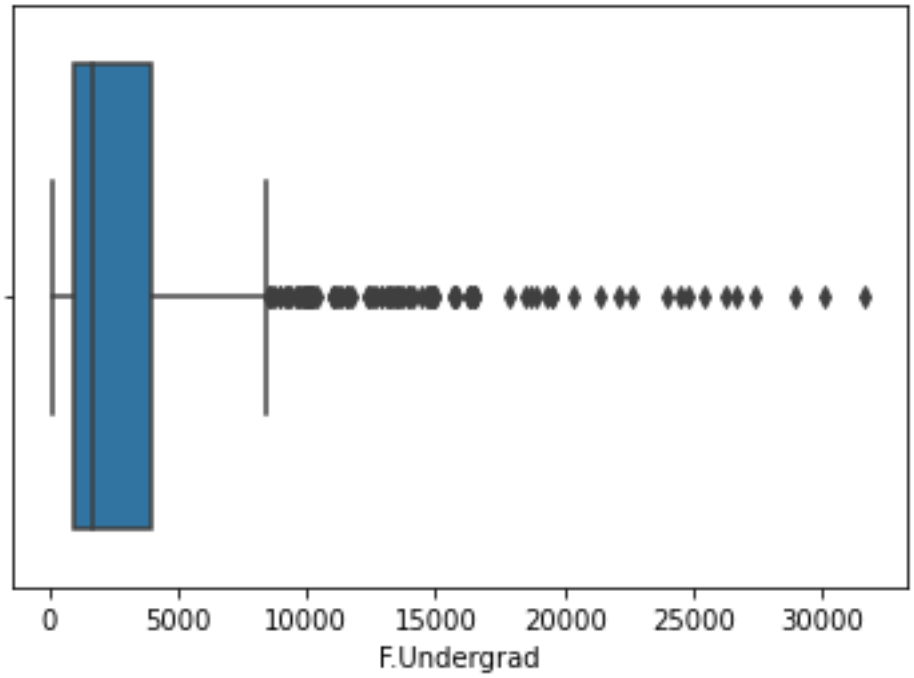

0
5000
10000
15000
20000
25000
Accept

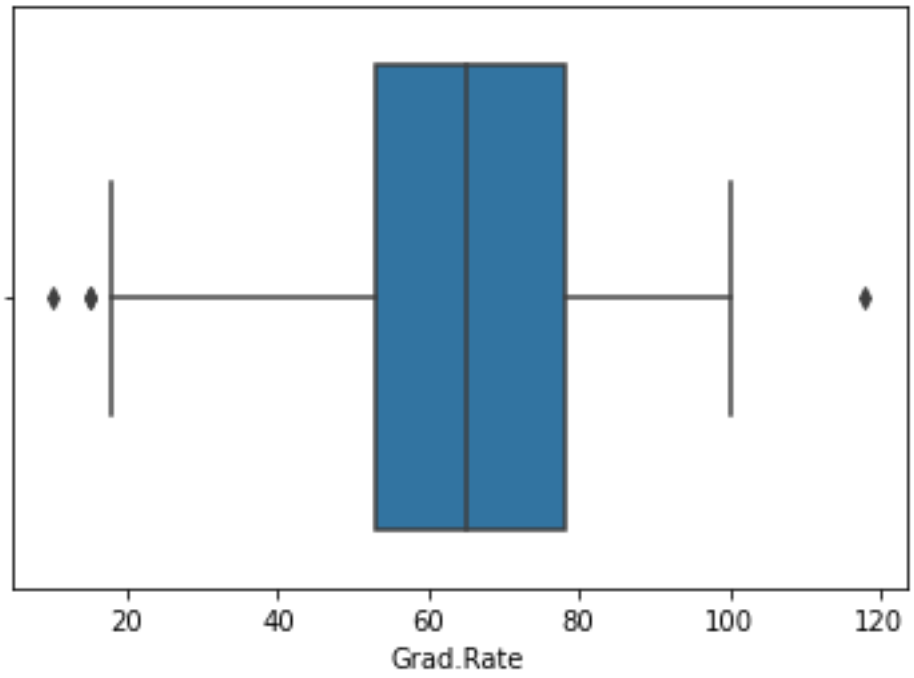

0
10000
20000
30000
40000
50000
Apps

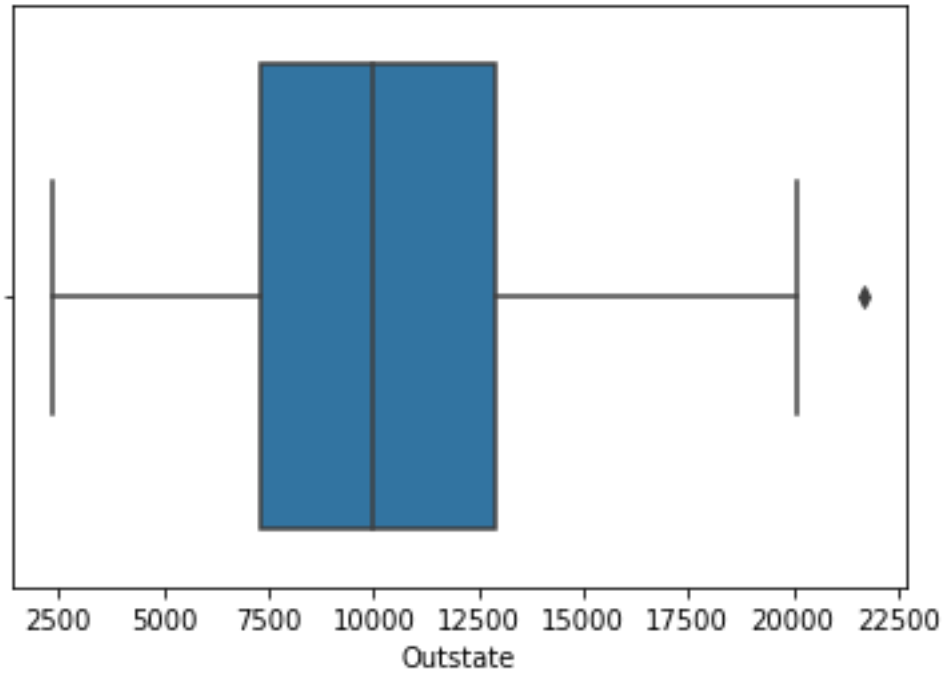

0
500
1000
1500
2000
Books

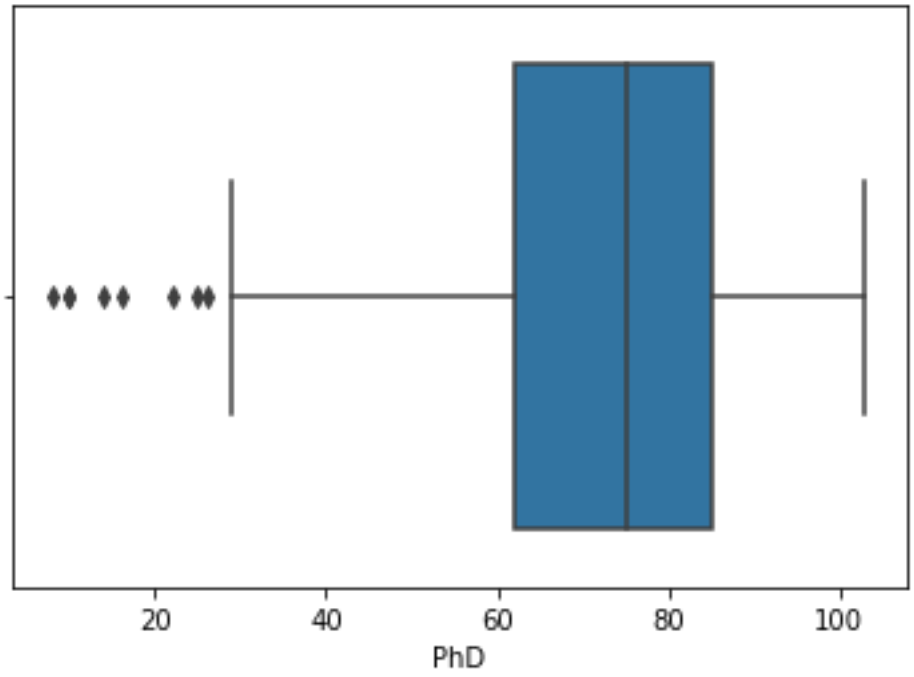

0
1000
2000
3000
4000
5000
6000
Enroll

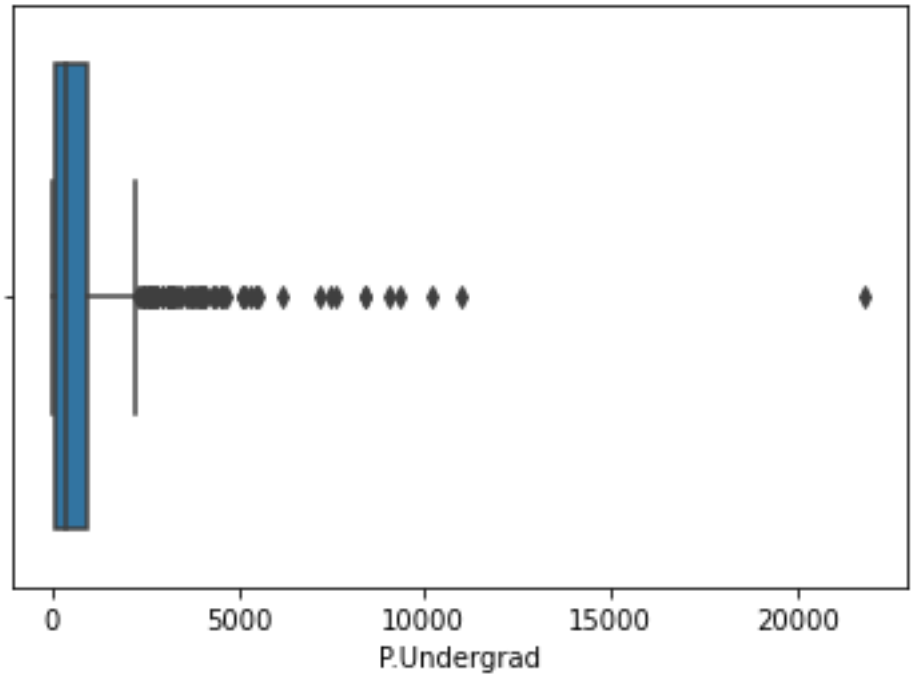

10000
20000
30000
40000
50000
Expend

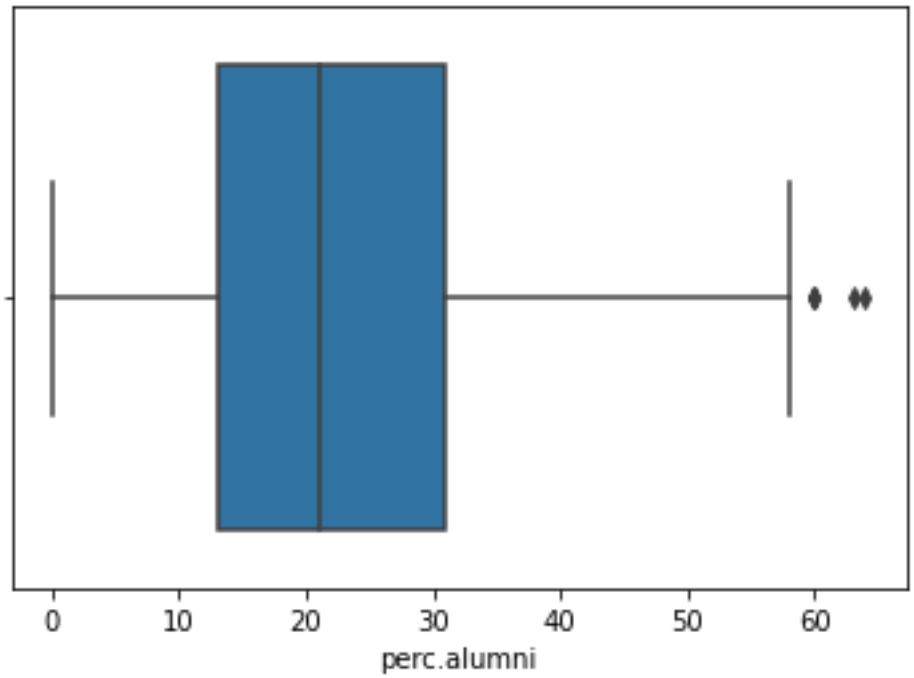

0
5000
10000
15000
20000
25000
30000
F.Undergrad

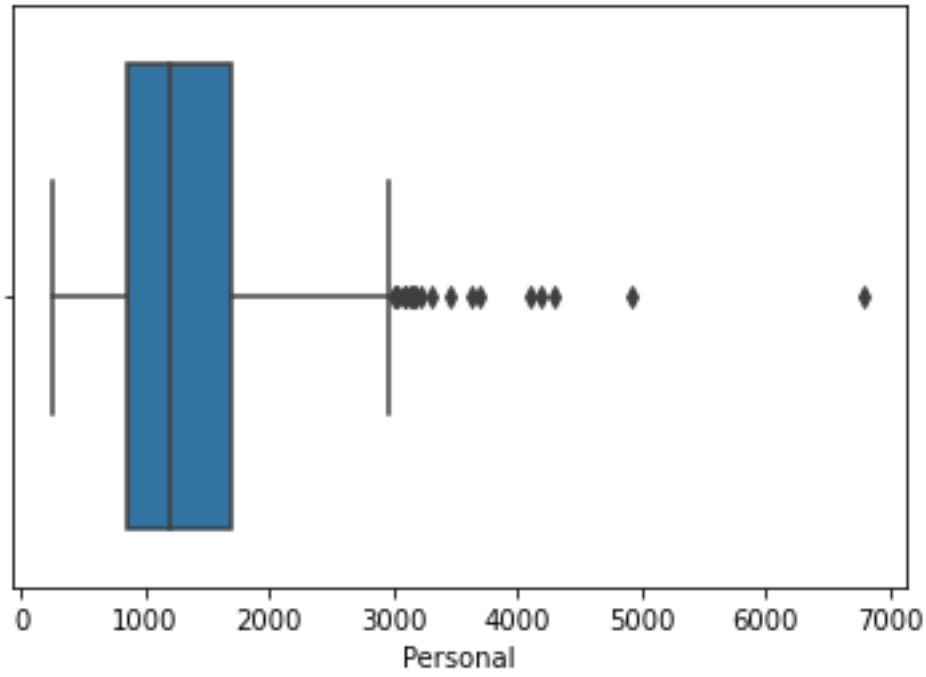

20
40
60
80
100
120
Grad.Rate

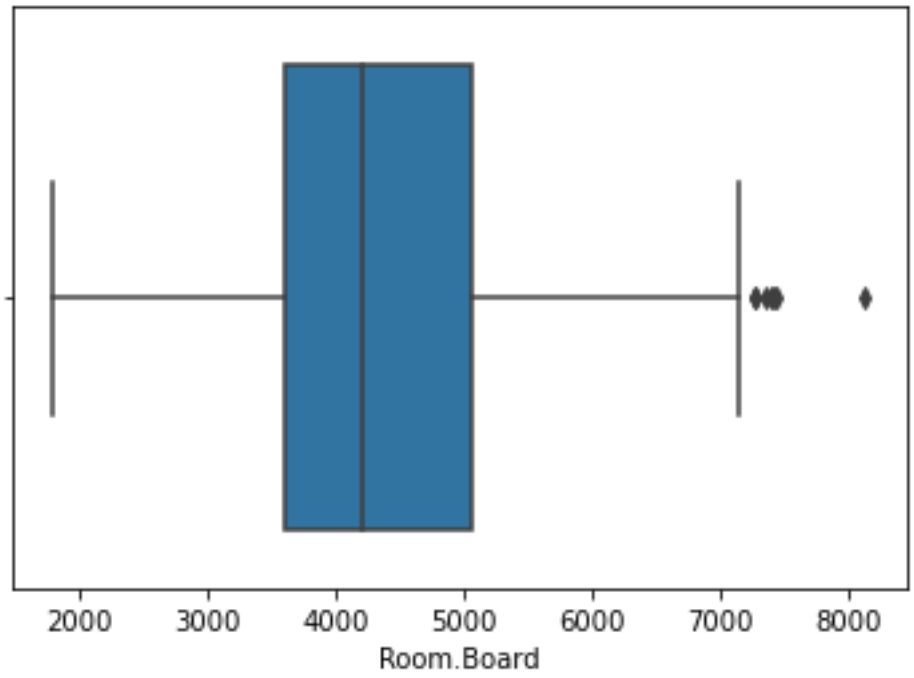

2500
5000
7500
10000
12500
15000
17500
20000
22500
Outstate

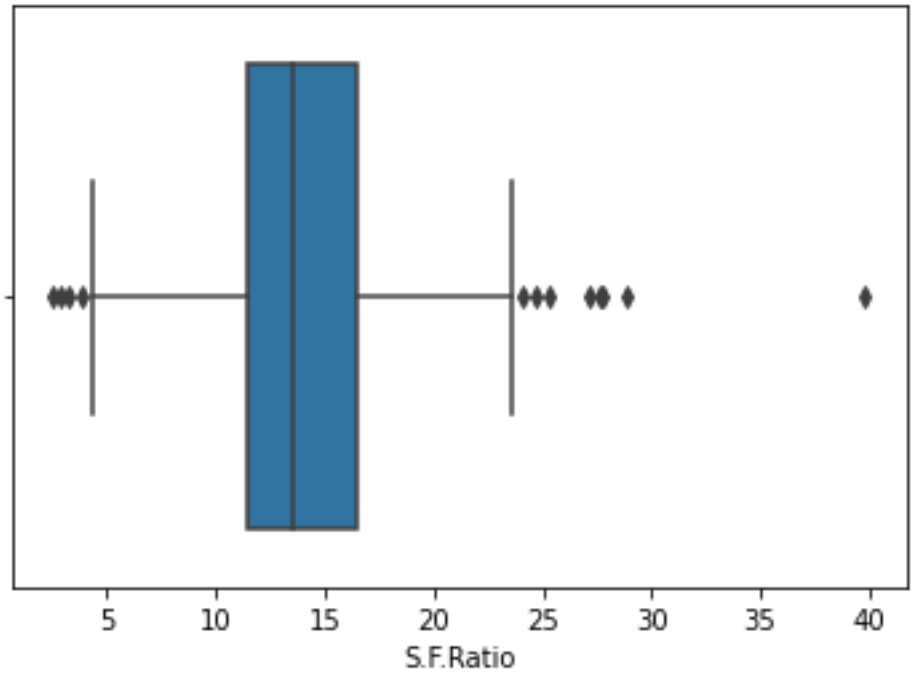

20
40
60
80
100
PhD

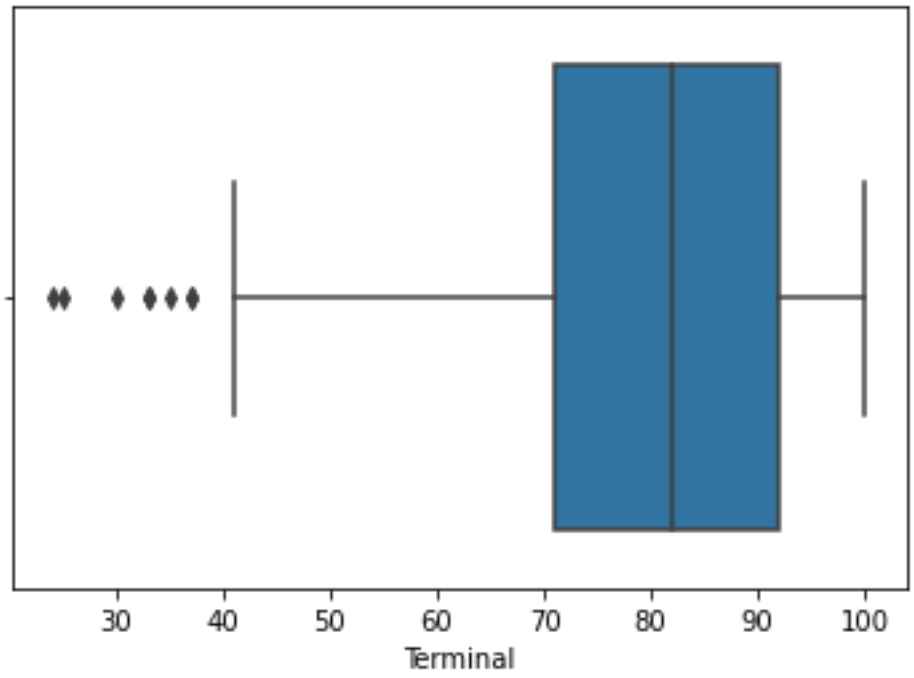

0
5000
10000
15000
20000
P.Undergrad

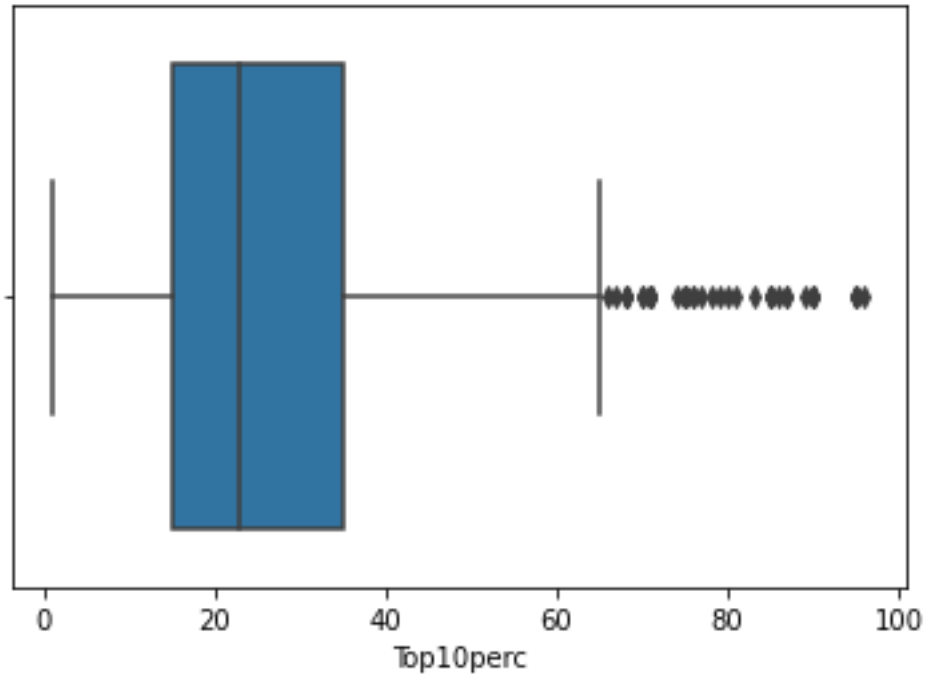

0
10
20
30
40
50
60
perc.alumni

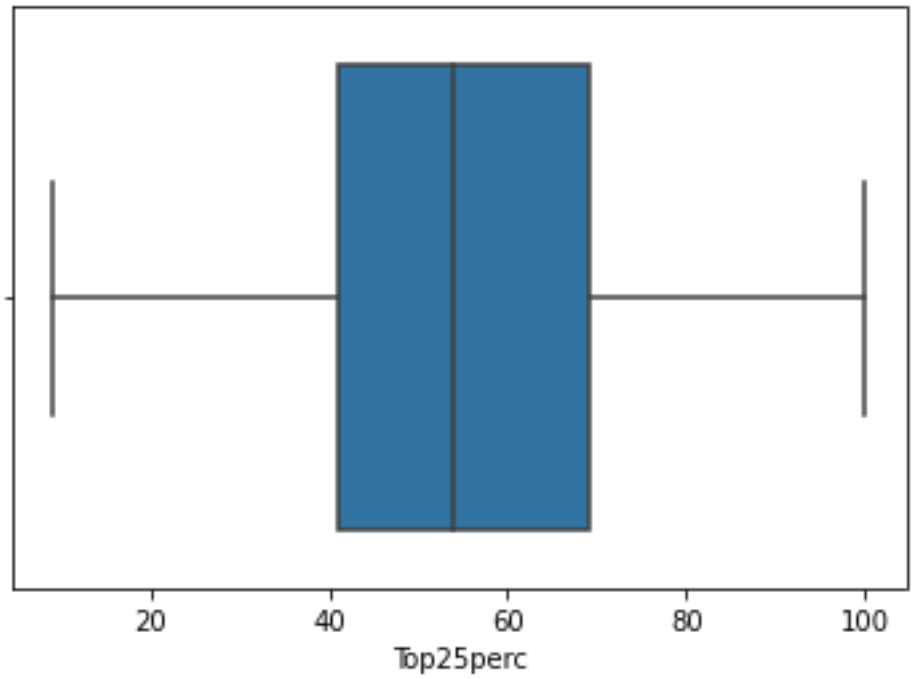

0
1000
2000
3000
4000
5000
6000
7000
Personal


2000
3000
4000
5000
6000
7000
8000
Room.Board


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


30
40
50
60
70
80
90
100
Terminal

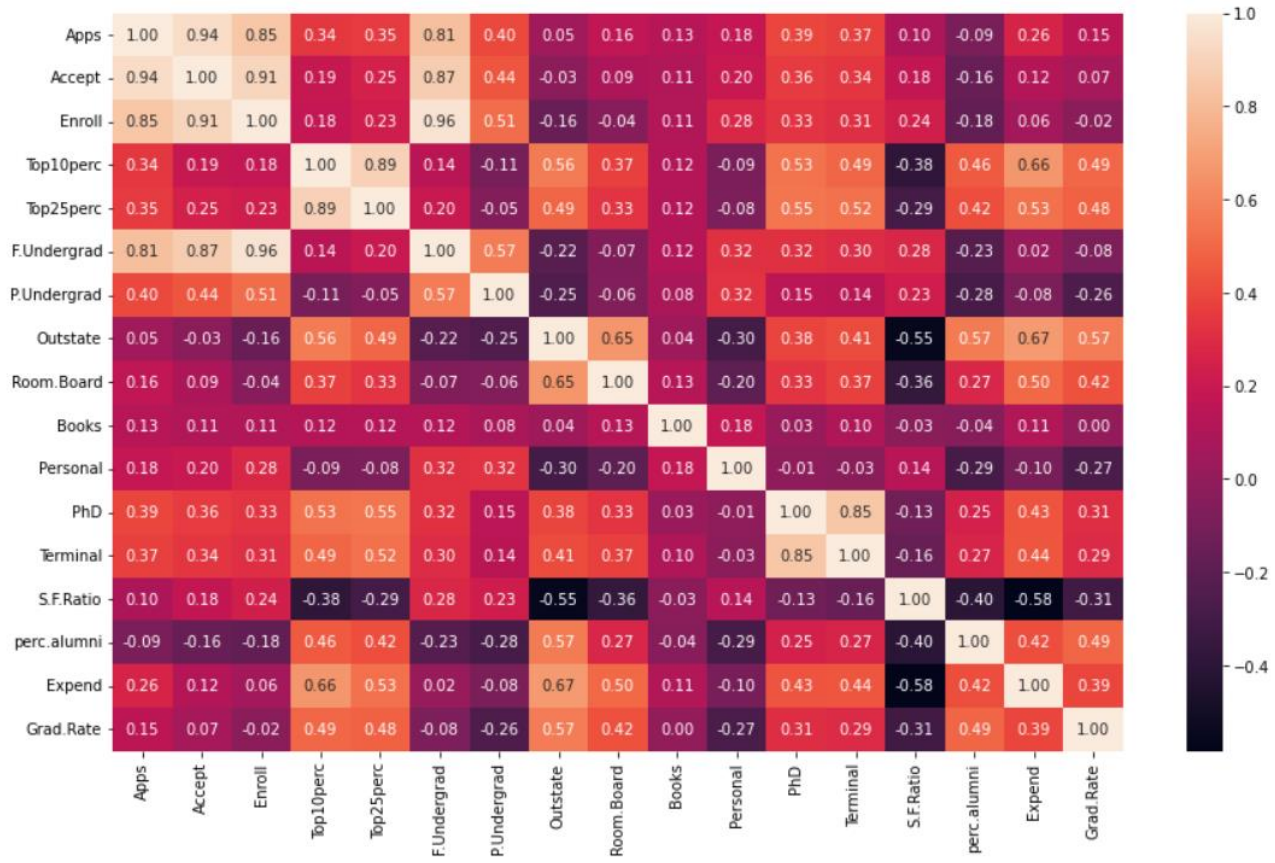

0
20
40
60
80
100
Top10perc



Insight: From the above Boxplots we observe that the only variable which has no outlier is Top25perc.

Multivariate Analysis

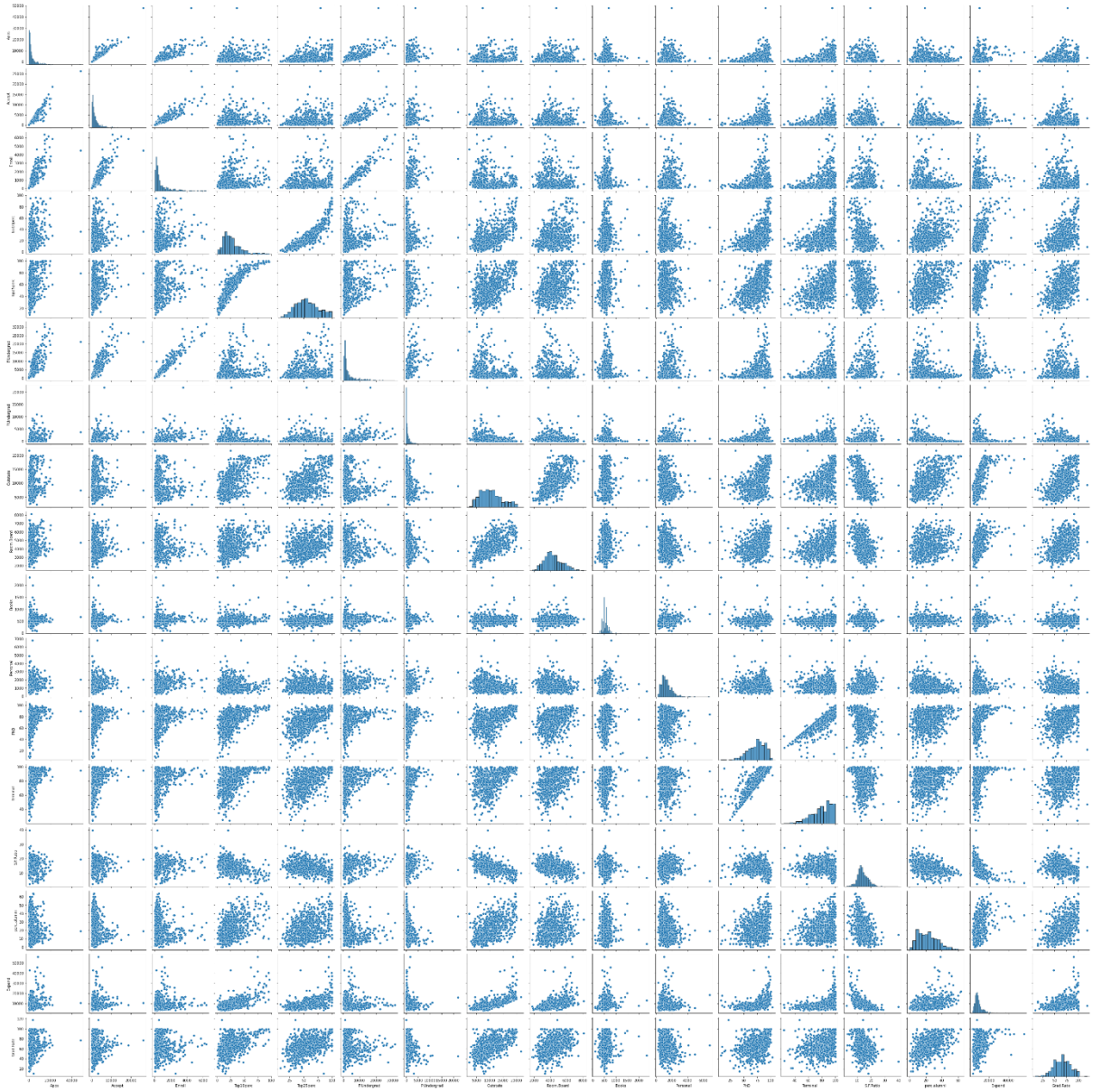
Figure 3 Correlation



From *figure 3* where we have plot a correlation plot, we can observe that Apps, Accept, Enroll, Top10perc, Top25perc and F Undergrad have high positive correlation.

S.F ratio has high negative correlation in the dataset.

Figure 4 Pairplot



Insight: Apps, Accept, Enroll and F undergrad are having a linear relationship.

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

Figure 5 Covariance of the data

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	F
Apps	14978459.53	8949859.81	3045255.99	23132.77	26952.66	15289702.47	2346620.15	780970.36	700072.87	84703.75	468346.83	2468
Accept	8949859.81	6007959.70	2076267.76	8321.12	12013.40	10393582.44	1646669.72	-253962.29	244347.15	45942.81	333556.63	1423
Enroll	3045255.99	2076267.76	863368.39	2971.58	4172.59	4347529.88	725790.67	-581188.48	-40997.06	17291.20	176737.97	502
Top10perc	23132.77	8321.12	2971.58	311.18	311.63	12089.11	-2829.47	39907.18	7186.71	346.18	-1114.55	153
Top25perc	26952.66	12013.40	4172.59	311.63	392.23	19158.95	-1615.41	38992.43	7199.90	377.76	-1083.61	176
F.Undergrad	15289702.47	10393582.44	4347529.88	12089.11	19158.95	23526579.33	4212910.09	-4209843.04	-366458.22	92535.76	1041709.09	2521
P.Undergrad	2346620.15	1646669.72	725790.67	-2829.47	-1615.41	4212910.09	2317798.85	-1552704.28	-102391.86	20410.45	329732.43	370
Outstate	780970.36	-253962.29	-581188.48	39907.18	38992.43	-4209843.04	-1552704.28	16184661.63	2886597.39	25808.24	-814673.72	2515
Room.Board	700072.87	244347.15	-40997.06	7186.71	7199.90	-366458.22	-102391.86	2886597.39	1202743.03	23170.31	-148083.77	589
Books	84703.75	45942.81	17291.20	346.18	377.76	92535.76	20410.45	25808.24	23170.31	27259.78	20043.03	72
Personal	468346.83	333556.63	176737.97	-1114.55	-1083.61	1041709.09	329732.43	-814673.72	-148083.77	20043.03	458425.75	-12

From figure 5 we are observing that the covariance of the data is widely different, hence scaling is necessary here. PCA works on total variance which is the sum of the variances in the data. If one variance is very high compared to the rest, it will dominate the construction of the PCs and all variables will not have proper representation.

We have performed scaling and here is the table after scaling

Figure 6 after scaling

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Rat
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	1.0137
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	-0.4777
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	-0.3007
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	-1.6152
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	-0.5535

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

Figure 7 covariance of scaled data

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.a
Apps	1.00	0.94	0.85	0.34	0.35	0.82	0.40	0.05	0.17	0.13	0.18	0.39	0.37	0.10	
Accept	0.94	1.00	0.91	0.19	0.25	0.88	0.44	-0.03	0.09	0.11	0.20	0.36	0.34	0.18	
Enroll	0.85	0.91	1.00	0.18	0.23	0.97	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.37	0.12	-0.09	0.53	0.49	-0.39	
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33	0.12	-0.08	0.55	0.53	-0.30	
F.Undergrad	0.82	0.88	0.97	0.14	0.20	1.00	0.57	-0.22	-0.07	0.12	0.32	0.32	0.30	0.28	
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	-0.06	0.08	0.32	0.15	0.14	0.23	
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.66	0.04	-0.30	0.38	0.41	-0.56	
Room.Board	0.17	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.66	1.00	0.13	-0.20	0.33	0.38	-0.36	
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.08	0.04	0.13	1.00	0.18	0.03	0.10	-0.03	
Personal	0.18	0.20	0.28	-0.09	-0.08	0.32	0.32	-0.30	-0.20	0.18	1.00	-0.01	-0.03	0.14	
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.03	-0.01	1.00	0.85	-0.13	
Terminal	0.37	0.34	0.31	0.49	0.53	0.30	0.14	0.41	0.38	0.10	-0.03	0.85	1.00	-0.16	
S.F.Ratio	0.10	0.18	0.24	-0.39	-0.30	0.28	0.23	-0.56	-0.36	-0.03	0.14	-0.13	-0.16	1.00	
perc.alumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.27	-0.04	-0.29	0.25	0.27	-0.40	
Expend	0.26	0.12	0.06	0.66	0.53	0.02	-0.08	0.67	0.50	0.11	-0.10	0.43	0.44	-0.58	
Grad.Rate	0.15	0.07	-0.02	0.50	0.48	-0.08	-0.26	0.57	0.43	0.00	-0.27	0.31	0.29	-0.31	

Figure 8 correlation of scaled data

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termin
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.3694
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758	0.3375
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469	0.3082
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828	0.4911
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862	0.5247
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.318337	0.3000
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.081200	0.319882	0.149114	0.1419
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.038855	-0.299087	0.382982	0.4079
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.127963	-0.199428	0.329202	0.3745
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.000000	0.179295	0.026906	0.0999
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.000000	-0.010936	-0.0306
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1.000000	0.8495
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.099955	-0.030613	0.849587	1.0000
S.F.Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628	-0.031929	0.136345	-0.130530	-0.1601
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009	0.2671
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739	0.112409	-0.097892	0.432762	0.4387
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.571290	0.424942	0.001061	-0.269344	0.305038	0.2895

Insight: Here we find that covariance and correlation matrix are same after scaling.

Scaling ensures that attributes have mean as zero and variance as one.

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

Figure 9 Outliers before scaling

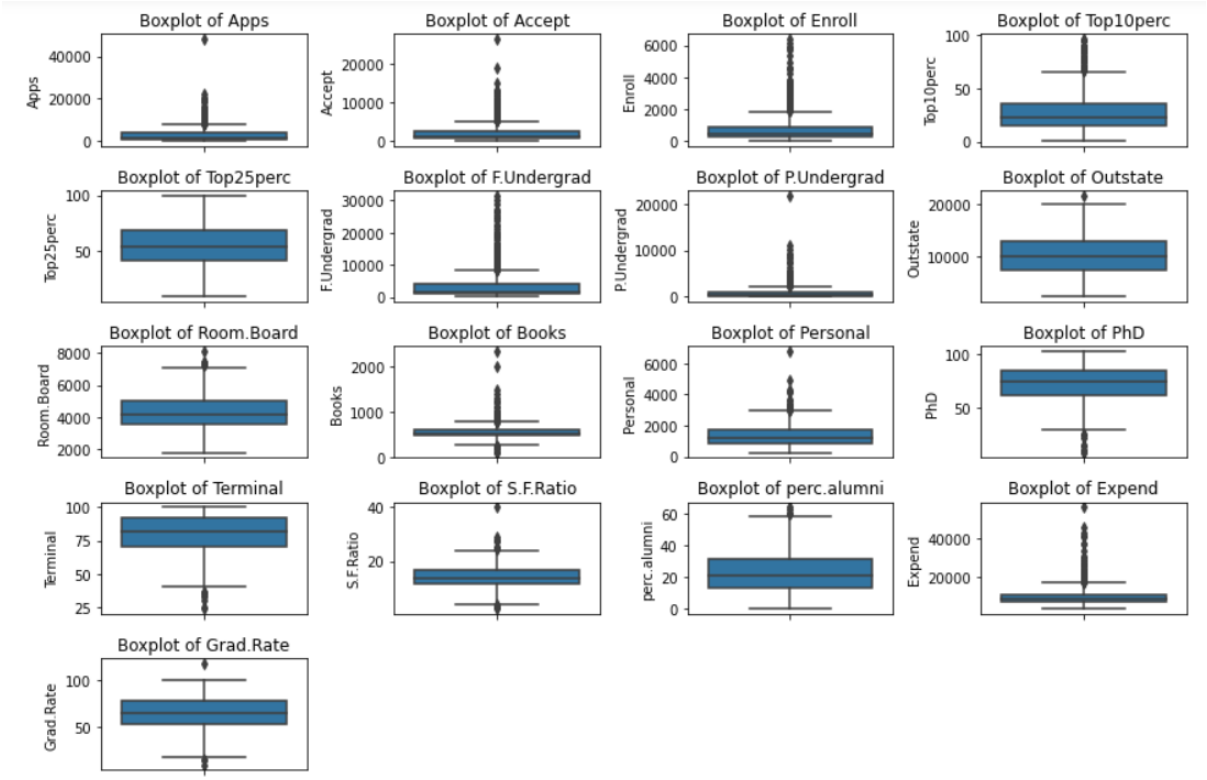
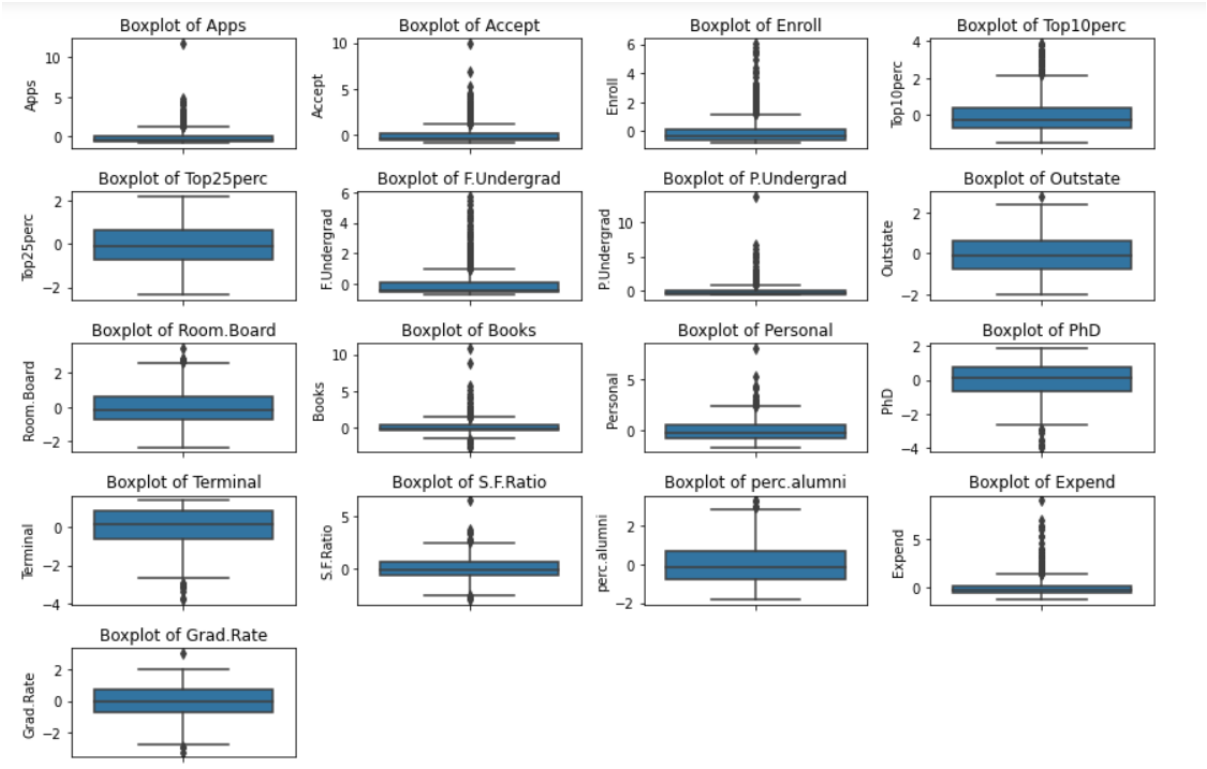


Figure 10 Outlier after scaling



Insight: Here we find that scaling does not treat outliers, it only reduces the magnitudes of the variables for better analysis.

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

Figure 11 Eigen Vectors

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

Figure 12 Eigen Values

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

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

Figure 13 Explained variance ratio

```
array([0.32020628, 0.26340214, 0.06900917, 0.05922989, 0.05488405,
       0.04984701, 0.03558871, 0.03453621, 0.03117234, 0.02375192,
       0.01841426, 0.01296041, 0.00985754, 0.00845842, 0.00517126,
       0.00215754, 0.00135284])
```

Here in *figure 13* we have plotted the explained variances of each individual PCs.

Figure 14 Data frame of PC with Eigen vectors

	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	0.0
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	0.0
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	-0.0
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	-0.1
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	0.1
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	-0.0
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	0.0
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	-0.0
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	-0.0
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	-0.0
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	0.0

In *figure 14* the PCs are plotted along with their Eigen vectors

Figure 15 Choosing Number of PC

```
array([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.        ])
```

here we will choose up to 6 PC which covers the 81% variance of the dataset

Figure 16 Selected PCs

	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

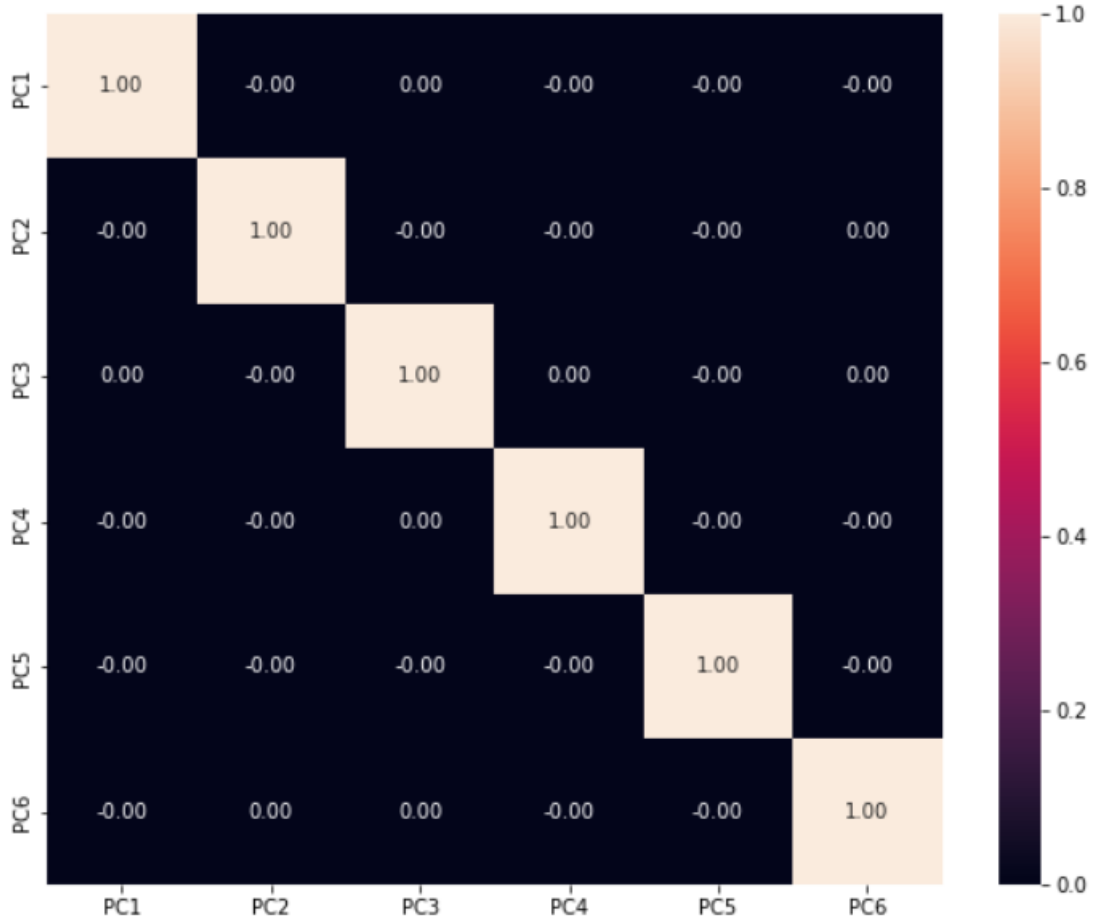
We have now reduced the dimension of the data from 17 to 6 and presented in *figure 16.*

Figure 17 Final PC Data Frame

	PC1	PC2	PC3	PC4	PC5	PC6
0	-1.592855	0.767334	-0.101074	-0.921749	-0.743975	-0.298306
1	-2.192402	-0.578830	2.278798	3.588918	1.059997	-0.177137
2	-1.430964	-1.092819	-0.438093	0.677241	-0.369613	-0.960592
3	2.855557	-2.630612	0.141722	-1.295486	-0.183837	-1.059508
4	-2.212008	0.021631	2.387030	-1.114538	0.684451	0.004918
5	-0.571665	-1.496325	0.024354	0.066944	-0.376261	-0.668343
6	0.241952	-1.506368	0.234194	-1.142024	1.546983	-0.009995
7	1.750474	-1.461412	-1.026589	-0.981184	0.217044	0.222924
8	0.769127	-1.984433	-1.426052	-0.071424	0.586380	-0.655179
9	-2.770721	-0.844611	1.627987	1.705091	-1.019826	-0.794401

As we know that PCs are the product of its Eigen vectors and normalized (scaled) variable, we have performed a dot product between the both and presented in *figure 17.* This is the data which can be taken to the next stage for further analysis.

Figure 18 Correlation between PCs



PC s are uncorrelated with each other and that we have showed in *figure 18.*

2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

Figure 19 First PC equation

PC1 =
0.24SApps+0.20SAccept+0.17SEnroll+0.35STop10perc+0.34STop25perc+0.15F.undergrad+0.02P.Undergrad+0.29Soutstate+0.24SRoom.Board+0.06SBooks-0.04SPersonal+0.31SPh.D+0.31STerminal-0.17SS.F.Ratio+0.20Sperc.alumni+0.31SExpend+0.25SGrad.Rate

The letter S indicates that the scaled (normalized) variable is used to construct the PCs.

In *figure 19* we have written the equation of first PC, this is nothing but the multiplication of Eigen vectors with the scaled variable. *Figure 17* can be referred for the rest PCs.

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?

Figure 20 Cumulative variance of PCs

```
array([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.        ])
```

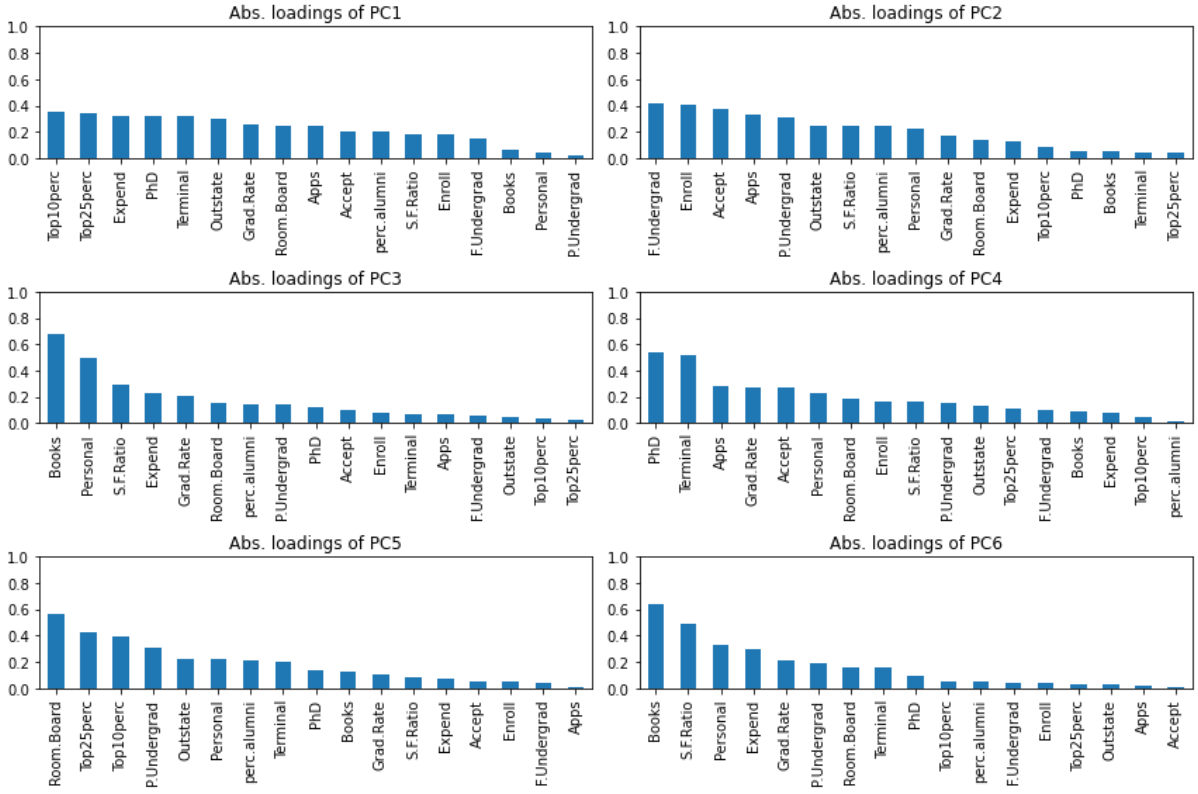
here we will choose up to 6 PC which covers the 81% variance of the dataset

In *figure 20* we have printed the cumulative variance explained by the PCs, and we have taken an 81.65% cut off to decide on the optimum number of PCs. Eigen vectors indicates the direction of PCs.

Cumulative values are helping us by calculating the number of PCs up to which the 80% explained variance cut off can be reached.

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]

Figure 21 Feature loading of PCs



From the *figure 21* it can be observed that

PC1 is having high weightage of attributes Top10perc, Top25perc, Expend and PhD (Topper students)

PC2 - F undergrad, Enroll and Accept (Applications)

PC3 – Books, Personal (Expenditure)

PC4 – PhD and Terminal (Faculty qualification)

PC5 – Room Board, Top25perc, Top10perc (Toppers expenditure)

PC6 – Books, SF ratio, Personal (Expenditure)

Now as we have reduced the dimensions it will help us in achieving accurate result at lesser training time.

If we closely analyse the PCs then they can be grouped as per the highest feature loading and mostly we have to look at the components as per No of Topper students, No of Applications, Expenditures, Qualification of the faculties and expenditure of topper students.

Assumption 1: We are trying to find out the admission preference of Toppers after 12th standard

Then we have to study the following components closely in the next steps,

No of applications, no of application of toppers, expenditure, and faculty qualification.

Assumption 2: We are trying to build a model to find out the best features of a university or college so as to help students choose better options.

Then we will proceed with the six principle components found and feed the data to our model for better outcomes.