Advanced Statistics - Business Report

Rohan R. Khade

Table of Contents

Back to Top

List of Tables

Back to Top

List of Figures

Back to Top

Chapter 1. Problem 1A

1.1 Problem Statement

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.]

1.2 Introduction

The dataset has 40 rows and 3 colums. The columns of the dataset include education, occupation, and salary. There are three levels of education provided namely high school graduate, bachelor, and doctorate. The dataset also includes four types of occupations namely administrative and clerical, sales, professional or specialty, and executive or managerial. Below are the details of the dataset provided

Table 1 Dataframe: salary (with head function)

	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
5	Doctorate	Sales	219420
6	Doctorate	Sales	237920
7	Doctorate	Sales	160540
8	Doctorate	Sales	180934
9	Doctorate	Prof-specialty	248156

Back to Top

Table 2 Dataframe: salary (with describe function)

	Salary
count	40.000000
mean	162186.875000
std	64860.407506
min	50103.000000
25%	99897.500000
50%	169100.000000
75%	214440.750000
max	260151.000000

Figure 1. Dataset information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Education   40 non-null     object
 1   Occupation  40 non-null     object
 2   Salary      40 non-null     int64
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

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

For Education

1. Null Hypothesis (H0) = The mean salaries for all the three education categories i.e., High School Graduate, Bachelor, and Doctorate are equal
2. Alternate Hypothesis (H1) = The mean salary is different for at least one of the three education categories i.e., High School Graduate, Bachelor, and Doctorate

For Occupation

1. Null Hypothesis (H0) = The mean salaries for all the four occupation categories i.e., Administrative and Clerical, Sales, Professional or Specialty, and Executive or Managerial are equal
2. Alternate Hypothesis (H1) = The mean salary is different for at least one of the four occupation categories i.e., Administrative and Clerical, Sales, Professional or Specialty, and Executive or Managerial

Back to Top

1.2.2 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.

	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 p value (1.257709e-08) is lower than the level of significance (alpha = 0.05), we reject the null hypothesis and infer that the mean salary of at least one education category is different.

1.2.3 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.

	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 p value (0.458508) is higher than the level of significance (alpha = 0.05), we accept the null hypothesis and infer that the mean salaries for all the four occupation categories is equal.

1.2.4 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)

Since we rejected the null hypothesis was rejected while comparing the means of education category, we can find out the difference in means using Tukey Honest Significant Difference test.

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

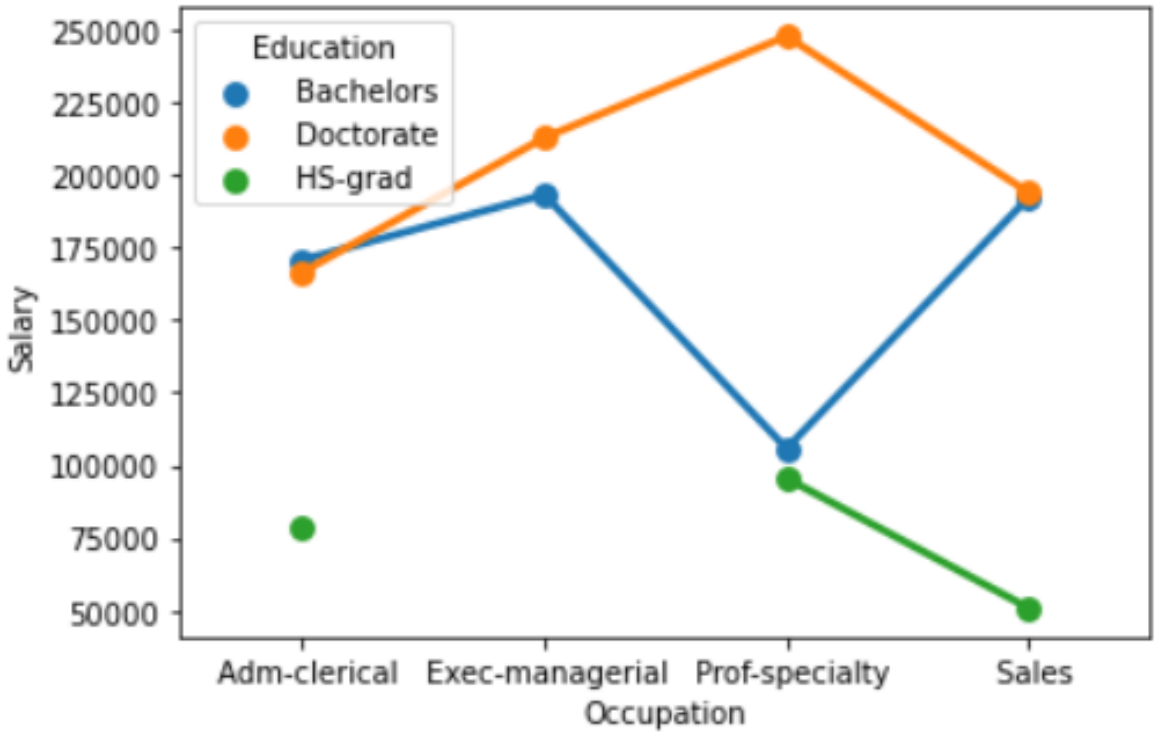
If we compare the first row with the other two, we can see that the p value of the first row differs significantly as compared to the p values of the other two. So, if we compare multiple means, there is a substantial differences across the comparison groups.

Back to Top

Chapter 2. Problem 1B

2.1 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]

Figure 2. Interaction pointplot



Based on the above plot, we can see that there is a significant interaction between the two variables namely occupation and education.

2.1.1 Interpretation

- Analyzing the above plot, high school graduates earn least when they work in sales; however, they earn more when they are working in the administrative & clerical sector and highest when they are working as professional or specialty. The salary difference of high school graduates is not so significant between administrative & clerical and professional or specialty occupation. However, as per the data indicates, high school graduates cannot work at executive or managerial positions.
- Doctorate education seems to fetch highest amount of salary especially when they are working as a professional or specialty occupation. Doctorate education also fetches high salary in the executive or managerial occupation. However, administrative & clerical occupation pays lowest salary to doctorate education.

Back to Top

- Bachelor education fetches highest salary in the sales occupation, and there is zero or negligible difference in the salary for bachelor degree holders in executive or managerial occupation. However, professional or specialty occupation fetches lowest salary for bachelor degree holders.
- High school graduates earn lowest whereas people with doctorates earn highest among the education category.

2.2 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?

2.2.1 Hypothesis

1. Null Hypothesis (H0) = The mean salaries do not vary due to the two variables namely occupation and education i.e., there is no interaction effect between the two variables

2. Alternative Hypothesis (H1) = There is an interaction effect between the variables education and occupation on the mean salary

Figure 3. Two-way ANOVA

	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

2.2.2 Interpretation

As p value (2.2325500e-05) is lower than the level of significance (alpha = 0.05), we reject the null hypothesis. We can also see that there is considerable interaction between the two variables. When we combine education and occupation it results in higher salaries. As we can infer that doctorate holders earn higher as compared to high school graduates. This clearly concludes that education and occupation has significant ipact on the salary.

Back to Top

Chapter 3. Problem 2

3.1 Problem Statement

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

3.2 Introduction

The data provides the information on the students, the cost associated with their education from different universities. The dataset provides cost of books and room and board. The dataset also provides personal spending estimates of the students and information on percentage of faculties with Ph.D.'s among others. Below is the data dictionary for your reference:

3.2.1 Data Dictionary

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

Back to Top

Table 3 Dataframe: edu (with head function)

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15
5	Albertson College	587	479	158	38	62	678	41	13500	3335	500	675	67	73	9.4	11	9727	55
6	Albertus Magnus College	353	340	103	17	45	416	230	13290	5720	500	1500	90	93	11.5	26	8861	63
7	Albion College	1899	1720	489	37	68	1594	32	13868	4826	450	850	89	100	13.7	37	11487	73
8	Albright College	1038	839	227	30	63	973	306	15595	4400	300	500	79	84	11.3	23	11644	80
9	Alderson-Broaddus College	582	498	172	21	44	799	78	10468	3380	660	1800	40	41	11.5	15	8991	52

Table 4 Dataframe: edu (with describe function)

	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

Back to Top

Table 5 Dataframe: edu (with info function)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 18 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Names        777 non-null    object
 1   Apps         777 non-null    int64
 2   Accept       777 non-null    int64
 3   Enroll       777 non-null    int64
 4   Top10perc    777 non-null    int64
 5   Top25perc    777 non-null    int64
 6   F.Undergrad  777 non-null    int64
 7   P.Undergrad  777 non-null    int64
 8   Outstate     777 non-null    int64
 9   Room.Board   777 non-null    int64
 10  Books        777 non-null    int64
 11  Personal     777 non-null    int64
 12  PhD          777 non-null    int64
 13  Terminal     777 non-null    int64
 14  S.F.Ratio    777 non-null    float64
 15  perc.alumni  777 non-null    int64
 16  Expend       777 non-null    int64
 17  Grad.Rate    777 non-null    int64
dtypes: float64(1), int64(16), object(1)
memory usage: 109.4+ KB
```

Table 6 Dataframe: edu (to check null values)

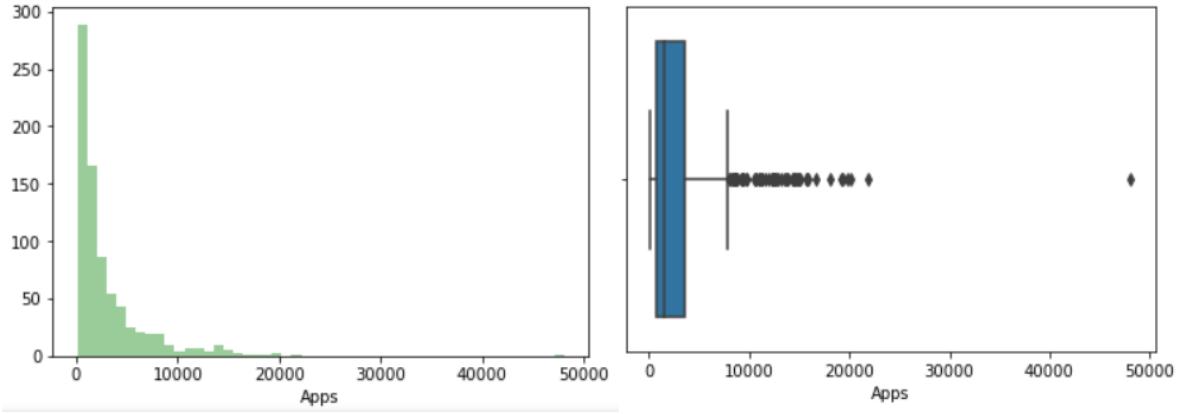
```
Names          0
Apps           0
Accept         0
Enroll         0
Top10perc      0
Top25perc      0
F.Undergrad    0
P.Undergrad    0
Outstate       0
Room.Board     0
Books          0
Personal       0
PhD            0
Terminal       0
S.F.Ratio      0
perc.alumni    0
Expend         0
Grad.Rate      0
dtype: int64
```

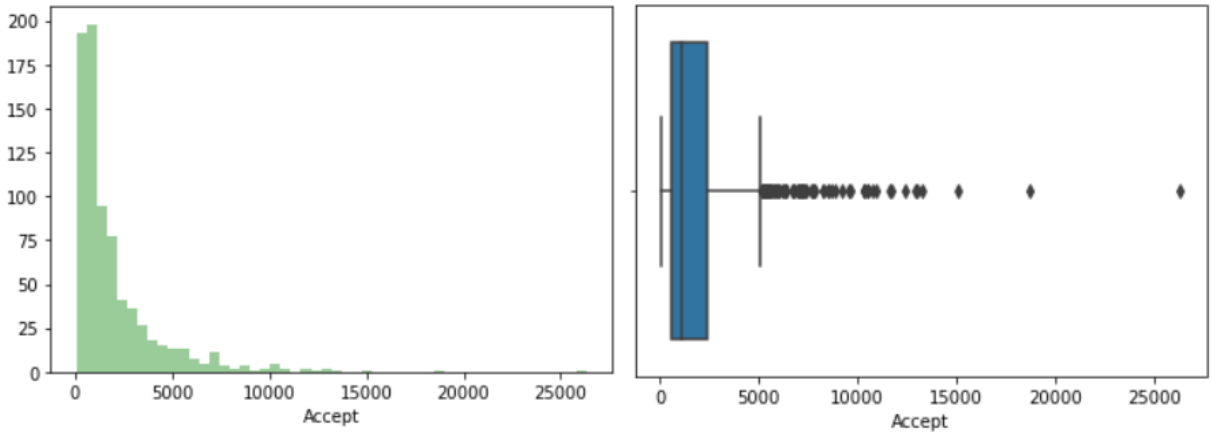
We have also checked if the dataset includes duplicate values and realized that there are no duplicate values.

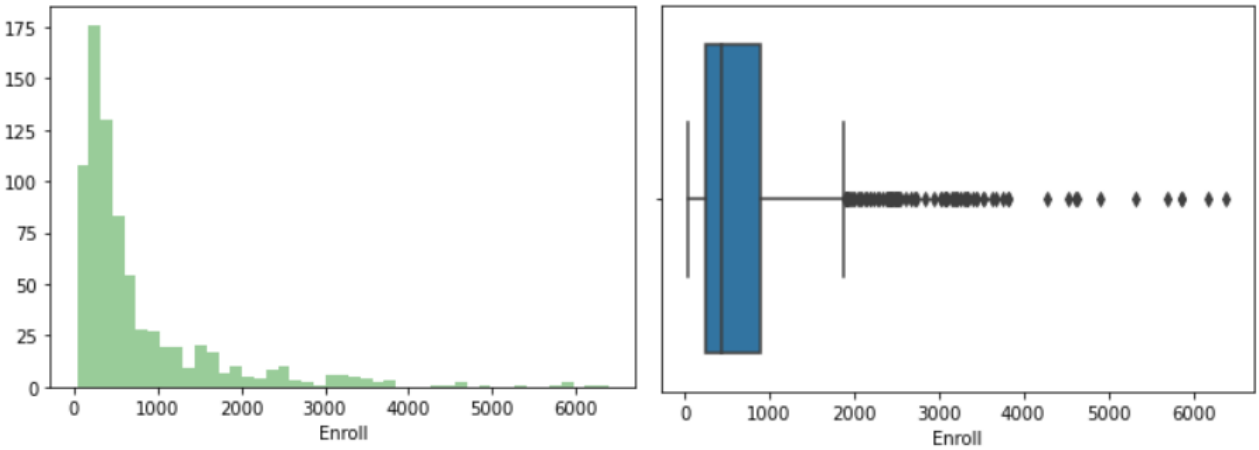
Back to Top

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

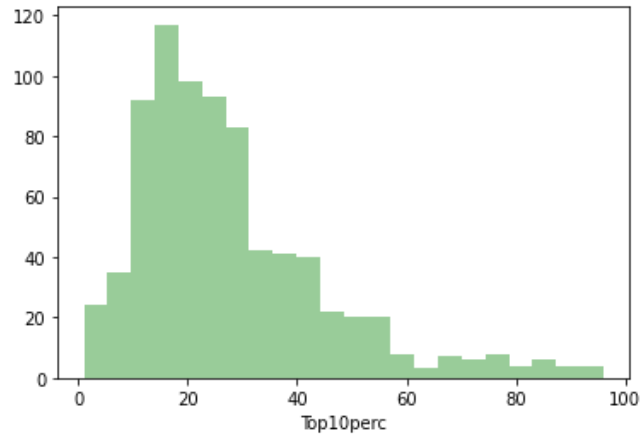
3.2.2.1 Univariate & Multivariate Analysis

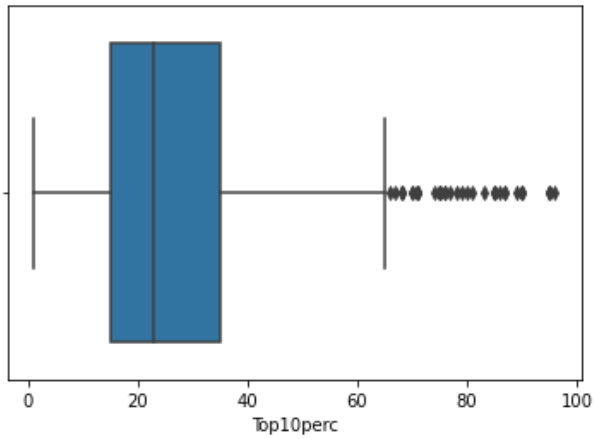


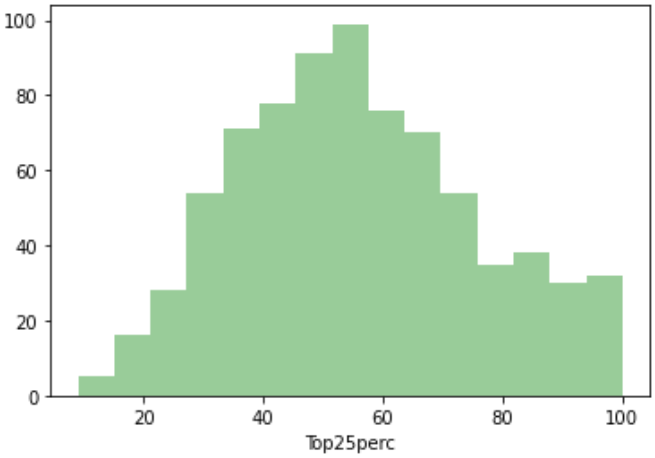


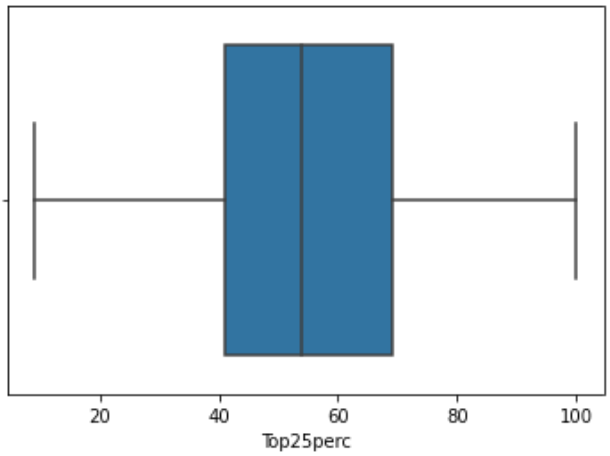


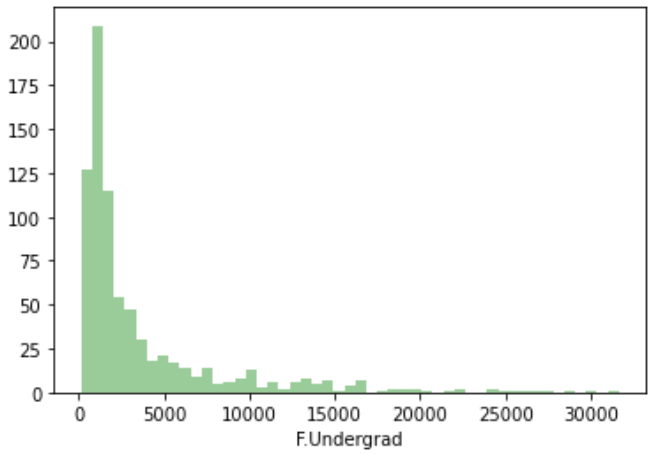
Back to Top

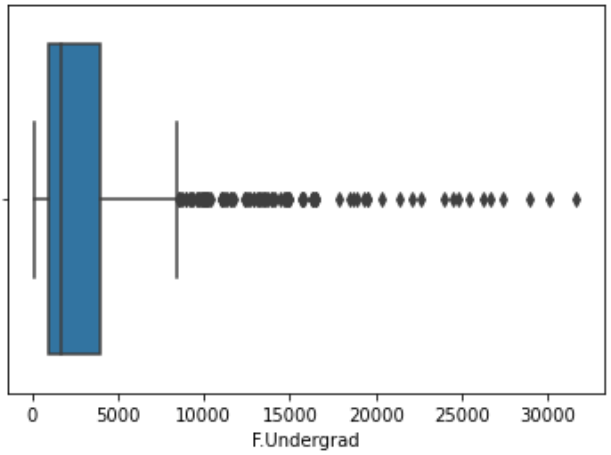

120
100
80
60
40
20
0
0
20
40
60
80
100
Top10perc


0
20
40
60
80
100
Top10perc

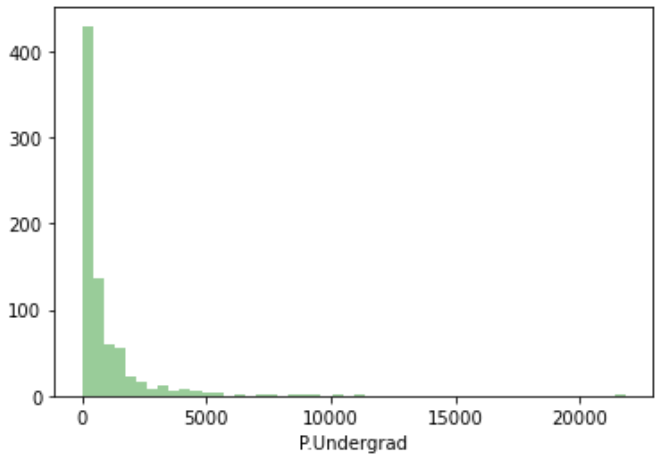

100
80
60
40
20
0
20
40
60
80
100
Top25perc

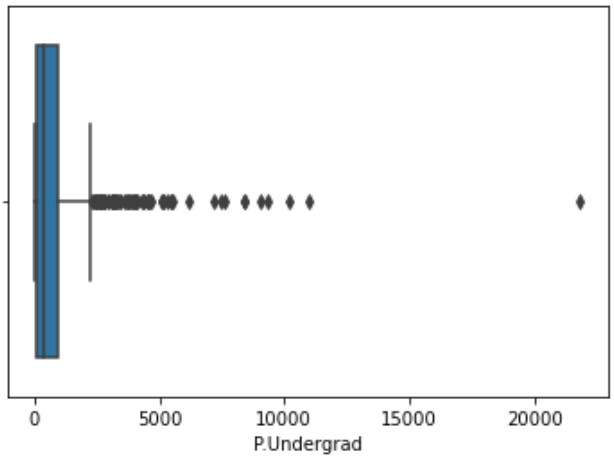

20
40
60
80
100
Top25perc

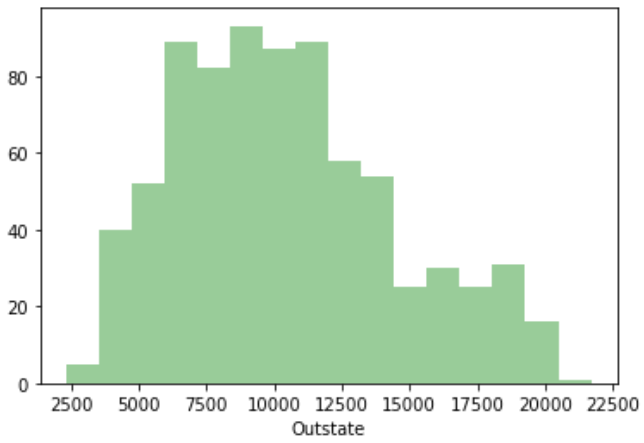

200
175
150
125
100
75
50
25
0
0
5000
10000
15000
20000
25000
30000
F.Undergrad

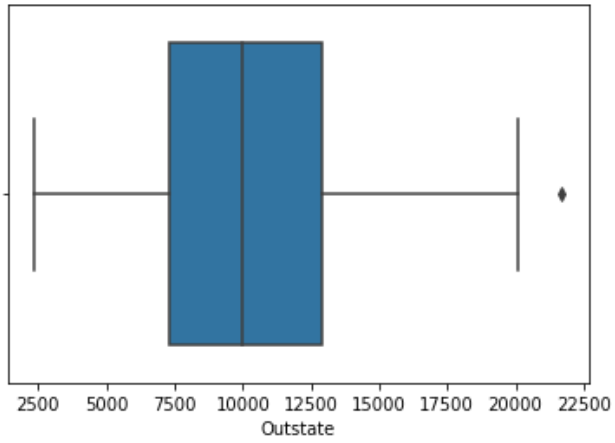

0
5000
10000
15000
20000
25000
30000
F.Undergrad

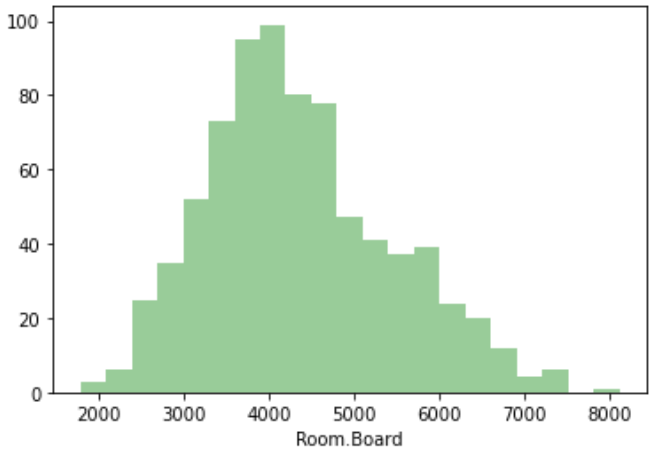
Back to Top

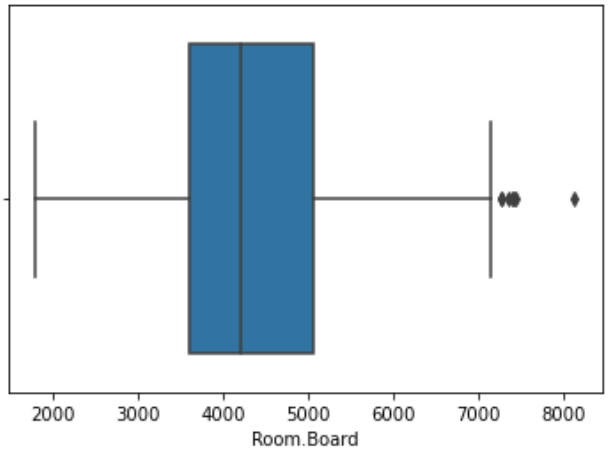

400
300
200
100
0
0
5000
10000
15000
20000
P.Undergrad


0
5000
10000
15000
20000
P.Undergrad

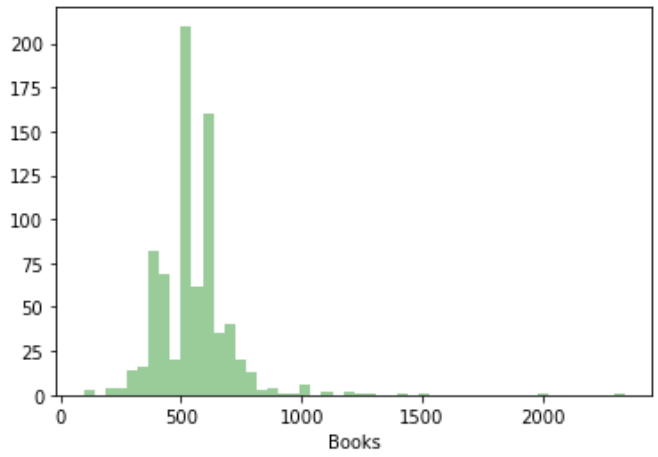

80
60
40
20
0
2500
5000
7500
10000
12500
15000
17500
20000
22500
Outstate

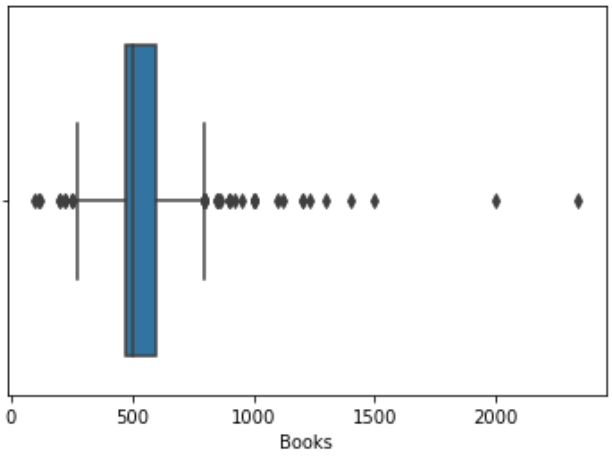

2500
5000
7500
10000
12500
15000
17500
20000
22500
Outstate

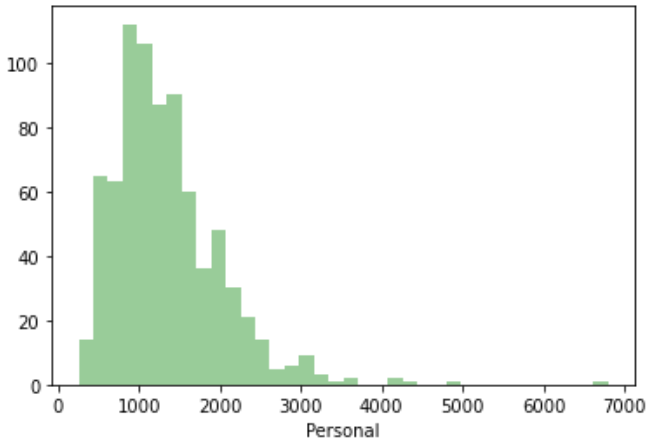

100
80
60
40
20
0
2000
3000
4000
5000
6000
7000
8000
Room.Board

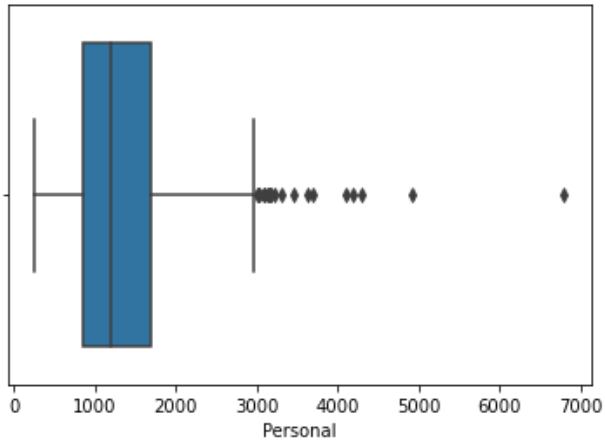

2000
3000
4000
5000
6000
7000
8000
Room.Board

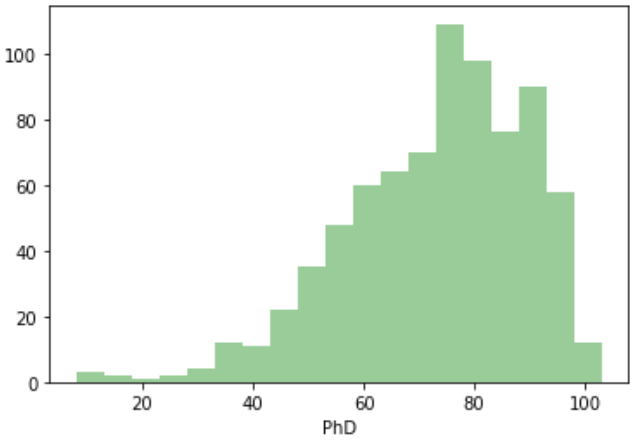
Back to Top

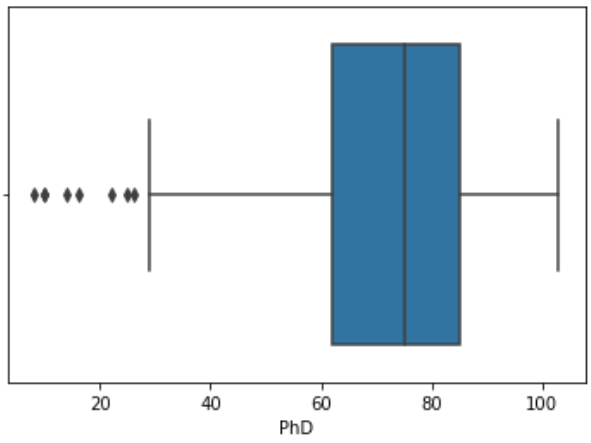

200
175
150
125
100
75
50
25
0
0
500
1000
1500
2000
Books


0
500
1000
1500
2000
Books

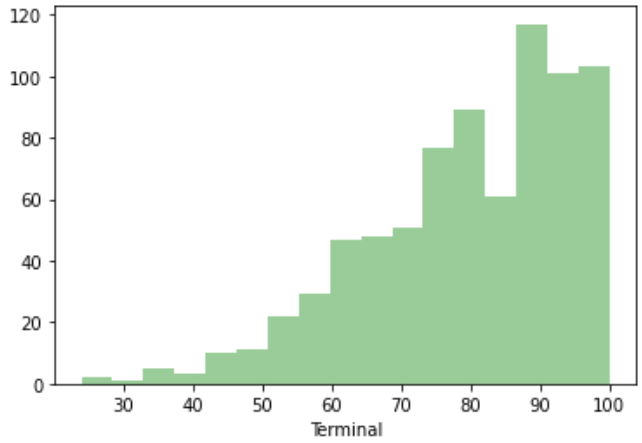

100
80
60
40
20
0
0
1000
2000
3000
4000
5000
6000
7000
Personal

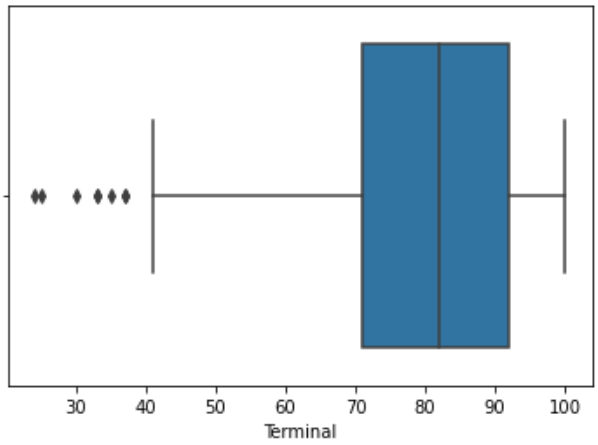

0
1000
2000
3000
4000
5000
6000
7000
Personal

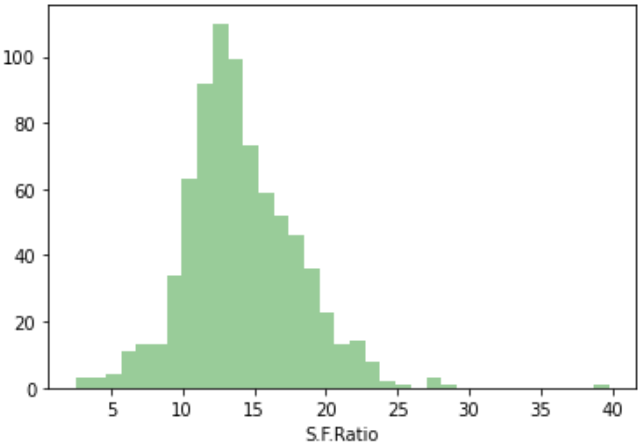

100
80
60
40
20
0
20
40
60
80
100
PhD

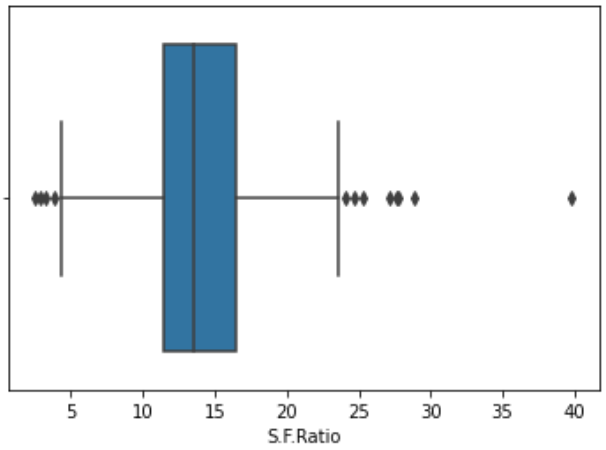

20
40
60
80
100
PhD

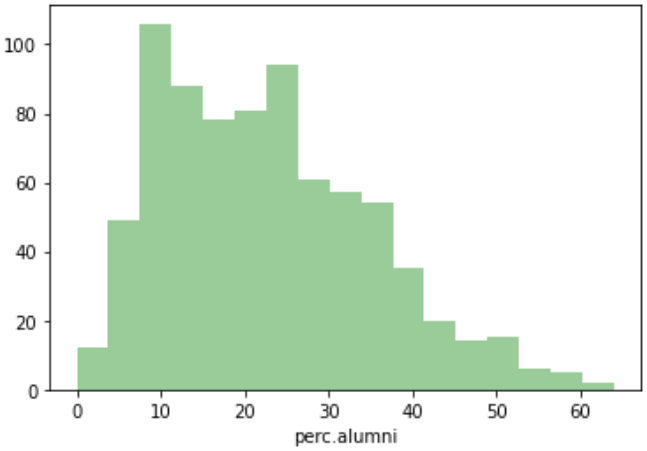
Back to Top

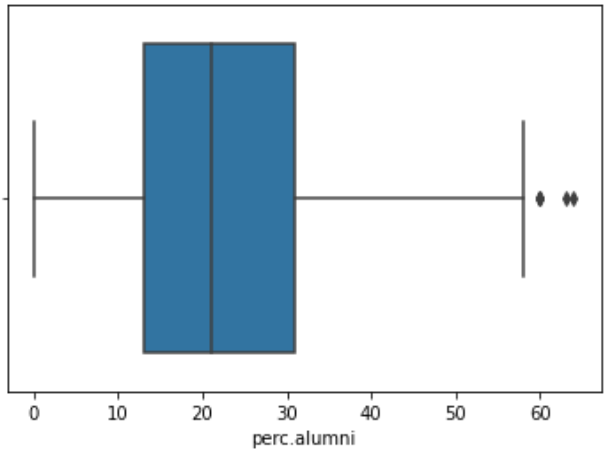

120
100
80
60
40
20
0
30
40
50
60
70
80
90
100
Terminal


30
40
50
60
70
80
90
100
Terminal

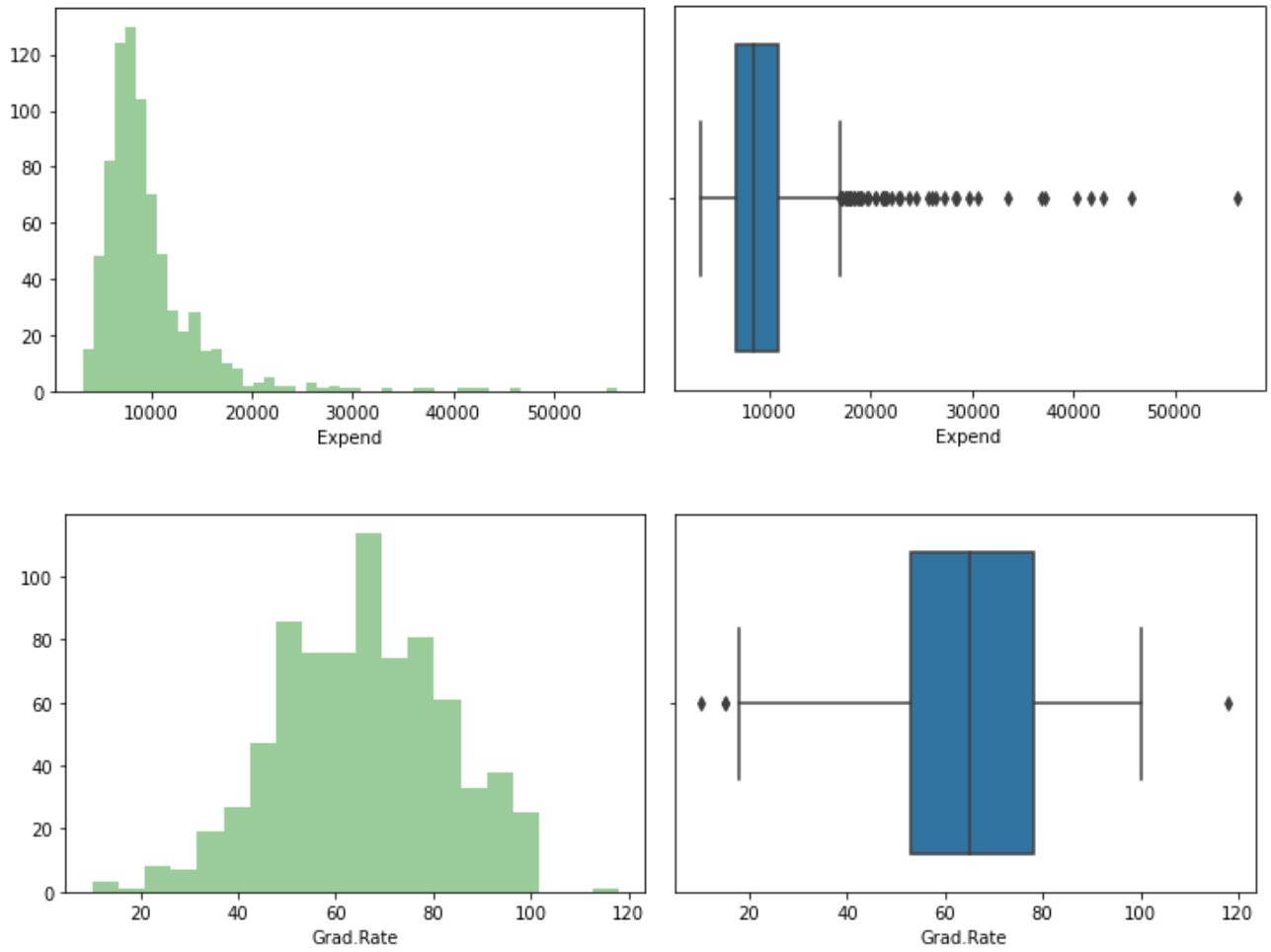

100
80
60
40
20
0
5
10
15
20
25
30
35
40
S.F.Ratio


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


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


0
10
20
30
40
50
60
perc.alumni

Back to Top

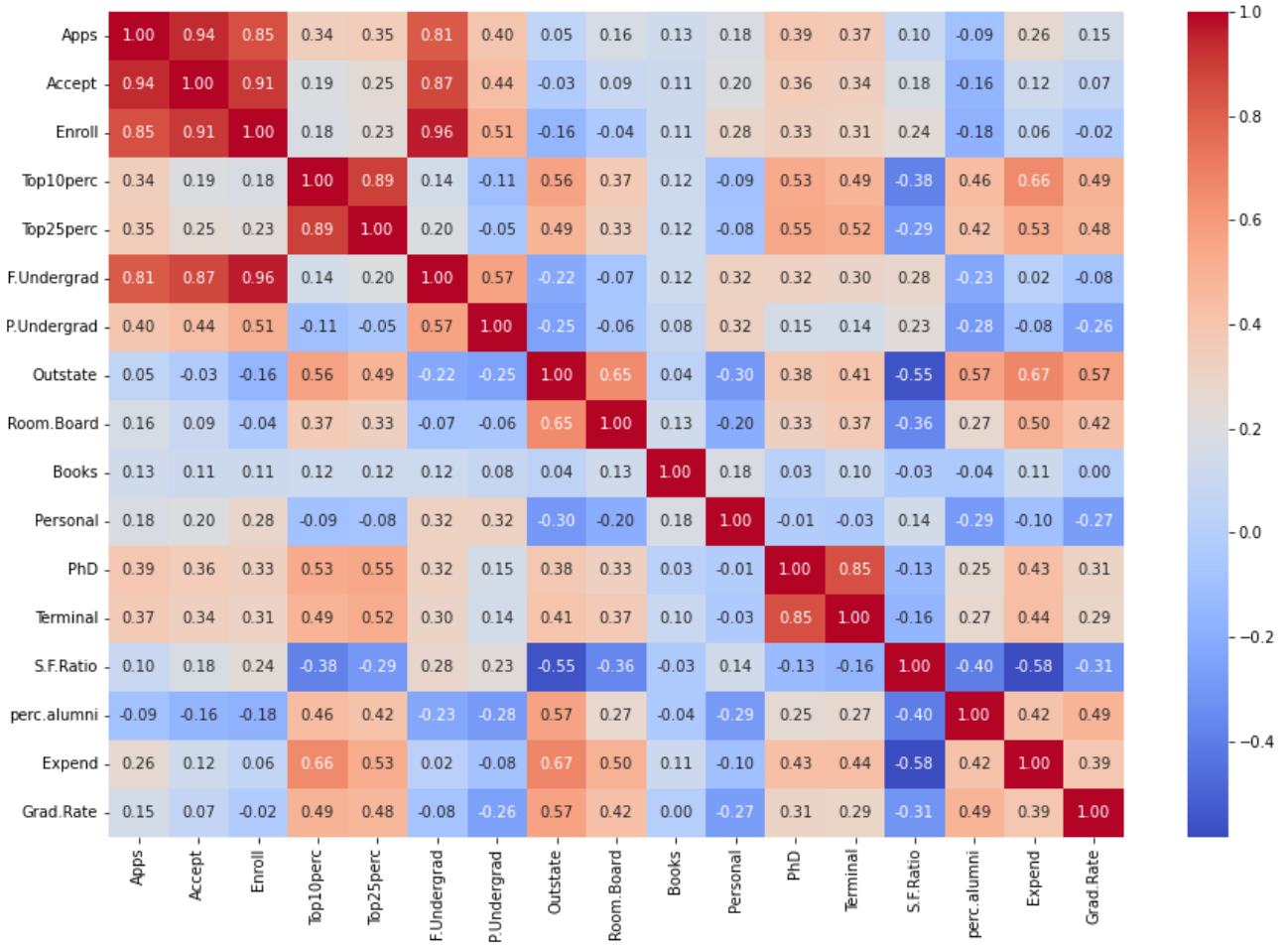


3.2.2.2 Interpretation

"Top 25 perc", "Room.Board", "Outstate" are nearly centrally distributed with a little rightly skewed. "Grad Rate" also is almost centrally distributed with little left skewed. The mentioned variables have lower number of outliers as compared to other variables. "Accept", “Apps”, “Books”, “Expend”, and “P.Undergrad” variables are heavily right skewed with lot of outliers on the right side of the whisker. Whereas, "Terminal" and "PhD" are heavily left-skewed with lot of outliers on the left side of the whisker.

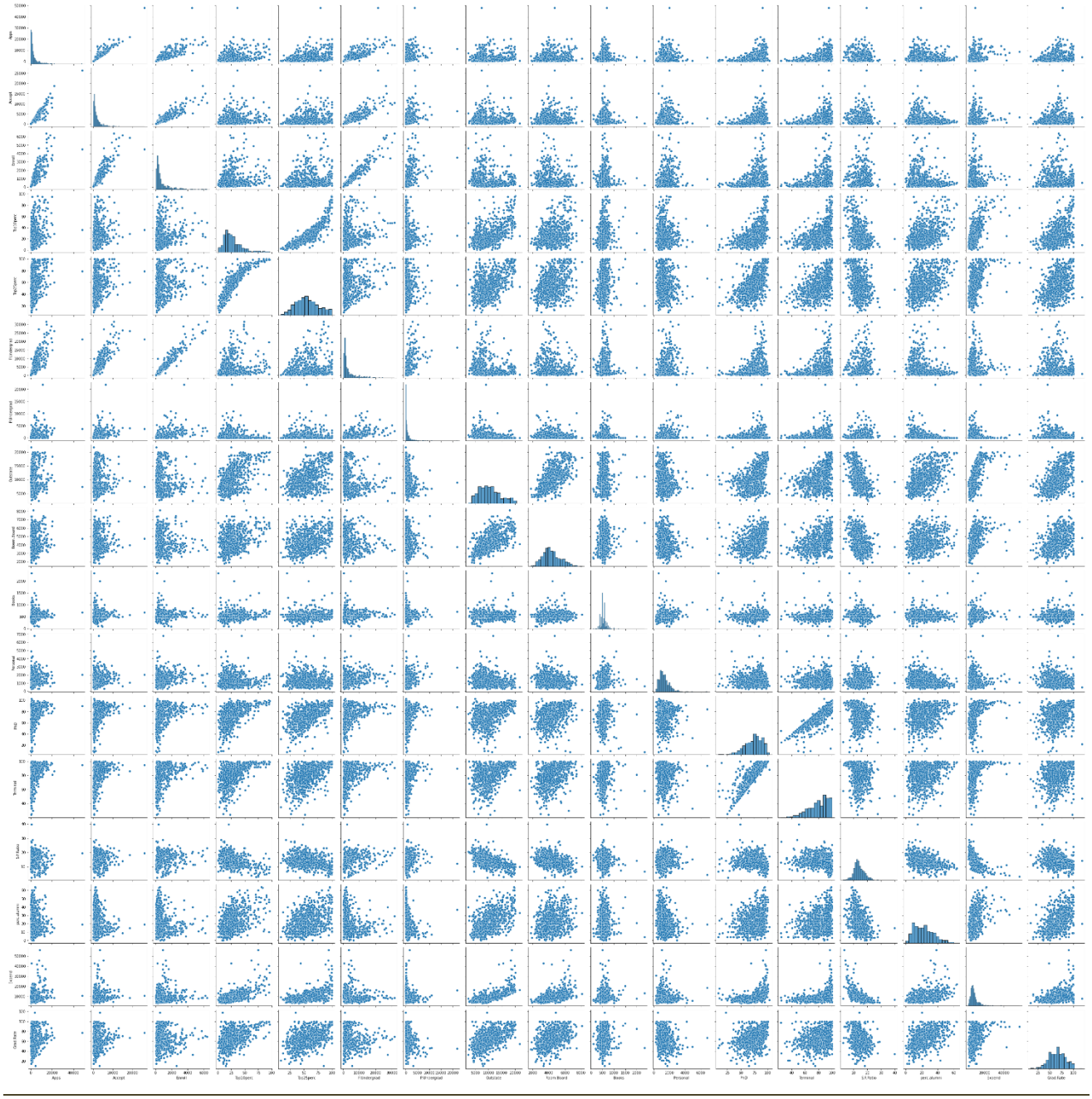
Back to Top

Figure 4. Heatmap



Back to Top

Figure 5. Pairplot



Highly negative correlation between variables

- Between “Top10Perc, Top25Perc” and “S.F.Ratio”
- Between “outstate” and “personal”
- Between “outstate” and “S.F.Ratio”
- Between “Room.Board” and “S.F.Ratio”
- Between “S.F. Ratio” and “Expend”
- Between “Grad Rate” and “S.F. Ratio”
- Between “Grad Rate” and “Personal”
- Between “Grad Rate” and “P.Undergrad”

Back to Top

Moderately high positive correlation (+0.4 to +0.69)

- Between “Top10Perc” and “Expend
- Between “Outstate” and “Room Board
- Between “Outstate” and “Expend
- Between “Top10Perc” and “Outstate”
- Between “Grad Rate” and “Top 10 Perc”
- Between “Grad Rate” and “Top 25 Perc”
- Between “Grad Rate” and “Outstate”
- Between “Grad Rate” and “Room Board”
- Between “Grad Rate” and “Perc Alumni”

High positive correlation (+0.7 to +0.84)

- Between “F.Undergrad” and “Apps”, “Accept”, “Enroll”

Very high positive correlation (+0.85 and above)

- Among the variables “Apps”, “Accept”, “Enroll”
- Between “PhD” and “Terminal
- Between “Top 10 perc” and “Top 25 perc”

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

Yes, scaling the data is necessary as it helps to normalize the data within a particular range which eases algorithmic calculations. There are three types of scaling methods which we can use on the original dataset (edu) as follows,

- Standard Scaler
- Minmax Scaler
- Logarithmic scaler

We have used standard scaler method to scale the data wherein we have transformed values into their corresponding zscore, below is the given table after performing the same and forming a new dataset called edu_num_scaled.

Table 7 Dataframe: edu_num_scaled (with corresponding zscore)

Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio
-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	1.013776
-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	-0.477704
-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	-0.300749
-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	-1.615274
-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	-0.553542
-0.624307	-0.628611	-0.669812	0.592287	0.313426	-0.623421	-0.535212	0.760947	-0.932970	-0.299280	-0.983753	-0.346878	-0.455567	-1.185526
-0.684808	-0.685356	-0.729043	-0.598931	-0.545505	-0.677472	-0.410988	0.708713	1.243144	-0.299280	0.235515	1.062639	0.903786	-0.654660
-0.285088	-0.121984	-0.313353	0.535563	0.616579	-0.434450	-0.541127	0.852479	0.427443	-0.602312	-0.725120	1.001356	1.379560	-0.098515
-0.507700	-0.481644	-0.595505	0.138490	0.363952	-0.562562	-0.361036	1.282036	0.038754	-1.511408	-1.242385	0.388522	0.292077	-0.705218
-0.625600	-0.620854	-0.654735	-0.372032	-0.596031	-0.598459	-0.510893	0.006798	-0.891911	0.670422	0.678885	-2.001529	-2.630532	-0.654660

Table 8 Dataframe: edu_num_scaled (with describe function)

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
count	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02
mean	6.355797e-17	6.774575e-17	-5.249269e-17	-2.753232e-17	-1.546739e-16	-1.661405e-16	-3.029180e-17	6.515595e-17	3.570717e-16	-2.192583e-16
std	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00
min	-7.551337e-01	-7.947645e-01	-8.022728e-01	-1.506526e+00	-2.364419e+00	-7.346169e-01	-5.615022e-01	-2.014878e+00	-2.351778e+00	-2.747779e+00
25%	-5.754408e-01	-5.775805e-01	-5.793514e-01	-7.123803e-01	-7.476067e-01	-5.586426e-01	-4.997191e-01	-7.762035e-01	-6.939170e-01	-4.810994e-01
50%	-3.732540e-01	-3.710108e-01	-3.725836e-01	-2.585828e-01	-9.077663e-02	-4.111378e-01	-3.301442e-01	-1.120949e-01	-1.437297e-01	-2.992802e-01
75%	1.609122e-01	1.654173e-01	1.314128e-01	4.221134e-01	6.671042e-01	6.294077e-02	7.341765e-02	6.179271e-01	6.318245e-01	3.067838e-01
max	1.165867e+01	9.924816e+00	6.043678e+00	3.882319e+00	2.233391e+00	5.764674e+00	1.378992e+01	2.800531e+00	3.436593e+00	1.085230e+01

Back to Top

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

Table 9 Correlation of the scaled dataset (edu_num_scaled)

	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

Table 10 Covariance of the scaled dataset (edu_num_scaled)

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termin
Apps	1.001289	0.944666	0.847913	0.339270	0.352093	0.815540	0.398777	0.050224	0.165152	0.132729	0.178961	0.391201	0.3699
Accept	0.944666	1.001289	0.912811	0.192695	0.247795	0.875350	0.441839	-0.025788	0.091016	0.113672	0.201248	0.356216	0.3380
Enroll	0.847913	0.912811	1.001289	0.181527	0.227037	0.965883	0.513730	-0.155678	-0.040284	0.112856	0.281291	0.331896	0.3086
Top10perc	0.339270	0.192695	0.181527	1.001289	0.893144	0.141471	-0.105492	0.563055	0.371959	0.119012	-0.093437	0.532513	0.4917
Top25perc	0.352093	0.247795	0.227037	0.893144	1.001289	0.199702	-0.053646	0.490024	0.331917	0.115676	-0.080914	0.546566	0.5254
F.Undergrad	0.815540	0.875350	0.965883	0.141471	0.199702	1.001289	0.571247	-0.216020	-0.068979	0.115699	0.317608	0.318747	0.3004
P.Undergrad	0.398777	0.441839	0.513730	-0.105492	-0.053646	0.571247	1.001289	-0.253839	-0.061405	0.081304	0.320294	0.149306	0.1420
Outstate	0.050224	-0.025788	-0.155678	0.563055	0.490024	-0.216020	-0.253839	1.001289	0.655100	0.038905	-0.299472	0.383476	0.4085
Room.Board	0.165152	0.091016	-0.040284	0.371959	0.331917	-0.068979	-0.061405	0.655100	1.001289	0.128128	-0.199685	0.329627	0.3750
Books	0.132729	0.113672	0.112856	0.119012	0.115676	0.115699	0.081304	0.038905	0.128128	1.001289	0.179526	0.026940	0.1000
Personal	0.178961	0.201248	0.281291	-0.093437	-0.080914	0.317608	0.320294	-0.299472	-0.199685	0.179526	1.001289	-0.010950	-0.0306
PhD	0.391201	0.356216	0.331896	0.532513	0.546566	0.318747	0.149306	0.383476	0.329627	0.026940	-0.010950	1.001289	0.8506
Terminal	0.369968	0.338018	0.308671	0.491768	0.525425	0.300406	0.142086	0.408509	0.375022	0.100084	-0.030653	0.850682	1.0012
S.F.Ratio	0.095756	0.176456	0.237577	-0.385370	-0.295009	0.280064	0.232830	-0.555536	-0.363095	-0.031970	0.136521	-0.130698	-0.1603
perc.alumni	-0.090342	-0.160196	-0.181027	0.456072	0.418403	-0.229758	-0.281154	0.566992	0.272714	-0.040260	-0.286337	0.249330	0.2674
Expend	0.259927	0.124878	0.064252	0.661765	0.528127	0.018676	-0.083676	0.673646	0.502386	0.112554	-0.098018	0.433319	0.4393
Grad.Rate	0.146944	0.067399	-0.022370	0.495627	0.477896	-0.078875	-0.257332	0.572026	0.425489	0.001062	-0.269691	0.305431	0.2899

Back to Top

3.2.4.1 Interpretation

Post scaling the dataset, i.e., normalizing it, we can see that the covariance and correlation matrix are pretty similar and identical, which shows that normalizing the data transforms covariance into correlation matrix.

3.2.5 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]

Figure 6. Boxplot of original dataset

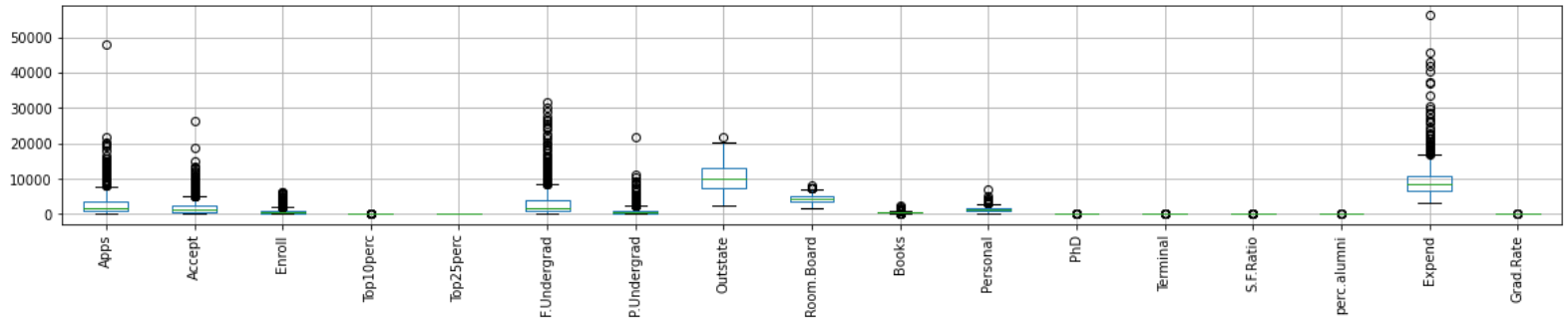
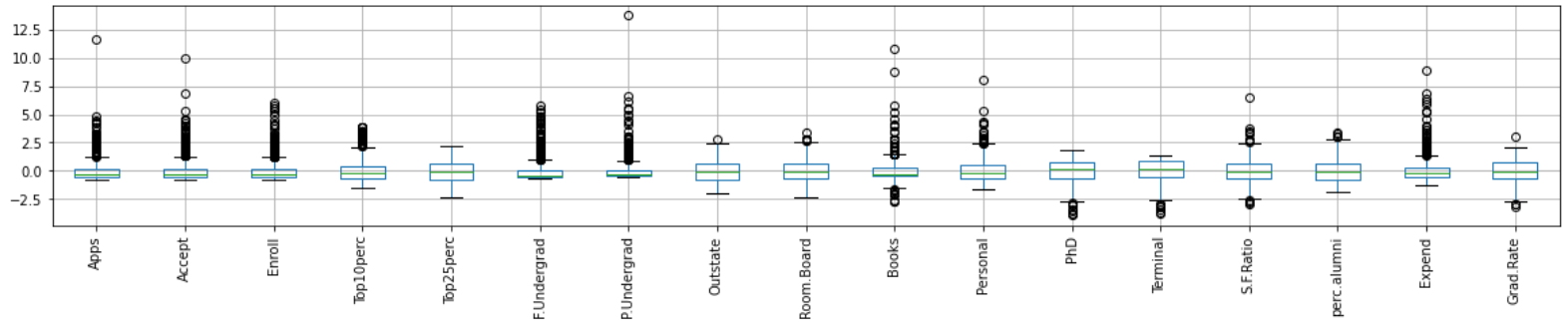


Figure 7. Boxplot of scaled dataset



3.2.5.1 Interpretation

We can see that post scaling the original data, there are several significant outliers for most of the variables. Scaling has definitely helped in treating the outliers to some extent; however, we can clearly say that outlier treatment is necessary before performing PCA.

Back to Top

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

```
Eigen Vectors
%s [[-2.48765602e-01  3.31598227e-01  6.30921033e-02 -2.81310530e-01
    5.74140964e-03  1.62374420e-02  4.24863486e-02  1.03090398e-01
    9.02270802e-02 -5.25098025e-02  3.58970400e-01 -4.59139498e-01
    4.30462074e-02 -1.33405806e-01  8.06328039e-02 -5.95830975e-01
    2.40709086e-02]
 [-2.07601502e-01  3.72116750e-01  1.01249056e-01 -2.67817346e-01
    5.57860920e-02 -7.53468452e-03  1.29497196e-02  5.62709623e-02
    1.77864814e-01 -4.11400844e-02 -5.43427250e-01  5.18568789e-01
   -5.84055850e-02  1.45497511e-01  3.34674281e-02 -2.92642398e-01
   -1.45102446e-01]
 [-1.76303592e-01  4.03724252e-01  8.29855709e-02 -1.61826771e-01
   -5.56936353e-02  4.25579803e-02  2.76928937e-02 -5.86623552e-02
    1.28560713e-01 -3.44879147e-02  6.09651110e-01  4.04318439e-01
   -6.93988831e-02 -2.95896092e-02 -8.56967180e-02  4.44638207e-01
    1.11431545e-02]
 [-3.54273947e-01 -8.24118211e-02 -3.50555339e-02  5.15472524e-02
   -3.95434345e-01  5.26927980e-02  1.61332069e-01  1.22678028e-01
   -3.41099863e-01 -6.40257785e-02 -1.44986329e-01  1.48738723e-01
   -8.10481404e-03 -6.97722522e-01 -1.07828189e-01 -1.02303616e-03
    3.85543001e-02]
 [-3.44001279e-01 -4.47786551e-02  2.41479376e-02  1.09766541e-01
   -4.26533594e-01 -3.30915896e-02  1.18485556e-01  1.02491967e-01
   -4.03711989e-01 -1.45492289e-02  8.03478445e-02 -5.18683400e-02
   -2.73128469e-01  6.17274818e-01  1.51742110e-01 -2.18838802e-02
   -8.93515563e-02]

[-1.54640962e-01  4.17673774e-01  6.13929764e-02 -1.00412335e-01
  -4.34543659e-02  4.34542349e-02  2.50763629e-02 -7.88896442e-02
   5.94419181e-02 -2.08471834e-02 -4.14705279e-01 -5.60363054e-01
  -8.11578181e-02 -9.91640992e-03 -5.63728817e-02  5.23622267e-01
   5.61767721e-02]
 [-2.64425045e-02  3.15087830e-01 -1.39681716e-01  1.58558487e-01
   3.02385408e-01  1.91198583e-01 -6.10423460e-02 -5.70783816e-01
  -5.60672902e-01  2.23105808e-01  9.01788964e-03  5.27313042e-02
   1.00693324e-01 -2.09515982e-02  1.92857500e-02 -1.25997650e-01
  -6.35360730e-02]
 [-2.94736419e-01 -2.49643522e-01 -4.65988731e-02 -1.31291364e-01
   2.22532003e-01  3.00003910e-02 -1.08528966e-01 -9.84599754e-03
   4.57332880e-03 -1.86675363e-01  5.08995918e-02 -1.01594830e-01
   1.43220673e-01 -3.83544794e-02 -3.40115407e-02  1.41856014e-01
  -8.23443779e-01]
 [-2.49030449e-01 -1.37808883e-01 -1.48967389e-01 -1.84995991e-01
   5.60919470e-01 -1.62755446e-01 -2.09744235e-01  2.21453442e-01
  -2.75022548e-01 -2.98324237e-01  1.14639620e-03  2.59293381e-02
  -3.59321731e-01 -3.40197083e-03 -5.84289756e-02  6.97485854e-02
   3.54559731e-01]
 [-6.47575181e-02  5.63418434e-02 -6.77411649e-01 -8.70892205e-02
  -1.27288825e-01 -6.41054950e-01  1.49692034e-01 -2.13293009e-01
   1.33663353e-01  8.20292186e-02  7.72631963e-04 -2.88282896e-03
   3.19400370e-02  9.43887925e-03 -6.68494643e-02 -1.14379958e-02
  -2.81593679e-02]
 [ 4.25285386e-02  2.19929218e-01 -4.99721120e-01  2.30710568e-01
  -2.22311021e-01  3.31398003e-01 -6.33790064e-01  2.32660840e-01
   9.44688900e-02 -1.36027616e-01 -1.11433396e-03  1.28904022e-02
  -1.85784733e-02  3.09001353e-03  2.75286207e-02 -3.94547417e-02
  -3.92640266e-02]
```


Back to Top

```
[-3.18312875e-01  5.83113174e-02  1.27028371e-01  5.34724832e-01
  1.40166326e-01 -9.12555212e-02  1.09641298e-03  7.70400002e-02
  1.85181525e-01  1.23452200e-01  1.38133366e-02 -2.98075465e-02
  4.03723253e-02  1.12055599e-01 -6.91126145e-01 -1.27696382e-01
  2.32224316e-02]
[-3.17056016e-01  4.64294477e-02  6.60375454e-02  5.19443019e-01
  2.04719730e-01 -1.54927646e-01  2.84770105e-02  1.21613297e-02
  2.54938198e-01  8.85784627e-02  6.20932749e-03  2.70759809e-02
 -5.89734026e-02 -1.58909651e-01  6.71008607e-01  5.83134662e-02
  1.64850420e-02]
[ 1.76957895e-01  2.46665277e-01  2.89848401e-01  1.61189487e-01
 -7.93882496e-02 -4.87045875e-01 -2.19259358e-01  8.36048735e-02
 -2.74544380e-01 -4.72045249e-01 -2.22215182e-03  2.12476294e-02
  4.45000727e-01  2.08991284e-02  4.13740967e-02  1.77152700e-02
 -1.10262122e-02]
[-2.05082369e-01 -2.46595274e-01  1.46989274e-01 -1.73142230e-02
 -2.16297411e-01  4.73400144e-02 -2.43321156e-01 -6.78523654e-01
  2.55334907e-01 -4.22999706e-01 -1.91869743e-02 -3.33406243e-03
 -1.30727978e-01  8.41789410e-03 -2.71542091e-02 -1.04088088e-01
  1.82660654e-01]
[-3.18908750e-01 -1.31689865e-01 -2.26743985e-01 -7.92734946e-02
  7.59581203e-02  2.98118619e-01  2.26584481e-01  5.41593771e-02
  4.91388809e-02 -1.32286331e-01 -3.53098218e-02  4.38803230e-02
  6.92088870e-01  2.27742017e-01  7.31225166e-02  9.37464497e-02
  3.25982295e-01]
[-2.52315654e-01 -1.69240532e-01  2.08064649e-01 -2.69129066e-01
 -1.09267913e-01 -2.16163313e-01 -5.59943937e-01  5.33553891e-03
 -4.19043052e-02  5.90271067e-01 -1.30710024e-02  5.00844705e-03
  2.19839000e-01  3.39433604e-03  3.64767385e-02  6.91969778e-02
  1.22106697e-01]]
```

```
  Eigen Values
 %s [5.44350679 4.47783645 1.17315581 1.00690817 0.93302887 0.84739916
  0.60500815 0.58711563 0.52992973 0.40378256 0.02299823 0.03667818
  0.31304247 0.08791135 0.1437932  0.1675782  0.22032704]
```

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

We have performed PCA for 5 components as we will see that there is an elbow that forms in the graph after the 5th component. As a result, while considering the same we performed rest of the functions

```
array([[-1.59285541e+00, -2.19240180e+00, -1.43096371e+00, ...,
        -7.32560609e-01,  7.91932738e+00, -4.69508074e-01],
       [ 7.67333505e-01, -5.78829992e-01, -1.09281889e+00, ...,
        -7.72352595e-02, -2.06832882e+00,  3.66660931e-01],
       [-1.01074098e-01,  2.27879717e+00, -4.38092790e-01, ...,
        -4.07886582e-04,  2.07356790e+00, -1.32891651e+00],
       [-9.21749098e-01,  3.58891883e+00,  6.77240513e-01, ...,
         5.43176160e-02,  8.52051441e-01, -1.08021761e-01],
       [-7.43975014e-01,  1.05999711e+00, -3.69613287e-01, ...,
        -5.16019788e-01, -9.47757176e-01, -1.13217511e+00]])
```


Back to Top

3.2.7.1 Variance

array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123])

3.2.7.2 PCA components

```
array([[ 0.2487656 ,  0.2076015 ,  0.17630359,  0.35427395,  0.34400128,
         0.15464096,  0.0264425 ,  0.29473642,  0.24903045,  0.06475752,
        -0.04252854,  0.31831287,  0.31705602, -0.17695789,  0.20508237,
         0.31890875,  0.25231565],
       [ 0.33159823,  0.37211675,  0.40372425, -0.08241182, -0.04477866,
         0.41767377,  0.31508783, -0.24964352, -0.13780888,  0.05634184,
         0.21992922,  0.05831132,  0.04642945,  0.24666528, -0.24659527,
        -0.13168986, -0.16924053],
       [-0.06309208, -0.10124908, -0.08298559,  0.03505553, -0.02414794,
        -0.06139295,  0.13968171,  0.04659888,  0.14896739,  0.67741165,
         0.49972112, -0.12702837, -0.06603755, -0.2898484 , -0.14698927,
         0.22674398, -0.20806465],
       [ 0.28131054,  0.26781734,  0.16182676, -0.05154725, -0.10976654,
         0.10041235, -0.15855849,  0.13129137,  0.18499599,  0.08708922,
        -0.23071057, -0.53472483, -0.51944302, -0.16118949,  0.01731422,
         0.07927349,  0.26912907],
       [ 0.00574153,  0.05578595, -0.05569371, -0.39543437, -0.42653359,
        -0.04345425,  0.3023854 ,  0.22253203,  0.56091946, -0.12728882,
        -0.22231102,  0.14016633,  0.20471973, -0.07938825, -0.21629741,
         0.07595811, -0.10926791]])
```

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

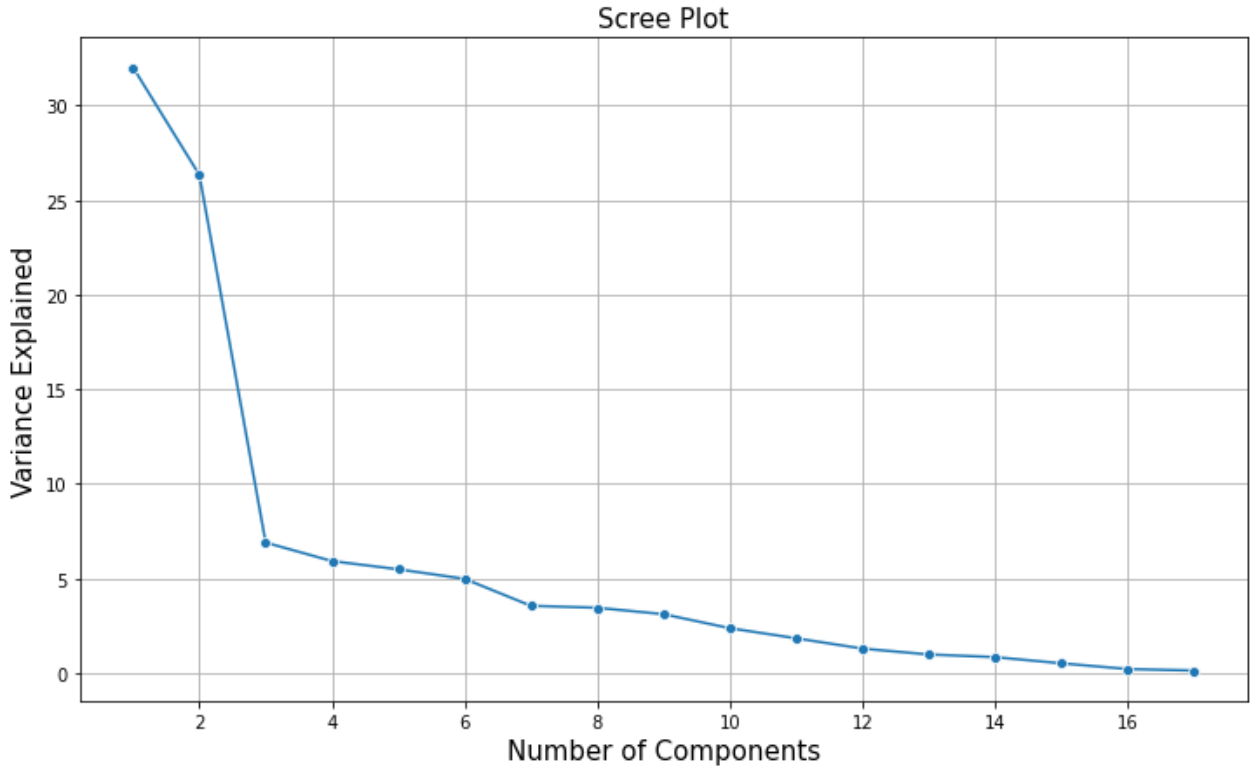
```
Linear eq of first PC:
0.25 = Apps
0.21 = Accept
0.18 = Enroll
0.35 = Top10perc
0.34 = Top25perc
0.15 = F.Undergrad
0.03 = P.Undergrad
0.29 = Outstate
0.25 = Room.Board
0.06 = Books
-0.04 = Personal
0.32 = PhD
0.32 = Terminal
-0.18 = S.F.Ratio
0.21 = perc.alumni
0.32 = Expend
0.25 = Grad.Rate
```


Back to Top

3.2.9 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?

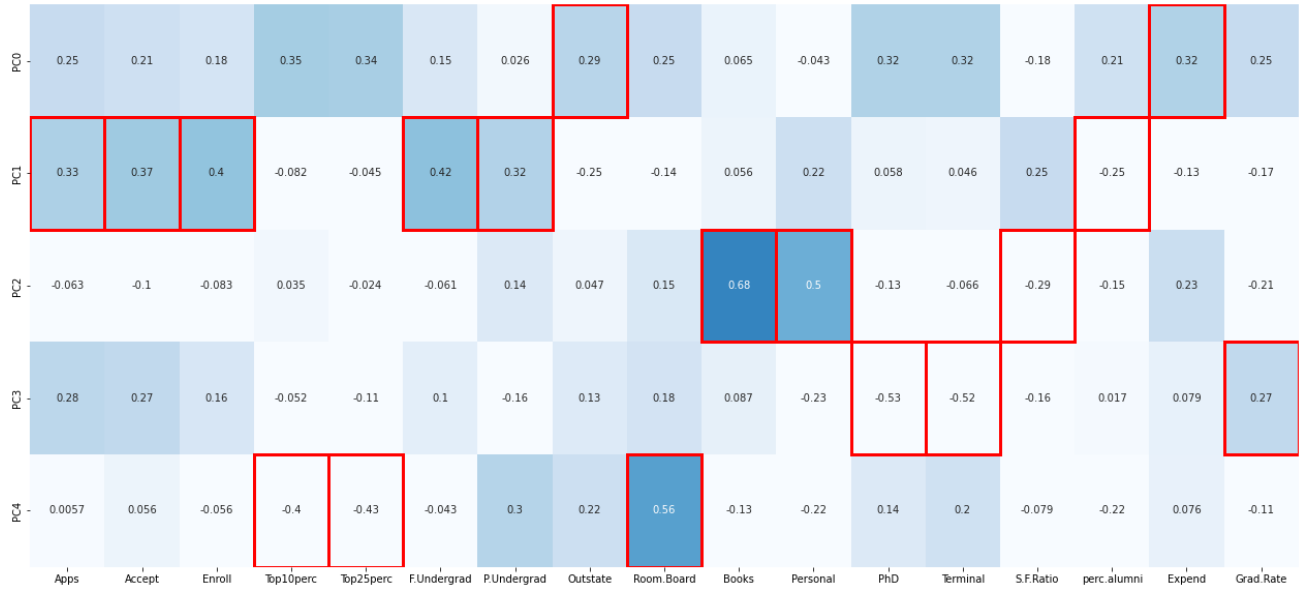
```
array([ 32.0206282 ,  58.36084263,  65.26175919,  71.18474841,
        76.67315352,  81.65785448,  85.21672597,  88.67034731,
        91.78758099,  94.16277251,  96.00419883,  97.30024023,
        98.28599436,  99.13183669,  99.64896227,  99.86471628,
       100.        ])
```

Figure 8. Scree Plot



Back to Top

Figure 9. Heat map for five principal components



3.2.9.1 Interpretation

As mentioned above, we can see an elbow forming after the 5th component, we have performed PCA for 5 components and other related functions. We can also determine from the cumulative eigen values that till the 5th component around 76% of the data is covered. This helps us in deciding the optimum level of principal components.

Eigen vector indicate the direction or rotation of the dataset due to the covariance vectors. They indicate the linear combination of the variables and the eigen vectors represent a direction which is nothing but (-2.48 e-01 multiplied by the first feature, i.e., “apps” ; 3.31 e-01 multiplied by the second feature, i.e., “accept”, and so on.

3.2.10 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]

Interpretations of the complete dataset are given below:

- High positive correlation between F. Undergrad and Apps, Accept, Enroll: This is an interesting observation. If students are full-time students, they are more likely to apply for these universities and lead to subsequent enrollment. This does not apply to part-time students.
- High positive correlation between Top 10 Percs and Top 25 Percs: Potential overlap between these two categories There is a positive correlation between these two categories that may contribute to higher degrees
- High negative correlation between Top10Perc, Top25Perc, and S.F. ratios: The higher the number of students in the upper grades category, the lower the student-to-facility ratio. The lower the S .F ratio, the more faculty members per student group. We can conclude that a better university (with

Back to Top

- senior students) has a large number of faculty members. You can also show small classrooms that can pay individual attention to each student.
- Very high negative correlation with S.F. Ratio, Personal, and P.Undergrad: S.F. Ratio indicates higher number of students means higher number of faculties for each group. Personal variable of personal expenses of students indicates that students who spend more on other expenses than studies tend to lose focus.
- The first principal component is positive correlated with “Top 10 Perc”, “Top 25 Perc”, “PhD”, “Terminal”, and “Expend”. This might indicate that these five features vary together. If one increases, the other also increases. Elite students (top 10% and 25%) get enrolled in top universities having elite faculty (PhD/terminal degree). This also increases the instructional expenditure, possibly due to higher salaries paid to the faculty as well as unique learning methods deployed.
- The second principal component is positive correlated with “Apps”, “Accept”, “Enroll”, “F.Undergrad” and “P.Undergrad”. This is an obvious correlation related to the admission procedure. The higher the applications submitted, the higher would be the subsequent acceptance and enrolment. It also tells us that the higher the enrolment, the higher is the number of full-time graduates, thus possibly hinting at who are the universities’ target audience.
- The third principal component is highly positive correlated with the “Books” and “Personal” expenses of a student. Can be labelled as books and personal expenses. The fourth principal component is highly negative correlated with “PhD” and “Terminal”. Can be labelled % of elite faculty.
- The fifth principal component is more related to the percentage of students belonging to the top 10% and 25% of higher secondary class. Can be labelled as % elite students

Business implications of using PCA:

- The original dataset has a lot of variables (17), and as observed there exists correlation between different pairs of variables.
- From an analyst’s perspective, it is essential that several variables exhibiting common characteristics are clubbed together, whereas some unnecessary variables are omitted from the dataset.
- PCA helps achieve this reduction in dimension from 17 variables to 5 principal components (which cover around 76% of data) describing these 17 variables.
- PCA also helps in determining the hidden, not so obvious relationship between the variables in the dataset.
- There are considerably higher number of full-time undergraduate students who apply to these universities/colleges.
- Having elite faculty with PhD and terminal degree results helps in achieving a better graduation rate.
- A lower student-to-faculty ratio (i.e., higher number of faculty per a given group of students), smaller classrooms, etc. enhance personal attention given to students, thereby increasing their chances of graduating.
- PCA also helps in dimension reduction/grouping of similar variables. This aids in better decision-making by concentrating on only a handful of variables.

Back to Top