Problem 1

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

Null and alternate hypothesis for 'Education' with respect to 'Salary'.
H0: Salary depends on Education.

HA: Salary does not depend on Education.

Confidence level= 0.05

Null and alternate hypothesis for 'Occupation' with respect to 'Salary'.
H0: Salary depends on Occupation.

HA: Salary does not depend on Occupation.

Confidence level= 0.05

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

P value is less than 0.05. So, the null hypothesis is rejected.

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

P value is greater than 0.05. So, we fail to reject 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.

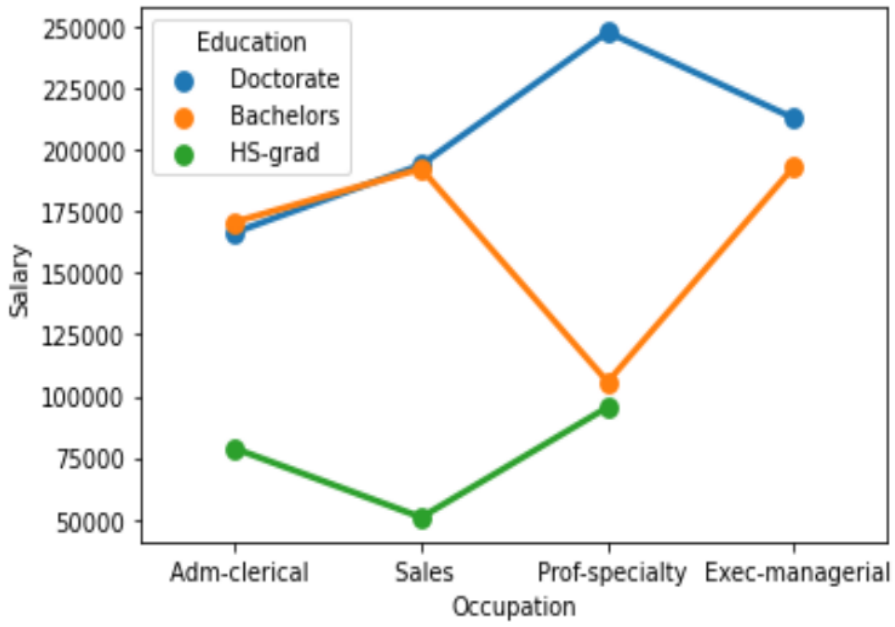
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Adm-clerical	Exec-managerial	55693.3	0.4133	-40414.1768	151800.7768	False
Adm-clerical	Prof-specialty	27528.8538	0.7478	-46276.6569	101334.3646	False
Adm-clerical	Sales	16180.1167	0.9374	-58950.5539	91310.7873	False
Exec-managerial	Prof-specialty	-28164.4462	0.8439	-120501.5231	64172.6308	False
Exec-managerial	Sales	-39513.1833	0.668	-132912.8623	53886.4956	False
Prof-specialty	Sales	-11348.7372	0.972	-81591.9315	58894.4572	False

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Bachelors	Doctorate	43274.0667	0.0146	7537.2745	79010.8589	True
Bachelors	HS-grad	-90114.1556	0.0	-132039.7353	-48188.5758	True
Doctorate	HS-grad	-133388.2222	0.0	-174819.5736	-91956.8709	True

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.



From the above interaction plot of Salary and Education, we find that the salary packages of Admn-clerical and Sales professionals with Bachelors and Doctorate degrees is almost similar whereas the salary packages of professionals with HS-grad degree is low in every Occupation as compared to Bachelors and Doctorate.

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?

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	31.257677	1.981539e-08
C(Occupation)	3.0	5.519946e+09	1.839982e+09	1.120080	3.545825e-01
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

Two-way ANOVA based on the Education and Occupation:

H0: Salary depends on both categories - Education and Occupation.

HA: Salary does not depend on at least one of the categories - Education and Occupation.

Confidence level = 0.05

Considering both education and Occupation, Education is a significant factor as P value is <0.05, Whereas Occupation is not significant variable as P value of it is > 0.05.

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

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

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

After performing ANOVA on this data set, we came to few conclusions:

- Salary is dependent on Occupation.
- The salary packages of Adm-clerical and Sales professionals with Bachelors and Doctorate degrees is almost same, Whereas the salary packages with HS-grad degree is low in every Occupation as compared to Bachelors and Doctorate.
- Education is a significant factor as P value is <0.05, whereas Occupation is not significant variable as P value of it is >0.05.
- While performing one-way ANOVA for Education with respect to the variable 'Salary', we found that the P value is less than 0.05. Hence, the null hypothesis is rejected which means that Salary does not depend on Education.
- While performing one-way ANOVA for Occupation with respect to the variable 'Salary', we found that the P value is greater than 0.05. Hence, we fail to reject the null hypothesis which means that Salary depends on Education.

Problem 2

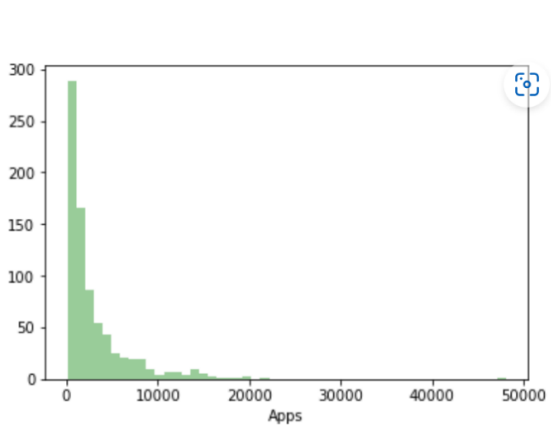
2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

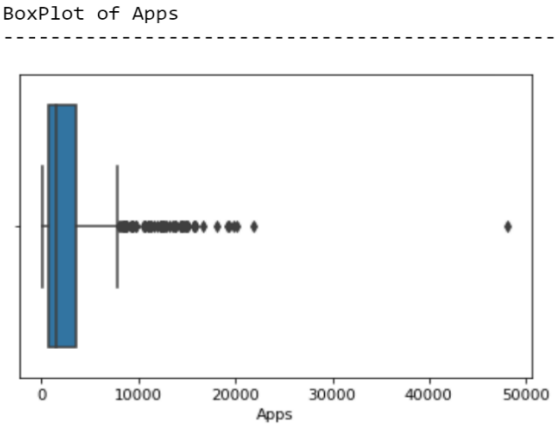
Performing univariate analysis of 17 numeric variables.

This Analysis includes:

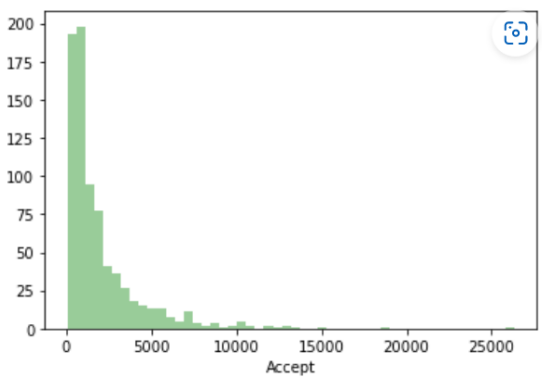
- Description of all the 17 numeric variables.
- Histogram or distplot showing the distribution of the variables.
- Boxplot showing the outlies if any.

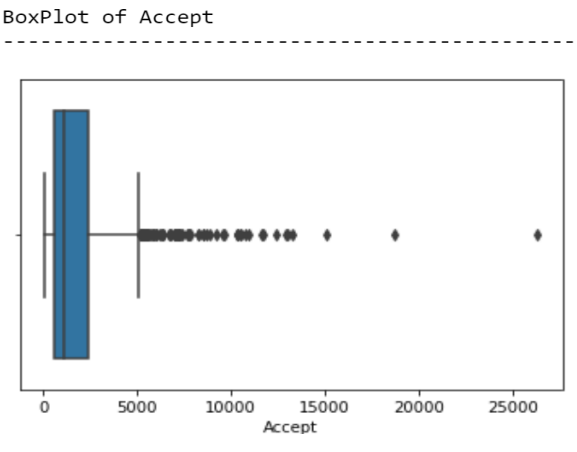
```
Description of Apps
-----------------------------------------------
count      777.000000
mean      3001.638353
std       3870.201484
min         81.000000
25%        776.000000
50%       1558.000000
75%       3624.000000
max      48094.000000
Name: Apps, dtype: float64 Distribution of Apps
```





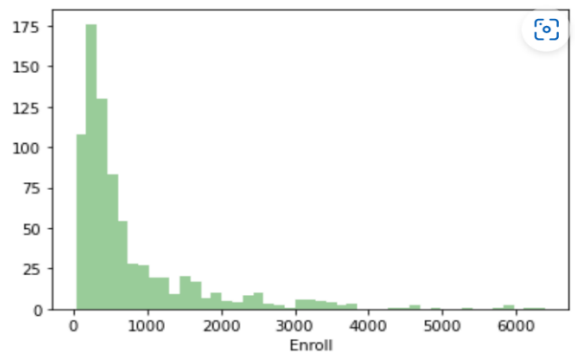

```
Description of Accept
-----------------------------------------------
count      777.000000
mean      2018.804376
std       2451.113971
min         72.000000
25%        604.000000
50%       1110.000000
75%       2424.000000
max      26330.000000
Name: Accept, dtype: float64 Distribution of Accept
```



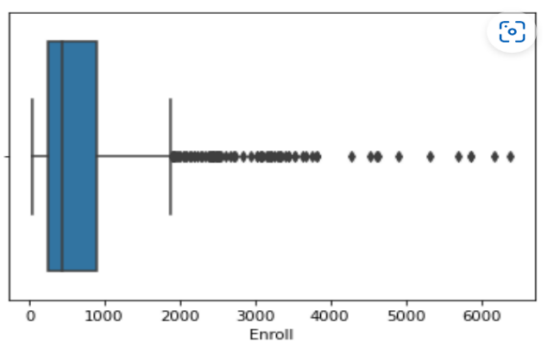


Description of Enroll

```
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 Distribution of Enroll
```

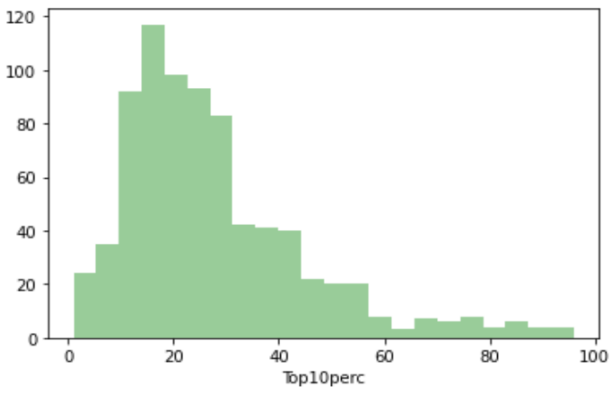


BoxPlot of Enroll

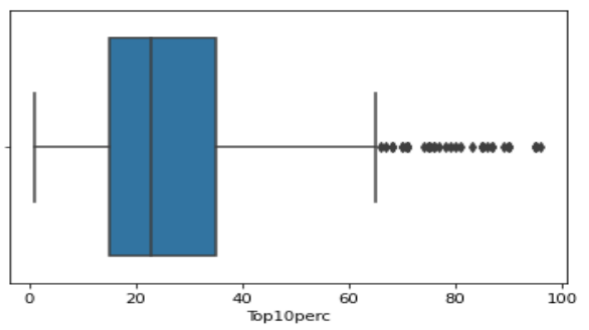


Description of Top10perc

```
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 Distribution of Top10perc
```

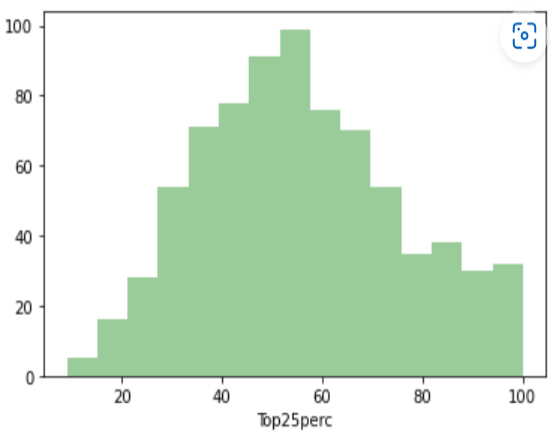


BoxPlot of Top10perc

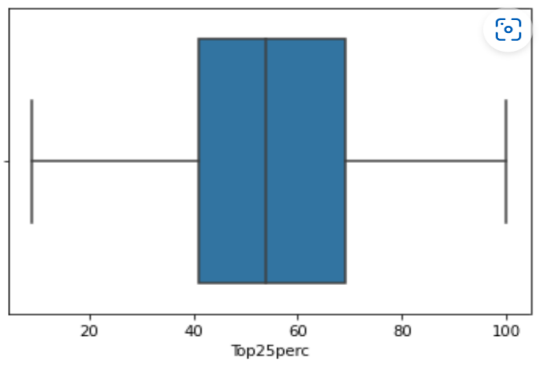


Description of Top25perc

```
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 Distribution of Top25perc
```

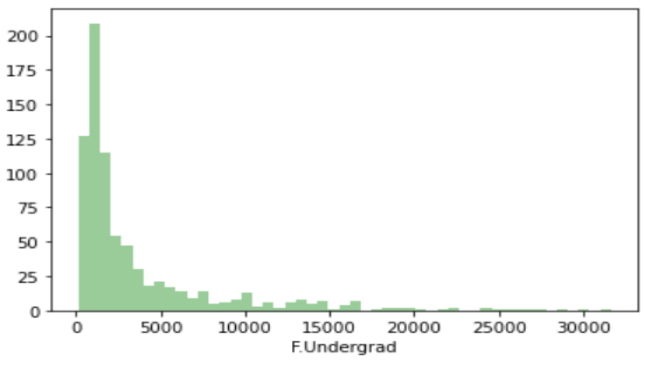


BoxPlot of Top25perc

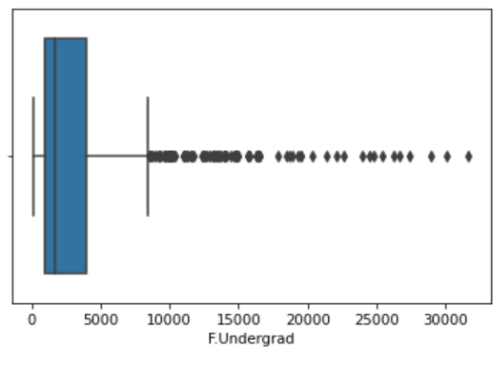


Description of F.Undergrad

```
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 Distribution of F.Undergrad
```

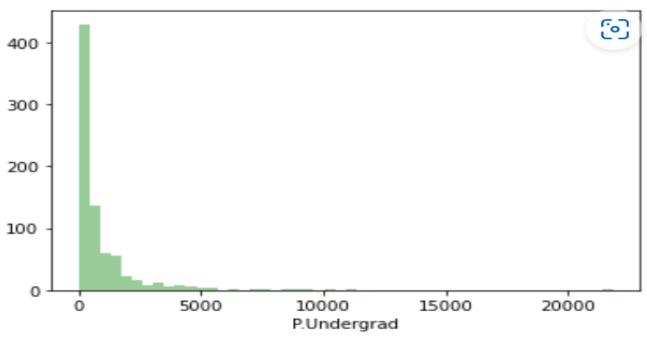


BoxPlot of F.Undergrad

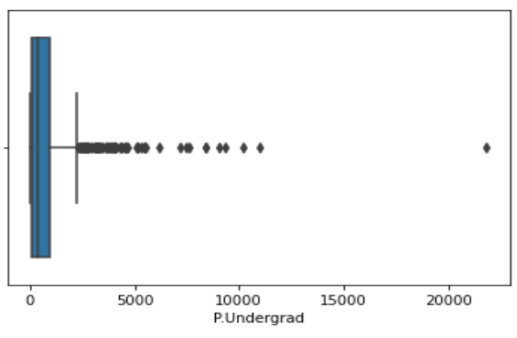


Description of P.Undergrad

```
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 Distribution of P.Undergrad
```

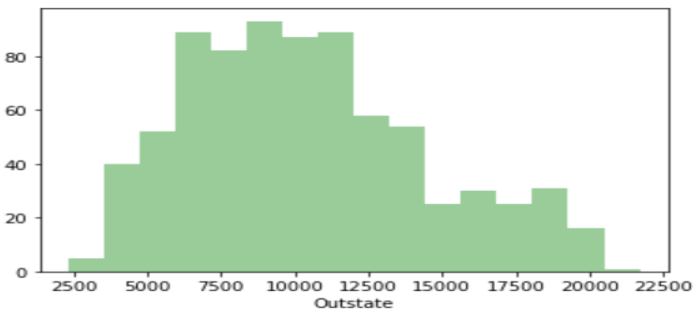


BoxPlot of P.Undergrad

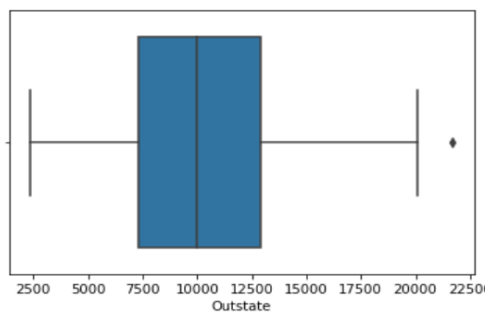


Description of Outstate

```
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 Distribution of Outstate
```



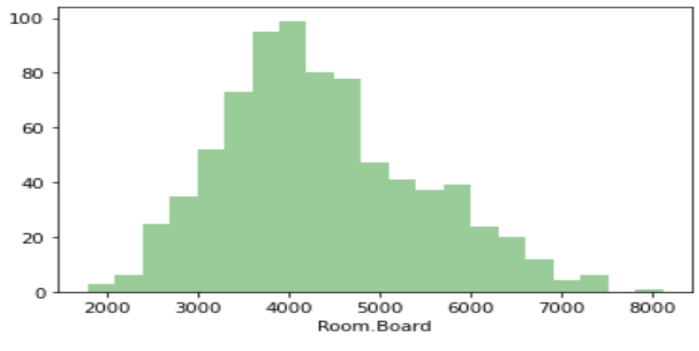
BoxPlot of Outstate



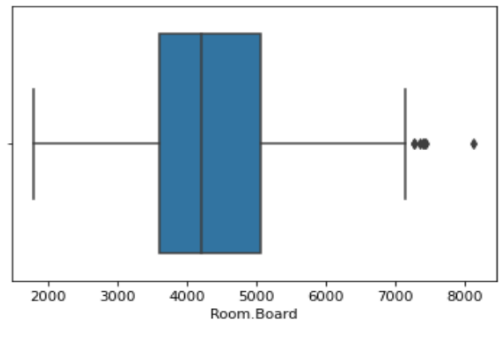
Description of Room.Board

---

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 Distribution of Room.Board



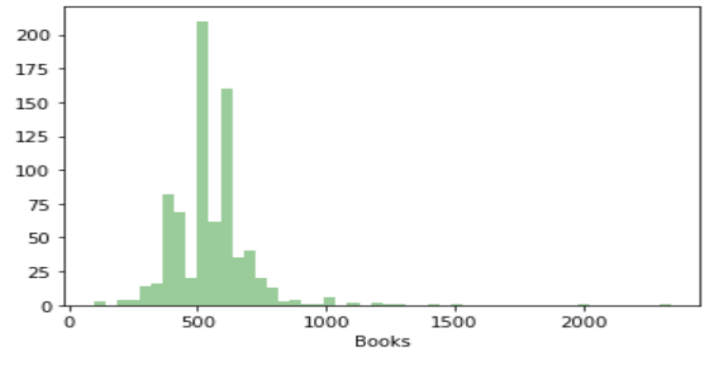
BoxPlot of Room.Board



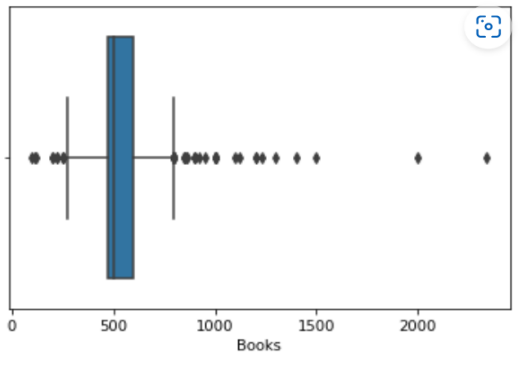
Description of Books

---

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 Distribution of Books



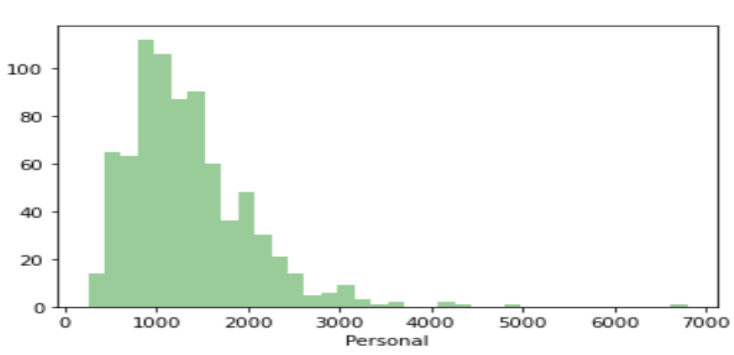
BoxPlot of Books



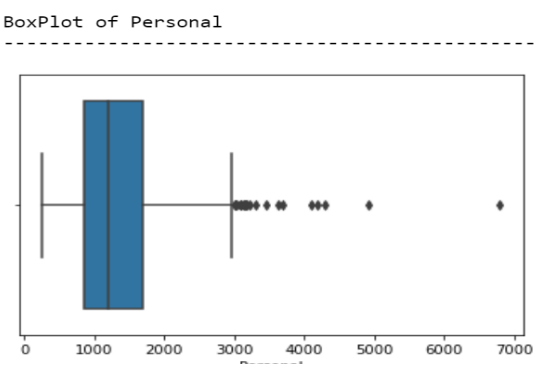
Description of Personal

---

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 Distribution of Personal



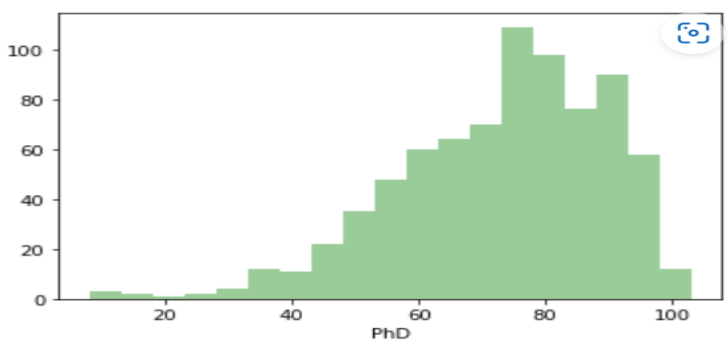
BoxPlot of Personal



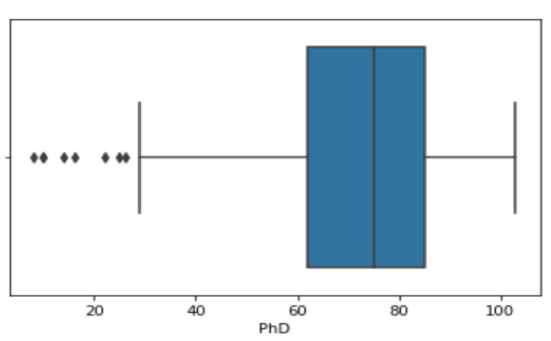
Description of PhD

---

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 Distribution of PhD



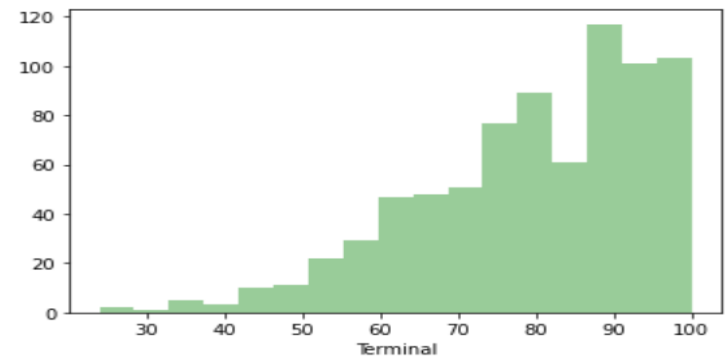
BoxPlot of PhD



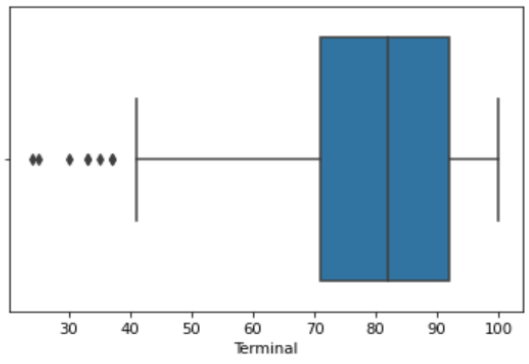
Description of Terminal

---

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 Distribution of Terminal



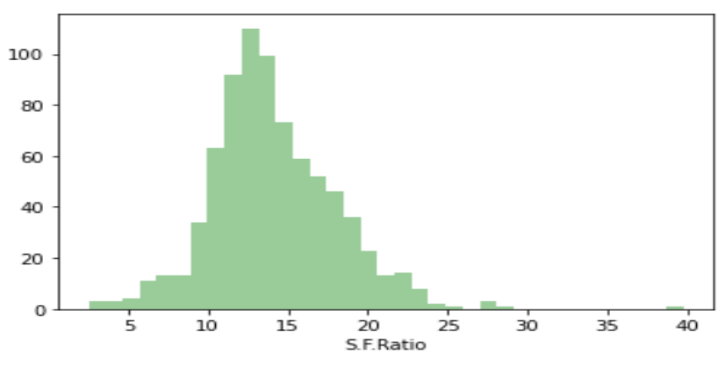
BoxPlot of Terminal



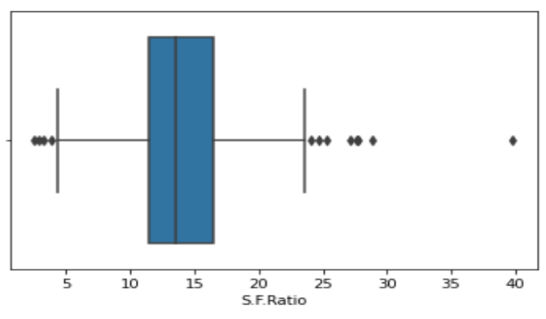
Description of S.F.Ratio

---

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 Distribution of S.F.Ratio

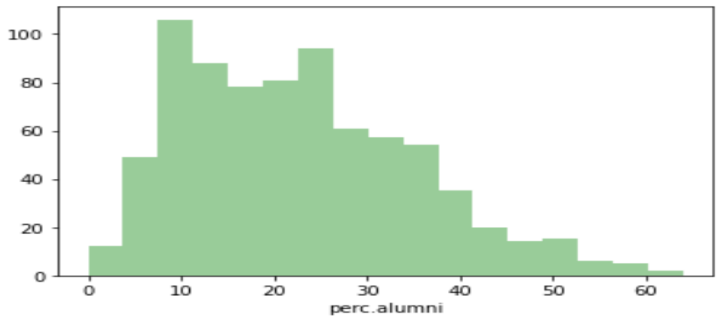


BoxPlot of S.F.Ratio

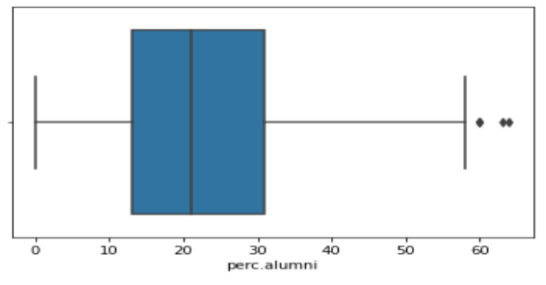


Description of perc.alumni

```
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 Distribution of perc.alumni
```

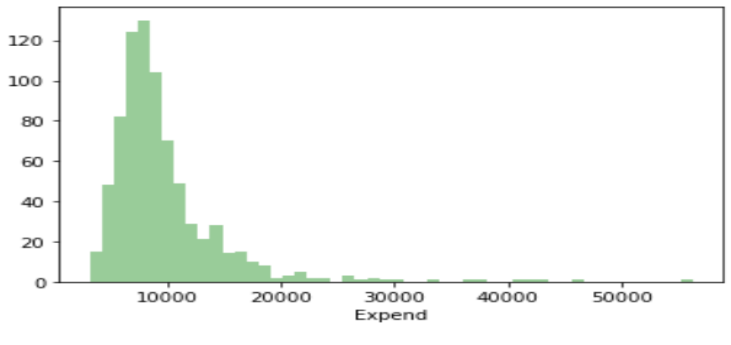


BoxPlot of perc.alumni

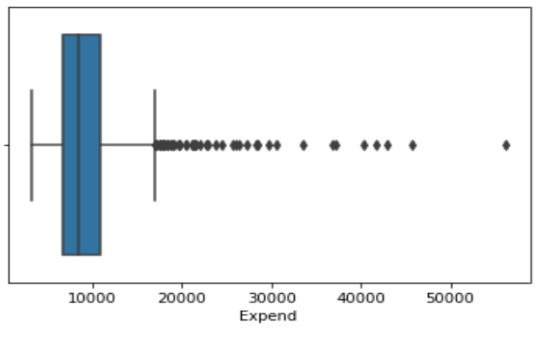


Description of Expend

```
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 Distribution of Expend
```

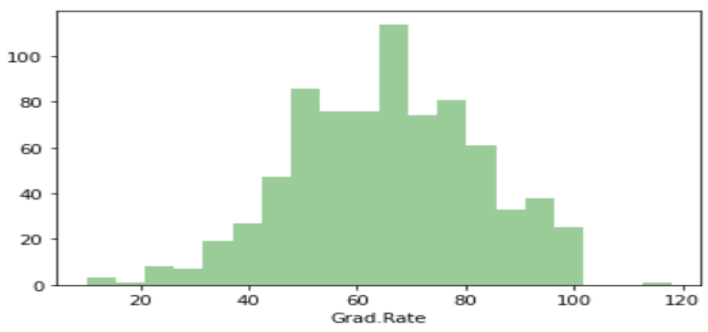


BoxPlot of Expend

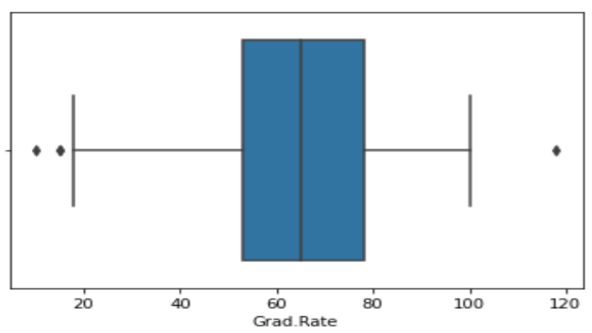


Description of Grad.Rate

```
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 Distribution of Grad.Rate
```

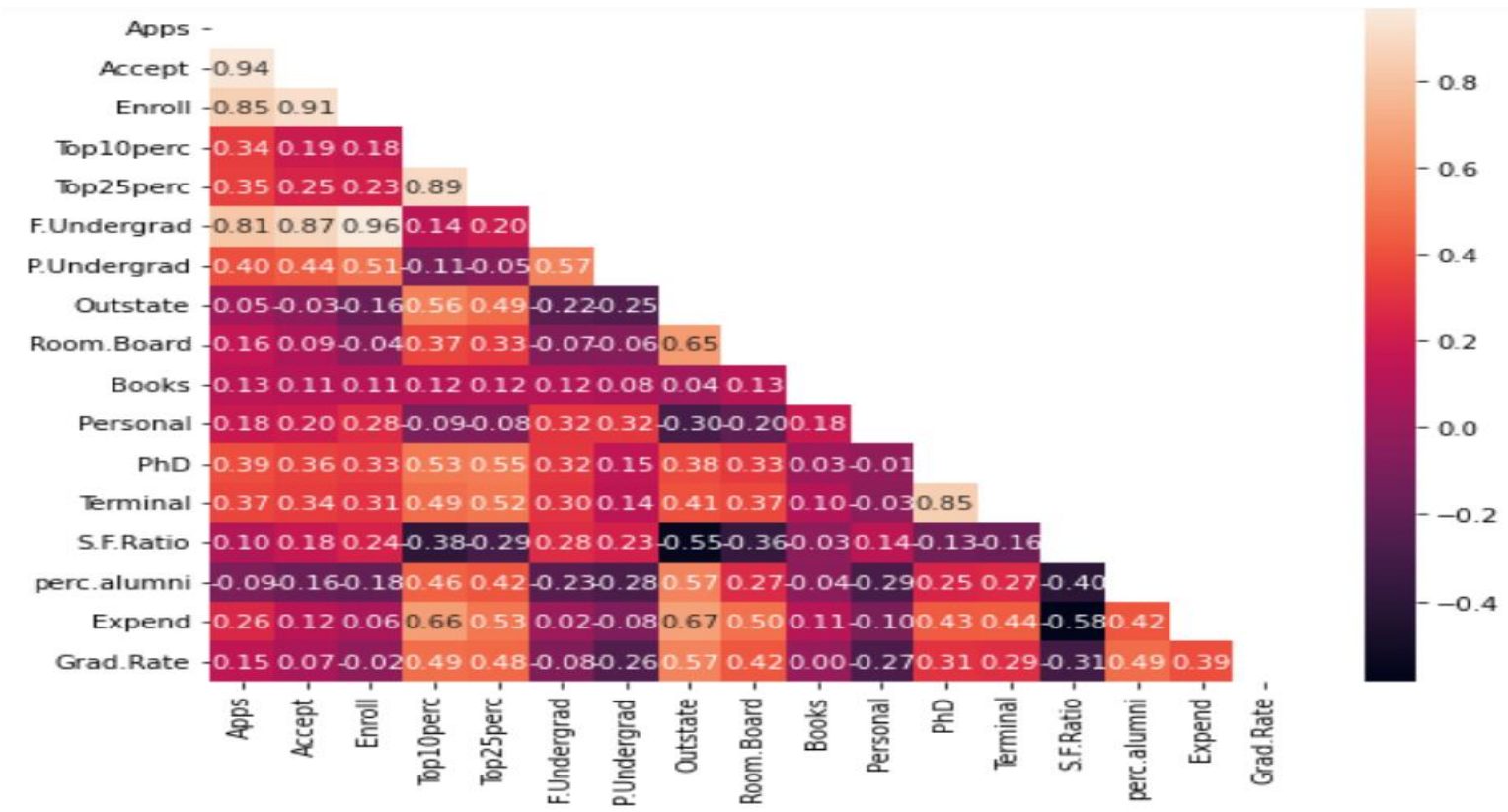


BoxPlot of Grad.Rate



The above output shows the total of 17*3 = 51 distinct charts/columns.

Next, we are to perform the multivariate analysis by using the function of correlation. Below is the heatmap for multivariate analysis.



Observation:

- The heatmap shows that there are considerable number of variables which are highly correlated.
- “Apps” has high correlation with “Accept”, and” Enroll”.
- Average book cost is around 550.
- The minimum S.F. ratio is around 2.5.

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

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369491
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758	0.337583
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469	0.308274
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828	0.491135
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862	0.524749
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.318337	0.300019
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.081200	0.319882	0.149114	0.141904
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.038855	-0.299087	0.382982	0.407983
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.127963	-0.199428	0.329202	0.374540
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.000000	0.179295	0.026906	0.099955
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.000000	-0.010936	-0.030613
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1.000000	0.849587
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.099955	-0.030613	0.849587	1.000000
S.F.Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628	-0.031929	0.136345	-0.130530	-0.160104
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009	0.267130
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739	0.112409	-0.097892	0.432762	0.438799
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.571290	0.424942	0.001061	-0.269344	0.305038	0.289527

	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	1.01377
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

Yes, Scaling is necessary for PCA in this case as scaling normalises the data because when sample variances of the original variables show differences by large order of magnitude, variables need to be normalized. Scaling ensures that attribute means are all 0 and variances 1.

Scaling is necessary due to various other reasons:

- PCA effectiveness depends upon the scales of the attributes. If attributes have different scales, PCA will pick variable with highest variance rather than picking up attributes based on correlation.
- Changing scales of the variables can change the PCA.
- Interpreting PCA can become challenging due to presence of discrete data.
- Presence of skew in data with long thick tail can impact the effectiveness of the PCA.
- PCA assumes linear relationship between attributes. It is ineffective when relationships are non-linear.

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

```
Covariance Matrix
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.81554018
   0.3987775   0.05022367  0.16515151  0.13272942  0.17896117  0.39120081
   0.36996762  0.09575627 -0.09034216  0.2599265   0.14694372]
 [ 0.94466636  1.00128866  0.91281145  0.19269493  0.24779465  0.87534985
   0.44183938 -0.02578774  0.09101577  0.11367165  0.20124767  0.35621633
   0.3380184   0.17645611 -0.16019604  0.12487773  0.06739929]
 [ 0.84791332  0.91281145  1.00128866  0.18152715  0.2270373   0.96588274
   0.51372977 -0.1556777  -0.04028353  0.11285614  0.28129148  0.33189629
   0.30867133  0.23757707 -0.18102711  0.06425192 -0.02236983]
 [ 0.33927032  0.19269493  0.18152715  1.00128866  0.89314445  0.1414708
  -0.10549205  0.5630552   0.37195909  0.1190116  -0.09343665  0.53251337
   0.49176793 -0.38537048  0.45607223  0.6617651   0.49562711]
 [ 0.35209304  0.24779465  0.2270373   0.89314445  1.00128866  0.19970167
  -0.05364569  0.49002449  0.33191707  0.115676   -0.08091441  0.54656564
   0.52542506 -0.29500852  0.41840277  0.52812713  0.47789622]
 [ 0.81554018  0.87534985  0.96588274  0.1414708   0.19970167  1.00128866
   0.57124738 -0.21602002 -0.06897917  0.11569867  0.31760831  0.3187472
   0.30040557  0.28006379 -0.22975792  0.01867565 -0.07887464]
 [ 0.3987775   0.44183938  0.51372977 -0.10549205 -0.05364569  0.57124738
   1.00128866 -0.25383901 -0.06140453  0.08130416  0.32029384  0.14930637
   0.14208644  0.23283016 -0.28115421 -0.08367612 -0.25733218]
 [ 0.05022367 -0.02578774 -0.1556777   0.5630552   0.49002449 -0.21602002
  -0.25383901  1.00128866  0.65509951  0.03890494 -0.29947232  0.38347594
   0.40850895 -0.55553625  0.56699214  0.6736456   0.57202613]
 [ 0.16515151  0.09101577 -0.04028353  0.37195909  0.33191707 -0.06897917
  -0.06140453  0.65509951  1.00128866  0.12812787 -0.19968518  0.32962651
   0.3750222  -0.36309504  0.27271444  0.50238599  0.42548915]
```

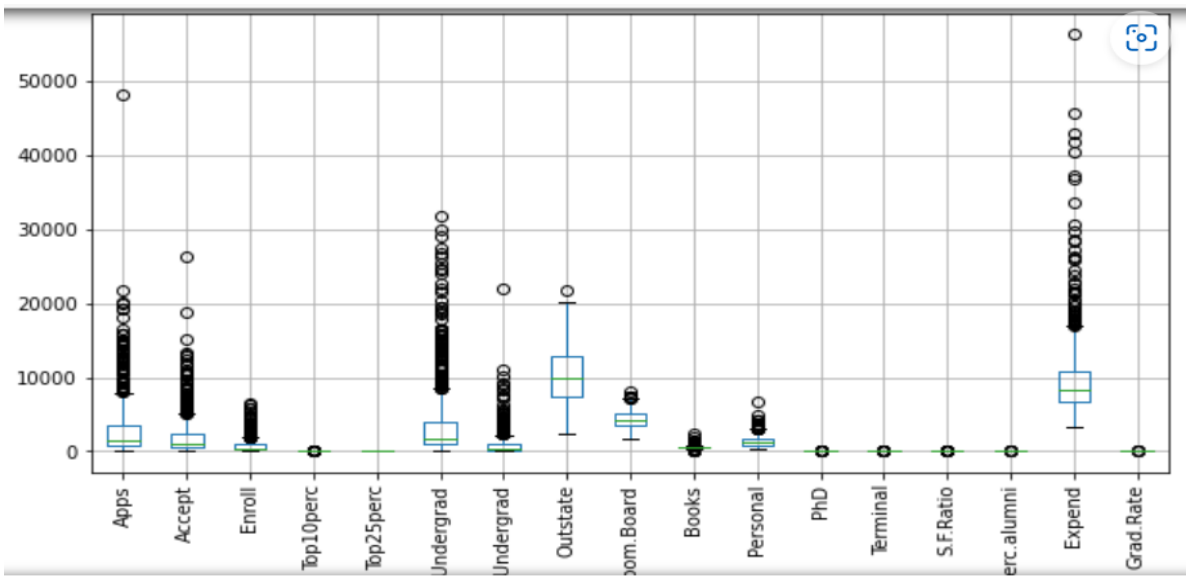
Correlation is a scaled version of covariance.

Covariance shows you how the two variables differ. Covariance refers to a systematic relationship between two random variables in which a change in the other reflects a change in one variable.

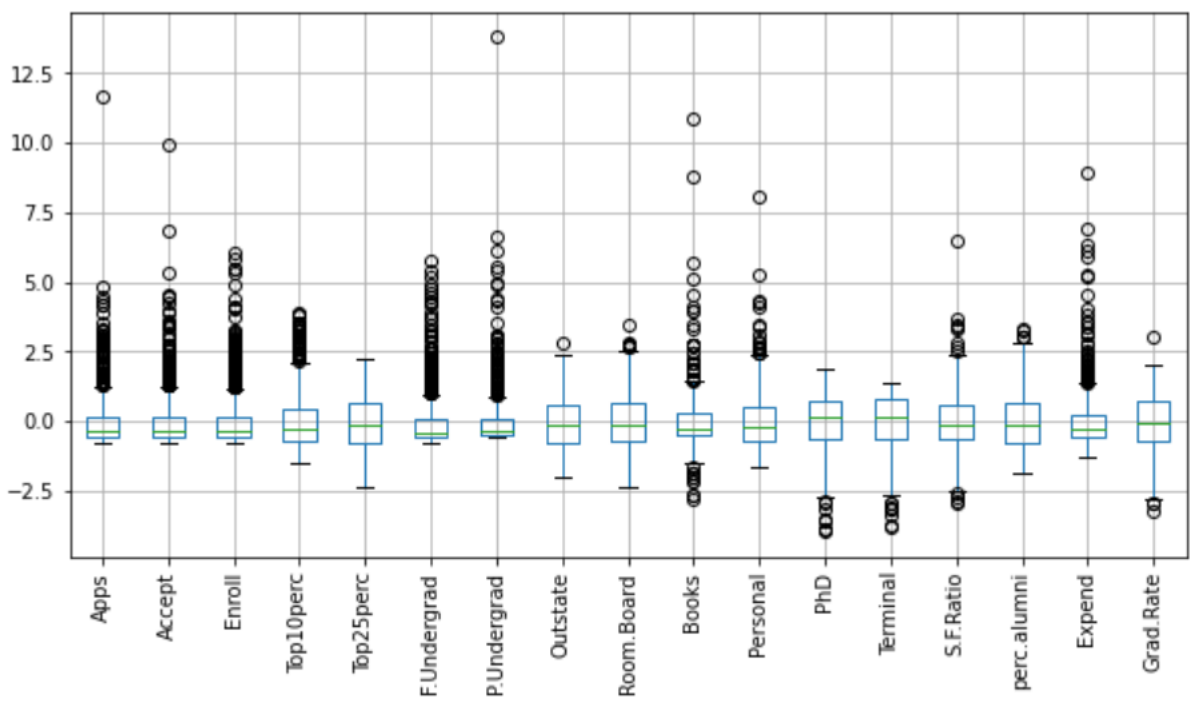
Correlation shows you how the two variables are related. Correlation measures both the strength and direction of the linear relationship between two variables.

These two parameters always have the same sign (positive, negative, or 0). When the sign is positive, the variables are said to be positively correlated; when the sign is negative, the variables are said to be negatively correlated; and when the sign is 0, the variables are said to be uncorrelated.

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



The above Boxplot shows outliers before scaling.



The above Boxplot shows outliers after scaling.

Insights:

- By scaling, all variables have the same standard deviation, thus all variables have the same weight and thus resulting in PCA calculating relevant axis.

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

Eigenvalues and eigenvectors can be extracted through using covariance matrix.

The values below show the 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])
```

The values below show the 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,
```

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

<u>Principal Component Analysis steps:</u>

- Step 1: Generate the covariance matrix.
- Step 2: Get eigenvalues and eigenvectors.
- Step 3: View scree plot to identify the number of components to be built.
- Step 4: Perform PCA on the scaled data set by importing PCA from sklearn decomposition.

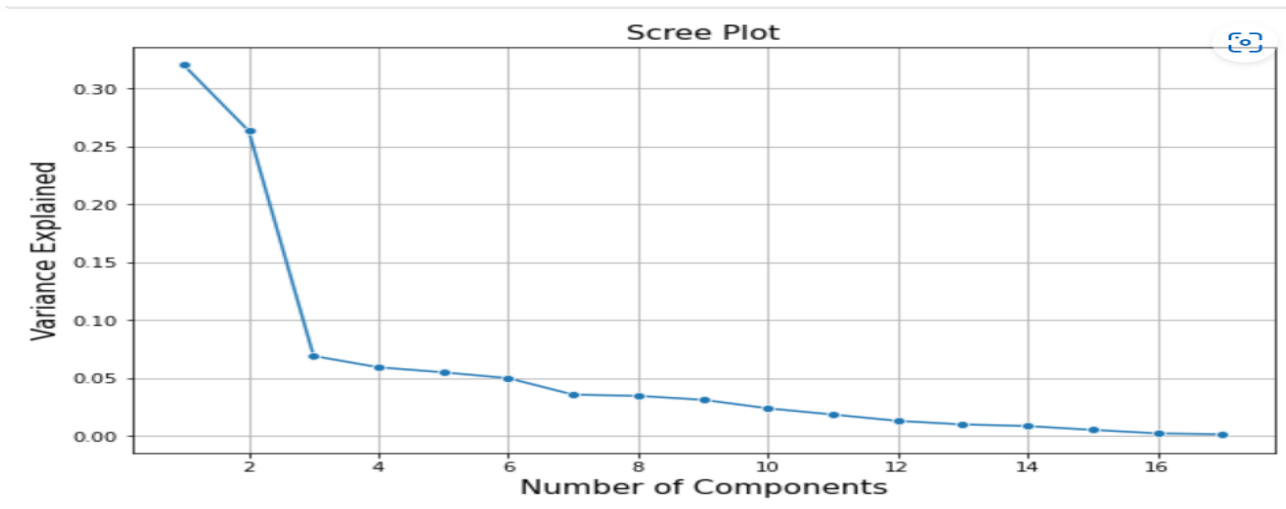
```
Covariance Matrix
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.81554018
    0.3987775   0.05022367  0.16515151  0.13272942  0.17896117  0.39120081
    0.36996762  0.09575627 -0.09034216  0.2599265   0.14694372]
 [ 0.94466636  1.00128866  0.91281145  0.19269493  0.24779465  0.87534985
    0.44183938 -0.02578774  0.09101577  0.11367165  0.20124767  0.35621633
    0.3380184   0.17645611 -0.16019604  0.12487773  0.06739929]
 [ 0.84791332  0.91281145  1.00128866  0.18152715  0.2270373   0.96588274
    0.51372977 -0.1556777  -0.04028353  0.11285614  0.28129148  0.33189629
    0.30867133  0.23757707 -0.18102711  0.06425192 -0.02236983]
 [ 0.33927032  0.19269493  0.18152715  1.00128866  0.89314445  0.1414708
   -0.10549205  0.5630552   0.37195909  0.1190116  -0.09343665  0.53251337
    0.49176793 -0.38537048  0.45607223  0.6617651   0.49562711]
 [ 0.35209304  0.24779465  0.2270373   0.89314445  1.00128866  0.19970167
   -0.05364569  0.49002449  0.33191707  0.115676   -0.08091441  0.54656564
    0.52542506 -0.29500852  0.41840277  0.52812713  0.47789622]
 [ 0.81554018  0.87534985  0.96588274  0.1414708   0.19970167  1.00128866
    0.57124738 -0.21602002 -0.06897917  0.11569867  0.31760831  0.3187472
    0.30040557  0.28006379 -0.22975792  0.01867565 -0.07887464]
 [ 0.3987775   0.44183938  0.51372977 -0.10549205 -0.05364569  0.57124738
    1.00128866 -0.25383901 -0.06140453  0.08130416  0.32029384  0.14930637
    0.14208644  0.23283016 -0.28115421 -0.08367612 -0.25733218]
 [ 0.05022367 -0.02578774 -0.1556777   0.5630552   0.49002449 -0.21602002
   -0.25383901  1.00128866  0.65509951  0.03890494 -0.29947232  0.38347594
    0.40850895 -0.55553625  0.56699214  0.6736456   0.57202613]
 [ 0.16515151  0.09101577 -0.04028353  0.37195909  0.33191707 -0.06897917
   -0.06140453  0.65509951  1.00128866  0.12812787 -0.19968518  0.32962651
    0.3750222  -0.36309504  0.27271444  0.50238599  0.42548915]
```

The below snapshots represent the extracted eigenvalues and eigenvectors.

```
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])
```

```
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,
```

The scree Plot below helps to identify the number of components to be built.



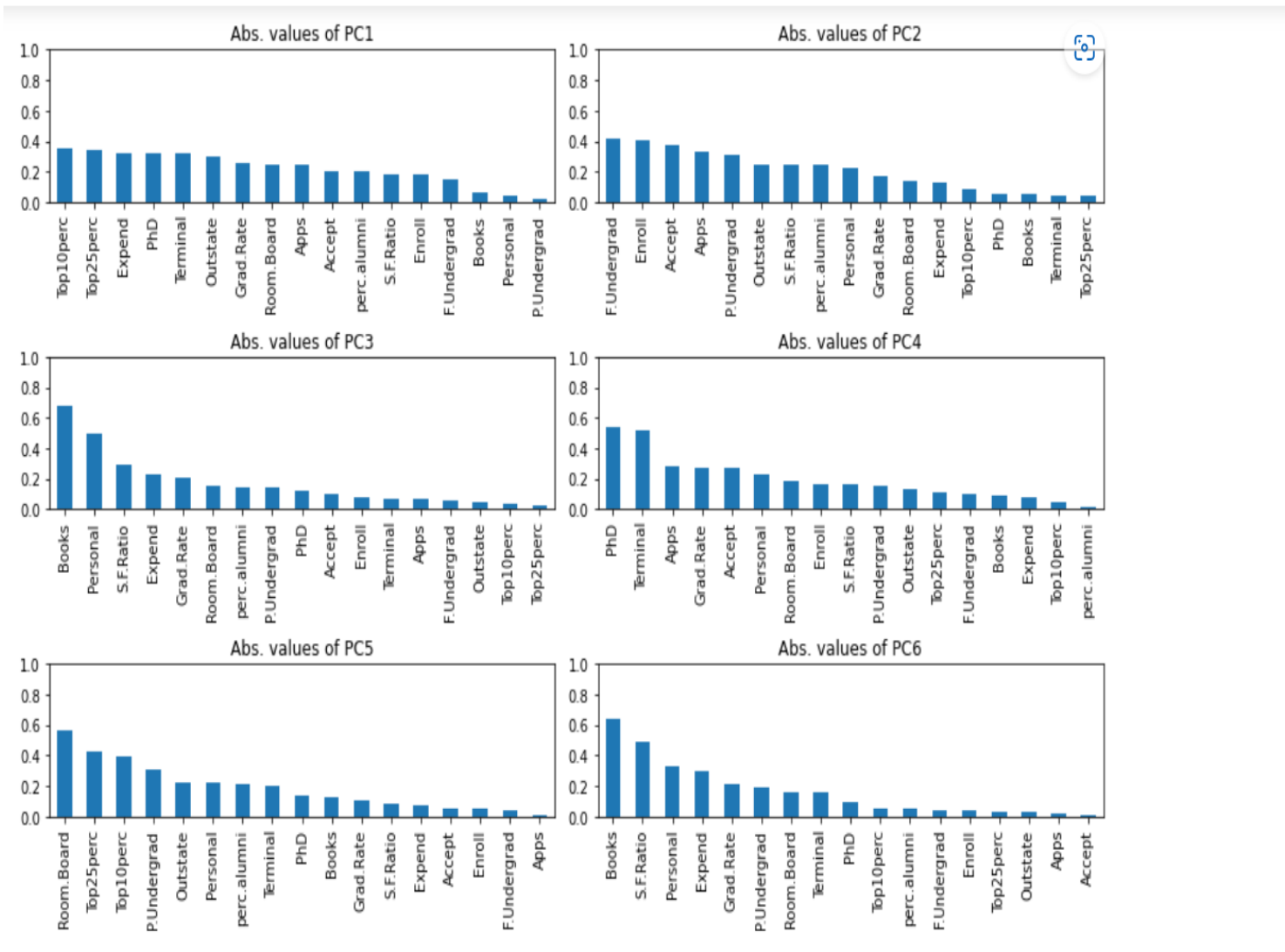
We then check the cumulative explained variance ratio to find a cut off for selecting the number of PCs.

```
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.        ])
```

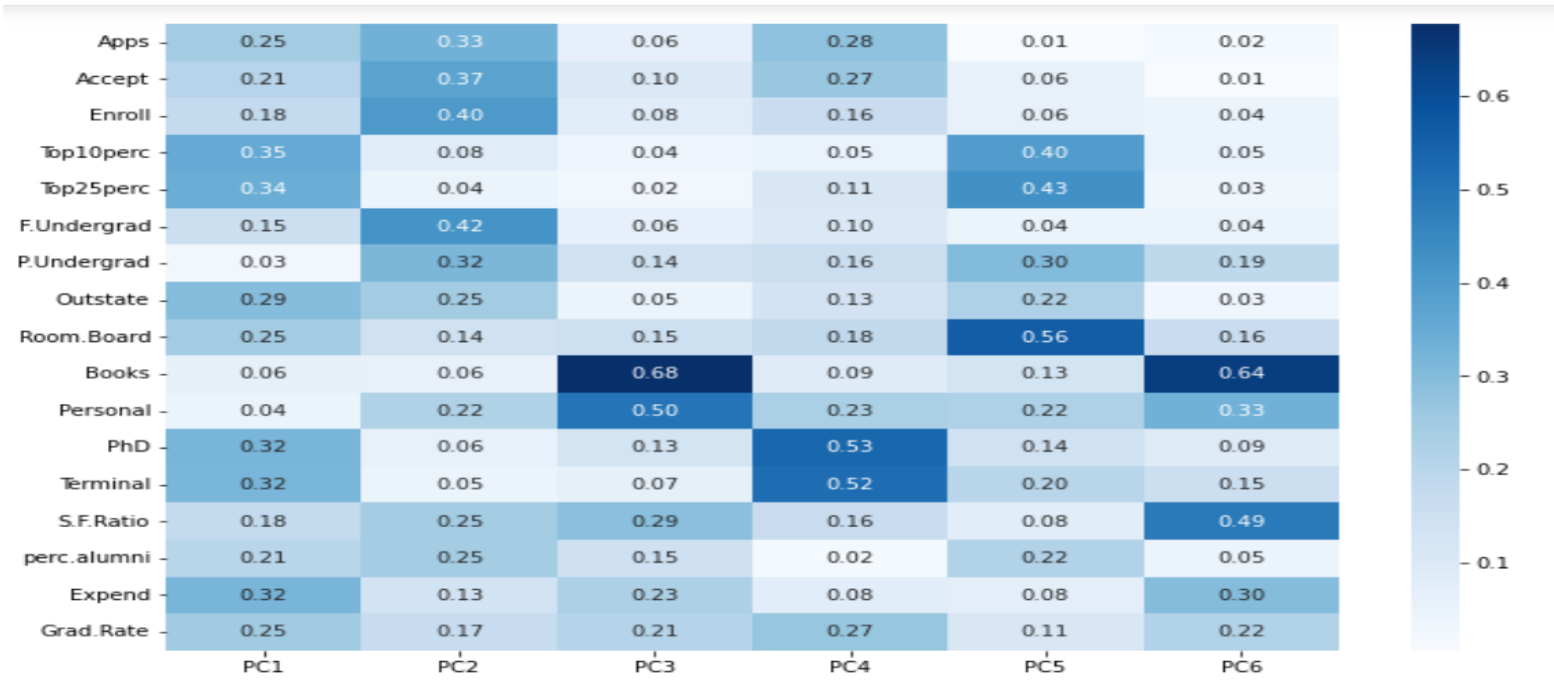
This is the new dataframe containing the loadings or coefficients of all PCs.

	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

This output given below gives us the abstract values of all the variables which checks as to how the original features matter to each PC.



The Heatmap helps us to Compare how the original features influence various PCs.



In order to calculate PC scores, we need loadings, below:

	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

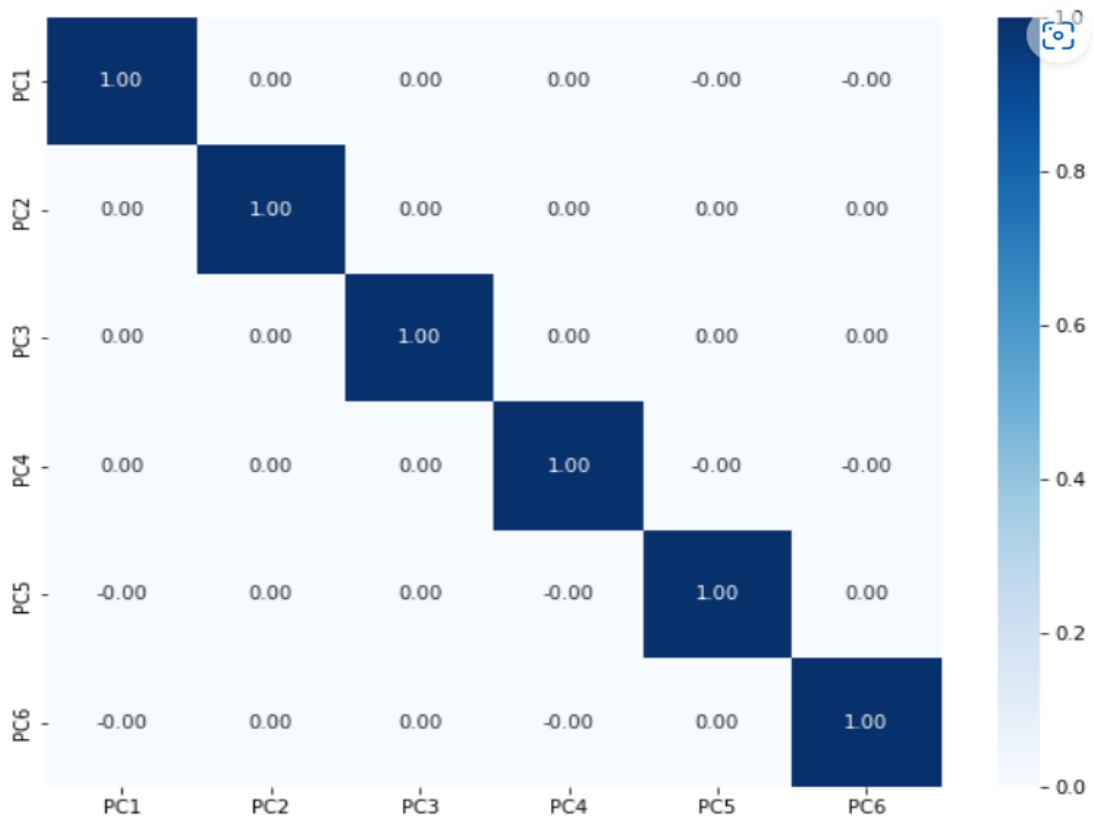
We also need the original scaled features which are given below.

```
Apps           -0.346882
Accept         -0.321205
Enroll         -0.063509
Top10perc      -0.258583
Top25perc      -0.191827
F.Undergrad    -0.168116
P.Undergrad    -0.209207
Outstate       -0.746356
Room.Board     -0.964905
Books          -0.602312
Personal        1.270045
PhD            -0.163028
Terminal       -0.115729
S.F.Ratio       1.013776
perc.alumni    -0.867574
Expend         -0.501910
Grad.Rate      -0.318252
Name: 0, dtype: float64
```

This new dataframe is formed out of fit_transformed scaled data.

	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

The below Heatmap is to check for presence of correlations among the PCs.



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]

```
The Linear eq of 1st component:
0.249 * Apps + 0.208 * Accept + 0.176 * Enroll + 0.354 * Top10perc + 0.344 * Top25perc + 0.155 * F.Undergrad + 0.026 * P.Underg
rad + 0.295 * Outstate + 0.249 * Room.Board + 0.065 * Books + -0.043 * Personal + 0.318 * PhD + 0.317 * Terminal + -0.177 * S.
F.Ratio + 0.205 * perc.alumni + 0.319 * Expend + 0.252 * Grad.Rate +
```

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?

Below are the cumulative values of the eigenvectors which shows us that there are around 8 principal components which explained more that 90% of the variance. Hence, the optimum number of principal components is 8.

```
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.        ])
```

```
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,
```

2.9 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]¶

- This business dataset about education contains the names of various colleges and university.
- Univariate analysis and Multivariate analysis is done which gives us the understanding about all the variables.
- By using Univariate analysis and Multivariate analysis, we can understand the distribution of the dataset, skew, and patterns in the dataset.
- From multivariate analysis we can get the correlation of variables. The multivariate analysis shows that there are multiple variables which are highly correlated with each other.
- The principal component analysis is used to reduce the multicollinearity between the variables.
- The scaling of the variables also helps the dataset to standardize the variable in one scale.