Advanced Statistics Project

Submitted by

Priyadharshini K

Problem 1A:

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

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

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

The null and the alternate hypothesis for conducting one-way ANOVA for Education

Null Hypothesis $H0$: The means of 'Salary' with respect to each category of Education is equal.

Alternate Hypothesis $H1$: At least one of the means of 'Salary' varies with respect to Education qualification

The null and the alternate hypothesis for conducting one-way ANOVA for Occupation

Null Hypothesis $H0$: The means of 'Salary' with respect to each category of Occupation is equal.

Alternate Hypothesis $H1$: At least one of the means of 'Salary' varies with respect to Occupation

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

Since the p value is less than the significance level α (0.05), we can say that we reject the Null Hypothesis ($H0$) and confirm that At least one of the means of 'Salary' varies with respect to Education qualification

.

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

Since the p value is more than the significance level, we fail to reject the null hypothesis and confirm that the means of 'Salary' with respect to each category of Occupation is equal.

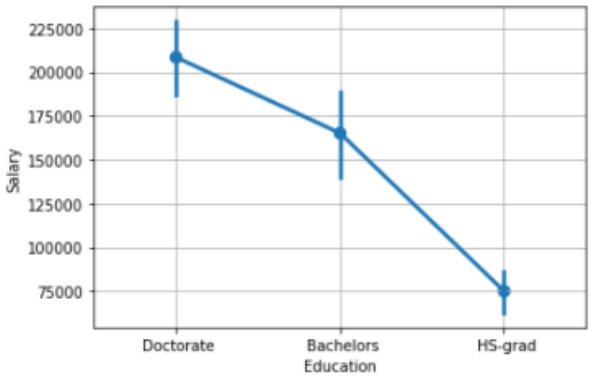
1.4 If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result.

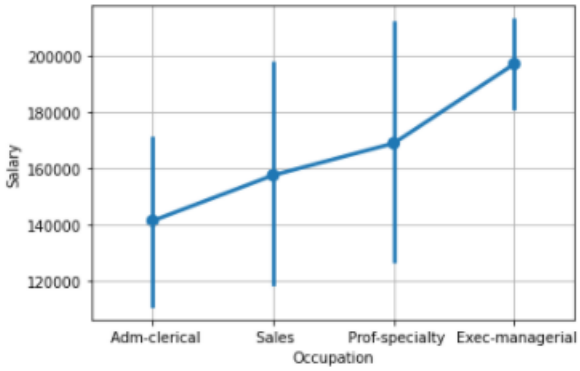
Null hypothesis is rejection in 1.2 and failed to reject in 1.3.
The class of means with respect to Education is significantly different.

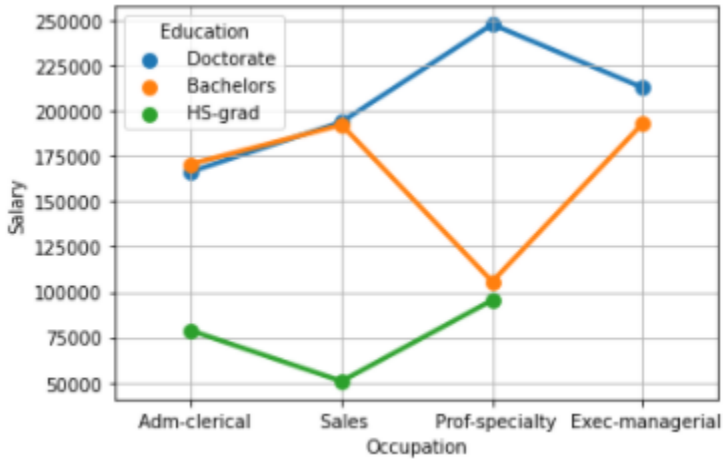
Problem 1B:

1. What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot. [hint: use the 'point plot' function from the 'seaborn' function]
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?
3. Explain the business implications of performing ANOVA for this particular case study.

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







From the above interaction plot,
we confirm that the salary of an individual increases with increased educational qualification;
we confirm that the salary of an individual is higher for Exec-managerial and reduces as we go to lower grade prof-specialty, Sales and Adm-clerical and
there seems to be statistical interaction between both the variables Occupation and Education.

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?

H_0: The means of 'Salary' variable with respect to each category of Education and Occupation is equal.

H_1: At least one of the means of 'Salary' variable with respect to each category of Education and Occupation is unequal.

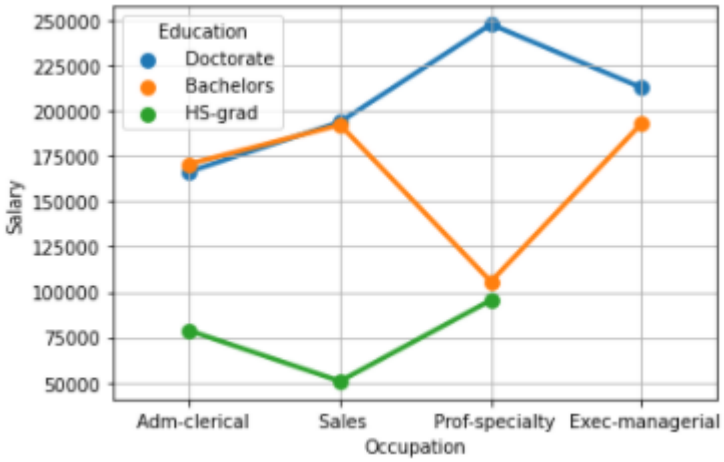
	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

As p value of interaction between Occupation and Education is 2.232500e-05 < 0.05 , there seems to be statistical interaction between both the variables.

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

By performing ANOVA on this case study, we conclude the following.

- With HS-Grad, people in Sales get minimum Salary and people in Prof-Specialty get higher salary
- Both Bachelors and Doctorate degree holders get almost same salary in Adm-Clerical and Sales roles
- Bachelor's degree holders get their lowest salary when employed in Prof-Specialty roles
- Doctorate degree holders get their highest salary in Prof-Specialty roles.
- Bachelor's degree holders get the maximum salary when employed in Exec-managerial and Sales roles, but Doctorates receive a slightly higher salary in Exec-managerial position than the Bachelor's degree holders



Problem 2:

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.

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

Observations from Univariate Analysis

Percentage of new students from top 25% of Higher Secondary class has no outliers

Rutgers at New Brunswick has received most number of applications and it is the college which has accepted most number of applications

Texas A&M Univ. at College Station has enrolled most number of students and it has the highest number of Number of full-time undergraduate students

University of Minnesota Twin Cities has the highest number of Number of part-time undergraduate students

Maximum Percentage of alumni who donate is from Williams College

University of Charleston has the best Student to Faculty Ratio

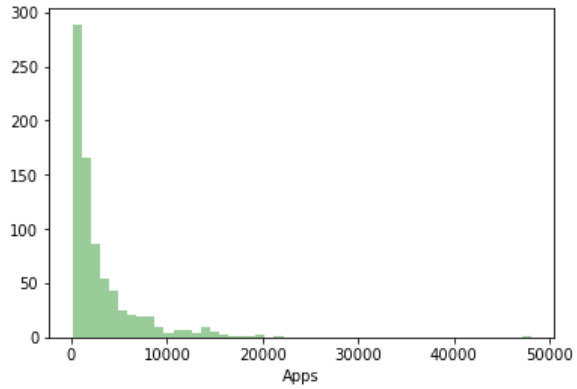
Observations from Bivariate Analysis

There are considerable number of features that are highly correlated.

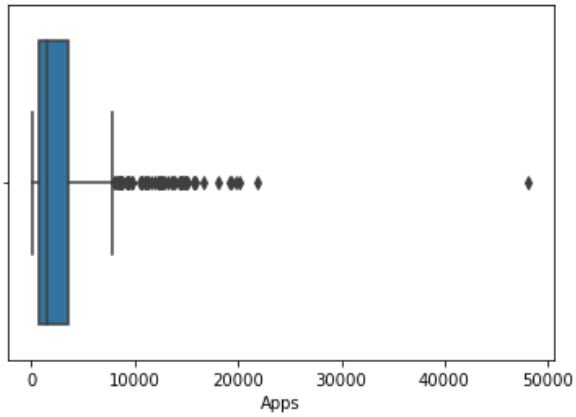
'Number of applications received' shows high correlation with 'Number of applications accepted','Number of new students enrolled'

'Number of full-time undergraduate students' shows high correlation with 'Number of new students enrolled'

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

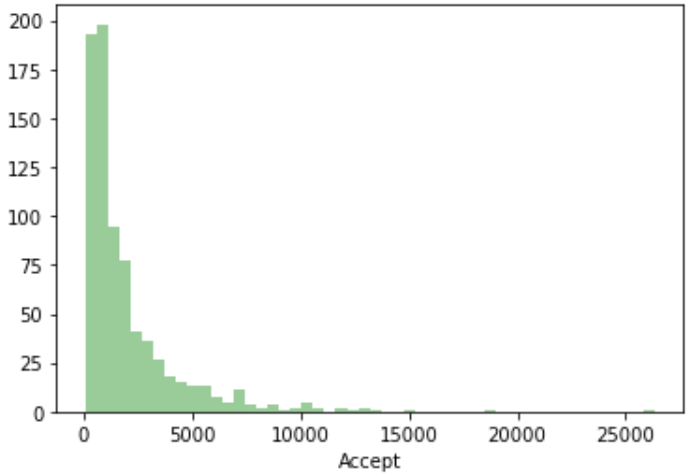


BoxPlot of Apps

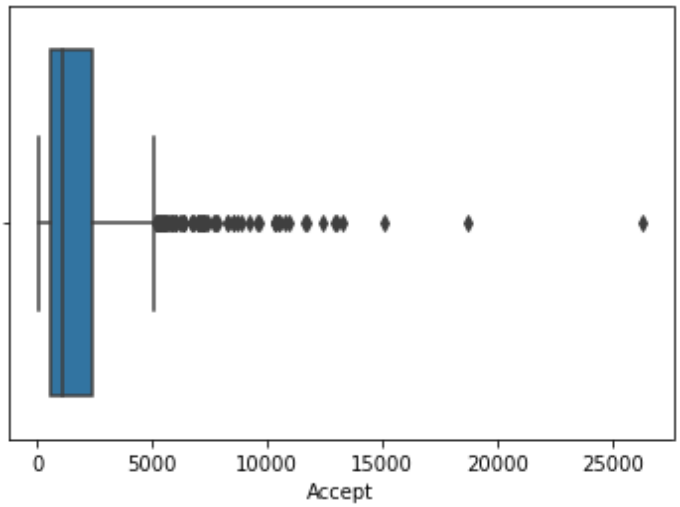


Description of Accept

```
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
```

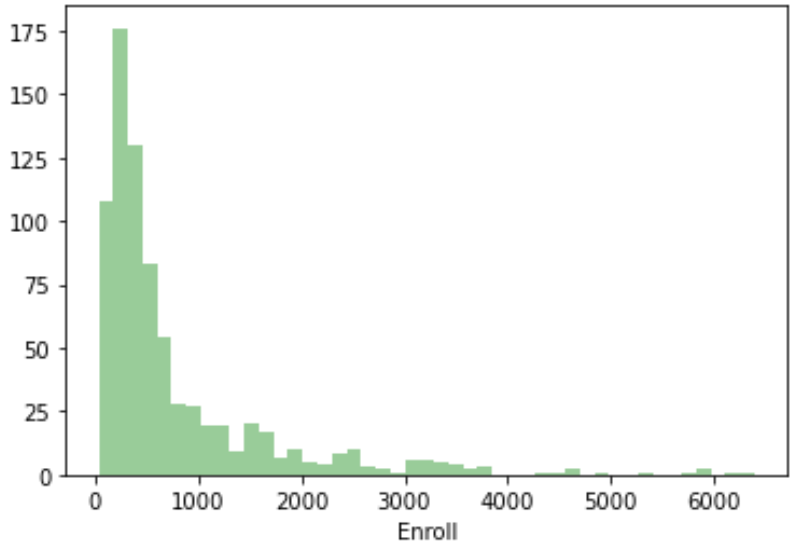


Box Plot of Accept

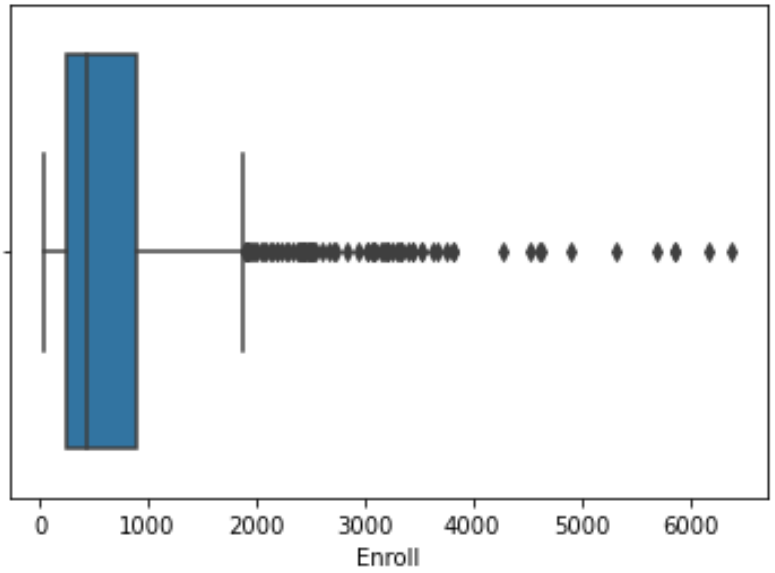


Description of Enroll

```
count     777.000000
mean      779.972973
std       929.176190
min        35.000000
25%       242.000000
50%       434.000000
75%       902.000000
max      6392.000000
```

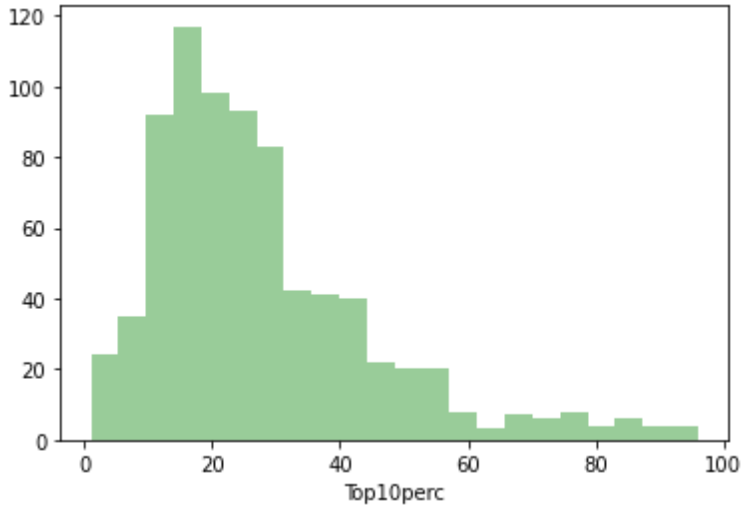


BoxPlot of Enroll

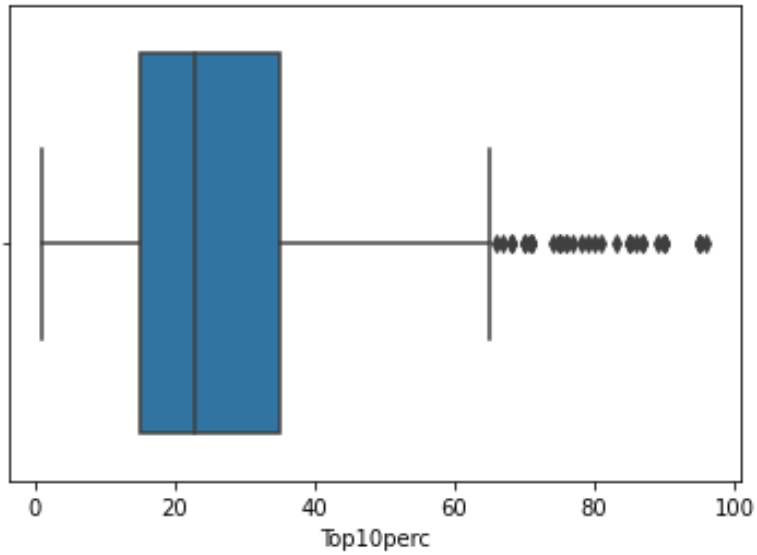


Description of Top10perc

```
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
```

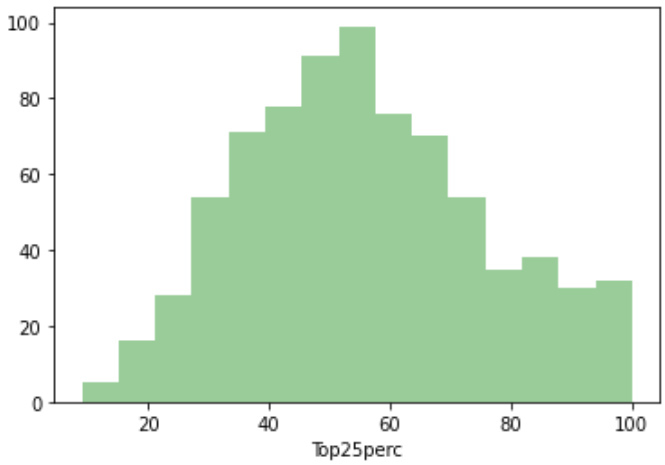


BoxPlot of Top10perc

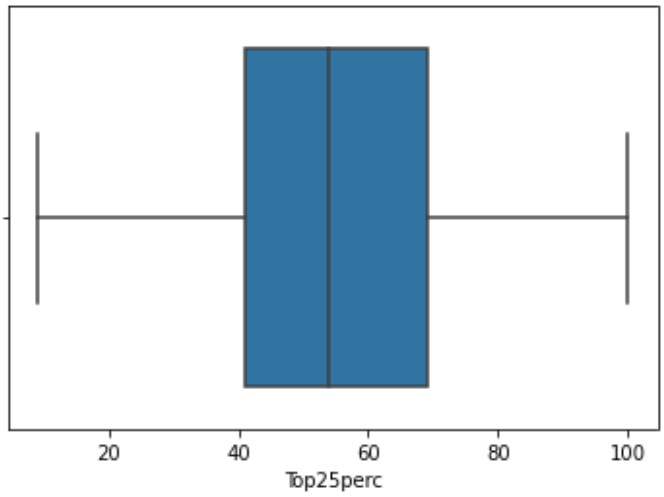


Description of Top25perc

```
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
```

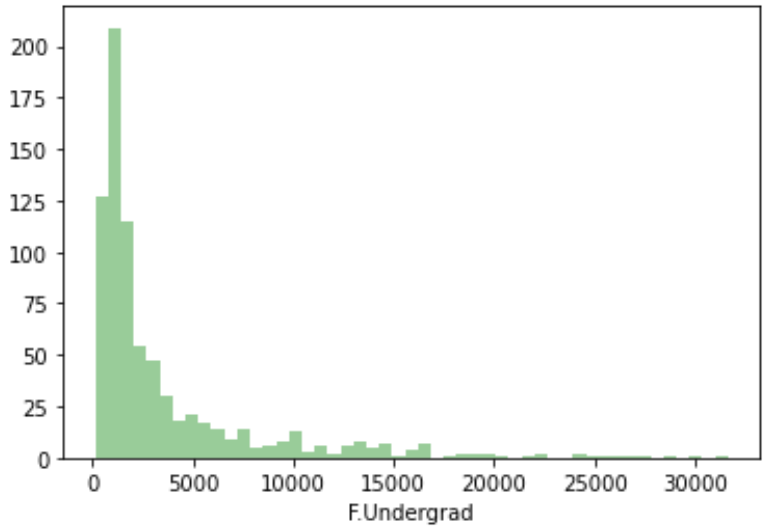


BoxPlot of Top25perc

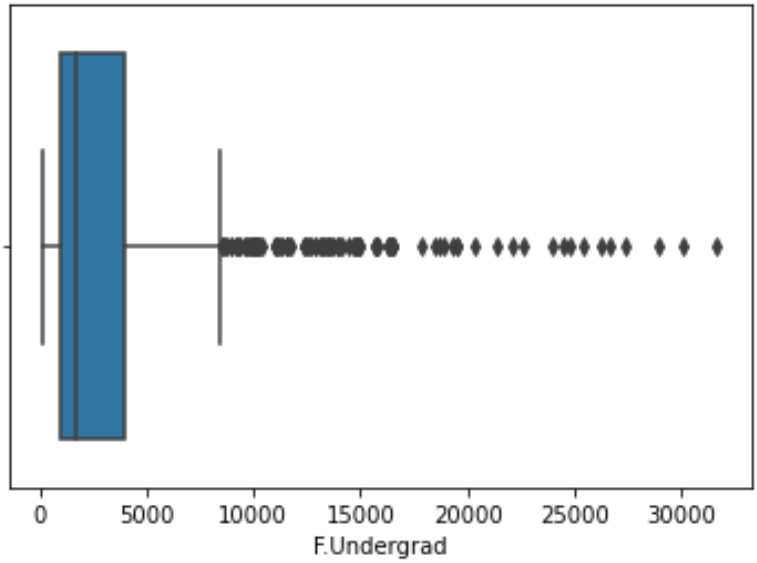


Description of F.Undergrad

```
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
```

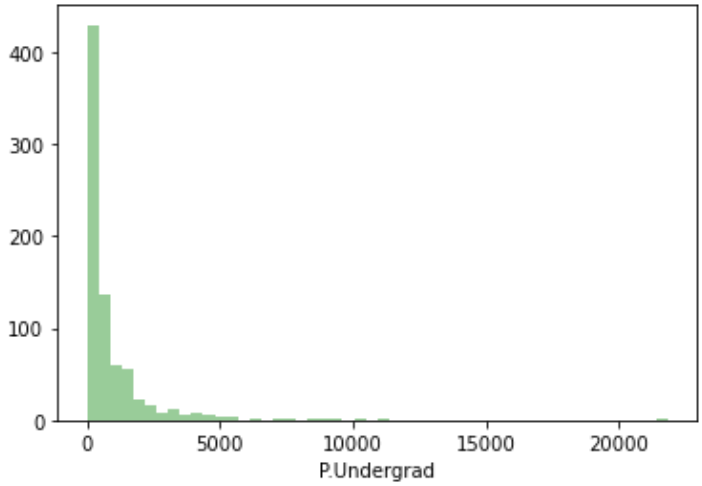


BoxPlot of F.Undergrad

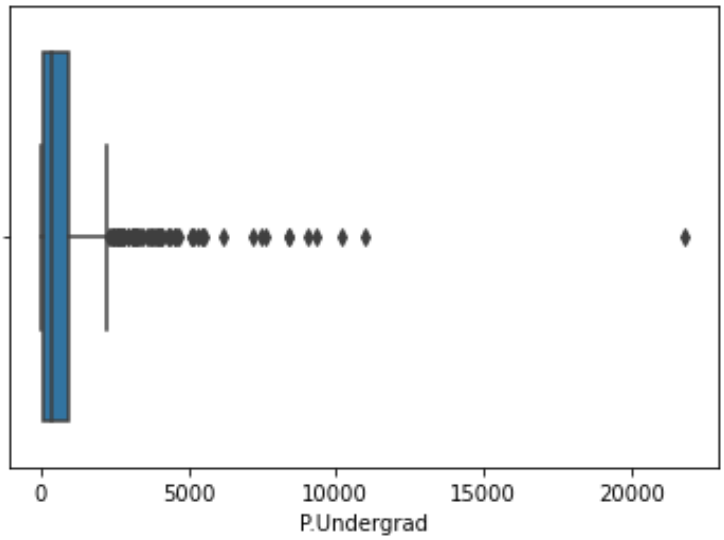


Description of P.Undergrad

```
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
```

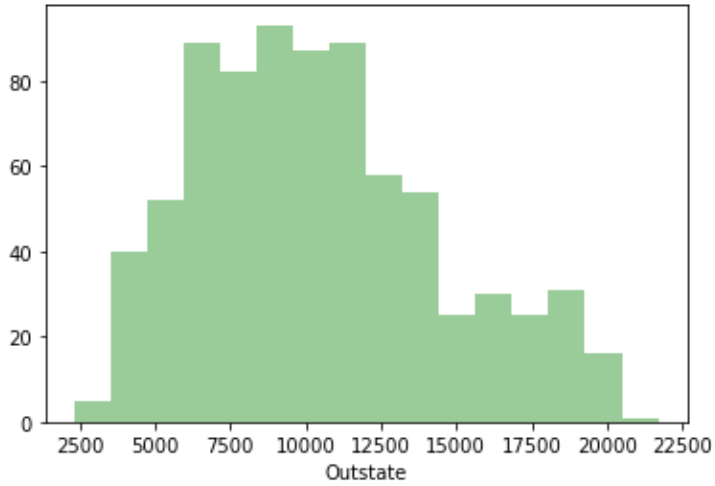


BoxPlot of P.Undergrad

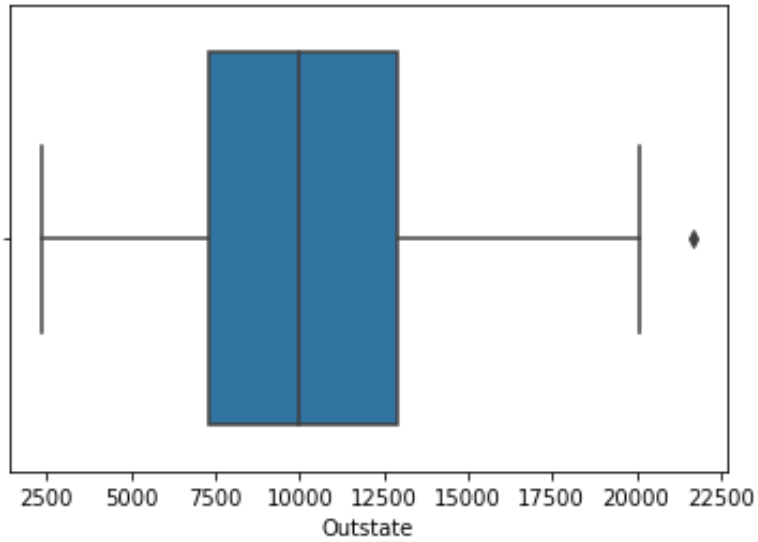


Description of Outstate

```
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
```

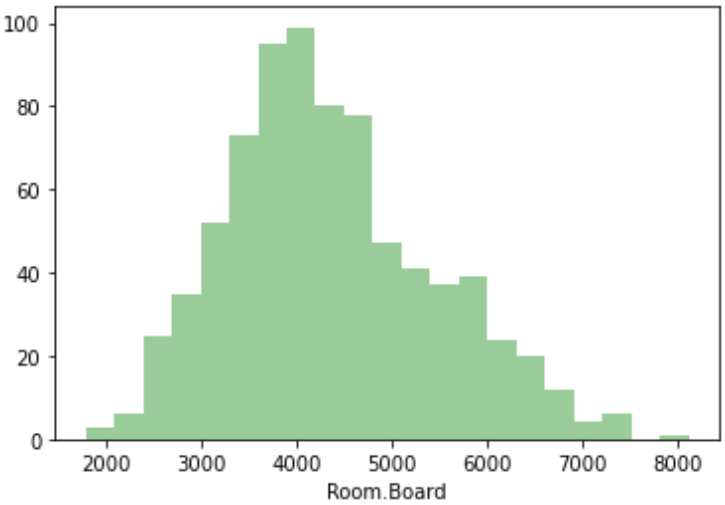


BoxPlot of Outstate

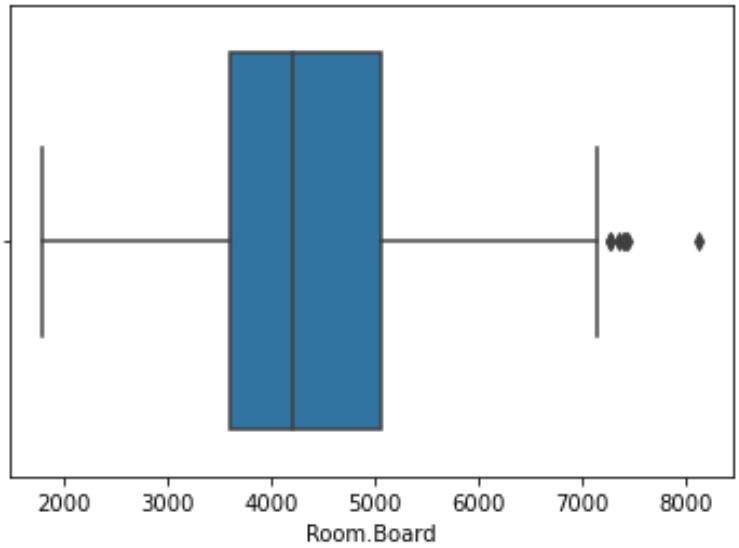


Description of Room.Board

```
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
```

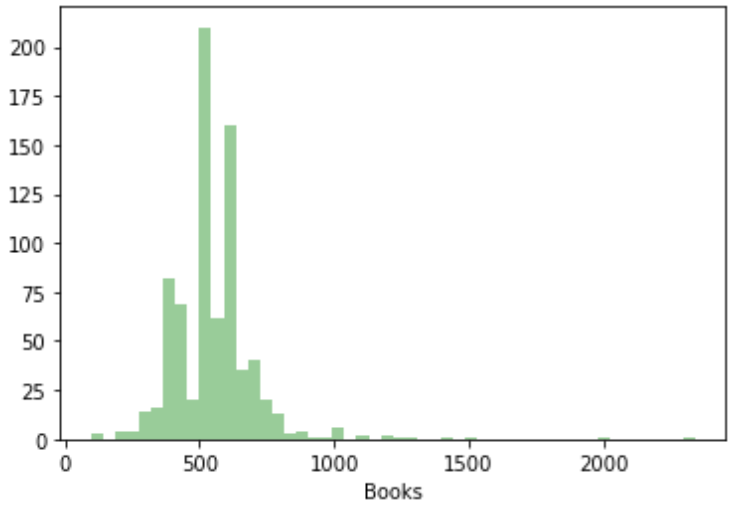


BoxPlot of Room.Board

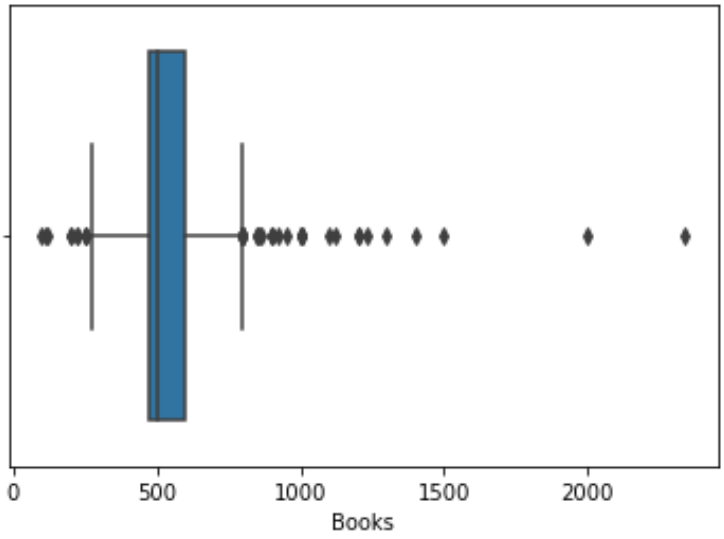


Description of Books

```
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
```

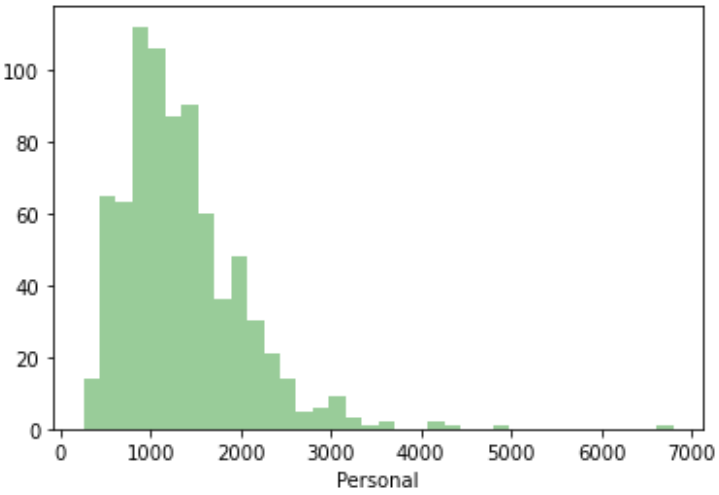


BoxPlot of Books

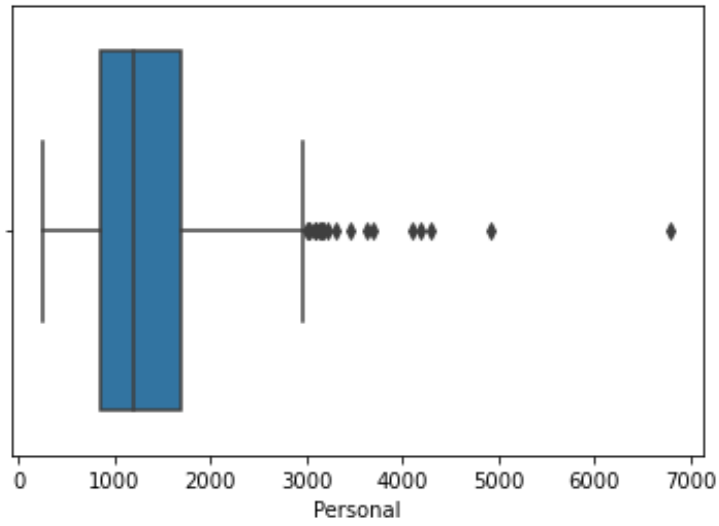


Description of Personal

```
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
```

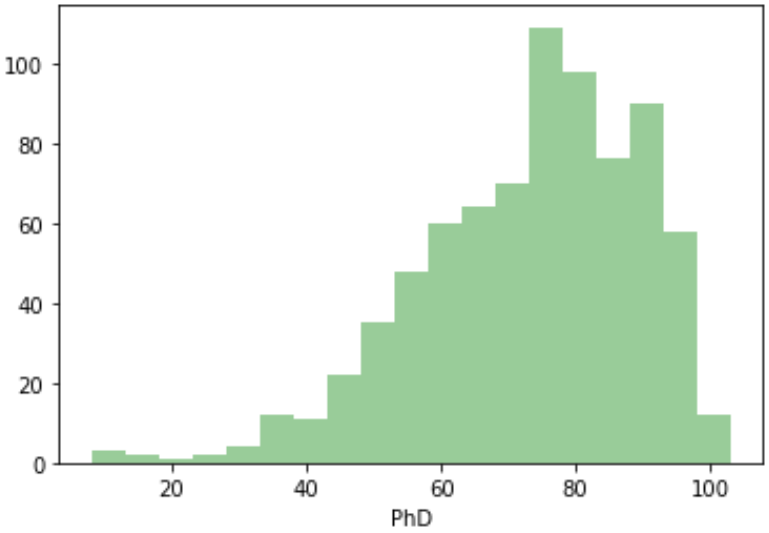


BoxPlot of Personal

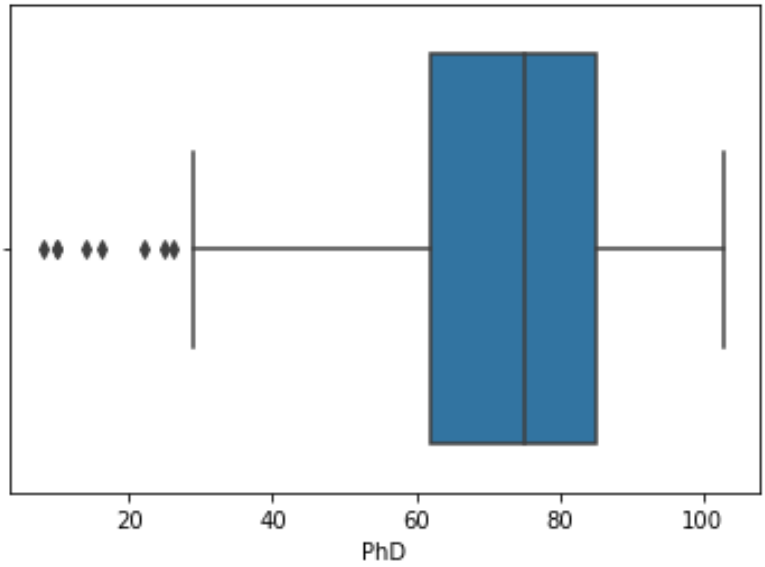


Description of PhD

```
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
```

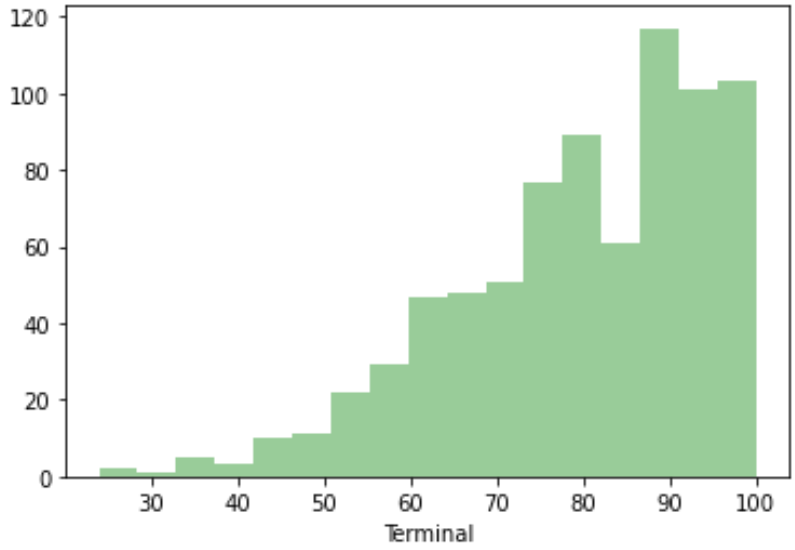


BoxPlot of PhD

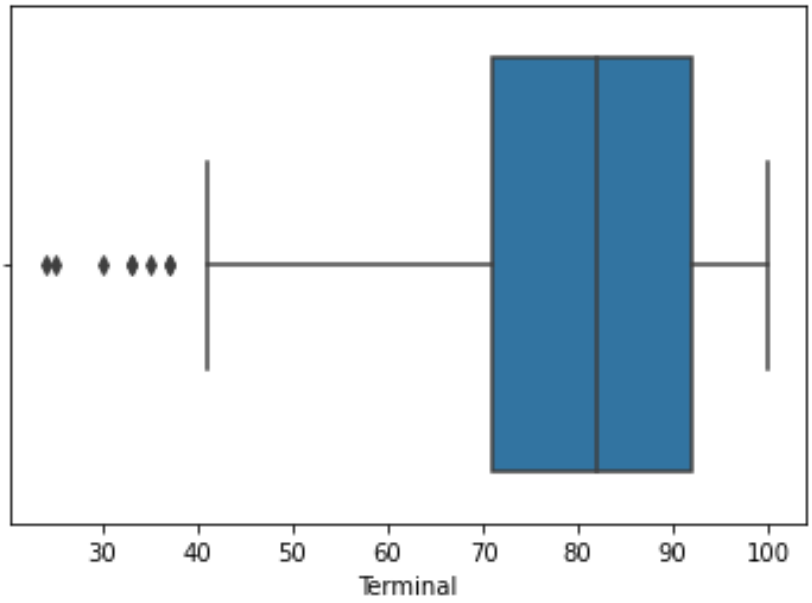


Description of Terminal

```
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
```

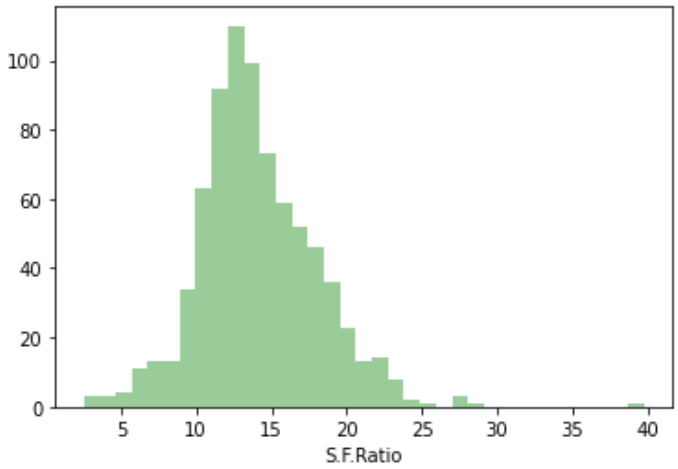


BoxPlot of Terminal

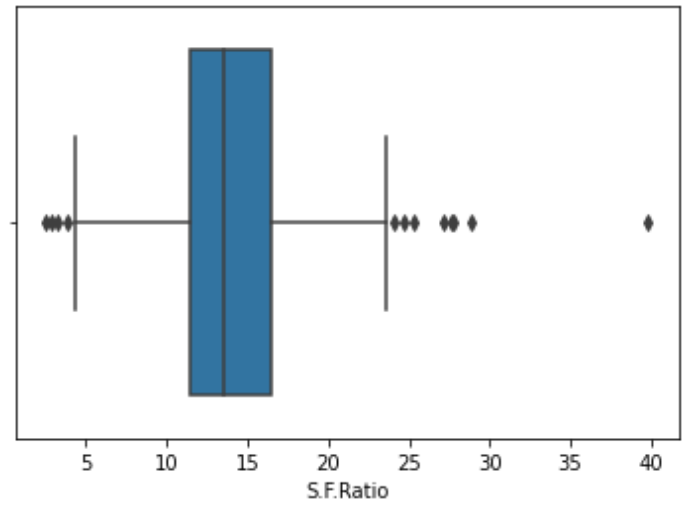


Description of S.F.Ratio

```
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
```

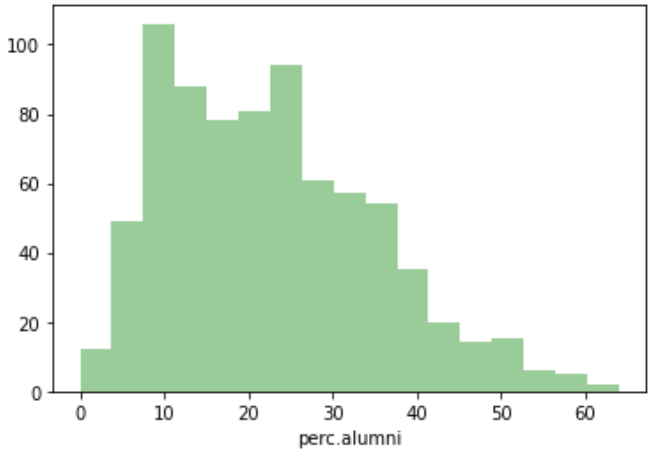


BoxPlot of S.F.Ratio

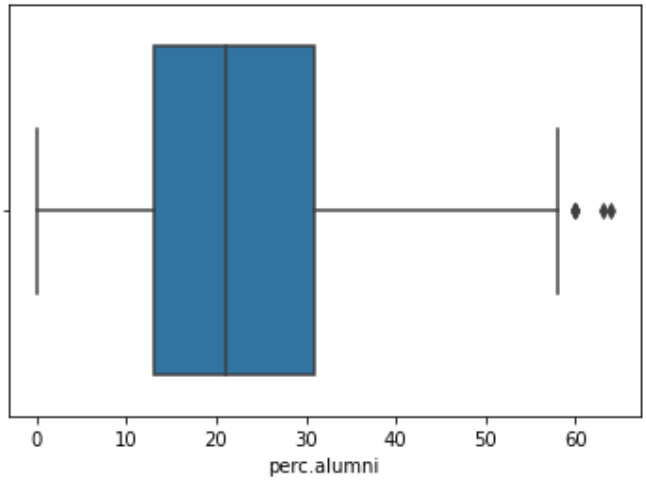


Description of perc.alumni

```
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
```

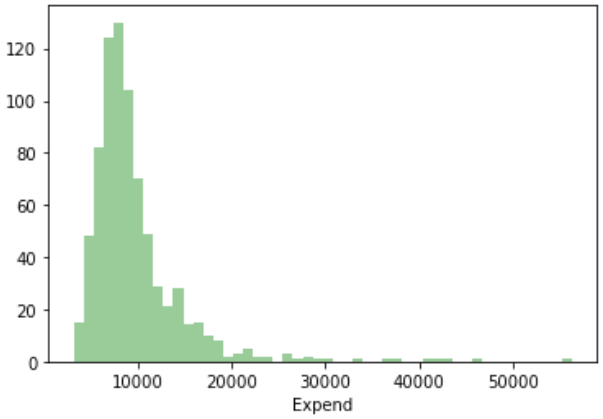


BoxPlot of perc.alumni

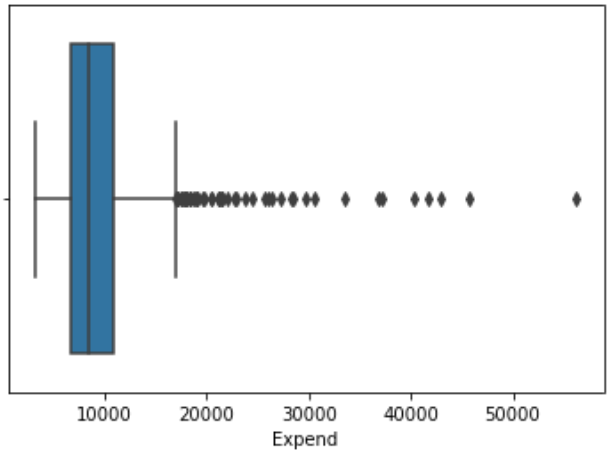


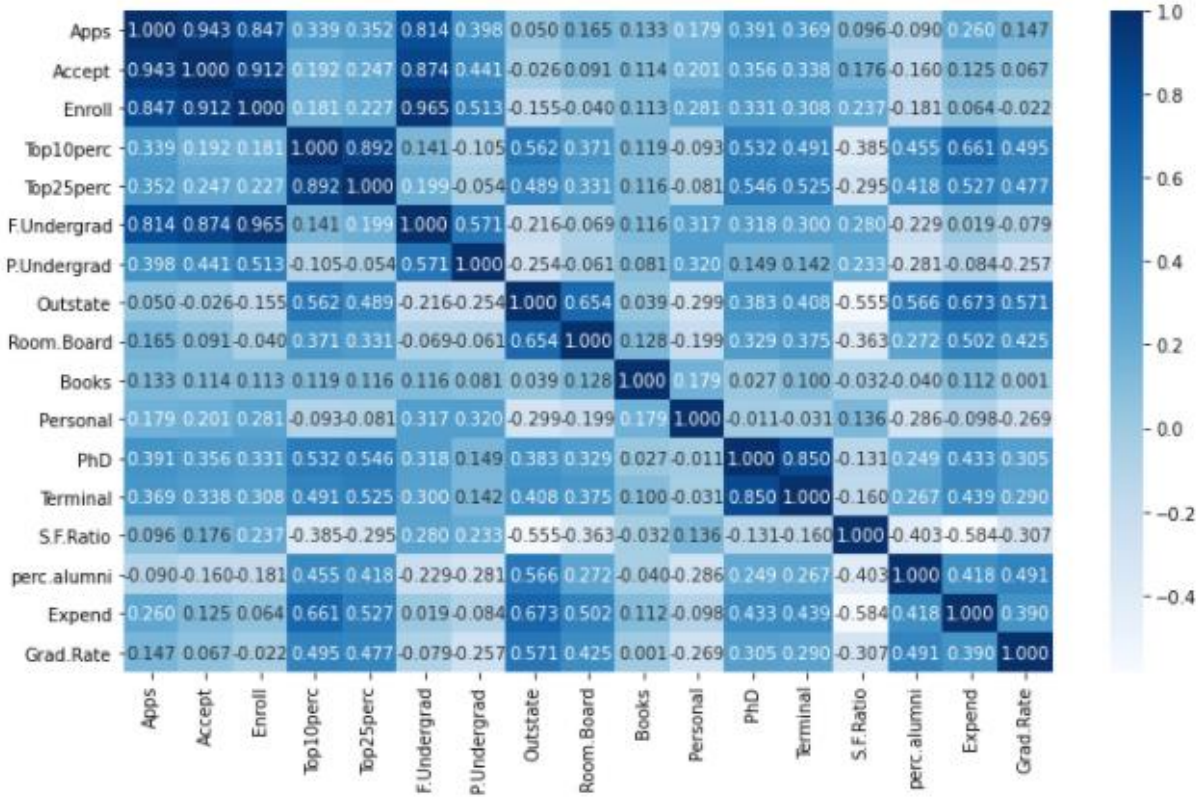
Description of Expend

```
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





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

Yes, scaling is necessary for PCA as comparison of colleges is done by various measures with different scales. Normalization of the variables is necessary so that all variables will be on a same scale to make comparison justified.

Before Scaling, data is spread across different weights

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9

After Scaling, data is centered towards the origin with same weight

```
from scipy.stats import zscore
df_num_scaled=df_num.apply(zscore)
df_num_scaled.head()
```

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.R
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	1.013
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	-0.477
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	-0.300
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	-1.615
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	-0.553

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

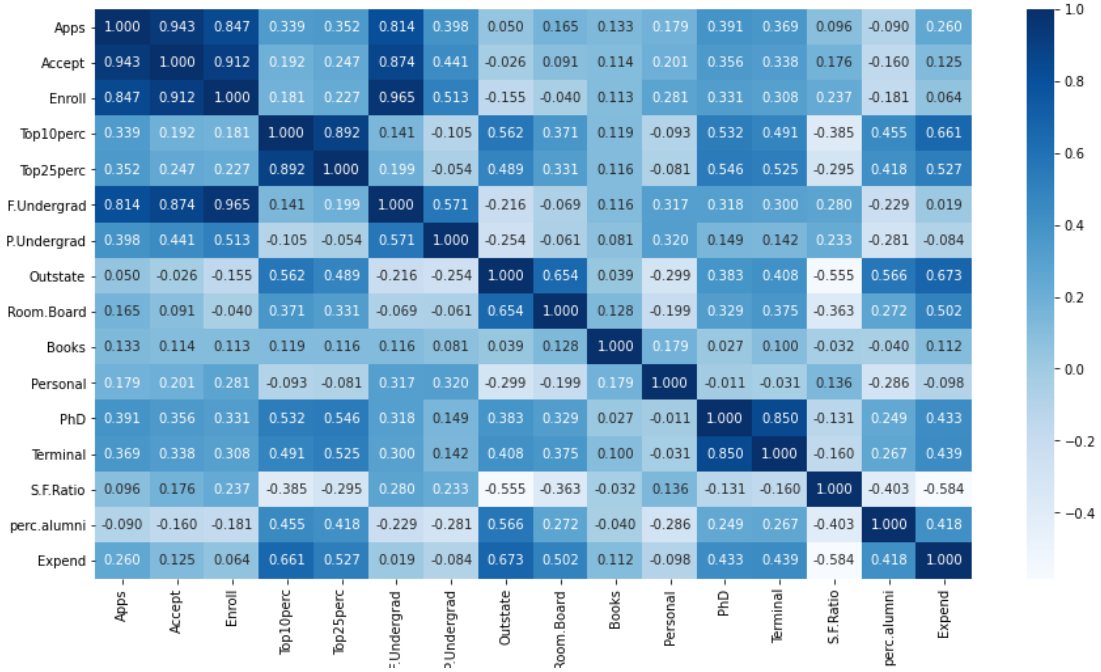
Correlation refers to the scaled form of covariance. Covariance indicates the direction of the linear relationship between variables; but Correlation on the other hand measures both the strength and direction of the linear relationship between two variables

Correlation is a measure used to represent how strongly two random variables are related to each other. Co-relation is mostly preferred as it remains unaffected by the change in dimensions, location, and scale, and can also be used to make a comparison between two pairs of variables

```
Covariance Matrix
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.8155401
8
   0.3987775   0.05022367  0.16515151  0.13272942  0.17896117  0.39120081
   0.36996762  0.09575627 -0.09034216  0.2599265 ]
 [ 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]
 [ 0.84791332  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]
 [ 0.33927032  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 ]
 [ 0.35209304  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]
[ 0.81554018  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]
[ 0.3987775   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]
[ 0.05022367 -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 ]
[ 0.16515151  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]
[ 0.13272942  0.11367165  0.11285614  0.1190116   0.115676    0.11569867
  0.08130416  0.03890494  0.12812787  1.00128866  0.17952581  0.0269404
  0.10008351 -0.03197042 -0.04025955  0.11255393]
[ 0.17896117  0.20124767  0.28129148 -0.09343665 -0.08091441  0.31760831
  0.32029384 -0.29947232 -0.19968518  0.17952581  1.00128866 -0.01094989
 -0.03065256  0.13652054 -0.2863366  -0.09801804]
[ 0.39120081  0.35621633  0.33189629  0.53251337  0.54656564  0.3187472
  0.14930637  0.38347594  0.32962651  0.0269404  -0.01094989  1.00128866
  0.85068186 -0.13069832  0.24932955  0.43331936]
[ 0.36996762  0.3380184   0.30867133  0.49176793  0.52542506  0.30040557
  0.14208644  0.40850895  0.3750222   0.10008351 -0.03065256  0.85068186
  1.00128866 -0.16031027  0.26747453  0.43936469]
[ 0.09575627  0.17645611  0.23757707 -0.38537048 -0.29500852  0.28006379
  0.23283016 -0.55553625 -0.36309504 -0.03197042  0.13652054 -0.13069832
 -0.16031027  1.00128866 -0.4034484  -0.5845844 ]
[-0.09034216 -0.16019604 -0.18102711  0.45607223  0.41840277 -0.22975792
 -0.28115421  0.56699214  0.27271444 -0.04025955 -0.2863366   0.24932955
  0.26747453 -0.4034484   1.00128866  0.41825001]
[ 0.2599265   0.12487773  0.06425192  0.6617651   0.52812713  0.01867565
 -0.08367612  0.6736456   0.50238599  0.11255393 -0.09801804  0.43331936
  0.43936469 -0.5845844   0.41825001  1.00128866]]
```

Correlation Matrix

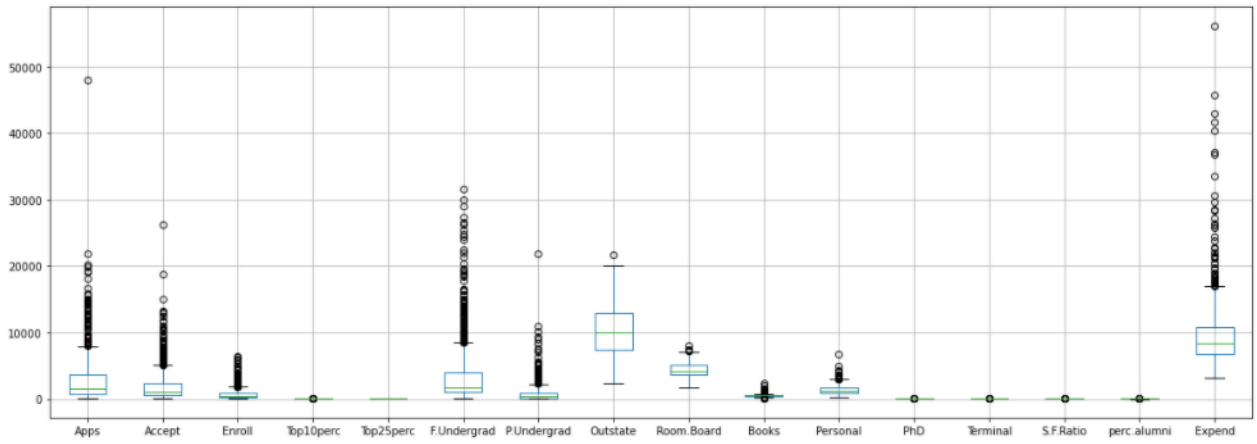


2.4 Check the dataset for outliers before and after scaling. What insight do you derive here?

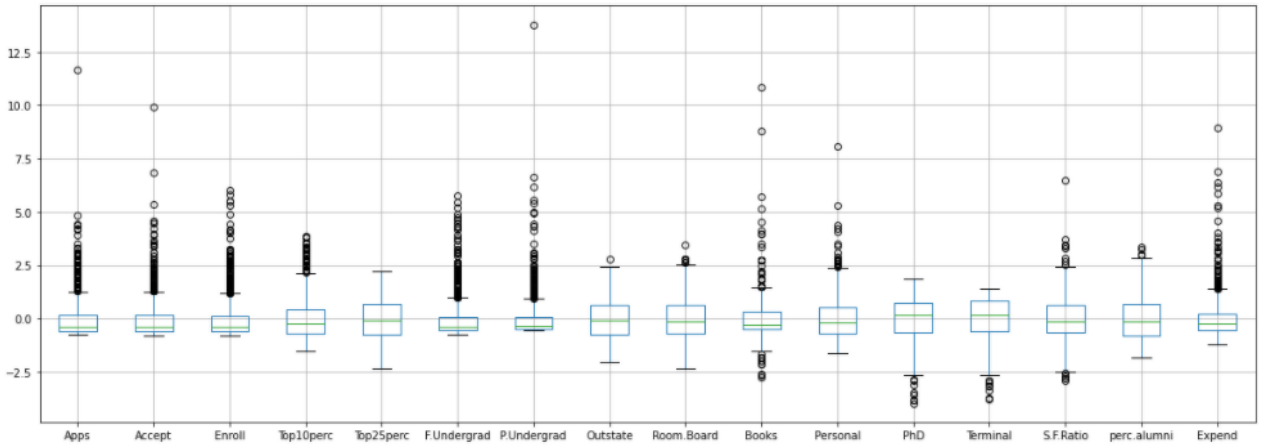
The Dataset has features with different "weights". It is recommended to transform the features so that all features are in same "scale" in order to do PCA. When sample variances of the original variables show differences by large order of magnitude, variables need to be normalized.

Before scaling the data, the outliers seem to look skewed towards the left which is not the correct way of interpretation.
For scaling the data, data on all the dimensions are subtracted from their means to shift the data points to the origin. i.e. the data is centered on the origins after scaling.
So, the outliers can be correctly interpreted from the box plot, once the data is scaled, i.e., outliers on either side of the mean can be clearly seen with the same weight across multiple variables.

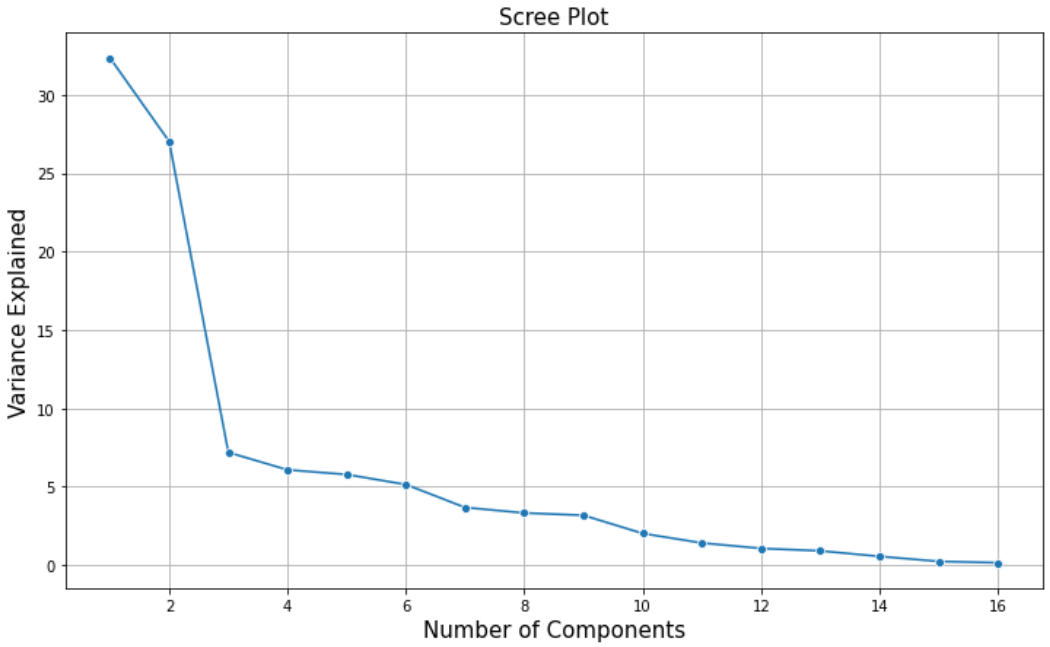
Outliers before scaling



Outliers after scaling



2.5 Perform PCA and export the data of the Principal Component scores into a data frame.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
0	-1.37	1.09	-0.18	-1.13	-0.42	-0.55	-0.87	-0.01
1	-2.21	-0.14	2.82	3.35	0.15	0.22	0.05	0.90
2	-1.52	-0.87	-0.43	0.86	-0.85	-0.52	0.31	-0.14
3	2.41	-3.44	-0.42	-0.70	-0.30	-0.71	-0.16	0.58
4	-1.55	-0.13	1.76	-0.61	0.74	0.28	-0.65	0.24
5	-0.74	-1.49	-0.14	0.30	-0.68	-0.18	-0.55	0.64
6	-0.03	-1.60	-0.08	-0.56	1.77	-0.47	-0.32	-0.76
7	1.36	-1.79	-1.24	-0.65	0.39	0.16	0.17	-0.39

2.6 Extract the eigenvalues, and eigenvectors.

Eigen Values

```
%s [5.18370633 4.32982721 1.15086658 0.97245765 0.925095   0.82220451
0.58787402 0.53169175 0.50798663 0.3222952  0.02311581 0.03673801
0.08802958 0.16938401 0.14443173 0.22491452]
```

Eigen Vectors

```
 %s [[-3.14525529e-01 -2.70550871e-01 -1.70182690e-02  2.80580801e-01
  -8.67487226e-02  6.06850978e-02  1.03369527e-01  9.20444027e-02
  -1.46208845e-02 -4.55476430e-02  3.56001005e-01  4.59248663e-01
  -1.32900015e-01  5.92119945e-01 -1.01451357e-01 -1.68743372e-04]
 [-2.81620660e-01 -3.21258052e-01 -5.48065186e-02  2.81243348e-01
  -3.13150224e-02  7.10965195e-02  5.62102996e-02  1.83691526e-01
  -1.87325427e-02  5.68164005e-02 -5.43860242e-01 -5.20576237e-01
   1.46372038e-01  2.77506082e-01 -4.95520272e-02 -1.45573271e-01]
 [-2.58189668e-01 -3.58727778e-01 -5.18688023e-02  1.51895234e-01
  -1.13481833e-01  1.24644140e-02 -5.84711059e-02  1.26300904e-01
   1.28156063e-02  4.95561066e-02  6.11229982e-01 -4.02576452e-01
  -3.00254489e-02 -4.42448793e-01  1.00713434e-01  3.72836422e-02]
 [-3.37987212e-01  1.71553974e-01  1.68355863e-02 -1.36056615e-01
  -3.89093760e-01  3.77693010e-02  1.24365937e-01 -3.53760828e-01
  -1.60844132e-02 -1.88110756e-02 -1.46242545e-01 -1.49741294e-01
  -6.97260239e-01 -2.83228053e-05  1.05186098e-01  4.94681325e-02]
 [-3.34482930e-01  1.30297586e-01 -4.24326279e-02 -2.11453115e-01
  -3.85768977e-01  1.07378753e-01  1.03698182e-01 -4.03766504e-01
  -7.01273879e-02  2.57211464e-01  7.85645687e-02  5.14382999e-02
   6.17718897e-01  1.64689623e-04 -1.59737272e-01 -5.92926744e-02]
 [-2.40326471e-01 -3.76799852e-01 -4.05240258e-02  9.37864412e-02
  -8.16273132e-02 -7.99146817e-03 -7.86628584e-02  5.56254438e-02
   1.56130317e-02  5.92500316e-02 -4.14917067e-01  5.59836026e-01
  -1.03742014e-02 -5.20847376e-01  7.43947113e-02  8.69883931e-02]
 [-9.53432697e-02 -3.01165003e-01  1.02236456e-01 -6.30591348e-02
   3.21045643e-01 -3.07653043e-01 -5.70860434e-01 -5.72342390e-01
   3.83926223e-02 -3.02351507e-02  1.09022741e-02 -5.20342234e-02
  -2.13760094e-02  1.36275297e-01 -1.79986529e-02 -1.00179783e-01]
 [-2.42213446e-01  3.20506976e-01  4.31467075e-02  2.07017218e-01
   1.59455590e-01 -2.19392238e-02 -1.12398503e-02  5.07702726e-02
  -2.19090249e-01 -1.43906484e-01  4.71827564e-02  1.00335671e-01
  -3.73152562e-02 -1.86925656e-01  2.18188201e-02 -8.04652736e-01]
 [-2.20144175e-01  1.98720162e-01  1.56137960e-01  3.05181746e-01
   5.16634245e-01  1.01089760e-01  2.18905394e-01 -1.86728640e-01
  -4.39288500e-01  2.62832211e-01 -5.19312095e-04 -2.65897896e-02
  -2.98135928e-03 -7.05172811e-02  5.75111121e-02  4.04429341e-01]
 [-7.88479206e-02 -3.63704229e-02  7.16514558e-01 -1.60318775e-01
   2.97187750e-02  6.05514105e-01 -2.11775289e-01  9.90758454e-02
   1.57256346e-01 -1.23758997e-02  1.01544048e-03  2.97716824e-03
   9.40426668e-03  1.55916815e-02  6.74339037e-02 -3.73919275e-02]
 [-6.56575400e-03 -2.24030456e-01  4.78810857e-01 -3.67291346e-01
  -1.14791945e-01 -5.11079169e-01  2.25968729e-01  2.17767506e-01
  -4.55907399e-01  1.76605432e-02 -2.72934465e-04 -1.25905067e-02
   2.92410997e-03  4.01663179e-02 -2.75665023e-02 -4.51328368e-02]
 [-3.33548932e-01  2.60849173e-02 -2.28523025e-01 -4.37332778e-01
   2.81805858e-01 -2.10914071e-02  7.72076288e-02  1.61432184e-01
   1.44753840e-01 -5.20013444e-03  1.27505331e-02  2.94124677e-02
   1.12628189e-01  1.56166029e-01  6.88179611e-01 -5.35710814e-03]
 [-3.30838506e-01  3.88675599e-02 -1.68161266e-01 -4.17282588e-01
   3.51481303e-01  3.82809226e-02  1.25791675e-02  2.26990738e-01
   1.60903729e-01  6.41764723e-02  7.45641944e-03 -2.65518308e-02
  -1.59558126e-01 -8.02345997e-02 -6.68060520e-01  3.12119824e-02]
 [ 1.30450185e-01 -2.95141381e-01 -2.68269979e-01 -2.38983030e-01
   5.95238899e-02  4.31972852e-01  8.08285950e-02 -1.73703151e-01
  -5.06076089e-01 -5.29756413e-01 -2.28221350e-03 -2.12954273e-02
   2.09194807e-02 -2.04613532e-02 -4.14763331e-02 -1.23617307e-02]
 [-1.51247389e-01  2.92413705e-01 -1.46203544e-01 -1.23719830e-02
```

```
  -2.35329320e-01   8.59831853e-04 -6.81631130e-01   3.47133444e-01
  -4.12783598e-01   2.04942331e-02 -2.20761400e-02   2.15944885e-03
   9.23565226e-03   9.34201784e-02  1.69458129e-02   2.21775874e-01]
 [-2.96896121e-01   2.17464850e-01  1.92610762e-01   1.54468500e-01
  -1.99215195e-02  -2.37434178e-01  5.67145124e-02  -6.76351318e-03
   2.13245712e-01  -7.38811708e-01 -3.32355069e-02  -4.31399727e-02
   2.27009067e-01  -7.09133273e-02 -6.06306213e-02   2.98409372e-01]]
```

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

```
 [ 0.31, 0.28,  0.26,  0.34,  0.33,  0.24,  0.1 ,  0.24,  0.22, 0.08,  0.01
,  0.33,  0.33, -0.13,  0.15,  0.3 ]
```

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?

Cumulative Value of the eigen Values is 16.02

Eigen Values

[5.18 4.33 1.15 0.97 0.93 0.82 0.59 0.53 0.51 0.32
0.02 0.04 0.09 0.17 0.14 0.22]

90% of 16.02 (Sum of Eigen Values) = 14.4 is obtained by adding first 8 Eigen values;

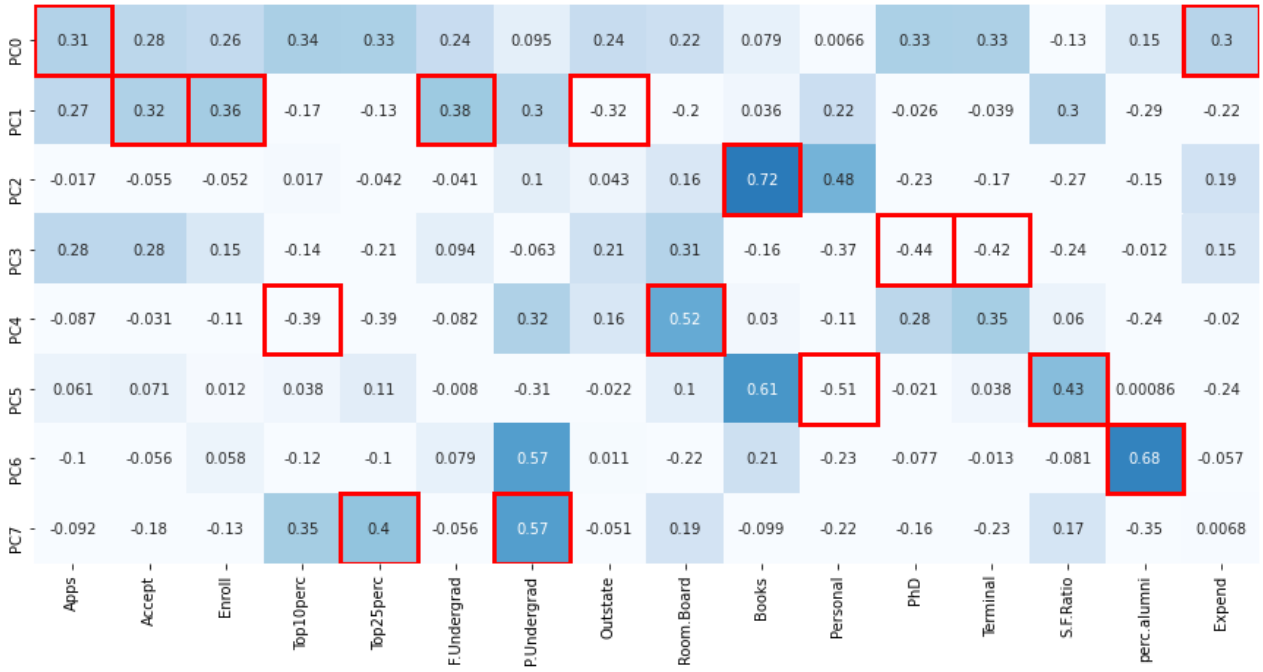
Thus it helps to decide on the optimum number of principal components for considering 90% variation; Number of variables reduced from 15 to 8

Eigen vectors are the principal components

Eigen vectors with insignificant contribution to total eigen values can be removed from analysis

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]

- ✓ Principal components help identify the appropriate college for a student.
- ✓ A student can zero down the factors he should consider when applying to these colleges to increase his chances of enrolment.
- ✓ With PCA, a student can have a view about the student to faculty ratio, faculties with Ph.D's and terminal degree and opt for the right mentor.
- ✓ Also, a student understands the expenses he may incur during his study period and plan his/her budget accordingly.
- ✓ With PCA, a student rightly understands the acceptance rate of a college and the graduation rate.
- ✓ Correlated attributes do not contribute to information in the data set. Further, correlated attributes create instability in the analysis of data. High dimensional data is not informative unless dimensions are orthogonal, i.e. uncorrelated or independent.



- For each feature, we find the maximum loading value across the components and mark the same with help of RED rectangular box
- Features marked with RED rectangular box are the ones having maximum loading on the respective component. We consider these marked features to decide the context that the component represents

- Number of applications received & The Instructional expenditure per student has max. loading on PC0
- Number of applications accepted, Number of new students enrolled, Number of full-time undergraduate students, Number of students for whom the particular college or university is Out-of-state tuition has max. loading on PC1
- Estimated book costs for a student has max. loading on PC2
- Percentage of faculties with Ph.D.'s and terminal degree has max. loading on PC3
- Percentage of new students from top 10% of Higher Secondary class, cost of Room and board has max. loading on PC4
- Estimated personal spending for a student, Student/faculty ratio has max. loading on PC5
- Percentage of alumni who donate has max. loading on PC6
- Percentage of new students from top 25% of Higher Secondary class, Number of part-time undergraduate students has max. loading on PC7