ADVANCED STATISTICS PROJECT

Balasubramaniyam.R

8/14/2022

Table of Contents

Table of Contents(Tables)

<u>Problem 1A:</u>

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.

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

Hypothesis for Education level and salary,

Null hypothesis, H_0: Mean salary is the same on all the education level.

Alternative hypothesis, H_a: Mean salary is not the same on all the education level.

Hypothesis for Occupation and Salary,

Null hypothesis, H_0: Mean salary is the same on all the Occupation levels.

Alternative hypothesis, H_a: Mean salary is not the same on all the Occupation levels.

1.2 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.

1.2.1 One-Way ANOVA table,

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.03e+11	5.13e+10	30.95628	1.26e-08
Residual	37.0	6.14e+10	1.66e+09	NaN	NaN

Since the p-value is less than 0.05, we reject the null hypothesis.

1.3 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.

1.3.1 One-Way ANOVA table,

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.13e+10	3.75e+09	0.884144	0.458508
Residual	36.0	1.53e+11	4.24e+09	NaN	NaN

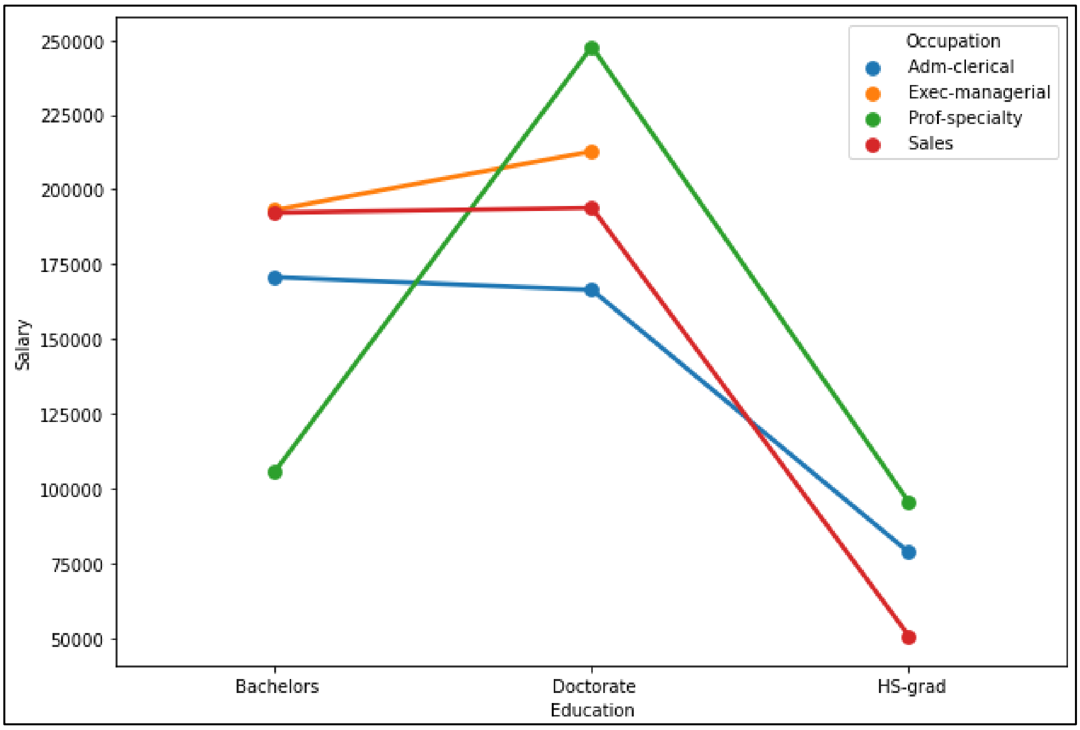
Since the p-value is more than 0.05, we accept the null hypothesis.

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

The null hypothesis is rejected in (1.2). The salary mean is significantly different across the different education levels.

Problem 1B:

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.



i. The employees with high school graduation are getting lower salaries than the employees with Bachelors, and doctorates.

ii. In sales occupation, the employees with both the Bachelors, and the Doctorates are getting a similar salary.

iii. Employees with doctorate education in Professional or specialty occupation is earning more salary than all other employees.

iv. Employees with both the Bachelors, and the Doctorates are working in all the four occupation levels whereas the employees with high school graduation are not working in Executive or managerial level positions.

1.6 **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?**

Hypothesis for interaction between education and occupation

Null hypothesis,

Null hypothesis for interaction, H_0: There is no interaction between education and Occupation.

Alternative hypothesis,

Alternative hypothesis for interaction, H_a: There is an interaction between the education and Occupation.

1.6.1 Two-Way ANOVA table,

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.03e+11	5.13e+10	72.21196	5.47e-12
C(Occupation)	3.0	5.52e+09	1.84e+09	2.587626	7.21e-02
C(Education):C(Occupation)	6.0	3.63e+10	6.06e+09	8.519815	2.23e-05
Residual	29.0	2.06e+10	7.11e+08	NaN	NaN

Results: Since the p-value for Education*Occupation is less than 0.05, we reject the null hypothesis. There is an interaction between the education and occupation on the mean salary.

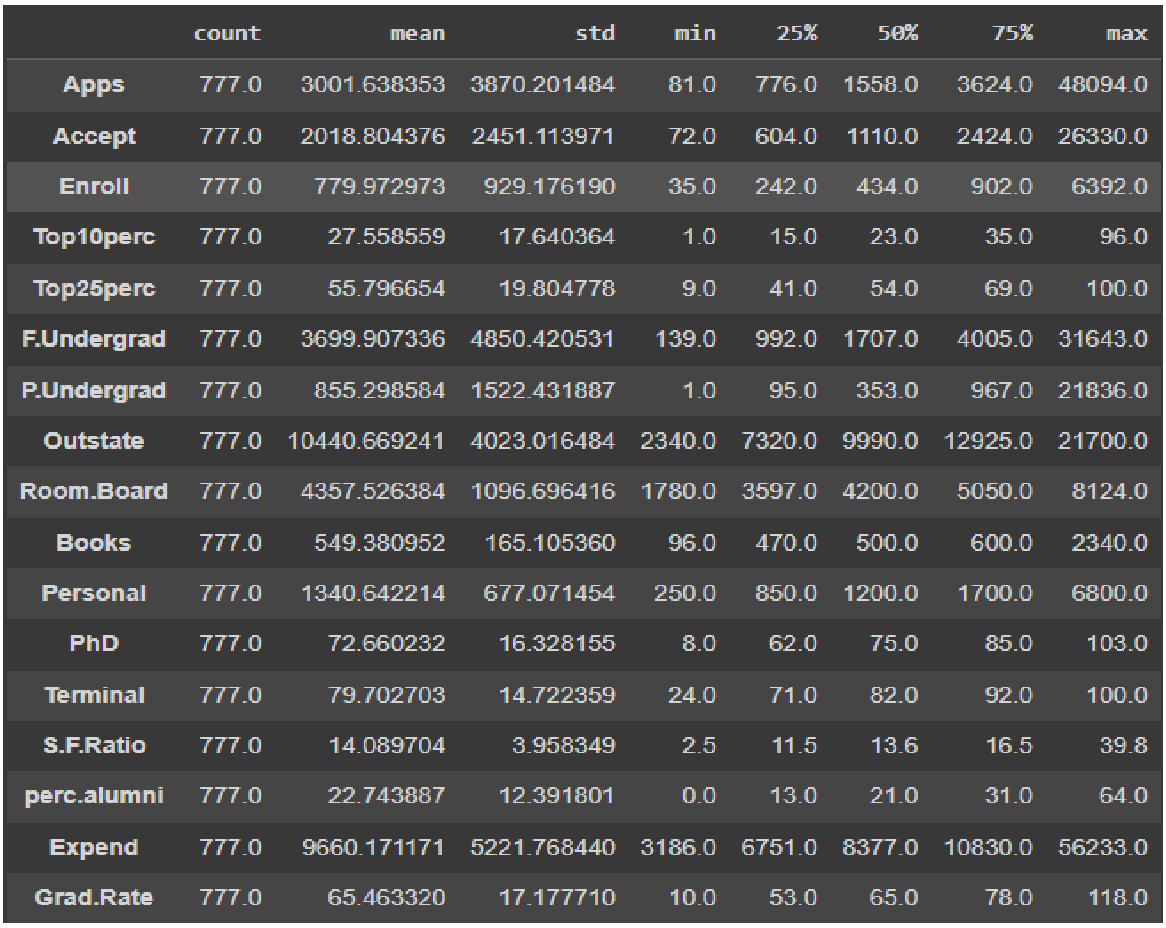
1.7 Explain the business implications of performing ANOVA for this particular case study.

By performing the ANOVA and observing the interaction between education and occupation, it is found that the education interacts with occupation, and it results in higher and better salaries among the highly educated employees. Employees with Doctorate are earning more salary than the employees with high school graduation.

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.

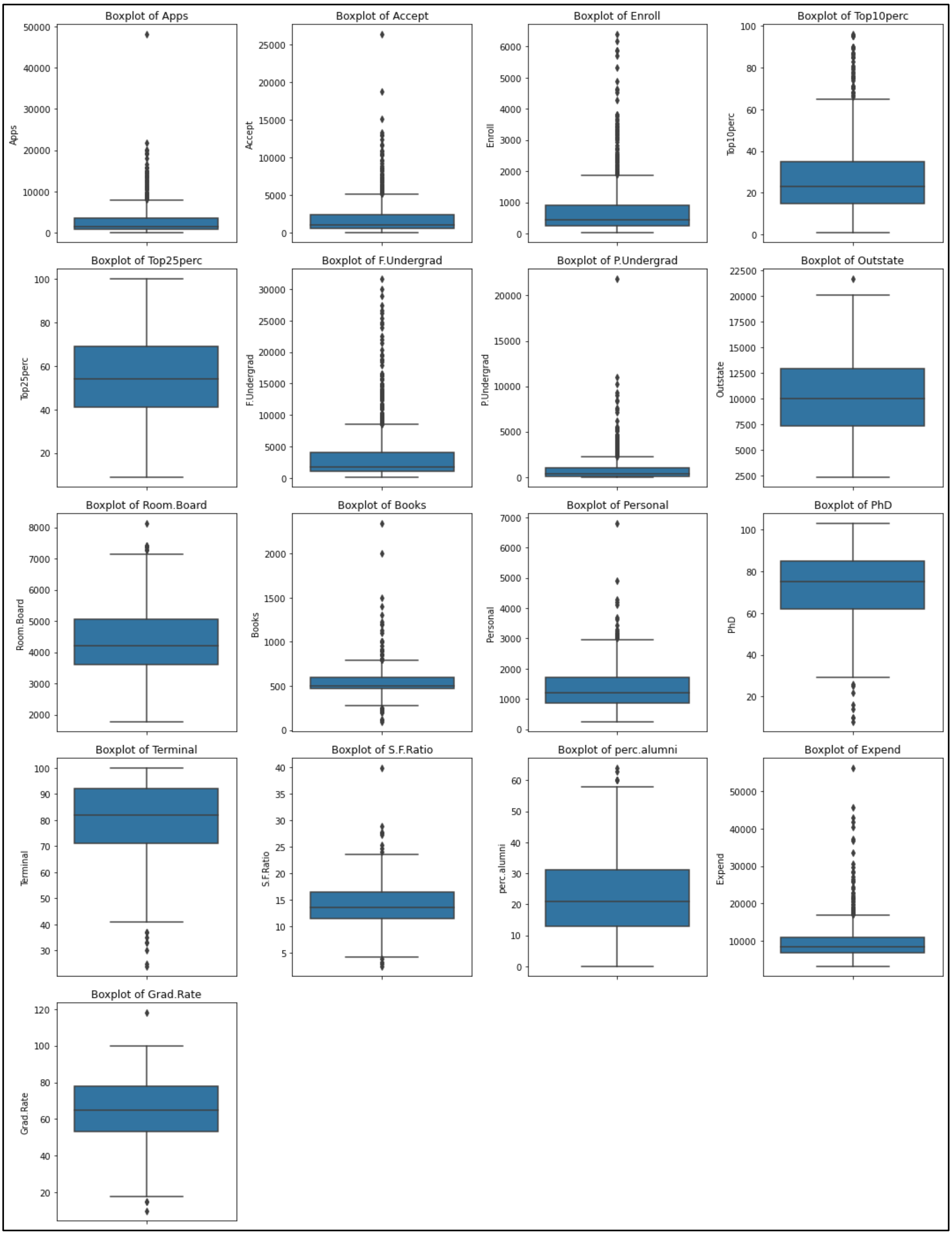
2.1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?



	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

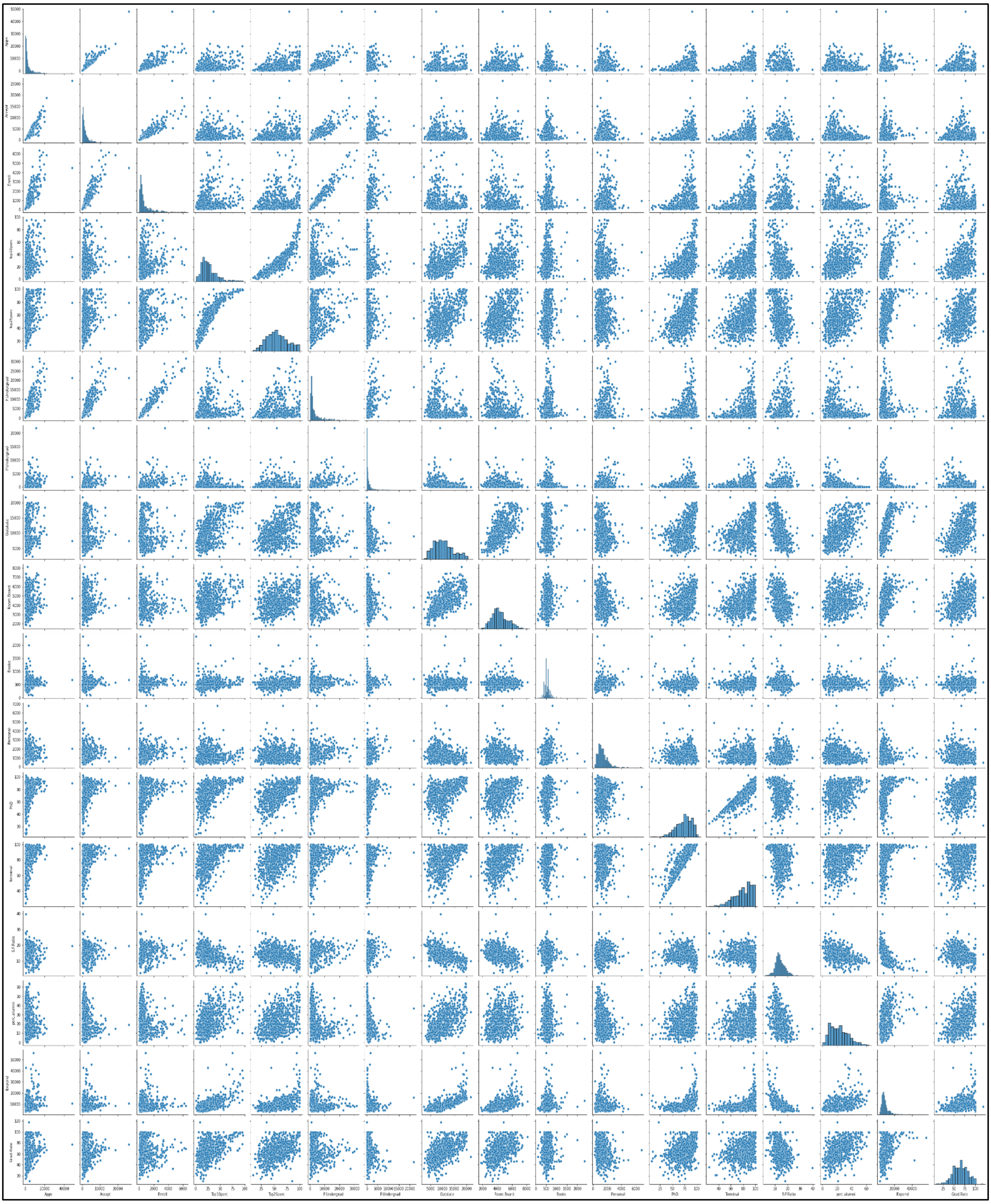
2.1.1. Univariate analysis (Box plot)

The univariate analysis helps us to understand the various features present in our dataset. It visually gives us an idea about how the data is distributed, skewness and all the features in our dataset except Top25perc feature are having outliers.



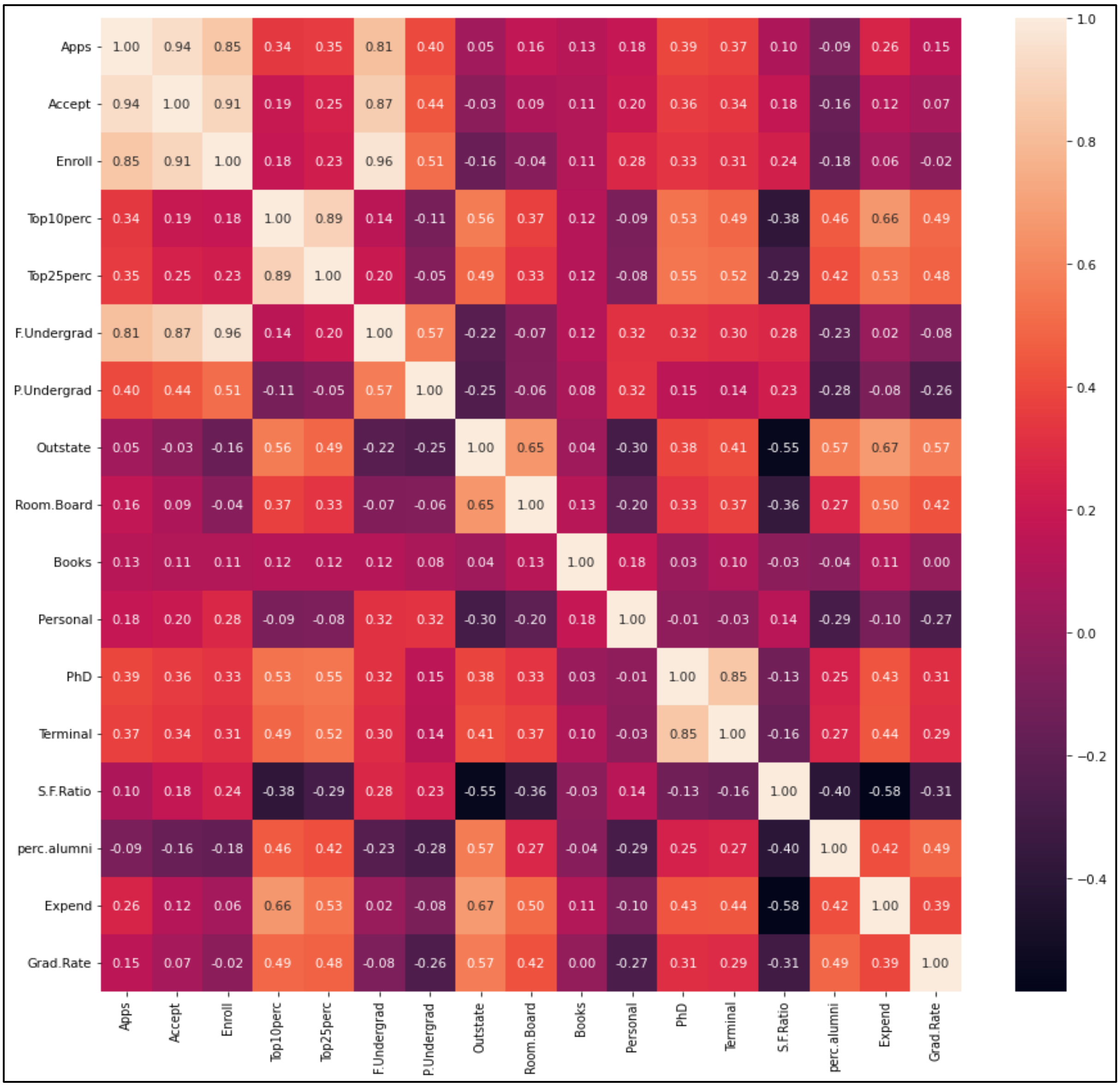
2.1.2. Multivariate analysis (Pair plot)

The pair plot helps us to understand the relationship and patterns between the features of the dataset.



2.1.3. Multivariate analysis (Heatmap)

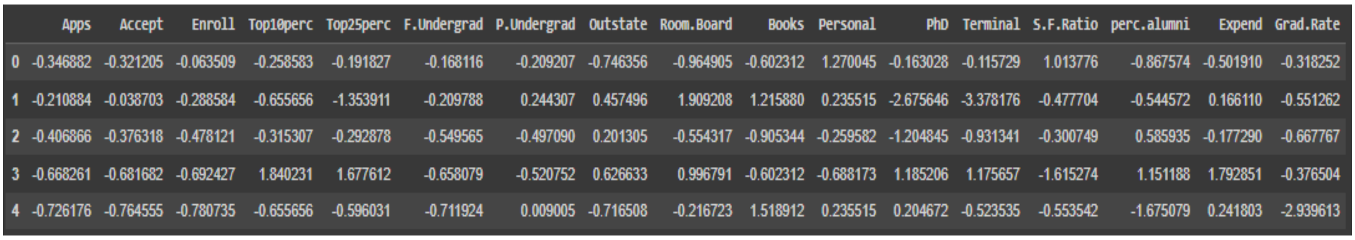
This heatmap gives us the correlation between the numerical variables. Here we can see the number of applications received is having high correlation with number of applications accepted (0.94) and number of new students enrolled (0.85).



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

The data in our dataset has highly varying data points with different magnitudes (like extreme values in each feature), scaling is done so that our data gets standardized within a specific scale, like (0-1). Model will treat the scaled data with equal weightage to all datapoints and higher magnitude data-points does not dominate over the lower magnitude data points. Therefore, scaling is required for our dataset to perform PCA (Scaling of data is done here by applying Z-score).

2.2.1. Scaled data table



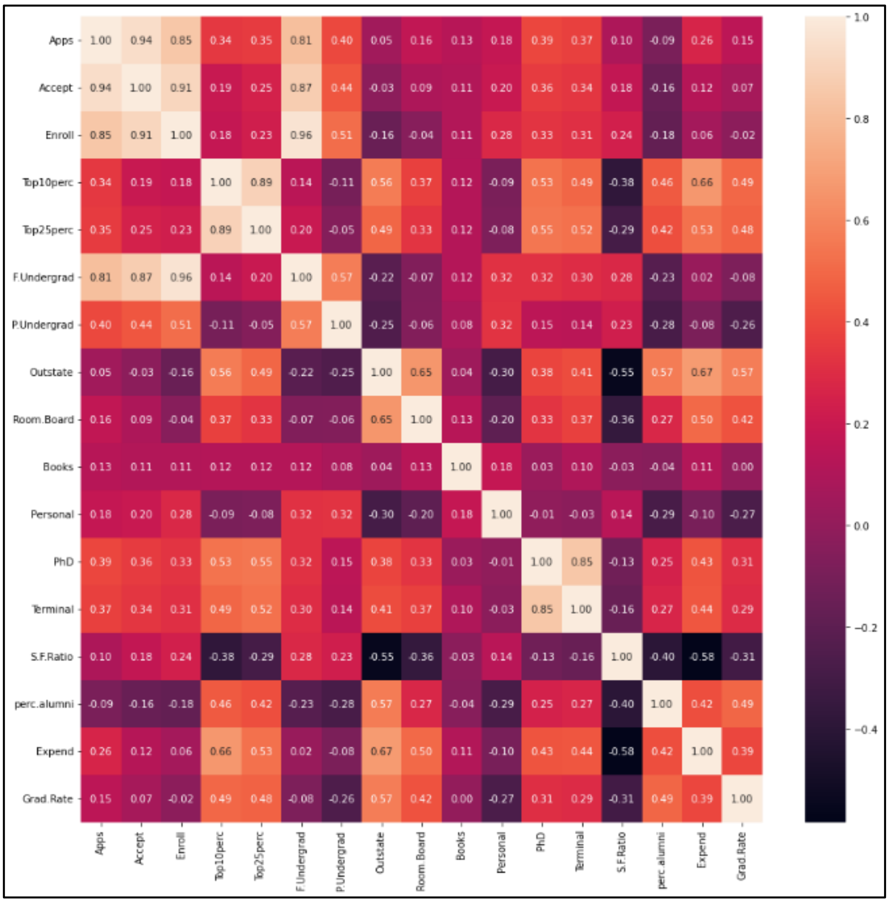
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.163028	-0.115729	1.013776	-0.867574	-0.501910	-0.318252
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	-0.477704	-0.544572	0.166110	-0.551262
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	-0.300749	0.585935	-0.177290	-0.667767
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	-1.615274	1.151188	1.792851	-0.376504
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	-0.553542	-1.675079	0.241803	-2.939613

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

Both the covariance and the correlation matrices are used to find the relationship between the features. The covariance matrix gives us the direction of relationship (positive or negative) and the correlation matrix gives us both direction and strength of the relationship.

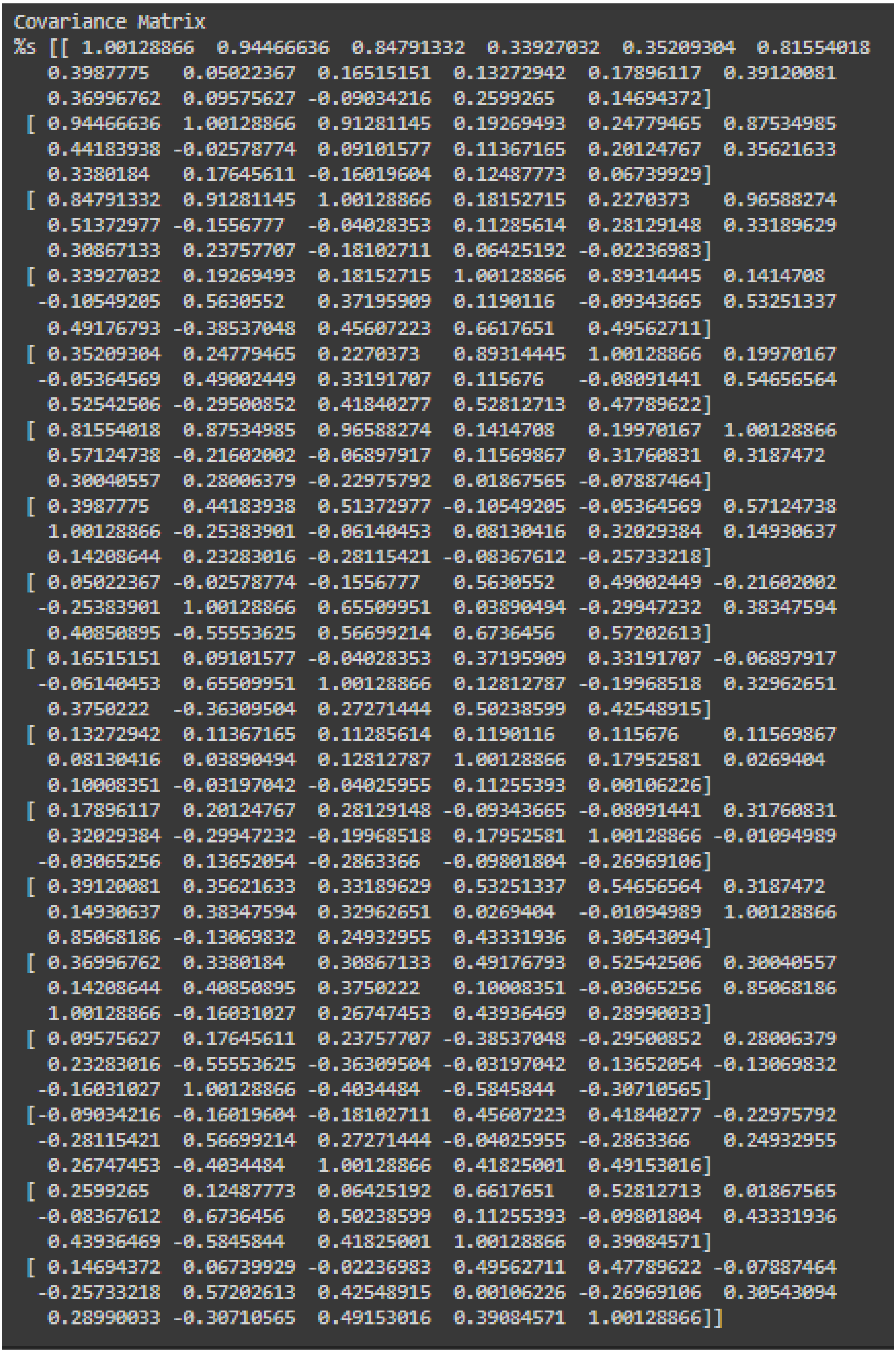
2.3.1. Correlation matrix

The correlation matrix/heatmap helps us to understand the strength and type (positive or negative) of relation between the variables. Here Top10% and Top25% are highly corelated.



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.00	0.94	0.85	0.34	0.35	0.81	0.40	0.05	0.16	0.13	0.18	0.39	0.37	0.10	-0.09	0.26	0.15
Accept	0.94	1.00	0.91	0.19	0.25	0.87	0.44	-0.03	0.09	0.11	0.20	0.36	0.34	0.18	-0.16	0.12	0.07
Enroll	0.85	0.91	1.00	0.18	0.23	0.96	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	-0.18	0.06	-0.02
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.37	0.12	-0.09	0.53	0.49	-0.38	0.46	0.66	0.49
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33	0.12	-0.08	0.55	0.52	-0.29	0.42	0.53	0.48
F.Undergrad	0.81	0.87	0.96	0.14	0.20	1.00	0.57	-0.22	-0.07	0.12	0.32	0.32	0.30	0.28	-0.23	0.02	-0.08
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	-0.06	0.08	0.32	0.15	0.14	0.23	-0.28	-0.08	-0.26
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.65	0.04	-0.30	0.38	0.41	-0.55	0.57	0.67	0.57
Room.Board	0.16	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.65	1.00	0.13	-0.20	0.33	0.37	-0.36	0.27	0.50	0.42
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.08	0.04	0.13	1.00	0.18	0.03	0.10	-0.03	-0.04	0.11	0.00
Personal	0.18	0.20	0.28	-0.09	-0.08	0.32	0.32	-0.30	-0.20	0.18	1.00	-0.01	-0.03	0.14	-0.29	-0.10	-0.27
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.03	-0.01	1.00	0.85	-0.13	0.25	0.43	0.31
Terminal	0.37	0.34	0.31	0.49	0.52	0.30	0.14	0.41	0.37	0.10	-0.03	0.85	1.00	-0.16	0.27	0.44	0.29
S.F.Ratio	0.10	0.18	0.24	-0.38	-0.29	0.28	0.23	-0.55	-0.36	-0.03	0.14	-0.13	-0.16	1.00	-0.40	-0.58	-0.31
perc.alumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.27	-0.04	-0.29	0.25	0.27	-0.40	1.00	0.42	0.49
Expend	0.26	0.12	0.06	0.66	0.53	0.02	-0.08	0.67	0.50	0.11	-0.10	0.43	0.44	-0.58	0.42	1.00	0.39
Grad.Rate	0.15	0.07	-0.02	0.49	0.48	-0.08	-0.26	0.57	0.42	0.00	-0.27	0.31	0.29	-0.31	0.49	0.39	1.00

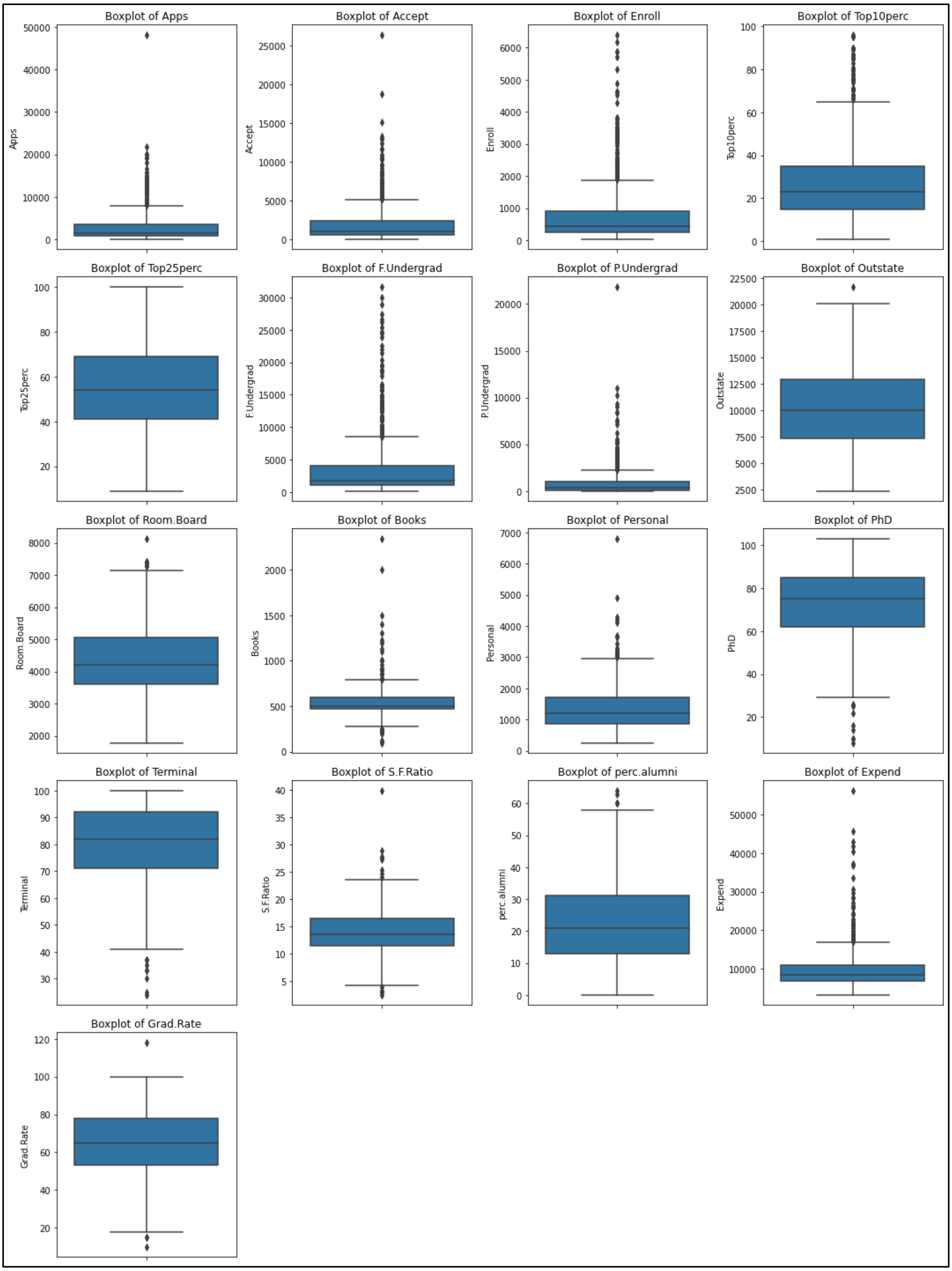
2.3.2. Covariance Matrix



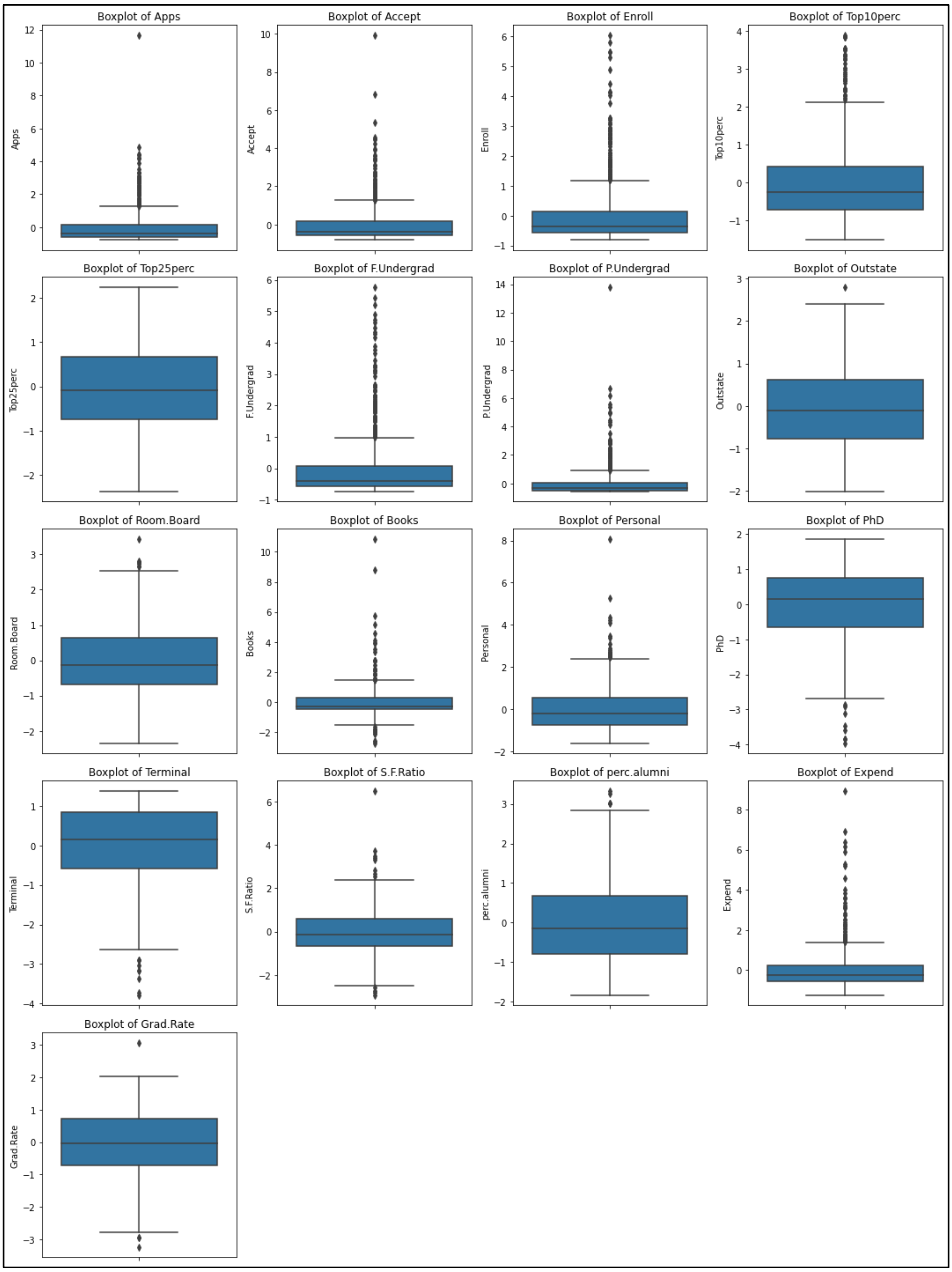
```
Covariance Matrix
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.81554018
   0.3987775   0.05022367  0.16515151  0.13272942  0.17896117  0.39120081
   0.36996762  0.09575627 -0.09034216  0.2599265   0.14694372]
 [ 0.94466636  1.00128866  0.91281145  0.19269493  0.24779465  0.87534985
   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]
 [ 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  0.00106226]
 [ 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 -0.26969106]
 [ 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  0.30543094]
 [ 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  0.28990033]
 [ 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  -0.30710565]
 [-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  0.49153016]
 [ 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  0.39084571]
 [ 0.14694372  0.06739929 -0.02236983  0.49562711  0.47789622 -0.07887464
  -0.25733218  0.57202613  0.42548915  0.00106226 -0.26969106  0.30543094
   0.28990033 -0.30710565  0.49153016  0.39084571  1.00128866]]
```

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

2.4.1. Box plot before scaling,



2.4.2. Box plot after scaling,



The outliers are present in the dataset both before and after the scaling of data. However, the range in the y-axis have changed to a comparable range after scaling the data. Scaling of data have not removed the outliers.

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

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,
2.04719730e-01, -7.93882496e-02, -2.16297411e-01, 7.59581203e-02, -
1.09267913e-01], [-1.62374420e-02, 7.53468452e-03, -4.25579803e-02, -
5.26927980e-02, 3.30915896e-02, -4.34542349e-02, -1.91198583e-01, -
3.00003910e-02, 1.62755446e-01, 6.41054950e-01, -3.31398003e-01,
9.12555212e-02, 1.54927646e-01, 4.87045875e-01, -4.73400144e-02, -
2.98118619e-01, 2.16163313e-01], [-4.24863486e-02, -1.29497196e-02, -
2.76928937e-02, -1.61332069e-01, -1.18485556e-01, -2.50763629e-02,
6.10423460e-02, 1.08528966e-01, 2.09744235e-01, -1.49692034e-01,
6.33790064e-01, -1.09641298e-03, -2.84770105e-02, 2.19259358e-01,
2.43321156e-01, -2.26584481e-01, 5.59943937e-01], [-1.03090398e-01, -
5.62709623e-02, 5.86623552e-02, -1.22678028e-01, -1.02491967e-01,
7.88896442e-02, 5.70783816e-01, 9.84599754e-03, -2.21453442e-01,
```

2.13293009e-01, -2.32660840e-01, -7.70400002e-02, -1.21613297e-02, -8.36048735e-02, 6.78523654e-01, -5.41593771e-02, -5.33553891e-03], [-9.02270802e-02, -1.77864814e-01, -1.28560713e-01, 3.41099863e-01, 4.03711989e-01, -5.94419181e-02, 5.60672902e-01, -4.57332880e-03, 2.75022548e-01, -1.33663353e-01, -9.44688900e-02, -1.85181525e-01, -2.54938198e-01, 2.74544380e-01, -2.55334907e-01, -4.91388809e-02, 4.19043052e-02], [5.25098025e-02, 4.11400844e-02, 3.44879147e-02, 6.40257785e-02, 1.45492289e-02, 2.08471834e-02, -2.23105808e-01, 1.86675363e-01, 2.98324237e-01, -8.20292186e-02, 1.36027616e-01, -1.23452200e-01, -8.85784627e-02, 4.72045249e-01, 4.22999706e-01, 1.32286331e-01, -5.90271067e-01], [4.30462074e-02, -5.84055850e-02, -6.93988831e-02, -8.10481404e-03, -2.73128469e-01, -8.11578181e-02, 1.00693324e-01, 1.43220673e-01, -3.59321731e-01, 3.19400370e-02, -1.85784733e-02, 4.03723253e-02, -5.89734026e-02, 4.45000727e-01, -1.30727978e-01, 6.92088870e-01, 2.19839000e-01], [2.40709086e-02, -1.45102446e-01, 1.11431545e-02, 3.85543001e-02, -8.93515563e-02, 5.61767721e-02, -6.35360730e-02, -8.23443779e-01, 3.54559731e-01, -2.81593679e-02, -3.92640266e-02, 2.32224316e-02, 1.64850420e-02, -1.10262122e-02, 1.82660654e-01, 3.25982295e-01, 1.22106697e-01], [5.95830975e-01, 2.92642398e-01, -4.44638207e-01, 1.02303616e-03, 2.18838802e-02, -5.23622267e-01, 1.25997650e-01, -1.41856014e-01, -6.97485854e-02, 1.14379958e-02, 3.94547417e-02, 1.27696382e-01, -5.83134662e-02, -1.77152700e-02, 1.04088088e-01, -9.37464497e-02, -6.91969778e-02], [8.06328039e-02, 3.34674281e-02, -8.56967180e-02, -1.07828189e-01, 1.51742110e-01, -5.63728817e-02, 1.92857500e-02, -3.40115407e-02, -5.84289756e-02, -6.68494643e-02, 2.75286207e-02, -6.91126145e-01, 6.71008607e-01, 4.13740967e-02, -2.71542091e-02, 7.31225166e-02, 3.64767385e-02], [1.33405806e-01, -1.45497511e-01, 2.95896092e-02, 6.97722522e-01, -6.17274818e-01, 9.91640992e-03, 2.09515982e-02, 3.83544794e-02, 3.40197083e-03, -9.43887925e-03, -3.09001353e-03, -1.12055599e-01, 1.58909651e-01, -2.08991284e-02, -8.41789410e-03, -2.27742017e-01, -3.39433604e-03], [4.59139498e-01, -5.18568789e-01, -4.04318439e-01, -1.48738723e-01, 5.18683400e-02, 5.60363054e-01, -5.27313042e-02, 1.01594830e-01, -2.59293381e-02, 2.88282896e-03, -1.28904022e-02, 2.98075465e-02, -2.70759809e-02, -2.12476294e-02, 3.33406243e-03, -4.38803230e-02, -5.00844705e-03], [3.58970400e-01, -5.43427250e-01, 6.09651110e-01, -1.44986329e-01, 8.03478445e-02, -4.14705279e-01, 9.01788964e-03, 5.08995918e-02, 1.14639620e-03, 7.72631963e-04, -1.11433396e-03, 1.38133366e-02, 6.20932749e-03, -2.22215182e-03, -1.91869743e-02, -3.53098218e-02, -1.30710024e-02]])

The Eigen Values,

Array ([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,

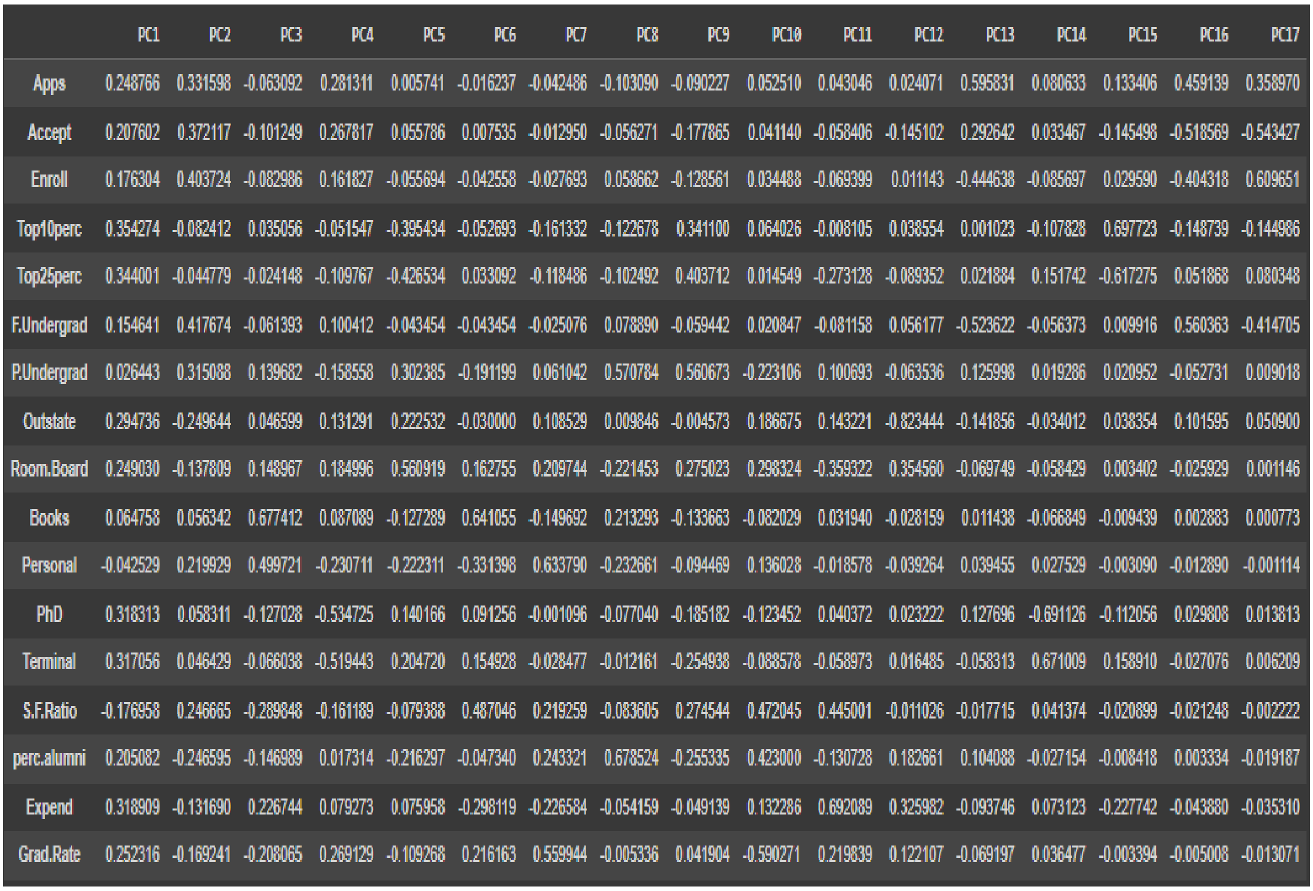
0.84849117, 0.6057878, 0.58787222, 0.53061262, 0.4043029,

0.31344588, 0.22061096, 0.16779415, 0.1439785, 0.08802464,

0.03672545, 0.02302787])

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

2.6.1. PCA data frame with original features,



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.103090	-0.090227	0.052510	0.043046	0.024071	0.595831	0.080633	0.133406	0.459139	0.358970
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950	-0.056271	-0.177865	0.041140	-0.058406	-0.145102	0.292642	0.033467	-0.145498	-0.518569	-0.543427
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693	0.058662	-0.128561	0.034488	-0.069399	0.011143	-0.444638	-0.085697	0.029590	-0.404318	0.609651
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678	0.341100	0.064026	-0.008105	0.038554	0.001023	-0.107828	0.697723	-0.148739	-0.144986
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486	-0.102492	0.403712	0.014549	-0.273128	-0.089352	0.021884	0.151742	-0.617275	0.051868	0.080348
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076	0.078890	-0.059442	0.020847	-0.081158	0.056177	-0.523622	-0.056373	0.009916	0.560363	-0.414705
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042	0.570784	0.560673	-0.223106	0.100693	-0.063536	0.125998	0.019286	0.020952	-0.052731	0.009018
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529	0.009846	-0.004573	0.186675	0.143221	-0.823444	-0.141856	-0.034012	0.038354	0.101595	0.050900
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744	-0.221453	0.275023	0.298324	-0.359322	0.354560	-0.069749	-0.058429	0.003402	-0.025929	0.001146
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692	0.213293	-0.133663	-0.082029	0.031940	-0.028159	0.011438	-0.066849	-0.009439	0.002883	0.000773
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790	-0.232661	-0.094469	0.136028	-0.018578	-0.039264	0.039455	0.027529	-0.003090	-0.012890	-0.001114
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096	-0.077040	-0.185182	-0.123452	0.040372	0.023222	0.127696	-0.691126	-0.112056	0.029808	0.013813
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477	-0.012161	-0.254938	-0.088578	-0.058973	0.016485	-0.058313	0.671009	0.158910	-0.027076	0.006209
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	0.219259	-0.083605	0.274544	0.472045	0.445001	-0.011026	-0.017715	0.041374	-0.020899	-0.021248	-0.002222
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321	0.678524	-0.255335	0.423000	-0.130728	0.182661	0.104088	-0.027154	-0.008418	0.003334	-0.019187
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584	-0.054159	-0.049139	0.132286	0.692089	0.325982	-0.093746	0.073123	-0.227742	-0.043880	-0.035310
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944	-0.005336	0.041904	-0.590271	0.219839	0.122107	-0.069197	0.036477	-0.003394	-0.005008	-0.013071

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 equation of 1st PC:

0.25 * Apps + 0.21 * Accept + 0.18 * Enrol + 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 +

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?

```
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.])
```

Cumulative values of eigen values are used to identify the optimum number of PC, if we look at the first 5 or 6 PC the valuable 76% to 81% information in our entire data is captured. Therefore, we can drop the remaining PC. If we add all the 17 PC, we will get 100%

Based on this we can select the first 5 PC which covers the 76% of valuable information. The eigen vectors indicate the weightage of the variables in the PC.

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]

The PCA helps us to understand the data given in our dataset. Initially based on the univariate and multivariate analysis we can visualize our data and understand the distribution and the skew nature of it. Further we also understood how the data values are correlated and how strong or weak is that correlation present in our dataset.

After understanding the data PCA is performed, and valuable principal components were identified. In our data we found first five PC to be valuable and can be used for further analysis.

2.9.1. Data Dictionary

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