ADVANCED STATISTICS

Report by

Table of Contents

INTRODUCTION

This report includes a detailed explanation of the approach taken, inferences, and insights addressing all three problems. It includes outputs such as graphs, tables, and all other relevant information. This Report does not include any codes.

Cases Covered

SALARY DATA ANALYSIS

EDUCATION DATA PCA

PROBLEM - 1
SUMMARY

SALARY DATA ANALYSIS

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

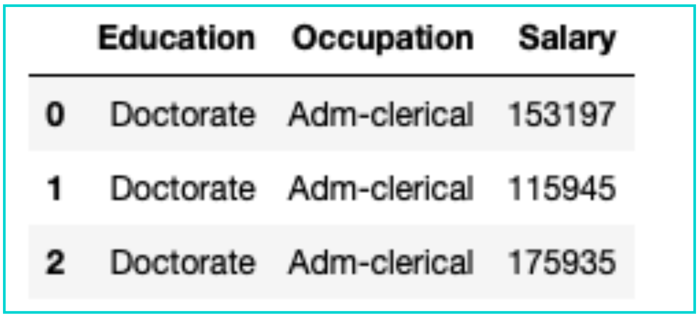
Assumption : The data follows a normal distribution.

Exploratory Data Analysis

Data Description

1. Education : Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate.
2. Occupation : Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial.
3. Salary : Salaries of different number of observations are in each level of education - occupation combination.

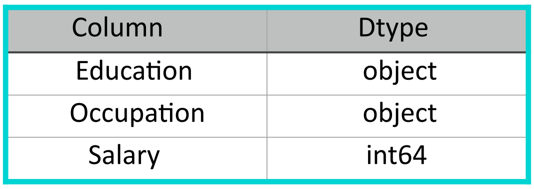
Sample of the dataset



	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935

Dataset Sample

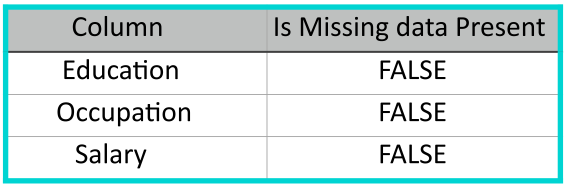
Data types of different variable



Column	Dtype
Education	object
Occupation	object
Salary	int64

Data Information

Missing data analysis for the variables



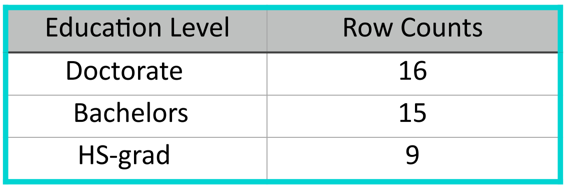
Column	Is Missing data Present
Education	FALSE
Occupation	FALSE
Salary	FALSE

Missing Data Analysis

Inference

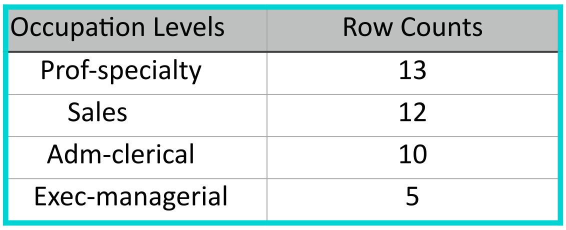
- Dataset has a total of 40 rows and 3 columns.
- Out of that 2 are objects and 1 is int type.
- There were no missing data in any of the rows.

Categorical Feature Analysis



Education Level	Row Counts
Doctorate	16
Bachelors	15
HS-grad	9

Education Level Analysis



Occupation Levels	Row Counts
Prof-specialty	13
Sales	12
Adm-clerical	10
Exec-managerial	5

Occupation Level Analysis

PROBLEMS 1A

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

Hypothesis for Education Levels¶

Null Hypothesis

All population mean of Salary is equal for all the levels of Educations i.e. Mean Salary of Doctorate = Mean Salary of Bachelors = Mean Salary of HS-grad

Alternative Hypothesis

Not all population mean of Salary is equal. At least one pair of population mean is not equal for all the levels of Education

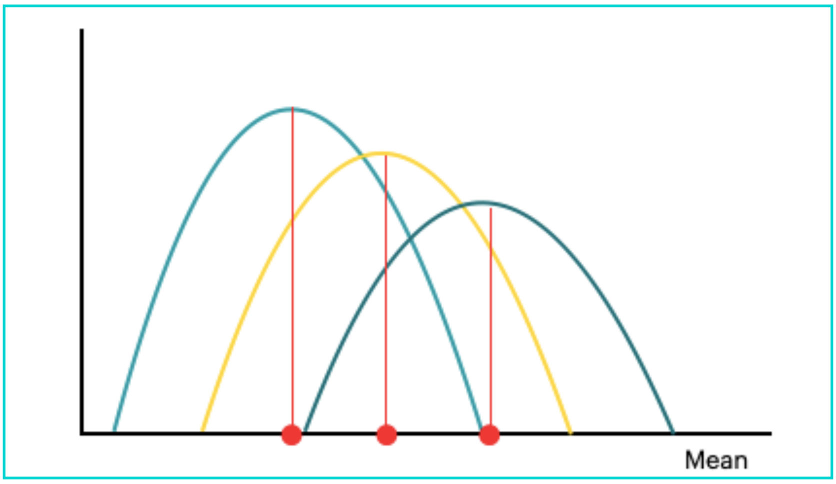
Hypothesis for Occupation Levels¶

Null Hypothesis

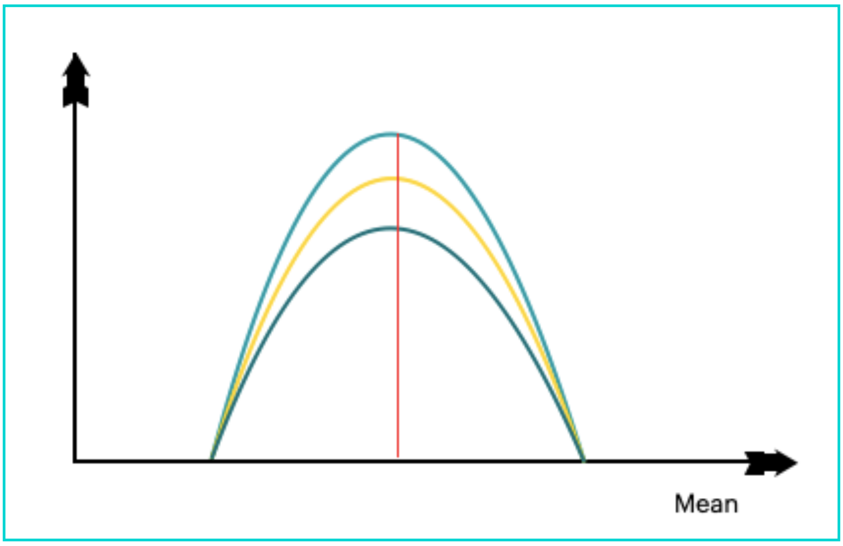
All population mean of Salary is equal for all the levels of Occupations i.e. Mean Salary of Prof-specialty = Mean Salary of Sales = Mean Salary of Adm-clerical = Mean Salary of Exec-managerial

Alternative Hypothesis

Not all population mean of Salary is equal. At least one pair of population mean is not equal for all the levels of Occupations



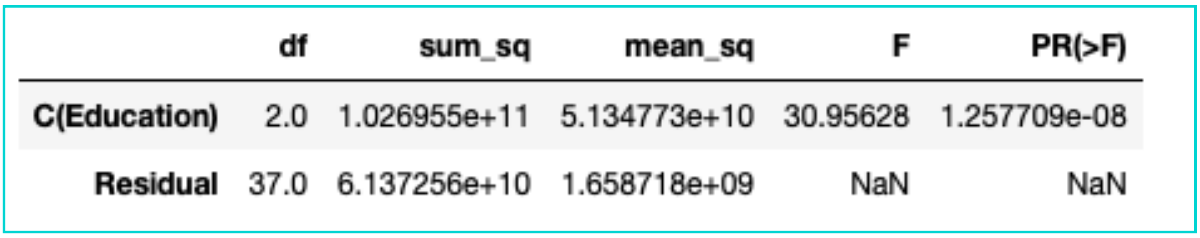
Null Hypothesis



Alternative Hypothesis

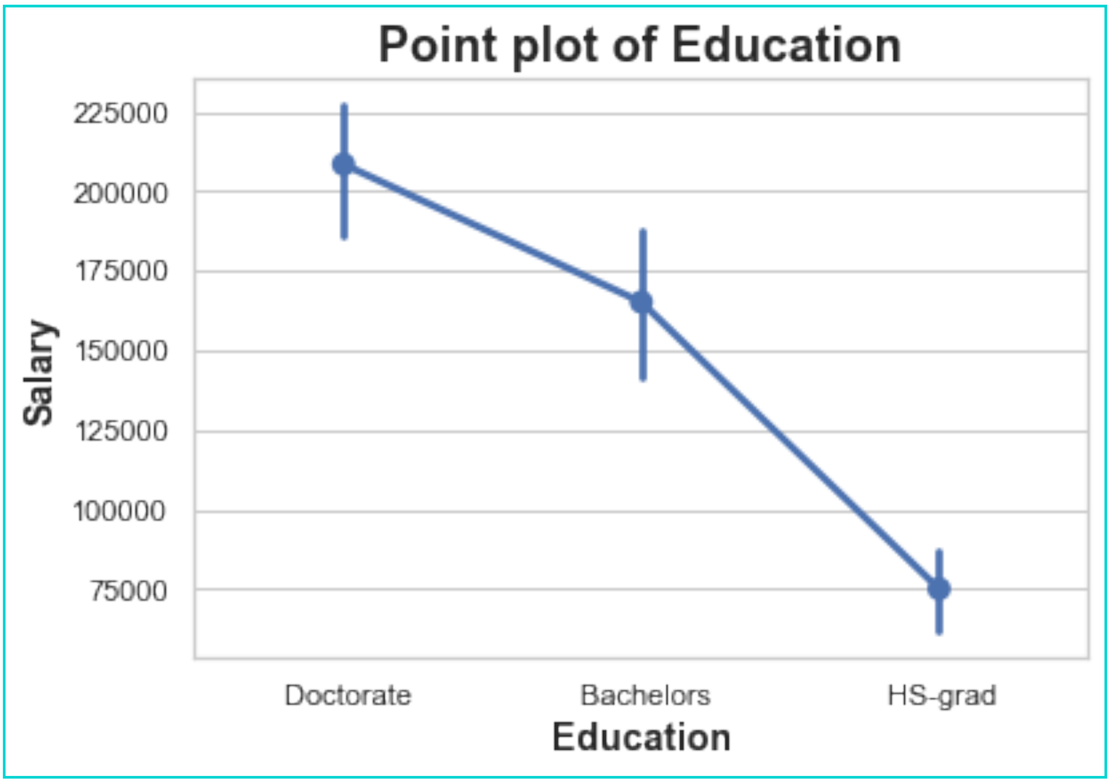
II. Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

ANOVA Table



	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

Anova table for Education

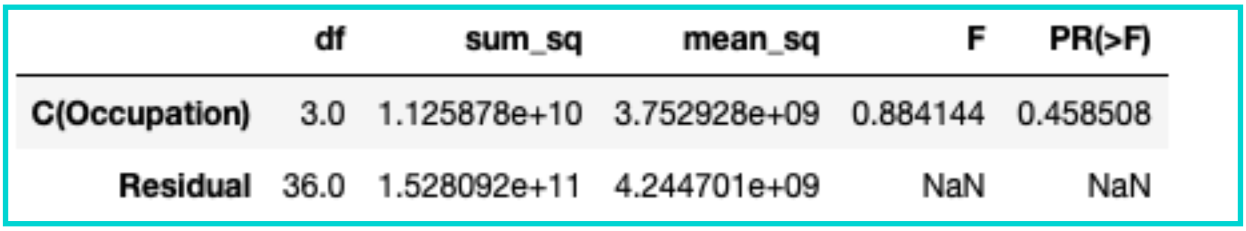


Inference

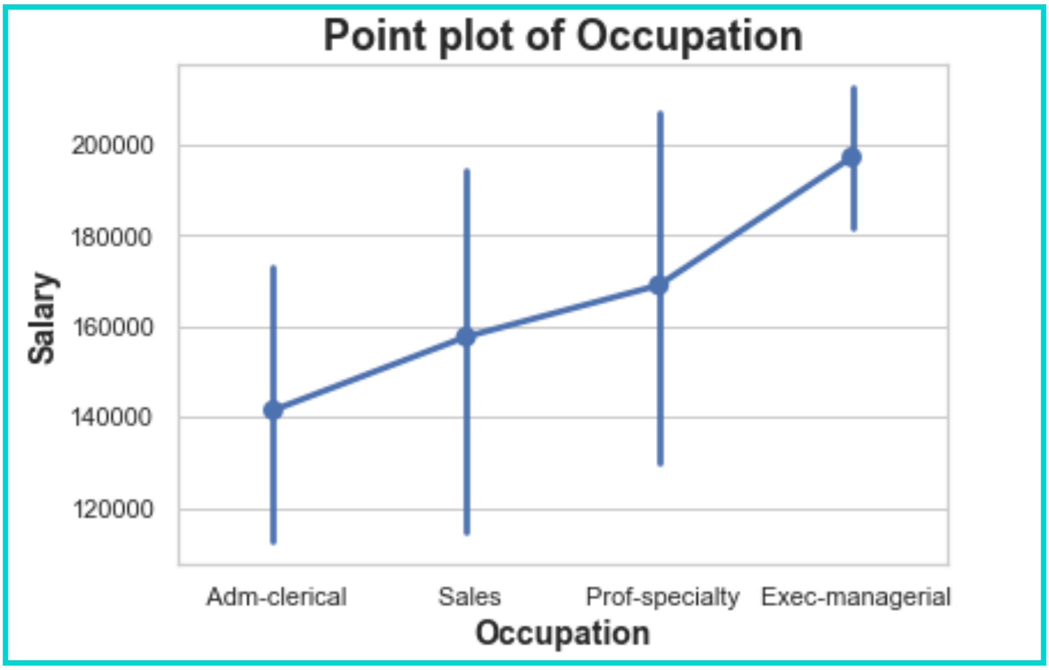
Here, we assumed that Alpha = 0.05. As P-Value is less than Alpha, hence we don't have enough evidence to accept the Null hypothesis.

III. 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.

ANOVA Table



	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



Point-plot of Salary vs Occupation

Inference

Here, we assumed that Alpha = 0.05. As P-Value is higher than Alpha, hence we have enough evidence to accept the Null hypothesis.

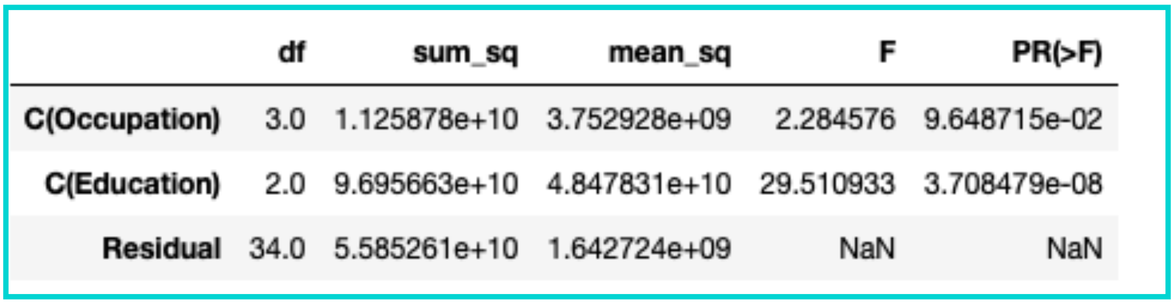
IV. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result.

The Null hypothesis is rejected when we did the ANOVA on salary means based on Education levels. Null hypothesis states here the salary means for all the levels of education is same. But since it's rejected, hence it means at least one pair of the Salary means based on different Education levels is not equal.

PROBLEMS 1B

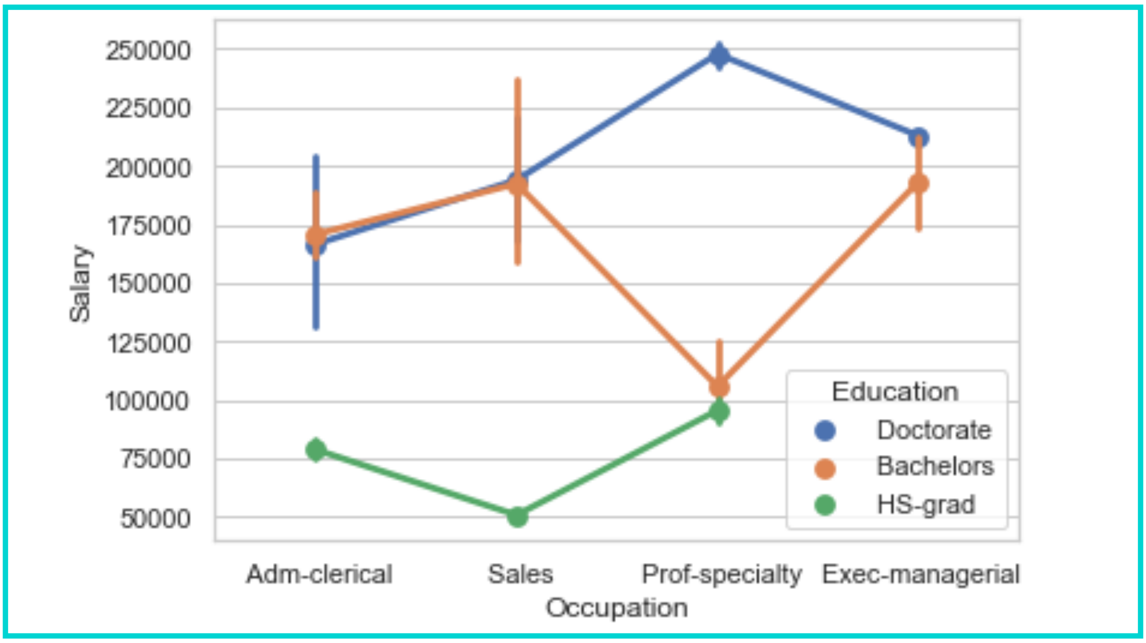
I. What is the interaction between two treatments? Analyse the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the ‘point-plot’ function from the ‘sea-born’ function]

The interaction between two treatments : Education and Occupation, can be observed using Two-Way ANOVA table and point-plot.



	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	2.284576	9.648715e-02
C(Education)	2.0	9.695663e+10	4.847831e+10	29.510933	3.708479e-08
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

Two-Way ANOVA table without Interaction

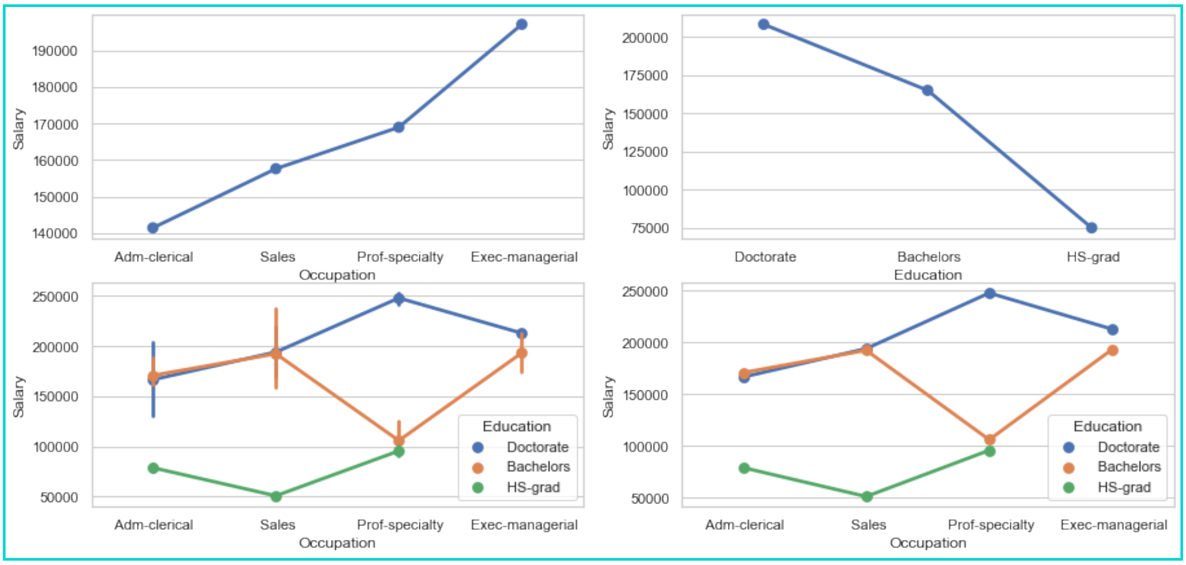


Point-plot with CI

Inference¶

1. Doctorate salary mean is approximately same as Bachelors for Smd-Clerical and Sales professionals.
2. Hs-grad and Bachelors salary mean is almost equal for Prof-Specialty professionals.
3. Hs-Grad are not working as Exec-managerial.
4. Exec-managerial's mean salaries are approximately close for Doctorate and Bachelors.
5. Salary mean trends are opposite for Prof-specialities who are Doctorate and Bachelors.

This graph shoes some relationships between the Education levels and Occupation levels interactions of salary means.



II. 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?

The Null hypothesis and the Alternate hypothesis is as follows:

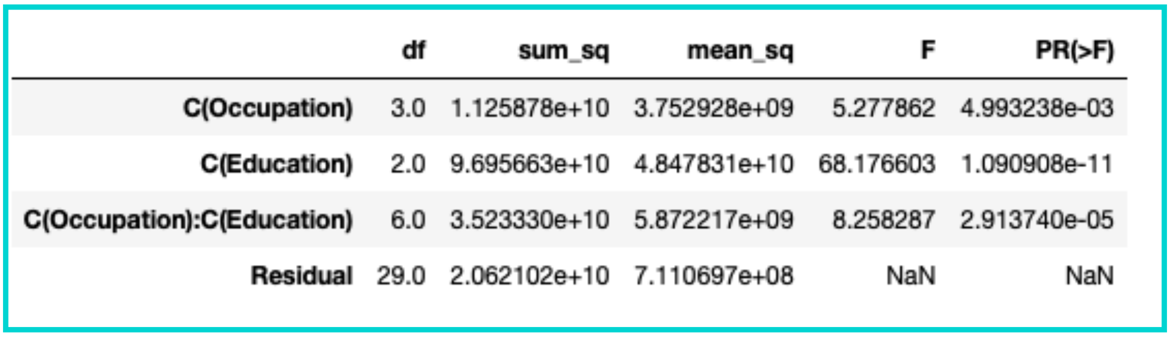
Null Hypothesis

1. All population mean of Salary is equal for all the levels of Educations i.e. Mean Salary of Doctorate = Mean Salary of Bachelors = Mean Salary of HS-grad
2. All population mean of Salary is equal for all the levels of Occupations i.e. Mean Salary of Prof-specialty = Mean Salary of Sales = Mean Salary of Adm-clerical = Mean Salary of Exec-managerial

Alternative Hypothesis

1. Not all population mean of Salary is equal. At least one pair of population mean is not equal for all the levels of Education.
2. Not all population mean of Salary is equal. At least one pair of population mean is not equal for all the levels of Occupations

Two-Way ANOVA

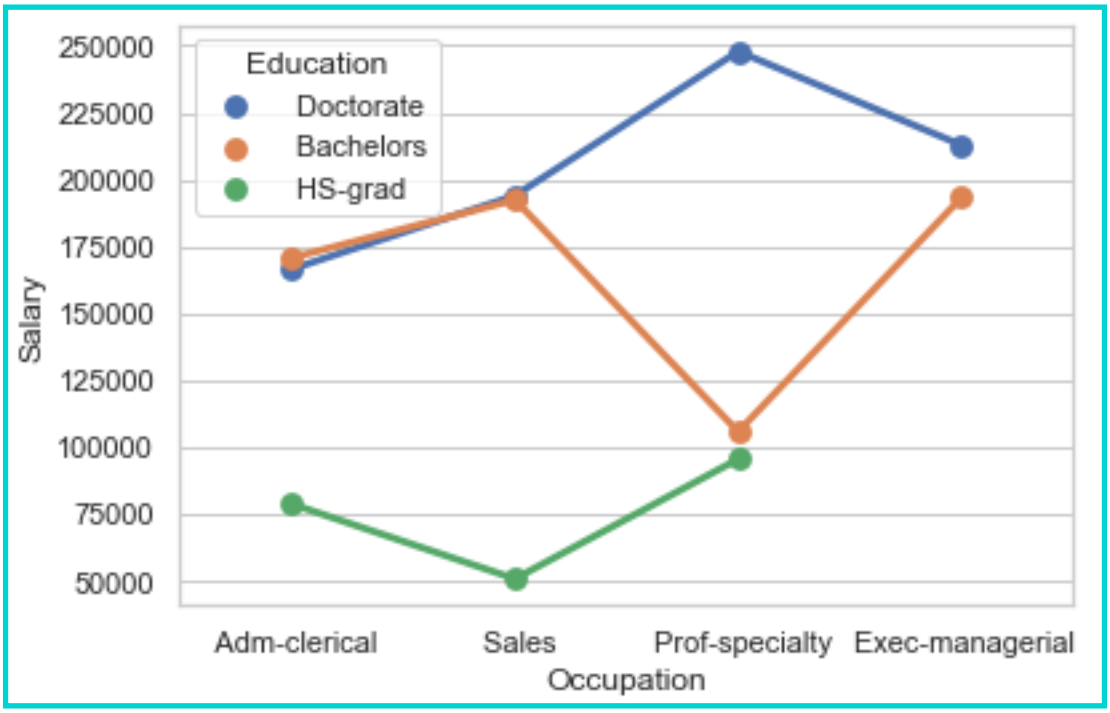


	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	5.277862	4.993238e-03
C(Education)	2.0	9.695663e+10	4.847831e+10	68.176603	1.090908e-11
C(Occupation):C(Education)	6.0	3.523330e+10	5.872217e+09	8.258287	2.913740e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Two-way table with Interaction

The two-way ANOVA based on Salary with respect to both Education and Occupation, can be observed using Two-Way ANOVA table.

Interaction Plot



point-Plot without CI

Interpretation

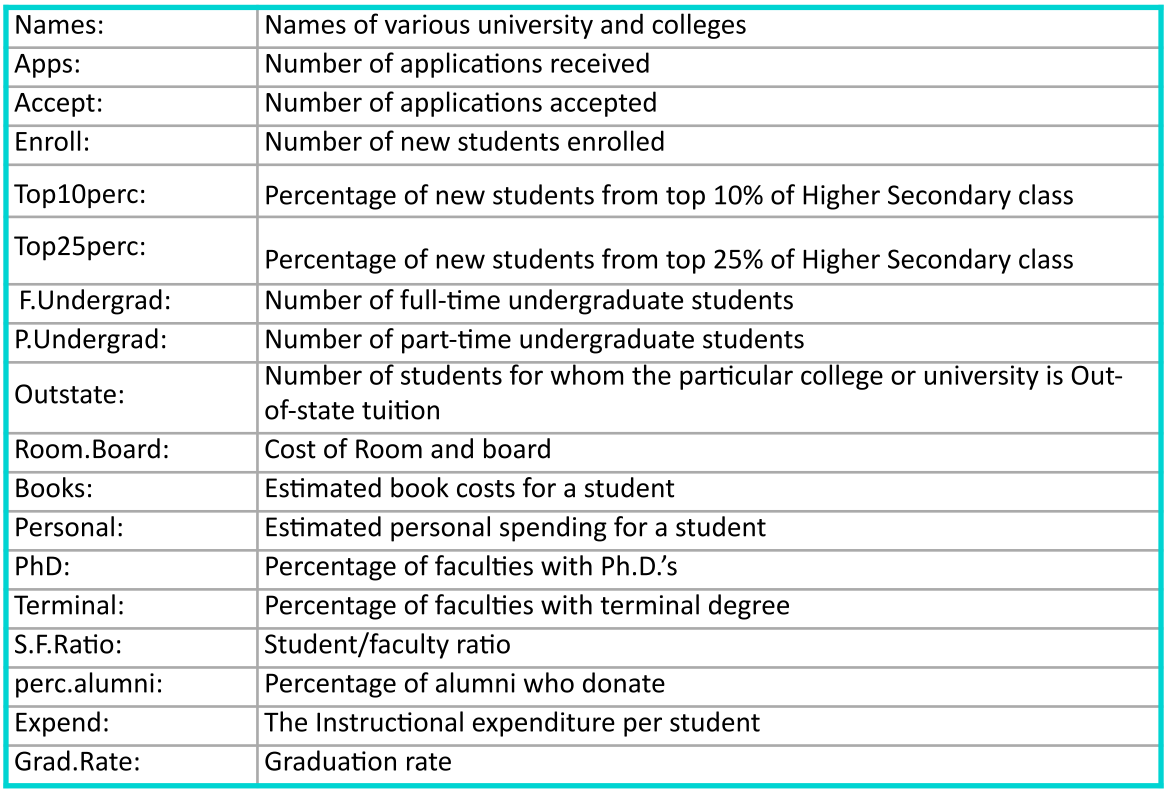
- **ANOVA based on Salary w.r.t Occupation :** As the P-Value is low, hence we do not have enough evidence to prove the Null hypothesis.
- **ANOVA based on Salary w.r.t Education :** As the P-Value is low, hence we do not have enough evidence to prove the Null hypothesis.
- **ANOVA based on Salary w.r.t Interactions of Education and Occupation :** As the P-Value is low, hence we do not have enough evidence to prove the Null hypothesis.

PROBLEM - 2 SUMMARY

EDUCATION DATA ANALYSIS

The dataset contains information on various colleges.It contains ***Names, applications, applications accepted, Students enrolled, Top10, Top25, Full time graduates, Part Time graduates, no. of students with Out of State tuition, Cost of room and board,Book cost for students, Personal spending of the students, Faculty percent of students, Faculty Percentage with Terminal degrees, student to faculty ratio, Percentage of alumni who donate, The Instructional expenditure per student and Graduation rate.***
We will analyse the data deduce inferences.

Data Description

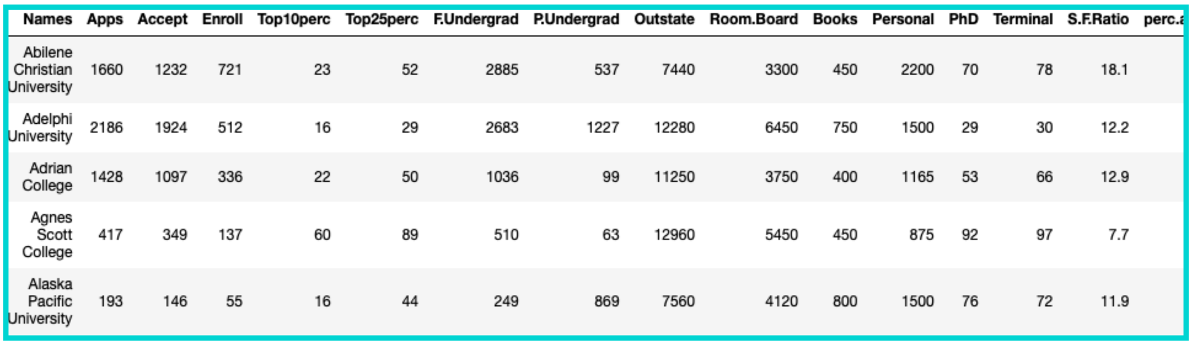


Names:	Names of various university and colleges
Apps:	Number of applications received
Accept:	Number of applications accepted
Enroll:	Number of new students enrolled
Top10perc:	Percentage of new students from top 10% of Higher Secondary class
Top25perc:	Percentage of new students from top 25% of Higher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

Data Description

Exploratory Data Analysis

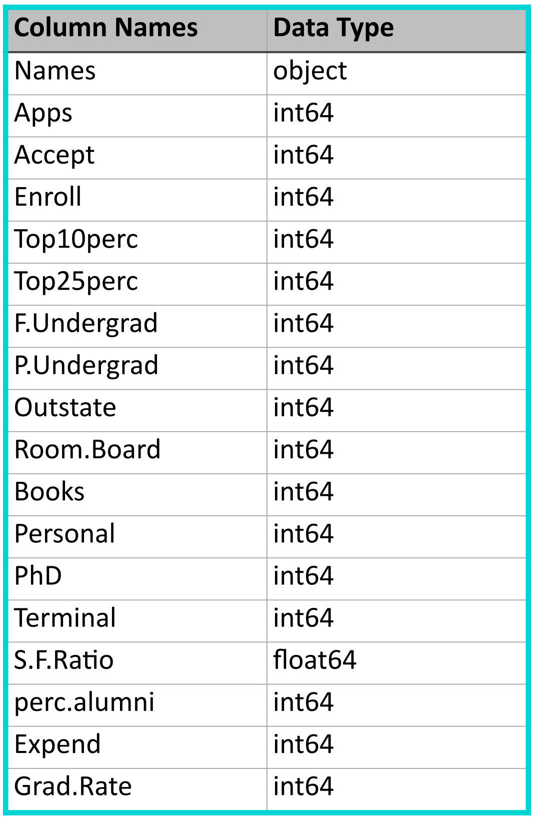
Sample of the dataset



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

Sample DataFrame

Data types of different variable



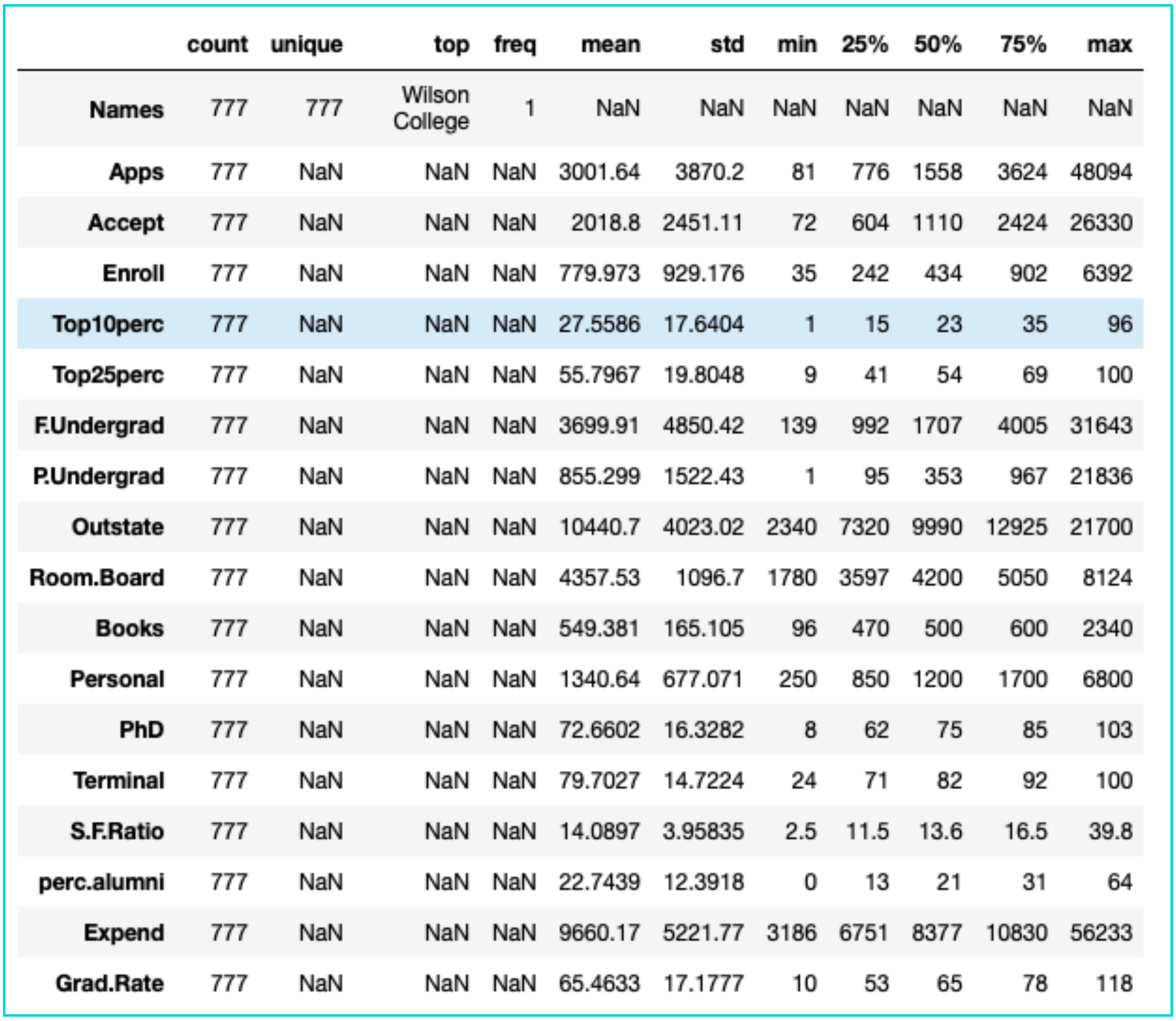
Column Names	Data Type
Names	object
Apps	int64
Accept	int64
Enroll	int64
Top10perc	int64
Top25perc	int64
F.Undergrad	int64
P.Undergrad	int64
Outstate	int64
Room.Board	int64
Books	int64
Personal	int64
PhD	int64
Terminal	int64
S.F.Ratio	float64
perc.alumni	int64
Expend	int64
Grad.Rate	int64

Data Types of fields

Details

- The Dataframe has 777 rows and 18 columns.
- Out of 18 Columns: 1 is Float, 1 is Object and 16 are Int.

Descriptive Data Analysis



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Names	777	777	Wilson College	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Apps	777	NaN	NaN	NaN	3001.64	3870.2	81	776	1558	3624	48094
Accept	777	NaN	NaN	NaN	2018.8	2451.11	72	604	1110	2424	26330
Enroll	777	NaN	NaN	NaN	779.973	929.176	35	242	434	902	6392
Top10perc	777	NaN	NaN	NaN	27.5586	17.6404	1	15	23	35	96
Top25perc	777	NaN	NaN	NaN	55.7967	19.8048	9	41	54	69	100
F.Undergrad	777	NaN	NaN	NaN	3699.91	4850.42	139	992	1707	4005	31643
P.Undergrad	777	NaN	NaN	NaN	855.299	1522.43	1	95	353	967	21836
Outstate	777	NaN	NaN	NaN	10440.7	4023.02	2340	7320	9990	12925	21700
Room.Board	777	NaN	NaN	NaN	4357.53	1096.7	1780	3597	4200	5050	8124
Books	777	NaN	NaN	NaN	549.381	165.105	96	470	500	600	2340
Personal	777	NaN	NaN	NaN	1340.64	677.071	250	850	1200	1700	6800
PhD	777	NaN	NaN	NaN	72.6602	16.3282	8	62	75	85	103
Terminal	777	NaN	NaN	NaN	79.7027	14.7224	24	71	82	92	100
S.F.Ratio	777	NaN	NaN	NaN	14.0897	3.95835	2.5	11.5	13.6	16.5	39.8
perc.alumni	777	NaN	NaN	NaN	22.7439	12.3918	0	13	21	31	64
Expend	777	NaN	NaN	NaN	9660.17	5221.77	3186	6751	8377	10830	56233
Grad.Rate	777	NaN	NaN	NaN	65.4633	17.1777	10	53	65	78	118

Inference

- Outliers can be observed in the fields: ***Apps, Accept, Enroll, F:Undergrad, P:Undergrad, Outstate, Books, Personal, Expend***.
- No null values are observed.
- All Names are Unique.

PROBLEMS

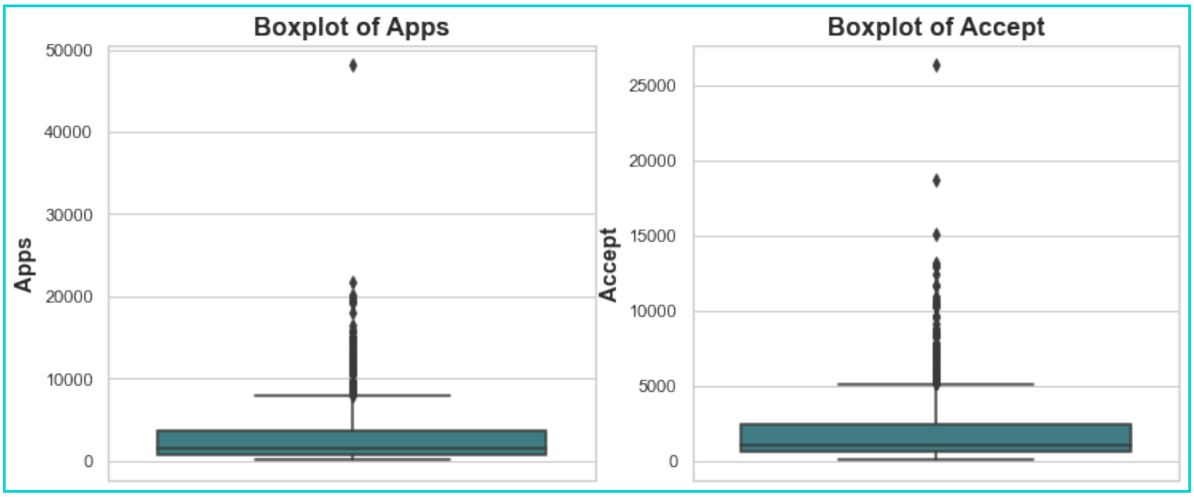
I. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

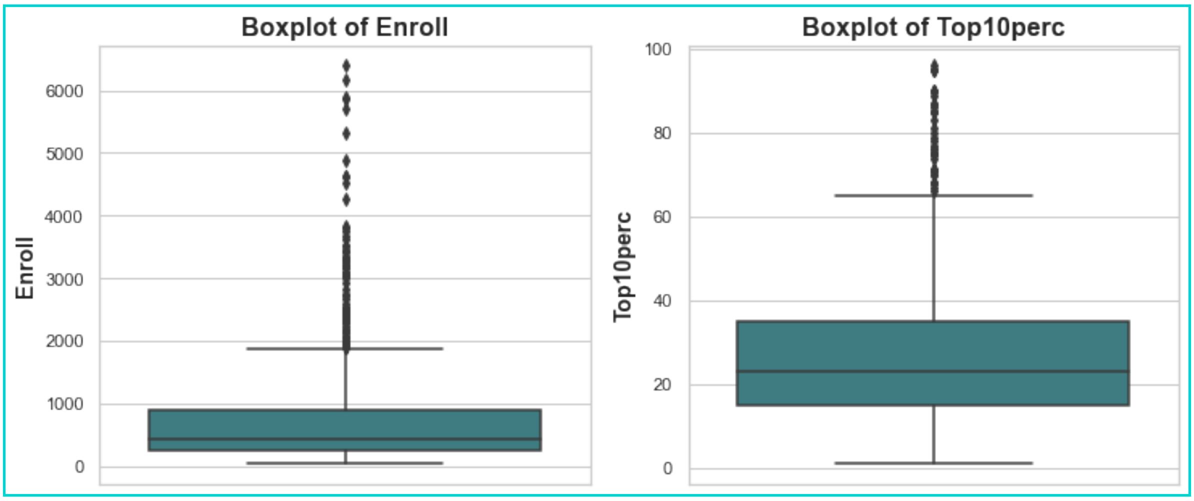
Step 1: Checking Duplicate rows

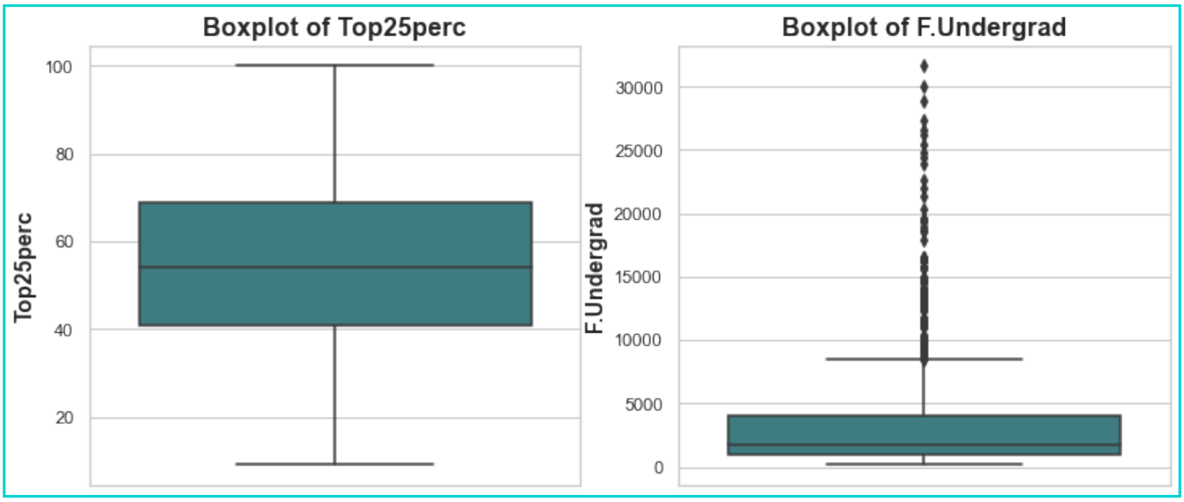
No duplicate rows present in the dataframe

Step 2: Checking Outliers

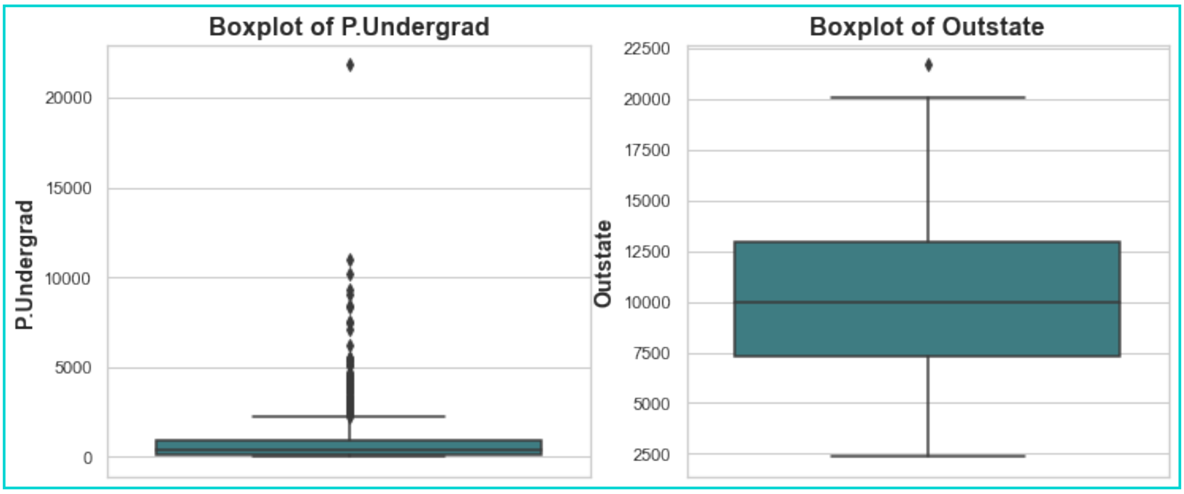
- Box-plots of different features for finding Outliers :



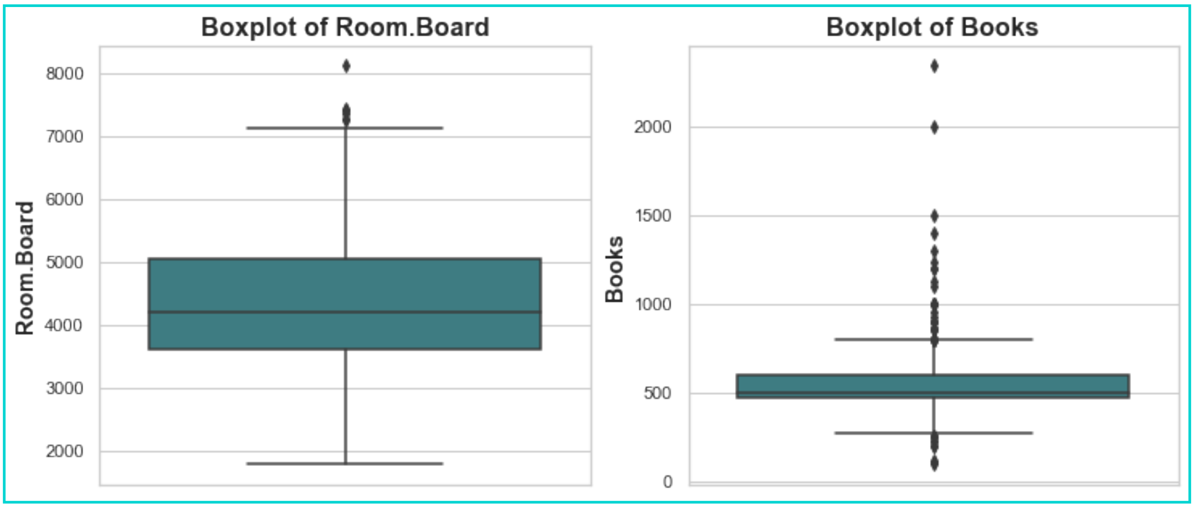



Boxplot of Top25perc

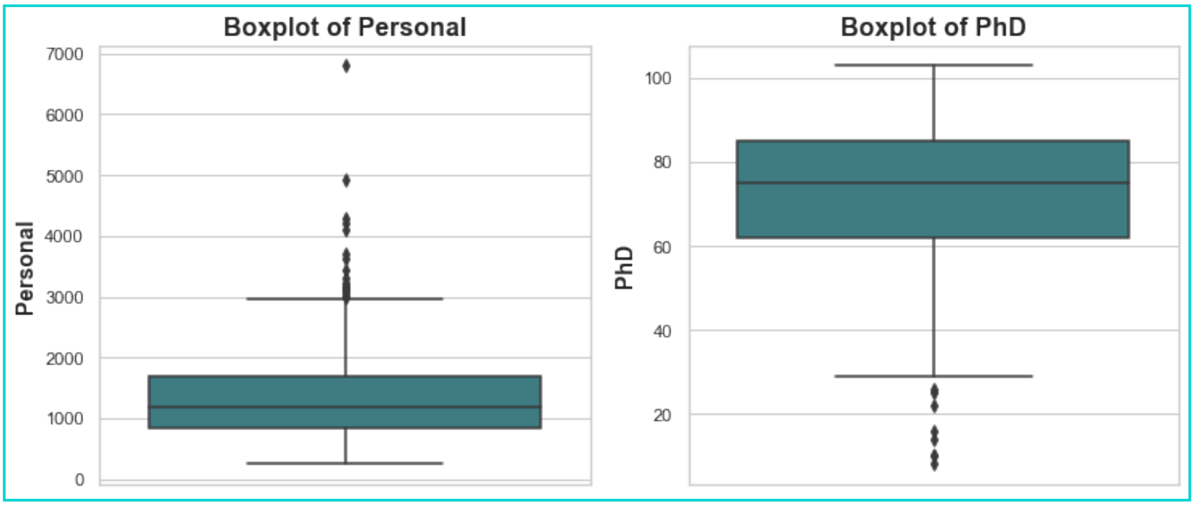
Boxplot of F.Undergrad

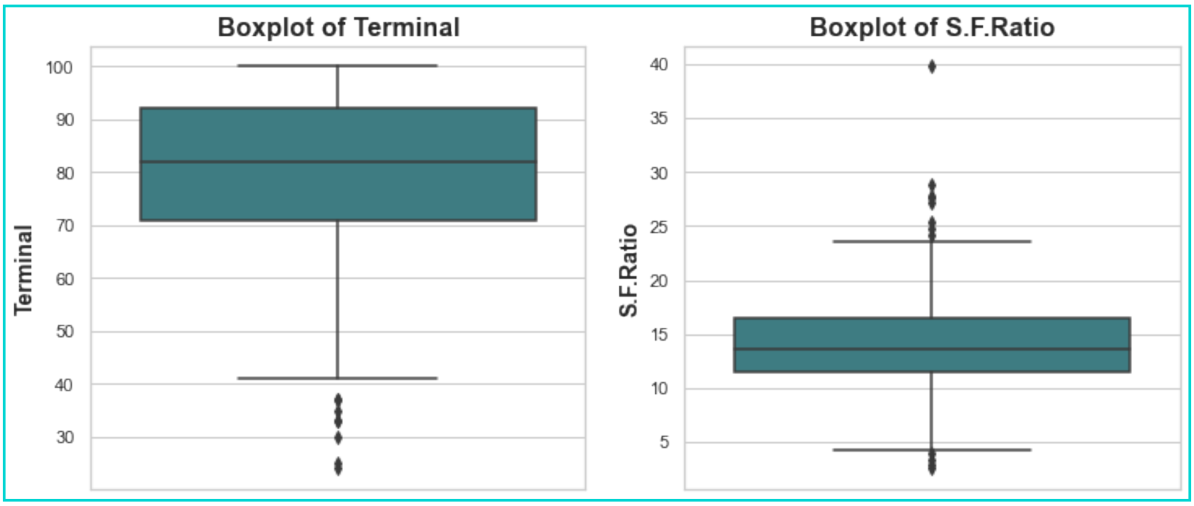

Boxplot of P.Undergrad

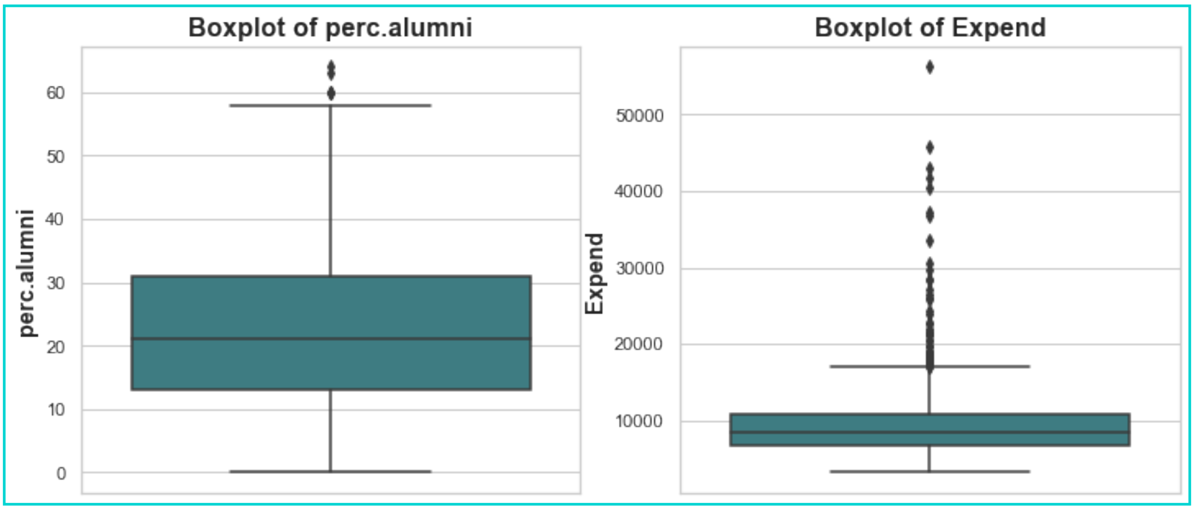
Boxplot of Outstate

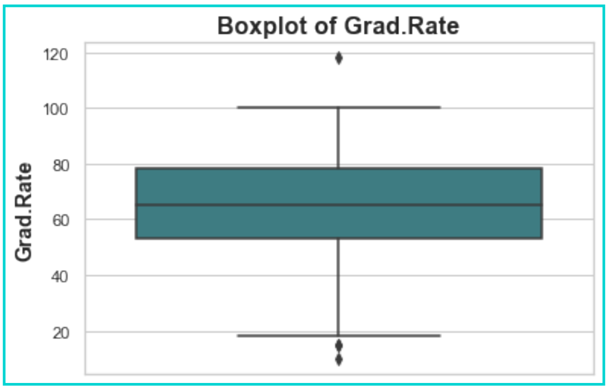

Boxplot of Room.Board

Boxplot of Books


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


Boxplot of Terminal
Terminal
100
90
80
70
60
50
40
30
Boxplot of S.F.Ratio
S.F.Ratio
40
35
30
25
20
15
10
5


Boxplot of perc.alumni
perc.alumni
60
50
40
30
20
10
0
Boxplot of Expend
Expend
50000
40000
30000
20000
10000

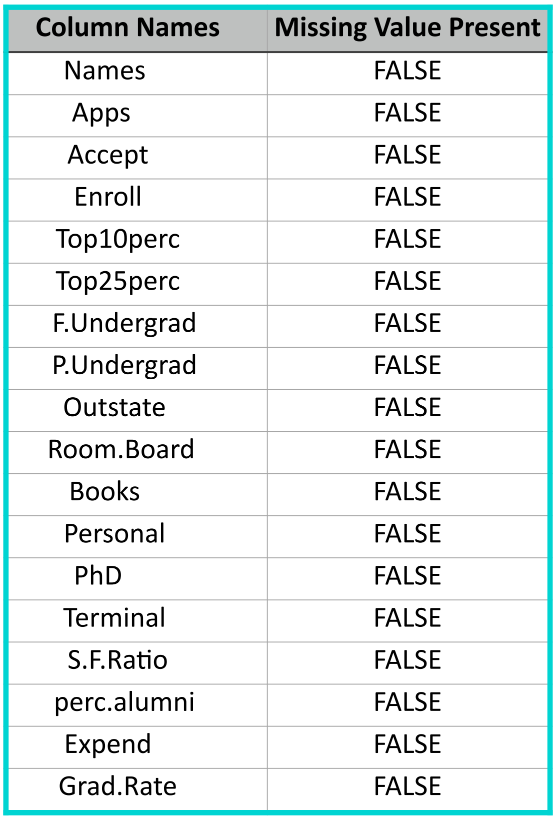


Inference from Box-plot

- Outliers are observed in all the features except ***Top25Perc***

Step 3: Checking for Missing Values

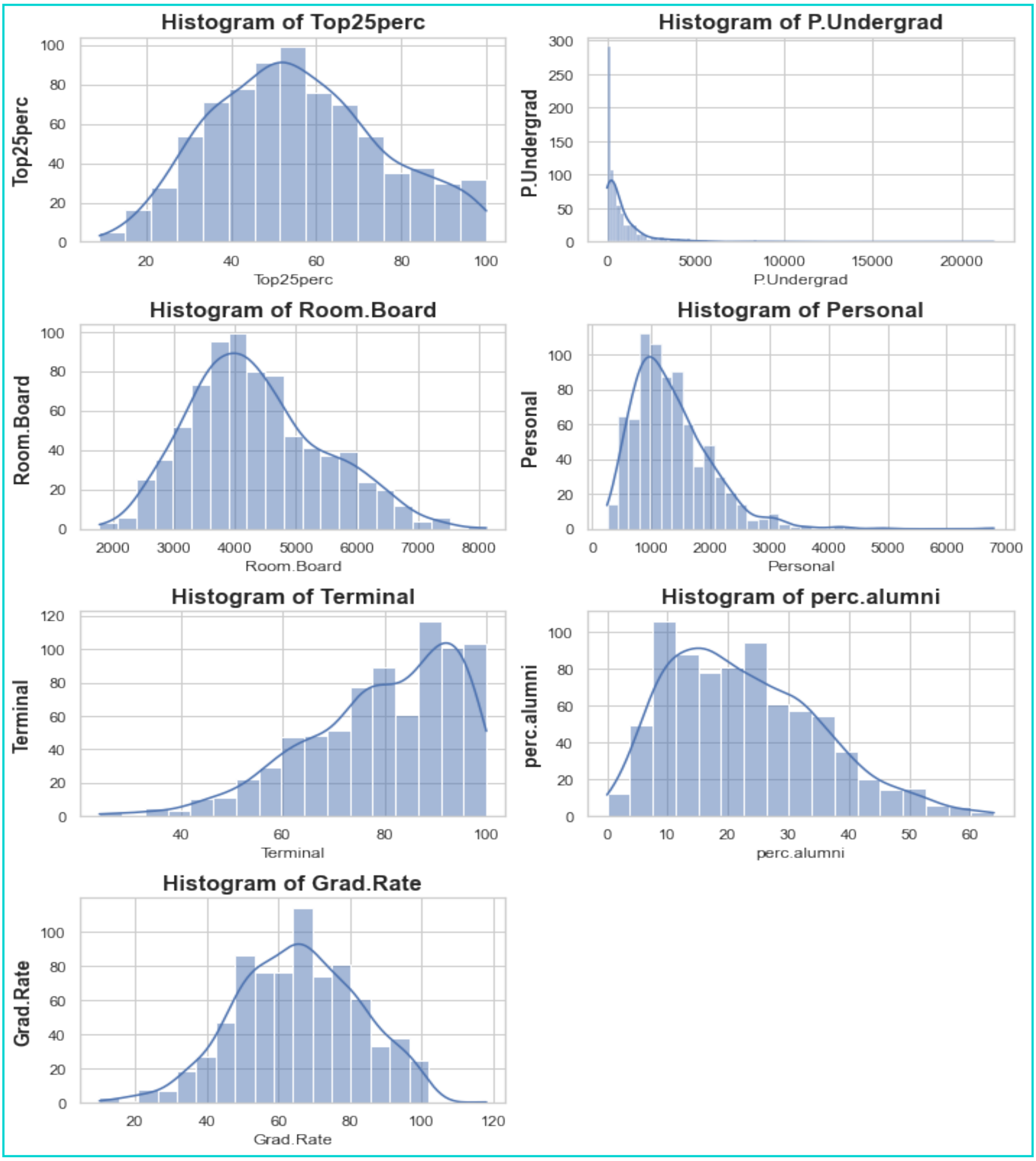
As per the observation, there is no missing value present.

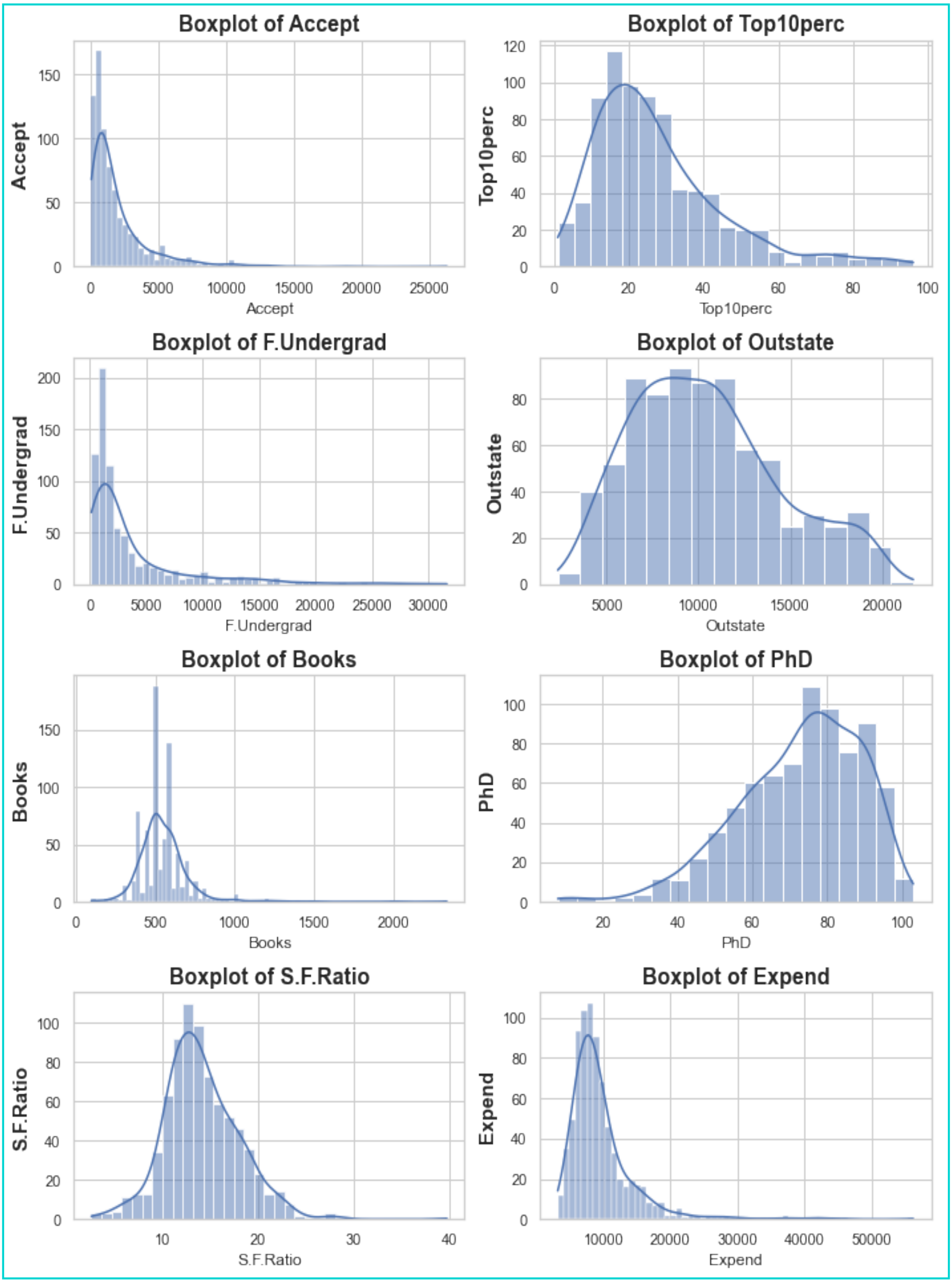


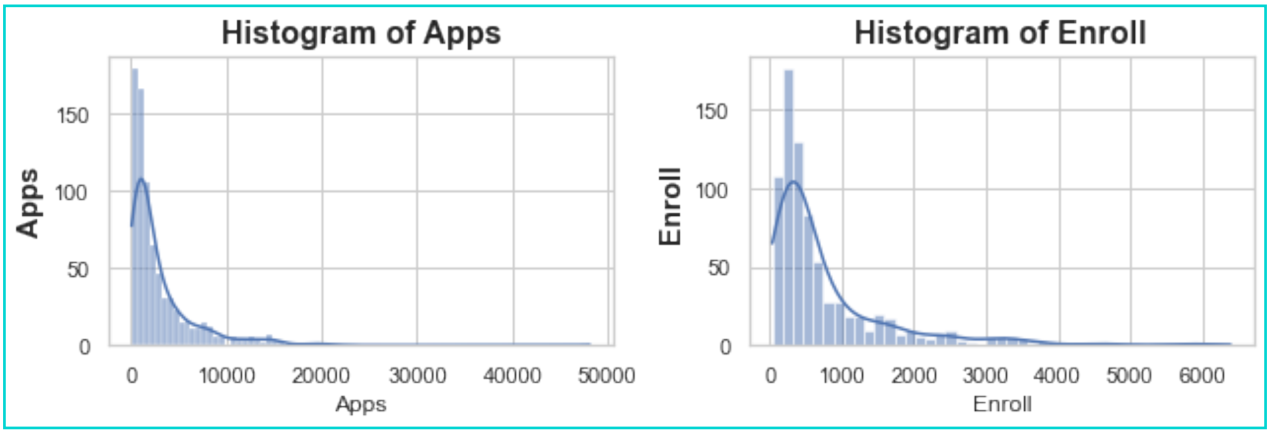
Column Names	Missing Value Present
Names	FALSE
Apps	FALSE
Accept	FALSE
Enroll	FALSE
Top10perc	FALSE
Top25perc	FALSE
F.Undergrad	FALSE
P.Undergrad	FALSE
Outstate	FALSE
Room.Board	FALSE
Books	FALSE
Personal	FALSE
PhD	FALSE
Terminal	FALSE
S.F.Ratio	FALSE
perc.alumni	FALSE
Expend	FALSE
Grad.Rate	FALSE

Step 4: Univariate Analysis

The histogram of all numerical fields are as follows:




Boxplot of Accept
Accept
0
50
100
150
0
5000
10000
15000
20000
25000
Accept
Boxplot of Top10perc
Top10perc
0
20
40
60
80
100
120
0
20
40
60
80
100
Top10perc
Boxplot of F.Undergrad
F.Undergrad
0
50
100
150
200
0
5000
10000
15000
20000
25000
30000
F.Undergrad
Boxplot of Outstate
Outstate
0
20
40
60
80
5000
10000
15000
20000
Outstate
Boxplot of Books
Books
0
50
100
150
0
500
1000
1500
2000
Books
Boxplot of PhD
PhD
0
20
40
60
80
100
20
40
60
80
100
PhD
Boxplot of S.F.Ratio
S.F.Ratio
0
20
40
60
80
100
10
20
30
40
S.F.Ratio
Boxplot of Expend
Expend
0
20
40
60
80
100
10000
20000
30000
40000
50000
Expend



Inference from Histogram

- *Apps, Enroll, Accept, Top10Perc, F.Undergrad, Expend, P.underGrad,* ***Personal*** seems to be ***Right Skewed.***
- *PHD* and *Terminal* seems to be ***Left-Skewed.***
- Rest are approximately normally distributed.
- We can also observe outliers in these graphs like Book, Enroll , etc.

<u>Skewness of the fields are as follows:</u>

Positive / Right Skewed

- Skewness of Apps is 3.7165574035202718
- Skewness of Accept is 3.4111258724395235
- Skewness of Enroll is 2.6852679191653412
- Skewness of Top10perc is 1.410487098842332
- Skewness of F.Undergrad is 2.6054157486361564
- Skewness of P.Undergrad is 5.681358169711681
- Skewness of Books is 3.478293278376379
- Skewness of Personal is 1.7391308384291781
- Skewness of S.F.Ratio is 0.6661461873546756
- Skewness of Expend is 3.4526399033472197
- Skewness of Outstate is 0.508294284359404
- Skewness of Room.Board is 0.4764335489968277

Negative / Left Skewed

- Skewness of PhD is -0.7666863621506335
- Skewness of Terminal is -0.8149651536781263
- Skewness of Grad.Rate is -0.11355752571272018

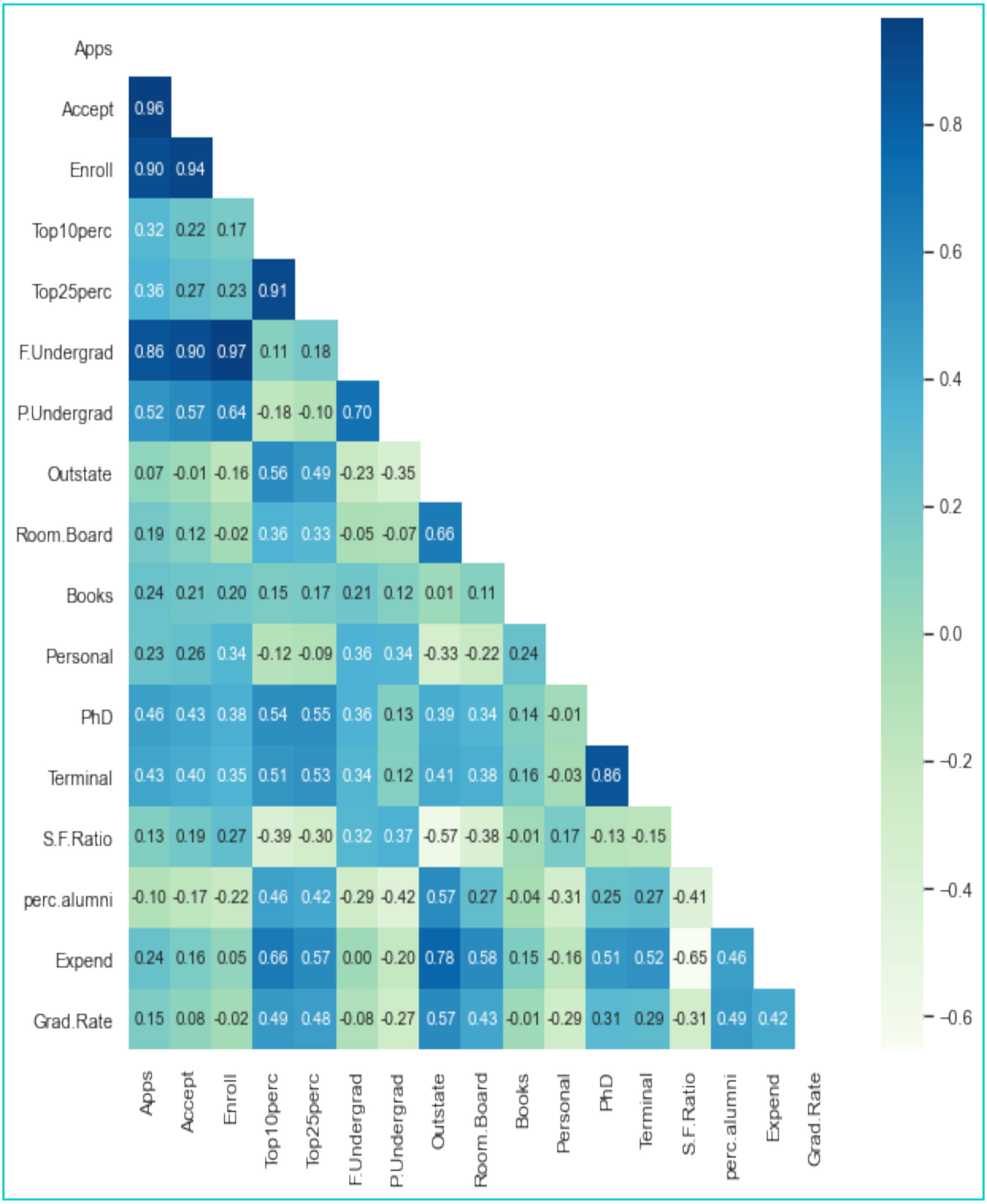
Approximately Normally distributed

- Skewness of Top25perc is 0.2588394269741162
- Skewness of perc.alumni is 0.6057189848601131

Step 5: Bivariate Analysis

For Bivariate analysis, we can find the Covariance and Correlation among various fields, using heat map.

HeatMap



	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																	
Accept	0.96																
Enroll	0.90	0.94															
Top10perc	0.32	0.22	0.17														
Top25perc	0.36	0.27	0.23	0.91													
F.Undergrad	0.86	0.90	0.97	0.11	0.18												
P.Undergrad	0.52	0.57	0.64	-0.18	-0.10	0.70											
Outstate	0.07	-0.01	-0.16	0.56	0.49	-0.23	-0.35										
Room.Board	0.19	0.12	-0.02	0.36	0.33	-0.05	-0.07	0.66									
Books	0.24	0.21	0.20	0.15	0.17	0.21	0.12	0.01	0.11								
Personal	0.23	0.26	0.34	-0.12	-0.09	0.36	0.34	-0.33	-0.22	0.24							
PhD	0.46	0.43	0.38	0.54	0.55	0.36	0.13	0.39	0.34	0.14	-0.01						
Terminal	0.43	0.40	0.35	0.51	0.53	0.34	0.12	0.41	0.38	0.16	-0.03	0.86					
S.F.Ratio	0.13	0.19	0.27	-0.39	-0.30	0.32	0.37	-0.57	-0.38	-0.01	0.17	-0.13	-0.15				
perc.alumni	-0.10	-0.17	-0.22	0.46	0.42	-0.29	-0.42	0.57	0.27	-0.04	-0.31	0.25	0.27	-0.41			
Expend	0.24	0.16	0.05	0.66	0.57	0.00	-0.20	0.78	0.58	0.15	-0.16	0.51	0.52	-0.65	0.46		
Grad.Rate	0.15	0.08	-0.02	0.49	0.48	-0.08	-0.27	0.57	0.43	-0.01	-0.29	0.31	0.29	-0.31	0.49	0.42	

Inference from Heat-Map

- Positive correlation is observed between various fields as : *Accept-Apps, Apps-Enroll, Enroll-Accept* etc.
- Negative Correlations can be observed in various fields like : *S.F.Ratio-Expend, Outstate-S.F.Ratio* etc

Inference from Pair-Plot

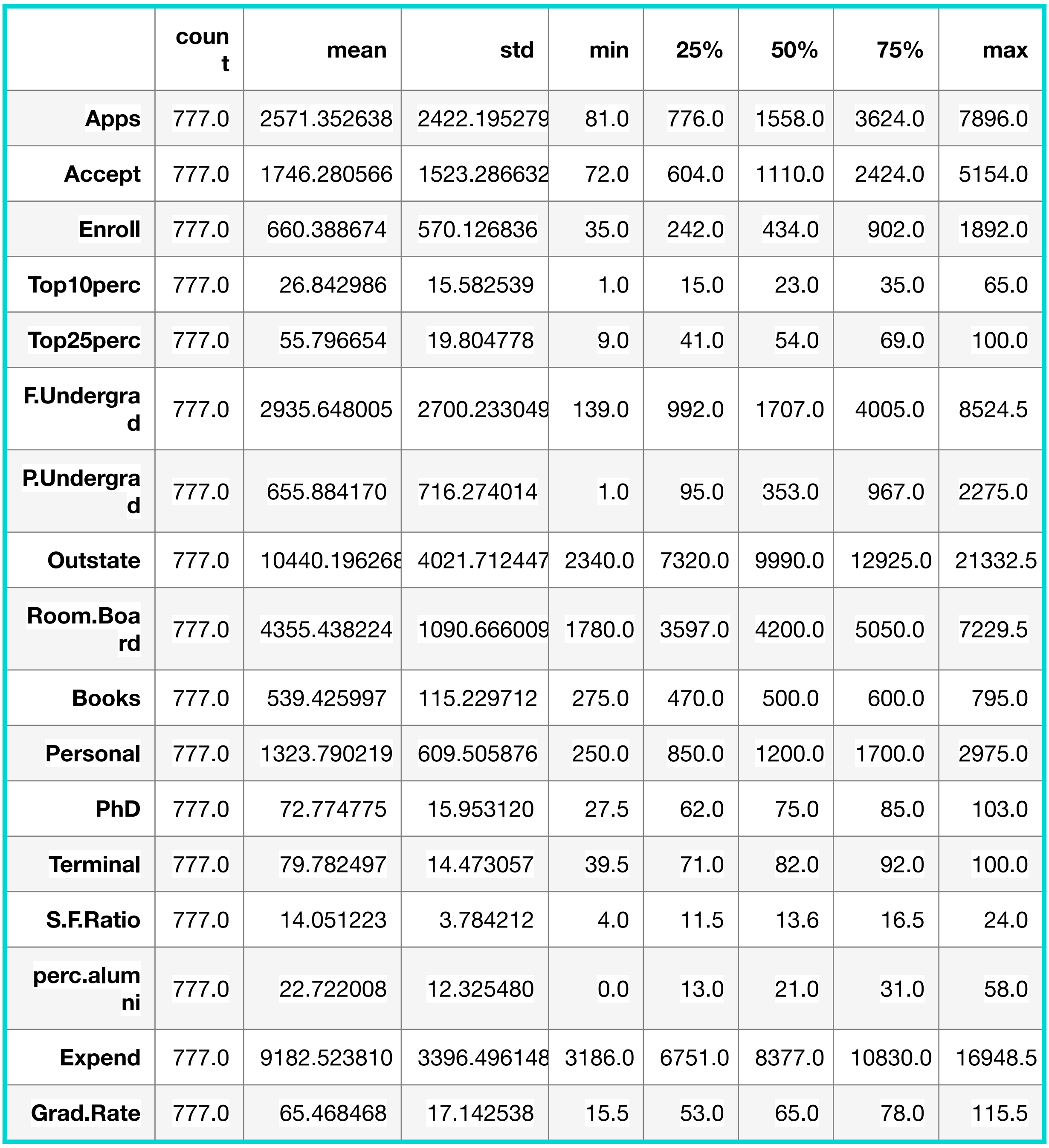
- Positive covariance is observed between various fields as: *Enroll-P.Undergrad, Accept-Enroll* , etc.
- Negative covariance can be observed in various fields like : *S.F.Ratio-Expend, Outstate-S.F.Ratio,* etc

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

To determine if Scaling is required or not, we need to find the summary of all the fields and check if all those are in approximately same range.

Descriptive Statistic

As observed in the above descriptive statistics, the range of fields vary from one another. Hence, we would require scaling for doing PCA.



	count	mean	std	min	25%	50%	75%	max
Apps	777.0	2571.352638	2422.195279	81.0	776.0	1558.0	3624.0	7896.0
Accept	777.0	1746.280566	1523.286632	72.0	604.0	1110.0	2424.0	5154.0
Enroll	777.0	660.388674	570.126836	35.0	242.0	434.0	902.0	1892.0
Top10perc	777.0	26.842986	15.582539	1.0	15.0	23.0	35.0	65.0
Top25perc	777.0	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	2935.648005	2700.233049	139.0	992.0	1707.0	4005.0	8524.5
P.Undergrad	777.0	655.884170	716.274014	1.0	95.0	353.0	967.0	2275.0
Outstate	777.0	10440.196268	4021.712447	2340.0	7320.0	9990.0	12925.0	21332.5
Room.Board	777.0	4355.438224	1090.666009	1780.0	3597.0	4200.0	5050.0	7229.5
Books	777.0	539.425997	115.229712	275.0	470.0	500.0	600.0	795.0
Personal	777.0	1323.790219	609.505876	250.0	850.0	1200.0	1700.0	2975.0
PhD	777.0	72.774775	15.953120	27.5	62.0	75.0	85.0	103.0
Terminal	777.0	79.782497	14.473057	39.5	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.051223	3.784212	4.0	11.5	13.6	16.5	24.0
perc.alumni	777.0	22.722008	12.325480	0.0	13.0	21.0	31.0	58.0
Expend	777.0	9182.523810	3396.496148	3186.0	6751.0	8377.0	10830.0	16948.5
Grad.Rate	777.0	65.468468	17.142538	15.5	53.0	65.0	78.0	115.5

Scaling

After performing scaling on the dataframe with the z-score method, we found the following summary which indicates the field values to be in similar range.

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	1.234534E-16	1.000644	-1.028801	-0.741686	-0.418631	0.434864	2.199689
Accept	777.0	1.340626E-16	1.000644	-1.099832	-0.750362	-0.417972	0.445193	2.238524
Enroll	777.0	1.521645E-16	1.000644	-1.097636	-0.734325	-0.397341	0.424058	2.161632
Top10perc	777.0	-2.250452E-18	1.000644	-1.659526	-0.760506	-0.246780	0.523809	2.450281
Top25perc	777.0	-1.546739E-16	1.000644	-2.364419	-0.747607	-0.090777	0.667104	2.233391
F.Undergrad	777.0	-1.911679E-16	1.000644	-1.036373	-0.720271	-0.455309	0.396277	2.071100
P.Undergrad	777.0	-9.573352E-17	1.000644	-0.914882	-0.783562	-0.423133	0.434633	2.261926
Outstate	777.0	-1.583175E-16	1.000644	-2.015414	-0.776337	-0.112014	0.618245	2.710119
Room.Board	777.0	-1.900382E-17	1.000644	-2.362866	-0.695838	-0.142609	0.637234	2.636841
Books	777.0	-4.465183E-16	1.000644	-2.296251	-0.602889	-0.342372	0.526019	2.219381
Personal	777.0	-9.605501E-17	1.000644	-1.762874	-0.777836	-0.203230	0.617635	2.710841
PhD	777.0	4.232636E-16	1.000644	-2.839817	-0.675837	0.139575	0.766815	1.895848
Terminal	777.0	2.460494E-16	1.000644	-2.785068	-0.607208	0.153315	0.844699	1.397806
S.F.Ratio	777.0	3.635016E-16	1.000644	-2.657805	-0.674610	-0.119315	0.647520	2.630716
perc.alumni	777.0	5.765444E-17	1.000644	-1.844686	-0.789281	-0.139801	0.672049	2.864044
Expend	777.0	1.148802E-16	1.000644	-1.766640	-0.716353	-0.237316	0.485364	2.287940
Grad.Rate	777.0	-2.743408E-16	1.000644	-2.916759	-0.727809	-0.027345	0.731490	2.920440

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

To find the covariance and the correlation matrices from this data, the matrices are as follows:

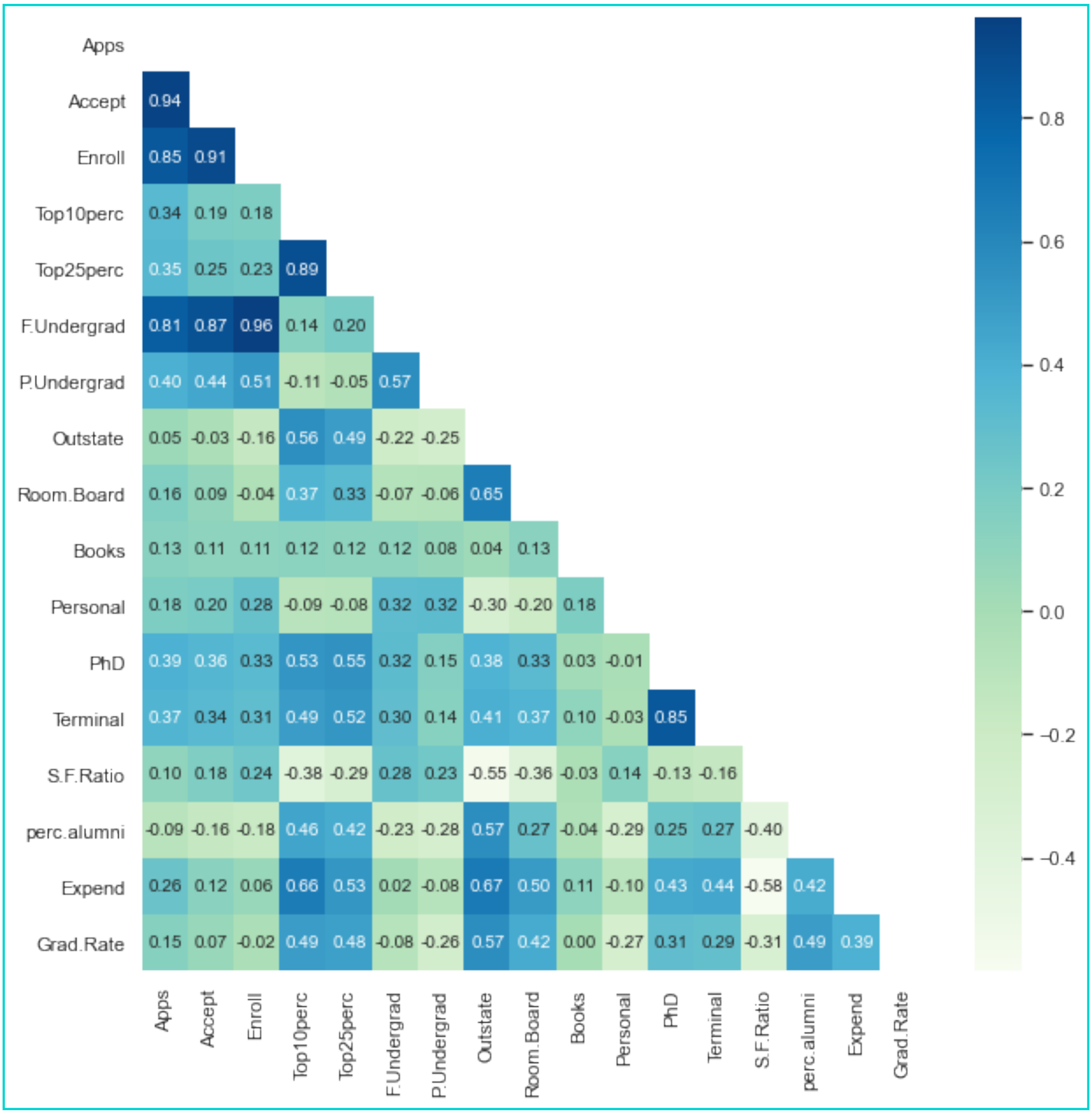
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate
Apps	1.000000	0.955307	0.896883	0.321342	0.364491	0.861002	0.519823	0.065337
Accept	0.955307	1.000000	0.935277	0.223298	0.273681	0.897034	0.572691	-0.005002
Enroll	0.896883	0.935277	1.000000	0.171756	0.230434	0.967302	0.641595	-0.155655
Top10perc	0.321342	0.223298	0.171756	1.000000	0.913875	0.111215	-0.180009	0.562160
Top25perc	0.364491	0.273681	0.230434	0.913875	1.000000	0.181196	-0.099295	0.489569
F.Undergrad	0.861002	0.897034	0.967302	0.111215	0.181196	1.000000	0.696130	-0.226166
P.Undergrad	0.519823	0.572691	0.641595	-0.180009	-0.099295	0.696130	1.000000	-0.354216
Outstate	0.065337	-0.005002	-0.155655	0.562160	0.489569	-0.226166	-0.354216	1.000000
Room.Board	0.187475	0.119586	-0.023846	0.357366	0.330987	-0.054476	-0.067638	0.655489
Books	0.236138	0.208705	0.202057	0.153452	0.169761	0.207879	0.122529	0.005110
Personal	0.229948	0.256346	0.339348	-0.116730	-0.086810	0.359783	0.344053	-0.325609
PhD	0.463924	0.427341	0.381540	0.544048	0.551461	0.361564	0.127663	0.391321
Terminal	0.434478	0.403409	0.354379	0.506748	0.527654	0.335054	0.122152	0.412579
S.F.Ratio	0.126411	0.188506	0.274269	-0.387926	-0.297233	0.324504	0.370607	-0.573683
perc.alumni	-0.101158	-0.165516	-0.222723	0.455797	0.416832	-0.285457	-0.419334	0.565736
Expend	0.242935	0.161808	0.054221	0.657039	0.572905	0.000371	-0.201929	0.775328
Grad.Rate	0.150803	0.078982	-0.023251	0.493670	0.478985	-0.082239	-0.265158	0.572458

	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	0.187475	0.236138	0.229948	0.463924	0.434478	0.126411	-0.101158	0.242935	0.150803
Accept	0.119586	0.208705	0.256346	0.427341	0.403409	0.188506	-0.165516	0.161808	0.078982
Enroll	-0.023846	0.202057	0.339348	0.381540	0.354379	0.274269	-0.222723	0.054221	-0.023251
Top10perc	0.357366	0.153452	-0.116730	0.544048	0.506748	-0.387926	0.455797	0.657039	0.493670
Top25perc	0.330987	0.169761	-0.086810	0.551461	0.527654	-0.297233	0.416832	0.572905	0.478985
F.Undergrad	-0.054476	0.207879	0.359783	0.361564	0.335054	0.324504	-0.285457	0.000371	-0.082239
P.Undergrad	-0.067638	0.122529	0.344053	0.127663	0.122152	0.370607	-0.419334	-0.201929	-0.265158
Outstate	0.655489	0.005110	-0.325609	0.391321	0.412579	-0.573683	0.565736	0.775328	0.572458
Room.Board	1.000000	0.108924	-0.219554	0.341469	0.379270	-0.376430	0.272393	0.580622	0.425790
Books	0.108924	1.000000	0.239863	0.136390	0.159318	-0.008536	-0.042832	0.149983	-0.008051
Personal	-0.219554	0.239863	1.000000	-0.011684	-0.031971	0.173913	-0.305753	-0.163271	-0.290894
PhD	0.341469	0.136390	-0.011684	1.000000	0.862928	-0.129390	0.248877	0.510529	0.310019
Terminal	0.379270	0.159318	-0.031971	0.862928	1.000000	-0.150993	0.266033	0.524068	0.292803
S.F.Ratio	-0.376430	-0.008536	0.173913	-0.129390	-0.150993	1.000000	-0.412101	-0.654376	-0.308525
perc.alumni	0.272393	-0.042832	-0.305753	0.248877	0.266033	-0.412101	1.000000	0.462922	0.491408
Expend	0.580622	0.149983	-0.163271	0.510529	0.524068	-0.654376	0.462922	1.000000	0.415291
Grad.Rate	0.425790	-0.008051	-0.290894	0.310019	0.292803	-0.308525	0.491408	0.415291	1.000000

<u>Inferences from the correlation table:</u>

- All the highlighted fields have correlation more than 0.5 in both positive and negative relationships.
- Positive correlation is observed between various highlighted fields as : *Accept- Apps, Apps-Enroll, Enroll-Accept* etc.
- Negative Correlations can be observed in various fields like : *S.F.Ratio-Expend, Outstate-S.F.Ratio* etc

Heat Map to represent it visually:

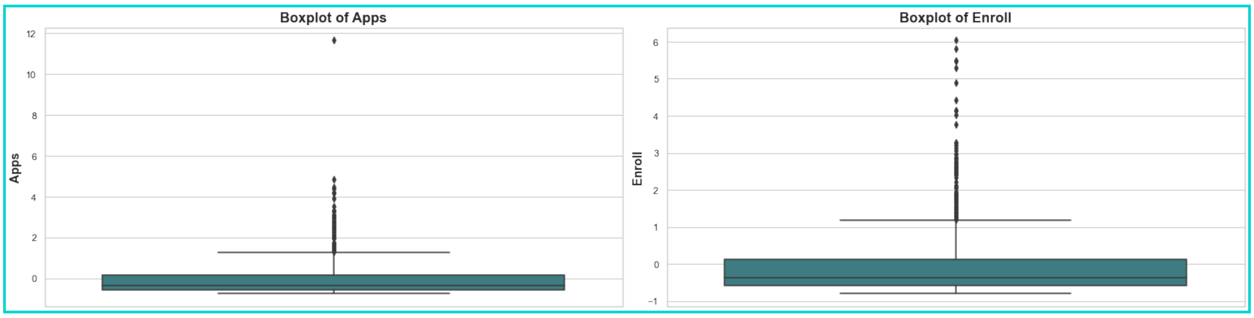


	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps																	
Accept	0.94																
Enroll	0.85	0.91															
Top10perc	0.34	0.19	0.18														
Top25perc	0.35	0.25	0.23	0.89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

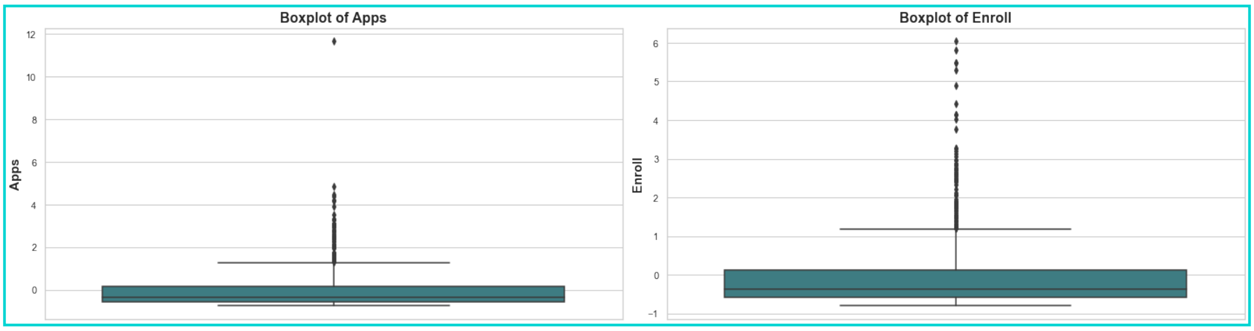
IV. Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]

We will use Box-plots of different features for finding Outliers :

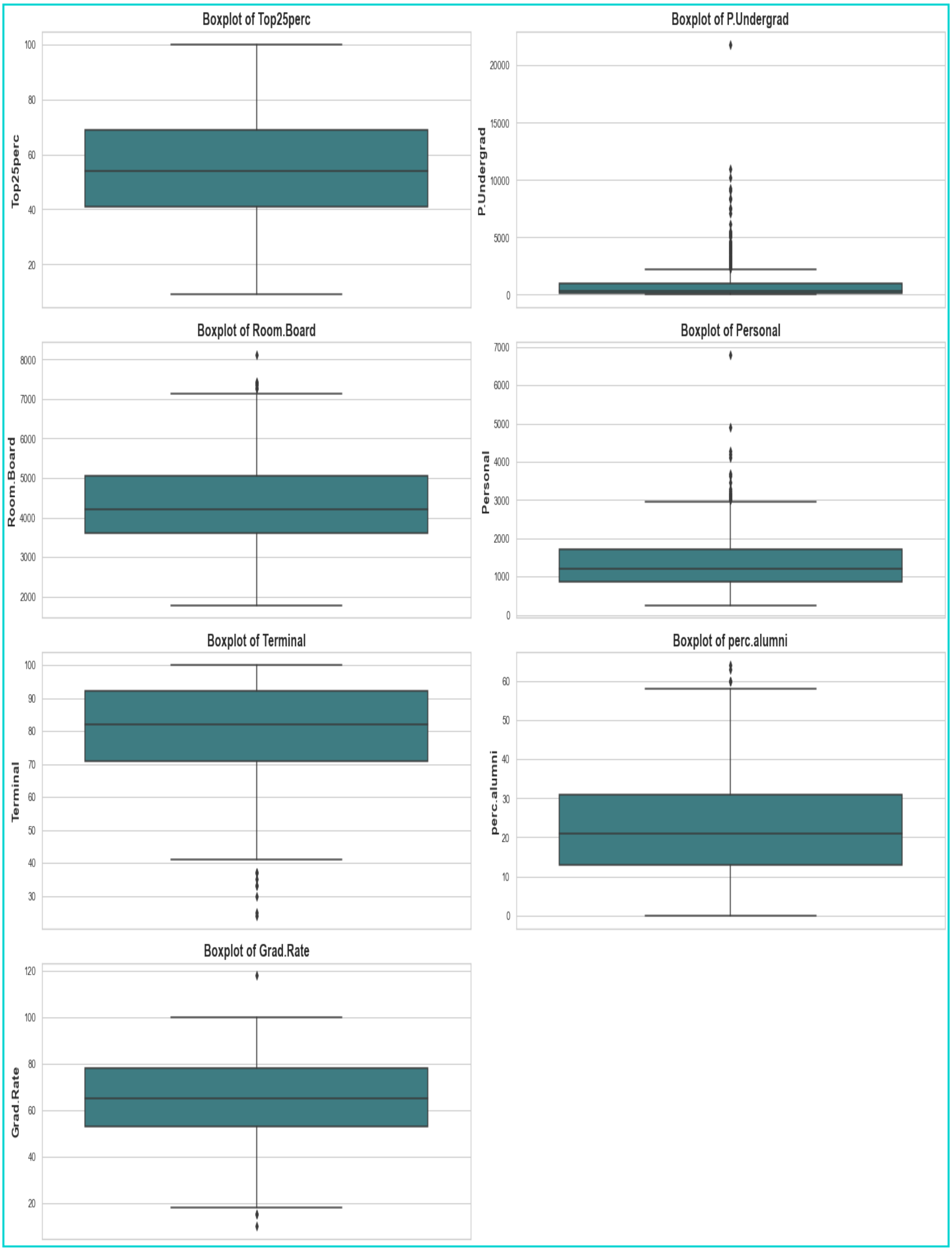
Box-Plots Before Scaling



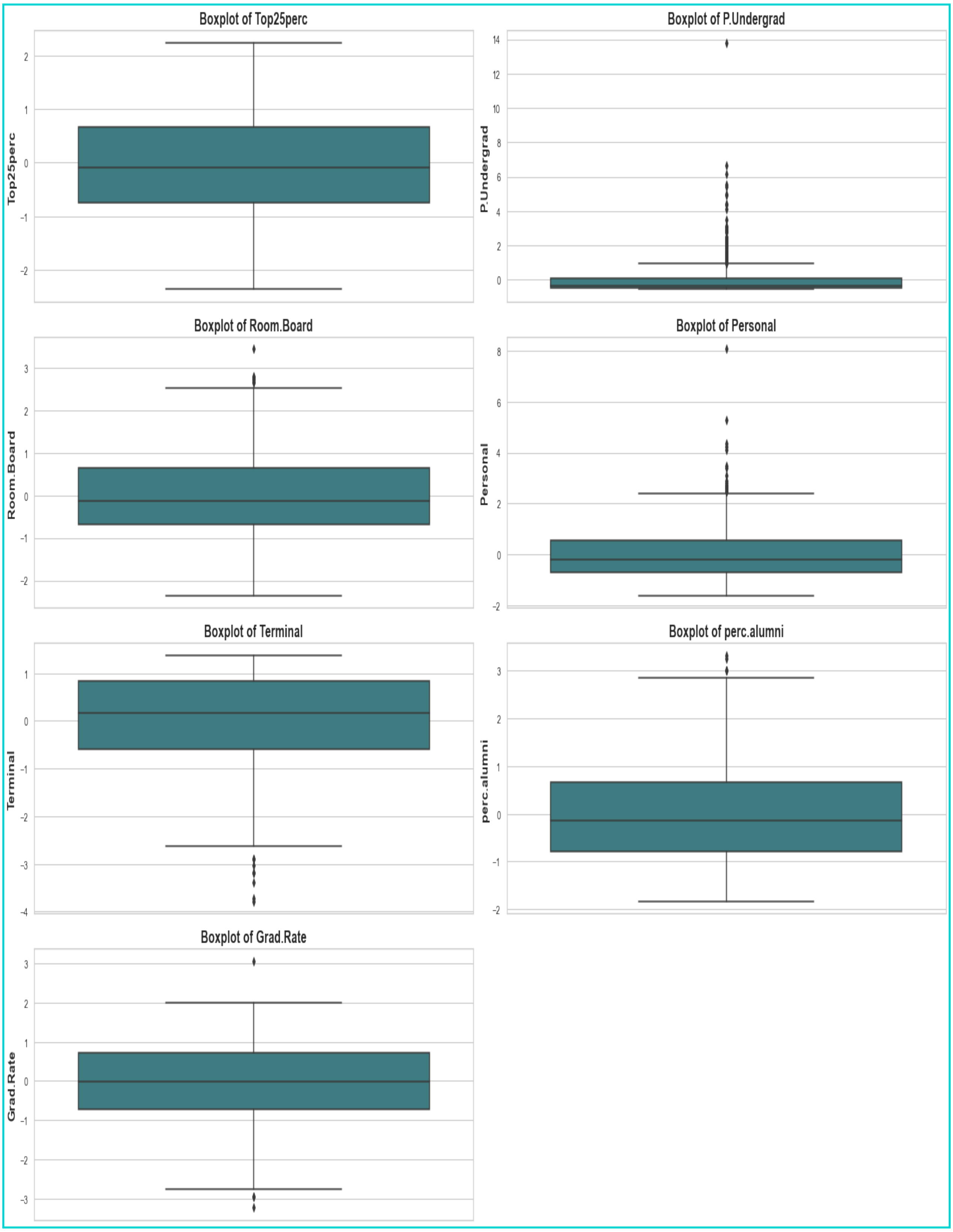
Box-Plots After Scaling



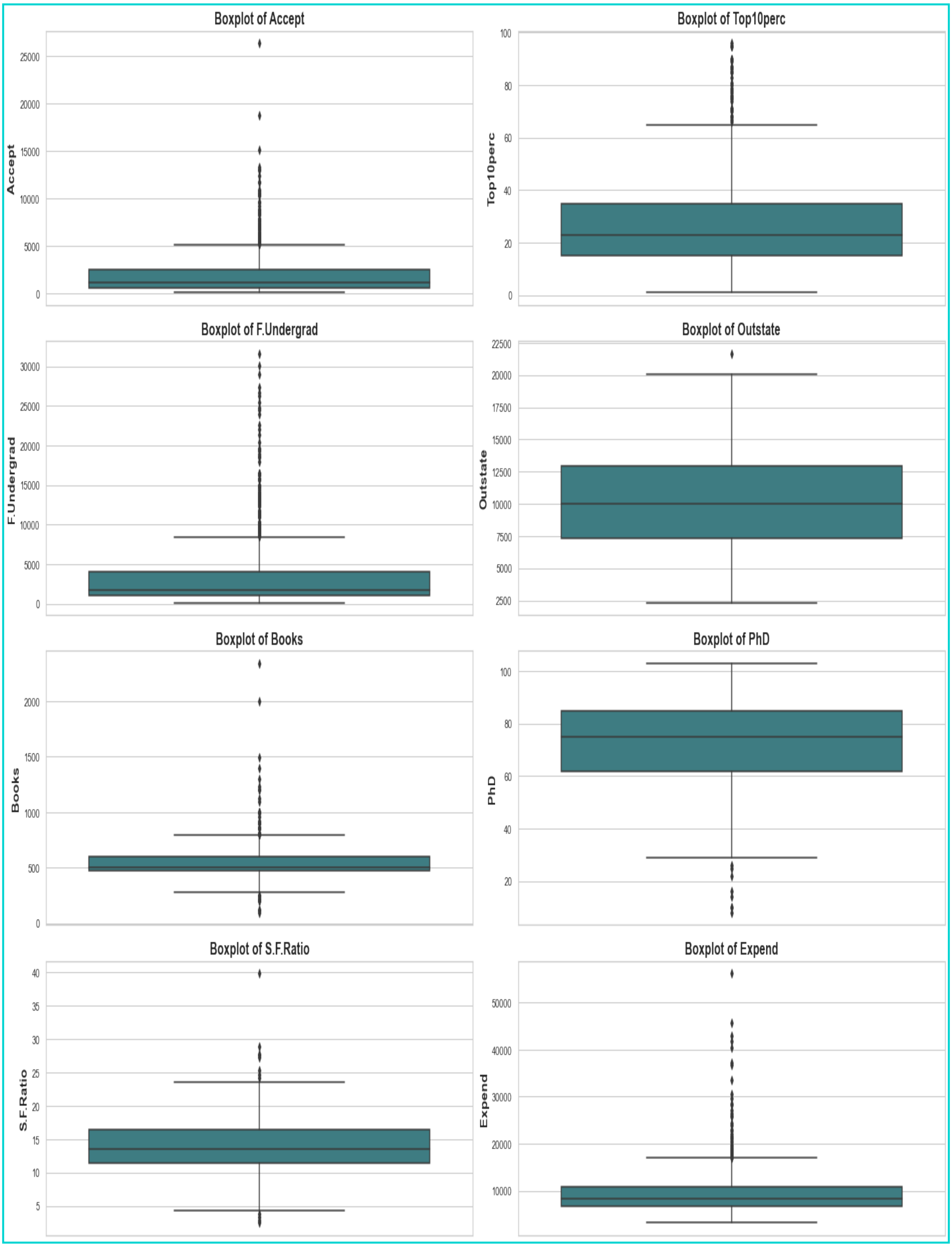
Box-Plots Before Scaling



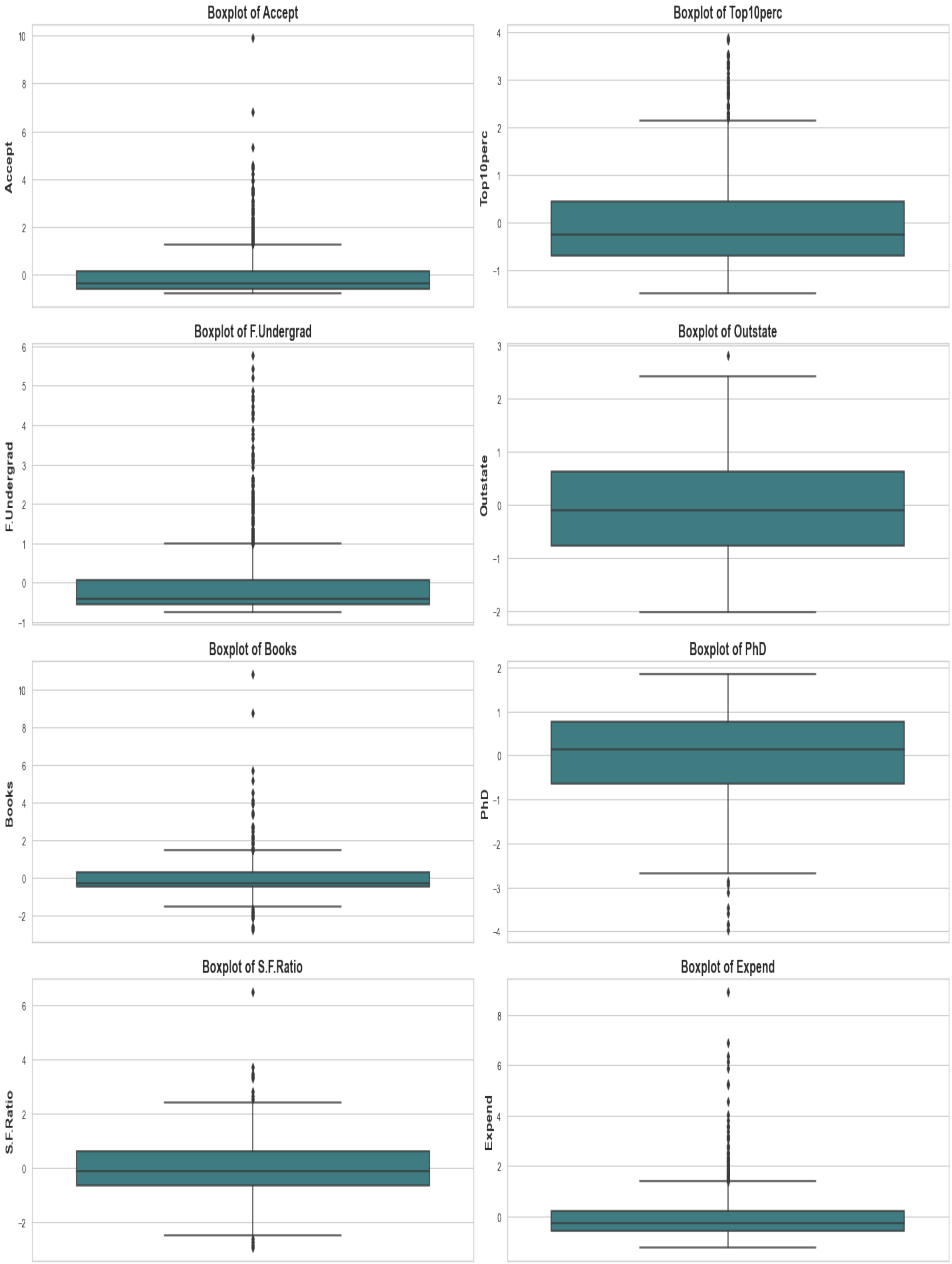
Box-Plots After Scaling



Box-Plots Before Scaling



Box-Plots After Scaling



Observation

Scaling doesn't effect the outliers, before scaling outliers were present and same is the case after scaling.

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

Statistical tests to be done before PCA¶

Bartletts Test of Sphericity¶

Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

- H0: All variables in the data are uncorrelated
- Ha: At least one pair of variables in the data are correlated

If the null hypothesis cannot be rejected, then PCA is not advisable.
If the p-value is small, then we can reject the null hypothesis and agree that there is at-least one pair of variables in the data which are correlated hence PCA is recommended.

Correlations are significant as P-Value = 0.0.

KMO Test¶

The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.
Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

Acceptable for kmo_model 0.8131251200373522.

Extract Eigen vectors

```
[[ 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]]
```

Extract Eigen Values

```
[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]
```

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

After performing PCA on all 17 features, we found the following.

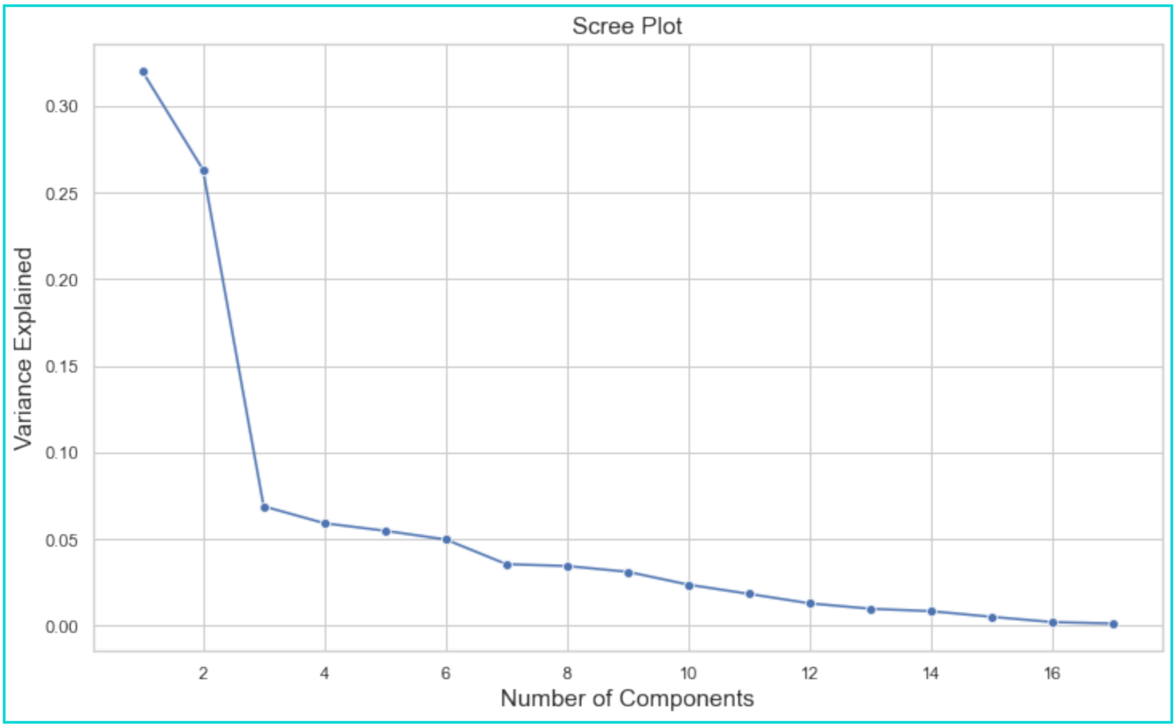
PC Data Frame

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Apps	0.25	0.33	-0.06	0.28	0.01	-0.02	-0.04	-0.10
Accept	0.21	0.37	-0.10	0.27	0.06	0.01	-0.01	-0.06
Enroll	0.18	0.40	-0.08	0.16	-0.06	-0.04	-0.03	0.06
Top10perc	0.35	-0.08	0.04	-0.05	-0.40	-0.05	-0.16	-0.12
Top25perc	0.34	-0.04	-0.02	-0.11	-0.43	0.03	-0.12	-0.10
F.Undergrad	0.15	0.42	-0.06	0.10	-0.04	-0.04	-0.03	0.08
P.Undergrad	0.03	0.32	0.14	-0.16	0.30	-0.19	0.06	0.57
Outstate	0.29	-0.25	0.05	0.13	0.22	-0.03	0.11	0.01
Room.Board	0.25	-0.14	0.15	0.18	0.56	0.16	0.21	-0.22
Books	0.06	0.06	0.68	0.09	-0.13	0.64	-0.15	0.21
Personal	-0.04	0.22	0.50	-0.23	-0.22	-0.33	0.63	-0.23
PhD	0.32	0.06	-0.13	-0.53	0.14	0.09	0.00	-0.08
Terminal	0.32	0.05	-0.07	-0.52	0.20	0.15	-0.03	-0.01
S.F.Ratio	-0.18	0.25	-0.29	-0.16	-0.08	0.49	0.22	-0.08
perc.alumni	0.21	-0.25	-0.15	0.02	-0.22	-0.05	0.24	0.68
Expend	0.32	-0.13	0.23	0.08	0.08	-0.30	-0.23	-0.05
Grad.Rate	0.25	-0.17	-0.21	0.27	-0.11	0.22	0.56	-0.01

	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	-0.09	0.05	0.04	0.02	0.60	0.08	0.13	0.46	0.36
Accept	-0.18	0.04	-0.06	-0.15	0.29	0.03	-0.15	-0.52	-0.54
Enroll	-0.13	0.03	-0.07	0.01	-0.44	-0.09	0.03	-0.40	0.61
Top10perc	0.34	0.06	-0.01	0.04	0.00	-0.11	0.70	-0.15	-0.14
Top25perc	0.40	0.01	-0.27	-0.09	0.02	0.15	-0.62	0.05	0.08
F.Undergrad	-0.06	0.02	-0.08	0.06	-0.52	-0.06	0.01	0.56	-0.41
P.Undergrad	0.56	-0.22	0.10	-0.06	0.13	0.02	0.02	-0.05	0.01
Outstate	0.00	0.19	0.14	-0.82	-0.14	-0.03	0.04	0.10	0.05
Room.Board	0.28	0.30	-0.36	0.35	-0.07	-0.06	0.00	-0.03	0.00
Books	-0.13	-0.08	0.03	-0.03	0.01	-0.07	-0.01	0.00	0.00
Personal	-0.09	0.14	-0.02	-0.04	0.04	0.03	0.00	-0.01	0.00
PhD	-0.19	-0.12	0.04	0.02	0.13	-0.69	-0.11	0.03	0.01
Terminal	-0.25	-0.09	-0.06	0.02	-0.06	0.67	0.16	-0.03	0.01
S.F.Ratio	0.27	0.47	0.45	-0.01	-0.02	0.04	-0.02	-0.02	0.00
perc.alumni	-0.26	0.42	-0.13	0.18	0.10	-0.03	-0.01	0.00	-0.02
Expend	-0.05	0.13	0.69	0.33	-0.09	0.07	-0.23	-0.04	-0.04
Grad.Rate	0.04	-0.59	0.22	0.12	-0.07	0.04	0.00	-0.01	-0.01

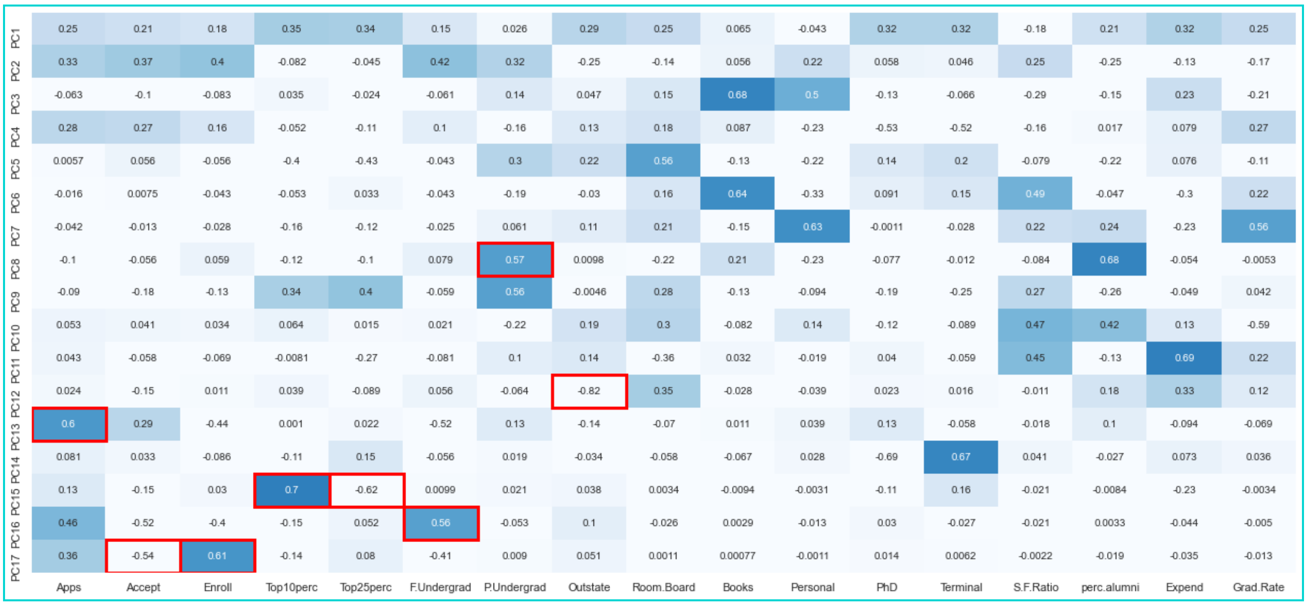
Then, we visualised the Scree Plot with the help of it.

Scree Plot



To identify which features have maximum loading across the components.
We will first plot the component loading on a heat-map.
For each feature, we find the maximum loading value across the components and mark the same with help of rectangular box.
Features marked with rectangular red box are the one having maximum loading on the respective component. We consider these marked features to decide the context that the component represents

Heat Map with Rectangles with maximum loading



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
PC1	0.25	0.21	0.18	0.35	0.34	0.15	0.026	0.29	0.25	0.065	-0.043	0.32	0.32	-0.18	0.21	0.32	0.25
PC2	0.33	0.37	0.4	-0.082	-0.045	0.42	0.32	-0.25	-0.14	0.056	0.22	0.058	0.046	0.25	-0.25	-0.13	-0.17
PC3	-0.063	-0.1	-0.083	0.035	-0.024	-0.061	0.14	0.047	0.15	0.68	0.5	-0.13	-0.066	-0.29	-0.15	0.23	-0.21
PC4	0.28	0.27	0.16	-0.052	-0.11	0.1	-0.16	0.13	0.18	0.087	-0.23	-0.53	-0.52	-0.16	0.017	0.079	0.27
PC5	0.0057	0.056	-0.056	-0.4	-0.43	-0.043	0.3	0.22	0.56	-0.13	-0.22	0.14	0.2	-0.079	-0.22	0.076	-0.11
PC6	-0.016	0.0075	-0.043	-0.053	0.033	-0.043	-0.19	-0.03	0.16	0.64	-0.33	0.091	0.15	0.49	-0.047	-0.3	0.22
PC7	-0.042	-0.013	-0.028	-0.16	-0.12	-0.025	0.061	0.11	0.21	-0.15	0.63	-0.0011	-0.028	0.22	0.24	-0.23	0.56
PC8	-0.1	-0.056	0.059	-0.12	-0.1	0.079	0.57	0.0098	-0.22	0.21	-0.23	-0.077	-0.012	-0.084	0.68	-0.054	-0.0053
PC9	-0.09	-0.18	-0.13	0.34	0.4	-0.059	0.56	-0.0046	0.28	-0.13	-0.094	-0.19	-0.25	0.27	-0.26	-0.049	0.042
PC10	0.053	0.041	0.034	0.064	0.015	0.021	-0.22	0.19	0.3	-0.082	0.14	-0.12	-0.089	0.47	0.42	0.13	-0.59
PC11	0.043	-0.058	-0.069	-0.0081	-0.27	-0.081	0.1	0.14	-0.36	0.032	-0.019	0.04	-0.059	0.45	-0.13	0.69	0.22
PC12	0.024	-0.15	0.011	0.039	-0.089	0.056	-0.064	-0.82	0.35	-0.028	-0.039	0.023	0.016	-0.011	0.18	0.33	0.12
PC13	0.6	0.29	-0.44	0.001	0.022	-0.52	0.13	-0.14	-0.07	0.011	0.039	0.13	-0.058	-0.018	0.1	-0.094	-0.069
PC14	0.081	0.033	-0.086	-0.11	0.15	-0.056	0.019	-0.034	-0.058	-0.067	0.028	-0.69	0.67	0.041	-0.027	0.073	0.036
PC15	0.13	-0.15	0.03	0.7	-0.62	0.0099	0.021	0.038	0.0034	-0.0094	-0.0031	-0.11	0.16	-0.021	-0.0084	-0.23	-0.0034
PC16	0.46	-0.52	-0.4	-0.15	0.052	0.56	-0.053	0.1	-0.026	0.0029	-0.013	0.03	-0.027	-0.021	0.0033	-0.044	-0.005
PC17	0.36	-0.54	0.61	-0.14	0.08	-0.41	0.009	0.051	0.0011	0.00077	-0.0011	0.014	0.0062	-0.0022	-0.019	-0.035	-0.013

VII. 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]

To write the first PC, we need to first check the values as below:

Features	PC1
Apps	0.25
Accept	0.21
Enroll	0.18
Top10perc	0.35
Top25perc	0.34
F.Undergrad	0.15
P.Undergrad	0.03
Outstate	0.29
Room.Board	0.25
Books	0.06
Personal	-0.04
PhD	0.32
Terminal	0.32
S.F.Ratio	-0.18
perc.alumni	0.21
Expend	0.32
Grad.Rate	0.25

Equation for PC1

0.25 * Apps + 0.21 * Accept + 0.18 * Enroll + 0.35 * Top10perc + 0.34 * Top25perc + 0.15 * F.Undergrad + 0.03 * P.Undergrad + 0.29 * Outstate + 0.25 * Room.Board + 0.06 * Books + -0.04 * Personal + 0.32 * PhD + 0.32 * Terminal + -0.18 * S.F.Ratio + 0.21 * perc.alumni + 0.32 * Expend + 0.25 * Grad.Rate

Equation for few other PCs

1. **PC1 =** 0.25 * Apps + 0.21 * Accept + 0.18 * Enroll + 0.35 * Top10perc + 0.34 * Top25perc + 0.15 * F.Undergrad + 0.03 * P.Undergrad + 0.29 * Outstate + 0.25 * Room.Board + 0.06 * Books + -0.04 * Personal + 0.32 * PhD + 0.32 * Terminal + -0.18 * S.F.Ratio + 0.21 * perc.alumni + 0.32 * Expend + 0.25 * Grad.Rate
2. **PC2 =** 0.33 * Apps + 0.37 * Accept + 0.40 * Enroll + -0.08 * Top10perc + -0.04 * Top25perc + 0.42 * F.Undergrad + 0.32 * P.Undergrad + -0.25 * Outstate + -0.14 * Room.Board + 0.06 * Books + 0.22 * Personal + 0.06 * PhD + 0.05 * Terminal + 0.25 * S.F.Ratio + -0.25 * perc.alumni + -0.13 * Expend + -0.17 * Grad.Rate
3. **PC3 =** -0.06 * Apps + -0.10 * Accept + -0.08 * Enroll + 0.04 * Top10perc + -0.02 * Top25perc + -0.06 * F.Undergrad + 0.14 * P.Undergrad + 0.05 * Outstate + 0.15 * Room.Board + 0.68 * Books + 0.50 * Personal + -0.13 * PhD + -0.07 * Terminal + -0.29 * S.F.Ratio + -0.15 * perc.alumni + 0.23 * Expend + -0.21 * Grad.Rate
4. **PC4 =** 0.28 * Apps + 0.27 * Accept + 0.16 * Enroll + -0.05 * Top10perc + -0.11 * Top25perc + 0.10 * F.Undergrad + -0.16 * P.Undergrad + 0.13 * Outstate + 0.18 * Room.Board + 0.09 * Books + -0.23 * Personal + -0.53 * PhD + -0.52 * Terminal + -0.16 * S.F.Ratio + 0.02 * perc.alumni + 0.08 * Expend + 0.27 * Grad.Rate
5. **PC5 =** 0.01 * Apps + 0.06 * Accept + -0.06 * Enroll + -0.40 * Top10perc + -0.43 * Top25perc + -0.04 * F.Undergrad + 0.30 * P.Undergrad + 0.22 * Outstate + 0.56 * Room.Board + -0.13 * Books + -0.22 * Personal + 0.14 * PhD + 0.20 * Terminal + -0.08 * S.F.Ratio + -0.22 * perc.alumni + 0.08 * Expend + -0.11 * Grad.Rate
6. **PC6 = -**0.02 * Apps + 0.01 * Accept + -0.04 * Enroll + -0.05 * Top10perc + 0.03 * Top25perc + -0.04 * F.Undergrad + -0.19 * P.Undergrad + -0.03 * Outstate + 0.16 * Room.Board + 0.64 * Books + -0.33 * Personal + 0.09 * PhD + 0.15 * Terminal + 0.49 * S.F.Ratio + -0.05 * perc.alumni + -0.30 * Expend + 0.22 * Grad.Rate

VIII.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?

The cumulative values of the eigenvalues¶

When PCA was performed with all the fields, then the corresponding cumulative values of the eigenvalues were:

```
[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. ]
```

From these Eigenvalues we can observe the following:

1. The first PC component alone contributes approx 32%.
2. If we choose first 6 components, cumulatively contributes to approx 81%.

Scree Plot for Selected Pcs¶

On the basis of cumulative explained variance ratio , we can choose first 6 PCs, which will cover 81% approximately

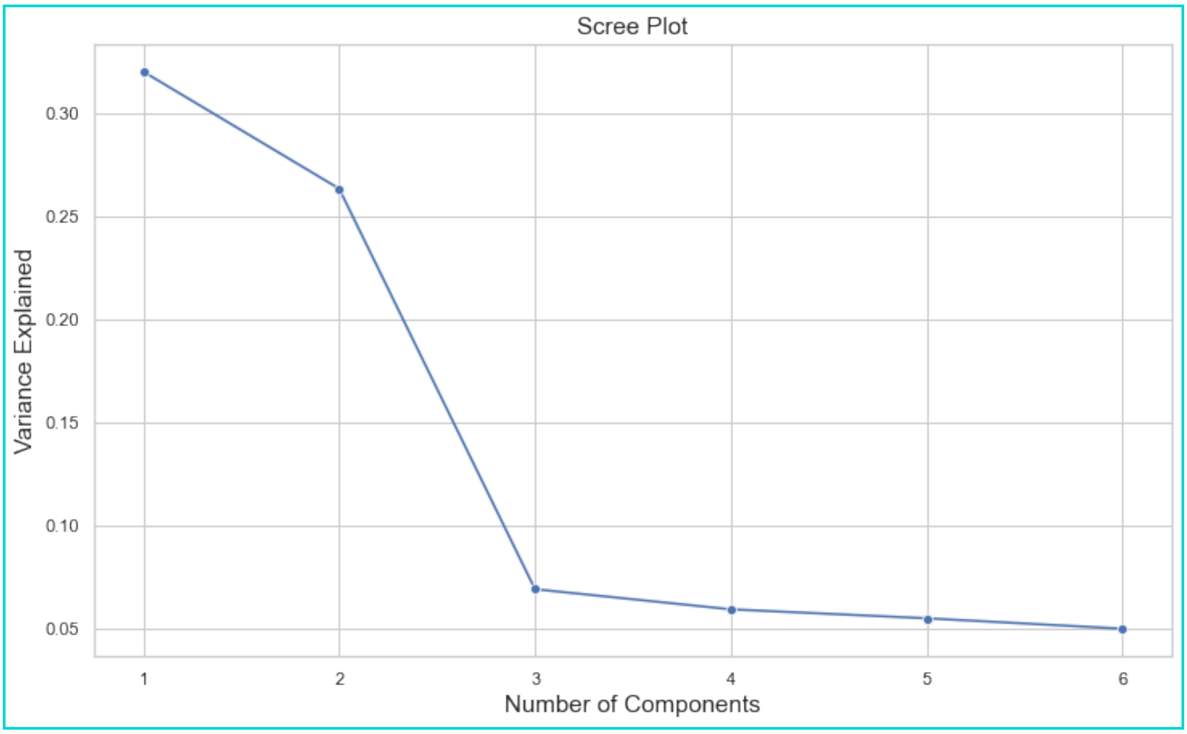
Extract Eigen vectors of 6 PC

```
[[ 0.2487656 , 0.2076015 , 0.17630359, 0.35427395, 0.34400128, 0.15464096,
0.0264425 , 0.29473642, 0.24903045, 0.06475752, -0.04252854, 0.31831287,
0.31705602, -0.17695789, 0.20508237, 0.31890875, 0.25231565],
[0.33159823, 0.37211675, 0.40372425, -0.08241182, -0.04477866, 0.41767377,
0.31508783, -0.24964352, -0.13780888, 0.05634184, 0.21992922, 0.05831132,
0.04642945, 0.24666528, -0.24659527, -0.13168986, -0.16924053],
[-0.06309209, -0.10124907, -0.08298558, 0.03505553,
-0.02414794,-0.06139296, 0.13968171, 0.04659888, 0.14896739, 0.67741165,
0.49972112, -0.12702837, -0.06603755, -0.2898484 , -0.14698927, 0.22674398,
-0.20806465],
[0.28131052, 0.26781736, 0.16182679, -0.05154725, -0.10976654, 0.10041231,
-0.15855849, 0.13129136, 0.18499599, 0.08708922, -0.23071057, -0.53472483,
-0.51944302, -0.16118949, 0.01731422, 0.0792735 , 0.26912907],
[0.00574142, 0.05578609, -0.05569364, -0.39543435, -0.42653359,
-0.04345436, 0.30238541, 0.222532 , 0.56091947, -0.12728883, -0.22231102,
0.14016633, 0.20471973, -0.07938825, -0.21629741, 0.07595812, -0.10926791],
[-0.01623744, 0.00753468, -0.04255798, -0.0526928 , 0.03309159,
-0.04345423, -0.19119858, -0.03000039, 0.16275545, 0.64105495, -0.331398,
0.09125552, 0.15492765, 0.48704587, -0.04734001, -0.29811862, 0.21616331]]
```

Extract Eigen values of 6 PC

[5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123, 0.84849117]

Scree Plot



Cumulative values of the eigenvalues of selected PCs

Hence, those 6 components have been selected further. The cumulative values of the eigenvalues of those selected components are : *[0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154, 0.81657854]*

Eigenvectors¶

Eigenvalues are simply the coefficients attached to eigenvectors, which give the axes magnitude. In this case, they are the measure of the data’s covariance. By ranking your eigenvectors in order of their eigenvalues, highest to lowest, you get the principal components in order of significance. The eigenvectors and eigenvalues of a covariance (or correlation) matrix represent the “core” of a PCA: The eigenvectors (principal components) determine the directions of the new feature space, and the eigenvalues determine their magnitude.

IX. 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]

In this case study, the original number of PCs in this case study was 17 but after PCA analysis could give us 6 PCs, based on above analysis. Hence, Principal Component Analysis has many benefits for this case study:

- Since there are many features in the dataframe, by doing PCA, we can reduce the number of features and thus running models on the same dataframe can be faster and more efficient.
- As most of the features are correlated, hence most of the features are redundant. So, those can be removed.
- The Principle components obtained has few number of features, which are not correlated and can be used for analysis efficiently.

Advantages of Principal Component Analysis

- Improves Algorithm Performance:

With so many features, the performance of our algorithm will drastically degrade. PCA is a very common way to speed up your Machine Learning algorithm by getting rid of correlated variables which don't contribute in any decision making. The training time of the algorithms reduces significantly with less number of features.

So, if the input dimensions are too high, then using PCA to speed up the algorithm is a reasonable choice.

- Improves Visualization:

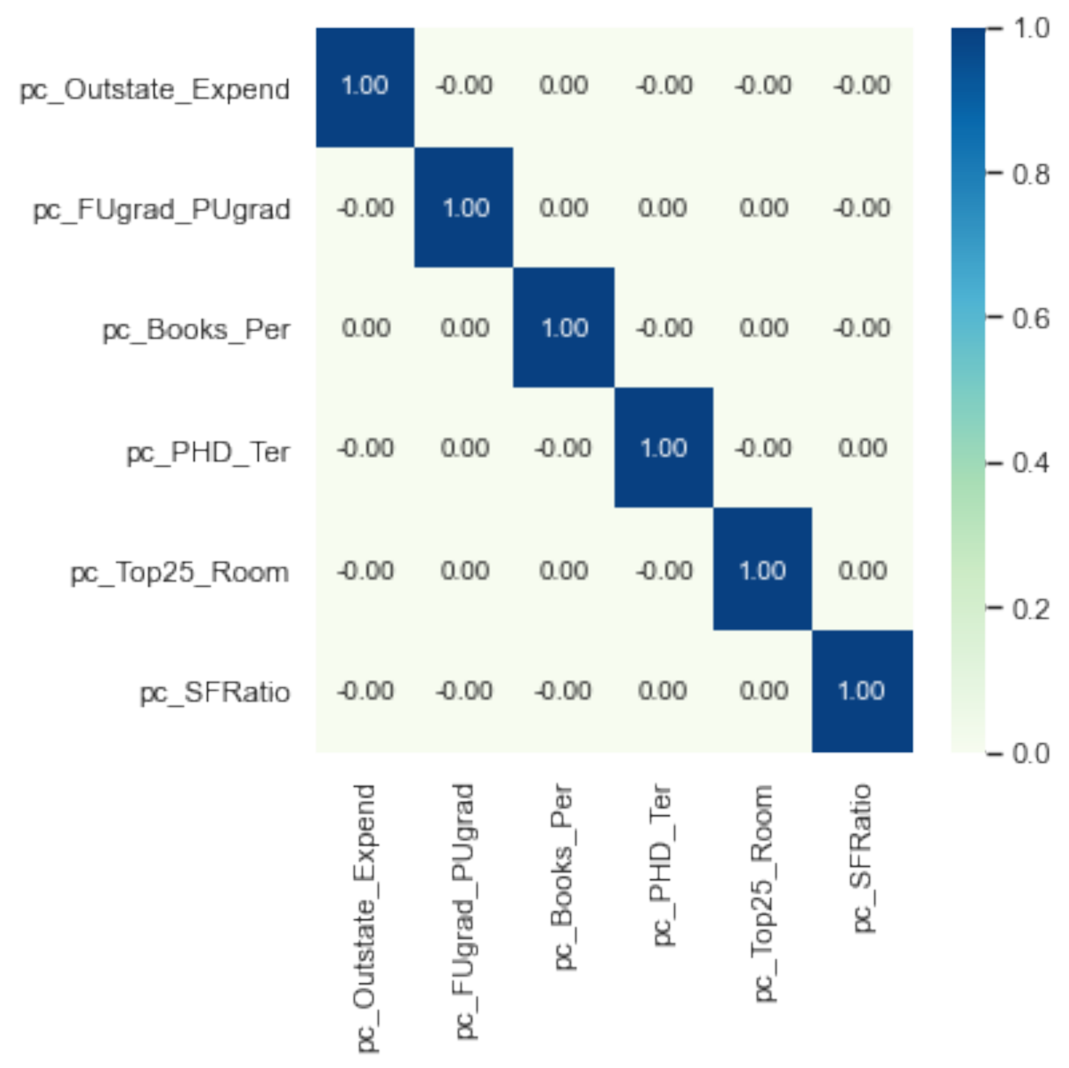
It is very hard to visualise and understand the data in high dimensions. PCA transforms a high dimensional data to low dimensional data, so that it can be visualised easily.

- Removes Correlated Features:

In this scenario, we had 17 features in our dataset. We should not run our algorithm on all the features as it will reduce the performance of your algorithm and it will not be easy to visualise that many features in any kind of graph. So, we MUST reduce the number of features in our dataset.

we need to find out the correlation among the features (correlated variables). Finding correlation manually in multiple features is dificult, frustrating and time-consuming. PCA does this for us efficiently.

After implementing the PCA on our dataset, all the Principal Components are independent of one another. There is no correlation among them.


pc_Outstate_Expend
1.00
-0.00
0.00
-0.00
-0.00
-0.00
pc_FUgrad_PUgrad
-0.00
1.00
0.00
0.00
0.00
-0.00
pc_Books_Per
0.00
0.00
1.00
-0.00
0.00
-0.00
pc_PHD_Ter
-0.00
0.00
-0.00
1.00
-0.00
0.00
pc_Top25_Room
-0.00
0.00
0.00
-0.00
1.00
0.00
pc_SFRatio
-0.00
-0.00
-0.00
0.00
0.00
1.00
pc_Outstate_Expend
pc_FUgrad_PUgrad
pc_Books_Per
pc_PHD_Ter
pc_Top25_Room
pc_SFRatio
1.0
0.8
0.6
0.4
0.2
0.0

Thank