ADVANCE STATISTICS ASSIGNMENT

TABLE OF CONTENTS

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?	25
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]	25

LIST OF FIGURES

LIST OF TABLES

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.

1A.1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

1a. Postulate the Null and Alternate Hypothesis for Education Qualification

H_0: The mean Salary for all Education field remains the same

H_1: The mean Salary of Education field is different

1b. Postulate the Null and Alternate Hypothesis for Occupaction

H_0: The mean Salary Across all Occupation field remains the same

H_1: The mean Salary of various Occupation field is different

1A.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.

For Education

Degree of Freedom between groups 2

Degree of Freedom within groups 37

	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

Table 1: one-way ANOVA on Salary with respect to Education

Since the p value is less than the significance level α(0.05), we can reject the null hupothesis and state that there is a difference in the mean salaries across various Education Field.

1A.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.

For Occupation

Degree of Freedom between groups 3

Degree of Freedom within groups 36

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	0.884144	**0.458508**

	df	sum_sq	mean_sq	F	PR(>F)
Residual	36.0	1.528092e+11	4.244701e+09	NaN	NaN

Table 2: one-way ANOVA on Salary with respect to Occupation

Now, we see that the corresponding p-value is greater than α(0.05). Thus, we fail to reject the Null Hypothesis (H0) and state that there is no difference in the mean salaries across various Occupation Field.

1A.4 . If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result.

Education	Salary
Bachelors	165152.93
Doctorate	208427.00
HS-grad	75038.78

Table 3: Mean Values of All Education levels

- We Can concure that the mean Salary of various Education field is different from the above table
- Doctorate Field seem to have the highest Salary package when compared to others
- This indicates that Higher the edcation field Better the Salary Expectency

Problem 1B

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.

SHAPIRO TEST

- For bachelors, ShapiroResult(statistic=0.9625411629676819, pvalue=0.7080602049827576)
- For doctorate, ShapiroResult(statistic=0.8669456243515015, pvalue=0.019654255360364914)
- For HS_grad, ShapiroResult(statistic=0.885286271572113, pvalue=0.1783432960510254)
- For adm_clerical, ShapiroResult(statistic=0.9105437994003296, pvalue=0.24753518402576447)
- For exec_managerial, ShapiroResult(statistic=0.8690852522850037, pvalue=0.22258324921131134)
- For prof_speciality, ShapiroResult(statistic=0.7667546272277832, pvalue=0.0019934121519327164)
- For sales, ShapiroResult(statistic=0.8897126913070679, pvalue=0.11683900654315948)

We see that, except for prof_speciality salary day, rest all are Normally Distributed. We won't be converting it for not but we can use transformation techniques to normalize data

LEVENE TEST

- For Education, LeveneResult(statistic=2.003165251179662, pvalue=0.14855191967482995)
- For Occupation, LeveneResult(statistic=2.4685686855381976, pvalue=0.07631646420034875)
- Here we infer that Variance across all levels are Equal since p value is greater than alpha

Let us now perform the Two Way ANOVA to analyse the effect of both the treatments on the 'Salary' variable.

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

Table 4: Two Way ANOVA to analyse the effect of both the treatments

We see that P value for both the treatments is less than Significant value α(0.05)

Let us check whether there is any interaction effect between the treatments.

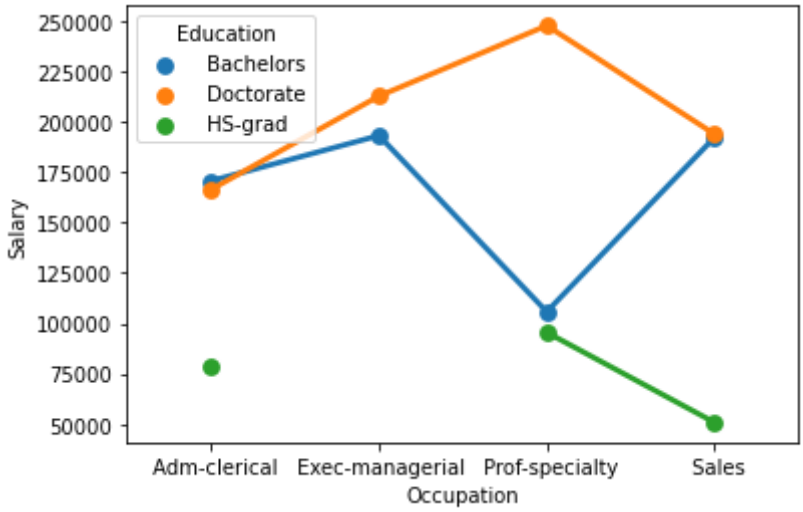


Fig1. Point Plot to check whether there is any interaction effect between the treatments

We can see that there is some sort of interaction between the two treatments

1B.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?

The Hypothesis for the Two Way ANOVA are:

H_0: The mean Salary across all the levels are same

H_a: For at least one level of Education or Occupation, mean Salary is different

	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

Table 5: two-way ANOVA based on Salary with respect to both Education and Occupation along with their interaction

Due to the inclusion of the interaction effect term, we can see a slight change in the p-value of the first two treatments as compared to the Two-Way ANOVA without the interaction effect terms.And we see that the p-value of the interaction effect term of 'Education' and 'Occupation' suggests that the Null Hypothesis is rejected in this case.

1B.3 . Explain the business implications of performing ANOVA for this particular case study.

- ANOVA is used to compare different levels of Treatment, here in this case is Education and Occupation, based on the samples collected.
- Education Field in itself had a variation when its levels are compared but we see a large change in F stats when Co-relation between Education and Occupation is introduced.
- We see that Occupation field in itself doesn't so much difference with Salary package but the introduction of corelation between Education field and occupation, there is a greater difference as compared with only Occupation field.
- We reject the null hypothesis and infer that Higher the Edcation field Better the Salary Expectency and this also affects the Occupation we choose with the level of Education a Sample has.

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.

Sample of the Dataset

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Roo
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	330(
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	645(
2	Adrian College	1428	1097	336	22	50	1036	99	11250	375(
3	Agnes Scott College	417	349	137	60	89	510	63	12960	545(
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	412(

Table 6: Sample dataset of Education - Post 12th Standard

Dataset has 777 different Colleges / Universities with each College / University having different fields such as

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

The Dataset contains no. of rows: 777 & no. of columns: 18

The Type of Variables in the Data Frame and a check for missing Values in the Dataset is as below:

There are a Totla of 777 Rows and 18 Columns in the Dataset. Out of 18 Columsn, we observe only 1 Object type and 1 Float Type. Rest are all Type Integer. Data columns (total 18 columns):

	Column	Non-Null	Count	Dtype
0	Names	777	non-null	object
1	Apps	777	non-null	int64
2	Accept	777	non-null	int64
3	Enroll	777	non-null	int64
4	Top10perc	777	non-null	int64
5	Top25perc	777	non-null	int64
6	F.Undergrad	777	non-null	int64
7	P.Undergrad	777	non-null	int64
8	Outstate	777	non-null	int64
9	Room.Board	777	non-null	int64
10	Books	777	non-null	int64
11	Personal	777	non-null	int64
12	PhD	777	non-null	int64
13	Terminal	777	non-null	int64
14	S.F.Ratio	777	non-null	float64
15	perc.alumni	777	non-null	int64
16	Expend	777	non-null	int64
17	Grad.Rate	777	non-null	int64

Table 7: Missing Value Check and Datatype check

We see that there are no missing values present in the Dataset

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.64	3870.20	81.0	776.0	1558.0	3624.0	48094.0

	count	mean	std	min	25%	50%	75%	max
Accept	777.0	2018.80	2451.11	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.97	929.18	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.56	17.64	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.80	19.80	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.91	4850.42	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.30	1522.43	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.67	4023.02	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.53	1096.70	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.38	165.11	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1340.64	677.07	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	72.66	16.33	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.70	14.72	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.09	3.96	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	22.74	12.39	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.17	5221.77	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.46	17.18	10.0	53.0	65.0	78.0	118.0

Table 8: Descriptive Stats

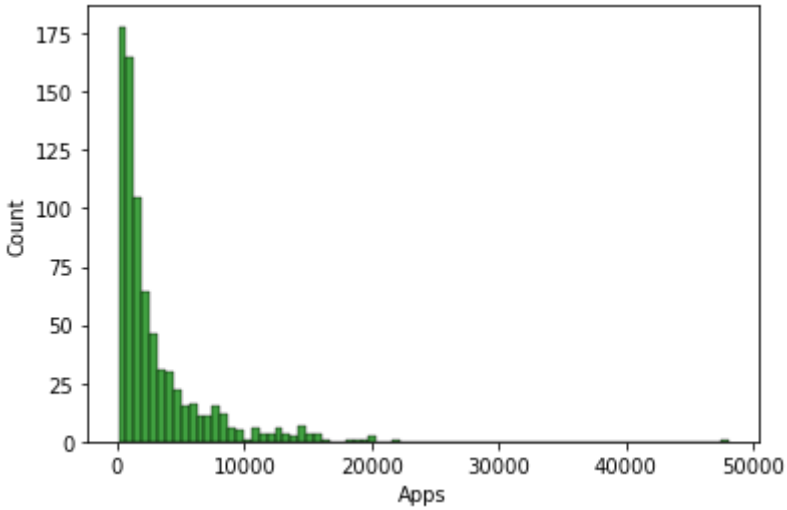
From the Dataset and descriptive stats, we can infer the below details.

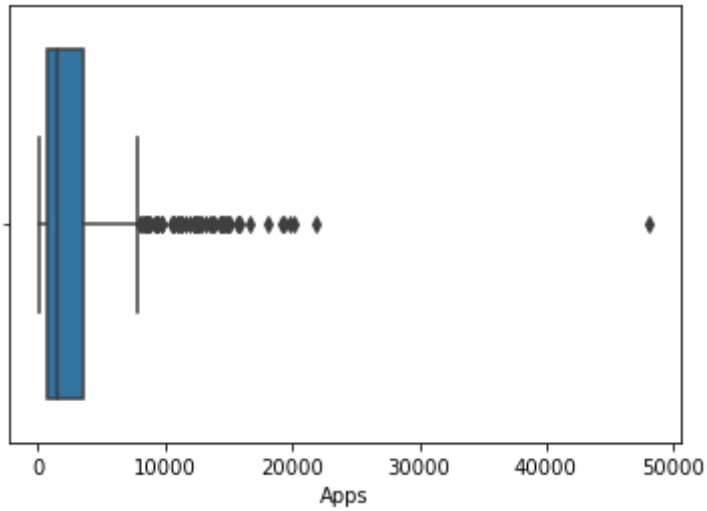
- On an Average there are nearly 3001 applications per College/University of which 2018 Applications are accepted on an Average
- On an Average 780 New Students enroll of which 27 are new students from top 10% of Higher Secondary class and 55 are new students from top 25% of Higher Secondary class.
- We can also see that on an Average per College / University, 3700 students are full-time undergraduate students and 855 students are part-time undergraduate students
- An estimate Average Spending per student is $1340.64
- Atleast 72 % professors Hold PhD Degree and 79% Hold terminal Degree.
- We see that Per College / University, the Grad rate is atleast 65.46%

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

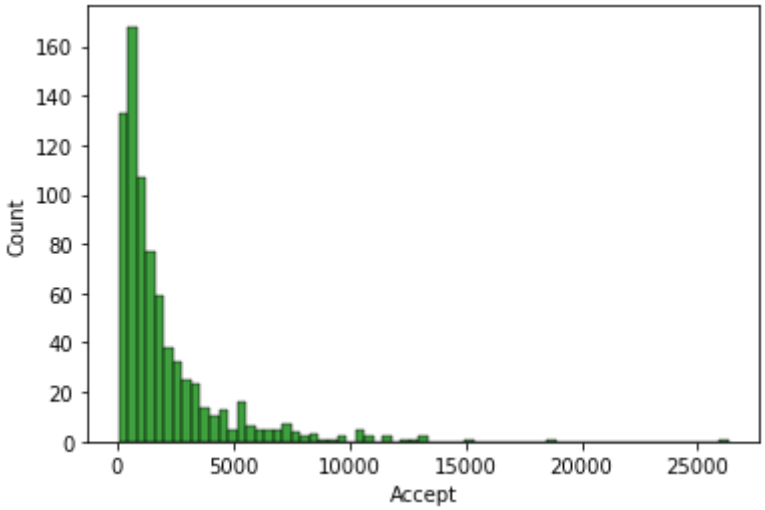
UNIVARIATE ANALYSIS

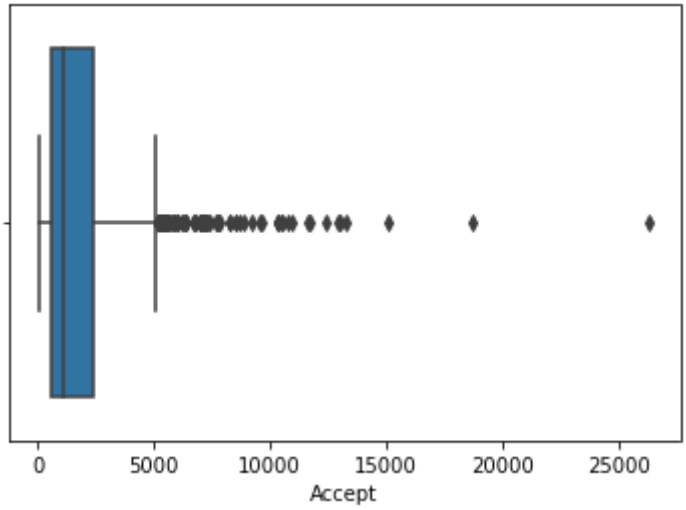
- Distribution & Boxplot of Apps



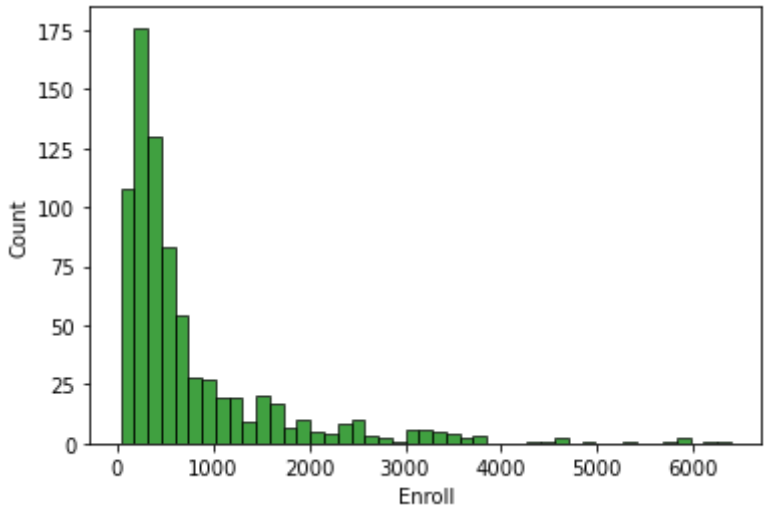


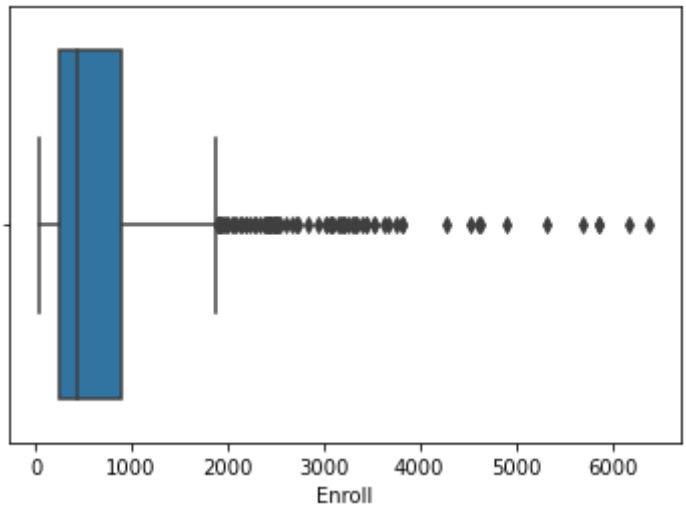
- Distribution & Boxplot of Accept



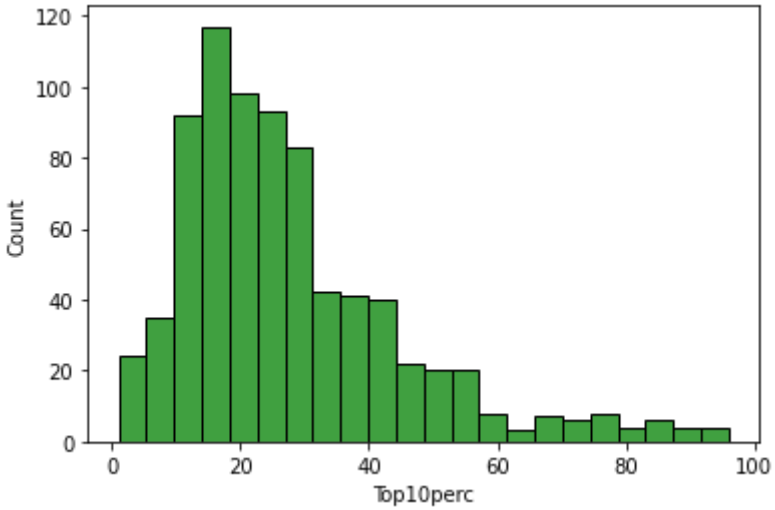


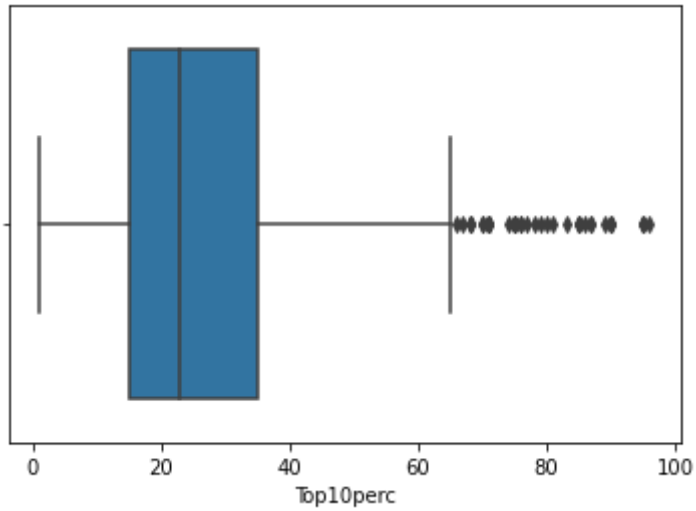
- Distribution & Boxplot of Enroll



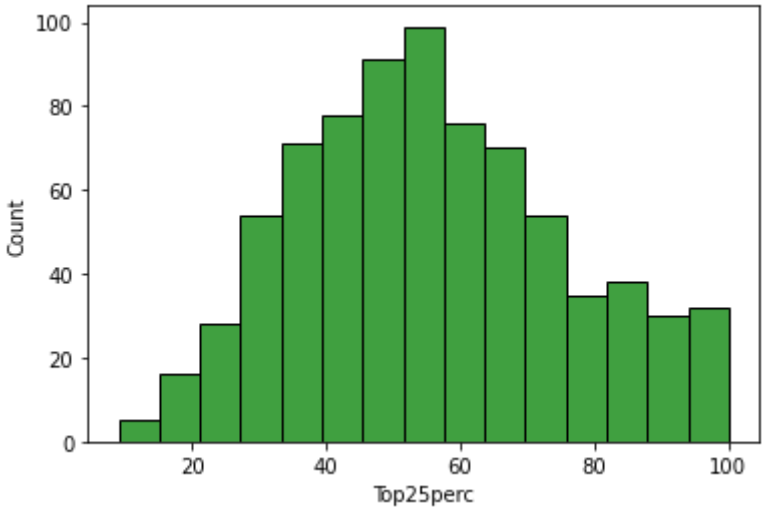


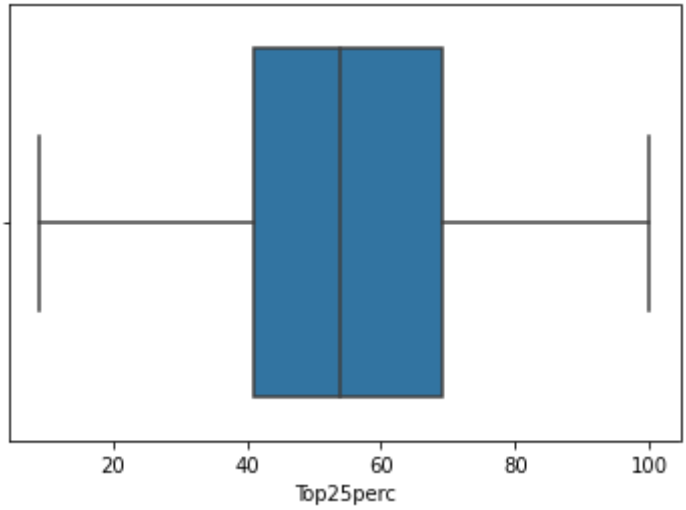
- Distribution & Boxplot of Top10perc



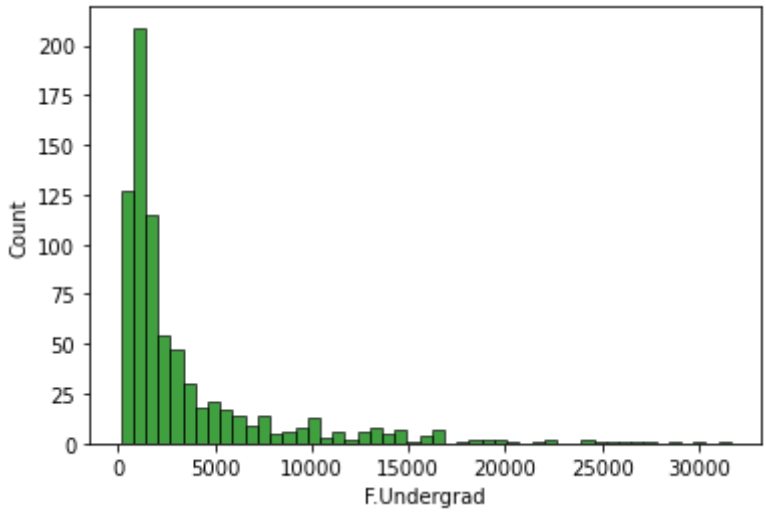


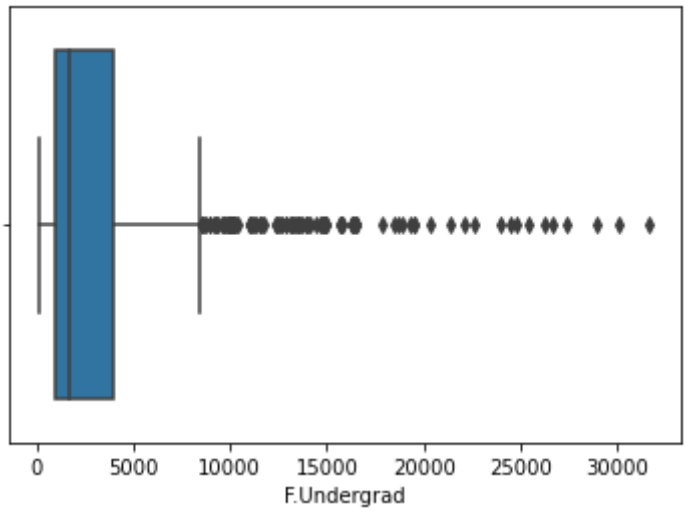
- Distribution & Boxplot of Top25perc



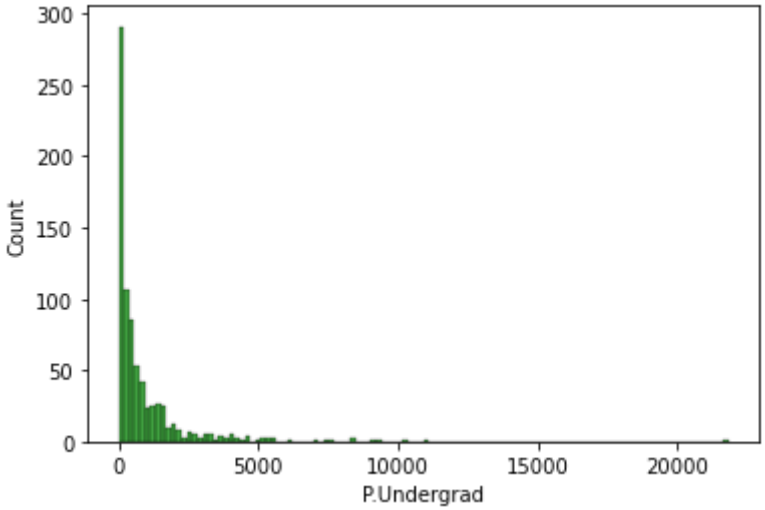


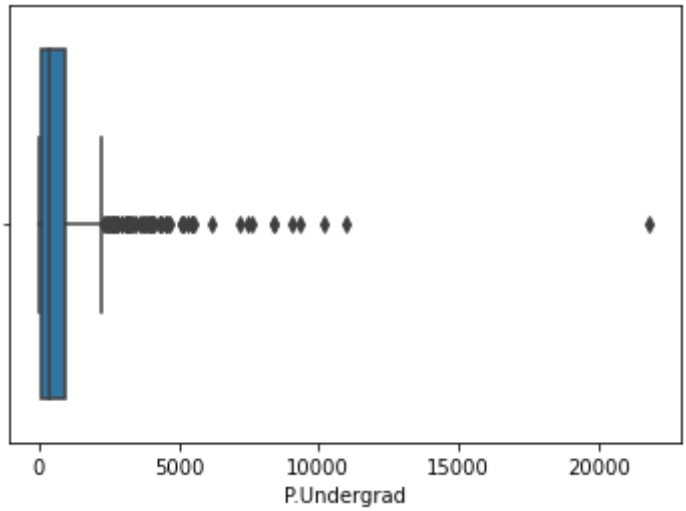
- Distribution & Boxplot of F.Undergrad



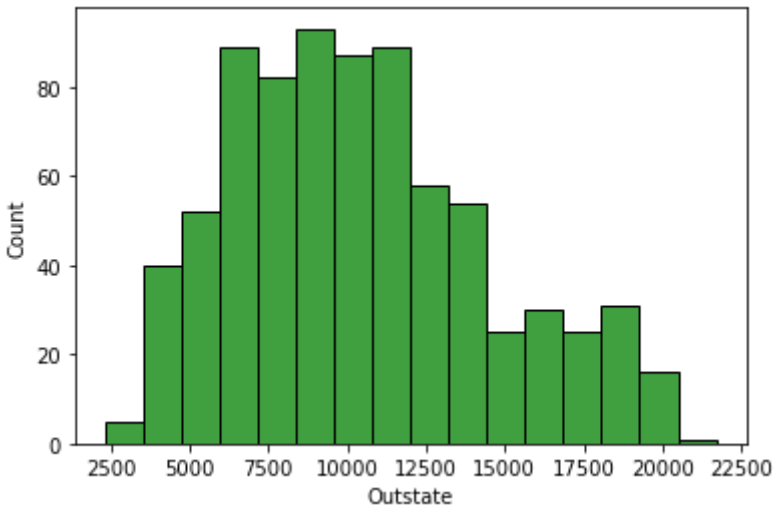


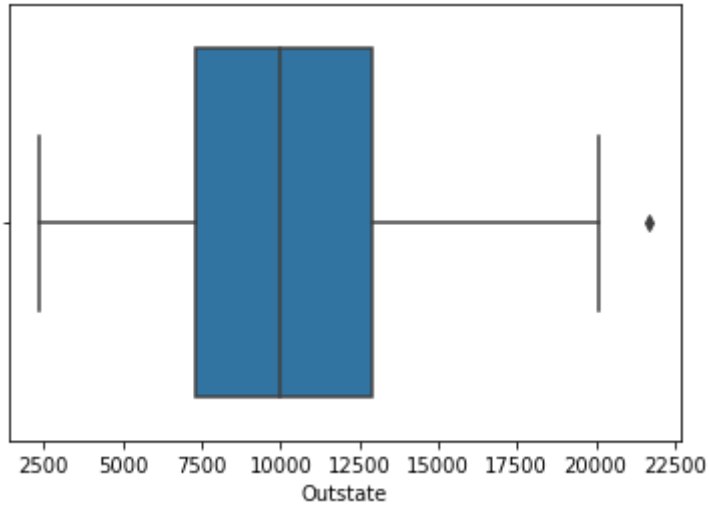
- Distribution & Boxplot of P.Undergrad



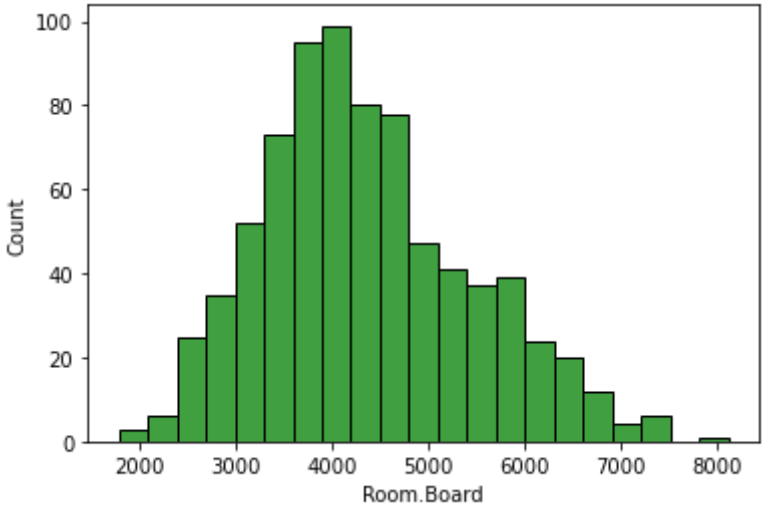


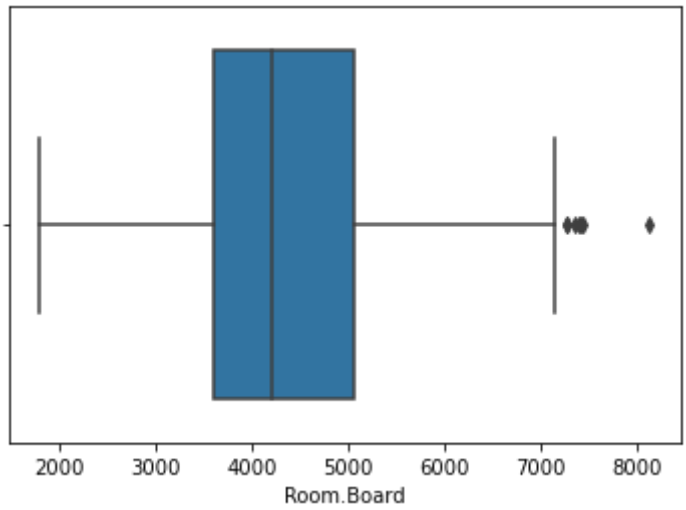
- Distribution & Boxplot of Outstate



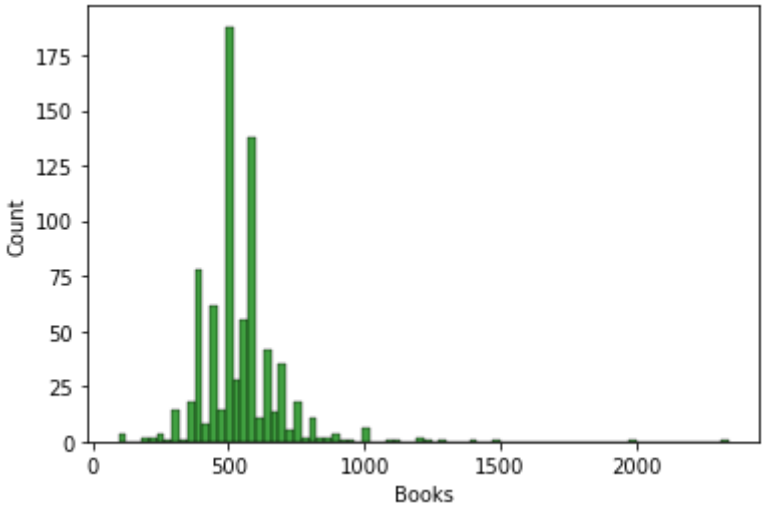


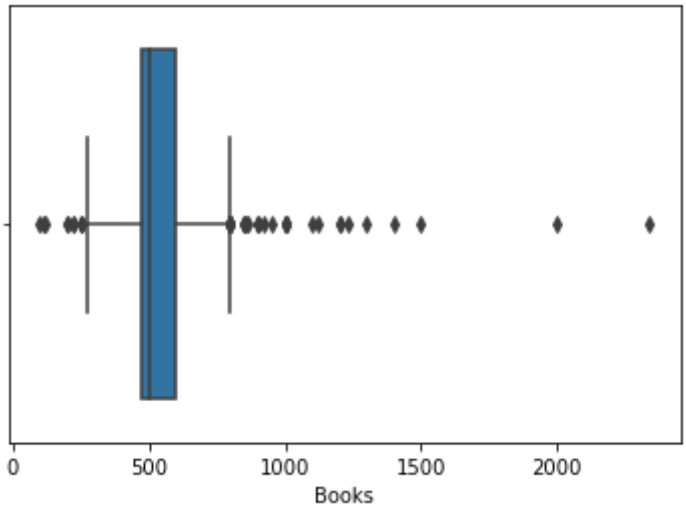
- Distribution & Boxplot of Room.Board



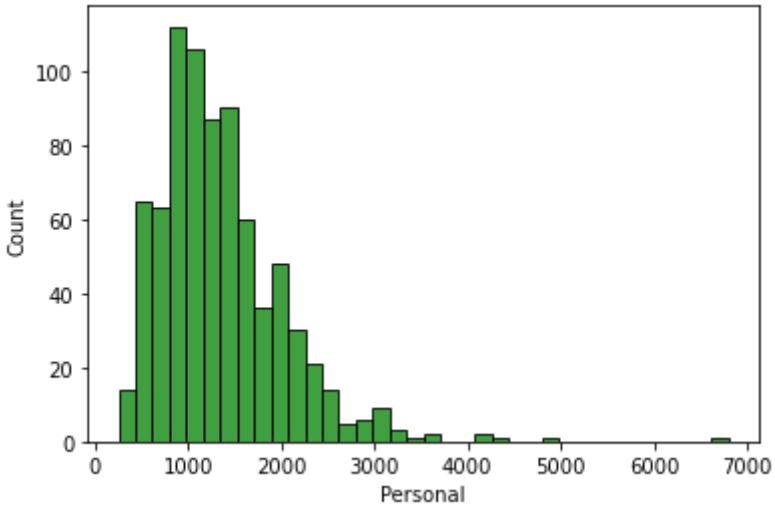


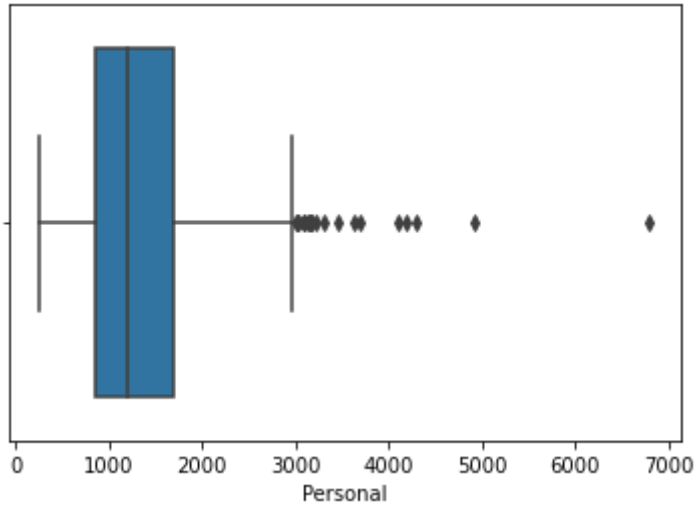
- Distribution & Boxplot of Books



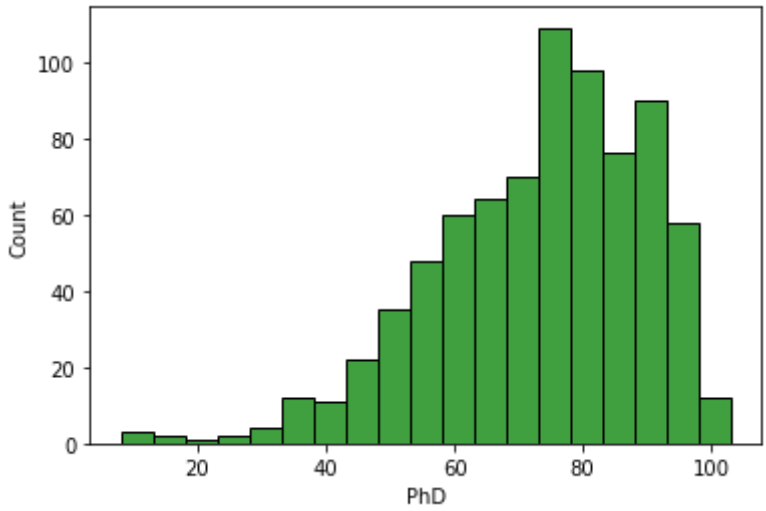


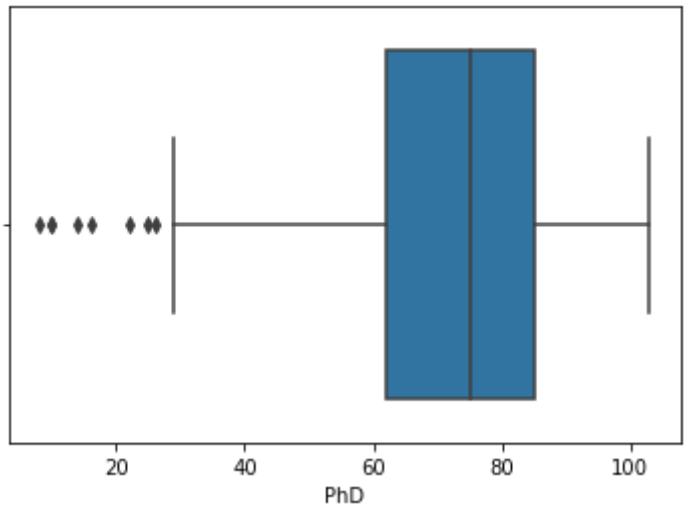
- Distribution & Boxplot of Personal



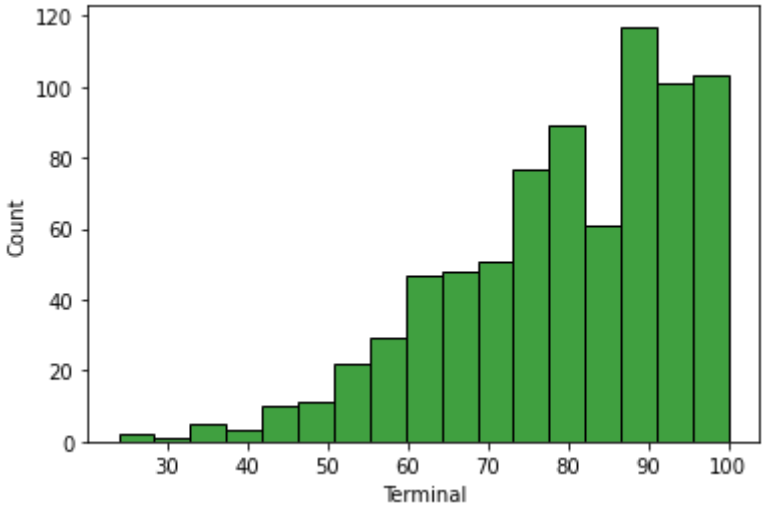


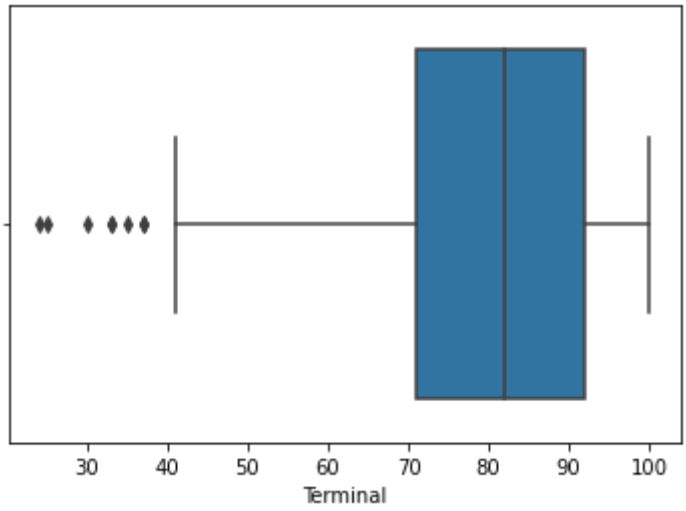
- Distribution & Boxplot of PhD



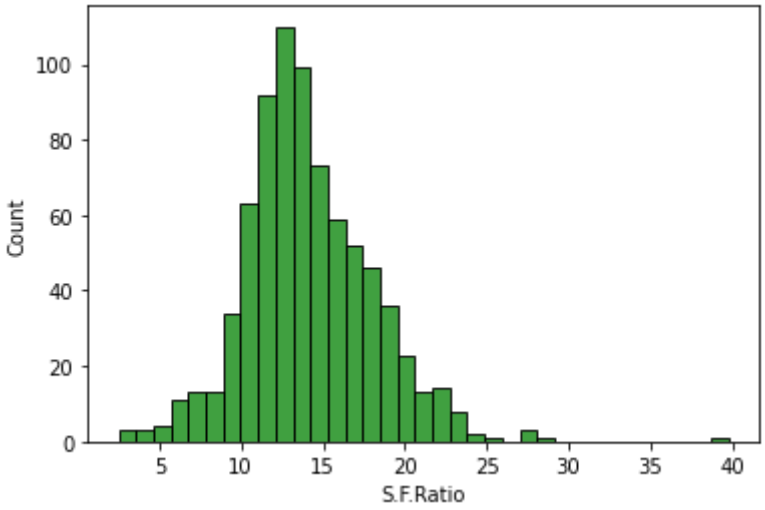


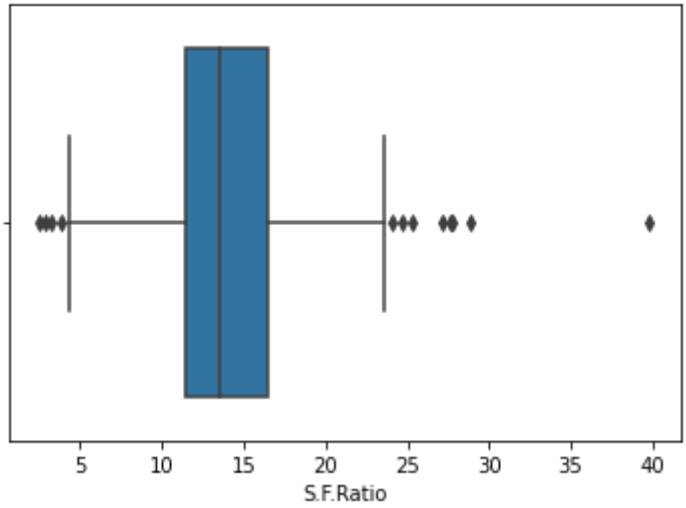
- Distribution & Boxplot of Terminal



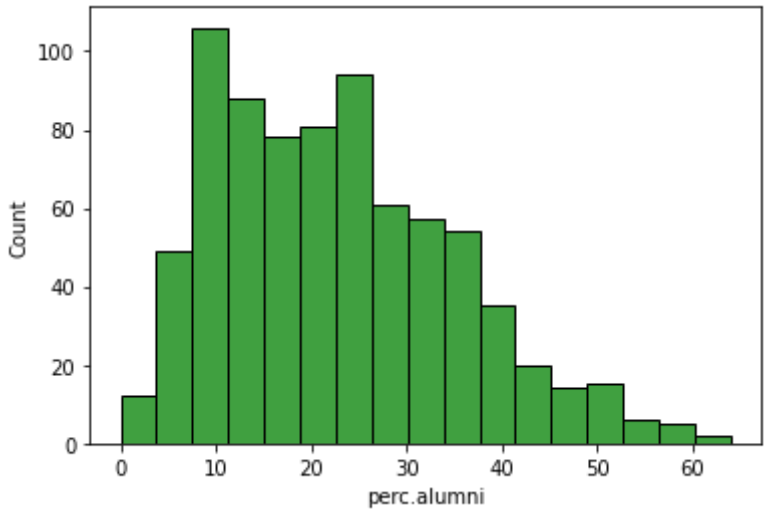


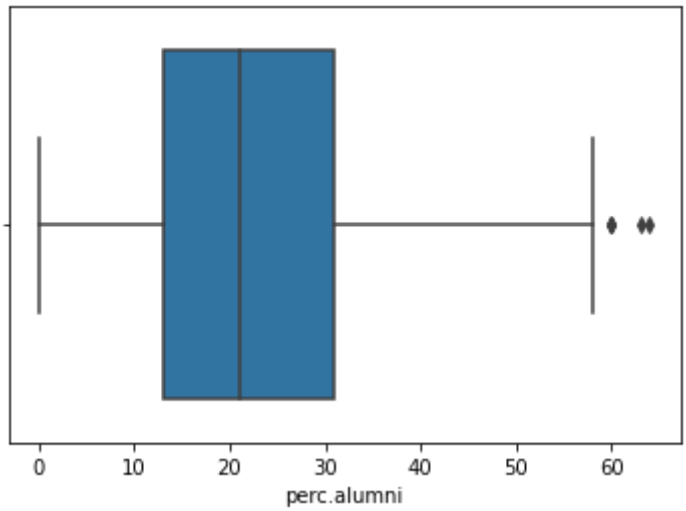
- Distribution & Boxplot of S.F.Ratio



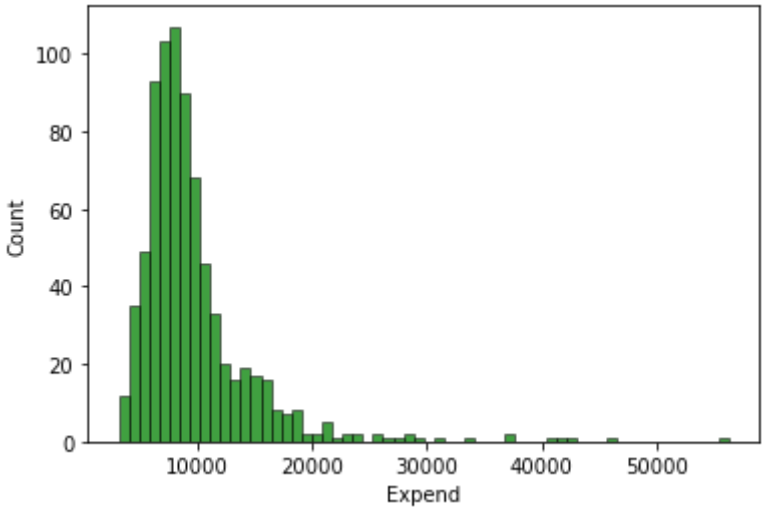


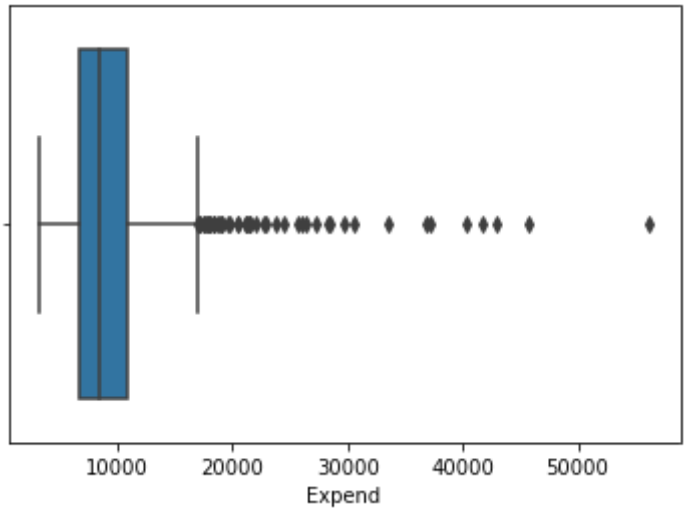
- Distribution & Boxplot of perc.alumni



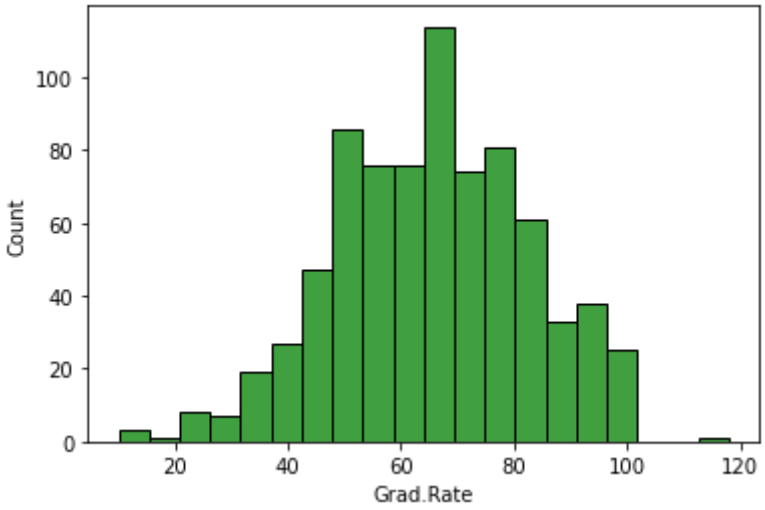


- Distribution & Boxplot of Expend





- Distribution & Boxplot of Grad.Rate



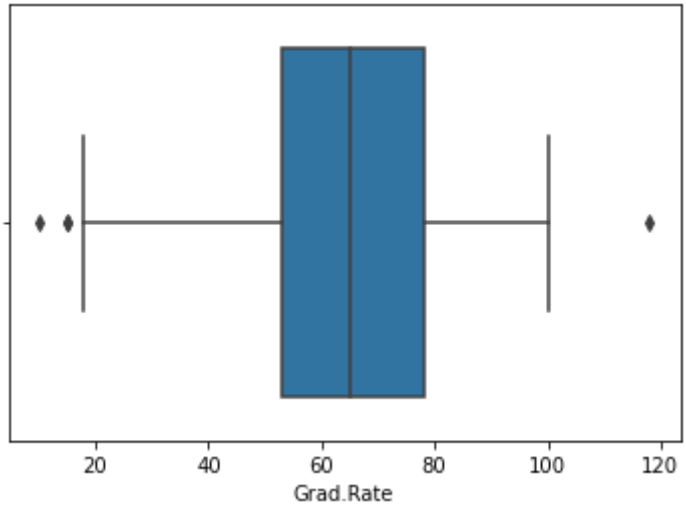


Fig 2: Univariate Analysis, Distribution and Bloxplot of all Numeric Fields

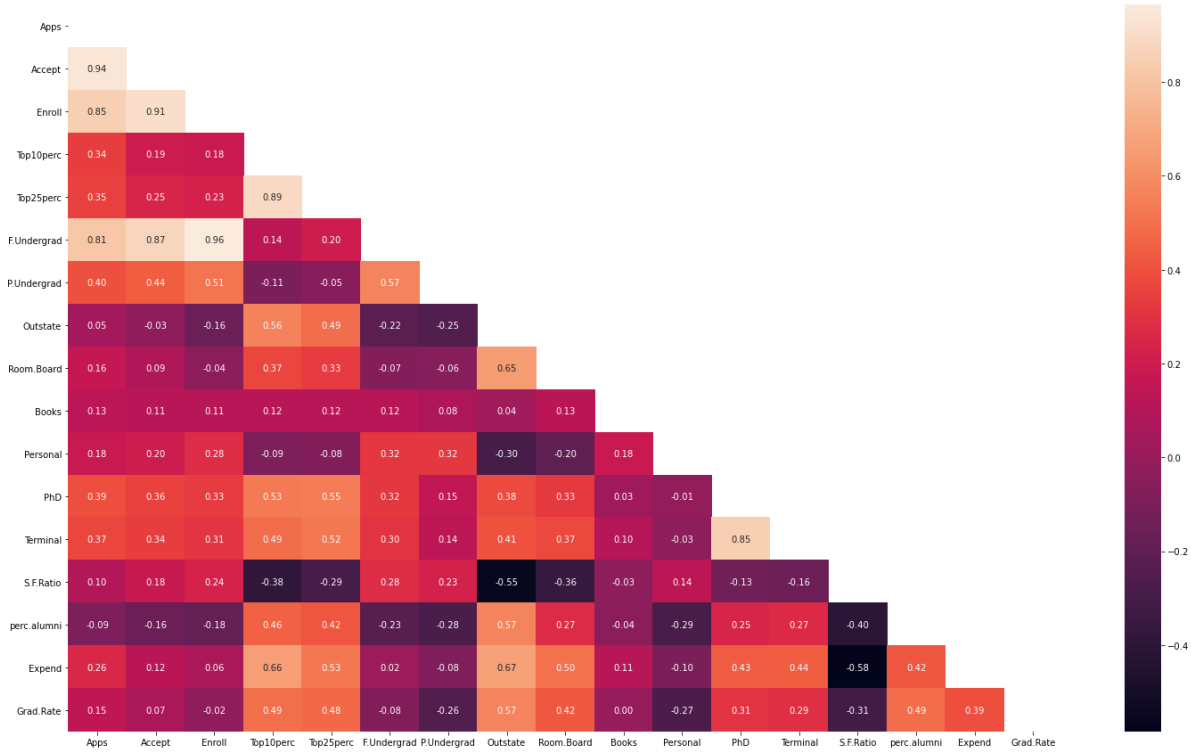


Fig 3 : Heat Map which shows Correlation of numeric variables with other numeric variables

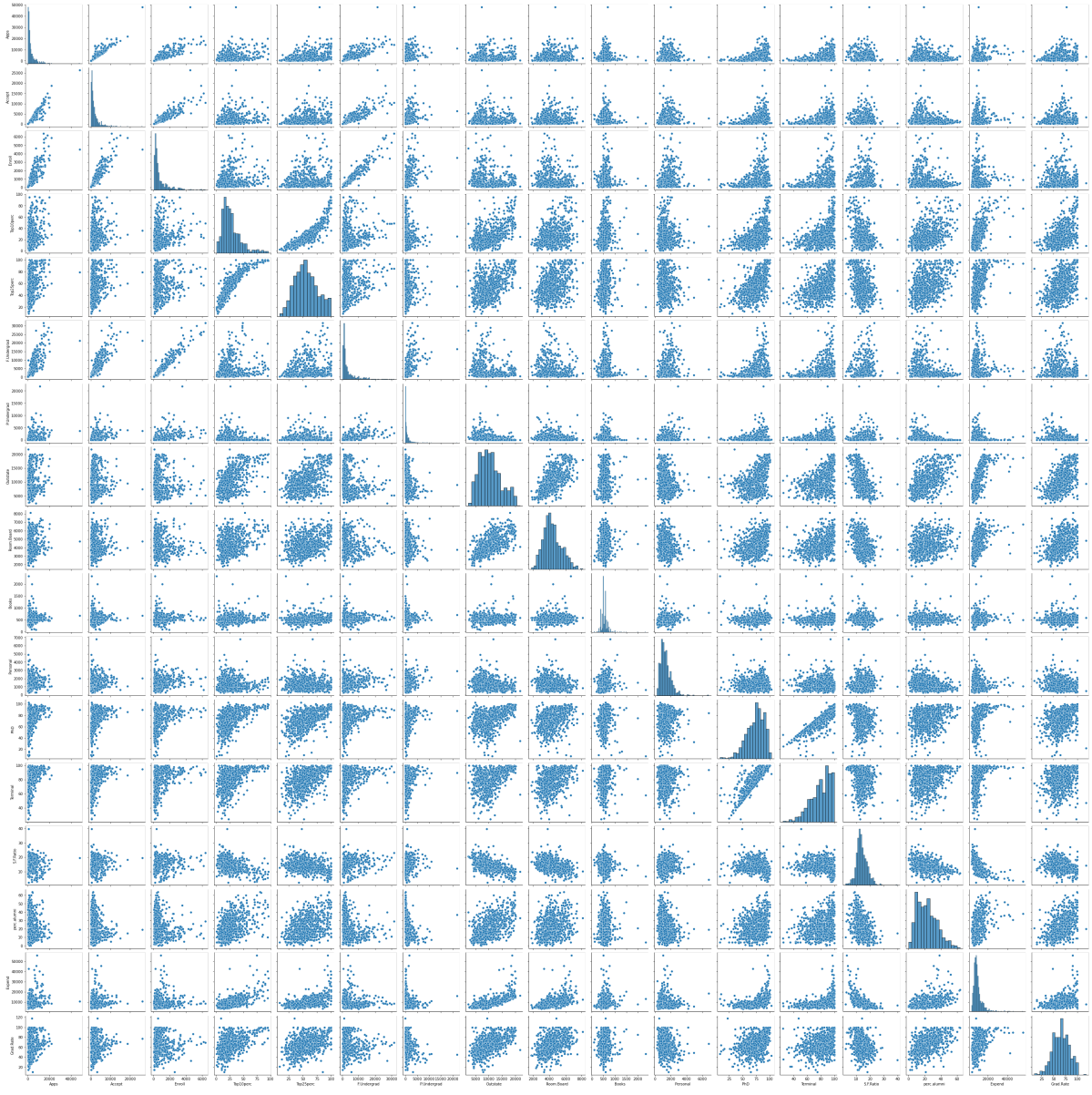


Fig 4: Pair Plot of numeric variables with other numeric variables

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

Scaling is necessary in this case as Dataset has features with different "weights".

Here we will be using Z-score Scaling Method

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Roor
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.96
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.55
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.21

Table 9: Scaled Dataset

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

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

P value is 0.0, p-value is small. We can reject the null hypothesis and agree that there is atleast one pair of variables in the data wihich are correlated hence PCA is recommended.

KMO Test

The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

MSA = 0.8131251200373522 is greater than 0.5 & MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components. Hence PCA is recommended.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Roon
Apps	1.00	0.94	0.85	0.34	0.35	0.82	0.40	0.05	0.17
Accept	0.94	1.00	0.91	0.19	0.25	0.88	0.44	-0.03	0.09
Enroll	0.85	0.91	1.00	0.18	0.23	0.97	0.51	-0.16	-0.04
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.37
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33
F.Undergrad	0.82	0.88	0.97	0.14	0.20	1.00	0.57	-0.22	-0.07
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	-0.06
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.66
Room.Board	0.17	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.66	1.00
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
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33
Terminal	0.37	0.34	0.31	0.49	0.53	0.30	0.14	0.41	0.38
S.F.Ratio	0.10	0.18	0.24	-0.39	-0.30	0.28	0.23	-0.56	-0.36
perc.alumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.27
Expend	0.26	0.12	0.06	0.66	0.53	0.02	-0.08	0.67	0.50
Grad.Rate	0.15	0.07	-0.02	0.50	0.48	-0.08	-0.26	0.57	0.43

Table 10: Variance-Covariance Matrix of the Scaled Data

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Roon
Apps	1.00	0.94	0.85	0.34	0.35	0.81	0.40	0.05	0.16
Accept	0.94	1.00	0.91	0.19	0.25	0.87	0.44	-0.03	0.09
Enroll	0.85	0.91	1.00	0.18	0.23	0.96	0.51	-0.16	-0.04

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

Table 11: New Correlation Matrix of the Scaled Data

We see that the Covariance Matrix and the Correlation matrix of the Scaled Data are identical to each other.

2.4. 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]

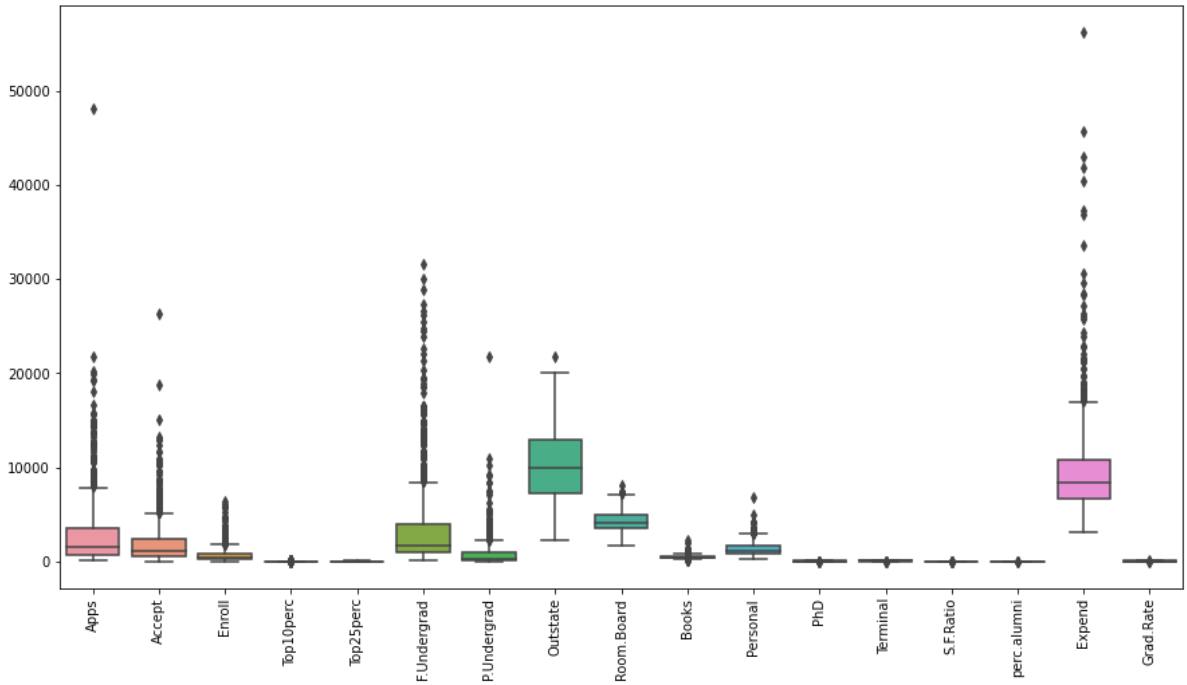


Fig 5 : Boxplot of all the Attributes of Original Dataset to check for Outliers

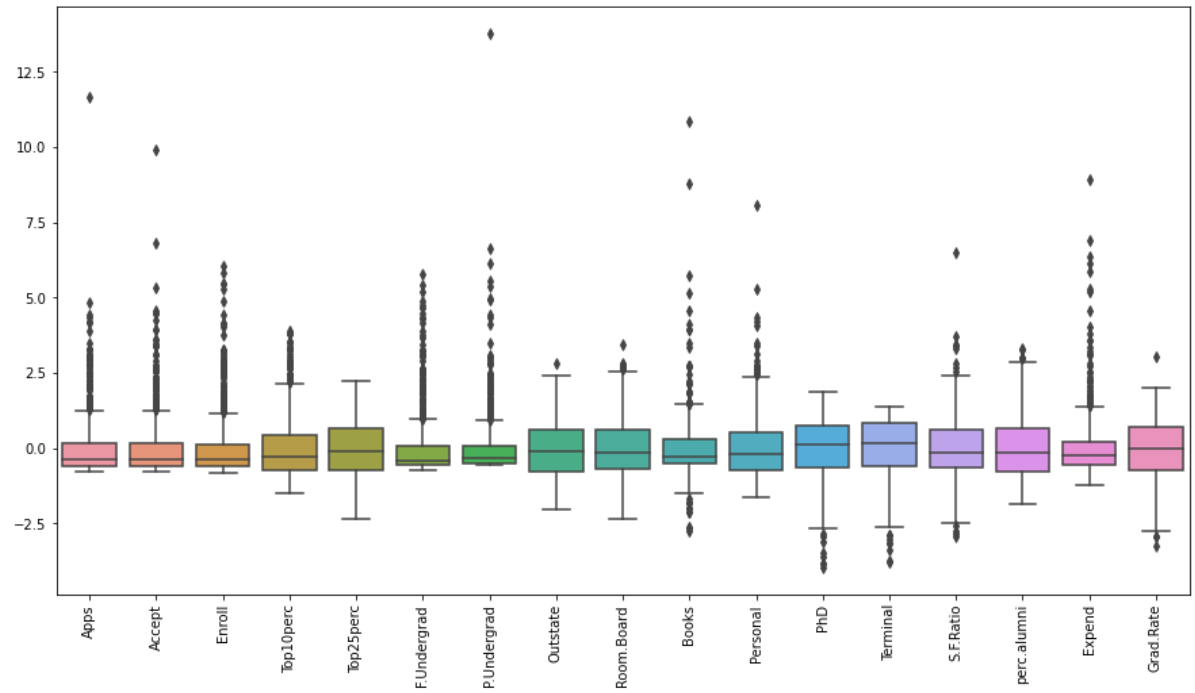


Fig 6 : Boxplot of all the Attributes of Scaled Dataset to check for Outliers

We can Inter that upon scaling the Dataset, We do not treat the Outliers but we ensure that the Attributes Means are all 0 and Variances 1.

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

Eigen Values(first 17 PCAs)
225.09375072515908
86.26531354873661
53.41961939946679
48.583579842205545
45.2510908430626
38.238826731597094
28.652708330647695
26.11287254512243
21.497086394781938
16.228140490471755
10.705379520702754
8.171740563657291
6.9831428512805225
4.282207511322702
1.7935302083639977
1.1255166831627155
1.4297919314602925e-14

Table 12: Eigen Values without using sklearn

	0	1	2	3	4
0	-0.022919+0.000000j	-0.018999+0.000000j	-0.015433+0.000000j	0.029136+0.000000j	-0.028335+0.00
1	-0.007299+0.000000j	-0.056117+0.000000j	0.081335+0.000000j	-0.117802+0.000000j	0.042004+0.000
2	0.007012+0.000000j	-0.008281+0.000000j	-0.005125+0.000000j	-0.022541+0.000000j	-0.015303+0.00
3	0.060368+0.000000j	0.033981+0.000000j	0.019605+0.000000j	0.057081+0.000000j	-0.014407+0.00
4	-0.015208+0.000000j	-0.025249+0.000000j	0.079143+0.000000j	0.057174+0.000000j	0.018533+0.000
...	...	...	...	...	...
772	-0.041957+0.000000j	-0.036483+0.000000j	-0.025293+0.000000j	-0.010533+0.000000j	0.032081+0.000
773	0.011221+0.000000j	-0.016373+0.000000j	-0.006555+0.000000j	-0.024601+0.000000j	0.017332+0.000
774	-0.003378+0.000000j	0.003735+0.000000j	0.004142+0.000000j	-0.001260+0.000000j	-0.019143+0.00
775	0.085569+0.000000j	0.054337+0.000000j	0.096314+0.000000j	-0.007773+0.000000j	-0.044637+0.00
776	-0.009372+0.000000j	-0.013009+0.000000j	-0.063788+0.000000j	-0.014326+0.000000j	-0.035182+0.00

Table 12: Eigen Vector without using sklearn

Using sklearn finding Eigenvalues and Eigenvectors

	Eigen Values
1	5.45052162
2	4.48360686
3	1.17466761
4	1.00820573
5	0.93423123
6	0.84849117
7	0.6057878
8	0.58787222

Table 14: Eigen Values using sklearn

	0	1	2	3	4	5	6	7	8	
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809	0
2	-0.063092	-0.101249	-0.082986	0.035056	-0.024148	-0.061393	0.139682	0.046599	0.148967	0
3	0.281311	0.267817	0.161827	-0.051547	-0.109767	0.100412	-0.158558	0.131291	0.184996	0
4	0.005741	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919	-(
5	-0.016237	0.007535	-0.042558	-0.052693	0.033092	-0.043454	-0.191199	-0.030000	0.162755	0
6	-0.042486	-0.012950	-0.027693	-0.161332	-0.118486	-0.025076	0.061042	0.108529	0.209744	-(
7	-0.103090	-0.056271	0.058662	-0.122678	-0.102492	0.078890	0.570784	0.009846	-0.221453	0

Table 15: Eigen Vector using sklearn

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

	0	1	2	3	4	5	6	7
0	-1.592855	0.767334	-0.101074	-0.921749	-0.743975	-0.298306	0.638443	-0.879386
1	-2.192402	-0.578830	2.278798	3.588918	1.059997	-0.177137	0.236753	0.046925
2	-1.430964	-1.092819	-0.438093	0.677241	-0.369613	-0.960592	-0.248276	0.308740
3	2.855557	-2.630612	0.141722	-1.295486	-0.183837	-1.059508	-1.249356	-0.147694
4	-2.212008	0.021631	2.387030	-1.114538	0.684451	0.004918	-2.159220	-0.624413
...	...	...	...	...	...	...	...	...
772	-3.328458	1.220255	-0.383388	0.108555	0.776996	0.309429	-0.165021	0.347435
773	0.199389	-0.686689	0.051564	0.562269	0.375191	0.373343	0.848453	0.626515
774	-0.732561	-0.077235	-0.000406	0.054316	-0.516021	0.468014	-1.317492	-0.128288
775	7.919327	-2.068329	2.073564	0.852054	-0.947755	-2.069937	0.083328	-0.552586
776	-0.469508	0.366661	-1.328915	-0.108023	-1.132176	0.839893	1.307313	0.627410

Table 16: Component Loadings

	0	1	2	3	4	5	6	7
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.103090
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950	-0.056271
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693	0.058662
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486	-0.102492
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076	0.078890
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042	0.570784
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529	0.009846
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744	-0.221453
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692	0.213293
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790	-0.232661
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096	-0.077040
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477	-0.012161
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	0.219259	-0.083605
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321	0.678524
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584	-0.054159
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944	-0.005336

Table 17: Data of principal Component into Data Frame with Original Features

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]

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* - *0.04* S.F.Ratio + 0.21 *perc.alumni* + *0.32* Expend + 0.25 * Grad.Rate

Explicit Form of the first PC

Similarly we can write Expilicit form for all PCs

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 Values of Eigenvalues
1	36.16518654
2	50.02519442
3	58.60797601
4	66.41376459
5	73.68413143
6	79.82785731
7	84.43140822
8	88.62689079
9	92.08076784
10	94.68809801
11	96.40810155
12	97.72103237
13	98.84299449
14	99.53100486
15	99.81916637
16	100.
17	100.

Table 18: Cumulative Values of Eigenvalues

- From the Above Cumulative Value table of Eigenvalues, we see that from the First 8 Components we have 88% total variation explained. Hence we consider only 8 Components.
- Thus with the help of Cumulative Value table of Eigenvalues, we will be able to decide on the optimum number of principal components
- Eigenvectors represent directions

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]

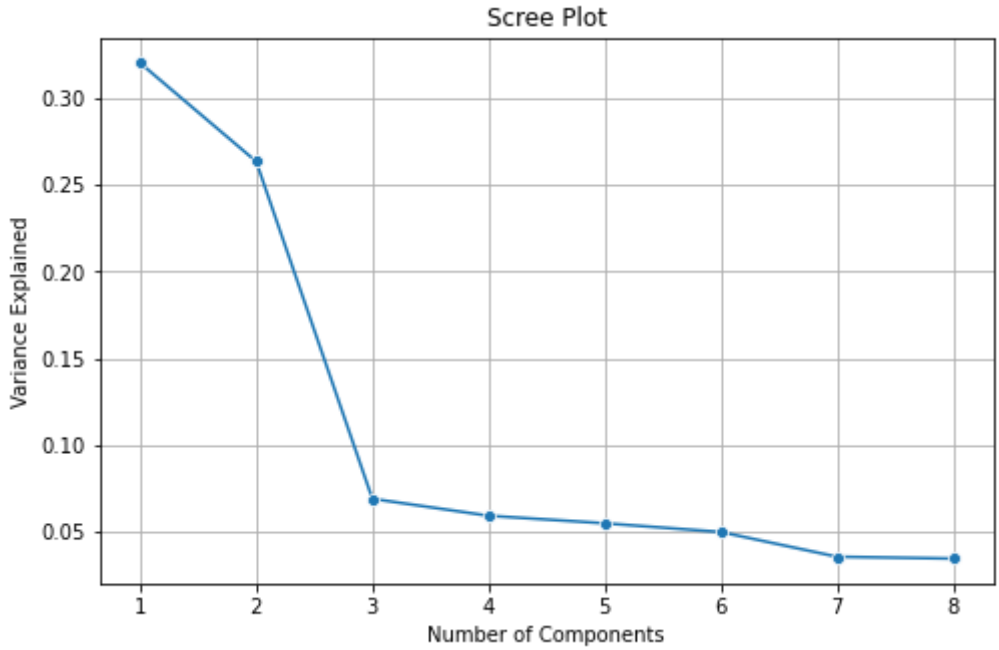


Fig 7: Scree Plot

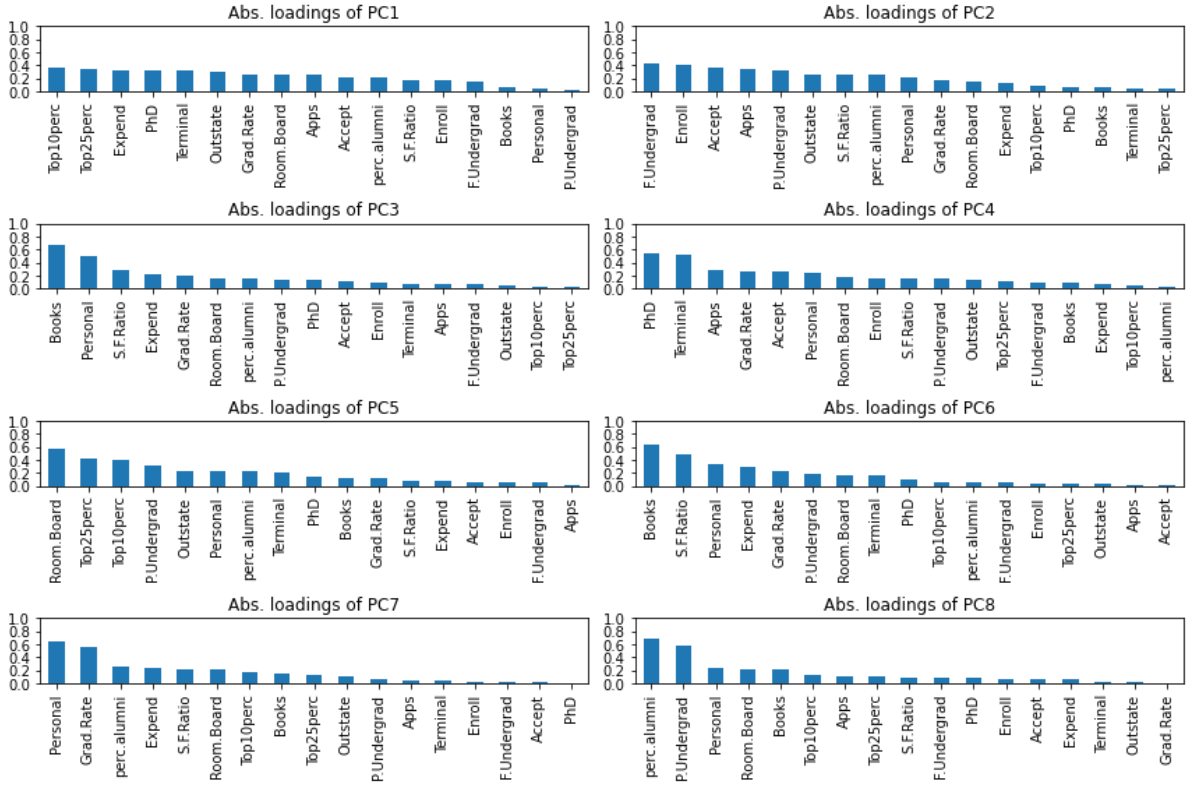


Fig 8: Absolute loadings of all PCs

With help of PCA we have been able to reduce 18 numeric features into 8 components which is able to explain 88.6% of variance in the data

Using the components, additional rules can be derived and analyzed.

Unsupervised learning like clustering can further be applied on the data to segment the Colleges / Universities based on the components & can further be analyzed.

The END