Advance Stats Business Report

- **Name**: Hemant Patidar
- **Batch**: PGPDSBA Online Sep_A 2021
- **Date**: 11/12/2021

Table of Contents

List of Figures

List of Tables

Salary Data Analysis

Executive Summary

Salary is hypothesized 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.

Introduction

Purpose of our exercise would be to check if any of the factors affect the salary in significant way. Occupation and Educational qualification are 2 different factors against which we have salary data of individuals.

Data Description

- Education – Qualification (High School Graduate, Bachelor and Doctorate)
- Occupation – Administrative and clerical, Sale, Professional or Specialty and Executive or managerial
- Salary – Salary received by individual

Visualization of Data

Education vs. Salary

To visualize average salary data and outliers, we have used boxplot and bar plot… At initial level (by looking at the graphs) we can say the salary increases with level of education. Hence,

Salary of person with Doctorate has higher salary compared to bachelor's and bachelors' have higher salary compared to high-school graduate.

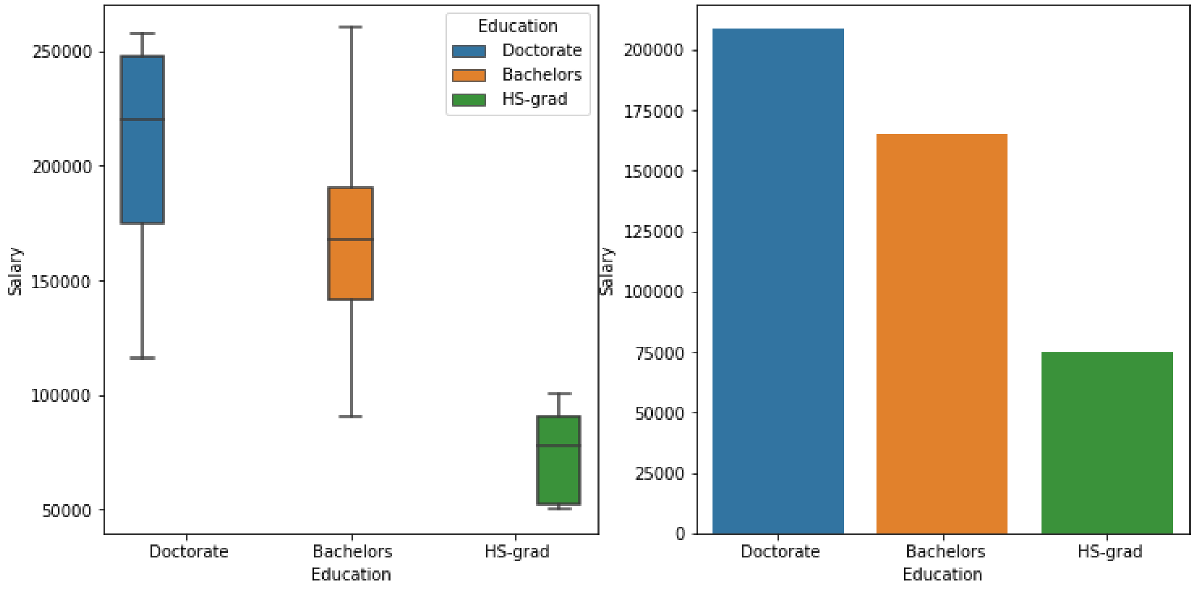


Figure I - Education vs. Salary Plots

Occupation vs. Salary

With below EDA, we can infer that executive-managerial occupation have higher salary bucket, and administrative-clerical have widespread in salary.

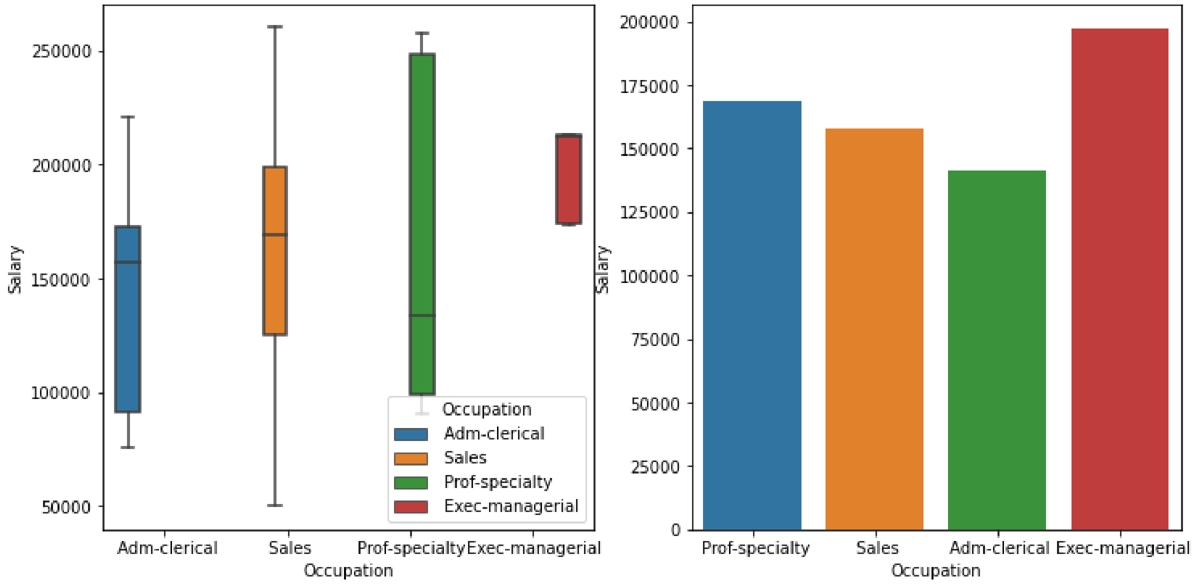


Figure II -Occupation vs. Salary Plots

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

One-way ANOVA tests the null hypothesis

$H0$: $\mu1 = \mu2 = \mu3 = \cdots = \mu c$

Against the alternate

Ha: At least one population mean is different from the rest

Assumptions for ANOVA –

- The populations have normal distribution
- The populations have equal variances
- Observations are independent (Randomly collected)

NULL and Alternate hypothesis for Education –

NULL Hypothesis (H0) - Average salary against different educational qualifications is equal

Alternate Hypothesis (Ha) – Average salary of at least one educational qualification is significantly different from the other qualifications.

NULL and Alternate hypothesis for Occupation –

NULL Hypothesis (H0) - Average salary against different occupations is equal

Alternate Hypothesis (Ha) – Average salary against at least one occupation is significantly different from the other occupation.

1.2 Perform one-way ANOVA for Education with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results

We would assume that our populations are following ANOVA assumptions (As stated above).

Deciding the significance level (α): *We are selecting as 0.5*

Final ANOVA Table (Education) –

Formula to feed into OLS algorithm – *Salary ~ C(Education)*

	Degree of Freedom	Sum of Square	Mean Sum of Square	F-Stat	PR(>F)
C(Education)	2	102695500000	51347730000	30.95628	0.00000001257709
Residual	37	61372560000	1658718000		

Table 1 - One-way ANOVA table (Education)

The P-values is way lesser than 0.05 (significance level), thus we have evidence to reject NULL hypothesis. We can say that –

**Average salary of at least one educational qualification is different from the rest.

1.3 Perform one-way ANOVA for variable Occupation with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results

We would assume that our populations are following ANOVA assumptions (As stated above).

Deciding the significance level (α): *We are selecting as 0.5*

Final ANOVA Table (Occupation) –

Formula to feed into OLS algorithm – *Salary ~ C(Occupation)*

	Degree Of Freedom	Sum Of Square	Mean Sum of Square	F-Stat	PR(>F)
C(Occupation)	3	11258780000	3752928000	0.884144	0.458508
Residual	36	152809200000	4244701000		

Table 2 - One-way ANOVA table (Occupation)

The P-values > 0.05 (significance level), thus we have no evidence to reject NULL hypothesis. We can say that –

**No significance difference found in average salary of different occupation.

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

We have evidence to reject NULL hypothesis when we performed one-way ANOVA for Education with respect to Salary,

To find out which class means are significantly different, we can perform Post-hoc (Tukey) test.

Stating hypothesis for Post-hoc test –

Since, we have only 3 pairs to be considered for Education (HS-grad, Bachelor and Doctorate), Post-hoc would be testing NULL hypothesis

H0 → $\mu1 = \mu2$ and $\mu1 = \mu3$ and $\mu2 = \mu3$

Against, Alternate hypothesis Ha → $\mu1 \neq \mu2$ or $\mu1 \neq \mu3$ or $\mu2 \neq \mu3$

$\mu1$ – Average salary of high school graduate

$\mu2$ – Average salary of Bachelor

$\mu3$ – Average salary of Doctorate

Comparison of Means table (Tukey HSD) –

group1	group2	meandiff	p-adj	lower	upper	reject
Bachelors	Doctorate	43274.0667	0.0146	7541.144	79006.9894	TRUE
Bachelors	HS-grad	-90114.1556	0.001	-132035	-48193.1153	TRUE
Doctorate	HS-grad	-133388.2222	0.001	-174815	-91961.3569	TRUE

Table 3 - Mean comparison table (Post-Hoc Analysis)

P-Adj is lower than significance level in all comparison, hence we can conclude that…

***Each education qualification has significantly different average salary than another qualification.

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

Interaction is a quantification of association of two factors. If one factor behaves differently at different levels of one or more factors, an interaction effect is said to exist.

It occurs when the pattern of the cell means in one row (going across columns) varies from the patterns of cell means in other rows.

Interaction Plot

We can infer from the plot that the lines are not all parallel, since they are interacting at a point, it indicates that Education and Occupation have interaction effect.

And mean of salary bachelors and doctorate are same in administrative-clerical and sales occupation.

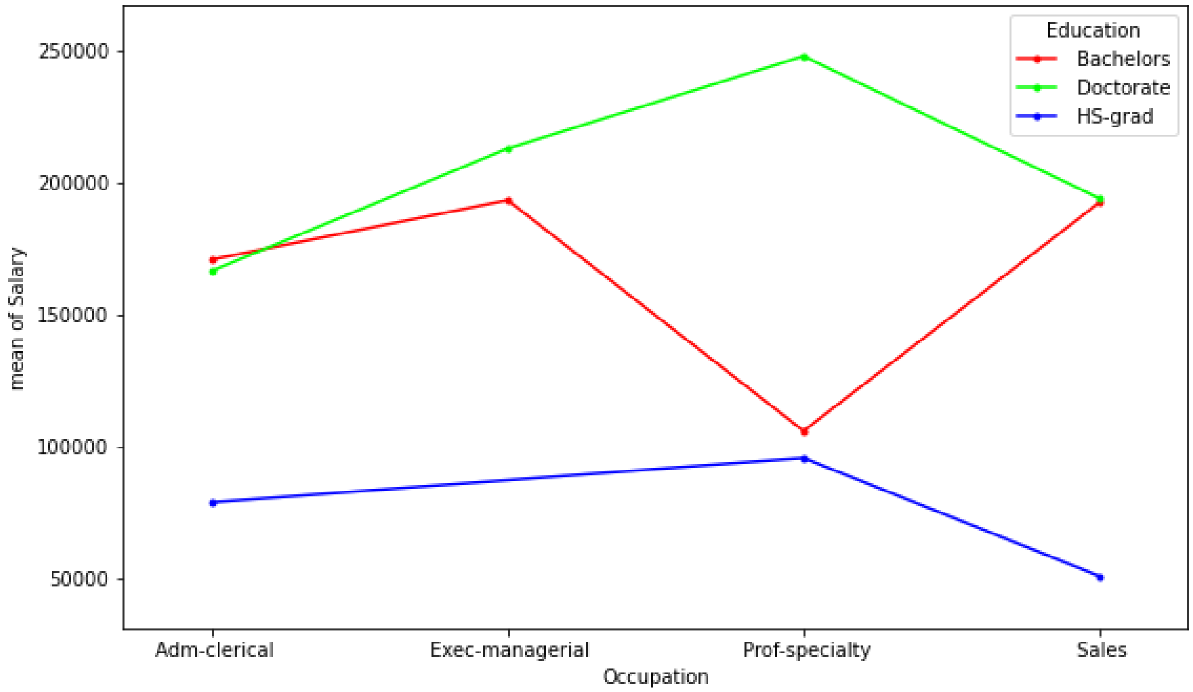


Figure III - Interaction Plot (Education vs. Occupation)

By above plot, we can conclude that high school graduate students do not reach at executive-managerial position.

1.6 Perform a two-way ANOVA based on the Education and Occupation (along with their interaction Education*Occupation) with the variable ‘Salary’. State the null and alternative hypotheses and state your results. How will you interpret this result?

Before performing two-way ANOVA for Education and Occupation with interaction, we would assume that the populations are following ANOVA assumptions (As stated above)

Defining the hypothesis

Two-way ANOVA tests NULL & Alternate hypothesis for each factor and interaction between them, it can be written as –

NULL Hypothesis (H0)

- The population means of Education groups are equal
- The population means of Occupation groups are equal
- There is no interaction between Education and Occupation

Alternate Hypothesis (Ha)

- The population mean of at least one Education group is different
- The population mean of at least one Occupation group is different
- There is significant interaction between both factor (Education and Occupation)

We would assume the significance level (α) → 0.05

Final Two-way ANOVA Table –

Formula to feed into OLS algorithm –

Salary ~ C(Education) + C(Occupation) + C(Education):C(Occupation)

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2	102695500000	51347730000	72.211958	0.000000000005466264
C(Occupation)	3	5519946000	1839982000	2.587626	0.0721158
C(Education):C(Occupation)	6	36349090000	6058182000	8.519815	0.000022325
Residual	29	20621020000	711069700		

Table 4 - Two-way ANOVA table

From the table above, we can conclude –

- P-value against Education is less than significance level (0.05), hence we have evidence to reject NULL hypothesis and alternate is true. Population means across different education level are not equal.
- P-value against Occupation is more than significance level (0.05), and we don't have any evidence to reject NULL hypothesis. Hence, we can say population means for different occupation are equal.
- P-value against both factors' interaction is less than significance level, hence we can infer that Education and Occupation has significant interaction with each other.

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

As we analysed by plots and ANOVA tests, we can infer that –

- Mean salary gets affected by educational qualification, and all 3 different qualification have different average salary.
- People tends to have good salary if they are highly Educated.
- Occupation does not much impact on how much salary people would get.
- Education and Occupation has significant interaction.

Education - Post 12th, PC Analysis

Executive Summary

University wise graduation rate, application received, and other metrics were collected for 777 universities, and how many of top students joining them.

Introduction

Purpose of our exercise would be to perform EDA and withdraw useful insights and perform PCA to seek most accurate data using lower dimensions.

Data Description

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

Describing the Data

	count	mean	std	min	25%	50%	75%	max
Apps	777	3,001.64	3,870.20	81	776	1,558	3,624	48,094
Accept	777	2,018.80	2,451.11	72	604	1,110	2,424	26,330
Enroll	777	779.97	929.18	35	242	434	902	6,392
Top10perc	777	27.56	17.64	1	15	23	35	96
Top25perc	777	55.80	19.80	9	41	54	69	100
F.Undergrad	777	3,699.91	4,850.42	139	992	1,707	4,005	31,643
P.Undergrad	777	855.30	1,522.43	1	95	353	967	21,836
Outstate	777	10,440.67	4,023.02	2,340	7,320	9,990	12,925	21,700
Room.Board	777	4,357.53	1,096.70	1,780	3,597	4,200	5,050	8,124
Books	777	549.38	165.11	96	470	500	600	2,340
Personal	777	1,340.64	677.07	250	850	1,200	1,700	6,800
PhD	777	72.66	16.33	8	62	75	85	103
Terminal	777	79.70	14.72	24	71	82	92	100
S.F.Ratio	777	14.09	3.96	3	12	14	17	40
perc.alumni	777	22.74	12.39	-	13	21	31	64
Expend	777	9,660.17	5,221.77	3,186	6,751	8,377	10,830	56,233
Grad.Rate	777	65.46	17.18	10	53	65	78	118

Table 5 - Describing University Data

From above, we can infer for our 17 variables –

- **Apps** has a mean 3001.64 with standard deviation 3870.2, with a minimum to maximum range is 81 to 48094.
- **Accept** has a mean 2018.8 with standard deviation 2451.11, with a minimum to maximum range is 72 to 26330.
- **Enroll** has a mean 779.97 with standard deviation 929.18, with a minimum to maximum range is 35 to 6392.
- **Top10perc** has a mean 27.56 with standard deviation 17.64, with a minimum to maximum range is 1 to 96.
- **Top25perc** has a mean 55.8 with standard deviation 19.8, with a minimum to maximum range is 9 to 100.
- **F.Undergrad** has a mean 3699.91 with standard deviation 4850.42, with a minimum to maximum range is 139 to 31643.
- **P.Undergrad** has a mean 855.3 with standard deviation 1522.43, with a minimum to maximum range is 1 to 21836.
- **Outstate** has a mean 10440.67 with standard deviation 4023.02, with a minimum to maximum range is 2340 to 21700.
- **Room.Board** has a mean 4357.53 with standard deviation 1096.7, with a minimum to maximum range is 1780 to 8124.
- **Books** has a mean 549.38 with standard deviation 165.11, with a minimum to maximum range is 96 to 2340.
- **Personal** has a mean 1340.64 with standard deviation 677.07, with a minimum to maximum range is 250 to 6800.
- **PhD** has a mean 72.66 with standard deviation 16.33, with a minimum to maximum range is 8 to 103. *(There cannot be more than 100% faculty so anything more than 100 would be incorrect data)*

- **Terminal** has a mean 79.7 with standard deviation 14.72, with a minimum to maximum range is 24 to 100.
- **S.F.Ratio** has a mean 14.09 with standard deviation 3.96, with a minimum to maximum range is 3 to 40.
- **perc.alumni** has a mean 22.74 with standard deviation 12.39, with a minimum to maximum range is - to 64.
- **Expend** has a mean 9660.17 with standard deviation 5221.77, with a minimum to maximum range is 3186 to 56233.
- **Grad.Rate** has a mean 65.46 with standard deviation 17.18, with a minimum to maximum range is 10 to 118 *(There cannot be more than 100% graduation rate so anything more than 100 would be incorrect data)*

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

We have 777 rows and 18 columns in dataframe with no null or duplicated values.

Univariate analysis would tell us individual field’s spread, skewness, and outliers’ presence in data.

Univariate Analysis –

Apps – Number of Application received

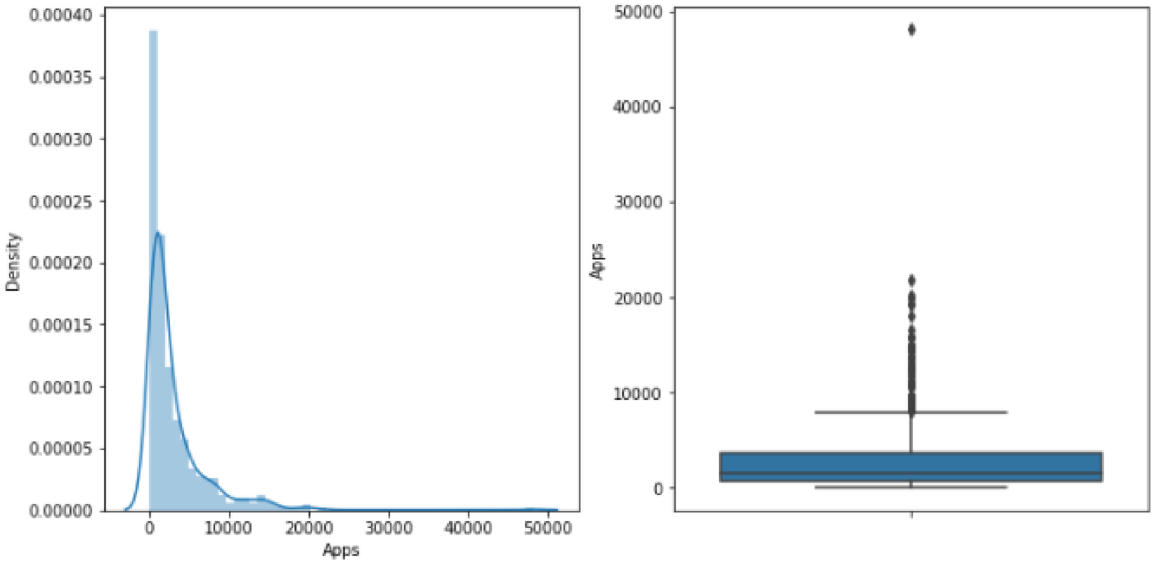


Figure IV - Univariate Analysis (Apps)

As evident by above figures, Apps data have outliers and data seems positively skewed. Also, we could see that an extreme number of applications (~50,000) received by certain university.

Accept – Number of applications accepted

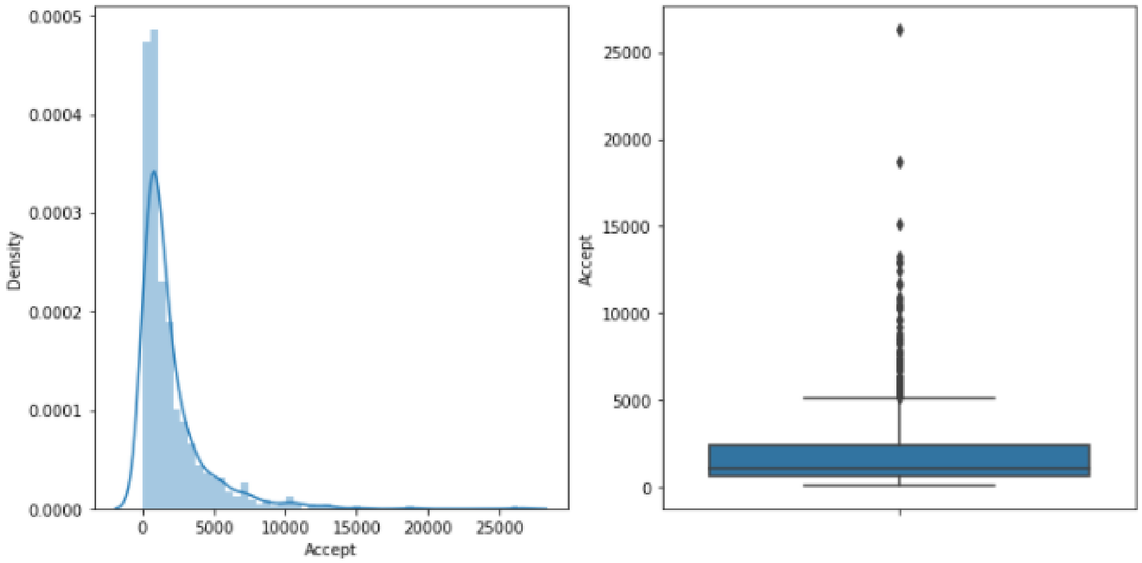


Figure V - Univariate Analysis (Accept)

Some universities accepted more application than the rest, and the data seems positively skewed with number of outlier present in Accept field.

Enroll – Number of new students enrolled

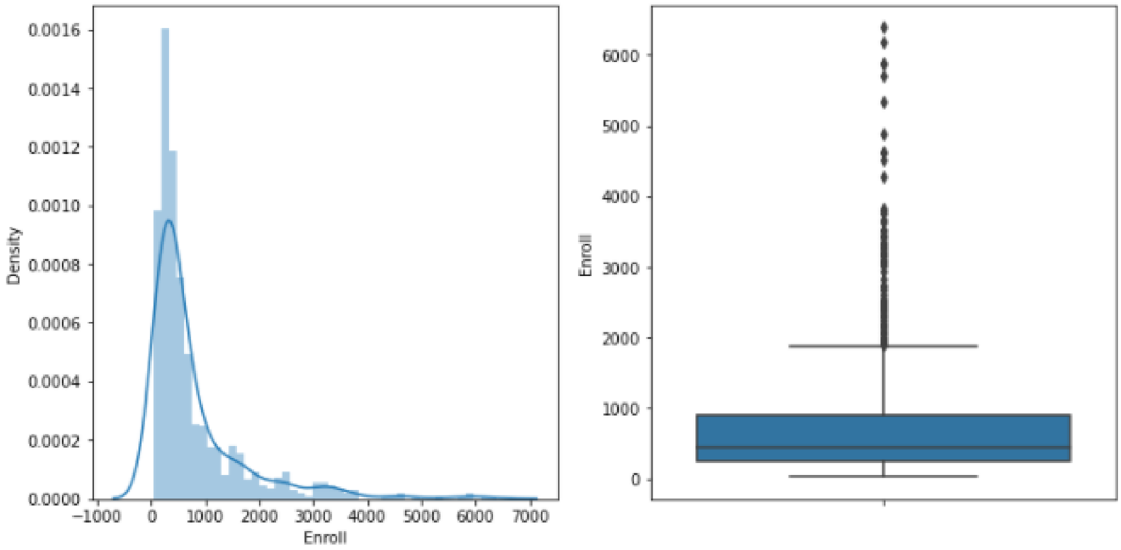


Figure VI - Univariate Analysis (Enroll)

The population of enrolled students has outliers and data seems positively skewed because of that.

Top10perc – Percentage of new students from top 10% of Higher Secondary class

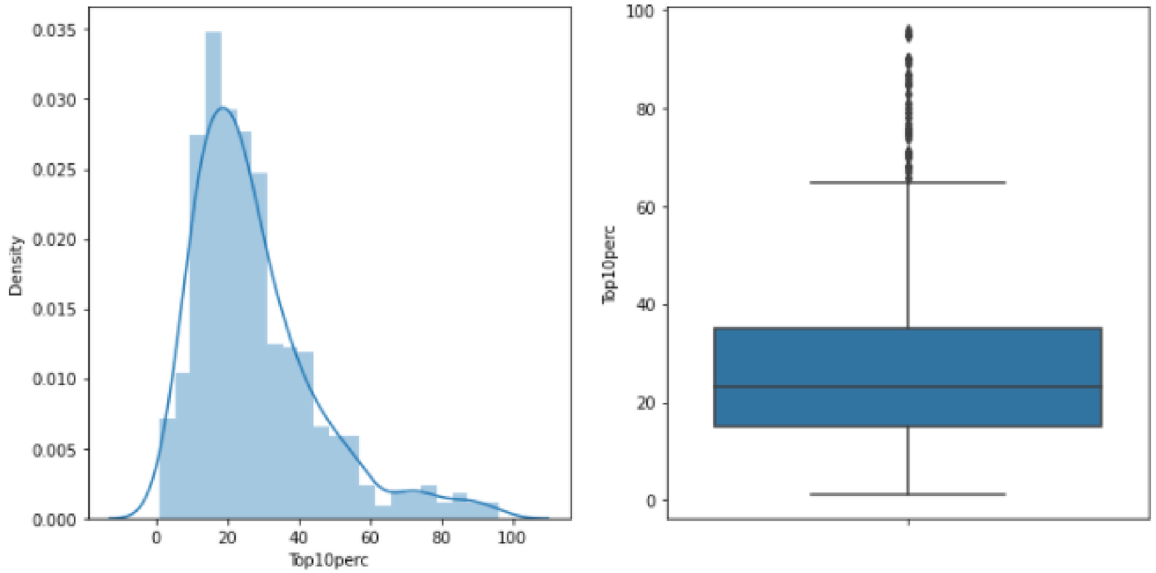


Figure VII - Univariate Analysis (Top 10 Percent)

Data seems normally distributed across all universities, but it has outliers up till almost 100% students joining same university which have been in top 10 % in higher secondary class.

Top25perc – Percentage of new students from top 25% of Higher Secondary class

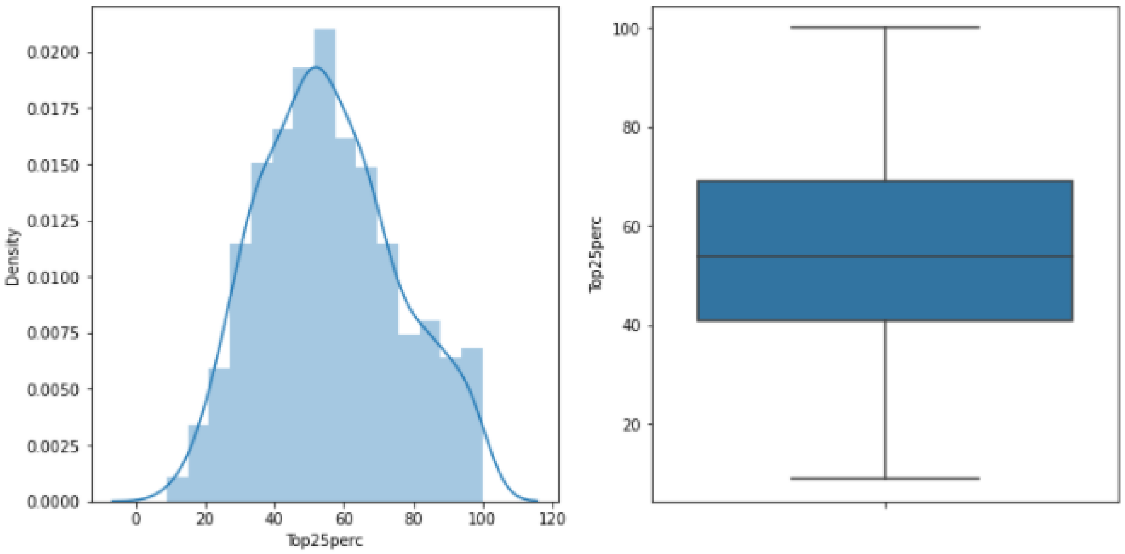


Figure VIII - Univariate Analysis (Top 25 Percent)

The data of top 25% new students of higher secondary class have joined university seems evenly distributed and has no outliers.

F.Undergrad – Number of full-time undergraduate students

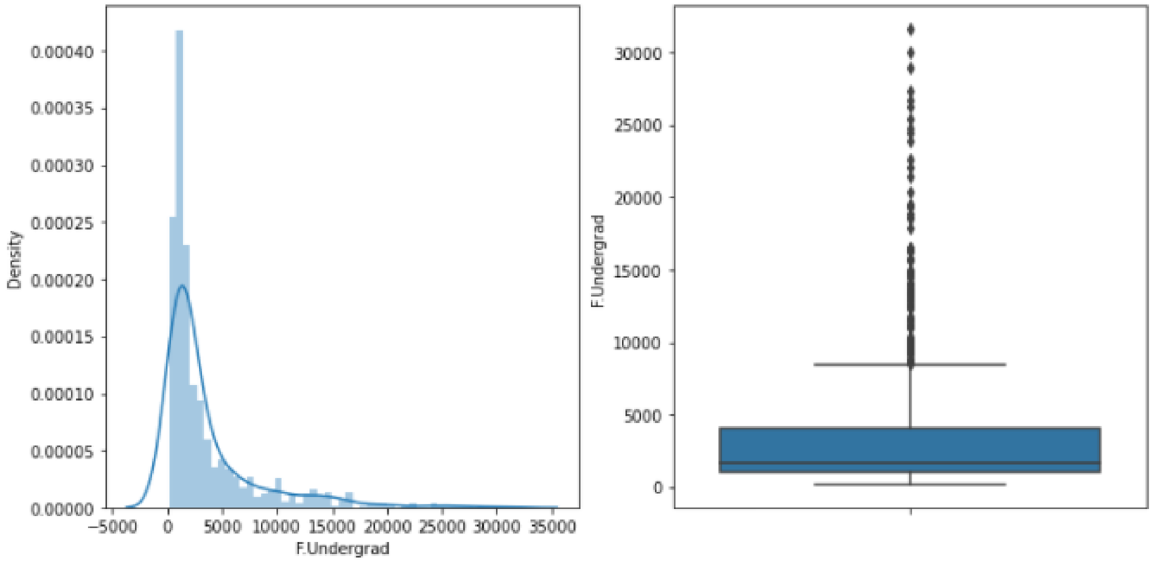


Figure IX - Univariate Analysis (Full time Undergraduate)

The data seems positively skewed and have outliers. The number of full-time undergraduate students in most of the university ranges between 0 – 5000.

P.Undergrad – Number of part-time undergraduate students

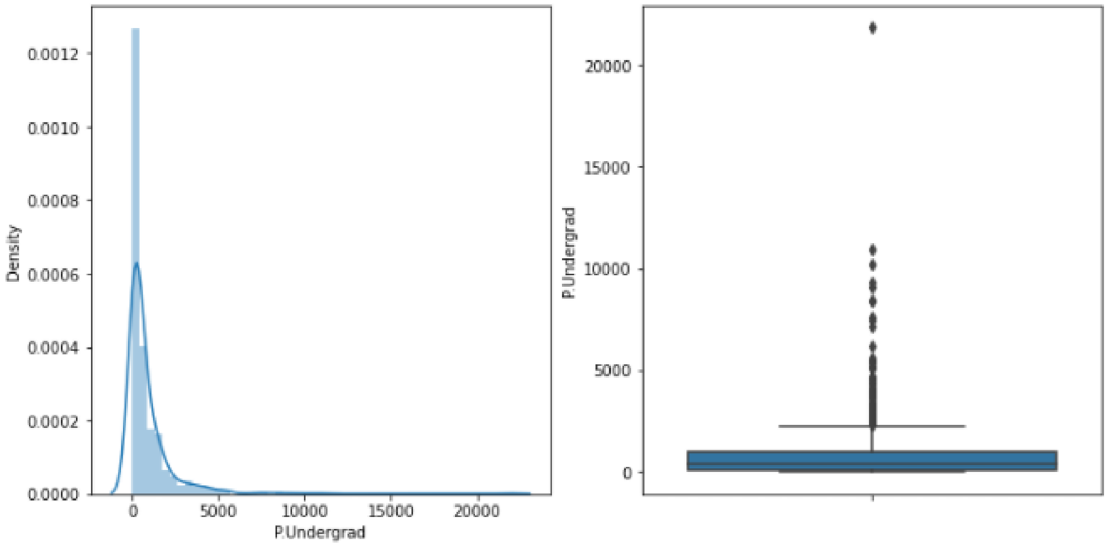


Figure X - Univariate Analysis (Part time Undergraduate)

A university has extreme number of part-time undergraduates (more than 20,000), most of the data seems in range 0 to 3000 with several outlier present in data.

Outstate – Number of students for whom the particular college or university is Out-of-state tuition

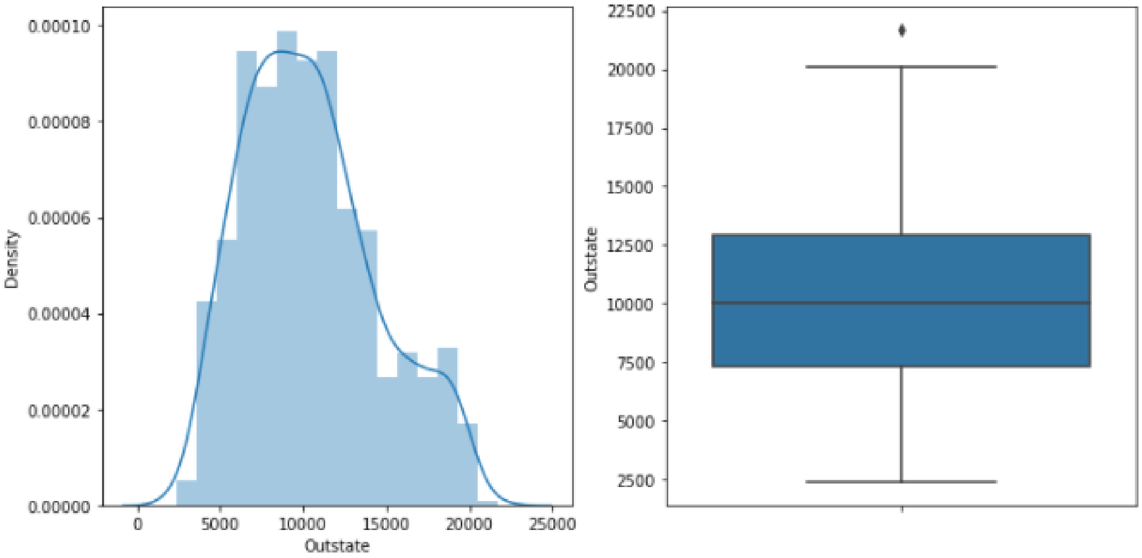


Figure XI - Univariate Analysis (Outstate)

The data seems normally distributed, and only one outlier present in outstate data. The median and mean looked around 10,000.

Room.Board – Cost of Room and board

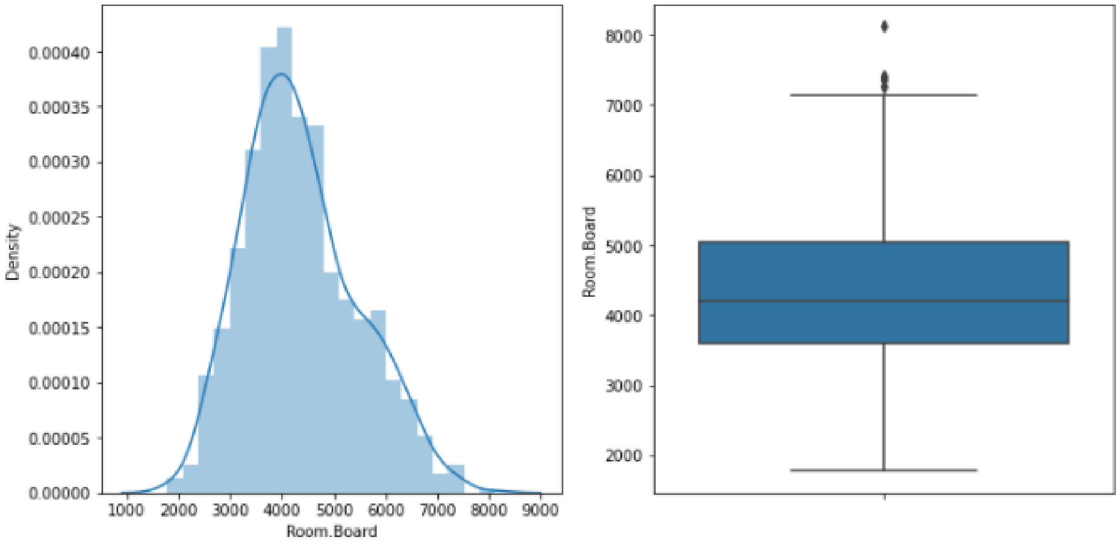


Figure XII - Univariate Analysis (Room & Board Cost)

The cost of room and board seem evenly distributed and have some outliers with an extreme value around 8,000.

Books – Estimated book costs for a student

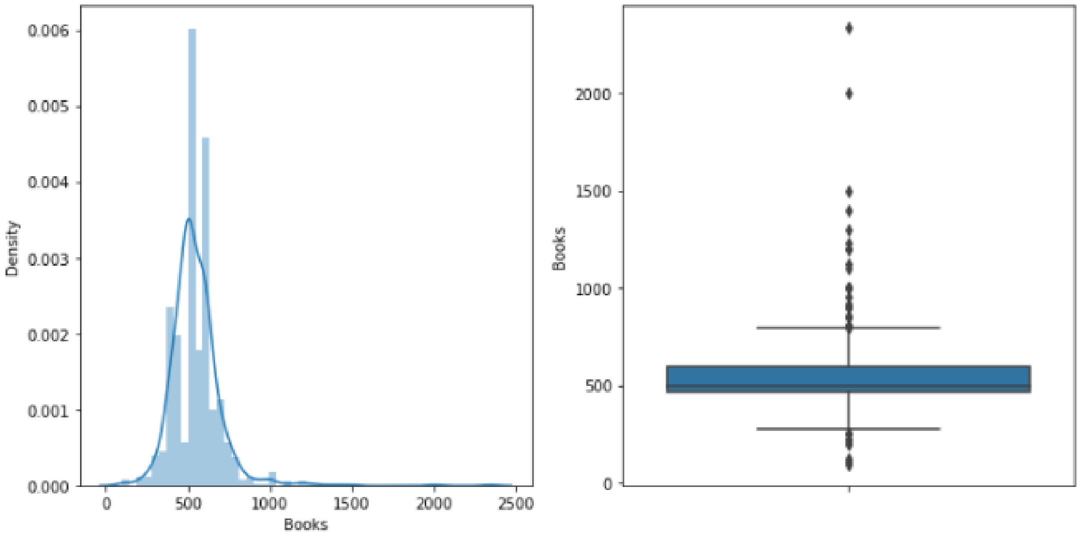


Figure XIII - Univariate Analysis (Books)

There seem to have some incorrect data (with negative cost), and several outliers presents as evident by box plot.

Personal – Estimated personal spending for a student

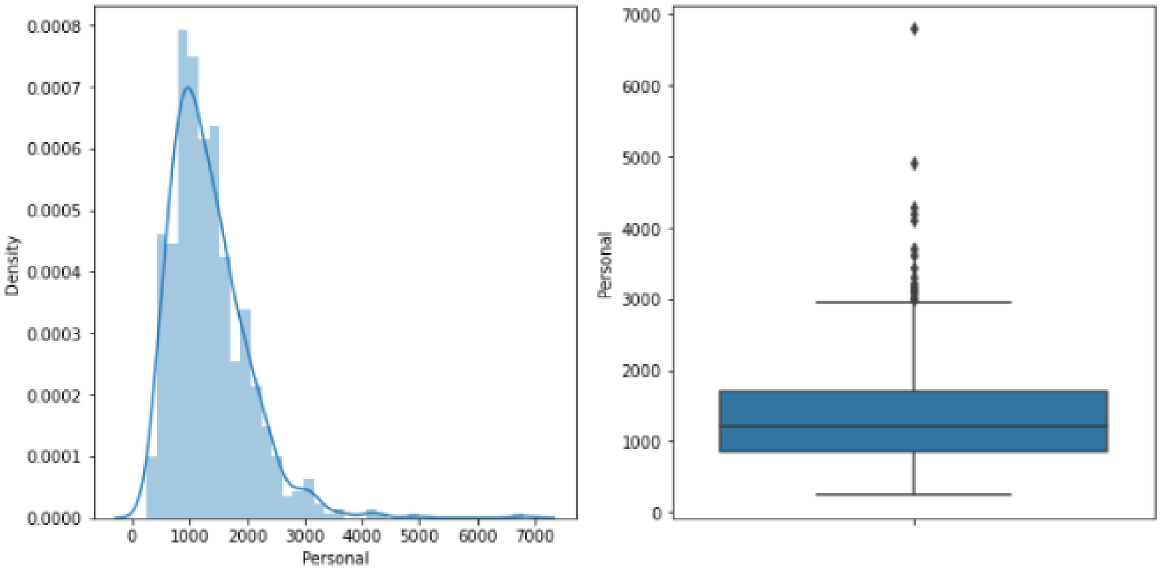


Figure XIV - Univariate Analysis (Personal Spending)

Spending data seems positively skewed and have outliers, with an extreme value around 7,000.

PhD – Percentage of faculties with Ph.D.'s

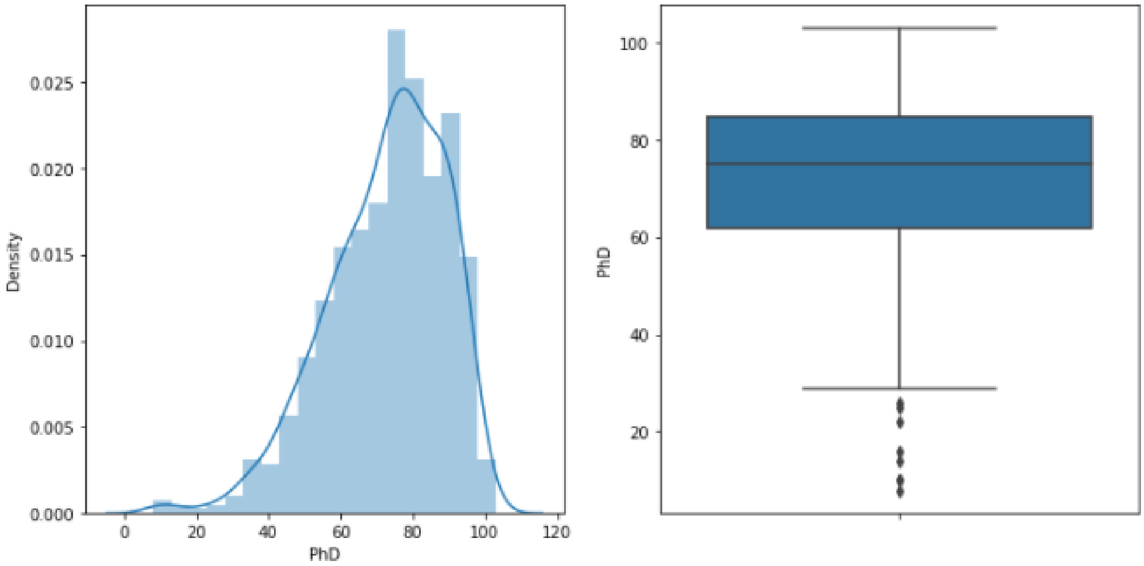


Figure XV - Univariate Analysis (PhD Faculties %)

There are number of universities having relatively less Ph.D.'s faculties compared to others, and those can be identified by outliers. We seem to have a university with 100% of Ph.D. faculties.

Terminal – Percentage of faculties with terminal degree

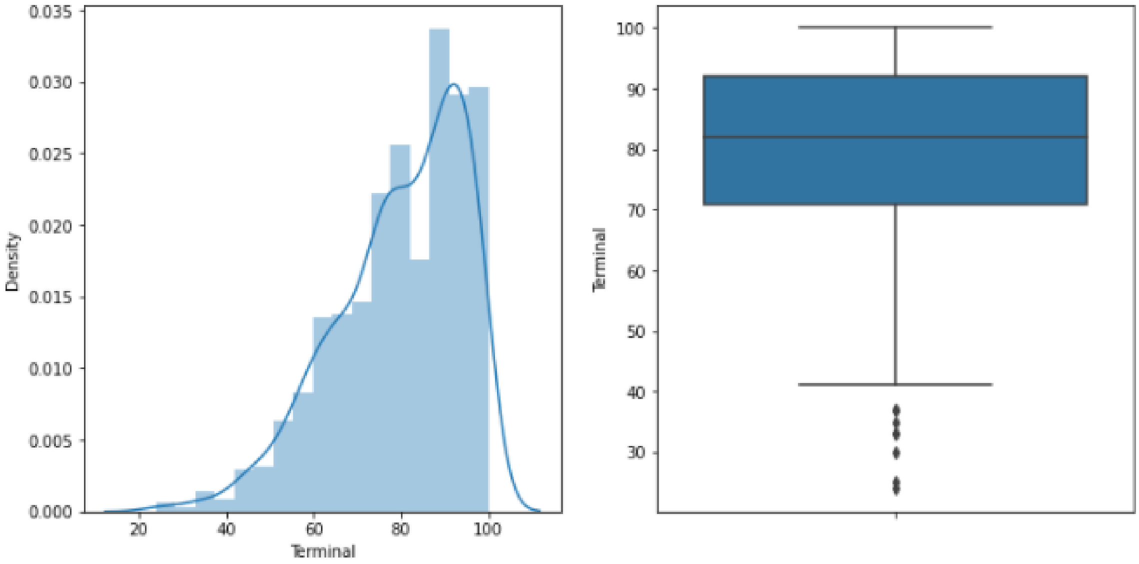


Figure XVI - Univariate Analysis (Terminal)

Data has outliers at lower limit, so we can infer that some universities have relatively less number of faculties with terminal degree and this is evident by histogram that most of them have faculties around 80 – 100% with terminal degree.

S.F.Ratio – Student/faculty ratio

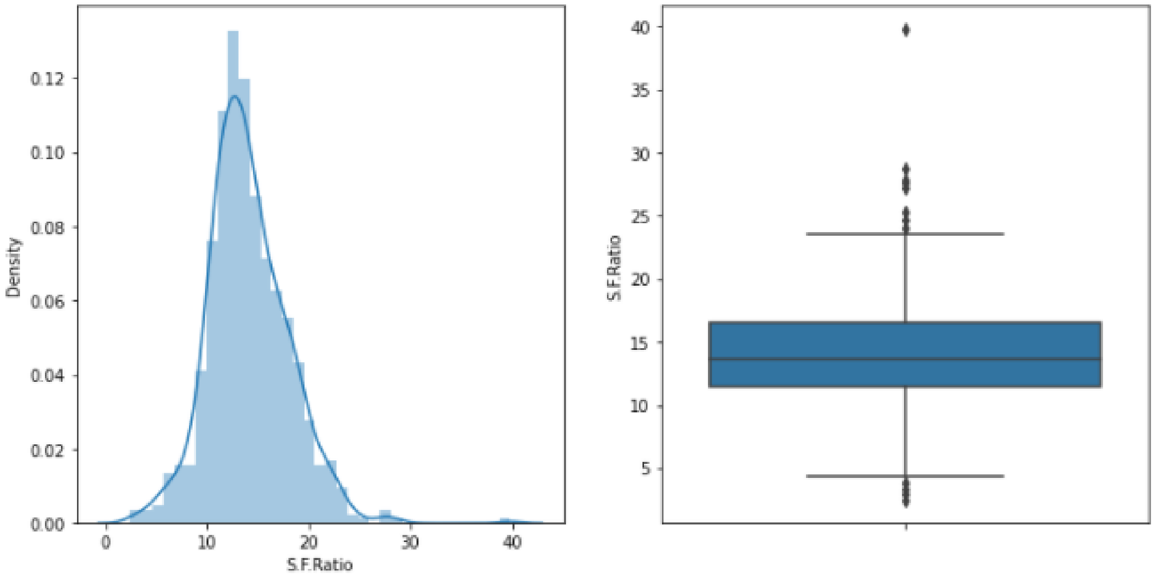


Figure XVII - Univariate Analysis (Student to Faculty Ratio)

Student to faculty ratio seems following normally distribution. By box plot, we are seeing there are outliers on upper and lower limit sides.

perc.alumni – Percentage of alumni who donate

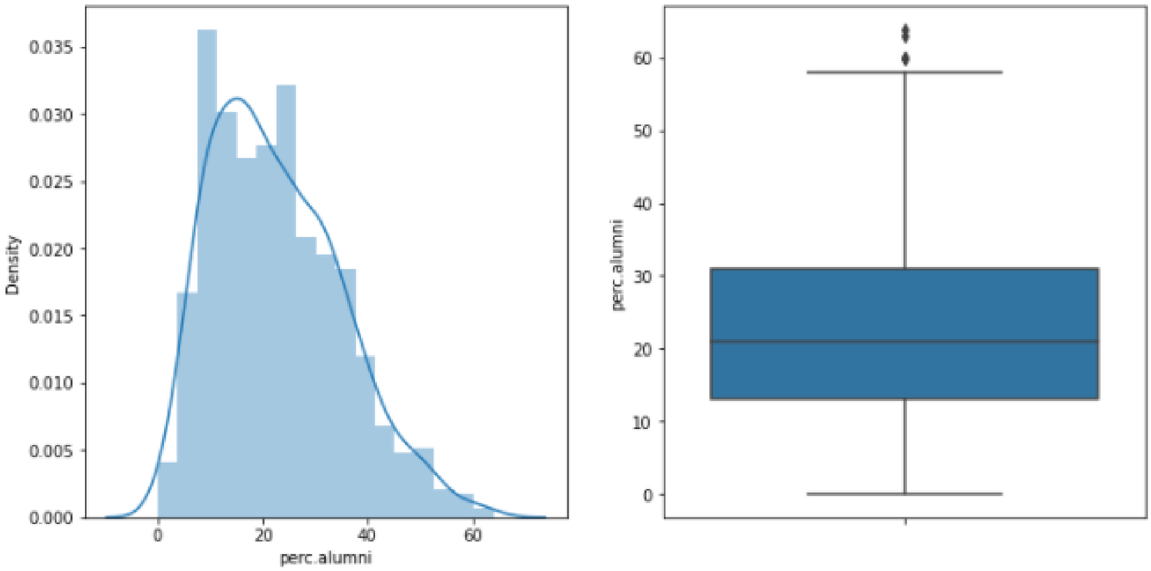


Figure XVIII - Univariate Analysis (Percentage of Alumni)

The data seems normally distributed and most of the data varies from 0 -60% and identified couple of outliers.

Expend – The Instructional expenditure per student

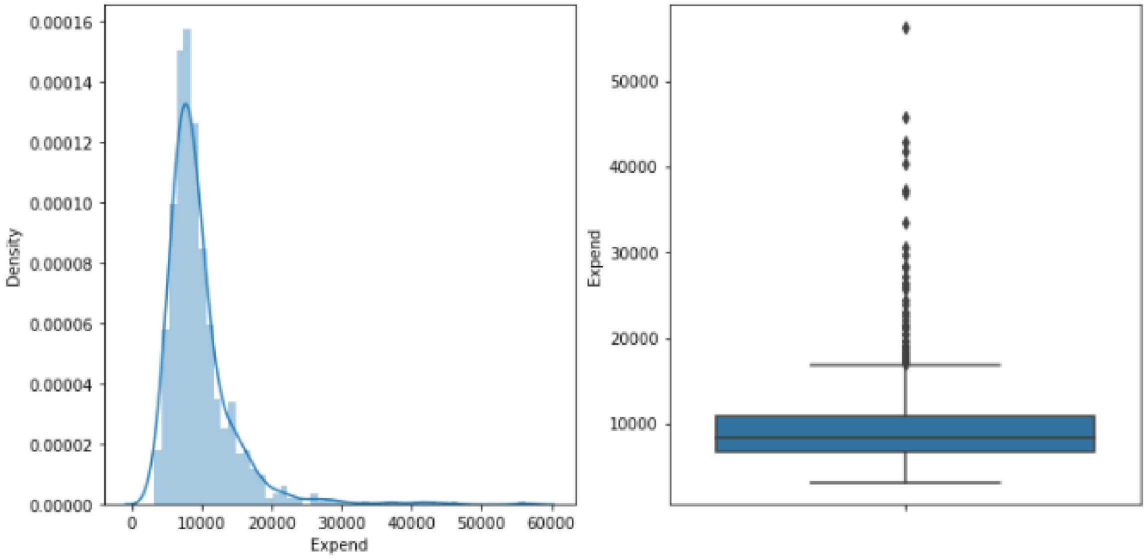


Figure XIX - Univariate Analysis (Instructional Expenditure)

The expenditure data seems positively skewed, and by box plot we can identify there are number of outlier present in the data.

Grad.Rate – Graduation rate

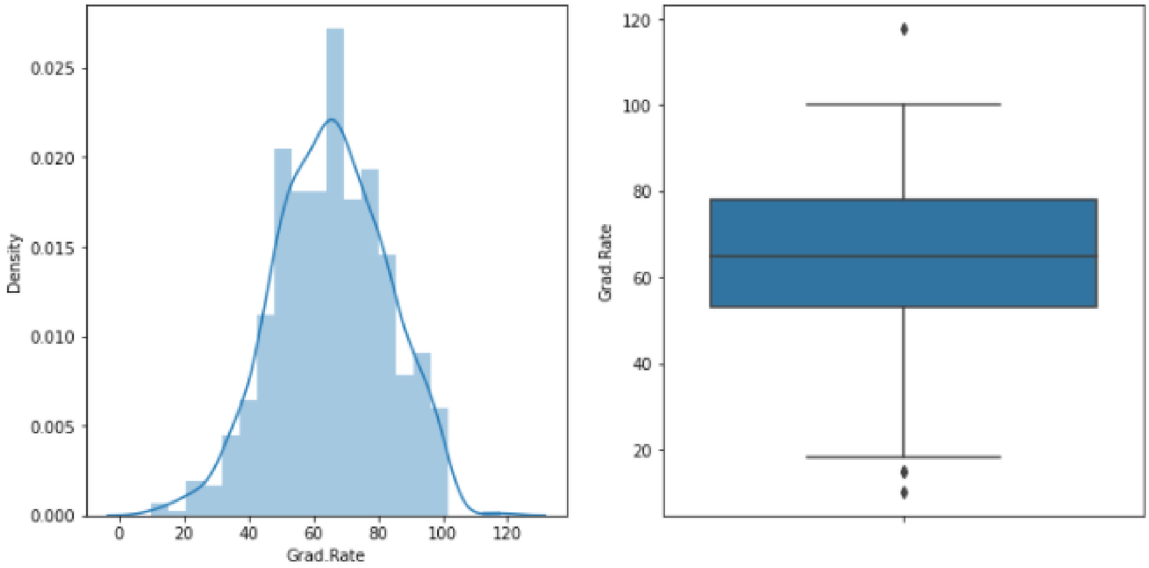


Figure XX - Univariate Analysis (Graduation Rate)

There are universities that have lower graduation rate than mean and high, the outlier on upper range seems incorrect as it is > 100%.

Bi-variate Analysis –

In Bi-variate analysis we'll perform analysis between two numeric variable, pair plot and correlation matrix would tell us how the variables are behaving and related against each other.

Pair Plot:

With pair plot we'll identify the pattern between 2 numerical variables.

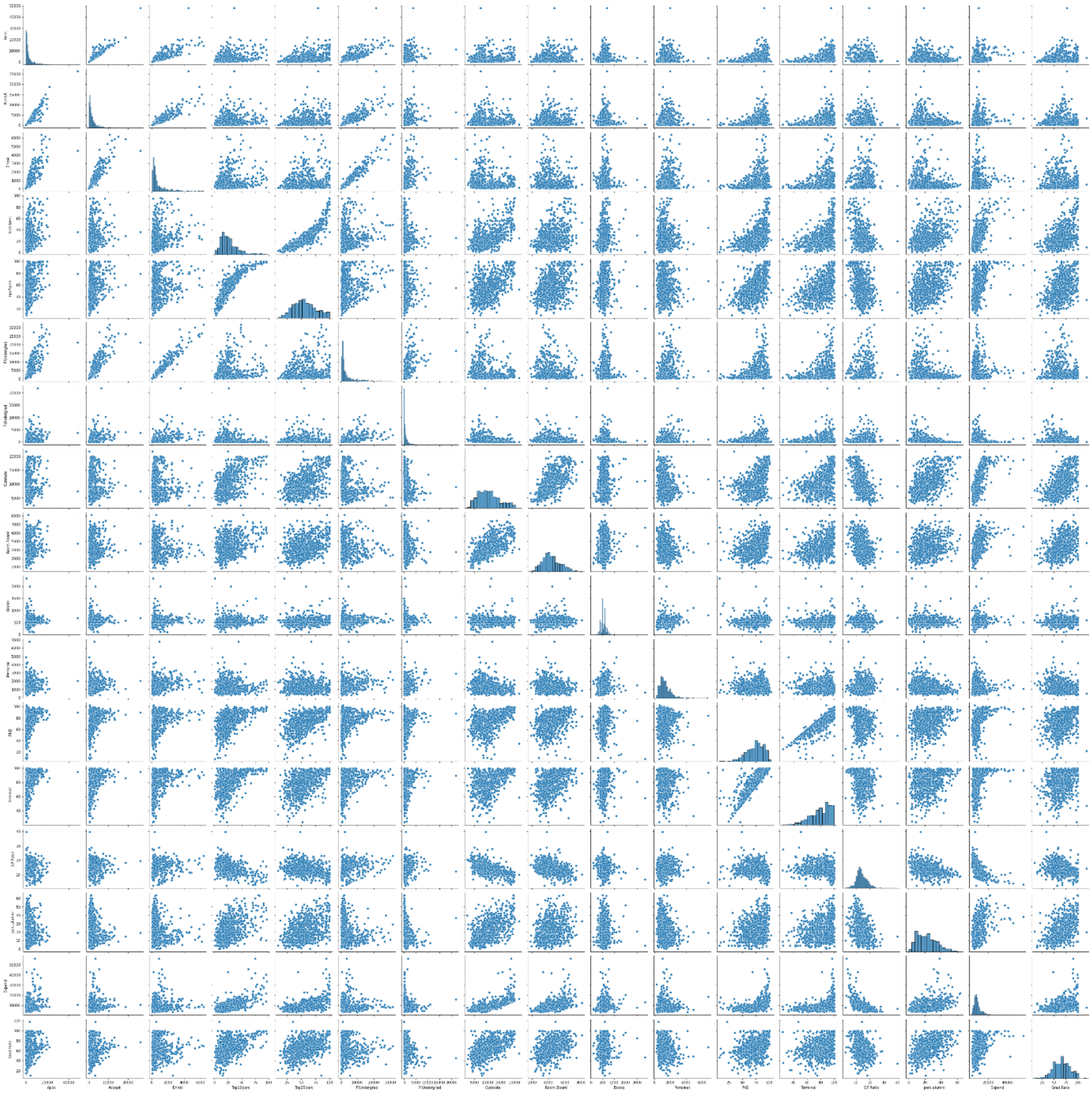


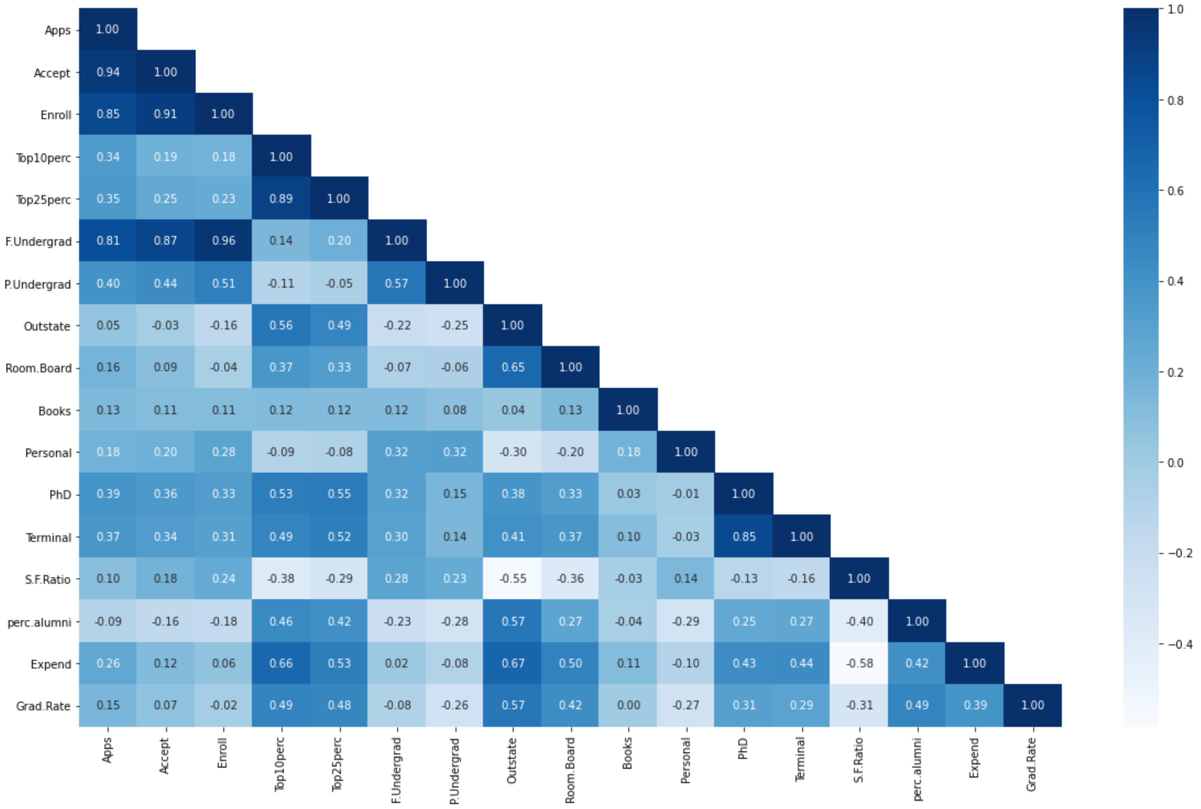
Figure XXI - Bi-Variate Analysis (Pair Plot)

Correlation Matrix (Heat Map):

Correlation would tell how significantly one variable follows another and type of their relationship.

Positive correlation – Both variables either increase together or decrease together

Negative correlation – If one variable increases other decreases and vice versa.



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

Figure XXII - Bi-Variate Analysis (Heat Map/Correlation)

By figure above, we can infer that

- Full time graduate relates with number of applications received, accepted and student enrolled in university.
- Faculties having Ph.D. and terminal degree tends to follow each other.
- Expenditure and student to faculty ratio seems inversely proportional to each other, can be assumed that as number of students per faculty increases, expenditure with respect to student drops.

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

While describing the data (here) we have seen the numeric fields have different range which cannot be fit into same scale. For example, top 10 or 25 percent students, student to faculty ratio, percentage of student graduated etc. would have range of 0 – 100, whereas expenditure, cost of room and board can vary more due to amount. Hence to bring them all in same scale we should have to scale these numeric fields' data.

Scaling Approach (Z-Score)

Z-Score tells us how many standard deviations the point is away from the mean, and as we know in a normal distribution, our 99.7% data lie between ± 3 standard deviation from the mean… so Z-score scaling would give us the result in same scale.

$$Z = \frac{x - \mu}{\sigma}$$

Columns	0	1	2	3	4
Apps	-0.346882	-0.210884	-0.406866	-0.668261	-0.726176
Accept	-0.321205	-0.038703	-0.376318	-0.681682	-0.764555
Enroll	-0.063509	-0.288584	-0.478121	-0.692427	-0.780735
Top10perc	-0.258583	-0.655656	-0.315307	1.840231	-0.655656
Top25perc	-0.191827	-1.353911	-0.292878	1.677612	-0.596031
F.Undergrad	-0.168116	-0.209788	-0.549565	-0.658079	-0.711924
P.Undergrad	-0.209207	0.244307	-0.49709	-0.520752	0.009005
Outstate	-0.746356	0.457496	0.201305	0.626633	-0.716508
Room.Board	-0.964905	1.909208	-0.554317	0.996791	-0.216723
Books	-0.602312	1.21588	-0.905344	-0.602312	1.518912
Personal	1.270045	0.235515	-0.259582	-0.688173	0.235515
PhD	-0.163028	-2.675646	-1.204845	1.185206	0.204672
Terminal	-0.115729	-3.378176	-0.931341	1.175657	-0.523535
S.F.Ratio	1.013776	-0.477704	-0.300749	-1.615274	-0.553542
perc.alumni	-0.867574	-0.544572	0.585935	1.151188	-1.675079
Expend	-0.50191	0.16611	-0.17729	1.792851	0.241803
Grad.Rate	-0.318252	-0.551262	-0.667767	-0.376504	-2.939613

Table 6 - Scaled Data Head (Transposed)

After scaling we would get data like above, it would be scaled and mostly vary between -3 to +3, anything beyond them are outliers.

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

Covariance Matrix (with heat map)

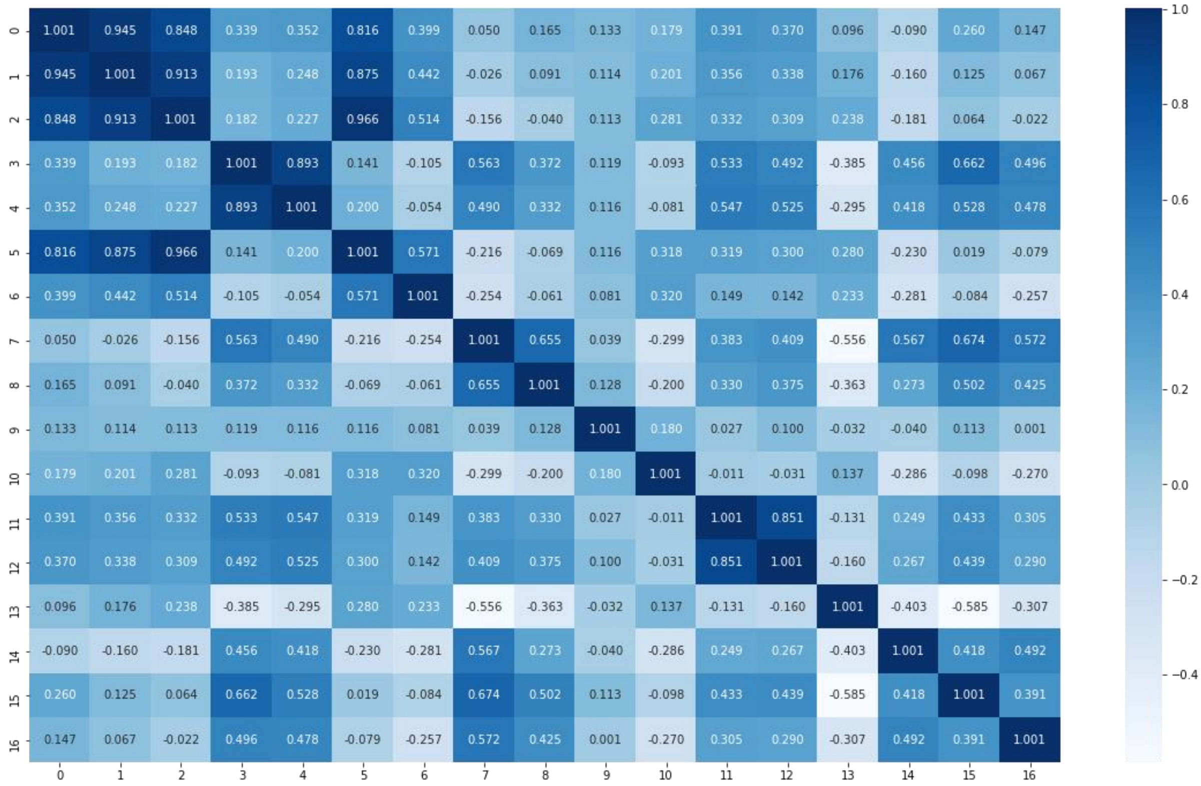
Covariance between two variables tells us the direction of linear relationship between them,

Positive means they increase or decrease together,

Negative means if one increases then other decreases (inversely proportional)

Covariance is identified as product of spread of each point (of 2 variables) from the mean.

$$\mathrm{Cov}\ (X, Y) = \frac{\sum(X_i - \overline{X})(Y_j - \overline{Y})}{n\ - 1}$$



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	1.001	0.945	0.848	0.339	0.352	0.816	0.399	0.050	0.165	0.133	0.179	0.391	0.370	0.096	-0.090	0.260	0.147
1	0.945	1.001	0.913	0.193	0.248	0.875	0.442	-0.026	0.091	0.114	0.201	0.356	0.338	0.176	-0.160	0.125	0.067
2	0.848	0.913	1.001	0.182	0.227	0.966	0.514	-0.156	-0.040	0.113	0.281	0.332	0.309	0.238	-0.181	0.064	-0.022
3	0.339	0.193	0.182	1.001	0.893	0.141	-0.105	0.563	0.372	0.119	-0.093	0.533	0.492	-0.385	0.456	0.662	0.496
4	0.352	0.248	0.227	0.893	1.001	0.200	-0.054	0.490	0.332	0.116	-0.081	0.547	0.525	-0.295	0.418	0.528	0.478
5	0.816	0.875	0.966	0.141	0.200	1.001	0.571	-0.216	-0.069	0.116	0.318	0.319	0.300	0.280	-0.230	0.019	-0.079
6	0.399	0.442	0.514	-0.105	-0.054	0.571	1.001	-0.254	-0.061	0.081	0.320	0.149	0.142	0.233	-0.281	-0.084	-0.257
7	0.050	-0.026	-0.156	0.563	0.490	-0.216	-0.254	1.001	0.655	0.039	-0.299	0.383	0.409	-0.556	0.567	0.674	0.572
8	0.165	0.091	-0.040	0.372	0.332	-0.069	-0.061	0.655	1.001	0.128	-0.200	0.330	0.375	-0.363	0.273	0.502	0.425
9	0.133	0.114	0.113	0.119	0.116	0.116	0.081	0.039	0.128	1.001	0.180	0.027	0.100	-0.032	-0.040	0.113	0.001
10	0.179	0.201	0.281	-0.093	-0.081	0.318	0.320	-0.299	-0.200	0.180	1.001	-0.011	-0.031	0.137	-0.286	-0.098	-0.270
11	0.391	0.356	0.332	0.533	0.547	0.319	0.149	0.383	0.330	0.027	-0.011	1.001	0.851	-0.131	0.249	0.433	0.305
12	0.370	0.338	0.309	0.492	0.525	0.300	0.142	0.409	0.375	0.100	-0.031	0.851	1.001	-0.160	0.267	0.439	0.290
13	0.096	0.176	0.238	-0.385	-0.295	0.280	0.233	-0.556	-0.363	-0.032	0.137	-0.131	-0.160	1.001	-0.403	-0.585	-0.307
14	-0.090	-0.160	-0.181	0.456	0.418	-0.230	-0.281	0.567	0.273	-0.040	-0.286	0.249	0.267	-0.403	1.001	0.418	0.492
15	0.260	0.125	0.064	0.662	0.528	0.019	-0.084	0.674	0.502	0.113	-0.098	0.433	0.439	-0.585	0.418	1.001	0.391
16	0.147	0.067	-0.022	0.496	0.478	-0.079	-0.257	0.572	0.425	0.001	-0.270	0.305	0.290	-0.307	0.492	0.391	1.001

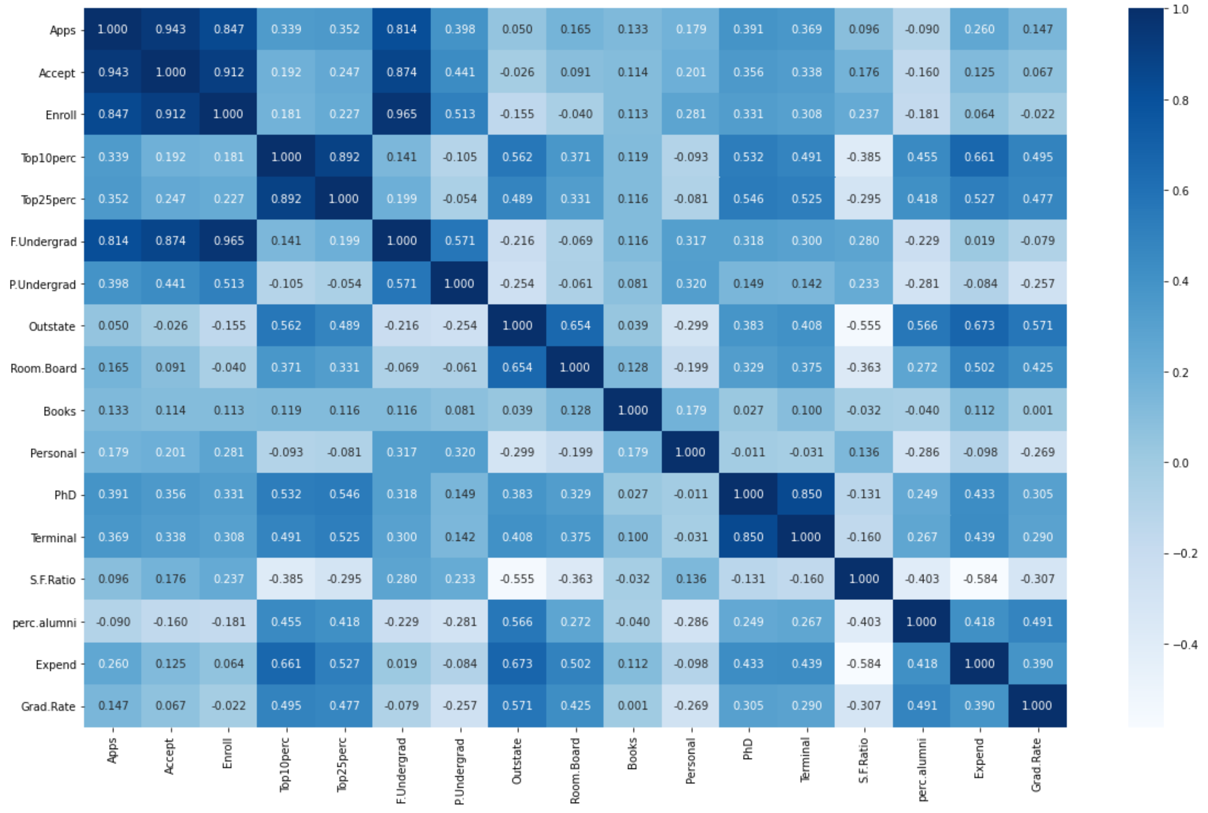
Figure XXIII - Covariance Matrix (After Scaling)

Correlation Matrix (with heat map)

As evident by correlation matrix before and after scaling, they both are same. Correlation gives the strength (how much significant the relationship is?) between two numeric variables.

Correlation (Pearson coefficient) is identified as, ratio of covariance against standard deviations of both variables.

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.000	0.943	0.847	0.339	0.352	0.814	0.398	0.050	0.165	0.133	0.179	0.391	0.369	0.096	-0.090	0.260	0.147
Accept	0.943	1.000	0.912	0.192	0.247	0.874	0.441	-0.026	0.091	0.114	0.201	0.356	0.338	0.176	-0.160	0.125	0.067
Enroll	0.847	0.912	1.000	0.181	0.227	0.965	0.513	-0.155	-0.040	0.113	0.281	0.331	0.308	0.237	-0.181	0.064	-0.022
Top10perc	0.339	0.192	0.181	1.000	0.892	0.141	-0.105	0.562	0.371	0.119	-0.093	0.532	0.491	-0.385	0.455	0.661	0.495
Top25perc	0.352	0.247	0.227	0.892	1.000	0.199	-0.054	0.489	0.331	0.116	-0.081	0.546	0.525	-0.295	0.418	0.527	0.477
F.Undergrad	0.814	0.874	0.965	0.141	0.199	1.000	0.571	-0.216	-0.069	0.116	0.317	0.318	0.300	0.280	-0.229	0.019	-0.079
P.Undergrad	0.398	0.441	0.513	-0.105	-0.054	0.571	1.000	-0.254	-0.061	0.081	0.320	0.149	0.142	0.233	-0.281	-0.084	-0.257
Outstate	0.050	-0.026	-0.155	0.562	0.489	-0.216	-0.254	1.000	0.654	0.039	-0.299	0.383	0.408	-0.555	0.566	0.673	0.571
Room.Board	0.165	0.091	-0.040	0.371	0.331	-0.069	-0.061	0.654	1.000	0.128	-0.199	0.329	0.375	-0.363	0.272	0.502	0.425
Books	0.133	0.114	0.113	0.119	0.116	0.116	0.081	0.039	0.128	1.000	0.179	0.027	0.100	-0.032	-0.040	0.112	0.001
Personal	0.179	0.201	0.281	-0.093	-0.081	0.317	0.320	-0.299	-0.199	0.179	1.000	-0.011	-0.031	0.136	-0.286	-0.098	-0.269
PhD	0.391	0.356	0.331	0.532	0.546	0.318	0.149	0.383	0.329	0.027	-0.011	1.000	0.850	-0.131	0.249	0.433	0.305
Terminal	0.369	0.338	0.308	0.491	0.525	0.300	0.142	0.408	0.375	0.100	-0.031	0.850	1.000	-0.160	0.267	0.439	0.290
S.F.Ratio	0.096	0.176	0.237	-0.385	-0.295	0.280	0.233	-0.555	-0.363	-0.032	0.136	-0.131	-0.160	1.000	-0.403	-0.584	-0.307
perc.alumni	-0.090	-0.160	-0.181	0.455	0.418	-0.229	-0.281	0.566	0.272	-0.040	-0.286	0.249	0.267	-0.403	1.000	0.418	0.491
Expend	0.260	0.125	0.064	0.661	0.527	0.019	-0.084	0.673	0.502	0.112	-0.098	0.433	0.439	-0.584	0.418	1.000	0.390
Grad.Rate	0.147	0.067	-0.022	0.495	0.477	-0.079	-0.257	0.571	0.425	0.001	-0.269	0.305	0.290	-0.307	0.491	0.390	1.000

Figure XXIV - Correlation Matrix (After Scaling)

Comparison of Covariance vs Correlation

As we know, Covariance gives direction of relation whereas Correlation gives strength of relation.

If we look at the matrices after scaling, both can be seen with same numbers. Since, we have applied z-score scaling the population rescaled with mean ~0 and standard deviation ~1.

Since correlation coefficient is ratio of covariance against standard deviation, there will not be significant difference between both matrices after scaling.

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

Before Scaling

We can see that each feature has a different scaling in plot.

The outliers before scaling can be identified as –

Values lesser than Q1 – 1.5*IQR or values more than Q3 + 1.5*IQR, are outliers.

IQR = Q3 – Q1

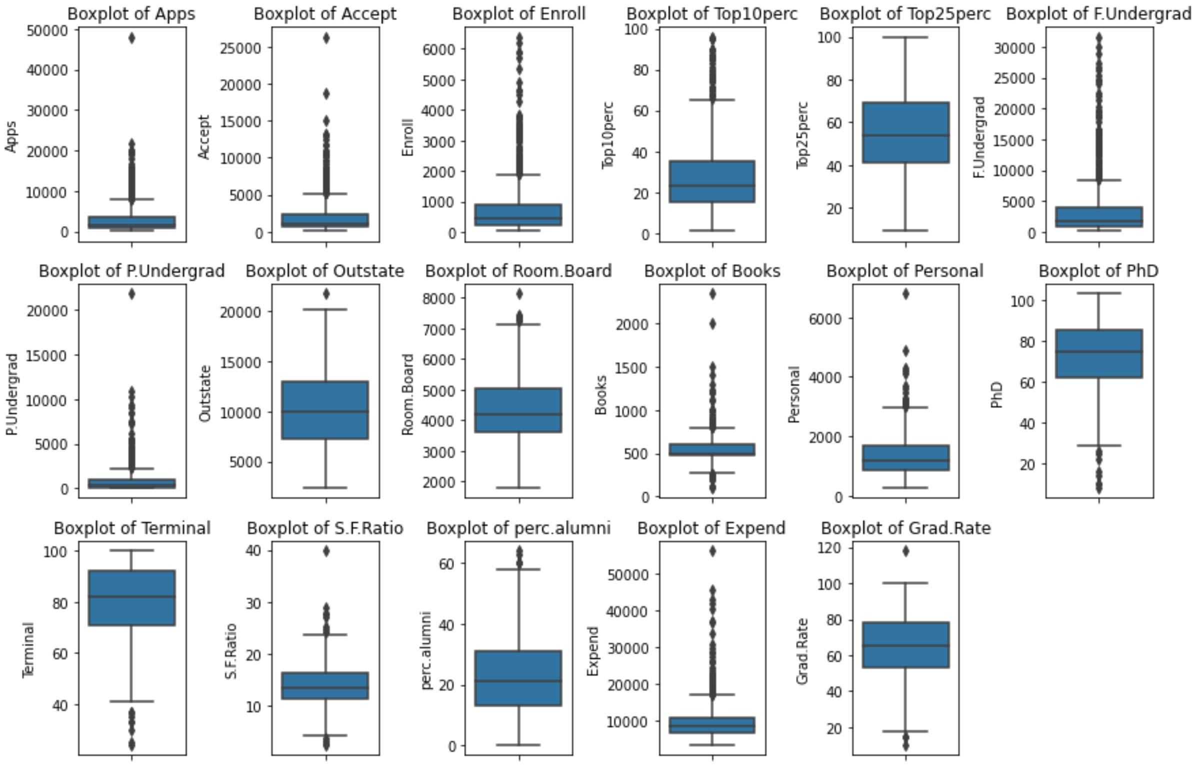


Figure XXV - Boxplot (Before Scaling)

After Scaling

After scaling (Z-score scaling), outlier would still be there, but the box plots would be more visible since they can fit in same axis (the plots are having same scale -3 to +3)

Outliers after scaling, can be identified as –

Values lesser than -3 or values more than +3

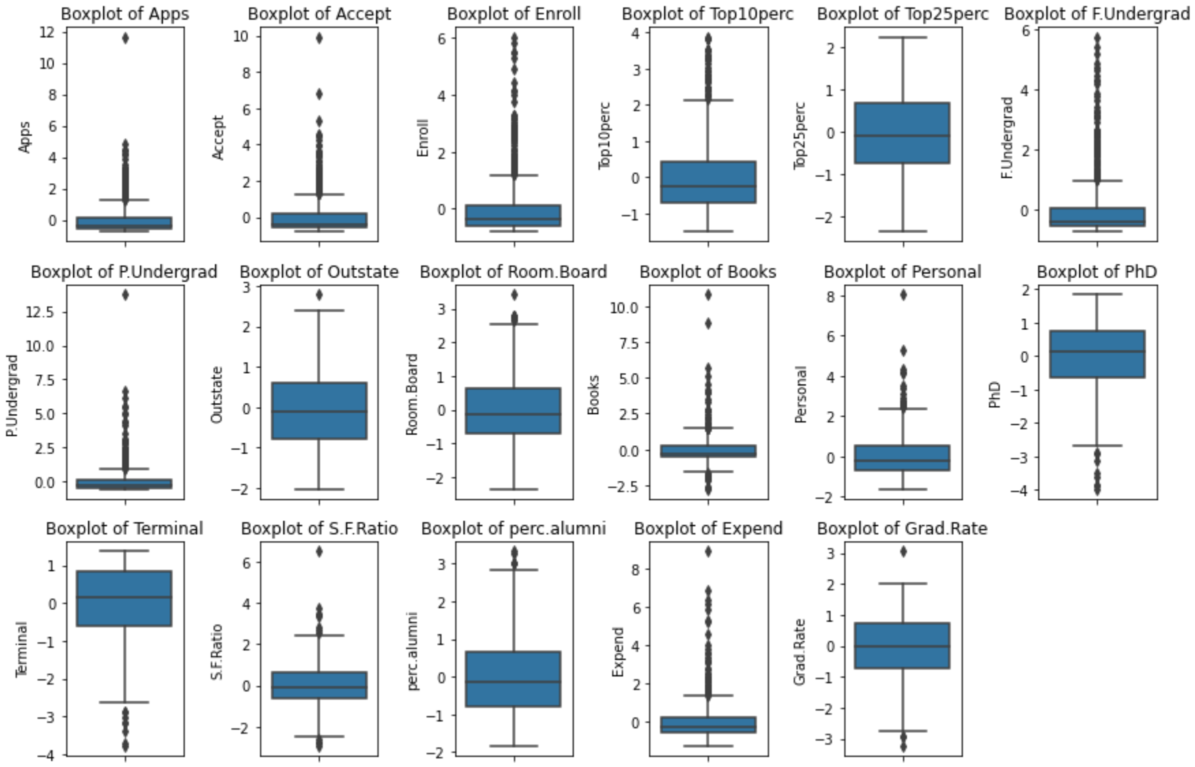

Boxplot of Apps
Boxplot of Accept
Boxplot of Enroll
Boxplot of Top10perc
Boxplot of Top25perc
Boxplot of F.Undergrad
Boxplot of P.Undergrad
Boxplot of Outstate
Boxplot of Room.Board
Boxplot of Books
Boxplot of Personal
Boxplot of PhD
Boxplot of Terminal
Boxplot of S.F.Ratio
Boxplot of perc.alumni
Boxplot of Expend
Boxplot of Grad.Rate

Figure XXVI - Boxplot (After Scaling)

Cleaning-up Incorrect Values

In EDA, we have seen for % of PhD faculties and Student's graduation rate, there are some incorrect values (>100 %)

We have removed them before treating the outliers so that they do not have adverse effect on our analysis.

Sno.	Names	PhD	Grad.Rate
95	Cazenovia College	22	118
582	Texam A&M University at Galveston	103	43

Table 7 - Incorrect Data Values

Treatment of Outliers

We can treat outliers by setting extreme values to their subsequent limit,

If value > Q3 + 1.5*IQR, we will set it as Q3 + 1.5*IQR

And if value < Q1 - 1.5*IQR, we will set it as Q1 - 1.5*IQR

Post treatment, our data would look like –

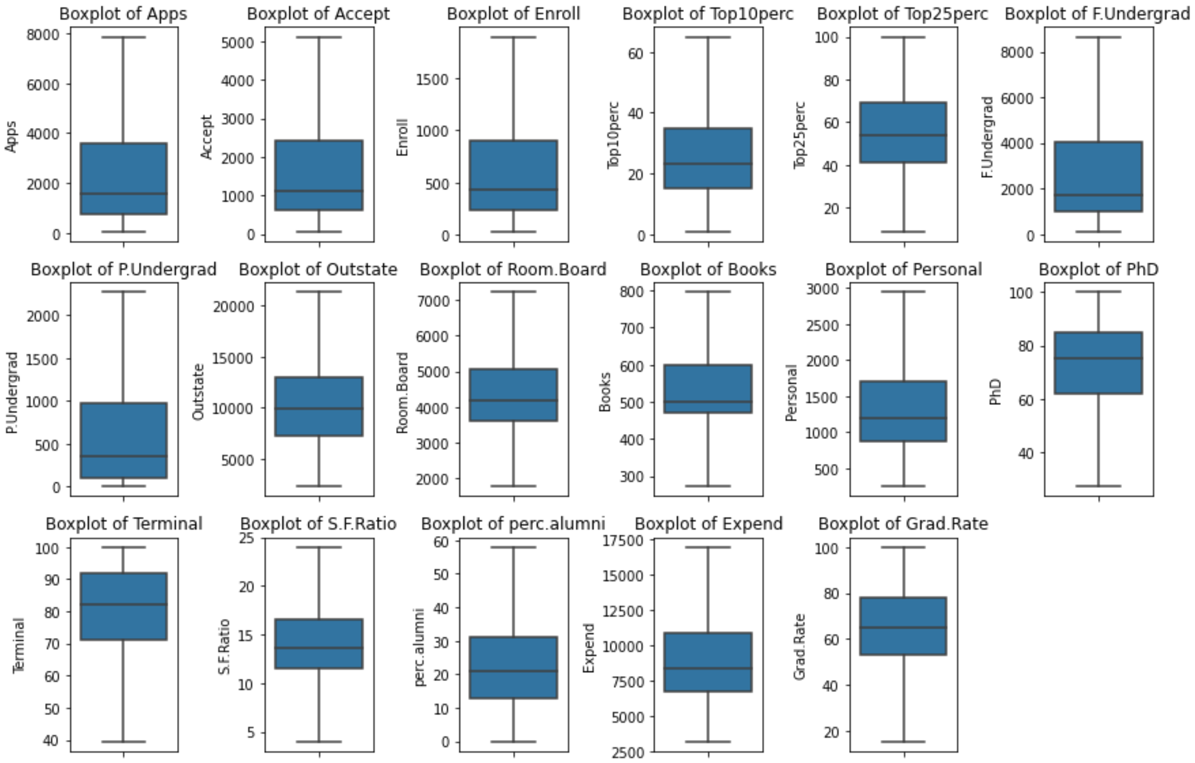


Figure XXVII - Boxplot (After Outlier treatment)

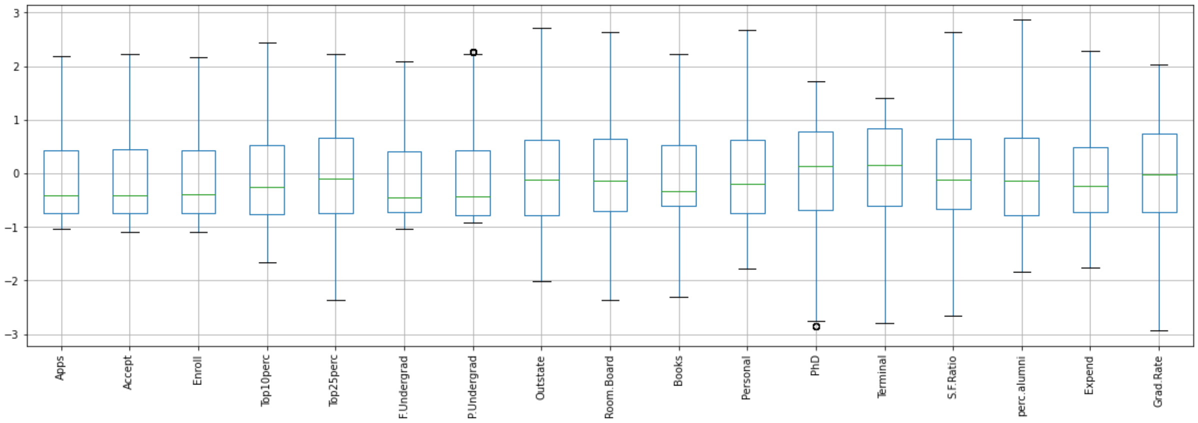


Figure XXVIII - Boxplot (After Outlier treatment & Scaling)

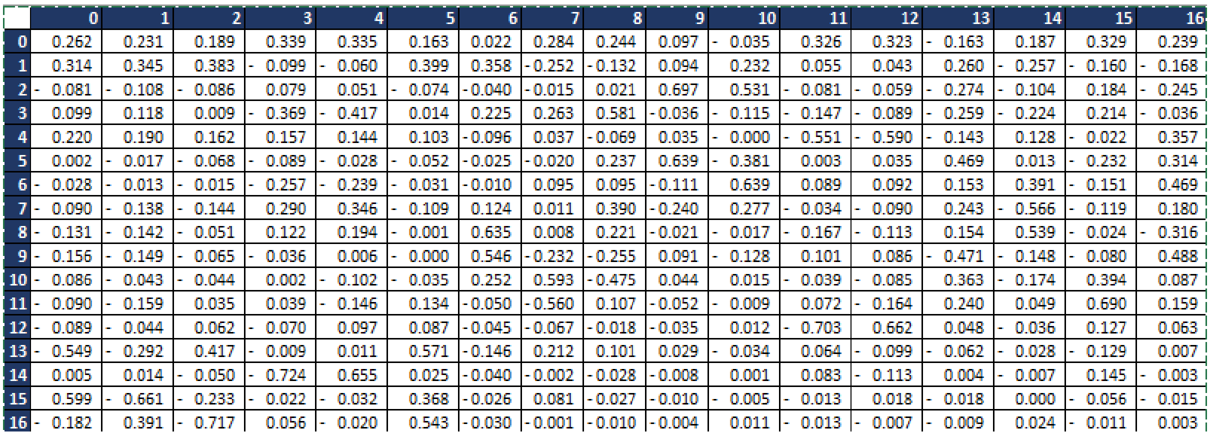
There were outliers in all numerical data points (except Top25perc), we have treated them to analysis further.

2.5 Extract the eigenvalues and eigenvectors

Eigen vectors

Eigen vectors represent principal components and how much weightage assigned to each feature.

```
[[ 0.262  0.231  0.189  0.339  0.335  0.163  0.022  0.284  0.244 0.097  -0.035  0.326  0.323  -0.163  0.187  0.329  0.239 ]
 [ 0.314  0.345  0.383  -0.099  -0.060  0.399  0.358  -0.252  -0.132  0.094  0.232  0.055  0.043  0.260  -0.257  -0.160  -0.168 ]
 [-0.081  -0.108  -0.086  0.079  0.051  -0.074  -0.040  -0.015  0.021  0.697  0.531  -0.081  -0.059  -0.274  -0.104  0.184  -0.245 ]
 [ 0.099  0.118  0.009  -0.369  -0.417  0.014  0.225  0.263  0.581  -0.036  -0.115  -0.147  -0.089  -0.259  -0.224  0.214  -0.036 ]
 [ 0.220  0.190  0.162  0.157  0.144  0.103  -0.096  0.037  -0.069  0.035  -0.000  -0.551  -0.590  -0.143  0.128  -0.022  0.357 ]
 [ 0.002  -0.017  -0.068  -0.089  -0.028  -0.052  -0.025  -0.020  0.237  0.639  -0.381  0.003  0.035  0.469  0.013  -0.232  0.314 ]
 [-0.028  -0.013  -0.015  -0.257  -0.239  -0.031  -0.010  0.095  0.095  -0.111  0.639  0.089  0.092  0.153  0.391  -0.151  0.469 ]
 [-0.090  -0.138  -0.144  0.290  0.346  -0.109  0.124  0.011  0.390  -0.240  0.277  -0.034  -0.090  0.243  -0.566  -0.119  0.180 ]
 [-0.131  -0.142  -0.051  0.122  0.194  -0.001  0.635  0.008  0.221  -0.021  -0.017  -0.167  -0.113  0.154  0.539  -0.024  -0.316 ]
 [-0.156  -0.149  -0.065  -0.036  0.006  -0.000  0.546  -0.232  -0.255  0.091  -0.128  0.101  0.086  -0.471  -0.148  -0.080  0.488 ]
 [-0.086  -0.043  -0.044  0.002  -0.102  -0.035  0.252  0.593  -0.475  0.044  0.015  -0.039  -0.085  0.363  -0.174  0.394  0.087 ]
 [-0.090  -0.159  0.035  0.039  -0.146  0.134  -0.050  -0.560  0.107  -0.052  -0.009  0.072  -0.164  0.240  0.049  0.690  0.159 ]
 [-0.089  -0.044  0.062  -0.070  0.097  0.087  -0.045  -0.067  -0.018  -0.035  0.012  -0.703  0.662  0.048  -0.036  0.127  0.063 ]
 [-0.549  -0.292  0.417  -0.009  0.011  0.571  -0.146  0.212  0.101  0.029  -0.034  0.064  -0.099  -0.062  -0.028  -0.129  0.007 ]
 [ 0.005  0.014  -0.050  -0.724  0.655  0.025  -0.040  -0.002  -0.028  -0.008  0.001  0.083  -0.113  0.004  -0.007  0.145  -0.003 ]
 [ 0.599  -0.661  -0.233  -0.022  -0.032  0.368  -0.026  0.081  -0.027  -0.010  -0.005  -0.013  0.018  -0.018  0.000  -0.056  -0.015 ]
 [-0.182  0.391  -0.717  0.056  -0.020  0.543  -0.030  -0.001  -0.010  -0.004  0.011  -0.013  -0.007  -0.009  0.024  -0.011  0.003 ]]
```



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	0.262	0.231	0.189	0.339	0.335	0.163	0.022	0.284	0.244	0.097	-0.035	0.326	0.323	-0.163	0.187	0.329	0.239
1	0.314	0.345	0.383	-0.099	-0.060	0.399	0.358	-0.252	-0.132	0.094	0.232	0.055	0.043	0.260	-0.257	-0.160	-0.168
2	-0.081	-0.108	-0.086	0.079	0.051	-0.074	-0.040	-0.015	0.021	0.697	0.531	-0.081	-0.059	-0.274	-0.104	0.184	-0.245
3	0.099	0.118	0.009	-0.369	-0.417	0.014	0.225	0.263	0.581	-0.036	-0.115	-0.147	-0.089	-0.259	-0.224	0.214	-0.036
4	0.220	0.190	0.162	0.157	0.144	0.103	-0.096	0.037	-0.069	0.035	-0.000	-0.551	-0.590	-0.143	0.128	-0.022	0.357
5	0.002	-0.017	-0.068	-0.089	-0.028	-0.052	-0.025	-0.020	0.237	0.639	-0.381	0.003	0.035	0.469	0.013	-0.232	0.314
6	-0.028	-0.013	-0.015	-0.257	-0.239	-0.031	-0.010	0.095	0.095	-0.111	0.639	0.089	0.092	0.153	0.391	-0.151	0.469
7	-0.090	-0.138	-0.144	0.290	0.346	-0.109	0.124	0.011	0.390	-0.240	0.277	-0.034	-0.090	0.243	-0.566	-0.119	0.180
8	-0.131	-0.142	-0.051	0.122	0.194	-0.001	0.635	0.008	0.221	-0.021	-0.017	-0.167	-0.113	0.154	0.539	-0.024	-0.316
9	-0.156	-0.149	-0.065	-0.036	0.006	-0.000	0.546	-0.232	-0.255	0.091	-0.128	0.101	0.086	-0.471	-0.148	-0.080	0.488
10	-0.086	-0.043	-0.044	0.002	-0.102	-0.035	0.252	0.593	-0.475	0.044	0.015	-0.039	-0.085	0.363	-0.174	0.394	0.087
11	-0.090	-0.159	0.035	0.039	-0.146	0.134	-0.050	-0.560	0.107	-0.052	-0.009	0.072	-0.164	0.240	0.049	0.690	0.159
12	-0.089	-0.044	0.062	-0.070	0.097	0.087	-0.045	-0.067	-0.018	-0.035	0.012	-0.703	0.662	0.048	-0.036	0.127	0.063
13	-0.549	-0.292	0.417	-0.009	0.011	0.571	-0.146	0.212	0.101	0.029	-0.034	0.064	-0.099	-0.062	-0.028	-0.129	0.007
14	0.005	0.014	-0.050	-0.724	0.655	0.025	-0.040	-0.002	-0.028	-0.008	0.001	0.083	-0.113	0.004	-0.007	0.145	-0.003
15	0.599	-0.661	-0.233	-0.022	-0.032	0.368	-0.026	0.081	-0.027	-0.010	-0.005	-0.013	0.018	-0.018	0.000	-0.056	-0.015
16	-0.182	0.391	-0.717	0.056	-0.020	0.543	-0.030	-0.001	-0.010	-0.004	0.011	-0.013	-0.007	-0.009	0.024	-0.011	0.003

Table 8 - Eigen Vector

Eigen Values

Eigen value also know as explained variance in principal component analysis, and it would be in decreasing order always since first eigen value represent variance of first principal component.

[5.663 , 4.895 , 1.126 , 1.004 , 0.872 , 0.766 , 0.585 , 0.545 , 0.424 , 0.381 , 0.247 , 0.147 , 0.134 , 0.099 , 0.075 , 0.038 , 0.022]

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

Before PCA, we will check if assumptions are holding up correctly or not, and if data is adequate to perform PCA.

- There is significant correlation between variables

With bartlett sphericity calculation, we will get a P-value 0.00 and we can reject the NULL hypothesis… Hence, it is evident that correlations are significant.

- Data adequacy

With KMO calculation, we can confirm the value of KMO test come up as 0.84, which is good to perform PCA.

Principal Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	0.262	0.314	-0.081	0.099	0.220	0.002	-0.028	-0.090	-0.131	-0.156	-0.086	-0.090	-0.089	-0.549	0.005	0.599	-0.182
Accept	0.231	0.345	-0.108	0.118	0.190	-0.017	-0.013	-0.138	-0.142	-0.149	-0.043	-0.159	-0.044	-0.292	0.014	-0.661	0.391
Enroll	0.189	0.383	-0.086	0.009	0.162	-0.068	-0.015	-0.144	-0.051	-0.065	-0.044	0.035	0.062	0.417	-0.050	-0.233	-0.717
Top10perc	0.339	-0.099	0.079	-0.369	0.157	-0.089	-0.257	0.290	0.122	-0.036	0.002	0.039	-0.070	-0.009	-0.724	-0.022	0.056
Top25perc	0.335	-0.060	0.051	-0.417	0.144	-0.028	-0.239	0.346	0.194	0.006	-0.102	-0.146	0.097	0.011	0.655	-0.032	-0.020
F.Undergrad	0.163	0.399	-0.074	0.014	0.103	-0.052	-0.031	-0.109	-0.001	0.000	-0.035	0.134	0.087	0.571	0.025	0.368	0.543
P.Undergrad	0.022	0.358	-0.040	0.225	-0.096	-0.025	-0.010	0.124	0.635	0.546	0.252	-0.050	-0.045	-0.146	-0.040	-0.026	-0.030
Outstate	0.284	-0.252	-0.015	0.263	0.037	-0.020	0.095	0.011	0.008	-0.232	0.593	-0.560	-0.067	0.212	-0.002	0.081	-0.001
Room.Board	0.244	-0.132	0.021	0.581	-0.069	0.237	0.095	0.390	0.221	-0.255	-0.475	0.107	-0.018	0.101	-0.028	-0.027	-0.010
Books	0.097	0.094	0.697	-0.036	0.035	0.639	-0.111	-0.240	-0.021	0.091	0.044	-0.052	-0.035	0.029	-0.008	-0.010	-0.004
Personal	-0.035	0.232	0.531	-0.115	0.000	-0.381	0.639	0.277	-0.017	-0.128	0.015	-0.009	0.012	-0.034	0.001	-0.005	0.011
PhD	0.326	0.055	-0.081	-0.147	-0.551	0.003	0.089	-0.034	-0.167	0.101	-0.039	0.072	-0.703	0.064	0.083	-0.013	-0.013
Terminal	0.323	0.043	-0.059	-0.089	-0.590	0.035	0.092	-0.090	-0.113	0.086	-0.085	-0.164	0.662	-0.099	-0.113	0.018	-0.007
S.F.Ratio	-0.163	0.260	-0.274	-0.259	-0.143	0.469	0.153	0.243	0.154	-0.471	0.363	0.240	0.048	-0.062	0.004	-0.018	-0.009
perc.alumni	0.187	-0.257	-0.104	-0.224	0.128	0.013	0.391	-0.566	0.539	-0.148	-0.174	0.049	-0.036	-0.028	-0.007	0.000	0.024
Expend	0.329	-0.160	0.184	0.214	-0.022	-0.232	-0.151	-0.119	-0.024	-0.080	0.394	0.690	0.127	-0.129	0.145	-0.056	-0.011
Grad.Rate	0.239	-0.168	-0.245	-0.036	0.357	0.314	0.469	0.180	-0.316	0.488	0.087	0.159	0.063	0.007	-0.003	-0.015	0.003

Table 9 - PCA Components

Explained Ratio by each component

Explained ratio tell us how much information each PC would explain, and it is represented in decreasing form (as first PC would always show more information compared to others)

Explained ratio = eigen value of component/sum of eigen values of all PCs

[0.333,0.288, 0.066, 0.059, 0.051, 0.045, 0.034, 0.032, 0.025, 0.022, 0.015, 0.009, 0.008, 0.006, 0.004, 0.002, 0.001]

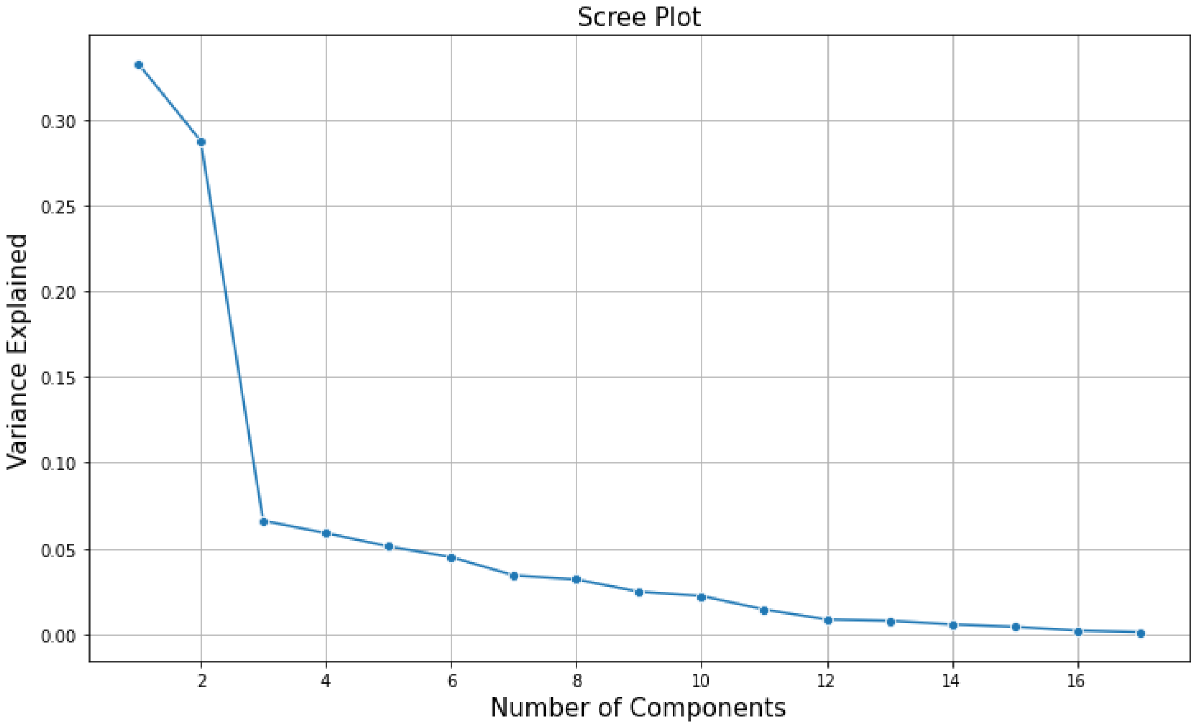


Figure XXIX - Scree Plot of all PCs

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

Each principal component assigns weightage to features and extract most of the information, the first PC can be written as –

PC1 ➔

(0.26*Apps) + (0.23*Accept) + (0.19*Enroll) + (0.34*Top10perc) + (0.33*Top25perc) + (0.16*F.Undergrad) + (0.02*P.Undergrad) + (0.28*Outstate) + (0.24*Room.Board) + (0.1*Books) + (-0.04*Personal) + (0.33*PhD) + (0.32*Terminal) + (-0.16*S.F.Ratio) + (0.19*perc.alumni) + (0.33*Expend) + (0.24*Grad.Rate)

2.8 Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?

Cumulative values of eigen values -

[0.333, 0.620, 0.686, 0.745, 0.797, 0.842, 0.876, 0.908, 0.933, 0.955, 0.970, 0.978, 0.986, 0.992, 0.996, 0.999, 1.000]

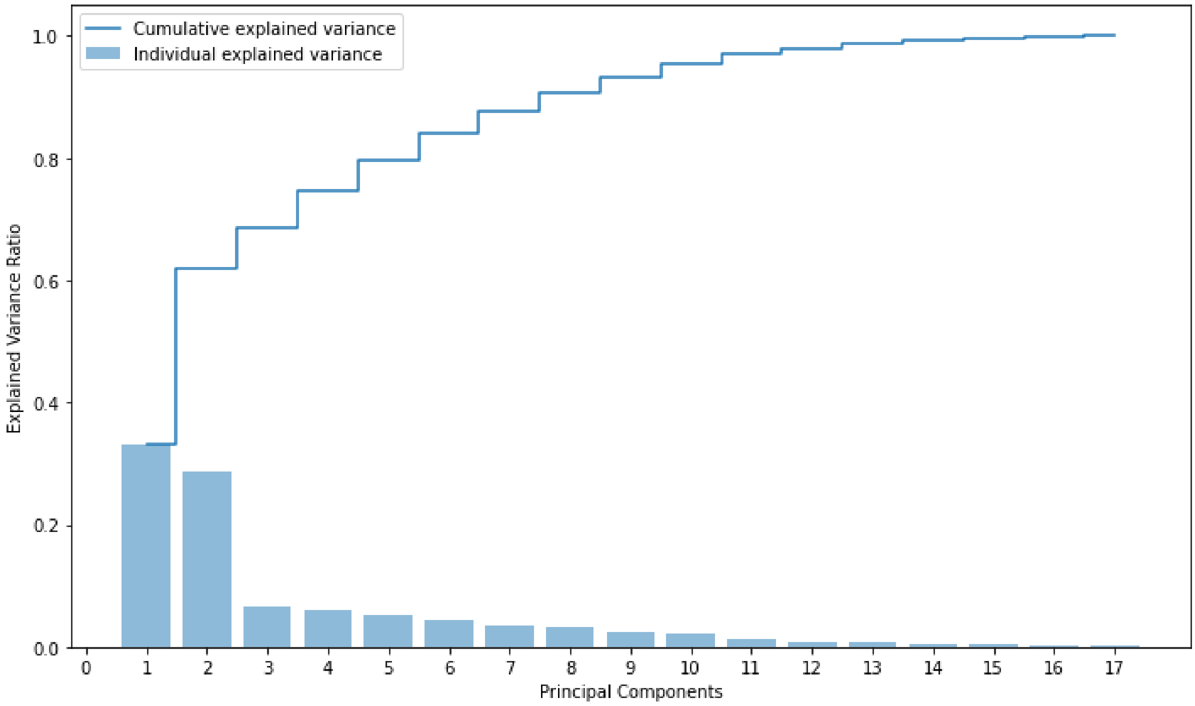


Figure XXX - Cumulative Explained Variance

There are multiple approach how optimum number of PCs can be chosen –

- We can pull number of components which will explain 95-99% of data.
- Or we can pull components up to where increment between 2 consecutive PC drops significantly.

1) We can choose first 8 PCs, by which 90% data can be explained, and on 9th PC the variance ratio not increases by 3%.
2) Or if we choose 6 PCs, 80% data can be explained by them.

And each eigen vector represents direction in principal component and weightage to each feature to explain most data in principal component.

2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis?

In this case study, we have derived following understanding from univariate analysis –

There are several outliers present in the features (except for Top25perc), a university has more part-time graduates.

From Bi-variate Analysis – we can see there is significant correlation between number of full-time graduates and application received, accepted or number of enrolled students.

By Principal component analysis, we reduced the dimensions and choose limited number of component where data is mostly spread (which holds most of the data variance).

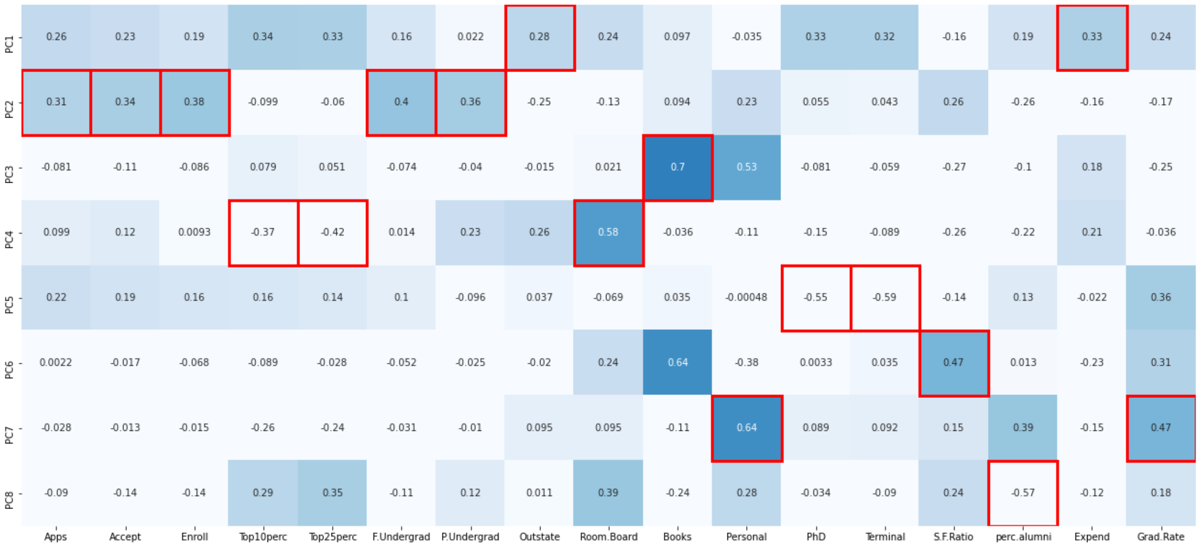
How may PCs help in the further analysis?

The number of PCs chosen depends on how much data we want to be explained by selected ones, with cumulative eigen values, we can see the variance those PCs are holding together.

- First PC can explain 33% of data
- First 4 PCs can explain 74.5% of data
- First 6 PCs can explain 84.2% of data
- First 8 PCs can explain 90.8% of data

We can choose either first 6 or 8 PCs based on amount of variance we are trying to put in our ML model.

We can also see which features have maximum loading across component(s) –



	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	0.26	0.23	0.19	0.34	0.33	0.16	0.022	0.28	0.24	0.097	-0.035	0.33	0.32	-0.16	0.19	0.33	0.24
PC2	0.31	0.34	0.38	-0.099	-0.06	0.4	0.36	-0.25	-0.13	0.094	0.23	0.055	0.043	0.26	-0.26	-0.16	-0.17
PC3	-0.081	-0.11	-0.086	0.079	0.051	-0.074	-0.04	-0.015	0.021	0.7	0.53	-0.081	-0.059	-0.27	-0.1	0.18	-0.25
PC4	0.099	0.12	0.0093	-0.37	-0.42	0.014	0.23	0.26	0.58	-0.036	-0.11	-0.15	-0.089	-0.26	-0.22	0.21	-0.036
PC5	0.22	0.19	0.16	0.16	0.14	0.1	-0.096	0.037	-0.069	0.035	-0.00048	-0.55	-0.59	-0.14	0.13	-0.022	0.36
PC6	0.0022	-0.017	-0.068	-0.089	-0.028	-0.052	-0.025	-0.02	0.24	0.64	-0.38	0.0033	0.035	0.47	0.013	-0.23	0.31
PC7	-0.028	-0.013	-0.015	-0.26	-0.24	-0.031	-0.01	0.095	0.095	-0.11	0.64	0.089	0.092	0.15	0.39	-0.15	0.47
PC8	-0.09	-0.14	-0.14	0.29	0.35	-0.11	0.12	0.011	0.39	-0.24	0.28	-0.034	-0.09	0.24	-0.57	-0.12	0.18

Figure XXXI - Feature weightage into Components

From above map, it can be derived that **PC1** has **Outstate** and **Expenditure** with maximum weightage, **PC2** has **Apps, Accept, Enroll, Part-time and full-time graduate feature** with maximum weight and so on…

END