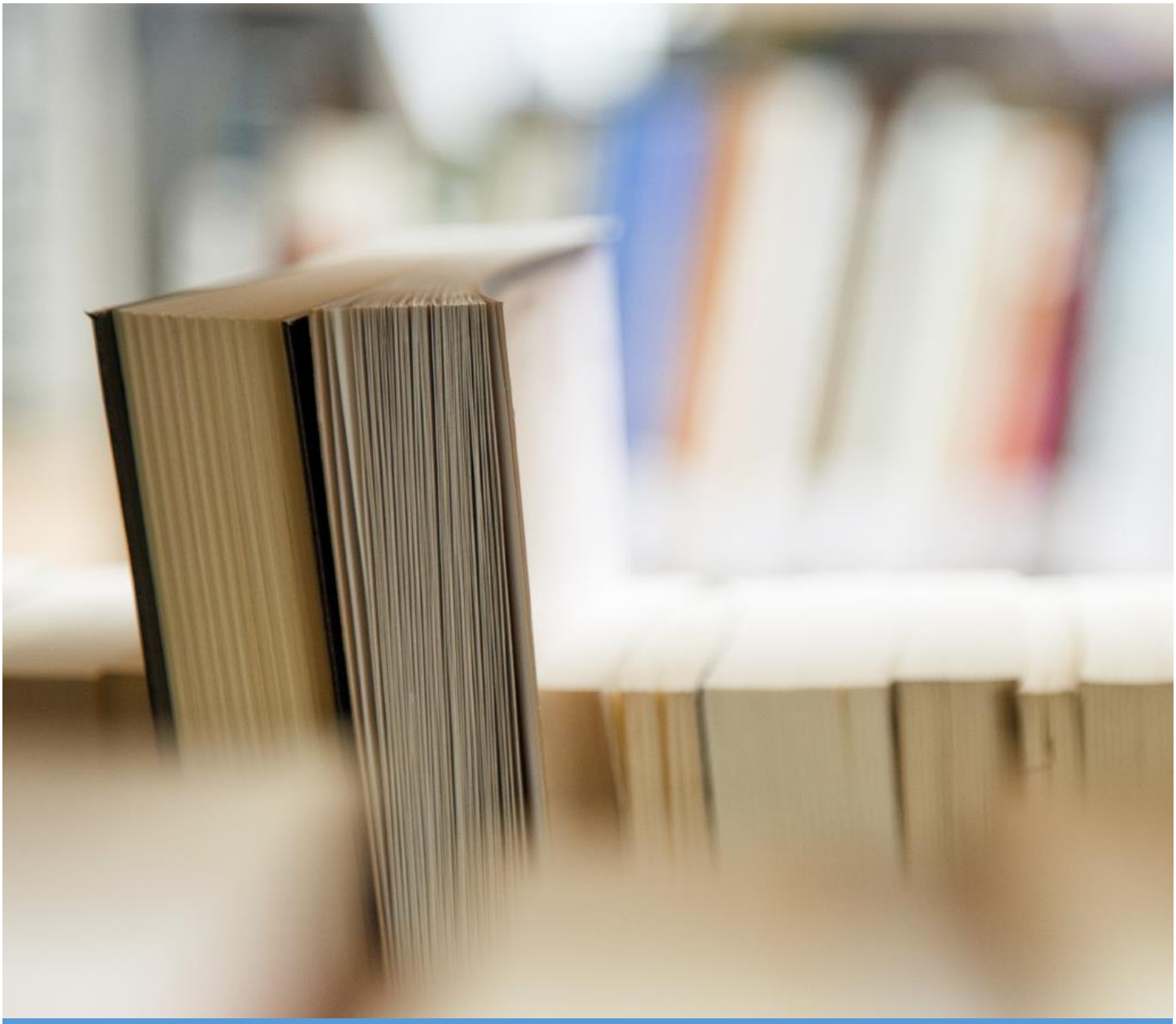


ADVANCED STATISTICS PROJECT REPORT

PGP-DSBA

srikanthpr.27@gmail.com

Table of Contents

1 Problem 1:

A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments. The data on hours of relief can be found in the following .csv file: Fever.csv

[Assume all of the ANOVA assumptions are satisfied]

One-way ANOVA Assumptions:

- The time taken [hours] for relief of hay fever samples drawn from individual ingredient 'A' and 'B' varied at 3 different levels are independent and random.
- The time taken [hours] for relief of hay fever samples drawn from individual ingredient 'A' and 'B' varied at 3 different levels are continuous and normally distributed. [Shapiro-test].
- The Homogeneity of variances [Variances are equal between groups].

1.1 State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables 'A' and 'B' individually. [Both statement and statistical form like Ho=mu, Ha>mu]

Solution:

State Null and Alternate Hypothesis:

For Individual Variable A:

- Null hypothesis:

H_0: The mean time taken [hours] for relief of hay fever by giving 'Ingredient-A' varied at 3 different Levels are same.

$$H_0: \mu_1 = \mu_2 = \mu_3$$

Where:
μ_1: The mean time taken [hours] for relief of hay fever using 'Ingredient-A' at Level-1.
μ_2: The mean time taken [hours] for relief of hay fever using 'Ingredient-A' at Level-2.
μ_3: The mean time taken [hours] for relief of hay fever using 'Ingredient-A' at Level-3.

- Alternate Hypothesis:

H_1: At least one of the mean time taken [hours] for relief of hay fever by using 'Ingredient-A' varied at 3 different levels are not same.

$$H_1: \mu_1 \neq \mu_2 = \mu_3 \text{ or}$$
$$H_1: \mu_1 = \mu_2 \neq \mu_3 \text{ or}$$
$$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \text{ or}$$

For Individual Variable B:

- Null hypothesis:

H_0: The mean time taken [hours] for relief of hay fever by using 'Ingredient-B' varied at 3 different Levels are same.

$$H_0: \mu_1 = \mu_2 = \mu_3$$

Where:
μ_1: The mean time taken [hours] for relief of hay fever using 'Ingredient-B' at Level-1.
μ_2: The mean time taken [hours] for relief of hay fever using 'Ingredient-B' at Level-2.
μ_3: The mean time taken [hours] for relief of hay fever using 'Ingredient-B' at Level-3.

- Alternate Hypothesis:

H_1: At least one of the mean time taken [hours] for relief of hay fever by using 'Ingredient-B' varied at 3 different levels are not same.

$$H_1: \mu_1 \neq \mu_2 = \mu_3 \text{ or}$$
$$H_1: \mu_1 = \mu_2 \neq \mu_3 \text{ or}$$
$$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \text{ or}$$

1.2 Perform one-way ANOVA for variable 'A' with respect to the variable 'Relief'. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.

Solution:

One-way ANOVA test results for ingredient 'A' with respect to 'Relief':

Output:

ANOVA Table:

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2.0	220.02	110.010000	23.465387	4.578242e-07
Residual	33.0	154.71	4.688182	NaN	NaN

Calculations:

Step1: Calculate mean sum of squares between the group [MSSB]:

- ✓ Sum of Squares between the Group; **SSB = 220.02** and
- ✓ Degrees of freedom [SSB] = c-1 = 2
  Where; C: Number of groups = 3
- ✓ Mean Sum of Squares between the Group [MSSB] = SSB ÷ Degrees of freedom [SSB]

MSSB= 110.01 ..... (1)

Step2: Calculate mean sum of squares within the group [MSSW]:

- ✓ Sum of Squares within the Group; **SSW = 154.71** and
- ✓ Degrees of freedom [SSW] = n-c = 33
  Where; n: Number of Observations = 36
- ✓ Mean Sum of Squares within the Group [MSSB] = SSW ÷ Degrees of freedom [SSW]

MSSW= 4.688182 ..... (2)

Step3: Calculate F-Ratio or F-statistics:

- ✓ F = MSSB ÷ MSSW

$$\mathbf{F_{stat} = 23.465} \text{ ..... } \mathbf{(3)}$$

Step4: Calculate P-value and compare with Level of significance, Alpha at 0.05:

P-value = 4.578242e-07

Final Interpretation:

- P-value obtained from ANOVA table is less than the level of significance Alpha, α = 0.05.
- Since P-value is less than Alpha; reject the null hypothesis, and therefore, we can conclude that the mean time taken [hours] for relief of hay fever by using individual ingredient 'A' varied at 3 different levels are significantly different.

1.3 Perform one-way ANOVA for variable 'B' with respect to the variable 'Relief'. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.

Solution:

One-way ANOVA test results for ingredient 'B' with respect to 'Relief':

Output:

ANOVA Table:

	df	sum_sq	mean_sq	F	PR(>F)
C(B)	2.0	123.66	61.830000	8.126777	0.00135
Residual	33.0	251.07	7.608182	NaN	NaN

Calculations:

Step1: Calculate mean sum of squares between the group [MSSB]:

- ✓ Sum of Squares between the Group; **SSB = 123.66** and
- ✓ Degrees of freedom [SSB] = c-1 = 2
  Where; C: Number of groups = 3
- ✓ Mean Sum of Squares between the Group [MSSB] = SSB ÷ Degrees of freedom [SSB]

MSSB= 61.83 ..... (1)

Step2: Calculate mean sum of squares within the group [MSSW]:

- ✓ Sum of Squares within the Group; **SSW = 251.07** and
- ✓ Degrees of freedom [SSW] = n-c = 33
  Where; n: Number of Observations = 36
- ✓ Mean Sum of Squares within the Group [MSSB] = SSW ÷ Degrees of freedom [SSW]

MSSW= 7.608182 ..... (2)

Step3: Calculate F-Ratio or F-statistics:

- ✓ F = MSSB ÷ MSSW = 23.465

F_{stat} =8.126777..... (3)

Step4: Calculate P-value and compare with Level of significance, Alpha at 0.05:

P-value = 0.00135

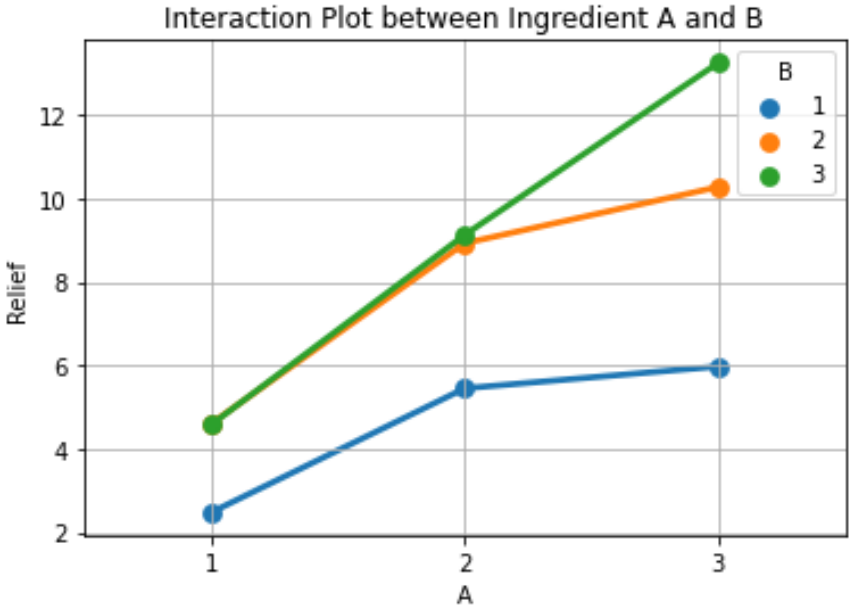
Final Interpretation:

- P-value obtained from ANOVA table is less than the level of significance Alpha, α = 0.05.
- Since P-value is less than Alpha; reject the null hypothesis, and therefore, we can conclude that the mean time taken [hours] for relief of hay fever by using individual ingredient 'B' varied at 3 different levels are significantly different.

1.4 Analyse the effects of one variable on another with the help of an interaction plot. What is the interaction between the two treatments? [Hint: use the 'point plot' function from the 'seaborn' function]

Solution:

Graph: Interaction Plot



Visual-Interpretation:

The graph shows that:

- The crossed lines between level-3 and level-2 of ingredient B shows that there is an interaction when the ingredient is A at Level-1 and Level-2.
- The time taken [hours] for relief of hay fever are almost same for Ingredient B at Level-3 and Level-2 when the Ingredient is 'A' at Level-1 and Level-2 respectively.
- There is no interaction for Ingredient B at Level-1 with ingredient 'A' at all 3 different levels.
- The time taken [hours] for relief of hay fever are higher for Ingredient B at Level-3 when the Ingredient is 'A' at Level-3.
- The time taken [hours] for relief of hay fever are lower for Ingredient B at Level-1 when the Ingredient is 'A' at Level-1.

1.5 Perform a two-way ANOVA based on the different ingredients (variable 'A' & 'B' along with their interaction 'A*B') with the variable 'Relief' and state your results.

Solution:

Two-way ANOVA test results for ingredient 'A' and 'B' with respect to 'Relief':

State the Null and Alternate Hypothesis:

- Null Hypothesis;
  H_0: There is no interaction between the levels of ingredient A and ingredient B.

- Alternate Hypothesis;
  H_1: There is an interaction between the levels of ingredient A and ingredient B.

Output:

ANOVA Table:

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2.0	220.020	110.010000	1827.858462	1.514043e-29
C(B)	2.0	123.660	61.830000	1027.329231	3.348751e-26
C(A):C(B)	4.0	29.425	7.356250	122.226923	6.972083e-17
Residual	27.0	1.625	0.060185	NaN	NaN

Final Interpretation:

- P-value obtained for interaction between the ingredients 'A' and 'B' from ANOVA table is less than the level of significance Alpha, α = 0.05.
- Since P-value is less than Alpha; reject the null hypothesis, and therefore, we can conclude that there is significant interaction between the levels of ingredient A and ingredient B.

1.6 Mention the business implications of performing ANOVA for this particular case study.

Solution:

- When one-way ANOVA applied separately to individual ingredient 'A' and 'B' with respect to variable 'Relief'. We can see that there is a significant different in time taken [hours] for relief of hay fever varied at all 3 different levels respectively.
- We can suggest to business to look into the data on hours of relief variables or increase the sample size, in order to get equal result for all the 3 different levels.
- ANOVA only suggest there is a difference, but will not say at which level there is a change in time taken for relief. In order to know the exact difference will have to perform Tukey's method which in turn will fetch more insights about performing ANOVA.
- When two-way ANOVA was performed, we could observe there is an interaction between the levels of ingredients 'A' and 'B'.
- From the interaction plot, we can also observe that the time taken [hours] for relief of hay fever are lower for Ingredient B at Level-1 when the Ingredient is 'A' at Level-1.By these we can suggest business to increase the production of Ingredient A and B at level-1 since the recovery time is very less.

2 Problem 2:

The dataset Education - Post 12th Standard.csv is a dataset that contains the names of various colleges. This particular case study is based on various parameters of various institutions. You are expected to do Principal Component Analysis for this case study according to the instructions given in the following rubric. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx.

Exploratory Data Analysis:

❖ View Head [Top 5 Observations] of the Data in Transpose View:

	0	1	2	3	4
Names	Abilene Christian University	Adelphi University	Adrian College	Agnes Scott College	Alaska Pacific University
Apps	1660	2186	1428	417	193
Accept	1232	1924	1097	349	146
Enroll	721	512	336	137	55
Top10perc	23	16	22	60	16
Top25perc	52	29	50	89	44
F.Undergrad	2885	2683	1036	510	249
P.Undergrad	537	1227	99	63	869
Outstate	7440	12280	11250	12960	7560
Room.Board	3300	6450	3750	5450	4120
Books	450	750	400	450	800
Personal	2200	1500	1165	875	1500
PhD	70	29	53	92	76
Terminal	78	30	66	97	72
S.F.Ratio	18.1	12.2	12.9	7.7	11.9
perc.alumni	12	16	30	37	2
Expend	7041	10527	8735	19016	10922
Grad.Rate	60	56	54	59	15

❖ Overview of the data:

Overview

Overview | Warnings 5 | Reproduction

Dataset statistics

Number of variables	18
Number of observations	777
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	163.4 KiB
Average record size in memory	215.3 B

Variable types

NUM	17
CAT	1

- There are total 777 Observations and 18 variables in the dataset.
- From the Dataset we have 17 Numerical Variables and one Categorical Variables.
- There are no Duplicate records for a given dataset.
- There are no missing values found for a given dataset.
- Summary of the Dataset before Outlier Treatment:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Names	777	777	LeTourneau University	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

2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented.

Solution:

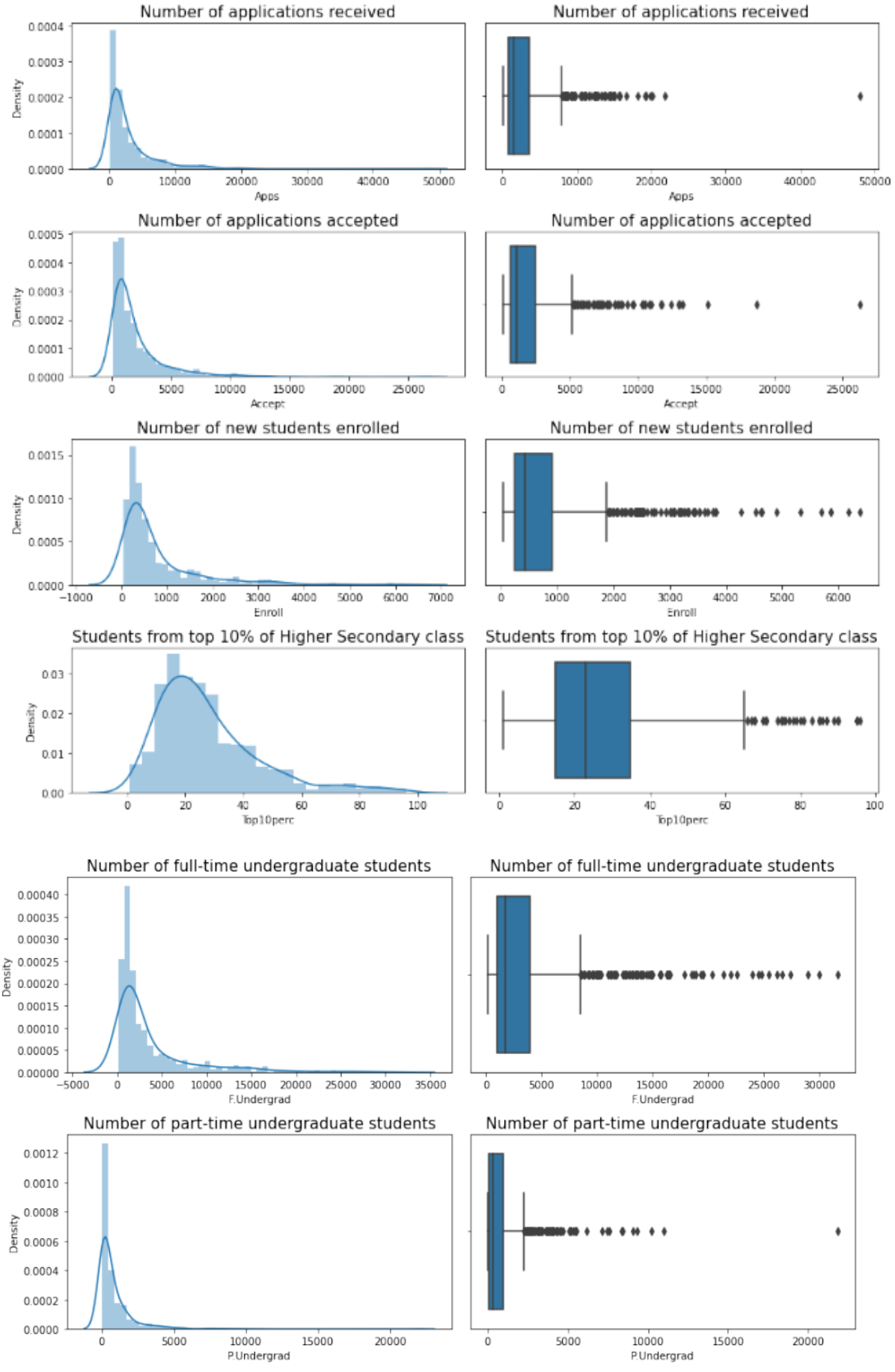
Univariate Analysis:

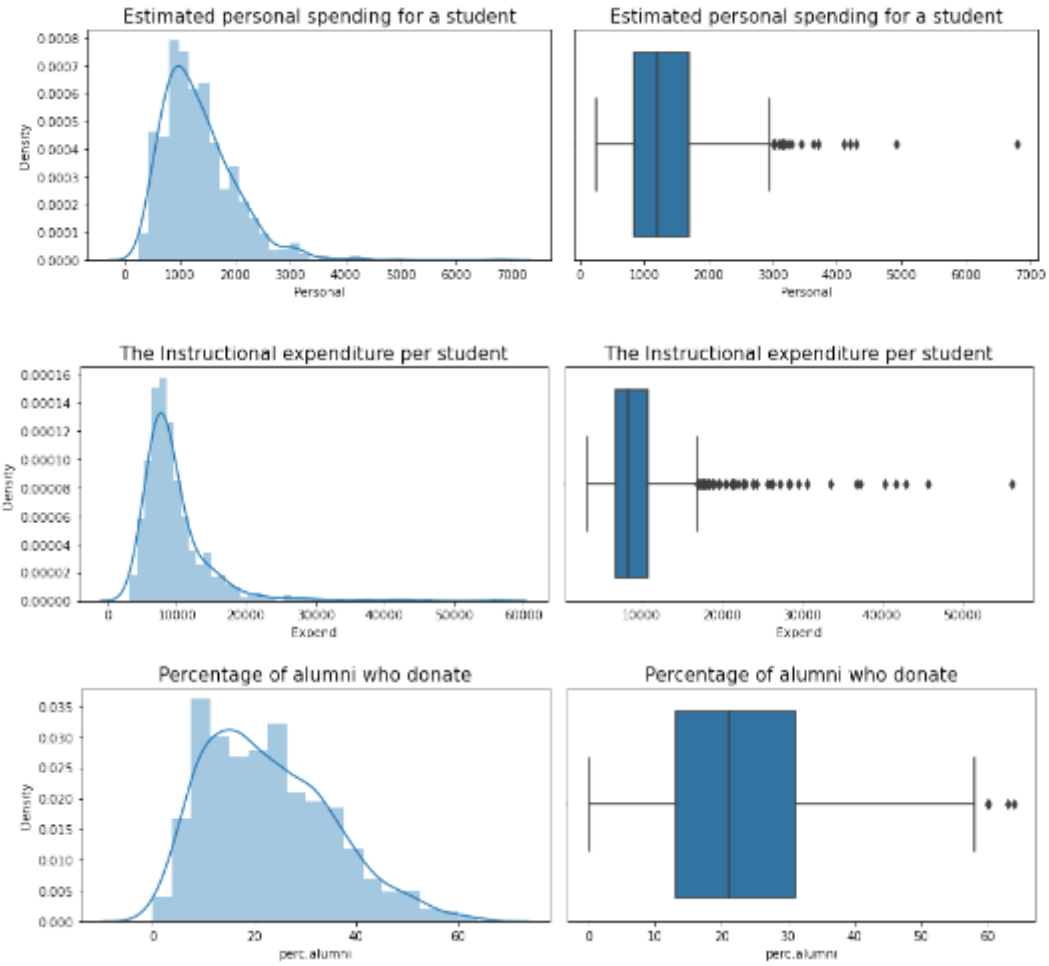
Positive or right skewed distribution: Mean > Median

By looking into the below distribution plot as well as box plot shows that the following variables are highly right skewed distribution and also there are outliers present towards high values for a given education dataset.

1. Apps: Number of applications received.
2. Accept: Number of applications accepted.
3. Enroll: Number of new students enrolled.
4. Top10perc: Percentage of new students from top 10% of Higher Secondary class.
5. F.Undergrad: Number of full-time undergraduate students.
6. P.Undergrad: Number of part-time undergraduate students.
7. Personal: Estimated personal spending for a student.

8. Expend: The Instructional expenditure per student.
9. perc.alumni: Percentage of alumni who donate.

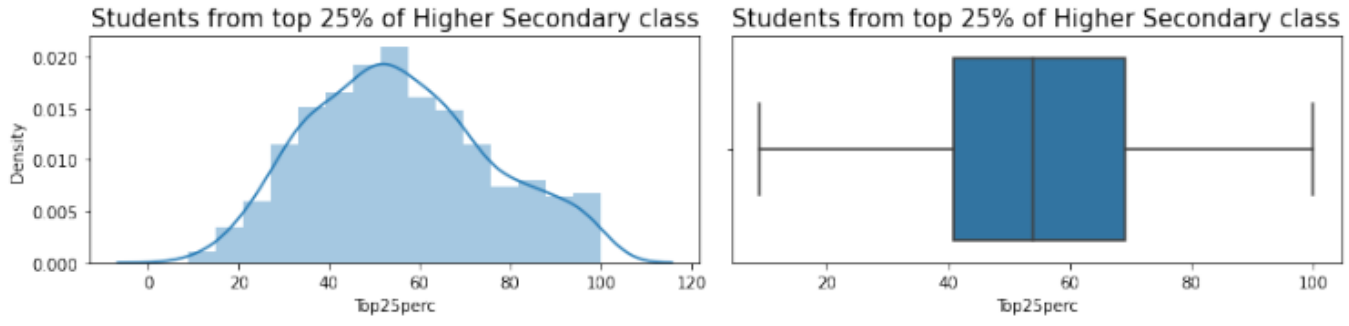


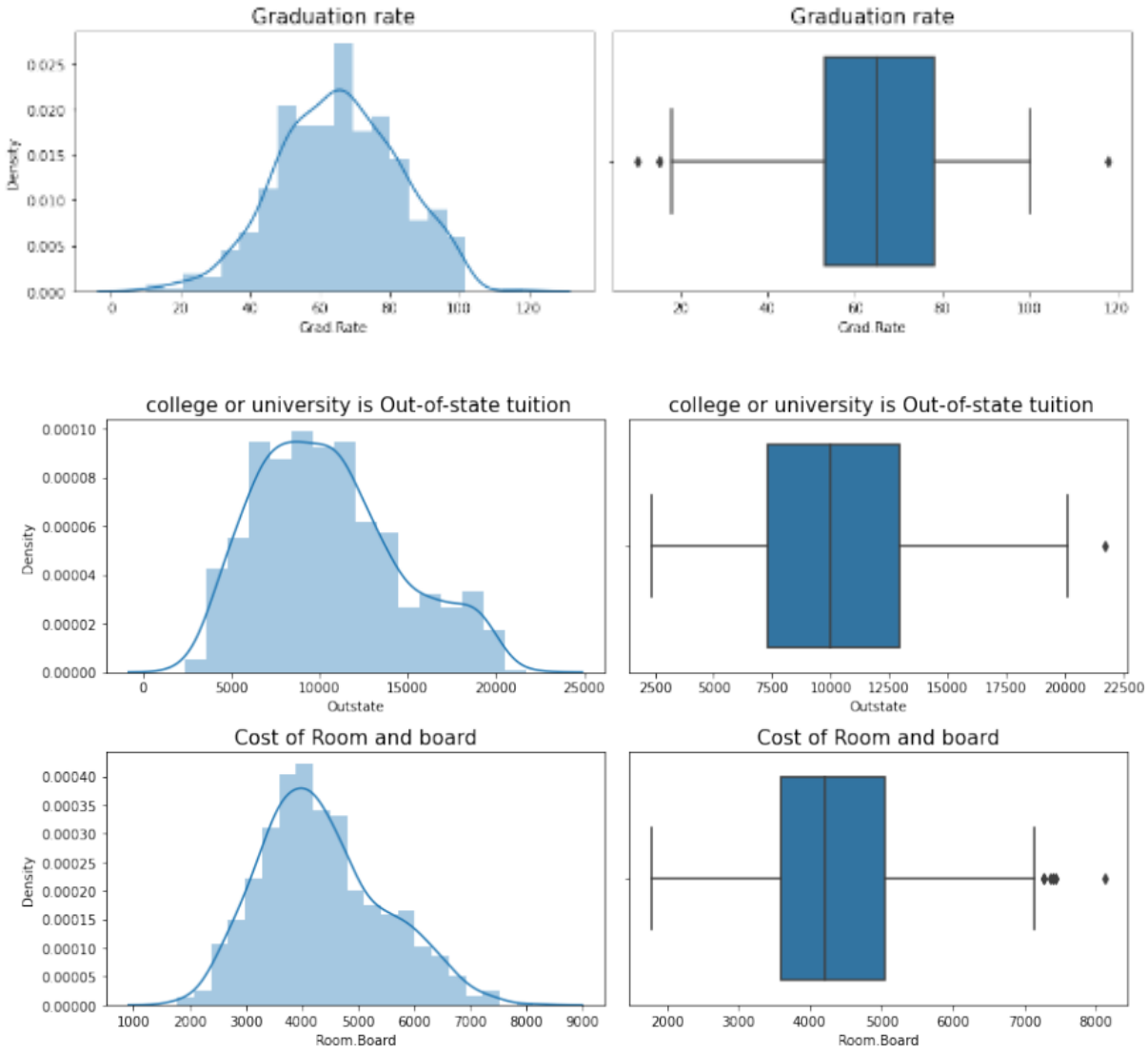


<u>Zero skewness or symmetrical distribution</u>: Mean = Median [Approx equal]

By looking into the below distribution plot as well as box plot shows that the following variables are approximately having zero skewness or symmetrical distribution and also there are very few outliers present for a given education dataset.

1. Top25perc: Percentage of new students from top 25% of Higher Secondary class.
2. Grad.Rate: Graduation rate
3. Outstate: Number of students for whom the particular college or university is Out-of-state tuition
4. Room.Board: Cost of Room and board

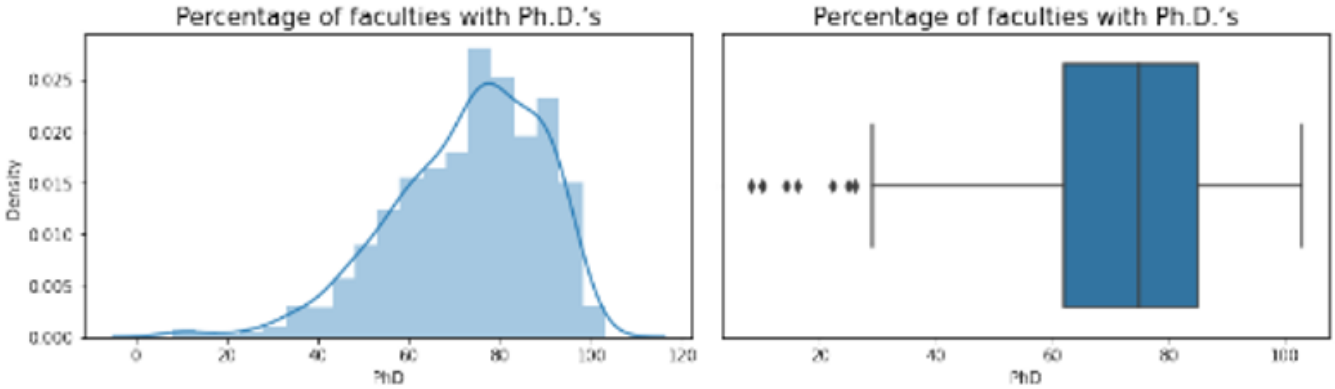


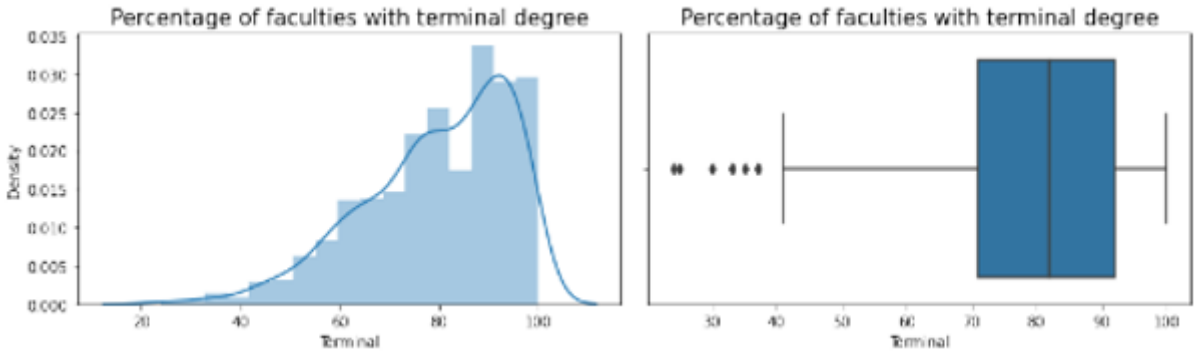


<u>Negative or left skewed distribution</u>: Mean < Median

By looking into the below distribution plot as well as box plot shows that the following variables are highly left skewed distribution and also there are outliers present towards lower values of a given education dataset.

1. PhD: Percentage of faculties with Ph.D.'s
2. Terminal: Percentage of faculties with terminal degree.

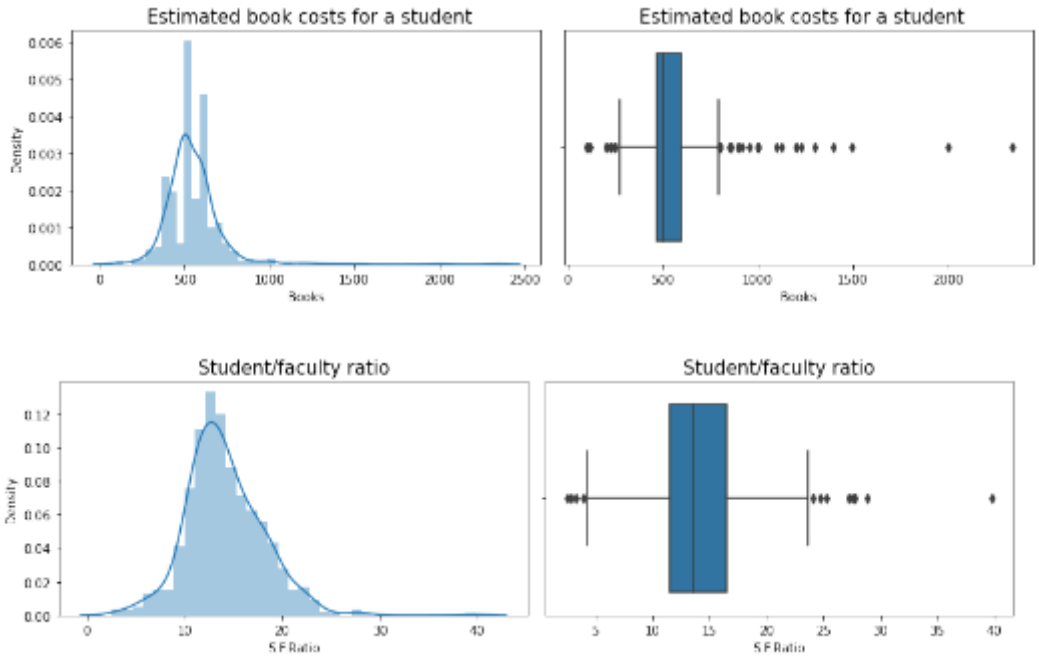




Outliers present on both the ends:

From the below graph, we can identify outliers are present on both the ends, lower tail as well higher tail for a given education dataset.

1. Books: Estimated book costs for a student.
2. S.F.Ratio: Student/faculty ratio



Skewness:

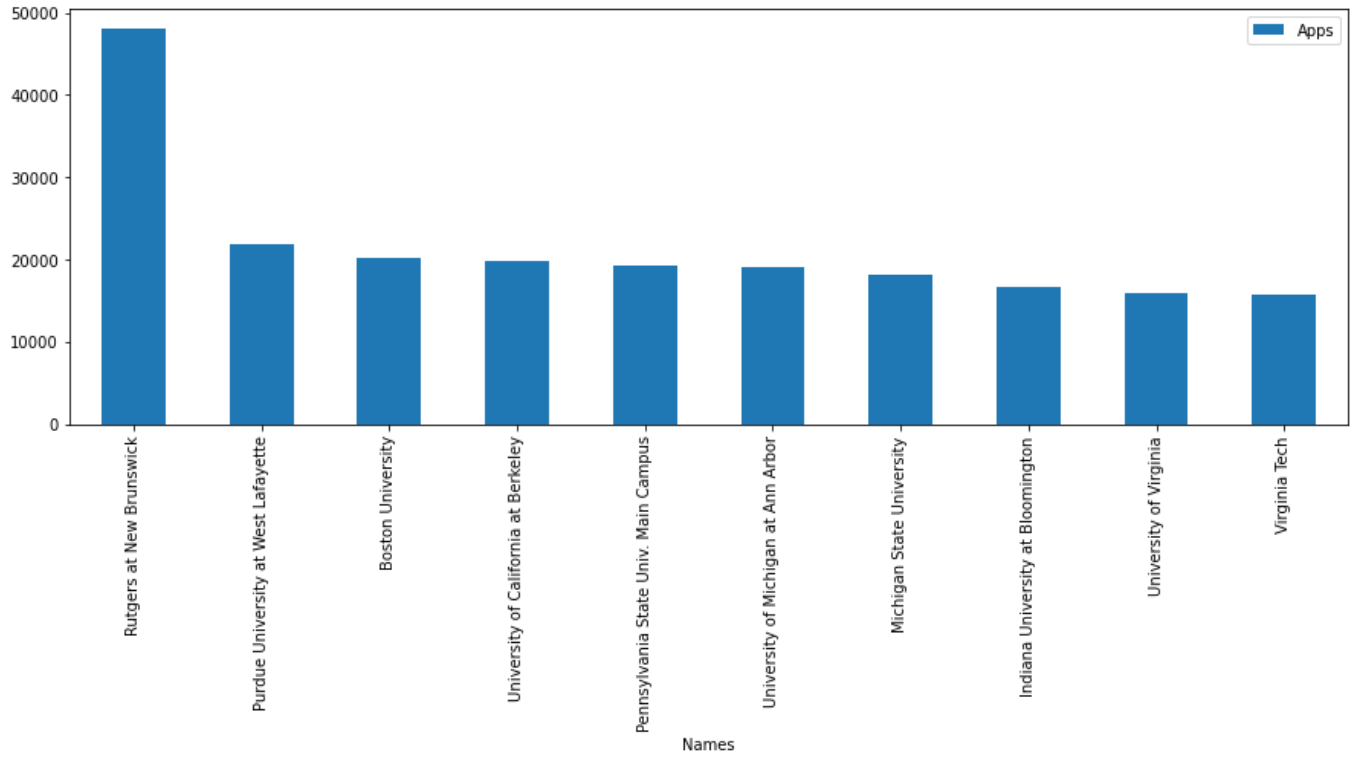
Positive or Right Skewed		Little skewness		Negative or Left skewed	
Top10perc	1.413217	Top25perc	0.259340	Terminal	-0.816542
Personal	1.742497	Room.Board	0.477356	PhD	-0.768170
F.Undergrad	2.610458	Outstate	0.509278	Grad.Rate	-0.113777
Enroll	2.690465	perc.alumni	0.606891		
Accept	3.417727	S.F.Ratio	0.667435		
Expend	3.459322				
Books	3.485025				
Apps	3.723750				
P.Undergrad	5.692353				

Final Interpretations drawn from Univariate Analysis:

- From the above univariate analysis, we can observe that except variable 'Top25perc' rest all the numerical variables contains outliers for a given dataset.
- Before further analysis like PCA, need to deal with outliers. Outliers need to be treated by imputing mean, median or IQR values into the dataset.
- We can also observe that 50% of the variables from a dataset is highly positive or right skewed data.
- We can also see that variables are not on the same scale. Before performing any predictive modelling, scaling needs to be done for a better results.

Multivariate Analysis:

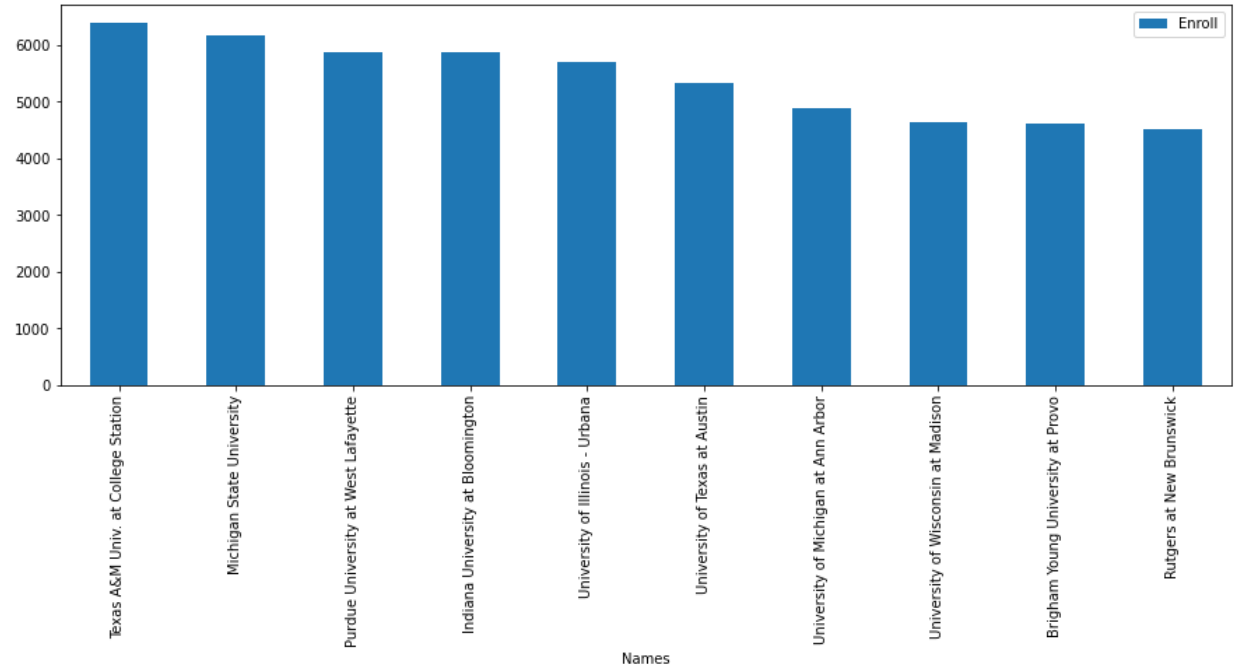
Output- The maximum number of applications received by the top 10 colleges



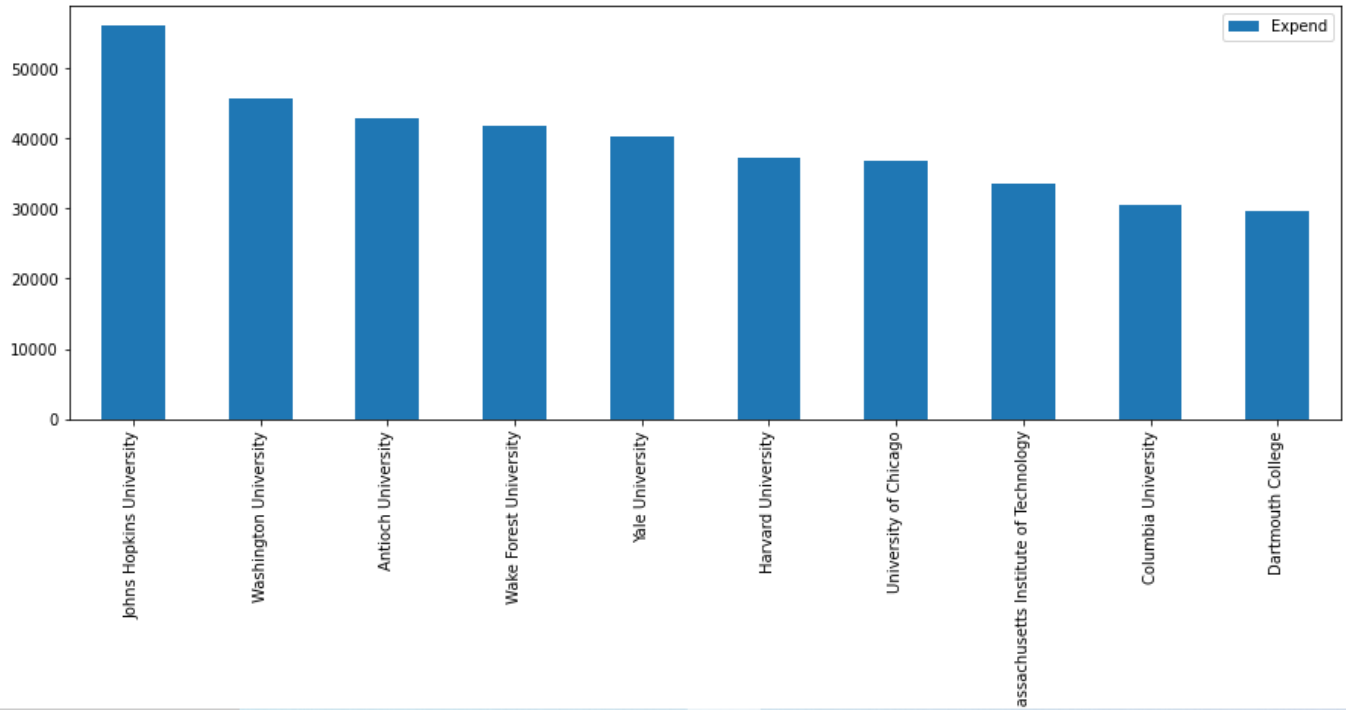
- From the above graph we can say that **Rutgers at New Brunswick** received the maximum number of applications.

- From the below graph we can say that the maximum students enrolled into Texas A&M university at college station.

Output- The maximum number of students enrolled into colleges



Output- The maximum instructional expenditure per students



- From the above graph we can see that the Johns Hopkins University has the high instructional expenditure per student compared to other colleges.

2.2 Scale the variables and write the inference for using the type of scaling function for this case study.

Solution:

- The variables of the dataset are of different scales, hence it is tough to compare these variables.
- Therefore, We have two types of scaling:

1. Feature Scaling-[Normalisation]:

Re-scaling the data or a variable to have values between 0 and 1. This technique is also known as Min-Max scaling.

Formula to achieve Feature scaling:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Summary of Scaled data post Outlier treatment:

We can observe that all the numerical variables lies within a range of min value 0 to maximum value 1.

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	0.318663	0.309942	0.0	0.088932	0.188996	0.453359	1.0
Accept	777.0	0.329453	0.299742	0.0	0.104683	0.204250	0.462810	1.0
Enroll	777.0	0.336774	0.307015	0.0	0.111470	0.214863	0.466882	1.0
Top10perc	777.0	0.403797	0.243477	0.0	0.218750	0.343750	0.531250	1.0
Top25perc	777.0	0.514249	0.217635	0.0	0.351648	0.494505	0.659341	1.0
F.Undergrad	777.0	0.333510	0.322012	0.0	0.101723	0.186989	0.461034	1.0
P.Undergrad	777.0	0.287988	0.314984	0.0	0.041337	0.154793	0.424802	1.0
Outstate	777.0	0.426494	0.211753	0.0	0.262209	0.402791	0.557325	1.0
Room.Board	777.0	0.472601	0.200141	0.0	0.333425	0.444077	0.600055	1.0
Books	777.0	0.508512	0.221596	0.0	0.375000	0.432692	0.625000	1.0
Personal	777.0	0.394051	0.223672	0.0	0.220183	0.348624	0.532110	1.0
PhD	777.0	0.599666	0.211300	0.0	0.456954	0.629139	0.761589	1.0
Terminal	777.0	0.665826	0.239224	0.0	0.520661	0.702479	0.867769	1.0
S.F.Ratio	777.0	0.502561	0.189211	0.0	0.375000	0.480000	0.625000	1.0
perc.alumni	777.0	0.391759	0.212508	0.0	0.224138	0.362069	0.534483	1.0
Expend	777.0	0.435715	0.246794	0.0	0.259037	0.377184	0.555422	1.0
Grad.Rate	777.0	0.499685	0.171425	0.0	0.375000	0.495000	0.625000	1.0

2. Standardization or Z-Transformation:

Transforming the data or a variable using a **z-score** to have a mean of zero and a standard deviation of 1. This technique is known as standardization or a Z-transformation.

Formula to achieve Feature scaling:

$$X' = \frac{X - \mu}{\sigma}$$

Summary of Scaled data post Outlier treatment:

We can see all the Numerical variables are now centered on the mean with a unit standard deviation.

	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

Final Interpretations:

- When the distribution of our data does not follow a Gaussian distribution, Feature scaling or Normalisation can be used.
- When the distribution of our data follow a Gaussian distribution, Standardization or Z-transformation can be used.
- So, selecting any type of the scaling technique depends on the problem statement or a kind predictive modelling we are using.

2.3 Comment on the comparison between covariance and the correlation matrix after scaling.

Solution:

Both Covariance and the correlation are used to determine the linear relationship and measure the dependency between the two random variables.

1. Covariance matrix:

- Covariance is nothing but a measure of correlation.
- Covariance indicates the direction of the linear relationship between variables.
- Covariance can vary between -∞ and +∞.
- Covariance is affected by the change in scale.

2. Correlation matrix:

- Correlation refers to the scaled form of covariance.
- Correlation measures both the strength and direction of the linear relationship between two variables.
- Correlation ranges between -1 and +1.
- Correlation is not influenced by the change in scale.

We can also state that the below three approaches yield the same eigenvectors and eigenvalue pairs:

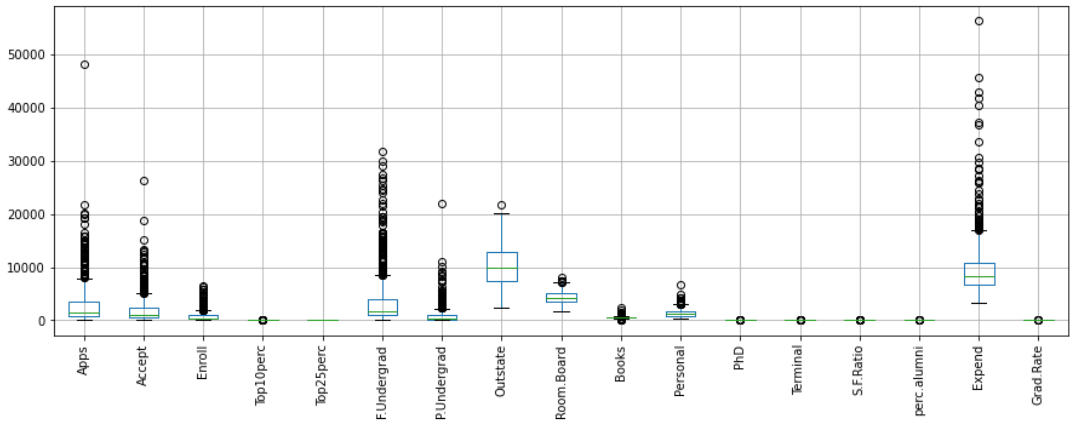
✓ Eigen decomposition of the covariance matrix after standardizing the data.
✓ Eigen decomposition of the correlation matrix.
✓ Eigen decomposition of the correlation matrix after standardizing the data.

Finally we can say that after scaling - the covariance and the correlation have the same values.

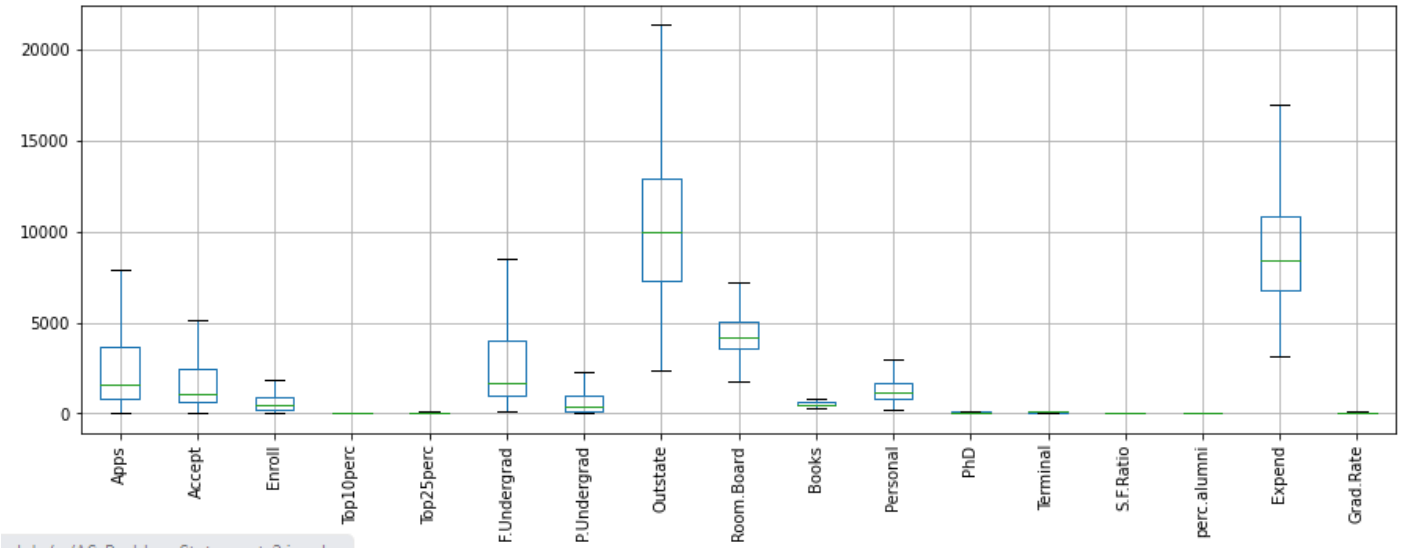
2.4 Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

Solution:

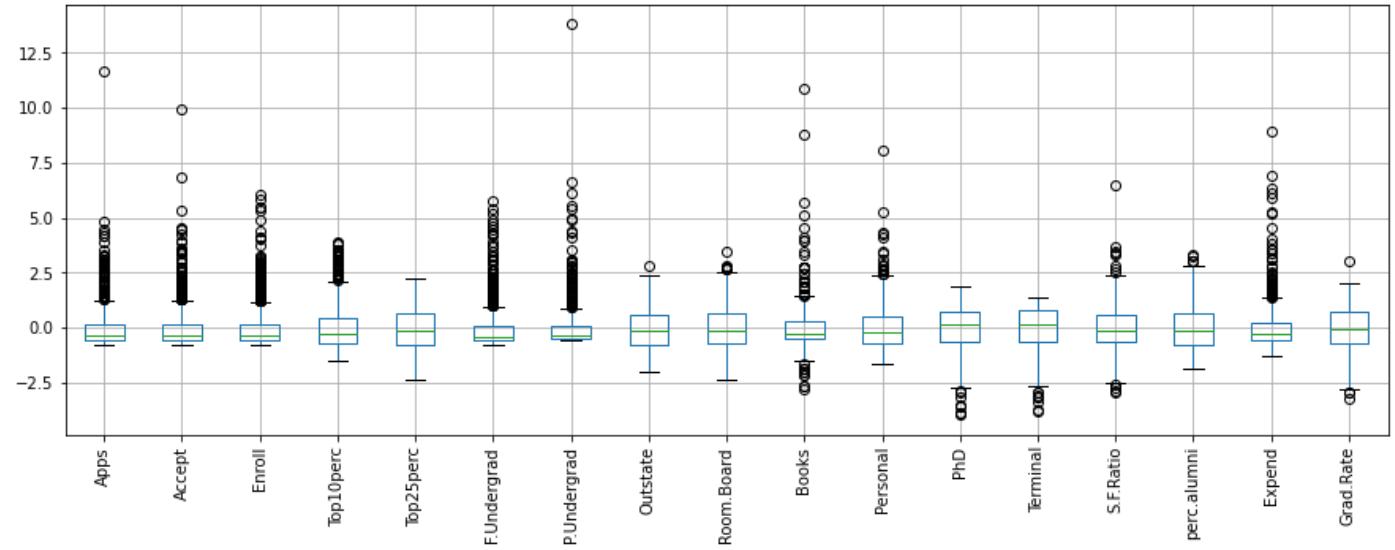
Output-1: Boxplot to identify Outliers for an Unscaled dataset:



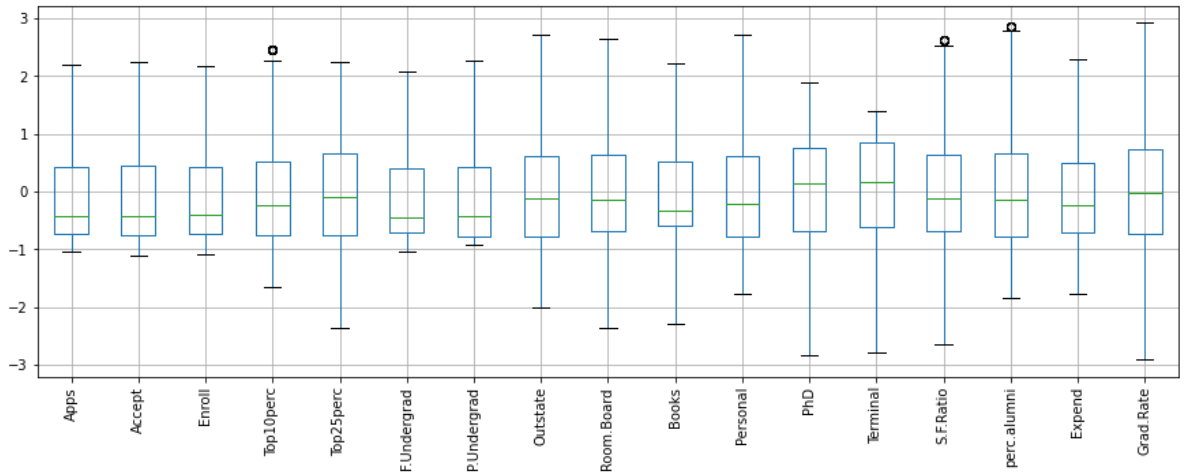
Output-1: Boxplot to identify Outliers for an Unscaled dataset post –Outlier Treatment:



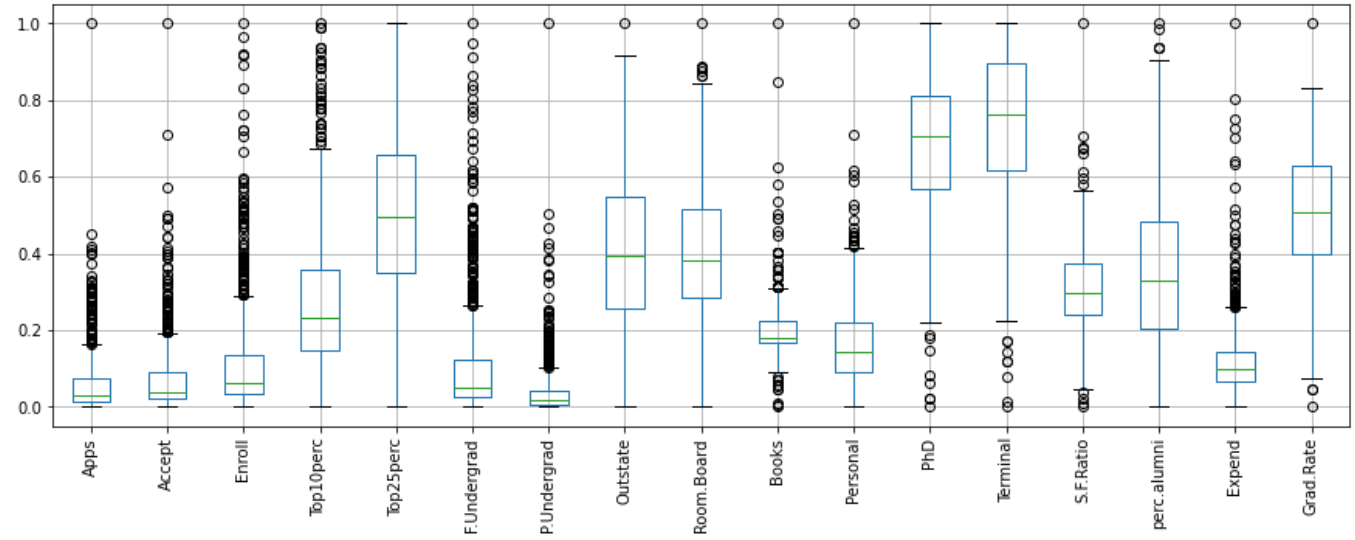
Output-2: Boxplot to identify Outliers for a Scaled data [Standardization]:



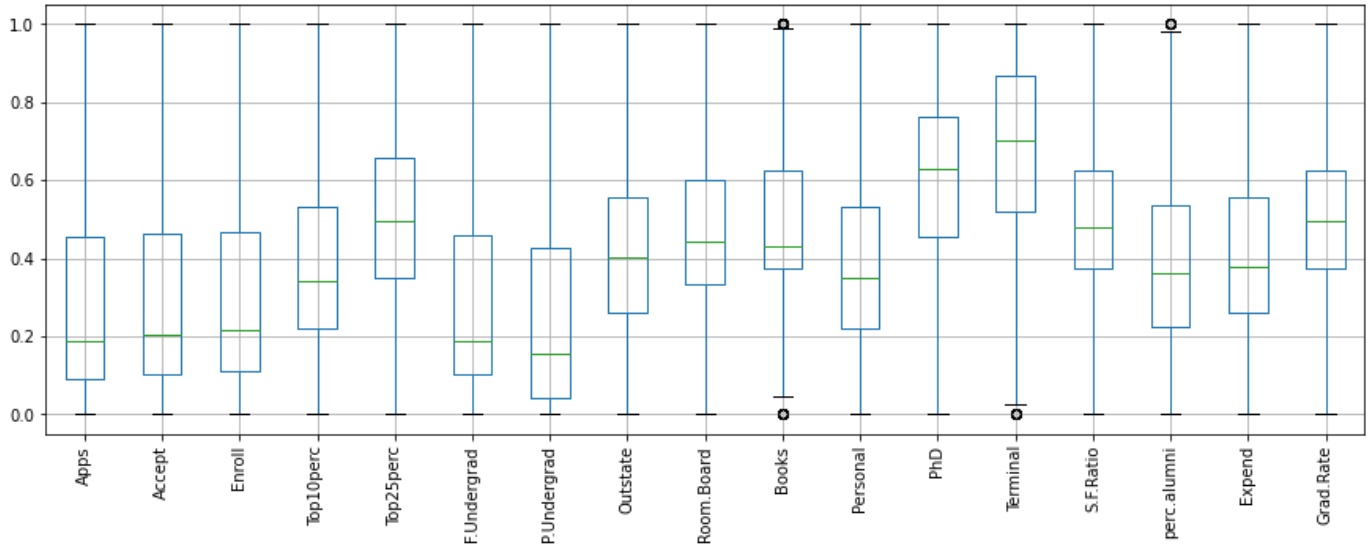
Output-2: Boxplot to identify Outliers for a Scaled data [Standardization] Post –Outlier Treatment:



Output-3: Boxplot to identify Outliers for a Scaled data [Normalisation]:



Output-3: Boxplot to identify Outliers for a Scaled data [Normalisation] Post –Outlier Treatment:



2.5 Build the covariance matrix, eigenvalues, and eigenvector.

Solution:

Covariance matrix:

Covariance matrix is a symmetric, square matrix where the variance of each feature and the cross-features covariance's are stored.

```
Covariance Matrix
%s [[ 1.00128866e+00  9.56537704e-01  8.98039052e-01  3.21756324e-01
   3.64960691e-01  8.62111140e-01  5.20492952e-01  6.54209711e-02
   1.87717056e-01  2.36441941e-01  2.30243993e-01  4.64521757e-01
   4.35037784e-01  1.26573895e-01 -1.01288006e-01  2.43248206e-01
   1.50997775e-01]
 [ 9.56537704e-01  1.00128866e+00  9.36482483e-01  2.23586208e-01
   2.74033187e-01  8.98189799e-01  5.73428908e-01 -5.00874847e-03
   1.19740419e-01  2.08974091e-01  2.56676290e-01  4.27891234e-01
   4.03929238e-01  1.88748711e-01 -1.65728801e-01  1.62016688e-01
   7.90839722e-02]
 [ 8.98039052e-01  9.36482483e-01  1.00128866e+00  1.71977357e-01
   2.30730728e-01  9.68548601e-01  6.42421828e-01 -1.55856056e-01
  -2.38762560e-02  2.02317274e-01  3.39785395e-01  3.82031198e-01
   3.54835877e-01  2.74622251e-01 -2.23009677e-01  5.42906862e-02
  -2.32810071e-02]
 [ 3.21756324e-01  2.23586208e-01  1.71977357e-01  1.00128866e+00
   9.15052977e-01  1.11358019e-01 -1.80240778e-01  5.62884044e-01
   3.57826139e-01  1.53650150e-01 -1.16880152e-01  5.44748764e-01
   5.07401238e-01 -3.88425719e-01  4.56384036e-01  6.57885921e-01
   4.94306540e-01]
 [ 3.64960691e-01  2.74033187e-01  2.30730728e-01  9.15052977e-01
   1.00128866e+00  1.81429267e-01 -9.94231153e-02  4.90200034e-01
   3.31413314e-01  1.69979808e-01 -8.69219644e-02  5.52172085e-01
   5.28333659e-01 -2.97616423e-01  4.17369123e-01  5.73643193e-01
   4.79601950e-01]
 [ 8.62111140e-01  8.98189799e-01  9.68548601e-01  1.11358019e-01
   1.81429267e-01  1.00128866e+00  6.97027420e-01 -2.26457040e-01
  -5.45459528e-02  2.08147257e-01  3.60246460e-01  3.62030390e-01
   3.35485771e-01  3.24921933e-01 -2.85825062e-01  3.71119607e-04
  -8.23447851e-02]
 [ 5.20492952e-01  5.73428908e-01  6.42421828e-01 -1.80240778e-01
  -9.94231153e-02  6.97027420e-01  1.00128866e+00 -3.54672874e-01
  -6.77252009e-02  1.22686416e-01  3.44495974e-01  1.27827147e-01
   1.22309141e-01  3.71084841e-01 -4.19874031e-01 -2.02189396e-01
  -2.65499420e-01]
 [ 6.54209711e-02 -5.00874847e-03 -1.55856056e-01  5.62884044e-01
   4.90200034e-01 -2.26457040e-01 -3.54672874e-01  1.00128866e+00
   6.56333564e-01  5.11656377e-03 -3.26028927e-01  3.91824814e-01
   4.13110264e-01 -5.74421963e-01  5.66465309e-01  7.76326650e-01
   5.73195743e-01]
 [ 1.87717056e-01  1.19740419e-01 -2.38762560e-02  3.57826139e-01
   3.31413314e-01 -5.45459528e-02 -6.77252009e-02  6.56333564e-01
   1.00128866e+00  1.09064551e-01 -2.19837042e-01  3.41908577e-01
   3.79759015e-01 -3.76915472e-01  2.72743761e-01  5.81370284e-01
   4.26338910e-01]
 [ 2.36441941e-01  2.08974091e-01  2.02317274e-01  1.53650150e-01
   1.69979808e-01  2.08147257e-01  1.22686416e-01  5.11656377e-03
   1.09064551e-01  1.00128866e+00  2.40172145e-01  1.36566243e-01
   1.59523091e-01 -8.54689129e-03 -4.28870629e-02  1.50176551e-01
  -8.06107505e-03]
```

```
[ 2.30243993e-01  2.56676290e-01  3.39785395e-01 -1.16880152e-01
 -8.69219644e-02  3.60246460e-01  3.44495974e-01 -3.26028927e-01
 -2.19837042e-01  2.40172145e-01  1.00128866e+00 -1.16986124e-02
 -3.20117803e-02  1.74136664e-01 -3.06146886e-01 -1.63481407e-01
 -2.91268705e-01]
[ 4.64521757e-01  4.27891234e-01  3.82031198e-01  5.44748764e-01
  5.52172085e-01  3.62030390e-01  1.27827147e-01  3.91824814e-01
  3.41908577e-01  1.36566243e-01 -1.16986124e-02  1.00128866e+00
  8.64040263e-01 -1.29556494e-01  2.49197779e-01  5.11186852e-01
  3.10418895e-01]
[ 4.35037784e-01  4.03929238e-01  3.54835877e-01  5.07401238e-01
  5.28333659e-01  3.35485771e-01  1.22309141e-01  4.13110264e-01
  3.79759015e-01  1.59523091e-01 -3.20117803e-02  8.64040263e-01
  1.00128866e+00 -1.51187934e-01  2.66375402e-01  5.24743500e-01
  2.93180212e-01]
[ 1.26573895e-01  1.88748711e-01  2.74622251e-01 -3.88425719e-01
 -2.97616423e-01  3.24921933e-01  3.71084841e-01 -5.74421963e-01
 -3.76915472e-01 -8.54689129e-03  1.74136664e-01 -1.29556494e-01
 -1.51187934e-01  1.00128866e+00 -4.12632056e-01 -6.55219504e-01
 -3.08922187e-01]
[-1.01288006e-01 -1.65728801e-01 -2.23009677e-01  4.56384036e-01
  4.17369123e-01 -2.85825062e-01 -4.19874031e-01  5.66465309e-01
  2.72743761e-01 -4.28870629e-02 -3.06146886e-01  2.49197779e-01
  2.66375402e-01 -4.12632056e-01  1.00128866e+00  4.63518674e-01
  4.92040760e-01]
 [ 2.43248206e-01  1.62016688e-01  5.42906862e-02  6.57885921e-01
   5.73643193e-01  3.71119607e-04 -2.02189396e-01  7.76326650e-01
   5.81370284e-01  1.50176551e-01 -1.63481407e-01  5.11186852e-01
   5.24743500e-01 -6.55219504e-01  4.63518674e-01  1.00128866e+00
   4.15826026e-01]
 [ 1.50997775e-01  7.90839722e-02 -2.32810071e-02  4.94306540e-01
   4.79601950e-01 -8.23447851e-02 -2.65499420e-01  5.73195743e-01
   4.26338910e-01 -8.06107505e-03 -2.91268705e-01  3.10418895e-01
   2.93180212e-01 -3.08922187e-01  4.92040760e-01  4.15826026e-01
   1.00128866e+00]]
```

Eigen Values:

```
Eigen Values
%s [5.6625219  4.89470815 1.12636744 1.00397659 0.87218426 0.7657541
0.58491404 0.5445048  0.42352336 0.38101777 0.24701456 0.02239369
0.03789395 0.14726392 0.13434483 0.09883384 0.07469003]
```

Eigen Vectors:

```
Eigen Vectors
 %s [[-2.62171542e-01  3.14136258e-01 -8.10177245e-02  9.87761685e-02
   2.19898081e-01 -2.18800617e-03  2.83715076e-02 -8.99498102e-02
   1.30566998e-01 -1.56464458e-01 -8.62132843e-02  1.82169814e-01
  -5.99137640e-01  8.99775288e-02 -8.88697944e-02  5.49428396e-01
   5.41453698e-03]
 [-2.30562461e-01  3.44623583e-01 -1.07658626e-01  1.18140437e-01
   1.89634940e-01  1.65212882e-02  1.29584896e-02 -1.37606312e-01
   1.42275847e-01 -1.49209799e-01 -4.25899061e-02 -3.91041719e-01
   6.61496927e-01  1.58861886e-01 -4.37945938e-02  2.91572312e-01
   1.44582845e-02]
 [-1.89276397e-01  3.82813322e-01 -8.55296892e-02  9.30717094e-03
   1.62314818e-01  6.80794143e-02  1.52403625e-02 -1.44216938e-01
   5.08712481e-02 -6.48997860e-02 -4.38408622e-02  7.16684935e-01
   2.33235272e-01 -3.53988202e-02  6.19241658e-02 -4.17001280e-01
  -4.97908902e-02]
 [-3.38874521e-01 -9.93191661e-02  7.88293849e-02 -3.69115031e-01
   1.57211016e-01  8.88656824e-02  2.57455284e-01  2.89538833e-01
  -1.22467790e-01 -3.58776186e-02  1.77837341e-03 -5.62053913e-02
   2.21448729e-02 -3.92277722e-02 -6.99599977e-02  8.79767299e-03
  -7.23645373e-01]
 [-3.34690532e-01 -5.95055011e-02  5.07938247e-02 -4.16824361e-01
   1.44449474e-01  2.76268979e-02  2.39038849e-01  3.45643551e-01
  -1.93936316e-01  6.41786425e-03 -1.02127328e-01  1.96735274e-02
   3.22646978e-02  1.45621999e-01  9.70282598e-02 -1.07779150e-02
   6.55464648e-01]

  [-1.63293010e-01  3.98636372e-01 -7.37077827e-02  1.39504424e-02
    1.02728468e-01  5.16468727e-02  3.11751439e-02 -1.08748900e-01
    1.45452749e-03 -1.63981359e-04 -3.49993487e-02 -5.42774834e-01
   -3.67681187e-01 -1.33555923e-01  8.71753137e-02 -5.70683843e-01
    2.53059904e-02]
  [-2.24797091e-02  3.57550046e-01 -4.03568700e-02  2.25351078e-01
   -9.56790178e-02  2.45375721e-02  1.00138971e-02  1.23841696e-01
   -6.34774326e-01  5.46346279e-01  2.52107094e-01  2.95029745e-02
    2.62494456e-02  5.02487566e-02 -4.45537493e-02  1.46321060e-01
   -3.97146972e-02]
  [-2.83547285e-01 -2.51863617e-01 -1.49394795e-02  2.62975384e-01
    3.72750885e-02  2.03860462e-02 -9.45370782e-02  1.12721477e-02
   -8.36648339e-03 -2.31799759e-01  5.93433149e-01  1.03393587e-03
   -8.14247697e-02  5.60392799e-01 -6.72405494e-02 -2.11561014e-01
   -1.59275617e-03]
  [-2.44186588e-01 -1.31909124e-01  2.11379165e-02  5.80894132e-01
   -6.91080879e-02 -2.37267409e-01 -9.45210745e-02  3.89639465e-01
   -2.20526518e-01 -2.55107620e-01 -4.75297296e-01  9.85725168e-03
    2.67779296e-02 -1.07365653e-01 -1.77715010e-02 -1.00935084e-01
   -2.82578388e-02]
  [-9.67082754e-02  9.39739472e-02  6.97121128e-01 -3.61562884e-02
    3.54056654e-02 -6.38604997e-01  1.11193334e-01 -2.39817267e-01
    2.10246624e-02  9.11624912e-02  4.35697999e-02  4.36086500e-03
    1.04624246e-02  5.16224550e-02 -3.54343707e-02 -2.86384228e-02
   -8.06259380e-03]
```

```
[ 3.52299594e-02  2.32439594e-01  5.30972806e-01 -1.14982973e-01
 -4.75358244e-04  3.81495854e-01 -6.39418106e-01  2.77206569e-01
  1.73715184e-02 -1.27647512e-01  1.51627393e-02 -1.08725257e-02
  4.54572099e-03  9.39409228e-03  1.18604404e-02  3.38197909e-02
  1.42590097e-03]
[-3.26410696e-01  5.51390195e-02 -8.11134044e-02 -1.47260891e-01
 -5.50786546e-01 -3.34444832e-03 -8.92320786e-02 -3.42628480e-02
  1.66510079e-01  1.00975002e-01 -3.91865961e-02  1.33146759e-02
  1.25137966e-02 -7.16590441e-02 -7.02656469e-01 -6.38096394e-02
  8.31471932e-02]
[-3.23115980e-01  4.30332048e-02 -5.89785929e-02 -8.90079921e-02
 -5.90407136e-01 -3.54121294e-02 -9.16985445e-02 -9.03076644e-02
  1.12609034e-01  8.60363025e-02 -8.48575651e-02  7.38135022e-03
 -1.79275275e-02  1.63820871e-01  6.62488717e-01  9.85019644e-02
 -1.13374007e-01]
[ 1.63151642e-01  2.59804556e-01 -2.74150657e-01 -2.59486122e-01
 -1.42842546e-01 -4.68752604e-01 -1.52864837e-01  2.42807562e-01
 -1.53685343e-01 -4.70527925e-01  3.63042716e-01  8.85797314e-03
  1.83059753e-02 -2.39902591e-01  4.79006197e-02  6.19970446e-02
  3.83160891e-03]
[-1.86610828e-01 -2.57092552e-01 -1.03715887e-01 -2.23982467e-01
  1.28215768e-01 -1.25669415e-02 -3.91400512e-01 -5.66073056e-01
 -5.39235753e-01 -1.47628917e-01 -1.73918533e-01 -2.40534190e-02
 -8.03169296e-05 -4.89753356e-02 -3.58875507e-02  2.80805469e-02
 -7.32598621e-03]

[-3.28955847e-01 -1.60008951e-01  1.84205687e-01  2.13756140e-01
 -2.24240837e-02  2.31562325e-01  1.50501305e-01 -1.18823549e-01
  2.42371616e-02 -8.04154875e-02  3.93722676e-01  1.05658769e-02
  5.60069250e-02 -6.90417042e-01  1.26667522e-01  1.28739213e-01
  1.45099786e-01]
[-2.38822447e-01 -1.67523664e-01 -2.45335837e-01 -3.61915064e-02
  3.56843227e-01 -3.13556243e-01 -4.68641965e-01  1.80458508e-01
  3.15812873e-01  4.88415259e-01  8.72638706e-02 -2.51028410e-03
  1.48410810e-02 -1.59332164e-01  6.30737002e-02 -7.09643331e-03
 -3.29024228e-03]]
```

2.6 Write the explicit form of the first PC (in terms of Eigen Vectors).

Solution:

To calculate the explicit form of the first PC, need to calculate matrix multiplication of the original data with the eigenvectors

Output: Top 10 Eigen vectors of PCA_1:

```
array([ 1.60249937,  1.80467545,  1.60828257, -2.80364434,  2.20086769,
        0.73016388, -0.00451649, -1.83606666, -0.61923075,  2.93435271])
```

The variance of the first principal component is 5.66 within our dataset.

2.7 Discuss 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?
Perform PCA and export the data of the Principal Component scores into a data frame.

Solution:

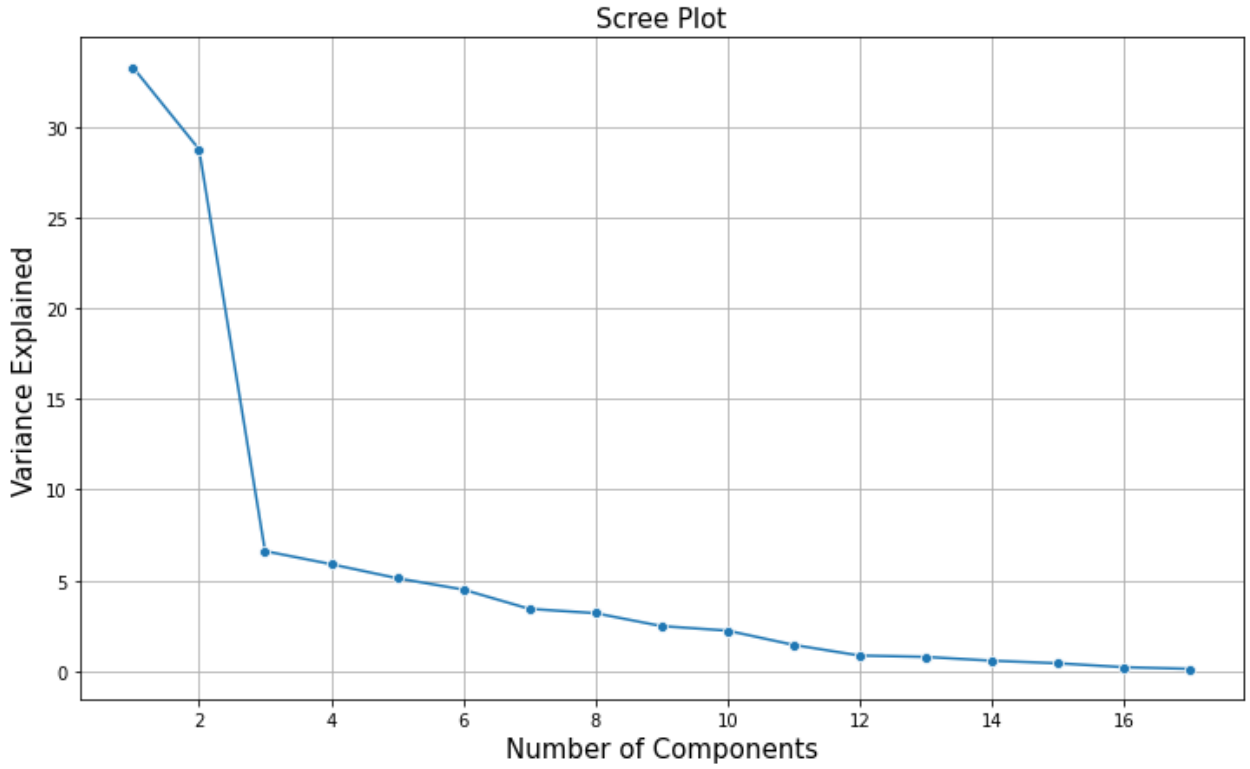
The Cumulative % gives the percentage of variance accounted for by the n components. For example, the cumulative percentage for the second component is the sum of the percentage of variance for the first and second components. It helps in deciding the number of components by selecting the components which explained the high variance.

Output: Cumulative Distribution of Eigenvalues

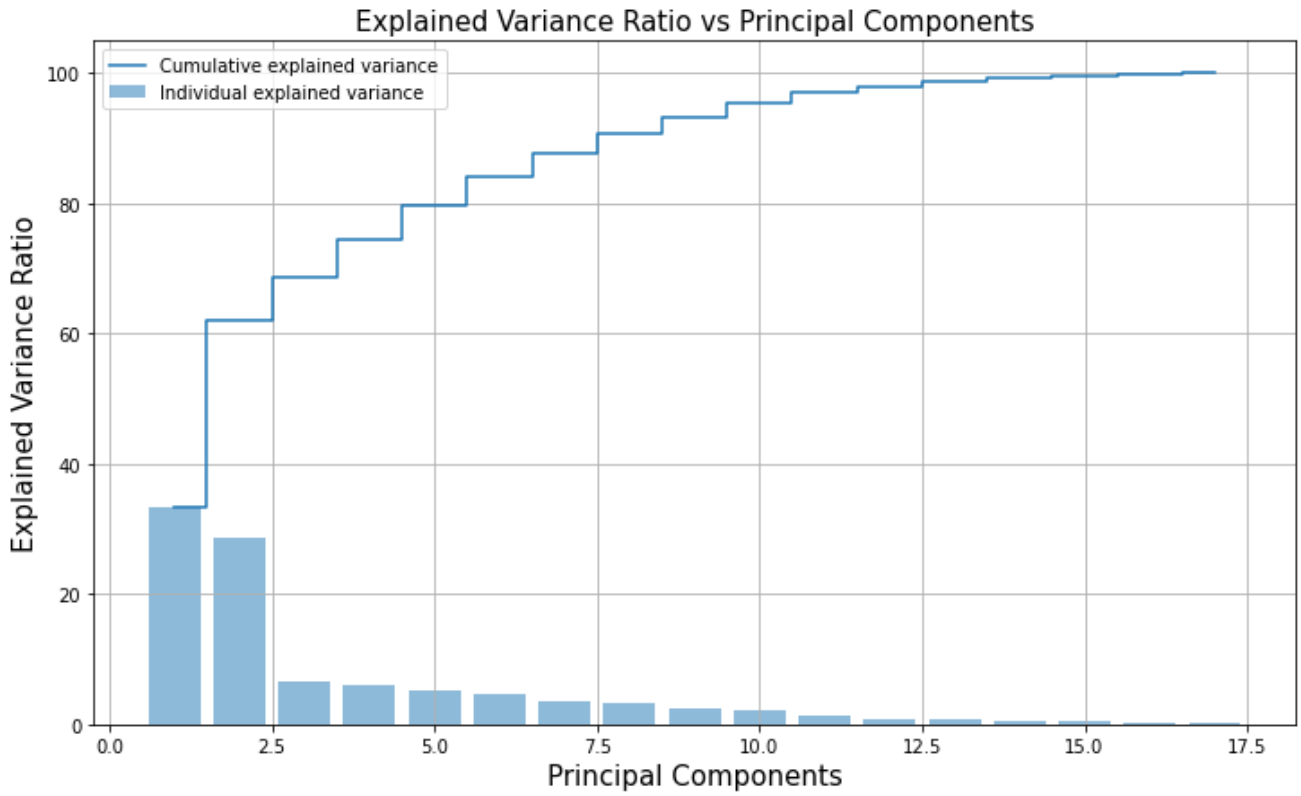
```
Cumulative Variance Explained [ 33.26608367  62.02142867  68.63859223  74.53673619  79.66062886
  84.15926753  87.59551019  90.79435736  93.28246491  95.52086136
  96.97201814  97.83716159  98.62640821  99.20703552  99.64582321
  99.86844192 100.        ]
```

In the above array we see that the first feature explains 33.26% of the variance within our data set while the first two explain 62.02 and so on. If we employ 5 features we capture ~ 80% of the variance within the dataset.

Output: Scree-Plot



Output: Plot Cumulative explained variance and individual explained variance vs Principal Components



From the above graph we can observe that first 5 Principal components explains 80% of the variance within our dataset.

- Generally, the Eigen values which are greater than one help's us to decide on the exact optimum number of principal components to be selected or considered.
- The eigenvectors represent the directions or components for the reduced subspace of original dataset, whereas the eigenvalues represent the magnitudes for the directions.

Output: Export the data of the Principal Component scores into a data frame.

	principal component 1	principal component 2	principal component 3	principal component 4	principal component 5
0	1.602499	0.993683	0.030045	-0.366886	0.697476
1	1.804675	-0.070415	2.122128	2.453212	-0.994859
2	1.608283	-1.382792	-0.501513	0.765997	1.026237
3	-2.803644	-3.367395	0.367768	-1.192601	1.457080
4	2.200868	-0.099348	3.122523	-1.828044	-0.140915
5	0.730164	-1.998741	0.237171	0.062740	0.821044
6	-0.004516	-1.884603	0.237183	-1.878438	0.132645
7	-1.836067	-1.733341	-0.995891	-0.996701	0.117335
8	-0.619231	-2.459100	-1.823771	-0.341261	0.977575
9	2.934353	-1.106131	2.142631	1.926359	0.320840

2.8 Mention the business implication of using the Principal Component Analysis for this case study. [Hint: Write Interpretations of the Principal Components Obtained]

Solution:

- The given dataset has 17 dimension, in order to reduce the dimensionality to fewer components, PCA needs to be done.
- After performing PCA, we could observe that first 4 principal components explains ~75% of variance within our data. Whereas remaining ~25% of the data would be loss. To reduce the data loss we can still add 2 more features i.e if we consider 6 principal components which explains about ~85% of variance within our data.
- So finally, by performing PCA, we can reduce the dimensionality from 17 features to either 4 or 6 features depending on the business needs.