Fever

Problem Statement:

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.

[Assume all of the ANOVA assumptions are satisfied]

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]

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.

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.

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 'pointplot' function from the 'seaborn' function]

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.

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

Exploratory Data Analysis:

The first 10 rows of the data:

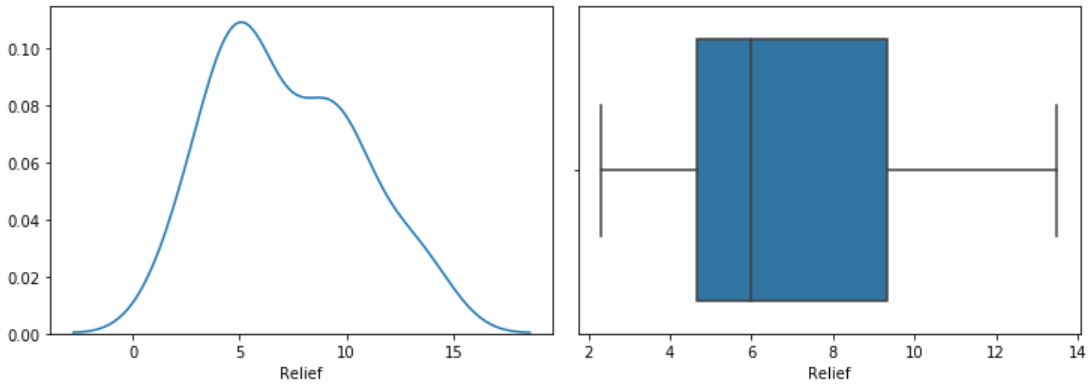
	A	B	Volunteer	Relief
0	1	1	1	2.4
1	1	1	2	2.7
2	1	1	3	2.3
3	1	1	4	2.5
4	1	2	1	4.6

The data has 36 entries (rows) and 4 variables (columns). From the definition of levels of compunds we can say that **A**, **B** are Categorical in nature, **Volunteer** and **Relief** are retained to be integer data type.

#	Column	Non-NullCount	Dtype
0	A	36 non-null	category
1	B	36 non-null	category
2	Volunteer	36 non-null	int64
3	Relief	36 non-null	float64

The summary for the Relief variable is shown below:

count	mean	std	min	25%	50%	75%	max
36	7.183333	3.27209	2.3	4.675	6	9.325	13.5



The Relief variable is slightly right skewed and has no outliers.

The summary for categorical variables 'A' and 'B' are shown below:

	count	unique	top	freq
A	36	3	3	12
B	36	3	3	12

Findings:

1. Column headers are 'A', 'B', 'Volunteer' and 'Relief'
2. There are 36 entries (rows) and 4 variables (columns)
3. The data types of the variables are:
4. There are no missing values
5. A and B variables have 3 categories each with 12 observations in each of the categories
6. The variable Relief is slightly right skewed
7. There are no outliers in the Relief variable

Question 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]

Answer:

1. Hypothesis for A (one-way):

   Numerical:

   - $H_0 : \mu_{A1} = \mu_{A2} = \mu_{A3}$
   - $H_a : \mu_{A1} = \mu_{A2} \neq \mu_{A3}$ or $\mu_{A1} \neq \mu_{A2} \neq \mu_{A3}$ or $\mu_{A1} \neq \mu_{A2} = \mu_{A3}$

   In lay man's terms

   - H_0 : The means of fever for each of the level in compound A are equal
   - H_a : There is at least one level has a mean different compared to the other two in compound A.

2. Hypothesis for B (one-way):

   Numerical:

- $H_0 : \mu_{B1} = \mu_{B2} = \mu_{B3}$
- $H_b : \mu_{B1} = \mu_{B2} \neq \mu_{B3}$ or $\mu_{B1} \neq \mu_{B2} \neq \mu_{B3}$ or $\mu_{B1} \neq \mu_{B2} = \mu_{B3}$

In lay man's terms:

- H_0 : The means of fever for each of the level in compound B are equal
- H_b : There is at least one level has a mean different compared to the other two in Compound B.

Question 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.

Answer:

Hypothesis:

- $H_0 : \mu_{A1} = \mu_{A2} = \mu_{A3}$
- $H_a : \mu_{A1} = \mu_{A2} \neq \mu_{A3}$ or $\mu_{A1} \neq \mu_{A2} \neq \mu_{A3}$ or $\mu_{A1} \neq \mu_{A2} = \mu_{A3}$

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2	220.02	110.01	23.46539	4.58E-07
Residual	33	154.71	4.688182	NaN	NaN

P= 4.578242e-07

Since P < 0.05, we can reject the null hypothesis.

Considering a confidence ratio of 95% we can say that the means for fever for each amount level of compound A are not equal.

Therefore, we can say that the means of relief for each level in compound A are not equal.

Question 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.

Answer:

Hypothesis:

- $H_0 : \mu_{B1} = \mu_{B2} = \mu_{B3}$
- $H_b : \mu_{B1} = \mu_{B2} \neq \mu_{B3}$ or $\mu_{B1} \neq \mu_{B2} \neq \mu_{B3}$ or $\mu_{B1} \neq \mu_{B2} = \mu_{B3}$

	df	sum_sq	mean_sq	F	PR(>F)
C(B)	2	123.66	61.83	8.126777	1.35E-03
Residual	33	251.07	7.608182	NaN	NaN

P=0.00135
Since P < 0.05, we can reject the null hypothesis.

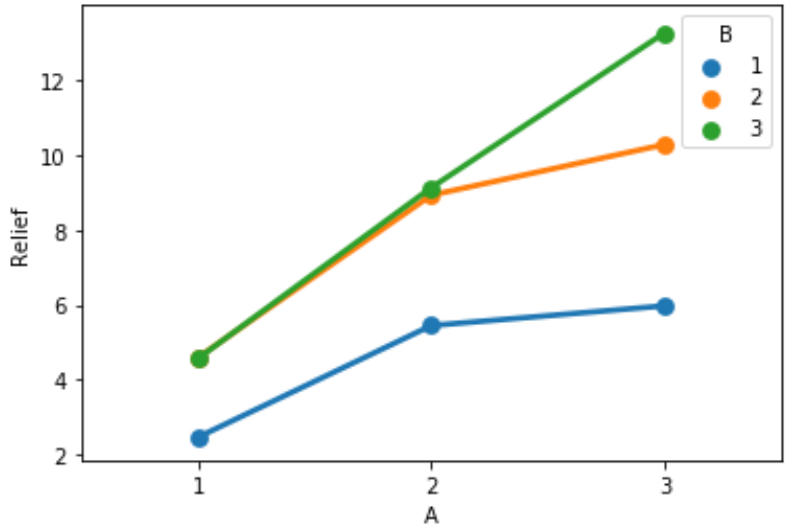
Considering a confidence ratio of 95% we can say that the means for fever for each amount level of compound A are not equal.

Therefore, we can say that the means of relief for each level in compound B are not equal.

Question 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 'pointplot' function from the 'seaborn' function]

Answer:
The interaction plot between Relief and the 2 compounds is shown below:



Findings:

1. On overview the order of relief is as follows for both the compounds: **1 < 2 < 3**
2. With increase in levels in compound A, all levels of compound B increase in relief.
3. Levels 1 and 2 of compound B, for all levels in compound A, show almost similar behaviour with respect to relief.
4. Level 3 of compound B shows the same character till level 2 of compound A but however, after that it continues to increase while the other two levels tend to saturate in relief.
5. There is not much interaction between level 1 and 2 across levels in compound A. While level 3 of compound B shows considerable interaction with the remaining 2 levels of compound B.

6. Overall there is very small amount of interaction which we will verify statistically in the next question.

Question 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.

Answer:

Hypothesis for compound A:

- $H_0 : \mu_{A1} = \mu_{A2} = \mu_{A3}$
- $H_a : \mu_{A1} = \mu_{A2} \neq \mu_{A3}$ or $\mu_{A1} \neq \mu_{A2} \neq \mu_{A3}$ or $\mu_{A1} \neq \mu_{A2} = \mu_{A3}$

Hypothesis for compound B:

- $H_0 : \mu_{B1} = \mu_{B2} = \mu_{B3}$
- $H_b : \mu_{B1} = \mu_{B2} \neq \mu_{B3}$ or $\mu_{B1} \neq \mu_{B2} \neq \mu_{B3}$ or $\mu_{B1} \neq \mu_{B2} = \mu_{B3}$

The hypothesis for interaction between the compounds:

- H_0: There is no interaction effect of Compound A and Compound B on Relief.
- $H_{A:B}$: There is interaction effect of Compound A and Compound B on Relief.

	df	sum_sq	mean_sq	F	PR(>F)
C(A)	2	220.02	110.01	1827.858	1.51E-29
C(B)	2	123.66	61.83	1027.329	3.35E-26
C(A):C(B)	4	29.425	7.35625	122.2269	6.97E-17
Residual	27	1.625	0.060185	NaN	NaN

P[C(A)] = 1.514043e-29
P[C(B)] = 3.348751e-26
P[C(A):C(B)] = 6.972083e-17

- Compound A has a significantly more effect on Relief when compared to B and interaction between the two.
- Compound B has a significantly lower effect on Relief when compared to A and higher when compared to the interaction between the two.
- At 95% confidence, we can say that there is interaction between the compounds. The model has to be balanced by adjusting for type 3 error.

Question 6:

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

Answer:

Findings:

3. On overview the order of relief is as follows for both the compounds: 1 < 2 < 3

4. With increase in levels in compound A, all levels of compound B increase in relief.
5. Levels 1 and 2 of compound B, for all levels in compound A, show almost similar behaviour with respect to relief.
6. Level 3 of compound B shows the same character till level 2 of compound A but however, after that it continues to increase while the other two levels tend to saturate in relief.
7. There is not much interaction between level 1 and 2 across levels in compound A. While level 3 of compound B shows considerable interaction with the remaining 2 levels of compound B.
8. Overall there is a considerable amount of interaction. We will verify statistically in the next question if this interaction is considerable enough.
9. Compound A has a significant higher effect on Relief compared to compound B.
10. Compound B has a significant higher effect on Relief compared to compound A.
11. At 95% confidence, we can say that there is interaction between the compounds.

Business Implications on this case study are:

- The data collected at each level is equal and the experiment design has been conducted in a fairly random and appropriate manner. 12 in each level for each compound.
- Compound A has a comparatively higher effect on the relief compared to compound B.
- Across all levels of compound B, relief for level 3 is greater than level 2 than level1 in compound A.
- Across all levels of A, levels 1 and 2 of compound B perform similarly. However, level 3 in compound B shows same results as level 2 in compound B until level 2 in compound A but for level 3 in compound A and B there is an abrupt increase in relief.
- This abruptly high relief is what may be causing the relief to skew slightly to the right. This needs special attention and more study so as to be proven to be more consistent.
- If proven consistent, the combination of Level 3 of compound A and B can be used to give high relief and the combination of Level 1 of compound A and B can be used to give every mild relief. All other combinations provide intermediate relief.

Education – Post 12th Standard

Problem Statement:

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.

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

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

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

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

2.5 Build the covariance matrix, eigenvalues, and eigenvector.

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

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.

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

Exploratory Data Analysis:

First 10 rows of the data:

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	Abilene Chr	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
1	Adelphi Uni	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
2	Adrian Colle	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
3	Agnes Scot	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
4	Alaska Paci	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15
5	Albertson C	587	479	158	38	62	678	41	13500	3335	500	675	67	73	9.4	11	9727	55
6	Albertus Ma	353	340	103	17	45	416	230	13290	5720	500	1500	90	93	11.5	26	8861	63
7	Albion Colle	1899	1720	489	37	68	1594	32	13868	4826	450	850	89	100	13.7	37	11487	73
8	Albright Col	1038	839	227	30	63	973	306	15595	4400	300	500	79	84	11.3	23	11644	80
9	Alderson-B	582	498	172	21	44	799	78	10468	3380	660	1800	40	41	11.5	15	8991	52

The data set has 777 rows and 18 columns in total. The data types of each of the row/variable is shown below. It can also be noted that there are 777 entries in all the rows and hence there are no blank cells/missing values.

#	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	iint64
14	S.F.Ratio	777 non-null	float64
15	perc.alumni	777 non-null	iint64
16	Expend	777 non-null	int64
17	Grad.Rate	777 non-null	int64

The five number summary for the variables is shown below:

	count	mean	std	min	25%	50%	75%	max
Apps	777	3001.638	3870.201	81	776	1558	3624	48094
Accept	777	2018.804	2451.114	72	604	1110	2424	26330
Enroll	777	779.973	929.1762	35	242	434	902	6392
Top10perc	777	27.55856	17.64036	1	15	23	35	96
Top25perc	777	55.79665	19.80478	9	41	54	69	100
F.Undergrad	777	3699.907	4850.421	139	992	1707	4005	31643
P.Undergrad	777	855.2986	1522.432	1	95	353	967	21836
Outstate	777	10440.67	4023.016	2340	7320	9990	12925	21700
Room.Board	777	4357.526	1096.696	1780	3597	4200	5050	8124
Books	777	549.381	165.1054	96	470	500	600	2340
Personal	777	1340.642	677.0715	250	850	1200	1700	6800
PhD	777	72.66023	16.32816	8	62	75	85	103
Terminal	777	79.7027	14.72236	24	71	82	92	100
S.F.Ratio	777	14.0897	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777	22.74389	12.3918	0	13	21	31	64
Expend	777	9660.171	5221.768	3186	6751	8377	10830	56233
Grad.Rate	777	65.46332	17.17771	10	53	65	78	118

1. There are total of 777 universities (rows) and 18 variables (rows)
2. The columns in the data are 'Names', 'Apps', 'Accept', 'Enroll', 'Top10perc', 'Top25perc', 'F.Undergrad', 'P.Undergrad', 'Outstate', 'Room.Board', 'Books', 'Personal', 'PhD', 'Terminal', 'S.F.Ratio', 'perc.alumni', 'Expend', 'Grad.Rate'.
3. 'Names' is Object data type, 'S.F.Ratio' is Float datatype and the remaining variables are integer type.
4. From number summary:
   - 'PhD', 'Terminal' and 'Grad,Rate' are Left skewed the remaining are Right skewed.
   - For 'PhD' and 'Terminal' variables there is a large number of observations int he 1st quartile. For the other variables there is a large number observations in the 4th quartile. This suggests that there could be outliers and must be handled carefully.

Question 1:

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

Answer:

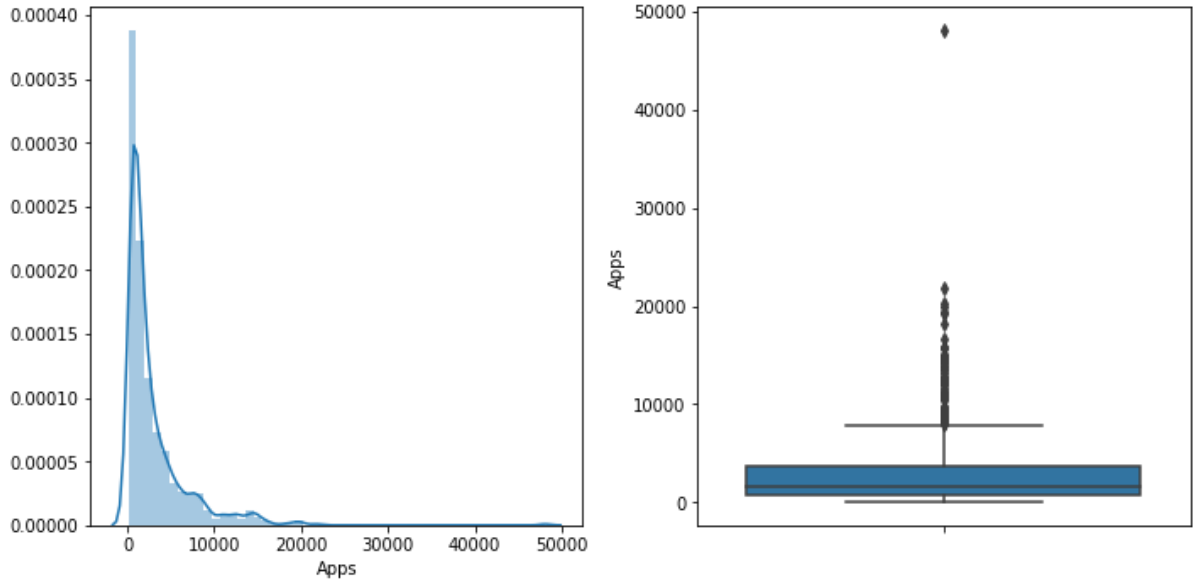
Univariate

The summary of univariate analysis is shown below:

	Number of outliers	% of outliers	Skewness
Apps	70	9.01	Right Skewed
Accept	73	9.4	Right Skewed
Enroll	79	10.17	Right Skewed
Top10perc	39	5.02	Right Skewed
Top25perc	0	0	Right Skewed
F.Undergrad	97	12.48	Right Skewed
P.Undergrad	67	8.62	Right Skewed
Outstate	1	0.13	Right Skewed
Room.Board	7	0.9	Right Skewed
Books	46	5.92	Right Skewed
Personal	20	2.57	Right Skewed
PhD	8	1.03	Left Skewed
Terminal	8	1.03	Left Skewed
S.F.Ratio	12	1.54	Right Skewed
perc.alumni	5	0.64	Right Skewed
Expend	48	6.18	Right Skewed
Grad.Rate	4	0.51	Left Skewed
Total	584	NaN	NaN

Variable wise detailed analysis is shown below:

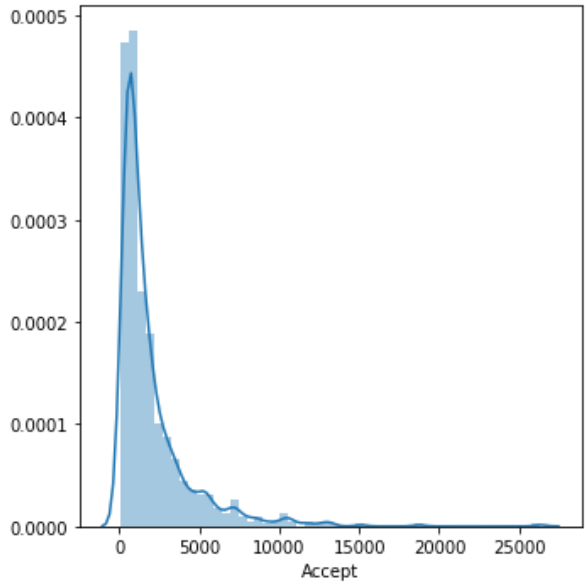
Apps

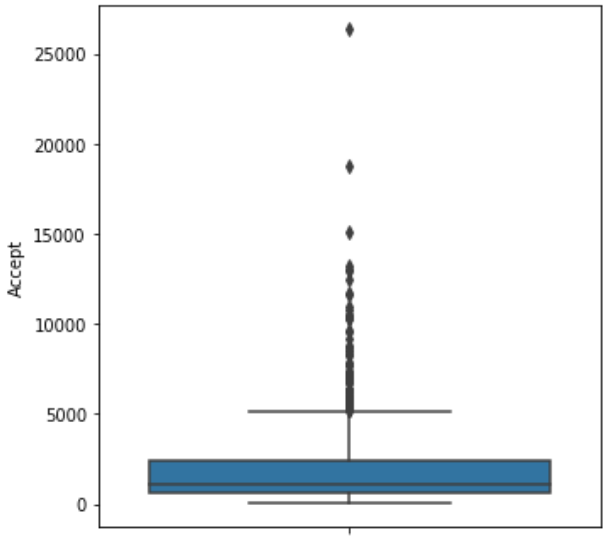


Apps is Positive or Right skewed.

The number of outliers in Apps is 70

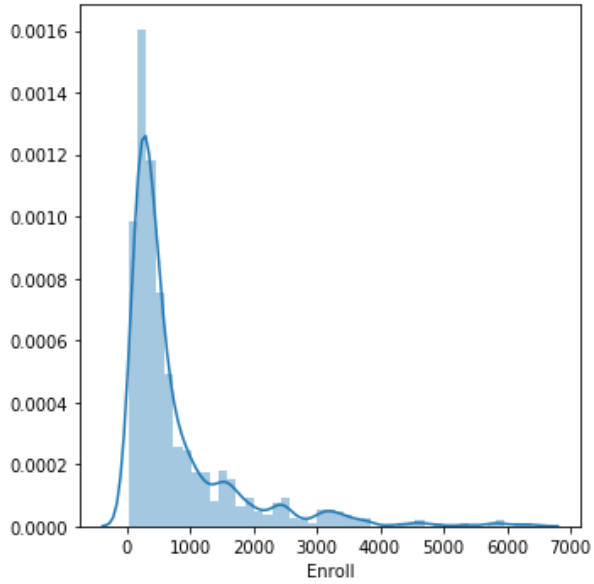
Accept

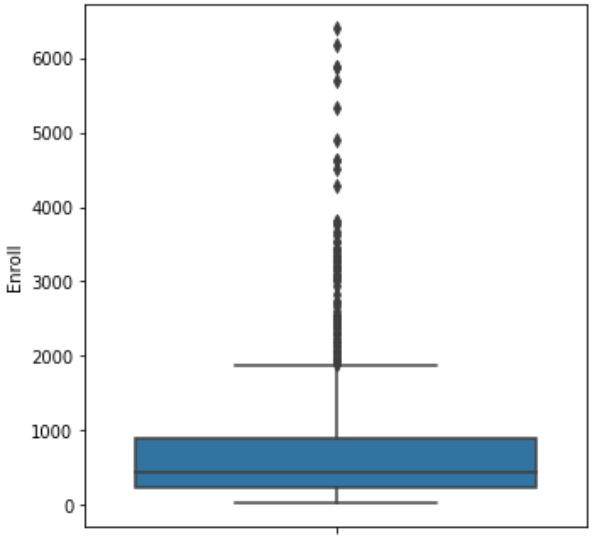




Accept is Positive or Right skewed.
The number of outliers in Accept is 73

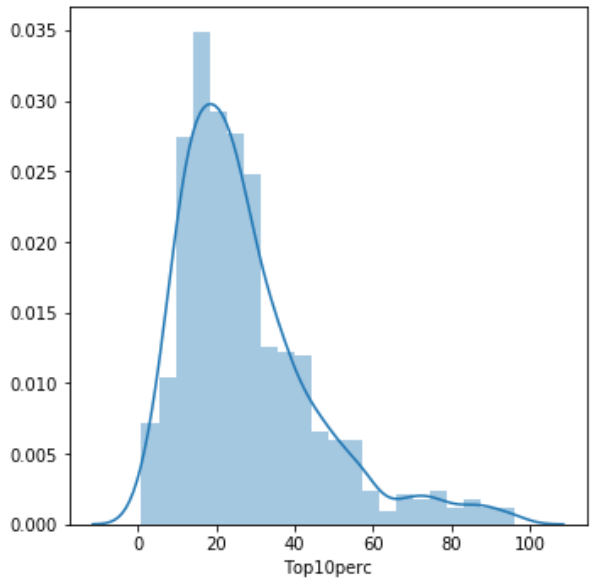
Enroll

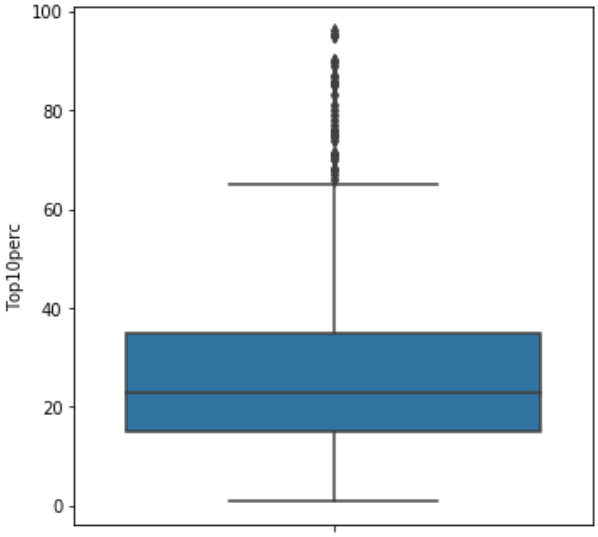




Enroll is Positive or Right skewed.
The number of outliers in Enroll is 79

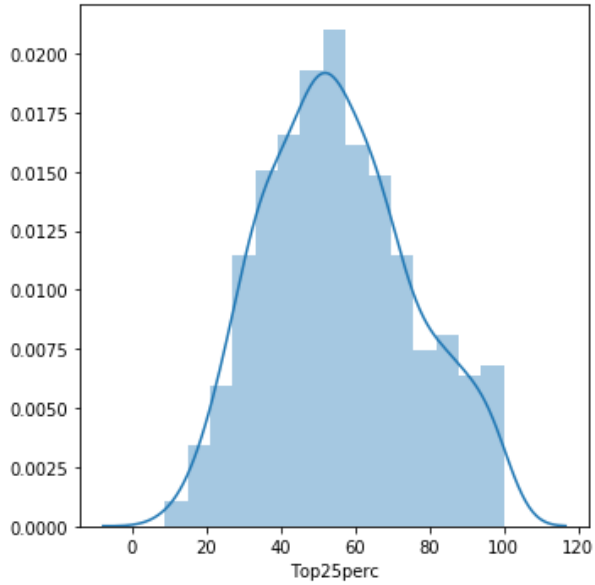
Top10perc

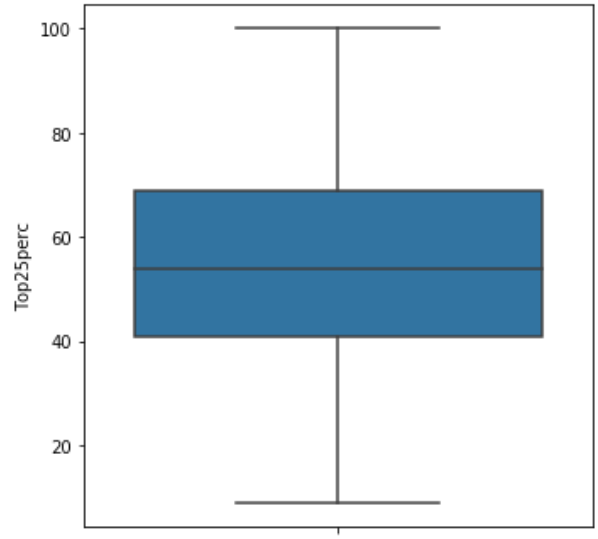




Top10perc is Positive or Right skewed.
The number of outliers in Top10perc is 39

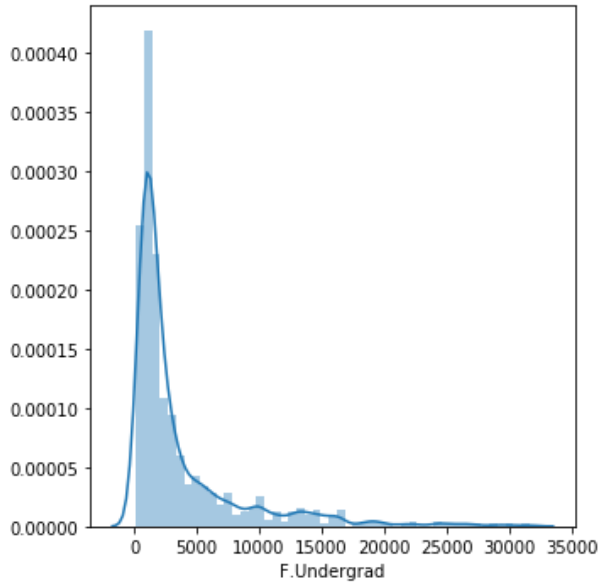
Top25perc

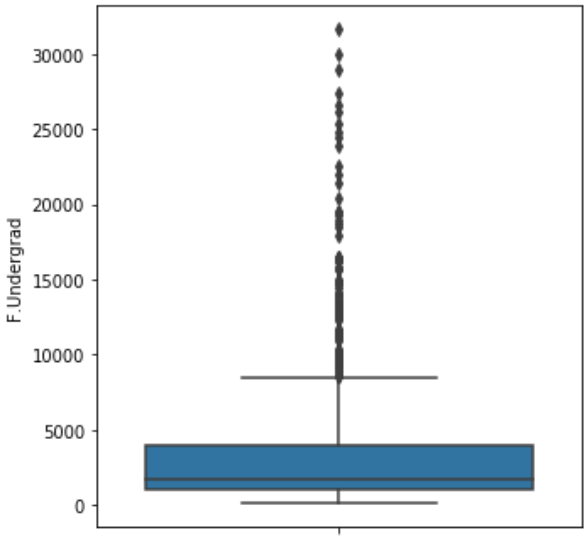




Top25perc is Positive or Right skewed.
The number of outliers in Top25perc is 0

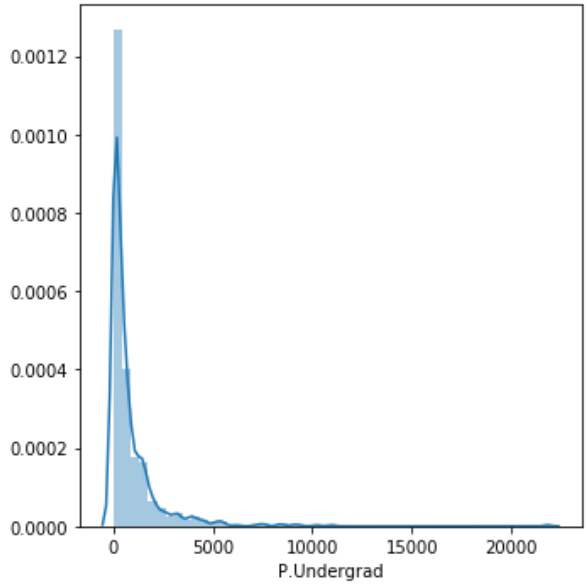
F.Undergrad

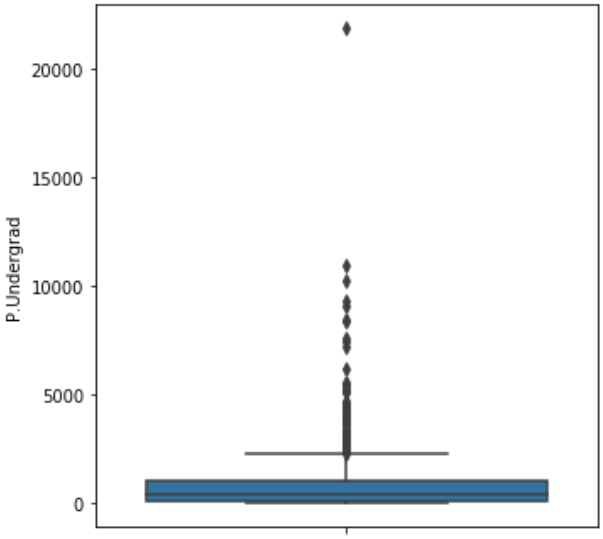




F.Undergrad is Positive or Right skewed.
The number of outliers in F.Undergrad is 97

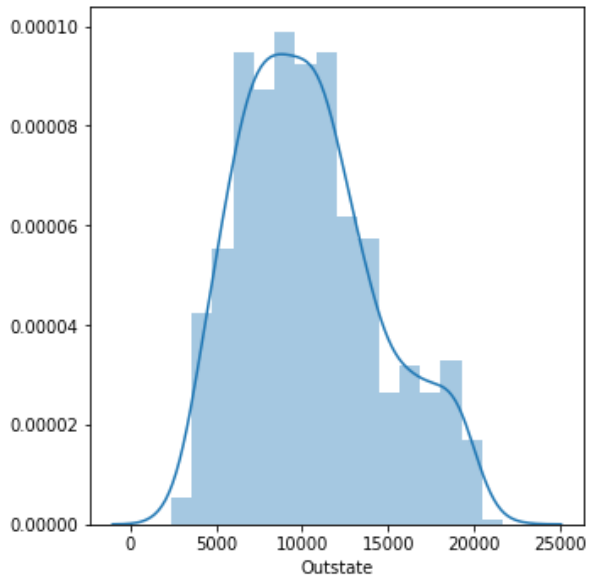
P.Undergrad

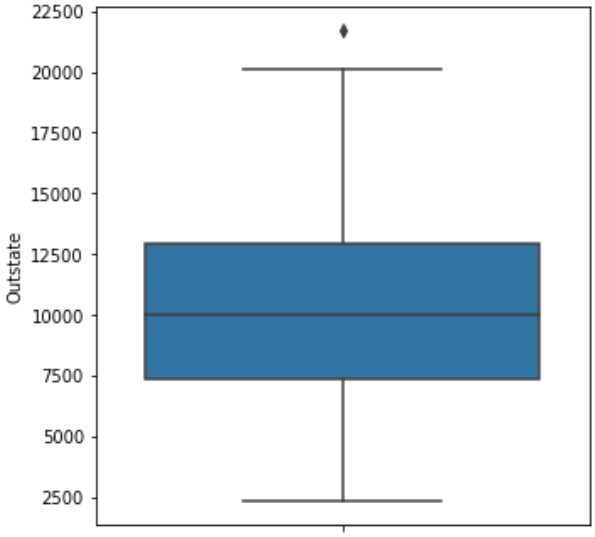




P.Undergrad is Positive or Right skewed.
The number of outliers in P.Undergrad is 67

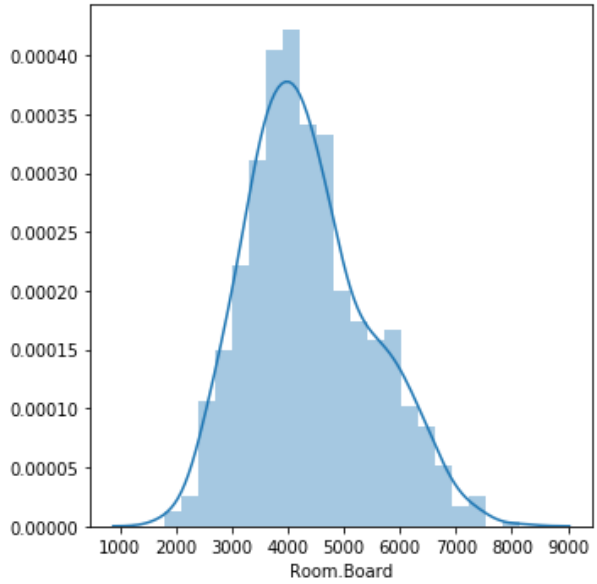
Outstate

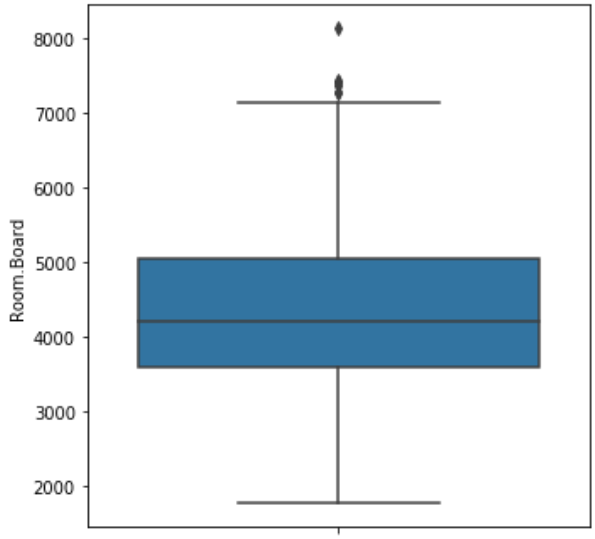




Outstate is Positive or Right skewed.
The number of outliers in Outstate is 1

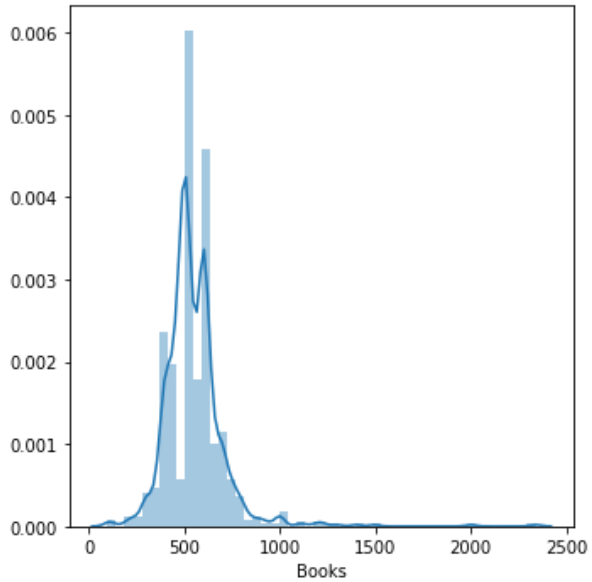
Room.Board

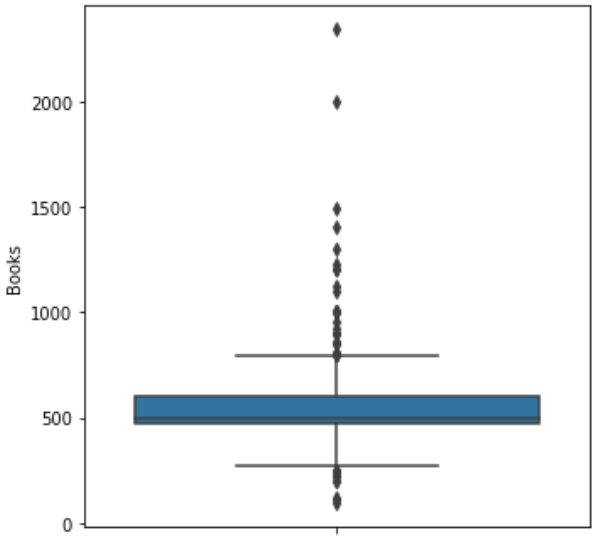




Room.Board is Positive or Right skewed.
The number of outliers in Room.Board is 7

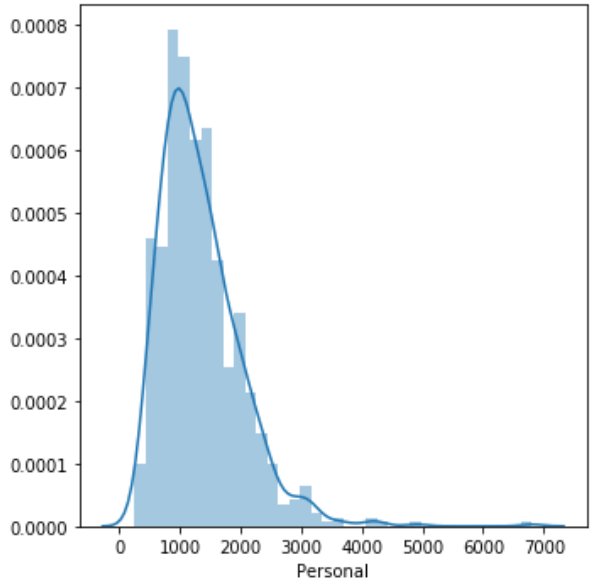
Books

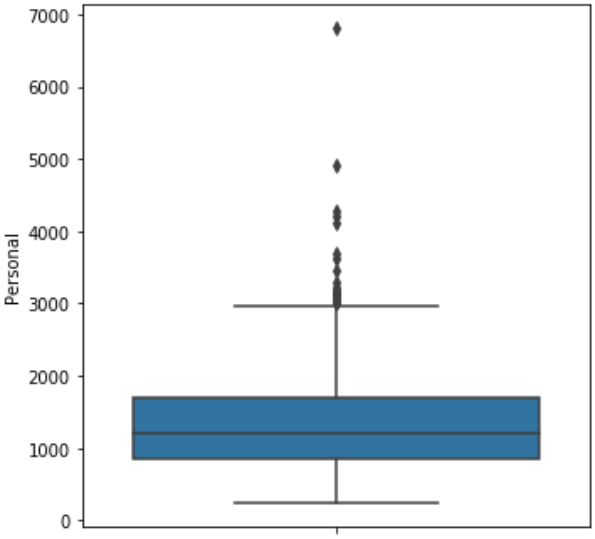




Books is Positive or Right skewed.
The number of outliers in Books is 46

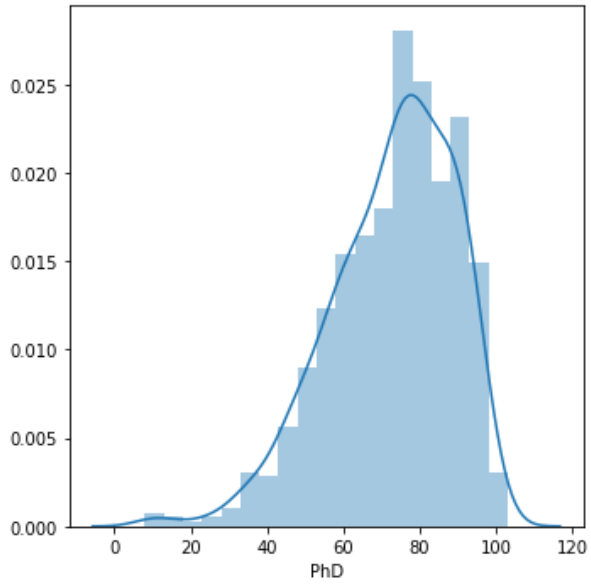
Personal

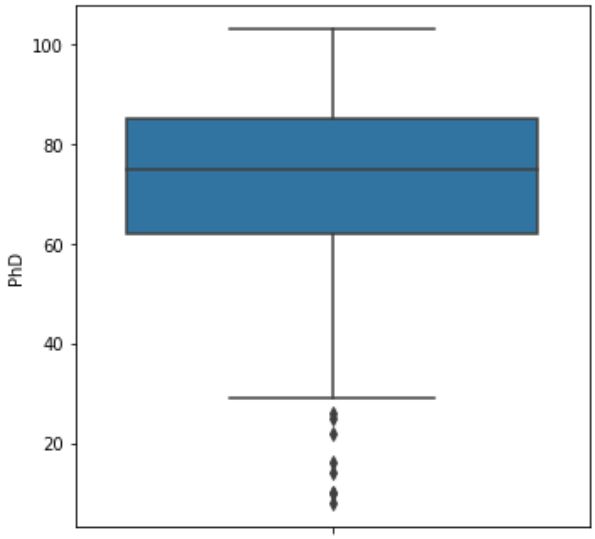




Personal is Positive or Right skewed.
The number of outliers in Personal is 20

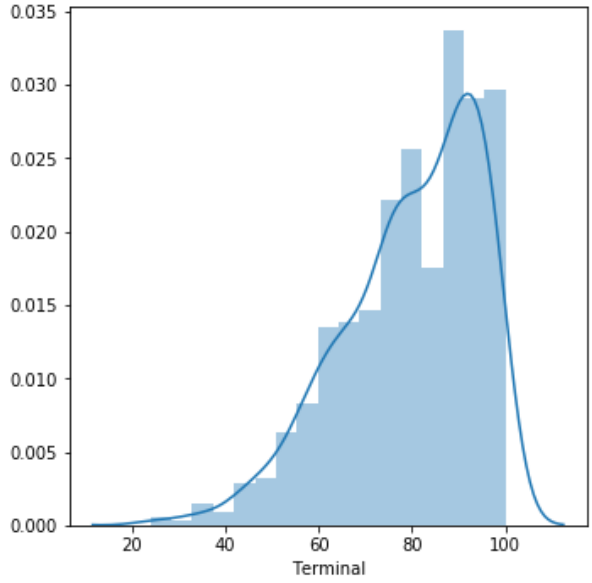
PhD

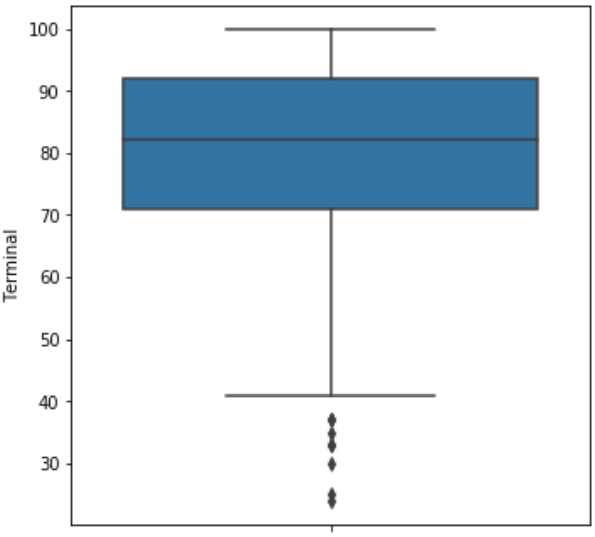




PhD is Negative or Left skewed.
The number of outliers in PhD is 8

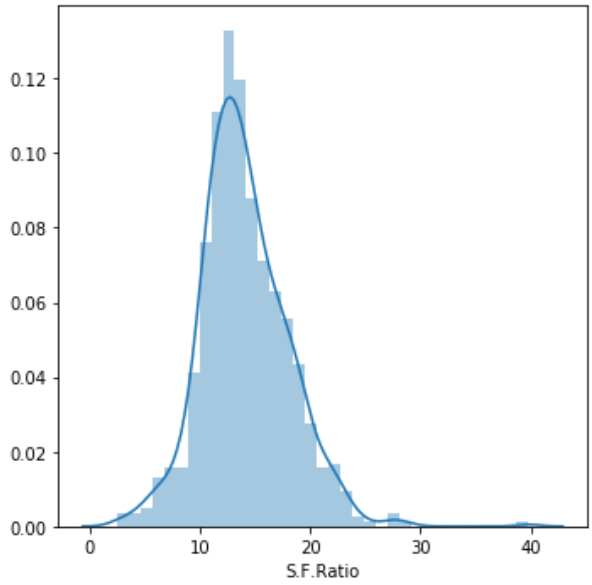
Terminal

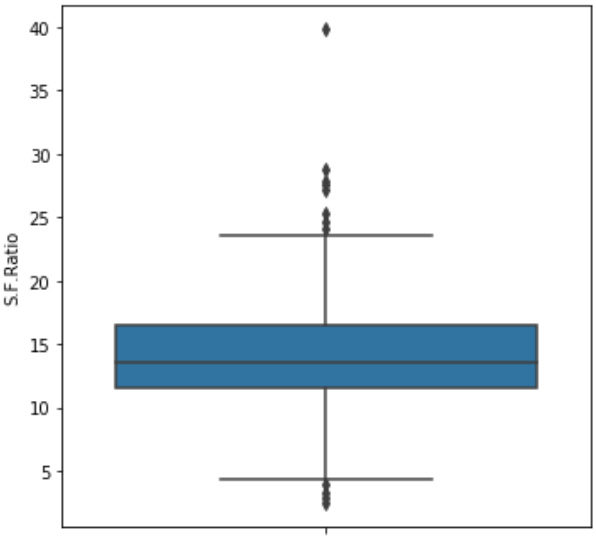




Terminal is Negative or Left skewed.
The number of outliers in Terminal is 8

S.F.Ratio

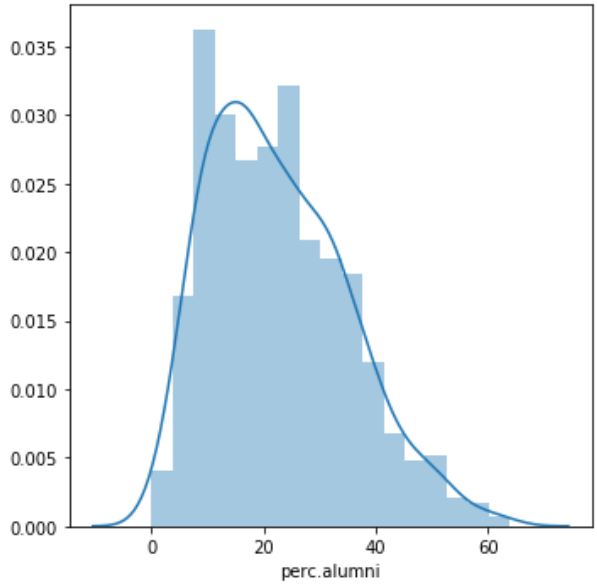


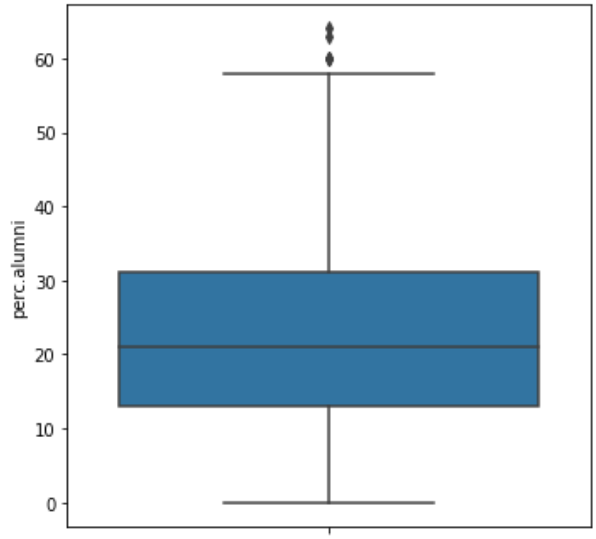


S.F.Ratio is Positive or Right skewed.
The number of outliers in S.F.Ratio is 12

---

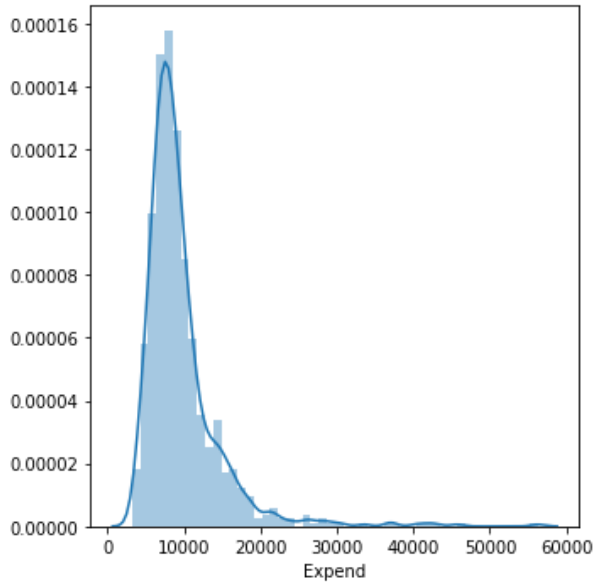
perc.alumni

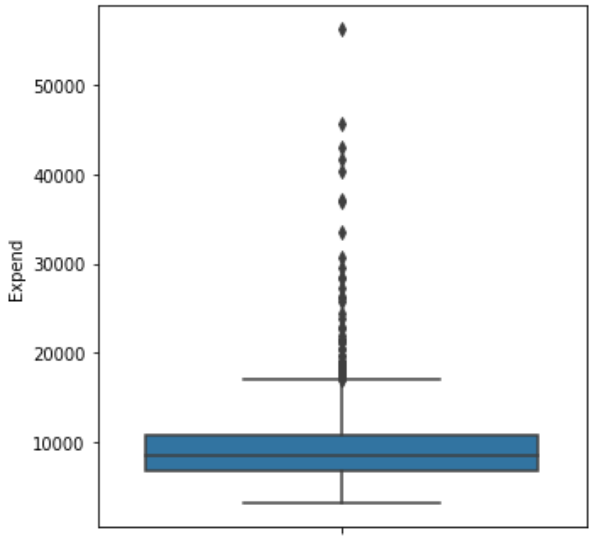




perc.alumni is Positive or Right skewed.
The number of outliers in perc.alumni is 5

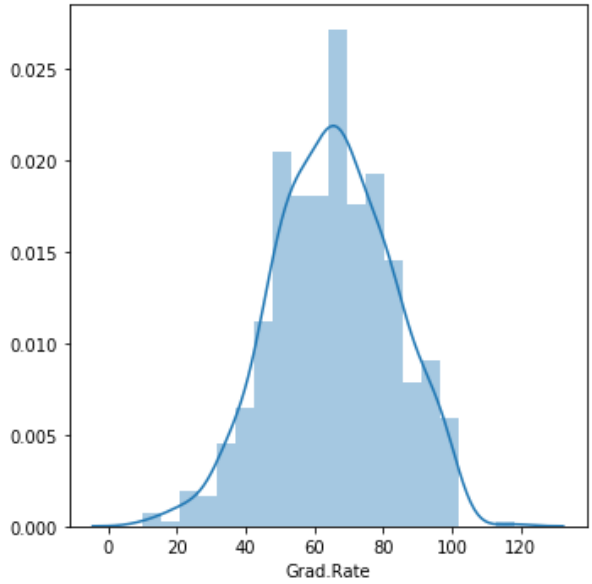
Expend

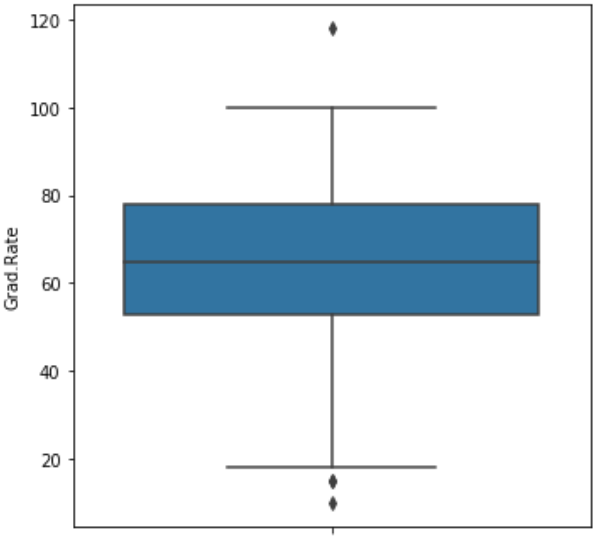




Expend is Positive or Right skewed.
The number of outliers in Expend is 48

Grad.Rate

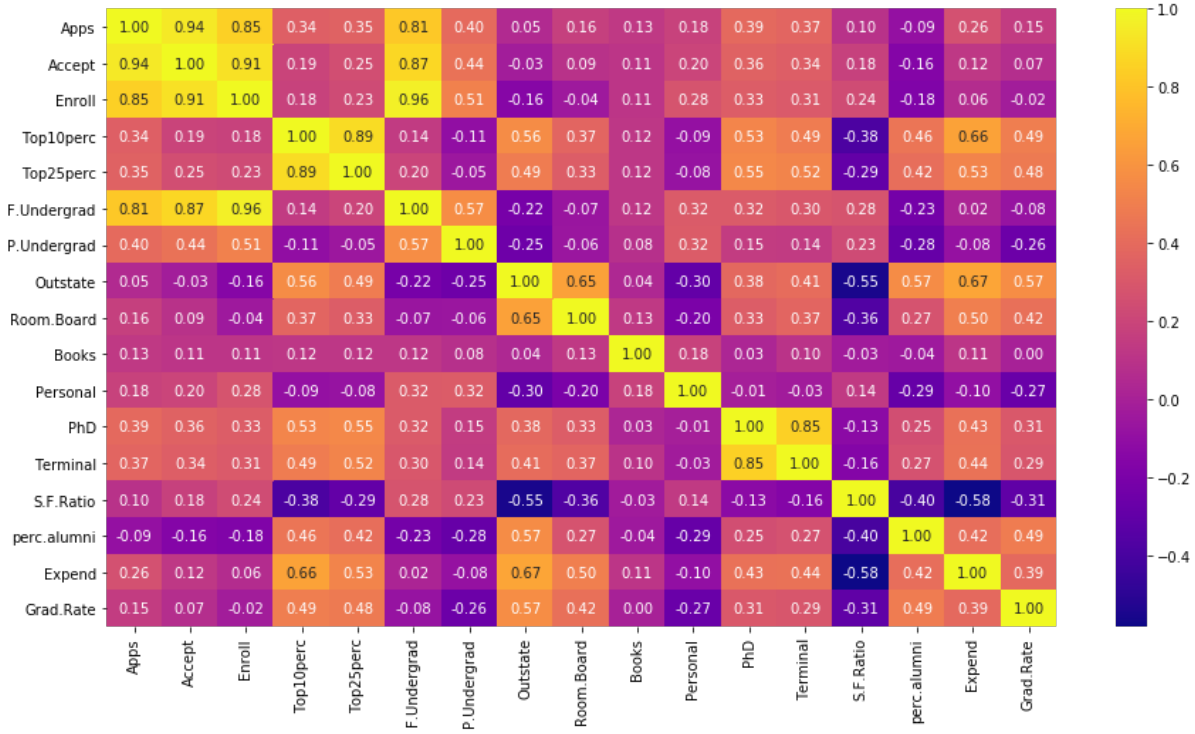




Grad.Rate is Negative or Left skewed.
The number of outliers in Grad.Rate is 4

Multivariate analysis:

The correlation heatmap is shown below:



Findings:

1. There are total of 777 universities (rows) and 18 variables (rows)
2. The columns in the data are 'Names', 'Apps', 'Accept', 'Enroll', 'Top10perc', 'Top25perc', 'F.Undergrad', 'P.Undergrad', 'Outstate', 'Room.Board', 'Books', 'Personal', 'PhD', 'Terminal', 'S.F.Ratio', 'perc.alumni', 'Expend', 'Grad.Rate'.
3. 'Names' is Object data type, 'S.F.Ratio' is Float datatype and the remaining variables are integer type.
4. From number summary:
   - 'PhD', 'Terminal' and ‘Grad,Rate’ are Left skewed the remaining are Right skewed.
   - For 'PhD' and 'Terminal' variables there is a large number of observations int he 1st quartile. For the other variables there is a large number observations in the 4th quartile. This suggests that there could be outliers and must be handled carefully.
5. Form the 5 number summary we can observe that most of the variables are right skewed. Also the range of each quartile is unequal which also suggests high number of outliers.
6. A brief summary of univariate analysis is shown below:

	Number of outliers	% of outliers	Skewness
Apps	70	9.01	Right Skewed
Accept	73	9.4	Right Skewed
Enroll	79	10.17	Right Skewed
Top10perc	39	5.02	Right Skewed
Top25perc	0	0	Right Skewed
F.Undergrad	97	12.48	Right Skewed
P.Undergrad	67	8.62	Right Skewed
Outstate	1	0.13	Right Skewed
Room.Board	7	0.9	Right Skewed
Books	46	5.92	Right Skewed
Personal	20	2.57	Right Skewed
PhD	8	1.03	Left Skewed
Terminal	8	1.03	Left Skewed
S.F.Ratio	12	1.54	Right Skewed
perc.alumni	5	0.64	Right Skewed
Expend	48	6.18	Right Skewed
Grad.Rate	4	0.51	Left Skewed
Total	584	NaN	NaN

7. Some of the pairs that have high positive correlations are:
   - F.Undergrad and Enroll
   - Top10perc and Top25perc
   - Terminal and PhD
   - Top10perc and Expend
8. Some of the pairs that have high negative correlations are:
   - S.F.Ratio and Expend
   - S.F.Ratio and Outstate

Question 2:

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

Answer:

Standard scalar will only standardize the values into a z-score but the number of outliers and the skewness will almost remain the same.
We have a total of 17 variables out of which 14 are right skewed. This means we have a few observations that have very high values. To pull these higher values towards the middle of the graph we will do a logarithmic transformation.

Doing so may reduce the outliers for the 14 variables that are right skewed. However, for the 3 variables that are already left skewed, there may be an increased number of outliers and hence needs special attention.

On the whole, when we consider the entire data set, this kind of scaling will most likely reduce the outliers considerably.

Question 3:

Comment on the comparison between covariance and the correlation matrix.

Answer:

The covariance matrix of the raw data is:

	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	1.50E+07	8.95E+06	3.05E+06	23132.77314	26952.66348	1.53E+07	2.35E+06	7.81E+05	7.00E+05	84703.75	4.68E+05	24689.43	21053.068	1465.0606	-4327.122381	5.25E+06	9756.42164
Accept	8.95E+06	6.01E+06	2.08E+06	8321.124872	12013.40476	1.04E+07	1.65E+06	-2.54E+05	2.44E+05	45942.81	3.34E+05	14238.2	12182.094	1709.8382	-4859.487022	1.60E+06	2834.16292
Enroll	3.05E+06	2.08E+06	8.63E+05	2971.583415	4172.592435	4.35E+06	7.26E+05	-5.81E+05	-4.10E+04	17291.2	1.77E+05	5028.961	4217.086	872.68477	-2081.693787	3.11E+05	-356.58798
Top10perc	2.31E+04	8.32E+03	2.97E+03	311.182456	311.63048	1.21E+04	-2.83E+03	3.99E+04	7.19E+03	346.1774	-1.11E+03	153.1849	127.55158	-26.87453	99.567208	6.09E+04	149.992164
Top25perc	2.70E+04	1.20E+04	4.17E+03	311.63048	392.229216	1.92E+04	-1.62E+03	3.90E+04	7.20E+03	377.7593	-1.08E+03	176.5184	153.00261	-23.0972	102.550946	5.45E+04	162.371398
F.Undergra	1.53E+07	1.04E+07	4.35E+06	12089.11368	19158.95278	2.35E+07	4.21E+06	-4.21E+06	-3.66E+05	92535.76	1.04E+06	25211.78	21424.242	5370.2086	-13791.92969	4.72E+05	-6563.3075
P.Undergra	2.35E+06	1.65E+06	7.26E+05	-2829.47498	-1615.41214	4.21E+06	2.32E+06	-1.55E+06	-1.02E+05	20410.45	3.30E+05	3706.756	3180.5966	1401.3026	-5297.33709	-6.64E+05	-6721.0625
Outstate	7.81E+05	-2.54E+05	-5.81E+05	39907.17983	38992.4275	-4.21E+06	-1.55E+06	1.62E+07	2.89E+06	25808.24	-8.15E+05	25157.52	24164.148	-8835.254	28229.55307	1.41E+07	39479.6818
Room.Boar	7.00E+05	2.44E+05	-4.10E+04	7186.705605	7199.903568	-3.66E+05	-1.02E+05	2.89E+06	1.20E+06	23170.31	-1.48E+05	5895.035	6047.2997	-1574.206	3701.431379	2.87E+06	8005.36018
Books	8.47E+04	4.59E+04	1.73E+04	346.177405	377.759266	9.25E+04	2.04E+04	2.58E+04	2.32E+04	27259.78	2.00E+04	72.53424	242.96392	-20.86721	-82.263132	9.69E+04	3.008837
Personal	4.68E+05	3.34E+05	1.77E+05	-1114.55119	-1083.60507	1.04E+06	3.30E+05	-8.15E+05	-1.48E+05	20043.03	4.58E+05	-120.899	-305.1542	365.41577	-2399.310824	-3.46E+05	-3132.6149
PhD	2.47E+04	1.42E+04	5.03E+03	153.18487	176.518449	2.52E+04	3.71E+03	2.52E+04	5.90E+03	72.53424	-1.21E+02	266.6086	204.23133	-8.436492	50.38323	3.69E+04	85.557109
Terminal	2.11E+04	1.22E+04	4.22E+03	127.551581	153.002612	2.14E+04	3.18E+03	2.42E+04	6.05E+03	242.9639	-3.05E+02	204.2313	216.74784	-9.330256	48.734327	3.37E+04	73.220396
S.F.Ratio	1.47E+03	1.71E+03	8.73E+02	-26.874525	-23.097199	5.37E+03	1.40E+03	-8.84E+03	-1.57E+03	-20.8672	3.65E+02	-8.43649	-9.330256	15.668528	-19.764109	-1.21E+04	-20.854888
perc.alumn	-4.33E+03	-4.86E+03	-2.08E+03	99.567208	102.550946	-1.38E+04	-5.30E+03	2.82E+04	3.70E+03	-82.2631	-2.40E+03	50.38323	48.734327	-19.76411	153.556744	2.70E+04	104.493815
Expend	5.25E+06	1.60E+06	3.11E+05	60879.3102	54546.48331	4.72E+05	-6.64E+05	1.41E+07	2.87E+06	96912.58	-3.46E+05	36898.06	33733.457	-12067.56	27028.92147	2.73E+07	35012.9683
Grad.Rate	9.76E+03	2.83E+03	-3.57E+02	149.992164	162.371398	-6.56E+03	-6.72E+03	3.95E+04	8.01E+03	3.008837	-3.13E+03	85.55711	73.220396	-20.85489	104.493815	3.50E+04	295.073717

The correlation matrix of the raw data is:

	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	1.00E+00	9.43E-01	8.47E-01	0.338834	0.35164	8.14E-01	3.98E-01	5.02E-02	1.65E-01	0.132559	1.79E-01	0.390697	0.369491	0.095633	-0.090226	2.60E-01	0.146755
Accept	9.43E-01	1.00E+00	9.12E-01	0.192447	0.247476	8.74E-01	4.41E-01	-2.58E-02	9.09E-02	0.113525	2.01E-01	0.355758	0.337583	0.176229	-0.15999	1.25E-01	0.067313
Enroll	8.47E-01	9.12E-01	1.00E+00	0.181294	0.226745	9.65E-01	5.13E-01	-1.55E-01	-4.02E-02	0.112711	2.81E-01	0.331469	0.308274	0.237271	-0.180794	6.42E-02	-0.022341
Top10perc	3.39E-01	1.92E-01	1.81E-01	1	0.891995	1.41E-01	-1.05E-01	5.62E-01	3.71E-01	0.118858	-9.33E-02	0.531828	0.491135	-0.384875	0.455485	6.61E-01	0.494989
Top25perc	3.52E-01	2.47E-01	2.27E-01	0.891995	1	1.99E-01	-5.36E-02	4.89E-01	3.31E-01	0.115527	-8.08E-02	0.545862	0.524749	-0.294629	0.417864	5.27E-01	0.477281
F.Undergra	8.14E-01	8.74E-01	9.65E-01	0.141289	0.199445	1.00E+00	5.71E-01	-2.16E-01	-6.89E-02	0.11555	3.17E-01	0.318337	0.300019	0.279703	-0.229462	1.87E-02	-0.078773
P.Undergra	3.98E-01	4.41E-01	5.13E-01	-0.105356	-0.053577	5.71E-01	1.00E+00	-2.54E-01	-6.13E-02	0.0812	3.20E-01	0.149114	0.141904	0.232531	-0.280792	-8.36E-02	-0.257001
Outstate	5.02E-02	-2.58E-02	-1.55E-01	0.562331	0.489394	-2.16E-01	-2.54E-01	1.00E+00	6.54E-01	0.038855	-2.99E-01	0.382982	0.407983	-0.554821	0.566262	6.73E-01	0.57129
Room.Boar	1.65E-01	9.09E-02	-4.02E-02	0.37148	0.33149	-6.89E-02	-6.13E-02	6.54E-01	1.00E+00	0.127963	-1.99E-01	0.329202	0.37454	-0.362628	0.272363	5.02E-01	0.424942
Books	1.33E-01	1.14E-01	1.13E-01	0.118858	0.115527	1.16E-01	8.12E-02	3.89E-02	1.28E-01	1	1.79E-01	0.026906	0.099955	-0.031929	-0.040208	1.12E-01	0.001061
Personal	1.79E-01	2.01E-01	2.81E-01	-0.093316	-0.08081	3.17E-01	3.20E-01	-2.99E-01	-1.99E-01	0.179295	1.00E+00	-0.01094	-0.030613	0.136345	-0.285968	-9.79E-02	-0.269344
PhD	3.91E-01	3.56E-01	3.31E-01	0.531828	0.545862	3.18E-01	1.49E-01	3.83E-01	3.29E-01	0.026906	-1.09E-02	1	0.849587	-0.13053	0.249009	4.33E-01	0.305038
Terminal	3.69E-01	3.38E-01	3.08E-01	0.491135	0.524749	3.00E-01	1.42E-01	4.08E-01	3.75E-01	0.099955	-3.06E-02	0.849587	1	-0.160104	0.26713	4.39E-01	0.289527
S.F.Ratio	9.56E-02	1.76E-01	2.37E-01	-0.384875	-0.294629	2.80E-01	2.33E-01	-5.55E-01	-3.63E-01	-0.03193	1.36E-01	-0.13053	-0.160104	1	-0.402929	-5.84E-01	-0.30671
perc.alumn	-9.02E-02	-1.60E-01	-1.81E-01	0.455485	0.417864	-2.29E-01	-2.81E-01	5.66E-01	2.72E-01	-0.04021	-2.86E-01	0.249009	0.26713	-0.402929	1	4.18E-01	0.490898
Expend	2.60E-01	1.25E-01	6.42E-02	0.660913	0.527447	1.87E-02	-8.36E-02	6.73E-01	5.02E-01	0.112409	-9.79E-02	0.432762	0.438799	-0.583832	0.417712	1.00E+00	0.390343
Grad.Rate	1.47E-01	6.73E-02	-2.23E-02	0.494989	0.477281	-7.88E-02	-2.57E-01	5.71E-01	4.25E-01	0.001061	-2.69E-01	0.305038	0.289527	-0.30671	0.490898	3.90E-01	1

When we compute the z-score standardized form of the raw data and then calculate the covariation, the obtained matrix is shown below:

	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	1.00E+00	9.45E-01	8.48E-01	0.33927	0.352093	8.16E-01	3.99E-01	5.02E-02	1.65E-01	0.132729	1.79E-01	0.391201	0.369968	0.095756	-0.090342	2.60E-01	0.146944
Accept	9.45E-01	1.00E+00	9.13E-01	0.192695	0.247795	8.75E-01	4.42E-01	-2.58E-02	9.10E-02	0.113672	2.01E-01	0.356216	0.338018	0.176456	-0.160196	1.25E-01	0.067399
Enroll	8.48E-01	9.13E-01	1.00E+00	0.181527	0.227037	9.66E-01	5.14E-01	-1.56E-01	-4.03E-02	0.112856	2.81E-01	0.331896	0.308671	0.237577	-0.181027	6.43E-02	-0.02237
Top10perc	3.39E-01	1.93E-01	1.82E-01	1.001289	0.893144	1.41E-01	-1.05E-01	5.63E-01	3.72E-01	0.119012	-9.34E-02	0.532513	0.491768	-0.38537	0.456072	6.62E-01	0.495627
Top25perc	3.52E-01	2.48E-01	2.27E-01	0.893144	1.001289	2.00E-01	-5.36E-02	4.90E-01	3.32E-01	0.115676	-8.09E-02	0.546566	0.525425	-0.295009	0.418403	5.28E-01	0.477896
F.Undergra	8.16E-01	8.75E-01	9.66E-01	0.141471	0.199702	1.00E+00	5.71E-01	-2.16E-01	-6.90E-02	0.115699	3.18E-01	0.318747	0.300406	0.280064	-0.229758	1.87E-02	-0.078875
P.Undergra	3.99E-01	4.42E-01	5.14E-01	-0.105492	-0.053646	5.71E-01	1.00E+00	-2.54E-01	-6.14E-02	0.081304	3.20E-01	0.149306	0.142086	0.23283	-0.281154	-8.37E-02	-0.257332
Outstate	5.02E-02	-2.58E-02	-1.56E-01	0.563055	0.490024	-2.16E-01	-2.54E-01	1.00E+00	6.55E-01	0.038905	-2.99E-01	0.383476	0.408509	-0.555536	0.566992	6.74E-01	0.572026
Room.Boar	1.65E-01	9.10E-02	-4.03E-02	0.371959	0.331917	-6.90E-02	-6.14E-02	6.55E-01	1.00E+00	0.128128	-2.00E-01	0.329627	0.375022	-0.363095	0.272714	5.02E-01	0.425489
Books	1.33E-01	1.14E-01	1.13E-01	0.119012	0.115676	1.16E-01	8.13E-02	3.89E-02	1.28E-01	1.001289	1.80E-01	0.02694	0.100084	-0.03197	-0.04026	1.13E-01	0.001062
Personal	1.79E-01	2.01E-01	2.81E-01	-0.093437	-0.080914	3.18E-01	3.20E-01	-2.99E-01	-2.00E-01	0.179526	1.00E+00	-0.01095	-0.030653	0.136521	-0.286337	-9.80E-02	-0.269691
PhD	3.91E-01	3.56E-01	3.32E-01	0.532513	0.546566	3.19E-01	1.49E-01	3.83E-01	3.30E-01	0.02694	-1.10E-02	1.001289	0.850682	-0.130698	0.24933	4.33E-01	0.305431
Terminal	3.70E-01	3.38E-01	3.09E-01	0.491768	0.525425	3.00E-01	1.42E-01	4.09E-01	3.75E-01	0.100084	-3.07E-02	0.850682	1.001289	-0.16031	0.267475	4.39E-01	0.2899
S.F.Ratio	9.58E-02	1.76E-01	2.38E-01	-0.38537	-0.295009	2.80E-01	2.33E-01	-5.56E-01	-3.63E-01	-0.03197	1.37E-01	-0.1307	-0.16031	1.001289	-0.403448	-5.85E-01	-0.307106
perc.alumn	-9.03E-02	-1.60E-01	-1.81E-01	0.456072	0.418403	-2.30E-01	-2.81E-01	5.67E-01	2.73E-01	-0.04026	-2.86E-01	0.24933	0.267475	-0.403448	1.001289	4.18E-01	0.49153
Expend	2.60E-01	1.25E-01	6.43E-02	0.661765	0.528127	1.87E-02	-8.37E-02	6.74E-01	5.02E-01	0.112554	-9.80E-02	0.433319	0.439365	-0.584584	0.41825	1.00E+00	0.390846
Grad.Rate	1.47E-01	6.74E-02	-2.24E-02	0.495627	0.477896	-7.89E-02	-2.57E-01	5.72E-01	4.25E-01	0.001062	-2.70E-01	0.305431	0.2899	-0.307106	0.49153	3.91E-01	1.001289

The comparison between covariance and correlation is:

- Covariance is the numerical covariance between 2 variables and the diagonal elements of the covariance matrix is the variance of the variable. This is unite dependent.
- Correlation is nothing but covariance but is unit independent. For this the covariances are divided by their respective standard deviations.
- The diagonal elements of the correlation is 1.
- There is no min or max as the variance/covariance will vary based on the values
- The correlation has a max of 1 and a min if -1. It is said that the 2 variables are positively or negatively correlated respectively.

Example:

If we compared heights with weights. The weights are recorded in kilograms and grams. Since gram is a smaller unit and numerically bigger, the covariance of height and kilograms will be smaller than height and grams.

However this problem can be solved by measuring in correlation. In correlation, we will get the same values for the two pairs because correlation is unit independent. While dividing by the standard deviation the units will get normalized.

Note:
If we converted the data to z-scale and then found covariance we will get the data's correlation.

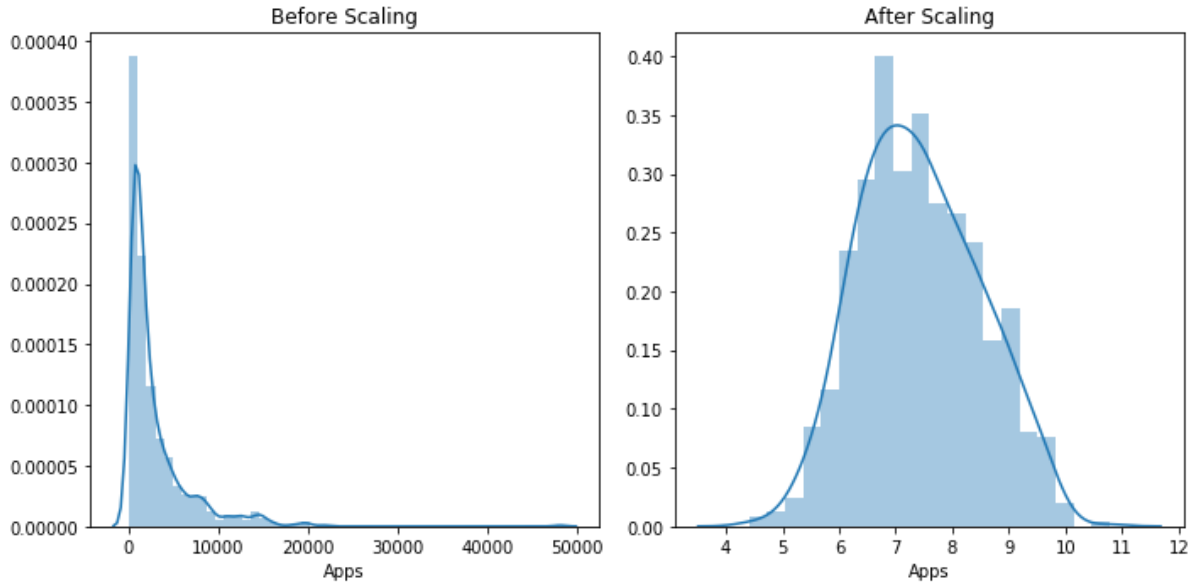
Question 4:

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

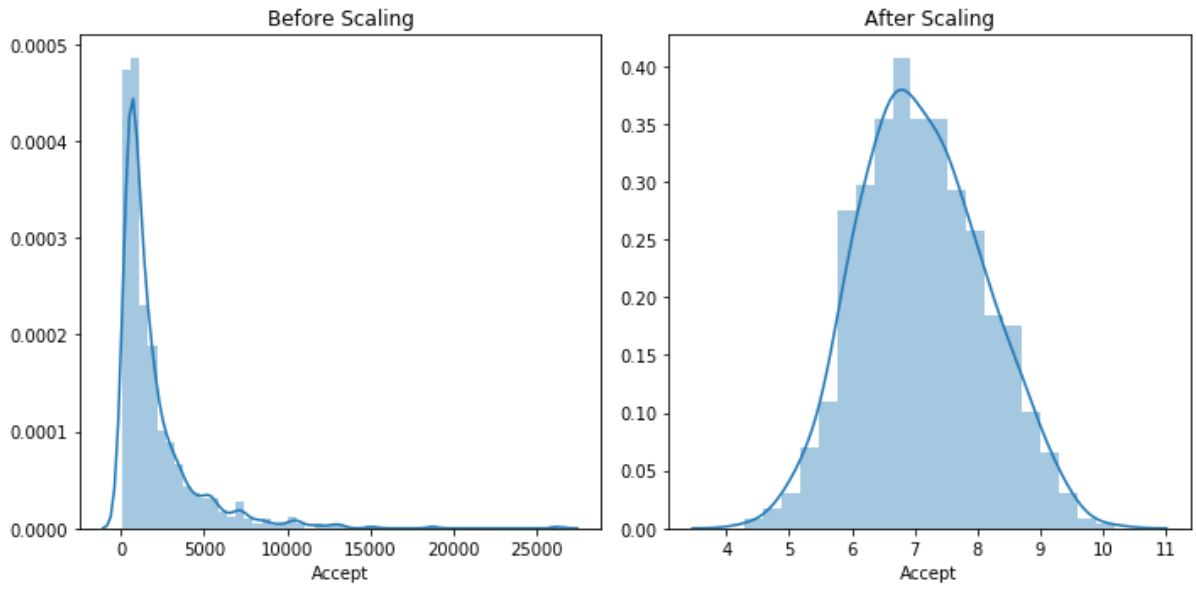
Answer:

The comparison of distribution before and after scaling is shown below:

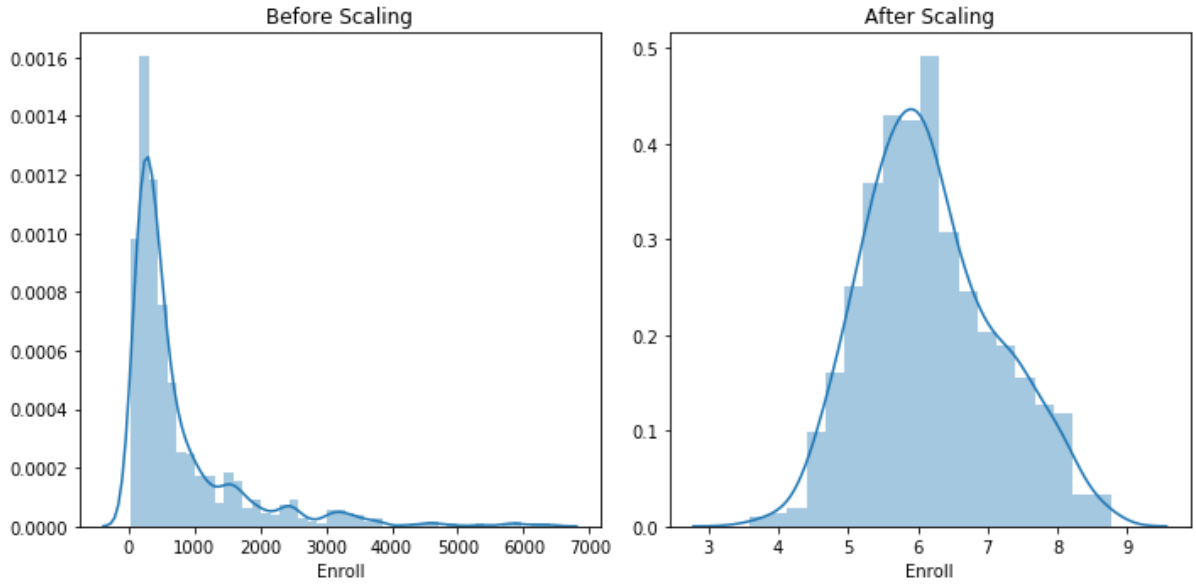
Apps



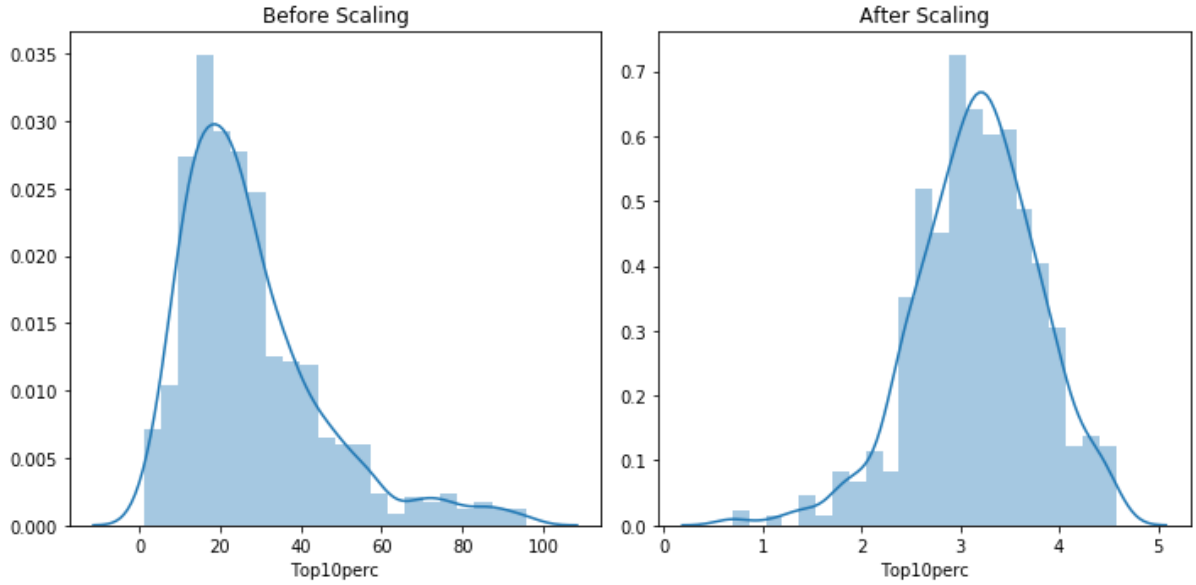
Accept



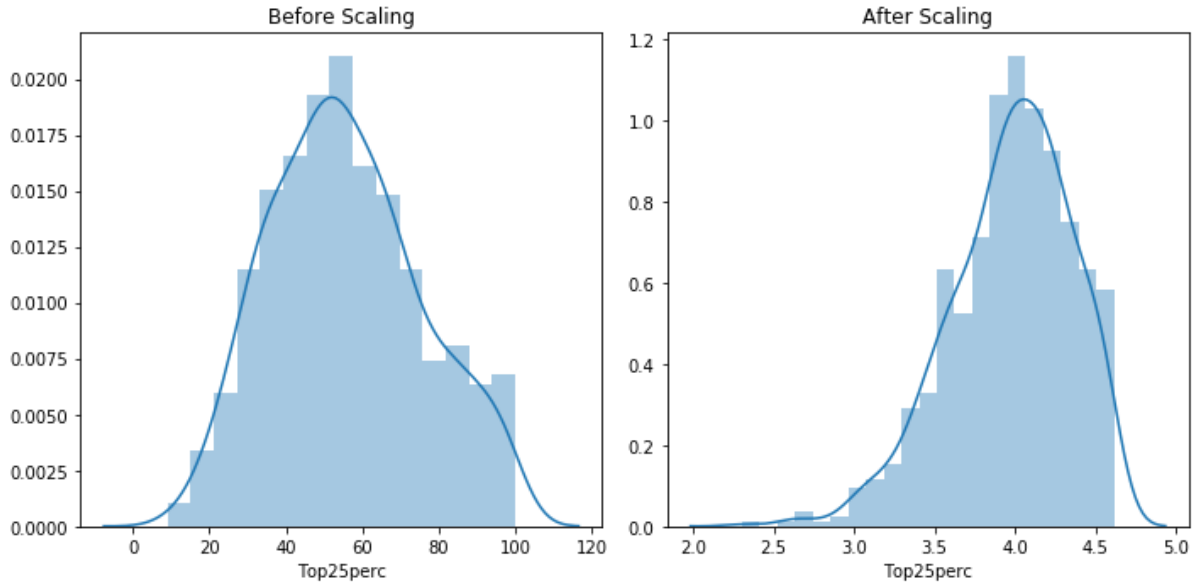
Enroll



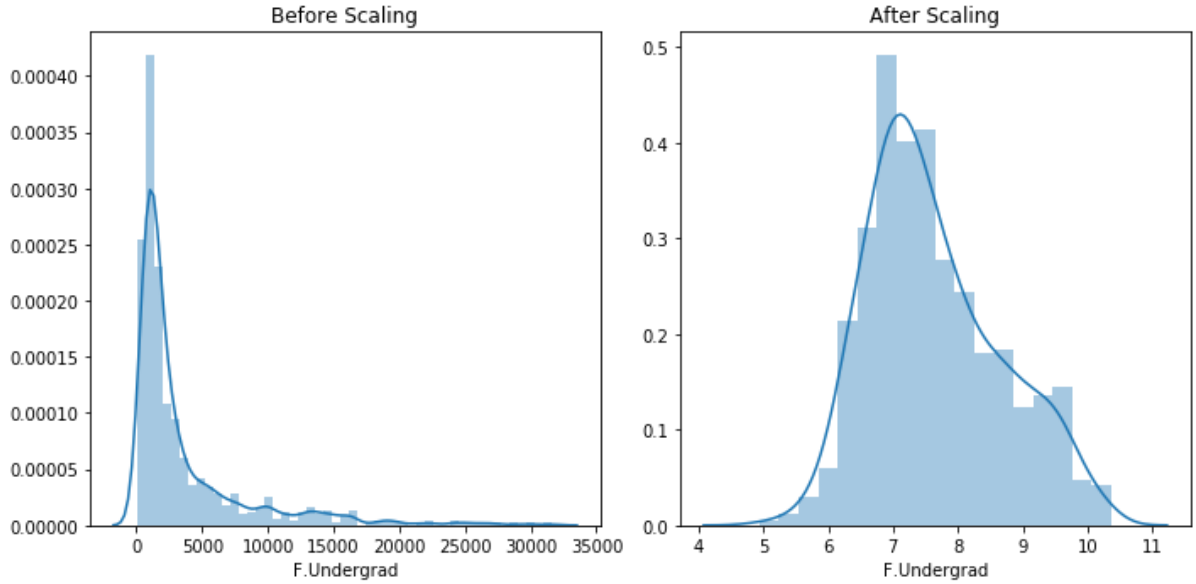
Top10perc



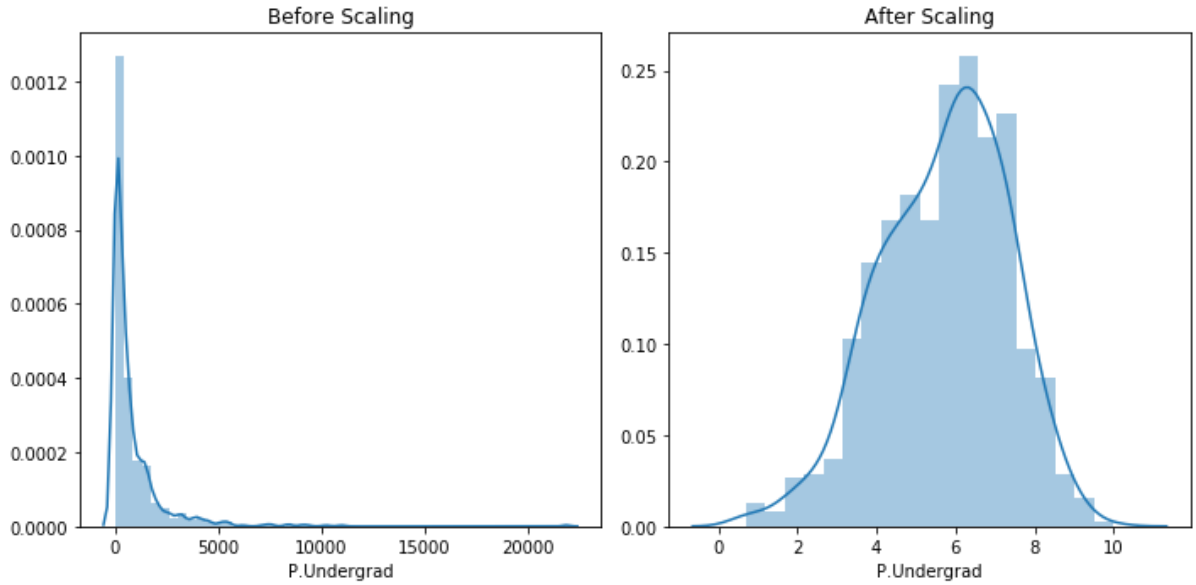
Top25perc



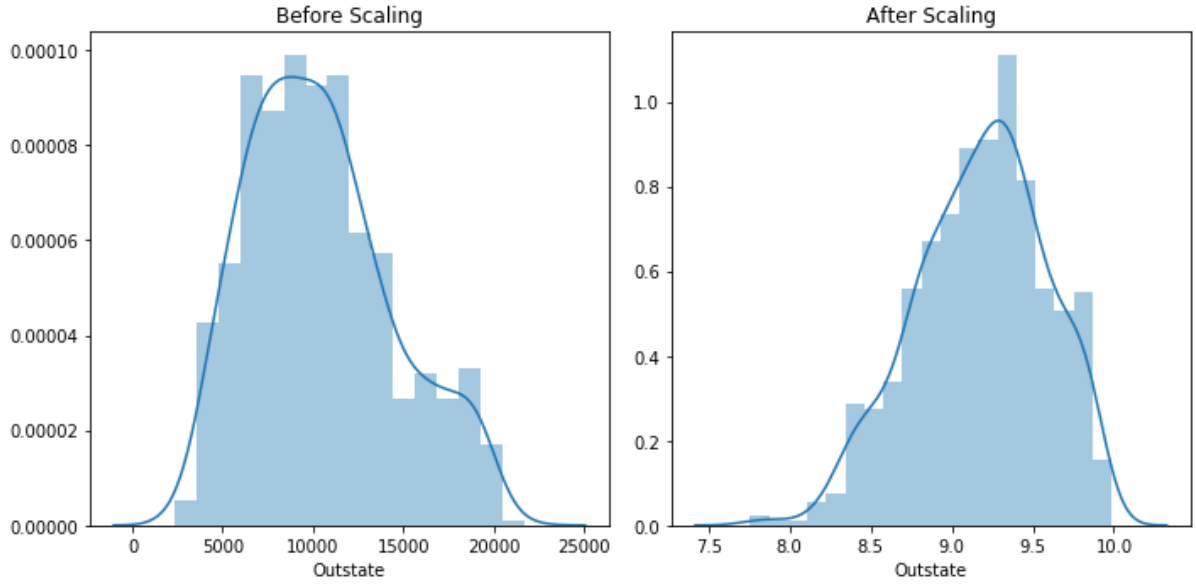
F.Undergrad



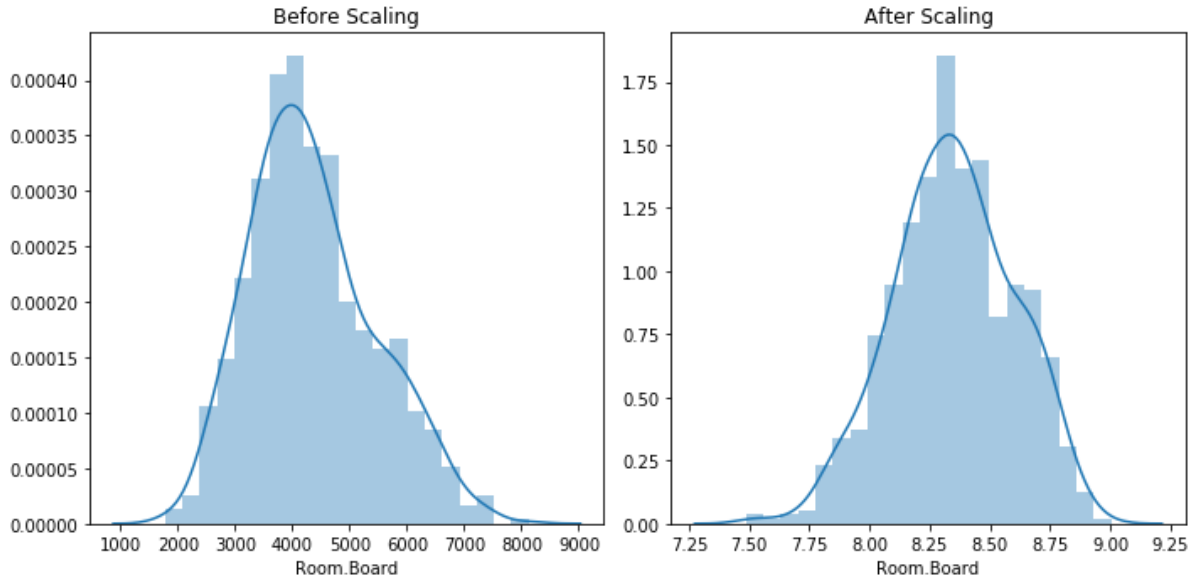
P.Undergrad



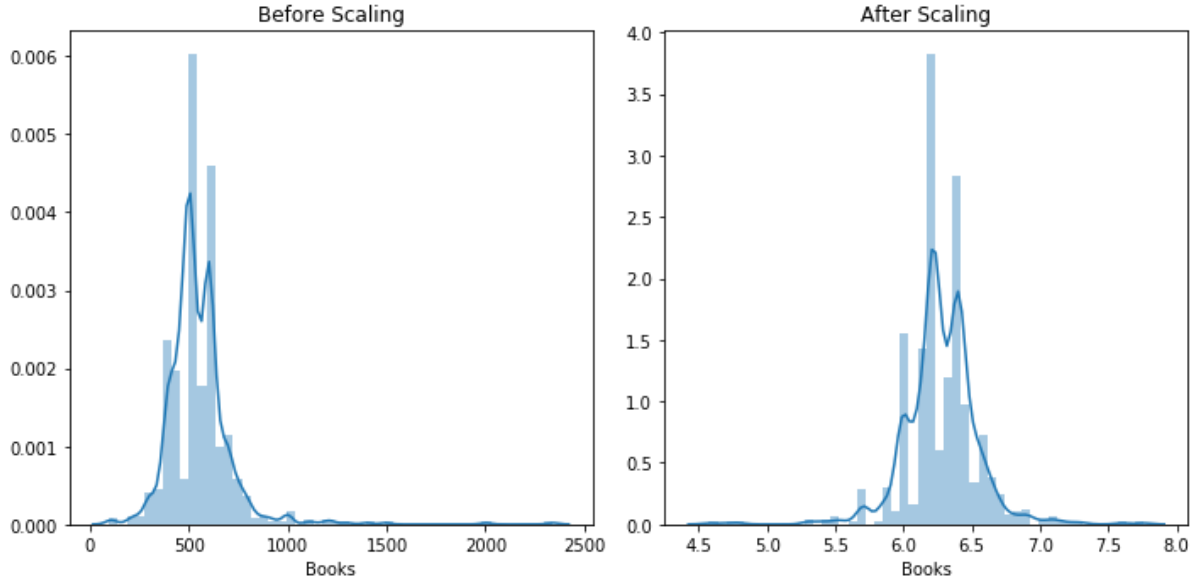
Outstate



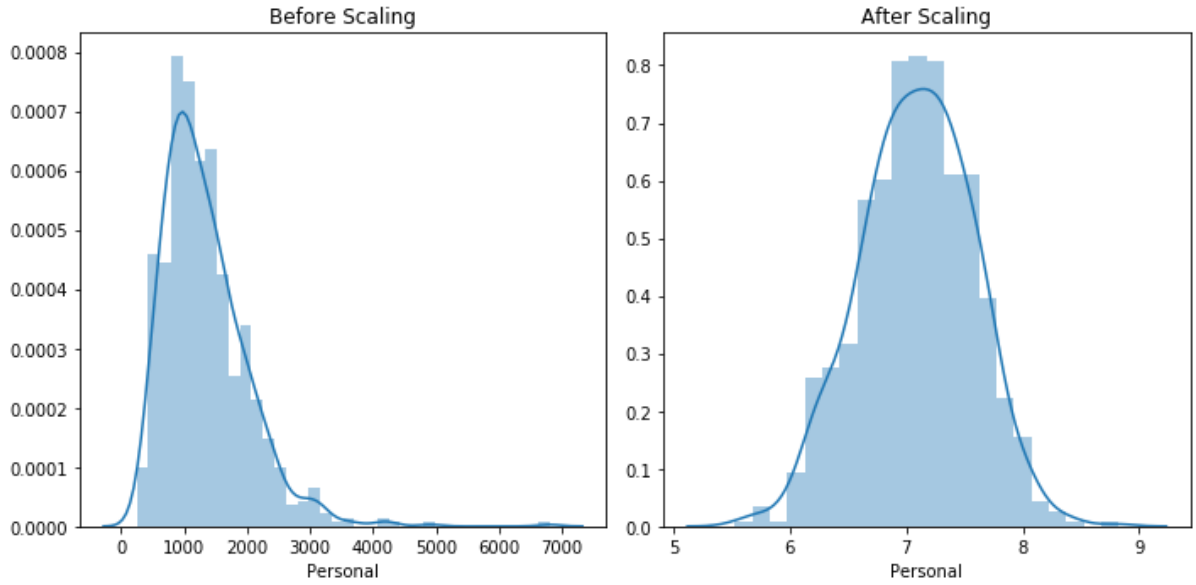
Room.Board



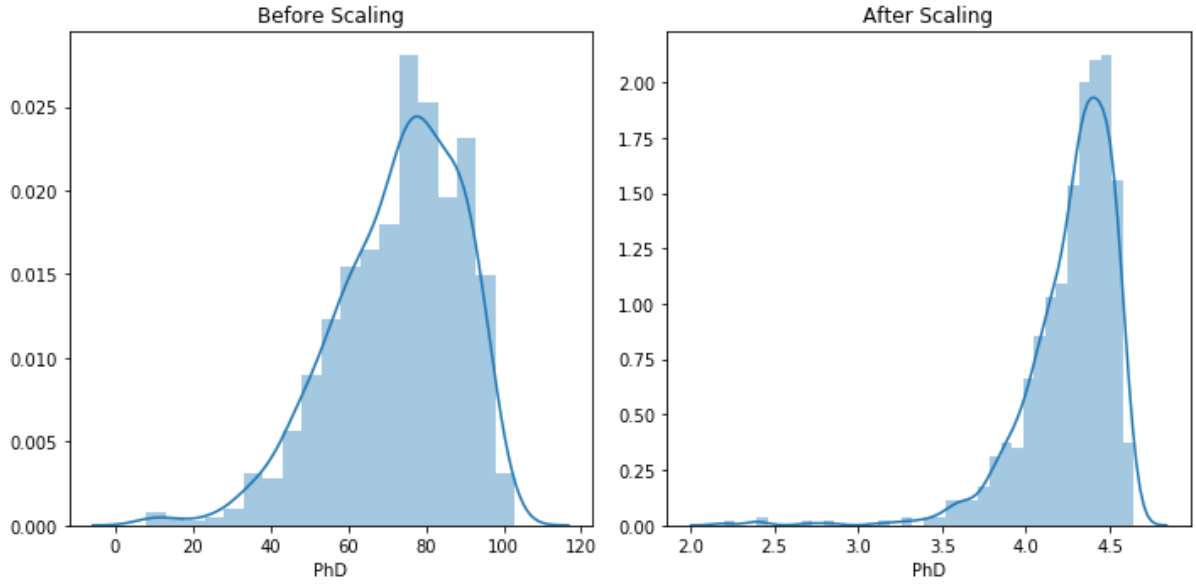
Books



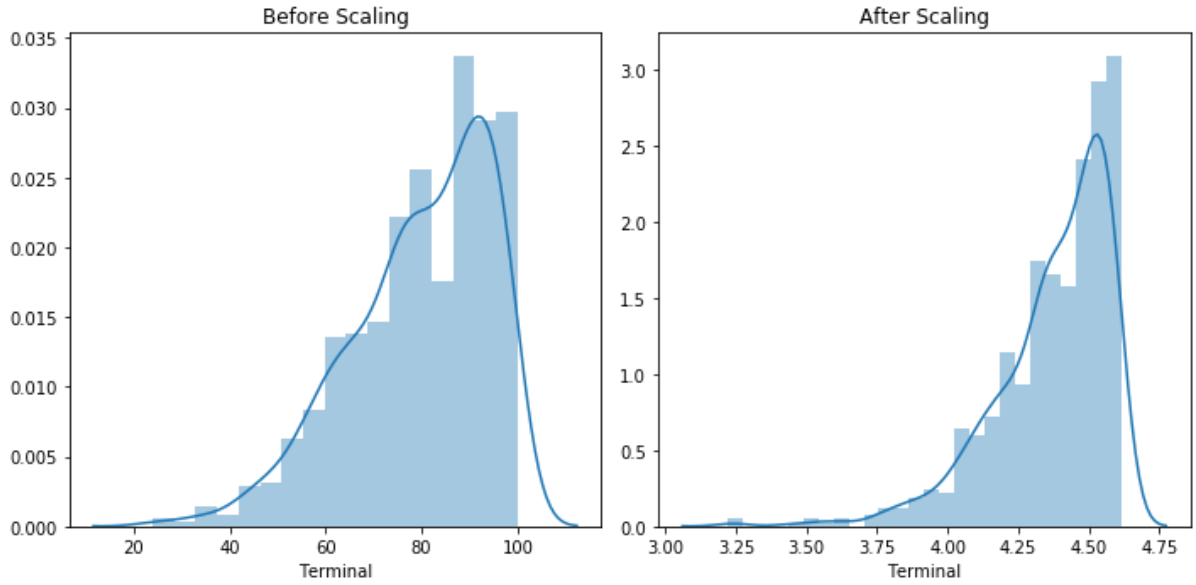
Personal



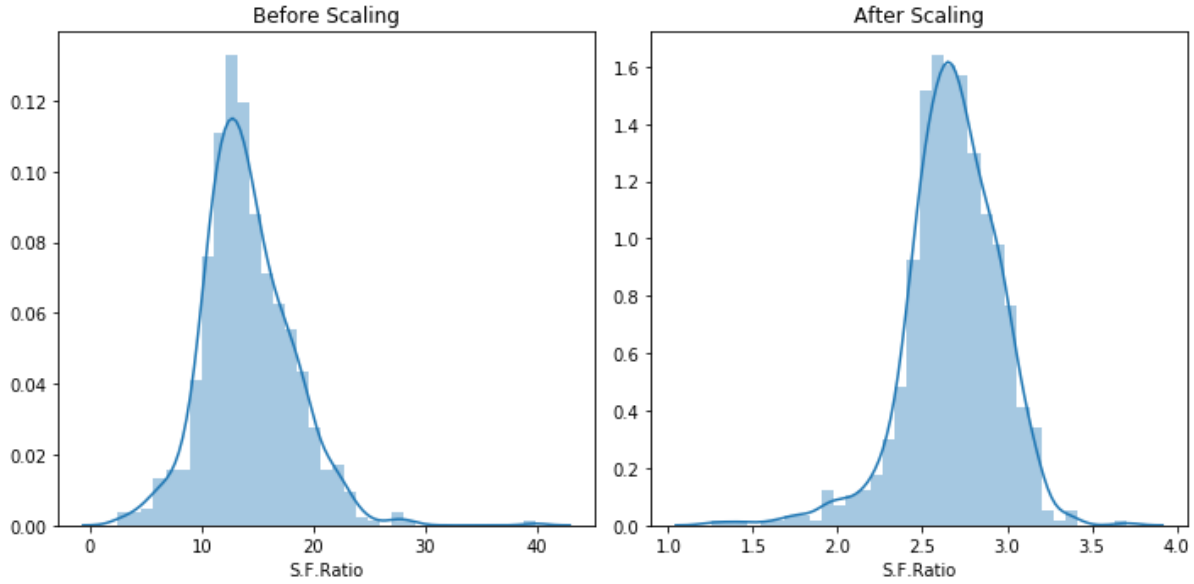
PhD



Terminal

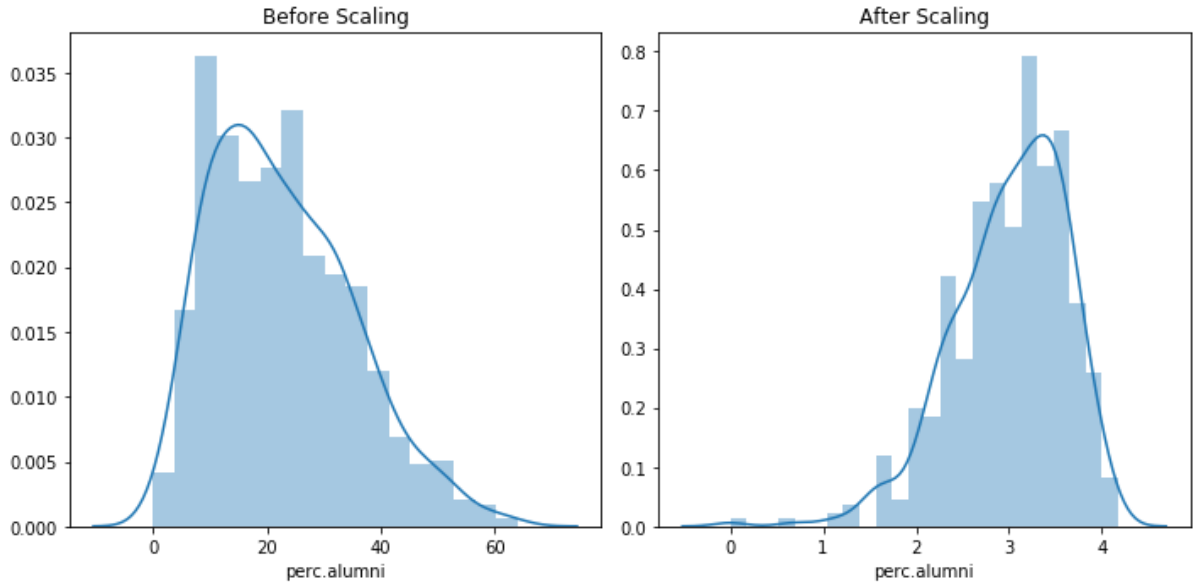


S.F.Ratio

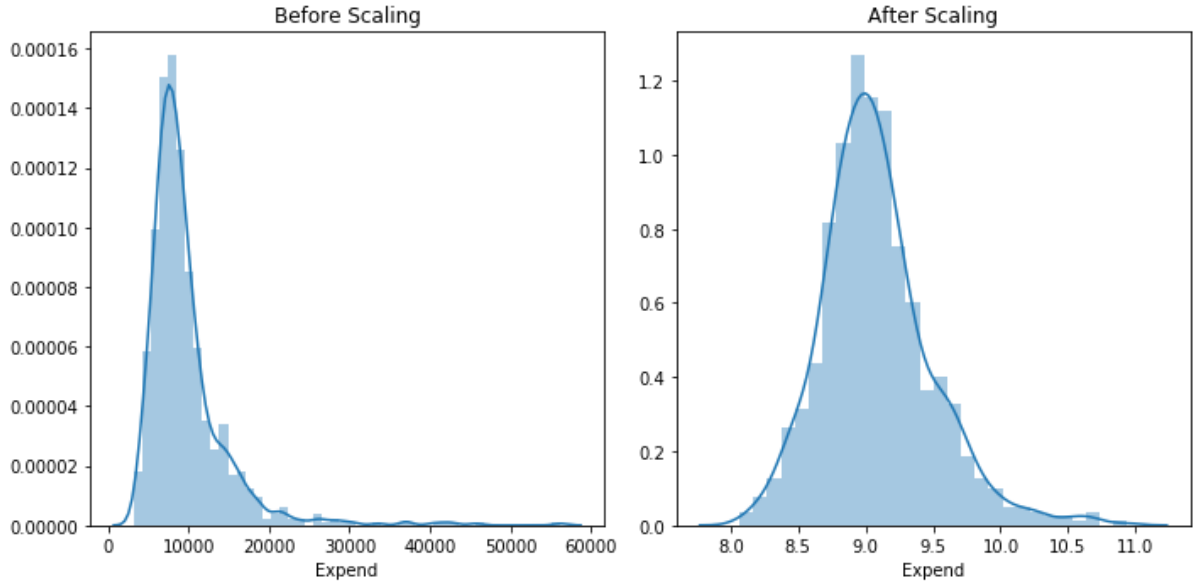


---

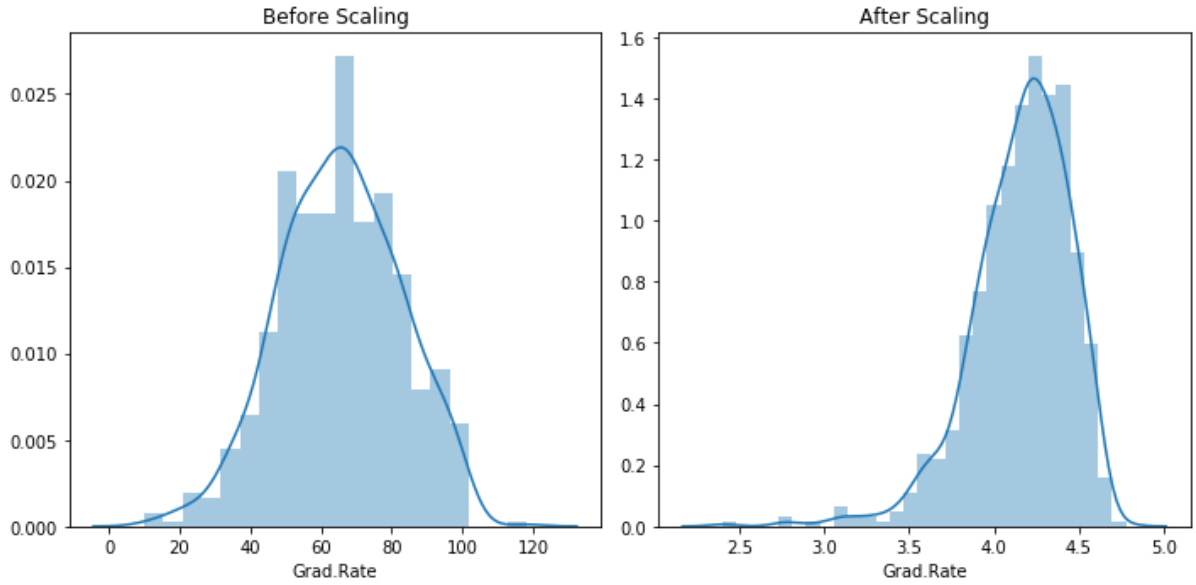
perc.alumni



Expend



Grad.Rate



The below is the comparison for number of outliers, % of number of outliers in each variable and the skewness for before scaling and after logarithmic scaling.

	Number of outliers (Raw Data)	% of outliers (Raw Data)	Skewness (Raw Data)	Number of outliers (After scaling)	% of outliers (After scaling)	Skewness (After scaling)
Apps	70	9.01	Right Skewed	1	0.13	Right Skewed
Accept	73	9.4	Right Skewed	2	0.26	Right Skewed
Enroll	79	10.17	Right Skewed	0	0	Right Skewed
Top10perc	39	5.02	Right Skewed	11	1.42	Left Skewed
Top25perc	0	0	Right Skewed	8	1.03	Left Skewed
F.Undergrad	97	12.48	Right Skewed	0	0	Right Skewed
P.Undergrad	67	8.62	Right Skewed	4	0.51	Left Skewed
Outstate	1	0.13	Right Skewed	4	0.51	Left Skewed
Room.Board	7	0.9	Right Skewed	4	0.51	Left Skewed
Books	46	5.92	Right Skewed	46	5.92	Left Skewed
Personal	20	2.57	Right Skewed	7	0.9	Left Skewed
PhD	8	1.03	Left Skewed	24	3.09	Left Skewed
Terminal	8	1.03	Left Skewed	28	3.6	Left Skewed
S.F.Ratio	12	1.54	Right Skewed	22	2.83	Left Skewed
perc.alumni	5	0.64	Right Skewed	12	1.54	Left Skewed
Expend	48	6.18	Right Skewed	22	2.83	Right Skewed
Grad.Rate	4	0.51	Left Skewed	15	1.93	Left Skewed
Total	584	NaN	NaN	210	NaN	NaN

The skewness of the variables before and after scaling is shown below:

	Raw Data	After Scaling
Apps	3.72375	0.188485
Accept	3.417727	0.179966
Enroll	2.690465	0.373329
Top10perc	1.413217	-0.433738
Top25perc	0.25934	-0.668823
F.Undergrad	2.610458	0.517054
P.Undergrad	5.692353	-0.362271
Outstate	0.509278	-0.345097
Room.Board	0.477356	-0.143616
Books	3.485025	-0.366866
Personal	1.742497	-0.105722
PhD	-0.76817	-2.409729
Terminal	-0.816542	-1.516196
S.F.Ratio	0.667435	-0.783541
perc.alumni	0.606891	-0.880574
Expend	3.459322	0.845072
Grad.Rate	-0.113777	-1.234452

The following inferences can be drawn from the table above:

- The amount of outliers have decreased to a great extent. The original number of outliers were 584 whereas after scaling the number of outliers have decreased to 210.

- There were 3 variables in the raw data that were left skewed and the remaining were right skewed. After scaling, there are 5 variables that are right skewed while the remaining are all left skewed.
- Variables like 'Apps', 'Accept' etc show great improvement in the skewness. i.e. they are tending more towards normal distribution.
- However the variables like 'PhD', which were already left skewed end up being more skewed. Hence they need more attention in further processing.

Question 5:

Build the covariance matrix, eigenvalues, and eigenvector.

Answer:

There are a few steps to be performed before we calculate the eigen values and eigen vectors:

1. We check for outliers and then correct them. In this case we impute the outliers with the whisker closest to them. i.e. values that are greater than (1.5*IQR + Q3) are capped at (1.5*IQR + Q3) and values lesser than (1.5*IQR – Q1) are capped at (1.5*IQR – Q1).
2. Then we will convert the data into standard scaler. i.e. the z-score for each element of the data.
3. Then we find the covariance matrix of the standardized data. (Verify the obtained output with the correlation matrix of the data before standardization)
4. We then compute the eigen values and eigen vectors for the data.

Covariance Matrix (of the raw data):

	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	1.50E+07	8.95E+06	3.05E+06	23132.77314	26952.66348	1.53E+07	2.35E+06	7.81E+05	7.00E+05	84703.75	4.68E+05	24689.43	21053.068	1465.0606	-4327.122381	5.25E+06	9756.42164
Accept	8.95E+06	6.01E+06	2.08E+06	8321.124872	12013.40476	1.04E+07	1.65E+06	-2.54E+05	2.44E+05	45942.81	3.34E+05	14238.2	12182.094	1709.8382	-4859.487022	1.60E+06	2834.16292
Enroll	3.05E+06	2.08E+06	8.63E+05	2971.583415	4172.592435	4.35E+06	7.26E+05	-5.81E+05	-4.10E+04	17291.2	1.77E+05	5028.961	4217.086	872.68477	-2081.693787	3.11E+05	-356.58798
Top10perc	2.31E+04	8.32E+03	2.97E+03	311.182456	311.63048	1.21E+04	-2.83E+03	3.99E+04	7.19E+03	346.1774	-1.11E+03	153.1849	127.55158	-26.87453	99.567208	6.09E+04	149.992164
Top25perc	2.70E+04	1.20E+04	4.17E+03	311.63048	392.229216	1.92E+04	-1.62E+03	3.90E+04	7.20E+03	377.7593	-1.08E+03	176.5184	153.00261	-23.0972	102.550946	5.45E+04	162.371398
F.Undergra	1.53E+07	1.04E+07	4.35E+06	12089.11368	19158.95278	2.35E+07	4.21E+06	-4.21E+06	-3.66E+05	92535.76	1.04E+06	2521.178	21424.242	5370.2086	-13791.92969	4.72E+05	-6563.3075
P.Undergra	2.35E+06	1.65E+06	7.26E+05	-2829.47498	-1615.41214	4.21E+06	2.32E+06	-1.55E+06	-1.02E+05	20410.45	3.30E+05	3706.756	3180.5966	1401.3026	-5297.33709	-6.64E+05	-6721.0625
Outstate	7.81E+05	-2.54E+05	-5.81E+05	39907.17983	38992.4275	-4.21E+06	-1.55E+06	1.62E+07	2.89E+06	25808.24	-8.15E+05	25157.52	24164.148	-8835.254	28229.55307	1.41E+07	39479.6818
Room.Boar	7.00E+05	2.44E+05	-4.10E+04	7186.705605	7199.903568	-3.66E+05	-1.02E+05	2.89E+06	1.20E+06	23170.31	-1.48E+05	5895.035	6047.2997	-1574.206	3701.431379	2.87E+06	8005.36018
Books	8.47E+04	4.59E+04	1.73E+04	346.177405	377.759266	9.25E+04	2.04E+04	2.58E+04	2.32E+04	27259.78	2.00E+04	72.53424	242.96392	-20.86721	-82.263132	9.69E+04	3.008837
Personal	4.68E+05	3.34E+05	1.77E+05	-1114.55119	-1083.60507	1.04E+06	3.30E+05	-8.15E+05	-1.48E+05	20043.03	4.58E+05	-120.899	-305.1542	365.41577	-2399.310824	-3.46E+05	-3132.6149
PhD	2.47E+04	1.42E+04	5.03E+03	153.18487	176.518449	2.52E+04	3.71E+03	2.52E+04	5.90E+03	72.53424	-1.21E+02	266.6086	204.23133	-8.436492	50.38323	3.69E+04	85.557109
Terminal	2.11E+04	1.22E+04	4.22E+03	127.551581	153.002612	2.14E+04	3.18E+03	2.42E+04	6.05E+03	242.9639	-3.05E+02	204.2313	216.74784	-9.330256	48.734327	3.37E+04	73.220396
S.F.Ratio	1.47E+03	1.71E+03	8.73E+02	-26.874525	-23.097199	5.37E+03	1.40E+03	-8.84E+03	-1.57E+03	-20.8672	3.65E+02	-8.43649	-9.330256	15.668528	-19.764109	-1.21E+04	-20.854888
perc.alumn	-4.33E+03	-4.86E+03	-2.08E+03	99.567208	102.550946	-1.38E+04	-5.30E+03	2.82E+04	3.70E+03	-82.2631	-2.40E+03	50.38323	48.734327	-19.76411	153.556744	2.70E+04	104.493815
Expend	5.25E+06	1.60E+06	3.11E+05	60879.3102	54546.48331	4.72E+05	-6.64E+05	1.41E+07	2.87E+06	96912.58	-3.46E+05	36898.06	33733.457	-12067.56	27028.92147	2.73E+07	35012.9683
Grad.Rate	9.76E+03	2.83E+03	-3.57E+02	149.992164	162.371398	-6.56E+03	-6.72E+03	3.95E+04	8.01E+03	3.008837	-3.13E+03	85.55711	73.220396	-20.85489	104.493815	3.50E+04	

Covariance Matrix (of the standardized data):

	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	1.001289	0.956538	0.898039	0.321756	0.364961	0.862111	0.520493	0.065421	0.187717	0.236442	0.230244	0.464522	0.435038	0.126574	-0.101288	0.243248	0.150998
Accept	0.956538	1.001289	0.936482	0.223586	0.274033	0.89819	0.573429	-0.005009	0.11974	0.208974	0.256676	0.427891	0.403929	0.188749	-0.165729	0.162017	0.079084
Enroll	0.898039	0.936482	1.001289	0.171977	0.230731	0.968549	0.642422	-0.155856	-0.023876	0.202317	0.339785	0.382031	0.354836	0.274622	-0.22301	0.054291	-0.023281
Top10perc	0.321756	0.223586	0.171977	1.001289	0.915053	0.111358	-0.180241	0.562884	0.357826	0.15365	-0.11688	0.544749	0.507401	-0.388426	0.456384	0.657886	0.494307
Top25perc	0.364961	0.274033	0.230731	0.915053	1.001289	0.181429	-0.099423	0.4902	0.331413	0.16998	-0.086922	0.552172	0.528334	-0.297616	0.417369	0.573643	0.479602
F.Undergra	0.862111	0.89819	0.968549	0.111358	0.181429	1.001289	0.697027	-0.226457	-0.054546	0.208147	0.360246	0.36203	0.335486	0.324922	-0.285825	0.000371	-0.082345
P.Undergra	0.520493	0.573429	0.642422	-0.180241	-0.099423	0.697027	1.001289	-0.354673	-0.067725	0.122686	0.344496	0.127827	0.122309	0.371085	-0.419874	-0.20219	-0.265499
Outstate	0.065421	-0.00501	-0.15586	0.562884	0.4902	-0.226457	-0.354673	1.001289	0.656334	0.005117	-0.326029	0.391825	0.41311	-0.574422	0.566465	0.776327	0.573196
Room.Boar	0.187717	0.11974	-0.02388	0.357826	0.331413	-0.054546	-0.067725	0.656334	1.001289	0.109065	-0.219837	0.341909	0.379759	-0.376915	0.272744	0.58137	0.426339
Books	0.236442	0.208974	0.202317	0.15365	0.16998	0.208147	0.122686	0.005117	0.109065	1.001289	0.240172	0.136566	0.159523	-0.008547	-0.042887	0.150177	-0.008061
Personal	0.230244	0.256676	0.339785	-0.11688	-0.086922	0.360246	0.344496	-0.326029	-0.219837	0.240172	1.001289	-0.0117	-0.032012	0.174137	-0.306147	-0.16348	-0.291269
PhD	0.464522	0.427891	0.382031	0.544749	0.552172	0.36203	0.127827	0.391825	0.341909	0.136566	-0.011699	1.001289	0.86404	-0.129556	0.249198	0.511187	0.310419
Terminal	0.435038	0.403929	0.354836	0.507401	0.528334	0.335486	0.122309	0.41311	0.379759	0.159523	-0.032012	0.86404	1.001289	-0.151188	0.266375	0.524744	0.29318
S.F.Ratio	0.126574	0.188749	0.274622	-0.388426	-0.297616	0.324922	0.371085	-0.574422	-0.376915	-0.00855	0.174137	-0.12956	-0.151188	1.001289	-0.412632	-0.65522	-0.308922
perc.alumn	-0.10129	-0.16573	-0.22301	0.456384	0.417369	-0.285825	-0.419874	0.566465	0.272744	-0.04289	-0.306147	0.249198	0.266375	-0.412632	1.001289	0.463519	0.492041
Expend	0.243248	0.162017	0.054291	0.657886	0.573643	0.000371	-0.202189	0.776327	0.58137	0.150177	-0.163481	0.511187	0.524744	-0.65522	0.463519	1.001289	0.415826
Grad.Rate	0.150998	0.079084	-0.02328	0.494307	0.479602	-0.082345	-0.265499	0.573196	0.426339	-0.00806	-0.291269	0.310419	0.29318	-0.308922	0.492041	0.415826	1.001289

Eigen Values:

Eigen Values
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 Vector:

	Apps	Accept	Enroll	Top10pe	Top25pe	F.Underg	P.Underg	Outstate	Room.Bc	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alur	Expend	Grad.Rate
0	-0.2622	0.3141	0.081	-0.0988	-0.2199	0.0022	-0.0284	-0.0899	0.1306	-0.1565	-0.0862	0.1822	-0.5991	0.09	-0.0889	0.5494	0.005415
1	-0.2306	0.3446	0.1077	-0.1181	-0.1896	-0.0165	-0.013	-0.1376	0.1423	-0.1492	-0.0426	-0.391	0.6615	0.1589	-0.0438	0.2916	0.014458
2	-0.1893	0.3828	0.0855	-0.0093	-0.1623	-0.0681	-0.0152	-0.1442	0.0509	-0.0649	-0.0438	0.7167	0.2332	-0.0354	0.0619	-0.417	-0.04979
3	-0.3389	-0.0993	-0.0788	0.3691	-0.1572	-0.0889	-0.2575	0.2895	-0.1225	-0.0359	0.0018	-0.0562	0.0221	-0.0392	-0.07	0.0088	-0.72365
4	-0.3347	-0.0595	-0.0508	0.4168	-0.1444	-0.0276	-0.239	0.3456	-0.1939	0.0064	-0.1021	0.0197	0.0323	0.1456	0.097	-0.0108	0.655465
5	-0.1633	0.3986	0.0737	-0.014	-0.1027	-0.0516	-0.0312	-0.1087	0.0015	-0.0002	-0.035	-0.5428	-0.3677	-0.1336	0.0872	-0.5707	0.025306
6	-0.0225	0.3576	0.0404	-0.2254	0.0957	-0.0245	-0.01	0.1238	-0.6348	0.5463	0.2521	0.0295	0.0262	0.0502	-0.0446	0.1463	-0.03971
7	-0.2835	-0.2519	0.0149	-0.263	-0.0373	-0.0204	0.0945	0.0113	-0.0084	-0.2318	0.5934	0.001	-0.0814	0.5604	-0.0672	-0.2116	-0.00159
8	-0.2442	-0.1319	-0.0211	-0.5809	0.0691	0.2373	0.0945	0.3896	-0.2205	-0.2551	-0.4753	0.0099	0.0268	-0.1074	-0.0178	-0.1009	-0.02826
9	-0.0967	0.094	-0.6971	0.0362	-0.0354	0.6386	-0.1112	-0.2398	0.021	0.0912	0.0436	0.0044	0.0105	0.0516	-0.0354	-0.0286	-0.00806
10	0.0352	0.2324	-0.531	0.115	0.0005	-0.3815	0.6394	0.2772	0.0174	-0.1276	0.0152	-0.0109	0.0045	0.0094	0.0119	0.0338	0.001426
11	-0.3264	0.0551	0.0811	0.1473	0.5508	0.0033	0.0892	-0.0343	0.1665	0.101	-0.0392	0.0133	0.0125	-0.0717	-0.7027	-0.0638	0.083147
12	-0.3231	0.043	0.059	0.089	0.5904	0.0354	0.0917	-0.0903	0.1126	0.086	-0.0849	0.0074	-0.0179	0.1638	0.6625	0.0985	-0.11337
13	0.1632	0.2598	0.2742	0.2595	0.1428	0.4688	0.1529	0.2428	-0.1537	-0.4705	0.363	0.0089	0.0183	-0.2399	0.0479	0.062	0.003832
14	-0.1866	-0.2571	0.1037	0.224	-0.1282	0.0126	0.3914	-0.5661	-0.5392	-0.1476	-0.1739	-0.0241	-8E-05	-0.049	-0.0359	0.0281	-0.00733
15	-0.329	-0.16	-0.1842	-0.2138	0.0224	-0.2316	-0.1505	-0.1188	0.0242	-0.0804	0.3937	0.0106	0.056	-0.6904	0.1267	0.1287	0.1451
16	-0.2388	-0.1675	0.2453	0.0362	-0.3568	0.3136	0.4686	0.1805	0.3158	0.4884	0.0873	-0.0025	0.0148	-0.1593	0.0631	-0.0071	-0.00329

Question 6:

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

Answer:

The explicit form of first PC is:

(-0.262171542236867)*(Apps)+(-0.230562460758459)*(Accept)+
(-0.189276397098432)*(Enroll)+(-0.338874521368412)*(Top10perc)+
(-0.334690531786966)*(Top25perc)+(-0.163293009839948)*(F.Undergrad)+
(-0.0224797090797236)*(P.Undergrad)+(-0.283547285249585)*(Outstate)+
(-0.244186587778693)*(Room.Board)+
(-0.0967082753803971)*(Books)+(0.0352299593764859)*(Personal)+
(-0.326410695575945)*(PhD)+(-
0.323115980266162)*(Terminal)+(0.163151641594741)*(S.F.Ratio)+
(-0.186610828188054)*(perc.alumni)+(-0.328955847333676)*(Expend)+
(-0.238822446648551)*(Grad.Rate)

The first row of the standardized data is:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	-0.376493	-0.3378299	0.10638048	-0.2467801	-0.1918274	-0.018769	-0.1660827	-0.7464802	-0.968324	-0.7765671	1.43849994	-0.1740451	-0.123239	1.07060212	-0.8704662	-0.6309158	-0.3192054

The value for the first element in the reduced data will be:

1.602499375

The first PC will be as follows:

	0	1	2	3	4	5	.....	771	772	773	774	775	776
PC1	1.602499	1.804675	1.608283	-2.803644	2.200868	0.730164	.....	-2.936392	3.395392	-0.31975	0.576883	-6.57095	0.477393

The only difference between using linear algebra and sklearn package is the sign, apart from that the values and the outcome remains the same.

Question 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.

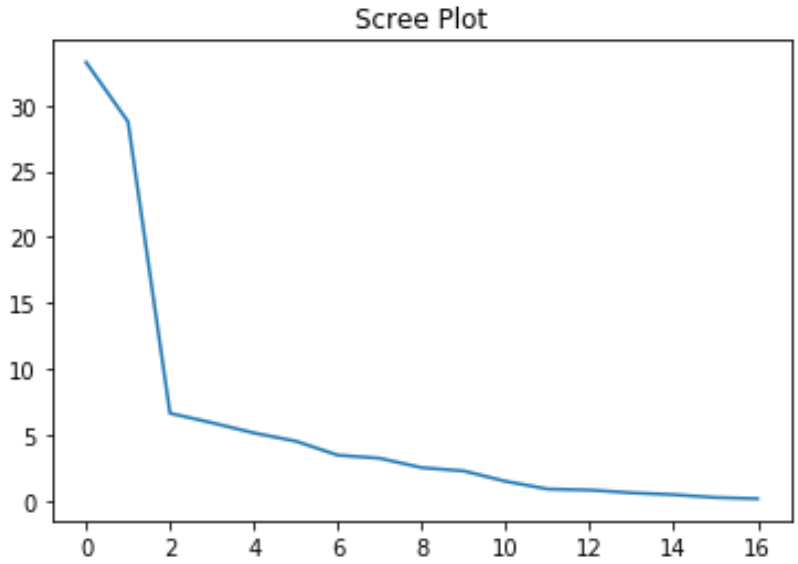
Answer:

The eigen values obtained are:

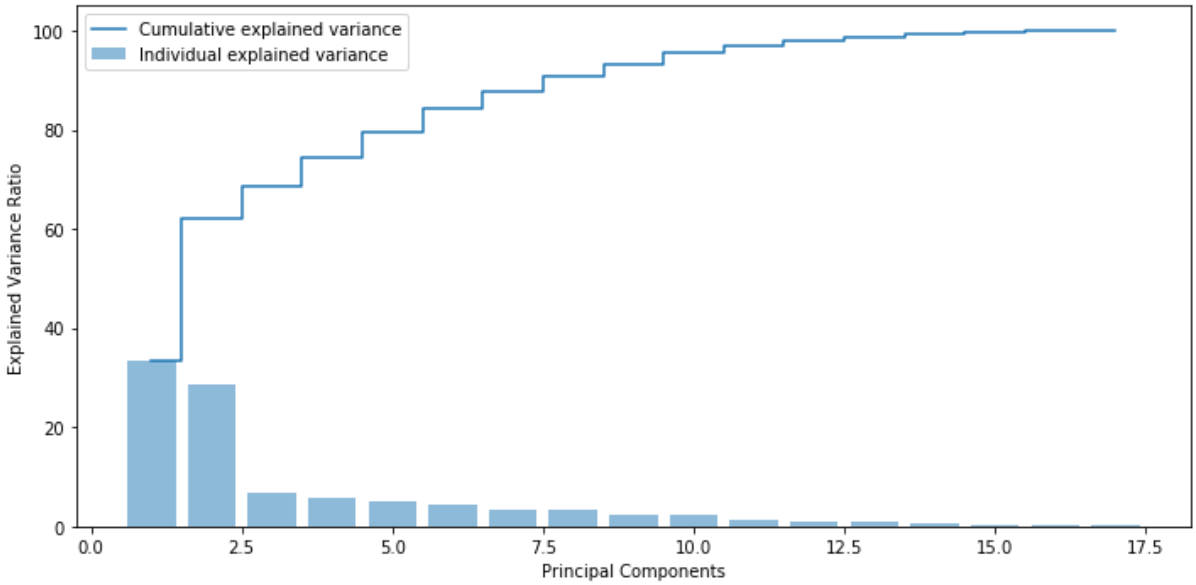
```
array([5.6625219 , 4.89470815, 1.12636744, 1.00397659, 0.87218426,
       0.7657541 , 0.58491404, 0.5445048 , 0.42352336, 0.38101777,
       0.24701456, 0.02239369, 0.03789395, 0.14726392, 0.13434483,
       0.09883384, 0.07469003])
```

On arranging in ascending and finding the % cumulative variance contributed by each PC, we get:

```
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.        ]
```



A knee point from the scree plot is observed when we consider 3 PCs. However it covers only 68% of the total variance of data.



We can decide on the optimum number of PC based on the requirement. In this instance let's assume we want to capture 90% variance or above for the entire data set. In that case we will have to go ahead with the first 8 PCs.

Eigen vectors indicate the direction of axis that covers the variance of that particular PC. All the PCs are orthogonal. Having said that, the PC1 vector picks up maximum variation in the data and its direction. The weightage/coefficient for the vectors are the respective standardized data. While PC2 picks up residual in the orthogonal direction. Similarly the residual left behind by PC1 and PC2 is picked by PC3. This goes on to continue till PC18

The PCA is performed and exported. The reduced data frame is exported to ''ReducedData.xlsx'' and the eigen vector is exported to "EigenVector.xlsx".

Question 8:

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

Answer:

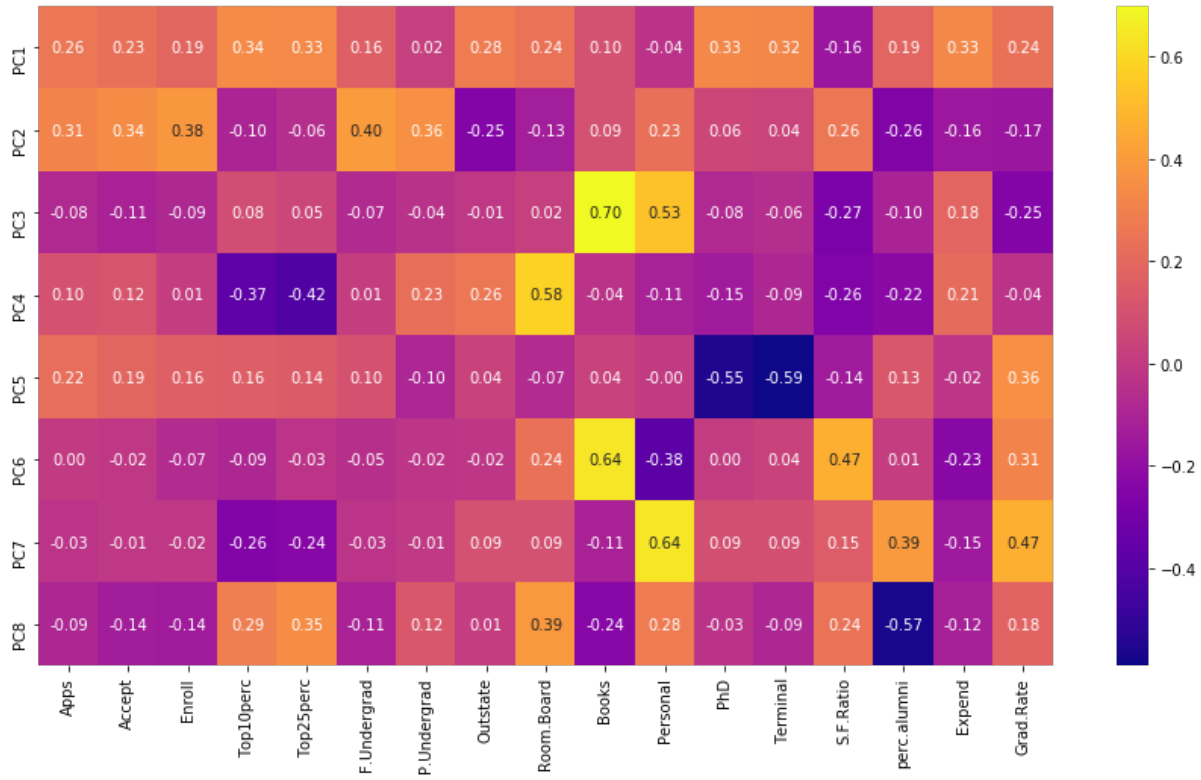
The below are the main business implications of PCA are:

- Here we have considered the 8 PCs to capture 90% of the variance. We have reduced the data from 18 dimensions to 8 dimensions.
- The correlation of the eigen vector is shown below:

	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	1.00E+00	4.94E-01	-1.37E-01	0.058562	-0.021447	2.81E-01	4.19E-01	1.14E-01	-3.14E-01	-0.13858	2.33E-01	-0.00829	0.018119	-0.102192	0.017897	-2.18E-02	-0.014092
Accept	4.94E-01	1.00E+00	3.36E-02	-0.014346	0.005254	-6.89E-02	-1.03E-01	-2.80E-02	7.69E-02	0.033949	-5.70E-02	0.00203	-0.004439	0.025035	-0.004384	5.34E-03	0.003452
Enroll	-1.37E-01	3.36E-02	1.00E+00	0.003984	-0.001459	1.91E-02	2.85E-02	7.77E-03	-2.14E-02	-0.00943	1.58E-02	-0.00056	0.001233	-0.006952	0.001218	-1.48E-03	-0.000959
Top10perc	5.86E-02	-1.43E-02	3.98E-03	1	0.000623	-8.17E-03	-1.22E-02	-3.32E-03	9.12E-03	0.004023	-6.76E-03	0.000241	-0.000526	0.002967	-0.00052	6.33E-04	0.000409
Top25perc	-2.14E-02	5.25E-03	-1.46E-03	0.000623	1	2.99E-03	4.45E-03	1.22E-03	-3.34E-03	-0.00147	2.48E-03	-8.8E-05	0.000193	-0.001086	0.00019	-2.32E-04	-0.00015
F.Undergra	2.81E-01	-6.89E-02	1.91E-02	-0.008165	0.00299	1.00E+00	-5.84E-02	-1.59E-02	4.38E-02	0.019322	-3.25E-02	0.001155	-0.002526	0.014249	-0.002495	3.04E-03	0.001965
P.Undergra	4.19E-01	-1.03E-01	2.85E-02	-0.01216	0.004454	-5.84E-02	1.00E+00	-2.37E-02	6.52E-02	0.028777	-4.84E-02	0.001721	-0.003762	0.02122	-0.003716	4.53E-03	0.002926
Outstate	1.14E-01	-2.80E-02	7.77E-03	-0.003317	0.001215	-1.59E-02	-2.37E-02	1.00E+00	1.78E-02	0.007849	-1.32E-02	0.000469	-0.001026	0.005788	-0.001014	1.23E-03	0.000798
Room.Boar	-3.14E-01	7.69E-02	-2.14E-02	0.009118	-0.003339	4.38E-02	6.52E-02	1.78E-02	1.00E+00	-0.02158	3.63E-02	-0.00129	0.002821	-0.015911	0.002786	-3.39E-03	-0.002194
Books	-1.39E-01	3.39E-02	-9.43E-03	0.004023	-0.001473	1.93E-02	2.88E-02	7.85E-03	-2.16E-02	1	1.60E-02	-0.00057	0.001245	-0.00702	0.001229	-1.50E-03	-0.000968
Personal	2.33E-01	-5.70E-02	1.58E-02	-0.00676	0.002476	-3.25E-02	-4.84E-02	-1.32E-02	3.63E-02	0.015996	1.00E+00	0.000956	-0.002091	0.011796	-0.002066	2.52E-03	0.001627
PhD	-8.29E-03	2.03E-03	-5.64E-04	0.000241	-0.000088	1.16E-03	1.72E-03	4.69E-04	-1.29E-03	-0.00057	9.56E-04	1	0.000074	-0.00042	0.000074	-9.00E-05	-0.000058
Terminal	1.81E-02	-4.44E-03	1.23E-03	-0.000526	0.000193	-2.53E-03	-3.76E-03	-1.03E-03	2.82E-03	0.001245	-2.09E-03	0.000074	1	0.000918	-0.000161	1.96E-04	0.000127
S.F.Ratio	-1.02E-01	2.50E-02	-6.95E-03	0.002967	-0.001086	1.42E-02	2.12E-02	5.79E-03	-1.59E-02	-0.00702	1.18E-02	-0.00042	0.000918	1	0.000907	-1.10E-03	-0.000714
perc.alumn	1.79E-02	-4.38E-03	1.22E-03	-0.00052	0.00019	-2.50E-03	-3.72E-03	-1.01E-03	2.79E-03	0.001229	-2.07E-03	0.000074	-0.000161	0.000907	1	1.93E-04	0.000125
Expend	-2.18E-02	5.34E-03	-1.48E-03	0.000633	-0.000232	3.04E-03	4.53E-03	1.23E-03	-3.39E-03	-0.0015	2.52E-03	-0.00009	0.000196	-0.001104	0.000193	1.00E+00	-0.000152
Grad.Rate	-1.41E-02	3.45E-03	-9.59E-04	0.000409	-0.00015	1.97E-03	2.93E-03	7.98E-04	-2.19E-03	-0.00097	1.63E-03	-5.8E-05	0.000127	-0.000714	0.000125	-1.52E-04	1

This shows that we have reduced the correlation between the variables to a considerable amount. This reduces the problem of multi collinearity when we try to implement models like regression model etc. The PCs are orthogonal and so varying one PC will not affect the output of the other PCs.

- Plotting the heat map for the PCs:



The PC1 captures maximum variance. This PC can be tweaked to get an overall increase in the output. Similarly, ‘Books’ and ‘Personal’ have high weightages in PC3 so if we needed to tweak these variables alone and not changing much of the other variables then we have to tweak PC3. Similarly, PC5 for ‘PhD’ and ‘Terminal’. So on and so forth.

This way we can individually tweak the variables and we can also analyse the change in output for a unit change in the variables.