Question 1:

1.1) State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables ‘A’ and ‘B’ individually.

#1.1 Part A
Hypothesis for one way Anova on A ingredient is as below

- $H0$: The means of 'Relief' variable with respect to each of the treatment levels for A ingredient is equal.
- $H1$: At least one of the means of 'Relief' variable with respect to each of the treatment levels for A ingredient is unequal.
- Alpha: 0.05

#1.1 Part B
Hypothesis for one way Anova on B ingredient is as below

- $H0$: The means of 'Relief' variable with respect to each of the treatment levels for B ingredient is equal.
- $H1$: At least one of the means of 'Relief' variable with respect to each of the treatment levels for B ingredient is unequal.
- Alpha: 0.05

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.

Going by the hypothesis stated in #1.1 Part A above, the one way ANOVA test conducted on the categorical variable A with the dependent/continous variable ‘Relief’ resulted in p_factor much lesser than the alpha (significance value) as below.

	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

Hence, with p_value low, null will go and we reject the null hypothesis. Meaning, at least one of the means of 'Relief' variable with respect to each of the treatment levels for A ingredient is unequal.

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.

Going by the hypothesis stated in #1.1 Part B above, the one way ANOVA test conducted on the categorical variable B with the dependent/continous variable ‘Relief’ resulted in p_factor much lesser than the alpha (significance value) as below.

	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

Hence, with p_value low, null will go and we reject the null hypothesis. Meaning, at least one of the means of 'Relief' variable with respect to each of the treatment levels for B ingredient is unequal.

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?

- In order to visualize the interaction effect of one categorical/independent variable over the other we can perform a point plot of two categorical variable against one continuous variable with one categorical variable (i.e A ingredient) on the x axis and other(i.e B ingredient) being the hue while the continuous variable (i.e Relief) can be plotted on the y axis.
- If the plots result in hue based lines for ingredient B connecting mean levels for each of its treatment levels showing up parallel to each other across the levels for A ingredient then we could infer that for no interaction effect. If they are not exactly parallel at least across any of the levels then we could infer that for a potential interaction effect of A and B ingredients for those levels on the relief variable.

Please find below the point plot as suggested above.

Figure 1: A ingredient plotted in x axis with B ingredient as hue

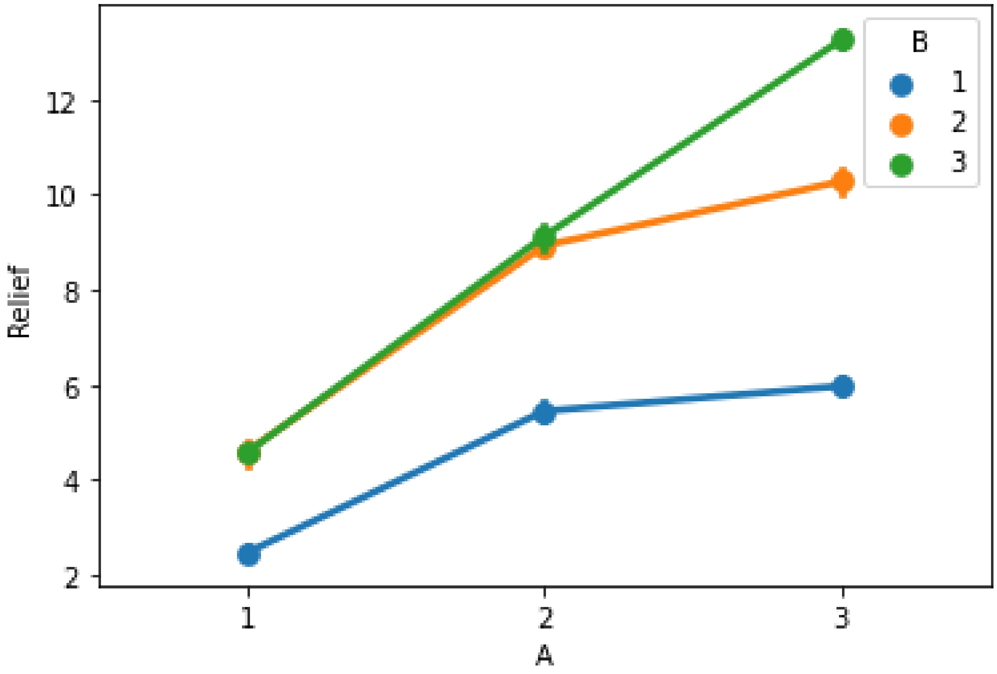
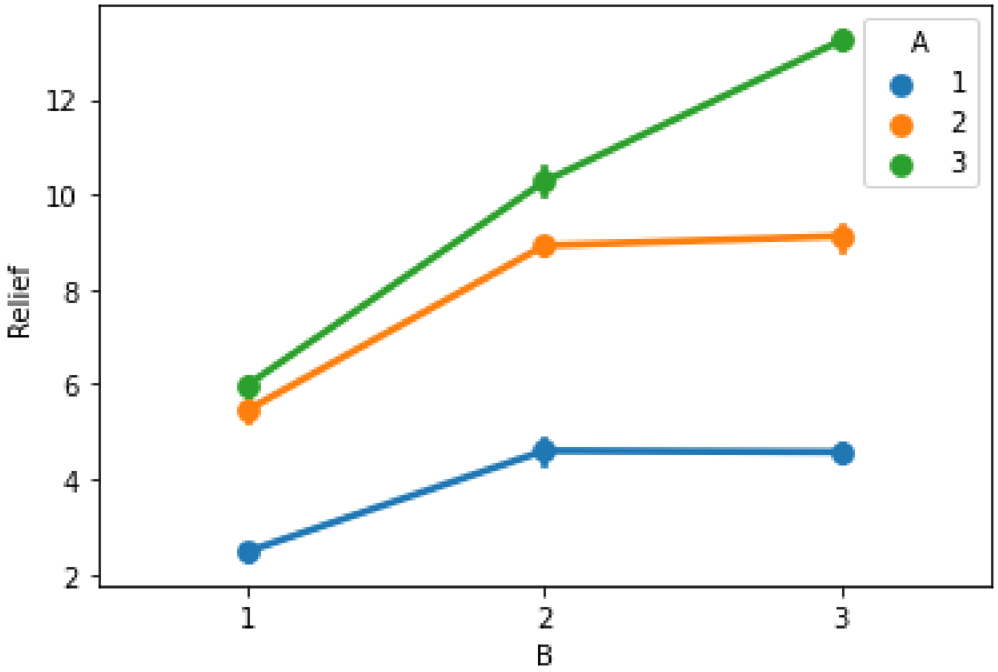


Figure 2: B ingredient plotted in x axis with A ingredient as hue.



Observation from the interaction plot as below:

- Based on the interaction plot above there seems to be some interaction effect between A and B on the dependent variable relief as we can see the lines are not exactly parallel between A2 and A3 within the range of B2 and B3.
- That is A3 and B3 combination of ingredients seems to be have interaction effect resulting in impacting the relief variable positively going by the plot above.

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

Hypothesis on 2 way ANOVA based on two factors (ingredients A and B) along with the interaction between these two factors/variables.

- $H0$: The means of 'Relief' variable with respect to each of the ingredient levels across A and B ingredients are equal.
- $H1$: At least one of the means of 'Relief' variable with respect to each of the ingredient levels across A and B ingredients is unequal.

A two way ANOVA test has been conducted including the interaction effect considering ingredient A and B as categorical variable with ‘Relief’ as dependent variable. Please find below the table that depicts the sum of squares among, its degrees of freedom and the respective mean squares among and the related F statistics for factors such as ingredient A and ingredient B along with their interaction in individual rows accordingly. Finally, p value has been derived based on the F statistics whose observation in the same row respectively.

2 way ANOVA observations as below

	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

Please note that the results in the above table is depicting the p statistics for ingredient A in the first row, ingredient B in the second row, interaction effect in the third row.
Also, residual sum of squares within, its degrees of freedom and its mean square within is given in 4th row.

Result:
1.Since the p value in both of the above scenarios are less than α(0.05), we can say that we reject the null hypothesis ($H0$).
2.Meaning at least one of the means of 'Relief' variable with respect to each of the ingredient levels across A and B ingredients is unequal.

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

ANOVA is the foundational technique for analytics and is also used for experimental design. In essence ANOVA compares means. if the means are organized for one factor/treatment they are one way ANOVA of if more than one like two it shall be called two way ANOVA with or without interactions across factors. Objectively ANOVA test observes effect of independent variables on the dependent variables to suggest potential impact from the treatments.

In this business case since we have two factors such as Ingredient A and B wherein treatments are designed to capture relief from the usage of these ingredients individually across multiple treatment level combinations as well as look for interaction effect of both the factors on the dependent variable 'Relief'. Each of those factors have 3 levels of treatment that has been designed to be applied across 4 volunteers for each of the treatment level combinations across both the ingredients thus making 36 different volunteers yielding 36 observations of the dependent variable 'Relief' from those treatments.

Two way ANOVA test on the sample data has provided the result that each of the ingredients A and B by itself produce difference in the mean effect ('Relief') at least for one pair of treatment level within each of the ingredients individually. Apart from that both the ingredients do have an interaction effect over each other that causes additional effect on the target variable 'Relief'. This is an indication that the both the ingredients does have combination effect in addition to the individual effect by respective factors on ‘Relief’ going by the treatment.

Scree plot presented earlier suggests that there is an increased mean in one of the observed levels across each of the ingredients indicating improved relief effected by each ingredients at those levels.

Question 2:
2.1) Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented.

- Based on the information on the dataset except for Names every other dimension are numeric ones.
- Based on the percentage of unique value counts for each of the integer variable in the dataset along with the data dictionary definitions we can safely assume columns other than Names are continues variables and not the discrete ones.
- There are no missing values across any of the variables
- Also since all the independent variables are integers we can rule out presence of junk characters or unusual values that needs to be imputed for PCA.

2.1 Part A: **Univariate analysis:**
Outliers:
Except for Top25percent every other variable has high amount of outliers with outstate having minimal outlier among them.

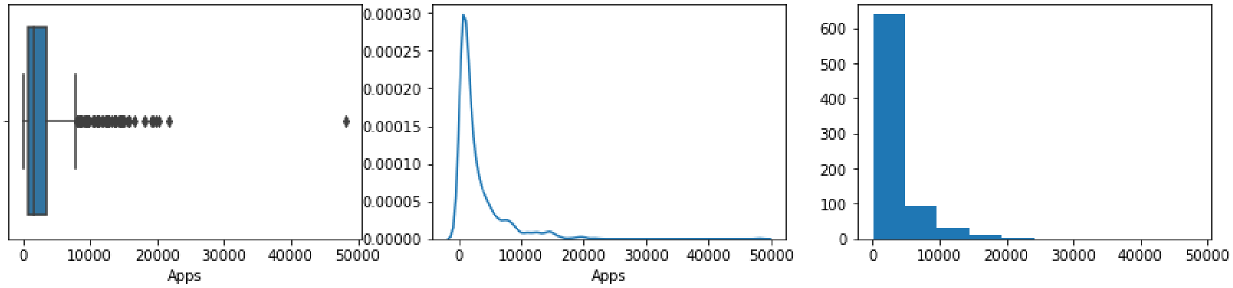
Distribution:
Top25percent while not perfectly normally distributed is somewhat closer to it while rest of the variables are mostly

Skewness:
All of the variables indicates skewness. Terminal, PhD and Grad.Rate being left skewed while rest of the variables are right skewed to the mean.

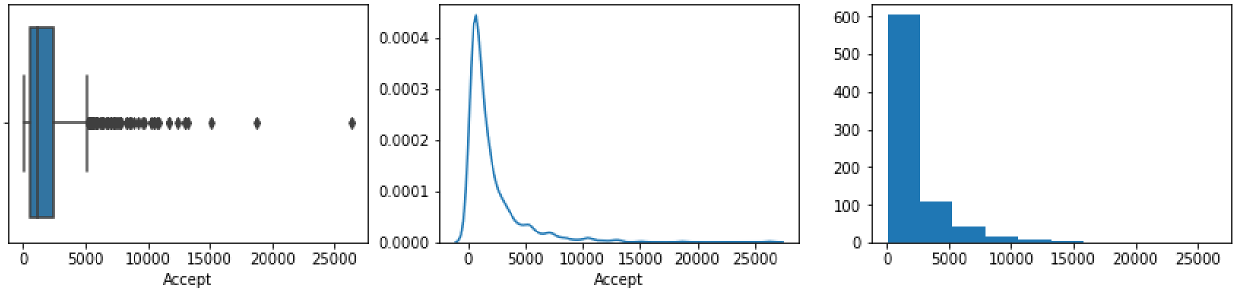
Plots:
Following depicts the univariate analysis for each of the continuous variables in the dataset showcasing boxplot, kernel density estimate diagram and histograms.

- Mean, median and modes are displayed for each variable.
- Shapiro test has been performed in each of the variable to test the normal distribution and accordingly results captured below.
- Outlier test has been done and results mentioned accordingly.

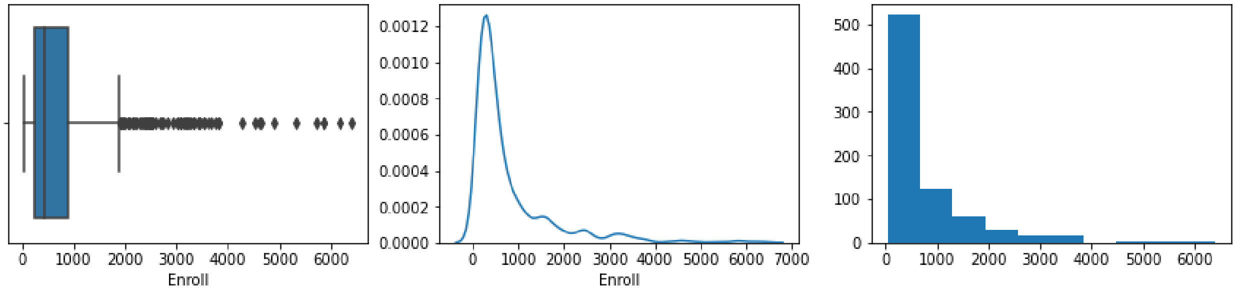
```
1. Univariate analysis for Apps

Mean is 3001.638353, Median is 1558.000000, Modes are [3]
Column Apps has outliers
```



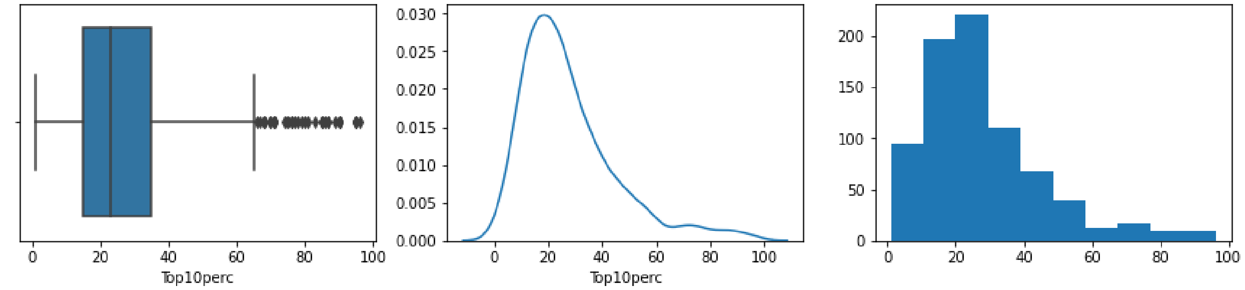
```
Column Apps is not normally distributed

 2. Univariate analysis for Accept

Mean is 2018.804376, Median is 1110.000000, Modes are [4]
Column Accept has outliers
```



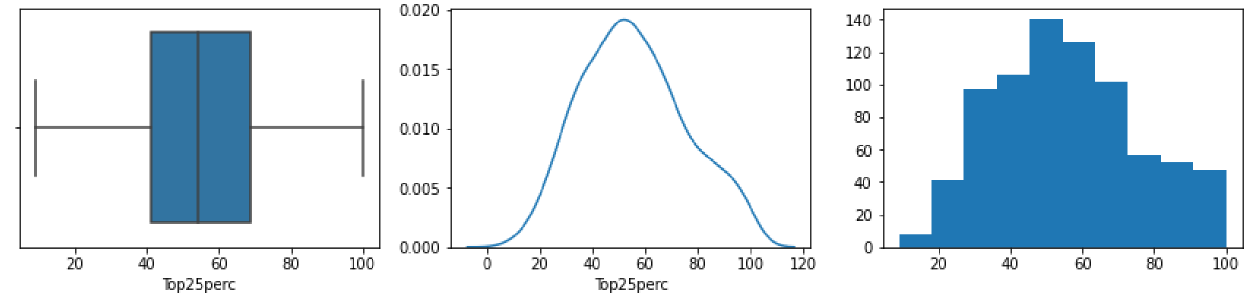
```
Column Accept is not normally distributed

 3. Univariate analysis for Enroll

Mean is 779.972973, Median is 434.000000, Modes are [5]
Column Enroll has outliers
```



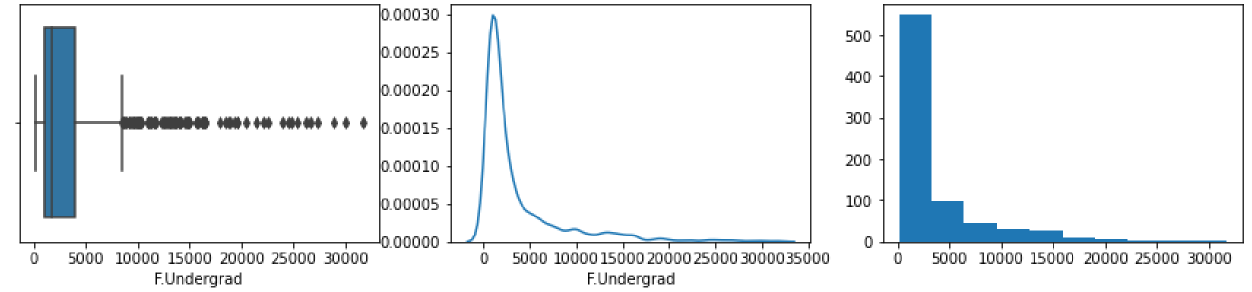
```
Column Enroll is not normally distributed

 4. Univariate analysis for Top10perc

Mean is 27.558559, Median is 23.000000, Modes are [37]
Column Top10perc has outliers
```



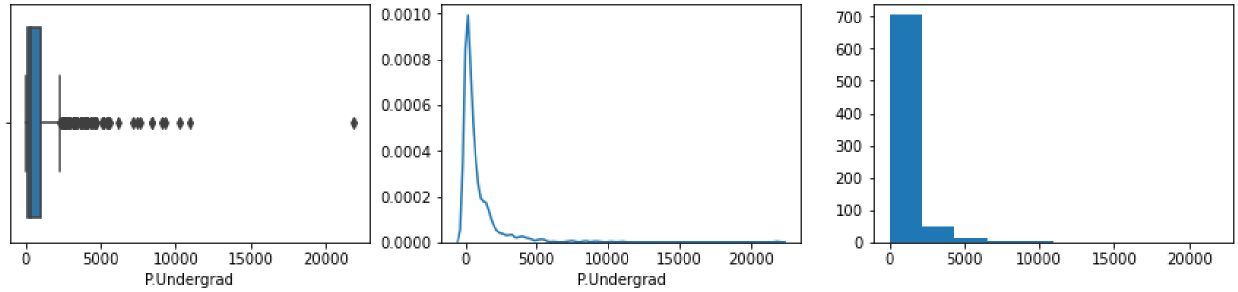
```
Column Top10perc is not normally distributed

 5. Univariate analysis for Top25perc

Mean is 55.796654, Median is 54.000000, Modes are [20]
Column Top25perc does not have outliers
```



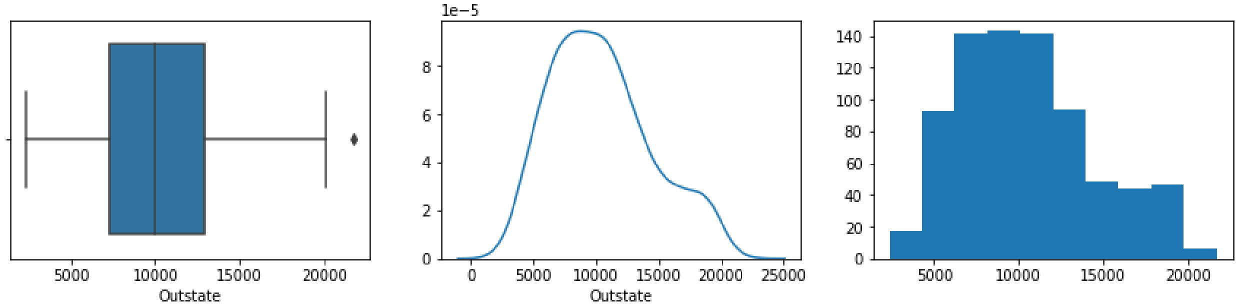
```
Column Top25perc is not normally distributed

 6. Univariate analysis for F.Undergrad

Mean is 3699.907336, Median is 1707.000000, Modes are [3]
Column F.Undergrad has outliers
```



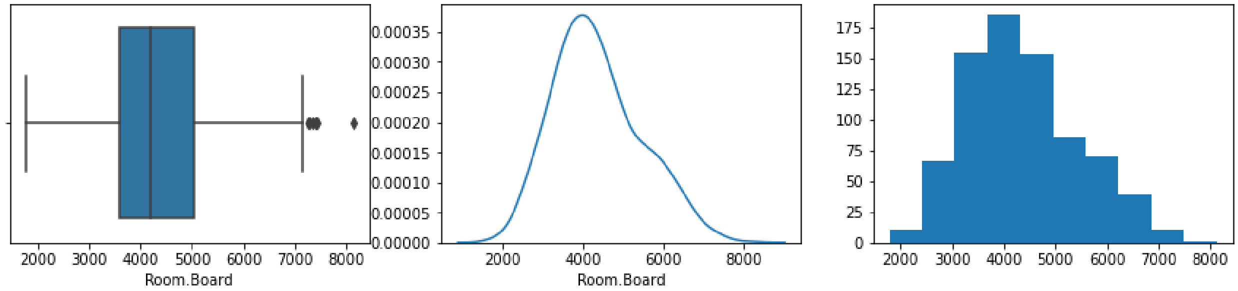
```
Column F.Undergrad is not normally distributed

 7. Univariate analysis for P.Undergrad

Mean is 855.298584, Median is 353.000000, Modes are [7]
Column P.Undergrad has outliers
```



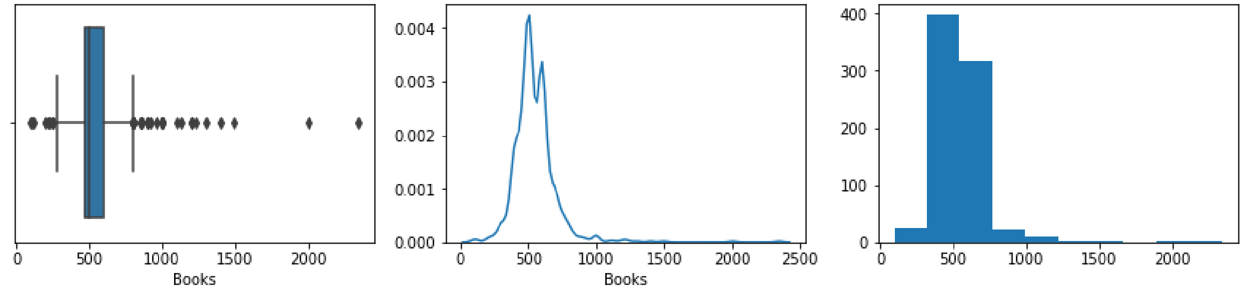
```
Column P.Undergrad is not normally distributed

 8. Univariate analysis for Outstate

Mean is 10440.669241, Median is 9990.000000, Modes are [13]
Column Outstate has outliers
```



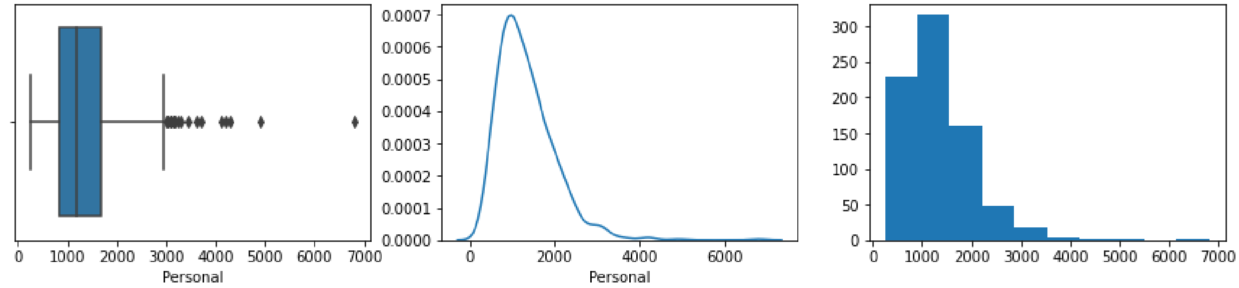
```
Column Outstate is not normally distributed

 9. Univariate analysis for Room.Board

Mean is 4357.526384, Median is 4200.000000, Modes are [9]
Column Room.Board has outliers
```



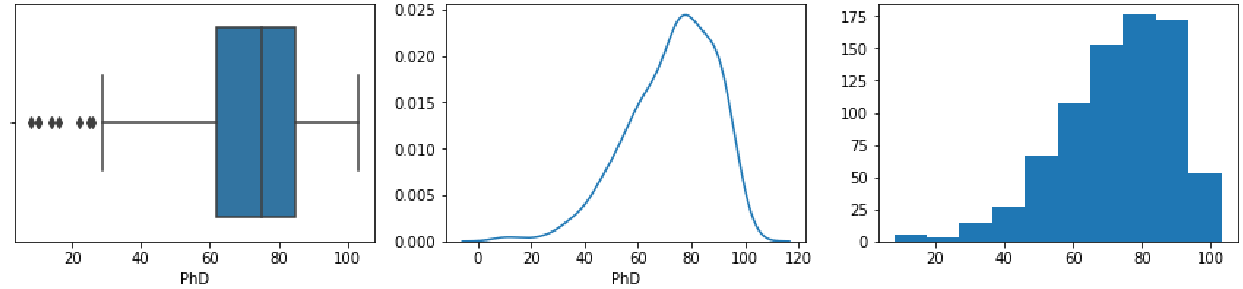
```
Column Room.Board is not normally distributed

 10. Univariate analysis for Books

Mean is 549.380952, Median is 500.000000, Modes are [178]
Column Books has outliers
```



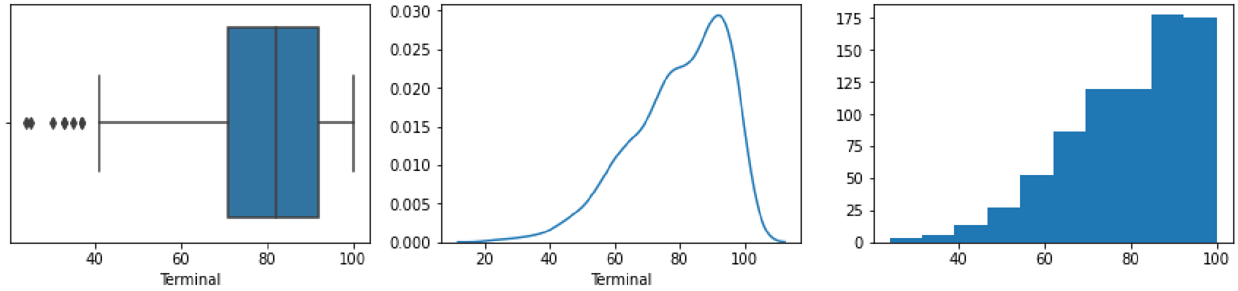
```
Column Books is not normally distributed

 11. Univariate analysis for Personal

Mean is 1340.642214, Median is 1200.000000, Modes are [45]
Column Personal has outliers
```



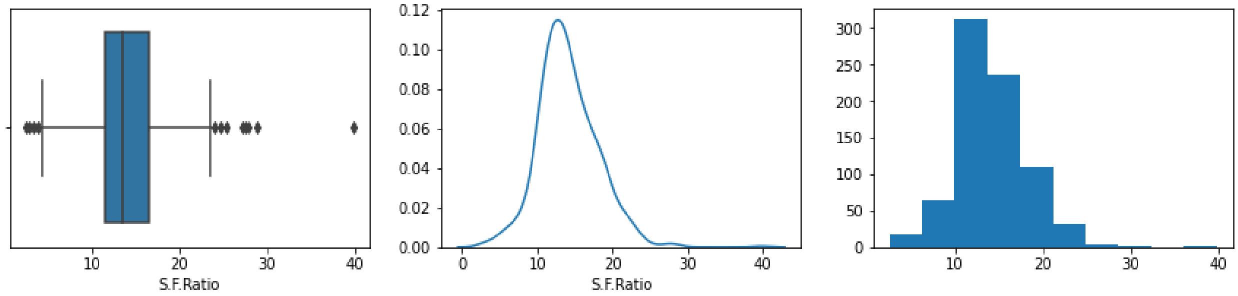
```
Column Personal is not normally distributed

 12. Univariate analysis for PhD

Mean is 72.660232, Median is 75.000000, Modes are [26]
Column PhD has outliers
```



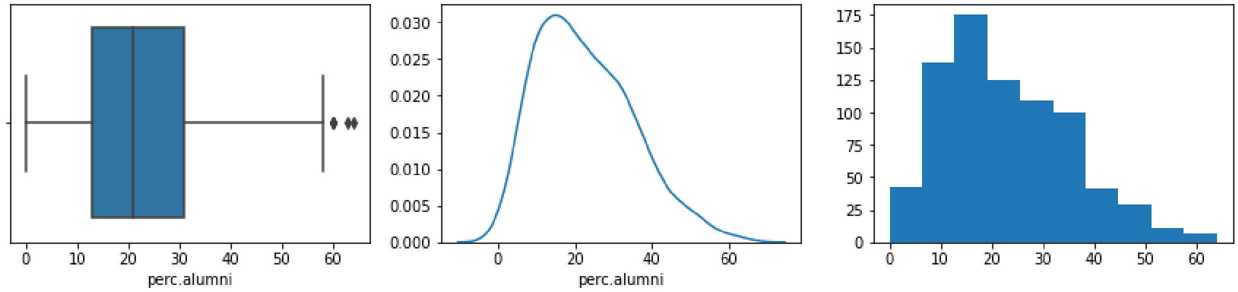
```
Column PhD is not normally distributed

 13. Univariate analysis for Terminal

Mean is 79.702703, Median is 82.000000, Modes are [30]
Column Terminal has outliers
```



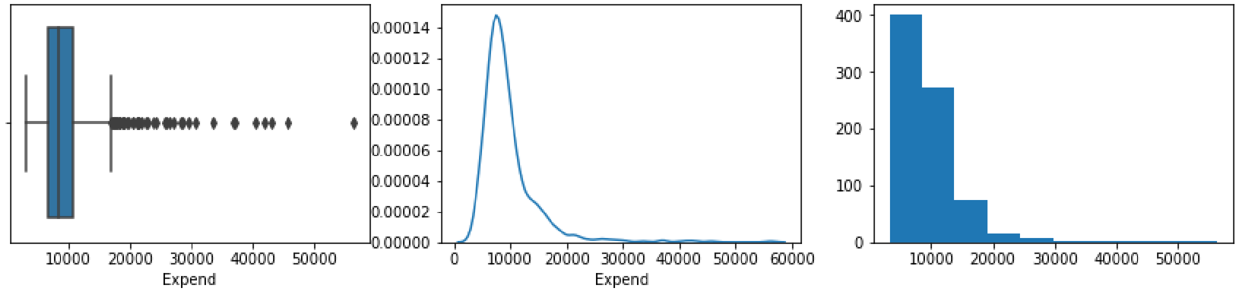
```
Column Terminal is not normally distributed

 14. Univariate analysis for S.F.Ratio

Mean is 14.089704, Median is 13.600000, Modes are [15]
Column S.F.Ratio has outliers
```



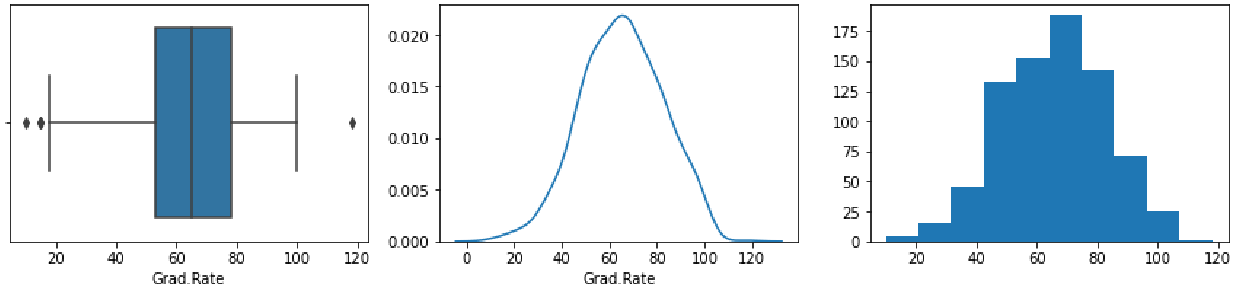
```
Column S.F.Ratio is not normally distributed

 15. Univariate analysis for perc.alumni

Mean is 22.743887, Median is 21.000000, Modes are [32]
Column perc.alumni has outliers
```



```
Column perc.alumni is not normally distributed

 16. Univariate analysis for Expend

Mean is 9660.171171, Median is 8377.000000, Modes are [2]
Column Expend has outliers
```



```
Column Expend is not normally distributed

 17. Univariate analysis for Grad.Rate

Mean is 65.463320, Median is 65.000000, Modes are [24]
Column Grad.Rate has outliers
```



```
Column Grad.Rate is not normally distributed
```

2.1 Part B: **Multi variate analysis**

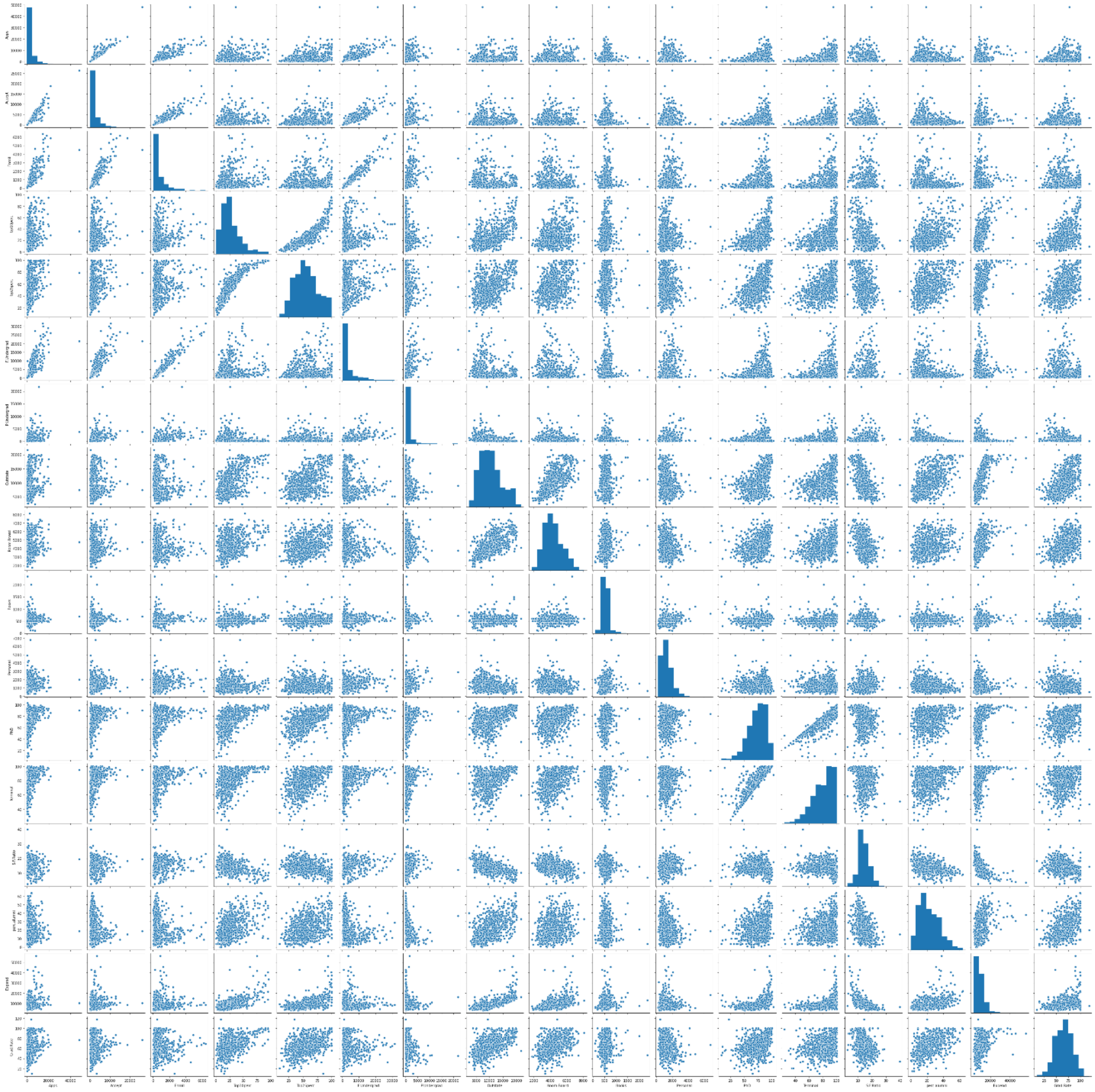
Multivariate observations based on the pair plot pasted below
While the pair plot displays the scatter plots for all combinations of pairs of columns from the original data set aiding to visually observe relative the degree of correlation between the variables, the actual magnitude of co relation can be found coefficient of co relation between the variables.

Based on the coefficient of correlation following observations are being made:

- *Highly and positively correlated:*
    - *Apps with Accept, Enroll and F_Undergrad*
    - *Accept with Enroll and F.Undergrad*
    - *Top10perc with Top25perc*
    - *Room.Board somewhat with Outstate*
    - *Terminal with PhD*

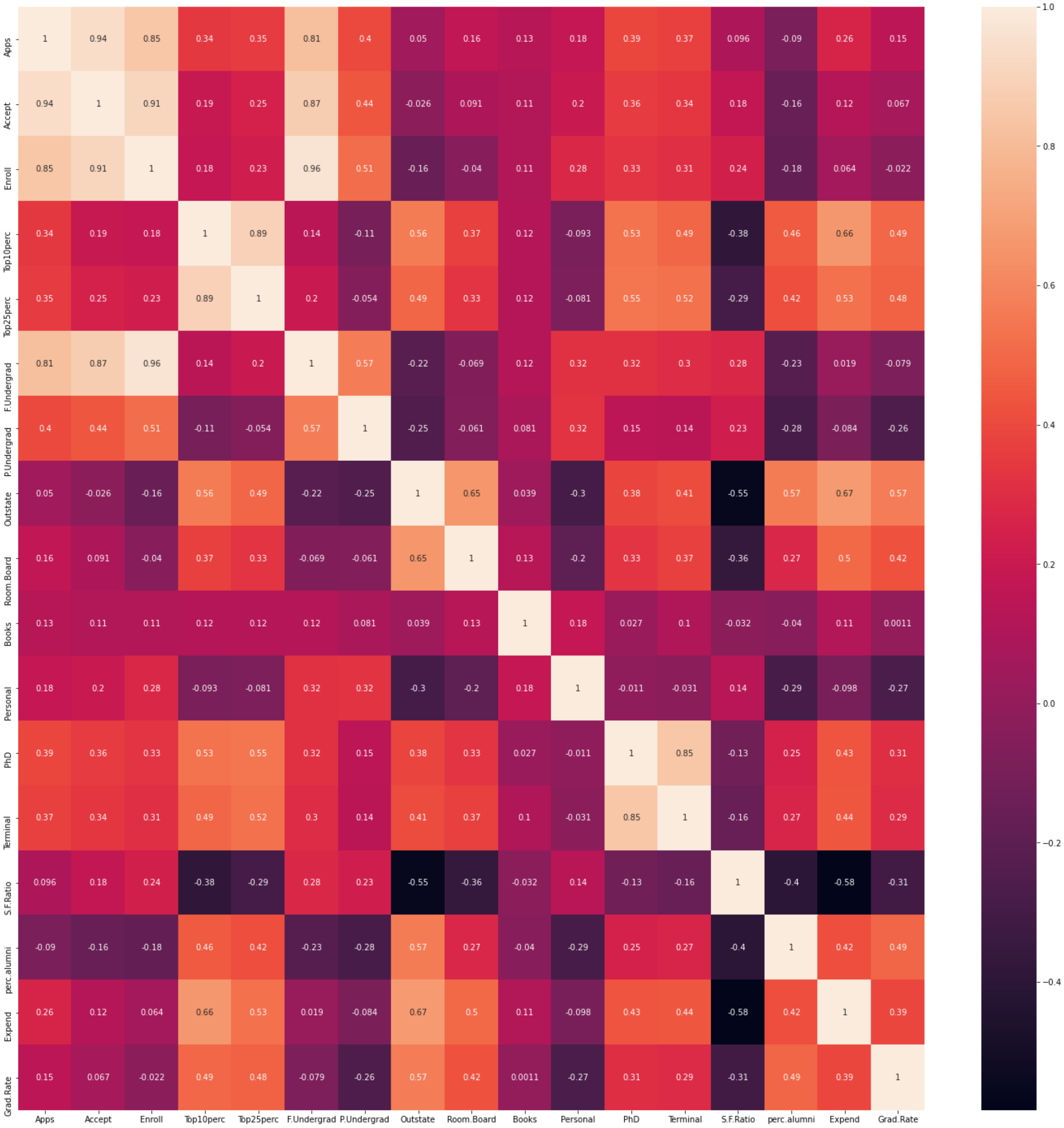
- *Low correlations with positive linearity:*
    - *Books and Personal are comparatively least corelated with every other variable.*

- *Negative correlations*
    - *S.F. Ratio has the maximum negative corelation with any variables namely Expend and Outstate*



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
20000
40000
0
10000
20000
0
2000
4000
6000
0
25
50
75
100
25
50
75
100
0
10000
20000
30000
0
5000
10000
15000
20000
5000
10000
15000
20000
2000
4000
6000
8000
0
500
1000
1500
2000
0
2000
4000
6000
25
50
75
100
40
60
80
100
10
20
30
40
0
20
40
60
20000
40000
25
50
75
100

The heatmap below reflects the observations narrated above about the correlation with dark shades indicating inverse correlation and light shades indicating positive correlation.



	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	0.94	0.85	0.34	0.35	0.81	0.4	0.05	0.16	0.13	0.18	0.39	0.37	0.096	-0.09	0.26	0.15
Accept	0.94	1	0.91	0.19	0.25	0.87	0.44	-0.026	0.091	0.11	0.2	0.36	0.34	0.18	-0.16	0.12	0.067
Enroll	0.85	0.91	1	0.18	0.23	0.96	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	-0.18	0.064	-0.022
Top10perc	0.34	0.19	0.18	1	0.89	0.14	-0.11	0.56	0.37	0.12	-0.093	0.53	0.49	-0.38	0.46	0.66	0.49
Top25perc	0.35	0.25	0.23	0.89	1	0.2	-0.054	0.49	0.33	0.12	-0.081	0.55	0.52	-0.29	0.42	0.53	0.48
F.Undergrad	0.81	0.87	0.96	0.14	0.2	1	0.57	-0.22	-0.069	0.12	0.32	0.32	0.3	0.28	-0.23	0.019	-0.079
P.Undergrad	0.4	0.44	0.51	-0.11	-0.054	0.57	1	-0.25	-0.061	0.081	0.32	0.15	0.14	0.23	-0.28	-0.084	-0.26
Outstate	0.05	-0.026	-0.16	0.56	0.49	-0.22	-0.25	1	0.65	0.039	-0.3	0.38	0.41	-0.55	0.57	0.67	0.57
Room.Board	0.16	0.091	-0.04	0.37	0.33	-0.069	-0.061	0.65	1	0.13	-0.2	0.33	0.37	-0.36	0.27	0.5	0.42
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.081	0.039	0.13	1	0.18	0.027	0.1	-0.032	-0.04	0.11	0.0011
Personal	0.18	0.2	0.28	-0.093	-0.081	0.32	0.32	-0.3	-0.2	0.18	1	-0.011	-0.031	0.14	-0.29	-0.098	-0.27
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.027	-0.011	1	0.85	-0.13	0.25	0.43	0.31
Terminal	0.37	0.34	0.31	0.49	0.52	0.3	0.14	0.41	0.37	0.1	-0.031	0.85	1	-0.16	0.27	0.44	0.29
S.F.Ratio	0.096	0.18	0.24	-0.38	-0.29	0.28	0.23	-0.55	-0.36	-0.032	0.14	-0.13	-0.16	1	-0.4	-0.58	-0.31
perc.alumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.27	-0.04	-0.29	0.25	0.27	-0.4	1	0.42	0.49
Expend	0.26	0.12	0.064	0.66	0.53	0.019	-0.084	0.67	0.5	0.11	-0.098	0.43	0.44	-0.58	0.42	1	0.39
Grad.Rate	0.15	0.067	-0.022	0.49	0.48	-0.079	-0.26	0.57	0.42	0.0011	-0.27	0.31	0.29	-0.31	0.49	0.39	1

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

Inference based on observations for scaling across various scaling options.

- *Scaling options considered: Standardization, Normalization and Logarithmic transformation*

Observations:

1. Given the significant outliers in the original data continues to be retained post to the scaling using Minmaxscaler and standardscaler these methods does not seem to scale the dimension efficiently. Also if the original dataset has good amount of skewness it needs to be scaled first before using standardscaler.

2. Logarithmic scaler results in decreased skewness of the dataset along with outliers ending up imputed within either side of the whiskers for the good part of it. Almost 66% of the outliers has been reduced as a result of logarithmic scaling.

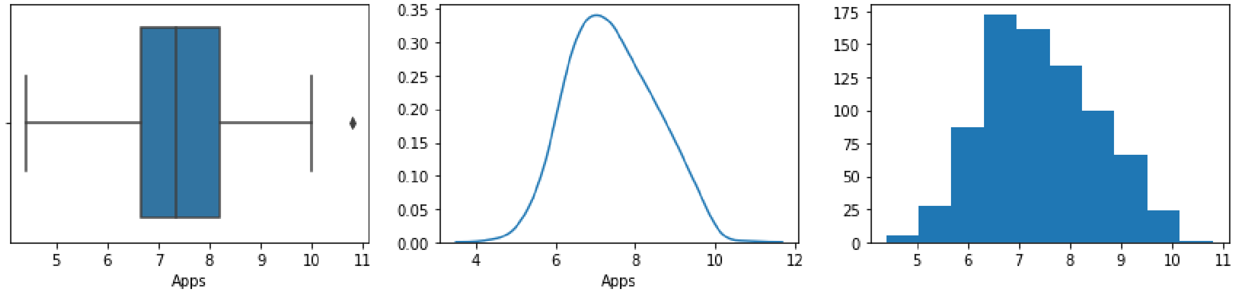
Scaling option finalized accordingly will be logarithmic transformation.

```
Univariate analysis after logarithmic scaling of the original dataset
----------------------------------------------------------------------
```

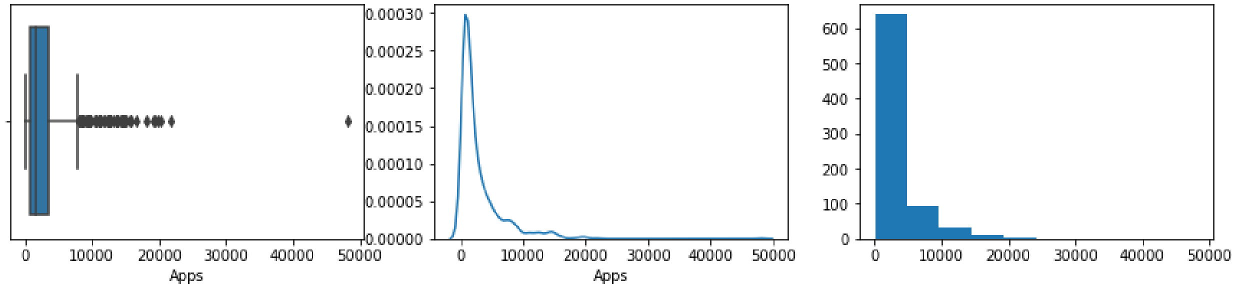
Note:

- Below analysis is done once for scaled dataset and original dataset for each variable
- Same serial numbers back to back depict comparison of same variable between scaled dataset and original in that order with original dataset being called out for clarity.

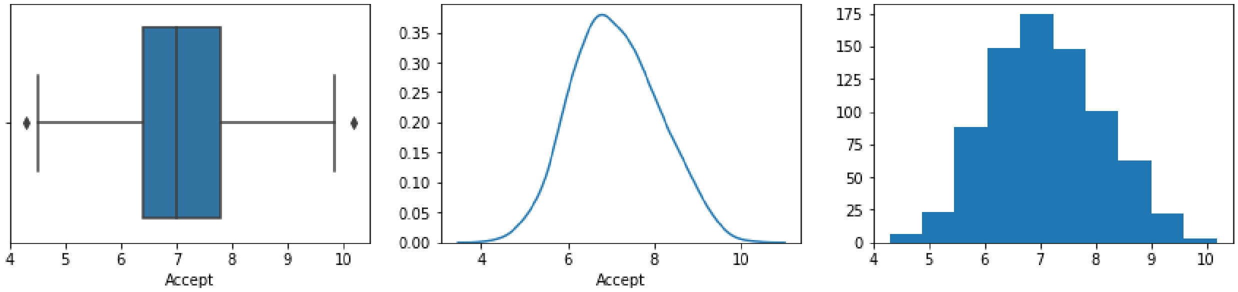
```
 1. Univariate analysis for Apps

Mean is 7.427593, Median is 7.351800, Modes are [3]
Column Apps has outliers
```



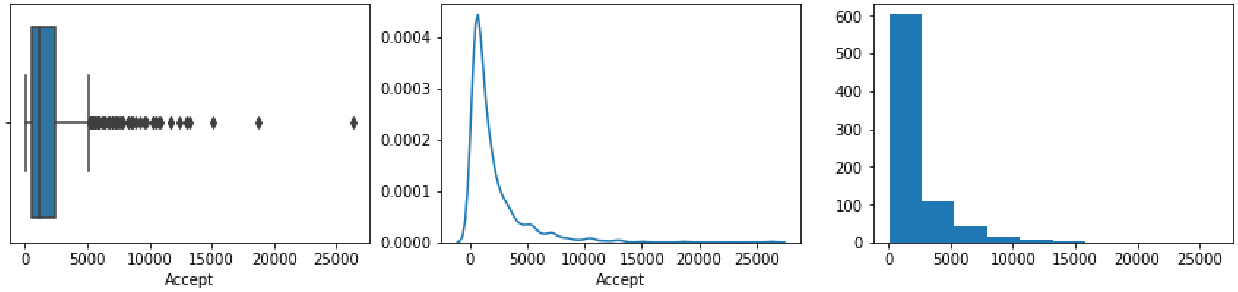
```
Column Apps is not normally distributed

 1. Original dataset: Univariate analysis for Apps

Mean is 3001.638353, Median is 1558.000000, Modes are [3]
Column Apps has outliers
```



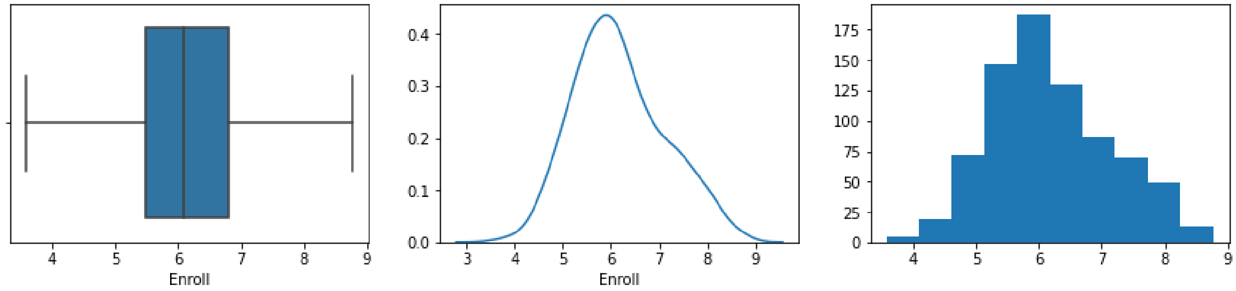
```
Column Apps is not normally distributed

 2. Univariate analysis for Accept

Mean is 7.110960, Median is 7.013016, Modes are [4]
Column Accept has outliers
```



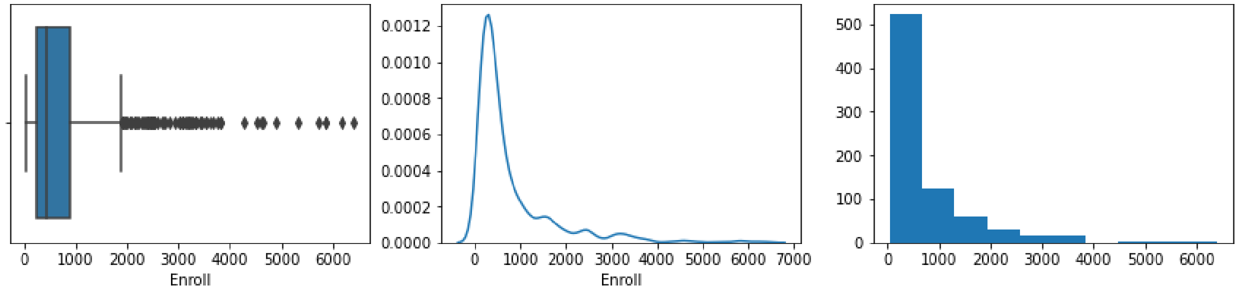
```
Column Accept is not normally distributed

 2. Original dataset: Univariate analysis for Accept

Mean is 2018.804376, Median is 1110.000000, Modes are [4]
Column Accept has outliers
```



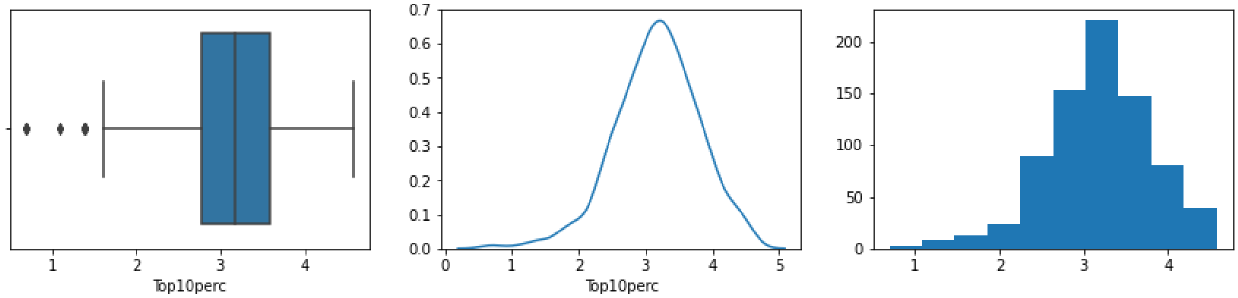
```
Column Accept is not normally distributed

 3. Univariate analysis for Enroll

Mean is 6.176126, Median is 6.075346, Modes are [5]
Column Enroll does not have outliers
```



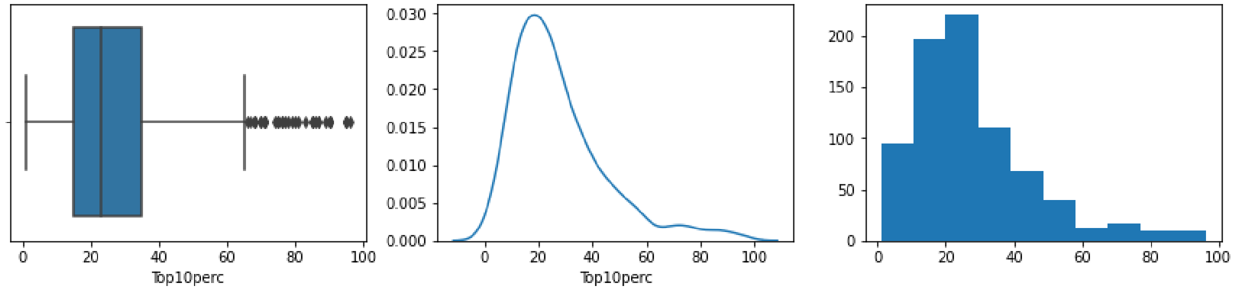
```
Column Enroll is not normally distributed

 3. Original dataset: Univariate analysis for Enroll

Mean is 779.972973, Median is 434.000000, Modes are [5]
Column Enroll has outliers
```



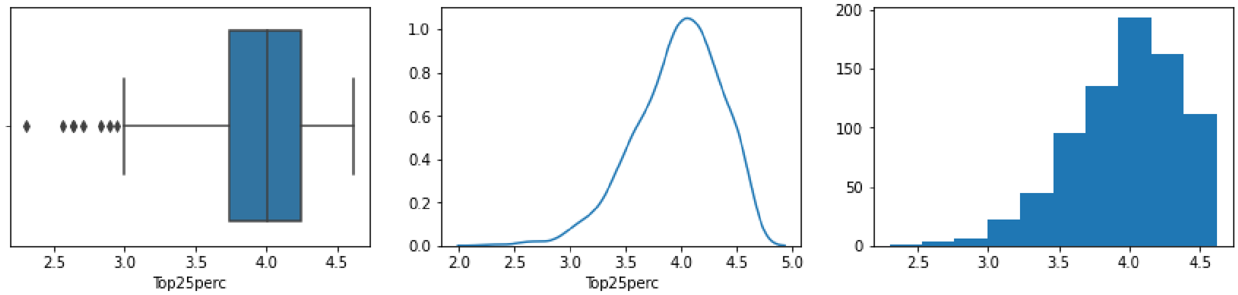
```
Column Enroll is not normally distributed

 4. Univariate analysis for Top10perc

Mean is 3.169247, Median is 3.178054, Modes are [37]
Column Top10perc has outliers
```



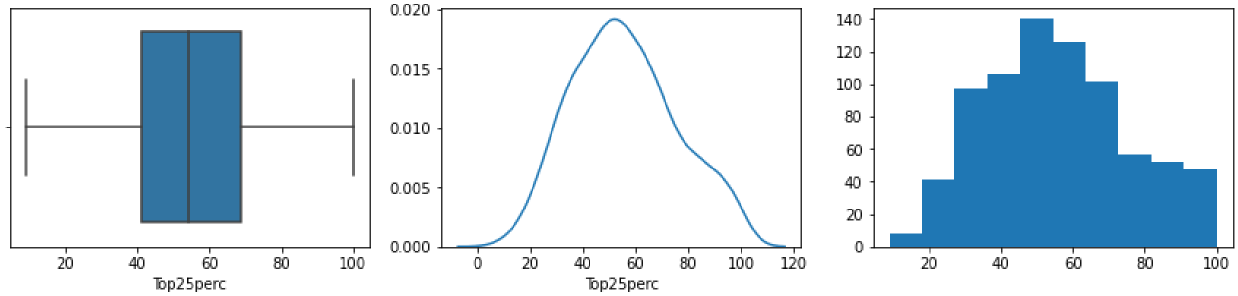
```
Column Top10perc is not normally distributed

 4. Original dataset: Univariate analysis for Top10perc

Mean is 27.558559, Median is 23.000000, Modes are [37]
Column Top10perc has outliers
```



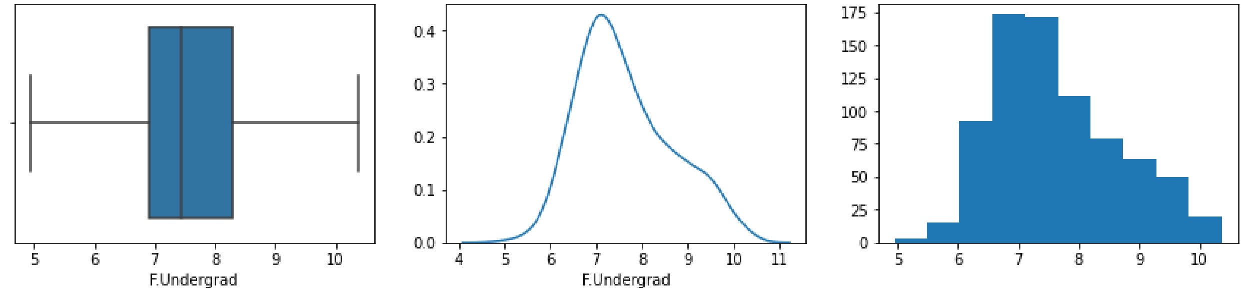
```
Column Top10perc is not normally distributed

 5. Univariate analysis for Top25perc

Mean is 3.972021, Median is 4.007333, Modes are [20]
Column Top25perc has outliers
```



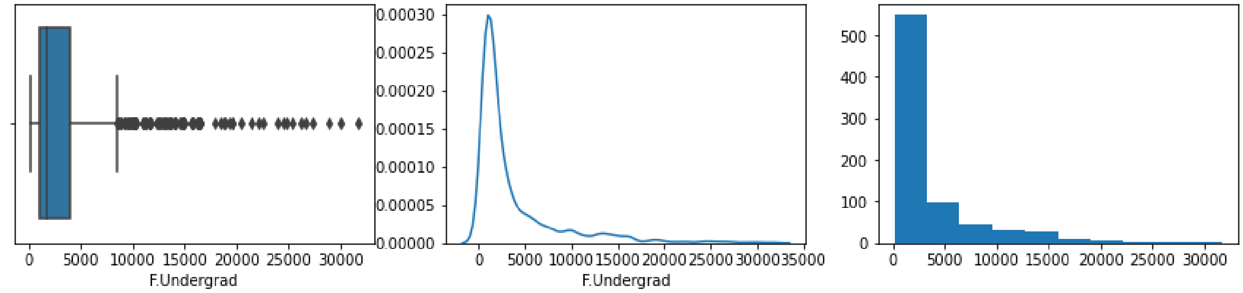
```
Column Top25perc is not normally distributed

 5. Original dataset: Univariate analysis for Top25perc

Mean is 55.796654, Median is 54.000000, Modes are [20]
Column Top25perc does not have outliers
```



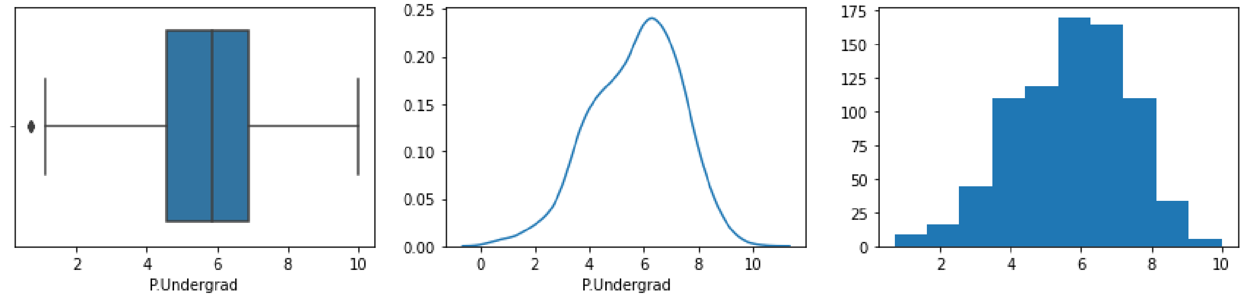
```
Column Top25perc is not normally distributed

 6. Univariate analysis for F.Undergrad

Mean is 7.635932, Median is 7.443078, Modes are [3]
Column F.Undergrad does not have outliers
```



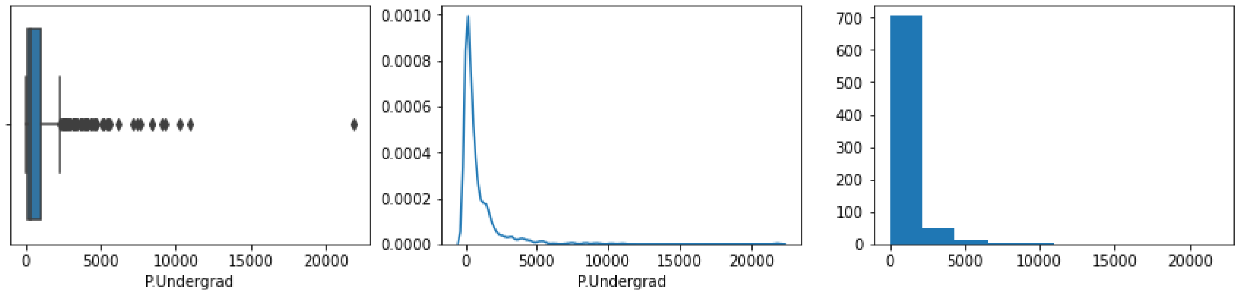
```
Column F.Undergrad is not normally distributed

 6. Original dataset: Univariate analysis for F.Undergrad

Mean is 3699.907336, Median is 1707.000000, Modes are [3]
Column F.Undergrad has outliers
```



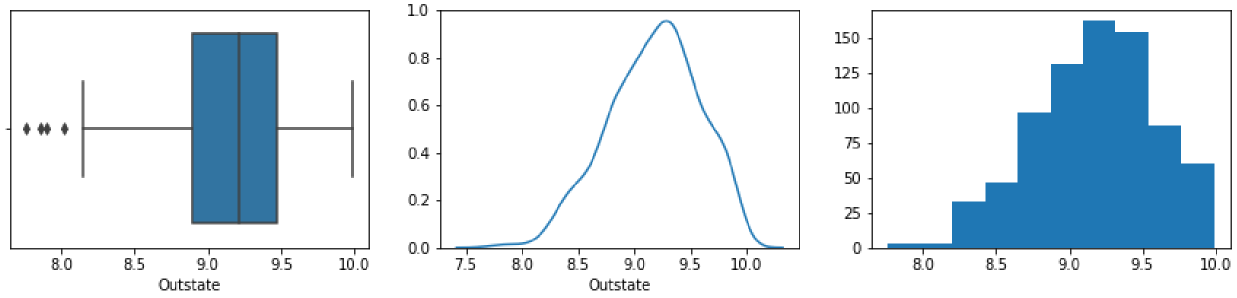
```
Column F.Undergrad is not normally distributed

 7. Univariate analysis for P.Undergrad

Mean is 5.706966, Median is 5.869297, Modes are [7]
Column P.Undergrad has outliers
```



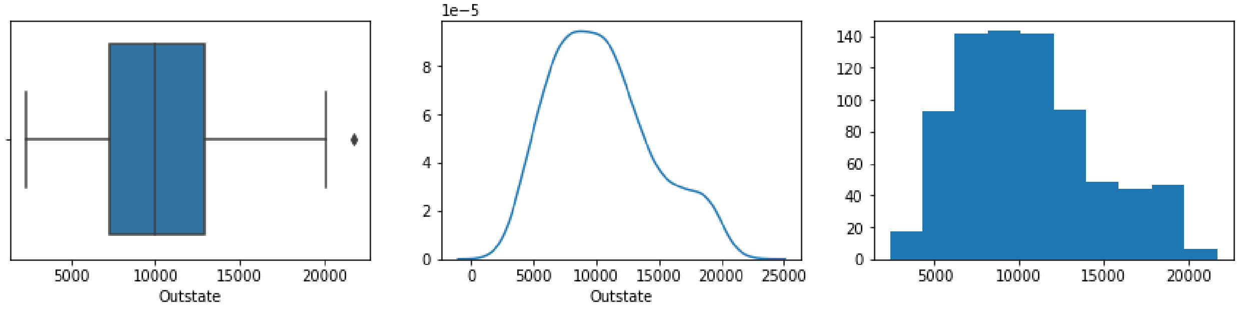
```
Column P.Undergrad is not normally distributed

 7. Original dataset: Univariate analysis for P.Undergrad

Mean is 855.298584, Median is 353.000000, Modes are [7]
Column P.Undergrad has outliers
```



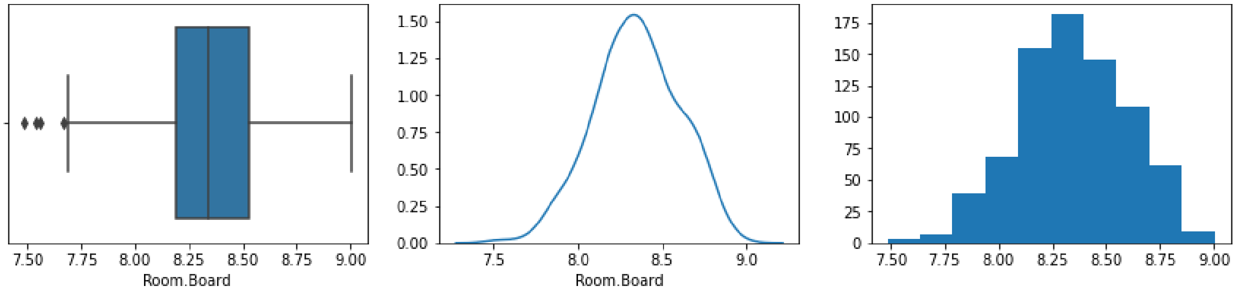
```
Column P.Undergrad is not normally distributed

 8. Univariate analysis for Outstate

Mean is 9.175706, Median is 9.209440, Modes are [13]
Column Outstate has outliers
```



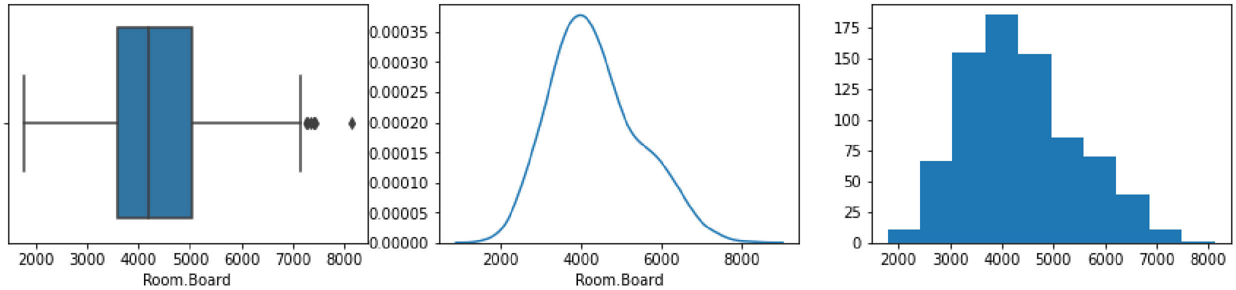
```
Column Outstate is not normally distributed

 8. Original dataset: Univariate analysis for Outstate

Mean is 10440.669241, Median is 9990.000000, Modes are [13]
Column Outstate has outliers
```



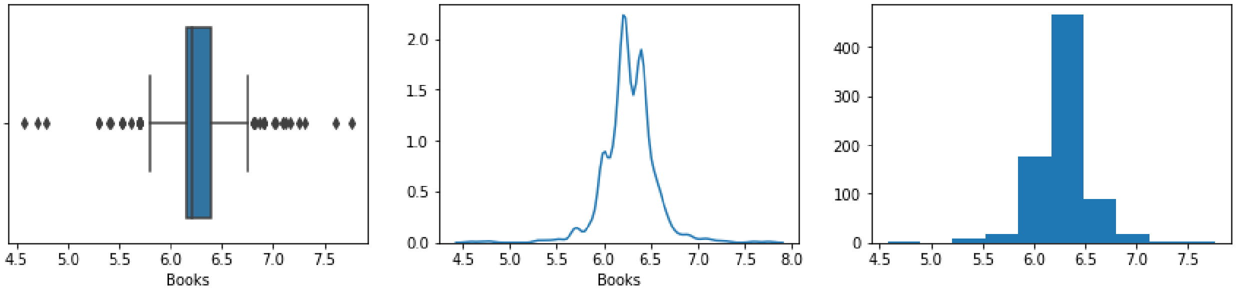
```
Column Outstate is not normally distributed

 9. Univariate analysis for Room.Board

Mean is 8.348207, Median is 8.343078, Modes are [9]
Column Room.Board has outliers
```



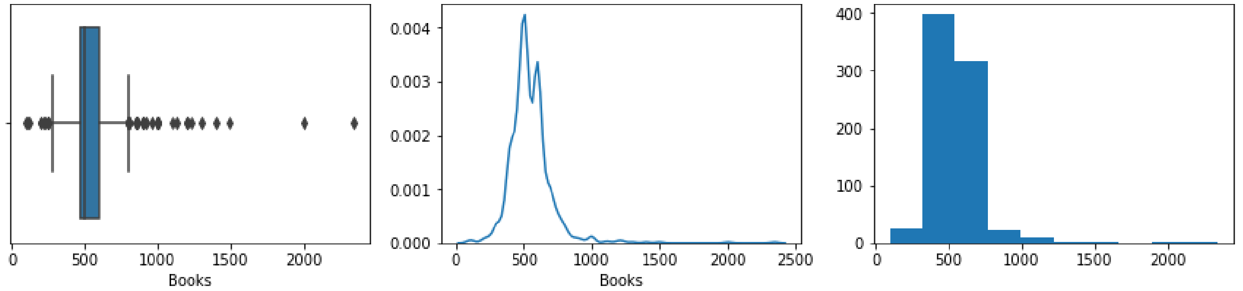
```
Column Room.Board is not normally distributed

 9. Original dataset: Univariate analysis for Room.Board

Mean is 4357.526384, Median is 4200.000000, Modes are [9]
Column Room.Board has outliers
```



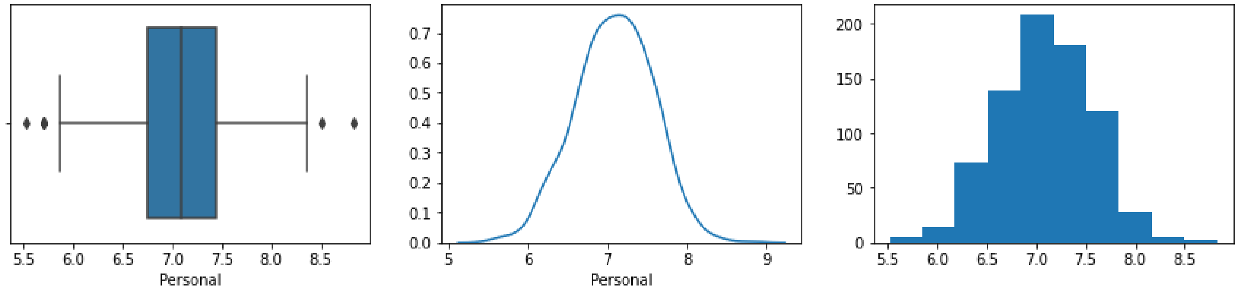
```
Column Room.Board is not normally distributed

 10. Univariate analysis for Books

Mean is 6.273912, Median is 6.216606, Modes are [178]
Column Books has outliers
```



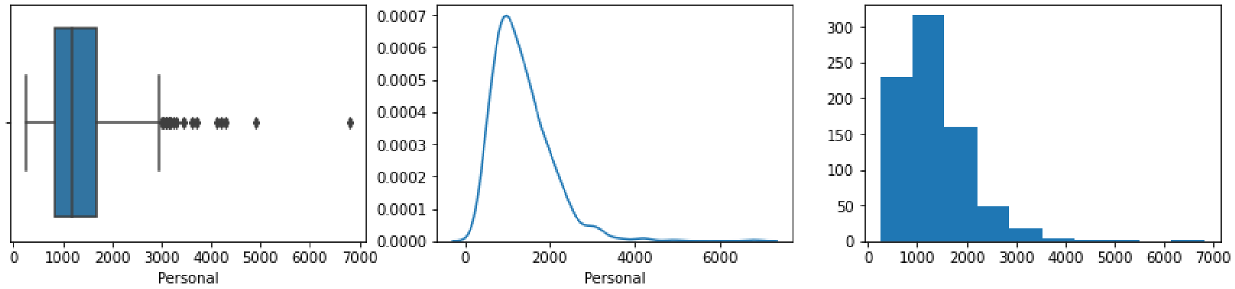
```
Column Books is not normally distributed

 10. Original dataset: Univariate analysis for Books

Mean is 549.380952, Median is 500.000000, Modes are [178]
Column Books has outliers
```



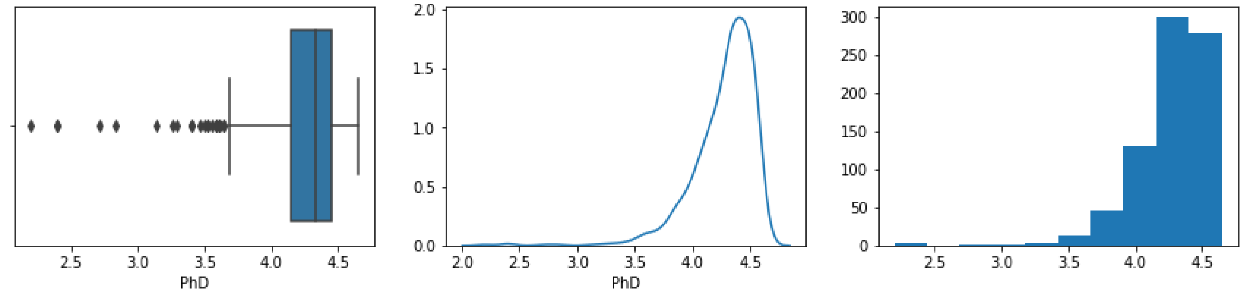
```
Column Books is not normally distributed

 11. Univariate analysis for Personal

Mean is 7.086012, Median is 7.090910, Modes are [45]
Column Personal has outliers
```



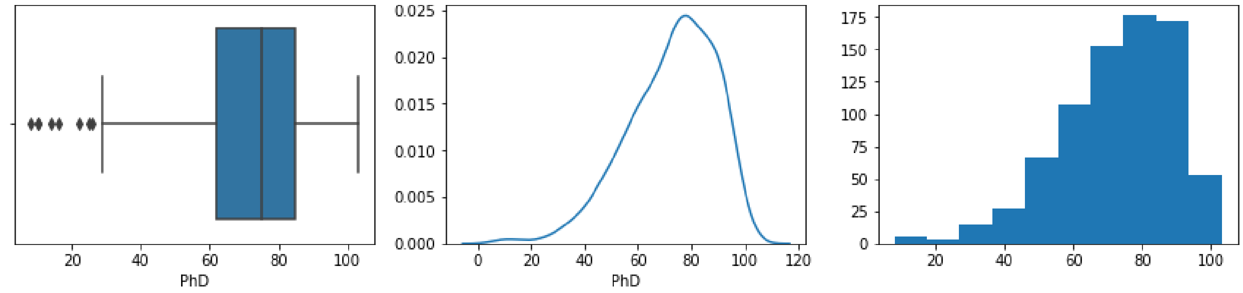
```
Column Personal is normally distributed

 11. Original dataset: Univariate analysis for Personal

Mean is 1340.642214, Median is 1200.000000, Modes are [45]
Column Personal has outliers
```



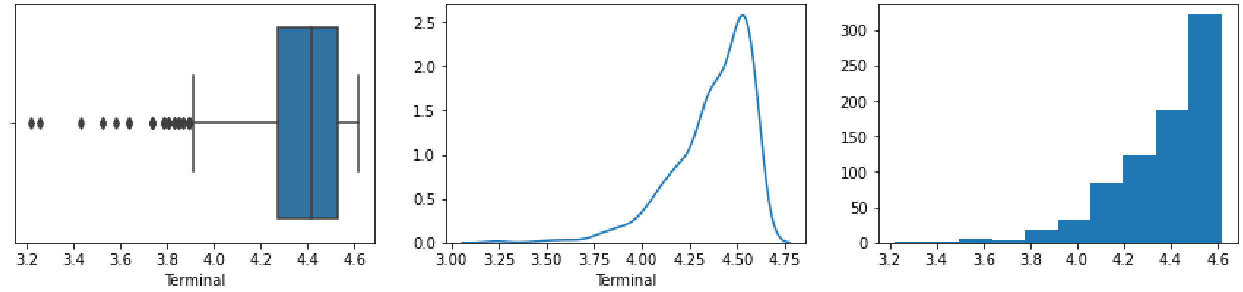
```
Column Personal is not normally distributed

 12. Univariate analysis for PhD

Mean is 4.267357, Median is 4.330733, Modes are [26]
Column PhD has outliers
```



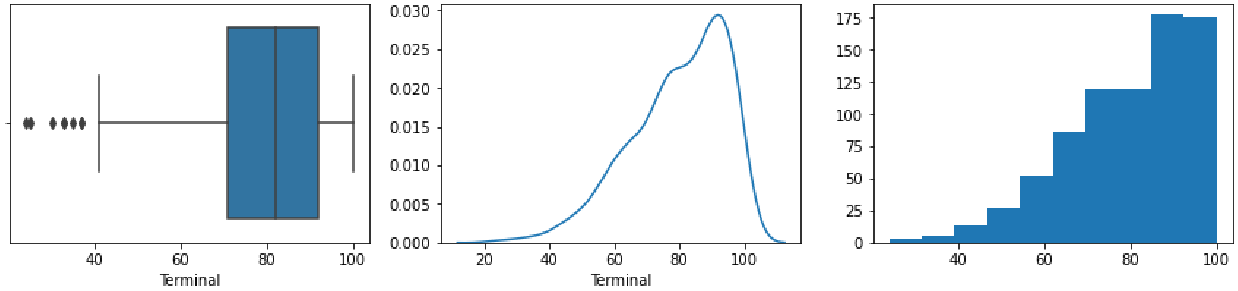
```
Column PhD is not normally distributed

 12. Original dataset: Univariate analysis for PhD

Mean is 72.660232, Median is 75.000000, Modes are [26]
Column PhD has outliers
```



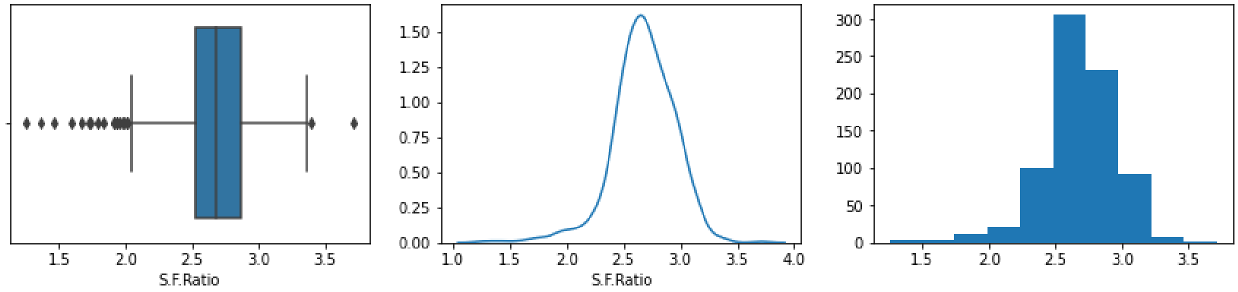
```
Column PhD is not normally distributed

 13. Univariate analysis for Terminal

Mean is 4.371075, Median is 4.418841, Modes are [30]
Column Terminal has outliers
```



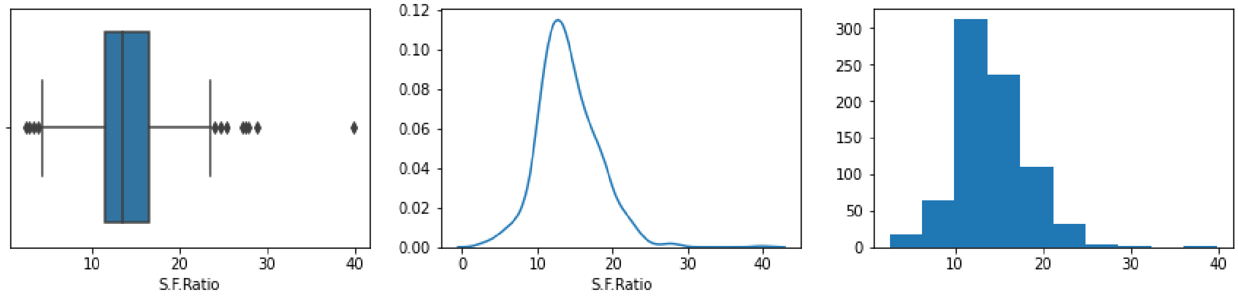
```
Column Terminal is not normally distributed

 13. Original dataset: Univariate analysis for Terminal

Mean is 79.702703, Median is 82.000000, Modes are [30]
Column Terminal has outliers
```



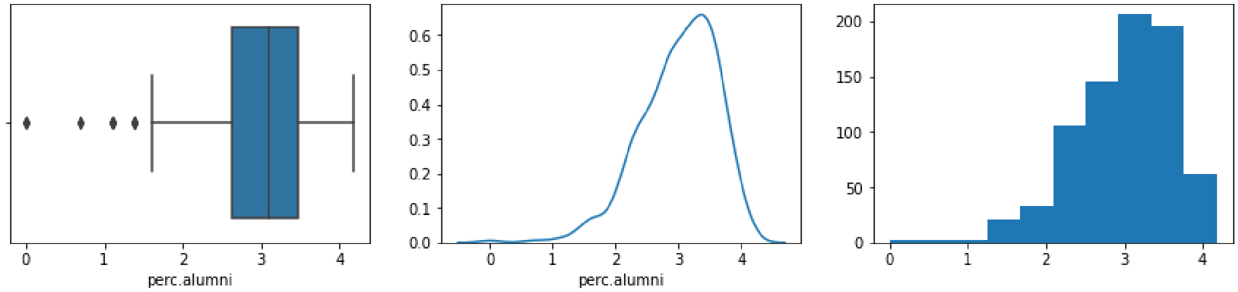
```
Column Terminal is not normally distributed

 14. Univariate analysis for S.F.Ratio

Mean is 2.678210, Median is 2.681022, Modes are [15]
Column S.F.Ratio has outliers
```



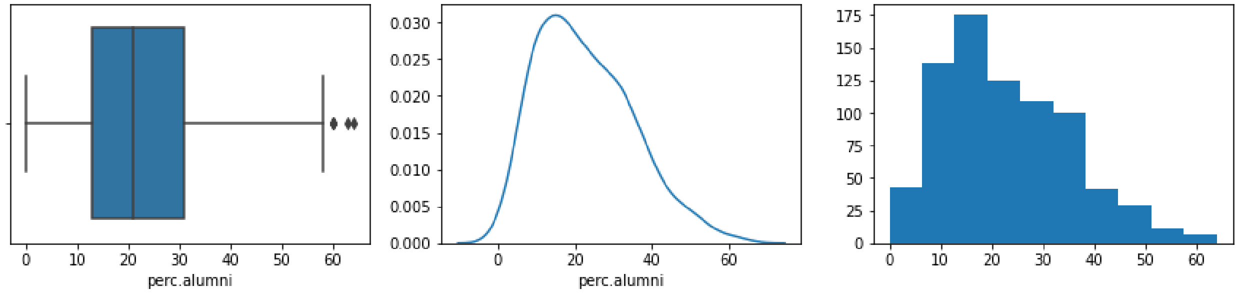
```
Column S.F.Ratio is not normally distributed

 14. Original dataset: Univariate analysis for S.F.Ratio

Mean is 14.089704, Median is 13.600000, Modes are [15]
Column S.F.Ratio has outliers
```



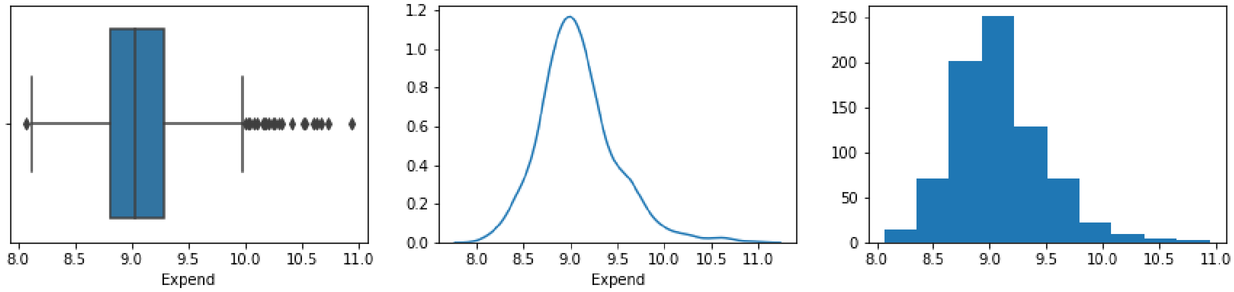
```
Column S.F.Ratio is not normally distributed

 15. Univariate analysis for perc.alumni

Mean is 3.007372, Median is 3.091042, Modes are [32]
Column perc.alumni has outliers
```



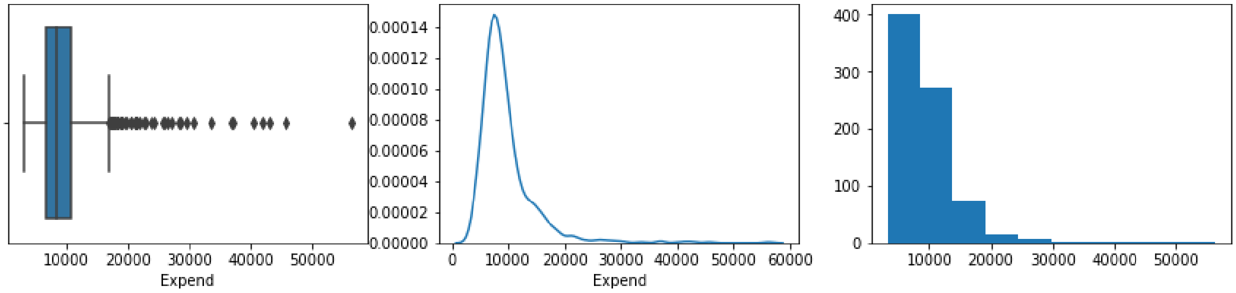
```
Column perc.alumni is not normally distributed

 15. Original dataset: Univariate analysis for perc.alumni

Mean is 22.743887, Median is 21.000000, Modes are [32]
Column perc.alumni has outliers
```



```
Column perc.alumni is not normally distributed

 16. Univariate analysis for Expend

Mean is 9.080763, Median is 9.033365, Modes are [2]
Column Expend has outliers
```



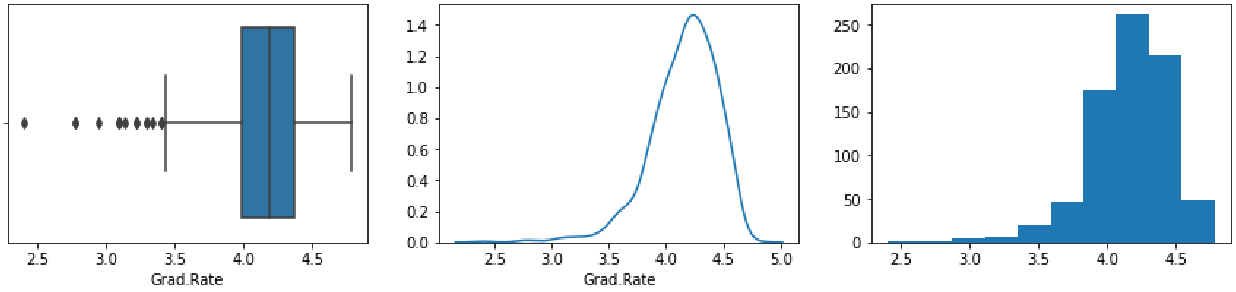
```
Column Expend is not normally distributed

 16. Original dataset: Univariate analysis for Expend

Mean is 9660.171171, Median is 8377.000000, Modes are [2]
Column Expend has outliers
```



```
Column Expend is not normally distributed
```

17. Univariate analysis for Grad.Rate

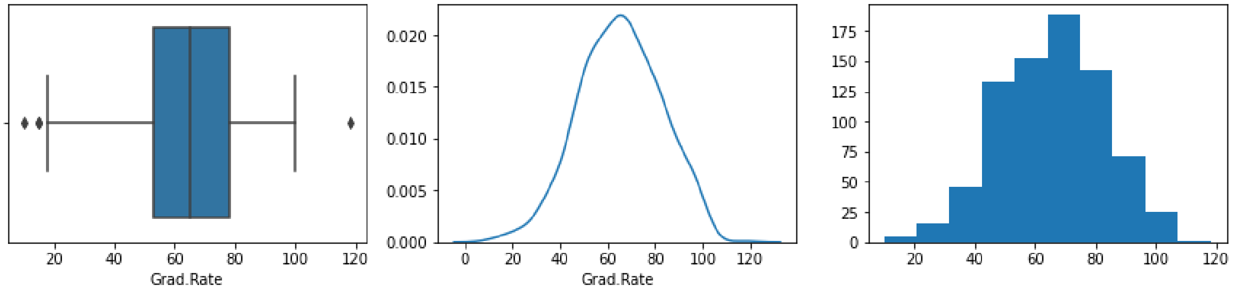
```
Mean is 4.158004, Median is 4.189655, Modes are [24]
Column Grad.Rate has outliers
```



```
Column Grad.Rate is not normally distributed
```

17. Original dataset: Univariate analysis for Grad.Rate

```
Mean is 65.463320, Median is 65.000000, Modes are [24]
Column Grad.Rate has outliers
```



```
Column Grad.Rate is not normally distributed
```

Outlier observations:

- There were 584 outliers across variables in the original unscaled data set
- There are 210 outliers across variables in the scaled dataset

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

Comparison between Covariance and Correlation:

1.Covariance is the variance measured among the dimension which denotes the direction of relationship between two independent variables. Covariance expresses a dimensions variance with itself as well as with other dimensions in the form of matrix. However when it comes to correlation apart from just the direction of relationship it also denotes the measure of strength of relationship between two independent variables.

2.Covariance are influenced by unit of variables in the original dataset while correlation which is derived by dividing product of variance between two independent variables with product of their standard deviation has values standardized between -1 and +1 which will be its range.

3.Hence coreleation is a unit free standardized measure with a known range of values while covariance can range between infinite values on either side of zero.

4.Due to robustness of Corelation measure between two variables it is more preferred than covariance for analysis.

5.On the other hand, covariance matrix of standardized variable is actually a correlation matrix

6.The sign of elements of covariance matrix and the equivalent co relation matrix is identical.corelation is a function of the covariance.

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

1. Variables other than Books, Expend, Top25perc, Top10perc, PhD, Terminal and S.F.Ration are significantly treated for their outliers automatically post scaling.

2. Out of those variables Enroll and F.Undergrad are completely treated for outliers.

3. Few observations on those variables that still retain outliers or introduced new outliers as below.

- *Books, Expend, continue to have good number of outliers despite scaling however they are equally spread across either side of whiskers compared just right side of the whisker in the original dataset.*
- *Top25perc has introduced new outliers while original dataset did not have any outliers*
- *Top10perc, P.Undergrad, Outstate, Room.Board, perc.alumni had outliers beyond maximum whisker in the original dataset. While scaled dataset does not have any of those outliers it has introduced few outliers to the left of minimum whisker.*
- *Personal variable got most of its outlier right of whisker fixed but couple of outliers introduced to the left of whisker.*
- *PhD, Terminal, S.F.Ratio and Grad.Rate ended up with more outliers to the left of whisker than the original dataset.*

4. Personal variable is normally distributed post scaling

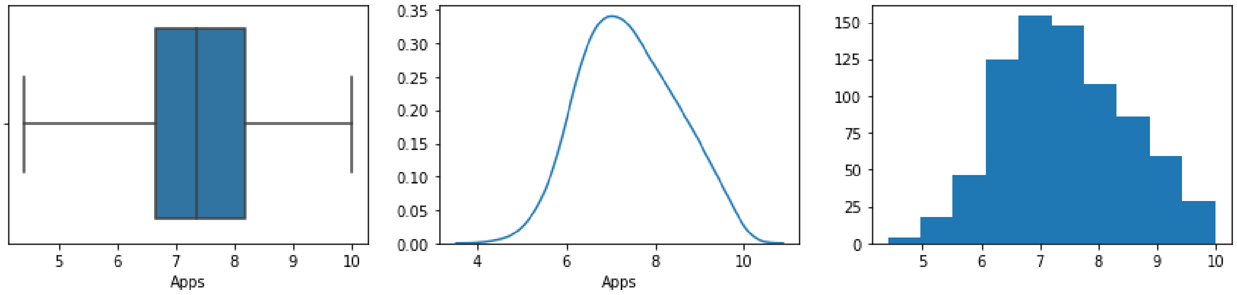
Please find below the analysis done on each of the variables post scaling.

Note:

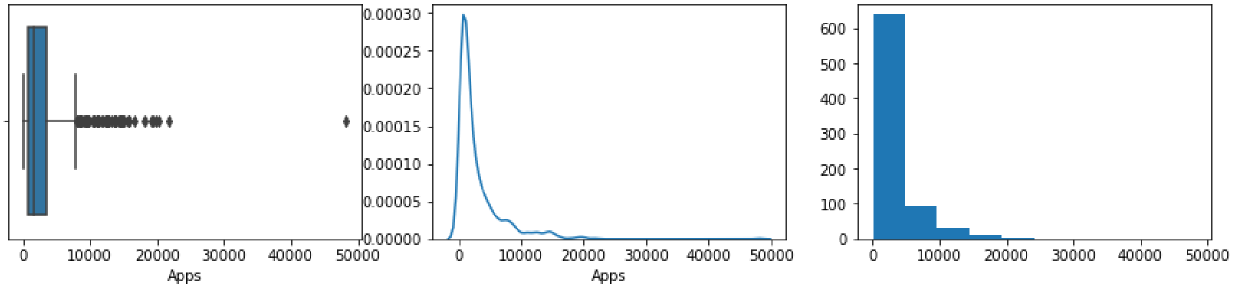
- Below analysis is done once for outlier imputed dataset and original dataset for each variable
- Same serial numbers back to back depict comparison of same variable between imputed dataset and original in that order with original dataset being called out for clarity.

```
1. Univariate analysis for Apps
```

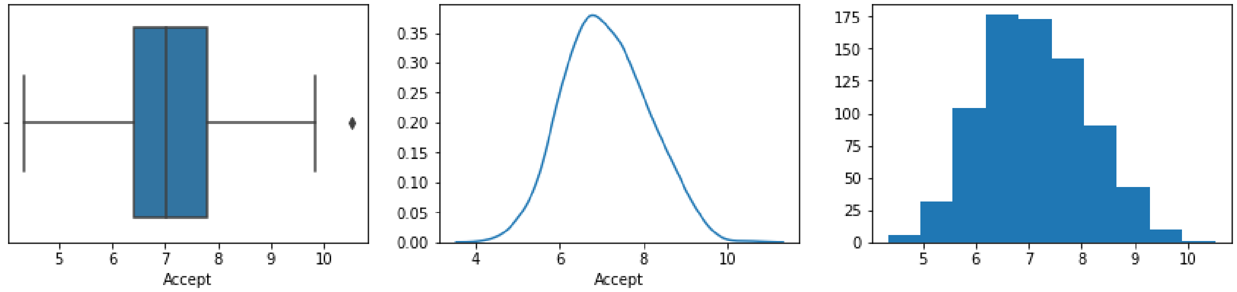
```
Mean is 7.420076, Median is 7.351158, Modes are [3]
Column Apps does not have outliers
```



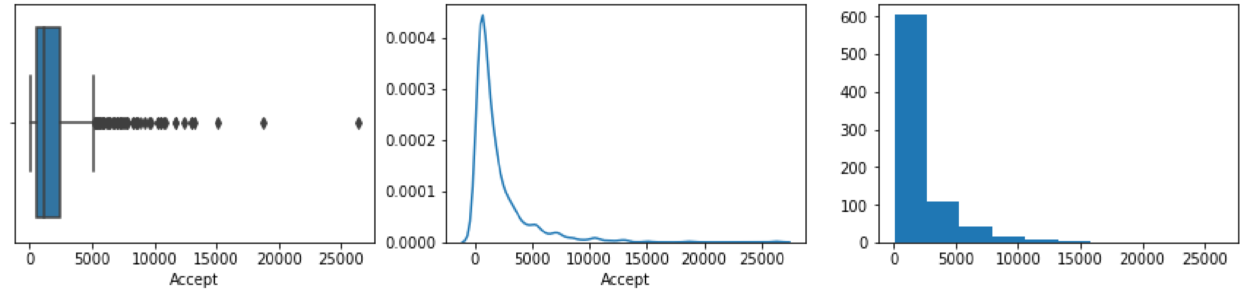
```
Column Apps is not normally distributed

 1. Original dataset: Univariate analysis for Apps

Mean is 3001.638353, Median is 1558.000000, Modes are [3]
Column Apps has outliers
```



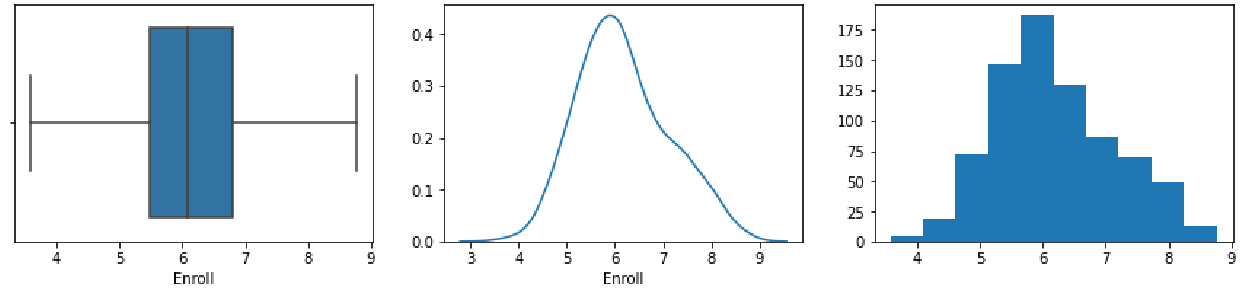
```
Column Apps is not normally distributed

 2. Univariate analysis for Accept

Mean is 7.111451, Median is 7.013016, Modes are [4]
Column Accept has outliers
```



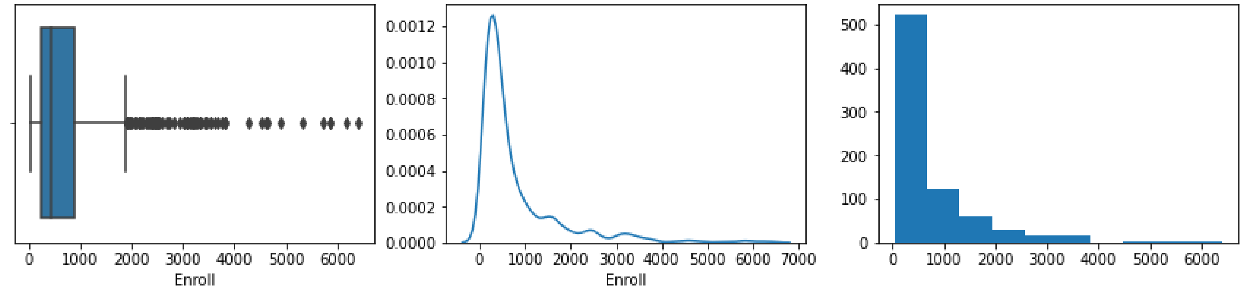
```
Column Accept is not normally distributed

 2. Original dataset: Univariate analysis for Accept

Mean is 2018.804376, Median is 1110.000000, Modes are [4]
Column Accept has outliers
```



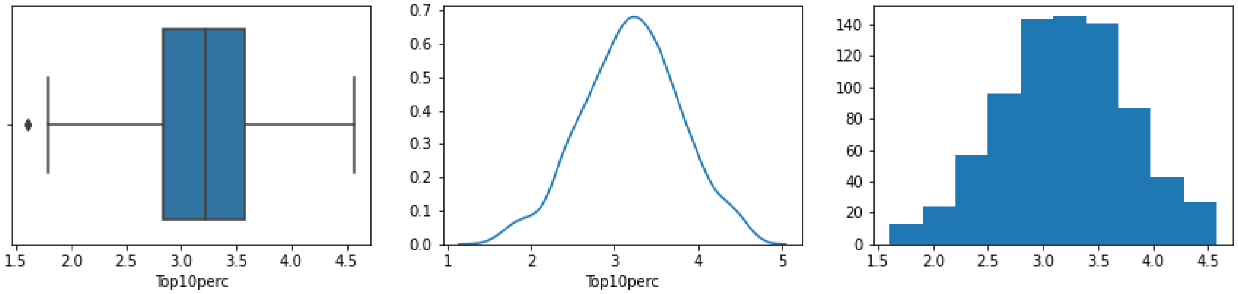
```
Column Accept is not normally distributed

 3. Univariate analysis for Enroll

Mean is 6.176126, Median is 6.075346, Modes are [5]
Column Enroll does not have outliers
```



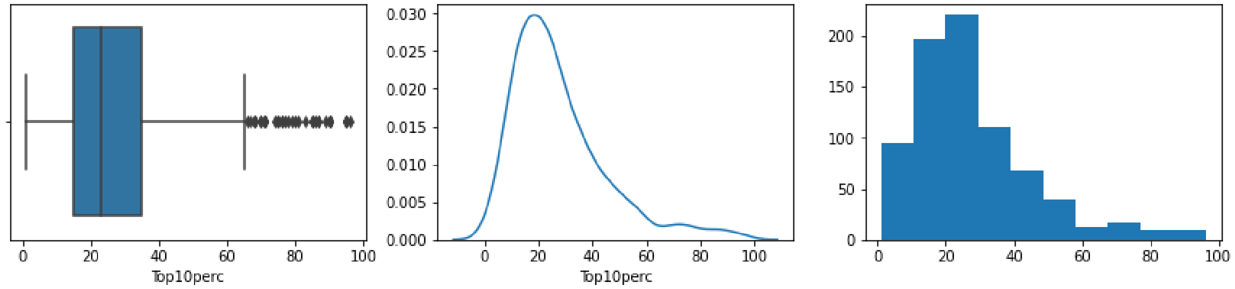
```
Column Enroll is not normally distributed

 3. Original dataset: Univariate analysis for Enroll

Mean is 779.972973, Median is 434.000000, Modes are [5]
Column Enroll has outliers
```



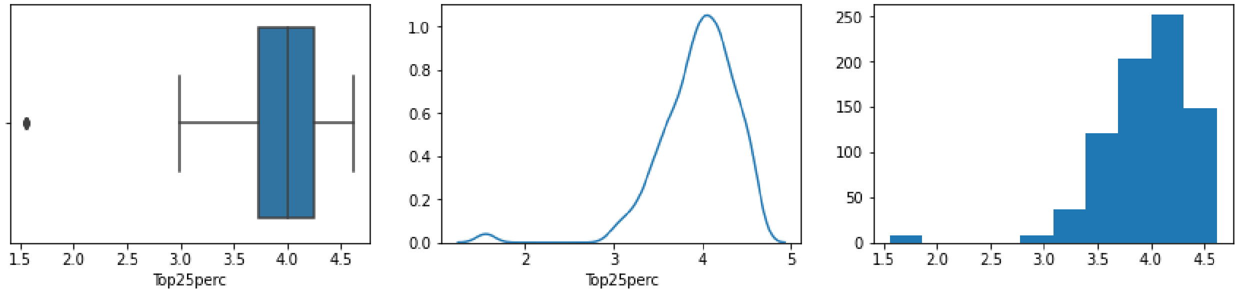
```
Column Enroll is not normally distributed

 4. Univariate analysis for Top10perc

Mean is 3.202929, Median is 3.218876, Modes are [37]
Column Top10perc has outliers
```



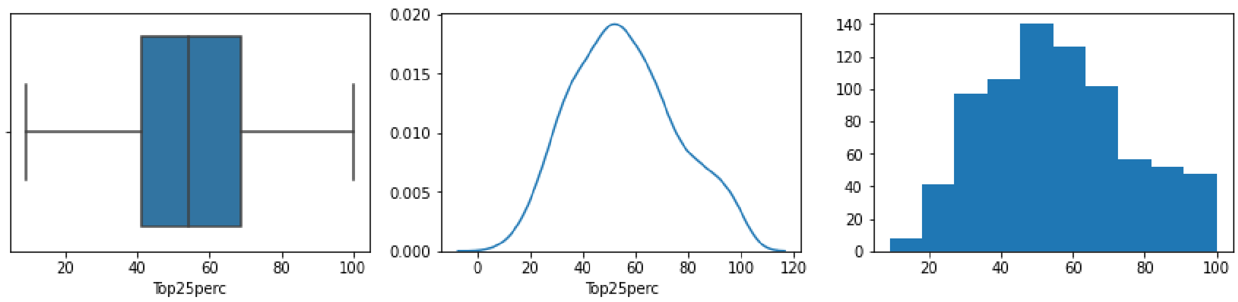
```
Column Top10perc is not normally distributed

 4. Original dataset: Univariate analysis for Top10perc

Mean is 27.558559, Median is 23.000000, Modes are [37]
Column Top10perc has outliers
```



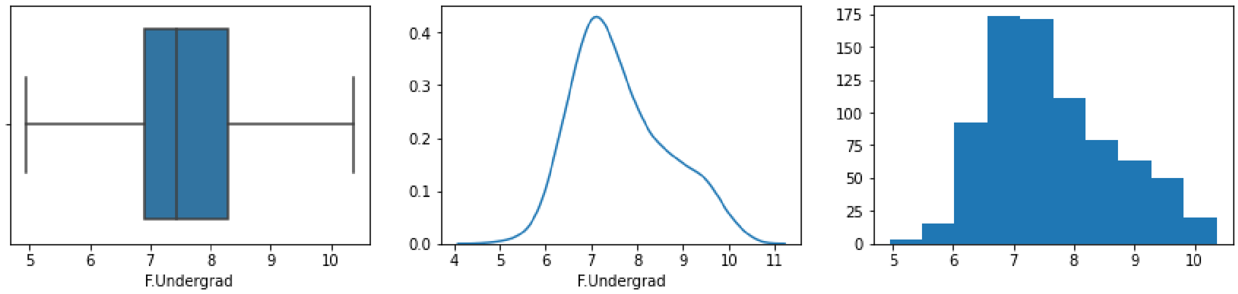
```
Column Top10perc is not normally distributed

 5. Univariate analysis for Top25perc

Mean is 3.960345, Median is 4.007333, Modes are [20]
Column Top25perc has outliers
```



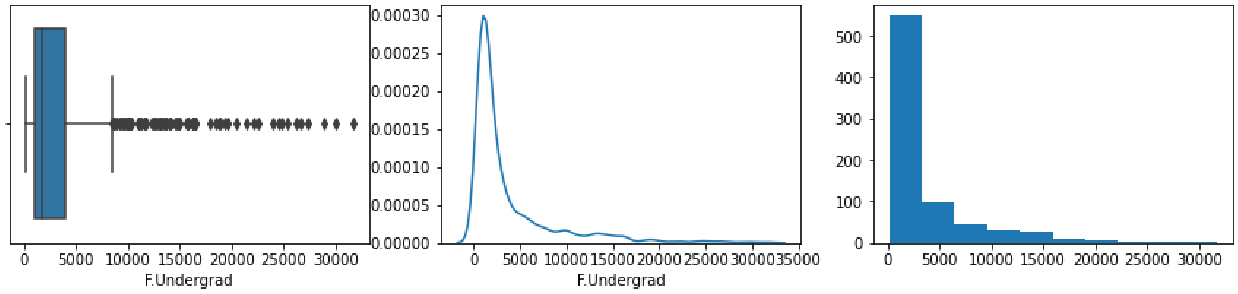
```
Column Top25perc is not normally distributed

 5. Original dataset: Univariate analysis for Top25perc

Mean is 55.796654, Median is 54.000000, Modes are [20]
Column Top25perc does not have outliers
```



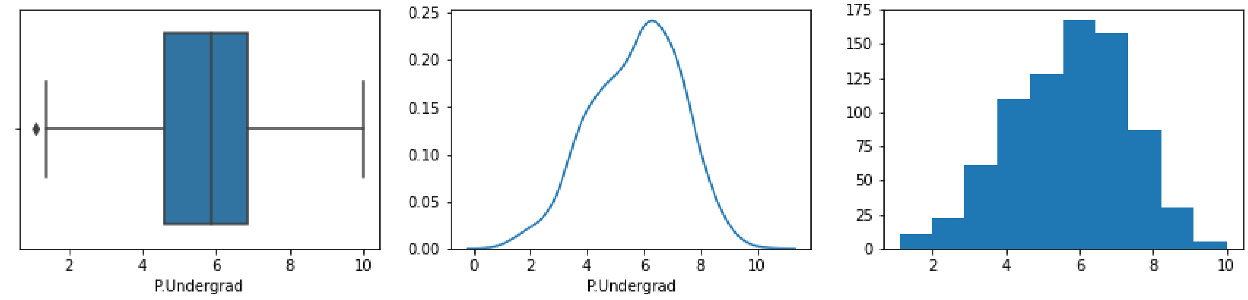
```
Column Top25perc is not normally distributed

 6. Univariate analysis for F.Undergrad

Mean is 7.635932, Median is 7.443078, Modes are [3]
Column F.Undergrad does not have outliers
```



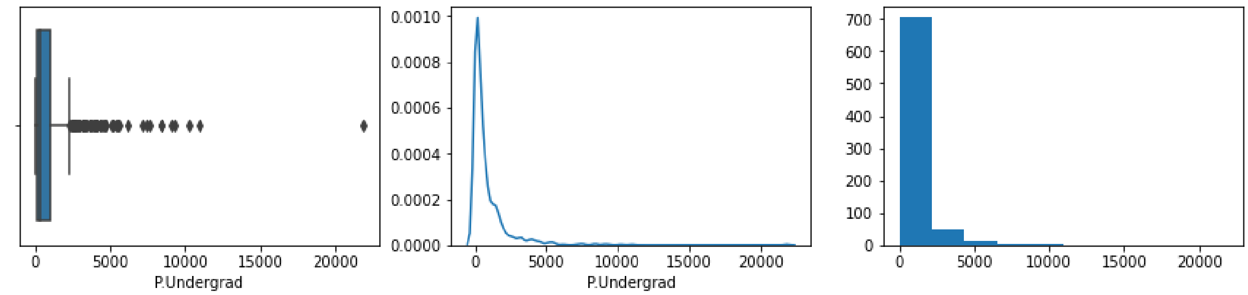
```
Column F.Undergrad is not normally distributed

 6. Original dataset: Univariate analysis for F.Undergrad

Mean is 3699.907336, Median is 1707.000000, Modes are [3]
Column F.Undergrad has outliers
```



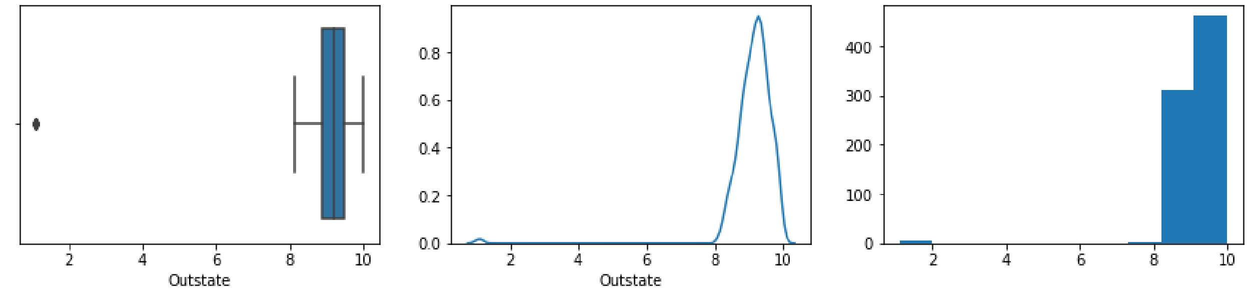
```
Column F.Undergrad is not normally distributed

 7. Univariate analysis for P.Undergrad

Mean is 5.728152, Median is 5.869297, Modes are [7]
Column P.Undergrad has outliers
```



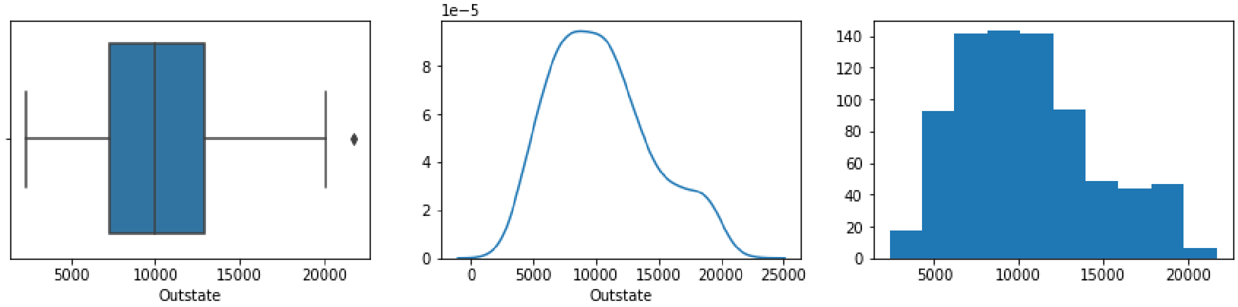
```
Column P.Undergrad is not normally distributed

 7. Original dataset: Univariate analysis for P.Undergrad

Mean is 855.298584, Median is 353.000000, Modes are [7]
Column P.Undergrad has outliers
```



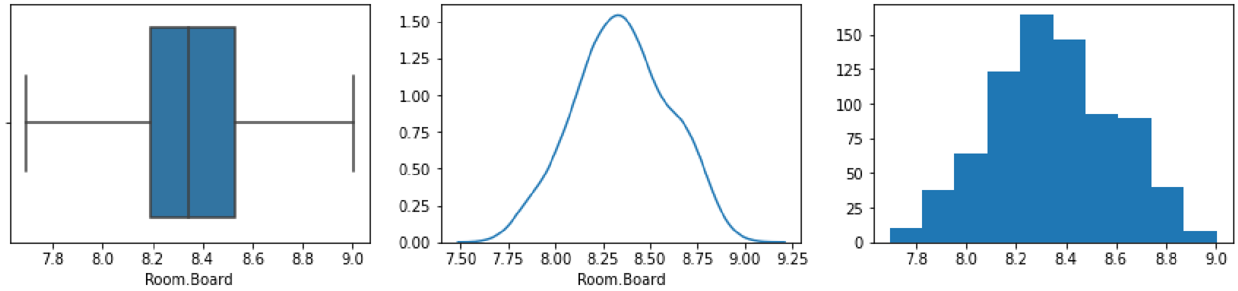
```
Column P.Undergrad is not normally distributed

 8. Univariate analysis for Outstate

Mean is 9.140772, Median is 9.209440, Modes are [13]
Column Outstate has outliers
```



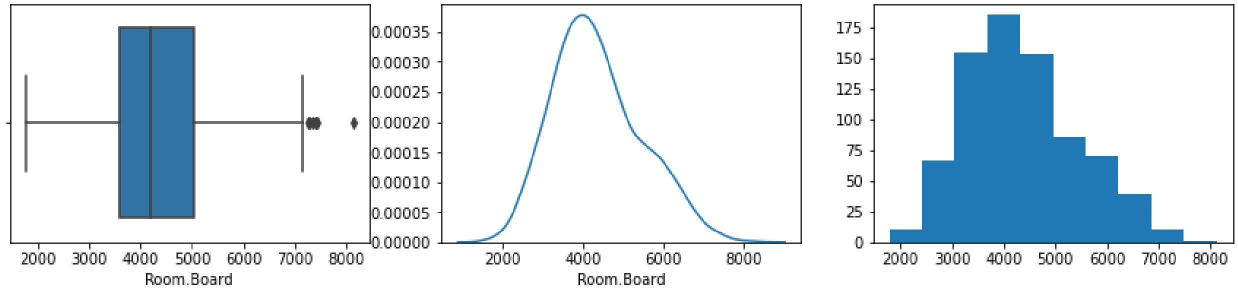
```
Column Outstate is not normally distributed

 8. Original dataset: Univariate analysis for Outstate

Mean is 10440.669241, Median is 9990.000000, Modes are [13]
Column Outstate has outliers
```



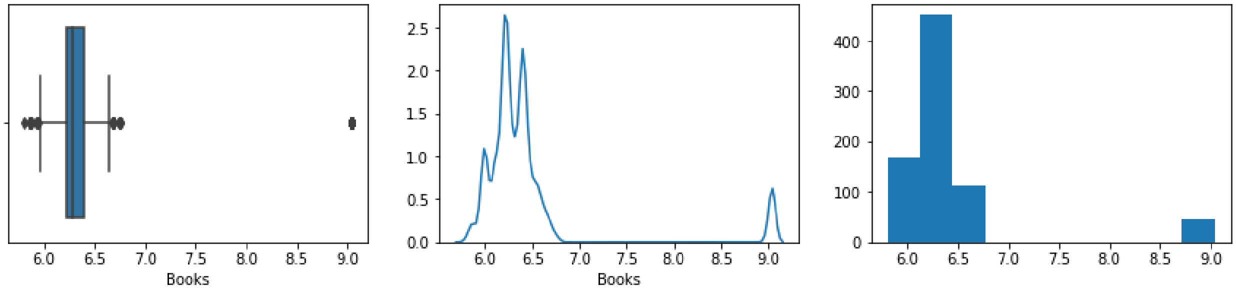
```
Column Outstate is not normally distributed

 9. Univariate analysis for Room.Board

Mean is 8.350686, Median is 8.343078, Modes are [9]
Column Room.Board does not have outliers
```



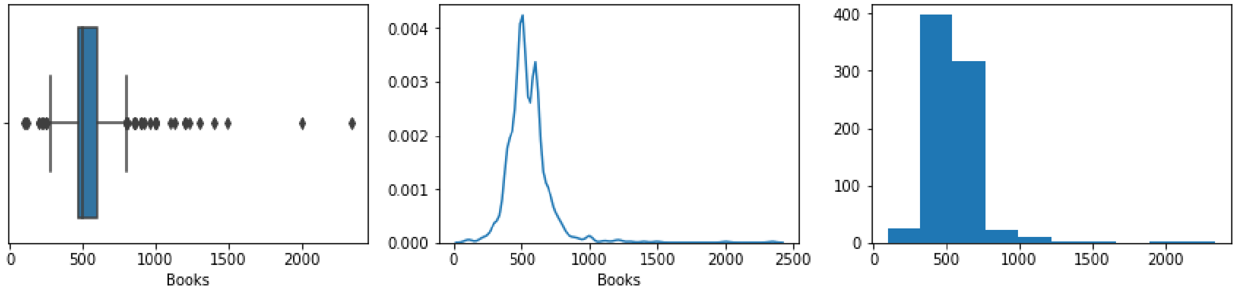
```
Column Room.Board is not normally distributed

 9. Original dataset: Univariate analysis for Room.Board

Mean is 4357.526384, Median is 4200.000000, Modes are [9]
Column Room.Board has outliers
```



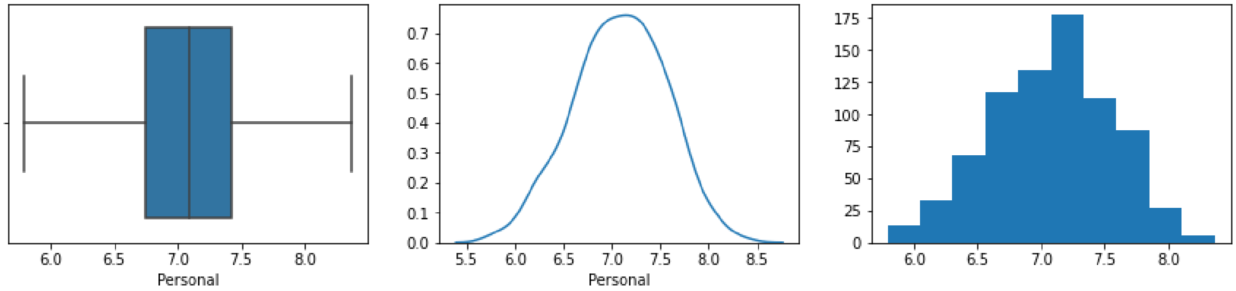
```
Column Room.Board is not normally distributed

 10. Univariate analysis for Books

Mean is 6.441340, Median is 6.284134, Modes are [178]
Column Books has outliers
```



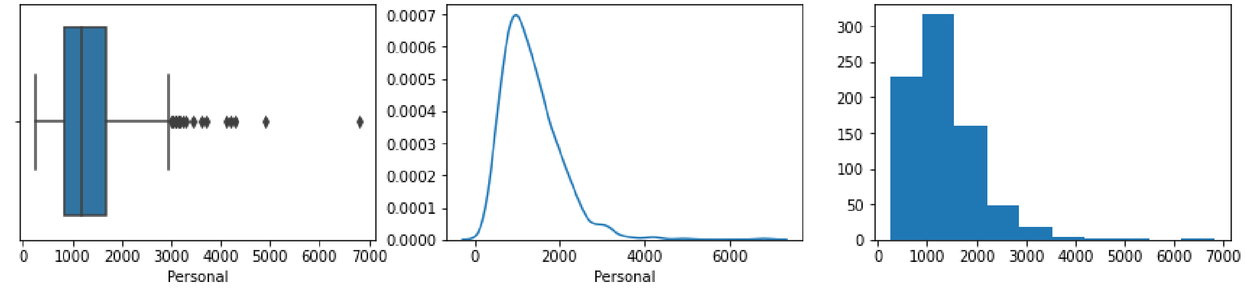
```
Column Books is not normally distributed

 10. Original dataset: Univariate analysis for Books

Mean is 549.380952, Median is 500.000000, Modes are [178]
Column Books has outliers
```



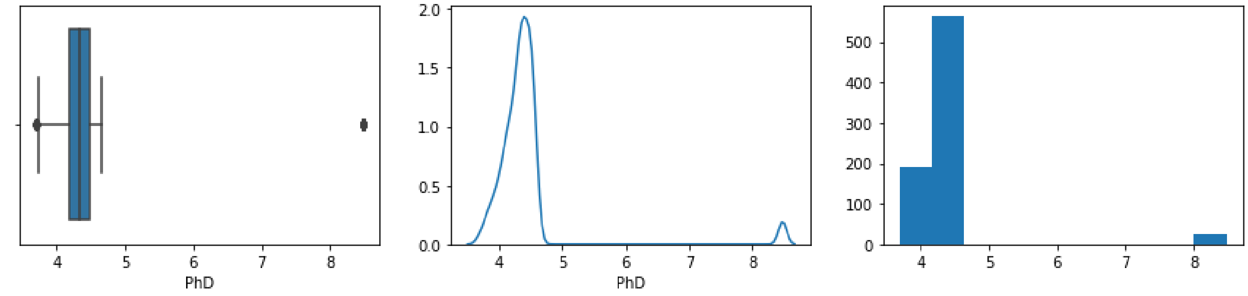
```
Column Books is not normally distributed

 11. Univariate analysis for Personal

Mean is 7.081889, Median is 7.090910, Modes are [45]
Column Personal does not have outliers
```



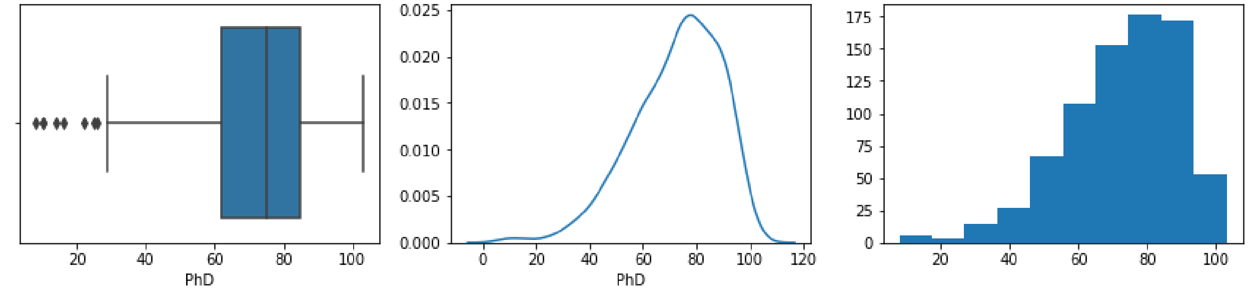
```
Column Personal is not normally distributed

 11. Original dataset: Univariate analysis for Personal

Mean is 1340.642214, Median is 1200.000000, Modes are [45]
Column Personal has outliers
```



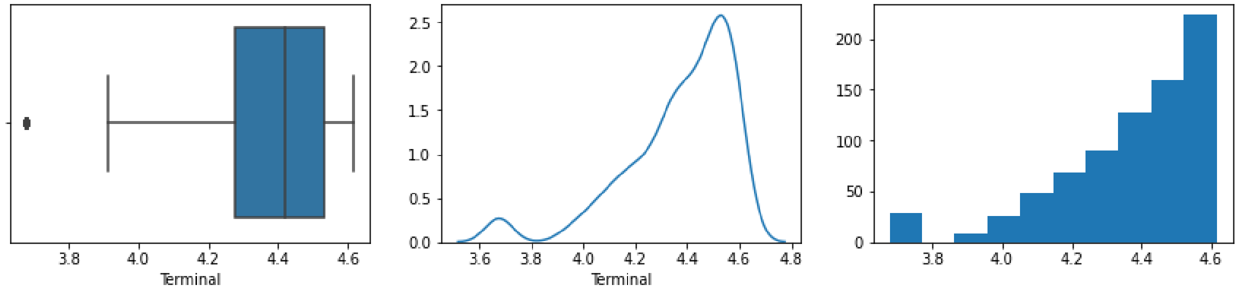
```
Column Personal is not normally distributed

 12. Univariate analysis for PhD

Mean is 4.427462, Median is 4.343805, Modes are [26]
Column PhD has outliers
```



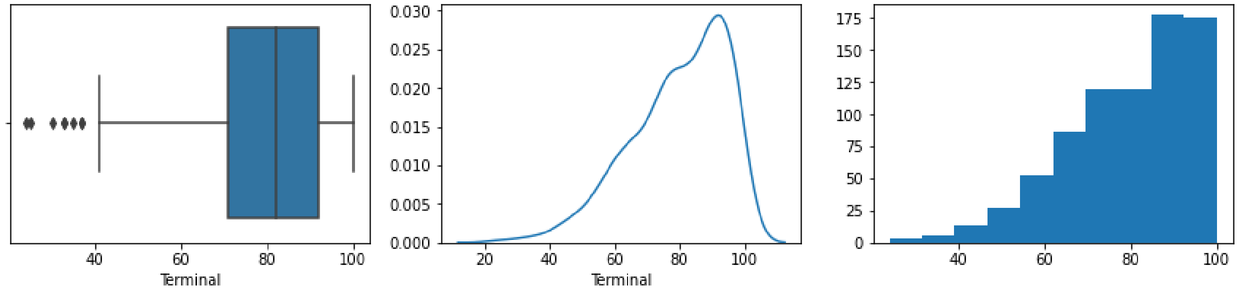
```
Column PhD is not normally distributed

 12. Original dataset: Univariate analysis for PhD

Mean is 72.660232, Median is 75.000000, Modes are [26]
Column PhD has outliers
```



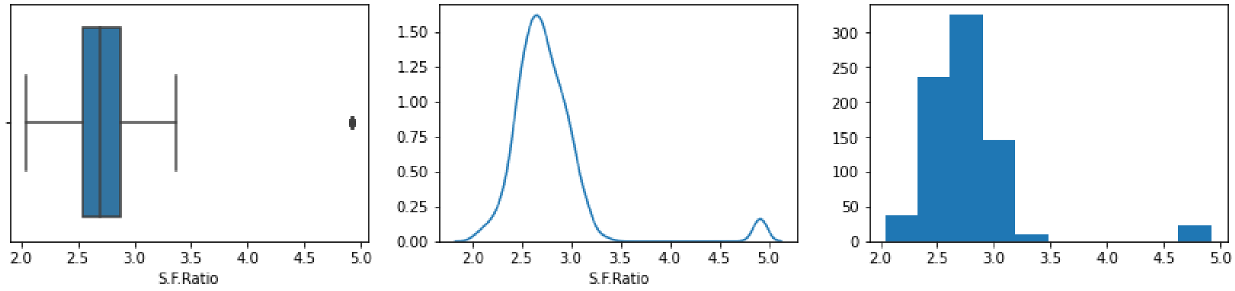
```
Column PhD is not normally distributed

 13. Univariate analysis for Terminal

Mean is 4.369178, Median is 4.418841, Modes are [30]
Column Terminal has outliers
```



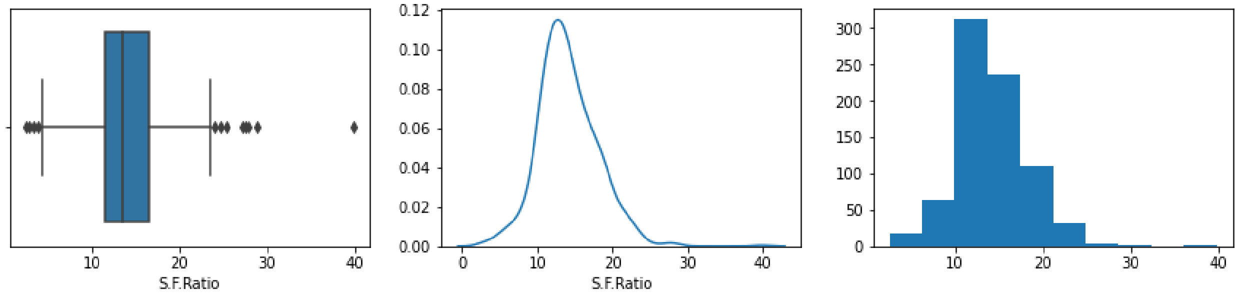
```
Column Terminal is not normally distributed

 13. Original dataset: Univariate analysis for Terminal

Mean is 79.702703, Median is 82.000000, Modes are [30]
Column Terminal has outliers
```



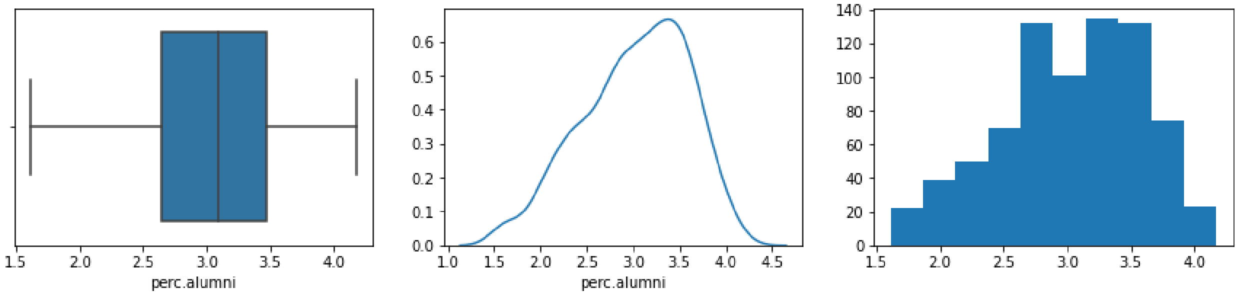
```
Column Terminal is not normally distributed

 14. Univariate analysis for S.F.Ratio

Mean is 2.762204, Median is 2.701361, Modes are [22]
Column S.F.Ratio has outliers
```



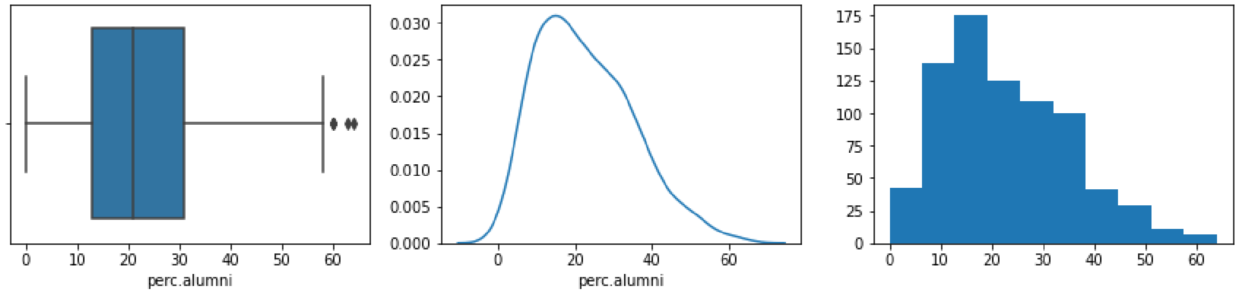
```
Column S.F.Ratio is not normally distributed

 14. Original dataset: Univariate analysis for S.F.Ratio

Mean is 14.089704, Median is 13.600000, Modes are [15]
Column S.F.Ratio has outliers
```



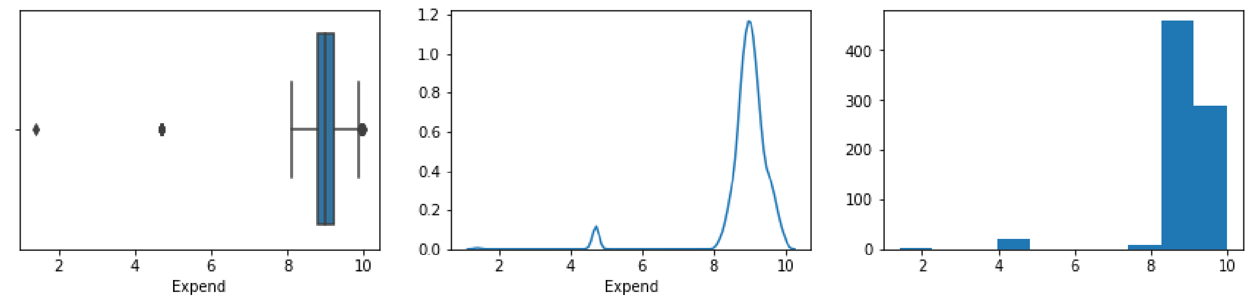
```
Column S.F.Ratio is not normally distributed

 15. Univariate analysis for perc.alumni

Mean is 3.023637, Median is 3.091042, Modes are [32]
Column perc.alumni does not have outliers
```



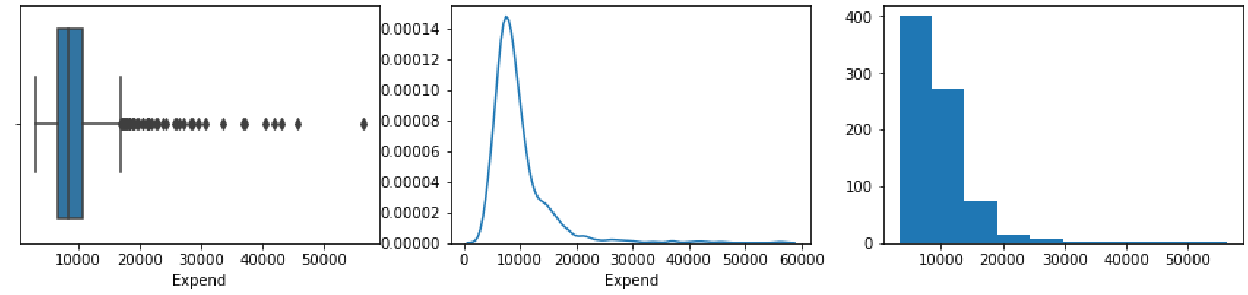
```
Column perc.alumni is not normally distributed

 15. Original dataset: Univariate analysis for perc.alumni

Mean is 22.743887, Median is 21.000000, Modes are [32]
Column perc.alumni has outliers
```



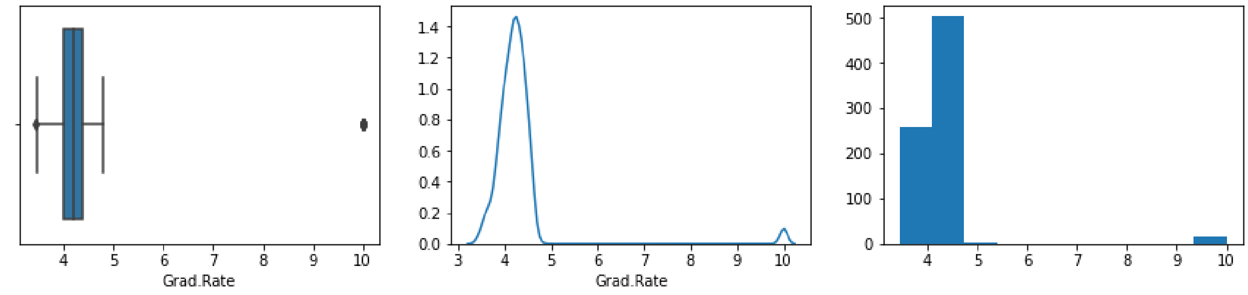
```
Column perc.alumni is not normally distributed

 16. Univariate analysis for Expend

Mean is 8.919903, Median is 9.004054, Modes are [21]
Column Expend has outliers
```



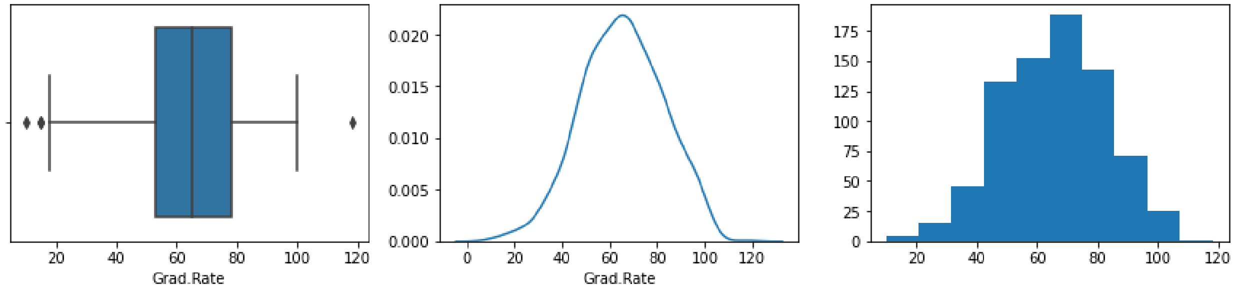
```
Column Expend is not normally distributed

 16. Original dataset: Univariate analysis for Expend

Mean is 9660.171171, Median is 8377.000000, Modes are [2]
Column Expend has outliers
```



```
Column Expend is not normally distributed

 17. Univariate analysis for Grad.Rate

Mean is 4.291241, Median is 4.204693, Modes are [24]
Column Grad.Rate has outliers
```



```
Column Grad.Rate is not normally distributed

 17. Original dataset: Univariate analysis for Grad.Rate

Mean is 65.463320, Median is 65.000000, Modes are [24]
Column Grad.Rate has outliers
```



```
Column Grad.Rate is not normally distributed
```

2.5) Build the covariance matrix and calculate the eigenvalues and the eigenvector.

```
Part A of 2.5: Covariance Matrix
```

```
The scaled dataset from the original data has been further imputed for its
outliers with whiskers. Subsequently it order to normal spread of the data
it has been standardized further wherein the resultant data set would have
mean as zero and standard deviation as 1.
The transpose of standardized dataset is used to build the below covarianc
e matrix.

%s [[ 1.00128866  0.95541955  0.91262543  0.32523071  0.32626151  0.874392
33
   0.37895685  0.01277525  0.18424231  0.05770185  0.19192928  0.01000893
   0.45265804  0.25188199 -0.06942519 -0.10662031 -0.00853013]
 [ 0.95541955  1.00128866  0.94871582  0.25983855  0.27964596  0.90664593
   0.43367266 -0.0099024   0.13195219  0.04534186  0.21298526 -0.00103516
   0.43002043  0.24795573 -0.10025888 -0.0555676  -0.03255256]
 [ 0.91262543  0.94871582  1.00128866  0.21786442  0.24810373  0.96294356
   0.50145483 -0.12729824 -0.0105333   0.05610706  0.28258486 -0.00752911
   0.3754636   0.27164187 -0.17069922 -0.10579141 -0.03251622]
 [ 0.32523071  0.25983855  0.21786442  1.00128866  0.73648415  0.16428874
  -0.26404496  0.30867023  0.33552143 -0.04263824 -0.12274647 -0.00587091
   0.43808623  0.01983607  0.42355493 -0.01234464  0.02152219]
 [ 0.32626151  0.27964596  0.24810373  0.73648415  1.00128866  0.22379513
  -0.13991007  0.2406126   0.29150266 -0.00429549 -0.06341502 -0.04900996
   0.4744419  -0.00554412  0.36811822  0.01705732 -0.06462958]
 [ 0.87439233  0.90664593  0.96294356  0.16428874  0.22379513  1.00128866
   0.58854327 -0.15676453 -0.02445247  0.067161    0.30994846 -0.01062726
   0.35644072  0.28324972 -0.24048969 -0.12384906 -0.07395962]
 [ 0.37895685  0.43367266  0.50145483 -0.26404496 -0.13991007  0.58854327
   1.00128866 -0.2519212  -0.07403701  0.01703668  0.32875726 -0.01453353
   0.03117041  0.2033387  -0.44139811 -0.04192537 -0.06854909]
 [ 0.01277525 -0.0099024  -0.12729824  0.30867023  0.2406126  -0.15676453
  -0.2519212   1.00128866  0.43515264 -0.0363087  -0.17736377  0.05952298
   0.17614089 -0.10801639  0.30250703  0.06556907  0.05252247]
 [ 0.18424231  0.13195219 -0.0105333   0.33552143  0.29150266 -0.02445247
  -0.07403701  0.43515264  1.00128866 -0.02949514 -0.21344729  0.07049765
   0.34510098 -0.01673382  0.2714961   0.05867713  0.02022646]
 [ 0.05770185  0.04534186  0.05610706 -0.04263824 -0.00429549  0.067161
   0.01703668 -0.0363087  -0.02949514  1.00128866  0.04229724  0.14263648
  -0.01420618  0.00908497 -0.09644504  0.02998151 -0.04908418]
 [ 0.19192928  0.21298526  0.28258486 -0.12274647 -0.06341502  0.30994846
   0.32875726 -0.17736377 -0.21344729  0.04229724  1.00128866 -0.00919825
  -0.01612421  0.08372049 -0.29104666 -0.05438935 -0.04653986]
 [ 0.01000893 -0.00103516 -0.00752911 -0.00587091 -0.04900996 -0.01062726
  -0.01453353  0.05952298  0.07049765  0.14263648 -0.00919825  1.00128866
  -0.18941157  0.00542516 -0.04878139  0.02743129  0.10286011]
 [ 0.45265804  0.43002043  0.3754636   0.43808623  0.4744419   0.35644072
   0.03117041  0.17614089  0.34510098 -0.01420618 -0.01612421 -0.18941157
   1.00128866  0.04270595  0.25265237  0.06445938 -0.06496179]
 [ 0.25188199  0.24795573  0.27164187  0.01983607 -0.00554412  0.28324972
   0.2033387  -0.10801639 -0.01673382  0.00908497  0.08372049  0.00542516
   0.04270595  1.00128866 -0.13060367 -0.50070988 -0.01847701]
 [-0.06942519 -0.10025888 -0.17069922  0.42355493  0.36811822 -0.24048969
  -0.44139811  0.30250703  0.2714961  -0.09644504 -0.29104666 -0.04878139
   0.25265237 -0.13060367  1.00128866  0.01277898  0.02201253]
 [-0.10662031 -0.0555676  -0.10579141 -0.01234464  0.01705732 -0.12384906
  -0.04192537  0.06556907  0.05867713  0.02998151 -0.05438935  0.02743129
   0.06445938 -0.50070988  0.01277898  1.00128866 -0.02113291]
 [-0.00853013 -0.03255256 -0.03251622  0.02152219 -0.06462958 -0.07395962
  -0.06854909  0.05252247  0.02022646 -0.04908418 -0.04653986  0.10286011
  -0.06496179 -0.01847701  0.02201253 -0.02113291  1.00128866]]
```

Part B of 2.5: Eigen values and Eigen vectors

The extraction of eigen vectors and eigen values follow the traditional method using linear algebra libraries. The resulting eigen values with its corresponding eigen vectors could be further processed to ascertain optimal number of principal components that enables representation of data from high dimensional space to low dimensional space based on the cumulative percentage coverage of variances (from the eigen values). This process is called dimension reduction.
Please find below the derived eigen values and related eigen vectors along with variances explained both at individual levels of eigen values and at its cumulative levels in its descending order.

```
Sorted Eigen values in descending order:
%s [4.683397133581554, 3.0926752033820497, 1.4150913593257115, 1.242765373
88643, 1.054608412107467, 0.9726867273112076, 0.8346439370126061, 0.797946
9470672212, 0.6226380761382637, 0.5650932810065527, 0.4850339454217155, 0.
4453577902878707, 0.4228799324849895, 0.23807285040562853, 0.0868029983201
3989, 0.03946682639945528, 0.022746422355994823]

Eigen Vectors as below:
[[-4.35748482e-01 -4.03364844e-02 -2.83934370e-02 -8.38665148e-02
  -5.58243663e-02 -1.62153504e-01 -5.49512118e-01  6.16611011e-01
  -1.28872642e-02  8.23794028e-02  6.67439775e-02 -4.28007045e-02
  -1.70834912e-01  3.50923636e-02 -1.65558605e-01  1.33633444e-01
   1.22852534e-02]
 [-4.39771363e-01 -3.21520279e-03 -7.03943912e-02 -6.12248159e-02
  -7.04273419e-02  4.76241061e-01  6.47794247e-01  2.38357134e-01
   3.33753348e-03  8.52593090e-02  9.30464419e-02 -7.07190940e-02
  -2.03528649e-01  4.17960752e-02 -1.46287423e-01  4.61732011e-02
   4.50330572e-02]
 [-4.42742378e-01  6.72421367e-02 -4.81800486e-02 -1.94443160e-02
  -3.42385008e-02 -7.36028533e-01  2.70420500e-01 -3.56236591e-01
  -8.82230632e-02  6.02056963e-02  1.00077819e-01 -1.40966918e-02
  -1.69960304e-01  1.22514605e-02 -5.41199141e-02  1.50016177e-02
  -1.81997122e-03]
 [-1.53181502e-01 -4.27673380e-01  8.35205231e-02  2.60166807e-02
   1.20616756e-01  2.92778730e-02  3.30085145e-02  8.25157332e-03
  -2.72752957e-01 -1.72553934e-01  2.93061507e-02  7.25928158e-01
   1.43830013e-01 -3.11989584e-01 -1.29362111e-01 -2.73623348e-03
  -5.14676906e-02]
 [-1.75065087e-01 -3.89705820e-01  1.24575309e-02  1.12819046e-01
   2.04426016e-01 -9.25470460e-03  2.16935607e-02  3.76873470e-02
  -2.36740194e-01 -1.94416764e-01  4.55357173e-03 -6.36897988e-01
   2.78259279e-01 -3.39832976e-01  1.18296056e-01 -1.25177656e-01
  -1.96850222e-01]
 [-4.38826310e-01  1.06366051e-01 -5.08358106e-02 -5.07174054e-03
  -1.31354705e-02  4.47911601e-01 -4.42319019e-01 -6.10760699e-01
  -3.00906828e-02  3.44674575e-02  8.27887344e-02 -8.57632440e-03
  -9.65206183e-02 -2.42411790e-02  3.52519378e-02 -1.14340558e-02
  -6.11339329e-02]
 [-2.42239117e-01  3.02850148e-01 -1.17403261e-01 -1.71751682e-03
  -1.43840503e-01 -3.69527979e-02  4.35265682e-02  1.86547610e-01
   2.75661388e-01 -4.06060148e-02 -4.14277812e-02  2.10244081e-01
   2.93125798e-01 -5.13457967e-02  3.67805276e-01 -4.87379443e-01
  -4.38326247e-01]
 [ 2.61764862e-02 -3.39135301e-01 -1.45497786e-02 -2.51126196e-01
  -1.47351014e-01 -2.21515063e-02 -2.03408096e-02 -3.84930883e-02
   3.64654010e-01 -2.16472683e-01 -3.01614547e-01 -1.49784701e-02
  -5.62444692e-01 -2.90857780e-01  2.71461900e-01 -1.77014328e-01
   1.41176929e-01]
 [-6.02407989e-02 -3.37426857e-01 -1.60873407e-02 -2.43086643e-01
  -1.63163659e-01 -2.89509652e-02  1.18372641e-02 -1.52745510e-01
   5.37077354e-01  5.71886373e-04 -1.47214016e-01 -5.78237233e-02
   3.65334474e-01  1.70559961e-01 -4.39255092e-01  2.03306619e-01
  -2.48849027e-01]
 [-2.78879876e-02  5.35469377e-02 -1.32100671e-01 -3.58669246e-01
   6.99241223e-01  1.16770430e-03  1.50373713e-02  1.98278399e-03
  -4.22828965e-02  4.11385902e-01 -4.14706473e-01  3.98088213e-02
  -3.40582276e-02  1.62839089e-02 -1.20883680e-02 -7.33424925e-02
  -9.77153094e-02]
 [-1.39590890e-01  2.35308094e-01 -1.00622913e-01  9.03605823e-02
   7.31304884e-02  8.02271175e-03  5.64092171e-03  1.69737647e-02
  -1.71472275e-01 -6.79208938e-01 -5.40040861e-01 -4.53514692e-03
   7.83679800e-03  2.88572411e-01 -1.78106670e-01  1.49668355e-02
   5.87235220e-02]
 [ 1.56994590e-02  2.28897263e-02 -3.76567713e-02 -7.18090484e-01
   1.23686617e-01  9.88302358e-04  7.06678499e-03  6.97959510e-03
  -8.66051344e-02 -3.64011232e-01  4.40089008e-01 -5.78582861e-03
   1.47790236e-01  1.95276707e-01  2.32747155e-01  5.80966377e-02
   1.24586360e-01]
 [-2.36795057e-01 -2.89779165e-01 -8.53482628e-02  2.07274585e-01
   5.85687853e-04 -8.52050350e-03  1.55368181e-02  9.80160434e-03
   9.28795883e-02  1.56419874e-01 -1.87158721e-01  7.38793949e-02
   2.98940076e-01  3.33475728e-01  5.49929901e-01  2.27366554e-01
   4.25474590e-01]
  [-1.57860597e-01  1.12474904e-01  6.26315466e-01 -5.90636070e-02
    8.22495487e-02 -5.02452309e-03 -1.77949704e-02 -4.78765177e-03
    1.54912294e-01  3.08319516e-02 -2.81378911e-02 -5.15241145e-02
    2.09099034e-01 -5.41466950e-02 -2.10494613e-01 -4.39105311e-01
    4.97362452e-01]
  [ 6.23355324e-02 -4.06670548e-01  1.10257738e-01  9.52933504e-02
    3.17826747e-02  1.17712909e-02 -3.66394011e-02 -2.15687946e-02
   -1.57064274e-01  4.22910864e-02  1.21103331e-01  4.74048903e-03
   -2.49165085e-01  6.51790973e-01 -3.69860459e-02 -4.74599046e-01
   -2.27741631e-01]
  [ 6.80083304e-02 -7.75821623e-02 -7.21243279e-01  1.16416916e-02
   -7.67873456e-02 -1.26689454e-02 -5.83267117e-02 -1.98823943e-02
   -1.61353207e-02  5.39621865e-02  9.67118166e-02 -1.44102007e-02
    1.43118523e-01 -7.85293907e-02 -2.81381763e-01 -4.13751770e-01
    4.11613009e-01]
  [ 3.18780070e-02 -1.95137290e-02  8.39662991e-02 -3.80946804e-01
   -5.84772711e-01  2.09063173e-02 -2.58710156e-03 -2.49395923e-02
   -5.16754380e-01  2.68074351e-01 -3.70992808e-01 -5.03148016e-02
    1.36456539e-01 -1.80591220e-02  3.25414474e-02 -6.20394403e-02
   -2.20630892e-02]]
```

```
Variance explained in terms of percentages from the eigen values as below:
---------------------------------------------------------------------------
 [27.513938796724073, 18.16879368479422, 8.31335373485314, 7.3009760779458
62, 6.195594880728244, 5.714322813184166, 4.903351465832249, 4.68776463717
2861, 3.6578631772525716, 3.3198000307448297, 2.8494688594689297, 2.616380
083756434, 2.4843275615743172, 1.3986261784750365, 0.5099487220563896, 0.2
3185901495932523, 0.13363028047734096]
```

```
Cumulative Variance Explained from the above that can assist in dimension
reduction:
---------------------------------------------------------------------------
 [ 27.5139388   45.68273248  53.99608622  61.29706229  67.49265718
  73.20697999  78.11033145  82.79809609  86.45595927  89.7757593
  92.62522816  95.24160824  97.7259358   99.12456198  99.6345107
  99.86636972 100.        ]
```

The highlighted portion above depicts that with the help top 9 variances explained by reverse sorted of eigen values will cover up to ~86% of data and hence we could reduce 16 continues variables/dimensions to just 9 dimensions.

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

Below are the principle components extracted from the scaled, outlier imputed and standardized data set from the given data set. Please note that the dimension reduction exercise from 16 to 9 has been achieved by constructing the eigen values and the respective eigen vectors that revealed the variances covered by following principle components would explain almost 86%

```
of the data.
```

Index	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	0.435748	0.439771	0.442742	0.153182	0.175065	0.438826	0.242239	-0.02618	0.060241	0.027888	0.139591	-0.0157	0.236795	0.157861	-0.06234	-0.06801	-0.03188
1	-0.04034	-0.00322	0.067242	-0.42767	-0.38971	0.106366	0.30285	-0.33914	-0.33743	0.053547	0.235308	0.02289	-0.28978	0.112475	-0.40667	-0.07758	-0.01951
2	-0.02839	-0.07039	-0.04818	0.083521	0.012458	-0.05084	-0.1174	-0.01455	-0.01609	-0.1321	-0.10062	-0.03766	-0.08535	0.626315	0.110258	-0.72124	0.083966
3	0.083867	0.061225	0.019444	-0.02602	-0.11282	0.005072	0.001718	0.251126	0.243087	0.358669	-0.09036	0.71809	-0.20728	0.059064	-0.09529	-0.01164	0.380947
4	0.055824	0.070427	0.034239	-0.12062	-0.20443	0.013135	0.143841	0.147351	0.163164	-0.69924	-0.07313	-0.12369	-0.00059	-0.08225	-0.03178	0.076787	0.584773
5	0.012887	-0.00334	0.088223	0.272753	0.23674	0.030091	-0.27566	-0.36465	-0.53708	0.042283	0.171472	0.086605	-0.09288	-0.15491	0.157064	0.016135	0.516754
6	0.082379	0.085259	0.060206	-0.17255	-0.19442	0.034467	-0.04061	-0.21647	0.000572	0.411386	-0.67921	-0.36401	0.15642	0.030832	0.042291	0.053962	0.268074
7	0.066744	0.093046	0.100078	0.029306	0.004554	0.082789	-0.04143	-0.30162	-0.14721	-0.41471	-0.54004	0.440089	-0.18716	-0.02814	0.121103	0.096712	-0.37099
8	-0.17084	-0.20353	-0.16996	0.14383	0.278259	-0.09652	0.293126	-0.56245	0.365334	-0.03406	0.007837	0.14779	0.29894	0.209099	-0.24917	0.143119	0.136457

The explicit form of PCA 1 would be nothing but multiplication of factors against each of the dimension from the first principle component with the values of corresponding each of the dimensions against the original data set.

The equivalent explicit form of equation for the above PCA 1 would be as follows:

```
0.435748 * Apps + 0.439771 * Accept + 0.442742 * Enroll + 0.153182 * Top10
perc + 0.175065 * Top25perc + 0.438826 * F.Undergrad + 0.242239 * P.Underg
rad + -0.026176 * Outstate + 0.060241 * Room.Board + 0.027888 * Books + 0.
139591 * Personal + -0.015699 * PhD + 0.236795 * Terminal + 0.157861 * S.F
.Ratio + -0.062336 * perc.alumni + -0.068008 * Expend + -0.031878 * Grad.R
ate
```

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.

2.7 Part A

Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components?

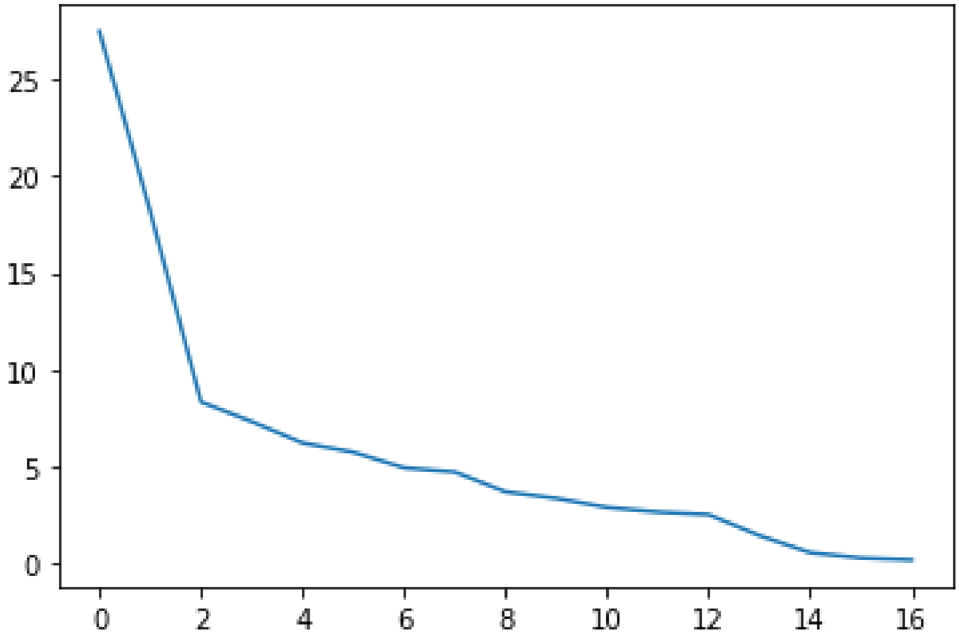
Cumulative values of eigen values and how it helps to decide on the optimum number of principle components:

- Eigen values are the variances mapped to the corresponding eigen vector.
- Highest eigen values based on its proportion to the total eigen values derived out of decomposition of the covariance matrix of the original dataset helps us in optimizing/determining the optimal number of principal components towards prediction analysis. This method is also referred to as dimension reduction.
- Hence cumulative values of eigen values in its descending order helps us to determine optimal number of corresponding eigen vectors to ensure appropriate percentage of coverage of total variance among the data in the original dataset.
- Eigen values assists in how many PCA to pick based on its cumulative percentage coverage compared to the total variance of the original data set.
- Each eigen value derived out of decomposition of the covariance matrix of the original dataset has corresponding eigen vector which is nothing but the principle component for the given eigen value.

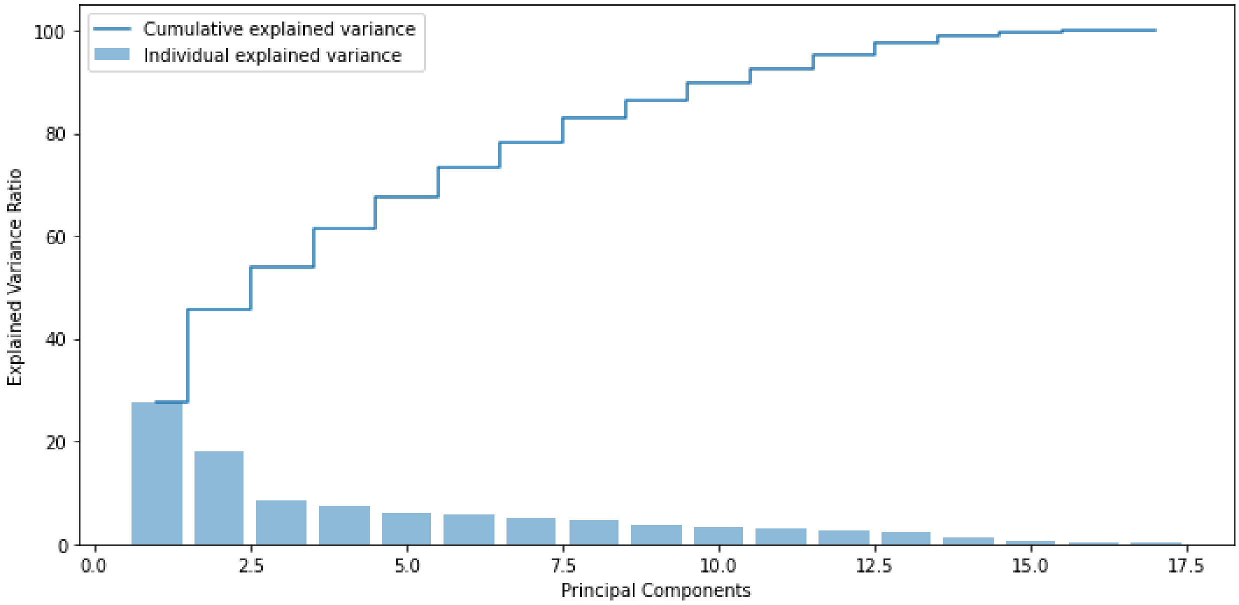
Based on the scree plot for variance explained and the step plot for cumulative eigen values it could be narrowed down that for the given data set we could use the top 9 Principle components towards reducing the number of dimensions from 17 to 9 covering almost 86% of the variation in the dataset. Despite the fact that the elbow point in the scree plot happens at the second principal component at

46th percent of cumulative variance explained, we are proceeding upto 9 principle components to explain up to 86% of the data.

Scree plot as below



Bar plot including the step plot suggesting the cumulative variance explained for the principal component starting from 1 to 17.



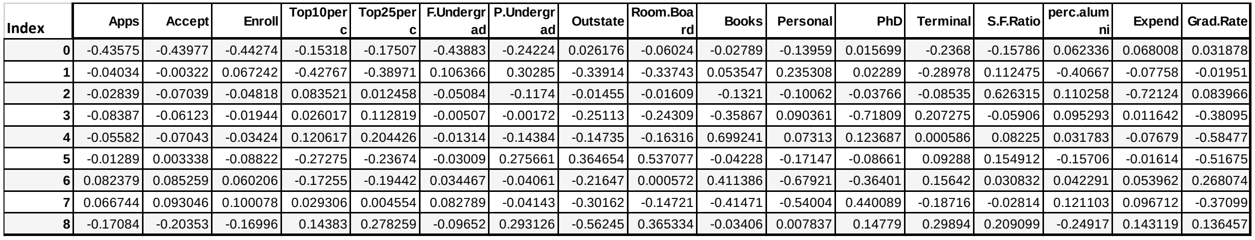
Based on the scree plot that depicts the trend of reverse sorted individual eigen values and step plot that depicts the cumulative trend of those sorted eigen values, it could be derived that the PCA could be limited to 9 components to cover around 86% of the total variance.
Note: Despite the fact that the elbow point in the scree plot happens at the second principal component at 46th percent of cumulative variance explained, we are proceeding up to 9 principle components to explain up to 86% of the data.

2.7 Part B
What do eigen vectors indicate?

- Eigen vector establishes a direction that captures most of the variances on its axis and there will be as many eigen vectors until all of the data could be explained in as many dimensions.
- Each eigen vector will correspond to an appropriate eigen value that explains the variance captured for the dataset equivalent to that of a single eigen vector (i.e Principal component analysis).
- Eigen values assists in how many PCA to pick based on its cumulative percentage coverage compared to the total variance of the original data set.
- Eigen vector can be used for loadings of factor analysis.Influence of principal component on a dimension is called weights or "loadings" which is nothing but eigen vector.
- Eigen vector reduces off diagonal elements of covariance metrics from which it is formed to 0. with every eigen vector corresponding to a Principal component this property of the eigen vector ensures orthogonality of every other principal component also enables to overcome the challenge of multi colinearity among variables when it comes to prediction.

2.7 Part C
Perform PCA and export the data of the Principal Component scores into a data frame..
After scaling the dataset with logarithmic transformation and imputing the outliers it has been standardized before generating covariance matrix for further decomposition of the same into eigen vectors and eigen values by traditional method as performed before.
The extraction of eigen vectors and eigen values following the traditional method allows for better analysis of eigen values and its cumulative variance coverage which enables us to narrow down to optimal number of principal components that better explains the original dataset.
Accordingly, as discussed in the 2.7 Part A we could narrow down to 9 principal components.

Eigen vectors
Accordingly, principle components has been narrowed down and please find below the top 9 principle components which are exported into the data frame.



Index	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	-0.43575	-0.43977	-0.44274	-0.15318	-0.17507	-0.43883	-0.24224	0.026176	-0.06024	-0.02789	-0.13959	0.015699	-0.2368	-0.15786	0.062336	0.068008	0.031878
1	-0.04034	-0.00322	0.067242	-0.42767	-0.38971	0.106366	0.30285	-0.33914	-0.33743	0.053547	0.235308	0.02289	-0.28978	0.112475	-0.40667	-0.07758	-0.01951
2	-0.02839	-0.07039	-0.04818	0.083521	0.012458	-0.05084	-0.1174	-0.01455	-0.01609	-0.1321	-0.10062	-0.03766	-0.08535	0.626315	0.110258	-0.72124	0.083966
3	-0.08387	-0.06123	-0.01944	0.026017	0.112819	-0.00507	-0.00172	-0.25113	-0.24309	-0.35867	0.090361	-0.71809	0.207275	-0.05906	0.095293	0.011642	-0.38095
4	-0.05582	-0.07043	-0.03424	0.120617	0.204426	-0.01314	-0.14384	-0.14735	-0.16316	0.699241	0.07313	0.123687	0.000586	0.08225	0.031783	-0.07679	-0.58477
5	-0.01289	0.003338	-0.08822	-0.27275	-0.23674	-0.03009	0.275661	0.364654	0.537077	-0.04228	-0.17147	-0.08661	0.09288	0.154912	-0.15706	-0.01614	-0.51675
6	0.082379	0.085259	0.060206	-0.17255	-0.19442	0.034467	-0.04061	-0.21647	0.000572	0.411386	-0.67921	-0.36401	0.15642	0.030832	0.042291	0.053962	0.268074
7	0.066744	0.093046	0.100078	0.029306	0.004554	0.082789	-0.04143	-0.30162	-0.14721	-0.41471	-0.54004	0.440089	-0.18716	-0.02814	0.121103	0.096712	-0.37099
8	-0.17084	-0.20353	-0.16996	0.14383	0.278259	-0.09652	0.293126	-0.56245	0.365334	-0.03406	0.007837	0.14779	0.29894	0.209099	-0.24917	0.143119	0.136457

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

By definition, Principal component analysis, or PCA, is a statistical procedure that allows to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed.

In this use case, with so much of statistical information available across various parameters for each and every educational institution we have rich content to work upon to be able to put the data to use and produce analytics for variety of use cases that could help either universities or students towards understanding what the historical trend means and also will enable data driven decision for planning their next action as applicable. It could include predicting the future spend rates for the given university that could help students to plan accordingly, analyzing what drives the expenditure differences between the colleges or the universities need to look at these data and evaluate needs for revisiting some of their operations or functions etc.

However in this case with 16 variables to work upon it becomes very cumbersome to handle so many dimensions to narrow down and focus on the right variables to investigate further for any given use case.

In technical terms it is important to reduce the dimension of the feature space to represent the data in higher dimensional space at the lower dimensional space. In this project as a result of performing PCA after narrowing down the Eigen values that makes up for top 86% of variances, we have reduced to total dimension of 16 to just 9 that explains almost 86% of the data.   And each of the 9 principle components which are orthogonal to each other (independent of each other) contains factors in each of the dimensions that represents the loading weights for the respective dimensions in the data for the respective variance that maps to the eigen value for that principle component.

Essentially these new variables across each components of PCA are all independent of one another. This is a benefit because the assumptions of a linear model require our independent variables to be independent of one another. If we decide to fit a linear regression model with these “new” variables assumption will necessarily be satisfied.