Problem 2:

The dataset 'Education Post 12th std' contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary

- The shape of the data is (777, 18) meaning the dataset has 777 rows and 18 columns.
- The dataset has no missing values or null values.
- The column 'Name' is of object data type while all the other columns are integer or float data type.
- There are no duplicates in the dataset.

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

Univariate Analysis:

1. **Apps:**

Table 1: Description of 'Apps'

```
Description of Apps
-----------------------------------------------
count      777.000000
mean      3001.638353
std       3870.201484
min         81.000000
25%        776.000000
50%       1558.000000
75%       3624.000000
max      48094.000000
Name: Apps, dtype: float64 Distribution of Apps
```

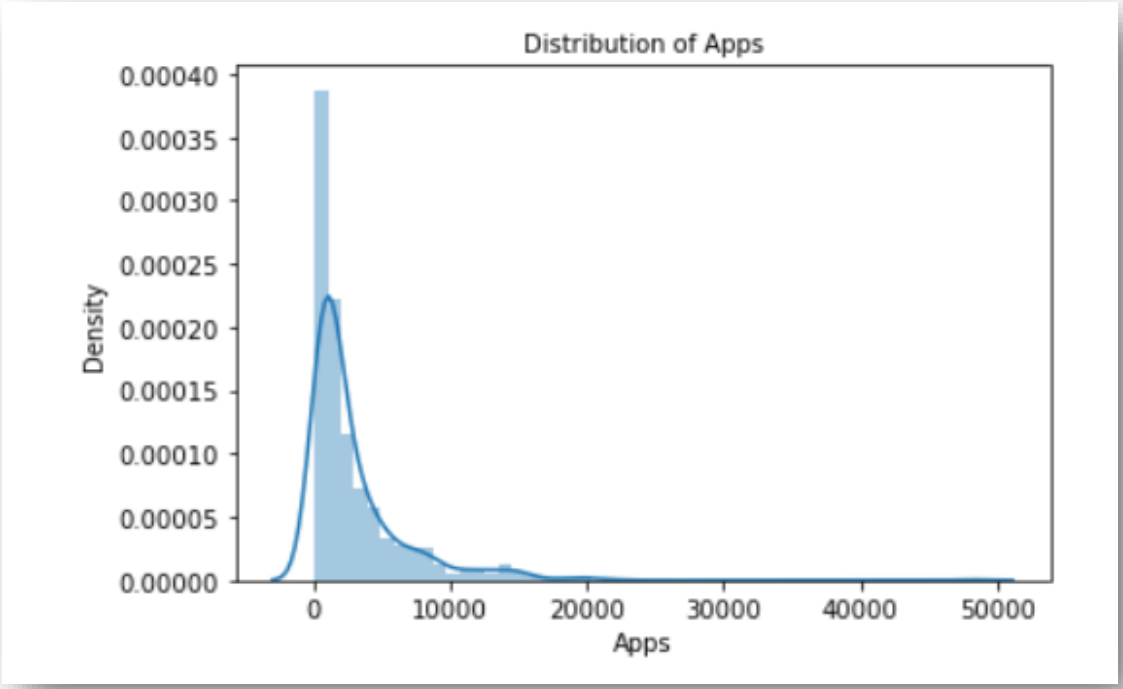

Figure 1: Univariate distribution of 'Apps'

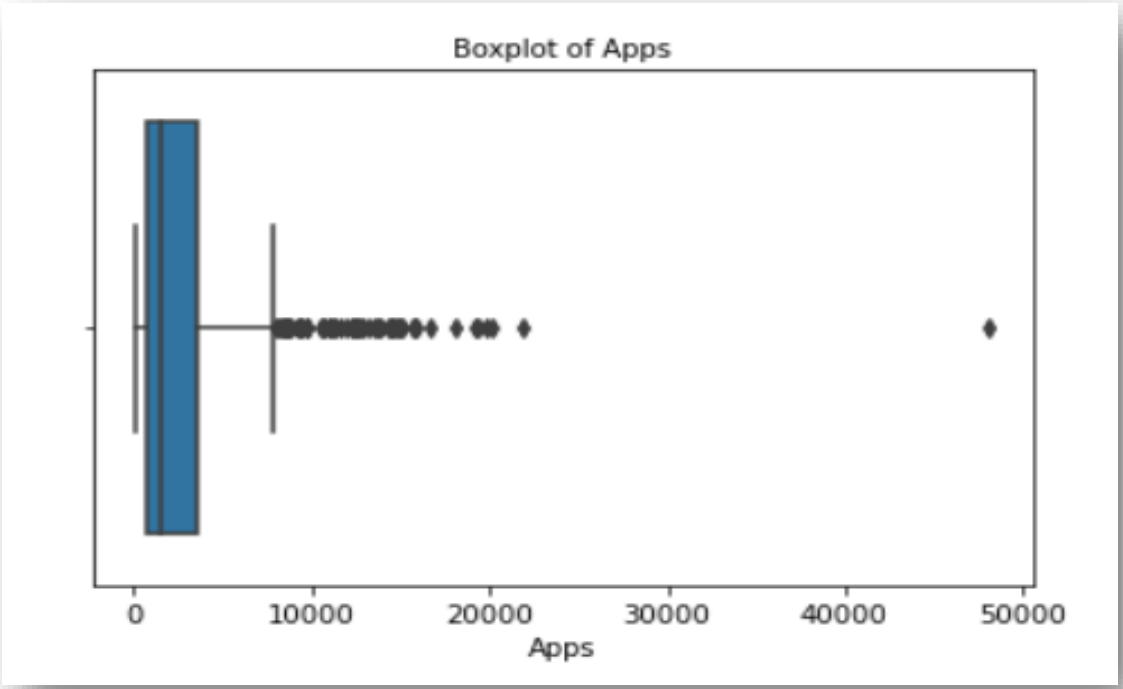

Figure 2: Boxplot showing the distribution of 'Apps'

Univariate analysis of 'Apps' is done to understand the patterns and distribution of the data. From figure 2, we can see that the Box plot of 'Apps' variable has outliers. However the distribution of the data is skewed which is seen in figure 2. From table 4, it is seen that the mean of the data is 3001. From this, we can come to an understanding that the college or university offers application mostly in the range of 3000. The maximum application is 48094.

2. **Accept:**

Table 2: Description of 'Accept'

```
Description of Accept
-----------------------------------------------------
count      777.000000
mean      2018.804376
std       2451.113971
min         72.000000
25%        604.000000
50%       1110.000000
75%       2424.000000
max      26330.000000
Name: Accept, dtype: float64 Distribution of Accept
```

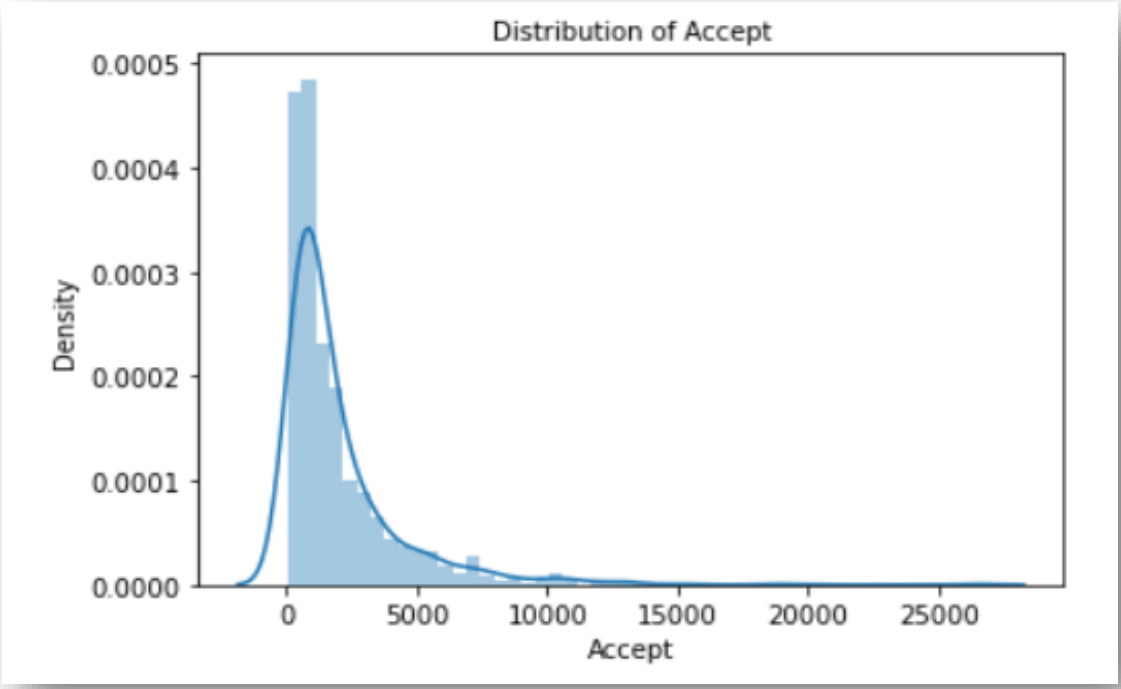

Figure 3: Univariate distribution of 'Accept'

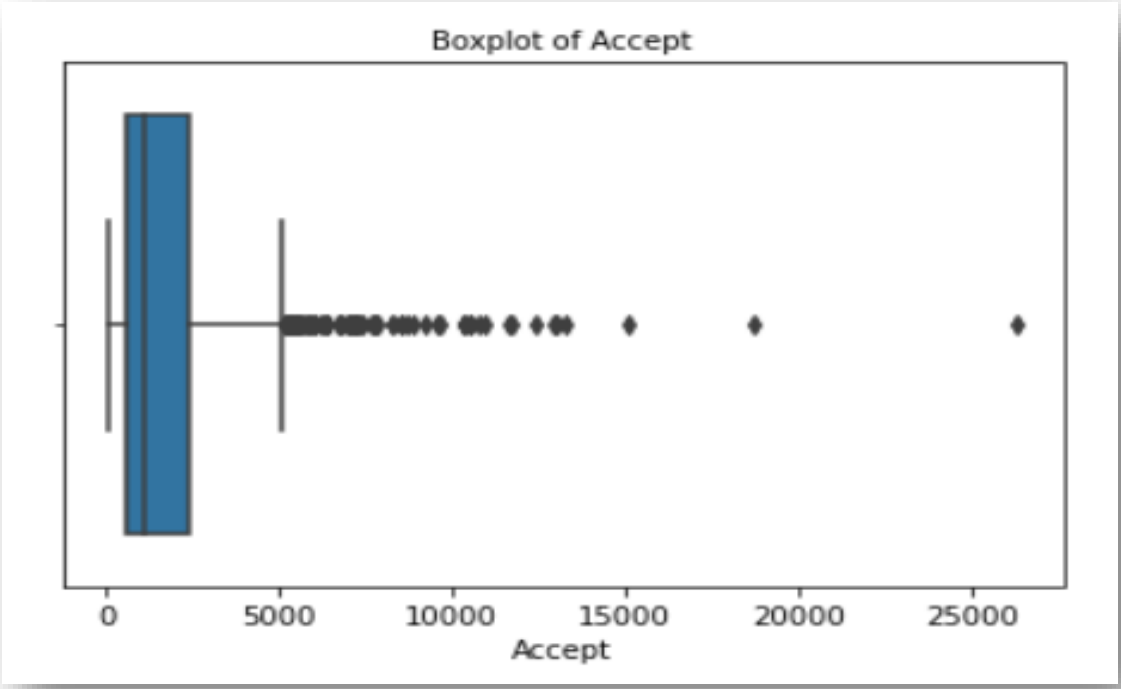

Figure 4: Boxplot showing the distribution of 'Accept'

Univariate analysis of 'Accept' is done to understand the patterns and distribution of the data. From figure 5, we can see that the Box plot of 'Accept' variable has outliers. The distribution of the 'Accept' data is positively skewed which is seen in figure 4. From table 5, it is seen that the applications accepted from each university range from 72 to 26330.

3. Enroll:

Table 3:Description of 'Enroll'

```
Description of Enroll
-----------------------------------------------------
count     777.000000
mean      779.972973
std       929.176190
min        35.000000
25%       242.000000
50%       434.000000
75%       902.000000
max      6392.000000
Name: Enroll, dtype: float64 Distribution of Enroll
```

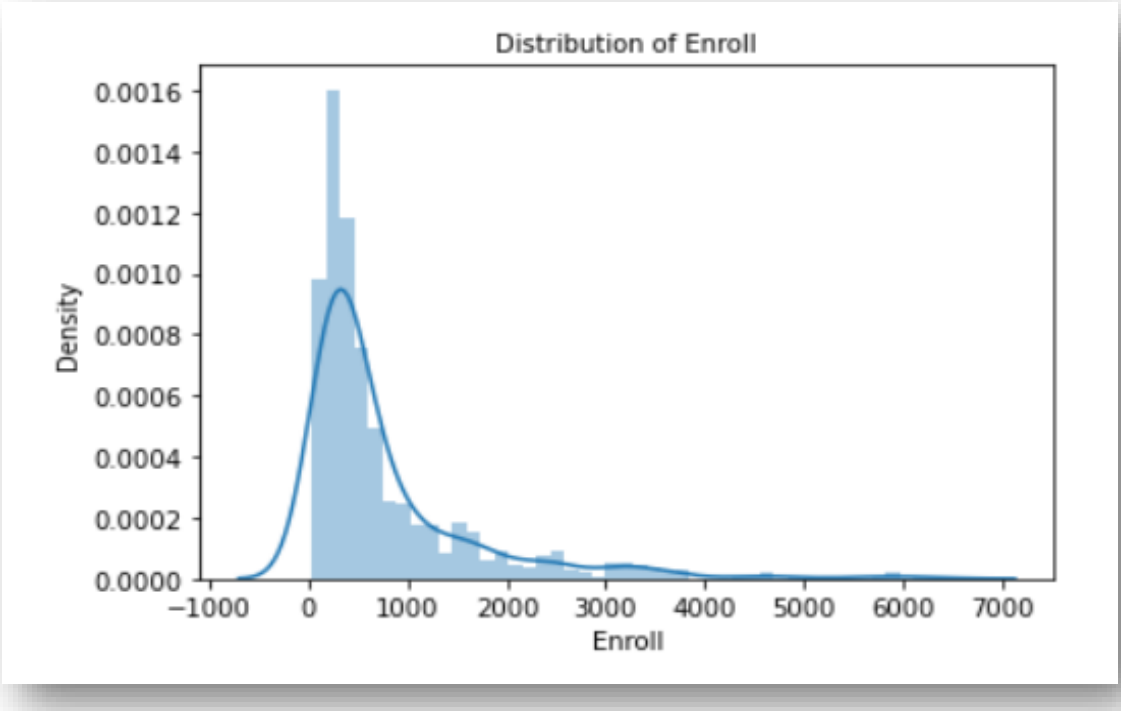

Figure 5: Univariate distribution of 'Enroll'

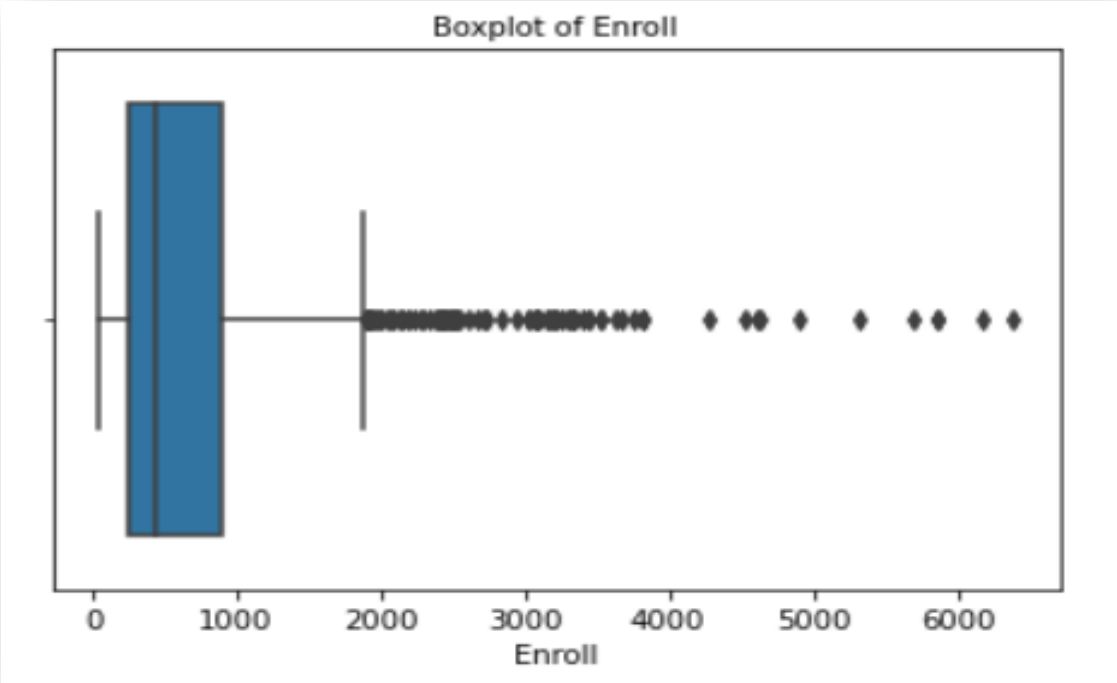

Figure 6: Boxplot showing the distribution of 'Enroll'

Univariate analysis of 'Enroll' is done to understand the patterns and distribution of the data. From figure 7, we can see that the Box plot of 'Enroll' variable has outliers. The distribution of the 'Enroll' data is positively skewed which is seen in figure 6. From table 6, it is seen that the colleges have enrolled students in the range of 35 to 6392.

4. **Top10perc:**

Table 4: Description of 'Top10perc'

```
Description of Top10perc
-------------------------------------------------------------
count    777.000000
mean      27.558559
std       17.640364
min        1.000000
25%       15.000000
50%       23.000000
75%       35.000000
max       96.000000
Name: Top10perc, dtype: float64 Distribution of Top10perc
```

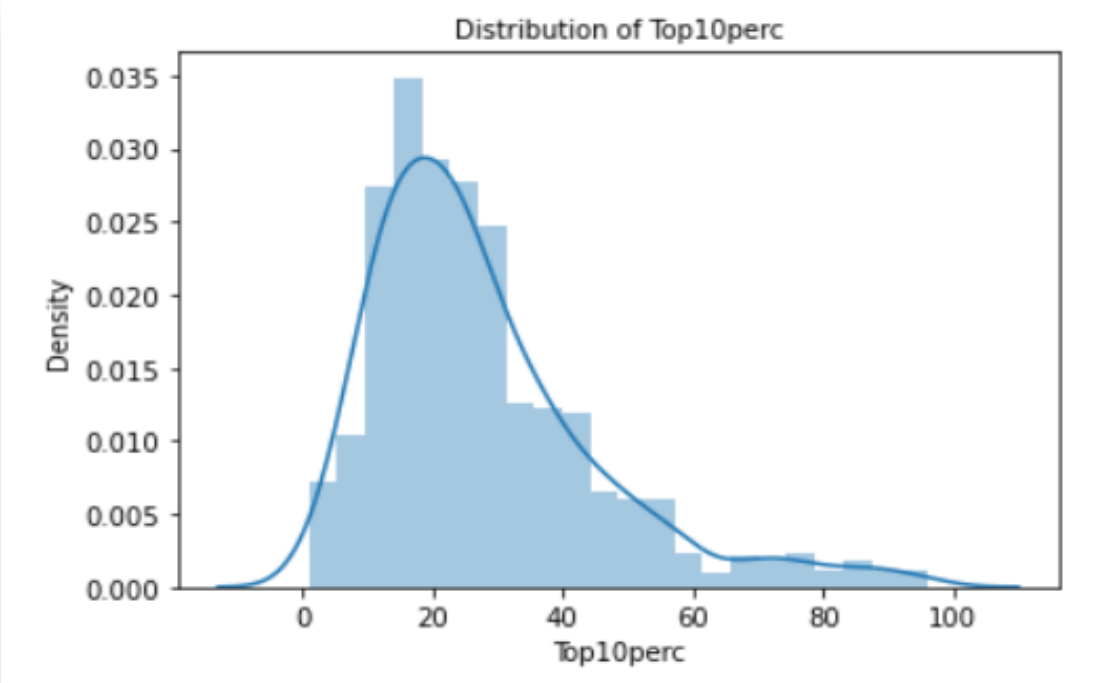

Figure 7: Univariate distribution of 'Top10perc'

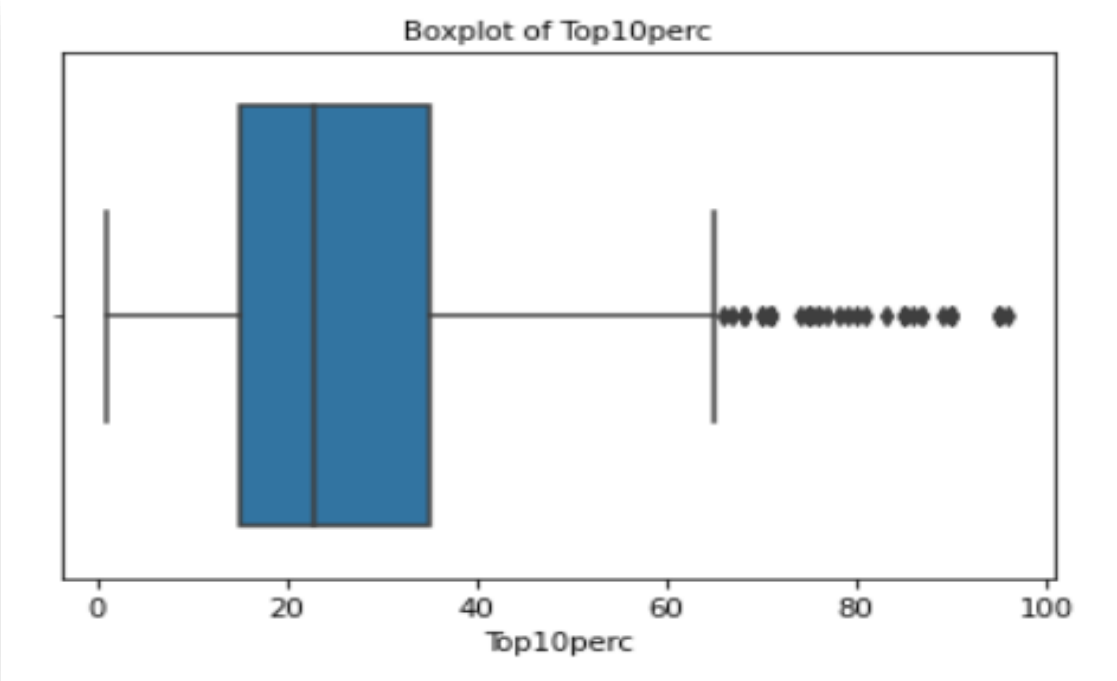

Figure 8: Boxplot showing the distribution of 'Top10perc'

Univariate analysis of 'Top10perc' (top 10 percentage) of higher secondary class is done to understand the patterns and distribution of the data. From figure 9, we can see that the Box plot of 'Top10perc' variable has outliers. The distribution of the 'Top10perc' data is positively skewed which is seen in figure 8. From table 7, it is seen that the intake of students from top 10 percentage of higher secondary class range from 1 to 96.

5. Top25perc:

Table 5: Description of 'Top25perc'

```
Description of Top25perc
----------------------------------------------------------
count    777.000000
mean      55.796654
std       19.804778
min        9.000000
25%       41.000000
50%       54.000000
75%       69.000000
max      100.000000
Name: Top25perc, dtype: float64 Distribution of Top25perc
```

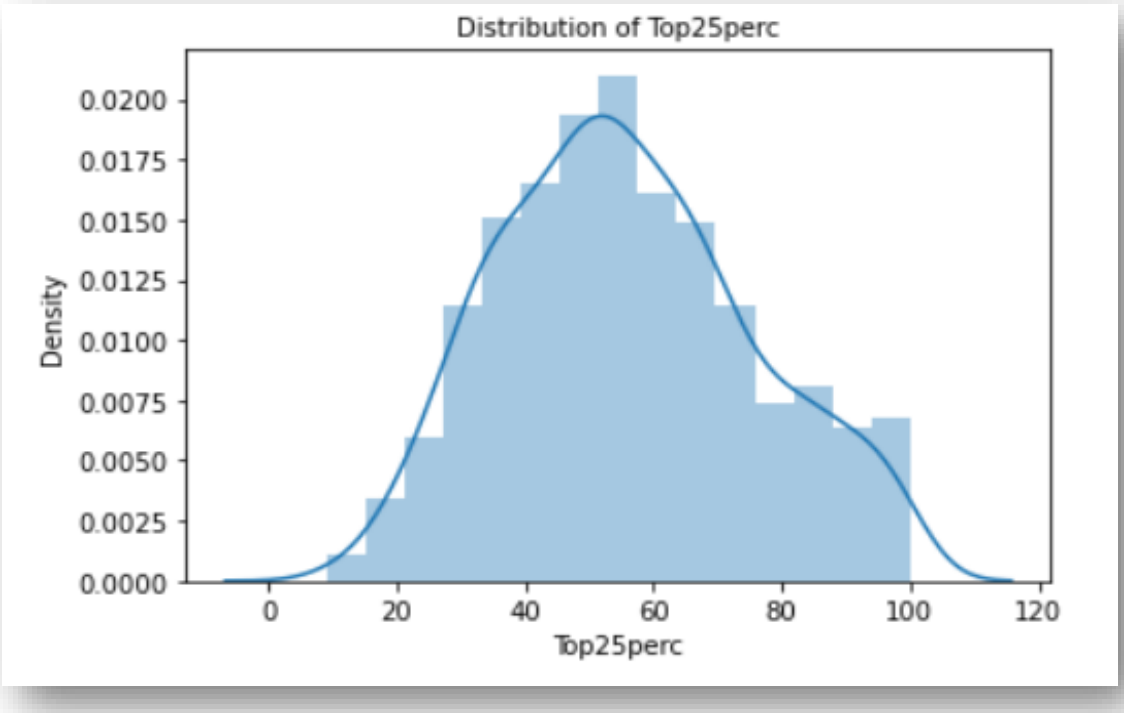

Figure 9: Univariate distribution of 'Top25perc'

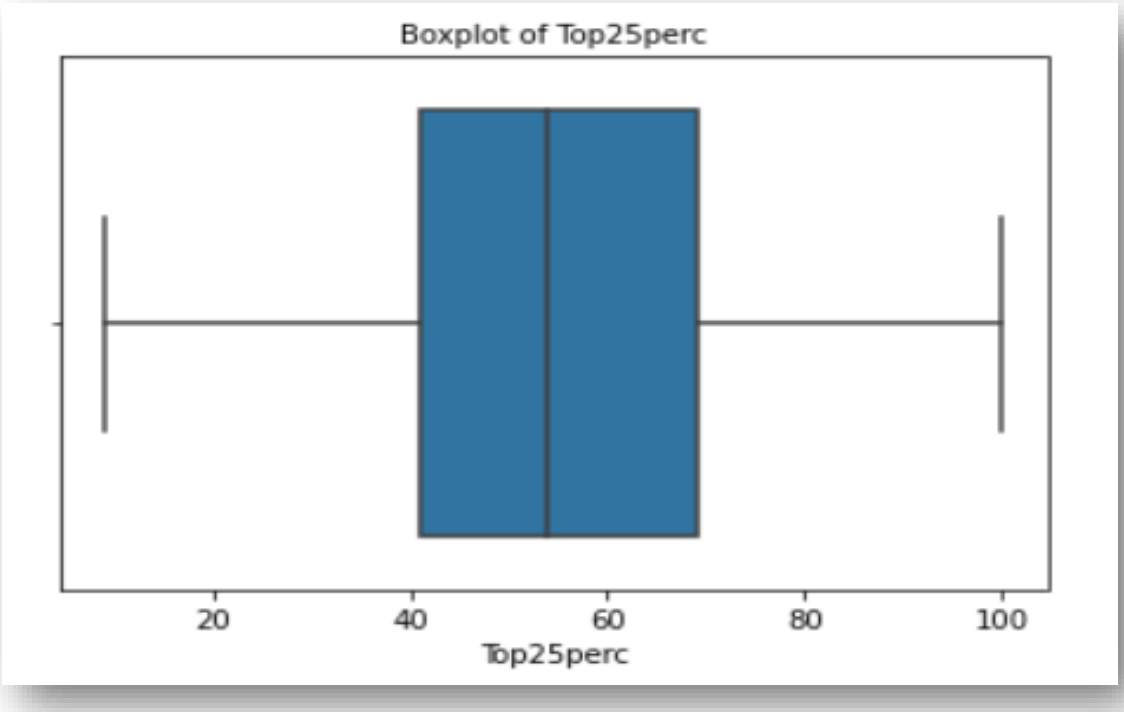

Figure 10: Boxplot showing the distribution of 'Top25perc'

Univariate analysis of 'Top25perc' (top 25 percentage) of higher secondary class is done to understand the patterns and distribution of the data. From figure 11, we can see that the Box plot of 'Top25perc' variable has no outliers. The distribution of the 'Top25perc' data is normally distributed which is seen in figure 10. From table 8, it is seen that the intake of students from top 25 percentage of higher secondary class range from 9 to 100.

6. F.Undergrad:

Table 6: Description of 'F.Undergrad'

```
Description of F.Undergrad
----------------------------------------------------------------------
count      777.000000
mean      3699.907336
std       4850.420531
min        139.000000
25%        992.000000
50%       1707.000000
75%       4005.000000
max      31643.000000
Name: F.Undergrad, dtype: float64 Distribution of F.Undergrad
```

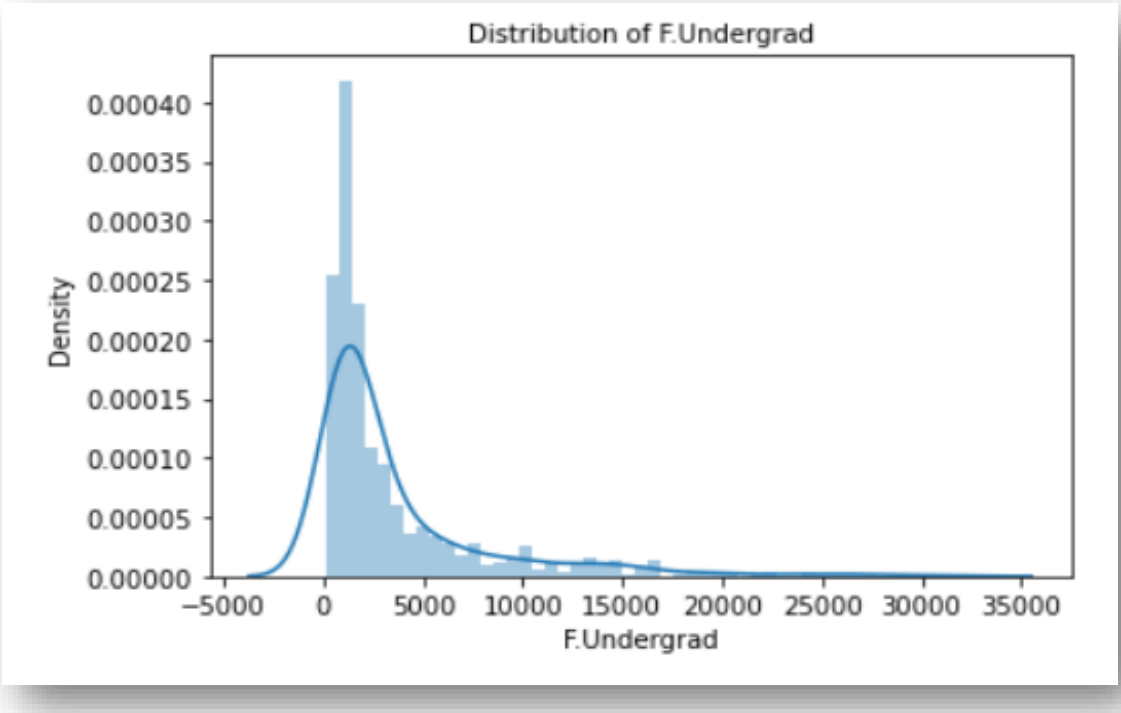

Figure 11: Univariate distribution of 'F. Undergrad'

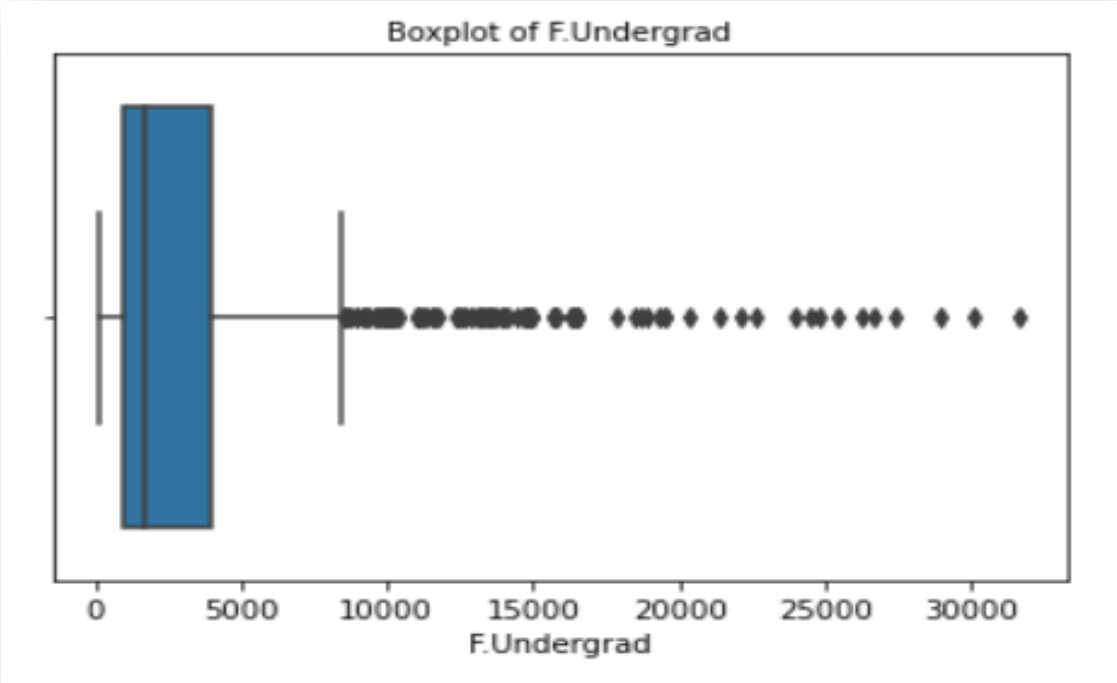

Figure 12: Boxplot showing the distribution of 'F. Undergrad'

Univariate analysis of 'F. Undergrad' (full time graduates) is done to understand the patterns and distribution of the data. From figure 12, we can see that the Box plot of 'F. Undergrad' variable has outliers. The distribution of the 'F. Undergrad' data is positively skewed which is seen in figure 12. From table 9, it is seen that 139 to 31643 full time graduates are studying in all the university.

7. P.Undergrad:

Table 7: Description of 'P.Undergrad'

Description of P.Undergrad	
count	777.000000
mean	855.298584
std	1522.431887
min	1.000000
25%	95.000000
50%	353.000000
75%	967.000000
max	21836.000000

Name: P.Undergrad, dtype: float64 Distribution of P.Undergrad

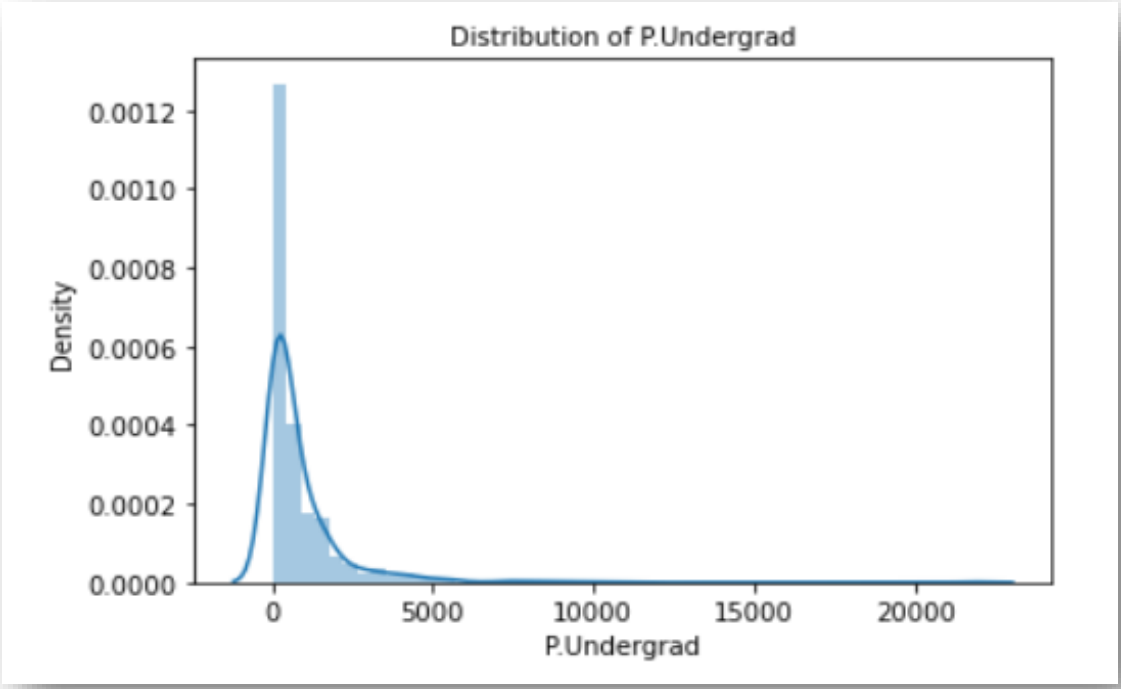

Figure 13: Univariate distribution of 'P. Undergrad'

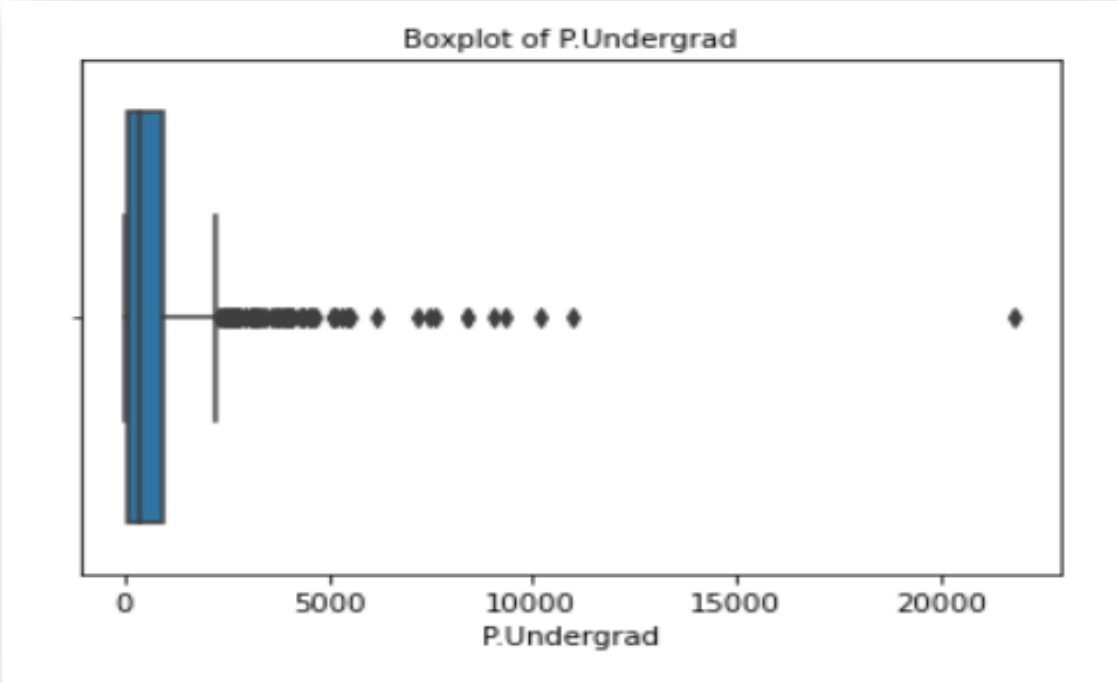

Figure 14: Boxplot showing the distribution of 'P. Undergrad'

Univariate analysis of 'P. Undergrad' (part time graduates) is done to understand the patterns and distribution of the data. From figure 15, we can see that the Box plot of 'P. Undergrad' variable has outliers. The distribution of the 'P. Undergrad' data is positively skewed which is seen in figure 14. From table 10, it is seen that 1 to 21836 part time graduates are studying in all the university.

8. Outstate:

Table 8: Description of 'Outstate'

```
Description of Outstate
-------------------------------------------------------
count      777.000000
mean     10440.669241
std       4023.016484
min       2340.000000
25%       7320.000000
50%       9990.000000
75%      12925.000000
max      21700.000000
Name: Outstate, dtype: float64 Distribution of Outstate
```

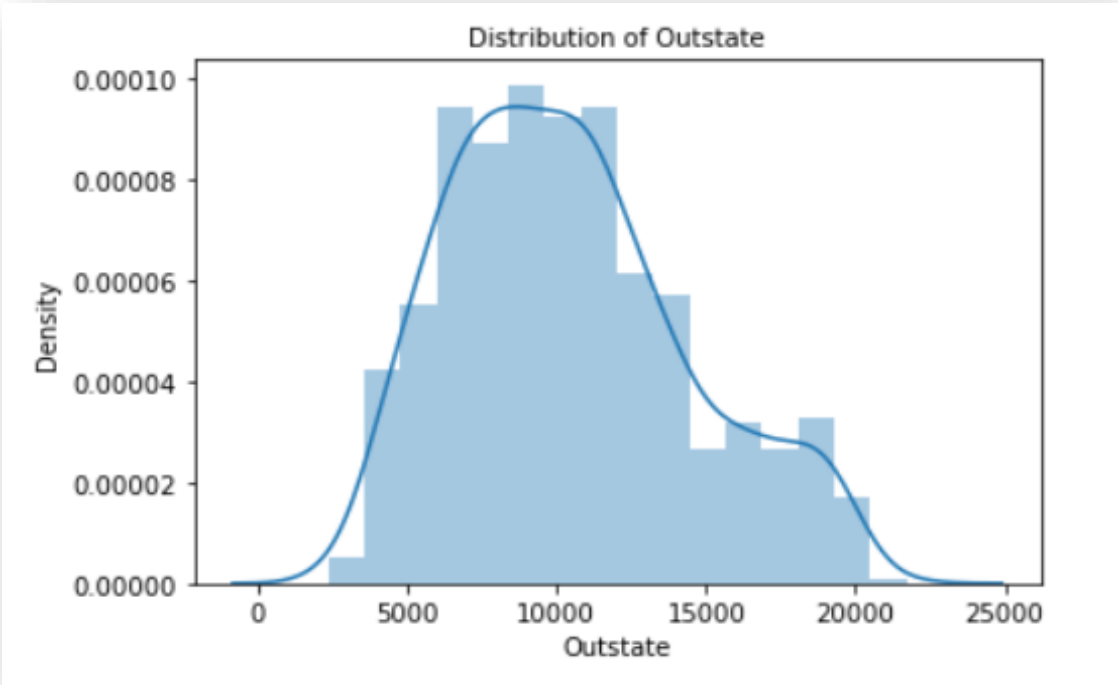

Figure 15: Univariate distribution of 'Outstate'

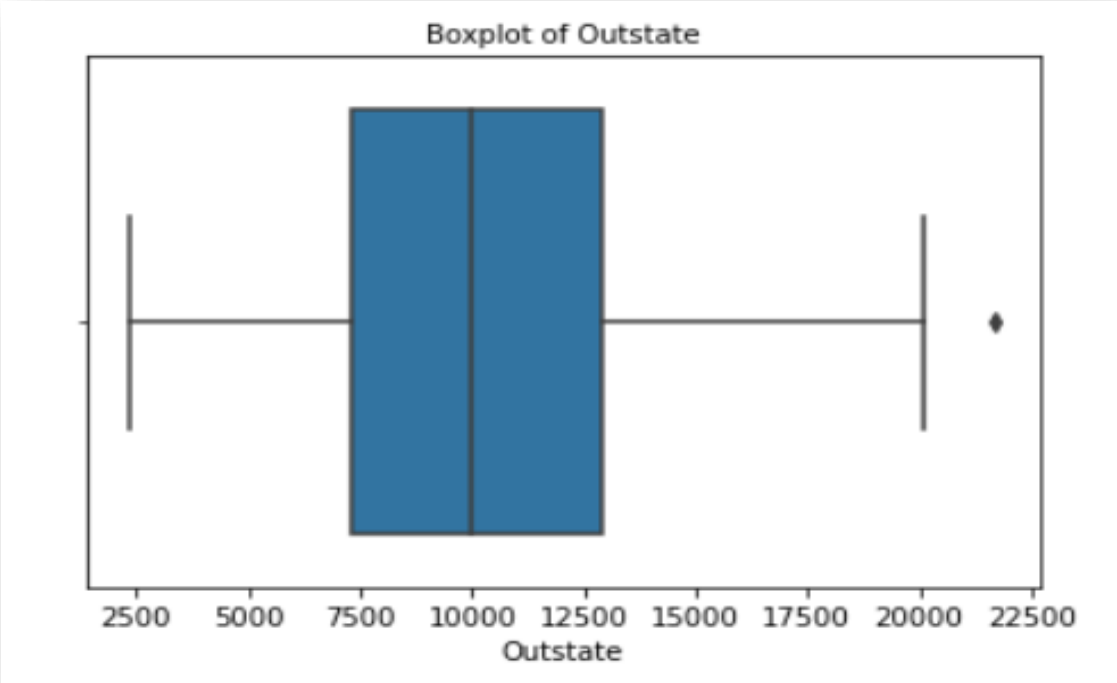

Figure 16: Boxplot showing the distribution of 'Outstate'

Univariate analysis of 'Outstate' is done to understand the patterns and distribution of the data. From figure 17, we can see that the Box plot of 'Outstate' variable has one outlier. The distribution of the 'Outstate' data is normally distributed which is seen in figure 16.

9. Room.Board:

Table 9: Description of 'Room.Board'

Description of Room.Board	
count	777.000000
mean	4357.526384
std	1096.696416
min	1780.000000
25%	3597.000000
50%	4200.000000
75%	5050.000000
max	8124.000000

Name: Room.Board, dtype: float64 Distribution of Room.Board

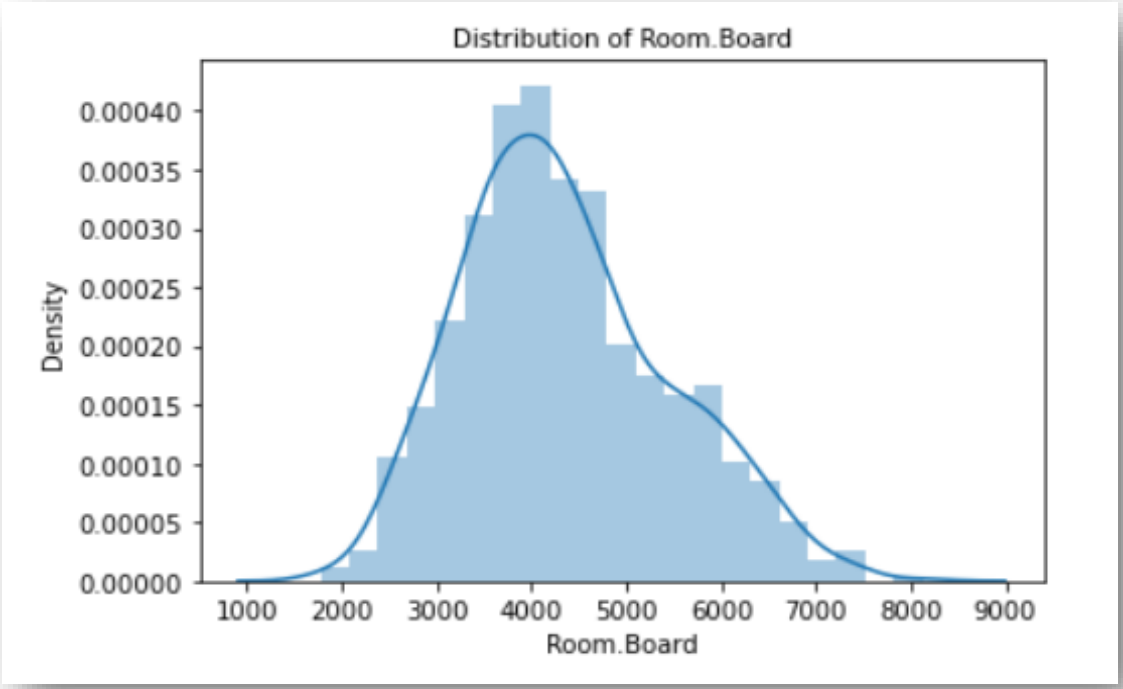

Figure 17: Univariate distribution of 'Room.Board'

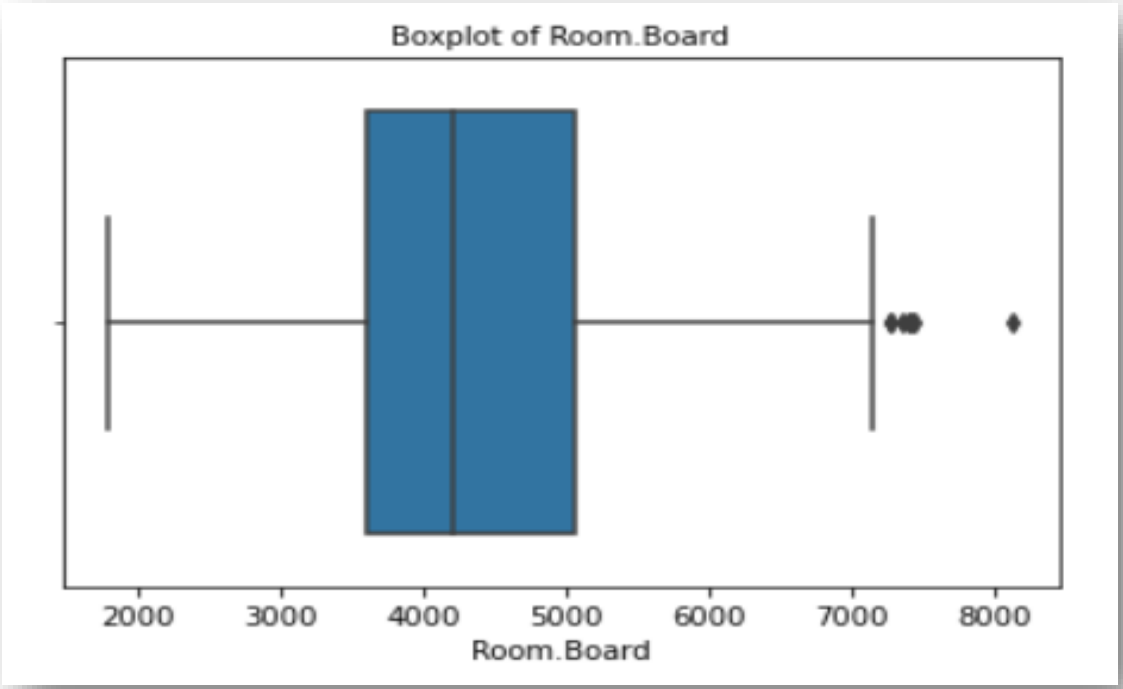

Figure 18: Boxplot showing the distribution of 'Room.Board'

Univariate analysis of 'Room.Board' is done to understand the patterns and distribution of the data. From figure 19, we can see that the Box plot of 'Room.Board' variable has outliers. The distribution of the 'Room.Board' data is normally distributed which is seen in figure 18.

10. Books:

Table 10: Description of 'Books'

```
Description of Books
----------------------------------------------------
count     777.000000
mean      549.380952
std       165.105360
min        96.000000
25%       470.000000
50%       500.000000
75%       600.000000
max      2340.000000
Name: Books, dtype: float64 Distribution of Books
```

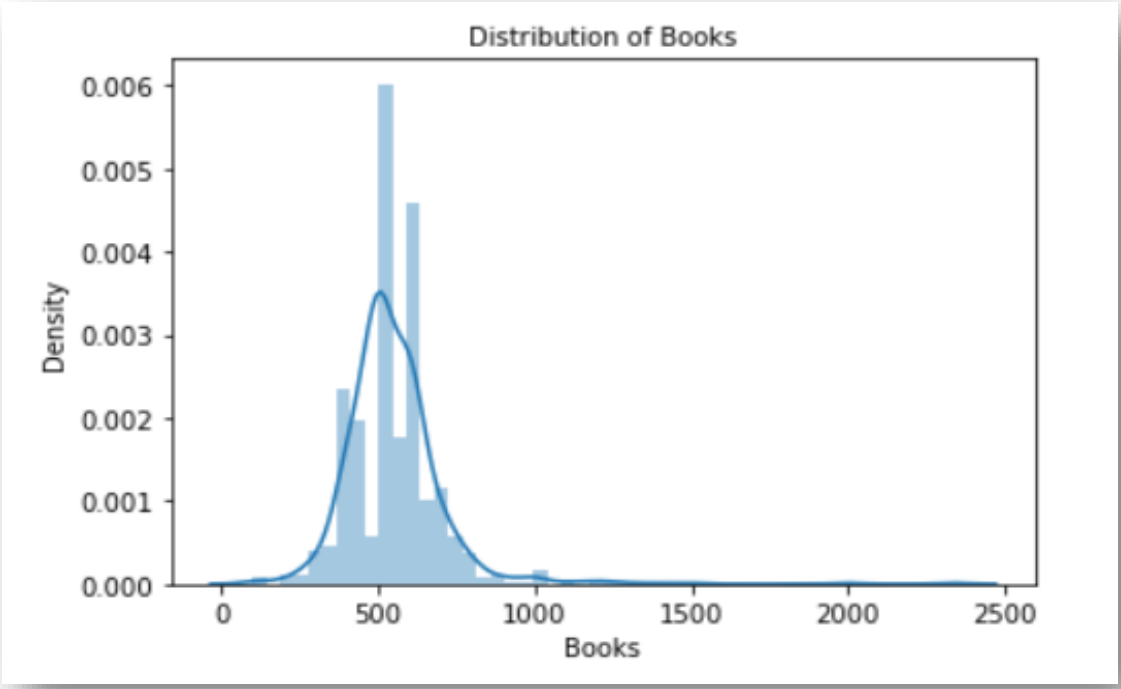

Figure 19: Univariate distribution of 'Books'

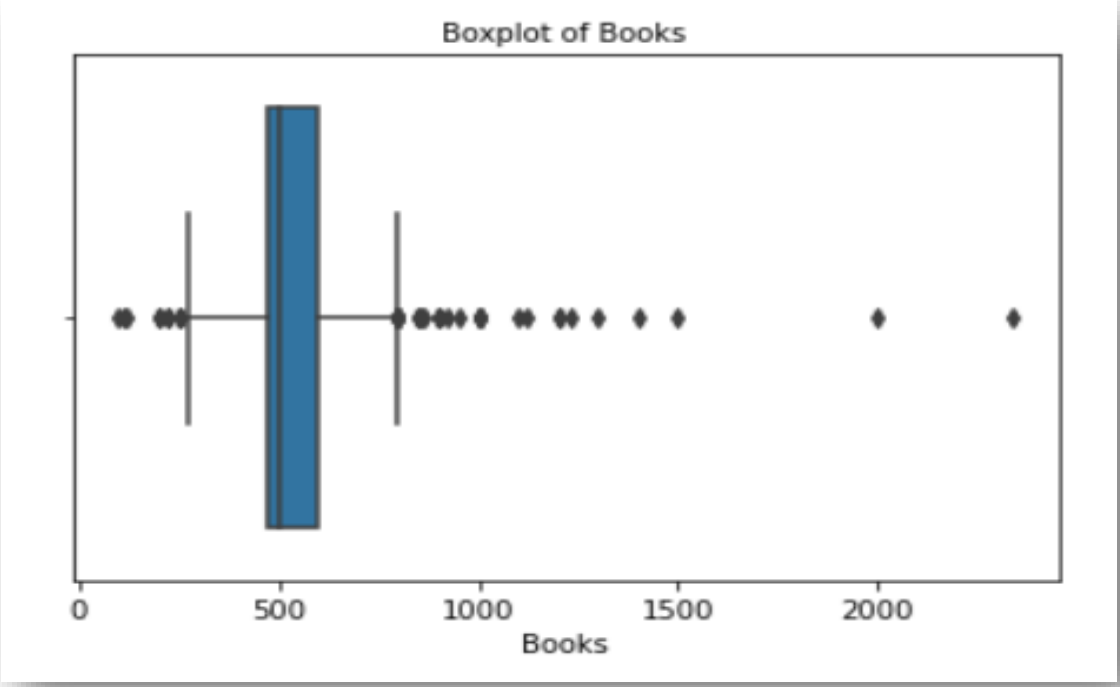

Figure 20: Boxplot showing the distribution of 'Books'

Univariate analysis of 'Books' is done to understand the patterns and distribution of the data. From figure 21 we can see that the Box plot of 'Books' variable has outliers. The distribution of the 'Books' data seems to be bimodal which is seen in figure 20. From table 13, it is seen that the cost of books per student seems to be in the range of 96 to 2340.

11. Personal:

Table 11: Description of 'Personal'

```
Description of Personal
-------------------------------------------------------
count     777.000000
mean     1340.642214
std       677.071454
min       250.000000
25%       850.000000
50%      1200.000000
75%      1700.000000
max      6800.000000
Name: Personal, dtype: float64 Distribution of Personal
```

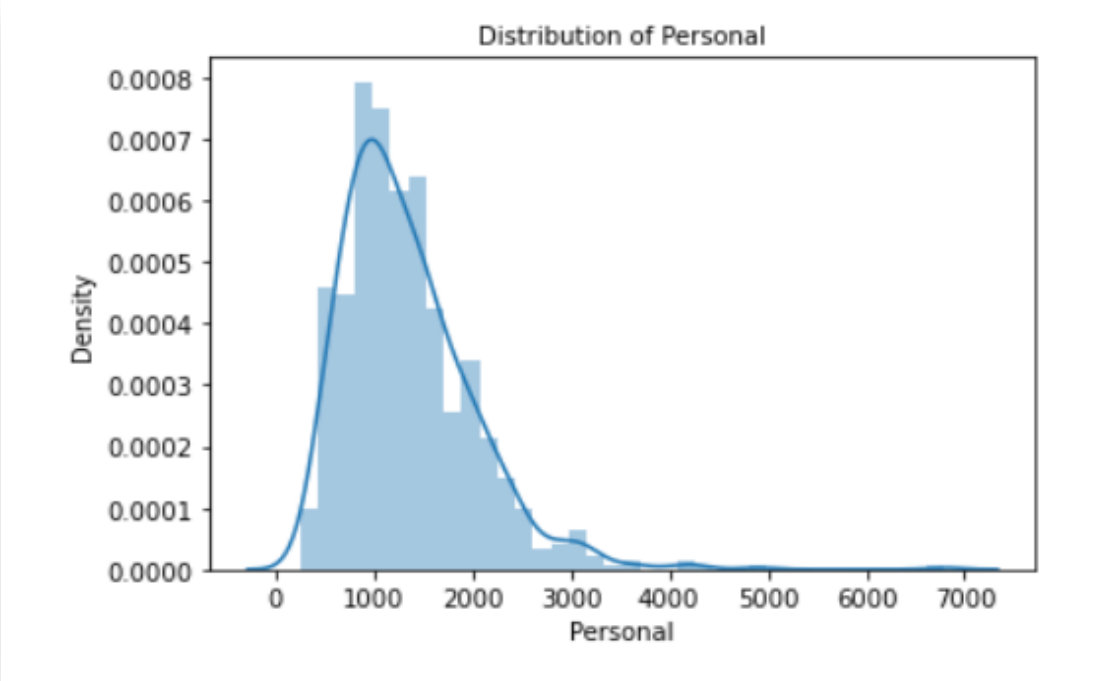

Figure 21: Univariate distribution of 'Personal'

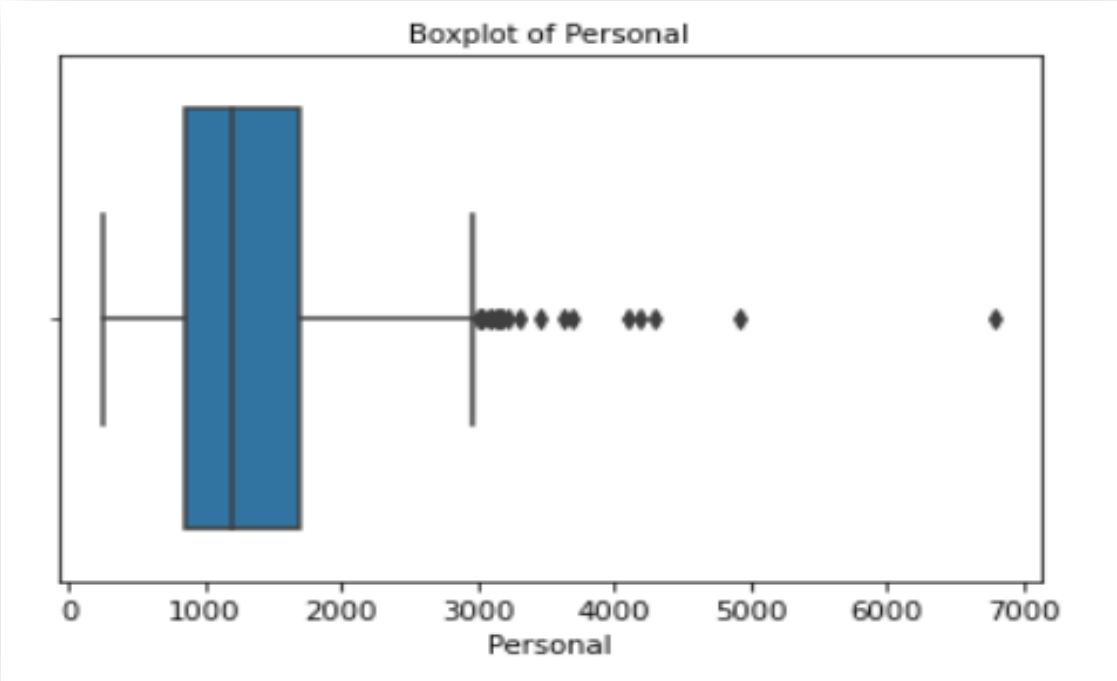

Figure 22: Boxplot showing the distribution of 'Personal'

Univariate analysis of ‘Personal’ is done to understand the patterns and distribution of the data. From figure 23, we can see that the Box plot of ‘Personal’ variable has outliers as some student’s personal expense are way bigger than the rest of the students. The distribution of the ‘Personal’ data is positively skewed which is seen in figure 22. From table 10, it is seen that the expense ranges from 250 to 6800.

12. PhD:

Table 12: Description of ‘PhD’

```
Description of PhD
-----------------------------------------------
count    777.000000
mean      72.660232
std       16.328155
min        8.000000
25%       62.000000
50%       75.000000
75%       85.000000
max      103.000000
Name: PhD, dtype: float64 Distribution of PhD
```

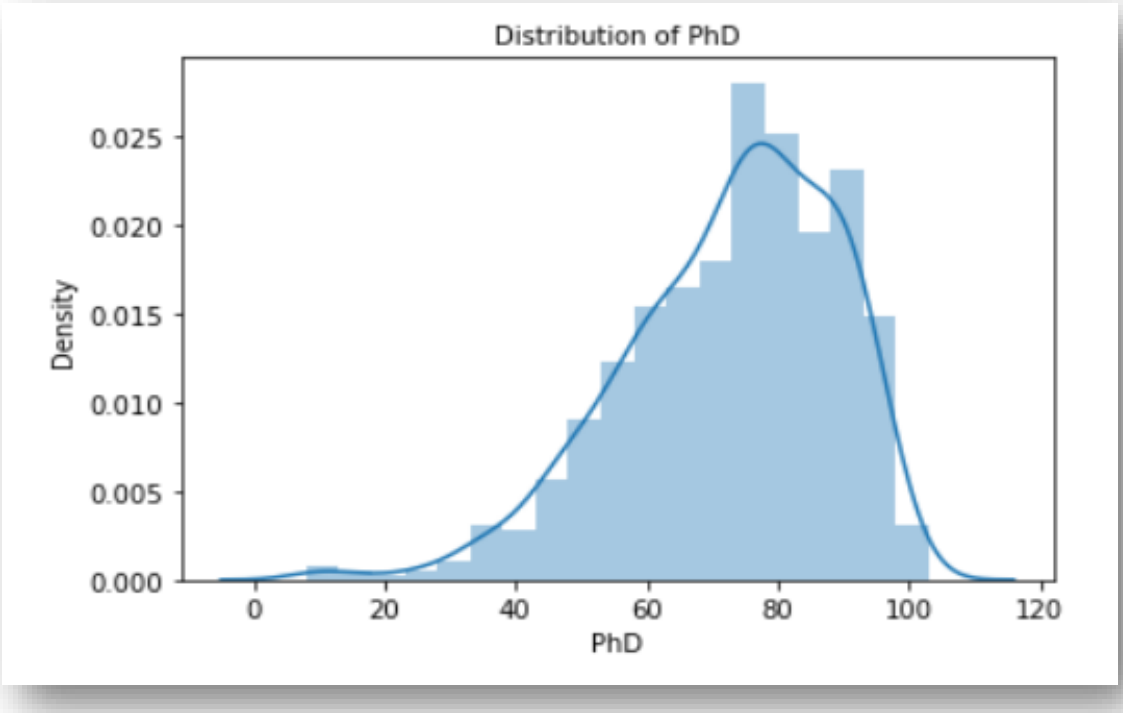

Figure 23: Univariate distribution of ‘PhD’

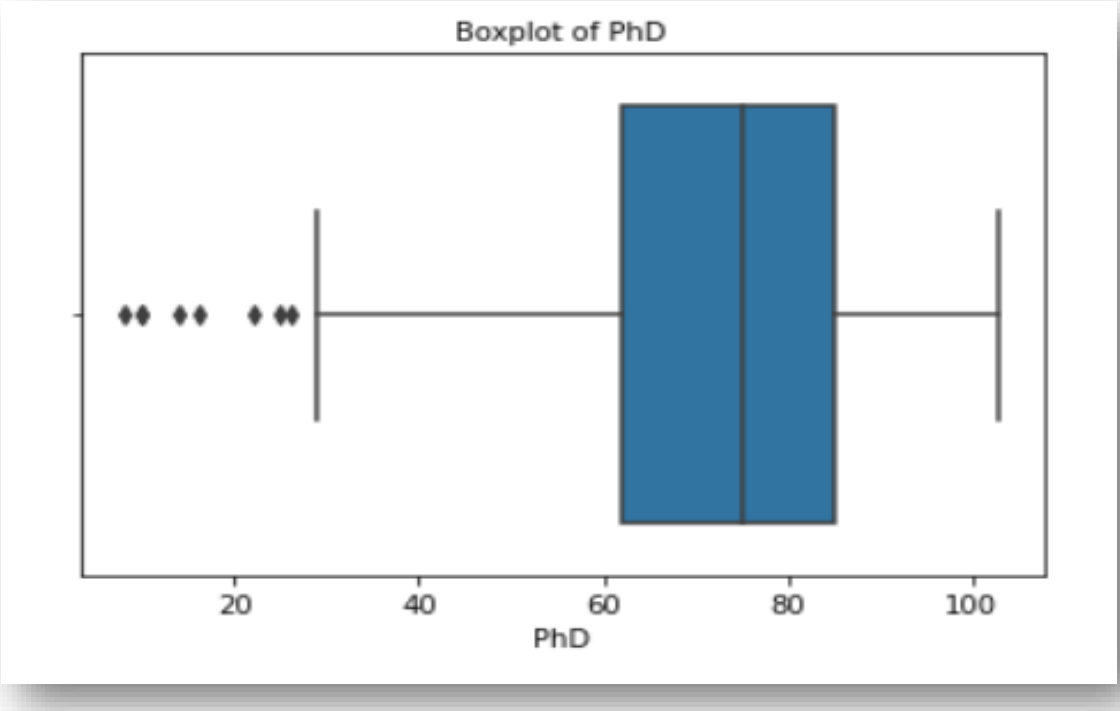

Figure 24: Boxplot showing the distribution of 'PhD'

Univariate analysis of 'PhD' is done to understand the patterns and distribution of the data. From figure 25, we can see that the Box plot of 'PhD' variable has outliers. The distribution of the 'PhD' data is negatively skewed which is seen in figure 24 and 25.

13. Terminal:

Table 13: Description of 'Terminal'

```
Description of Terminal
-----------------------------------------------------------
count    777.000000
mean      79.702703
std       14.722359
min       24.000000
25%       71.000000
50%       82.000000
75%       92.000000
max      100.000000
Name: Terminal, dtype: float64 Distribution of Terminal
```

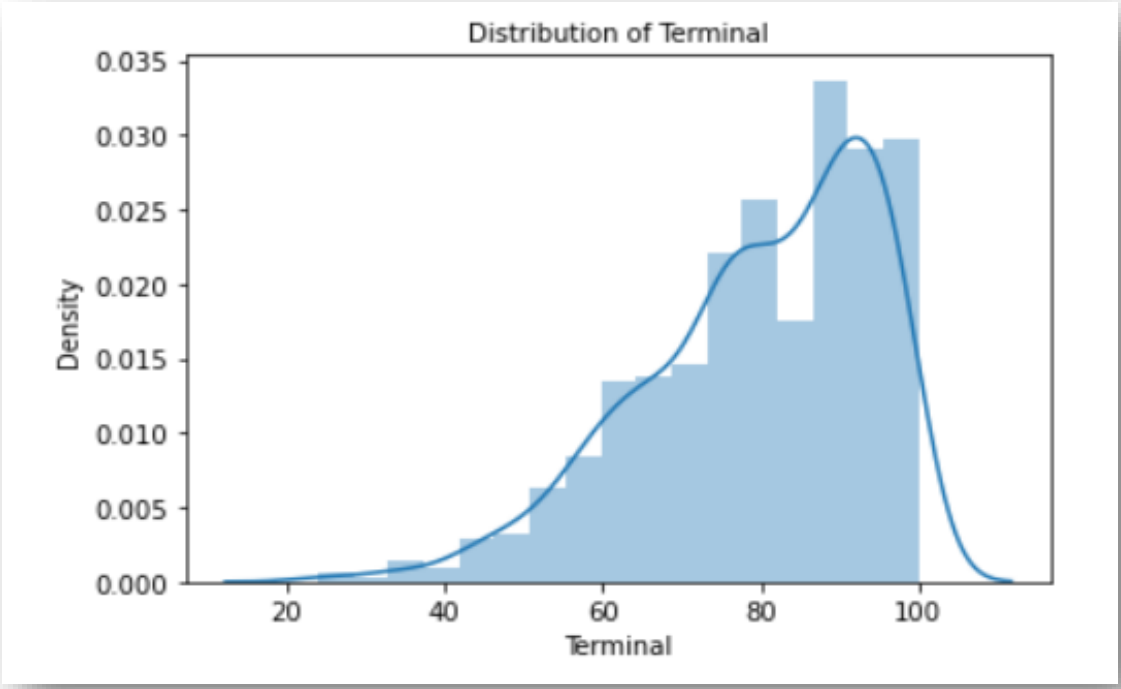

Figure 25: Univariate distribution of 'Terminal'

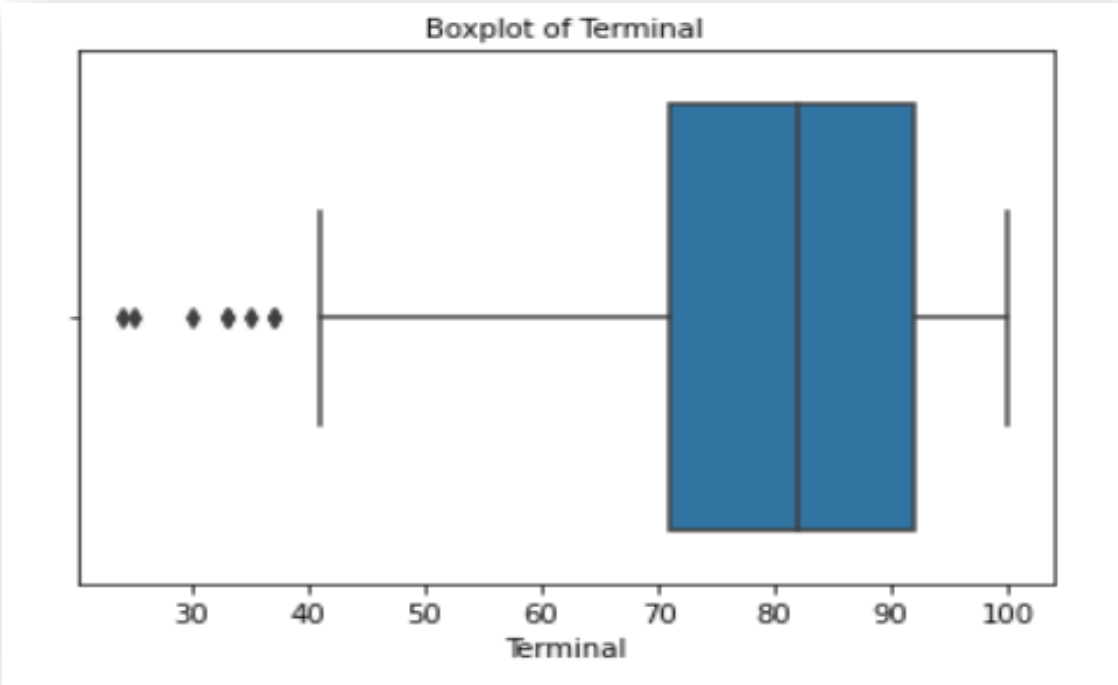

Figure 26: Boxplot showing the distribution of 'Terminal'

Univariate analysis of 'Terminal' is done to understand the patterns and distribution of the data. From figure 27, we can see that the Box plot of 'Terminal' variable has outliers. The distribution of the 'Terminal' data is negatively skewed which is seen in figure 26 and 27.

14. S.F.Ratio:

Table 14: Description of 'S.F.Ratio'

```
Description of S.F.Ratio
----------------------------------------------------------
count    777.000000
mean      14.089704
std        3.958349
min        2.500000
25%       11.500000
50%       13.600000
75%       16.500000
max       39.800000
Name: S.F.Ratio, dtype: float64 Distribution of S.F.Ratio
```

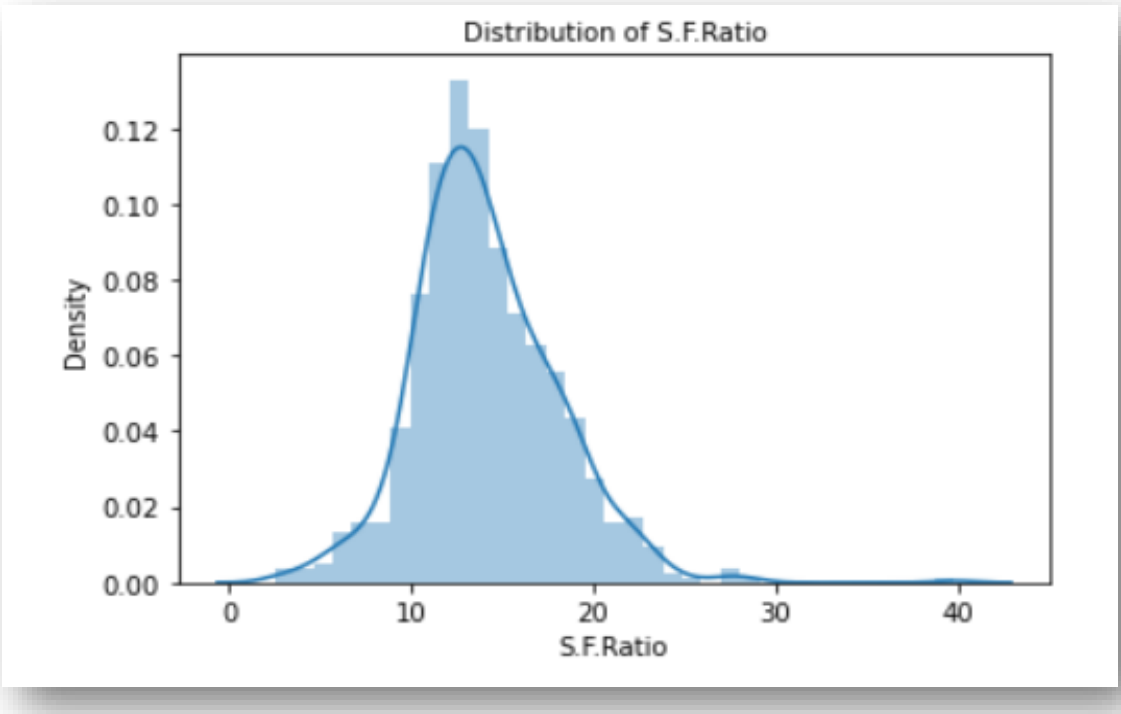

Figure 27: Univariate distribution of 'S.F.Ratio'

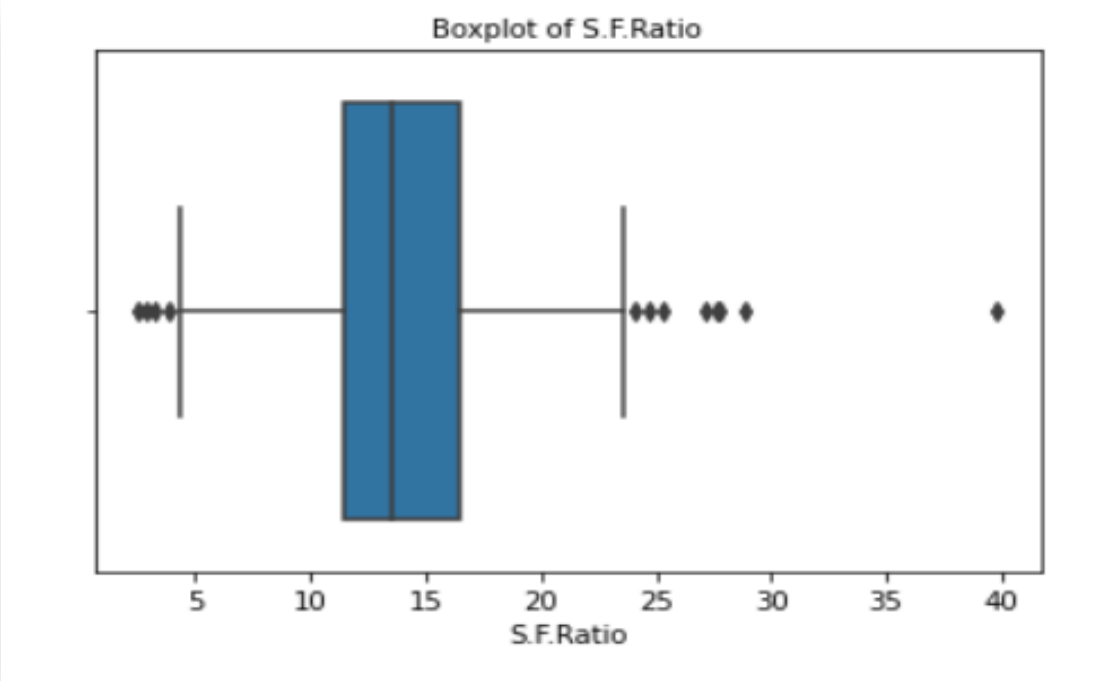

Figure 28: Boxplot showing the distribution of 'S.F.Ratio'

Univariate analysis of 'S.F.Ratio' (student faculty ratio) is done to understand the patterns and distribution of the data. From figure 29, we can see that the Box plot of 'S.F.Ratio' variable has outliers. The distribution of the 'S.F.Ratio' data is almost normally distributed which is seen in figure 28. The student faculty ratio is almost same in all the university and colleges.

15. perc.alumni:

Table 15: Description of 'perc.alumni'

```
Description of perc.alumni
-----------------------------------------------------------
count    777.000000
mean      22.743887
std       12.391801
min        0.000000
25%       13.000000
50%       21.000000
75%       31.000000
max       64.000000
Name: perc.alumni, dtype: float64 Distribution of perc.alumni
```

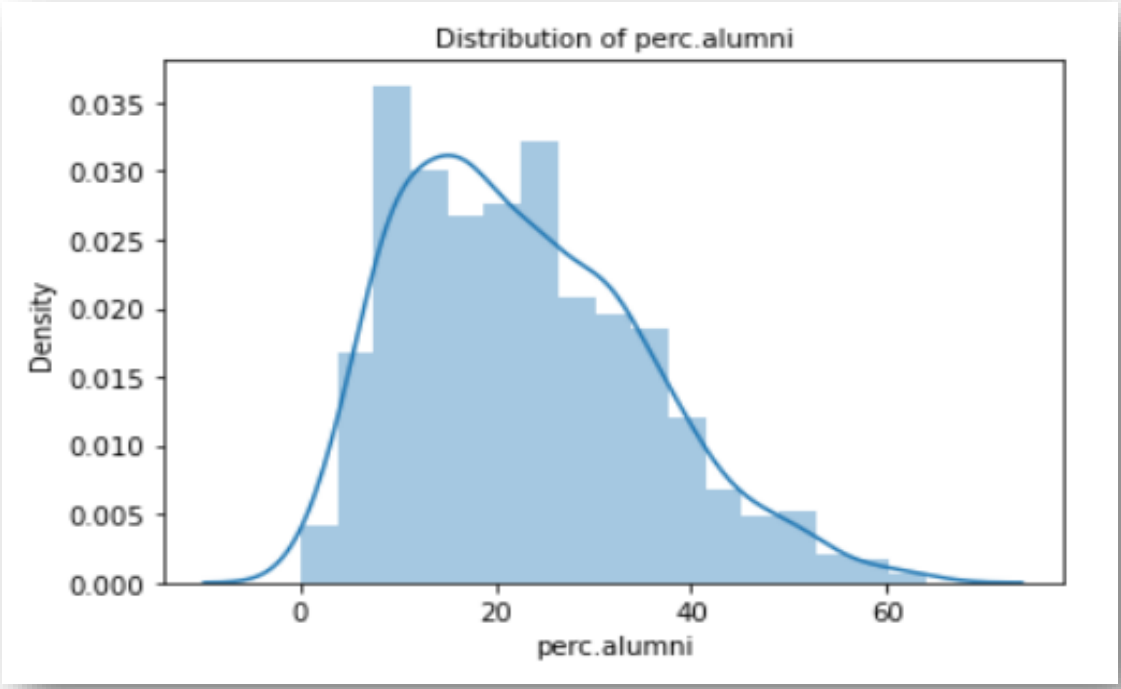

Figure 29: Univariate distribution of 'perc.alumni'

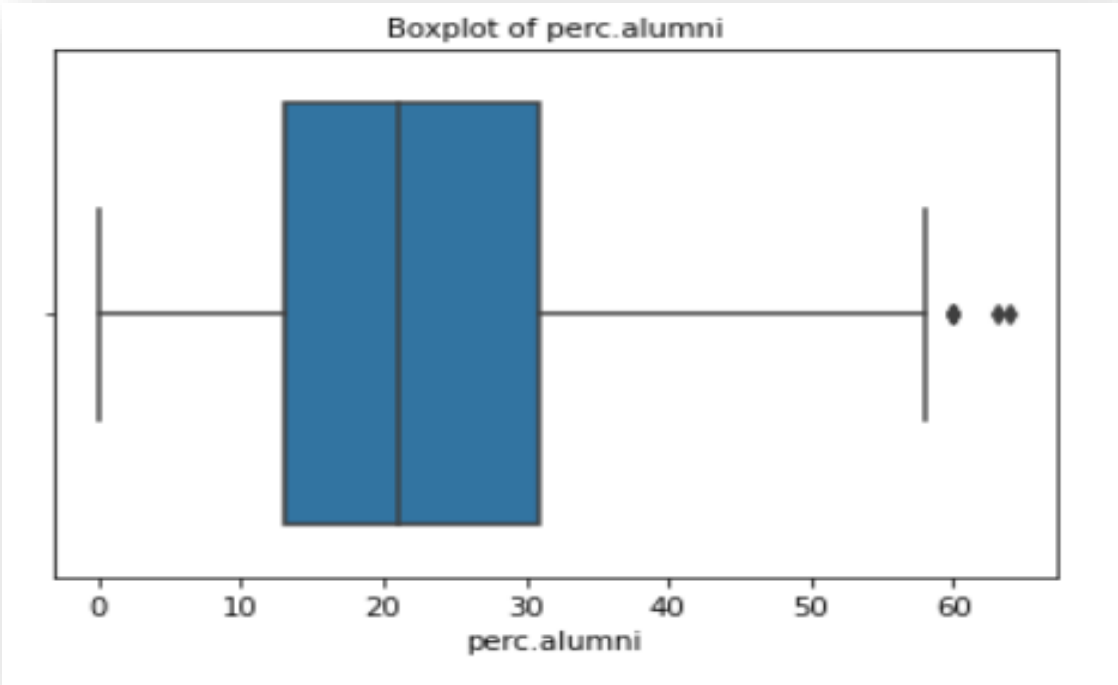

Figure 30: Boxplot showing the distribution of 'perc.alumni'

Univariate analysis of ‘perc.alumni’ (percentage of alumni) is done to understand the patterns and distribution of the data. From figure 31, we can see that the Box plot of ‘perc.alumni’ variable has outliers. The distribution of the ‘perc.alumni’ data is almost normally distributed which is seen in figure 30.

16. Expend:

Table 16: Description of ‘Expend’

```
Description of Expend
-----------------------------------------------------
count      777.000000
mean      9660.171171
std       5221.768440
min       3186.000000
25%       6751.000000
50%       8377.000000
75%      10830.000000
max      56233.000000
Name: Expend, dtype: float64 Distribution of Expend
```

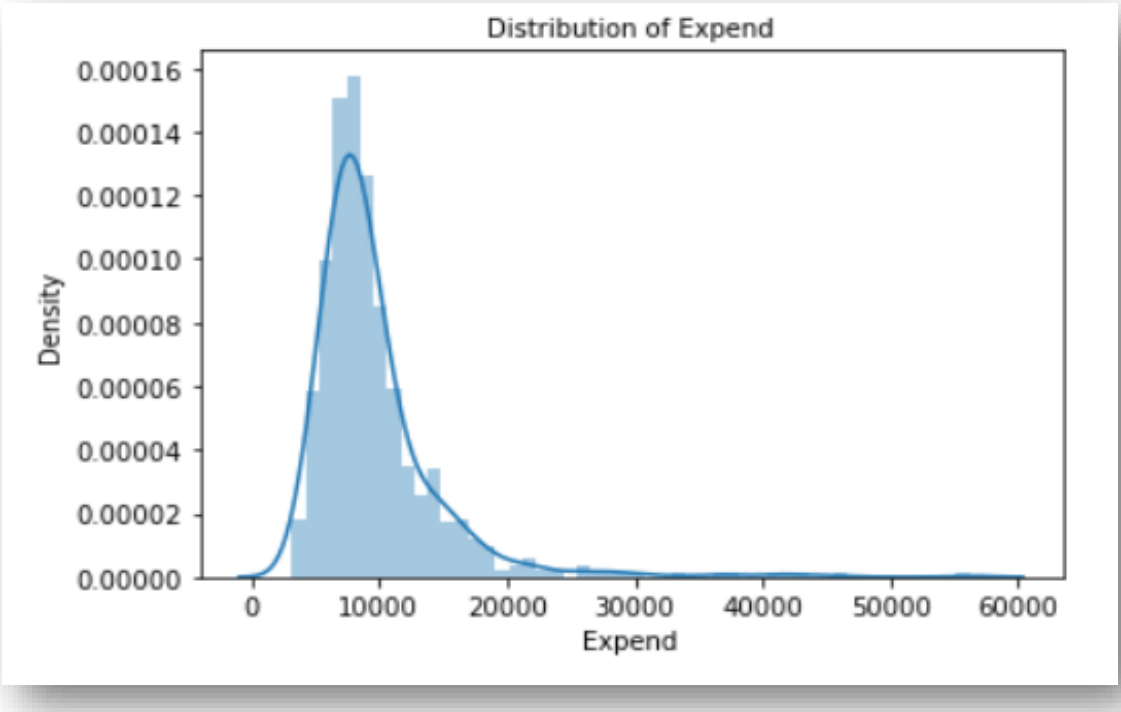

Figure 31: Univariate distribution of ‘Expend’

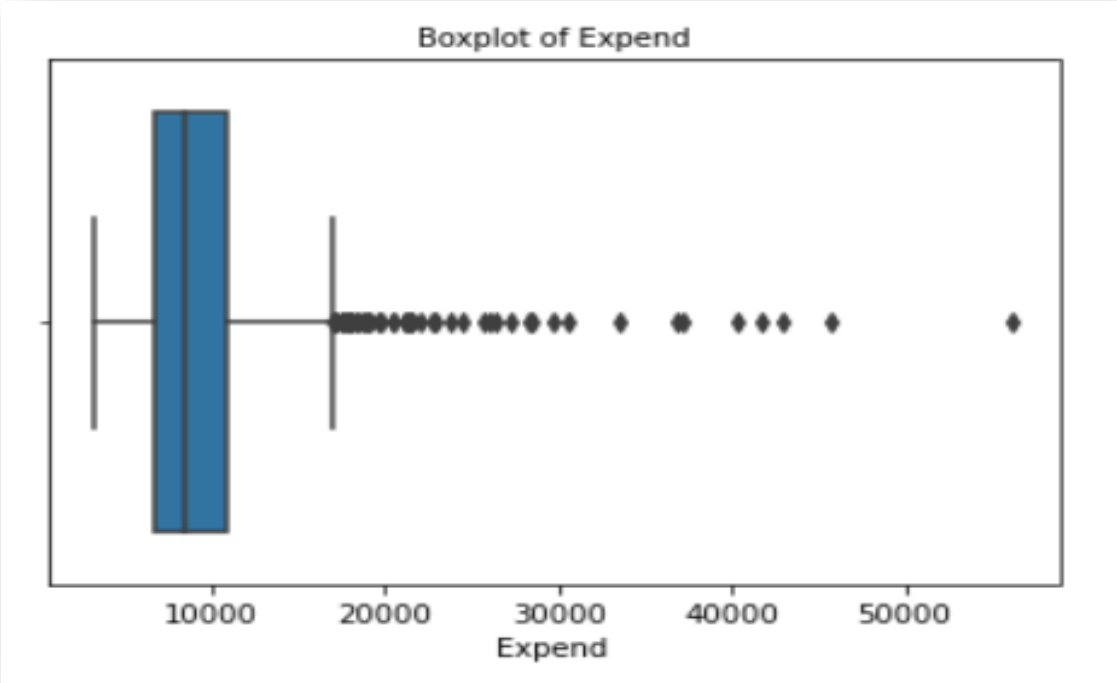

Figure 32: Boxplot showing the distribution of 'Expend'

Univariate analysis of 'Expend' (expenditure) is done to understand the patterns and distribution of the data. From figure 33, we can see that the Box plot of 'Expend' variable has outliers. The distribution of the 'Expend' data is positively skewed which is seen in figure 32. From table 19, it is seen that the expenditure ranges from 3186 to 56233.

17. Grad.Rate:

Table 17: Description of 'Grad.Rate'

Description of Grad.Rate	
count	777.00000
mean	65.46332
std	17.17771
min	10.00000
25%	53.00000
50%	65.00000
75%	78.00000
max	118.00000

Name: Grad.Rate, dtype: float64 Distribution of Grad.Rate

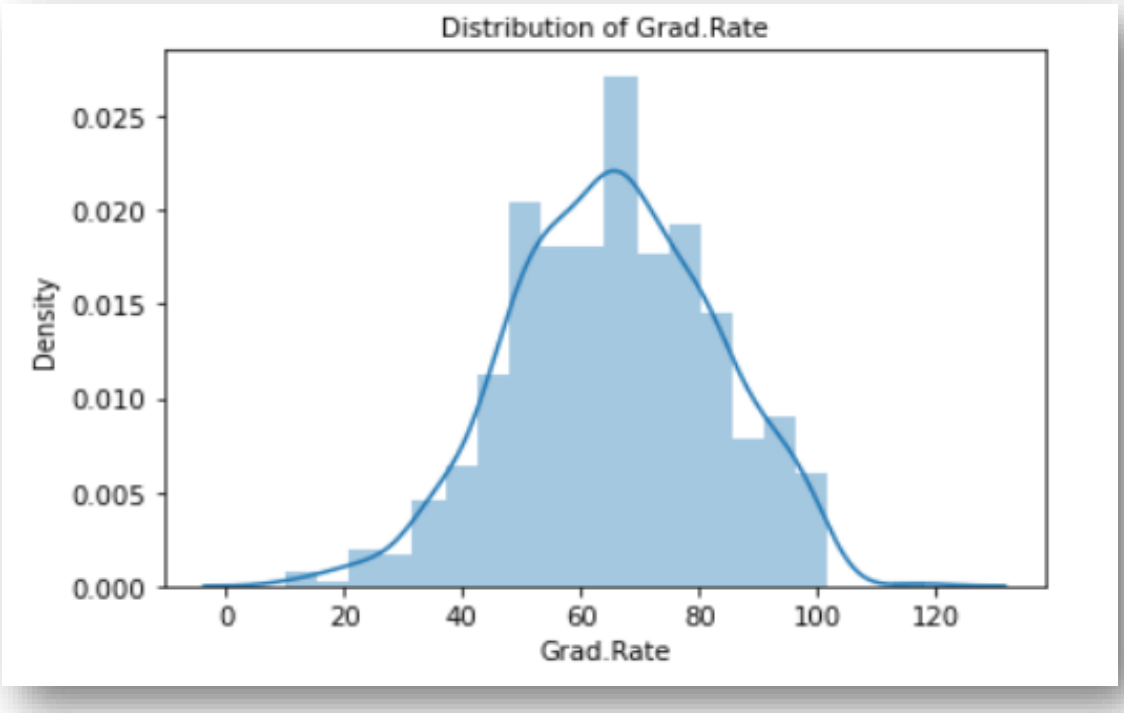

Figure 33: Univariate distribution of 'Grad.Rate'

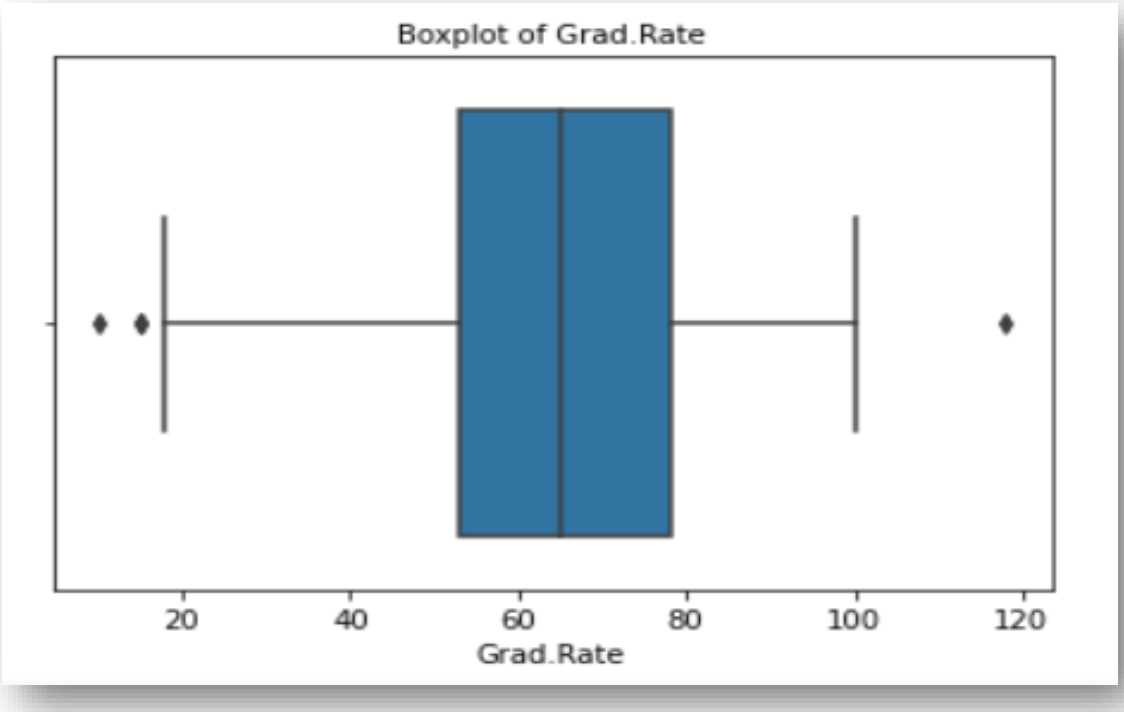

Figure 34: Boxplot showing the distribution of 'Grad.Rate'

Univariate analysis of 'Grad.Rate' (graduation rate) is done to understand the patterns and distribution of the data. From figure 35, we can see that the Box plot of 'Grad.Rate' variable has outliers. The distribution of the 'Grad.Rate' data is normally distributed which is seen in figure 34. From table 20, it is seen that the graduation rate ranges from 10 to 118.

Bivariate Analysis:

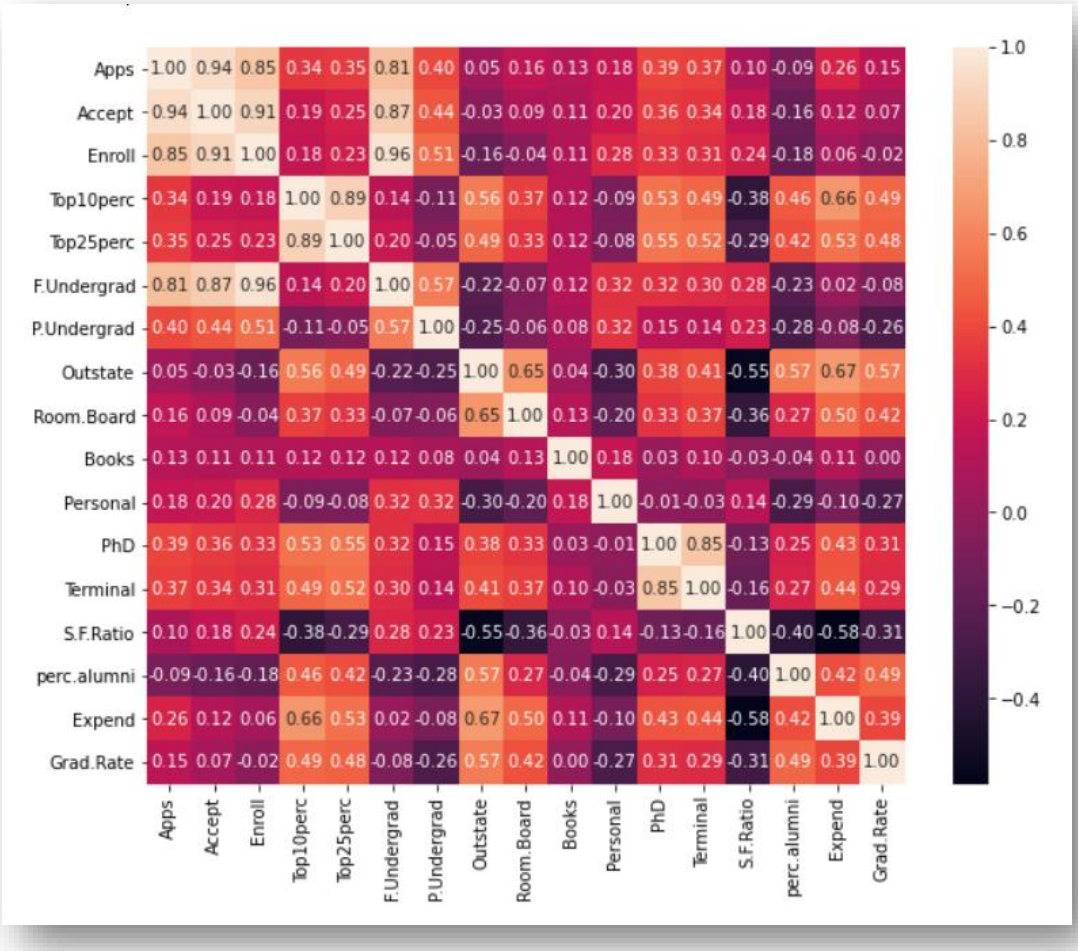

Figure 35: Heat map showing the bivariate analysis of the dataset

Bivariate analysis is done using the help of a heat map. A heat map is used to understand the correlation between two numerical values in a dataset. Figure 36 shows the heat map of the dataset.

Multivariate Analysis:

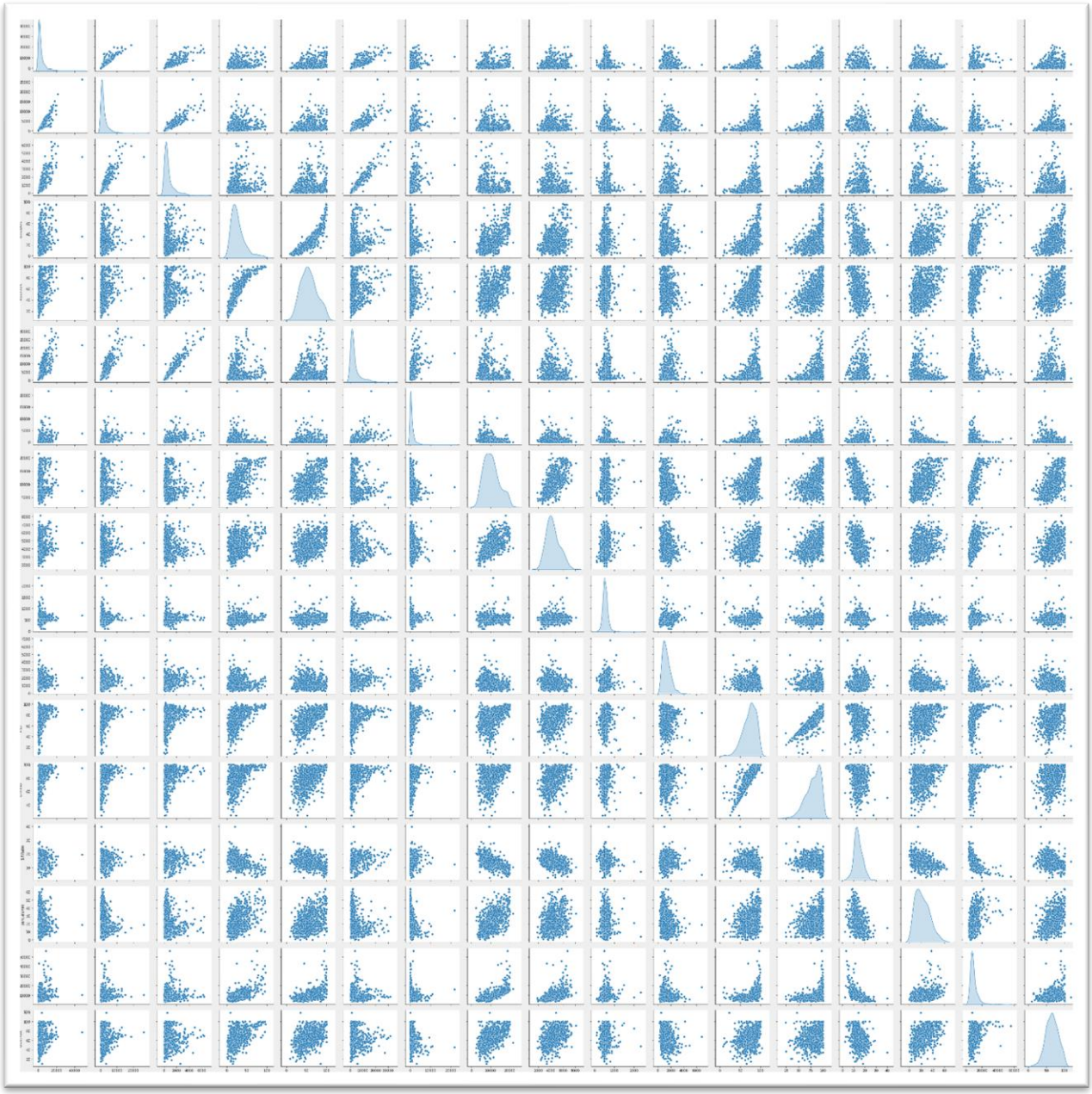

Figure 36: Pair plot showing the multivariate analysis of the dataset

Multivariate analysis is done using the help of a pair plot to understand the relationship between all the numerical values in the dataset. Pair plot can be used to compare all the variables with each other to understand the patterns or trends in the dataset. Figure 37 shows the pair plot of the dataset.

Insights:

From the exploratory data analysis (EDA) done above, we could understand that the 'Apps' (application) variable is highly positively correlated with 'Accept' (application accepted), 'Enroll' (students enrolled) and 'F.Undergrad' (full time graduates). So this relationship gives the insights on when student submits the application it is accepted and the student is enrolled as fulltime graduate. We can find negative correlation between 'Apps' (application) and 'perc.alumni' (percentage of

alumni). This indicates us not all students are part of alumni of their college or university. The application with ‘Top10perc’, ‘Top25perc’ of higher secondary class, ‘Outstate’, ‘Room.Board’, ‘Books’, ‘Personal’, ‘PhD’, ‘Terminal’, ‘S.F.Ratio’, ‘Expend’ and ‘Grad.Rate’ are all positively correlated.

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

Scaling is a way of representing a dataset. Scaling needs to be done as a dataset has features or variables with different ‘weights’ for each feature. In such cases, it is suggested to transform the features so that all the features are in the same ‘scale’. This is called scaling. Scaling also makes it easier to compare the features now that the weightage on each feature is the same.

Scaling can be done by Z-score method and Min-max method. Z-score method is used when you want to centralize the data (weight-based). Example: principal component analysis (PCA), neural network etc., Min-max method is used when the data is distance based. Example: Clustering, KNN etc., Models have to be implemented only after scaling is done.

In the given dataset, there is 1 categorical variable and 17 numerical variables. Therefore the categorical variable ‘Name’ is dropped before performing scaling. Scaling is done on the 17 numerical variables. Among the 17 numerical variables, application, accepted application, enrolled, full time graduates, part-time graduates, outstate are the variables in which the values represent the number of students. The top10 percent and top25 percent are students in which the values are given in percentage. Room board, books, and personal are variables in which the values are associated with money. The phd, sf ratio, percentage of alumni are percentage values of different combinations of students teachers alumini these are percentage values. The graduation rate is also percentage value of graduates who get graduated every year.

Z-score method is used for scaling this dataset.

Table 18: Description of the dataset before scaling

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.380952	165.105360	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.171171	5221.768440	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.463320	17.177710	10.0	53.0	65.0	78.0	118.0

Table 19: Description of the dataset after scaling using Z-score method

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	6.355797e-17	1.000644	-0.755134	-0.575441	-0.373254	0.160912	11.658671
Accept	777.0	6.774575e-17	1.000644	-0.794764	-0.577581	-0.371011	0.165417	9.924816
Enroll	777.0	-5.249269e-17	1.000644	-0.802273	-0.579351	-0.372584	0.131413	6.043678
Top10perc	777.0	-2.753232e-17	1.000644	-1.506526	-0.712380	-0.258583	0.422113	3.882319
Top25perc	777.0	-1.546739e-16	1.000644	-2.364419	-0.747607	-0.090777	0.667104	2.233391
F.Undergrad	777.0	-1.661405e-16	1.000644	-0.734617	-0.558643	-0.411138	0.062941	5.764674
P.Undergrad	777.0	-3.029180e-17	1.000644	-0.561502	-0.499719	-0.330144	0.073418	13.789921
Outstate	777.0	6.515595e-17	1.000644	-2.014878	-0.776203	-0.112095	0.617927	2.800531
Room.Board	777.0	3.570717e-16	1.000644	-2.351778	-0.693917	-0.143730	0.631824	3.436593
Books	777.0	-2.192583e-16	1.000644	-2.747779	-0.481099	-0.299280	0.306784	10.852297
Personal	777.0	4.765243e-17	1.000644	-1.611860	-0.725120	-0.207855	0.531095	8.068387
PhD	777.0	5.954768e-17	1.000644	-3.962596	-0.653295	0.143389	0.756222	1.859323
Terminal	777.0	-4.481615e-16	1.000644	-3.785982	-0.591502	0.156142	0.835818	1.379560
S.F.Ratio	777.0	-2.057556e-17	1.000644	-2.929799	-0.654660	-0.123794	0.609307	6.499390
perc.alumni	777.0	-6.022638e-17	1.000644	-1.836580	-0.786824	-0.140820	0.666685	3.331452
Expend	777.0	1.213101e-16	1.000644	-1.240641	-0.557483	-0.245893	0.224174	8.924721
Grad.Rate	777.0	3.886495e-16	1.000644	-3.230876	-0.726019	-0.026990	0.730293	3.060392

From table 21 and 22, we are able to see how the values have changed in scale. The values may seem to be different but however they are only scaled. The dataset is brought into one unit of comparison. To prove this, a histogram is plotted.

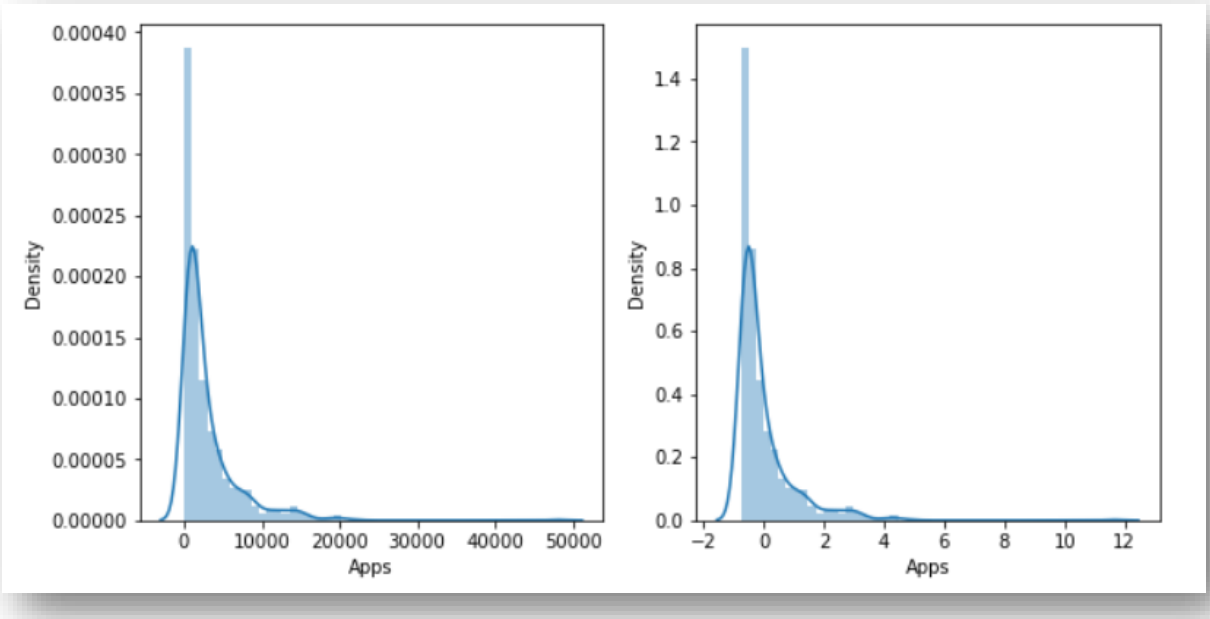

Figure 37: Graph showing the histogram of 'Apps' variable before scaling (left) and the histogram of 'Apps' variable after scaling (right)

From figure 38, we can see that the plot of the 'Apps' variable before and after scaling is the same but the scale has changed. This is the case for all the other numerical variables that has been scaled using the Z-score method.

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

Covariance and correlation matrices are both measures of the relationship and the dependency between two variables.

Covariance indicates the direction of the linear relationship between the variables whether it is positive or negative. By direction means it is directly proportional or inversely proportional.

Correlation measures both the strength and direction of the linear relationship between two variables. Correlation is a function of the covariance. Positive correlation means one variable increases as the other increases and negative correlation means one variable decreases as the other increases. The correlation matrix before scaling and after scaling will remain the same.

Covariance matrix is done on the scaled dataset. Scaling is done using Z-score method.

```
Covariance Matrix
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.815
54018
0.3987775   0.05022367  0.16515151  0.13272942  0.17896117  0.39120081
0.36996762  0.09575627 -0.09034216  0.2599265   0.14694372]
[ 0.94466636  1.00128866  0.91281145  0.19269493  0.24779465  0.8753498
5
0.44183938 -0.02578774  0.09101577  0.11367165  0.2
0124767  0.35621633
0.3380184   0.17645611 -0.16019604  0.12487773  0.06739929]
[ 0.84791332  0.91281145  1.00128866  0.18152715  0.2270373   0.9658827
4
```

```
0.51372977 -0.1556777  -0.04028353  0.11285614  0.28129148  0.33189629
0.30867133  0.23757707 -0.18102711  0.06425192 -0.02236983]
[ 0.33927032  0.19269493  0.18152715  1.00128866  0.89314445  0.1414708
-0.10549205  0.5630552   0.37195909  0.1190116  -0.09343665  0.53251337
0.49176793 -0.38537048  0.45607223  0.6617651   0.49562711]
[ 0.35209304  0.24779465  0.2270373   0.89314445  1.00128866  0.1997016
7
-0.05364569  0.49002449  0.33191707  0.115676   -0.08091441  0.54656564
0.52542506 -0.29500852  0.41840277  0.52812713  0.47789622]
[ 0.81554018  0.87534985  0.96588274  0.1414708   0.19970167  1.0012886
6
0.57124738 -0.21602002 -0.06897917  0.11569867  0.31760831  0.3187472
0.30040557  0.28006379 -0.22975792  0.01867565 -0.07887464]
[ 0.3987775   0.44183938  0.51372977 -0.10549205 -0.05364569  0.5712473
8
1.00128866 -0.25383901 -0.06140453  0.08130416  0.32029384  0.14930637
0.14208644  0.23283016 -0.28115421 -0.08367612 -0.25733218]
[ 0.05022367 -0.02578774 -0.1556777   0.5630552   0.49002449 -0.2160200
2
-0.25383901  1.00128866  0.65509951  0.03890494 -0.29947232  0.38347594
0.40850895 -0.55553625  0.56699214  0.6736456   0.57202613]
[ 0.16515151  0.09101577 -0.04028353  0.37195909  0.33191707 -0.0689791
7
-0.06140453  0.65509951  1.00128866  0.12812787 -0.19968518  0.32962651
0.3750222  -0.36309504  0.27271444  0.50238599  0.42548915]
[ 0.13272942  0.11367165  0.11285614  0.1190116   0.115676    0.1156986
7
0.08130416  0.03890494  0.12812787  1.00128866  0.17952581  0.0269404
0.10008351 -0.03197042 -0.04025955  0.11255393  0.00106226]
[ 0.17896117  0.20124767  0.28129148 -0.09343665 -0.08091441  0.3176083
1
0.32029384 -0.29947232 -0.19968518  0.17952581  1.00128866 -0.01094989
-0.03065256  0.13652054 -0.2863366  -0.09801804 -0.26969106]
[ 0.39120081  0.35621633  0.33189629  0.53251337  0.54656564  0.3187472
0.14930637  0.38347594  0.32962651  0.0269404  -0.01094989  1.00128866
0.85068186 -0.13069832  0.24932955  0.43331936  0.30543094]
[ 0.36996762  0.3380184   0.30867133  0.49176793  0.52542506  0.3004055
7
0.14208644  0.40850895  0.3750222   0.10008351 -0.03065256  0.85068186
1.00128866 -0.16031027  0.26747453  0.43936469  0.28990033]
[ 0.09575627  0.17645611  0.23757707 -0.38537048 -0.29500852  0.2800637
9
0.23283016 -0.55553625 -0.36309504 -0.03197042  0.13652054 -0.13069832
-0.16031027  1.00128866 -0.4034484  -0.5845844  -0.30710565]
[-0.09034216 -0.16019604 -0.18102711  0.45607223  0.41840277 -0.2297579
2
-0.28115421  0.56699214  0.27271444 -0.04025955 -0.2863366   0.24932955
0.26747453 -0.4034484   1.00128866  0.41825001  0.49153016]
[ 0.2599265   0.12487773  0.06425192  0.6617651   0.52812713  0.0186756
5
-0.08367612  0.6736456   0.50238599  0.11255393 -0.09801804  0.43331936
0.43936469 -0.5845844   0.41825001  1.00128866  0.39084571]
[ 0.14694372  0.06739929 -0.02236983  0.49562711  0.47789622 -0.0788746
4
-0.25733218  0.57202613  0.42548915  0.00106226 -0.26969106  0.30543094
0.28990033 -0.30710565  0.49153016  0.39084571  1.00128866]]
```

Above is the covariance matrix on the scaled dataset. The values in the covariance matrix show the distribution and direction of multivariate data in multidimensional space.

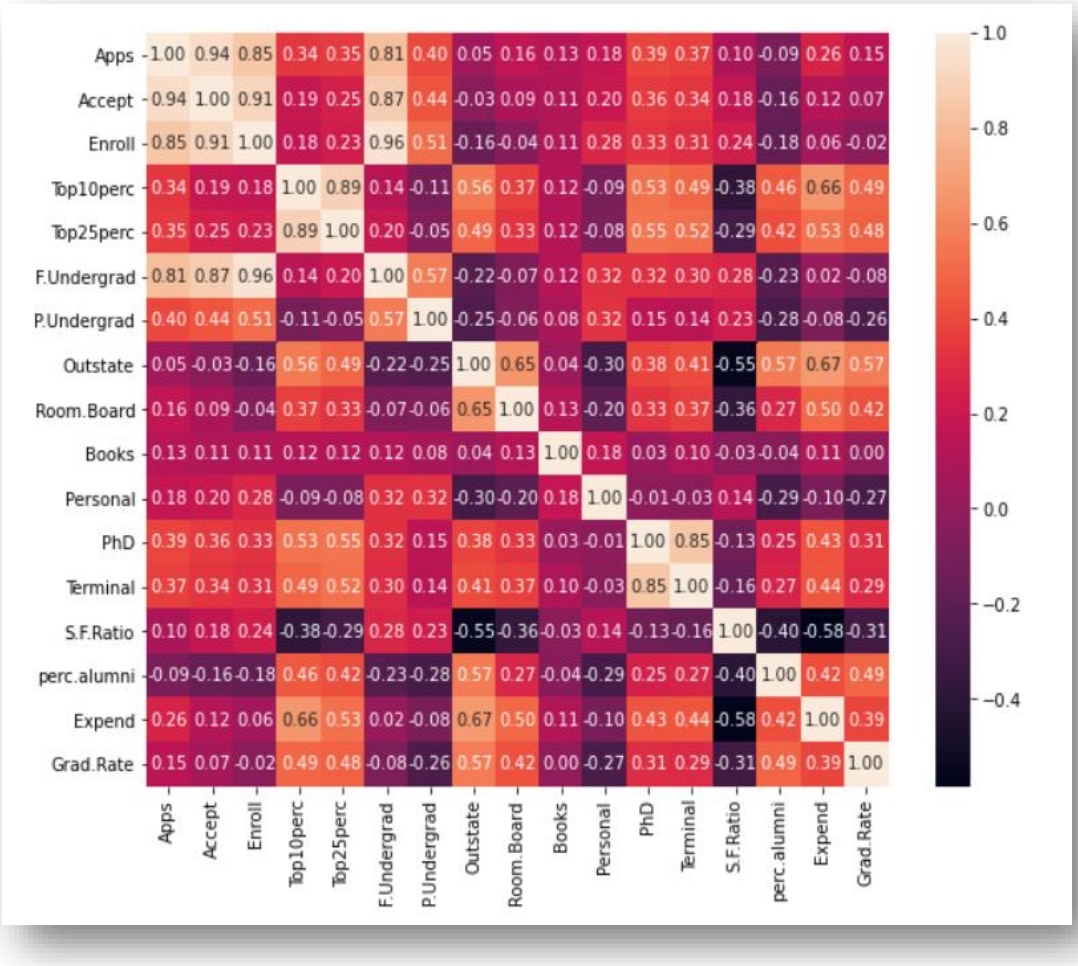

Figure 38: Heat map representing the correlation of the dataset

Correlation matrix is represented by a heat map shown in figure 39. From figure 39, we can see the variables that are highly positively correlated, variables that are highly negatively correlated and also the variables that are moderately correlated with each other. We can also see that the variables ‘Apps’, ‘Accept’, ‘Enroll’ and ‘F.Undergrad’ are highly positively correlated. The variables ‘Top10perc’ and ‘Top25perc’ are highly positively correlated.

2.4 Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]

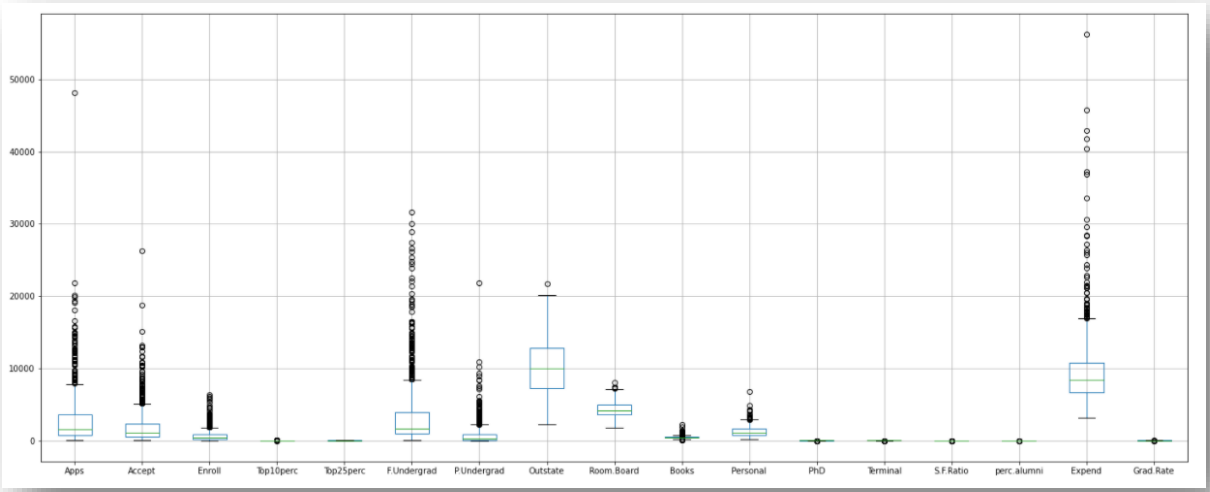

Figure 39: Boxplot showing the dataset with outliers before scaling

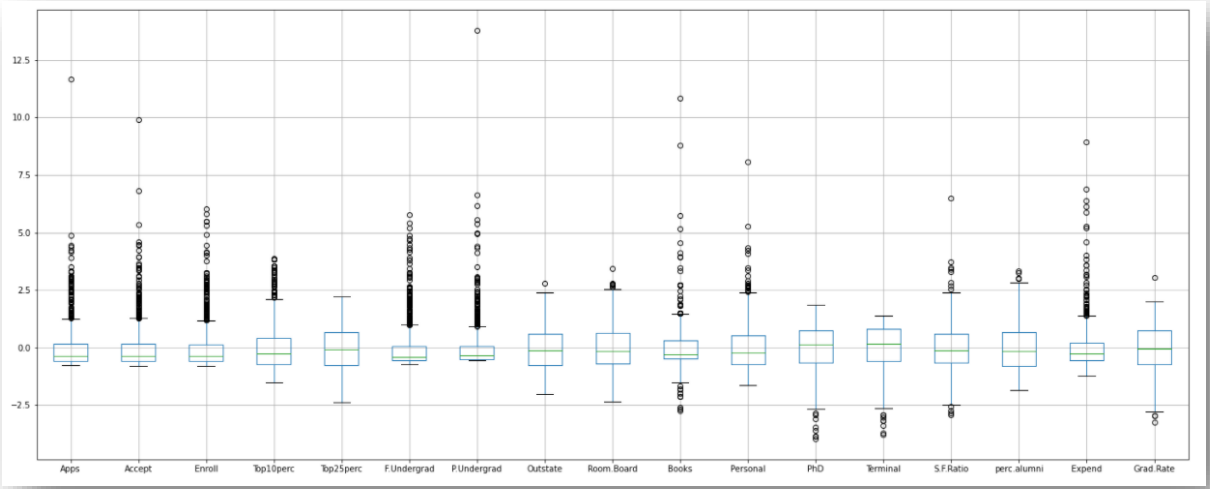

Figure 40: Boxplot showing the dataset with outliers after scaling

Insights:

From figure 40 and 41, it is seen that both the graphs have outliers present. This is because scaling does not treat or remove outliers. Scaling is only a way of representing the dataset. The values may seem to be different but however they are only scaled. The dataset is brought into one unit of comparison. This can be seen in figure 38. The plot of the 'Apps' variable before and after scaling is the same but the scale has changed.

Outliers have to be further treated using other methods.

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

Eigen vectors:

```
array([[ 2.48765602e-01,  2.07601502e-01,  1.76303592e-01,
         3.54273947e-01,  3.44001279e-01,  1.54640962e-01,
         2.64425045e-02,  2.94736419e-01,  2.49030449e-01,
         6.47575181e-02, -4.25285386e-02,  3.18312875e-01,
         3.17056016e-01, -1.76957895e-01,  2.05082369e-01,
         3.18908750e-01,  2.52315654e-01],
       [ 3.31598227e-01,  3.72116750e-01,  4.03724252e-01,
        -8.24118211e-02, -4.47786551e-02,  4.17673774e-01,
         3.15087830e-01, -2.49643522e-01, -1.37808883e-01,
         5.63418434e-02,  2.19929218e-01,  5.83113174e-02,
         4.64294477e-02,  2.46665277e-01, -2.46595274e-01,
        -1.31689865e-01, -1.69240532e-01],
       [-6.30921033e-02, -1.01249056e-01, -8.29855709e-02,
         3.50555339e-02, -2.41479376e-02, -6.13929764e-02,
         1.39681716e-01,  4.65988731e-02,  1.48967389e-01,
         6.77411649e-01,  4.99721120e-01, -1.27028371e-01,
        -6.60375454e-02, -2.89848401e-01, -1.46989274e-01,
         2.26743985e-01, -2.08064649e-01],
       [ 2.81310530e-01,  2.67817346e-01,  1.61826771e-01,
        -5.15472524e-02, -1.09766541e-01,  1.00412335e-01,
        -1.58558487e-01,  1.31291364e-01,  1.84995991e-01,
         8.70892205e-02, -2.30710568e-01, -5.34724832e-01,
        -5.19443019e-01, -1.61189487e-01,  1.73142230e-02,
         7.92734946e-02,  2.69129066e-01],
       [ 5.74140964e-03,  5.57860920e-02, -5.56936353e-02,
        -3.95434345e-01, -4.26533594e-01, -4.34543659e-02,
         3.02385408e-01,  2.22532003e-01,  5.60919470e-01,
        -1.27288825e-01, -2.22311021e-01,  1.40166326e-01,
         2.04719730e-01, -7.93882496e-02, -2.16297411e-01,
         7.59581203e-02, -1.09267913e-01],
       [-1.62374420e-02,  7.53468452e-03, -4.25579803e-02,
        -5.26927980e-02,  3.30915896e-02, -4.34542349e-02,
        -1.91198583e-01, -3.00003910e-02,  1.62755446e-01,
         6.41054950e-01, -3.31398003e-01,  9.12555212e-02,
         1.54927646e-01,  4.87045875e-01, -4.73400144e-02,
        -2.98118619e-01,  2.16163313e-01],
       [-4.24863486e-02, -1.29497196e-02, -2.76928937e-02,
        -1.61332069e-01, -1.18485556e-01, -2.50763629e-02,
         6.10423460e-02,  1.08528966e-01,  2.09744235e-01,
        -1.49692034e-01,  6.33790064e-01, -1.09641298e-03,
        -2.84770105e-02,  2.19259358e-01,  2.43321156e-01,
        -2.26584481e-01,  5.59943937e-01],
       [-1.03090398e-01, -5.62709623e-02,  5.86623552e-02,
        -1.22678028e-01, -1.02491967e-01,  7.88896442e-02,
         5.70783816e-01,  9.84599754e-03, -2.21453442e-01,
         2.13293009e-01, -2.32660840e-01, -7.70400002e-02,
        -1.21613297e-02, -8.36048735e-02,  6.78523654e-01,
        -5.41593771e-02, -5.33553891e-03],
       [-9.02270802e-02, -1.77864814e-01, -1.28560713e-01,
         3.41099863e-01,  4.03711989e-01, -5.94419181e-02,
         5.60672902e-01, -4.57332880e-03,  2.75022548e-01,
        -1.33663353e-01, -9.44688900e-02, -1.85181525e-01,
        -2.54938198e-01,  2.74544380e-01, -2.55334907e-01,
        -4.91388809e-02,  4.19043052e-02],
       [ 5.25098025e-02,  4.11400844e-02,  3.44879147e-02,
```

```
         6.40257785e-02,  1.45492289e-02,  2.08471834e-02,
        -2.23105808e-01,  1.86675363e-01,  2.98324237e-01,
        -8.20292186e-02,  1.36027616e-01, -1.23452200e-01,
        -8.85784627e-02,  4.72045249e-01,  4.22999706e-01,
         1.32286331e-01, -5.90271067e-01],
       [ 4.30462074e-02, -5.84055850e-02, -6.93988831e-02,
        -8.10481404e-03, -2.73128469e-01, -8.11578181e-02,
         1.00693324e-01,  1.43220673e-01, -3.59321731e-01,
         3.19400370e-02, -1.85784733e-02,  4.03723253e-02,
        -5.89734026e-02,  4.45000727e-01, -1.30727978e-01,
         6.92088870e-01,  2.19839000e-01],
       [ 2.40709086e-02, -1.45102446e-01,  1.11431545e-02,
         3.85543001e-02, -8.93515563e-02,  5.61767721e-02,
        -6.35360730e-02, -8.23443779e-01,  3.54559731e-01,
        -2.81593679e-02, -3.92640266e-02,  2.32224316e-02,
         1.64850420e-02, -1.10262122e-02,  1.82660654e-01,
         3.25982295e-01,  1.22106697e-01],
       [ 5.95830975e-01,  2.92642398e-01, -4.44638207e-01,
         1.02303616e-03,  2.18838802e-02, -5.23622267e-01,
         1.25997650e-01, -1.41856014e-01, -6.97485854e-02,
         1.14379958e-02,  3.94547417e-02,  1.27696382e-01,
        -5.83134662e-02, -1.77152700e-02,  1.04088088e-01,
        -9.37464497e-02, -6.91969778e-02],
       [ 8.06328039e-02,  3.34674281e-02, -8.56967180e-02,
        -1.07828189e-01,  1.51742110e-01, -5.63728817e-02,
         1.92857500e-02, -3.40115407e-02, -5.84289756e-02,
        -6.68494643e-02,  2.75286207e-02, -6.91126145e-01,
         6.71008607e-01,  4.13740967e-02, -2.71542091e-02,
         7.31225166e-02,  3.64767385e-02],
       [ 1.33405806e-01, -1.45497511e-01,  2.95896092e-02,
         6.97722522e-01, -6.17274818e-01,  9.91640992e-03,
         2.09515982e-02,  3.83544794e-02,  3.40197083e-03,
        -9.43887925e-03, -3.09001353e-03, -1.12055599e-01,
         1.58909651e-01, -2.08991284e-02, -8.41789410e-03,
        -2.27742017e-01, -3.39433604e-03],
       [ 4.59139498e-01, -5.18568789e-01, -4.04318439e-01,
        -1.48738723e-01,  5.18683400e-02,  5.60363054e-01,
        -5.27313042e-02,  1.01594830e-01, -2.59293381e-02,
         2.88282896e-03, -1.28904022e-02,  2.98075465e-02,
        -2.70759809e-02, -2.12476294e-02,  3.33406243e-03,
        -4.38803230e-02, -5.00844705e-03],
       [ 3.58970400e-01, -5.43427250e-01,  6.09651110e-01,
        -1.44986329e-01,  8.03478445e-02, -4.14705279e-01,
         9.01788964e-03,  5.08995918e-02,  1.14639620e-03,
         7.72631963e-04, -1.11433396e-03,  1.38133366e-02,
         6.20932749e-03, -2.22215182e-03, -1.91869743e-02,
        -3.53098218e-02, -1.30710024e-02]])
```

Eigen values:

```
array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,
       0.84849117, 0.6057878 , 0.58787222, 0.53061262, 0.4043029 ,
       0.31344588, 0.22061096, 0.16779415, 0.1439785 , 0.08802464,
       0.03672545, 0.02302787])
```

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

Explained variance:

```
array([0.32020628, 0.26340214, 0.06900917, 0.05922989, 0.05488405,
       0.04984701, 0.03558871, 0.03453621, 0.03117234, 0.02375192,
       0.01841426, 0.01296041, 0.00985754, 0.00845842, 0.00517126,
       0.00215754, 0.00135284])
```

Creating a dataframe:

Table 20: Dataframe containing the loadings or coefficients of all PCs

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
PC1	0.25	0.21	0.18	0.35	0.34	0.15	0.03	0.29	0.25	0.06	-0.04	0.32	0.32	-0.18	0.21	0.32	0.252
PC2	0.33	0.37	0.40	-0.08	-0.04	0.42	0.32	-0.25	-0.14	0.06	0.22	0.06	0.05	0.25	-0.25	-0.13	-0.189
PC3	-0.06	-0.10	-0.08	0.04	-0.02	-0.06	0.14	0.05	0.15	0.68	0.50	-0.13	-0.07	-0.29	-0.15	0.23	-0.208
PC4	0.28	0.27	0.16	-0.05	-0.11	0.10	-0.16	0.13	0.18	0.09	-0.23	-0.53	-0.52	-0.16	0.02	0.08	0.269
PC5	0.01	0.06	-0.06	-0.40	-0.43	-0.04	0.30	0.22	0.56	-0.13	-0.22	0.14	0.20	-0.08	-0.22	0.08	-0.109
PC6	-0.02	0.01	-0.04	-0.05	0.03	-0.04	-0.19	-0.03	0.16	0.64	-0.33	0.09	0.15	0.49	-0.05	-0.30	0.216
PC7	-0.04	-0.01	-0.03	-0.16	-0.12	-0.03	0.06	0.11	0.21	-0.15	0.63	0.00	-0.03	0.22	0.24	-0.23	0.560
PC8	-0.10	-0.06	0.06	-0.12	-0.10	0.08	0.57	0.01	-0.22	0.21	-0.23	-0.08	-0.01	-0.08	0.68	-0.05	-0.005
PC9	-0.09	-0.18	-0.13	0.34	0.40	-0.06	0.56	0.00	0.28	-0.13	-0.09	-0.19	-0.25	0.27	-0.26	-0.05	0.042
PC10	0.05	0.04	0.03	0.06	0.01	0.02	-0.22	0.19	0.30	-0.08	0.14	-0.12	-0.09	0.47	0.42	0.13	-0.590
PC11	0.04	-0.06	-0.07	-0.01	-0.27	-0.08	0.10	0.14	-0.36	0.08	-0.02	0.04	-0.06	0.45	-0.13	0.69	0.220
PC12	0.02	-0.15	0.01	0.04	-0.09	0.06	-0.06	-0.82	0.35	-0.03	-0.04	0.02	0.02	-0.01	0.18	0.33	0.122
P13	0.60	0.29	-0.44	0.00	0.02	-0.52	0.13	-0.14	-0.07	0.01	0.04	0.13	-0.06	-0.02	0.10	-0.09	-0.069
P14	0.08	0.03	-0.09	-0.11	0.15	-0.06	0.02	-0.03	-0.06	-0.07	0.03	-0.69	0.67	0.04	-0.03	0.07	0.036
P15	0.13	-0.15	0.03	0.70	-0.62	0.01	0.02	0.04	0.00	-0.01	0.00	-0.11	0.16	-0.02	-0.01	-0.23	-0.003
P16	0.46	-0.52	-0.40	-0.15	0.05	0.56	-0.05	0.10	-0.03	0.00	-0.01	0.03	-0.03	-0.02	0.00	-0.04	-0.005
P17	0.36	-0.54	0.61	-0.14	0.08	-0.41	0.01	0.05	0.00	0.00	0.00	0.01	0.01	0.00	-0.02	-0.04	-0.013

2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

```
The Linear eq of 1st component:
0.25 * Apps + 0.21 * Accept + 0.18 * Enroll + 0.35 * Top10perc + 0.34 *
Top25perc + 0.15 * F.Undergrad + 0.03 * P.Undergrad + 0.29 * Outstate +
0.25 * Room.Board + 0.06 * Books + -0.04 * Personal + 0.32 * PhD + 0.32
* Terminal + -0.18 * S.F.Ratio + 0.21 * perc.alumni + 0.32 * Expend + 0
.25 * Grad.Rate
```

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

Explained variance:

```
array([0.32020628, 0.26340214, 0.06900917, 0.05922989, 0.05488405,
       0.04984701, 0.03558871, 0.03453621, 0.03117234, 0.02375192,
       0.01841426, 0.01296041, 0.00985754, 0.00845842, 0.00517126,
       0.00215754, 0.00135284])
```

Cumulative explained variance ratio:

```
array([0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154,
       0.81657854, 0.85216726, 0.88670347, 0.91787581, 0.94162773,
       0.96004199, 0.9730024 , 0.98285994, 0.99131837, 0.99648962,
       0.99864716, 1.        ])
```

The cumulative explained variance ratio is used to find a cut off for selecting the number of Pcs. For example, 0.32020628 means 32.02% of the data is captured in PC1. In other words the first component explains 32.02% variance in data. Therefore in this study we are selecting 7 components to explain 85.21% variance in data.

Optimum number of principal components:

The optimum choice of the number of principal components (k) is subjective. The set of principal components effectively substitute the original number of variables in the dataset. General rule of thumb is to choose the number of principal components so as to explain 70% -90% of the total variance. Often a screeplot is used to determine the number of principal components.

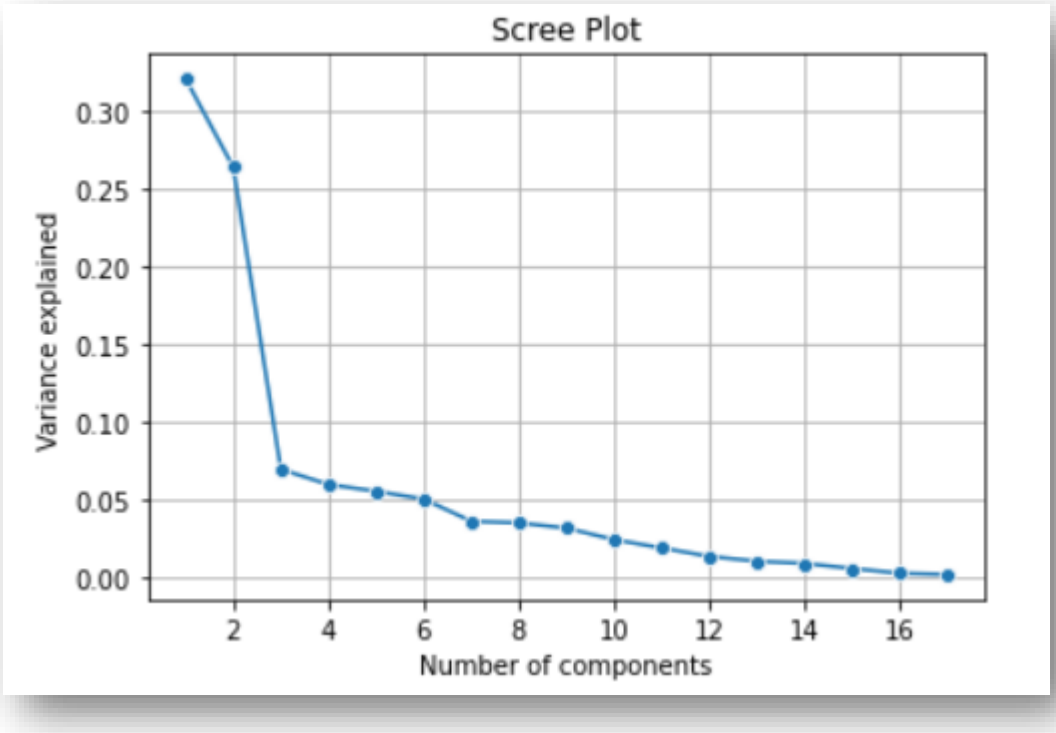

Figure 41: Scree Plot

The scree plot is a useful visual tool to select k. On the X-axis are shown the number of components of the PCs and on the Y-axis are shown the variances. If there is a distinct break point in the line

joining the variances (elbow point) beyond which the line becomes approximately horizontal, then that point may be taken as the value of k, provided other conditions are also satisfied.

In figure 42, there is a distinct break at 3. However, k cannot be taken to be 3 since the first three PCs explain only 65% of total variance. The PCs must be taken so as to explain between 70% -90% of the total variance. If k=7, then the first 7 PCs explain 85% of the total variance. One choice of k could have been 5 or 6. However, we have taken k= 7 so that the explained variance is above 80%.

Table 21: Dataframe containing the selected Pcs

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	0.219259
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944

2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]

PCA is performed and it is exported into a data frame. After PCA the multicollinearity is highly reduced.

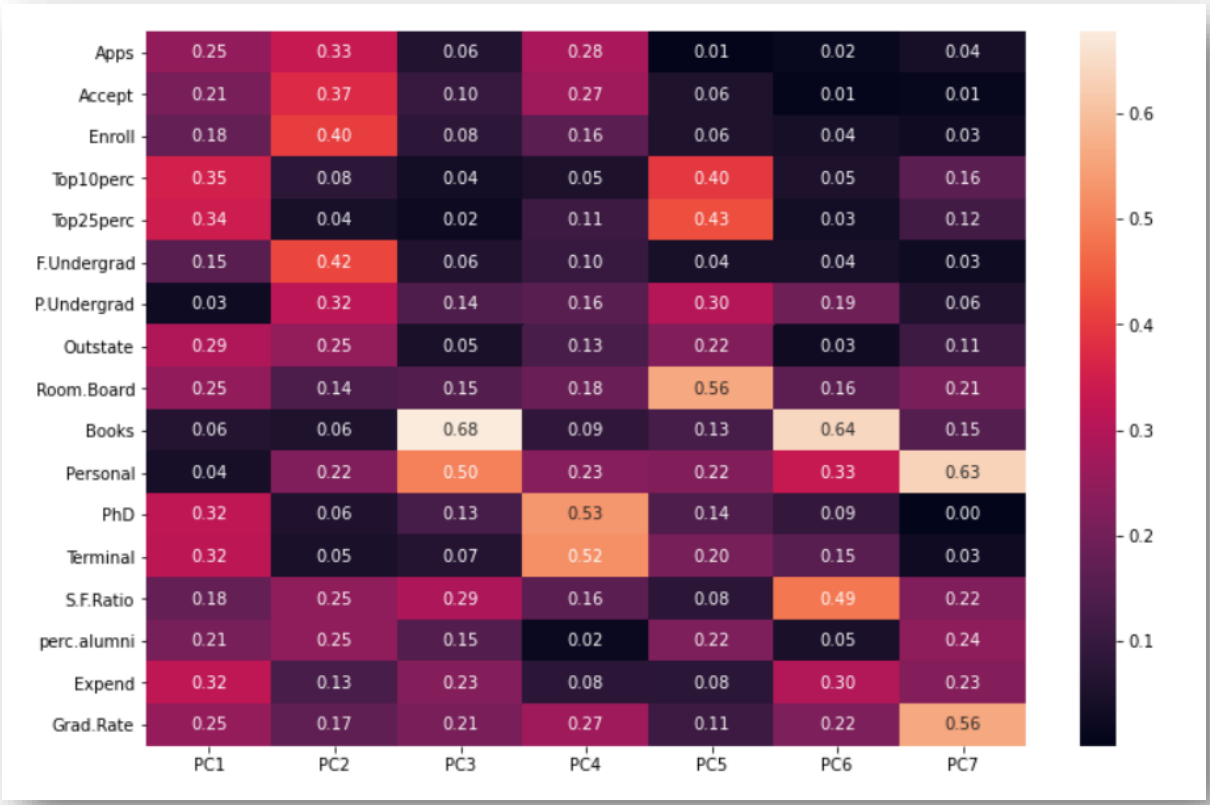

Figure 42: Heat map showing the correlation of the selected PCs

This case study is related to education and the dataset has information related to the students, names of universities etc. The dataset is analysed through various techniques like univariate analysis and multivariate analysis which helps us understand the variables better. Through univariate analysis we are able to know the distribution of the dataset and also about the outliers of the dataset. Through bivariate and multivariate analysis in figure 36 and 37, we can understand the correlation of variables. From figure 36 and 37, we can understand that multiple variables highly correlated with each other.

Further scaling is performed. This helps the dataset to standardize the variable in one unit. The principal component analysis is used reduce the multicollinearity between the variables. There are 17 variables in this dataset. Depending on the variance of the dataset we can reduce the PCA components. The PCA components for this business case is 7 where we could understand the maximum variance of the dataset. Using the components we can now understand the reduced multicollinearity in the dataset. From figure 43, we can see that the multicollinearity has reduced after performing PCA.