College University: Principal Component Analysis

Table of Contents

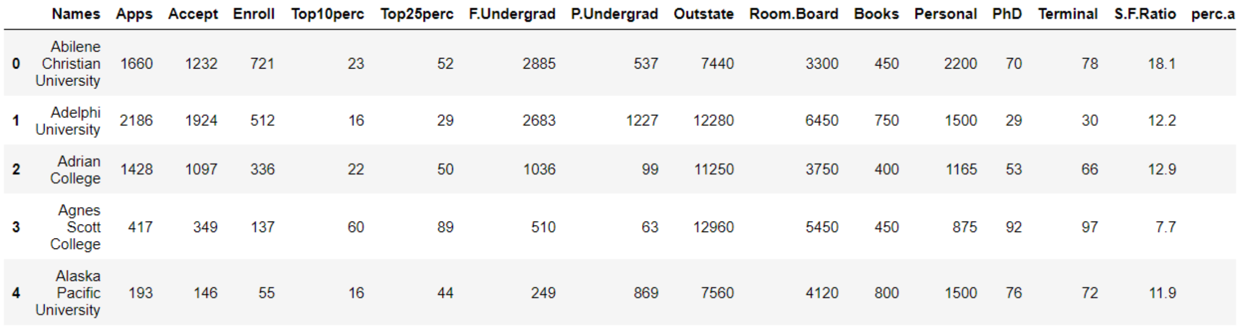
Contents

Education Principal Component Analysis

The dataset contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given

Basic Data Exploration

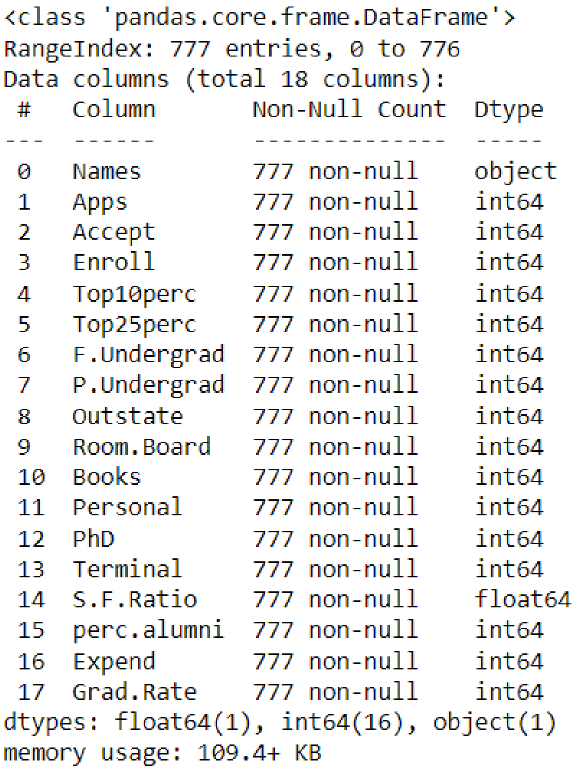
Sample of the dataset:



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

Figure 1. Education_Dataset_Sample

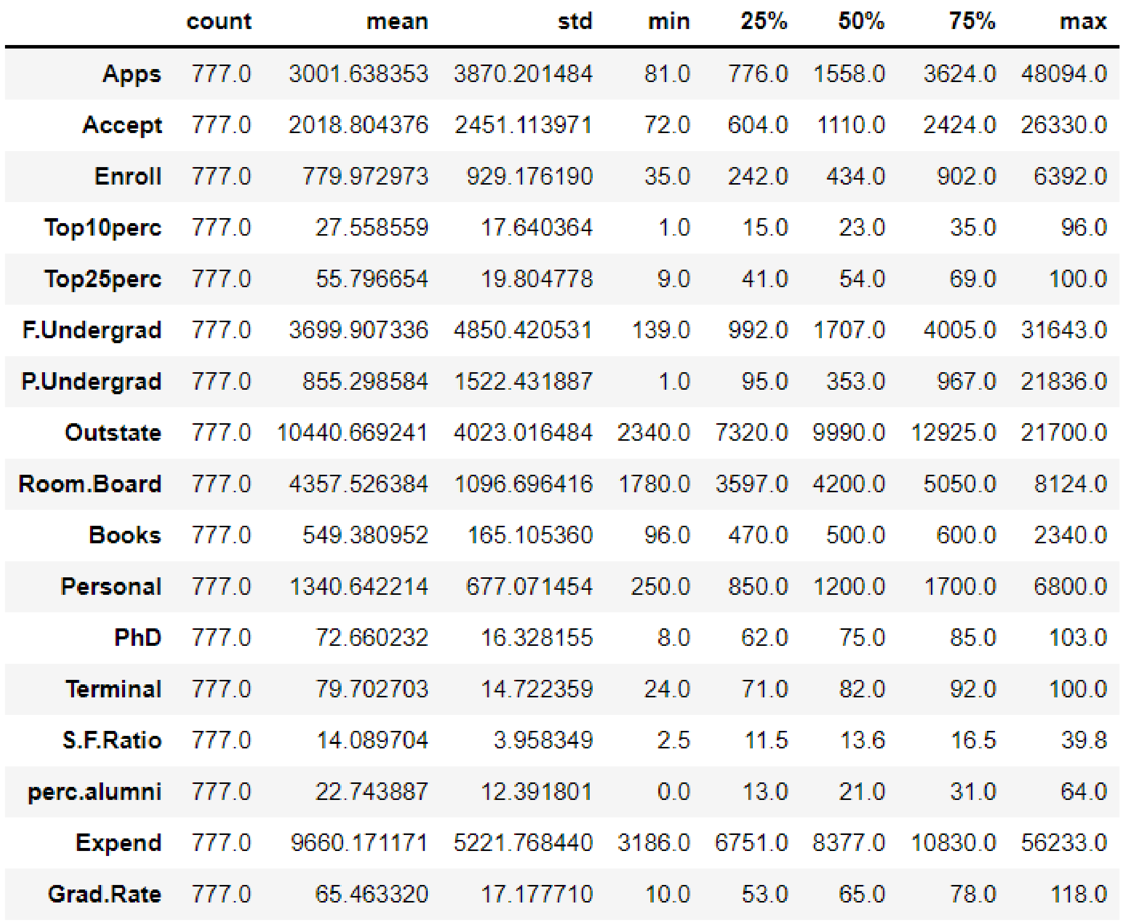
Let us check the basic info of the data frame.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 18 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Names        777 non-null    object 
 1   Apps         777 non-null    int64  
 2   Accept       777 non-null    int64  
 3   Enroll       777 non-null    int64  
 4   Top10perc    777 non-null    int64  
 5   Top25perc    777 non-null    int64  
 6   F.Undergrad  777 non-null    int64  
 7   P.Undergrad  777 non-null    int64  
 8   Outstate     777 non-null    int64  
 9   Room.Board   777 non-null    int64  
 10  Books        777 non-null    int64  
 11  Personal     777 non-null    int64  
 12  PhD          777 non-null    int64  
 13  Terminal     777 non-null    int64  
 14  S.F.Ratio    777 non-null    float64
 15  perc.alumni  777 non-null    int64  
 16  Expend       777 non-null    int64  
 17  Grad.Rate    777 non-null    int64  
dtypes: float64(1), int64(16), object(1)
memory usage: 109.4+ KB
```

Figure 2. Education_Dataset_Info

Below is the statistical summary of the dataset

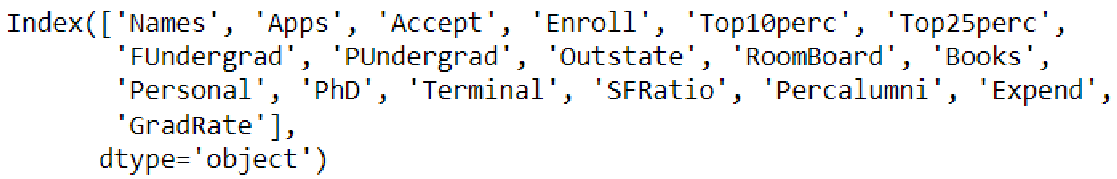


	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	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

Figure 3. Education_Dataset_Statistical Summary

Data Cleanup

The column names will be checked for special characters (‘.’ , ’^’ , ‘,’ , ‘-‘) and make it uniform (either all in lowercase or uppercase)



```
Index(['Names', 'Apps', 'Accept', 'Enroll', 'Top10perc', 'Top25perc',
       'FUndergrad', 'PUndergrad', 'Outstate', 'RoomBoard', 'Books',
       'Personal', 'PhD', 'Terminal', 'SFRatio', 'Percalumni', 'Expend',
       'GradRate'],
      dtype='object')
```

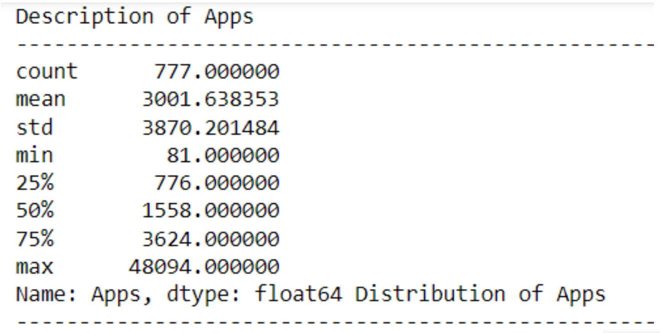
Figure 4. Education_Dataset_Column Names

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

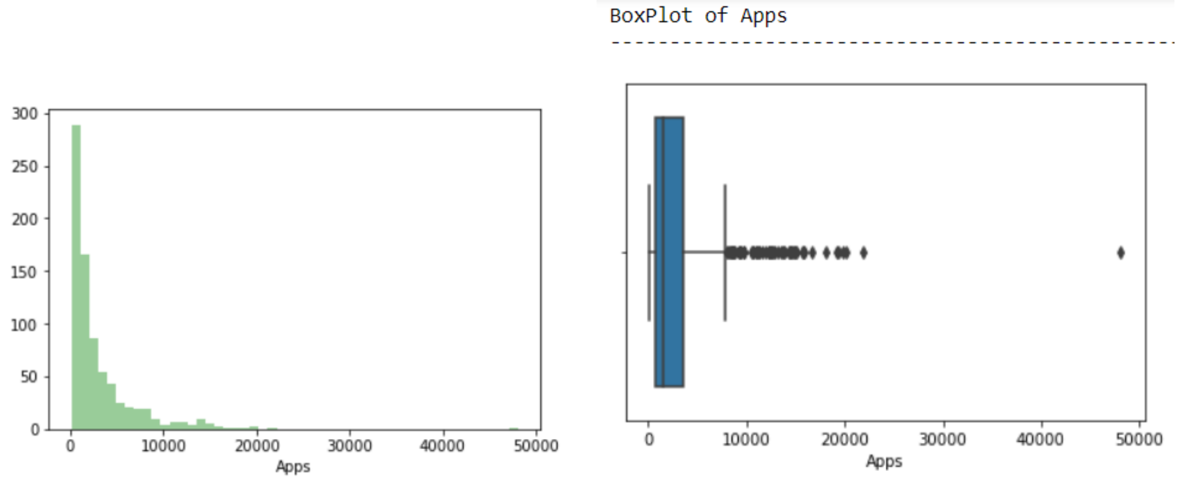
Univariate Analysis

This analysis will display the statistical description of the numeric variable to view 5 point summary, histogram or distplot to view the distribution and the box plot to view outliers if any

- Apps



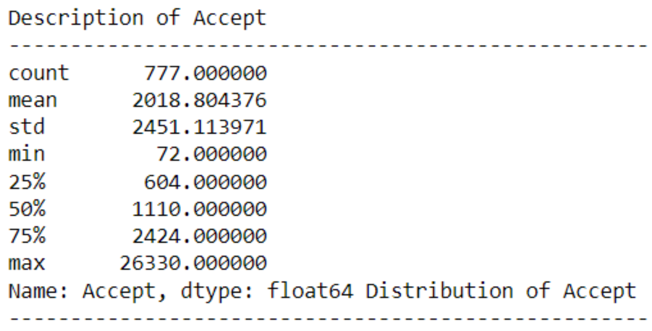
```
Description of Apps
---------------------------------------------------
count      777.000000
mean      3001.638353
std       3870.201484
min         81.000000
25%        776.000000
50%       1558.000000
75%       3624.000000
max      48094.000000
Name: Apps, dtype: float64 Distribution of Apps
---------------------------------------------------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



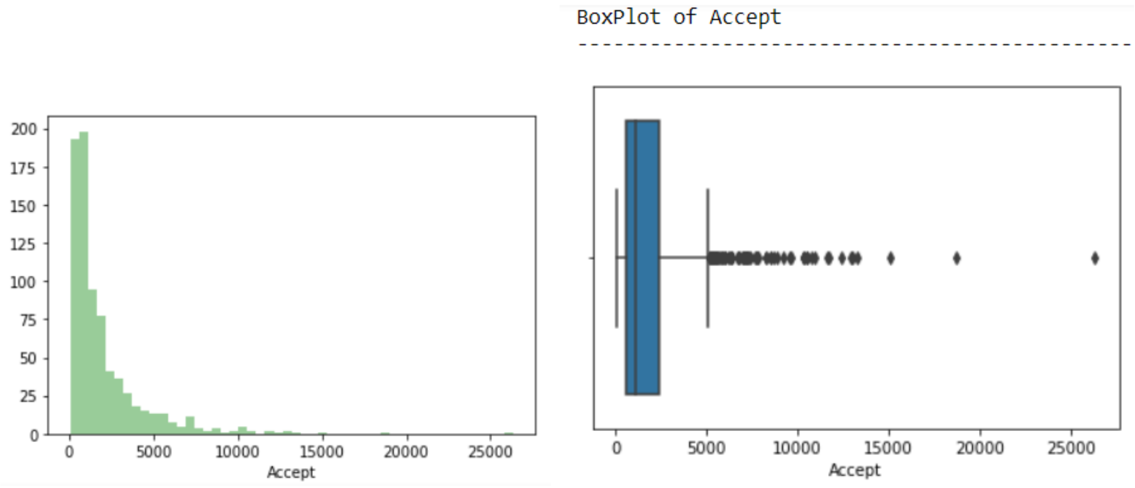
```
BoxPlot of Apps
---------------------------------------------
```

Figure 5. Apps_EDA

- Accept



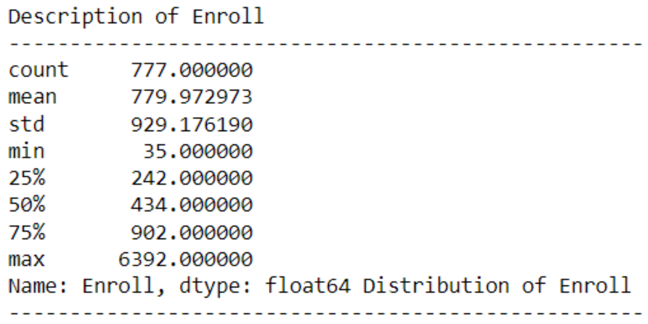
```
Description of Accept
---------------------------------------------------
count      777.000000
mean      2018.804376
std       2451.113971
min         72.000000
25%        604.000000
50%       1110.000000
75%       2424.000000
max      26330.000000
Name: Accept, dtype: float64 Distribution of Accept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



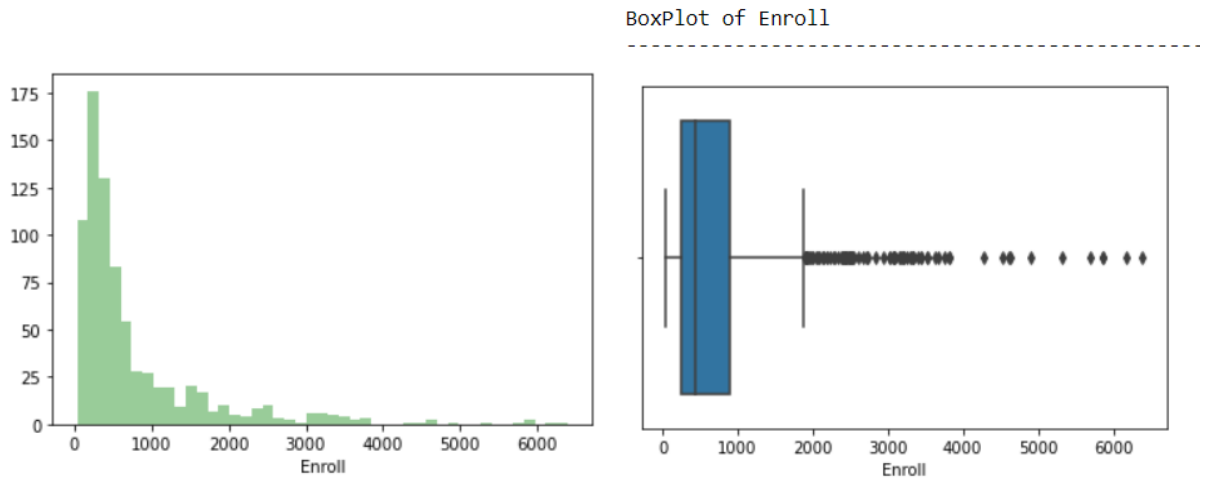
```
BoxPlot of Accept
----------------------------------------------
```

Figure 6. Accept_EDA

- Enroll



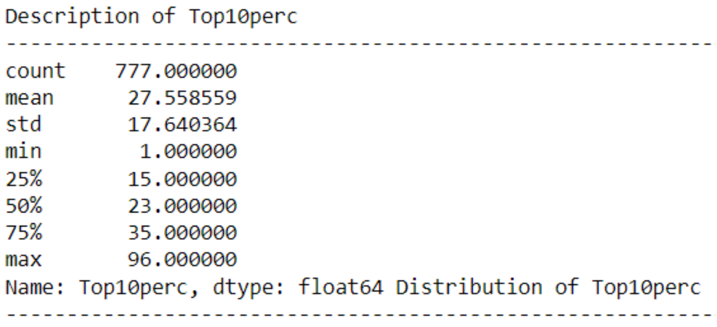
```
Description of Enroll
----------------------------------------------------
count     777.000000
mean      779.972973
std       929.176190
min        35.000000
25%       242.000000
50%       434.000000
75%       902.000000
max      6392.000000
Name: Enroll, dtype: float64 Distribution of Enroll
----------------------------------------------------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



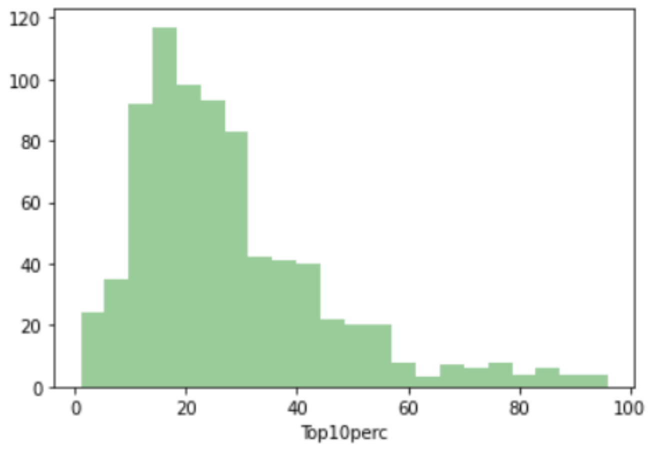
```
BoxPlot of Enroll
-----------------------------------------------
```

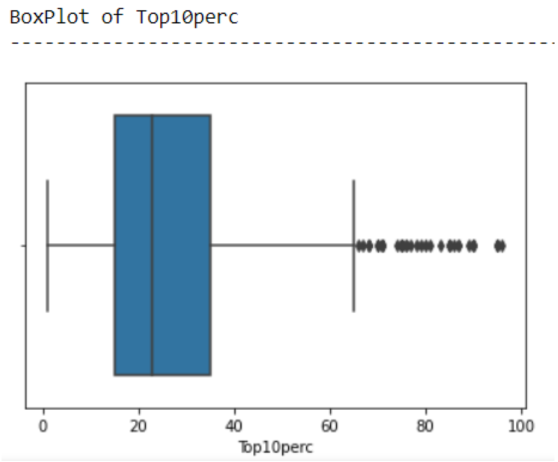
Figure 7. Enroll_EDA

- 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
------------------------------------------------------------
```

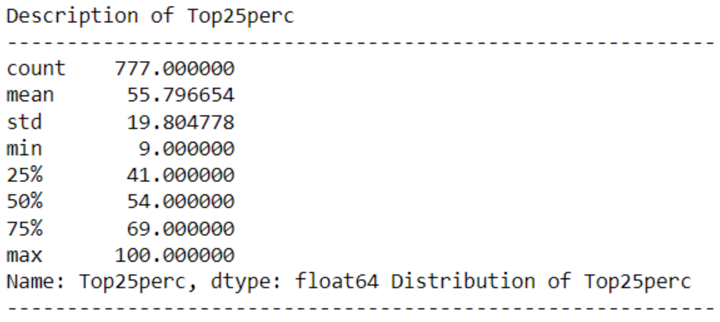




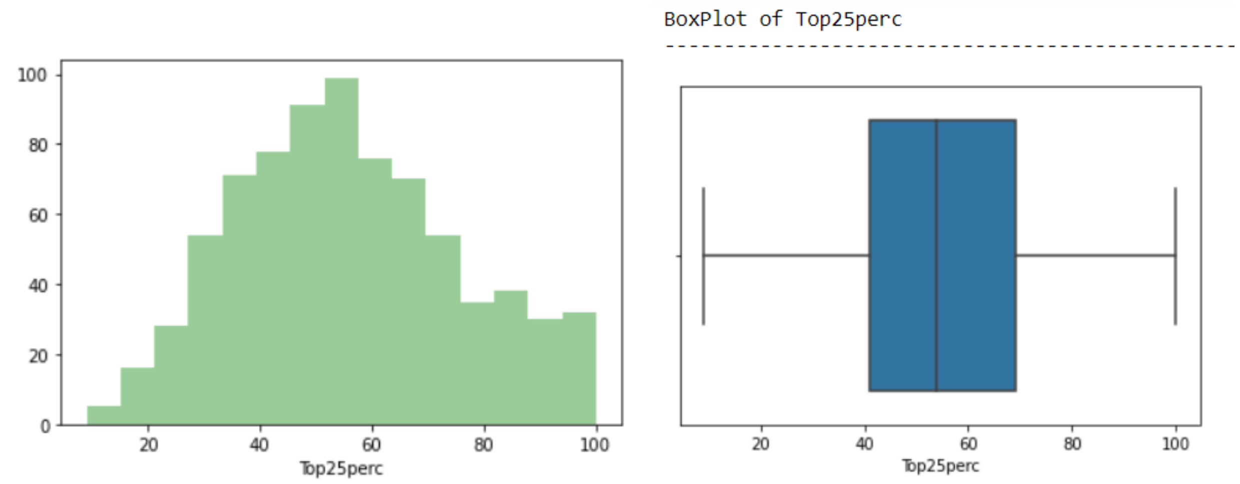
```
BoxPlot of Top10perc
-----------------------------------------------
```

Figure 8. Top10perc_EDA

- Top25perc



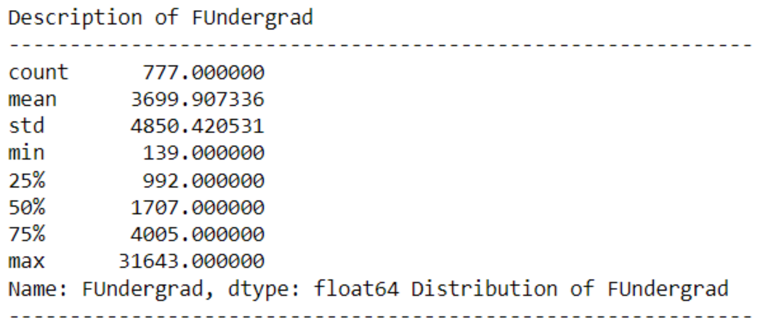
```
Description of Top25perc
------------------------------------------------------------
count    777.000000
mean      55.796654
std       19.804778
min        9.000000
25%       41.000000
50%       54.000000
75%       69.000000
max      100.000000
Name: Top25perc, dtype: float64 Distribution of Top25perc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



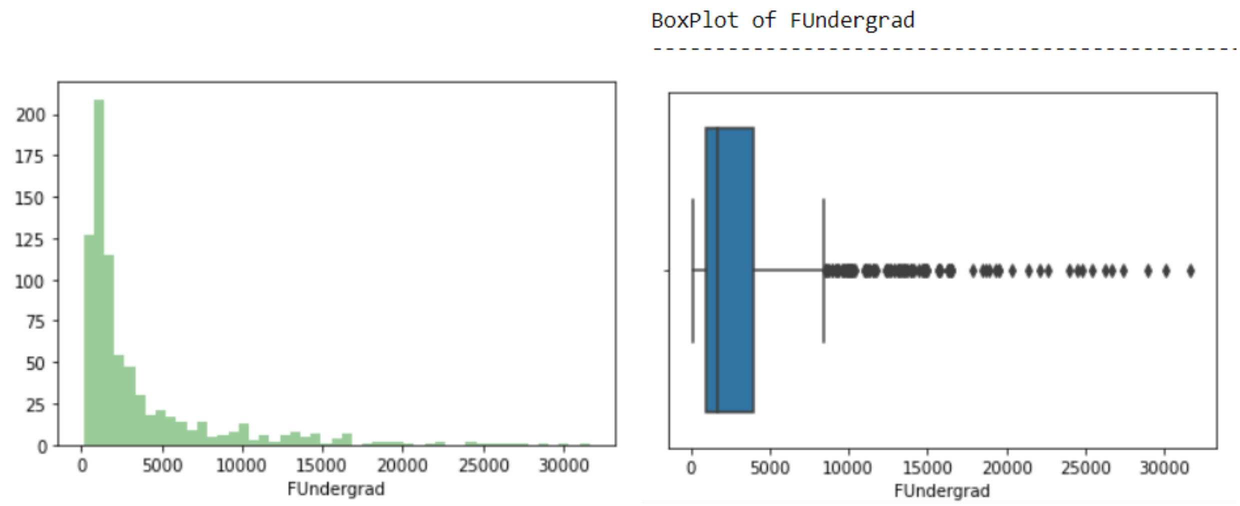
```
BoxPlot of Top25perc
----------------------------------------------
```

Figure 9. Top25perc_EDA

- FUndergrad



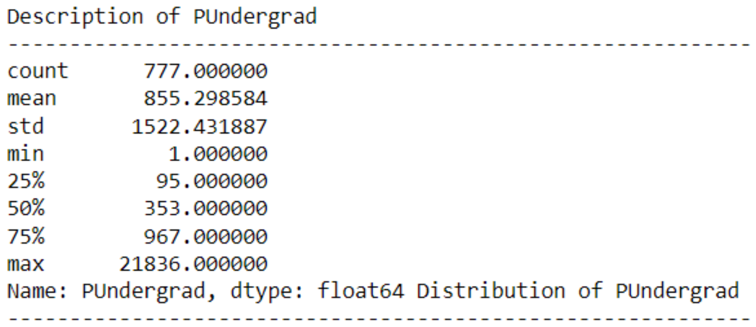
```
Description of FUndergrad
------------------------------------------------------------
count      777.000000
mean      3699.907336
std       4850.420531
min        139.000000
25%        992.000000
50%       1707.000000
75%       4005.000000
max      31643.000000
Name: FUndergrad, dtype: float64 Distribution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



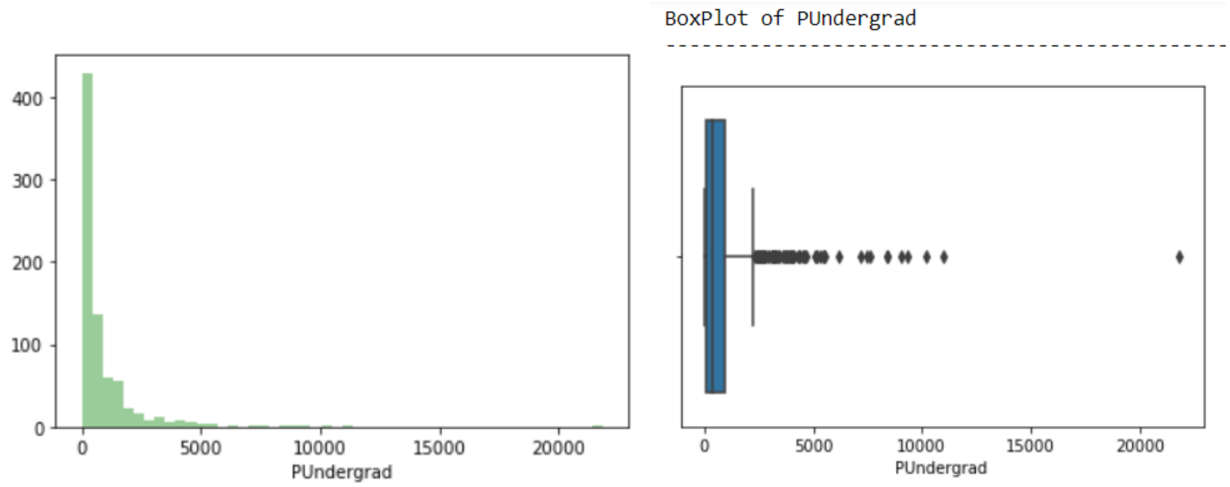
```
BoxPlot of FUndergrad
-----------------------------------------------
```

Figure 10. FUndergrad_EDA

- PUndergrad



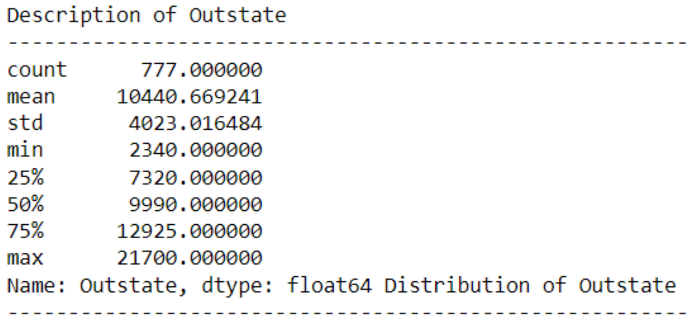
```
Description of PUndergrad
------------------------------------------------------------
count      777.000000
mean       855.298584
std       1522.431887
min          1.000000
25%         95.000000
50%        353.000000
75%        967.000000
max      21836.000000
Name: PUndergrad, dtype: float64 Distribution of PUndergrad
------------------------------------------------------------
```



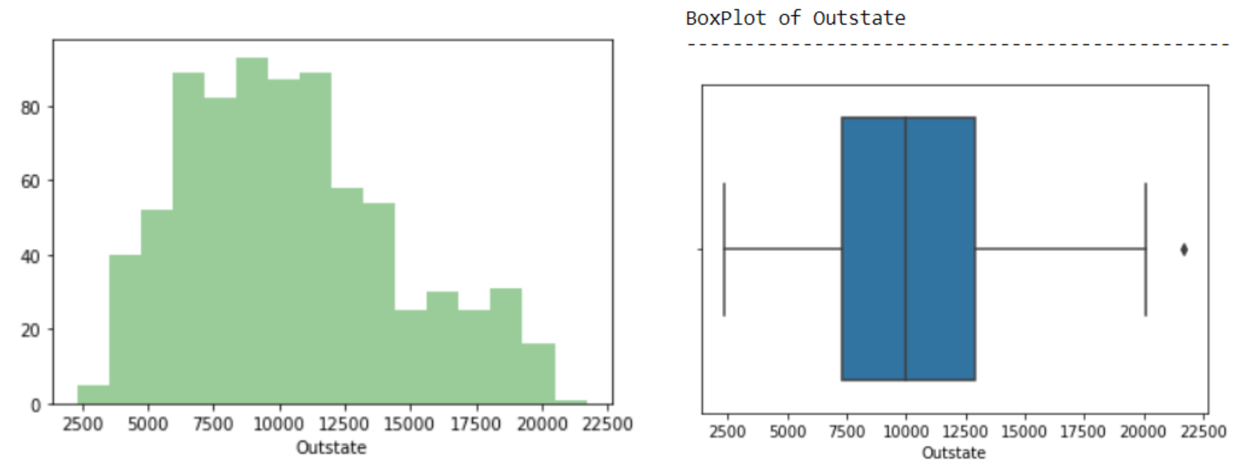
```
BoxPlot of PUndergrad
---------------------------------------------
```

Figure 11. PUndergrad_EDA

- Outstate



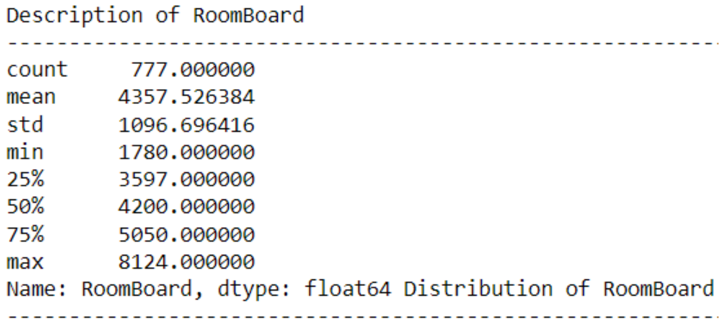
```
Description of Outstate
-------------------------------------------------------
count      777.000000
mean     10440.669241
std       4023.016484
min       2340.000000
25%       7320.000000
50%       9990.000000
75%      12925.000000
max      21700.000000
Name: Outstate, dtype: float64 Distribution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



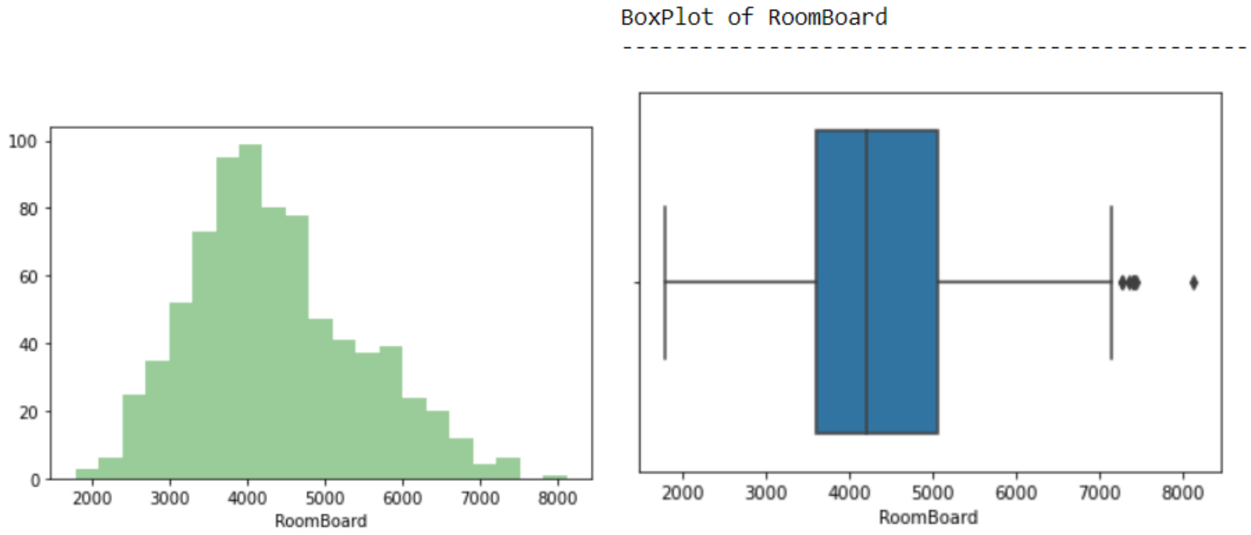
```
BoxPlot of Outstate
----------------------------------------------
```

Figure 12. Outstate_EDA

- RoomBoard



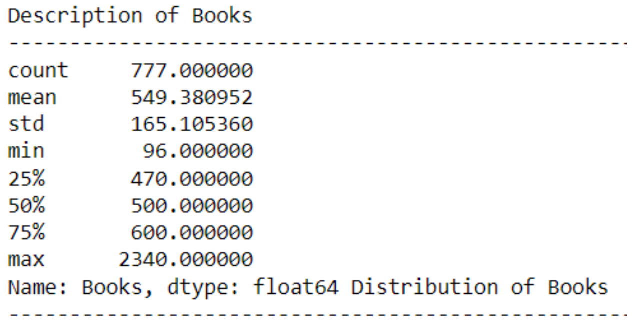
```
Description of RoomBoard
------------------------------------------------------------
count      777.000000
mean      4357.526384
std       1096.696416
min       1780.000000
25%       3597.000000
50%       4200.000000
75%       5050.000000
max       8124.000000
Name: RoomBoard, dtype: float64 Distribution of RoomBoard
------------------------------------------------------------
```



```
BoxPlot of RoomBoard
----------------------------------------------------
```

Figure 13. RoomBoard_EDA

- 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
-----------------------------------------------------
```

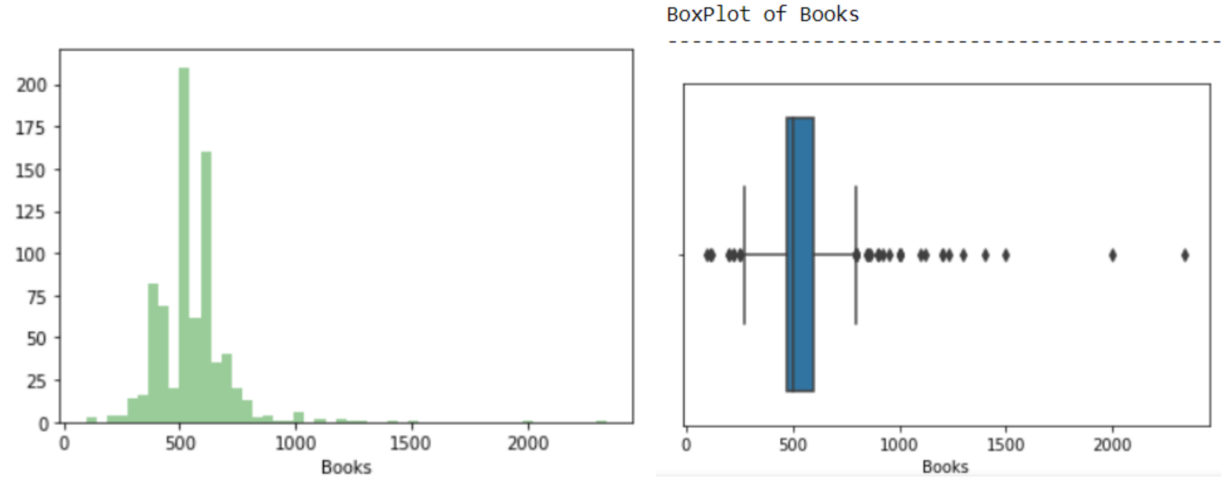
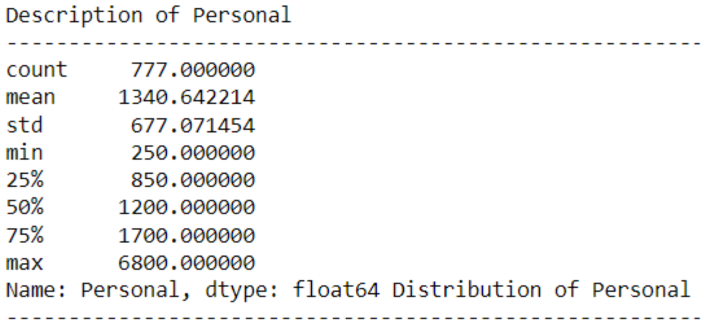


Figure 14. Books_EDA

- 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
-----------------------------------------------------------
```

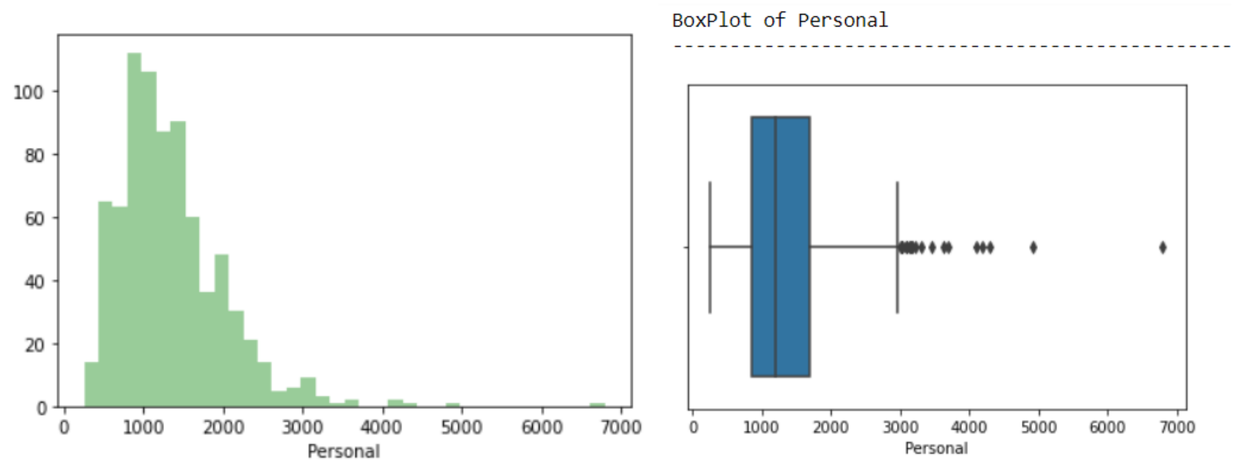
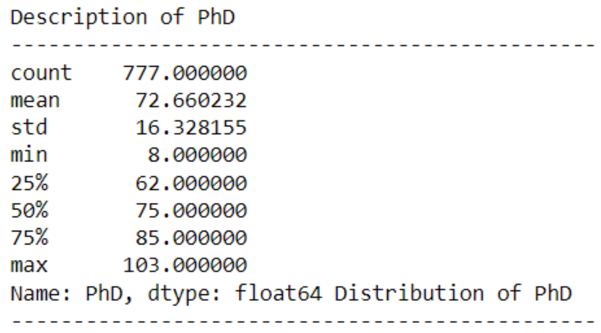


Figure 15. Personal_EDA

- 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
----------------------------------------------
```

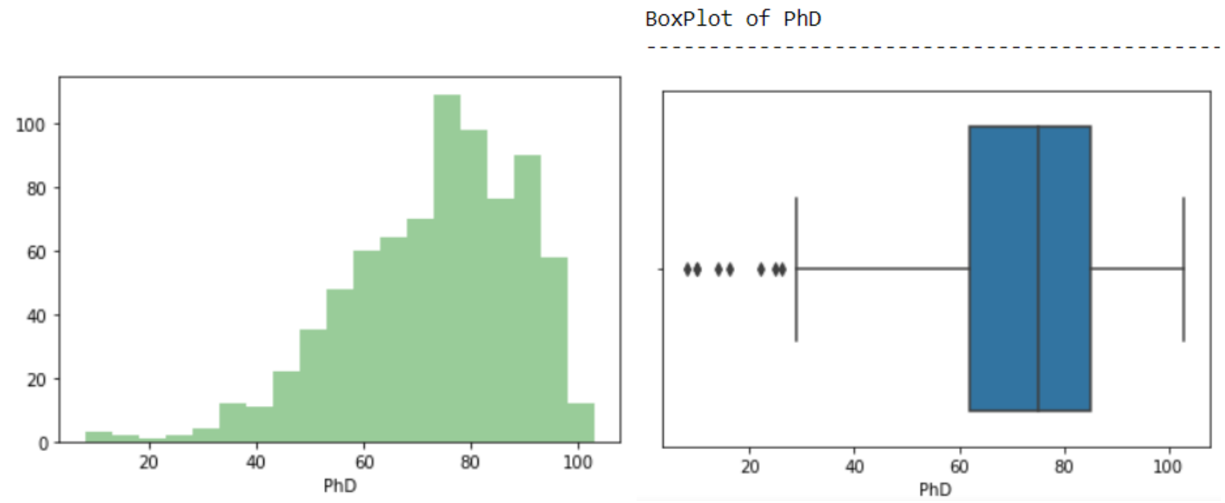
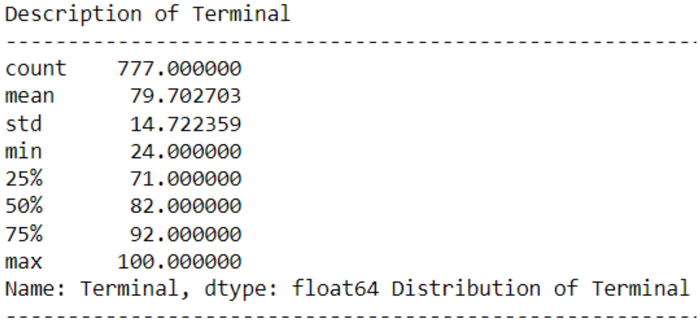
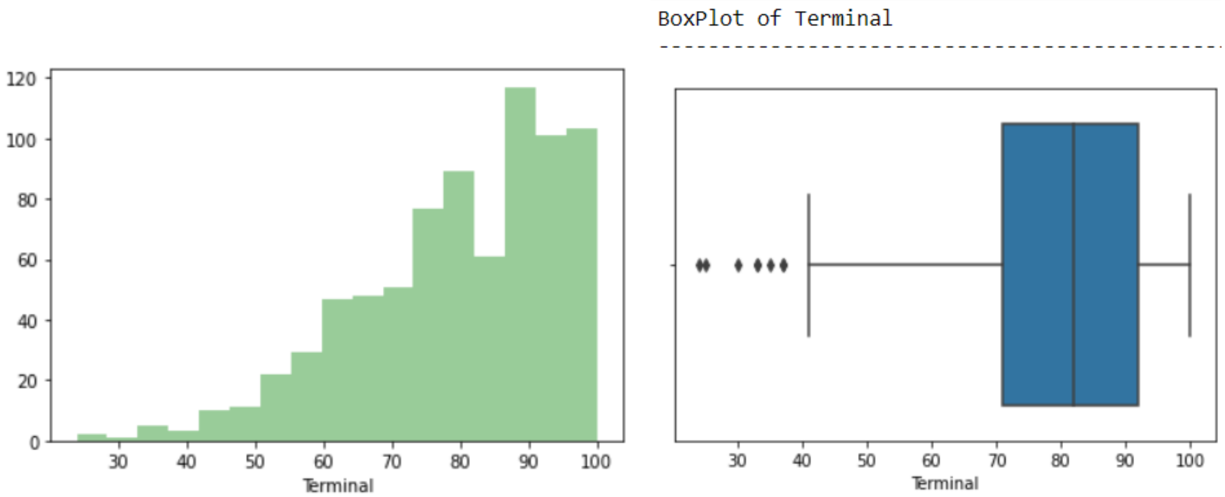


Figure 16. PhD_EDA

- Terminal



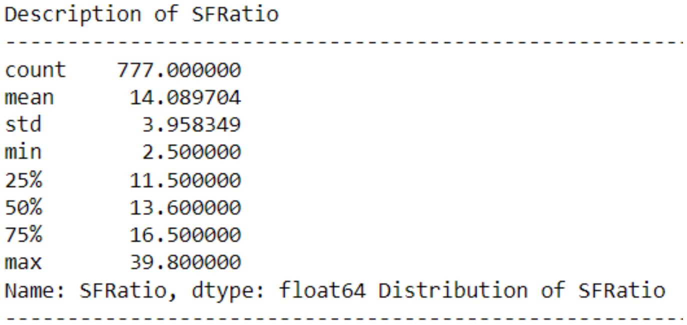
```
Description of Terminal
-------------------------------------------------------
count    777.000000
mean      79.702703
std       14.722359
min       24.000000
25%       71.000000
50%       82.000000
75%       92.000000
max      100.000000
Name: Terminal, dtype: float64 Distribution of Terminal
-------------------------------------------------------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



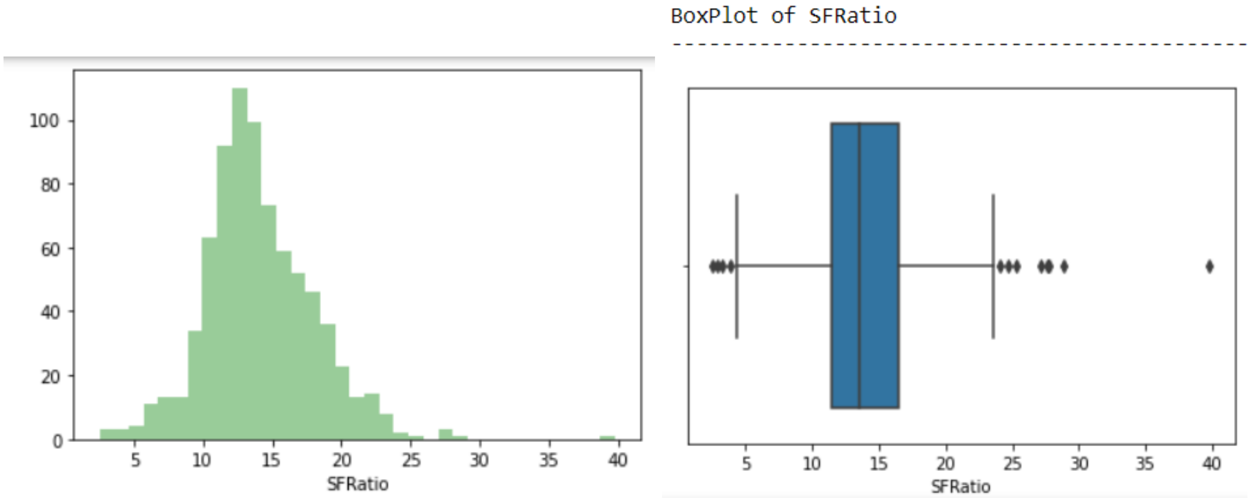
```
BoxPlot of Terminal
---------------------------------------------
```

Figure 17. Terminal_EDA

- SFRatio



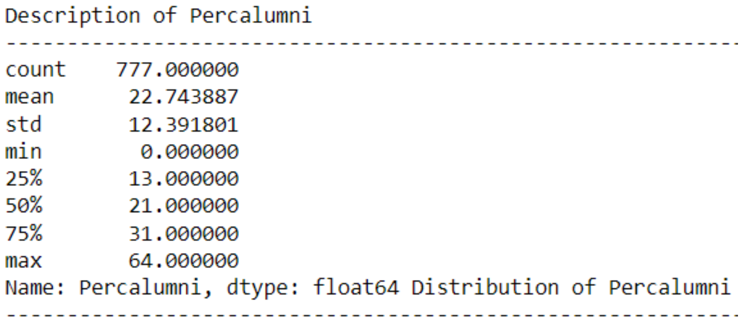
```
Description of SFRatio
-------------------------------------------------------
count    777.000000
mean      14.089704
std        3.958349
min        2.500000
25%       11.500000
50%       13.600000
75%       16.500000
max       39.800000
Name: SFRatio, dtype: float64 Distribution of SFRatio
-------------------------------------------------------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



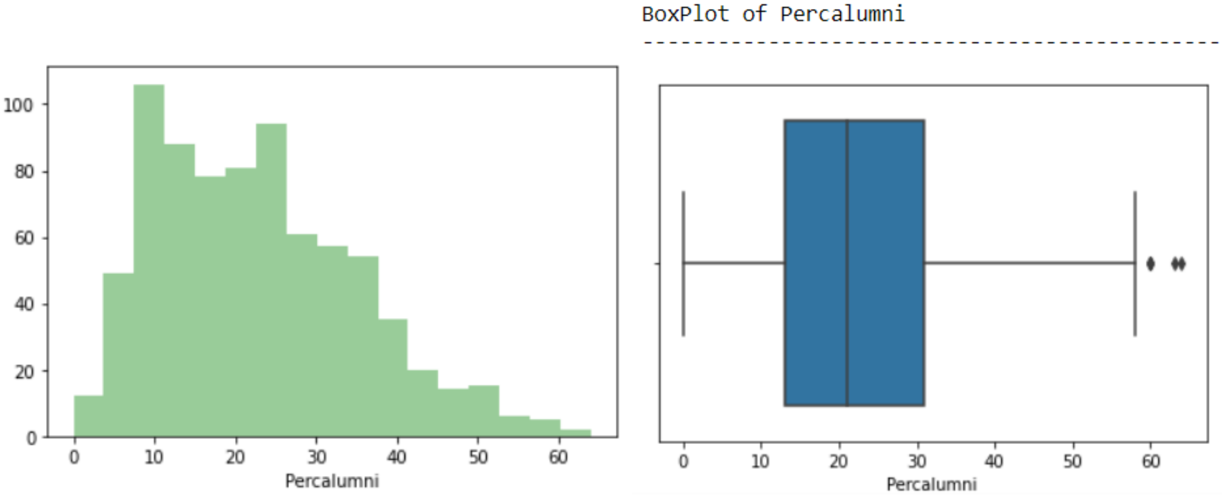
```
BoxPlot of SFRatio
--------------------------------------------------
```

Figure 18. SFRatio_EDA

- Percalumni



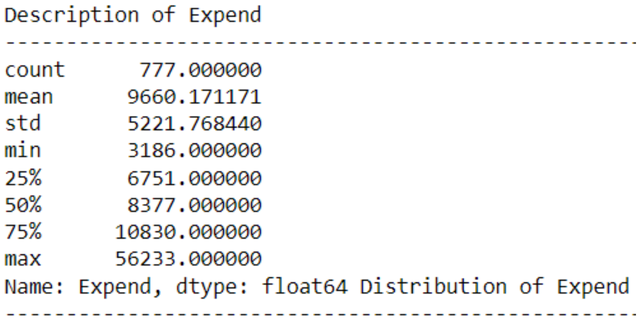
```
Description of Percalumni
----------------------------------------------------------------
count    777.000000
mean      22.743887
std       12.391801
min        0.000000
25%       13.000000
50%       21.000000
75%       31.000000
max       64.000000
Name: Percalumni, dtype: float64 Distribution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



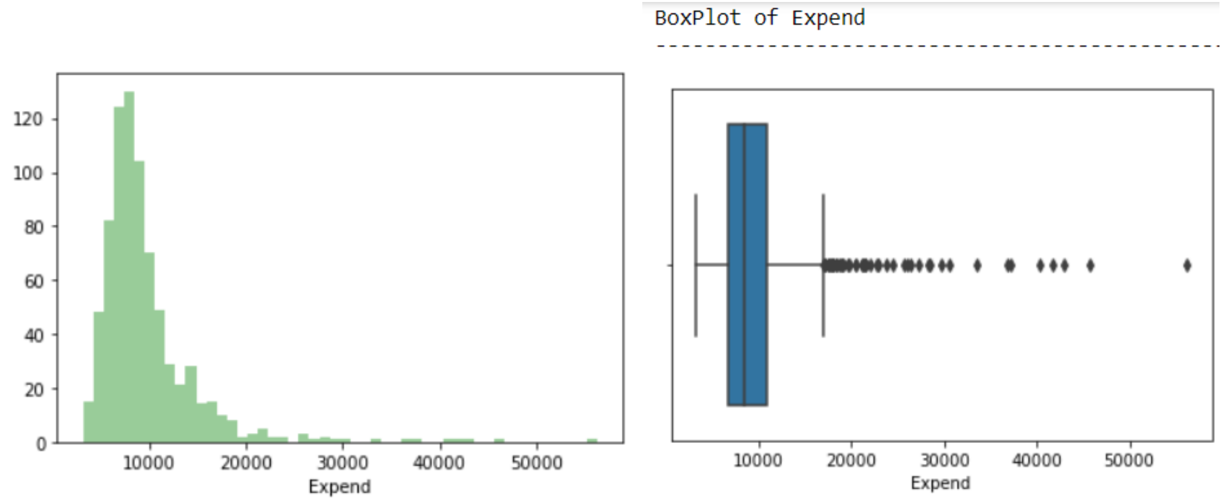
```
BoxPlot of Percalumni
--------------------------------------------------
```

Figure 19. Percalumni_EDA

- Expend



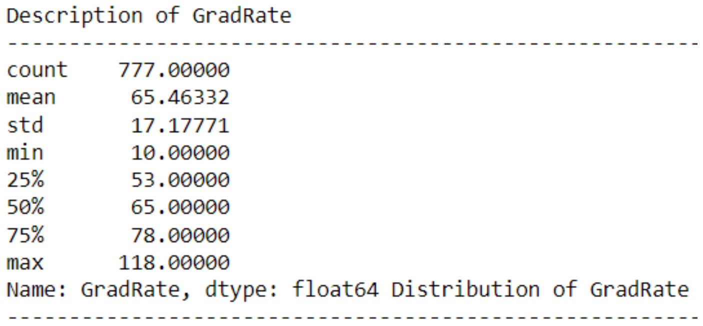
```
Description of Expend
----------------------------------------------------
count      777.000000
mean      9660.171171
std       5221.768440
min       3186.000000
25%       6751.000000
50%       8377.000000
75%      10830.000000
max      56233.000000
Name: Expend, dtype: float64 Distribution of Expend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



```
BoxPlot of Expend
-----------------------------------------------
```

Figure 20. Expend_EDA

- GradRate



```
Description of GradRate
--------------------------------------------------------
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: GradRate, dtype: float64 Distribution of GradRate
--------------------------------------------------------
```

```
BoxPlot of GradRate
-------------------------------------------------
```

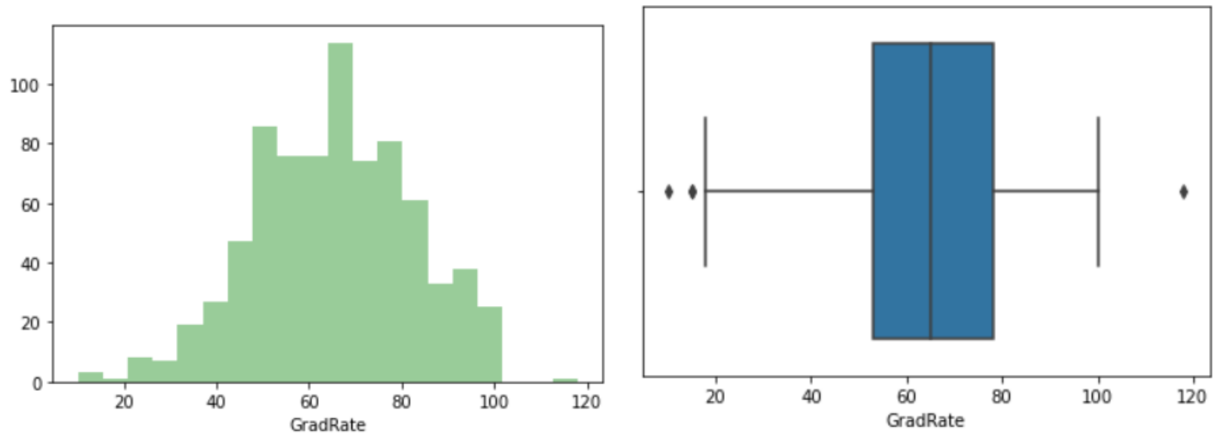
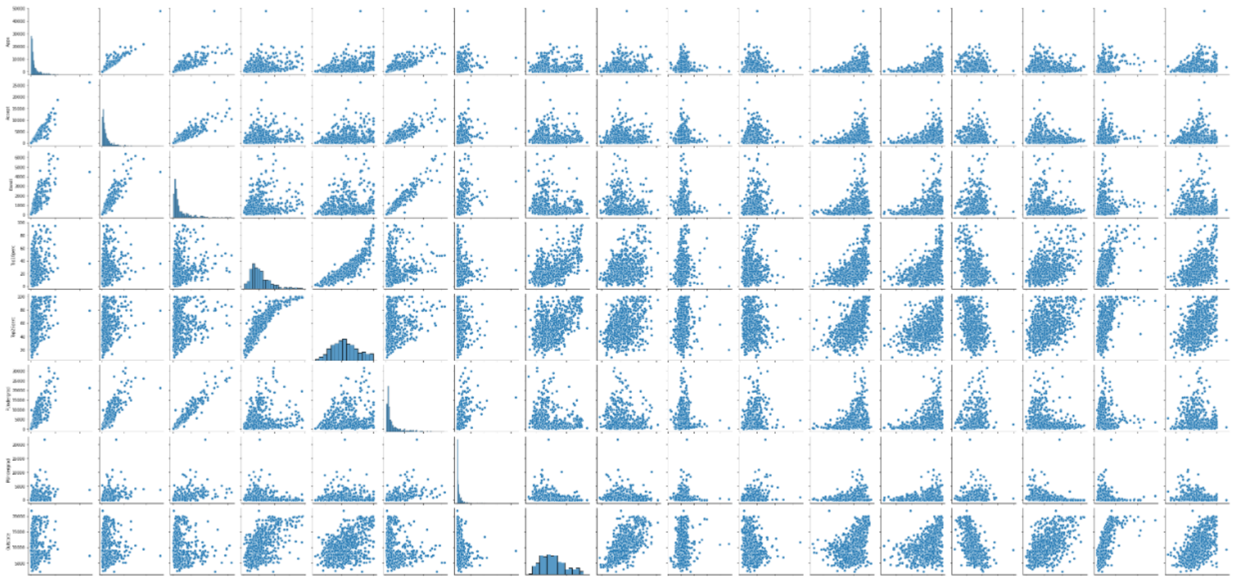


Figure 21. GradRate_EDA

Multivariate Analysis



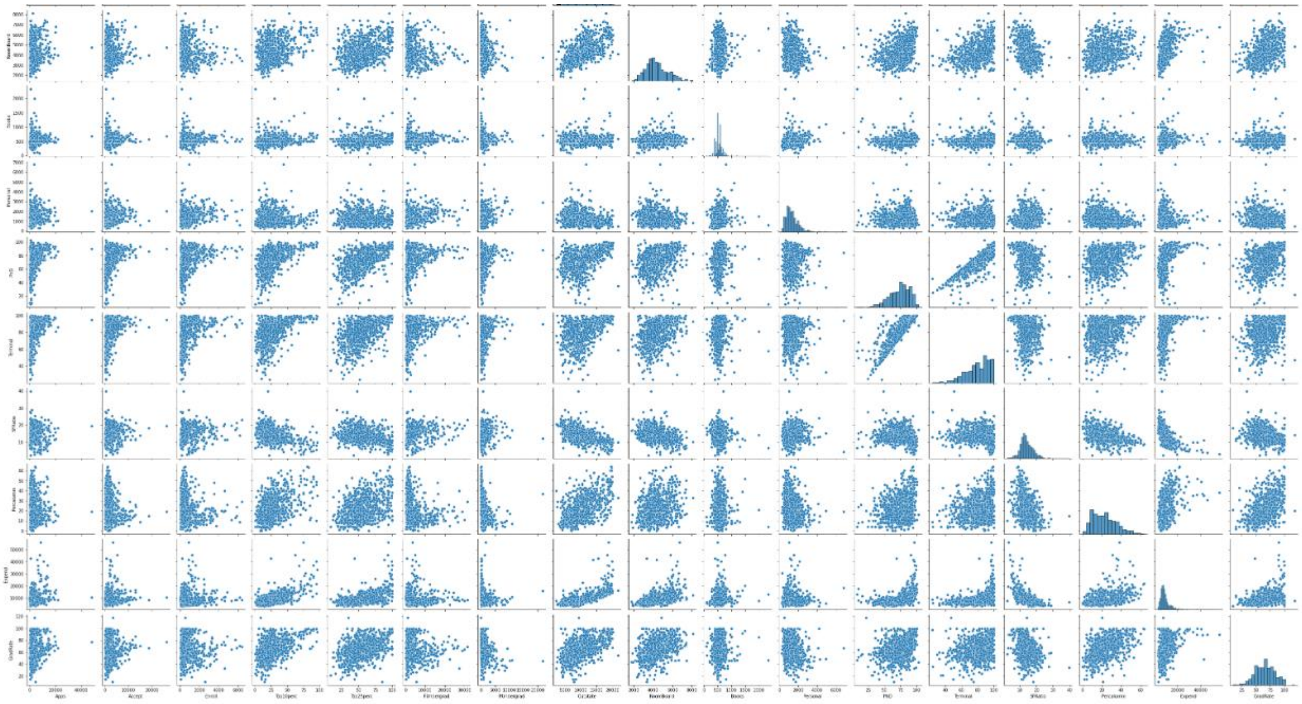


Figure 22. Education_Dataset_Pairplot

Correlation Heatmap

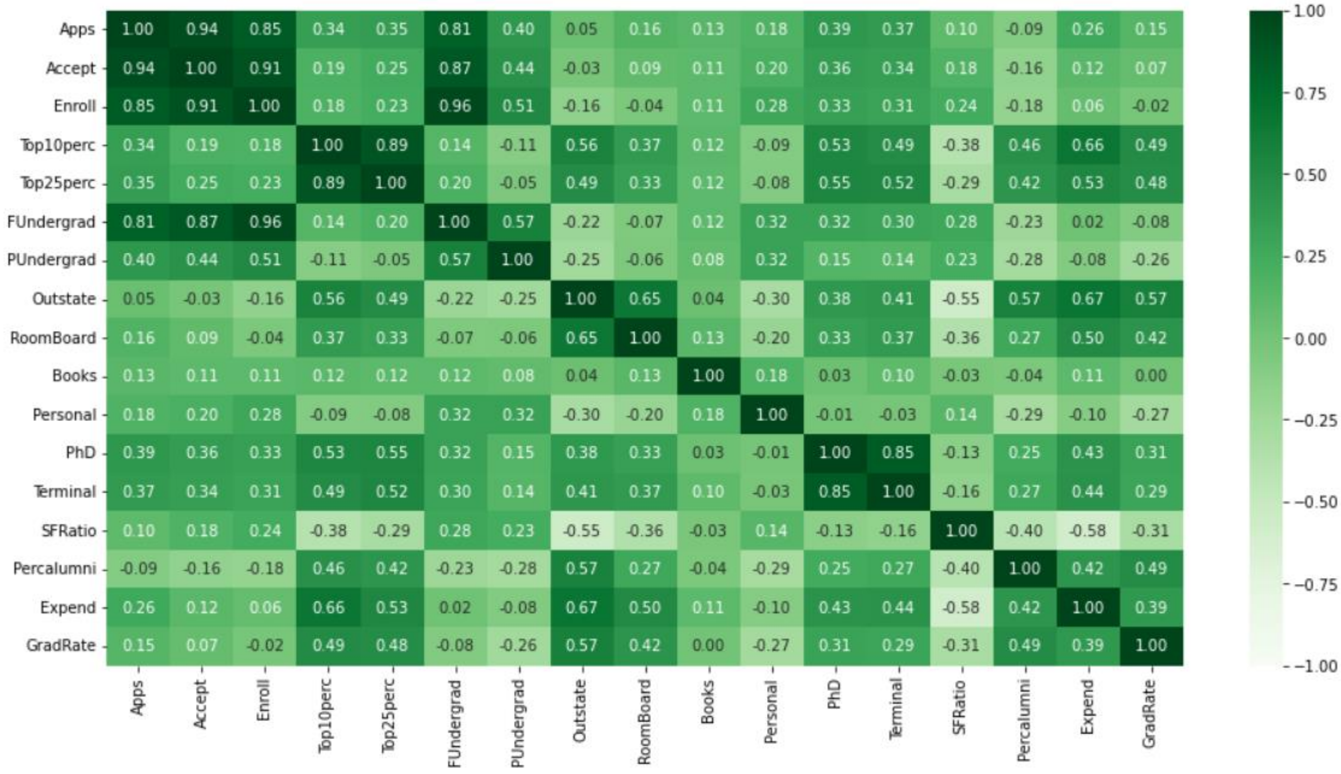


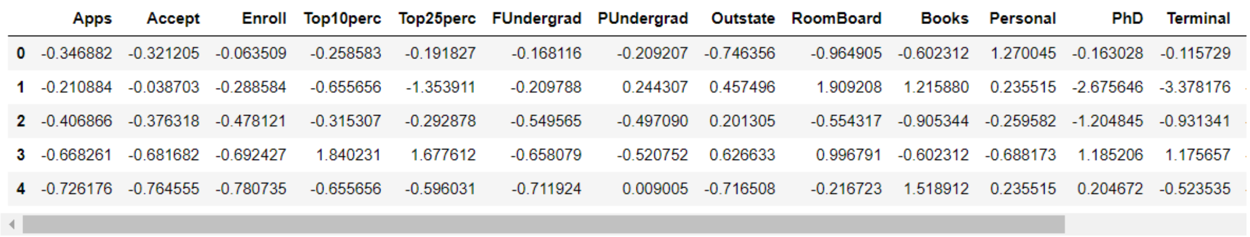
Figure 23. Education_Dataset_Correlation_1

Insights: --

- The dataset has 18 columns and 777 rows
- There are no Null values as indicated by non-null values
- 16 columns are of the integer data type, 1 each of object and float data type
- There are no duplicate rows in the dataset
- A lot of the features are right skewed (Apps, Accept, Enroll, FUndergrad, PUndergrad, Expend)
- A few features are left skewed (PhD, Terminal)
- A few features seem normally distributed (Top25perc, Outstate, RoomBoard, GradRate)
- All features apart from Top25perc have Outliers as demonstrated by the box plots
- There are a few columns with inconsistent name and need to be amended; F.Undergrad; P.Undergrad; Room.Board; S.F.Ratio have '.', perc.alumni starts with lowercase unlike others
- We have plotted scatter diagrams for all the numerical columns in the dataset. A scatter plot is a visual representation of the degree of correlation between any two columns
- We've also plotted a heatmap to display the numerical values of the degree of correlation between any two columns

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

- Often the variables of the data set are of different scales i.e., one variable is in millions and other in only 100. For e.g., in our data set Applications is having values in thousands and PhD, Terminal, GradRate in just two digits. Since the data in these variables are of different scales, it is tough to compare these variables
- Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing
- In this method, we convert variables with different scales of measurements into a single scale and will be doing this only for the numerical variables
- StandardScaler normalizes the data using the formula (x-mean)/standard deviation
- We can either use StandardScaler for each and every feature or apply the z-score (both methods will give us the same result)
- Below is the dataset after applying the z-scores



	Apps	Accept	Enroll	Top10perc	Top25perc	FUndergrad	PUndergrad	Outstate	RoomBoard	Books	Personal	PhD	Terminal
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.163028	-0.115729
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.675646	-3.378176
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1.204845	-0.931341
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.185206	1.175657
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0.204672	-0.523535

Figure 24. Education_Dataset_Scaled

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

- Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency.
- Correlation is a statistical measure that indicates how strongly two variables are related. The value of covariance lies in the range of -∞ and +∞
- As we can see from the below 2 charts, both are same

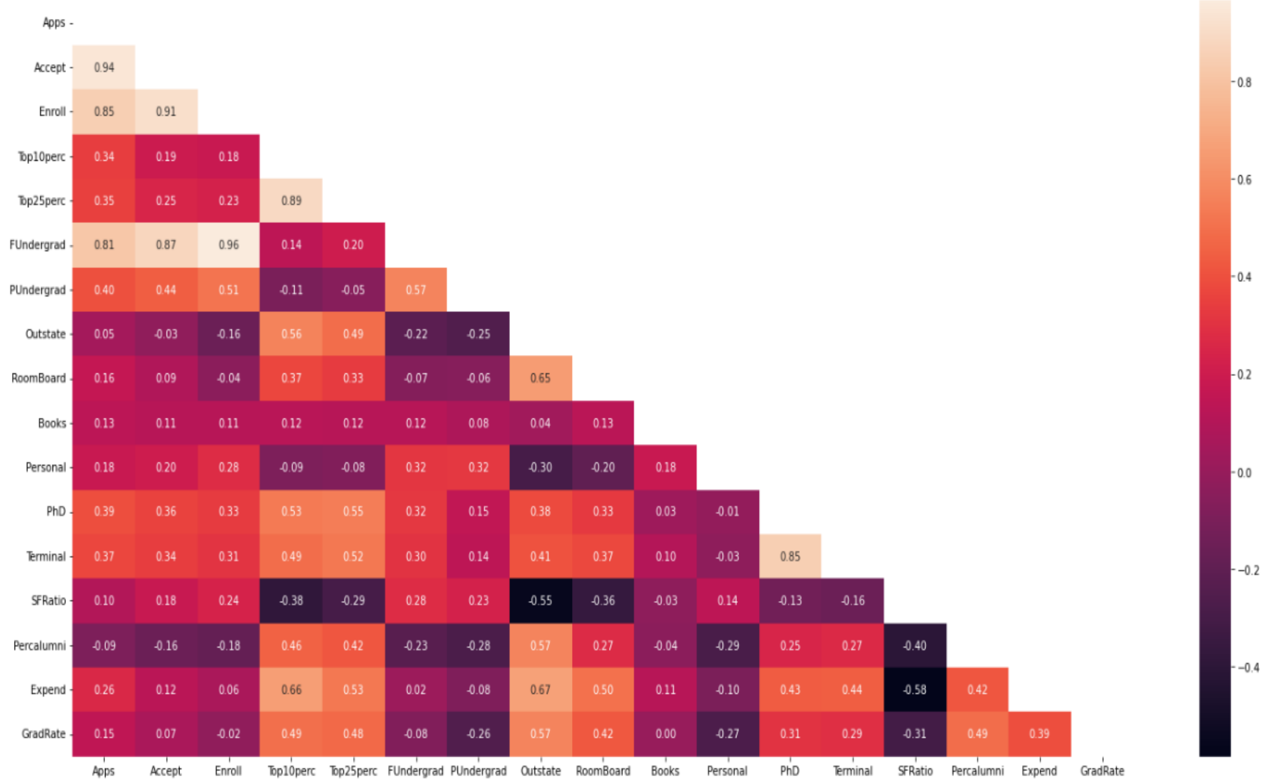
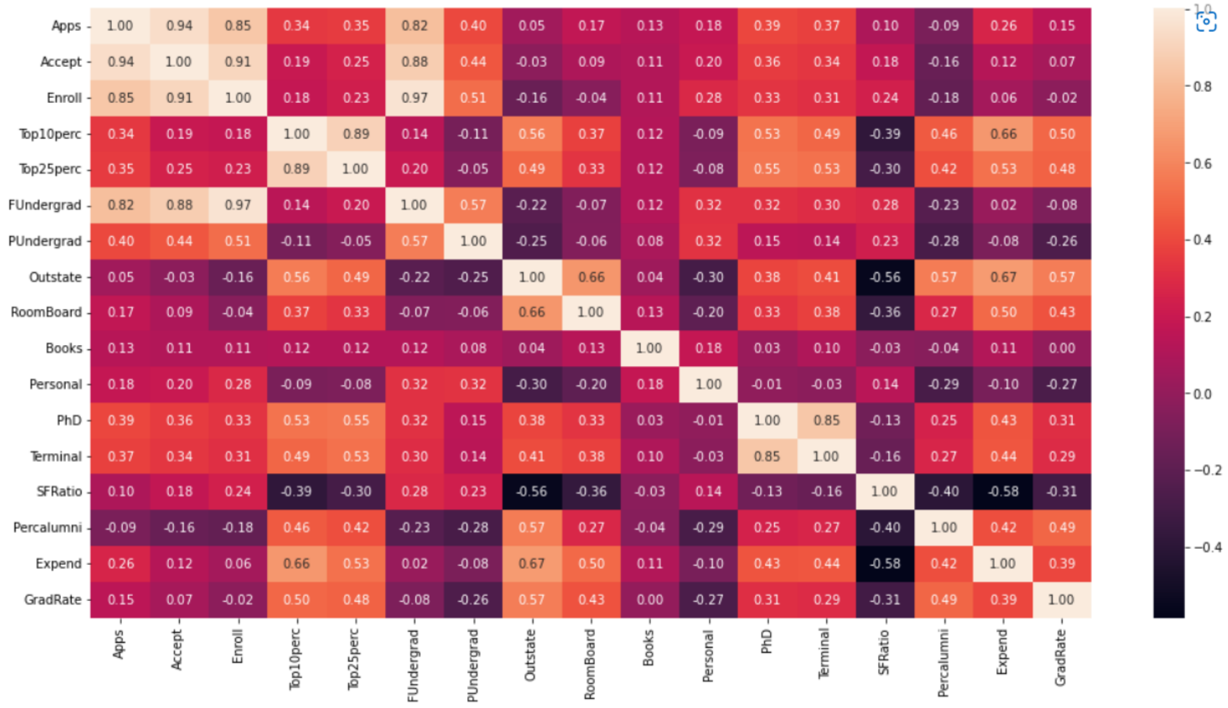


Figure 25. Education_Dataset_ Correlation_2

Covariance matrix for the scaled data



	Apps	Accept	Enroll	Top10perc	Top25perc	FUndergrad	PUndergrad	Outstate	RoomBoard	Books	Personal	PhD	Terminal	SFRatio	Percalumni	Expend	GradRate
Apps	1.00	0.94	0.85	0.34	0.35	0.82	0.40	0.05	0.17	0.13	0.18	0.39	0.37	0.10	-0.09	0.26	0.15
Accept	0.94	1.00	0.91	0.19	0.25	0.88	0.44	-0.03	0.09	0.11	0.20	0.36	0.34	0.18	-0.16	0.12	0.07
Enroll	0.85	0.91	1.00	0.18	0.23	0.97	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	-0.18	0.06	-0.02
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.37	0.12	-0.09	0.53	0.49	-0.39	0.46	0.66	0.50
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33	0.12	-0.08	0.55	0.53	-0.30	0.42	0.53	0.48
FUndergrad	0.82	0.88	0.97	0.14	0.20	1.00	0.57	-0.22	-0.07	0.12	0.32	0.32	0.30	0.28	-0.23	0.02	-0.08
PUndergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	-0.06	0.08	0.32	0.15	0.14	0.23	-0.28	-0.08	-0.26
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.66	0.04	-0.30	0.38	0.41	-0.56	0.57	0.67	0.57
RoomBoard	0.17	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.66	1.00	0.13	-0.20	0.33	0.38	-0.36	0.27	0.50	0.43
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.08	0.04	0.13	1.00	0.18	0.03	0.10	-0.03	-0.04	0.11	0.00
Personal	0.18	0.20	0.28	-0.09	-0.08	0.32	0.32	-0.30	-0.20	0.18	1.00	-0.01	-0.03	0.14	-0.29	-0.10	-0.27
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.03	-0.01	1.00	0.85	-0.13	0.25	0.43	0.31
Terminal	0.37	0.34	0.31	0.49	0.53	0.30	0.14	0.41	0.38	0.10	-0.03	0.85	1.00	-0.16	0.27	0.44	0.29
SFRatio	0.10	0.18	0.24	-0.39	-0.30	0.28	0.23	-0.56	-0.36	-0.03	0.14	-0.13	-0.16	1.00	-0.40	-0.58	-0.31
Percalumni	-0.09	-0.16	-0.18	0.46	0.42	-0.23	-0.28	0.57	0.27	-0.04	-0.29	0.25	0.27	-0.40	1.00	0.42	0.49
Expend	0.26	0.12	0.06	0.66	0.53	0.02	-0.08	0.67	0.50	0.11	-0.10	0.43	0.44	-0.58	0.42	1.00	0.39
GradRate	0.15	0.07	-0.02	0.50	0.48	-0.08	-0.26	0.57	0.43	0.00	-0.27	0.31	0.29	-0.31	0.49	0.39	1.00

Figure 26. Education_Dataset_ Covariance Matrix

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

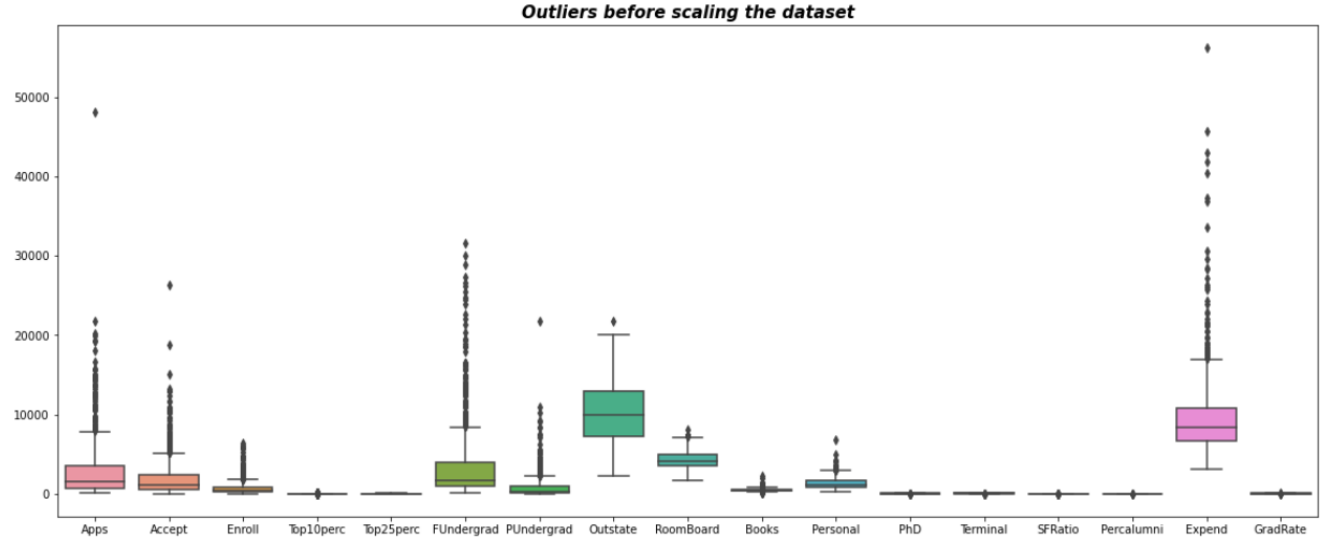

Outliers before scaling the dataset

Figure 27. Education_Dataset_Outliers before Scaling

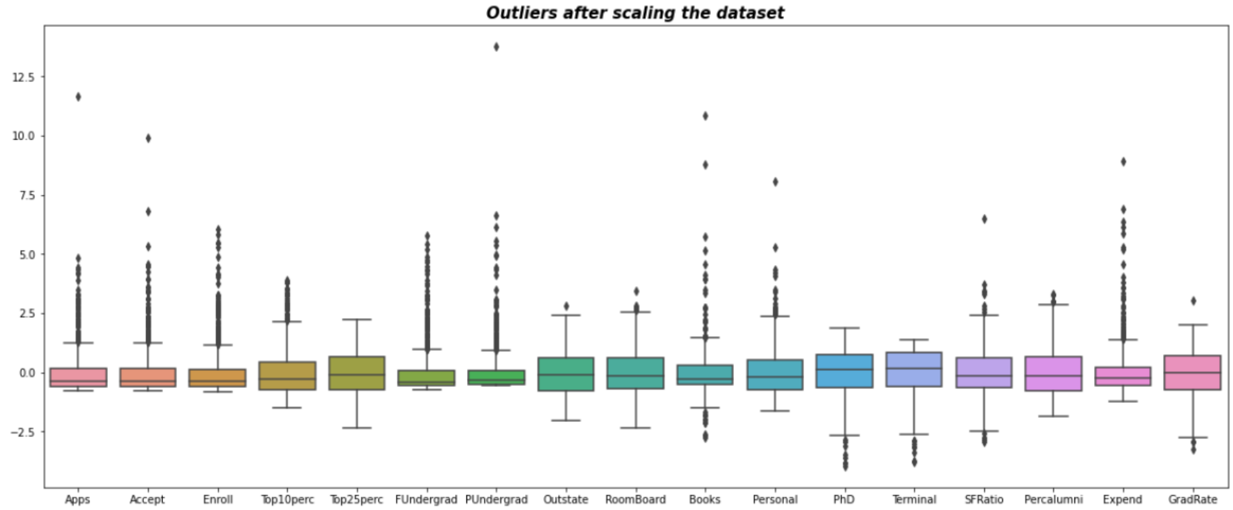
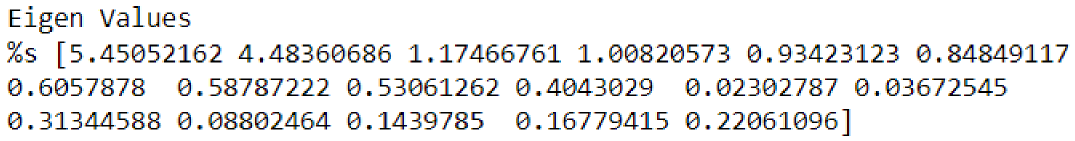


Figure 28. Education_Dataset_Outliers after Scaling

- We can see that the Outliers have reduced in magnitude after scaling
- This is due to applying the z-score
- Once the data is scaled the values are in the range of -3 to +3
- As we haven't treated the dataset for Outliers, we don't see this

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



```
Eigen Values
%s [5.45052162 4.48360686 1.17466761 1.00820573 0.93423123 0.84849117
0.6057878  0.58787222 0.53061262 0.4043029  0.02302787 0.03672545
0.31344588 0.08802464 0.1439785  0.16779415 0.22061096]
```

Figure 29. Education_Dataset_Eigen Values

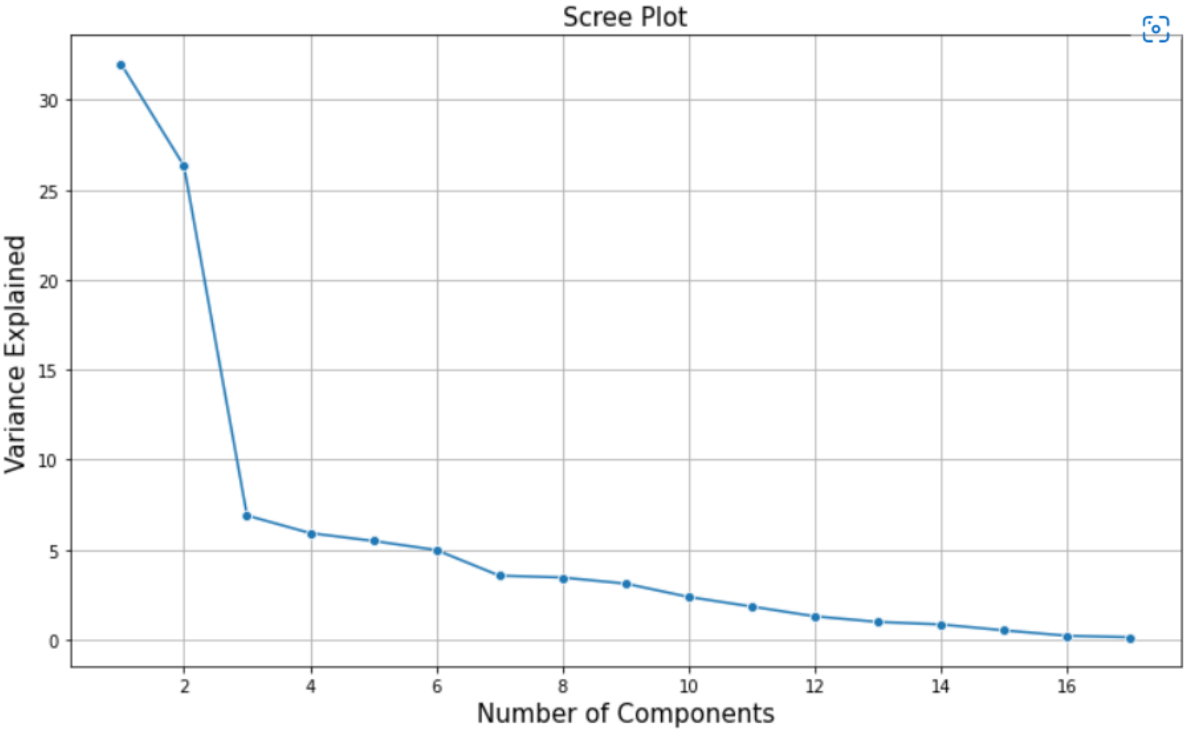
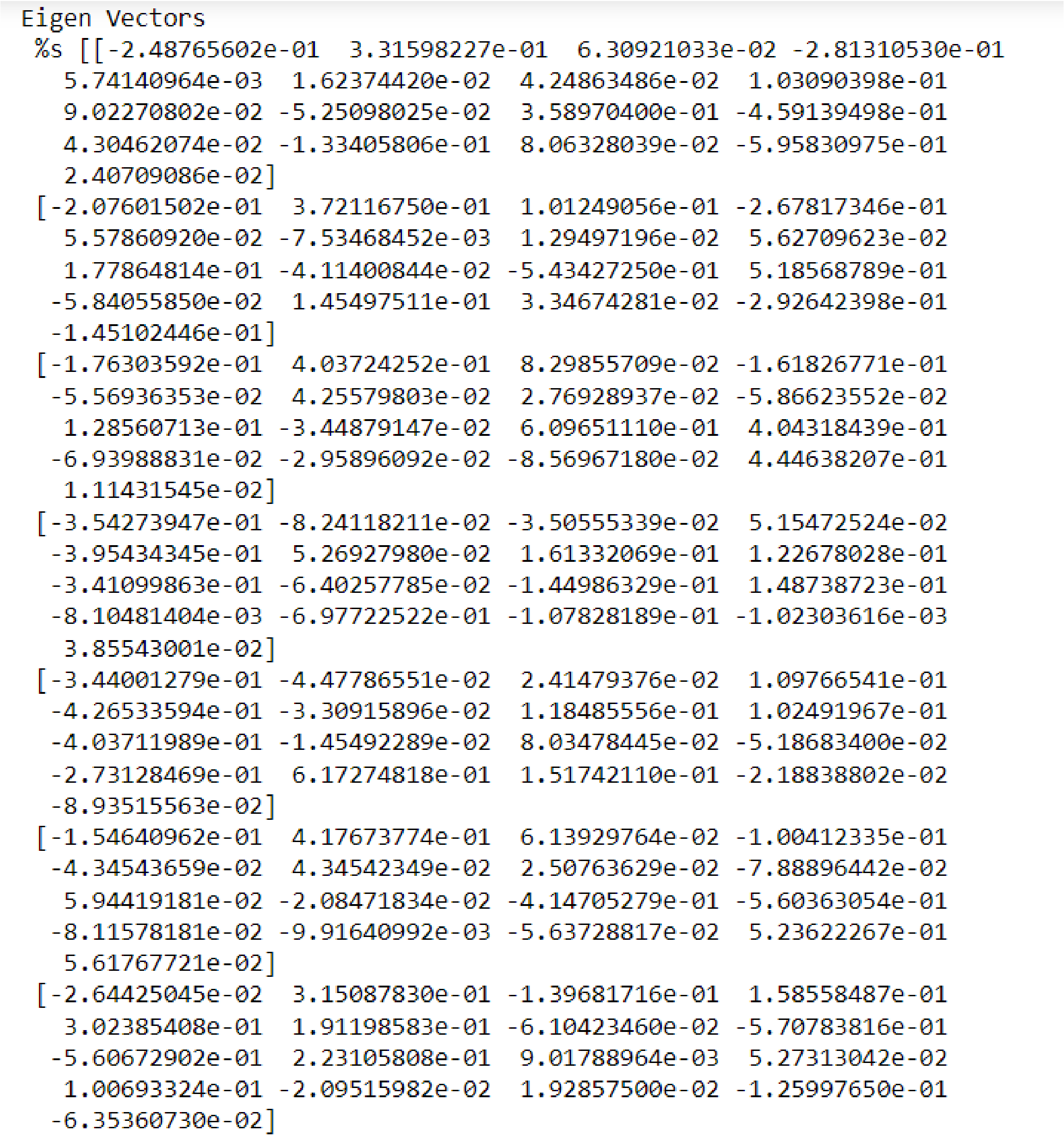
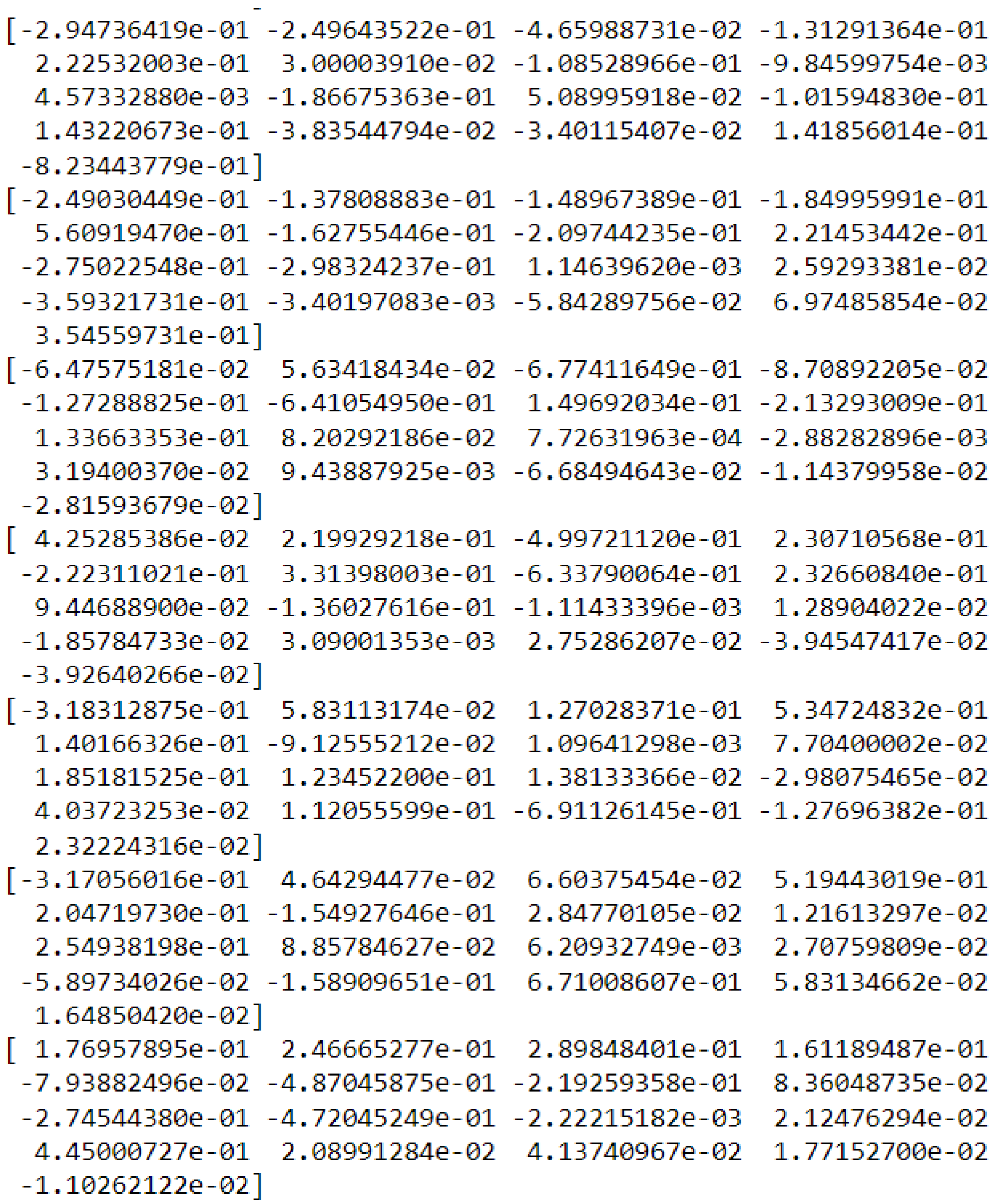


Figure 30. Education_Dataset_Scree Plot

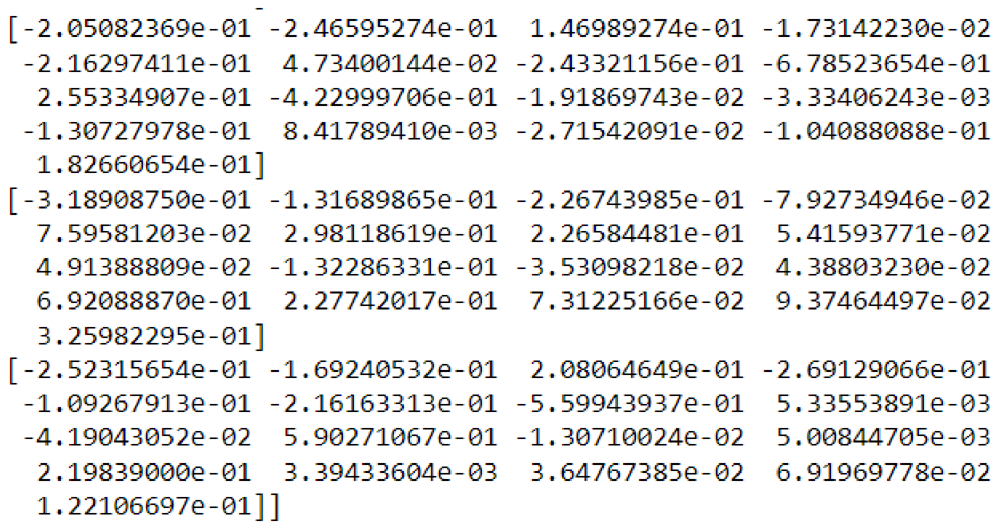
- A scree plot always displays the eigenvalues in a downward curve, ordering the eigenvalues from largest to smallest



```
Eigen Vectors
 %s [[-2.48765602e-01  3.31598227e-01  6.30921033e-02 -2.81310530e-01
   5.74140964e-03  1.62374420e-02  4.24863486e-02  1.03090398e-01
   9.02270802e-02 -5.25098025e-02  3.58970400e-01 -4.59139498e-01
   4.30462074e-02 -1.33405806e-01  8.06328039e-02 -5.95830975e-01
   2.40709086e-02]
 [-2.07601502e-01  3.72116750e-01  1.01249056e-01 -2.67817346e-01
   5.57860920e-02 -7.53468452e-03  1.29497196e-02  5.62709623e-02
   1.77864814e-01 -4.11400844e-02 -5.43427250e-01  5.18568789e-01
  -5.84055850e-02  1.45497511e-01  3.34674281e-02 -2.92642398e-01
  -1.45102446e-01]
 [-1.76303592e-01  4.03724252e-01  8.29855709e-02 -1.61826771e-01
  -5.56936353e-02  4.25579803e-02  2.76928937e-02 -5.86623552e-02
   1.28560713e-01 -3.44879147e-02  6.09651110e-01  4.04318439e-01
  -6.93988831e-02 -2.95896092e-02 -8.56967180e-02  4.44638207e-01
   1.11431545e-02]
 [-3.54273947e-01 -8.24118211e-02 -3.50555339e-02  5.15472524e-02
  -3.95434345e-01  5.26927980e-02  1.61332069e-01  1.22678028e-01
  -3.41099863e-01 -6.40257785e-02 -1.44986329e-01  1.48738723e-01
  -8.10481404e-03 -6.97722522e-01 -1.07828189e-01 -1.02303616e-03
   3.85543001e-02]
 [-3.44001279e-01 -4.47786551e-02  2.41479376e-02  1.09766541e-01
  -4.26533594e-01 -3.30915896e-02  1.18485556e-01  1.02491967e-01
  -4.03711989e-01 -1.45492289e-02  8.03478445e-02 -5.18683400e-02
  -2.73128469e-01  6.17274818e-01  1.51742110e-01 -2.18838802e-02
  -8.93515563e-02]
 [-1.54640962e-01  4.17673774e-01  6.13929764e-02 -1.00412335e-01
  -4.34543659e-02  4.34542349e-02  2.50763629e-02 -7.88896442e-02
   5.94419181e-02 -2.08471834e-02 -4.14705279e-01 -5.60363054e-01
  -8.11578181e-02 -9.91640992e-03 -5.63728817e-02  5.23622267e-01
   5.61767721e-02]
 [-2.64425045e-02  3.15087830e-01 -1.39681716e-01  1.58558487e-01
   3.02385408e-01  1.91198583e-01 -6.10423460e-02 -5.70783816e-01
  -5.60672902e-01  2.23105808e-01  9.01788964e-03  5.27313042e-02
   1.00693324e-01 -2.09515982e-02  1.92857500e-02 -1.25997650e-01
  -6.35360730e-02]
```



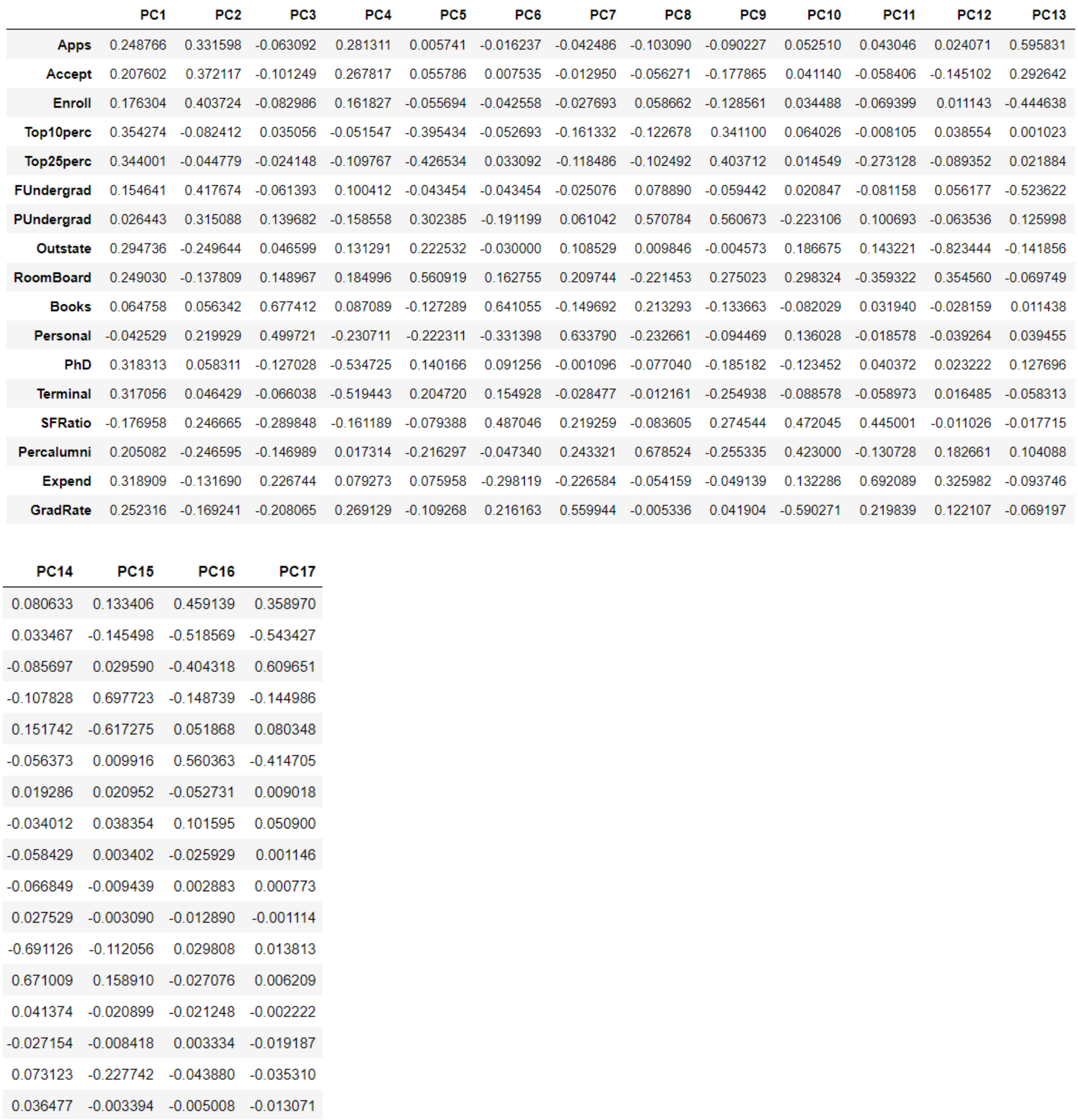
```
[-2.94736419e-01 -2.49643522e-01 -4.65988731e-02 -1.31291364e-01
  2.22532003e-01  3.00003910e-02 -1.08528966e-01 -9.84599754e-03
  4.57332880e-03 -1.86675363e-01  5.08995918e-02 -1.01594830e-01
  1.43220673e-01 -3.83544794e-02 -3.40115407e-02  1.41856014e-01
 -8.23443779e-01]
[-2.49030449e-01 -1.37808883e-01 -1.48967389e-01 -1.84995991e-01
  5.60919470e-01 -1.62755446e-01 -2.09744235e-01  2.21453442e-01
 -2.75022548e-01 -2.98324237e-01  1.14639620e-03  2.59293381e-02
 -3.59321731e-01 -3.40197083e-03 -5.84289756e-02  6.97485854e-02
  3.54559731e-01]
[-6.47575181e-02  5.63418434e-02 -6.77411649e-01 -8.70892205e-02
 -1.27288825e-01 -6.41054950e-01  1.49692034e-01 -2.13293009e-01
  1.33663353e-01  8.20292186e-02  7.72631963e-04 -2.88282896e-03
  3.19400370e-02  9.43887925e-03 -6.68494643e-02 -1.14379958e-02
 -2.81593679e-02]
[ 4.25285386e-02  2.19929218e-01 -4.99721120e-01  2.30710568e-01
 -2.22311021e-01  3.31398003e-01 -6.33790064e-01  2.32660840e-01
  9.44688900e-02 -1.36027616e-01 -1.11433396e-03  1.28904022e-02
 -1.85784733e-02  3.09001353e-03  2.75286207e-02 -3.94547417e-02
 -3.92640266e-02]
[-3.18312875e-01  5.83113174e-02  1.27028371e-01  5.34724832e-01
  1.40166326e-01 -9.12555212e-02  1.09641298e-03  7.70400002e-02
  1.85181525e-01  1.23452200e-01  1.38133366e-02 -2.98075465e-02
  4.03723253e-02  1.12055599e-01 -6.91126145e-01 -1.27696382e-01
  2.32224316e-02]
[-3.17056016e-01  4.64294477e-02  6.60375454e-02  5.19443019e-01
  2.04719730e-01 -1.54927646e-01  2.84770105e-02  1.21613297e-02
  2.54938198e-01  8.85784627e-02  6.20932749e-03  2.70759809e-02
 -5.89734026e-02 -1.58909651e-01  6.71008607e-01  5.83134662e-02
  1.64850420e-02]
[ 1.76957895e-01  2.46665277e-01  2.89848401e-01  1.61189487e-01
 -7.93882496e-02 -4.87045875e-01 -2.19259358e-01  8.36048735e-02
 -2.74544380e-01 -4.72045249e-01 -2.22215182e-03  2.12476294e-02
  4.45000727e-01  2.08991284e-02  4.13740967e-02  1.77152700e-02
 -1.10262122e-02]
```



```
[-2.05082369e-01 -2.46595274e-01  1.46989274e-01 -1.73142230e-02
 -2.16297411e-01  4.73400144e-02 -2.43321156e-01 -6.78523654e-01
  2.55334907e-01 -4.22999706e-01 -1.91869743e-02 -3.33406243e-03
 -1.30727978e-01  8.41789410e-03 -2.71542091e-02 -1.04088088e-01
  1.82660654e-01]
[-3.18908750e-01 -1.31689865e-01 -2.26743985e-01 -7.92734946e-02
  7.59581203e-02  2.98118619e-01  2.26584481e-01  5.41593771e-02
  4.91388809e-02 -1.32286331e-01 -3.53098218e-02  4.38803230e-02
  6.92088870e-01  2.27742017e-01  7.31225166e-02  9.37464497e-02
  3.25982295e-01]
[-2.52315654e-01 -1.69240532e-01  2.08064649e-01 -2.69129066e-01
 -1.09267913e-01 -2.16163313e-01 -5.59943937e-01  5.33553891e-03
 -4.19043052e-02  5.90271067e-01 -1.30710024e-02  5.00844705e-03
  2.19839000e-01  3.39433604e-03  3.64767385e-02  6.91969778e-02
  1.22106697e-01]]
```

Figure 31. Education_Dataset_Eigen Vectors

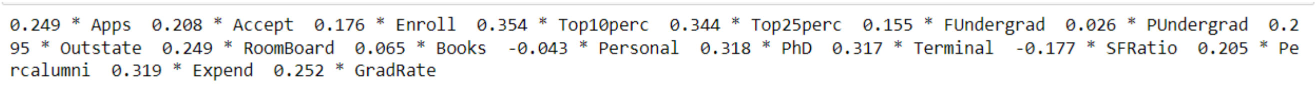
2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.103090	-0.090227	0.052510	0.043046	0.024071	0.595831	0.080633	0.133406	0.459139	0.358970
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950	-0.056271	-0.177865	0.041140	-0.058406	-0.145102	0.292642	0.033467	-0.145498	-0.518569	-0.543427
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693	0.058662	-0.128561	0.034488	-0.069399	0.011143	-0.444638	-0.085697	0.029590	-0.404318	0.609651
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678	0.341100	0.064026	-0.008105	0.038554	0.001023	-0.107828	0.697723	-0.148739	-0.144986
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486	-0.102492	0.403712	0.014549	-0.273128	-0.089352	0.021884	0.151742	-0.617275	0.051868	0.080348
FUndergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076	0.078890	-0.059442	0.020847	-0.081158	0.056177	-0.523622	-0.056373	0.009916	0.560363	-0.414705
PUndergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042	0.570784	0.560673	-0.223106	0.100693	-0.063536	0.125998	0.019286	0.020952	-0.052731	0.009018
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529	0.009846	-0.004573	0.186675	0.143221	-0.823444	-0.141856	-0.034012	0.038354	0.101595	0.050900
RoomBoard	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744	-0.221453	0.275023	0.298324	-0.359322	0.354560	-0.069749	-0.058429	0.003402	-0.025929	0.001146
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692	0.213293	-0.133663	-0.082029	0.031940	-0.028159	0.011438	-0.066849	-0.009439	0.002883	0.000773
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790	-0.232661	-0.094469	0.136028	-0.018578	-0.039264	0.039455	0.027529	-0.003090	-0.012890	-0.001114
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096	-0.077040	-0.185182	-0.123452	0.040372	0.023222	0.127696	-0.691126	-0.112056	0.029808	0.013813
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477	-0.012161	-0.254938	-0.088578	-0.058973	0.016485	-0.058313	0.671009	0.158910	-0.027076	0.006209
SFRatio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	0.219259	-0.083605	0.274544	0.472045	0.445001	-0.011026	-0.017715	0.041374	-0.020899	-0.021248	-0.002222
Percalumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321	0.678524	-0.255335	0.423000	-0.130728	0.182661	0.104088	-0.027154	-0.008418	0.003334	-0.019187
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584	-0.054159	-0.049139	0.132286	0.692089	0.325982	-0.093746	0.073123	-0.227742	-0.043880	-0.035310
GradRate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944	-0.005336	0.041904	-0.590271	0.219839	0.122107	-0.069197	0.036477	-0.003394	-0.005008	-0.013071

Figure 32. Education_Dataset_PCA

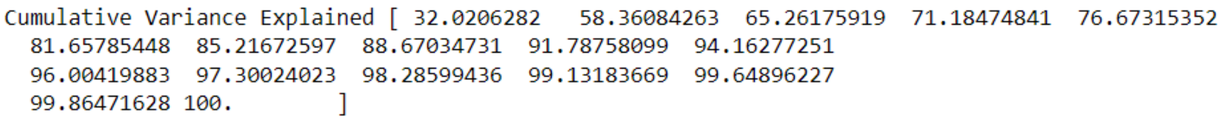
2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]



```
0.249 * Apps  0.208 * Accept  0.176 * Enroll  0.354 * Top10perc  0.344 * Top25perc  0.155 * FUndergrad  0.026 * PUndergrad  0.2
95 * Outstate  0.249 * RoomBoard  0.065 * Books  -0.043 * Personal  0.318 * PhD  0.317 * Terminal  -0.177 * SFRatio  0.205 * Pe
rcalumni  0.319 * Expend  0.252 * GradRate
```

Figure 33. Education_Dataset_PCA_Linear Equation

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



```
Cumulative Variance Explained [ 32.0206282   58.36084263  65.26175919  71.18474841  76.67315352
  81.65785448  85.21672597  88.67034731  91.78758099  94.16277251
  96.00419883  97.30024023  98.28599436  99.13183669  99.64896227
  99.86471628 100.        ]
```

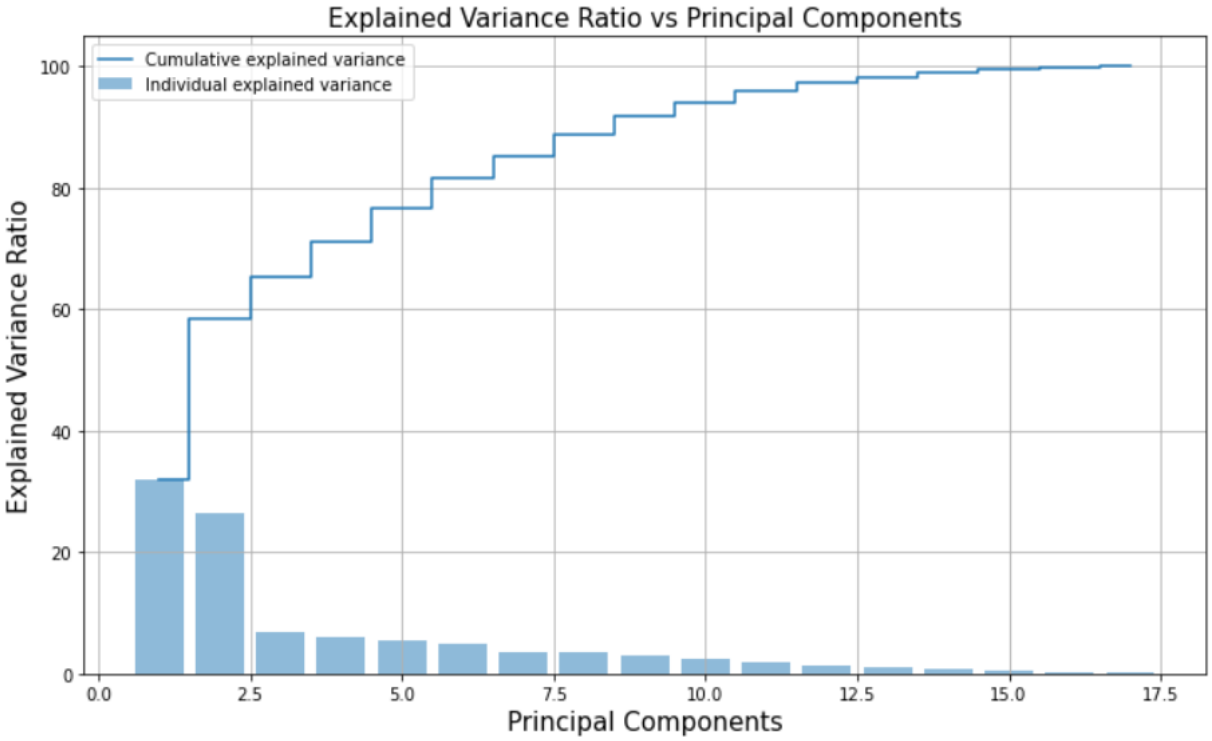


Figure 34. Education_Dataset_Cumulative Explained Variance

- PCA is a method that:
  - Measures how each variable is associated with one another using a Covariance matrix
  - Understands the directions of the spread of our data using Eigenvectors
  - Brings out the relative importance of these directions using Eigenvalues

    - Eigenvalues are coefficients applied to eigenvectors that give the vectors their length or magnitude
- The Cumulative values of the eigen values explain how much of the variance is explained by each component
- As we can see from the above table and chart, the first 7 components explain 85.22% of the variance
- Hence, although we have 17 features, we only need 7 of them to explain the variance in the dataset without losing too much information

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 Summary:--

- PCA forms the basis of multivariate data analysis
- Principal Component Analysis (PCA) is a dimensionality reduction technique, which is used for identification of a smaller number of uncorrelated variables known as Principal Components from a larger set of data
- PCA is important as it not only helps in cleaning up the data, but also in reducing the number of features used to observe trends, clusters, outliers and draw conslusions from a dataset
- It uncovers the relationships between observations and variables, and among the variables
- The column names are all different; some have '.' in them and some are in lowercase; getting them in a uniform format is a good start to analysing the data
- Once the data clean-up is done, we create another dataframe with only the numeric columns (float and integer data type) as columns with text don't really help
- However Categorical data which are repetitive in nature can be used; e.g. Gender, Blood type, Country, Education
- With this new dataframe we check for any duplicate records, null values in any of the columns and treat it for Outliers
- We now scale the data so all columns are in a uniform format and easily comparable
- Now, we can get the eigen values and eigen vectors, chart the Scree plot; Cumulative explained variance and individual explained variance vs Principal Components
- As depicted by the eigenvalues, 7 out of the 17 components explain 85.22% of the variance and we'll only need these 7 features to observe trends and patterns

Business Implications:--

- The first 3 columns 'No. of applications', 'No. of students accepted', 'No. of students enrolled' is a good place to start
    - 67% of the applicants are accepted into the colleges
    - 26% of the applicants and 39% of the accepted students are new, which is a good sign
    - Thereby Accept(0.94) and Enroll(0.85) are highly correlated with Apps

  - The business should look at ways to increase the number of new applicants by offerring more scholarships, providing testimonials and case studies, placement history and opportunities are good measures
- Enroll, Top10perc and Top25perc are inter-related as the latter two are part of the new students enrolled; one can drop the latter 2 or all 3 columns before proceeding with scaling the data
- There seem to be 1 part-time student for every 4 full-time graduate student, which is reasonable
- The 'Outstate' is another column that can be removed from our analysis for 2 reasons
  - The number is way higher than the applications received suggesting that this value is for the overall students in a college rather than the new applicants
  - This data is unrelated to our analysis which focusses more on new applicants
- The next 3 columns, RoomBoard, Books and Personal are vital to our research as these are important considerations for a student before taking up a course
  - The colleges should help students by providing various jobs within the campus, give recommendation letters, ensure food is available at reasonable prices
- The next 2 columns 'PhD' and 'Terminal' are again inter-related as PhD is the highest achievable degree in most academic fields; either of the 2 columns may be dropped in this case before proceeding with scaling the data
  - 73% of the faculty have a PhD and 80% have a terminal degree, which bodes really well for the colleges as these are high numbers
- The Student/faculty ratio is again a good criteria to evaluate a college from the students perspective
  - 14 students for every professor seems to be the average, and 75% of the values are below 17 students
  - Both are reasonable ratios, however there is scope for improvement here to get this closer to 10
- The percalumni, the percentage of alumni who donate is good at 23%; however the colleges could do more to increase this average above the 30% mark
- The Instructional expenditure per student, Expend is a measure of the public investment that a country devotes annually, on average, to each student's education
  - The average of 9,660 is really good and indicates it's in the best interests of the country to spend on educating it's people
- The Graduation rate at 65%, although good, can be improved upon by the colleges
  - A more in-depth research needs to be conducted on the reasons for students to not graduate
  - The colleges can set-up counselling sessions for students to help manage their time better, have doubt clearing sessions from seniors and professors and assign mentors who can guide the students to do well in their assignments