STUDENT PCA-ANALYSIS

& HYPOTHESIS TESTING USING ANOVA

Business report

S NITIN KUMAR

nithinkumar650@gmai.com

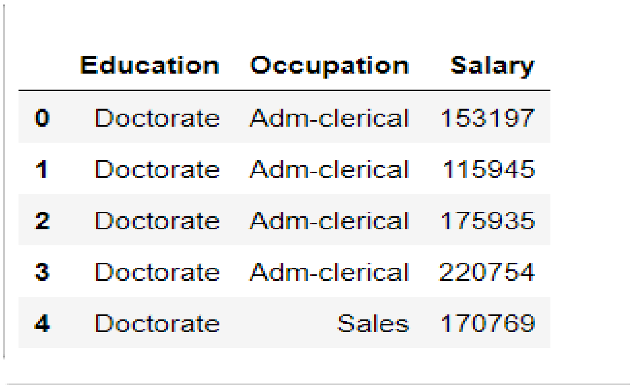
Table of Contents

Problem 1A:

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals. are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

Data description

Sample of the dataset:



	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769

Data set has 3 variables with 3 Education levels and 4 Occupation levels along with their Salary. In total the data set contain 40 rows and 3 columns

1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

One way ANOVA(Education)

Null Hypothesis $H0$: The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, HS-Grad).

Alternate Hypothesis $H1$: The mean salary is different in at least one category of education.

One way ANOVA(**Occupation**)

Null Hypothesis $H0$: The mean salary is the same across all the 4 categories of occupation(Prof-Specialty, Sales, Adm-clerical, Exec-Managerial).

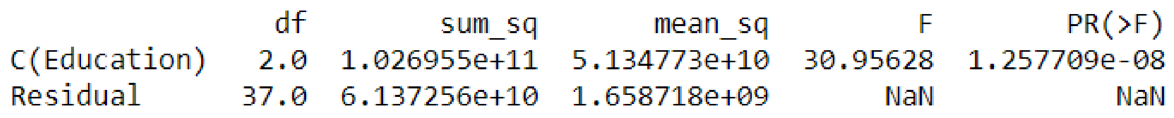
Alternate Hypothesis $H1$: The mean salary is different in at least one category of occupation.

1.2 Perform one-way ANOVA for Education with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

One way ANOVA(Education)

Null Hypothesis $H0$: The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, HS-Grad).

Alternate Hypothesis $H1$: The mean salary is different in at least one category of education.



```
                 df        sum_sq       mean_sq          F        PR(>F)
C(Education)    2.0  1.026955e+11  5.134773e+10  30.95628  1.257709e-08
Residual       37.0  6.137256e+10  1.658718e+09       NaN           NaN
```

The above is the ANOVA table for Education variable.

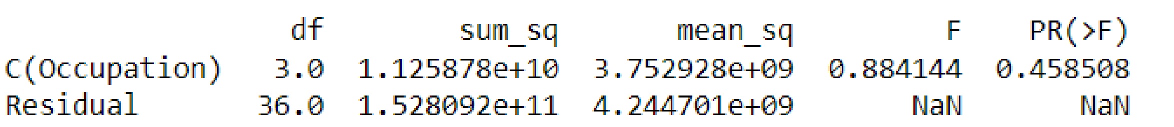
Since the p value = 1.257709e-08 is less than the significance level (alpha = 0.05), we can reject the null hypothesis and conclude that there is a significant difference in the mean salaries for at least one category of education.

1.3 Perform one-way ANOVA for variable Occupation with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

One way ANOVA(**Occupation**)

Null Hypothesis H0: The mean salary is the same across all the 4 categories of occupation (Prof-Specialty, Sales, Adm-clerical, Exec-Managerial).

Alternate Hypothesis H1: The mean salary is different in at least one category of occupation.

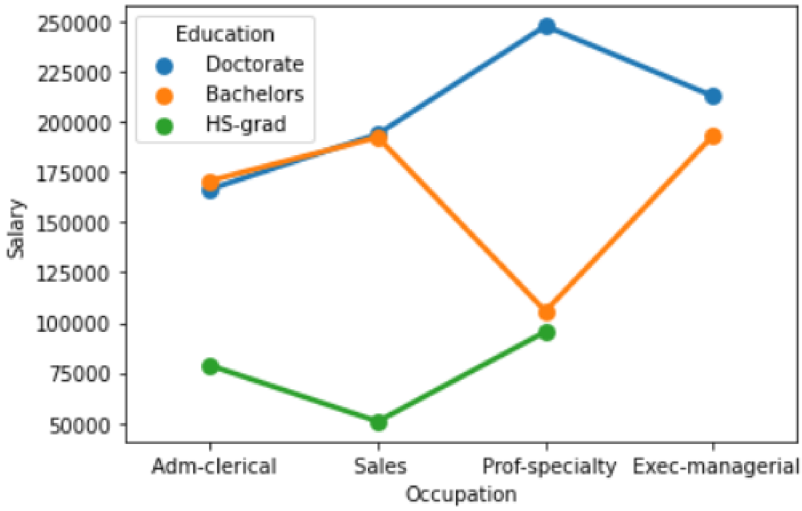


```
                  df        sum_sq       mean_sq         F    PR(>F)
C(Occupation)    3.0  1.125878e+10  3.752928e+09  0.884144  0.458508
Residual        36.0  1.528092e+11  4.244701e+09       NaN       NaN
```

Since the p value = 0.458508 is greater than the significance level (alpha = 0.05), we fail to reject the null hypothesis (i.e., we accept H0) and conclude that there is no significant difference in the mean salaries across the 4 categories of occupation

1.4 If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result.

1.5 What is the interaction between the two treatments? Analyse the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.

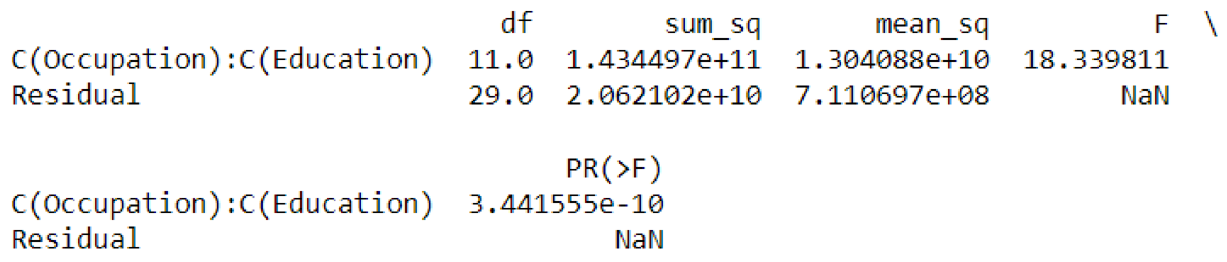


- From the above plot we can say that people who have done education as doctorate and choosing the occupation as adm-clerical and sales earn same as the people who done education in bachelors
- People who have done education in HS-grad and choosing the occupation as AD-clerical earn lesser then others and they have no access to do exec-managerial post
- people who have done education as doctorate and choosing the occupation as prof-speciality earn way higher than the people who have done education in bachelors and choosing the same occupation as prof-specialty
- people who have done education as doctorate and choosing the occupation as exec-managerial earn slightly higher the people who have done education in bachelor and choosing the occupation as exec-managerial

1.6 Perform a two-way ANOVA based on the Education and Occupation (along with their interaction Education*Occupation) with the variable 'Salary'. State the null and alternative hypotheses and state your results. How will you interpret this result?

$H0$: The effect of the independent variable 'education' on the mean 'salary' does not depend on the effect of the other independent variable 'occupation' (i. e. there is no interaction effect between the 2 independent variables, education and occupation).

$H1$: There is an interaction effect between the independent variable 'education' and the independent variable 'occupation' on the mean salary



```
                              df        sum_sq       mean_sq          F  \
C(Occupation):C(Education)  11.0  1.434497e+11  1.304088e+10  18.339811
Residual                    29.0  2.062102e+10  7.110697e+08        NaN

                                  PR(>F)
C(Occupation):C(Education)  3.441555e-10
Residual                             NaN
```

As p value = 2.232500e-05 is lesser than the significance level (alpha = 0.05), we reject the null hypothesis.

Thus, we see that there is an interaction effect between education and occupation on the mean salary.

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

Business implementation of performing ANOVA for this particular case study is that from ANOVA and interaction plot we can see that education and occupation both combined yield the higher and good salary among the people

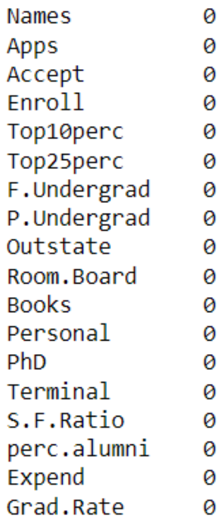
We can see clearly that people with education as doctorate draw the maximum salaries when compares to the people with education HS-grade who earn the least.

Problem 2:

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. The data dictionary of the 'Education

Data description

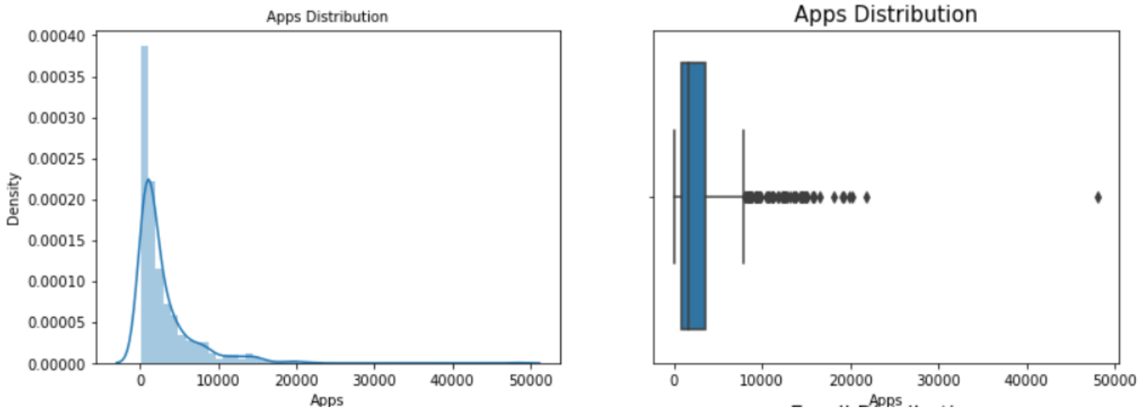
- The data set contain 777 rows and 18 columns
- The names columns are a categorical value
- The are no duplicates in the data set
- There are total of 18 variables in date set
- The data set does not contain any null values or any missing values



```
Names          0
Apps           0
Accept         0
Enroll         0
Top10perc      0
Top25perc      0
F.Undergrad    0
P.Undergrad    0
Outstate       0
Room.Board     0
Books          0
Personal       0
PhD            0
Terminal       0
S.F.Ratio      0
perc.alumni    0
Expend         0
Grad.Rate      0
```

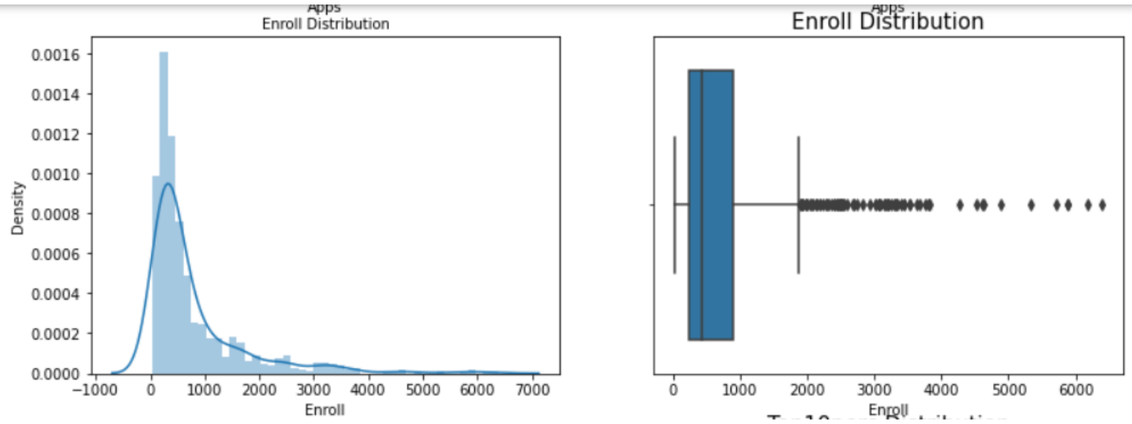
2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

UNIVARIENT ANALYSIS
APPS



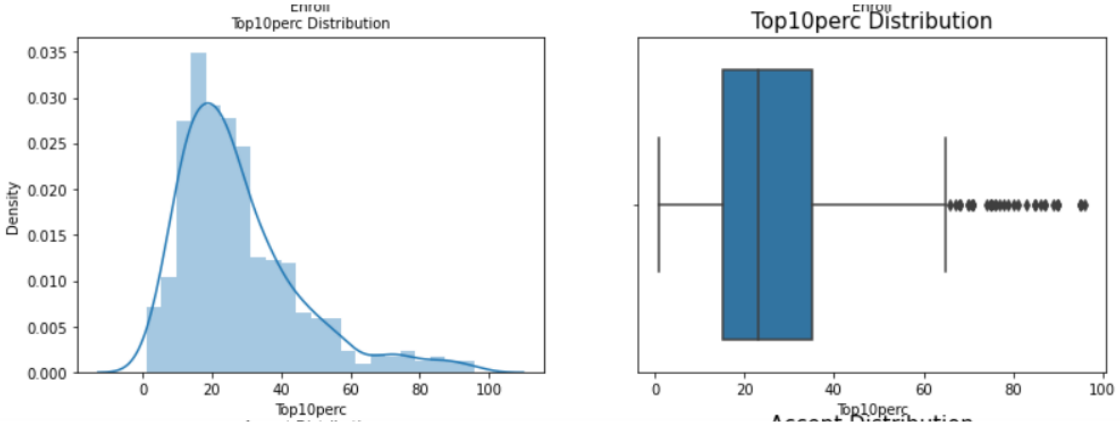
The box plot for apps seems to be an outlier. From the plot we could see that the university offers applications in range 1000 to 3000 the max applications seems to be around 50.000

ENROLL



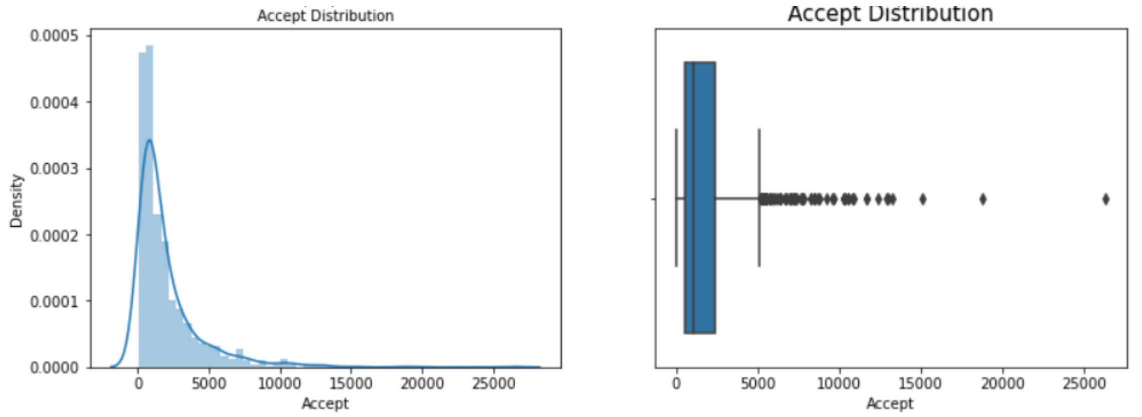
The box plot for apps seems to be an outlier for enrol. From this plot we can understand that most of the students have enrolled from the college in range 200 to 600

TOP 10



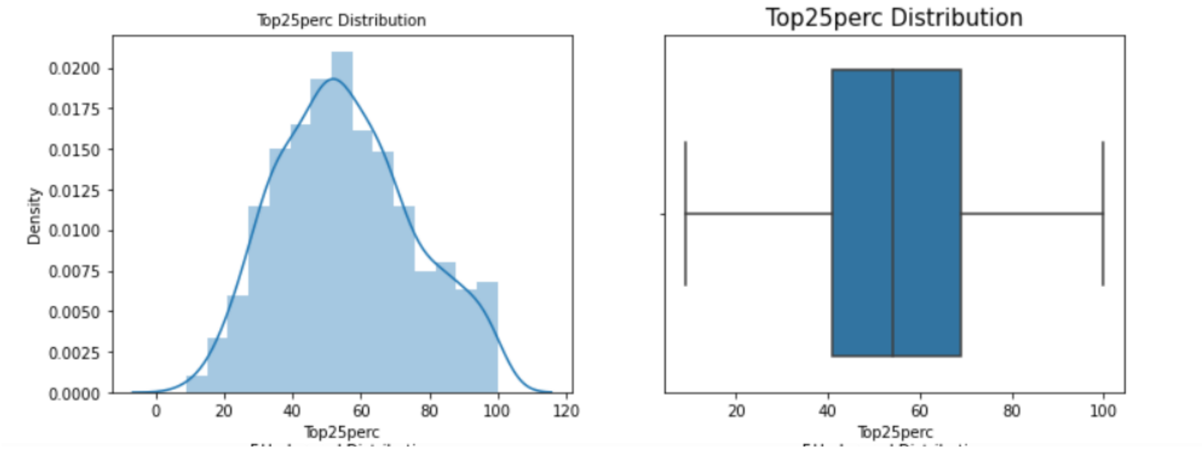
The box plot of the students from top 10 percentage of higher secondary class seems to have outliers. The distribution seems to be positively skewed. There is good amount of intake about 30 to 50 students from top 10 percentage of higher secondary class.

Accept



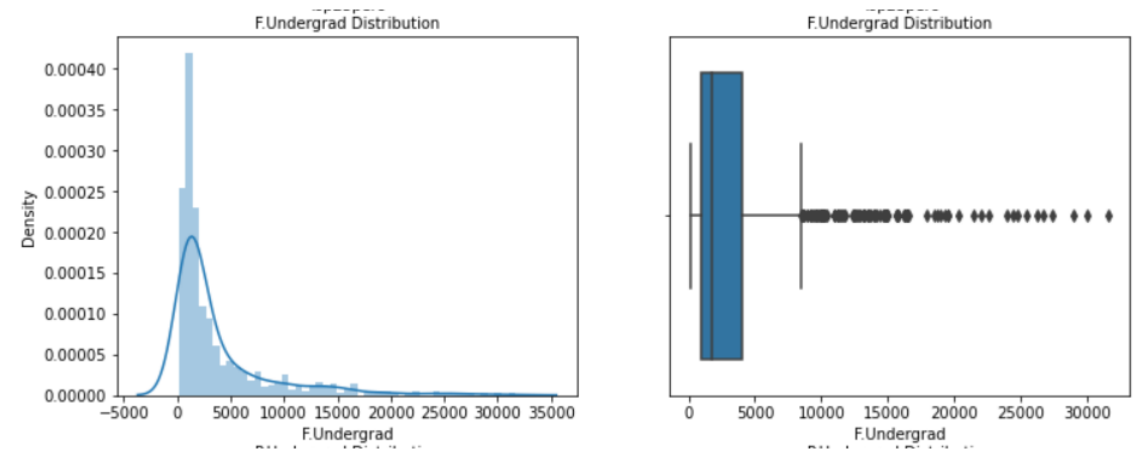
The accept variable seems to have outliers. The dist. plot shows us the majority of applications accepted from each university are in the range from 70 to 1500.The accept variable seems to be positively skewed

TOP 25 PERC



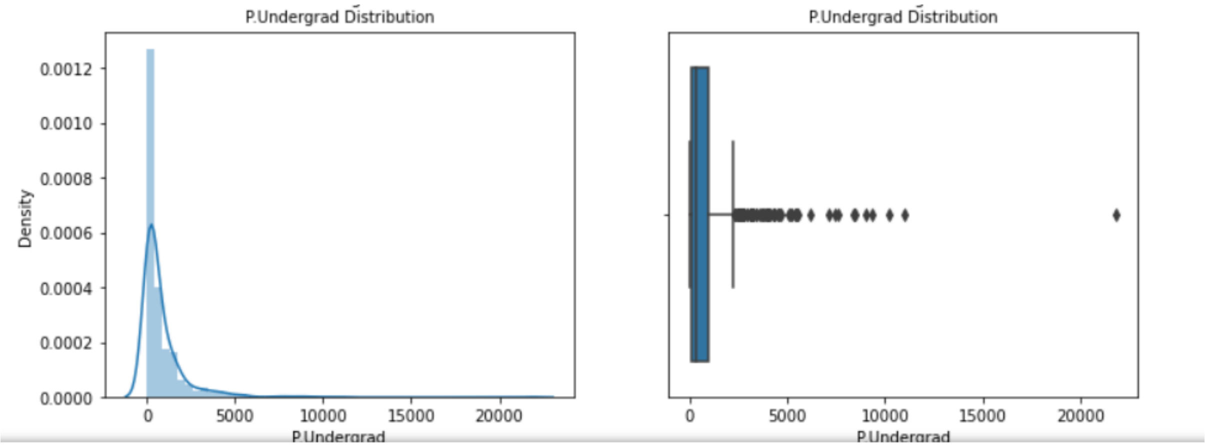
The distribution is almost normally distributed. Majority of the students are from top 25% of higher secondary class.

F.UNDERGRADE



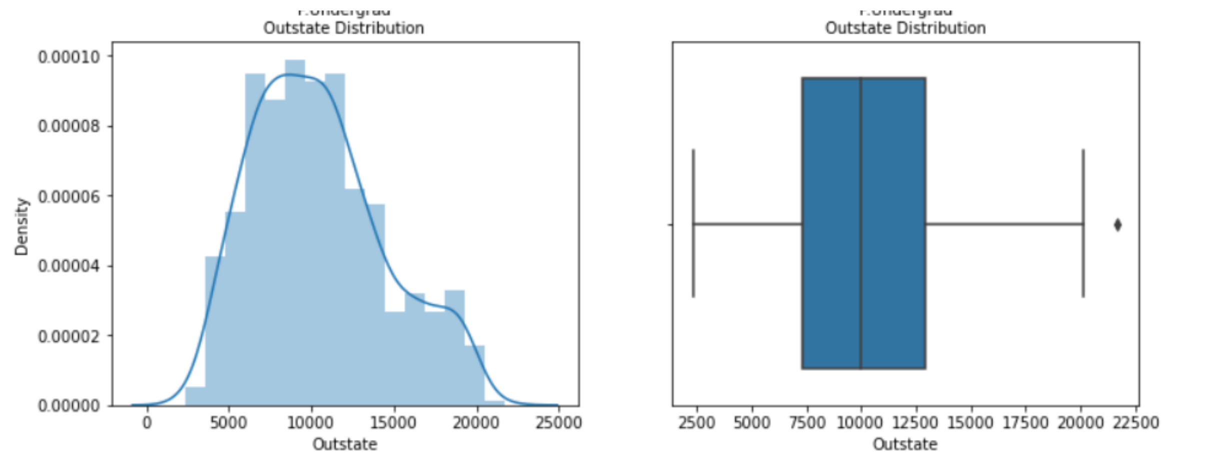
The distribution of the data is positively skewed. In the range about 3000 to 5000 they are full time graduates studying in all the university.

P.UNDERGRADE



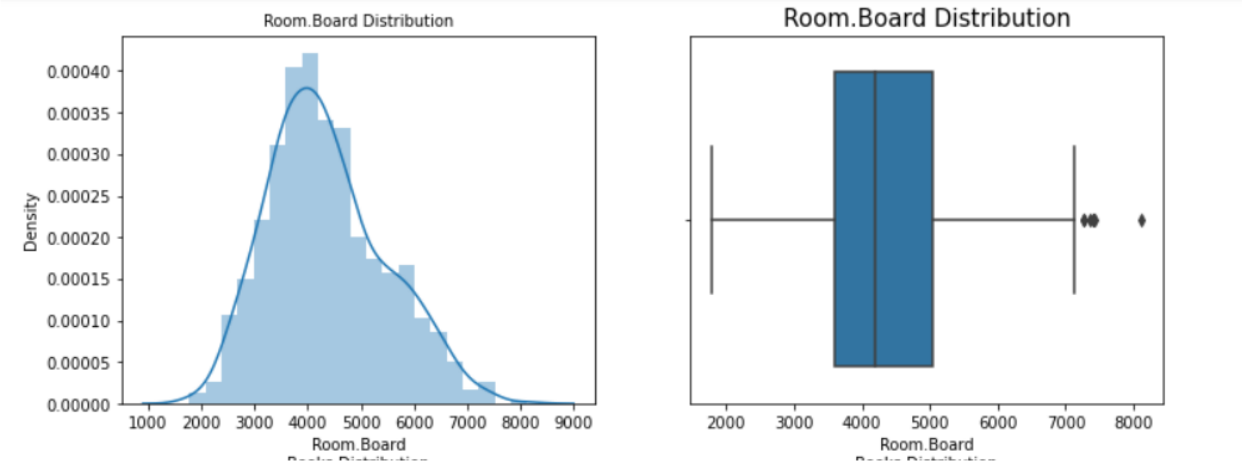
The distribution of the data is positively skewed. In the range about 1000 to 3000 they are part-time graduates studying in all the university

OUTSTATE



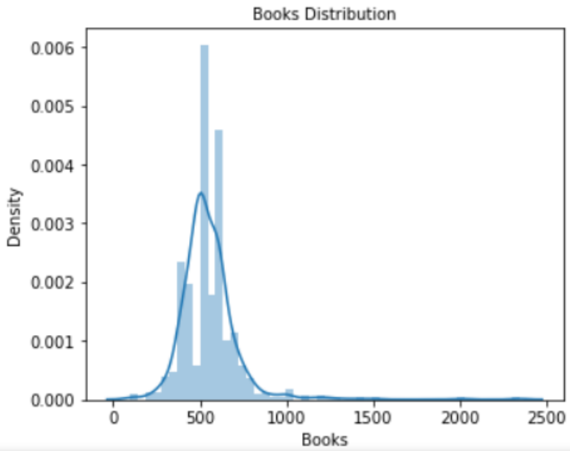
The box plot of outstate has only one outlier. The distribution is almost normally distributed.

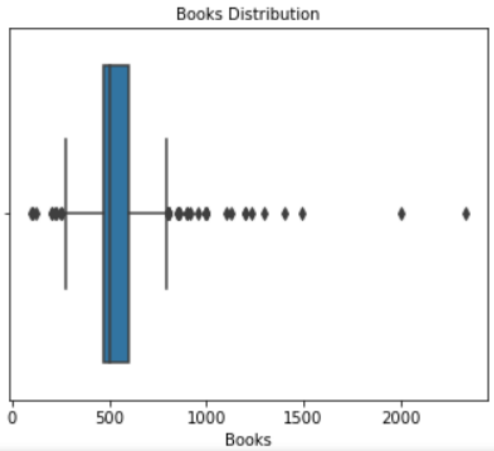
ROOM BOARD



The Room board has few outliers. The distribution is normally distributed.

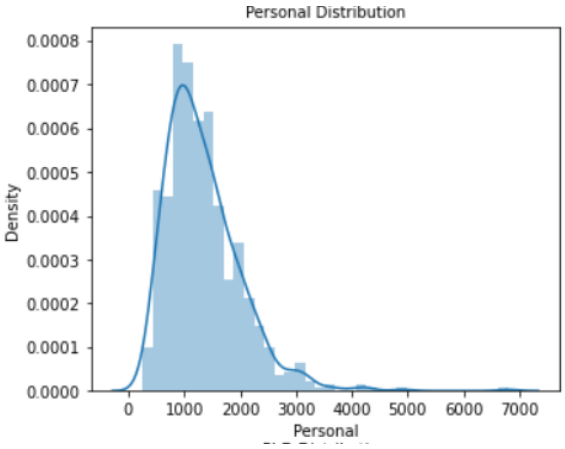
BOOKS

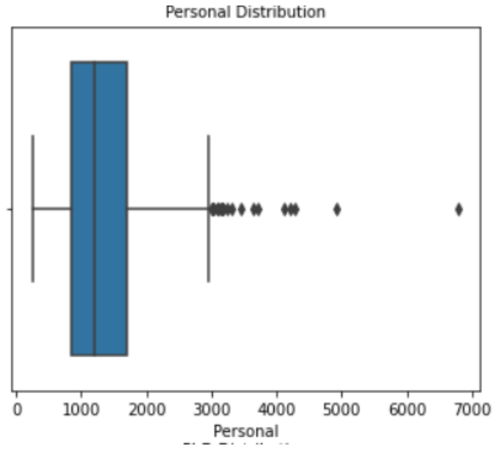




The Room board has few outliers. The distribution is normally distributed.

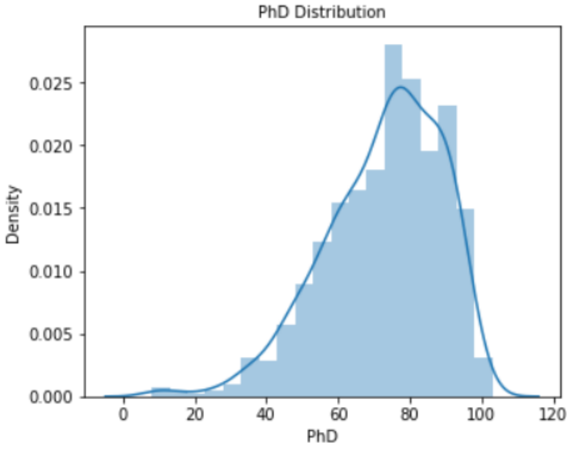
PEERSONAL

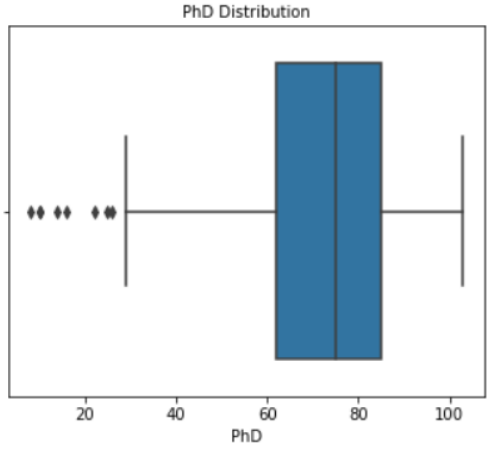




Some student’s personal expense are way bigger than the rest of the students. The distribution seems to be positively skewed.

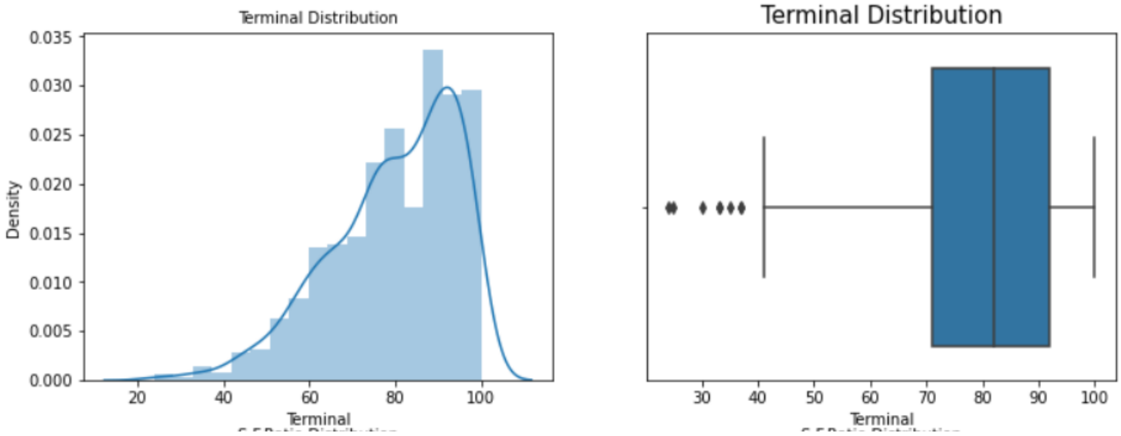
PHD





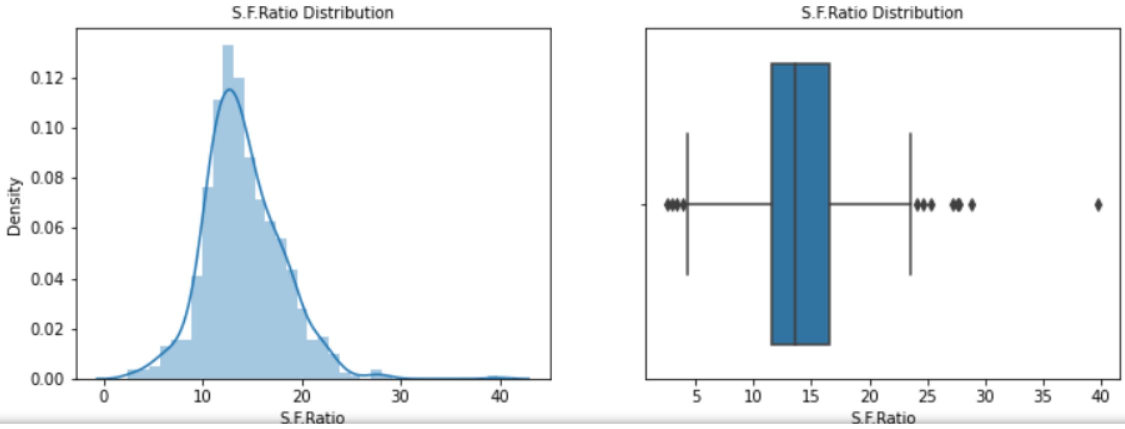
.The distribution seems to be negatively skewed

Terminal distribution



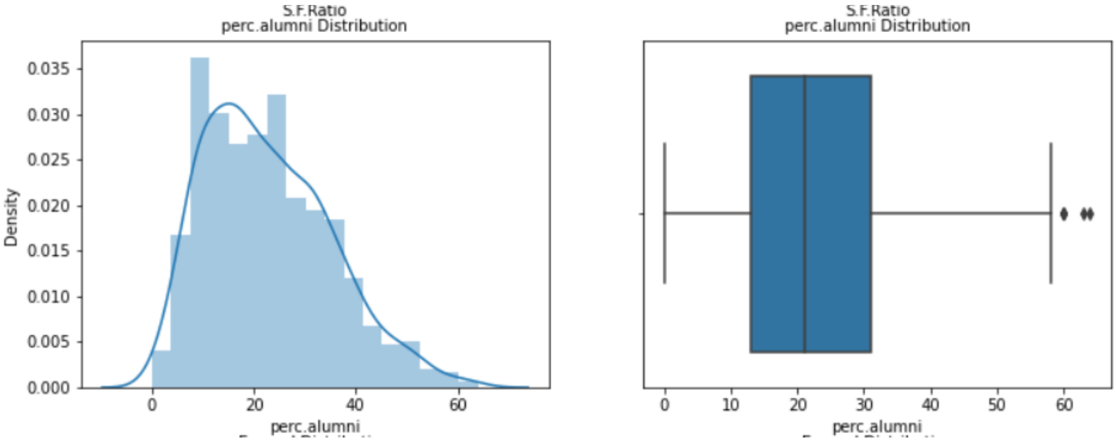
The box plot of terminal seems to have outliers in the dataset. The distribution for the terminal also seems to be negatively skewed.

SF ratio



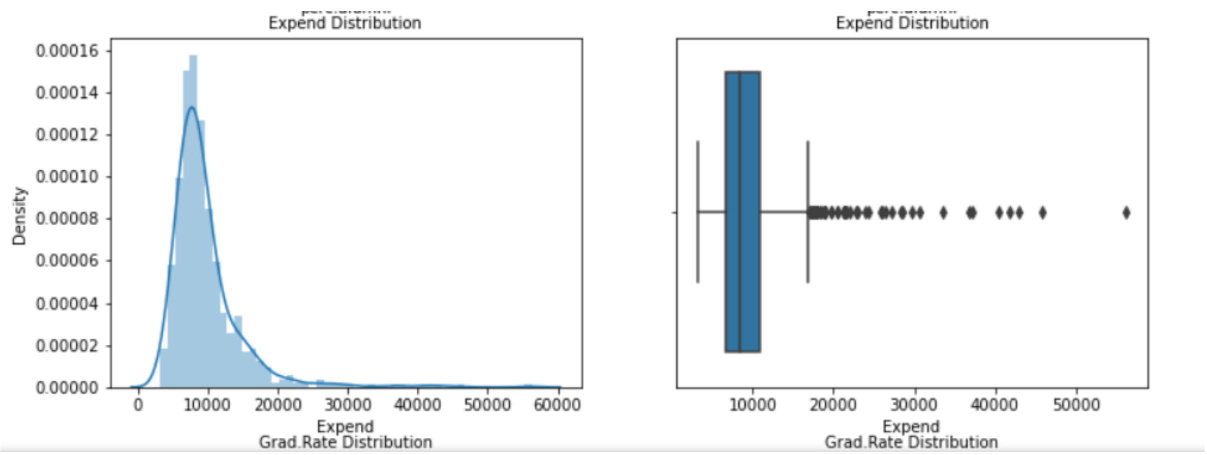
The distribution is almost normally distributed. The student faculty ratio is almost same in all the university and colleges.

Perc. Alumni distribution



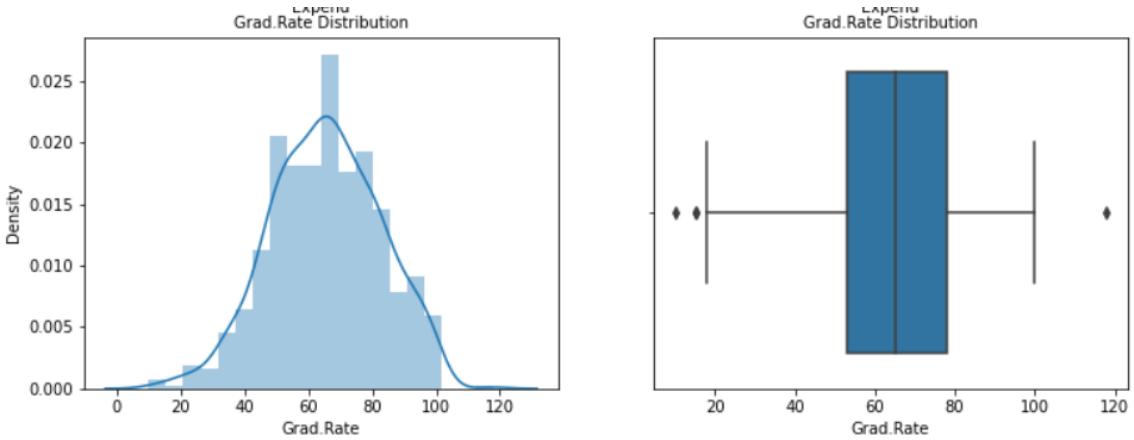
The percentage of alumni box plot seems to have outliers in the dataset. The distribution is almost normally distributed.

Expend distribution



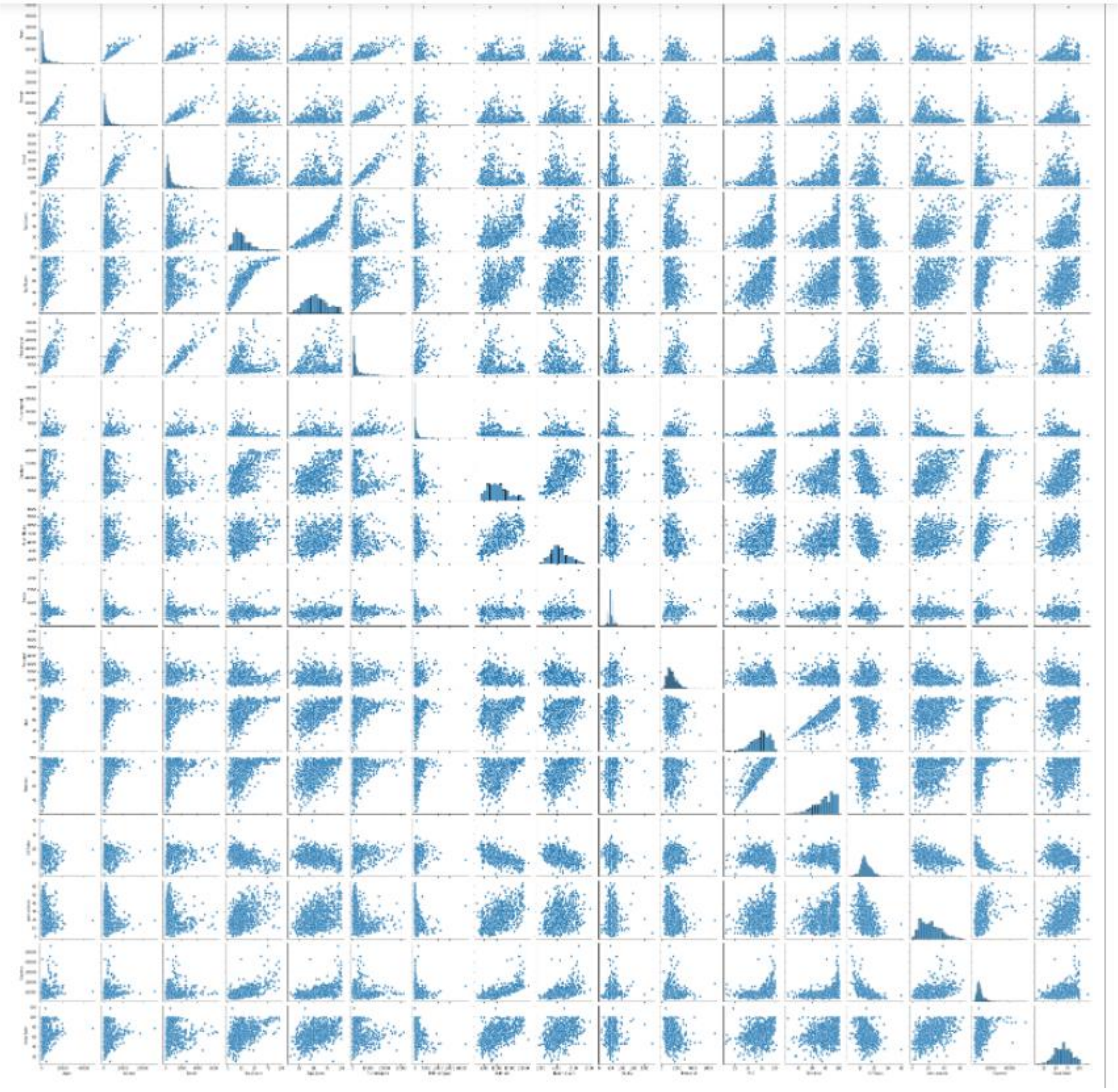
The expenditure variable also has outliers in the dataset. The distribution of the expenditure is positively skewed.

Grade. Rate Distribution



The graduation rate among the students in all the university above 60%. The box plot of the graduation rate has outliers in the dataset. The distribution is normally distributed.

MULTIVARIATE ANALYSIS



The pair plot helps us to understand the relationship between all the numerical values in the dataset. On comparing all the variables with each other we could understand the patterns or trends in the dataset.

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

The application, accepted application, enrolled full-time graduates, part-time graduates, outstate are numbers of students

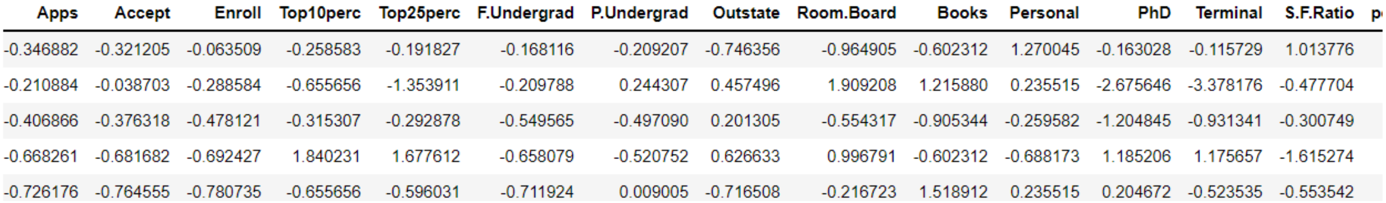
The top-10 percent and top 20percent are students in which the values are in percentage

Room board, boos and personal are values associated with money.

The PhD, sf ratio, percentage of alumni are percentage values of different combinations of student's teachers' alumina all these are parentage values

The graduation ratio is also a percentage value

The scaling or the z score tells how many standard deviations is the point away from mean and also the direction. Which is needed in order to perform PCA



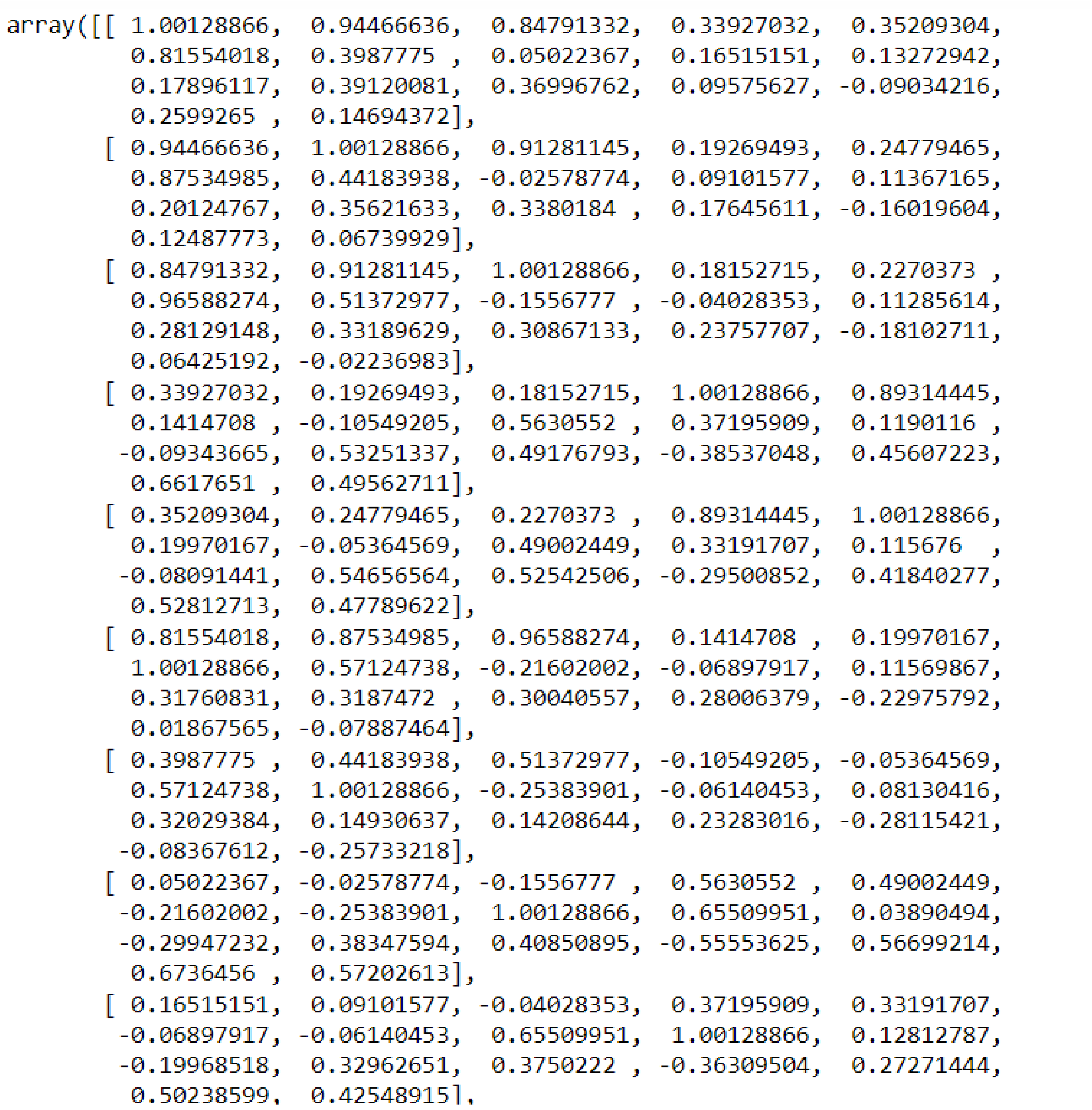
Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	p
-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.163028	-0.115729	1.013776	
-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.675646	-3.378176	-0.477704	
-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	-0.300749	
-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	-1.615274	
-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	-0.553542	

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

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.

The comparison between the covariance and correlation matrix is that both of the terms measure the relationship and the dependency between two variables.

Covariance matrix:



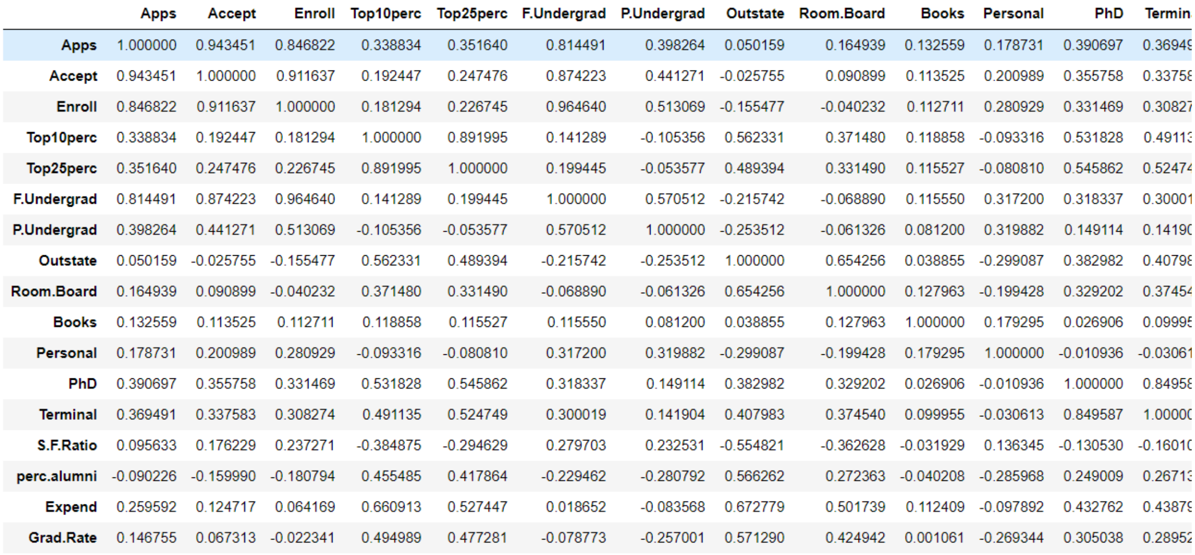
```
array([[ 1.00128866,  0.94466636,  0.84791332,  0.33927032,  0.35209304,
         0.81554018,  0.3987775 ,  0.05022367,  0.16515151,  0.13272942,
         0.17896117,  0.39120081,  0.36996762,  0.09575627, -0.09034216,
         0.2599265 ,  0.14694372],
       [ 0.94466636,  1.00128866,  0.91281145,  0.19269493,  0.24779465,
         0.87534985,  0.44183938, -0.02578774,  0.09101577,  0.11367165,
         0.20124767,  0.35621633,  0.3380184 ,  0.17645611, -0.16019604,
         0.12487773,  0.06739929],
       [ 0.84791332,  0.91281145,  1.00128866,  0.18152715,  0.2270373 ,
         0.96588274,  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,
         0.19970167, -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,
         1.00128866,  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,
         0.57124738,  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,
        -0.21602002, -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,
        -0.06897917, -0.06140453,  0.65509951,  1.00128866,  0.12812787,
        -0.19968518,  0.32962651,  0.3750222 , -0.36309504,  0.27271444,
         0.50238599,  0.42548915],
```

Correlation matrix:

The correlation matrix before scaling and after scaling will remain the same

From this snippet we can understand variables which are highly positively correlated and the variables which are highly negatively correlated. We can also understand the variables which are moderately correlated with each other

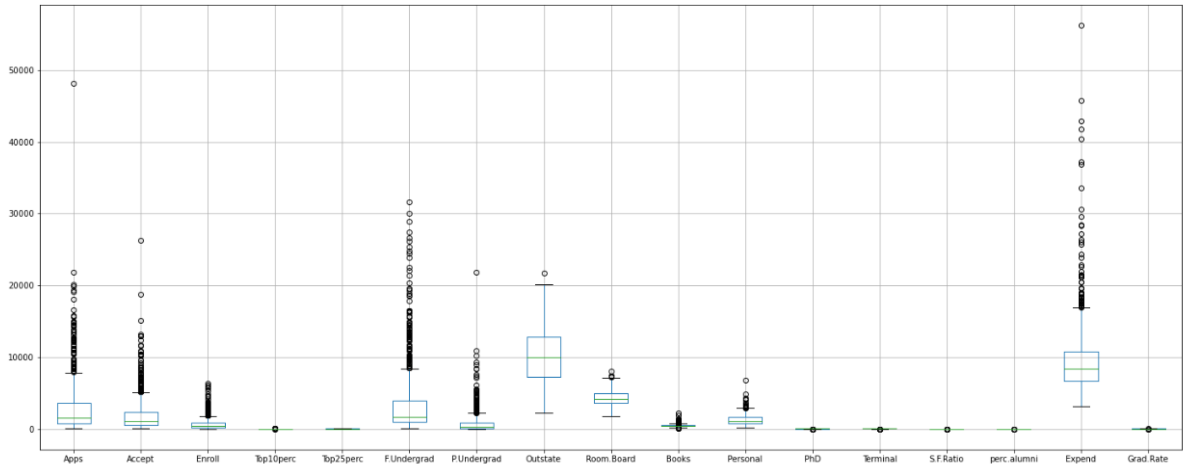
Also, the top 10 percentage and top 25 percentage are highly positively correlated



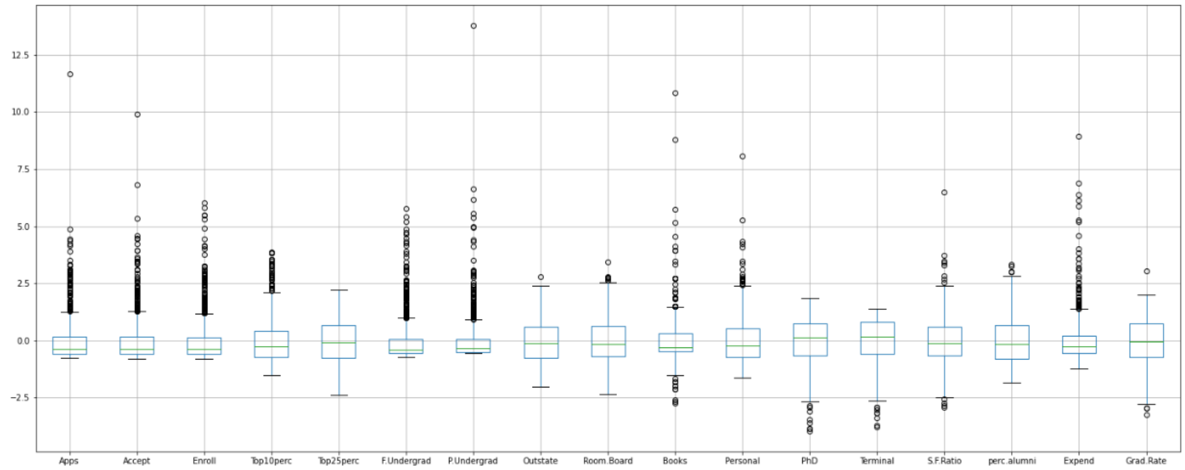
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termin
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.3694
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758	0.3375
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469	0.3082
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828	0.4911
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862	0.5247
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.318337	0.3000
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.081200	0.319882	0.149114	0.1419
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.038855	-0.299087	0.382982	0.4079
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.127963	-0.199428	0.329202	0.3745
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.000000	0.179295	0.026906	0.0999
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.000000	-0.010936	-0.0306
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1.000000	0.8495
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.099955	-0.030613	0.849587	1.0000
S.F.Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628	-0.031929	0.136345	-0.130530	-0.1601
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009	0.2671
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739	0.112409	-0.097892	0.432762	0.4387
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.571290	0.424942	0.001061	-0.269344	0.305038	0.2895

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

Data before scaling:

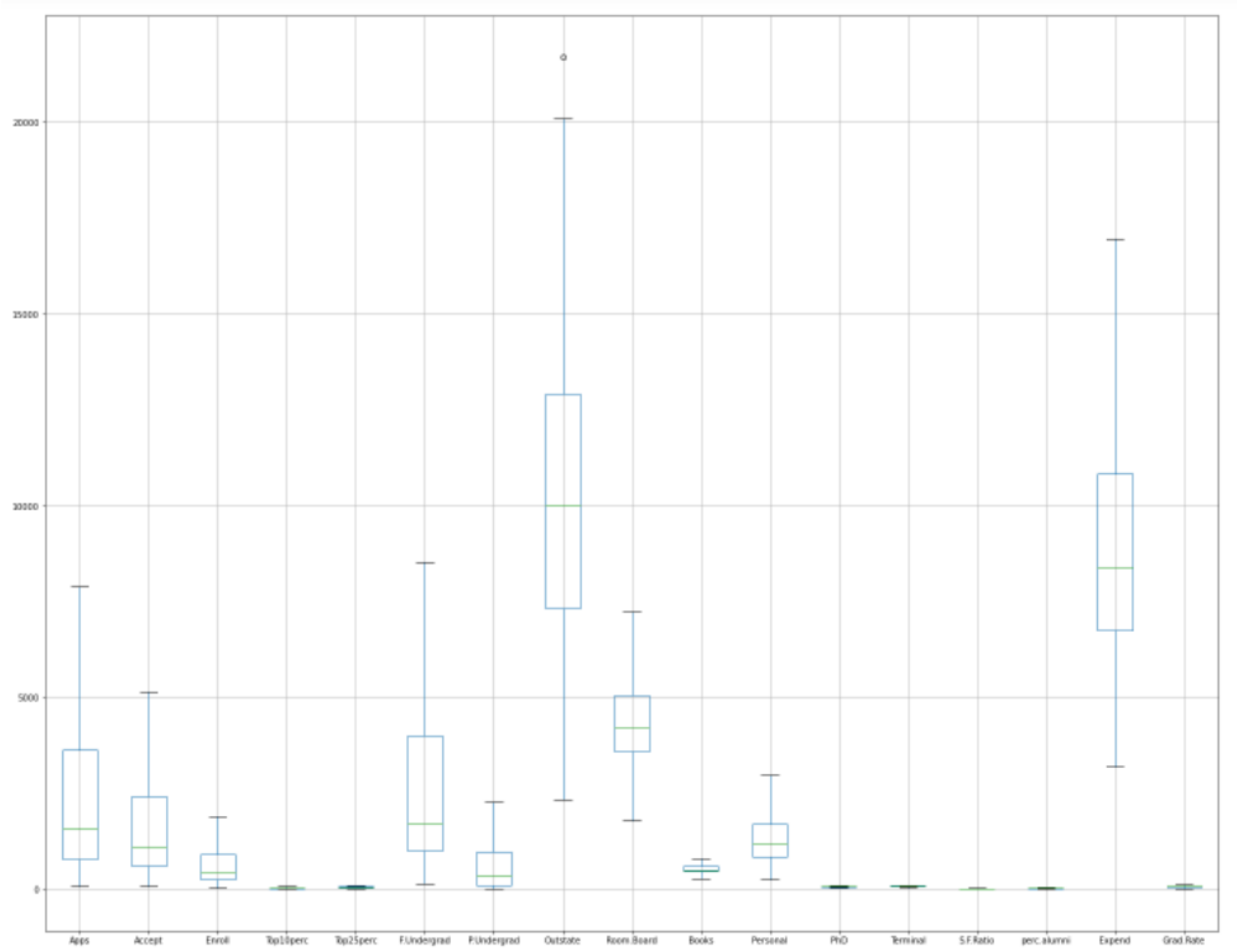


Data after scaling:



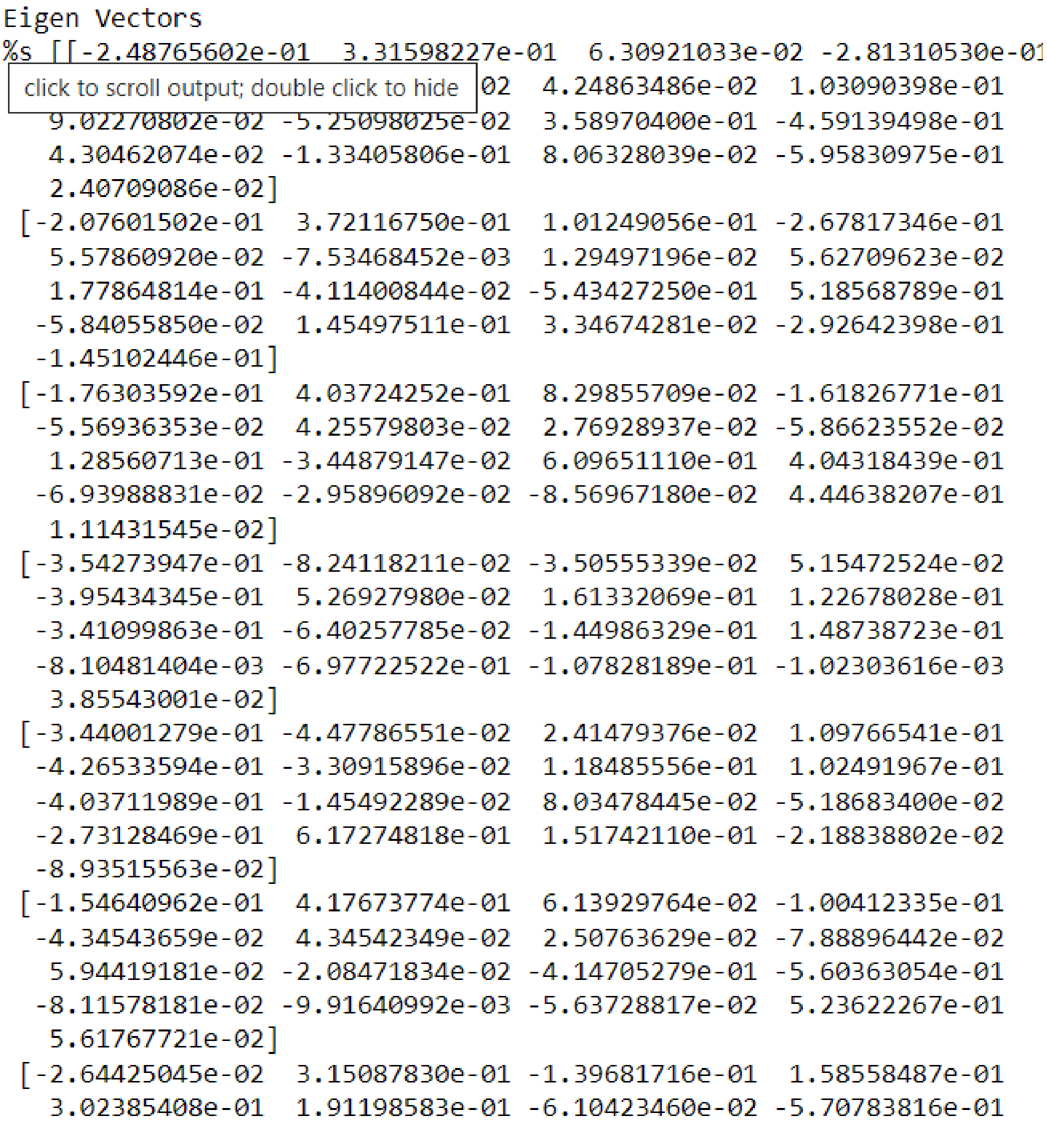
The outliers are still present in the dataset. Scaling does not remove the outliers scaling scales value on a z score distribution, we can use any one method to remove outliers for further processes

after treating outliers



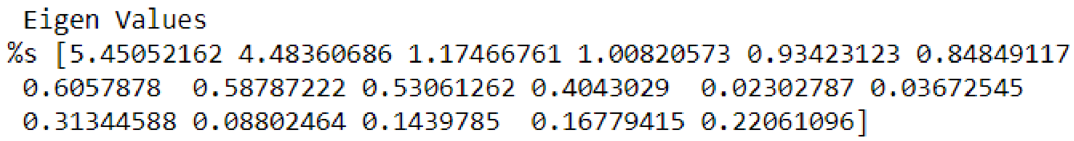
2.5 Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both]

Eigen values:



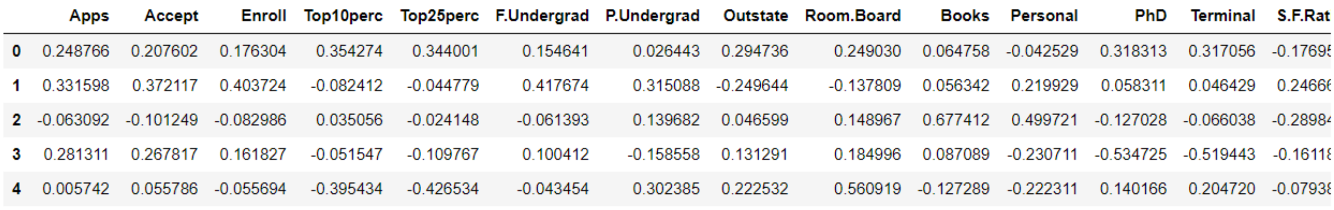
```
Eigen Vectors
%s [[-2.48765602e-01  3.31598227e-01  6.30921033e-02 -2.81310530e-01
  click to scroll output; double click to hide 02  4.24863486e-02  1.03090398e-01
   9.02270802e-02 -5.25098025e-02  3.58970400e-01 -4.59139498e-01
   4.30462074e-02 -1.33405806e-01  8.06328039e-02 -5.95830975e-01
   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
```

Eigen vectors:



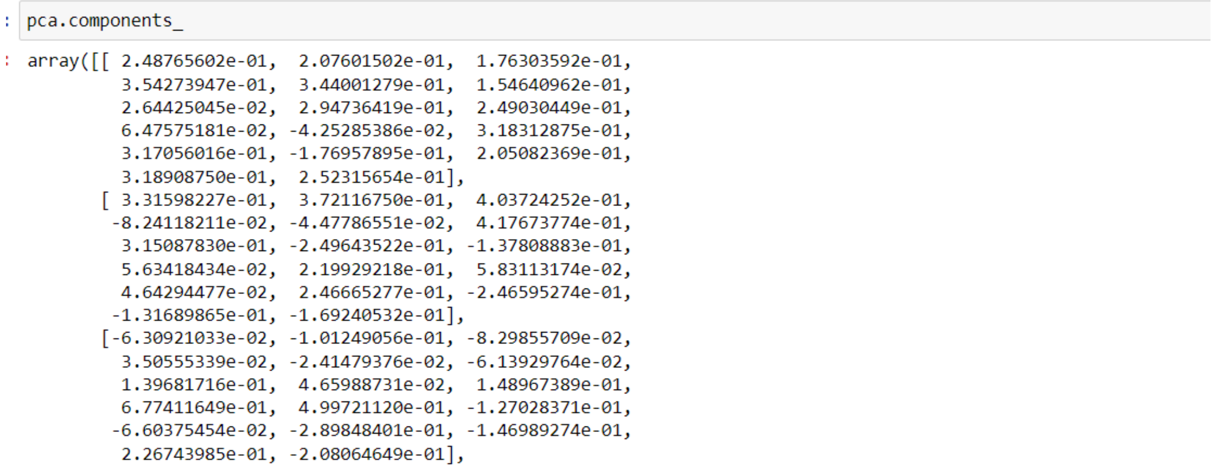
```
 Eigen Values
%s [5.45052162 4.48360686 1.17466761 1.00820573 0.93423123 0.84849117
 0.6057878  0.58787222 0.53061262 0.4043029  0.02302787 0.03672545
 0.31344588 0.08802464 0.1439785  0.16779415 0.22061096]
```

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



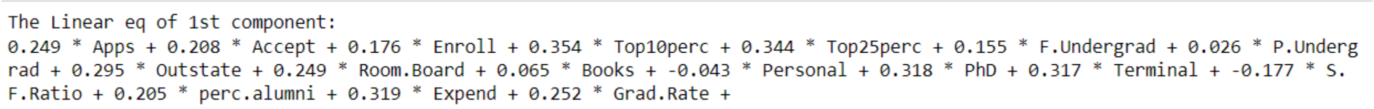
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Rat
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0.064758	-0.042529	0.318313	0.317056	-0.1769
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809	0.056342	0.219929	0.058311	0.046429	0.2466
2	-0.063092	-0.101249	-0.082986	0.035056	-0.024148	-0.061393	0.139682	0.046599	0.148967	0.677412	0.499721	-0.127028	-0.066038	-0.2898
3	0.281311	0.267817	0.161827	-0.051547	-0.109767	0.100412	-0.158558	0.131291	0.184996	0.087089	-0.230711	-0.534725	-0.519443	-0.1611
4	0.005742	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919	-0.127289	-0.222311	0.140166	0.204720	-0.0793

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]



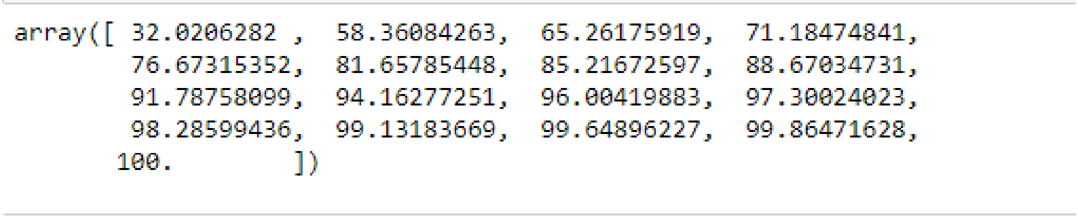
```
: pca.components_

: 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],
```

linear equation of PC in terms of eigenvectors and corresponding features



```
The Linear eq of 1st component:
0.249 * Apps + 0.208 * Accept + 0.176 * Enroll + 0.354 * Top10perc + 0.344 * Top25perc + 0.155 * F.Undergrad + 0.026 * P.Underg
rad + 0.295 * Outstate + 0.249 * Room.Board + 0.065 * Books + -0.043 * Personal + 0.318 * PhD + 0.317 * Terminal + -0.177 * S.
F.Ratio + 0.205 * perc.alumni + 0.319 * Expend + 0.252 * 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?



```
array([ 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.        ])
```

The first components explain 32.02% variance in data

The first two components explain 58.36% variance in data

The first three components explain 65.26% variance in data

The first four components explain 71.18% variance in data

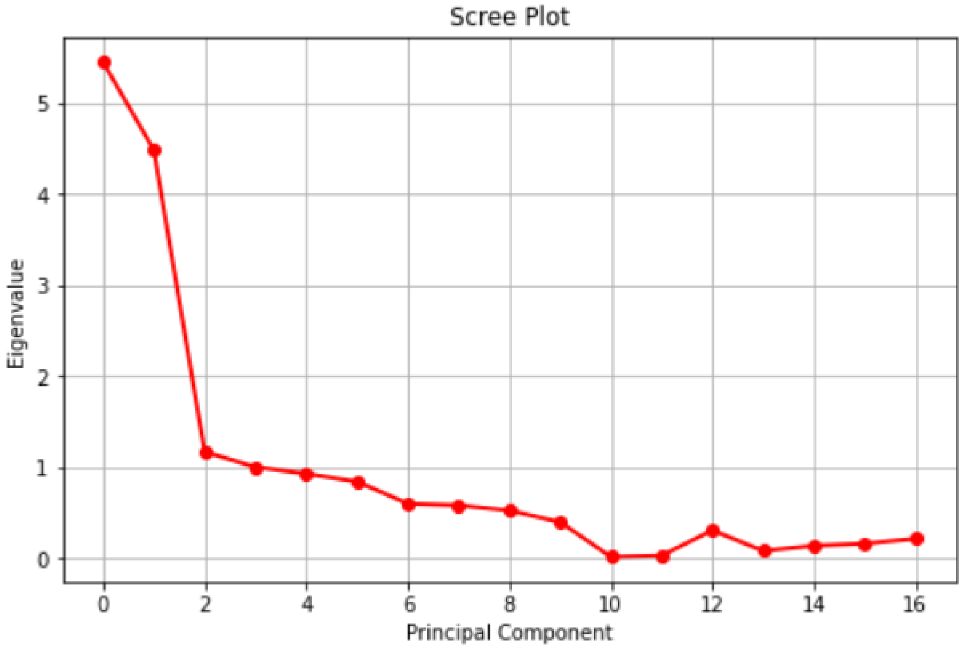
The first five components explain 76.67% variance in data

To decide the optimum number of principal components

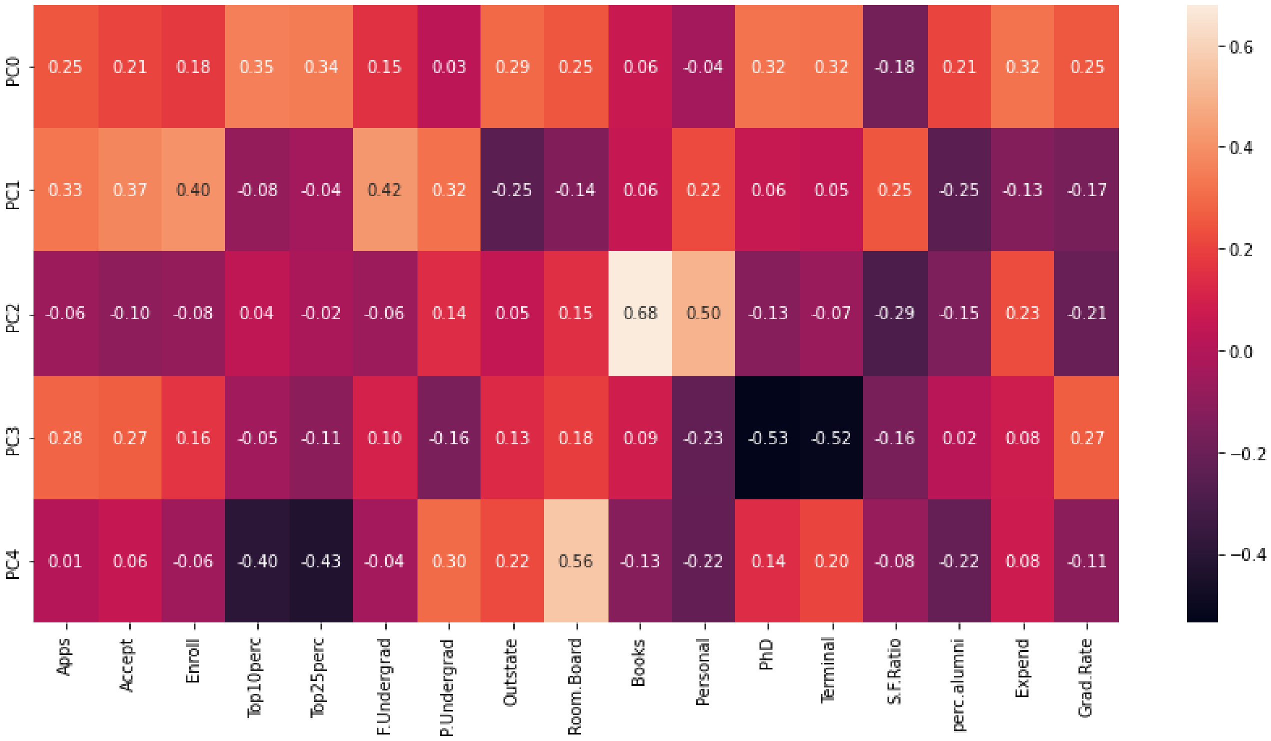
1. Check for cumulative variance up to 90%, check the corresponding associated with 90%

2. The incremental value between the components should not be less than five percent. 18 So basis on this we can decide the optimum number of principal components as 6, because after this the incremental value between the is less than 5%

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]



This business case study is about education dataset which contain the names of various colleges, which has various details of colleges and university. To understand more about the dataset, we perform univariate analysis and multivariate analysis which gives us the understanding about the variables. From analysis we can understand the distribution of the dataset, skew, and patterns in the dataset. From multivariate analysis we can understand the correlation of variables. Inference of multivariate analysis shows we can understand multiple variables highly correlated with each other. The scaling helps the dataset to standardize the variable in one scale. Outliers are imputed using IQR values once the values are imputed, we can perform PCA. The principal component analysis is used reduce the multicollinearity between the variables. Depending on the variance of the dataset we can reduce the PCA components. The PCA components for this business case is 5 where we could understand the maximum variance of the dataset. Using the components, we can now understand the reduced multicollinearity in the dataset. with this analysis we can perform further analysis and model building PCA will improve the efficiency of machine learning models.


Apps
Accept
Enroll
Top10perc
Top25perc
F.Undergrad
P.Undergrad
Outstate
Room.Board
Books
Personal
PhD
Terminal
S.F.Ratio
perc.alumni
Expend
Grad.Rate
PC0 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 0.25
PC1 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 -0.17
PC2 -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 -0.21
PC3 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 0.27
PC4 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 -0.11
0.6
0.4
0.2
0.0
−0.2
−0.4