BUSINESS REPORT ON ADVANCED STATISTICS

By Kshitij Nishant

Problem 1A

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csv] 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.

[Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.]

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

2. Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

3. Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

4. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. (Non-Graded)

1A.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

Answer:
1)Hypothesis for Education's One-Way ANOVA are:

Null Hypothesis H(0):
The mean salary is same at three level of education,i.e, High school graduate, Bachelor, and Doctorate

Alternate Hypothesis H(alt):
At least on one level of education, the mean salary is different

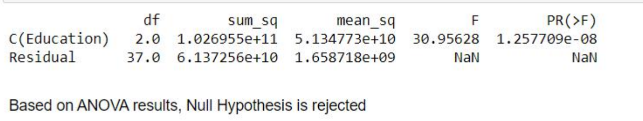
2)Hypothesis for Ocuupation's One way ANOVA are:

Null Hypothesis H(0):
The mean salary is same at four levels of Occupation,i.e,Administrative and clerical, Sales, Professional or specialty, and Executive or managerial

Alternate Hypothesis H(alt):
At least one level Occupation, the mean salary is different

1A.2 Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

Answer:



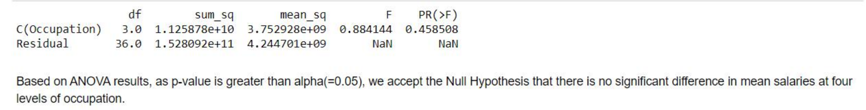
```
               df        sum_sq       mean_sq         F        PR(>F)
C(Education)   2.0  1.026955e+11  5.134773e+10  30.95628  1.257709e-08
Residual      37.0  6.137256e+10  1.658718e+09       NaN           NaN
```

Based on ANOVA results, Null Hypothesis is rejected

Since, p-value<alpha, we reject the Null Hypothesis.

1A.3 Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

Answer:



```
                 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
```

Based on ANOVA results, as p-value is greater than alpha(=0.05), we accept the Null Hypothesis that there is no significant difference in mean salaries at four levels of occupation.

Since, p-value is greater than alpha(=0.05), we accept the Null Hypothesis.

1A.4If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result.

Answer:

The null hypothesis for 1A.2 is rejected.

As ANOVA can tell us only about the significant difference in mean for at least one pair, we cannot figure out which class mean is significantly different. To find the difference in class means other methods are used.

Problem 1B

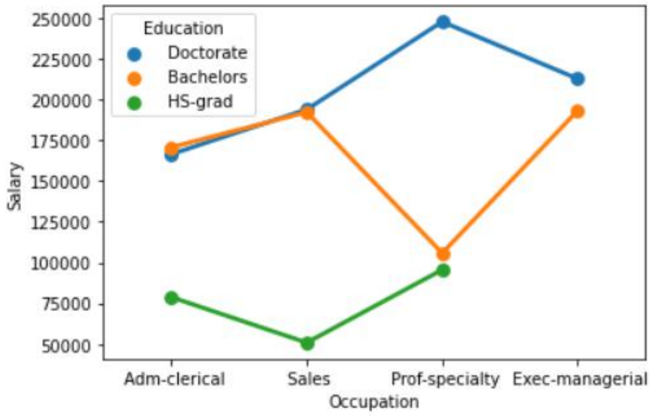
1. What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the ‘pointplot’ function from the ‘seaborn’ function]

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

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

1B.1 If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result.

Answer:



From above plot we can make out that the interaction between people with:

1.Adm-Clerical job with Bachelors and Doctorates is almost good.
2.Sales job with Bachelors and Doctorates is good.
3.Prof-Speciality job with HS-grad and Bachelors is a bit.
4.All four occupations with educational level HS-grad and Doctorate is absolutely NIL.
5.Exec-Managerial job role has no interactions with any other educational background.

From above plot we can figure out that people with educational level:

1.Doctorates : are into higher salary brackets and mostly Prof-speciality roles or Exec-managerial roles or in sales profile, very few are doing Adm-clerical jobs

2.Bachlores: fall in mid income range and found mostly working as an Exec -managers , Adm-clerks or into sales but very few are found in Prof- speciality profile.

3.HS-grads : are in low income brackets, mostly doing Prof-speciality or Adm -clerical work and few are doing Sales but hardly any in Exec-managerial role.

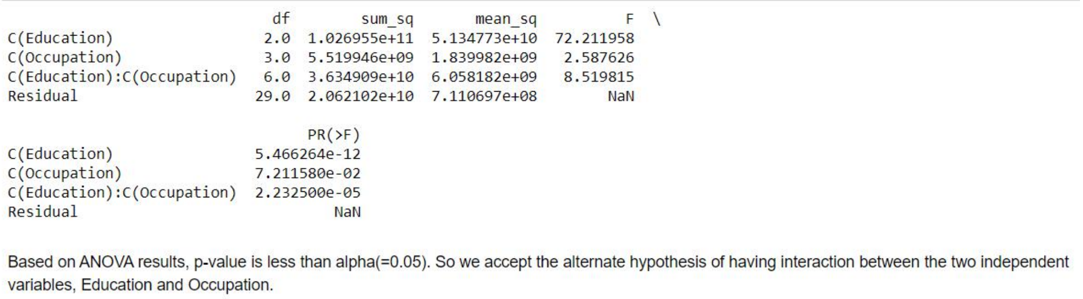
1B.2 Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?

Answer:

The Null and Alternate Hypothesis for Two-way ANOVA for Education and Occupation are:

Null Hypothesis H(0): Mean salary for each Education and Occupation type are equal

Alternate Hypothesis H(alt): For at least one of means of salary is not equal for each Education and Occupation type



```
                               df        sum_sq       mean_sq          F  \
C(Education)                  2.0  1.026955e+11  5.134773e+10  72.211958
C(Occupation)                 3.0  5.519946e+09  1.839982e+09   2.587626
C(Education):C(Occupation)    6.0  3.634909e+10  6.058182e+09   8.519815
Residual                     29.0  2.062102e+10  7.110697e+08        NaN

                                   PR(>F)
C(Education)                 5.466264e-12
C(Occupation)                7.211580e-02
C(Education):C(Occupation)   2.232500e-05
Residual                              NaN
```

Based on ANOVA results, p-value is less than alpha(=0.05). So we accept the alternate hypothesis of having interaction between the two independent variables, Education and Occupation.

1B.3 Explain the business implications of performing ANOVA for this particular case study.

Answer:

1.ANOVA is used in a business context to help manage income /salary by comparing your education to occupation here in this case to help manage the Salary.

2.ANOVA can also be used to forecast Salary trends by analyzing patterns in data to better understand the future hike in Salary.

3.It's also a widely used statistical technique for comparing the relationship between factors that cause a rise in Salary, assuming this report is for HR department or HR consulting firm. Some of the key takeaways as below:

i. As the Education level upgrades Salary increases. On an average Doctorate earns higher salary than Bachelors and HS-Grads. However, it might be possibility that being Doctorate may not necessarily mean significant high salary than HS-Grad or Bachelors employees. So that means Doctorates are suitable for all job role or not always preferred above other education levels,maybe they can be considered some times as over qualified for certain job roles

ii. Though there is lesser significance of Occupation than education on Salary but at certain levels it impacts Salary.

iii. We must also take note of that high salaries are offered to Bachelor's degree holders than Doctorates for few occupations. So, we can say that there are some shortcomings of data set provided which reduces accuracy of the test and analysis done, as there can be few more other important variables which can impact salary such as years of experience, specialization,industry/domain etc.

iv. HR department plays more comprehensive role while setting up salary bands. As similar job titles with different industries demands varying salary package as per job profile, plus years of experience for the job matters here deciding scale of a person.

v. ANOVA test indicates that the Education level coupled with Occupation has significant influence over salary than alone occupation type with comparison to Educational background.

Problem 2

The dataset Education - Post 12th Standard.csv 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.xlsx.

1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?
2. Is scaling necessary for PCA in this case? Give justification and perform scaling.
3. Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].
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]
5. Extract the eigenvalues and eigenvectors.[Using Sklearn PCA Print Both]
6. Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features
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]
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?
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]

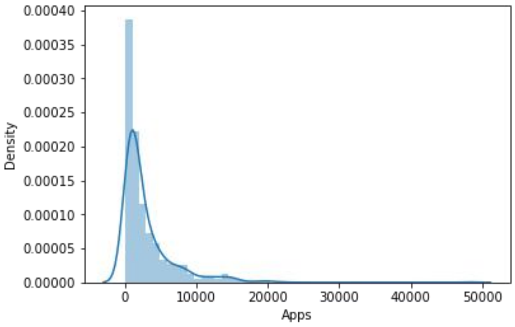
2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

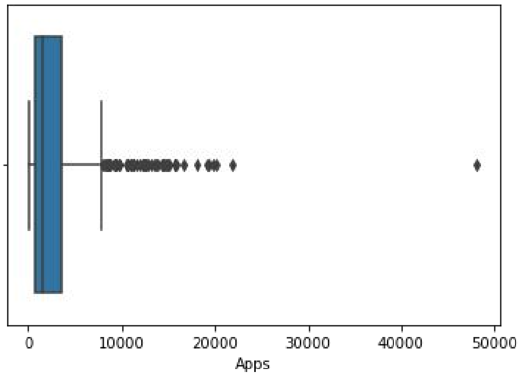
Answer:

A. For Univariate:

Main purpose of univariate data analysis is to summarize and find patterns in data.

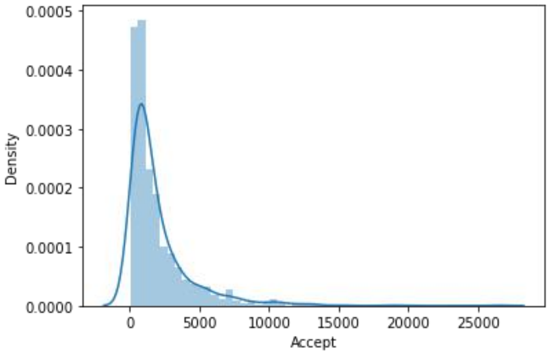
APPS

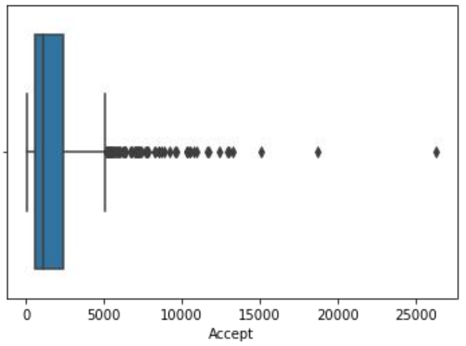




We have outliers in data set.

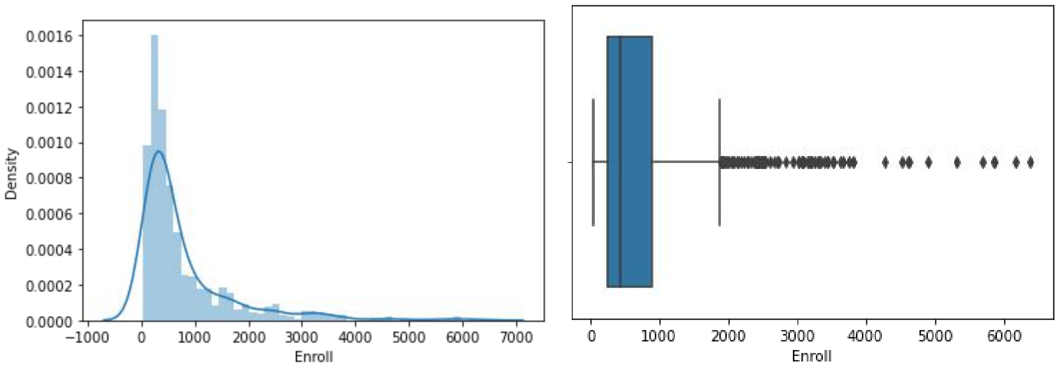
Accept





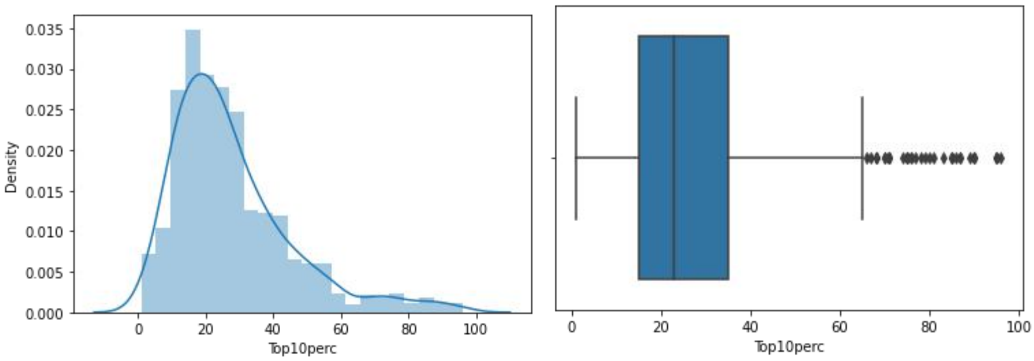
Distribution is positively skewed.

Enroll

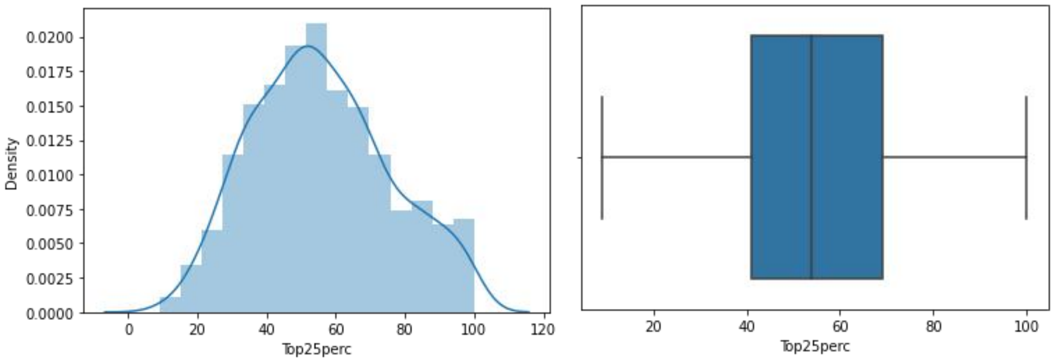


Distribution of data is positively skewed.

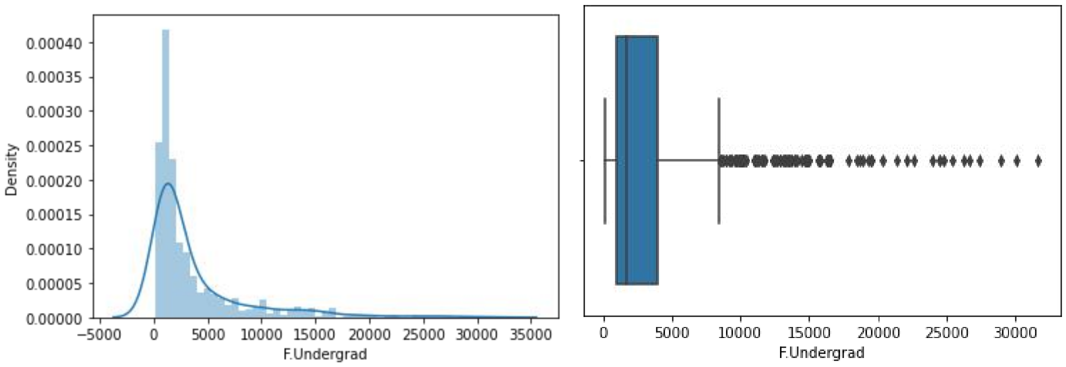
Top 10%



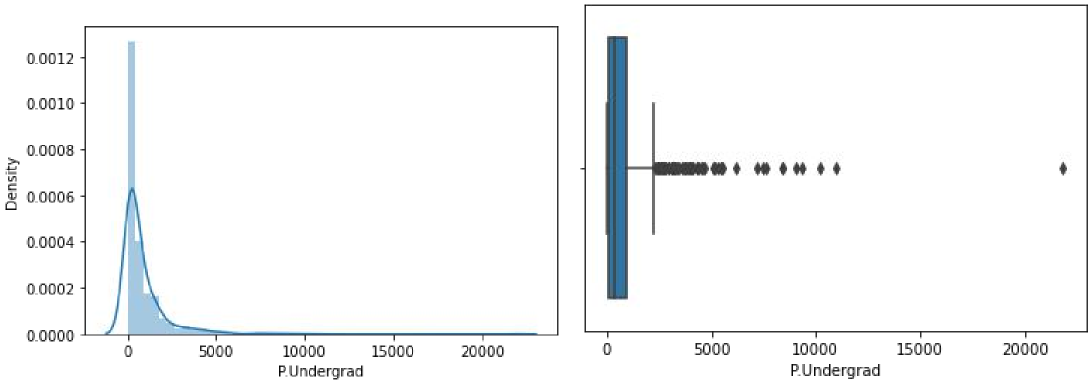
Top 25%



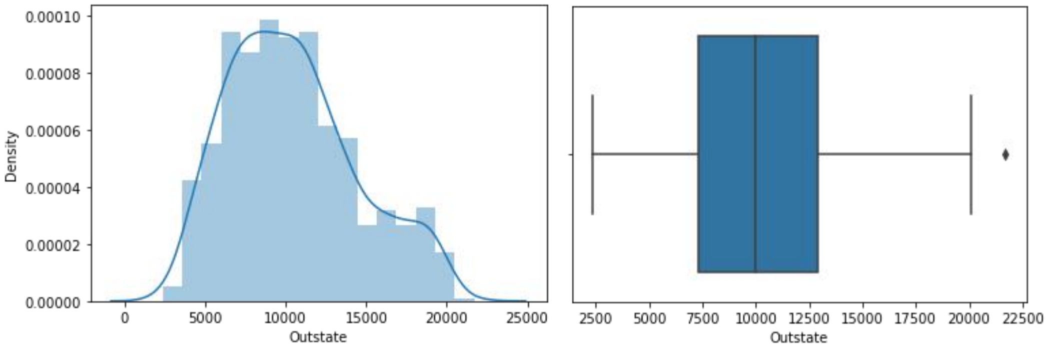
Full Time UnderGraduate



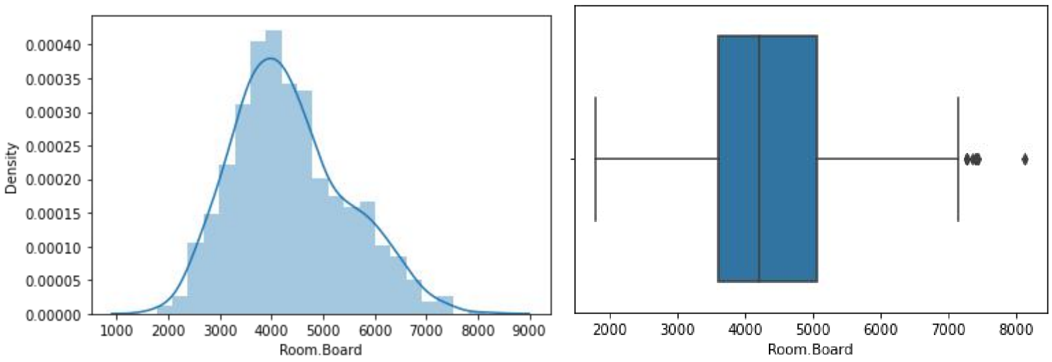
Part Time UnderGraduate



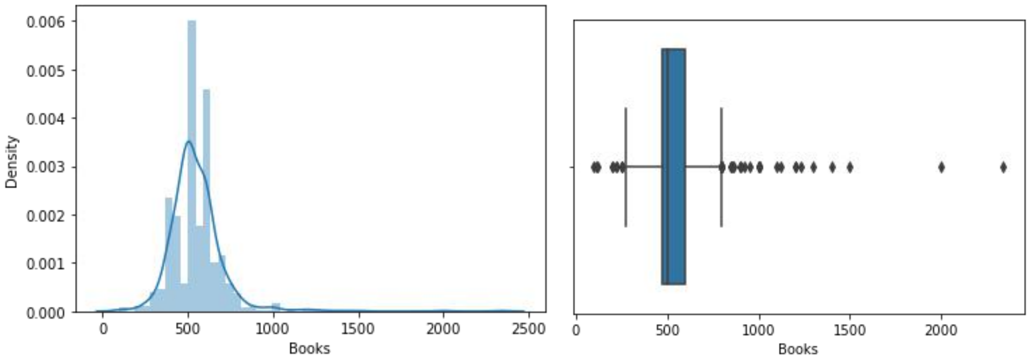
Outstate



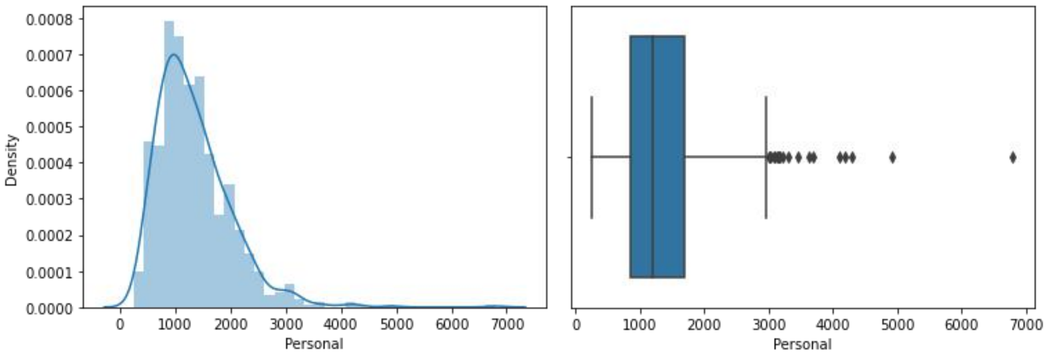
Room Board



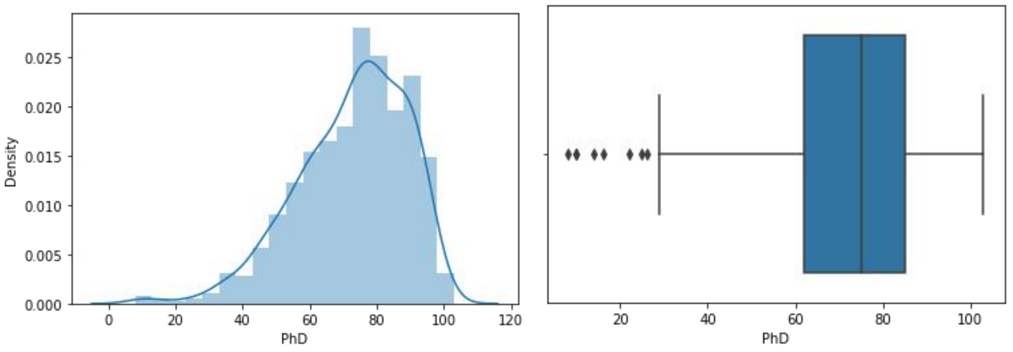
Books



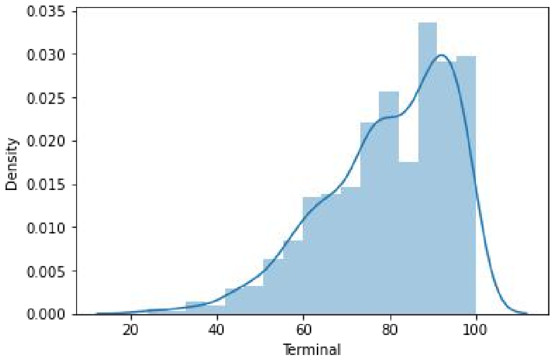
Personal

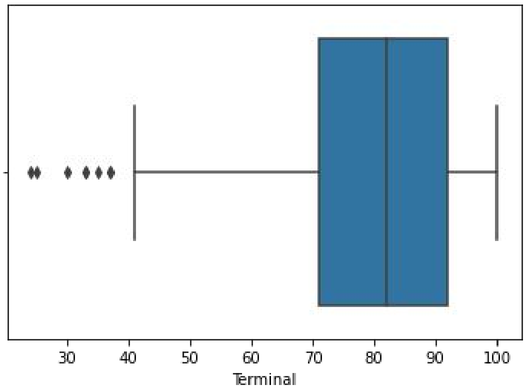


PhD

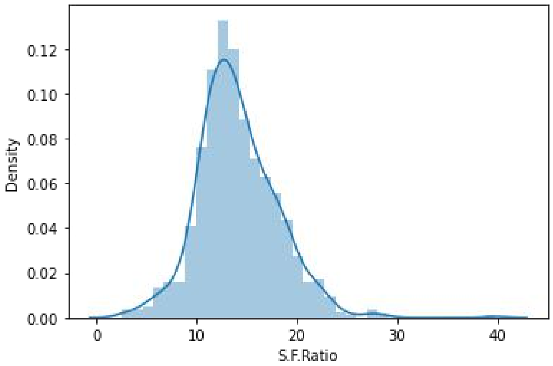


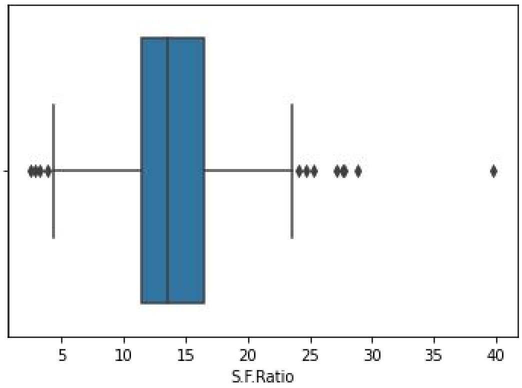
Terminal



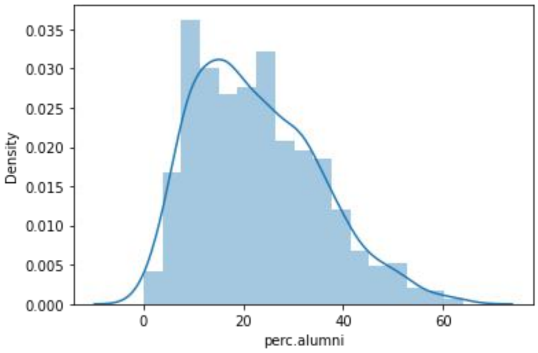


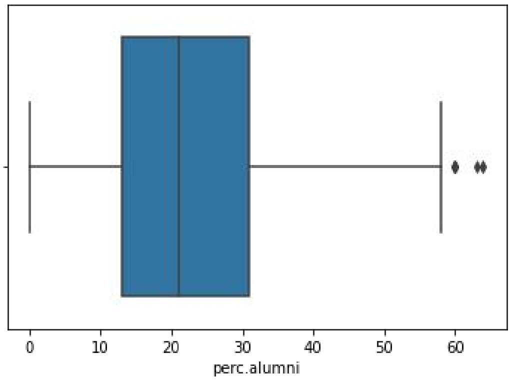
SF Ratio



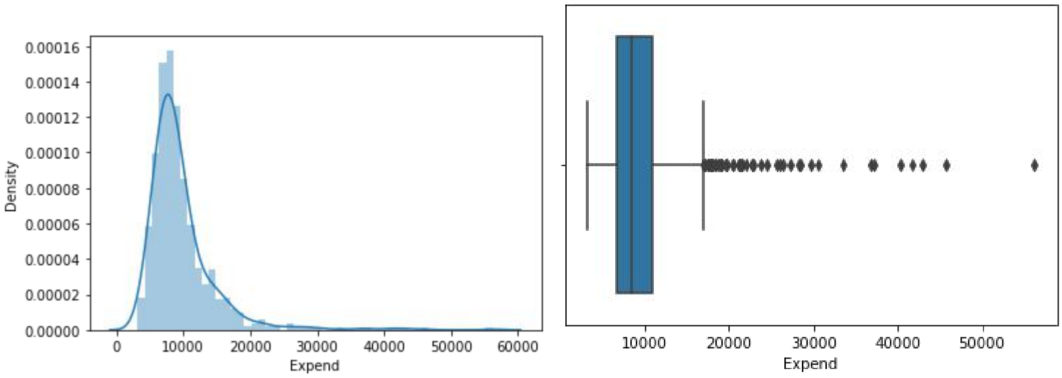


Perci Alumini

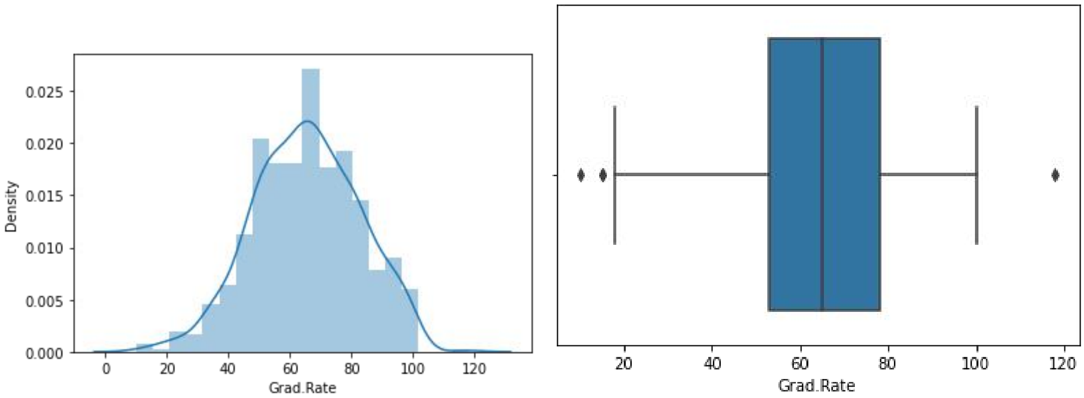




Expenditure



Grad Rate

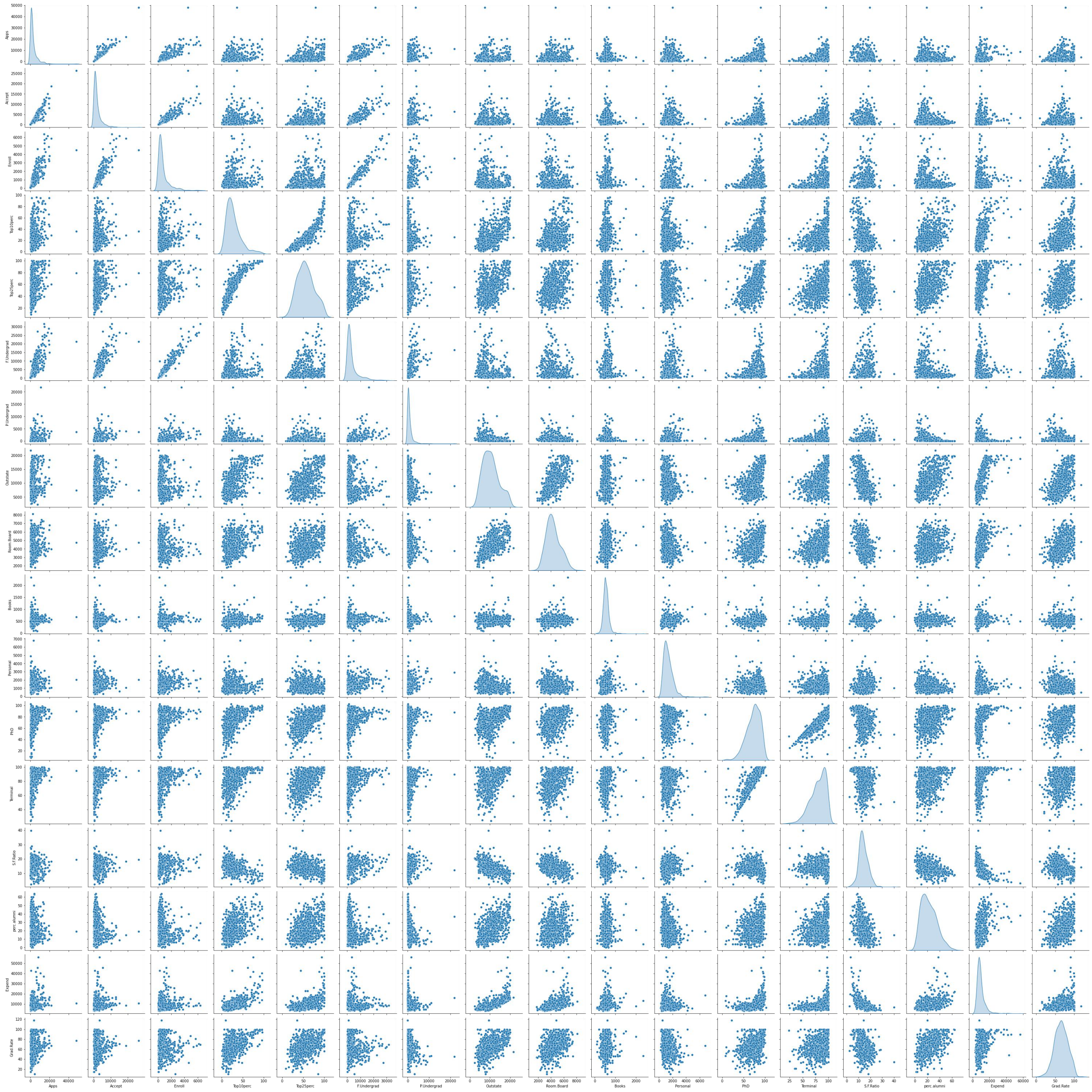


a. Right/Positively Skewed: PhD, Terminal

b. Normally distributed data: Top 25%, Outstate, Grad Rate

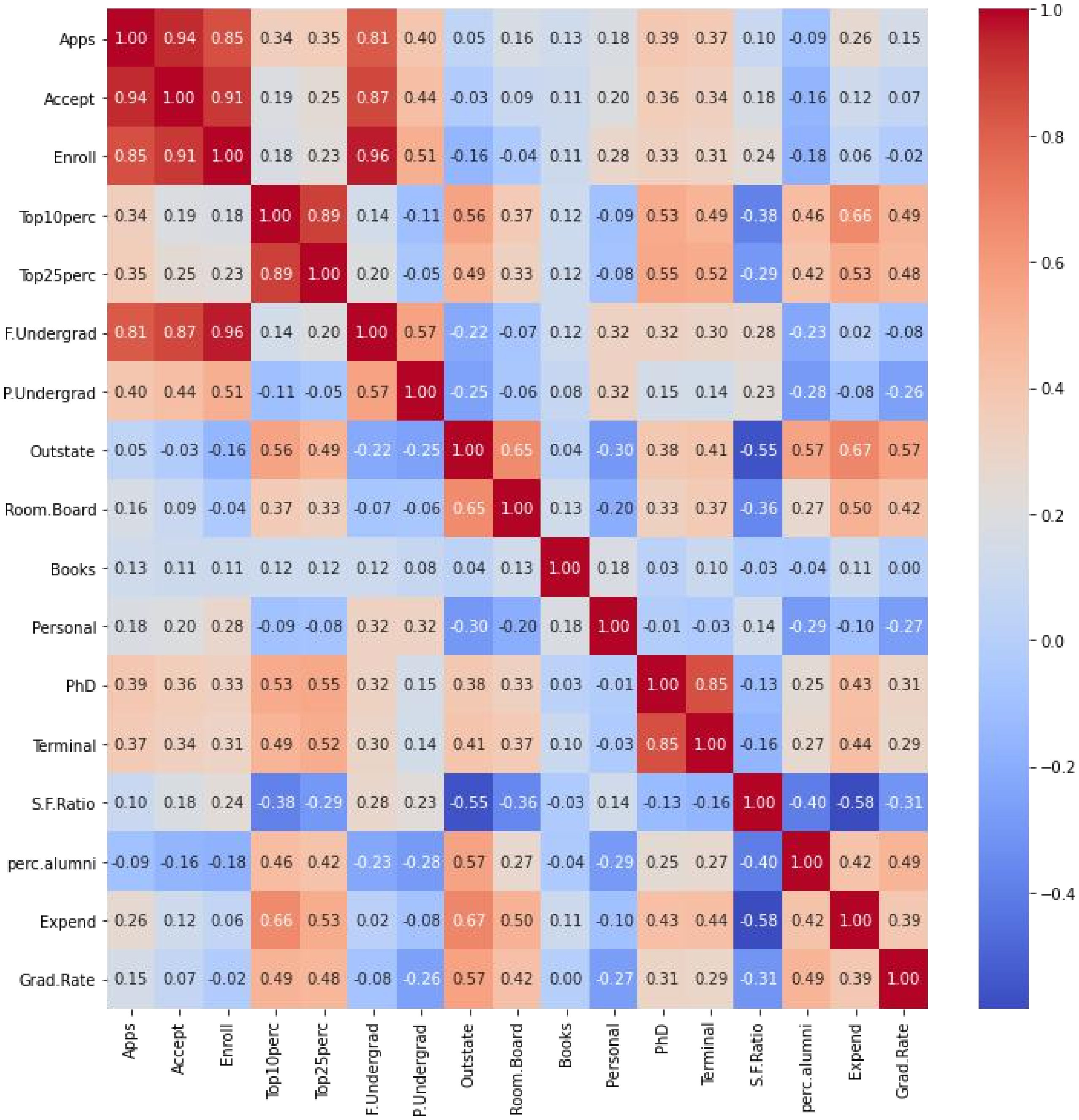
c. Left/Negatively Skewed: Apps, Accept, Top10%, F.Graduate, P.Graduate, Room Board, Books, Personal, SF Ratio, Perci Alumini, Expenditure

B. For Multivariate:



On comparing all values with each other we can understand trends in the dataset.

C. HEATMAP



The heatmap gives the correlation between numerical values.
We can understand the application variable is highly positively correlated with application accepted, students enrolled and full time graduates. So this relationship tells us when student submits the application it is accepted and the student is enrolled as full time graduate.

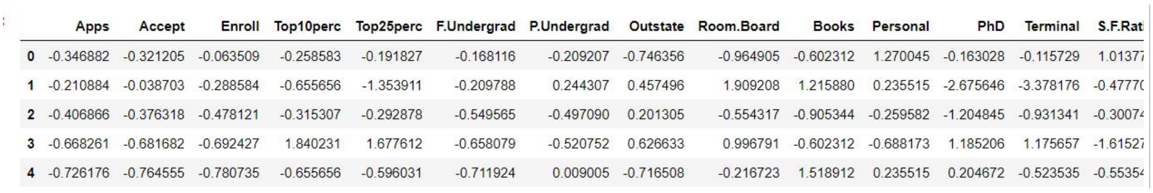
We can find negative correlation between application and percentage of alumni. This indicates us not all students are part of alumni of their college or university.

The application with top 10, 25 of higher secondary class, outstate, room board, books, personal, PhD, terminal. S.F ratio, expenditure and Graduation ratio are positively correlated.

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

<u>Answer:</u>
Since, there are 18 attributes, we got 18 principal components.
Once we get amount of variance of each PC we can know how many PCs are needed for PCA.
The PCA calculates new projections on the data set.
If we normalize our data, all variables have the same standard deviation, thus all variables have same weight and our PCA calculate relevant axis.

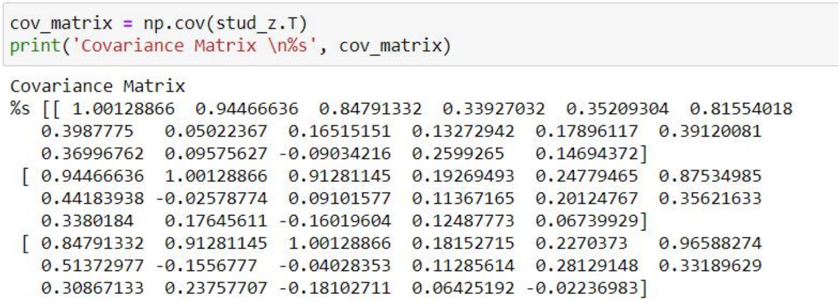
Scaling can be done using Z-score method



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Rat
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	1.01377
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	-0.4777
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	-0.3007
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	-1.6152
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	-0.5535

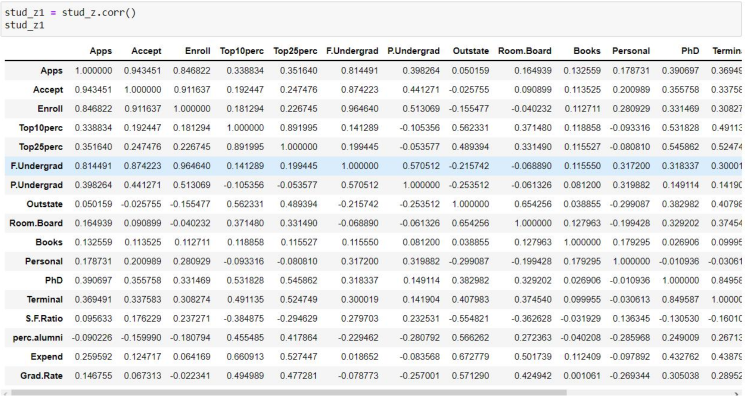
After Scaling Standard deviation is 1.0 for all variables.
Post scaling, Q1(25%) value and minimum values difference is lesser than original data set in most of the variables.

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



```
cov_matrix = np.cov(stud_z.T)
print('Covariance Matrix \n%s', cov_matrix)
```

```
Covariance Matrix
%s [[ 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]
```



```
stud_z1 = stud_z.corr()
stud_z1
```

	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	0.36949
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	0.33758
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	0.30827
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	0.49113
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	0.52474
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	0.30001
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	0.14190
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	0.40798
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	0.37454
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	0.09995
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	-0.03061
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	0.84958
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	1.00000
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	-0.16010
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	0.26713
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	0.43879
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	0.28952

Both measures relationship and dependency between two variables.

Highest corelation is seen among :

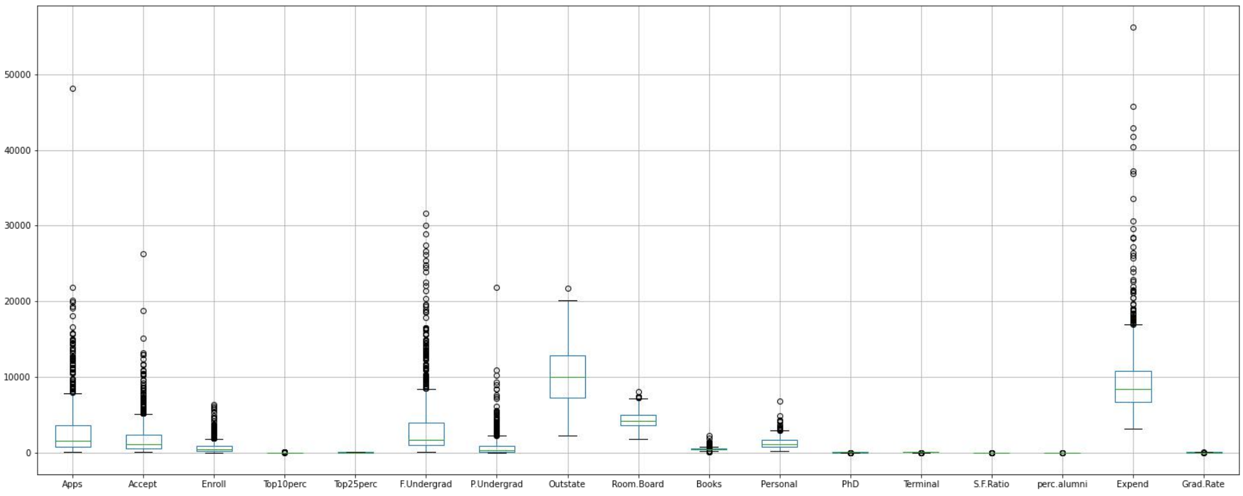
Enroll variable with F.Undergrad, Enroll with Accept, Apps with Accept and Apps, P.Undergrad, Terminal

Least corelations seen with SF Ratio variable with:
Expend, Outstate, Grad Rate, perc.alumni, Room board and Top10perc.

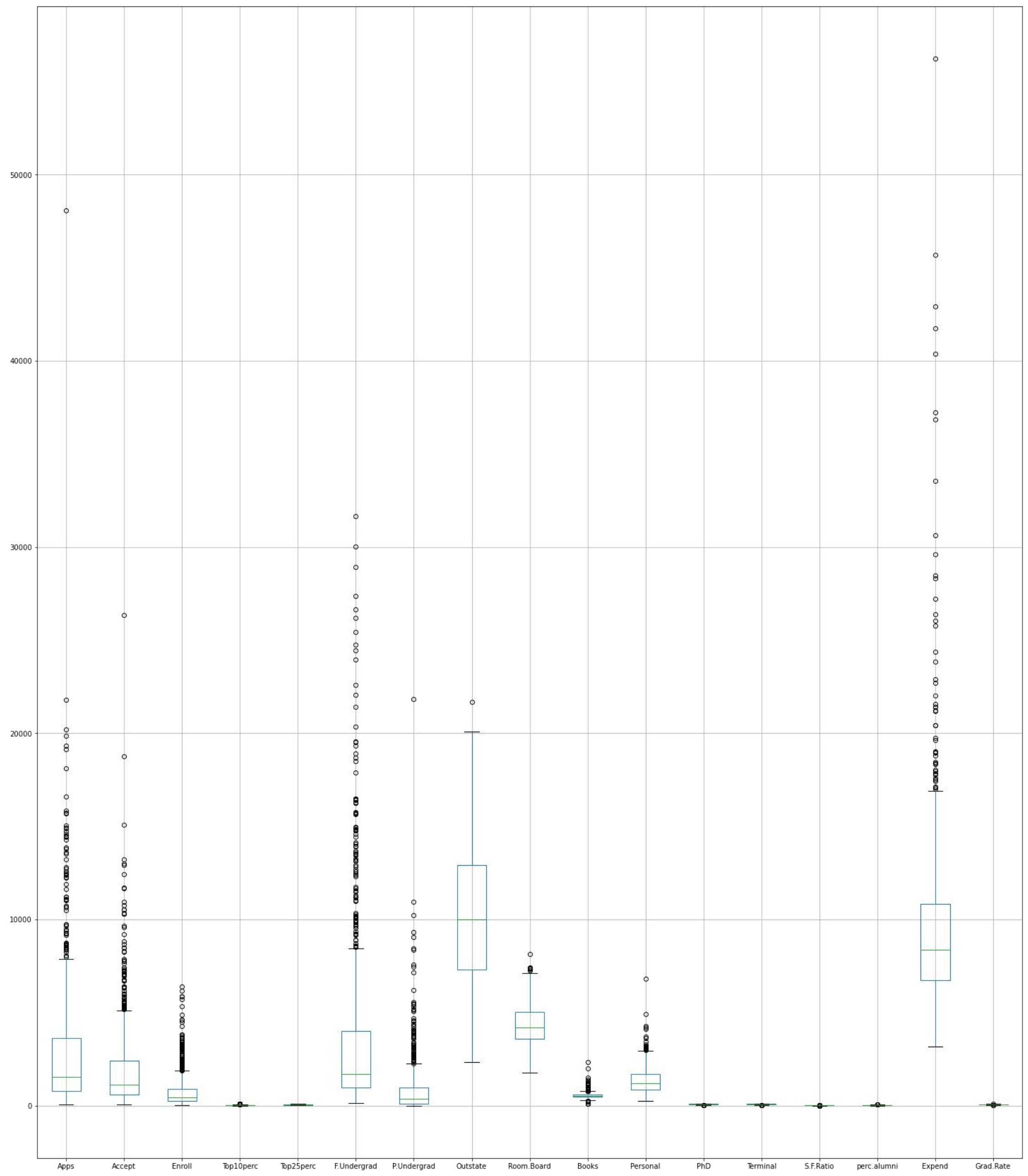
2.4 Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]

Answer:

Checking data before scaling



Checking data after scaling



The scaling shrinks the range of the feature values as shown in the left figure below. However, the outliers have an influence when computing the empirical mean and standard deviation.Standard Scalar therefore cannot guarantee balanced feature scales in the presence of outliers.

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

Answer:

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]]
```

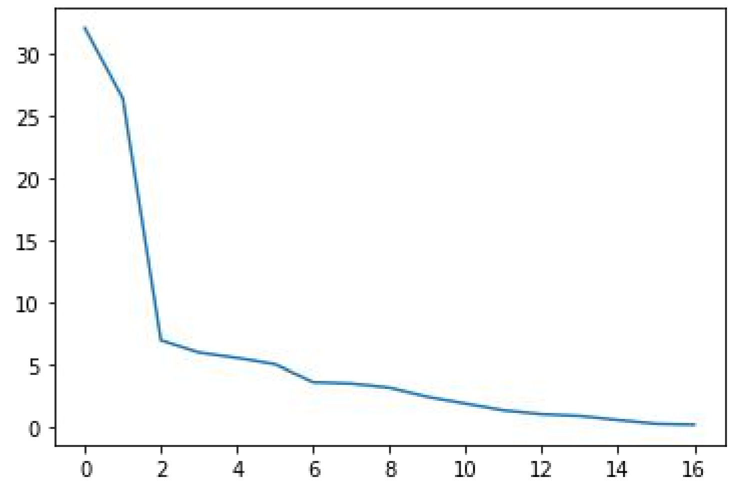
Eigen Values

```
%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]
```

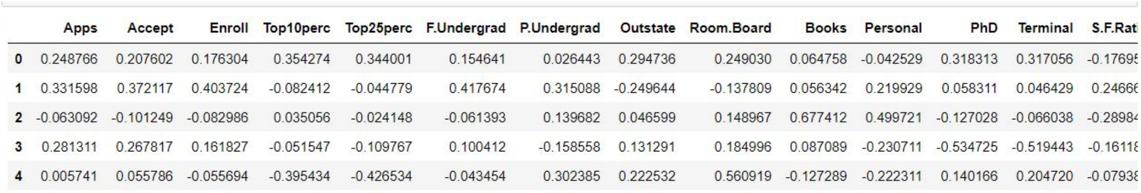
2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features

Answer:

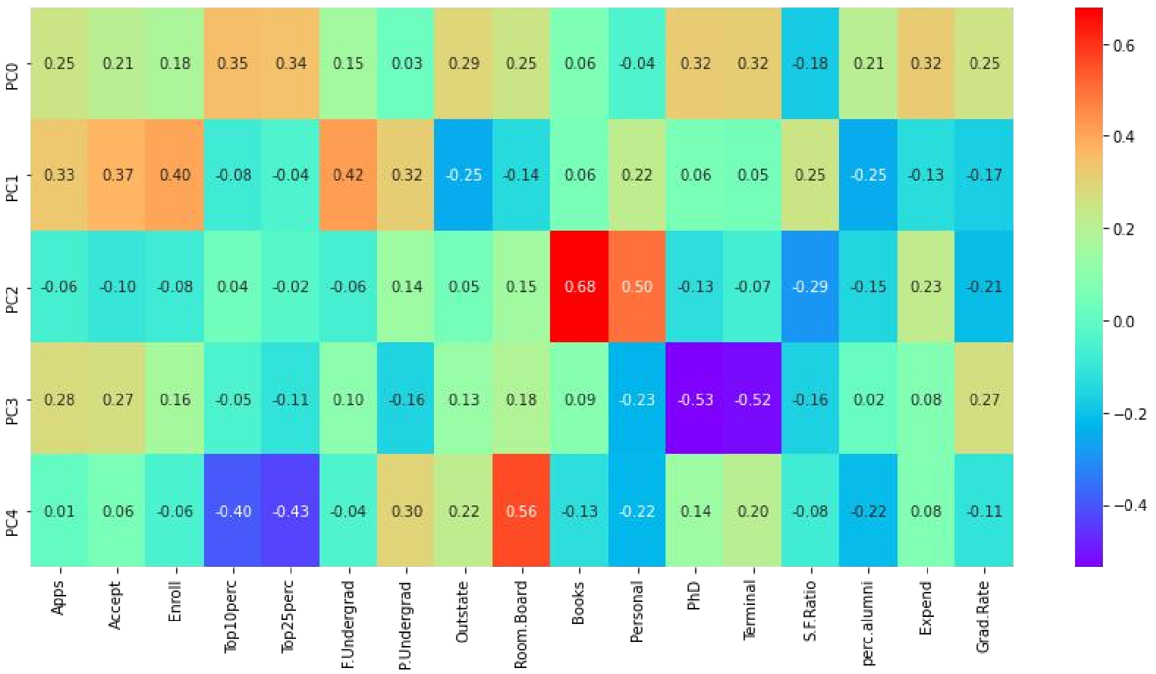
we can see that first PC or Array explains 33.12% variance in our dataset, while first seven features captures 70.12% variance.



Using Z-score we have done dimension reduction from 17 PCAs to 9 PCAs.

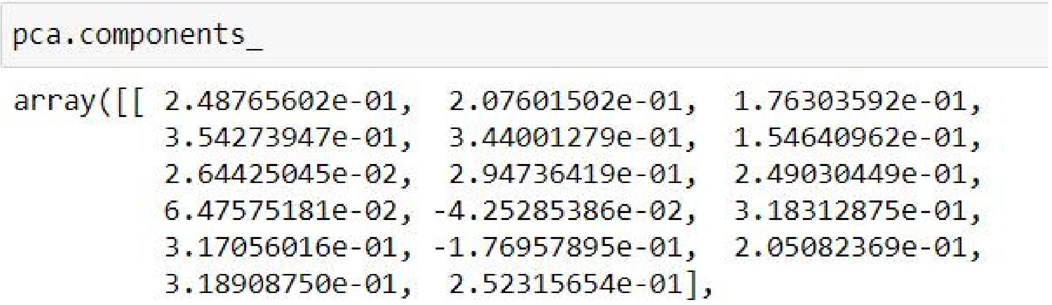


	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Rat
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0.064758	-0.042529	0.318313	0.317056	-0.17695
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809	0.056342	0.219929	0.058311	0.046429	0.24666
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	-0.28984
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	-0.16118
4	0.005741	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919	-0.127289	-0.222311	0.140166	0.204720	-0.07938


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
Apps
Accept
Enroll
Top10perc
Top25perc
F.Undergrad
P.Undergrad
Outstate
Room.Board
Books
Personal
PhD
Terminal
S.F.Ratio
perc.alumni
Expend
Grad.Rate
0.6
0.4
0.2
0.0
−0.2
−0.4

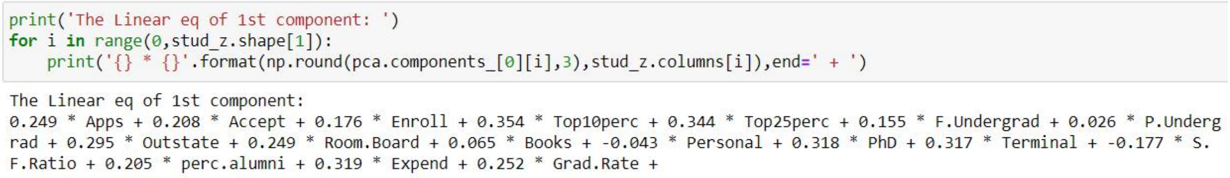
2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

Answer:



```
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],
```



```
print('The Linear eq of 1st component: ')
for i in range(0,stud_z.shape[1]):
    print('{} * {}'.format(np.round(pca.components_[0][i],3),stud_z.columns[i]),end=' + ')
```

```
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?

Answer:

PCA uses the eigenvectors of the covariance matrix to figure out how we should rotate the data. Because rotation is a kind of linear transformation, new dimensions will be sums of the old ones. The eigen-vectors, determine the direction or Axes along which linear transformation acts, stretching or compressing input vectors. They are the lines of change that represent the action of the larger matrix, the very “line” in linear transformation.

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]

Answer:

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.