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.***

Solution:

Formulation of hypothesis for conducting one-way ANOVA for education qualification w.r.t salary

- H0: Salary depend on education qualification
- Ha: Salary does not depend on education
- Confidence level = 0.05

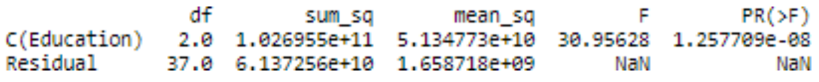
Formulation of hypothesis for conducting one-way ANOVA for occupation w.r.t salary

- H0: Salary depend on occupation
- Ha: Salary does not depend on occupation
- Confidence level = 0.05

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.***

Solution:

To perform one-way ANOVA for education w.r.t the variable 'Salary', we apply the ANOVA formula inthe Jupyter notebook and run the AOV table. We get following output:



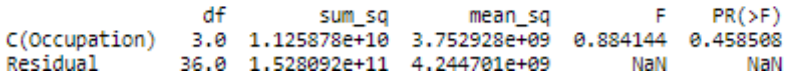
```
               df        sum_sq       mean_sq         F        PR(>F)
C(Education)  2.0  1.026955e+11  5.134773e+10  30.95628  1.257709e-08
Residual     37.0  6.137256e+10  1.658718e+09       NaN           NaN
```

From the above table, we find that the P value is less than 0.05, hence we reject the null hypothesis.

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.

Solution:

To perform one-way ANOVA for occupation w.r.t the variable 'Salary', we apply the ANOVA formula inthe Jupyter notebook and run the AOV table. We get following output:



```
                 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
```

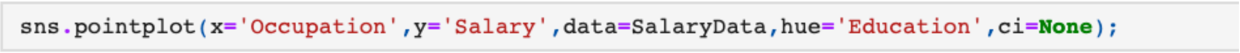
From the above table, we find that the P value is greater than 0.05, hence we do not reject the null hypothesis.

Problem 1B:

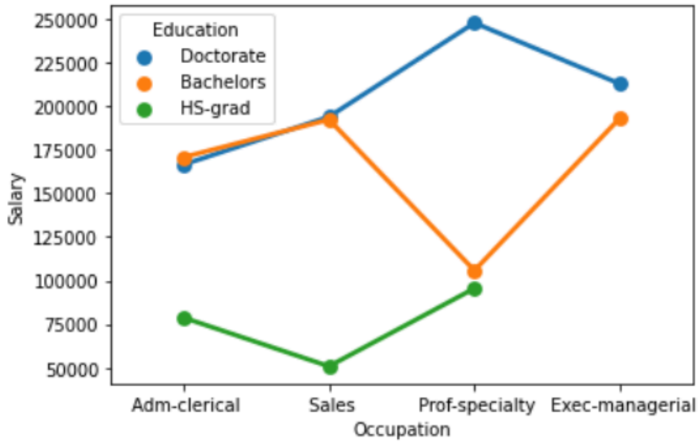
5. What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.

Solution:

As seen from the below interaction plots, there seems to be moderate interaction between the two categorical variables.



```
sns.pointplot(x='Occupation',y='Salary',data=SalaryData,hue='Education',ci=None);
```



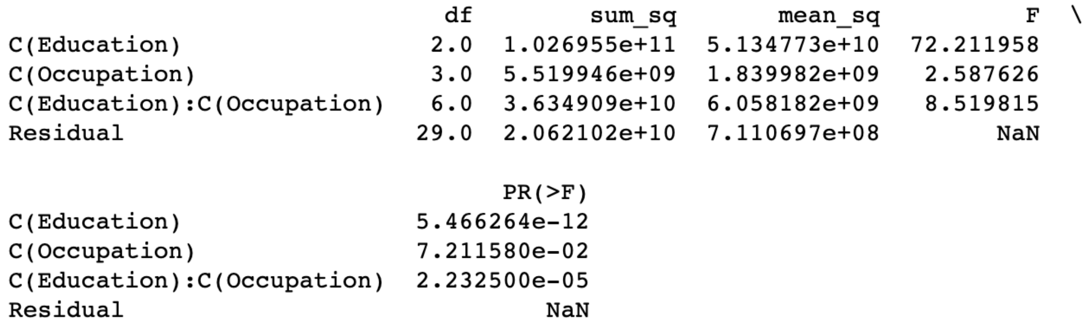
Adm-clerical and sales professionals with bachelors and doctorate degrees earn almost similar salary packages.

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?***

Solution:

Formulation of hypothesis for conducting two-way ANOVA based on education and occupation w.r.t salary.

- H0: Salary depends on both categories - education and occupation
- Ha: Salary does not depend on at least one of the categories - education and occupation
- Confidence level = 0.05



```
                               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  
```

Considering both education and occupation, education is a significant factor as the P value is <0.05, whereas occupation is not a significant variable as P value of it is >0.05

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

Solution:

By performing ANOVA on the given data set, we can conclude that salary is dependent on occupation.

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.

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

Solution:

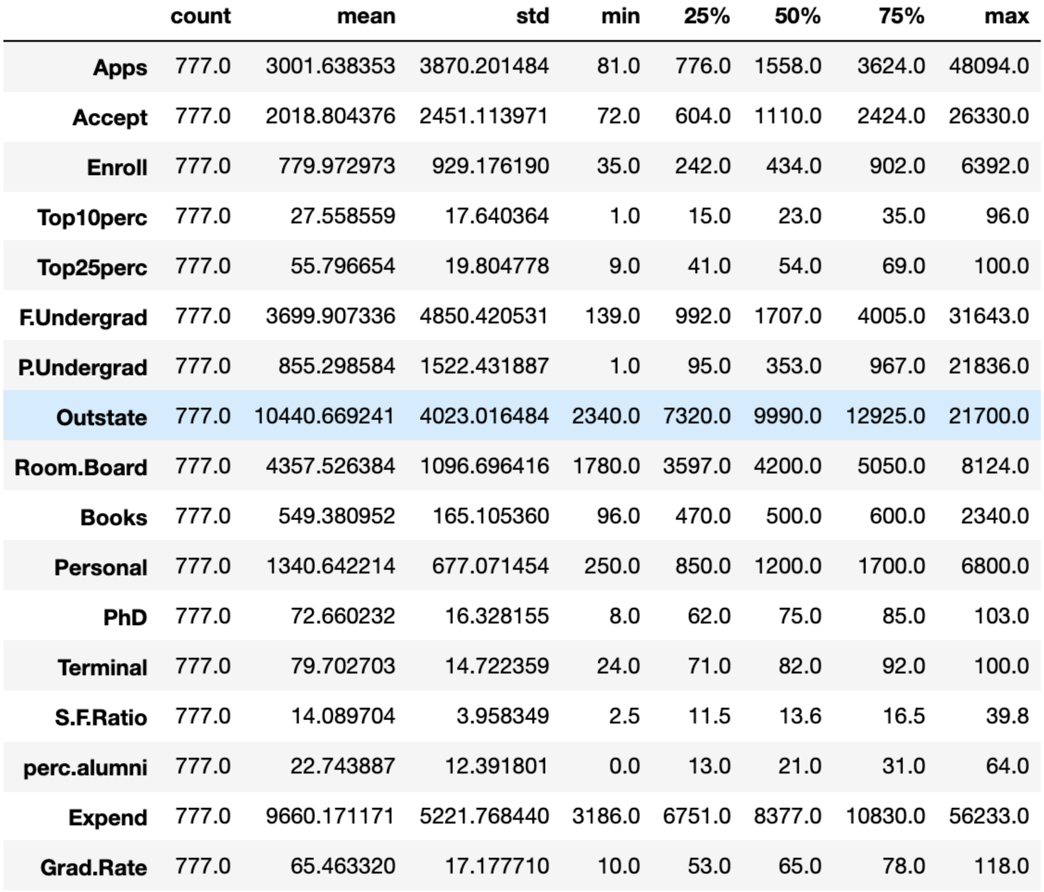
Firstly, after importing all the relevant libraries on Jupyter notebook, we load the data set. Then, we perform EDA to extract and see patterns in the given data set.

Advance Statistics Assignment

Observation from the initial analysis:

1. Data consists of 777 university / college information with 18 features.
2. There are no missing values in the data set.
3. Names of various university and colleges is the only column with categorical field.
4. 17 other columns are numerical fields.
5. Almost for all columns there are difference in mean and median values.which shows data set has some skewness.
6. Data set has outliers.

We then check the statistical summary of the data set, which is represented below



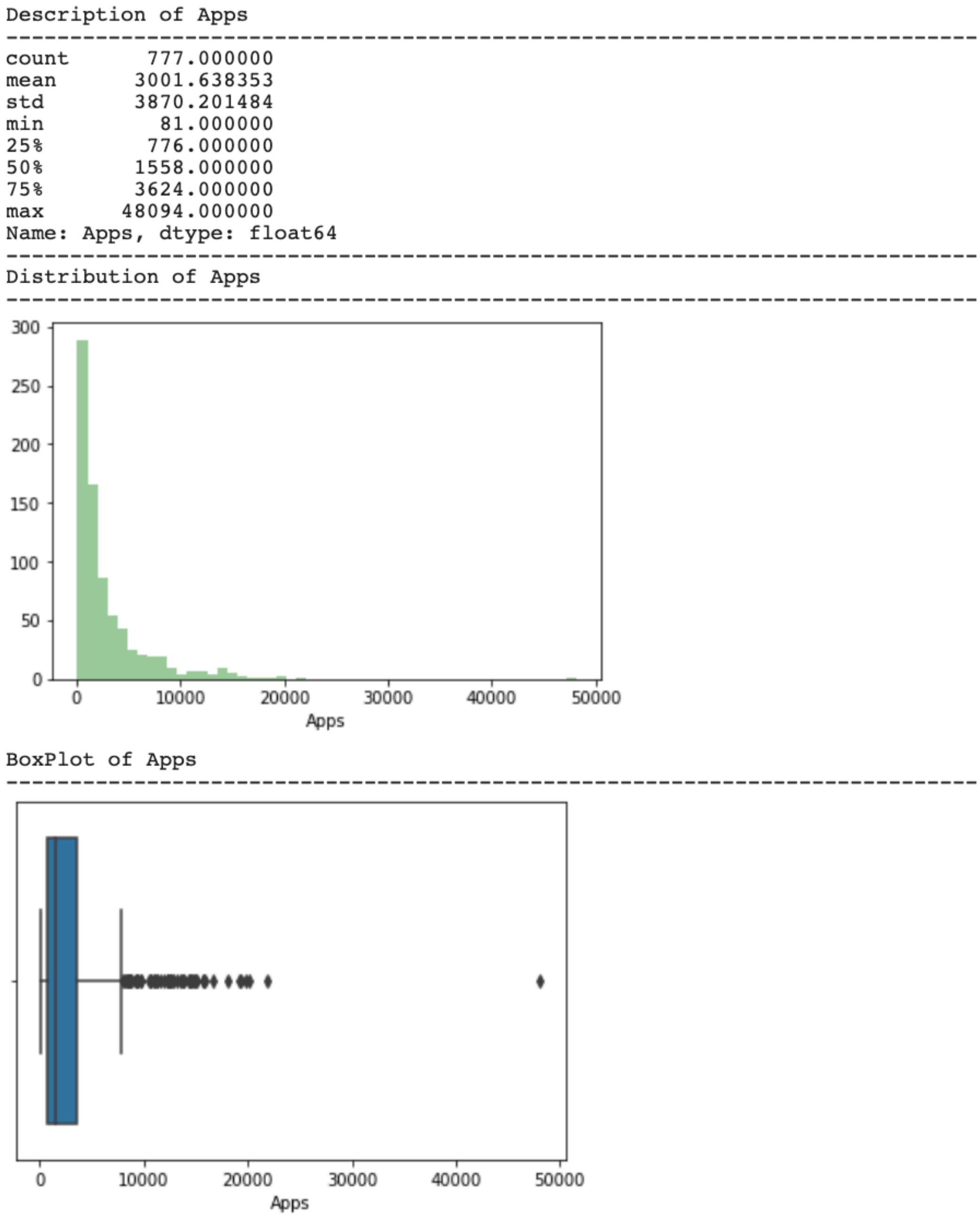
	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc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

Post this, we need to perform univariate analysis by defining “univariateAnalysis_numeric”, which includes 17 numeric variables and then adding a loop to perform the analysis for all the 17 numeric variables. The function will accept column name and number of bins as arguments.

The analysis of all these variables includes:

- Statistical description of the numeric variable
- Distribution of the column with histogram or distplot

- Boxplot representation of the column - 5 point summary and outliers if any

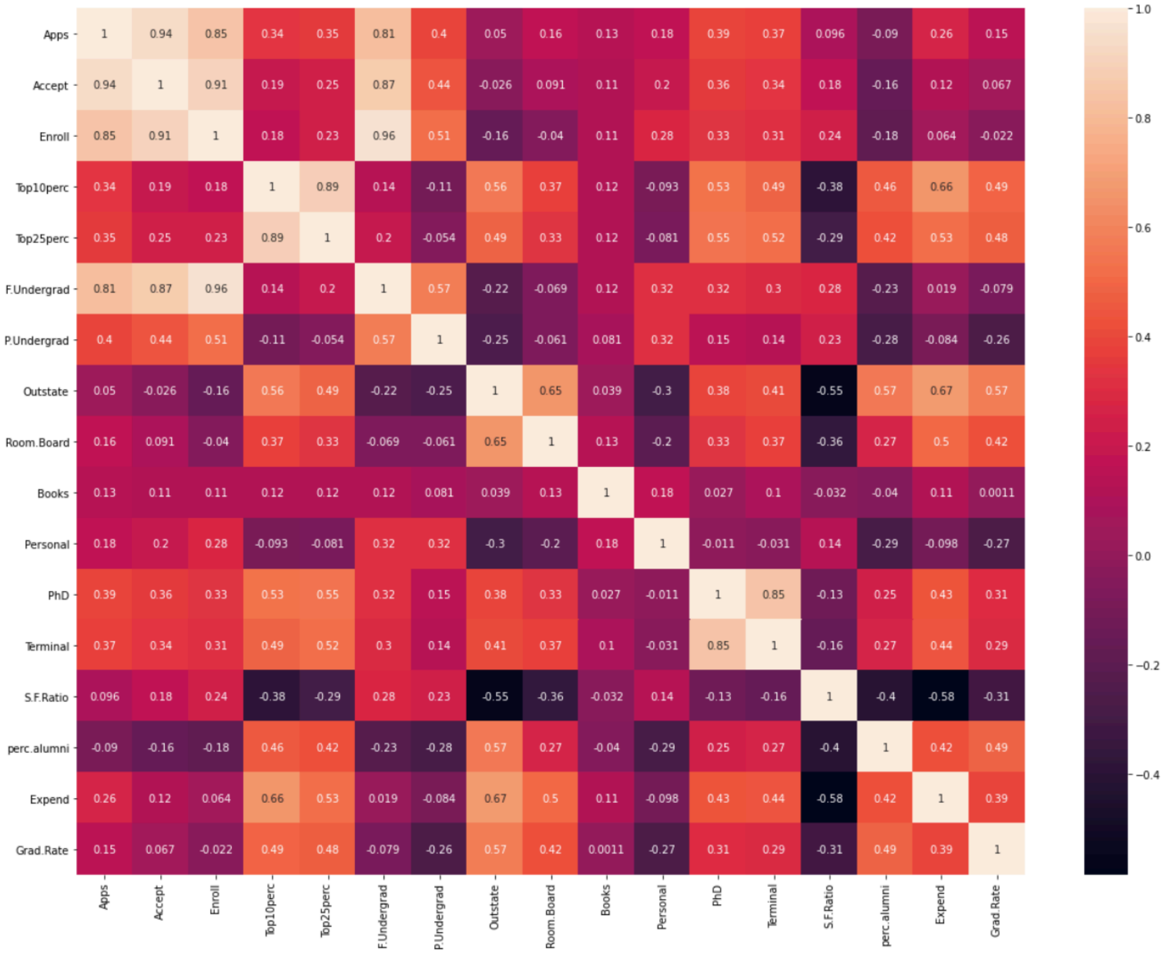


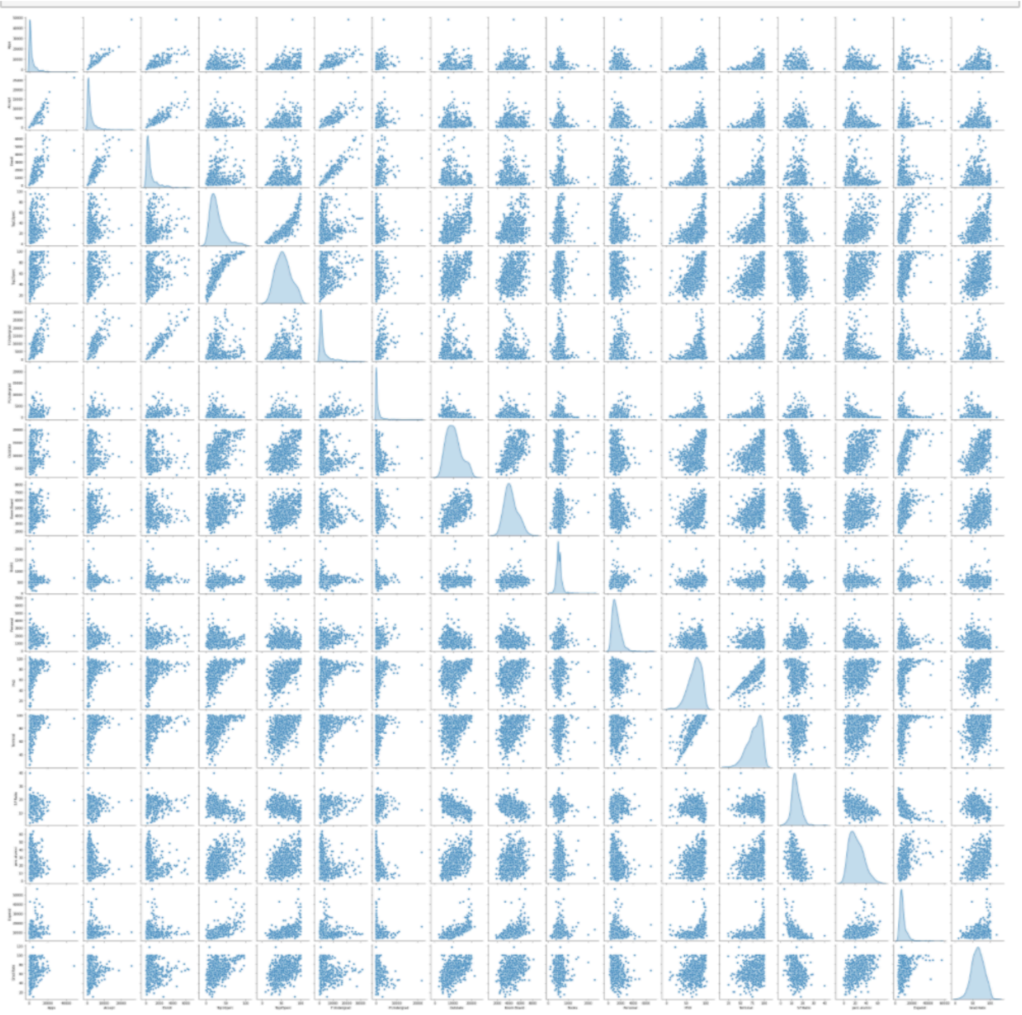
```
Description of Apps
----------------------------------------------------------------------------
count      777.000000
mean      3001.638353
std       3870.201484
min         81.000000
25%        776.000000
50%       1558.000000
75%       3624.000000
max      48094.000000
Name: Apps, dtype: float64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

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

The output displays, total 17*3 = 51 distinct charts/columns. Hence I have put the screenshot of onlyone variable i.e. apps (Please refer the python notebook for your perusal).

Further, we perform multivariate analysis, using correlation function & pair-plot in which we get below output.


Apps 1 0.94 0.85 0.34 0.35 0.81 0.4 0.05 0.16 0.13 0.18 0.39 0.37 0.096 -0.09 0.26 0.15
Accept 0.94 1 0.91 0.19 0.25 0.87 0.44 -0.026 0.091 0.11 0.2 0.36 0.34 0.18 -0.16 0.12 0.067
Enroll 0.85 0.91 1 0.18 0.23 0.96 0.51 -0.16 -0.04 0.11 0.28 0.33 0.31 0.24 -0.18 0.064 -0.022
Top10perc 0.34 0.19 0.18 1 0.89 0.14 -0.11 0.56 0.37 0.12 -0.093 0.53 0.49 -0.38 0.46 0.66 0.49
Top25perc 0.35 0.25 0.23 0.89 1 0.2 -0.054 0.49 0.33 0.12 -0.081 0.55 0.52 -0.29 0.42 0.53 0.48
F.Undergrad 0.81 0.87 0.96 0.14 0.2 1 0.57 -0.22 -0.069 0.12 0.32 0.32 0.3 0.28 -0.23 0.019 -0.079
P.Undergrad 0.4 0.44 0.51 -0.11 -0.054 0.57 1 -0.25 -0.061 0.081 0.32 0.15 0.14 0.23 -0.28 -0.084 -0.26
Outstate 0.05 -0.026 -0.16 0.56 0.49 -0.22 -0.25 1 0.65 0.039 -0.3 0.38 0.41 -0.55 0.57 0.67 0.57
Room.Board 0.16 0.091 -0.04 0.37 0.33 -0.069 -0.061 0.65 1 0.13 -0.2 0.33 0.37 -0.36 0.27 0.5 0.42
Books 0.13 0.11 0.11 0.12 0.12 0.12 0.081 0.039 0.13 1 0.18 0.027 0.1 -0.032 -0.04 0.11 0.0011
Personal 0.18 0.2 0.28 -0.093 -0.081 0.32 0.32 -0.3 -0.2 0.18 1 -0.011 -0.031 0.14 -0.29 -0.098 -0.27
PhD 0.39 0.36 0.33 0.53 0.55 0.32 0.15 0.38 0.33 0.027 -0.011 1 0.85 -0.13 0.25 0.43 0.31
Terminal 0.37 0.34 0.31 0.49 0.52 0.3 0.14 0.41 0.37 0.1 -0.031 0.85 1 -0.16 0.27 0.44 0.29
S.F.Ratio 0.096 0.18 0.24 -0.38 -0.29 0.28 0.23 -0.55 -0.36 -0.032 0.14 -0.13 -0.16 1 -0.4 -0.58 -0.31
perc.alumni -0.09 -0.16 -0.18 0.46 0.42 -0.23 -0.28 0.57 0.27 -0.04 -0.29 0.25 0.27 -0.4 1 0.42 0.49
Expend 0.26 0.12 0.064 0.66 0.53 0.019 -0.084 0.67 0.5 0.11 -0.098 0.43 0.44 -0.58 0.42 1 0.39
Grad.Rate 0.15 0.067 -0.022 0.49 0.48 -0.079 -0.26 0.57 0.42 0.0011 -0.27 0.31 0.29 -0.31 0.49 0.39 1
Apps Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4



Insights:

- The average (mean) number of applications received by the listed universities is around 3,001
- The number of applications accepted ranges from 72 to 26,330
- Average student enrolment is around ~880
- Median of new students from top 10% of higher secondary class is 23%
- Average book cost is around 550
- The minimum S.F. ratio is around 2.5
- Average percentage of faculties with Ph.D.'s is 72.6
- There are considerable number of variables that are highly correlated
- "Apps" has high correlation with "Accept", and "Enroll"

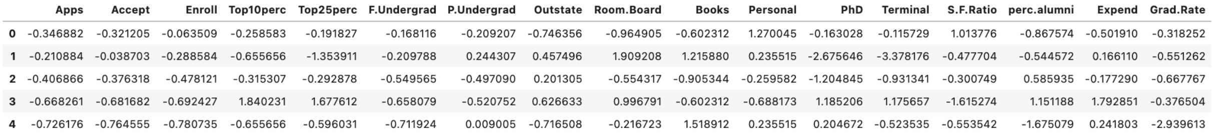
2. Is scaling necessary for PCA in this case? Give justification and perform scaling.

<u>Solution:</u>

Yes, it is necessary to perform scaling for PCA. For instance, in given data set, applications and other variables are having values in thousands and few variables such as percentile is in just two digits. So, the data in these variables are of different scales, it is tough to compare these variables.

The PCA calculates a new projection of the given data set and the new axis are based on the standard deviation of the variables. So a variable with a high standard deviation in the data set will have a higher weight for the calculation of axis than a variable with a low standard deviation. By performing scaling, we can easily compare these variables.

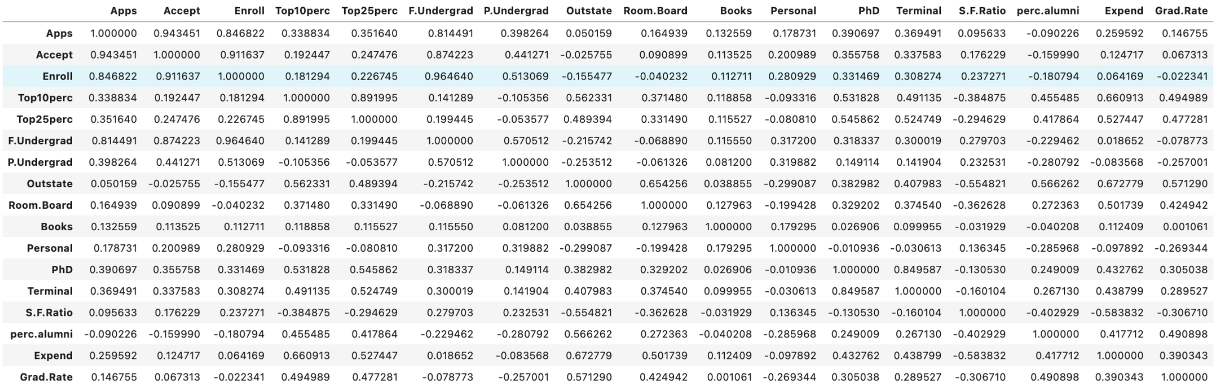
We get the following output, post we perform scaling **using Z score**..



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	-0.346882	-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	-0.867574	-0.501910	-0.318252
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	-0.477704	-0.544572	0.166110	-0.551262
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	-0.300749	0.585935	-0.177290	-0.667767
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	-1.615274	1.151188	1.792851	-0.376504
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	-0.553542	-1.675079	0.241803	-2.939613

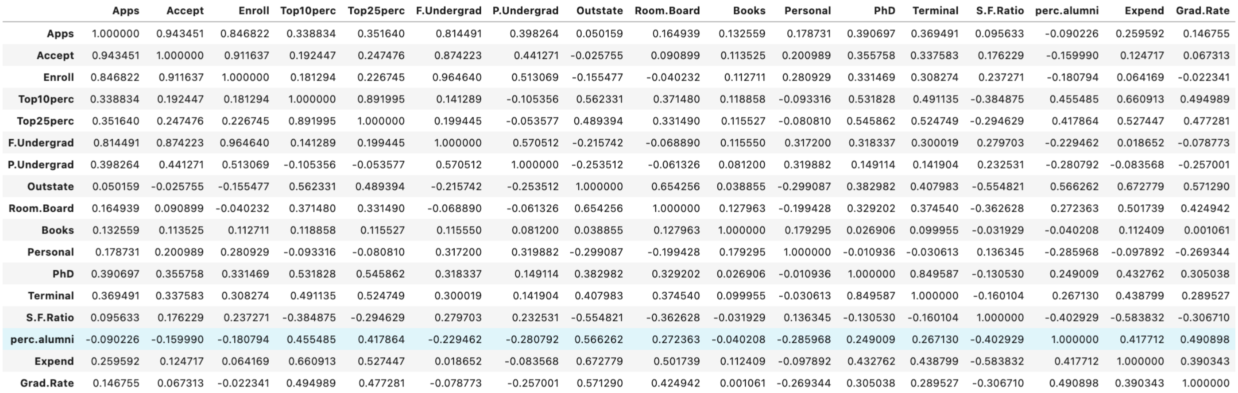
3. Comment on the comparison between the covariance and the correlation matrices from this data.

Without scaling the correlation matrix



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369491	0.095633	-0.090226	0.259592	0.146755
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	0.337583	0.176229	-0.159990	0.124717	0.067313
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	0.308274	0.237271	-0.180794	0.064169	-0.022341
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	0.491135	-0.384875	0.455485	0.660913	0.494989
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	0.524749	-0.294629	0.417864	0.527447	0.477281
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	0.300019	0.279703	-0.229462	0.018652	-0.078773
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	0.141904	0.232531	-0.280792	-0.083568	-0.257001
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	0.407983	-0.554821	0.566262	0.672779	0.571290
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	0.374540	-0.362628	0.272363	0.501739	0.424942
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	0.099955	-0.031929	-0.040208	0.112409	0.001061
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	-0.030613	0.136345	-0.285968	-0.097892	-0.269344
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	0.849587	-0.130530	0.249009	0.432762	0.305038
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	1.000000	-0.160104	0.267130	0.438799	0.289527
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	-0.160104	1.000000	-0.402929	-0.583832	-0.306710
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	0.267130	-0.402929	1.000000	0.417712	0.490898
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	0.438799	-0.583832	0.417712	1.000000	0.390343
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	0.289527	-0.306710	0.490898	0.390343	1.000000

With scaling the correlation matrix



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369491	0.095633	-0.090226	0.259592	0.146755
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	0.337583	0.176229	-0.159990	0.124717	0.067313
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	0.308274	0.237271	-0.180794	0.064169	-0.022341
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	0.491135	-0.384875	0.455485	0.660913	0.494989
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	0.524749	-0.294629	0.417864	0.527447	0.477281
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	0.300019	0.279703	-0.229462	0.018652	-0.078773
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	0.141904	0.232531	-0.280792	-0.083568	-0.257001
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	0.407983	-0.554821	0.566262	0.672779	0.571290
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	0.374540	-0.362628	0.272363	0.501739	0.424942
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	0.099955	-0.031929	-0.040208	0.112409	0.001061
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	-0.030613	0.136345	-0.285968	-0.097892	-0.269344
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	0.849587	-0.130530	0.249009	0.432762	0.305038
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	1.000000	-0.160104	0.267130	0.438799	0.289527
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	-0.160104	1.000000	-0.402929	-0.583832	-0.306710
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	0.267130	-0.402929	1.000000	0.417712	0.490898
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	0.438799	-0.583832	0.417712	1.000000	0.390343
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	0.289527	-0.306710	0.490898	0.390343	1.000000

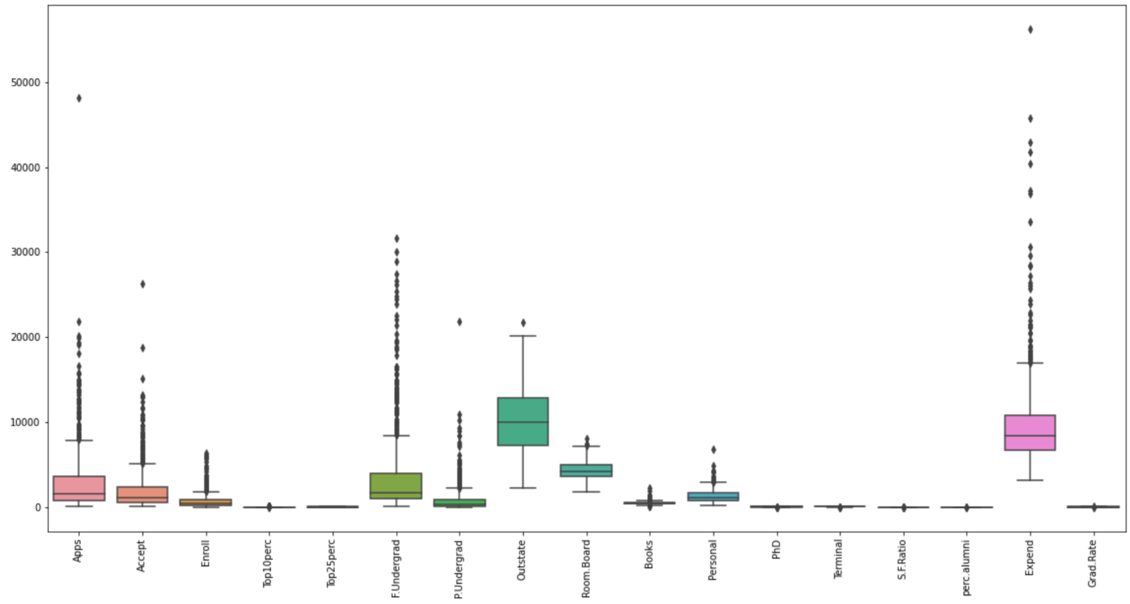
Covariance tells us the direction of the linear relationship between two variables. Correlation indicates the measures of both the strength and direction of the linear relationship between two variables. Correlation is function of the covariance. We can find the correlation coefficient of two variables by dividing the covariance of these variables by the product of the standard deviations of the same values. Now without Scaling let's check out correlation matrix and after scaling also

correlation matrix yields the same result.

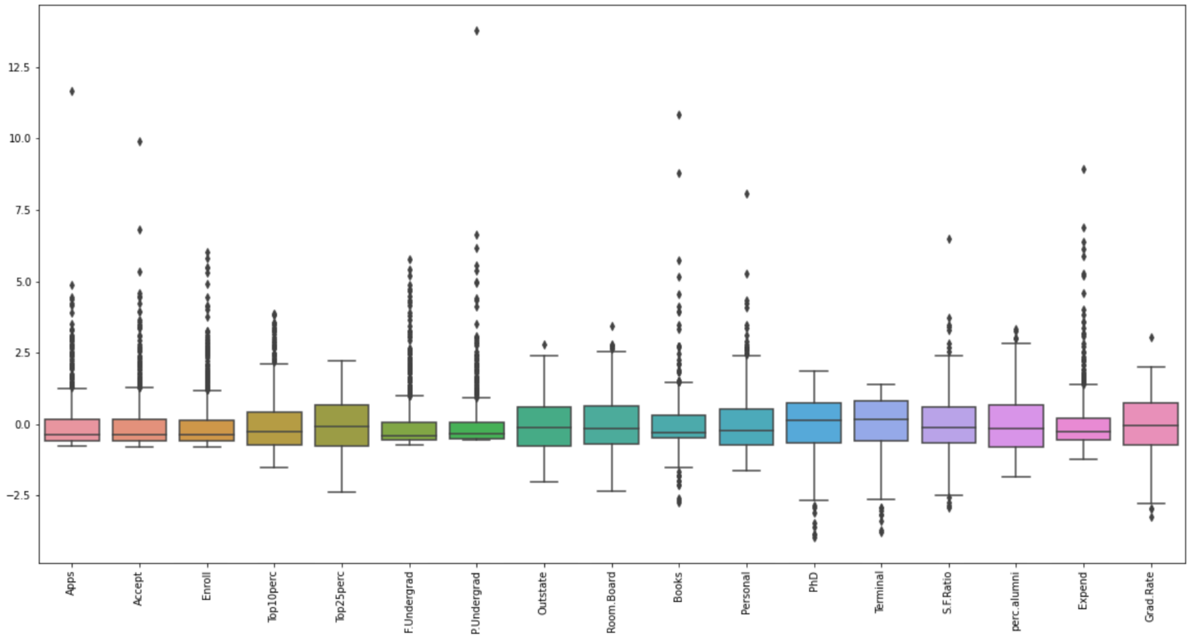
4. Check the dataset for outliers before and after scaling. What insight do you derive here?

Solution:

Before scaling, let's plot a boxplot to check the outliers in all the variables. We get the following output:



Post scaling, let's plot a boxplot to check the outliers in all the variables. We get the following output:



Insights:

- By scaling, all variables have the same standard deviation, thus all variables have the same weight and thus resulting in PCA calculating relevant axis.
- Before scaling, we only had one variable with no outliers (top25 perc); Post scaling, we have multiple variables with negligible outliers – this is achieved by normalizing the scale of the variables

5. Perform PCA and export the data of the Principal Component scores into a data frame.

Solution:

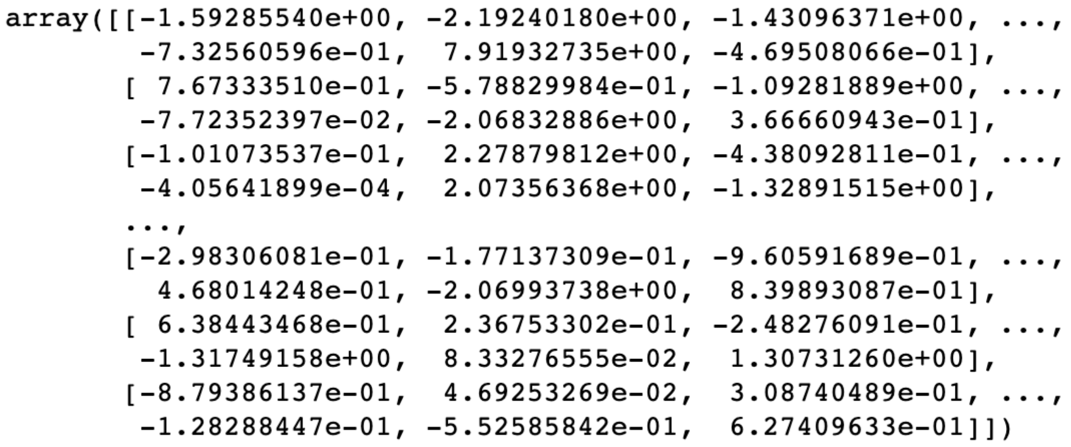
For performing PCA, we need to follow below steps:

# Step 1: Generate the covariance matrix

# Step 2: Get eigenvalues and eigenvector

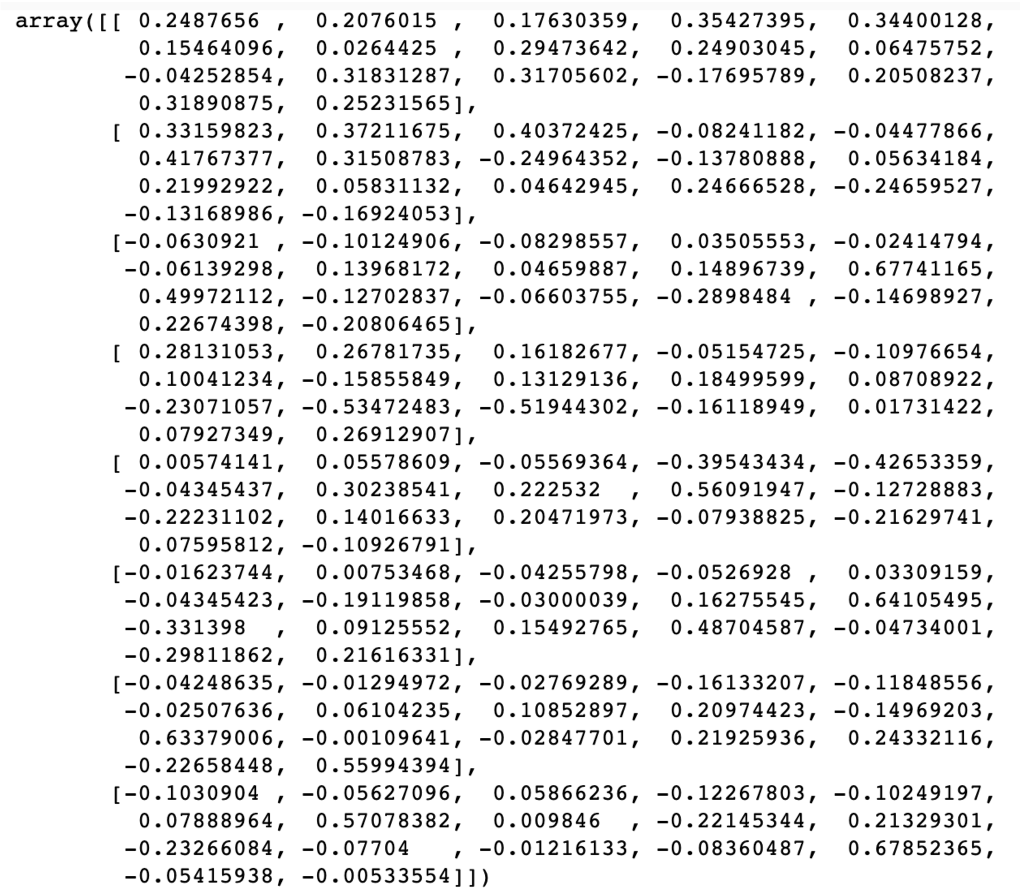
# Step 3: View Scree Plot to identify the number of components to be built

# Step 4: We can perform PCA on the scaled data set by importing PCA from sklearn.decomposition. We get following component output:



```
array([[-1.59285540e+00, -2.19240180e+00, -1.43096371e+00, ...,
        -7.32560596e-01,  7.91932735e+00, -4.69508066e-01],
       [ 7.67333510e-01, -5.78829984e-01, -1.09281889e+00, ...,
        -7.72352397e-02, -2.06832886e+00,  3.66660943e-01],
       [-1.01073537e-01,  2.27879812e+00, -4.38092811e-01, ...,
        -4.05641899e-04,  2.07356368e+00, -1.32891515e+00],
       ...,
       [-2.98306081e-01, -1.77137309e-01, -9.60591689e-01, ...,
         4.68014248e-01, -2.06993738e+00,  8.39893087e-01],
       [ 6.38443468e-01,  2.36753302e-01, -2.48276091e-01, ...,
        -1.31749158e+00,  8.33276555e-02,  1.30731260e+00],
       [-8.79386137e-01,  4.69253269e-02,  3.08740489e-01, ...,
        -1.28288447e-01, -5.52585842e-01,  6.27409633e-01]])
```

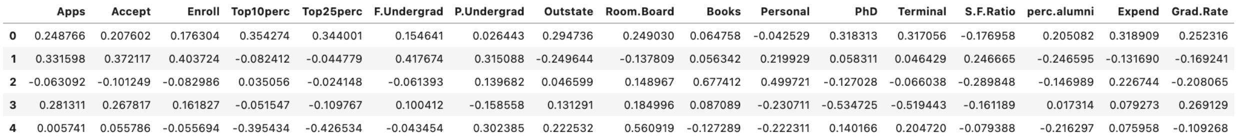
Then, we can do loading of each feature on the components



```
array([[ 0.2487656 ,  0.2076015 ,  0.17630359,  0.35427395,  0.34400128,
         0.15464096,  0.0264425 ,  0.29473642,  0.24903045,  0.06475752,
        -0.04252854,  0.31831287,  0.31705602, -0.17695789,  0.20508237,
         0.31890875,  0.25231565],
       [ 0.33159823,  0.37211675,  0.40372425, -0.08241182, -0.04477866,
         0.41767377,  0.31508783, -0.24964352, -0.13780888,  0.05634184,
         0.21992922,  0.05831132,  0.04642945,  0.24666528, -0.24659527,
        -0.13168986, -0.16924053],
       [-0.0630921 , -0.10124906, -0.08298557,  0.03505553, -0.02414794,
        -0.06139298,  0.13968172,  0.04659887,  0.14896739,  0.67741165,
         0.49972112, -0.12702837, -0.06603755, -0.2898484 , -0.14698927,
         0.22674398, -0.20806465],
       [ 0.28131053,  0.26781735,  0.16182677, -0.05154725, -0.10976654,
         0.10041234, -0.15855849,  0.13129136,  0.18499599,  0.08708922,
        -0.23071057, -0.53472483, -0.51944302, -0.16118949,  0.01731422,
         0.07927349,  0.26912907],
       [ 0.00574141,  0.05578609, -0.05569364, -0.39543434, -0.42653359,
        -0.04345437,  0.30238541,  0.222532  ,  0.56091947, -0.12728883,
        -0.22231102,  0.14016633,  0.20471973, -0.07938825, -0.21629741,
         0.07595812, -0.10926791],
       [-0.01623744,  0.00753468, -0.04255798, -0.0526928 ,  0.03309159,
        -0.04345423, -0.19119858, -0.03000039,  0.16275545,  0.64105495,
        -0.331398  ,  0.09125552,  0.15492765,  0.48704587, -0.04734001,
        -0.29811862,  0.21616331],
       [-0.04248635, -0.01294972, -0.02769289, -0.16133207, -0.11848556,
        -0.02507636,  0.06104235,  0.10852897,  0.20974423, -0.14969203,
         0.63379006, -0.00109641, -0.02847701,  0.21925936,  0.24332116,
        -0.22658448,  0.55994394],
       [-0.1030904 , -0.05627096,  0.05866236, -0.12267803, -0.10249197,
         0.07888964,  0.57078382,  0.009846  , -0.22145344,  0.21329301,
        -0.23266084, -0.07704   , -0.01216133, -0.08360487,  0.67852365,
        -0.05415938, -0.00533554]])
```

Post that, we can load these components into a dataframe along with the list of columns we had earlier considered in Educationdata_scaled.

Below is the representative screenshot of Educationdata_pca_comp in which we had exported the principalcomponent scores into a data frame.



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0.064758	-0.042529	0.318313	0.317056	-0.176958	0.205082	0.318909	0.252316
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	0.246665	-0.246595	-0.131690	-0.169241
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	-0.289848	-0.146989	0.226744	-0.208065
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	-0.161189	0.017314	0.079273	0.269129
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	-0.079388	-0.216297	0.075958	-0.109268

6. Extract the eigenvalues, and eigenvectors.

Solution:

We can extract the above represented eigenvalues and eigenvectors using covariance matrix.The below snapshot represent extracted eigenvalues and eigenvectors:

```
Eigen Vectors :

[[-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 :

```
[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]
```

7. Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).

Solution:

First PC in terms of Eigen Vector : eig_vecs[0]

Eigen Vector of First PC

[2.42 3.24 9.77 -1.02

2.28 -4.76 1.23 -3.41

-1.84 -1.34 -6.79 -1.51

5.73 2.54 -3.50 4.76

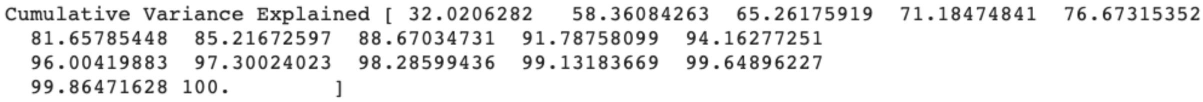
-2.73]

If we sort the eigenvectors in descending order with respect to their eigenvalues, we will have that the first eigenvector accounts for the largest spread among data, the second one for the second largest spread and so on.

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?

Solution:

From the below screenshot of cumulative values of the eigenvalues, we can see that around 8 principal components explained almost 90% of the variance. Thus, the optimum number of principal components can be 8.



```
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.        ]
```

Furthermore, eigenvectors indicate the direction of the principal components, we can multiply the original data by the eigenvectors to re-orient our data onto the new axes.

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]

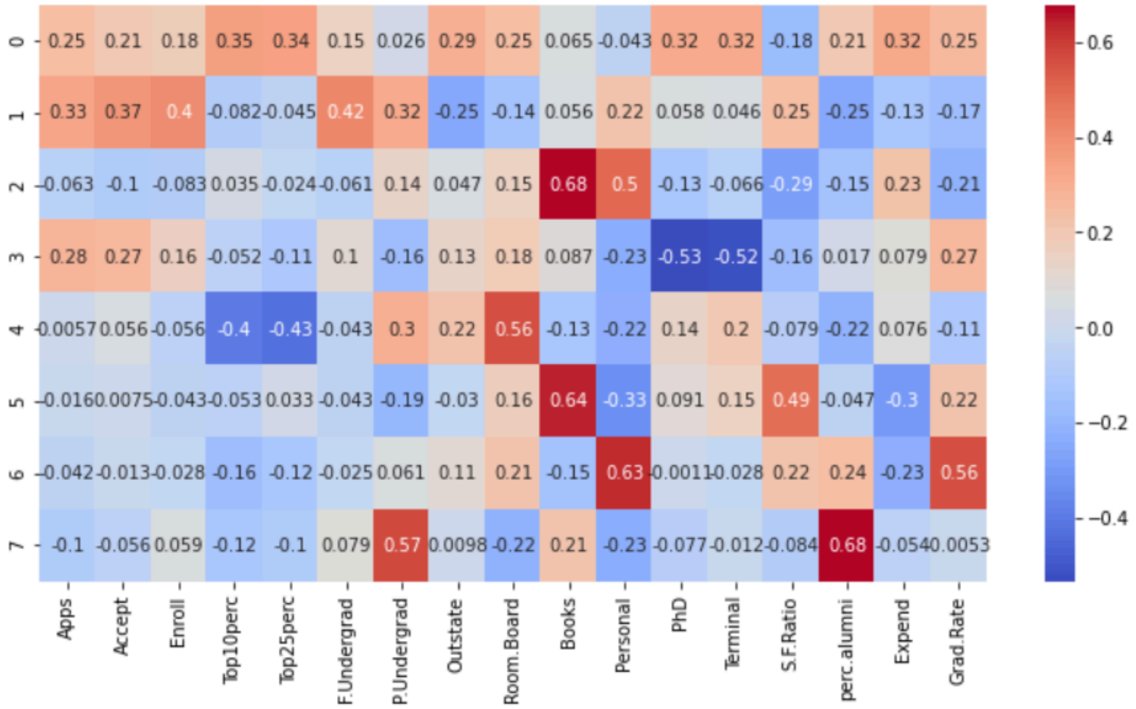
Solution:

We know that the principal components describe the amount of the total variance that can be explained by a single dimension of the data. As mentioned above, we have generated only 8 PCA dimensions. These 8 PCA can be used for further analysis, representing almost 90% of the variance.

In this case study, we had 17 numeric variables to be assessed, with PCA we did dimensionality reduction from 17 to 8 (representing almost 90% of the variance).

However, we can see from the above mentioned cumulative variance that even 5 PCA dimensions represent around 80% of the variance. But, to be on a safer side, we have considered to go with 90% variance.

Thus, as far as business implication of using PCA is concerned, in this case, we are reducing a high-dimensional space (with 17 variables) and converting it to a lower dimensional space without (theoretically) losing much of the explanatory power.



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	0.25	0.21	0.18	0.35	0.34	0.15	0.026	0.29	0.25	0.065	-0.043	0.32	0.32	-0.18	0.21	0.32	0.25
1	0.33	0.37	0.4	-0.082	-0.045	0.42	0.32	-0.25	-0.14	0.056	0.22	0.058	0.046	0.25	-0.25	-0.13	-0.17
2	-0.063	-0.1	-0.083	0.035	-0.024	-0.061	0.14	0.047	0.15	0.68	0.5	-0.13	-0.066	-0.29	-0.15	0.23	-0.21
3	0.28	0.27	0.16	-0.052	-0.11	0.1	-0.16	0.13	0.18	0.087	-0.23	-0.53	-0.52	-0.16	0.017	0.079	0.27
4	-0.0057	0.056	-0.056	-0.4	-0.43	-0.043	0.3	0.22	0.56	-0.13	-0.22	0.14	0.2	-0.079	-0.22	0.076	-0.11
5	-0.016	0.0075	-0.043	-0.053	0.033	-0.043	-0.19	-0.03	0.16	0.64	-0.33	0.091	0.15	0.49	-0.047	-0.3	0.22
6	-0.042	-0.013	-0.028	-0.16	-0.12	-0.025	0.061	0.11	0.21	-0.15	0.63	-0.0011	-0.028	0.22	0.24	-0.23	0.56
7	-0.1	-0.056	0.059	-0.12	-0.1	0.079	0.57	0.0098	-0.22	0.21	-0.23	-0.077	-0.012	-0.084	0.68	-0.054	0.0053

Following are the interpretations from the obtained PCs

- PC1: Explains No. of students for whom the particular college or university is Out-of-state tuition and instructional expenditure per student
- PC2: Represents the highly correlated variables such as Apps, Enroll and Accept
- PC3: Highlights the estimated cost of books for a student
- PC4: Represents % of faculties with Ph.D.'s and terminal degree
- PC5: Explains percentage of new students from top 10% and 25% of higher secondary class including cost of room and board
- PC6: Details about student/faculty ratio
- PC7: Highlights estimated personal spending for a student and graduation rate
- PC8: Explains number of part-time undergraduate students and alumni who donate