SMDM Project Report

Contents

Plots and tables

Plots / Page no.

Problem 1:

Problem Statement:

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.]

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

Solution:

Null and alternate hypothesis for conducting one-way ANOVA for 'Education' with respect to 'Salary':

Ho : Salary depend on Education

Ha : Salary does not depend on Education

Confidence level = 0.05

Null and alternate hypothesis for conducting one-way ANOVA for 'Occupation' with respect to

'Salary':
Ho: Salary depend on Education
Ha: Salary does not depend on Occupation
Confidence level = 0.05

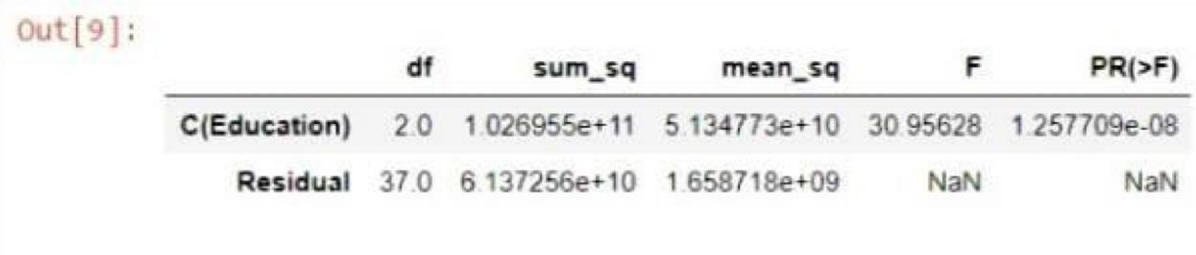
Question 1.2 Perform one-way ANOVA for Education with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

Solution:
One-way ANOVA ,Education w.r.to Salary:
From above table, we found that the P value is less than 0.05, hence the null hypothesis is rejected.

Question 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.

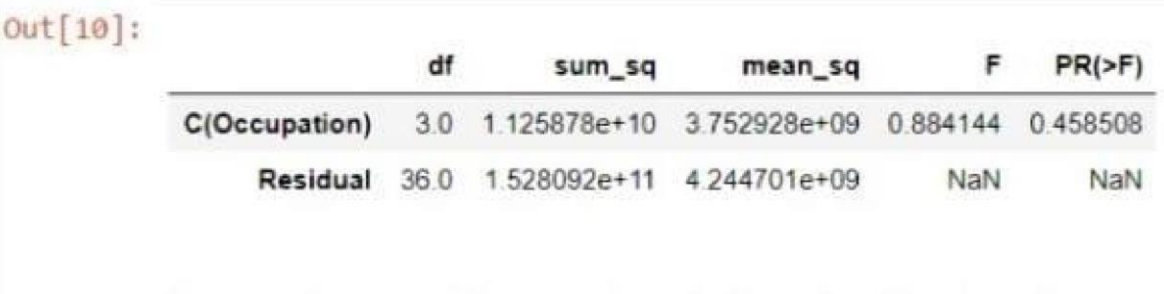
Solution:
One-way ANOVA, Occupation w.r.to Salary:
From above table, we found that the P value is greater than 0.05, Hence we fail to reject the null hypothesis.

Question 1.5 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.



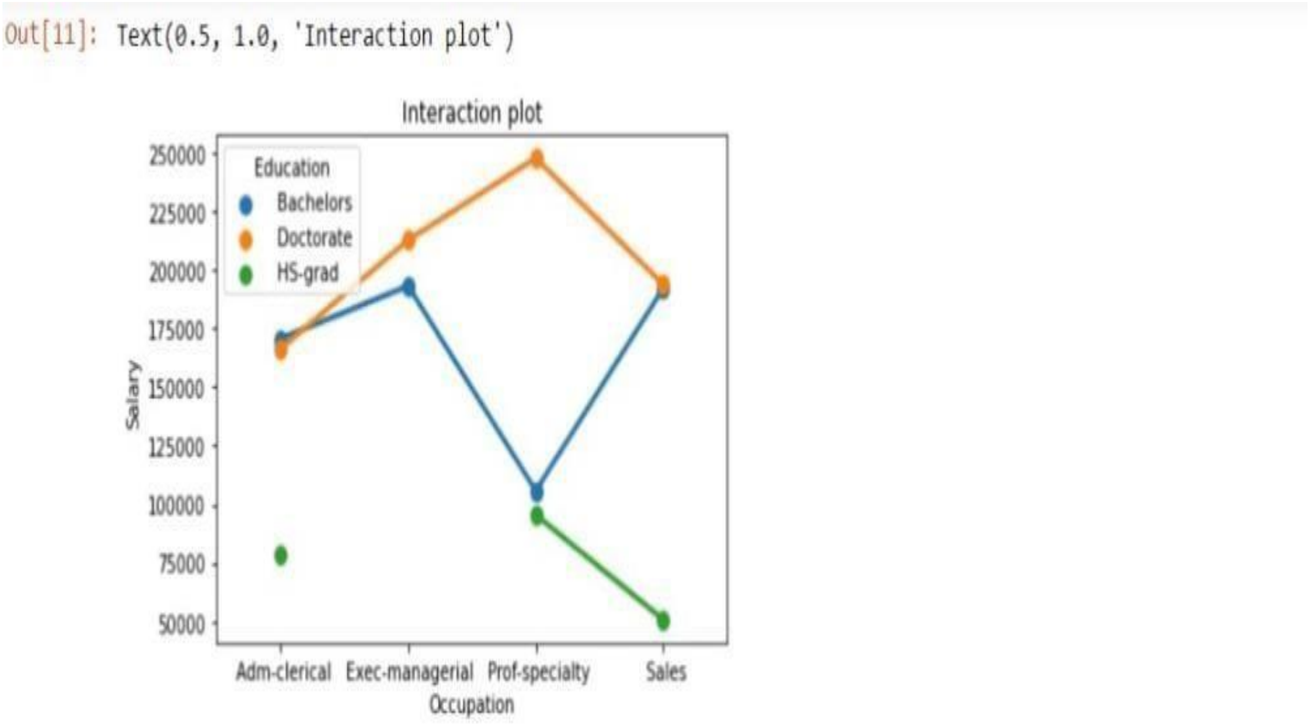
Out[9]:

	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



Out[10]:

	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



```
Out[11]: Text(0.5, 1.0, 'Interaction plot')
```

Adm-Clerical and Sales professionals with Bachelors and Doctorate degrees earn almost similar salary packages, where HS-grad has low salary packages in every Occupation as compare to Bachelors and Doctorate.

Qustion 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?

Solution:

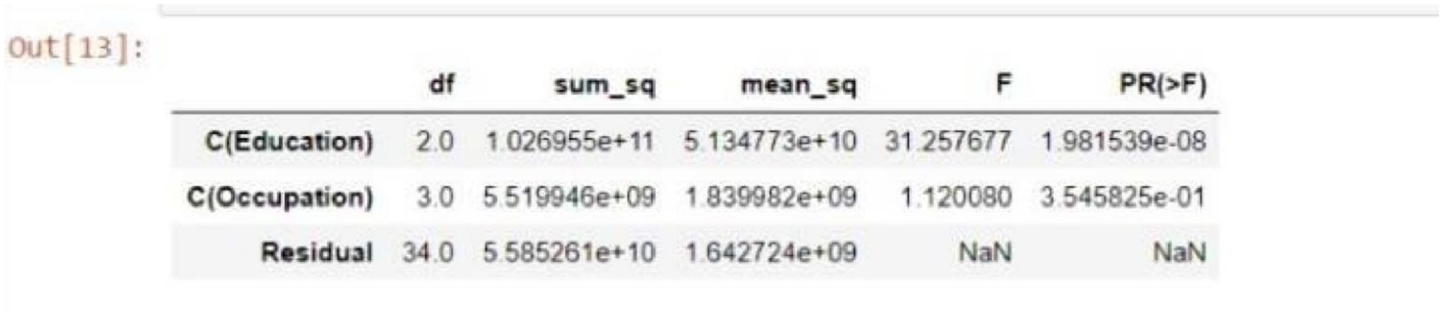
Two-way ANOVA based on the Education and Occupation:

Ho : 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

Considering both education and Occupation, Education is a significant factor as P value is <0.05, Whereas Occupation is not significant variable as P value of it is >0.05 .



Out[13]:

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	31.257677	1.981539e-08
C(Occupation)	3.0	5.519946e+09	1.839982e+09	1.120080	3.545825e-01
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

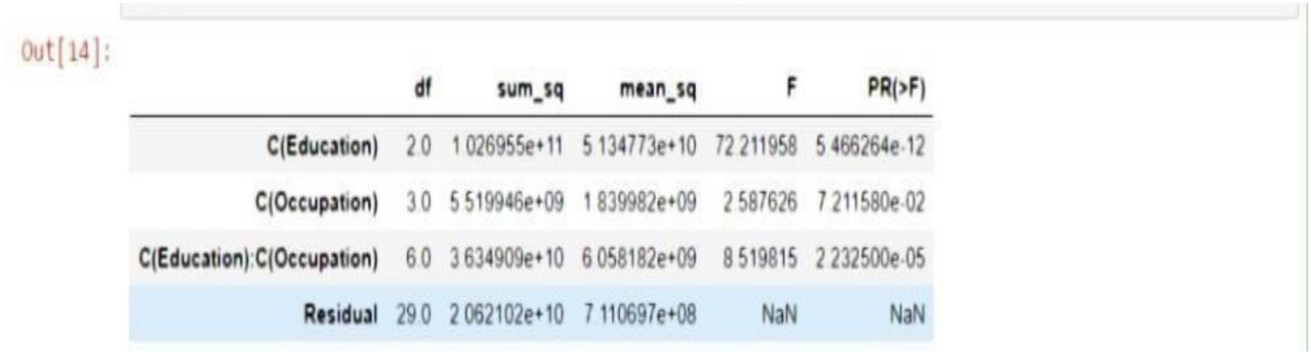
Question 1.7 Explain the business implications of performing ANOVA for this particular case study.

Solution:

By performing ANOVA on the given data set, we can conclude that Salary is dependent on Occupation. Along with the interaction of Education*Occupation, Considering both education and Occupation, Education is a significant factor as P value is <0.05, Whereas Occupation is not significant variable as P value of it is >0.05 . Adm-Clerical and Sales professionals with Bachelors and Doctorate degrees earn almost similar salary packages, where HS-grad has low salary packages in every Occupation as compare to Bachelors and Doctorate .

while performing one-way ANOVA for Education with respect to the variable 'Salary' we found that the P value is less than 0.05 ,Hence the null hypothesis is rejected that is Salary does not depend on Education. And While performing one-way ANOVA for Occupation with respect to the variable 'Salary'we found that the P value is greater than 0.05 ,Hence we fail to reject the null hypothesis,that is Salary depend on Education.

If we perform a two-way ANOVA based on the Education and Occupation with their interaction Education*Occupation with the variable 'Salary' we will get following result:



Out[14]:

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Problem 2:

Problem Statement:

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.

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

Solution:

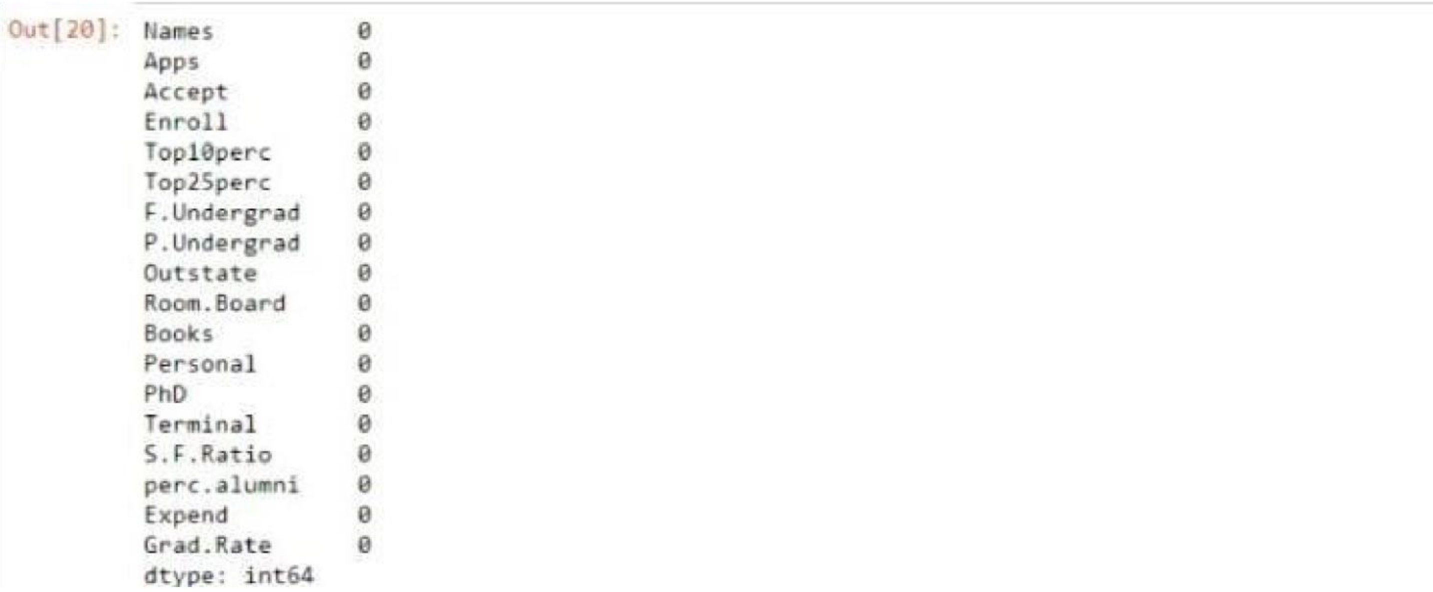
Firstly, after importing all the relevant libraries on Jupyter notebook, we load the data set.

Then, we perform EDA to extract and see patterns in the given data set.

The given data set has a shape of (777, 18). Also, we check the top 5 rows of the data set then Checking for missing values:

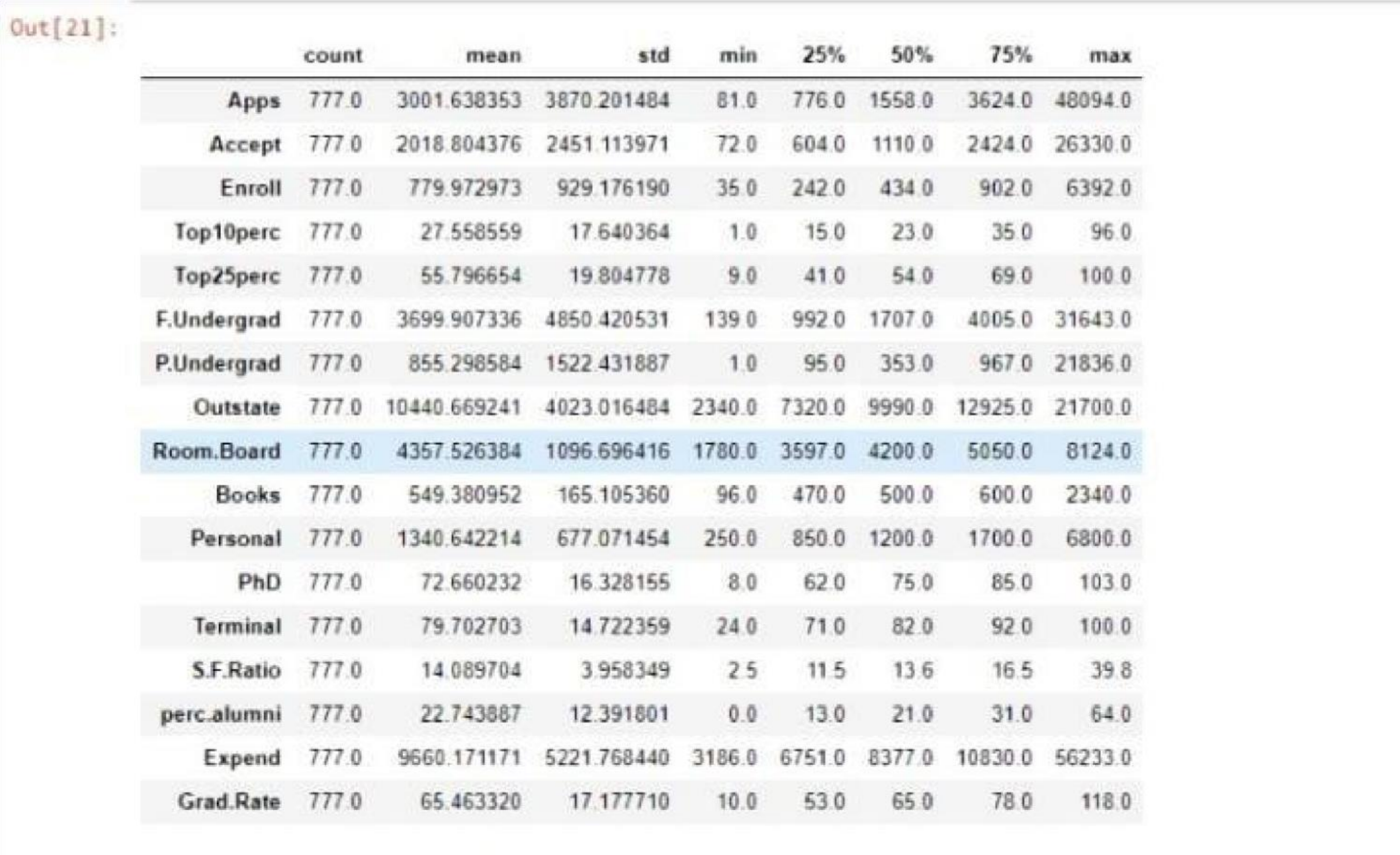
No missing values found.

Checking summary of data:



```
Out[20]: 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
         dtype: int64
```

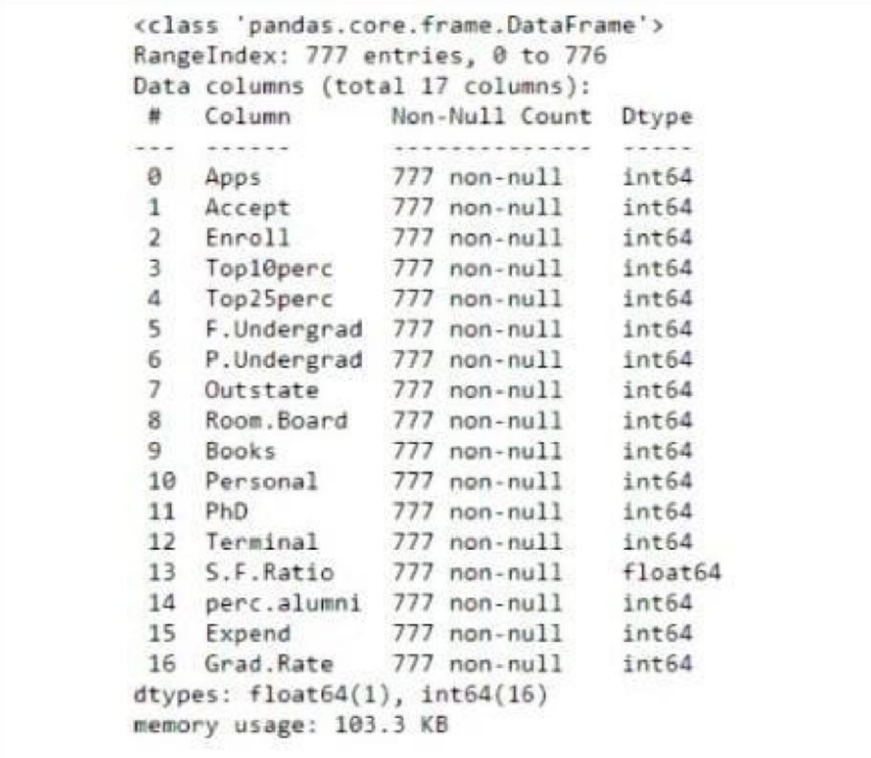
We need to perform univariate analysis which includes 17 numaric variables .
The analysis of all these variables includes:



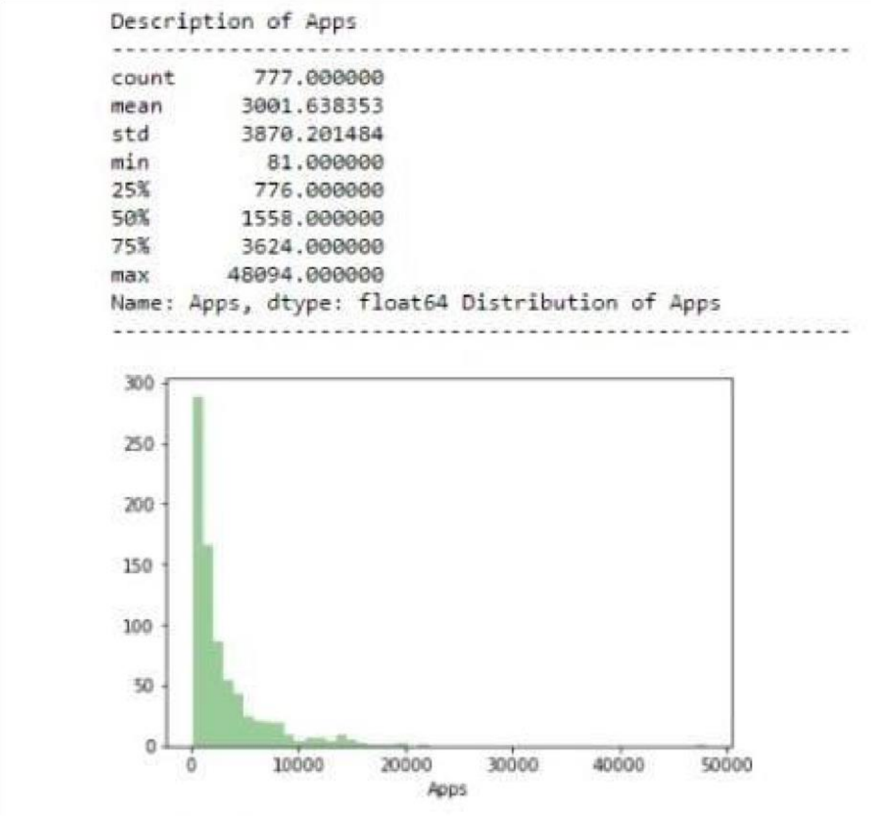
Out[21]:

	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

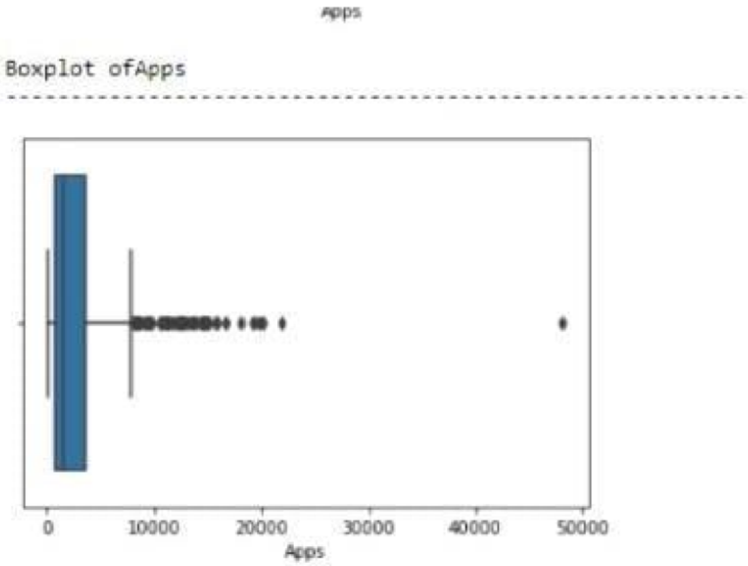
- Statistical description of the numeric variable
- Distribution of the column with histogram or distplot
- Boxplot representation of the column - 5 point summary and outliers if any Checking info of Numaric data frame only:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 17 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Apps         777 non-null    int64
 1   Accept       777 non-null    int64
 2   Enroll       777 non-null    int64
 3   Top10perc    777 non-null    int64
 4   Top25perc    777 non-null    int64
 5   F.Undergrad  777 non-null    int64
 6   P.Undergrad  777 non-null    int64
 7   Outstate     777 non-null    int64
 8   Room.Board   777 non-null    int64
 9   Books        777 non-null    int64
 10  Personal     777 non-null    int64
 11  PhD          777 non-null    int64
 12  Terminal     777 non-null    int64
 13  S.F.Ratio    777 non-null    float64
 14  perc.alumni  777 non-null    int64
 15  Expend       777 non-null    int64
 16  Grad.Rate    777 non-null    int64
dtypes: float64(1), int64(16)
memory usage: 103.3 KB
```

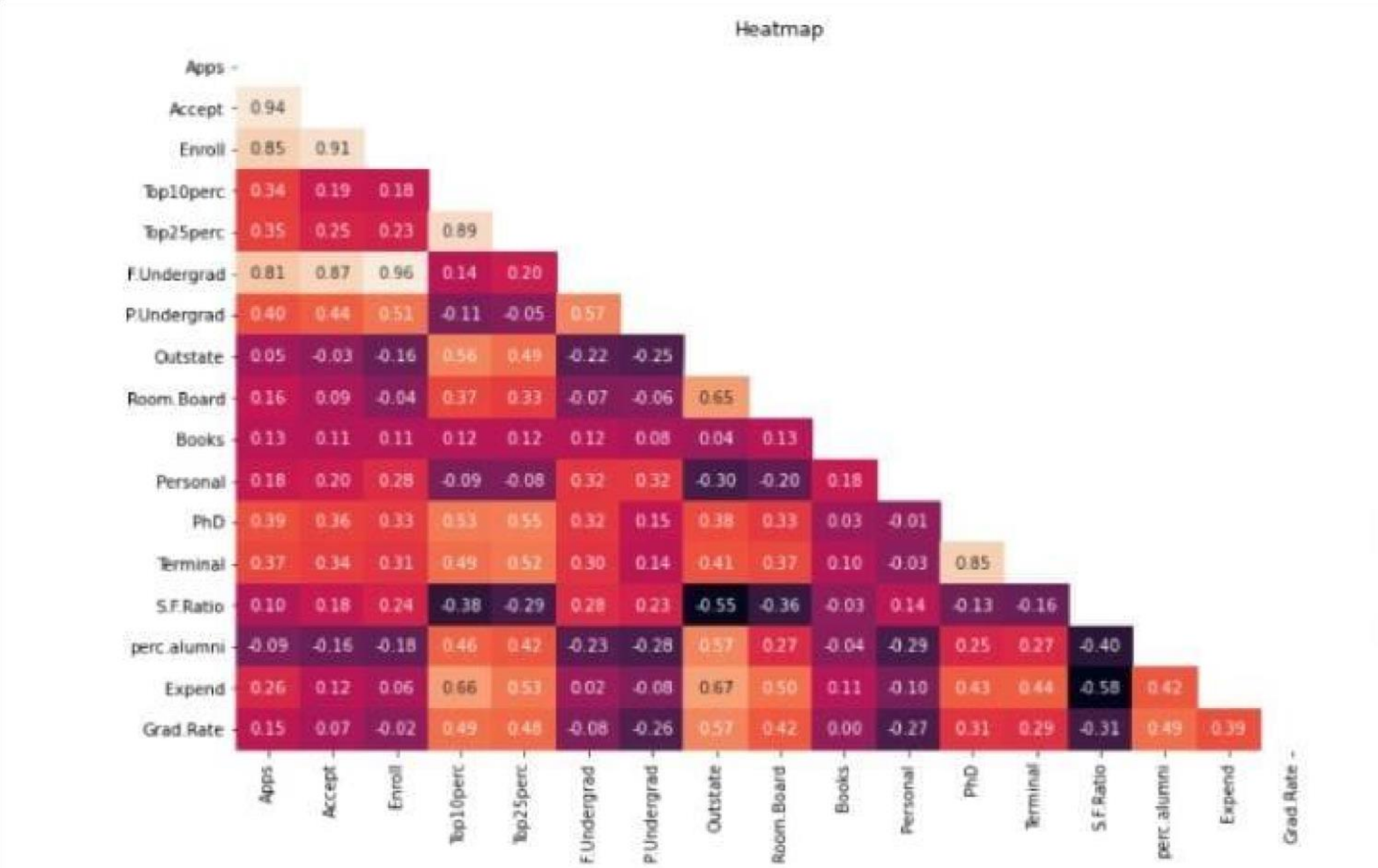


```
Description of Apps
----------------------------------------------------------
count      777.000000
mean      3001.638353
std       3870.201484
min         81.000000
25%        776.000000
50%       1558.000000
75%       3624.000000
max      48094.000000
Name: Apps, dtype: float64 Distribution of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



```
Boxplot ofApps
-------------------------------------------------------
```

The output displays, total 17*3 = 51 distinct charts/columns. Hence I have put the screenshot of only one variable i.e. apps.



Further, we perform multivariate analysis, using correlation function in which we get below output.

Insights:

- 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.
- Average percentage of faculties with Ph.D.’s is 72.66. The minimum S.F. ratio is around 2.5.
- There are considerable number of variables that are highly correlated.
- “Apps” has high correlation with “Accept”, and ”Enroll”.

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

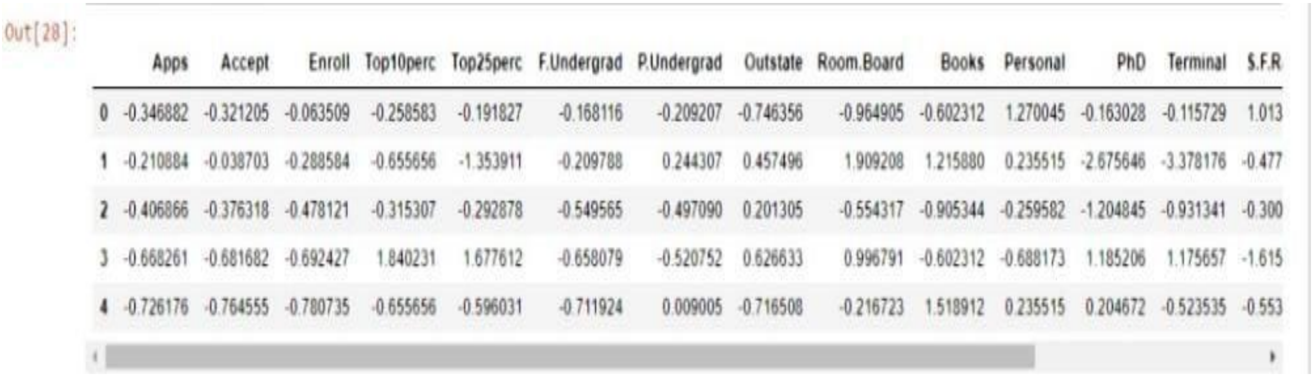
Solution:

Yes, it is necessary to perform scaling for PCA.

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.

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data.

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



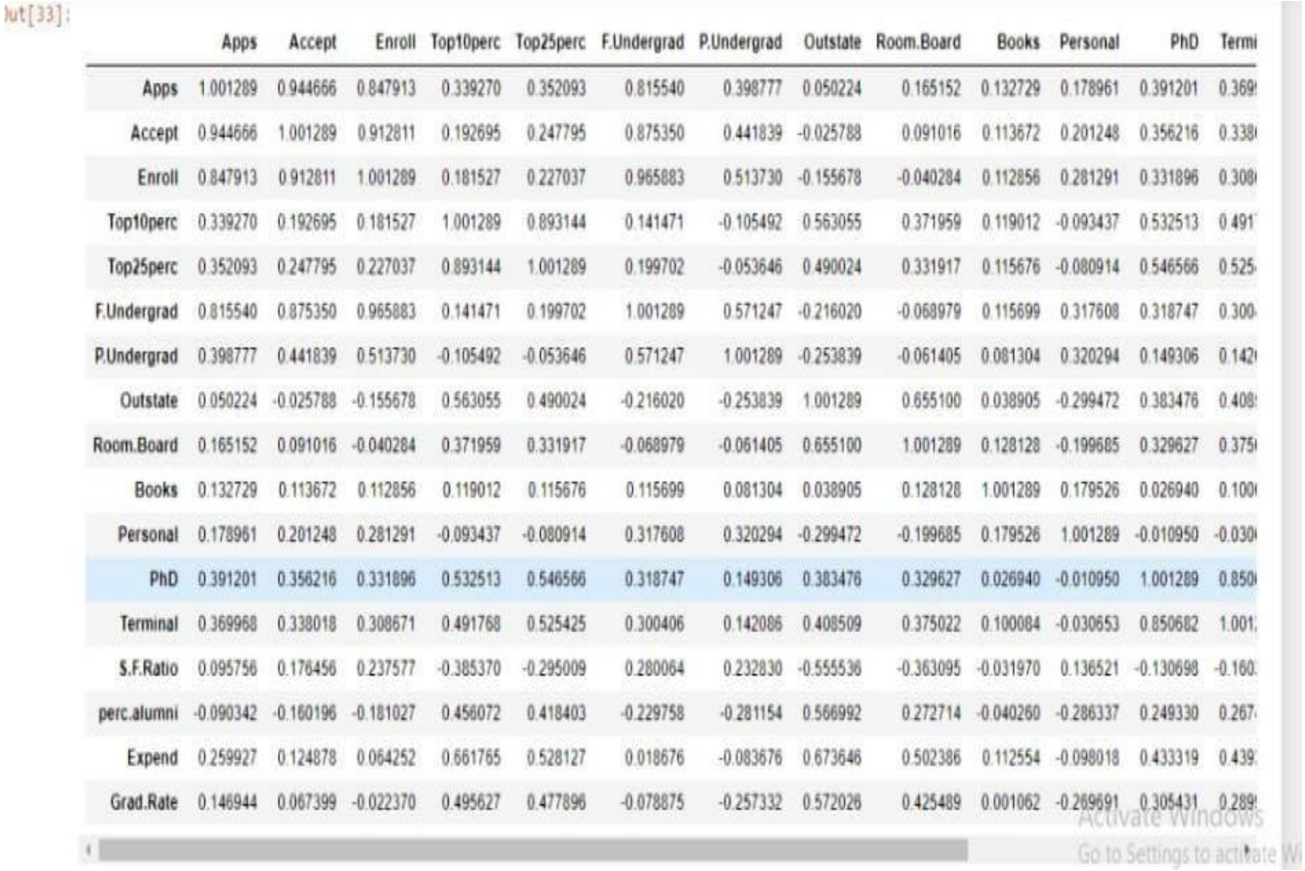
Out[28]:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.R
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	1.013
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	-0.477
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	-0.300
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	-1.615
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	-0.553

Question 2.3 Comment on the comparison between the covariance and the correlation matrices from this data.(on scaled data)

Solution:

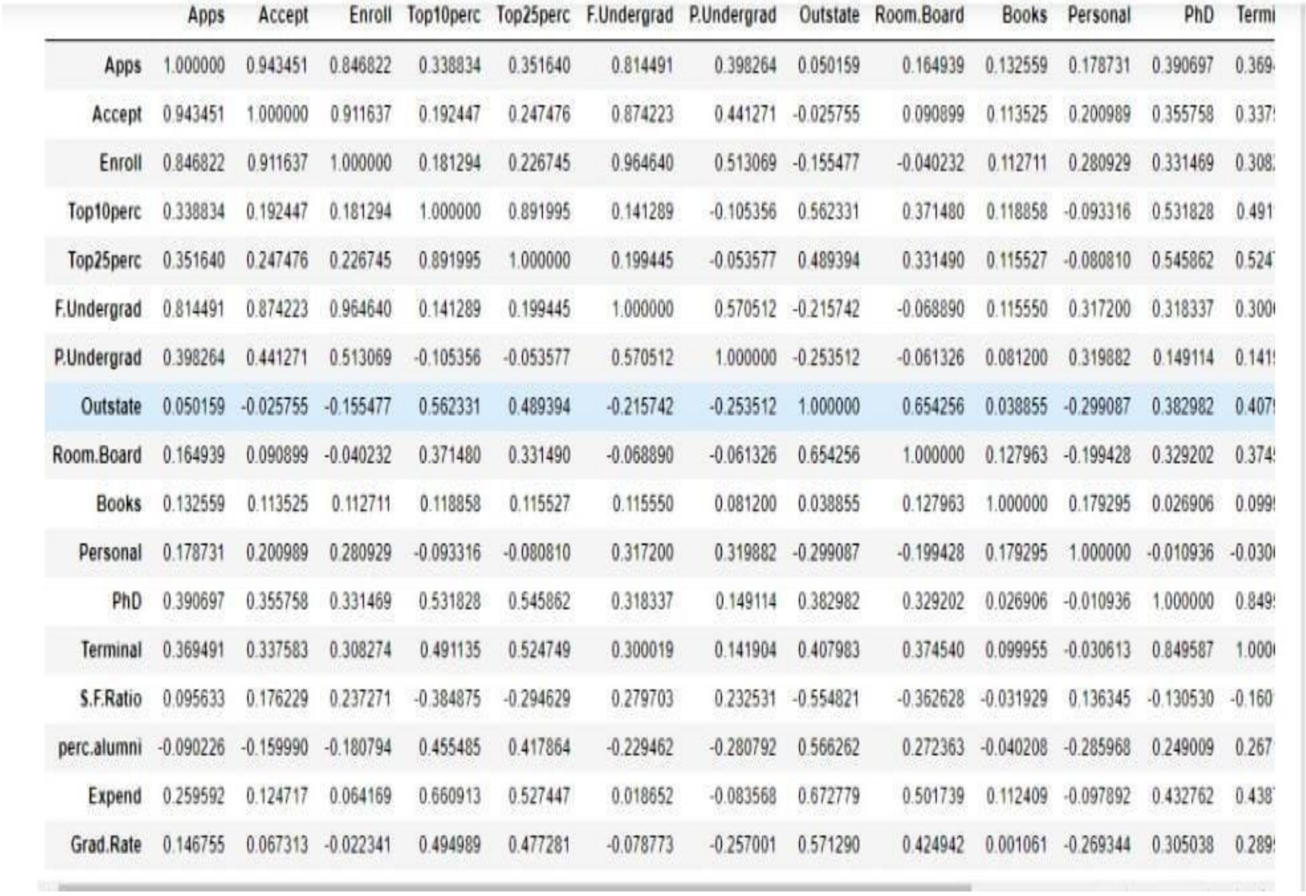
Correlation is a scaled version of covariance; note that the two parameters always have the same sign (positive, negative, or 0). When the sign is positive, the variables are said to be positively correlated; when the sign is negative, the variables are said to be negatively correlated; and when the sign is 0, the variables are said to be uncorrelated. covariance matrix:



)ut[33]:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termi
Apps	1.001289	0.944666	0.847913	0.339270	0.352093	0.815540	0.398777	0.050224	0.165152	0.132729	0.178961	0.391201	0.369
Accept	0.944666	1.001289	0.912811	0.192695	0.247795	0.875350	0.441839	-0.025788	0.091016	0.113672	0.201248	0.356216	0.338
Enroll	0.847913	0.912811	1.001289	0.181527	0.227037	0.965883	0.513730	-0.155678	-0.040284	0.112856	0.281291	0.331896	0.308
Top10perc	0.339270	0.192695	0.181527	1.001289	0.893144	0.141471	-0.105492	0.563055	0.371959	0.119012	-0.093437	0.532513	0.491
Top25perc	0.352093	0.247795	0.227037	0.893144	1.001289	0.199702	-0.053646	0.490024	0.331917	0.115676	-0.080914	0.546566	0.525
F.Undergrad	0.815540	0.875350	0.965883	0.141471	0.199702	1.001289	0.571247	-0.216020	-0.068979	0.115699	0.317608	0.318747	0.300
P.Undergrad	0.398777	0.441839	0.513730	-0.105492	-0.053646	0.571247	1.001289	-0.253839	-0.061405	0.081304	0.320294	0.149306	0.142
Outstate	0.050224	-0.025788	-0.155678	0.563055	0.490024	-0.216020	-0.253839	1.001289	0.655100	0.038905	-0.299472	0.383476	0.408
Room.Board	0.165152	0.091016	-0.040284	0.371959	0.331917	-0.068979	-0.061405	0.655100	1.001289	0.128128	-0.199685	0.329627	0.375
Books	0.132729	0.113672	0.112856	0.119012	0.115676	0.115699	0.081304	0.038905	0.128128	1.001289	0.179526	0.026940	0.100
Personal	0.178961	0.201248	0.281291	-0.093437	-0.080914	0.317608	0.320294	-0.299472	-0.199685	0.179526	1.001289	-0.010950	-0.030
PhD	0.391201	0.356216	0.331896	0.532513	0.546566	0.318747	0.149306	0.383476	0.329627	0.026940	-0.010950	1.001289	0.850
Terminal	0.369968	0.338018	0.308671	0.491768	0.525425	0.300406	0.142086	0.408509	0.375022	0.100084	-0.030653	0.850682	1.001
S.F.Ratio	0.095756	0.176456	0.237577	-0.385370	-0.295009	0.280064	0.232830	-0.555536	-0.363095	-0.031970	0.136521	-0.130698	-0.160
perc.alumni	-0.090342	-0.160196	-0.181027	0.456072	0.418403	-0.229758	-0.281154	0.566992	0.272714	-0.040260	-0.286337	0.249330	0.267
Expend	0.259927	0.124878	0.064252	0.661765	0.528127	0.018676	-0.083676	0.673646	0.502386	0.112554	-0.098018	0.433319	0.439
Grad.Rate	0.146944	0.067399	-0.022370	0.495627	0.477896	-0.078875	-0.257332	0.572026	0.425489	0.001062	-0.269691	0.305431	0.289

Corelation matrix:



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termi
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	0.369
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	0.337
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	0.308
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	0.491
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	0.524
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	0.300
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	0.141
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	0.407
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	0.374
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	0.099
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	-0.030
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	0.849
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	1.000
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	-0.160
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	0.267
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	0.438
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	0.289

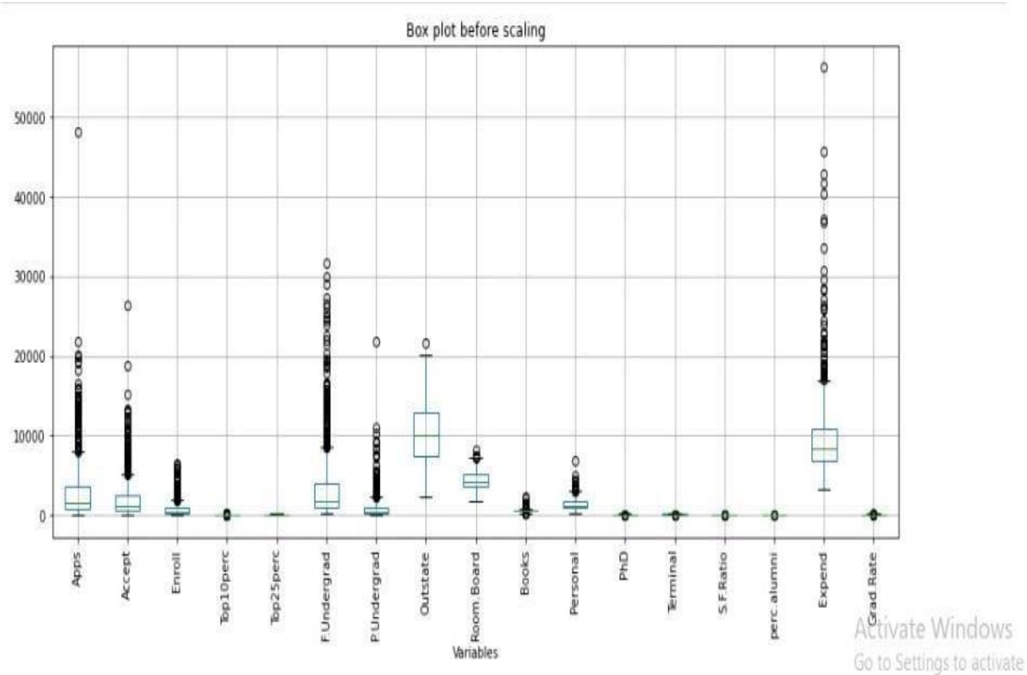
In simple sense correlation, measures both the strength and direction of the linear relationship between two variables.

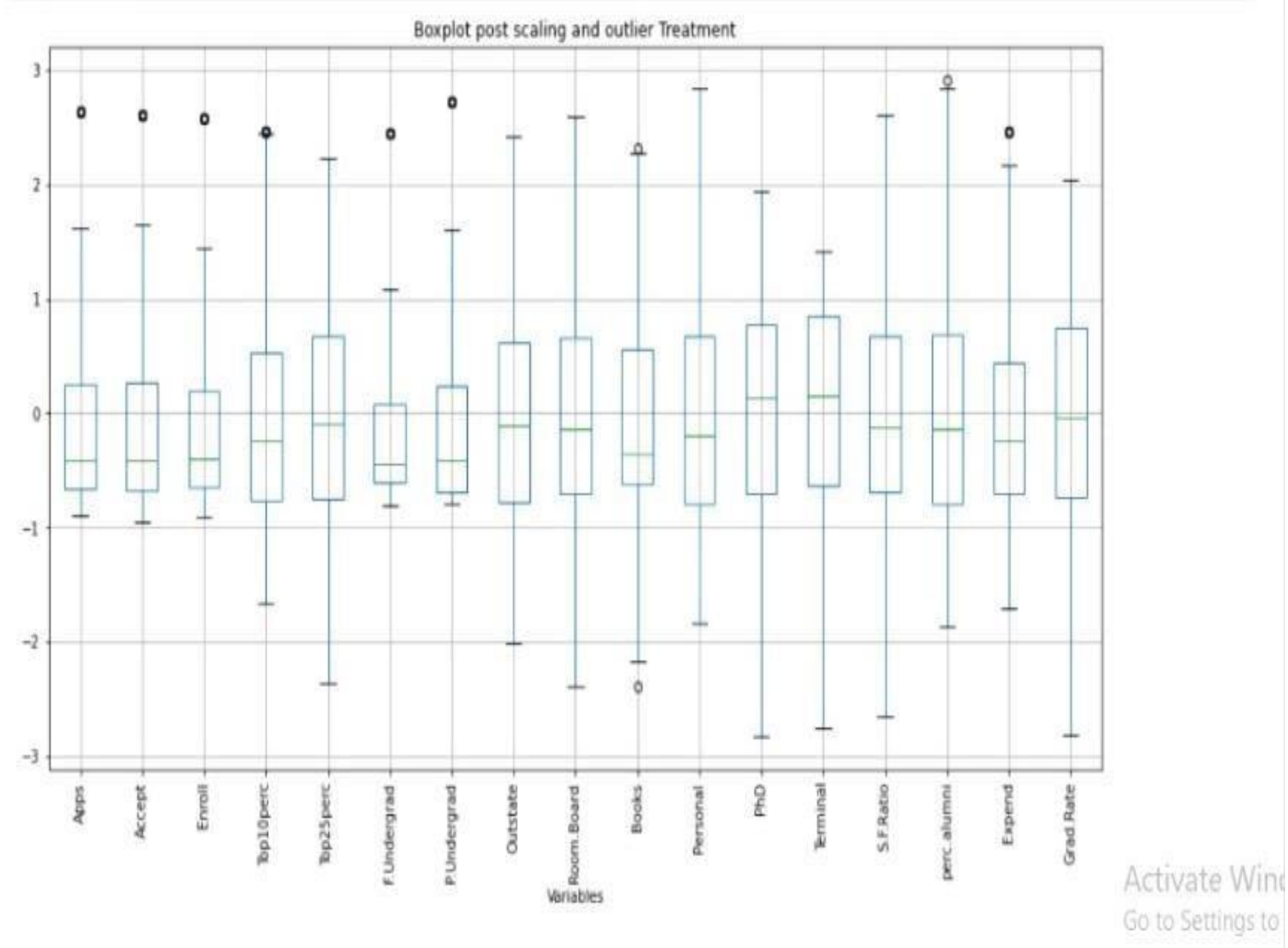
Covariance is a measure used to determine how much two variables change in tandem. It indicates the direction of the linear relationship between variables.

Question 2.4 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:





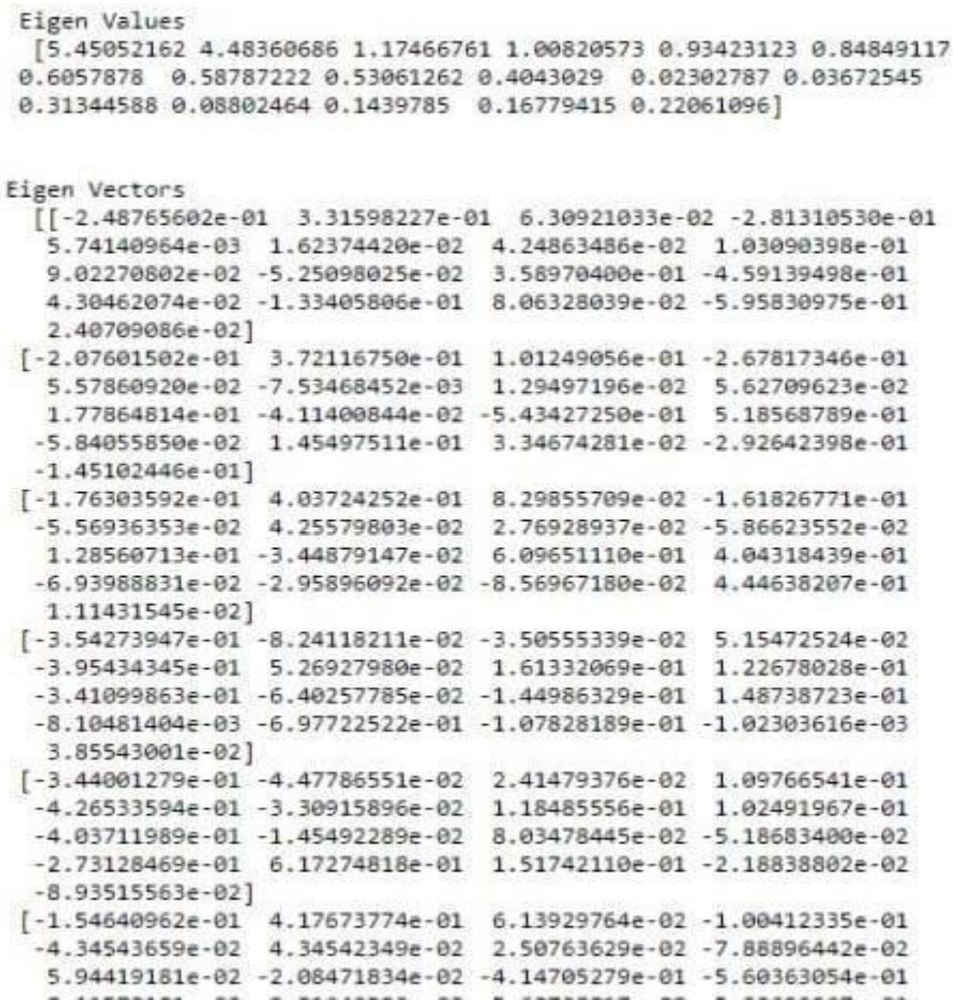
Insights:

- 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

Question 2.5 Extract the eigenvalues and eigenvectors ?

Solution: We can extract eigenvalues and eigenvectors using covariance matrix.

We have already found the Covariance matrix before Extracting eigenvalues and eigenvectors. The below snapshot represent extracted eigenvalues and eigenvectors



```
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]


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
```

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

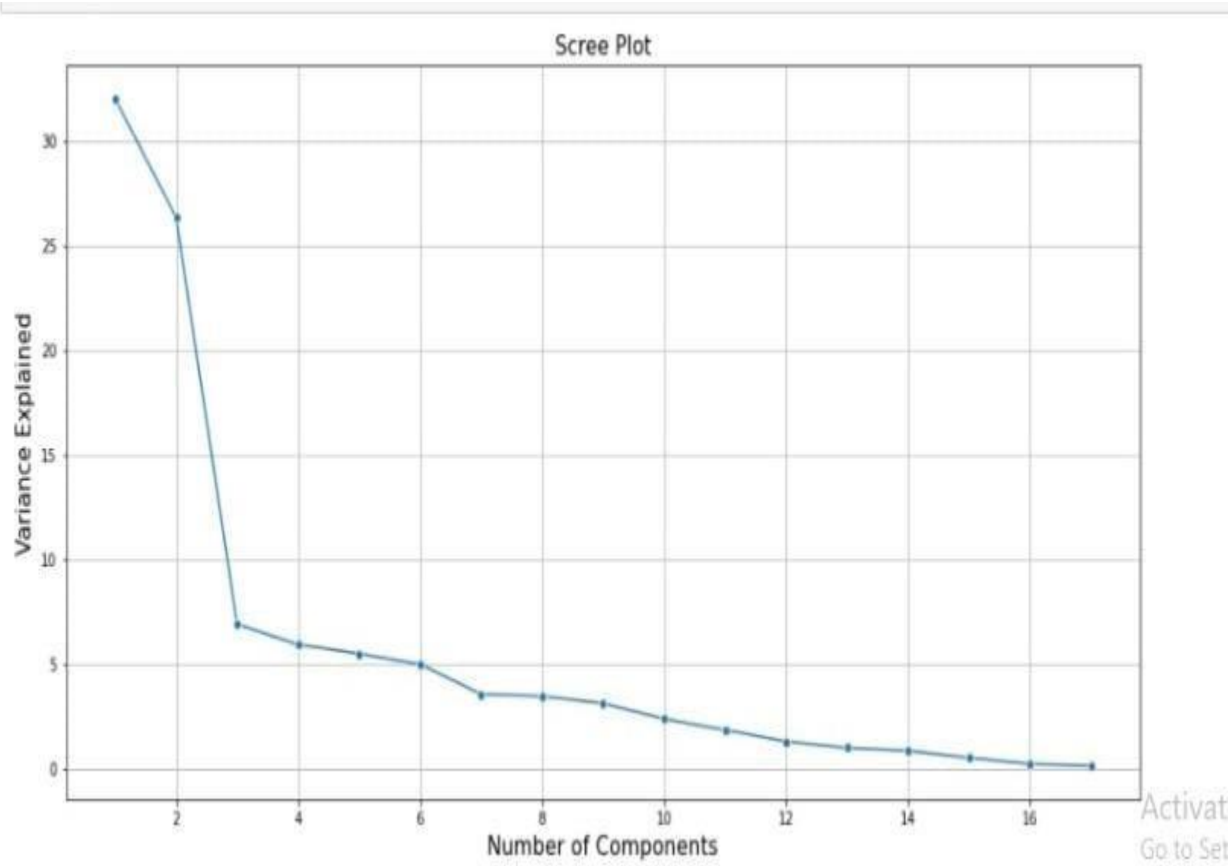
Solution:

For performing PCA , we need to follow below steps:

- Step 1: Generate the covariance matrix
- Step 2: Get eigenvalues and eiigenvectors
- 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.

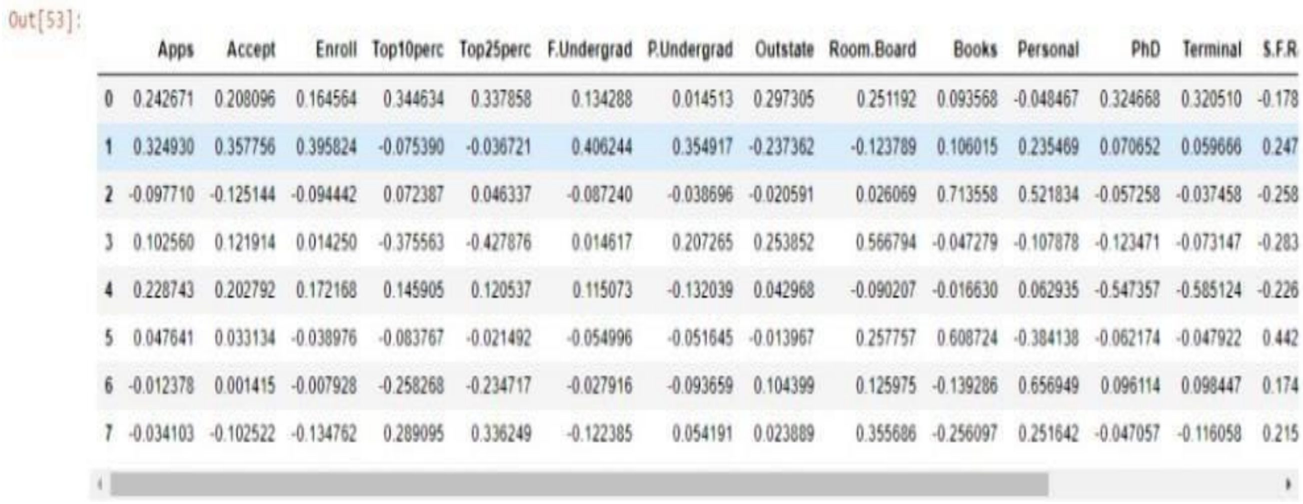
Covariance matrix, Eigenvalues and Eigenvectors are already generated above , so moving straight to the next step:

Step 3: Plotting Scree Plot to identify the number of components to be built.



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

Below is the representative screenshot of df_pca_loading in which we had exported the principal component scores into a data frame.

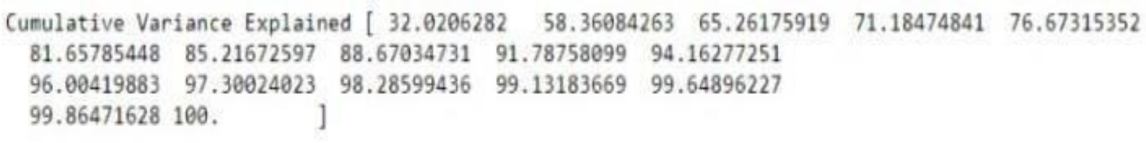

Out[53]:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.R
0	0.242671	0.208096	0.164564	0.344634	0.337858	0.134288	0.014513	0.297305	0.251192	0.093568	-0.048467	0.324668	0.320510	-0.178
1	0.324930	0.357756	0.395824	-0.075390	-0.036721	0.406244	0.354917	-0.237362	-0.123789	0.106015	0.235469	0.070652	0.059666	0.247
2	-0.097710	-0.125144	-0.094442	0.072387	0.046337	-0.087240	-0.038696	-0.020591	0.026069	0.713558	0.521834	-0.057258	-0.037458	-0.258
3	0.102560	0.121914	0.014250	-0.375563	-0.427876	0.014617	0.207265	0.253852	0.566794	-0.047279	-0.107878	-0.123471	-0.073147	-0.283
4	0.228743	0.202792	0.172168	0.145905	0.120537	0.115073	-0.132039	0.042968	-0.090207	-0.016630	0.062935	-0.547357	-0.585124	-0.226
5	0.047641	0.033134	-0.038976	-0.083767	-0.021492	-0.054996	-0.051645	-0.013967	0.257757	0.608724	-0.384138	-0.062174	-0.047922	0.442
6	-0.012378	0.001415	-0.007928	-0.258268	-0.234717	-0.027916	-0.093659	0.104399	0.125975	-0.139286	0.656949	0.096114	0.098447	0.174
7	-0.034103	-0.102522	-0.134762	0.289095	0.336249	-0.122385	0.054191	0.023889	0.355686	-0.256097	0.251642	-0.047057	-0.116058	0.215

Question 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?

Solution:

Cumulative Distribution of Eigen values Cumulative
variance explained:



```
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.        ]
```

We can see that around 8 principal components explained over 90% of the variance. Thus, the optimum number of principal components can be 8. 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.

Question 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]

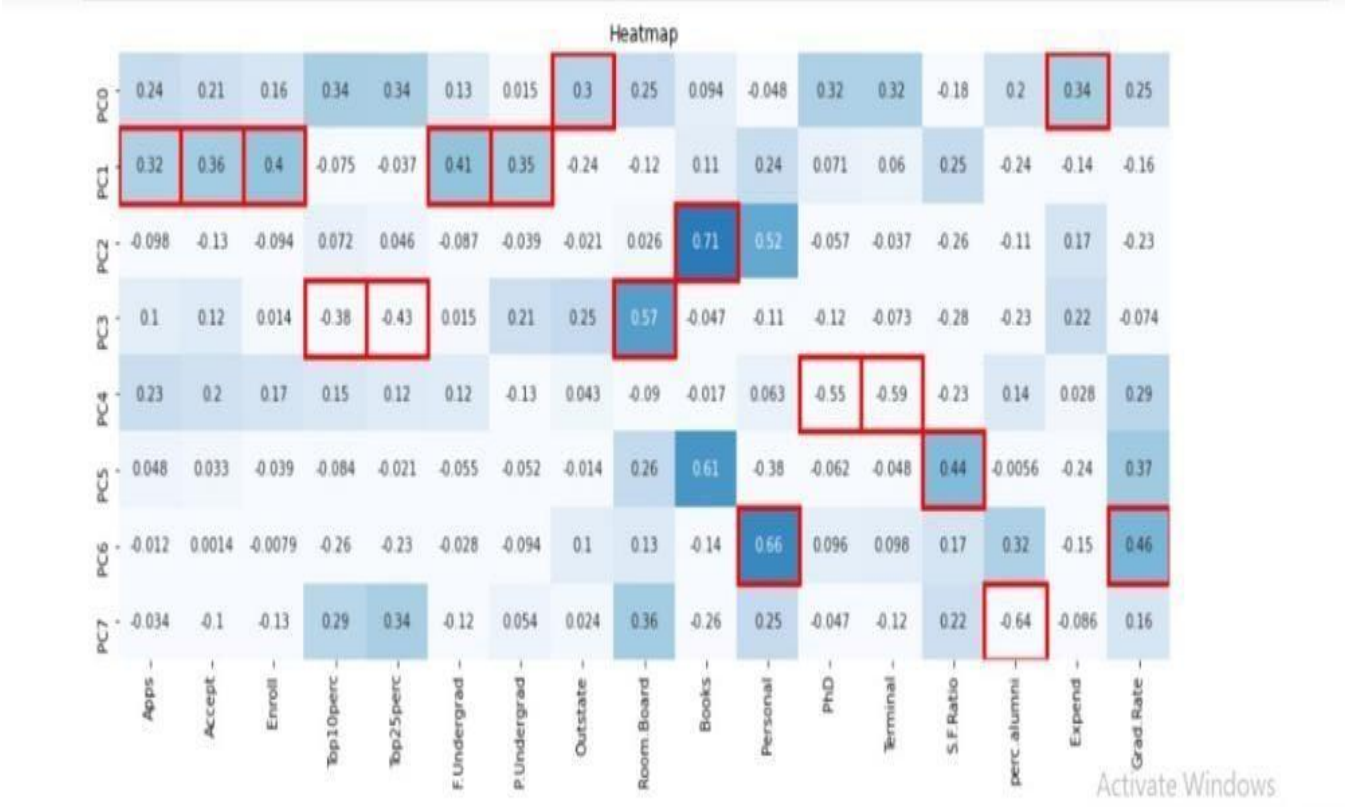
Solution:

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 more than 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 more than 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 highdimensional space (with 17 variables) and converting it to a lower dimensional space without (theoretically) losing much of the explanatory power



Following are the interpretations from the obtained PCs

- PC0: Explains No. of students for whom the particular college or university is Out-of-state tuition and instructional expenditure per student
- PC1: Represents the highly correlated variables such as Apps, Enroll and Accept
- PC2: Highlights the estimated cost of books for a student
- PC3: Explains percentage of new students from top 10% and 25% of higher secondary class including cost of room and board
- PC4: Represents % of faculties with Ph.D.’s and terminal degree
- PC5: Details about student/faculty ratio
- PC6: Highlights estimated personal spending for a student and graduation rate PC7: Explains number of alumni who donate