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.

Question 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':

$H0$: The mean salary is same across all the three categories of education.

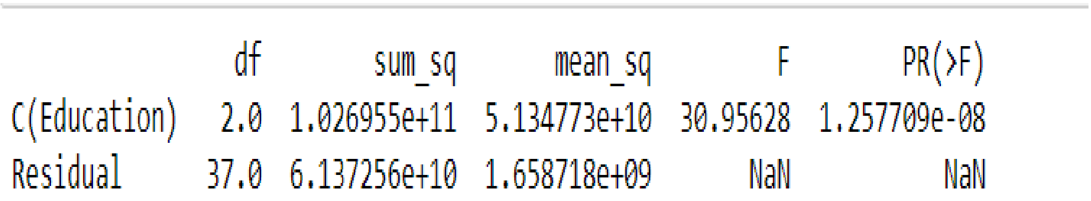
Ha: The mean salary is different in at least one category of education.

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

$H0$: The mean salary in all the four categories of Occupation is same.

Ha: The mean salary is different in at least one category of education.

Question 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.

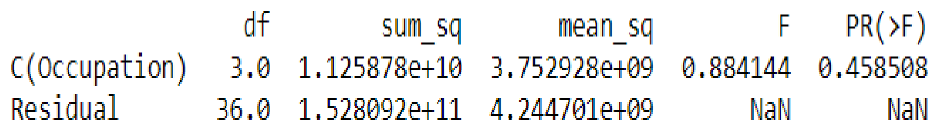


```
                df        sum_sq       mean_sq         F        PR(>F)
C(Education)   2.0  1.026955e+11  5.134773e+10  30.95628  1.257709e-08
Residual      37.0  6.137256e+10  1.658718e+09       NaN           NaN
```

Solution:
We performed one-way ANOVA in python and got the above output. Since pvalue i.e., 1.257709e-08 <alpha (0.05), null hypothesis is rejected. It means there is significant difference in mean salaries in a t least one category of education.

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

Solution:



```
                  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
```

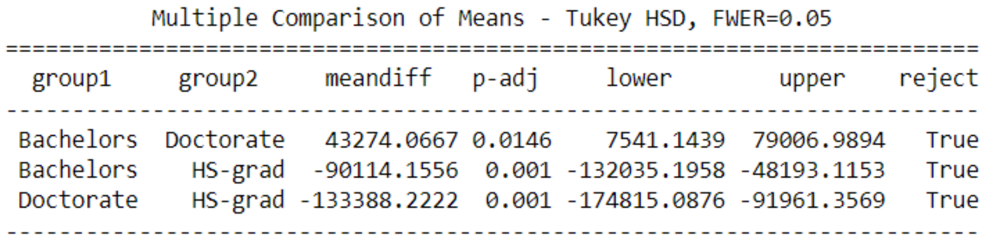
From the above table, we can see that pvalue >alpha (0.05). It means there is no significant difference in mean salaries across all the four categories of occupation.

Question 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).

Solution: Tukey’s honestly significant difference (HSD) test performs pairwise comparison of means for a set of samples. It is a post-hoc statistical test used to determine whether the means of two sets of data are statistically different from each other. This test is based on the studentized range distribution and is performed after an ANOVA test has indicated a significant difference in means of three or more sets of data.

Null Hypothesis: Means are equal.

Alternate Hypothesis: Means are not equal.



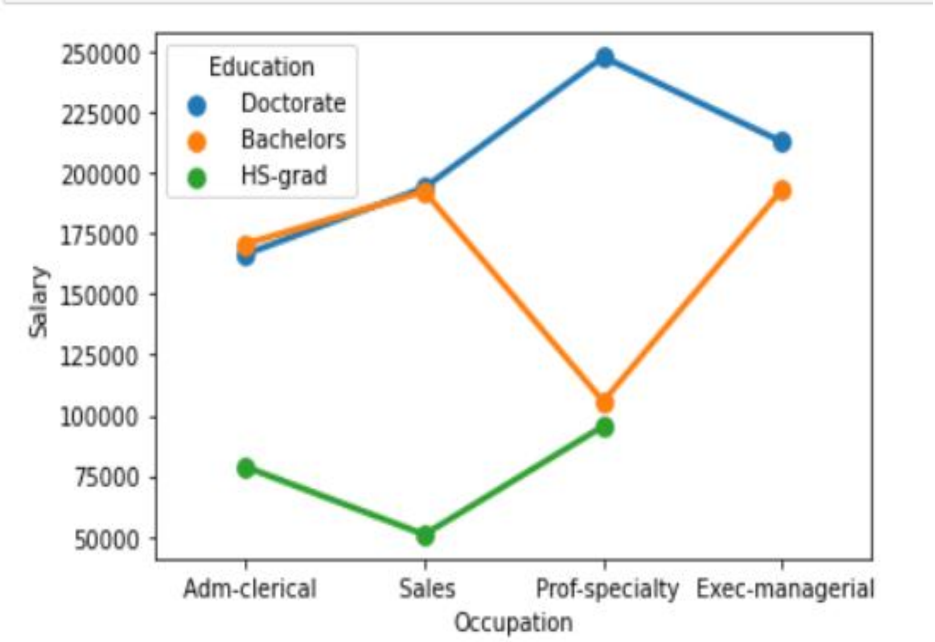
```
           Multiple Comparison of Means - Tukey HSD, FWER=0.05
==========================================================================
  group1     group2     meandiff   p-adj     lower        upper     reject
--------------------------------------------------------------------------
 Bachelors  Doctorate   43274.0667 0.0146    7541.1439  79006.9894   True
 Bachelors    HS-grad  -90114.1556  0.001 -132035.1958 -48193.1153   True
 Doctorate    HS-grad -133388.2222  0.001 -174815.0876 -91961.3569   True
--------------------------------------------------------------------------
```

If reject is "True”, it means we can reject the null hypothesis. And their means are significantly different.

If reject is "False”, it means we fail to reject the null hypothesis. And their means are almost equal.

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

Solution:



HS-Grad in Sales has minimum salary while doctorate in Prof-speciality has maximum salary.

Adm-Clerical and Sales professionals with bachelor's and Doctorate degrees earn almost similar salary packages.

HS-grad has lesser salary in every Occupation compared to bachelor's and Doctorate.

HS-grad people do not reach to the exec-Managerial position and they hold all the other positions (Adm-clerical,Sales,Prof-speciality)

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?

Solution. 1) Null and alternate hypothesis for conducting one-way ANOVA for 'Education' with respect to 'Salary':

$H0$: The mean salary is same across all the three categories of education.

Ha: The mean salary is different in at least one category of education.

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

$H0$: The mean salary in all the four categories of Occupation is same.

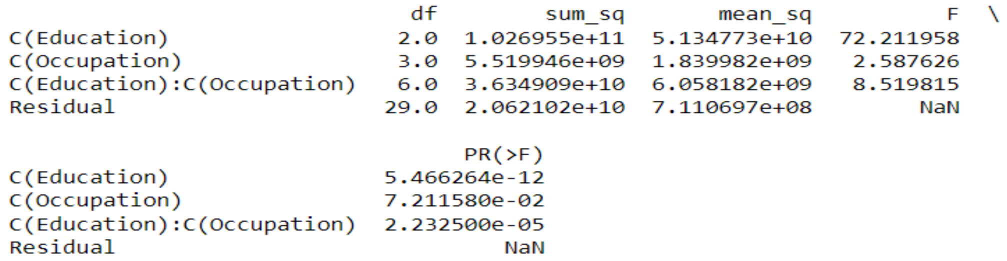
Ha: The mean salary is different in at least one category of education.

3)Interaction effect

$H0$: There is no interaction effect between Education and Occupation.

Ha: There is an interaction effect between Education and Occupation on the mean salary.

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
```

Conclusions:

1) Education is a significant factor as pvalue<0.05. We reject the null hypothesis. It means the mean salary is different in at least one category of education.

2)Occupation is not significant variable as P value >0.05. We fail to reject the null hypothesis. The mean salary in all the four categories of Occupation is same.

3) There is an interaction effect between Education and Occupation on the mean salary as pvalue<0.05.

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

Solution. From the ANOVA and the interaction plot, we can say that education combined with occupation results in higher salary. People with education as doctorate get maximum salary whereas people with education as HS-Grad earn the minimum. Grad in Sales has minimum salary while doctorate in Prof-speciality has maximum salary. By performing ANOVA, we can conclude that Education is a significant factor as pvalue<0.05. It means the mean salary is different in at least one category of education. Occupation is not significant variable as P value >0.05. We fail to reject the null hypothesis. The mean salary in all the four categories of Occupation is same. There is an interaction effect between Education and Occupation on the mean salary as pvalue<0.05.

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?

Solution:

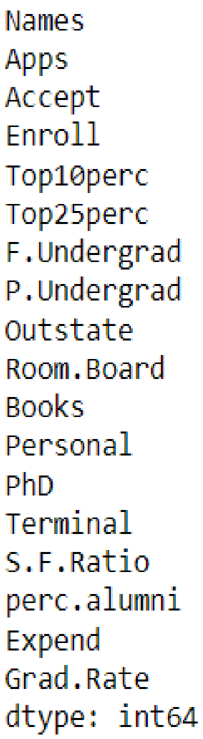
INFERENCE OF THE DATASET GIVEN:

The dataset has 777 rows and 18 columns.

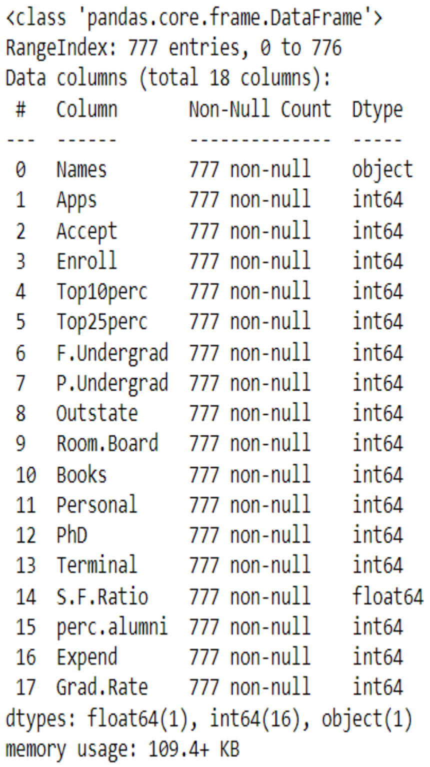
The “Names” column is categorical, “S.F. Ratio” is float while all other columns are of integer datatype.

The dataset has no missing/null value.

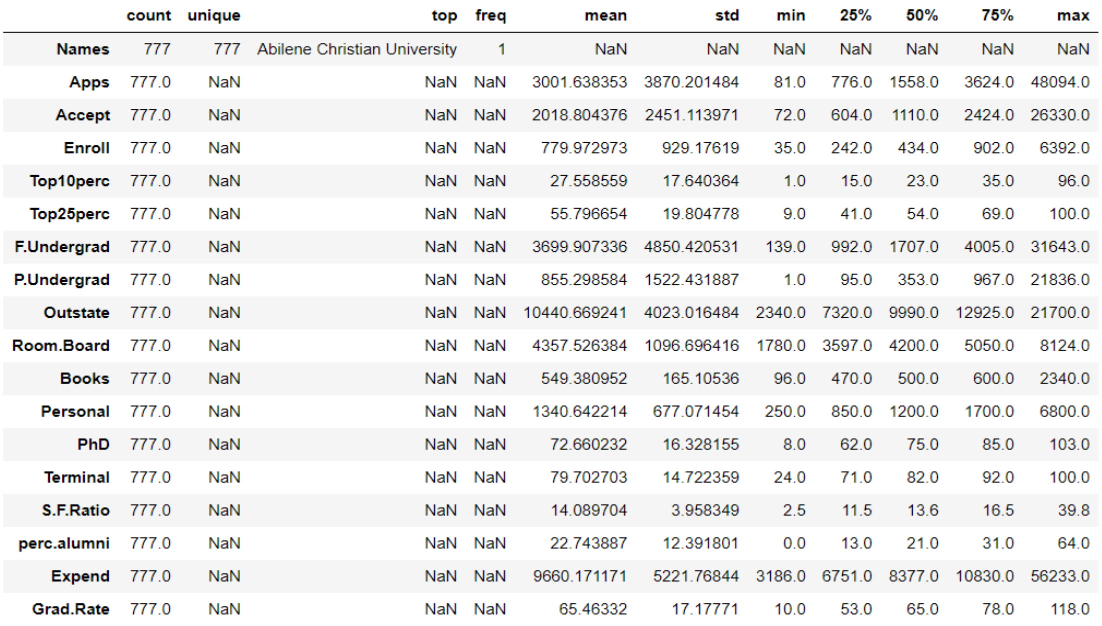
There are no duplicates in the dataset.



```
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
```



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

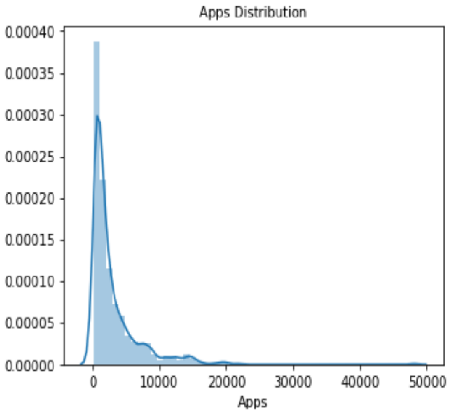


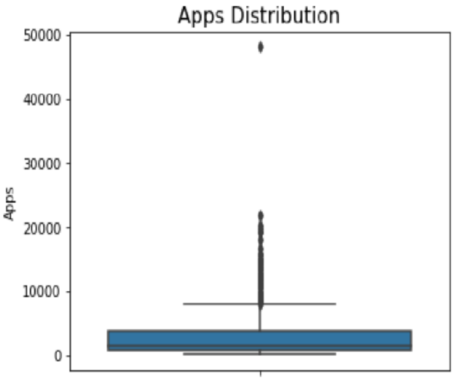
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Names	777	777	Abilene Christian University	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Apps	777.0	NaN	NaN	NaN	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	NaN	NaN	NaN	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	NaN	NaN	NaN	779.972973	929.17619	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	NaN	NaN	NaN	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	NaN	NaN	NaN	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	NaN	NaN	NaN	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	NaN	NaN	NaN	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	NaN	NaN	NaN	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	NaN	NaN	NaN	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	NaN	NaN	NaN	549.380952	165.10536	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	NaN	NaN	NaN	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	NaN	NaN	NaN	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	NaN	NaN	NaN	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	NaN	NaN	NaN	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	NaN	NaN	NaN	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	NaN	NaN	NaN	9660.171171	5221.76844	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	NaN	NaN	NaN	65.46332	17.17771	10.0	53.0	65.0	78.0	118.0

UNIVARIATE ANALYSIS

The term **univariate analysis** refers to the analysis of one variable. The purpose of the univariate analysis is to understand the distribution of values for a single variable. For univariate analysis, we are using a distplot and a boxplot. Distplot is a histogram with a line on it. Distplot is used for single variable distribution. A Box plot is used to visualize the descriptive statistics of a variable. It is used to detect outliers. It represents the five-point summary.

APPS:

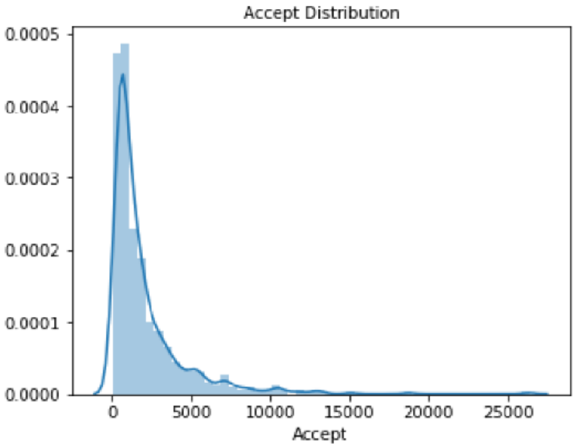


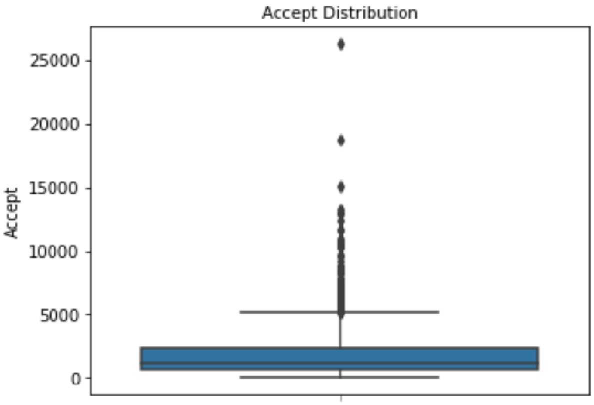


The distribution of "APPS" seems to be skewed. From the boxplot, we clearly see there are outliers in the dataset. The college offers application in the range 3000 to 5000. 5000 is the median value. The maximum application seems to be around 50000.

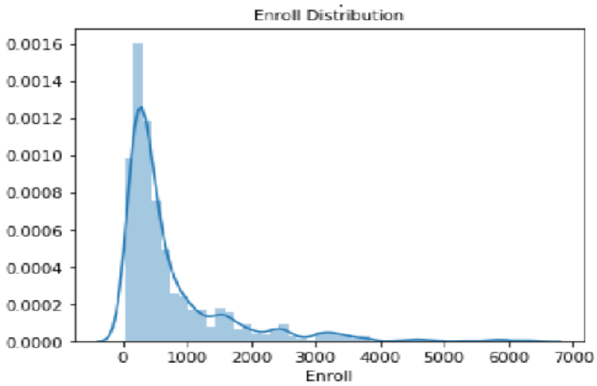
ACCEPT:

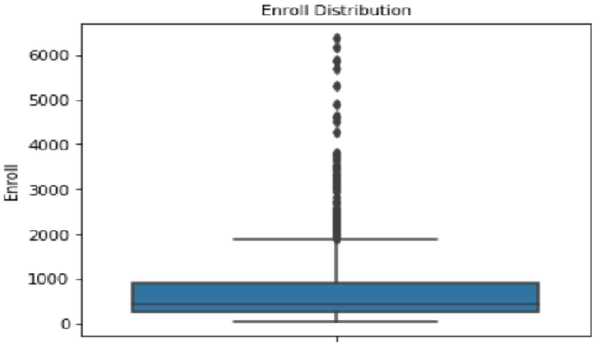
From the distplot, we see that most of the applications accepted are in the range 80 to 1500.It seems to be positively skewed and contain outliers.





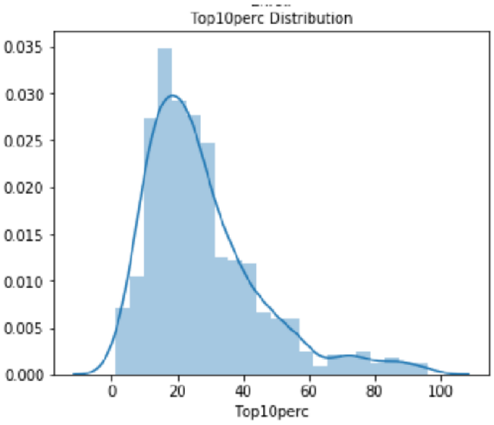
ENROLL:

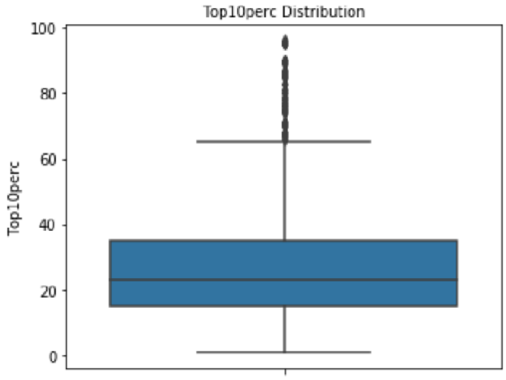




Most of the colleges enrolled students are in the range 200 to 500. The distribution seems to be positively skewed and the boxplot seems to contain lots of outliers.

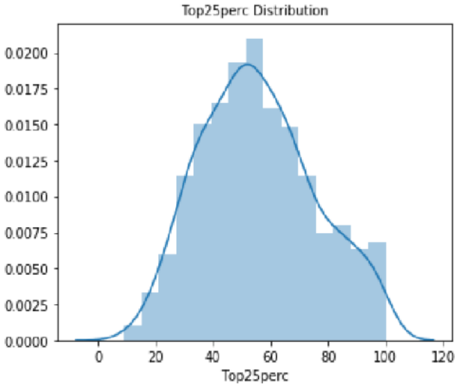
TOP 10PERC:

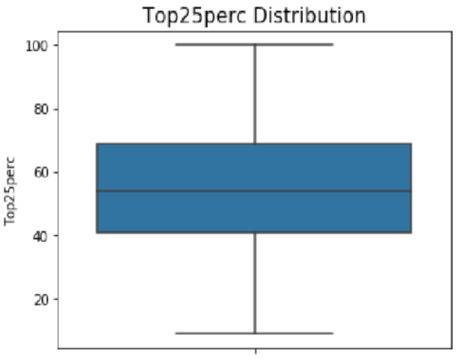




The distribution seems to be positively skewed. Here also the boxplot has a lot of outliers. The median value seems to be around 25.

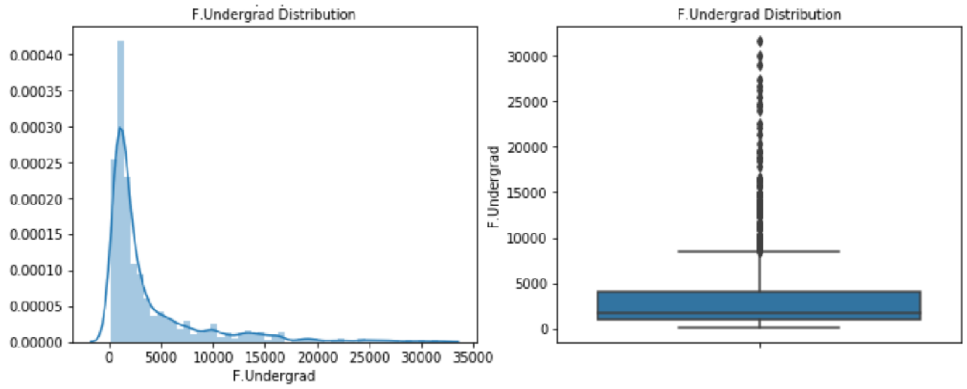
TOP25 PERC:





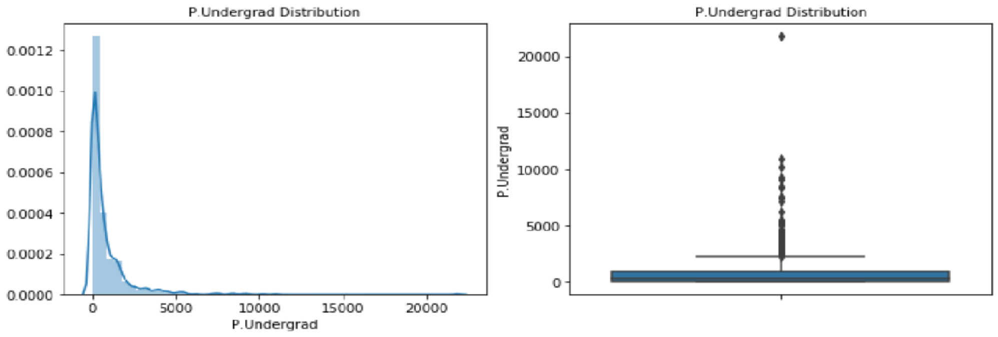
The boxplot has no outliers. Most of the students are from the top 25 % of the class. The distribution seems to be almost normal here.

FULL-TIME UNDERGRADUATE:



The distribution is positively skewed. Most of the full-time graduates studying are in the range 3000 to 5000. The boxplot contains many outliers.

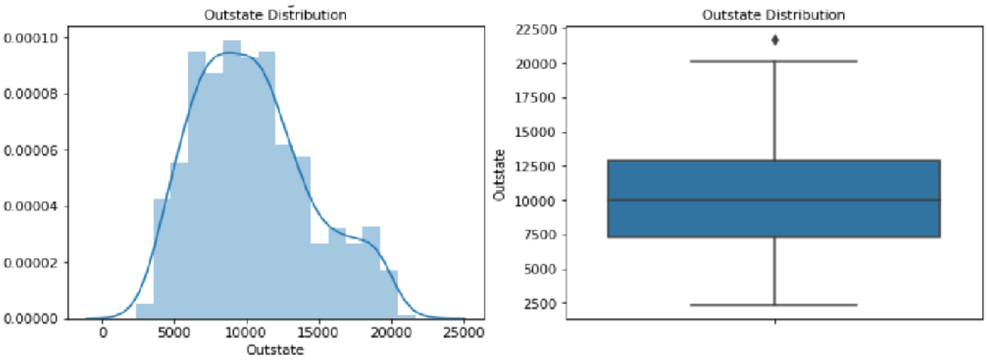
PART-TIME UNDERGRADUATE:



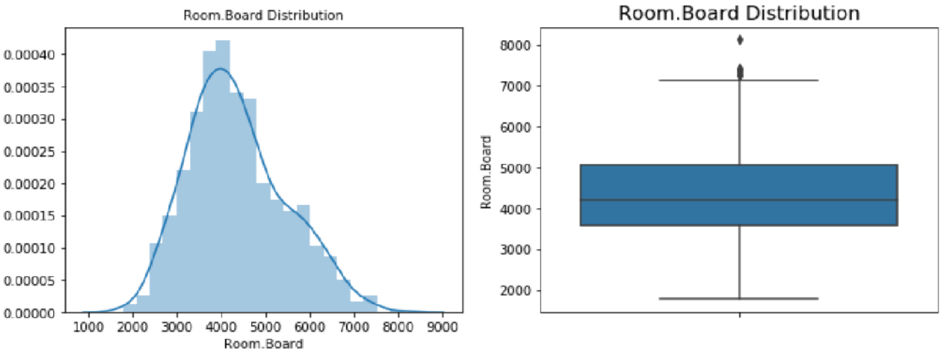
The distribution is positively skewed. The boxplot has lots of outliers. Part time undergraduates studying are in the range 1000 to 3000.

OUTSTATE:

The distribution here seems to be normal. The boxplot has only one outlier.

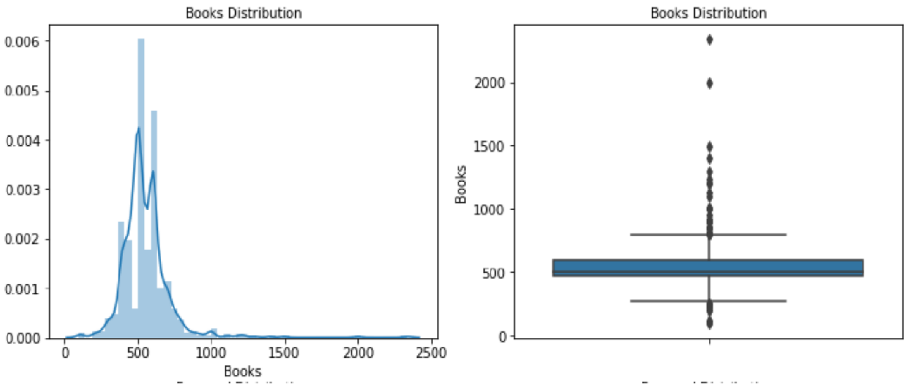


ROOM BOARD:



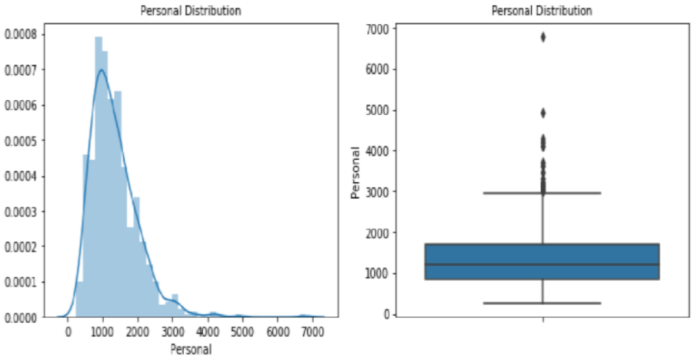
The distribution is normally distributed. There are only few outliers in the data.

BOOKS:

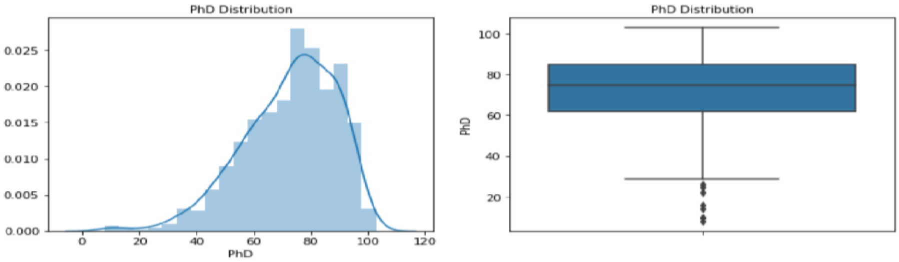


The boxplot contains outliers. Books distribution seems to be in the range of 250 to 750.

PERSONAL: The distribution seems to be positively skewed. The boxplot contains outliers.

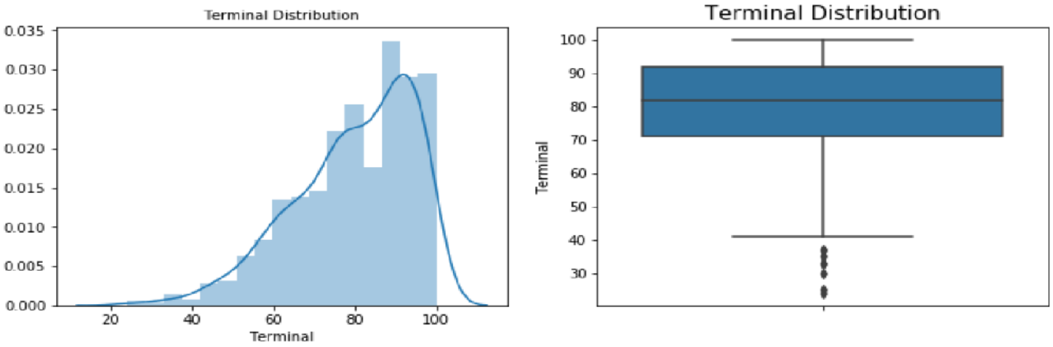


PHD:



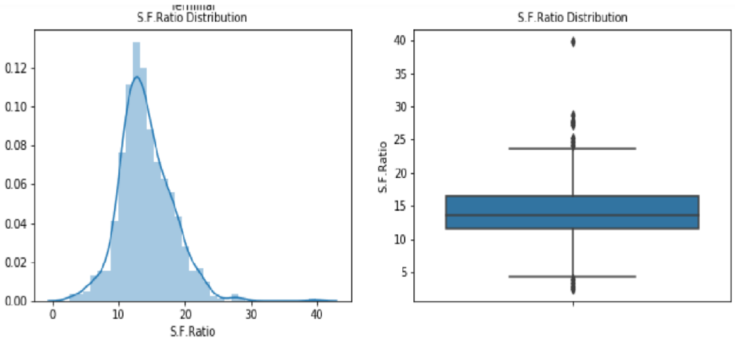
The boxplot has outliers. The distribution is negatively skewed.

TERMINAL:



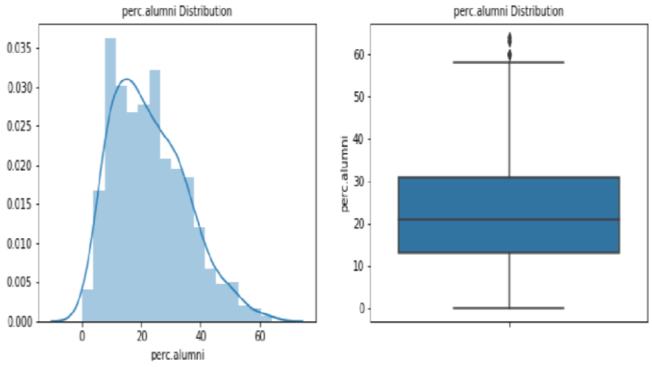
The distribution is negatively skewed. The boxplot has many outliers.

SF RATIO:



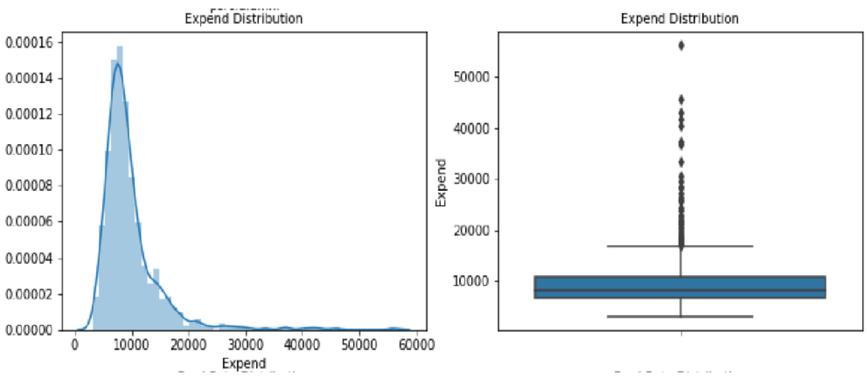
The distribution is almost normally distributed. It also has outliers in the dataset.

PERCI ALUMINI:



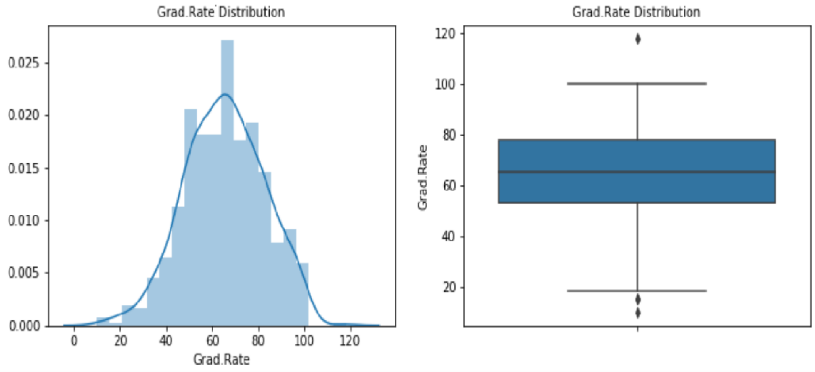
The distribution is almost normal, and it seems to have outliers also.

EXPENDITURE:



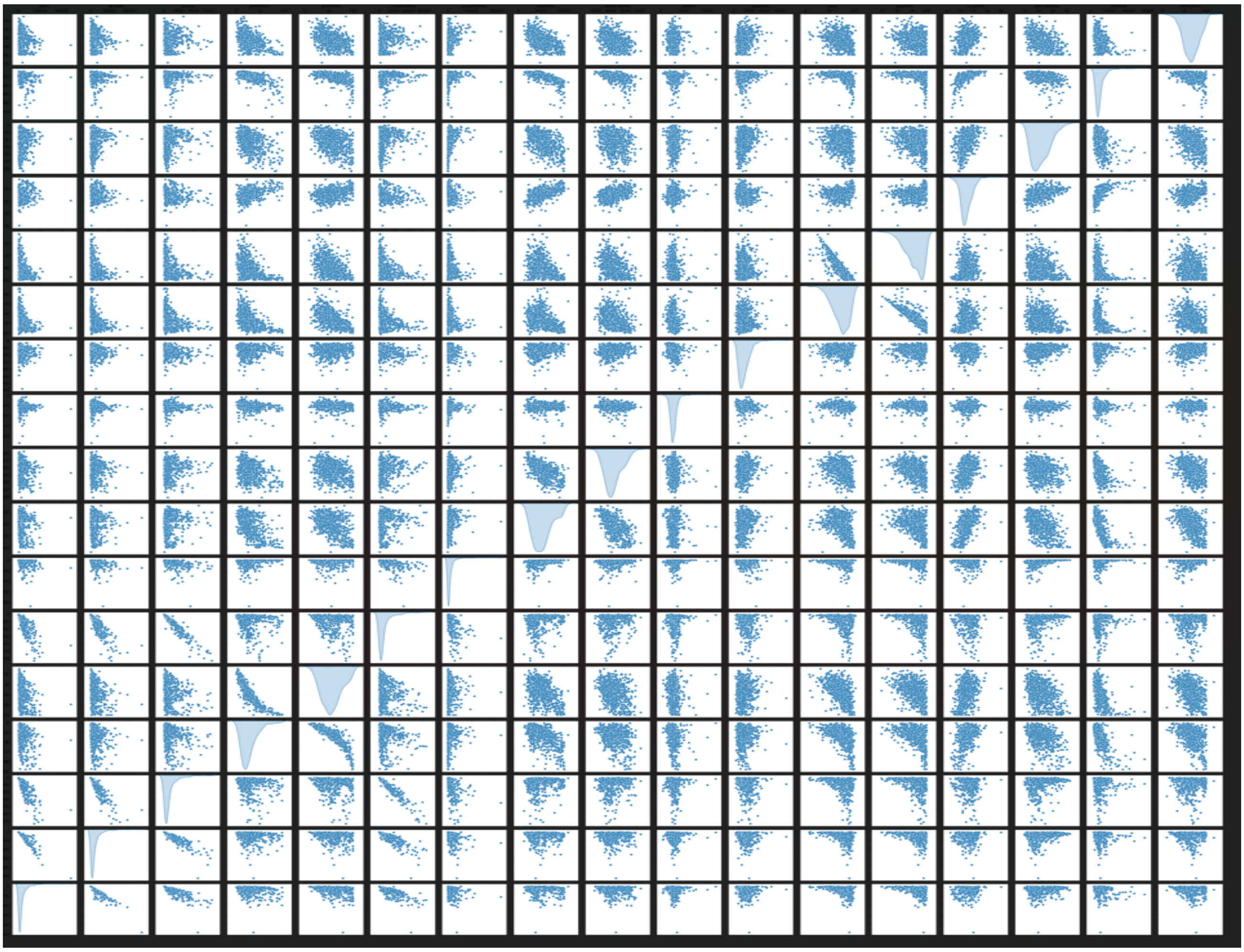
The distribution is positively skewed, and it also has outliers in the dataset.

GRAD RATE:



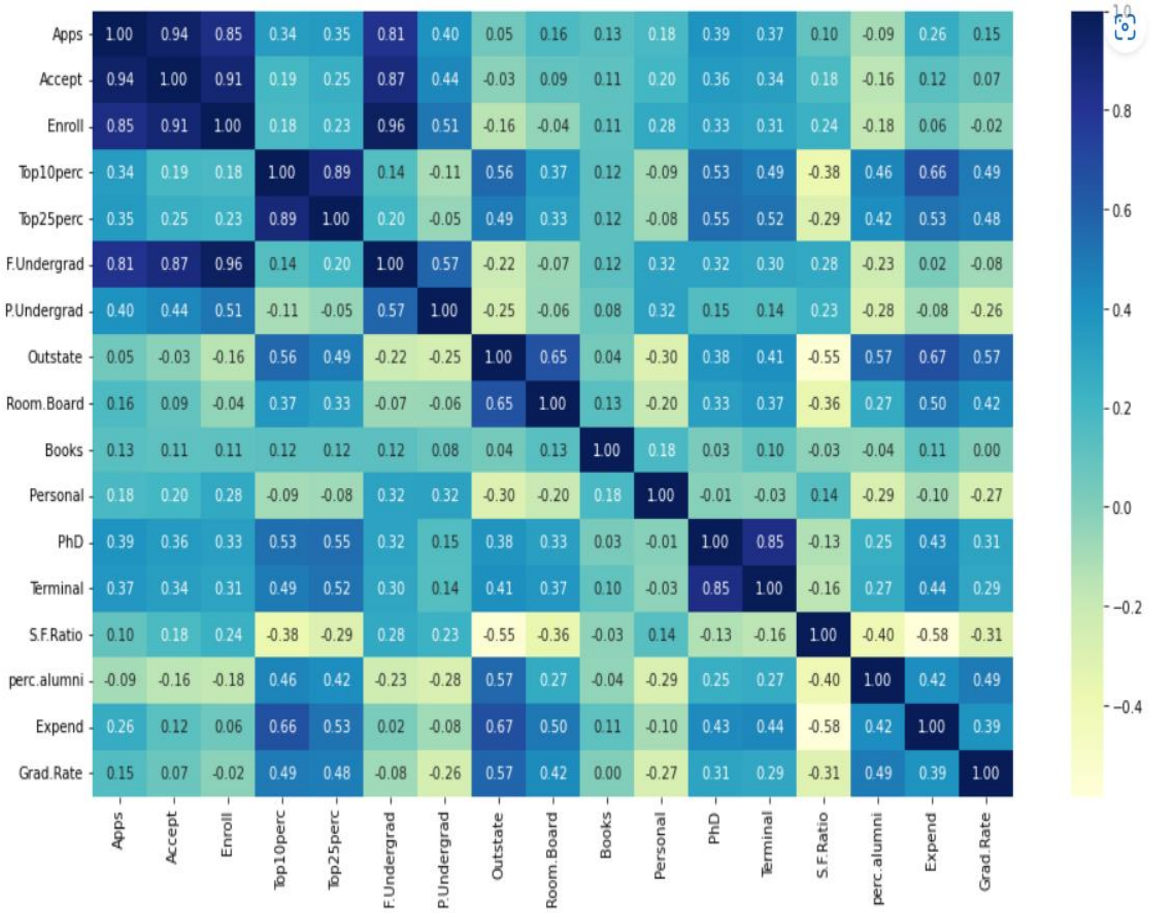
The distribution is normally distributed. The boxplot has outliers in the dataset. The median is around 65.

MULTIVARIATE ANALYSIS



The **pairplot()** function can also be used to showcase the subset of variables, or we can plot different types of variables on rows and columns. It shows the relationship between variables in the form of a scatterplot and a histogram.

HEATMAP:



	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	1.00	0.94	0.85	0.34	0.35	0.81	0.40	0.05	0.16	0.13	0.18	0.39	0.37	0.10	-0.09	0.26	0.15
Accept	0.94	1.00	0.91	0.19	0.25	0.87	0.44	-0.03	0.09	0.11	0.20	0.36	0.34	0.18	-0.16	0.12	0.07
Enroll	0.85	0.91	1.00	0.18	0.23	0.96	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	-0.18	0.06	-0.02
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.37	0.12	-0.09	0.53	0.49	-0.38	0.46	0.66	0.49
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33	0.12	-0.08	0.55	0.52	-0.29	0.42	0.53	0.48
F.Undergrad	0.81	0.87	0.96	0.14	0.20	1.00	0.57	-0.22	-0.07	0.12	0.32	0.32	0.30	0.28	-0.23	0.02	-0.08
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	-0.06	0.08	0.32	0.15	0.14	0.23	-0.28	-0.08	-0.26
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.65	0.04	-0.30	0.38	0.41	-0.55	0.57	0.67	0.57
Room.Board	0.16	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.65	1.00	0.13	-0.20	0.33	0.37	-0.36	0.27	0.50	0.42
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.08	0.04	0.13	1.00	0.18	0.03	0.10	-0.03	-0.04	0.11	0.00
Personal	0.18	0.20	0.28	-0.09	-0.08	0.32	0.32	-0.30	-0.20	0.18	1.00	-0.01	-0.03	0.14	-0.29	-0.10	-0.27
PhD	0.39	0.36	0.33	0.53	0.55	0.32	0.15	0.38	0.33	0.03	-0.01	1.00	0.85	-0.13	0.25	0.43	0.31
Terminal	0.37	0.34	0.31	0.49	0.52	0.30	0.14	0.41	0.37	0.10	-0.03	0.85	1.00	-0.16	0.27	0.44	0.29
S.F.Ratio	0.10	0.18	0.24	-0.38	-0.29	0.28	0.23	-0.55	-0.36	-0.03	0.14	-0.13	-0.16	1.00	-0.40	-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	-0.40	1.00	0.42	0.49
Expend	0.26	0.12	0.06	0.66	0.53	0.02	-0.08	0.67	0.50	0.11	-0.10	0.43	0.44	-0.58	0.42	1.00	0.39
Grad.Rate	0.15	0.07	-0.02	0.49	0.48	-0.08	-0.26	0.57	0.42	0.00	-0.27	0.31	0.29	-0.31	0.49	0.39	1.00

Heatmaps are a method of representing data graphically where values are depicted by color, making it easy to visualize complex data and understand it briefly. The shade of color in a heat map typically indicates the correlation between two numerical variables.

We can positive correlation between "Apps" and Accept, Enroll, Top10perc, phD, Terminal, SF ratio. There is negative correlation between "Apps" and "perc alumni".

The value of correlation takes any value from -1 to 1. Correlation near to 1 means they're highly positively correlated. A correlation of -1 means the two variables are perfectly negatively correlated, which means as one variable increases, the other decreases.

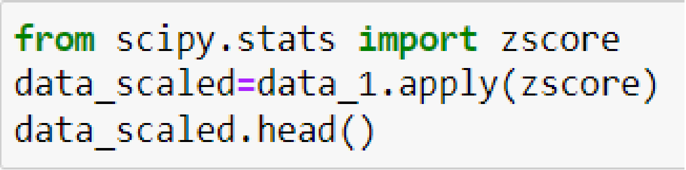
2. Is scaling necessary for PCA in this case? Give justification and perform scaling.

Solution: It is recommended to center data before performing PCA since the transformation relies on the data being around the origin. Some data might already follow a standard normal distribution with a mean zero and standard deviation of one and so would not have to be scaled before PCA. If we are getting several PCA components for multiple features, it is best to scale them otherwise our algorithm might interpret one as more important than the others without any real reason.

Before PCA, I dropped the column" Names" which is categorical. I did z-score scaling here for the normalisation of data. Z-score is a variation of scaling that represents the number of standard deviations away from the mean. We use z-score to ensure our feature distributions have mean = 0 and std = 1.

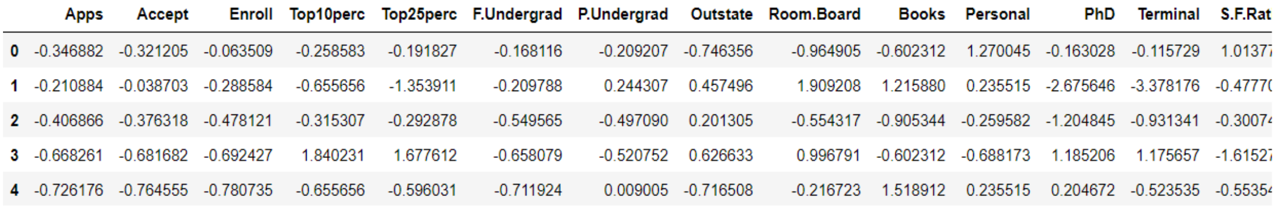
The formula for calculating the z-score of a point, *x*, is as follows:

x′=(x−μ)/σ



```
from scipy.stats import zscore
data_scaled=data_1.apply(zscore)
data_scaled.head()
```

Scaled Data:



	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	-1.61527
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

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

Solution: Covariance is when the two variables vary with each other. It signifies the direction of the linear relationship between the two variables. By direction, we mean if the two variables are directly or inversely proportional to each other.

Positive covariance: Indicates that two variables tend to move in the same direction.

Negative covariance: Reveals that two variables tend to move in inverse directions.

Population Covariance Formula: Cov (x, y) = Σ ((xi – x) * (Yi -y) / N

Sample Covariance Formula: Cov (x, y) = Σ ((xi – x) * (Yi –y) / (N – 1)

where, xi= Data variable of x, Yi= Data variable of y, x = Mean of x, y = Mean of y, N= Number of data variables.

Covariance matrix:

```
array ([[ 1.00128866, 0.94466636, 0.84791332, 0.33927032, 0.35209304,
     0.81554018, 0.3987775, 0.05022367, 0.16515151, 0.13272942,
     0.17896117, 0.39120081, 0.36996762, 0.09575627, -0.09034216,
     0.2599265, 0.14694372],
    [ 0.94466636, 1.00128866, 0.91281145,   0.19269493,   0.24779465,
     0.87534985, 0.44183938, -0.02578774, 0.09101577,   0.11367165,
     0.20124767, 0.35621633, 0.3380184, 0.17645611, -0.16019604,
     0.12487773, 0.06739929],
    [ 0.84791332, 0.91281145, 1.00128866, 0.18152715, 0.2270373,
     0.96588274, 0.51372977, -0.1556777, -0.04028353, 0.11285614,
     0.28129148, 0.33189629, 0.30867133, 0.23757707, -0.18102711,
     0.06425192, -0.02236983],
    [ 0.33927032, 0.19269493, 0.18152715, 1.00128866,    0.89314445,
     0.1414708, -0.10549205, 0.5630552,    0.37195909, 0.1190116 ,
    -0.09343665, 0.53251337, 0.49176793, -0.38537048,  0.45607223,
     0.6617651, 0.49562711],
    [ 0.35209304, 0.24779465, 0.2270373,  0.89314445,  1.00128866,
     0.19970167, -0.05364569, 0.49002449, 0.33191707, 0.115676  ,
    -0.08091441, 0.54656564, 0.52542506, -0.29500852,  0.41840277,
     0.52812713, 0.47789622],
    [ 0.81554018, 0.87534985, 0.96588274,  0.1414708 ,  0.19970167,
     1.00128866, 0.57124738, -0.21602002, -0.06897917,  0.11569867,
     0.31760831,  0.3187472 ,  0.30040557, 0.28006379, -0.22975792,
     0.01867565, -0.07887464],
    [ 0.3987775 ,  0.44183938,  0.51372977, -0.10549205, -0.05364569,
     0.57124738,  1.00128866, -0.25383901, -0.06140453, 0.08130416,
     0.32029384,  0.14930637,  0.14208644,  0.23283016, -0.28115421,
    -0.08367612, -0.25733218],
    [ 0.05022367, -0.02578774, -0.1556777 ,  0.5630552 ,  0.49002449,
    -0.21602002, -0.25383901,  1.00128866,  0.65509951,  0.03890494,
    -0.29947232,  0.38347594,  0.40850895, -0.55553625,  0.56699214,
     0.6736456 ,  0.57202613],
    [ 0.16515151,  0.09101577, -0.04028353,  0.37195909,  0.33191707,
    -0.06897917, -0.06140453,  0.65509951,  1.00128866,  0.12812787,
    -0.19968518,  0.32962651,  0.3750222 , -0.36309504,  0.27271444,
     0.50238599,  0.42548915],
    [ 0.13272942,  0.11367165,  0.11285614,  0.1190116 ,  0.115676  ,
     0.11569867,  0.08130416,  0.03890494,  0.12812787,  1.00128866,
     0.17952581,  0.0269404 ,  0.10008351, -0.03197042, -0.04025955,
     0.11255393,  0.00106226],
    [ 0.17896117,  0.20124767,  0.28129148, -0.09343665, -0.08091441,
     0.31760831,  0.32029384, -0.29947232, -0.19968518,  0.17952581,
     1.00128866, -0.01094989, -0.03065256,  0.13652054, -0.2863366 ,
    -0.09801804, -0.26969106],
    [ 0.39120081,  0.35621633,  0.33189629,  0.53251337,  0.54656564,
     0.3187472 ,  0.14930637,  0.38347594,  0.32962651,  0.0269404 ,
    -0.01094989, 1.00128866,  0.85068186, -0.13069832,  0.24932955,
     0.43331936, 0.30543094],
    [ 0.36996762, 0.3380184,  0.30867133, 0.49176793,  0.52542506,
     0.30040557, 0.14208644, 0.40850895, 0.3750222,  0.10008351,
    -0.03065256, 0.85068186, 1.00128866, -0.16031027, 0.26747453,
     0.43936469,  0.28990033],
    [ 0.09575627,  0.17645611,  0.23757707, -0.38537048, -0.29500852,
     0.28006379,  0.23283016, -0.55553625, -0.36309504, -0.03197042,
     0.13652054, -0.13069832, -0.16031027,  1.00128866, -0.4034484 ,
    -0.5845844 , -0.30710565],
    [-0.09034216, -0.16019604, -0.18102711,  0.45607223,  0.41840277,
    -0.22975792, -0.28115421, 0.56699214, 0.27271444, -0.04025955,
    -0.2863366, 0.24932955, 0.26747453, -0.4034484,  1.00128866,
     0.41825001, 0.49153016],
    [ 0.2599265, 0.12487773, 0.06425192,  0.6617651 ,  0.52812713,
     0.01867565, -0.08367612, 0.6736456, 0.50238599,   0.11255393,
    -0.09801804,  0.43331936, 0.43936469, -0.5845844 ,  0.41825001,
     1.00128866, 0.39084571],
    [ 0.14694372,  0.06739929, -0.02236983,  0.49562711,  0.47789622,
    -0.07887464, -0.25733218,  0.57202613,  0.42548915,  0.00106226,
    -0.26969106,  0.30543094,  0.28990033, -0.30710565,  0.49153016,
     0.39084571,  1.00128866]])
```

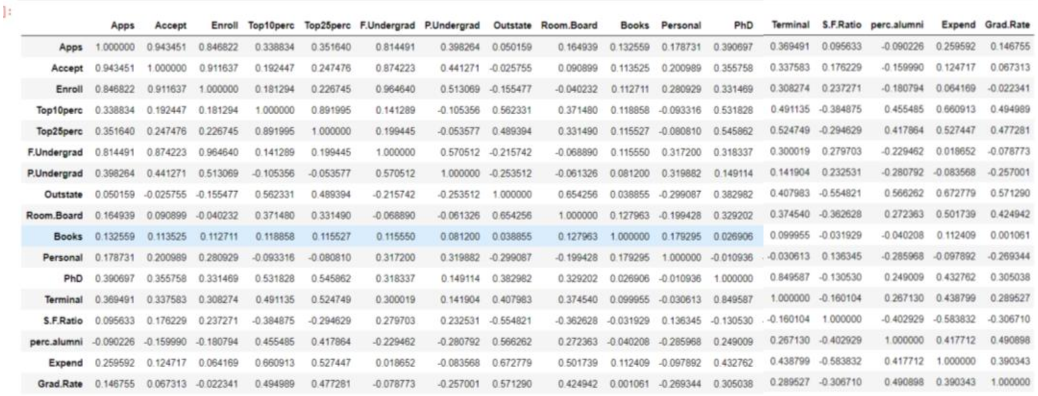
A correlation is a statistical measure of the relationship between two variables. The measure is best used in variables that demonstrate a linear relationship between each other. The fit of the data can be visually represented in a scatterplot.

- Positive Correlation – when the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.
- Negative Correlation – when the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.
- No Correlation – when there is no linear dependence or no relation between the two variables.

Correlation Formula: The covariance of two variables divided by the product of their standard deviations gives correlation. It is usually represented by ρ (rho).

ρ (X, Y) = cov (X, Y) / σX.σY, where σX and σY are the standard deviation of x and standard deviation of y respectively.

Correlation matrix:

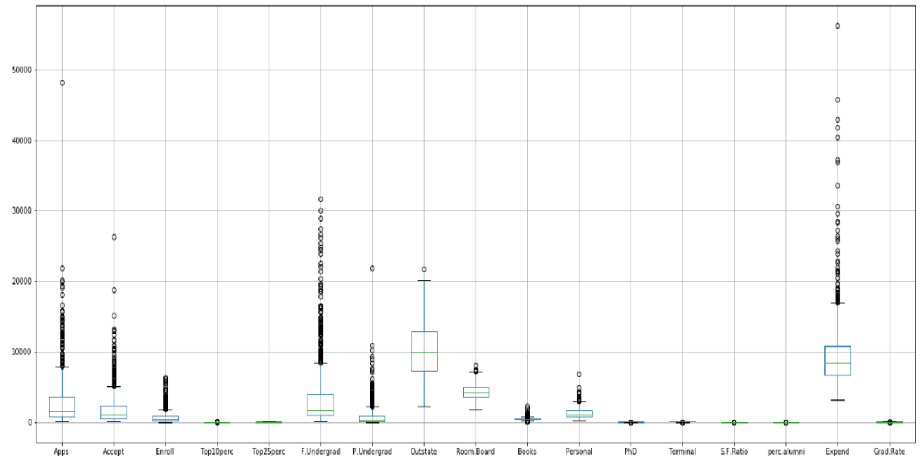


	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369491	0.095633	-0.090226	0.259592	0.146755
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

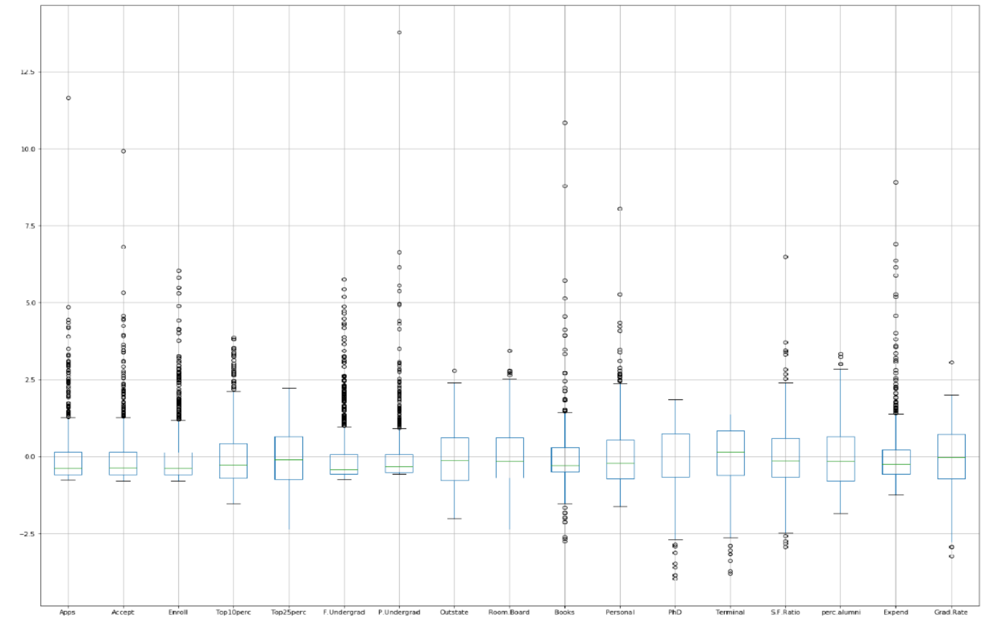
4. Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]

Solution:

Data before scaling:



Data after scaling:



Scaling does not remove outliers. Even after scaling, the outliers are still present in our dataset. To remove outliers, either drop the outlier value or replace the outlier value using the Inter Quartile Range**.** We can use Min max scaler/ log transformation for scaling.

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

Solution: 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]]
```

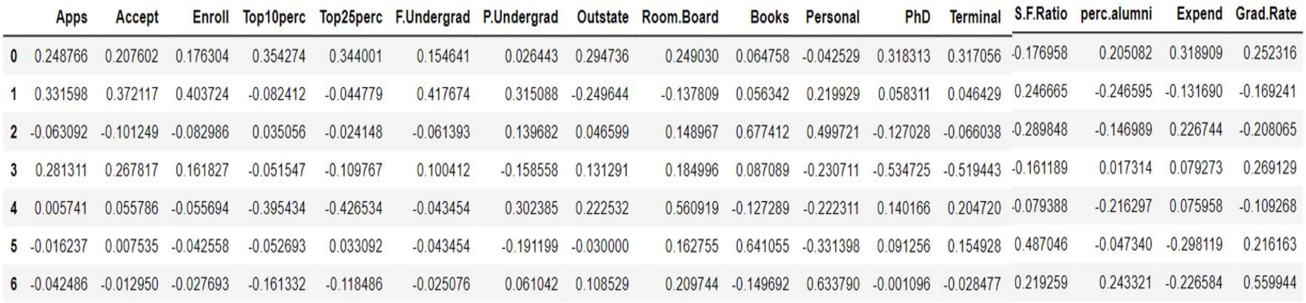
Eigen Values:

```
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]
```

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

Solution:

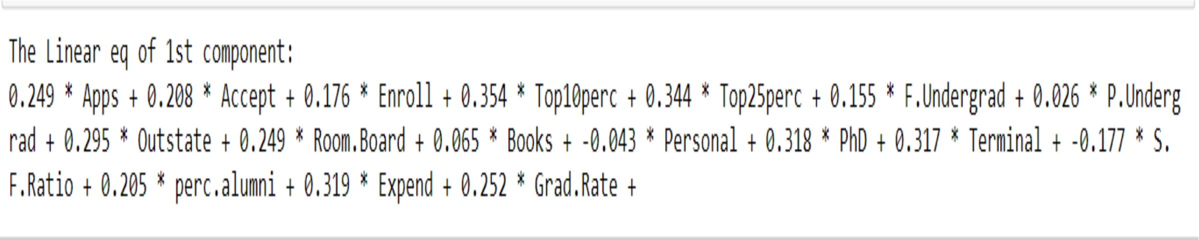
Principal Components Analysis (PCA) is an algorithm to transform the columns of a dataset into a new set of features called Principal Components. By doing this, a large chunk of the information across the full dataset is effectively compressed into fewer feature columns. This enables dimensionality reduction and the ability to visualize the separation of classes or clusters if any.



	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
5	-0.016237	0.007535	-0.042558	-0.052693	0.033092	-0.043454	-0.191199	-0.030000	0.162755	0.641055	-0.331398	0.091256	0.154928	0.487046	-0.047340	-0.298119	0.216163
6	-0.042486	-0.012950	-0.027693	-0.161332	-0.118486	-0.025076	0.061042	0.108529	0.209744	-0.149692	0.633790	-0.001096	-0.028477	0.219259	0.243321	-0.226584	0.559944

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]

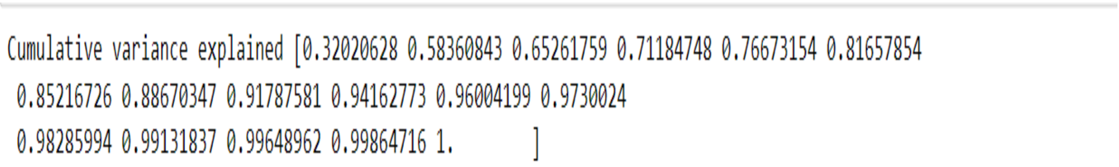
Solution:



```
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 +
```

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?

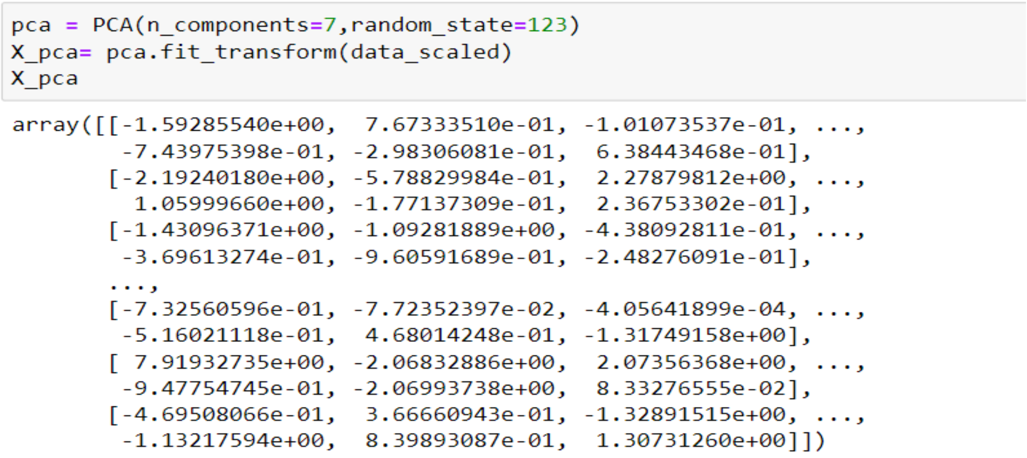
Solution:



```
Cumulative variance explained [0.32020628 0.58360843 0.65261759 0.71184748 0.76673154 0.81657854
 0.85216726 0.88670347 0.91787581 0.94162773 0.96004199 0.9730024
 0.98285994 0.99131837 0.99648962 0.99864716 1.        ]
```

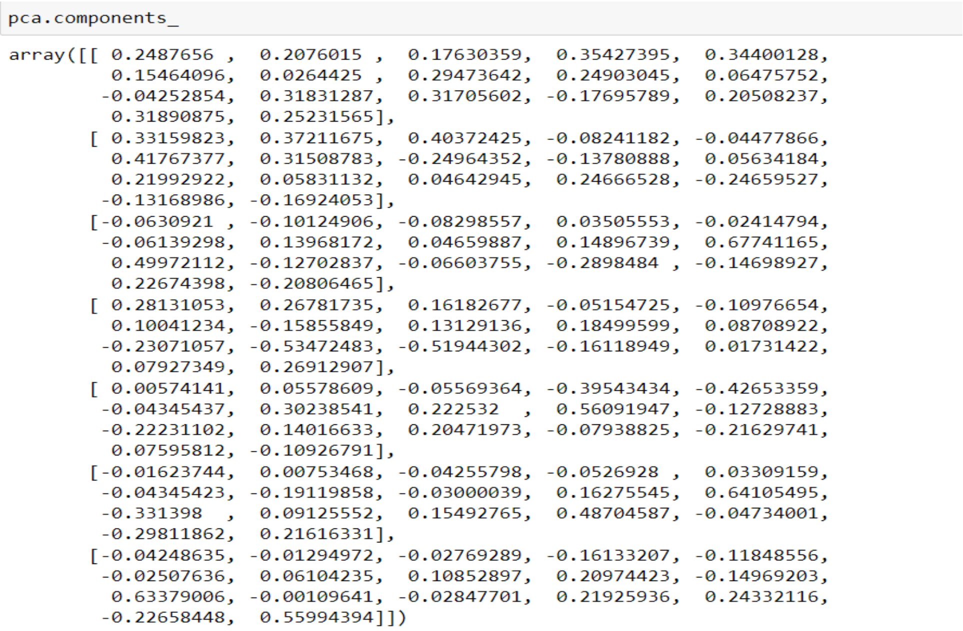
Cumulative variance of all the 17 components can explain 100% variance in data. Cumulative variance helps us to analyse that we're able to get 85.21 percent of the data only in 7 principal components. The first two, three, four, five, six principal components explain 58%,65%,71%,76%,85% cumulative variances respectively. We can now transform into another dataset having lower dimensions and we'll not lose much information when we're transferring our data from high to low dimension.

Eigen vectors indicates us the direction of our main axes (principal components) of our data. The greater the eigen value, the greater the variation along this axis.



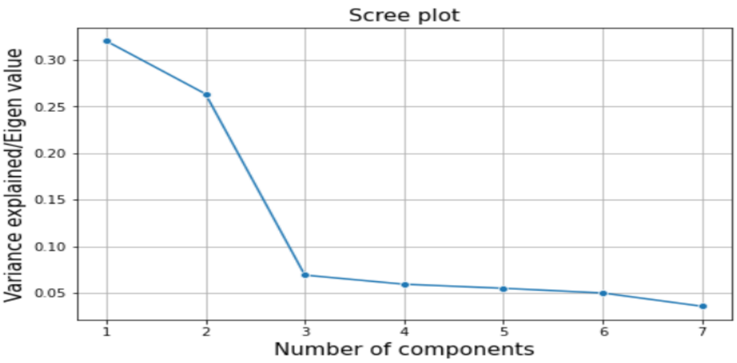
```
pca = PCA(n_components=7,random_state=123)
X_pca= pca.fit_transform(data_scaled)
X_pca
```

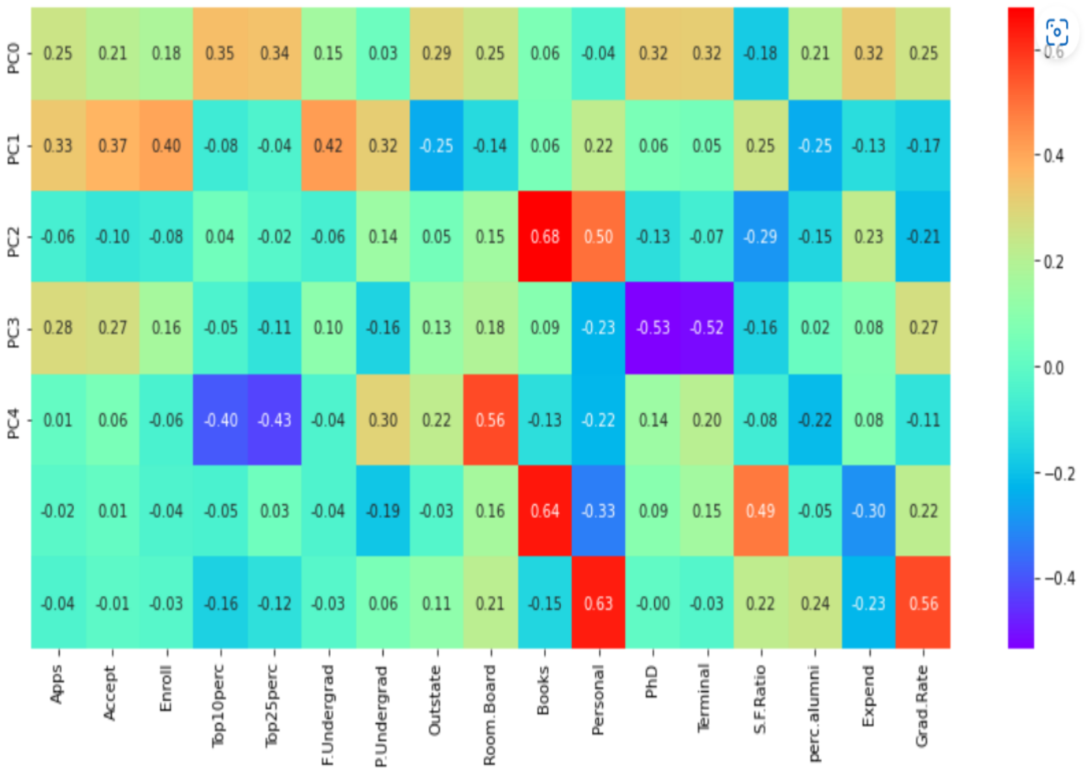
```
array([[-1.59285540e+00,  7.67333510e-01, -1.01073537e-01, ...,
        -7.43975398e-01, -2.98306081e-01,  6.38443468e-01],
       [-2.19240180e+00, -5.78829984e-01,  2.27879812e+00, ...,
         1.05999660e+00, -1.77137309e-01,  2.36753302e-01],
       [-1.43096371e+00, -1.09281889e+00, -4.38092811e-01, ...,
        -3.69613274e-01, -9.60591689e-01, -2.48276091e-01],
       ...,
       [-7.32560596e-01, -7.72352397e-02, -4.05641899e-04, ...,
        -5.16021118e-01,  4.68014248e-01, -1.31749158e+00],
       [ 7.91932735e+00, -2.06832886e+00,  2.07356368e+00, ...,
        -9.47754745e-01, -2.06993738e+00,  8.33276555e-02],
       [-4.69508066e-01,  3.66660943e-01, -1.32891515e+00, ...,
        -1.13217594e+00,  8.39893087e-01,  1.30731260e+00]])
```



```
pca.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]])
```





	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
PC0	0.25	0.21	0.18	0.35	0.34	0.15	0.03	0.29	0.25	0.06	-0.04	0.32	0.32	-0.18	0.21	0.32	0.25
PC1	0.33	0.37	0.40	-0.08	-0.04	0.42	0.32	-0.25	-0.14	0.06	0.22	0.06	0.05	0.25	-0.25	-0.13	-0.17
PC2	-0.06	-0.10	-0.08	0.04	-0.02	-0.06	0.14	0.05	0.15	0.68	0.50	-0.13	-0.07	-0.29	-0.15	0.23	-0.21
PC3	0.28	0.27	0.16	-0.05	-0.11	0.10	-0.16	0.13	0.18	0.09	-0.23	-0.53	-0.52	-0.16	0.02	0.08	0.27
PC4	0.01	0.06	-0.06	-0.40	-0.43	-0.04	0.30	0.22	0.56	-0.13	-0.22	0.14	0.20	-0.08	-0.22	0.08	-0.11
	-0.02	0.01	-0.04	-0.05	0.03	-0.04	-0.19	-0.03	0.16	0.64	-0.33	0.09	0.15	0.49	-0.05	-0.30	0.22
	-0.04	-0.01	-0.03	-0.16	-0.12	-0.03	0.06	0.11	0.21	-0.15	0.63	-0.00	-0.03	0.22	0.24	-0.23	0.56

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:

This business case study is of the dataset that contains the names of various colleges, which contains information of colleges ,students, alumni, books etc. and we performed univariate analysis and multivariate analysis which gives us an understanding about the different variables. From univariate analysis, we understand the distribution of the dataset, skewness and patterns in the dataset. From multivariate analysis, we understand the correlation of variables. The scaling helps us to make data points generalized so that the distance between them will be lower. Outliers are treated using IQR values once the values are imputed before we perform PCA. We can also remove them if they're only a few. The principal component analysis is used for dimension reduction which lets us identify correlation and patterns so that they can be transferred into another dataset with lower dimensions without losing much information. Depending on the variance of the dataset we can reduce the PCA components. The PCA component for this business case is 7 where we could understand the maximum variance of the dataset.