ADVANCED STATISTICS PROJECT REPORT

ANOVA
EDA
PCA

Contents:

Problem 1A:

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

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

1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.
2. Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.
3. Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.
4. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. (Non-Graded)

First five rows of the dataset:

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

Shape of the dataset is: (40, 3)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Education   40 non-null     object
 1   Occupation  40 non-null     object
 2   Salary      40 non-null     int64
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

There are 3 variables in the given dataset. Datatypes of 2 variables is object and the other one is numerical. We have 1 dependent variable and 2 independent variables.

In the below step, we will change the datatype of 2 independent variables to categorical variables.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Education   40 non-null     category
 1   Occupation  40 non-null     category
 2   Salary      40 non-null     int64
dtypes: category(2), int64(1)
memory usage: 864.0 bytes
```

```
Below are the categories in Education variable:


 Doctorate    16
 Bachelors    15
 HS-grad       9
Name: Education, dtype: int64
```

```
Below are the categories in Occupation variable:


 Prof-specialty     13
 Sales              12
 Adm-clerical       10
 Exec-managerial     5
Name: Occupation, dtype: int64
```

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

a) Variable: 'Education'

Null hypothesis states that the mean salary for all categories of education are equal.

Alternative hypothesis states that there will be an effect of Education qualification on the Salary. The mean salary for atleast one level of educational qualification is unequal.

$H0$: μ_salary_highschoolgraduate = μ_salary_bachelor = μ_salary_doctorate $H0$: The means of 'Salary' for the three education levels are equal.

$H1$: μ_salary_highschoolgraduate ≠ μ_salary_bachelor = μ_salary_doctorate or

$H1$: μ_salary_highschoolgraduate = μ_salary_bachelor ≠ μ_salary_doctorate or

$H1$: μ_salary_highschoolgraduate = μ_salary_doctorate ≠ μ_salary_bachelor or

$H1$: μ_salary_highschoolgraduate ≠ μ_salary_bachelor ≠ μ_salary_graduate

$H1$: Atleast one of the means of 'Salary' with respect to the various education levels is unequal.

b) Variable: 'Occupation'

Null hypothesis states that the mean salary for all categories of occupation are equal.

Alternative hypothesis states that there will be an effect on Salary with respect to the type of occupation. The mean salary for atleast one category of Occupation is unequal.

$H0$: μ_sal_exec = μ_sal_adm = μ_sal_sales = μ_sal_prof $H0$: The means of Salary between various occupations are equal

$H1$: μ_sal_exec ≠ μ_sal_adm = μ_sal_sales = μ_sal_prof or

$H1$: μ_sal_exec = μ_sal_adm ≠ μ_sal_sales = μ_sal_prof or

$H1$: μ_sal_exec = μ_sal_sales ≠ μ_sal_prof = μ_sal_adm or

$H1$: μ_sal_prof ≠ μ_sal_exec = μ_sal_adm = μ_sal_sales or

$H1$: μ_sal_exec ≠ μ_sal_adm ≠ μ_sal_sales ≠ μ_sal_prof

$H1$: The means of Salary between various Occupations are unequal (Atleast one of the means between various Occupation is unequal)

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.

Level of signficance α = 0.05

Mean of the variable Salary for all the 40 individuals: 162186.875

```
Salary mean for all the categories under Education variable:

Education
 Bachelors    165152.933333
 Doctorate    208427.000000
 HS-grad       75038.777778
Name: Salary, dtype: float64
```

One way ANOVA on Salary with respect to Education:

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

We see that the p-value(1.257709e-08) is less than alpha (0.05). Thus, we Reject the Null Hypothesis (H0). This means at least one particular category in the 'Education' variable has different mean of Salary as compared to the other categories.

Since p-value is less than 0.05, we have sufficient evidence to say that the mean salary across each category for education are not equal.

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

mean of the variable Salary for all the 40 individuals: 162186.875

```
Salary mean for all the categories under Occupation variable:

Occupation
 Adm-clerical       141424.300000
 Exec-managerial    197117.600000
 Prof-specialty     168953.153846
 Sales              157604.416667
Name: Salary, dtype: float64
```

One way ANOVA on Salary with respect to Occupation:

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

We see that the p-value(0.4585) is greater than alpha (0.05). Thus, we failed to Reject the Null Hypothesis (H0). Hence, mean salary for all categories of occupation are equal.

Since p-value is greater than 0.05, we have sufficient evidence to say that the mean salary across each category for occupation are equal.

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)

To find out which class means are significantly different, the Tukey Honest Significant Difference test is performed. Null Hypothesis is rejected when one way ANOVA is conducted on Salary with respect to Education.

Tukey Honest Significant Difference is performed below for Education variable to check which category means are significantly different.

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

P-value for the difference in means between Bachelors and Doctorate : 0.0146

P-value for the difference in means between Bachelors and HS-grad : 0.001

P-value for the difference in means between Doctorate and HS-grad : 0.001
Thus, we can say that there is a statistically significant difference between the means of Bachelors and Doctorate; Bachelors and HS-grad and also Doctorate and HS-grad.

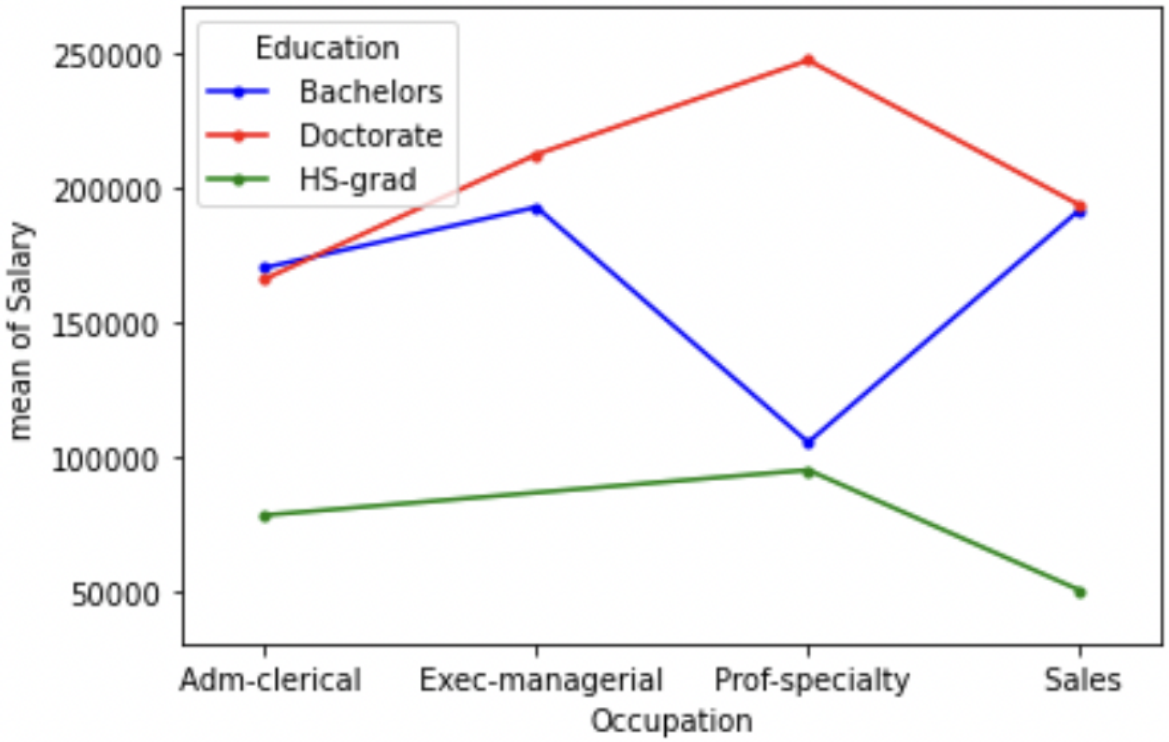
We can conclude that there is significant difference between the means for all the levels of Education.

Problem 1B:

1. 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.[hint: Use Interaction_Plot function from statsmodel.graphics.factorplots module and give inference from the plot.]
2. Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?
3. Explain the business implications of performing ANOVA for this particular case study.

	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769
5	Doctorate	Sales	219420
6	Doctorate	Sales	237920
7	Doctorate	Sales	160540
8	Doctorate	Sales	180934
9	Doctorate	Prof-specialty	248156

1. **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.[hint: Use Interaction_Plot function from statsmodel.graphics.factorplots module and give inference from the plot.]**



The interaction plot shows that there is significant amount of interaction between the categorical variables, Education and Occupation.

From the interaction plot, we can say that:

1. People having HS-grad education level are not in the position of Exec-managerial and they are able to get only Adm-clerical, Sales and Prof-Specialty occupations
2. People who have a Doctorate and work as a Prof-specialty earn the highest salaries when compared to all other levels in both Education and Occupation.
3. People who have HS-grad education level and working in Sales earn the least salaries when compared to all other levels in both Education and Occupation.
4. As per the interaction, it is clear that people with education level as Bachelors or Doctorate and occupation as Adm-clerical and Sales almost earn the same salaries.
5. People with education as HS-Grad earn the minimum salaries in all the categories of Occupation.

6. When calculating the total salary of all categories in Occupation, people with education level as Doctorate tend to earn the maximum.
7. People with education as Bachelors and occupation as Sales earn higher than people with education as Bachelors and occupation Prof-Specialty. We see a 'v' curve for this.
8. People with education as Doctorate and occupation Sales earn lesser than people with Doctorate and occupation Prof-Specialty. We see a reverse 'v' curve in this part of the graph.

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?

Null Hypothesis(H0): The effect of the independent variable ‘education’ on the mean ‘salary’ does not depend on the effect of the other independent variable ‘occupation’. There is no interaction effect between the 2 independent variables, education and occupation.

Alternative Hypothesis(H1): There is an interaction effect between the independent variables ‘education’ and ‘occupation’ on the mean Salary.

Two way annova with Education and Occupation along with Interaction effect:

	df	sum_sq	mean_sq	F	PR(>F)
Education	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
Occupation	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
Education: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

P-value for Education = 5.466264e-12 < 0.05

P-value for Occupation = 7.211580e-02 > 0.05

P-value for Interaction = 2.232500e-05 < 0.05

The P-value obtained from ANOVA for Education is statistically significant (P<0.05). The P-value obtained from ANOVA for Occupation is not statistically significant (P>0.05).

Also the interaction effect using both Education and Occupation is statistically significant (P<0.05).

We can conclude that Education significantly affect the Salary variable but not Occupation. However, Education and Occupation in interaction, significantly affect the Salary variable.

For Education, atleast one of the means is different for all the categories with respect to Salary For Occupation, means of Salary is same for all the categories For interaction term, at least one of the means is different for the combination of Education and Occupation with respect to Salary.

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

1. From ANOVA method, education individually has effect on Salary variable. Whereas Occupation individually doesn't effect the Salary variable.
2. From the ANOVA method and the interaction plot, we see that education combined with occupation results in higher and better salaries among the individuals.
3. It is clear that individuals with education as Doctorate have maximum salaries and individuals with education as HS-grad earn the least salaries.
4. Thus, we can conclude that Salary is dependent on educational qualifications and different occupations.

Problem 2:

The dataset Education - Post 12th Standard.csv contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx.

1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?
2. Is scaling necessary for PCA in this case? Give justification and perform scaling.
3. Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].
4. Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]
5. Extract the eigenvalues and eigenvectors.
6. Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features
7. Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]
8. Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?
9. Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]

First 5 rows of the dataset:

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alu
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	

Shape of the dataset is: (777, 18)

Checking the datatypes of the dataset:

```
Names            object
Apps              int64
Accept            int64
Enroll            int64
Top10perc         int64
Top25perc         int64
F.Undergrad       int64
P.Undergrad       int64
Outstate          int64
Room.Board        int64
Books             int64
Personal          int64
PhD               int64
Terminal          int64
S.F.Ratio       float64
perc.alumni       int64
Expend            int64
Grad.Rate         int64
dtype: object
```

```
<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
```

Viewing the data summary of numeric features (.T just transposes the information and is done for better readability)

Describe function gives all the information such as mean, standard deviation, minimum value, maximum value,
first quartile, median and third quartile

	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

Checking for missing values:

```
Names           0
Apps            0
Accept          0
Enroll          0
Top10perc       0
Top25perc       0
F.Undergrad     0
P.Undergrad     0
Outstate        0
Room.Board      0
Books           0
Personal        0
PhD             0
Terminal        0
S.F.Ratio       0
perc.alumni     0
Expend          0
Grad.Rate       0
dtype: int64
```

Observations:

We have 777 transactions and 18 features captured for the same.

We observe that the column ‘Names’ is the only variable which doesn't have numerical variables and is a categorical variable.

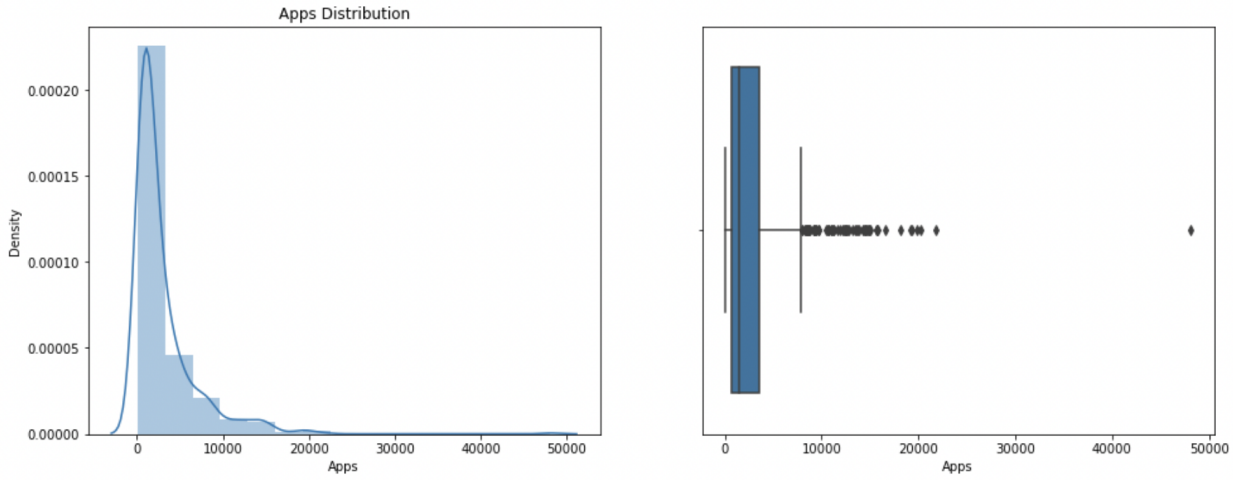
All other columns seems to be integer or float values. In all, we have 17 numeric features and 1 non-numeric.

There are no missing values or null values in the dataset. There are no duplicates in the dataset.

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

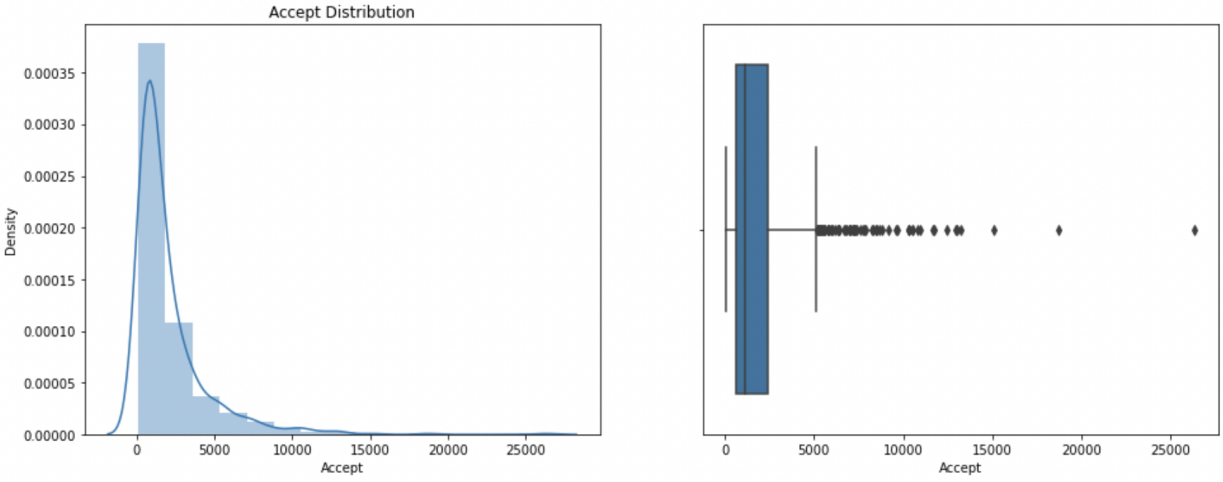
Univariate Analysis: To understand the distribution of data in the dataset. We can find patterns and summarize the data for all the variables.

Apps:



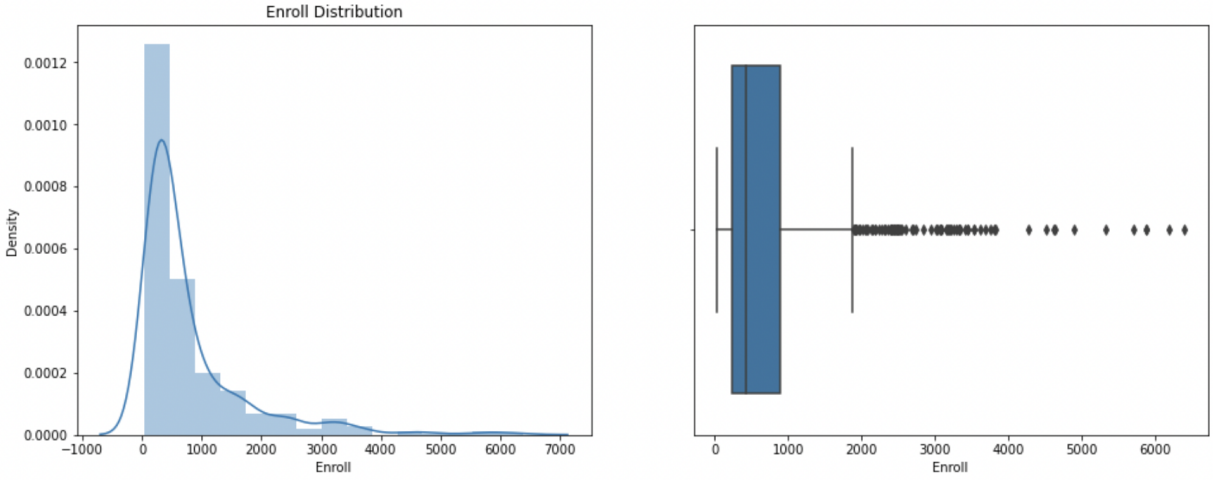
The boxplot of Apps variable has outliers. The distribution is right skewed, we could see that each college offers applications from 3000 to 5000. The max applications are around 50,000.

Accept:



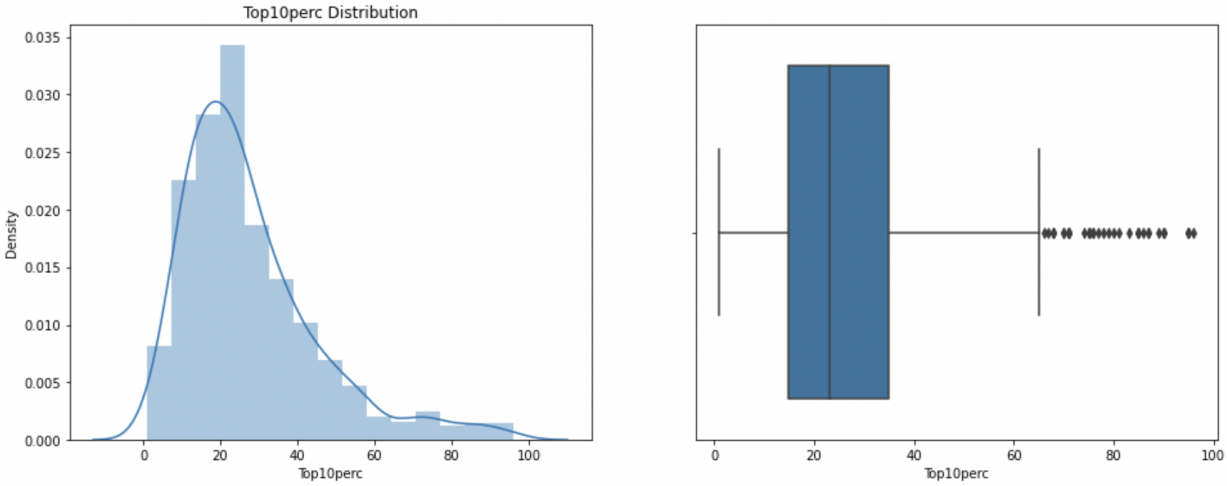
As per the boxplot, Accept variable seems to have outliers. Majority of applications accepted are in the range of 70 to 1500. It is right skewed.

Enroll:



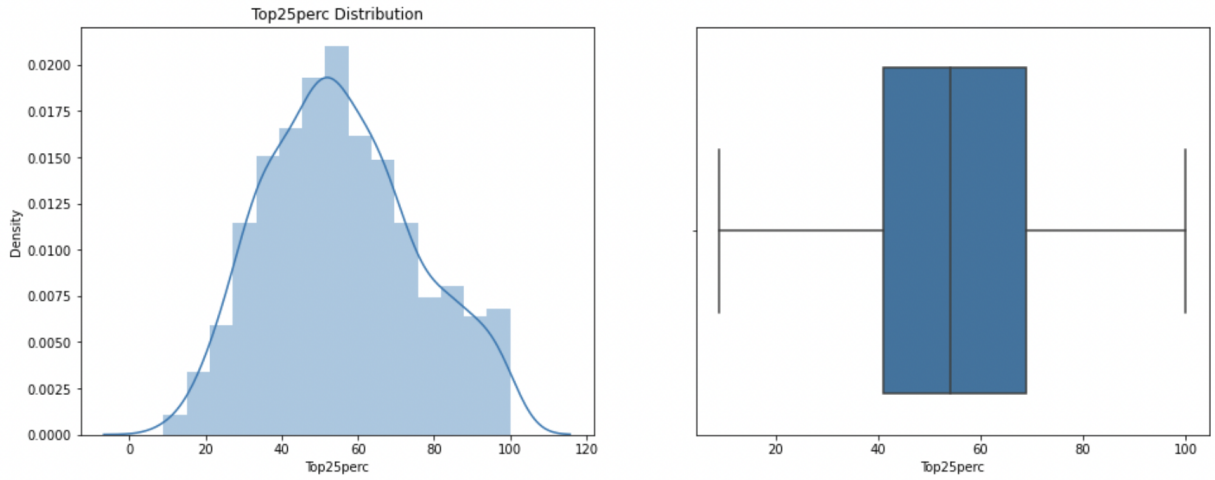
Enroll variable also has outliers as per the boxplot. Dist plot shows that it is right skewed. As per dist plot, we can say that majority of colleges have enrolled students in the range of 100 to 500.

Top 10 perc:



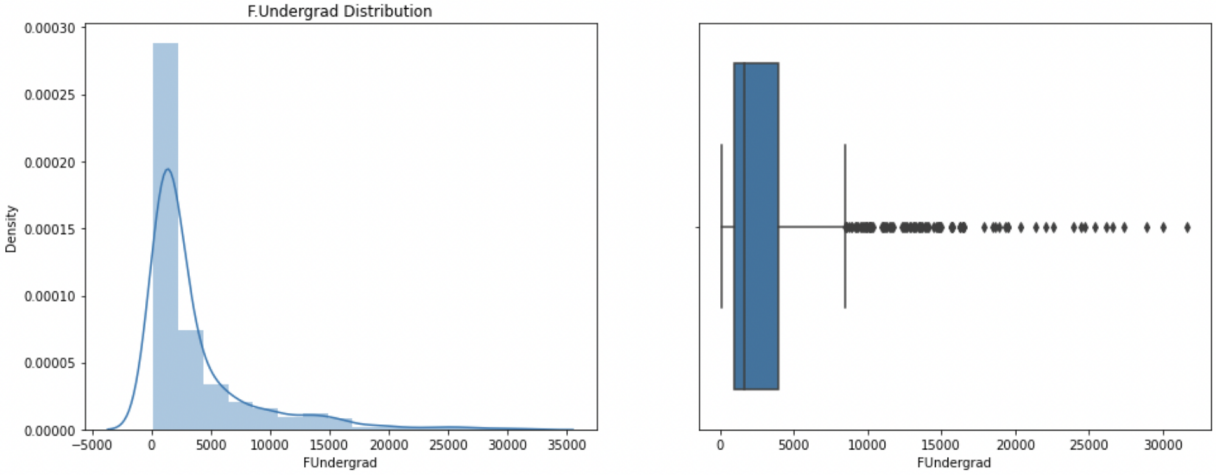
Top 10 perc have outliers as per the boxplot. The distribution here is slightly right skewed.

Top 25 perc:



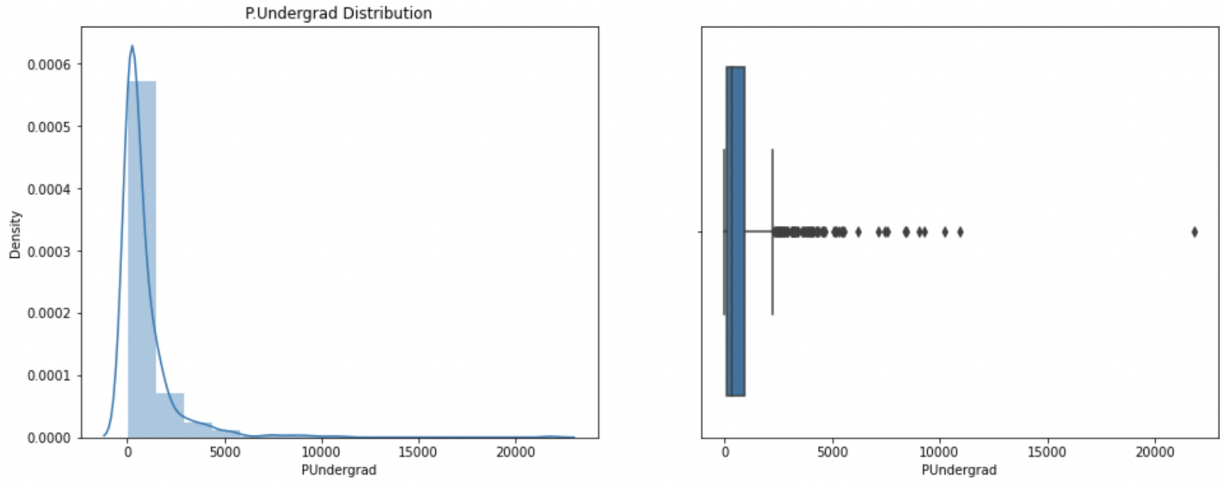
The box plot for the top 25% has no outliers. The distribution is almost normally distributed.

F Undergrad:



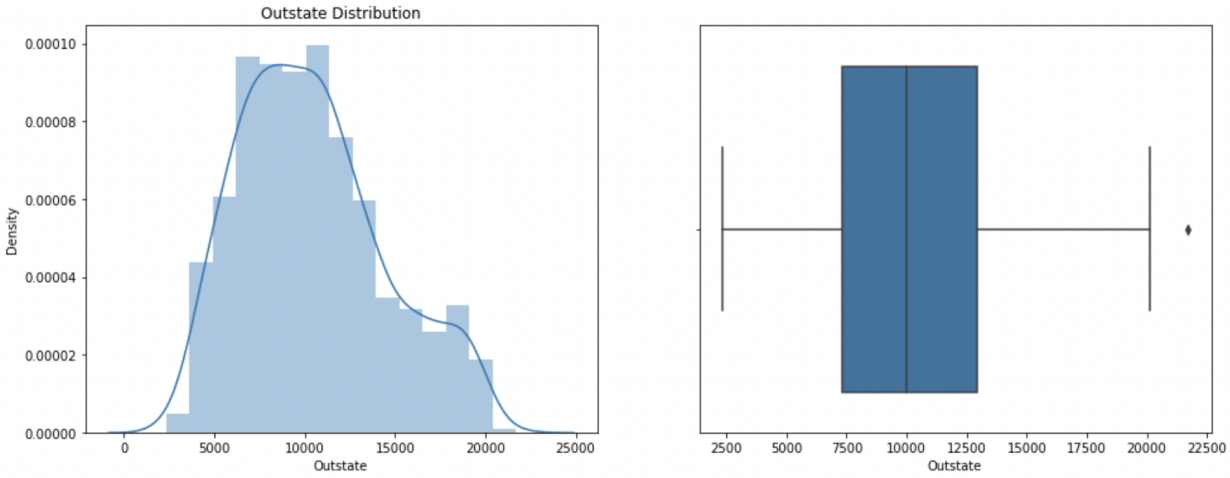
Full time Undergrad variable has outliers as per the box plot. The distribution of data is positively skewed.

P Undergrad:



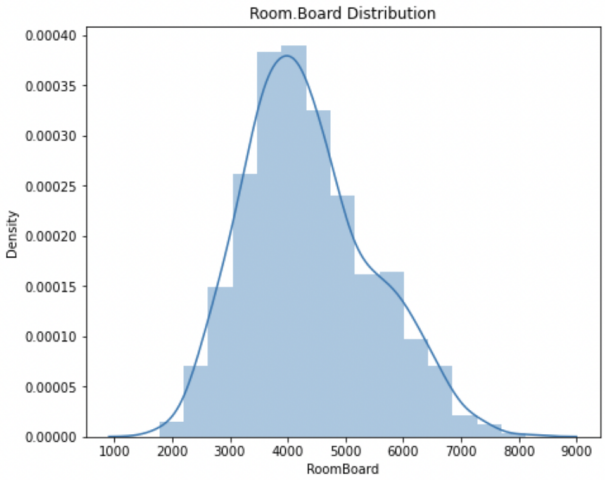
The boxplot of the part time graduates has outliers. As per the dist plot, the distribution of data is right skewed.

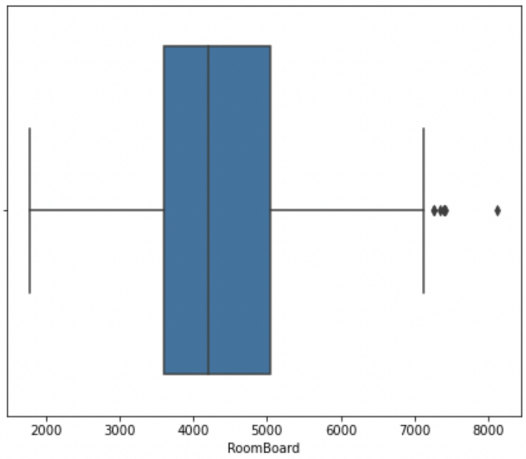
Outstate:



The boxplot of Outstate has only one outlier. As per the dist plot, we can say that the data is a lmost normally distributed.

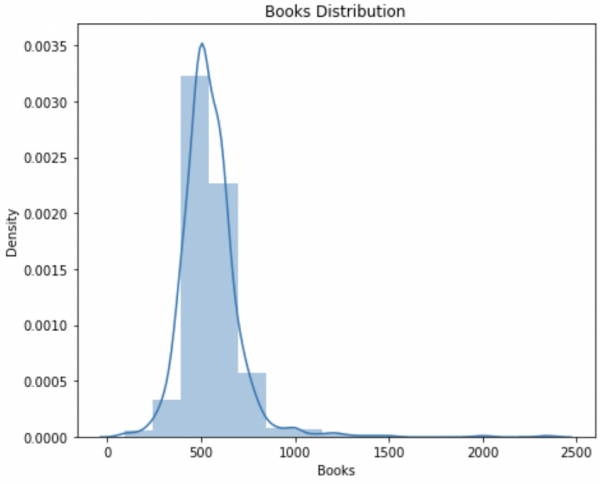
Room Board:

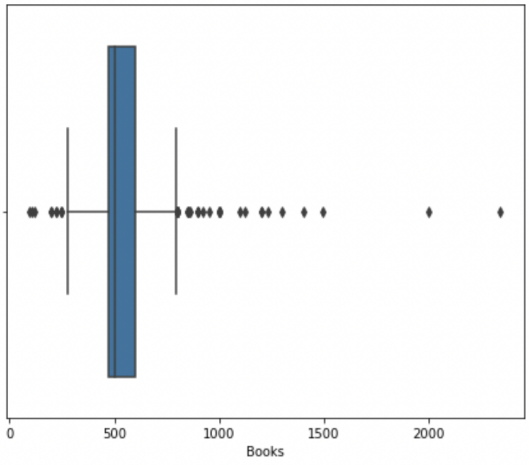




Room Board has very few outliers as per the box plot. Data is normally distributed as per the above dist plot.

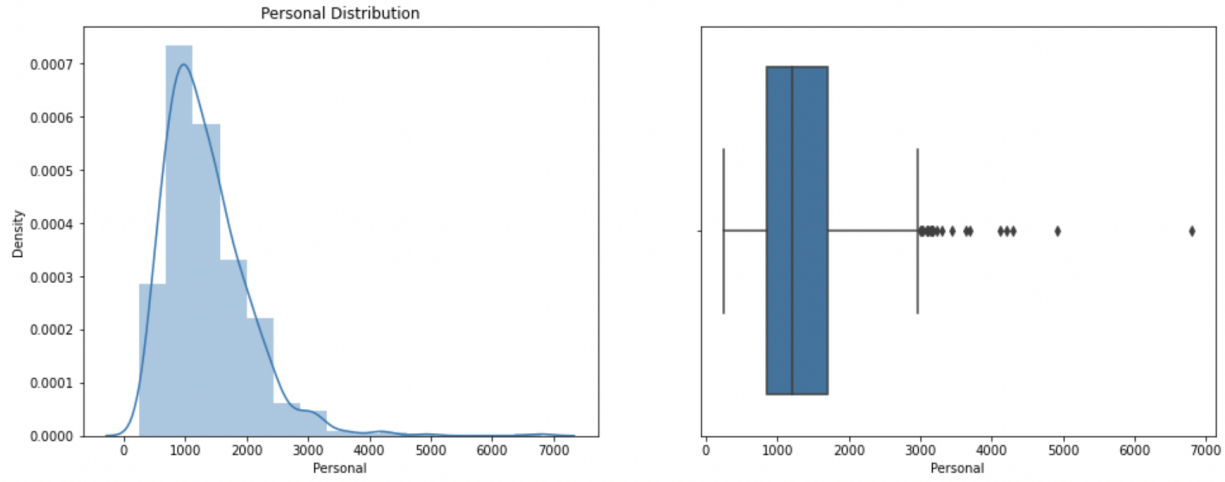
Books:





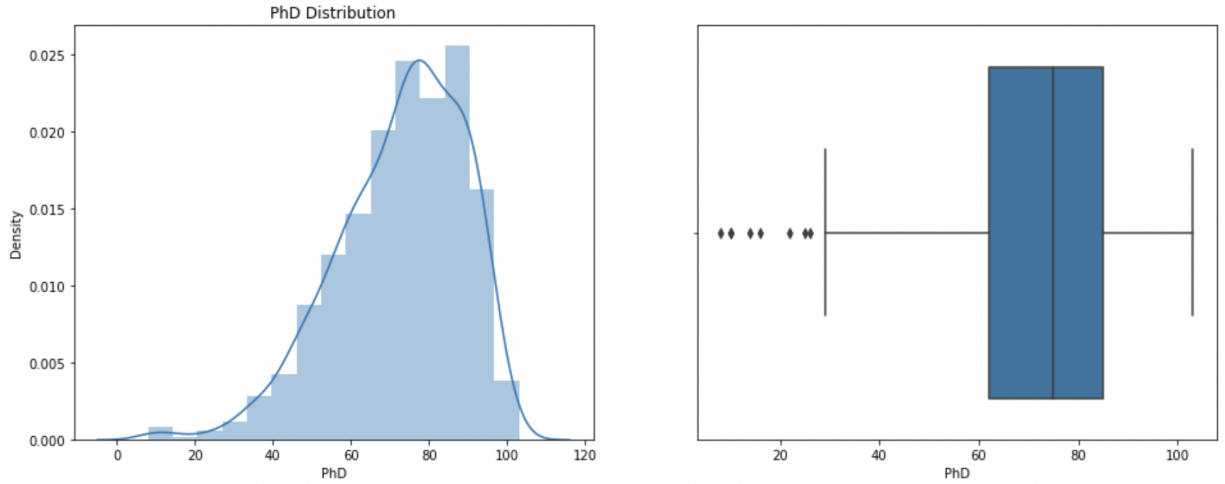
The boxplot of Books has outliers.

Personal:



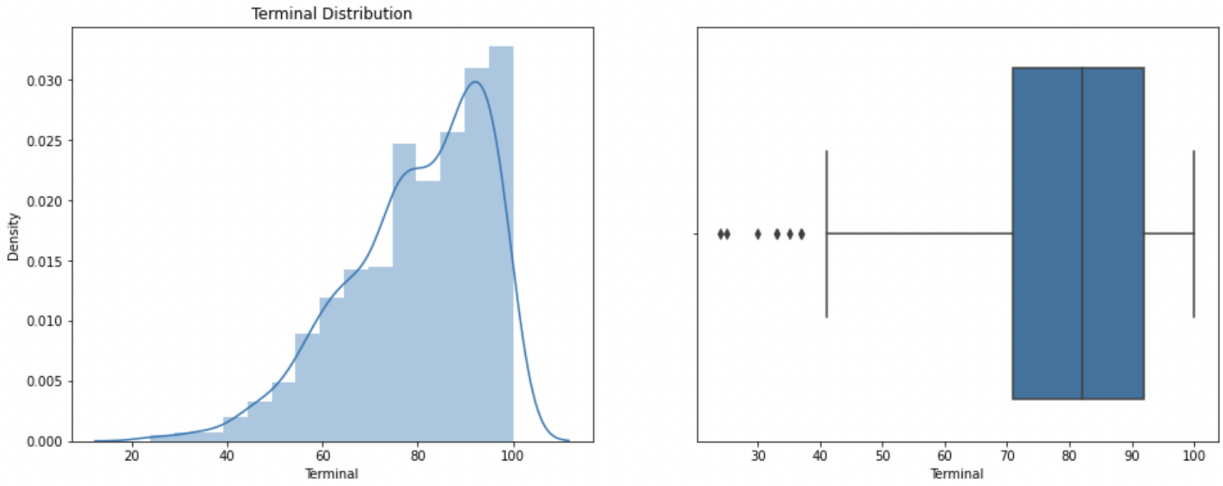
Personal variable has outliers as per the boxplot. The distribution seems to be right skewed. As per the box plot, we can say that some students personal expense is way higher than others.

PhD:



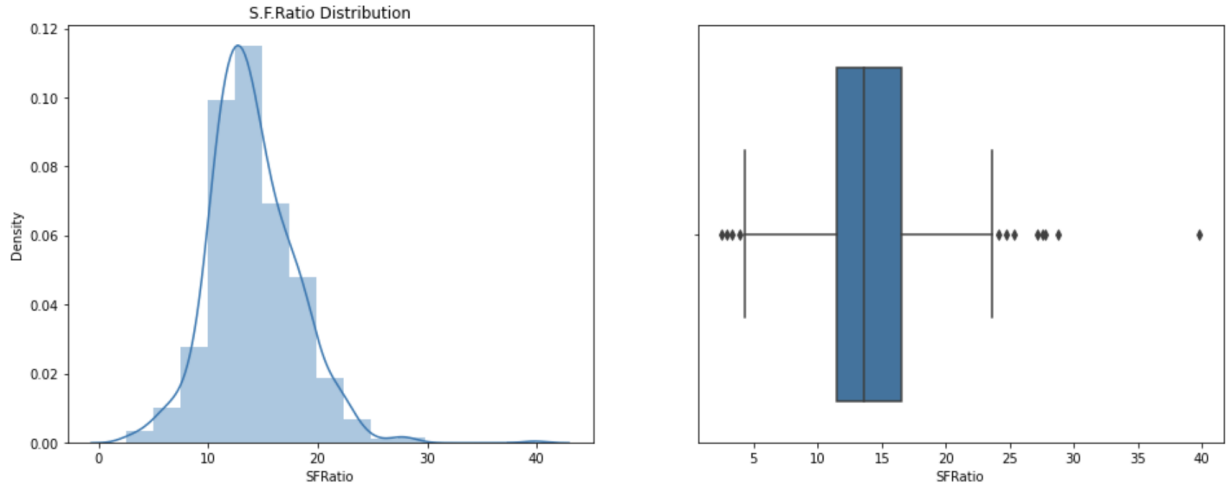
The boxplot of PhD has outliers. The distribution seems to be negatively skewed.

Terminal:



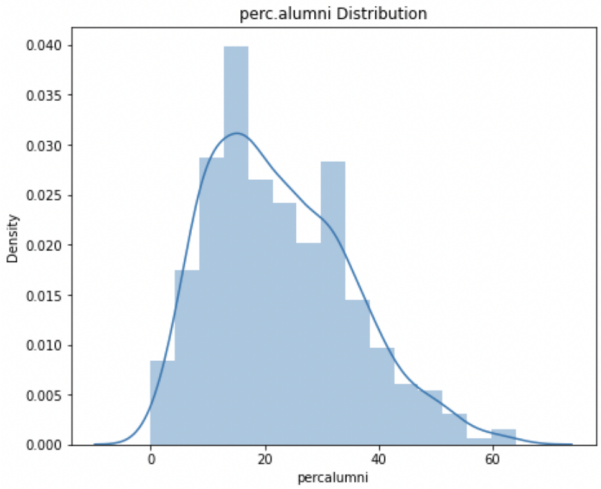
Terminal variable has outliers as per the boxplot. Data is left skewed as per the dist plot.

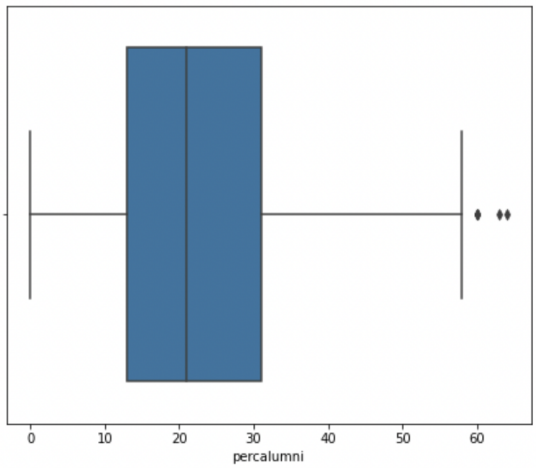
SF Ratio:



SF Ratio has outliers as per the box plot. We can say that the distribution is almost normally distributed.

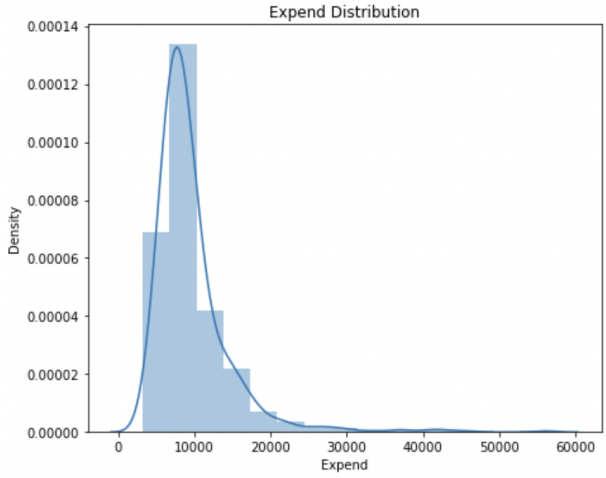
Perc Alumni:

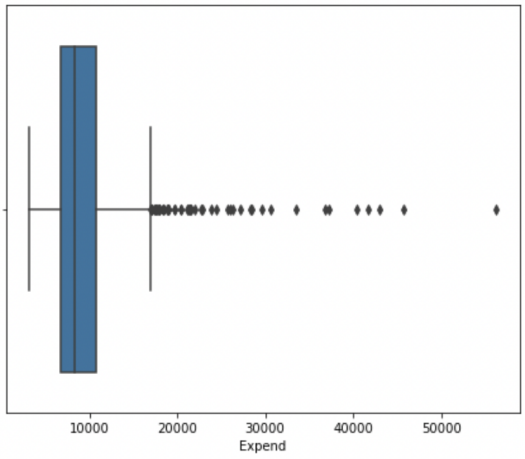




The percentage of alumni has outliers in the data. The distribution is almost normally distributed.

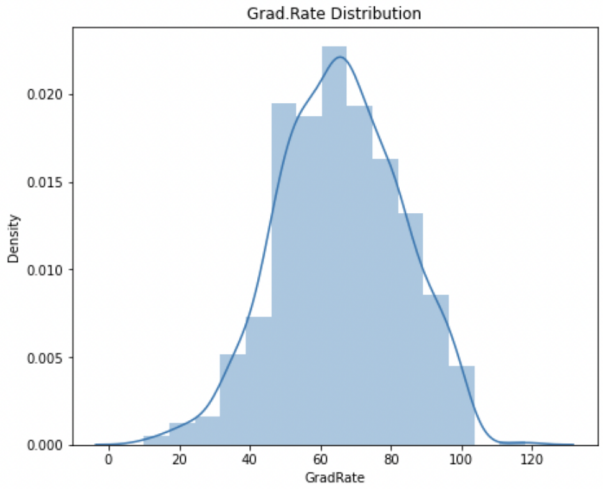
Expend:

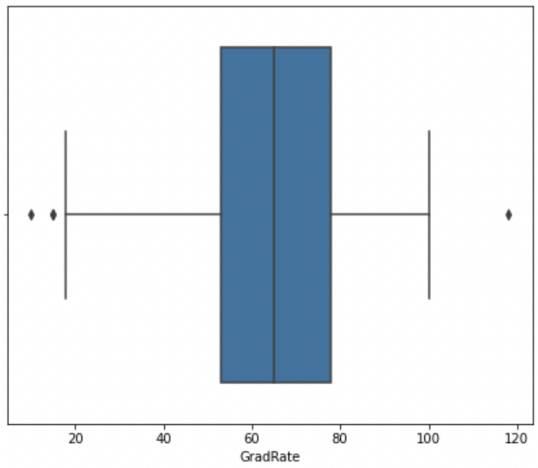




Expend variable has outliers in the dataset as per the boxplot. The distribution of data is a bit right skewed.

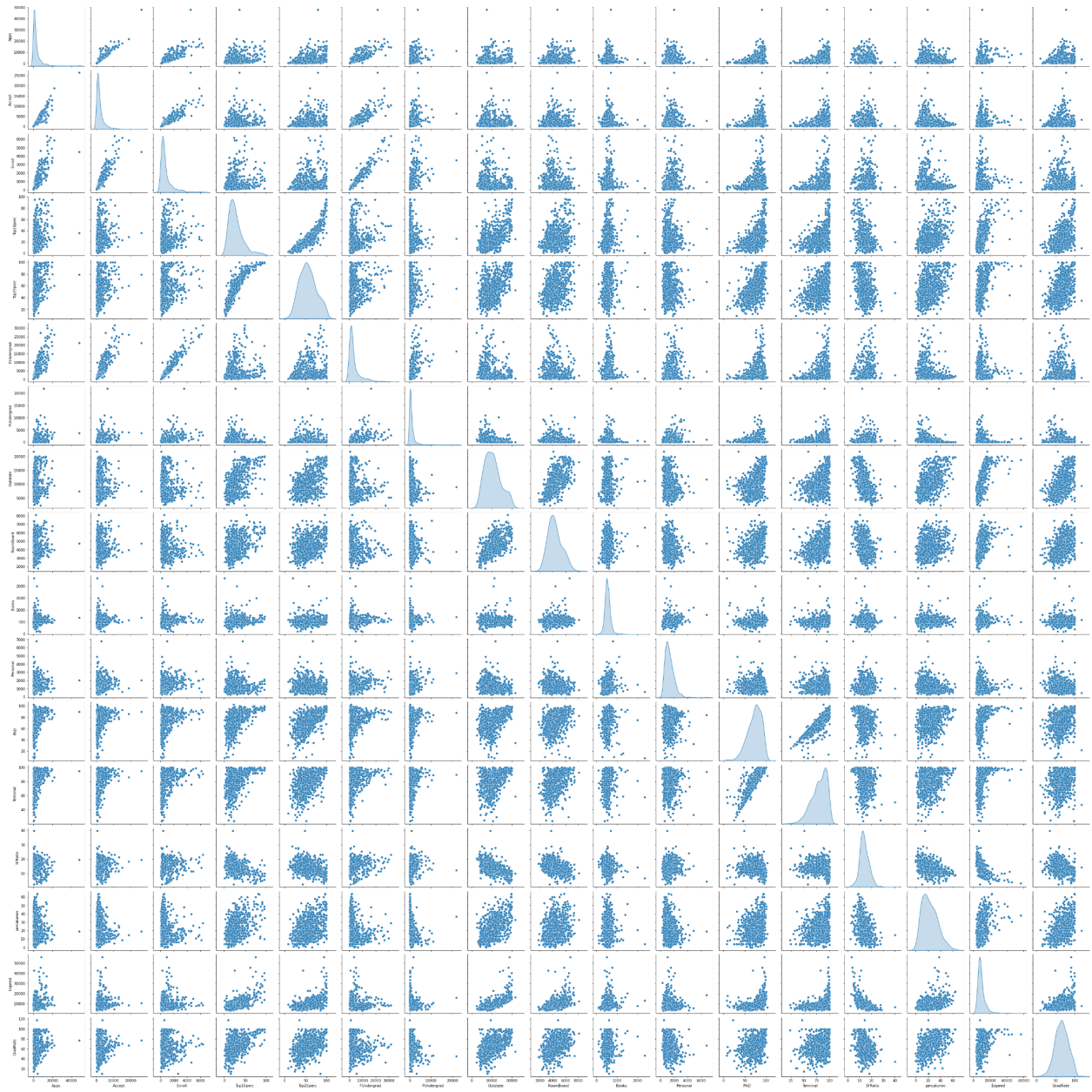
Grad Rate:





The boxplot for Grad Rate show very few outliers. Data is normally distributed.

Multivariate Analysis:

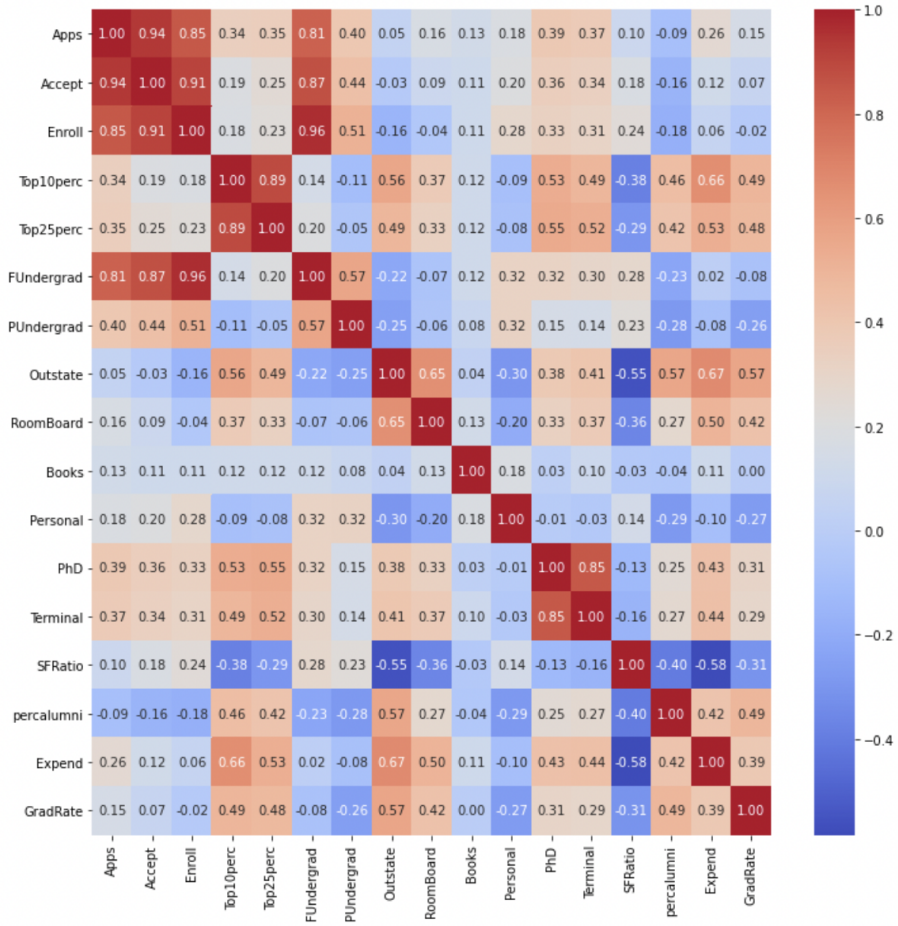


Above pair plot helps us to understand the relationship between all the numerical values in the dataset. By comparing all the variables, we can understand the patterns in the data.

Correlation:

	Apps	Accept	Enroll	Top10perc	Top25perc	FUndergrad	PUndergrad	Outstate	RoomBoard	Books	Personal	PhD	Terminal
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
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
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
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
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
FUndergrad	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
PUndergrad	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
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
RoomBoard	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
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
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
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
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
SFRatio	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
percalumni	-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
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
GradRate	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

Heatmap:



Above heatmap gives us the correlation between two numerical variables.
Application variable is highly positively correlated with application accepted, students enrolled and full time graduates. We can get some information on this when a student submits application and it is accepted and also whether the student is enrolled as a full time graduate.
There is a negative correlation between application and percentage of alumni. We can conclude that not all students who enrolled into the college are part of the alumni of college.
The application with top 10,25 of higher secondary class, outstate, room board, books, personal, PhD, terminal, S.F Ratio, expenditure, graduation ratio are positively correlated.

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

Below scaling is performed after dropping the 'Names' variable as it is a categorical variable. Dataset with only numerical values is used.

The dataset contains 17 numerical columns with different scales.
Application, accepted application, enrolled full time graduates, part time graduates, outstate: There variables represent number of students.
Top10 percent and Top25 percent are given in percentages.
Room board, books, personal: These variables represent cost/money
PhD, sf ratio, percentage of alumni, graduation rate: Represent percentage values
As per the above inferences, some of the variables have different scales.
Hence, scaling is necessary for PCA in this case.

Using z-score to scale the variables in the data. Z-score gives us how many standard deviation the point is away from the mean. Below is the scaled data:

	0	1	2	3	4	5	6	7	8	9	...	767	768	769	
Apps	-0.346882	-0.210884	-0.406866	-0.668261	-0.726176	-0.624307	-0.684808	-0.285088	-0.507700	-0.625600	...	-0.176238	-0.736777	-0.264404	-0.
Accept	-0.321205	-0.038703	-0.376318	-0.681682	-0.764555	-0.628611	-0.685356	-0.121984	-0.481644	-0.620854	...	-0.087284	-0.771903	-0.114227	-0.
Enroll	-0.063509	-0.288584	-0.478121	-0.692427	-0.780735	-0.669812	-0.729043	-0.313353	-0.595505	-0.654735	...	-0.011817	-0.759196	-0.220739	-0.
Top10perc	-0.258583	-0.655656	-0.315307	1.840231	-0.655656	0.592287	-0.598931	0.535563	0.138490	-0.372032	...	-0.201858	-0.598931	0.819186	1.
Top25perc	-0.191827	-1.353911	-0.292878	1.677612	-0.596031	0.313426	-0.545505	0.616579	0.363952	-0.596031	...	0.262901	-0.747607	0.616579	1.
FUndergrad	-0.168116	-0.209788	-0.549565	-0.658079	-0.711924	-0.623421	-0.677472	-0.434450	-0.562562	-0.598459	...	-0.062903	-0.705116	-0.354818	-0.
PUndergrad	-0.209207	0.244307	-0.497090	-0.520752	0.009005	-0.535212	-0.410988	-0.541127	-0.361036	-0.510893	...	-0.121791	-0.547700	-0.467513	-0.
Outstate	-0.746356	0.457496	0.201305	0.626633	-0.716508	0.760947	0.708713	0.852479	1.282036	0.006798	...	-1.005035	-0.333464	1.369837	0.
RoomBoard	-0.964905	1.909208	-0.554317	0.996791	-0.216723	-0.932970	1.243144	0.427443	0.038754	-0.891911	...	-0.880962	-0.599938	0.042403	-0.
Books	-0.602312	1.215880	-0.905344	-0.602312	1.518912	-0.299280	-0.299280	-0.602312	-1.511408	0.670422	...	0.185571	-0.299280	-0.905344	0.
Personal	1.270045	0.235515	-0.259582	-0.688173	0.235515	-0.983753	0.235515	-0.725120	-1.242385	0.678885	...	1.196150	0.087725	-0.799015	0.
PhD	-0.163028	-2.675646	-1.204845	1.185206	0.204672	-0.346878	1.062639	1.001356	0.388522	-2.001529	...	-0.101745	-1.511262	0.572372	1.
Terminal	-0.115729	-3.378176	-0.931341	1.175657	-0.523535	-0.455567	0.903786	1.379560	0.292077	-2.630532	...	0.020207	-2.154758	1.039721	0.
SFRatio	1.013776	-0.477704	-0.300749	-1.615274	-0.553542	-1.185526	-0.654660	-0.098515	-0.705218	-0.654660	...	-0.326029	-1.413040	-0.326029	0.
percalumni	-0.867574	-0.544572	0.585935	1.151188	-1.675079	-0.948325	0.262933	1.151188	0.020681	-0.625323	...	0.262933	0.262933	0.505184	1.
Expend	-0.501910	0.166110	-0.177290	1.792851	0.241803	0.012806	-0.153145	0.350074	0.380160	-0.128233	...	-0.561698	-0.134173	0.144456	-0.
GradRate	-0.318252	-0.551262	-0.667767	-0.376504	-2.939613	-0.609514	-0.143495	0.439030	0.846798	-0.784272	...	-0.376504	-0.900777	0.730293	0.

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

Covariance matrix on scaled dataset:

```
Covariance Matrix
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.81554018
   0.3987775   0.05022367  0.16515151  0.13272942  0.17896117  0.39120081
   0.36996762  0.09575627 -0.09034216  0.2599265   0.14694372]
 [ 0.94466636  1.00128866  0.91281145  0.19269493  0.24779465  0.87534985
   0.44183938 -0.02578774  0.09101577  0.11367165  0.20124767  0.35621633
   0.3380184   0.17645611 -0.16019604  0.12487773  0.06739929]
 [ 0.84791332  0.91281145  1.00128866  0.18152715  0.2270373   0.96588274
   0.51372977 -0.1556777  -0.04028353  0.11285614  0.28129148  0.33189629
   0.30867133  0.23757707 -0.18102711  0.06425192 -0.02236983]
 [ 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]]
```

Correlation matrix: It remains same for before scaling and after scaling

	Apps	Accept	Enroll	Top10perc	Top25perc	FUndergrad	PUndergrad	Outstate	RoomBoard	Books	Personal	PhD	Terminal	SFRatio	percalumni	Expend	GradRate
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
FUndergrad	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
PUndergrad	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
RoomBoard	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
SFRatio	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
percalumni	-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
GradRate	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 and Correlation determines the dependency between two variables.

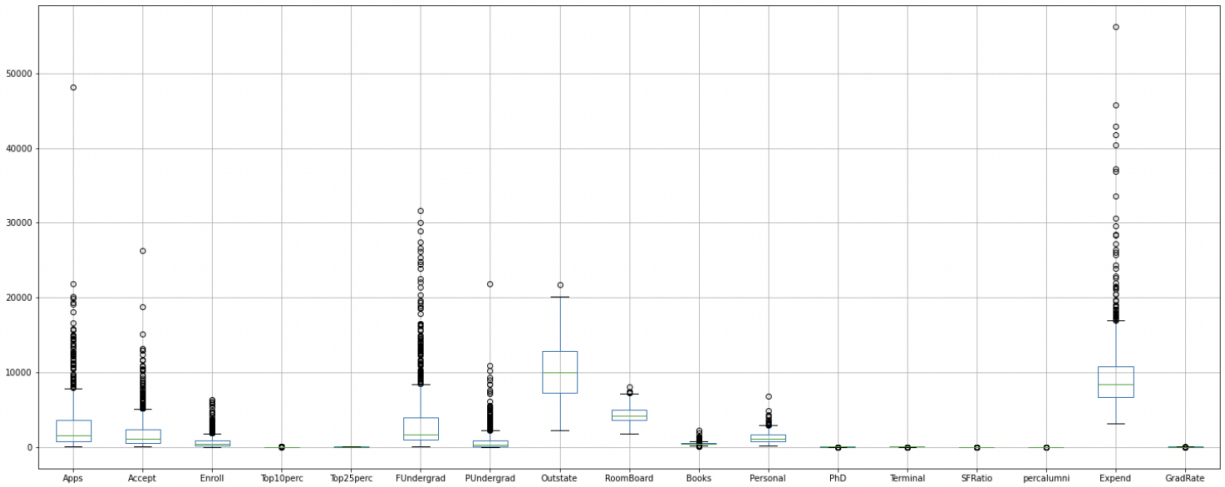
Covariance matrix indicates direction of the linear relationship between the variables. Correlation measures the strength and direction of the linear relationship between two variables. To check if it is positively correlated or negatively correlated.

From the above correlation matrix, we see that few variables are highly positively correlated and few are highly negatively correlated.

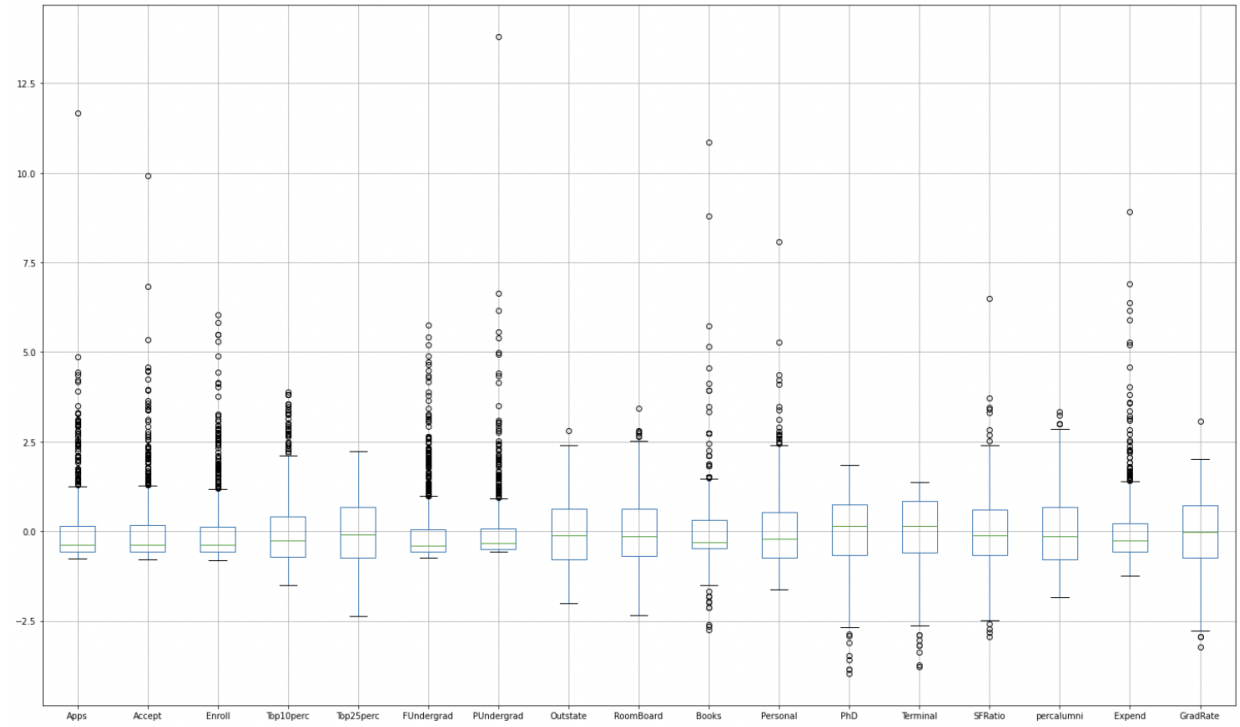
Application, acceptance, enrollment, full-time graduates, top 10 percent and top 25 percent are highly positively correlated.

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]

Before Scaling:



After Scaling:



There are outliers in the box plot before scaling the data. After scaling the data, outliers are still present.
Scaling did not remove outliers. Scaling scales the values on a z-score distribution.
Here, we can wish to treat outliers too. If we wish to treat outliers, we can take 3 standard deviations as outliers or impute them with IQR values

5. Extract the eigenvalues and eigenvectors.

Eigen Vectors:

	eigenvec_00	eigenvec_01	eigenvec_02	eigenvec_03	eigenvec_04	eigenvec_05	eigenvec_06	eigenvec_07	eigenvec_08	eigenvec_09	eigenvec_10	eigenvec
0	-0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	0.042486	-0.103090	0.090227	-0.052510	-0.043046	-0.024
1	-0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	0.012950	-0.056271	0.177865	-0.041140	0.058406	0.145
2	-0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	0.027693	0.058662	0.128561	-0.034488	0.069399	-0.011
3	-0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	0.161332	-0.122678	-0.341100	-0.064026	0.008105	-0.038
4	-0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	0.118486	-0.102492	-0.403712	-0.014549	0.273128	0.089
5	-0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	0.025076	0.078890	0.059442	-0.020847	0.081158	-0.056
6	-0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	-0.061042	0.570784	-0.560673	0.223106	-0.100693	0.063
7	-0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	-0.108529	0.009846	0.004573	-0.186675	-0.143221	0.823
8	-0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	-0.209744	-0.221453	-0.275023	-0.298324	0.359322	-0.354
9	-0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	0.149692	0.213293	0.133663	0.082029	-0.031940	0.028
10	0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	-0.633790	-0.232661	0.094469	-0.136028	0.018578	0.039
11	-0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	0.001096	-0.077040	0.185182	0.123452	-0.040372	-0.023
12	-0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	0.028477	-0.012161	0.254938	0.088578	0.058973	-0.016
13	0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	-0.219259	-0.083605	-0.274544	-0.472045	-0.445001	0.011
14	-0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	-0.243321	0.678524	0.255335	-0.423000	0.130728	-0.182
15	-0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	0.226584	-0.054159	0.049139	-0.132286	-0.692089	-0.325
16	-0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	-0.559944	-0.005336	-0.041904	0.590271	-0.219839	-0.122

Eigen Values:

```
array([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

Data after scaling:

	Apps	Accept	Enroll	Top10perc	Top25perc	FUndergrad	PUndergrad	Outstate	RoomBoard	Books	Personal	PhD	Terminal	SFRati
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	-0.47770
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	-0.30074
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	-0.55354

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]])
```

Eigen Vectors :

	eigenvec_00	eigenvec_01	eigenvec_02	eigenvec_03	eigenvec_04	eigenvec_05	eigenvec_06	eigenvec_07	eigenvec_08	eigenvec_09	eigenvec_10	eigenvec_11	eigenvec_12	eigenvec_13	eigenvec_14	eigenvec_15	eigenvec_16
0	-0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	0.042486	-0.103090	0.090227	-0.052510	-0.043046	-0.024071	0.595831	-0.080633	-0.133406	0.459139	-0.358970
1	-0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	0.012950	-0.056271	0.177865	-0.041140	0.058406	0.145102	0.292642	-0.033467	0.145498	-0.518569	0.543427
2	-0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	0.027693	0.058662	0.128561	-0.034488	0.069399	-0.011143	-0.444638	0.085697	-0.029590	-0.404318	-0.609651
3	-0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	0.161332	-0.122678	-0.341100	-0.064026	0.008105	-0.038554	0.001023	0.107828	-0.697723	-0.148739	0.144986
4	-0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	0.118486	-0.102492	-0.403712	-0.014549	0.273128	0.089352	0.021884	-0.151742	0.617275	0.051868	-0.080348
5	-0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	0.025076	0.078890	0.059442	-0.020847	0.081158	-0.056177	-0.523622	0.056373	-0.009916	0.560363	0.414705
6	-0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	-0.061042	0.570784	-0.560673	0.223106	-0.100693	0.063536	0.125998	-0.019286	-0.020952	-0.052731	-0.009018
7	-0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	-0.108529	0.009846	0.004573	-0.186675	-0.143221	0.823444	-0.141856	0.034012	-0.038354	0.101595	-0.050900
8	-0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	-0.209744	-0.221453	-0.275023	-0.298324	0.359322	-0.354560	-0.069749	0.058429	-0.003402	-0.025929	-0.001146
9	-0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	0.149692	0.213293	0.133663	0.082029	-0.031940	0.028159	0.011438	0.066849	0.009439	0.002883	-0.000773
10	0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	-0.633790	-0.232661	0.094469	-0.136028	0.018578	0.039264	0.039455	-0.027529	0.003090	-0.012890	0.001114
11	-0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	0.001096	-0.077040	0.185182	0.123452	-0.040372	-0.023222	0.127696	0.691126	0.112056	0.029808	-0.013813
12	-0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	0.028477	-0.012161	0.254938	0.088578	0.058973	-0.016485	-0.058313	-0.671009	-0.158910	-0.027076	-0.006209
13	0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	-0.219259	-0.083605	-0.274544	-0.472045	-0.445001	0.011026	-0.017715	-0.041374	0.020899	-0.021248	0.002222
14	-0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	-0.243321	0.678524	0.255335	-0.423000	0.130728	-0.182661	0.104088	0.027154	0.008418	0.003334	0.019187
15	-0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	0.226584	-0.054159	0.049139	-0.132286	-0.692089	-0.325982	-0.093746	-0.073123	0.227742	-0.043880	0.035310
16	-0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	-0.559944	-0.005336	-0.041904	0.590271	-0.219839	-0.122107	-0.069197	-0.036477	0.003394	-0.005008	0.013071

Eigen Values:

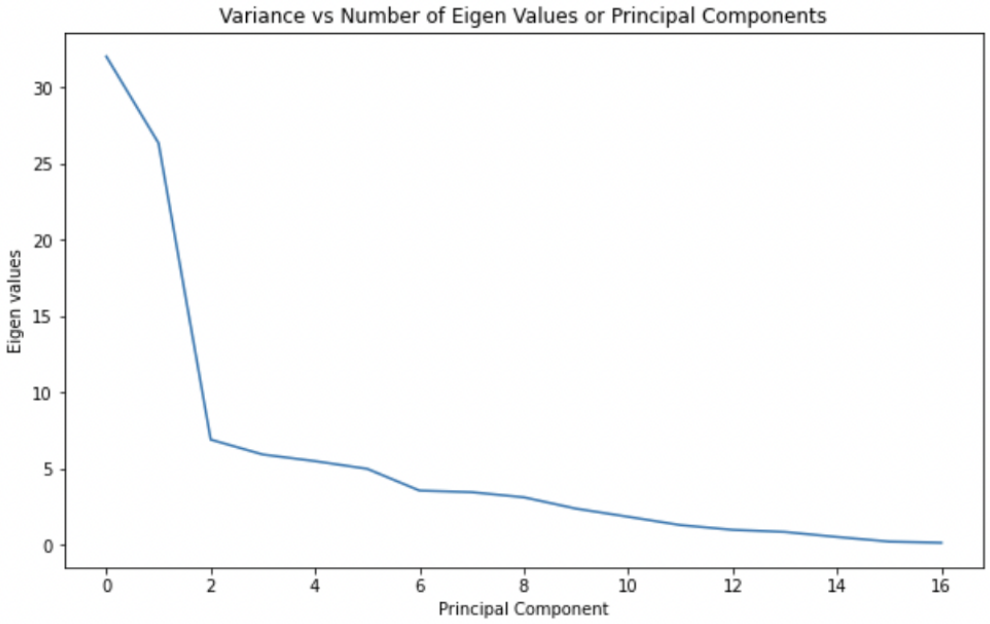
```
array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,
       0.84849117, 0.6057878 , 0.58787222, 0.53061262, 0.4043029 ,
       0.02302787, 0.03672545, 0.31344588, 0.08802464, 0.1439785 ,
       0.16779415, 0.22061096])
```

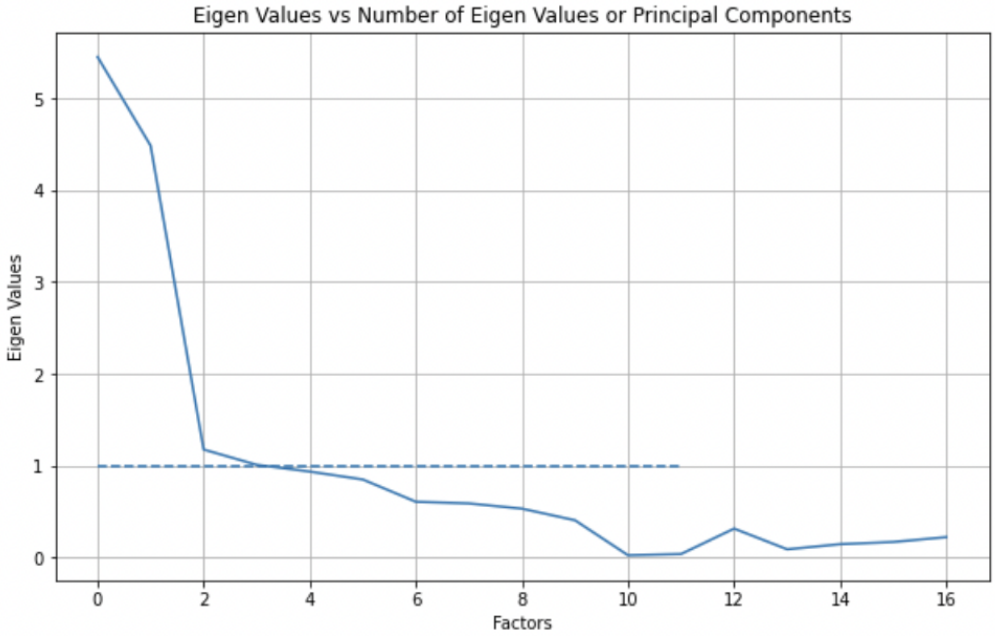
The variance explained by each of eigen values in order is :

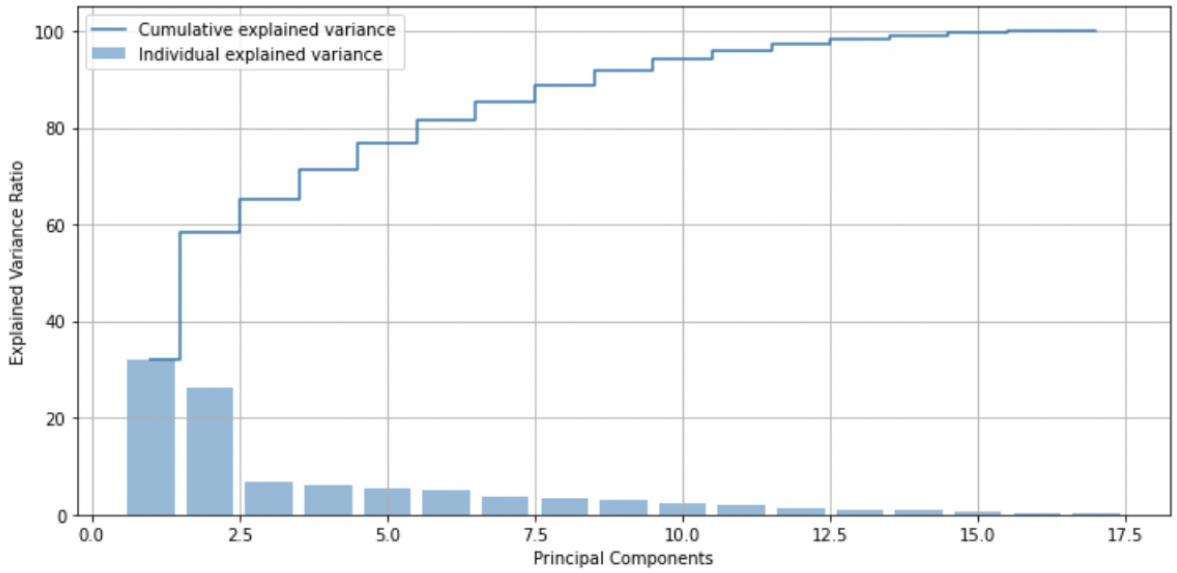
```
[32.02062819886914,
 26.340214436112475,
 6.900916554222497,
 5.92298922292629,
 5.488405110358494,
 4.9847009545574466,
 3.558871491746659,
 3.453621336999261,
 3.117233679821723,
 2.3751915258938006,
 1.8414263209386872,
 1.2960414001235336,
 0.9857541228001171,
 0.8458423350830021,
 0.5171255833731938,
 0.215754010072759,
 0.13528371610094883]
```

Cumulative Variance Explained :

```
array([ 32.0206282 ,  58.36084263,  65.26175919,  71.18474841,
        76.67315352,  81.65785448,  85.21672597,  88.67034731,
        91.78758099,  94.16277251,  96.00419883,  97.30024023,
        98.28599436,  99.13183669,  99.64896227,  99.86471628,
       100.        ])
```







Calculating the Principal Components:

	comp_00	comp_01	comp_02	comp_03	comp_04	comp_05	comp_06	comp_07	comp_08	comp_09	comp_10	comp_11	comp_12	comp_
Apps	-0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	0.042486	-0.103090	0.090227	-0.052510	-0.043046	-0.024071	0.595831	-0.0806
Accept	-0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	0.012950	-0.056271	0.177865	-0.041140	0.058406	0.145102	0.292642	-0.0334
Enroll	-0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	0.027693	0.058662	0.128561	-0.034488	0.069399	-0.011143	-0.444638	0.0856
Top10perc	-0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	0.161332	-0.122678	-0.341100	-0.064026	0.008105	-0.038554	0.001023	0.1078
Top25perc	-0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	0.118486	-0.102492	-0.403712	-0.014549	0.273128	0.089352	0.021884	-0.1517
FUndergrad	-0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	0.025076	0.078890	0.059442	-0.020847	0.081158	-0.056177	-0.523622	0.0563
PUndergrad	-0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	-0.061042	0.570784	-0.560673	0.223106	-0.100693	0.063536	0.125998	-0.0192
Outstate	-0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	-0.108529	0.009846	0.004573	-0.186675	-0.143221	0.823444	-0.141856	0.0340
RoomBoard	-0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	-0.209744	-0.221453	-0.275023	-0.298324	0.359322	-0.354560	-0.069749	0.0584
Books	-0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	0.149692	0.213293	0.133663	0.082029	-0.031940	0.028159	0.011438	0.0668
Personal	0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	-0.633790	-0.232661	0.094469	-0.136028	0.018578	0.039264	0.039455	-0.0275
PhD	-0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	0.001096	-0.077040	0.185182	0.123452	-0.040372	-0.023222	0.127696	0.6911
Terminal	-0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	0.028477	-0.012161	0.254938	0.088578	0.058973	-0.016485	-0.058313	-0.6710
SFRatio	0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	-0.219259	-0.083605	-0.274544	-0.472045	-0.445001	0.011026	-0.017715	-0.0413
percalumni	-0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	-0.243321	0.678524	0.255335	-0.423000	0.130728	-0.182661	0.104088	0.0271
Expend	-0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	0.226584	-0.054159	0.049139	-0.132286	-0.692089	-0.325982	-0.093746	-0.0731
GradRate	-0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	-0.559944	-0.005336	-0.041904	0.590271	-0.219839	-0.122107	-0.069197	-0.0364

	comp_00	comp_01	comp_02	comp_03	comp_04	comp_05	comp_06	comp_07	comp_08	comp_09	comp_10	comp_11	comp_12	comp_13	comp
0	1.592855	0.767334	-0.101074	-0.921749	-0.743975	-0.298306	-0.638443	-0.879386	-0.093084	-0.048593	-0.399747	0.089690	-0.052098	-0.180140	-0.001
1	2.192402	-0.578830	2.278798	3.588918	1.059997	-0.177137	-0.236753	0.046925	-1.113780	-0.965154	0.212509	-0.097239	-0.243518	0.744204	-0.103
2	1.430964	-1.092819	-0.438093	0.677241	-0.369613	-0.960592	0.248276	0.308740	0.105452	-0.640660	0.154993	0.344731	0.097551	-0.227527	0.022
3	-2.855557	-2.630612	0.141722	-1.295486	-0.183837	-1.059508	1.249356	-0.147694	-0.378997	-0.461244	0.420651	-0.687143	-0.075461	0.003380	0.073
4	2.212008	0.021631	2.387030	-1.114538	0.684451	0.004918	2.159220	-0.624413	0.160383	-0.363428	0.153339	0.050552	0.267207	0.614409	0.273

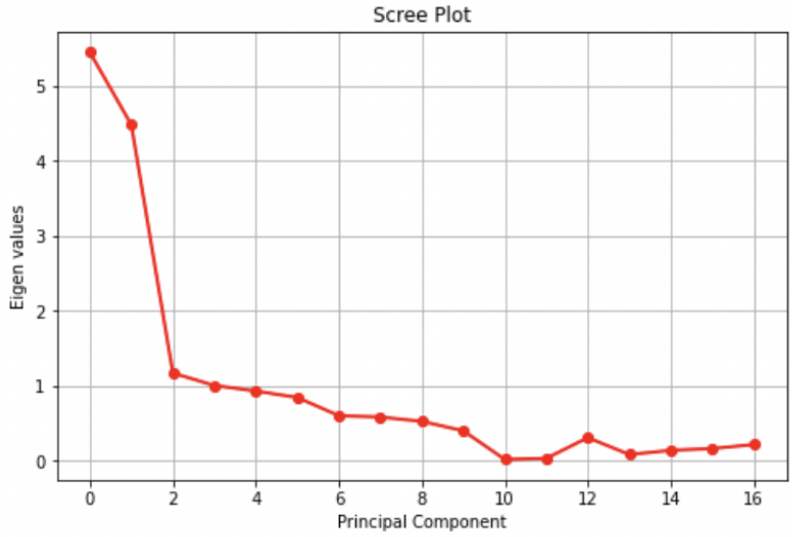
7. Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

The Linear equation of 1st principal component:
0.25 * Apps + 0.21 * Accept + 0.18 * Enroll + 0.35 * Top10perc + 0.34 * Top25perc + 0.15 * FUndergrad + 0.03 * PUndergrad + 0.29 * Outstate + 0.25 * RoomBoard + 0.06 * Books + -0.04 * Personal + 0.32 * PhD + 0.32 * Terminal + -0.18 * SFRatio + 0.21 * percalumni + 0.32 * Expend + 0.25 * GradRate +

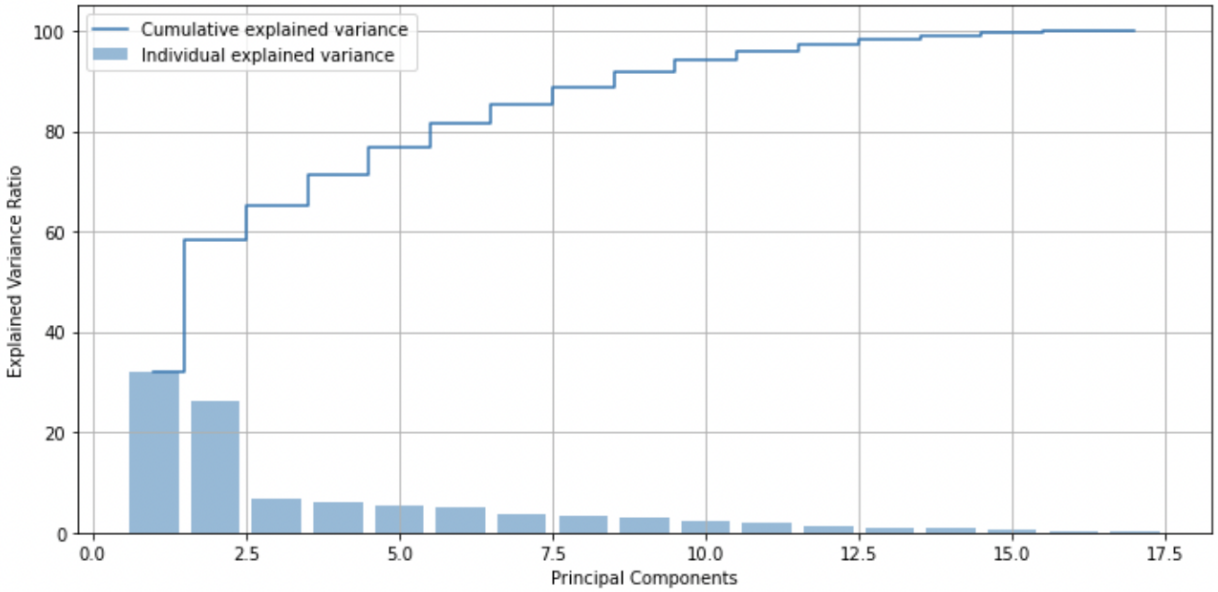
8. Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?

```
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.        ]
```

Adding the Eigen Values, we get the sum as 100.



From the above plot, we can see that the number of components that we can probably take is 2.
We also see that if we take 6 components the total amount of variance explained is 81.65785447704636 %



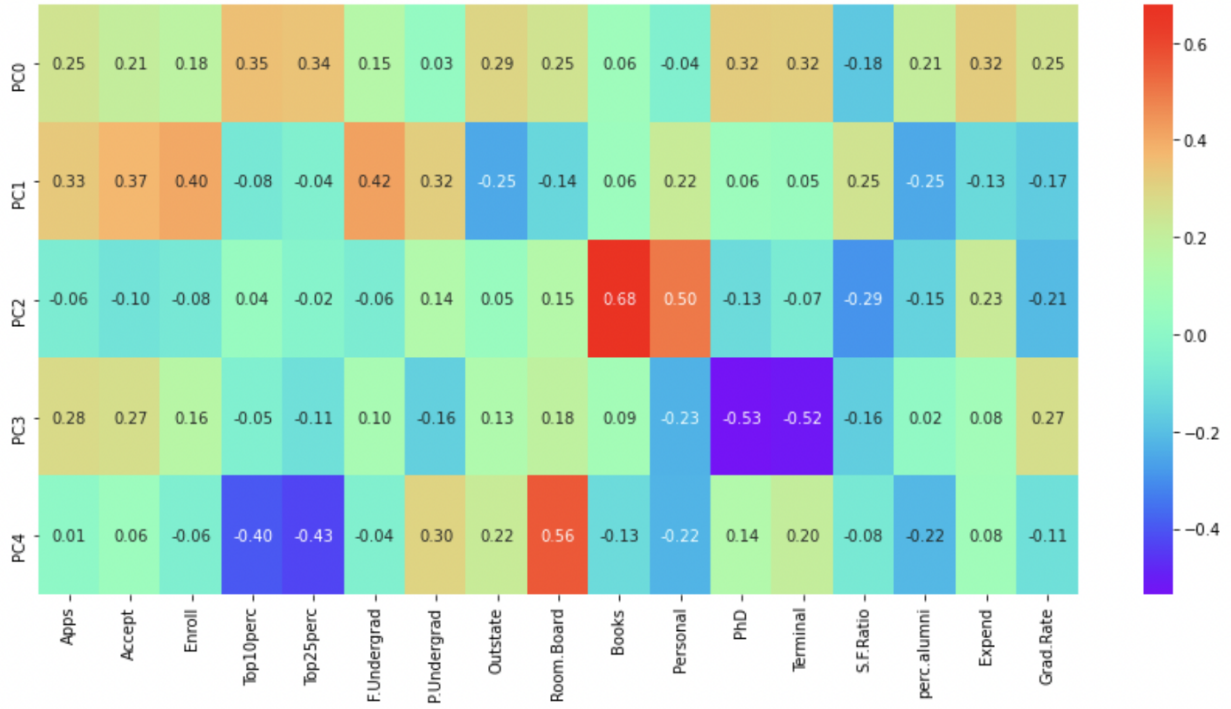
Principal Components:

```
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, -0.05569363, -0.39543435, -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],
       [-0.01623745,  0.0075347 , -0.042558  , -0.05269279,  0.03309159,
        -0.04345422, -0.19119858, -0.03000039,  0.16275545,  0.64105495,
        -0.331398  ,  0.09125552,  0.15492765,  0.48704587, -0.04734001,
        -0.29811862,  0.21616331]])
```

The first components explain 32.02% variance in data
The first 2 components explains 58.36% variance in data
The first 3 components explains 65.26% variance in data
The first 4 components explains 71.18% variance in data
The first 5 components explains 76.67% variance in data

The Eigen vectors or PC for this case study is 6. We can see how much each variable contributes to the principal components. PCA reduces multicollinearity.

	Apps	Accept	Enroll	Top10perc	Top25perc	FUndergrad	PUndergrad	Outstate	RoomBoard	Books	Personal	PhD	Terminal	SFRati
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0.064758	-0.042529	0.318313	0.317056	-0.17695
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809	0.056342	0.219929	0.058311	0.046429	0.24666
2	-0.063092	-0.101249	-0.082986	0.035056	-0.024148	-0.061393	0.139682	0.046599	0.148967	0.677412	0.499721	-0.127028	-0.066038	-0.28984
3	0.281311	0.267817	0.161827	-0.051547	-0.109767	0.100412	-0.158558	0.131291	0.184996	0.087089	-0.230711	-0.534725	-0.519443	-0.16118
4	0.005741	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919	-0.127289	-0.222311	0.140166	0.204720	-0.07938
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	0.48704



9. Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]

This business case study is about education dataset which contain the names of various colleges, which has various details of colleges and university. To understand more about the dataset we perform univariate analysis and multivariate analysis which gives us the understanding about the variables. From analysis we can understand the distribution of the dataset, skew, and patterns in the dataset. From multivariate analysis we can understand the correlation of variables. Inference of multivariate analysis shows we can understand multiple variables highly correlated with each other. The scaling helps the dataset to standardize the variable in one scale. Outliers are imputed using IQR values once the values are imputed we can perform PCA. The principal component analysis is used reduce the multicollinearity between the variables. Depending on the variance of the dataset we can reduce the PCA components. The PCA components for this business case is 6 where we could understand the maximum variance of the dataset. Using the components we can now understand the reduced multicollinearity in the dataset.