SMDM Project: Advanced Statistics

ANOVA, EDA AND PCA

Student’s Name: THILAK RAJ | Batch: 18 FEB 2022

INDEX:

Executive Summary (PROBLEM-1)

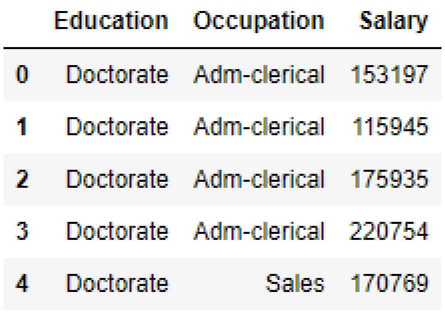
A Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csv] are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, 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.]

Data Description

1. Education: Education of the sample individual.
2. Occupation: Occupation of the sample individual.
3. Salary: Salary of the sample individual.

Sample of the dataset:

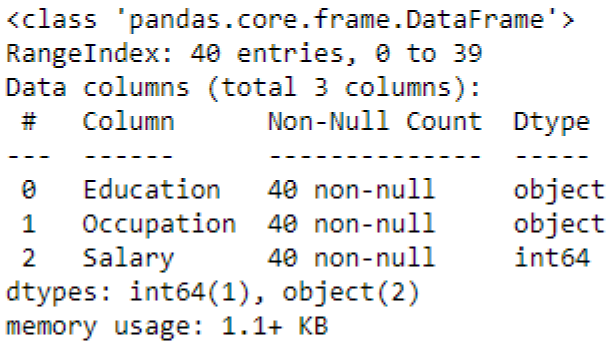


	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769

Dataset has 3 columns. 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. Dataset is consisting of 40 individual's data.

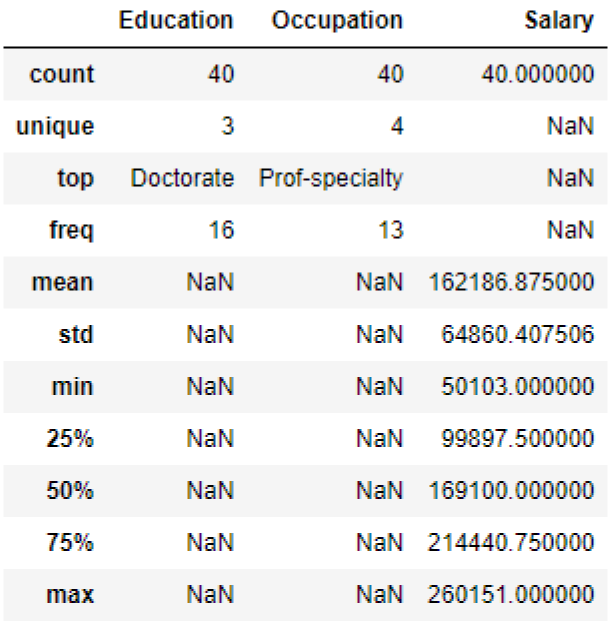
Exploratory Data Analysis:

Let us check the types of variables and Missing Values in the data frame.



```
<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
```

- We found that datatypes of Education and Occupation is Object type. We are going to change the datatype to Category before we are proceeding. The Salary column is of int64 datatype.
- We found No null values and no duplicate entries in the data frame.



	Education	Occupation	Salary
count	40	40	40.000000
unique	3	4	NaN
top	Doctorate	Prof-specialty	NaN
freq	16	13	NaN
mean	NaN	NaN	162186.875000
std	NaN	NaN	64860.407506
min	NaN	NaN	50103.000000
25%	NaN	NaN	99897.500000
50%	NaN	NaN	169100.000000
75%	NaN	NaN	214440.750000
max	NaN	NaN	260151.000000

- Education column has 40 entries with 3 unique values. The Doctorate level is the most frequently shown data in the column with 16 occurrences.
- Occupation column has 40 entries with 4 unique values. The Prof-speciality is the most frequently shown data in the column with 13 occurrences.
- Salary column has 40 entries with the mean value of 1662186.9. The maximum and minimum salary are 50103.0 and 260151.0 respectively.

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

One way ANOVA(Education)
Null Hypothesis $H0$: The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, HS-Grad).
Alternate Hypothesis $H1$: The mean salary is different in at least one category of education.

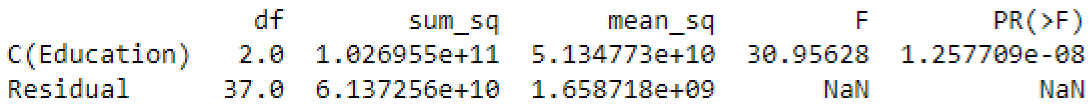
One way ANOVA(Occupation)
Null Hypothesis $H0$: The mean salary is the same across all the 4 categories of occupation (Prof-Specialty, Sales, Adm-clerical, Exec-Managerial).
Alternate Hypothesis $H1$: The mean salary is different in at least one category of occupation.

Where Alpha = 0.05

If the p-value is < 0.05, then we reject the null hypothesis.
If the p-value is >= 0.05, then we fail to reject the null hypothesis

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

Below is the result of one-way ANOVA for Education with respect the variable 'Salary':

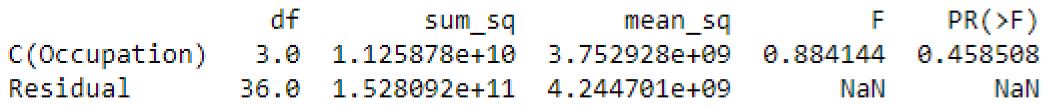


```
                 df        sum_sq       mean_sq          F        PR(>F)
C(Education)    2.0  1.026955e+11  5.134773e+10   30.95628  1.257709e-08
Residual       37.0  6.137256e+10  1.658718e+09        NaN           NaN
```

Since the p value = 1.257709e-08 is less than the significance level (alpha = 0.05), we can reject the null hypothesis and conclude that there is a significant difference in the mean salaries for at least one category of education.

Q 1.3 Perform one-way ANOVA for variable Occupation with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

Below is the result of one-way ANOVA for Occupation with respect the variable 'Salary':



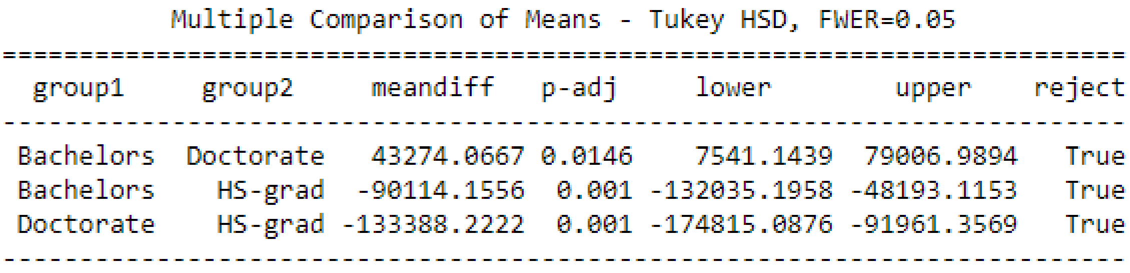
```
                  df        sum_sq       mean_sq         F    PR(>F)
C(Occupation)    3.0  1.125878e+10  3.752928e+09  0.884144  0.458508
Residual        36.0  1.528092e+11  4.244701e+09       NaN       NaN
```

Since the p value = 0.458508 is greater than the significance level (alpha = 0.05), we fail to reject the null hypothesis (i.e., we accept H0) and conclude that there is no significant difference in the mean salaries across the 4 categories of occupation.

Q1.4 If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result.

ANOVA tells us if our results or significant or not but does not tell us where the results are significant. But the interpretability of statistical significance is crucial to figure out in order to guide us. So, a Tukey Test allows us to interpret the statistical significance of our ANOVA test and find out which specific groups' means (compared with each other) are different. So, after performing each round of ANOVA, we can use a Tukey Test to find out where the statistical significance is occurring in our data.

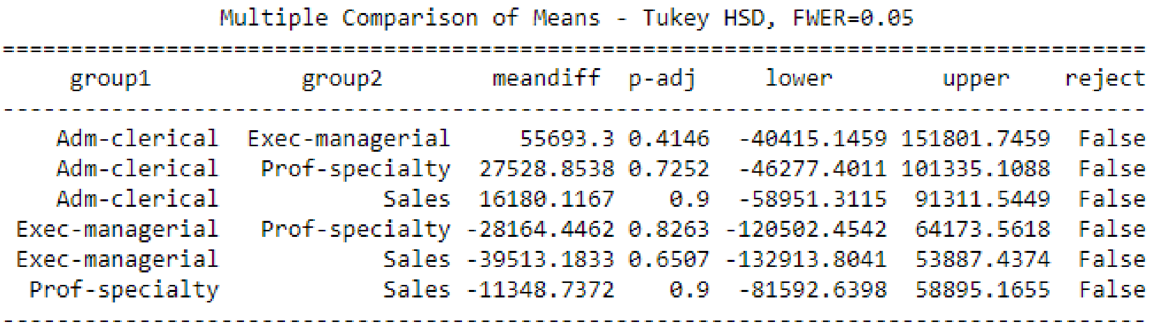
Using, the Tukey Honest Significant Difference test, we get the following table for the category education:



```
           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
------------------------------------------------------------------------
```

Since the p- values (p-adj in the table) are lesser than the significance level for all the three categories of education, this implies that the mean salaries across all categories of education are different.

Using, the Tukey Honest Significant Difference test, we get the following table for the category Occupation:

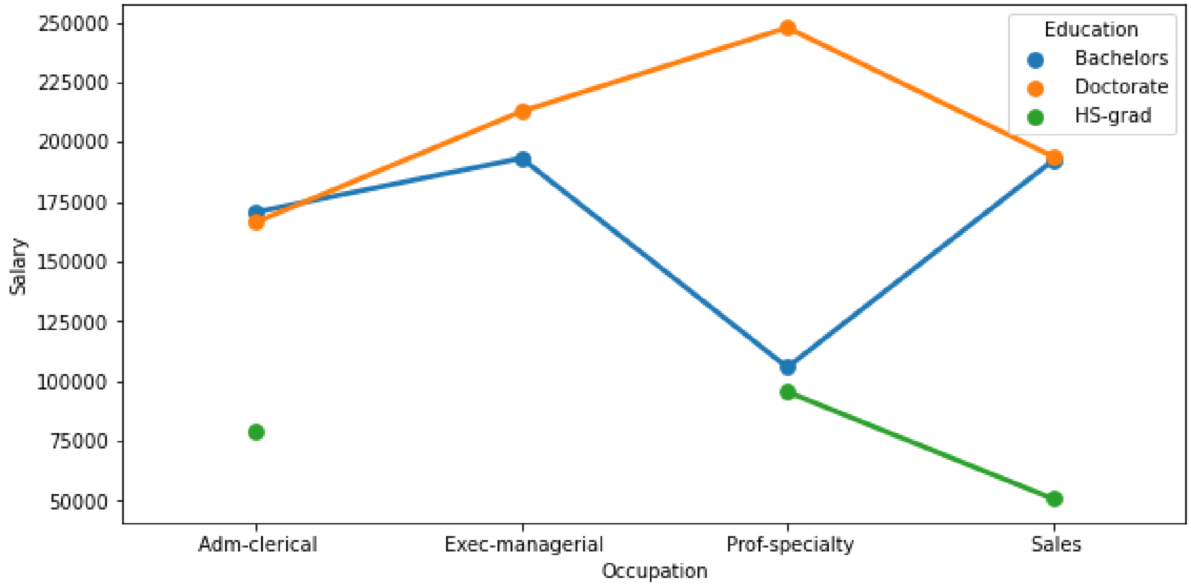


```
                    Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================================================
     group1           group2         meandiff  p-adj     lower         upper    reject
-------------------------------------------------------------------------------------
    Adm-clerical  Exec-managerial     55693.3 0.4146   -40415.1459 151801.7459  False
    Adm-clerical   Prof-specialty  27528.8538 0.7252   -46277.4011 101335.1088  False
    Adm-clerical            Sales  16180.1167    0.9   -58951.3115  91311.5449  False
 Exec-managerial   Prof-specialty -28164.4462 0.8263  -120502.4542  64173.5618  False
 Exec-managerial            Sales -39513.1833 0.6507  -132913.8041  53887.4374  False
  Prof-specialty            Sales -11348.7372    0.9   -81592.6398  58895.1655  False
-------------------------------------------------------------------------------------
```

For the category occupation, the Tukey Honest Significant Difference test has further confirmed that the mean salaries across all occupation classes are significantly same. The table below confirms the same, wherein we see that all p-values are greater than 0.05.

Q 1.5 What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.

Below is the pointplot to validate the interaction between two treatments.



Observation:

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

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

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

- Doctorates : are into higher salary brackets and mostly Prof-speciality roles or Exec-managerial roles or in sales profile, very few are doing Adm-clerical jobs
- Bachlores: fall in mid income rangeand found mostly working as an Exec-managers , Adm-clerks or into sales but very few are found in Prof- speciality profile.
- HS-grads : are in low income brackets, mostly doing Prof-speciality or Adm - clerical work and few are doing Sales but hardly any in Exec-managerial role.

Q1.6 Perform a two-way ANOVA based on the Education and Occupation (along with their interaction Education*Occupation) with the variable ‘Salary’. State the null and alternative hypotheses and state your results. How will you interpret this result?

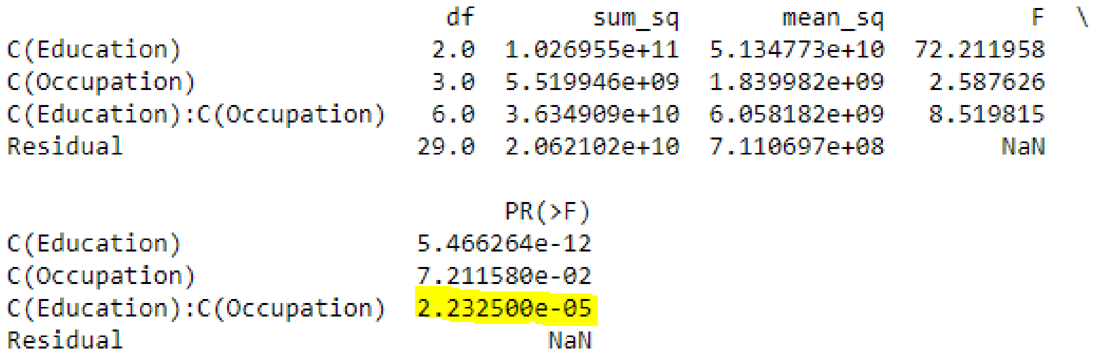
H0: The effect of the independent variable ‘education’ on the mean ‘salary’ does not depend on the effect of the other independent variable ‘occupation’ (i.e., there is no interaction effect between the 2 independent variables, education and occupation).

H1: There is an interaction effect between the independent variable ‘education’ and the independent variable ‘occupation’ on the mean salary.

Where Alpha = 0.05

- If the p-value is < 0.05, then we reject the null hypothesis.
- If the p-value is >= 0.05, then we fail to reject the null hypothesis.

Below is the Two-way ANOVA result:



```
                                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  
```

- **we see that there is a significant amount of interaction between the variables, Education and Occupation.**
- **As p value = 2.232500e-05 is lesser than the significance level (alpha = 0.05), we reject the null hypothesis. Thus, we see that there is an interaction effect between education and occupation on the mean salary.**

Q1.7 Explain the business implications of performing ANOVA for this case study.

Observation:

- **ANOVA is used in a business context to help manage salary by comparing the education to occupation in this case to help manage salary.**
- **From the ANOVA method and the interaction plot, we see that education combined with occupation results in higher and better salaries among the people.**
- **It is clearly seen that people with education as Doctorate draw the maximum salaries and people with education HS-grad earn the least. Thus, we can conclude that Salary is dependent**

on educational qualifications and occupation.

- **Though there is lesser significance of Occupation than education on Salary but at certain levels it impacts Salary.**
- **We can also take see that high salaries are offered to Bachelor's degree holders than Doctorates for few occupations. So, we can say that there are some shortcomings of dataset provided which reduces accuracy of the test and analysis done, as there can be few more other important variables which can impact salary such as years of experience, specialization, industry/domain etc.**

Executive Summary (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.

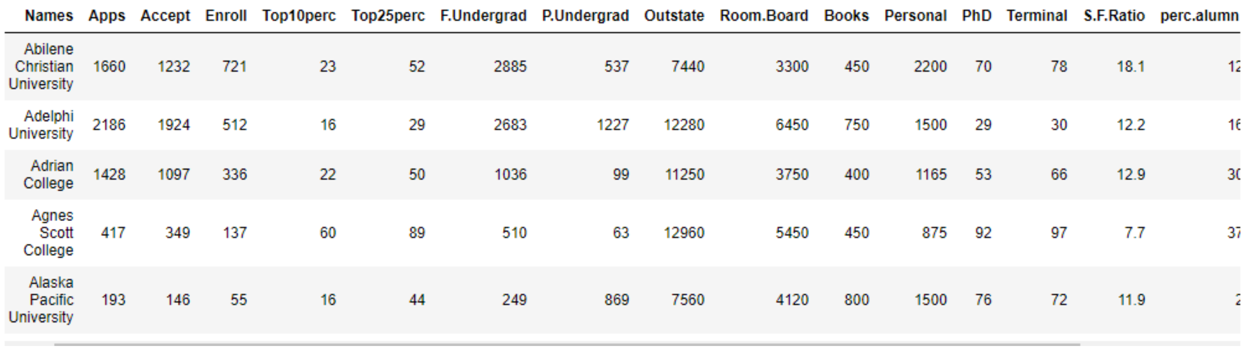
Introduction

The given dataset consists of data points of names of various university and college which has number of application received, accepted, and enrolled, percentage of new students from top 10% of higher secondary class, percentage of new students from top 25% of higher secondary class, Number of fulltime undergraduates, Number of parttime undergraduate students, Number of students for whom the particular college is out of state tuition, cost of room and board, estimated book costs for a student, estimated personal spending for a student, percentage of faculties with PHD, percentage of faculties with terminal degree, student/faculty ratio, percentage of alumni who donate, The instructional expenditure per student, Graduation Rate.

Data Description

1. Names: Names of various university and colleges
2. Apps: Number of applications received
3. Accept: Number of applications accepted
4. Enroll: Number of new students enrolled
5. Top10perc: Percentage of new students from top 10% of Higher Secondary class
6. Top25perc: Percentage of new students from top 25% of Higher Secondary class
7. F.Undergrad: Number of full-time undergraduate students
8. P.Undergrad: Number of part-time undergraduate students
9. Outstate: Number of students for whom the particular college or university is Out-of-state tuition
10. Room.Board: Cost of Room and board
11. Books: Estimated book costs for a student
12. Personal: Estimated personal spending for a student
13. PhD: Percentage of faculties with Ph.D.’s
14. Terminal: Percentage of faculties with terminal degree
15. S.F.Ratio: Student/faculty ratio
16. perc.alumni: Percentage of alumni who donate
17. Expend: The Instructional expenditure per student
18. Grad.Rate: Graduation rate

Sample of the dataset:

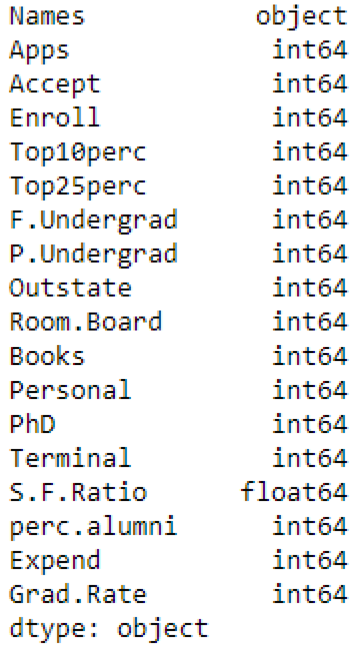


Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumn
Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12
Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16
Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30
Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37
Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2

Dataset contains 18 columns and 777 rows. Columns are the name of the university or college which has number of application received, accepted, and enrolled, percentage of new students from top 10% of higher secondary class, percentage of new students from top 25% of higher secondary class, Number of fulltime undergraduates, Number of parttime undergraduate students, Number of students for whom the particular college is out of state tuition, cost of room and board, estimated book costs for a student, estimated personal spending for a student, percentage of faculties with PHD, percentage of faculties with terminal degree, student/faculty ratio, percentage of alumni who donate,
The instructional expenditure per student, Graduation Rate.

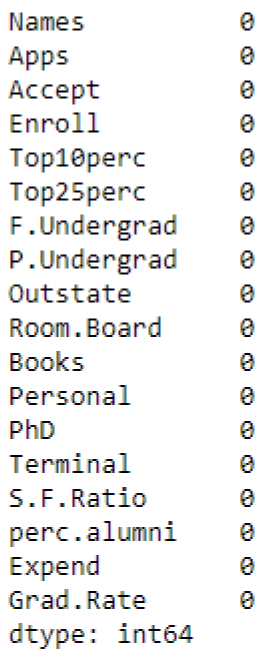
Exploratory Data Analysis:

Let us check the types of variables and Missing Values in the data frame.



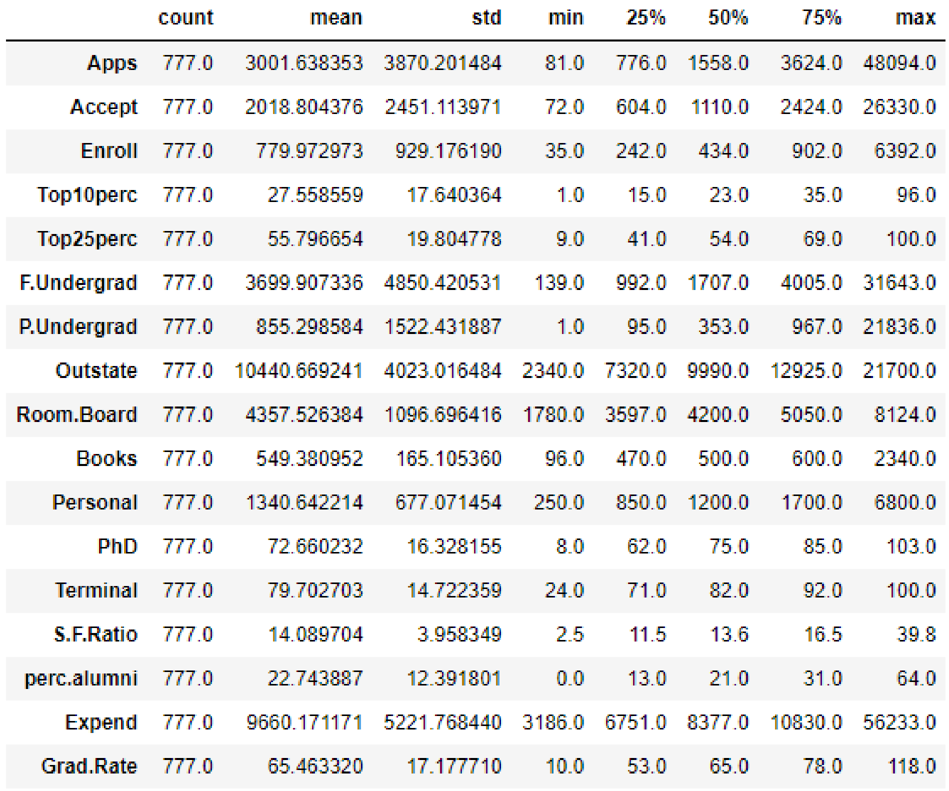
```
Names            object
Apps              int64
Accept            int64
Enroll            int64
Top10perc         int64
Top25perc         int64
F.Undergrad       int64
P.Undergrad       int64
Outstate          int64
Room.Board        int64
Books             int64
Personal          int64
PhD     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

Above data shows the datatypes of all the columns.



```
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
```

Above data shows the null values in each column.



	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

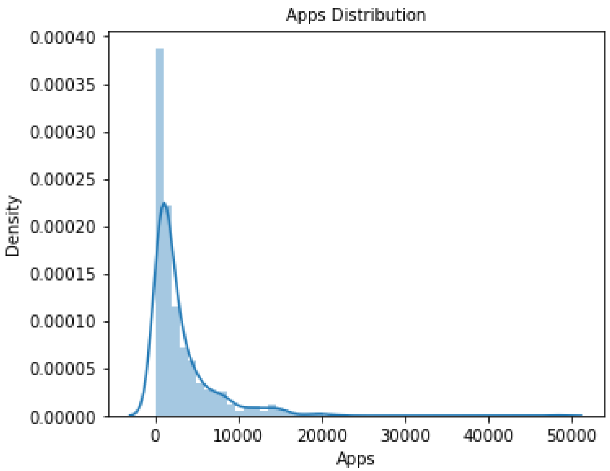
- **The shape of the dataset seems to be with 777 rows and 18 columns.**
- **All the columns seem to be integer or float values.**
- **The Names column alone is a categorical value.**
- **We also can see they are no duplicates in the dataset. T**
- **The entire dataset does not have missing values or null values.**

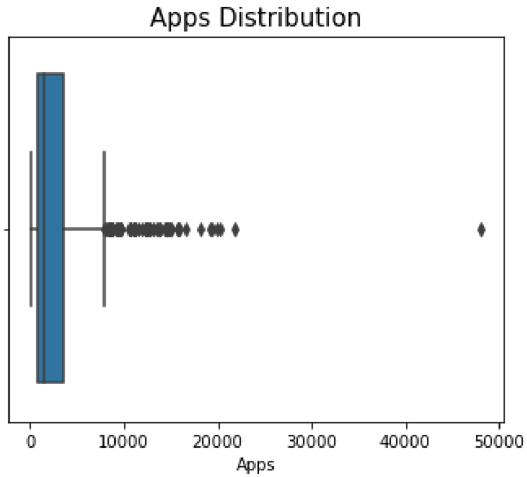
Q2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

UNIVARIATE ANALYSIS:

Helps us to understand the distribution of data in the dataset. With univariate analysis we can find patterns and we can summarize the data for

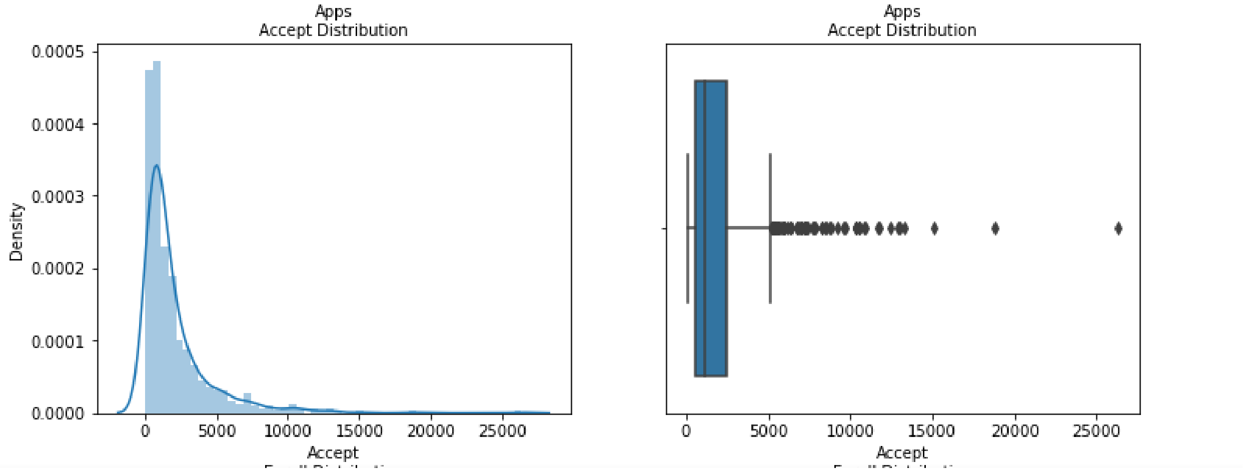
APPS:





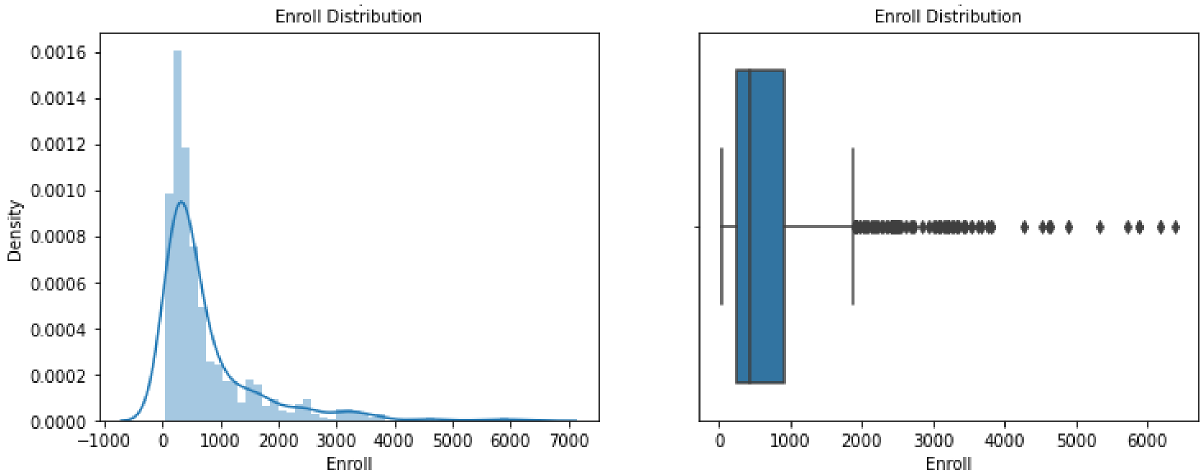
- **The Box plot of Apps variable seems to have outliers, the distribution of the data is skewed**
- **We could also understand that each college or university offers application in the range 80 and 48094.**
- **For univariate analysis of Apps we are using box plot and distplot to find information or patterns in the data.**
- **So we can clearly understand from the box plot we have outliers in the dataset.**

ACCEPT:



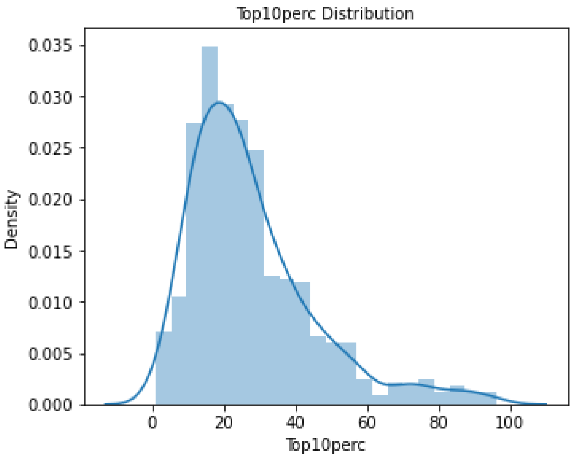
- **The accept variable seems to have outliers.**
- **The accept variable seems to be positively skewed.**

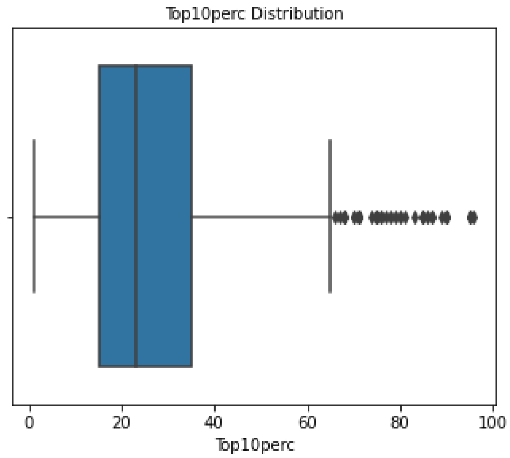
ENROLL:



- **The box plot of the Enroll variable also has outliers.**
- **The distribution of the data is positively skewed.**

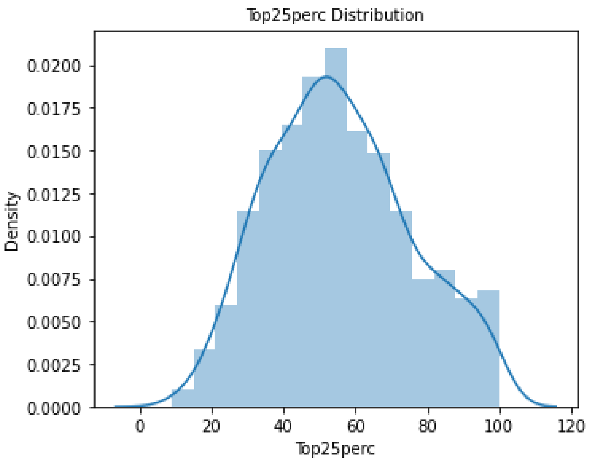
TOP10 PERC:

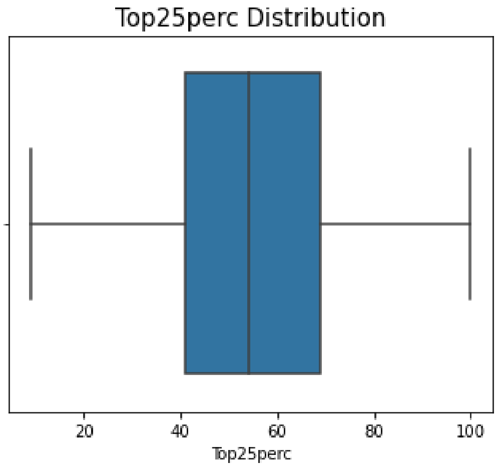




- **The box plot of the students from top 10 percentage of higher secondary class seems to have outliers.**
- **The distribution seems to be positively skewed.**
- **There is good amount of intake about 30 to 50 students from top 10 percentage of higher secondary class.**

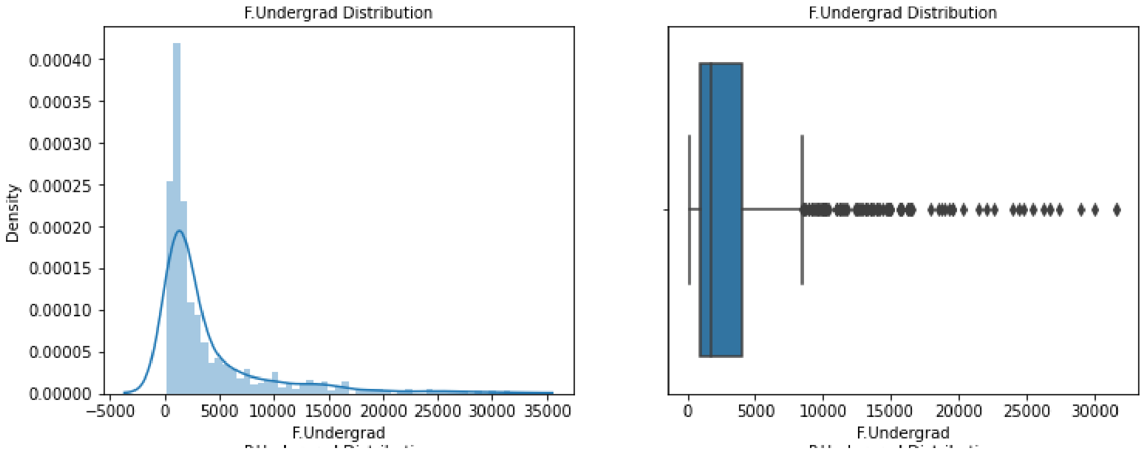
TOP25 PERC:





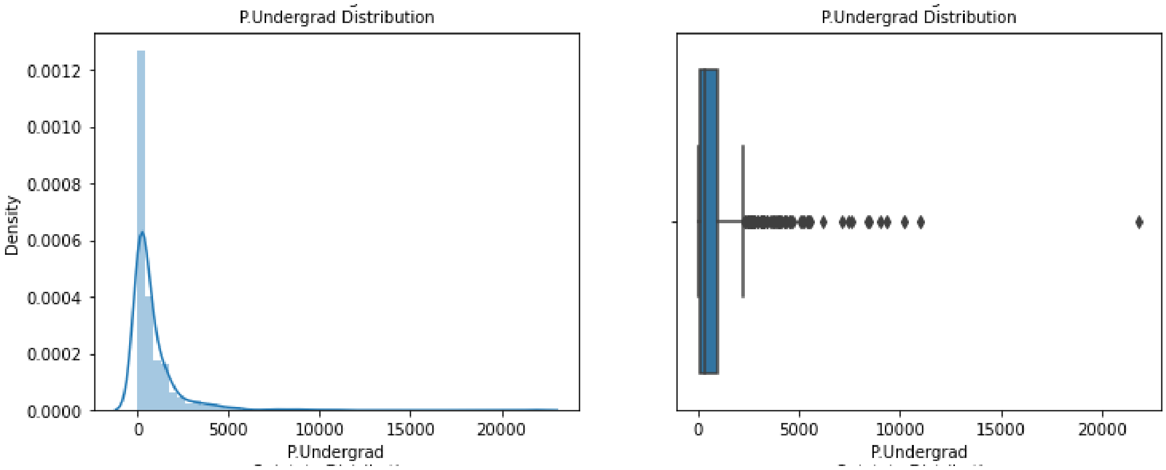
- **The box plot for the top 25% has no outliers.**
- **The distribution is almost normally distributed.**

FULL TIME UNDERGRADUATE:



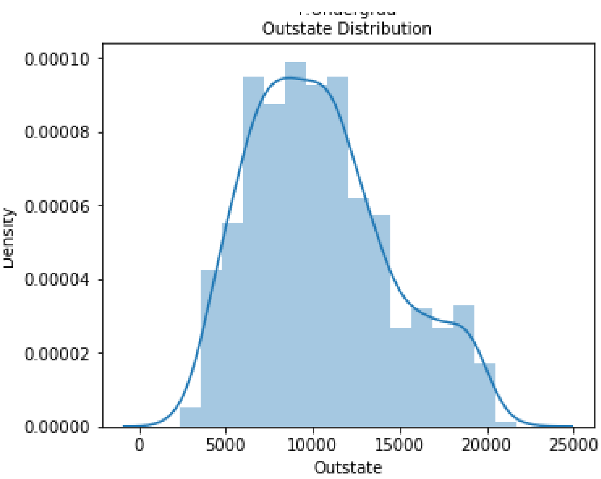
- **The box plot of the full-time graduates has outliers.**
- **The distribution of the data is positively skewed.**
- **In the range about 3000 to 5000 they are full time graduates studying in all the university.**

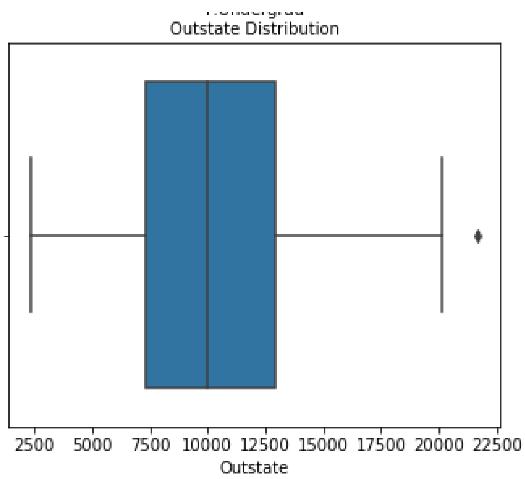
PART TIME UNDERGRADUATE:



- **The box plot of the part time graduates has outliers.**
- **The distribution of the data is positively skewed.**

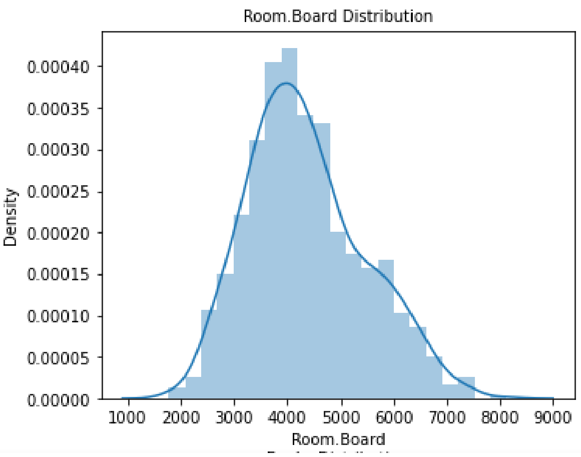
OUTSTATE:

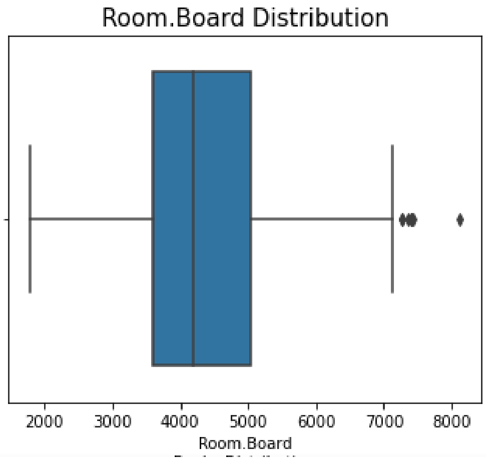




- **The box plot of outstate has only one outlier.**
- **The distribution is almost normally distributed.**

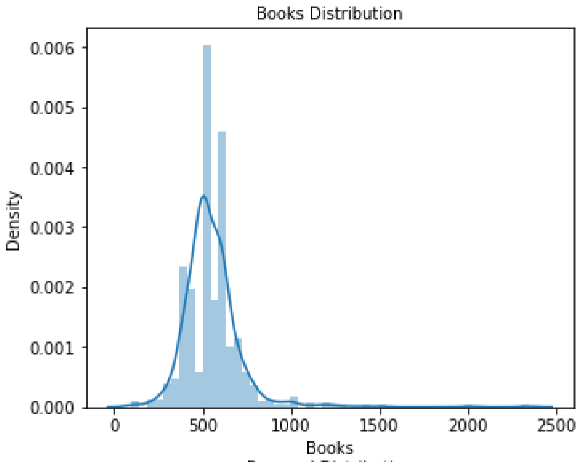
ROOM BOARD:

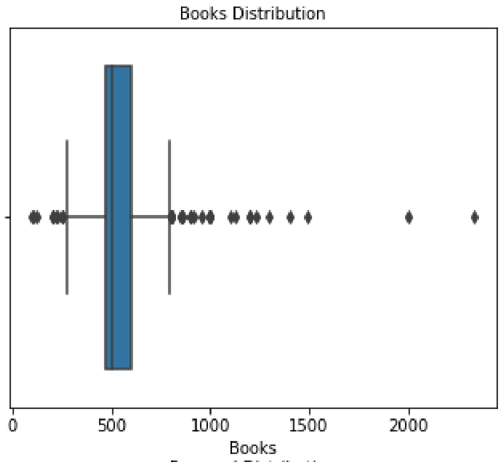




- **The Room board has few outliers.**
- **The distribution is normally distributed.**

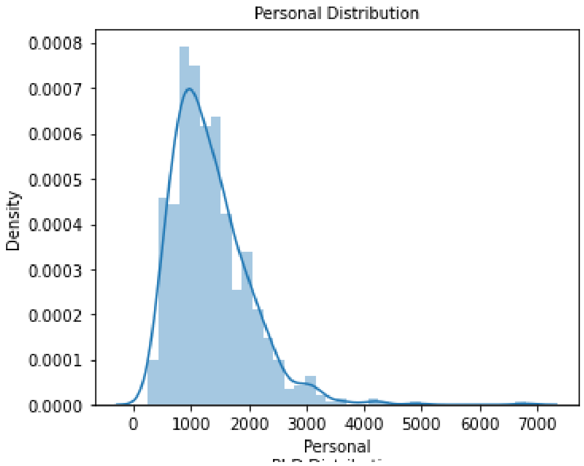
BOOKS:

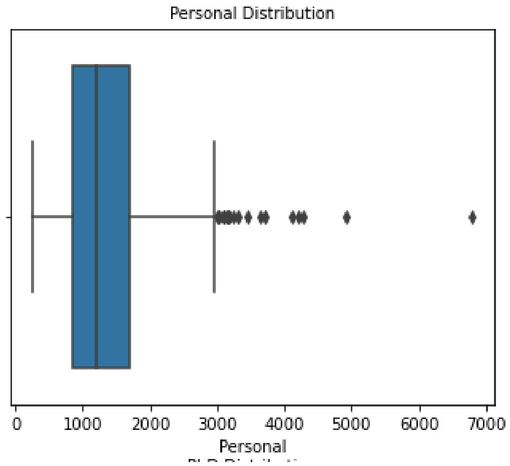




- **The box plot of books has outliers.**
- **The distribution seems to be normally distributed.**

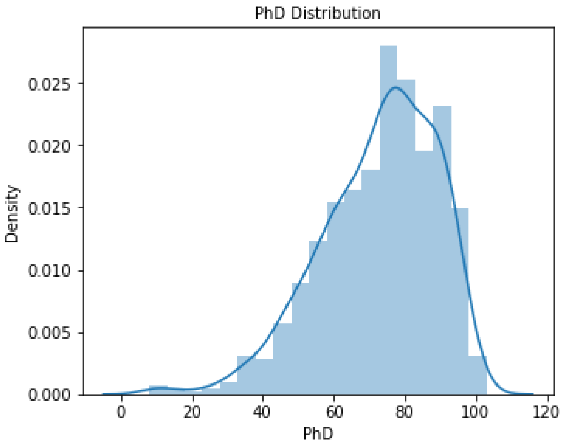
PERSONAL:

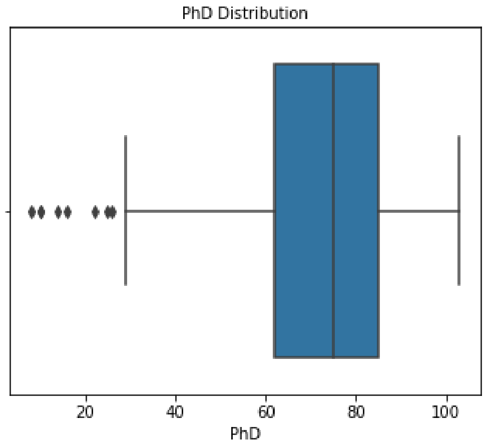




- **The box plot of personal expense has outliers.**
- **The distribution seems to be positively skewed.**

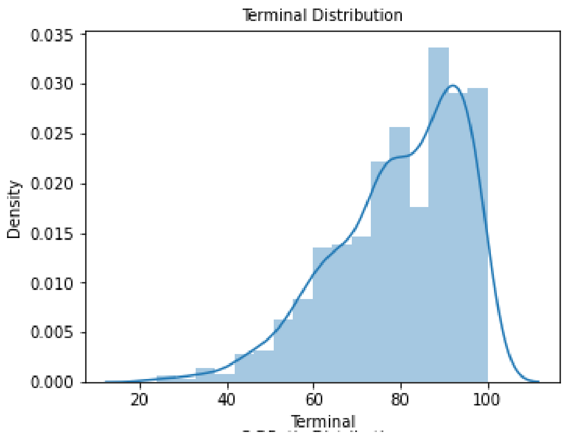
PHD:

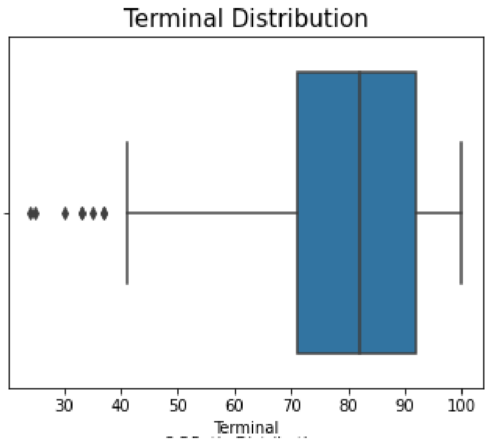




- **The box plot of PHD has outliers.**
- **The distribution seems to be negatively skewed.**

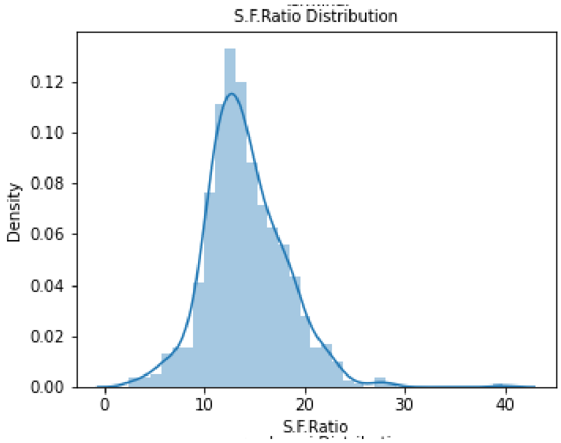
TERMINAL:

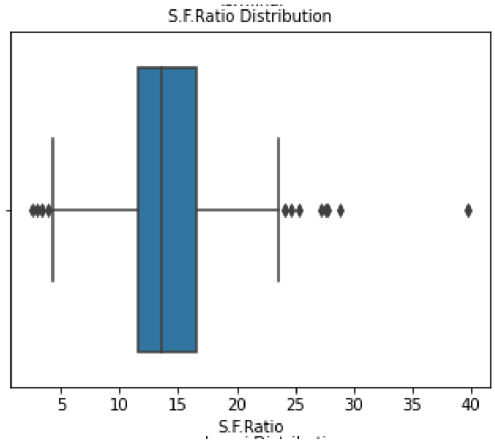




- **The box plot of terminal seems to have outliers in the dataset.**
- **The distribution for the terminal also seems to be negatively skewed.**

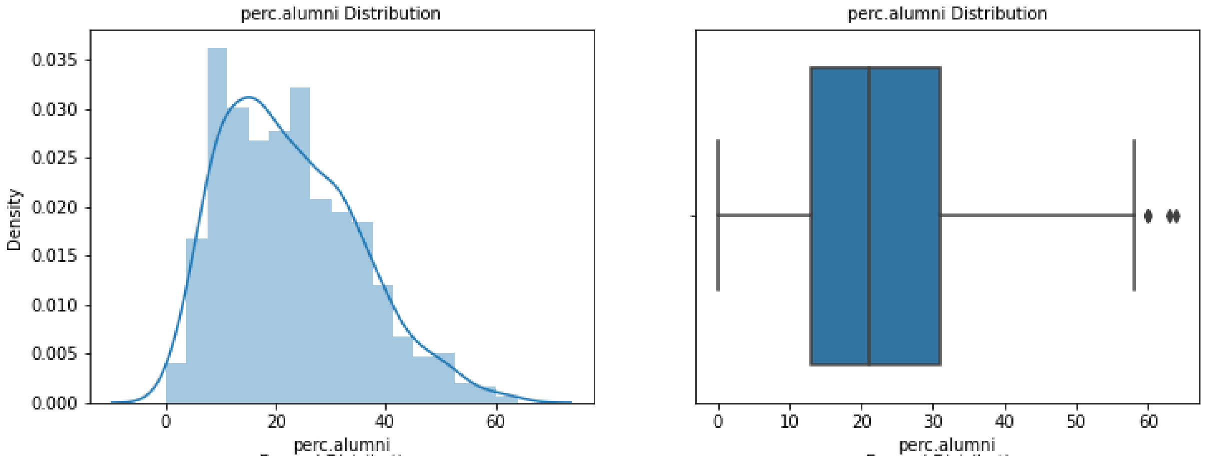
SF RATIO:





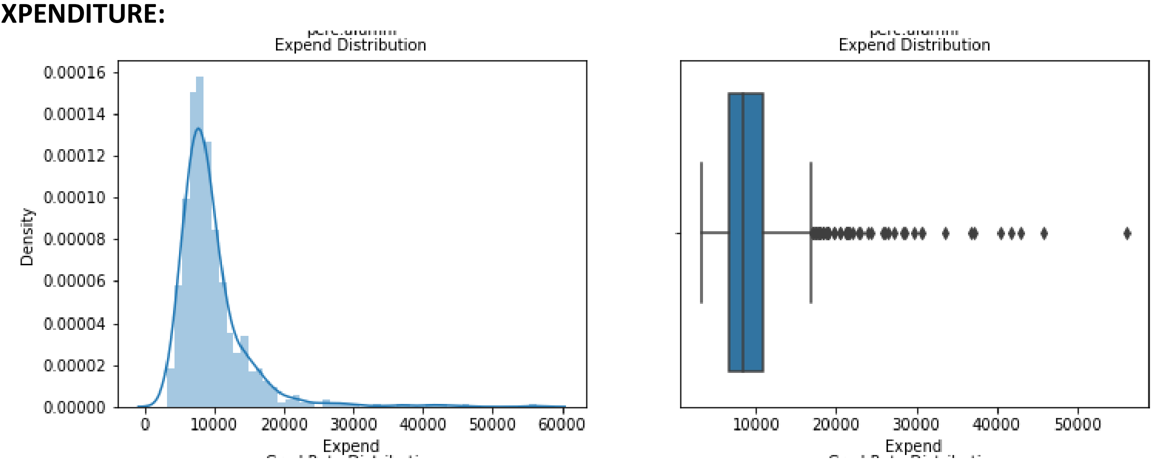
- **The SF ratio variable also has outliers in the dataset.**
- **The distribution is almost normally distributed.**
- **The student faculty ratio is almost same in all the university and colleges.**

PERCI ALUMINI:



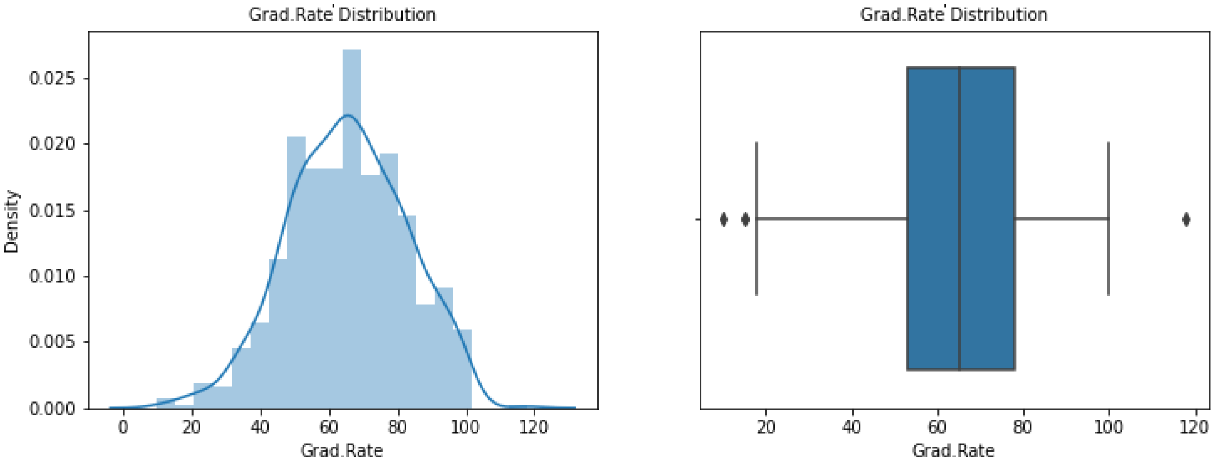
- **The percentage of alumni box plot seems to have outliers in the dataset.**
- **The distribution is almost normally distributed.**

EXPENDITURE:

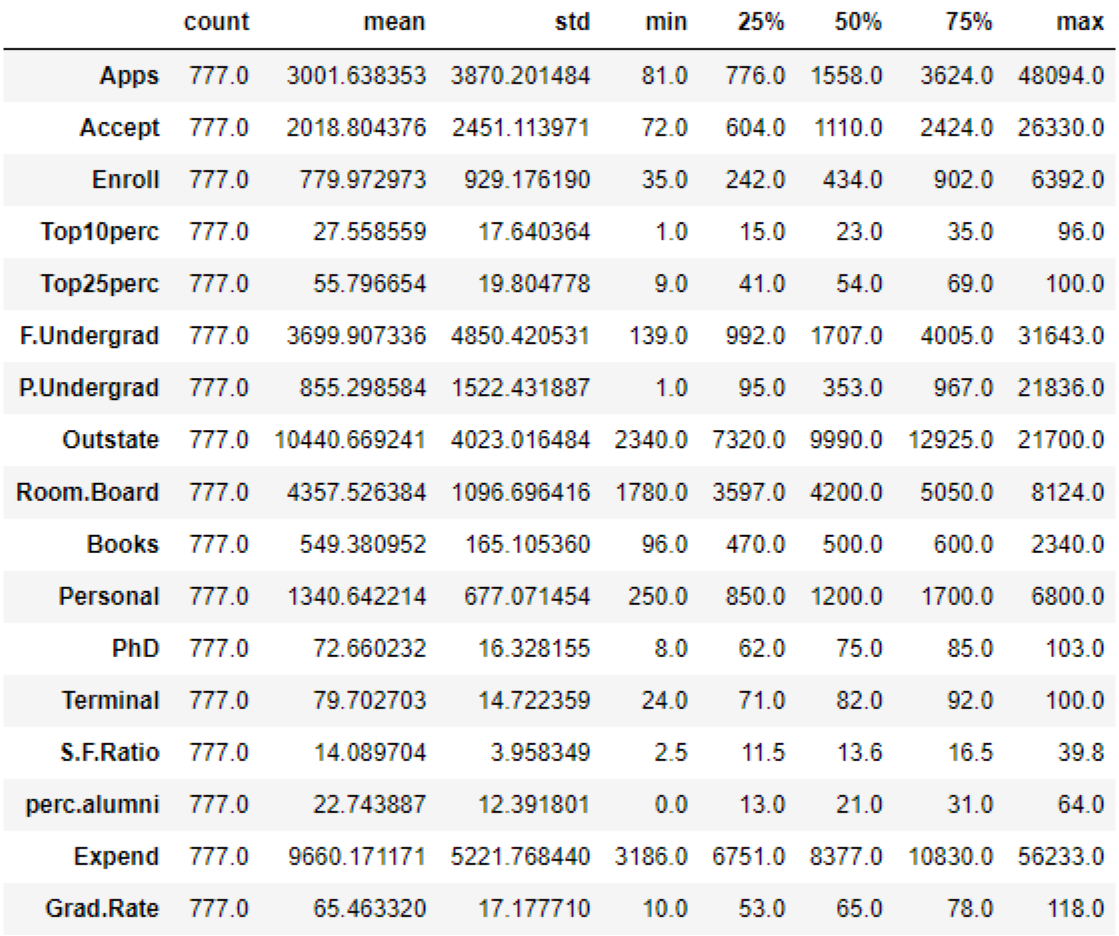


- **The expenditure variable also has outliers in the dataset.**
- **The distribution of the expenditure is positively skewed.**

GRAD RATE:



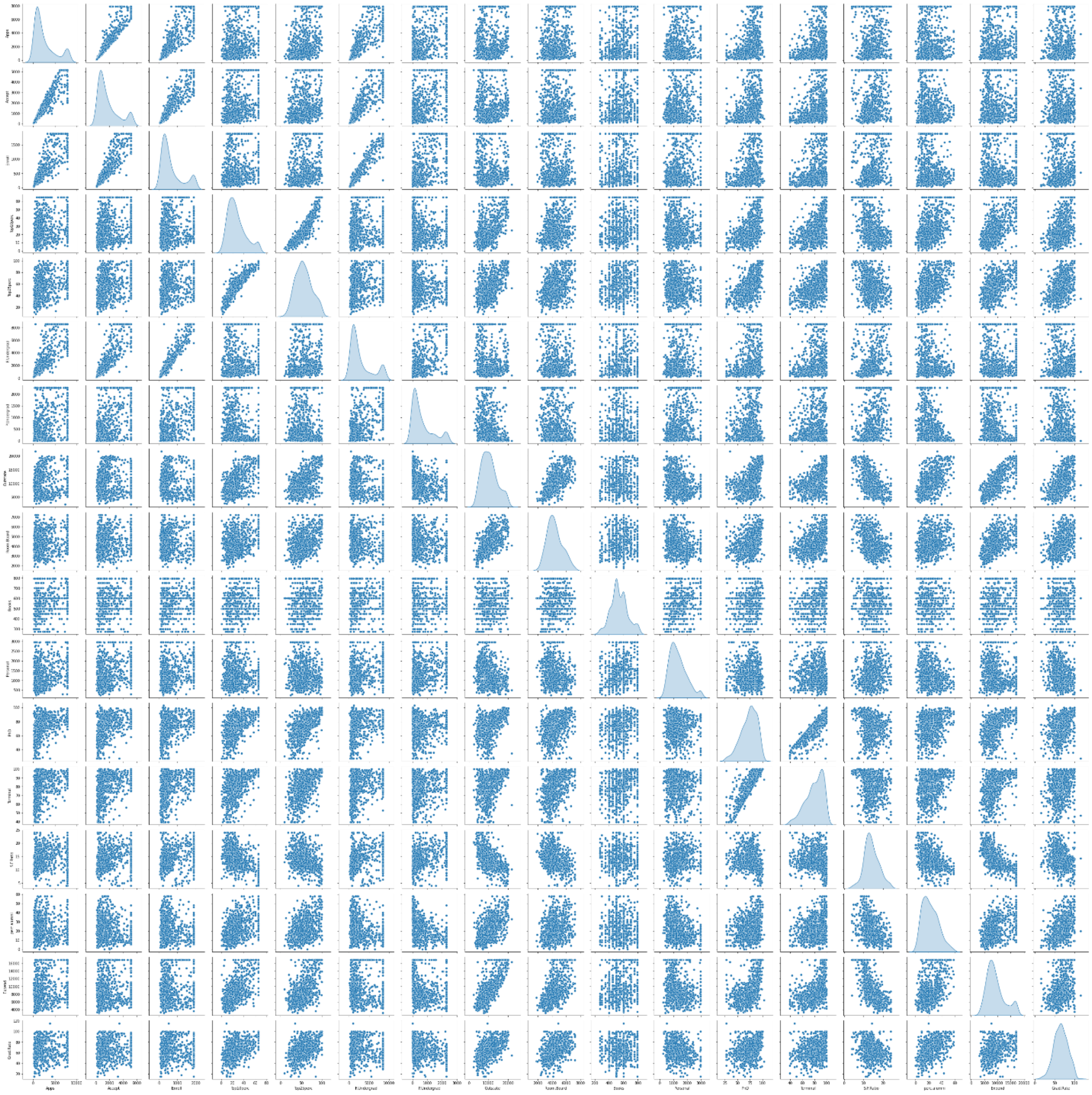
- **The box plot of the graduation rate has outliers in the dataset.**
- **The distribution is normally distributed.**
- **The graduation rate among the students in all the university above 60%.**



	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

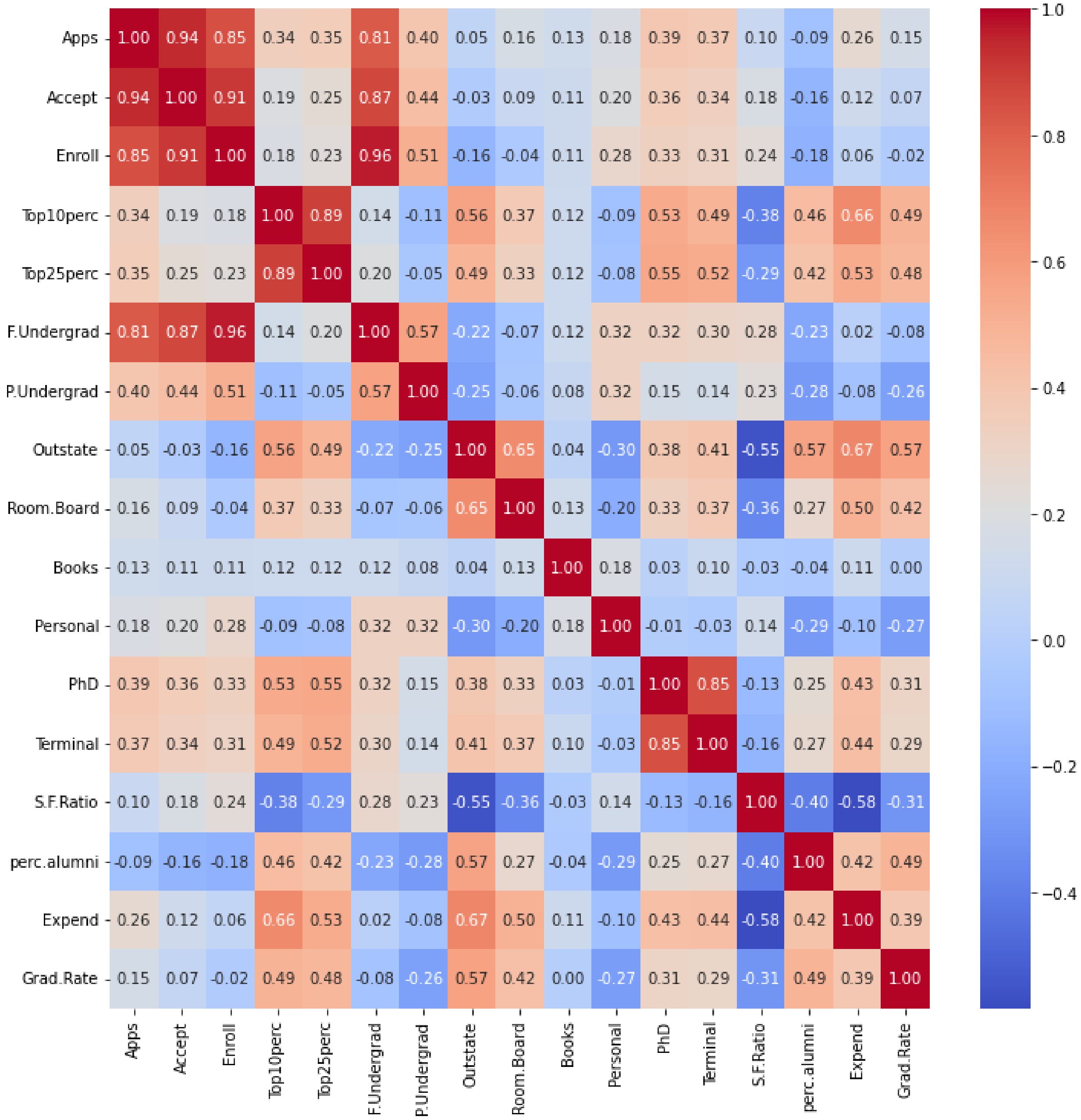
MULTIVARIENT ANALYSIS:

Pair plot to see relationship of all Variables among each other.



- **The pair plot helps us to understand the relationship between all the numerical values in the dataset. On comparing all the variables with each other we could understand the patterns or trends in the dataset.**
- **Few pairs have very high co-relation:**
    - **Application and acceptance**
    - **Students from top 10% schools and from top 25% schools**
    - **Students from top 10% schools and Graduation rate**
    - **Enrollment and Full-time undergrad students**
    - **PHD faculties and Terminal.**

Below Heatmap exhibits multicollinearity issue as significant number of high co-relation variables pairs / features.
When the statistical significance of independent variable is undermined Multicollinearity is observed.



	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

- **This Heat map gives us the correlation between two numerical values.**
- **We could understand the application variable is highly positively correlated with application accepted, students enrolled and full-time graduates. So, this relationship gives the insights on when student submits the application, it is accepted, and the student is enrolled as fulltime graduate.**
- **We can find negative correlation between application and percentage of alumni. This indicates us not all students are part of alumni of their college or university.**
- **The application with top 10, 25 of higher secondary class, outstate, room board, books,**

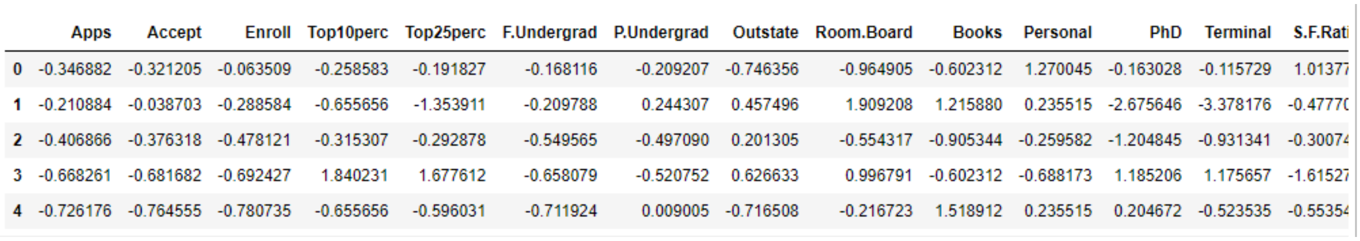
personal, PhD, terminal, S.F ratio, expenditure and Graduation ratio are positively correlated.

Q2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling.

- Our dataset has 18 attributes initially hence we get 18 principal components.
- Once we get the amount of variance explained by each principal component, we can decide how many components we need for our model based on the amount of information we want to retain.
- Hence, it is necessary to normalize data before performing PCA.
- The PCA calculates a new projection to our data set.
- Scaling of Data can be done using Z-Score method or Standard Scalar in SkLearn
  Formula for Z-score:

  $$Z = \frac{x - \mu}{\sigma}$$

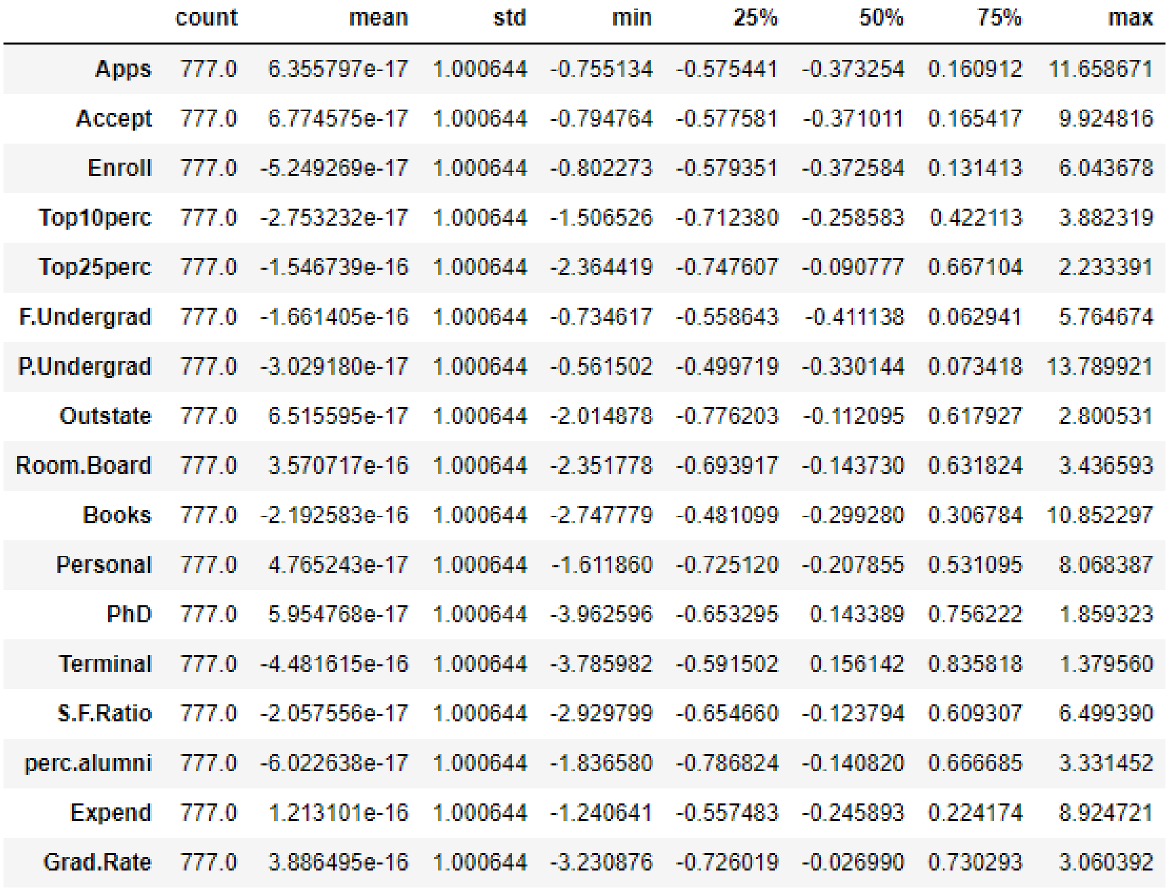
- Z score tells us how many standard deviations the point is away from the mean and also the direction.
- Before scaling I have dropped the names variable which is categorical. Now, the dataset consists of only numerical values, I have applied z-score method for this case study



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Rati
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

Now, we can understand that all the variables are scaled by using z score function. Scaling is one of the most important methods to follow before implementing models.

Below is the Statistical Description of Scaled Dataset.



	count	mean	std	min	25%	50%	75%	max
Apps	777.0	6.355797e-17	1.000644	-0.755134	-0.575441	-0.373254	0.160912	11.658671
Accept	777.0	6.774575e-17	1.000644	-0.794764	-0.577581	-0.371011	0.165417	9.924816
Enroll	777.0	-5.249269e-17	1.000644	-0.802273	-0.579351	-0.372584	0.131413	6.043678
Top10perc	777.0	-2.753232e-17	1.000644	-1.506526	-0.712380	-0.258583	0.422113	3.882319
Top25perc	777.0	-1.546739e-16	1.000644	-2.364419	-0.747607	-0.090777	0.667104	2.233391
F.Undergrad	777.0	-1.661405e-16	1.000644	-0.734617	-0.558643	-0.411138	0.062941	5.764674
P.Undergrad	777.0	-3.029180e-17	1.000644	-0.561502	-0.499719	-0.330144	0.073418	13.789921
Outstate	777.0	6.515595e-17	1.000644	-2.014878	-0.776203	-0.112095	0.617927	2.800531
Room.Board	777.0	3.570717e-16	1.000644	-2.351778	-0.693917	-0.143730	0.631824	3.436593
Books	777.0	-2.192583e-16	1.000644	-2.747779	-0.481099	-0.299280	0.306784	10.852297
Personal	777.0	4.765243e-17	1.000644	-1.611860	-0.725120	-0.207855	0.531095	8.068387
PhD	777.0	5.954768e-17	1.000644	-3.962596	-0.653295	0.143389	0.756222	1.859323
Terminal	777.0	-4.481615e-16	1.000644	-3.785982	-0.591502	0.156142	0.835818	1.379560
S.F.Ratio	777.0	-2.057556e-17	1.000644	-2.929799	-0.654660	-0.123794	0.609307	6.499390
perc.alumni	777.0	-6.022638e-17	1.000644	-1.836580	-0.786824	-0.140820	0.666685	3.331452
Expend	777.0	1.213101e-16	1.000644	-1.240641	-0.557483	-0.245893	0.224174	8.924721
Grad.Rate	777.0	3.886495e-16	1.000644	-3.230876	-0.726019	-0.026990	0.730293	3.060392

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

Q2.3 Comment on the comparison between the covariance and the correlation matrices from this data.[on scaled data]

The comparison between the covariance and correlation matrix is that both of the terms measures the relationship and the dependency between two variables.

Scaling in general means representation of the dataset. The numbers will not change. We are bringing the dataset into one unit.

Covariance indicates the direction of the linear relationship between the variables whether it is positive or negative. By direction means it is directly proportional or inversely proportional.

This below snippet is the covariance matrix on scaled dataset. We can clearly understand covariance matrix indicates direction of the linear relationship between the variables. By direction means it is directly proportional or inversely proportional.

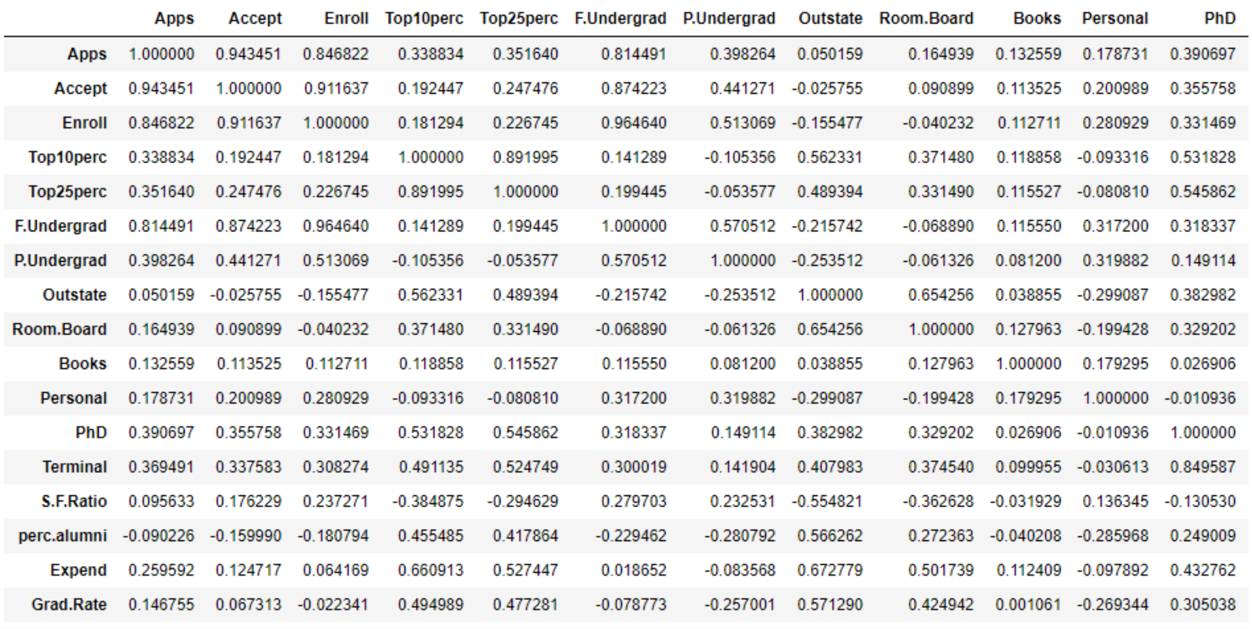
Covariance Matrix

```
%s [[ 1.00128866  0.94466636  0.84791332  0.33927032  0.35209304  0.815540
18
   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]]
```

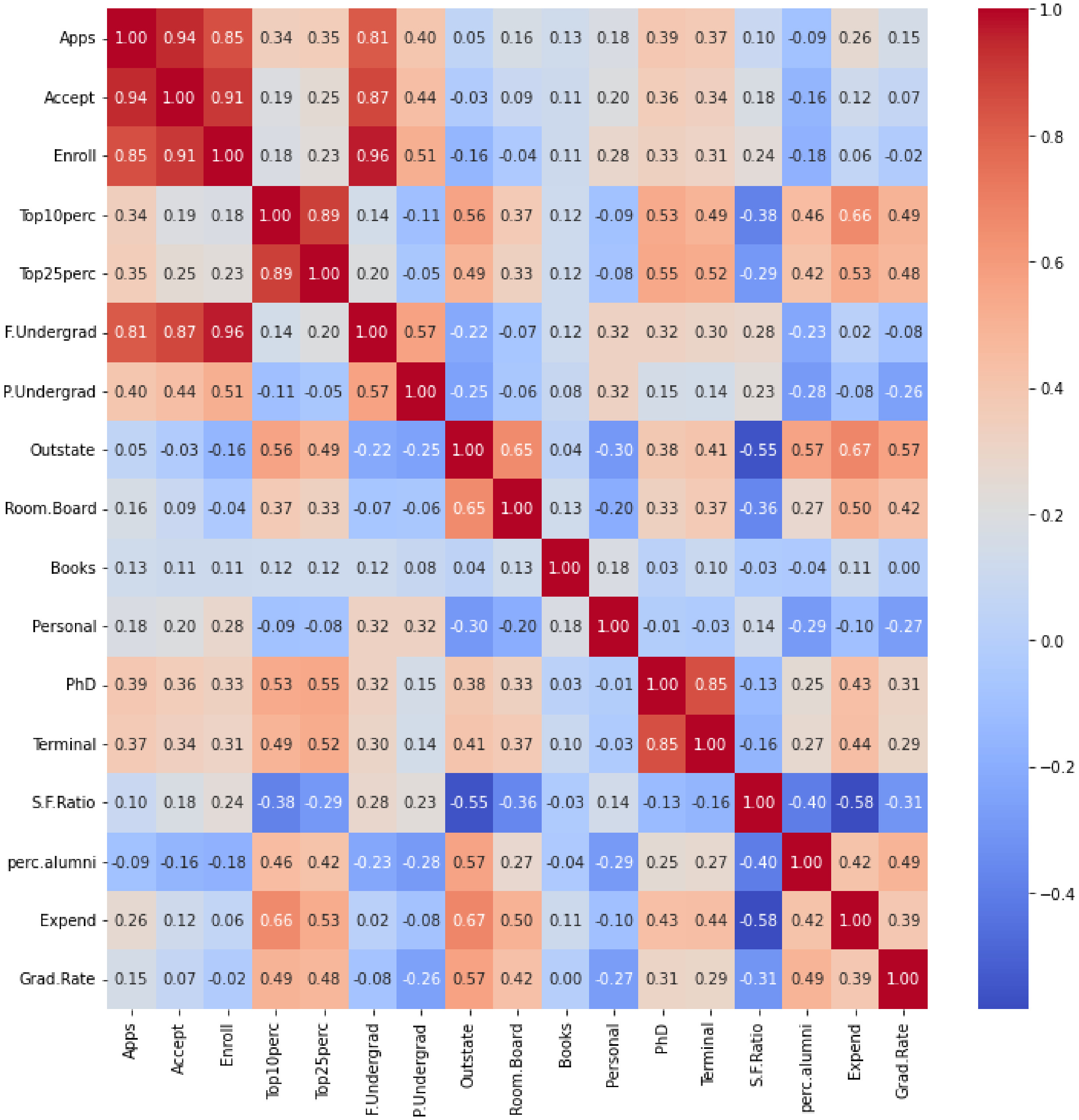
Correlation measures the strength and the direction of the linear relationship between two variables. Strength is that is that positively correlated or negatively correlated.

This below snippet is the correlation matrix. We can clearly understand the correlation matrix which gives the strength and the relationship between the variables.
The correlation matrix before scaling and after scaling will remain the same.



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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

Below is the Heatmap of correlation between Variables after performing Scaling (Before PCA).

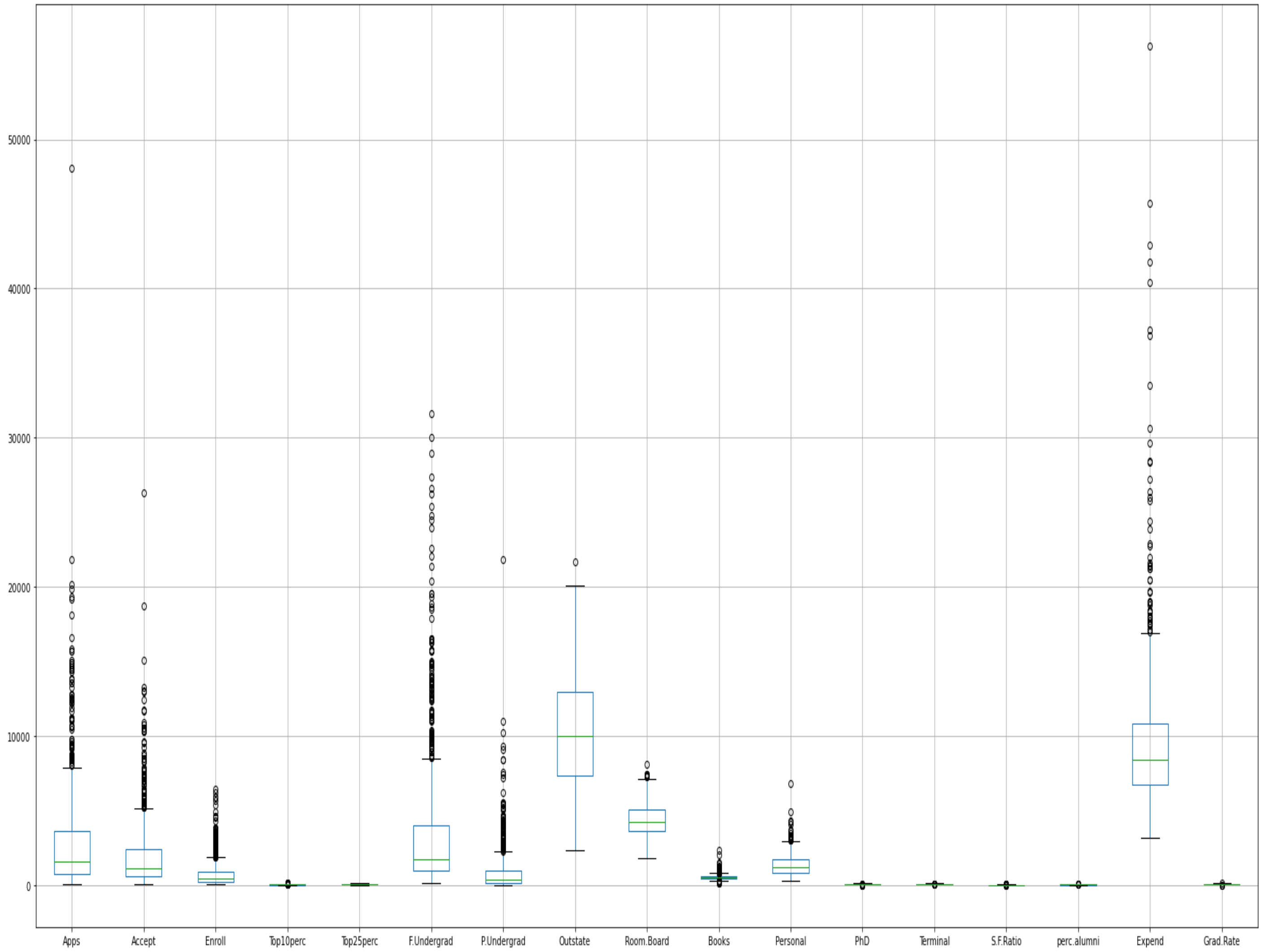


	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.00	0.94	0.85	0.34	0.35	0.81	0.40	0.05	0.16	0.13	0.18	0.39	0.37	0.10	-0.09	0.26	0.15
Accept	0.94	1.00	0.91	0.19	0.25	0.87	0.44	-0.03	0.09	0.11	0.20	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

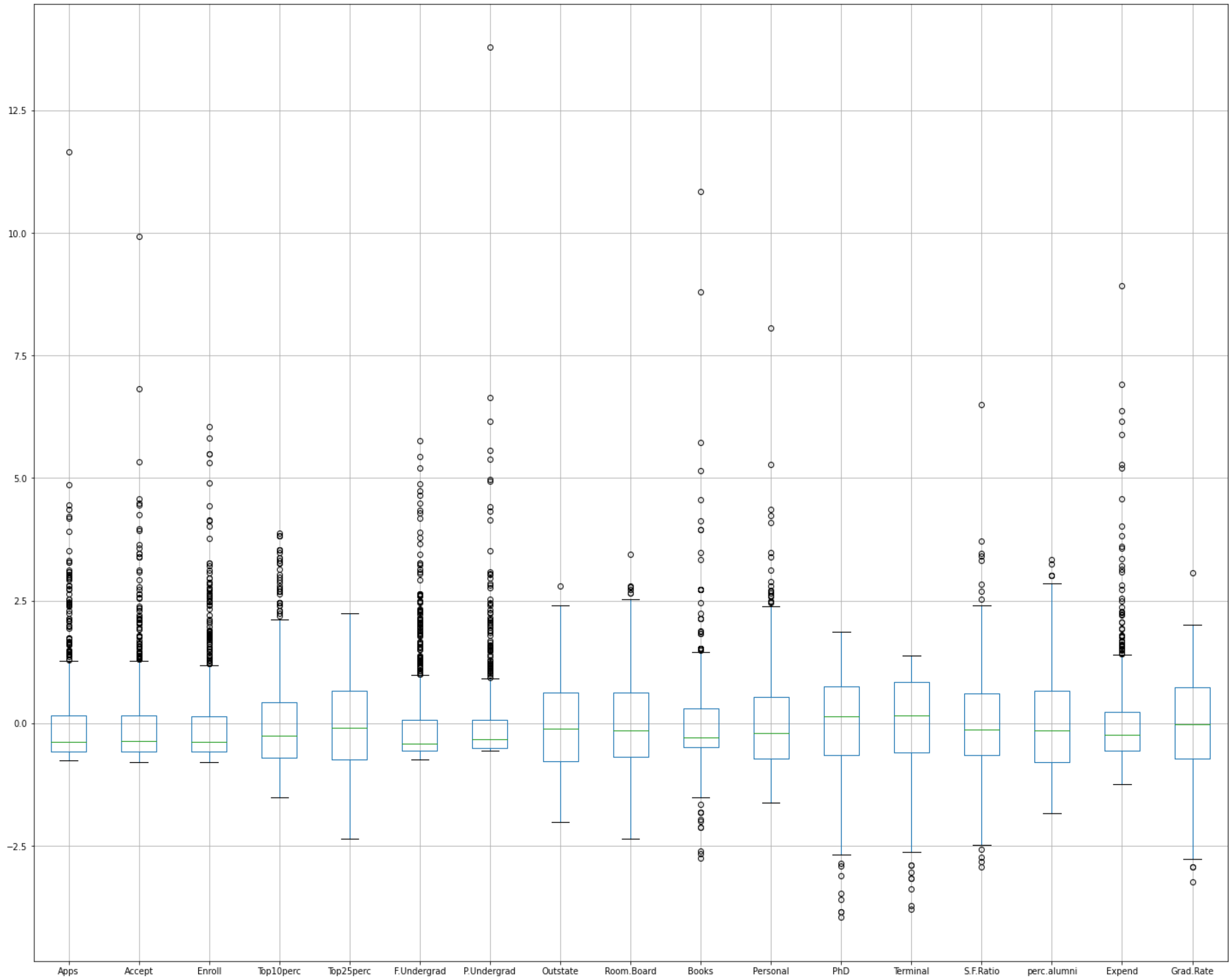
- **From this snippet we can understand variables which are highly positively correlated and the variables which are highly negatively correlated. We can also understand the variables which are moderately correlated with each other. We can see that application, acceptance, enrollment and fulltime graduates are highly positively correlated.**
- **Also, the top 10 percentage and top 25 percentage are highly positively correlated.**
- **Least correlations observed with SF Ratio variable with Expend, Outstate, Grad Rate,**
- **perc.alumni , Room board and Top10perc.**

Q2.4 Check the dataset for outliers before and after scaling. What insight do you derive here?

Checking the data before scaling:



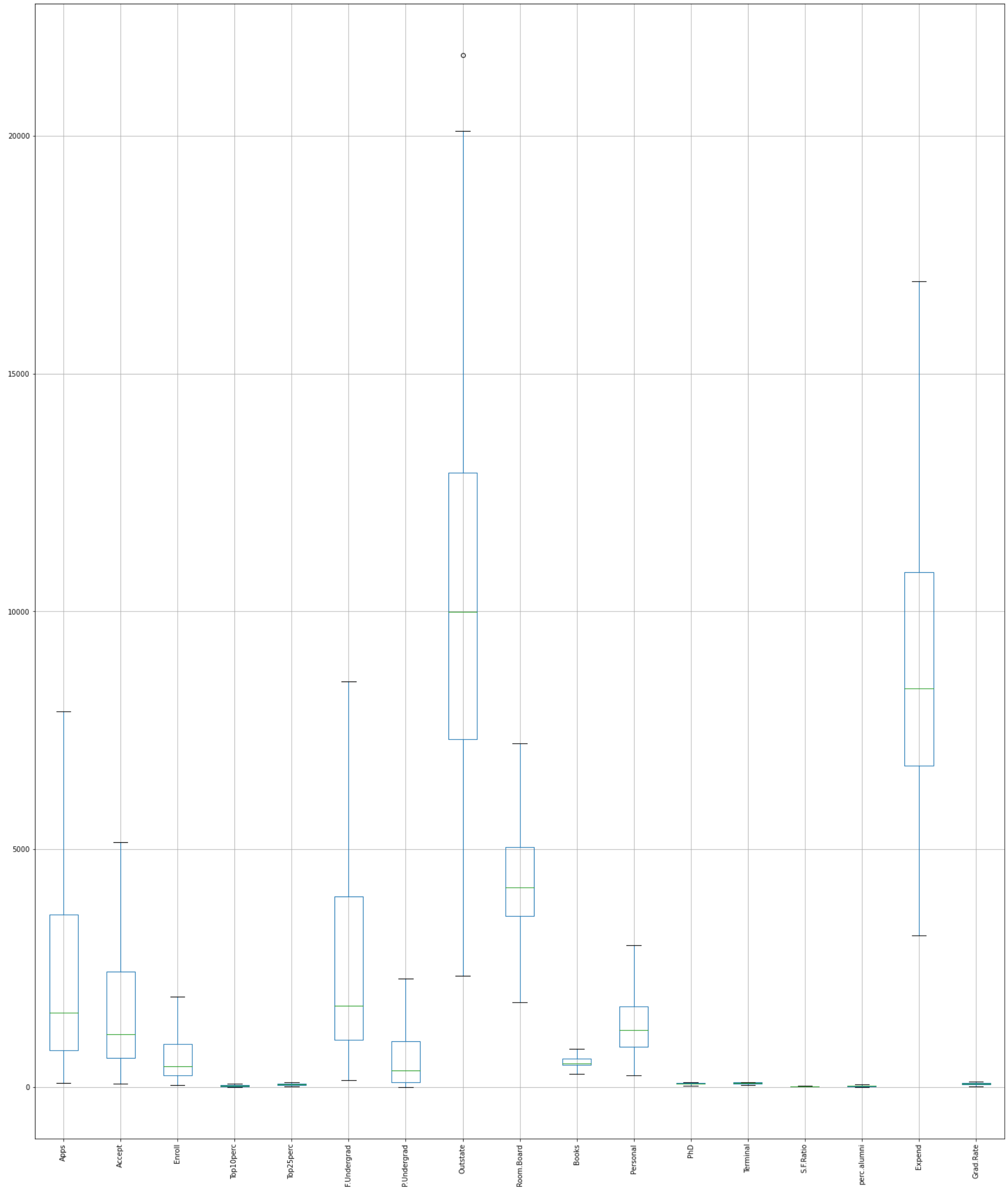
Checking the data after scaling:



Inference:
The outliers are still present in dataset.

Reason:
scaling does not remove outliers scaling scales the values on a Z score distribution. We can use any one method to remove outliers for further processes.

We are using **IQR imputation method** to remove outliers. Below is the data after removing outliers.



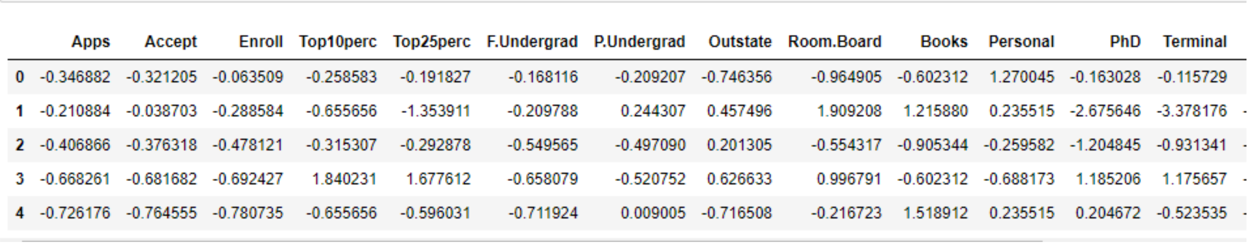
Q2.5 Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both].

Covariance Matrix:

```
[[ 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]]
```

Dataframe after zscore:



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
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	-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
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
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
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

Eigen Vectors:

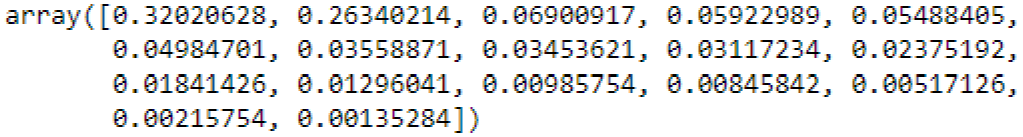
```
array([[ 2.48765602e-01,  2.07601502e-01,  1.76303592e-01,
         3.54273947e-01,  3.44001279e-01,  1.54640962e-01,
         2.64425045e-02,  2.94736419e-01,  2.49030449e-01,
         6.47575181e-02, -4.25285386e-02,  3.18312875e-01,
         3.17056016e-01, -1.76957895e-01,  2.05082369e-01,
         3.18908750e-01,  2.52315654e-01],
       [ 3.31598227e-01,  3.72116750e-01,  4.03724252e-01,
        -8.24118211e-02, -4.47786551e-02,  4.17673774e-01,
         3.15087830e-01, -2.49643522e-01, -1.37808883e-01,
         5.63418434e-02,  2.19929218e-01,  5.83113174e-02,
         4.64294477e-02,  2.46665277e-01, -2.46595274e-01,
        -1.31689865e-01, -1.69240532e-01],
       [-6.30921033e-02, -1.01249056e-01, -8.29855709e-02,
         3.50555339e-02, -2.41479376e-02, -6.13929764e-02,
         1.39681716e-01,  4.65988731e-02,  1.48967389e-01,
         6.77411649e-01,  4.99721120e-01, -1.27028371e-01,
        -6.60375454e-02, -2.89848401e-01, -1.46989274e-01,
         2.26743985e-01, -2.08064649e-01],
       [ 2.81310530e-01,  2.67817346e-01,  1.61826771e-01,
        -5.15472524e-02, -1.09766541e-01,  1.00412335e-01,
        -1.58558487e-01,  1.31291364e-01,  1.84995991e-01,
         8.70892205e-02, -2.30710568e-01, -5.34724832e-01,
        -5.19443019e-01, -1.61189487e-01,  1.73142230e-02,
         7.92734946e-02,  2.69129066e-01],
       [ 5.74140964e-03,  5.57860920e-02, -5.56936353e-02,
        -3.95434345e-01, -4.26533594e-01, -4.34543659e-02,
         3.02385408e-01,  2.22532003e-01,  5.60919470e-01,
        -1.27288825e-01, -2.22311021e-01,  1.40166326e-01,
         2.04719730e-01, -7.93882496e-02, -2.16297411e-01,
         7.59581203e-02, -1.09267913e-01],
       [-1.62374420e-02,  7.53468452e-03, -4.25579803e-02,
        -5.26927980e-02,  3.30915896e-02, -4.34542349e-02,
        -1.91198583e-01, -3.00003910e-02,  1.62755446e-01,
         6.41054950e-01, -3.31398003e-01,  9.12555212e-02,
         1.54927646e-01,  4.87045875e-01, -4.73400144e-02,
        -2.98118619e-01,  2.16163313e-01],
       [-4.24863486e-02, -1.29497196e-02, -2.76928937e-02,
        -1.61332069e-01, -1.18485556e-01, -2.50763629e-02,
         6.10423460e-02,  1.08528966e-01,  2.09744235e-01,
        -1.49692034e-01,  6.33790064e-01, -1.09641298e-03,
        -2.84770105e-02,  2.19259358e-01,  2.43321156e-01,
        -2.26584481e-01,  5.59943937e-01],
       [-1.03090398e-01, -5.62709623e-02,  5.86623552e-02,
        -1.22678028e-01, -1.02491967e-01,  7.88896442e-02,
         5.70783816e-01,  9.84599754e-03, -2.21453442e-01,
         2.13293009e-01, -2.32660840e-01, -7.70400002e-02,
        -1.21613297e-02, -8.36048735e-02,  6.78523654e-01,
        -5.41593771e-02, -5.33553891e-03],
       [-9.02270802e-02, -1.77864814e-01, -1.28560713e-01,
         3.41099863e-01,  4.03711989e-01, -5.94419181e-02,
         5.60672902e-01, -4.57332880e-03,  2.75022548e-01,
        -1.33663353e-01, -9.44688900e-02, -1.85181525e-01,
        -2.54938198e-01,  2.74544380e-01, -2.55334907e-01,
        -4.91388809e-02,  4.19043052e-02],
       [ 5.25098025e-02,  4.11400844e-02,  3.44879147e-02,
         6.40257785e-02,  1.45492289e-02,  2.08471834e-02,
        -2.23105808e-01,  1.86675363e-01,  2.98324237e-01,
        -8.20292186e-02,  1.36027616e-01, -1.23452200e-01,
        -8.85784627e-02,  4.72045249e-01,  4.22999706e-01,
         1.32286331e-01, -5.90271067e-01],
       [ 4.30462074e-02, -5.84055850e-02, -6.93988831e-02,
        -8.10481404e-03, -2.73128469e-01, -8.11578181e-02,
         1.00693324e-01,  1.43220673e-01, -3.59321731e-01,
         3.19400370e-02, -1.85784733e-02,  4.03723253e-02,
        -5.89734026e-02,  4.45000727e-01, -1.30727978e-01,
         6.92088870e-01,  2.19839000e-01],
       [ 2.40709086e-02, -1.45102446e-01,  1.11431545e-02,
         3.85543001e-02, -8.93515563e-02,  5.61767721e-02,
        -6.35360730e-02, -8.23443779e-01,  3.54559731e-01,
        -2.81593679e-02, -3.92640266e-02,  2.32224316e-02,
         1.64850420e-02, -1.10262122e-02,  1.82660654e-01,
         3.25982295e-01,  1.22106697e-01],
       [ 5.95830975e-01,  2.92642398e-01, -4.44638207e-01,
         1.02303616e-03,  2.18838802e-02, -5.23622267e-01,
         1.25997650e-01, -1.41856014e-01, -6.97485854e-02,
         1.14379958e-02,  3.94547417e-02,  1.27696382e-01,
        -5.83134662e-02, -1.77152700e-02,  1.04088088e-01,
        -9.37464497e-02, -6.91969778e-02],
       [ 8.06328039e-02,  3.34674281e-02, -8.56967180e-02,
        -1.07828189e-01,  1.51742110e-01, -5.63728817e-02,
         1.92857500e-02, -3.40115407e-02, -5.84289756e-02,
        -6.68494643e-02,  2.75286207e-02, -6.91126145e-01,
         6.71008607e-01,  4.13740967e-02, -2.71542091e-02,
         7.31225166e-02,  3.64767385e-02],
       [ 1.33405806e-01, -1.45497511e-01,  2.95896092e-02,
         6.97722522e-01, -6.17274818e-01,  9.91640992e-03,
         2.09515982e-02,  3.83544794e-02,  3.40197083e-03,
        -9.43887925e-03, -3.09001353e-03, -1.12055599e-01,
         1.58909651e-01, -2.08991284e-02, -8.41789410e-03,
        -2.27742017e-01, -3.39433604e-03],
       [ 4.59139498e-01, -5.18568789e-01, -4.04318439e-01,
        -1.48738723e-01,  5.18683400e-02,  5.60363054e-01,
        -5.27313042e-02,  1.01594830e-01, -2.59293381e-02,
         2.88282896e-03, -1.28904022e-02,  2.98075465e-02,
        -2.70759809e-02, -2.12476294e-02,  3.33406243e-03,
        -4.38803230e-02, -5.00844705e-03],
       [ 3.58970400e-01, -5.43427250e-01,  6.09651110e-01,
        -1.44986329e-01,  8.03478445e-02, -4.14705279e-01,
         9.01788964e-03,  5.08995918e-02,  1.14639620e-03,
         7.72631963e-04, -1.11433396e-03,  1.38133366e-02,
         6.20932749e-03, -2.22215182e-03, -1.91869743e-02,
        -3.53098218e-02, -1.30710024e-02]])
```

Eigan Values:

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

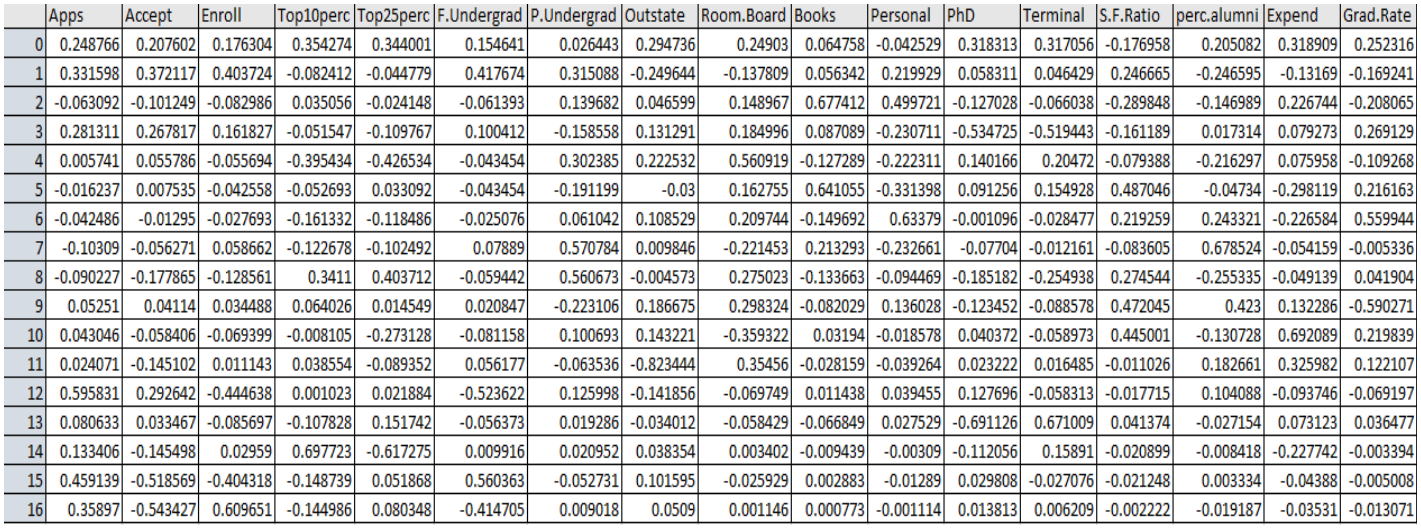
Q2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features.

Explained varience of each PC:



```
array([0.32020628, 0.26340214, 0.06900917, 0.05922989, 0.05488405,
       0.04984701, 0.03558871, 0.03453621, 0.03117234, 0.02375192,
       0.01841426, 0.01296041, 0.00985754, 0.00845842, 0.00517126,
       0.00215754, 0.00135284])
```

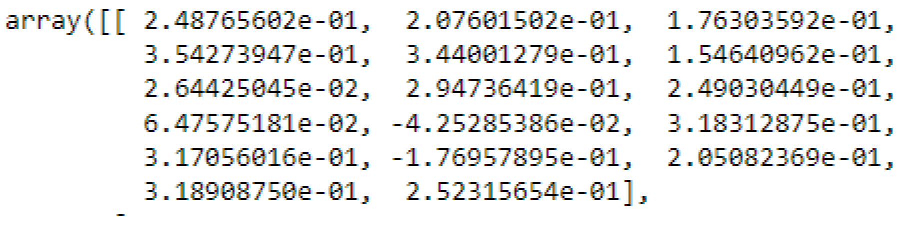
The Loading Score(pca.components_) after performing PCA (in Dataframe format):



	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	0.24903	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	-0.13169	-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	0.20472	-0.079388	-0.216297	0.075958	-0.109268
5	-0.016237	0.007535	-0.042558	-0.052693	0.033092	-0.043454	-0.191199	-0.03	0.162755	0.641055	-0.331398	0.091256	0.154928	0.487046	-0.04734	-0.298119	0.216163
6	-0.042486	-0.01295	-0.027693	-0.161332	-0.118486	-0.025076	0.061042	0.108529	0.209744	-0.149692	0.63379	-0.001096	-0.028477	0.219259	0.243321	-0.226584	0.559944
7	-0.10309	-0.056271	0.058662	-0.122678	-0.102492	0.07889	0.570784	0.009846	-0.221453	0.213293	-0.232661	-0.07704	-0.012161	-0.083605	0.678524	-0.054159	-0.005336
8	-0.090227	-0.177865	-0.128561	0.3411	0.403712	-0.059442	0.560673	-0.004573	0.275023	-0.133663	-0.094469	-0.185182	-0.254938	0.274544	-0.255335	-0.049139	0.041904
9	0.05251	0.04114	0.034488	0.064026	0.014549	0.020847	-0.223106	0.186675	0.298324	-0.082029	0.136028	-0.123452	-0.088578	0.472045	0.423	0.132286	-0.590271
10	0.043046	-0.058406	-0.069399	-0.008105	-0.273128	-0.081158	0.100693	0.143221	-0.359322	0.03194	-0.018578	0.040372	-0.058973	0.445001	-0.130728	0.692089	0.219839
11	0.024071	-0.145102	0.011143	0.038554	-0.089352	0.056177	-0.063536	-0.823444	0.35456	-0.028159	-0.039264	0.023222	0.016485	-0.011026	0.182661	0.325982	0.122107
12	0.595831	0.292642	-0.444638	0.001023	0.021884	-0.523622	0.125998	-0.141856	-0.069749	0.011438	0.039455	0.127696	-0.058313	-0.017715	0.104088	-0.093746	-0.069197
13	0.080633	0.033467	-0.085697	-0.107828	0.151742	-0.056373	0.019286	-0.034012	-0.058429	-0.066849	0.027529	-0.691126	0.671009	0.041374	-0.027154	0.073123	0.036477
14	0.133406	-0.145498	0.02959	0.697723	-0.617275	0.009916	0.020952	0.038354	0.003402	-0.009439	-0.00309	-0.112056	0.15891	-0.020899	-0.008418	-0.227742	-0.003394
15	0.459139	-0.518569	-0.404318	-0.148739	0.051868	0.560363	-0.052731	0.101595	-0.025929	0.002883	-0.01289	0.029808	-0.027076	-0.021248	0.003334	-0.04388	-0.005008
16	0.35897	-0.543427	0.609651	-0.144986	0.080348	-0.414705	0.009018	0.0509	0.001146	0.000773	-0.001114	0.013813	0.006209	-0.002222	-0.019187	-0.03531	-0.013071

Q2.7 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 explicit form of the first PC is below:



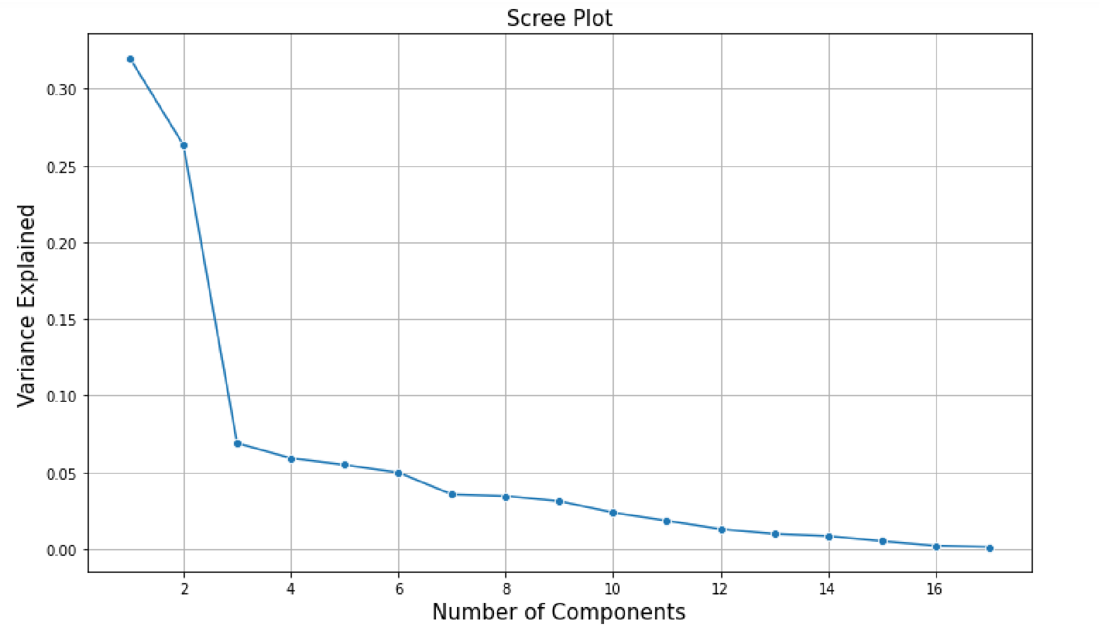
```
array([[ 2.48765602e-01,  2.07601502e-01,  1.76303592e-01,
         3.54273947e-01,  3.44001279e-01,  1.54640962e-01,
         2.64425045e-02,  2.94736419e-01,  2.49030449e-01,
         6.47575181e-02, -4.25285386e-02,  3.18312875e-01,
         3.17056016e-01, -1.76957895e-01,  2.05082369e-01,
         3.18908750e-01,  2.52315654e-01],
```

The Linear equation of 1st component:

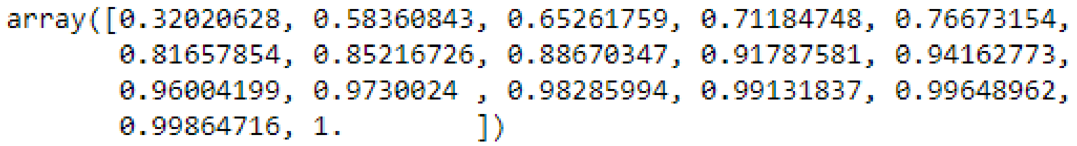
0.25 * Apps + 0.21 * Accept + 0.18 * Enroll + 0.35 * Top10perc + 0.34 * Top25perc + 0.15 * F.Undergrad + 0.03 * P.Undergrad + 0.29 * Outstate + 0.25 * Room.Board + 0.06 * Books + -0.04 * Personal + 0.32 * PhD + 0.32 * Terminal + -0.18 * S.F.Ratio + 0.21 * perc.alumni + 0.32 * Expend + 0.25 * Grad.Rate

Q2.8 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?

Below is the Scree Plot explaining the comparison between Cumulative and Individual explained Variance.



The cumulative explained variance ratio to find a cut off for selecting the number of PCs.



```
array([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.        ])
```

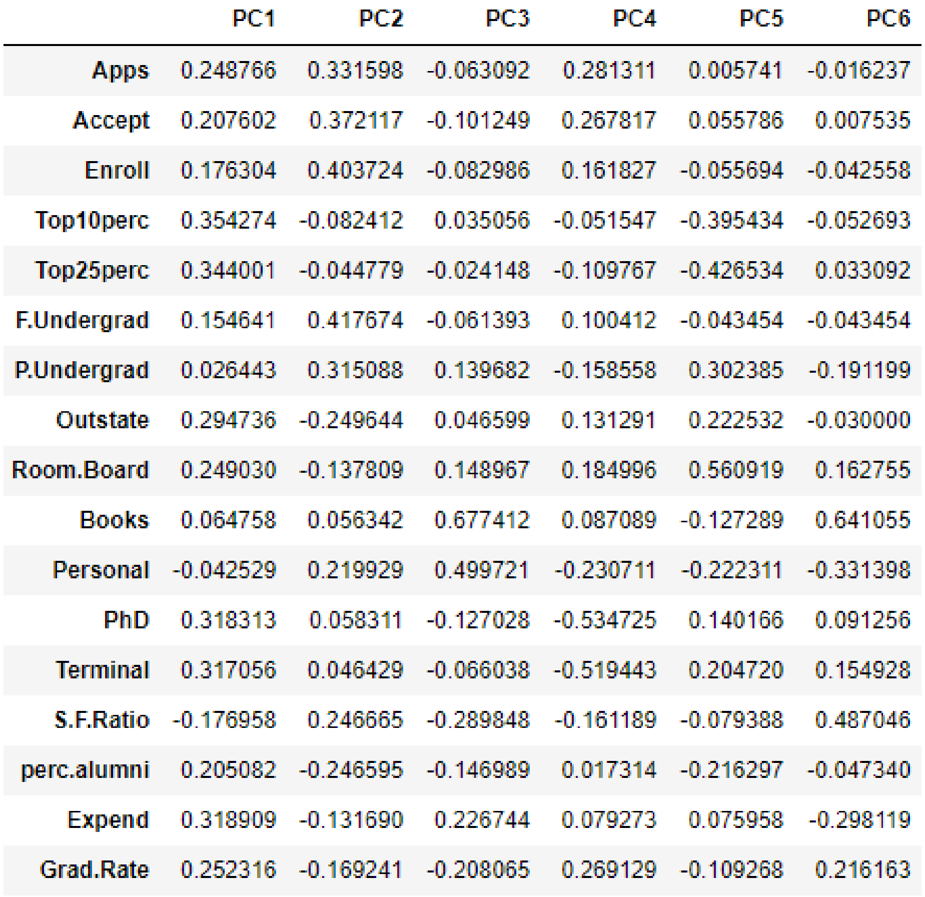
To decide the optimum number of principal components,

1. Check for cumulative variance up to 80%, check the corresponding associated with 80%
2. The incremental value between the components should not be less than five percent.

So, basis on this **we can decide the optimum number of principal components as 6**, because,

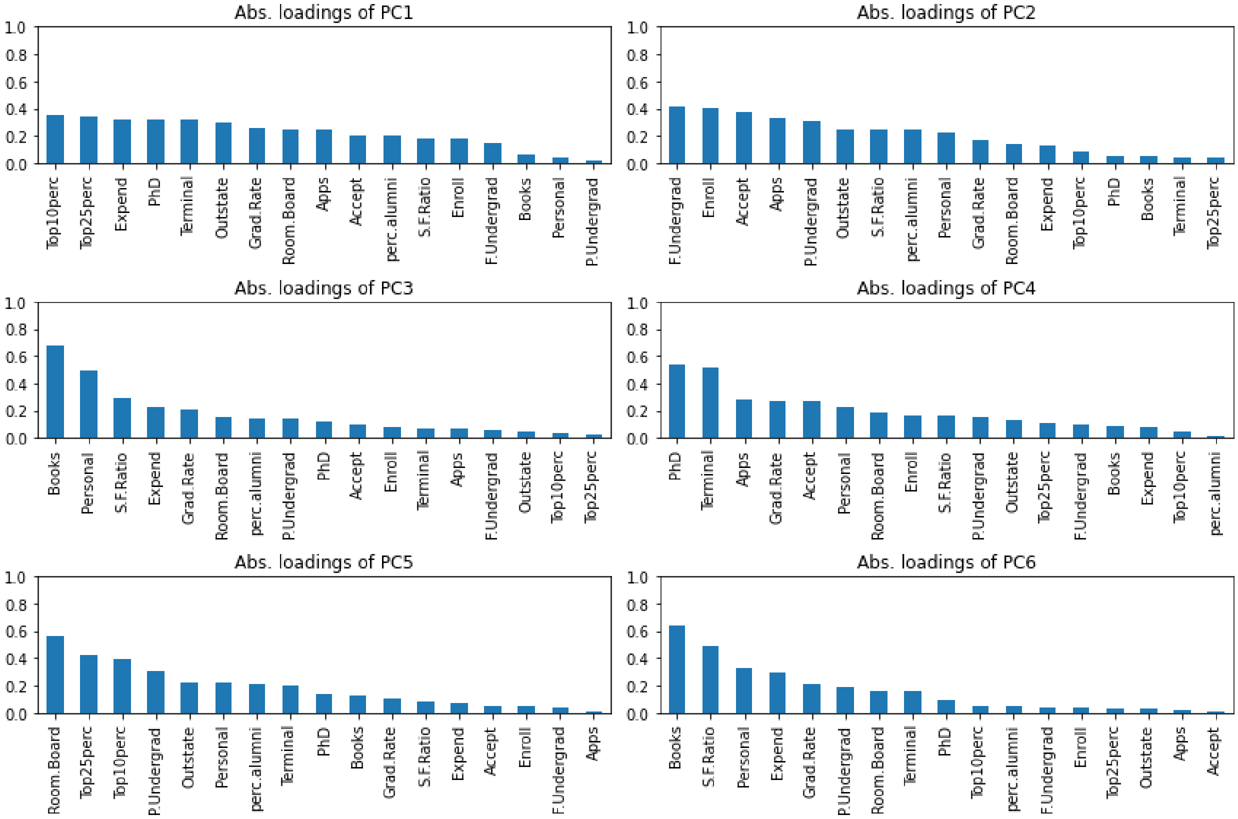
The first components explain 32.02% variance in data
The first two components explain 58.36% variance in data
The first three components explain 65.26% variance in data
The first four components explain 71.18% variance in data
The first five components explain 76.67% variance in data
And **the 6 numbers of principal components cover 81.65% of the variances.**

PCA is performed and it is exported into a data frame. After PCA the multi collinearity is highly reduced.

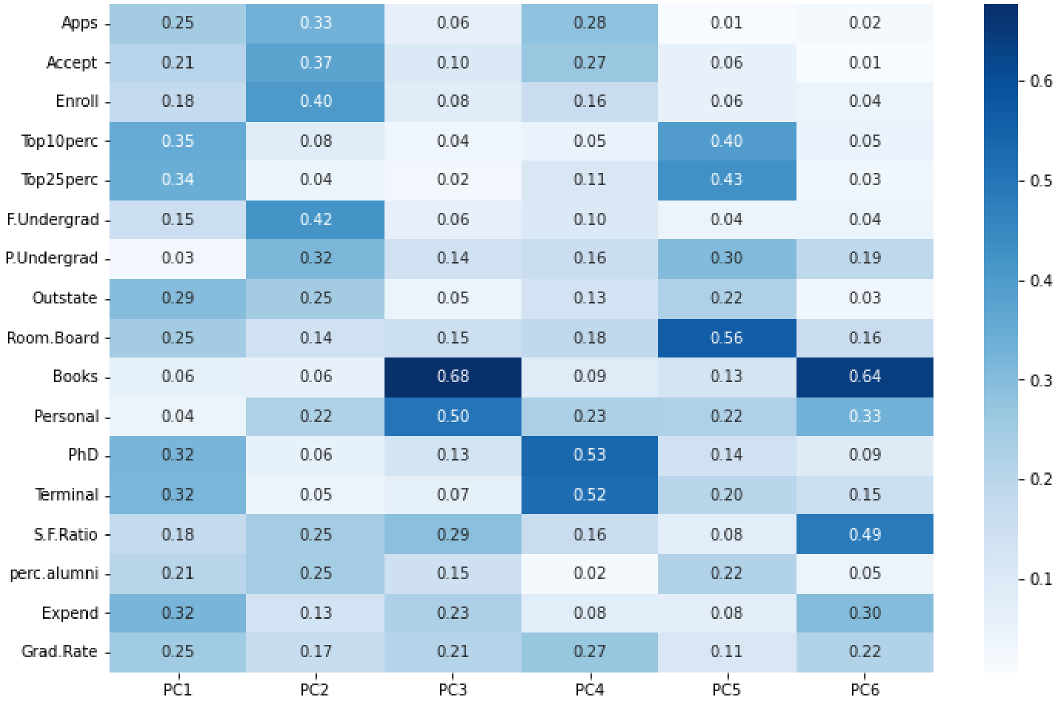


	PC1	PC2	PC3	PC4	PC5	PC6
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163

Subplot on exported data frame:

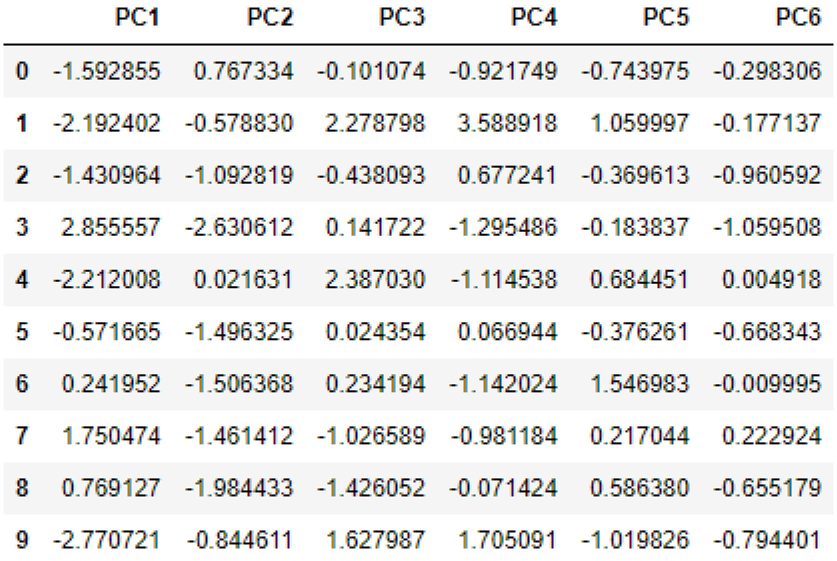


Heatmap:



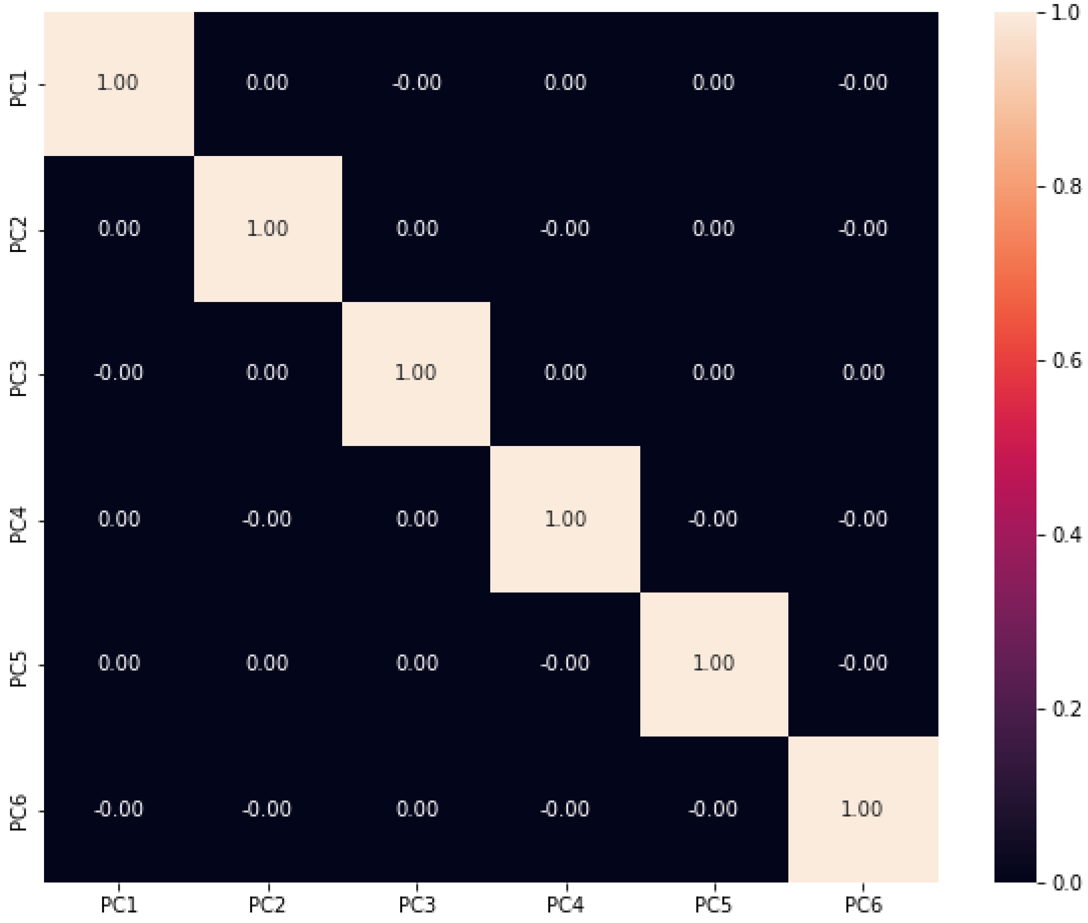
	PC1	PC2	PC3	PC4	PC5	PC6
Apps	0.25	0.33	0.06	0.28	0.01	0.02
Accept	0.21	0.37	0.10	0.27	0.06	0.01
Enroll	0.18	0.40	0.08	0.16	0.06	0.04
Top10perc	0.35	0.08	0.04	0.05	0.40	0.05
Top25perc	0.34	0.04	0.02	0.11	0.43	0.03
F.Undergrad	0.15	0.42	0.06	0.10	0.04	0.04
P.Undergrad	0.03	0.32	0.14	0.16	0.30	0.19
Outstate	0.29	0.25	0.05	0.13	0.22	0.03
Room.Board	0.25	0.14	0.15	0.18	0.56	0.16
Books	0.06	0.06	0.68	0.09	0.13	0.64
Personal	0.04	0.22	0.50	0.23	0.22	0.33
PhD	0.32	0.06	0.13	0.53	0.14	0.09
Terminal	0.32	0.05	0.07	0.52	0.20	0.15
S.F.Ratio	0.18	0.25	0.29	0.16	0.08	0.49
perc.alumni	0.21	0.25	0.15	0.02	0.22	0.05
Expend	0.32	0.13	0.23	0.08	0.08	0.30
Grad.Rate	0.25	0.17	0.21	0.27	0.11	0.22

The dataframe out of fit_transformed scaled data:



	PC1	PC2	PC3	PC4	PC5	PC6
0	-1.592855	0.767334	-0.101074	-0.921749	-0.743975	-0.298306
1	-2.192402	-0.578830	2.278798	3.588918	1.059997	-0.177137
2	-1.430964	-1.092819	-0.438093	0.677241	-0.369613	-0.960592
3	2.855557	-2.630612	0.141722	-1.295486	-0.183837	-1.059508
4	-2.212008	0.021631	2.387030	-1.114538	0.684451	0.004918
5	-0.571665	-1.496325	0.024354	0.066944	-0.376261	-0.668343
6	0.241952	-1.506368	0.234194	-1.142024	1.546983	-0.009995
7	1.750474	-1.461412	-1.026589	-0.981184	0.217044	0.222924
8	0.769127	-1.984433	-1.426052	-0.071424	0.586380	-0.655179
9	-2.770721	-0.844611	1.627987	1.705091	-1.019826	-0.794401

The final presence of correlations among the PCs:



	PC1	PC2	PC3	PC4	PC5	PC6
PC1	1.00	0.00	-0.00	0.00	0.00	-0.00
PC2	0.00	1.00	0.00	-0.00	0.00	-0.00
PC3	-0.00	0.00	1.00	0.00	0.00	0.00
PC4	0.00	-0.00	0.00	1.00	-0.00	-0.00
PC5	0.00	0.00	0.00	-0.00	1.00	-0.00
PC6	-0.00	-0.00	0.00	-0.00	-0.00	1.00

Q2.9 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**

---END--