Table Of Contents

About the Data

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals 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.

Data Description

- *Education: - Education taken by the candidates and it is categorized into 3 categories*
- *Doctorate, Bachelors and H.S Grade. It is in categorical data type.*
- *Occupation: - Occupation is the occupation of them and it is categorized into 4*
- *categories Adm-clerical, Sales, Prof-specialty and Exec-managerial. It is in categorical*
- *data type.*
- *Salary: - The salary is the salary of them in integer data type.*

	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

Sample of the dataset

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Education	40	3	Doctorate	16	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Occupation	40	4	Prof-specialty	13	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	40.00	NaN	NaN	NaN	162186.88	64860.41	50103.00	99897.50	169100.00	214440.75	260151.00

Description of the dataset

The above Dataset has 3 variables with 3 different varieties of education and 4 different occupation types.

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

a. The Null and Alternate Hypothesis for Education

H_0: The Mean Salary is same across all the level of education
H_1: The Mean Salary is different at least in one level of education

b. The Null and Alternate Hypothesis for Occupation

H_0: The Mean Salary is the same across all the categories of occupation
H_1: The Mean Salary is different at least in one categories of occupation

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.

H_0: The Mean Salary is same across all the level of education
H_1: The Mean Salary is different at least in one level of education

	df	Sum_Sq	Mean_Sq	F	PR(>F)
Education	2.00	1,02,69,54,66,735.89	51,34,77,33,367.94	30.96	0
Residual	37.00	61,37,25,59,274.49	1,65,87,17,818.23	NaN	NaN

ANOVA Table 1: Salary with respect to Education

Since the p value = 0.00 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.

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.

H_0: The Mean Salary is the same across all the categories of occupation

H_1: The Mean Salary is different at least in one categories of occupation

	df	Sum_Sq	Mean_Sq	F	PR(>F)
Occupation	3.00	1.1259E+10	3.753E+09	0.88	0.46
Residual	36.00	1.5281E+11	4.245E+09	NaN	NaN

ANOVA Table 2: Salary with respect to Occupation

Since the p value = 0.46 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.

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)

	Education	Occupation
Mean_Sq	**5.135E+10**	**3.753E+09**
Residual	**1.659E+09**	**4.245E+09**
Name: mean_sq, dtype: float64		

Table 3: Mean Square

Since mean_sq difference is more between educational qualification category which tells that Education category means are significantly different.

Problem 1B:

1. What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the 'pointplot' function from the 'seaborn' function]

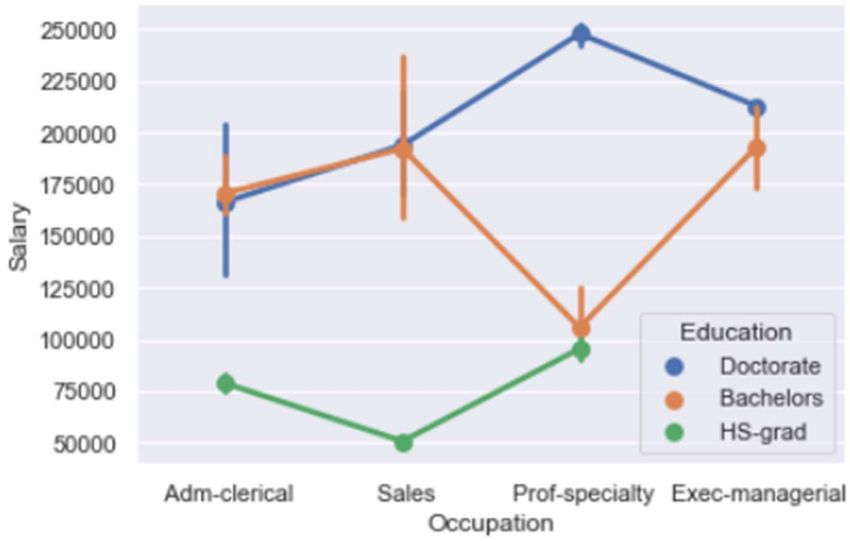


Figure 1: Interaction Between Education and Occupation

From the above Figure we can infer the following points

- *People with a HS-grad can only find employment as admin clerks, salespeople, and professors with specialties; they cannot advance to executive management positions.*
- *People who have a Bachelor's or Doctoral degree and work in sales or administration with incomes between $170,000 and $190,000 make approximately the same amount of money.*
- *Those who with a bachelor's degree and work in the prof-specialty field make less money than those who hold a bachelor's degree and work in the admin-clerical and sales fields.*
- *Bachelor's degree holders pursuing a career in sales make more money than those with a Bachelor's degree holders having a career in academia, while those with a doctorate having a career in sales make less money than those with a doctorate pursuing a career in academia.*
- *Similarly, those who hold a bachelor's degree and work in a prof-specialty earn less than those who don't.*

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?

H_0: 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).

H_1: There is an interaction effect between the independent variable 'education' and the independent variable 'occupation' on the mean Salary.

	df	Sum_sq	Mean_sq	F	PR(>F)
C(Education)	2.00	1.027E+11	5.13E+10	72.21	-
C(Occupation)	3.00	5.52E+09	1.84E+09	2.59	0.07
C(Education):C(Occupation)	6.00	3.635E+10	6.06E+09	8.52	-
Residual	29.00	2.062E+10	7.11E+08	NaN	NaN

ANOVA Table 4: Two-way test based on Salary with respect to Education and Occupation

From the table, we see that there is a significant amount of interaction between the variables, Education and Occupation.

As p value = 0 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.

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

With a Bachelor's degree and a job as a Prof-Specialty earn less than people with a Bachelor's degree and a job as an Exec-Managerial, whereas people with a Doctorate degree and a job as a Prof-Specialty earn more than people with a Doctorate degree and a job as an Exec-Managerial. There is also a plot twist in this section.

But, 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, considering the both category we can conclude that Salary is moderately dependent on educational qualifications and occupation.

About the Data

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

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	perc.alumni	Expend	Grad.Rate
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	12	7041	60
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	16	10527	56
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	30	8735	54
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	37	19016	59
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	2	10922	15

Sample of the dataset

	count	mean	std	min	25%	50%	75%	max
Apps	777.00	3001.64	3870.20	81.00	776.00	1558.00	3624.00	48094.00
Accept	777.00	2018.80	2451.11	72.00	604.00	1110.00	2424.00	26330.00
Enroll	777.00	779.97	929.18	35.00	242.00	434.00	902.00	6392.00
Top10perc	777.00	27.56	17.64	1.00	15.00	23.00	35.00	96.00
Top25perc	777.00	55.80	19.80	9.00	41.00	54.00	69.00	100.00
F.Undergrad	777.00	3699.91	4850.42	139.00	992.00	1707.00	4005.00	31643.00
P.Undergrad	777.00	855.30	1522.43	1.00	95.00	353.00	967.00	21836.00
Outstate	777.00	10440.67	4023.02	2340.00	7320.00	9990.00	12925.00	21700.00
Room.Board	777.00	4357.53	1096.70	1780.00	3597.00	4200.00	5050.00	8124.00
Books	777.00	549.38	165.11	96.00	470.00	500.00	600.00	2340.00
Personal	777.00	1340.64	677.07	250.00	850.00	1200.00	1700.00	6800.00
PhD	777.00	72.66	16.33	8.00	62.00	75.00	85.00	103.00
Terminal	777.00	79.70	14.72	24.00	71.00	82.00	92.00	100.00
S.F.Ratio	777.00	14.09	3.96	2.50	11.50	13.60	16.50	39.80
perc.alumni	777.00	22.74	12.39	0.00	13.00	21.00	31.00	64.00
Expend	777.00	9660.17	5221.77	3186.00	6751.00	8377.00	10830.00	56233.00
Grad.Rate	777.00	65.46	17.18	10.00	53.00	65.00	78.00	118.00

Description of the dataset

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.

The entire dataset does not have any missing values or null values.

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

UNIVARIATE ANALYSIS

The main purpose of univariate analysis is to describe the data, summarize and finds pattern,

it doesn't deal with causes and relationships unlike regression.

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

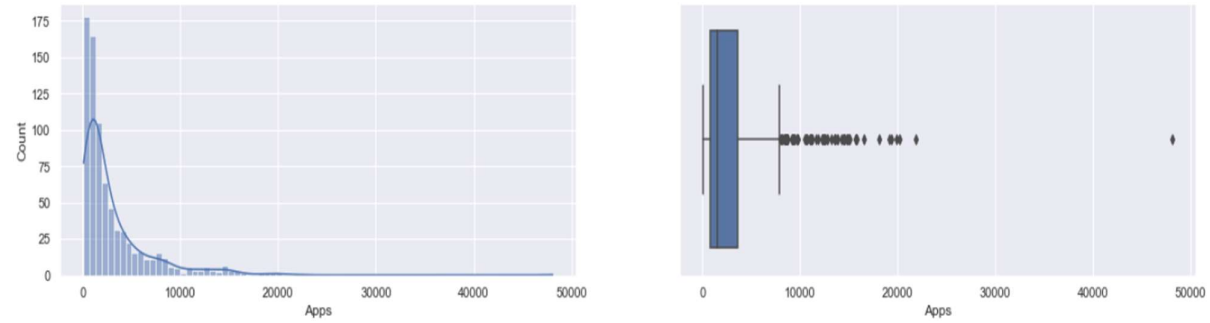


Figure 2 : Apps

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 3000 to 5000. The max applications seems to be around 50,000.

For univariate analysis of apps we are using box plot and dist plot to find information or patterns in the data.

So, we can clearly understand from the box plot we have outliers in the dataset.

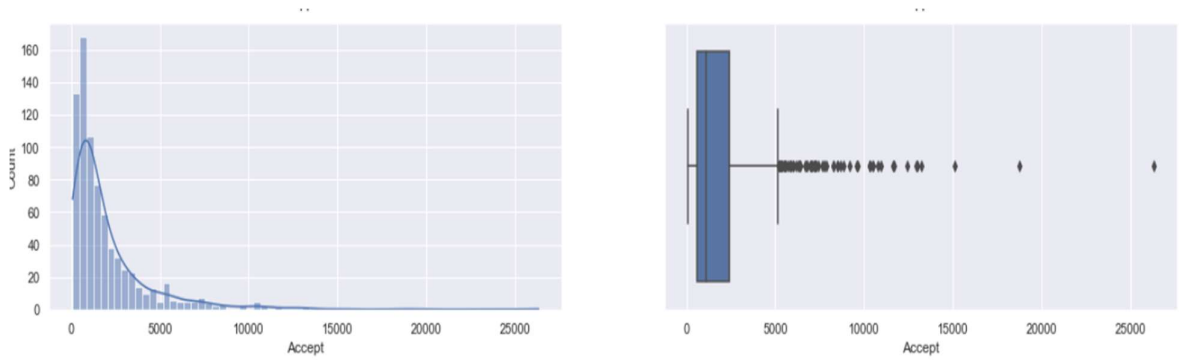


Figure 3 : Accept

The accept variable seems to have outliers. The dist plot shows us the majority of applications accepted from each university are in the range from 70 to 1500.The accept variable seems to be positively skewed.

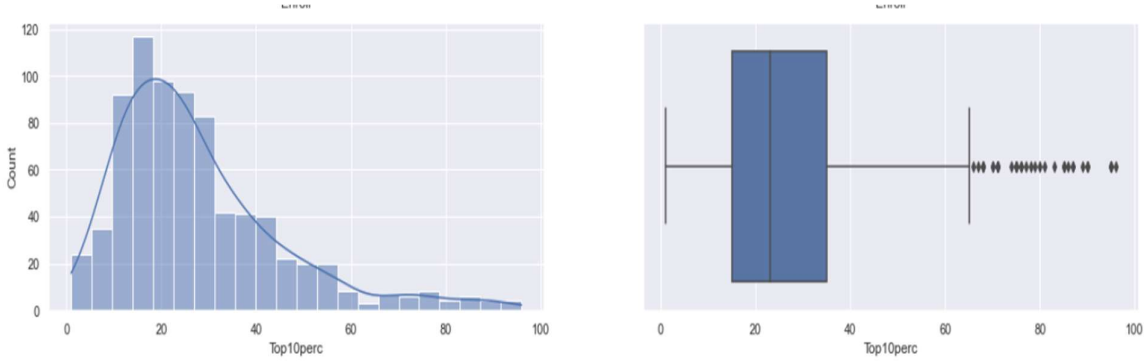


Figure 4 : Top 10 Perc

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.

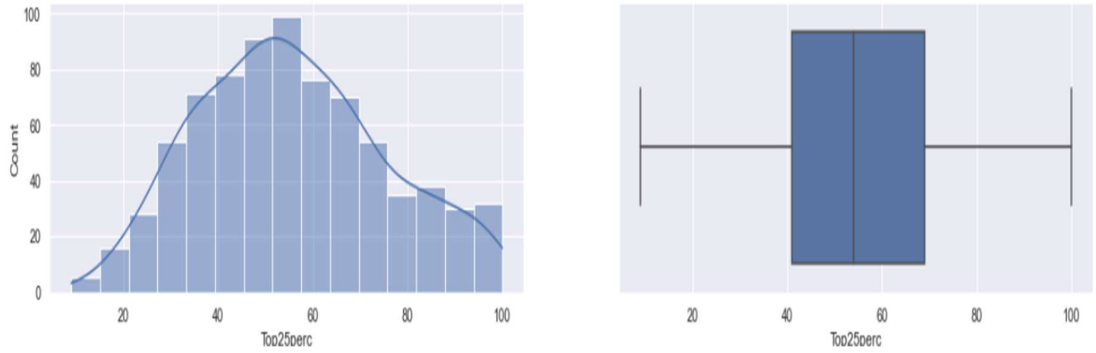


Figure 5 : Top 25 Perc

The box plot for the top 25% has no outliers. The distribution is almost normally distributed. Majority of the students are from top 25% of higher secondary class.

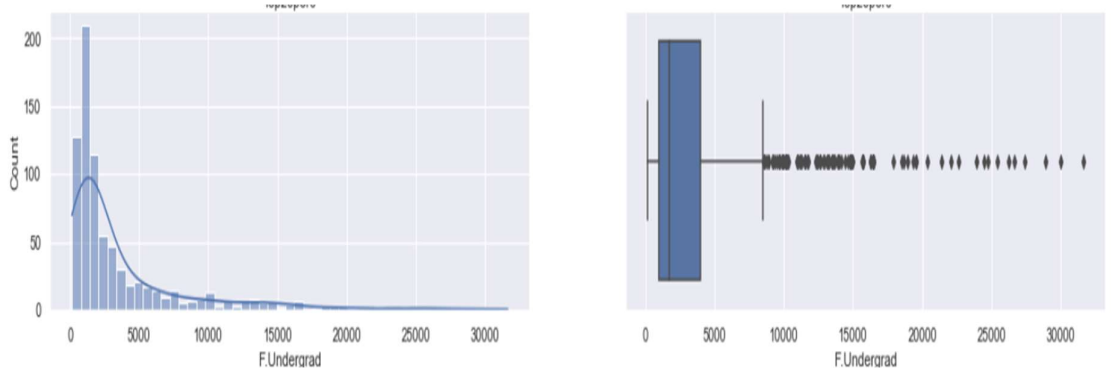


Figure 6 : Full time Undergraduate

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.

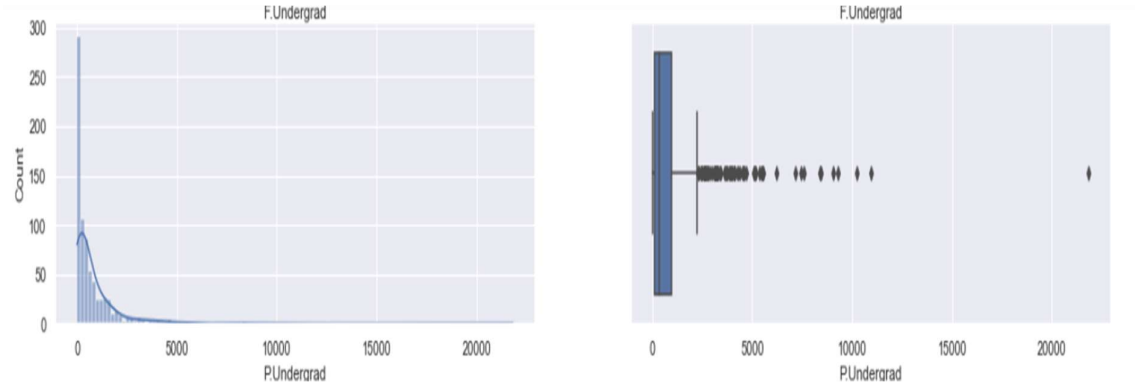


Figure 7 : Part time Undergraduate

The box plot of the part time graduates has outliers. The distribution of the data is positively skewed. In the range about 1000 to 3000 they are part-time graduates studying in all the university.

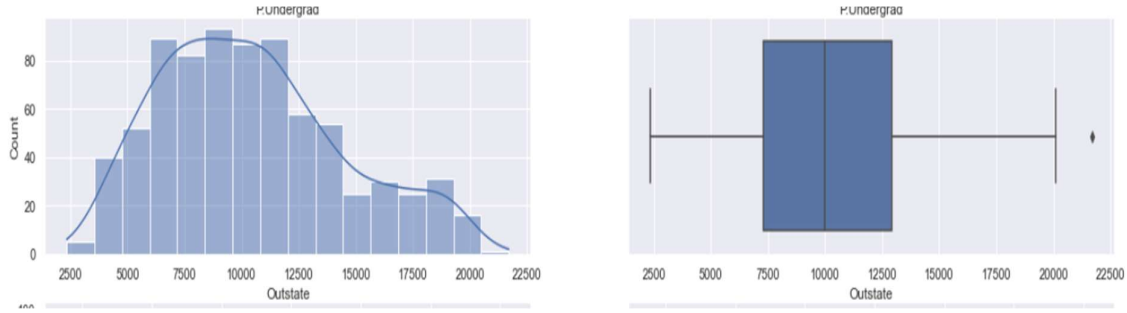


Figure 8 : Outstate

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

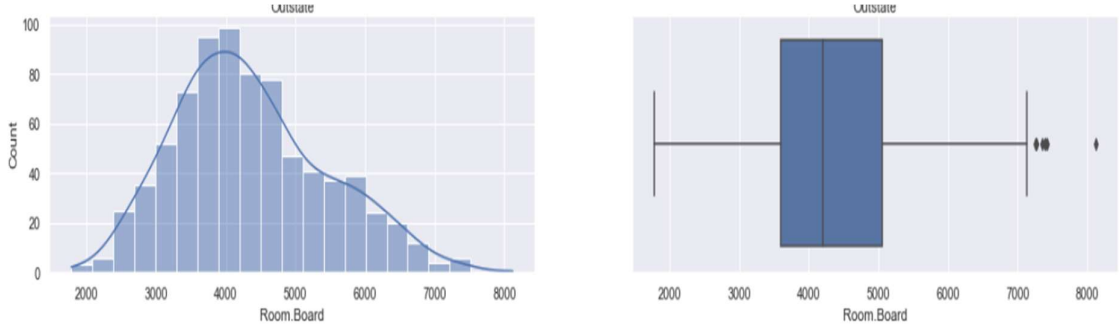


Figure 9 : Room Board

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

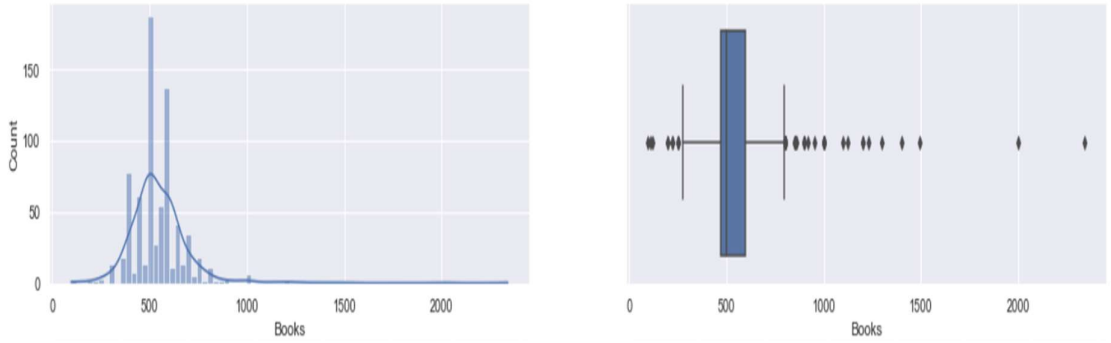


Figure 10 : Books

The box plot of books has outliers. The distribution seems to be bimodal. The cost of books per student seems to be in the range of 500 to 100.

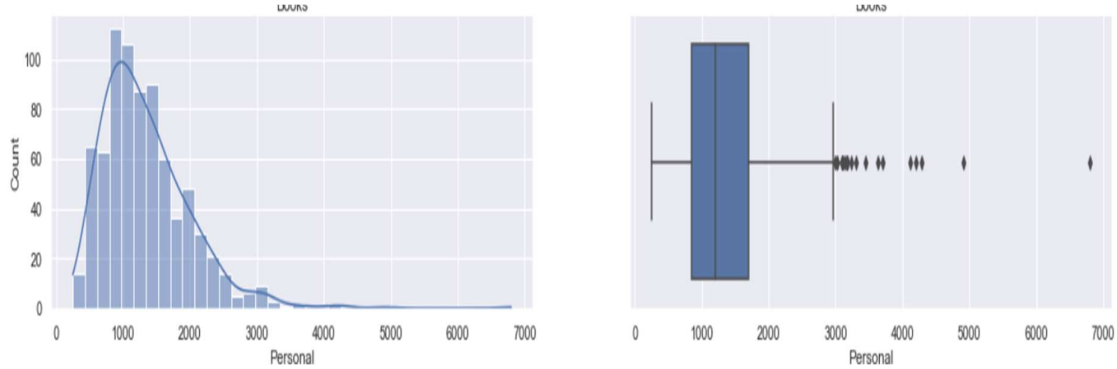


Figure 11 : Personal

The box plot of personal expense has outliers. Some student's personal expense is way bigger than the rest of the students. The distribution seems to be positively skewed.

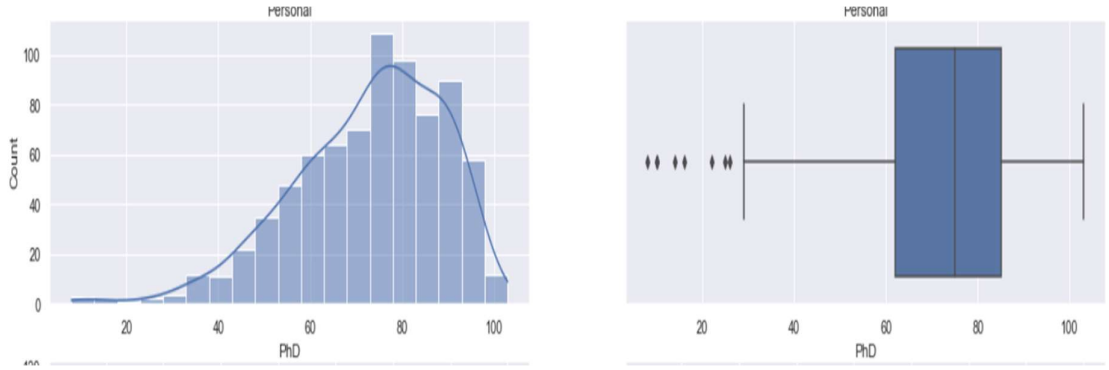


Figure 12 : PhD

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

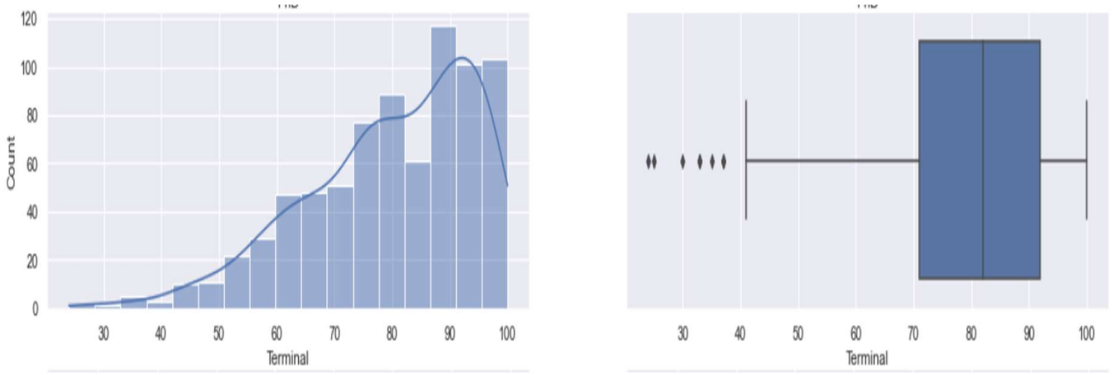


Figure 13 : Terminal

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

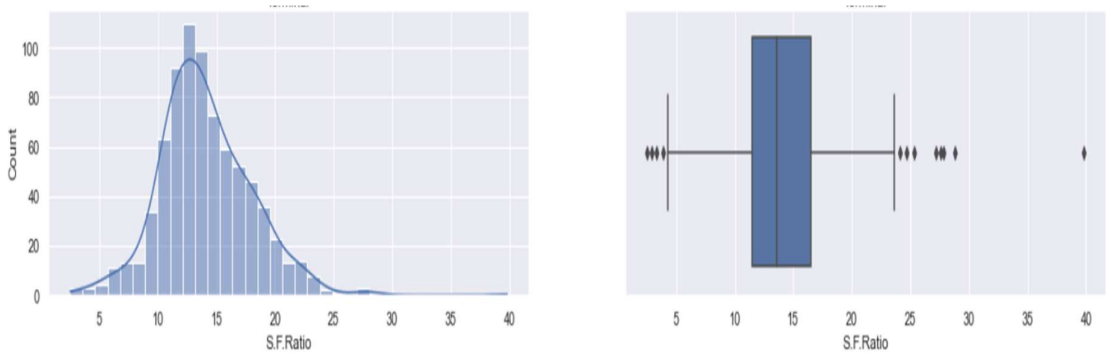


Figure 14 : SF Ratio

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.

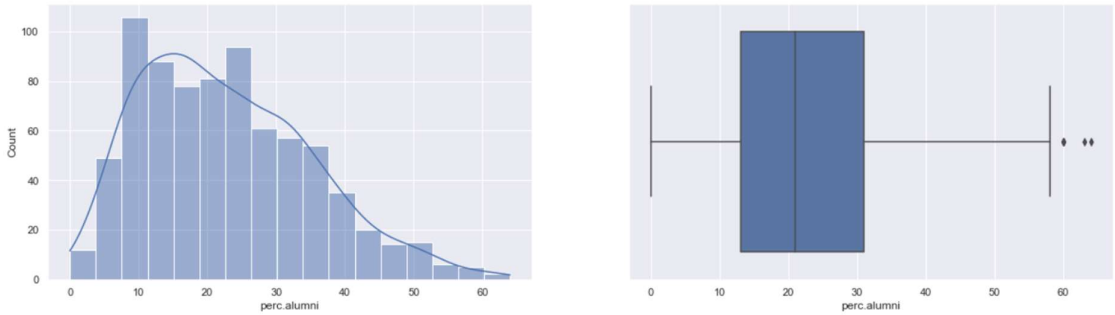


Figure 15 : Perci Alumini

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

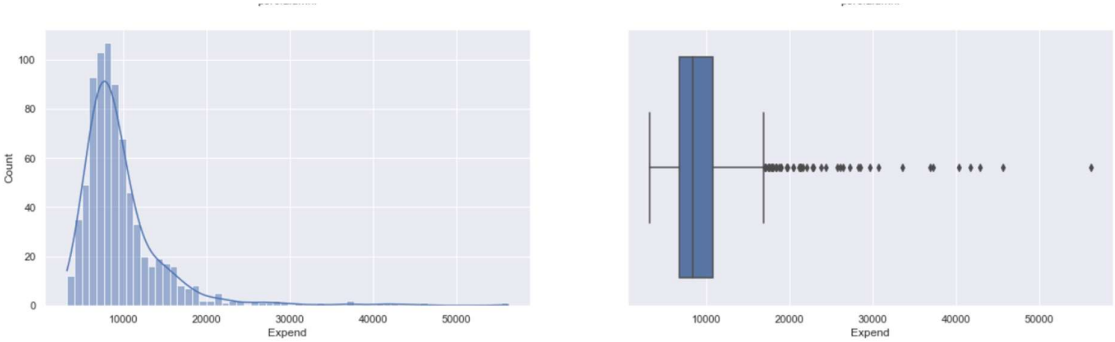


Figure 16 : Expenditure

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

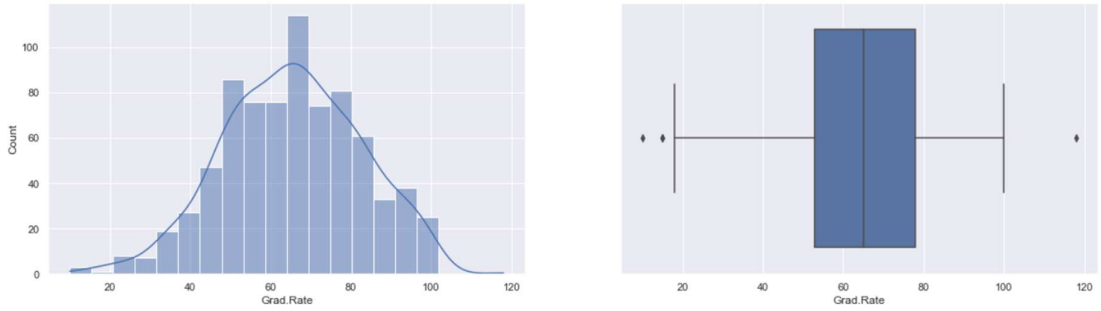


Figure 17 : Grad Rate

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

MULTIVARIATE ANALYSIS

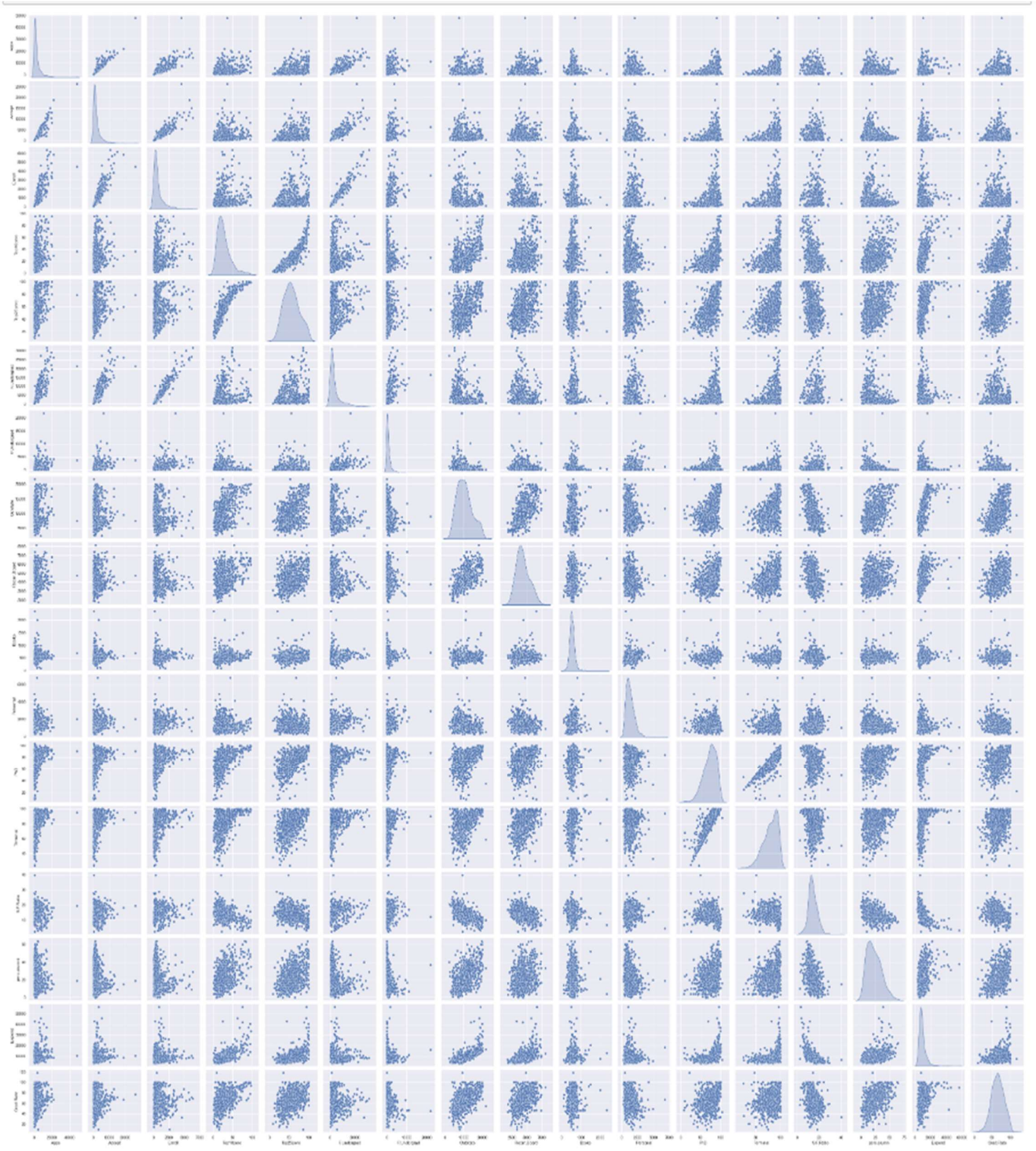


Figure 18 : Pair plot

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.

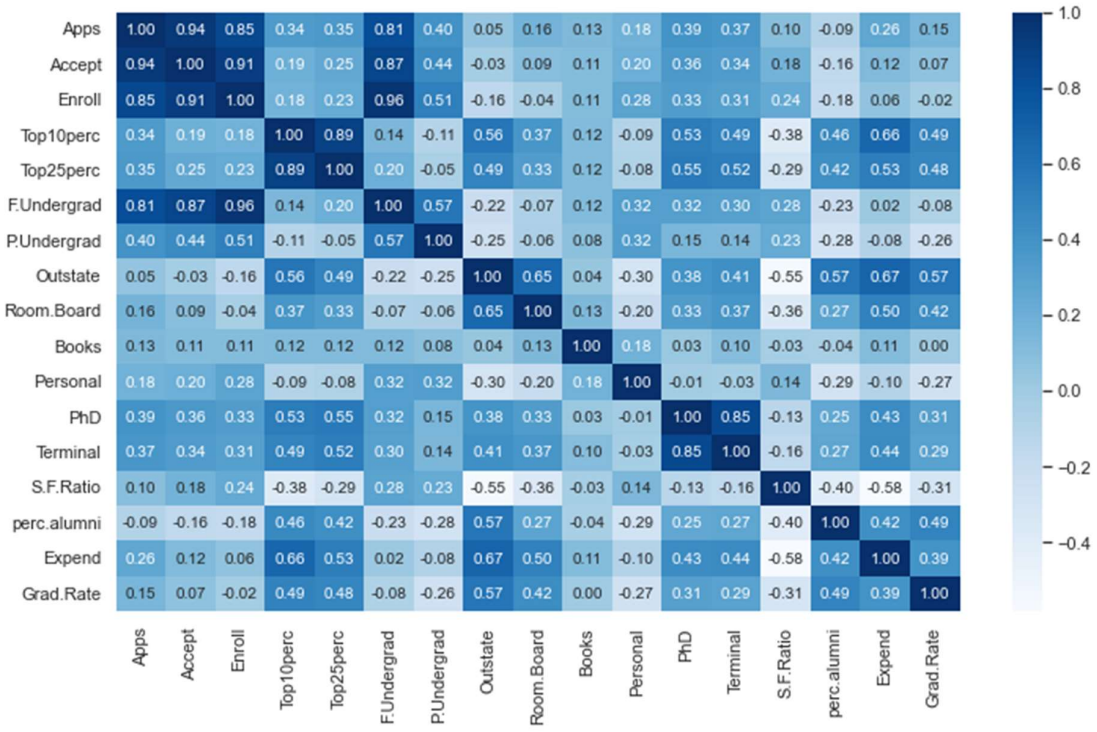


Figure 19 : Pair plot

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. Hence 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.

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

If one component varies less than another because of their respective scales, PCA might determine that the direction of maximal variance are more closely corresponds with each other, if those features are not scaled.

Feature scaling through standardization (or Z-score normalization) can be an important pre-processing step for many machine learning algorithms. Standardization involves rescaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one.

	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	-0.35	-0.32	-0.06	-0.26	-0.19	-0.17	-0.21	-0.75	-0.96	-0.60	1.27	-0.16	-0.12	1.01	-0.87	-0.50	-0.32
1	-0.21	-0.04	-0.29	-0.66	-1.35	-0.21	0.24	0.46	1.91	1.22	0.24	-2.68	-3.38	-0.48	-0.54	0.17	-0.55
2	-0.41	-0.38	-0.48	-0.32	-0.29	-0.55	-0.50	0.20	-0.55	-0.91	-0.26	-1.20	-0.93	-0.30	0.59	-0.18	-0.67
3	-0.67	-0.68	-0.69	1.84	1.68	-0.66	-0.52	0.63	1.00	-0.60	-0.69	1.19	1.18	-1.62	1.15	1.79	-0.38
4	-0.73	-0.76	-0.78	-0.66	-0.60	-0.71	0.01	-0.72	-0.22	1.52	0.24	0.20	-0.52	-0.55	-1.68	0.24	-2.94

Figure 20 : Scaled Data

Z Score tells us how many SD is the point away from the mean and also the direction. Now, we can see that all the variables are scaled by standardizing

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

Both covariance and correlation measure the relationship and the dependency between two variables. Covariance indicates the direction of the linear relationship between variables. Correlation measures both the strength and direction of the linear relationship between two variables.

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

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

```
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]])
```

This above 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

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

$$Correlation = \frac{Cov\ (x, y)}{\sigma x * \sigma y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

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.

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, enrolment and fulltime graduates are highly positively correlated

Also, the top 10 percentage and top 25 percentage are highly positively correlated

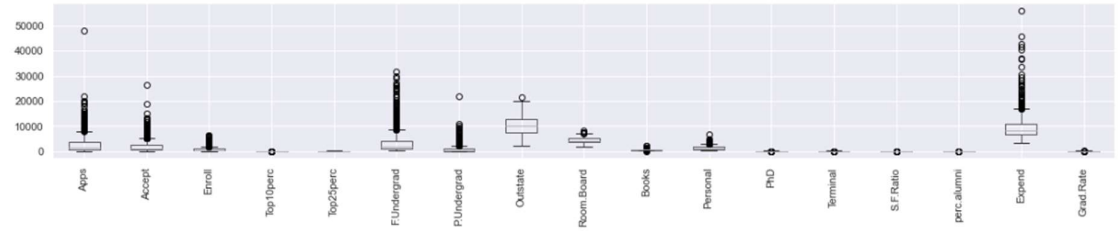
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.00	0.94	0.85	0.34	0.35	0.81	0.40	0.05	0.16	0.13	0.18	0.39	0.37	0.10	-0.09	0.26	0.15
Accept	0.94	1.00	0.91	0.19	0.25	0.87	0.44	-0.03	0.09	0.11	0.20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

Figure 21 : Correlation Matrix

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

Checking the data before scaling

Figure 22 : Outliers before scaling



Checking the dataset after scaling

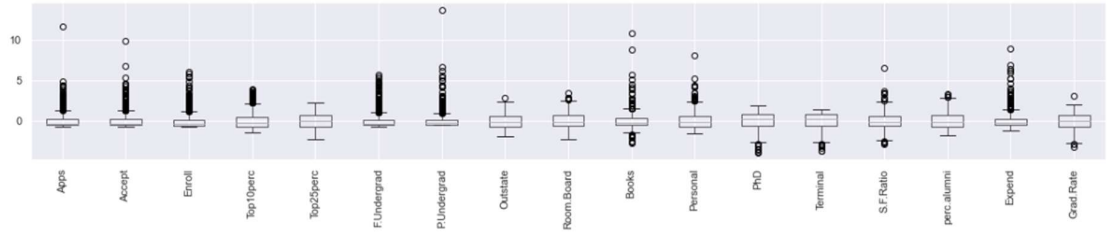


Figure 23 : Outliers after scaling

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.

For example, if we wish to remove outliers, we can consider taking 3 standard deviations as outliers or either we can remove them or impute them with IQR values.

v. Extract the eigenvalues and eigenvectors

<u>Eigen Vectors</u>

```
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]])
```

Eigen 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])
```

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

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

```
array([0.32020628, 0.26340214, 0.06900917, 0.05922989, 0.05488405,
       0.04984701, 0.03558871])
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Apps	0.25	0.33	-0.06	0.28	0.01	-0.02	-0.04
Accept	0.21	0.37	-0.10	0.27	0.06	0.01	-0.01
Enroll	0.18	0.40	-0.08	0.16	-0.06	-0.04	-0.03
Top10perc	0.35	-0.08	0.04	-0.05	-0.40	-0.05	-0.16
Top25perc	0.34	-0.04	-0.02	-0.11	-0.43	0.03	-0.12
F.Undergrad	0.15	0.42	-0.06	0.10	-0.04	-0.04	-0.03
P.Undergrad	0.03	0.32	0.14	-0.16	0.30	-0.19	0.06
Outstate	0.29	-0.25	0.05	0.13	0.22	-0.03	0.11
Room.Board	0.25	-0.14	0.15	0.18	0.56	0.16	0.21
Books	0.06	0.06	0.68	0.09	-0.13	0.64	-0.15
Personal	-0.04	0.22	0.50	-0.23	-0.22	-0.33	0.63
PhD	0.32	0.06	-0.13	-0.53	0.14	0.09	-0.00
Terminal	0.32	0.05	-0.07	-0.52	0.20	0.15	-0.03
S.F.Ratio	-0.18	0.25	-0.29	-0.16	-0.08	0.49	0.22
perc.alumni	0.21	-0.25	-0.15	0.02	-0.22	-0.05	0.24
Expend	0.32	-0.13	0.23	0.08	0.08	-0.30	-0.23
Grad.Rate	0.25	-0.17	-0.21	0.27	-0.11	0.22	0.56

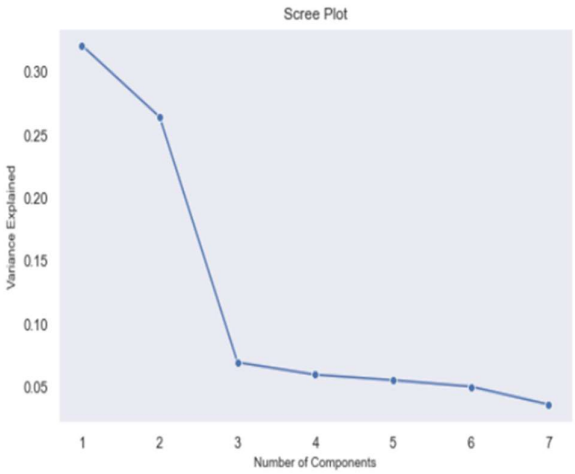


Figure 24 : Reduction of Multi-Collinerity

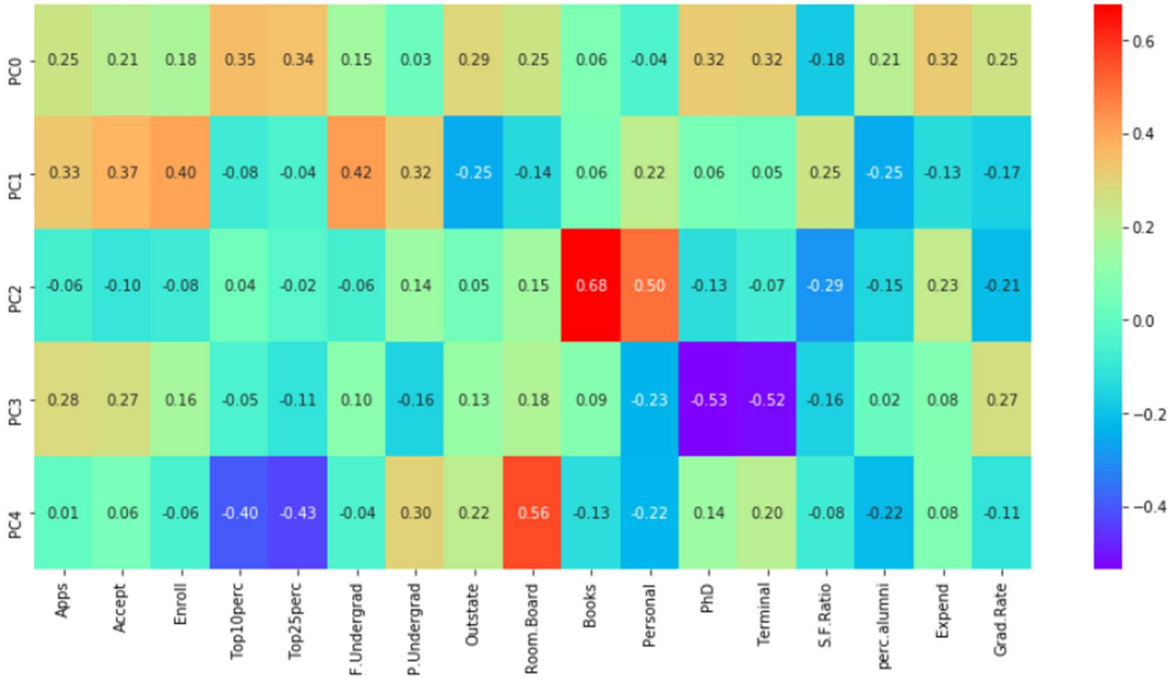


Figure 25 : Heatmap after PCA

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

The Linear eq of 1st component:

0.249 * Apps + 0.208 * Accept + 0.176 * Enroll + 0.354 * Top10perc + 0.344 * Top25perc + 0.155 * F.Undergrad + 0.026 * P.Undergrad + 0.295 * Outstate + 0.249 * Room.Board + 0.065 * Books + -0.043 * Personal + 0.318 * PhD + 0.317 * Terminal + -0.177 * S.F.Ratio + 0.205 * perc.alumni + 0.319 * Expend + 0.252 * Grad.Rate +

Figure 26 : Linear equation

viii. 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?

```
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.        ])
```

Adding the Eigen values, we will get sum of 100.

To decide the optimum number of principal components

1. Check for cumulative variance up to 90%, check the corresponding associated with 90%

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 after this the incremental value between the is less than 5%.

Therefore, we select 6 principal components for this case study.

ix. Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis?

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 5 where we could understand the maximum variance of the dataset. Using the components we can now understand the reduced multicollinearity in the dataset. with this analysis we can perform further analysis and model building PCA will improve the efficiency of machine learning models