ADVANCE STATISTICS PROJECT

NAME: SHOUNACK MANDAL

COURSE: PGP - DSBA Online Sep.

Date: 12/ December / 2021

Table of Content

LIST OF FIGURE

LIST OF TABLES

1. Problem 1 A :- (ANOVA)

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.

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

Null and the alternate hypothesis one way ANOVA for Education:-

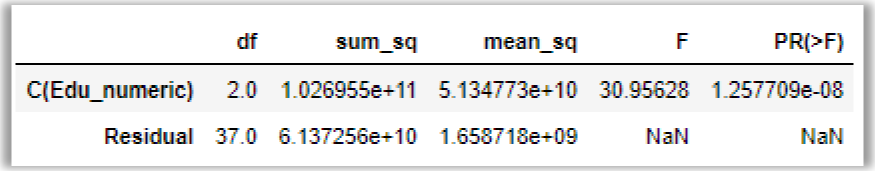
As there are 3 categories in education these are Doctorate, Bachelors, HS-Grad. The Null Hypothesis can be stated as H0: The mean of salary is the same across the 3 categories of education. On the other hand Alternate Hypothesis H1: The mean of salary is different in at least one category of education.

Null and the alternate hypothesis one way ANOVA for Occupation:-

There are 4 categories in occupation these are Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. The Null Hypothesis can be stated as H0: The mean of salary is the same across the 4 categories of education. On the other hand Alternate Hypothesis H1: The mean salary is different in at least one category of education.

1.2: 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.

The table given below is the one way ANOVA on Salary with respect to Education variable: -



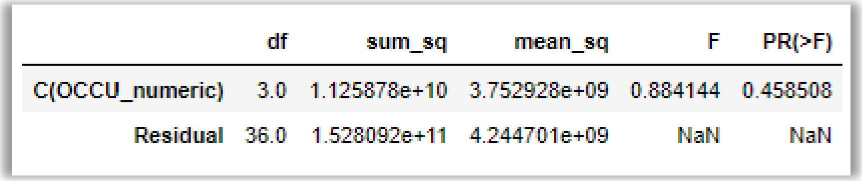
	df	sum_sq	mean_sq	F	PR(>F)
C(Edu_numeric)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

Table 1:- ANOVA for education.

As in the table, P-value (1.257709e-08) i.e. 0.0000000125 which is less than the significance level (0.05). Thus, the Null Hypothesis (H0) is rejected and Alternate Hypothesis (H1) is accepted as concluded there is significant difference in the mean salary for at least one category of education.

1.3: 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.

The one way ANOVA education table given below for Occupation variable: -



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

Table 2:- ANOVA for Occupation

As in the table P-value is 0.458508 which is greater than the significance level (0.05). As per the occupation category Null Hypothesis (H0) is accepted and Alternate Hypothesis (H1) is rejected as concluded there is no significant difference in the mean salary among any of categories of the occupation.

2. Problem 1B :- (ANOVA)

2.1: 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]

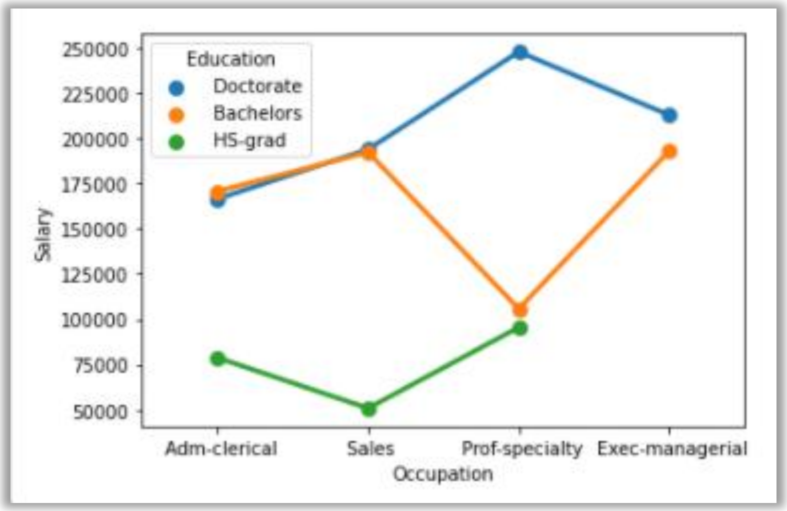


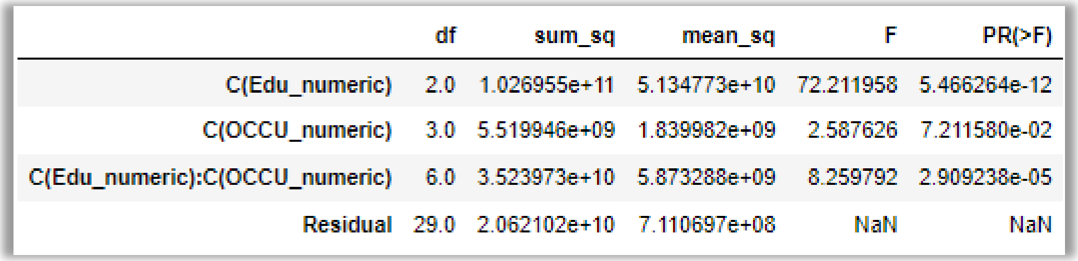
Figure 1: Interaction between the treatments, Education and Occupation.

Few interactions we can observe from this pointplot of salary with Occupation and Education:-

A. Administrative-clerical job is somewhat having same salary for people who are doctors or Bachelors.
B. And, sales job is also having same salary for people who are doctors or Bachelors.
C. Highest salary is being paid for Prof-specialty job role is to doctorates and very minimal salary to people with both the other education levels.
D. For the executive and managerial job role only candidature are who has education level of doctorates and bachelors, and comparatively doctors have better salary than bachelors.
E. High school graduates have very minimal salary i.e. less than 100000 and also the lowest as under sales job with just 50000 and none is working in the occupation for executive or managerial.
F. People with education level of bachelors earning same salary for the occupation sales and executive or managerial.

2.2: 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?

In combination of both the treatments we can perform the **two way ANOVA** which will show interaction among occupation and education also the cause and effect of it on the salary :-



	df	sum_sq	mean_sq	F	PR(>F)
C(Edu_numeric)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(OCCU_numeric)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Edu_numeric):C(OCCU_numeric)	6.0	3.523973e+10	5.873288e+09	8.259792	2.909238e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Table 3: Two way ANOVA showing the interaction effect.

The null and alternate hypothesis for the two way ANOVA of Salary with respect to both Education and Occupation is:-

Null hypothesis (Ho): There is no interaction effect between the treatments, education and occupation.

Alternate hypothesis (H1): There is an interaction effect between the treatments, education and occupation.

INTERPRETATION: -

The 2 way ANOVA shows the combination of (education and occupation) has P- value of 2.909238e-05 is less than 0.05 means rejecting the null hypothesis (H0). Thus in cognizance of this fact we get that there is some interactions between occupation and education. There is proper evidence that there is an impact on the salary as due to any change in the treatments. So we can say as the treatments have significant interaction effect which causes salary as optically we have seen that with combination of occupation and education, Professional or specialty and doctorate respectively results higher and better salary.

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

Business implication for performing ANOVA for this particular case study are:-

1. It shows the dependency relations among the variable and that can be visually understand through univariate, bivariate and multi-variate analysis of the data.
2. We get to se the prominent factor among the treatments is education which causes more impacts on the salary. Followed by occupation which has lesser significance.
3. So, ANOVA mainly shows the dependency in this case and with following outputs we can predict the constrain effect more to the salary.

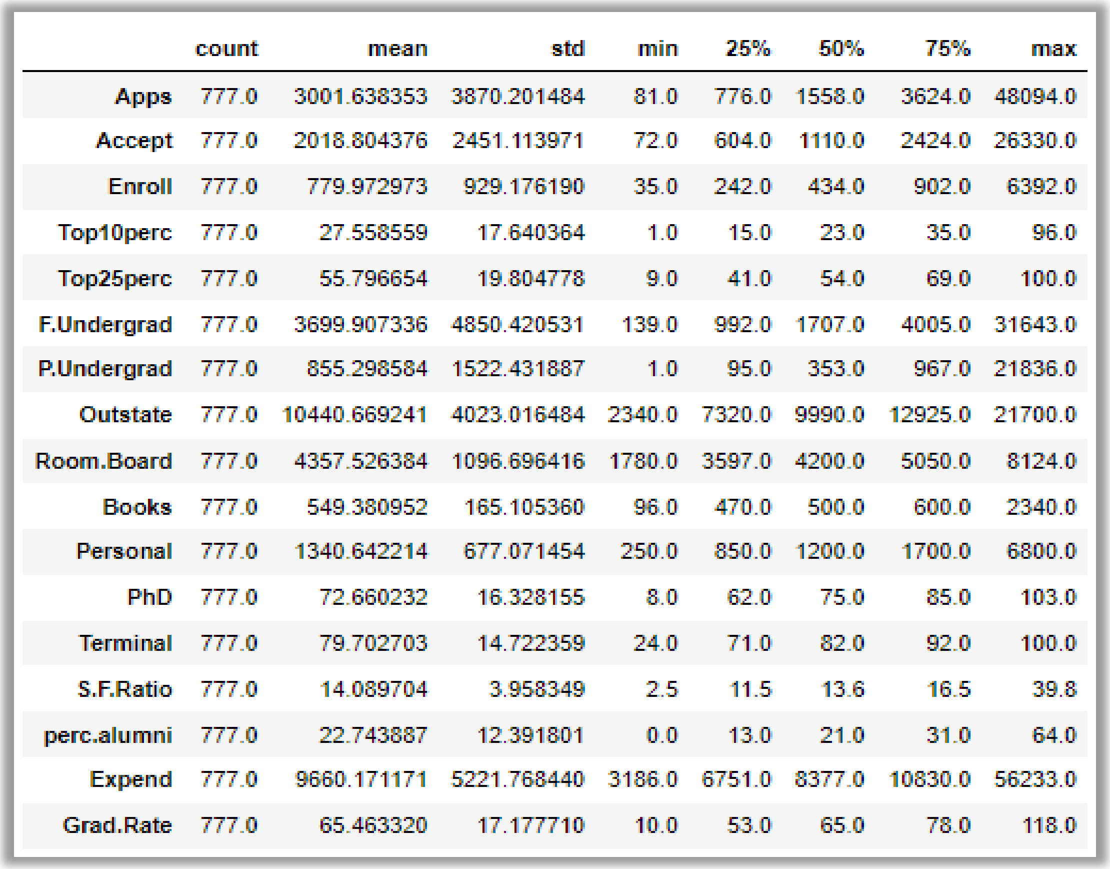
3: Problem 2 : (EDA & PCA)

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.

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

Univariate Analysis

To summarize the data here is the statistical description given:-



	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

Table 4: Statistical description of the data.

Every variable has 777 counts except the name variable which is categorical one. There isn`t any missing values in any of the columns or variables. For the exploratory data analysis we have 18 columns/variables. Figures below shows the univatiate analysis with the histograph and boxplots:-

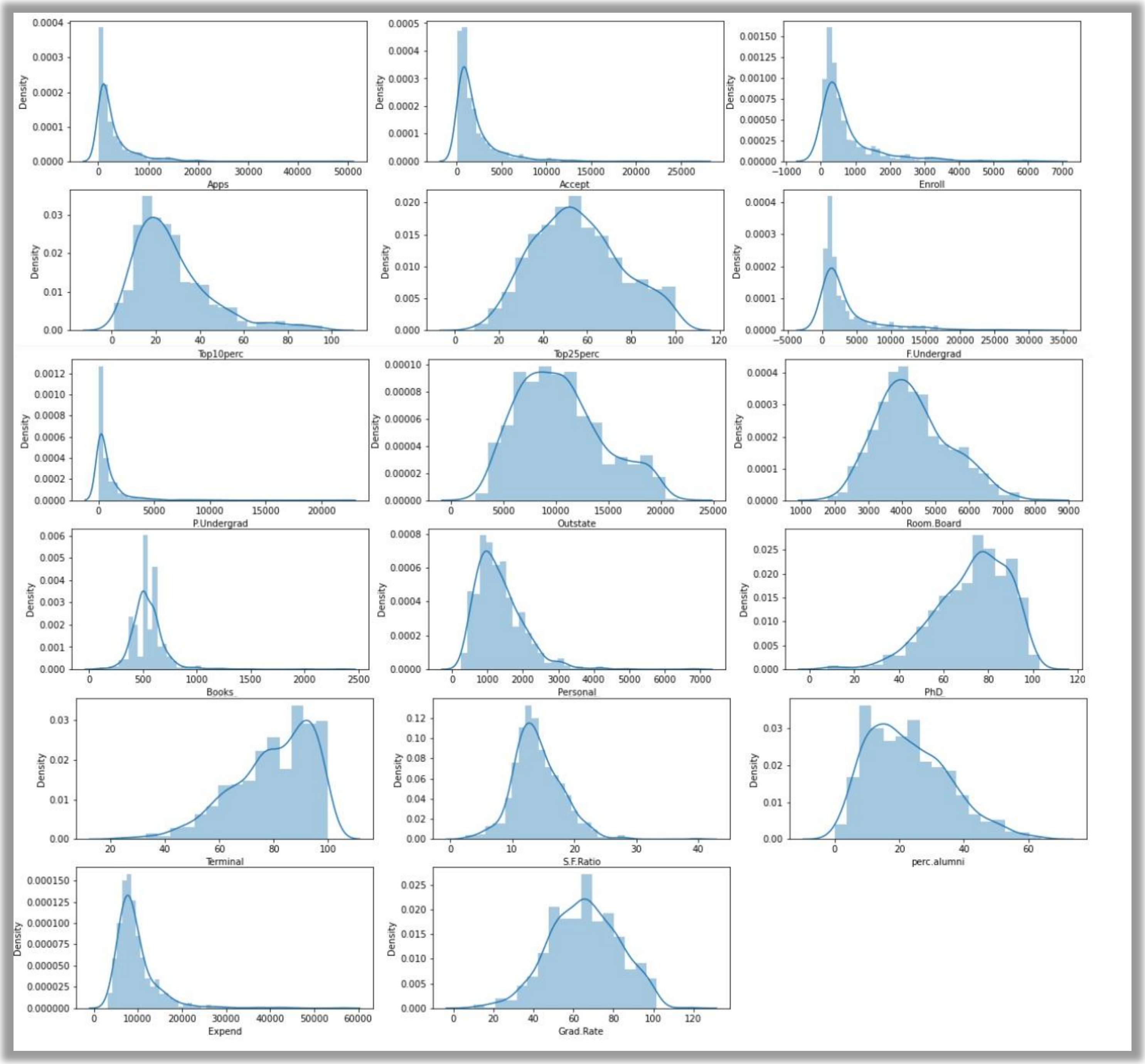


Figure2: histograph showing the univariate variables

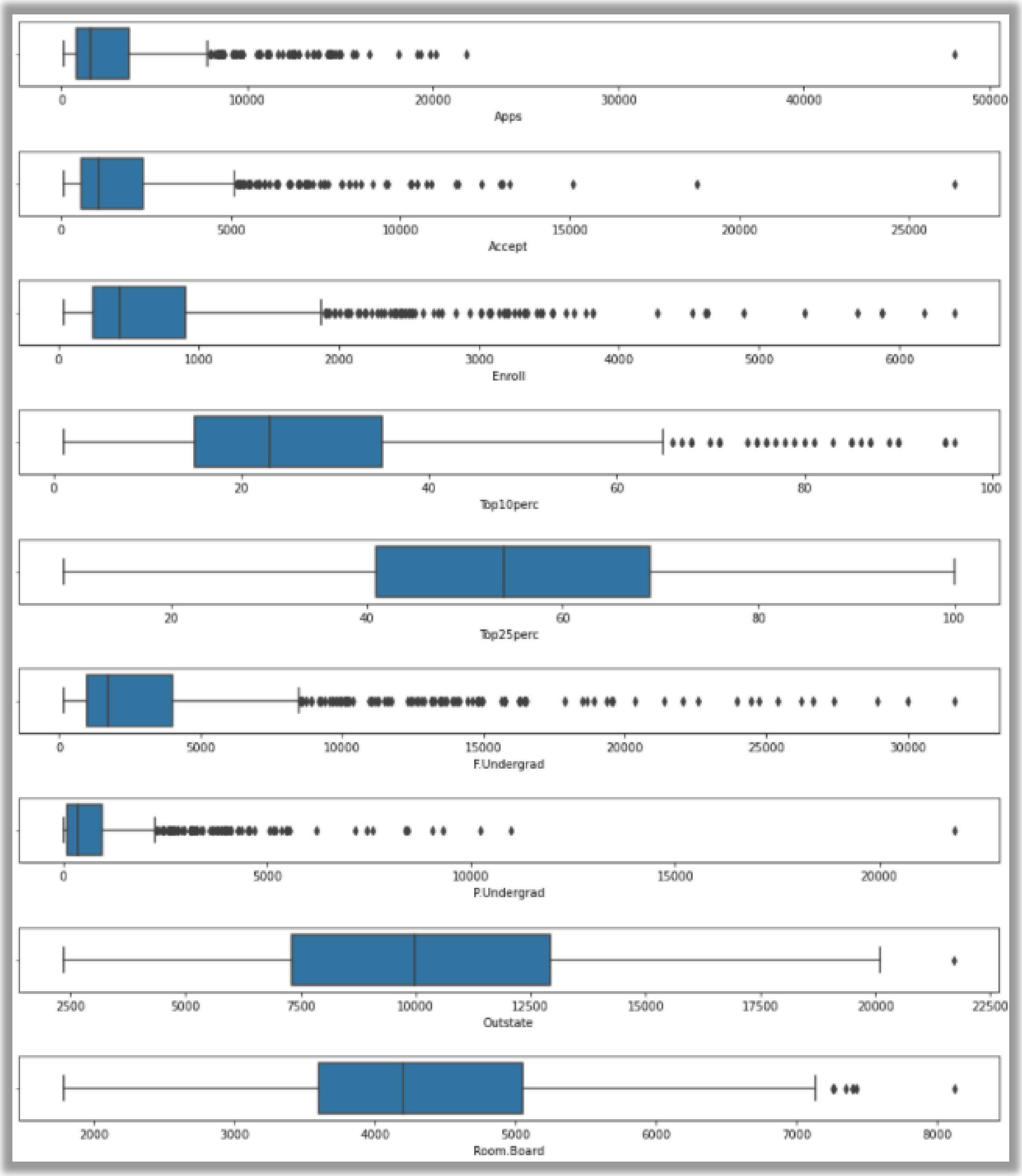


Figure 3: Boxplot for scaled data part 1

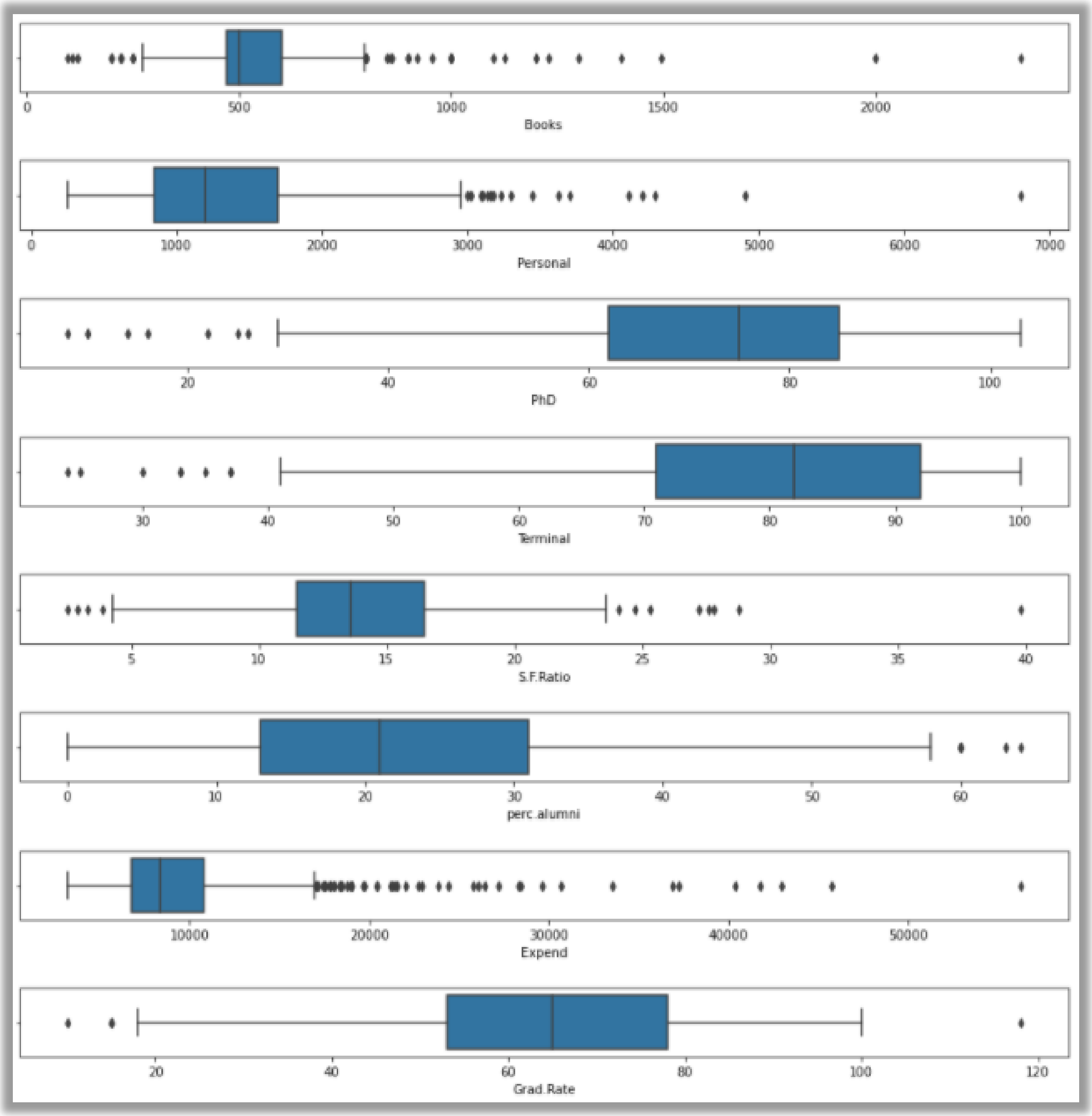


Figure 4: Boxplot for scaled data part 2

All the variables have outliers except the variable **top 25 perc.** The maximum number of applicant for any college went up to **50,000**. **Application** for all the colleges has formed a distribution of positive skewed histogram. From all the applicants almost half candidates have been selected into the colleges as showing the **acceptance (accept)** number went up to 25000 with positively skewed histogram. The **enrolment (enroll)** went up to 7000 and majority of the colleges has only 1000 students enrolled. Among those 1000 students on an average we can see only 100 students able to rank and place themselves above the **top 10 percentage.** Followed by 120 students above the **top 25 percentage**. Full time undergraduate and part time undergraduates has students in all universities as approx 35000 and 20000 approx, respectively. **Outstation** students are seems to be present with a normal distribution in all universities. As followed the students become graduates with at least 60 % from all the universities.

MULTIVARIATE ANALYSIS :-

A. Pairplot:-

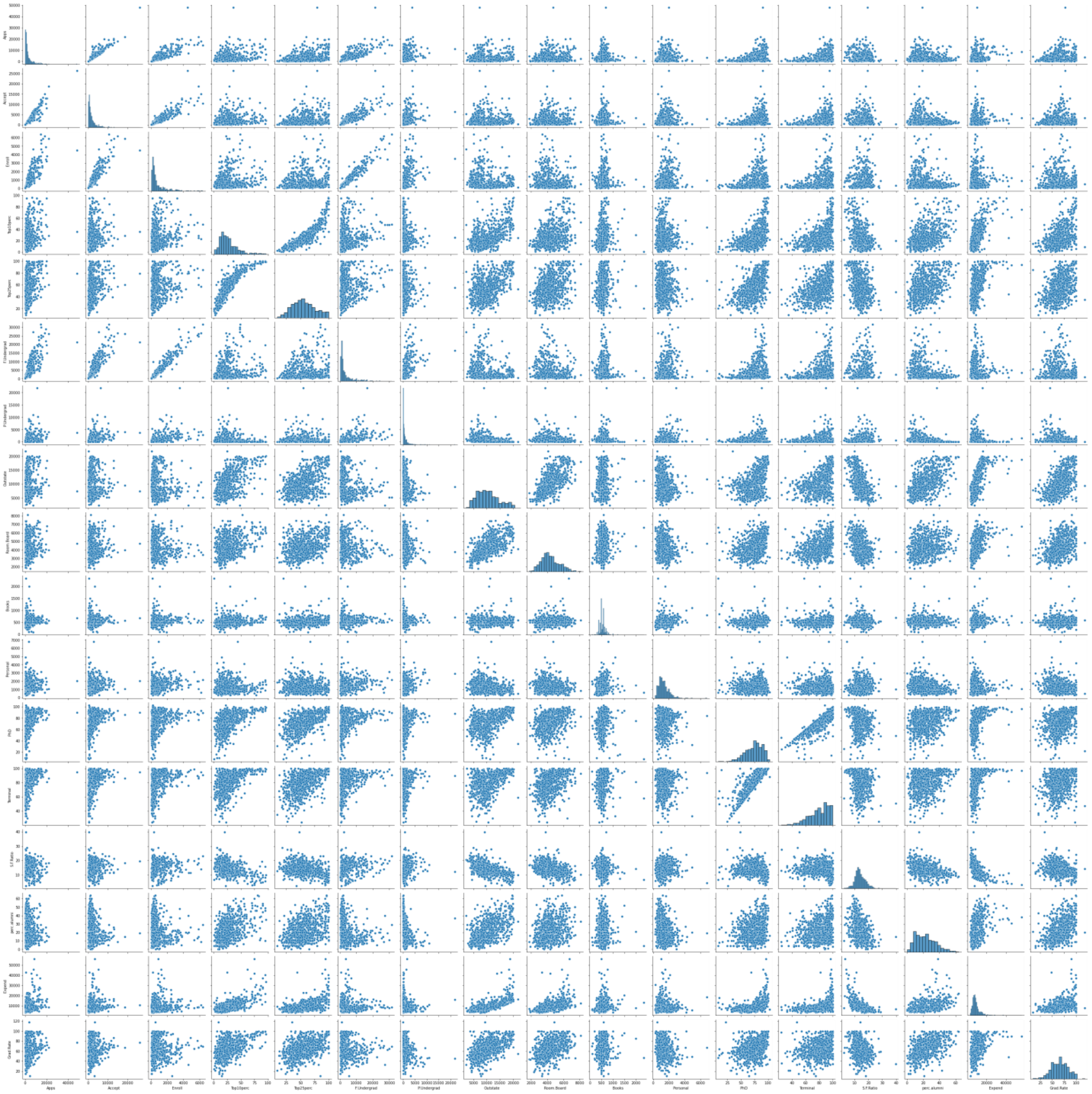


Figure 5 Multivariate Analysis

B. Heatmap:-

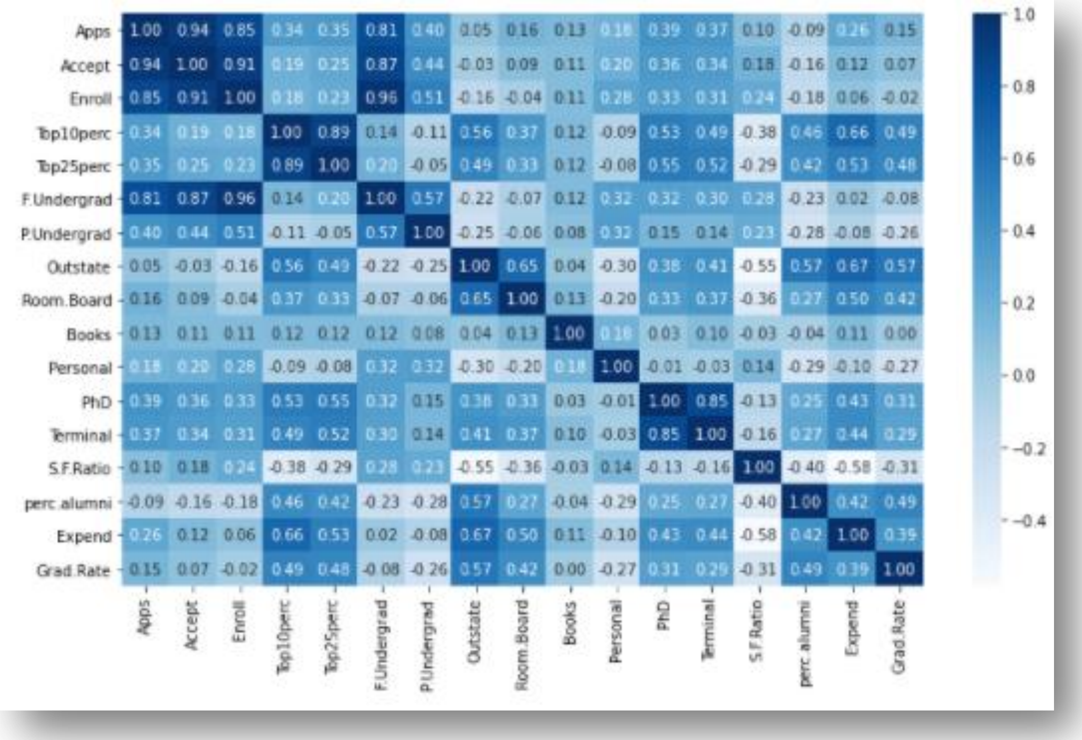


Figure 6: Heat Map

Heat map shows multicollinearity. **Applicants, Acceptance, Enroll and Full time undergraduates** are highly positively correlated among each other as the value is **above 0.75,** also the value **between top 25 and top 10 percentages**. As none apart from these variables has such high values for correlation with each other the other correlation values are fine and can be move forward with for the further modeling and analysis. So either way we can drop the values and once the values are dropped then we can resolve the issues of multicollinearity.

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

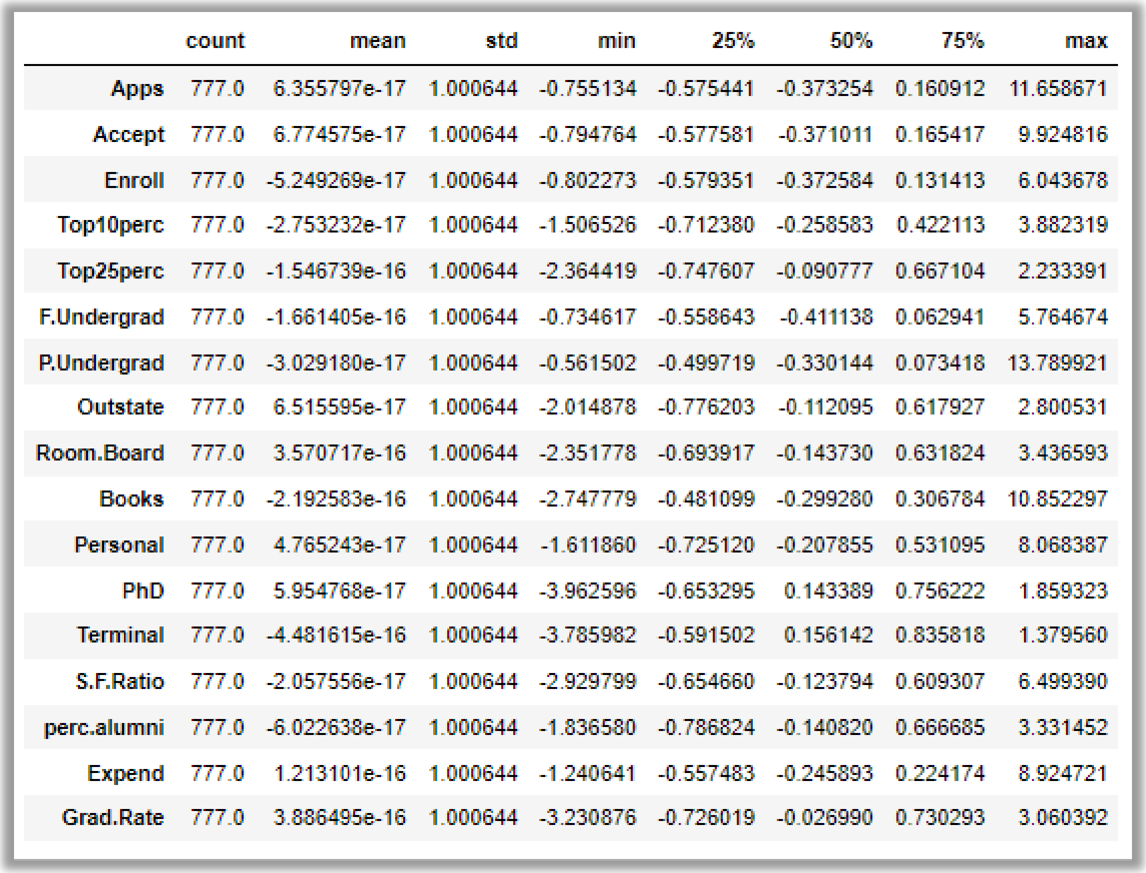
The value need to be numerical so we already dropped the categorical values from the original data. Now we have all 17 variables data in numerical values. Now we can perform the normalization te3chniques, particularly I used the z-score normalization with the given formulae:-

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

Figure 7: Z-Score formulae

After scaling the data description has certain changes from the original data:-

1. The mean values were very high, now each mean value is below 1.
2. Standard deviation value is same for all the 17 variables i.e. 1.
3. The quarter values changes from the original data, it equalize and minimize the values.



	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

Table 5: Statistical description of the scaled dataet.

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

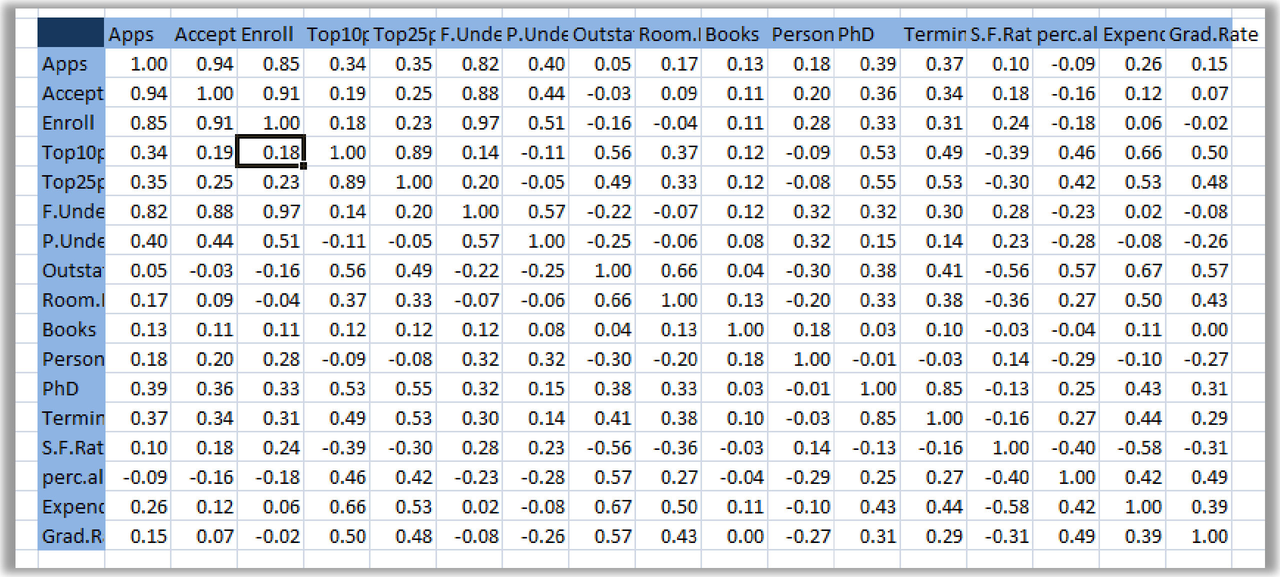
In comparison to the covariance and correlation matrix, we have following knowledge of it. Both covariance and correlation measure the relationship and the dependency between two variables. **Covariance** indicates the direction of the linear relationship between variables. **Correlation** measures both the strength and direction of the linear relationship between two variables. Correlation values are standardized. Direction meaning weather the relationship between the two variable directly proportional or inversely proportional.

$$Cov(X,Y) = \frac{\Sigma(X_i - \bar{X})\,(Y_i - \bar{Y})}{n}$$

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

Figure 8: Covariance and correlation formulae..

Even after the data is normalized the covariance and correlation matrix shows same values this is because after scaling also the value indicates the actual value. Hence the correlation or covariance values doesn`t change. The given below table for correlation and covariance is performed after the scaling is done on the actual data.



	Apps	Accept	Enroll	Top10p	Top25p	F.Unde	P.Unde	Outsta	Room.I	Books	Person	PhD	Termin	S.F.Rat	perc.al	Expenc	Grad.Rate
Apps	1.00	0.94	0.85	0.34	0.35	0.82	0.40	0.05	0.17	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	0.88	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	0.97	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31	0.24	-0.18	0.06	-0.02
Top10p	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	-0.39	0.46	0.66	0.50
Top25p	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33	0.12	-0.08	0.55	0.53	-0.30	0.42	0.53	0.48
F.Unde	0.82	0.88	0.97	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.Unde	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
Outsta	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.66	0.04	-0.30	0.38	0.41	-0.56	0.57	0.67	0.57
Room.I	0.17	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.66	1.00	0.13	-0.20	0.33	0.38	-0.36	0.27	0.50	0.43
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
Person	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
Termin	0.37	0.34	0.31	0.49	0.53	0.30	0.14	0.41	0.38	0.10	-0.03	0.85	1.00	-0.16	0.27	0.44	0.29
S.F.Rat	0.10	0.18	0.24	-0.39	-0.30	0.28	0.23	-0.56	-0.36	-0.03	0.14	-0.13	-0.16	1.00	-0.40	-0.58	-0.31
perc.al	-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
Expenc	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.R	0.15	0.07	-0.02	0.50	0.48	-0.08	-0.26	0.57	0.43	0.00	-0.27	0.31	0.29	-0.31	0.49	0.39	1.00

Table 6: Correlation and covariance matrix

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

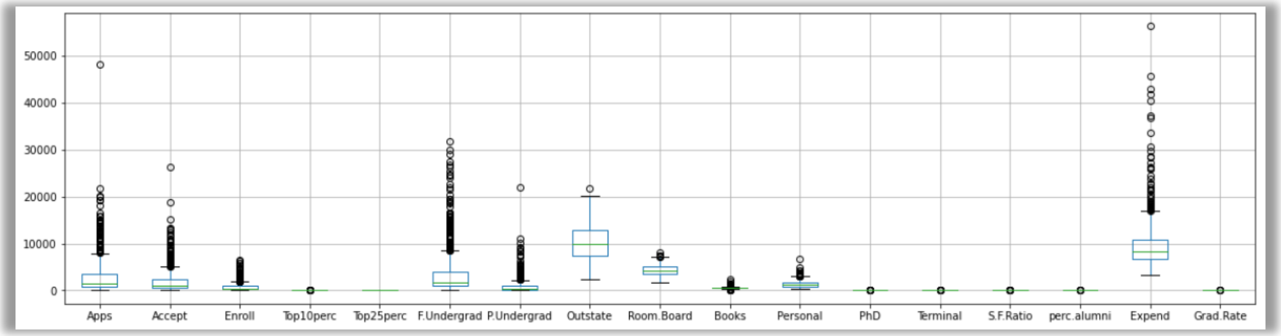


Figure 9: Boxplot on unscaled dataset.

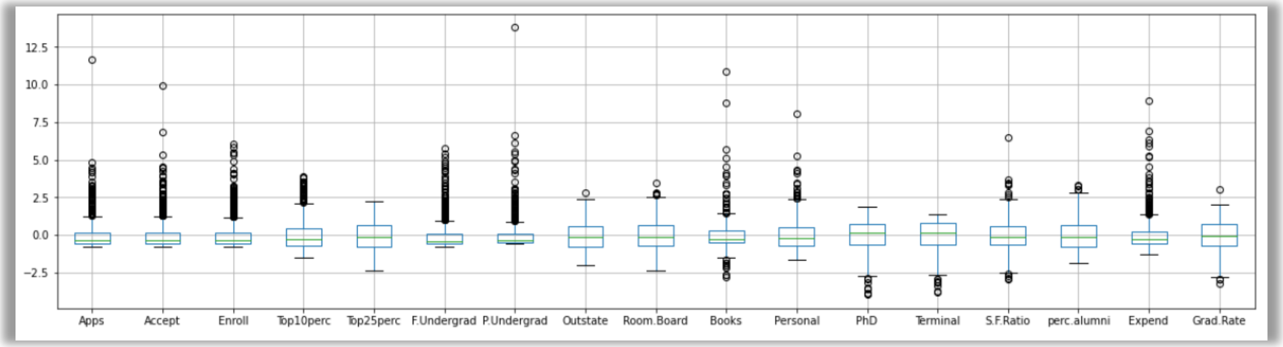
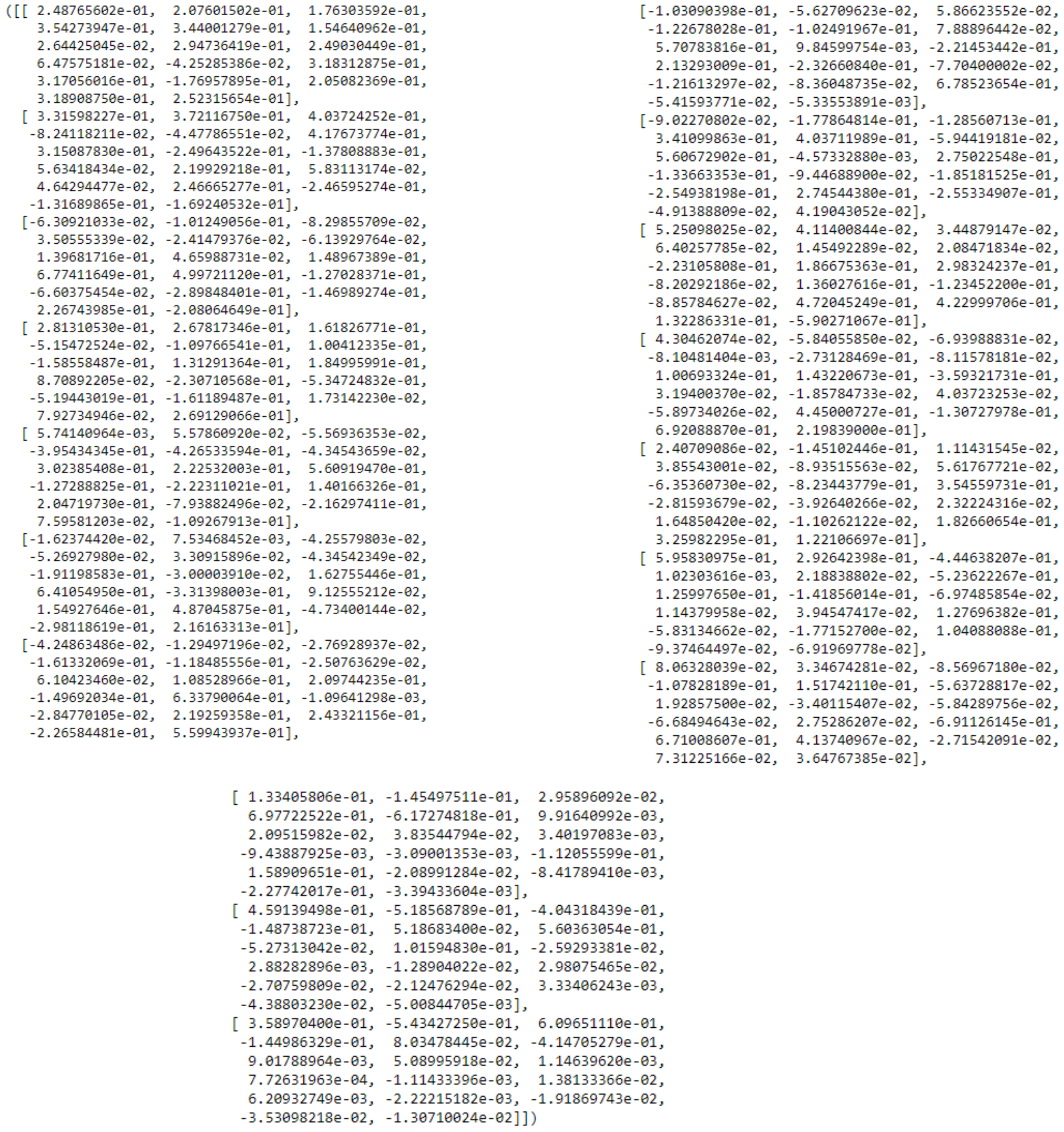


Figure 10: Boxplot on scaled dataset.

Above given figure is before and after the scaling technique is processed, i.e. using z score method of standardization. We can observe in the figure prepared before the scaling method done is having the entire variable in different scales but after the standardization we get the variables aligned with the same median values and the inner quartile ranges are also quite clearly visible. So one scale is showing the values for all variables including its IQR and outliers.

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

Eigen Vectors:



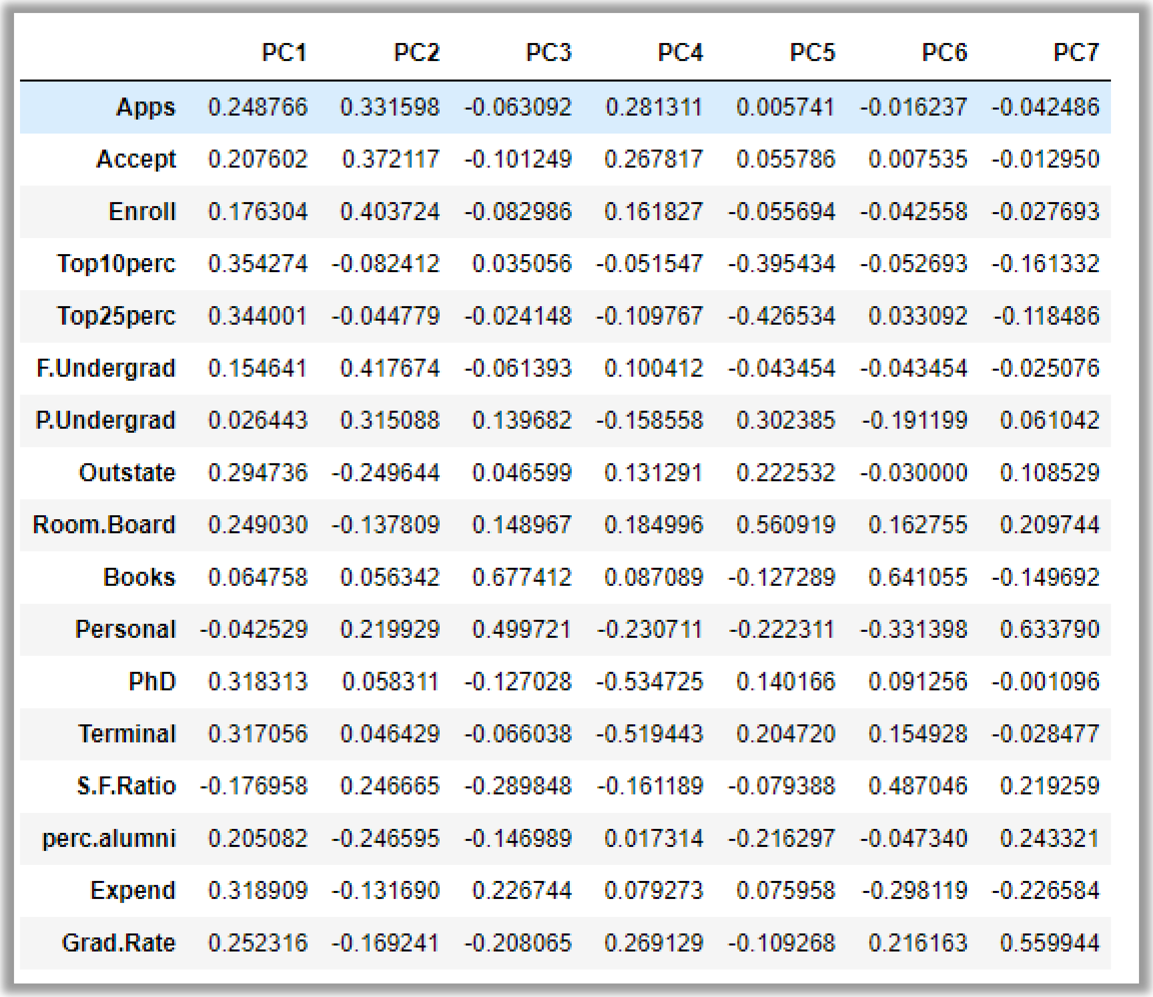
```
([[ 2.48765602e-01,  2.07601502e-01,  1.76303592e-01,
    3.54273947e-01,  3.44001279e-01,  1.54640962e-01,
    2.64425045e-02,  2.94736419e-01,  2.49030449e-01,
    6.47575181e-02, -4.25285386e-02,  3.18312875e-01,
    3.17056016e-01, -1.76957895e-01,  2.05082369e-01,
    3.18908750e-01,  2.52315654e-01],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

Figure 11: Eigen vectors.

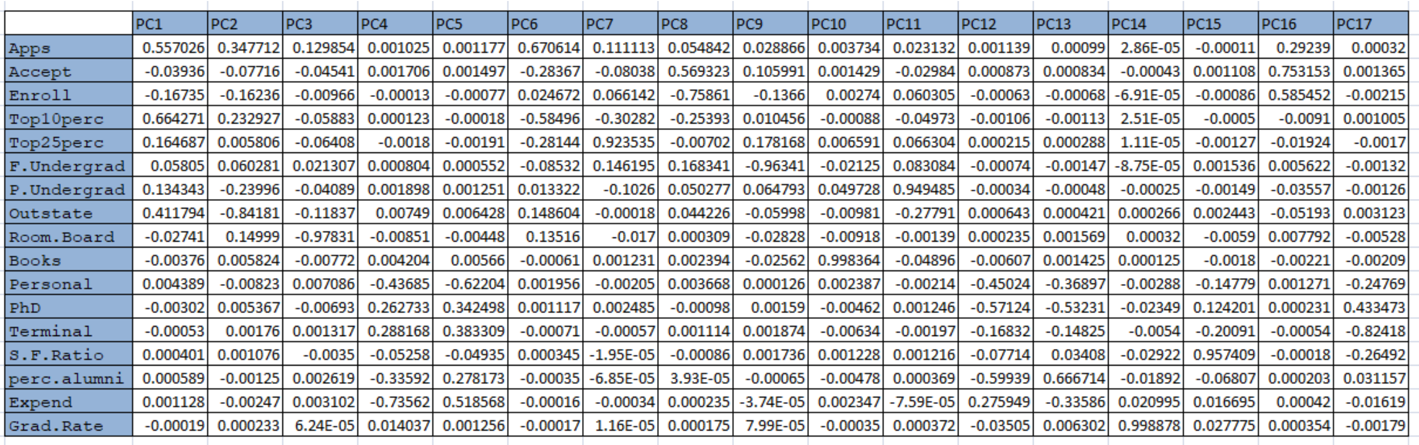
Eigenvalues:-



	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	0.219259
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944

Table 7 Eigen Values.

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



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	0.557026	0.347712	0.129854	0.001025	0.001177	0.670614	0.111113	0.054842	0.028866	0.003734	0.023132	0.001139	0.00099	2.86E-05	-0.00011	0.29239	0.00032
Accept	-0.03936	-0.07716	-0.04541	0.001706	0.001497	-0.28367	-0.08038	0.569323	0.105991	0.001429	-0.02984	0.000873	0.000834	-0.00043	0.001108	0.753153	0.001365
Enroll	-0.16735	-0.16236	-0.00966	-0.00013	-0.00077	0.024672	0.066142	-0.75861	-0.1366	0.00274	0.060305	-0.00063	-0.00068	-6.91E-05	-0.00086	0.585452	-0.00215
Top10perc	0.664271	0.232927	-0.05883	0.000123	-0.00018	-0.58496	-0.30282	-0.25393	0.010456	-0.00088	-0.04973	-0.00106	-0.00113	2.51E-05	-0.0005	-0.0091	0.001005
Top25perc	0.164687	0.005806	-0.06408	-0.0018	-0.00191	-0.28144	0.923535	-0.00702	0.178168	0.006591	0.066304	0.000215	0.000288	1.11E-05	-0.00127	-0.01924	-0.0017
F.Undergrad	0.05805	0.060281	0.021307	0.000804	0.000552	-0.08532	0.146195	0.168341	-0.96341	-0.02125	0.083084	-0.00074	-0.00147	-8.75E-05	0.001536	0.005622	-0.00132
P.Undergrad	0.134343	-0.23996	-0.04089	0.001898	0.001251	0.013322	-0.1026	0.050277	0.064793	0.049728	0.949485	-0.00034	-0.00048	-0.00025	-0.00149	-0.03557	-0.00126
Outstate	0.411794	-0.84181	-0.11837	0.00749	0.006428	0.148604	-0.00018	0.044226	-0.05998	-0.00981	-0.27791	0.000643	0.000421	0.000266	0.002443	-0.05193	0.003123
Room.Board	-0.02741	0.14999	-0.97831	-0.00851	-0.00448	0.13516	-0.017	0.000309	-0.02828	-0.00918	-0.00139	0.000235	0.001569	0.00032	-0.0059	0.007792	-0.00528
Books	-0.00376	0.005824	-0.00772	0.004204	0.00566	-0.00061	0.001231	0.002394	-0.02562	0.998364	-0.04896	-0.00607	0.001425	0.000125	-0.0018	-0.00221	-0.00209
Personal	0.004389	-0.00823	0.007086	-0.43685	-0.62204	0.001956	-0.00205	0.003668	0.000126	0.002387	-0.00214	-0.45024	-0.36897	-0.00288	-0.14779	0.001271	-0.24769
PhD	-0.00302	0.005367	-0.00693	0.262733	0.342498	0.001117	0.002485	-0.00098	0.00159	-0.00462	0.001246	-0.57124	-0.53231	-0.02349	0.124201	0.000231	0.433473
Terminal	-0.00053	0.00176	0.001317	0.288168	0.383309	-0.00071	-0.00057	0.001114	0.001874	-0.00634	-0.00197	-0.16832	-0.14825	-0.0054	-0.20091	-0.00054	-0.82418
S.F.Ratio	0.000401	0.001076	-0.0035	-0.05258	-0.04935	0.000345	-1.95E-05	-0.00086	0.001736	0.001228	0.001216	-0.07714	0.03408	-0.02922	0.957409	-0.00018	-0.26492
perc.alumni	0.000589	-0.00125	0.002619	-0.33592	0.278173	-0.00035	-6.85E-05	3.93E-05	-0.00065	-0.00478	0.000369	-0.59939	0.666714	-0.01892	-0.06807	0.000203	0.031157
Expend	0.001128	-0.00247	0.003102	-0.73562	0.518568	-0.00016	-0.00034	0.000235	-3.74E-05	0.002347	-7.59E-05	0.275949	-0.33586	0.020995	0.016695	0.00042	-0.01619
Grad.Rate	-0.00019	0.000233	6.24E-05	0.014037	0.001256	-0.00017	1.16E-05	0.000175	7.99E-05	-0.00035	0.000372	-0.03505	0.006302	0.998878	0.027775	0.000354	-0.00179

Table 8: Exported data frame of principal components.

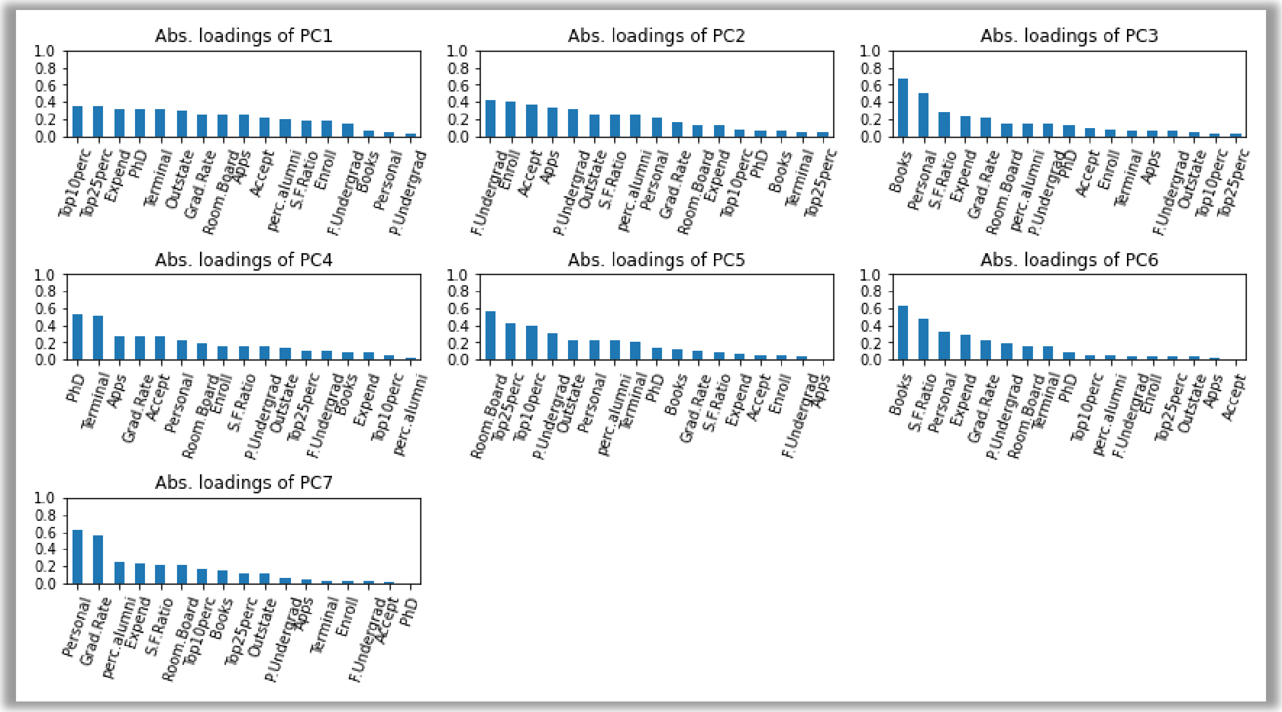
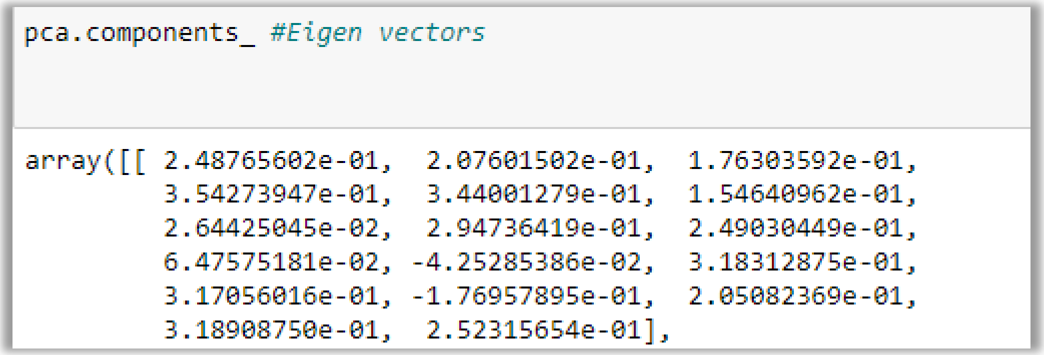


Figure12: Loading of each selected principal component.

3.7: 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]



```
pca.components_ #Eigen vectors
```

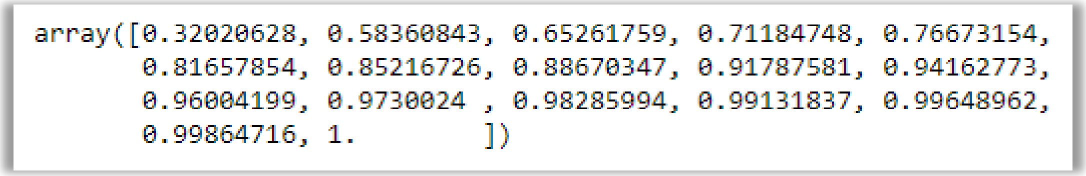
```
array([[ 2.48765602e-01,  2.07601502e-01,  1.76303592e-01,
         3.54273947e-01,  3.44001279e-01,  1.54640962e-01,
         2.64425045e-02,  2.94736419e-01,  2.49030449e-01,
         6.47575181e-02, -4.25285386e-02,  3.18312875e-01,
         3.17056016e-01, -1.76957895e-01,  2.05082369e-01,
 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

Figure 13: First principal component.

The Linear equation for the 1st principal component is = A1X1 +A2X2 +A3X3… .AnXn.

‘Apps’ * 0.25 + 'Accept' * 0.21 , 'Enroll' * 0.18, 'Top10perc' *0.35, 'Top25perc' * 0.34, 'F.Undergrad' * 0.15, 'P.Undergrad' * 0.02, 'Outstate' * 0.29, 'Room.Board' * 0.25, 'Books' *0.06 , 'Personal' * 0.04, 'PhD' *0.32, 'Terminal' *0.32 , 'S.F.Ratio' * 0.18, 'perc.alumni' * 0.20, 'Expend'* 0.32, 'Grad.Rate*0.25.

3.8: 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?



```
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.        ])
```

Figure 14: Cumulative values for the principal

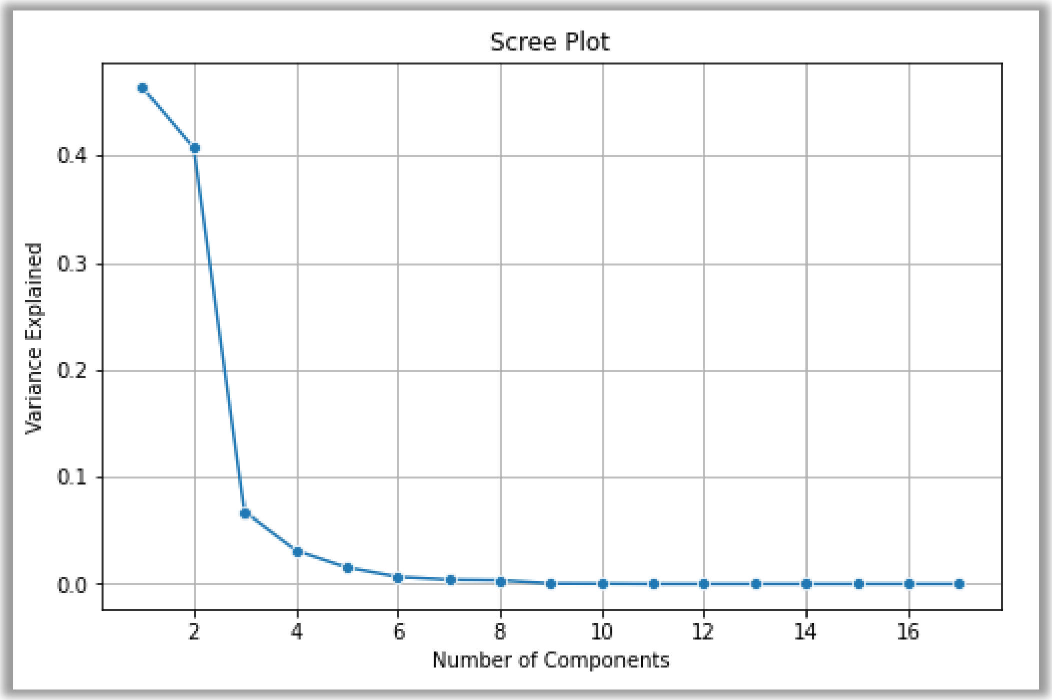


Figure 15 Scree plot.

1. This data shows the individual PCs and their ratio of explaining the variance of the original data.
2. The very first component can reduce the variance among the variable of the original dataset by 32.02%, and followed by 58% by PC2, 65.26% by PC3, 71.18% by PC4, 76.67% by PC5, 81.65% by PC6, 85.22% by PC7.
3. Eigen vectors indicate that which variable needs more weightage to reduce the variability among the variables.
4. Scree plot shows the first 7 principal components explained majority of the variance as we know i.e. 85%.

3.9: 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]

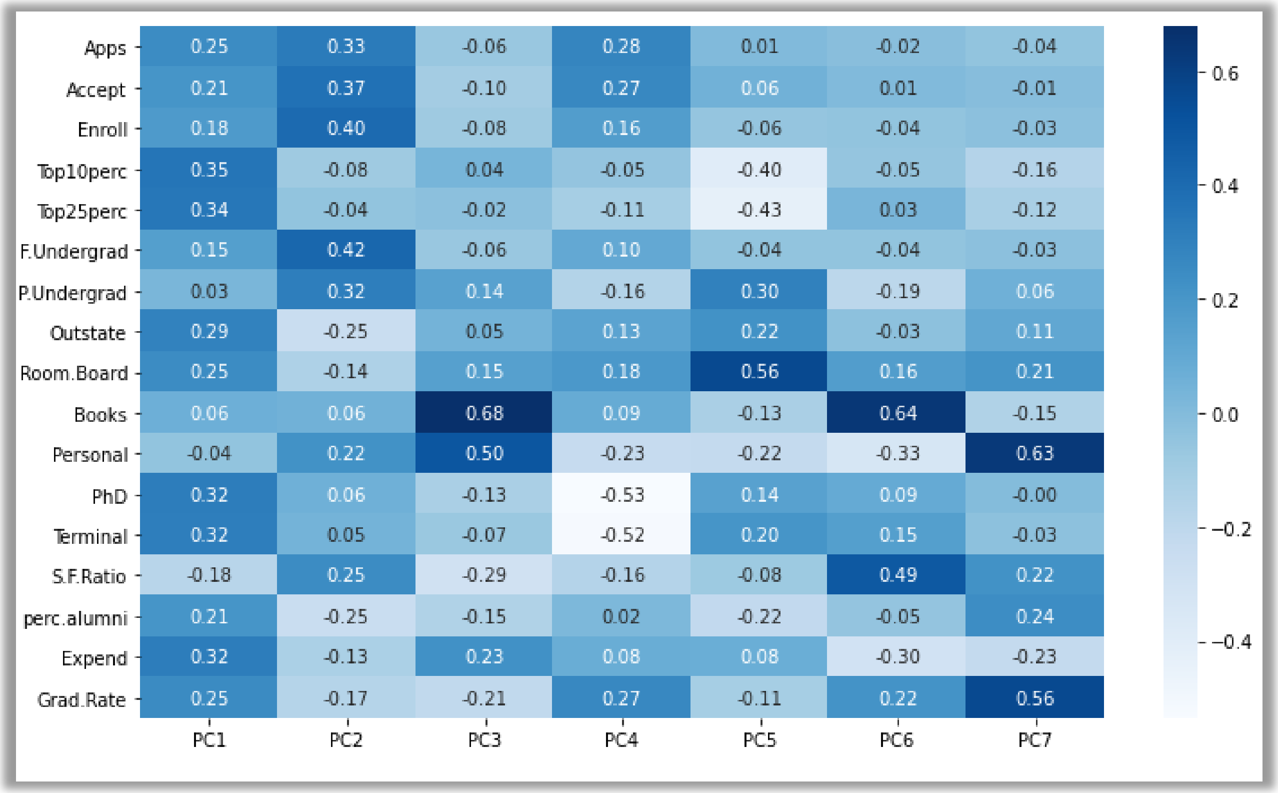


Figure 16: PCA comparisons matrix on original dataset.

Here the most important is PC1, having the largest value of **0.35 in Top 10 perc** and not having such high coefficient in any of the PCs. Likewise we do have other high coefficient values which are not lying in the PC1 are **Books and Personal** at PC3, and PC6 and PC7 respectively. For S.F Ratio and Grad Rate, and PC7 respectively.

The business implications:-

1. The PCA considered as the dimension reduction technique, so in this case study we have the reduced dimensions from 17 to 7 which explains 85% variability in the original dataset.
2. Taking about the original dataset as we have data for various colleges and universities.
3. We can perform the PCA test after removing the outliers, using IQR range to the actual dataset.
4. That helps in the reduction of multicollinearity among the variable which has variance in it.
5. The cumulative Eigen value shows the number of PCs required for the reduction of multicollinearity weather by 85% or by 90 %.

THANK YOU.