Advanced Statistics

Pradeep Kumar Mishra
PGP-DSBA Online
Jun_B_21
Date: 11:Sep:2021

Content View

List of Figures :

List of Tables :

Problem 1A:

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

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

Exploratory Data Analysis

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

Table-1 Dataset Sample

Let us check the types of variables and missing values in the dataset

- From the below results we can see that there is no missing value present in the dataset.
- There are a total of 40 rows and 3 columns in the dataset. Out of 3, 2 columns are of object type and the rest 1 are int64 data types.

```
<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
```

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

Ho : mu1 = mu2 = m3 = ...muk All the means are equals.

H1 : at-least one means is unequal.

Hypothesis with Education:

Null Hypothesis (Ho): The mean salary earned by the People is the same with different categories of educational qualification (High school graduate, Bachelor, and Doctorate.)

Alternate Hypothesis (Ha): The mean salary earned by the People is different in at-least one category of educational qualification.

Hypothesis with Occupation:

Null Hypothesis (Ho): The mean salary earned by the People is the same with different categories of occupation .

Alternate Hypothesis (Ha): The mean salary earned by the People is different in at least one category 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.

Hypothesis with Education:

Null Hypothesis (Ho): The mean salary earned by the People is the same with different categories of educational qualification (High school graduate, Bachelor, and Doctorate.)

Alternate Hypothesis (Ha): The mean salary earned by the People is different in at-least one category of educational qualification.

Assumptions involved in using ANOVA:

- The samples drawn from different populations are independent and random.
- The variances of all the populations are equal.
- The response variables of all the populations are normally distributed. Central Limit Theorem(CLT) asserts that sample mean follows normal distribution, even if the population distribution is not normal, when sample size is at least 30.

By using the below formula :

```
formula = 'Salary ~ C(Education)'
model = ols(formula, data).fit()
aov_table = anova_lm(model)
```

We get the below table :

	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

Table-2 anova

Conclusion : We have evidence to reject the null hypothesis, since p value < Level of significance (0.05) Accept the Alternate hypothesis (Ha): The mean salary earned by the People is different in at-least one category of educational qualification.

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.

Hypothesis with Occupation:

Null Hypothesis (Ho): The mean salary earned by the People is the same with different categories of occupation .

Alternate Hypothesis (Ha): The mean salary earned by the People is different in at least one category of occupation.

Assumptions involved in using ANOVA:

- The samples drawn from different populations are independent and random.
- The variances of all the populations are equal.
- The response variables of all the populations are normally distributed. Central Limit Theorem(CLT) asserts that sample mean follows normal distribution, even if the population distribution is not normal, when sample size is at least 30.

By using the below formula :

```
formula = 'Salary ~ C(Occupation)'
model = ols(formula, data).fit()
aov_table = anova_lm(model)
```

We get the below table :

	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

Table-3 anova

Conclusion : We have no evidence to reject the null hypothesis, since p value > Level of significance(0.05) Accept the Null Hypothesis (Ho) : The mean salary earned by the People is the same with different 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).

Salary with education where I found the null hypothesis is rejected I use the tukeyhsd to check in which class means are significantly different.

By using the below formula :

```
mc= MultiComparison(data.Salary , data.Education)
result=mc.tukeyhsd()
```

We get the below table :

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

Table-4 anova

Conclusion: In the above result I observe that all class means with each other have significant differences. reject the null hypothesis.

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]

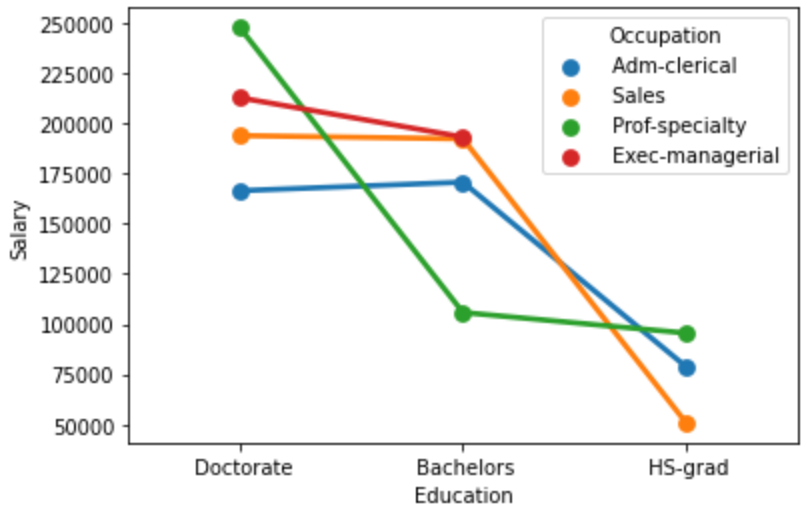

Fig.1 Pointplot

As seen from the above plot, there seems to be interaction among the four categorical variables.

- Doctorate education salary is high if there occupation is Prof-specialty and low if there occupation is Adm-clerical.
- Sales salary is very low in HS-grad education but if we compare in Bachelors education then sales salary is equal to Exe-managerial occupation which is highest in bachelors education.
- Sales salary is very low in HS-grad education but if we compare in Bachelors education then sales salary is equal to Exe-managerial occupation which is highest in bachelors education.

If we also do the statistical analysis we see below table C(Education):C(Occupation) 4.227791e+10 6.0 9.909463 1.323371e-05 p < 0.05 it means there is interaction between variables Education and Occupation.

By using the below formula:

```
formula = 'Salary ~ C(Education) + C(Occupation) + C(Education) * C(Occupation) '

model = ols(formula, data).fit()

aov_table = anova_lm(model,typ=3)
```

	sum_sq	df	F	PR(>F)
Intercept	8.742674e+10	1.0	122.951013	6.022884e-12
C(Education)	1.686040e+10	2.0	11.855657	1.726495e-04
C(Occupation)	2.028454e+10	3.0	9.508931	1.556271e-04
C(Education):C(Occupation)	4.227791e+10	6.0	9.909463	1.323371e-05
Residual	2.062102e+10	29.0	NaN	NaN

Table-5 anova

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?

If we perform along with interaction Education*Occupation We have to use typ=3

Null Hypothesis (Ho): The means 'Salary' variable for Education, occupation and interaction level is equal.

Alternate hypothesis (Ha) : At least one of the means "Salary" variables for Education, occupation and interaction level is unequal.

We use the below formula:

```
formula = 'Salary ~ C(Education) + C(Occupation) + C(Education) * C(Occupation)'
model = ols(formula, data).fit()
aov_table = anova_lm(model,typ=3)
```

	sum_sq	df	F	PR(>F)
Intercept	8.742674e+10	1.0	122.951013	6.022884e-12
C(Education)	1.686040e+10	2.0	11.855657	1.726495e-04
C(Occupation)	2.028454e+10	3.0	9.508931	1.556271e-04
C(Education):C(Occupation)	4.227791e+10	6.0	9.909463	1.323371e-05
Residual	2.062102e+10	29.0	NaN	NaN

Table-6 anova

Conclusion : we have evidence to reject the null Hypothesis, since p<Level of significance(0.05) so we reject the null hypothesis and accept the Alternate hypothesis (Ha) : At least one of the means "Salary" variable for Education, occupation and interaction level is unequal.

If we Perform a two-way ANOVA based on Salary with respect to both Education and Occupation. We use typ=2

Null Hypothesis (Ho): The means of 'Salary' variable with respect to each Education category and occupation is equal.

Alternate hypothesis (Ha) : At least one of the means of the "Salary" variable with respect to each Education category and occupation is unequal.

By using the below formula :

```
formula = 'Salary ~ C(Education) + C(Occupation)'

model = ols(formula, data).fit()

aov_table = anova_lm(model,typ=2)
```

We get the below table:

	sum_sq	df	F	PR(>F)
C(Education)	9.695663e+10	2.0	29.510933	3.708479e-08
C(Occupation)	5.519946e+09	3.0	1.120080	3.545825e-01
Residual	5.585261e+10	34.0	NaN	NaN

Table-7 anova

Conclusion : we have evidence to reject the null Hypothesis, since p<Level of significance(0.05) so we reject the null hypothesis and accept the Alternate hypothesis(Ha):At least one of the means of "Salary" variable with respect to each Education category and occupation is unequal.

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

In this problem We are trying to compare whether the mean salary earned by people with occupation and educational qualification is different or not.

We found that the salary is the same with different categories of occupation(Administrative and clerical, Sales, Professional or specialty, and Executive or managerial).

It does not matter in which occupation you are working, the salary is the same for that.

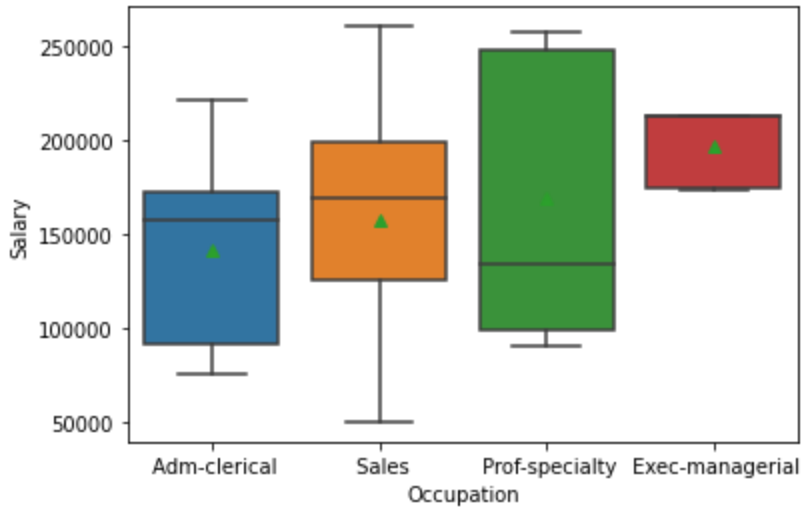

Fig.2 Boxplot

If we compare with educational qualifications I found that the mean salary is different for each educational qualification(High school graduate, Bachelor, and Doctorate).

Your salary depends on your education qualification.

If your education is Doctorate then your salary is higher than a High school graduate and Bachelor.

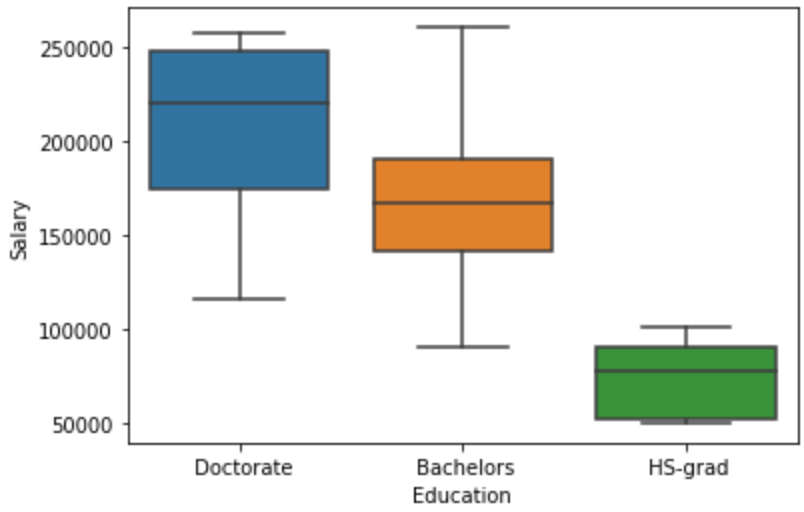

Fig.3 Boxplot

If we compare the salary earned by people with Doctorate education is impacting with different occupations.

We observed that with each education impacting each occupation.

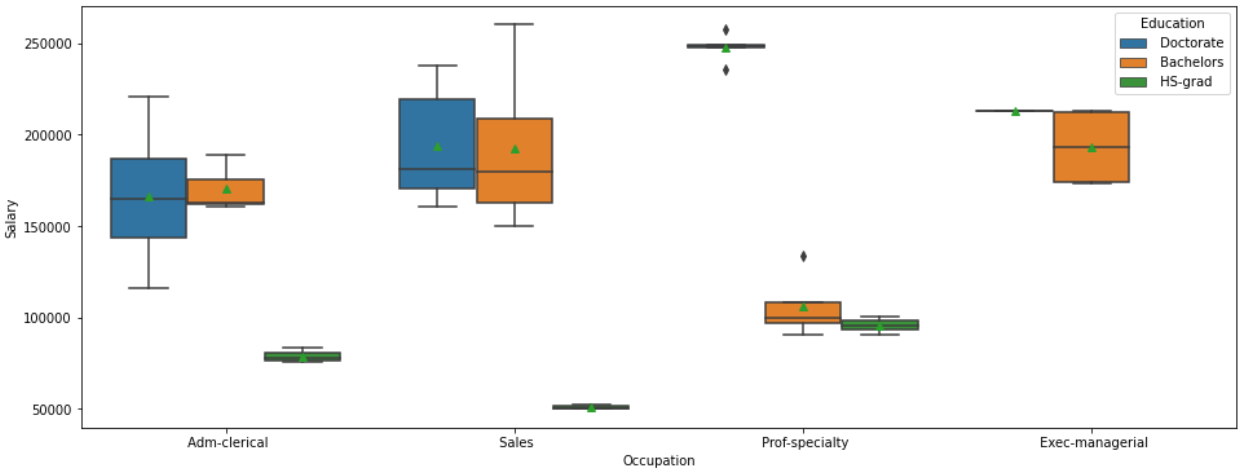

Fig.4 Boxplot

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.

Exploratory Data Analysis:

Describe the dataset:

	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-8 Describe the Dataset

Let us check the types of variables and missing values in the dataset

- From the below results we can see that there is no missing value present in the dataset.
- There are a total of 777 rows and 18 columns in the dataset. Out of 18, 1 column is of object type, 1 column is of float64 and the rest 16 are int64 data types.

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

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

Univariate Analysis :

Describe the data:

	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-9 Describe the data

Insight:

- Data consists of 777 colleges with 17 features.
- All fields have equal count. It means there is no missing value.
- Fields Apps, Accent, Enroll, F.Undergrad and P.Undergrad mean and median difference is high so there are outliers.
- No. of part time undergraduate students is less than full time undergraduate.

Plot the histplot :

Insight:

- fields Apps, Accept, Enroll, F.Undergrad, P.Undergrad, Personal, perc.alumni and Expend have right skewness.
- fields Phd, Terminal and Grad.Rate have left skewness.
- fields Top25perc is seems normal distribution.

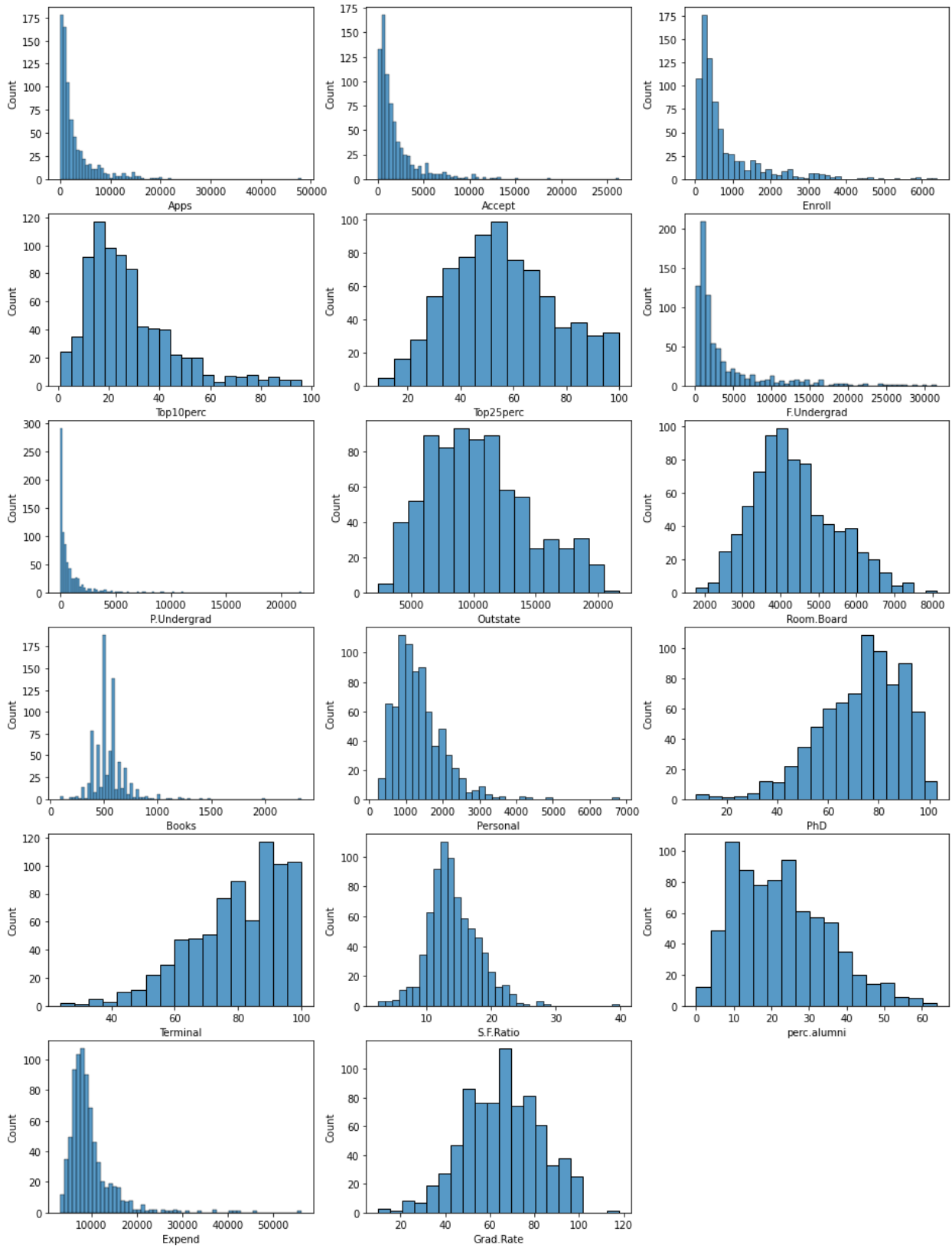

Fig.5 Histplot

Plot the Boxplot:

Insight:

- In the boxplot most of the variables seem outliers except Top25perc.
- in some of the fields seems less outiers outstates, Room.Board, perc.alumni and Grad.Rate

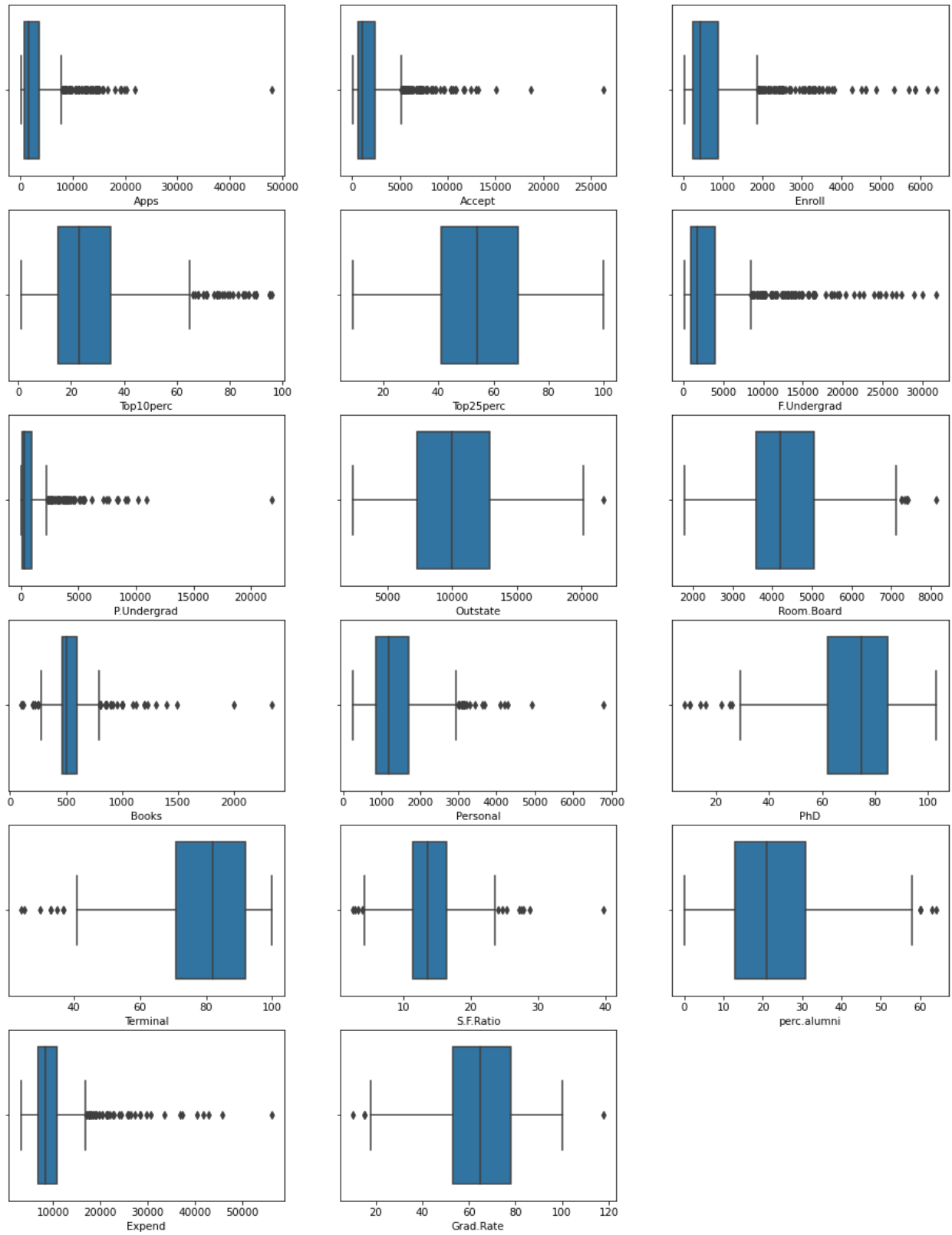

Fig.6 Boxplot

Bivariate Analysis:

Insight:

- There are a considerable number of features that are highly correlated.
- Accept variables show highly correlated with Apps and Enroll. F.undergraduate is highly correlated with Enroll.

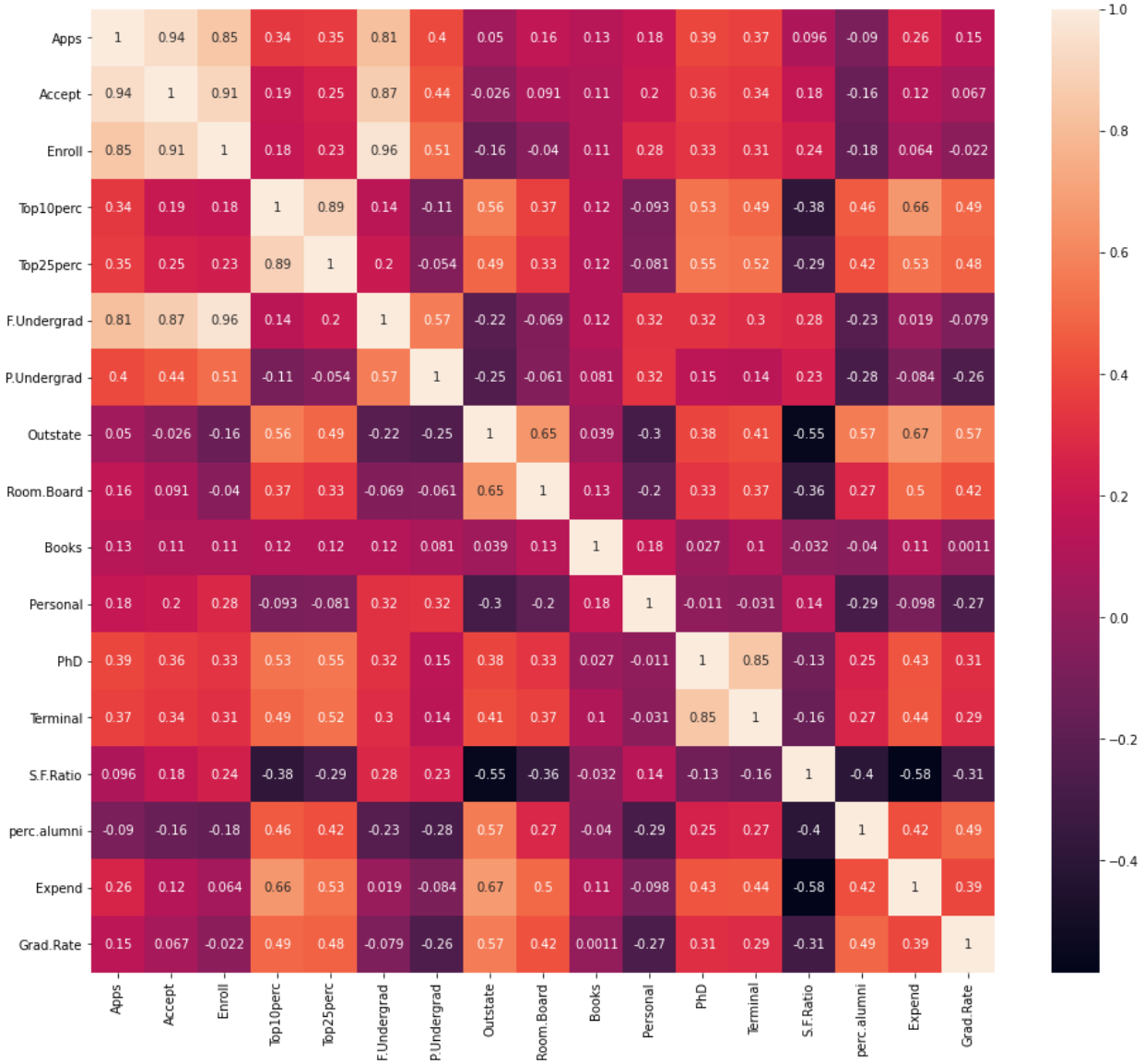

Fig.7 heatmap

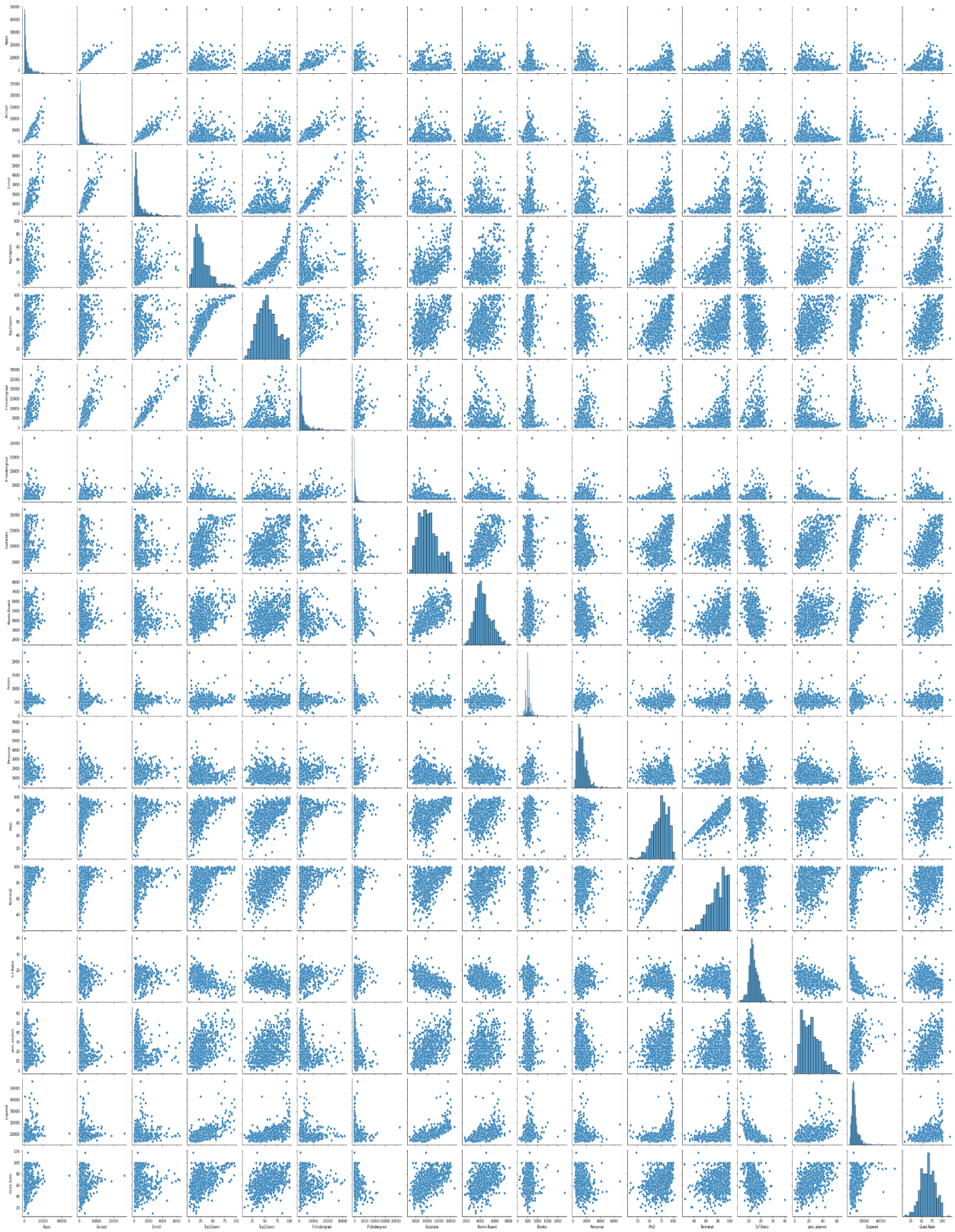

Fig.8 pairplot

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

	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-10 Describe the data

- Scaling is necessary for PCA in this case because I can see in the above described data and say that data is with different weights. It is recommended to transform the features so that all features are in the same scale.
- Using Z-score for scaling and can see the below table that seems the same scale.

	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 -11 Describe data after scaling

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

comparison between the covariance and the correlation :

- Both covariance and correlation measure the relationship and the dependency between two variables.
- Covariance indicates the direction of the linear relationship between variables while Correlation measures both the strength and direction of the linear relationship between two variables.
- Correlation values are standardized. Covariance values are not standardized.
- You can obtain the correlation coefficient of two variables by dividing the covariance of these variables by the product of the standard deviations of the same values.

We can state that approaches yield the same eigenvectors and eigenvalue pairs:

- Eigen decomposition of the covariance matrix after standardizing the data.
- Eigen decomposition of the correlation matrix.
- Eigen decomposition of the correlation matrix after standardizing the data.
- Finally we can say that after scaling - the covariance and the correlation have the same values.

Corr. matrix.

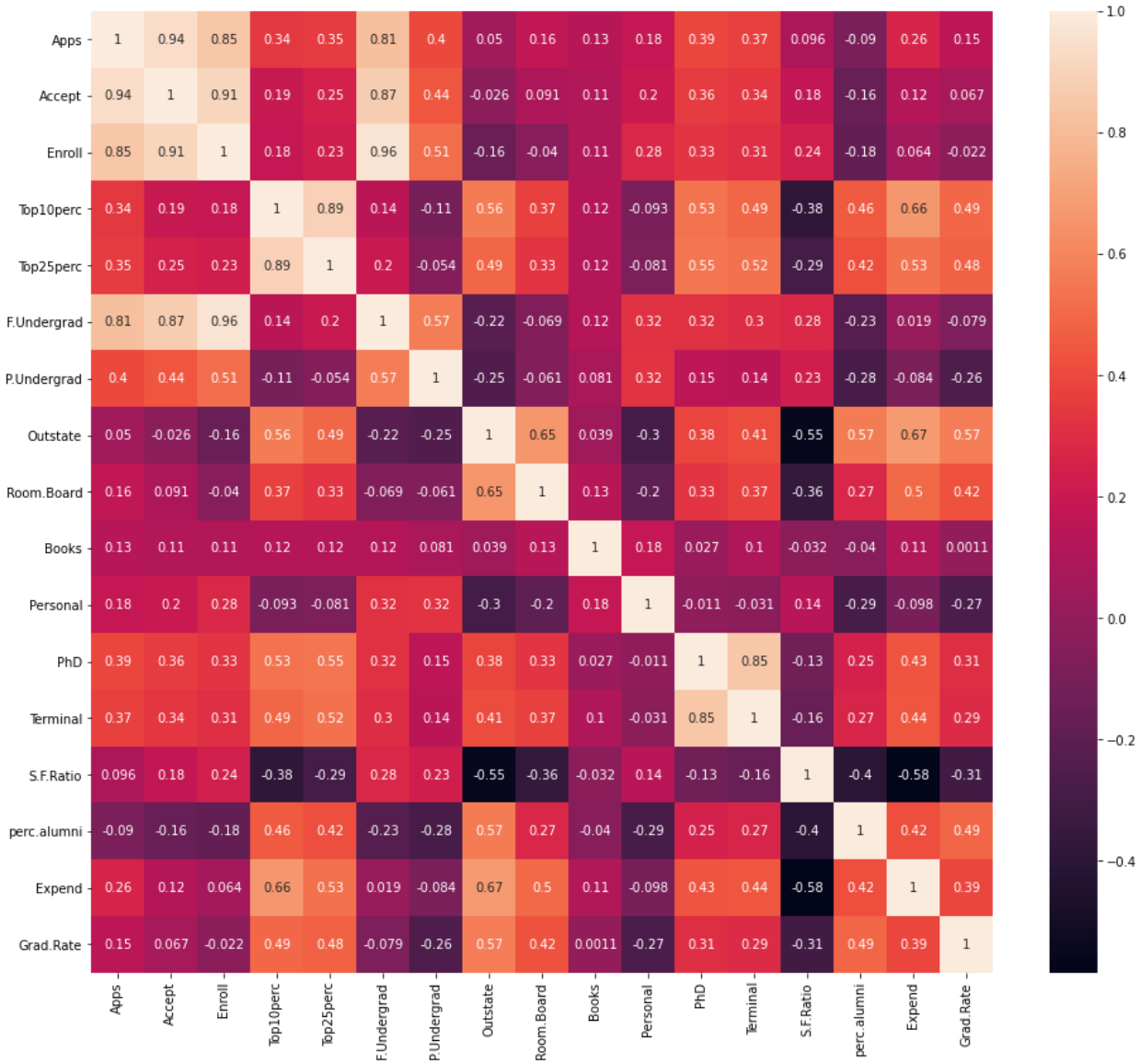

Fig.9 heatmap

Cov. 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
```

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

Boxplot before scaling.

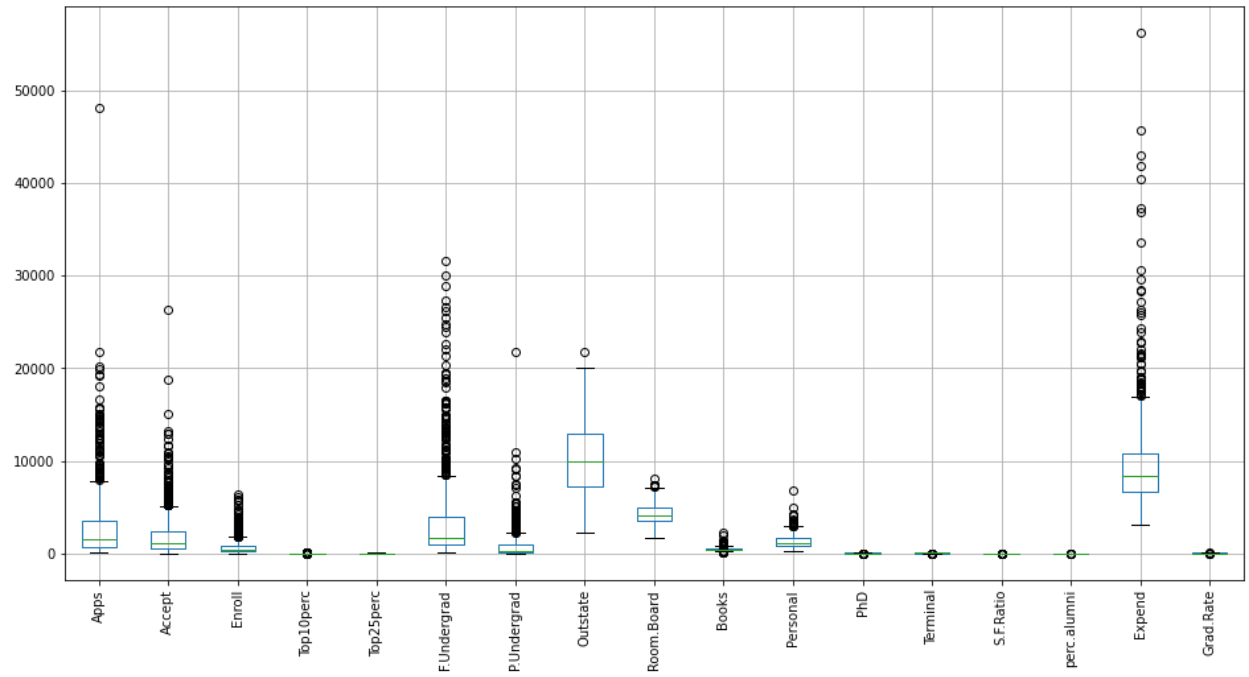

Fig10. Boxplot

After the scaling.

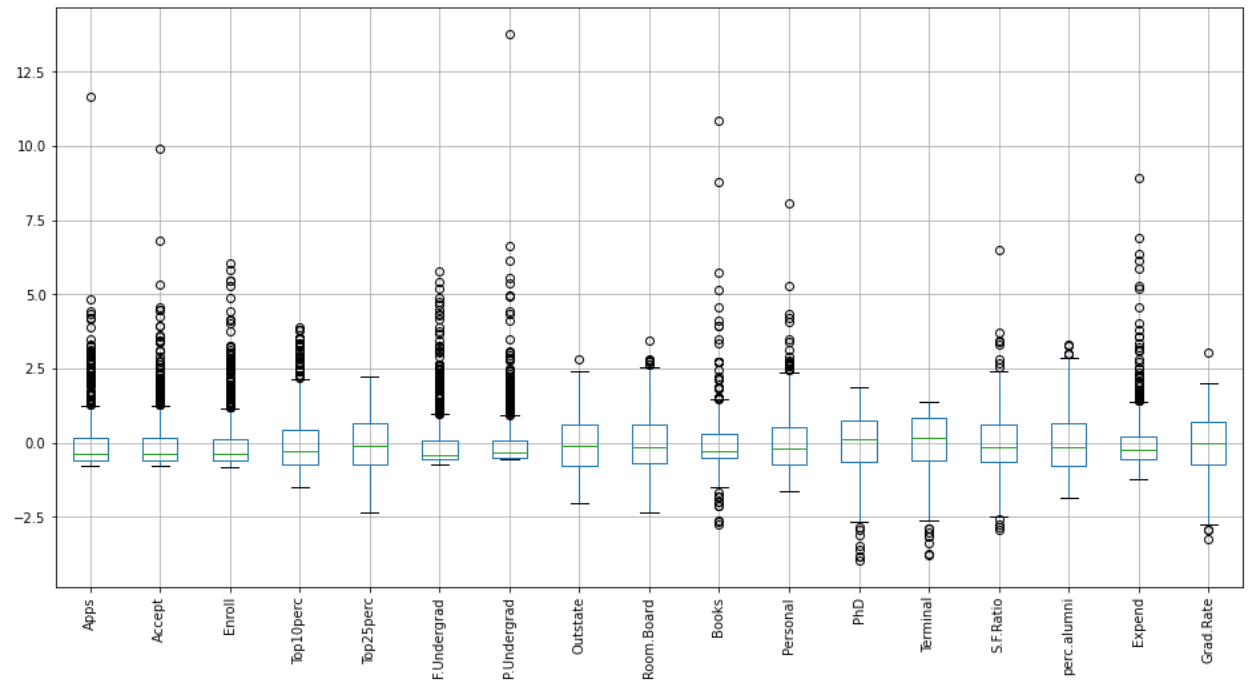

Fig. 11 Boxplot

Insight

- If we look at the boxplot after the scaling, all the boxplot is comparable.

- Before, the scaling boxplot look was not comparable because some variable boxplot looked very big and some were very small so there is no sense to compare with each other in boxplot.
- We can see the Outstate variable this is the big boxplot but if we see the Top10perc, Top25perc, PhD. etc. these are the very small boxplot.

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

Statistical tests to be done before PCA

1.Bartlett's Test of Sphericity

Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

H0: All variables in the data are uncorrelated.

Ha: At least one pair of variables in the data are correlated.

If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is small, then we can reject the null hypothesis and agree that there is at least one pair of variables in the data which are correlated hence PCA is recommended.

By using the below formula:

chi_square_value,p_value=calculate_bartlett_sphericity(data_scaled)

p_value=0.0

Conclusion : We have evidence to reject the null hypothesis, since p value < Level of significance (0.05) Accept the Alternative hypothesis Ha: At least one pair of variables in the data are correlated.

2.KMO Test

The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction in the dimension and extraction of meaningful components.

kmo_model= 0.8131251200373522

MSA > 0.7 is expected to provide a considerable reduction in the dimension and extraction of meaningful components.

As per the selection criteria of eigenvalues :

- Consider all eigens which have values of >=1.
- Consider all eigens where the cumm variance is atleast 80%.

We are going to use only 6 components as per given selection criteria.

eigenvector/ coefficient of pc1 ..... a1,a2,a3.....

pca.components_

```
array([[ 0.2487656 ,  0.2076015 ,  0.17630359,  0.35427395,  0.34400128,
         0.15464096,  0.0264425 ,  0.29473642,  0.24903045,  0.06475752,
        -0.04252854,  0.31831287,  0.31705602, -0.17695789,  0.20508237,
         0.31890875,  0.25231565],
       [ 0.33159823,  0.37211675,  0.40372425, -0.08241182, -0.04477866,
         0.41767377,  0.31508783, -0.24964352, -0.13780888,  0.05634184,
         0.21992922,  0.05831132,  0.04642945,  0.24666528, -0.24659527,
        -0.13168986, -0.16924053],
       [-0.06309209, -0.10124907, -0.08298558,  0.03505553, -0.02414794,
        -0.06139296,  0.13968171,  0.04659888,  0.14896739,  0.67741165,
         0.49972112, -0.12702837, -0.06603755, -0.2898484 , -0.14698927,
         0.22674398, -0.20806465],
       [ 0.28131052,  0.26781736,  0.16182679, -0.05154725, -0.10976654,
         0.10041231, -0.15855849,  0.13129136,  0.18499599,  0.08708922,
        -0.23071057, -0.53472483, -0.51944302, -0.16118949,  0.01731422,
         0.0792735 ,  0.26912907],
       [ 0.00574142,  0.05578609, -0.05569364, -0.39543435, -0.42653359,
        -0.04345436,  0.30238541,  0.222532  ,  0.56091947, -0.12728883,
        -0.22231102,  0.14016633,  0.20471973, -0.07938825, -0.21629741,
         0.07595812, -0.10926791],
       [-0.01623744,  0.00753468, -0.04255798, -0.0526928 ,  0.03309159,
        -0.04345423, -0.19119858, -0.03000039,  0.16275545,  0.64105495,
        -0.331398  ,  0.09125552,  0.15492765,  0.48704587, -0.04734001,
        -0.29811862,  0.21616331]])
```

eigenvalues This is always return is descending order

pca.explained_variance_

[5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123, 0.84849117]

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

Statistical tests to be done before PCA

1.Bartlett's Test of Sphericity

Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

H0: All variables in the data are uncorrelated.

Ha: At least one pair of variables in the data are correlated.

If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is small, then we can reject the null hypothesis and agree that there is at least one pair of variables in the data which are correlated hence PCA is recommended.

By using the below formula:

chi_square_value,p_value=calculate_bartlett_sphericity(data_scaled)

p_value=0.0

Conclusion : We have evidence to reject the null hypothesis, since p value < Level of significance (0.05) Accept the Alternative hypothesis Ha: At least one pair of variables in the data are correlated.

2.KMO Test

The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction in the dimension and extraction of meaningful components.

kmo_model= 0.8131251200373522

MSA > 0.7 is expected to provide a considerable reduction in the dimension and extraction of meaningful components.

As per the selection criteria of eigenvalues :

- Consider all eigens which have values of >=1.
- Consider all eigens where the cumm variance is atleast 80%.

We are going to use only 6 components as per given selection criteria.

	0	1	2	3	4	5
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

Tabel 12. Dataset

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].

linear equation of PC in terms of eigenvectors and corresponding features

PC= a1x1+a2x2.....anxn.

We use the above formula and find the below output :

PC1 = 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

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?

Remark: the no. of Eigenvalues = no. of variable = default no. of principle components.

for the dimension reduction we going to select the couple of them selection criteria of eigenvalues :

- Consider all eigens which have values of >=1.
- Consider all eigens where the cumm variance is atleast 80%.

Pca.explained_variance_

[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]

For the above eigen values we can consider the first 4 eigen values because this satisfies the first criteria to select the principal component. but before moving to conclusion we must check the second criteria.

Total variance explained by the first eigenvalues for checking the second criteria we follow the below formula.

percentage contribution by the first eigenvalues= (first eigenvalue/total sum of eigen value)*100 as we are looking for a cumulative sum we add and found that adding the first four eigenvalues we reach only 71.1% [32. , 58.3, 65.2, 71.1]

but we need at-least 80% so we add 2 more eigen values to reach that.

[32. , 58.3, 65.2, 71.1, 76.6, 81.6]

Remarks: if adding eigens which have value 1 or greater than 1 but not reaching cumm variance at-least 80% then we go with eigens value which is less than 1 to reach cumm variance 80% .

eigenvectors indicate:

eigen vectors indicate the coefficients or Factor loadings eg. PC= a1x1 + a2x2 + a3x3 + a4x4 + ... +anxn.

x= variables

a= coefficients/ factor loadings

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].

- PCA helps in Dimensionality reduction. Converts a set of correlated variables to non-correlated variables.
- It finds a sequence of linear combinations of variables.
- PCA= Linear Combination of Variable.

  PCA= a1x1+a2x2+a3x3.....anxn

  a= Coefficient / Factor Loading /EigenVectors

  x=Variables/features

- It is often used to help in dealing with multi- collinearity before a model is developed.
- Assumptions of PCA :
  1. Independent variables are highly correlated to each other.
  2. Variables included are metric level or nominal level.

3. Independent variables are numeric in nature.
4. Bartlett-Test: The Bartlett test is statistically significant as.

   H0: Variables are uncorrelated.

   H1: Variables are correlated.

In this case I checked the correlation between each variable and see below figure that each variable is highly correlated with each other.

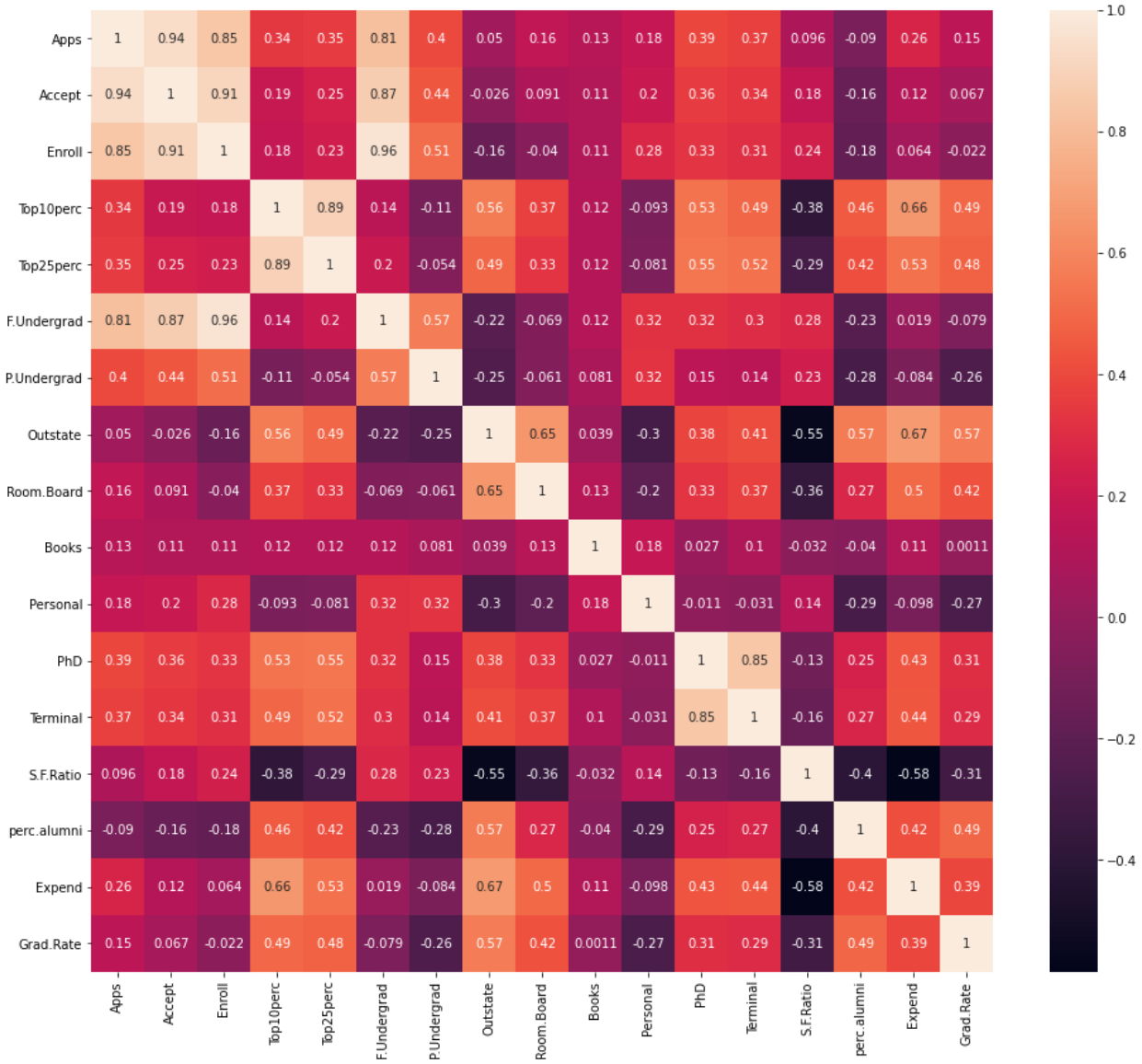

Fig.12 heatmap

I performed the PCA and now you can see the figure below.

There are only 6 variables and they are not correlated to each other.

By using the PCA we found two things.

1. 1. removed the multicollinearity
2. 2. reduced the dimension.

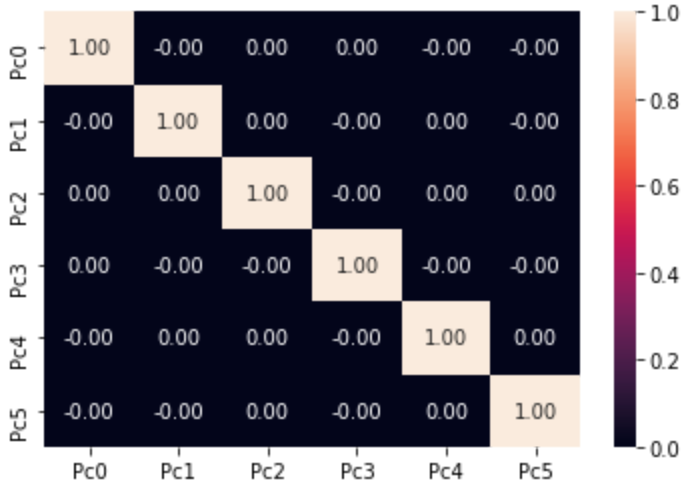

Fig.13 heatmap

PCs help in the further analysis:

- PCA improves the performance of the model as it eliminates correlated variables that don't contribute to any decision making.
- PCA helps in overcoming data overfitting issues by decreasing the number of features.
- PCA results in high variance and thus improves visualization.