Advance Statistics PROJECT

Name: Swetha Kunapuli

Batch & Course: PGP-DSBA

Online June Batch

Date: 11/09/2021

Table of Contents

Problem 1

Problem 2

List of Figures

PROBLEM 1

Executive Summary:

The wholesale customer data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories.

Introduction:

The purpose of this exercise is recommended for learning and practicing skills in exploratory data analysis by performing hypothesis testing using ANOVA. The expectation is to execute the hypothesis test and interpret the result based on the given hypothesis.

The dataset has total 40 rows and 3 columns.

Data Description:

One way ANOVA(Education)

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

One way ANOVA(Occupation)

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

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

Exploratory Data Analysis:

Check for types of variables in the data frame:

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

Check for missing values in the dataset:

```
Education     0
Occupation    0
Salary        0
dtype: int64
```

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

One way ANOVA(Education)

Null Hypothesis $H0$: The mean salary is the same across all the 3 categories of education (Doctorate, Bachelors, HS-Grad).

Alternate Hypothesis $H1$: The mean salary is different in at least one category of education.

One way ANOVA(Occupation)

Null Hypothesis $H0$: The mean salary is the same across all the 4 categories of occupation (Prof-Specialty, Sales, Adm-clerical, Exec-Managerial).

Alternate Hypothesis $H1$: The mean salary is different in at least one category of occupation.

2A. 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.

	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

The above is the ANOVA table for Education variable.

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

3A. 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.

	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

The above is the ANOVA table for Occupation variable.

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

4A. If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result.

To find out which class means are significantly different, the Tukey Honest Significant Difference test is performed.

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

```
         Multiple Comparison of Means - Tukey HSD, FWER=0.05
=========================================================================
  group1     group2    meandiff  p-adj     lower         upper     reject
-------------------------------------------------------------------------
 Bachelors Doctorate   43274.0667 0.0146     7541.1439  79006.9894   True
 Bachelors   HS-grad  -90114.1556  0.001 -132035.1958 -48193.1153   True
 Doctorate   HS-grad -133388.2222  0.001 -174815.0876 -91961.3569   True
-------------------------------------------------------------------------
```

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

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

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

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

We analyse the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.

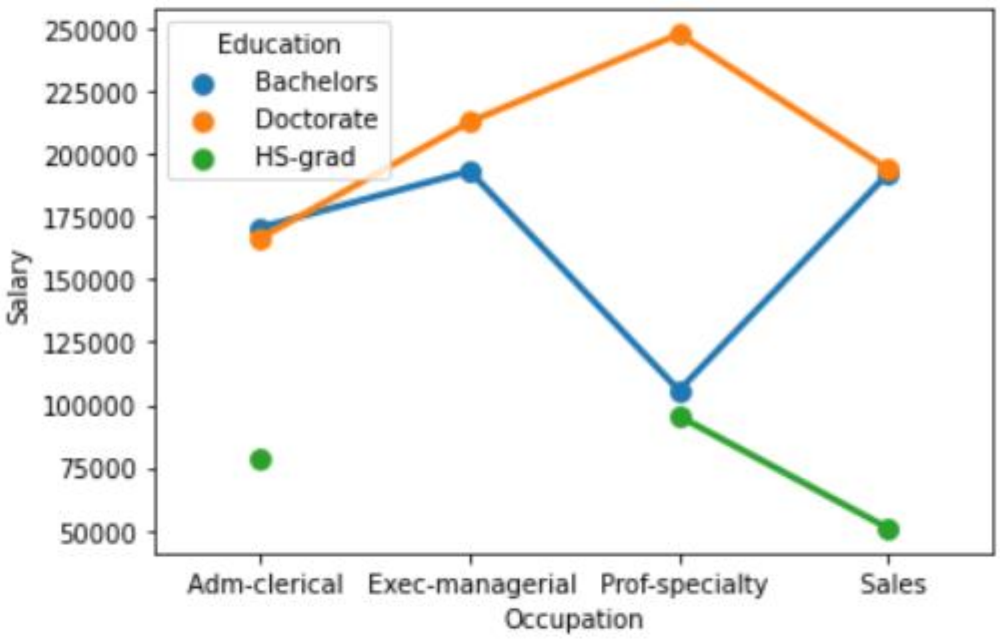

Fig.1

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

The following are some of the observations from the interaction plot:

- People with HS-grad education do not reach the position of Exec-managerial and they hold only Adm-clerk, Sales and Prof-Specialty occupations.

- People with education as Bachelors or Doctorate and occupation as Adm-clerical and Sales almost earn the same salaries(salaries ranging from 170000–190000).

- People with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupations as Adm-clerical and Sales.

- People with education as Bachelors and occupation Sales earn higher than people with education as Bachelors and occupation Prof-Specialty whereas people with education as Doctorate and occupation Sales earn lesser than people with Doctorate and occupation Prof-Specialty. We see a reversal in this part of the plot.

- Similarly, people with education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupation Exec-Managerial whereas people with education as Doctorate and occupation as Prof-Specialty earn higher than people with education as Doctorate and occupation Exec-Managerial. There is a reversal in this part of the plot too.

- Salespeople with Bachelors or Doctorate education earn the same salaries and earn higher than people with education as HS-grad.

- Adm clerical people with education as HS-grad earn the lowest salaries when compared to people with education as Bachelors or Doctorate.

- Prof-Specialty people with education as Doctorate earn maximum salaries and people with education as HS-Grad earn the minimum.

- People with education as HS -Grad earn the minimum salaries.

- There are no people with education as HS -grad who hold Exec-managerial occupation.

- People with education as Bachelors and occupation, Sales and Exec-Managerial earn the same salaries.

2B. 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?

Two-Way Anova:

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

H1: There is an interaction effect between the independent variable 'education' and the independent variable 'occupation' on the mean salary. By performing two-way ANOVA, we get the following table:

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

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

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

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

From the ANOVA method and the interaction plot, we see that education combined with occupation results in higher and better salaries among the people. It is clearly seen that people with education as Doctorate draw the maximum salaries and people with education HS-grad earn the least.

Thus, we can conclude that Salary is dependent on educational qualifications and occupation.

PROBLEM 2

Executive Summary:

The dataset Education - Post 12th Standard.csv contains information on various colleges.

Introduction:

The purpose of this whole exercise is to perform Principal Component Analysis for this case study according to the instructions given.

The survey dataset has 777 rows and 18 columns about student's details.

Data Description:

The dataset Education - Post 12th Standard.csv contains one categorical column and rest are numerical columns out of 18 columns.

We must perform Exploratory Data Analysis, if necessary, before performing Principal Component Analysis.

Sample of the dataset:

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

Exploratory Data Analysis

Check for types of variables in the data frame:

```
RangeIndex: 777 entries, 0 to 776
Data columns (total 17 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Apps         777 non-null    float64
 1   Accept       777 non-null    float64
 2   Enroll       777 non-null    float64
 3   Top10perc    777 non-null    float64
 4   Top25perc    777 non-null    int64
 5   F.Undergrad  777 non-null    float64
 6   P.Undergrad  777 non-null    float64
 7   Outstate     777 non-null    float64
 8   Room.Board   777 non-null    float64
 9   Books        777 non-null    float64
 10  Personal     777 non-null    float64
 11  PhD          777 non-null    float64
 12  Terminal     777 non-null    float64
 13  S.F.Ratio    777 non-null    float64
 14  perc.alumni  777 non-null    float64
 15  Expend       777 non-null    float64
 16  Grad.Rate    777 non-null    float64
dtypes: float64(16), int64(1)
memory usage: 103.3 KB
```

Check for missing values in the dataset:

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

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

Exploratory Data Analysis is to perform basic checks as in the above figures, to check for missing values, check for duplicates, check for data types for variables and checking mean, median, standard deviation, min, max values.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000
mean	2571.352638	1746.280566	660.388674	26.842986	55.796654	2935.648005	655.884170	10440.196268	4355.438224
std	2422.195279	1523.286632	570.126836	15.582539	19.804778	2700.233049	716.274014	4021.712447	1090.666009
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	2340.000000	1780.000000
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	7320.000000	3597.000000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	9990.000000	4200.000000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	12925.000000	5050.000000
max	7896.000000	5154.000000	1892.000000	65.000000	100.000000	8524.500000	2275.000000	21332.500000	7229.500000

Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000
539.425997	1323.790219	72.774775	79.782497	14.051223	22.722008	9182.523810	65.468468
115.229712	609.505876	15.953120	14.473057	3.784212	12.325480	3396.496148	17.142538
275.000000	250.000000	27.500000	39.500000	4.000000	0.000000	3186.000000	15.500000
470.000000	850.000000	62.000000	71.000000	11.500000	13.000000	6751.000000	53.000000
500.000000	1200.000000	75.000000	82.000000	13.600000	21.000000	8377.000000	65.000000
600.000000	1700.000000	85.000000	92.000000	16.500000	31.000000	10830.000000	78.000000
795.000000	2975.000000	103.000000	100.000000	24.000000	58.000000	16948.500000	115.500000

Check for Duplicates

```
Number of duplicate rows = 0
```

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

Bivariate Analysis

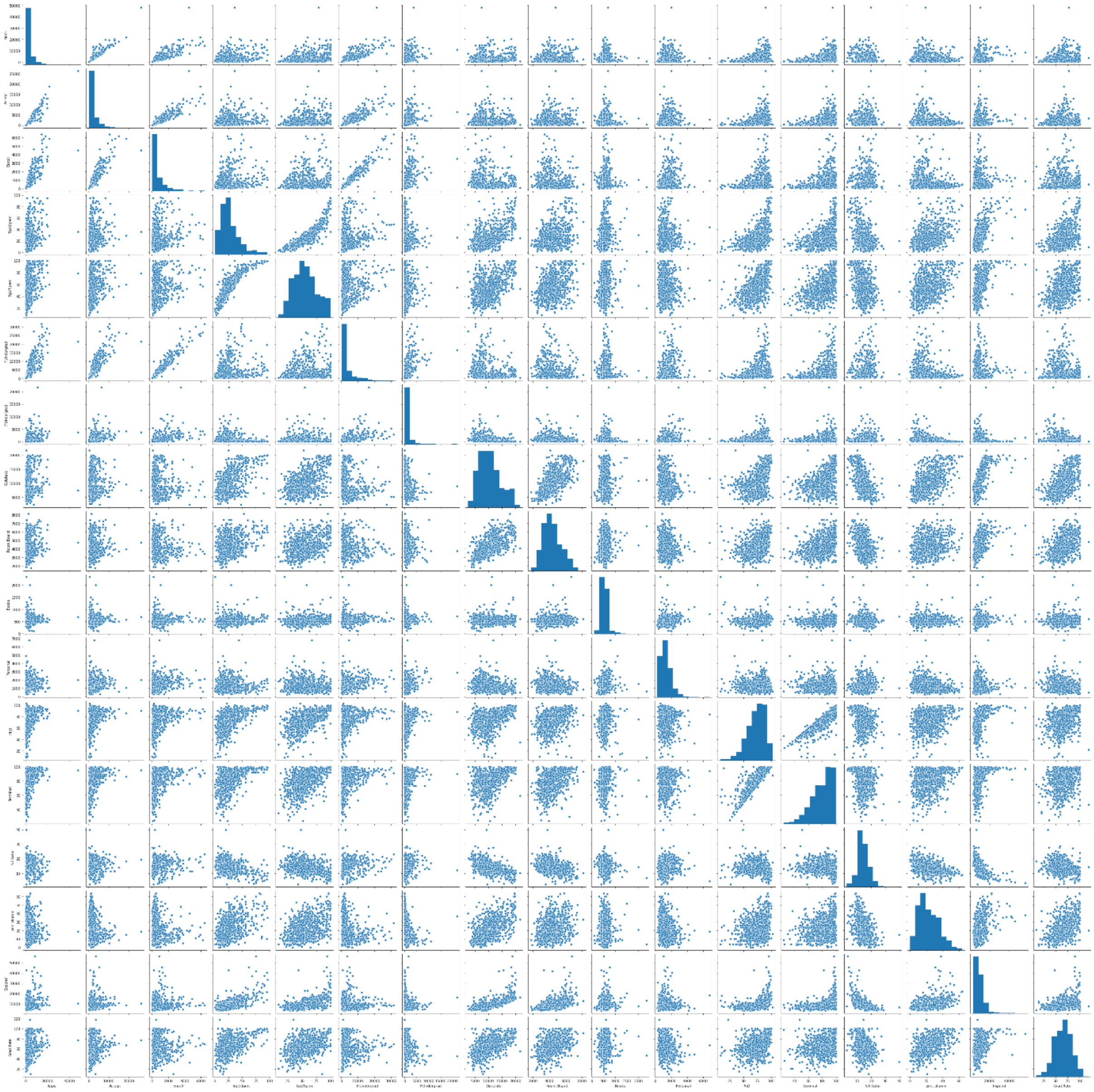

Fig. 2

Distribution of variables shows most of the values are concentrated on lower side.

Relationship between variables shows come correlation.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428

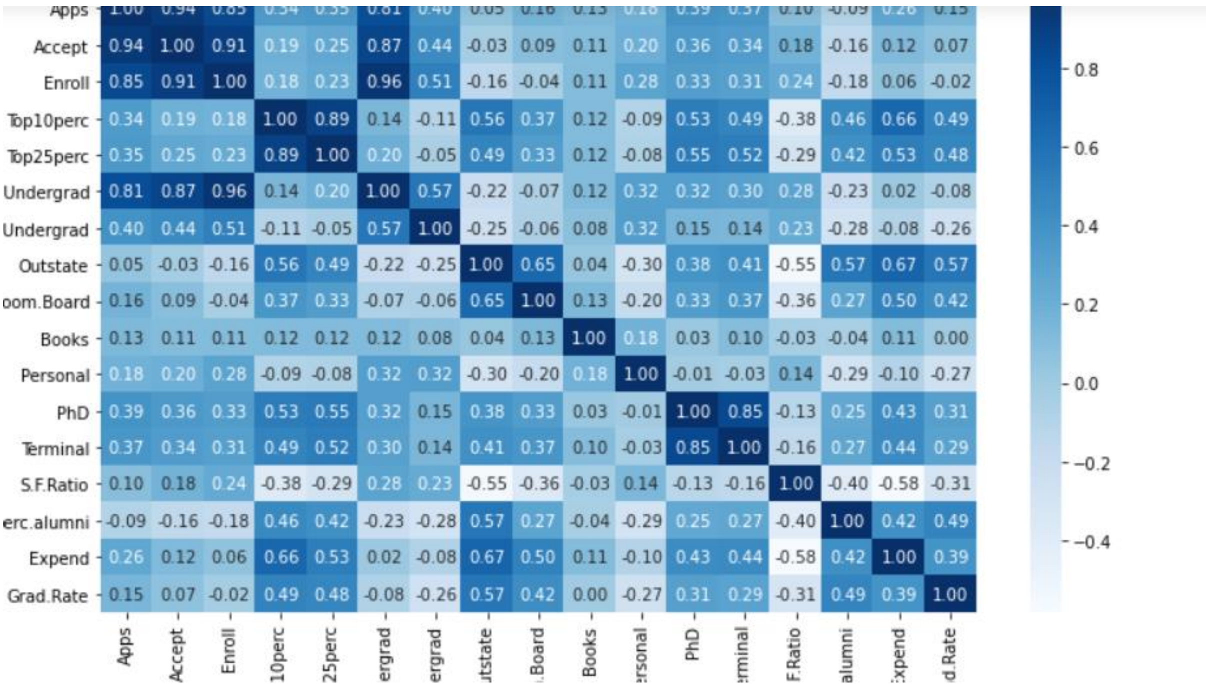

Fig. 3

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

Categorical column to be dropped and use z score for scaling the data to standardize the process.

Data after dropping categorical column.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	T
0	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	
1	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	
2	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	
3	417	349	137	60	89	510	63	12960	5450	450	875	92	
4	193	146	55	16	44	249	869	7560	4120	800	1500	76	

Below is the table after applying z score to the data after dropping categorical field.

```
To standardize the process we will use zscaler
```

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912

Representing boxplot

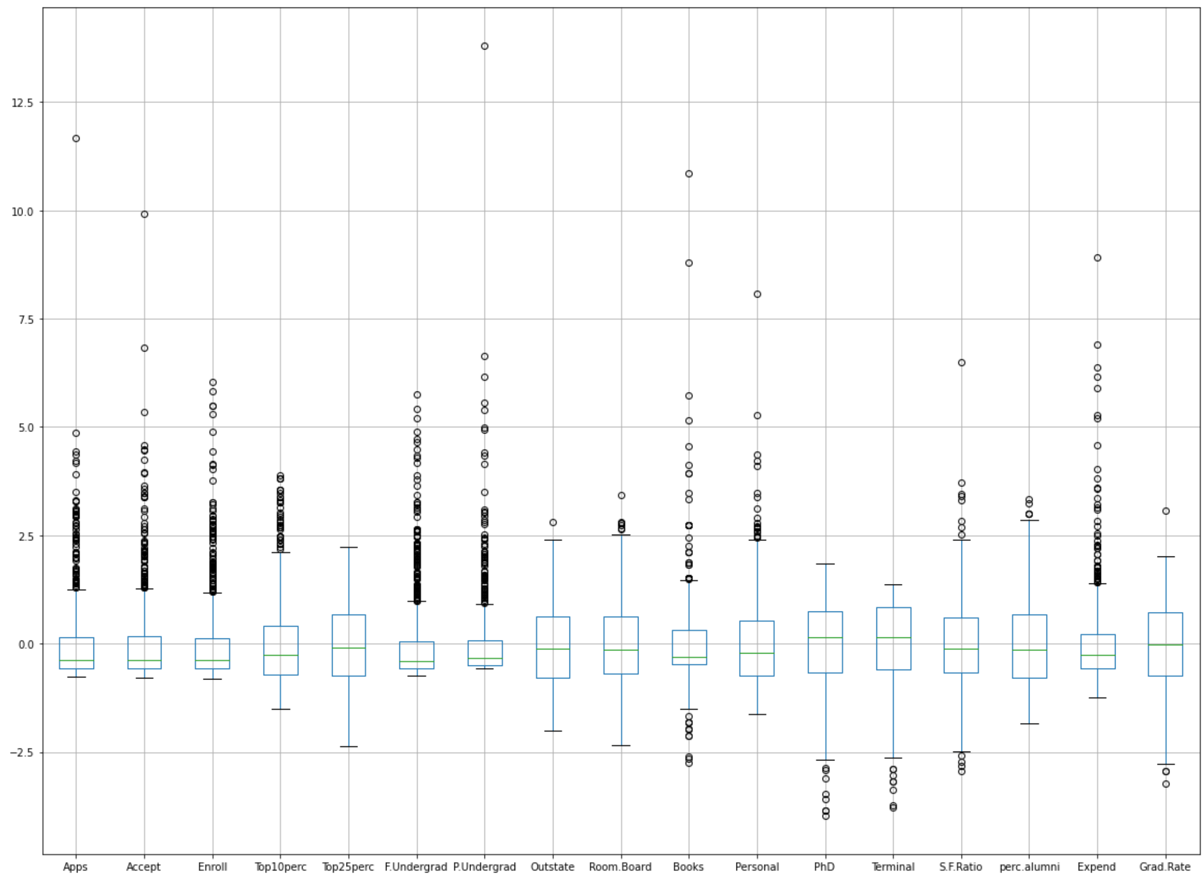

Fig. 4

Reference table after applying z score to the data with numeric columns.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate
count	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02
mean	6.355797e-17	6.774575e-17	-5.249269e-17	-2.753232e-17	-1.546739e-16	-1.661405e-16	-3.029180e-17	6.515595e-17
std	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00
min	-7.551337e-01	-7.947645e-01	-8.022728e-01	-1.506526e+00	-2.364419e+00	-7.346169e-01	-5.615022e-01	-2.014878e+00
25%	-5.754408e-01	-5.775805e-01	-5.793514e-01	-7.123803e-01	-7.476067e-01	-5.586426e-01	-4.997191e-01	-7.762035e-01
50%	-3.732540e-01	-3.710108e-01	-3.725836e-01	-2.585828e-01	-9.077663e-02	-4.111378e-01	-3.301442e-01	-1.120949e-01
75%	1.609122e-01	1.654173e-01	1.314128e-01	4.221134e-01	6.671042e-01	6.294077e-02	7.341765e-02	6.179271e-01
max	1.165867e+01	9.924816e+00	6.043678e+00	3.882319e+00	2.233391e+00	5.764674e+00	1.378992e+01	2.800531e+00

Inference we used ZSCALER to standardize the data into single scale. Now all variables are in in between the scale of -2.5 to 12.5

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

Correlation Table

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Bool
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.1325
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.1135
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.1127
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.1188
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.1155
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.1155
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.0812
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.0388
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.1279
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.0000
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.1792
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.0269
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.0999
S.F.Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628	-0.0319
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.0402

Covariance Table

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outs
Apps	1.497846e+07	8.949860e+06	3.045256e+06	23132.773138	26952.663479	1.528970e+07	2.346620e+06	7.809704e
Accept	8.949860e+06	6.007960e+06	2.076268e+06	8321.124872	12013.404757	1.039358e+07	1.646670e+06	-2.539623e
Enroll	3.045256e+06	2.076268e+06	8.633684e+05	2971.583415	4172.592435	4.347530e+06	7.257907e+05	-5.811885e
Top10perc	2.313277e+04	8.321125e+03	2.971583e+03	311.182456	311.630480	1.208911e+04	-2.829475e+03	3.990718e
Top25perc	2.695266e+04	1.201340e+04	4.172592e+03	311.630480	392.229216	1.915895e+04	-1.615412e+03	3.899243e
F.Undergrad	1.528970e+07	1.039358e+07	4.347530e+06	12089.113681	19158.952782	2.352658e+07	4.212910e+06	-4.209843e
P.Undergrad	2.346620e+06	1.646670e+06	7.257907e+05	-2829.474981	-1615.412144	4.212910e+06	2.317799e+06	-1.552704e
Outstate	7.809704e+05	-2.539623e+05	-5.811885e+05	39907.179832	38992.427500	-4.209843e+06	-1.552704e+06	1.618466e
Room.Board	7.000729e+05	2.443471e+05	-4.099706e+04	7186.705605	7199.903568	-3.664582e+05	-1.023919e+05	2.886597e
Books	8.470375e+04	4.594281e+04	1.729120e+04	346.177405	377.759266	9.253576e+04	2.041045e+04	2.580824e
Personal	4.683468e+05	3.335566e+05	1.767380e+05	-1114.551186	-1083.605065	1.041709e+06	3.297324e+05	-8.146737e
PhD	2.468943e+04	1.423820e+04	5.028961e+03	153.184870	176.518449	2.521178e+04	3.706756e+03	2.515752e
Terminal	2.105307e+04	1.218209e+04	4.217086e+03	127.551581	153.002612	2.142424e+04	3.180597e+03	2.416415e
S.F.Ratio	1.465061e+03	1.709838e+03	8.726848e+02	-26.874525	-23.097199	5.370209e+03	1.401303e+03	-8.835254e

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 is a function of the covariance. What sets them apart is the fact that correlation values are standardized whereas, covariance values are not.

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

Box plot for data before scaling with outliers:

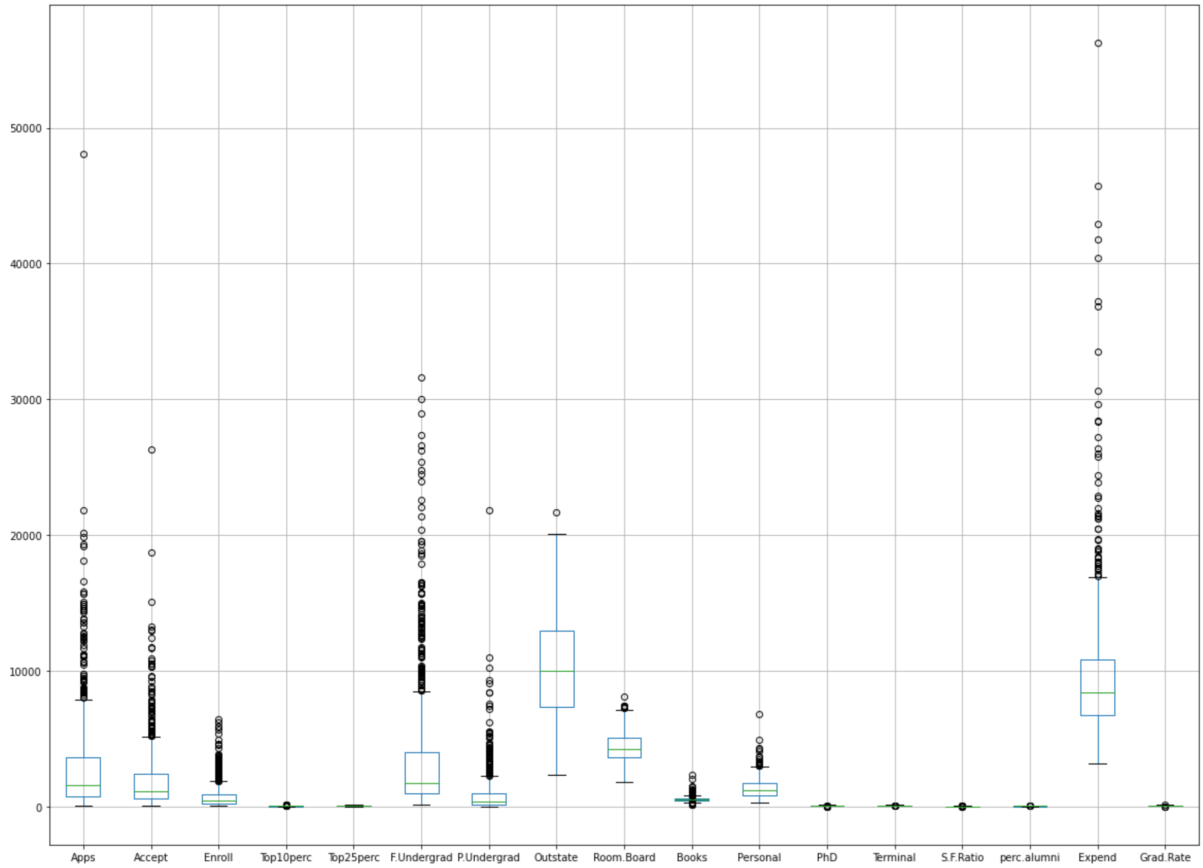

Fig. 5

Box plot for data after scaling with outliers:

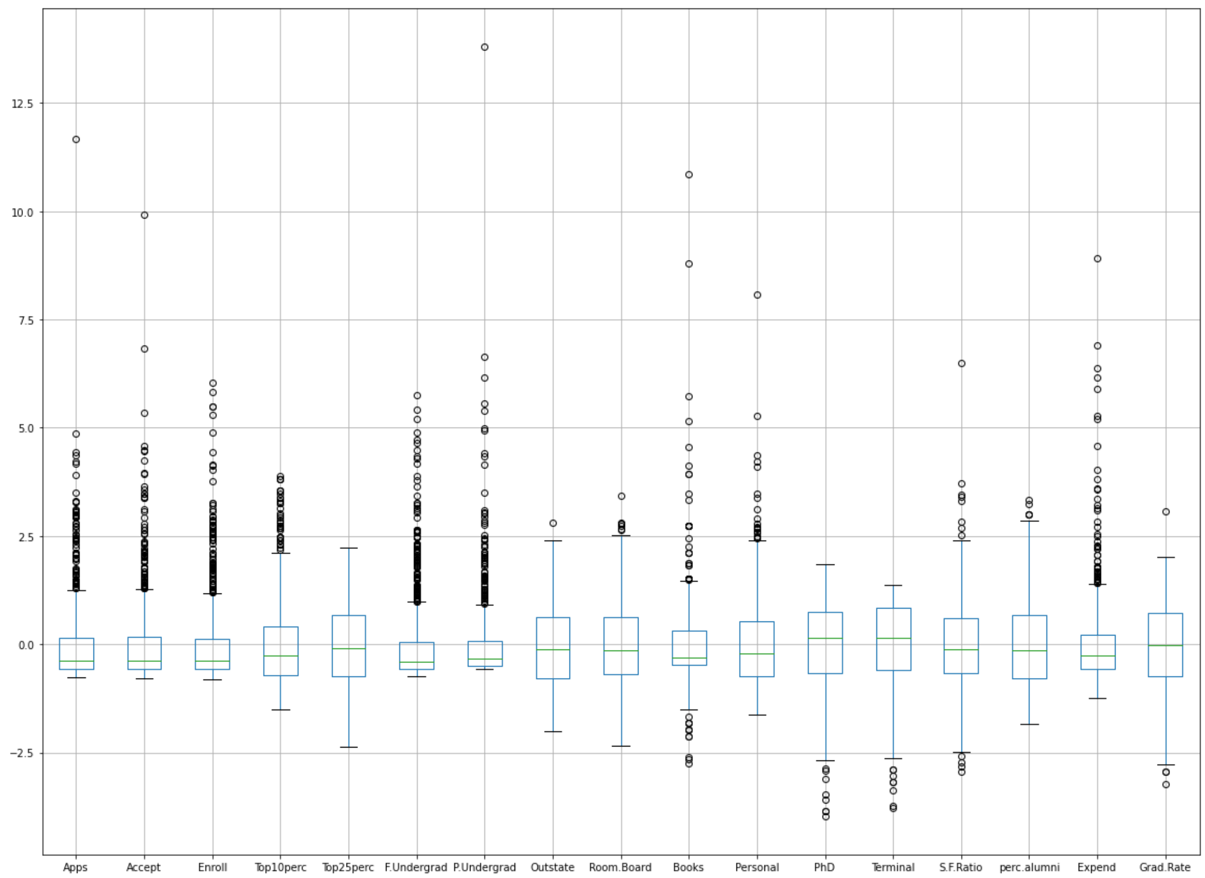

Fig. 6

Table for scaled data with outliers:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate
count	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02
mean	6.355797e-17	6.774575e-17	-5.249269e-17	-2.753232e-17	-1.546739e-16	-1.661405e-16	-3.029180e-17	6.515595e-17
std	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00
min	-7.551337e-01	-7.947645e-01	-8.022728e-01	-1.506526e+00	-2.364419e+00	-7.346169e-01	-5.615022e-01	-2.014878e+00
25%	-5.754408e-01	-5.775805e-01	-5.793514e-01	-7.123803e-01	-7.476067e-01	-5.586426e-01	-4.997191e-01	-7.762035e-01
50%	-3.732540e-01	-3.710108e-01	-3.725836e-01	-2.585828e-01	-9.077663e-02	-4.111378e-01	-3.301442e-01	-1.120949e-01
75%	1.609122e-01	1.654173e-01	1.314128e-01	4.221134e-01	6.671042e-01	6.294077e-02	7.341765e-02	6.179271e-01
max	1.165867e+01	9.924816e+00	6.043678e+00	3.882319e+00	2.233391e+00	5.764674e+00	1.378992e+01	2.800531e+00

without removing the outliers if we scale the data using z score it will affect the mean and the standard deviation of the data from the above analysis, we can see that the standard deviation for scaled data with outliers is 1.00644.

Table for data after removing outliers:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000
mean	2856.956242	1917.760103	748.335907	26.853024	55.796654	3678.852767	744.579408	10436.548263	4347.803089
std	3120.470980	1942.822994	781.271463	15.607194	19.804778	4414.345270	940.269547	4013.095875	1073.326060
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	2340.000000	1780.000000
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	7320.000000	3597.000000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	9990.000000	4200.000000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	12925.000000	5050.000000
max	11066.200000	6979.200000	2757.000000	65.200000	100.000000	14477.800000	3303.600000	20100.000000	7131.000000

Scaling data after removing outliers:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Persor
0	-0.383829	-0.353198	-0.035012	-0.247034	-0.191827	-0.179951	-0.220908	-0.747173	-0.976849	-0.807146	1.5340
1	-0.215156	0.003214	-0.302696	-0.695834	-1.353911	-0.225740	0.513397	0.459655	1.959843	1.912681	0.3257
2	-0.458225	-0.422730	-0.528115	-0.311148	-0.292878	-0.599082	-0.687032	0.202830	-0.557322	-1.260450	-0.2524
3	-0.782423	-0.807984	-0.782992	2.125195	1.677612	-0.718316	-0.725344	0.629209	1.027560	-0.807146	-0.7530
4	-0.854253	-0.912539	-0.888017	-0.695834	-0.596031	-0.777479	0.132410	-0.717252	-0.212377	2.054112	0.3257

Box plot before scaling the data without outliers:

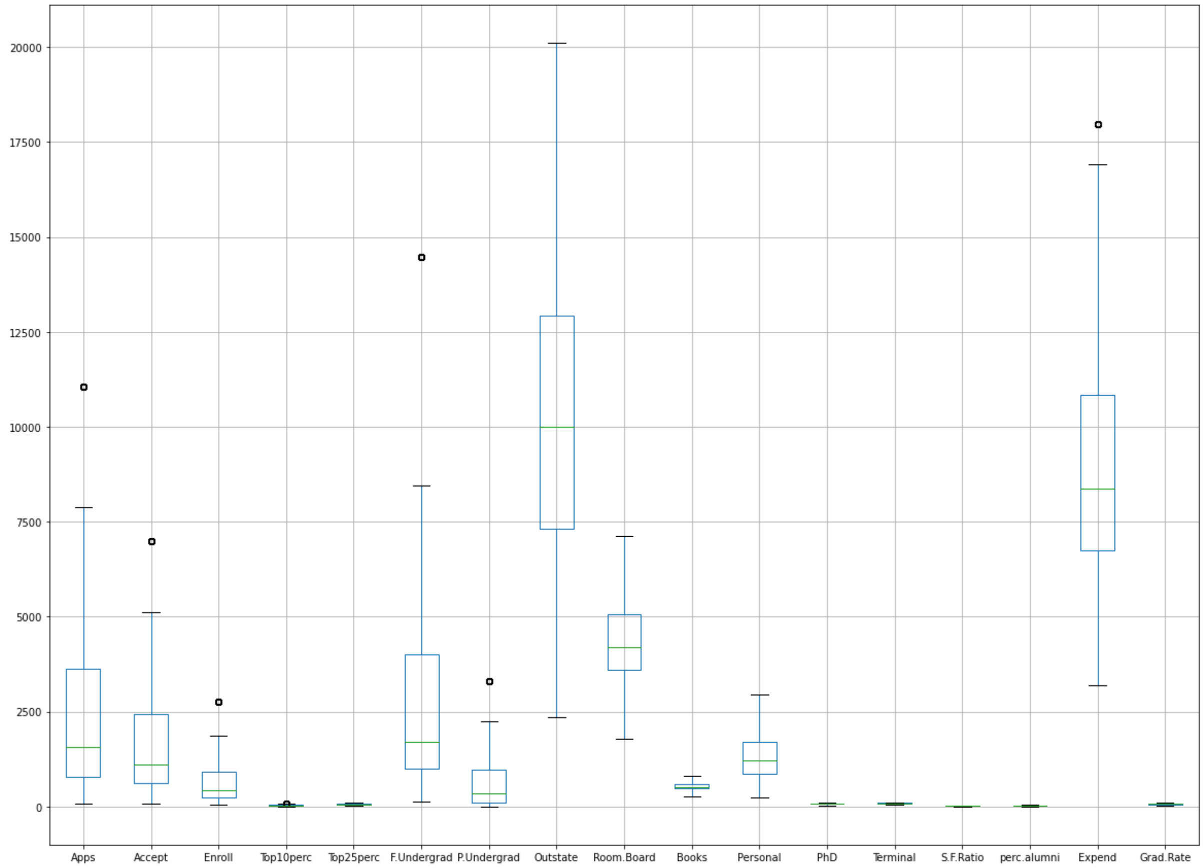

Fig. 7

Box plot after scaling the data without outliers:

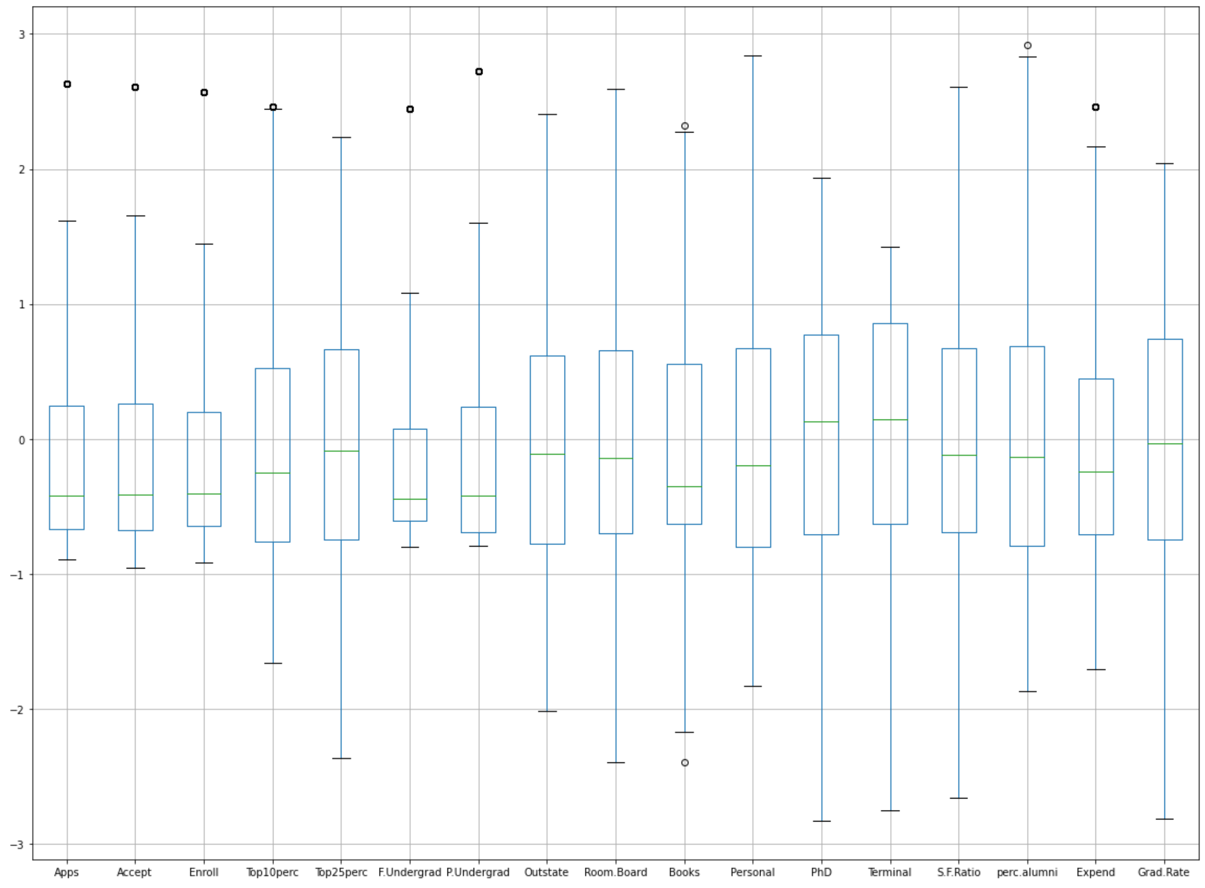

Fig. 8

The outliers are reduced / replaced from the data set.

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

Eigen Vectors

```
%s [[-1.51051724e-01  5.73869368e-01  2.54721171e-02  3.50002377e-01
   4.76265776e-01 -2.73993248e-02 -6.79408155e-02 -1.34049566e-01
   1.84655032e-01 -3.41030317e-02  1.23782113e-02  4.76414519e-02
  -2.28743180e-01 -1.02559773e-01  9.77100175e-02 -3.24930495e-01
  -2.42671239e-01]
 [ 4.52766958e-01 -6.43625404e-01 -4.08143058e-02  1.12837998e-01
   2.08677137e-01 -1.27528369e-01 -2.86891699e-02 -1.23207526e-01
```

```
  1.89697047e-01 -1.02521665e-01 -1.41529768e-03  3.31338141e-02
 -2.02792107e-01 -1.21914245e-01  1.25144023e-01 -3.57755851e-01
 -2.08095876e-01]
[-7.50067816e-01 -2.58381892e-01  3.37484396e-02 -2.25003975e-01
 -2.65981931e-01 -1.80558174e-02 -2.29745788e-02 -4.79563882e-02
  5.20184210e-02 -1.34762063e-01  7.92830517e-03 -3.89761143e-02
 -1.72168365e-01 -1.42497171e-02  9.44419384e-02 -3.95824297e-01
 -1.64564266e-01]
[ 5.89947774e-02 -5.31897461e-02  7.23553559e-01 -3.22924466e-02
  1.62488072e-02  4.57358763e-02 -6.57319491e-03 -7.14429611e-02
 -1.10851590e-01  2.89094711e-01  2.58267694e-01 -8.37673857e-02
 -1.45905144e-01  3.75563233e-01 -7.23866450e-02  7.53900839e-02
 -3.44633526e-01]
[-1.47356588e-02 -3.70257583e-03 -6.58266244e-01  2.64582929e-02
 -3.47432747e-02 -1.58456329e-01 -1.32078205e-01 -4.53255044e-02
 -1.89924670e-01  3.36249057e-01  2.34717438e-01 -2.14918233e-02
 -1.20536687e-01  4.27876370e-01 -4.63368319e-02  3.67211412e-02
 -3.37858398e-01]
[ 4.51829780e-01  4.13880625e-01 -1.05340454e-02 -3.50240396e-01
 -5.20754661e-01  7.88826178e-02 -3.63762678e-02  1.12660606e-02
 -1.41252801e-04 -1.22385171e-01  2.79162755e-02 -5.49956869e-02
 -1.15073146e-01 -1.46165800e-02  8.72397333e-02 -4.06243667e-01
 -1.34287678e-01]
[ 4.97285352e-03 -3.24986907e-02  3.82640339e-02  1.01785907e-01
  1.61437628e-01 -3.58599650e-02  1.89391557e-01  4.23776360e-01
 -7.36103386e-01  5.41905661e-02  9.36586774e-02 -5.16448338e-02
  1.32038801e-01 -2.07265372e-01  3.86964803e-02 -3.54916637e-01
 -1.45128920e-02]
[-4.74030188e-03  9.51907262e-02  2.55089170e-03 -2.23348292e-01
 -7.53206365e-03 -5.57302446e-01  6.09931131e-01 -1.87206448e-01
 -1.46112057e-02  2.38893511e-02 -1.04399025e-01 -1.39668813e-02
 -4.29684243e-02 -2.53851713e-01  2.05908405e-02  2.37362415e-01
 -2.97304568e-01]
[-1.79855036e-02 -2.24494212e-02  3.34522167e-02 -9.10703410e-02
 -8.88393882e-02  1.05909326e-01 -4.62002002e-01 -3.04566995e-01
 -2.17093330e-01  3.55685905e-01 -1.25975104e-01  2.57756666e-01
  9.02072957e-02 -5.66793784e-01 -2.60693995e-02  1.23789047e-01
 -2.51192093e-01]
[ 2.61824511e-03  2.76914586e-03  8.27989701e-03 -4.23639027e-02
  9.11163317e-03 -4.85902891e-02  5.14384110e-02  7.45947123e-02
  1.62931047e-02 -2.56097327e-01  1.39285992e-01  6.08723983e-01
  1.66299240e-02  4.72789590e-02 -7.13557985e-01 -1.06015391e-01
 -9.35681745e-02]
[ 1.80670889e-02 -1.08413791e-02  1.34656529e-03  2.94677640e-02
  2.34506560e-02 -9.33480418e-03  1.75508664e-02 -9.21496211e-02
 -3.28224085e-02  2.51641980e-01 -6.56948779e-01 -3.84138124e-01
 -6.29349774e-02  1.07878431e-01 -5.21834336e-01 -2.35469217e-01
  4.84668755e-02]
[ 3.31773567e-04  5.14472664e-03 -5.66570968e-02 -5.33328919e-01
  4.34191485e-01  1.86806043e-01 -3.61407075e-02  1.25222037e-01
```

```
 1.67560289e-01 -4.70574563e-02 -9.61136496e-02 -6.21741995e-02
 5.47356939e-01  1.23470983e-01  5.72580979e-02 -7.06517594e-02
-3.24667558e-01]
[-1.52860888e-02 -1.06241335e-03  8.90053732e-02  5.22450951e-01
-3.67220609e-01 -2.64231329e-01 -1.04786025e-01  7.52852907e-02
 1.29240181e-01 -1.16057935e-01 -9.84467080e-02 -4.79218101e-02
 5.85124026e-01  7.31469402e-02  3.74577785e-02 -5.96664001e-02
-3.20509921e-01]
[ 1.26724133e-03 -1.49408711e-02 -8.57331081e-03  8.61575853e-02
-4.40006223e-02  2.33516509e-01  3.83414110e-01 -4.58497282e-01
-1.21020302e-01  2.15537935e-01 -1.74587187e-01  4.42093906e-01
 2.26758818e-01  2.83024041e-01  2.58375559e-01 -2.47834896e-01
 1.78476677e-01]
[ 2.63795357e-02  3.68623907e-03  8.85282560e-03  1.17084626e-02
 6.97525565e-02  5.84510444e-02 -1.69270316e-01 -2.50492792e-01
-4.57694691e-01 -6.35277046e-01 -3.21857779e-01 -5.59562277e-03
-1.38310455e-01  2.29944433e-01  1.09906654e-01  2.43261851e-01
-1.98617542e-01]
[-9.78083104e-03 -6.46344811e-02 -1.59986586e-01  2.25352711e-01
-1.17933520e-01  6.64009736e-01  3.96898901e-01 -6.52884786e-02
 5.05354371e-02 -8.64143063e-02  1.50523749e-01 -2.38206017e-01
-2.83072831e-02 -2.20176259e-01 -1.72929690e-01  1.35747859e-01
-3.40157000e-01]
[-2.27570853e-03 -1.81770778e-02  7.18415214e-03  5.27965354e-02
-8.65713354e-02  1.41880695e-01  7.39141063e-02  5.80722250e-01
 1.29825028e-01  1.62585522e-01 -4.63707778e-01  3.72585008e-01
-2.87880353e-01  7.41465533e-02  2.31028150e-01  1.60607758e-01
-2.48644778e-01]]
```

Eigen Values

```
array([5.64307841, 4.82973672, 1.10030644, 0.9966849 , 0.8977433 ,
       0.76549205, 0.58709565, 0.55450358, 0.44319291, 0.38222641,
       0.24563729, 0.14684496, 0.13603844, 0.12376406, 0.07466871,
       0.05597992, 0.03891348])
```

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

Check for presence of correlation:

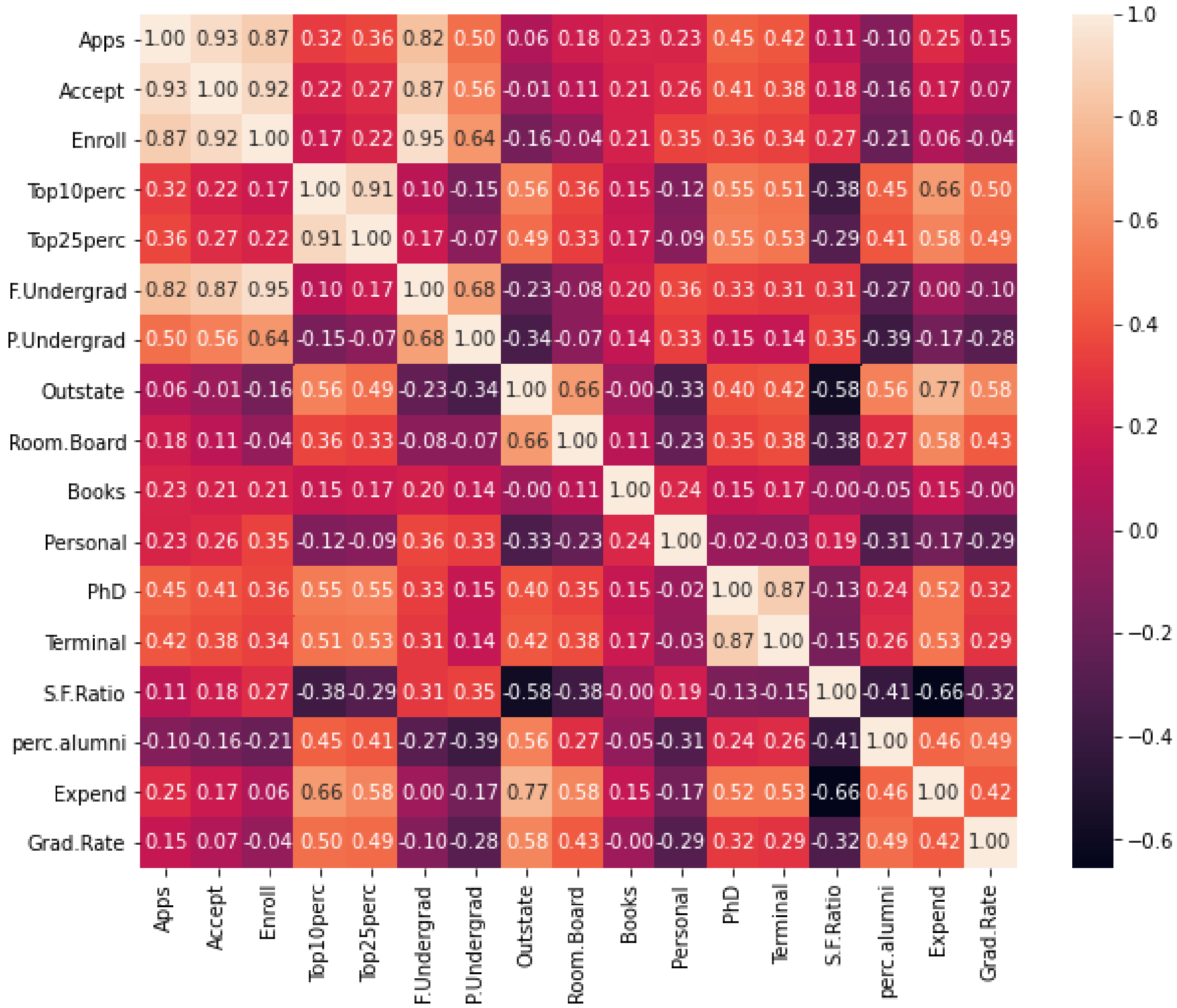

Fig. 9

Confirm the statistical significance of correlations using Bartlett Sphericity test

H0: Correlations are not significant, H1: There are significant correlations

Reject H0 if p-value < 0.05

After running Bartlett Sphericity test p-value is 0, Hence reject H0 and We can confirm there are significant correlations.

After checking significant correlations confirm the adequacy of sample size using KMO Model test.

Note: If value is above 0.7 is good and if value is below 0.5 it is not acceptable.

After performing KMO test value is 0.8561625269639279 which is greater than 0.7.

Next step is to perform PCA for all Components i.e., total 17 columns and below are the values.

pca.components_

```
array([[ 2.42671239e-01,  2.08095876e-01,  1.64564266e-01,
         3.44633526e-01,  3.37858398e-01,  1.34287678e-01,
         1.45128920e-02,  2.97304568e-01,  2.51192093e-01,
         9.35681745e-02, -4.84668755e-02,  3.24667558e-01,
         3.20509921e-01, -1.78476677e-01,  1.98617542e-01,
         3.40157000e-01,  2.48644778e-01],
       [ 3.24930495e-01,  3.57755851e-01,  3.95824297e-01,
        -7.53900839e-02, -3.67211412e-02,  4.06243667e-01,
         3.54916637e-01, -2.37362415e-01, -1.23789047e-01,
         1.06015391e-01,  2.35469217e-01,  7.06517594e-02,
         5.96664001e-02,  2.47834896e-01, -2.43261851e-01,
        -1.35747859e-01, -1.60607758e-01],
       [-9.77100175e-02, -1.25144023e-01, -9.44419384e-02,
         7.23866450e-02,  4.63368319e-02, -8.72397333e-02,
        -3.86964803e-02, -2.05908405e-02,  2.60693995e-02,
         7.13557985e-01,  5.21834336e-01, -5.72580979e-02,
        -3.74577785e-02, -2.58375559e-01, -1.09906654e-01,
         1.72929690e-01, -2.31028150e-01],
       [ 1.02559773e-01,  1.21914245e-01,  1.42497171e-02,
        -3.75563233e-01, -4.27876370e-01,  1.46165800e-02,
         2.07265372e-01,  2.53851713e-01,  5.66793784e-01,
        -4.72789590e-02, -1.07878431e-01, -1.23470983e-01,
        -7.31469402e-02, -2.83024041e-01, -2.29944433e-01,
         2.20176259e-01, -7.41465533e-02],
       [ 2.28743180e-01,  2.02792107e-01,  1.72168365e-01,
         1.45905144e-01,  1.20536687e-01,  1.15073146e-01,
        -1.32038801e-01,  4.29684243e-02, -9.02072957e-02,
        -1.66299240e-02,  6.29349774e-02, -5.47356939e-01,
        -5.85124026e-01, -2.26758818e-01,  1.38310455e-01,
         2.83072831e-02,  2.87880353e-01],
       [ 4.76414519e-02,  3.31338141e-02, -3.89761143e-02,
        -8.37673857e-02, -2.14918233e-02, -5.49956869e-02,
        -5.16448338e-02, -1.39668813e-02,  2.57756666e-01,
         6.08723983e-01, -3.84138124e-01, -6.21741995e-02,
        -4.79218101e-02,  4.42093906e-01, -5.59562277e-03,
        -2.38206017e-01,  3.72585008e-01],
       [-1.23782113e-02,  1.41529768e-03, -7.92830517e-03,
        -2.58267694e-01, -2.34717438e-01, -2.79162755e-02,
        -9.36586774e-02,  1.04399025e-01,  1.25975104e-01,
        -1.39285992e-01,  6.56948779e-01,  9.61136496e-02,
         9.84467080e-02,  1.74587187e-01,  3.21857779e-01,
        -1.50523749e-01,  4.63707778e-01],
       [-3.41030317e-02, -1.02521665e-01, -1.34762063e-01,
         2.89094711e-01,  3.36249057e-01, -1.22385171e-01,
         5.41905661e-02,  2.38893511e-02,  3.55685905e-01,
        -2.56097327e-01,  2.51641980e-01, -4.70574563e-02,
```

```
 -1.16057935e-01,  2.15537935e-01, -6.35277046e-01,
 -8.64143063e-02,  1.62585522e-01],
[-1.84655032e-01, -1.89697047e-01, -5.20184210e-02,
  1.10851590e-01,  1.89924670e-01,  1.41252801e-04,
  7.36103386e-01,  1.46112057e-02,  2.17093330e-01,
 -1.62931047e-02,  3.28224085e-02, -1.67560289e-01,
 -1.29240181e-01,  1.21020302e-01,  4.57694691e-01,
 -5.05354371e-02, -1.29825028e-01],
[-1.34049566e-01, -1.23207526e-01, -4.79563882e-02,
 -7.14429611e-02, -4.53255044e-02,  1.12660606e-02,
  4.23776360e-01, -1.87206448e-01, -3.04566995e-01,
  7.45947123e-02, -9.21496211e-02,  1.25222037e-01,
  7.52852907e-02, -4.58497282e-01, -2.50492792e-01,
 -6.52884786e-02,  5.80722250e-01],
[-6.79408155e-02, -2.86891699e-02, -2.29745788e-02,
 -6.57319491e-03, -1.32078205e-01, -3.63762678e-02,
  1.89391557e-01,  6.09931131e-01, -4.62002002e-01,
  5.14384110e-02,  1.75508664e-02, -3.61407075e-02,
 -1.04786025e-01,  3.83414110e-01, -1.69270316e-01,
  3.96898901e-01,  7.39141063e-02],
[-2.73993248e-02, -1.27528369e-01, -1.80558174e-02,
  4.57358763e-02, -1.58456329e-01,  7.88826178e-02,
 -3.58599650e-02, -5.57302446e-01,  1.05909326e-01,
 -4.85902891e-02, -9.33480418e-03,  1.86806043e-01,
 -2.64231329e-01,  2.33516509e-01,  5.84510444e-02,
  6.64009736e-01,  1.41880695e-01],
[-4.76265776e-01, -2.08677137e-01,  2.65981931e-01,
 -1.62488072e-02,  3.47432747e-02,  5.20754661e-01,
 -1.61437628e-01,  7.53206365e-03,  8.88393882e-02,
 -9.11163317e-03, -2.34506560e-02, -4.34191485e-01,
  3.67220609e-01,  4.40006223e-02, -6.97525565e-02,
  1.17933520e-01,  8.65713354e-02],
[ 3.50002377e-01,  1.12837998e-01, -2.25003975e-01,
 -3.22924466e-02,  2.64582929e-02, -3.50240396e-01,
  1.01785907e-01, -2.23348292e-01, -9.10703410e-02,
 -4.23639027e-02,  2.94677640e-02, -5.33328919e-01,
  5.22450951e-01,  8.61575853e-02,  1.17084626e-02,
  2.25352711e-01,  5.27965354e-02],
[-2.54721171e-02,  4.08143058e-02, -3.37484396e-02,
 -7.23553559e-01,  6.58266244e-01,  1.05340454e-02,
 -3.82640339e-02, -2.55089170e-03, -3.34522167e-02,
 -8.27989701e-03, -1.34656529e-03,  5.66570968e-02,
 -8.90053732e-02,  8.57331081e-03, -8.85282560e-03,
  1.59986586e-01, -7.18415214e-03],
[ 5.73869368e-01, -6.43625404e-01, -2.58381892e-01,
 -5.31897461e-02, -3.70257583e-03,  4.13880625e-01,
 -3.24986907e-02,  9.51907262e-02, -2.24494212e-02,
  2.76914586e-03, -1.08413791e-02,  5.14472664e-03,
 -1.06241335e-03, -1.49408711e-02,  3.68623907e-03,
 -6.46344811e-02, -1.81770778e-02],
```

```
[ 1.51051724e-01, -4.52766958e-01,  7.50067816e-01,
 -5.89947774e-02,  1.47356588e-02, -4.51829780e-01,
 -4.97285352e-03,  4.74030188e-03,  1.79855036e-02,
 -2.61824511e-03, -1.80670889e-02, -3.31773567e-04,
  1.52860888e-02, -1.26724133e-03, -2.63795357e-02,
  9.78083104e-03,  2.27570853e-03]])
```

pca.explained_variance_(Eigen Values):

```
array([5.64307841, 4.82973672, 1.10030644, 0.9966849 , 0.8977433 ,
       0.76549205, 0.58709565, 0.55450358, 0.44319291, 0.38222641,
       0.24563729, 0.14684496, 0.13603844, 0.12376406, 0.07466871,
       0.05597992, 0.03891348])
```

Check the explained variance for each PC

Note: Explained variance = (eigen value of each PC)/(sum of eigen values of all PCs)

pca.explained_variance_ratio_

Below is the output of explained variance for each PC:

```
array([0.33151857, 0.28373652, 0.06464061, 0.05855307, 0.05274046,
       0.04497099, 0.03449059, 0.03257588, 0.02603662, 0.02245497,
       0.01443066, 0.00862682, 0.00799196, 0.00727087, 0.00438662,
       0.0032887 , 0.00228608])
```

Create a data frame containing the loadings or coefficients of all PCs and below is the table.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Apps	0.242671	0.324930	-0.097710	0.102560	0.228743	0.047641	-0.012378	-0.034103	-0.184655	-0.134050	-0.067941	-0.027399
Accept	0.208096	0.357756	-0.125144	0.121914	0.202792	0.033134	0.001415	-0.102522	-0.189697	-0.123208	-0.028689	-0.127528
Enroll	0.164564	0.395824	-0.094442	0.014250	0.172168	-0.038976	-0.007928	-0.134762	-0.052018	-0.047956	-0.022975	-0.018056
Top10perc	0.344634	-0.075390	0.072387	-0.375563	0.145905	-0.083767	-0.258268	0.289095	0.110852	-0.071443	-0.006573	0.045736
Top25perc	0.337858	-0.036721	0.046337	-0.427876	0.120537	-0.021492	-0.234717	0.336249	0.189925	-0.045326	-0.132078	-0.158456
F.Undergrad	0.134288	0.406244	-0.087240	0.014617	0.115073	-0.054996	-0.027916	-0.122385	0.000141	0.011266	-0.036376	0.078883
P.Undergrad	0.014513	0.354917	-0.038696	0.207265	-0.132039	-0.051645	-0.093659	0.054191	0.736103	0.423776	0.189392	-0.035860
Outstate	0.297305	-0.237362	-0.020591	0.253852	0.042968	-0.013967	0.104399	0.023889	0.014611	-0.187206	0.609931	-0.557302
Room.Board	0.251192	-0.123789	0.026069	0.566794	-0.090207	0.257757	0.125975	0.355686	0.217093	-0.304567	-0.462002	0.105909
Books	0.093568	0.106015	0.713558	-0.047279	-0.016630	0.608724	-0.139286	-0.256097	-0.016293	0.074595	0.051438	-0.048590
Personal	-0.048467	0.235469	0.521834	-0.107878	0.062935	-0.384138	0.656949	0.251642	0.032822	-0.092150	0.017551	-0.009335

PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
0.047641	-0.012378	-0.034103	-0.184655	-0.134050	-0.067941	-0.027399	-0.476266	0.350002	-0.025472	0.573869	0.151052
0.033134	0.001415	-0.102522	-0.189697	-0.123208	-0.028689	-0.127528	-0.208677	0.112838	0.040814	-0.643625	-0.452767
0.038976	-0.007928	-0.134762	-0.052018	-0.047956	-0.022975	-0.018056	0.265982	-0.225004	-0.033748	-0.258382	0.750068
0.083767	-0.258268	0.289095	0.110852	-0.071443	-0.006573	0.045736	-0.016249	-0.032292	-0.723554	-0.053190	-0.058995
0.021492	-0.234717	0.336249	0.189925	-0.045326	-0.132078	-0.158456	0.034743	0.026458	0.658266	-0.003703	0.014736
0.054996	-0.027916	-0.122385	0.000141	0.011266	-0.036376	0.078883	0.520755	-0.350240	0.010534	0.413881	-0.451830
0.051645	-0.093659	0.054191	0.736103	0.423776	0.189392	-0.035860	-0.161438	0.101786	-0.038264	-0.032499	-0.004973
0.013967	0.104399	0.023889	0.014611	-0.187206	0.609931	-0.557302	0.007532	-0.223348	-0.002551	0.095191	0.004740
0.257757	0.125975	0.355686	0.217093	-0.304567	-0.462002	0.105909	0.088839	-0.091070	-0.033452	-0.022449	0.017986
0.608724	-0.139286	-0.256097	-0.016293	0.074595	0.051438	-0.048590	-0.009112	-0.042364	-0.008280	0.002769	-0.002618
0.384138	0.656949	0.251642	0.032822	-0.092150	0.017551	-0.009335	-0.023451	0.029468	-0.001347	-0.010841	-0.018067

Create a scree plot and below is the screen plot.

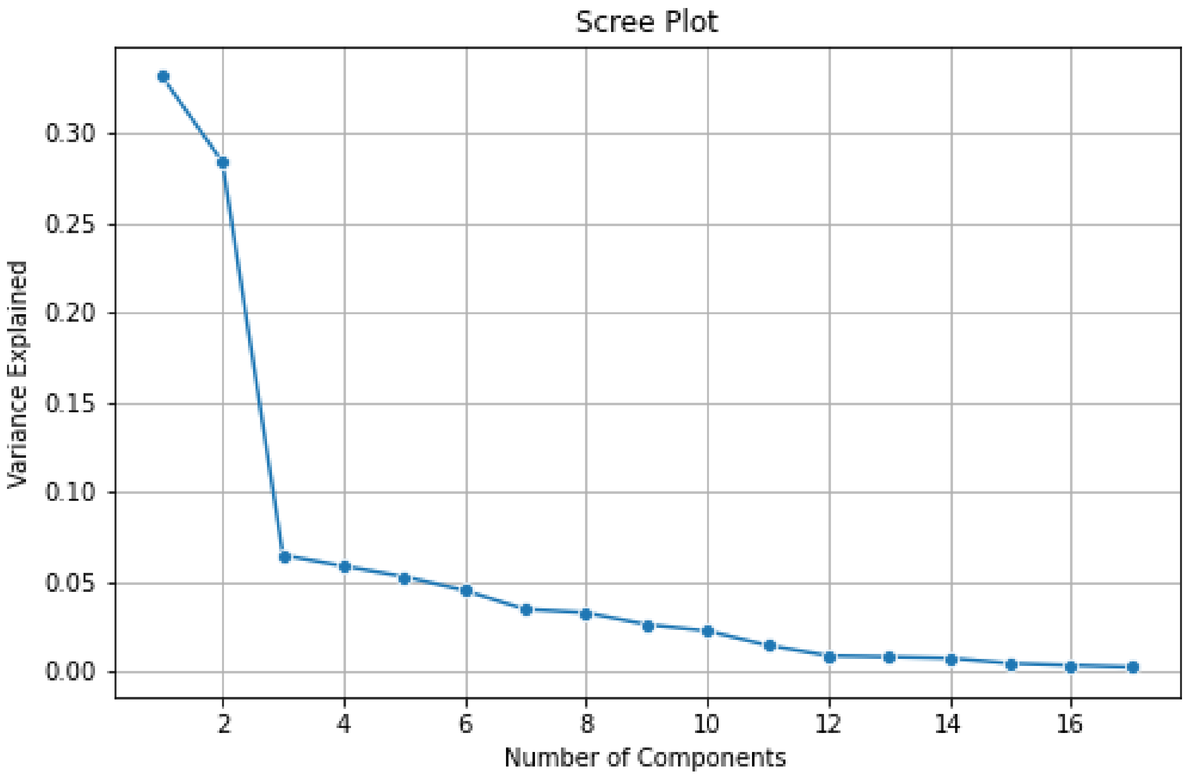

Fig. 10

2.7 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]

In generic first PC can be represented using linear combination of features and its coefficients/weights

PC1=w11 * X1+w12 * X2+w13 * X3+w14 * X4....... where X1,X2,X3,X4... are original variables/features before transformation.
The explicit form of the first PC :

```
array([[ 1.73690056,  1.59813592,  1.54279982, -3.18198787,  1.78588136,
         0.54961821, -0.23204615, -1.90442505, -0.79778763,  2.83704769,
        -1.92917206, -2.19751868,  0.08621937, -0.88461566,  2.20172327,
         1.50532369, -5.22639492,  2.21614773,  2.02850424,  2.98218645,
        -0.16647671, -0.42486665,  1.79587893, -1.14116234, -0.69064956,
         3.45608305, -1.26658595, -1.35107274,  1.65014044, -1.02727695,
         0.94417962, -1.16286652,  2.65509454,  1.96644492,  0.03207601,
         1.15445763, -3.82111755, -3.73537582,  0.47199807, -1.35922637,
        -0.45281261,  0.05659088,  1.70842694,  1.34404146, -0.93545893,
         3.64699493,  2.20110218,  0.40224536, -1.26850553,  0.34056911,
         2.04117675,  1.46787903,  3.21901363,  3.46314232, -1.14836404,
         1.16531024,  1.04781343,  3.13600604,  1.27843908, -4.91626359,
        -5.31141129, -0.72430214,  0.56153994, -0.81583818, -4.20667789,
         1.60678943,  3.93267614,  2.29319743,  0.51838451, -1.32546715,
        -6.59793242, -4.93923926, -4.26581836,  0.57161781, -1.283442  ,
         1.48631412,  0.64358518,  0.4265354 , -1.1451771 ,  0.09866348,
        -0.08561853,  1.52918759,  2.6049984 , -0.05809533, -0.20807246,
         3.38106163, -4.18769947, -5.28607588,  0.62075597,  1.72104422,
         0.72967797, -4.16516146,  1.31424779,  1.30291269, -1.66424398,
         1.62913274, -0.42145742,  1.74463892,  1.72581055, -0.07005301,
         1.86031257, -0.28250191,  2.33118364,  2.26897182,  1.40699704,
         3.28925711, -2.69693384, -0.07606795, -1.54185602,  0.08081483,
        -0.19885417,  1.01382209,  2.32049143,  2.77772781, -4.7839659 ,
        -2.15479279,  1.41194604, -2.35309793, -2.17591392,  3.18717336,
        -0.28956555,  0.83495321, -3.92186372, -4.30582766,  0.21415541,
         1.04757479,  1.84894822, -0.28811643,  1.03895758,  0.12310671,
        -0.18191919,  0.37466172,  0.32386196,  0.65390341,  1.8569272 ,
         2.63498017,  0.60430396, -4.51416199, -3.44208885, -2.16700461,
        -3.46988023, -1.67700432,  2.58019343,  0.63734249, -5.70019554,
         2.27036761,  2.66326344,  1.26870982,  1.69484144, -3.96747211,
        -0.14467483, -0.59535398, -2.44370953,  2.38696944,  3.00668178,
         1.44918346,  3.02853597,  2.25849728, -6.53048149, -4.59684469,
         2.73892451,  3.42540434, -2.41609599, -2.36306709, -2.85470229,
         3.90636028,  1.39094928,  1.60518149,  3.81350428,  1.87077896,
         1.93433894, -1.47345134, -3.66810547, -0.25206698, -6.64135243,
        -2.2066802 , -0.71046121,  1.53200204,  3.78558501,  1.93091022,
         1.69064966,  1.52426423,  1.32334612,  2.64279982, -1.279083  ,
        -0.51041586,  0.0767806 ,  0.64233941,  1.34205088,  2.76880035,
        -0.53228367, -6.31513317,  1.47153133, -0.39317967,  0.58670513,
         1.22839897, -2.64230727,  3.50939423,  3.04251373,  3.20406126,
        -1.07238558, -1.02610389,  1.73657237, -2.28301602,  3.32667337,
        -2.13016508,  1.61403804,  2.80610189,  1.60146762,  0.22913626,
         2.39809165,  2.0405452 ,  0.71167465, -2.63915891,  1.5271081 ,
         2.85790204,  1.71064658,  1.93513291, -0.31802421, -4.11238204,
         0.90790878, -6.06208616, -3.80617907,  1.72399962,  1.84623546,
        -3.27830834,  4.66919982, -0.78664816,  0.41663239,  1.04041768,
        -2.39604654,  2.23913772,  3.32389793,  1.36770308,  1.87242734,
```

2.13662206, 1.98598829, -3.80424423, 0.30990106, 0.15918291,
-1.54935099, 1.24251809, -3.85468208, -1.12739263, -1.28541486,
0.67691215, 0.01785441, 2.36006251, 0.27844912, -0.97215011,
-6.45122619, -5.37442309, 2.21116955, -0.81447022, -0.87435449,
-1.36918636, -2.7611951 , -2.0723166 , -1.44686632, -1.38892581,
-0.79635468, 0.03096386, 1.96725289, 1.37185965, 4.84645824,
4.15377537, -0.01913202, 0.94460077, -1.50119179, 0.09707688,
-2.0244582 , 1.38855617, 0.85365374, 1.02774593, -1.98854423,
3.13077595, 0.09545367, -1.53320528, -1.77309303, -2.32071851,
3.2129836 , 2.78070363, 1.58305584, -1.33383394, -6.11333944,
2.26673242, 2.88469844, -1.77230155, -0.14312723, 2.567682 ,
2.05156724, 1.72425841, -3.33349817, 1.11205935, 0.65002374,
0.97045388, -2.1047155 , 2.19886724, -0.65600216, -3.50575677,
1.27303784, -2.02361073, 2.47059268, 3.30227085, 3.12608606,
2.92524166, -2.58822788, -1.02373212, -0.60670518, -4.69793359,
0.28279537, 1.39702386, 1.84812107, -2.04456229, 2.98999645,
1.94127339, 2.21015667, 1.85084855, -1.39696214, 0.88298054,
2.5713972 , 1.46204944, 1.50148796, 1.49917705, -0.18137951,
2.35861226, -1.58310036, -1.92762182, -0.79730189, -2.0051273 ,
-0.9490778 , 0.49512404, 1.46340032, 3.07567792, -3.00007994,
2.18045408, 3.27606403, 1.03714415, -1.49054947, -1.49839459,
2.52911556, 2.87261975, -1.07761892, 0.02164143, -2.30595358,
0.32207042, -1.09878367, -0.53789135, -0.11957169, 0.67105815,
0.98026727, 0.0132106 , 2.19304619, 0.65552666, -5.82708863,
3.78793178, 2.15461899, 2.79677098, 2.38247633, -1.48338622,
2.04036737, 0.7254916 , 0.35954834, 4.20634029, -0.20437783,
-1.89078517, -2.55309641, -0.72526303, 3.64298877, 0.33552459,
1.60440702, 0.22132666, -0.63029784, 1.81162293, 1.37773269,
0.38907535, 3.52386607, 2.96595648, 3.6936194 , -0.70313631,
0.67814363, 0.75021703, 1.38487562, 0.68061644, 3.73909212,
1.92636339, -0.7256306 , 0.01208241, 2.09537304, 3.67962898,
-3.97897212, 3.19711926, 2.63981286, 2.43883811, 3.50882584,
-0.2388621 , 1.23585411, -0.46543762, -0.07628002, 3.12154995,
-1.56615286, 1.24534686, 0.10200342, 1.76940474, -0.54849479,
-1.30292282, 0.12367594, -6.12157318, 2.11955304, -1.62004837,
2.80973328, 2.14204182, -3.63665797, 1.85357566, -0.58927983,
1.38363199, 0.56569748, 0.08461451, -1.96154826, 0.41515347,
0.06880922, 3.13241131, 2.18678501, 1.60085837, -6.71256985,
0.75915552, 3.23989383, 0.49824259, -4.33928372, -3.22968929,
-1.59404003, -1.11588938, -1.19667536, -2.36689458, 1.9775436 ,
2.99279362, -0.43197194, 0.26272963, 1.68627186, 2.15310943,
-2.44923898, -0.82615818, 1.118989 , -0.21933072, 2.56606863,
-3.02043142, -4.2340279 , 3.1704897 , 2.96319226, 1.08469245,
1.86552065, 1.60314325, 2.86207426, -2.71988612, 1.31027787,
1.98332798, -2.09709989, 2.9469816 , -0.59720461, -6.34326418,
-0.94664993, -2.03336878, 0.1410854 , 0.77513321, -0.6367393 ,
0.96435284, 1.98871559, -0.19855341, -2.06414288, -3.17746669,

-0.6194946 , -4.02487431, -3.2149828 , -0.78045575, -1.07487425,
2.09913114, -0.61628558, 0.93140836, 1.99060892, 2.07272134,
-1.57362333, -0.1056797 , 0.87111529, -3.16966393, -0.48819806,
-0.87915935, 0.37372154, 2.15763886, 0.08241169, 1.97437423,
2.91584624, 1.5759083 , 0.03230475, 1.75339995, 1.99075863,
-1.23625547, -0.39882772, -2.8674525 , -1.14144808, 2.04593782,
1.45156099, -1.03541094, -1.93067397, 0.58604876, 1.1958066 ,
-0.08891941, 1.77529858, -0.11475757, 1.23310808, 0.2548118 ,
-1.64165919, -2.67610943, -3.72188761, 0.86353768, 1.77586486,
-4.19228994, -1.03794776, -1.04055563, -0.73132522, 0.45077953,
0.65348423, 1.79053232, -1.07610296, 3.22672988, -1.55877038,
0.47493077, 2.53726202, -2.6941781 , -4.4406798 , 2.22305299,
2.13258709, 3.56726517, 2.54993974, 1.95715039, -1.83756084,
3.43832836, 1.0979323 , 2.73709154, 3.51847869, 2.177698 ,
-1.72513354, 2.05963418, -0.88441556, 3.23467745, -0.15833685,
-0.4841842 , 0.12202023, -2.91756179, 3.0516152 , -1.75204999,
-0.7726868 , 0.44021456, -0.26693552, 4.05342754, 1.97018594,
1.75400468, -1.1841815 , -2.62462076, 0.76176321, -1.32572879,
-2.04612023, -3.48540179, -3.58893334, -2.72092591, 1.10850954,
-0.17029583, 1.65435743, 0.7113155 , 0.54526005, -1.92148215,
-0.11030747, 0.72357719, 1.15246721, 1.17434865, -1.32612427,
-2.90528195, -3.76059845, 2.51052704, 2.23611059, -0.13905325,
3.36929211, -2.55658317, 1.44209222, -0.6524969 , 2.31000768,
0.85587976, 1.84301035, 1.03388032, 1.06127056, 3.01158075,
-1.62812704, -1.80855431, 2.19965353, -4.28523067, -1.48358266,
0.70776836, -3.54141075, -4.49587568, 2.91161114, 0.88561431,
2.21195798, -4.16903425, 1.5476178 , -0.75962467, 0.36012379,
-5.26580998, -5.25268182, -0.46661987, 2.48633709, -5.79885336,
-2.40287957, -2.94233654, -1.17059355, -1.63781801, -2.12734837,
-1.00847019, 0.79598641, 1.37752308, -0.2469104 , -4.06624932,
-2.59582747, -0.01261397, -0.91234316, -3.15458577, -1.86755044,
1.39624264, -1.82061742, 1.12008256, 0.07496428, 1.8251573 ,
3.31799029, 3.17429498, 0.32552589, -2.04568808, -1.52118317,
0.88830922, -4.28355457, -5.59434025, 0.71013645, 0.05547099,
-2.92855077, 0.58591586, -1.62626613, -0.92048156, 1.66665929,
2.90944782, 2.23565566, -1.6139704 , 0.46192509, -2.15043663,
0.95094954, -4.99971911, -0.1288452 , 0.19020783, 0.3208212 ,
0.15345819, 0.20035019, 0.81515632, 1.52200415, 0.13214566,
-5.26884862, -1.11857138, -1.41987572, -7.13975733, -3.23402806,
-0.69976082, -1.91553636, -1.63303678, -2.48695727, -5.674443 ,
-2.07476957, -0.69273266, 3.80249379, -1.94681834, 3.60030886,
-1.47764749, -0.86184235, -5.0554999 , 2.79777338, 3.5100708 ,
0.97454105, -0.20602572, 0.39227363, -1.29952046, 2.52378432,
-2.49816651, 0.34575892, 2.12075869, -2.07854364, -3.14069714,
-1.20110592, -0.98959544, -2.52090952, -5.12706154, -4.09887053,
1.23990814, 2.21426918, 2.22829043, 1.09682608, 1.98279776,
-3.56343407, 0.50724973, 0.53681439, 2.42216411, -2.05899455,

```
 2.08192528,  4.80398277, -1.57542076, -5.89778682, -4.62546932,
-3.50729574, -0.58909167,  2.94008499, -2.24685904,  2.60908294,
 0.78434249,  1.88434331,  4.78455853, -1.91851702,  0.36438053,
-4.6529743 ,  1.12847435,  1.40815424, -0.07971701, -1.70432339,
-3.83015724, -1.13967762, -2.19308103, -5.97858449,  3.65421649,
 2.65912633,  3.49484069,  2.27674014, -5.05929865, -2.8291165 ,
 1.04086912,  2.45299464, -4.57831608,  0.69702515,  4.23938746,
 0.36370993,  2.35453756, -0.88773896, -0.72152928,  1.19080985,
 2.8432703 , -0.50895166,  2.20644326,  1.45056972,  0.16666725,
 0.83241466, -0.78903614, -1.96816746, -2.73144128,  1.04588602,
-2.56085188, -1.28062868, -0.51339419, -0.7049714 ,  0.23122233,
-2.06283521,  0.802929  ,  2.58288919, -5.28184487,  0.23238862,
 1.66030907,  2.87573106,  0.8529374 ,  2.59607281, -1.6959675 ,
-1.67594748, -2.90984372,  3.58454345, -0.26911621,  0.67443642,
-6.6604879 ,  0.62211979]])
```

2.8 Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?
cumulative values are the created by adding all the eigen values and finding the proportion of each .

By knowing the cumulative values, we can know how many percent of information is captured in each principal components and decide the number of optimum numbers. eigen vector indicate the coefficient of the features or numerical columns.

Check the selected PCs.

	PC1	PC2	PC3	PC4	PC5	PC6
Apps	0.242671	0.324930	-0.097710	0.102560	0.228743	0.047641
Accept	0.208096	0.357756	-0.125144	0.121914	0.202792	0.033134
Enroll	0.164564	0.395824	-0.094442	0.014250	0.172168	-0.038976
Top10perc	0.344634	-0.075390	0.072387	-0.375563	0.145905	-0.083767
Top25perc	0.337858	-0.036721	0.046337	-0.427876	0.120537	-0.021492
F.Undergrad	0.134288	0.406244	-0.087240	0.014617	0.115073	-0.054996
P.Undergrad	0.014513	0.354917	-0.038696	0.207265	-0.132039	-0.051645
Outstate	0.297305	-0.237362	-0.020591	0.253852	0.042968	-0.013967
Room.Board	0.251192	-0.123789	0.026069	0.566794	-0.090207	0.257757
Books	0.093568	0.106015	0.713558	-0.047279	-0.016630	0.608724
Personal	-0.048467	0.235469	0.521834	-0.107878	0.062935	-0.384138
PhD	0.324668	0.070652	-0.057258	-0.123471	-0.547357	-0.062174
Terminal	0.320510	0.059666	-0.037458	-0.073147	-0.585124	-0.047922
S.F.Ratio	-0.178477	0.247835	-0.258376	-0.283024	-0.226759	0.442094
perc.alumni	0.198618	-0.243262	-0.109907	-0.229944	0.138310	-0.005596
Expend	0.340157	-0.135748	0.172930	0.220176	0.028307	-0.238206
Grad.Rate	0.248645	-0.160608	-0.231028	-0.074147	0.287880	0.372585

Check as to how the original features matter to each PC

Note: Here we are only considering the absolute values

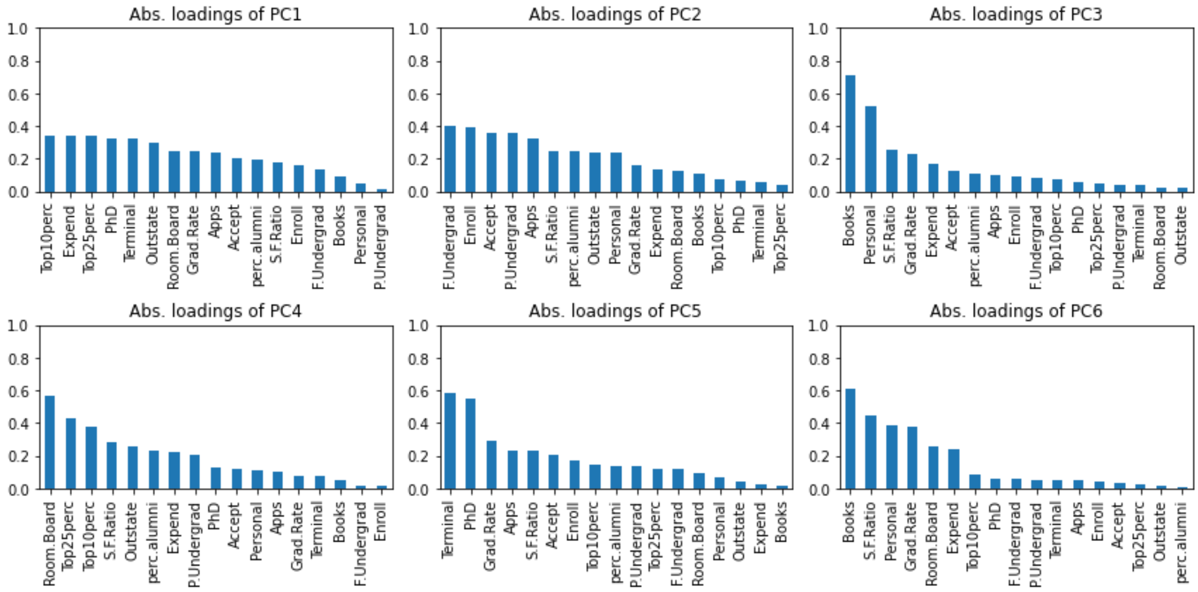

Fig. 11

Compare how the original features influence various PCs

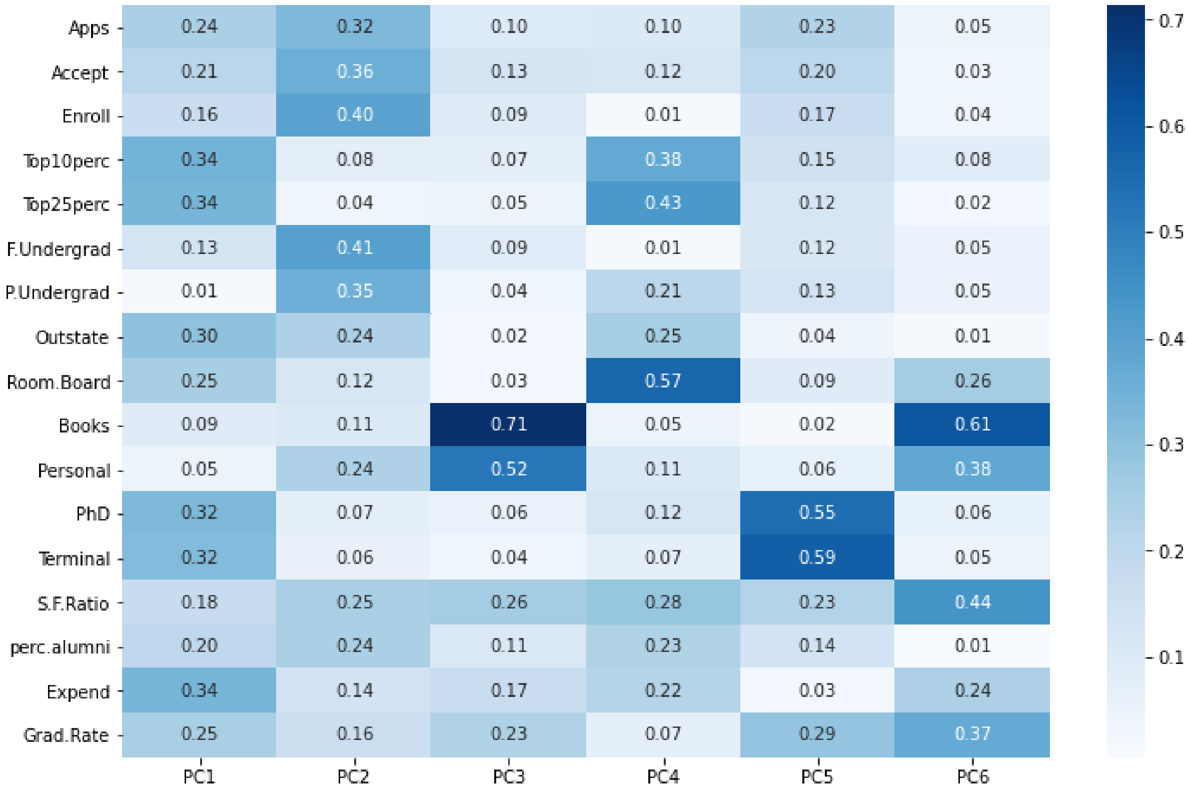

Fig. 12

Extract the required(as per the cumulative explained variance) number of PCs by creating a data frame out of fit transformed scaled data

	PC1	PC2	PC3	PC4	PC5	PC6
0	-1.736901	0.786523	0.091333	-1.018149	-0.351402	-0.765610
1	-1.598136	-0.332040	2.129008	2.898618	1.927793	1.364934
2	-1.542800	-1.379268	-0.602489	0.005509	0.955652	-0.965602
3	3.181988	-2.993983	0.335529	-0.456312	-0.915075	-1.753029
4	-1.785881	-0.202226	2.731234	0.689054	-1.194913	0.174538
5	-0.549618	-1.823884	0.164432	-0.211133	0.244816	-0.839955
6	0.232046	-1.661746	0.276294	0.957245	-1.712301	-0.370757
7	1.904425	-1.642138	-0.988321	-0.497628	-1.039238	-0.255905
8	0.797788	-2.344255	-1.933845	0.354534	-0.240210	-0.984291
9	-2.837048	-1.026997	2.106880	0.260420	2.173966	-0.123553

Check for presence of correlations among the PCs

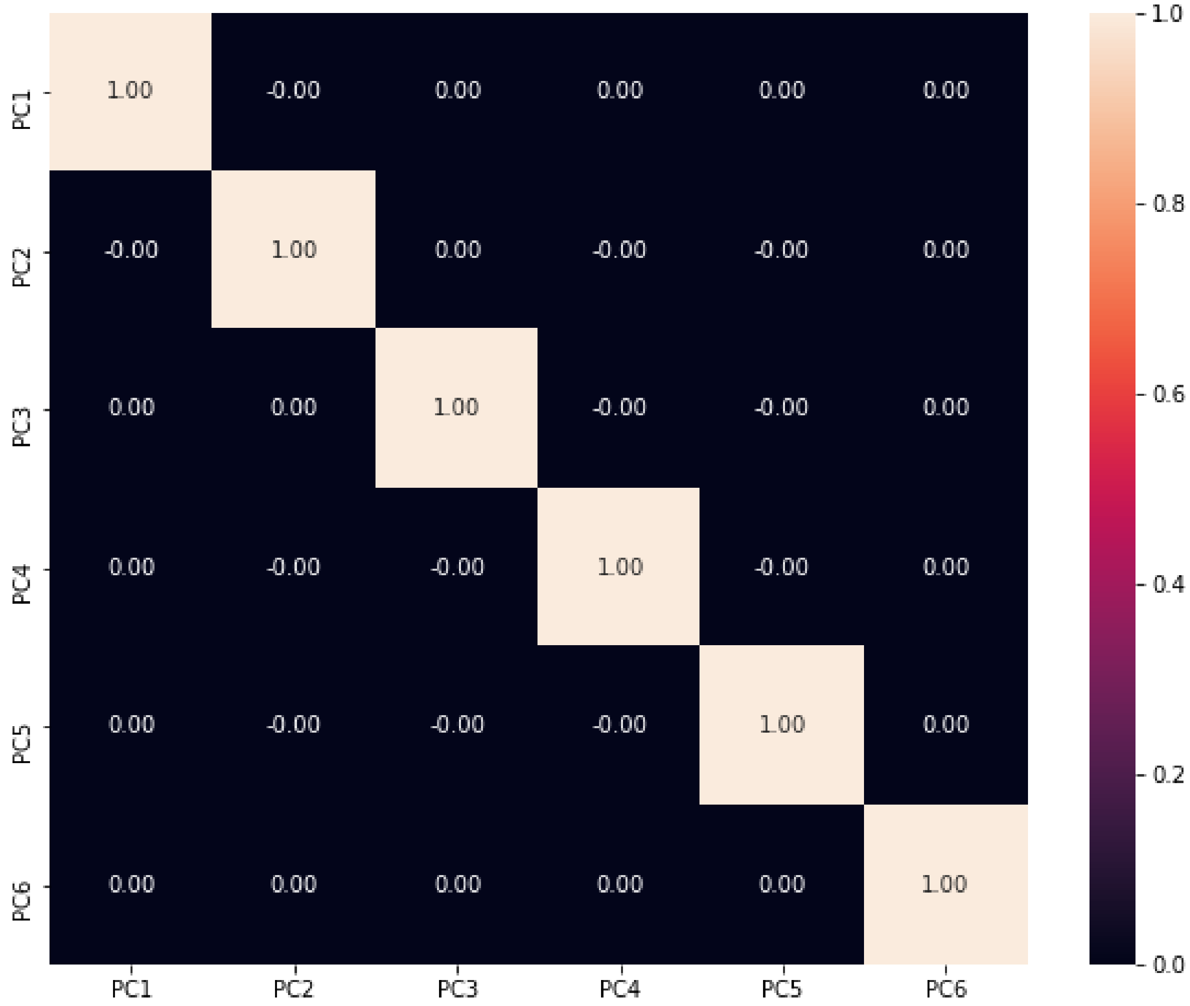
PC1
PC2
PC3
PC4
PC5
PC6
1.00
-0.00
0.00
0.00
0.00
0.00
-0.00
1.00
0.00
-0.00
-0.00
0.00
0.00
0.00
1.00
-0.00
-0.00
0.00
0.00
-0.00
-0.00
1.00
-0.00
0.00
0.00
-0.00
-0.00
-0.00
1.00
0.00
0.00
0.00
0.00
0.00
0.00
1.00
PC1
PC2
PC3
PC4
PC5
PC6
1.0
0.8
0.6
0.4
0.2
0.0

Fig. 13

2.9 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]

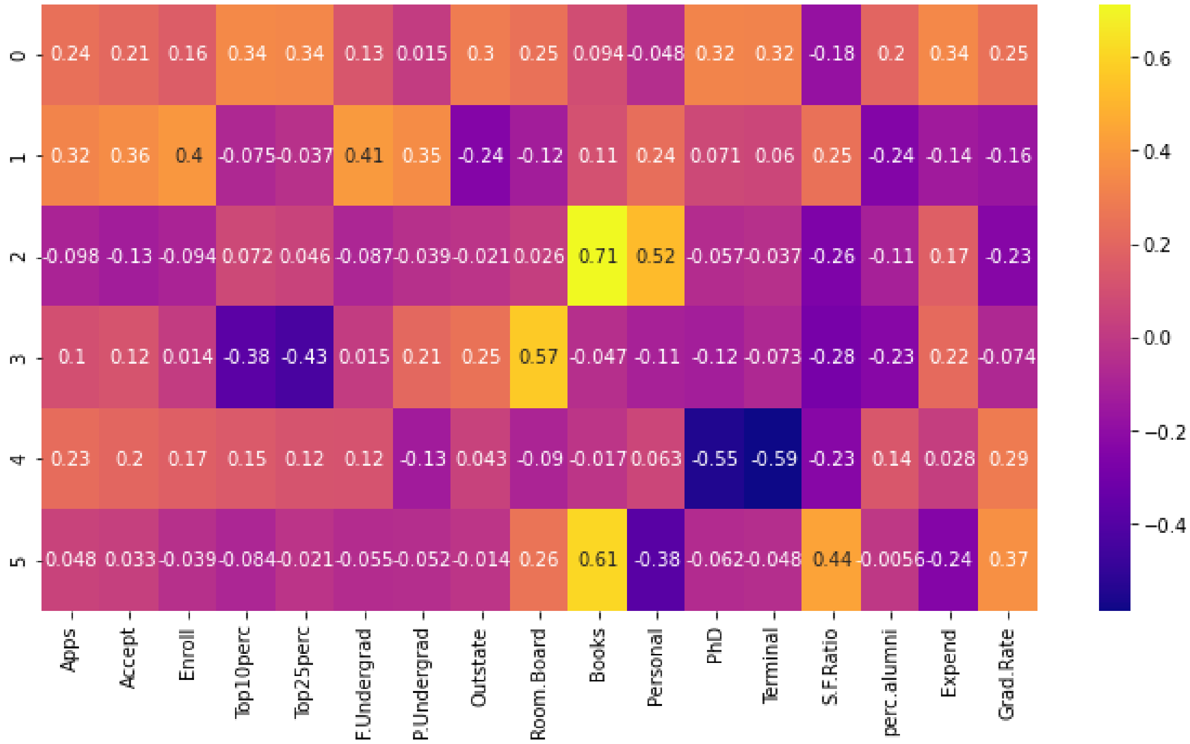

Fig. 14

The Six Principal components created are free from multicollinearity

Just Six PCA (out of 17) components is picking up around 68 % of variability .

PC0 explains most of variables at average level of .24 with good explanity for top 10 perc , top 20 perct,expend,phd, terminal, outstate variables.

PC1 has good explanity for f.undergrad ,enroll ,accept,punderground , accept and apps.

PC2 has highest explanity for Books and personal.

PC3 has good explanity for Room.board

PC5 has highest explanity for Books.

XXXXX