ADVANCED STATISTICS
PROJECT -3 REPORT

Piyush Kumar Singh
PGP – DSBA Online
May-21 Batch

Date: 10/08/2021

Table of Contents:

List of Figures

List of Tables

Problem 1A :

Problem Statement

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.

Dataset Head

	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 Variables

```
<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
```

Inference – There is no missing value in the dataset.

Education, Occupation Variables are of Object type and Salary Variable is an integer type.

Summary of Dataset

	Salary
count	40.000000
mean	162186.875000
std	64860.407506
min	50103.000000
25%	99897.500000
50%	169100.000000
75%	214440.750000
max	260151.000000

Table.2

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

Salary is dependent variable as it is continuous and normally distributed.

Education and Occupation are independent variables.

One Way Anova (Education)

H0: Mean salary of different education levels are equal.

H1: The mean salary is different for at least one education level.

One Way Anova (Occupation)

H0: Mean salary of different Occupation is equal.

H1: The mean salary is different for at least one Occupation level.

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

One Way Anova (Education)

H0: Mean salary of different education levels are equal.

H1: The mean salary is different for at least one education level.

	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.3

Since p_value = 1.257709e-08 is less than aplha (0.05) we reject H0 and conclude that there is significant difference between mean salaries across different education levels.

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

One Way Anova (Occupation)

H0: Mean salary of different Occupation is equal.

H1: The mean salary is different for at least one Occupation level.

	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.4

Since p_value = 0.458508 is greater than aplha (0.05) we fail to reject H0 means there is not significant difference between mean salaries across different occupation levels.

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

We rejected the null hypothesis in (1.2) for Education concluding that there is significant difference between mean salaries across different education levels.

To find out which class means are significantly different the Tukey Honesty Significant Difference test is performed. When tukeyhsd is performed on category Education we get this 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.5

The 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. Bachelors and Doctorate shows maximum difference of all three.

Problem 1B:

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

We analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot like point plot from seaborn library.

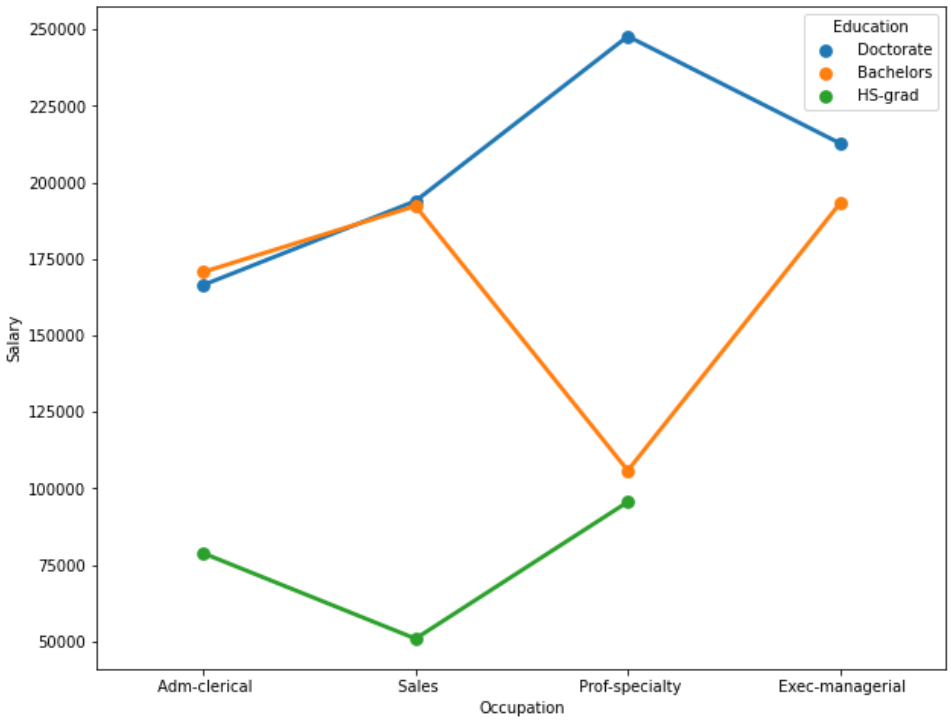

Fig. 1 Point Plot

From the point plot we can clearly see there significant amount of interaction between the category “Education” and “Occupation”.

Observation from the point plot that can be inferred are as follows:

- People with HS-grad earn lowest salaries when compared to all others.
- Prof-speciality people with Doctorate degree earn the highest salary and Sales people with HS-grad earn the lowest salaries.
- Sales people with Bachelor degree and Exec-Managerial people with bachelor degree ear almost same salaries.
- Exec-Managerial person with Doctorate degree earn more salary than Exec-Managerial person with Bachelor degree.
- Sales people with Bachelors and Doctorate degree earn almost same salaries and more than HS-grad people.
- Adm-clerical people with Bachelors and Doctorate degree earn almost same salaries more than HS-grad people.
- Only people with Bachelors and Doctorate degree get post Exec-managerial.
- Adm-clerical, sales, Prof-speciality is hold by people with HS-grad, Bachelors and Doctorate.

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

Hypothesis formulation for 2-Way Anova is as follows:

H0: There is no interaction effect between the 2 independent variables, education and occupation when it comes to the mean Salary

H1: There is an interaction effect between the independent variable education and occupation when it comes to the mean Salary

Now when we perform 2-Way Anova we get the following table

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	5.277862	4.993238e-03
C(Education)	2.0	9.695663e+10	4.847831e+10	68.176603	1.090908e-11
C(Occupation):C(Education)	6.0	3.523330e+10	5.872217e+09	8.258287	2.913740e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Table.6

Now row 3 shows the interaction values between the two variables.

Here p_value = 2.913740e-05(PR(>F)) is less than 0.05 so we can reject H0 hypothesis. This means there is significant interaction effect between the independent variable education and occupation on the mean salary.

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

From the 1-ANOVA and 2-ANOVA performed for this particular case we can conclude that there is interaction between the independent variable education and occupation with respect to mean salary.

Also from the line plot as show above we can clearly see that people with higher education and occupation earn more salary. Hence salary is dependent on both education and occupation.

Problem 2:

Problem Statement

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.

Dataset Head

	0	1	2	3	4
Names	Abilene Christian University	Adelphi University	Adrian College	Agnes Scott College	Alaska Pacific University
Apps	1660	2186	1428	417	193
Accept	1232	1924	1097	349	146
Enroll	721	512	336	137	55
Top10perc	23	16	22	60	16
Top25perc	52	29	50	89	44
F.Undergrad	2885	2683	1036	510	249
P.Undergrad	537	1227	99	63	869
Outstate	7440	12280	11250	12960	7560
Room.Board	3300	6450	3750	5450	4120
Books	450	750	400	450	800
Personal	2200	1500	1165	875	1500
PhD	70	29	53	92	76
Terminal	78	30	66	97	72
S.F.Ratio	18.1	12.2	12.9	7.7	11.9
perc.alumni	12	16	30	37	2
Expend	7041	10527	8735	19016	10922
Grad.Rate	60	56	54	59	15

Table.7

Dataset Variables

```
<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
```

Inference – There is no missing value in the dataset. Also no duplicate values are present in dataset. Except Name which is object type data all rest are integer type data.

Summary of 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

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

Please refer to output of in codes.

After analysis the summary of each column, their boxplot and distplot for each columns individually we can draw the following conclusions.

Insights of Univariate Analysis -

Apps	skeweness - right, outliners - yes
Accept	skeweness - right, outliners - yes
Enroll	skeweness - right, outliners - yes
Top10perc	skeweness - right, outliners - yes
F.Undergrad	skeweness - right, outliners - yes
P.Undergrad	skeweness - right, outliners - yes
Books	skeweness - right, outliners - yes
Personal	skeweness - right, outliners - yes
Expend	skeweness - right, outliners - yes
Top25perc	noramally distributed, outliners - no
Outstate	noramally distributed, outliners - yes
Room.Board	noramally distributed, outliners - yes
Grad.Rate	noramally distributed, outliners - yes
S.F.Ratio	noramally distributed, outliners - yes
perc.alumni	noramally distributed, outliners - yes
PhD	skeweness - left, outliners - yes
Terminal	skeweness - left, outliners - yes

Insights of Bivariant Analysis -

Apps, Accept, Enroll, F.Undergrad are highly correlated with each other.

Top25perc is highly correlated with Top10perc.

Terminal is highly correlated with Ph.d.

Note : Graphs of Bivariant Analysis are shown in next page.

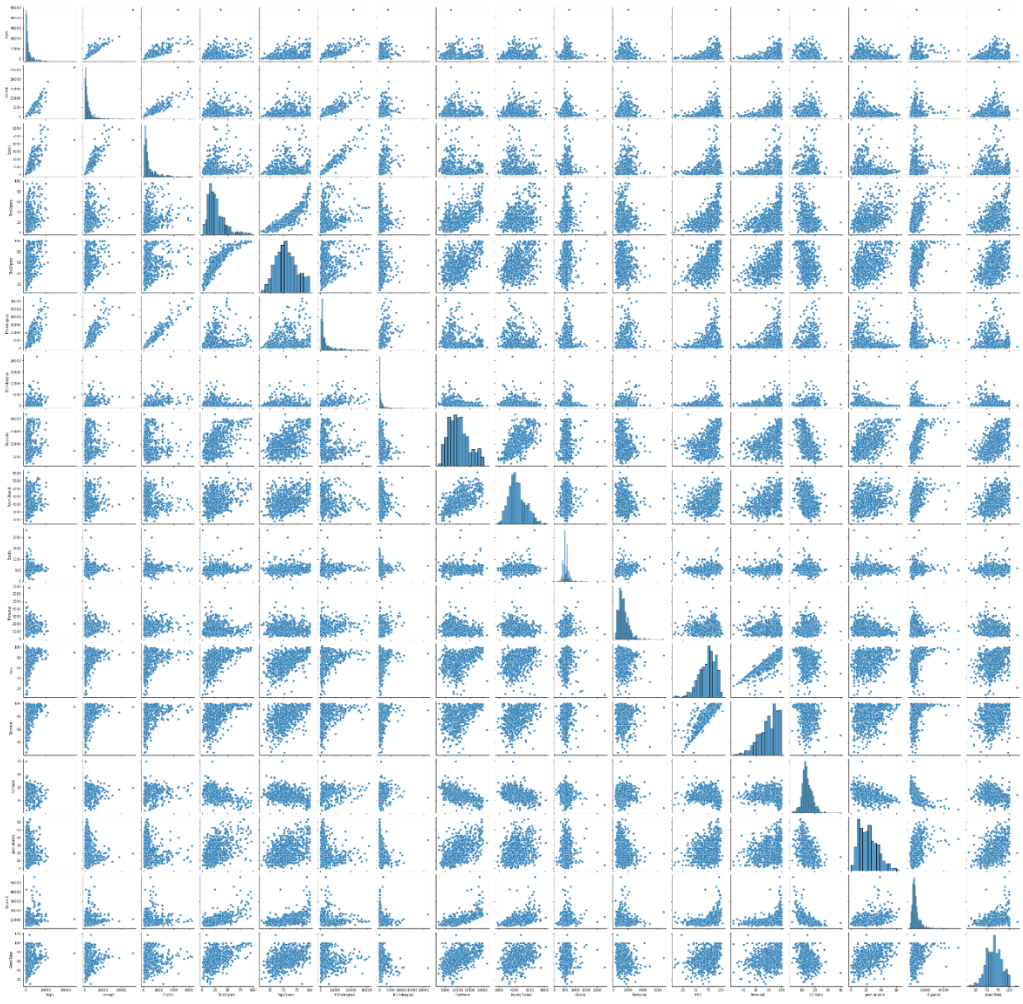

Fig. 2 Scatter Plot

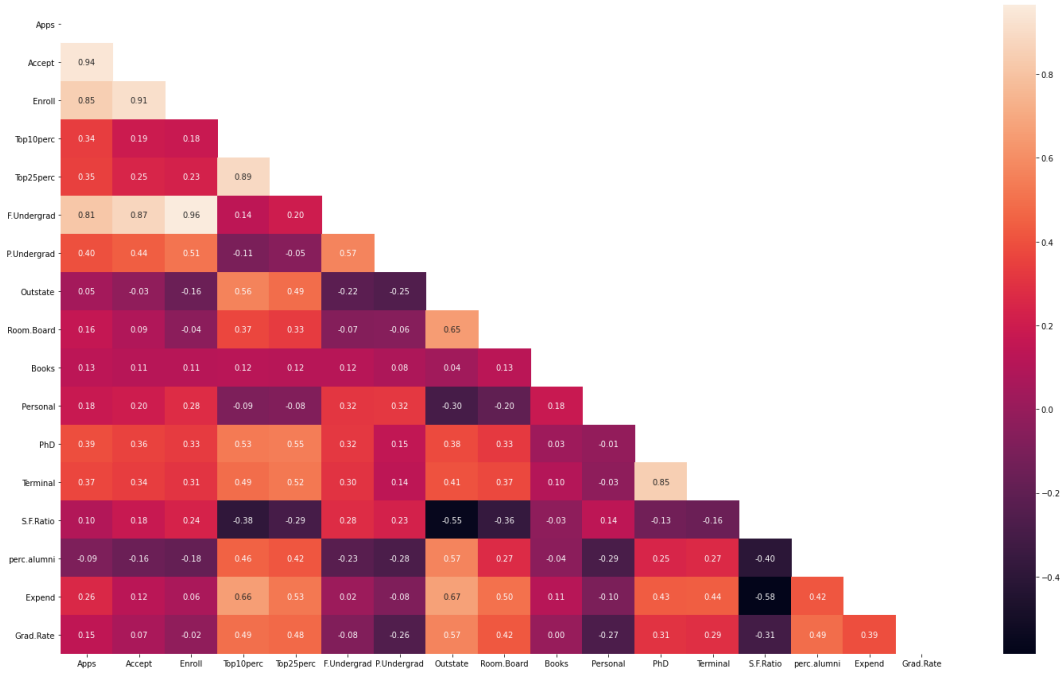

Fig.3 Scatter Heatmap

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

We create a new dataset after removing the name column and analyse the summary of this new dataset. It can be seen range of all the values of all columns vary highly from 0 to 56233. So scaling needs to be performed to make the range of values of columns consistent and to be able to perform PCA on the resulting data.

Dataset before 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.9

Dataset After Scaling –

Here we are applying Z score scaling to scale the dataset.

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

Table.10

Now all the variable are consistent in range and are in between -4 and 14.

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

Correlation Matrix: [on scaled data]

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369491	0.095633	-0.090226	0.259592	0.146755
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758	0.337583	0.176229	-0.159990	0.124717	0.067313
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469	0.308274	0.237271	-0.180794	0.064169	-0.022341
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828	0.491135	-0.384875	0.455485	0.660913	0.494989
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862	0.524749	-0.294629	0.417864	0.527447	0.477281
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.318337	0.300019	0.279703	-0.229462	0.018652	-0.078773
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.081200	0.319882	0.149114	0.141904	0.232531	-0.280792	-0.083568	-0.257001
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.038855	-0.299087	0.382982	0.407983	-0.554821	0.566262	0.672779	0.571290
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.127963	-0.199428	0.329202	0.374540	-0.362628	0.272363	0.501739	0.424942
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.000000	0.179295	0.026906	0.099955	-0.031929	-0.040208	0.112409	0.001061
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.000000	-0.010936	-0.030613	0.136345	-0.285968	-0.097892	-0.269344
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1.000000	0.849587	-0.130530	0.249009	0.432762	0.305038
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.099955	-0.030613	0.849587	1.000000	-0.160104	0.267130	0.438799	0.289527
S.F.Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628	-0.031929	0.136345	-0.130530	-0.160104	1.000000	-0.402929	-0.583832	-0.306710
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009	0.267130	-0.402929	1.000000	0.417712	0.490898
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739	0.112409	-0.097892	0.432762	0.438799	-0.583832	0.417712	1.000000	0.390343
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.571290	0.424942	0.001061	-0.269344	0.305038	0.289527	-0.306710	0.490898	0.390343	1.000000

Table.11

Covariance Matrix: [on scaled data]

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.001289	0.944666	0.847913	0.339270	0.352093	0.815540	0.398777	0.050224	0.165152	0.132729	0.178961	0.391201	0.369968	0.095756	-0.090342	0.259927	0.146944
Accept	0.944666	1.001289	0.912811	0.192695	0.247795	0.875350	0.441839	-0.025788	0.091016	0.113672	0.201248	0.356216	0.338018	0.176456	-0.160196	0.124878	0.067399
Enroll	0.847913	0.912811	1.001289	0.181527	0.227037	0.965883	0.513730	-0.155678	-0.040284	0.112856	0.281291	0.331896	0.308671	0.237577	-0.181027	0.064252	-0.022370
Top10perc	0.339270	0.192695	0.181527	1.001289	0.893144	0.141471	-0.105492	0.563055	0.371959	0.119012	-0.093437	0.532513	0.491768	-0.385370	0.456072	0.661765	0.495627
Top25perc	0.352093	0.247795	0.227037	0.893144	1.001289	0.199702	-0.053646	0.490024	0.331917	0.115676	-0.080914	0.546566	0.525425	-0.295009	0.418403	0.528127	0.477896
F.Undergrad	0.815540	0.875350	0.965883	0.141471	0.199702	1.001289	0.571247	-0.216020	-0.068979	0.115699	0.317608	0.318747	0.300406	0.280064	-0.229758	0.018676	-0.078875
P.Undergrad	0.398777	0.441839	0.513730	-0.105492	-0.053646	0.571247	1.001289	-0.253839	-0.061405	0.081304	0.320294	0.149306	0.142086	0.232830	-0.281154	-0.083676	-0.257332
Outstate	0.050224	-0.025788	-0.155678	0.563055	0.490024	-0.216020	-0.253839	1.001289	0.655100	0.038905	-0.299472	0.383476	0.408509	-0.555536	0.566992	0.673646	0.572026
Room.Board	0.165152	0.091016	-0.040284	0.371959	0.331917	-0.068979	-0.061405	0.655100	1.001289	0.128128	-0.199685	0.329627	0.375022	-0.363095	0.272714	0.502386	0.425489
Books	0.132729	0.113672	0.112856	0.119012	0.115676	0.115699	0.081304	0.038905	0.128128	1.001289	0.179526	0.026940	0.100084	-0.031970	-0.040260	0.112554	0.001062
Personal	0.178961	0.201248	0.281291	-0.093437	-0.080914	0.317608	0.320294	-0.299472	-0.199685	0.179526	1.001289	-0.010950	-0.030653	0.136521	-0.286337	-0.098018	-0.269691
PhD	0.391201	0.356216	0.331896	0.532513	0.546566	0.318747	0.149306	0.383476	0.329627	0.026940	-0.010950	1.001289	0.850682	-0.130698	0.249330	0.433319	0.305431
Terminal	0.369968	0.338018	0.308671	0.491768	0.525425	0.300406	0.142086	0.408509	0.375022	0.100084	-0.030653	0.850682	1.001289	-0.160310	0.267475	0.439365	0.289900
S.F.Ratio	0.095756	0.176456	0.237577	-0.385370	-0.295009	0.280064	0.232830	-0.555536	-0.363095	-0.031970	0.136521	-0.130698	-0.160310	1.001289	-0.403448	-0.584584	-0.307106
perc.alumni	-0.090342	-0.160196	-0.181027	0.456072	0.418403	-0.229758	-0.281154	0.566992	0.272714	-0.040260	-0.286337	0.249330	0.267475	-0.403448	1.001289	0.418250	0.491530
Expend	0.259927	0.124878	0.064252	0.661765	0.528127	0.018676	-0.083676	0.673646	0.502386	0.112554	-0.098018	0.433319	0.439365	-0.584584	0.418250	1.001289	0.390846
Grad.Rate	0.146944	0.067399	-0.022370	0.495627	0.477896	-0.078875	-0.257332	0.572026	0.425489	0.001062	-0.269691	0.305431	0.289900	-0.307106	0.491530	0.390846	1.001289

Table.12

Insights:

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. But for our case since we have performed on a scaled data set there is no such difference present in both the matrices.

All three approaches will yield the same eigenvectors and eigenvalue pairs:

Eigendecomposition of the covariance matrix after standardizing the data.

Eigendecomposition of the correlation matrix.

Eigendecomposition of the correlation matrix after standardizing the data.

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

Outliers before scaling

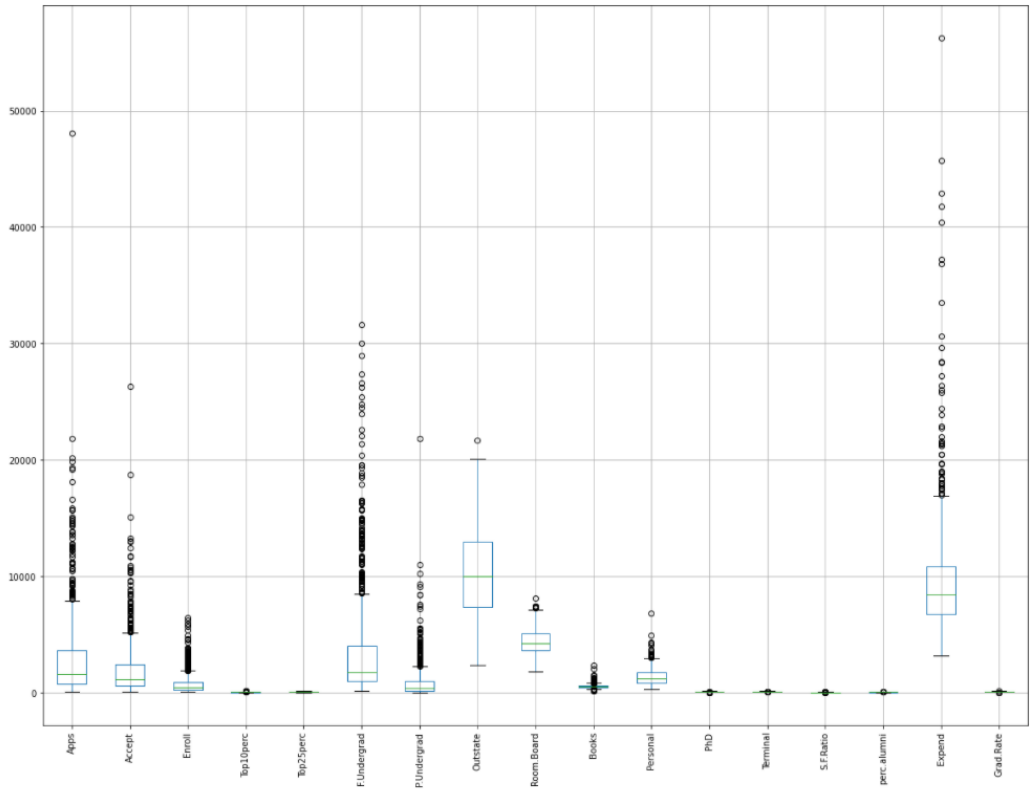

Fig. 4 Box Plot

Outliers after scaling

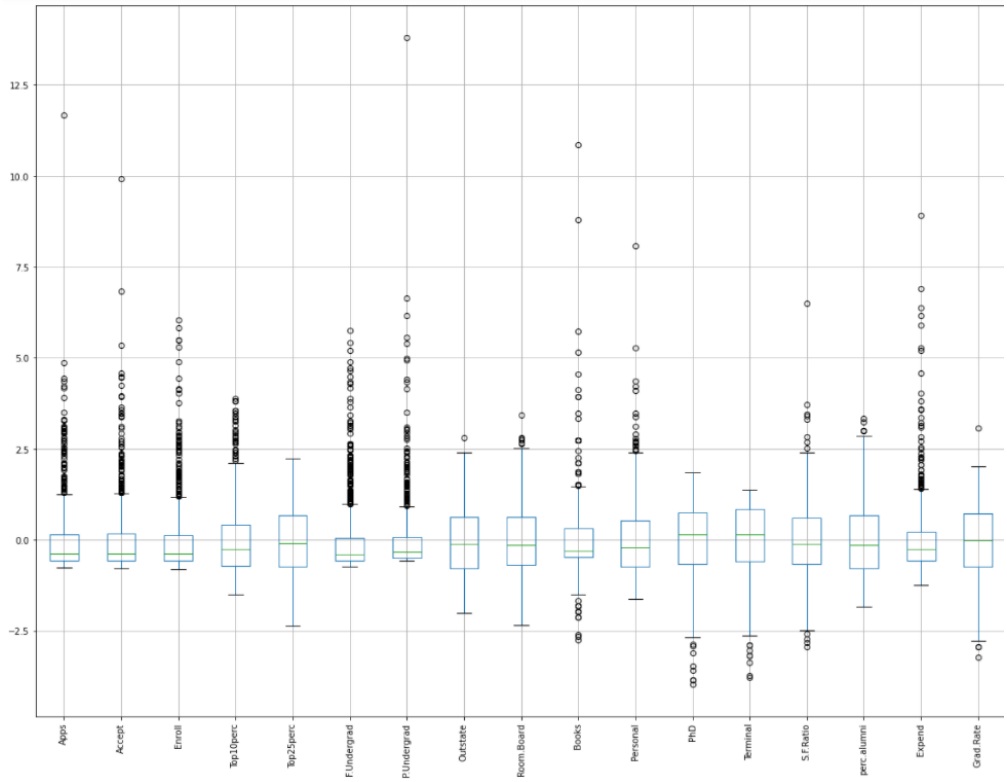

Fig. 5 Box Plot

Insights:

Without removing the outliers if we scale the data using z-score it will affect the mean and the standard deviation of the data. If we analysis the summary of scaled data we can see that the standard deviation for scaled data with outliers is 1.00644. Please refer to summary of scaled data from question 2.2.

2.5 Extract the eigenvalues and eigenvectors.

We first perform Bartletts Test of Sphericity check whether correlation is significant or not.

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

Result - For our dataset it passes the test for p-value=0.

We then perform The Kaiser-Meyer-Olkin to test the sample adequacy.

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 is the dimension and extraction of meaningful components.

Result - For our dataset it passes the test for kmo_model = 0.8131251200373522

Eigenvectors :

```
array ([[ 2.48765602e-01,  2.07601502e-01,  1.76303592e-01,
         3.54273947e-01,  3.44001279e-01,  1.54640962e-01,
         2.64425045e-02,  2.94736419e-01,  2.49030449e-01,
         6.47575181e-02, -4.25285386e-02,  3.18312875e-01,
         3.17056016e-01, -1.76957895e-01,  2.05082369e-01,
         3.18908750e-01,  2.52315654e-01],
        [ 3.31598227e-01,  3.72116750e-01,  4.03724252e-01,
        -8.24118211e-02, -4.47786551e-02,  4.17673774e-01,
         3.15087830e-01, -2.49643522e-01, -1.37808883e-01,
         5.63418434e-02,  2.19929218e-01,  5.83113174e-02,
         4.64294477e-02,  2.46665277e-01, -2.46595274e-01,
        -1.31689865e-01, -1.69240532e-01],
        [-6.30921033e-02, -1.01249056e-01, -8.29855709e-02,
         3.50555339e-02, -2.41479376e-02, -6.13929764e-02,
         1.39681716e-01,  4.65988731e-02,  1.48967389e-01,
         6.77411649e-01,  4.99721120e-01, -1.27028371e-01,
```

-6.60375454e-02, -2.89848401e-01, -1.46989274e-01,
2.26743985e-01, -2.08064649e-01],
[2.81310530e-01, 2.67817346e-01, 1.61826771e-01,
-5.15472524e-02, -1.09766541e-01, 1.00412335e-01,
-1.58558487e-01, 1.31291364e-01, 1.84995991e-01,
8.70892205e-02, -2.30710568e-01, -5.34724832e-01,
-5.19443019e-01, -1.61189487e-01, 1.73142230e-02,
7.92734946e-02, 2.69129066e-01],
[5.74140964e-03, 5.57860920e-02, -5.56936353e-02,
-3.95434345e-01, -4.26533594e-01, -4.34543659e-02,
3.02385408e-01, 2.22532003e-01, 5.60919470e-01,
-1.27288825e-01, -2.22311021e-01, 1.40166326e-01,
2.04719730e-01, -7.93882496e-02, -2.16297411e-01,
7.59581203e-02, -1.09267913e-01],
[-1.62374420e-02, 7.53468452e-03, -4.25579803e-02,
-5.26927980e-02, 3.30915896e-02, -4.34542349e-02,
-1.91198583e-01, -3.00003910e-02, 1.62755446e-01,
6.41054950e-01, -3.31398003e-01, 9.12555212e-02,
1.54927646e-01, 4.87045875e-01, -4.73400144e-02,
-2.98118619e-01, 2.16163313e-01],
[-4.24863486e-02, -1.29497196e-02, -2.76928937e-02,
-1.61332069e-01, -1.18485556e-01, -2.50763629e-02,
6.10423460e-02, 1.08528966e-01, 2.09744235e-01,
-1.49692034e-01, 6.33790064e-01, -1.09641298e-03,
-2.84770105e-02, 2.19259358e-01, 2.43321156e-01,
-2.26584481e-01, 5.59943937e-01],
[-1.03090398e-01, -5.62709623e-02, 5.86623552e-02,
-1.22678028e-01, -1.02491967e-01, 7.88896442e-02,
5.70783816e-01, 9.84599754e-03, -2.21453442e-01,
2.13293009e-01, -2.32660840e-01, -7.70400002e-02,
-1.21613297e-02, -8.36048735e-02, 6.78523654e-01,
-5.41593771e-02, -5.33553891e-03],
[-9.02270802e-02, -1.77864814e-01, -1.28560713e-01,
3.41099863e-01, 4.03711989e-01, -5.94419181e-02,
5.60672902e-01, -4.57332880e-03, 2.75022548e-01,
-1.33663353e-01, -9.44688900e-02, -1.85181525e-01,
-2.54938198e-01, 2.74544380e-01, -2.55334907e-01,
-4.91388809e-02, 4.19043052e-02],
[5.25098025e-02, 4.11400844e-02, 3.44879147e-02,
6.40257785e-02, 1.45492289e-02, 2.08471834e-02,
-2.23105808e-01, 1.86675363e-01, 2.98324237e-01,
-8.20292186e-02, 1.36027616e-01, -1.23452200e-01,
-8.85784627e-02, 4.72045249e-01, 4.22999706e-01,
1.32286331e-01, -5.90271067e-01],
[4.30462074e-02, -5.84055850e-02, -6.93988831e-02,
-8.10481404e-03, -2.73128469e-01, -8.11578181e-02,
1.00693324e-01, 1.43220673e-01, -3.59321731e-01,
3.19400370e-02, -1.85784733e-02, 4.03723253e-02,
-5.89734026e-02, 4.45000727e-01, -1.30727978e-01,
6.92088870e-01, 2.19839000e-01],
[2.40709086e-02, -1.45102446e-01, 1.11431545e-02,
3.85543001e-02, -8.93515563e-02, 5.61767721e-02,
-6.35360730e-02, -8.23443779e-01, 3.54559731e-01,
-2.81593679e-02, -3.92640266e-02, 2.32224316e-02,
1.64850420e-02, -1.10262122e-02, 1.82660654e-01,
3.25982295e-01, 1.22106697e-01],
[5.95830975e-01, 2.92642398e-01, -4.44638207e-01,
1.02303616e-03, 2.18838802e-02, -5.23622267e-01,
1.25997650e-01, -1.41856014e-01, -6.97485854e-02,
1.14379958e-02, 3.94547417e-02, 1.27696382e-01,
-5.83134662e-02, -1.77152700e-02, 1.04088088e-01,
-9.37464497e-02, -6.91969778e-02],

[8.06328039e-02, 3.34674281e-02, -8.56967180e-02,
 -1.07828189e-01, 1.51742110e-01, -5.63728817e-02,
  1.92857500e-02, -3.40115407e-02, -5.84289756e-02,
 -6.68494643e-02, 2.75286207e-02, -6.91126145e-01,
  6.71008607e-01, 4.13740967e-02, -2.71542091e-02,
  7.31225166e-02, 3.64767385e-02],
[1.33405806e-01, -1.45497511e-01, 2.95896092e-02,
  6.97722522e-01, -6.17274818e-01, 9.91640992e-03,
  2.09515982e-02, 3.83544794e-02, 3.40197083e-03,
 -9.43887925e-03, -3.09001353e-03, -1.12055599e-01,
  1.58909651e-01, -2.08991284e-02, -8.41789410e-03,
 -2.27742017e-01, -3.39433604e-03],
[4.59139498e-01, -5.18568789e-01, -4.04318439e-01,
 -1.48738723e-01, 5.18683400e-02, 5.60363054e-01,
 -5.27313042e-02, 1.01594830e-01, -2.59293381e-02,
  2.88282896e-03, -1.28904022e-02, 2.98075465e-02,
 -2.70759809e-02, -2.12476294e-02, 3.33406243e-03,
 -4.38803230e-02, -5.00844705e-03],
[3.58970400e-01, -5.43427250e-01, 6.09651110e-01,
 -1.44986329e-01, 8.03478445e-02, -4.14705279e-01,
  9.01788964e-03, 5.08995918e-02, 1.14639620e-03,
  7.72631963e-04, -1.11433396e-03, 1.38133366e-02,
  6.20932749e-03, -2.22215182e-03, -1.91869743e-02,
 -3.53098218e-02, -1.30710024e-02]])

Eigenvalues :

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

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

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.103090	-0.090227	0.052510	0.043046	0.024071	0.595831	0.080633	0.133406	0.459139	0.358970
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950	-0.056271	-0.177865	0.041140	-0.058406	-0.145102	0.292642	0.033467	-0.145498	-0.518569	-0.543427
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693	0.058662	-0.128561	0.034488	-0.069399	0.011143	-0.444638	-0.085697	0.029590	-0.404318	0.609651
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678	0.341100	0.064026	-0.008105	0.038554	0.001023	-0.107828	0.697723	-0.148739	-0.144986
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486	-0.102492	0.403712	0.014549	-0.273128	-0.089352	0.021884	0.151742	-0.617275	0.051868	0.080348
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454	-0.043454	-0.025076	0.078890	-0.059442	0.020847	-0.081158	0.056177	-0.523622	-0.056373	0.009916	0.560363	-0.414705
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385	-0.191199	0.061042	0.570784	0.560673	-0.223106	0.100693	-0.063536	0.125998	0.019286	0.020952	-0.052731	0.009018
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532	-0.030000	0.108529	0.009846	-0.004573	0.186675	0.143221	-0.823444	-0.141856	-0.034012	0.038354	0.101595	0.050900
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919	0.162755	0.209744	-0.221453	0.275023	0.298324	-0.359322	0.354560	-0.069749	-0.058429	0.003402	-0.025929	0.001146
Books	0.064758	0.056342	0.677412	0.087089	-0.127289	0.641055	-0.149692	0.213293	-0.133663	-0.082029	0.031940	-0.028159	0.011438	-0.066849	-0.009439	0.002883	0.000773
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311	-0.331398	0.633790	-0.232661	-0.094469	0.136028	-0.018578	-0.039264	0.039455	0.027529	-0.003090	-0.012890	-0.001114
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166	0.091256	-0.001096	-0.077040	-0.185182	-0.123452	0.040372	0.023222	0.127696	-0.691126	-0.112056	0.029808	0.013813
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720	0.154928	-0.028477	-0.012161	-0.254938	-0.088578	-0.058973	0.016485	-0.058313	0.671009	0.158910	-0.027076	0.006209
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388	0.487046	0.219259	-0.083605	0.274544	0.472045	0.445001	-0.011026	-0.017715	0.041374	-0.020899	-0.021248	-0.002222
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297	-0.047340	0.243321	0.678524	-0.255335	0.423000	-0.130728	0.182661	0.104088	-0.027154	-0.008418	0.003334	-0.019187
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958	-0.298119	-0.226584	-0.054159	-0.049139	0.132286	0.692089	0.325982	-0.093746	0.073123	-0.227742	-0.043880	-0.035310
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268	0.216163	0.559944	-0.005336	0.041904	-0.590271	0.219839	0.122107	-0.069197	0.036477	-0.003394	-0.005008	-0.013071

Table.13

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

```
The the explicit form of the first PC :
array([[ 1.59285540e+00,  2.19240180e+00,  1.43096371e+00,
        -2.85555743e+00,  2.21200763e+00,  5.71665290e-01,
        -2.41952194e-01, -1.75047435e+00, -7.69126967e-01,
         2.77072109e+00, -1.78458338e+00, -2.04593024e+00,
         9.27889650e-02, -8.69119448e-01,  2.10737261e+00,
         1.39808738e+00, -5.52219259e+00,  2.12992232e+00,
         2.09537648e+00,  2.75495789e+00, -1.44690572e+00,
        -4.65513246e-02,  1.55096615e+00, -1.84319188e+00,
        -4.73231510e-01,  3.19911106e+00, -1.25403504e+00,
        -1.25744524e+00,  1.55427885e+00, -9.50770170e-01,
         8.67345541e-01, -1.08967107e+00,  2.56280258e+00,
         1.73119974e+00,  1.78939136e-02,  1.10581081e+00,
        -3.40792240e+00, -3.83322430e+00,  4.86863403e-01,
        -8.14962228e-01, -4.82406760e-01,  7.99914211e-02,
         1.51248072e+00,  1.32549014e+00, -8.20260242e-01,
         3.40615566e+00,  1.98763942e+00,  3.84777542e-01,
        -1.07423958e+00,  3.18668114e-01,  1.94092825e+00,
         1.37730601e+00,  3.12740663e+00,  3.32988374e+00,
        -1.19754655e+00,  1.08373057e+00,  1.01179971e+00,
         2.91101208e+00,  1.19311027e+00, -5.60708293e+00,
        -5.37279852e+00, -4.62703394e-01,  7.22643658e-01,
        -1.13558028e+00, -3.68693101e+00,  1.45231096e+00,
         3.71460293e+00,  2.18642704e+00,  4.99087352e-01,
        -1.68363951e+00, -6.51617919e+00, -4.57340816e+00,
        -3.77171621e+00,  4.35933693e-01, -1.19565744e+00,
         1.26302353e+00,  5.71771684e-01,  4.51794277e-01,
        -9.20586442e-01,  1.49294827e-01, -1.59439401e-01,
         1.35862470e+00,  2.43677133e+00, -6.88025408e-02,
        -1.82105180e-01,  3.73046474e+00, -4.13631929e+00,
        -4.69724631e+00,  5.27831684e-01,  1.55270445e+00,
         6.67718514e-01, -4.09382308e+00,  1.18313317e+00,
         1.28086682e+00, -1.43649575e+00,  1.69943280e+00,
        -3.83974261e-01,  1.62824472e+00,  1.67245575e+00,
         3.74399572e-04,  2.12177363e+00, -2.10189418e-01,
         2.18793129e+00,  2.08481461e+00,  1.36103063e+00,
         3.02256380e+00, -2.64377042e+00,  9.64720291e-02,
        -1.30299738e+00,  9.89535752e-02, -1.60576244e-01,
         1.00366106e+00,  2.08381531e+00,  2.64141646e+00,
        -4.46661234e+00, -1.94330141e+00,  1.24078594e+00,
        -2.08091416e+00, -1.31038706e+00,  2.96810053e+00,
        -2.75372233e-01,  7.21253161e-01, -3.54435136e+00,
        -3.88319728e+00,  1.93390410e-01,  1.01309680e+00,
         1.61567708e+00, -2.62053681e-01,  1.06404371e+00,
         8.64987482e-02, -1.78776058e-01,  2.57295092e-01,
         2.62414308e-01,  6.06621890e-01,  1.79307334e+00,
```

```
 2.43240494e+00,  5.30057671e-01, -4.37347250e+00,
-3.18773077e+00, -1.89931362e+00, -3.16967463e+00,
-1.22156663e+00,  2.30818480e+00,  5.50134004e-01,
-6.11063916e+00,  2.10768778e+00,  2.48246575e+00,
 1.20024446e+00,  1.51419619e+00, -3.59468357e+00,
-1.03542302e-01, -5.05046842e-01, -2.25940363e+00,
 2.16455701e+00,  2.84409221e+00,  1.38368214e+00,
 2.85456230e+00,  2.14272625e+00, -6.85352991e+00,
-4.31345222e+00,  2.54066216e+00,  3.16708499e+00,
-2.18014104e+00, -2.15098195e+00, -2.53360970e+00,
 3.63666769e+00,  1.28378189e+00,  1.41689534e+00,
 3.57823517e+00,  1.68204697e+00,  1.94887559e+00,
-1.33953337e+00, -3.30149066e+00, -2.29519590e-01,
-7.14050412e+00, -1.89785191e+00, -1.95666653e-01,
 1.62474775e+00,  3.55378150e+00,  1.78551343e+00,
 1.62647322e+00,  1.64785701e+00,  1.31106194e+00,
 2.42522954e+00, -9.87518052e-01, -4.49693161e-01,
 1.44126809e-02,  5.59772492e-01,  1.23854209e+00,
 2.87420782e+00, -4.92153800e-01, -6.26750224e+00,
 1.30207245e+00, -4.55084651e-01,  5.20417890e-01,
 1.20740013e+00, -2.33302057e+00,  3.31124202e+00,
 2.92096156e+00,  2.89935954e+00, -8.82075664e-01,
-8.94174777e-01,  1.56333359e+00, -2.37100338e+00,
 2.93443154e+00, -1.88968681e+00,  1.49451610e+00,
 2.56199145e+00,  1.48107832e+00,  2.79677535e-01,
 2.33376902e+00,  1.90796975e+00,  6.90164570e-01,
-2.33387776e+00,  1.43782771e+00,  2.68745147e+00,
 1.54887252e+00,  1.74980249e+00, -6.20013579e-02,
-3.52576903e+00,  8.65752671e-01, -5.56840118e+00,
-3.52403850e+00,  1.74184513e+00,  1.68007825e+00,
-2.92335256e+00,  5.66290540e+00, -7.14106580e-01,
 3.45920377e-01,  1.04210711e+00, -2.07182044e+00,
 2.06693657e+00,  3.09061993e+00,  1.28844883e+00,
 1.68904918e+00,  2.04446384e+00,  1.83220538e+00,
-3.47098881e+00,  2.04209511e-01,  1.32432280e-01,
-1.51478504e+00,  1.10511173e+00, -3.72193751e+00,
-1.06099853e+00, -1.06792080e+00,  7.29394786e-01,
 4.29892858e-02,  2.30393058e+00,  2.68785067e-01,
-8.34721687e-01, -7.69419904e+00, -5.63947883e+00,
 2.06122014e+00, -7.84512936e-01, -7.71521001e-01,
-1.27050379e+00, -2.44964409e+00, -1.65753799e+00,
-1.26503643e+00, -1.26421415e+00, -7.69653526e-01,
-4.16180234e-03,  1.87624400e+00,  1.25068199e+00,
 5.10404587e+00,  3.76238417e+00,  1.30457227e-02,
 8.99229242e-01, -1.29805615e+00,  4.00858648e-01,
-1.94399531e+00,  1.24358975e+00,  7.92289215e-01,
 1.20330192e+00, -3.51149466e+00,  3.84024754e+00,
 1.82131182e-01, -1.36851478e+00, -1.33913522e+00,
-1.74564453e+00,  2.96583917e+00,  2.67114955e+00,
```

```
 1.44092540e+00, -1.23949821e+00, -8.04718221e+00,
 2.15492097e+00,  2.69526318e+00, -1.61165504e+00,
-1.17022202e-01,  2.38572596e+00,  1.95731291e+00,
 1.57612300e+00, -2.92633200e+00,  1.04988584e+00,
 5.08304352e-01,  9.69709141e-01, -1.93311612e+00,
 2.07919595e+00, -5.98767441e-01, -3.08153362e+00,
 1.33830396e+00, -1.69473860e+00,  2.19462976e+00,
 3.11375970e+00,  2.94202711e+00,  2.75312876e+00,
-2.34218540e+00, -1.07275423e+00, -6.53393348e-01,
-4.15217110e+00,  2.42719465e-01,  1.63779676e+00,
 1.72139339e+00, -1.80449428e+00,  2.82490858e+00,
 1.77726893e+00,  2.16676367e+00,  1.81585720e+00,
-1.34475565e+00,  9.62445051e-01,  2.43290953e+00,
 1.33070428e+00,  1.40838904e+00,  1.43169260e+00,
-1.88368645e-01,  2.23029843e+00, -1.39830292e+00,
-1.68361569e+00, -6.85581349e-01, -1.64681429e+00,
-8.64718266e-01,  4.46721494e-01,  1.39645639e+00,
 2.91316640e+00, -2.69189610e+00,  2.15921500e+00,
 2.97799543e+00,  9.59637554e-01, -1.39796918e+00,
-1.37663162e+00,  2.49624833e+00,  2.97919260e+00,
-1.01682581e+00,  9.00777279e-02, -2.04772518e+00,
 5.61737704e-01, -1.23159847e+00, -5.46547248e-01,
-5.89418255e-02,  6.35792160e-01,  9.54536640e-01,
-5.48819369e-03,  1.96014066e+00,  6.63016034e-01,
-6.74121550e+00,  3.63072045e+00,  2.00219328e+00,
 2.57903691e+00,  2.30698798e+00, -1.43606469e+00,
 1.85594634e+00,  5.92238070e-01,  2.95674418e-01,
 4.24933479e+00, -2.54760926e-01, -1.64785806e+00,
-4.36322951e+00, -6.44259204e-01,  3.90771000e+00,
 3.40333105e-01,  1.54089120e+00,  1.66075294e-01,
-5.18438032e-01,  1.65611046e+00,  1.29791249e+00,
 4.69899530e-01,  3.31353754e+00,  2.68952672e+00,
 3.36685466e+00, -6.48226940e-01,  7.18079576e-01,
 6.34400389e-01,  1.51062277e+00,  7.40770818e-01,
 3.99087478e+00,  1.81911335e+00, -6.73667079e-01,
 1.69928901e-02,  1.97053149e+00,  3.53340728e+00,
-3.47404030e+00,  2.94344811e+00,  2.46650761e+00,
 2.21297616e+00,  3.60959919e+00, -2.24443578e-01,
 1.07558656e+00, -3.93419073e-01, -1.84336634e-01,
 2.98431747e+00, -1.39606474e+00,  1.22814165e+00,
 1.58290694e-01,  1.72496643e+00, -5.04332110e-01,
-1.04922551e+00,  2.20765923e-01, -5.95138487e+00,
 1.91291369e+00, -1.57123552e+00,  2.65486461e+00,
 2.14743757e+00, -3.38670473e+00,  1.88617660e+00,
-8.61126466e-01,  1.33487068e+00,  5.50100958e-01,
 9.50770392e-02, -2.00832649e+00,  7.44084897e-01,
 2.39033814e-01,  2.91573698e+00,  1.99632130e+00,
 1.34101426e+00, -6.50992132e+00,  7.63632477e-01,
 2.93857207e+00,  4.08188074e-01, -3.86845217e+00,
```

-2.83506950e+00, -1.42121822e+00, -9.69110720e-01,
-1.14332826e+00, -2.12824739e+00, 1.81265798e+00,
 2.80846319e+00, -9.60682242e-02, 1.95649055e-01,
 1.53778656e+00, 2.08179131e+00, -1.91491110e+00,
-7.43104985e-01, 1.11575407e+00, -1.46116797e-01,
 2.44128802e+00, -4.12858899e+00, -4.19041903e+00,
 2.92128538e+00, 2.77314474e+00, 1.05219295e+00,
 1.83959984e+00, 1.58465127e+00, 2.65378977e+00,
-2.35558448e+00, 1.33395932e+00, 1.80583021e+00,
-1.87810944e+00, 2.85744193e+00, -5.94080954e-01,
-6.92787140e+00, -9.28124853e-01, -4.38925294e+00,
 1.54474909e-01, 6.19682606e-01, -5.60888849e-01,
 9.30925182e-01, 1.93166485e+00, -1.56126186e-01,
-1.80369757e+00, -2.82397944e+00, -5.46248803e-01,
-3.62952044e+00, -2.91154095e+00, -5.09930373e-01,
-9.94058233e-01, 1.96389709e+00, -5.69211563e-01,
 8.39036835e-01, 1.87612725e+00, 2.02205412e+00,
-1.39024438e+00, -8.95274142e-02, 9.02092753e-01,
-7.32097607e+00, -3.33588654e-01, -6.00409253e-01,
 2.82699376e-01, 1.99477253e+00, -1.55991502e-02,
 2.02739643e+00, 2.73370449e+00, 1.39831414e+00,
 5.69249816e-02, 1.67192736e+00, 2.02525540e+00,
-1.25649677e+00, -3.51363726e-01, -2.05906733e+00,
-1.05924938e+00, 1.79652112e+00, 1.35530286e+00,
-9.73735267e-01, -1.69605960e+00, 5.76739011e-01,
 1.11877635e+00, -4.44898609e-02, 2.25319707e+00,
-1.55003635e-01, 1.16389998e+00, 2.99456676e-01,
-1.25420760e+00, -2.40812052e+00, -3.33467710e+00,
 1.25015944e+00, 1.66448982e+00, -3.94709993e+00,
-9.21106790e-01, -9.61703636e-01, -5.06972632e-01,
 3.70472638e-01, 6.16184387e-01, 1.63246356e+00,
-1.03883723e+00, 3.09760182e+00, -1.36403944e+00,
 3.98014378e-01, 2.41076126e+00, -2.35815594e+00,
-4.18574918e+00, 2.05544943e+00, 1.99504248e+00,
 3.26725008e+00, 2.46780300e+00, 1.87832271e+00,
-1.54377878e+00, 3.13960609e+00, 1.19370548e+00,
 2.59798722e+00, 3.23950787e+00, 2.05588451e+00,
-1.48601433e+00, 1.90071093e+00, -7.28636151e-01,
 3.12295146e+00, -1.25925768e-01, -3.64347326e-01,
 4.81632835e-02, -2.53079617e+00, 2.85457669e+00,
-1.60594992e+00, -6.05151821e-01, 4.78963934e-01,
-3.11997195e-01, 3.95194094e+00, 1.80784404e+00,
 1.66156370e+00, -1.04283538e+00, -2.43929538e+00,
 7.67441643e-01, -1.30848442e+00, -1.89580075e+00,
-3.04288584e+00, -3.60796649e+00, -2.13365307e+00,
 1.12506071e+00, 5.09105731e-02, 1.66937671e+00,
 7.59232595e-01, 5.77891700e-01, -1.57244551e+00,
 1.83283417e-01, 7.48228023e-01, 1.08122046e+00,
 1.11932687e+00, -1.26920428e+00, -2.57726942e+00,

```
-2.97507741e+00,  2.31255480e+00,  2.13897968e+00,
-2.09384770e-01,  3.14304048e+00, -3.99474853e+00,
 1.39219377e+00, -4.94695788e-01,  2.12958096e+00,
 1.17635294e+00,  1.79785099e+00,  9.51404486e-01,
 1.12969253e+00,  2.80863698e+00, -1.52506018e+00,
-1.58648819e+00,  2.05001643e+00, -3.83293834e+00,
-1.36735580e+00,  6.98431787e-01, -3.16859183e+00,
-3.98257585e+00,  2.63888509e+00,  9.91615036e-01,
 2.06174536e+00, -3.81842540e+00,  1.46446930e+00,
-3.53779659e-01,  6.11512524e-01, -6.16624572e+00,
-5.39877323e+00, -1.75584534e-01,  2.40805112e+00,
-6.40919025e+00, -1.64363001e+00, -2.25703639e+00,
-1.13439670e+00, -1.26970689e+00, -2.40458249e+00,
-8.94165629e-01,  8.57381327e-01,  1.27983178e+00,
-1.83736442e-01, -4.17805724e+00, -2.64792546e+00,
 1.46835619e-01, -4.75493967e-01, -4.39572391e+00,
-9.83180519e-01,  1.27953061e+00, -1.43020978e+00,
 1.15131408e+00,  3.06586853e-01,  1.62443130e+00,
 3.13461752e+00,  2.92750329e+00,  3.25428675e-01,
-2.51307183e+00, -2.13761930e+00,  8.72241802e-01,
-3.43754473e+00, -6.68784769e+00,  7.75319021e-01,
 4.66655310e-02, -2.79095682e+00,  6.56126358e-01,
-1.39568625e+00, -7.29429376e-01,  1.60108955e+00,
 2.69594987e+00,  1.97327235e+00, -1.59652280e+00,
 5.06103265e-01, -1.65261452e+00,  9.39032286e-01,
-4.78566161e+00,  6.10730283e-02,  2.87984077e-01,
 3.06698701e-01,  2.14586509e-01,  1.34031718e-01,
 8.62690567e-01,  1.49532037e+00,  2.57932383e-01,
-4.87667986e+00, -6.26171499e-01, -8.52507610e-01,
-7.17776353e+00, -2.47546758e+00, -5.94806885e-01,
-1.66075917e+00, -9.77456437e-01, -2.14368345e+00,
-5.15039211e+00, -1.81661135e+00, -5.08224119e-01,
 3.53663249e+00, -1.75197130e+00,  3.36354580e+00,
-1.06412878e+00, -7.79606627e-01, -4.38206036e+00,
 2.59900376e+00,  3.26577832e+00,  1.05443333e+00,
-1.79644186e-01,  3.15609336e-01, -8.83364336e-01,
 2.35064128e+00, -3.60162756e+00,  5.69958938e-01,
 2.58465574e+00, -1.74231362e+00, -2.82255360e+00,
-1.08369264e+00, -7.02644632e-01, -2.03833294e+00,
-4.78980331e+00, -3.82889464e+00,  1.20922865e+00,
 2.02067219e+00,  2.09076602e+00,  9.55413531e-01,
 1.85483699e+00, -4.37577380e+00,  7.75701790e-01,
 6.04614668e-01,  2.14557719e+00, -1.88989167e+00,
 1.90484417e+00,  4.50518886e+00, -1.46524723e+00,
-5.65091007e+00, -4.04029735e+00, -3.08162970e+00,
-2.42922642e-01,  2.80397348e+00, -2.96676289e+00,
 2.46551938e+00,  7.30798210e-01,  1.73943723e+00,
 4.50767451e+00, -1.68124746e+00,  3.61942598e-01,
-5.74822970e+00,  9.25188837e-01,  1.32239736e+00,
```

```
-1.57969870e-01, -1.67585798e+00, -3.43428448e+00,
-1.03047494e+00, -1.65810725e+00, -6.93228531e+00,
 3.40542198e+00,  2.47794046e+00,  3.26028671e+00,
 2.12089613e+00, -5.08258588e+00, -2.54428584e+00,
 2.03384834e+00,  2.25898150e+00, -4.27563284e+00,
 6.48072824e-01,  4.40937913e+00,  2.85057372e-01,
 2.29314615e+00, -8.18300289e-01, -3.61133409e-01,
 1.14931122e+00,  2.64609725e+00, -2.55335845e-01,
 2.03456270e+00,  1.39248002e+00,  1.00440086e-01,
 7.63463001e-01, -7.88804150e-01, -1.79764028e+00,
-2.55298336e+00,  9.87248178e-01, -2.23372968e+00,
-1.14672909e+00, -4.64674649e-01, -6.42666966e-01,
 2.52582289e-01, -1.95229896e+00,  7.15641707e-01,
 2.49339745e+00, -5.64559475e+00,  2.11358971e-01,
 1.56984038e+00,  2.60836650e+00,  8.01020642e-01,
 2.47139681e+00, -1.59981058e+00, -1.61192176e+00,
-2.69832869e+00,  3.32845760e+00, -1.99389495e-01,
 7.32560596e-01, -7.91932735e+00,  4.69508066e-01]])
```

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?

To find cumulative values of eigen values we first find explained variance for each PC which is eigen value of each PC divided by the sum of all the eigen values of all PC's.

Variance Explained:

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

Now the cumulative sum of explained variance is as follows.

Cumulative Variance Explained:

array ([0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154,

0.81657854, 0.85216726, 0.88670347, 0.91787581, 0.94162773,

0.96004199, 0.9730024 , 0.98285994, 0.99131837, 0.99648962,

0.99864716, 1.])

Cumulative of Variance Explained represents how much data/information is captured by each PC. Here for our output we can see that by considering only the first 6 PC's are able to capture 81% of information. So Cumulative of Variance Explained helps in choosing the optimum number of PC's which represents the maximum amount of data.

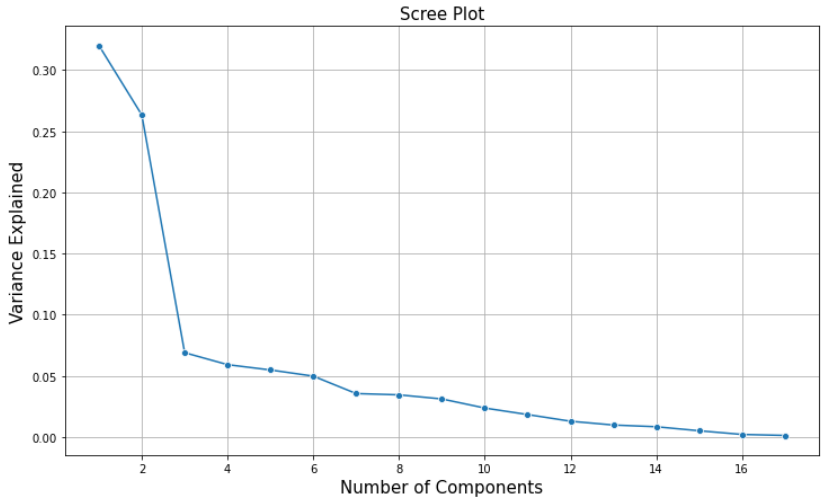

Fig. 6 Scree Plot

We can represent Cumulative of Variance Explained against its PC's using a scree plot as well.

Insight - For our dataset optimum number of PC we have chosen is 6.

Heatmap of all our selected PC shows that none of the Pc's are correlated to each other.

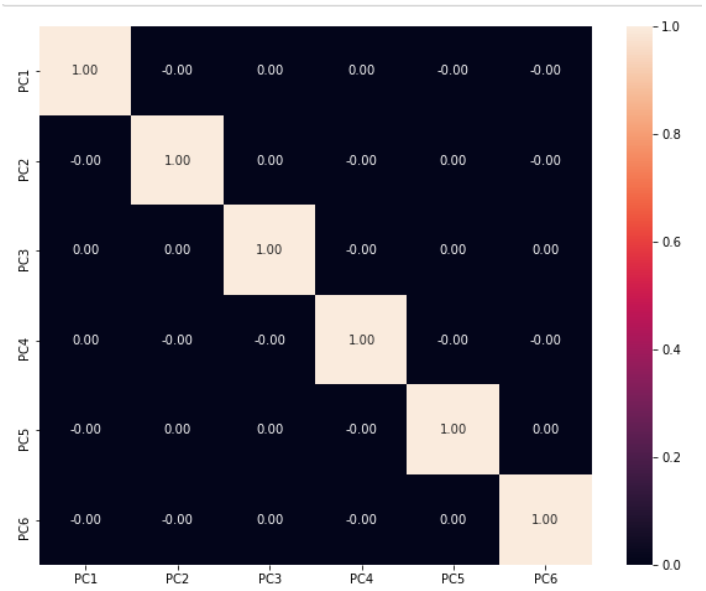

Fig. 7 Heatmap

Eigen Vectors are dimensions in mathematical space where the maximum spread of data is captured. All eigen vectors are orthogonal to each other to become uncorrelated and reduce multicollinearity.

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

After preforming PCA on the dataset containing 18 variables we selected only 6 PC which can explain most of information in our dataset (capturing 81.65 % information of original dataset). Hence we have significantly reduced the number of variables required for further analysis of data using machine learning algorithms and pattern recognition. Since the number of variables if less now our result from machine learning algorithms will be more accurate.

We now identify which features have maximum loading across the components.

We will first plot the component loading on a heatmap.

For each feature, we find the maximum loading value across the components and mark the same with help of rectangular box.

Features marked with rectangular red box are the one having maximum loading on the respective component. We consider these marked features to decide the context that the component represents

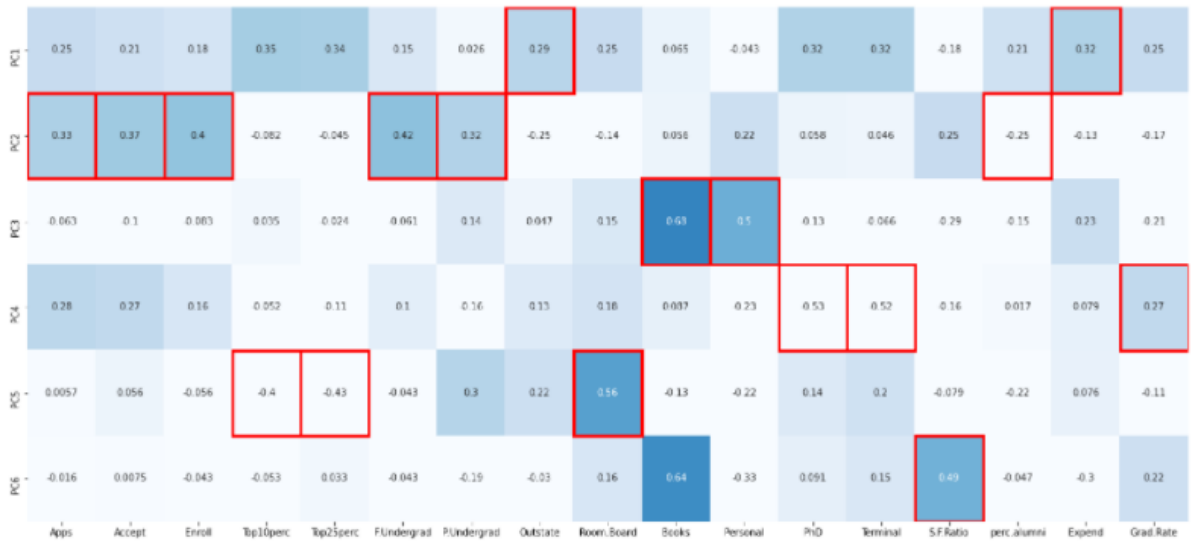

Fig. 8 Heatmap

Insights:

PC1 explains most for outstate and Expend variables

PC2 explains most for Apps, Accept, Enroll, F.Undergrad, P.Undergrad

PC3 explains most for Books and Personal

PC4 explains most for Grad.Rate

PC5 explains most for Room.Board

PC6 explains most for Books, S.F Ratio, Grad.Rate