ADVANCED STATISTICS PROJECT

Amrita Jena

PGPDSBA Online May B 2021

Table of contents-

Problem 1A:

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

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

1. **State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.**
2. **Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.**
3. **Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.**
4. **If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result.** (Non-Graded)

Problem 1B:

1. **What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the 'pointplot' function from the 'seaborn' function]**
2. **Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?**
3. **Explain the business implications of performing ANOVA for this particular case study.**

Problem 2:

The dataset Education - Post 12th Standard.csv contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx.

1. **Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?**
2. **Is scaling necessary for PCA in this case? Give justification and perform scaling.**
3. **Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].**
4. **Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so]**
5. **Extract the eigenvalues and eigenvectors.[Using Sklearn PCA Print Both]**
6. **Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features**
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]**
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?**
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]**

EXECUTIVE SUMMARY: Problem 1-

1- 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. In this problem statement we will analyse the different categories of **Education** and **Occupation** and its impact on **Salary**.

INTRODUCTION:

The purpose of this whole exercise is to explore the dataset. Do the ANOVA analysis to interpret the result and also Utilize the knowledge gained by it in real world work environment as well as

DATA DESCRIPTION:

	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
5	Doctorate	Sales	219420
6	Doctorate	Sales	237920
7	Doctorate	Sales	160540
8	Doctorate	Sales	180934
9	Doctorate	Prof-specialty	248156

Problem 1A:

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

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

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

Ans- **Null and Alternate Hypothesis for Education:**

H0-The means of 'Salary' with respect to each education is equal.

H1- At least one of means of 'Salary' with respect to each education is unequal.

Null and Alternate Hypothesis for Occupation:

H0-The means of 'Salary' with respect to each Occupation is equal.

H1- At least one of means of 'Salary' with respect to each Occupation is unequal.

2.Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

Ans -

	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

Since in the result we could see that the p value is less than alpha(0.05) we reject the null hypothesis.

Since in the result we could see that the p value is less than alpha (0.05) we reject the null hypothesis.

3.Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

Ans-

	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

Since in the result we could observe that the p value is greater than alpha(0.05) we fail to reject the null hypothesis.

Since in the result we could observe that the p value is greater than alpha (0.05) we fail to reject the null hypothesis.

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

Ans -

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

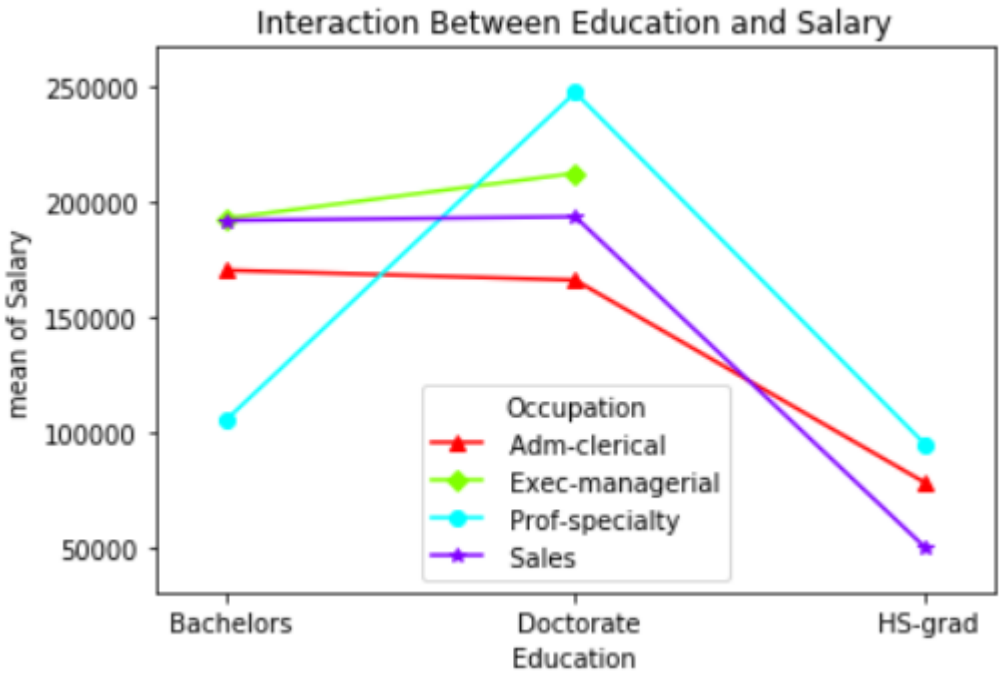
Here we use Tukey's Test, is the most commonly used post hoc tests which allows us to make pairwise comparisons between the means of each group while controlling for the family-wise error rate.

From here it could be inferred that means of Doctorate, Bachelors, HS-grad means are significantly different.

Problem 1A:

1.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 plot. [hint: use the 'point plot' function from the 'seaborn' function]

Ans –



From the above interaction plot, a very good interaction between Doctorate and bachelors in the occupation of Adm-clerical and Sales could be observed.

An analogous interaction between Bachelors and HS-grad in the occupation of Prof-specialty is seen.

No interaction between Doctorate and HS-grad in any of the occupation could be observed.

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

Ans –

Null and Alternate Hypothesis:

H0-The means of Salary with respect to each Education category and Occupation is equal

H1- At least one of means of Salary with respect to each Education category and Occupation is unequal.

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	31.257677	1.981539e-08
C(Occupation)	3.0	5.519946e+09	1.839982e+09	1.120080	3.545825e-01
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

From the result it could be interpreted that-

For Education category p value is less than alpha (0.05) so we reject null hypothesis establishing that Education has a significant impact on mean salary.

For Occupation category p value is greater than alpha (0.05) so we fail to reject null hypothesis establishing that Occupation does not have a significant impact on mean salary.

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

Ans –

From the above ANOVA analysis, we could state that:

An employee or a graduate's salary is significantly dependent on their level of education as compared to their occupation.

From the statistical conclusion about the interaction effect of education and occupation on salary we could say that despite occupation's lesser significance, there is some level of impact of job role on salary

It is obvious that on an average a Doctorate should probably earn higher salary than Bachelors and HS-grads. However, it is also true that being a Doctorate may not necessarily mean Significantly higher salary.

Hence, there should be more comprehensive approach towards setting of salary ranges.

EXECUTIVE SUMMARY: Problem 2-

The dataset Education - Post 12th Standard.csv contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx. In this problem statement we will apply EDA(Exploratory Data Analysis) and PCA (Principle Component Analysis)

DATA DESCRIPTION:

1) Names: Names of various university and colleges
2) Apps: Number of applications received
3) Accept: Number of applications accepted
4) Enroll: Number of new students enrolled
5) Top10perc: Percentage of new students from top 10% of Higher Secondary class
6) Top25perc: Percentage of new students from top 25% of Higher Secondary class
7) F.Undergrad: Number of full-time undergraduate students
8) P.Undergrad: Number of part-time undergraduate students
9) Outstate: Number of students for whom the particular college or university is Out-of-state tuition
10) Room.Board: Cost of Room and board
11) Books: Estimated book costs for a student
12) Personal: Estimated personal spending for a student
13) PhD: Percentage of faculties with Ph.D.'s
14) Terminal: Percentage of faculties with terminal degree
15) S.F.Ratio: Student/faculty ratio
16) perc.alumni: Percentage of alumni who donate
17) Expend: The Instructional expenditure per student
18) Grad.Rate: Graduation rate

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

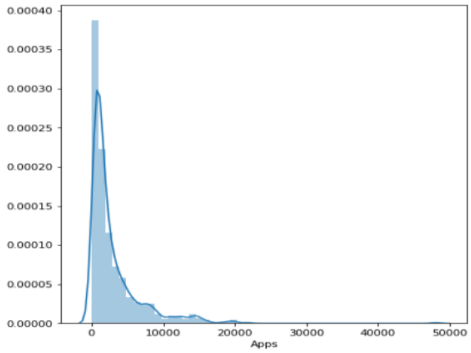
Ans-

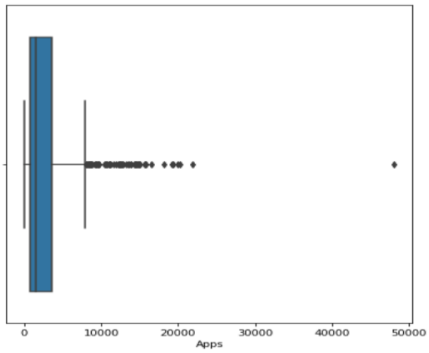
After performing univariate analysis it is concluded that:

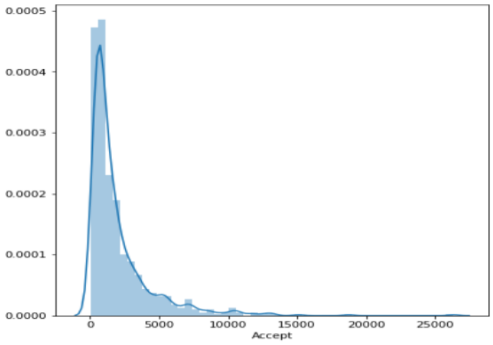
The Education data set has 777 rows and 18 columns
One categorical column (Names) and 17 numerical columns.
All the values are on int64 type except 'Names' which is of object datatype and 'S.F.Ratio' which is of float64 datatype.
We don't have any null/missing values

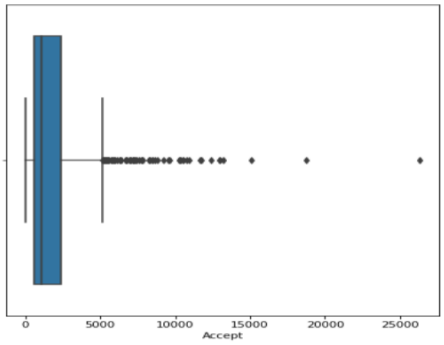
There are no duplicated values present in the dataset.

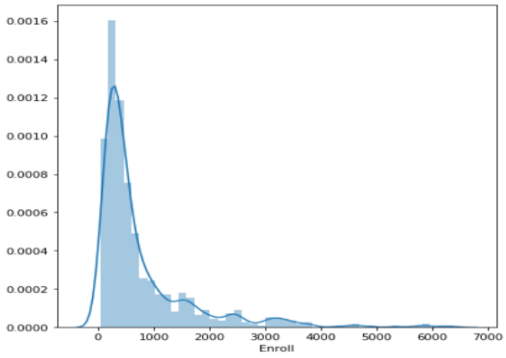
Then we used distplot and boxplot to check whether the data is normally distributed or not and if outliers are present in the dataset.

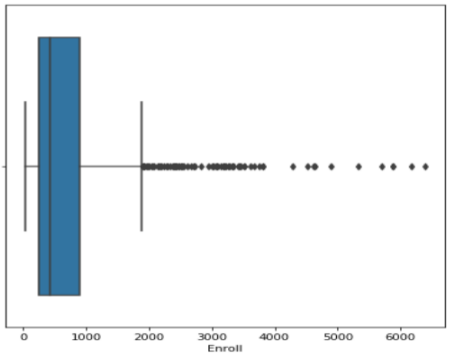


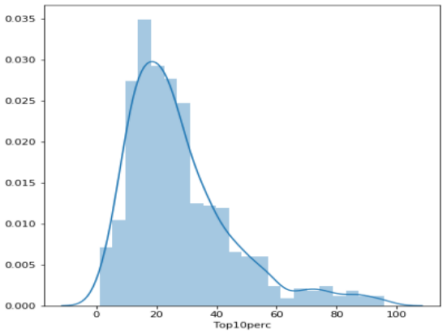


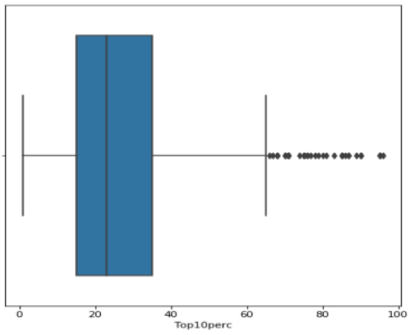


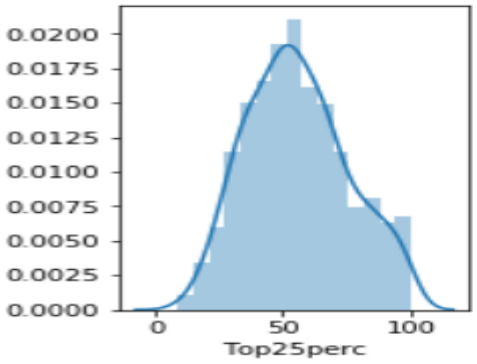


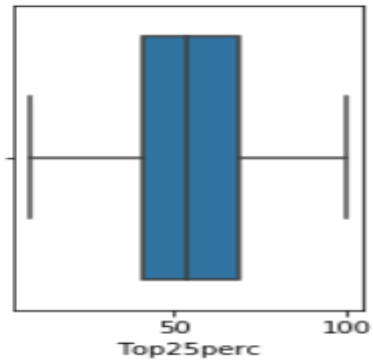


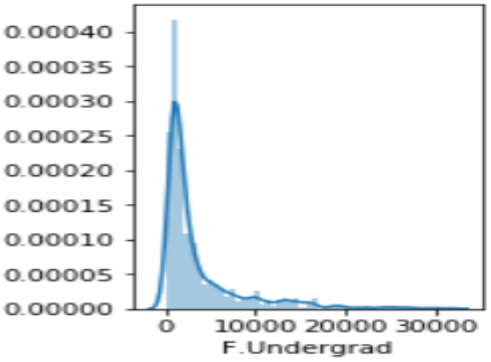


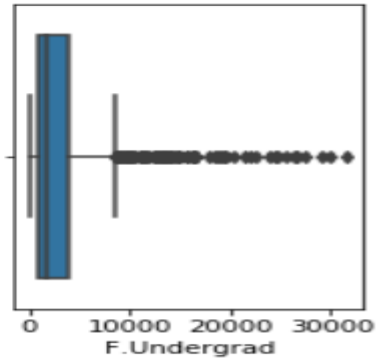

0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
Top10perc

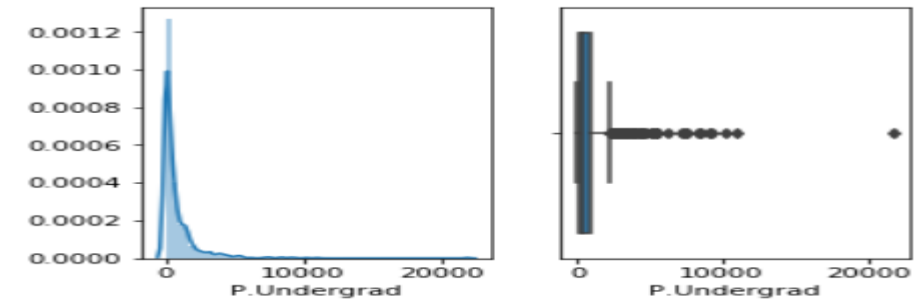

0
20
40
60
80
100
Top10perc

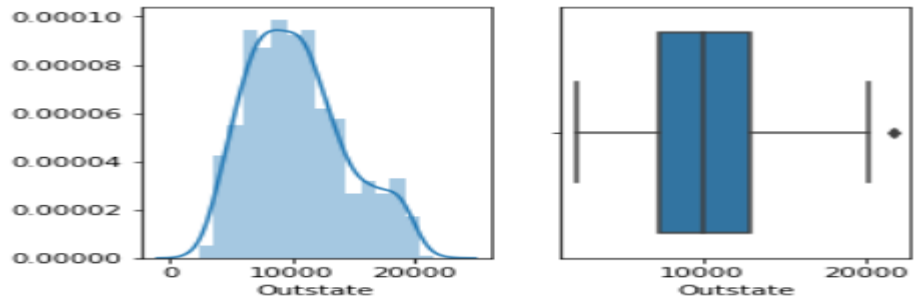

0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
50
100
Top25perc

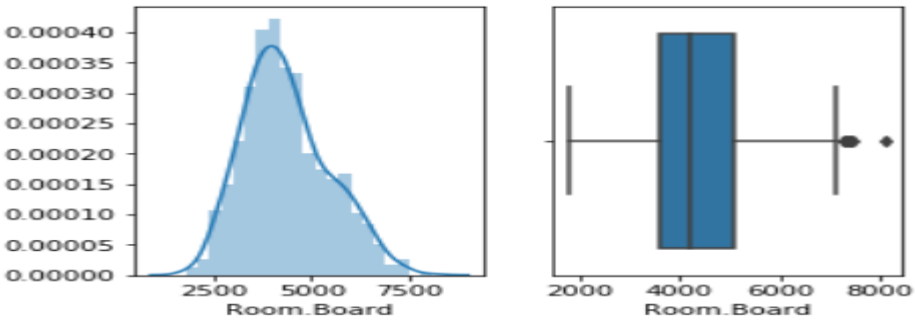

50
100
Top25perc

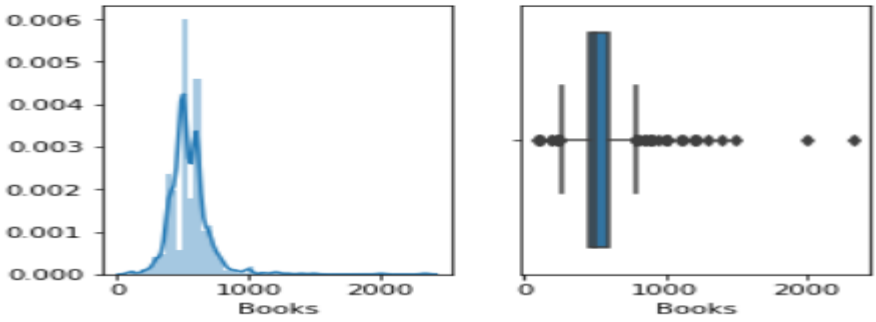

0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
10000
20000
30000
F.Undergrad

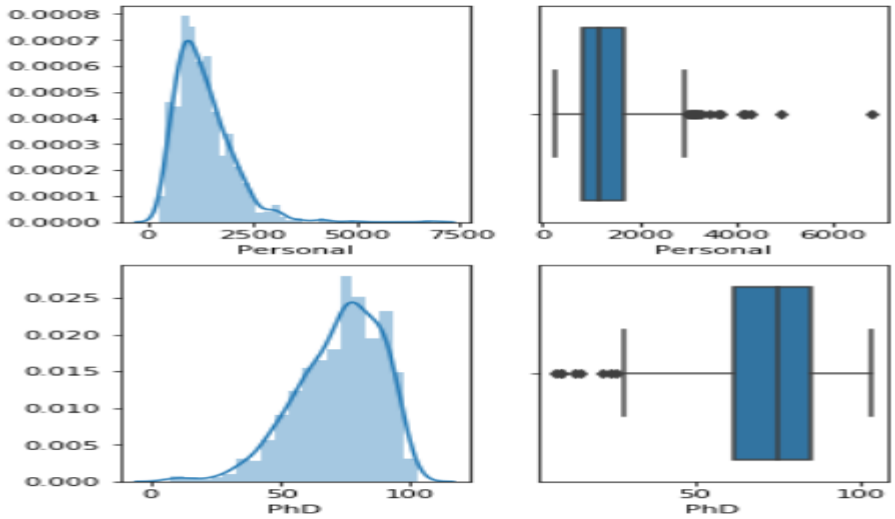

0
10000
20000
30000
F.Undergrad

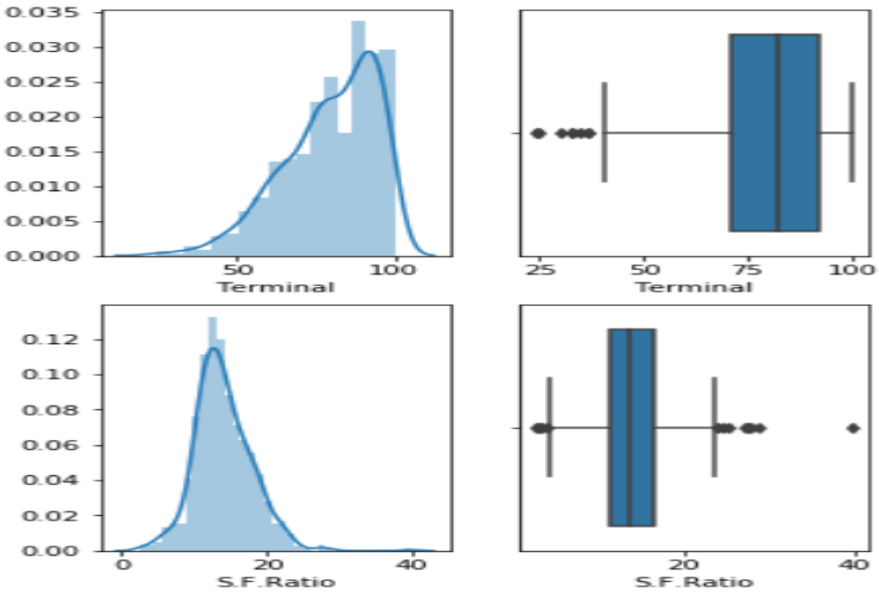

0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
10000
20000
P.Undergrad
0
10000
20000
P.Undergrad

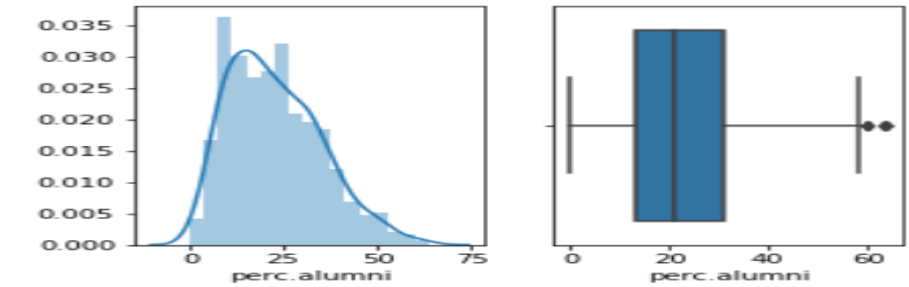

0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
10000
20000
Outstate
10000
20000
Outstate

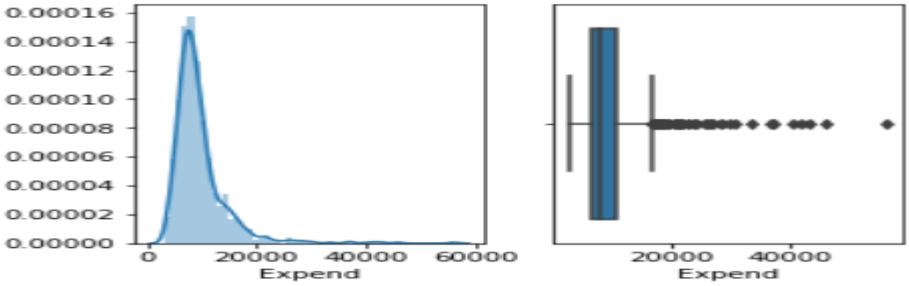

0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
2500
5000
7500
Room.Board
2000
4000
6000
8000
Room.Board

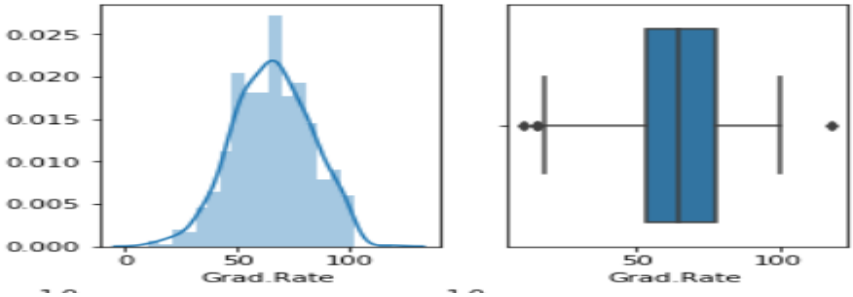

0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
1000
2000
Books
0
1000
2000
Books


0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
2500
5000
7500
Personal
0
2000
4000
6000
Personal
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
PhD
50
100
PhD


0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
50
100
Terminal
25
50
75
100
Terminal
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
S.F.Ratio
20
40
S.F.Ratio


0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
25
50
75
perc.alumni
0
20
40
60
perc.alumni


0.00016
0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
20000
40000
60000
Expend
20000
40000
Expend


0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
Grad.Rate
50
100
Grad.Rate

```
Apps           3.723750
Accept         3.417727
Enroll         2.690465
Top10perc      1.413217
Top25perc      0.259340
F.Undergrad    2.610458
P.Undergrad    5.692353
Outstate       0.509278
Room.Board     0.477356
Books          3.485025
Personal       1.742497
PhD           -0.768170
Terminal      -0.816542
S.F.Ratio      0.667435
perc.alumni    0.606891
Expend         3.459322
Grad.Rate     -0.113777
dtype: float64
```

From above distplots it is clear that the data is highly skewed.

we use skewness to understand which data set is normally distributed and which is not. If the skewness =0, It is said to be normally distributed, if it is >0 it is left skewed and if it <O it is skewed towards right.

Skewness calculated suggests that we have right skewed, left skewed and symmetrical data

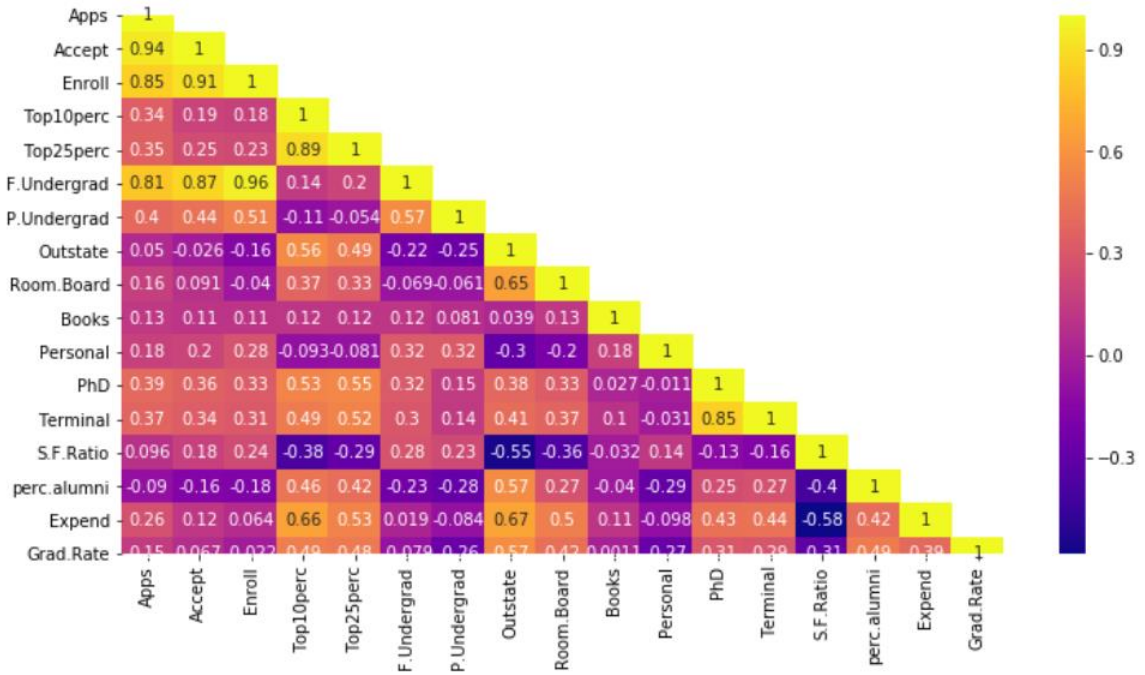
3 - Symmetrical data about its axis

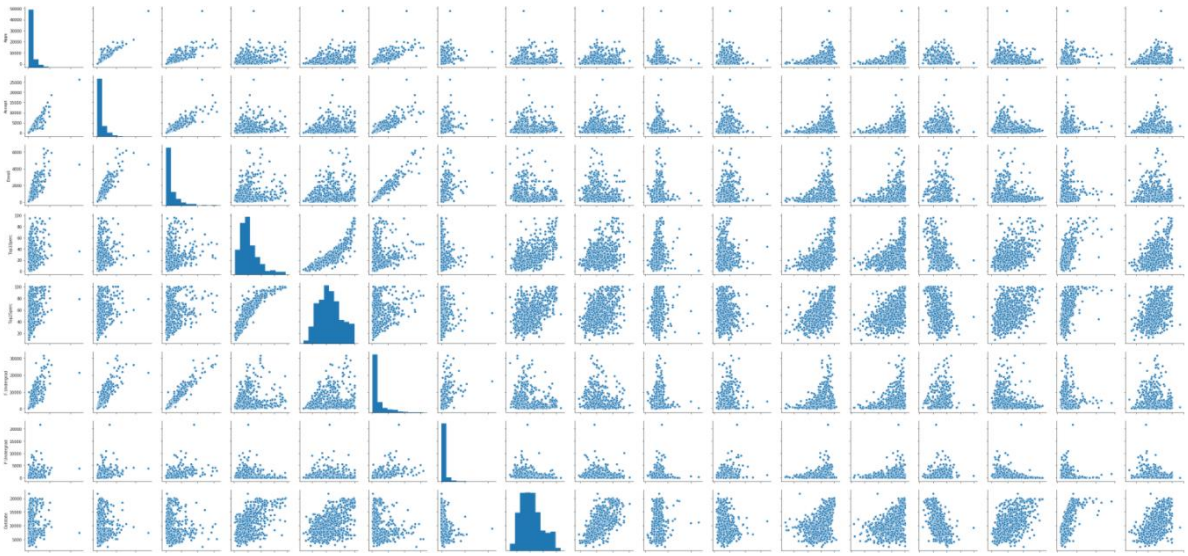
12- Right skewed

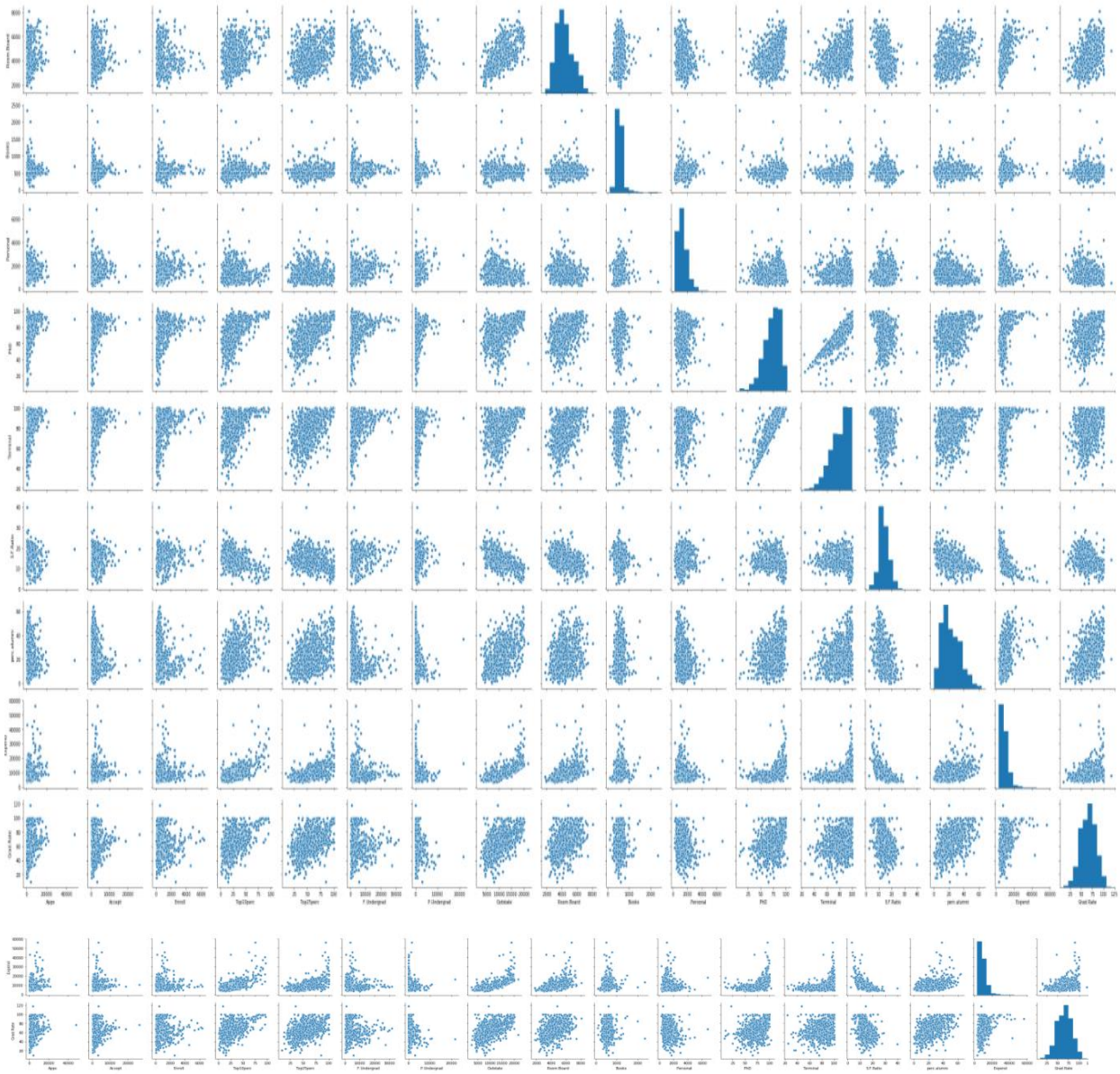
2 - Left skewed

it is also observed from the above boxplots that outliers present in every column value in the dataset except in 'Top25perc'.

Multivariate Analysis


Apps
Accept
Enroll
Top10perc
Top25perc
F.Undergrad
P.Undergrad
Outstate
Room.Board
Books
Personal
PhD
Terminal
S.F.Ratio
perc.alumni
Expend
Grad.Rate
0.9
0.6
0.3
0.0
−0.3





The heatmap shows strong correlation (>.85) between;

- Accept/Apps
- Accept/Enroll
- Apps/Enrol
- Top25perc/Top10perc
- Undergrad/Apps/Accept/Enroll
- Terminal/PhD

Similarly from the pair plot we observe the same by establishing a linear relationship between the above mentioned strongly correlated variables and a highly scattered plot in all other scenarios.

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

Yes, we need to scale the data here. Feature scaling through standardization (or Z-score normalization) is an important pre-processing step as it involves rescaling the features of the data, such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one hence normalizing a data within a particular range. It also helps in speeding up the calculations in an algorithm.

	0	1	2	3	4	5	6	7	8	9
Apps	-0.346882	-0.210884	-0.406866	-0.668261	-0.726176	-0.624307	-0.684808	-0.285088	-0.507700	-0.625600
Accept	-0.321205	-0.038703	-0.376318	-0.681682	-0.764555	-0.628611	-0.685356	-0.121984	-0.481644	-0.620854
Enroll	-0.063509	-0.288584	-0.478121	-0.692427	-0.780735	-0.669812	-0.729043	-0.313353	-0.595505	-0.654735
Top10perc	-0.258583	-0.655656	-0.315307	1.840231	-0.655656	0.592287	-0.598931	0.535563	0.138490	-0.372032
Top25perc	-0.191827	-1.353911	-0.292878	1.677612	-0.596031	0.313426	-0.545505	0.616579	0.363952	-0.596031
F.Undergrad	-0.168116	-0.209788	-0.549565	-0.658079	-0.711924	-0.623421	-0.677472	-0.434450	-0.562562	-0.598459
P.Undergrad	-0.209207	0.244307	-0.497090	-0.520752	0.009005	-0.535212	-0.410988	-0.541127	-0.361036	-0.510893
Outstate	-0.746356	0.457496	0.201305	0.626633	-0.716508	0.760947	0.708713	0.852479	1.282036	0.006798
Room.Board	-0.964905	1.909208	-0.554317	0.996791	-0.216723	-0.932970	1.243144	0.427443	0.038754	-0.891911
Books	-0.602312	1.215880	-0.905344	-0.602312	1.518912	-0.299280	-0.299280	-0.602312	-1.511408	0.670422
Personal	1.270045	0.235515	-0.259582	-0.688173	0.235515	-0.983753	0.235515	-0.725120	-1.242385	0.678885
PhD	-0.163028	-2.675646	-1.204845	1.185206	0.204672	-0.346878	1.062639	1.001356	0.388522	-2.001529
Terminal	-0.115729	-3.378176	-0.931341	1.175657	-0.523535	-0.455567	0.903786	1.379560	0.292077	-2.630532
S.F.Ratio	1.013776	-0.477704	-0.300749	-1.615274	-0.553542	-1.185526	-0.654660	-0.098515	-0.705218	-0.654660
perc.alumni	-0.867574	-0.544572	0.585935	1.151188	-1.675079	-0.948325	0.262933	1.151188	0.020681	-0.625323
Expend	-0.501910	0.166110	-0.177290	1.792851	0.241803	0.012806	-0.153145	0.350074	0.380160	-0.128233
Grad.Rate	-0.318252	-0.551262	-0.667767	-0.376504	-2.939613	-0.609514	-0.143495	0.439030	0.846798	-0.784272

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

```
array([[ 1.00128866,  0.94466636,  0.84791332,  0.33927032,  0.35209304,
         0.81554018,  0.3987775 ,  0.05022367,  0.16515151,  0.13272942,
         0.17896117,  0.39120081,  0.36996762,  0.09575627, -0.09034216,
         0.2599265 ,  0.14694372],
       [ 0.94466636,  1.00128866,  0.91281145,  0.19269493,  0.24779465,
         0.87534985,  0.44183938, -0.02578774,  0.09101577,  0.11367165,
         0.20124767,  0.35621633,  0.3380184 ,  0.17645611, -0.16019604,
         0.12487773,  0.06739929],
       [ 0.84791332,  0.91281145,  1.00128866,  0.18152715,  0.2270373 ,
         0.96588274,  0.51372977, -0.1556777 , -0.04028353,  0.11285614,
         0.28129148,  0.33189629,  0.30867133,  0.23757707, -0.18102711,
         0.06425192, -0.02236983],
       [ 0.33927032,  0.19269493,  0.18152715,  1.00128866,  0.89314445,
         0.1414708 , -0.10549205,  0.5630552 ,  0.37195909,  0.1190116 ,
        -0.09343665,  0.53251337,  0.49176793, -0.38537048,  0.45607223,
         0.6617651 ,  0.49562711],
       [ 0.35209304,  0.24779465,  0.2270373 ,  0.89314445,  1.00128866,
         0.19970167, -0.05364569,  0.49002449,  0.33191707,  0.115676  ,
        -0.08091441,  0.54656564,  0.52542506, -0.29500852,  0.41840277,
         0.52812713,  0.47789622],
       [ 0.81554018,  0.87534985,  0.96588274,  0.1414708 ,  0.19970167,
         1.00128866,  0.57124738, -0.21602002, -0.06897917,  0.11569867,
         0.31760831,  0.3187472 ,  0.30040557,  0.28006379, -0.22975792,
         0.01867565, -0.07887464],
       [ 0.3987775 ,  0.44183938,  0.51372977, -0.10549205, -0.05364569,
         0.57124738,  1.00128866, -0.25383901, -0.06140453,  0.08130416,
         0.32029384,  0.14930637,  0.14208644,  0.23283016, -0.28115421,
        -0.08367612, -0.25733218],
```

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

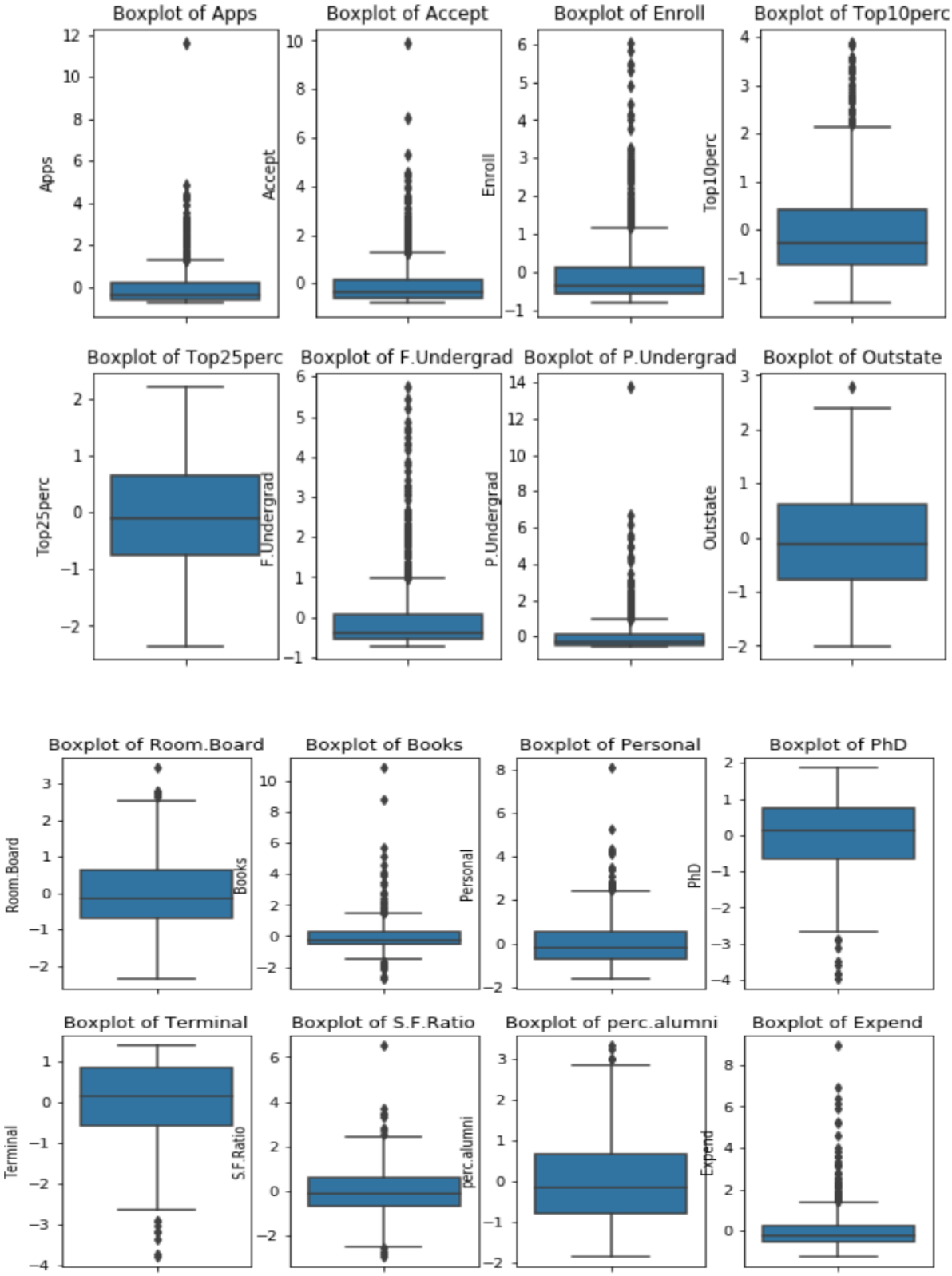
PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0.390697	0.369491	0.095633	-0.090226	0.259592	0.146755
0.355758	0.337583	0.176229	-0.159990	0.124717	0.067313
0.331469	0.308274	0.237271	-0.180794	0.064169	-0.022341
0.531828	0.491135	-0.384875	0.455485	0.660913	0.494989
0.545862	0.524749	-0.294629	0.417864	0.527447	0.477281
0.318337	0.300019	0.279703	-0.229462	0.018652	-0.078773
0.149114	0.141904	0.232531	-0.280792	-0.083568	-0.257001
0.382982	0.407983	-0.554821	0.566262	0.672779	0.571290
0.329202	0.374540	-0.362628	0.272363	0.501739	0.424942
0.026906	0.099955	-0.031929	-0.040208	0.112409	0.001061
0.010936	-0.030613	0.136345	-0.285968	-0.097892	-0.269344
1.000000	0.849587	-0.130530	0.249009	0.432762	0.305038
0.849587	1.000000	-0.160104	0.267130	0.438799	0.289527
0.130530	-0.160104	1.000000	-0.402929	-0.583832	-0.306710
0.249009	0.267130	-0.402929	1.000000	0.417712	0.490898
0.432762	0.438799	-0.583832	0.417712	1.000000	0.390343
0.305038	0.289527	-0.306710	0.490898	0.390343	1.000000

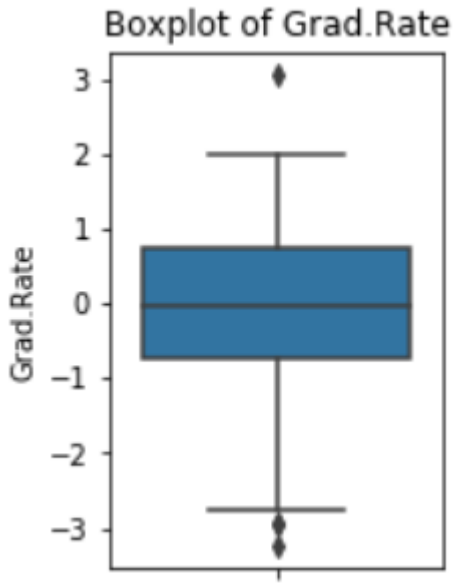
Correlation is a scaled version of covariance the two parameters always have the same sign positive negative or 0.

Positive sign indicates positively corelated variables, negative sign indicates negatively corelated variables and 0 indicates uncorrelated.

Correlation measures both strength and direction of the linear relationship between two variables whereas covariance indicates the direction of the linear relationship of the variables

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





Initially while doing univariate analysis, we have checked for outliers and except Top25perc presence of outliers were found in all other 16 variables.

After scaling again, the data was analysed for outliers we could observe that there was not much difference as presence of outliers were still found in the data.

Outliers could be removed by treating them. (But as it is not specifically asked in the question, we will skip the step of treating the outliers)

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

```
Eigen Values
%s [5.45052162 4.48360686 1.17466761 1.00820573 0.93423123 0.84849117
0.6057878  0.58787222 0.53061262 0.4043029  0.02302787 0.03672545
0.31344588 0.08802464 0.1439785  0.16779415 0.22061096]
```

```
Eigen Vectors
%s [[-2.48765602e-01  3.31598227e-01  6.30921033e-02 -2.81310530e-01
   5.74140964e-03  1.62374420e-02  4.24863486e-02  1.03090398e-01
   9.02270802e-02 -5.25098025e-02  3.58970400e-01 -4.59139498e-01
   4.30462074e-02 -1.33405806e-01  8.06328039e-02 -5.95830975e-01
   2.40709086e-02]
[-2.07601502e-01  3.72116750e-01  1.01249056e-01 -2.67817346e-01
   5.57860920e-02 -7.53468452e-03  1.29497196e-02  5.62709623e-02
   1.77864814e-01 -4.11400844e-02 -5.43427250e-01  5.18568789e-01
  -5.84055850e-02  1.45497511e-01  3.34674281e-02 -2.92642398e-01
  -1.45102446e-01]
[-1.76303592e-01  4.03724252e-01  8.29855709e-02 -1.61826771e-01
  -5.56936353e-02  4.25579803e-02  2.76928937e-02 -5.86623552e-02
   1.28560713e-01 -3.44879147e-02  6.09651110e-01  4.04318439e-01
  -6.93988831e-02 -2.95896092e-02 -8.56967180e-02  4.44638207e-01
   1.11431545e-02]
[-3.54273947e-01 -8.24118211e-02 -3.50555339e-02  5.15472524e-02
  -3.95434345e-01  5.26927980e-02  1.61332069e-01  1.22678028e-01
  -3.41099863e-01 -6.40257785e-02 -1.44986329e-01  1.48738723e-01
  -8.10481404e-03 -6.97722522e-01 -1.07828189e-01 -1.02303616e-03
   3.85543001e-02]
[-3.44001279e-01 -4.47786551e-02  2.41479376e-02  1.09766541e-01
  -4.26533594e-01 -3.30915896e-02  1.18485556e-01  1.02491967e-01
  -4.03711989e-01 -1.45492289e-02  8.03478445e-02 -5.18683400e-02
  -2.73128469e-01  6.17274818e-01  1.51742110e-01 -2.18838802e-02
  -8.93515563e-02]
```

```
[-1.54640962e-01  4.17673774e-01  6.13929764e-02 -1.00412335e-01
 -4.34543659e-02  4.34542349e-02  2.50763629e-02 -7.88896442e-02
  5.94419181e-02 -2.08471834e-02 -4.14705279e-01 -5.60363054e-01
 -8.11578181e-02 -9.91640992e-03 -5.63728817e-02  5.23622267e-01
  5.61767721e-02]
[-2.64425045e-02  3.15087830e-01 -1.39681716e-01  1.58558487e-01
  3.02385408e-01  1.91198583e-01 -6.10423460e-02 -5.70783816e-01
 -5.60672902e-01  2.23105808e-01  9.01788964e-03  5.27313042e-02
  1.00693324e-01 -2.09515982e-02  1.92857500e-02 -1.25997650e-01
 -6.35360730e-02]
[-2.94736419e-01 -2.49643522e-01 -4.65988731e-02 -1.31291364e-01
  2.22532003e-01  3.00003910e-02 -1.08528966e-01 -9.84599754e-03
  4.57332880e-03 -1.86675363e-01  5.08995918e-02 -1.01594830e-01
  1.43220673e-01 -3.83544794e-02 -3.40115407e-02  1.41856014e-01
 -8.23443779e-01]
[-2.49030449e-01 -1.37808883e-01 -1.48967389e-01 -1.84995991e-01
  5.60919470e-01 -1.62755446e-01 -2.09744235e-01  2.21453442e-01
 -2.75022548e-01 -2.98324237e-01  1.14639620e-03  2.59293381e-02
 -3.59321731e-01 -3.40197083e-03 -5.84289756e-02  6.97485854e-02
  3.54559731e-01]
[-6.47575181e-02  5.63418434e-02 -6.77411649e-01 -8.70892205e-02
 -1.27288825e-01 -6.41054950e-01  1.49692034e-01 -2.13293009e-01
  1.33663353e-01  8.20292186e-02  7.72631963e-04 -2.88282896e-03
  3.19400370e-02  9.43887925e-03 -6.68494643e-02 -1.14379958e-02
 -2.81593679e-02]
[ 4.25285386e-02  2.19929218e-01 -4.99721120e-01  2.30710568e-01
 -2.22311021e-01  3.31398003e-01 -6.33790064e-01  2.32660840e-01
  9.44688900e-02 -1.36027616e-01 -1.11433396e-03  1.28904022e-02
 -1.85784733e-02  3.09001353e-03  2.75286207e-02 -3.94547417e-02
 -3.92640266e-02]
```

```
[-3.18312875e-01  5.83113174e-02  1.27028371e-01  5.34724832e-01
  1.40166326e-01 -9.12555212e-02  1.09641298e-03  7.70400002e-02
  1.85181525e-01  1.23452200e-01  1.38133366e-02 -2.98075465e-02
  4.03723253e-02  1.12055599e-01 -6.91126145e-01 -1.27696382e-01
  2.32224316e-02]
[-3.17056016e-01  4.64294477e-02  6.60375454e-02  5.19443019e-01
  2.04719730e-01 -1.54927646e-01  2.84770105e-02  1.21613297e-02
  2.54938198e-01  8.85784627e-02  6.20932749e-03  2.70759809e-02
 -5.89734026e-02 -1.58909651e-01  6.71008607e-01  5.83134662e-02
  1.64850420e-02]
[ 1.76957895e-01  2.46665277e-01  2.89848401e-01  1.61189487e-01
 -7.93882496e-02 -4.87045875e-01 -2.19259358e-01  8.36048735e-02
 -2.74544380e-01 -4.72045249e-01 -2.22215182e-03  2.12476294e-02
  4.45000727e-01  2.08991284e-02  4.13740967e-02  1.77152700e-02
 -1.10262122e-02]
[-2.05082369e-01 -2.46595274e-01  1.46989274e-01 -1.73142230e-02
 -2.16297411e-01  4.73400144e-02 -2.43321156e-01 -6.78523654e-01
  2.55334907e-01 -4.22999706e-01 -1.91869743e-02 -3.33406243e-03
 -1.30727978e-01  8.41789410e-03 -2.71542091e-02 -1.04088088e-01
  1.82660654e-01]
[-3.18908750e-01 -1.31689865e-01 -2.26743985e-01 -7.92734946e-02
  7.59581203e-02  2.98118619e-01  2.26584481e-01  5.41593771e-02
  4.91388809e-02 -1.32286331e-01 -3.53098218e-02  4.38803230e-02
  6.92088870e-01  2.27742017e-01  7.31225166e-02  9.37464497e-02
  3.25982295e-01]
[-2.52315654e-01 -1.69240532e-01  2.08064649e-01 -2.69129066e-01
 -1.09267913e-01 -2.16163313e-01 -5.59943937e-01  5.33553891e-03
 -4.19043052e-02  5.90271067e-01 -1.30710024e-02  5.00844705e-03
  2.19839000e-01  3.39433604e-03  3.64767385e-02  6.91969778e-02
  1.22106697e-01]]
```

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

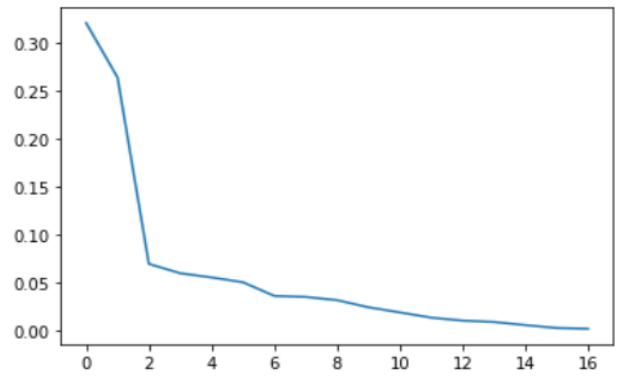
If we don't have any strict constraints, then we should plot the cumulative sum of eigenvalues.

If we divide each value by the total sum of eigenvalues prior to plotting, then our plot will show the fraction of total variance retained vs. number of eigenvalues.

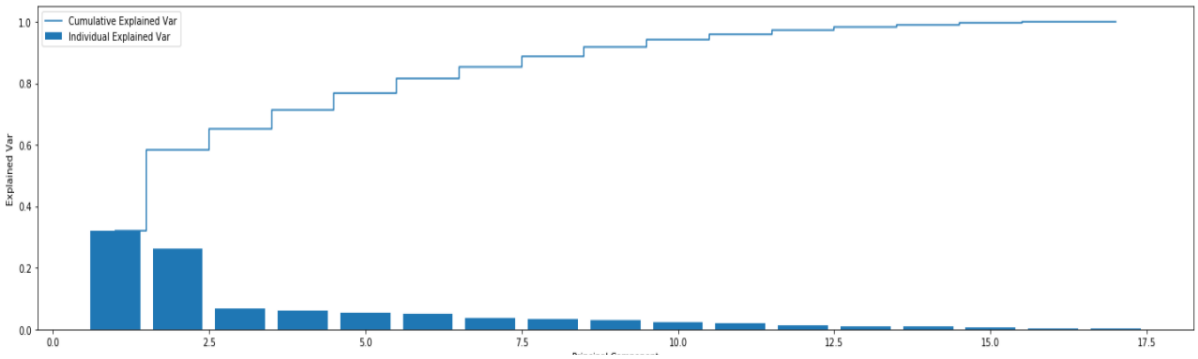
The plot will then provide a good indication of when we will he point of diminishing returns.

```
[0.3202062819886915,
 0.26340214436112463,
 0.06900916554222497,
 0.05922989222926289,
 0.054884051103584804,
 0.0498470095455745,
 0.0355887149174665,
 0.03453621336999264,
 0.031172336798217196,
 0.023751915258937994,
 0.01841426320938688,
 0.012960414001235345,
 0.00985754122800116,
 0.008458423350830023,
 0.005171255833731941,
 0.002157540100727585,
 0.0013528371610095184]
```

```
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.        ])
```



we will perform a scree plot as a scree plot assists in visualizing the relative importance of the factors a sharp drop in the plot signals that subsequent factors are ignorable.



To find PCA components we use PCA commands from SKlearn

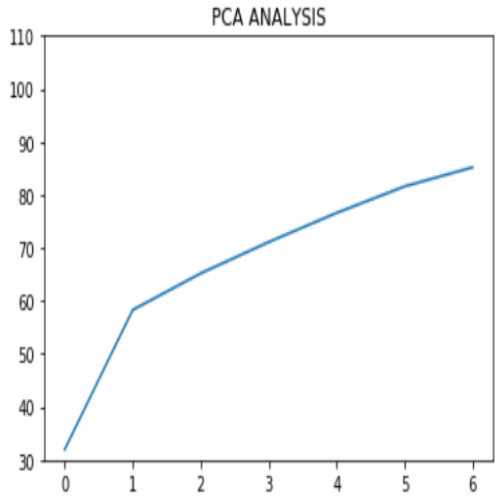
```
array([[-1.59285540e+00,  7.67333510e-01, -1.01073537e-01, ...,
        -7.43975398e-01, -2.98306081e-01,  6.38443468e-01],
       [-2.19240180e+00, -5.78829984e-01,  2.27879812e+00, ...,
         1.05999660e+00, -1.77137309e-01,  2.36753302e-01],
       [-1.43096371e+00, -1.09281889e+00, -4.38092811e-01,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

To Check the explained variance for each PC we perform the below mentioned formula.

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

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

```
[<matplotlib.lines.Line2D at 0x26dcf23c588>]
```



This plot represents all the 7 features of the dataset contributing~ 85.2% of the variance within the dataset

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.012950
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486

```
array([[ 0.2487656 ,  0.2076015 ,  0.17630359,  0.35427395,  0.34400128,
         0.15464096,  0.0264425 ,  0.29473642,  0.24903045,  0.06475752,
        -0.04252854,  0.31831287,  0.31705602, -0.17695789,  0.20508237,
         0.31890875,  0.25231565],
       [ 0.33159823,  0.37211675,  0.40372425, -0.08241182, -0.04477866,
         0.41767377,  0.31508783, -0.24964352, -0.13780888,  0.05634184,
         0.21992922,  0.05831132,  0.04642945,  0.24666528, -0.24659527,
        -0.13168986, -0.16924053],
       [-0.0630921 , -0.10124906, -0.08298557,  0.03505553, -0.02414794,
        -0.06139298,  0.13968172,  0.04659887,  0.14896739,  0.67741165,
         0.49972112, -0.12702837, -0.06603755, -0.2898484 , -0.14698927,
         0.22674398, -0.20806465],
       [ 0.28131053,  0.26781735,  0.16182677, -0.05154725, -0.10976654,
         0.10041234, -0.15855849,  0.13129136,  0.18499599,  0.08708922,
        -0.23071057, -0.53472483, -0.51944302, -0.16118949,  0.01731422,
         0.07927349,  0.26912907],
       [ 0.00574141,  0.05578609, -0.05569364, -0.39543434, -0.42653359,
        -0.04345437,  0.30238541,  0.222532  ,  0.56091947, -0.12728883,
        -0.22231102,  0.14016633,  0.20471973, -0.07938825, -0.21629741,
         0.07595812, -0.10926791],
       [-0.01623744,  0.00753468, -0.04255798, -0.0526928 ,  0.03309159,
        -0.04345423, -0.19119858, -0.03000039,  0.16275545,  0.64105495,
        -0.331398  ,  0.09125552,  0.15492765,  0.48704587, -0.04734001,
        -0.29811862,  0.21616331],
       [-0.04248635, -0.01294972, -0.02769289, -0.16133207, -0.11848556,
        -0.02507636,  0.06104235,  0.10852897,  0.20974423, -0.14969203,
         0.63379006, -0.00109641, -0.02847701,  0.21925936,  0.24332116,
        -0.22658448,  0.55994394]])
```

```
array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,
       0.84849117, 0.6057878 ])
```

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]

Ans-

```
[-2.48765602e-01  3.31598227e-01  6.30921033e-02 -2.81310530e-01
   5.74140964e-03  1.62374420e-02  4.24863486e-02  1.03090398e-01
   9.02270802e-02 -5.25098025e-02  3.58970400e-01 -4.59139498e-01
   4.30462074e-02 -1.33405806e-01  8.06328039e-02 -5.95830975e-01
   2.40709086e-02]
```

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?

Ans-

```
array([32. , 58.3, 65.2, 71.1, 76.6, 81.6, 85.2])
```

The Cumulative % gives the percentage of variance accounted for by the n components. It aids in deciding the number of components by selecting the components which explained the high variance.To decide how many eigenvalues/eigenvectors to keep.

In the above array we see that the first feature explains 32.0 % of the variance within our data set while the second and third one explain 58.3% ,65.2% respectively and so on. If we take on all the 7 features we capture ~ 85.2% of the variance within the dataset.

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]

Ans-Principal component analysis (PCA) is a technique for reducing the dimensionality of large datasets and increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

It uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables.

Here in the dataset there were 18 variable where after applying PCA they were reduced to only 7 components capturing ~ 85.2% of the variance within the dataset. -