Business Report

Sanjay Srinivasan
PGP-DSBA Online
JULY' 21 Batch
Date: 09-10-2021

INDEX

List Of Tables

List Of Figures

Problem - 1

Summary

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

Introduction

The purpose of this exercise is to explore the dataset across the mean difference of two or more independent variable. The exploratory data analysis of the dataset as the salary details across the individual with different educational and occupation. From this three different education level has different occupation level has the salary difference. This assignment helps in exploring the dataset with anova test.

Data Description

1. Education: Individual with three different level of education.
2. Occupation: Individual with four different level of occupation.
3. Salary: Salary details for every individual.

Sample of the dataset:

	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769

Table 1.1 Dataset Sample

Dataset has 3 variables with 2 different types of the variable. Each variable has different level of education and occupation. Based on the characteristic salary of each individual on each individual is defined.

Exploratory Data Analysis

Let us check the types of variables in the data frame.

```
Education     object
Occupation    object
Salary         int64
dtype: object
```

Table- 1.2. Datatypes of the variable

There are total 40 rows and 3 columns in the dataset. Out of 3, 2 columns (Education and Occupation) are of object type and rest is of either integer data type.

Check for missing values in the dataset:

```
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
Education     40 non-null object
Occupation    40 non-null object
Salary        40 non-null int64
dtypes: int64(1), object(2)
memory usage: 1.0+ KB
```

Table- 1.3. Check null values

Problem 1A:

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

For Education:

The Hypothesis for the One Way ANOVA are:

H_0***: The mean salary is same across 3 levels of Education***

H_a***: For at least one level of Education, mean salary of Education level is different***

Fig – 1.1.1 Hypothesis for Salary vs. Education

For Occupation:

The Hypothesis for the One Way ANOVA are:

H_0***: The mean number of salary is the same at 3 levels of Occupation***

H_a***: For at least one level of Occupation, mean number of salary of Occupation level is different***

Fig – 1.1.2 Hypothesis for Salary vs. Occupation

From this, Null hypothesis and the Alternative hypothesis are framed for the Salary against Education and Salary against Occupation.

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

The one-way anova for salary with Education is framed.

	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 – 1.4 One-way Anova for Salary vs. Education

Now, we see that the corresponding p-value (1.257709e-08) is less than alpha (0.05). Thus, we **eject** the **Null Hypothesis** ($H0$).

1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

The one-way anova for salary with Occupation is framed.

	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 – 1.5 One-way Anova for Salary vs. Occupation

Now, we see that the corresponding p-value (0.458508) is greater than alpha (0.05). Thus, we **fail to reject** the **Null Hypothesis** (0).

1.4 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)

From the one-way anova result.

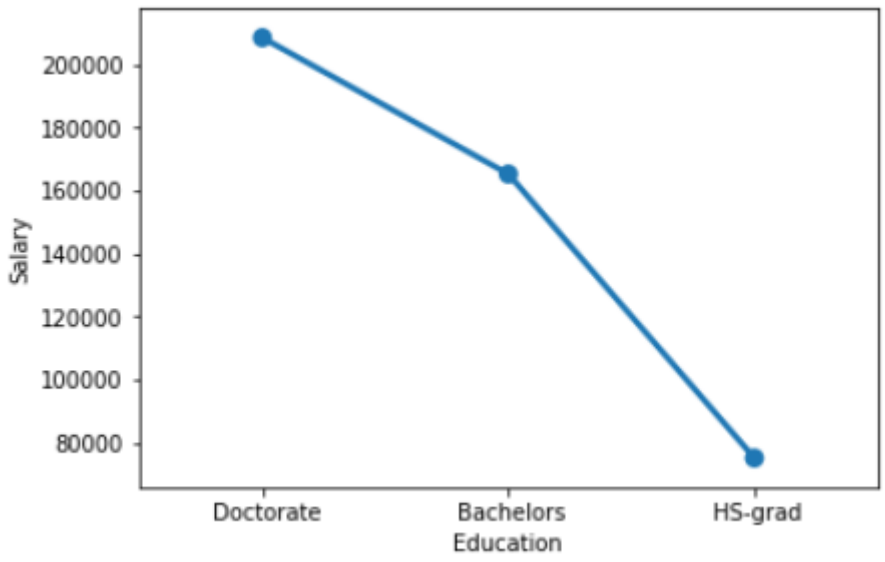

Fig – 1.4 Pointplot for Salary vs. Education

The null hypothesis is rejected for Salary Vs. Education. From this above Pointplot, the Mean are significantly different for all classes.

Problem 1B:

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

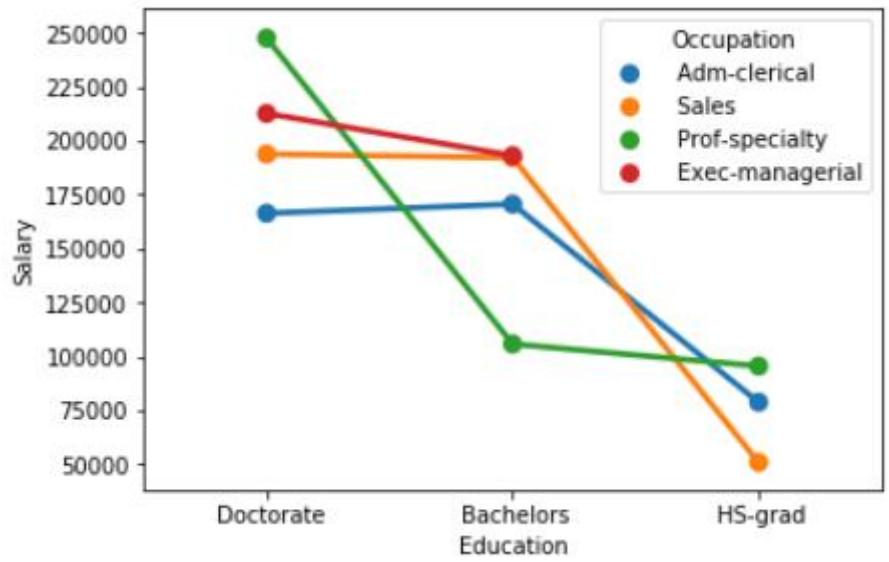

Fig – 1.5 Pointplot for Education vs. Salary vs. Occupation

Inference from the above pointplot, Doctorate education in the prof-speciality has the highest salary of 250000, whereas HS-grad education with the sales occupation has the least salary of around 50000.Bachelors education with the occupation (Sales and Exec-managerial) has the same salary of 200000.

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?

The Hypothesis for the two Way ANOVA are:

H_0***: The mean salary is the same at interaction with Education and Occupation***

H_a***: For at least one level of Occupation, mean number of salary of Occupation and Education Interaction is different***

Fig – 1.6.1 Hypothesis for two way anova

	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

Fig – 1.6.2 probability of interaction between Occupation and Education

The p-Value (2.913740e-05) is less than the alpha (significant value - 0.05). We failed to reject the null hull hypothesis. The mean salary of interacted Education and occupation is same.

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

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

Problem – 2

Summary

The data is gathered about the students, joined at different College/University after completing the12th grade. This dataset consist of data of 777 students who have enrolled for the different university. In this problem statement, we will explore the dataset and perform Principle Component Analysis (PCA).

Introduction

The purpose of this exercise is to perform PCA by reducing the dimensionality without losing the data. The Principle Component Analysis of this dataset will reduce the dimensionality by reducing the variable of the dataset, while preserving as much information as possible. This dataset consist of 777 rows and 18 columns, by reducing dimensionality, for eg.(from 3d to 2d) without losing of data.

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

Sample of the dataset:

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

Table 2.1 Dataset Sample

Dataset has 18 variables with student details. Based on the Student details, who have enrolled in different colleges is defined.

Exploratory Data Analysis

Let us check the types of variables in the data frame.

```
Names           object
Apps             int64
Accept           int64
Enroll           int64
Top10perc        int64
Top25perc        int64
F.Undergrad      int64
P.Undergrad      int64
Outstate         int64
Room.Board       int64
Books            int64
Personal         int64
PhD              int64
Terminal         int64
S.F.Ratio      float64
perc.alumni      int64
Expend           int64
Grad.Rate        int64
dtype: object
```

Table 2.2 Datatypes of the variable

There are total 777 rows and 18 columns in the dataset. Out of 18, 1 column is of object type, 1 column is of float (Decimal value) type and rest 16 are of integer data type.

Check for missing values in the dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 18 columns):
Names          777 non-null object
Apps           777 non-null int64
Accept         777 non-null int64
Enroll         777 non-null int64
Top10perc      777 non-null int64
Top25perc      777 non-null int64
F.Undergrad    777 non-null int64
P.Undergrad    777 non-null int64
Outstate       777 non-null int64
Room.Board     777 non-null int64
Books          777 non-null int64
Personal       777 non-null int64
PhD            777 non-null int64
Terminal       777 non-null int64
S.F.Ratio      777 non-null float64
perc.alumni    777 non-null int64
Expend         777 non-null int64
Grad.Rate      777 non-null int64
dtypes: float64(1), int64(16), object(1)
memory usage: 109.3+ KB
```

Table 2.3 Check null values

From this, it is clear that there are no null values present in the dataset.

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

Univariate Analysis:

Univariate analysis is the simplest form of analysing data. Analysing each variable in detail.

Insights :

Both Enroll and Accept rate have outliers in upper values.

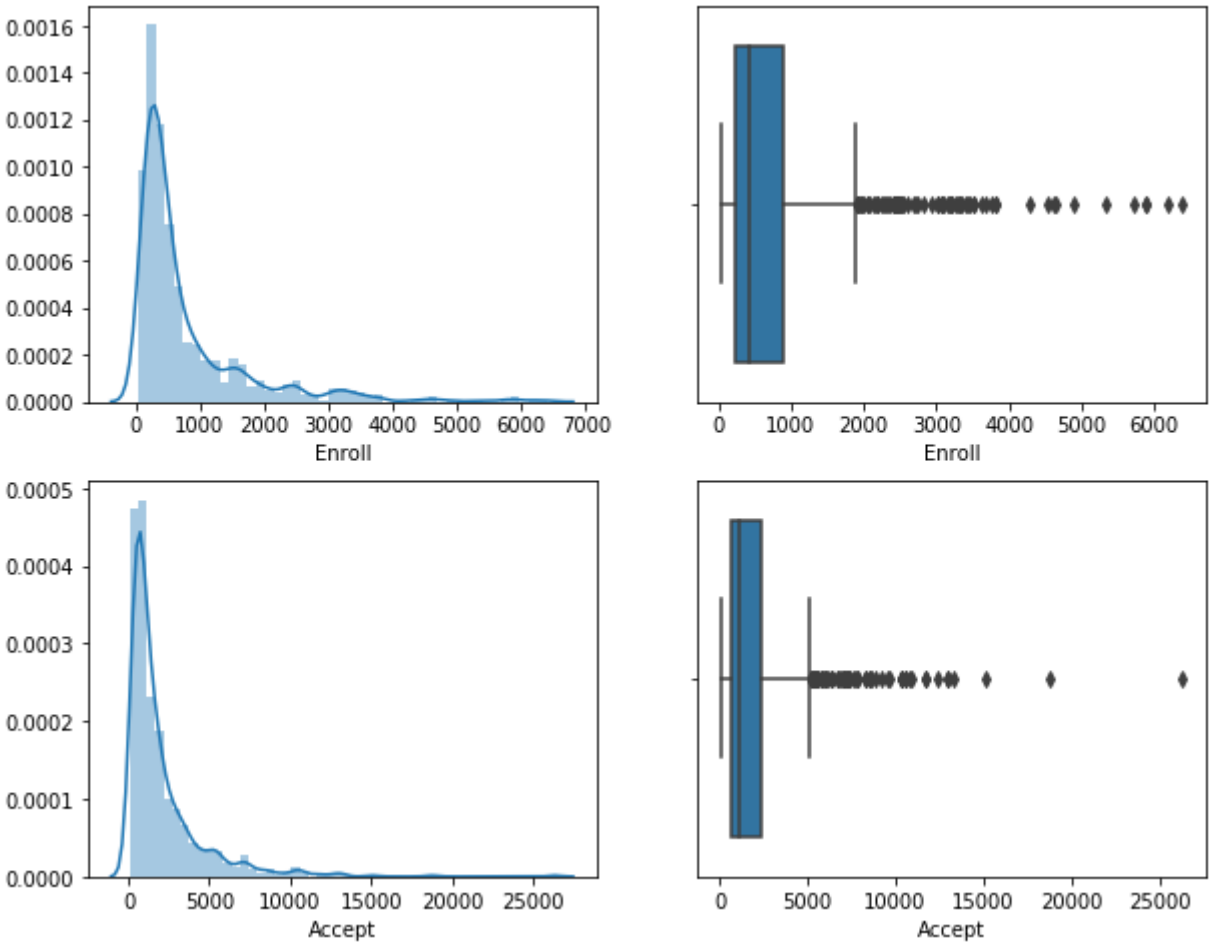

Fig – 2.1.1 Univariate analysis distplot and boxplot

Multivariate Analysis:

Analysing the data with two variables.

Insights :

There is a strong correlation observed between few fields. 'Apps' is highly correlated to 'Accept' and 'Enroll'

Also, 'Apps' shows high correlation with 'F.undergrad'

Whereas, 'S.F.Ratio' shows least correlation with 'Outstate'

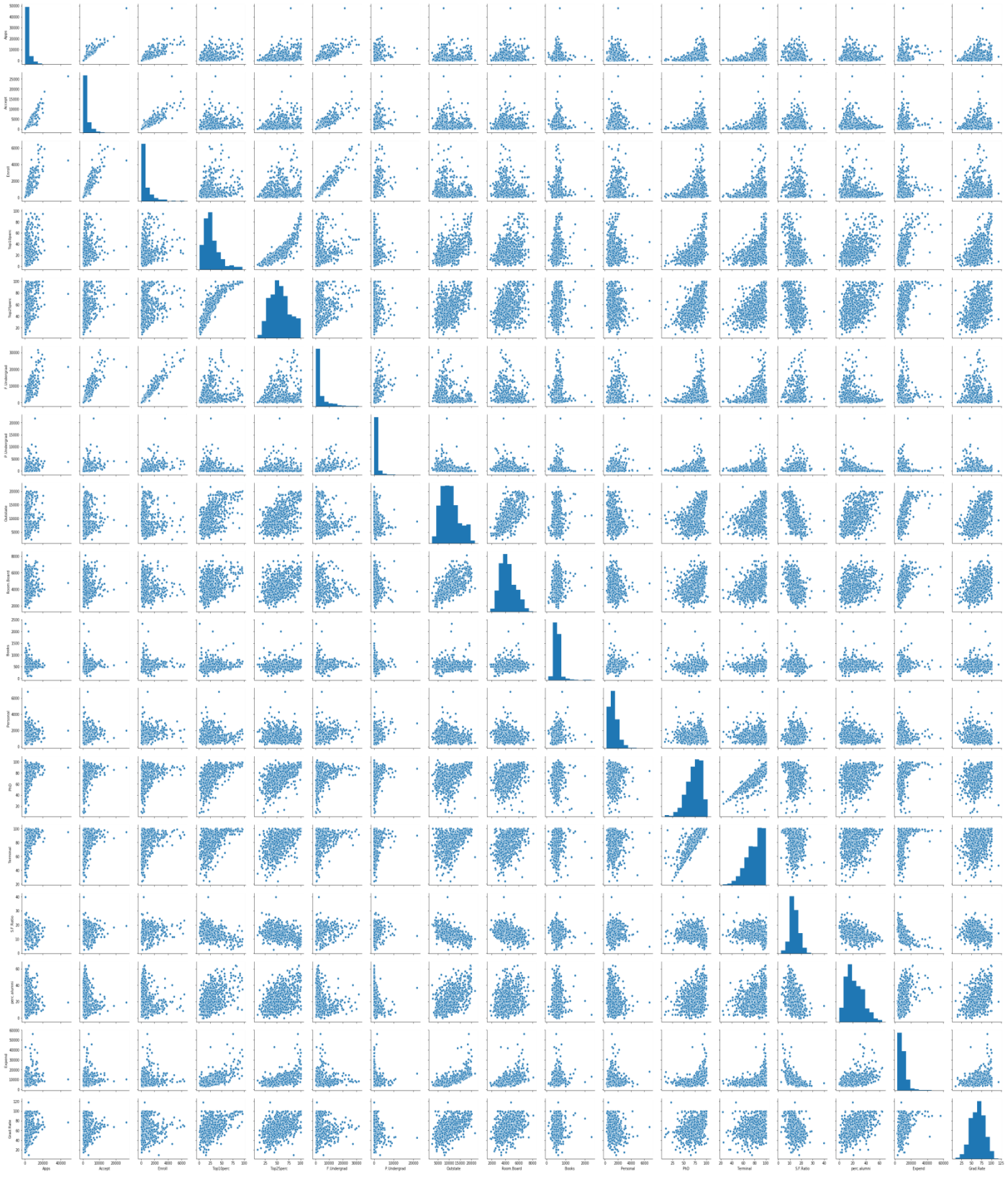

Fig – 2.1.2 Multivariate analysis of pairplot

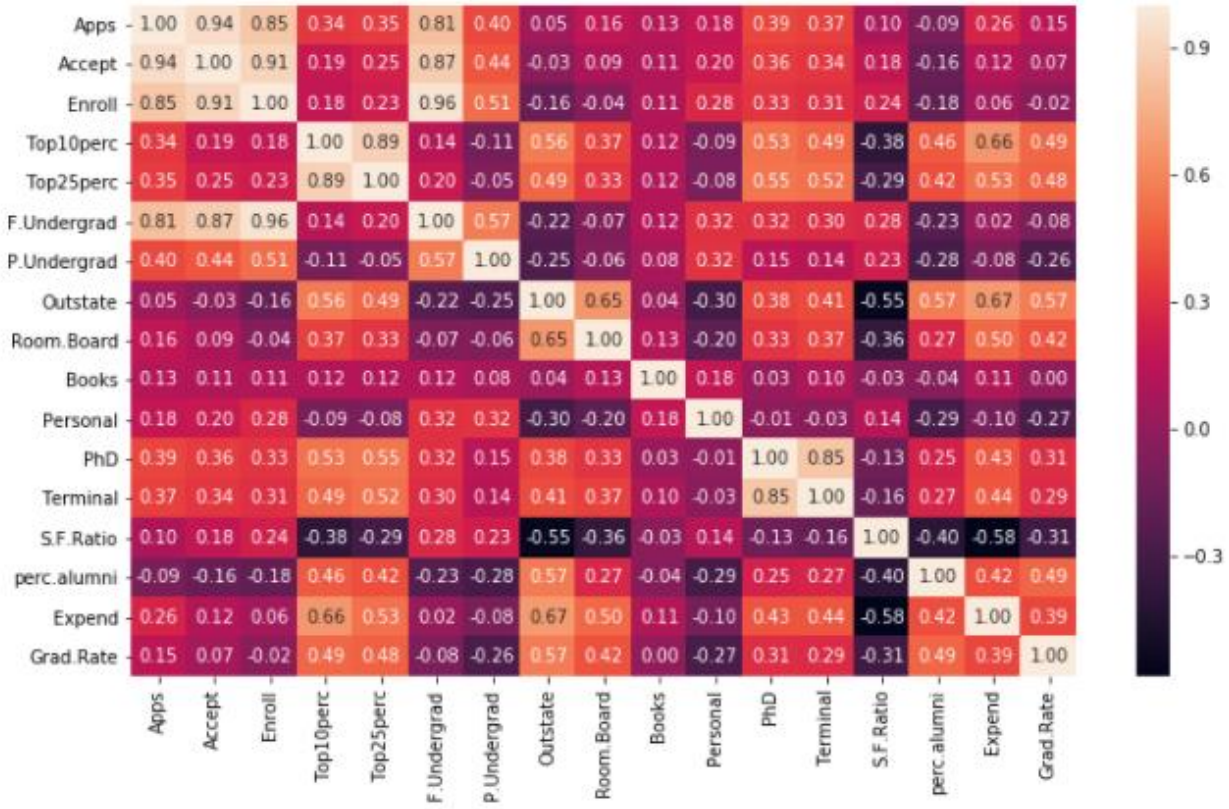

Fig – 2.1.3 Multivariate analysis heatmap

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

Yes, it is necessary to normalize data before performing PCA. The PCA calculates a new projection of your data set. ... If you normalize your data, all variables have the same standard deviation, thus all variables have the same weight and your PCA calculates relevant axis.

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

Table 2.2.1 Data Before scaling

Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	p
-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.163028	-0.115729	1.013776	
-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.675646	-3.378176	-0.477704	
-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1.204845	-0.931341	-0.300749	
-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.185206	1.175657	-1.615274	
-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0.204672	-0.523535	-0.553542	

Table 2.2.2 Data after scaling

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

Both covariance and correlation measure the relationship and the dependency between two variables.

Covariance	Correlation
1. Covariance indicates the direction of the linear relationship between variables.	1. Correlation measures both the strength and direction of the linear relationship between two variables.
2. Covariance values are not standardized.	2. Correlation values are standardized.
3. Covariance can vary between - infinity to + infinity.	3. Correlation can vary between - 1 to + 1.

Covariance Matrix:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termina
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	0.36996
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	0.33801
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	0.30867
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	0.49176
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	0.52542
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	0.30040
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	0.14208
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	0.40850
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	0.37502
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	0.10008
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	-0.03065
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	0.85068
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	1.00128
S.F.Ratio	0.095756	0.176456	0.237577	-0.385370	-0.295009	0.260064	0.232830	-0.555536	-0.363095	-0.031970	0.136521	-0.130698	-0.16031
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	0.26747
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	0.43936
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	0.28990

Table 2.3.1 covariance matrix

Correlation Matrix:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Termina
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	0.36949
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	0.33758
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	0.30827
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	0.49113
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	0.52474
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	0.30001
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	0.14190
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	0.40798
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	0.37454
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	0.09995
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	-0.03061
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	0.84958
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	1.00000
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	-0.16010
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	0.26713
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	0.43879
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	0.28952

Table 2.3.2 correlation matrix

2.4) 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]

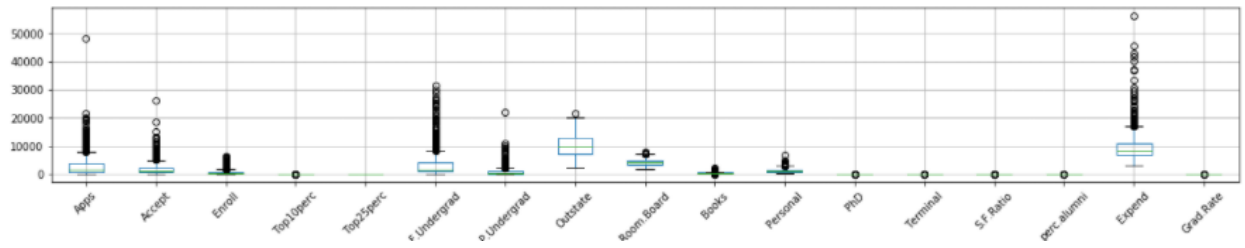

Fig – 2.4.1 Boxplot before scaling

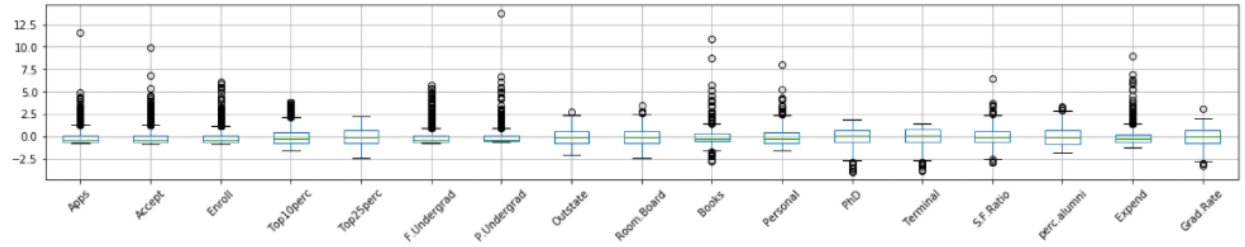

Fig – 2.4.2 Boxplot after scaling

The range in y axis is in 10000 difference in the original dataframe, whereas the range of scaled data where in 2.5 difference. The data are in the different scaling. So, the values in 'Top10perc' and 'Top25perc' box plot is not visible in the original dataframe, whereas, after scaling all the values are converted to the same range.

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

Bartletts Test of Sphericity:

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

- H0: All variables in the data are uncorrelated

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

```
0.0
```

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

Fig – 2.5.1 Output for probability of Bartletts Test

KMO Test:

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

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

```
0.8131251200373503
```

MSA(0.8131251200373503) > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

Fig – 2.5.2 Output for probability of KMO Test

From this Bartletts test and KMO test, the data is sufficient and good to perform PCA.

The Component Output is

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

Fig – 2.5.3 Output for PCA component

```
Eigen vector are -
 [[ 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]]
```

Fig – 2.5.4 Output for Eigen vector

```
Eigen values are -
 [0.32020628 0.26340214 0.06900917 0.05922989 0.05488405 0.04984701
 0.03558871]
```

Fig – 2.5.5 Output for Eigen value

Eigen values are plotted as Scree plot. From Screeplot we can infer that around 85% of data lies in the first 8 components of the Eigen values in PCA

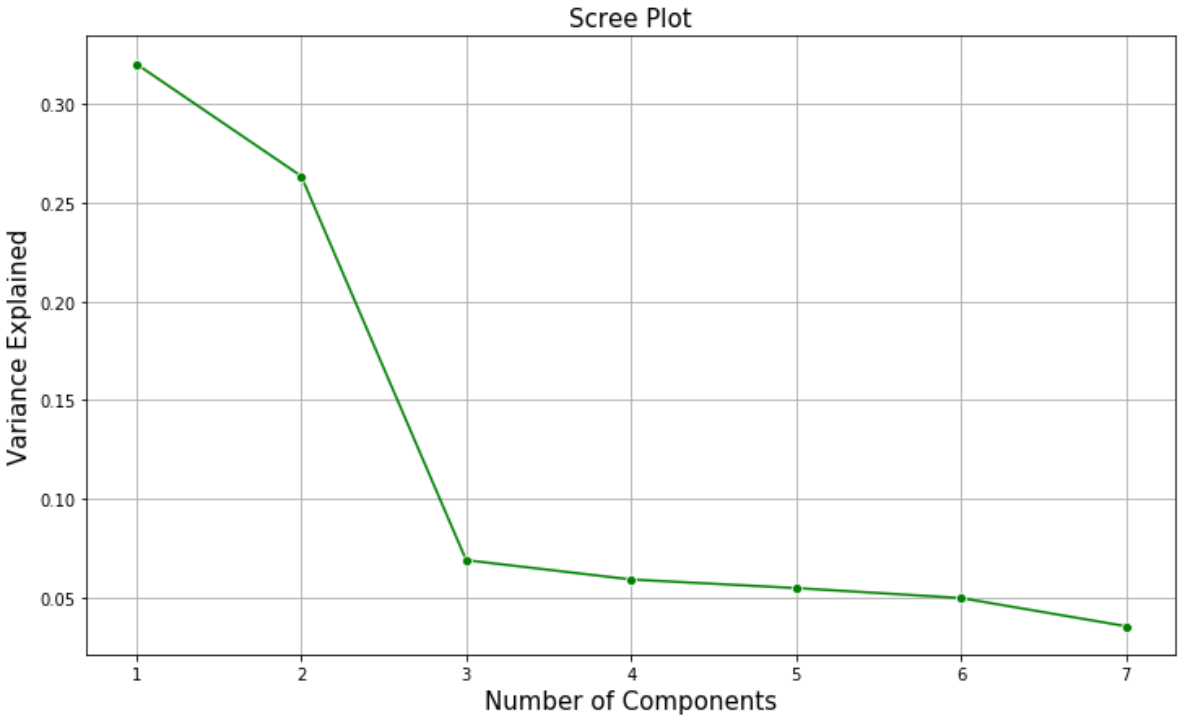

Fig – 2.5.6 Scree Plot for Eigen value

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

To perform PCA, All the Eigen vectors are loaded into a dataframe with 7 rows and 17 columns (excluding of categorical column from the original dataframe)

The Eigen vectors are converted as the dataframe. Sample dataframe image is attached below.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Rati
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0.064758	-0.042529	0.318313	0.317056	-0.17695
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809	0.056342	0.219929	0.058311	0.046429	0.24666
2	-0.063092	-0.101249	-0.082986	0.035056	-0.024148	-0.061393	0.139682	0.046599	0.148967	0.677412	0.499721	-0.127028	-0.066038	-0.28984
3	0.281311	0.267817	0.161827	-0.051547	-0.109767	0.100412	-0.158558	0.131291	0.184996	0.087089	-0.230711	-0.534725	-0.519443	-0.16118
4	0.005741	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919	-0.127289	-0.222311	0.140166	0.204720	-0.07938
5	-0.016237	0.007535	-0.042558	-0.052693	0.033092	-0.043454	-0.191199	-0.030000	0.162755	0.641055	-0.331398	0.091256	0.154928	0.48704
6	-0.042486	-0.012950	-0.027693	-0.161332	-0.118486	-0.025076	0.061042	0.108529	0.209744	-0.149692	0.633790	-0.001096	-0.028477	0.21925

Table 2.6 Eigen vector as Dataframe

2.7)Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [Hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

First row from the Eigen vector dataframe

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Exp
0	0.25	0.21	0.18	0.35	0.34	0.15	0.03	0.29	0.25	0.06	-0.04	0.32	0.32	-0.18	0.21	

Table 2.7.1 Eigen vector dataframe first row

Linear equation formula:

PC = m1x1+m2x2+m3x3+....m17x17
Where,

m - Represents Eigen vector
x – Represents the variable

Explicit Form of first PC:

PC1= 0.25*Apps + 0.21*Accept + 0.18*Enroll + 0.35*Top10perc + 0.34*Top25perc + 0.15*F.Undergrad+ 0.03*P.Undergrad + 0.29*Outstate + 0.25*Room.Board + 0.06*Books + -0.04*Personal + 0.32*PhD +0.32*Terminal + -0.18*S.F.Ratio + 0.21*perc.alumni + 0.32*Expend + 0.25*Grad.Rate

	column	PC1
0	Apps	0.0618843
1	Accept	0.0430984
2	Enroll	0.031083
3	Top10perc	0.12551
4	Top25perc	0.118337
5	F.Undergrad	0.0239138
6	P.Undergrad	0.000699206
7	Outstate	0.0868696
8	Room.Board	0.0620162
9	Books	0.00419354
10	Personal	0.00180868
11	PhD	0.101323
12	Terminal	0.100525
13	S.F.Ratio	0.0313141
14	perc.alumni	0.0420588
15	Expend	0.101703
16	Grad.Rate	0.0636632

Table 2.7.2 Explicit form of first PC

	ratio	PC
0	0.32	PC1
1	0.58	PC2
2	0.53	PC3
3	0.33	PC4
4	0.32	PC5
5	0.32	PC6
6	0.31	PC7

Table 2.7.3 PC values

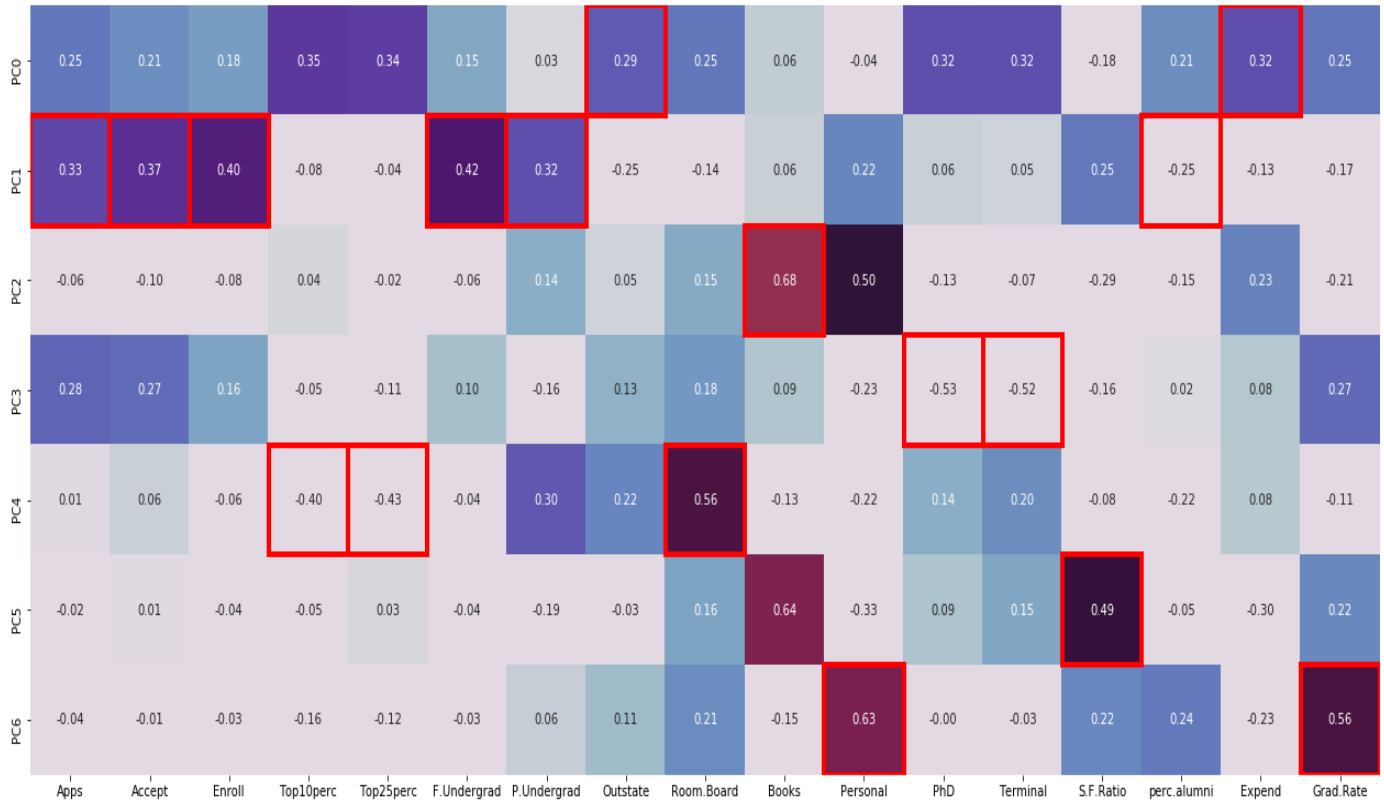

Fig – 2.7.1 Heatmap for PC values

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?

The cumulative value up to seventh Principal Component is 85.21. General rule of thumb is to choose first k PC's such that the first k PC's explaining 70-90% of the total variance. Hence from the cumulative values of Eigen values, help in selecting the required no. of PC's. In this case first seven PC's have been selected capturing 85.2% of variation and thereby reducing our dimension by half.

```
[0.3202062819886914, 0.5836084263498161, 0.6526175918920409, 0.7118474841213039, 0.7667315352248886, 0.8165785447704629, 0.8521
672596879294]

Cumulative Variance Explained 85.22 %
```

Fig – 2.8.1 Cumulative Eigen value

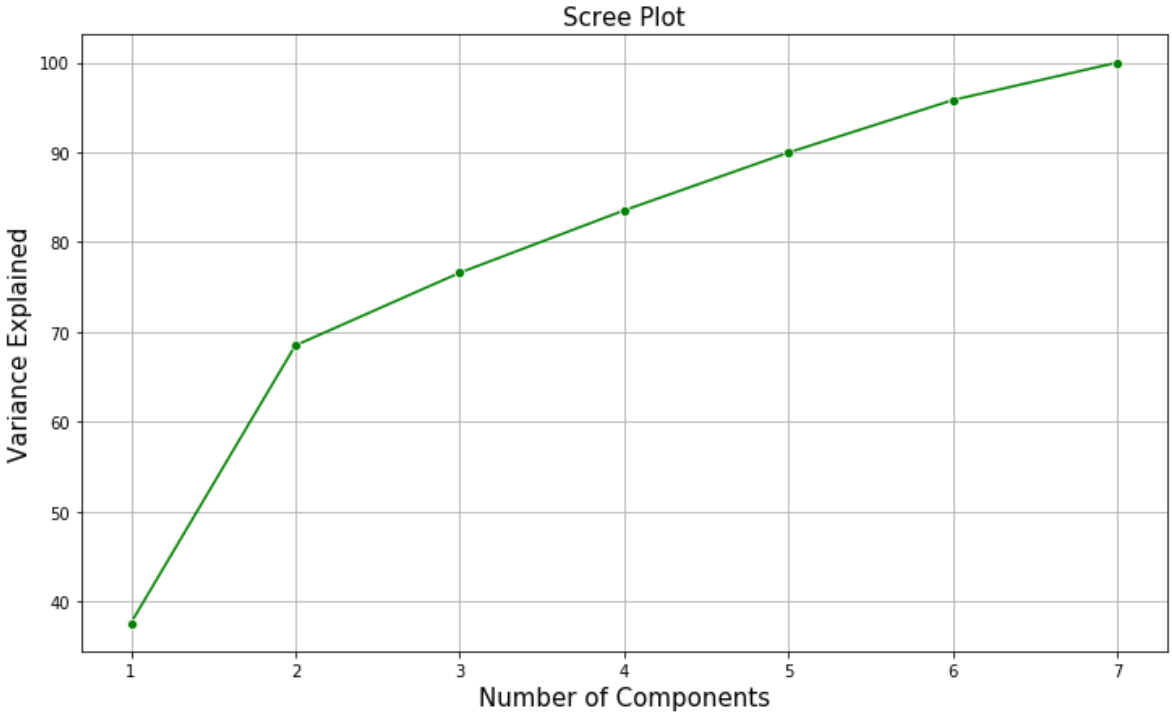

Fig – 2.8.2 Scree Plot for Cumulative Eigen value

The above Scree plot shows, the sum of Eigen values are added and plotted in the graph.

2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]

PCA is a “dimensionality reduction” method. It reduces the number of variables that are correlated to each other into fewer independent variables without losing the essence of these variables. It provides an overview of linear relationships between inputs and variables.

Business Implication:

From this dataset, 17 continuous variables have been reduced to 7 variables without losing of the data. The output of the PCA from this dataset, after performing the PCA 85% data have been is available in the 7 variable by reducing t he dimensionality.

After completing the PCA, 7 PC's are used for the further process.

The sample new dataframe after concatenating the categorical and numerical (After PCA) dataframe.

	Names	Accommodation_expenses	Student_profiling	Books	Faculty_profiling	Merit_lodging	S.F.ratio	Graduation_expenses
0	Abilene Christian University	-1.592855	0.767334	-0.101074	-0.921749	-0.743975	-0.298306	0.638443
1	Adelphi University	-2.192402	-0.578830	2.278798	3.588918	1.059997	-0.177137	0.236753
2	Adrian College	-1.430964	-1.092819	-0.438093	0.677241	-0.369613	-0.960592	-0.248276
3	Agnes Scott College	2.855557	-2.630612	0.141722	-1.295486	-0.183837	-1.059508	-1.249356
4	Alaska Pacific University	-2.212008	0.021631	2.387030	-1.114538	0.684451	0.004918	-2.159220

Table 2.9.1 Sample data of new dataframe after PCA

	Names	Accommodation_expenses	Student_profiling	Books	Faculty_profiling	Merit_lodging	S.F.ratio	Graduation_expenses
count	777	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02
unique	777	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	Massachusetts Institute of Technology	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	9.773392e-17	1.428858e-18	6.201246e-17	-8.751758e-17	2.743408e-17	-1.857516e-18	-7.915876e-17
std	NaN	2.334635e+00	2.117453e+00	1.083821e+00	1.004094e+00	9.665564e-01	9.211358e-01	7.783237e-01
min	NaN	-5.662905e+00	-3.590891e+00	-2.941286e+00	-2.943103e+00	-2.690124e+00	-3.822954e+00	-2.810575e+00
25%	NaN	-1.731200e+00	-1.348075e+00	-6.663045e-01	-6.558315e-01	-6.998497e-01	-5.229543e-01	-5.082039e-01
50%	NaN	-2.994567e-01	-6.268330e-01	-1.010735e-01	-5.842776e-02	-5.112391e-02	-2.959210e-03	3.816859e-02
75%	NaN	1.339533e+00	6.924474e-01	4.943709e-01	5.987764e-01	6.313480e-01	4.553863e-01	4.729330e-01
max	NaN	8.047182e+00	1.200237e+01	9.006415e+00	5.177648e+00	4.248195e+00	5.991244e+00	4.350537e+00

Table 2.9.2 Data description of new dataframe after PCA