GREAT LEARNING

POST GRADUATE PROGRAM IN DATA SCIENCE & BUSINESS ANALYTICS

BUSINESS REPORT

CASE STUDIES ON:

- Salary Data Analysis Using ANOVA
- College Survey post 12th (EDA & PCA)

Submitted By:

STEFFIN JOHN

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Case Study 1 - Salary Data Analysis Using ANOVA

Overview:

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.]

Summary:

This business report provides detailed explanation on the approach to each problem definition, solution to those the problems provides some key insights/recommendations to the business.

Problem 1A:

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

The Hypothesis of One-Way ANOVA for 'Education' with respect to 'Salary'

HO: The mean salary of individuals is same for all 3 levels of Education.

HA: For at least one level of Education, mean salary of individuals is different.

The Hypothesis of One-Way ANOVA for 'Occupation' with respect to 'Salary'

HO: The mean salary of individuals is same for all 4 levels of Occupation.

HA: For at least one level of Occupation, mean salary of individuals is different.

Were,

HO = Null Hypothesis

HA = Alternate Hypothesis

Also, it is given that the dataset qualifies all the assumptions for ANOVA.

a) Each group sample is drawn from a normally distributed population

b) All populations have a common variance

c) All samples are independently of each other

d) Within each sample, the observations are sampled randomly and independently of each other

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.

Before performing One-Way ANOVA, we convert the variable 'Education' from object to categorical datatype and subdivide the dataset according to categories of variable 'Education' (i.e., HS-Grad, Doctorate and Bachelors).

Now, we perform One-Way ANOVA

The Hypothesis of One-Way ANOVA for 'Education' with respect to 'Salary'

HO: The mean salary of individuals is same for all 3 levels of Education.

HA: For at least one level of Education, mean salary of individuals is different.

Below is the result from python code:

	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: ANOVA results for variable 'Education' with respect to variable 'Salary'

As shown by the table above, the relevant p-value is lower than alpha (0.05). As a result, we accept the alternative hypothesis and reject the null hypothesis.

Therefore, the mean salary of individuals varies for at least one degree of education.

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.

Before performing One-Way ANOVA, we convert the variable 'Occupation' from object to categorical datatype and we subdivide the dataset according to categories of variable 'Occupation' (i.e. Adm-clerical, Sales, Prof-specialty, Exec-managerial).

Now, we perform One-Way ANOVA

The Hypothesis of One Way ANOVA for 'Occupation' with respect to 'Salary'

HO: The mean salary of individuals is same for all 4 levels of Occupation.

HA: For at least one level of Occupation, mean salary of individuals is different.

Below is the result from python code:

	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 2: ANOVA results for variable 'Occupation' with respect to variable 'Salary'

From the above table, we see that the corresponding p-value is greater than alpha (0.05). Thus, we fail to reject the Null hypothesis.

Therefore, the mean salary of individuals is same for all 4 levels of Occupation.

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

We know the null hypothesis is rejected in (2). We can find the difference in class means using two methods

Method-1: Tukey Honest Significance Test

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Bachelors	Doctorate	43274.0667	0.0146	7537.2745	79010.8589	True
Bachelors	HS-grad	-90114.1556	0.0	-132039.7353	-48188.5758	True
Doctorate	HS-grad	-133388.2222	0.0	-174819.5736	-91956.8709	True

Table 3: Tukey Honest Significance Test

From the table above, it is clear that:

a) Those with doctorates and bachelor's degrees make significantly more money on average than those with only an HS diploma.

b) The mean earnings of people with doctorates and bachelor's degrees differ moderately.

Method-2: Point Plot of Education vs Salary

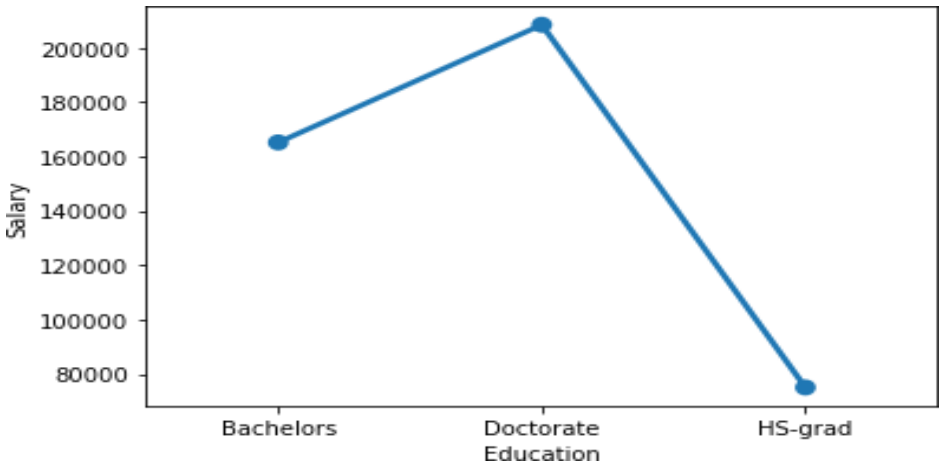

Figure 1: Plot of Education vs Salary

Observations:

From the table above, it is clear that:

a) Those with doctorates and bachelor's degrees make significantly more money on average than those with only an HS diploma.

b) The mean earnings of people with doctorates and bachelor's degrees differ moderately.

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]**

If the response of the continuous measure to one categorical variable depends on another categorical variable, then there is an interaction between two treatments (in this case, the categorical variables Occupation & Education) with respect to the continuous measure (in this case, the salary variable).

- When an interaction effect is present, the influence of one factor is dependent on the value of the other component.

- Interaction effects show the combined impacts of factors on the dependent measure.

- The presence of interaction effects indicates that the main effects' interpretation is incomplete or misleading.

Point plot shows level of interaction by the number of intersection points:

- More the number of intersection points in the graph, higher the interaction level between concerned variables and vice-versa.

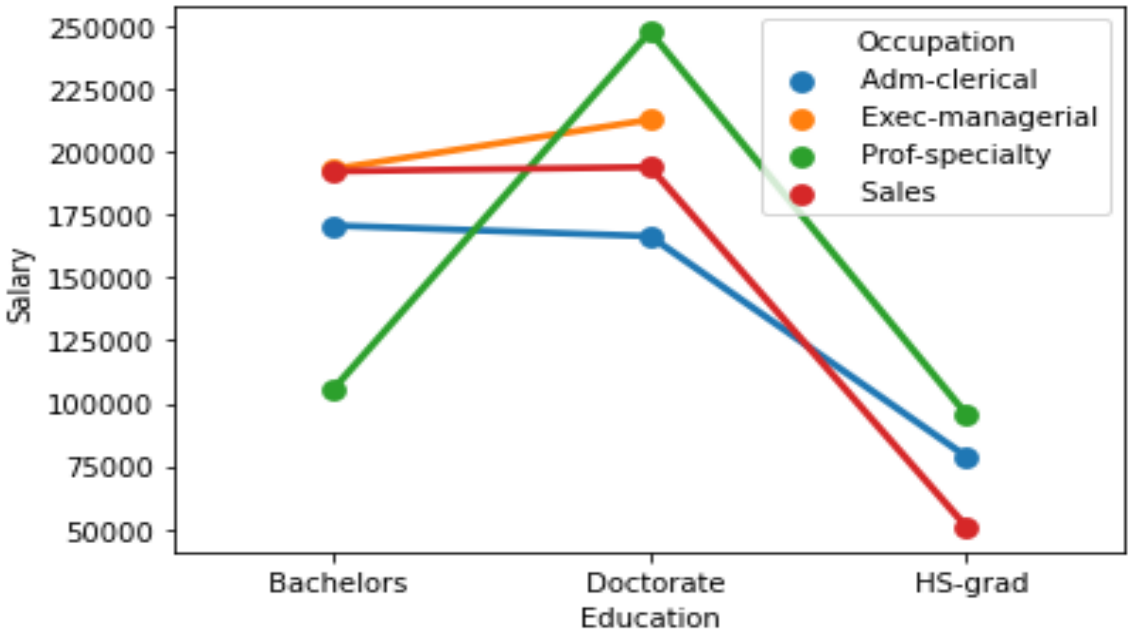

Figure 2: Interaction Plot between Education & Occupation variables

Observations:

From the graph above, we can say that:

There are several intersection points in the graph which shows there is a decent level of interaction between Occupation & Education variable.

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?

A two-way ANOVA with interaction tests three null hypotheses at the same time:

Null Hypothesis (HO):

a) There is no difference in mean salary of individuals at any level of Education.

b) There is no difference in mean salary of individuals for any type of Occupation.

c) There is no interaction effect between Education and Occupation on average salary.

Alternate Hypothesis (HA):

a) There is a difference in mean salary of individuals at any level of Education.

b) There is a difference in mean salary of individuals for any type of Occupation.

c) There is an interaction effect between Education and Occupation on average salary.

A two-way ANOVA without interaction only tests the first two of these hypotheses.

Below is the result from python code:

	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

Table 4: ANOVA results of variables 'Education' & 'Occupation' with respect to variable 'Salary' without their interaction

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

Table 5: ANOVA results of variables 'Education' & 'Occupation' with respect to variable 'Salary' along with their interaction

The p-value of the first two dependencies has changed little when the interaction effect term is included compared to the Two-Way ANOVA without the interaction effect terms.

We are only interested in the findings of the third hypothesis here because One-Way ANOVA for the Education & Occupation variable above has already been conducted separately (interaction test between Education & Occupation with respect to Salary)

Because the null hypothesis is rejected in this instance and the p-value of the interaction effect term of "education" and "occupation" is less than 0.05, we can accept the alternative hypothesis.

Therefore,

a) There is a difference in mean salary of individuals at any level of Education.

b) There is no difference in mean salary of individuals for any type of Occupation.

c) There is an interaction effect between Education and Occupation on average salary.

This means Education & Occupation variable when combined together influence the mean salaries of individuals.

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

By using ANOVA for the dataset, we came to know that –

a) Occupations of people alone do not influence mean salaries of individuals.

b) Educational qualifications of people alone on the other hand do influence mean salaries of individuals.

c) However, Occupation & Education when combined together do influence mean salaries of individuals.

d) Additionally, we learn from the pointplot that people with higher secondary degrees make significantly less money on average than those with bachelor's degrees. Additionally, the mean incomes for those with bachelor's degrees and doctorates are nearly identical. If a business choice is to be made based on the population's mean salaries, we must ensure that the decision has a favorable impact on the entire population and does not result in any bias for any group.

Case Study 2 -College Survey Post 12 th (EDA & PCA)

Overview:

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.)

Problem 2:

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

Observations of basic Data Exploration:

a) Dataset has 18 columns and 777 rows.

b) The entire dataset is of integer data type. However, column 'Names' is object datatype & S.F. Ratio is float datatype.

c) No null values.

d) Grad.Rate has a maximum value of 118

Univariate Analysis

1) Boxplot for outlier identification

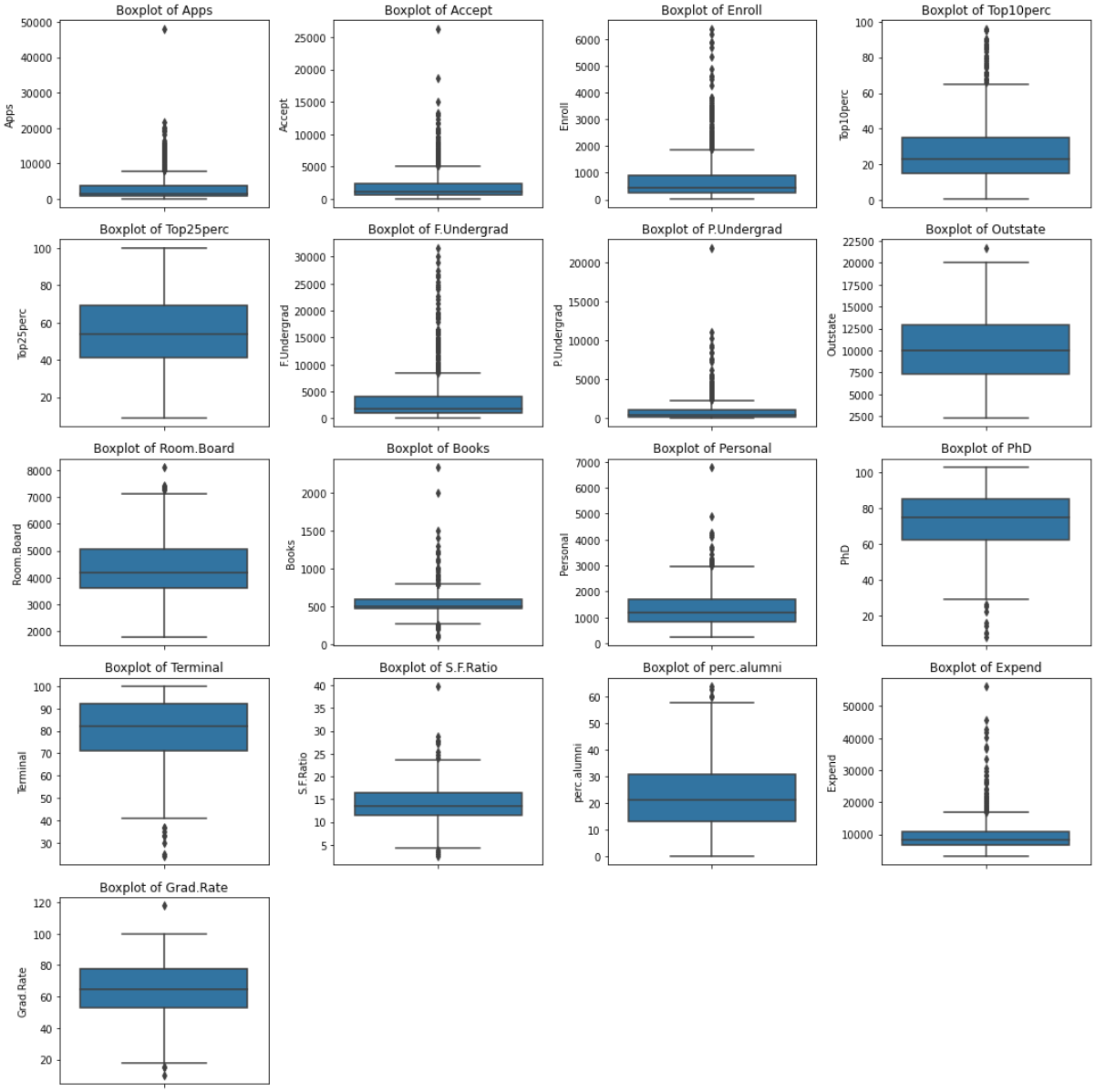

Figure 3: Boxplot for Outlier Identification

Observations:

From the graph above, we can clearly say that, there are a lot of outliers in the dataset. Hence, we are going to replace them with either the maximum or the minimum value based on which side of the boxplot they lie.

2) Distplot for studying variable distribution

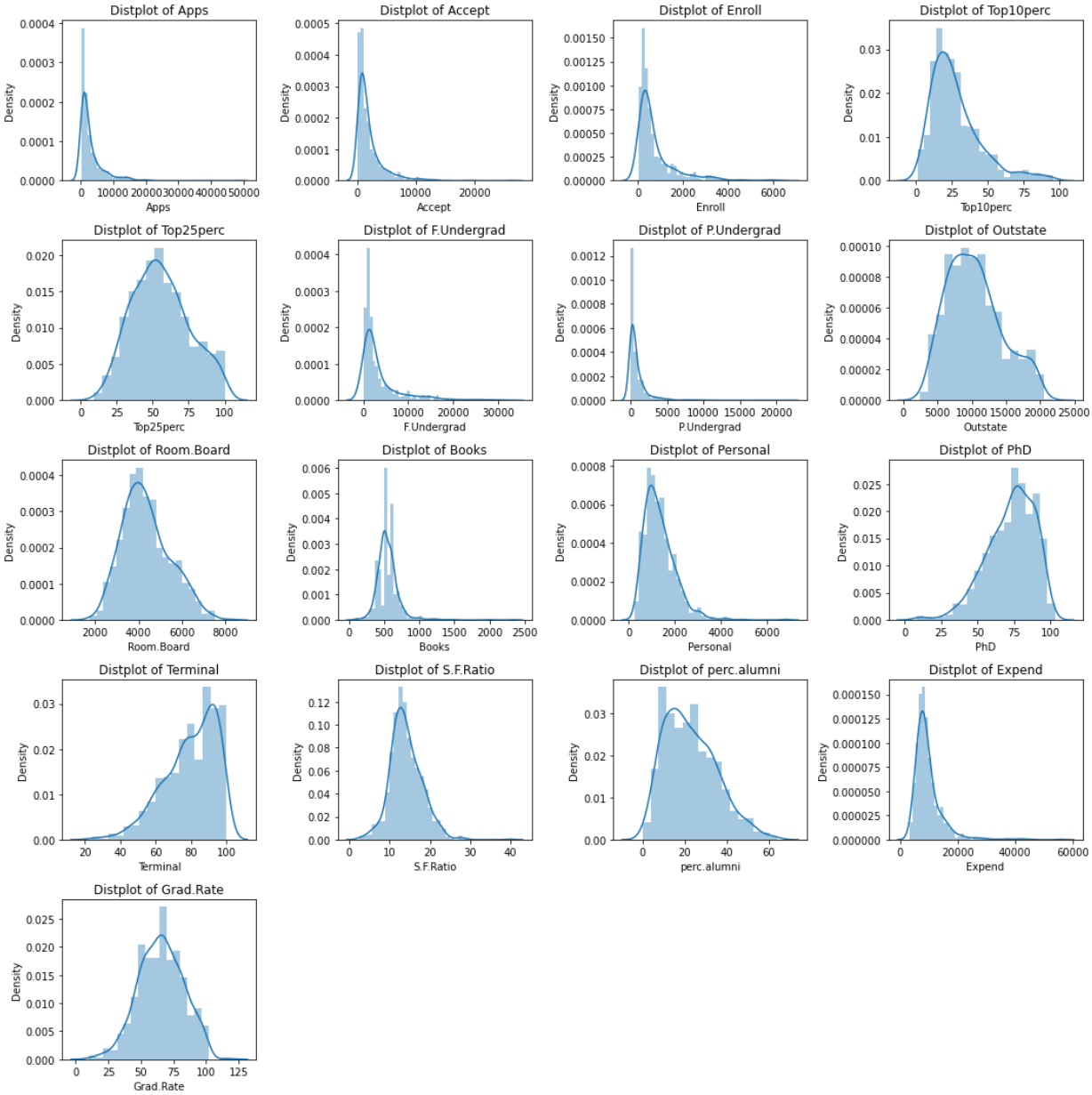

Figure 4: Distplot for studying variable distribution.

Observations:

a) The variables Apps, Accept, Enroll, Top10%, F.Undergrad, P.Undergrad, Books, Personal, % Alumni and Expend have a right-skewed distribution.

b) The variables PHD & Terminal have a left-skewed distribution.

c) The variables Top25%, Outstate, RoomBoard, S.F.Ratio, Grad Rate have a normal distribution.

Multivariate Analysis

Heatmap to study correlation between variables

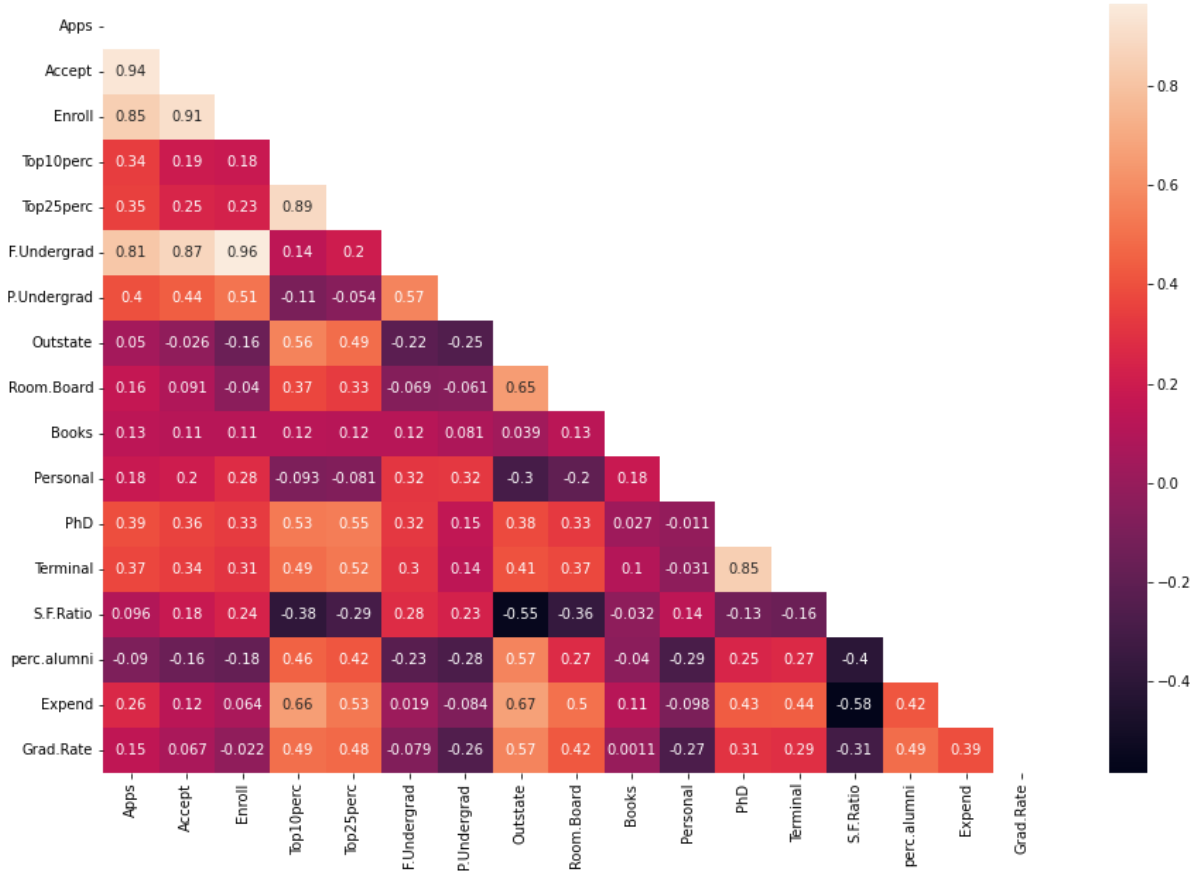

Figure 5: Correlation Heatmap

Observations:

a) Variable Apps, Accept, Enroll & F.Undergrad have a strong correlation with one another.

b) Variable Top10% has a very strong correlation with Top25%.

c) Variable PhD has a strong correlation with Terminal.

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

Yes, in this situation, normalising the data prior to doing PCA is important since some of the variables are in percentages and others are counts (number of students).

The standard deviation of our variables serves as the basis for the new axis that the PCA creates when projecting the dataset in a new way. In order to calculate the axis, a variable with a high standard deviation will be given a higher weight than a variable with a low standard deviation. If your data is normalised, all of the variables will have the same standard deviation and weight when our PCA creates the relevant axis.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
0	-0.348882	-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	-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
2	-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
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.185206	1.175667
4	-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

Table 6: Sample of data after scaling

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 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.

```
In [41]: from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
         chi_square_value,p_value=calculate_bartlett_sphericity(df_pca_scaled)
         p_value

Out[41]: 0.0
```

Table 7: Bartletts Test of Sphericity

Thus, we reject the Null Hypothesis (value = 0.0) and agree that there is at least one pair of variables in the data which are correlated

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.

```
In [42]: from factor_analyzer.factor_analyzer import calculate_kmo
         kmo_all,kmo_model=calculate_kmo(df_pca_scaled)
         kmo_model

Out[42]: 0.8131251200373522
```

Table 8: KMO Test of Sampling Adequacy

Thus, PCA is recommended as MSA (0.813) > 0.7

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

Correlation Matrix:

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

Table 9: Correlation Matrix of the scaled dataset

Covariance matrix:

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

Table 10: Covariance Matrix of the scaled dataset

Observation:

We can conclude from the two tables above that there is just a small difference in their values. This is due to the scaled nature of the dataset used to calculate covariance and correlation. The covariance matrix would have varies significantly if the dataset hadn't been resized.

Although correlation and covariance have a strong relationship, they are significantly different from one another. The latter stands to be the first choice when deciding between covariance and correlation because it is unaffected by changes in dimensions, position, and scale and may also be used to compare two sets of variables.

Pair plot:

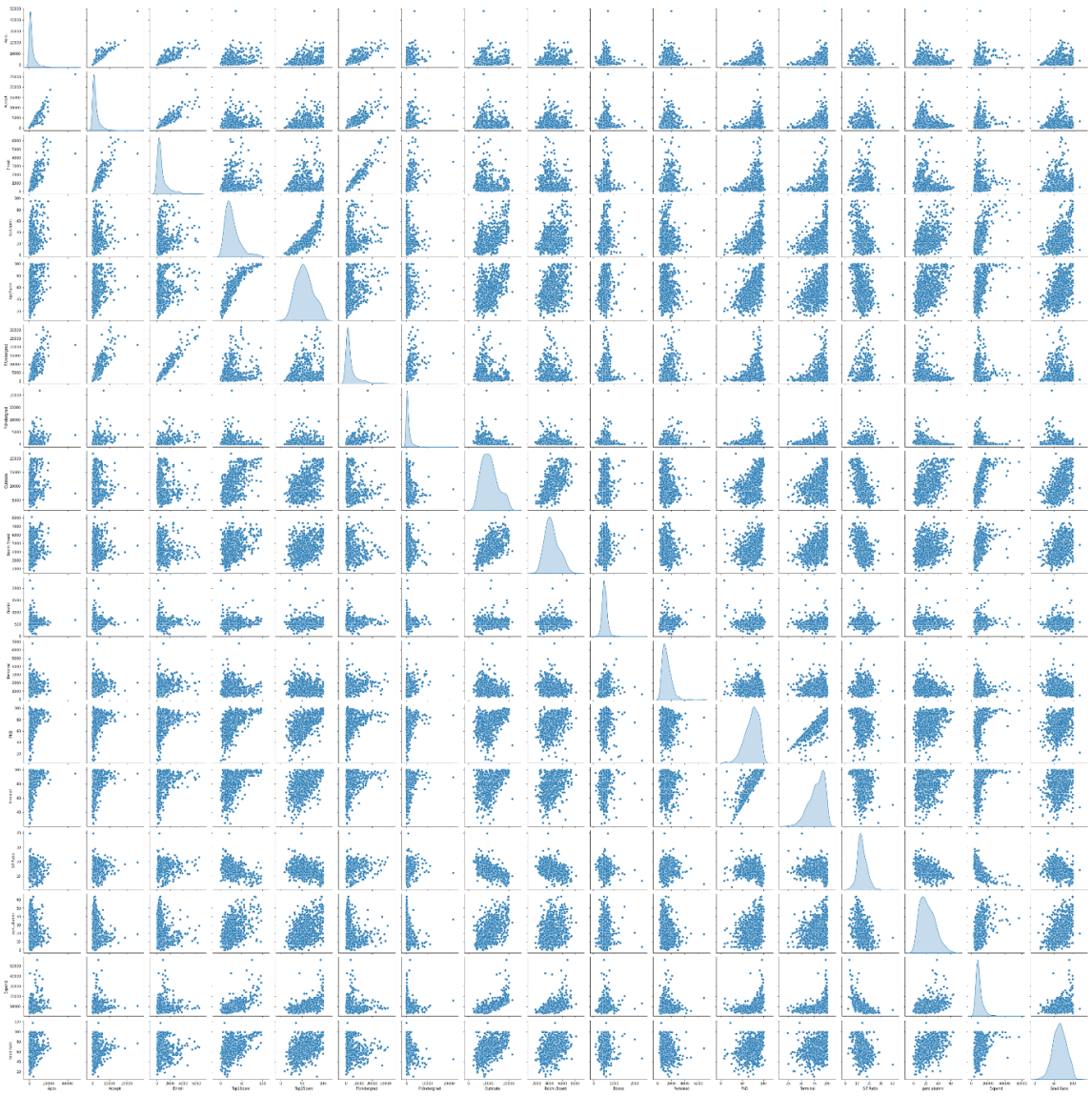

Figure 6: PairPlot

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]

Outliers in the dataset before scaling:

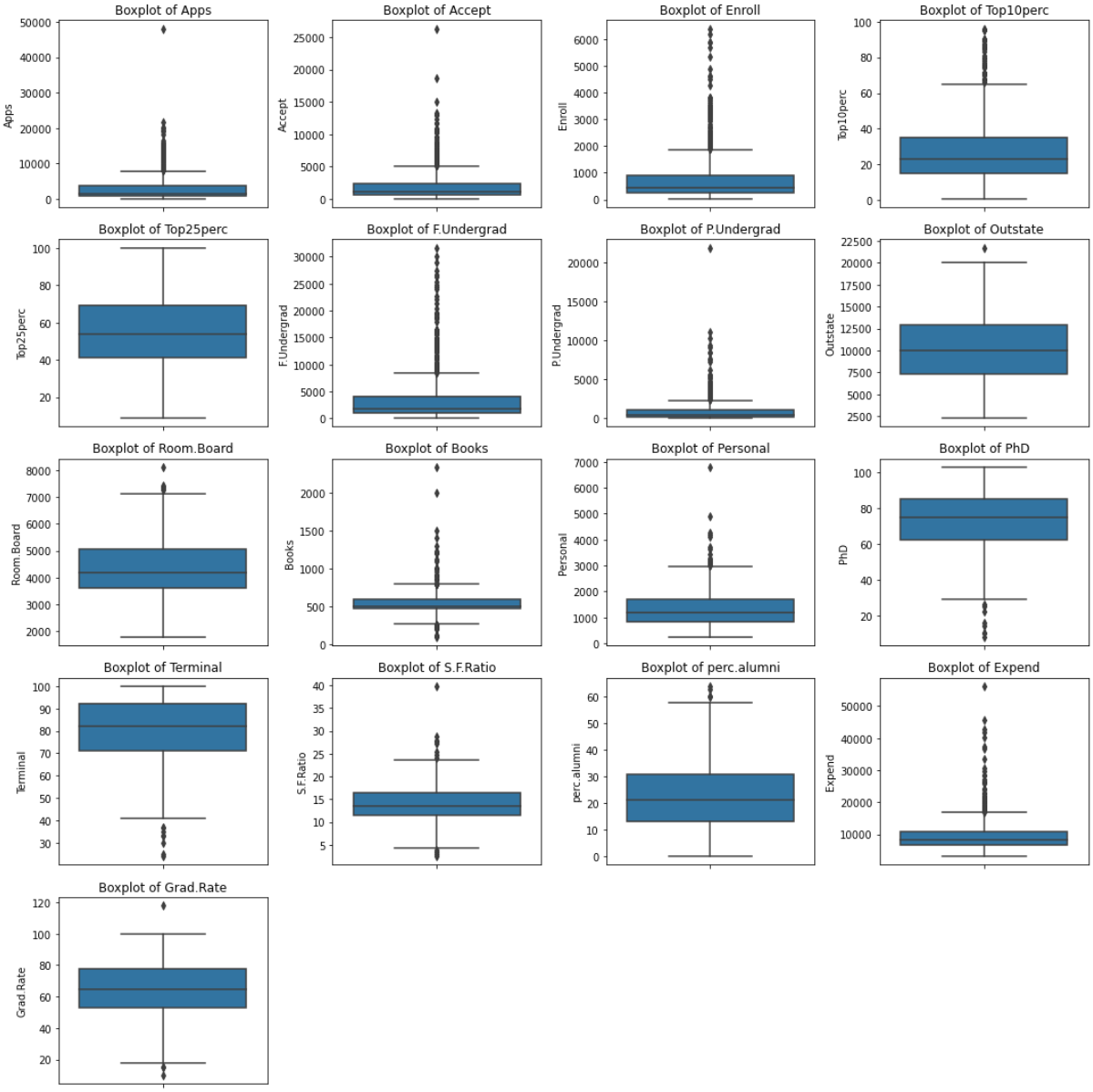

Figure 7: Outliers before scaling the dataset

Outliers in the dataset after scaling:

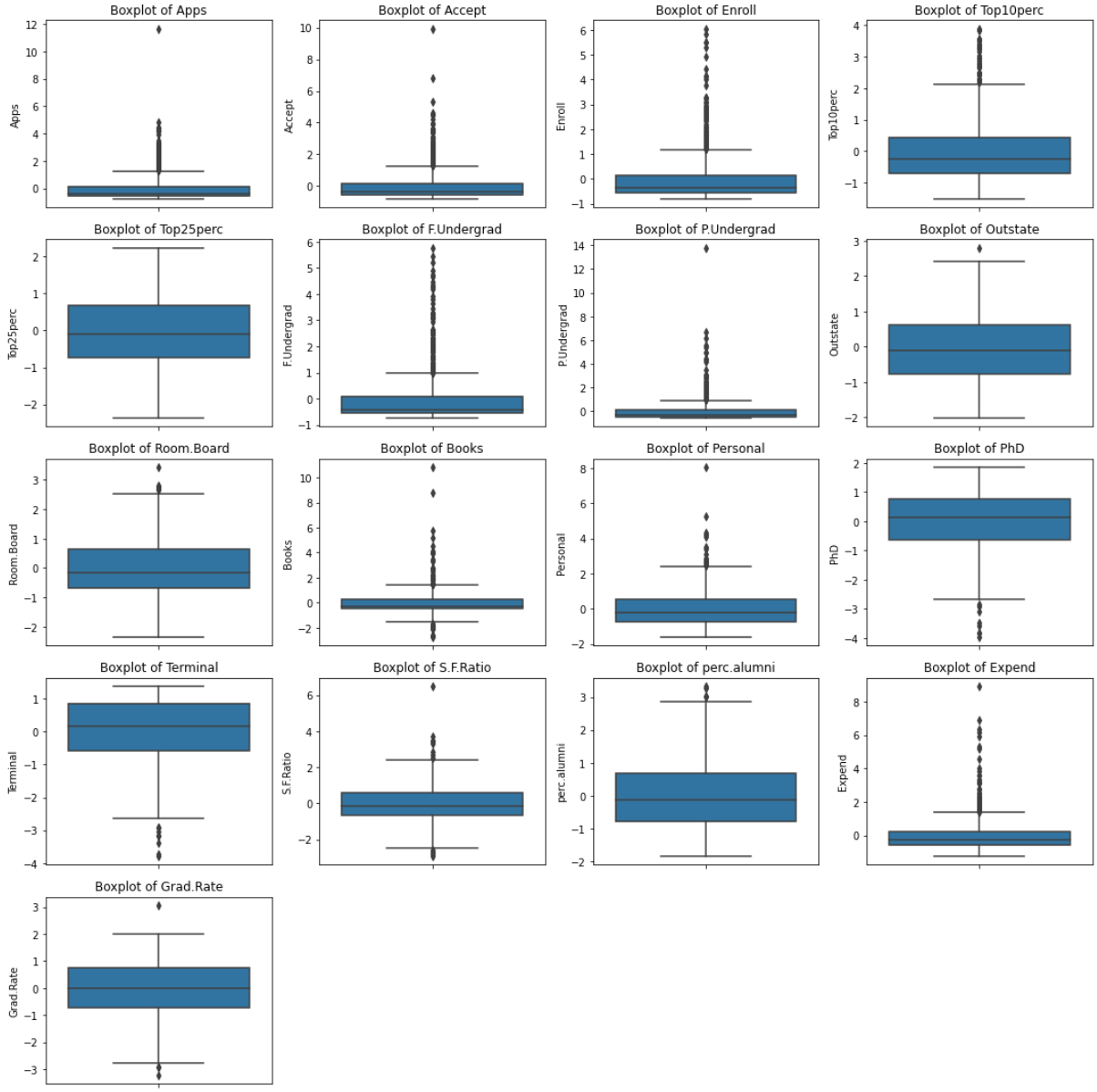

Figure 8: Outliers after scaling the dataset

Observations:

From the above two graphs, we can say that:

There is no difference in outliers of the dataset before & after scaling. Scaling of data just transforms all variables in the dataset to a same range and it has no effect whatsoever on outliers in the dataset.

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

Eigen vectors:

Below is the output of python code:

```
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]])
```

Eigen Values:

Below is the output of python code:

```
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])
```

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
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.103090	-0.090227	0.052610	0.043046	0.024071	0.595831	0.080633
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
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
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678	0.341100	0.084026	-0.008105	0.038554	0.001023	-0.107828
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
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
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
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
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
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
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
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
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
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
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
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
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

Table 11: PCs into a Dataframe exported with 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]

The explicit form of the first PC ($a_1x_1 + a_2x_2 + ... + a_nx_n$)

Below is the output from python code:

```
(0.25 * -0.35)+ (0.21 * -0.32)+ (0.18 * -0.06)+ (0.35 * -0.26)+ (0.34 *
-0.19)+ (0.15 * -0.17)+ (0.03 * -0.21)+ (0.29 * -0.75)+ (0.25 * -0.96)+
(0.06 * -0.6)+ (-0.04 * 1.27)+ (0.32 * -0.16)+ (0.32 * -0.12)+ (-0.18 *
1.01)+ (0.21 * -0.87)+ (0.32 * -0.5)+ (0.25 * -0.32)+
```

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?

Below is the cumulative of eigenvalues output in python code:

```
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.        ])
```

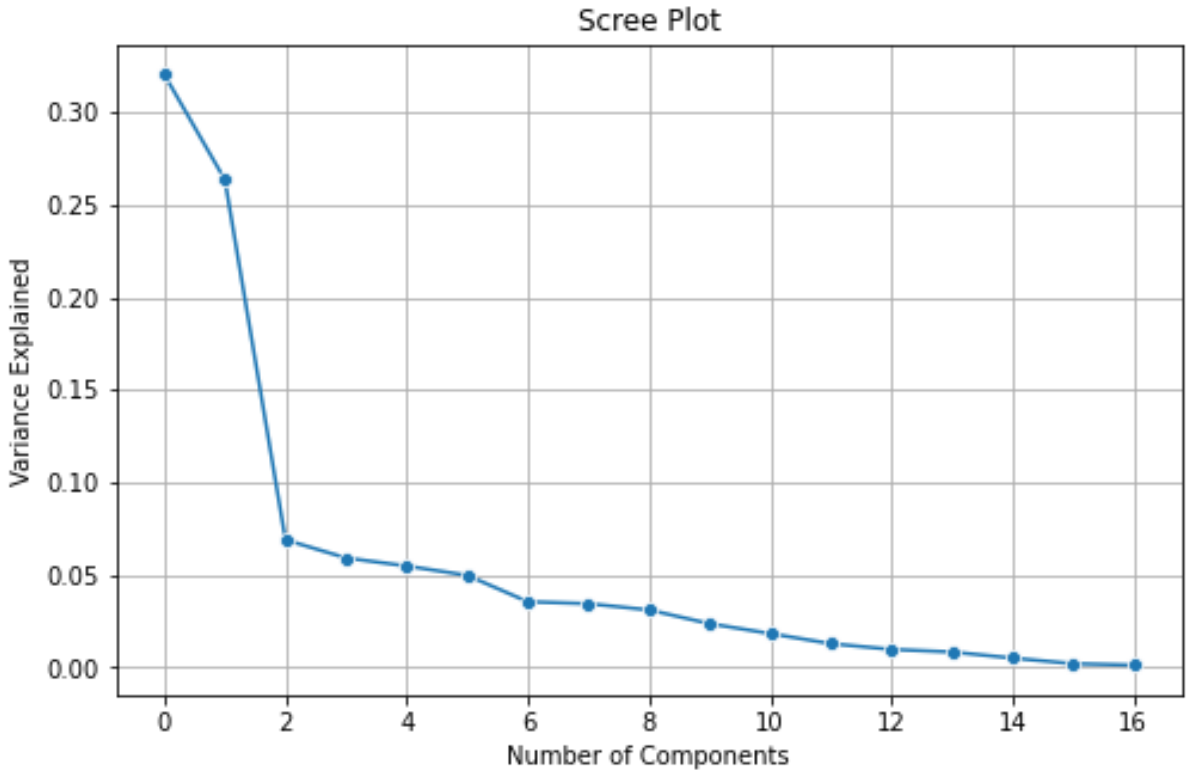

Figure 9: Scree Plot

Observation:

By taking into account the cumulative explained variance ratio with a specific confidence interval, the cumulative values of eigen values help us in determining the ideal number of principal components.

Because we set the confidence level in this scenario at 85%, we use 7 major components.

Let us now plot another graph with both the variances explained by each eigen value and the Cumulative Variance explained

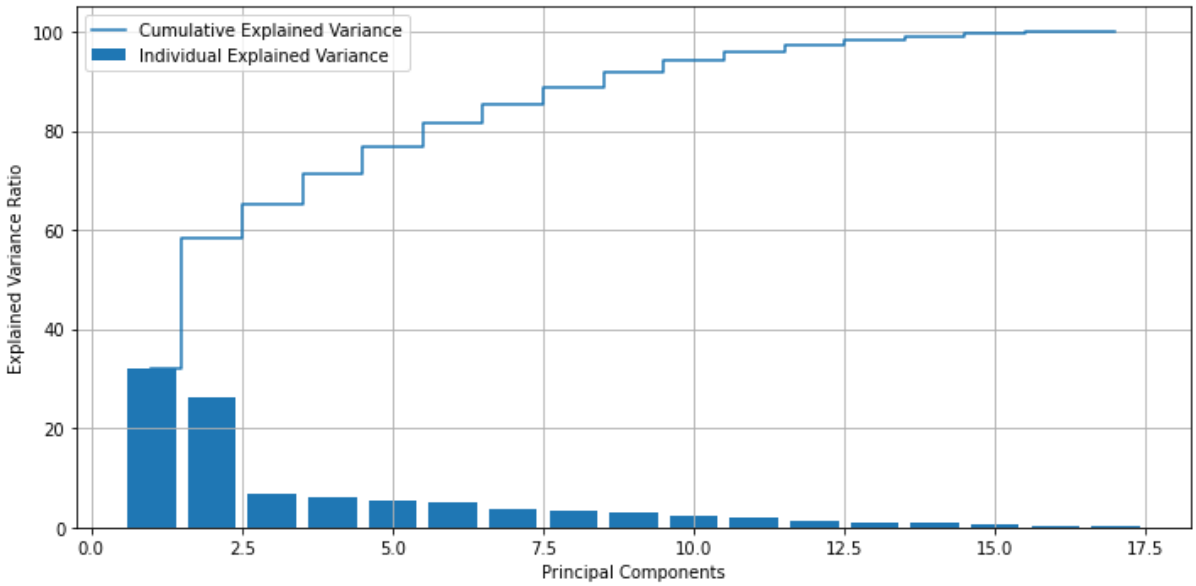

Figure 10: Bar & Step Plot

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]

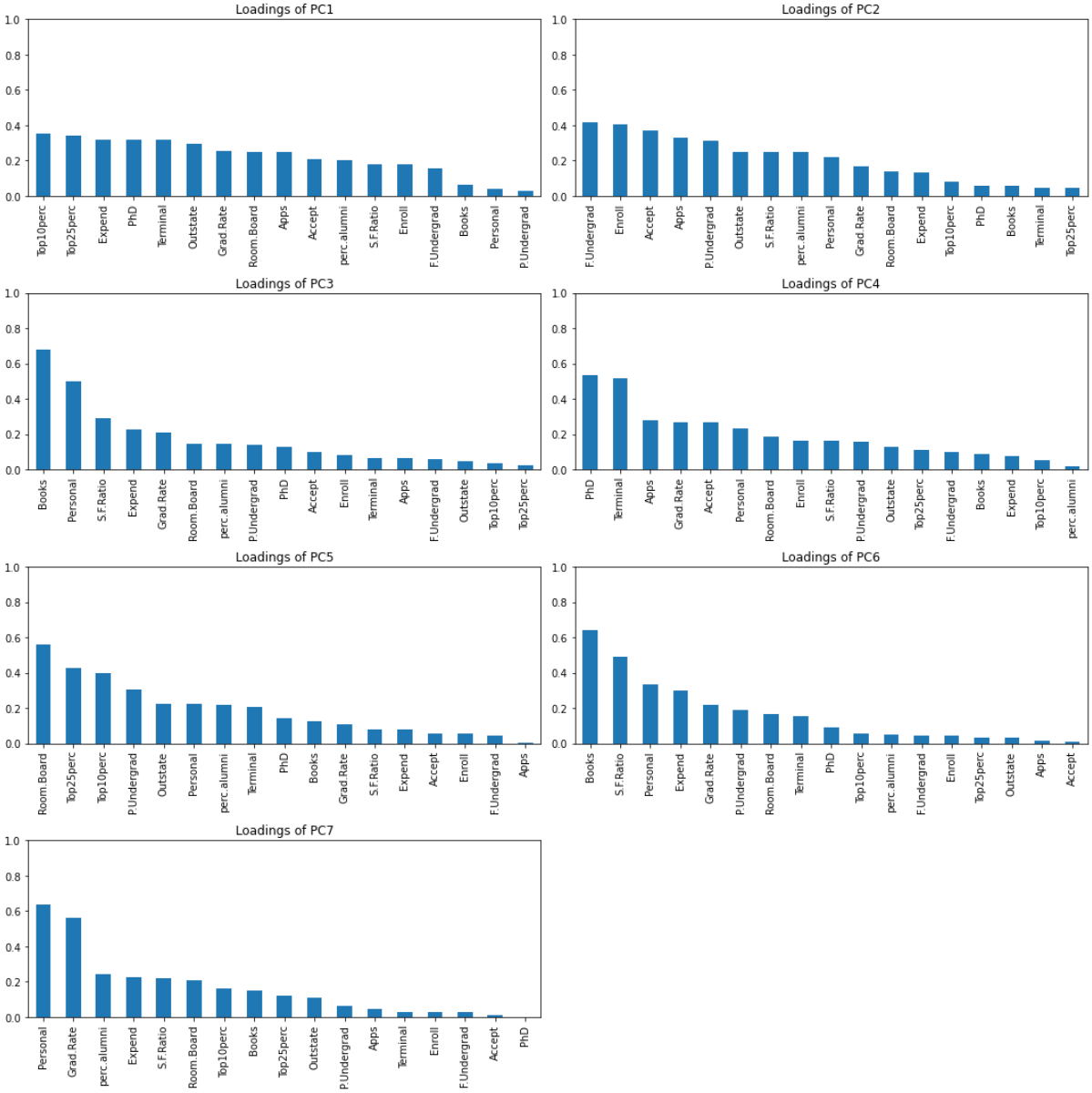

Figure 11: Loadings of 7 PCs

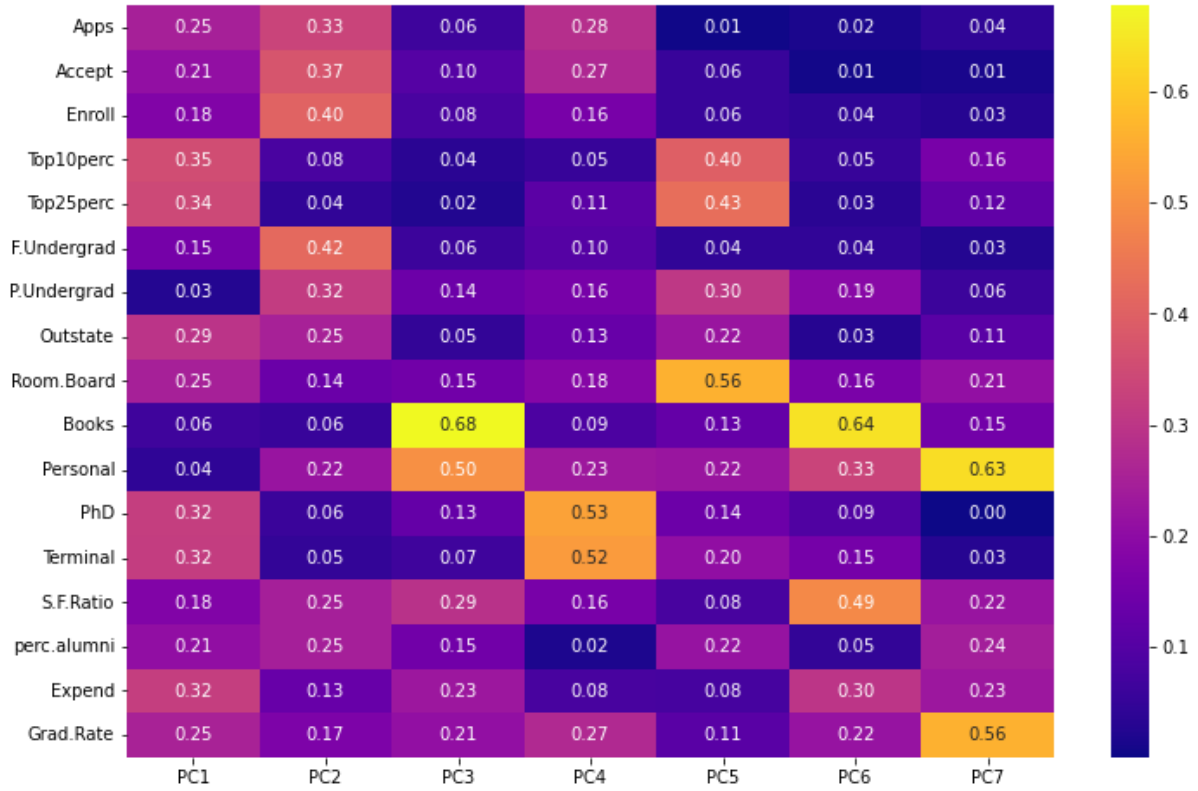

Figure 12: Heatmap of 7 PCs

Observations:

a) Principal Component Analysis in this case study reduced the dimensionality of the dataset from 17 to 7 as it gives us better perspective and less complexity.

b) It helps in minimizing redundant data and helps in refining useful data as when we use process-intensive algorithms (like many supervised algorithms) on the data so we need to get rid of redundancy.

c) PCA gave us linearly independent and different combinations of features which we can further to describe our data differently as it gives a whole new perspective.

7 Principal Components are enough to perform further analysis (as they cover 85% of variance of the dataset).