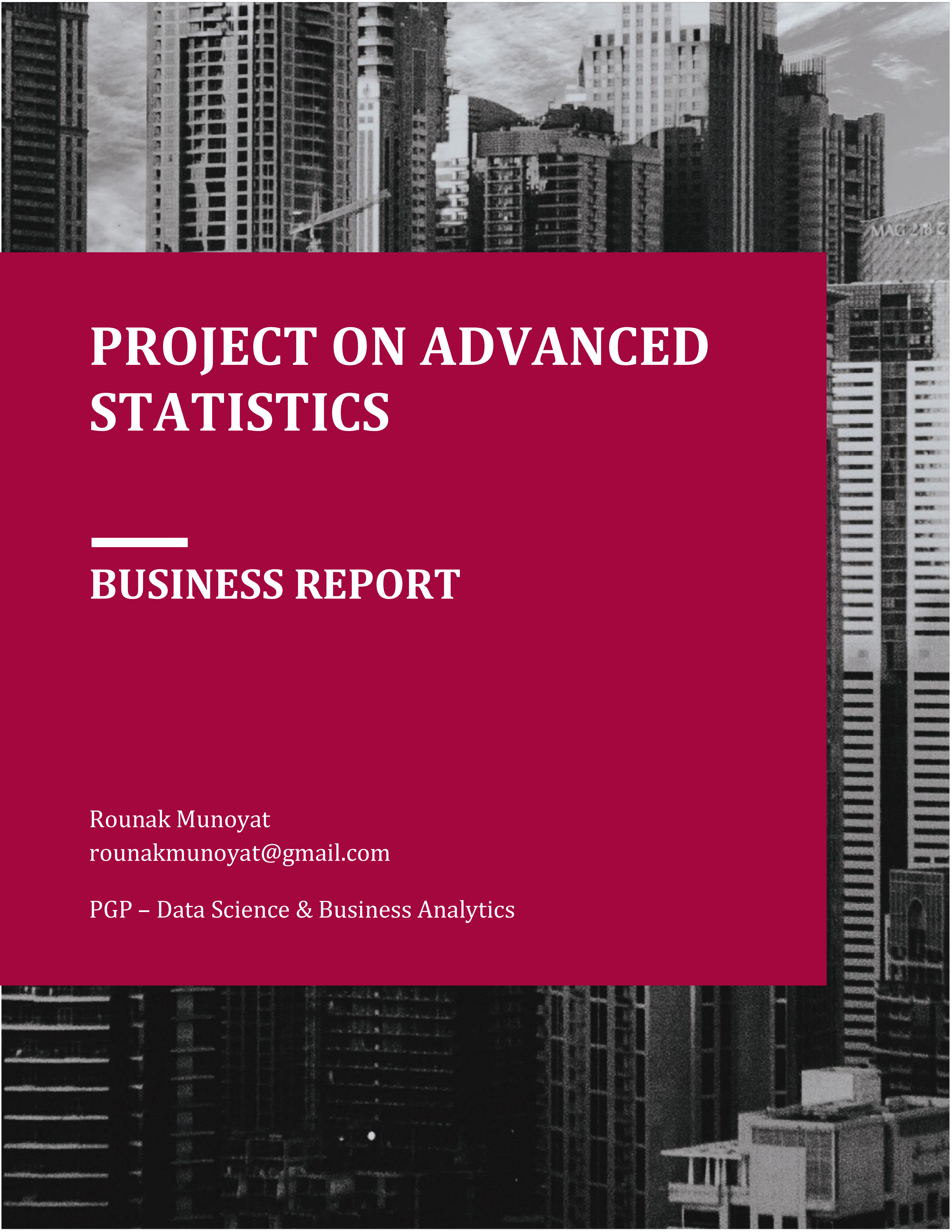


PROJECT ON ADVANCED STATISTICS

BUSINESS REPORT

Rounak Munoyat
rounakmunoyat@gmail.com

PGP – Data Science & Business Analytics

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 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.

Q1.1) State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually.

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

H_0: The mean salary of individuals is same for all 3 levels of Education.
H_A: For at least one level of Education, mean salary of individuals is different.

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

H_0: The mean salary of individuals is same for all 4 levels of Occupation.
H_A: For at least one level of Occupation, mean salary of individuals is different.

Where,
H_0 = Null Hypothesis
H_A = Alternate Hypothesis

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

- Each group sample is drawn from a normally distributed population
- All populations have a common variance
- All samples are drawn independently of each other
- Within each sample, the observations are sampled randomly and independently of each other
- Factor effects are additive

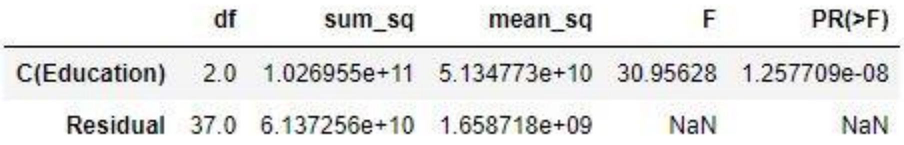
Q1.2) Perform one-way ANOVA for Education with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

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' -
H_0: The mean salary of individuals is same for all 3 levels of Education.
H_A: 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'

From above, we can say that the corresponding p-value is less than alpha (0.05). Thus, we reject the Null Hypothesis and accept the alternate hypothesis.

Therefore, for at least one level of Education, mean salary of individuals is different.

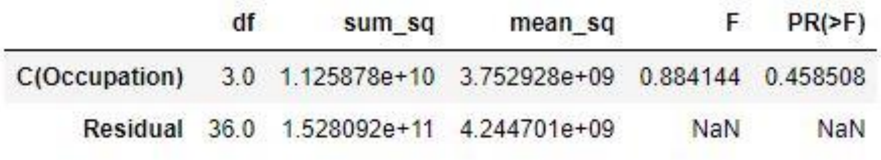
Q1.3) Perform one-way ANOVA for Occupation with respect to the variable ‘Salary’. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

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'
H_0: The mean salary of individuals is same for all 4 levels of Occupation.
H_A: 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’

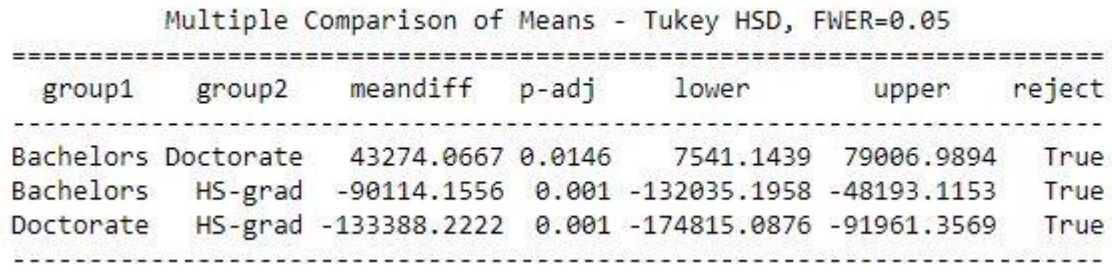
From above, we can say that the corresponding p-value is greater than alpha (0.05). Thus, we fail to reject the Null hypothesis.

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

Q1.4) If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result.

We know the null hypothesis is rejected in (Q1.2). We can find the difference in class means using two methods –

- 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    7541.1439  79006.9894   True
Bachelors    HS-grad  -90114.1556  0.001 -132035.1958 -48193.1153   True
Doctorate    HS-grad -133388.2222  0.001 -174815.0876 -91961.3569   True
-----------------------------------------------------------------------
```

Table 3: Tukey Honest Significance Test

From the above table, we can say that:

- The mean salary of individuals who are Doctorate & Bachelors is very large compared to those who are HS-grad.
- There is a moderate difference in mean salaries of individuals who are Doctorates and Bachelors.

- Point Plot of Education vs Salary

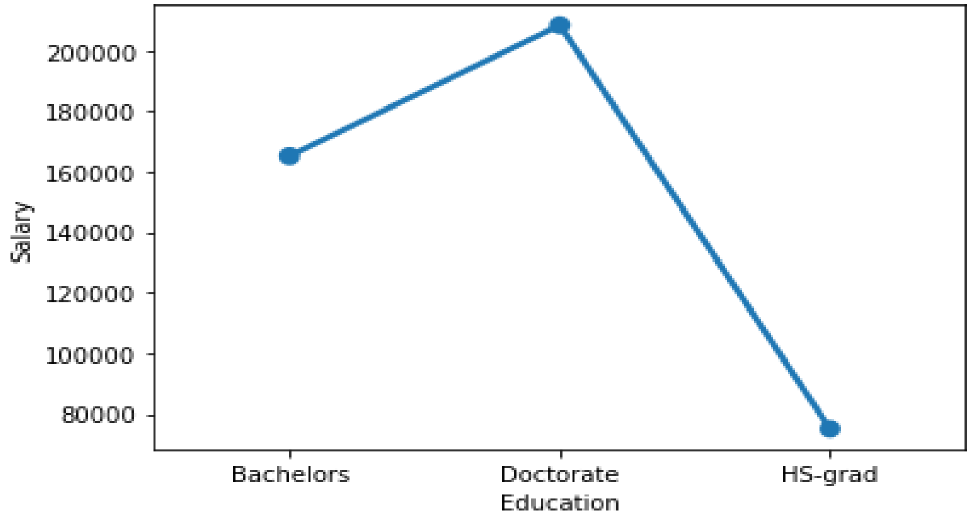


Figure 1: Plot of Education vs Salary

From the above graph, we can say that:

- The mean salary of individuals who are Doctorate & Bachelors is very large compared to those who are HS-grad.
- There is a moderate difference in mean salaries of individuals who are Doctorates and Bachelors.

Q1.5) What is the interaction between the two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.

The interaction between two treatments (categorical variables Occupation & Education in this case) with respect to continuous measure (Salary variable in this case) is said to exist, if, the response of continuous measure to one categorical variable depends on another categorical variable.

- Interaction effects represent the combined effects of factors on the dependent measure.
- When an interaction effect is present, the impact of one factor depends on the level of the other factor.
- When interaction effects are present, it means that interpretation of the main effects is incomplete or misleading.

Interaction 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.

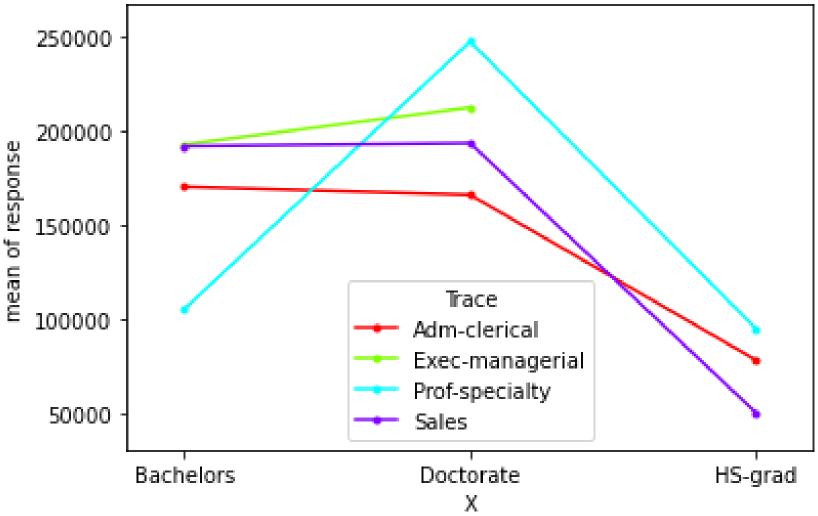


Figure 2: Interaction Plot between Education & Occupation variables

From the graph above, we can say that:

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

Q1.6) Perform a two-way ANOVA based on the Education and Occupation (along with their interaction Education*Occupation) with the variable 'Salary'. State the null and alternative hypotheses and state your results. How will you interpret this result?

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

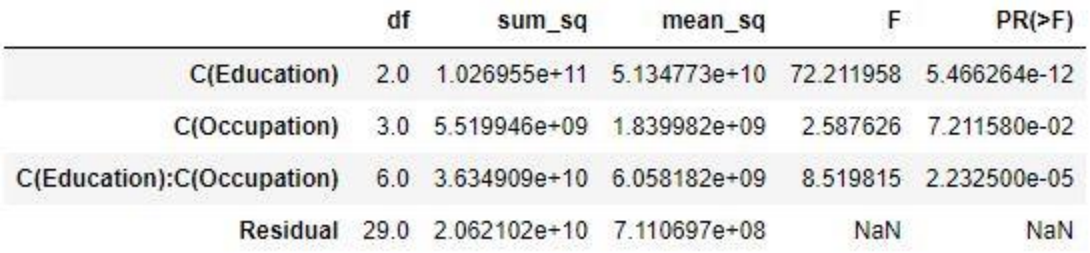
Null Hypothesis (H_0):

- There is no difference in mean salary of individuals at any level of Education.
- There is no difference in mean salary of individuals for any type of Occupation.
- There is no interaction effect between Education and Occupation on average salary.

Alternate Hypothesis (H_A):

- There is a difference in mean salary of individuals at any level of Education.
- There is a difference in mean salary of individuals for any type of Occupation.
- There is an interaction effect between Education and Occupation on average salary.

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

Since, One-Way ANOVA has already been performed individually for Education & Occupation variable above, we are concerned only with the results of third hypothesis here (interaction test between Education & Occupation with respect to Salary)

And, since, the p-value of the interaction effect term of 'Education' and 'Occupation' is less than 0.05 the Null Hypothesis is rejected in this case and we can accept the alternate hypothesis.

Therefore,

- There is a difference in mean salary of individuals at any level of Education.
- There is no difference in mean salary of individuals for any type of Occupation.
- 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.

Q1.7) Explain the business implications of performing ANOVA for this particular case study.

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

- Occupation type alone do not influence mean salaries of individuals.
- Educational qualification type alone on the other hand do influence mean salaries of individuals.
- However, Occupation type & Educational qualification type when combined together do influence mean salaries of individuals.
- Also, from the interaction plot we come to know that the mean salaries of individuals with high school degree is very less compared to the ones with Bachelor's degree. Also, the mean salaries of individuals with Bachelor's and Doctorates are almost same. If a business decision is to be taken based on the mean salaries for the entire population, we need to make sure the decision taken affects the entire population in a positive way and it does not lead to any kind of bias for any group.

Case Study 2- College Survey post 12th (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'.

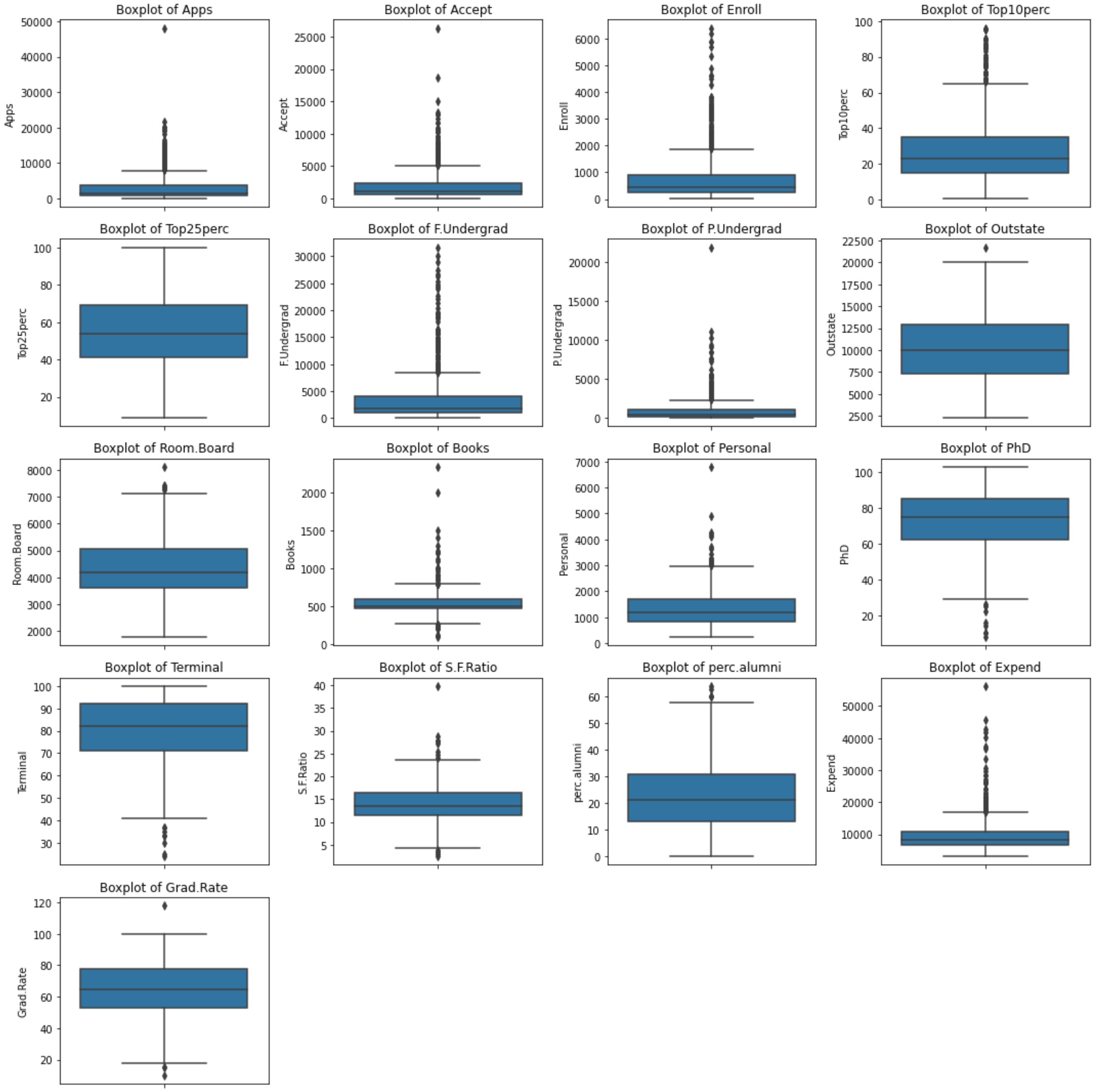
Q2.1) Perform Exploratory Data Analysis (both univariate and multivariate analysis to be performed). What insight do you draw from the EDA?

Observations of basic Data Exploration:

- Dataset has 18 columns and 777 rows.
- The entire dataset is of integer data type. However, column 'Names' is object datatype & S.F. Ratio is float datatype.
- No duplicate records.
- No null values.
- Grad.Rate has a maximum value of 118. This has to cleaned.
- We have to investigate further for outliers.

Univariate Analysis

1) Boxplot for outlier identification



Boxplot of Apps

Boxplot of Accept

Boxplot of Enroll

Boxplot of Top10perc

Boxplot of Top25perc

Boxplot of F.Undergrad

Boxplot of P.Undergrad

Boxplot of Outstate

Boxplot of Room.Board

Boxplot of Books

Boxplot of Personal

Boxplot of PhD

Boxplot of Terminal

Boxplot of S.F.Ratio

Boxplot of perc.alumni

Boxplot of Expend

Boxplot of Grad.Rate

Figure 3: Boxplot for Outlier Identification

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

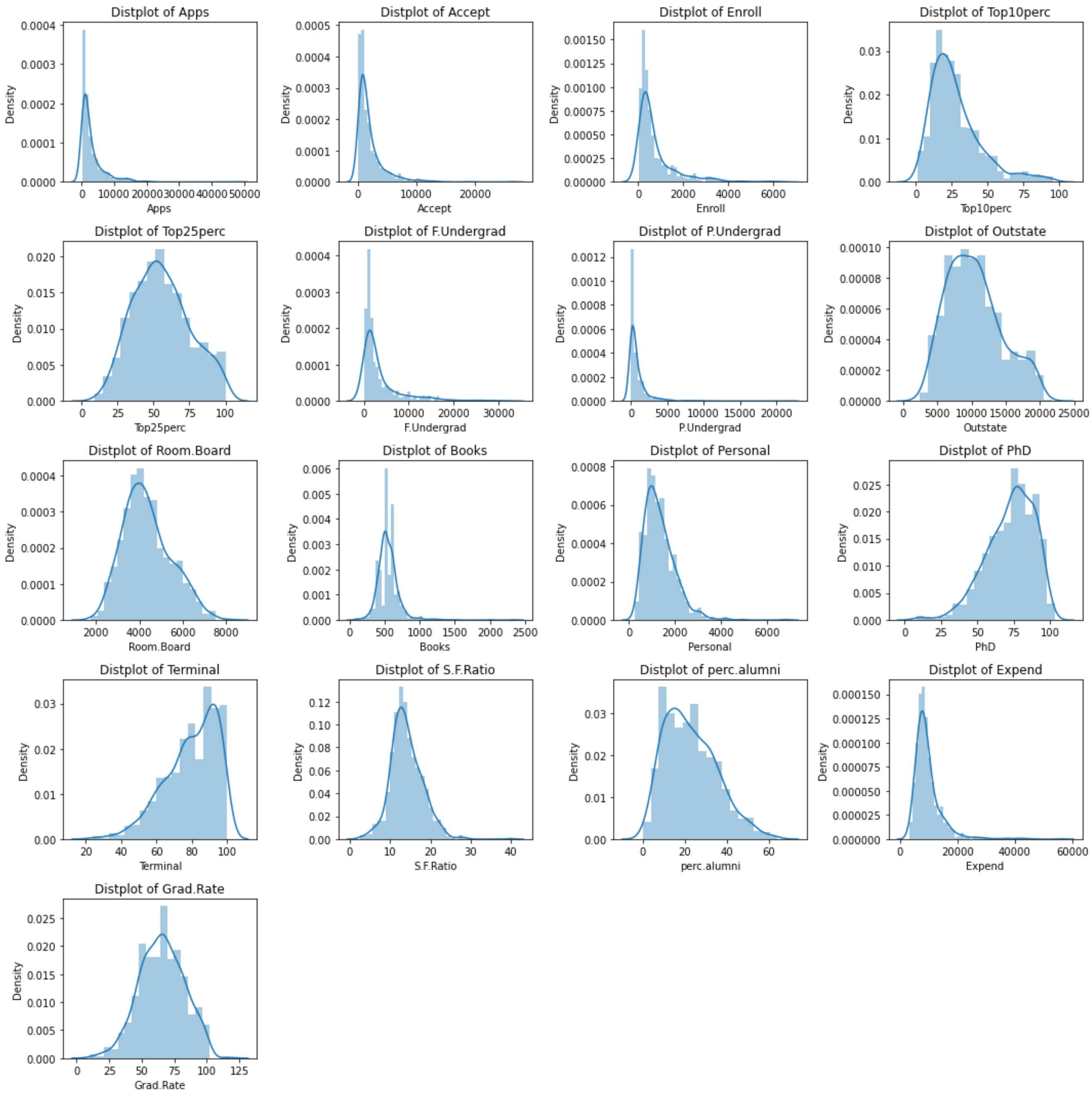


Figure 4: Distplot for studying variable distribution.

Observations:

- The variables Apps, Accept, Enroll, Top10%, F.Undergrad, P.Undergrad, Books, Personal, % Alumni and Expend have a right-skewed distribution.
- The variables PHD & Terminal have a left-skewed distribution.
- The variables Top25%, Outstate, RoomBoard, S.F.Ratio, Grad Rate have a normal distribution.

Multivariate Analysis

1) Heatmap to study correlation between variables

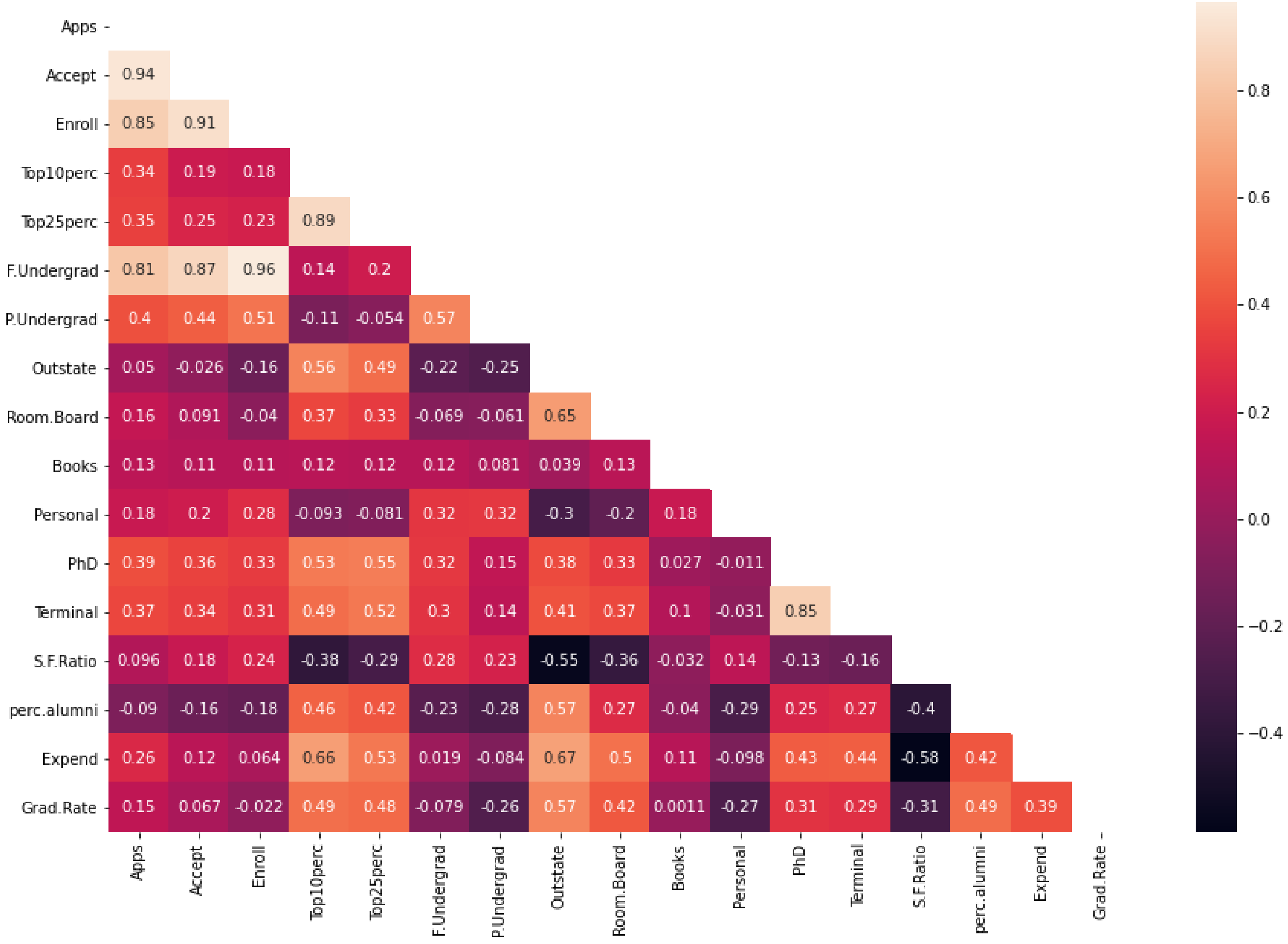


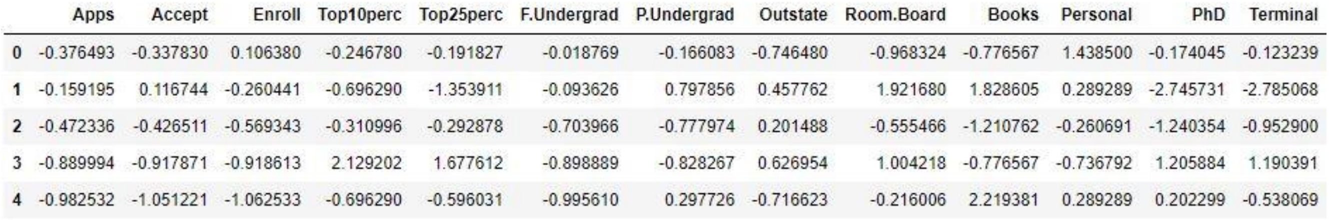
Figure 5: Correlation Heatmap

Observations:

- Variable Apps, Accept, Enroll & F.Undergrad have a strong correlation with one another.
- Variable Top10% has a very strong correlation with Top25%.
- Variable PhD has a strong correlation with Terminal.

Q2.2) Is scaling necessary for PCA in this case? Give justification and perform scaling.

- Yes, it is necessary to normalize data before performing PCA in this case because there are some variables which are in percentages and some are in counts (number of students).
- The PCA calculates a new projection of the dataset and the new axis is based on the standard deviation of our variables. So, a variable with a high standard deviation will have a higher weight for the calculation of axis than a variable with a low standard deviation. If we normalize your data, all variables will have the same standard deviation, thus all variables will have the same weight and our PCA calculates relevant axis.

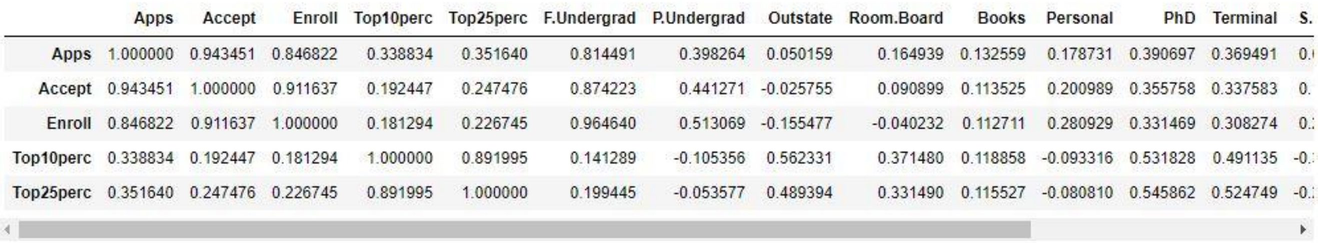


	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
0	-0.376493	-0.337830	0.106380	-0.246780	-0.191827	-0.018769	-0.166083	-0.746480	-0.968324	-0.776567	1.438500	-0.174045	-0.123239
1	-0.159195	0.116744	-0.260441	-0.696290	-1.353911	-0.093626	0.797856	0.457762	1.921680	1.828605	0.289289	-2.745731	-2.785068
2	-0.472336	-0.426511	-0.569343	-0.310996	-0.292878	-0.703966	-0.777974	0.201488	-0.555466	-1.210762	-0.260691	-1.240354	-0.952900
3	-0.889994	-0.917871	-0.918613	2.129202	1.677612	-0.898889	-0.828267	0.626954	1.004218	-0.776567	-0.736792	1.205884	1.190391
4	-0.982532	-1.051221	-1.062533	-0.696290	-0.596031	-0.995610	0.297726	-0.716623	-0.216006	2.219381	0.289289	0.202299	-0.538069

Table 5: Sample of data after scaling

Q2.3) Comment on the comparison between the covariance and the correlation matrices from this data. (on scaled data)

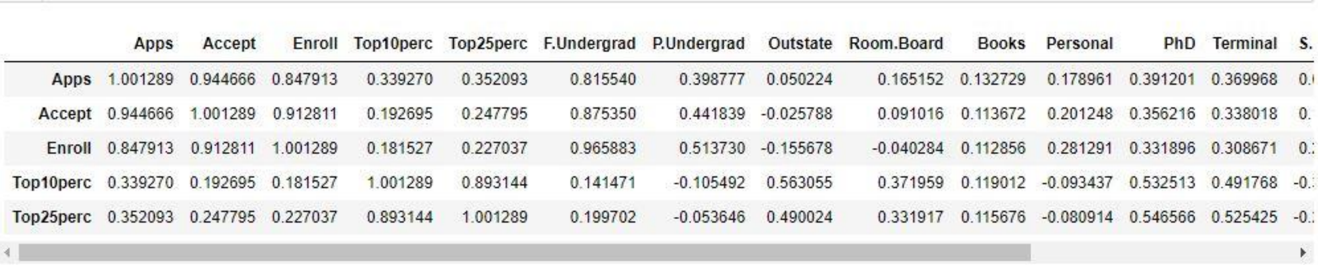
Correlation Matrix:



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369491	0.
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758	0.337583	0.
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469	0.308274	0.
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828	0.491135	-0.
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862	0.524749	-0.

Table 6: Correlation Matrix of the scaled dataset

Covariance Matrix



	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.
Apps	1.001289	0.944666	0.847913	0.339270	0.352093	0.815540	0.398777	0.050224	0.165152	0.132729	0.178961	0.391201	0.369968	0.
Accept	0.944666	1.001289	0.912811	0.192695	0.247795	0.875350	0.441839	-0.025788	0.091016	0.113672	0.201248	0.356216	0.338018	0.
Enroll	0.847913	0.912811	1.001289	0.181527	0.227037	0.965883	0.513730	-0.155678	-0.040284	0.112856	0.281291	0.331896	0.308671	0.
Top10perc	0.339270	0.192695	0.181527	1.001289	0.893144	0.141471	-0.105492	0.563055	0.371959	0.119012	-0.093437	0.532513	0.491768	-0.
Top25perc	0.352093	0.247795	0.227037	0.893144	1.001289	0.199702	-0.053646	0.490024	0.331917	0.115676	-0.080914	0.546566	0.525425	-0.

Table 7: Covariance Matrix of the scaled dataset

From the above two tables we can say that there is a slight or negligible difference in their values. This is because the dataset considered to calculate covariance & correlation is already scaled. If the dataset wouldn't have been scaled the covariance matrix would have differed a lot.

Both Correlation and Covariance are very closely related to each other and yet they differ a lot. When it comes to choosing between Covariance vs Correlation, the latter stands to be the first choice as it remains unaffected by the change in dimensions, location, and scale, and can also be used to make a comparison between two pairs of variables.

Q2.4) Check the dataset for outliers before and after scaling. What insight do you derive here?

Outliers in the dataset before scaling:

Figure 3 shows the outliers before scaling

Outliers in the dataset after scaling:

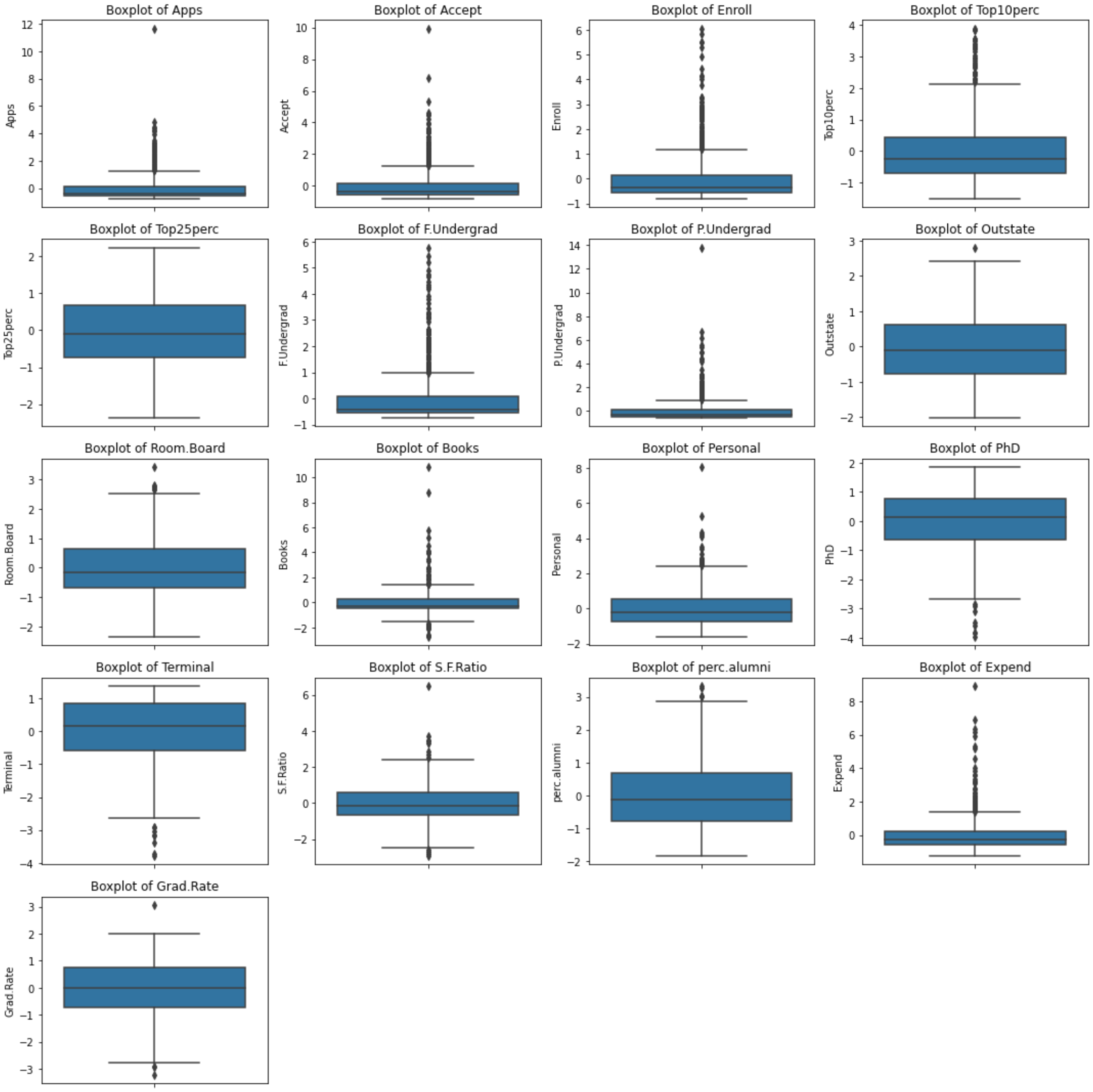


Figure 6: Outliers after scaling the dataset

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 the presence of outliers in the dataset or treats them in any way.

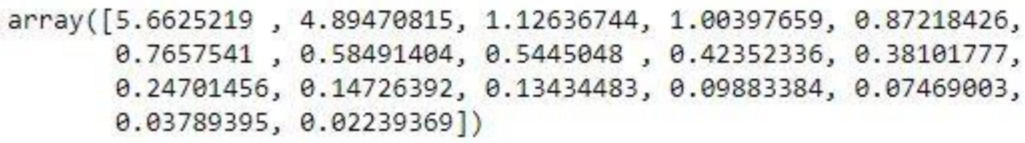
Q2.5) Extract the eigenvalues and eigenvectors. (Using Sklearn PCA print both)

Eigen vectors:
Below is the output of python code:

```
array([[ 2.62171542e-01,  2.30562461e-01,  1.89276397e-01,
         3.38874521e-01,  3.34690532e-01,  1.63293010e-01,
         2.24797091e-02,  2.83547285e-01,  2.44186588e-01,
         9.67082754e-02, -3.52299594e-02,  3.26410696e-01,
         3.23115980e-01, -1.63151642e-01,  1.86610828e-01,
         3.28955847e-01,  2.38822447e-01],
       [ 3.14136258e-01,  3.44623583e-01,  3.82813322e-01,
        -9.93191661e-02, -5.95055011e-02,  3.98636372e-01,
         3.57550046e-01, -2.51863617e-01, -1.31909124e-01,
         9.39739472e-02,  2.32439594e-01,  5.51390195e-02,
         4.30332048e-02,  2.59804556e-01, -2.57092552e-01,
        -1.60008951e-01, -1.67523664e-01],
       [-8.10177245e-02, -1.07658626e-01, -8.55296892e-02,
         7.88293849e-02,  5.07938247e-02, -7.37077827e-02,
        -4.03568700e-02, -1.49394795e-02,  2.11379165e-02,
         6.97121128e-01,  5.30972806e-01, -8.11134044e-02,
        -5.89785929e-02, -2.74150657e-01, -1.03715887e-01,
         1.84205687e-01, -2.45335837e-01],
       [ 9.87761685e-02,  1.18140437e-01,  9.30717094e-03,
        -3.69115031e-01, -4.16824361e-01,  1.39504424e-02,
         2.25351078e-01,  2.62975384e-01,  5.80894132e-01,
        -3.61562884e-02, -1.14982973e-01, -1.47260891e-01,
        -8.90079921e-02, -2.59486122e-01, -2.23982467e-01,
         2.13756140e-01, -3.61915064e-02],
       [ 2.19898081e-01,  1.89634940e-01,  1.62314818e-01,
         1.57211016e-01,  1.44449474e-01,  1.02728468e-01,
        -9.56790178e-02,  3.72750885e-02, -6.91080879e-02,
         3.54056654e-02, -4.75358244e-04, -5.50786546e-01,
        -5.90407136e-01, -1.42842546e-01,  1.28215768e-01,
        -2.24240837e-02,  3.56843227e-01],
       [ 2.18800617e-03, -1.65212882e-02, -6.80794143e-02,
        -8.88656824e-02, -2.76268979e-02, -5.16468727e-02,
        -2.45375721e-02, -2.03860462e-02,  2.37267409e-01,
         6.38604997e-01, -3.81495854e-01,  3.34444832e-03,
         3.54121294e-02,  4.68752604e-01,  1.25669415e-02,
        -2.31562325e-01,  3.13556243e-01],
       [-2.83715076e-02, -1.29584896e-02, -1.52403625e-02,
        -2.57455284e-01, -2.39038849e-01, -3.11751439e-02,
        -1.00138971e-02,  9.45370782e-02,  9.45210745e-02,
        -1.11193334e-01,  6.39418106e-01,  8.92320786e-02,
         9.16985445e-02,  1.52864837e-01,  3.91400512e-01,
        -1.50501305e-01,  4.68641965e-01],
       [-8.99498102e-02, -1.37606312e-01, -1.44216938e-01,
         2.89538833e-01,  3.45643551e-01, -1.08748900e-01,
         1.23841696e-01,  1.12721477e-02,  3.89639465e-01,
        -2.39817267e-01,  2.77206569e-01, -3.42628480e-02,
        -9.03076644e-02,  2.42807562e-01, -5.66073056e-01,
        -1.18823549e-01,  1.80458508e-01],
       [-1.30566998e-01, -1.42275847e-01, -5.08712481e-02,
         1.22467790e-01,  1.93936316e-01, -1.45452749e-03,
         6.34774326e-01,  8.36648339e-03,  2.20526518e-01,
        -2.10246624e-02, -1.73715184e-02, -1.66510079e-01,
        -1.12609034e-01,  1.53685343e-01,  5.39235753e-01,
        -2.42371616e-02, -3.15812873e-01],
       [-1.56464458e-01, -1.49209799e-01, -6.48997860e-02,
        -3.58776186e-02,  6.41786425e-03, -1.63981359e-04,
         5.46346279e-01, -2.31799759e-01, -2.55107620e-01,
         9.11624912e-02, -1.27647512e-01,  1.00975002e-01,
         8.60363025e-02, -4.70527925e-01, -1.47628917e-01,
        -8.04154875e-02,  4.88415259e-01],
       [-8.62132843e-02, -4.25899061e-02, -4.38408622e-02,
         1.77837341e-03, -1.02127328e-01, -3.49993487e-02,
         2.52107094e-01,  5.93433149e-01, -4.75297296e-01,
         4.35697999e-02,  1.51627393e-02, -3.91865961e-02,
        -8.48575651e-02,  3.63042716e-01, -1.73918533e-01,
         3.93722676e-01,  8.72638706e-02],
       [-8.99775288e-02, -1.58861886e-01,  3.53988202e-02,
         3.92277722e-02, -1.45621999e-01,  1.33555923e-01,
        -5.02487566e-02, -5.60392799e-01,  1.07365653e-01,
        -5.16224550e-02, -9.39409228e-03,  7.16590441e-02,
        -1.63820871e-01,  2.39902591e-01,  4.89753356e-02,
         6.90417042e-01,  1.59332164e-01],
       [-8.88697944e-02, -4.37945938e-02,  6.19241658e-02,
        -6.99599977e-02,  9.70282598e-02,  8.71753137e-02,
        -4.45537493e-02, -6.72405494e-02, -1.77715010e-02,
        -3.54343707e-02,  1.18604404e-02, -7.02656469e-01,
         6.62488717e-01,  4.79006197e-02, -3.58875507e-02,
         1.26667522e-01,  6.30737002e-02],
       [-5.49428396e-01, -2.91572312e-01,  4.17001280e-01,
        -8.79767299e-03,  1.07779150e-02,  5.70683843e-01,
        -1.46321060e-01,  2.11561014e-01,  1.00935084e-01,
         2.86384228e-02, -3.38197909e-02,  6.38096394e-02,
        -9.85019644e-02, -6.19970446e-02, -2.80805469e-02,
        -1.28739213e-01,  7.09643331e-03],
       [ 5.41453698e-03,  1.44582845e-02, -4.97908902e-02,
        -7.23645373e-01,  6.55464648e-01,  2.53059904e-02,
        -3.97146972e-02, -1.59275617e-03, -2.82578388e-02,
        -8.06259380e-03,  1.42590097e-03,  8.31471932e-02,
        -1.13374007e-01,  3.83160891e-03, -7.32598621e-03,
         1.45099786e-01, -3.29024228e-03],
       [ 5.99137640e-01, -6.61496927e-01, -2.33235272e-01,
        -2.21448729e-02, -3.22646978e-02,  3.67681187e-01,
        -2.62494456e-02,  8.14247697e-02, -2.67779296e-02,
        -1.04624246e-02, -4.54572099e-03, -1.25137966e-02,
         1.79275275e-02, -1.83059753e-02,  8.03169296e-05,
        -5.60069250e-02, -1.48410810e-02],
       [-1.82169814e-01,  3.91041719e-01, -7.16684935e-01,
         5.62053913e-02, -1.96735274e-02,  5.42774834e-01,
        -2.95029745e-02, -1.03393587e-03, -9.85725168e-03,
        -4.36086500e-03,  1.08725257e-02, -1.33146759e-02,
        -7.38135022e-03, -8.85797314e-03,  2.40534190e-02,
        -1.05658769e-02,  2.51028410e-03]])
```

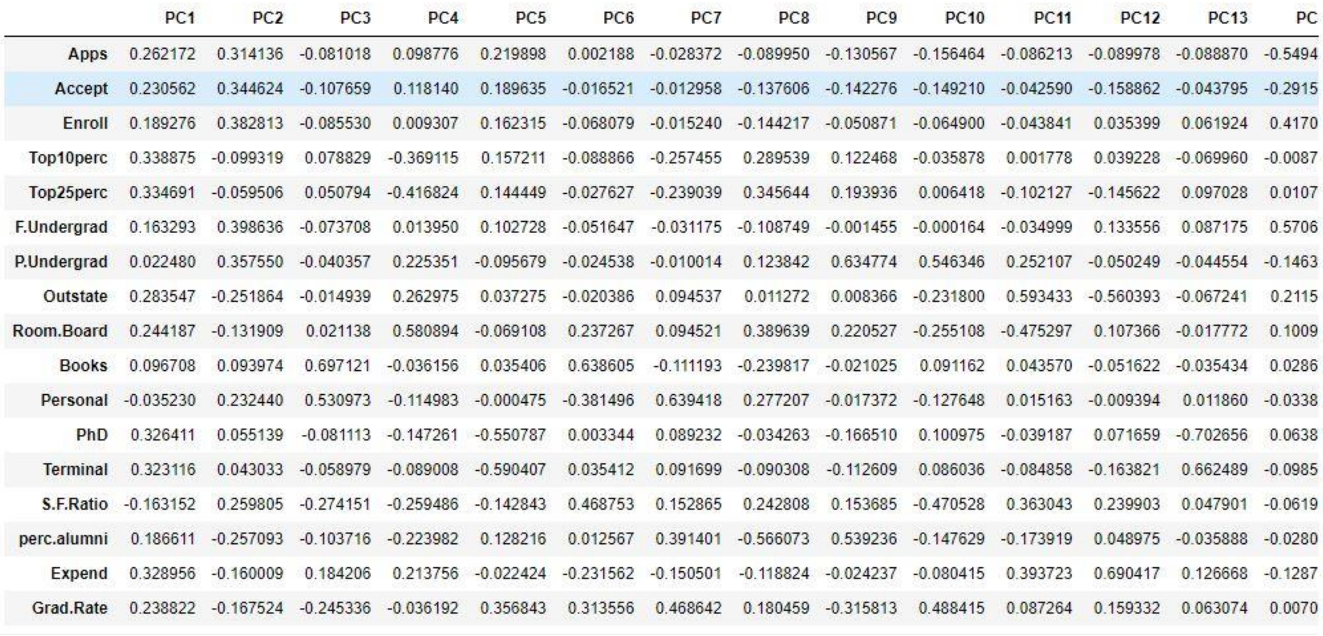
Eigen Values:

Below is the output of python code:



```
array([5.6625219 , 4.89470815, 1.12636744, 1.00397659, 0.87218426,
       0.7657541 , 0.58491404, 0.5445048 , 0.42352336, 0.38101777,
       0.24701456, 0.14726392, 0.13434483, 0.09883384, 0.07469003,
       0.03789395, 0.02239369])
```

Q2.6) Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features



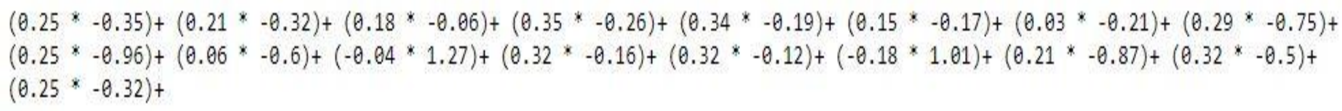
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC
Apps	0.262172	0.314136	-0.081018	0.098776	0.219898	0.002188	-0.028372	-0.089950	-0.130567	-0.156464	-0.086213	-0.089978	-0.088870	-0.5494
Accept	0.230562	0.344624	-0.107659	0.118140	0.189635	-0.016521	-0.012958	-0.137606	-0.142276	-0.149210	-0.042590	-0.158862	-0.043795	-0.2915
Enroll	0.189276	0.382813	-0.085530	0.009307	0.162315	-0.068079	-0.015240	-0.144217	-0.050871	-0.064900	-0.043841	0.035399	0.061924	0.4170
Top10perc	0.338875	-0.099319	0.078829	-0.369115	0.157211	-0.088866	-0.257455	0.289539	0.122468	-0.035878	0.001778	0.039228	-0.069960	-0.0087
Top25perc	0.334691	-0.059506	0.050794	-0.416824	0.144449	-0.027627	-0.239039	0.345644	0.193936	0.006418	-0.102127	-0.145622	0.097028	0.0107
F.Undergrad	0.163293	0.398636	-0.073708	0.013950	0.102728	-0.051647	-0.031175	-0.108749	-0.001455	-0.000164	-0.034999	0.133556	0.087175	0.5706
P.Undergrad	0.022480	0.357550	-0.040357	0.225351	-0.095679	-0.024538	-0.010014	0.123842	0.634774	0.546346	0.252107	-0.050249	-0.044554	-0.1463
Outstate	0.283547	-0.251864	-0.014939	0.262975	0.037275	-0.020386	0.094537	0.011272	0.008366	-0.231800	0.593433	-0.560393	-0.067241	0.2115
Room.Board	0.244187	-0.131909	0.021138	0.580894	-0.069108	0.237267	0.094521	0.389639	0.220527	-0.255108	-0.475297	0.107366	-0.017772	0.1009
Books	0.096708	0.093974	0.697121	-0.036156	0.035406	0.638605	-0.111193	-0.239817	-0.021025	0.091162	0.043570	-0.051622	-0.035434	0.0286
Personal	-0.035230	0.232440	0.530973	-0.114983	-0.000475	-0.381496	0.639418	0.277207	-0.017372	-0.127648	0.015163	-0.009394	0.011860	-0.0338
PhD	0.326411	0.055139	-0.081113	-0.147261	-0.550787	0.003344	0.089232	-0.034263	-0.166510	0.100975	-0.039187	0.071659	-0.702656	0.0638
Terminal	0.323116	0.043033	-0.058979	-0.089008	-0.590407	0.035412	0.091699	-0.090308	-0.112609	0.086036	-0.084858	-0.163821	0.662489	-0.0985
S.F.Ratio	-0.163152	0.259805	-0.274151	-0.259486	-0.142843	0.468753	0.152865	0.242808	0.153685	-0.470528	0.363043	0.239903	0.047901	-0.0619
perc.alumni	0.186611	-0.257093	-0.103716	-0.223982	0.128216	0.012567	0.391401	-0.566073	0.539236	-0.147629	-0.173919	0.048975	-0.035888	-0.0280
Expend	0.328956	-0.160009	0.184206	0.213756	-0.022424	-0.231562	-0.150501	-0.118824	-0.024237	-0.080415	0.393723	0.690417	0.126668	-0.1287
Grad.Rate	0.238822	-0.167524	-0.245336	-0.036192	0.356843	0.313556	0.468642	0.180459	-0.315813	0.488415	0.087264	0.159332	0.063074	0.0070

Table 8: Sample of Dataframe of PCA data exported into one with original features

Q2.7) Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).

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

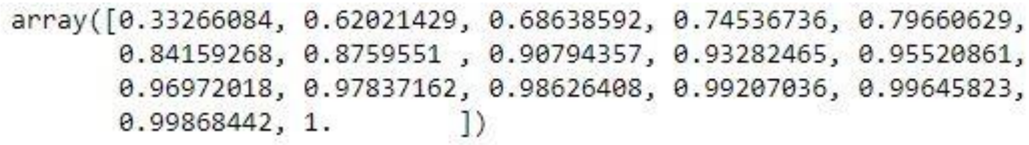
Below is the output from python code



```
(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)+
```

Q2.8) 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 output of python code:



```
array([0.33266084, 0.62021429, 0.68638592, 0.74536736, 0.79660629,
       0.84159268, 0.8759551 , 0.90794357, 0.93282465, 0.95520861,
       0.96972018, 0.97837162, 0.98626408, 0.99207036, 0.99645823,
       0.99868442, 1.        ])
```

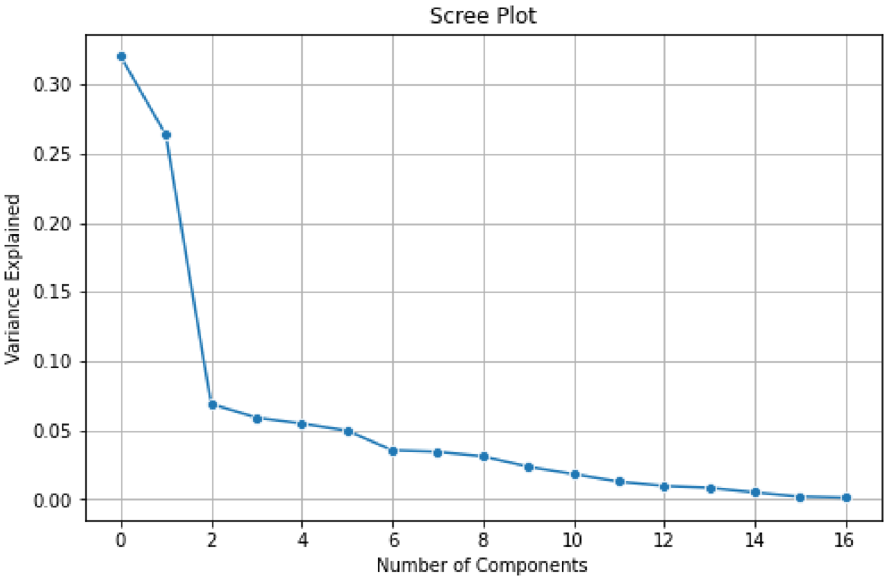


Figure 7: Scree Plot

Cumulative values of eigen values help us decide the optimum number of principal components by considering the cumulative explained variance ratio with a certain confidence interval.

In this case, we take confidence level as 85%, hence, we use 7 principal components.

The Eigenvectors indicate the direction of the principal components (new axes)

Q2.9) Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis?

- 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.
- 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.
- 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).