Advanced Statistics Project (DSBA)-17th July, 2022

Submitted By,

Deepa. K

Table of Contents

List of Figures:

Introduction:

The business report explains two different problems based on ANOVA and EDA. The exploration of data set using various attributes helps in analyzing the required information provided.

Problem Statement 1:

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

Data Description:

- Education: Bachelors/Doctorate/HS-Grad
- Occupation: Adm.-clerical/Sales/Prof-Specialty/Exec-Managerial
- Salary: Range from 50k to 2.6L

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

Solution:

Education:

- Null hypothesis (H_0): Mean salaries of all individuals are same at all 3 levels of Education.
- Alternate hypothesis (Ha): For at least one level of Education, mean salaries of individuals are different.

Occupation:

- Null hypothesis (H_0): Mean salaries of all individuals are same at all 4 levels of Occupation.

- Alternate hypothesis (Ha): For at least one level of Occupation, mean salaries of individuals are different.

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

Solution:

Null hypothesis (H_0): Mean salaries of all individuals are same at all 3 levels of Education.

Alternate hypothesis (Ha): For at least one level of Education, mean salaries of individuals are different.

	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

Figure 4:One-way ANOVA (Edu)

Since the p value 1.257709e-^{08} is less than alpha ($\alpha = 0.05$), we reject the null hypothesis (H_0).

Inference:

There is a significant difference in the mean salaries for at least one category of education.

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

Solution:

Null hypothesis (H_0): Mean salaries of all individuals are same at all 4 levels of Occupation.

Alternate hypothesis (Ha): For at least one level of Occupation, mean salaries of individuals are different.

	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

Figure 5:One-way ANOVA (Occ)

Since the p value 0.458508 is greater than alpha ($\alpha = 0.05$), we fail to reject the null hypothesis (H_0).

Inference:

There is no significant difference in the mean salaries across the 4 categories of occupation.

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

Solution:

The mean salaries are different for education as shown in 1.3.The P value is lesser than alpha so we reject null hypothesis under education with respect to salary.

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

Solution:

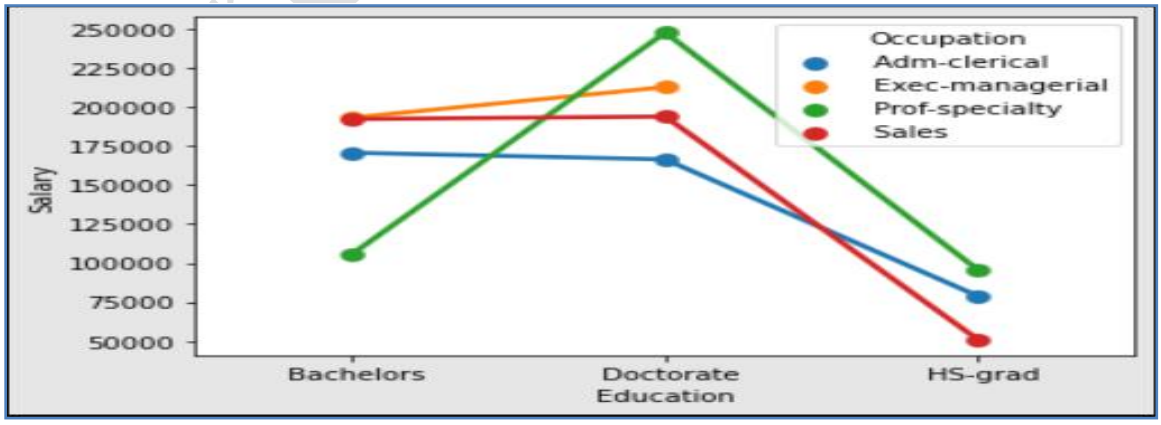

Figure 6: Edu Vs Salary

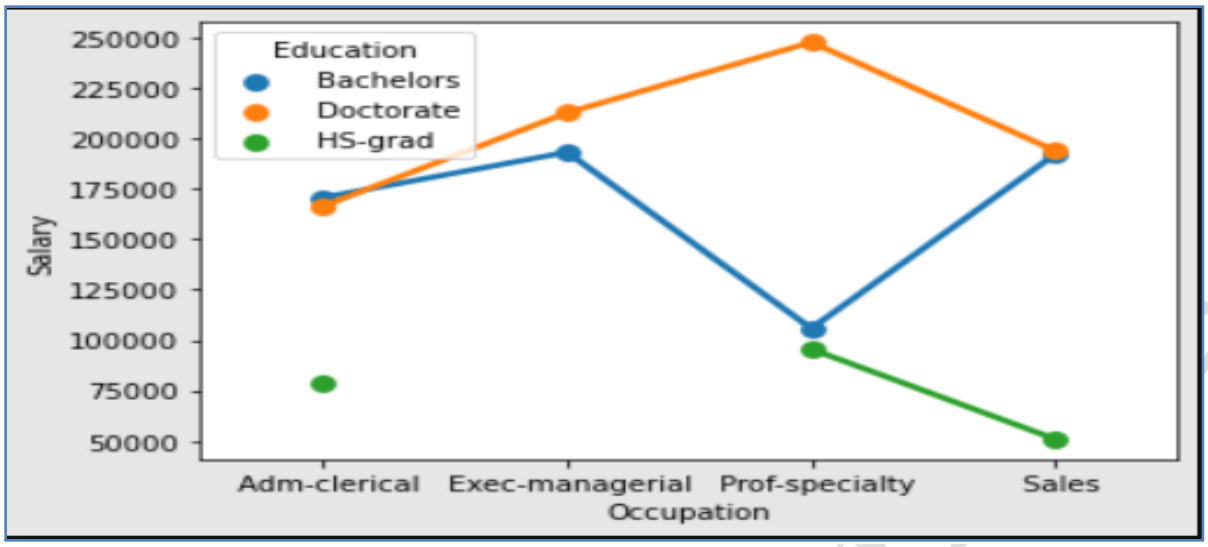

Figure 4: Occ Vs Salary

Inference:

- People with Education as Bachelors or Doctorate and Occupation as Admin-clerical and Sales almost earn the same salaries (salaries ranging from 170000–190000).
- People with Education as Bachelors and occupation as Prof-Specialty earn lesser than people with education as Bachelors and occupations as Admin-clerical and Sales.
- People with Education as HS -Grad earn the minimum salaries.

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

Solution:

Null hypothesis (H_0): There is no interaction effect between the 2 independent variables- Education and Occupation based on mean salaries.

Alternate hypothesis (H_a): There is an interaction effect between the 2 independent variables- Education and Occupation based on mean salaries.

	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

Figure 5: Two way ANOVA

Inference:

As p value ($2.232500e-^{05}$) is lesser than the significance level ($\alpha = 0.05$) we reject the null hypothesis. So there is an interaction effect between education and occupation based on the mean salaries.

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

Inference:

The ANOVA method and interaction plot explains the relevancy between educational qualification and their salaries earned.

It is clearly visible that people who have higher degrees (Doctorate) earn the maximum salary and people with only Higher Secondary- graduation earns the least.

So Salary is dependent based on Educational qualification and Occupation.

Problem Statement 2:

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

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

Solution:

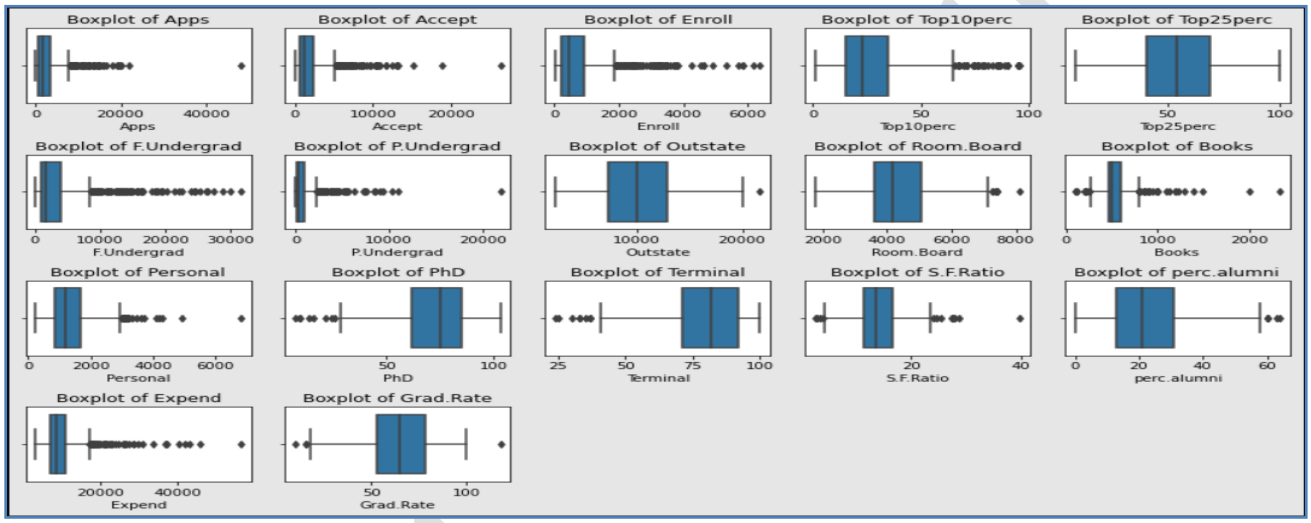

Figure 6: Univariate Analysis

Inference:

The above figure has the information about outlier presence in the dataset. Almost all the columns have outliers in it.

The column ‘Top25perc’ has no outlier and the column ‘Outstate’ has minimum outlier value.

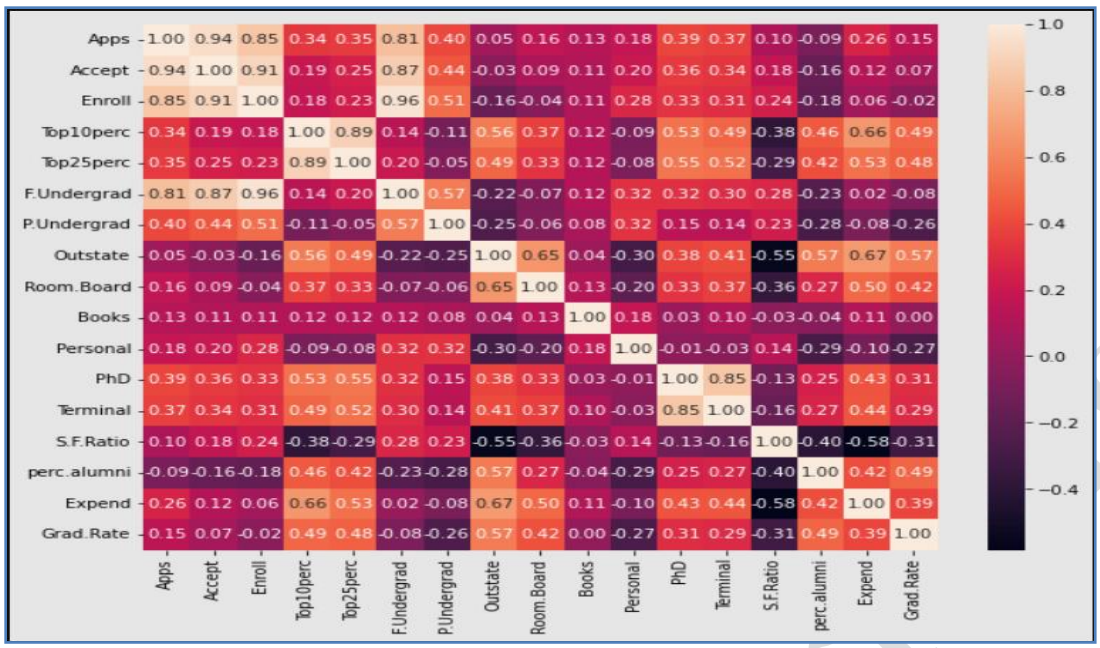

Figure 7: Bivariate Analysis

Inference:

Pair plot interpretation: The plot shows that most variables are right skewed and as scattered plot doesn't provide the correlation co-efficient, heat map is used.

The few stronger correlations are between the number of applications received and the number of applications accepted (Apps Vs Accept) & the number of applications accepted and number of new students enrolled (Accept Vs Enroll).This indicates that the number of applications accepted, received and enrolled students are highly correlated with each other.

The few weakest correlations are between the number of outstate students and Student-Faculty ratio (outstate Vs S.F ratio) & the instructional expenditure per student and Student-Faculty ratio (Expend Vs S.F ratio).This indicates that there is a higher risk of low student faculty ratio in the college.

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

Solution:

In this case scaling is necessary as it is used to change the values of numerical or categorical variables in a dataset to follow a common scale.

Here all the numerical variables are scaled with Z-score. Z-score removes the mean and scales the data to unit variance, shrinks the range from -1 to +1 values.

The column 'Names' is dropped as it is the only categorical variable which cannot be converted to Z-score / Min-max scaling.

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

Solution:

The Covariance matrix shows how two variables differ and the correlation matrix shows relation between two variables.

Pair plot interpretation: The plot shows that most variables are right skewed and as scattered plot doesn't provide the correlation co-efficient, heat map is used.

The stronger correlations are between the number of applications received and the number of applications accepted (Apps Vs Accept).

The few weakest correlations are between the number of outstate students and Student-Faculty ratio (outstate Vs S.F ratio).

Heat map shows both positive and negative correlation among the variables. The range varies from darker to lighter shade. The weakest correlation is under darker shades and the stronger correlation is under lighter shades.

Covariance	Correlation
Covariance is a measure of how closely two random variables change at the same time.	Correlation is a measure of how closely two random variables are connected.
Covariance is nothing more than a correlation measure.	The scaled version of covariance is referred to as correlation.
The direction of a linear relationship between two variables is indicated by covariance.	The intensity and direction of the linear link between two variables are measured by correlation.
Covariance can range from -∞ to +∞.	The coefficient of correlation lies between -1 to +1.
The shift in size has an impact on covariance. The covariance is modified when all of the values of one variable are multiplied by a constant, and all of the values of another variable are multiplied by a similar or different constant.	The change in scale does not affect correlation.
The units of covariance are assumed to be the product of the units of the two variables.	Correlation is a unit-free measure of the connection between variables since it is dimensionless.
The average covariance of two dependent variables measures how much they fluctuate in actual amount.	The proportion of how much two dependent variables differ on average from one another is measured by correlation.

Figure 8: Covariance and Correlation difference

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

Solution:

The dataset for outliers has been removed and used further for scaling in the following figures.

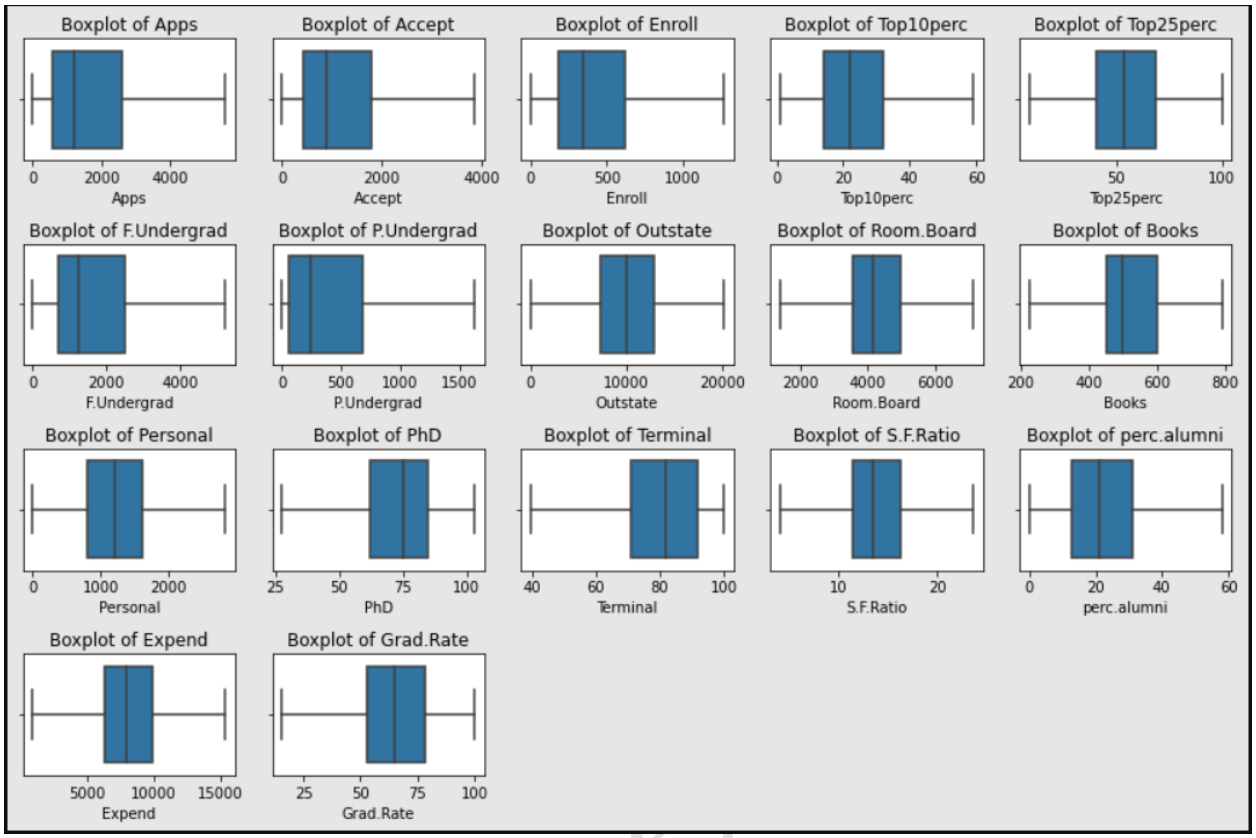

Figure 9: No outlier

	Apps	Accept	Enroll	Top10perc	Top25perc
count	777.000000	777.000000	777.000000	777.000000	777.000000
mean	2571.352638	1746.280566	660.388674	26.842986	55.796654
std	2422.195279	1523.286632	570.126836	15.582539	19.804778
min	81.000000	72.000000	35.000000	1.000000	9.000000
25%	776.000000	604.000000	242.000000	15.000000	41.000000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000
max	7896.000000	5154.000000	1892.000000	65.000000	100.000000

Figure 10: Before scaling

	Apps	Accept	Enroll	Top10perc	Top25perc
count	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02
mean	6.355797e-17	6.774575e-17	-5.249269e-17	-2.753232e-17	-1.546739e-16
std	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00
min	-7.551337e-01	-7.947645e-01	-8.022728e-01	-1.506526e+00	-2.364419e+00
25%	-5.754408e-01	-5.775805e-01	-5.793514e-01	-7.123803e-01	-7.476067e-01
50%	-3.732540e-01	-3.710108e-01	-3.725836e-01	-2.585828e-01	-9.077663e-02
75%	1.609122e-01	1.654173e-01	1.314128e-01	4.221134e-01	6.671042e-01
max	1.165867e+01	9.924816e+00	6.043678e+00	3.882319e+00	2.233391e+00

Figure 11: After scaling

Inference:

The outlier treatment before and after scaling shows the variation in the range of values (from Min to Max).

For example, the above figures (only for few variables) show the change in distribution and range of values before and after scaling the outliers using Z-score.

Z-score scaling helps in standardizing the values in same scale and using this technique helps us understand the number of standard deviations above and below the mean that each value falls.

2.5: Extract the eigen values and eigen vectors. [Using Sklearn PCA Print Both]

Solution:

Eigen values: It is the unit of variability captured by each principal component.

❶ 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])

Eigen vectors: The direction in which principal components are aligned.

❷ 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]])
```

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

Solution:

The PC's (PC1-PC17) are exported to a new dataset and a scree plot is obtained as shown below.

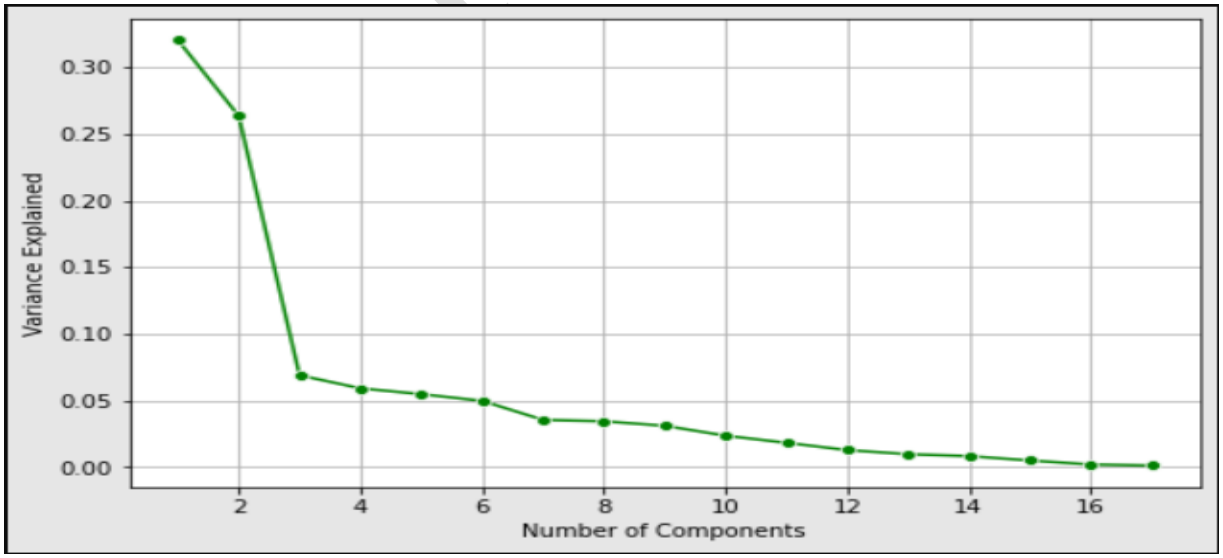

Figure 12: Scree plot

Inference:

The above graph shows the distribution of explained variance for each PC. The explained variance of each PC can be found by dividing the Eigen value of each PC by sum of Eigen values of all PC's.

PC1 holds almost 35% of data and less than 5% is captured by PC17.

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]

Solution:

Apps	0.25
Accept	0.33
Enroll	-0.06
Top10perc	0.28
Top25perc	0.01
F.Undergrad	-0.02
P.Undergrad	-0.04
Outstate	-0.10
Room.Board	-0.09
Books	0.05
Personal	0.04
PhD	0.02
Terminal	0.60
S.F.Ratio	0.08
perc.alumni	0.13
Expend	0.46
Grad.Rate	0.36

Figure 13: PC1

Linear Equation:

(0.25*Apps)+(0.33*Accept)+(-0.06*Enroll)+(0.28*Top10perc)+(0.01*Top25perc)

+(-0.02*F.undergrad)+(-0.04*P.undergrad)+(-0.10*outstate)+(-0.09*Room.board)

+(0.05*Books)+(0.04*personal)+(0.02*phd)+(0.60*terminal)+(0.08*S.F Ratio)

+(0.13*perc alumni)+(0.46*expend)+(0.36*Grad rate)

2.8: Consider the cumulative values of the eigen values. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate?

Solution:

Cumulative sum of the Eigen values:

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])

- The optimum number of principal components are decided (**PC1-PC7**) as it holds **85%** of goodness of the data. Hence the remaining PC's are dropped out.
- The Eigen vector indicates the direction in which PC's are aligned with new dimension (reduced dimension).

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]

Inference:

PCA helped to reduce 17 numeric features into 7 components which explain 85% of variance in the data.

Absolute loadings of each component with their corresponding distribution of data are explained below.

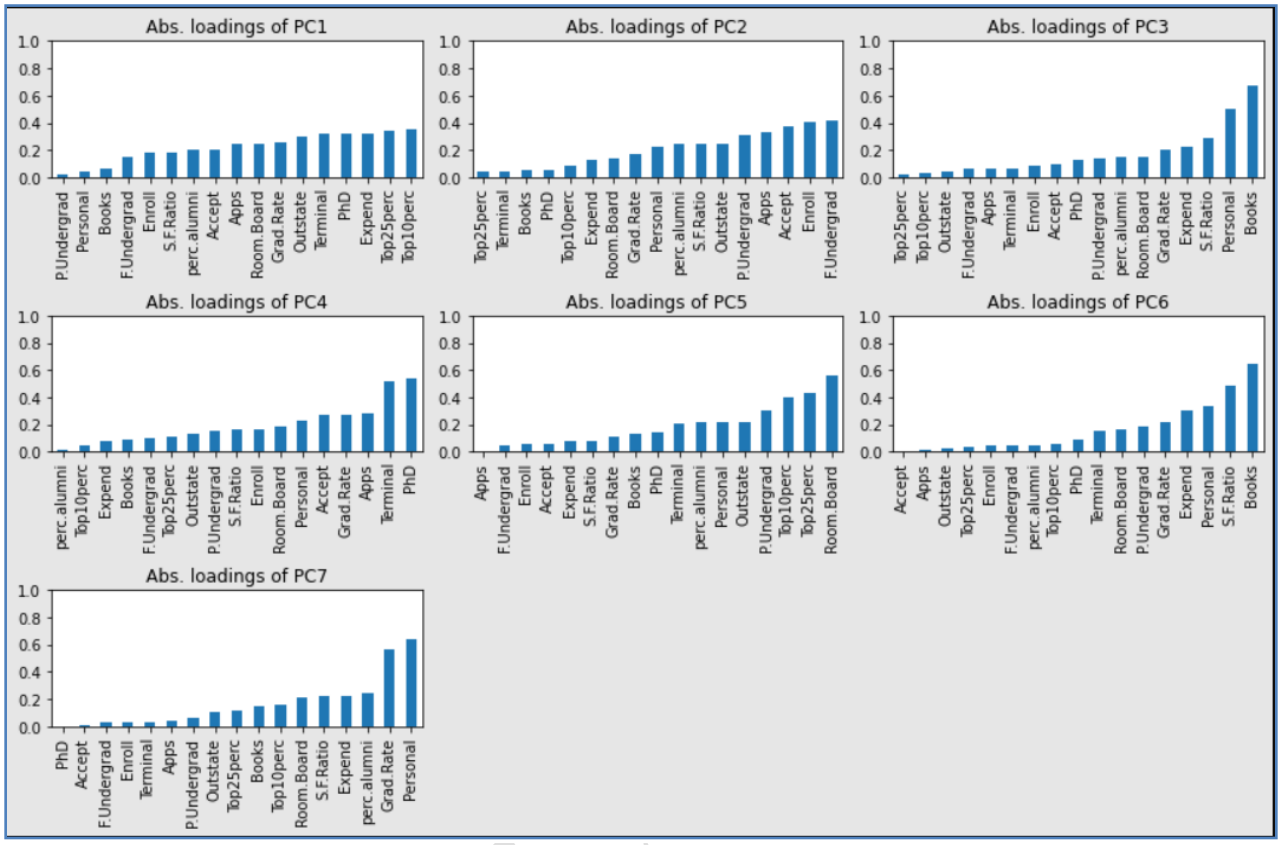

Figure 14:PC's obtained

From the above figure, the absolute characteristics of each PC's are clustered.

PC1: The Percentage of new students from top 10% of Higher Secondary class has higher value among other variables.

PC2: The full time undergraduate students hold about 50% of value.

PC3: Books hold the higher value compared to other attributes.

PC4: Percentage of faculties with PhD & terminal degree holding share almost same value.

PC5: The higher percentage (nearly 60%) is captured by the cost of Room and board.

PC6: The student-faculty ratio is comparatively higher in this component.

PC7: Estimated personal spending for a student is higher than other expenses.

Thank you!