May 2021 | PG-DSBA-Online

PROJECT REPORT
ADVANCE STATISTICS

SUBMITTED BY
DEV TRIPATHI

Contents

List of Figures

List of Tables

Problem 1A

Salary Dataset Analysis

Information about dataset-

In this dataset '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. Before diving into further analysis, an assumption is made that is the data follows a normal distribution, though in reality the normality assumption may not hold for such small sample size.

Now, let's start by preliminary exploration of the dataset.

EDA

Sample of Dataset-

	Education	Occupation	Salary
28	HS-grad	Sales	52242
6	Doctorate	Sales	237920
15	Bachelors	Adm-clerical	160910
36	Bachelors	Exec-managerial	212448
9	Doctorate	Prof-specialty	248156

Figure 1: Sample of Salary Dataset

Variable Information-

```
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Education   40 non-null     object
 1   Occupation  40 non-null     object
 2   Salary      40 non-null     int64
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

Figure 2: Variable Info of Salary Dataset

'Education' and 'Occupation' variables are object data-type variables and 'Salary' is integer data-type variable. There are total 40 entries in the dataset and none of these are null.

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

Sol:

For Education-

Null Hypothesis : H_0 = For all the education levels mean salary is equal

Alternative Hypothesis: H_1 = For at least one education level mean salary is not equal

For Occupation-

Null Hypothesis : H_0 = For all the occupations mean salary is equal

Alternative Hypothesis: H_1 = For at least one occupation mean salary is not equal

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.

Sol:

After performing one way ANOVA for Education with respect to the variable 'Salary' the following results are obtained:

	df	sum_sq	mean_sq	F	PR(>F)
Education	2	1.03E+11	5.13E+10	30.95628	1.26E-08
Residual	37	6.14E+10	1.66E+09	NaN	NaN

Table 1: One way ANOVA results for Education w.r.t. Salary

Since the p-value is less than the significance level ($\alpha = 0.05$), we can reject the null hypothesis and states that there is a difference in the mean salaries of employees with different education levels.

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.

Sol:

	df	sum_sq	mean_sq	F	PR(>F)
Occupation	3	1.13E+10	3.75E+09	0.884144	4.59E-01
Residual	36	1.53E+11	4.24E+09	NaN	NaN

Table 2: One way ANOVA for Occupation w.r.t. Salary

Since the p-value is greater than the significance level($\alpha = 0.05$), we cannot reject the null hypothesis and states that there is no difference in the mean salaries of employees with different occupations.

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.

Sol: As we can observe from the above results, for 'Education' w.r.t. 'Salary', null hypothesis can be rejected which means that for various 'Education' levels the mean 'Salary' is not equal. For checking which of these 'Education' differ significantly we can perform Tukey Test, results of which are shown here:

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

Figure 3: Tukey test results for Education w.r.t. Salary

As we can clearly observe from the above results, at the significance level of 5%, null hypothesis can be rejected for each combination that means none of means are equal to other.

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.

Sol:

Interaction occurs when the pattern of the cell means in one row (going across columns) varies from the patterns of cell means in other rows. In simple words, we can say that if the lines obtained are not parallel to each other then there is some interaction effect present.

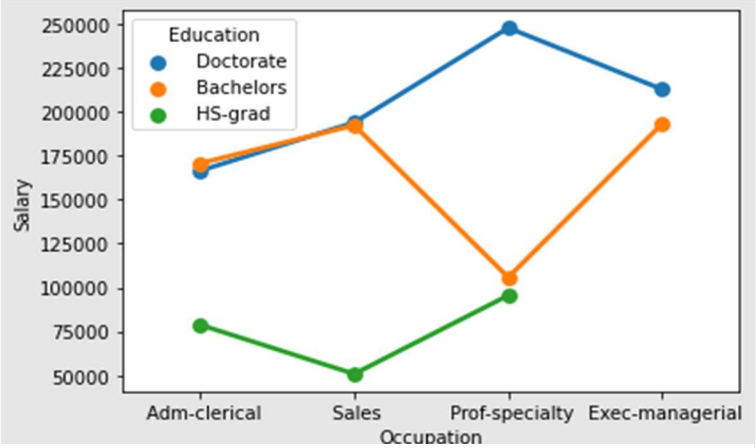

Figure 4: Interaction plot for Education and occupation variables

By using point plot function in Seaborn library we can clearly visualize that there is some kind of interaction present between the 'Education' and 'Occupation' variables. Hence, we should take this interaction effect into account while performing ANOVA for these variables.

Problem 1B

Salary Dataset Analysis

Two way ANOVA

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?

Sol: Hypotheses for Two-way ANOVA-

Occupation	Adm-clerical		Exec-managerial		Prof-specialty		Sales		All	
Education	**Count**	**Mean**	**Count**	**Mean**	**Count**	**Mean**	**Count**	**Mean**	**Count**	**Mean**
Bachelors	3	170711	4	193201.8	4	105787.8	4	192300.8	15	165152.9
Doctorate	4	166457.8	1	212781	6	247772.8	5	193916.6	16	208427
HS-grad	3	78759.67	0	NaN	3	95534.33	3	50822.33	9	75038.78
All	10	141424.3	5	197117.6	13	168953.2	12	157604.4	40	162186.9

Table 3: Salary at each combination of Education and Occupation

For Education-

Null Hypothesis : $H_0 \rightarrow \quad \mu_{1..} = \mu_{2..} = \mu_{3..}$

Alternative Hypothesis: $H_1 \rightarrow$ Not all $\mu_{i..}$ are equal

Where 1, 2, 3 refers to Doctorate, Bachelors and HS-grad education level respectively.

For Occupation-

Null Hypothesis : $H_0 \rightarrow \quad \mu_{..1} = \mu_{..2} = \mu_{..3}$

Alternative Hypothesis: $H_1 \rightarrow$ Not all $\mu_{..j}$ are equal

Where 1,2,3,4 refers to Prof-specialty, Sales, Adm-clerical and Exec-managerial respectively.

For Interaction effect-

Null Hypothesis : $H_0 =$ Interaction effect does not exist

Alternative Hypothesis: $H_1 =$ An Interaction effect exists

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2	1.03E+11	5.13E+10	72.21196	5.47E-12
C(Occupation)	3	5.52E+09	1.84E+09	2.587626	7.21E-02
C(Education):C(Occupation)	6	3.63E+10	6.06E+09	8.519815	2.23E-05
Residual	29	2.06E+10	7.11E+08	NaN	NaN

Table 4: Two way ANOVA results considering the Interaction effect

When only Education is the predictor, (102695500000/165185556000=) 62.2% of total variability is explained by it. When only manufacturer is the predictor (11258780000/165185556000 =) 6.8% of total variability is explained by it. However, when both the factors are in the model (144564536000/165185556000=) 87.5% of total variability is explained by both main effects and their interaction effects. Hence, two way ANOVA was beneficial for our analysis.

Note that all three hypotheses are significant at 5% level. Therefore, our conclusion based on two-way ANOVA test, we reject the null hypothesis that all group means are equal for different Education levels; we reject the hypothesis that all group means are equal for different Occupations. Similarly, equality of means at each combination of Education level and Occupation is also rejected.

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

Sol: In this particular case study we found out that:

1. The Salary of individuals is very different for different Education levels which is understandable because the Education level impacts the Salary of individuals a lot.
2. By comparing combinations of means for different Education levels, we came to know that none of means are equal to other
3. For Occupations of individuals with respect to Salary we came to the conclusion that the variability in Salary of individuals is not significant for individuals with different Occupations.
4. After checking for the interaction effect, we found out that there is some sort of interaction present within the Education and Occupation levels.
5. Finally, after considering the interaction effects we were able to determine the cause of 87.5% variability within the response variable (which in our case study is ‘Salary’).

Problem 2

Education-Post Dataset Analysis (PCA)

Information about Dataset

The dataset Education - Post 12th Standard.csv contains information on various colleges. Since, we are expected to perform only Principal Component Analysis, no response variable is there in the dataset. The dataset contains 18 variables in total. The data dictionary is mentioned below:

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.alumni	Expend	Grad.Rate
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15

Figure 5: Sample of Education-Post 12th standard Dataset

Variable Information

```
RangeIndex: 777 entries, 0 to 776
Data columns (total 18 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Names        777 non-null    object
 1   Apps         777 non-null    int64
 2   Accept       777 non-null    int64
 3   Enroll       777 non-null    int64
 4   Top10perc    777 non-null    int64
 5   Top25perc    777 non-null    int64
 6   F.Undergrad  777 non-null    int64
 7   P.Undergrad  777 non-null    int64
 8   Outstate     777 non-null    int64
 9   Room.Board   777 non-null    int64
 10  Books        777 non-null    int64
 11  Personal     777 non-null    int64
 12  PhD          777 non-null    int64
 13  Terminal     777 non-null    int64
 14  S.F.Ratio    777 non-null    float64
 15  perc.alumni  777 non-null    int64
 16  Expend       777 non-null    int64
 17  Grad.Rate    777 non-null    int64
dtypes: float64(1), int64(16), object(1)
memory usage: 109.4+ KB
```

Figure 6: Variable information for Education-Post dataset

As we can clearly see there are 18 variables present and all of them are numeric type variables except the ‘Names’ variable which is object data type. There are total 777 data entries present for these variables and no null entries are there within the dataset.

Check for duplicate records

No duplicate records were found to be present in the data set.

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

Sol:

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.380952	165.105360	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.171171	5221.768440	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.463320	17.177710	10.0	53.0	65.0	78.0	118.0

Figure 7: Description of Education-Post Dataset

For univariate analysis, we can look at the descriptive statistics and draw following insights for this dataset:

1) Most of the features are having their means greater than median values, indicating that the distributions are right skewed except 'PhD', 'Terminal'.
2) Using the median values, we can say that acceptance rate in the universities is around 70%
3) Among these universities, 54% or more new students are within the top 25 and 23% or more new students are within the top 10 in Higher Secondary Class.
4) Student vs Faculty ratio ranges from as low as 2.5 to 39.8, which is quite high.
5) Graduation rates for 75% of the universities is up to 78.0 which is a good sign for the education level for these institutions.

By using boxplots, for these variable, we can conclude that nearly all of these variables contain outliers but for our case study we will not be treating these outliers. (Though it is important to treat outliers before going for PCA in real world situations)

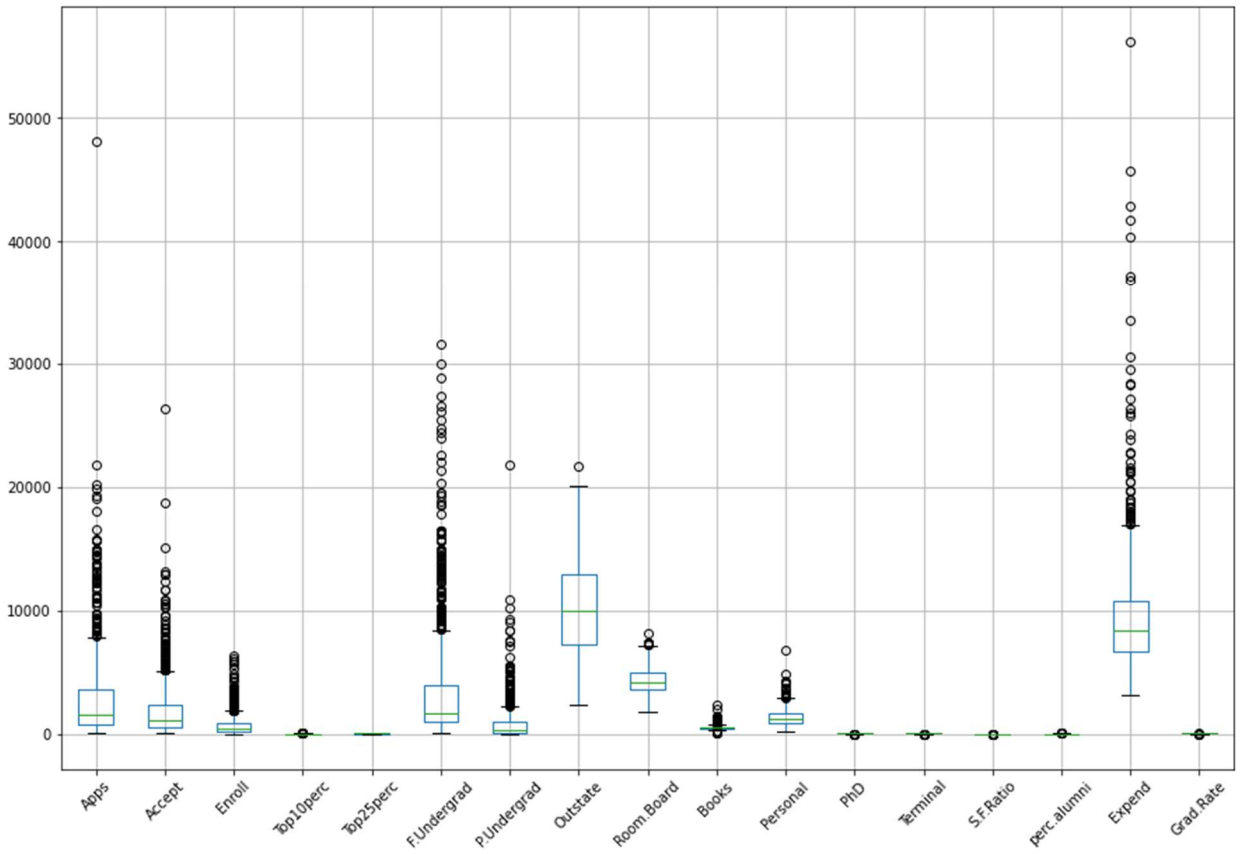

Figure 8: Boxplots for variables present in Education-Post dataset

Bivariate analysis-

Insights:

1. 'F.Undergrad' feature shows high positive correlation with the 'Apps', 'Accept' and 'Enroll' features which means that most of the students are applying for Undergrad programs.
2. Intrestingly, 'Outstate' feature shows moderately high positive correlation with 'Top10perc' and 'Top25perc' feature which could be interpreted as the students from "Out-states" are performing well in their universities.

3. Also, 'Outstate' feature depicts negative correlation with the 'S.F. Ratio' feature which shows that as higher as the number of students from "Out-states" lower is the student faculty ratio.
4. 'Expend' feature has also shown negative correlation with 'S.F. ratio' which could mean that the students in the universities with low Student vs Facult ratio are spending more than the rest

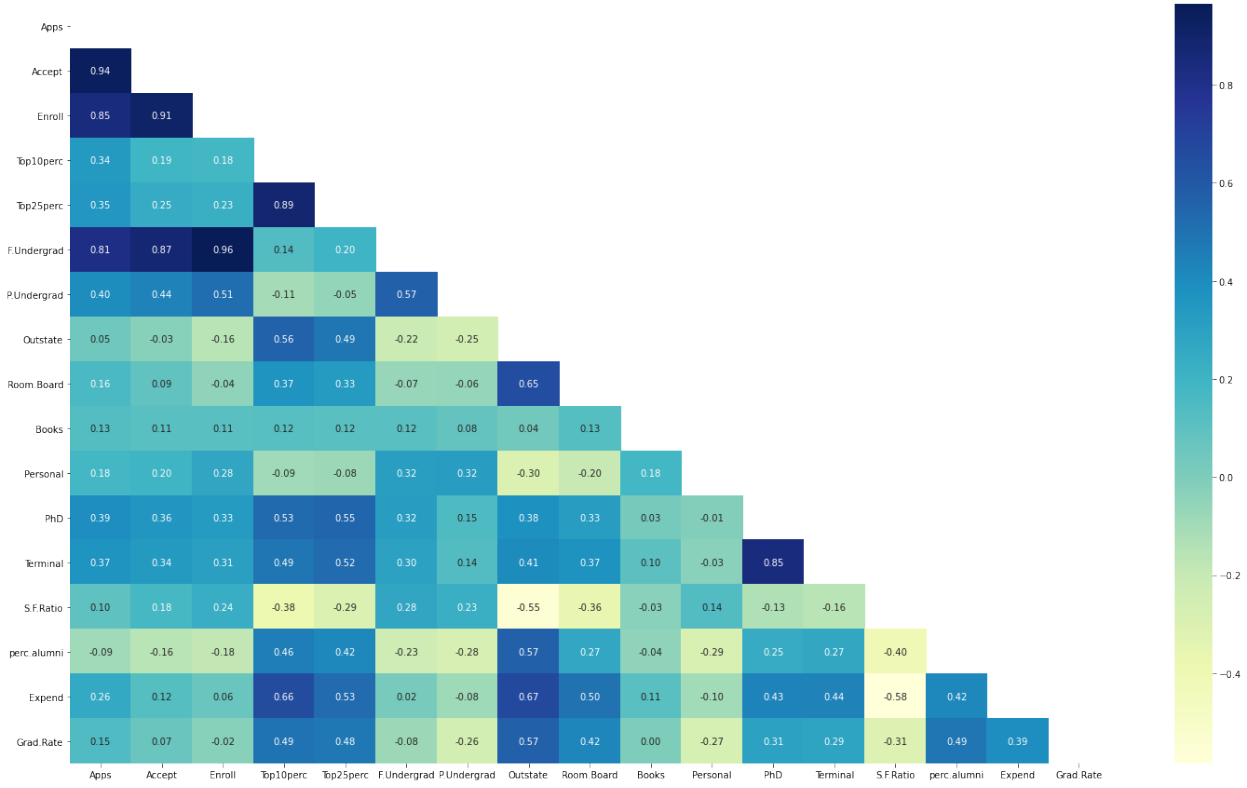

Figure 9: Correlation coefficient values for all the numeric variables

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

Sol:

In this case, if we look at the variance values for these variables ranges from 27266870 (for Expend variable) to as low as 15.6 (for S.F. ratio variable). PCA works on the total variance which is the sum of the variances in the data. If one (or more) variances is/are very high compared to the rest, it will dominate the construction of the PCs and all variables will not have proper representation.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	14978460.00	8949860.00	3045256.00	23132.77	26952.66	15289700.00	2346620.00	780970.40	700072.90	84703.75	468346.80	24689.43	21053.07	1465.06	-4327.12	5246171.00	9756.42
Accept	8949860.00	6007960.00	2076268.00	8321.12	12013.40	10393580.00	1646670.00	-253962.30	244347.10	45942.81	333556.60	14238.20	12182.09	1709.84	-4859.49	1596272.00	2834.16
Enroll	3045256.00	2076268.00	863368.40	2971.58	4172.59	4347530.00	725790.70	-581188.50	-40997.06	17291.20	176738.00	5028.96	4217.09	872.68	-2081.69	311345.40	-356.59
Top10perc	23132.77	8321.13	2971.58	311.18	311.63	12089.11	-2829.48	39907.18	7186.71	346.18	-1114.55	153.18	127.55	-26.87	99.57	60879.31	149.99
Top25perc	26952.66	12013.40	4172.59	311.63	392.23	19158.95	-1615.41	38992.43	7199.90	377.76	-1083.61	176.52	153.00	-23.10	102.55	54546.48	162.37
F.Undergrad	15289700.00	10393580.00	4347530.00	12089.11	19158.95	23526580.00	4212910.00	-4209843.00	-366458.20	92535.76	1041709.00	25211.78	21424.24	5370.21	-13791.93	472404.00	-6563.31
P.Undergrad	2346620.00	1646670.00	725790.70	-2829.47	-1615.41	4212910.00	2317799.00	-1552704.00	-102391.90	20410.45	329732.40	3706.76	3180.60	1401.30	-5297.34	-664351.20	-6721.06
Outstate	780970.40	-253962.30	-581188.50	39907.18	38992.43	-4209843.00	-1552704.00	16184660.00	2886597.00	25808.24	-814673.70	25157.52	24164.15	-8835.25	28229.55	14133240.00	39479.68
Room.Board	700072.90	244347.10	-40997.06	7186.71	7199.90	-366458.20	-102391.90	2886597.00	1202743.00	23170.31	-148083.80	5895.03	6047.30	-1574.21	3701.43	2873308.00	8005.36
Books	84703.75	45942.81	17291.20	346.18	377.76	92535.76	20410.45	25808.24	23170.31	27259.78	20043.03	72.53	242.96	-20.87	-82.26	96912.58	3.01
Personal	468346.80	333556.60	176738.00	-1114.55	-1083.61	1041709.00	329732.40	-814673.70	-148083.80	20043.03	458425.80	-120.90	-305.15	365.42	-2399.31	-346097.80	-3132.61
PhD	24689.43	14238.20	5028.96	153.18	176.52	25211.78	3706.76	25157.52	5895.04	72.53	-120.90	266.61	204.23	-8.44	50.38	36898.06	85.56
Terminal	21053.07	12182.09	4217.09	127.55	153.00	21424.24	3180.60	24164.15	6047.30	242.96	-305.15	204.23	216.75	-9.33	48.73	33733.46	73.22
S.F.Ratio	1465.06	1709.84	872.68	-26.87	-23.10	5370.21	1401.30	-8835.25	-1574.21	-20.87	365.42	-8.44	-9.33	15.67	-19.76	-12067.56	-20.85
perc.alumni	4327.12	-4859.49	-2081.69	99.57	102.55	-13791.93	-5297.34	28229.55	3701.43	-82.26	-2399.31	50.38	48.73	-19.76	153.56	27028.92	104.49
Expend	5246171.00	1596272.00	311345.40	60879.31	54546.48	472404.00	-664351.20	14133240.00	2873308.00	96912.58	-346097.80	36898.06	33733.46	-12067.56	27028.92	27266870.00	35012.97
Grad.Rate	9756.42	2834.16	-356.59	149.99	162.37	-6563.31	-6721.06	39479.68	8005.36	3.01	-3132.62	85.56	73.22	-20.85	104.49	35012.97	295.07

Figure 10: Covariance matrix for unscaled numeric variables in the data

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

Sol:

For scaled data, the values for covariance and correlation matrix were found to be same up to two digits after the decimal which shows that the standardization was effective. (Please refer to notebook for actual values).

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

Sol:

After checking the dataset for outliers before and after scaling we can observe that there is huge number of outliers present in the dataset in some variables such as ‘Apps’, ‘Books’, ‘Expend’ etc. The boxplot obtained for the same are following:

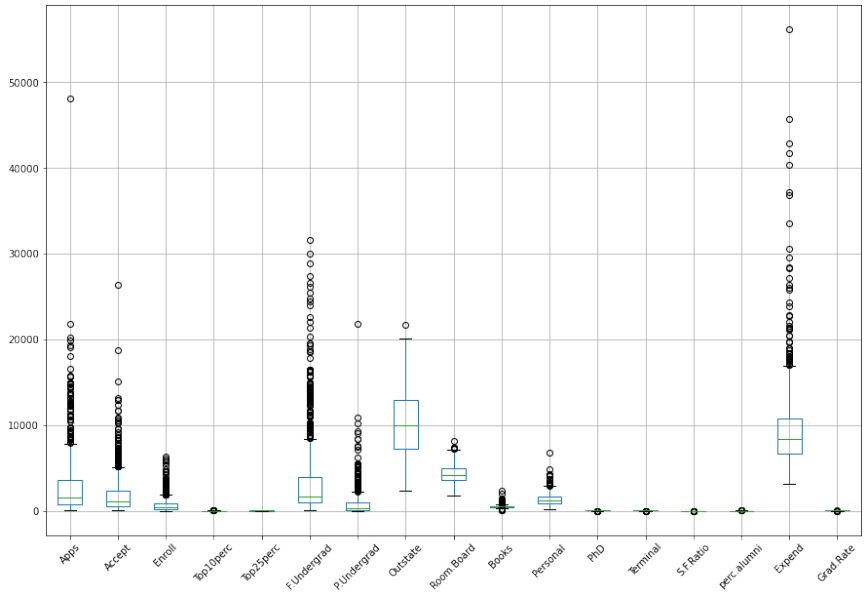

Figure 11: Boxplots for each variable in unscaled data

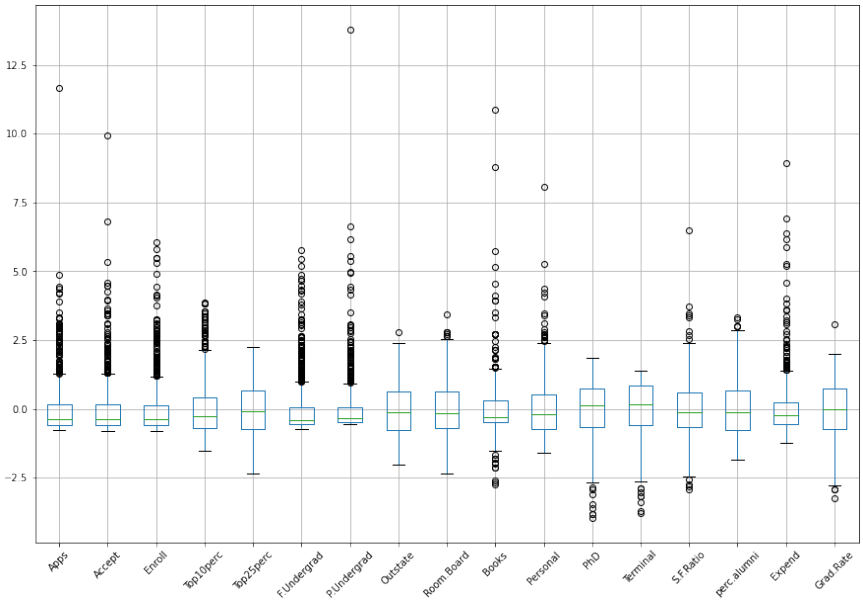

Figure 12: Boxplots for each variable in scaled data

2.5 Extract the eigenvalues and eigenvectors. [Print both]

Sol:

For the given dataset, the Eigen values and Eigen vectors of covariance matrix for scaled data can be given as:

Index	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate	eig_values
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	-0.32	-0.25	5.45
1	0.33	0.37	0.40	-0.08	-0.04	0.42	0.32	-0.25	-0.14	0.06	0.22	0.06	0.05	0.25	-0.25	-0.13	-0.17	4.48
2	-0.06	-0.10	-0.08	0.04	-0.02	-0.06	0.14	0.05	0.15	0.68	0.50	-0.13	-0.07	-0.29	-0.15	0.23	-0.21	1.17
3	0.28	0.27	0.16	-0.05	-0.11	0.10	-0.16	0.13	0.18	0.09	-0.23	-0.53	-0.52	-0.16	0.02	0.08	0.27	1.01
4	-0.01	-0.06	0.06	0.40	0.43	0.04	-0.30	-0.22	-0.56	0.13	0.22	-0.14	-0.20	0.08	0.22	-0.08	0.11	0.93
5	-0.02	0.01	-0.04	-0.05	0.03	-0.04	-0.19	-0.03	0.16	0.64	-0.33	0.09	0.15	0.49	-0.05	-0.30	0.22	0.85
6	-0.04	-0.01	-0.03	-0.16	-0.12	-0.03	0.06	0.11	0.21	-0.15	0.63	0.00	-0.03	0.22	0.24	-0.23	0.56	0.61
7	-0.10	-0.06	0.06	-0.12	-0.10	0.08	0.57	0.01	-0.22	0.21	-0.23	-0.08	-0.01	-0.08	0.68	-0.05	-0.01	0.59
8	-0.09	-0.18	-0.13	0.34	0.40	-0.06	0.56	0.00	0.28	-0.13	-0.09	-0.19	-0.25	0.27	-0.26	-0.05	0.04	0.53
9	0.05	0.04	0.03	0.06	0.01	0.02	-0.22	0.19	0.30	-0.08	0.14	-0.12	-0.09	0.47	0.42	0.13	-0.59	0.40
12	-0.04	0.06	0.07	0.01	0.27	0.08	-0.10	-0.14	0.36	-0.03	0.02	-0.04	0.06	-0.45	0.13	-0.69	-0.22	0.31
16	0.02	-0.15	0.01	0.04	-0.09	0.06	-0.06	-0.82	0.35	-0.03	-0.04	0.02	0.02	-0.01	0.18	0.33	0.12	0.22
15	-0.60	-0.29	0.44	0.00	-0.02	0.52	-0.13	0.14	0.07	-0.01	-0.04	-0.13	0.06	0.02	-0.10	0.09	0.07	0.17
14	-0.08	-0.03	0.09	0.11	-0.15	0.06	-0.02	0.03	0.06	0.07	-0.03	0.69	-0.67	-0.04	0.03	-0.07	-0.04	0.14
13	0.13	-0.15	0.03	0.70	-0.62	0.01	0.02	0.04	0.00	-0.01	0.00	-0.11	0.16	-0.02	-0.01	-0.23	0.00	0.09
11	0.46	-0.52	-0.40	-0.15	0.05	0.56	-0.05	0.10	-0.03	0.00	-0.01	0.03	-0.03	-0.02	0.00	-0.04	-0.01	0.04
10	-0.36	0.54	-0.61	0.14	-0.08	0.41	-0.01	-0.05	0.00	0.00	0.00	-0.01	-0.01	0.00	0.02	0.04	0.01	0.02

Figure 13: Eigen vector components for each dimensions (variable) and corresponding Eigen values (last column)

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

Sol:

By looking at Cumulative variance explained by each principal component we can decide how many principal components one should consider for given variance level. In our case, we assumed that 85% variance level should be considered. Hence corresponding to this variance explained level we found out, there are 7 principal component required for the same. They are following:

Index	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	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	0.32	0.25
1	0.33	0.37	0.40	-0.08	-0.04	0.42	0.32	-0.25	-0.14	0.06	0.22	0.06	0.05	0.25	-0.25	-0.13	-0.17
2	-0.06	-0.10	-0.08	0.04	-0.02	-0.06	0.14	0.05	0.15	0.68	0.50	-0.13	-0.07	-0.29	-0.15	0.23	-0.21
3	0.28	0.27	0.16	-0.05	-0.11	0.10	-0.16	0.13	0.18	0.09	-0.23	-0.53	-0.52	-0.16	0.02	0.08	0.27
4	0.01	0.06	-0.06	-0.40	-0.43	-0.04	0.30	0.22	0.56	-0.13	-0.22	0.14	0.20	-0.08	-0.22	0.08	-0.11
5	-0.02	0.01	-0.04	-0.05	0.03	-0.04	-0.19	-0.03	0.16	0.64	-0.33	0.09	0.15	0.49	-0.05	-0.30	0.22
6	-0.04	-0.01	-0.03	-0.16	-0.12	-0.03	0.06	0.11	0.21	-0.15	0.63	0.00	-0.03	0.22	0.24	-0.23	0.56

Figure 14: Obtained Principal Components

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

Sol:

Explicit form of First PC:

$$0.25 * Apps + 0.21 * Accept + 0.18 * Enroll + 0.35 * Top10perc + 0.34 * Top25perc + 0.15 * F.Undergrad + 0.03 * P.Undergrad + 0.29 * Outstate + 0.25 * Room.Board + 0.06 * Books + (-0.04) * Personal + 0.32 * PhD + 0.32 * Terminal + (-0.18) * S.F.Ratio + 0.21 * perc.alumni + 0.32 * Expend + 0.25 * Grad.Rate = New\ value\ for\ Transformed\ Data\ for\ 1st\ variable\ (i.e.\ 1st\ PC)$$

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?

Sol:

The Eigen values help us understand how much variance of Data is explained by the corresponding Eigen Vector. These Eigen vectors are nothing but the weights which we assign to our actual variables given in Dataset to transform data into new variable such that variance explained is maximum.

Now, by dividing these Eigen values with their summation we can obtain the fraction of explained variance. After taking the cumulative sum and plotting it, we can visualize actual values of the cumulative variance explained. This plot is also known as Scree plot. For our case the Scree plot obtained is following:

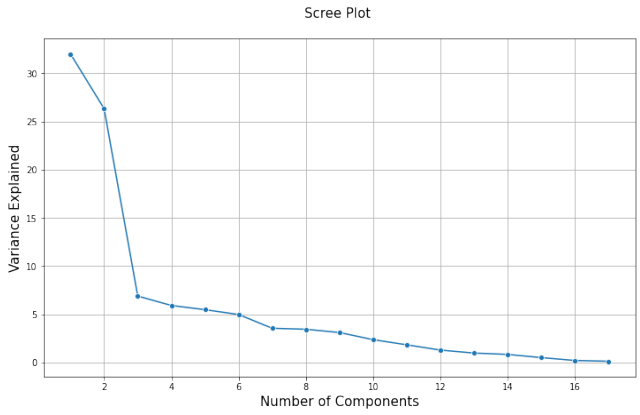

Figure 15: Scree Plot for obtained Eigen values

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

Sol:

The Principal Component Analysis is used for reducing the dimensionality and increasing interpretability at the same time reducing the information loss. In our case study, after performing the PCA we were able to reduce the number of numeric variables/features from 17 to 7 which explains around 85% of variance of data. The reduced number variable will be beneficial for creating models or other algorithms since one will have to deal with lesser number of features. It will also reduce the computational power required for further analysis or creating models.

The heatmap displays the correlation among the obtained variables after performing PCA on the dataset.

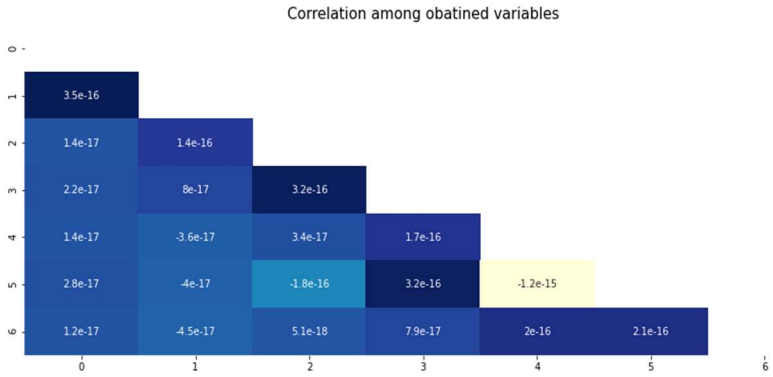

Figure 16: Correlation among new variables

We can clearly observe that the values are very close to 0 which can be interpreted as, the obtained variables are independent in nature which has to be true because by the definition principal components, they are always orthogonal to each other in multi-dimensional space.

Hence, with this we can say that the PCA performed on the dataset, worked pretty well.