TABLE OF CONTENTS

TABLES

FIGURES

Problem 1:

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.

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

Solution:

Null and Alternate Hypothesis for Education.

H_0 : The mean Salary is same across all Education level.

H_A : The mean Salary is different for at least one Education level.

Null and Alternate Hypothesis for Occupation.

H_0: The mean Salary is same across all Occupation.

H_A : The mean of Salary is different for at least one Occupation.

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

Solution: Anova Result for Education

	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. 1 - Anova Result for Education

Alpha = 0.05

P_value= 1.26E-08 < Alpha

We have enough evidence to reject the null hypothesis, The mean salary is different for at least one education level.

Q-1.3 : 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.

Solution:

Anova result for Occupation

	df	sum_sq	mean_sq	F	PR(>F)
(Occupation)	3	1.13E+10	3.75E+09	884144	0.458508
Residual	36	1.53E+11	4.24E+09	NaN	NaN

Table 1. 2 – Anova Result for Occupation.

Alpha = 0.05

P Value = 0.458508 > Alpha

We don't have enough evidence to reject the null hypothesis, The mean salary is same across all Occupation.

Q1.4: If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result.

Solution:

As Null hypothesis is been rejected for Education case not we test it for multiple comparison check mean differences.

Multiple Comparison for Education

Multiple Comparison of Means - Tukey HSD, FWER = 0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Bachelors	Doctorate	43274.07	.0146	7541.14	79006.99	TRUE
Bachelors	HS-grad	-90114.16	0.001	132035.20	48193.12	TRUE
Doctorate	HS-grad	-133388.22	0.001	174815.09	91961.36	TRUE

As we can interpret from the above result

1. Doctorate mean salary is higher than Bachelors by 43274.0667
2. Bachelors mean salary is higher than High school graduate by 90114.1556

We can conclude that Education level of Doctorate mean salary is higher as compared with Bachelors and High School Graduate.

Q-1.5: 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 'point plot' function from the 'seaborn' function]

Solution:

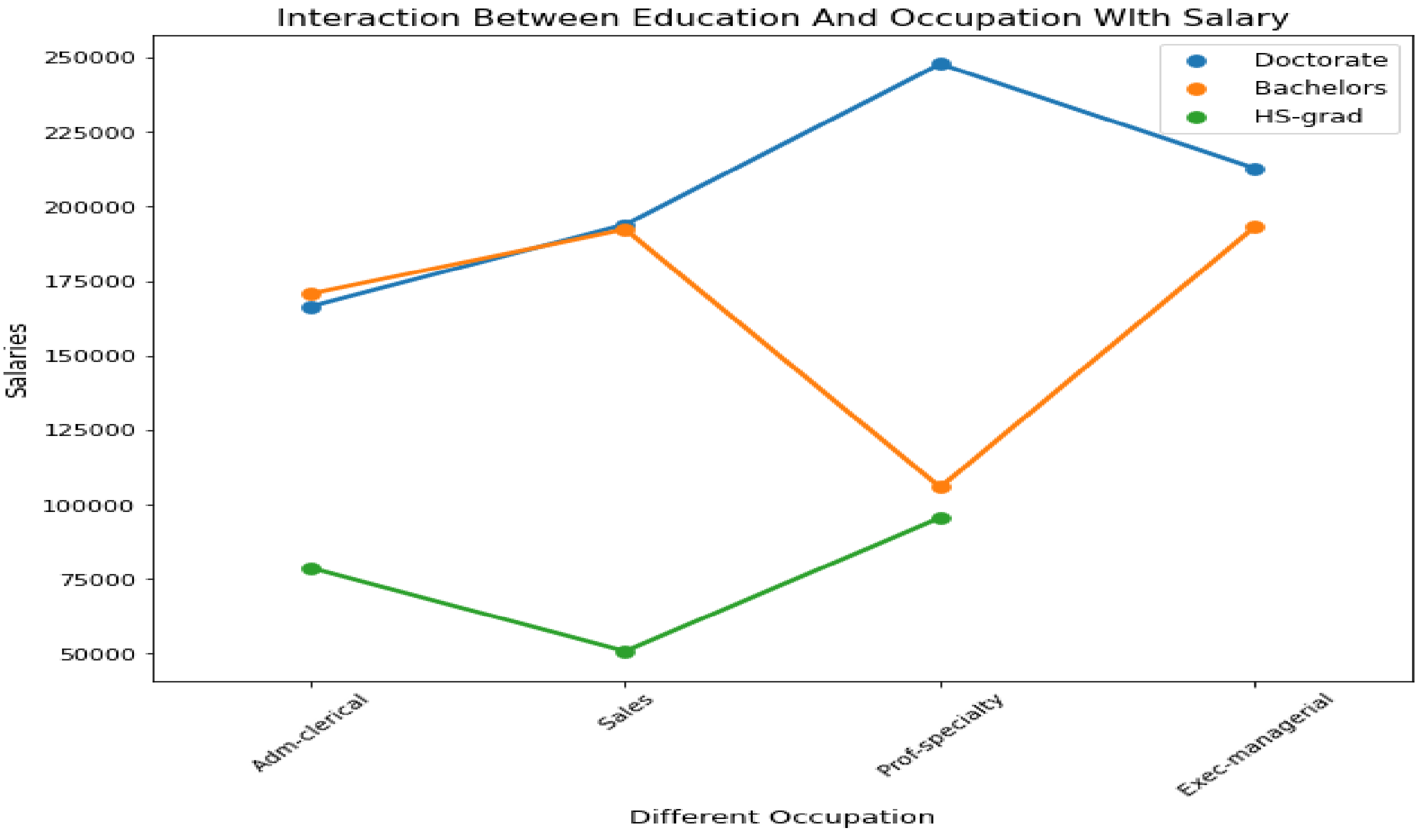


Figure 1. 1 – POINT PLOT

As from the above plot we can say there is an interaction of different education level with their occupations. The information that we infer are as follows,

1. In administrative and clerical occupation the people with education level of Doctorate and Bachelors salaries are almost similar, Even though Bachelors have higher salary than Doctorate but there is very slight difference. Whereas Salaries of High School Graduate is significantly lower than Doctorate and Bachelors.
2. For People working in Sales tends to get paid higher salaries than admin-clerical for Doctorate and bachelors Education level, whereas High School Graduate gets paid even lesser than the adm-clerical Occupation.
3. There is a higher demand of Doctorate Education in Prof-specialty Occupation as compare to Bachelors and High school graduate as Doctorate are getting remarkably higher salaries than Bachelors and High school Graduate.
4. In Prof-specialty People with bachelor's education have lowest salaries as compare to other occupation whereas high school graduates have highest salaries in this occupation amongst all other.

For Exec-Managerial Occupation roles Doctorate and Bachelors are preferred. But Salaries of people having education of Doctorate are higher than Bachelors.

Q-1.6: 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?

Solution:

Null and Alternate Hypothesis for Education.

H_0 : The mean Salary is same across all Education level.
H_A : The mean Salary is different for at least one Education level.

Null and Alternate Hypothesis for Occupation.

H_0 : The mean Salary is same across all Occupation.
H_A : The mean of Salary is different for at least one Occupation

Null and Alternate Hypothesis for Interaction between Education and Occupation.

H_0 : There is no Interaction between Education and Occupation.
H_A : There is an Interaction between Education and Occupation.

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2	1.03E+11	5.13E+10	72.211958	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 1. 3 – TWO WAY ANOVA TEST RESULT TABLE

Results for Education

Alpha = 0.05

P value for Education = 5.47E-12 < Alpha

We have enough evidence to reject the null hypothesis, The mean salary is different for at least one Education Level.

Result for Occupation:
Alpha = 0.05
P value = 7.21E-02 > Alpha

We don't have enough evidence to reject the null hypothesis, The mean salary is same across all Occupation.

Result for Interaction between Education and Occupation:
Alpha = 0.05
P value = 2.23E-05 < Alpha

We have enough evidence to reject the null hypothesis, There is an Interaction Between Education and Occupation.

CONCLUSION

As we can see from the result of Two-way Anova test that Education alone is an factor that's impacting the salary, but Occupation alone is not an impacting factor on salary whereas the results shows that Education and Occupation have some interaction which is impacting the salary.

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

Solution:

There are many business implications of performing Anova for this particular study like

1. Education is one the most deriving factor for salary.
2. Highly Educated people can get higher salaries.
3. High School Graduate have certain limit for getting salaries, they get paid lesser salaries as compare other higher education levels. But there is an occupation Prof-specialty which can pay them higher salary than the other two.
4. Certain Occupation like Adm-Clerical and Sales does not pay higher salaries than bachelors for higher education like Doctorate.
5. Doctorate is more preferred education level in Prof-specialty over Bachelors and High school graduate and can get higher salaries than other occupation. So people trying to excel in the same occupation should go for Doctorate.
6. Exec-managerial Jobs prefer both Education level bachelors and Doctorate but doctorate can get paid higher salaries than bachelors. High school graduates are not considered for Exec-managerial jobs.

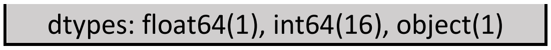
PROBLEM 2

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.

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

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

	count	mean	std	min	25%	50%	75%	max
Apps	777	3001.6384	3870.2015	81	776	1558	3624	48094
Accept	777	2018.8044	2451.114	72	604	1110	2424	26330
Enroll	777	779.97297	929.17619	35	242	434	902	6392
Top10perc	777	27.558559	17.640364	1	15	23	35	96
Top25perc	777	55.796654	19.804778	9	41	54	69	100
F.Undergrad	777	3699.9073	4850.4205	139	992	1707	4005	31643
P.Undergrad	777	855.29858	1522.4319	1	95	353	967	21836
Outstate	777	10440.669	4023.0165	2340	7320	9990	12925	21700
Room.Board	777	4357.5264	1096.6964	1780	3597	4200	5050	8124
Books	777	549.38095	165.10536	96	470	500	600	2340
Personal	777	1340.6422	677.07145	250	850	1200	1700	6800
PhD	777	72.660232	16.328155	8	62	75	85	103
Terminal	777	79.702703	14.722359	24	71	82	92	100
S.F.Ratio	777	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777	22.743887	12.391801	0	13	21	31	64
Expend	777	9660.1712	5221.7684	3186	6751	8377	10830	56233
Grad.Rate	777	65.46332	17.17771	10	53	65	78	118



```
dtypes: float64(1), int64(16), object(1)
```

Table 2. 1 – DATA INFORMATION.

Table 2. 2 – DATA DESCRIPTION

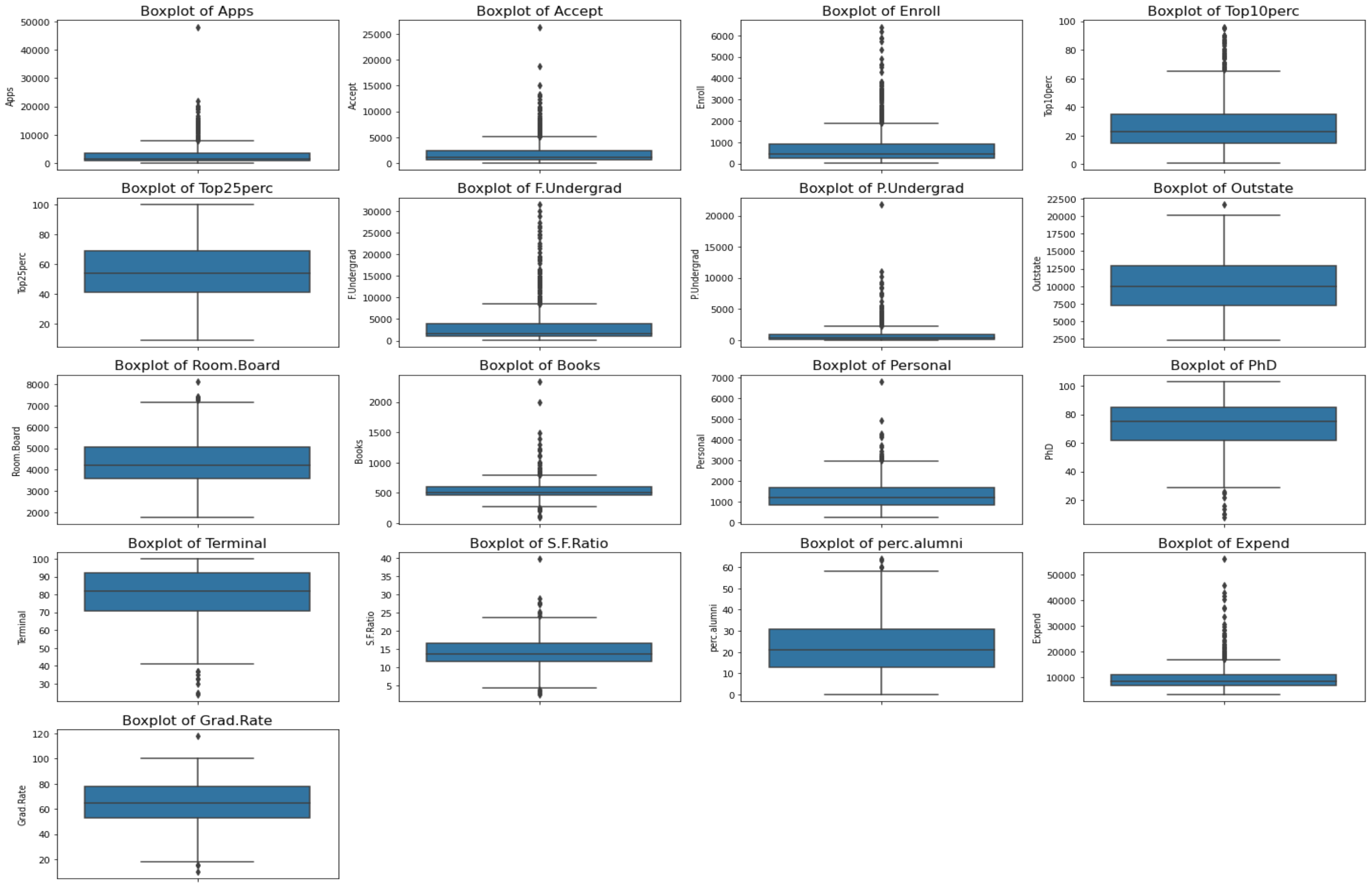


Figure 2. 1 – BOX PLOT OF DATA FEATURES

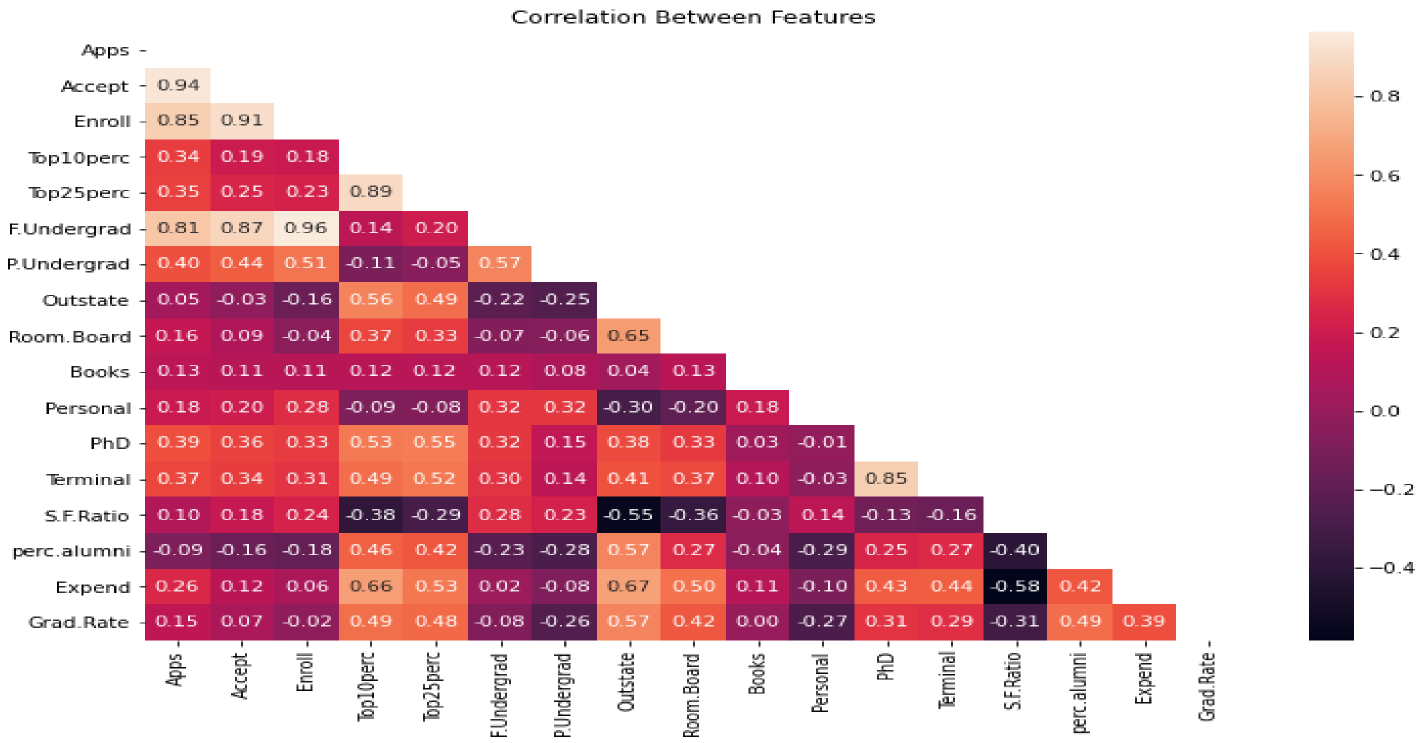


Figure 2. 2 – CORRELATION BETWEEN FEATURES.

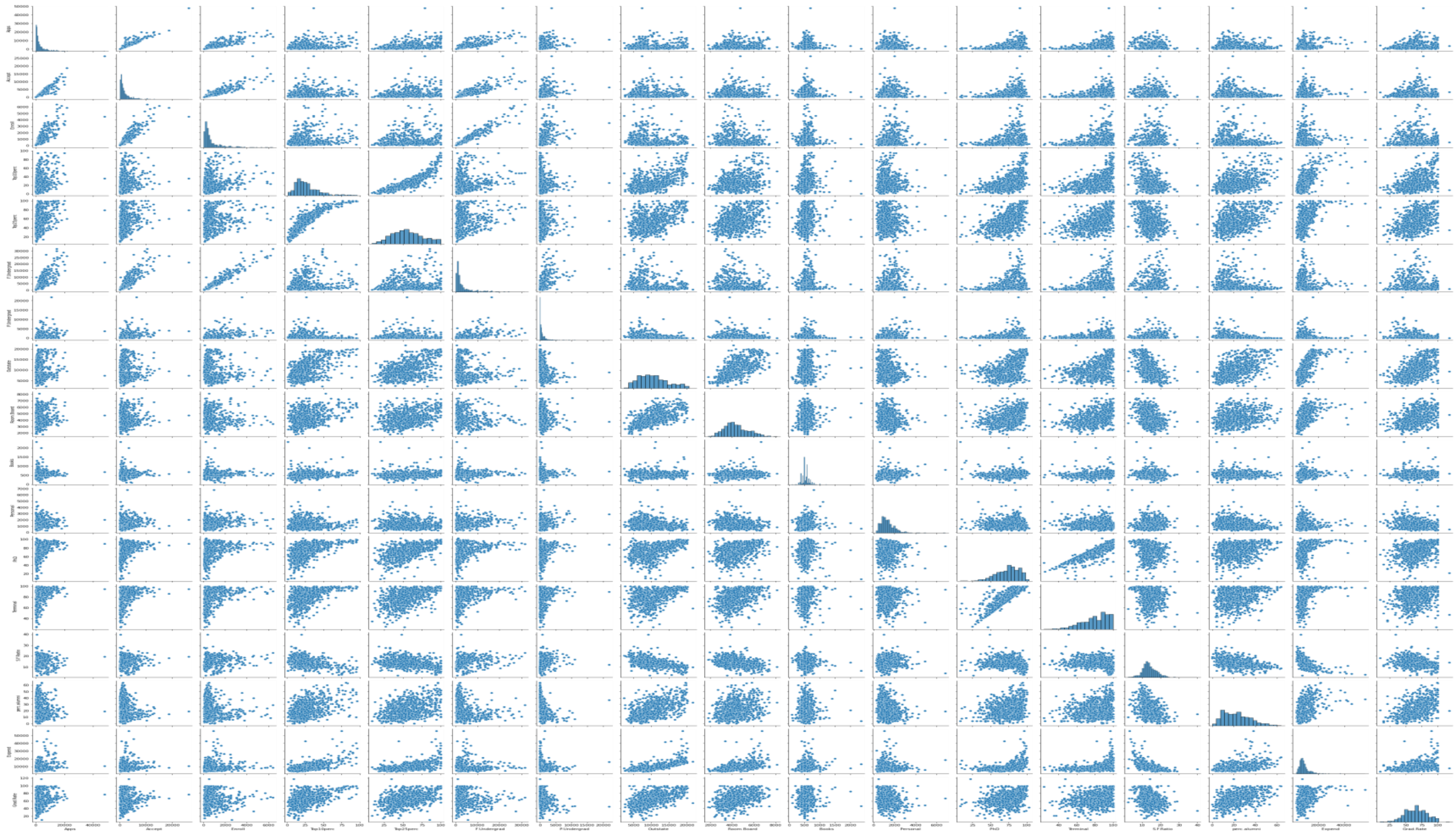


Figure 2. 3 – PAIR PLOT OF FEATURES

By performing EDA on the Data we can infer following things

1. There are 777 rows and 18 Features in the Data which consist of 17 numeric features.

2. There are no Null values and bad data present in the data frame that needs treatment although there are outliers present in almost every features that needs to be treated.
3. Application received by each college and Universities ranges from 81 to 48094.
4. The Estimated personal expenditure for a student can start from 250 and can goes up to 6800.
5. The Student faculty ratios are well maintained by some colleges and universities which can range from 2.5 to 39.8.
6. In colleges and universities faculties with degree can range from 8 to 103.
7. There are some features like Apps, Accept, Enroll, Personal are right skewed whereas like Terminal, PhD are left skewed.
8. There are some features like Top25perc, Outstate and Room board that are well distributed (almost normally distributed).
9. There are Few Features like - Apps-Accept, Apps-Enroll, Accept-Enroll, F.undergraduate - Apps, Accept, Enroll, Top10perc-Top25perc and Terminal-PhD that are very highly correlated.
10. This information we can also see from pair plot that as number of application increases acceptance rate and enrollment rate also increases.
11. We can also infer from the pair plot that universities and colleges receive more applications from full time graduates and the get accepted and enrolled by the universities and colleges.

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

Solution:

Yes it is necessary to scale the data as all the features are very vast and different from each other, like App is application received that is in 1000's , where as PhD teachers is in 100's similarly like there are multiple features like personal is expenditure in currency and enroll is in numbers. So it is necessary to bring all this data to one scale to perform further analysis.

	0	**1**	**2**	**3**	**4**	**...**	**772**	**773**	**774**	**775**	**776**
Apps	-0.346882	-0.210884	-0.406866	-0.668261	-0.726176	...	-0.20804	-0.269575	-0.233895	1.991711	-0.003268
Accept	-0.321205	-0.038703	-0.376318	-0.681682	-0.764555	...	-0.205673	-0.087284	-0.042377	0.177256	-0.066872
Enroll	-0.063509	-0.288584	-0.478121	-0.692427	-0.780735	...	-0.2552	-0.091509	-0.091509	0.578333	-0.095816
Top10perc	-0.258583	-0.655656	-0.315307	1.840231	-0.655656	...	-1.336352	-0.201858	0.365389	3.825595	0.025041
Top25perc	-0.191827	-1.353911	-0.292878	1.677612	-0.596031	...	-1.505488	-0.444454	0.262901	2.182866	0.363952
F.Undergrad	-0.168116	-0.209788	-0.549565	-0.658079	-0.711924	...	-0.12603	-0.175543	-0.187095	0.312977	-0.146867
P.Undergrad	-0.209207	0.244307	-0.49709	-0.520752	0.009005	...	0.771435	0.165435	-0.453053	-0.507606	0.572283
Outstate	-0.746356	0.457496	0.201305	0.626633	-0.716508	...	-0.906289	0.268462	-0.88067	2.337894	-1.355744
Room.Board	-0.964905	1.909208	-0.554317	0.996791	-0.216723	...	-0.417455	0.549707	-0.14373	1.963953	-0.727676
Books	-0.602312	1.21588	-0.905344	-0.602312	1.518912	...	-0.29928	0.306784	0.409815	0.488603	-0.29928
Personal	1.270045	0.235515	-0.259582	-0.688173	0.235515	...	-0.207855	-0.13396	-0.827095	1.144424	-0.13396
PhD	-0.163028	-2.675646	-1.204845	1.185206	0.204672	...	-0.775861	0.020822	-0.346878	1.430339	0.143389
Terminal	-0.115729	-3.378176	-0.931341	1.175657	-0.523535	...	-1.339146	-0.319632	-0.319632	1.107689	-0.319632
S.F.Ratio	1.013776	-0.477704	-0.300749	-1.615274	-0.553542	...	1.746877	-0.199632	0.078441	-2.095582	1.013776
perc.alumni	-0.867574	-0.544572	0.585935	1.151188	-1.675079	...	-0.706073	0.666685	-0.22157	2.120194	0.424434
Expend	-0.50191	0.16611	-0.17729	1.792851	0.241803	...	-0.994781	-0.09029	-0.256241	5.887971	-0.987116
Grad.Rate	-0.318252	-0.551262	-0.667767	-0.376504	-2.939613	...	-1.483301	1.021555	-0.959029	1.953595	1.953595

Table 2. 3 – SCALED DATA

Values showing data frame are after standardization (Z Score) with 777 Rows and 17 Columns.

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

Q-2.4: Check the dataset for outliers before and after scaling. What insight do you derive here? [Please do not treat Outliers unless specifically asked to do so].

Solution:

BOX PLOT OF DATA BEFORE SCALING

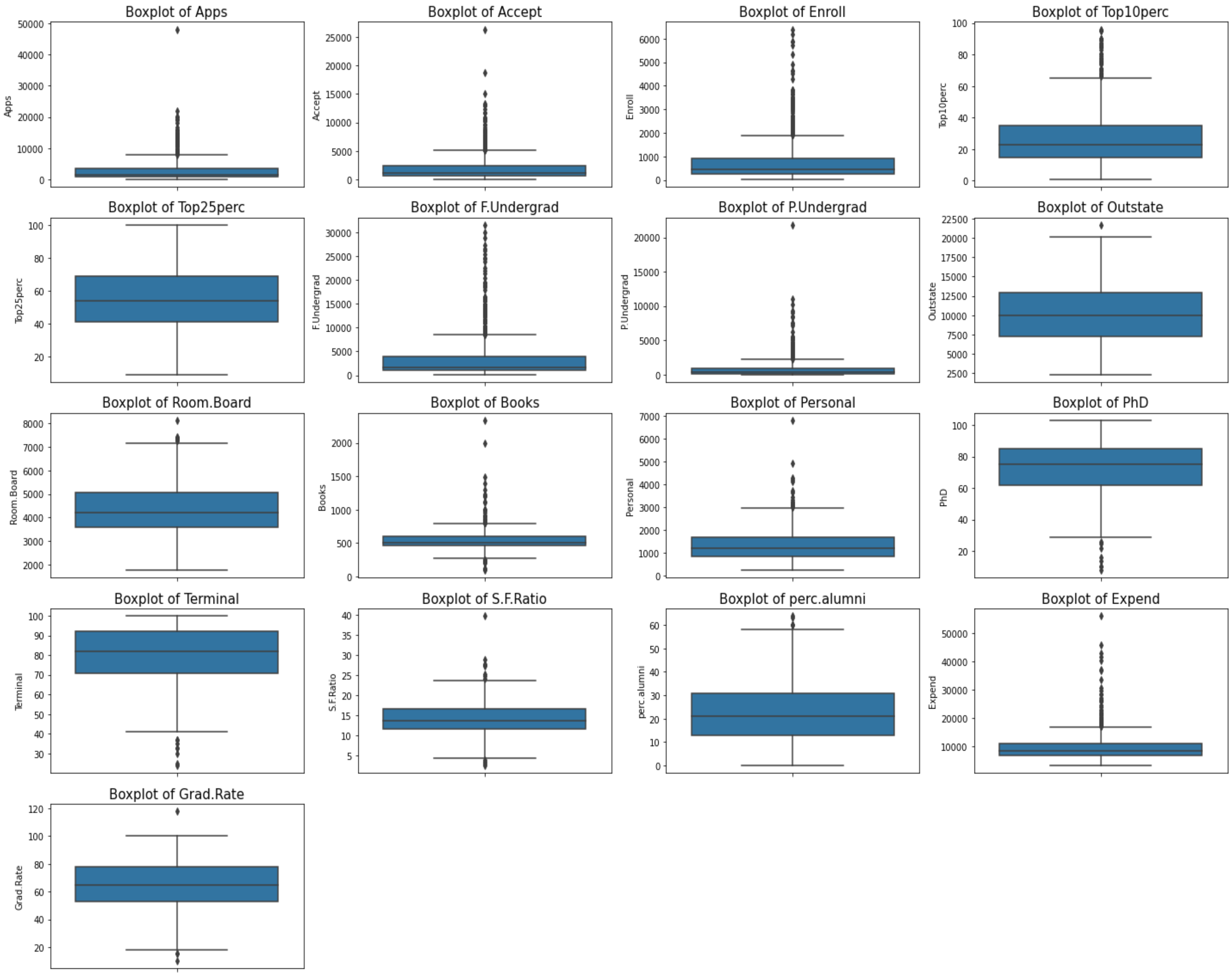


Figure 2. 4 – BOX PLOT BEFORE SCALING

BOX PLOT OF DATA AFTER SCALING

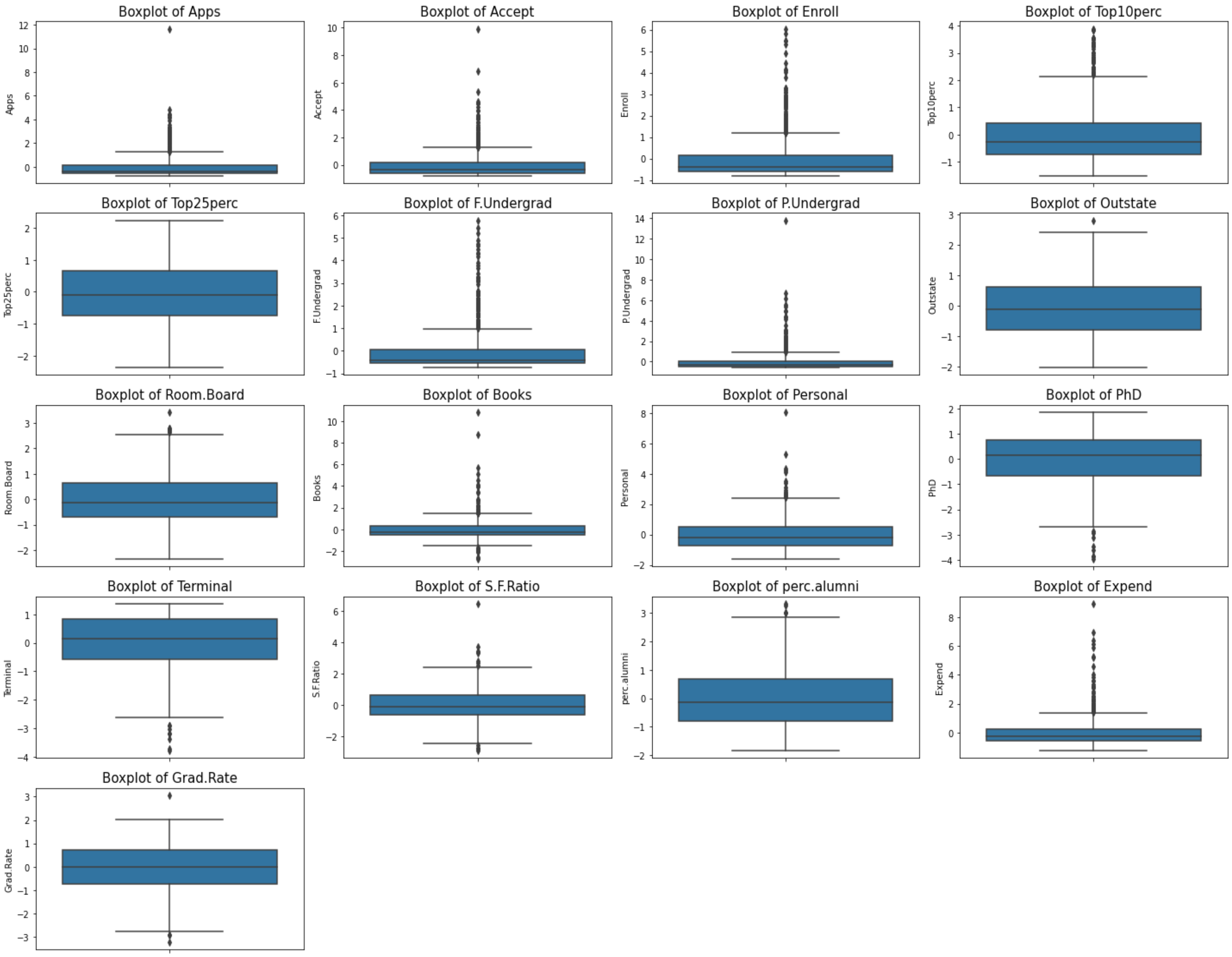


Figure 2. 5- BOX PLOT AFTER SCALING

Box plot after scaling the data.

Insights before and after scaling of data are as following.

1. Before scaling and after scaling the data, the range of displaying has changed.
2. After scaling the features has reduced to mean -0 and standard deviation -1.
3. Before scaling the range for Apps was 0 to 50000 but after scaling it has changed to 0 to 12.
4. As range has changed after scaling, we can easily interpret the distance of outliers from the upper and lower limit.
5. We can interpret that outliers are how many standard deviation away from the mean.
6. This can help us to understand how to treat the outliers in better way.

Scaling will not affect the outliers but it will help us to locate and position the outliers which will help us to understand better and more efficient ways to treat the outliers without affecting the data.

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

Solution:

Extracted Eigen values and Eigen Vectors using Python.

Eigen Vectors for PCA 1

```
[ 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]
```

Eigen Vectors for PCA 2

```
[ 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]
```

Eigen Vectors for PCA 3

```
[-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]
```

Eigen Vectors for PCA 4

```
[ 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]
```

Eigen Vectors for PCA 5

```
[ 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]
```

Eigen Vectors for PCA 6

```
[-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]
```

Eigen Vectors for PCA 7

```
[-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],
```

Eigen Vectors for PCA 8

```
[-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],
```

Eigen Vectors for PCA 9

```
[-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]
```

Eigen Vectors for PCA 10

```
[ 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]
```

Eigen Vectors for PCA 11

```
[ 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],
```

Eigen Vectors for PCA 12

```
[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],
```

Eigen Vectors for PCA 13

```
[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],
```

Eigen Vectors for PCA 14

```
[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],
```

Eigen Vectors for PCA 15

```
[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],
```

Eigen Vectors for PCA 17

```
[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],
```

Eigen Vectors for PCA 18

```
[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]
```

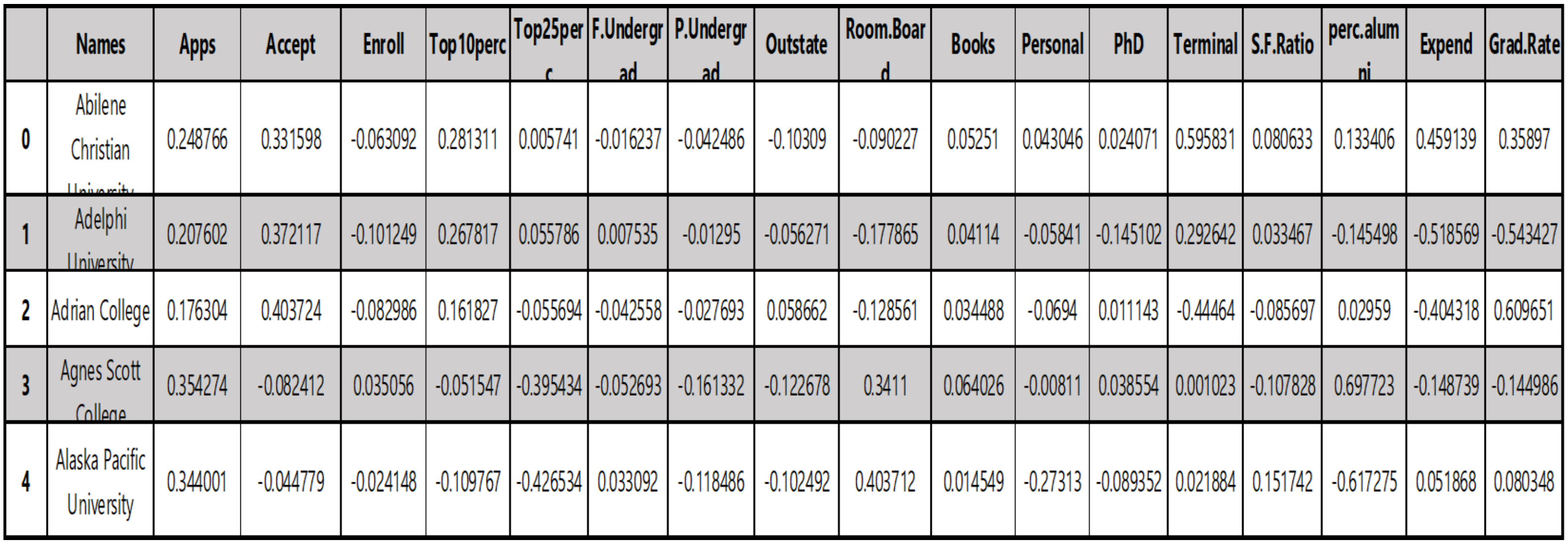
Eigen Values:

```
[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]
```

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

Solution:

Displaying top 5 rows after performing PCA and exporting the Eigen Vectors(Principal component) with original features of data frame



	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	0.248766	0.331598	-0.063092	0.281311	0.005741	-0.016237	-0.042486	-0.10309	-0.090227	0.05251	0.043046	0.024071	0.595831	0.080633	0.133406	0.459139	0.35897
1	Adelphi University	0.207602	0.372117	-0.101249	0.267817	0.055786	0.007535	-0.01295	-0.056271	-0.177865	0.04114	-0.05841	-0.145102	0.292642	0.033467	-0.145498	-0.518569	-0.543427
2	Adrian College	0.176304	0.403724	-0.082986	0.161827	-0.055694	-0.042558	-0.027693	0.058662	-0.128561	0.034488	-0.0694	0.011143	-0.44464	-0.085697	0.02959	-0.404318	0.609651
3	Agnes Scott College	0.354274	-0.082412	0.035056	-0.051547	-0.395434	-0.052693	-0.161332	-0.122678	0.3411	0.064026	-0.00811	0.038554	0.001023	-0.107828	0.697723	-0.148739	-0.144986
4	Alaska Pacific University	0.344001	-0.044779	-0.024148	-0.109767	-0.426534	0.033092	-0.118486	-0.102492	0.403712	0.014549	-0.27313	-0.089352	0.021884	0.151742	-0.617275	0.051868	0.080348

TABLE 2. 4 – PCA WITH DATAFRAME ORIGNAL 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). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

Solution:

After Performing PCA, Displaying First Row

-1.59, 0.77, -0.1, -0.92, -0.74, -0.3, 0.64, -0.88, 0.09, 0.05, 0.4, -0.09, -0.05, 0.18, 0.0, -0.09, 0.09

Now performing explicit form of first PC

First PC = $a_1 * x_1 + a_2 * x_2 + a_3 * x_3 + a_4 * x_4 +$……………………..+ $a_x * x_n$

First PC =(-0.35) * (0.25) + (-0.32) * (0.21) + (-0.06) * (0.18) + (-0.25) * (0.35)+ (-0.19) * (0.34)+(-0.16) * (0.15) + (-0.21) * (0.03)+(-0.75) * (0.29) + (-0.96) * (0.24)+(-0.60) * (0.06) + (1.27) * (-0.04) + (-0.16) * (0.31) + (-0.11) * (0.31)+ (1.01) * (-0.18) + (-0.86) * (0.21) + (-0.50) * (0.32) + (-0.32) * (0.25)

This value gives the first value of PC which is also same as above results;

First value of PC= -1.59.

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?

Solution:

Below Plot and Table represents the cumulative value contains by each PC,

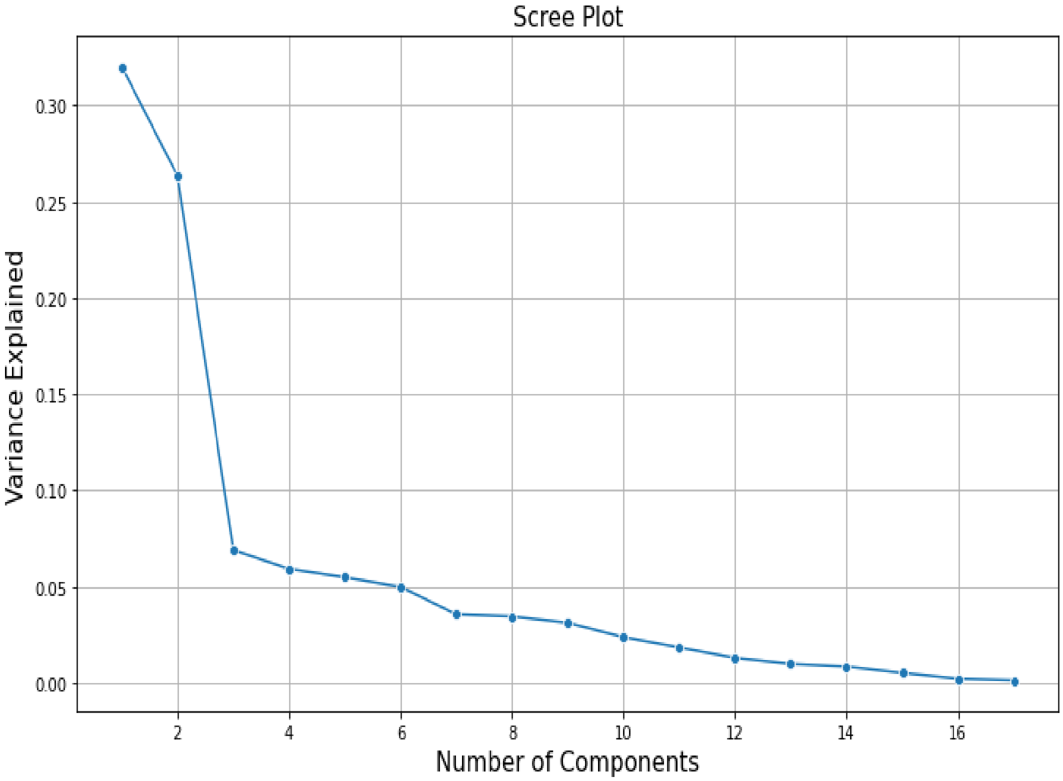


FIGURE 2. 6 – SCREE PLOT

PC	Cumm Value
PC 1	0.32020628
PC 2	0.58360843
PC 3	0.65261759
PC 4	0.71184748
PC 5	0.76673154
PC 6	0.81657854
PC 7	0.85216726
PC 8	0.88670347
PC 9	0.91787581
PC 10	0.94162773
PC 11	0.96004199
PC 12	0.9730024
PC 13	0.98285994
PC 14	0.99131837
PC 15	0.99648962
PC 16	0.99864716
PC 17	1

TABLE 2. 5 – CUMULATIVE EIGEN VALUES

After Checking the cumulative value of Eigen values, mathematically and Visually we can easily come to a conclusion to choose 8 Principal components to move further for our analysis on the data because 8 Pc's itself explain the 88.5% data, rest just to get more 10 % data we will have to add 4 more PC's which is not feasible for dimension reduction of the data.

Eigen Vectors Geometrically, an eigenvector, corresponding to a real nonzero value, points in a direction in which it is stretched by the transformation and the eigenvalue is the factor by which it is stretched. The eigenvectors (principal components) determine the directions of the new feature space, and the eigenvalues determine their magnitude.

Q2.9: 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].

Solutions:

Principal component analysis can serve much business implication for this case study such as:

1. PCA helped to reduce the dimensions for data in for this case study.
2. The entire PC obtained is independent of each other. There is no correlation present between the data.
3. Most of the variances was present in the original data has been captured in all PCs.
4. Reduction of the data will help in further and better analysis of the data.
5. PCA will also help in building better business models and improve the performance of algorithms.
6. It will also improve the accuracy of the results obtained after further analysis.

After Checking the cumulative value of Eigen values, mathematically and Visually we can easily come to a conclusion to choose 8 Principal components to move further for our analysis on the data because 8 Pc's itself explain the 88.5% data.