8/15/2021

Advanced Statistics Project Report PGP -DSBA

Saloni Juwatkar
PGP – DATA SCIENCE AND BUSINESS ANALYTICS

Table of Contents

List of Figures

1 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: has three levels: Doctorate, Bachelors and high school graduate.
- Occupation: has 4 levels: Administrative and clerical, Sales, Professional or specialty, and Executive or managerial.
- Salary: Integer data type showing salary of 40 individuals.

Sample Dataset

	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769

Figure 1: Problem Statement 1: Sample Dataset.

Dataset has salary of 40 individuals where salary is dependent on Education qualification and occupation. The dependency of the two components on the salary of the individual is to be determined by doing hypothesis testing and the conclusion has to be derived.

```
Education     0
Occupation    0
Salary        0
dtype: int64
```

Figure 2: Sample containing null data

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

The hypothesis for conducting one way ANOVA for Education:

H0: 3 Education levels have the same mean salary
H1: At least 1 education level has different mean salary.
Confidence Level: 0.05

The hypothesis for conducting one way ANOVA for occupation:

H0: All 4 occupation levels have the same mean salary.
H1: At least 1 occupation has different mean salary.
Confidence Level: 0.05

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.

The hypothesis for conducting one way ANOVA for Education:

H0: 3 Education levels have the same mean salary
H1: At least 1 education level has different mean salary.
Confidence Level: 0.05

To find whether all three education qualifications have same mean salary, we perform one way ANOVA on salary with respect to Education. Below are the results for one way ANOVA.

	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 3: One Way Anova: Salary and Education.

Additional test of Tukeyhsd test is performed to confirm the Anova results.

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 4: Tukeyhsd

From the above analysis it can be inferred that since the p-value (1.257709e-08) is less than the significance level (0.05), we reject the null hypothesis and we say that the different education qualifications affect the salary mean and hence at least 1 education gives different salary.

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.

The hypothesis for conducting one way ANOVA for occupation:

H0: All 4 occupation levels have the same mean salary.
H1: At least 1 occupation has different mean salary.
Confidence Level: 0.05

To find whether all 4 occupations have same mean salary, we perform one way ANOVA on salary with respect to occupation. Below are the results for one way ANOVA.

	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- Salary and Occupation

Additional test of Tukeyhsd test is performed to confirm the Anova results.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Adm-clerical	Exec-managerial	55693.3	0.4146	-40415.1459	151801.7459	False
Adm-clerical	Prof-specialty	27528.8538	0.7252	-46277.4011	101335.1088	False
Adm-clerical	Sales	16180.1167	0.9	-58951.3115	91311.5449	False
Exec-managerial	Prof-specialty	-28164.4462	0.8263	-120502.4542	64173.5618	False
Exec-managerial	Sales	-39513.1833	0.6507	-132913.8041	53887.4374	False
Prof-specialty	Sales	-11348.7372	0.9	-81592.6398	58895.1655	False

Figure 6: Tukeyhsd

From the above analysis it can be inferred that since the p-value (0.458508) is greater than the significance level (0.05), we fail to reject the null hypothesis and we say that we have no dependency of occupation on the salary of the individual and all the 4 occupations have the same mean salary.

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

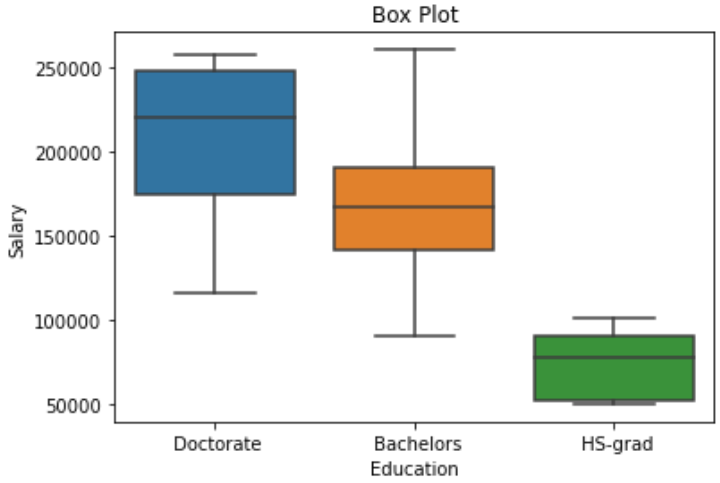


Figure 7: Boxplot salary vs Education.

The null hypothesis is rejected in (2) where from the above boxplot we can see that the HS-grad mean is significantly low that that of other qualifications.

1.5 What is the interaction between two treatments? Analyse 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]

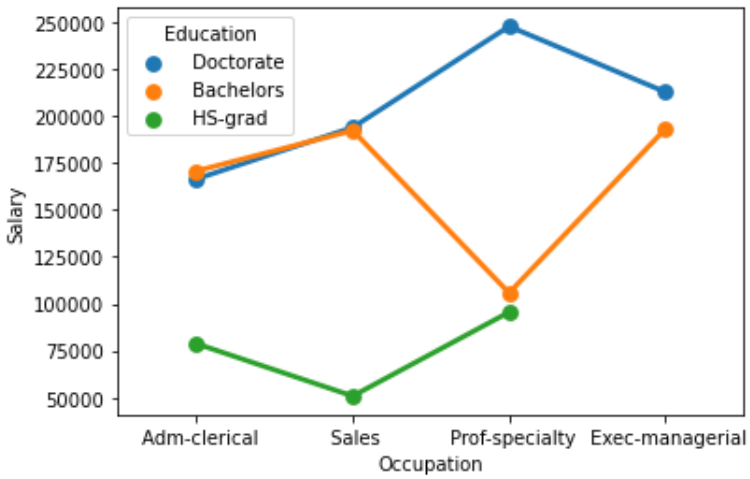


Figure 8: Point plot interaction between education and occupation.

From the plot we can infer that the education and occupation both affect the salary of an individual.

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?

The hypothesis for **Two Way ANOVA** is:

H0: There is no relationship between Salary, Education and Occupation

H1: There is relationship between Salary, Education and Occupation.

Confidence Level: 0.05.

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	5.277862	4.993238e-03
C(Education)	2.0	9.695663e+10	4.847831e+10	68.176603	1.090908e-11
C(Occupation):C(Education)	6.0	3.523330e+10	5.872217e+09	8.258287	2.913740e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Figure 9: Two Way ANOVA

From the above analysis, it can be seen that the p-values are less than the confidence level (0.05) and hence the null hypothesis is rejected in this case.

Hence, we can say that there is dependency between Education and occupation which tells us that the salary is dependent upon education and occupation.

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

- Performing one way ANOVA on Salary and Education as well as Salary and Occupation help us check the relationship between education and Occupation.
- We are doing ANOVA to find relationship between Salary and other two categories.
- Salary being a dependent variable we check whether the salary is dependent upon Education and Occupation.
- From the question 1.2 we can see that the null hypothesis is rejected stating that the Mean salary is different due to different educations.
- From the question 1.3 we can see that the null hypothesis is not rejected stating that the mean salary is not different due to different occupations.
- From the question 1.6, we can see that the null hypothesis is rejected stating that there is interaction between education and occupation which in turn affects the salary.

2 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.

Data Description

```
<class 'pandas.core.frame.DataFrame'>
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 10 : Problem 2 - Data Info

- The purpose of the dataset is to study the data obtained from different colleges where we have to perform the exploratory data analysis to deduct some inferences.
- Also, principal component analysis is to be performed to reduce the dimensions (remove redundant dimensions) for better analysis of data with the data giving the highest variance so the final dimensions will show 0 correlation in them.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000
mean	3001.638353	2018.804376	779.972973	27.558559	55.796654	3699.907336	855.298584	10440.669241	4357.526384	549.380952	1340.642214	72.660232	79.702703
std	3870.201484	2451.113971	929.176190	17.640364	19.804778	4850.420531	1522.431887	4023.016484	1096.696416	165.105360	677.071454	16.328155	14.722359
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	2340.000000	1780.000000	96.000000	250.000000	8.000000	24.000000
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	7320.000000	3597.000000	470.000000	850.000000	62.000000	71.000000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	9990.000000	4200.000000	500.000000	1200.000000	75.000000	82.000000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	12925.000000	5050.000000	600.000000	1700.000000	85.000000	92.000000
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000	31643.000000	21836.000000	21700.000000	8124.000000	2340.000000	6800.000000	103.000000	100.000000

Figure 12: Problem 2- Data Description.

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Gra
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	

Figure 11:Problem 2 - Sample Dataset.

Exploratory Data Analysis

Null Data Detection

```
Names          0
Apps           0
Accept         0
Enroll         0
Top10perc      0
Top25perc      0
F.Undergrad    0
P.Undergrad    0
Outstate       0
Room.Board     0
Books          0
Personal       0
PhD            0
Terminal       0
S.F.Ratio      0
perc.alumni    0
Expend         0
Grad.Rate      0
dtype: int64
```

Figure 13: Problem 2 - Null Data

- The “Names” column can be removed from the dataset to do exploratory data analysis.
- There is no null data in the dataset.

Duplicate Detection.

```
Number of duplicate rows = 0
```

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

Figure 14: Problem 2 - Duplicate Data

- There are no duplicate records to fix.

Outlier Treatment.

- From the boxplot (), we can see that we need to treat outliers expect for “Top25perc” as below.

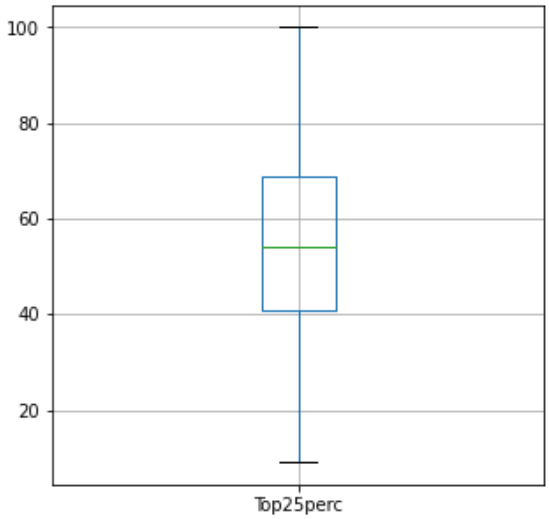


Figure 15: Top25Percent Boxplot.

- We shall treat outliers by treating the data beyond the IQR to the lower and upper bounds.
- For the higher outliers we will treat it to get it at 95 percentile values.
- Lower-level outliers will be treated to get it at 5 percentile values.

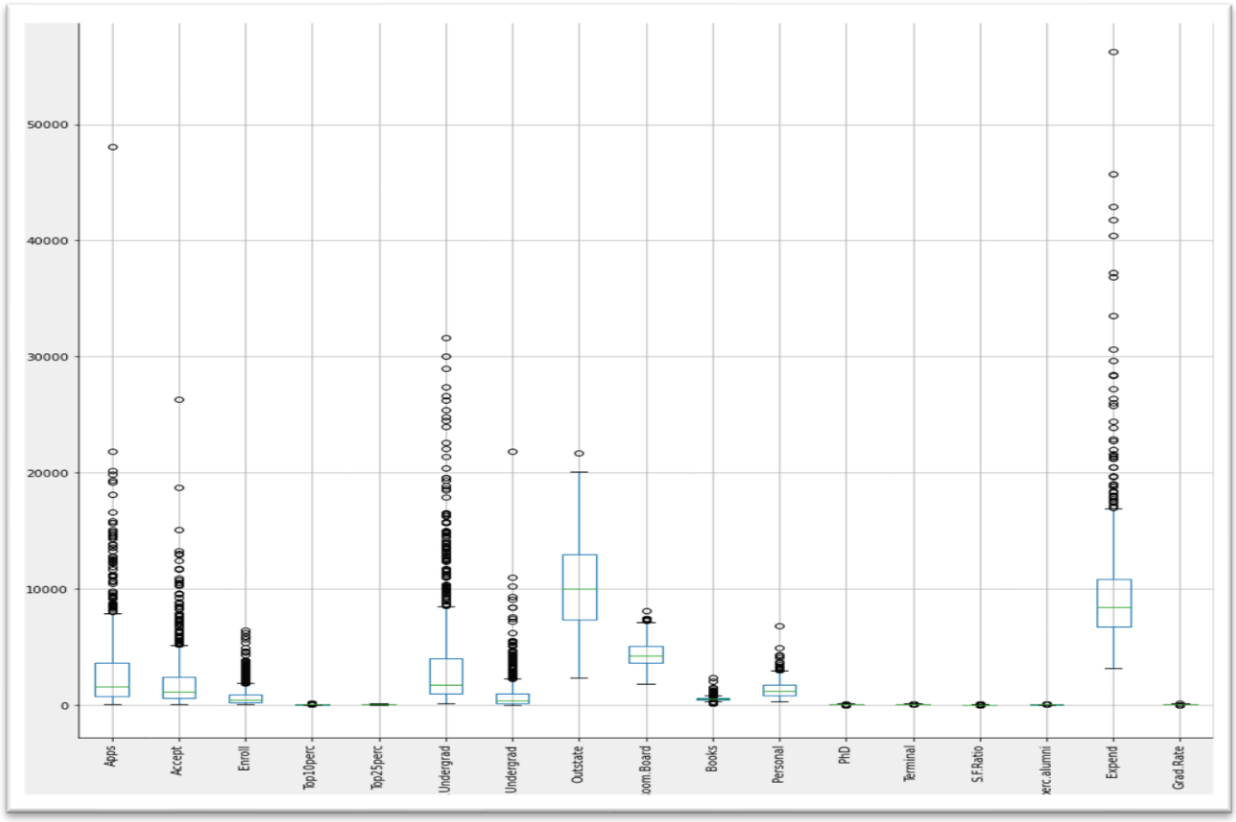


Figure 16: Before Outlier Treatment boxplot.

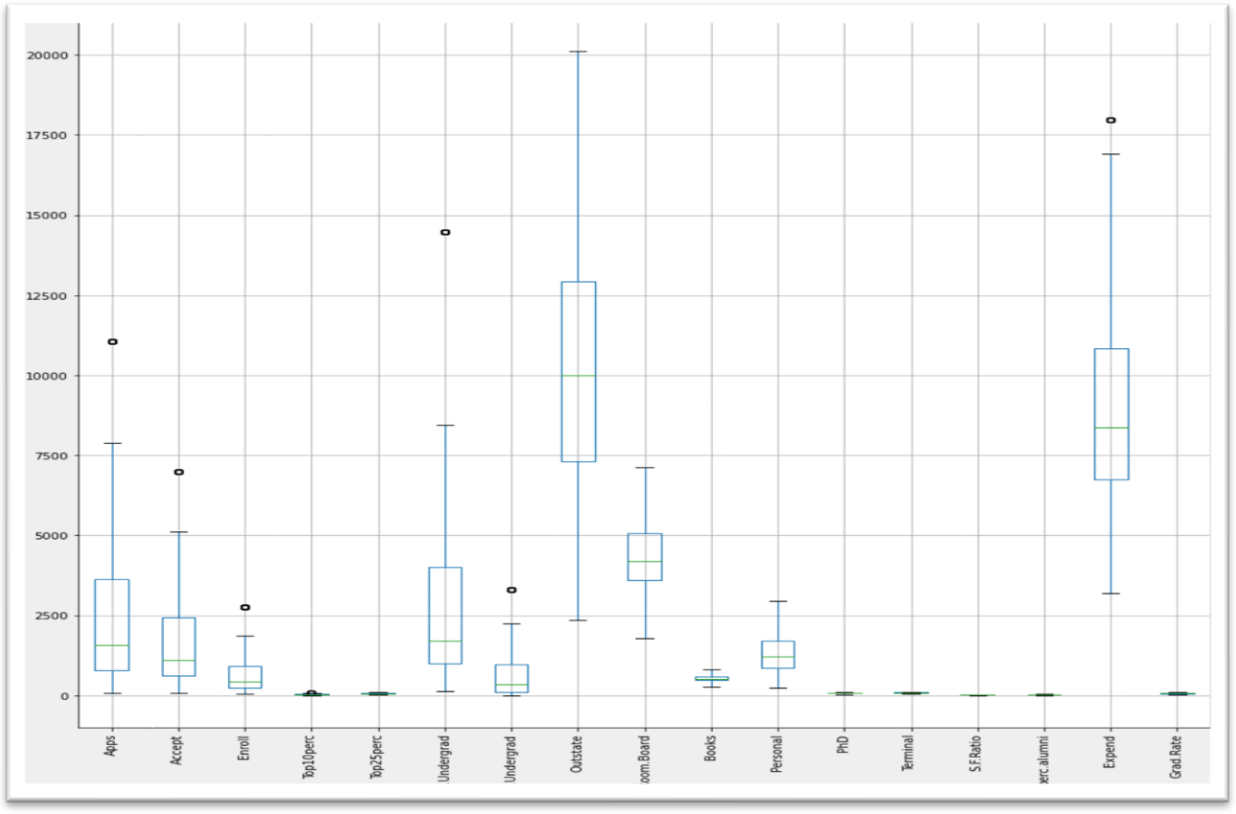


Figure 17: After Outlier Treatment.

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

From the Treatment done above we have a dataset with no extreme outliers on which univariate and multivariate analysis can be performed.

The further analysis would be done on the outlier treated data.

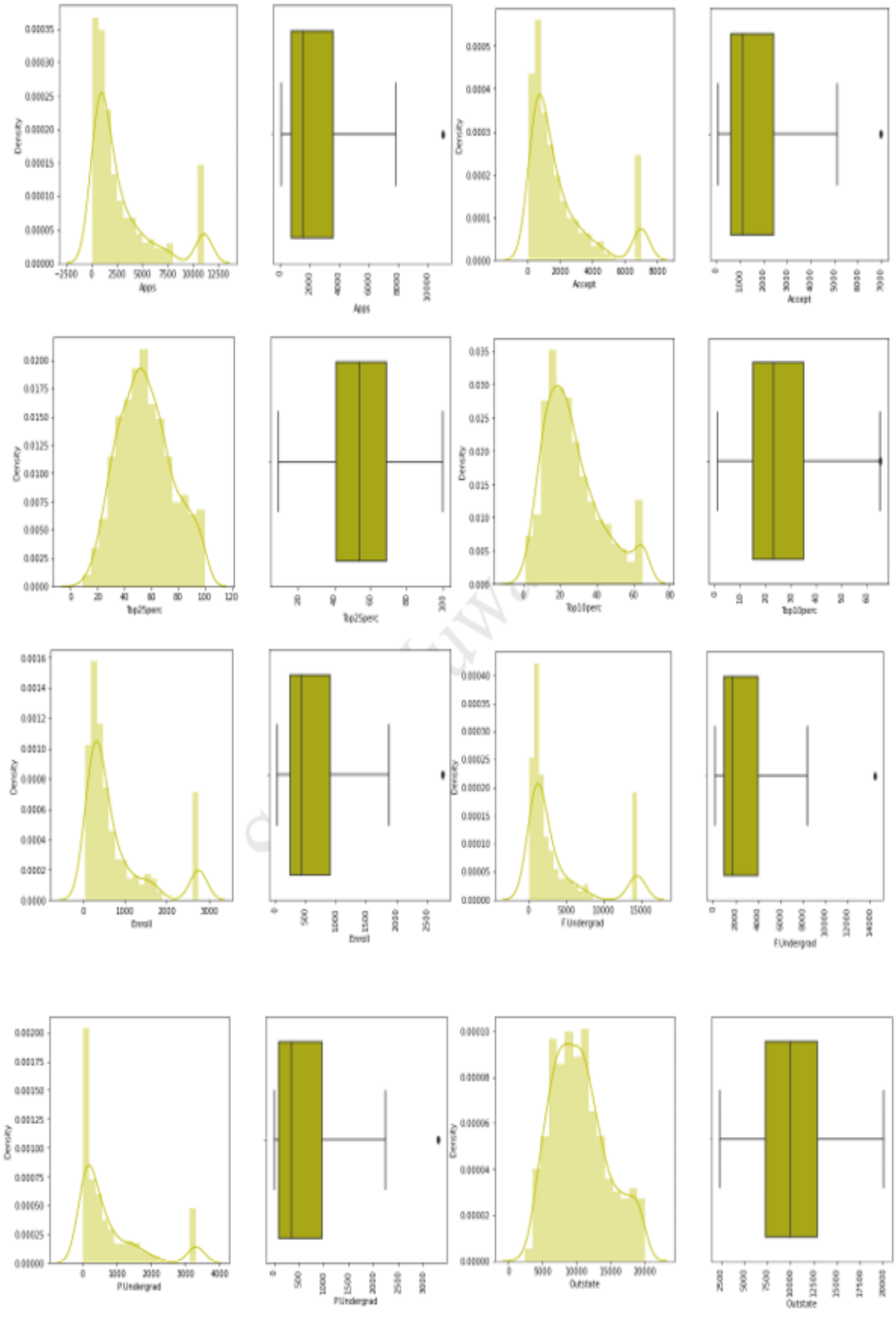
Univariate Analysis.

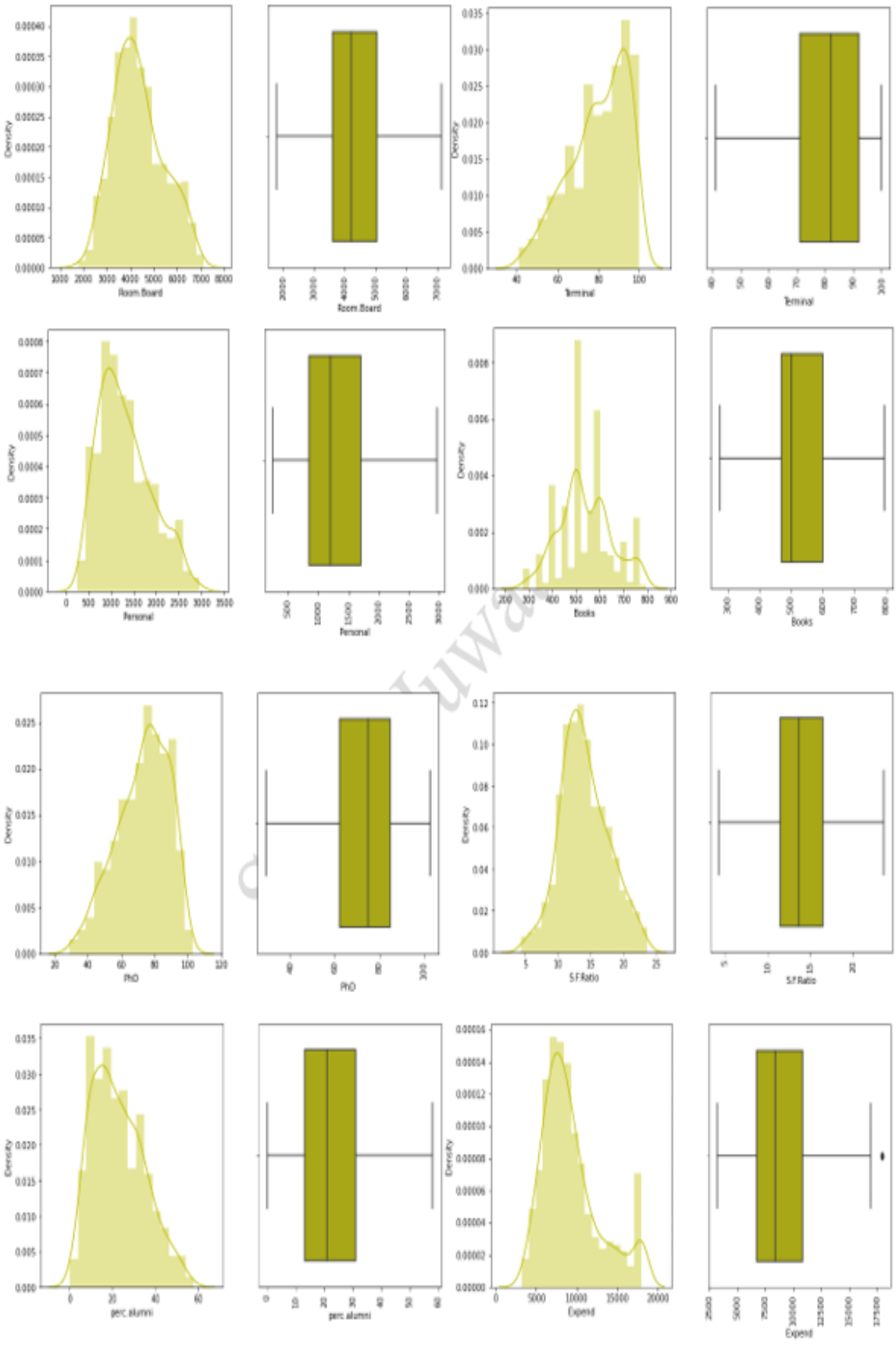
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	7
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	
mean	2856.956242	1917.760103	748.335907	26.853024	55.796654	3678.852767	744.579408	10436.548263	4347.803089	539.029086	1311.275418	
std	3120.470980	1942.822994	781.271463	15.607194	19.804778	4414.345270	940.269547	4013.095875	1073.326060	110.372183	579.698842	
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	2340.000000	1780.000000	275.000000	250.000000	
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	7320.000000	3597.000000	470.000000	850.000000	
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	9990.000000	4200.000000	500.000000	1200.000000	
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	12925.000000	5050.000000	600.000000	1700.000000	
max	11066.200000	6979.200000	2757.000000	65.200000	100.000000	14477.800000	3303.600000	20100.000000	7131.000000	795.000000	2958.000000	1

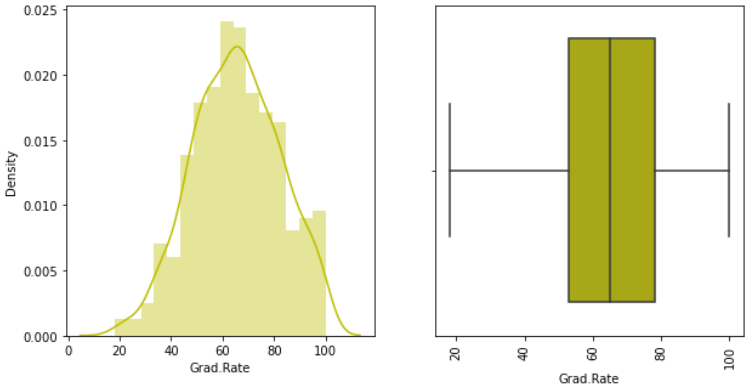
Figure 18: Univariate Analysis - Data Description

Inference from the below graphs and the data description.

- Apps: The distribution is right skewed with a range of 81 - 11066.2.
- Accept: It is right skewed
- Enrol: It is right Skewed.
- Top10perc: It is normally distributed.
- Top25perc: It is normally distributed.
- F. Undergrad: It is right skewed.
- P.Undergrad: It is right skewed.
- Outstate: It is normally distributed.
- Room.Board: It is normally distributed.
- Books: It is normally distributed.
- Personal: It is right skewed.
- PhD: it is left skewed.
- Terminal: It is left skewed.
- S.F.Ratio: It is right skewed.
- perc.alumni: It is right skewed.
- Expend: It is right skewed.
- Grad.Rate: It is normally distributed


Density
Apps
Accept
Top25perc
Top10perc
Enroll
F.Undergrad
P.Undergrad
Outstate


Density
Room.Board
Terminal
Personal
Books
PhD
S.F.Ratio
perc.alumni
Expend



Bi-Variate Analysis

Correlation matrix:

	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	1	0.933721	0.869927	0.32385	0.362402	0.816797	0.501112	0.063205	0.182724	0.232991	0.229238	0.446195	0.41673	0.114067	-0.100481	0.254206	0.146968
Accept	0.933721	1	0.921169	0.220338	0.267999	0.870428	0.556221	-0.014237	0.110183	0.21363	0.25602	0.407037	0.383561	0.182299	-0.1624	0.166054	0.069452
Enroll	0.869927	0.921169	1	0.167349	0.22457	0.947269	0.641672	-0.160551	-0.037202	0.213123	0.347087	0.360814	0.33565	0.267559	-0.213467	0.057268	-0.040492
Top10perc	0.32385	0.220338	0.167349	1	0.913717	0.103675	-0.146407	0.563581	0.35506	0.153943	-0.120176	0.54649	0.510105	-0.382899	0.450047	0.662904	0.498954
Top25perc	0.362402	0.267999	0.22457	0.913717	1	0.168115	-0.068976	0.490749	0.329582	0.17175	-0.088003	0.554801	0.527806	-0.290843	0.412259	0.576125	0.486738
F.Undergrad	0.816797	0.870428	0.947269	0.103675	0.168115	1	0.682826	-0.233126	-0.079608	0.202379	0.364438	0.331159	0.311108	0.308149	-0.270638	0.000221	-0.103195
P.Undergrad	0.501112	0.556221	0.641672	-0.146407	-0.068976	0.682826	1	-0.335734	-0.07481	0.136037	0.331455	0.150676	0.142498	0.350515	-0.393735	-0.166944	-0.276314
Outstate	0.063205	-0.014237	-0.160551	0.563581	0.490749	-0.233126	-0.335734	1	0.659314	-0.00434	-0.330512	0.401629	0.419211	-0.578969	0.562757	0.773354	0.580301
Room.Board	0.182724	0.110183	-0.037202	0.35506	0.329582	-0.079608	-0.07481	0.659314	1	0.107052	-0.226698	0.352936	0.383446	-0.382253	0.272121	0.581051	0.430189
Books	0.232991	0.21363	0.213123	0.153943	0.17175	0.202379	0.136037	-0.00434	0.107052	1	0.237367	0.152585	0.16894	-0.003635	-0.048824	0.145813	-0.003494
Personal	0.229238	0.25602	0.347087	-0.120176	-0.088003	0.364438	0.331455	-0.330512	-0.226698	0.237367	1	-0.017993	-0.034483	0.186528	-0.30905	-0.167221	-0.289016
PhD	0.446195	0.407037	0.360814	0.54649	0.554801	0.331159	0.150676	0.401629	0.352936	0.152585	-0.017993	1	0.866535	-0.132716	0.243272	0.523549	0.318781
Terminal	0.41673	0.383561	0.33565	0.510105	0.527806	0.311108	0.142498	0.419211	0.383446	0.16894	-0.034483	0.866535	1	-0.152132	0.263103	0.527427	0.293958
S.F.Ratio	0.114067	0.182299	0.267559	-0.382899	-0.290843	0.308149	0.350515	-0.578969	-0.382253	-0.003635	0.186528	-0.132716	-0.152132	1	-0.413352	-0.655492	-0.317034
perc.alumni	-0.100481	-0.1624	-0.213467	0.450047	0.412259	-0.270638	-0.393735	0.562757	0.272121	-0.048824	-0.30905	0.243272	0.263103	-0.413352	1	0.459584	0.491641
Expend	0.254206	0.166054	0.057268	0.662904	0.576125	0.000221	-0.166944	0.773354	0.581051	0.145813	-0.167221	0.523549	0.527427	-0.655492	0.459584	1	0.424709
Grad.Rate	0.146968	0.069452	-0.040492	0.498954	0.486738	-0.103195	-0.276314	0.580301	0.430189	-0.003494	-0.289016	0.318781	0.293958	-0.317034	0.491641	0.424709	1

Figure 19: Correlation Matrix

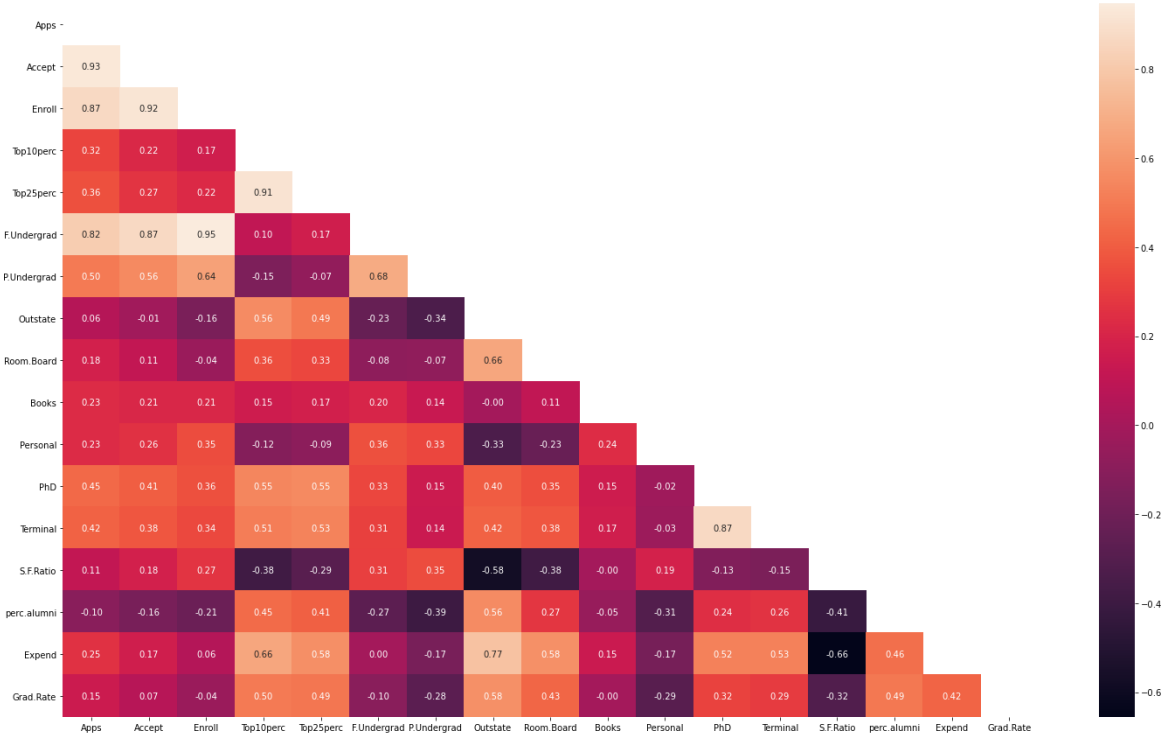


Figure 20: Corelation Heatmap

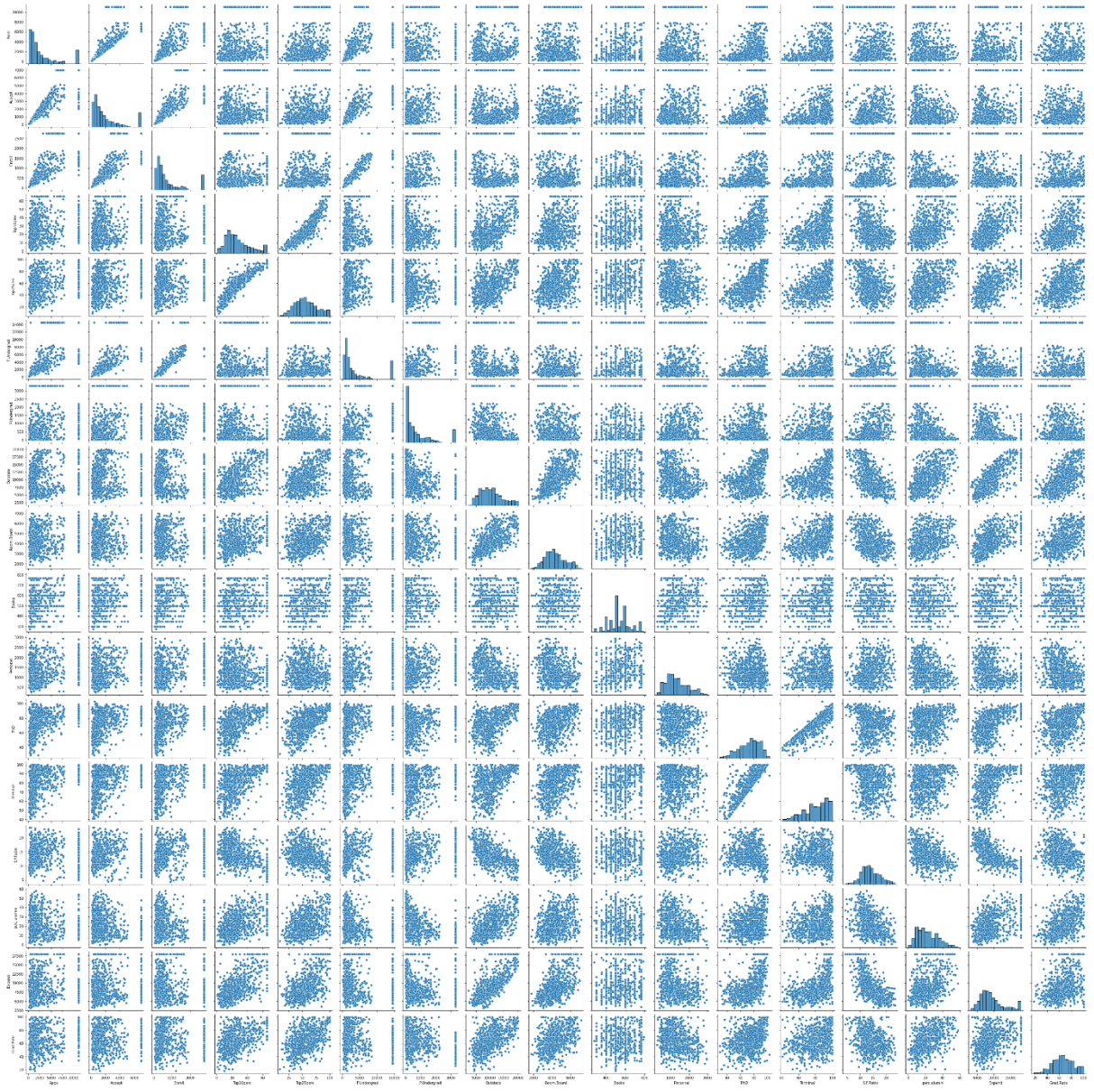


Figure 21: Pair Plot

- Bivariate analysis is done using the heatmap of the corelation matrix wherein the dependency between two variables is checked.
- “Apps” have high correlation between “Accept” and “Enrol”.
- The correlation indicates how two variables are dependent on one another and to what extent.
- The objective of PCA is to reduce this correlation.

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

From the boxplot of the outlier treated data Figure 17: After Outlier Treatment. We can see that there are some extreme variations in the range of data for e.g., Grad.Rate and Outstate where there is difference in the range making it difficult to compare the two data. Also, a variable in the dataset with high standard deviation will have higher weight of calculation than that of low standard deviation variable. Hence, it is necessary to scale the data to standardisation the range of all the dimensions.

PCA calculates new data axis depending on the deviation of the data.

To get the output we use Z-Score method to scale the data.

	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	-0.38383	-0.3532	-0.03501	-0.247034	-0.191827	-0.179951	-0.220908	-0.74717	-0.976849	-0.80715	1.534067	-0.18928	-0.13574	1.107193	-0.877388	-0.62198	-0.32732
1	-0.21516	0.003214	-0.3027	-0.695834	-1.353911	-0.22574	0.513397	0.459655	1.959843	1.912681	0.325766	-2.82658	-1.91778	-0.50276	-0.547691	0.361374	-0.56434
2	-0.45823	-0.42273	-0.52812	-0.311148	-0.292878	-0.599082	-0.687032	0.20283	-0.557322	-1.26045	-0.25249	-1.28279	-0.98433	-0.31175	0.606247	-0.14413	-0.68285
3	-0.78242	-0.80798	-0.78299	2.125195	1.677612	-0.718316	-0.725344	0.629209	1.02756	-0.80715	-0.75308	1.225853	1.207869	-1.73069	1.183216	2.462296	-0.38658
4	-0.85425	-0.91254	-0.88802	-0.695834	-0.596031	-0.777479	0.13241	-0.71725	-0.212377	2.054112	0.325766	0.196665	-0.56003	-0.58462	-1.701629	0.472798	-1.69017

Figure 22: Scaled data Sample

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
count	7.77E+02	7.77E+02	7.77E+02	7.77E+02	7.77E+02	7.77E+02	7.77E+02	7.77E+02	7.77E+02	7.77E+02	7.77E+02	7.77E+02	7.77E+02	7.77E+02	7.77E+02	7.77E+02	7.77E+02
mean	-1.61E-16	2.29E-16	-3.67E-16	2.41E-16	-1.55E-16	4.24E-17	-1.48E-16	1.96E-16	-1.33E-16	5.22E-16	8.17E-17	7.41E-17	-5.38E-16	-8.69E-17	4.42E-17	1.10E-16	5.51E-16
std	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
min	-8.90E-01	-9.51E-01	-9.14E-01	-1.66E+00	-2.36E+00	-8.02E-01	-7.91E-01	-2.02E+00	-2.39E+00	-2.39E+00	-1.83E+00	-2.83E+00	-2.75E+00	-2.66E+00	-1.87E+00	-1.71E+00	-2.82E+00
25%	-6.67E-01	-6.77E-01	-6.49E-01	-7.60E-01	-7.48E-01	-6.09E-01	-6.91E-01	-7.77E-01	-7.00E-01	-6.26E-01	-7.96E-01	-7.04E-01	-6.31E-01	-6.94E-01	-7.95E-01	-7.04E-01	-7.42E-01
50%	-4.17E-01	-4.16E-01	-4.03E-01	-2.47E-01	-9.08E-02	-4.47E-01	-4.17E-01	-1.11E-01	-1.38E-01	-3.54E-01	-1.92E-01	1.32E-01	1.47E-01	-1.21E-01	-1.36E-01	-2.45E-01	-3.11E-02
75%	2.46E-01	2.61E-01	1.97E-01	5.22E-01	6.67E-01	7.39E-02	2.37E-01	6.20E-01	6.55E-01	5.53E-01	6.71E-01	7.76E-01	8.54E-01	6.71E-01	6.89E-01	4.47E-01	7.39E-01
max	2.63E+00	2.61E+00	2.57E+00	2.46E+00	2.23E+00	2.45E+00	2.72E+00	2.41E+00	2.59E+00	2.32E+00	2.84E+00	1.93E+00	1.42E+00	2.61E+00	2.91E+00	2.46E+00	2.04E+00

Figure 23: Scaled data description

From the above description and sample data of the scaled dataset we can see that the scale of each data is standardised when the mean tends to 0 and the standard deviation tends to 1.

Also, the data is now evenly spread out making to easier to derive deductions from it.

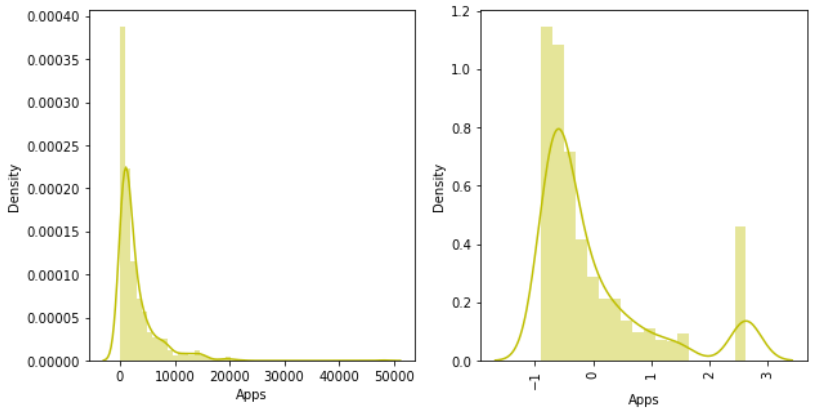


Figure 24: Scaled Data Plot Example.

2.3 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.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1	0.933721	0.869927	0.32385	0.362402	0.816797	0.501112	0.063205	0.182724	0.232991	0.229238	0.446195	0.41673	0.114067	-0.100481	0.254206	0.146968
Accept	0.933721	1	0.921169	0.220338	0.267999	0.870428	0.556221	-0.014237	0.110183	0.21363	0.25602	0.407037	0.383561	0.182299	-0.1624	0.166054	0.069452
Enroll	0.869927	0.921169	1	0.167349	0.22457	0.947269	0.641672	-0.160551	-0.037202	0.213123	0.347087	0.360814	0.33565	0.267559	-0.213467	0.057268	-0.040492
Top10perc	0.32385	0.220338	0.167349	1	0.913717	0.103675	-0.146407	0.563581	0.35506	0.153943	-0.120176	0.54649	0.510105	-0.382899	0.450047	0.662904	0.498954
Top25perc	0.362402	0.267999	0.22457	0.913717	1	0.168115	-0.068976	0.490749	0.329582	0.17175	-0.088003	0.554801	0.527806	-0.290843	0.412259	0.576125	0.486738
F.Undergrad	0.816797	0.870428	0.947269	0.103675	0.168115	1	0.682826	-0.233126	-0.079608	0.202379	0.364438	0.331159	0.311108	0.308149	-0.270638	0.000221	-0.103195
P.Undergrad	0.501112	0.556221	0.641672	-0.146407	-0.068976	0.682826	1	-0.335734	-0.07481	0.136037	0.331455	0.150676	0.142498	0.350515	-0.393735	-0.166944	-0.276314
Outstate	0.063205	-0.014237	-0.160551	0.563581	0.490749	-0.233126	-0.335734	1	0.659314	-0.00434	-0.330512	0.401629	0.419211	-0.578969	0.562757	0.773354	0.580301
Room.Board	0.182724	0.110183	-0.037202	0.35506	0.329582	-0.079608	-0.07481	0.659314	1	0.107052	-0.226698	0.352936	0.383446	-0.382253	0.272121	0.581051	0.430189
Books	0.232991	0.21363	0.213123	0.153943	0.17175	0.202379	0.136037	-0.00434	0.107052	1	0.237367	0.152585	0.16894	-0.003635	-0.048824	0.145813	-0.003494
Personal	0.229238	0.25602	0.347087	-0.120176	-0.088003	0.364438	0.331455	-0.330512	-0.226698	0.237367	1	-0.017993	-0.034483	0.186528	-0.30905	-0.167221	-0.289016
PhD	0.446195	0.407037	0.360814	0.54649	0.554801	0.331159	0.150676	0.401629	0.352936	0.152585	-0.017993	1	0.866535	-0.132716	0.243272	0.523549	0.318781
Terminal	0.41673	0.383561	0.33565	0.510105	0.527806	0.311108	0.142498	0.419211	0.383446	0.16894	-0.034483	0.866535	1	-0.152132	0.263103	0.527427	0.293958
S.F.Ratio	0.114067	0.182299	0.267559	-0.382899	-0.290843	0.308149	0.350515	-0.578969	-0.382253	-0.003635	0.186528	-0.132716	-0.152132	1	-0.413352	-0.655492	-0.317034
perc.alumni	-0.100481	-0.1624	-0.213467	0.450047	0.412259	-0.270638	-0.393735	0.562757	0.272121	-0.048824	-0.30905	0.243272	0.263103	-0.413352	1	0.459584	0.491641
Expend	0.254206	0.166054	0.057268	0.662904	0.576125	0.000221	-0.166944	0.773354	0.581051	0.145813	-0.167221	0.523549	0.527427	-0.655492	0.459584	1	0.424709
Grad.Rate	0.146968	0.069452	-0.040492	0.498954	0.486738	-0.103195	-0.276314	0.580301	0.430189	-0.003494	-0.289016	0.318781	0.293958	-0.317034	0.491641	0.424709	1

Figure 25: corelation matrix of scaled data.

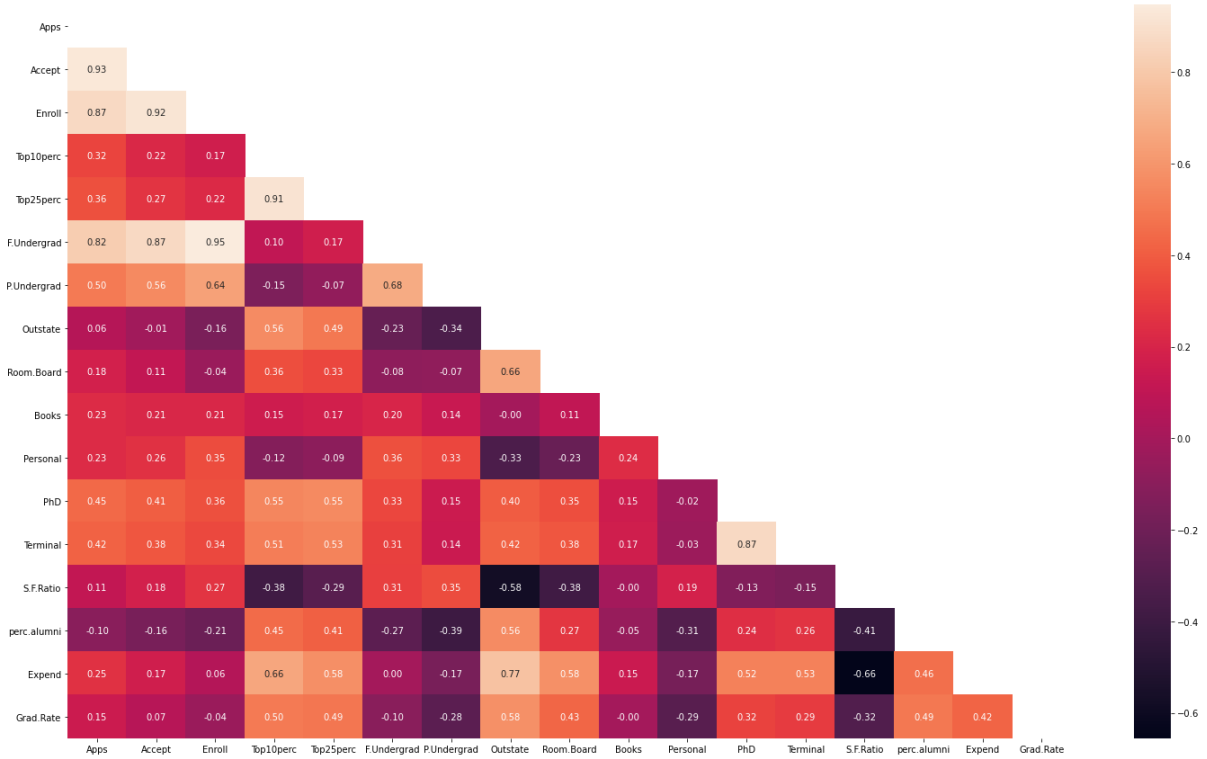


Figure 26: Heatmap of correlation of Scaled data.

Covariance Matrix:

```
np.round(std_df_scaled.cov(),2)
```

	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	1	0.93	0.87	0.32	0.36	0.82	0.5	0.06	0.18	0.23	0.23	0.45	0.42	0.11	-0.1	0.25	0.15
Accept	0.93	1	0.92	0.22	0.27	0.87	0.56	-0.01	0.11	0.21	0.26	0.41	0.38	0.18	-0.16	0.17	0.07
Enroll	0.87	0.92	1	0.17	0.22	0.95	0.64	-0.16	-0.04	0.21	0.35	0.36	0.34	0.27	-0.21	0.06	-0.04
Top10perc	0.32	0.22	0.17	1	0.91	0.1	-0.15	0.56	0.36	0.15	-0.12	0.55	0.51	-0.38	0.45	0.66	0.5
Top25perc	0.36	0.27	0.22	0.91	1	0.17	-0.07	0.49	0.33	0.17	-0.09	0.56	0.53	-0.29	0.41	0.58	0.49
F.Undergrad	0.82	0.87	0.95	0.1	0.17	1	0.68	-0.23	-0.08	0.2	0.36	0.33	0.31	0.31	-0.27	0	-0.1
P.Undergrad	0.5	0.56	0.64	-0.15	-0.07	0.68	1	-0.34	-0.07	0.14	0.33	0.15	0.14	0.35	-0.39	-0.17	-0.28
Outstate	0.06	-0.01	-0.16	0.56	0.49	-0.23	-0.34	1	0.66	0	-0.33	0.4	0.42	-0.58	0.56	0.77	0.58
Room.Board	0.18	0.11	-0.04	0.36	0.33	-0.08	-0.07	0.66	1	0.11	-0.23	0.35	0.38	-0.38	0.27	0.58	0.43
Books	0.23	0.21	0.21	0.15	0.17	0.2	0.14	0	0.11	1	0.24	0.15	0.17	0	-0.05	0.15	0
Personal	0.23	0.26	0.35	-0.12	-0.09	0.36	0.33	-0.33	-0.23	0.24	1	-0	-0.03	0.19	-0.31	-0.17	-0.29
PhD	0.45	0.41	0.36	0.55	0.56	0.33	0.15	0.4	0.35	0.15	-0.02	1	0.87	-0.13	0.24	0.52	0.32
Terminal	0.42	0.38	0.34	0.51	0.53	0.31	0.14	0.42	0.38	0.17	-0.03	0.87	1	-0.15	0.26	0.53	0.29
S.F.Ratio	0.11	0.18	0.27	-0.38	-0.29	0.31	0.35	-0.58	-0.38	0	0.19	-0.1	-0.15	1	-0.41	-0.66	-0.32
perc.alumni	-0.1	-0.16	-0.21	0.45	0.41	-0.27	-0.39	0.56	0.27	-0.05	-0.31	0.24	0.26	-0.41	1	0.46	0.49
Expend	0.25	0.17	0.06	0.66	0.58	0	-0.17	0.77	0.58	0.15	-0.17	0.52	0.53	-0.66	0.46	1	0.43
Grad.Rate	0.15	0.07	-0.04	0.5	0.49	-0.1	-0.28	0.58	0.43	0	-0.29	0.32	0.29	-0.32	0.49	0.43	1

Inference:

- Covariance shows the direction of the linear relationship between variables.
- Correlation measures both the strength and direction of the linear relationship between two variables.
- Correlation is a function of the covariance.
- Covariance indicates the relationship of two variables whenever one variable changes. If an increase in one variable results in an increase in the other variable, both variables are said to have a positive covariance. Decreases in one variable also cause a decrease in the other.

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

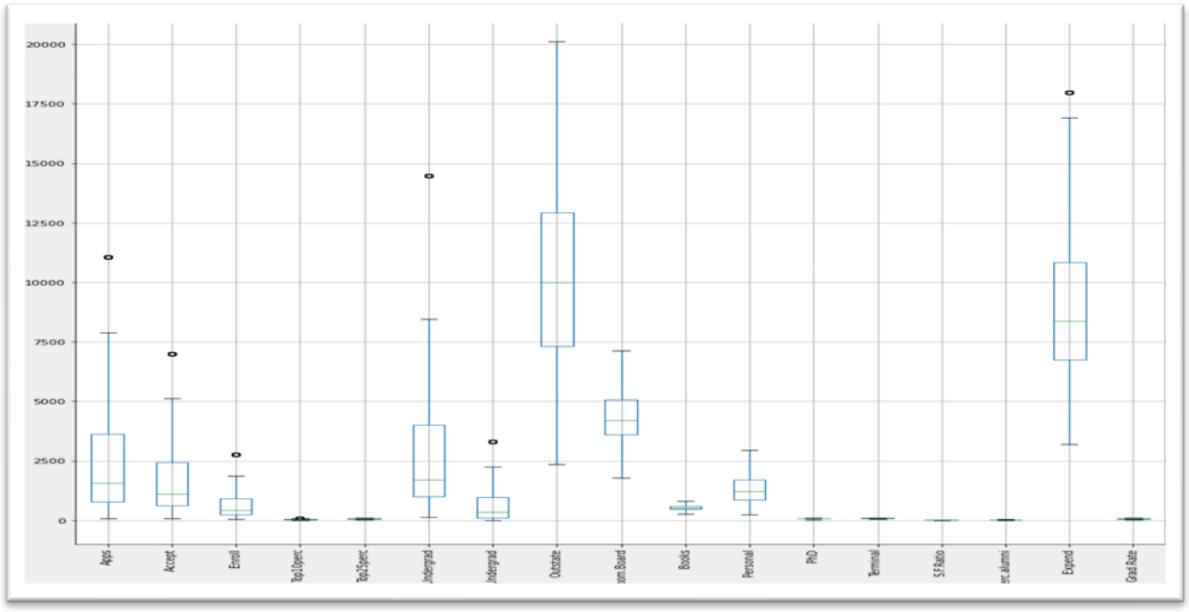


Figure 27: Outlier Treated Data.

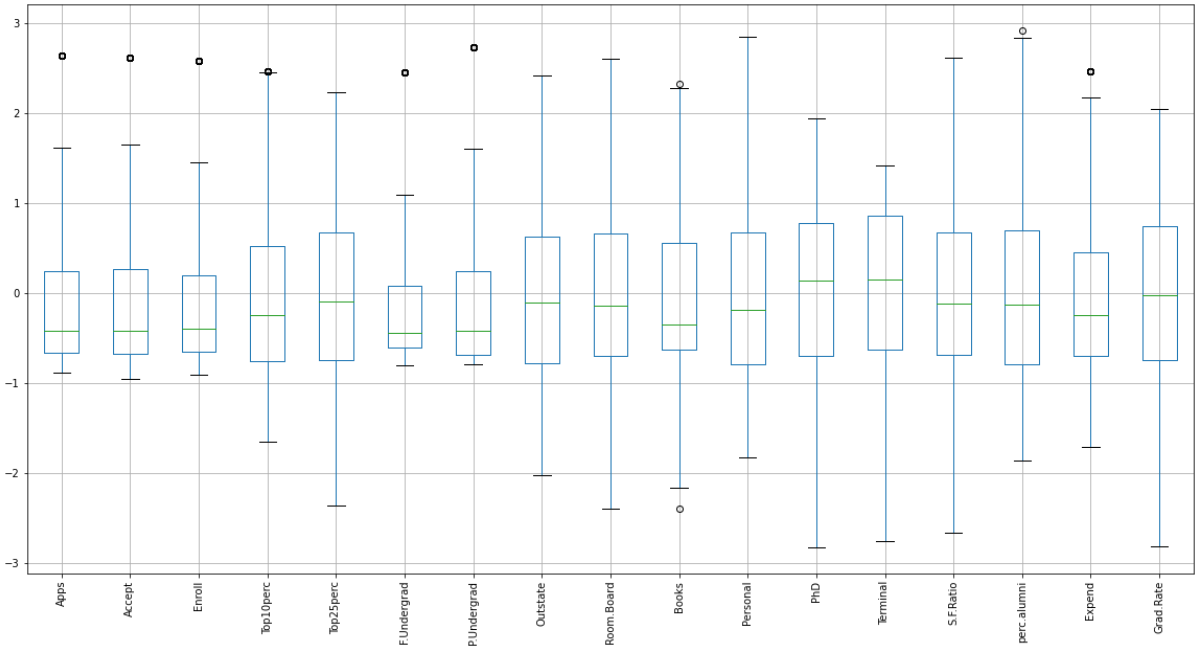


Figure 28: After outlier and Scaling Treatment on Data.

From the Figure 27: Outlier Treated Data. And Figure 28: After outlier and Scaling Treatment on Data.

- Before the treatment the dataset had many outliers except one dimension. Post treatment we have negligible number of outliers.
- Scaling the data helped to standardise the data thus giving same weight to all and further supplementing by giving PCA relevant axis.

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

Bartlett's test of sphericity

To perform PCA we need to perform Bartlett's test of sphericity to test the hypothesis that the variables are uncorrelated in the population.

- H0: All variables in the data are uncorrelated

- Ha: At least one pair of variables in the data are correlated

After performing the test, the p-value is 0 which means that the null hypothesis is rejected and at least one pair of variables are correlated hence PCA is recommended.

Kaiser-Meyer-Olkin (KMO)

The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

Here for the dataset, the p-value is 0.86 which indicates that after performing PCA there would be a considerable reduction of dimensionality.

Step 1: Create Covariance matrix as question 2.4

Step 2: Get Eigen vector and Eigen Values.

Performing PCA on the scaled and treated data using the Sklearn, we find the following outputs.

Eigen Vectors:

	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	0.242671	0.208096	0.164564	0.344634	0.337858	0.134288	0.014513	0.297305	0.251192	0.093568	-0.048467	0.324668	0.32051	-0.178477	0.198618	0.340157	0.248645
1	0.32493	0.357756	0.395824	-0.07539	-0.036721	0.406244	0.354917	-0.237362	-0.123789	0.106015	0.235469	0.070652	0.059666	0.247835	-0.243262	-0.135748	-0.160608
2	-0.09771	-0.125144	-0.094442	0.072387	0.046337	-0.08724	-0.038696	-0.020591	0.026069	0.713558	0.521834	-0.057258	-0.037458	-0.258376	-0.109907	0.17293	-0.231028
3	0.10256	0.121914	0.01425	-0.375563	-0.427876	0.014617	0.207265	0.253852	0.566794	-0.047279	-0.107878	-0.123471	-0.073147	-0.283024	-0.229944	0.220176	-0.074147
4	0.228743	0.202792	0.172168	0.145905	0.120537	0.115073	-0.132039	0.042968	-0.090207	-0.01663	0.062935	-0.547357	-0.585124	-0.226759	0.13831	0.028307	0.28788
5	0.047641	0.033134	-0.038976	-0.083767	-0.021492	-0.054996	-0.051645	-0.013967	0.257757	0.608724	-0.384138	-0.062174	-0.047922	0.442094	-0.005596	-0.238206	0.372585
6	-0.012378	0.001415	-0.007928	-0.258268	-0.234717	-0.027916	-0.093659	0.104399	0.125975	-0.139286	0.656949	0.096114	0.098447	0.174587	0.321858	-0.150524	0.463708
7	-0.034103	-0.102522	-0.134762	0.289095	0.336249	-0.122385	0.054191	0.023889	0.355686	-0.256097	0.251642	-0.047057	-0.116058	0.215538	-0.635277	-0.086414	0.162586
8	-0.184655	-0.189697	-0.052018	0.110852	0.189925	0.000141	0.736103	0.014611	0.217093	-0.016293	0.032822	-0.16756	-0.12924	0.12102	0.457695	-0.050535	-0.129825
9	-0.13405	-0.123208	-0.047956	-0.071443	-0.045326	0.011266	0.423776	-0.187206	-0.304567	0.074595	-0.09215	0.125222	0.075285	-0.458497	-0.250493	-0.065288	0.580722
10	-0.067941	-0.028689	-0.022975	-0.006573	-0.132078	-0.036376	0.189392	0.609931	-0.462002	0.051438	0.017551	-0.036141	-0.104786	0.383414	-0.16927	0.396899	0.073914
11	-0.027399	-0.127528	-0.018056	0.045736	-0.158456	0.078883	-0.03586	-0.557302	0.105909	-0.04859	-0.009335	0.186806	-0.264231	0.233517	0.058451	0.66401	0.141881
12	-0.476266	-0.208677	0.265982	-0.016249	0.034743	0.520755	-0.161438	0.007532	0.088839	-0.009112	-0.023451	-0.434191	0.367221	0.044001	-0.069753	0.117934	0.086571
13	0.350002	0.112838	-0.225004	-0.032292	0.026458	-0.35024	0.101786	-0.223348	-0.09107	-0.042364	0.029468	-0.533329	0.522451	0.086158	0.011708	0.225353	0.052797
14	-0.025472	0.040814	-0.033748	-0.723554	0.658266	0.010534	-0.038264	-0.002551	-0.033452	-0.00828	-0.001347	0.056657	-0.089005	0.008573	-0.008853	0.159987	-0.007184
15	0.573869	-0.643625	-0.258382	-0.05319	-0.003703	0.413881	-0.032499	0.095191	-0.022449	0.002769	-0.010841	0.005145	-0.001062	-0.014941	0.003686	-0.064634	-0.018177
16	0.151052	-0.452767	0.750068	-0.058995	0.014736	-0.45183	-0.004973	0.00474	0.017986	-0.002618	-0.018067	-0.000332	0.015286	-0.001267	-0.02638	0.009781	0.002276

Figure 29: Eigen Vectors Matrix

Eigen values:

Eigen Values: [0.33151857 0.28373652 0.06464061 0.05855307 0.05274046 0.04497099
 0.03449059 0.03257588 0.02603662 0.02245497 0.01443066 0.00862682

0.00799196 0.00727087 0.00438662 0.0032887 0.00228608]

Step 3: Plot Scree plot and find cumulative explained variance

Refer question no **2.8** .

Step 4: Apply PCA to the number of decided Components to get loading and component output.

Refer question no **2.9**

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

After performing PCA and exporting the data of the eigenvectors from question 2.5 in a data frame with its original features we the following data frame

Below is the sample of this data frame from loading with all the components.

	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	0.242671	0.208096	0.164564	0.344634	0.337858	0.134288	0.014513	0.297305	0.251192	0.093568	-0.048467	0.324668	0.32051	-0.178477	0.198618	0.340157	0.248645
1	0.32493	0.357756	0.395824	-0.07539	-0.036721	0.406244	0.354917	-0.237362	-0.123789	0.106015	0.235469	0.070652	0.059666	0.247835	-0.243262	-0.135748	-0.160608
2	-0.09771	-0.125144	-0.094442	0.072387	0.046337	-0.08724	-0.038696	-0.020591	0.026069	0.713558	0.521834	-0.057258	-0.037458	-0.258376	-0.109907	0.17293	-0.231028
3	0.10256	0.121914	0.01425	-0.375563	-0.427876	0.014617	0.207265	0.253852	0.566794	-0.047279	-0.107878	-0.123471	-0.073147	-0.283024	-0.229944	0.220176	-0.074147
4	0.228743	0.202792	0.172168	0.145905	0.120537	0.115073	-0.132039	0.042968	-0.090207	-0.01663	0.062935	-0.547357	-0.585124	-0.226759	0.13831	0.028307	0.28788
5	0.047641	0.033134	-0.038976	-0.083767	-0.021492	-0.054996	-0.051645	-0.013967	0.257757	0.608724	-0.384138	-0.062174	-0.047922	0.442094	-0.005596	-0.238206	0.372585
6	-0.012378	0.001415	-0.007928	-0.258268	-0.234717	-0.027916	-0.093659	0.104399	0.125975	-0.139286	0.656949	0.096114	0.098447	0.174587	0.321858	-0.150524	0.463708
7	-0.034103	-0.102522	-0.134762	0.289095	0.336249	-0.122385	0.054191	0.023889	0.355686	-0.256097	0.251642	-0.047057	-0.116058	0.215538	-0.635277	-0.086414	0.162586
8	-0.184655	-0.189697	-0.052018	0.110852	0.189925	0.000141	0.736103	0.014611	0.217093	-0.016293	0.032822	-0.16756	-0.12924	0.12102	0.457695	-0.050535	-0.129825
9	-0.13405	-0.123208	-0.047956	-0.071443	-0.045326	0.011266	0.423776	-0.187206	-0.304567	0.074595	-0.09215	0.125222	0.075285	-0.458497	-0.250493	-0.065288	0.580722
10	-0.067941	-0.028689	-0.022975	-0.006573	-0.132078	-0.036376	0.189392	0.609931	-0.462002	0.051438	0.017551	-0.036141	-0.104786	0.383414	-0.16927	0.396899	0.073914
11	-0.027399	-0.127528	-0.018056	0.045736	-0.158456	0.078883	-0.03586	-0.557302	0.105909	-0.04859	-0.009335	0.186806	-0.264231	0.233517	0.058451	0.66401	0.141881
12	-0.476266	-0.208677	0.265982	-0.016249	0.034743	0.520755	-0.161438	0.007532	0.088839	-0.009112	-0.023451	-0.434191	0.367221	0.044001	-0.069753	0.117934	0.086571
13	0.350002	0.112838	-0.225004	-0.032292	0.026458	-0.35024	0.101786	-0.223348	-0.09107	-0.042364	0.029468	-0.533329	0.522451	0.086158	0.011708	0.225353	0.052797
14	-0.025472	0.040814	-0.033748	-0.723554	0.658266	0.010534	-0.038264	-0.002551	-0.033452	-0.00828	-0.001347	0.056657	-0.089005	0.008573	-0.008853	0.159987	-0.007184
15	0.573869	-0.643625	-0.258382	-0.05319	-0.003703	0.413881	-0.032499	0.095191	-0.022449	0.002769	-0.010841	0.005145	-0.001062	-0.014941	0.003686	-0.064634	-0.018177
16	0.151052	-0.452767	0.750068	-0.058995	0.014736	-0.45183	-0.004973	0.00474	0.017986	-0.002618	-0.018067	-0.000332	0.015286	-0.001267	-0.02638	0.009781	0.002276

Figure 30: Loading DF with eigenvectors

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]

The linear or explicit form of first PC is obtained by using the eigenvectors for PC1.

Below is the linear form of the first PC. Refer Figure 30: Loading DF with eigenvectors

PC1 = 0.24*Apps + 0.21*Accept + 0.16*Enroll + 0.34*Top10perc + 0.34*Top25perc + 0.13*F.Undergrad + 0.01* P.Undergrad + 0.30* Outstate + 0.25*Room.Board +0.09*Books + -0.05*Personal + 0.32*PhD + 0.32*Terminal + -0.18* S.F.Ratio +0.20*perc.alumni + 0.34*Expend + 0.25*Grad.Rate

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?

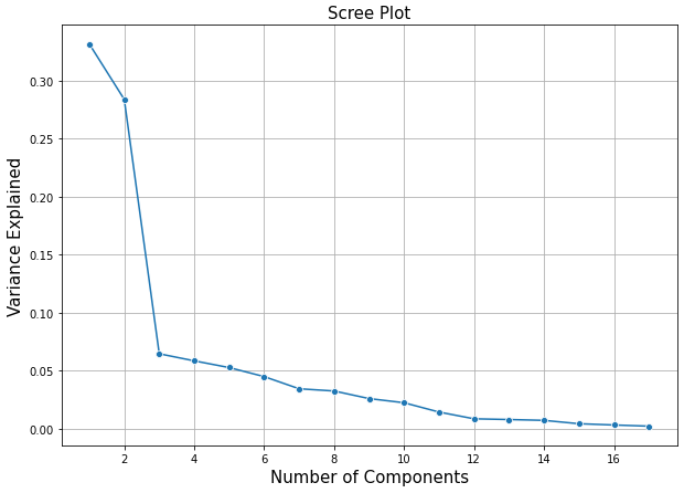


Figure 31: Scree Plot.

Cumulative Variance:

array([0.33151857, 0.61525509, 0.6798957 , 0.73844877, 0.79118924,
0.83616023, 0.87065082, 0.9032267 , 0.92926332, 0.95171829,
0.96614894, 0.97477576, 0.98276773, 0.9900386 , 0.99442522,
0.99771392, 1.])

Cumulative Plot :

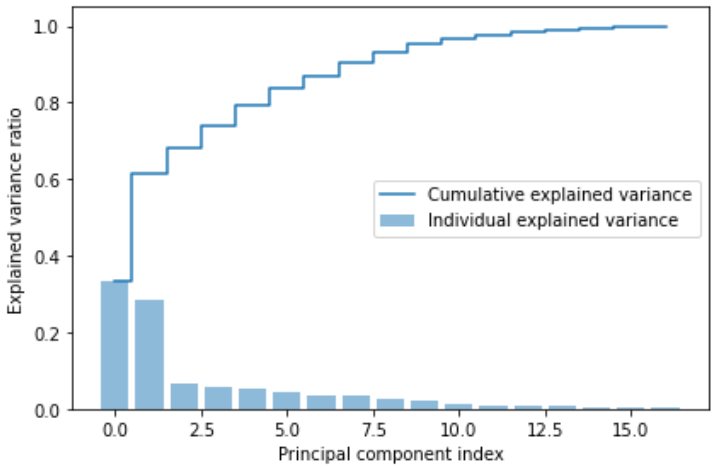


Figure 32: Cumulative Variance

From the above scree plot and the cumulative eigen variance, we can see that around 8 components give about 90% of the variance. Thus, the dimensions can be reduced from 17 to 8 given optimal solution.

The data can now be reoriented onto new axes by transforming the principal components into the direction of the eigenvectors.

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]

- After PCA we found out that the dimensions can be reduced to 8 components which indicates that the correlation between them 0 as per Figure 34: Correlation matrix of the selected PC.
- This indicates that the redundant dimensions are removed.
- Now the 8 dimensions/PCA represent almost 90% of the data.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
0	-1.736901	0.786523	0.091334	-1.018149	-0.351402	-0.765610	0.879426	0.976811
1	-1.598136	-0.332040	2.129009	2.898618	1.927792	1.364933	-0.337346	0.219223
2	-1.542800	-1.379268	-0.602489	0.005509	0.955652	-0.965602	-0.174358	-0.332625
3	3.181988	-2.993983	0.335530	-0.456311	-0.915076	-1.753031	-1.261560	0.241643
4	-1.785881	-0.202226	2.731233	0.689054	-1.194913	0.174538	-1.369389	0.101444
5	-0.549618	-1.823884	0.164431	-0.211133	0.244816	-0.839955	-1.791114	0.352368
6	0.232046	-1.661746	0.276293	0.957245	-1.712301	-0.370756	0.988602	0.062343
7	1.904425	-1.642138	-0.988320	-0.497628	-1.039238	-0.255906	0.226435	-0.304651
8	0.797788	-2.344255	-1.933846	0.354534	-0.240210	-0.984291	-0.302141	0.497219
9	-2.837048	-1.026997	2.106879	0.260420	2.173966	-0.123553	-0.535384	0.088376

Figure 33: DF after performing PCA

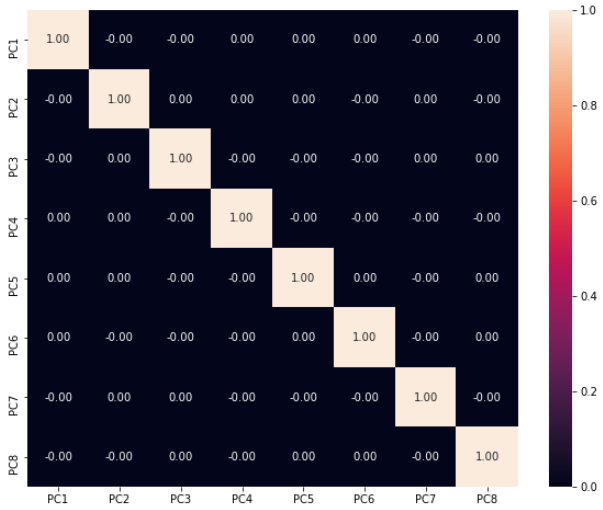


Figure 34: Correlation matrix of the selected PC

- We now have to apply the new PCA to the scaled data giving the below result as per Figure 35: PCA Transformed Data frame Heatmap.

Inference from the transformed data or the PC's:

- PC1: Show the no of students who have to pay outstate tuition.
- PC2: Show that the student admission depends highly on application, acceptance and enrolment where these components are highly correlated.
- PC3: Shows the cost of books for a student.
- PC4: Indicates the top 10 and top 25 percentage.
- PC5: Represents % of faculties with Ph.D and terminal degree.
- PC6: Represents the student/faculty ratio.
- PC7: Highlights estimated personal spending for a student and graduation rate
- PC8: Highlights the alumni members.

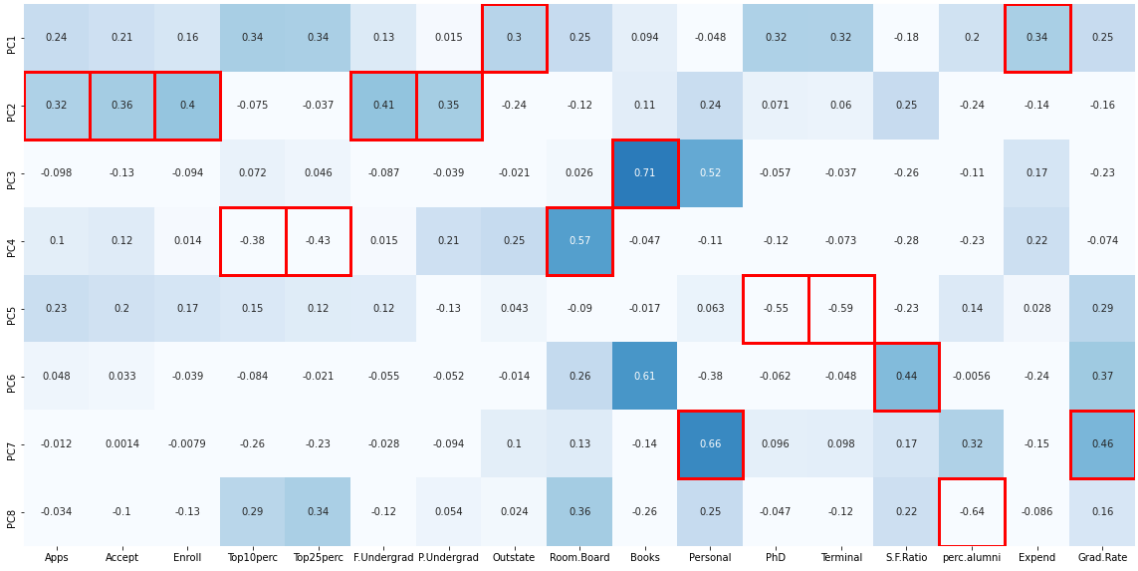


Figure 35: PCA Transformed Data frame Heatmap.

The End!