Advanced Statistics Project

Nandha Keshore Utti
PG-DSBA Online
March’ 22

Date: 26/06/2022

Contents

List of figures

List of tables

Problem 1 (Salary Analysis)

Problem Statement:

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.

One way ANOVA:

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

- **Education:**

H_0= The mean salary of students for different 'Education' levels is same
H_1= The mean salary of students is different in at least one of the 'Education' levels

- **Occupation:**

H_0= The mean salary of students for four different 'Occupation' levels is same
H_1= The mean salary of students is different in at least one of the 'Occupation' levels

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.

- **One-way ANOVA on Salary with respect to Education:**

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

Table.1

- **Conclusion:** Since the p value is less than the significance level (0.05), we can reject the null hypothesis and conclude that, at 95% confidence level, there is a difference in the mean salary of students in at least one 'Education' level.

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.

- **One-way ANOVA on Salary with respect to Occupation:**

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	0.884144	0.458508
Residual	36.0	1.528092e+11	4.244701e+09	NaN	NaN

Table.2

- **Conclusion:** Since the p value is greater than the significance level (0.05), we fail to reject the null hypothesis. So, at 95% confidence level, there is sufficient evidence to prove that the mean salary of students across four different Occupation levels is same

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)

- Null hypothesis is rejected for 'Education' feature.

Education	Salary
Bachelors	165152.933333
Doctorate	208427.000000
HS-grad	75038.777778

Table.3

- It can be observed that HS-grad students have significantly different mean salary compared to other 'Education' levels.

Two-way ANOVA

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]

- Let's do two-way ANOVA to check the interaction b/w two treatments.

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	31.257677	1.981539e-08
C(Occupation)	3.0	5.519946e+09	1.839982e+09	1.120080	3.545825e-01
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

Table.4

➢ **Insights:**

Education:

- F-stat value for 'Education' feature is ~31, which says that variability b/w education levels is 31 times than variability within the education levels.
- p-value is less than 0.05, so we can reject the null hypothesis and conclude that, at 95% confidence level, there is a difference in the mean salary of students in at least one 'Education' level.

Occupation:

- F-stat value for 'Occupation' feature is ~1, which says that variability b/w occupation levels is less compared variability within the occupation levels.
- p-value is greater than 0.05, so we fail to reject the null hypothesis. So, at 95% confidence level, there is sufficient evidence to prove that the mean salary of students across four different Occupation levels is same

Conclusion: 'Education' is significant cause for the effect on the salaries, not 'Occupation'

Point plot:

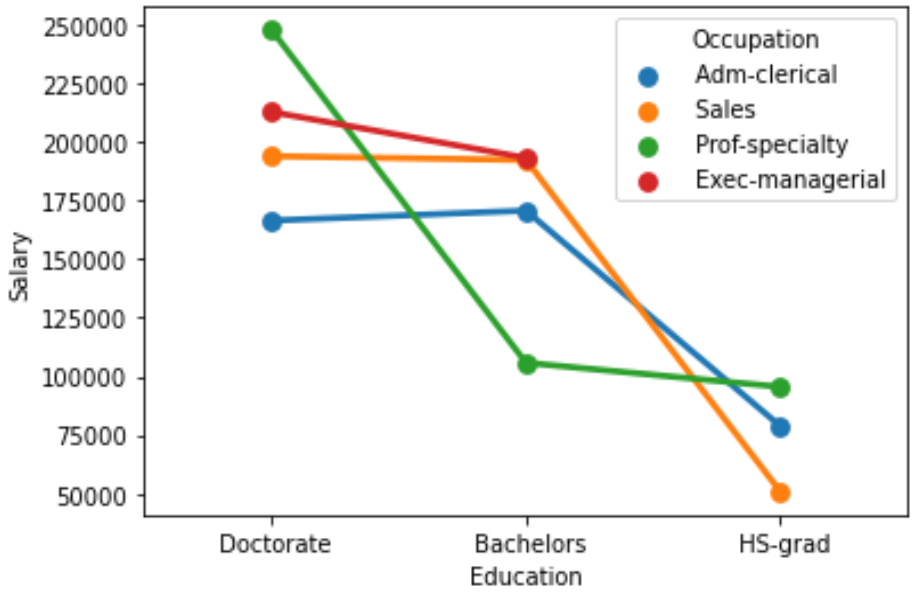

Fig.1

➢ **Insights:**

- We can see significant drop in salaries from 'Bachelors' level to 'HS-grad' for 'Sales' & 'Adm-clerical' occupation levels.
- One more significant drop observed from 'Doctorate' level to 'Bachelors' for 'Prof-specialty' occupation level.
- There are no significant changes in salaries observed from 'Doctorate' level to 'Bachelors' for 'Adm-Clerical', 'Sales' & 'Exec-managerial' occupation levels.

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?

- Two-way ANOVA to check effect of interaction b/w two features on the salary.

H_0= There is interaction b/w 'Education' and 'Occupation' features

H_1=There is no interaction b/w 'Education' and 'Occupation' features

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Table.5

- **Interpretation:**

- p-value is less than 0.05 for the interaction. We reject the null hypothesis i.e., at 95% confidence level, there is no interaction b/w 'Education' and 'Occupation' features
- F-stat value for ‘Education’ also increased after considering the interaction. That means, variability b/w education level is ~72 times than variability within the education levels when we consider the interaction b/w two features

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

- Salary dependency is more on 'Education' background level compared to 'Occupation' level. Variability in salary is high b/w 'Education' levels, on which company should concentrate more on.
- Mean salary for 'Prof specialty' is highly varying for 'Doctorate' & 'Bachelor' levels. Company should concentrate on reducing the variability in this case.
- We observed that there is not much interaction b/w 'Education' & 'Occupation' on salary. But there should be some interaction b/w them in deciding the salary of the student. Company should work on correlating the 'Education' to 'Occupation' some more.

Problem 2 (Colleges’ analysis)

Problem statement:

The dataset Education contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given.

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

Exploratory Data Analysis:

- **Data description:**

Reading the data file and loading first five records:

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

Table.6

- Total 606039 students enrolled out of 2332273 applications received by various colle ges
- Percentage of new students from top 10% of the higher secondary class is 3.53%
- Percentage of new students from top 25% of the higher secondary class is 7.15%

Dataset information:

```
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
```

Table.7

- There are 777 records and 18 features without any missing values and duplicated re cords in the dataset.
- 'Names' column is categorical and remaining all other columns are numerical

Dataset description:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Names	777	777	Abilene Christian University	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Apps	777.0	NaN	NaN	NaN	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	NaN	NaN	NaN	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	NaN	NaN	NaN	779.972973	929.17619	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	NaN	NaN	NaN	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	NaN	NaN	NaN	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	NaN	NaN	NaN	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	NaN	NaN	NaN	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	NaN	NaN	NaN	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	NaN	NaN	NaN	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	NaN	NaN	NaN	549.380952	165.10536	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	NaN	NaN	NaN	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0

Table.8

➢ **Insights:**

- Percentage values of ‘Phd’ and ‘Grad.Rate’ show 103% and 118% as maximum value s respectively. It is an anomaly because percentage value should always be in 0-100% range.
- On an average, terminal degree faculties are more than PhD faculties.
- As per S.F.Ratio, on an average there is one faculty for every ~14 students.
- On an average, full-time graduate students are more than part-time graduate students.
- Spendings on ‘Room&Board’ is more than ‘Book’ and ‘Personal’ features.
- Variability is more for instructional expenditure per student across all the colleges.

➢ **Data pre-processing:**

Checking for missing values and NA values:

- We have already seen above that there are no missing values, NA entries and duplicat ed records

Anomaly check and treatment:

- Anomalies are observed in Percentage of ‘Phd’ and ‘Grad.Rate’ features as shown bel ow.

F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend
1206	134	4860	3122	600	650	103	88	17.4	16	6415

Table.9

Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
9384	4840	600	500	22	47	14.3	20	7697	118

Table.10

- Percentages have been observed more than 100%. So, let's impute the values with ma ximum possible value i.e., 100% and check the results.

Names	Texas A&M University at Galveston
Apps	529
Accept	481
Enroll	243
Top10perc	22
Top25perc	47
F.Undergrad	1206
P.Undergrad	134
Outstate	4860
Room.Board	3122
Books	600
Personal	650
PhD	100
Terminal	88
S.F.Ratio	17.4
perc.alumni	16
Expend	6415
Grad.Rate	43

Name: 582, dtype: object

Table.11

Names	Cazenovia College
Apps	3847
Accept	3433
Enroll	527
Top10perc	9
Top25perc	35
F.Undergrad	1010
P.Undergrad	12
Outstate	9384
Room.Board	4840
Books	600
Personal	500
PhD	22
Terminal	47
S.F.Ratio	14.3
perc.alumni	20
Expend	7697
Grad.Rate	100

Name: 95, dtype: object

Table.12

- Anomalies are treated successfully

➢ **Data visualization:**

Univariate analysis:

- Let's visualize all the numeric columns using hist plot and check the distribution natur e of the features.

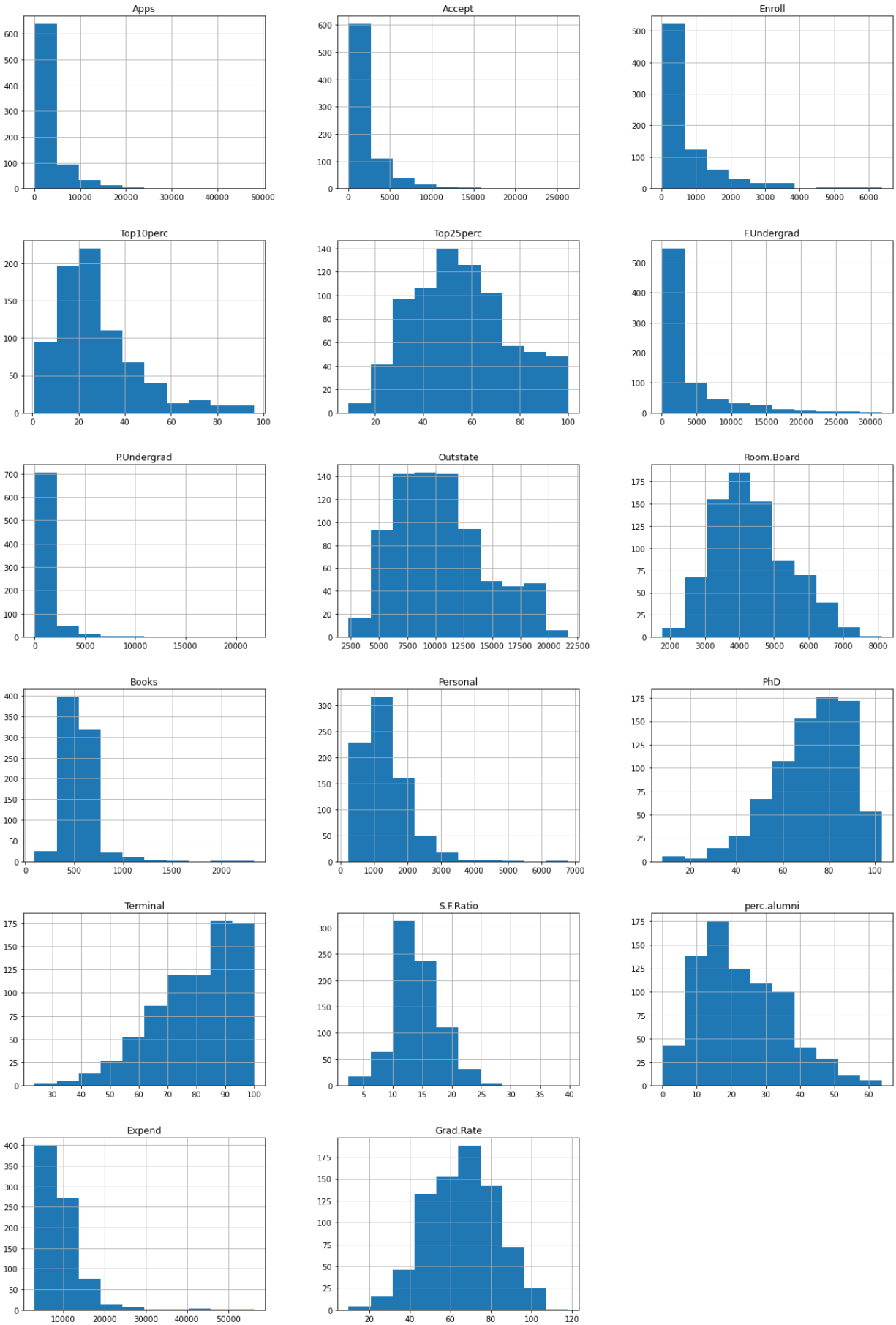
Apps
Accept
Enroll
Top10perc
Top25perc
F.Undergrad
P.Undergrad
Outstate
Room.Board
Books
Personal
PhD
Terminal
S.F.Ratio
perc.alumni
Expend
Grad.Rate

Fig.2

Checking skewness:

Apps	3.72
Accept	3.42
Enroll	2.69
Top10perc	1.41
Top25perc	0.26
F.Undergrad	2.61
P.Undergrad	5.69
Outstate	0.51
Room.Board	0.48
Books	3.49
Personal	1.74
PhD	-0.77
Terminal	-0.82
S.F.Ratio	0.67
perc.alumni	0.61
Expend	3.46
Grad.Rate	-0.14

dtype: float64

Table.13

- **Interpretations:** From both histplot and skewness data, we can state the below con clusions:

- Normally distributed features: ‘Top25perc’, ‘Outstate’, ‘Room.Board’, ‘Grad.Rate’
- Highly right skewed distributions: ‘Apps’, ‘Accept’, ‘Enroll’, ‘Top10perc’, ‘F.Underg rad’, ‘P.Undergrad’, ‘Books’, ‘Personal’, ‘Expend’
- Moderately right skewed distributions: ‘S.F.Ratio’, ‘perc.alumni’
- Moderately left skewed distribution: ‘PhD’, ‘Terminal’

Bivariate analysis:

- Let’s plot the pair plot and heatmap to check correlation b/w the data features

Pair plot:

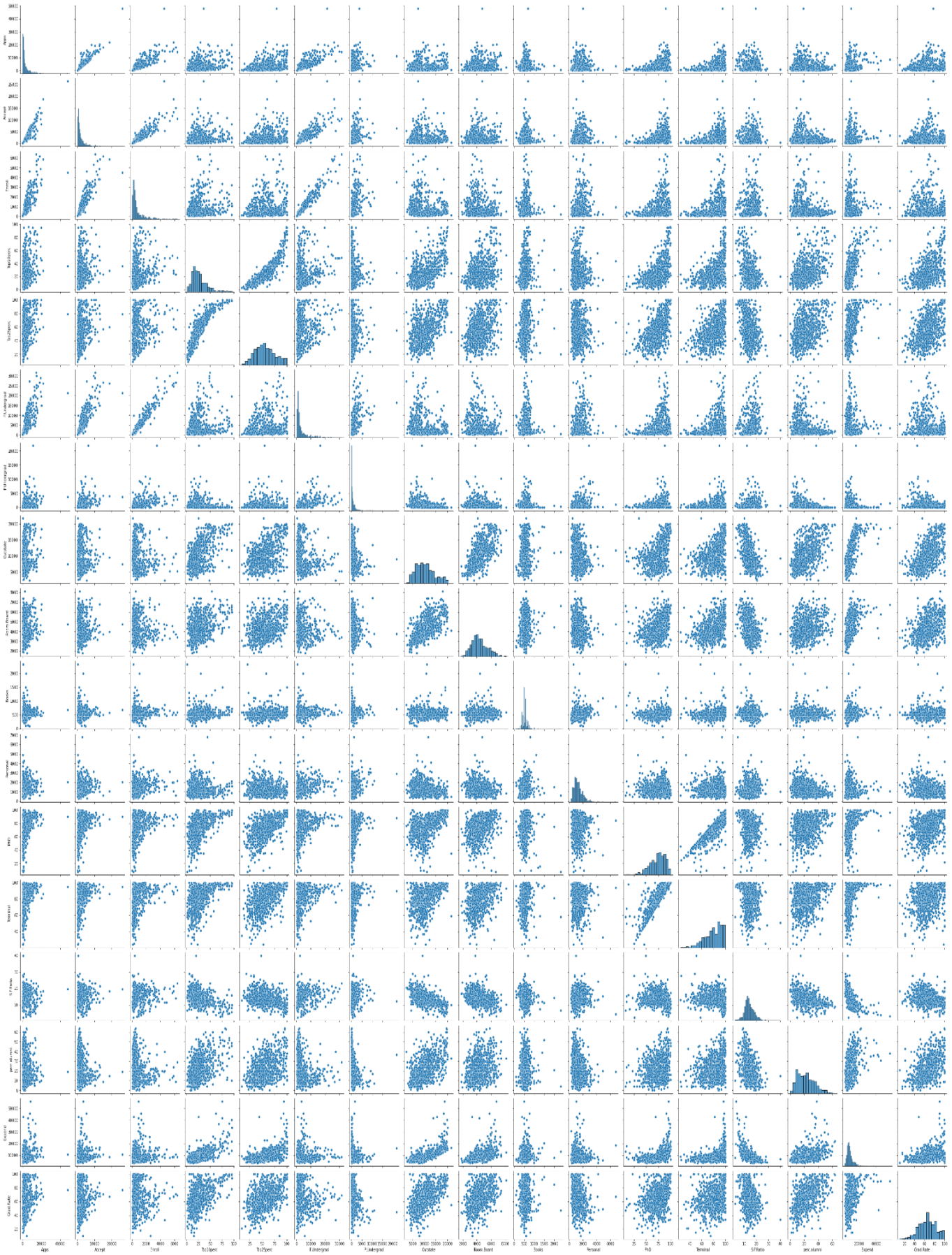

Fig.3

Heatmap:

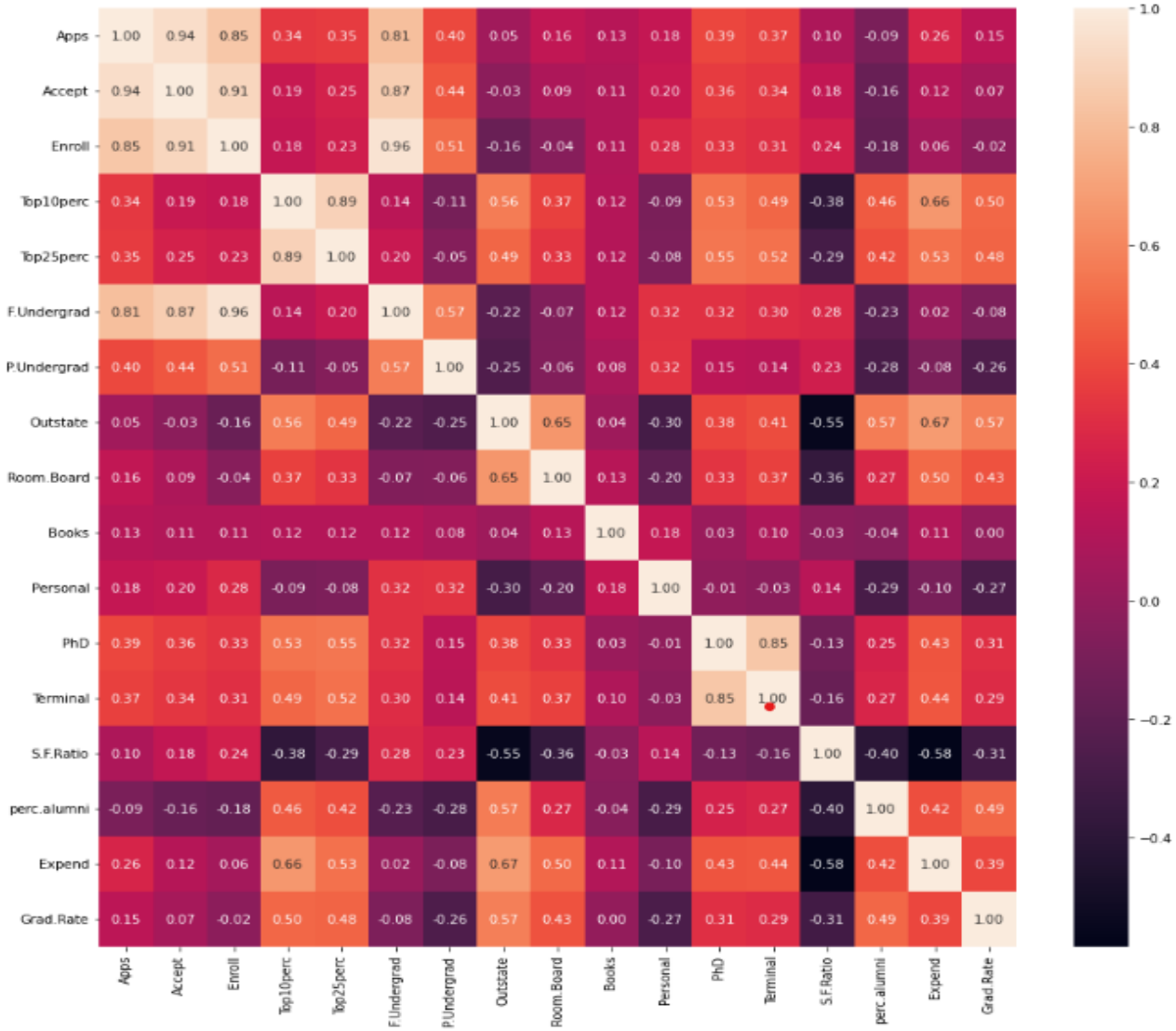

Fig.4

- **Insignts (From both pairplot and heatmap):** We can observe that there is strong cor relation between below mentioned features:

    - ‘Accept’, ‘Enroll’, ‘F.Undergrad’ vs ‘Apps’
    - ‘Apps’, ‘Enroll’, ‘F.Undergrad’ vs ‘Accept’
    - ‘Accept’, ‘Apps’, ‘F.Undergrad’ vs ‘Enroll’
    - ‘Top10perc’ vs ‘Top25perc’
    - ‘PhD’ vs ‘Terminal’

Multivariate analysis: It is not possible for the given dataset as we have only one categorical variable

- **Data preparation:**

- Scaling, Outlier detection are done below as the answers to the given questions
- Data encoding is not possible for the given dataset

Principle component analysis:

- Before performing PCA, let's drop the categorical available (Names) in the dataset.
- Sample data frame shown below after dropping 'Names' column:

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

Table.14

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

- Scaling is necessary for PCA in this case because all of the numerical features are not on same weight. For example,

1) Features like 'Top10perc', 'Top25perc', 'Grad.Rate' are in percentage scale
2) Features like 'Apps', 'Accept' are on number scale ranging from 2 digits to 5 digits

- So, we need to transform the features onto same scale.
- Scaling of the data using z-score: Sample data frame is shown below after performing the scaling

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.162859	-0.115729
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.676529	-3.378176
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1.205112	-0.931341
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.185939	1.175657
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0.204995	-0.523535

Table.15

- We can see that data has been scaled

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

First, Let's find the correlation and covariance matrices of the scaled data.

Correlation matrix:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
Apps	1.00	0.94	0.85	0.34	0.35	0.81	0.40	0.05	0.16	0.13	0.18	0.39	0.37
Accept	0.94	1.00	0.91	0.19	0.25	0.87	0.44	-0.03	0.09	0.11	0.20	0.36	0.34
Enroll	0.85	0.91	1.00	0.18	0.23	0.96	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.37	0.12	-0.09	0.53	0.49
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33	0.12	-0.08	0.55	0.52
F.Undergrad	0.81	0.87	0.96	0.14	0.20	1.00	0.57	-0.22	-0.07	0.12	0.32	0.32	0.30
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	-0.06	0.08	0.32	0.15	0.14
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.65	0.04	-0.30	0.38	0.41
Room.Board	0.16	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.65	1.00	0.13	-0.20	0.33	0.37
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.08	0.04	0.13	1.00	0.18	0.03	0.10
Personal	0.18	0.20	0.28	-0.09	-0.08	0.32	0.32	-0.30	-0.20	0.18	1.00	-0.01	-0.03

Table.16

Covariance matrix:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
Apps	1.00	0.94	0.85	0.34	0.35	0.82	0.40	0.05	0.17	0.13	0.18	0.39	0.37
Accept	0.94	1.00	0.91	0.19	0.25	0.88	0.44	-0.03	0.09	0.11	0.20	0.36	0.34
Enroll	0.85	0.91	1.00	0.18	0.23	0.97	0.51	-0.16	-0.04	0.11	0.28	0.33	0.31
Top10perc	0.34	0.19	0.18	1.00	0.89	0.14	-0.11	0.56	0.37	0.12	-0.09	0.53	0.49
Top25perc	0.35	0.25	0.23	0.89	1.00	0.20	-0.05	0.49	0.33	0.12	-0.08	0.55	0.53
F.Undergrad	0.82	0.88	0.97	0.14	0.20	1.00	0.57	-0.22	-0.07	0.12	0.32	0.32	0.30
P.Undergrad	0.40	0.44	0.51	-0.11	-0.05	0.57	1.00	-0.25	-0.06	0.08	0.32	0.15	0.14
Outstate	0.05	-0.03	-0.16	0.56	0.49	-0.22	-0.25	1.00	0.66	0.04	-0.30	0.38	0.41
Room.Board	0.17	0.09	-0.04	0.37	0.33	-0.07	-0.06	0.66	1.00	0.13	-0.20	0.33	0.38
Books	0.13	0.11	0.11	0.12	0.12	0.12	0.08	0.04	0.13	1.00	0.18	0.03	0.10
Personal	0.18	0.20	0.28	-0.09	-0.08	0.32	0.32	-0.30	-0.20	0.18	1.00	-0.01	-0.03

Table.17

Conclusion:

- After scaling (z-score) the data given, both covariance and correlation matrices are same.
- From above results, we can see correlation and correlation matrices are same.

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]

Let’s check for the outliers before scaling: (by box plot method)

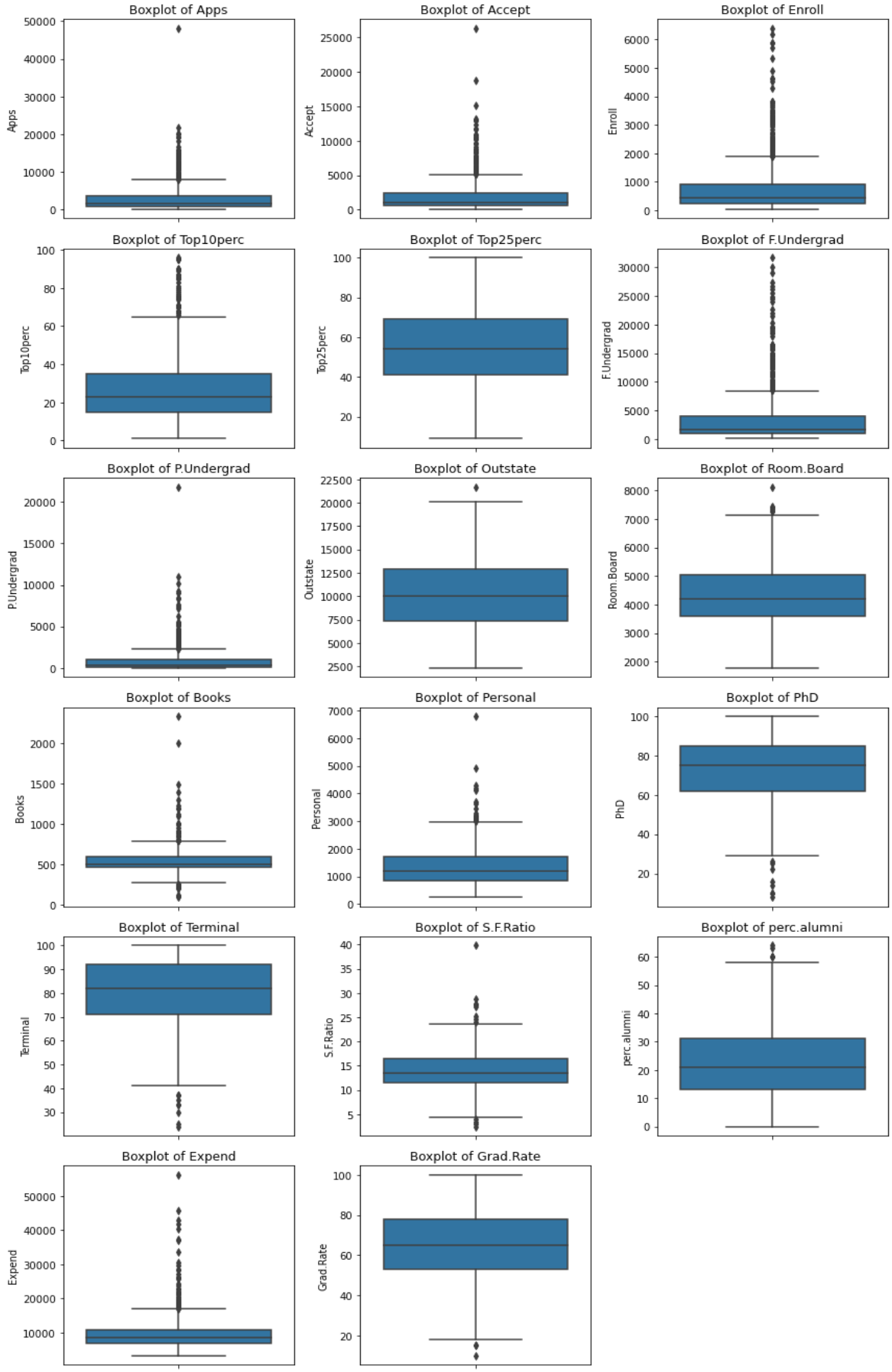
Boxplot of Apps
Apps
Boxplot of Accept
Accept
Boxplot of Enroll
Enroll
Boxplot of Top10perc
Top10perc
Boxplot of Top25perc
Top25perc
Boxplot of F.Undergrad
F.Undergrad
Boxplot of P.Undergrad
P.Undergrad
Boxplot of Outstate
Outstate
Boxplot of Room.Board
Room.Board
Boxplot of Books
Books
Boxplot of Personal
Personal
Boxplot of PhD
PhD
Boxplot of Terminal
Terminal
Boxplot of S.F.Ratio
S.F.Ratio
Boxplot of perc.alumni
perc.alumni
Boxplot of Expend
Expend
Boxplot of Grad.Rate
Grad.Rate

Fig.5

- After seeing above representation, we can say there are outliers in every feature except 'Top25perc' feature
- ‘Grad.Rate’, ‘perc.alumni’, ‘Room.Board’, ‘Outstate’ are having very minimum no. of outliers

Let’s check for the outliers after scaling: (by box plot method)

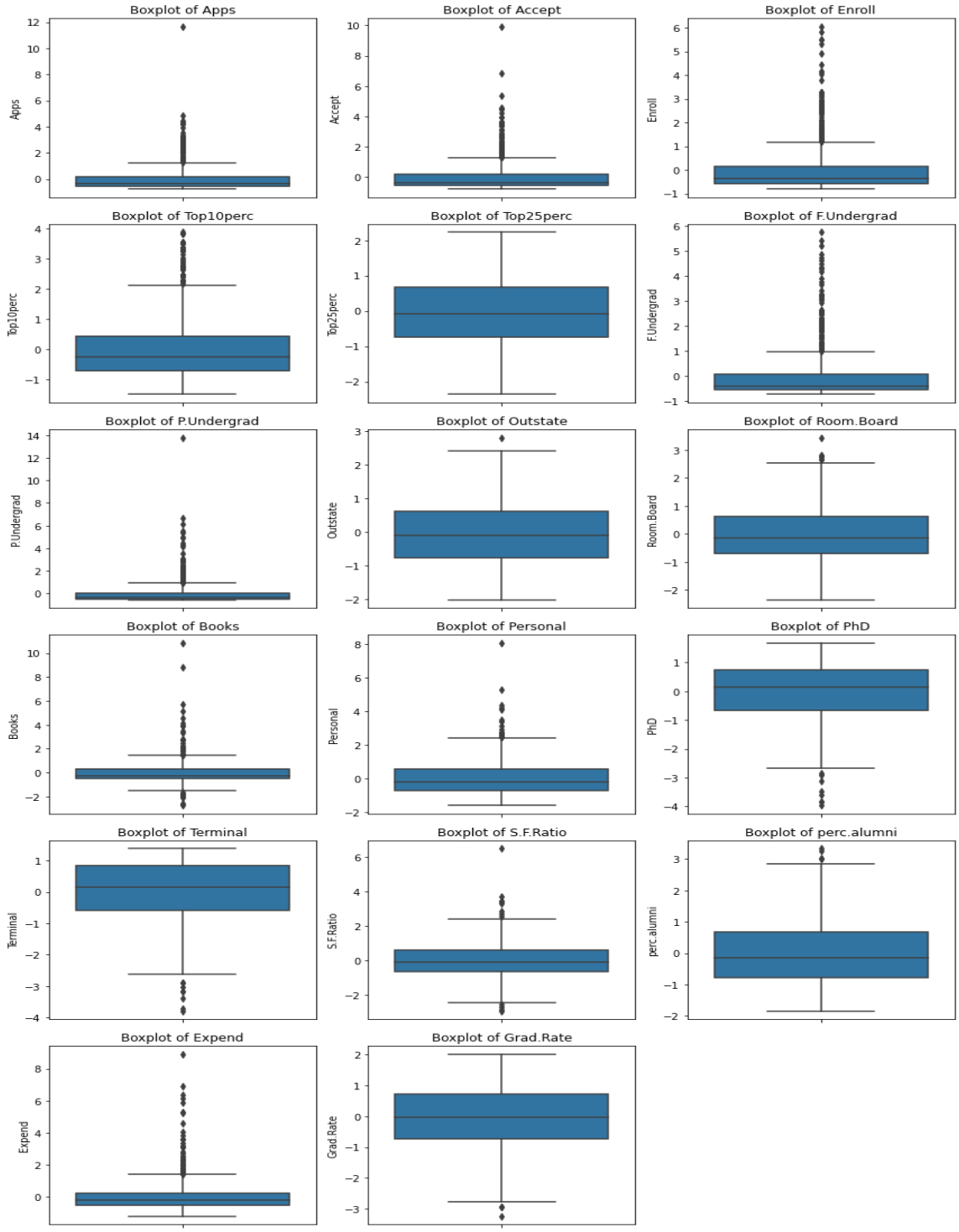

Fig.6

- There are same kind of outliers observed for scaled data as observed for unscaled data

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

Let's extract the eigenvectors for the scaled data using Sklearn:

```
[[ 2.48183494e-01  2.06969666e-01  1.75694894e-01  3.54244165e-01
   3.43942665e-01  1.54038269e-01  2.60218245e-02  2.94935587e-01
   2.49048365e-01  6.45834383e-02 -4.27212841e-02  3.18442536e-01
   3.16946970e-01 -1.77143862e-01  2.05340243e-01  3.18874326e-01
   2.53800631e-01]
 [ 3.32025554e-01  3.72491384e-01  4.04001896e-01 -8.18674207e-02
  -4.42586593e-02  4.17901438e-01  3.15112160e-01 -2.49149211e-01
  -1.37349991e-01  5.64823263e-02  2.19795159e-01  5.86942062e-02
   4.68443038e-02  2.46336958e-01 -2.46268610e-01 -1.31140311e-01
  -1.69072352e-01]
 [-6.16148087e-02 -9.97638535e-02 -8.22289219e-02  3.47636016e-02
  -2.46996496e-02 -6.09712128e-02  1.39028188e-01  4.73522670e-02
   1.50278691e-01  6.78096050e-01  4.98147475e-01 -1.29562447e-01
  -6.83497434e-02 -2.90605689e-01 -1.46940854e-01  2.27294587e-01
  -2.06564710e-01]
```

Fig.7

Note: First three PCs are shown above

Let's extract the eigenvectors for the scaled data using Sklearn:

```
[5.45485033 4.48406663 1.17470127 1.00536006 0.9343371  0.84817556
 0.60551358 0.58787041 0.53053165 0.40349818 0.31326156 0.22048561
 0.16780564 0.14368317 0.08802439 0.03672101 0.02302105]
```

Fig.8

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

PCA has been performed before extracting eigen vectors and eigen values. Now let's export the data of PCs to a data frame with original features.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Apps	0.25	0.33	-0.06	0.28	0.00	-0.01	-0.04	-0.10	-0.09	0.05	0.04	0.02	0.60	0.08	0.13	0.46	0.36
Accept	0.21	0.37	-0.10	0.27	0.05	0.01	-0.01	-0.06	-0.18	0.04	-0.06	-0.15	0.29	0.03	-0.15	-0.52	-0.54
Enroll	0.18	0.40	-0.08	0.16	-0.06	-0.04	-0.03	0.06	-0.13	0.03	-0.07	0.01	-0.45	-0.08	0.03	-0.40	0.61
Top10perc	0.35	-0.08	0.03	-0.05	-0.39	-0.05	-0.16	-0.12	0.34	0.06	-0.01	0.04	0.00	-0.11	0.70	-0.15	-0.14
Top25perc	0.34	-0.04	-0.02	-0.11	-0.43	0.03	-0.12	-0.10	0.40	0.01	-0.27	-0.09	0.02	0.15	-0.62	0.05	0.08
F.Undergrad	0.15	0.42	-0.06	0.10	-0.04	-0.04	-0.02	0.08	-0.06	0.02	-0.08	0.06	-0.52	-0.05	0.01	0.56	-0.41
P.Undergrad	0.03	0.32	0.14	-0.16	0.30	-0.19	0.05	0.57	0.56	-0.22	0.10	-0.06	0.13	0.02	0.02	-0.05	0.01
Outstate	0.29	-0.25	0.05	0.13	0.22	-0.03	0.11	0.01	-0.00	0.18	0.14	-0.82	-0.14	-0.03	0.04	0.10	0.05
Room.Board	0.25	-0.14	0.15	0.19	0.56	0.16	0.21	-0.22	0.28	0.30	-0.36	0.35	-0.07	-0.06	0.00	-0.03	0.00
Books	0.06	0.06	0.68	0.08	-0.13	0.64	-0.15	0.21	-0.13	-0.08	0.03	-0.03	0.01	-0.07	-0.01	0.00	0.00
Personal	-0.04	0.22	0.50	-0.24	-0.22	-0.33	0.63	-0.23	-0.09	0.14	-0.02	-0.04	0.04	0.03	-0.00	-0.01	-0.00
PhD	0.32	0.06	-0.13	-0.53	0.14	0.09	-0.00	-0.08	-0.19	-0.12	0.04	0.02	0.12	-0.69	-0.11	0.03	0.01
Terminal	0.32	0.05	-0.07	-0.52	0.21	0.15	-0.03	-0.01	-0.26	-0.08	-0.06	0.02	-0.05	0.67	0.16	-0.03	0.01
S.F.Ratio	-0.18	0.25	-0.29	-0.16	-0.08	0.49	0.22	-0.08	0.28	0.47	0.44	-0.01	-0.02	0.04	-0.02	-0.02	-0.00
perc.alumni	0.21	-0.25	-0.15	0.02	-0.22	-0.05	0.24	0.68	-0.25	0.42	-0.13	0.18	0.10	-0.03	-0.01	0.00	-0.02
Expend	0.32	-0.13	0.23	0.08	0.08	-0.30	-0.23	-0.06	-0.05	0.14	0.69	0.33	-0.09	0.07	-0.23	-0.04	-0.04
Grad.Rate	0.25	-0.17	-0.21	0.26	-0.11	0.22	0.56	-0.00	0.04	-0.59	0.22	0.12	-0.07	0.04	-0.00	-0.01	-0.01

Table.18

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]

Explicit form of the first PC:

PC1= 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

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?

Cumulative values of eigenvalues:

```
[0.32046058 0.58388974 0.65290088 0.7119636  0.76685387 0.81668234
 0.85225494 0.88679105 0.91795863 0.94166327 0.9600667  0.97301975
 0.98287797 0.99131904 0.99649028 0.99864756 1.        ]
```

Fig.9

- **Interpretation:**
  - Cumulative values of the eigenvalues show the percentage of data covered after each PC. It also helps us to reduce the dimensions by deciding how many PC's have to be taken.
  - Generally, it reduces the dimensions by half after selecting first 5 PC's which cover 70-80% of the data.

- For the given case, selecting 5, 6 PC's will cover the 76%, 81% of the data respectively.
- Eigenvector is a line that is drawn in mathematical space which shows the direction of maximum variance in the dataset. These are new linear combinations of the initial variables.

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]

➢ Interpretations:

- For this case study, we can take first 6 PCs which covers ~81% of the data.

	PC1	PC2	PC3	PC4	PC5	PC6
Apps	0.248183	0.332026	-0.061615	0.282569	0.004158	-0.014238
Accept	0.206970	0.372491	-0.099764	0.269149	0.054318	0.009380
Enroll	0.175695	0.404002	-0.082229	0.162611	-0.056603	-0.041347
Top10perc	0.354244	-0.081867	0.034764	-0.052469	-0.394960	-0.053064
Top25perc	0.343943	-0.044259	-0.024700	-0.111389	-0.425700	0.032288
F.Undergrad	0.154038	0.417901	-0.060971	0.100948	-0.044035	-0.042658
P.Undergrad	0.026022	0.315112	0.139028	-0.158473	0.303348	-0.192744
Outstate	0.294936	-0.249149	0.047352	0.133122	0.221797	-0.028860
Room.Board	0.249048	-0.137350	0.150279	0.186480	0.559744	0.164168
Books	0.064583	0.056482	0.678096	0.079488	-0.128343	0.641364
Personal	-0.042721	0.219795	0.498147	-0.236126	-0.221267	-0.333976
PhD	0.318443	0.058694	-0.129562	-0.533695	0.143265	0.087686
Terminal	0.316947	0.046844	-0.068350	-0.519385	0.207796	0.151537
S.F.Ratio	-0.177144	0.246337	-0.290606	-0.163696	-0.078557	0.486124
perc.alumni	0.205340	-0.246269	-0.146941	0.017638	-0.216206	-0.047228
Expend	0.318874	-0.131140	0.227295	0.081376	0.075677	-0.297658
Grad.Rate	0.253801	-0.169072	-0.206565	0.260607	-0.111222	0.216104

Table.19

- Generally, each PC influences (or variates) the given factors in different levels.
- Now, let's check how the selected PCs influence the factors given.

1) PC1 influences most of the factors positively except 'P.Undergrad', 'Books', 'Personal', 'S.F.Ratio'

2) PC2 influences Apps, Accept, Enroll, F.Undergrad, P.Undergrad, Personal, S.F.Ratio

3) PC3 influences P.Undergrad, Room.Board, S.F.Ratio, Books, Personal,Expend

4) PC4 influences Apps, Accept, Enroll, F.Undergrad, Room.Board, Grad.Rate

5) PC5 influences P.Graduate, Outstae, Room.Board, PhD, Terminal

6) PC6 influences Room.Board, Books, Terminal, S.F.Ratio

- By above influencial analysis of factors by PCs, we can study the different factors w.r.t to the data which is dimensionally reduced to 6 PCs. This will ease the colleges to study and understand about the admissions, expenditure, faculty, graduation rate etc.

THE END