AS PROJECT BUSSINESS REPORT

Kratik Mehta

PGP-DSBA Feb 2022

Table of Contents

List of Figures

List of Tables

Problem 1

Executive Summary

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals is 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 individuals or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

Sample of the Salary 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

Table 1: Salary Dataset Sample

Checking the types of variables in the dataset.

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

From the above output we can see that:

- There are 40 observations of different individuals in the data.
- There are 3 variables, out of which, 1 is of integer type and 2 are of object(categorical) type.
- **Salary** is a *continuous numerical variable*.
- The dataset does not have any missing values.

Descriptive Statistics of the dataset

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Education	40	3	Doctorate	16	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Occupation	40	4	Prof-specialty	13	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	40.0	NaN	NaN	NaN	162186.875	64860.407506	50103.0	99897.5	169100.0	214440.75	260151.0

Table 2: Description of the dataset

Checking the counts of individuals in various levels of Education and Occupation

```
 Doctorate    16
 Bachelors    15
 HS-grad       9
Name: Education, dtype: int64

 Prof-specialty     13
 Sales              12
 Adm-clerical       10
```

```
 Exec-managerial     5
Name: Occupation, dtype: int64
```

Distribution of the Salary variable.

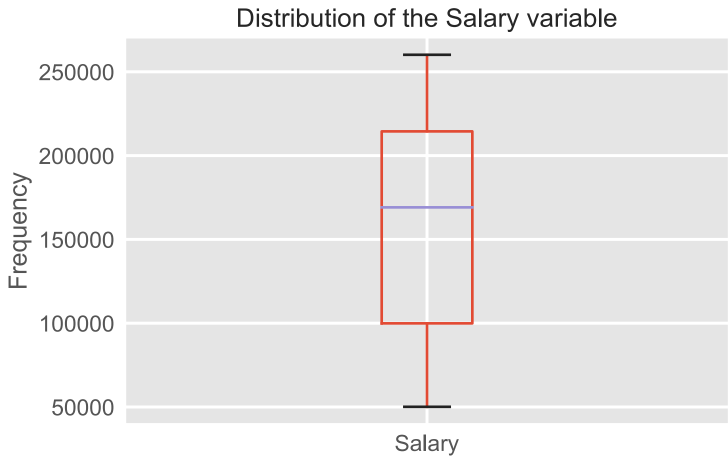

Figure 1: Distribution of the Salary variable.

Assumptions for ANOVA

1. All populations under consideration have normal distribution.
2. All populations under consideration have equal variances.
3. The sample is a random sample, i.e., the observations are collected independently of each other.

1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually. Descriptive Statistics of the Wholesale Customers Dataset.

The Null and Alternate Hypothesis for **Education** for conducting one-way ANOVA are:

$$\mathrm{H_0}: \mu_{\mathrm{Doctorate}} = \mu_{\mathrm{Bachelors}} = \mu_{\mathrm{HS\_grad}}$$

$$H_A: Atleast\ one\ Education\ level\ is\ different\ from\ the\ rest.$$

The Null and Alternate Hypothesis for **Occupation** for conducting one-way ANOVA are:

$$H_0: \mu_{Prof-specialty} = \mu_{Sales} = \mu_{Adm-clerical} = \mu_{Exec-managerial}$$

$$H_A: Atleast\ one\ Occupation\ level\ is\ different\ from\ the\ rest.$$

The means in the above equations are **Salary** means for respective levels.

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

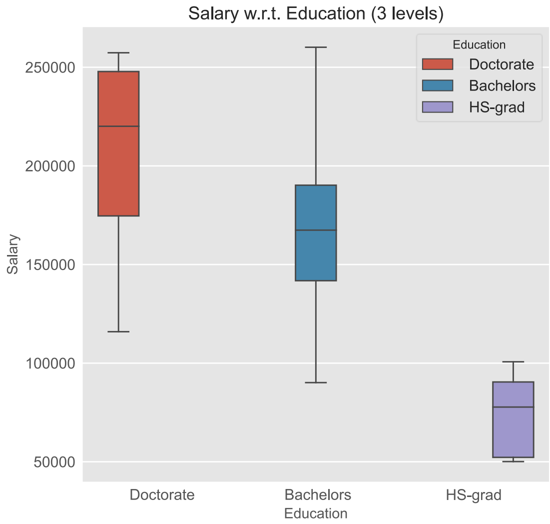

Figure 2: Salary w.r.t. Education (3 levels)

	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 3: One-way ANOVA of Education w.r.t. Salary

From the above table we can see that the ***p value*** for the one-way ANOVA test for **Education** is almost equal to ***zero***. Assuming a ***significance level of 0.05*** we can ***reject the null hypothesis***, which means that, we have enough evidence that at least one Education level is different from the rest in terms of Salary.

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

Figure 3: Salary w.r.t. Occupation (4 levels)

	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 4: One-way ANOVA of Occupation w.r.t. Salary

From the above table we can see that the ***p value*** for the one-way ANOVA test for **Occupation** is equal to ***0.45***. Assuming a ***significance level of 0.05*** we ***cannot reject the null hypothesis***, which means that, we have enough evidence that all Occupation levels are the same in terms of Salary.

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

From the one-way ANOVA test of the **Education w.r.t. Salary**, we have enough evidence that atleast one level in **Education** is different from the rest. From the ***Boxplot in Figure 2*** we can see that the means of all levels in **Education** are significantly different from each other.

Doctorate level has the ***highest median*** **Salary** of around ***225000***, while **HS grad** level has the ***lowest median*** **Salary** of around ***80000***. The **Salary** for **Bachelors** is ***widely spread out***.

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

Interaction Plots helps us to analyze the effects of one variable on the other variable. If the lines in the plot are approximately parallel to each other we can say that the two variables don't have any interaction.

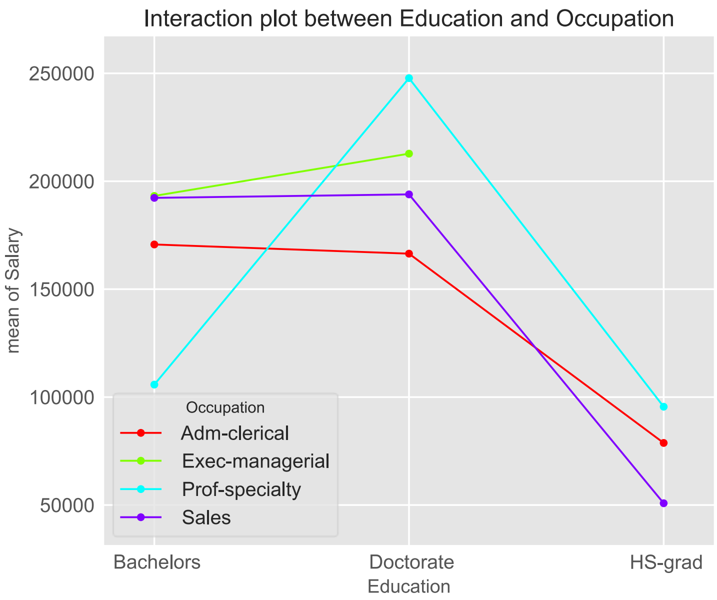

Figure 4: Interaction plot between Education and Occupation

The lines in the above interaction plot are not parallel, hence we can say that there is some interaction between **Education** and **Occupation**. In particular:

1. For **Prof-speciality** level in **Occupation** the **Salary** increases significantly from **Bachelors** to **Doctorate** degree, while it remains the same for other Occupation levels for **Bachelors** and **Doctorate** degree.
2. For **Sales** level in **Occupation** the **Salary** decreases more steeply from **Doctorate** to **HS-grad** degree compared to other levels.

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

Null and Alternate Hypothesis for two-way ANOVA.

For factor **Education**:

$$\mathrm{H}_0: \mu_{\mathrm{Doctorate}} = \mu_{\mathrm{Bachelors}} = \mu_{\mathrm{HS\_grad}}$$

$$H_A: Atleast\ one\ Education\ level\ is\ different\ from\ the\ rest.$$

For factor **Occupation**:

$$H_0: \mu_{Prof-specialty} = \mu_{Sales} = \mu_{Adm-clerical} = \mu_{Exec-managerial}$$

$$H_A: Atleast\ one\ Occupation\ level\ is\ different\ from\ the\ rest.$$

For **Interaction** between **Education** and **Occupation**:

$$H_0: The\ interaction\ effect\ does\ not\ exist.$$

$$H_A: An\ interaction\ effect\ exist.$$

Results

	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: Result table of two-way ANOVA

```
Total sum of squares: 165185524257.17
Variance explained by factor Education: 62.2%
Variance explained by factor Occupation: 3.3%
Variance explained by interaction effect: 22.0%
```

Interpretations

1. The ***p value*** for the factor **Education** is almost ***equal to zero***. Therefore, for a ***significance level of 0.05*** we ***reject the Null hypothesis*** for **Education**. Which means that the levels in Education are significantly different from each other.
2. The ***p value*** for the factor **Occupation** is 0.072. Therefore, for a ***significance level of 0.05*** we ***cannot reject the Null hypothesis*** for **Occupation**. Which means that the levels in Occupation are not significantly different from each other.
3. The ***p value*** for the ***interaction factor*** between Education and Occupation *is* ***less than the significance level of 0.05***. Therefore, we ***reject the Null hypothesis*** for **Interaction**. Hence, there is an interaction between Education and Occupation. This can also be confirmed from the interaction plot in Figure 4.
4. The ***p value*** for **Education** and **Occupation** has ***decreased significantly from the one-way ANOVA test*** above. Which means that these two factors taken together explain more variance than taken individually.

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

For this particular case study:

1. **Education** is a ***very good predictor*** of **Salary** of an individual. Around ***62.2% of the total variance*** in **Salary** is explained by **Education**.
2. **Occupation** is ***not a very good predictor*** of **Salary** of an individual. Only ***3.3% of the total variance*** in **Salary** is explained by **Occupation**. Therefore, it is not useful to use **Occupation** alone as a predictor of Salary.
3. There ***exists an interaction*** between **Education** and **Occupation**. This interaction factor helps explain ***22% of the total variance*** in **Salary**. Though **Occupation** is not a good predictor for **Salary**, it's interaction with **Education** becomes a good predictor.

Problem 2

Executive Summary

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 Dictionary

1. Names: Names of various university and colleges
2. Apps: Number of applications received
3. Accept: Number of applications accepted
4. Enroll: Number of new students enrolled
5. Top10perc: Percentage of new students from top 10% of Higher Secondary class
6. Top25perc: Percentage of new students from top 25% of Higher Secondary class
7. F.Undergrad: Number of full-time undergraduate students
8. P.Undergrad: Number of part-time undergraduate students
9. Outstate: Number of students for whom the particular college or university is Out-of-state tuition
10. Room.Board: Cost of Room and board
11. Books: Estimated book costs for a student
12. Personal: Estimated personal spending for a student
13. PhD: Percentage of faculties with Ph.D.'s
14. Terminal: Percentage of faculties with terminal degree
15. S.F.Ratio: Student/faculty ratio
16. perc.alumni: Percentage of alumni who donate
17. Expend: The Instructional expenditure per student
18. Grad.Rate: Graduation rate

Sample of the Education Dataset

	Names	Apps	Accept	Enroll	Top10 perc	Top25 perc	F.Under grad	P.Under grad	Outstate	Room. Board	Books	Personal	PhD	Terminal	S.F. Ratio	perc. alumni	Expend	Grad. Rate
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15

Table 6: Sample of the Education Dataset.

Checking the types of variables in the dataset.

```
<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
```

From the above output we can see that:

- There are ***777 observations*** of different colleges in the data.
- There are ***18 variables***, out of which, ***1 is of object type*** and ***17 are of integer/float type***.
- The numerical variables have proper integer/float type of data.
- The dataset does not have any missing values.

Descriptive Statistics of the Education Dataset

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

Table 7: Descriptive Statistics of the Education Dataset

There seems to be no bad data in the dataset. The **scale of numerical variables is different** from each other; hence we will have to **standardize the data** before performing PCA.

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

Univariate Analysis of Apps variable

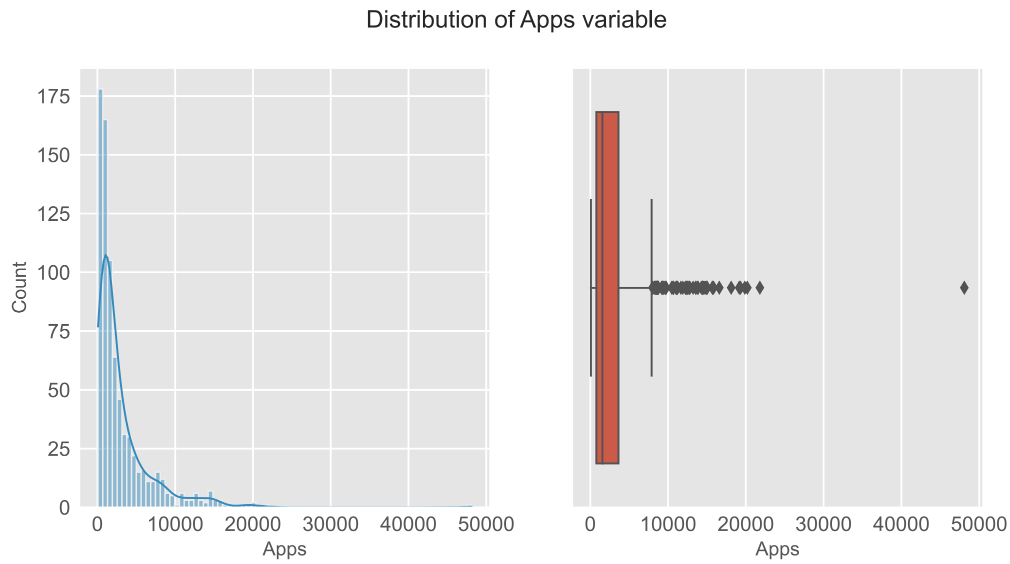

Figure 5: Distribution of Apps variable.

The data in the **Apps** variable is ***highly skewed to the right***. The **Apps** variable also ***contains outliers***.

Univariate Analysis of Accept variable

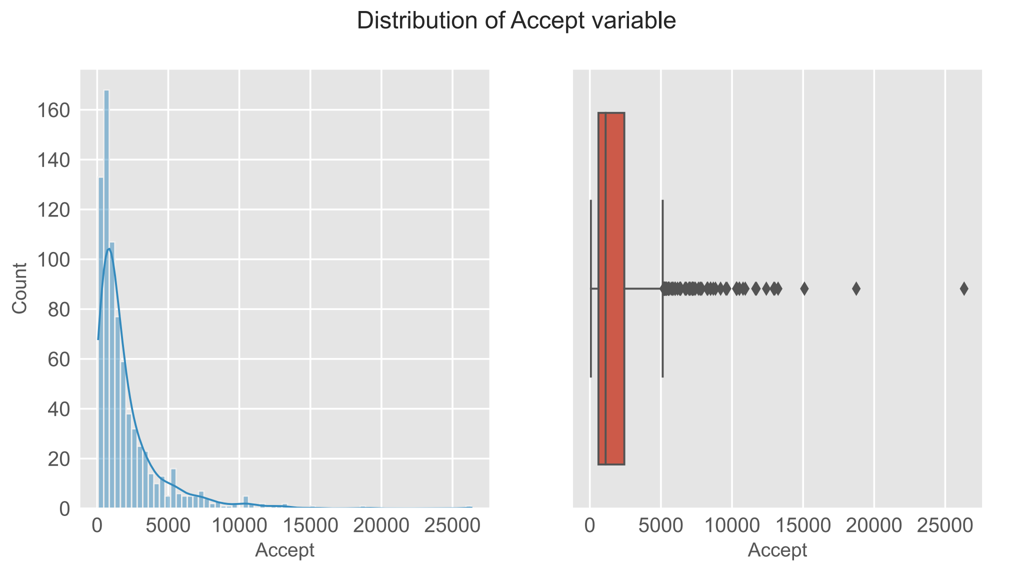

Figure 6: Distribution of Accept variable.

The data in the **Accept** variable is ***highly skewed to the right***. The **Accept** variable also ***contains outliers***.

Univariate Analysis of Enroll variable

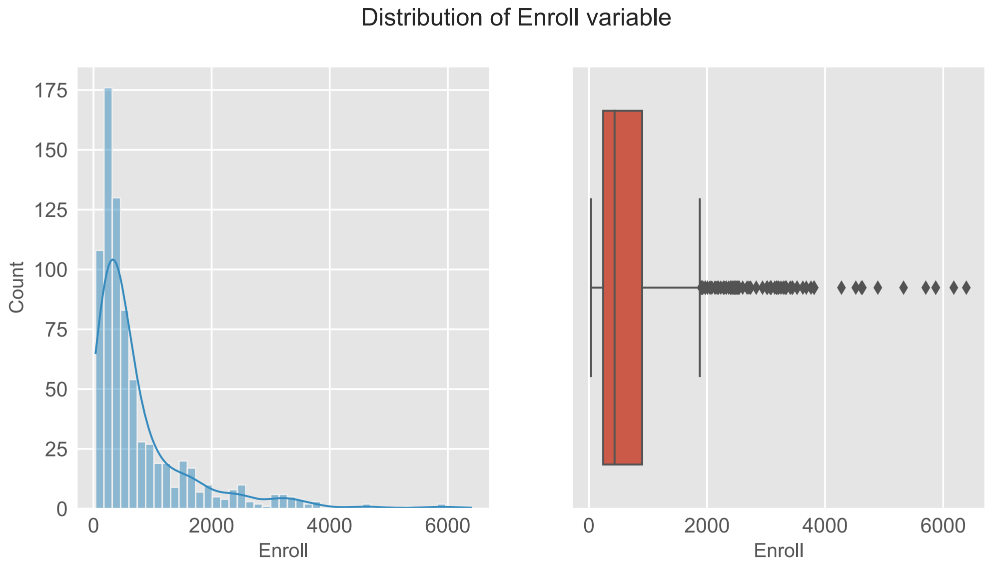

Figure 7: Distribution of Enroll variable.

The data in the **Enroll** variable is ***highly skewed to the right***. The **Enroll** variable also ***contains outliers***.

Univariate Analysis of Top10perc variable

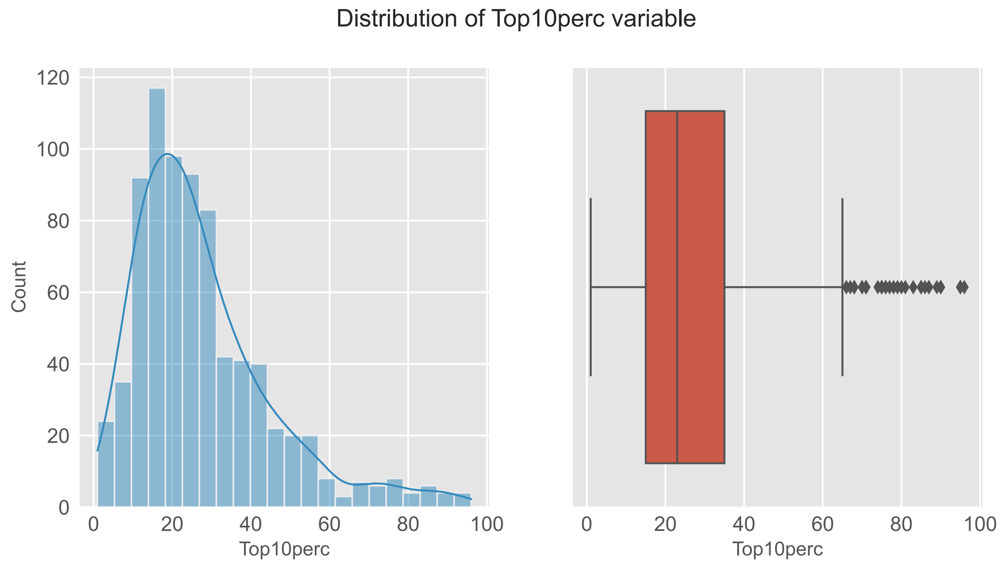

Figure 8: Distribution of Top10perc variable.

The data in the **Top10perc** variable is ***slightly skewed to the right***. The **Top10perc** variable also ***contains some outliers***.

Univariate Analysis of Top25perc variable

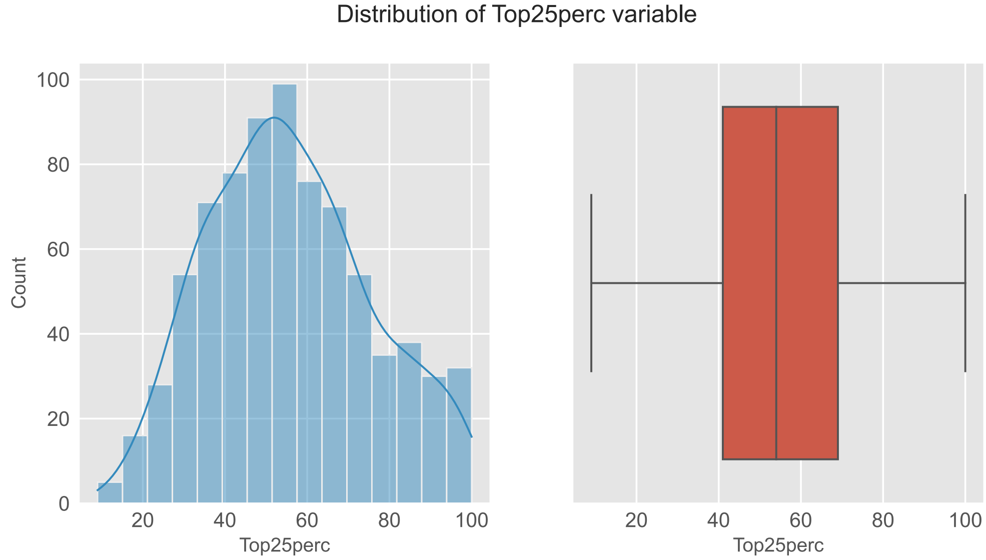

Figure 9: Distribution of Top25perc variable.

The data in the **Top25perc** variable is ***approximately normally distributed***. The variable ***does not contain any outliers***.

Univariate Analysis of F.Undergrad variable

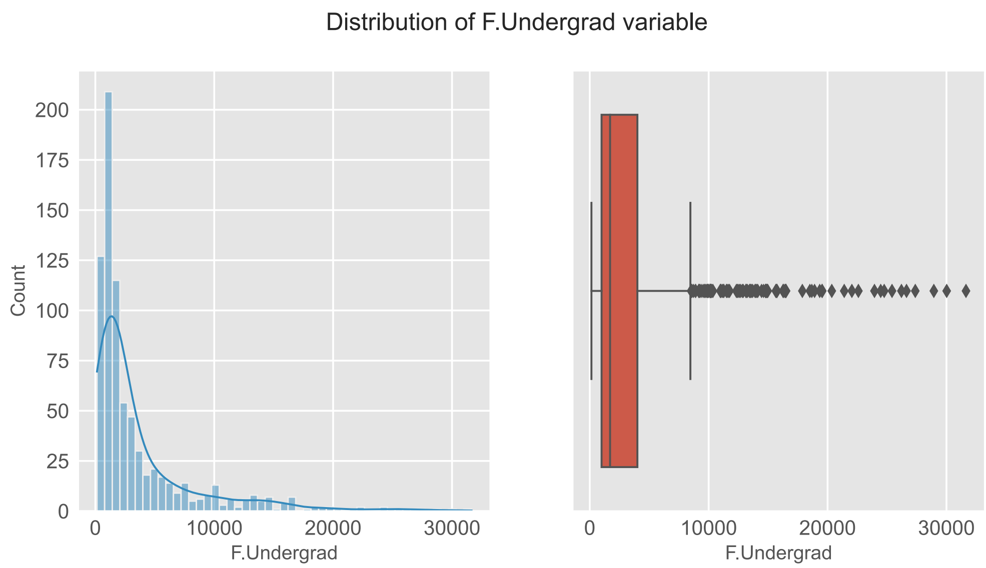

Figure 10: Distribution of F.Undergrad variable.

The data in the **F.Undergrad** variable is ***highly skewed to the right***. The variable also ***contains outliers***.

Univariate Analysis of P.Undergrad variable

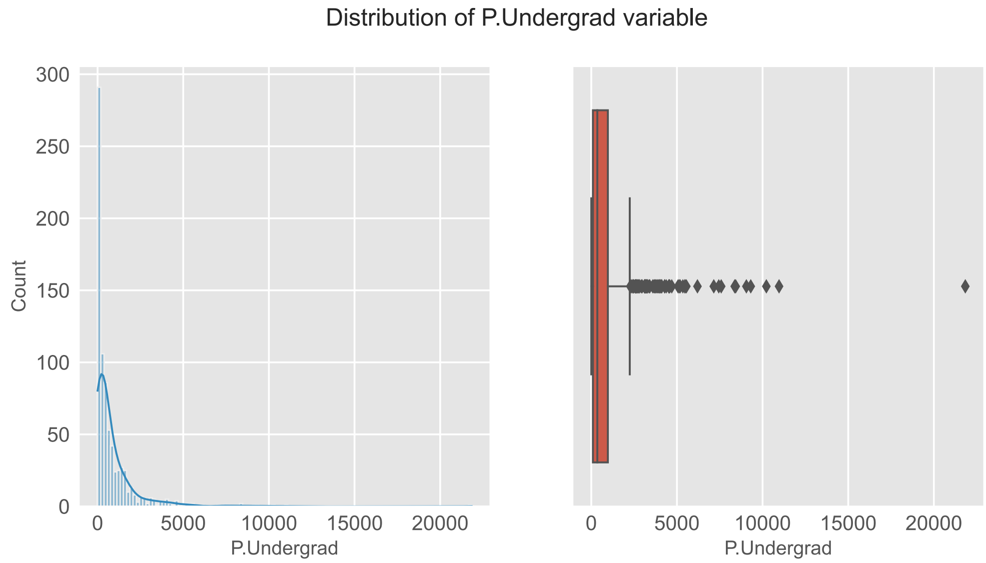

Figure 11: Distribution of P.Undergrad variable.

The data in the **P.Undergrad** variable is ***highly skewed to the right***. The variable also ***contains outliers***.

Univariate Analysis of Outstate variable

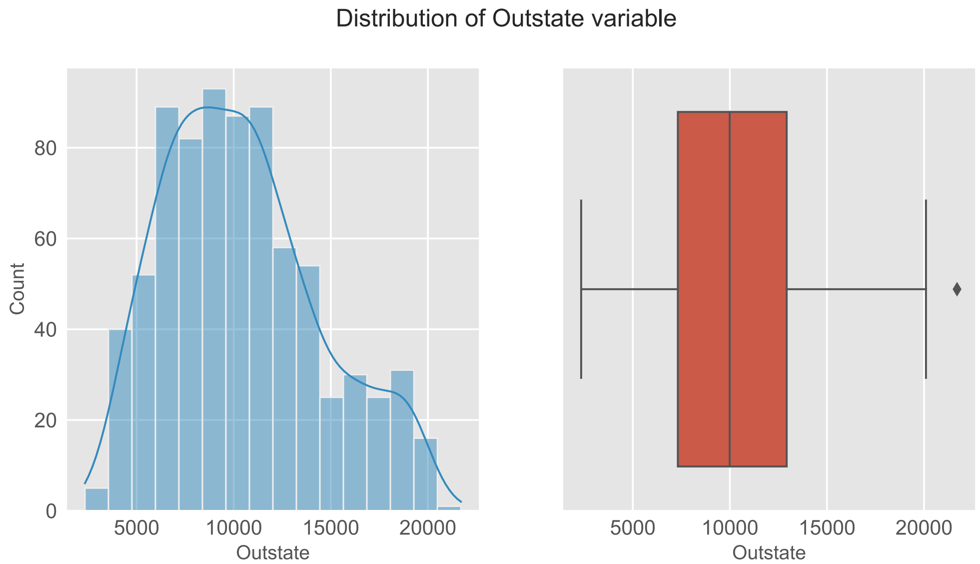

Figure 12: Distribution of Outstate variable.

The data in the **Outstate** variable is ***approximately normally distributed***. The variable ***contains one outlier***.

Univariate Analysis of Room.Board variable

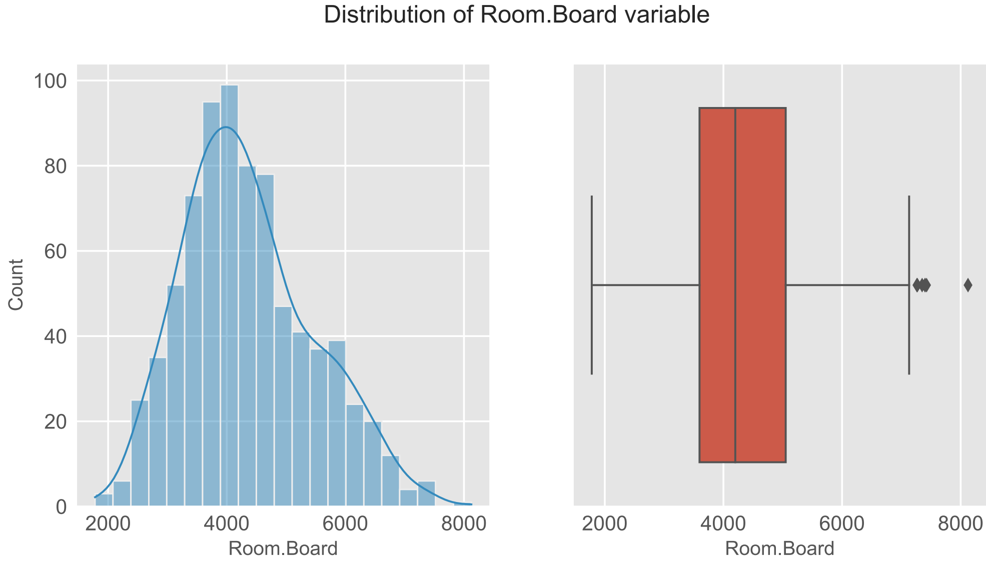

Figure 13: Distribution of Room.Board variable.

The data in the **Room.Board** variable is ***approximately normally distributed***. The variable ***contains some outliers***.

Univariate Analysis of Books variable

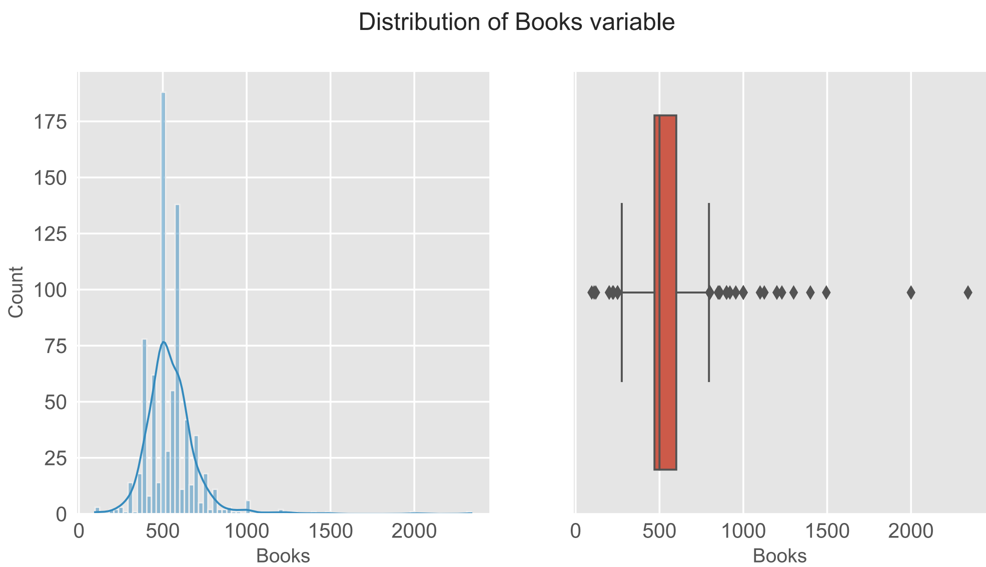

Figure 14: Distribution of Books variable.

The data in the **Books** variable is ***slightly skewed to the right***. The variable also ***contains outliers on both sides***.

Univariate Analysis of Personal variable

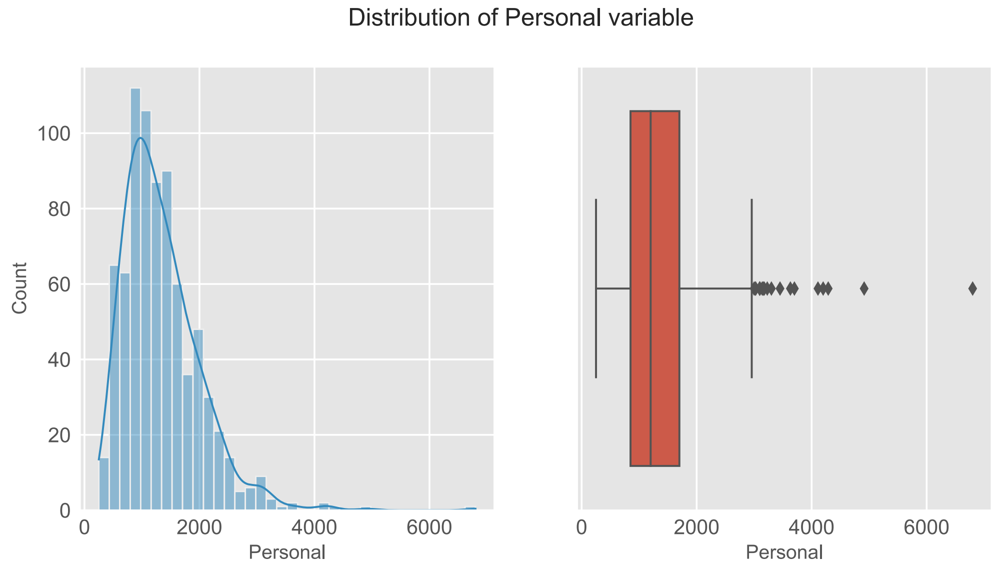

Figure 15: Distribution of Personal variable.

The data in the **Personal** variable is ***slightly skewed to the right***. The variable also ***contains outliers***.

Univariate Analysis of PhD variable

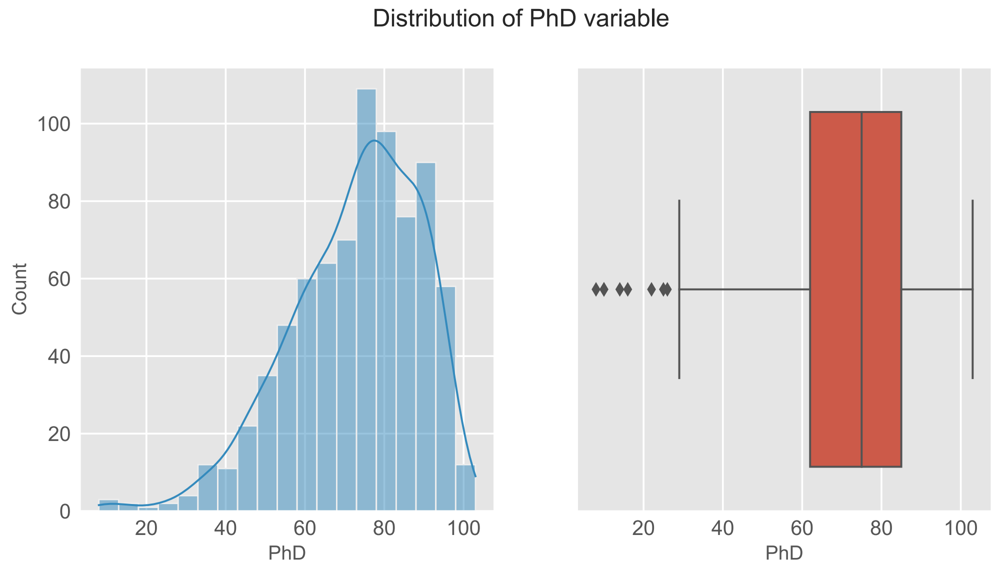

Figure 16: Distribution of PhD variable.

The data in the **PhD** variable is ***slightly skewed to the left***. The variable also ***contains outliers***.

Univariate Analysis of Terminal variable

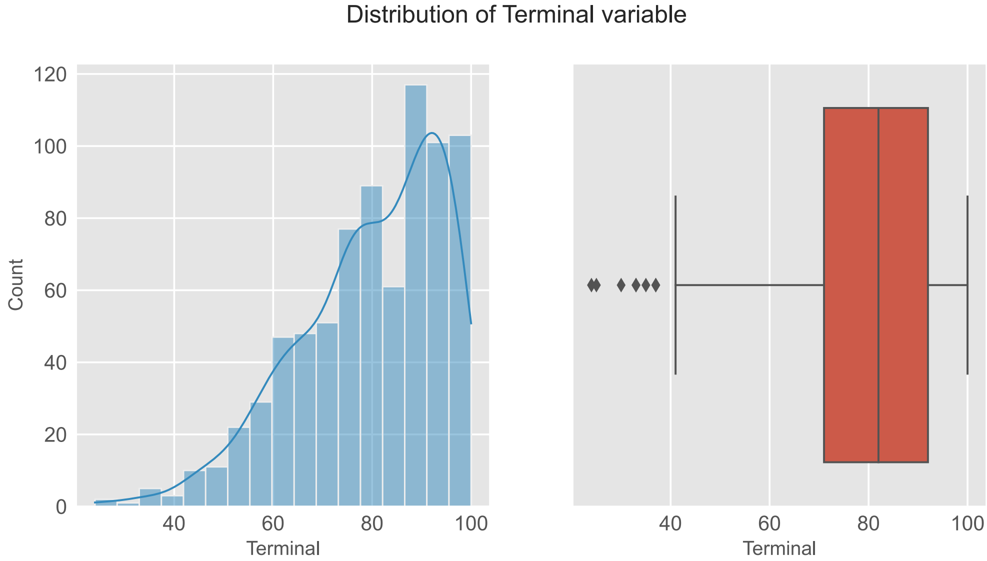

Figure 17: Distribution of Terminal variable.

The data in the **Terminal** variable is ***slightly skewed to the left***. The variable also ***contains outliers***.

Univariate Analysis of S.F.Ratio variable

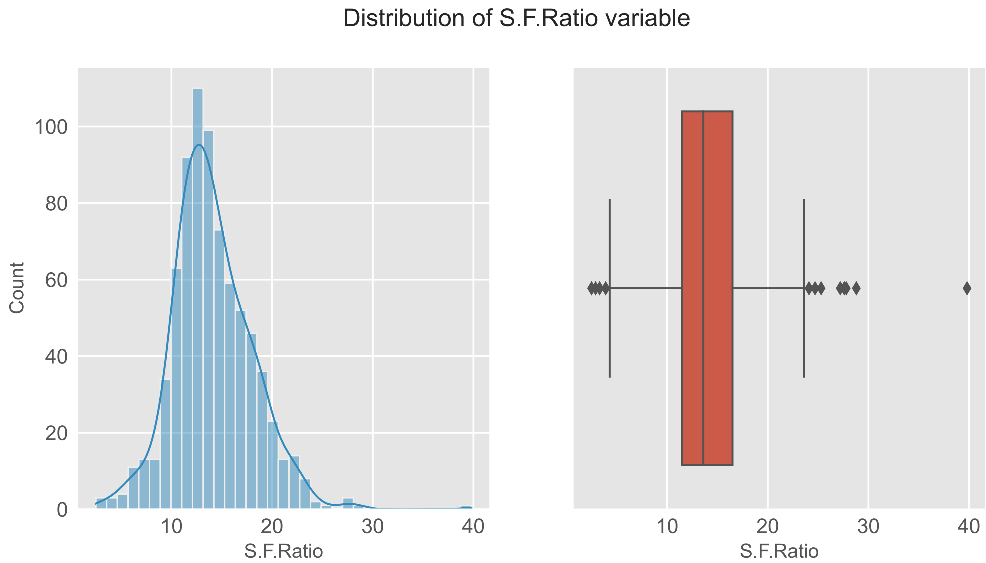

Figure 18: Distribution of S.F.Ratio variable.

The data in the **S.F.Ratio** variable is ***approximately normally distributed***. The variable ***contains some outliers***.

Univariate Analysis of perc.alumni variable

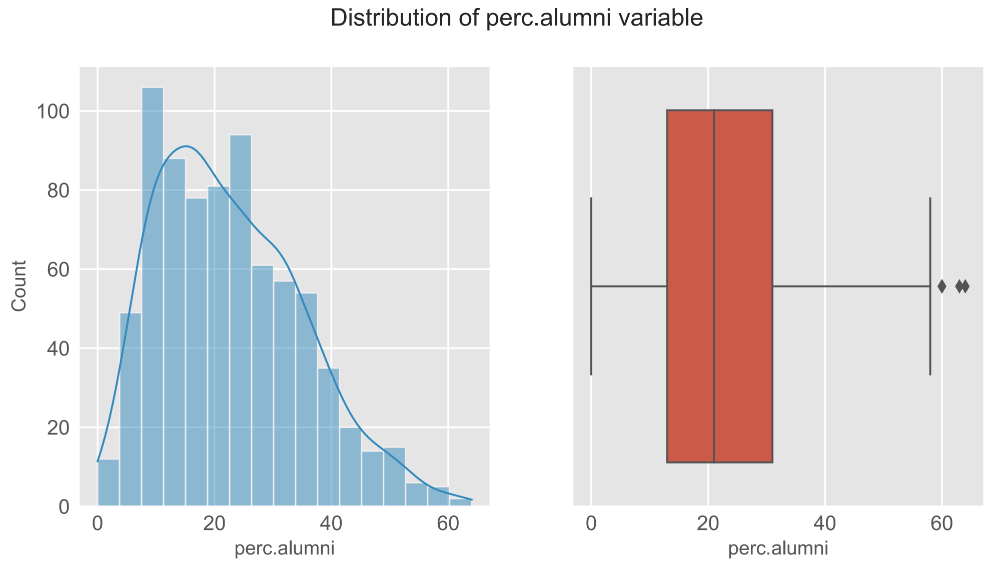

Figure 19: Distribution of perc.alumni variable.

The data in the **perc.alumni** variable is ***approximately normally distributed***. The variable ***contains few outliers***.

Univariate Analysis of Expend variable

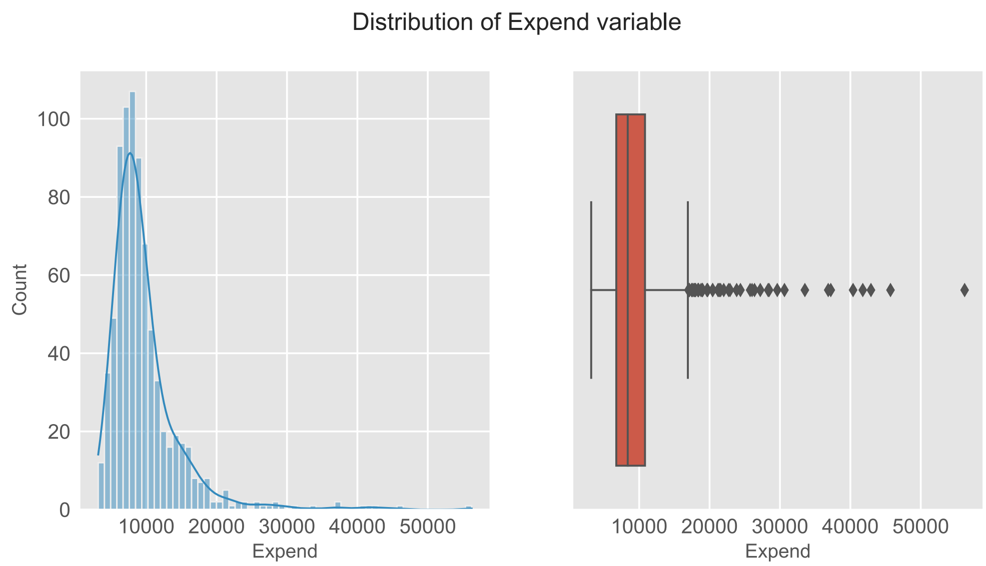

Figure 20: Distribution of Expend variable.

The data in the **Expend** variable is ***highly skewed to the right***. The variable also ***contains outliers***.

Univariate Analysis of Grad.Rate variable

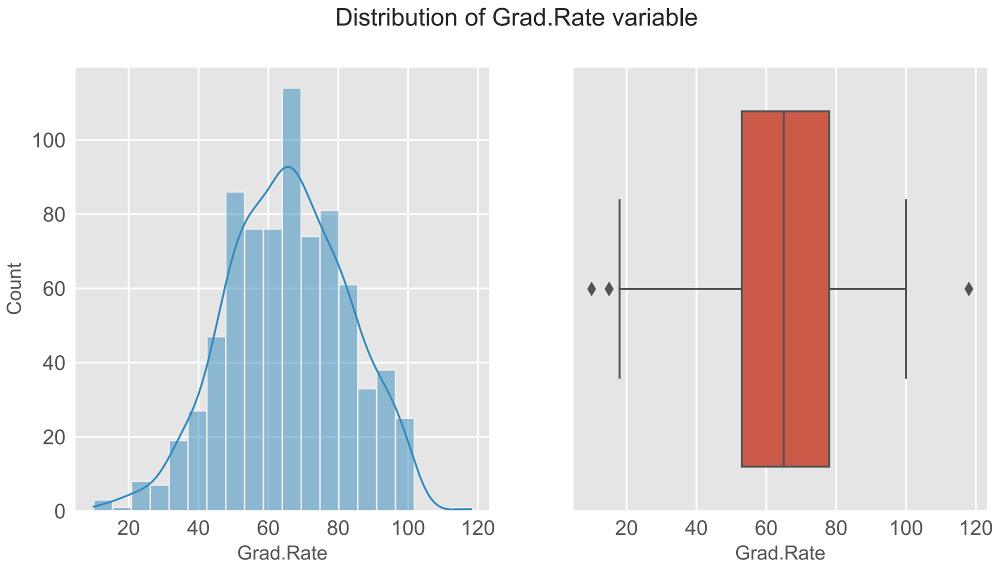

Figure 21: Distribution of Grad.Rate variable.

The data in the **Grad.Rate** variable is ***approximately normally distributed***. The variable ***contains few outliers***.

Multivariate Analysis of Education Dataset

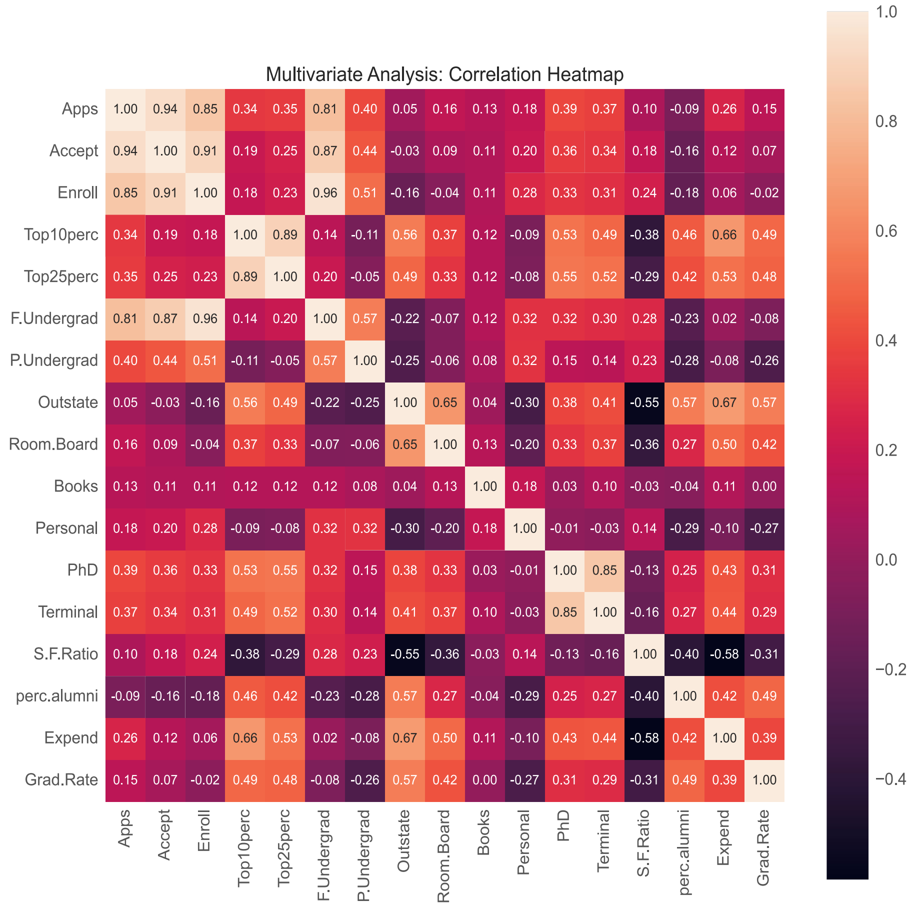

Figure 22: Correlation Heatmap

From the above correlation heatmap, we find that:

1. Variables **Apps**, **Accept**, **Enroll** and **F.Undergrad** are ***highly correlated***.
2. Variables **Top10perc** and **Top25perc** are ***highly correlated*** to each other.
3. Variables **F.Undergrad** and **P.Undergrad** are ***slightly correlated***.
4. **S.F.Rat**io has the ***lowest negatively correlation*** with **Outstate** and **Expend**.
5. Variables **Terminal** and **PhD** are ***highly correlated*** to each other.

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

PCA is sensitive to scale of the variables. It gives higher importance to variables with larger scales, i.e., the PCA components are greater for these variables. Feature scaling is required for data with different scales. In feature scaling, the mean is subtracted from the data points and then divided by the standard deviation for a particular variable. Also, subtracting the mean centres the data around the origin, which is required for PCA.

From the above Univariate analysis and also from the description in Table 7, we can see that the scales of the variables in the Education dataset is very different from each other. Therefore, scaling is required for this data to perform PCA.

	Apps	Accept	Enroll	Top10 perc	Top25 perc	F.Under grad	P.Under grad	Outstate	Room. Board	Books	Personal	PhD	Terminal	S.F. Ratio	perc. alumni	Expend	Grad. Rate
0	-0.347	-0.321	-0.064	-0.259	-0.192	-0.168	-0.209	-0.746	-0.965	-0.602	1.270	-0.163	-0.116	1.014	-0.868	-0.502	-0.318
1	-0.211	-0.039	-0.289	-0.656	-1.354	-0.210	0.244	0.457	1.909	1.216	0.236	-2.676	-3.378	-0.478	-0.545	0.166	-0.551
2	-0.407	-0.376	-0.478	-0.315	-0.293	-0.550	-0.497	0.201	-0.554	-0.905	-0.260	-1.205	-0.931	-0.301	0.586	-0.177	-0.668
3	-0.668	-0.682	-0.692	1.840	1.678	-0.658	-0.521	0.627	0.997	-0.602	-0.688	1.185	1.176	-1.615	1.151	1.793	-0.377
4	-0.726	-0.765	-0.781	-0.656	-0.596	-0.712	0.009	-0.717	-0.217	1.519	0.236	0.205	-0.524	-0.554	-1.675	0.242	-2.940

Table 8: Scaled Education dataset

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

$$r = \frac{cov(X, X)}{\sigma_X \sigma_X}$$

For scaled data, the variance is equal to 1. Hence, the correlation matrix and covariance matrix are equal to each other for scaled data. This can be confirmed from below heatmaps.

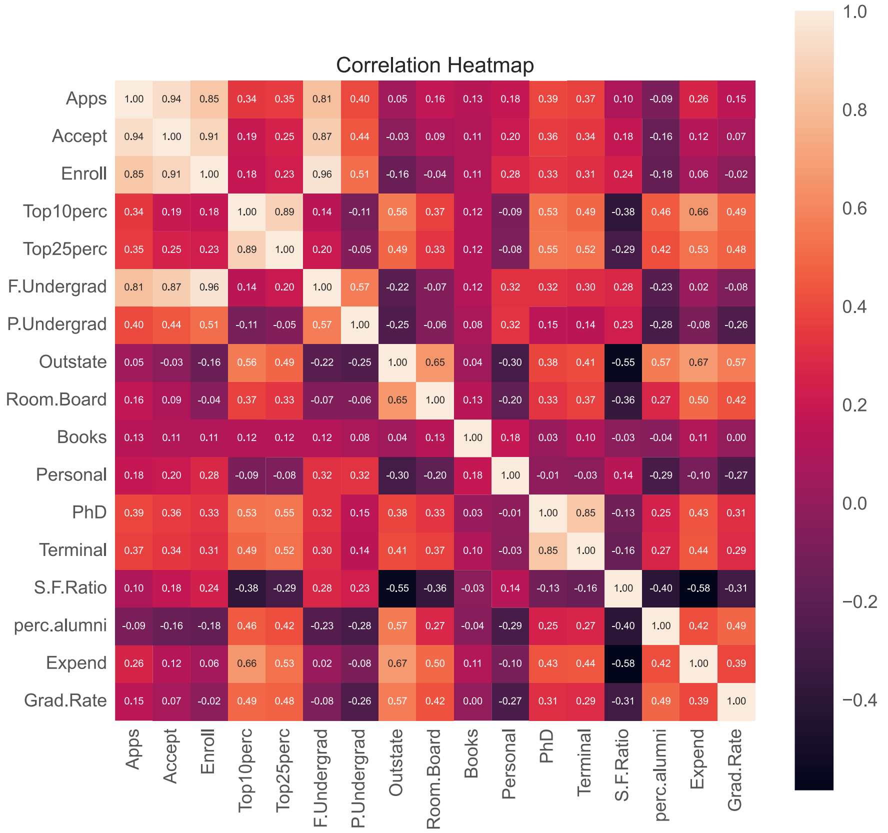

Figure 23: Correlation Heatmap of Scaled data

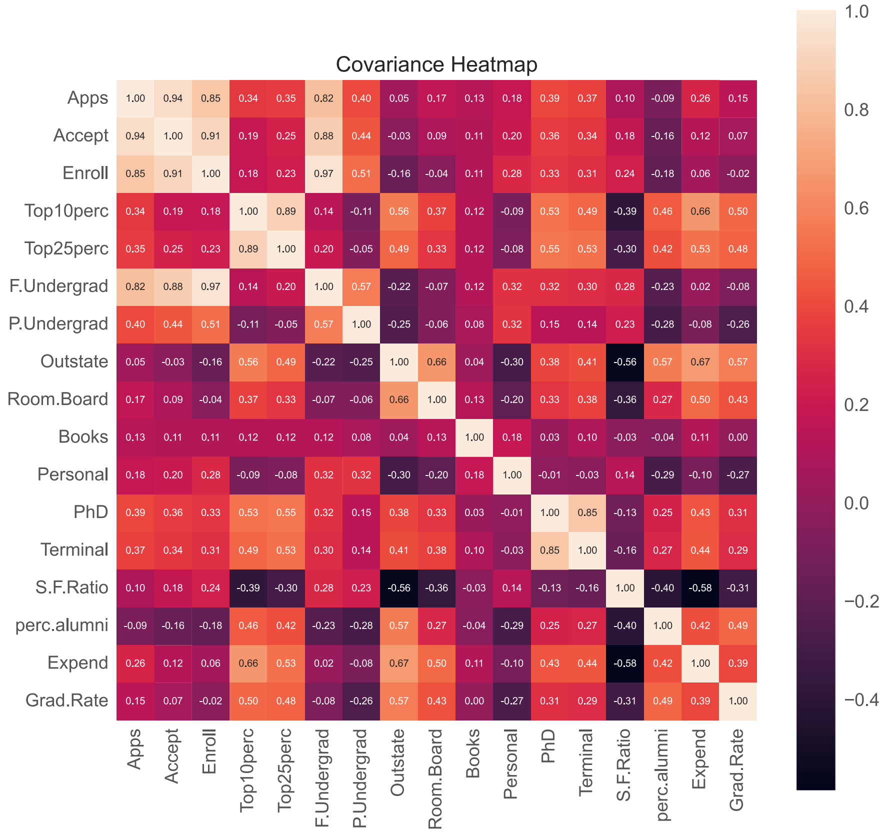

Figure 24: Covariance Heatmap of Scaled data

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

From the above Univariate analysis, we saw that the **original (unscaled) data had outliers** for many variables. **Scaling the data does not treat or remove the outliers**. Therefore, we expect **outliers to be present in the scaled data** as well. This can be confirmed from the below figure.

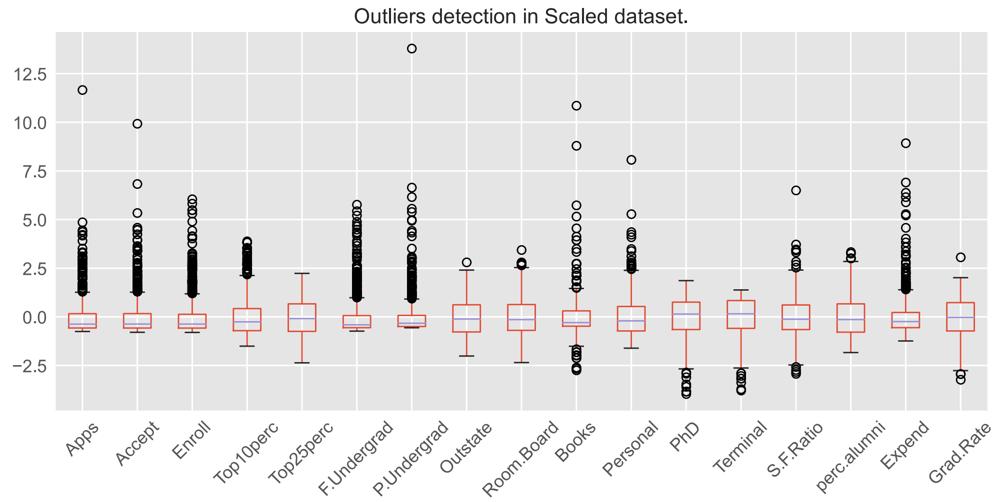

Figure 25: Outliers detection in Scaled dataset.

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

In PCA decomposition, **the eigenvectors help us understand the directions of spread of our data**, while **eigenvalues measures variance or the relative importance of these directions**.

```
Eigenvectors:
        [[ 2.488e-01  2.076e-01  1.763e-01  3.543e-01  3.440e-01  1.546e-01
           2.644e-02  2.947e-01  2.490e-01  6.476e-02 -4.253e-02  3.183e-01
           3.171e-01 -1.770e-01  2.051e-01  3.189e-01   2.523e-01]
         [ 3.316e-01  3.721e-01  4.037e-01 -8.241e-02 -4.478e-02  4.177e-01
           3.151e-01 -2.496e-01 -1.378e-01  5.634e-02  2.199e-01  5.831e-02
           4.643e-02  2.467e-01 -2.466e-01 -1.317e-01 -1.692e-01]
         [-6.309e-02 -1.012e-01 -8.299e-02  3.506e-02 -2.415e-02 -6.139e-02
           1.397e-01  4.660e-02  1.490e-01  6.774e-01  4.997e-01 -1.270e-01
          -6.604e-02 -2.898e-01 -1.470e-01  2.267e-01 -2.081e-01]
         [ 2.813e-01  2.678e-01  1.618e-01 -5.155e-02 -1.098e-01  1.004e-01
          -1.586e-01  1.313e-01  1.850e-01  8.709e-02 -2.307e-01 -5.347e-01
          -5.194e-01 -1.612e-01  1.731e-02  7.927e-02  2.691e-01]
         [ 5.741e-03  5.579e-02 -5.569e-02 -3.954e-01 -4.265e-01 -4.345e-02
           3.024e-01  2.225e-01  5.609e-01 -1.273e-01 -2.223e-01  1.402e-01
           2.047e-01 -7.939e-02 -2.163e-01  7.596e-02 -1.093e-01]
         [-1.624e-02  7.535e-03 -4.256e-02 -5.269e-02  3.309e-02 -4.345e-02
          -1.912e-01 -3.000e-02  1.628e-01  6.411e-01 -3.314e-01  9.126e-02
           1.549e-01  4.870e-01 -4.734e-02 -2.981e-01  2.162e-01]
         [-4.249e-02 -1.295e-02 -2.769e-02 -1.613e-01 -1.185e-01 -2.508e-02
           6.104e-02  1.085e-01  2.097e-01 -1.497e-01  6.338e-01 -1.096e-03
          -2.848e-02  2.193e-01  2.433e-01 -2.266e-01  5.599e-01]
         [-1.031e-01 -5.627e-02  5.866e-02 -1.227e-01 -1.025e-01  7.889e-02
           5.708e-01  9.846e-03 -2.215e-01  2.133e-01 -2.327e-01 -7.704e-02
          -1.216e-02 -8.360e-02  6.785e-01 -5.416e-02 -5.336e-03]
         [-9.023e-02 -1.779e-01 -1.286e-01  3.411e-01  4.037e-01 -5.944e-02
           5.607e-01 -4.573e-03  2.750e-01 -1.337e-01 -9.447e-02 -1.852e-01
          -2.549e-01  2.745e-01 -2.553e-01 -4.914e-02  4.190e-02]
         [ 5.251e-02  4.114e-02  3.449e-02  6.403e-02  1.455e-02  2.085e-02
          -2.231e-01  1.867e-01  2.983e-01 -8.203e-02  1.360e-01 -1.235e-01
          -8.858e-02  4.720e-01  4.230e-01  1.323e-01 -5.903e-01]
```

```
 [ 4.305e-02 -5.841e-02 -6.940e-02 -8.105e-03 -2.731e-01 -8.116e-02
   1.007e-01  1.432e-01 -3.593e-01  3.194e-02 -1.858e-02  4.037e-02
  -5.897e-02  4.450e-01 -1.307e-01  6.921e-01  2.198e-01]
 [ 2.407e-02 -1.451e-01  1.114e-02  3.855e-02 -8.935e-02  5.618e-02
  -6.354e-02 -8.234e-01  3.546e-01 -2.816e-02 -3.926e-02  2.322e-02
   1.649e-02 -1.103e-02  1.827e-01  3.260e-01  1.221e-01]
 [ 5.958e-01  2.926e-01 -4.446e-01  1.023e-03  2.188e-02 -5.236e-01
   1.260e-01 -1.419e-01 -6.975e-02  1.144e-02  3.945e-02  1.277e-01
  -5.831e-02 -1.772e-02  1.041e-01 -9.375e-02 -6.920e-02]
 [ 8.063e-02  3.347e-02 -8.570e-02 -1.078e-01  1.517e-01 -5.637e-02
   1.929e-02 -3.401e-02 -5.843e-02 -6.685e-02  2.753e-02 -6.911e-01
   6.710e-01  4.137e-02 -2.715e-02  7.312e-02  3.648e-02]
 [ 1.334e-01 -1.455e-01  2.959e-02  6.977e-01 -6.173e-01  9.916e-03
   2.095e-02  3.835e-02  3.402e-03 -9.439e-03 -3.090e-03 -1.121e-01
   1.589e-01 -2.090e-02 -8.418e-03 -2.277e-01 -3.394e-03]
 [ 4.591e-01 -5.186e-01 -4.043e-01 -1.487e-01  5.187e-02  5.604e-01
  -5.273e-02  1.016e-01 -2.593e-02  2.883e-03 -1.289e-02  2.981e-02
  -2.708e-02 -2.125e-02  3.334e-03 -4.388e-02 -5.008e-03]
 [ 3.590e-01 -5.434e-01  6.097e-01 -1.450e-01  8.035e-02 -4.147e-01
   9.018e-03  5.090e-02  1.146e-03  7.726e-04 -1.114e-03  1.381e-02
   6.209e-03 -2.222e-03 -1.919e-02 -3.531e-02 -1.307e-02]]

Eigenvalues:
        [5.451 4.484 1.175 1.008 0.934 0.848 0.606 0.588 0.531 0.404 0.313 0.221
         0.168 0.144 0.088 0.037 0.023]
```

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

	Apps	Accept	Enroll	Top10 perc	Top25 perc	F.Under grad	P.Under grad	Outstate	Room. Board	Books	Personal	PhD	Terminal	S.F. Ratio	perc. alumni	Expend	Grad. Rate
PC1	0.249	0.208	0.176	0.354	0.344	0.155	0.026	0.295	0.249	0.065	-0.043	0.318	0.317	-0.177	0.205	0.319	0.252
PC2	0.332	0.372	0.404	-0.082	-0.045	0.418	0.315	-0.250	-0.138	0.056	0.220	0.058	0.046	0.247	-0.247	-0.132	-0.169
PC3	-0.063	-0.101	-0.083	0.035	-0.024	-0.061	0.140	0.047	0.149	0.677	0.500	-0.127	-0.066	-0.290	-0.147	0.227	-0.208
PC4	0.281	0.268	0.162	-0.052	-0.110	0.100	-0.159	0.131	0.185	0.087	-0.231	-0.535	-0.519	-0.161	0.017	0.079	0.269
PC5	0.006	0.056	-0.056	-0.395	-0.427	-0.043	0.302	0.223	0.561	-0.127	-0.222	0.140	0.205	-0.079	-0.216	0.076	-0.109
PC6	-0.016	0.008	-0.043	-0.053	0.033	-0.043	-0.191	-0.030	0.163	0.641	-0.331	0.091	0.155	0.487	-0.047	-0.298	0.216
PC7	-0.042	-0.013	-0.028	-0.161	-0.118	-0.025	0.061	0.109	0.210	-0.150	0.634	-0.001	-0.028	0.219	0.243	-0.227	0.560
PC8	-0.103	-0.056	0.059	-0.123	-0.102	0.079	0.571	0.010	-0.221	0.213	-0.233	-0.077	-0.012	-0.084	0.679	-0.054	-0.005
PC9	-0.090	-0.178	-0.129	0.341	0.404	-0.059	0.561	-0.005	0.275	-0.134	-0.094	-0.185	-0.255	0.275	-0.255	-0.049	0.042
PC10	0.053	0.041	0.034	0.064	0.015	0.021	-0.223	0.187	0.298	-0.082	0.136	-0.123	-0.089	0.472	0.423	0.132	-0.590
PC11	0.043	-0.058	-0.069	-0.008	-0.273	-0.081	0.101	0.143	-0.359	0.032	-0.019	0.040	-0.059	0.445	-0.131	0.692	0.220
PC12	0.024	-0.145	0.011	0.039	-0.089	0.056	-0.064	-0.823	0.355	-0.028	-0.039	0.023	0.016	-0.011	0.183	0.326	0.122
PC13	0.596	0.293	-0.445	0.001	0.022	-0.524	0.126	-0.142	-0.070	0.011	0.039	0.128	-0.058	-0.018	0.104	-0.094	-0.069
PC14	0.081	0.033	-0.086	-0.108	0.152	-0.056	0.019	-0.034	-0.058	-0.067	0.028	-0.691	0.671	0.041	-0.027	0.073	0.036
PC15	0.133	-0.145	0.030	0.698	-0.617	0.010	0.021	0.038	0.003	-0.009	-0.003	-0.112	0.159	-0.021	-0.008	-0.228	-0.003
PC16	0.459	-0.519	-0.404	-0.149	0.052	0.560	-0.053	0.102	-0.026	0.003	-0.013	0.030	-0.027	-0.021	0.003	-0.044	-0.005
PC17	0.359	-0.543	0.610	-0.145	0.080	-0.415	0.009	0.051	0.001	0.001	-0.001	0.014	0.006	-0.002	-0.019	-0.035	-0.013

Table 9: Principal Components with Original features.

Each Principal Component is a linear combination of the original scaled features in the data. For each PC, the row of length 17 gives the weights with which the corresponding variables need to be multiplied to get the PC.

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]

$$\begin{aligned} PC_1 = {} & 0.25 \times Apps + 0.21 \times Accept + 0.18 \times Enroll + 0.35 \times Top10perc + 0.34 \times Top25perc \\ & + 0.16 \times F.Undergrad + 0.03 \times P.Undergrad + 0.29 \times Outstate \\ & + 0.25 \times Room.Board + 0.06 \times Books - 0.04 \times Personal + 0.32 \times PhD \\ & + 0.32 \times Terminal - 0.18 \times S.F.Ratio + 0.21 \times perc.alumni + 0.32 \times Expend \\ & + 0.25 \times Grad.Rate \end{aligned}$$

Similarly, the other PCs can also be expressed in terms of the scaled variables.

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?

	PCs	Proportion Of Variance	Cumulative Proportion
0	PC1	0.32	0.32
1	PC2	0.26	0.58
2	PC3	0.07	0.65
3	PC4	0.06	0.71
4	PC5	0.05	0.77
5	PC6	0.05	0.82
6	PC7	0.04	0.85
7	PC8	0.03	0.89
8	PC9	0.03	0.92
9	PC10	0.02	0.94
10	PC11	0.02	0.96
11	PC12	0.01	0.97
12	PC13	0.01	0.98
13	PC14	0.01	0.99
14	PC15	0.01	1.00
15	PC16	0.00	1.00
16	PC17	0.00	1.00

Table 10: Cumulative Proportion of Variance

Eigenvalue associated with a particular PC, measures the variance explained by that PC. Calculating the cumulative proportion of variance helps us to decide on the optimum number of Principal Components.

For example, from the above table, around 92% of the variance is explained by the first 9 principal components only. Therefore, we can select these PCs for further analysis, thereby reducing the dimensions of the original data from 17 to 9 without losing a lot of information.

Eigenvectors are unit vectors in the directions of the PCs. All the PCs are orthogonal to each other i.e., the correlation of PCs with each other is zero.

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]

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Apps	0.17	0.93	0.04	0.12	0.12	0.05	0.03	-0.03	0.03
Accept	0.04	0.97	0.03	0.13	0.07	-0.02	0.03	-0.03	0.04
Enroll	0.07	0.94	0.03	0.13	-0.09	-0.07	0.09	-0.02	0.14
Top10perc	0.85	0.14	0.05	0.24	0.15	0.30	-0.03	0.17	-0.06
Top25perc	0.87	0.17	0.05	0.28	0.13	0.16	-0.03	0.18	-0.03
F.Undergrad	0.07	0.91	0.04	0.14	-0.11	-0.11	0.12	-0.06	0.21
P.Undergrad	-0.07	0.40	0.02	0.07	-0.02	-0.11	0.14	-0.11	0.87
Outstate	0.26	-0.10	0.00	0.24	0.56	0.49	-0.15	0.32	-0.10
Room.Board	0.11	0.01	0.09	0.20	0.86	0.28	-0.10	0.01	0.05
Books	0.06	0.08	0.99	0.01	0.06	0.03	0.09	-0.02	0.02
Personal	-0.04	0.17	0.10	-0.01	-0.13	-0.03	0.95	-0.13	0.12
PhD	0.27	0.24	-0.04	0.87	0.14	0.09	0.01	0.09	0.03
Terminal	0.20	0.22	0.06	0.89	0.16	0.12	-0.02	0.10	0.04
S.F.Ratio	-0.12	0.17	-0.00	-0.01	-0.13	-0.87	0.00	-0.16	0.08
perc.alumni	0.21	-0.17	-0.01	0.17	0.03	0.31	-0.15	0.83	-0.03
Expend	0.38	0.10	0.05	0.23	0.28	0.72	-0.03	0.09	-0.02
Grad.Rate	0.34	0.06	-0.06	0.06	0.54	0.01	-0.05	0.59	-0.27

Table 11: Factor loadings of original variables.

From *Table 10*, we see that 92% of the total variance in the data is explained by the first 9 Principal components. Therefore, we use these 9 PCs for further analysis.

From the above correlation table:

1. **PC1** gives maximum loadings to **Top10perc** and **Top25perc**. This PC represents **Top Students**.
2. **PC2** gives maximum loadings to **Apps**, **Accept**, **Enroll** and **F.Undergrad**. This PC may represent **Admission Process**.
3. **PC3** gives maximum loadings to **Books**. This PC may represent **Books Expenditure**.
4. **PC4** gives maximum loadings to **PhD** and **Terminal**. This PC may represent **Faculty Qualification**.
5. **PC5** gives maximum loadings to **Room.Board**. This PC may represent **Living Expenditure**.
6. **PC6** gives maximum loadings to **S.F.Ratio** and **Expend**. This PC may represent **Teaching Cost**.
7. **PC7** gives maximum loadings to **Personal**. This PC may represent **Personal Expenditure**.
8. **PC8** gives maximum loadings to **perc.alumni**. This PC may represent **Alumni Donations**.
9. **PC9** gives maximum loadings to **P.Undergrad**. This PC may represent **Part-Time Admissions**.