|  |
| --- |
|  |
| AS PROJECT REPORT |
| DSBA |
|  |
| **Shripad Anwekar / PGPDSBA Online Mar\_A2021** |
| **6/13/2021** |

|  |
| --- |
|  |

Contents

[1. Problem 1A 4](#_Toc74496658)

[1.1. State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually. 4](#_Toc74496659)

[Hypothesis for conducting one-way ANOVA for Education Level as a Factor 4](#_Toc74496660)

[Hypothesis for conducting one-way ANOVA for Occupation Level as a Factor 4](#_Toc74496661)

[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. 5](#_Toc74496662)

[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. 5](#_Toc74496663)

[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) 5](#_Toc74496664)

[Problem 1B 6](#_Toc74496665)

[1.5. The Student News Service at Clear Mountain State University (CMSU) has decided to gather What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the ‘pointplot’ function from the ‘seaborn’ function] 6](#_Toc74496666)

[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? 7](#_Toc74496667)

[1.7. Explain the business implications of performing ANOVA for this particular case study. 8](#_Toc74496668)

[One-Way Anova 8](#_Toc74496669)

[Two-Way Anova 8](#_Toc74496670)

[2. Problem 2 9](#_Toc74496671)

[2.1. Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA? 9](#_Toc74496672)

[Describe 9](#_Toc74496673)

[Data Pre-processing 9](#_Toc74496674)

[Data Visualization 9](#_Toc74496675)

[Edu [Apps] 10](#_Toc74496676)

[Edu [Accept] 12](#_Toc74496677)

[Edu [Enroll] 14](#_Toc74496678)

[Edu [Top10perc] 16](#_Toc74496679)

[Edu [Top25perc] 18](#_Toc74496680)

[Edu [F.Undergrad] 20](#_Toc74496681)

[Edu [P.Undergrad] 22](#_Toc74496682)

[Edu [Outstate] 24](#_Toc74496683)

[Edu [Room.Board] 26](#_Toc74496684)

[Edu [Books] 28](#_Toc74496685)

[Edu [Personal] 30](#_Toc74496686)

[Edu [PhD] 32](#_Toc74496687)

[Edu [Terminal] 34](#_Toc74496688)

[Edu [S.F.Ratio] 36](#_Toc74496689)

[Edu [perc.alumni] 38](#_Toc74496690)

[Edu [Expend] 40](#_Toc74496691)

[Edu [Grad.Rate] 42](#_Toc74496692)

[Edu [Apps Vs Accept] 44](#_Toc74496693)

[Edu [Apps Vs Enroll] 44](#_Toc74496694)

[Edu [Apps Vs Top10perc] 45](#_Toc74496695)

[Edu [Apps Vs Top25perc] 45](#_Toc74496696)

[Edu [Apps Vs F.Undergrad] 46](#_Toc74496697)

[Edu [Apps Vs P.Undergrad] 46](#_Toc74496698)

[Edu [Apps Vs Outstate] 47](#_Toc74496699)

[Edu [Apps Vs Room.Board] 47](#_Toc74496700)

[Edu [Apps Vs Books] 48](#_Toc74496701)

[Edu [Apps Vs Personal] 48](#_Toc74496702)

[Edu [Apps Vs PhD] 49](#_Toc74496703)

[Edu [Apps Vs Terminal] 49](#_Toc74496704)

[Edu [Enroll Vs Top10perc] 50](#_Toc74496705)

[Edu [Enroll Vs Top25perc] 50](#_Toc74496706)

[Edu [Enroll Vs F.Undergrad] 51](#_Toc74496707)

[Edu [Enroll Vs P.Undergrad] 51](#_Toc74496708)

[Edu [Enroll Vs Outstate] 52](#_Toc74496709)

[Edu [Enroll Vs Room.Board] 52](#_Toc74496710)

[Edu [Books] 53](#_Toc74496711)

[Edu [Personal] 53](#_Toc74496712)

[Edu [PhD] 54](#_Toc74496713)

[Edu [Terminal] 54](#_Toc74496714)

[2.2. Is scaling necessary for PCA in this case? Give justification and perform scaling. 55](#_Toc74496715)

[2.3. Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data]. 56](#_Toc74496716)

[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] 56](#_Toc74496717)

[2.5. Extract the eigenvalues and eigenvectors.[print both] 59](#_Toc74496718)

[2.6. Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features 61](#_Toc74496719)

[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] 63](#_Toc74496720)

[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? 64](#_Toc74496721)

[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] 66](#_Toc74496722)

[Appendices 68](#_Toc74496723)

[List of Plots 68](#_Toc74496724)

[List of Tables 69](#_Toc74496725)

# Problem 1A

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [[SalaryData.csv](https://olympus.greatlearning.in/courses/39185/files/2922634/download?verifier=5goWBurNldK9FNqSe9IYvlQvOzwn5P2kHRqyJ36U&wrap=1)] 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.

 [Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.]

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

### Hypothesis for conducting one-way ANOVA for Education Level as a Factor

#### Null hypothesis H0 : The mean Salary received is same across different Education levels i.e. Doctorate, Bachelors and HS-grad

#### Alternate hypothesis Ha : The mean Salary received is different in at-least one category of Education than others where Education category as provided in data set are ' Doctorate', ' Bachelors' and ' HS-grad'

### Hypothesis for conducting one-way ANOVA for Occupation Level as a Factor

#### Null hypothesis H0 : The mean Salary received is same across different Occupation levels ' Adm-clerical', ' Sales', ' Prof-specialty' and ' Exec-managerial'

#### Alternate hypothesis Ha : The mean Salary received is different in at-least one category of Occupation levels which are ' Adm-clerical', ' Sales', ' Prof-specialty' and ' Exec-managerial'

## 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 returns the p-value of 0.00 which is less than the level of significance (0.05), we reject the null hypothesis and conclude that there is a difference in the mean salary received by the education levels by at least one level of Education and that Mean Salary paid to Education levels ' Doctorate', ' Bachelors' and ' HS-grad' varies.

#### The decision is again supported by the F Value of (30.96) which is fairly high as well the F-Critical Value of 3.252

## 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 Occupation returns the p-value (0.46 ) which is Greater than the level of significance (0.05), we fail to reject the null hypothesis and conclude that there is no a difference in the mean salary received by any one of the Occupations, and that the Mean Salary paid to Occupations ' Adm-clerical', ' Sales', ' Prof-specialty' and ' Exec-managerial' remains same.

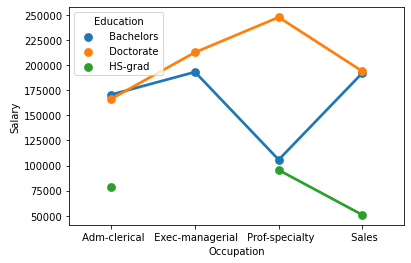
#### The decision is again supported by the F Value of (0.88) which is fairly low as well the F-Critical Value of 2.8663

## 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)

#### We performed a one-way ANOVA on Salary with respect to Education (Question 1.2). The null hypothesis is REJECTED based on the ANOVA results. Data shows that the mean salary to "HS-GRAD" differs significantly

# Problem 1B

## The Student News Service at Clear Mountain State University (CMSU) has decided to gather What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.[hint: use the ‘pointplot’ function from the ‘seaborn’ function]



***Plot 1.5.1***

#### Point chart (Plot 1.5.1) as well as 0.00 P=Value of variance analysis shows a really high degree of interaction between Education and Occupation impacting the Salary.

#### 'HS-Grad' at every occupation earns less than 'Doctorate' or 'Bachelors'.

#### 'Doctorate' at every occupation earns More than 'HS-Grad' or 'Bachelors'.

#### 'Bachelors' at every occupation earn More than 'HS-Grad'.

#### There is no evidence that The Salaries are independent of combination of 'Education' and 'Occupation'.

## 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?

#### Hypothesis for conducting Two-way ANOVA for Education Level as a Factor / Treatment

#### Null hypothesis H0 : The mean Salary received is same across different Education levels i.e. Doctorate, Bachelors and HS-grad

#### Alternate hypothesis Ha : The mean Salary received is different in at-least one category of Education than others where Education category as provided in data set are ' Doctorate', ' Bachelors' and ' HS-grad'

#### Hypothesis for conducting Two-way ANOVA for Occupation Level as a Factor / Treatment

#### Null hypothesis H0 : The mean Salary received is same across different Occupation levels ' Adm-clerical', ' Sales', ' Prof-specialty' and ' Exec-managerial'

#### Alternate hypothesis Ha : The mean Salary received is different in at-least one category of Occupation levels which are ' Adm-clerical', ' Sales', ' Prof-specialty' and ' Exec-managerial'

#### Hypothesis for conducting Two-way ANOVA for Interaction between Factors Education and Occupation to understand impact on Salary

#### Null hypothesis H0 : There is no impact of Interaction between factors Education and Occupation on Salary

#### Alternate hypothesis Ha : There is impact of Interaction between factors Education and Occupation on Salary

Result Interpretation

A 2-Way ANOVA was run on a sample of 40 individuals to examine the effect of Education and Occupation on Salary.

* There was a significant interaction as the P-Value of 0.00 < 0.05 between the effects of Education and Occupation on Salary.
* Simple main effects analysis showed that Occupation has no significant effect on Salary as the P-Value of 0.07 > 0.05
* Treatment level (Education) has significant effect on Salary as the P-Value of 0.00 < 0.05

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

### One-Way Anova

One-Way Anova is performed to determine if there are any statistically significant differences between means of two or more independent (Unrelated) groups. Ideally minimum 3. In current situation for Education levels when we performed this analysis it revealed that that there is difference in the mean of Salary received by Education levels 'Doctorate', 'Bachelors' and 'HS-grad'. This outcome of test is logically correct as it is expected that Salaries will vary for staff with varied Education levels.

One-Way Anova is performed to determine if there are any statistically significant differences between means of two or more independent (Unrelated) groups. Ideally minimum 3. In current situation for Occupation levels when we performed this analysis it revealed that that there is No difference in the mean of Salary received by Occupation levels 'Adm-clerical', 'Exec-managerial', 'Prof-specialty' and 'Sales'. This outcome of test is logically correct as it is expected that Salaries Spend will be similar for staff employed with different Departments. IN order to strengthen each of the department equally.

### Two-Way Anova

Two-Way is extension of One-way ANOVA that examines the influence of two different categorical independent variables on one continuous dependent variable. A 2-Way ANOVA was run on a sample of 40 individuals to examine the effect of Education and Occupation on Salary.

1) Simple main effects analysis showed that Occupation has no significant effect on Salary as the P-Value of 0.07 > 0.05; which is logically correct as Each occupation has staff with varied Education background and they are paid differently summing up to similar pay-out per occupation.

2) There was a significant interaction as the P-Value of 0.00 < 0.05 between the effects of Education and Occupation on Salary. This is logically correct since the Education levels within same occupation are paid differently.

3) Treatment level (Education) has significant effect on Salary as the P-Value of 0.00 < 0.05. This is logically correct since the Education levels within same occupation are expected to be paid differently.

# Problem 2

The dataset [Education - Post 12th Standard.csv](https://olympus.greatlearning.in/courses/39185/files/2196843/download?verifier=56SJ0y5mRJxgZSDC65L5avKL2cGJlT2ySEHZlyos&wrap=1) 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](https://olympus.greatlearning.in/courses/39185/files/2196842/download?verifier=TWsl0uGmUA2leKxplEfuskkpKCAPM7U9ukhvKWrD&wrap=1).

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

### Describe

Supplied Data set has 18 columns, with 1 categorical and 17 numerical values. Supplied data dictionary is reviewed against the data set and all Data types are found correct.

Quantity and relevance of data seem to be correct from Exercise stand point.

### Data Pre-processing

***Treat BAD Data***

Data types are in line with the supporting data dictionary, Presense of no Bad Data is noticed

***Treat Anomalies***

There are apparently no noticeable Anomalies, with the definition of Existing Business rules that are available

***Treat Missing Values***

There are no Missing values in the supplied DataSet

***Checking for Duplicates***

There are no Duplicate values in the supplied DataSet

### Data Visualization

***Univariate Analysis***

Following reviews are done for All 17 numeric variables to understand the trend

1. 5 Number Analysis
2. Histogram Review
3. BoxPlot
4. Empirical Rule
5. mean, median and mode

We will go through each of these variables to understand the inference of above reviews as ‘Insight’ for Each of the analysis Point.

### Edu [Apps]

* 5 number Summary Suggests that -

Distribution is Right skewed -

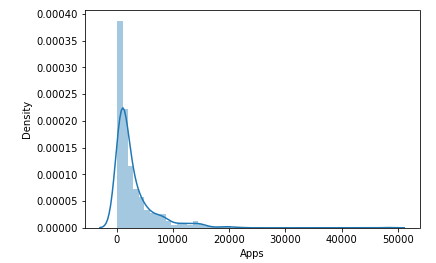
(Q1 - Min) which is 695.00 < Max - Q3 which is 44470.00

Following points support the symmetry -

(Q2 - Min) which is 1477.00 < (Max - Q2) which is 46536.00

(Q2 - Q1) which is 782.00 < (Q3 - Q2) which is 2066.00

* Histogram Review



***Plot 2.1.1***

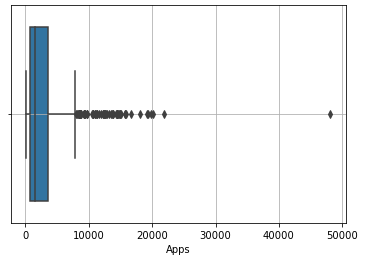
Histogram Review for Edu['Apps'] Shows -

1) 1 Peak

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.2***

BoxPlot Review for Edu ['Apps'] Shows -

1) There are many outliers

2) Distribution is Right Skewed due to longer Right Whisker

* Empirical Rule

Edu['Apps'] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Edu['Apps'] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (Apps):**

Number of applications received range from 81 to 48094

75% of the institutes have received less than 3624 applications.

From above figure, we can say that the Applications received across institutes is right skewed

Average Number of applications received is 3001.64 which are much higher than the median value 1558.00 indicating that the distribution is right tailed.

There are Outliers towards higher side which indicates there are few institutes that have received outstanding response in terms of number of applications.

### Edu [Accept]

* 5 number Summary Suggests that -

Distribution is Right skewed -

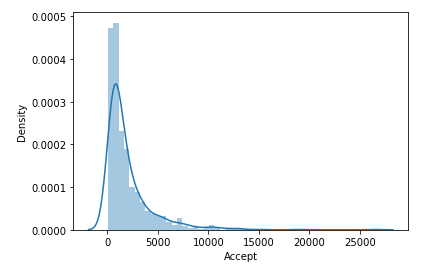
(Q1 - Min) which is 532.000 < Max - Q3 which is 23906.00

Following points support the symmetry -

(Q2 - Min) which is 1038.00 < (Max - Q2) which is 25220.00

(Q2 - Q1) which is 506.00 < (Q3 - Q2) which is 1314.00

* Histogram Review



***Plot 2.1.3***

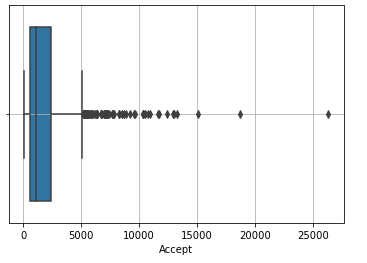
Histogram Review for Edu['Accept'] Shows -

1) 1 Peak

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.4***

BoxPlot Review for Edu['Accept'] Shows -

1) There are many outliers

2) Distribution is Right Skewed due to longer Right Whisker

* Empirical Rule

Edu ['Accept'] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Edu['Accept'] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights ('Accept') :**

Number of applications Accepted range from 72 to 26330

75% of the institutes have accepted less than 2424 applications.

From above figure, we can say that the Applications Accepted across institutes is right skewed

Average Number of applications Accepted is 2018.80 which is much higher than the median value 1110.00 indicating that the distribution is right tailed.

There are Outliers towards higher side which indicates there are few institutes that have accepted more applications and have more capacity.

### Edu [Enroll]

* 5 number Summary Suggests that -

Distribution is Right skewed -

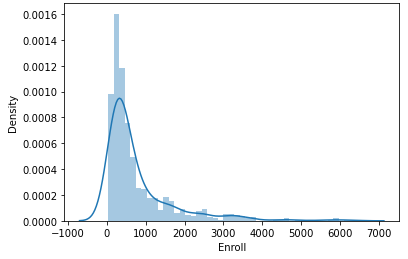
(Q1 - Min) which is 207.00 < Max - Q3 which is 5490.00

Following points support the symmetry -

(Q2 - Min) which is 399.00 < (Max - Q2) which is 5958.00

(Q2 - Q1) which is 192.00 < (Q3 - Q2) which is 468.00

* Histogram Review



***Plot 2.1.5***

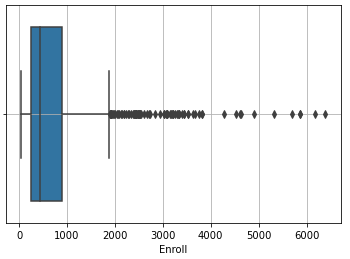
Histogram Review for Edu['Enroll '] Shows -

1) 1 Peak

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.6***

BoxPlot Review for Edu['Enroll'] Shows -

1) There are many outliers

2) Distribution is Right Skewed due to longer Right Whisker

* Empirical Rule

Edu[' Enroll '] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Edu[' Enroll '] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights ('Enroll’):**

Number of students enrolled range from 35.00 to 6392.00 which is much less than accepted. This means the institutes will still have capacity.

75% of the institutes have less than 902 Enrolments.

From above figure, we can say that the Enrolments across institutes is right skewed which means there are More institutes with lower number of Enrolments

Average Number of Enrolments is 779.97 which are much higher than the median value 434.00 indicating that the distribution is right tailed. There are Outliers towards higher side which indicates there are few institutes that have enrolled more applications and have more capacity.

### Edu [Top10perc]

* 5 number Summary Suggests that -

Distribution is Right skewed -

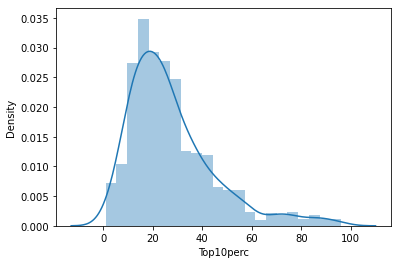
(Q1 - Min) which is 14.00 < Max - Q3 which is 61.00

Following points support the symmetry -

(Q2 - Min) which is 8.00 < (Max - Q2) which is 73.00

(Q2 - Q1) which is 192.00 < (Q3 - Q2) which is 12.00

* Histogram Review



***Plot 2.1.7***

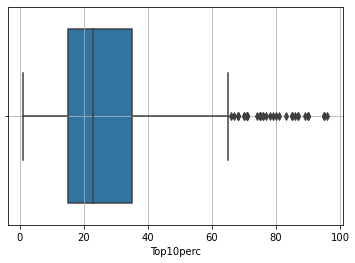
Histogram Review for Edu[‘Top10perc'] Shows -

1) 1 Peak

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.8***

BoxPlot Review for Edu [‘Top10perc’] Shows -

1) There are many outliers

2) Distribution is Right Skewed due to longer Right Whisker

* Empirical Rule

Edu [‘Top10perc’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Edu[‘Top10perc’] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (‘Top10perc’):**Number of Percentage of new students from top 10% of Higher Secondary class, range from 1.00 to 96.00

75% of the institutes have less than 35 new students from top 10% of Higher Secondary class.

From above figure, we can say that the new students from top 10% of Higher Secondary class across institutes is right skewed which means there are More institutes with lower number of new students from top 10% of Higher Secondary class

Average Number of new students from top 10% of Higher Secondary class is 27.56 which are higher than the median value 23.00 indicating that the distribution is right tailed.

There are Outliers towards higher side which indicates there are few institutes that have Enrolled more new students from top 10% of Higher Secondary class.

### Edu [Top25perc]

* 5 number Summary Suggests that -

Distribution is Almost Normal -

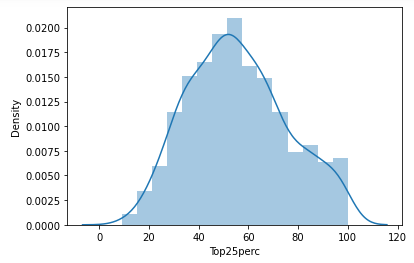
(Q1 - Min) which is 32.00 approximately equal to Max - Q3 which is 31.00

Following points support the symmetry -

(Q2 - Min) which is 45.00 approximately equal to (Max - Q2) which is 46.00

(Q2 - Q1) which is 13.00 approximately equal to (Q3 - Q2) which is 15.00

* Histogram Review



***Plot 2.1.9***

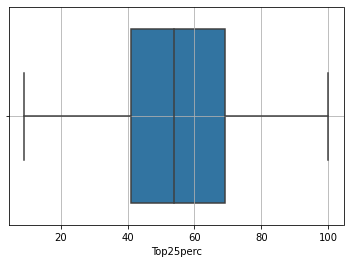
Histogram Review for Edu [‘Top25perc'] Shows -

1) 1 Peak

2) Approximate Normal Distribution

3) It shows a close to perfect 'Bell Curve

* BoxPlot Review



***Plot 2.1.10***

BoxPlot Review for Edu[‘Top25perc’] Shows -

1) There are No outliers

2) Distribution is Approximate Normal

* Empirical Rule

Edu [‘Top25perc’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is almost symmetrical

2) '2 standard deviation' from Mean is almost symmetrical

3) '3 standard deviation' from Mean is almost symmetrical

* Edu[‘Top25perc’] review of Mean, Median and Mode suggests that –

In this case Mean = Median = Mode hence the distribution is Approximate Normal

**Insights (‘Top25perc’):** Number of Percentage of new students from top 25% of Higher Secondary class, range from 9.00 to 100.00

75% of the institutes have less than 69 new students from top 25% of Higher Secondary class.

From above figure, we can say that the new students from top 25% of Higher Secondary class across institutes are spread normally across the listed university.

Average Number of new students from top 25% of Higher Secondary class is 55.80 which are approximately equal to median value 54.00 indicating that the distribution is approximately normal.

There are no Outliers which indicate there are no institutes that have enrolled higher / lower number of students from top 25% of Higher Secondary class.

### Edu [F.Undergrad]

* 5 number Summary Suggests that -

Distribution is Right skewed -

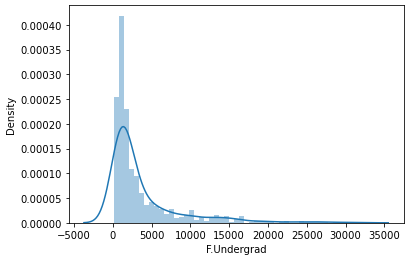
(Q1 - Min) which is 853.00 < Max - Q3 which is 27638.00

Following points support the symmetry -

(Q2 - Min) which is 1568.00 < (Max - Q2) which is 29936.00

(Q2 - Q1) which is 715.00 < (Q3 - Q2) which is 2298.00

* Histogram Review



***Plot 2.1.11***

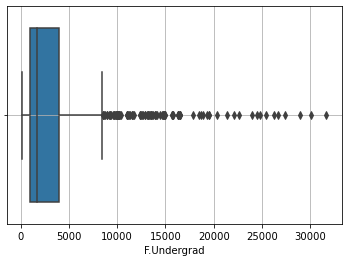
Histogram Review for Edu [‘F.Undergrad'] Shows -

1) 1 Peak

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.12***

BoxPlot Review for Edu [‘F.Undergrad’] Shows -

1) There are many outliers

2) Distribution is Right Skewed due to longer Right Whisker

* Empirical Rule

Edu [‘F.Undergrad’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Edu[‘F.Undergrad’] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (‘F.Undergrad’):**

Number of full-time undergraduate students, range from 139.00 to 31643.00

75% of the institutes have less than 4005 students full-time undergraduate.

From above figure, we can say that the new full-time undergraduate students institutes is right skewed which means there are More institutes with lower number of full-time undergraduate students

Average Number of full-time undergraduate students is 3699.91 which are higher than the median value 1707.00 indicating that the distribution is right tailed.

There are Outliers towards higher side which indicates there are few institutes with full-time undergraduate students

### Edu [P.Undergrad]

* 5 number Summary Suggests that -

Distribution is Right skewed -

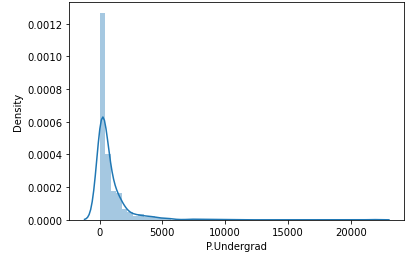
(Q1 - Min) which is 94.00 < Max - Q3 which is 20869.00

Following points support the symmetry -

(Q2 - Min) which is 352.00 < (Max - Q2) which is 21483.00

(Q2 - Q1) which is 258.00 < (Q3 - Q2) which is 614.00

* Histogram Review



***Plot 2.1.13***

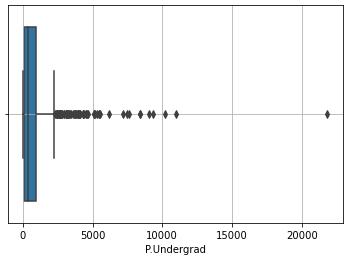
Histogram Review for Edu[‘P.Undergrad'] Shows -

1) 1 Peak

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.14***

BoxPlot Review for Edu[‘P.Undergrad’] Shows -

1) There are many outliers

2) Distribution is Right Skewed due to longer Right Whisker

* Empirical Rule

Edu[‘P.Undergrad’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Edu[‘P.Undergrad’] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (‘P.Undergrad’):**

Number of Part-time undergraduate students, range from 1.00 to 21836.00

75% of the institutes have less than 967.00 students part-time undergraduate.

From above figure, we can say that the new Part-time undergraduate students per institutes is right skewed which means there are More institutes with lower Count of full-time undergraduate students.

Average Number of Part-time undergraduate students is 855.30 which is higher than the median value 353.00 indicating that the distribution is right tailed.

There are Outliers towards higher side which indicates there are few institutes with relatively higher count of part-time undergraduate students

### Edu [Outstate]

* 5 number Summary Suggests that -

Distribution is Right skewed -

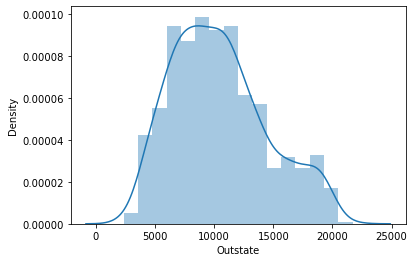
(Q1 - Min) which is 4980.00 < Max - Q3 which is 8775.00

Following points support the symmetry -

(Q2 - Min) which is 7650.00 < (Max - Q2) which is 11710.00

(Q2 - Q1) which is 2670.00 < (Q3 - Q2) which is 2935.00

* Histogram Review



***Plot 2.1.15***

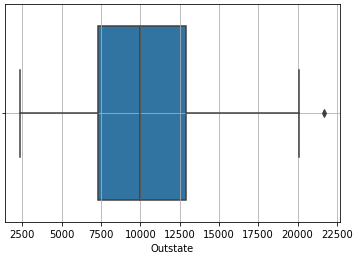
Histogram Review for Edu [‘Outstate’] Shows -

1) 1 Peak

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.16***

BoxPlot Review for Edu [‘Outstate’] Shows -

1) There is Only one outlier

2) Distribution is Right Skewed due to longer Right Whisker, though there are not any significant outliers.

* Empirical Rule

Edu [‘Outstate’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed), But the degree of Asymmetry is not so significant.

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed), But the degree of Asymmetry is not so significant.

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed), But the degree of Asymmetry is not so significant.

* Edu [‘Outstate’] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (‘Outstate’):**

Number of students for whom the particular college or university is Out-of-state tuition, range from 2340.00 to 21700.00

75% of the institutes have less than 12925.00 students for whom the particular college or university is Out-of-state tuition.

From above figure, we can say that the students for whom the particular college or university is Out-of-state tuition institutes is right skewed which means there are More institutes with lower Count of such students

Average Number of students for whom the particular college or university is Out-of-state tuition is 10440.67 which is higher than the median value 9990.00 indicating that the distribution is right tailed. Though the gap is marginal.

### Edu [Room.Board]

* 5 number Summary Suggests that -

Distribution is Right skewed -

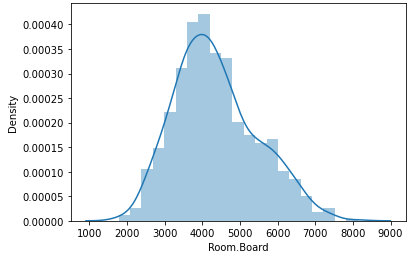
(Q1 - Min) which is 1817.00 < Max - Q3 which is 3074.00

Following points support the symmetry -

(Q2 - Min) which is 2420.00 < (Max - Q2) which is 3924.00

(Q2 - Q1) which is 603.00 < (Q3 - Q2) which is 850.00

* Histogram Review



***Plot 2.1.17***

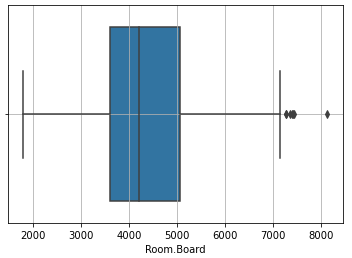
Histogram Review for Edu [‘Room.Board’] Shows -

1) 1 Peak

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.18***

BoxPlot Review for Edu [‘Room.Board’] Shows -

1) There are few outliers

2) Distribution is Right Skewed due to longer Right Whisker, though there are not any significant outliers.

* Empirical Rule

Edu [‘Room.Board’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) ‘2 standard deviation’ from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Edu [‘Room.Board’] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (‘Room.Board’):**

**Insights:**

Cost of Room and board within the listed institutes ranges from 1780.00 to 8124.00

75% of the institutes have less than 5050.00 Cost of Room and board

From above figure, we can say that the Cost of Room and board across institutes is right skewed which means there are More institutes with lower cost for this parameter

Average Cost of Room and Board College or university is 4357.53 which is higher than the median value 4200.00 indicating that the distribution is right tailed. Though, the gap is marginal.

### Edu [Books]

* 5 number Summary Suggests that -

Distribution is Right skewed -

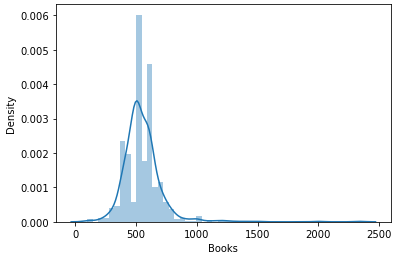
(Q1 - Min) which is 374.00 < Max - Q3 which is 1740.00

Following points support the symmetry -

(Q2 - Min) which is 404.00 < (Max - Q2) which is 1840.00

(Q2 - Q1) which is 30.00 < (Q3 - Q2) which is 100.00

* Histogram Review



***Plot 2.1.19***

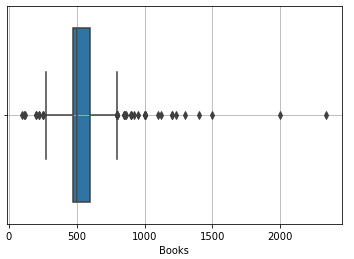
Histogram Review for Edu [‘Books’] Shows -

1) 1 Peak

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.20***

BoxPlot Review for Edu [‘Books’] Shows -

1) There are many outliers

2) Distribution is Right Skewed due to longer Right Whisker.

* Empirical Rule

Edu [‘Books’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) ‘2 standard deviation’ from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Edu [‘Books’] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (‘Books’):**

Cost of Books within the listed institutes ranges from 96.00 to 2340.00

75% of the institutes have less than 600.00 Cost of Books

From above figure, we can say that the Cost of Books across institutes is right skewed which means there are More institutes with lower / reasonable Cost of Books

Average Cost of Room and Board College or university is 549.38 which is slightly higher than the median value 500.00 indicating that the distribution is right tailed, though the gap is marginal.

### Edu [Personal]

* 5 number Summary Suggests that -

Distribution is Right skewed -

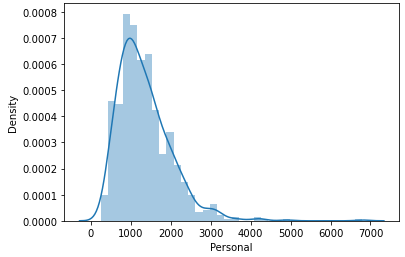
(Q1 - Min) which is 600.00 < Max - Q3 which is 5100.00

Following points support the symmetry -

(Q2 - Min) which is 950.00 < (Max - Q2) which is 5600.00

(Q2 - Q1) which is 350.00 < (Q3 - Q2) which is 500.00

* Histogram Review



***Plot 2.1.21***

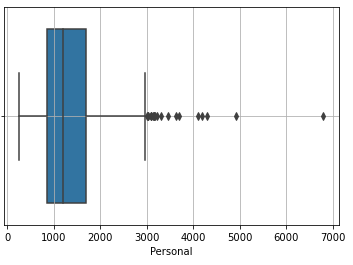
Histogram Review for Edu [‘Personal’] Shows -

1) 1 Peak

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.21***

BoxPlot Review for Edu [‘Personal’] Shows -

1) There are many outliers

2) Distribution is Right Skewed due to longer Right Whisker.

* Empirical Rule

Edu [‘Personal’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) ‘2 standard deviation’ from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Edu [‘Personal’] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (‘Personal’):**

Estimated personal spending for a student within the listed institutes ranges from 250.00 to 6800.00

75% of the students spend less than 1700.00 on personal needs

From above figure, we can say that the personal spend across institutes is right skewed which means there are More institutes where students can study by spending less on personal needs.

Average personal spend is 1340.64 which is slightly higher than the median value 1200.00 indicating that the distribution is right tailed, though the gap is marginal.

### Edu [PhD]

* 5 number Summary Suggests that -

Distribution is Left skewed -

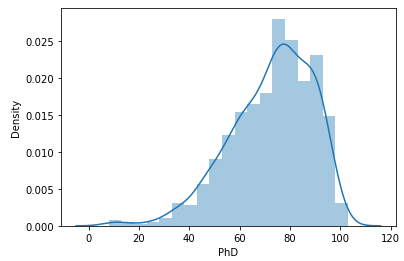
(Q1 - Min) which is 54.00 > Max - Q3 which is 18.00

Following points support the symmetry -

(Q2 - Min) which is 67.00 > (Max - Q2) which is 28.00

(Q2 - Q1) which is 13.00 > (Q3 - Q2) which is 100.00

* Histogram Review



***Plot 2.1.23***

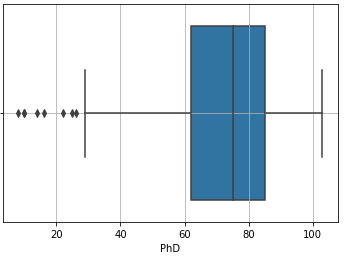
Histogram Review for Edu [‘PhD’] Shows -

1) 1 Peak

2) Left Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.24***

BoxPlot Review for Edu [‘PhD’] Shows -

1) There are many outliers lower than Q1

2) Distribution is Left Skewed due to longer Left Whisker.

* Empirical Rule

Edu [‘PhD’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Left Skewed)

2) ‘2 standard deviation’ from Mean is asymmetrical (Left Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Left Skewed)

* Edu [‘PhD’] review of Mean, Median and Mode suggests that –

In this case Mean < Median < Mode hence the distribution is Left Skewed

**Insights (‘PhD’):**

Number of Faculty members with PhD qualification ranges from 8.00 to 103.00 across the institutes

75% of the institutes / universities have less 85.00 Faculty members with PhD qualification

From above figure, we can say that the Faculty members with PhD qualification across institutes is left skewed which means there are More institutes with Faculty members with PhD qualification. there are very few without adequate PhD qualifies

Average count of Faculty members with PhD qualification is 72.66 which is slightly lesser than the median value 75.00 indicating that the distribution is left tailed, though the gap is marginal.

### Edu [Terminal]

* 5 number Summary Suggests that -

Distribution is Left skewed -

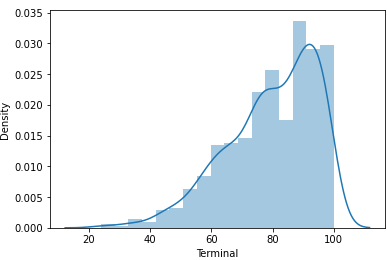
(Q1 - Min) which is 47.00 > Max - Q3 which is 8.00

Following points support the symmetry -

(Q2 - Min) which is 58.00 > (Max - Q2) which is 18.00

(Q2 - Q1) which is 11.00 > (Q3 - Q2) which is 10.00

* Histogram Review



***Plot 2.1.25***

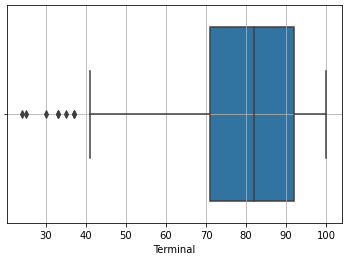
Histogram Review for Edu [‘Terminal’] Shows -

1) 1 Peak

2) Left Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.26***

BoxPlot Review for Edu [‘Terminal’] Shows -

1) There are many outliers lower than Q1

2) Distribution is Left Skewed due to longer Left Whisker.

* Empirical Rule

Edu [‘Terminal’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Left Skewed)

2) ‘2 standard deviation’ from Mean is asymmetrical (Left Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Left Skewed)

* Edu [‘Terminal’] review of Mean, Median and Mode suggests that –

In this case Mean < Median < Mode hence the distribution is Left Skewed

**Insights (‘Terminal’):**

Number of Faculty members with Terminal Degree ranges from 24.00 to 100.00 across the institutes

75% of the institutes / universities have less 92.00 Faculty members with Terminal Degree

From above figure, we can say that the Faculty members with Terminal Degree across institutes is left skewed which means there are institutes with more Faculty members with Terminal Degree, there are very few without Terminal Degree

Average count of Faculty members with Terminal qualification is 79.70 which is slightly lesser than the median value 82.00 indicating that the distribution is left tailed, though the gap is marginal.

### Edu [S.F.Ratio]

* 5 number Summary Suggests that -

Distribution is Right skewed -

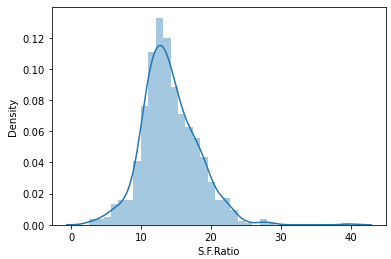
(Q1 - Min) which is 9.00 < Max - Q3 which is 23.30

Following points support the symmetry -

(Q2 - Min) which is 11.10< (Max - Q2) which is 26.20

(Q2 - Q1) which is 2.10 < (Q3 - Q2) which is 2.90

* Histogram Review



***Plot 2.1.27***

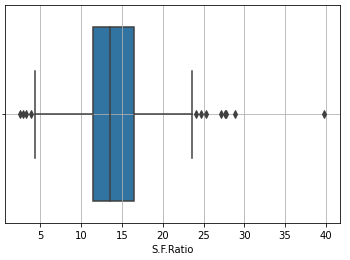
Histogram Review for Edu [‘S.F.Ratio’] Shows -

1) 1 Peak

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.28***

BoxPlot Review for Edu [‘S.F.Ratio’] Shows -

1) There are many outliers on both left as well as right side

2) Distribution is Little Right Skewed due to longer Right Whisker, though not so significant

* Empirical Rule

Edu [‘S.F.Ratio’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) ‘2 standard deviation’ from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Edu [‘S.F.Ratio’] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (‘S.F.Ratio’):**

Student/faculty ratio within the listed institutes ranges from 2.50 to 39.80

75% of the institutes have less than 16.50 Students /faculty member

From above figure, we can say that the Student/faculty ratio across institutes is right skewed which means there are More institutes where students are paid much attention as the number of Students / Faculty member is less.

Average Student/faculty ratio is 14.09 which are slightly higher than the median value 13.60 indicating that the distribution is right tailed, though the gap is marginal.

### Edu [perc.alumni]

* 5 number Summary Suggests that -

Distribution is Right skewed -

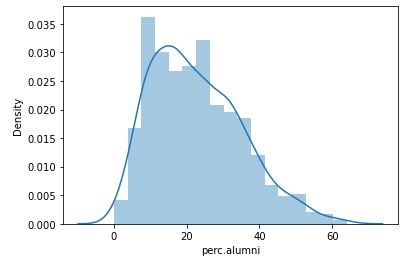
(Q1 - Min) which is 13.00 < Max - Q3 which is 33.00

Following points support the symmetry -

(Q2 - Min) which is 21.00< (Max - Q2) which is 43.00

(Q2 - Q1) which is 8.00 < (Q3 - Q2) which is 10.00

* Histogram Review



***Plot 2.1.29***

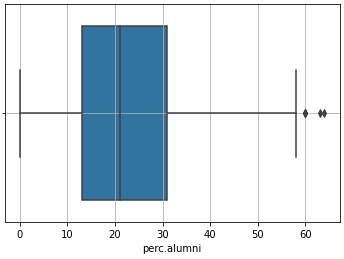
Histogram Review for Edu [‘perc.alumni’] Shows -

1) 1 Peak

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.30***

BoxPlot Review for Edu [‘perc.alumni’] Shows -

1) There are Few outliers on as right side

2) Distribution is RightSkewed due to longer Right Whisker.

* Empirical Rule

Edu [‘perc.alumni’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) ‘2 standard deviation’ from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Edu [‘perc.alumni’] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (‘perc.alumni’):**

Percentage of alumni who donate within the listed institutes ranges from 0.00 to 64.00

75% of the institutes receive donation from less than 31.00% of their Alums

From above figure, we can say that the Percentage of alumni who donate across institutes is right skewed which means there are less institutes where Percentage of alumni who donate is more.

Average Percentage of alumni who donate is 22.74 which are slightly higher than the median value 21.00 indicating that the distribution is right tailed, though the gap is marginal.

### Edu [Expend]

* 5 number Summary Suggests that -

Distribution is Right skewed -

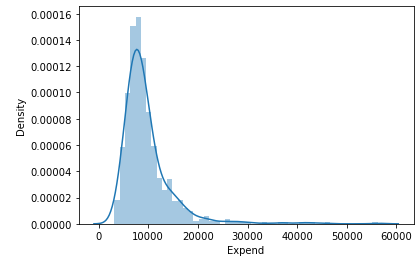
(Q1 - Min) which is 3565.00 < Max - Q3 which is 45403.00

Following points support the symmetry -

(Q2 - Min) which is 5191.00< (Max - Q2) which is 47856.00

(Q2 - Q1) which is 1626.00 < (Q3 - Q2) which is 2453.00

* Histogram Review



***Plot 2.1.31***

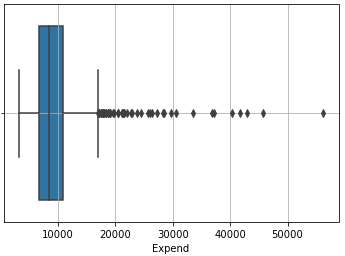
Histogram Review for Edu [‘Expend’] Shows -

1) 1 Peak

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.32***

BoxPlot Review for Edu [‘Expend’] Shows -

1) There are many outliers on as right side

2) Distribution is Right Skewed due to longer Right Whisker.

* Empirical Rule

Edu [‘Expend’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) ‘2 standard deviation’ from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Edu [‘Expend’] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (‘Expend’):**

The Instructional expenditure per student ranges from 3186.00 to 56233.00

75% of the institutes spend less than 10830.00 as Instructional expenditure per student

From above figure, we can say that The Instructional expenditure per student is right skewed which means there are very few institutes where The Instructional expenditure per student is more.

Average Percentage of The Instructional expenditure per student is 9660.17 which is slightly higher than the median value 8377.00 indicating that the distribution is right tailed, though the gap is marginal.

### Edu [Grad.Rate]

* 5 number Summary Suggests that -

Distribution is Right skewed -

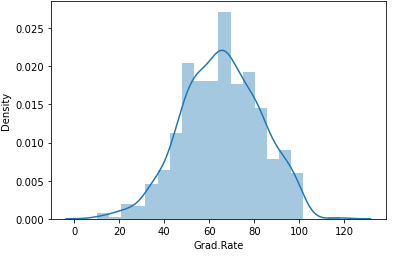
(Q1 - Min) which is 43.00 < Max - Q3 which is 40.00

Following points support the symmetry -

(Q2 - Min) which is 55.00< (Max - Q2) which is 53.00

(Q2 - Q1) which is 12.00 < (Q3 - Q2) which is 13.00

* Histogram Review



***Plot 2.1.33***

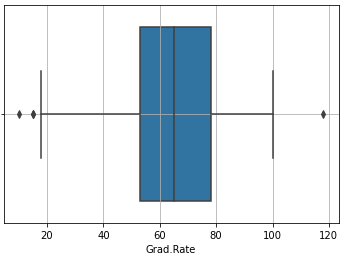
Histogram Review for Edu [‘Grad.Rate’] Shows -

1) 1 Peak

2) A very little Right Skewness almost normal

3) It does Show a near perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.34***

BoxPlot Review for Edu [‘Grad.Rate’] Shows -

1) There is handful of outliers on each side

2) Distribution appears little Left Skewed due to longer left Whisker however outliers on both the sides seem to centralize the distribution making it close to Normal.

* Empirical Rule

Edu [‘Grad.Rate’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is near Normal.

2) ‘2 standard deviation’ from Mean is near Normal.

3) '3 standard deviation' from Mean is near Normal.

* Edu [‘Grad.Rate’] review of Mean, Median and Mode suggests that –

In this case Mean = Median < Mode hence the distribution is Near Normal

**Insights (‘Grad.Rate’):**

The Graduation rate across institutes / Universities ranges from 10.00% to 118.00

75% of the institutes have graduation rate less than 78.00 % as Instructional expenditure per student

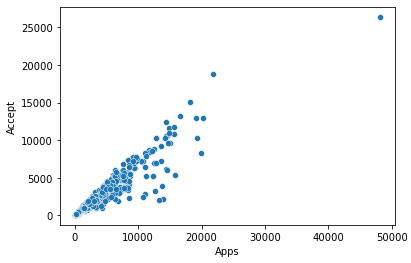
From above figure, we can say that The Graduation rate per institutes / Universities is normally distributed

Average The Graduation rate across institutes / Universities is 65.46 which are almost equal to the median value 65.00 indicating that the distribution is near normal.

***Bivariate Analysis***

Please find BiVariate Analysis performed on Few Pairs to demonstrate the approach as well as illustrate typical insights that can be drawn. Please note we only have continuous numerical fields as part of the supplied DataSet.

### Edu [Apps Vs Accept]

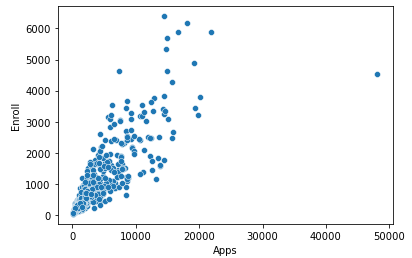


**INSIGHT**

Ratio of Accepted Applications is fairly linear, towards the Mean

***Plot 2.1.35***

### Edu [Apps Vs Enroll]

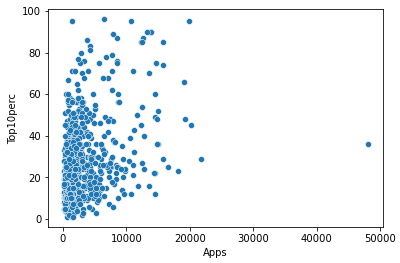


**INSIGHT**

Universities / Institutes that have received less application have enrolled fewer students

***Plot 2.1.36***

### Edu [Apps Vs Top10perc]

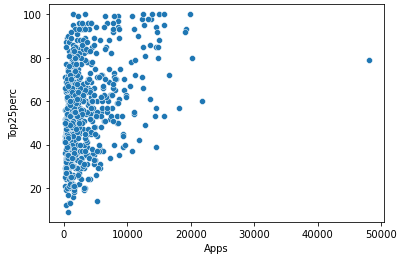


**INSIGHT**

Colleges / Institutes that have received less application have more Percentage of new students from top 10% of Higher Secondary class

***Plot 2.1.37***

### Edu [Apps Vs Top25perc]

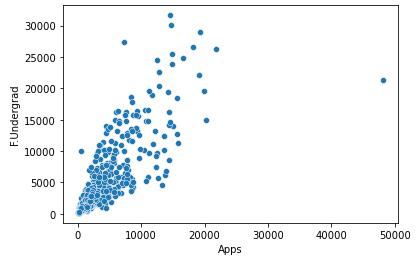


**INSIGHT**

Colleges / Institutes that have received less application have more Percentage of new students from top 25% of Higher Secondary class

***Plot 2.1.38***

### Edu [Apps Vs F.Undergrad]

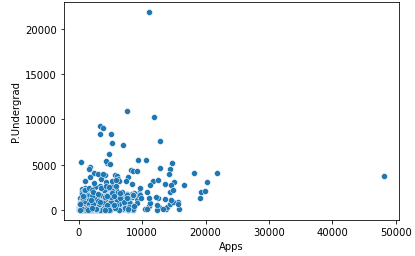


**INSIGHT**

Colleges / Institutes that have received less application have less Number of full-time undergraduate students

***Plot 2.1.39***

### Edu [Apps Vs P.Undergrad]

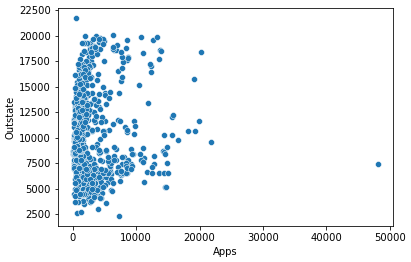


**INSIGHT**

Colleges / Institutes that have received less application have less Number of Part-time undergraduate students

***Plot 2.1.40***

### Edu [Apps Vs Outstate]

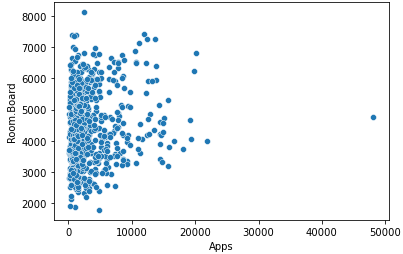


**INSIGHT**

Colleges / Institutes that have received less application have more Number of students for whom the particular college or university is Out-of-state tuition

***Plot 2.1.41***

### Edu [Apps Vs Room.Board]

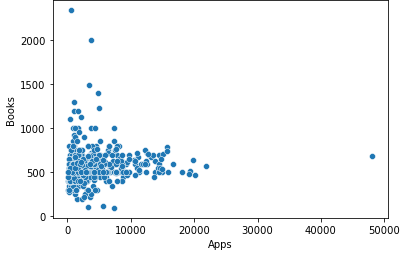


**INSIGHT**

Colleges / Institutes that have received less application have more Cost of Room and board

***Plot 2.1.42***

### Edu [Apps Vs Books]

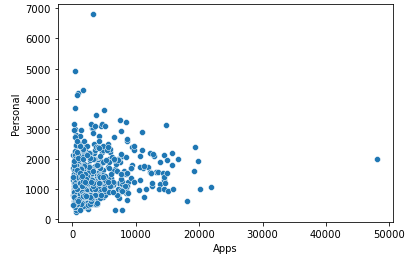


**INSIGHT**

Cost of Books is generally same across the Universities/ Institutes

***Plot 2.1.43***

### Edu [Apps Vs Personal]

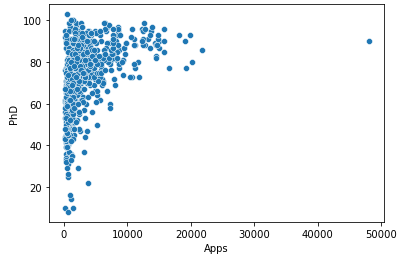


**INSIGHT**

Cost of Personal expenses is generally same across the Universities/ Institutes

***Plot 2.1.44***

### Edu [Apps Vs PhD]

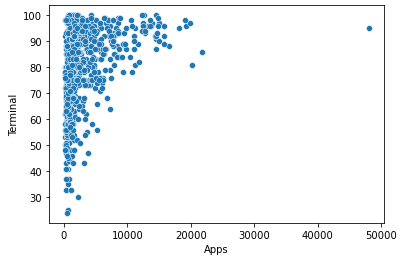


**INSIGHT**

There is Good Number of Faculty Members with PhD qualification across the Universities/ Institutes

***Plot 2.1.45***

### Edu [Apps Vs Terminal]

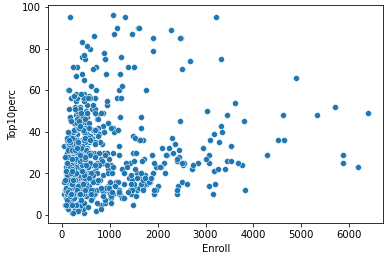


**INSIGHT**

There is Good Number of Faculty Members with Terminal qualification across the Universities/ Institutes

***Plot 2.1.46***

### Edu [Enroll Vs Top10perc]

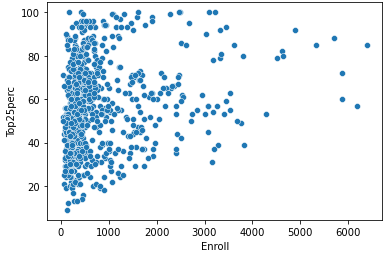


**INSIGHT**

Percentage of ENROLLMENT of new students from top 10% of Higher Secondary class is more within the Universities that offered less Enrolment.

***Plot 2.1.47***

### Edu [Enroll Vs Top25perc]

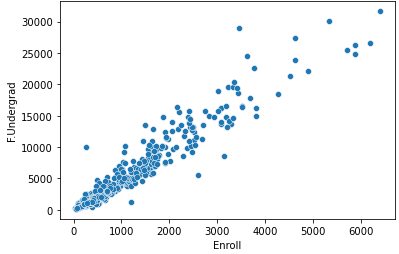


**INSIGHT**

Percentage of ENROLLMENT of new students from top 25% of Higher Secondary class is more within the Universities that offered less Enrolment.

***Plot 2.1.48***

### Edu [Enroll Vs F.Undergrad]

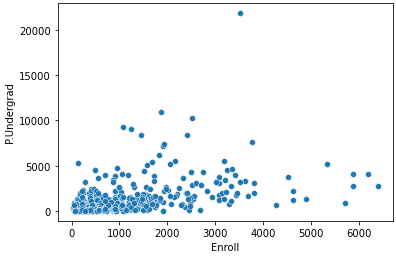


**INSIGHT**

Number of full-time undergraduate students is proportionate to the Enrollments across universities i.e. linear

***Plot 2.1.49***

### Edu [Enroll Vs P.Undergrad]

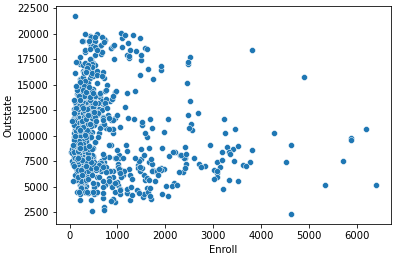


**INSIGHT**

Number of part-time undergraduate students is also less in the universities Enrolling less students and the number is in the similar range too.

***Plot 2.1.50***

### Edu [Enroll Vs Outstate]

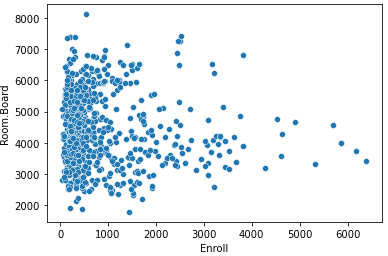


**INSIGHT**

Number of students for whom the particular college or university is Out-of-state tuition is evenly spread an within the college or university that offer Less Enrolments however as the enrolments increase we the spread turns uneven.

***Plot 2.1.51***

### Edu [Enroll Vs Room.Board]

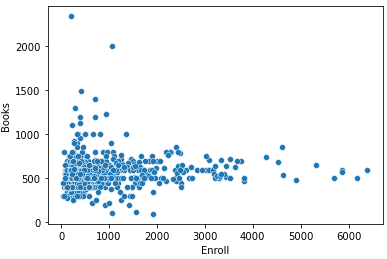


**INSIGHT**

Cost of Room and board is Evenly spread an within the college or university that offer Less Enrolments however as the enrolments increase we see this cost spread turns uneven.

***Plot 2.1.52***

### Edu [Books]

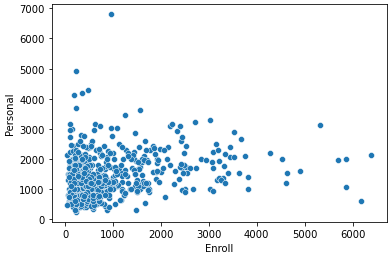


**INSIGHT**

Cost of Books is generally within 250-750 across irrespective of number of Enrolments

***Plot 2.1.53***

### Edu [Personal]

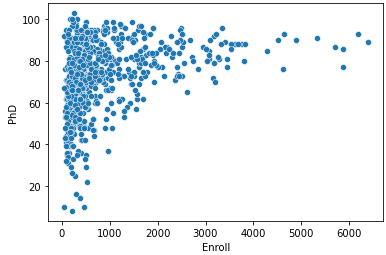


**INSIGHT**

Estimated personal spending for a student is generally within 250-2000across irrespective of number of Enrolments which means this Estimate is apparently location based.

***Plot 2.1.54***

### Edu [PhD]

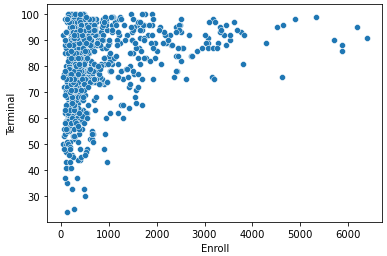


**INSIGHT**

Percentage of faculties with Ph.D.’s is generally higher irrespective of Enrolments.

***Plot 2.1.55***

### Edu [Terminal]



**INSIGHT**

Percentage of faculties with Terminal Degrees is generally higher irrespective of Enrolments.

***Plot 2.1.56***

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

PCA projects calculate a new projection based on variance/Variability (SD Deviation) of the variable in consideration. This means variables with higher (SD Deviation) will have higher weightage. In the normalized data where we have 'Zero' as mean and 'One' as Std deviation we will have same weight magnitude and the PCA calculates relevant axis to consider.

Secondly the unit of measurement of the variables includes (Numbers, Currency, Percentage, Ratio) getting a lot of difference within their magnitude

To Summarize, Reason that we must scale the data is

* In the un-scaled data standard deviation varies from 4 to 5000.
* In the un-scaled data Mean varies from 14 to 10000.
* Variable and their varied Measurement of Units

|  |  |
| --- | --- |
| **Variable** | **Unit Of Measurement** |
| Apps | Number |
| Accept | Number |
| Enroll | Number |
| Top10perc | Percentage |
| Top25perc | Percentage |
| F.Undergrad | Number |
| P.Undergrad | Number |
| Outstate | Number |
| Room.Board | Cost |
| Books | Cost |
| Personal | Cost |
| PhD | Percentage |
| Terminal | Percentage |
| S.F.Ratio | Ratio |
| perc.alumni | Percentage |
| Expend | Cost |
| Grad.Rate | Percentage |

***Table 2.2.1***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Pre Scaling** | | **Post Scaling** | |
|  | **mean** | **std** | **mean** | **std** |
| **Apps** | 3001.64 | 3870.2 | 0 | 1 |
| **Accept** | 2018.8 | 2451.11 | 0 | 1 |
| **Enroll** | 779.97 | 929.18 | 0 | 1 |
| **Top10perc** | 27.56 | 17.64 | 0 | 1 |
| **Top25perc** | 55.8 | 19.8 | 0 | 1 |
| **F.Undergrad** | 3699.91 | 4850.42 | 0 | 1 |
| **P.Undergrad** | 855.3 | 1522.43 | 0 | 1 |
| **Outstate** | 10440.67 | 4023.02 | 0 | 1 |
| **Room.Board** | 4357.53 | 1096.7 | 0 | 1 |
| **Books** | 549.38 | 165.11 | 0 | 1 |
| **Personal** | 1340.64 | 677.07 | 0 | 1 |
| **PhD** | 72.66 | 16.33 | 0 | 1 |
| **Terminal** | 79.7 | 14.72 | 0 | 1 |
| **S.F.Ratio** | 14.09 | 3.96 | 0 | 1 |
| **perc.alumni** | 22.74 | 12.39 | 0 | 1 |
| **Expend** | 9660.17 | 5221.77 | 0 | 1 |
| **Grad.Rate** | 65.46 | 17.18 | 0 | 1 |

***Table 2.2.2***

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

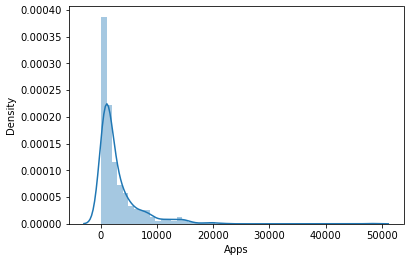
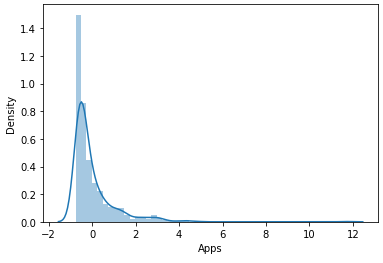
Covariance Matrix shows how two variables vary together, if they are directly or inversely proportionate. It doesn’t show the dependency of one variable on other. Since this exercise is done on the un-scaled data it doesn't tell us anything about magnitude. Covariance with the same variable varies between 15 to as high as 23526579.

Correlation Matrix shows how two variables vary together, if they are directly or inversely proportionate as well as strength of their relationship. Since this exercise is done on the Scaled data magnitude becomes insignificant. Correlation with the same variable is always 1.

Negative Co-variance remains negative even within the correlation matrix so scaling doesn’t change the direction but it does tell us how strong the relationship is.

## 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]

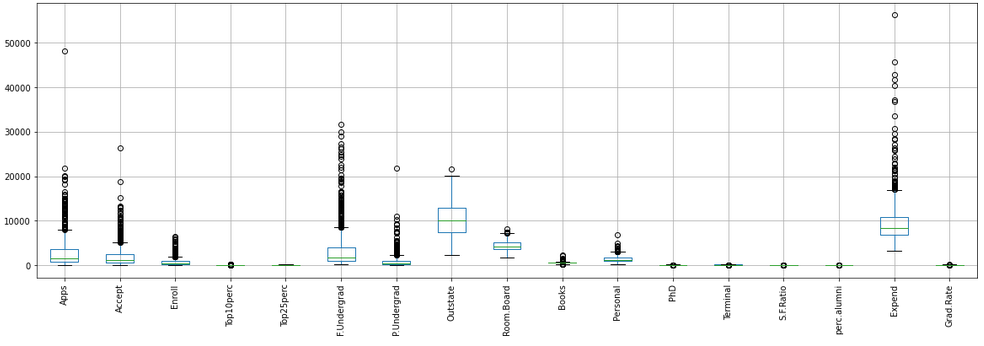
Histogram plot of one picked variable (Apps) which had Outliers clearly shows that there is no difference in the presence of outliers as well as the Right Skewed behaviour of Data. Please refer Plot 2.4.1

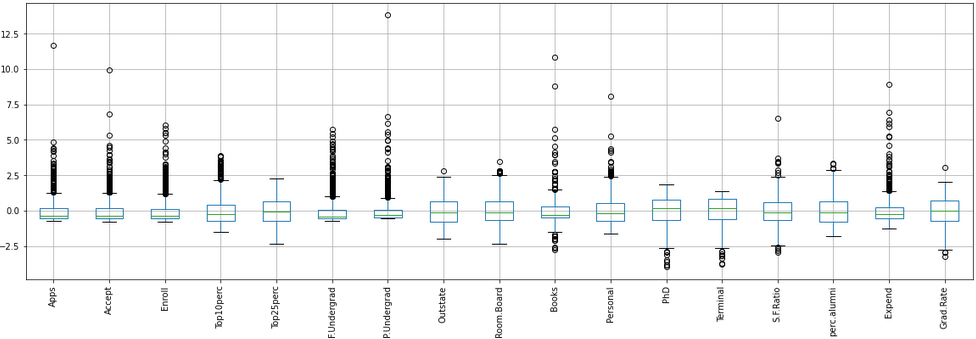
***Plot 2.4.1***

Further review of Box plot of data before and after scaling also confirms that the presense of Outliers irrespective of SCALING Please refer Plot 2.4.2 and Plot 2.4.3

Though the number of outliers remain the same they are scaled in such a way that they are brought close to the Q3 or Q1 mark. Thus reducing the impact.



***Plot 2.4.2***



***Plot 2.4.3***

Further Analysis shows following results confirming no change in the count of Outliers, Please refer Table 2.2.2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Before Scaling | | Before Scaling | |
| **Variable** | **Outlier Count** | **Outlier %** | **Outlier Count** | **Outlier %** |
| Apps | 70 | 9.01 | 70 | 9.01 |
| Accept | 73 | 9.4 | 73 | 9.4 |
| Enroll | 79 | 10.17 | 79 | 10.17 |
| Top10perc | 39 | 5.02 | 39 | 5.02 |
| Top25perc | 0 | 0 | 0 | 0 |
| F.Undergrad | 97 | 12.48 | 97 | 12.48 |
| P.Undergrad | 67 | 8.62 | 67 | 8.62 |
| Outstate | 1 | 0.13 | 1 | 0.13 |
| Room.Board | 7 | 0.9 | 7 | 0.9 |
| Books | 48 | 6.18 | 48 | 6.18 |
| Personal | 20 | 2.57 | 20 | 2.57 |
| PhD | 8 | 1.03 | 8 | 1.03 |
| Terminal | 8 | 1.03 | 8 | 1.03 |
| S.F.Ratio | 12 | 1.54 | 12 | 1.54 |
| perc.alumni | 5 | 0.64 | 5 | 0.64 |
| Expend | 48 | 6.18 | 48 | 6.18 |
| Grad.Rate | 4 | 0.51 | 4 | 0.51 |

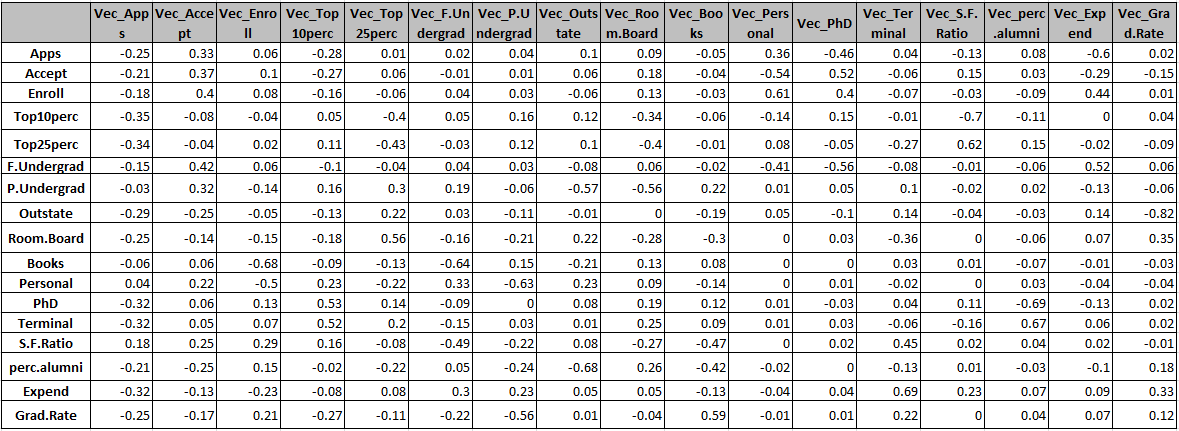
***Table 2.4.1***

## Extract the eigenvalues and eigenvectors.[print both]

Please refer to tables 2.5.1 and 2.5.2 respectively for the Extracted Eigen Values and Eigen vectors for the scaled version of supplied Data Set.

|  |  |
| --- | --- |
|  | **Eigen Values** |
| **Apps** | 5.45 |
| **Accept** | 4.48 |
| **Enroll** | 1.17 |
| **Top10perc** | 1.01 |
| **Top25perc** | 0.93 |
| **F.Undergrad** | 0.85 |
| **P.Undergrad** | 0.61 |
| **Outstate** | 0.59 |
| **Room.Board** | 0.53 |
| **Books** | 0.4 |
| **Personal** | 0.02 |
| **PhD** | 0.04 |
| **Terminal** | 0.31 |
| **S.F.Ratio** | 0.09 |
| **perc.alumni** | 0.14 |
| **Expend** | 0.17 |
| **Grad.Rate** | 0.22 |

***Table 2.5.1***



***Table 2.5.2***

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

* Generated 9 PCA dimensions (dimensionality reduction from 17 to 9) as part of PCA exercise.
* We get a Data Frame with the reduced number of columns
* Following explained Variance ratios are covered by these 9 components, Please refer Table 2.6.1

|  |  |
| --- | --- |
| **Principal Component** | **Var\_Ratio** |
| **Component-1** | 0.32 |
| **Component-2** | 0.26 |
| **Component-3** | 0.07 |
| **Component-4** | 0.06 |
| **Component-5** | 0.05 |
| **Component-6** | 0.05 |
| **Component-7** | 0.04 |
| **Component-8** | 0.03 |
| **Component-9** | 0.03 |

***Table 2.6.1***

* Loading of Features is obtained using Python Function across 9 components, Please refer Table 2.6.2
* Identify which features have maximum loading across the components.
* Highlighted Features are the one having maximum loading on the respective component. These marked features are considered to decide the context that the component represents; these features are then used as revised column names.
* In this case, table header of Table 2.6.2 has the most significant features

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Top10perc** | **F.Undergrad** | **Books** | **Apps** | **Room.Board** | **S.F.Ratio** | **Personal** | **perc.alumni** | **P.Undergrad** |
|  | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| **Apps** | 0.25 | 0.33 | -0.06 | 0.28 | 0.01 | -0.02 | -0.04 | -0.1 | -0.09 |
| **Accept** | 0.21 | 0.37 | -0.1 | 0.27 | 0.06 | 0.01 | -0.01 | -0.06 | -0.18 |
| **Enroll** | 0.18 | 0.4 | -0.08 | 0.16 | -0.06 | -0.04 | -0.03 | 0.06 | -0.13 |
| **Top10perc** | 0.35 | -0.08 | 0.04 | -0.05 | -0.4 | -0.05 | -0.16 | -0.12 | 0.34 |
| **Top25perc** | 0.34 | -0.04 | -0.02 | -0.11 | -0.43 | 0.03 | -0.12 | -0.1 | 0.4 |
| **F.Undergrad** | 0.15 | 0.42 | -0.06 | 0.1 | -0.04 | -0.04 | -0.03 | 0.08 | -0.06 |
| **P.Undergrad** | 0.03 | 0.32 | 0.14 | -0.16 | 0.3 | -0.19 | 0.06 | 0.57 | 0.56 |
| **Outstate** | 0.29 | -0.25 | 0.05 | 0.13 | 0.22 | -0.03 | 0.11 | 0.01 | 0 |
| **Room.Board** | 0.25 | -0.14 | 0.15 | 0.18 | 0.56 | 0.16 | 0.21 | -0.22 | 0.28 |
| **Books** | 0.06 | 0.06 | 0.68 | 0.09 | -0.13 | 0.64 | -0.15 | 0.21 | -0.13 |
| **Personal** | -0.04 | 0.22 | 0.5 | -0.23 | -0.22 | -0.33 | 0.63 | -0.23 | -0.09 |
| **PhD** | 0.32 | 0.06 | -0.13 | -0.53 | 0.14 | 0.09 | 0 | -0.08 | -0.19 |
| **Terminal** | 0.32 | 0.05 | -0.07 | -0.52 | 0.2 | 0.15 | -0.03 | -0.01 | -0.25 |
| **S.F.Ratio** | -0.18 | 0.25 | -0.29 | -0.16 | -0.08 | 0.49 | 0.22 | -0.08 | 0.27 |
| **perc.alumni** | 0.21 | -0.25 | -0.15 | 0.02 | -0.22 | -0.05 | 0.24 | 0.68 | -0.26 |
| **Expend** | 0.32 | -0.13 | 0.23 | 0.08 | 0.08 | -0.3 | -0.23 | -0.05 | -0.05 |
| **Grad.Rate** | 0.25 | -0.17 | -0.21 | 0.27 | -0.11 | 0.22 | 0.56 | -0.01 | 0.04 |

***Table 2.6.2***

* Please refer to Data Frame ‘Edu\_new’ which is the revised Data Frame after performing PCA

## 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]

* First Component Weightages to apply are taken using Python PCA function, Please refer to Table 2.7.1

|  |  |
| --- | --- |
| **Feature No.** | **Weightage** |
| **0** | 0.25 |
| **1** | 0.21 |
| **2** | 0.18 |
| **3** | 0.35 |
| **4** | 0.34 |
| **5** | 0.15 |
| **6** | 0.03 |
| **7** | 0.29 |
| **8** | 0.25 |

***Table 2.7.1***

* Significant Features that Weightage need to be applied are the columns from the Revised DataFrame, Please refer to Table 2.7.2

|  |  |
| --- | --- |
| **Feature No.** | **Feature Name** |
| **0** | pc\_Top10perc |
| **1** | pc\_F.Undergrad |
| **2** | pc\_Books |
| **3** | pc\_Apps |
| **4** | pc\_Room.Board |
| **5** | pc\_S.F.Ratio |
| **6** | pc\_Personal |
| **7** | pc\_perc.alumni |
| **8** | pc\_P.Undergrad |

***Table 2.7.2***

* Linear Equation for 1st Component will be sum of columns with Weightage applied, Please refer to tables 2.7.1 and 2.7.2

0.25 \* pc\_Top10perc + 0.21 \* pc\_F.Undergrad + 0.18 \* pc\_Books + 0.35 \* pc\_Apps + 0.34 \* pc\_Room.Board + 0.15 \* pc\_S.F.Ratio + 0.03 \* pc\_Personal + 0.29 \* pc\_perc.alumni + 0.25 \* pc\_P.Undergrad +

## 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 value of eigenvalues tells the variance and the correlation captured till a specific component.
* We have already seen how the weightages are reviewed in the HeatMap.
* This gives us fair idea as to which if significant components have had a higher weightage as part of the Heat map.
* They also tell us the features that are Vital for the further projections or analysis
* Please refer to Table 2.8.1 for Cumulative % of explained Variance

|  |  |
| --- | --- |
| **Principal components** | **Cumulative Variance Explained** |
| **0** | 32.02 |
| **1** | 58.36 |
| **2** | 65.26 |
| **3** | 71.18 |
| **4** | 76.67 |
| **5** | 81.66 |
| **6** | 85.22 |
| **7** | 88.67 |
| **8** | 91.79 |
| **9** | 94.16 |
| **10** | 96 |
| **11** | 97.3 |
| **12** | 98.29 |
| **13** | 99.13 |
| **14** | 99.65 |
| **15** | 99.86 |
| **16** | 100 |

***Table 2.8.1***

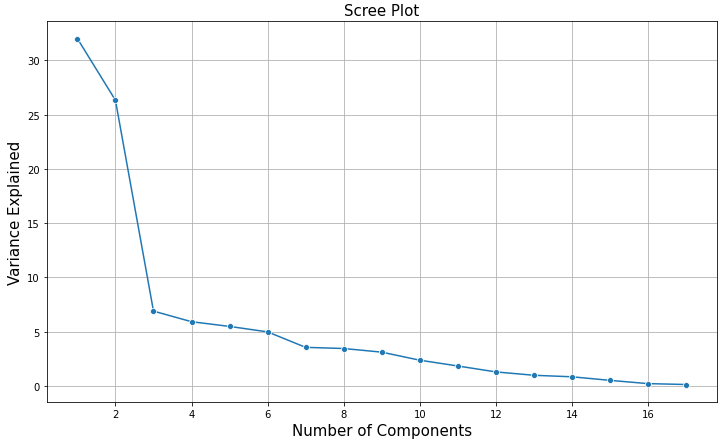
* As we see these is done, degree of Trade off that have to live with as we reduce the data by applying Principal Component approach. It varies from case to case on what is optimum % that we should consider in order to be able to comfortably complete the Analysis post PC optimization.
* It is equally important that we avoid any Principal Component that has less than ‘1%’ variance explained individually.
* Please refer to Table 2.8.2 where PC12-PC16 is listed covering less than 1% variance.
* In the current situation we can go with 9 PCs that cumulatively offer us 92% variability covered.

|  |  |
| --- | --- |
| **Principal components** | **Variance Explained** |
| **0** | 32.02 |
| **1** | 26.34 |
| **2** | 6.9 |
| **3** | 5.92 |
| **4** | 5.49 |
| **5** | 4.98 |
| **6** | 3.56 |
| **7** | 3.45 |
| **8** | 3.12 |
| **9** | 2.38 |
| **10** | 1.84 |
| **11** | 1.3 |
| **12** | 0.99 |
| **13** | 0.85 |
| **14** | 0.52 |
| **15** | 0.22 |
| **16** | 0.14 |

***Table 2.8.2***

Visual Technique to check the components that can be considered is also explained by SCREE PLOT.

Which graphically tells us the Components and variance, Please refer to



## 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]

* Weightage Range Covered through the selection is -0.53 and .68 for all the 17 variables ( Refer:Table 2.6.2)
* 92% of Cumulative Variance Explained within the 9 out of 17 PCs. ( Refer:Table 2.8.1)
* None of the components selected have less than 1 variance explained ( Refer:Table 2.8.2)

Following variables have the most weightage within each of the selected components; Their business significance as mentioned will help with further Analysis by the business will help address the decision making, their business significance is as follows-

* Top10perc - Reflects on the Quality of Intake
* F.Undergrad - Number of Positions in Fulltime UG
* Books - Cost of Books
* Apps - Number of Applications received is critical parameter to evaluate the College/ University
* Room.Board - cost of Room / Board
* S.F.Ratio - An important parameter that Applicant check before applying
* Personal - An important parameter that Applicant consider before applying
* perc.alumni - This reflects on the past performance of the College/ University and it helps reduce cost of operation and thus to the students, An important parameter that Applicant consider before applying
* P.Undergrad - Number of Positions in part time UG

# Appendices

## List of Plots

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Plot Details** | | | | |
| Plot 1.5.1 | Anova 1.5 |  | Plot 2.1.30 | perc.alumni Variable BoxPlot |
| Plot 2.1.1 | Apps Variable Histogram |  | Plot 2.1.31 | Expend Variable Histogram |
| Plot 2.1.2 | Apps Variable BoxPlot |  | Plot 2.1.32 | Expend Variable BoxPlot |
| Plot 2.1.3 | Accept Variable Histogram |  | Plot 2.1.33 | Grad.Rate Variable Histogram |
| Plot 2.1.4 | Accept Variable BoxPlot |  | Plot 2.1.34 | Grad.Rate Variable BoxPlot |
| Plot 2.1.5 | Enroll Variable Histogram |  | Plot 2.1.35 | ScatterPLot Apps Vs Accept |
| Plot 2.1.6 | Enroll Variable BoxPlot |  | Plot 2.1.36 | ScatterPLot Apps Vs Enroll |
| Plot 2.1.7 | Top10perc Variable Histogram |  | Plot 2.1.37 | ScatterPLot Apps Vs Top10perc |
| Plot 2.1.8 | Top10perc Variable BoxPlot |  | Plot 2.1.38 | ScatterPLot Apps Vs Top25perc |
| Plot 2.1.9 | Top25perc Variable Histogram |  | Plot 2.1.39 | ScatterPLot Apps Vs F.Undergrad |
| Plot 2.1.10 | Top25perc Variable BoxPlot |  | Plot 2.1.40 | ScatterPLot Apps Vs P.Undergrad |
| Plot 2.1.11 | F.Undergrad Variable Histogram |  | Plot 2.1.41 | ScatterPLot Apps Vs Outstate |
| Plot 2.1.12 | F.Undergrad Variable BoxPlot |  | Plot 2.1.42 | ScatterPLot Apps Vs Room.Board |
| Plot 2.1.13 | P.Undergrad Variable Histogram |  | Plot 2.1.43 | ScatterPLot Apps Vs Books |
| Plot 2.1.14 | P.Undergrad Variable BoxPlot |  | Plot 2.1.44 | ScatterPLot Apps Vs Personal |
| Plot 2.1.15 | Outstate Variable Histogram |  | Plot 2.1.45 | ScatterPLot Apps Vs PhD |
| Plot 2.1.16 | Outstate Variable BoxPlot |  | Plot 2.1.46 | ScatterPLot Apps Vs Terminal |
| Plot 2.1.17 | Room.Board Variable Histogram |  | Plot 2.1.47 | ScatterPLot Enroll Vs Top10perc |
| Plot 2.1.18 | Room.Board Variable BoxPlot |  | Plot 2.1.48 | ScatterPLot Enroll Vs Top25perc |
| Plot 2.1.19 | Books Variable Histogram |  | Plot 2.1.49 | ScatterPLot Enroll Vs F.Undergrad |
| Plot 2.1.20 | Books Variable BoxPlot |  | Plot 2.1.50 | ScatterPLot Enroll Vs P.Undergrad |
| Plot 2.1.21 | Personal Variable Histogram |  | Plot 2.1.51 | ScatterPLot Enroll Vs Outstate |
| Plot 2.1.22 | PersonalVariable BoxPlot |  | Plot 2.1.52 | ScatterPLot Enroll Vs Room.Board |
| Plot 2.1.23 | PhD Variable Histogram |  | Plot 2.1.53 | ScatterPLot Enroll Vs Books |
| Plot 2.1.24 | PhD Variable BoxPlot |  | Plot 2.1.54 | ScatterPLot Enroll Vs Personal |
| Plot 2.1.25 | Terminal Variable Histogram |  | Plot 2.1.55 | ScatterPLot Enroll Vs PhD |
| Plot 2.1.26 | Terminal Variable BoxPlot |  | Plot 2.1.56 | ScatterPLot Enroll Vs Terminal |
| Plot 2.1.27 | S.F.Ratio Variable Histogram |  | Plot 2.4.1 | Scaling Comparison |
| Plot 2.1.28 | S.F.Ratio Variable BoxPlot |  | Plot 2.4.2 | Before Scaling BoxPlot |
| Plot 2.1.29 | perc.alumni Variable Histogram |  | Plot 2.4.3 | After Scaling BoxPlot |
|  |  |  | Plot 2.8.1 | Scree Plot for PC |

## List of Tables

|  |  |
| --- | --- |
| **Table Details** | |
| Table 2.2.1 | Unit Of Measurement |
| Table 2.2.2 | Pre and Post Scaling Summary |
| Table 2.4.1 | Outliers Before/ After Scaling |
| Table 2.5.1 | Eigen value |
| Table 2.5.2 | Eigen Vector |
| Table 2.6.1 | Variance Ratio |
| Table 2.6.2 | Correlation HEATMAP |
| Table 2.7.1 | PC Feature Weightage |
| Table 2.7.2 | PC Feature List |
| Table 2.8.1 | Cumulative Variance |
| Table 2.8.2 | PC Variance |

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*End of Document \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*