***Statistical analysis of the Graduate Admissions data***

***Conor Clarke – D00183971***

Contents

[Description about dataset 3](#_Toc7701526)

[Visualisation of the Data 4](#_Toc7701527)

[Descriptive Statistics 8](#_Toc7701528)

[Research Questions 9](#_Toc7701529)

[Correlation R Code with Chance of Admittance 11](#_Toc7701530)

[Linear Regression – Chance of Admittance and GRE Score 12](#_Toc7701531)

[Linear Regression – Chance of Admittance and TOEFL Score 13](#_Toc7701532)

[Linear Regression – Chance of Admittance and CGPA 14](#_Toc7701533)

[Hypothesis Testing 15](#_Toc7701534)

[Hypothesis Tests: Chance of Admittance and Sex 15](#_Toc7701535)

[Hypothesis Tests: Chance of Admittance and Research 15](#_Toc7701536)

[Hypothesis Tests: Chance of Admittance and CGPA 16](#_Toc7701537)

# Description about dataset

The dataset contains several parameters which are considered important during the application for Masters Programs from Indian students. There are 400 observations and 10 variables.

> dim(mydata)

[1] 400 10

|  |  |  |
| --- | --- | --- |
| Name | Variable Type | Description |
| GRE Scores | Continuous Numeric | General Test Score (out of 340) |
| TOEFL Scores | Continuous Numeric | Test of English as a Foreign Language Scores (out of 120) |
| University Rating | Nominal Categorical | University Rating (out of 5) |
| SOP | Continuous Numeric | Statement of Purpose Strength (out of 5) |
| LOR | Continuous Numeric | Letter of Recommendation Strength (out of 5) |
| CGPA | Continuous Numeric | Undergraduate GPA (out of 10) |
| Research | Nominal Categorical | Research Experience (either 0 or 1) |
| Chance of Admit. | Continuous Numeric | Chance of Admit (range from 0 to 1) |
| Sex | Nominal Categorical | Sex (Male or Female) |

> which(is.na(mydata))

[1] 3612 3688 3782 3859 3970

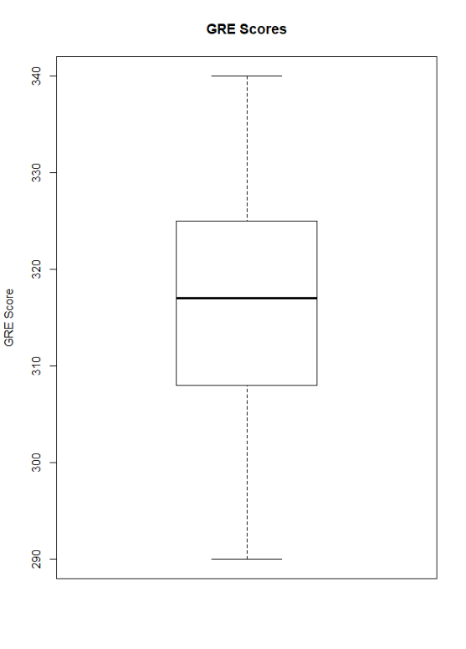
> which(is.na(mydata$Sex))

[1] 12 88 182 259 370

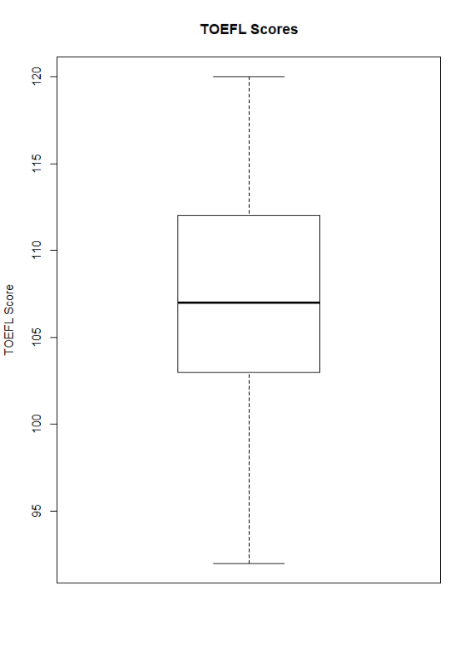
There are 5 missing data elements in this dataset, they occur in the Sex category where Male or Female was not entered in, you can remove the missing data, but here we are just going to ignore it as the influence of Sex won’t impact the data a great deal.

We will also examine outliers by looking at plots for each dataset.

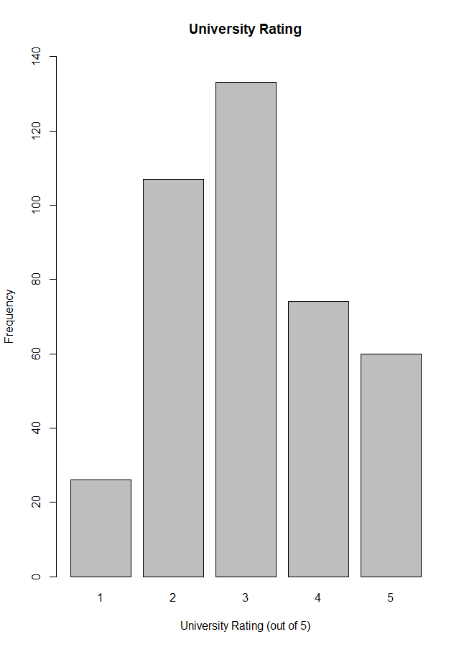
# Visualisation of the Data



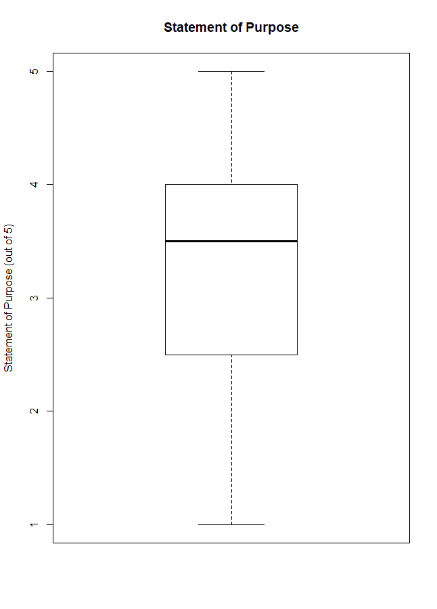
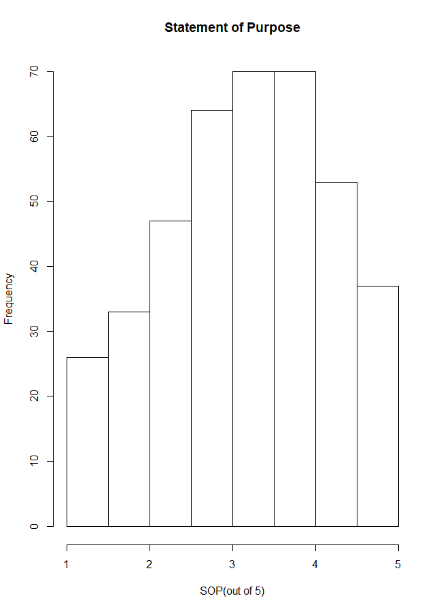
* GRE Score looks roughly symmetric from plot,
* contains no outliers.



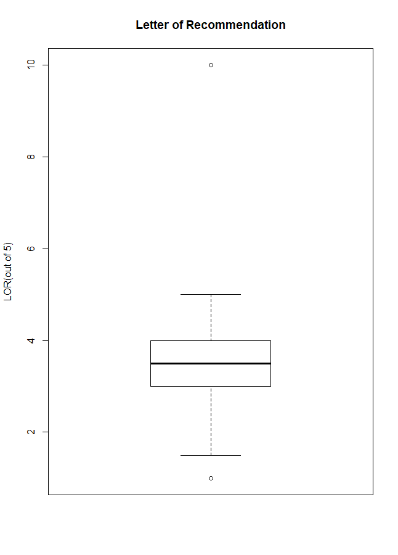
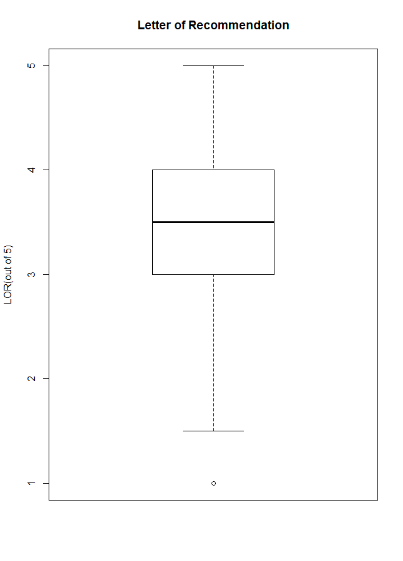
* TOEFL Score looks roughly symmetric,
* contains no outliers.



* University Rating looks to be symmetric
* contains no outliers



* Statement of Purpose (out of 5) looks symmetric with a of skew to the left (look at histogram),
* contains no outliers

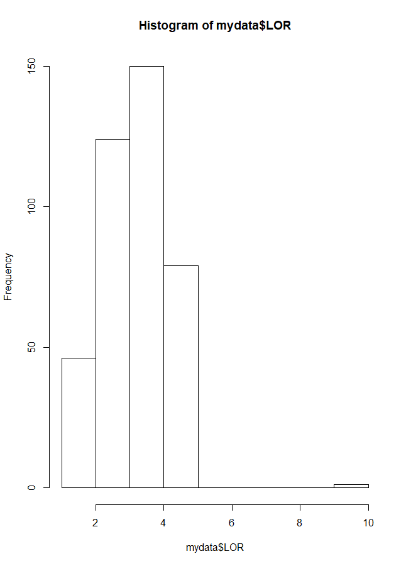
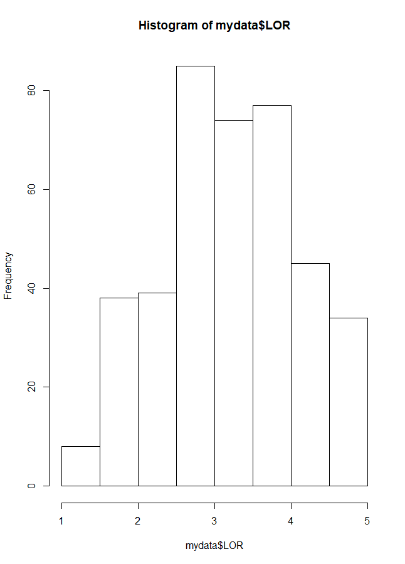


* Boxplot and histogram on the left contains an outlier that should not be present in the data, Letter of Recommendation was given a score out of 5, there is one with a 10.
* To solve this, I got the value and made it equal to the median of this data column.

> mydata$LOR[mydata$LOR == 10] <- median(mydata$LOR)

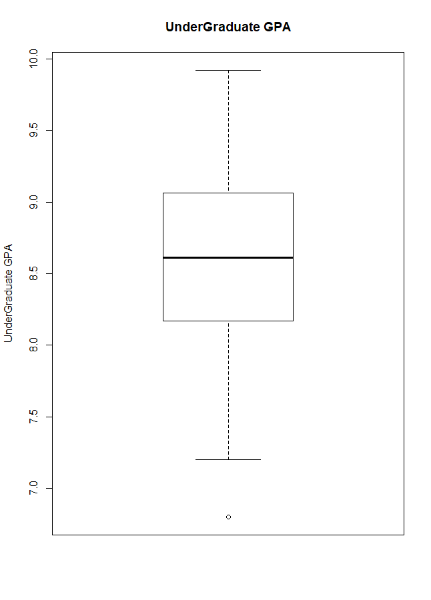
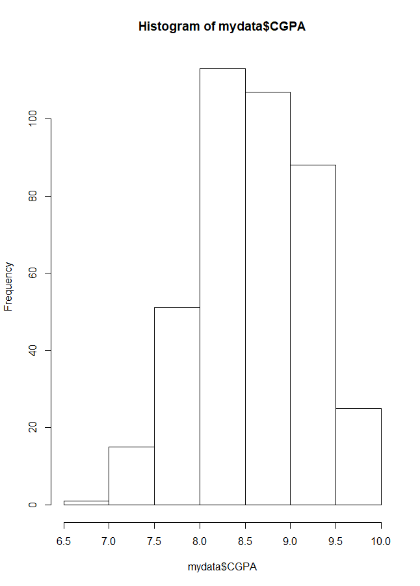
> hist(mydata$LOR)

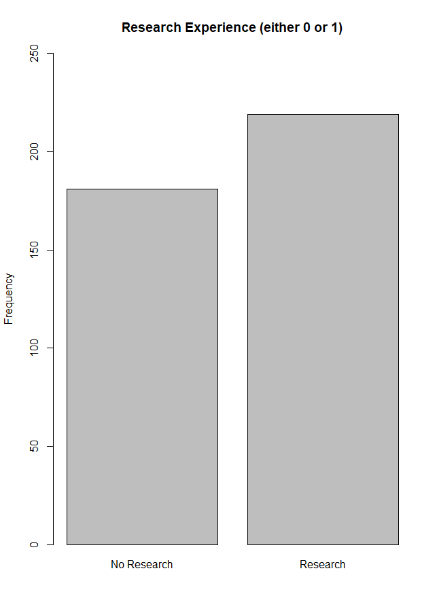
*Before* *After*



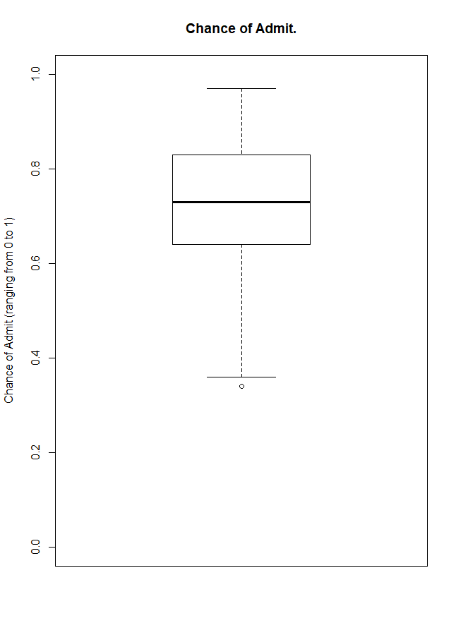
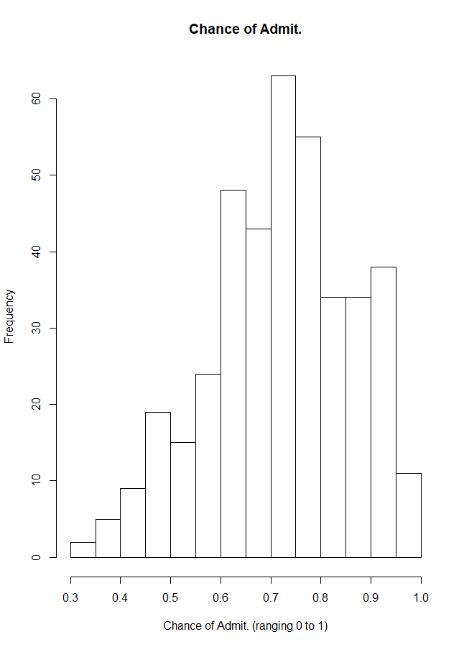
* As you can see there is a much better shape, it is now roughly symmetric
* Does not contain an outlier as it has been replaced by the median value of the data column

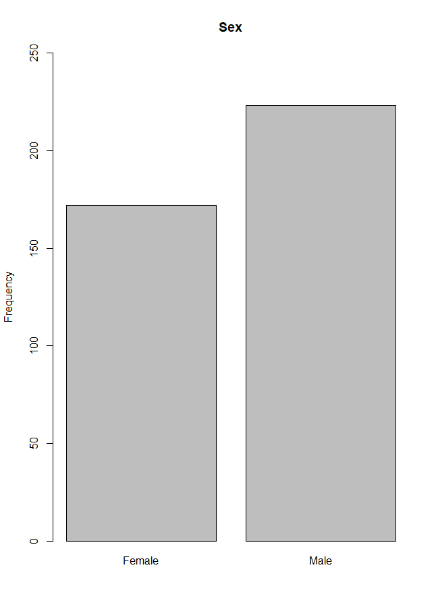
*Before* *After*

* Undergraduate GPA looks to be roughly symmetric, a bit skewed to the left
* Contains an outlier, but will leave in as it is a plausible value



* Most common research experience is 1



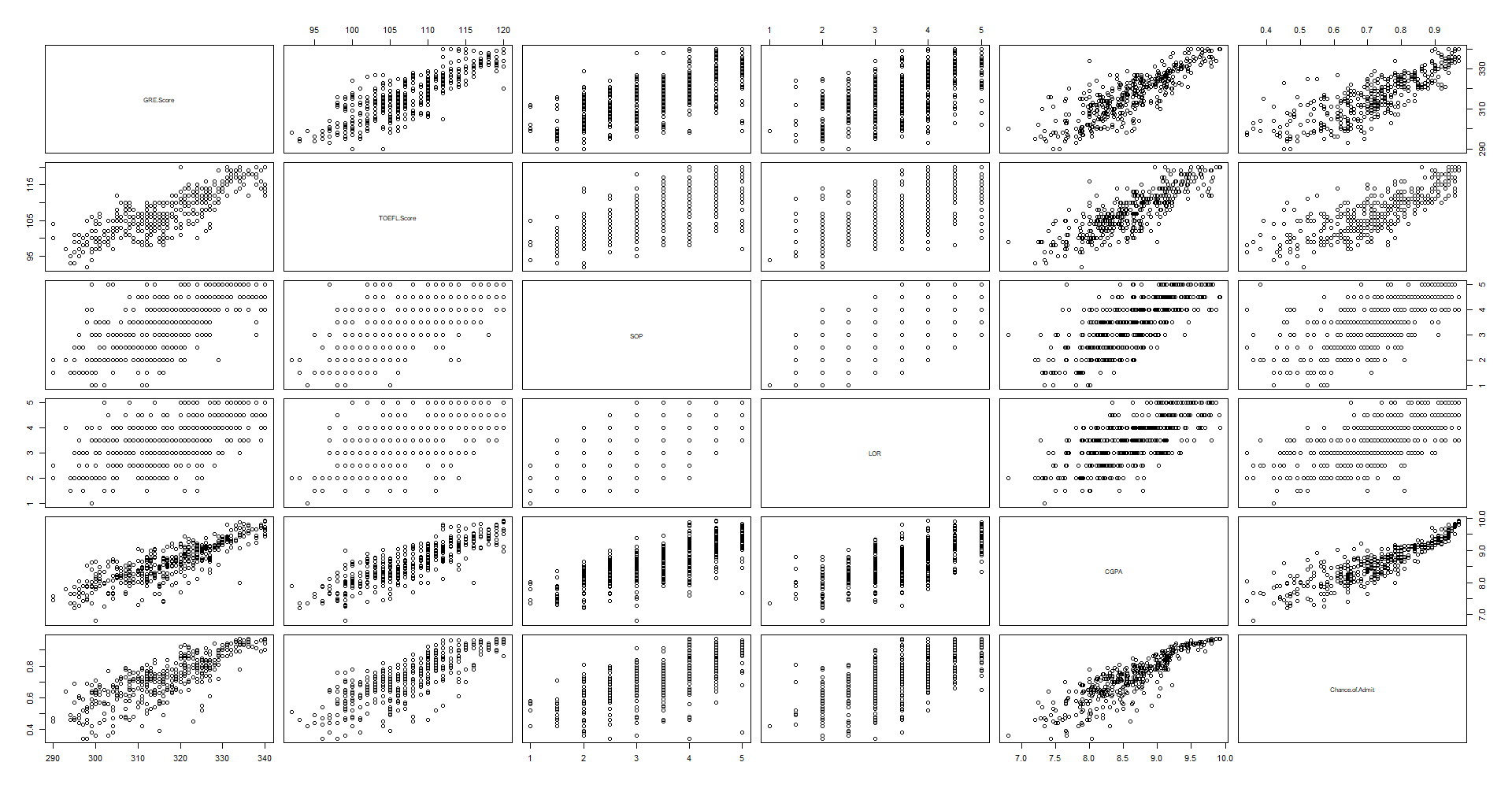
* Chance of admittance looks to be symmetric
* Contains an outlier, but will leave it as it is a plausible value
* Most common sex is male

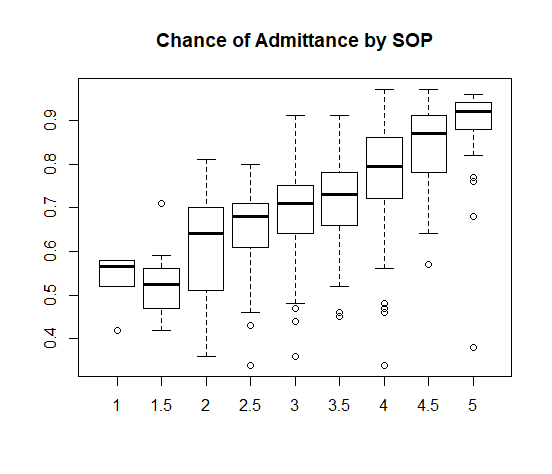
# Descriptive Statistics

TABLE FOR MEAN, MEDIAN, STANDARD DEVIATION AND IQR

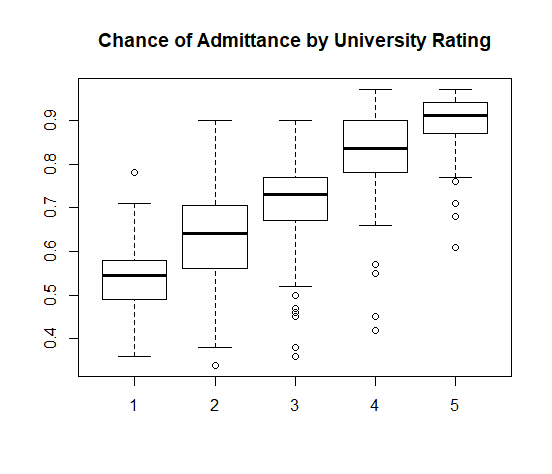
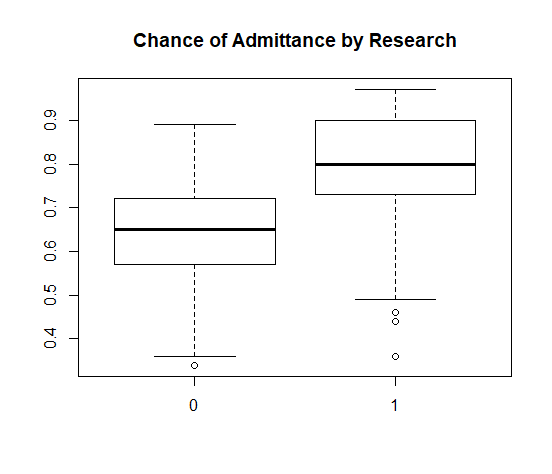
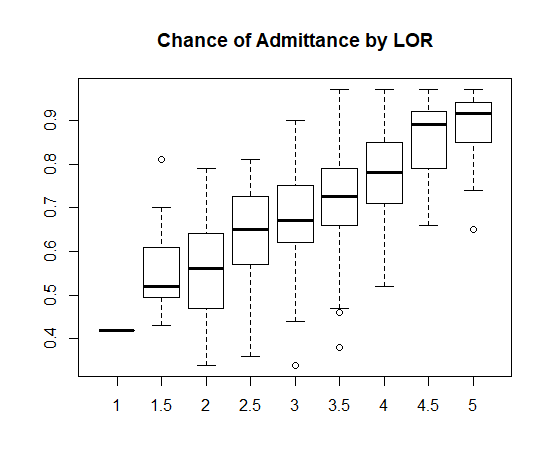
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Name | Mean | Median | Standard Deviation | IQR |
| GRE.Scores | General Test Score (out of 340) | 316.8075 | 317 | 11.47365 | 17 |
| TOEFL.Scores | Test of English as a Foreign Language Scores (out of 120) | 107.41 | 107 | 6.069514 | 9 |
| University.Rating | University Rating (out of 5) | 3.0875 | 3 | 1.143728 | 2 |
| SOP | Statement of Purpose Strength (out of 5) | 3.4 | 3.5 | 1.006869 | 1.5 |
| LOR | Letter of Recommendation Strength (out of 5) | 3.44875 | 3.5 | 0.8951266 | 1 |
| CGPA | Undergraduate GPA (out of 10) | 8.598925 | 8.61 | 0.5963171 | 0.8925 |
| Research | Research Experience (either 0 or 1) | 0.5475 | 1 | 0.498362 | 1 |
| Chance.of.Admit. | Chance of Admit (range from 0 to 1) | 0.72435 | 0.73 | 0.1426093 | 0.19 |
| Sex | Sex (Male or Female) | N/A | N/A | N/A | N/A |

# Research Questions

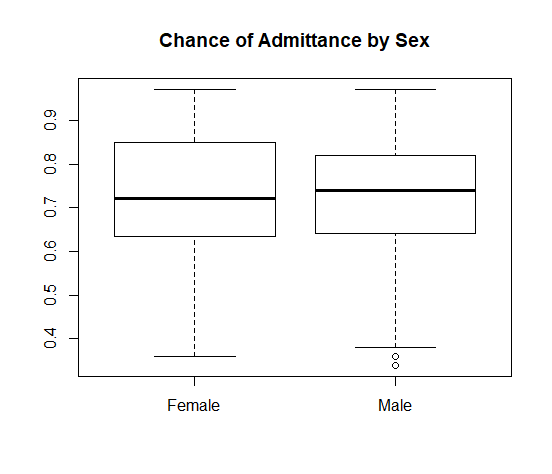
* Data is clearly collected to see what influences the chance of admittance; therefore, we are going to mostly treat this as the response variable.
* Linear Relationships:
* Chance of Admittance and GRE Scores (positive)
* Chance of Admittance and TOEFL Scores (positive)
* Chance of Admittance and CGPA (positive)
* GRE Scores and TOEFL Scores (positive)
* GRE Scores and CGPA (positive)
* TOEFL Scores and CGPA (positive)



* Seems to be a relationship between chance of admittance and Statement of Purpose
* But there are nine groups only know how to test the difference in the means for two groups



* Seems to be a relationship between chance of admittance and Letter of Recommendation
* But there are nine groups only know how to test the difference in the means for two groups
* Seems to be a relationship between chance of admittance and University Rating
* But there are five groups only know how to test the difference in the means for two groups
* Seems to be a relationship between chance of admittance and Research
* But there are nine groups only know how to test the difference in the means for two groups



* Seems to be a relationship between chance of admittance and Sex

# Correlation R Code with Chance of Admittance

> cor(mydata$Chance.of.Admit, mydata$GRE.Score)

[1] 0.8026105 #very strong positive linear relationship

> cor.test(mydata$Chance.of.Admit, mydata$GRE.Score)

Pearson's product-moment correlation

data: mydata$Chance.of.Admit and mydata$GRE.Score

t = 26.843, df = 398, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.7647419 0.8349536

sample estimates:

cor

0.8026105

> cor(mydata$Chance.of.Admit, mydata$TOEFL.Score)

[1] 0.791594 #very strong positive linear relationship

> cor.test(mydata$Chance.of.Admit, mydata$TOEFL.Score)

Pearson's product-moment correlation

data: mydata$Chance.of.Admit and mydata$TOEFL.Score

t = 25.845, df = 398, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.7519028 0.8255675

sample estimates:

cor

0.791594

> cor(mydata$Chance.of.Admit, mydata$CGPA)

[1] 0.8732891 #very strong positive linear relationship

> cor.test(mydata$Chance.of.Admit, mydata$CGPA)

Pearson's product-moment correlation

data: mydata$Chance.of.Admit and mydata$CGPA

t = 35.759, df = 398, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.8478354 0.8947275

sample estimates:

cor

0.8732891

## Linear Regression – Chance of Admittance and GRE Score

* Res vs Fits => equal var assumption, happy with as the spread of the residuals is roughly the same along the x-axis
* Qqplot => normality assumption for residuals, happy with because majority of data along diagonal with some deviation from the diagonal from the two tails

lm(formula = Chance.of.Admit ~ GRE.Score, data = mydata)

Residuals:

Min 1Q Median 3Q Max

-0.33613 -0.04604 0.00408 0.05644 0.18339

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.4360842 0.1178141 -20.68 <2e-16 \*\*\*

GRE.Score 0.0099759 0.0003716 26.84 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08517 on 398 degrees of freedom

Multiple R-squared: 0.6442, Adjusted R-squared: 0.6433

F-statistic: 720.6 on 1 and 398 DF, p-value: < 2.2e-16

* Significant relationship between chance of admittance and GRE Score, (p-value = )
* Equation: chance of admittance = -2.4360842 + 0.0099759(GRE Score)
* Therefore, as GRE Score increase by 1, chance of admittance increases by 0.0099759
* 64.42% of variation of chance of admittance is explained by GRE Score
* Residual standard error is 0.08517

## Linear Regression – Chance of Admittance and TOEFL Score

* Res vs Fits => equal var assumption, happy with as the spread of the residuals is roughly the same along the x-axis
* Qqplot => normality assumption for residuals, happy with because majority of data along diagonal with some deviation from the diagonal from the two tails

lm(formula = Chance.of.Admit ~ TOEFL.Score, data = mydata)

Residuals:

Min 1Q Median 3Q Max

-0.31252 -0.05128 0.01328 0.05453 0.21067

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.2734005 0.0774217 -16.45 <2e-16 \*\*\*

TOEFL.Score 0.0185993 0.0007197 25.84 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08725 on 398 degrees of freedom

Multiple R-squared: 0.6266, Adjusted R-squared: 0.6257

F-statistic: 667.9 on 1 and 398 DF, p-value: < 2.2e-16

* Significant relationship between Chance of Admittance and TOEFL Score (p-value = )
* Equation: chance of admittance = -1.2734005 + 0.0185993(TOEFL Score)
* Therefore, as TOEFL increases by 1, chance of admittance increases by 0.0185993
* 62.66% of variation of chance of admittance is being explained by TOEFL Score
* Residual standard error is 0.08725

## Linear Regression – Chance of Admittance and CGPA

* Res vs Fits => equal var assumption, happy with as the spread of the residuals is roughly the same along the x-axis
* Qqplot => normality assumption for residuals, happy with because majority of data along diagonal with some deviation from the diagonal from the two tails

lm(formula = Chance.of.Admit ~ CGPA, data = mydata)

Residuals:

Min 1Q Median 3Q Max

-0.274575 -0.030084 0.009443 0.041954 0.180734

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.07151 0.05034 -21.29 <2e-16 \*\*\*

CGPA 0.20885 0.00584 35.76 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06957 on 398 degrees of freedom

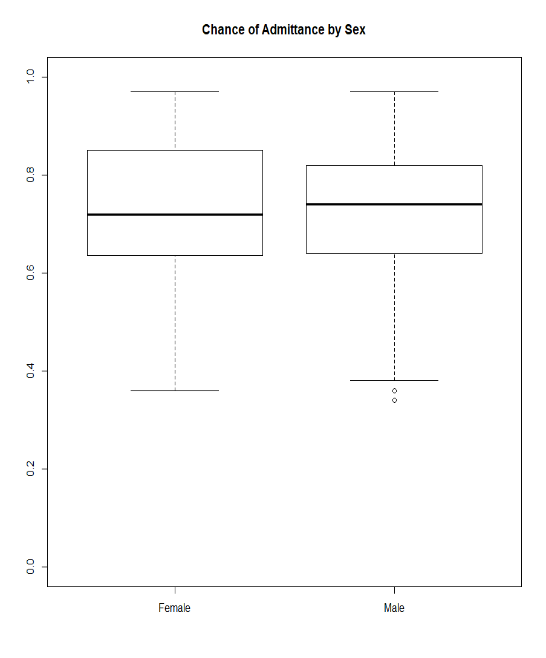
Multiple R-squared: 0.7626, Adjusted R-squared: 0.762

F-statistic: 1279 on 1 and 398 DF, p-value: < 2.2e-16

* Significant relationship between Chance of Admittance and CGPA, (p-value = )
* Equation: Chance of Admittance = -1.07151 + 0.20885(CGPA)
* Therefore, as CGPA increases by 1, chance of admittance increases by 0.20885
* 76.26% of variation of chance of admittance is being explained by CGPA
* Residual standard error is 0.06957
* Since is higher and standard error is lower for this model, this model is a better model for predicting chance of admittance

# Hypothesis Testing

## Hypothesis Tests: Chance of Admittance and Sex



Roughly happy with normality assumption and equal variance assumption.

> t.test(mydata$Chance.of.Admit~mydata$Sex)

Welch Two Sample t-test

data: mydata$Chance.of.Admit by mydata$Sex

t = -0.11669, df = 359.96, p-value = 0.9072

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.03050237 0.02708543

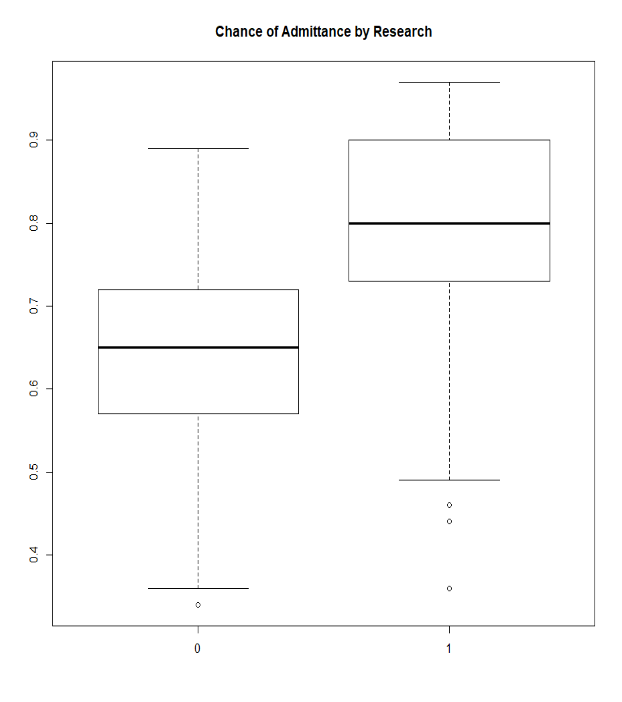
sample estimates:

mean in group Female mean in group Male

0.7233140 0.7250224

* There does not seem to be a significant relationship between chance of admittance and sex (looking at p-value)

## Hypothesis Tests: Chance of Admittance and Research



Roughly happy with normality assumption and equal variance assumption.

> t.test(mydata$Chance.of.Admit~mydata$Research)

Welch Two Sample t-test

data: mydata$Chance.of.Admit by mydata$Research

t = -13.347, df = 392.98, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.1816199 -0.1349844

sample estimates:

mean in group 0 mean in group 1

0.6376796 0.7959817

* Significant relationship between chance of admittance and research (looking at p-value and 95% CI)

## Hypothesis Tests: Chance of Admittance and CGPA

> t.test(mydata$Chance.of.Admit,mydata$GRE.Score)

Welch Two Sample t-test

data: mydata$Chance.of.Admit and mydata$GRE.Score

t = -550.93, df = 399.12, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-317.2111 -314.9552

sample estimates:

mean of x mean of y

0.72435 316.80750

* Significant relationship between chance of admittance and GRE Score (looking at p-value and 95% CI)

***Conclusion***

Many aspects of Graduate Admissions influence the chance of admittance of Graduates including GRE Scores, TOEFL Scores, and CGPA.

Also, there are other influences that were not tested to see if there is a significant relationship but examining plots and statistics, chance of admittance appears to have a relationship with statement of purpose, letter of recommendation, research, and University Rating.

Only Sex has no significant relationship with chance of admittance.