"Prediction of Admission in Graduate School"

Sara Parizi

Contents

1 Introduction **2**

2 Problem Statement **2**

3 Purpose Statement **2**

4 Exploring and Understanding Data **2**

4.1 Categorical Variables 3

4.2 Numerical Variables 7

4.3 Exploring relationships among features 15

5 Data Preparation - Creating Training and Test Datasets **17**

6 Building the Regression Model **17**

6.1 Model 1 (Included All Of The Features) 17

6.2 Model 2 (Including the Features with High Significant Levels) 22

7 Building the Regression Trees **29**

8 Neural Network Medels **30**

8.1 Neural Network Model (3 hidden nodes and 1 hidden layer) 31

8.2 Neural Network Model (3 hidden nodes and 2 hidden layers) 33

9 Conclusion **34**

References **36**

1 Introduction

Graduate programs aim to foster the development of the next generation of researchers, scholars, professionals, and leaders within a multitude of fields. The masters and doctoral students (graduate students) nurtured within these programs are future thinkers, leaders, researchers, and scholars who create policy, opportunities, and solutions for humanity through study, contemplation, and knowledge creation(Nesbitt 2021). Many students apply to graduate school each year. They can apply to multiple universities and file separate applications to each one. Applying to graduate school is costly and can be stressful. The outcome of the admission process may affect a student's life and career trajectory considerably.
There are several academic performance measurements such as CGOP, GRE, TOEFL ,LOR and etc can effect to admission of a student in a graduate school, but which factor affect more?

2 Problem Statement

Finding a model to predict the chance of admissions in graduate school using machine learning algorithms.

3 Purpose Statement

This project is aimed at formulating a predictive model for forecasting the chances of admission to universities based on the explanatory variables such as 'GRE Score', 'TOEFL Score', 'University Rating', 'Statement of Purpose', 'Letter of Recommendation Strength', 'CGPA' and 'Research Experience'. How likely is a student can get an admission from a graduate school? Does one specific variable hold more weight than another? whether a particular variable that students should focus on more, when their goal is to attend a graduate school. Does the GRE hold more weight than have research experience? Are letters of recommendation more helpful than a higher GPA?

4 Exploring and Understanding Data

The dataset has downloaded from 'Kaggle' website and was built with the purpose of helping students in shortlisting universities with their profiles. This dataset is inspired by

the 'UCLA Graduate Dataset' and is owned by Mohan Acharya, Asfia Armaan, Aneeta Antony. It contains 400 sets of student's information ranging from test scores to research experience(Acharya 2018).

The dataset has 9 columns and 400 rows with 0 NA values. It contains several parameters which are considered important during the application for Masters Programs.The parameters included are:'GRE Score', 'TOEFL Score', 'University Rating', 'Statement of Purpose', 'Letter of Recommendation Strength', 'CGPA', 'Research Experience' and 'Chance of Admission'.

The top 5 rows of the data has shown in the below table :

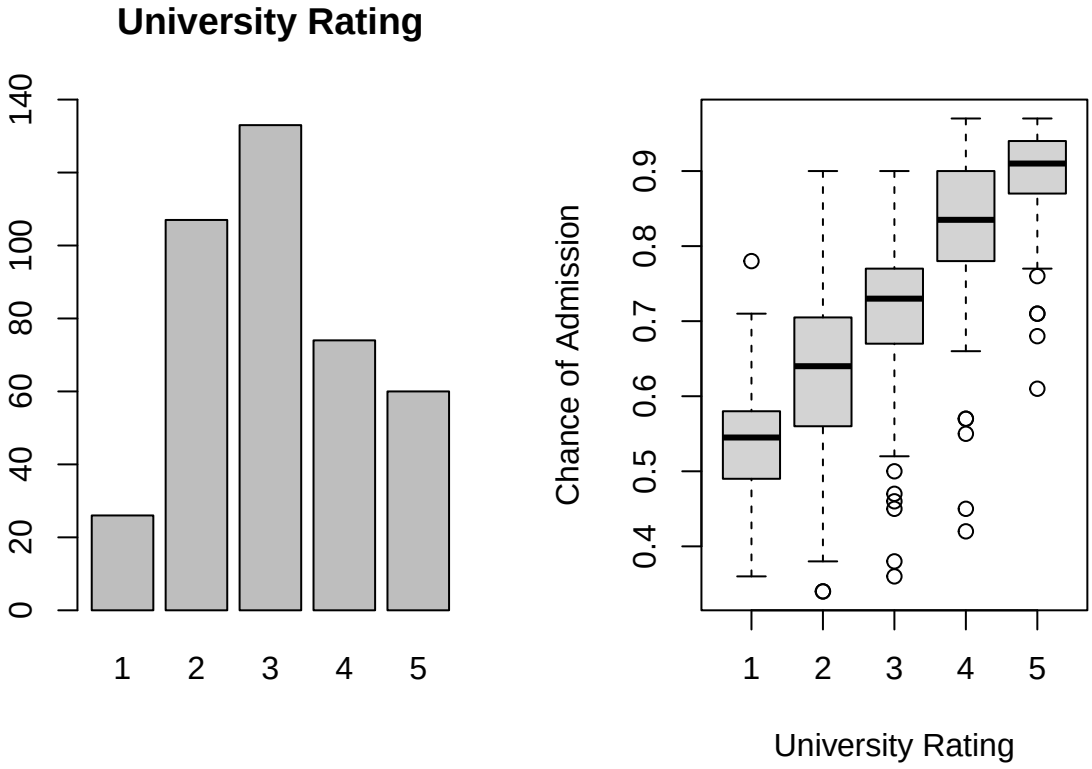
Serial No	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admission
1	337	118	4	4.5	4.5	9.65	1	0.92
2	324	107	4	4.0	4.5	8.87	1	0.76
3	316	104	3	3.0	3.5	8.00	1	0.72
4	322	110	3	3.5	2.5	8.67	1	0.80
5	314	103	2	2.0	3.0	8.21	0	0.65

4.1 Categorical Variables

(1) **University Rating**: The rating of the university (out of 5)
College and university rankings order the best institutions in higher education based on factors that vary depending on the ranking. Some rankings evaluate institutions within a single country, while others assess institutions worldwide. Rankings are typically conducted by magazines, newspapers, websites, government, or academics. In addition to ranking entire institutions, specific programs, departments, and schools can be ranked. Some rankings consider measures of wealth, excellence in research selective admissions, and alumni success. Rankings may also consider various combinations of measures of specialization expertise, student options, award numbers, internationalization, graduate employment, industrial linkage, historical reputation, and other criteria.
There is much debate about rankings' interpretation, accuracy, and usefulness. The expanding diversity in rating methodologies and accompanying criticisms of each indicate the lack of consensus in the field. Further, it seems possible to game the ranking systems through excessive self-citations or by researchers supporting each other in surveys. UNESCO has questioned whether rankings "do more harm than good", while acknowledging that "Rightly or wrongly, they are perceived as a measure of quality and so create intense competition between universities all over the world"(wiki, n.d.b).
The spread of the data is shown in the below bar plot:

```
  1   2   3   4   5
 26 107 133  74  60
```

```
   1    2    3    4    5
0.06 0.27 0.33 0.18 0.15
```



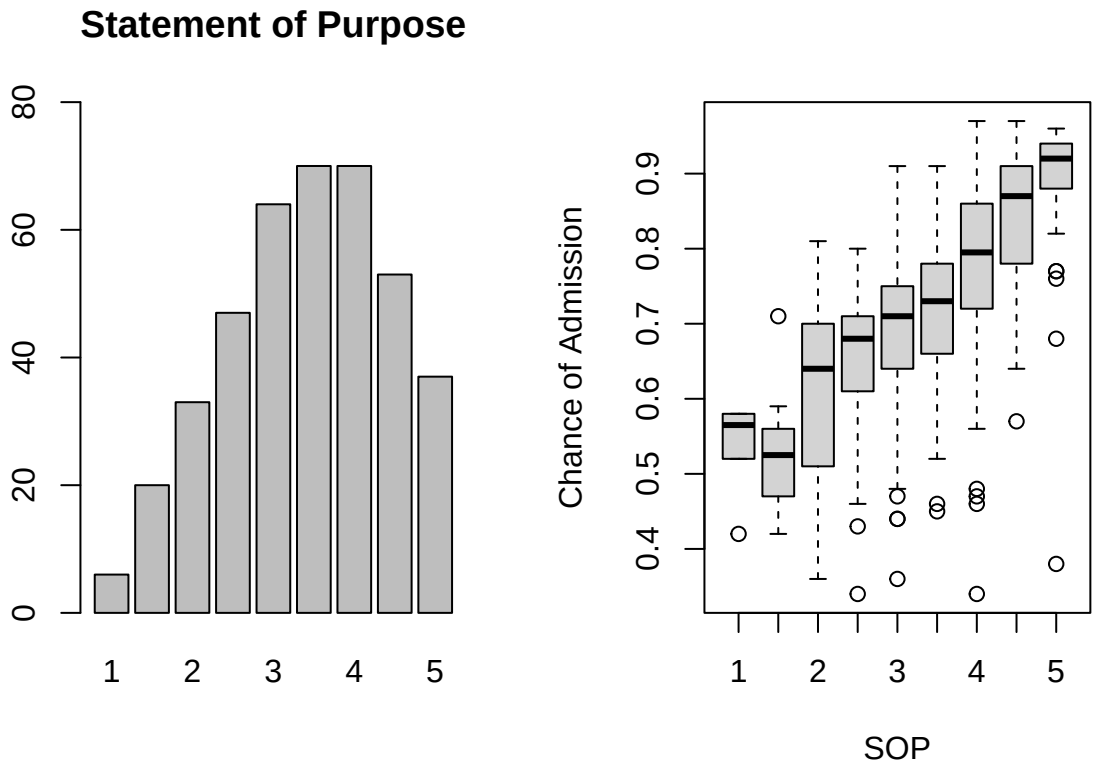
	Var1	Freq
0.065	1	26
0.2675	2	107
0.3325	3	133
0.185	4	74
0.15	5	60

As we can see in the table and plot above, the proportion of universities with low rating of 1 with 0.06 (6%) is the least proportion among others. Then higher ratings of 4 and 5 with 18% and 15% have the shorter bars in the plot while the universities of rating 3 with 33% has almost $\approx \frac{1}{3}$ of the whole and the UR of 2 with 26% has the second tall bar among the ratings.

(2) **SOP**: The statement of purpose strength (out of 5)
An admissions or application essay, sometimes also called a personal statement or a statement of purpose, is an essay or other written statement written by an applicant, often a prospective student applying to some college, university, or graduate school. The application essay is a common part of the university and college admissions process.(wiki, n.d.a)

```
  1 1.5   2 2.5   3 3.5   4 4.5   5
  6  20  33  47  64  70  70  53  37
```

```
    1   1.5     2   2.5     3   3.5     4   4.5     5
0.01 0.05 0.08 0.12 0.16 0.17 0.17 0.13 0.09
```



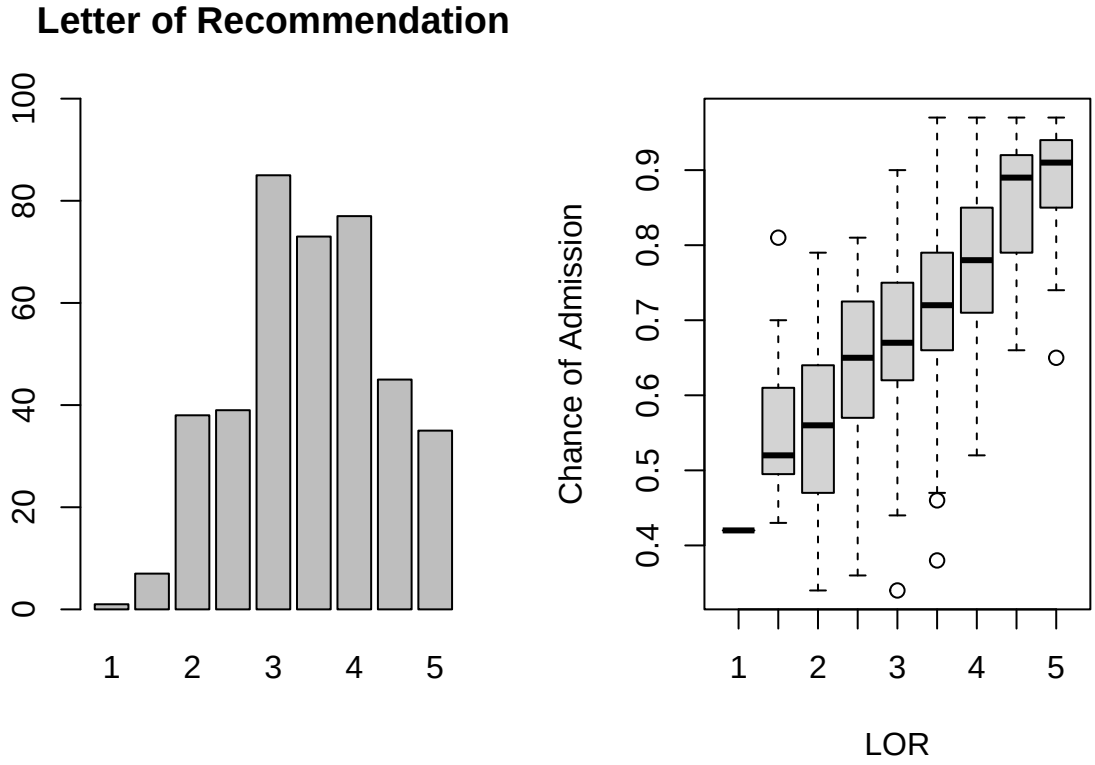
In this data set, the students with 1 SOP are just 1% of the whole while the students with 3.5 and 4 have the highest proportion with 17%.

(3) **LOR**: The letter of recommendation strength (out of 5)
A letter of recommendation or recommendation letter, also known as a letter of reference, reference letter or simply reference, is a document in which the writer

assesses the qualities, characteristics, and capabilities of the person being recommended in terms of that individual's ability to perform a particular task or function. Letters of recommendation are typically related to employment, admission to institutions of higher education, or scholarship eligibility. They are usually written by someone who worked with or taught the person, such as a supervisor, a colleague or teacher.(wiki, n.d.e)

```
##
##   1 1.5   2 2.5   3 3.5   4 4.5   5
##   1   7  38  39  85  73  77  45  35
```

```
##
##      1    1.5      2    2.5      3    3.5      4    4.5      5
## 0.0025 0.0175 0.0950 0.0975 0.2125 0.1825 0.1925 0.1125 0.0875
```

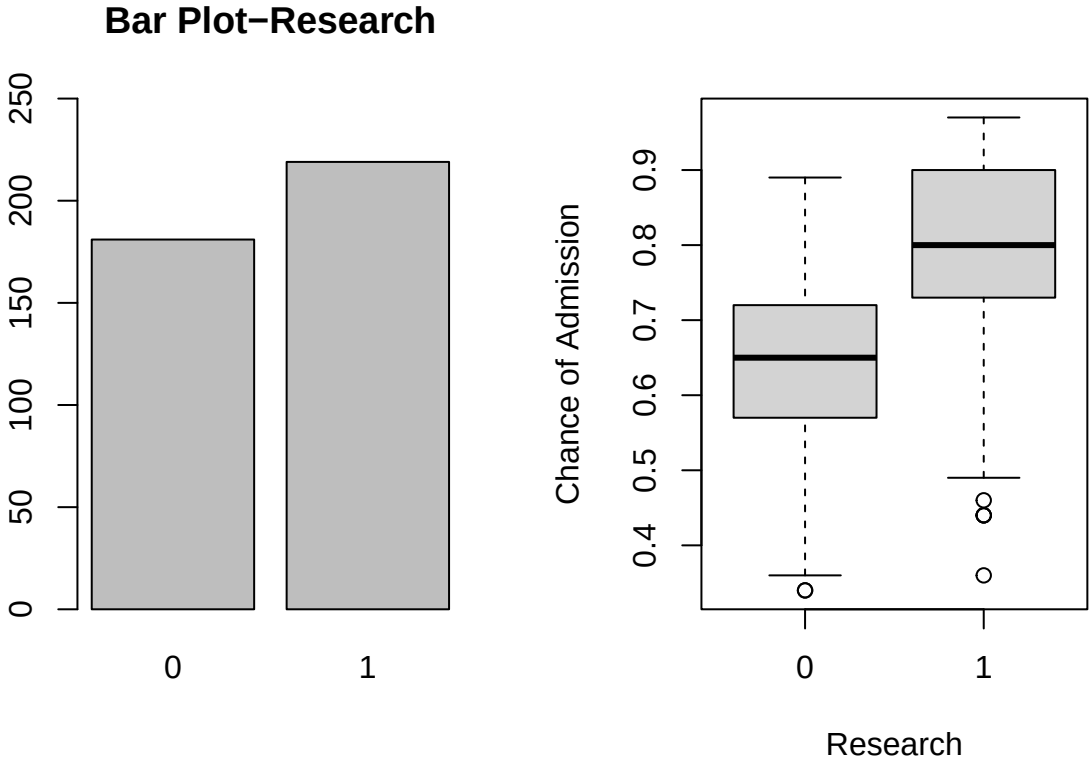


As it is obvious in the above tables and plot, the least number of letter of recommendations is 0.2% for the students with 1 letter while the students with 3 and 4 letters have highest ratio of $\approx 20\%$ among other groups.

(4) **Research**: Research experience (0 or 1)
This is a binary classification that consideres 0 for those who did not have research experience and 1 for students that have research experience in their resume.

```
##
##   0   1
## 181 219
```

```
##
##      0      1
## 0.4525 0.5475
```

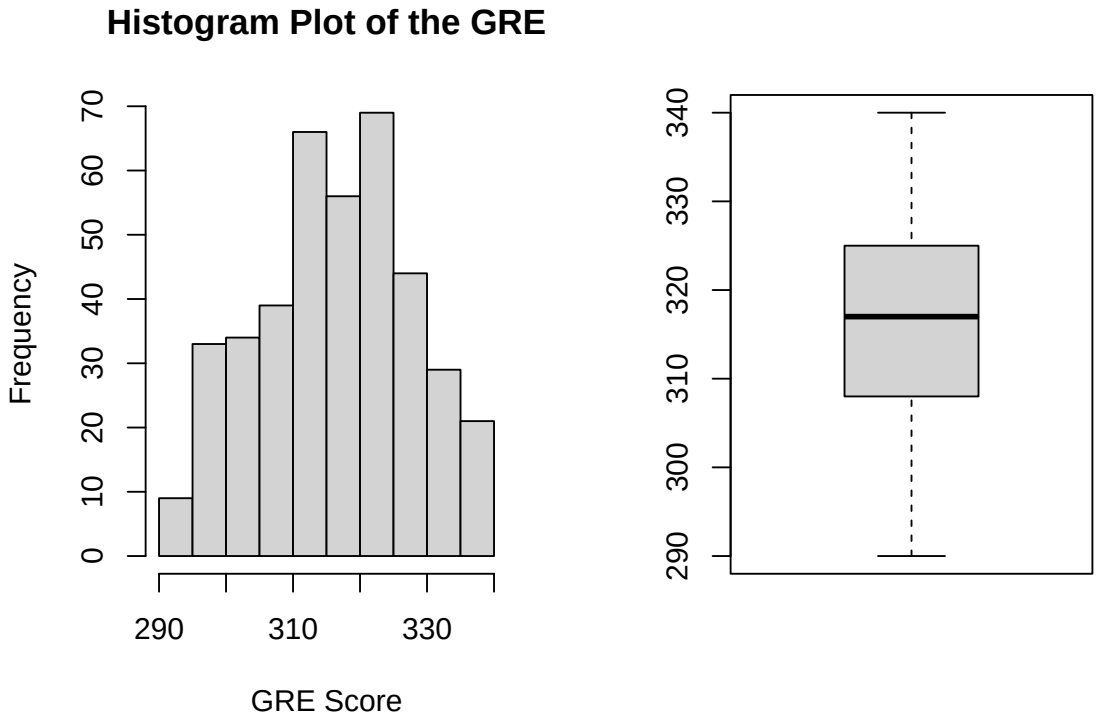


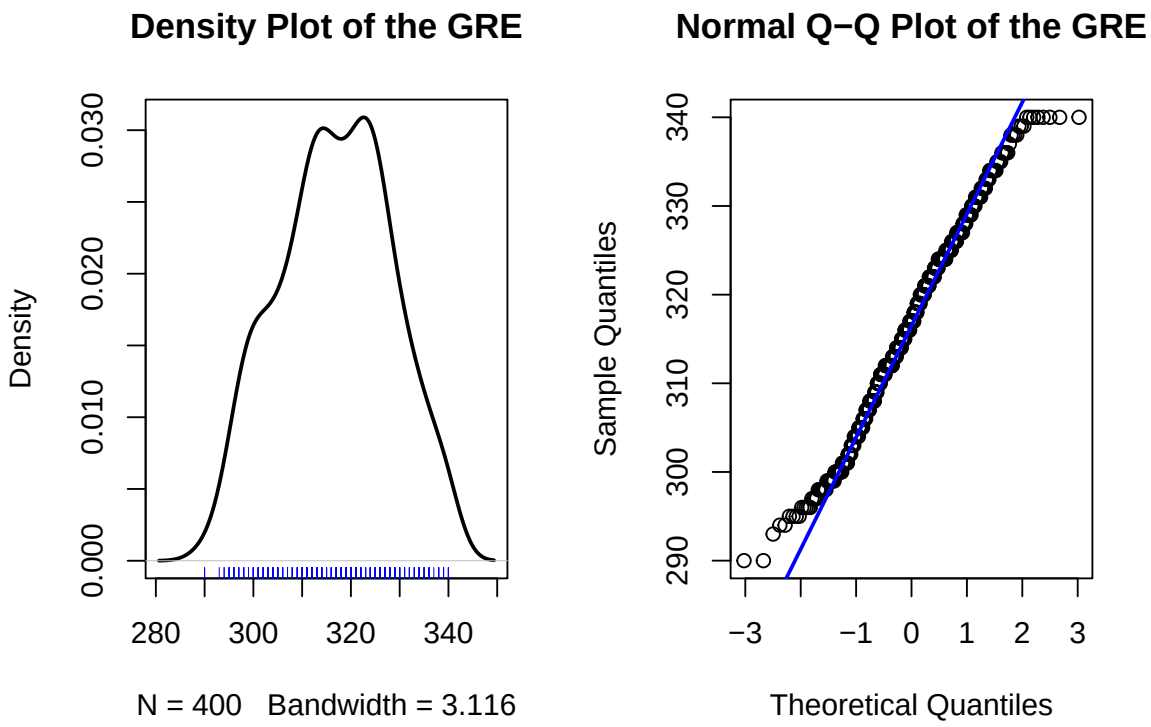
As we can see $\approx 55\%$ of students in this data set have research experience and $\approx 45\%$ of them do not have it.

4.2 Numerical Variables

(5) **GRE Score**: This is the students score in GRE (out of 340)
The Graduate Record Examinations (GRE) is a standardized test that is an admissions requirement for many graduate school in the United States and Canada and a few other countries. The GRE is owned and administered by Educational Testing Service (ETS). The test was established in 1936 by the Carnegie Foundation for the Advancement of Teaching.According to ETS, the GRE aims to measure verbal reasoning, quantitative reasoning, analytical writing, and critical thinking skills that have been acquired over a long period of learning. The content of the GRE consists

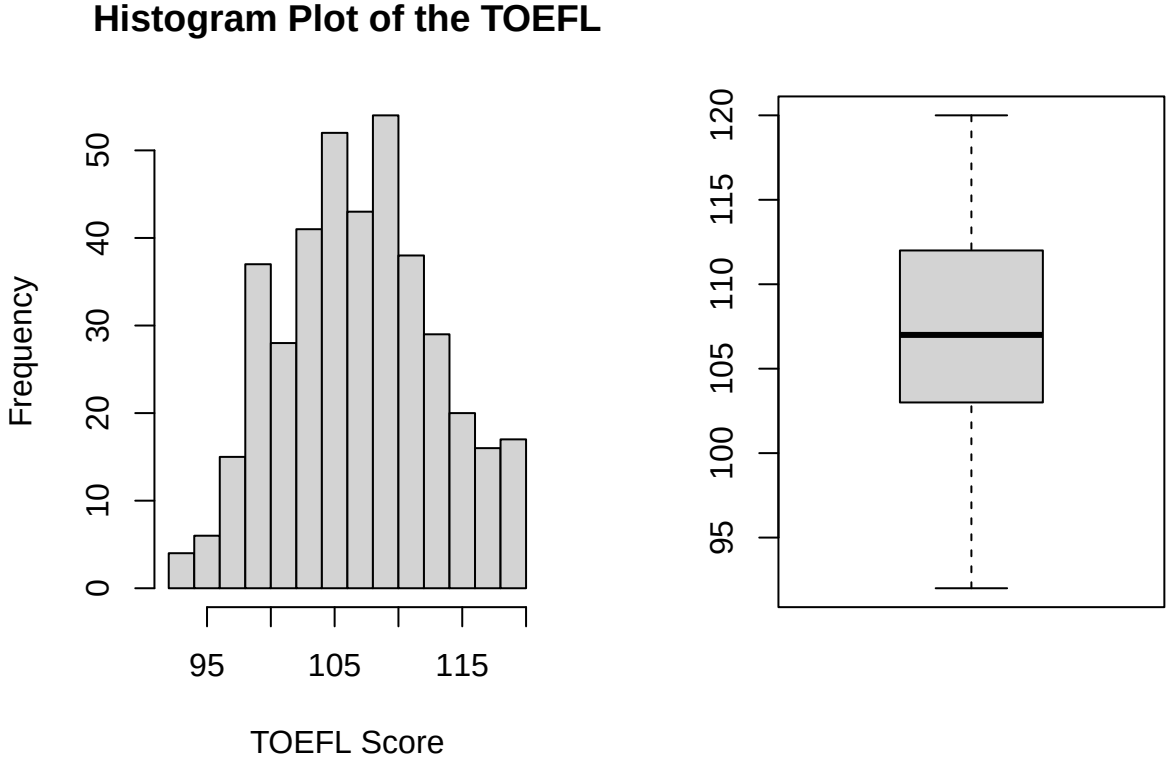
of certain specific algebra, geometry, arithmetic, and vocabulary sections. The GRE General Test is offered as a computer-based exam administered at testing centers and institution owned or authorized by Prometric. In the graduate school admissions process, the level of emphasis that is placed upon GRE scores varies widely between schools and departments within schools. The importance of a GRE score can range from being a mere admission formality to an important selection factor(wiki, n.d.d).

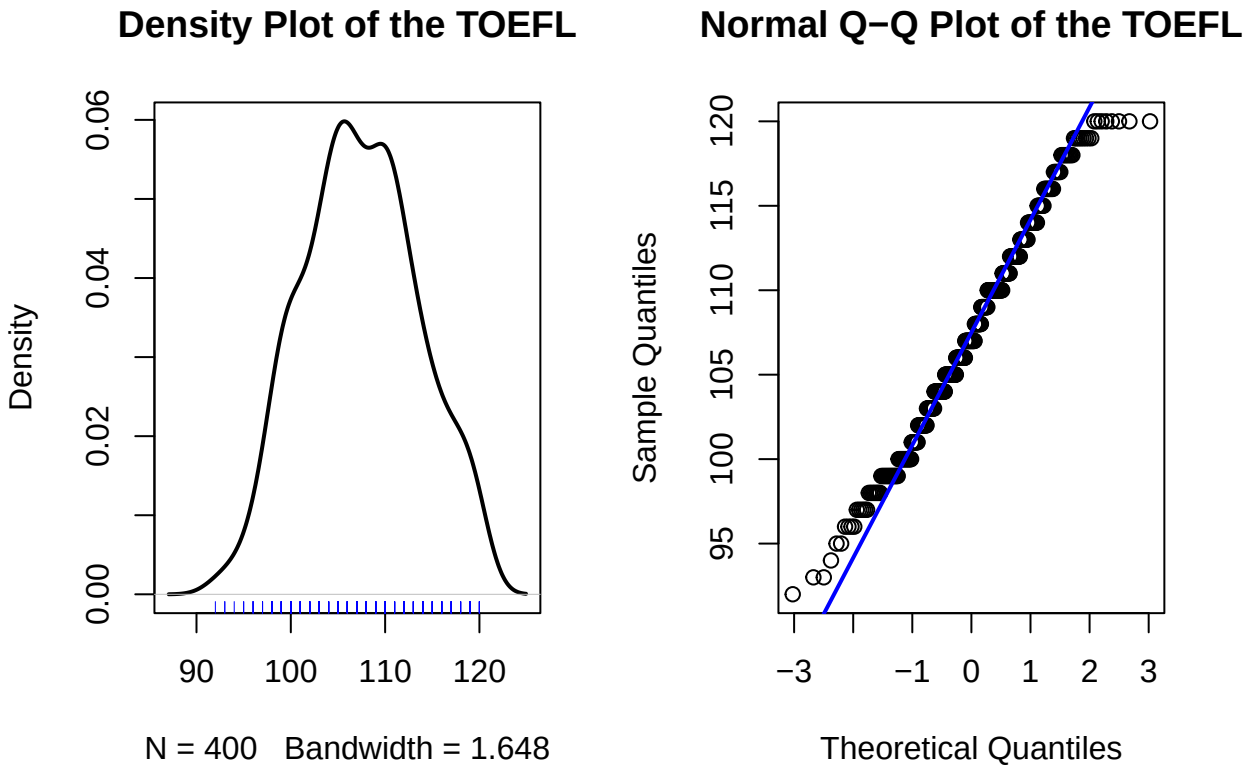




As we see in the above Plots for GRE score, the distribution is close to the bell shape of normal distribution and Q-Q plot for residuals close to the qqline and apparently the GRE data can follow the normal distribution.In order to test the normality “Asymptotic one-sample Kolmogorov-Smirnov test” has been used and the result is as follows:The p_value is 0.268 since it is more than 0.05, it means there are sufficient evidence to support the null hypothesis which is similarity of distribution to the Normal.

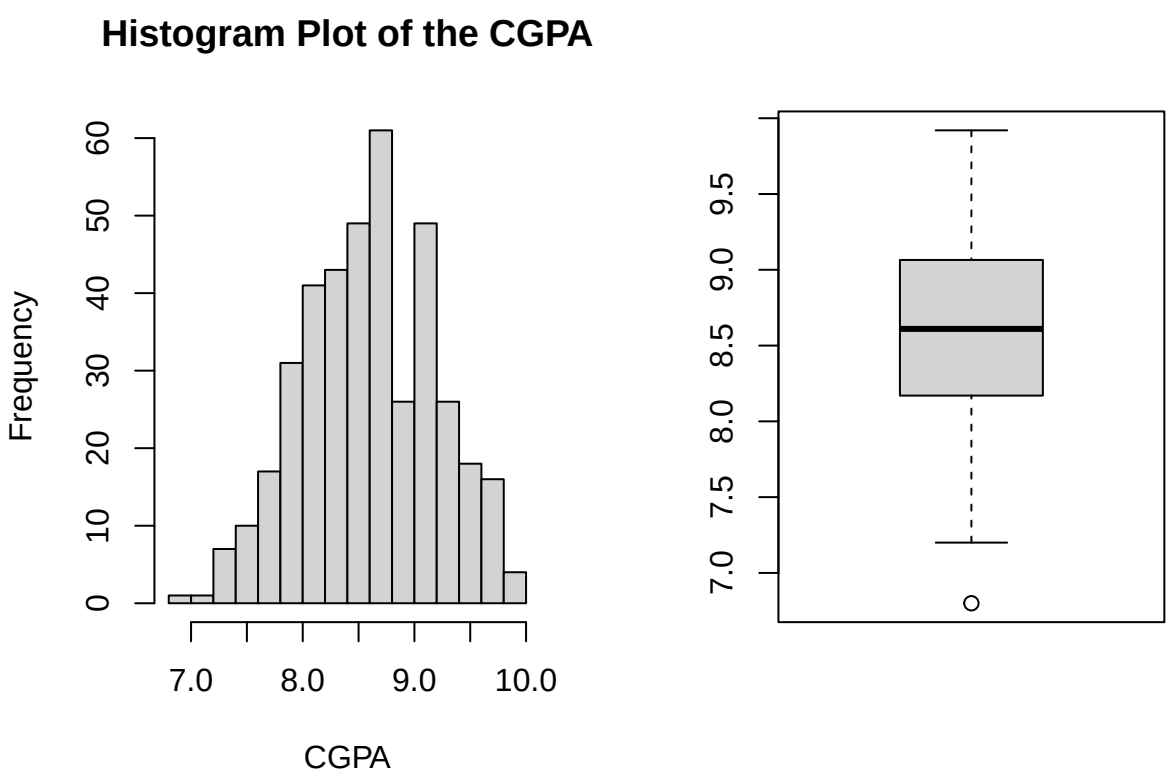
(6) **TOEFL Score**: TOEFL score of the students (out of 120)
Test of English as a Foreign Language is a standardized test to measure the English language ability of non-native speakers wishing to enroll in English-speaking universities. The test is accepted by more than 11,000 universities and other institutions in over 190 countries and territories The TOEFL Internet-based test (iBT) measures all four academic English skills- reading, listening, speaking, and writing. Since its introduction in late 2005, the Internet-based Test format has progressively replaced the computer-based tests (CBT) and paper-based tests (PBT), although paper-based testing is still used in select areas. The TOEFL iBT test is scored on a scale of 0 to 120 points. Each of the four sections (Reading, Listening, Speaking, and Writing) receives a scaled score from 0 to 30. The scaled scores from the four sections are added together to determine the total score. Most colleges use TOEFL scores as only one factor in their admission process, with a college or program within a college often setting a minimum TOEFL score required. The minimum TOEFL iBT scores range from 64 to 110(wiki, n.d.f).

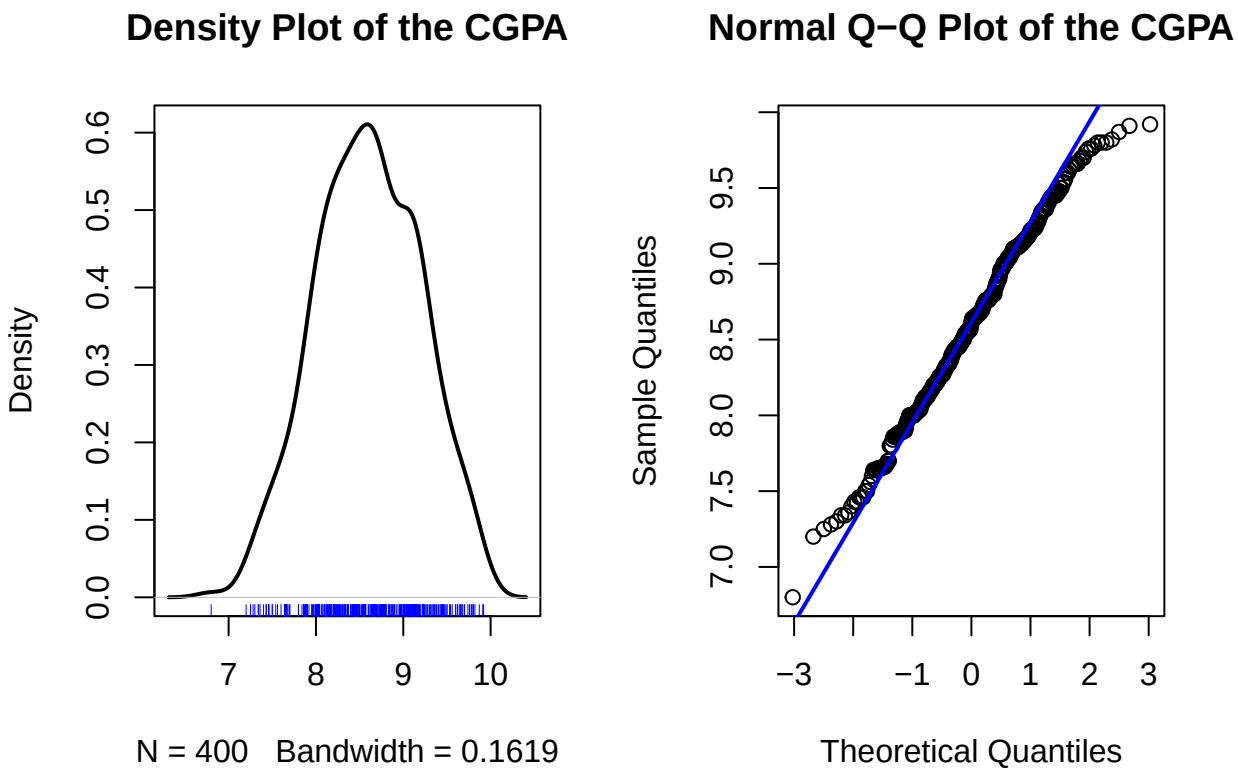

Histogram Plot of the TOEFL
Frequency
0
10
20
30
40
50
95
105
115
TOEFL Score
95
100
105
110
115
120



As we see in the above Plots for TOEFL score, the distribution is close to the bell shape of normal distribution and Q-Q plot for residuals close to the qqline and apparently the TOEFL data can follow the normal distribution.In order to test the normality “Asymptotic one-sample Kolmogorov-Smirnov test” has been used and the result is as follows:The p_value is 0.139 since it is more than 0.05, we have sufficient evidence to say that the data follows the Normal distribution.

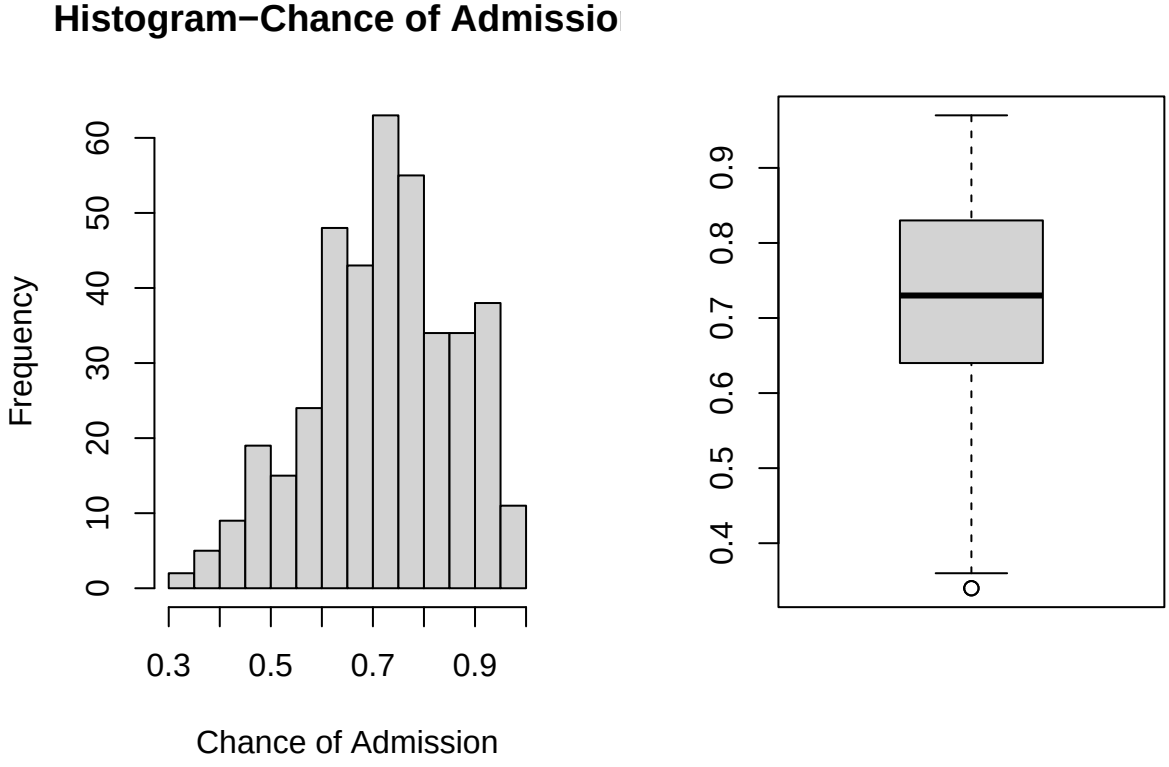
(7) **CGPA**: Cumulative grade point average (out of 10)
Grading in education is the process of applying standardized measurements for varying levels of achievements in a course. Grades can be assigned as letters (usually A through F), as a range (for example, 1 to 6), as a percentage, or as a number out of a possible total (often out of 100). In some countries, grades are averaged to create a grade point average (GPA). GPA is calculated by using the number of grade points a student earns in a given period of time. GPAs are often calculated for high school, undergraduate, and graduate students, and can be used by potential employers or educational institutions to assess and compare applicants. A cumulative grade point average (CGPA), sometimes referred to as just GPA, is a measure of performance for all of a student’s courses(wiki, n.d.c).

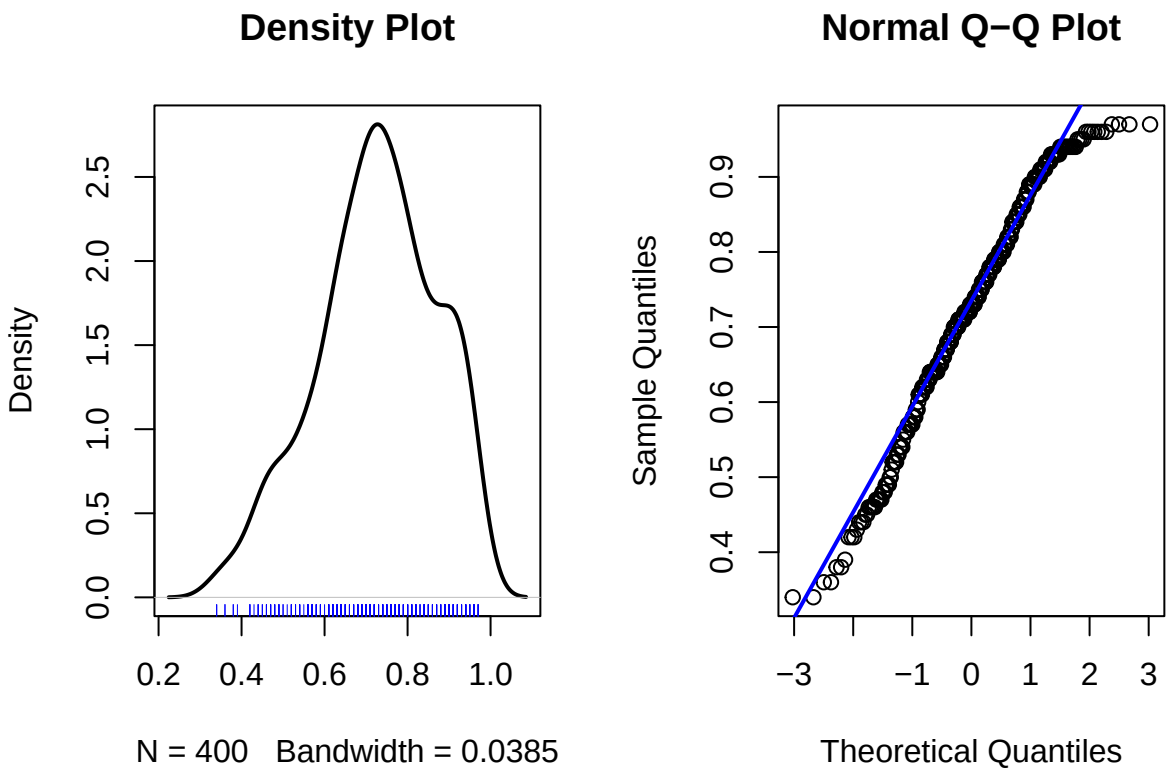

Histogram Plot of the CGPA
Frequency
0
10
20
30
40
50
60
7.0
8.0
9.0
10.0
CGPA
7.0
7.5
8.0
8.5
9.0
9.5



The density plot of the CGPA shows the distribution is close to the bell shape of normal distribution and Q-Q plot for residuals close to the qqline and it seems the CGPA data can follow the normal distribution. Same as the two previous variables to test the normality "Asymptotic one-sample Kolmogorov-Smirnov test" has been used and the p_value is 0.41 and obviously is more than 0.05, then we have strong evidence to support the Normality assumption of CGPA distribution.

(8) **Chance of Admission**: Chance of admit(ranging from 0 to 1)
The Chance of Admission is the dependent variable which shows the probability of enrollment for a student. Prior to building a regression model, it is often helpful to check for normality. Although linear regression does not strictly require a normally distributed dependent variable, the model often fits better when this is true.(Lantz 2019)
Same as other numeric variables in this paper, the density plot, Q-Q plot and Kolmogorov Smirnov test will use to test the normality of data.


Histogram−Chance of Admissio
Frequency
0
10
20
30
40
50
60
0.3
0.5
0.7
0.9
Chance of Admission
0.4
0.5
0.6
0.7
0.8
0.9



The density plot shows the distribution is close to bell shape of the normal distribution and Q-Q plot for residuals close to the qqline and it seems the Chance of Admission data can follow the normal distribution. The p_value of Kolmogorov Smirnov test is 0.276 and greater than 0.05, then we have strong evidence to support the Normality assumption of dependent variable.

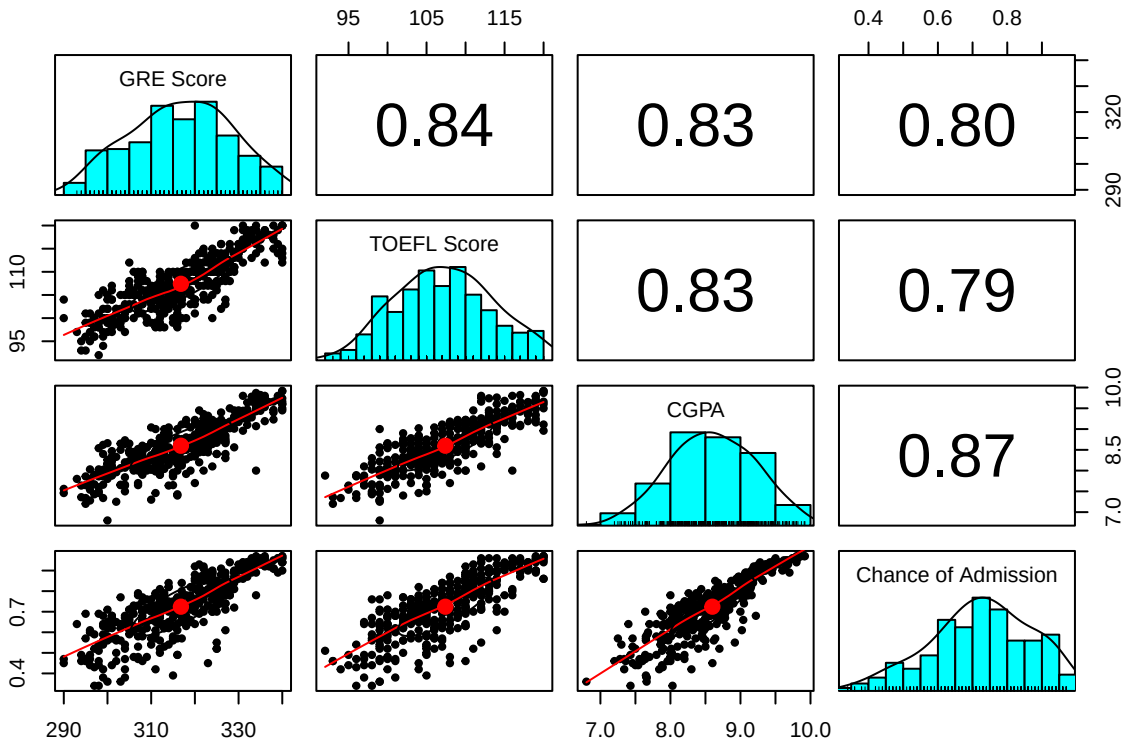
4.3 Exploring relationships among features

Before fitting a regression model to data, it can be useful to determine how the independent variables are related to the dependent variable and each other. A correlation matrix provides a quick overview of these relationships. Given a set of variables, it provides a correlation for each pairwise relationship.
The correlation matrix between numeric variables are as follows:

	GRE Score	TOEFL Score	CGPA	Chance of Admission
GRE Score	1.00	0.84	0.83	0.80
TOEFL Score	0.84	1.00	0.83	0.79
CGPA	0.83	0.83	1.00	0.87
Chance of Admission	0.80	0.79	0.87	1.00

It can also be helpful to visualize the relationship among features specially numeric

features with scatterplots.



In the $pairs.panels()$ output, the scatterplots above the diagonal are replaced with a correlarion matrix. The diagonal now contains histograms depicting the distribution of values for each feature. The scatterplots below the diagonal are presented with additional visual information.

The stretched ellipse in the scatterplots below the diagonal indicate the strong correlation among the numeric variables. The correlation between independent variable (Chance of Admission) and numeric independent variables (GRE Score, TOEFL Score and CGPA) are 80% , 79% and 87%, these high correlation can be a positive sign to start a regression model but the high correlations among the independent variables can be problematic and needs to be discussed.

A key difference between regression modeling and other machine learning approaches is that regression typically leaves feature selection and model specification to the user(Lantz 2019). Consequently it is not necessary to eliminate the features that there are high correlation among them. In this data set, the GRE score ,TOEFL Score and CGPA are in the requirements of most universities for graduate admissions and since each feature measures in a different way and evaluates vary knowledge, then considering all of them in the regression model seems reasonable.

5 Data Preparation - Creating Training and Test Datasets

It is critical to partition the data into training and testing sets when using supervised learning algorithms. Training data trains the model while testing checks whether this built model works correctly or not. As a result 90 percent of the data will use for training and 10 percent for testing. Although if the data is sorted in a random order, we can simply divide the dataset into two portions by taking the first 90 percent of records for training and the remaining 10 percent for testing. In contrast, if we are not sure about the randomness of data, obviously, this could be problematic.
This problem will solve by training the model on a random sample of the data. However, before putting it in action, a common practice is to set a **seed** value, which causes the randomization process to follow a sequence that can be replicated later. It may seem that this defeats the purpose of generating random numbers, but there is a good reason for doing it this way. Providing a seed value via the $set.seed()$ function ensures that if the analysis is repeated in the future, an identical result is obtained(Lantz 2019).
Hence, in order to provide a random sample, the $sample$ function will use to select 360 random numbers between 1 to 400, then extract these row numbers and save them in a data frame with the name $df - train$, the rest 40 rows will store in another data frame called $df - test$ for the future use to measure the accuracy of the model.

6 Building the Regression Model

To fit a linear regression model to data, the lm function can be used and at the beginning all of the independent variables are included in the model, then based on the evaluation of model performance, the improvement of model and model specification will apply.

6.1 Model 1 (Included All Of The Features)

The original model to start the regression analysis on dependent variable will include all of other features as independent variables.Therefore, Model1 (m1 in codes) will search for the linear regression relationship between the predictive variable and 7 regressors for 360 records of the training dataset , while the significant of the effect can be seen in the result of the $summary()$ function.

6.1.1 Equation - Model 1

	x
(Intercept)	0.020

	x
GRE.Score	0.168
TOEFL.Score	0.127
University.Rating	0.040
SOP	-0.014
LOR	0.146
CGPA	0.536
Research	0.037

The linear Model1 can be written like a following equation:

m1 = 0.02 + (GRE.Score $\times$ 0.168) + (TOEFL.Score $\times$ 0.127) + (University.Rating $\times$ 0.04) + (SOP $\times$ -0.014) + (LOR $\times$ 0.146) + (CGPA $\times$ 0.536) + (Research $\times$ 0.037)

6.1.2 Evaluating Model Performance - Model 1

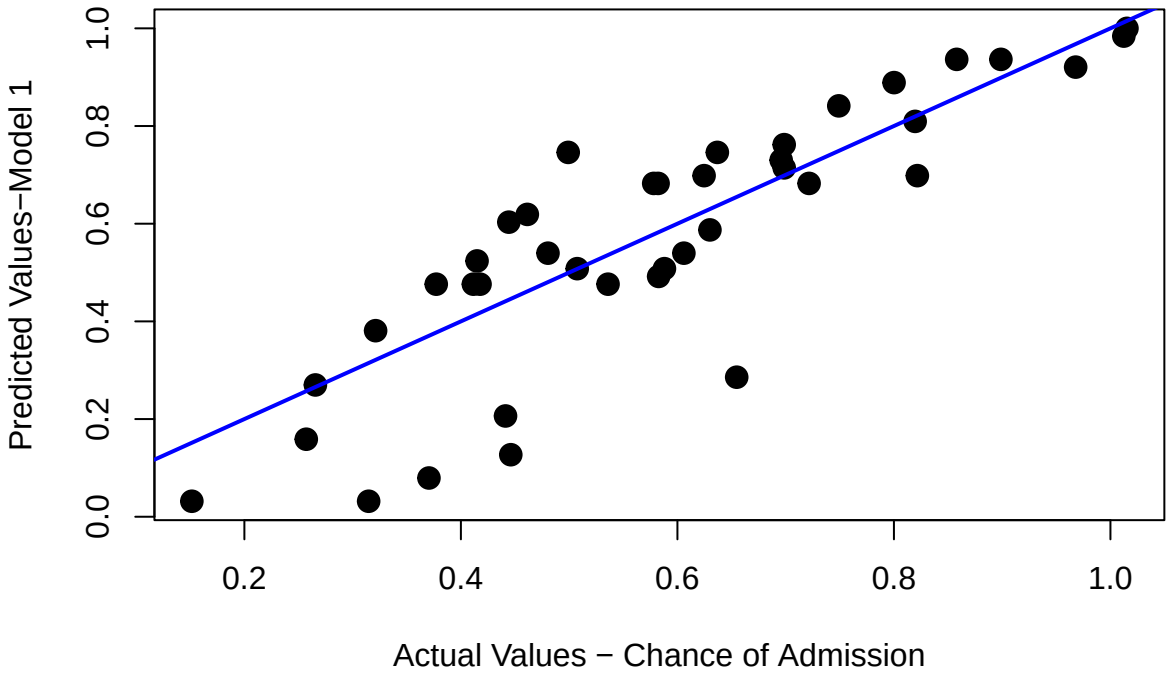
```
Call:
lm(formula = Chance.of.Admission ~ GRE.Score + TOEFL.Score +
    University.Rating + SOP + LOR + CGPA + Research, data = df_train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42266 -0.03281  0.01469  0.05319  0.25420

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.02027    0.01826   1.110 0.267831
GRE.Score          0.16770    0.04807   3.488 0.000548 ***
TOEFL.Score        0.12733    0.04883   2.608 0.009498 **
University.Rating  0.04046    0.02976   1.359 0.174872
SOP               -0.01420    0.03571  -0.398 0.691131
LOR                0.14554    0.03532   4.120 4.73e-05 ***
CGPA               0.53635    0.06195   8.658  < 2e-16 ***
Research           0.03654    0.01298   2.815 0.005153 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09663 on 352 degrees of freedom
Multiple R-squared:  0.8133,    Adjusted R-squared:  0.8096
F-statistic:   219 on 7 and 352 DF,  p-value: < 2.2e-16

     Model1 Train  Test
```

```
1      RMSE 0.096 0.136
2       MAE 0.068 0.103
3 R squared 0.813 0.860
```

Most of the coefficients in the Model1 are statistically significant, except University Rating and SOP, since their p_values are more than 0.05 which is indicate that the relationship between the dependent and these independent (University Rating and SOP) variables is not significant at the 95% certainty level. The highest coefficient is CGPA with the amount of 0.536 and it shows that based on this model, CGPA has the highest weight to effect on the Chance of Admission. After CGPA, the two other independent variables have more weights on dependent variable are Research and LOR with 0.037.
Since the model 0 is a linear regression, we might better check the **multicollinearity** between independent variables of the model with vif function. VIF is Variance Inflation Factor is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model.

	x
GRE.Score	4.66
TOEFL.Score	4.23
University.Rating	2.81
SOP	3.08
LOR	2.40
CGPA	5.26
Research	1.61

The table above shows the low multicollinearity between almost all of the independent variables (except GCPA) of the model 0 since all of the vif results are less than 5. The VIF for GCPA is equal to 5.26 and it is the indication of moderate collinearity between this variable and other features.
R^2 is equal to 0.81 and it means almost 81% of the variation in the independent variables can be explained by this model. Because models with more features always explain more variation, the Adjusted R-squared value corrects R-squared by penalizing models with a large number of independent variables.In this model, we have Adjusted R-squared with almost the same percent as the R-Squared 80% which is actually quite good.
Other items that need to calculates in this step is the correlation between the predicted values based on the regression model and actual values of the dependent variable. For this purpose another column is created including the predicted values of 'Chance of Admission' with the name of $prdct\_m1$ using $predict$ function. Then this column will apply in cor function to calculate the mentioned correlation. Obviously our interest is closer number to 1 which shows how best the regression model can predict the 'Chance of Admission'. The correlation - Model 0 = 0.86



The correlation is almost 86% and as we can see in the above plot the dots are vary close to the blue line. It means in general the predicted values of the model 1 is almost close to the actual values in the data.

Best Subsets Regression

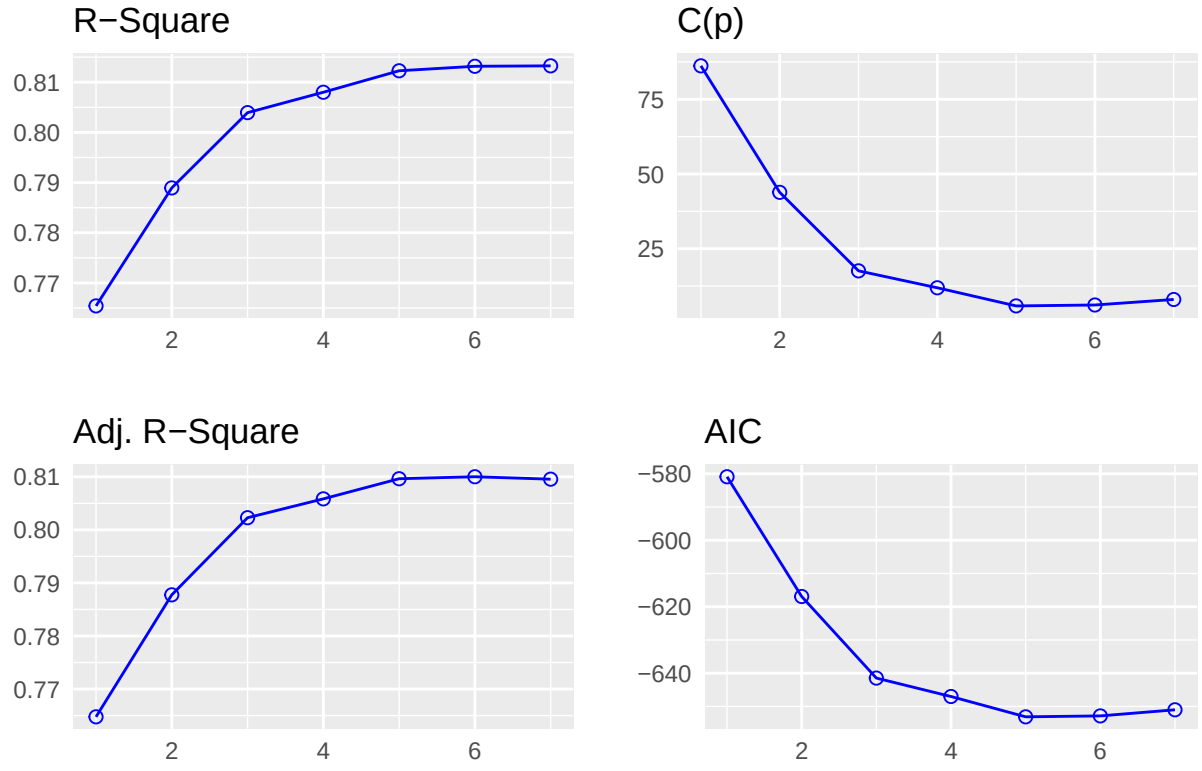
Model Index	Predictors
1	CGPA
2	GRE.Score CGPA
3	GRE.Score LOR CGPA
4	GRE.Score LOR CGPA Research
5	GRE.Score TOEFL.Score LOR CGPA Research
6	GRE.Score TOEFL.Score University.Rating LOR CGPA Research
7	GRE.Score TOEFL.Score University.Rating SOP LOR CGPA Research

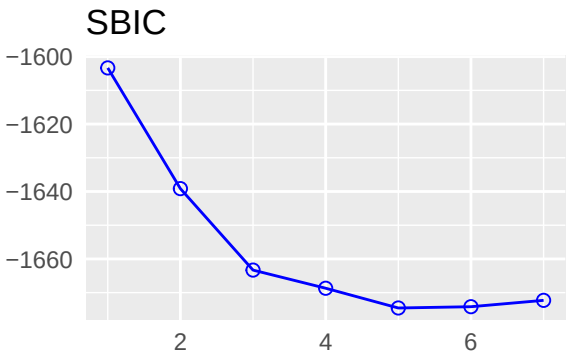
Subsets Regression Summary

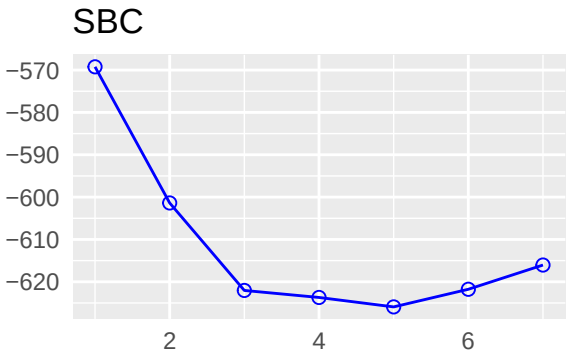
Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SB

1	0.7654	0.7648	0.7629	86.1903	-580.8912	-1603.3395	-569.
2	0.7889	0.7878	0.7854	43.8581	-616.9234	-1639.1482	-601.
3	0.8039	0.8023	0.7996	17.5779	-641.4676	-1663.3120	-622.
4	0.8080	0.8058	0.8022	11.9160	-647.0093	-1668.6981	-623.
5	0.8123	0.8096	0.8057	5.8482	-653.1253	-1674.5530	-625.
6	0.8132	0.8100	0.8057	6.1581	-652.8488	-1674.1738	-621.
7	0.8133	0.8096	0.8041	8.0000	-651.0105	-1672.2836	-616.

AIC: Akaike Information Criteria
SBIC: Sawa's Bayesian Information Criteria
SBC: Schwarz Bayesian Criteria
MSEP: Estimated error of prediction, assuming multivariate normality
FPE: Final Prediction Error
HSP: Hocking's Sp
APC: Amemiya Prediction Criteria







6.2 Model 2 (Including the Features with High Significant Levels)

The next step is to improve the model performance. Due to the low significance level of two coefficients (SOP and University Rating) in the Model 1, In order to improve the model performance, the linear regression will run included all of the regressors except 'SOP' and 'University Rating' .

6.2.1 Equation - Model 2

	x
(Intercept)	0.014
GRE.Score	0.171
TOEFL.Score	0.135
LOR	0.153
CGPA	0.551
Research	0.037

The linear Model1 can be written like a following equation:

m2 = 0.014 + (GRE.Score $\times$ 0.171) + (TOEFL.Score $\times$ 0.135) + (LOR $\times$ 0.153) + (CGPA $\times$ 0.551) + (Research $\times$ 0.037)

6.2.2 Evaluating Model Performance - Model 2

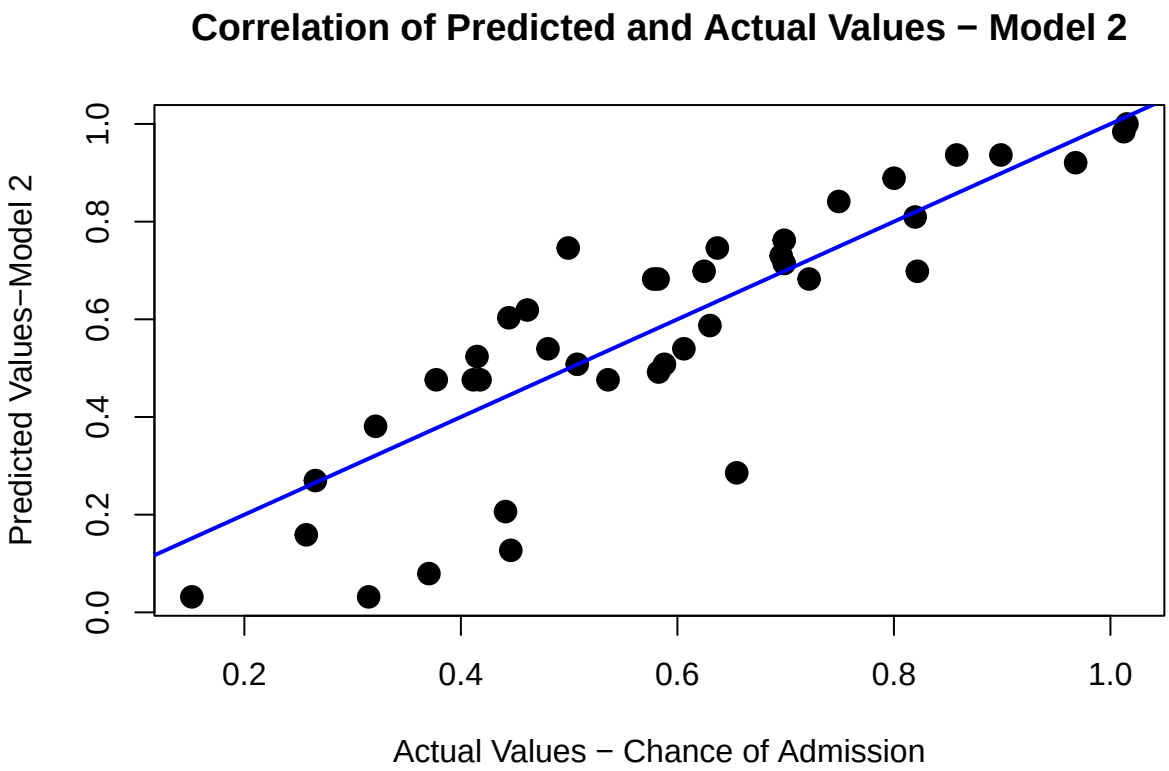
```
Call:
lm(formula = Chance.of.Admission ~ GRE.Score + TOEFL.Score +
    LOR + CGPA + Research, data = df_train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42356 -0.03431  0.01505  0.05502  0.25605

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01355    0.01755   0.772 0.440799
GRE.Score    0.17142    0.04774   3.591 0.000376 ***
TOEFL.Score  0.13496    0.04750   2.841 0.004757 **
LOR          0.15293    0.03094   4.942  1.2e-06 ***
CGPA         0.55102    0.05945   9.269  < 2e-16 ***
Research     0.03680    0.01292   2.849 0.004640 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09661 on 354 degrees of freedom
Multiple R-squared:  0.8123,    Adjusted R-squared:  0.8096
F-statistic: 306.4 on 5 and 354 DF,  p-value: < 2.2e-16


     Model2 Train  Test
1      RMSE 0.096 0.136
2       MAE 0.068 0.104
3 R squared 0.812 0.861
```

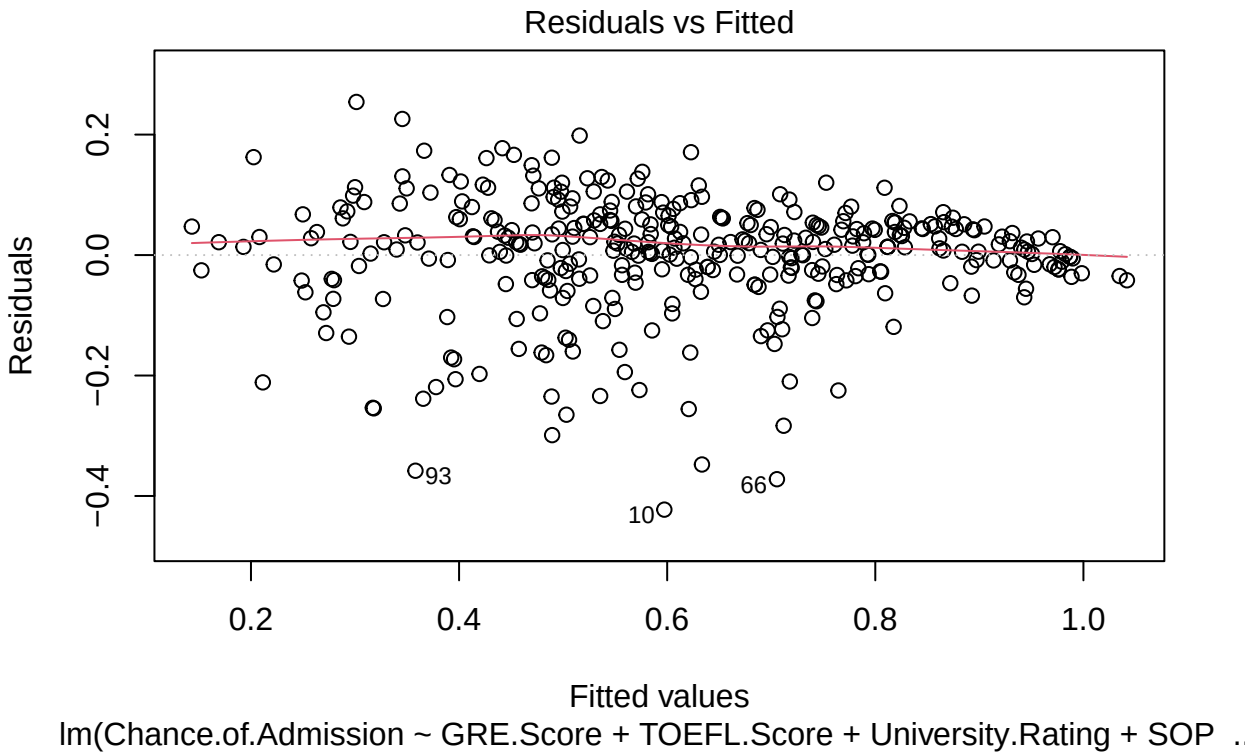
All of the coefficients in the Model1 are statistically significant,since their p_values are less than 0.05 which is indicate that the relationship between independent variable and all of the independent variables included in the Model1 are more likely linear. The highest coefficient is CGPA with the amount of 0.812 and it shows that based on this model, CGPA has the highest weight to effect on the 'Chance of Admission'. After CGPA, the two other independent variables have more weights on dependent variable are Research and LOR with NA.

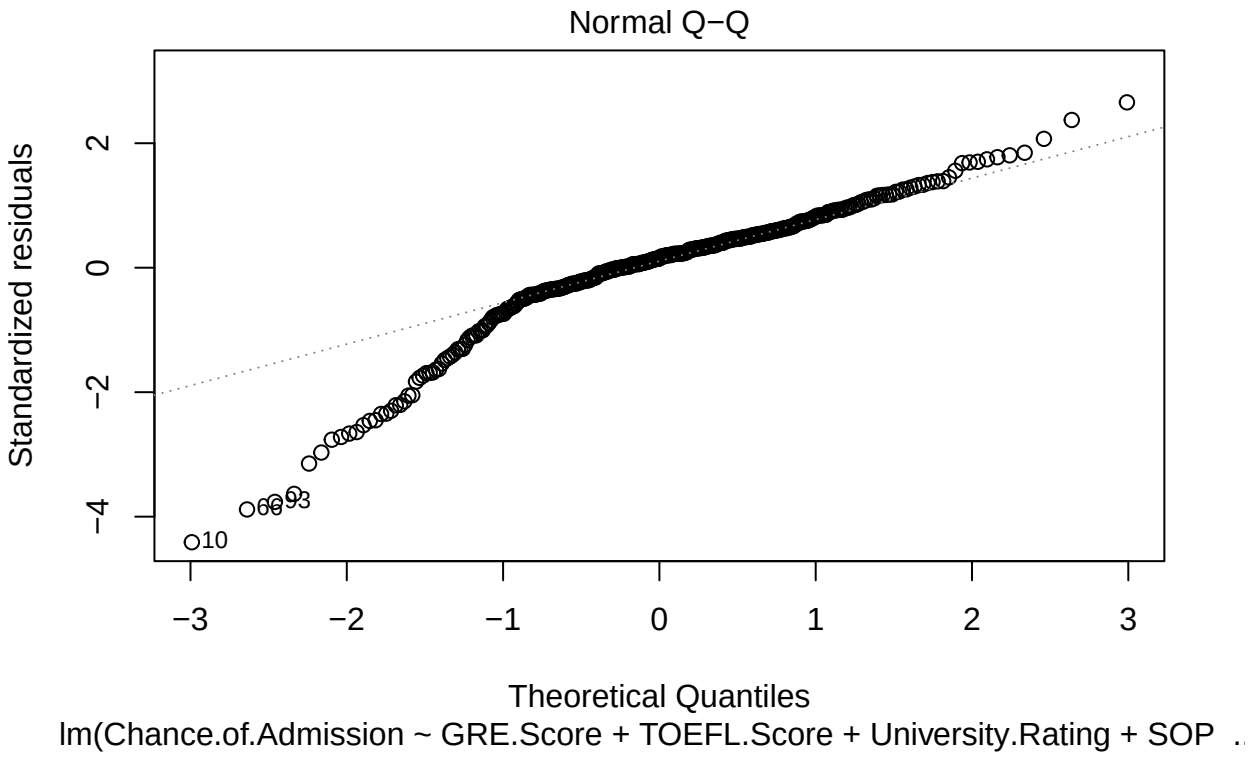
Since the model 1 is a linear regression, we might better check the **multicollinearity** between independent variables of the model with vif function.

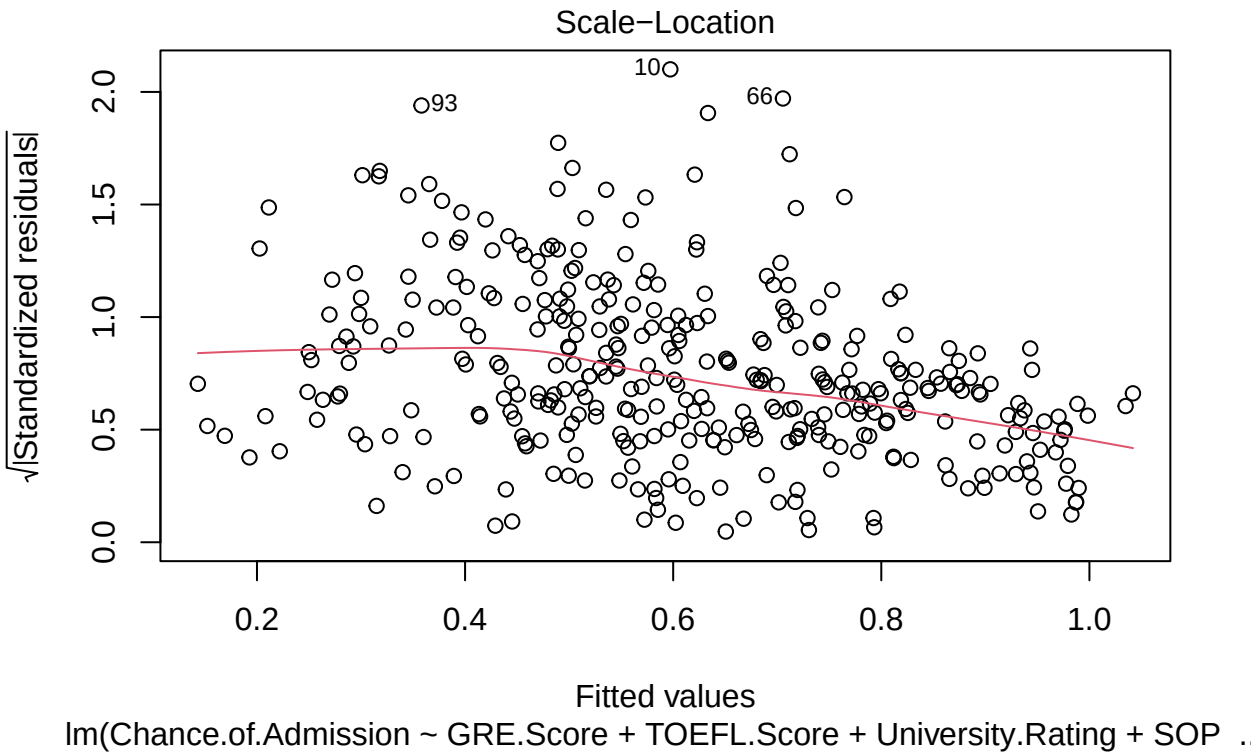
	x
GRE.Score	4.60
TOEFL.Score	4.00
LOR	1.84
CGPA	4.85
Research	1.59

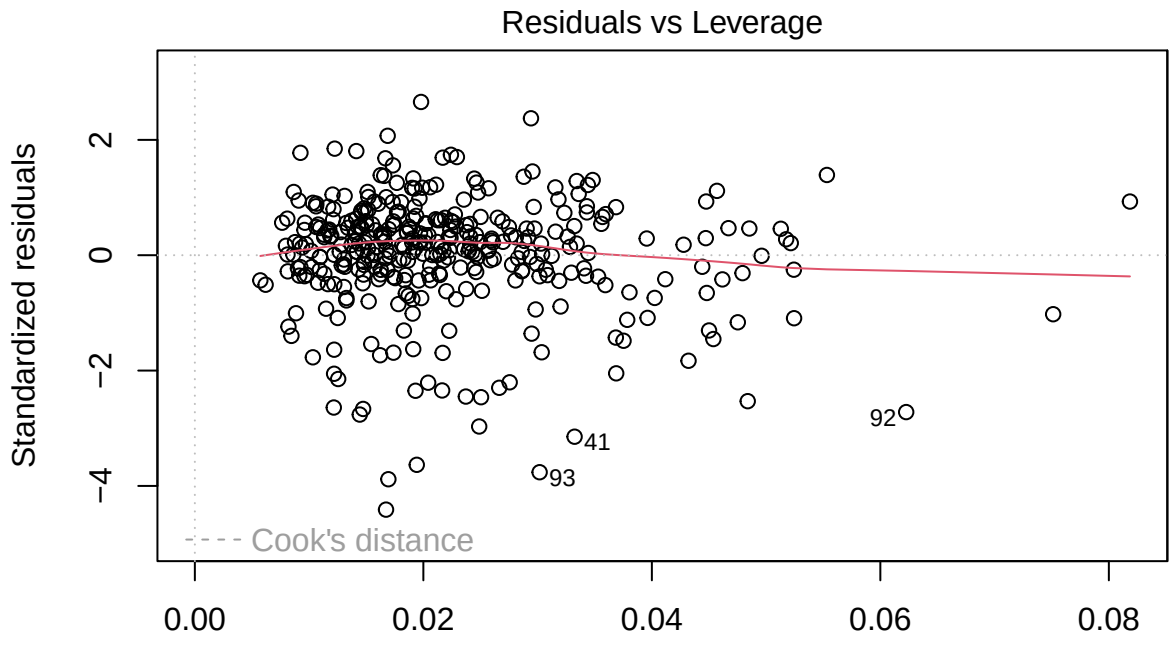
The table above shows the low multicollinearity between independent variables of the model1 since all of the vif results are less than 5.
R^2 is equal to 0.80 and it means almost 80% of the variation in the independent variables can be explained by this model.In this model, the Adjusted R-squared with 80% is almost equal to the R-Squared, which is actually quite good.
Similarly, we need to calculate the correlation between the predicted values based on the regression model and actual values of the dependent variable. For this purpose another column is created including the predicted values of 'Chance of Admission' with the name of $prdct\_m2$ using $predict$ function. Then this column will apply in cor function to calculate the mentioned correlation. The correlation - Model 1 = 0.86




Residuals vs Fitted
Residuals
0.2
0.0
−0.2
−0.4
93
66
10
0.2
0.4
0.6
0.8
1.0
Fitted values
lm(Chance.of.Admission ~ GRE.Score + TOEFL.Score + University.Rating + SOP


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Chance.of.Admission ~ GRE.Score + TOEFL.Score + University.Rating + SOP
10
93


Scale−Location
10
93
66
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0.2
0.4
0.6
0.8
1.0
Fitted values
lm(Chance.of.Admission ~ GRE.Score + TOEFL.Score + University.Rating + SOP



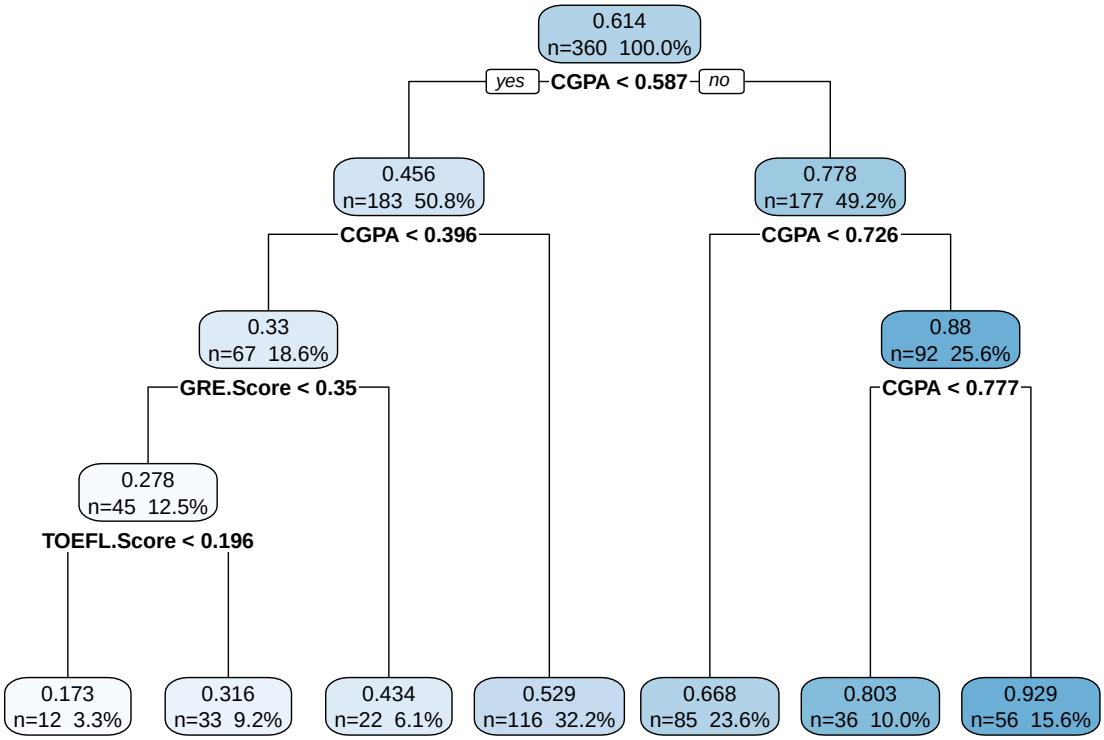
The correlation is almost 90% and as we can see in the above plot the dots are vary close to the blue line. However, the correlation only measures how strongly the predictions are related to the true value; it is not a measure of how far off the predictions were from the true values.

```
Call:
lm(formula = Chance.of.Admission ~ GRE.Score + LOR + CGPA + Research,
    data = df_train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42817 -0.03607  0.01336  0.05591  0.26741

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01635    0.01770   0.924  0.35625
GRE.Score    0.23024    0.04344   5.300 2.04e-07 ***
LOR          0.15931    0.03117   5.111 5.24e-07 ***
CGPA         0.61501    0.05556  11.070  < 2e-16 ***
Research     0.03574    0.01304   2.741  0.00643 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09757 on 355 degrees of freedom
Multiple R-squared:  0.808, Adjusted R-squared:  0.8058
F-statistic: 373.5 on 4 and 355 DF,  p-value: < 2.2e-16


[1] 0.1832237


[1] 0.1366429


[1] 0.1062831


[1] 0.859763


GRE.Score       LOR      CGPA  Research
 3.731673  1.832686  4.151018  1.592924
```

7 Building the Regression Trees

Trees that can perform numeric prediction are called regression trees. Trees for numeric prediction are built in much the same way as they are for classification. Beginning at the root node, the data is partitioned using a divide and conquer strategy according to the feature that will result in the greatest increase in homogeneity in the outcome after a split is performed. In classification, homogeneity is measured by entropy. This is undefined for numeric data. Instead for numeric decision trees, homogeneity is measured by statistics such as variance, standard deviation, or absolute deviation from the mean(**mlr?**).
Linear regression and logistic regression models fail in situations where the relationship between features and outcome is nonlinear or where features interact with each other. Time to shine for the decision tree! Tree based models split the data multiple times according to certain cutoff values in the features. Through splitting, different subsets of the dataset are created, with each instance belonging to one subset. The final subsets are called terminal or leaf nodes and the intermediate subsets are called internal nodes or split nodes. To predict the outcome in each leaf node, the average outcome of the training data in this node is used. Trees can be used for classification and regression.




```
      CGPA       GRE.Score     TOEFL.Score University.Rating
13.47790474     8.98286431      8.68768724      6.59069719
       SOP            LOR        Research
 6.50358926     5.78847432      0.08176368
```

```
[1] 0.1739606
```

```
[1] 0.160767
```

```
[1] 0.1117959
```

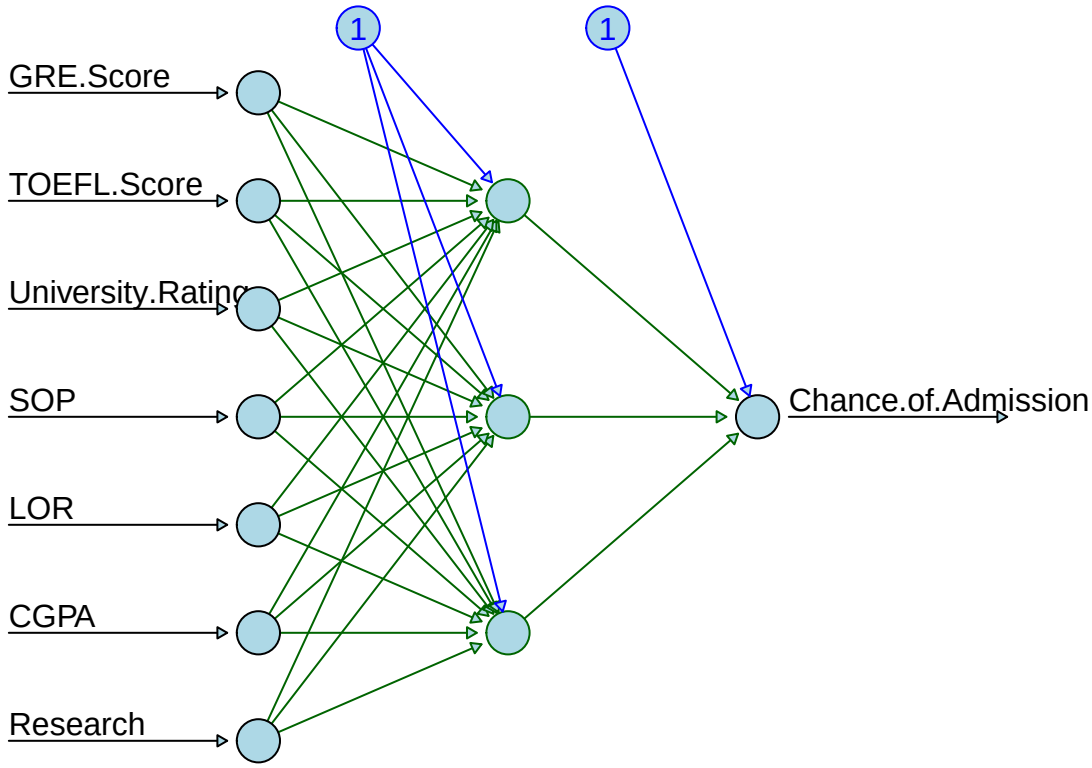
```
[1] 0.7992228
```

8 Neural Network Medels

A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure

that resembles the human brain.Neural networks can be applied to a broad range of problems and can assess many different types of input, including images, videos, files, databases, and more. They also do not require explicit programming to interpret the content of those inputs.

8.1 Neural Network Model (3 hidden nodes and 1 hidden layer)



Error: 1.567641 Steps: 1470

	[,1]
error	1.568
reached.threshold	0.010
steps	1470.000
Intercept.to.1layhid1	0.645
GRE.Score.to.1layhid1	-0.725
TOEFL.Score.to.1layhid1	0.631
University.Rating.to.1layhid1	1.121
SOP.to.1layhid1	0.479
LOR.to.1layhid1	-0.874
CGPA.to.1layhid1	-2.417
Research.to.1layhid1	-0.005
Intercept.to.1layhid2	-2.981

```
GRE.Score.to.1layhid2                0.235
TOEFL.Score.to.1layhid2              0.669
University.Rating.to.1layhid2        0.689
SOP.to.1layhid2                      0.592
LOR.to.1layhid2                      0.016
CGPA.to.1layhid2                     0.333
Research.to.1layhid2                 0.863
Intercept.to.1layhid3               -3.067
GRE.Score.to.1layhid3               -0.659
TOEFL.Score.to.1layhid3             10.404
University.Rating.to.1layhid3        7.357
SOP.to.1layhid3                     -5.447
LOR.to.1layhid3                     -0.057
CGPA.to.1layhid3                     4.581
Research.to.1layhid3               -10.030
Intercept.to.Chance.of.Admission    -0.327
1layhid1.to.Chance.of.Admission     -3.109
1layhid2.to.Chance.of.Admission      6.063
1layhid3.to.Chance.of.Admission      0.848
```

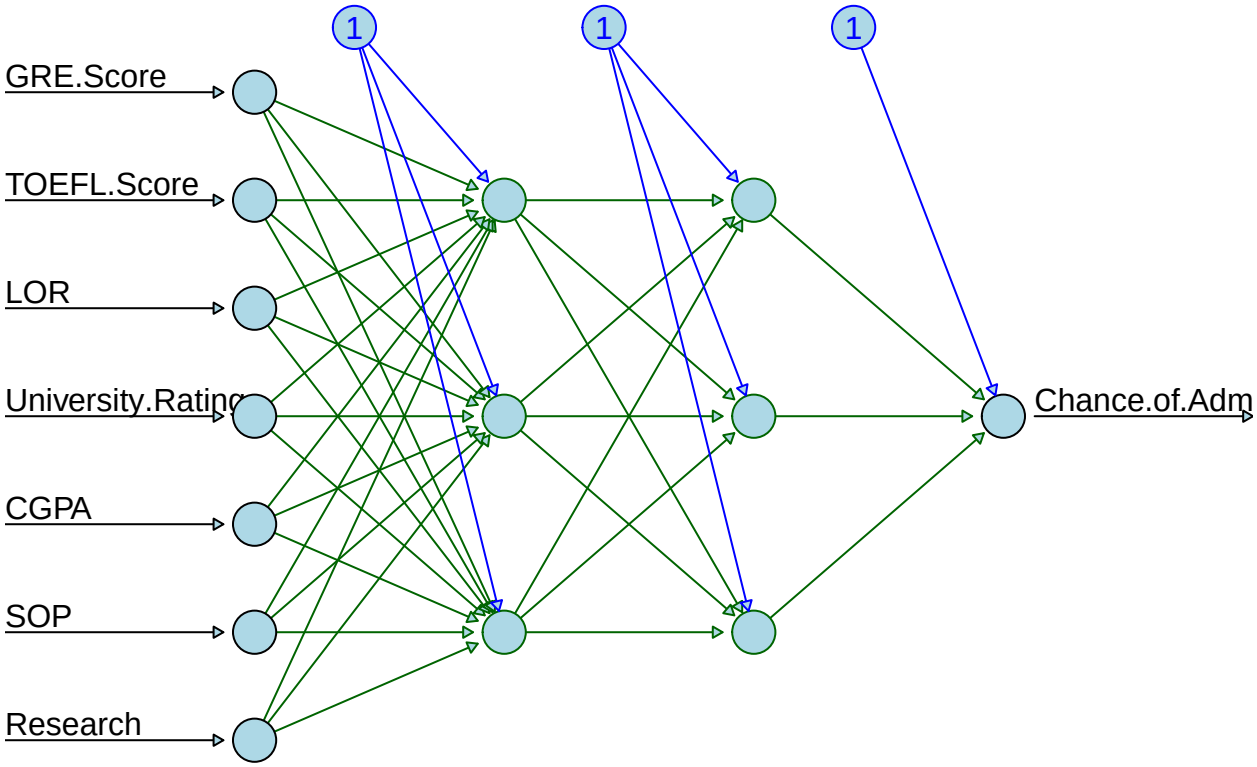
```
[1] 0.1833099
```

```
[1] 0.1381831
```

```
[1] 0.1015629
```

```
          [,1]
[1,] 0.8587673
```

8.2 Neural Network Model (3 hidden nodes and 2 hidden layers)



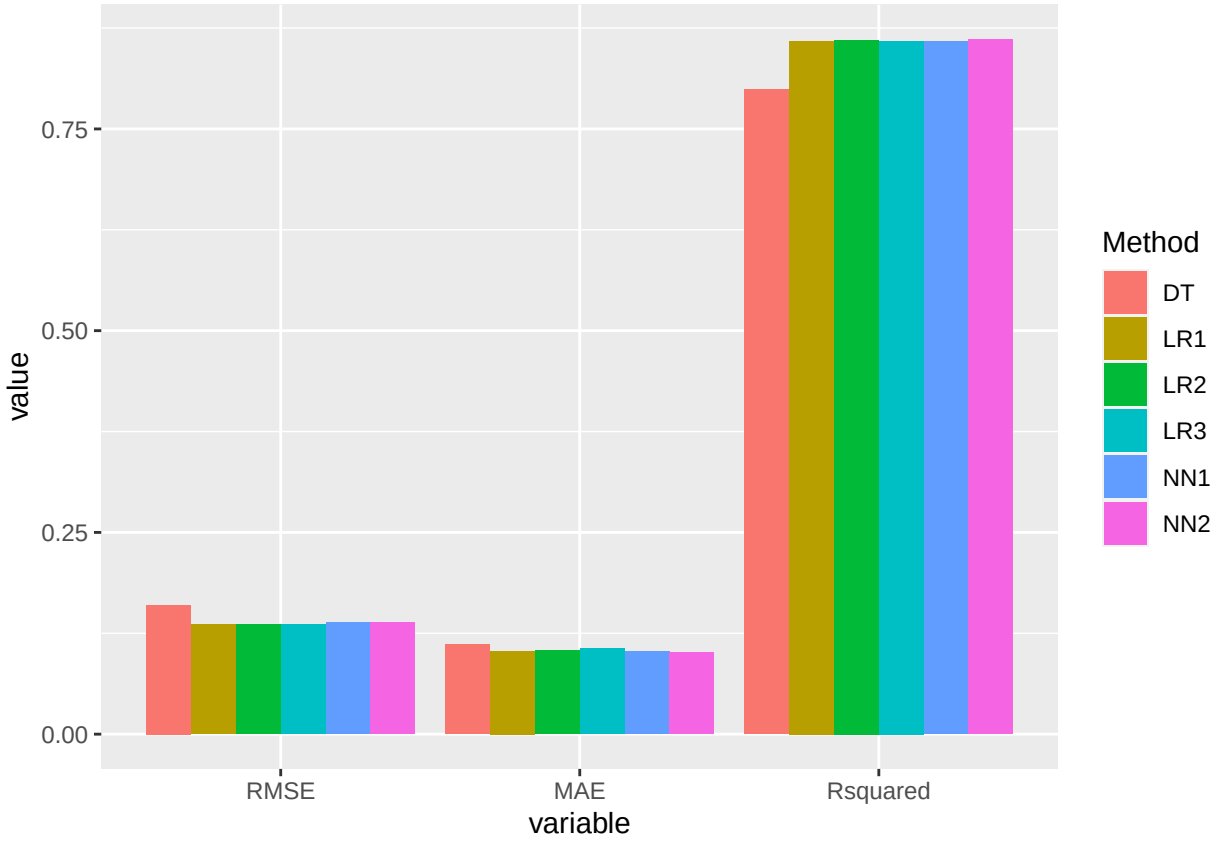
Error: 1.501379 Steps: 986

	[,1]
error	1.501
reached.threshold	0.008
steps	986.000
Intercept.to.1layhid1	1.897
GRE.Score.to.1layhid1	-3.210
TOEFL.Score.to.1layhid1	-20.777
LOR.to.1layhid1	-3.308
University.Rating.to.1layhid1	8.615
CGPA.to.1layhid1	-13.445
SOP.to.1layhid1	3.509
Research.to.1layhid1	0.954
Intercept.to.1layhid2	-3.187
GRE.Score.to.1layhid2	-0.557
TOEFL.Score.to.1layhid2	1.800
LOR.to.1layhid2	0.575
University.Rating.to.1layhid2	-0.328
CGPA.to.1layhid2	2.305
SOP.to.1layhid2	-0.527

```
Research.to.1layhid2              -0.011
Intercept.to.1layhid3              2.352
GRE.Score.to.1layhid3             -1.614
TOEFL.Score.to.1layhid3            1.353
LOR.to.1layhid3                   -0.280
University.Rating.to.1layhid3     -0.824
CGPA.to.1layhid3                  -1.483
SOP.to.1layhid3                   -0.677
Research.to.1layhid3              -0.279
Intercept.to.2layhid1             -0.551
1layhid1.to.2layhid1             -36.357
1layhid2.to.2layhid1               2.295
1layhid3.to.2layhid1              -1.475
Intercept.to.2layhid2             -0.507
1layhid1.to.2layhid2              54.951
1layhid2.to.2layhid2              -1.458
1layhid3.to.2layhid2               1.802
Intercept.to.2layhid3              0.326
1layhid1.to.2layhid3              59.449
1layhid2.to.2layhid3             -25.840
1layhid3.to.2layhid3              25.352
Intercept.to.Chance.of.Admission   0.538
2layhid1.to.Chance.of.Admission    5.686
2layhid2.to.Chance.of.Admission   -1.874
2layhid3.to.Chance.of.Admission   -0.780

[1] 0.1922341

[1] 0.1338376

[1] 0.1013352

          [,1]
[1,] 0.8616211
```

9 Conclusion

```
Attaching package: 'ggplot2'

The following objects are masked from 'package:psych':

    %+%, alpha
```



Though traditional regression methods are typically the first choice for numeric prediction tasks, in some cases, numeric decision trees offer distinct advantages. For instance, decision trees maybe suited for tasks with many features or many complex, nonlinear relationships among features and the outcome; These situations present challenges for regression. Regression modeling also makes assumptions about the data that are often violated in real-world data; this is not the case for trees(Lantz 2019).

When looking at being admitted to a graduate school, the criteria that are considered can vary. If only one criterion were to be considered, CGPA is the most important. If two will be considered together it would be the GRE Score and the CGPA. When three regressors are used, those regressors are (in order) GRE, LOR, and CGPA. If the model is to contain four regressors, in order of significance, those regressors would be GRE, TOEFEL, LOR, and CGPA. In the final model, all remaining regressors will be included.

Looking at the model indices and that subsets regression summary as one moves from the top of the table to the bottom, each successive model seems to be a little bit better. The Rˆ2 adjusted values are increasing, the C(p) value decrease, and the AIC values decrease. The increase from model 2 through model 5 in the R^2 adjusted value is minimal. Furthermore, the values of C(p) and AIC do decrease, but relative to the data in this column, we must ask ourselves, what amount of drop is enough? The decrease in successive values continues, but is not necessarily significant.

Thus, to continue our analysis below we will run the summary on each model to see if there is one that stands out and is better than the rest.

After exploratory and statistical analysis modeling, we identified the best fit model for our data. The model determined the best fit was model 3 using all the regressors. This data was transformed using the logistical transfrmation. After running the summary for this model, we are left with this model equation.

This data set was not very small but I think a research topic such as this would benefit from a larger data frame. For future research I might suggest gathering more data and possibly individualizing the schools. A researcher could determine which regressor is most important for that particular school as opposed to in a general sense.

DT Linear regression and logistic regression models fail in situations where the relationship between features and outcome is nonlinear or where features interact with each other.

References

Acharya, Mohan S. 2018. “Graduate Admission 2.” https://www.kaggle.com/datasets.

Lantz, Brett. 2019. *Machine Learning with r*. Third edition. Birmingham B3 2PB, UK: Packt, ISBN = 978-1-78829-586-4.

Nesbitt, Heather. 2021. “Conceptualizing Thriving.” https://www.frontiersin.org/articles/10.3389/feduc.2021.704135/full.

wiki. n.d.a. “Application Essay.” https://en.wikipedia.org/wiki/Application_essay.

———. n.d.b. “College and University Rankings.” https://en.wikipedia.org/wiki/College_and_university_rankings.

———. n.d.c. “Grading in Education.” https://en.wikipedia.org/wiki/Grading_in_education.

———. n.d.d. “Graduate Record Examinations.” https://en.wikipedia.org/wiki/Graduate_Record_Examinations.

———. n.d.e. “Letter of Recommendation.” https://en.wikipedia.org/wiki/Letter_of_recommendation.

———. n.d.f. “Test of English as a Foreign Language.” https://en.wikipedia.org/wiki/Test_of_English_as_a_Foreign_Language.