About Dataset (copy pasted from Kaggle)

Dataset contains under given important parameters which are considered mainly during application for Masters Programs.

Parameters description:

GRE Scores (out of 340)

TOEFL Scores (out of 120)

University Rating (out of 5)

Statement of Purpose -(SOP) Strength (out of 5)

Letter of Recommendation-(LOR) Strength (out of 5)

Undergraduate GPA-CGPA (out of 10)

Research Experience (either 0 or 1)

Chance of Admit (ranging from 0 to 1)

Sports Involvement (either 0 or 1)

Data Preprocessing

First, all the libraries used were defined, and then we read the CSV file after setting the working directory.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.0.4
## ── Conflicts ────────────────────────────────────────── 
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(ggplot2)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(here)
df <- read_csv(data1)
```

```
## Rows: 400 Columns: 9
## ── Column specification
──────────────────────────────────────────────────────
## Delimiter: ","
## dbl (9): GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA,
Research...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

No Na values are present in the file since the dim function gave the same output The aim of removing rows with missing values is to ensure our data is complete, avoid errors and improve the quality of our data

```
dim(df)
```

```
## [1] 400   9
```

```
df <- drop_na(df)
dim(df)
```

```
## [1] 400   9
```

Here is an overview of the predictors and the response variable “Chance of Admit” we are working with

```
colnames(df)
```

```
## [1] "GRE Score"         "TOEFL Score"       "University Rating"
## [4] "SOP"               "LOR"               "CGPA"
## [7] "Research"          "Sport Involvement" "Chance of Admit"
```

All predictors are of numerical type Sort Involvement and Research should me converted to factor

```
str(df)
```

```
## tibble [400 × 9] (S3: tbl_df/tbl/data.frame)
##  $ GRE Score        : num [1:400] 337 324 316 322 314 330 321 308 302 323
...
##  $ TOEFL Score      : num [1:400] 118 107 104 110 103 115 109 101 102 108
```

```
...
##  $ University Rating: num [1:400] 4 4 3 3 2 5 3 2 1 3 ...
##  $ SOP              : num [1:400] 4.5 4 3 3.5 2 4.5 3 3 2 3.5 ...
##  $ LOR              : num [1:400] 4.5 4.5 3.5 2.5 3 3 4 4 1.5 3 ...
##  $ CGPA             : num [1:400] 9.65 8.87 8 8.67 8.21 9.34 8.2 7.9 8 8.6
...
##  $ Research         : num [1:400] 1 1 1 1 0 1 1 0 0 0 ...
##  $ Sport Involvement: num [1:400] 1 1 1 1 1 1 1 1 0 0 ...
##  $ Chance of Admit  : num [1:400] 0.92 0.76 0.72 0.8 0.65 0.9 0.75 0.68
0.5 0.45 ...
```

To check if the values are withing the range specified in the description of the dataset. We check specifically the Min and Max

```
summary(df)
```

```
##    GRE Score      TOEFL Score    University Rating      SOP
##  Min.   :290.0   Min.   : 92.0   Min.   :1.000     Min.   :1.0
##  1st Qu.:308.0   1st Qu.:103.0   1st Qu.:2.000     1st Qu.:2.5
##  Median :317.0   Median :107.0   Median :3.000     Median :3.5
##  Mean   :316.8   Mean   :107.4   Mean   :3.087     Mean   :3.4
##  3rd Qu.:325.0   3rd Qu.:112.0   3rd Qu.:4.000     3rd Qu.:4.0
##  Max.   :340.0   Max.   :120.0   Max.   :5.000     Max.   :5.0
##       LOR             CGPA          Research      Sport Involvement
##  Min.   :1.000   Min.   :6.800   Min.   :0.0000   Min.   :0.000
##  1st Qu.:3.000   1st Qu.:8.170   1st Qu.:0.0000   1st Qu.:1.000
##  Median :3.500   Median :8.610   Median :1.0000   Median :1.000
##  Mean   :3.453   Mean   :8.599   Mean   :0.5475   Mean   :0.815
##  3rd Qu.:4.000   3rd Qu.:9.062   3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :5.000   Max.   :9.920   Max.   :1.0000   Max.   :1.000
##  Chance of Admit
##  Min.   :0.3400
##  1st Qu.:0.6400
##  Median :0.7300
##  Mean   :0.7244
##  3rd Qu.:0.8300
##  Max.   :0.9700
```

We did not detect any duplicates in our dataset. However, if duplicates were present, it is advisable to remove them, as they can introduce correlated error terms, leading to an undeserved sense of confidence in our model in subsequent analyses.

```
df[duplicated(df), ]
```

```
## # A tibble: 0 × 9
## # ℹ 9 variables: GRE Score <dbl>, TOEFL Score <dbl>, University Rating
<dbl>,
## #   SOP <dbl>, LOR <dbl>, CGPA <dbl>, Research <dbl>, Sport Involvement
<dbl>,
## #   Chance of Admit <dbl>
```

```r
View(df)
```

The values are logical we do not have any sentinel or unexpected random values

```r
unique(df$`University Rating`)
```

```
## [1] 4 3 2 5 1
```

```r
unique(df$SOP)
```

```
## [1] 4.5 4.0 3.0 3.5 2.0 5.0 1.5 1.0 2.5
```

```r
unique(df$LOR)
```

```
## [1] 4.5 3.5 2.5 3.0 4.0 1.5 2.0 5.0 1.0
```

```r
unique(df$Research)
```

```
## [1] 1 0
```

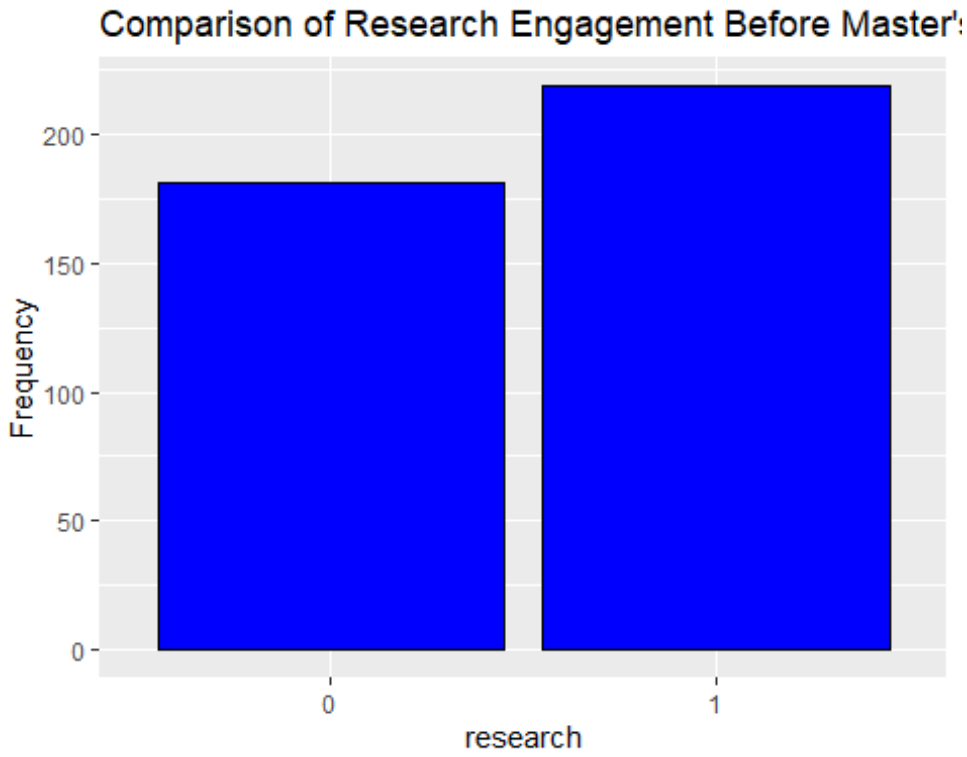
```r
unique(df$`Sport Involvement`)
```

```
## [1] 1 0
```

Data Visualization

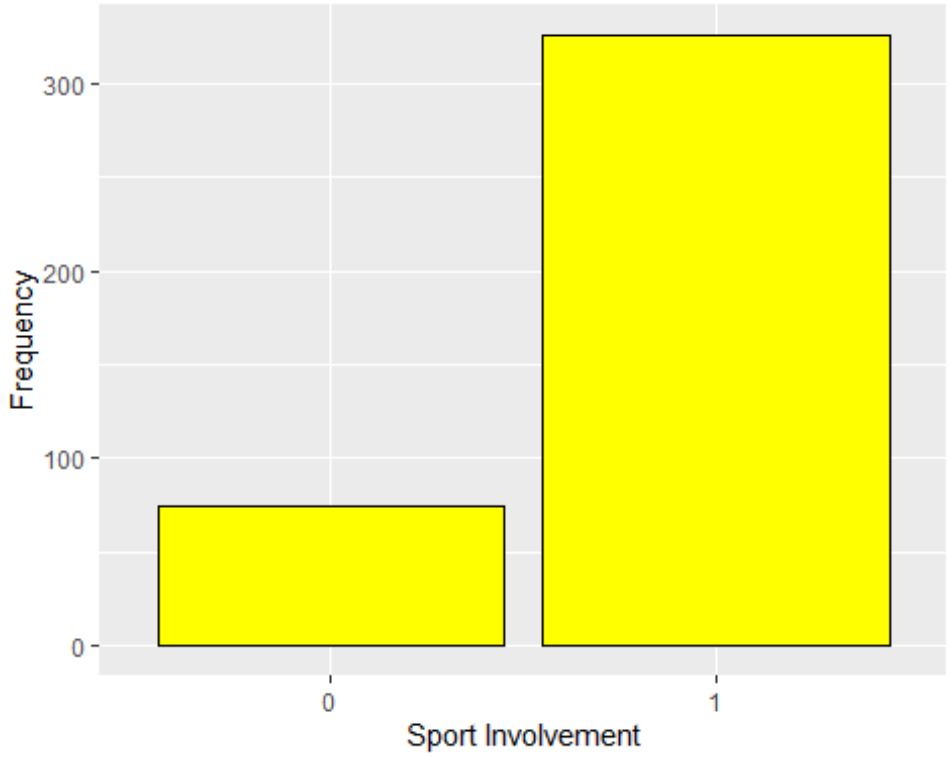
The purpose of the plot is to visualize the frequency of individuals who engage in research before applying to a master's program. The result indicates that the frequencies of those who do research and those who don't do not differ remarkably

```r
ggplot(data = df, aes(x = as.factor(Research))) +
  geom_bar(fill = "blue", color = "black") +
  labs(title = "Comparison of Research Engagement Before Master's Program
Application", x = "research", y = "Frequency")
```



The purpose of the plot is to visualize the frequency of individuals who participate in a certain sport before applying to a master's program. The result shows that the frequency of those who are involved in the sport is significantly higher compared to those who are not.

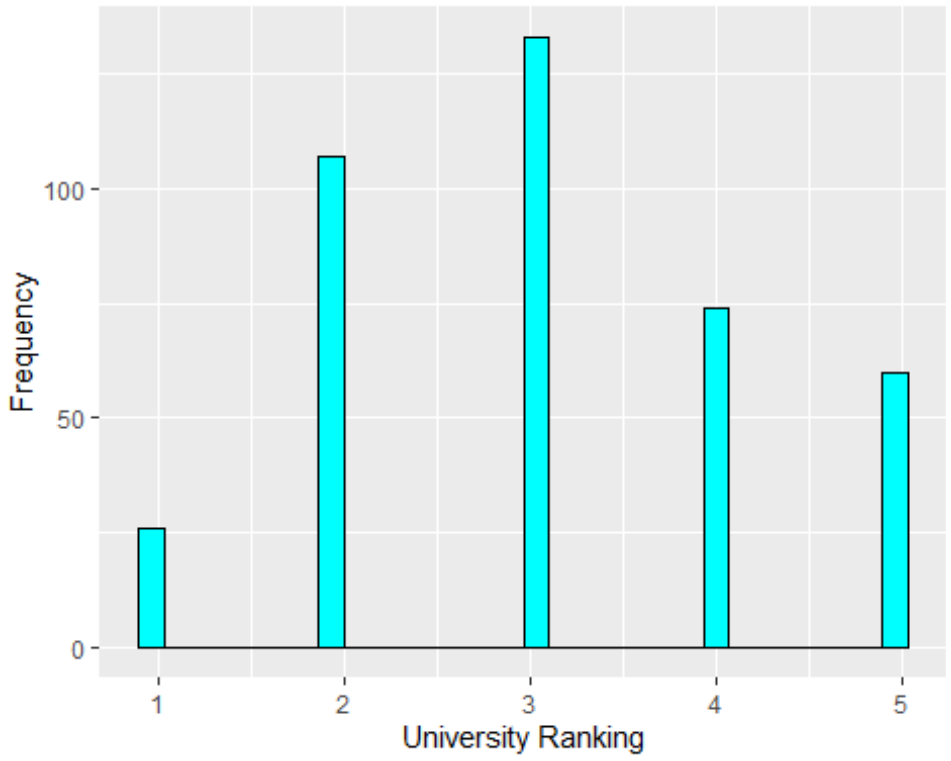
```
ggplot(data = df, aes(x = as.factor(`Sport Involvement`))) +
  geom_bar(fill = "yellow", color = "black") +
  labs(x = "Sport Involvement", y = "Frequency")
```



Most of the universities to which students in this dataset apply for master's programs have a rating of 3.

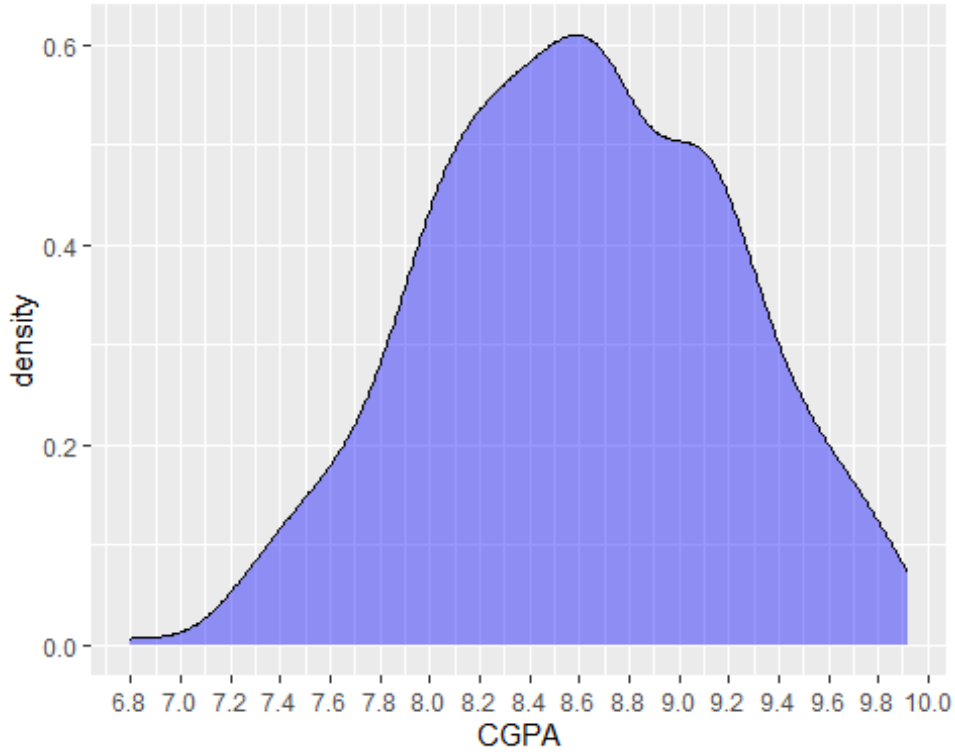
```
ggplot(data = df, aes(x = `University Rating`)) +
  geom_histogram(fill = "cyan", color = "black") +
  labs(x = "University Ranking", y = "Frequency")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Students with a CGPA between 8.5 and 8.7 are the most likely to apply for a master's program.

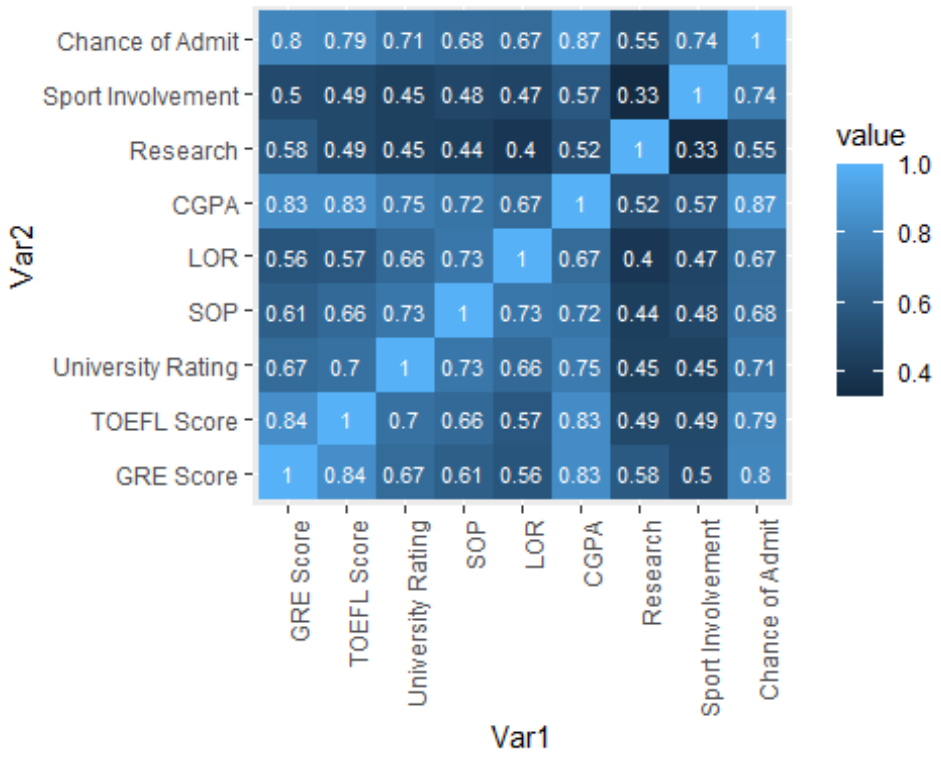
```
ggplot(data = df, aes(x = CGPA)) +
  geom_density(fill = "blue", color = "black", alpha = 0.4) +
  scale_x_continuous(breaks = seq(6.8, 10, by = 0.2))
```



The Correlation matrix will not work if research and sport involvement are categorical We will initially perform the correlation analysis and subsequently convert these predictors into categorical variables. Additionally, the correlation matrix will provide insights into the potential plots we can generate.

The minimum correlation value 0.33 is between research and sport involvement which is still significant We can conclude that the predictors are correlated with each other and each predictor is correlated with the response “chance of getting admit”

```
cormatrix <- round(cor(df),2)
melted <- melt(cormatrix)
ggplot(melted,aes(Var1,Var2,fill=value))+
  geom_tile()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  geom_text(aes(Var2, Var1, label = value),
            color = "white", size = 3)
```



From the correlation matrix we can hypothesize that a relationship is present between each predictor and the response “Chance of admit”.

Convert categorical predictors to factor

```
df$Research <- as.factor(df$Research)
df$`Sport Involvement` <- as.factor(df$`Sport Involvement`)
```

Simple Linear Models

Next, we will create individual plots for each predictor variable against the response variable for the purpose of univariate visualization. We anticipate significant results based on the previously generated correlation matrix.

GRE Score against Chance of Admit A linear relationship is present in this plot between the two attributes Hence the chance of admission increases with the GRE score

```
ggplot(data=df)+
  geom_point(mapping=aes(x=`GRE Score`,y=`Chance of Admit`))+
  geom_smooth(mapping=aes(x=`GRE Score`,y=`Chance of Admit`))

## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

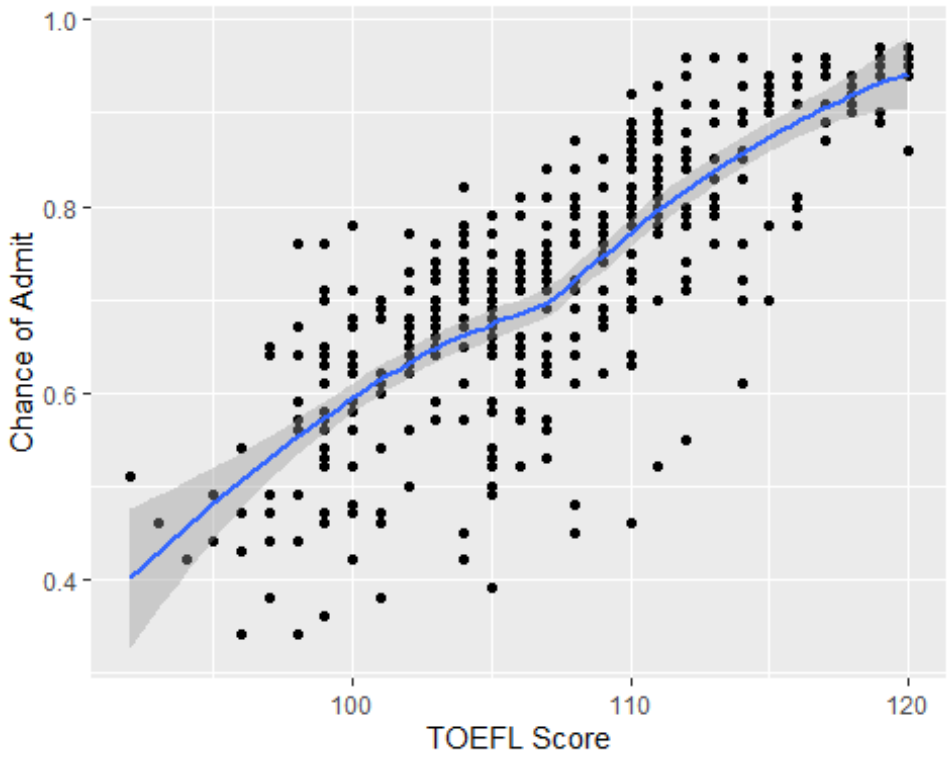

```
(gre <- summary(lm(`Chance of Admit`~`GRE Score`,df)))
```

```
##
## Call:
## lm(formula = `Chance of Admit` ~ `GRE Score`, data = df)
##
## 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
```

TOEFL Score against Chance of Admit A linear relationship is present in this plot between the two attributes Hence the chance of admission increases with the TOEFL score

```
ggplot(data=df)+
  geom_point(mapping=aes(x=`TOEFL Score`,y=`Chance of Admit`))+
  geom_smooth(mapping=aes(x=`TOEFL Score`,y=`Chance of Admit`))
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



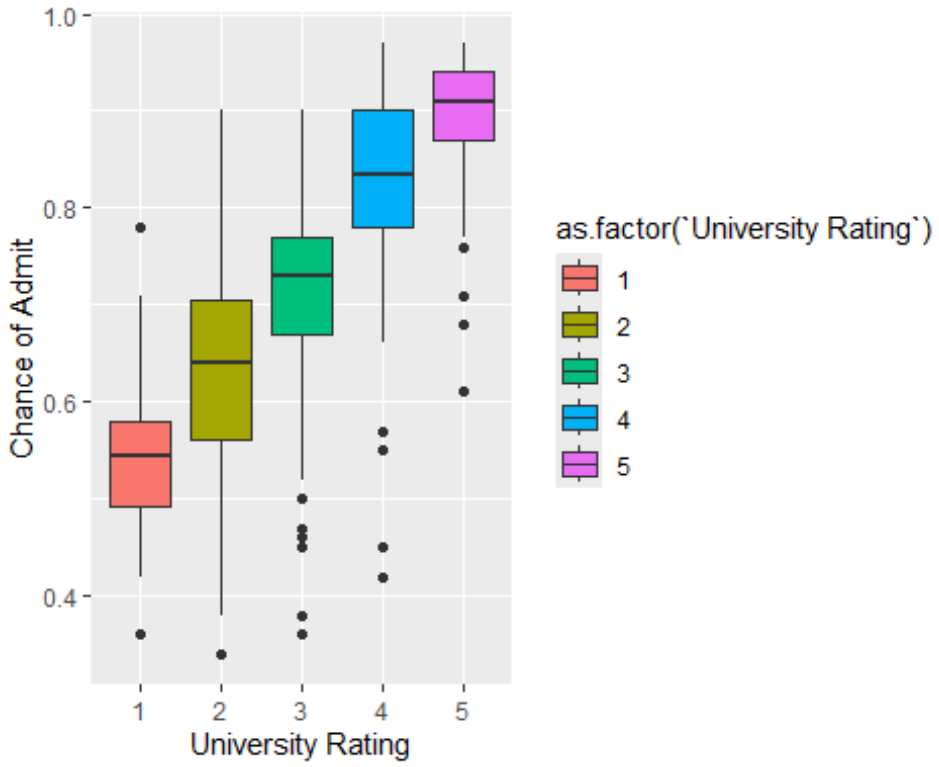

```
(toefl<- summary(lm(`Chance of Admit`~`TOEFL Score`,df)))

##
## Call:
## lm(formula = `Chance of Admit` ~ `TOEFL Score`, data = df)
##
## 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
```

The two previous plots display a confidence interval around the smooth curve. This interval appears wider at the beginning and end of the plot, which could be interpreted as a higher uncertainty in those regions

University Rating against Chance of Admit Boxplot was chosen because the quantitative variable is discrete

```
ggplot(data=df)+
  geom_boxplot(mapping = aes(x=as.factor(`University Rating`),y=`Chance of
Admit`,fill=as.factor(`University Rating`)))+
   xlab("University Rating")
```




```
(rate<- summary(lm(`Chance of Admit`~`University Rating`,df)))

##
## Call:
## lm(formula = `Chance of Admit` ~ `University Rating`, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38527 -0.04560  0.01473  0.06341  0.27209
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.450537   0.014463   31.15   <2e-16 ***
## `University Rating` 0.088684   0.004393   20.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1004 on 398 degrees of freedom
```

```
## Multiple R-squared:  0.5059, Adjusted R-squared:  0.5046
## F-statistic: 407.5 on 1 and 398 DF,  p-value: < 2.2e-16
```

There is a peculiar trend where, in general, higher-rated universities have lower admission rates than lower-rated universities. Shouldn't the trend be reversed, indicating that the higher the university's rating, the lower the chance of getting admitted to that university.

To better comprehend this trend, we calculated the average of the predictors in two cases: first, for universities rated higher than 3, and second, for universities rated below 3

```
avgHigherRate <- df %>%
  filter(as.numeric(df$`University Rating`) > 3) %>%
  summarize(avgTOEFL=mean(`TOEFL Score`),
            avgGRE=mean(`GRE Score`),
            avgSOP=mean(SOP),
            avgLOR=mean(LOR),
            avgCGPA=mean(CGPA))
avgLowerRate <- df %>%
  filter(as.numeric(df$`University Rating`) < 3) %>%
  summarize(avgTOEFL=mean(`TOEFL Score`),
            avgGRE=mean(`GRE Score`),
            avgSOP=mean(SOP),
            avgLOR=mean(LOR),
            avgCGPA=mean(CGPA))
average <- rbind(avgHigherRate,avgLowerRate)
new_row_names <- c(">3", "<3")
average<- as.data.frame(average)
rownames(average) <- new_row_names
average
```

```
##    avgTOEFL   avgGRE   avgSOP   avgLOR  avgCGPA
## >3 112.6493 326.3955 4.283582 4.164179 9.142313
## <3 102.6541 308.0000 2.545113 2.785714 8.098120
```

Analyzing the averages allows us to gain deeper insights into the prevailing trends. These averages exhibit significant differences especially in SOP and LOR

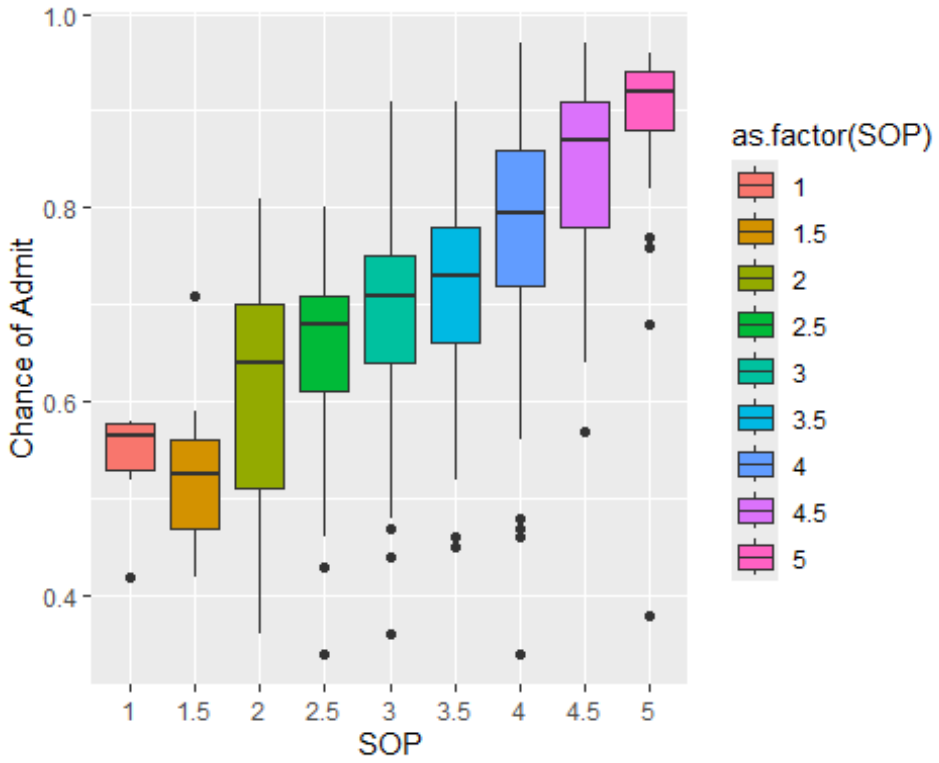
When comparing the averages of all features for universities rated below 3 to those rated above 3, a consistent pattern emerges. The averages for all features tend to be lower in the universities with lower ratings

In other words, we are not comparing the same students with the same predictor values for their chances of getting admitted to high and low-rated universities. Instead, we are comparing different students with varying features or characteristics.

In general, students applying to master's programs at higher-rated universities tend to have better qualifications than those applying to lower-rated universities. This is why they have a higher chance of getting admitted, even if the university they are applying to is rated higher than those who are applying to universities with lower ratings and have lower qualifying features.

SOP against Chance of Admit The chance of admission increases with the SOP, hence it seems it is a linear trend

```
ggplot(data = df, aes(x = as.factor(SOP), y = `Chance of Admit`, fill =
as.factor(SOP))) +
  geom_boxplot() +
  xlab("SOP")
```



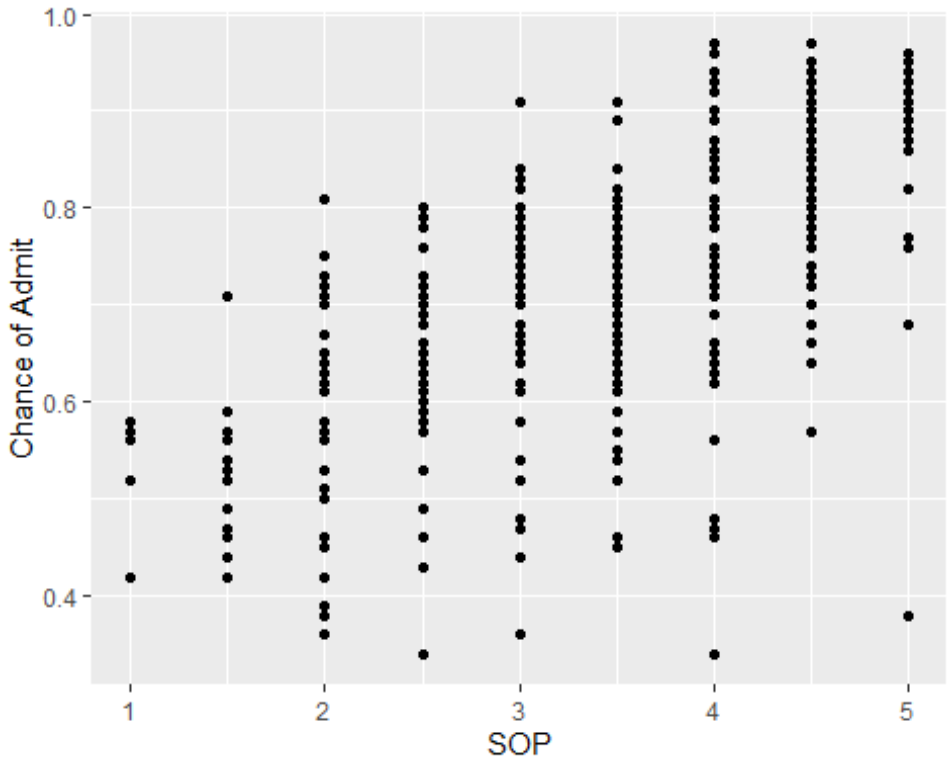

```
(sop <- summary(lm(`Chance of Admit`~ SOP,df)))

##
## Call:
## lm(formula = `Chance of Admit` ~ SOP, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49748 -0.05392  0.01823  0.07037  0.22393
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.398942   0.018556   21.50   <2e-16 ***
## SOP         0.095708   0.005233   18.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1053 on 398 degrees of freedom
```

```
## Multiple R-squared:  0.4566, Adjusted R-squared:  0.4552
## F-statistic: 334.4 on 1 and 398 DF,  p-value: < 2.2e-16
```

The average of getting admitted seems higher when the SOP is 1 compared to an SOP of 1.5 Is this really the case. will will present the same graph as a scatter plot instead of a boxplot.

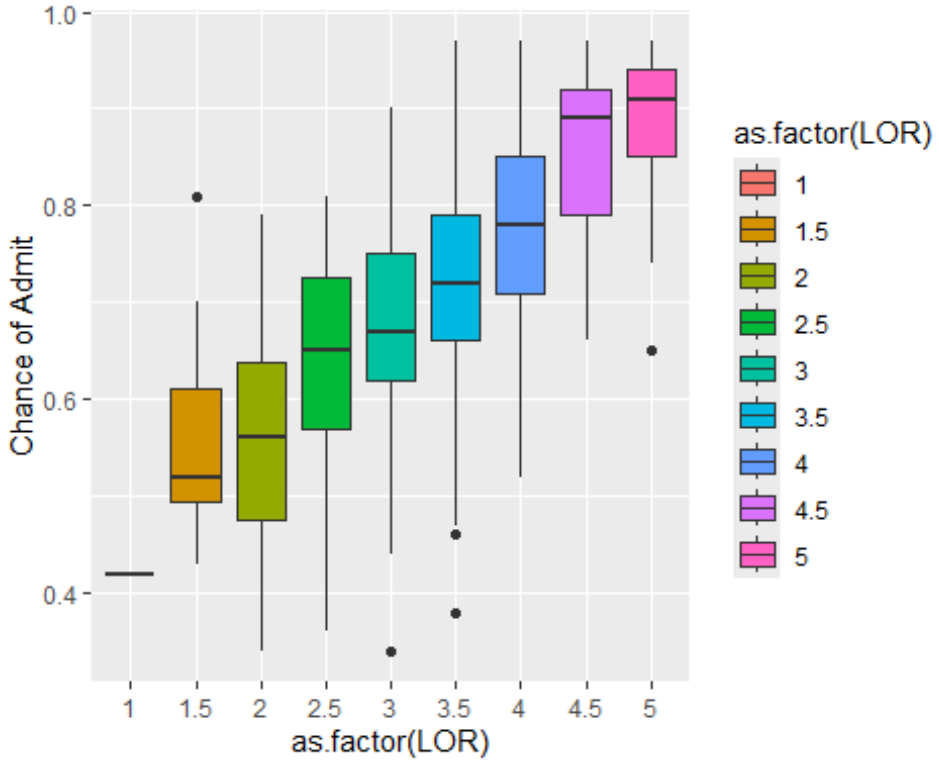
```
ggplot(data = df, aes(x = SOP, y = `Chance of Admit`)) +
  geom_point()
```



The number of students with an SOP of 1.5 is higher than those with an SOP of 1. Among the five students with an SOP of 1, three have values near 0.6 on the y-axis. In contrast, students with an SOP of 1.5 are more numerous and widely distributed. However, it's evident that the chance of getting admitted for a student with an SOP of 1 does not surpass that of a student with an SOP of 1.5. That why the average of getting admitted seems higher when the SOP is 1 compared to an SOP of 1.5 but it is not the case. The higher the SOP score the higher the chance of Admit, hence it seems we have a linear trend

LOR against Chance of Admit The higher the LOR score the higher the chance of Admit, hence it seems we have a linear trend

```
ggplot(data = df, aes(x = as.factor(LOR), y = `Chance of Admit`, fill =
as.factor(LOR))) +
  geom_boxplot()
```



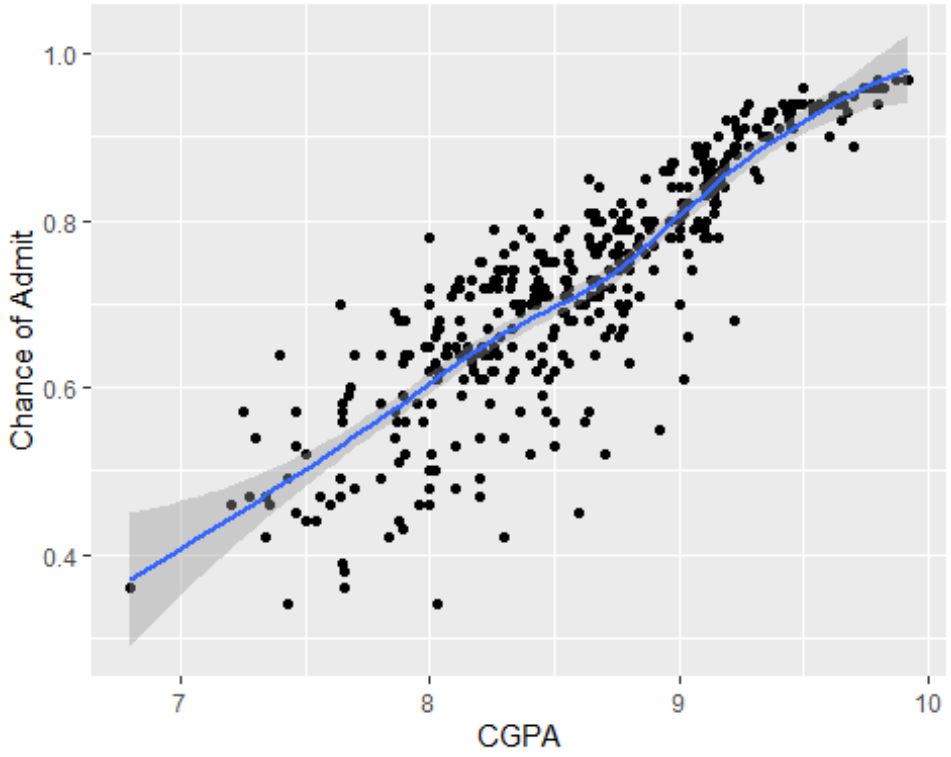

```
(lor <- summary(lm(`Chance of Admit`~LOR,df)))

##
## Call:
## lm(formula = `Chance of Admit` ~ LOR, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34940 -0.06256  0.00060  0.07389  0.29325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.357256   0.021072   16.95   <2e-16 ***
## LOR         0.106327   0.005907   18.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.106 on 398 degrees of freedom
## Multiple R-squared:  0.4488, Adjusted R-squared:  0.4474
## F-statistic:   324 on 1 and 398 DF,  p-value: < 2.2e-16
```

CGPA against Chance of Admit A linear trend is evident; the chance of admission increases with CGPA. However, there is a high level of uncertainty at the beginning of the curve, and a lower one towards the end.

```
ggplot(data=df)+
  geom_point(mapping=aes(x=CGPA,y=`Chance of Admit`))+
  geom_smooth(mapping=aes(x=CGPA,y=`Chance of Admit`))
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

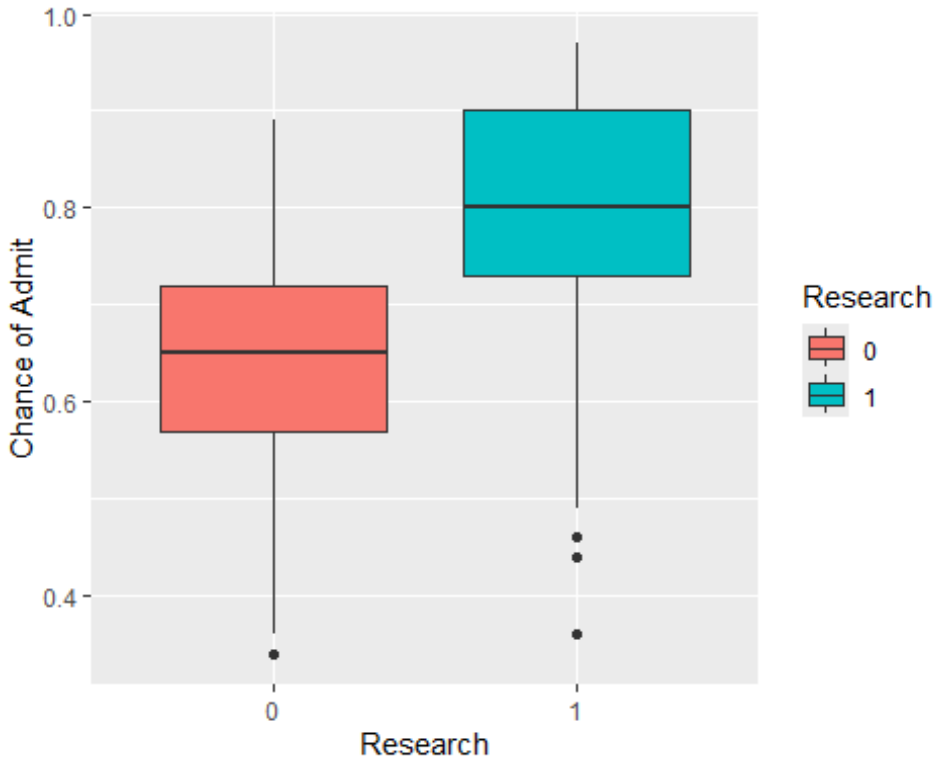



```
(cgpa<- summary(lm(`Chance of Admit`~ CGPA,df)))
```

```
##
## Call:
## lm(formula = `Chance of Admit` ~ CGPA, data = df)
##
## 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
```

Research against Chance of Admit

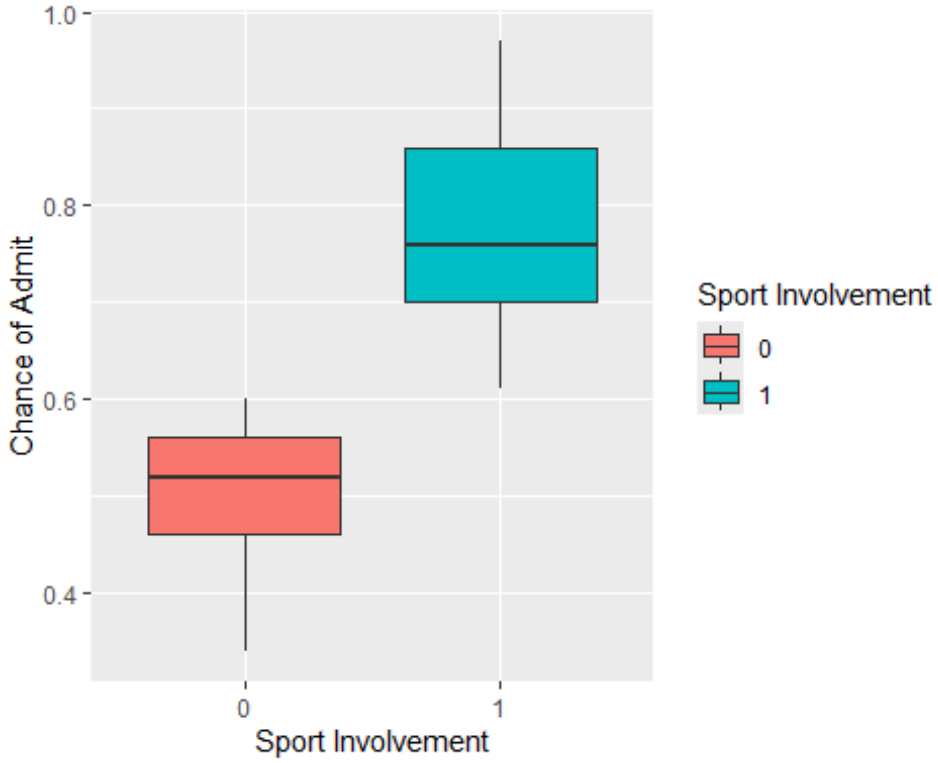
```
ggplot(data=df)+
  geom_boxplot(mapping=aes(x=Research,y=`Chance of Admit`,fill=Research))
```




```
research<- summary(lm(`Chance of Admit`~ Research,df))
```

Sport involvement against Chance of Admit

```
ggplot(data=df)+
  geom_boxplot(mapping=aes(x=`Sport Involvement`,y=`Chance of
Admit`,fill=`Sport Involvement`))
```




```
(sport<- summary(lm(`Chance of Admit`~ `Sport Involvement`,df)))

##
## Call:
## lm(formula = `Chance of Admit` ~ `Sport Involvement`, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16454 -0.07454 -0.00454  0.06676  0.19546
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.50324    0.01117   45.05   <2e-16 ***
## `Sport Involvement`1  0.27130    0.01237   21.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0961 on 398 degrees of freedom
## Multiple R-squared:  0.547,  Adjusted R-squared:  0.5459
## F-statistic: 480.6 on 1 and 398 DF,  p-value: < 2.2e-16
```

It is worth noting that the previous two boxplots displays a very pronounced relationship between the each predictor “Research” and “Sport Involvement” and the response “Chance of Admit”. Hence, “Research” and “Sport Involvement” corresponds to good attributes to include in our model.

Note that in every simple linear regression previously presented the p-value is significant, the RSE is low and the R^2 is notable and the F- statistic is for from 1. Hence, we can conclude that there is a relationship between each predictor and the response. Our previous hypothesis is present.

Generating the model

We present a model between sport and research after we assumed that these attributes are good to include in our model. All p-values are significant and the values of R^2 and RSE are 0.65 and 0.08407 respectively. No evidence till now that these attributes shouldn't belong to our model

```
(sport_research<- summary(lm(`Chance of Admit`~ `Sport
Involvement`+Research,df)))

##
## Call:
## lm(formula = `Chance of Admit` ~ `Sport Involvement` + Research,
##     data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.222385 -0.062424 -0.002424  0.068313  0.177576
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.483122   0.009939   48.61   <2e-16 ***
## `Sport Involvement`1 0.229302   0.011468   20.00   <2e-16 ***
## Research1            0.099263   0.008946   11.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08407 on 397 degrees of freedom
## Multiple R-squared:  0.6542, Adjusted R-squared:  0.6525
## F-statistic: 375.6 on 2 and 397 DF,  p-value: < 2.2e-16
```

Clearly, CGPA is a crucial feature to include in our model. Acceptance into a master's program is notably dependent on the CGPA.

```
(model2 <- summary(lm(`Chance of Admit`~ `Sport
Involvement`+Research+CGPA,df)))

##
## Call:
## lm(formula = `Chance of Admit` ~ `Sport Involvement` + Research +
##     CGPA, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.180908 -0.032727  0.003963  0.036244  0.165961
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.653942   0.048105 -13.594  < 2e-16 ***
## `Sport Involvement`1  0.128575   0.008483  15.157  < 2e-16 ***
## Research1             0.034111   0.006356   5.367 1.37e-07 ***
## CGPA                  0.145928   0.006119  23.847  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05393 on 396 degrees of freedom
## Multiple R-squared:  0.8581, Adjusted R-squared:  0.857
## F-statistic: 798.1 on 3 and 396 DF,  p-value: < 2.2e-16
```

We observe significant increase in R^2 and decrease in RSE These three predictors were evident features to be put in the model

After trying many models containing these three features, and checking for significant interaction terms between them. Considering the R^2 and the RSE and the p-values for each feature and for the interactions,this model yielded the highest R^2 with the smallest RSE.

```
(model3 <- summary(lm(`Chance of Admit`~ `Sport
Involvement`+Research*CGPA,df)))
```

```
##
## Call:
## lm(formula = `Chance of Admit` ~ `Sport Involvement` + Research *
##     CGPA, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.184151 -0.031279  0.003796  0.033882  0.124221
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.157427   0.075994  -2.072    0.039 *
## `Sport Involvement`1  0.150850   0.008339  18.090  < 2e-16 ***
## Research1            -0.704019   0.091634  -7.683 1.24e-13 ***
## CGPA                  0.083979   0.009546   8.797  < 2e-16 ***
## Research1:CGPA        0.086818   0.010755   8.072 8.42e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05003 on 395 degrees of freedom
## Multiple R-squared:  0.8782, Adjusted R-squared:  0.8769
## F-statistic: 711.8 on 4 and 395 DF,  p-value: < 2.2e-16
```

These predictors were chosen based on our initial analysis as clear candidates for inclusion in our model. However, the process of determining which additional features to include is not as evident. We initially started by including all available features and evaluated how the model performed. From there, we used the model's performance as a basis for analysis and made modifications as necessary.

```
(MLR <-summary(lm(`Chance of Admit`~`Sport Involvement`+`GRE Score`+`TOEFL
Score`+`University Rating`+SOP+LOR+`Research`*CGPA,df)))
```

```
##
## Call:
## lm(formula = `Chance of Admit` ~ `Sport Involvement` + `GRE Score` +
##     `TOEFL Score` + `University Rating` + SOP + LOR + Research *
##     CGPA, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.160162 -0.024694  0.005319  0.029793  0.119076
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.4600859  0.1143437  -4.024 6.88e-05 ***
## `Sport Involvement`1  0.1432561  0.0079817  17.948  < 2e-16 ***
## `GRE Score`           0.0011206  0.0004445   2.521 0.012106 *
## `TOEFL Score`         0.0027638  0.0008066   3.427 0.000676 ***
## `University Rating`   0.0046960  0.0035515   1.322 0.186850
## SOP                  -0.0062141  0.0041312  -1.504 0.133344
## LOR                   0.0157681  0.0041239   3.824 0.000153 ***
## Research1            -0.6462053  0.0873740  -7.396 8.68e-13 ***
## CGPA                  0.0393005  0.0112797   3.484 0.000550 ***
## Research1:CGPA        0.0789401  0.0102776   7.681 1.29e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04721 on 390 degrees of freedom
## Multiple R-squared:  0.8929, Adjusted R-squared:  0.8904
## F-statistic: 361.2 on 9 and 390 DF,  p-value: < 2.2e-16
```

The p-values of University Rating and SOP are no longer significant, despite their significance in the two simple linear regression models. In simple linear models, we isolate a single predictor, ignoring other factors. However, in multiple linear models, we fix all predictors except the one under investigation. The change in significance can be attributed to the correlations between University Rating and SOP with other predictors. These two predictors were likely getting "credit" for the effect of other predictors on "chance of Admit". It's challenging to specify which predictor(s) precisely, as the correlation matrix indicates that all predictors are correlated.

We will remove these two predictors

```
(model4 <- summary(lm(`Chance of Admit`~`Sport Involvement`+`GRE
Score`+`TOEFL Score`+LOR+`Research`*CGPA,df)))
```

```
##
## Call:
## lm(formula = `Chance of Admit` ~ `Sport Involvement` + `GRE Score` +
##     `TOEFL Score` + LOR + Research * CGPA, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.161454 -0.026017  0.005065  0.031932  0.119914
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.4697255  0.1114680  -4.214 3.12e-05 ***
## `Sport Involvement`1  0.1428447  0.0079824  17.895  < 2e-16 ***
## `GRE Score`           0.0011784  0.0004438   2.655 0.008247 **
## `TOEFL Score`         0.0027317  0.0007903   3.456 0.000607 ***
## LOR                   0.0144506  0.0036010   4.013 7.18e-05 ***
## Research1            -0.6605146  0.0870156  -7.591 2.35e-13 ***
## CGPA                  0.0384742  0.0110854   3.471 0.000577 ***
## Research1:CGPA        0.0805777  0.0102349   7.873 3.43e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04727 on 392 degrees of freedom
## Multiple R-squared:  0.892,  Adjusted R-squared:  0.8901
## F-statistic: 462.7 on 7 and 392 DF,  p-value: < 2.2e-16
```

When comparing models 3 and 4, we observed that the RSE increased from 0.04721 to 0.04727, and the R^2 decreased from 0.8929 to 0.892. This slight decrease in R^2 suggests a scenario of overfitting, providing stronger evidence for the removal of these two predictors.

We hypothesize that the chance of admission is higher for students with acceptable grades in both TOEFL and GRE scores compared to students who performed well on one test while performing poorly on the other, hence an interaction between these two scores.

```
model5 <-lm(`Chance of Admit`~`Sport Involvement`+`GRE Score`*`TOEFL
Score`+LOR+`Research`*CGPA,df)
summary(model5)
```

```
##
## Call:
## lm(formula = `Chance of Admit` ~ `Sport Involvement` + `GRE Score` *
##     `TOEFL Score` + LOR + Research * CGPA, data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168275 -0.026260  0.004815  0.032242  0.121947
```

```
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.788e+00  1.241e+00   2.247 0.025211 *
## `Sport Involvement`1        1.485e-01  8.209e-03  18.091  < 2e-16 ***
## `GRE Score`                -9.178e-03  3.954e-03  -2.321 0.020781 *
## `TOEFL Score`              -2.836e-02  1.182e-02  -2.399 0.016910 *
## LOR                         1.462e-02  3.575e-03   4.089 5.26e-05 ***
## Research1                  -5.615e-01  9.418e-02  -5.962 5.57e-09 ***
## CGPA                        4.221e-02  1.109e-02   3.805 0.000165 ***
## `GRE Score`:`TOEFL Score`  9.765e-05  3.705e-05   2.636 0.008728 **
## Research1:CGPA              6.880e-02  1.110e-02   6.200 1.44e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04692 on 391 degrees of freedom
## Multiple R-squared:  0.8939, Adjusted R-squared:  0.8918
## F-statistic: 411.9 on 8 and 391 DF,  p-value: < 2.2e-16
```

Comparing model 4 and 5, we observed that the R^2 increased from 0.892 to 0.8939, and the RSE decreased from 0.04727 to 0.04692. Our hypothesis is valid.

No polynomial regression model provided better results than Model 5, which aligns with our expectations as the predictors demonstrate a linear trend in the simple linear models. We provide a slight example

```
(model6 <- summary(lm(`Chance of Admit`~ I(LOR^2)+LOR+`Sport
Involvement`+`GRE Score`*`TOEFL Score`+`Research`*CGPA,df)))
```

```
##
## Call:
## lm(formula = `Chance of Admit` ~ I(LOR^2) + LOR + `Sport Involvement` +
##     `GRE Score` * `TOEFL Score` + Research * CGPA, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15923 -0.02587  0.00425  0.03194  0.11425
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.190e+00  1.251e+00   1.750  0.08082 .
## I(LOR^2)                    7.272e-03  2.717e-03   2.677  0.00775 **
## LOR                        -3.518e-02  1.894e-02  -1.858  0.06398 .
## `Sport Involvement`1        1.505e-01  8.180e-03  18.403  < 2e-16 ***
## `GRE Score`                -7.126e-03  3.997e-03  -1.783  0.07542 .
## `TOEFL Score`              -2.214e-02  1.196e-02  -1.851  0.06492 .
## Research1                  -5.232e-01  9.453e-02  -5.535 5.72e-08 ***
## CGPA                        4.427e-02  1.103e-02   4.012 7.21e-05 ***
## `GRE Score`:`TOEFL Score`  7.828e-05  3.747e-05   2.089  0.03731 *
## Research1:CGPA              6.430e-02  1.114e-02   5.773 1.59e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04655 on 390 degrees of freedom
## Multiple R-squared:  0.8958, Adjusted R-squared:  0.8934
## F-statistic: 372.7 on 9 and 390 DF,  p-value: < 2.2e-16
```

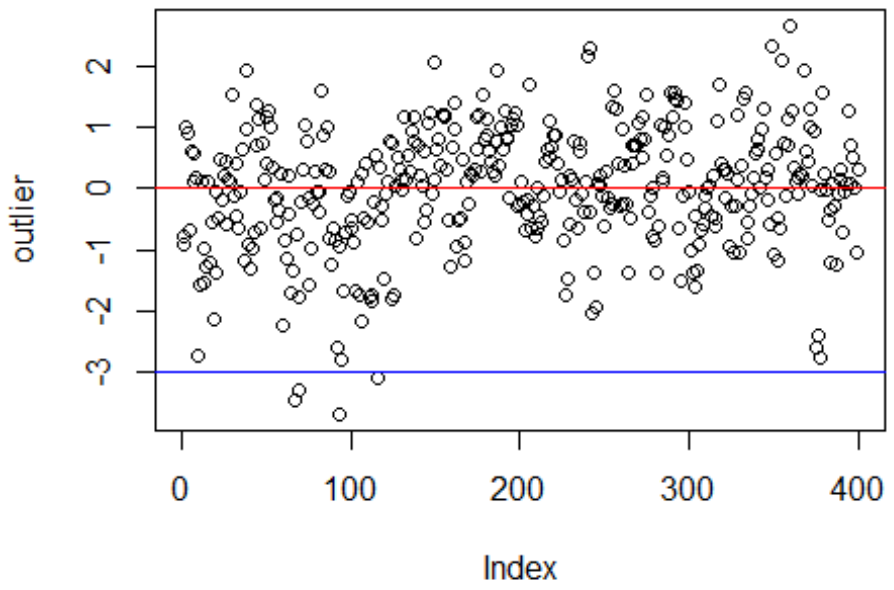
Even if this model gave a better R^2 and RSE, but this is not a good model to consider, since most p-values are now insignificant.

So our model is model5 model5 <- lm(`Chance of Admit~Sport Involvement+GRE Score*` TOEFL Score +LOR+`Research`*CGPA,df))

Checking for outliers

We should assess whether there are outliers in our data. If outliers are present, it's important to consider how our model will perform when these outliers are removed from the dataset. To do so, we can plot the studentized residuals. Observations whose studentized residuals are greater than 3 in absolute value are possible outliers.

```
outlier <- rstudent(model5)
plot(outlier)
abline(h = 0, col = "red")
abline(h = 3, col = "blue")
abline(h = -3, col = "blue")
```



Three observations were below -3, we will remove them and see how our model is performing. First, we find the indices of outliers, and then extract these points points from our original dataset.

```
threshold <- -3
outlier_indices <- which(outlier < threshold)
outliers <- df[outlier_indices, ]
print(outliers)
```

```
## # A tibble: 4 × 9
##   `GRE Score` `TOEFL Score` `University Rating`   SOP   LOR  CGPA Research
##         <dbl>         <dbl>               <dbl> <dbl> <dbl> <dbl> <fct>
## 1         327           114                   3   3     3    9.02 0
## 2         318           109                   3   3.5   4    9.22 1
## 3         298            98                   2   4     3    8.03 0
## 4         310           106                   4   4.5   4.5  9.04 1
## # ℹ 2 more variables: `Sport Involvement` <fct>, `Chance of Admit` <dbl>
```

```
cleaned_df <- df[-outlier_indices, ]
```

```
(Model<-summary(lm(`Chance of Admit`~`Sport Involvement`+`GRE Score`*`TOEFL Score`+LOR+`Research`*CGPA,cleaned_df)))
```

```
##
## Call:
## lm(formula = `Chance of Admit` ~ `Sport Involvement` + `GRE Score` *
##     `TOEFL Score` + LOR + Research * CGPA, data = cleaned_df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.129742 -0.026992  0.002875  0.029465  0.122526
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.791e+00  1.181e+00   2.364   0.0186 *
## `Sport Involvement`1       1.459e-01  7.782e-03  18.748  < 2e-16 ***
## `GRE Score`               -9.339e-03  3.762e-03  -2.482   0.0135 *
## `TOEFL Score`             -2.827e-02  1.126e-02  -2.510   0.0125 *
## LOR                        1.455e-02  3.391e-03   4.292 2.24e-05 ***
## Research1                 -5.508e-01  8.993e-02  -6.125 2.23e-09 ***
## CGPA                       4.991e-02  1.055e-02   4.729 3.16e-06 ***
## `GRE Score`:`TOEFL Score`  9.694e-05  3.529e-05   2.747   0.0063 **
## Research1:CGPA             6.753e-02  1.061e-02   6.366 5.51e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04437 on 387 degrees of freedom
## Multiple R-squared:  0.9041, Adjusted R-squared:  0.9021
## F-statistic: 456.2 on 8 and 387 DF,  p-value: < 2.2e-16
```

Outliers usually do not have a big impact on the least square line but removing them cause a lower RSE and an increase in R^2, i.e we get a different interpretation of the fit. This is the case here after removing the outliers RSE decreased from 0.04692 to 0.0451 and R^2 increased from 0.8958 to 0.8998. We conclude that our model became better after removing the outliers.

Further Hypothesis

There is a relationship between the strength of Statement of Purpose (SOP) and the Chance of Admit, but only in higher rated universities.

Students who do not have a remarkable CGPA are more likely to gain admission to higherrated universities if they actively participate in sports and engage in research activities.

Enrollment in a sport will impact a student's cumulative CGPA, as they may need to allocate time for practice, potentially affecting their academic performance.