DATA ANALYSIS AND REGRESSION ASSIGNMENT 3

Mrunali Vikas Patil

PROBLEM SET - 1

Load the dataset

```
data <- read.csv("college.csv")
```

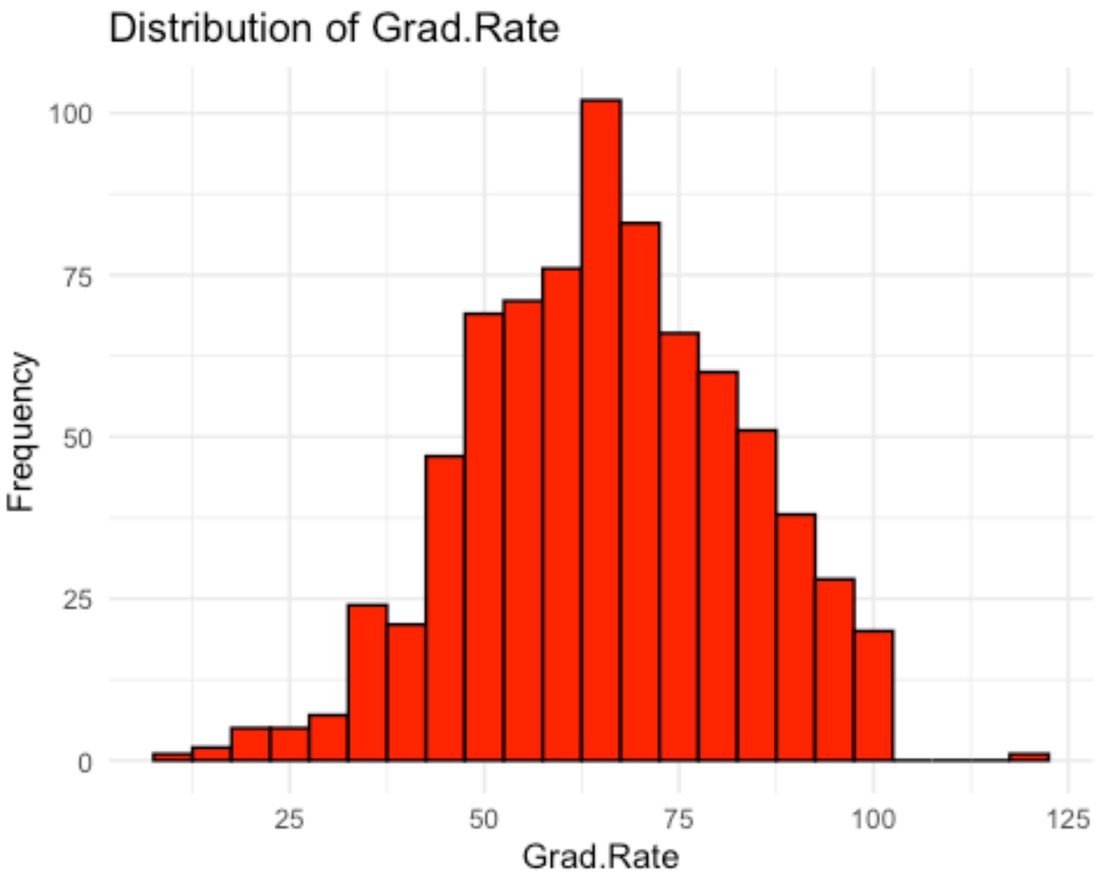
Question A)

Load the libraries

```
library(ggplot2)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Histogram for Grad.Rate

```
ggplot(data, aes(x = Grad.Rate)) +
  geom_histogram(binwidth = 5, fill = "red", color = "black") +
  labs(title = "Distribution of Grad.Rate", x = "Grad.Rate", y =
"Frequency") +
  theme_minimal()
```

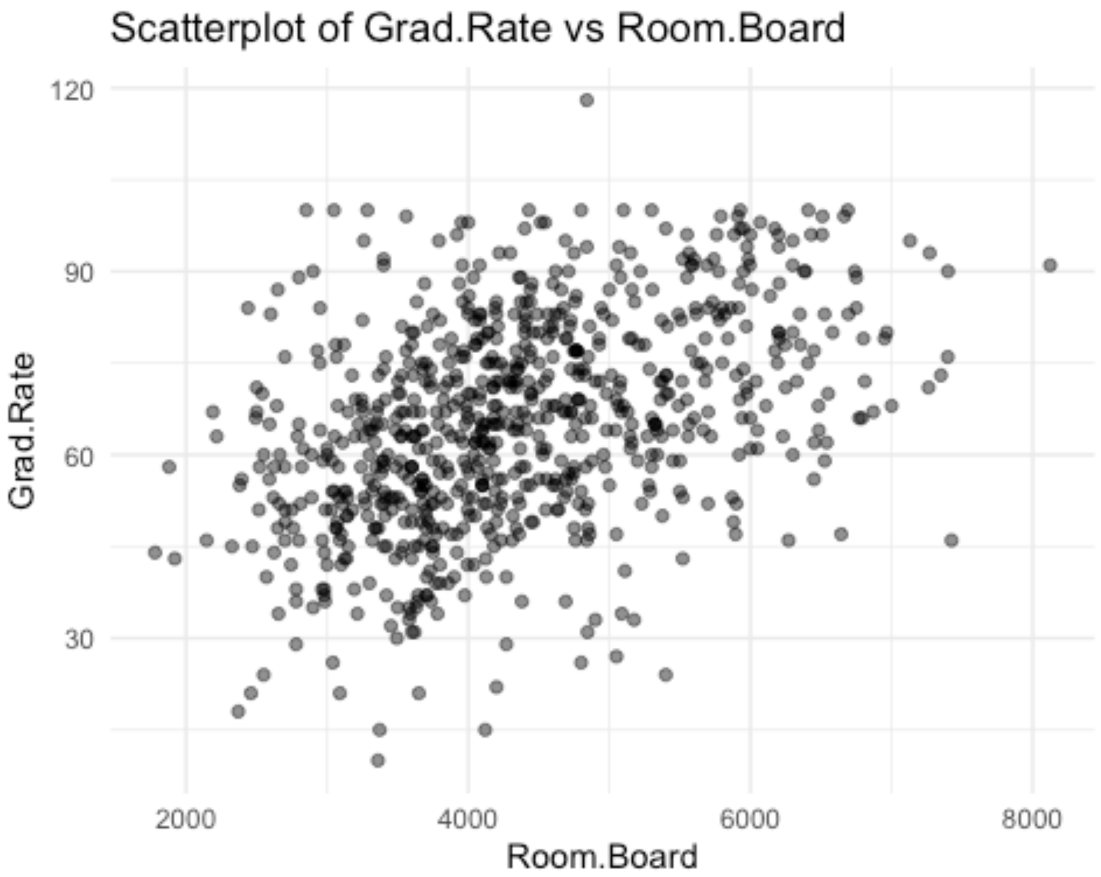


The distribution of Grad.Rate appears to be slightly left-skewed, with a tail on the left side. While the distribution is not perfectly symmetric, it doesn't seem severely skewed.

Question B)

Scatterplot for Grad.Rate vs each of the independent variables #Independent Variable - “Room.board”

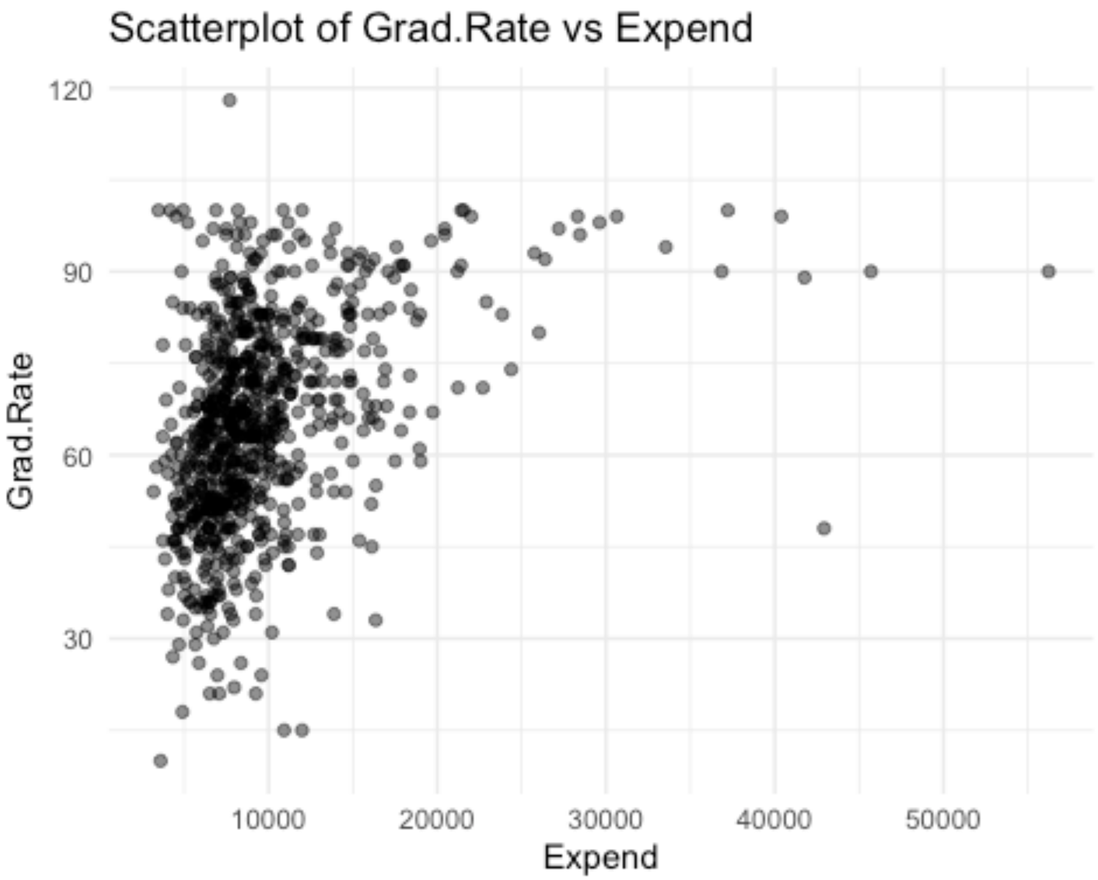
```r
ggplot(data, aes(x = Room.Board, y = Grad.Rate)) +
  geom_point(alpha = 0.5) +
  labs(title = "Scatterplot of Grad.Rate vs Room.Board", x =
"Room.Board", y = "Grad.Rate") +
  theme_minimal()
```



By looking at the Scatterplot, we can conclude that there is positive linear relations between the vairables

Independent Variable - “Expend”

```
ggplot(data, aes(x = Expend, y = Grad.Rate)) +
  geom_point(alpha = 0.5) +
  labs(title = "Scatterplot of Grad.Rate vs Expend", x = "Expend", y =
"Grad.Rate") +
  theme_minimal()
```



The Scatterplot shows no linear relationship between the Variables

To find correlation between Grad.Rate and all other numeric variables:

```
cor_data <- cor(data[sapply(data, is.numeric)])
cor_data["Grad.Rate", ]

##   Accept.pct      Elite10  F.Undergrad  P.Undergrad     Outstate 
Room.Board 
## -0.286971504  0.348732550 -0.078773129 -0.257000991  0.571289928 
0.424941541 
##        Books     Personal          PhD     Terminal    S.F.Ratio 
perc.alumni 
##  0.001060894 -0.269343964  0.305037850  0.289527232 -0.306710405 
0.490897562
```

```
##       Expend    Grad.Rate 
##  0.390342696  1.000000000
```

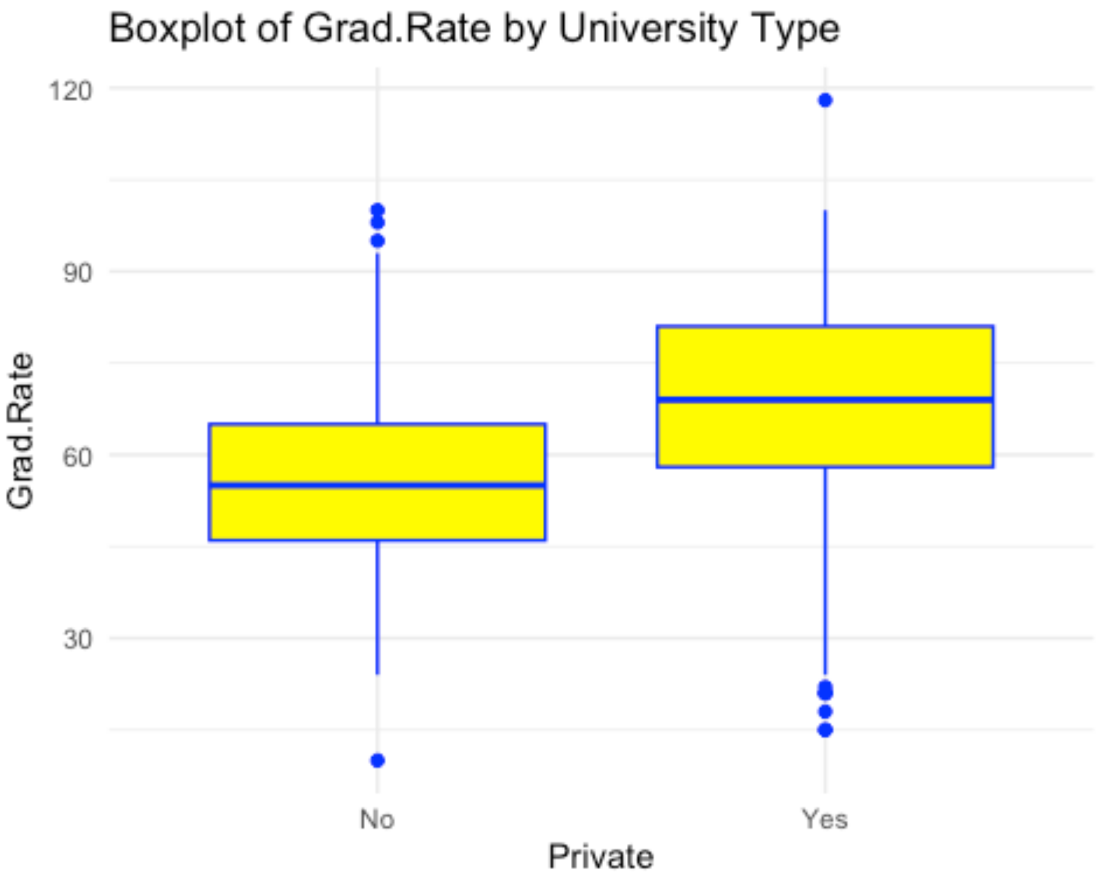
By looking at the Correlation Values, Variable Grad.Rate and Outstate have the maximum correlation Value, whereas variables Book and Grad.Rate have the lowest Correlation value

Question C)

Boxplots to evaluate if graduation rates vary by university type (private vs public) and by status (elite vs not elite)

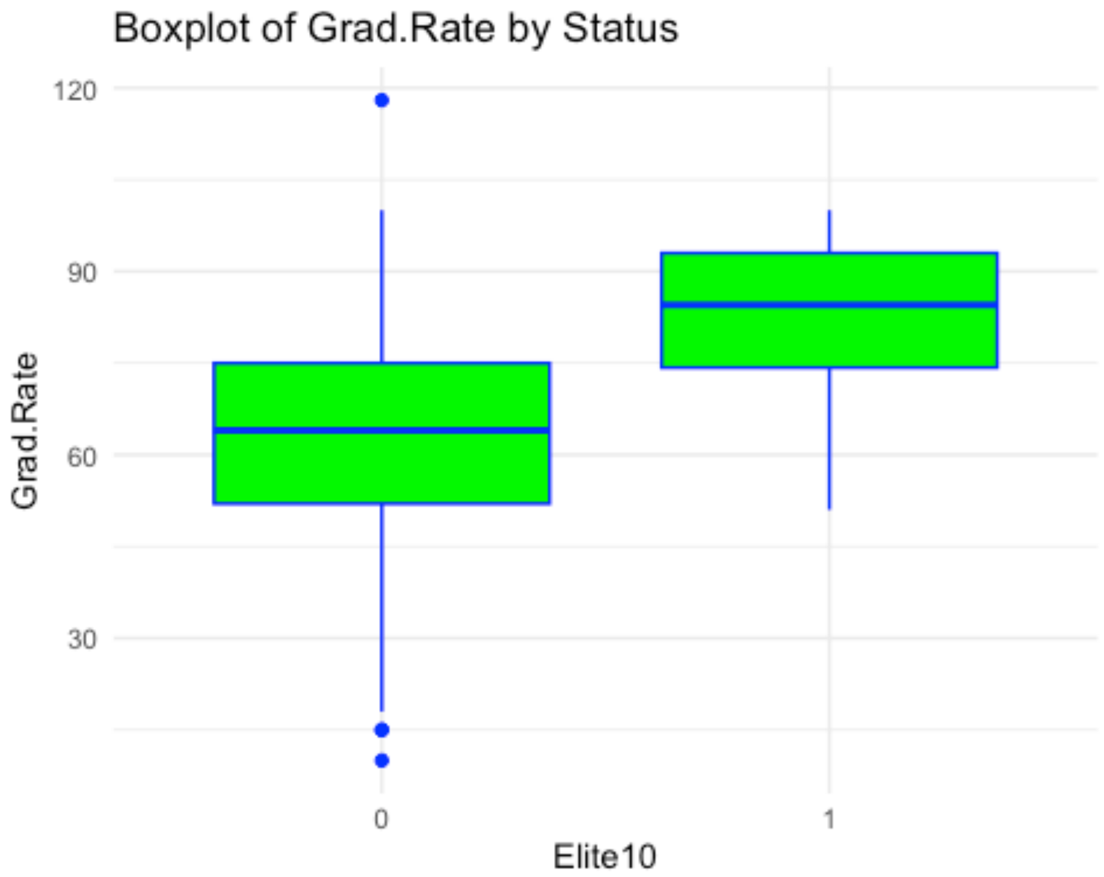
Boxplot for Grad.Rate vs Private

```
ggplot(data, aes(x = Private, y = Grad.Rate)) +
  geom_boxplot(fill = "yellow",color = "blue") +
  labs(title = "Boxplot of Grad.Rate by University Type", x = 
"Private", y = "Grad.Rate") +
  theme_minimal()
```



Boxplot for Grad.Rate vs Elite

```r
data$Elite10 <- as.factor(data$Elite10)

ggplot(data, aes(x = Elite10, y = Grad.Rate)) +
  geom_boxplot(fill = "green",color = "blue") +
  labs(title = "Boxplot of Grad.Rate by Status", x = "Elite10", y =
"Grad.Rate") +
  theme_minimal()
```



By looking at the Boxplots, we can see Grad.Rate vary by status that is elite Vs non elite as we can see there is separation between two classes

We can also conclude that there is slight overlapping of Private University and Pubic University hence the variation of Grad.Rate is not as good as Status

Question D)

Remove the 'school' column and convert 'Private' to a dummy variable

```
data_adj <- data %>%
  select(-school) %>%
  mutate(Private = ifelse(Private == "Yes", 1, 0))
```

#Fit a full model (with all independent variables) to predict Grad.Rate

```
model1 <- lm(Grad.Rate ~ ., data = data_adj)
```

Display the model summary

```
summary(model1)

## 
## Call:
## lm(formula = Grad.Rate ~ ., data = data_adj)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -46.200  -6.777  -0.707   7.217  57.907 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  5.140e+01  6.124e+00   8.393 2.30e-16 ***
## Private      4.620e+00  1.722e+00   2.683  0.00746 ** 
## Accept.pct  -1.811e+01  3.843e+00  -4.712 2.91e-06 ***
## Elite101     4.017e+00  2.003e+00   2.005  0.04527 *  
## F.Undergrad  6.809e-04  1.429e-04   4.767 2.24e-06 ***
## P.Undergrad -1.956e-03  3.904e-04  -5.009 6.80e-07 ***
## Outstate     1.235e-03  2.286e-04   5.401 8.88e-08 ***
## Room.Board   1.667e-03  5.944e-04   2.805  0.00517 ** 
## Books       -2.524e-03  2.966e-03  -0.851  0.39511    
## Personal    -1.718e-03  7.781e-04  -2.208  0.02753 *  
## PhD          1.306e-01  5.621e-02   2.324  0.02037 *  
## Terminal    -7.284e-02  6.257e-02  -1.164  0.24469    
## S.F.Ratio    1.003e-03  1.619e-01   0.006  0.99506    
## perc.alumni  3.092e-01  4.839e-02   6.390 2.89e-10 ***
## Expend      -4.365e-04  1.518e-04  -2.875  0.00415 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12.92 on 762 degrees of freedom
## Multiple R-squared:  0.4448, Adjusted R-squared:  0.4346 
## F-statistic: 43.61 on 14 and 762 DF,  p-value: < 2.2e-16
```

Question E)

View the first few rows of the adjusted data

```
head(data_adj)
```

```
##   Private Accept.pct Elite10 F.Undergrad P.Undergrad Outstate
Room.Board Books
## 1       1  0.7421687       0        2885         537     7440
3300   450
## 2       1  0.8801464       0        2683        1227    12280
6450   750
## 3       1  0.7682073       0        1036          99    11250
3750   400
## 4       1  0.8369305       1         510          63    12960
5450   450
## 5       1  0.7564767       0         249         869     7560
4120   800
## 6       1  0.8160136       0         678          41    13500
3335   500
##   Personal PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate
## 1     2200  70       78      18.1          12   7041        60
## 2     1500  29       30      12.2          16  10527        56
## 3     1165  53       66      12.9          30   8735        54
## 4      875  92       97       7.7          37  19016        59
## 5     1500  76       72      11.9           2  10922        15
## 6      675  67       73       9.4          11   9727        55
```

Vif Statistics

Load necessary library

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

Compute VIF

```
vif(model1)
```

```
##     Private  Accept.pct     Elite10 F.Undergrad P.Undergrad
Outstate
```

```
##    2.739521    1.486633    1.687903    2.233117    1.643393 
3.935059 
##  Room.Board       Books    Personal         PhD    Terminal 
S.F.Ratio 
##    1.976762    1.115823    1.290983    3.917716    3.946581 
1.909722 
## perc.alumni      Expend 
##    1.672367    2.922643
```

As all the variables have VIF value less than 10 hence there is no multi-collinearity in any of the variable.

Question F)

Applying two Variable Selection Procedures

Backward selection

```
backward_model <- step(model1, direction = "backward")

## Start:  AIC=3990.75
## Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad + 
P.Undergrad + 
##     Outstate + Room.Board + Books + Personal + PhD + Terminal + 
##     S.F.Ratio + perc.alumni + Expend
## 
##               Df Sum of Sq    RSS    AIC
## - S.F.Ratio    1       0.0 127126 3988.8
## - Books        1     120.8 127247 3989.5
## - Terminal     1     226.1 127352 3990.1
## <none>                     127126 3990.8
## - Elite10      1     671.0 127797 3992.8
## - Personal     1     813.4 127939 3993.7
## - PhD          1     901.2 128027 3994.2
## - Private      1    1200.9 128327 3996.1
## - Room.Board   1    1312.3 128438 3996.7
## - Expend       1    1379.4 128505 3997.1
## - Accept.pct   1    3704.3 130830 4011.1
## - F.Undergrad  1    3790.8 130917 4011.6
## - P.Undergrad  1    4185.8 131312 4013.9
## - Outstate     1    4866.1 131992 4017.9
## - perc.alumni  1    6812.2 133938 4029.3
```

```
##
## Step:  AIC=3988.75
## Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +
P.Undergrad +
##     Outstate + Room.Board + Books + Personal + PhD + Terminal +
##     perc.alumni + Expend
##
##                Df Sum of Sq    RSS    AIC
## - Books        1     120.8 127247 3987.5
## - Terminal     1     226.3 127352 3988.1
## <none>                     127126 3988.8
## - Elite10      1     671.0 127797 3990.8
## - Personal     1     818.0 127944 3991.7
## - PhD          1     903.9 128030 3992.3
## - Private      1    1227.8 128354 3994.2
## - Room.Board   1    1312.3 128438 3994.7
## - Expend       1    1642.3 128768 3996.7
## - Accept.pct   1    3734.1 130860 4009.2
## - F.Undergrad  1    3854.2 130980 4010.0
## - P.Undergrad  1    4186.5 131313 4011.9
## - Outstate     1    4891.0 132017 4016.1
## - perc.alumni  1    6848.2 133974 4027.5
##
## Step:  AIC=3987.49
## Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +
P.Undergrad +
##     Outstate + Room.Board + Personal + PhD + Terminal + perc.alumni
+
##     Expend
##
##                Df Sum of Sq    RSS    AIC
## - Terminal     1     274.2 127521 3987.2
## <none>                     127247 3987.5
## - Elite10      1     664.8 127912 3989.5
## - Personal     1     960.8 128208 3991.3
## - PhD          1    1018.4 128265 3991.7
## - Private      1    1188.8 128436 3992.7
## - Room.Board   1    1254.2 128501 3993.1
## - Expend       1    1672.3 128919 3995.6
## - Accept.pct   1    3625.5 130872 4007.3
## - F.Undergrad  1    3781.9 131029 4008.2
## - P.Undergrad  1    4171.4 131418 4010.6
## - Outstate     1    4937.7 132185 4015.1
## - perc.alumni  1    6929.8 134177 4026.7
##
## Step:  AIC=3987.16
## Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad + 
P.Undergrad + 
##     Outstate + Room.Board + Personal + PhD + perc.alumni + Expend
##
##                Df Sum of Sq    RSS    AIC
## <none>                      127521 3987.2
## - Elite10      1     672.6 128194 3989.3
## - PhD          1     861.3 128382 3990.4
## - Personal     1     946.5 128467 3990.9
## - Room.Board   1    1135.3 128656 3992.1
## - Private      1    1329.5 128851 3993.2
## - Expend       1    1719.0 129240 3995.6
## - Accept.pct   1    3655.7 131177 4007.1
## - F.Undergrad  1    3680.7 131202 4007.3
## - P.Undergrad  1    4219.0 131740 4010.5
## - Outstate     1    4773.9 132295 4013.7
## - perc.alumni  1    6758.1 134279 4025.3
```

```
forward_model <- step(model1, direction = "forward")
```

```
## Start:  AIC=3990.75
## Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad + 
P.Undergrad + 
##     Outstate + Room.Board + Books + Personal + PhD + Terminal + 
##     S.F.Ratio + perc.alumni + Expend
```

Display the summary for both models

```
summary(backward_model)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 + 
F.Undergrad + 
##     P.Undergrad + Outstate + Room.Board + Personal + PhD + 
perc.alumni +
##     Expend, data = data_adj)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.085  -6.932  -0.775   7.325  57.598
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.840e+01  4.621e+00  10.475  < 2e-16 ***
## Private      4.770e+00  1.689e+00   2.824  0.00486 **
## Accept.pct  -1.778e+01  3.797e+00  -4.683 3.34e-06 ***
## Elite101     4.022e+00  2.002e+00   2.009  0.04492 *
## F.Undergrad  6.631e-04  1.411e-04   4.699 3.10e-06 ***
## P.Undergrad -1.963e-03  3.901e-04  -5.031 6.09e-07 ***
## Outstate     1.215e-03  2.270e-04   5.352 1.15e-07 ***
## Room.Board   1.534e-03  5.878e-04   2.610  0.00924 **
## Personal    -1.820e-03  7.638e-04  -2.383  0.01742 *
## PhD          8.424e-02  3.706e-02   2.273  0.02329 *
## perc.alumni  3.060e-01  4.806e-02   6.367 3.32e-10 ***
## Expend      -4.465e-04  1.390e-04  -3.211  0.00138 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.91 on 765 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.4351
## F-statistic: 55.33 on 11 and 765 DF,  p-value: < 2.2e-16
```

```
summary(forward_model)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 +
F.Undergrad +
##     P.Undergrad + Outstate + Room.Board + Books + Personal +
##     PhD + Terminal + S.F.Ratio + perc.alumni + Expend, data =
data_adj)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.200  -6.777  -0.707   7.217  57.907
```

```
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  5.140e+01  6.124e+00   8.393 2.30e-16 ***
## Private      4.620e+00  1.722e+00   2.683  0.00746 ** 
## Accept.pct  -1.811e+01  3.843e+00  -4.712 2.91e-06 ***
## Elite101     4.017e+00  2.003e+00   2.005  0.04527 *  
## F.Undergrad  6.809e-04  1.429e-04   4.767 2.24e-06 ***
## P.Undergrad -1.956e-03  3.904e-04  -5.009 6.80e-07 ***
## Outstate     1.235e-03  2.286e-04   5.401 8.88e-08 ***
## Room.Board   1.667e-03  5.944e-04   2.805  0.00517 ** 
## Books       -2.524e-03  2.966e-03  -0.851  0.39511    
## Personal    -1.718e-03  7.781e-04  -2.208  0.02753 *  
## PhD          1.306e-01  5.621e-02   2.324  0.02037 *  
## Terminal    -7.284e-02  6.257e-02  -1.164  0.24469    
## S.F.Ratio    1.003e-03  1.619e-01   0.006  0.99506    
## perc.alumni  3.092e-01  4.839e-02   6.390 2.89e-10 ***
## Expend      -4.365e-04  1.518e-04  -2.875  0.00415 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12.92 on 762 degrees of freedom
## Multiple R-squared:  0.4448, Adjusted R-squared:  0.4346 
## F-statistic: 43.61 on 14 and 762 DF,  p-value: < 2.2e-16
```

Adjusted R-squared for backward Model: 0.4351

Adjusted R-squared for forward Model : 0.4346

Hence we can conclude that Backward model performed better as it has more significant Variables

Question G)

We will choose Backward model as it has more significant Variables and the adjusted R-squared is grater than the forward model.

```
summary(backward_model)
```

```
## 
## Call:
## lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 + 
F.Undergrad + 
##     P.Undergrad + Outstate + Room.Board + Personal + PhD + 
perc.alumni + 
##     Expend, data = data_adj)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -45.085  -6.932  -0.775   7.325  57.598 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.840e+01  4.621e+00  10.475  < 2e-16 ***
## Private      4.770e+00  1.689e+00   2.824  0.00486 ** 
## Accept.pct  -1.778e+01  3.797e+00  -4.683 3.34e-06 ***
## Elite101     4.022e+00  2.002e+00   2.009  0.04492 *  
## F.Undergrad  6.631e-04  1.411e-04   4.699 3.10e-06 ***
## P.Undergrad -1.963e-03  3.901e-04  -5.031 6.09e-07 ***
## Outstate     1.215e-03  2.270e-04   5.352 1.15e-07 ***
## Room.Board   1.534e-03  5.878e-04   2.610  0.00924 ** 
## Personal    -1.820e-03  7.638e-04  -2.383  0.01742 *  
## PhD          8.424e-02  3.706e-02   2.273  0.02329 *  
## perc.alumni  3.060e-01  4.806e-02   6.367 3.32e-10 ***
## Expend      -4.465e-04  1.390e-04  -3.211  0.00138 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12.91 on 765 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.4351 
## F-statistic: 55.33 on 11 and 765 DF,  p-value: < 2.2e-16
```

Using the provided coefficients, the equation becomes:

Grad.Rate=51.40+4.62(Private)−18.11(Accept.pct)
+4.017(Elite101)+0.0006809(F.Undergrad)−0.001956(P.Undergrad)+0.001235(Outstate)
+0.001667(Room.Board)−0.002524(Books)−0.001718(Personal)+0.1306(PhD)
−0.07284(Terminal)+0.001003(S.F.Ratio)+0.3092(perc.alumni)−0.0004365(Expend)
Grad.Rate=51.40+4.62(Private)−18.11(Accept.pct)
+4.017(Elite101)+0.0006809(F.Undergrad)−0.001956(P.Undergrad)+0.001235(Outstate)
+0.001667(Room.Board)−0.002524(Books)−0.001718(Personal)+0.1306(PhD)
−0.07284(Terminal)+0.001003(S.F.Ratio)+0.3092(perc.alumni)−0.0004365(Expend)

Fit a final regression model M1

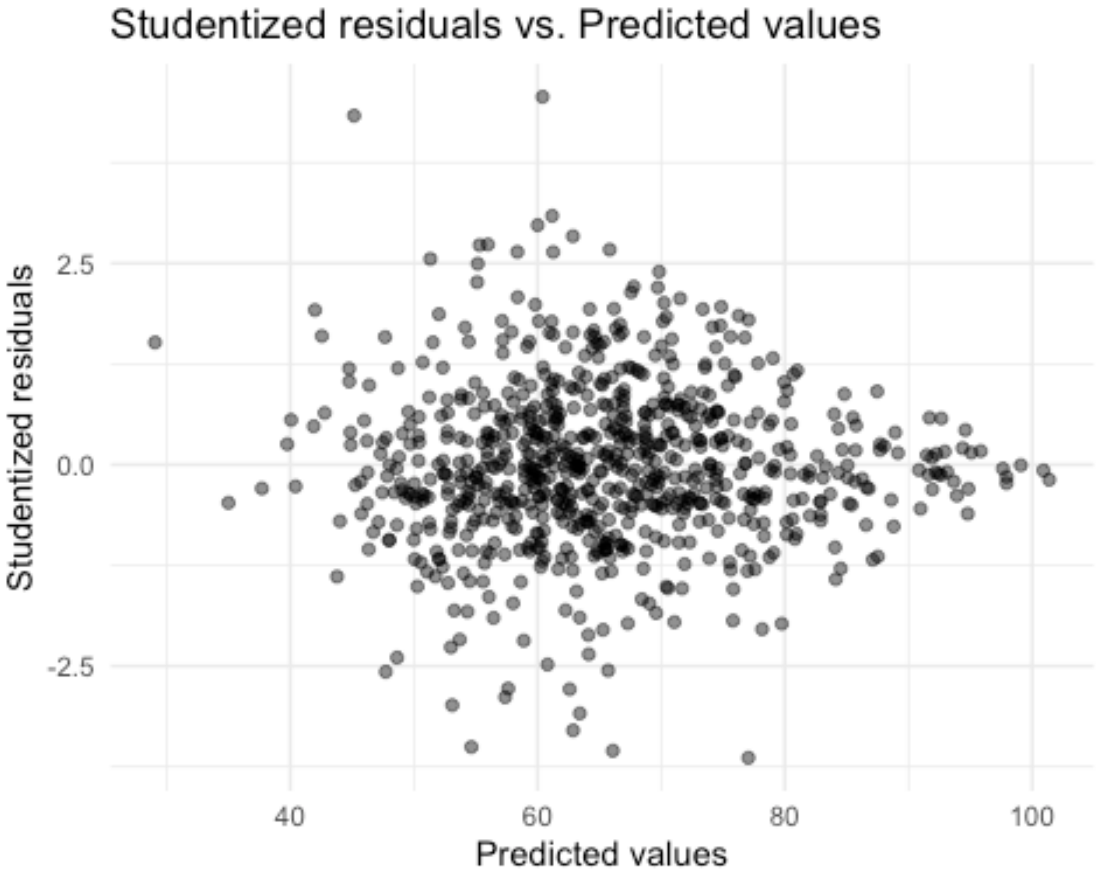
```
final_model <- lm(Grad.Rate ~Private + Accept.pct + Elite10 +
F.Undergrad +
                    P.Undergrad + Outstate + Room.Board + Personal +
PhD + perc.alumni +
                    Expend , data=data_adj)
summary(final_model)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 +
F.Undergrad +
##     P.Undergrad + Outstate + Room.Board + Personal + PhD +
perc.alumni +
##     Expend, data = data_adj)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.085  -6.932  -0.775   7.325  57.598
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.840e+01  4.621e+00  10.475  < 2e-16 ***
## Private      4.770e+00  1.689e+00   2.824  0.00486 **
## Accept.pct  -1.778e+01  3.797e+00  -4.683 3.34e-06 ***
## Elite101     4.022e+00  2.002e+00   2.009  0.04492 *
## F.Undergrad  6.631e-04  1.411e-04   4.699 3.10e-06 ***
## P.Undergrad -1.963e-03  3.901e-04  -5.031 6.09e-07 ***
## Outstate     1.215e-03  2.270e-04   5.352 1.15e-07 ***
## Room.Board   1.534e-03  5.878e-04   2.610  0.00924 **
## Personal    -1.820e-03  7.638e-04  -2.383  0.01742 *
## PhD          8.424e-02  3.706e-02   2.273  0.02329 *
## perc.alumni  3.060e-01  4.806e-02   6.367 3.32e-10 ***
## Expend      -4.465e-04  1.390e-04  -3.211  0.00138 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.91 on 765 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.4351
## F-statistic: 55.33 on 11 and 765 DF,  p-value: < 2.2e-16
```

Question H)

Scatter plot of the studentised residuals against the predicted values

```
studentized_residuals <- rstudent(final_model)
predicted_values <- fitted(final_model)

ggplot(data.frame(studentized_residuals, predicted_values), aes(x =
predicted_values, y = studentized_residuals)) +
  geom_point(alpha = 0.5) +
  labs(title = "Studentized residuals vs. Predicted values", x =
"Predicted values", y = "Studentized residuals") +
  theme_minimal()
```

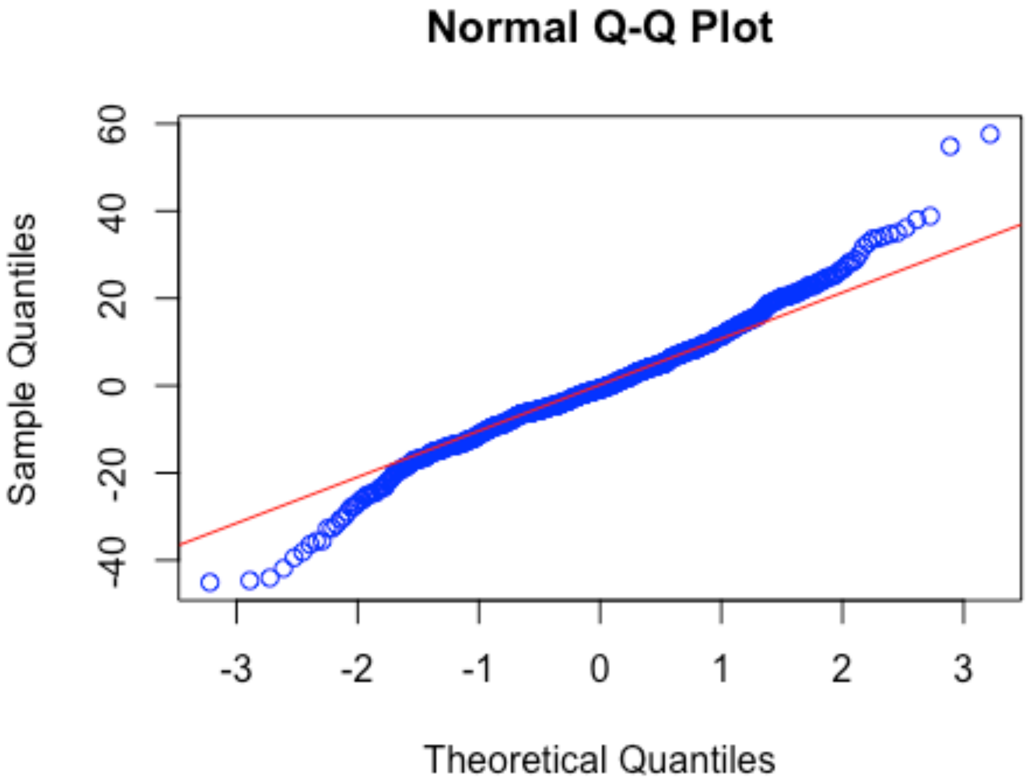


The residuals appear reasonably random, where variation first increased and then decreased hence we can see there is no constant Variation

Question I)

Analyze normal probability plot of residuals

```r
qqnorm(residuals(final_model), col = "blue")
qqline(residuals(final_model), col= "red")
```

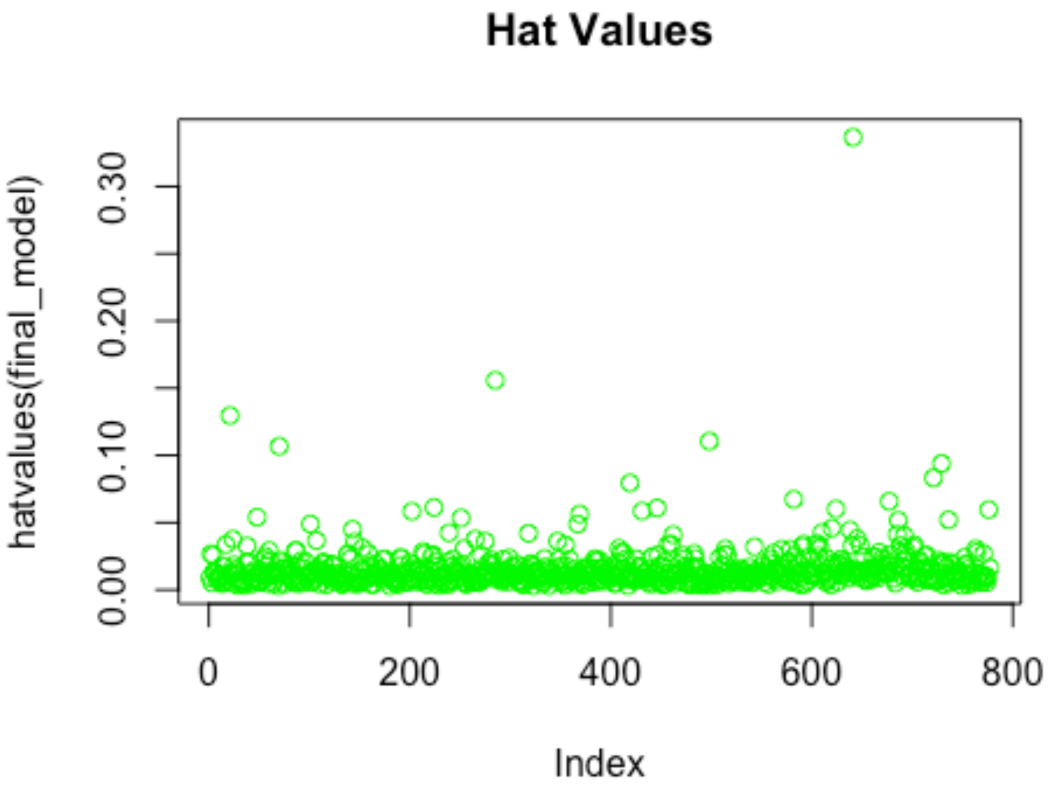


Upon examining the normal probability plot, we can observe that several spots deviate from the line. As a result, the model fits poorly.

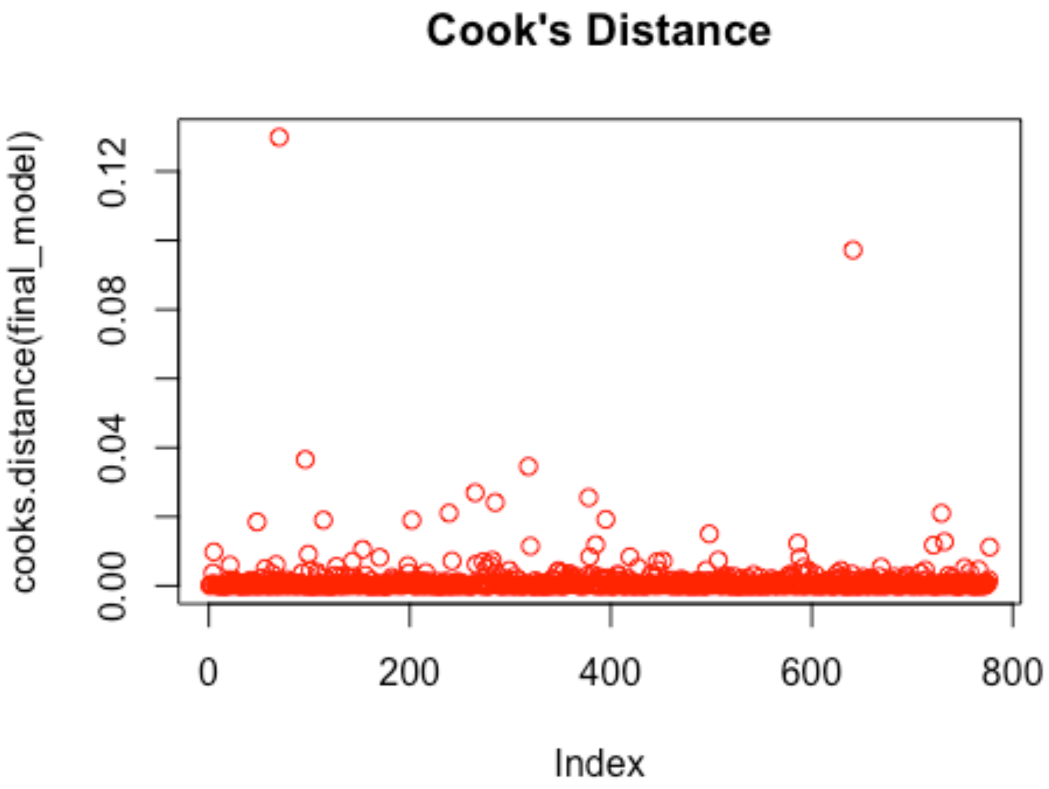
Question J)

Check for outliers or influential points

```r
plot(hatvalues(final_model), main="Hat Values", col = "green")
```

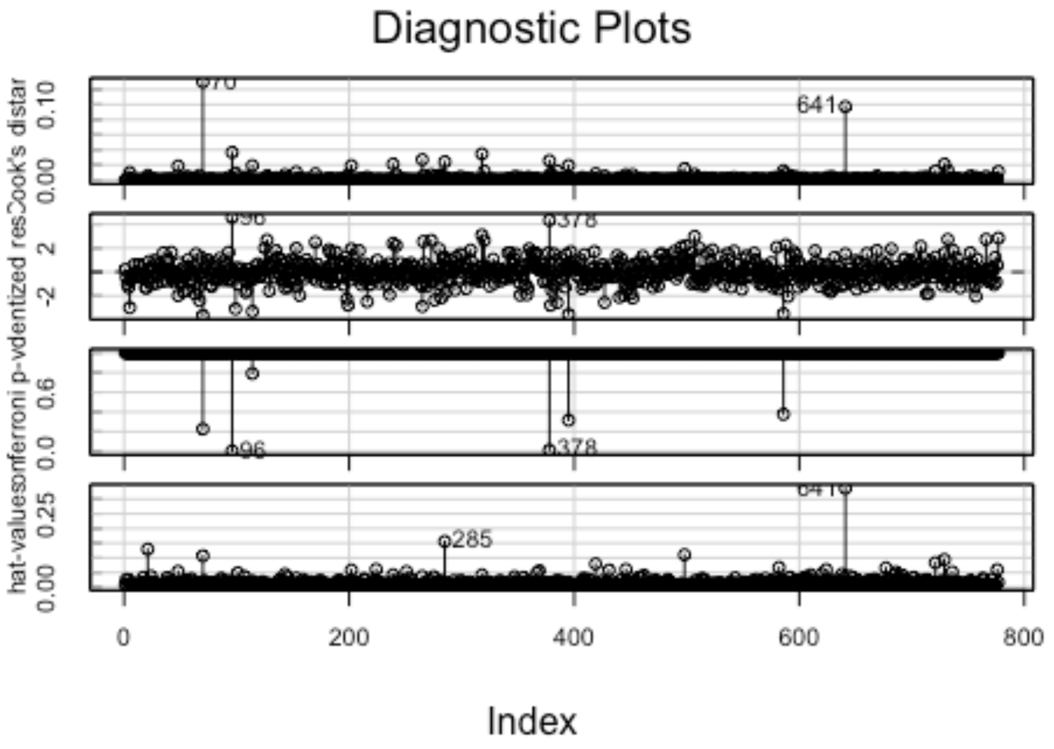


```r
plot(cooks.distance(final_model), main="Cook's Distance",col = "red")
```



Find indices of influential points with Cook’s distance > 1

```
influenceIndexPlot(final_model)
```



Positions in which outliers or influential points present are 70, 96, 285, 378, 641.

Question K)

Analyze the R2 value for the final model

```
summary(final_model)$r.squared
```

```
## [1] 0.4430843
```

Adjusted R-squared: 0.4351 #F-statistic: 55.33 #p-value: < 2.2e-16

R-squared: 0.4431 - suggest that it accounts for around 44.31% of the difference in university graduation rates. For forecasting graduation rates, the model has a modest level of explanatory power.

Question L)

Larger absolute coefficients and statistically significant (p-values) predictors may be regarded as more influential. Top 4 Significant predictors include perc.alumni, Outstate, P.Undergrad and Accept.pct.

Comparing Graduation Rates at Private and Public Universities: The coefficient for Private is positive (4.770) and statistically significant, meaning that private institutions typically have graduation rates that are 4.7 percentage points higher than those of public universities.

University Graduation Rates for "Elite" Universities: Elite10 has a positive (4.022) and statistically significant coefficient. This indicates that, in comparison to non-elite institutions, "elite" universities have graduation rates that are, on average, 4.02 percentage points higher.

PROBLEM SET 2

Question A)

Load necessary libraries

```
library(tidyverse)

## ── Attaching core tidyverse packages ────────────────────────
tidyverse 2.0.0 ──
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ lubridate 1.9.3     ✔ tibble    3.2.1
## ✔ purrr     1.0.2     ✔ tidyr     1.3.0
## ✔ readr     2.1.4
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ car::recode()   masks dplyr::recode()
## ✖ purrr::some()   masks car::some()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to
force all conflicts to become errors
```

Fit the regression model

```
interaction_model <- lm(Grad.Rate ~ Elite10 + Accept.pct + Outstate +
perc.alumni + Expend +
              Elite10:Accept.pct + Elite10:Outstate +
Elite10:perc.alumni + Elite10:Expend, data=data)
```

Display the summary of the model to check the significance of interaction terms

```
summary(interaction_model)

## 
## Call:
## lm(formula = Grad.Rate ~ Elite10 + Accept.pct + Outstate +
perc.alumni + 
##     Expend + Elite10:Accept.pct + Elite10:Outstate +
Elite10:perc.alumni + 
##     Elite10:Expend, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -53.787  -7.785  -0.400   7.769  57.177 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            5.316e+01  3.592e+00  14.801  < 2e-16 ***
## Elite101               3.763e+01  1.000e+01   3.762 0.000181 ***
## Accept.pct            -1.519e+01  4.129e+00  -3.678 0.000251 ***
## Outstate               2.296e-03  1.991e-04  11.532  < 2e-16 ***
## perc.alumni            3.505e-01  5.030e-02   6.968 6.95e-12 ***
## Expend                -9.536e-04  2.073e-04  -4.601 4.93e-06 ***
## Elite101:Accept.pct   -2.274e+01  9.822e+00  -2.315 0.020881 *  
## Elite101:Outstate     -2.054e-03  5.390e-04  -3.811 0.000150 ***
## Elite101:perc.alumni -1.227e-01  1.347e-01  -0.911 0.362485    
## Elite101:Expend        1.050e-03  2.889e-04   3.635 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 13.12 on 767 degrees of freedom
## Multiple R-squared:  0.4234, Adjusted R-squared:  0.4167 
## F-statistic: 62.58 on 9 and 767 DF,  p-value: < 2.2e-16
```

we will remove variable Elite10:perc.alumni as it is not significant.

Question B)

Simplified model, removing non-significant interaction terms

```
model_simplified <- lm(Grad.Rate ~ Elite10 + Accept.pct + Outstate +
perc.alumni + Expend +
                         Elite10:Accept.pct+Elite10:Outstate +
Elite10:Expend, data=data)
```

Display the summary of the simplified model

```
summary(model_simplified)

##
## Call:
## lm(formula = Grad.Rate ~ Elite10 + Accept.pct + Outstate +
perc.alumni +
##     Expend + Elite10:Accept.pct + Elite10:Outstate +
Elite10:Expend,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.724  -7.744  -0.468   7.727  57.150
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.314e+01  3.591e+00  14.797  < 2e-16 ***
## Elite101              3.585e+01  9.808e+00   3.655 0.000275 ***
## Accept.pct           -1.505e+01  4.126e+00  -3.647 0.000283 ***
## Outstate              2.322e-03  1.970e-04  11.786  < 2e-16 ***
## perc.alumni           3.334e-01  4.666e-02   7.145 2.09e-12 ***
## Expend               -9.506e-04  2.072e-04  -4.587 5.24e-06 ***
## Elite101:Accept.pct -2.164e+01  9.747e+00  -2.220 0.026705 *
## Elite101:Outstate   -2.253e-03  4.926e-04  -4.575 5.56e-06 ***
## Elite101:Expend      1.057e-03  2.888e-04   3.661 0.000268 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 13.12 on 768 degrees of freedom
## Multiple R-squared:  0.4228, Adjusted R-squared:  0.4168
## F-statistic: 70.32 on 8 and 768 DF,  p-value: < 2.2e-16
```

$$\mathrm{Grad.Rate}=53.14+35.85\times \mathrm{Elite10}-15.05\times \mathrm{Accept.pct}+0.002322\times \mathrm{Outstate}+0.3334\times \mathrm{perc.alumni}-0.0009506\times \mathrm{Expend}-21.64\times \mathrm{Elite10}\times \mathrm{Accept.pct}-0.002253\times \mathrm{Elite10}\times \mathrm{Outstate}+0.001057\times \mathrm{Elite10}\times \mathrm{Expend}$$

Question C)

Extract coefficients from the simplified model

```
coefficients(model_simplified)
```

```
##          (Intercept)             Elite101           Accept.pct 
Outstate 
##         5.313531e+01         3.584905e+01        -1.504755e+01 
2.322220e-03 
##          perc.alumni               Expend Elite101:Accept.pct 
Elite101:Outstate 
##         3.333751e-01        -9.505835e-04        -2.163809e+01 
-2.253419e-03 
##      Elite101:Expend 
##         1.057323e-03
```

The coefficient for Elite10 is 35.85. This means, all else being constant, being an "Elite10" university is associated with an increase of 35.85 units in graduation rate compared to a non-Elite10 university.

The coefficient for the interaction term Elite10:Accept.pct is -21.64. This means that the relationship between Accept.pct and Grad.Rate is moderated by whether the university is an Elite10. For Elite10 universities, a unit increase in Accept.pct is associated with a decrease in Grad.Rate of (15.05 - 21.64) = -36.69 units, whereas for non-Elite10 universities, the decrease is 15.05 units. This shows that the negative effect of Accept.pct on Grad.Rate is stronger for Elite10 universities.

The coefficient for Elite10:Outstate is -0.002253. While a unit increase in Outstate is associated with an increase of 0.002322 units in Grad.Rate for non-Elite10 universities, this effect is reduced by 0.002253 units for Elite10 universities, resulting in an almost negligible increase of 0.000069 units. Thus, being an Elite10 university dampens the positive effect of Outstate on Grad.Rate.

There is no interaction term between Elite10 and perc.alumni in the model. This means that the effect of perc.alumni on Grad.Rate is the same regardless of whether a university is an Elite10 or not. The effect is an increase of 0.3334 units in Grad.Rate for every unit increase in perc.alumni.

The coefficient for Elite10:Expend is 0.001057. This suggests that for Elite10 universities, a unit increase in Expend increases the Grad.Rate by (0.001057 - 0.0009506) = 0.0001064 units more than non-Elite10 universities. In other words, the negative effect of Expend on Grad.Rate is slightly mitigated for Elite10 universities.

The association between Grad.Rate and the predictors is impacted differently by being a “Elite10” university. It reduces the good impact of Outstate, increases the bad impact of Accept.pct, and somewhat lessens the negative impact of Expend. Regardless of whether the university is ranked in the Elite10, perc.alumni has the same impact.

Question D)

Load required library

```
library(caret)

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

Split the data into training and testing sets (70-30 split)

```
set.seed(123)
splitIndex <- createDataPartition(data$Grad.Rate, p=0.7, list=FALSE)
train_data <- data[splitIndex, ]
test_data <- data[-splitIndex, ]
```

Fit the simplified model on training data

```
model_train <- lm(Grad.Rate ~ Elite10 + Accept.pct + Outstate +
perc.alumni + Expend +
                    Elite10:Outstate + Elite10:Expend,
data=train_data)
```

Predict on test data

```
predictions <- predict(model_train, newdata=test_data)
```

Compute the MAPE statistic

```
MAPE <- mean(abs((predictions - test_data$Grad.Rate) /
test_data$Grad.Rate))

MAPE

## [1] 0.1753778
```

Question E)

```
set.seed(345)
```

Define the control parameters for cross-validation

```
library(Metrics)

##
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':
##
##     precision, recall

ctrl <- trainControl(method = "cv", number = 5)

ModelM2_CV <- train(Grad.Rate ~ Elite10+ Accept.pct+ Outstate+
perc.alumni + Expend + Elite10 * Accept.pct + Elite10 * Outstate +
Elite10 * Expend,
                                data = data,method = "lm", trControl =
ctrl)
```

```r
summary(ModelM2_CV)

## 
## Call:
## lm(formula = .outcome ~ ., data = dat)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -53.724  -7.744  -0.468   7.727  57.150 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            5.314e+01  3.591e+00  14.797  < 2e-16 ***
## Elite101               3.585e+01  9.808e+00   3.655 0.000275 ***
## Accept.pct            -1.505e+01  4.126e+00  -3.647 0.000283 ***
## Outstate               2.322e-03  1.970e-04  11.786  < 2e-16 ***
## perc.alumni            3.334e-01  4.666e-02   7.145 2.09e-12 ***
## Expend                -9.506e-04  2.072e-04  -4.587 5.24e-06 ***
## `Elite101:Accept.pct` -2.164e+01  9.747e+00  -2.220 0.026705 *  
## `Elite101:Outstate`   -2.253e-03  4.926e-04  -4.575 5.56e-06 ***
## `Elite101:Expend`      1.057e-03  2.888e-04   3.661 0.000268 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 13.12 on 768 degrees of freedom
## Multiple R-squared:  0.4228, Adjusted R-squared:  0.4168 
## F-statistic: 70.32 on 8 and 768 DF,  p-value: < 2.2e-16

predictions_M2 <- predict(ModelM2_CV, newdata = data)

mape_M2 <- mape(data$Grad.Rate, predictions_M2)
cat("MAPE for Model M2:", mape_M2, "\n")

## MAPE for Model M2: 0.1857689
```