DATA ANALYSIS AND REGRESSION

ASSIGNMENT_2

Mrunali Vikas Patil

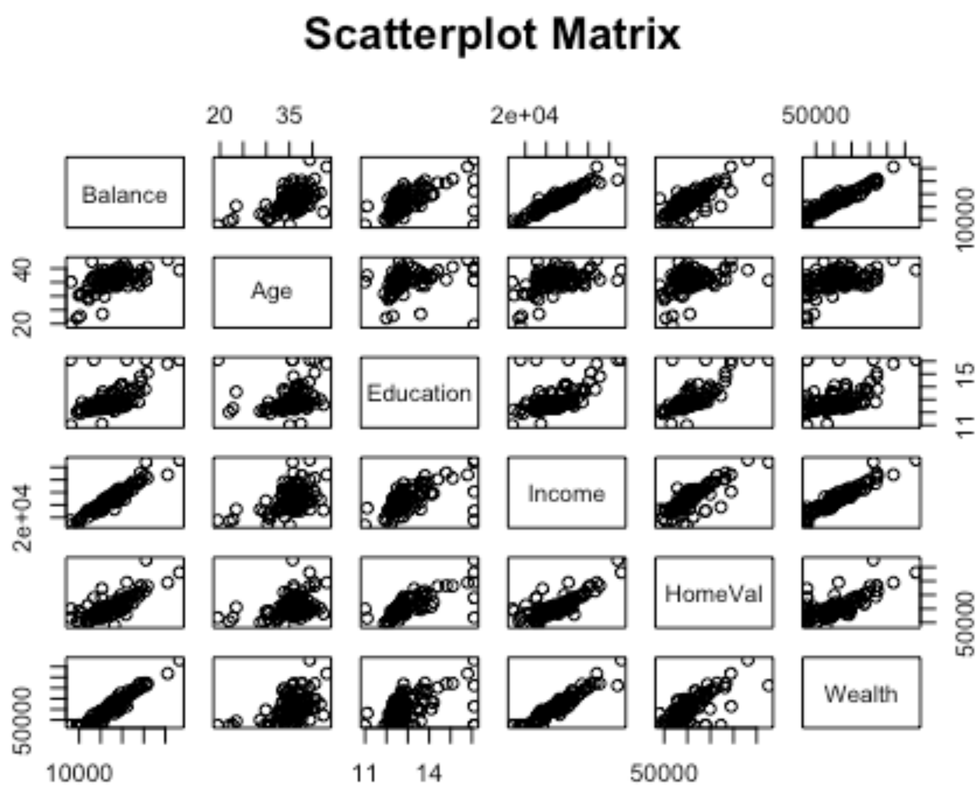
#PROBLEM SET 1

Importing data in R

```
data=read.table("Bankingfull.txt",header = TRUE)
dim(data)

## [1] 102   6

head(data)

##    Age Education Income HomeVal Wealth Balance
## 1 35.9      14.8  91033  183104 220741   38517
## 2 37.7      13.8  86748  163843 223152   40618
## 3 36.8      13.8  72245  142732 176926   35206
## 4 35.3      13.2  70639  145024 166260   33434
## 5 35.3      13.2  64879  135951 148868   28162
## 6 34.8      13.7  75591  155334 188310   36708
```

Question A) Scatterplots

```
pairs(~Balance + Age + Education + Income + HomeVal + Wealth,
data=data, main="Scatterplot Matrix")
```



The scatterplot matrix indicates that the variables wealth and income have a significant positive linear relationship with balance. Additionally, we cannot detect any outliers in those two variables. # Variable HomeVal and Balance have a solid linear relationship. # Although not as strongly as other variables, age and education also exhibit a linear relationship with balance. These two variables also contain outliers. # There is a strong or weak linear relationship between all variables.

Question B) Correlations

```
cor(data$Balance, data$Age)

## [1] 0.5654668

cor(data$Balance, data$Education)

## [1] 0.5548807

cor(data$Balance, data$Income)
```

```
## [1] 0.9516845

cor(data$Balance, data$HomeVal)

## [1] 0.7663871

cor(data$Balance, data$Wealth)

## [1] 0.9487117
```

The correlation matrix reveals a high positive correlation between the variables Wealth and Income and the target variable Balance.

The dependent variable Balance and the independent variables Age, Education, and HomeVal show a moderately positive connection.

There is a weak correlation between Variable Age and Education.

Question C) Regression Model M1 #Fit the Regression Model

```
Model_M1 <- lm(Balance ~ Age + Education + Income + HomeVal + Wealth,
data=data)
summary(Model_M1)

##
## Call:
## lm(formula = Balance ~ Age + Education + Income + HomeVal + Wealth,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5376.9 -1110.8   -77.2   872.3  7732.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.071e+04  4.261e+03  -2.514 0.013613 *
## Age          3.187e+02  6.099e+01   5.225 1.01e-06 ***
## Education    6.219e+02  3.190e+02   1.950 0.054135 .
```

```
## Income       1.463e-01  4.078e-02   3.588 0.000527 ***
## HomeVal      9.183e-03  1.104e-02   0.832 0.407505
## Wealth       7.433e-02  1.119e-02   6.643 1.85e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2056 on 96 degrees of freedom
## Multiple R-squared:  0.9469, Adjusted R-squared:  0.9441
## F-statistic: 342.4 on 5 and 96 DF,  p-value: < 2.2e-16
```

Load necessary library for VIF computation

```
library(car)

## Loading required package: carData
```

Compute VIF Statistics

```
vif<- vif(Model_M1)
```

VIF Value above 10 indicates high multicollinearity. # We examined the VIF statistics for the model M1 and discovered that the VIF factors for Variable Income and Wealth were greater than 10. #Thus, we can draw the conclusion that the variables Income and Wealth have a multi-collinearity issue.

Question D) Improved Regression Model M2

1] The variable Income will be removed from the model first because it has the greatest VIF value. # We will also remove the variable HomeVal because it has a higher P value.

```
Model_M2 <- lm(Balance ~ Age + Education + Wealth, data=data)
summary(Model_M2)

##
## Call:
## lm(formula = Balance ~ Age + Education + Wealth, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7330.6 -1096.7    -5.5   872.9  7087.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -1.773e+04  3.802e+03  -4.664 9.80e-06 ***
## Age          3.678e+02  6.460e+01   5.694 1.30e-07 ***
## Education    1.300e+03  2.500e+02   5.202 1.08e-06 ***
## Wealth       1.165e-01  4.680e-03  24.887  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2226 on 98 degrees of freedom
## Multiple R-squared:  0.9365, Adjusted R-squared:  0.9345
## F-statistic: 481.5 on 3 and 98 DF,  p-value: < 2.2e-16
```

Compute VIF Statistics

```
vif<- vif(Model_M2)
```

R-squared and adjusted R-squared

```
summary(Model_M1)$adj.r.squared

## [1] 0.9441433

summary(Model_M2)$adj.r.squared

## [1] 0.9345196
```

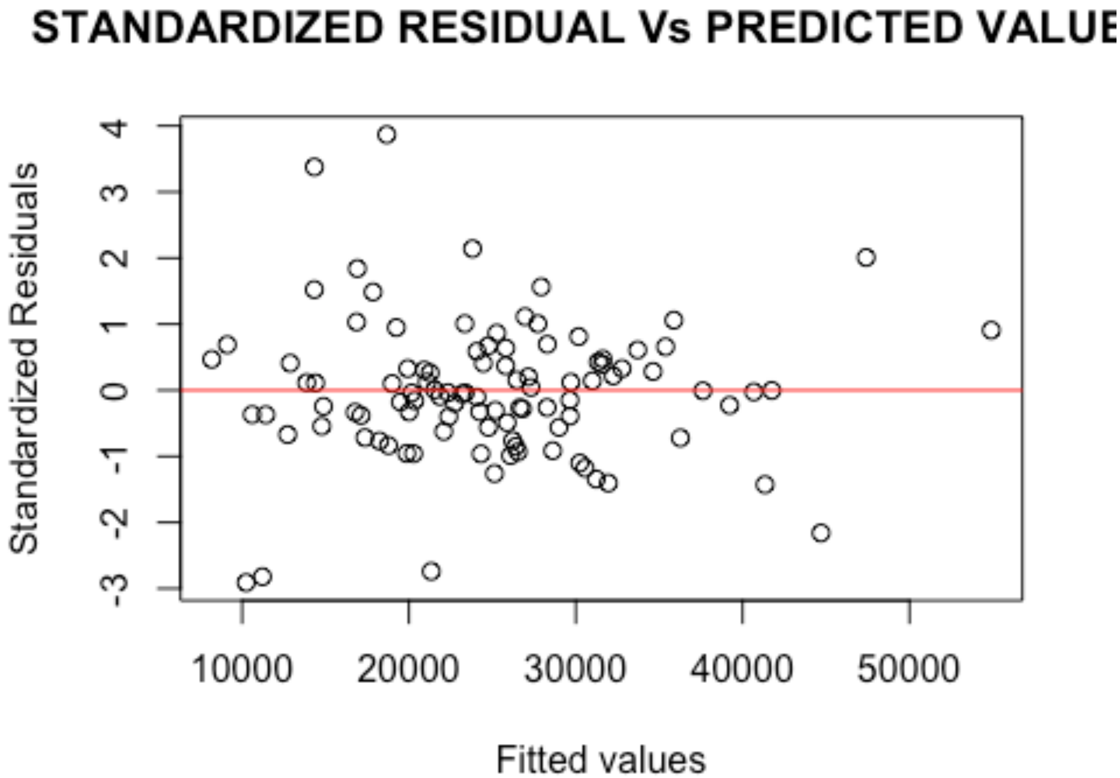
We can determine that adjusted R-Squared for the Model_M1 has higher values than Model_M2 after refitting the Model.

2]Residual Analysis

```
par(mfrow=c(2,2))
```

Standardized Residuals Vs Predicted Values

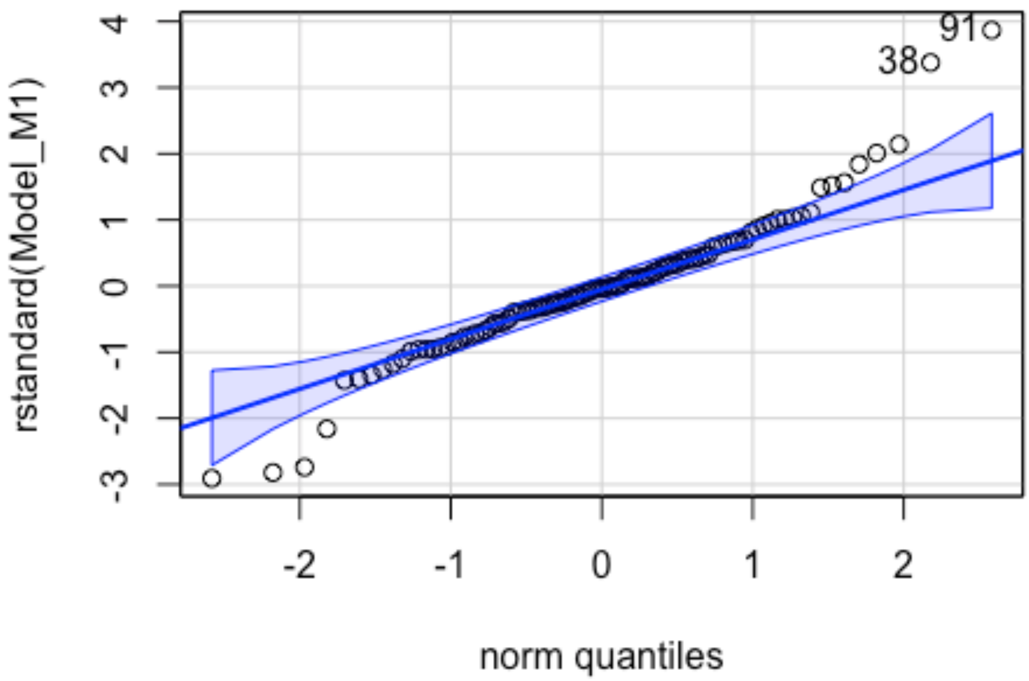
```
plot(fitted(Model_M1), rstandard(Model_M1), main="STANDARDIZED
RESIDUAL Vs PREDICTED VALUES ", xlab="Fitted values",
ylab="Standardized Residuals")
abline(h=0, col="red")
```



The plot indicates that the residuals show less variation, thus we can conclude that the model is good. Additionally, there are 2 to 3 outlier points.

Normal plot of residuals

```
qqPlot(rstandard(Model_M1))
```




```
## [1] 91 38
```

3]Finding Outliers and Influential Points

```
outliers <- which(abs(rstandard(Model_M1)) > 3)
print(outliers)

## 38 91
## 38 91
```

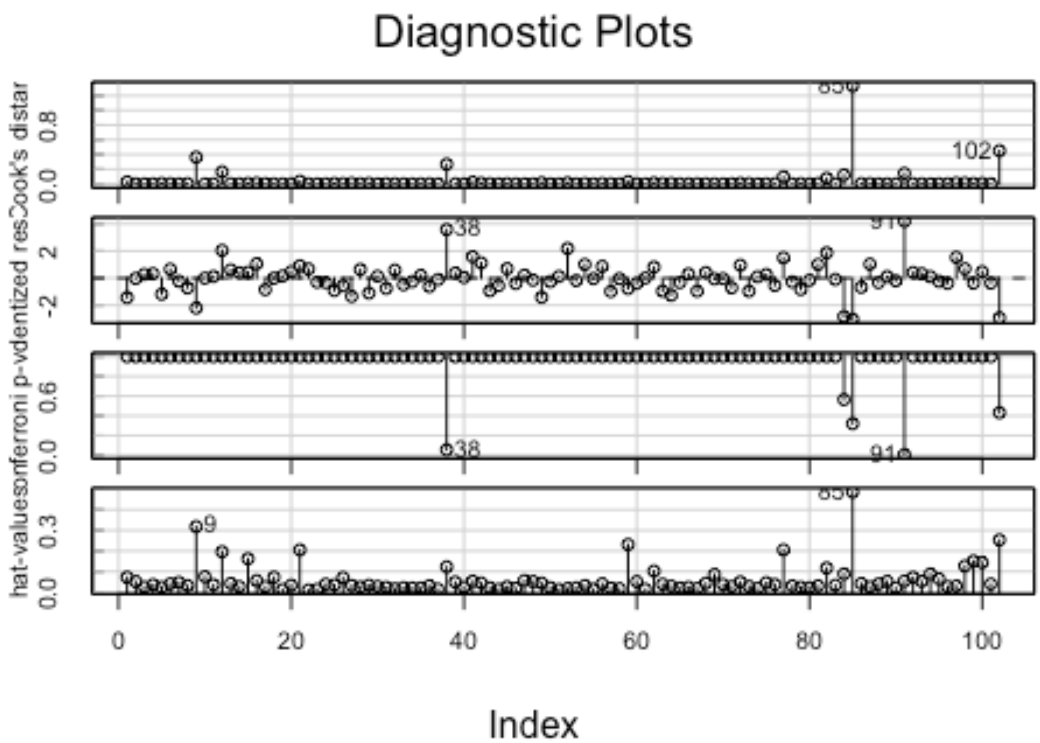
We can see there are less number of outliers also we are not able to see only one influential point in the fitted Model.

```
influential_points <- which(cooks.distance(Model_M1) > 1)
print(influential_points)
```

```
## 85
## 85
```

Graph for the Influence Point

```r
influenceIndexPlot(Model_M1)
```



4]Standardized Coefficients

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

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

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

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

library(QuantPsyc)

## Loading required package: boot

##
## Attaching package: 'boot'

## The following object is masked from 'package:car':
##
##     logit

## Loading required package: purrr

##
## Attaching package: 'purrr'

## The following object is masked from 'package:car':
##
##     some

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:base':
##
##     norm

lm.beta(Model_M1)
```

```
##        Age  Education     Income    HomeVal     Wealth
## 0.14239029 0.07186393 0.32572524 0.04095974 0.51136385
```

The standardized coefficients show that the variable "Wealth" has the greatest impact on the target variable "Balance."

Question E)Prediction # New data for prediction

```
new_data <- data.frame(Age = 34, Education = 13, Income = 64000,
HomeVal = 140000, Wealth = 160000)
predicted_balance <- predict(Model_M1, newdata=new_data,
interval="confidence", level=0.95)
print(predicted_balance)

##        fit      lwr      upr
## 1 30751.53 29952.27 31550.78
```

#Predicted average bank balance = 30751.53 #Lower 95% Confidence Interval = 29952.27 #Upper 95% Confidence Interval = 31550.78

#PROBLEM SET 2

Importing data in R

```
pgatour=read.csv("pgatour2006_small.csv",header = TRUE)
dim(pgatour)

## [1] 196   7

head(pgatour)

##                Name PrizeMoney DrivingAccuracy   GIR PuttingAverage
## 1   Aaron Baddeley      60661           60.73 58.26          1.745
## 2       Adam Scott     262045           62.00 69.12          1.767
## 3      Alex Aragon       3635           51.12 59.11          1.787
## 4       Alex Cejka      17516           66.40 67.70          1.777
## 5      Arjun Atwal      16683           63.24 64.04          1.761
## 6 Arron Oberholser     107294           62.53 69.27          1.775
##   BirdieConversion PuttsPerRound
## 1            31.36         27.96
## 2            30.39         29.28
```

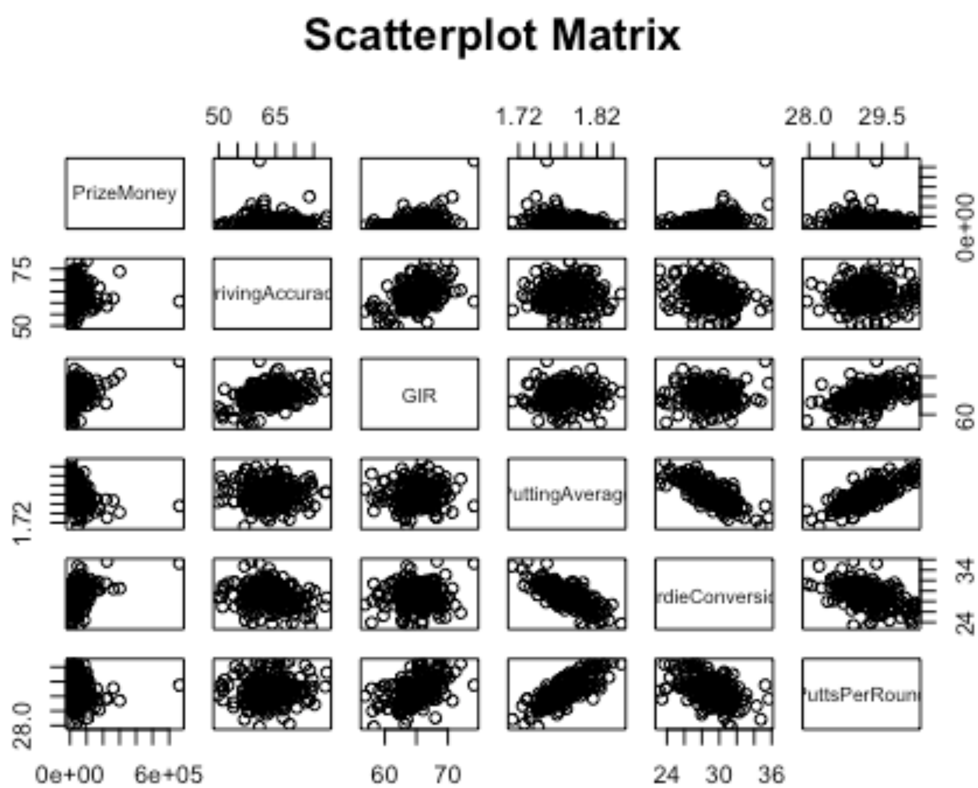
```
## 3              29.89        29.20
## 4              29.33        29.46
## 5              29.32        28.93
## 6              29.20        29.56
```

Remove the variable "Name" as it has unique values.

```
pgatour <- pgatour[, !names(data) %in% "Name"]
```

#Question 1)Scatterplots of PrizeMoney vs. other variables

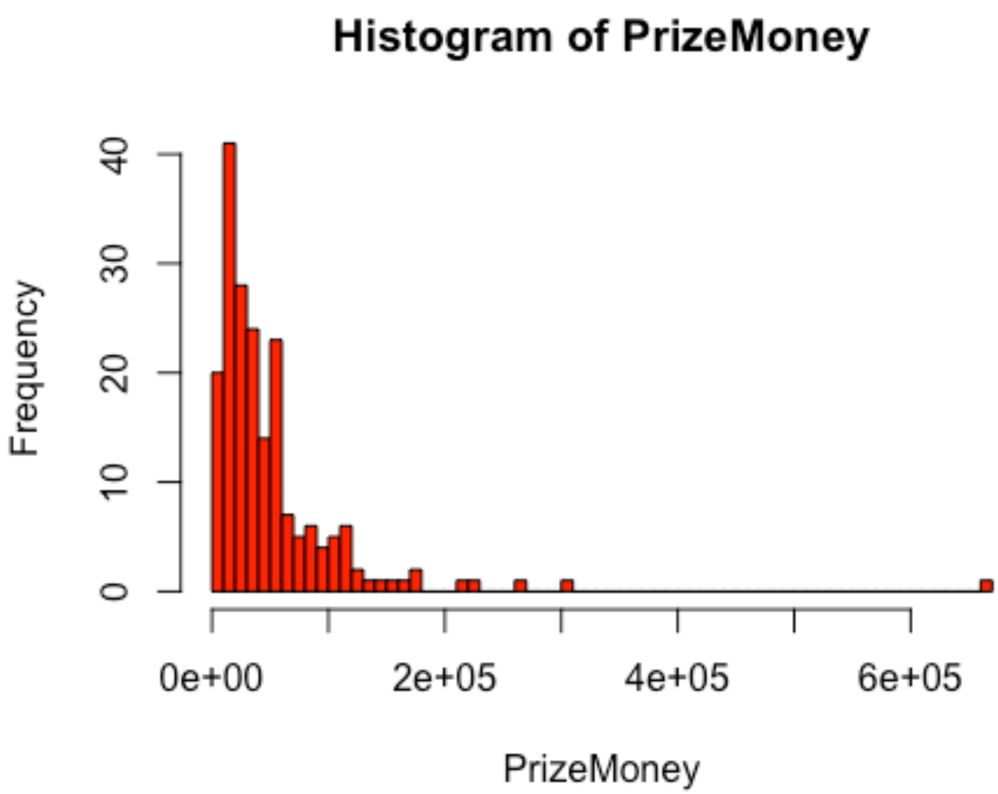
```
pairs(~PrizeMoney+DrivingAccuracy+GIR+PuttingAverage+BirdieConversion+
PuttsPerRound,data = pgatour,main = "Scatterplot Matrix")
```



We cannot find a linear relationship between the target variable "PrizeMoney" and the other variables from the scatterplot.

Question 2)Histogram of PrizeMoney

```
par(mfrow=c(1,1))
hist(pgatour$PrizeMoney, main="Histogram of PrizeMoney",
xlab="PrizeMoney", breaks=50, col = "red", border = "black")
```

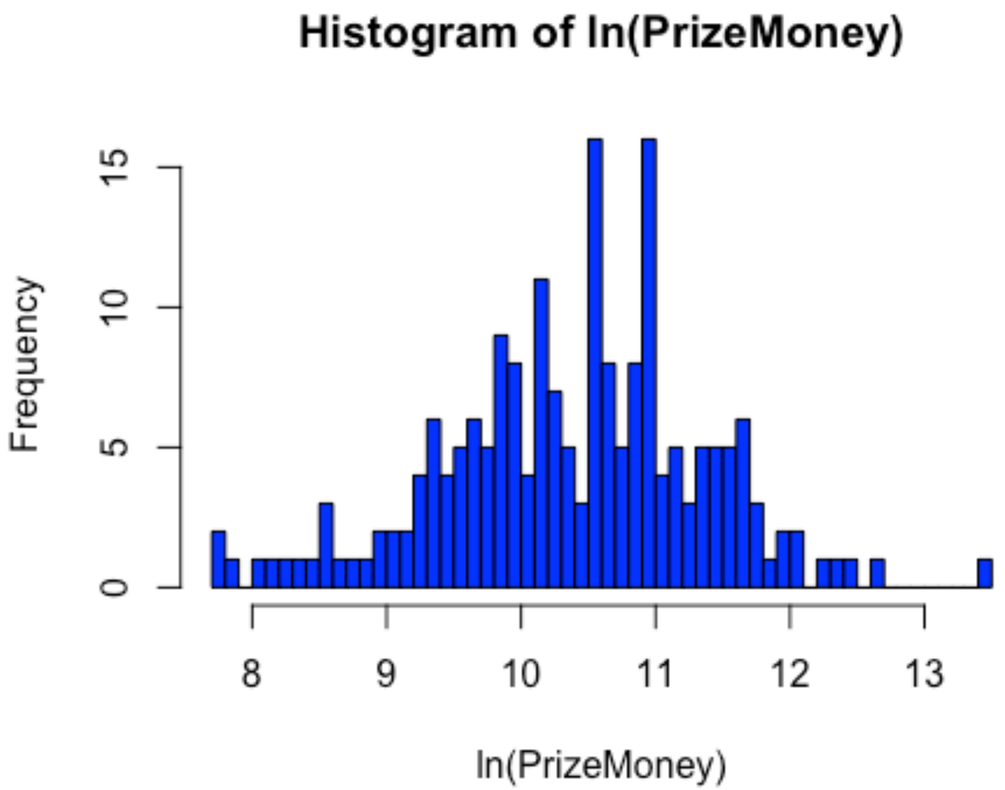


The histogram of PrizeMoney shows a right-skewed distribution

This indicates that while the majority of players receive lesser prize amounts, a small number of players receive much greater amounts, resulting to the long tail on the right side of the distribution.

Question 3)Applying log transformation

```
pgatour$ln_Prize <- log(pgatour$PrizeMoney)
hist(pgatour$ln_Prize, main="Histogram of ln(PrizeMoney)",
xlab="ln(PrizeMoney)", breaks=50, col = "blue", border = "black")
```



The histogram of ln (PrizeMoney) looks to be more symmetric after log transformation of variable “PrizeMoney”.

The distribution is now more bell-shaped, which is closer to a normal distribution, even though there is still a tiny right skew.

Question 4)

Fit the regression model to predict ln(PrizeMoney) and evaluating the significance of each predictor

```
Model1 <- lm(ln_Prize ~ DrivingAccuracy + GIR + BirdieConversion +
PuttingAverage + PuttsPerRound, data=pgatour)
```

1]

```
summary(Model1)
```

```
##
## Call:
## lm(formula = ln_Prize ~ DrivingAccuracy + GIR + BirdieConversion +
##     PuttingAverage + PuttsPerRound, data = pgatour)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.55696 -0.51250 -0.08005  0.45090  2.11898
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.2410192  7.1611241   1.151 0.251261
## DrivingAccuracy  -0.0007584  0.0116109  -0.065 0.947992
## GIR               0.2687898  0.0287938   9.335  < 2e-16 ***
## BirdieConversion  0.1523018  0.0408329   3.730 0.000253 ***
## PuttingAverage    8.7467774  5.3734220   1.628 0.105228
## PuttsPerRound    -1.2094847  0.2672761  -4.525 1.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6725 on 190 degrees of freedom
## Multiple R-squared:  0.5414, Adjusted R-squared:  0.5293
## F-statistic: 44.86 on 5 and 190 DF,  p-value: < 2.2e-16
```

By Looking at the summary, The adjusted R-Squared is 0.5293

DrivingAccuracy has a high p-value of 0.948, showing that it is not a significant predictor #Hence we will remove the DrivingAccuracy and then refit the model

```
Model2 <- lm(ln_Prize ~
GIR+PuttingAverage+BirdieConversion+PuttsPerRound, data=pgatour)
summary(Model2)
```

```
##
## Call:
## lm(formula = ln_Prize ~ GIR + PuttingAverage + BirdieConversion +
##     PuttsPerRound, data = pgatour)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.55608 -0.51122 -0.08109  0.45250  2.12227
```

```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         8.02738    6.35383   1.263   0.2080
## GIR                 0.26791    0.02536  10.563  < 2e-16 ***
## PuttingAverage      8.81065    5.26991   1.672   0.0962 .
## BirdieConversion    0.15360    0.03561   4.314 2.57e-05 ***
## PuttsPerRound      -1.20702    0.26391  -4.574 8.61e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6707 on 191 degrees of freedom
## Multiple R-squared:  0.5414, Adjusted R-squared:  0.5318
## F-statistic: 56.37 on 4 and 191 DF,  p-value: < 2.2e-16
```

PuttingAverage has a p-value of 0.0962, indicating that it might not be a strong predictor Hence we will remove the PuttingAverage and then refit the model.

```r
Model3 <- lm(ln_Prize ~ GIR+BirdieConversion+PuttsPerRound,
data=pgatour)
summary(Model3)
```

```
##
## Call:
## lm(formula = ln_Prize ~ GIR + BirdieConversion + PuttsPerRound,
##     data = pgatour)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6140 -0.5152 -0.0761  0.4540  2.0583
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         15.8102     4.3446   3.639 0.000352 ***
## GIR                  0.2454     0.0216  11.360  < 2e-16 ***
## BirdieConversion     0.1145     0.0270   4.243 3.43e-05 ***
## PuttsPerRound       -0.8476     0.1538  -5.512 1.13e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6738 on 192 degrees of freedom
```

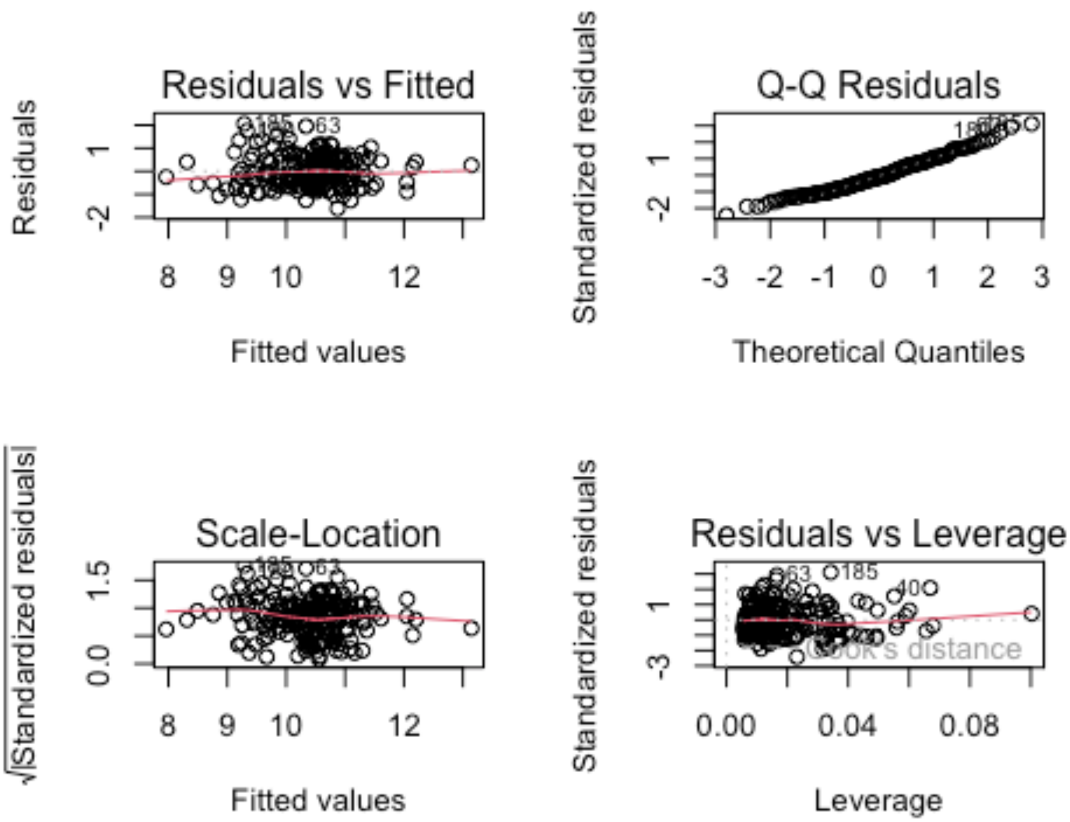
```
## Multiple R-squared:  0.5347, Adjusted R-squared:  0.5274
## F-statistic: 73.54 on 3 and 192 DF,  p-value: < 2.2e-16
```

After Refiting the model, all variables now appear to be significant.

AdjustedR-squared : 0.5274

2] Residual plots

```
par(mfrow=c(2,2))
plot(Model3)
```



By examining the Residuals vs Fitted plot, we can conclude that the model appears to be valid because variation is considerably lower and there are fewer outliers. # The Q-Q plot shows that there are several spots that follow the line.

3]Outliers and Influential points

```
residuals_standardize <- rstandard(Model3)
outliers_position <- which(residuals_standardize > 3)
```

Outliers

```
residuals_standardize[outliers_position]

##      185
## 3.108311
```

We can call these points as outliers because these points are located 3 standard deviations away from the mean.

Question 5)

```
coefficients <- coef(Model3)
coefficients

##      (Intercept)              GIR BirdieConversion    PuttsPerRound
##       15.8101628        0.2454205        0.1145444       -0.8475661
```

For every 1% rise in GIR, the coefficient for GIR indicates the change in PrizeMoney. Keeping all other variables constant, we expect an average rise in PrizeMoney of exp(`0.2454205`) times.

Question 6)Prediction

```
new_data <- data.frame(DrivingAccuracy = 64, GIR = 67,
BirdieConversion = 28, PuttingAverage = 1.77, PuttsPerRound = 29.16)
```

Predictions alongside with the 95% Prediction Interval

```
prediction <- predict(Model3, newdata = new_data, interval =
"prediction", level = 0.95)
print(prediction)

##        fit      lwr      upr
## 1 10.74555 9.407982 12.08312
```

```r
Predicted_Balance <- exp(10.74555)
lower_Limit <- exp(9.407982)
upper_limit <- exp(12.08312)

print(Predicted_Balance)

## [1] 46422.99

print(lower_Limit)

## [1] 12185.26

print(upper_limit)

## [1] 176861.1
```