FinalPPTRcode.R

Jadhav

2023-12-11

```
#Final Project

library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-7
```

```
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following objects are masked from 'package:Matrix':
##
##     expand, pack, unpack
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.2
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
## x tidyr::pack()   masks Matrix::pack()
## x tidyr::unpack() masks Matrix::unpack()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(ggplot2)
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 4.3.2
```

```r
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.3.2
```

```r
library(leaps)
library(MASS)
```

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

```r
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
#import the data set
shaq <- read.csv("E:\\GMU\\STAT 515\\Final Project\\shaq-nba-career-regular-season-stats-by-game.csv")

#Checking the dataset
head(shaq, 10)
```

```
##    Season SeasGm CarrGm  Date     Age  Tm Home Opp Win teamdiff GS Minutes FG
## 1       1      1      1 33914 20.6708 ORL    1 MIA   1       10  1      32  4
## 2       1      2      2 33915 20.6735 ORL    0 WSB   1        5  1      40  8
## 3       1      3      3 33918 20.6817 ORL    1 CHH   0       -4  1      34 15
## 4       1      4      4 33920 20.6872 ORL    1 WSB   1       27  1      36 12
## 5       1      5      5 33922 20.6927 ORL    0 NJN   0      -11  1      35  9
## 6       1      6      6 33926 20.7036 ORL    0 PHI   1       10  1      34 12
## 7       1      7      7 33927 20.7064 ORL    1 GSW   1       24  1      41  8
## 8       1      8      8 33929 20.7118 ORL    0 NYK   0      -15  1      44  7
## 9       1      9      9 33933 20.7228 ORL    1 HOU   1       13  1      42  6
## 10      1     10     10 33935 20.7283 ORL    0 IND   1       14  1      30  7
##    FGA   FG. X3P X3PA X3P. FT FTA   FT. ORB DRB TRB AST STL BLK TOV PF PTS GmSc
## 1    8 0.500   0    0   NA  4   7 0.571   5  13  18   2   1   3   8  6  12  8.3
## 2   16 0.500   0    0   NA  6  11 0.545   5  10  15   1   0   4   4  5  22 16.0
## 3   25 0.600   0    0   NA  5   8 0.625   4   9  13   1   1   3   4  4  35 26.0
## 4   19 0.632   0    0   NA  7  12 0.583   9  12  21   1   0   4   6  4  31 26.3
## 5   16 0.563   0    0   NA 11  16 0.688   5  10  15   1   1   3   2  4  29 26.1
## 6   19 0.632   0    0   NA  5  11 0.455   7  12  19   1   1   3   3  5  29 25.4
## 7   15 0.533   0    0   NA 13  18 0.722   5  11  16   3   2   3   4  3  29 27.5
## 8   18 0.389   0    0   NA  4  11 0.364   3  14  17   2   1   3   7  4  18  7.6
## 9   12 0.500   0    0   NA  0   0    NA   1  12  13   2   1   3   2  3  12 11.6
```

```
## 10  10 0.700   0    0   NA  7  13 0.538   3   8  11   1   1   4   4  4  21 17.8
##    Pls.Mns
## 1       NA
## 2       NA
## 3       NA
## 4       NA
## 5       NA
## 6       NA
## 7       NA
## 8       NA
## 9       NA
## 10      NA
```

```
tail(shaq, 10)
```

```
##      Season SeasGm CarrGm  Date     Age  Tm Home Opp Win teamdiff GS Minutes FG
## 1198     19     28   1198 40553 38.8487 BOS    1 HOU   0       -6  1   20.58  3
## 1199     19     29   1199 40555 38.8542 BOS    1 SAC   1       24  1   13.42  1
## 1200     19     30   1200 40557 38.8597 BOS    1 CHA   1        5  1   35.22 10
## 1201     19     31   1201 40560 38.8679 BOS    1 ORL   1        3  1   25.65  5
## 1202     19     32   1202 40562 38.8734 BOS    1 DET   1        4  1   25.02  5
## 1203     19     33   1203 40564 38.8789 BOS    1 UTA   1       24  1    6.32  1
## 1204     19     34   1204 40571 38.8980 BOS    0 PHO   0      -17  1   15.00  2
## 1205     19     35   1205 40573 38.9035 BOS    0 LAL   1       13  1   12.70  0
## 1206     19     36   1206 40575 38.9090 BOS    0 SAC   1        5  1   15.78  1
## 1207     19     37   1207 40636 39.0767 BOS    1 DET   1       11  0    5.48  3
##      FGA   FG. X3P X3PA X3P. FT FTA FT. ORB DRB TRB AST STL BLK TOV PF PTS GmSc
## 1198   3 1.000   0    0   NA  4   4 1.0   0   2   2   2   0   0   0  1  10 10.7
## 1199   2 0.500   0    0   NA  0   0  NA   1   0   1   1   1   1   3  4   2 -0.5
## 1200  12 0.833   0    0   NA  3   3 1.0   2   3   5   2   0   5   3  4  23 21.2
## 1201   7 0.714   0    0   NA  2   2 1.0   0   2   2   1   1   2   2  5  12  8.8
## 1202   9 0.556   0    0   NA  2   5 0.4   5   7  12   0   3   2   0  2  12 15.7
## 1203   1 1.000   0    0   NA  0   0  NA   0   0   0   1   1   0   0  1   2  3.0
## 1204   4 0.500   0    0   NA  1   2 0.5   0   4   4   1   1   2   0  3   5  5.7
## 1205   2 0.000   0    0   NA  0   0  NA   2   4   6   0   0   2   2  5   0 -1.4
## 1206   3 0.333   0    0   NA  1   2 0.5   0   4   4   0   0   1   3  3   3 -1.4
## 1207   3 1.000   0    0   NA  0   0  NA   0   1   1   0   0   0   1  0   6  4.4
##      Pls.Mns
## 1198      -8
## 1199      11
## 1200      19
## 1201       4
## 1202      -3
## 1203       7
## 1204       0
## 1205      -4
## 1206      -2
## 1207      -2
```

```
str(shaq)
```

```
## 'data.frame':    1207 obs. of  32 variables:
##  $ Season  : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ SeasGm  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ CarrGm  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Date    : int  33914 33915 33918 33920 33922 33926 33927 33929 33933 33935 ...
##  $ Age     : num  20.7 20.7 20.7 20.7 20.7 ...
##  $ Tm      : chr  "ORL" "ORL" "ORL" "ORL" ...
##  $ Home    : int  1 0 1 1 0 0 1 0 1 0 ...
##  $ Opp     : chr  "MIA" "WSB" "CHH" "WSB" ...
##  $ Win     : int  1 1 0 1 0 1 1 0 1 1 ...
##  $ teamdiff: int  10 5 -4 27 -11 10 24 -15 13 14 ...
##  $ GS      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Minutes : num  32 40 34 36 35 34 41 44 42 30 ...
##  $ FG      : int  4 8 15 12 9 12 8 7 6 7 ...
##  $ FGA     : int  8 16 25 19 16 19 15 18 12 10 ...
##  $ FG.     : num  0.5 0.5 0.6 0.632 0.563 0.632 0.533 0.389 0.5 0.7 ...
##  $ X3P     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X3PA    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X3P.    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ FT      : int  4 6 5 7 11 5 13 4 0 7 ...
##  $ FTA     : int  7 11 8 12 16 11 18 11 0 13 ...
##  $ FT.     : num  0.571 0.545 0.625 0.583 0.688 0.455 0.722 0.364 NA 0.538 ...
##  $ ORB     : int  5 5 4 9 5 7 5 3 1 3 ...
##  $ DRB     : int  13 10 9 12 10 12 11 14 12 8 ...
##  $ TRB     : int  18 15 13 21 15 19 16 17 13 11 ...
##  $ AST     : int  2 1 1 1 1 1 3 2 2 1 ...
##  $ STL     : int  1 0 1 0 1 1 2 1 1 1 ...
##  $ BLK     : int  3 4 3 4 3 3 3 3 3 4 ...
##  $ TOV     : int  8 4 4 6 2 3 4 7 2 4 ...
##  $ PF      : int  6 5 4 4 4 5 3 4 3 4 ...
##  $ PTS     : int  12 22 35 31 29 29 29 18 12 21 ...
##  $ GmSc    : num  8.3 16 26 26.3 26.1 25.4 27.5 7.6 11.6 17.8 ...
##  $ Pls.Mns : int  NA NA NA NA NA NA NA NA NA NA ...
```

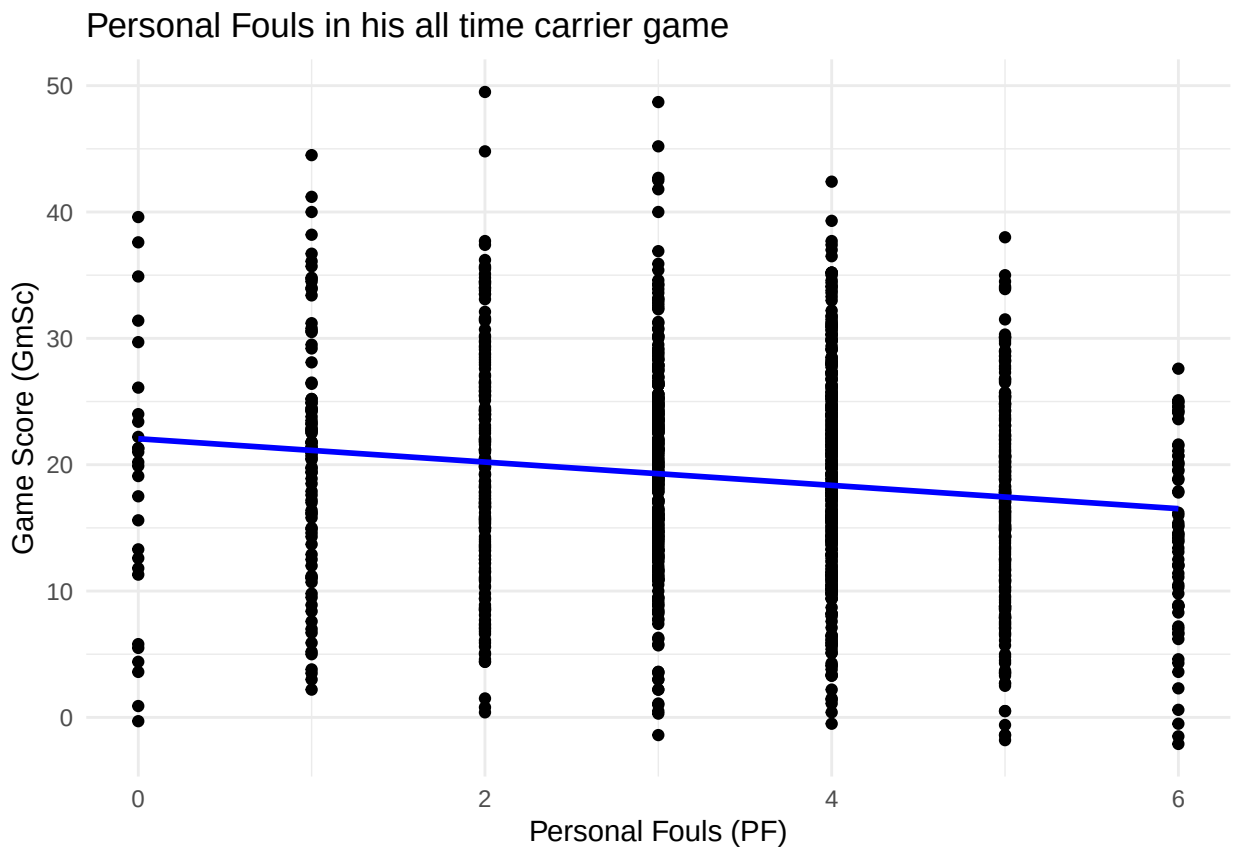
```
summary(shaq)
```

```
##      Season          SeasGm          CarrGm            Date
##  Min.   : 1.000   Min.   : 1.00   Min.   :   1.0   Min.   :33914
##  1st Qu.: 5.000   1st Qu.:16.00   1st Qu.: 302.5   1st Qu.:35382
##  Median : 9.000   Median :32.00   Median : 604.0   Median :36989
##  Mean   : 9.483   Mean   :33.64   Mean   : 604.0   Mean   :37095
##  3rd Qu.:14.000   3rd Qu.:49.00   3rd Qu.: 905.5   3rd Qu.:38740
##  Max.   :19.000   Max.   :81.00   Max.   :1207.0   Max.   :40636
##
##       Age             Tm                Home            Opp
##  Min.   :20.67   Length:1207        Min.   :0.0000   Length:1207
##  1st Qu.:24.69   Class :character   1st Qu.:0.0000   Class :character
##  Median :29.09   Mode  :character   Median :1.0000   Mode  :character
##  Mean   :29.38                      Mean   :0.5079
##  3rd Qu.:33.88                      3rd Qu.:1.0000
##  Max.   :39.08                      Max.   :1.0000
##
##       Win            teamdiff            GS            Minutes
##  Min.   :0.0000   Min.   :-42.000   Min.   :0.0000   Min.   : 2.00
##  1st Qu.:0.0000   1st Qu.: -4.000   1st Qu.:1.0000   1st Qu.:30.78
##  Median :1.0000   Median :  6.000   Median :1.0000   Median :36.00
```

```
##  Mean   :0.6785   Mean   :  4.963   Mean   :0.9917   Mean   :34.73
##  3rd Qu.:1.0000   3rd Qu.: 13.000   3rd Qu.:1.0000   3rd Qu.:40.00
##  Max.   :1.0000   Max.   : 48.000   Max.   :1.0000   Max.   :55.00
##
##        FG              FGA             FG.              X3P
##  Min.   : 0.000   Min.   : 0.00   Min.   :0.000   Min.   :0.0000000
##  1st Qu.: 7.000   1st Qu.:12.00   1st Qu.:0.500   1st Qu.:0.0000000
##  Median : 9.000   Median :16.00   Median :0.588   Median :0.0000000
##  Mean   : 9.387   Mean   :16.12   Mean   :0.589   Mean   :0.0008285
##  3rd Qu.:12.000   3rd Qu.:20.00   3rd Qu.:0.667   3rd Qu.:0.0000000
##  Max.   :24.000   Max.   :40.00   Max.   :1.000   Max.   :1.0000000
##                                   NA's   :1
##       X3PA             X3P.              FT              FTA
##  Min.   :0.00000   Min.   :0.0000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: 2.000   1st Qu.: 6.000
##  Median :0.00000   Median :0.0000   Median : 4.000   Median : 9.000
##  Mean   :0.01823   Mean   :0.0455   Mean   : 4.917   Mean   : 9.322
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.: 7.000   3rd Qu.:12.000
##  Max.   :1.00000   Max.   :1.0000   Max.   :19.000   Max.   :31.000
##                    NA's   :1185
##       FT.              ORB              DRB              TRB
##  Min.   :0.0000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.00
##  1st Qu.:0.4000   1st Qu.: 2.000   1st Qu.: 5.000   1st Qu.: 8.00
##  Median :0.5000   Median : 3.000   Median : 7.000   Median :11.00
##  Mean   :0.5232   Mean   : 3.487   Mean   : 7.365   Mean   :10.85
##  3rd Qu.:0.6670   3rd Qu.: 5.000   3rd Qu.: 9.000   3rd Qu.:14.00
##  Max.   :1.0000   Max.   :14.000   Max.   :25.000   Max.   :28.00
##  NA's   :25
##       AST              STL              BLK              TOV
##  Min.   : 0.000   Min.   :0.0000   Min.   : 0.000   Min.   :0.000
##  1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.: 1.000   1st Qu.:2.000
##  Median : 2.000   Median :0.0000   Median : 2.000   Median :3.000
##  Mean   : 2.507   Mean   :0.6123   Mean   : 2.263   Mean   :2.742
##  3rd Qu.: 4.000   3rd Qu.:1.0000   3rd Qu.: 3.000   3rd Qu.:4.000
##  Max.   :10.000   Max.   :5.0000   Max.   :15.000   Max.   :9.000
##
##        PF              PTS              GmSc           Pls.Mns
##  Min.   :0.000   Min.   : 0.00   Min.   :-2.10   Min.   :-27.000
##  1st Qu.:3.000   1st Qu.:18.00   1st Qu.:12.90   1st Qu.: -3.000
##  Median :4.000   Median :24.00   Median :18.90   Median :  5.000
##  Mean   :3.435   Mean   :23.69   Mean   :18.89   Mean   :  4.526
##  3rd Qu.:4.000   3rd Qu.:30.00   3rd Qu.:24.50   3rd Qu.: 13.000
##  Max.   :6.000   Max.   :61.00   Max.   :49.50   Max.   : 34.000
##                                                  NA's   :534
```

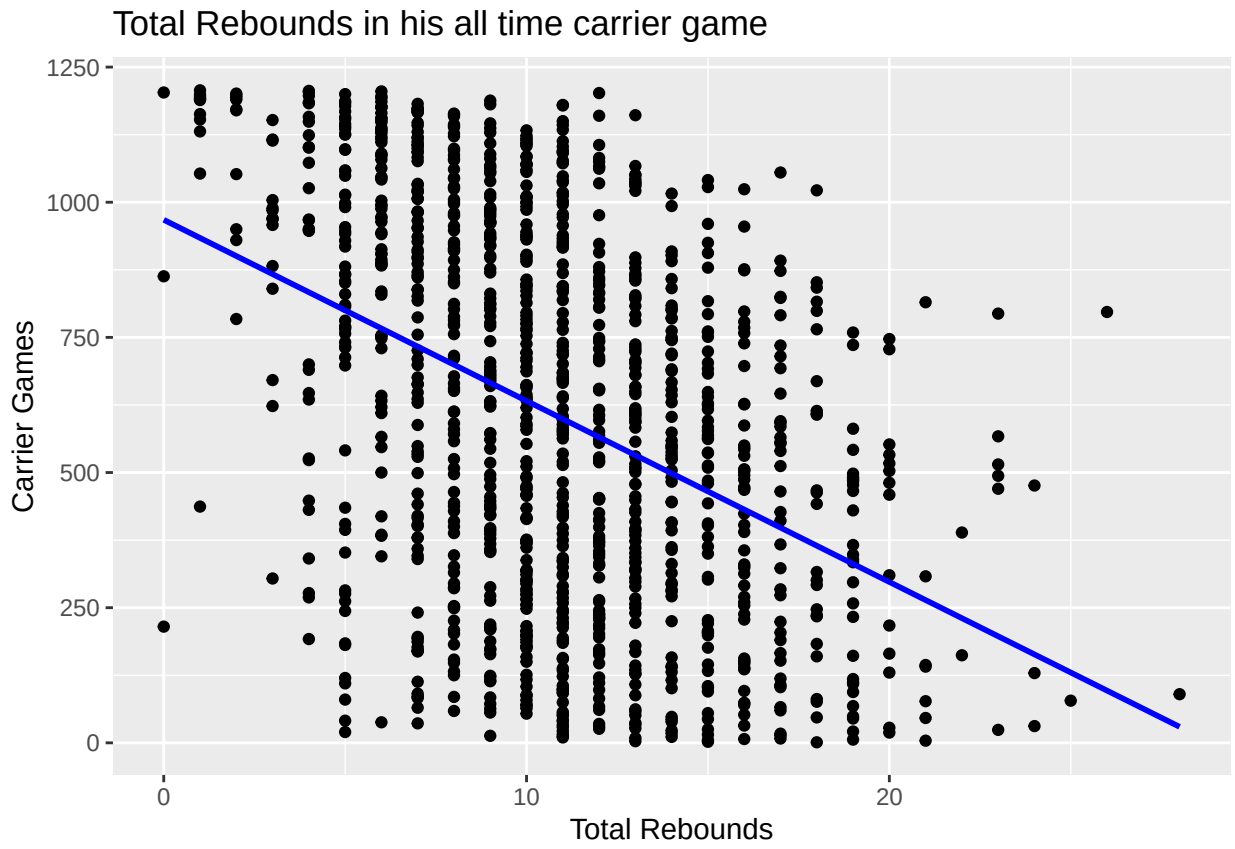
```
#Personal Fouls in his all time Game Score
Fouls <- ggplot(aes(x=PF,y=GmSc),data=shaq)+
  geom_point()+
  xlab("Personal Fouls (PF)")+
  ylab("Game Score (GmSc)")+
  ggtitle("Personal Fouls in his all time carrier game")+
  geom_smooth(method = "lm", se = FALSE, color = "blue")+
  theme_minimal()
Fouls
```

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



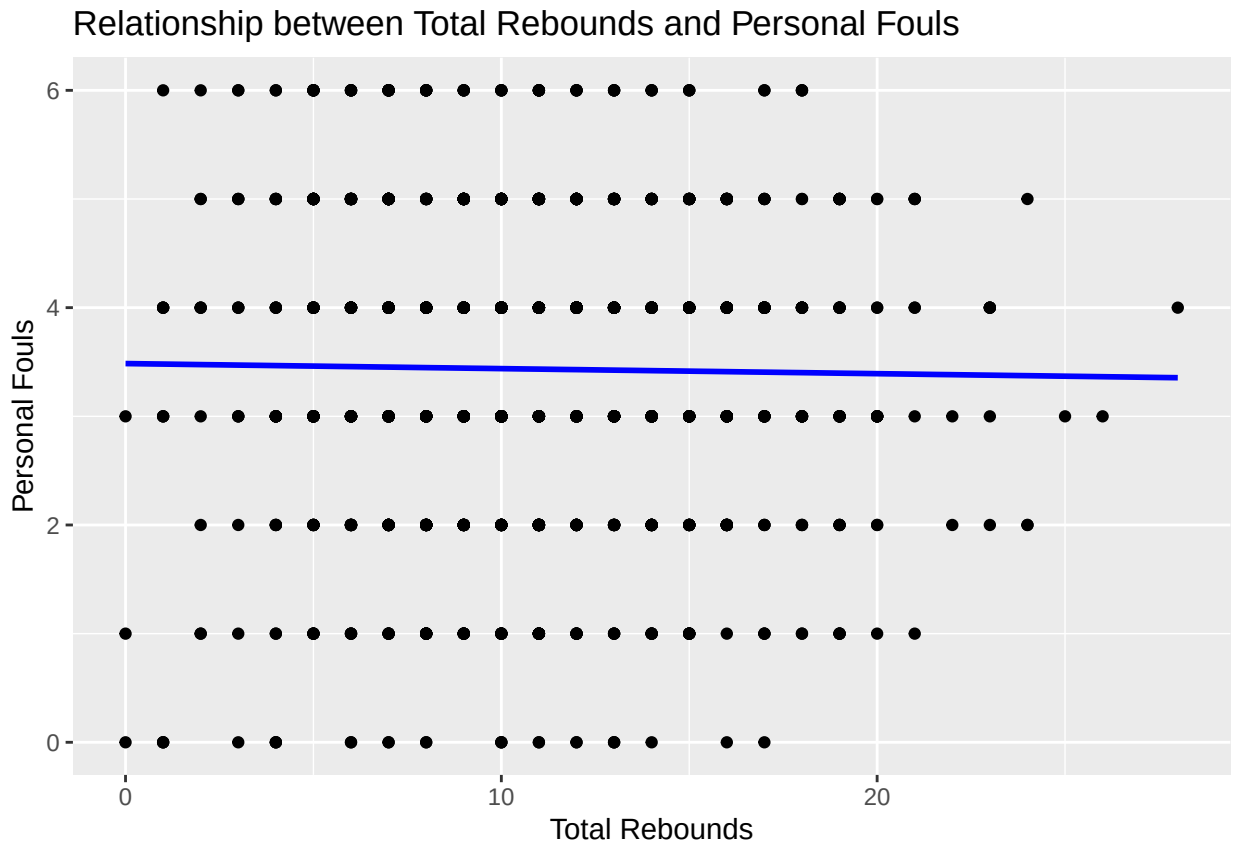

```
#Total Rebounds in his all time carrier game
Rebounds <- ggplot(aes(x=TRB,y=CarrGm),data=shaq)+
  geom_point()+
  xlab("Total Rebounds")+
  ylab("Carrier Games")+
  geom_smooth(method = "lm", se = FALSE, color = "blue")+
  ggtitle("Total Rebounds in his all time carrier game")
Rebounds
```

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



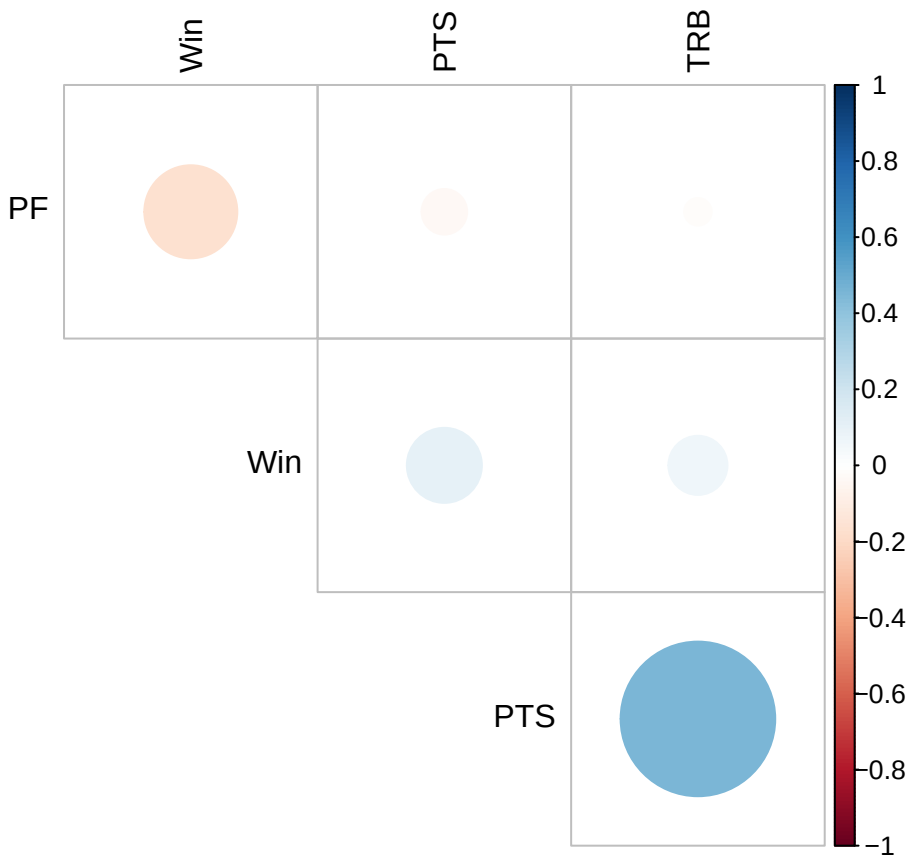

```
#Relation of Total rebounds and Personal fouls
Relation <- ggplot(aes(x=TRB,y=PF),data=shaq)+
  geom_point()+
  xlab("Total Rebounds")+
  ylab("Personal Fouls")+
  ggtitle("Relationship between Total Rebounds and Personal Fouls")+
  geom_smooth(method = "lm", se = FALSE, color = "blue")
Relation
```

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




```
#Correlation Plot
# Compute the correlation matrix
cor_matrix <- cor(shaq[, c("PF", "PTS", "TRB", "Win")])

# Plot the correlation matrix using corrplot
corrplot(cor_matrix, type = "upper", method = "circle",tl.col = 'black',
         order = "hclust", diag = FALSE,sig.level = 0.01, insig = "blank")
```




```
#Linear Regression  (including best-subset, stepwise, and AIC-based selection methods)

set.seed(42)
train_indices <- sample(1:nrow(shaq), 0.8 * nrow(shaq))
train_data <- shaq[train_indices, ]
test_data <- shaq[-train_indices, ]

# Linear Regression with best-subset selection
best_subset_model <- regsubsets(PTS ~ PF + TRB, data = train_data, method = "exhaustive")
best_subset_summary <- summary(best_subset_model)
best_subset_features <- rownames(best_subset_summary$outmat)[which.min(best_subset_summary$aic)]

cat("Best Subset Features:", best_subset_features, "\n")
```

```
## Best Subset Features:
```

```
cat("Best Subset Model AIC:", best_subset_summary$aic[which.min(best_subset_summary$aic)], "\n")
```

```
## Best Subset Model AIC:
```

```
# Linear Regression with stepwise selection
stepwise_model <- stepAIC(lm(PTS ~ 1, data = train_data), direction = "both", scope = list(lower = ~1, u
```

```
## Start:  AIC=4238.14
```

```
## PTS ~ 1
##
##        Df Sum of Sq   RSS    AIC
## + TRB   1   16579.0 61222 4008.9
## <none>              77801 4238.1
## + PF    1      36.7 77764 4239.7
##
## Step:  AIC=4008.88
## PTS ~ TRB
##
##        Df Sum of Sq   RSS    AIC
## <none>              61222 4008.9
## + PF    1      25.9 61196 4010.5
## - TRB   1   16579.0 77801 4238.1
```

```
stepwise_features <- names(stepwise_model$coefficients)

cat("Stepwise Features:", stepwise_features, "\n")
```

```
## Stepwise Features: (Intercept) TRB
```

```
cat("Stepwise Model AIC:", AIC(stepwise_model), "\n")
```

```
## Stepwise Model AIC: 6749.433
```

```
# Linear Regression without feature selection
full_model <- lm(PTS ~ PF + TRB, data = train_data)
cat("Full Model AIC:", AIC(full_model), "\n")
```

```
## Full Model AIC: 6751.025
```

```
# Visualize the linear regression models

# Extracting the best subset features
best_subset_features <- names(which(summary(best_subset_model)$which[which.min(summary(best_subset_mode

# Predictions on the training set for best-subset model
train_data$Predicted_BestSubset <- predict(lm(PTS ~ ., data = train_data[, c('PTS', best_subset_features

# Predictions on the training set for stepwise model
train_data$Predicted_Stepwise <- predict(stepwise_model, newdata = train_data)

# Predictions on the training set for full model
train_data$Predicted_Full <- predict(full_model, newdata = train_data)


ggplot(train_data, aes(x = TRB, y = PTS)) +
  geom_point() +
  geom_line(aes(y = Predicted_BestSubset), color = "blue", linetype = "dashed", size = 1) +
  geom_line(aes(y = Predicted_Stepwise), color = "red", linetype = "dashed", size = 1) +
  geom_line(aes(y = Predicted_Full), color = "green", linetype = "dashed", size = 1) +
```

```
  labs(title = "Linear Regression Models",
       x = "TRB",
       y = "PTS") +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

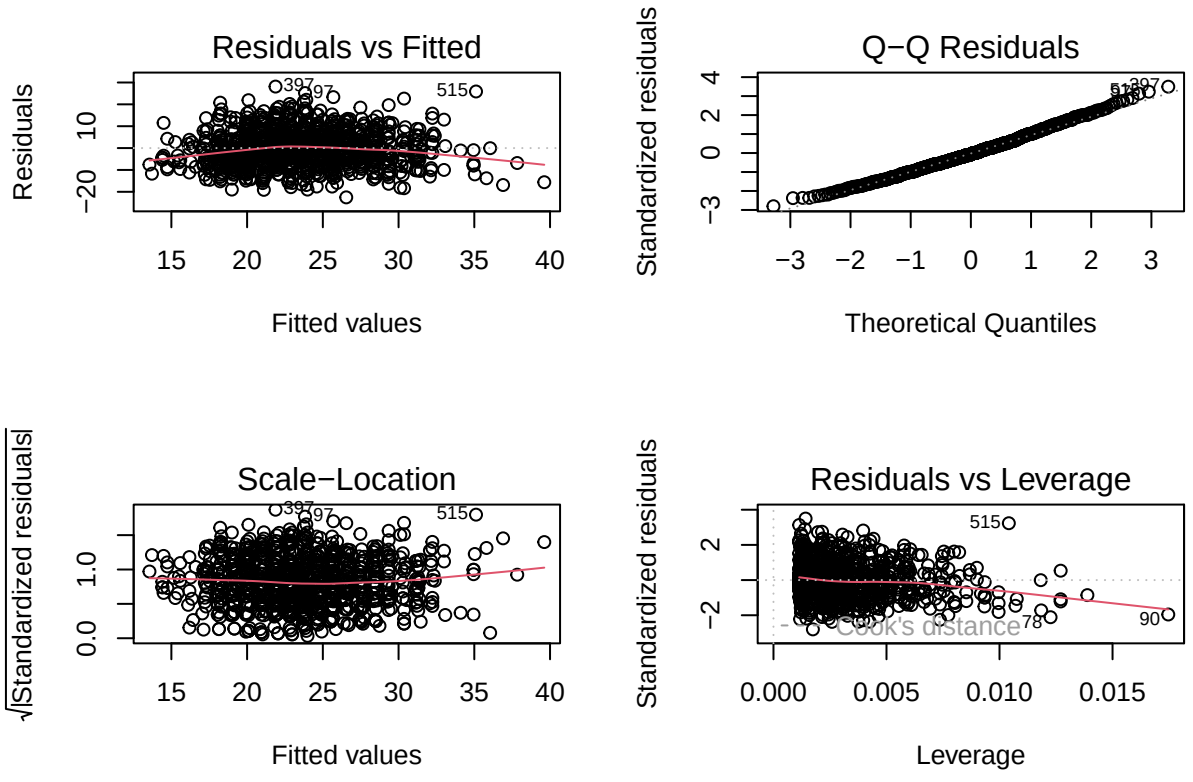

```
#Normal Linear regression plot
# Split the data into training and testing sets
set.seed(8495)  # for reproducibility
sample_indices <- sample(nrow(shaq), 0.8 * nrow(shaq))
train_data <- shaq[sample_indices, ]
test_data <- shaq[-sample_indices, ]


#creating a linear regression model
model <- lm(PTS~PF + TRB, data=train_data)
summary(model)
```

```
##
```

```
## Call:
## lm(formula = PTS ~ PF + TRB, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.5462  -5.4061  -0.4959   5.0569  28.1230
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.79998    0.94718  14.569   <2e-16 ***
## PF          -0.08193    0.18810  -0.436    0.663
## TRB          0.93386    0.06008  15.545   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.05 on 962 degrees of freedom
## Multiple R-squared:  0.2008, Adjusted R-squared:  0.1991
## F-statistic: 120.8 on 2 and 962 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))
plot(model)
```



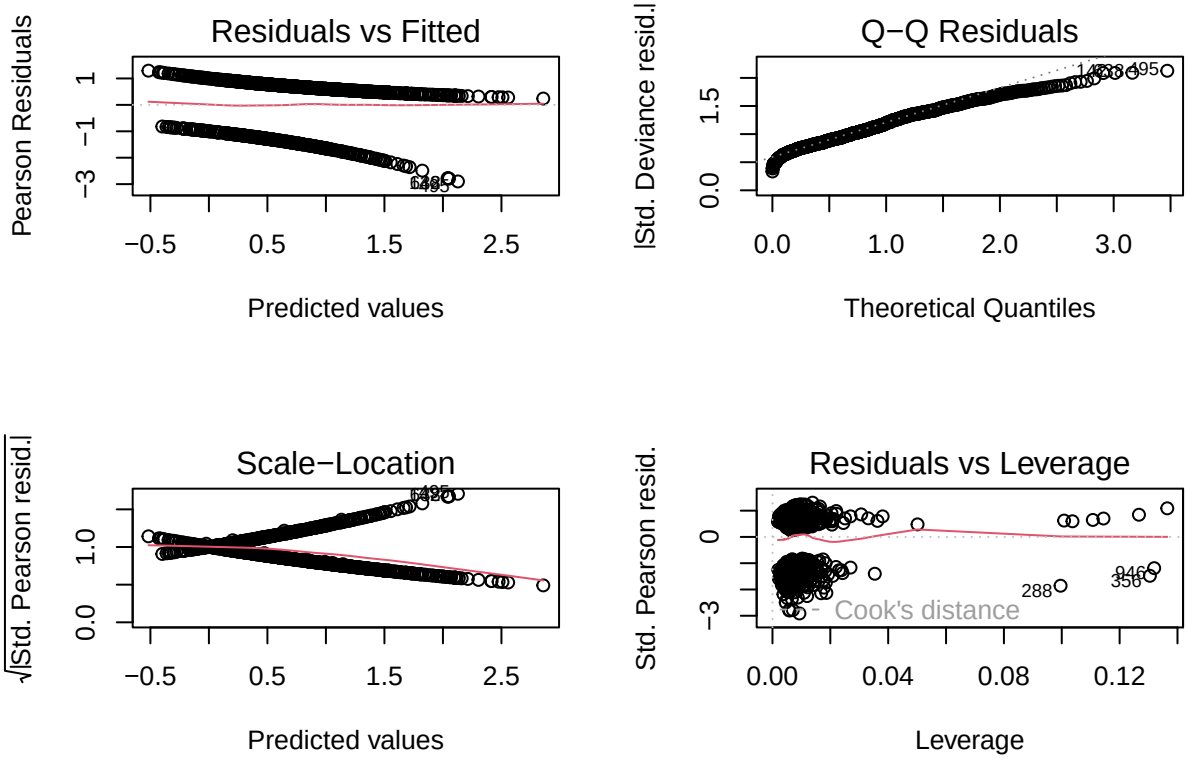

```r
# Build a logistic regression model
logistic_model <- glm(Win ~ TRB + PF + PTS + AST + BLK + STL + FT + GS ,
                      data = train_data,
```

```
                      family = "binomial")

# Summary of the model
summary(logistic_model)
```

```
##
## Call:
## glm(formula = Win ~ TRB + PF + PTS + AST + BLK + STL + FT + GS,
##     family = "binomial", data = train_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.932600   0.755557   1.234   0.2171
## TRB         -0.008710   0.019088  -0.456   0.6482
## PF          -0.268214   0.053451  -5.018 5.22e-07 ***
## PTS          0.010979   0.010878   1.009   0.3128
## AST          0.184265   0.043065   4.279 1.88e-05 ***
## BLK          0.071579   0.042934   1.667   0.0955 .
## STL          0.051520   0.085329   0.604   0.5460
## FT          -0.007493   0.028460  -0.263   0.7923
## GS          -0.027004   0.727472  -0.037   0.9704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1220.5  on 964  degrees of freedom
## Residual deviance: 1163.4  on 956  degrees of freedom
## AIC: 1181.4
##
## Number of Fisher Scoring iterations: 4
```

```
#Plotting the logistic model
par(mfrow = c(2, 2))
plot(logistic_model)
```




```
# Make predictions on the test set
predictions <- predict(logistic_model, newdata = test_data, type = "response")

# Convert probabilities to class labels (Win or Loss)
predicted_classes <- ifelse(predictions > 0.5, "Win", "Loss")

# Create a confusion matrix to evaluate the model
confusion_matrix <- table(test_data$Win, predicted_classes)
print(confusion_matrix)
```

```
##    predicted_classes
##     Loss Win
##   0   16  56
##   1   11 159
```

```
# Calculate accuracy
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
print(paste("Accuracy: ", accuracy))
```

```
## [1] "Accuracy:  0.723140495867769"
```