MTH-245 Final Project

Maria Balderas and Pratik Shrestha

Dr. Kevin Hutson - Spring 2023

Table of Contents

1 Abstract

Background: The National Basketball Association (NBA) is a professional basketball league with 29 teams in the United States and Canada, with a regular season running from October till April, consisting of 82 games in total. Having so many games in a season makes one wonder what variables predict a win and which ones do not. Therefore, the objective of this study was to create a model that best predicts a win in an NBA basketball game.

Methods: The study utilized multiple linear regression to model the relationship between Wins and various independent variables such as PTS, oppPTS, FG, FGA, X2P, X2PA, X3P, X3PA, FT, FTA, ORB, DRB, AST, STL, BLK, and TOV. The dataset was adjusted, and several models were created to identify the model with the most significant decrease in variance inflation factors (VIFs).

Bottom Line: After analyzing the data, it was found that the model where explanatory variables were centered and scaled, and RFG and RX2P were taken out had the most significant decrease in VIFs. This suggests that these variables had the least impact on the model and could be excluded. The final model identified the significant predictors of a win in an NBA game, which included PTS, oppPTS, FG, FGA, X2P, X2PA, X3P, X3PA, FT, FTA, ORB, DRB, AST, STL, BLK, and TOV. This model provides a useful tool for coaches and players to make strategic decisions that may increase their chances of winning. In conclusion, this study aimed to identify the factors that predict a win in an NBA game by creating a model through multiple linear regression analysis. The final model identified the significant predictors of a win in an NBA game and provides valuable insights that can be used to make informed decisions by coaches and players.un

2 Introduction

The success of Furman's basketball team this season sparked our interest in understanding what makes a basketball team successful. We wanted to identify the predictor variables that have the most influence on a team's wins. Initially, we thought college basketball was a good place to start looking for data, but then we found data from the National Basketball Associations (NBA) teams, a professional basketball league in the United States and Canada.

We selected the NBA dataset because it covers a wide range of years and includes multiple variables. Additionally, as a national professional organization, the dataset can be applied to real-world scenarios, helping coaches and players to improve in areas that can potentially help with more wins and playoff appearances.

Another reason why we chose this over college basketball data is because the NBA only has 29 teams, while Division I college basketball has over 300 basketball teams, and multiple conferences, adding more confounding variables that could potentially affect our analysis. In order to answer some questions that we have, we will create a multiple linear regression model to predict wins.

2.1 The Data

The data used in the project was found on the Massachusetts Institute of Technology website that contains every season from every NBA team from 1980 to 2011-2012, but it excludes teams who did not have 82 games. It comprises 835 observations and 17 variables(a combination of sixteen discrete quantitative and one categorical variable. Below there is a preview of the dataset.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2
--
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.2.1     v dplyr   1.1.2
## v tidyr   1.2.1     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts()
--
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(patchwork)
nba <- read.csv("~/Downloads/NBA_train.csv")
nba %>% head(10)
```

```
##    SeasonEnd                 Team Playoffs  W  PTS oppPTS   FG  FGA  X2P X2PA
## 1       1980         Atlanta Hawks        1 50 8573   8334 3261 7027 3248 6952
## 2       1980        Boston Celtics        1 61 9303   8664 3617 7387 3455 6965
## 3       1980         Chicago Bulls        0 30 8813   9035 3362 6943 3292 6668
## 4       1980   Cleveland Cavaliers        0 37 9360   9332 3811 8041 3775 7854
## 5       1980        Denver Nuggets        0 30 8878   9240 3462 7470 3379 7215
```

```
## 6       1980       Detroit Pistons        0 16 8933   9609 3643 7596 3586 7377
## 7       1980 Golden State Warriors        0 24 8493   8853 3527 7318 3500 7197
## 8       1980       Houston Rockets        1 41 9084   9070 3599 7496 3495 7117
## 9       1980        Indiana Pacers        0 37 9119   9176 3639 7689 3551 7375
## 10      1980     Kansas City Kings        1 47 8860   8603 3582 7489 3557 7375
##    X3P X3PA   FT  FTA  ORB  DRB  AST STL BLK  TOV
## 1   13   75 2038 2645 1369 2406 1913 782 539 1495
## 2  162  422 1907 2449 1227 2457 2198 809 308 1539
## 3   70  275 2019 2592 1115 2465 2152 704 392 1684
## 4   36  187 1702 2205 1307 2381 2108 764 342 1370
## 5   83  255 1871 2539 1311 2524 2079 746 404 1533
## 6   57  219 1590 2149 1226 2415 1950 783 562 1742
## 7   27  121 1412 1914 1155 2437 2028 779 339 1492
## 8  104  379 1782 2326 1394 2217 2149 782 373 1565
## 9   88  314 1753 2333 1398 2326 2148 900 530 1517
## 10  25  114 1671 2250 1187 2429 2123 863 356 1439
```

2.2 Data Descriptions

There is only one categorical variable in this dataset which is *Playoffs*. The following information indicates the category of this variable.

- **Playoffs:** The only categorical variable is whether a team made it into the playoffs or not. With 0 meaning that the team didn't make it and 1 meaning that they did make it into the playoffs.

The remaining variables are discrete quantitative.

- **Defensive Rebounds (DRB):** A *defensive rebound* is a rebound made by a player on defense. It is measured by the number of defensive rebounds a team gets in that season.
- **Blocks(BLK):** A *block* is where a defensive player deflects a shot attempt from an offensive player to prevent a score. This was measured by the total number of blocks a team successfully attempted during the season.
- **Points scored during regular season(PTS):** *Points scored during regular season* is measured by the number of points a team scores during a regular basketball season.
- **Opponent points scored during the season(oppPTS):** *Opponent points scored during the season* is the total number of points scored by the opposing teams during each season.
- **Successful field goals (FG):** *Successful field goals* are any shots scored other than a free throw, these shots scored can range between 2-3 points depending on the location where it is shot. This is measured by the number of successful field baskets made during the season.
- **Field Goals Attempted (FGA):** *Attempted field goals* are any attempted baskets other than free throws in the seasons. It is measured by the number of successful field goals and unsuccessful field goals during the season.

- **Successful Two Pointers (X2P):** *Successful two pointers* is a basket scored anywhere inside the three-point arc. All the successful attempts inside the three-point arc are worth two points and are added to the total points of the team within its season. The number of two pointers is what was added to the data set.
- **Two pointers attempted (X2PA):** *Two pointers attempted* is any attempt to score a basket from anywhere inside the three point arc. This includes successful and unsuccessful baskets. The total count of attempted two pointers is added to the total in the dataset.
- **Successful three pointers (X3PA):** *Successful three pointers* refers to a successful shot that is made beyond the three-point arc. All the successful attempts outside the three-point arc are worth three points and added to the total points of the team within its season. The number of three pointers is what was added to the data set.
- **Three pointers attempted (X3PA):** *Three pointers attempted* is any attempt to score a basket from anywhere outisde the thee point arc. This includes successful and unsuccessful baskets.The total count of attempted three pointers is added to the total in the dataset.
- **Successful free throws(FT):** A *successful free throw* A free throw is usually a scoring attempt since the defense is not allowed to interfere with the shot. Free throws are usually given after a foul and are worth one point. A successful free throw are free throws that successfully made it into the baskets. The number of successful free shows are added to the season total.
- **Attempted free throws(FTA):** *Attempted free throws* refers to the number of free throws attempted by a team during the season.
- **Offensive Rebound(ORB):** *Offensive rebound* is where a player from the offensive team retrieves the ball from an unsuccessful shot attempt by their own team. This is measured by the number of times a team member is able to retrieve a ball from missed attempts.
- **Assist(AST):** An *assist* is where a player passes the ball to another teammate in order to score. This is measured by the number of assists in each season.
- **Steals(STL):** A *steal* is where a defensive player takes possession of the ball from the opposing team. Steal is measured by the number of steals within the season.
- **Turnovers(TOV):** *Turnovers* happen whenever an offensive player loses ball possession not including taking a shot or causing a foul. Turnovers are measured by the amount of turnovers within the season.

2.3 Summary Statistics

The summary statistics of the response variable *Wins(W)* was evaluated.

```
summary(nba$W)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    11.0    31.0    42.0    41.0    50.5    72.0
```

The response variable, *Wins (W)* which is measured by the amount of wins the team has each season. There are 82 games in each NBA season, and after each game the team with the highest score receives the win which is added to the win total for that season. The summary statistics point out that the mean and median are close in value meaning that the data is fairly evenly distributed. Even though the mean and median are similar, there can still be variations.

The summary statistics of the rest of the variables is shown below.

```
summary(nba)
```

```
##    SeasonEnd        Team              Playoffs            W
##  Min.   :1980   Length:835         Min.   :0.0000   Min.   :11.0
##  1st Qu.:1989   Class :character   1st Qu.:0.0000   1st Qu.:31.0
##  Median :1996   Mode  :character   Median :1.0000   Median :42.0
##  Mean   :1996                      Mean   :0.5749   Mean   :41.0
##  3rd Qu.:2005                      3rd Qu.:1.0000   3rd Qu.:50.5
##  Max.   :2011                      Max.   :1.0000   Max.   :72.0
##       PTS            oppPTS            FG            FGA            X2P
##  Min.   : 6901   Min.   : 6909   Min.   :2565   Min.   :5972   Min.   :1981
##  1st Qu.: 7934   1st Qu.: 7934   1st Qu.:2974   1st Qu.:6564   1st Qu.:2510
##  Median : 8312   Median : 8365   Median :3150   Median :6831   Median :2718
##  Mean   : 8370   Mean   : 8370   Mean   :3200   Mean   :6873   Mean   :2881
##  3rd Qu.: 8784   3rd Qu.: 8768   3rd Qu.:3434   3rd Qu.:7157   3rd Qu.:3296
##  Max.   :10371   Max.   :10723   Max.   :3980   Max.   :8868   Max.   :3954
##       X2PA           X3P             X3PA             FT            FTA
##  Min.   :4153   Min.   : 10.0   Min.   :  75.0   Min.   :1189   Min.   :1475
##  1st Qu.:5269   1st Qu.:131.5   1st Qu.: 413.0   1st Qu.:1502   1st Qu.:2008
##  Median :5706   Median :329.0   Median : 942.0   Median :1628   Median :2176
##  Mean   :5956   Mean   :319.0   Mean   : 916.9   Mean   :1650   Mean   :2190
##  3rd Qu.:6754   3rd Qu.:481.5   3rd Qu.:1347.5   3rd Qu.:1781   3rd Qu.:2352
##  Max.   :7873   Max.   :841.0   Max.   :2284.0   Max.   :2388   Max.   :3051
##       ORB             DRB            AST            STL
##  Min.   : 639.0   Min.   :2044   Min.   :1423   Min.   : 455.0
##  1st Qu.: 953.5   1st Qu.:2346   1st Qu.:1735   1st Qu.: 599.0
##  Median :1055.0   Median :2433   Median :1899   Median : 658.0
##  Mean   :1061.6   Mean   :2427   Mean   :1912   Mean   : 668.4
##  3rd Qu.:1167.0   3rd Qu.:2516   3rd Qu.:2078   3rd Qu.: 729.0
##  Max.   :1520.0   Max.   :2753   Max.   :2575   Max.   :1053.0
##       BLK             TOV
##  Min.   :204.0   Min.   : 931
##  1st Qu.:359.0   1st Qu.:1192
##  Median :410.0   Median :1289
##  Mean   :419.8   Mean   :1303
##  3rd Qu.:469.5   3rd Qu.:1396
##  Max.   :716.0   Max.   :1873
```

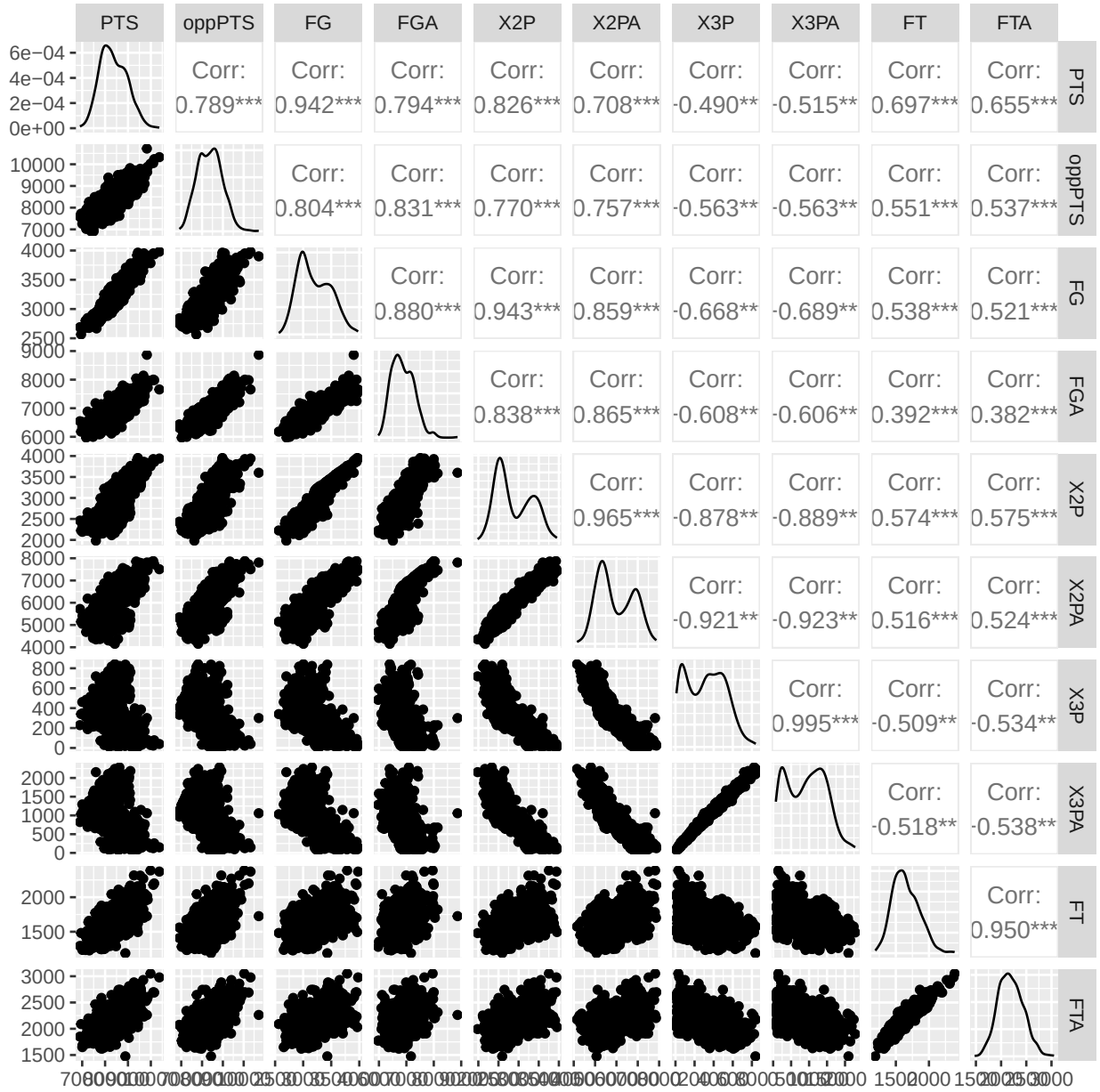
The explanatory variables all have different summary statistics from one another but there

was an issue when evaluating some variables.

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(nba, columns = 5:14)
```



When we evaluated the variables, we noticed that there's a strong correlation (>=0.70) among *FG,FGA, X2P,X2PA, X3P,X3PA, FT,FTA* independent variables. For *FG* and *FGA* the correlation is 0.879, *X2P* and *X2PA* the correlation is 0.965, *X3P* and X3PA the

correlation is 0.994, *FT* and *FTA* the correlation is 0.950. These variables are high correlated which could cause issues if we did not create the ratio.

We decided to ratio the variables of *FG,FGA, X2P,X2PA, X3P,X3PA, FT,FTA* into *FG/FGA(RFG), X2P/X2PA(RX2P), X3P/X3PA(RX3P) and FT/FTA(RFT)*. In the summary below it shows the ratio variables into new columns.

```
nba <- nba %>% mutate(RFG = FG/FGA, RX2P = X2P/X2PA, RX3P = X3P/X3PA, RFT = FT/FTA)
summary(nba)
```

```
##    SeasonEnd        Team              Playoffs            W
##  Min.   :1980   Length:835         Min.   :0.0000   Min.   :11.0
##  1st Qu.:1989   Class :character   1st Qu.:0.0000   1st Qu.:31.0
##  Median :1996   Mode  :character   Median :1.0000   Median :42.0
##  Mean   :1996                      Mean   :0.5749   Mean   :41.0
##  3rd Qu.:2005                      3rd Qu.:1.0000   3rd Qu.:50.5
##  Max.   :2011                      Max.   :1.0000   Max.   :72.0
##       PTS            oppPTS            FG            FGA            X2P
##  Min.   : 6901   Min.   : 6909   Min.   :2565   Min.   :5972   Min.   :1981
##  1st Qu.: 7934   1st Qu.: 7934   1st Qu.:2974   1st Qu.:6564   1st Qu.:2510
##  Median : 8312   Median : 8365   Median :3150   Median :6831   Median :2718
##  Mean   : 8370   Mean   : 8370   Mean   :3200   Mean   :6873   Mean   :2881
##  3rd Qu.: 8784   3rd Qu.: 8768   3rd Qu.:3434   3rd Qu.:7157   3rd Qu.:3296
##  Max.   :10371   Max.   :10723   Max.   :3980   Max.   :8868   Max.   :3954
##       X2PA           X3P             X3PA             FT            FTA
##  Min.   :4153   Min.   : 10.0   Min.   :  75.0   Min.   :1189   Min.   :1475
##  1st Qu.:5269   1st Qu.:131.5   1st Qu.: 413.0   1st Qu.:1502   1st Qu.:2008
##  Median :5706   Median :329.0   Median : 942.0   Median :1628   Median :2176
##  Mean   :5956   Mean   :319.0   Mean   : 916.9   Mean   :1650   Mean   :2190
##  3rd Qu.:6754   3rd Qu.:481.5   3rd Qu.:1347.5   3rd Qu.:1781   3rd Qu.:2352
##  Max.   :7873   Max.   :841.0   Max.   :2284.0   Max.   :2388   Max.   :3051
##       ORB             DRB            AST            STL
##  Min.   : 639.0   Min.   :2044   Min.   :1423   Min.   : 455.0
##  1st Qu.: 953.5   1st Qu.:2346   1st Qu.:1735   1st Qu.: 599.0
##  Median :1055.0   Median :2433   Median :1899   Median : 658.0
##  Mean   :1061.6   Mean   :2427   Mean   :1912   Mean   : 668.4
##  3rd Qu.:1167.0   3rd Qu.:2516   3rd Qu.:2078   3rd Qu.: 729.0
##  Max.   :1520.0   Max.   :2753   Max.   :2575   Max.   :1053.0
##       BLK             TOV            RFG             RX2P
##  Min.   :204.0   Min.   : 931   Min.   :0.4093   Min.   :0.4275
##  1st Qu.:359.0   1st Qu.:1192   1st Qu.:0.4490   1st Qu.:0.4694
##  Median :410.0   Median :1289   Median :0.4640   Median :0.4826
##  Mean   :419.8   Mean   :1303   Mean   :0.4651   Mean   :0.4831
##  3rd Qu.:469.5   3rd Qu.:1396   3rd Qu.:0.4803   3rd Qu.:0.4962
##  Max.   :716.0   Max.   :1873   Max.   :0.5448   Max.   :0.5550
##       RX3P             RFT
##  Min.   :0.1042   Min.   :0.6622
##  1st Qu.:0.3066   1st Qu.:0.7368
```

```
##  Median :0.3408   Median :0.7545
##  Mean   :0.3256   Mean   :0.7536
##  3rd Qu.:0.3606   3rd Qu.:0.7714
##  Max.   :0.4276   Max.   :0.8319
```

3 Exploratory Analysis

3.1 Visualizing the Variables

In this section, data visualizations were made in order to show the distributions of each variable including response and explanatory variables.

3.1.1 Histograms of Response Variable

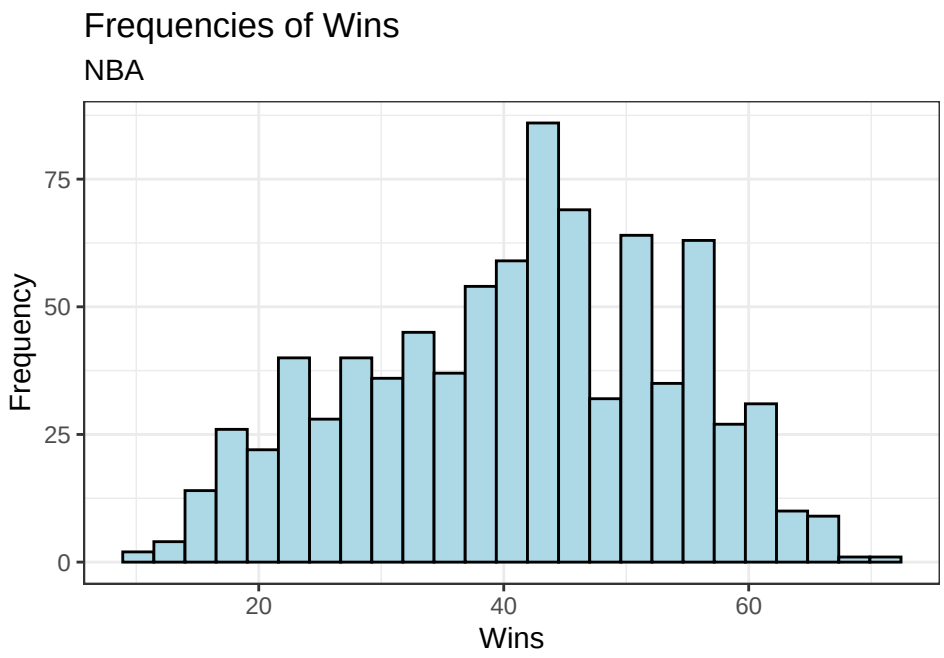


Figure 1: A histogram of frequencies of Wins

From Figure 1, one can see that the response variable's distribution is normally distributed

3.1.2 Histograms of the Predictors

After examining the distribution of the response variable, we will now examine the distribution of the explanatory variables which are all discrete quantitative

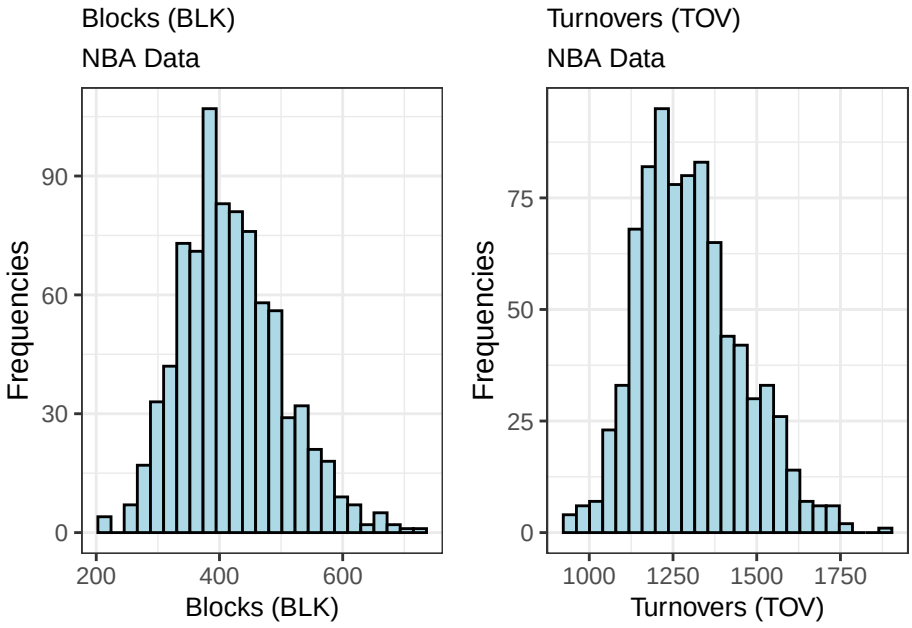


Figure 2: A histogram of frequencies of BLK and TOV

Figure 2: Shows that the histograms of *Blocks(BLK)* and *Turnovers(TOV)* are both normally distributed.

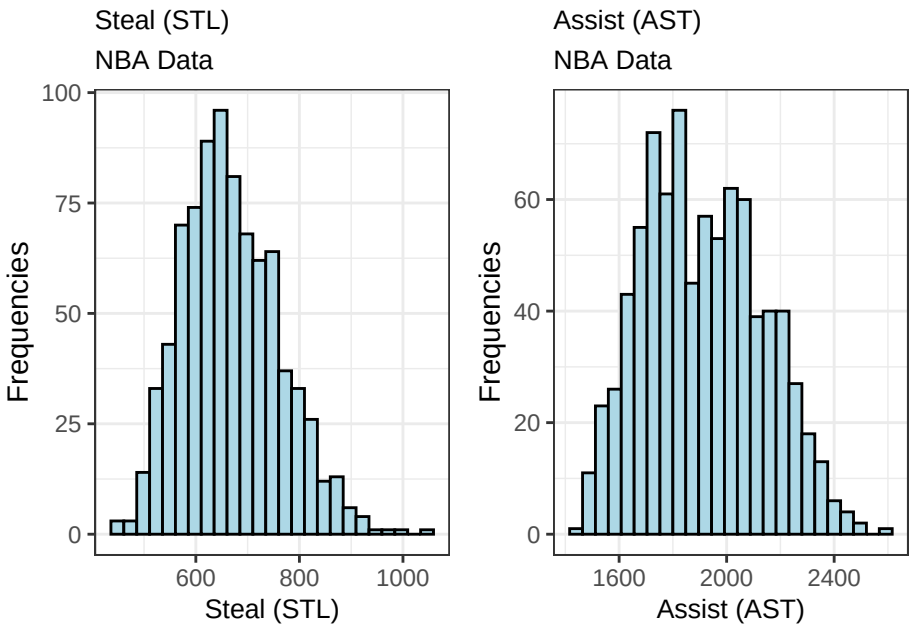


Figure 3: A histogram of frequencies of STL and AST

Figure 3: Shows that *Steal(STL)* and *Assist(AST)* are normally distributed.

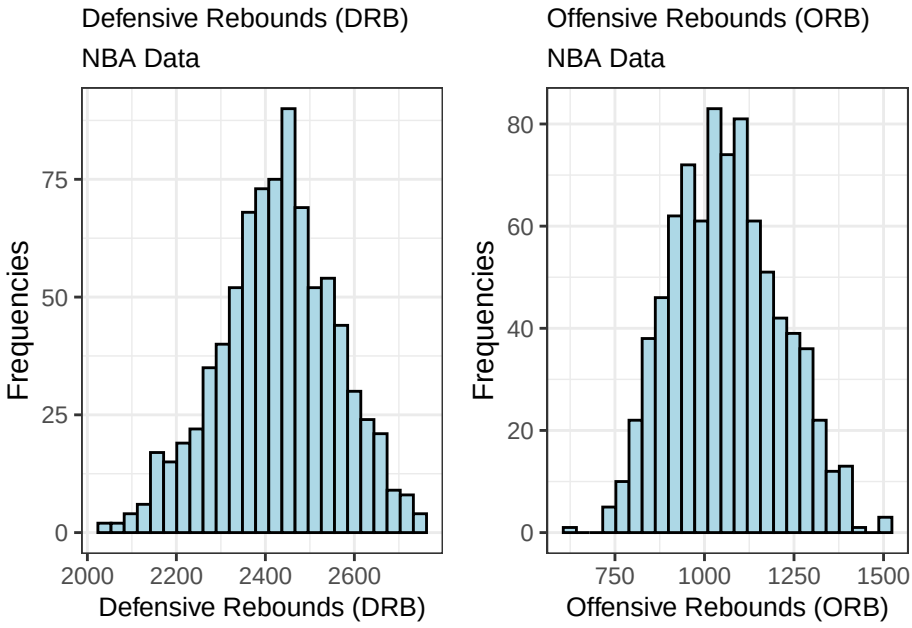


Figure 4: A histogram of frequencies of DRB and ORB

Figure 4: Shows that *Defensive Rebounds(DRB)* and *Offensive Rebounds(ORB)* are normally distributed.

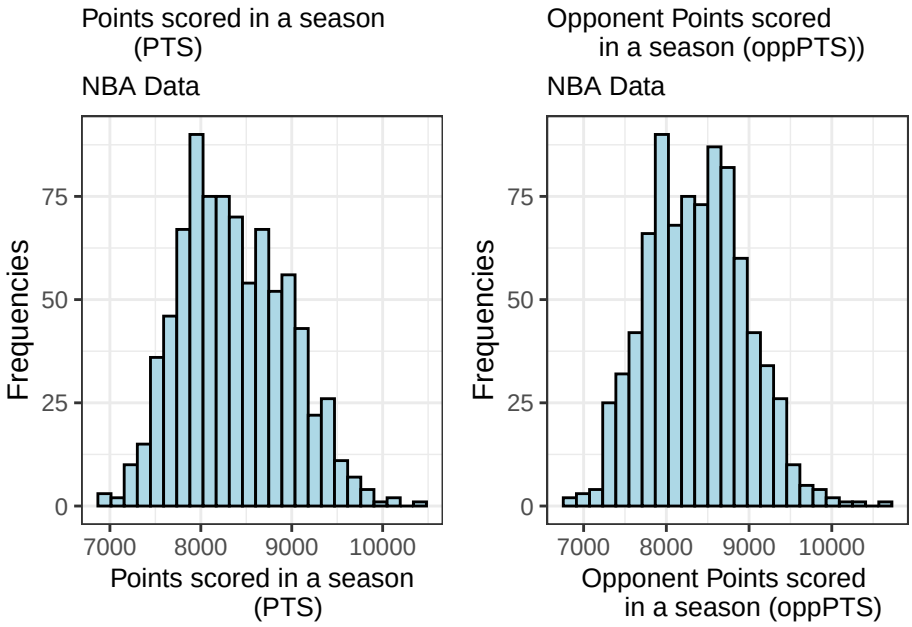


Figure 5: A histogram of frequencies of PTS and oppPTS

Figure 5: Shows that *Points scored in a season (PTS)* and *Opponent Points scored in a season(oppPTS)* are normally distributed.

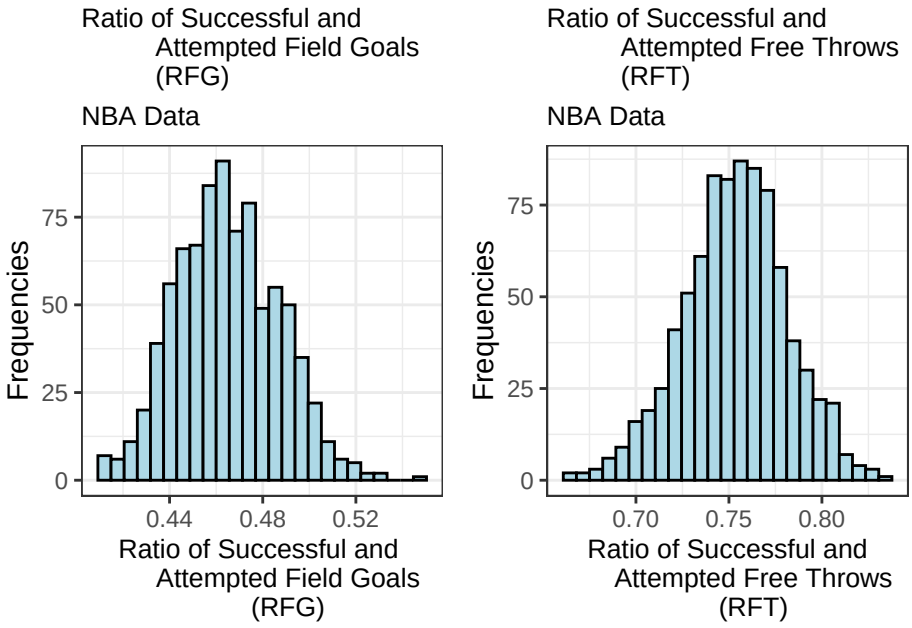


Figure 6: A histogram of frequencies of RFG and RFT

Figure 6: Shows that *Ratio of Successful and Attempted Field Goals(RFG)* and *Ratio of Succcessful and Attempted Free Throws* are normally distributed.

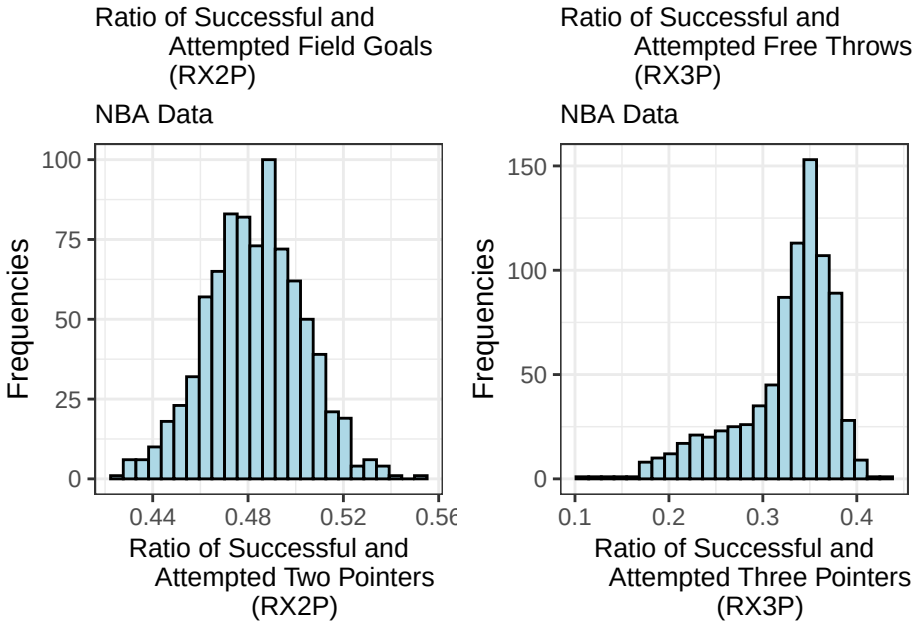


Figure 7: A histogram of frequencies of RX2P and RX3P

Figure 7: Shows that *Ratio of Successful and Attempted Field Goals(RX2p)* and *Ratio of Successful and Attempted Free Throws (RX3P)* are normally distributed.

3.1.3 Scatter Plot Matrix

Next, we created a scatter plot matrix of all the variable in the dataset.

```
library(GGally)
ggpairs(nba, columns = c(5, 6:24))
```

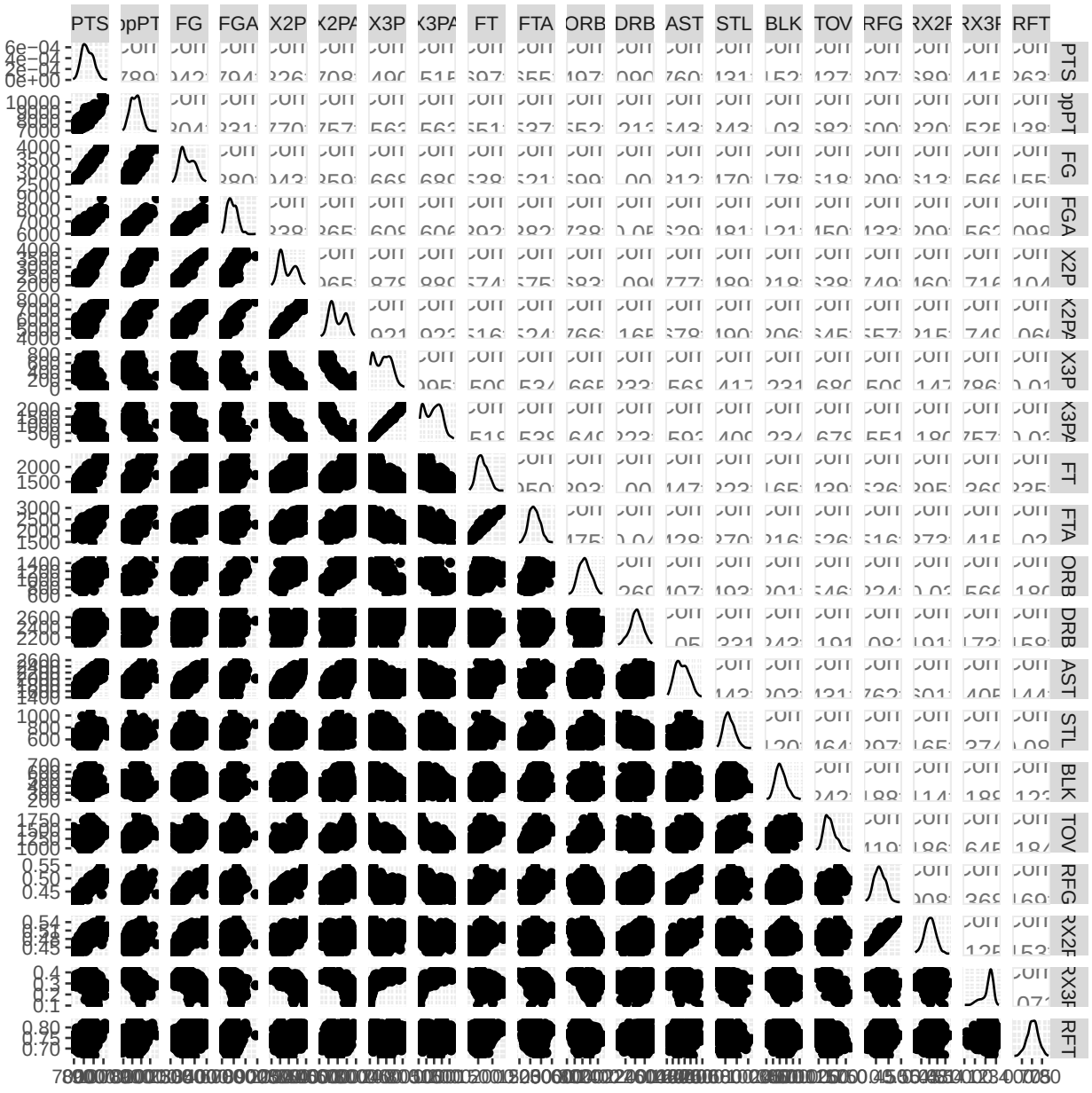


Figure 8: A scatter plot matrix of all 12 variables (ratio)

This scatter plot matrix allows one to see all the variables all at once alongside the correlation between variables.

4 Methods

4.1 Full First Order Model

This section involves generating a complete first-order linear model as a reference point. The model assumptions are evaluated, and a summary of the model is presented, followed by conducting a type 2 ANOVA test and constructing residual plots.

```
nba.model1 <- lm(W ~ PTS + oppPTS + FG +
                   FGA + X2P + X2PA + X3P +
                   X3PA + FT + FTA + ORB + DRB +
                   AST + STL + BLK + TOV,
                 data = nba)
summary(nba.model1)
```

```
##
## Call:
## lm(formula = W ~ PTS + oppPTS + FG + FGA + X2P + X2PA + X3P +
##     X3PA + FT + FTA + ORB + DRB + AST + STL + BLK + TOV, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6442 -2.0533 -0.1379  1.9371 10.8969
##
## Coefficients: (3 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.4570895  3.5960379  10.694  < 2e-16 ***
## PTS          0.0320387  0.0019993  16.025  < 2e-16 ***
## oppPTS      -0.0311545  0.0006777 -45.974  < 2e-16 ***
## FG          -0.0073013  0.0085327  -0.856  0.39242
## FGA          0.0007056  0.0026055   0.271  0.78659
## X2P          0.0081818  0.0063944   1.280  0.20107
## X2PA        -0.0030590  0.0023183  -1.319  0.18738
## X3P                 NA         NA      NA       NA
## X3PA                NA         NA      NA       NA
## FT                  NA         NA      NA       NA
## FTA         -0.0010690  0.0016989  -0.629  0.52937
## ORB          0.0019638  0.0017382   1.130  0.25890
## DRB          0.0026881  0.0014251   1.886  0.05961 .
## AST          0.0012175  0.0008993   1.354  0.17617
## STL          0.0012109  0.0019445   0.623  0.53363
## BLK          0.0038771  0.0014509   2.672  0.00768 **
## TOV         -0.0020945  0.0013992  -1.497  0.13481
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.045 on 821 degrees of freedom
## Multiple R-squared:  0.9438,Adjusted R-squared:  0.9429
```

```
## F-statistic:  1060 on 13 and 821 DF,  p-value: < 2.2e-16
```

```
nba.model1 <- lm(W ~ PTS + oppPTS + FG +
                   FGA + X2P + X2PA + FTA +
                   ORB + DRB +
                   AST + STL + BLK + TOV,
                 data = nba)
summary(nba.model1)
```

```
##
## Call:
## lm(formula = W ~ PTS + oppPTS + FG + FGA + X2P + X2PA + FTA +
##     ORB + DRB + AST + STL + BLK + TOV, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6442 -2.0533 -0.1379  1.9371 10.8969
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.4570895  3.5960379  10.694  < 2e-16 ***
## PTS          0.0320387  0.0019993  16.025  < 2e-16 ***
## oppPTS      -0.0311545  0.0006777 -45.974  < 2e-16 ***
## FG          -0.0073013  0.0085327  -0.856  0.39242
## FGA          0.0007056  0.0026055   0.271  0.78659
## X2P          0.0081818  0.0063944   1.280  0.20107
## X2PA        -0.0030590  0.0023183  -1.319  0.18738
## FTA         -0.0010690  0.0016989  -0.629  0.52937
## ORB          0.0019638  0.0017382   1.130  0.25890
## DRB          0.0026881  0.0014251   1.886  0.05961 .
## AST          0.0012175  0.0008993   1.354  0.17617
## STL          0.0012109  0.0019445   0.623  0.53363
## BLK          0.0038771  0.0014509   2.672  0.00768 **
## TOV         -0.0020945  0.0013992  -1.497  0.13481
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.045 on 821 degrees of freedom
## Multiple R-squared:  0.9438,Adjusted R-squared:  0.9429
## F-statistic:  1060 on 13 and 821 DF,  p-value: < 2.2e-16
```

From the regression model above, the estimated (fitted) linear regression equation is

$$\widehat{Y}$$

$$=38.4570895 + 0.0320387(\text{PTS}) - 0.0311545(\text{oppPTS}) - 0.0073013(\text{FG}) + 0.0007056(\text{FGA}) + 0.0081818(\text{X2P}) - 0.0030590(\text{X2PA}) - 0.0010690\ (\text{FTA}) + 0.0019638(\text{ORB}) +$$

0.0026881(DRB) + 0.00122175(AST) + 0.0012109(STL) + 0.0038771(BLK) - 0.0020945(TOV)

This is a preliminary model and the assumptions for Multiple Linear Regression have not yet been assessed.

```
library("car")
```

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

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

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

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

```
Anova(nba.model1, type = 2)
```

```
## Anova Table (Type II tests)
##
## Response: W
##            Sum Sq  Df   F value    Pr(>F)
## PTS         2380.6   1  256.8039 < 2.2e-16 ***
## oppPTS     19593.5   1 2113.6374 < 2.2e-16 ***
## FG             6.8   1    0.7322  0.392419
## FGA            0.7   1    0.0734  0.786587
## X2P           15.2   1    1.6372  0.201072
## X2PA          16.1   1    1.7410  0.187375
## FTA            3.7   1    0.3959  0.529371
## ORB           11.8   1    1.2764  0.258902
## DRB           33.0   1    3.5579  0.059615 .
## AST           17.0   1    1.8328  0.176166
## STL            3.6   1    0.3878  0.533633
## BLK           66.2   1    7.1408  0.007684 **
## TOV           20.8   1    2.2406  0.134810
## Residuals   7610.7 821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.2 Assessing Assumptions

We are now going to assess the assumptions for a multiple linear regression model, the assumptions that we are going to measure are for the first model that was shown earlier. Assessing assumptions in a multiple linear regression is important because it helps one to interpret data. A data set either fully meets the assumptions,not fully satisfies assumptions or not meet assumptions. Depending on how the assumptions are answered it determines

how one interprets the data.

4.2.1 Linear Relationship

The first assumption is that there is a linear relationship between the explanatory variable and the response variable. In order to assess this assumption a scatter matrix plot was generated in order to check linear relationship.

```
library(GGally)
ggpairs(nba, columns = c(4:20))
```

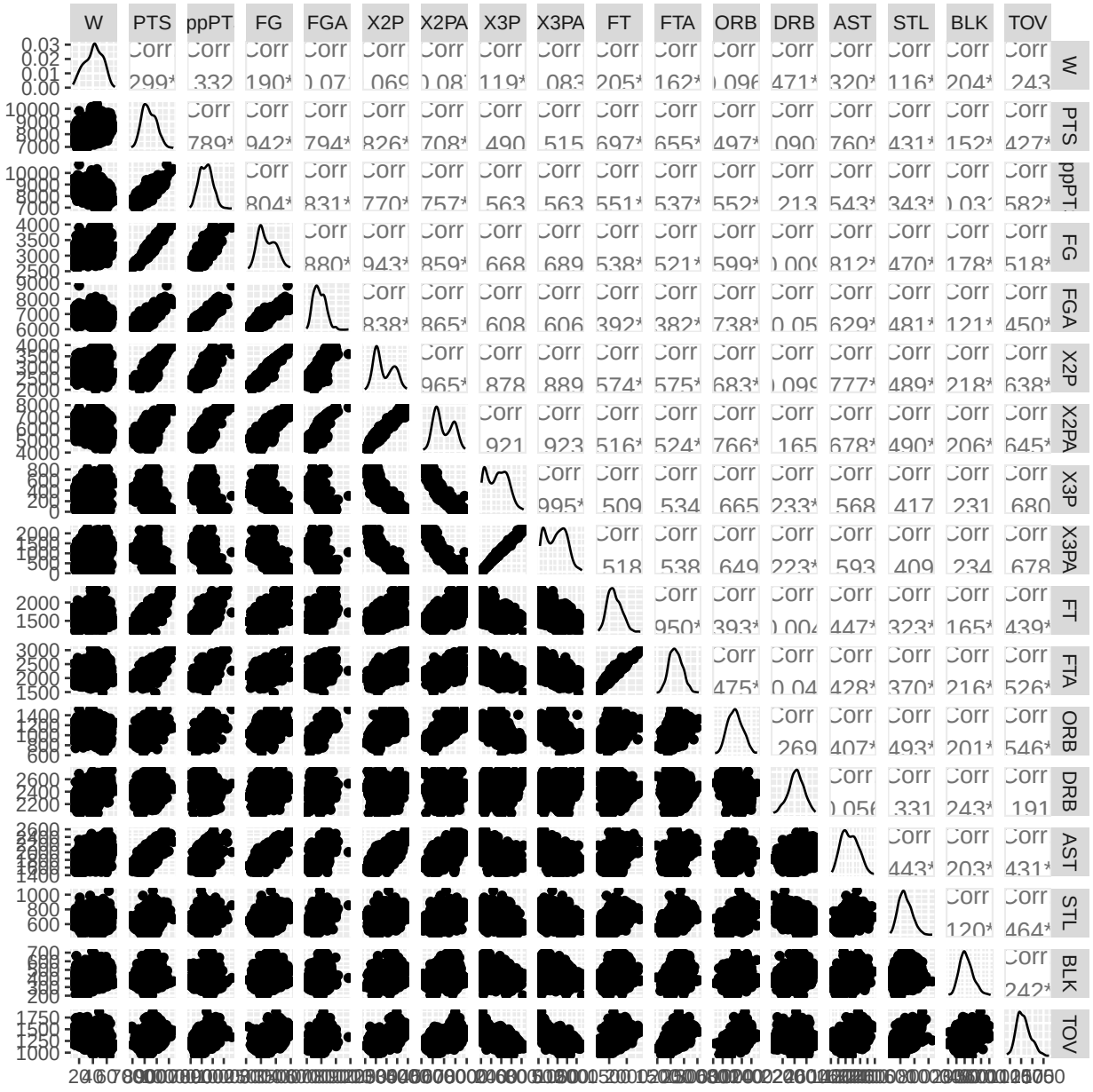


Figure 9: Scatter Matrix plot with W

The scatter plot matrix above shows that their is a linear relationship within variables.

4.2.2 Constant Error Variance

The second assumptions is that residuals have a constant variance σ^2. To test this assumption we will generate different plots to check variance.

```
library("qqplotr")
```

```
##
## Attaching package:  'qqplotr'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     stat_qq_line, StatQqLine
```

```
source("https://cipolli.com/students/code/plotResiduals.R")
plotResiduals(nba.model1)
```

```
## 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.
```

```
## Warning:  The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

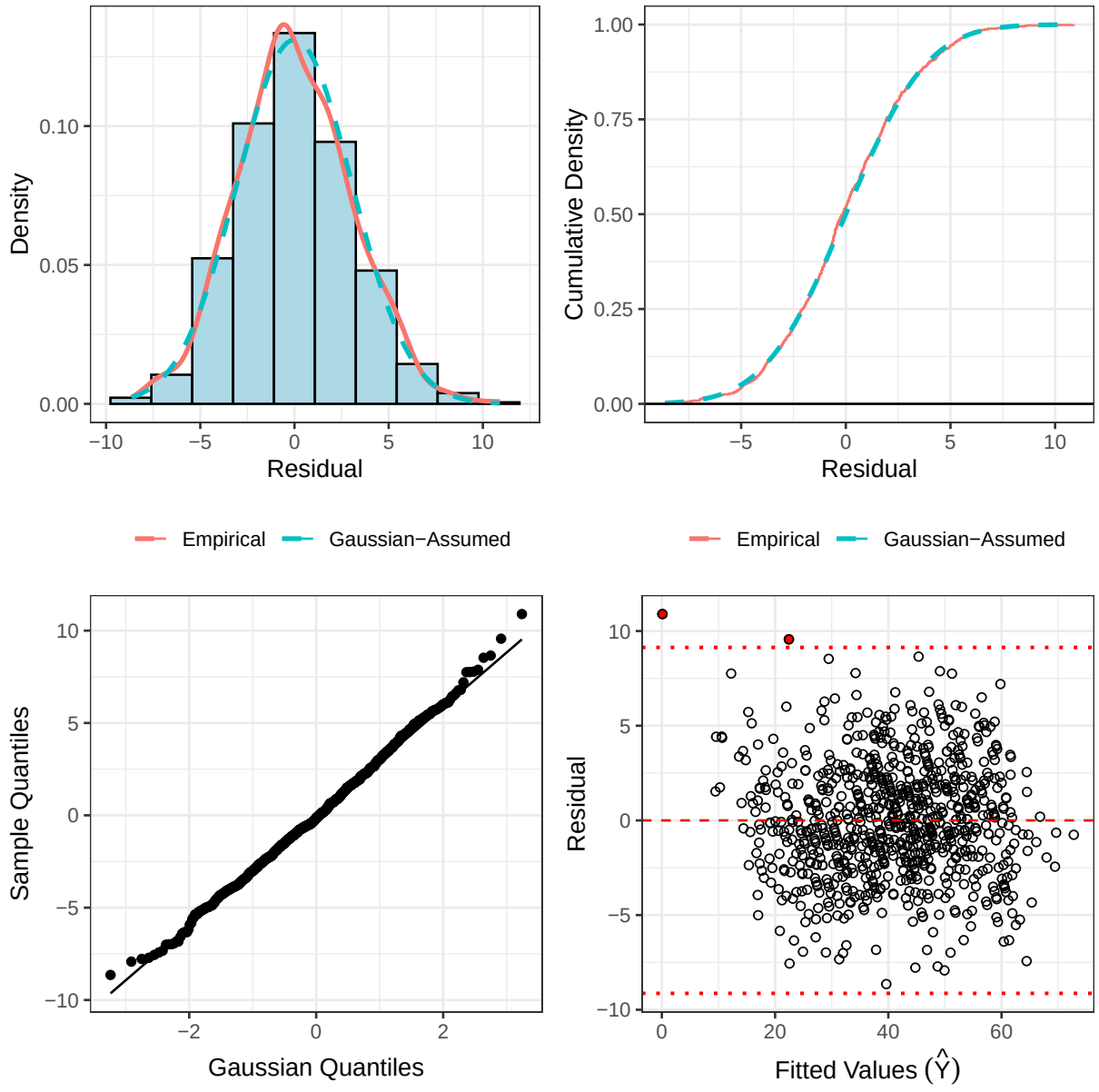


Figure 10: Residual diagnostic plots for the first order model.

Looking at Figure 10, we can determine that residuals (errors) have constant variance. Even though there is test that we could run to assess the constant variance assumptions, these test usually reject the variance assumptions when ran. The residuals are scattered with any pattern allowing there be constant variance.

4.2.3 Residuals are Normal

The third assumption is that the errors(residuals) are approximately Gaussian distributed with mean 0. We will again use Figure 10 but this time focus on Gaussian Quantilnes which has Q-Q plot, we can see it meets normality.

4.2.4 Residuals are Independent

The fourth assumption is that the errors (residuals) are independent. We addressed this assumption earlier because some variables were highly correlated with one another. To avoid this high correlation we decided to ratio the two variables because they are both dependent to each other. The more points scored the higher points attempted will be. For the rest of the variables, we can determine that they are independent because they do not have high correlation with other explanatory variables.

4.2.5 Representative Sample

The fifth assumption is that the sample is representative which is assessed on how the data was collected. Even though the data collected on each team is not necessarily random, the game it self is random making the data representative.

4.2.6 Multicollinearity

The last assumption is that no multicollinearity exist between the independent variable. To assess this assumption, we will calculate Variance Inflation Factors (VIF's) for every explanatory variables.

```
vif(nba.model1)
```

```
##         PTS     oppPTS          FG         FGA         X2P        X2PA         FTA
## 121.407373  14.261876 540.215950  98.220142 732.049563 333.580292  15.522610
##         ORB         DRB         AST         STL         BLK         TOV
##   6.134618   3.119902   3.573578   2.967062   1.281995   4.175844
```

Rule of thumb of measuring multicollinearity is that higher the number, higher the multicollinearity and lower the number, lower the multicollinearity is. If a variable has a multicollinearity of one it means that there is no multicollinearity.

After

4.3 Transformation

We decided to center and scale all of the variables so they can all be on the same playing field. Since center and scale makes every variable to have the same unit i.e, standard deviation, it makes it easier to compare parameters.

```
nba.model2 <- lm(scale(W, center = T, scale = T) ~
                   scale(PTS, center = T, scale = T) +
                   scale(oppPTS, center = T, scale = T) +
                   scale(RFG, center = T, scale = T)+
                   scale(RX3P, center = T, scale = T) +
                   scale(RX2P, center = T, scale = T)+
                   scale(RFT, center = T, scale = T) +
                   scale(ORB, center = T, scale = T) +
                   scale(DRB , center = T, scale = T) +
                   scale(AST, center = T, scale = T) +
```

```
                  scale(STL, center = T, scale = T) +
                  scale(BLK, center = T, scale = T) +
                  scale(TOV, center = T, scale = T),
                data = nba)
summary(nba.model2)
```

```
##
## Call:
## lm(formula = scale(W, center = T, scale = T) ~ scale(PTS, center = T,
##     scale = T) + scale(oppPTS, center = T, scale = T) + scale(RFG,
##     center = T, scale = T) + scale(RX3P, center = T, scale = T) +
##     scale(RX2P, center = T, scale = T) + scale(RFT, center = T,
##     scale = T) + scale(ORB, center = T, scale = T) + scale(DRB,
##     center = T, scale = T) + scale(AST, center = T, scale = T) +
##     scale(STL, center = T, scale = T) + scale(BLK, center = T,
##     scale = T) + scale(TOV, center = T, scale = T), data = nba)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69401 -0.16576 -0.01368  0.15290  0.86615
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -1.445e-15  8.255e-03   0.000  1.00000
## scale(PTS, center = T, scale = T)    1.394e+00  4.038e-02  34.534  < 2e-16 ***
## scale(oppPTS, center = T, scale = T) -1.450e+00  2.483e-02 -58.393  < 2e-16 ***
## scale(RFG, center = T, scale = T)    -3.675e-02  3.548e-02  -1.036  0.30054
## scale(RX3P, center = T, scale = T)   -1.624e-02  1.220e-02  -1.332  0.18337
## scale(RX2P, center = T, scale = T)    6.707e-02  3.011e-02   2.228  0.02617 *
## scale(RFT, center = T, scale = T)     8.617e-03  1.054e-02   0.817  0.41394
## scale(ORB, center = T, scale = T)     1.710e-02  1.698e-02   1.007  0.31439
## scale(DRB, center = T, scale = T)     2.286e-02  1.311e-02   1.744  0.08153 .
## scale(AST, center = T, scale = T)     2.131e-02  1.501e-02   1.420  0.15598
## scale(STL, center = T, scale = T)     5.093e-03  1.306e-02   0.390  0.69659
## scale(BLK, center = T, scale = T)     2.536e-02  9.310e-03   2.724  0.00659 **
## scale(TOV, center = T, scale = T)    -2.285e-02  1.512e-02  -1.511  0.13107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2385 on 822 degrees of freedom
## Multiple R-squared:  0.9439,Adjusted R-squared:  0.9431
## F-statistic:  1153 on 12 and 822 DF,  p-value: < 2.2e-16
```

Now that the model has been centered and scaled, we will now compare R-Squared, Adjusted R-Squared, and RSE from the original model.

In the new model, there is a slight improvement with the R-Squared and RSE values. In the

transformation model we have a higher R-Squared and a lower RSE but a higher Adjusted R-Squred.

```
library("qqplotr")
source("https://cipolli.com/students/code/plotResiduals.R")
plotResiduals(nba.model2)
```

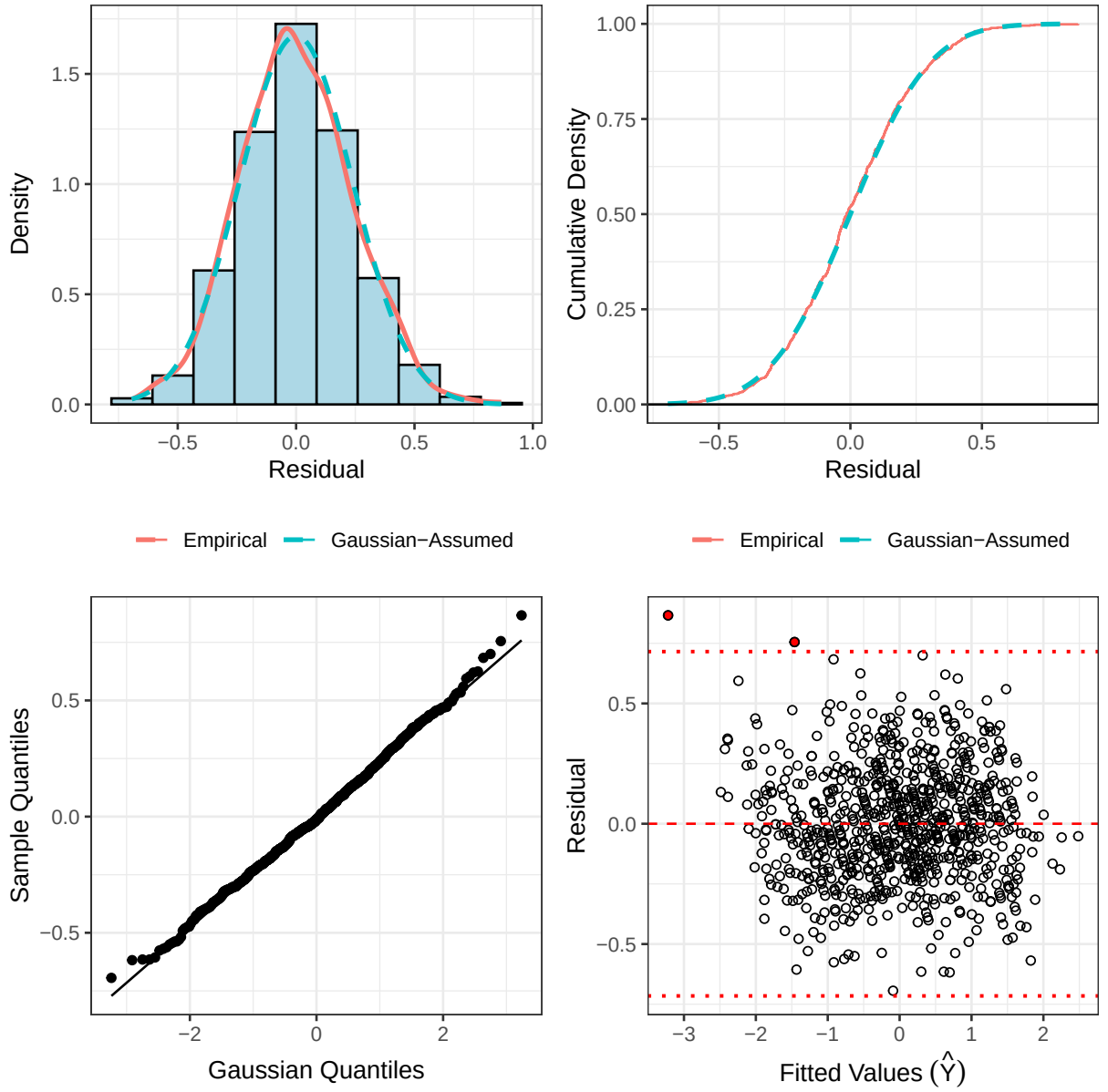


Figure 11: Residual diagnostic plots for scaled and centered model.

In Figure 11, the residual plots are very similar to one another meaning that the data in the first model is pretty good since not much changed.

```
library(GGally)
ggpairs(nba, columns = c(5, 6, 15:24))
```

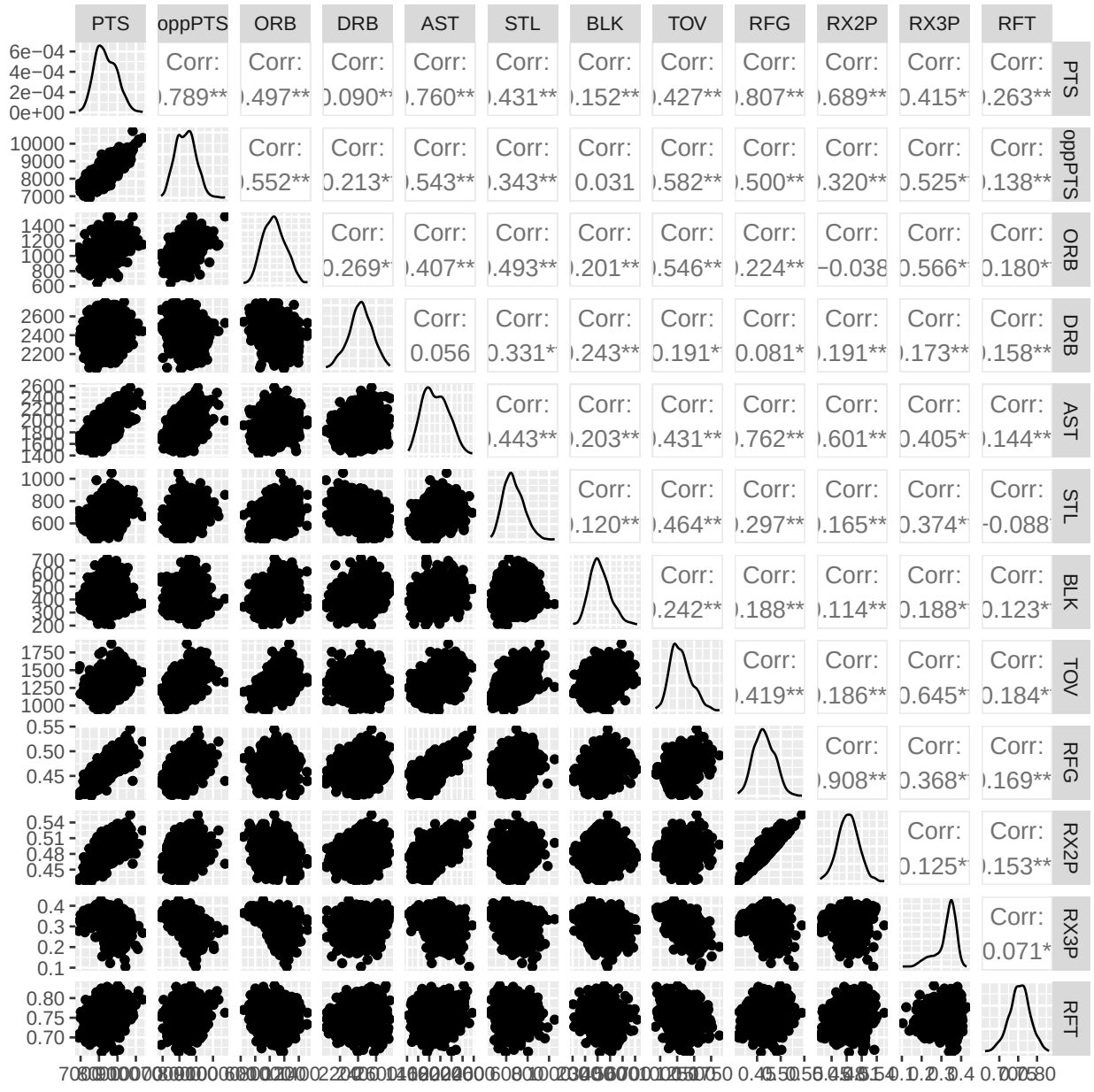


Figure 12: Scatter plot Matrix of scaled and centered data

Figure 12: Shows the linear relationship with other centered and scaled independent variables alongside with the ratio variables.

We also added a VIF figure to compare the difference between the original dataset and the one with centered and scaled data with ratios.

```
vif(nba.model2)
```

```
##    scale(PTS, center = T, scale = T) scale(oppPTS, center = T, scale = T)
##                             23.896054                                9.038733
##    scale(RFG, center = T, scale = T)   scale(RX3P, center = T, scale = T)
##                             18.449589                                2.180346
##   scale(RX2P, center = T, scale = T)    scale(RFT, center = T, scale = T)
##                             13.288187                                1.629120
##    scale(ORB, center = T, scale = T)    scale(DRB, center = T, scale = T)
##                              4.228370                                2.518945
##    scale(AST, center = T, scale = T)    scale(STL, center = T, scale = T)
##                              3.302017                                2.499331
##    scale(BLK, center = T, scale = T)    scale(TOV, center = T, scale = T)
##                              1.270568                                3.350863
```

We noticed that the multicollinearity for all of the variable drastically decreased with the highest collinearity being at 23 now instead of 750.

4.4 Removing Faulty Data

We decided to refit the regression model and take out the variable of *RFG* and *RX2P*

```
nba.model3 <- lm(scale(W, center = T, scale = T) ~
                   scale(PTS, center = T, scale = T) +
                   scale(oppPTS, center = T, scale = T) +
                   scale(RX3P, center = T, scale = T) +
                   scale(RFT, center = T, scale = T) +
                   scale(ORB, center = T, scale = T) +
                   scale(DRB , center = T, scale = T) +
                   scale(AST, center = T, scale = T) +
                   scale(STL, center = T, scale = T) +
                   scale(BLK, center = T, scale = T) +
                   scale(TOV, center = T, scale = T),
                 data = nba)
summary(nba.model3)
```

```
##
## Call:
## lm(formula = scale(W, center = T, scale = T) ~ scale(PTS, center = T,
##     scale = T) + scale(oppPTS, center = T, scale = T) + scale(RX3P,
##     center = T, scale = T) + scale(RFT, center = T, scale = T) +
##     scale(ORB, center = T, scale = T) + scale(DRB, center = T,
##     scale = T) + scale(AST, center = T, scale = T) + scale(STL,
##     center = T, scale = T) + scale(BLK, center = T, scale = T) +
##     scale(TOV, center = T, scale = T), data = nba)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70281 -0.16243 -0.00787  0.15317  0.85711
##
```

```
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          1.312e-16  8.273e-03   0.000  1.00000
## scale(PTS, center = T, scale = T)     1.450e+00  2.175e-02  66.682  < 2e-16 ***
## scale(oppPTS, center = T, scale = T) -1.474e+00  1.851e-02 -79.660  < 2e-16 ***
## scale(RX3P, center = T, scale = T)   -1.444e-02  1.170e-02  -1.234  0.21761
## scale(RFT, center = T, scale = T)    -1.994e-03  9.565e-03  -0.208  0.83492
## scale(ORB, center = T, scale = T)    -8.447e-03  1.215e-02  -0.695  0.48703
## scale(DRB, center = T, scale = T)     1.598e-02  1.115e-02   1.433  0.15230
## scale(AST, center = T, scale = T)     1.872e-02  1.357e-02   1.380  0.16797
## scale(STL, center = T, scale = T)    -6.853e-04  1.151e-02  -0.060  0.95253
## scale(BLK, center = T, scale = T)     2.465e-02  9.288e-03   2.654  0.00811 **
## scale(TOV, center = T, scale = T)    -1.963e-02  1.313e-02  -1.496  0.13512
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2391 on 824 degrees of freedom
## Multiple R-squared:  0.9435,Adjusted R-squared:  0.9428
## F-statistic:  1377 on 10 and 824 DF,  p-value: < 2.2e-16
```

```
Anova(nba.model3, type = 2)
```

```
## Anova Table (Type II tests)
##
## Response: scale(W, center = T, scale = T)
##                                     Sum Sq  Df   F value    Pr(>F)
## scale(PTS, center = T, scale = T)    254.12   1 4446.4880 < 2.2e-16 ***
## scale(oppPTS, center = T, scale = T) 362.67   1 6345.6774 < 2.2e-16 ***
## scale(RX3P, center = T, scale = T)     0.09   1    1.5224  0.217614
## scale(RFT, center = T, scale = T)      0.00   1    0.0435  0.834921
## scale(ORB, center = T, scale = T)      0.03   1    0.4835  0.487027
## scale(DRB, center = T, scale = T)      0.12   1    2.0528  0.152303
## scale(AST, center = T, scale = T)      0.11   1    1.9043  0.167973
## scale(STL, center = T, scale = T)      0.00   1    0.0035  0.952530
## scale(BLK, center = T, scale = T)      0.40   1    7.0438  0.008107 **
## scale(TOV, center = T, scale = T)      0.13   1    2.2371  0.135120
## Residuals                             47.09 824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When we compared the R-Square, Adjusted R-Square, and RSE we noticed that the scores were not as good as the second model and were slightly a bit lower in R-Square and Adjusted R-Square and higher in RSE. Even though, the values are not as good as Model 2 the VIF scores is much better for every variable.

4.5 Reassessing Assumptions

We will now reassess the assumptions with the previous model shown.

4.5.1 Linear Relationship

The first assumption is that there is a linear relationship between each quantitative independent variable and the response variable. In order to assess this assumption a scatter plot was generated.

```
library(GGally)
ggpairs(nba, columns = c(5, 6, 15:19, 20, 23, 24))
```

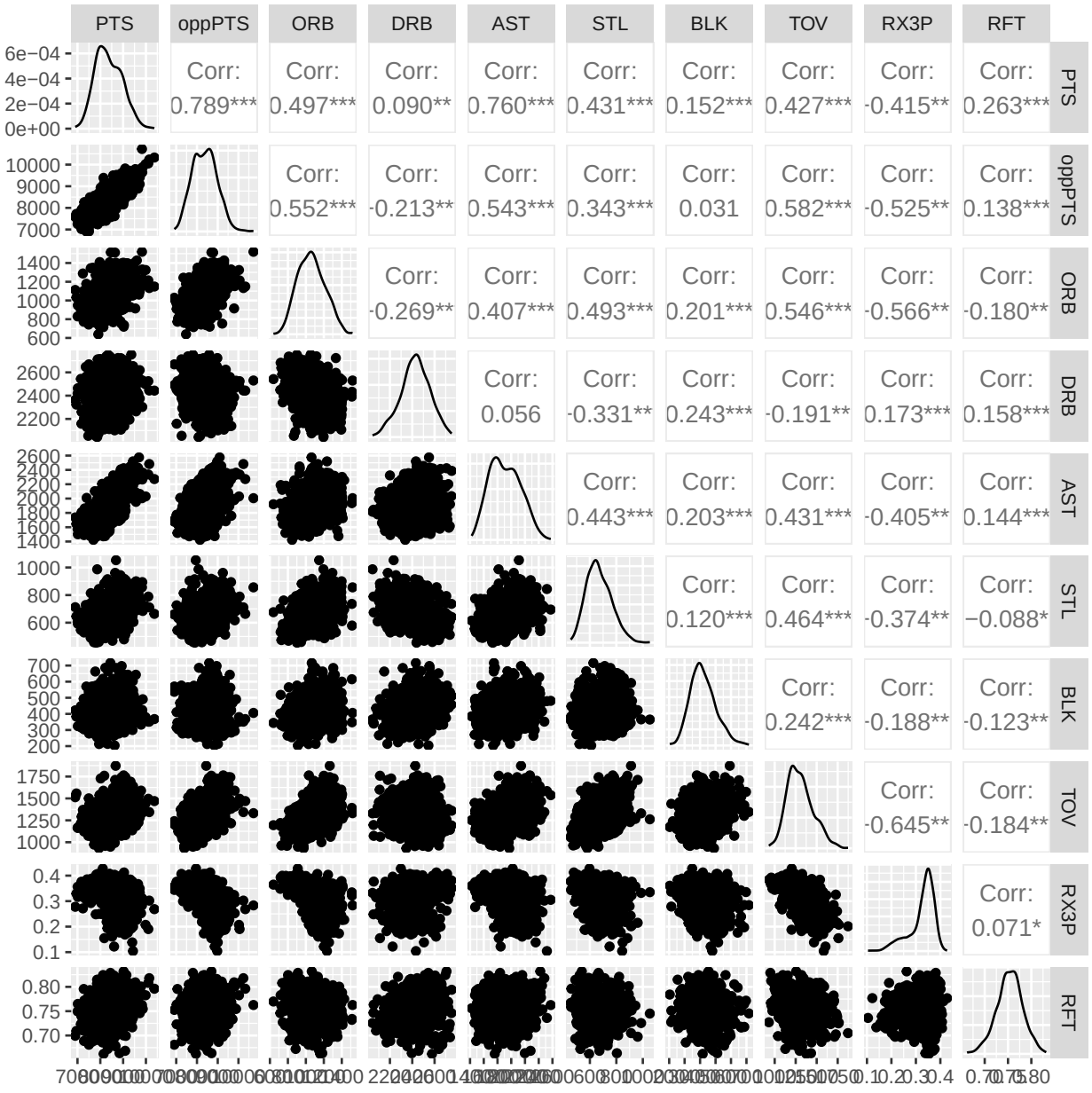


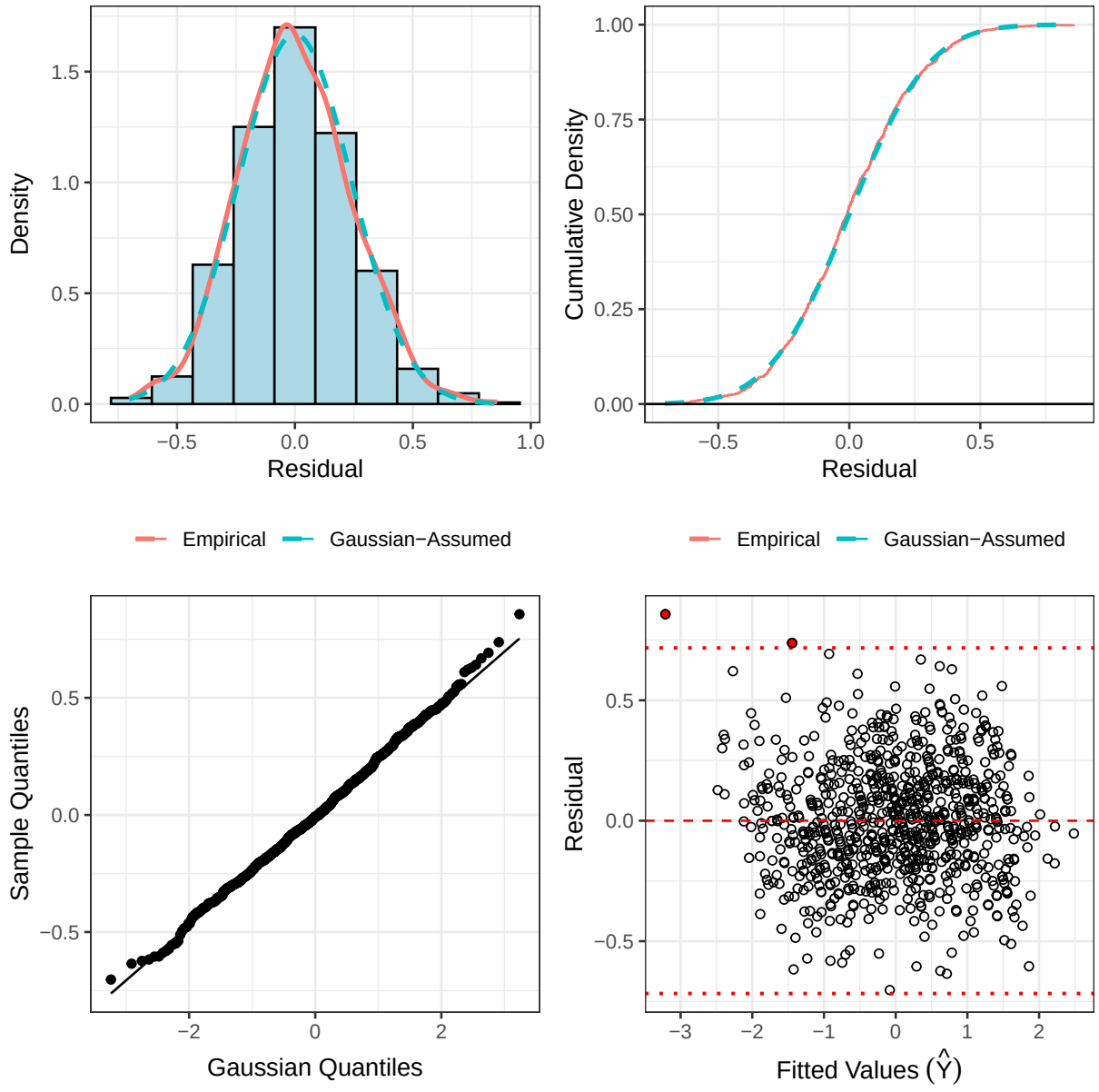
Figure 13: Scatter Plot for Model 3

We see that there is a linear relationship between the variables.

4.5.2 Constant Error Variance

The second assumptions is that residuals have a constant variance. To test this assumption we will generate different plots to check variance.

```
plotResiduals(nba.model3)
```



Looking at Figure 14 we can determine that residuals (errors) have constant variance. Even though, there's test that we could run to assess the constant variance assumptions, these

test usually reject the variance assumptions when ran. The residuals are scattered with any pattern allowing there be constant variance.

4.5.3 Residuals are Normal

The third assumption is that the errors(residuals) are approximately Gaussian distributed with mean 0. We will again use Figure 10 but this time focus on Gaussian Quantilnes which has Q-Q plot, we can see it meets normality.

4.5.4 Residuals are Independent

The fourth assumption is that the errors (residuals) are independent. We center and scaled the variables so that we can interpret the data easier. We also removed *RFG* because it had a very high multicollinearity.

4.5.5 Representative Sample

The fifth assumption is that the sample is representative which is assessed on how the data was collected. Even though, the data collected on each team is not necessarily random but the game it self is random making the data representative.

4.5.6 Multicollinearity

We again use VIF to assess the multicollineartity assumption.

```
vif(nba.model3)
```

```
##    scale(PTS, center = T, scale = T) scale(oppPTS, center = T, scale = T)
##                             6.902282                               4.999071
##   scale(RX3P, center = T, scale = T)    scale(RFT, center = T, scale = T)
##                             1.998215                               1.335089
##    scale(ORB, center = T, scale = T)    scale(DRB, center = T, scale = T)
##                             2.153147                               1.814504
##    scale(AST, center = T, scale = T)    scale(STL, center = T, scale = T)
##                             2.685714                               1.932853
##    scale(BLK, center = T, scale = T)    scale(TOV, center = T, scale = T)
##                             1.258862                               2.514708
```

The multicollinearity for the variables is the best than it has been for any previous models. The highest multicollinearity is at 6.9 which is slightly high collinearity but it has being the best we have seen. The rest of the variables have moderate multicollinearity.

4.6 Model Selection

We are taking nba.model3 as our model because it doesn't have multicollinearity and good adjusted R-squared. The coefficients are rounded to 4 decimal places for the summary of the regression coefficients.

```
nba.best <- lm(W ~ PTS + oppPTS +
                 RX3P +RFT + ORB + DRB +
```

```
                  AST + STL + BLK + TOV,
                data = nba)
round(summary(nba.best)$coefficients,4)
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  39.7812     4.8170   8.2585   0.0000
## PTS           0.0318     0.0005  66.6820   0.0000
## oppPTS       -0.0320     0.0004 -79.6598   0.0000
## RX3P         -3.5551     2.8813  -1.2338   0.2176
## RFT          -0.9089     4.3603  -0.2085   0.8349
## ORB          -0.0007     0.0010  -0.6954   0.4870
## DRB           0.0016     0.0011   1.4328   0.1523
## AST           0.0011     0.0008   1.3800   0.1680
## STL          -0.0001     0.0016  -0.0595   0.9525
## BLK           0.0038     0.0014   2.6540   0.0081
## TOV          -0.0016     0.0011  -1.4957   0.1351
```

We performed a best subset selection based on AIC (Akaike Information Criterion) where the TopModels argument specifies the maximum number of top models to output (which is 5 in our case).

```
library("bestglm")
```

```
## Loading required package:  leaps
```

```
y <- nba$W
x <- model.matrix(nba.best) [,-1]
xy <- as.data.frame(cbind(x,y))
best.subsets.aic <- bestglm(xy, IC = "AIC", TopModels = 5)
```

```
best.subsets.aic$BestModel
```

```
##
## Call:
## lm(formula = y ~ ., data = data.frame(Xy[, c(bestset[-1], FALSE),
##     drop = FALSE], y = y))
##
## Coefficients:
## (Intercept)          PTS       oppPTS          DRB          BLK
##   35.991775     0.032141    -0.032206     0.001685     0.003490
```

```
summary(best.subsets.aic$BestModel)
```

```
##
## Call:
## lm(formula = y ~ ., data = data.frame(Xy[, c(bestset[-1], FALSE),
##     drop = FALSE], y = y))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -9.0053 -2.0740 -0.0828  2.0787 10.6136
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.9917754  2.8211733  12.758  < 2e-16 ***
## PTS          0.0321406  0.0003292  97.637  < 2e-16 ***
## oppPTS      -0.0322061  0.0003300 -97.587  < 2e-16 ***
## DRB          0.0016853  0.0009310   1.810  0.07061 .
## BLK          0.0034902  0.0013343   2.616  0.00906 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.044 on 830 degrees of freedom
## Multiple R-squared:  0.9432,Adjusted R-squared:  0.9429
## F-statistic:  3445 on 4 and 830 DF,  p-value: < 2.2e-16
```

By examining the summary output, we can see the predictor variables that were included in the best model based on AIC. These variables are the ones that were deemed to have the strongest association with the response variable "W" based on the AIC value.

```
library("leaps")
regsubset.out <- regsubsets(W ~ PTS + oppPTS +
                   RX3P +RFT + ORB + DRB +
                  AST + STL + BLK + TOV,
                 data = nba, nbest = 1)

as.data.frame(summary(regsubset.out)$outmat)
```

```
##          PTS oppPTS RX3P RFT ORB DRB AST STL BLK TOV
## 1  ( 1 )                             *
## 2  ( 1 )   *      *
## 3  ( 1 )   *      *                               *
## 4  ( 1 )   *      *                 *             *
## 5  ( 1 )   *      *                 *   *         *
## 6  ( 1 )   *      *                 *   *         *   *
## 7  ( 1 )   *      *    *            *   *         *   *
## 8  ( 1 )   *      *    *       *    *   *         *   *
```

We use "leap" package to perform a best subset selection of the predictor variables for predicting the response variable "W". The regsubsets function is used to perform the best subset selection, where the nbest argument is set to 1 to select the best subset of predictor variables based on R-squared value.

```
fit.stats <- data.frame(num.variables = 1:8,
                        adjr2 = summary(regsubset.out)$adjr2,
                        bic = summary(regsubset.out)$bic)
fit.stats
```

```
##   num.variables     adjr2        bic
```

```
## 1             1 0.2208102  -195.8799
## 2             2 0.9422064 -2362.3047
## 3             3 0.9427592 -2364.6063
## 4             4 0.9429156 -2361.1693
## 5             5 0.9429520 -2355.9820
## 6             6 0.9430082 -2351.0845
## 7             7 0.9430202 -2345.5421
## 8             8 0.9429835 -2339.2872
```

The "adjr2" column contains the adjusted R-squared values for each model. The adjusted R-squared is a modification of the regular R-squared that takes into account the number of predictor variables in the model. The "bic" column contains the Bayesian information criterion (BIC) values for each model. The BIC is a measure of the goodness of fit of a statistical model, adjusted for the number of predictor variables. It is similar to the AIC but places a greater penalty on models with more predictor variables. It is better to have higher ADJIR2 value and lower BIC value. By creating this data frame, we can identify the optimal number of predictor variables to include in the model based on these criteria.

```
which.min(summary(regsubset.out)$bic)

## [1] 3

min(summary(regsubset.out)$bic)

## [1] -2364.606

coef(regsubset.out, 4)

##  (Intercept)          PTS       oppPTS          DRB          BLK
## 35.991775405  0.032140616 -0.032206069  0.001685318  0.003490228
```

According to BIC, the best model is with the variables PTS, oppPTS, BLK and DRB.

```
which.max(summary(regsubset.out)$adjr2)

## [1] 7

max(summary(regsubset.out)$adjr2)

## [1] 0.9430202

coef(regsubset.out, 5)

##   (Intercept)           PTS        oppPTS           DRB           AST
## 36.0875079639  0.0318188860 -0.0321371460  0.0018269886  0.0009276258
##           BLK
##  0.0032584824
```

According to ADJR2, the best model is with the variables PTS, oppPTS, BLK, AST and DRB.

4.7 Cross Validation

Cross-validation is done to assess the performance of a statistical model and to evaluate its ability to generalize to new data. Cross-validation can be used to to select the best model from a set of candidate models. By evaluating the performance of different models on multiple independent samples of data, we can select the model with the best performance on average, which is likely to generalize well to new data.

```
library(caret)
```

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

```
##
## Attaching package:  'caret'
```

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

```
specs <- trainControl(method = "CV",number = 10)
```

```
model.2 <- train(W ~ PTS + oppPTS + BLK,
                 data = nba,
                 method = "lm",
                 trControl = specs,
                 na.action = na.omit)
model.2
```

```
## Linear Regression
##
## 835 samples
##   3 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 752, 751, 752, 752, 752, 751, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   3.052546  0.9441537  2.436223
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
model.3 <- train(W ~ PTS + oppPTS + BLK + DRB,
                 data = nba,
                 method = "lm",
                 trControl = specs,
                 na.action = na.omit)

model.3
```

```
## Linear Regression
##
## 835 samples
##   4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 752, 751, 751, 752, 752, 752, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   3.054268  0.9437634  2.433111
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
model.4 <- train(W ~ PTS + oppPTS + BLK + DRB + AST,
                 data = nba,
                 method = "lm",
                 trControl = specs,
                 na.action = na.omit)

model.4
```

```
## Linear Regression
##
## 835 samples
##   5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 751, 752, 752, 750, 752, 752, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   3.047607  0.9432163  2.424678
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

We performed 10-fold cross-validation on three different linear regression models (model.2, model.3, and model.4) that predict the number of wins in the NBA based on different combinations of predictor variables.We got summary of the model performance, including the mean squared error (MSE) and the R-squared value (Rsquared) averaged over the 10 cross-validation folds. Model.2 and Model.3 has very similar R-squared, MSE and MAE but Model.3 is slightly better than Model.2. So, we will use PTS, oppPTS, DRB and BLK as our predictive variables for our prediction of win.

4.8 Influence Analysis

We used influence analysis to identify and investigate the potential impact of outliers or high leverage observations on the results of a statistical model.

4.8.1 Leverage Values

```
leverage <- nba %>% mutate(h.values = hatvalues(nba.model3))
p <- 11
n <- nrow(nba)
high.leverage <- leverage %>% filter(h.values > 2*p/n)
nrow(high.leverage)
```

```
## [1] 21
```

```
very.high.leverage <- leverage %>% filter(h.values > 3*p/n)
nrow(very.high.leverage)
```

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

This variable contains the hat values, which are the diagonal elements of the "hat matrix" that maps the observed response variable to the predicted values based on the predictor variables in the model. The second line of code sets the number of predictor variables p to 11 (based on the number of predictors in nba.model3). We calculated the leverage of each observation in a linear regression model, which measures how much an individual observation contributes to the fit of the model. High-leverage observations can potentially have a large influence on the estimated regression coefficients and can affect the overall performance of the model, so it is important to identify and investigate them.

4.8.2 Outliers

```
new.residuals <- nba %>% mutate(stdres = rstandard(nba.model3), studres = rstudent(nba.m
strong.outliers.stdres <-new.residuals %>% filter(abs(stdres)>3)
nrow(strong.outliers.stdres)
```

```
## [1] 2
```

```
strong.outliers.studres <-new.residuals %>% filter(abs(studres)>3)
nrow(strong.outliers.studres)
```

```
## [1] 2
```

4.8.3 Influencial Data Points

```
cooks.values <- nba %>% mutate(cooks = cooks.distance(nba.model3))
cooks.strong <- cooks.values %>% filter(cooks>1)
nrow(cooks.strong)
```

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

4.9 Regression Results

4.9.1 Research Questions

1. **What are the variables to predict the win in a game?**

   Based on the dataset, we developed a linear model to predict the number of wins using all available variables. We filtered out irrelevant data and addressed multicollinearity issues before performing model selection. To assess the model's performance, we conducted cross-validation and ultimately identified the best-performing model.

```
final.model <- lm(W ~ PTS + oppPTS + BLK + DRB, data = nba)
summary(final.model)
```

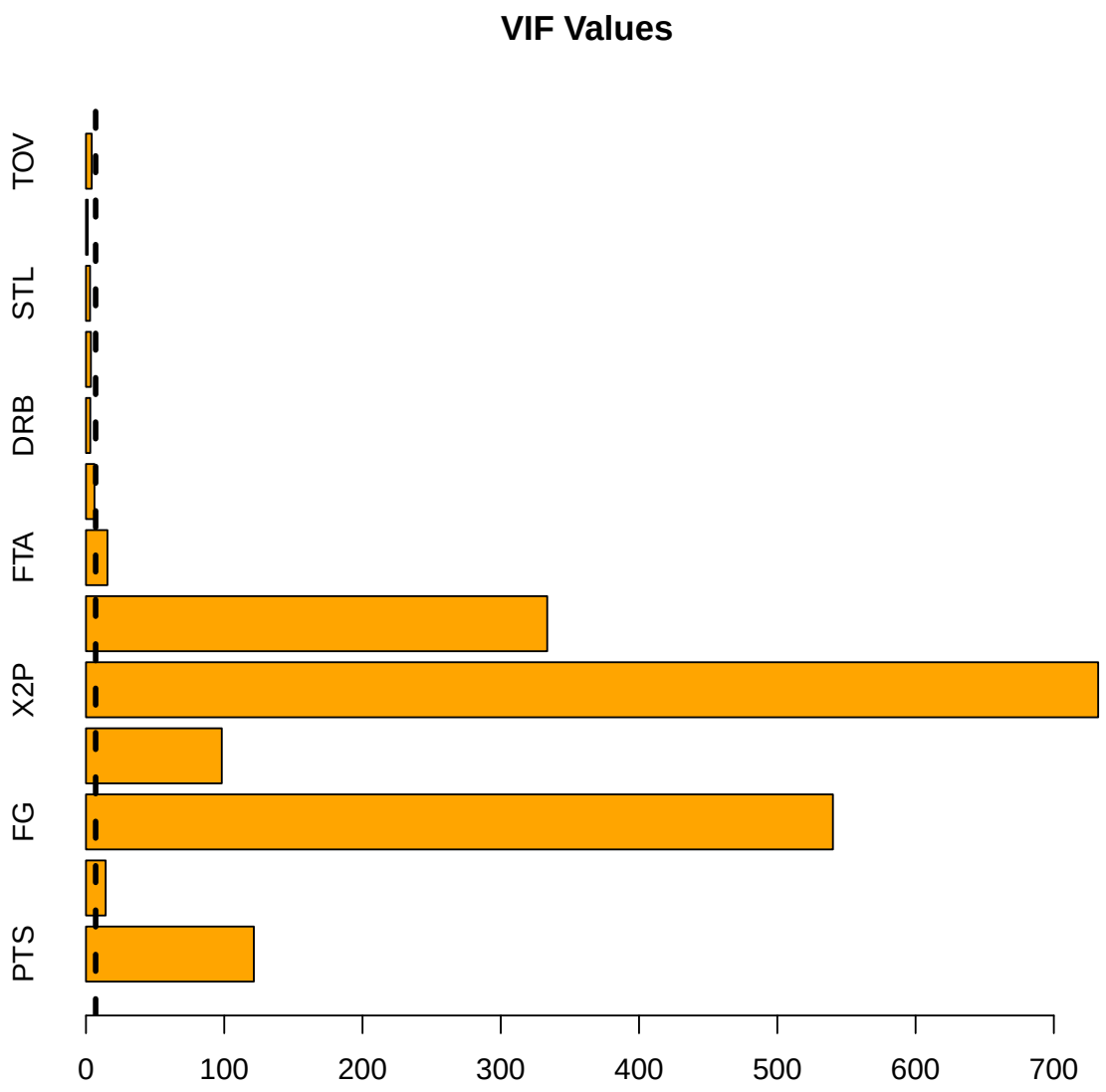
```
##
## Call:
## lm(formula = W ~ PTS + oppPTS + BLK + DRB, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0053 -2.0740 -0.0828  2.0787 10.6136
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.9917754  2.8211733  12.758  < 2e-16 ***
## PTS          0.0321406  0.0003292  97.637  < 2e-16 ***
## oppPTS      -0.0322061  0.0003300 -97.587  < 2e-16 ***
## BLK          0.0034902  0.0013343   2.616  0.00906 **
## DRB          0.0016853  0.0009310   1.810  0.07061 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.044 on 830 degrees of freedom
## Multiple R-squared:  0.9432,Adjusted R-squared:  0.9429
## F-statistic:  3445 on 4 and 830 DF,  p-value: < 2.2e-16
```

2. **Is there evidence of multicollinearity among the variables??**

Wins = 0.0323780 * PTS + -0.0324730 * oppPTS + 0.0039443 * BLK + 40.1389267 + 0.0016853 * DRB

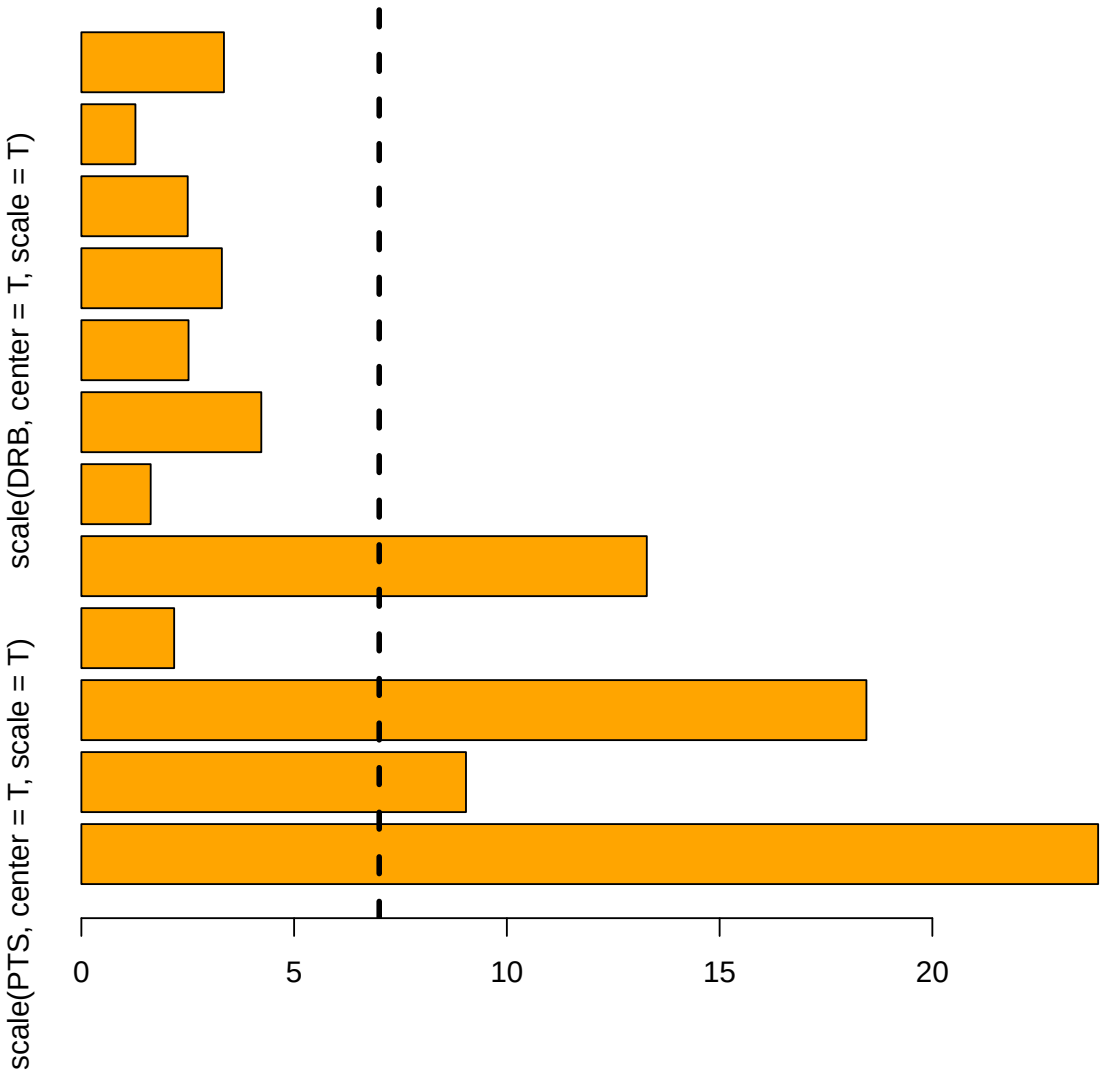
In our initial model, we observed a significant degree of collinearity among several variables such as X2P and X2PA, FG and FGA, X3P and X3PA, and FT and FTA.

```
vif <- vif(nba.model1)
barplot(vif, main = "VIF Values", horiz = TRUE, col = "orange")
abline(v = 7, lwd = 3, lty = 2)
```



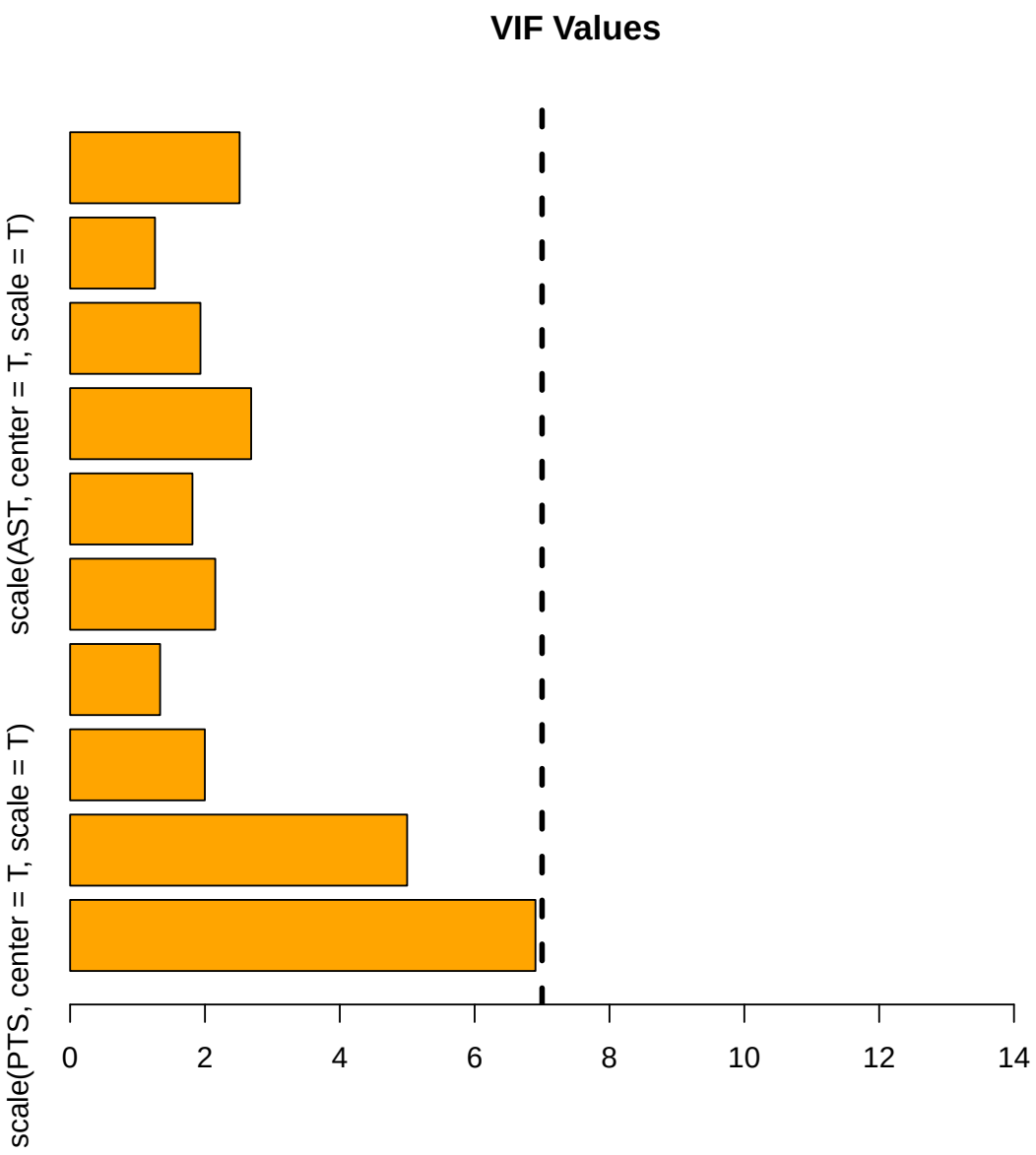
To address this issue, we calculated ratios and created a new variable, but multicollinearity persisted between RX2P, RFG, and PTS.

```
vif <- vif(nba.model2)
barplot(vif, main = "VIF Values", horiz = TRUE, col = "orange")
abline(v = 7, lwd = 3, lty = 2)
```



As RX2P and RFG were strongly correlated with PTS, we decided to eliminate them, which helped us to eliminate multicollinearity in the model.

```
vif <- vif(nba.model3)
barplot(vif, main = "VIF Values", horiz = TRUE, col = "orange", xlim = c(0, 15))
abline(v = 7, lwd = 3, lty = 2)
```



4.9.2 Model Interpretations

During the initial phase of our linear model, we noticed that there were some variables such as *X2P and X2PA, FG and FGA, X3P and X3PA, and FT and FTA* that had a high degree of correlation among them. This multicollinearity issue can cause problems in regression analysis as it can lead to unstable and unreliable coefficients. To tackle this issue, we calculated ratios for the variables that were highly correlated and created a new variable. However, even after doing so, the multicollinearity issue persisted, particularly between *RX2P, RFG, and PTS.*

To resolve this problem, we decided to remove the RX2P and RFG variables, which were

strongly correlated with PTS. This helped us to eliminate multicollinearity in the model, as these variables were no longer creating any overlap with the other predictor variables. We checked the correlation matrix plot multiple times to ensure that we have eliminated the multicollinearity problem and obtained reliable results.

After eliminating the problematic variables, we found a good correlation value between points scored and opponent points scored, which is natural in a basketball game. Although the correlation is high, it doesn't pose a concern for our model interpretation, as the score of the game is directly related to the opponent's score, and it's an essential factor to consider when analyzing a basketball game.

In summary, our initial model had a problem with multicollinearity among predictor variables, and we had to take several steps to eliminate it. We created new variables by calculating ratios and ultimately removed the highly correlated variables to ensure the reliability of our regression coefficients. Although there is a high correlation between points scored and opponent points scored, we decided to keep it in our model as it is an important factor to consider in basketball games.

5 Conclusion

In conclusion, to create a multiple linear regression model to predict the win for a team in basketball, we had to carefully examine and analyze the provided dataset. We encountered a problem of multicollinearity among some predictor variables, which we resolved by calculating ratios and removing highly correlated variables. This ensured the reliability of our regression coefficients and the overall model.

The process of building the multiple linear regression model involved trial and error as we had to make several attempts to obtain the perfect model. We had to explore and understand the data and propose research questions that motivated our analysis. The insights gained from the model can be significant in understanding the factors that lead to the win of a basketball game.

Data analytics plays a crucial role in sports, and our study highlights the importance of using statistical methods to analyze the data and gain insights into the game's dynamics. The model can be a useful tool for coaches and team managers to analyze the team's performance and make informed decisions based on the factors that contribute to the team's win. Overall, our study demonstrates the potential of data analytics in sports and its importance in decision-making processes.