Modeling NBA Player Statistics in the 2020-2021 NBA Regular Season

Introduction:

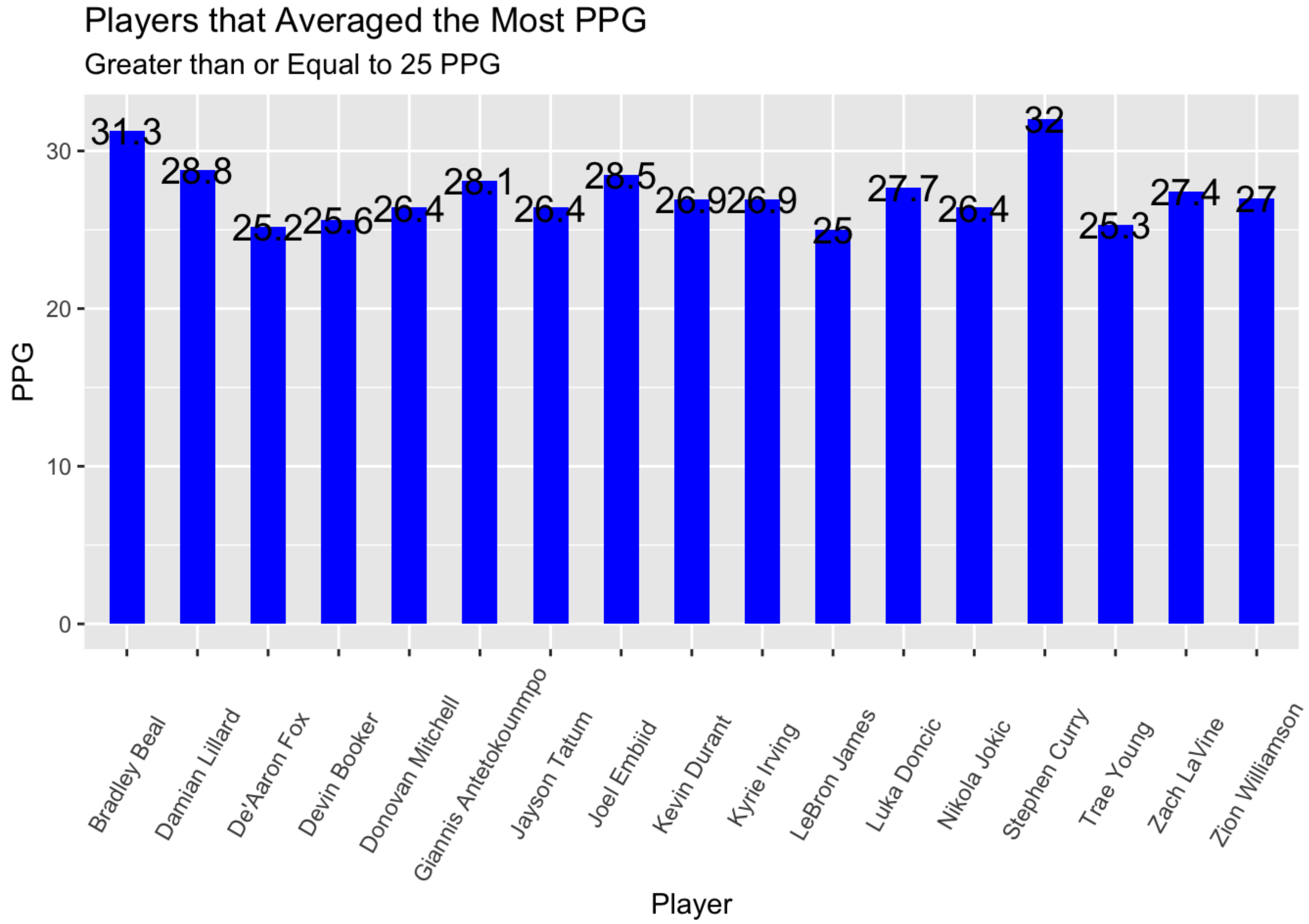
The National Basketball Association(NBA) is a professional basketball league that includes 30 teams around the United States including one team from Toronto, Canada. The league consists of the best basketball players from around the world where they all compete for the NBA Championship. Fans from around the world cheer for their favorite teams and players. There are different statistics in basketball. Points is when a player makes a basket and scores points for their team. Rebounds is when a player misses a shot and another player retrieves the ball. Assists is when a player passes the ball to a teammate and the teammate scores a basket.

Overview

In this project, I focused on NBA players statistics from the 2020-2021 NBA season. I found a data set that includes different statistics from each NBA player from the 2020-2021 season. I found this data set on Kaggle.com. The data is in a csv format. I used the read.csv function to read the data file so I can use it in R. I used the ggplot2 and “dplyr” libraries. I also added new columns to the data set which were calculated with columns already is the data. For example, I divided total points in the season by a player by the total number of games that player played in the season to get a new column which was Points Per Game(PPG). I analyzed different statistics especially Points Per Game(PPG), Rebounds Per Game(RPG), Assists Per Game(APG). I made different plots and analyzed them with a summary. The domain My research question was which players averaged the most per game in points, rebounds, and assists and how it compared to the rest of the league. I also wanted to check the effect of these different stats on how many Minuted Per Game(MPG) a player plays.

Players that Averaged the Most Points Per Game

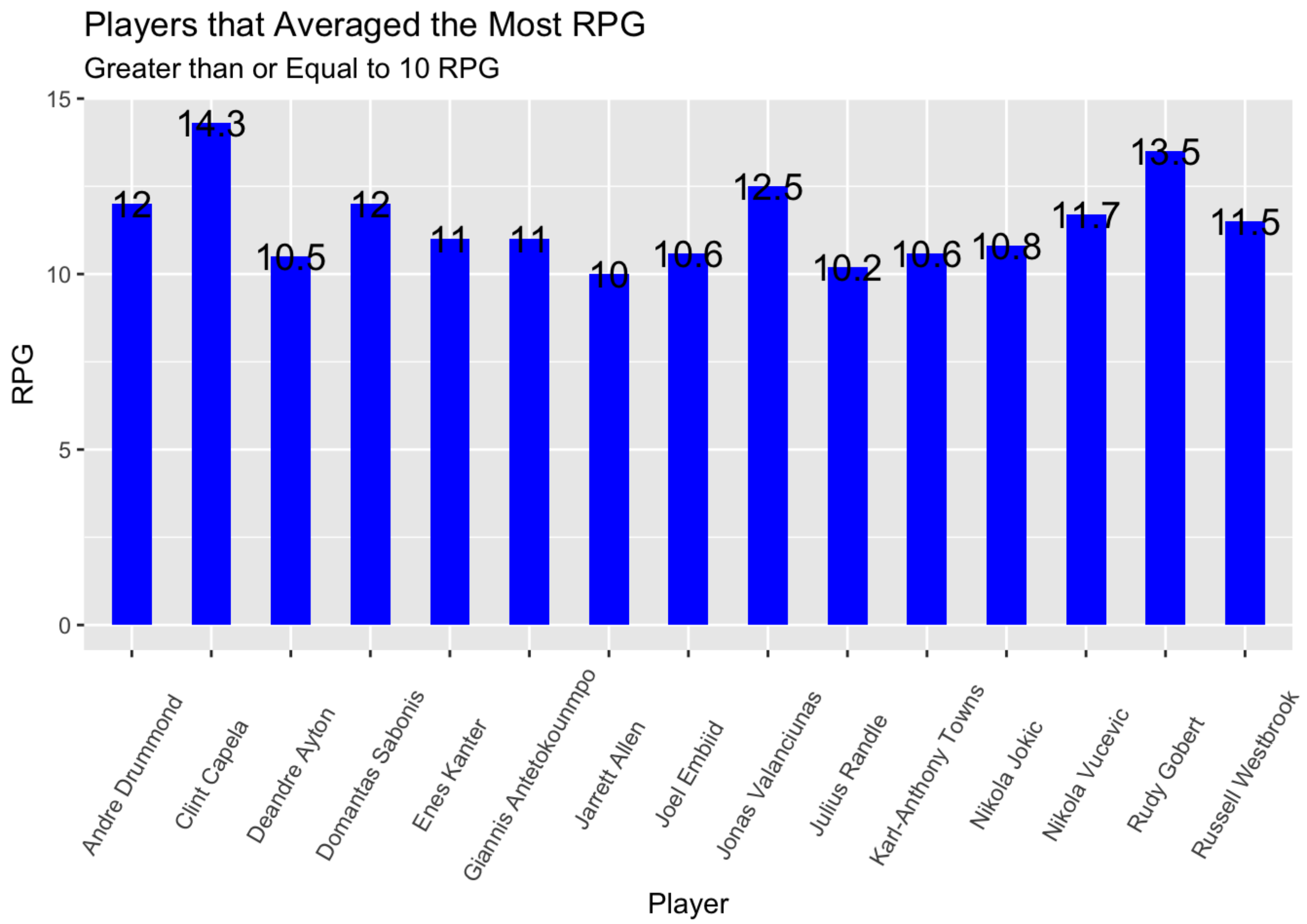
```
nba.df$PPG <- round((nba.df$PTS) / (nba.df$GP), digits = 1)

nba.df2 <- subset(nba.df, PPG >= 25)

ggplot(nba.df2, aes(x = Player, y = PPG)) +
  geom_bar(stat="identity", width=.5, fill="blue") +
  labs(title = "Players that Averaged the Most PPG",
       subtitle="Greater than or Equal to 25 PPG",
       xlab = "Player",
       ylab = "PPG") +
  theme(axis.text.x = element_text(angle=60, vjust=0.5)) +
  geom_text(aes(label = PPG), size = 5)
```



This bar chart includes the top scorers in the NBA. These are the only players that scored more than or equal to 25 points per game. The top 3 scorers include Stephen Curry(32 PPG), Bradley Beal(31.3 PPG), and Damian Lillard(28.8 PPG).

Players that Averaged the Most Rebounds Per Game

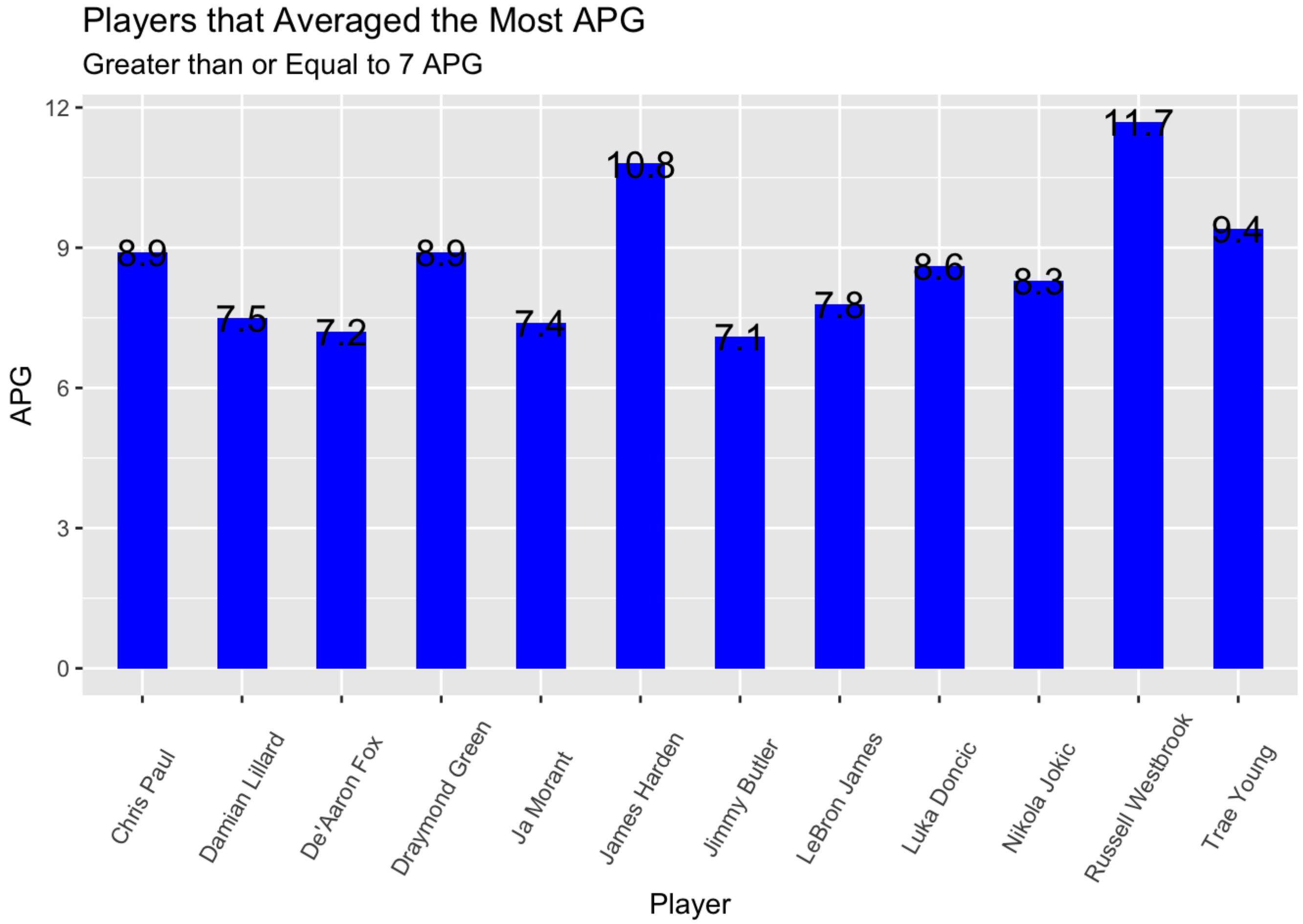
```
nba.df$RPG <- round((nba.df$REB) / (nba.df$GP), digits = 1)

nba.df3 <- subset(nba.df, RPG >= 10)

ggplot(nba.df3, aes(x = Player, y = RPG)) +
  geom_bar(stat="identity", width=.5, fill="blue") +
  labs(title = "Players that Averaged the Most RPG",
       subtitle="Greater than or Equal to 10 RPG",
       xlab = "Player",
       ylab = "RPG") +
  theme(axis.text.x = element_text(angle=60, vjust=0.5)) +
  geom_text(aes(label = RPG), size = 5)
```



This bar chart includes the players that averaged the most rebounds per game. These are the only players that averaged more than or equal to 10 rebounds per game. The top 3 rebounders include Clint Capela(14.3 RPG), Rudy Gobert(13.5 RPG), and Jonas Valanciunas(12.5 RPG).

Players that Averaged the Most Assists Per Game

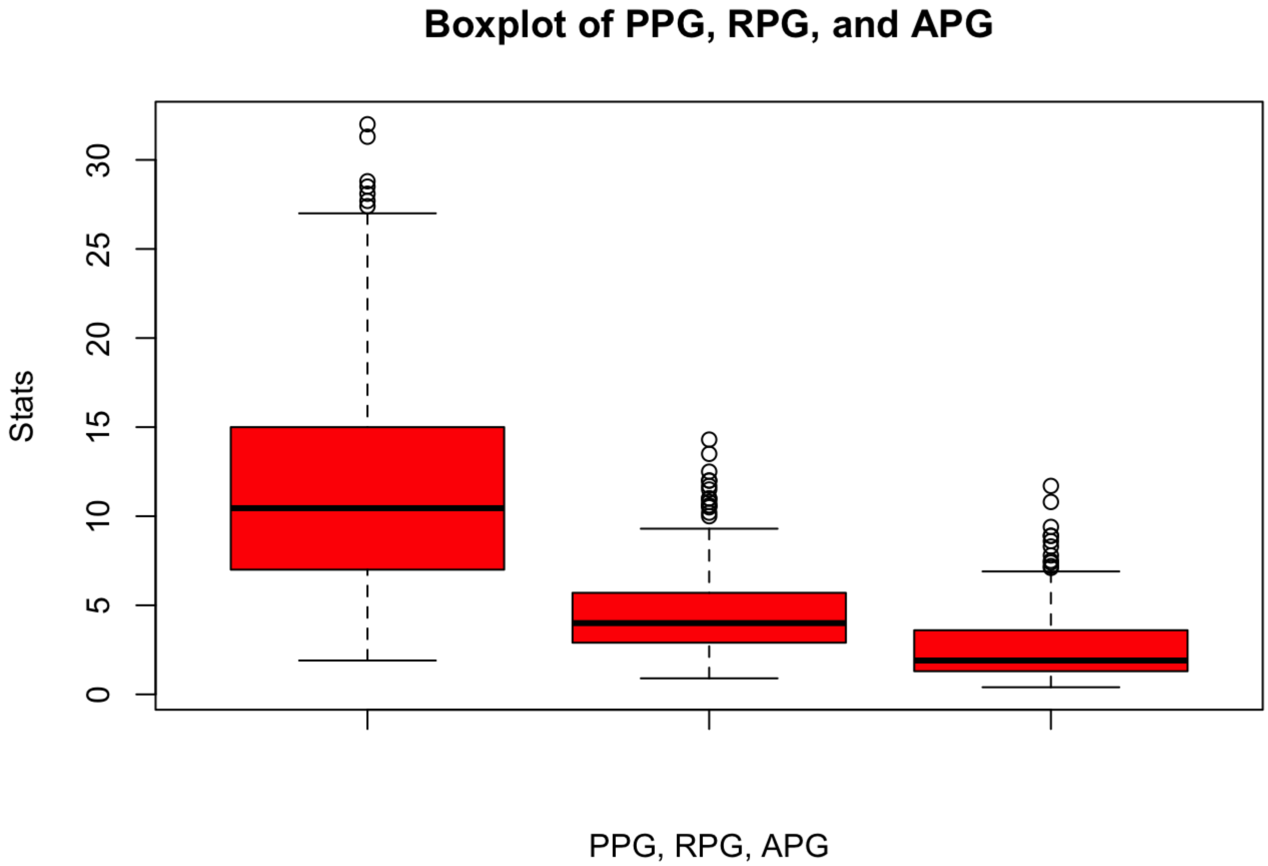
```
nba.df$APG <- round((nba.df$AST) / (nba.df$GP), digits = 1)

nba.df4 <- subset(nba.df, APG >= 7)

ggplot(nba.df4, aes(x = Player, y = APG)) +
  geom_bar(stat="identity", width=.5, fill="blue") +
  labs(title = "Players that Averaged the Most APG",
       subtitle="Greater than or Equal to 7 APG",
       xlab = "Player",
       ylab = "APG") +
  theme(axis.text.x = element_text(angle=60, vjust=0.5)) +
  geom_text(aes(label = APG), size = 5)
```



This bar chart includes the players that averaged the most assists per game. These are the only players that averaged more than or equal to 7 assists per game. The top 3 playmakers include Russell Westbrook(11.7 APG), James Harden(10.8 APG), and Trae Young(9.4 APG).

Box Plot of PPG, RPG, and APG

```
x <- boxplot(nba.df$PPG, nba.df$RPG, nba.df$APG,
        main="Boxplot of PPG, RPG, and APG",
        xlab="PPG, RPG, APG",
        ylab="Stats",
        col = "red")
```



```
summary(nba.df$PPG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.90    7.00   10.45   11.99   14.95   32.00
```

```
summary(nba.df$RPG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   2.900   4.000   4.584   5.700  14.300
```

```
summary(nba.df$APG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.400   1.300   1.900   2.715   3.600  11.700
```

In this box plot, it shows the distribution of the data for PPG, RPG, and APG. The five lines in each plot includes the minimum, Q1, median, Q3, and maximum. The circles outside of the the minimum and maximum are outliers. We can see that the medians are around 10 PPG, 4 RPG, and 2 APG. We can see that top scorer in the league, Stephen Curry(32 PPG), is much greater than the median and also greater than the maximum. We can see it is a outlier as 32 is above the maximum line. We can see that the top rebounder, Clint Capela(14.3 RPG), is also an outlier. We can also see that the player that averaged the most APG, Russell Westbrook(11.7 APG), is also an outlier.

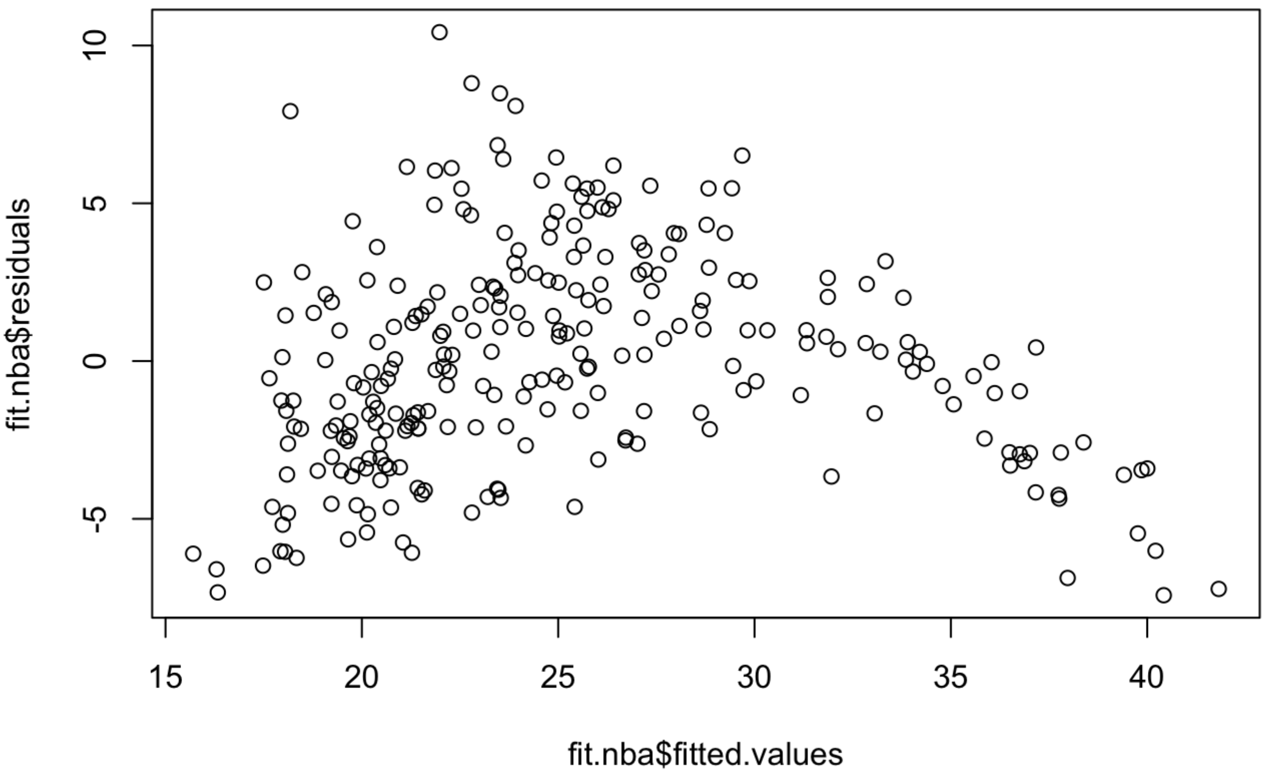
Linear Regression

```
nba.df$MPG <- round((nba.df$MIN) / (nba.df$GP), digits = 1)
nba.df$TPG <- round((nba.df$TOV) / (nba.df$GP), digits = 1)
fit.nba <- lm(MPG ~ PPG + RPG + APG + TPG, data = nba.df)
summary(fit.nba)
```

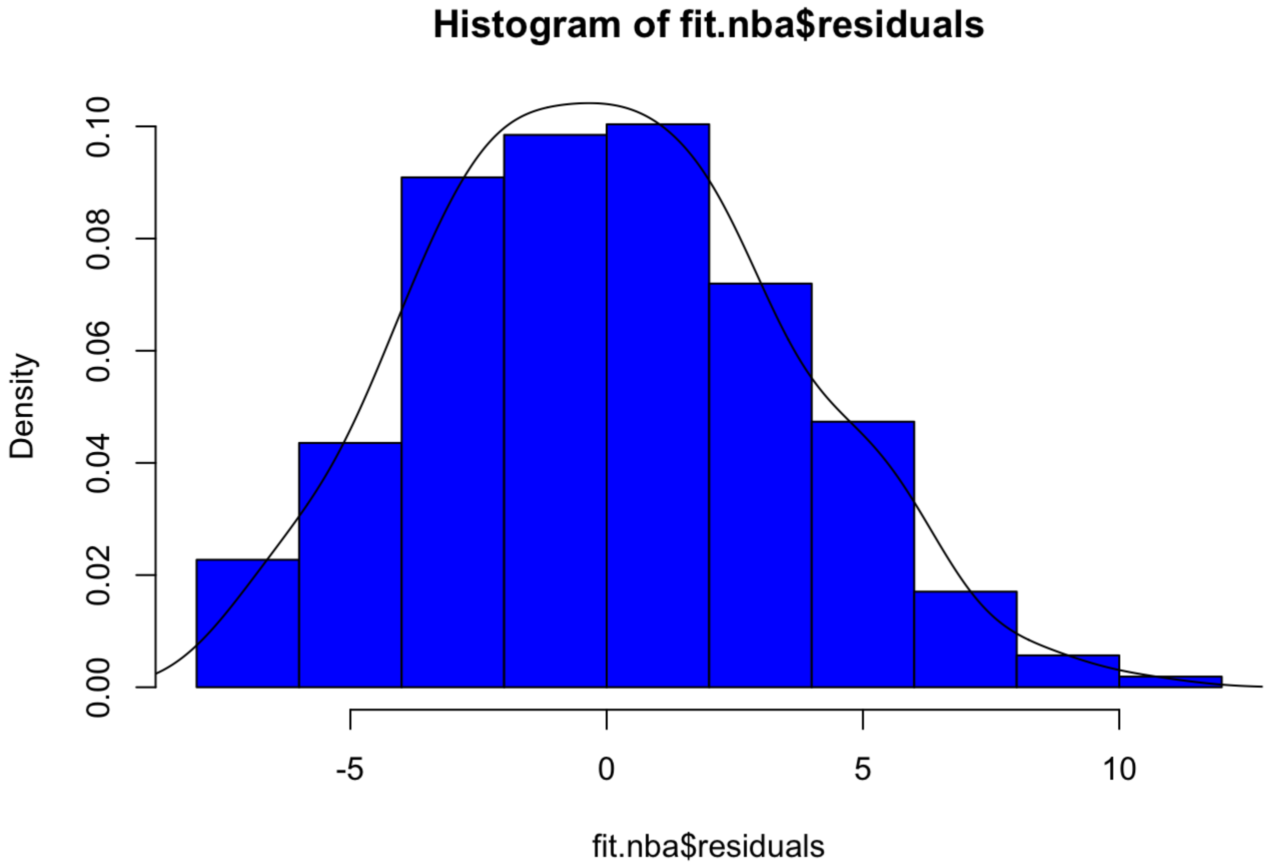
```
## 
## Call:
## lm(formula = MPG ~ PPG + RPG + APG + TPG, data = nba.df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.4208 -2.4712 -0.1644  2.3953 10.4218 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 13.42725    0.51867  25.888  < 2e-16 ***
## PPG          0.72941    0.05837  12.495  < 2e-16 ***
## RPG          0.53759    0.10669   5.039 8.80e-07 ***
## APG          0.95854    0.20407   4.697 4.28e-06 ***
## TPG         -1.49169    0.65682  -2.271    0.024 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.486 on 259 degrees of freedom
## Multiple R-squared:  0.7367, Adjusted R-squared:  0.7326 
## F-statistic: 181.2 on 4 and 259 DF,  p-value: < 2.2e-16
```

I fitted a linear model using PPG, RPG, APG, and TPG as the predictors and MPG as the predicted variable. I wanted to see the effect on how many minutes per game(MPG) a player plays depending on their game averages. In this summary, we can see that adding 1 PPG, a player plays an additional 0.660385 MPG. We can also see that adding 1 RPG, a player plays an additional 0.454139 MPG. We can also see that adding 1 APG,a player plays an additional 0.643647 MPG. We also checked how turnovers per game(TPG) effects a players MPG. We can see that adding 1 TPG, a player loses 1.49169 MPG. TPG has a negative effect on a players MPG unlike any other statistic that we tested. APG has the greatest positive effect on a players MPG. This model also explains for 73.26% variance.

```
plot(fit.nba$residuals~fit.nba$fitted.values)
```



```
hist(fit.nba$residuals, prob = TRUE, col = "blue")
lines(density(fit.nba$residuals))
```



In these plots, I plotted the residuals. The fit is good as the errors are fairly evenly distributed around 0.

```
fit.nba.step <- step(fit.nba)
```

```
## Start:  AIC=664.33
## MPG ~ PPG + RPG + APG + TPG
##
##        Df Sum of Sq    RSS    AIC
## <none>              3147.9 664.33
## - TPG   1     62.69 3210.6 667.54
## - APG   1    268.16 3416.1 683.92
## - RPG   1    308.58 3456.5 687.02
## - PPG   1   1897.64 5045.5 786.88
```

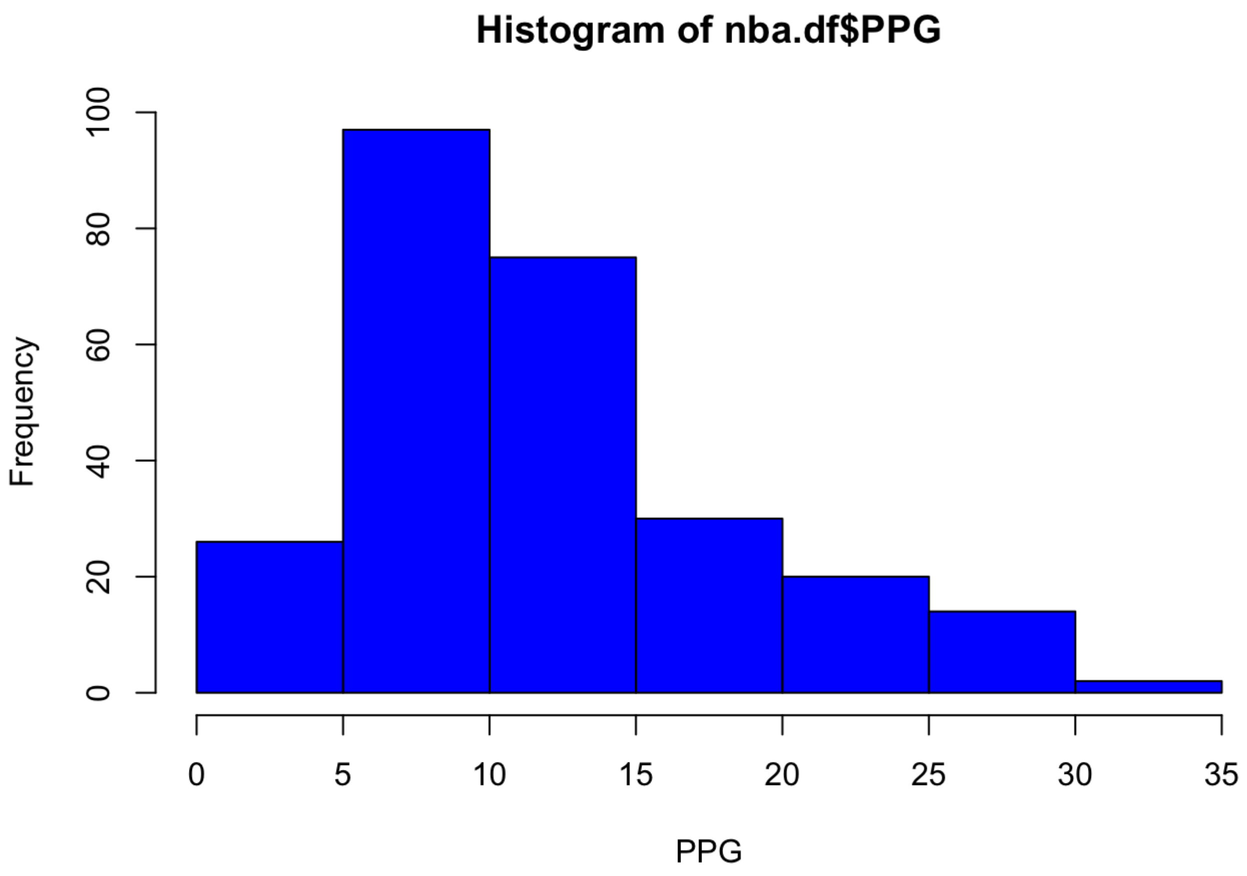
```
fit.nba.step
```

```
##
## Call:
## lm(formula = MPG ~ PPG + RPG + APG + TPG, data = nba.df)
##
## Coefficients:
## (Intercept)          PPG          RPG          APG          TPG
##     13.4272       0.7294       0.5376       0.9585      -1.4917
```

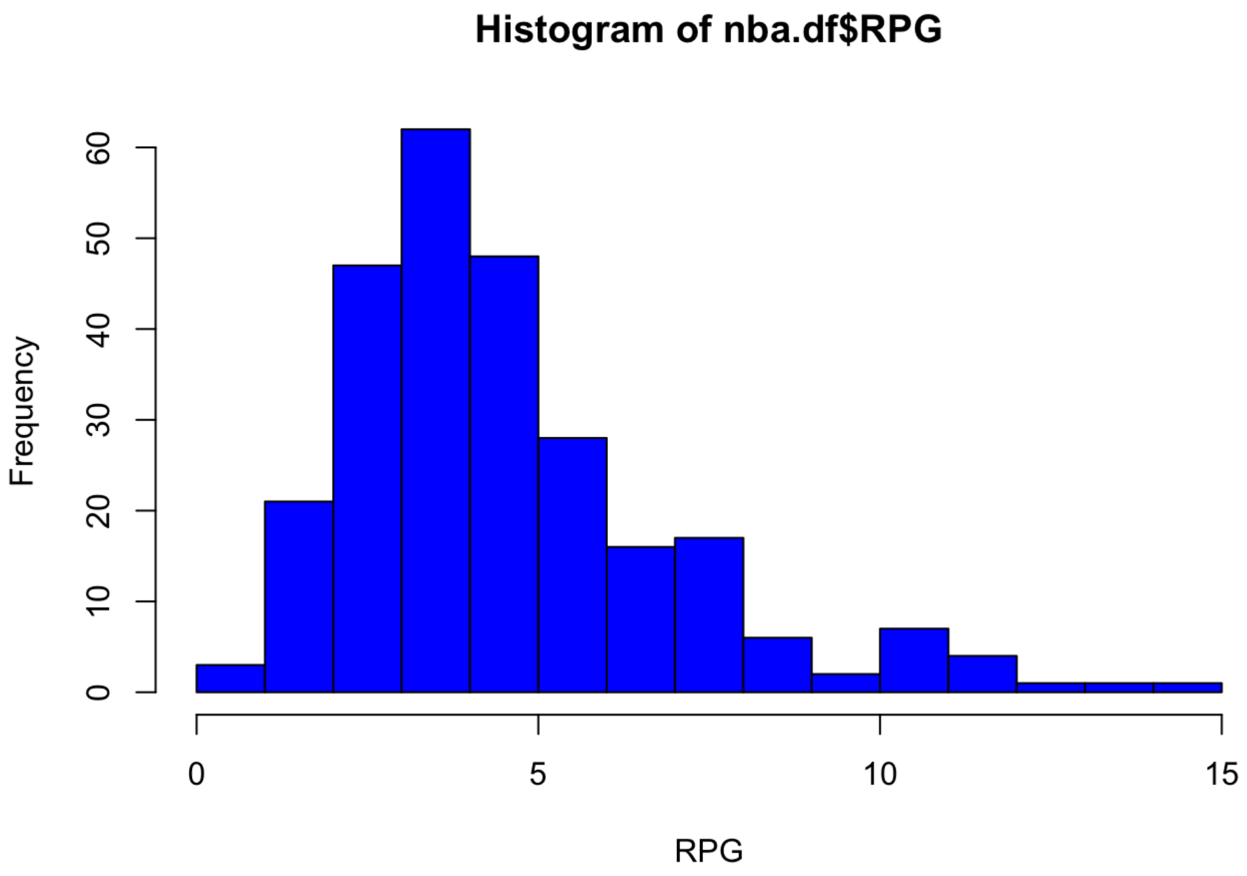
I used the step function to see which model is the “best”. The model that I used is the best model that I can use.

Histograms for PPG, RPG, and APG

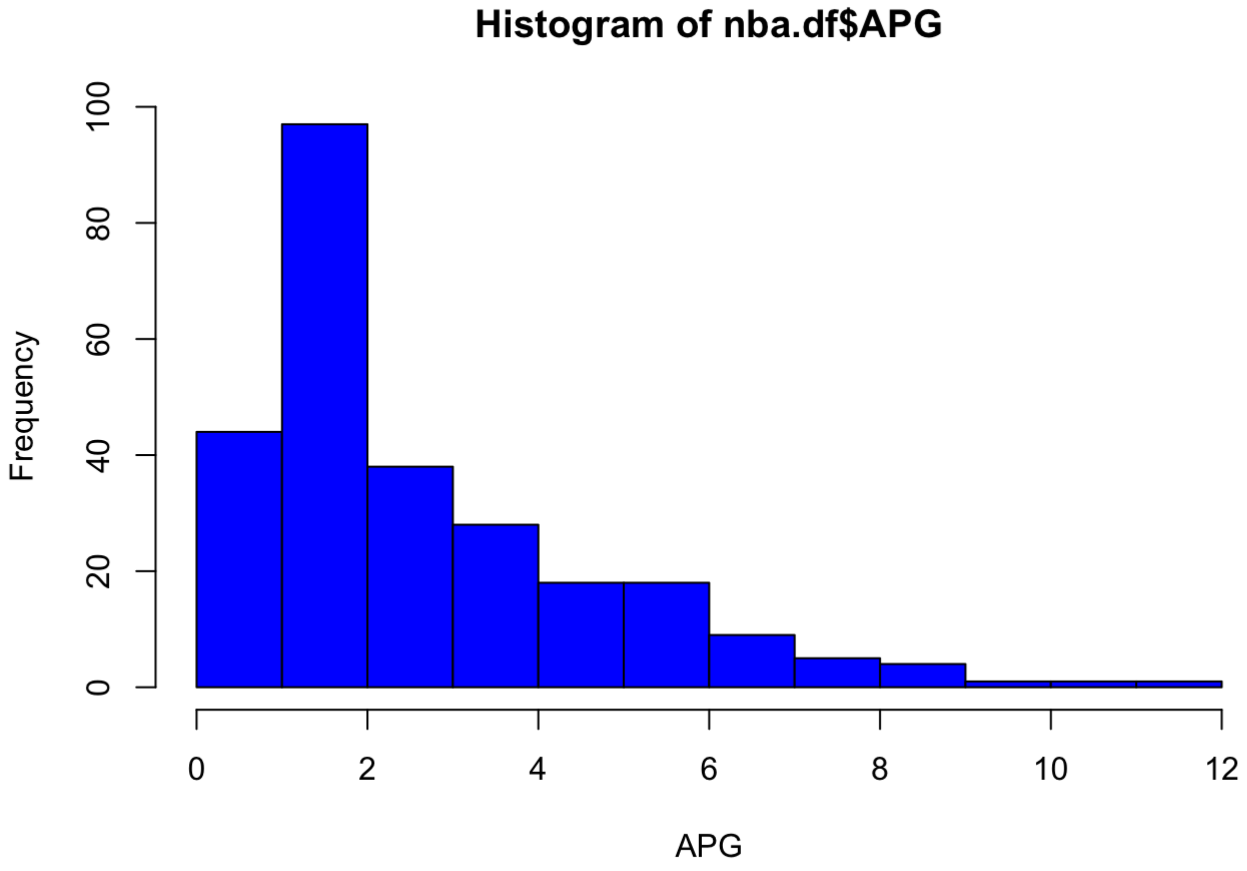
```
hist(nba.df$PPG, xlab = "PPG", col = "blue")
```



```
hist(nba.df$RPG, xlab = "RPG", col = "blue")
```



```
hist(nba.df$APG, xlab = "APG", col = "blue")
```



These histograms show the frequency of different statistical averages. For example, it shows that almost 100 different players averaged 5 to 10 PPG in the season. It also shows that very few players averaged above 10 RPG in the season. It also shows that almost 40 different players averaged 2 to 3 APG. All three of these histograms are skewed to the right.

Conclusion

After analyzing the data and making many different plots, my questions were answer. Stephen Curry averaged the most points per game with 32 PPG. This is very good compared to the rest of league as it is an outlier in the box plot. Clint Capela averaged the most rebounds per game with 14.3 RPG. This was also an outlier which means it is much greater than other players average rebounds. Russell Westbrook averaged the most assists per game with 11.7 APG which is also an outlier and very impressive.

We even answered my other question on how a players averages in PPG, RPG, APG, and TPG effect a players MPG. PPG, RPG, and APG all have positve effect on MPG while TPG has a negative effect on TPG. TVP has the overall greatest effect on MPG as its coefficient was the greatest even though it was negative. APG has the greatest positive effect on MPG followed by PPG and then RPG. This suggests that if a player wants to play more minutes per game, they need to average more assists, points, and rebounds and average less turnovers per game.