S&DS 230 Final Project: Predicting Effective Field Goal Percentage (eFG%) in the NBA

Mark Ayiah

August 7, 2022

Part I: Introduction

As the NBA has evolved, statistics have become an increasingly important part of how players, staff, and fans understand the game. The introduction of advanced statistics, detailed play-by-play tracking, and entire team departments dedicated to data analysis points towards the notion that analyzing the game from a statistical lens is an effective way to reach more basketball success. With this project, my aim is to participate in the basketball community's statistical discourse, and more specifically, to examine a particular statistic: effective field goal percentage (eFG%). My goal is to investigate the relationships between eFG% and various other metrics in order to eventually create a model that can accurately predict a player's eFG% based on how they perform in other facets of the game.

Part II: Data Collection | Web Scraping

I sought to use data from the most recent, complete NBA cycle, so I used Basketball Reference's catalog of statistics from 2021-22 NBA season. The link to the site is here. The data contains over 20 different categories of player statistics from over 500 unique NBA players. The statistics are calculated on a per game scale, with a complete player season being at most 82 games. To access this data in R, I had to manually scrape the values from the website, parse through the html encoding, and create a data frame/table of the values. Here is how I accomplished this:

(All data and statistical descriptions are courtesy of Basketball Reference)

```
# Data Scraping
nbaURL <- "https://www.basketball-
reference.com/leagues/NBA_2022_per_game.html"
nbaWebpage <- read_html(nbaURL)
nbaHTML <- html_nodes(nbaWebpage, "td")
nbaData <- html_text(nbaHTML)
nbaColNamesHTML <- html_nodes(nbaWebpage, ".center+ .poptip , .center+
.poptip")
nbaColNames <- html_text(nbaColNamesHTML)

# Data Frame Creation
nba <- data.frame(matrix(ncol = 29, nrow = 0))
colnames(nba) <- nbaColNames
```

```
# Creates a row with each player's information and adds it to the overall nba
data frame
playerIndex <- 1
for (i in 1:605) {
  playerRow <- nbaData[playerIndex:(playerIndex+28)]
  nba[(i), ] <- playerRow
  playerIndex <- playerIndex + 29
}
```

For the remainder of this paper, I will only be using a select few of the variables from the original data set. They are listed here along with their descriptions:

- Pos = Position
- Age = Player's age on February 1 of the season
- FG% = Field Goal Percentage
- eFG% = Effective Field Goal Percentage
  - This statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal.
- FT% = Free Throw Percentage
- TRB = Total Rebounds Per Game
- AST = Assists Per Game
- STL = Steals Per Game
- BLK = Blocks Per Game
- PTS = Points Per Game

Here is a glimpse at what the raw data looks like:

Player	Pos	Age	Tm	G	GS	MP	FG	FGA	FG%
Precious Achiuwa	C	22	TOR	73	28	23.6	3.6	8.3	.439
Steven Adams	C	28	MEM	76	75	26.3	2.8	5.1	.547
Bam Adebayo	C	24	MIA	56	56	32.6	7.3	13.0	.557
Santi Aldama	PF	21	MEM	32	0	11.3	1.7	4.1	.402
LaMarcus Aldridge	C	36	BRK	47	12	22.3	5.4	9.7	.550
Nickeil Alexander-Walker	SG	23	TOT	65	21	22.6	3.9	10.5	.372

Part III: Data Cleaning

There was a bit of work to do to make the data suitable for use. The most immediate issue was that all of the data was loaded in as strings/characters as opposed to numbers (i.e the computer was registering values like Precious Achiuwa's field goal percentage as the string ".439" instead of the numeric value .439). Thus, I corrected the data types for each column. Here is an example of how I accomplished this:

```
# Specifying the columns that I wanted to change to numeric values
numeric <- c(7:29)
```

```r
# Loop that iterates through the specified columns and changes their data
type to numeric
for (i in numeric) {
  nba[, i] <- as.numeric(nba[, i])
}
```

Next, I noticed that for players who were traded or signed to different teams during the season, the table had multiple rows of data for their statistics with each team that they played for and their total season averages. To avoid having to deal with these duplicate players, I decided to only use players who played at least 50 games for one team during the season. Thus, the frame of reference for the data would only include players who regularly made the rotation for their team. This was done with this line of code:

```r
nba <- nba[nba$G >= 50 & !(nba$Tm %in% "TOT"), ]
```

Finally, I decided to categorize the Age and Position variables. I put players in age groups of 4 year intervals (Ages 18-22, 23-27, 28-32, 33-37, and 38+) so that I could later see how these age groups relate to eFG%. I also put their positions in play style groups. Point guards, shooting guards, and small forwards tend to shoot further from the basket in comparison to power forwards and centers, so I split them into groups of guards/wings (guards & small forwards) and bigs (power forwards and centers) to later see if position was relevant when discussing eFG%. Finally, I narrowed down the data to only include the relevant metrics to be used throughout the rest of this paper. Here is what the data looks like now:

NOTE: The column names "eFG%", "FT%", and "FG%" were changed to "eFGpct" and "FTpct", and "FGpct" for compatibility reasons.

	FGpct	eFGpct	Pos	Age	FTpct	TRB	AST	STL	BLK	PTS
Precious Achiuwa	0.439	0.486	Big	18-22	0.595	6.5	1.1	0.5	0.6	9.1
Steven Adams	0.547	0.547	Big	28-32	0.543	10.0	3.4	0.9	0.8	6.9
Bam Adebayo	0.557	0.557	Big	23-27	0.753	10.1	3.4	1.4	0.8	19.1
Nickeil Alexander-Walker	0.375	0.450	Guard/Wing	23-27	0.722	3.3	2.8	0.8	0.4	12.8
Grayson Allen	0.448	0.588	Guard/Wing	23-27	0.865	3.4	1.5	0.7	0.3	11.1
Jarrett Allen	0.677	0.678	Big	23-27	0.708	10.8	1.6	0.8	1.3	16.1

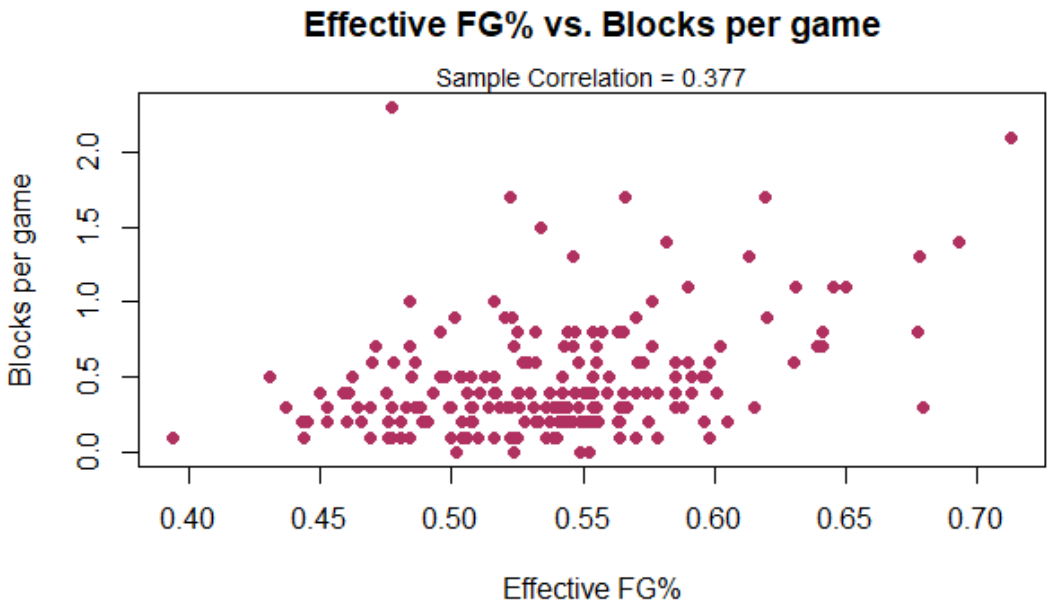
Part IV: Basic Testing

A: Finding Correlations

Now that the data has been cleaned and is ready to use, we can first do some basic analysis to get a better feel for the data set. Finding the correlation between each continuous variable is helpful because it gives us a glimpse at what predictors may end up being significant down the road when trying to fit a regression model to predict the response variable, eFG%. The correlations with eFG% are as follows:

```
##           FTpct       TRB        AST         STL       BLK         PTS
## [1,] -0.162159 0.3588543 -0.1481746 -0.05112717 0.3768247 0.08128397
```

Blocks per game had the strongest correlation with eFG% out of the selected predictors. Here is a scatterplot to better visualize the relationship:



As shown by both the correlation value and the scatterplot, there is a positive, moderate, linear relationship between a player's eFG% and their blocks per game.

B: T-Testing

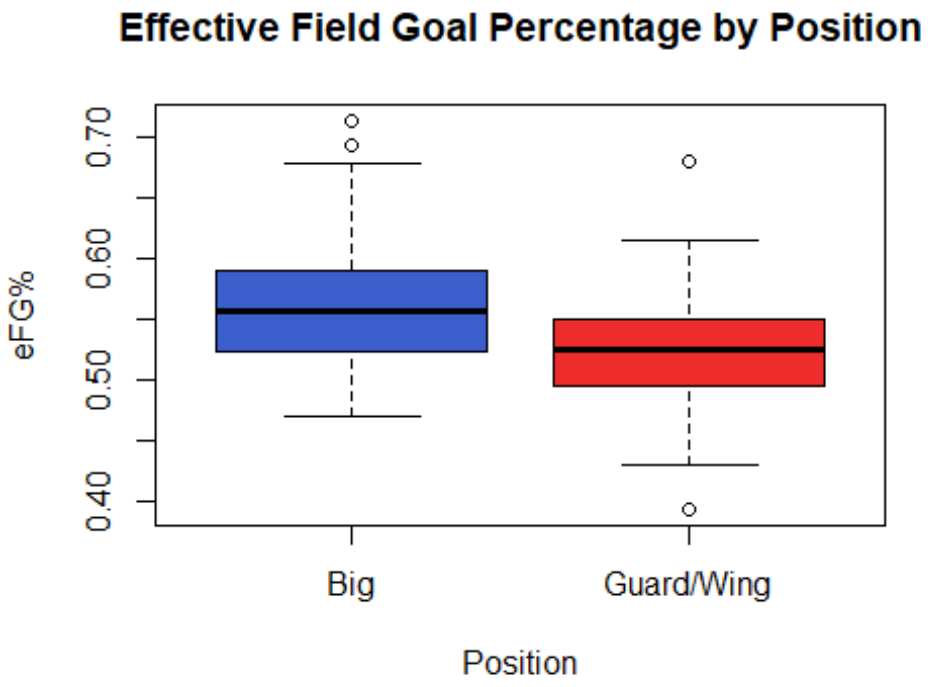
Next, we can test to see if a player's position had any measurable relationship with their eFG%. With regular field goal percentage, taller players (mostly forwards and centers) tend to shoot higher percentages because more of their looks come closer to the basket and thus are less likely to miss. A quick one-sided 2 Sample T-Test confirms this:

```
##
##  Welch Two Sample t-test
##
```

```
## data:  FGpct by Pos
## t = 7.0504, df = 100.49, p-value = 1.151e-10
## alternative hypothesis: true difference in means between group Big and
group Guard/Wing is greater than 0
## 95 percent confidence interval:
##  0.05280911        Inf
## sample estimates:
##        mean in group Big mean in group Guard/Wing
##                0.5096757                0.4406016
```

In this case, the null hypothesis is that the mean difference in FG% between the positional groups is less than or equal to zero, and the alternative hypothesis is that the mean FG% for bigs is greater than the mean FG% for guards/wings. Since the p-value very closely approaches 0 and is statistically significant at the .05 significance level, we reject the null hypothesis and conclude that, on average, bigs shoot higher FG% than guards & wings.

However, effective field goal percentage is meant to account for the fact that three-pointers are worth more (and are harder to make). First, we can look at a boxplot to see how the two distributions stack up against each other.



The boxplot showcases a clear difference between the groups' distributions. On average, it would appear that bigs shoot higher effective field goal percentages. However, we can do another T-Test to see if there is as much of a difference as we observed with regular FG% (or any difference at all) and its statistical significance:

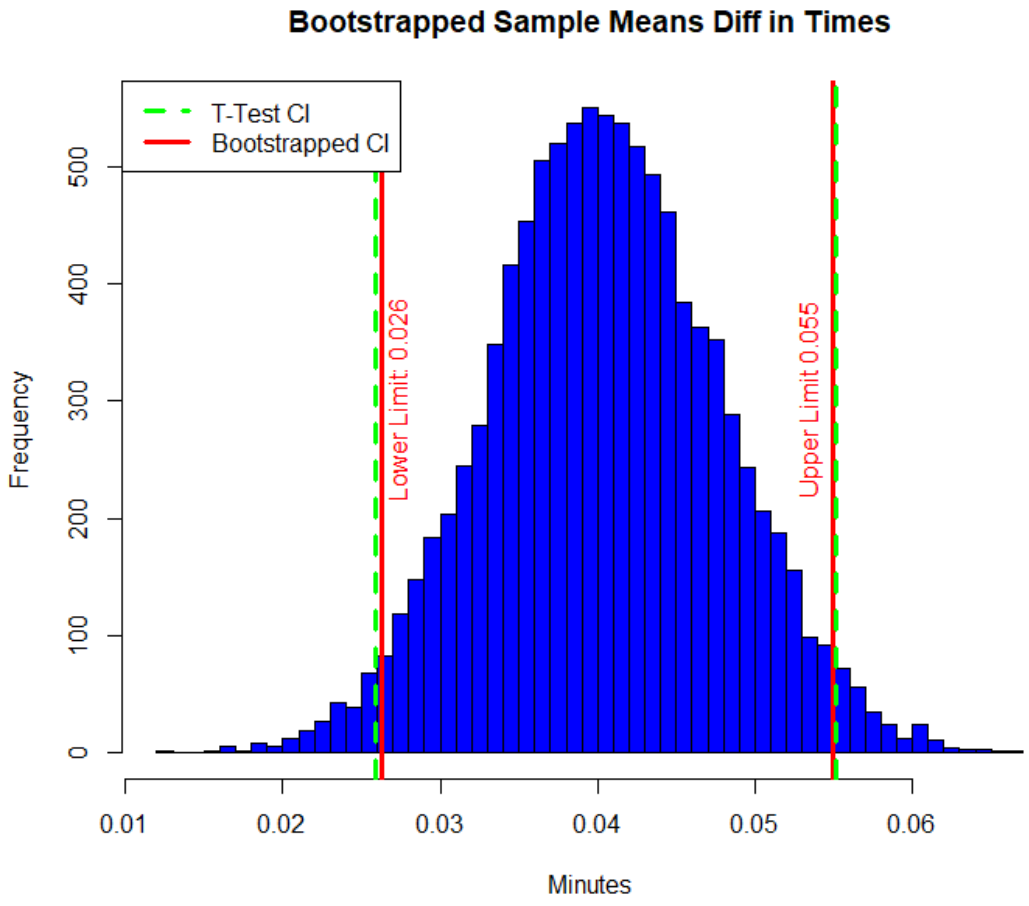
```
##
##  Welch Two Sample t-test
```

```
##
## data:  eFGpct by Pos
## t = 5.4681, df = 128.93, p-value = 1.135e-07
## alternative hypothesis: true difference in means between group Big and
group Guard/Wing is greater than 0
## 95 percent confidence interval:
##  0.02823052        Inf
## sample estimates:
##        mean in group Big mean in group Guard/Wing
##                0.5624865                0.5219844

## [1] "Difference in mean FG% between groups: 0.0690741131756757"

## [1] "Difference in mean eFG% between groups: 0.0405021114864864"
```

Here, we also get a p-value that very closely approaches 0 and is statistically significant at the .05 significance level. Thus, we can conclude that there is a statistically significant difference in the mean eFG% between guards/wings and bigs and that, on average, bigs shoot at a higher clip. This may also indicate that more advanced metrics such as true shooting percentage (which adjusts for three-pointers and free throws) may be more appropriate to comprehensively assess a player's shooting efficiency. However, the difference in mean eFG% between positional groups was not as large as the mean FG% between positional groups, which indicates that eFG% could be a better measure of efficiency than standard FG%.

C: Bootstrapping

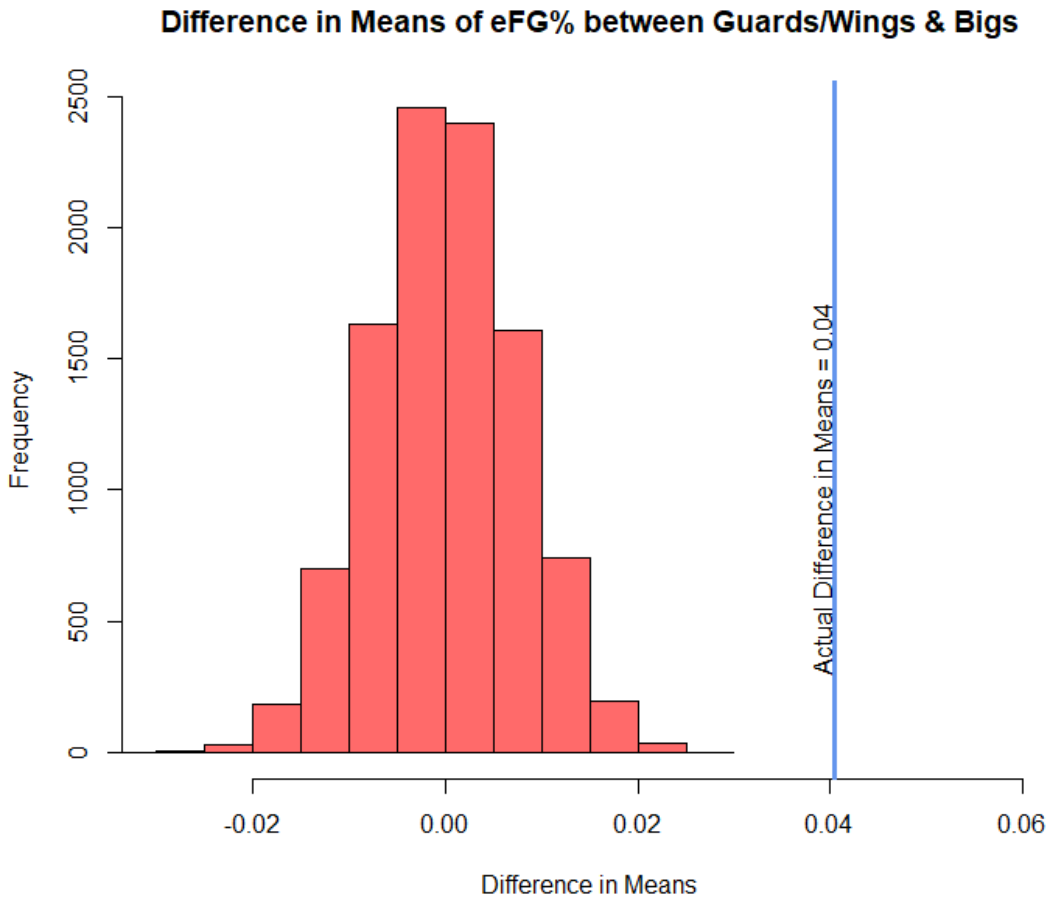
We can also use bootstrapping to better visualize the true difference in the means of eFG% between groups of positions. Bootstrapping essentially treats the data as overall population data and creates a sampling distribution of means based on samples from data. Then, based on the sampling distribution (which turns out to be normally distributed), we can create confidence intervals for the true population difference of means. The results of our bootstrap are as follows:



These results tells us that we are 95% confidence that the true population difference in mean eFG% between guards/wings and bigs is between 0.026% and 0.055% according to the bootstrapped test. This interval is a bit more narrow than the confidence interval provided by a two-sided t-test, which is a benefit of the added accuracy of bootstrapping.

Part V: Permutation Testing

We can also use a permutation test to quantify how likely it is that the difference in the means of eFG% between groups was simply due to chance. Permutation testing repeatedly shuffles the order of the categorical variable (in this case the player's position), creating new pairs of data to find an overall mean from, and then creates a sampling distribution of the mean from each repetition. Then, based on that distribution (which turns out to be normally distributed), we can calculate the probability that we would see a value at least as extreme as the observed value purely due to chance. For this test, the null hypothesis states that there is not sufficient evidence to suggests that the difference in means between the two categorical groups is not due to chance. At a 95% confidence level, if the p-value (probability) is less than .05, we reject the null hypothesis and conclude that the difference in means is statistically significant and not due to chance. The results of our permutation test are as follows:

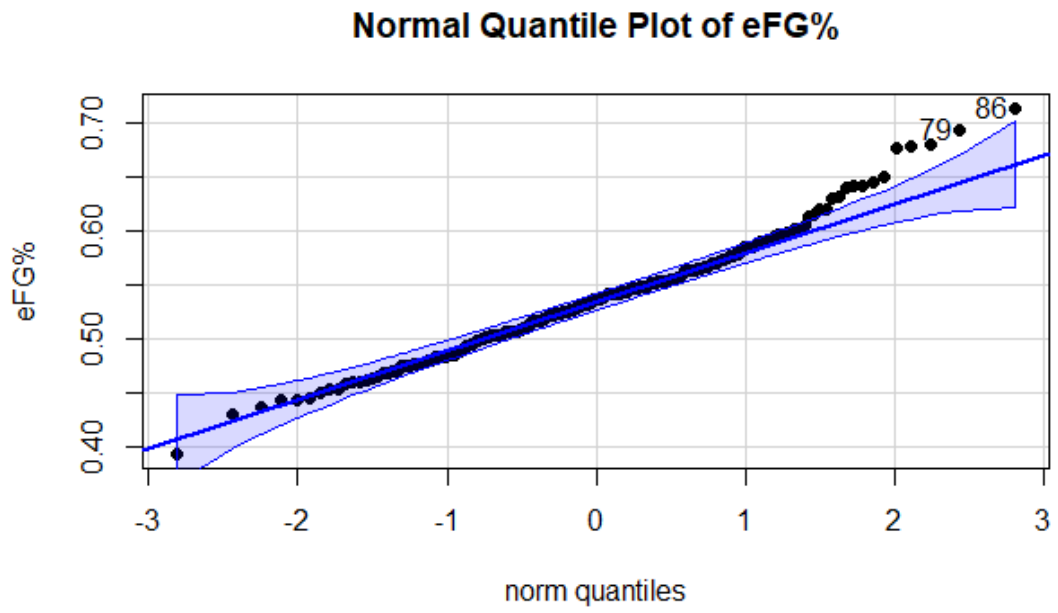



```
## [1] "p-value: 0"
```

The p-value is effectively 0 which is a very strong indication that the observed difference in the means of eFG% between position group was not due to chance.

Part VI: Multiple Regression

Now, we can fit a model that best predicts eFG% based on a combination of the other statistics in our data set. First, we can look at the distribution of eFG% among players using a normal quantile plot. This plot essentially shows how normally distributed a set of values is; if the values fit reasonably well and make a line, we can assume that the data is normally distributed.

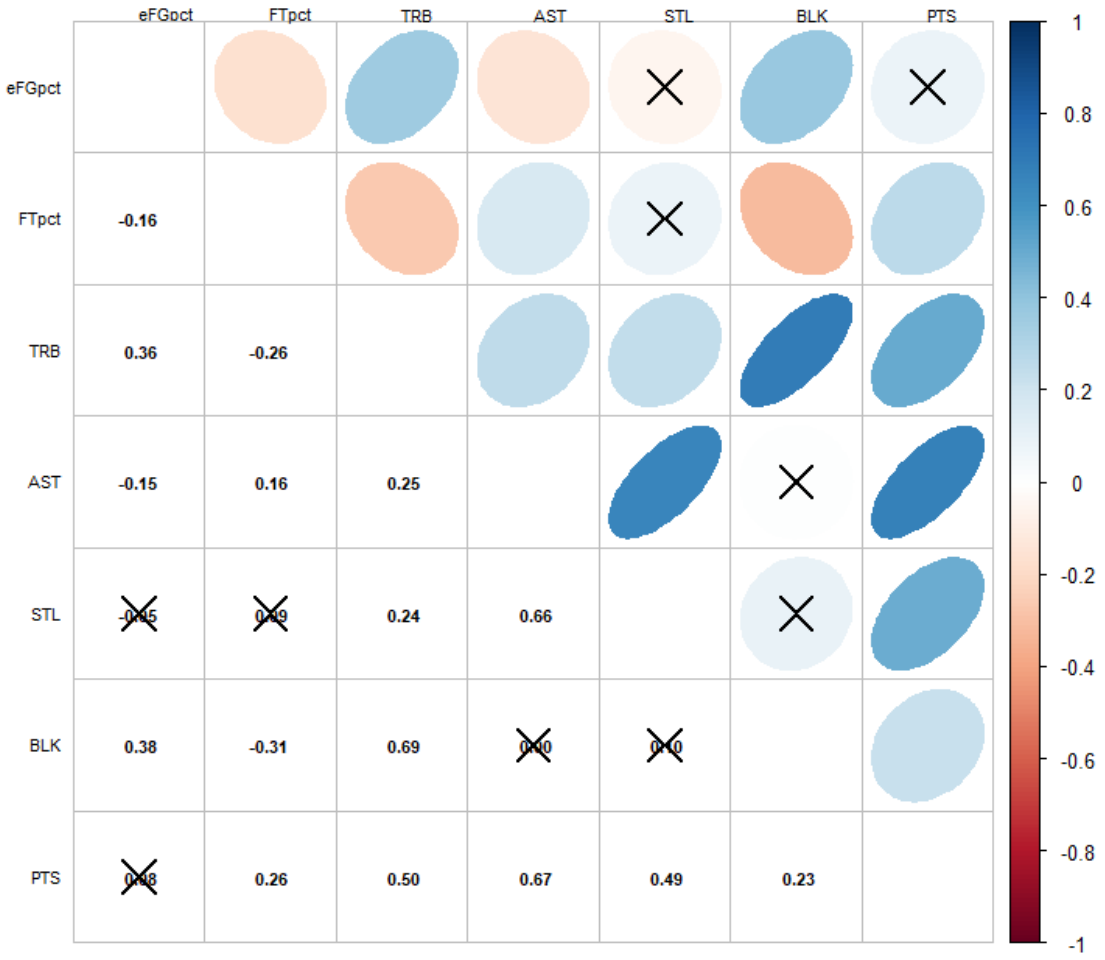



```
## [1] 86 79
```

As we can see, eFG% is approximately normally distributed; although it is not a requirement that the response variable is normally distributed for this analysis, it is a good sign that we may be able to avoid issues like heteroskedasticity (when the variances of the errors are unequal) or otherwise problematic occurences going forward.

Next, we can analyze the pairwise relationships between our continuous predictors using correlation charts.

```
# Creates a data frame with only the continous predictors
nbaCont <- nba[, c("eFGpct", "FTpct", "TRB", "AST", "STL", "BLK", "PTS")]

# Finds the pairwise correlations of the continuous predictors
sigcorr <- cor.mtest(nbaCont)

# Plots the correlations
corrplot.mixed(cor(nbaCont), lower.col = "black", upper = "ellipse", tl.col =
"black", number.cex=.7,
               tl.pos = "lt", tl.cex=.7, p.mat = sigcorr$p, sig.level = .05,
srt = 0)
```



As depicted, some of the predictors are related to each other. This indicates multicollinearity, which is essentially when predictors have substantial levels of correlation with each other and, in turn, "battle" to explain the variability in the response variable. This is nice to be aware of before doing our regression because it explains why the significance of some predictors may change when others are removed as we work towards the ideal model.

Now, we can fit a generalized linear model to predict eFG% using ANOVA (Analysis of Variance) and Type III Sum of Squares. This ensures that the significance of each predictor is calculated within the context of all the other terms in the model. The results are as follows:

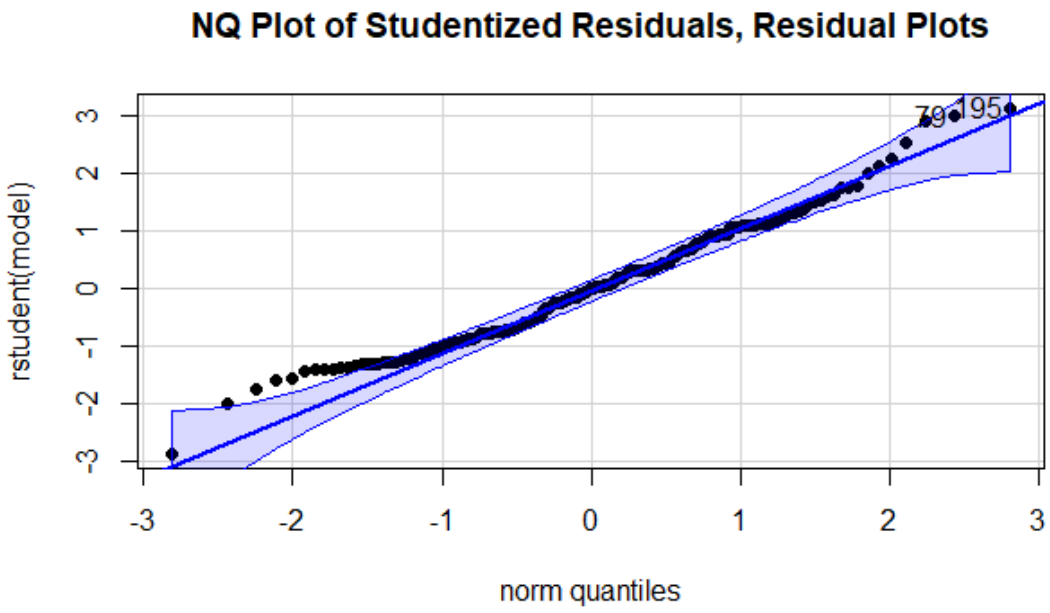
```
## Anova Table (Type III tests)
##
## Response: eFGpct
##              Sum Sq  Df  F value    Pr(>F)
## (Intercept) 0.46922   1 229.6229 < 2.2e-16 ***
## Pos         0.00032   1   0.1566  0.692722
```

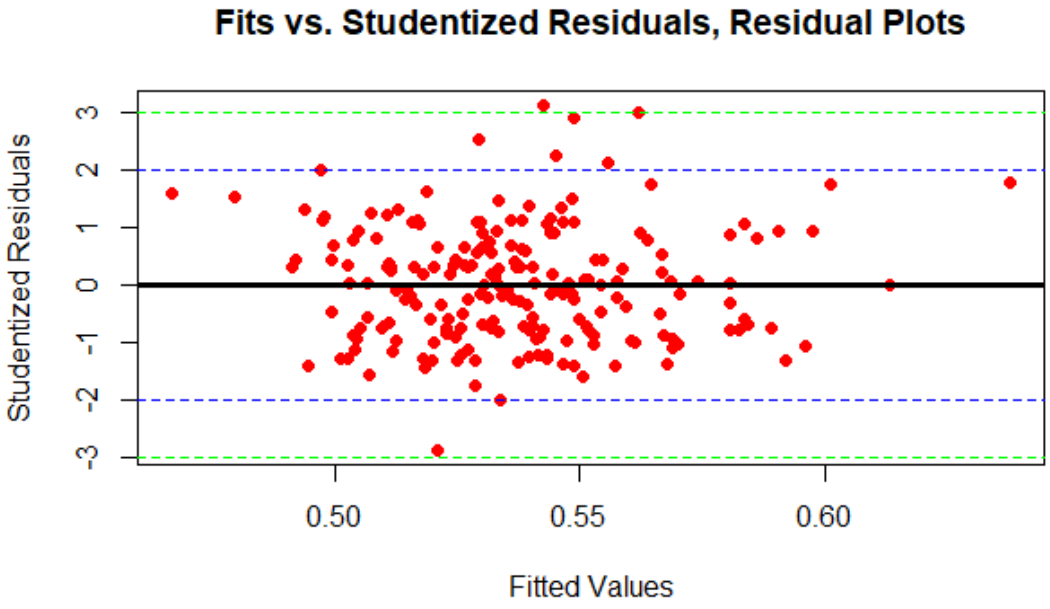
```
## Age         0.03035   3   4.9502  0.002479 **
## FTpct       0.00149   1   0.7273  0.394827
## TRB         0.00574   1   2.8113  0.095239 .
## AST         0.01230   1   6.0201  0.015040 *
## STL         0.00006   1   0.0312  0.859974
## BLK         0.00569   1   2.7856  0.096751 .
## PTS         0.00394   1   1.9267  0.166742
## Residuals   0.39029 191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the coefficients section, stars by the probability column indicate statistical significance. As shown, some of the variables are not statistically significant predictors of eFG% with this model. To fix this problem, we can do backwards stepwise regression, which essentially involves removing the predictor with the highest p-value one by one until your model only has statistically significant predictors left. The results with only statistically significant predictors are as follows:

```
## Anova Table (Type III tests)
##
## Response: eFGpct
##             Sum Sq  Df   F value    Pr(>F)
## (Intercept) 6.0312   1 2938.1572 < 2.2e-16 ***
## Age         0.0339   3    5.5102 0.0011829 **
## TRB         0.0803   1   39.1013 2.469e-09 ***
## AST         0.0296   1   14.4153 0.0001954 ***
## Residuals   0.4023 196
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As shown, after backwards stepwise regression we get a model for eFG% as predicted by a player's age, total rebounds per game, and assists per game. Now, we must check the residual plots of the model to ensure that the assumptions for the regression are met. The main assumptions are that the errors/residuals are normally distributed, the variances of the errors are equal across the board, and that there aren't outliers that are heavily affecting the data. Here are the plots:





According to these plots, the errors are approximately normally distributed, and the fits vs. studentized residual plot does not indicate that there is heteroskedasticity or that there are enough outliers (values with absolute values of their studentized residuals that are more than 3) to be concerned with. Therefore, we can continue with our analysis and look at more summary statistics of the model.

```
##
## Call:
## lm(formula = eFGpct ~ Age + TRB + AST)
```

```
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.126947 -0.034748 -0.000347  0.030301  0.136400
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.494321   0.009120  54.205  < 2e-16 ***
## Age23-27     0.024171   0.008132   2.973 0.003324 **
## Age28-32     0.023935   0.009294   2.575 0.010754 *
## Age33-37     0.042111   0.011363   3.706 0.000274 ***
## TRB          0.008639   0.001382   6.253 2.47e-09 ***
## AST         -0.006548   0.001725  -3.797 0.000195 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04531 on 196 degrees of freedom
## Multiple R-squared:  0.253,  Adjusted R-squared:  0.2339
## F-statistic: 13.27 on 5 and 196 DF,  p-value: 3.822e-11
```

These summary statistics tell us that the R-squared value for our model is .253. The R-squared value is, in essence, a metric estimating the overall predictive power of a model. The higher the value, the better a model is at predicting the response variable. It can also be interpreted as the percentage of the variance in the response variable that can be explained by the predictors of the model. Thus, in this case, roughly 26% of the variance in eFG% across the league can be explained by a player's age, total rebounds per game, and assists per game. Furthermore, the signs of the coefficients tell us about their individual relationships with eFG%. In this case, each age group along with total rebounds all have positive coefficients which suggests that as a player's age or total rebounding per game increases, their eFG% increases as well. However, assists has a negative coefficient which suggests that the more assists per game a player averages, the worse their eFG% is.

Part VII: Conclusion

In this project, I used data from the 2021-22 NBA season to conduct an analysis that focused on the relationships between eFG% and various other statistics in order to ultimately create a model that can predict eFG% based on a series of other metrics. Throughout the experiment, a few intermediate conclusions were drawn as well. Using T-Tests and bootstrapping, I discovered that, on average, there is a statistically significant relationship between eFG% and a player's position. Furthermore, I created a model to predict eFG% based on the statistically significant predictors of player age, total rebounds per game, and assists per game. A more extensive data set with a larger catalog of statistics may be able to provide deeper insight as to what metrics can be used to most accurately predict and quantify a player's offensive efficiency. With that being said, this experiment is a solid building block for further discussion of the significance of statistics in today's basketball culture.