883 Final Project - NBA

Kaitlyn Arnold, James Cuozzo, Greg Gibson, Christian Matye, and Erika Brittingham

4/13/2023

Data Import

We have NBA data from the following seasons: 17-18, 18-19, 19-20, 20-21, 21-22, 22-23. We were able to get the data by doing some web scraping on the NBA website. We then edited the csv file to add a few categorical variables such as conference and playoff outcomes.

```
nba_17_18 <- read.csv("NBA_17_18.csv",header=TRUE)
nba_18_19 <- read.csv("NBA_18_19.csv",header=TRUE)
nba_19_20 <- read.csv("NBA_18_19.csv",header=TRUE)
nba_20_21 <- read.csv("NBA_20_21.csv",header=TRUE)
nba_21_22 <- read.csv("NBA_21_22.csv",header=TRUE)
nba_22_23 <- read.csv("NBA_22_23.csv",header=TRUE)
```

Data Cleaning

```
nba_17_18$Playoffs_dummy <- ifelse(!is.na(nba_17_18$Playoffs), 1, 0)
nba_18_19$Playoffs_dummy <- ifelse(!is.na(nba_18_19$Playoffs), 1, 0)
nba_20_21$Playoffs_dummy <- ifelse(!is.na(nba_20_21$Playoffs), 1, 0)
nba_21_22$Playoffs_dummy <- ifelse(!is.na(nba_21_22$Playoffs), 1, 0)
nba_22_23$Playoffs_dummy <- ifelse(!is.na(nba_22_23$Playoffs), 1, 0)
```

We created a dummy variable to take the playoff results and indicate if the team made the playoffs or not, regardless of their results in the playoffs.

```
nba_17_23 <- rbind(nba_17_18, nba_18_19, nba_20_21, nba_21_22, nba_22_23)
nba_17_23 <- nba_17_23[, (colnames(nba_17_23)) %in% c('Season', 'TEAM_NAME', 'Conference', 'Division',
```

We combined the data from each season to create one data frame with all the data. We did not include the 2019-2020 season when joining the data frames due to this being the covid year. That year the NBA was sent into a bubble format after the season was put on pause.
We also removed the ranks for all of the categories. For all of the variables, the data has the actual variable and then the rank of the variable. Since we have the true variables, we do not need the rank because we can just sort the variables if need be.

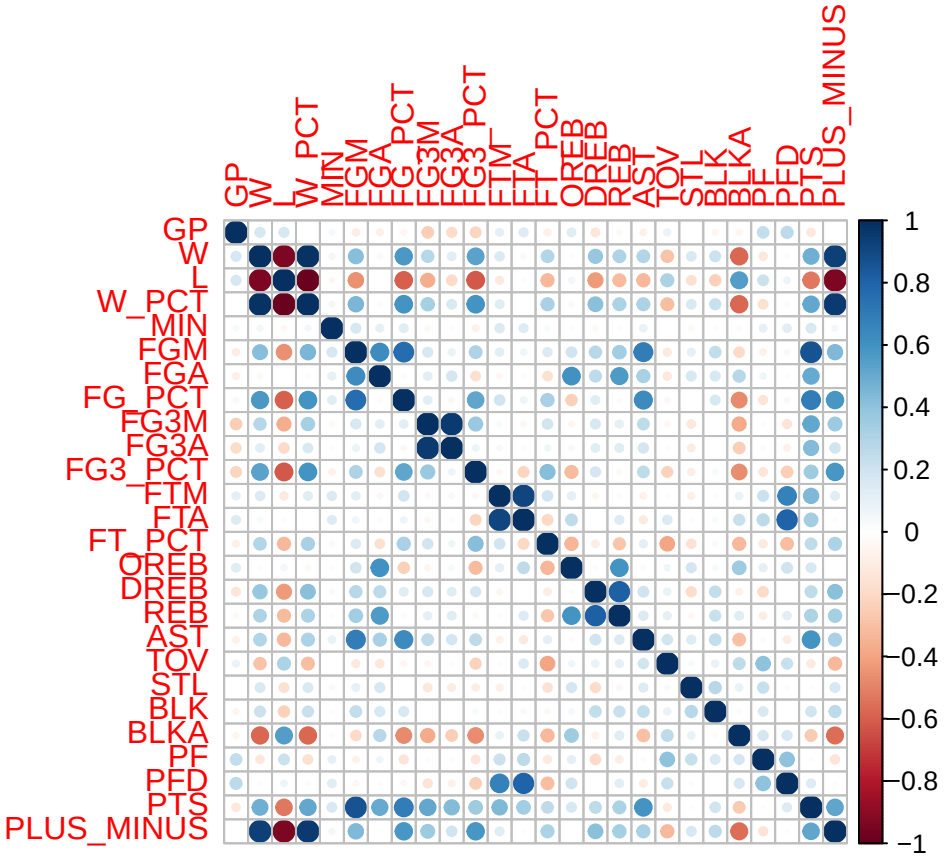
```
head(nba_17_23)
```

```
##   Season           TEAM_NAME Conference  Division Playoffs GP  W  L W_PCT
## 1  17-18       Atlanta Hawks    Eastern Southeast     <NA> 82 24 58 0.293
## 2  17-18      Boston Celtics    Eastern  Atlantic       CF 82 55 27 0.671
## 3  17-18       Brooklyn Nets    Eastern  Atlantic     <NA> 82 28 54 0.341
## 4  17-18   Charlotte Hornets    Eastern Southeast     <NA> 82 36 46 0.439
## 5  17-18       Chicago Bulls    Eastern   Central     <NA> 82 27 55 0.329
## 6  17-18 Cleveland Cavaliers    Eastern   Central      2nd 82 50 32 0.610
##    MIN  FGM  FGA FG_PCT FG3M FG3A FG3_PCT  FTM  FTA FT_PCT OREB DREB  REB
## 1 48.1 38.2 85.5  0.446 11.2 31.0   0.360 15.8 20.2  0.785  9.1 32.8 41.9
## 2 48.3 38.3 85.1  0.450 11.5 30.4   0.377 16.0 20.7  0.771  9.4 35.1 44.5
## 3 48.4 38.2 86.8  0.441 12.7 35.7   0.356 17.4 22.6  0.772  9.7 34.8 44.4
## 4 48.2 39.0 86.7  0.450 10.0 27.2   0.369 20.2 27.0  0.747 10.1 35.4 45.5
## 5 48.4 38.7 88.8  0.435 11.0 31.1   0.355 14.6 19.2  0.759  9.6 35.0 44.7
```

```
## 6 48.1 40.4 84.8  0.476 12.0 32.1   0.372 18.1 23.3  0.779  8.5 33.7 42.1
##    AST  TOV STL BLK BLKA   PF  PFD   PTS PLUS_MINUS Playoffs_dummy
## 1 23.7 15.5 7.8 4.2  5.5 19.6 20.3 103.4       -5.5              0
## 2 22.5 14.0 7.4 4.5  4.4 19.7 19.2 104.0        3.6              1
## 3 23.7 15.2 6.2 4.8  5.5 20.6 19.7 106.6       -3.7              0
## 4 21.6 12.7 6.8 4.5  4.9 17.2 22.4 108.2        0.3              0
## 5 23.5 14.0 7.6 3.5  5.2 19.2 17.4 102.9       -7.0              0
## 6 23.4 13.7 7.1 3.8  4.1 18.6 20.7 110.9        0.9              1
```

Data Exploration

Correlation



We can see that there are a hand full of variables that have a high correlation between others while there are many others that have very little to no correlation between each other.

```
sort(cor(nba_17_23[, !(colnames(nba_17_23)) %in% c('Season', 'TEAM_NAME', 'Conference', 'Division', 'Pla
```

```
##             L          BLKA           TOV            PF          OREB
## -0.9827662282 -0.5771070831 -0.3087852825 -0.1663038854 -0.0364089385
##           PFD           FGA            GP           FTA           MIN
## -0.0259336629 -0.0225247628 -0.0001938075  0.0079056508  0.0519362737
##           FTM           STL          FG3A           BLK           REB
##  0.1363405269  0.1642337341  0.1645716782  0.2234245158  0.3125273177
##           AST        FT_PCT          FG3M          DREB           FGM
##  0.3151931719  0.3201495106  0.3330126071  0.4123630207  0.4501242348
##           PTS       FG3_PCT        FG_PCT    PLUS_MINUS             W
##  0.5134947297  0.5957705498  0.5967180591  0.9554754552  0.9826970369
```

```
##         W_PCT
##  1.0000000000
```

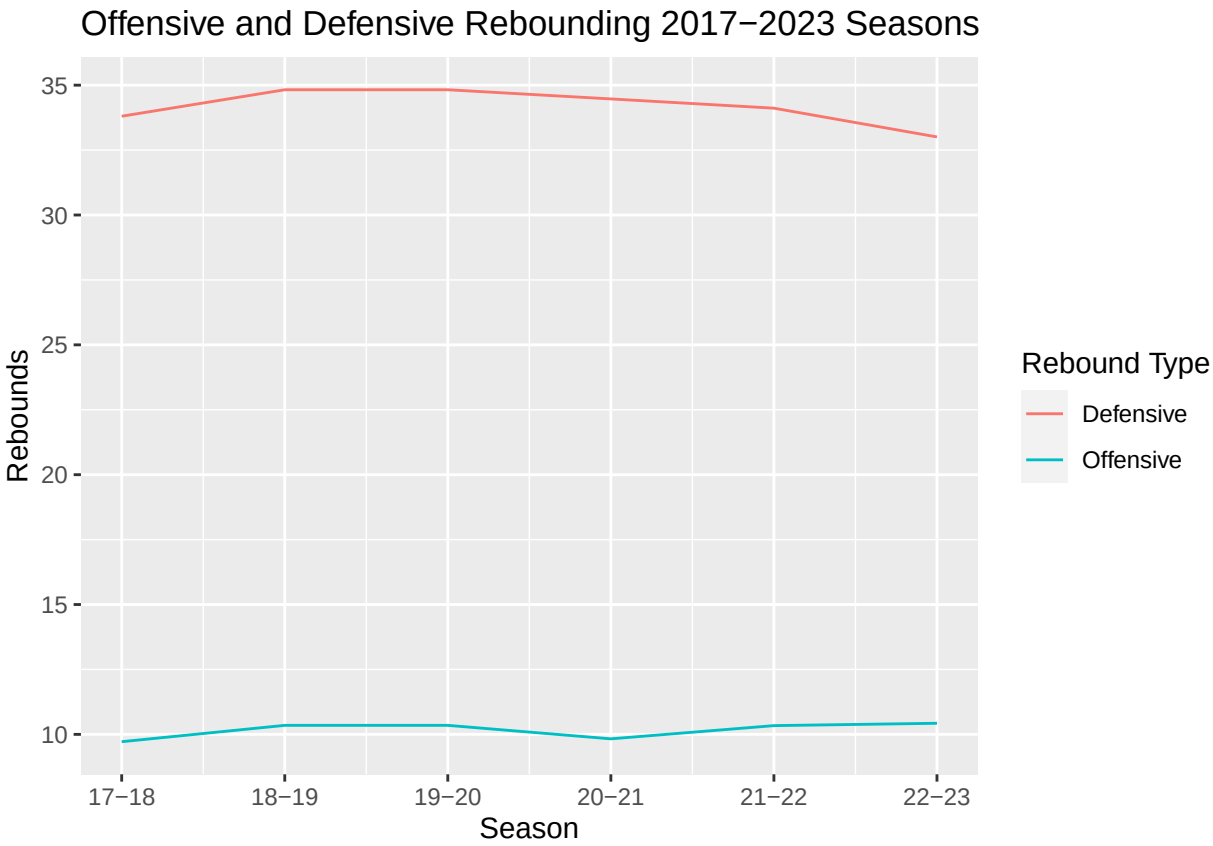
Based on the correlation matrix, the predictors that have the highest correlation to win percentage (positive or negative) are PLUS_MINUS, FG_PCT, FG3_PCT, PTS, FGM, DREB, REB, FT_PCT, FG3M, AST, BLKA, and TOV.

```
sort(cor(nba_17_23[, !(colnames(nba_17_23)) %in% c('Season', 'TEAM_NAME', 'Conference', 'Division', 'Pla
```

```
##           L        BLKA          GP         TOV         STL          PF
## -0.52414615 -0.25871692 -0.12181232 -0.11022433  0.01015746  0.01398089
##         PFD         MIN        OREB         BLK      FT_PCT        DREB
##  0.14239219  0.16972519  0.17237441  0.19762656  0.25193532  0.25970384
##         REB         FTA     FG3_PCT        FG3A         FTM           W
##  0.31047697  0.34995628  0.35333077  0.43903697  0.44783948  0.48065793
##         FGA       W_PCT        FG3M  PLUS_MINUS         AST      FG_PCT
##  0.50199511  0.51349473  0.51989232  0.52443005  0.59712407  0.69194380
##         FGM         PTS
##  0.86248188  1.00000000
```

The predictors with highest correlation to points scored are FGM, FG_PCT, PLUS_MINUS, AST, W_PCT, FGA, FG3M, REB, DREB, FG3A, FTM, FG3_PCT, and FTA.

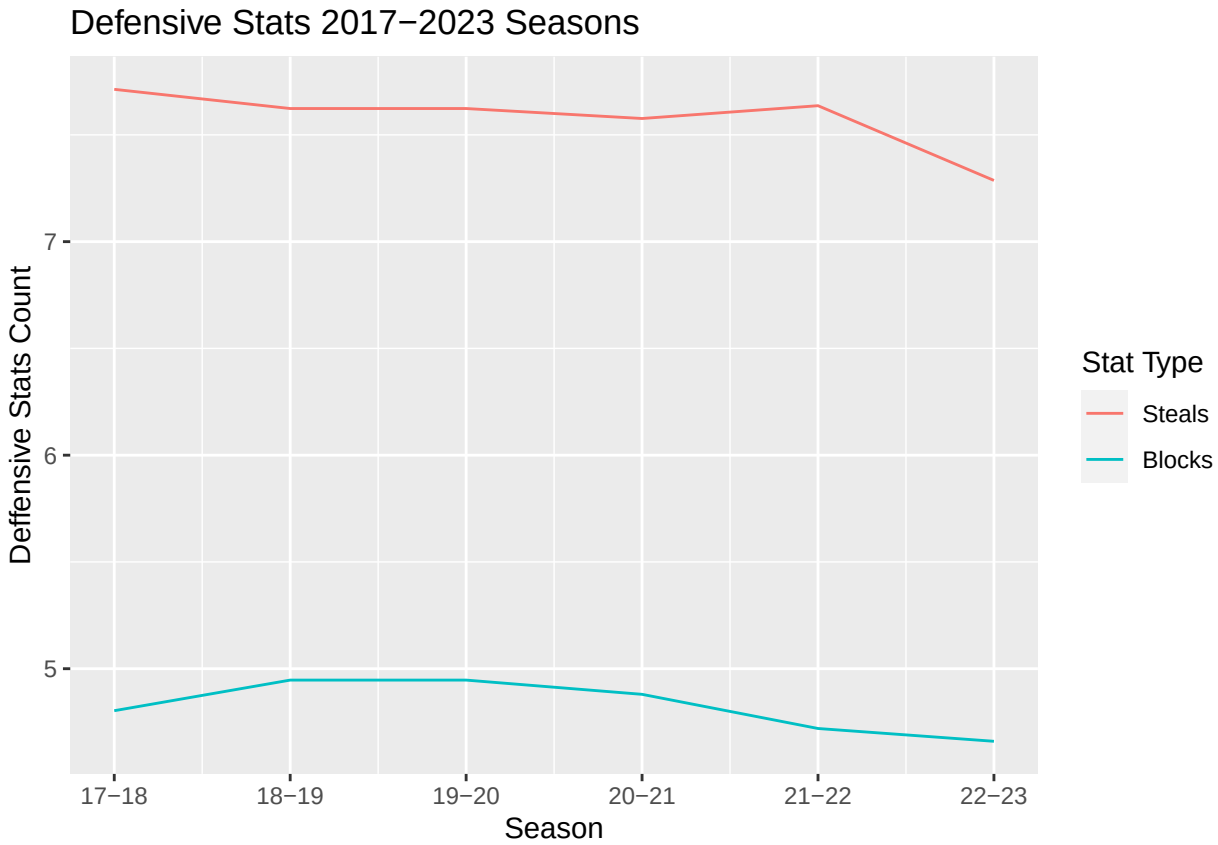
Rebounding



This graph shows us the trend of offensive rebounding vs defensive rebounding across the seasons. Higher values for offensive rebounding can show more of a defensive push while still on offense. Low offensive rebounding would mean not really caring much after the shot went up. This value remained fairly consistent across all 6 seasons. Defensive rebound has shown a decrease on the past few seasons. At the same time

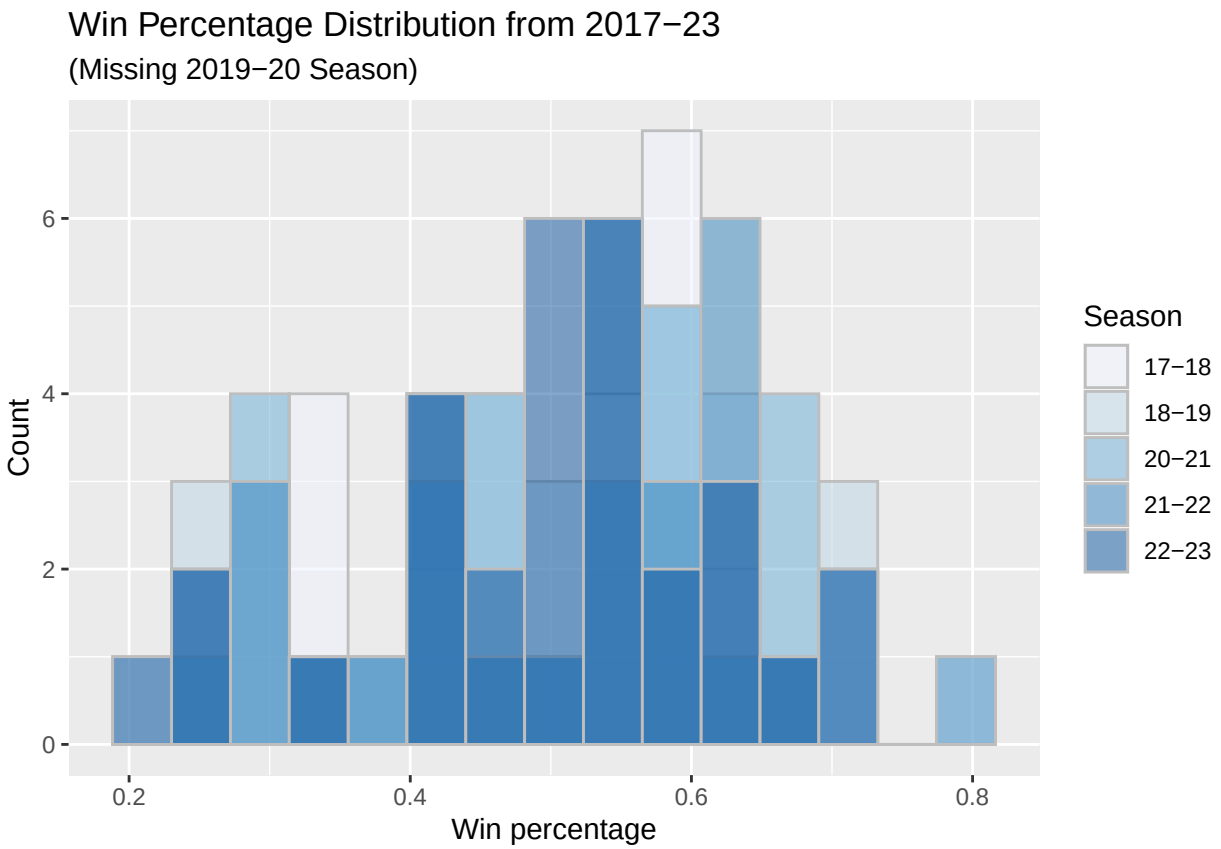
though there isn't much of an increase in offensive rebounds. This could mean that fewer shots in general are being taken or that players are making more of their shots.

Defensive Stats

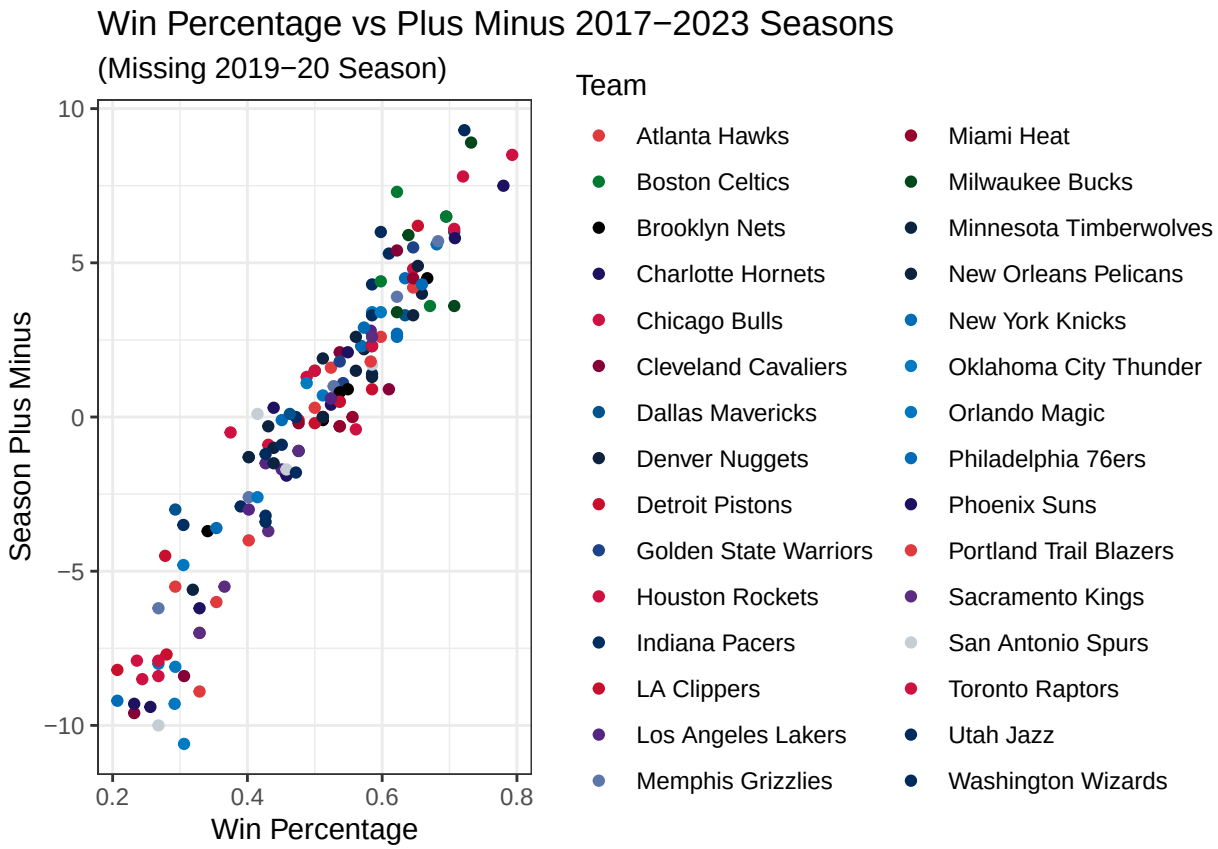


The NBA is usually known for not playing much defense. This graph helps to show that there has been a decrease in defensive plays in recent years. Steals and blocks are the two biggest defensive statistics (besides rebounds), and we can see that the number of steals and blocks in recent years has been on a general decline.

Win Percentage



This helps us to look at the overall win percentage distribution along with the distribution per season. Most seasons, most teams tend to fall between 50% and 70% with a few outliers on each side. There are a few years where the distribution is more condensed and some where it is more spread out. You can also see more teams having below average win percentages than above average. This indicates that there is usually a more distinct gap between the bad teams and average teams than the average teams and good teams. This is due to many teams tanking at the end of the season when they know they are out of the playoffs with the hopes of getting a better draft pick. This is part of the reason the play in part of the playoffs was recently added, to make it so less teams were tanking towards the end of the season.



There are no huge outliers here on this graph. It was expected that win percentage would highly correlate to plus minus on the season. We were more interested in looking at this graph to see if there were many teams that had a higher or lower plus minus than it appeared they should. IE, teams that had a higher plus minus than teams with higher win percentages than them. This may indicate that while they didn't win as many games, they may have won their games by a higher margin or some of their losses may have been very close. Same can go for teams that get blown out in losses, and have some close wins; they may have relatively low plus minuses in comparison to others with similar win percentages.

Models

Model 1- Predict Win Percentage

The goal of this model is model to be able to predict a team win percentage based on various different stats. For this model, we start by splitting the data into training and testing data. We are using the data from 2017-2021 for training and the seasons from 2021-2023 to test the data.

```
# Split train and test by season - 60% / 3 seasons train
data_sub <- nba_17_23[, !(colnames(nba_17_23)) %in% c('TEAM_NAME', 'Division', 'Playoffs', 'W', 'L')]
train_data_1721 <- subset(data_sub, Season != '22-23' & Season != '21-22')
test_data_2123 <- subset(data_sub, Season == '22-23' | Season == '21-22')
```

We then selected the stats with the higher correlation values to win percentage along with what conference the team was in try and create a model to predict win percentage.

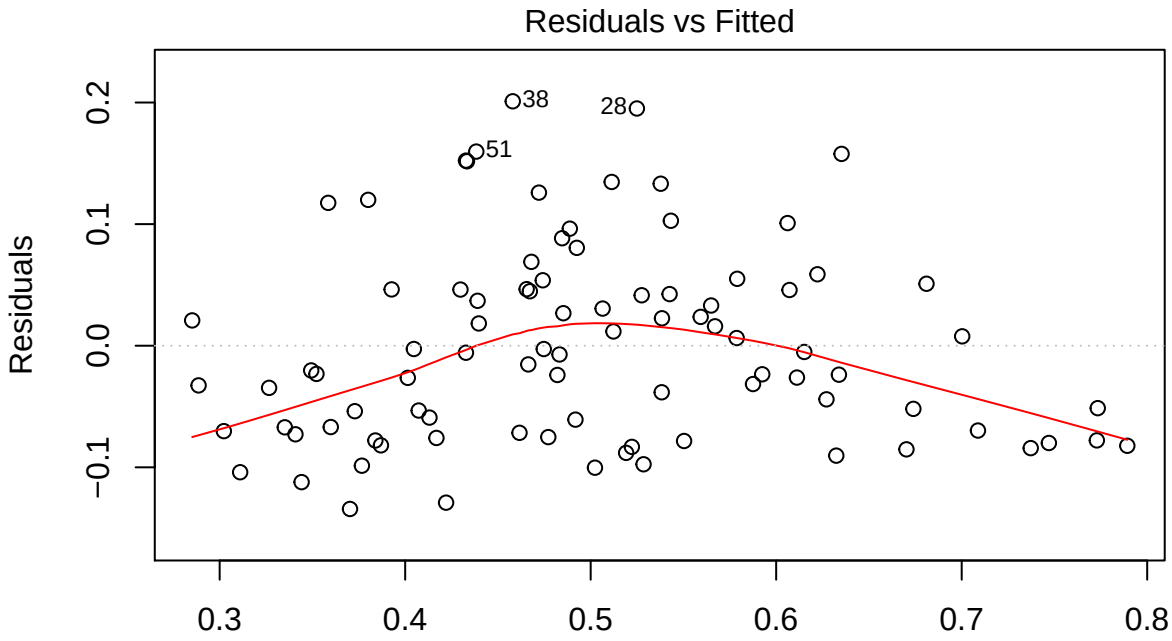
```
# Cols w/ high corr values and conference
model1 <- lm(W_PCT ~ Conference+FGM+FG_PCT+FG3_PCT+DREB+BLKA+PTS, data=train_data_1721)
summary(model1)
```

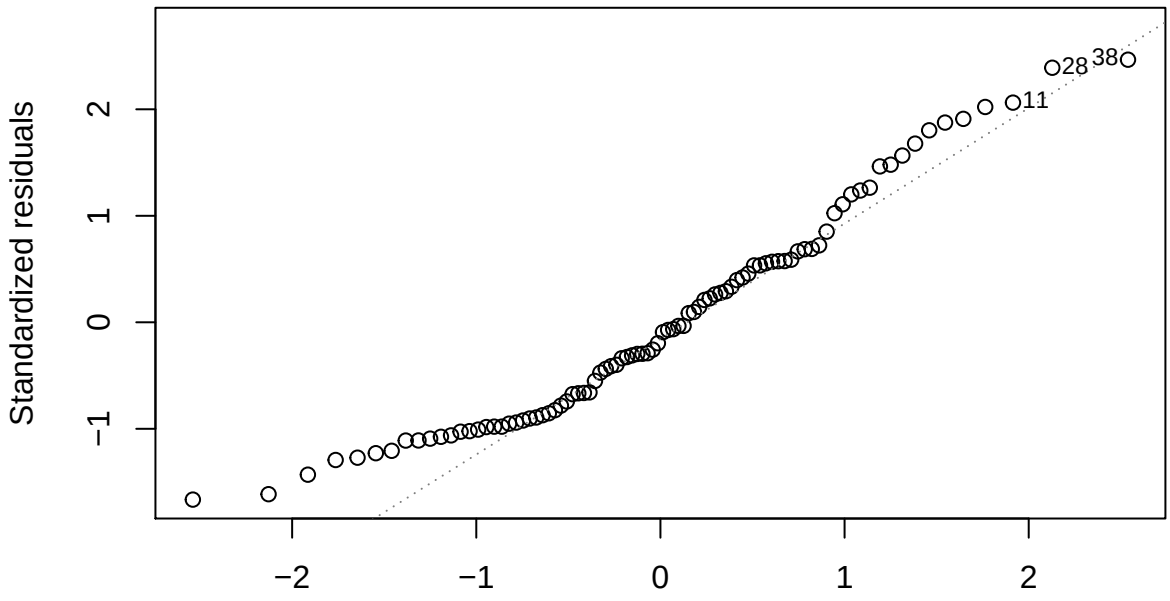
```
##
## Call:
```

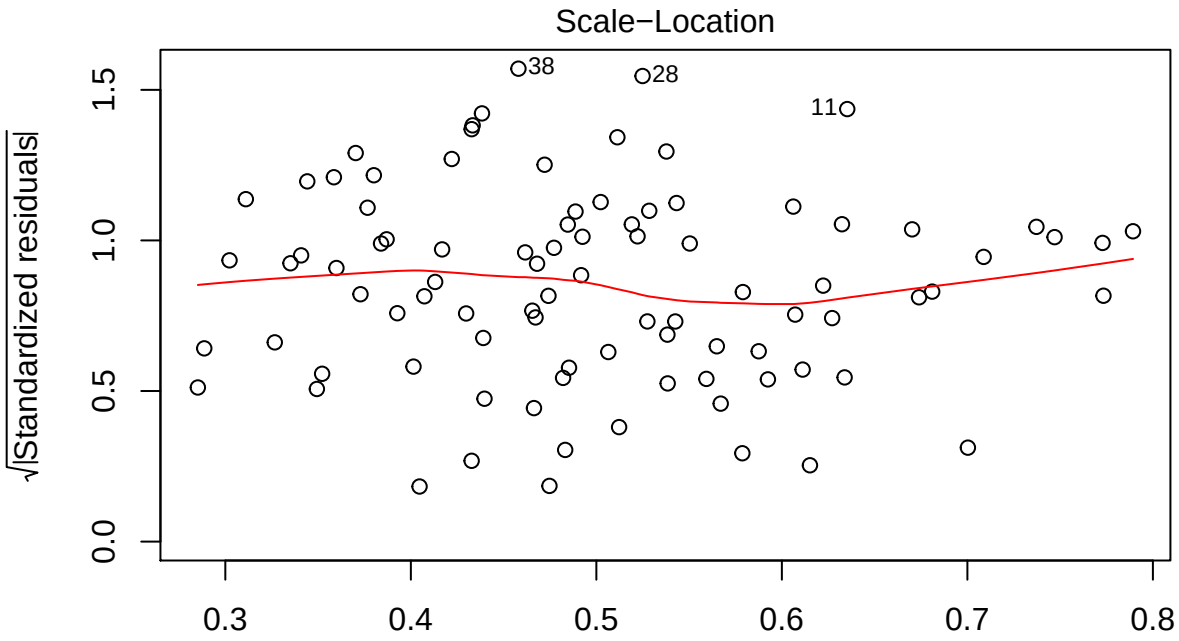
```
## lm(formula = W_PCT ~ Conference + FGM + FG_PCT + FG3_PCT + DREB +
##     BLKA + PTS, data = train_data_1721)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13426 -0.07013 -0.01128  0.04627  0.20101
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.398009   0.421557  -5.688 1.92e-07 ***
## ConferenceWestern  0.006281   0.018251   0.344 0.731633
## FGM               -0.037936   0.012894  -2.942 0.004235 **
## FG_PCT             4.252861   1.203410   3.534 0.000676 ***
## FG3_PCT            1.797004   0.699048   2.571 0.011958 *
## DREB               0.022808   0.006636   3.437 0.000925 ***
## BLKA              -0.067449   0.018651  -3.616 0.000515 ***
## PTS                0.012439   0.003816   3.260 0.001622 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08332 on 82 degrees of freedom
## Multiple R-squared:  0.6888, Adjusted R-squared:  0.6623
## F-statistic: 25.93 on 7 and 82 DF,  p-value: < 2.2e-16
```

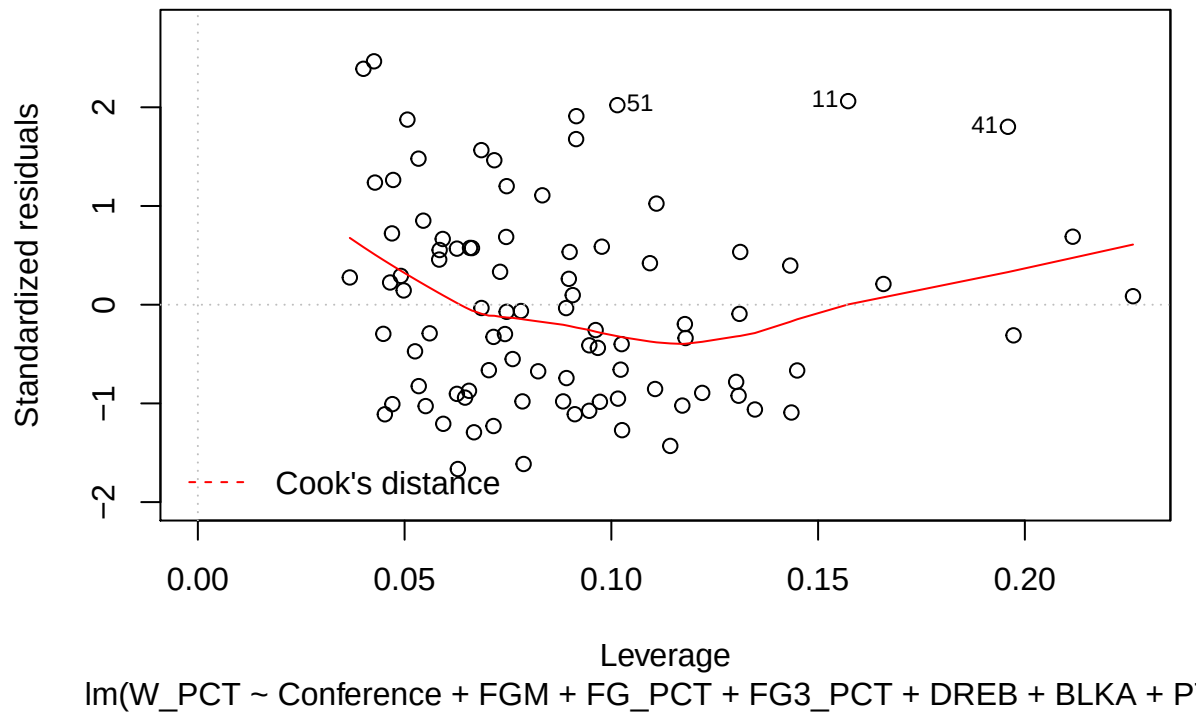
For the above model, we are able predict a teams win percentage based off of the conference they are in, teams average field goals made, teams field goal percent, teams 3 point percentage, how many defensive rebounds they have, how many attempts are made to block that teams shots, and how many points the team scores on average. The only one of these variables that is not significant is the teams conference. If we control for all other team statistics, there is not a statistically significant difference in win percentage when comparing a team in the western conference to a team in the eastern conference. When field goal percent, 3 point percent, defensive rebounding, and points per game increase, this leads to an increase in a teams season win percentage. When field goals made and block attempts against the team increase, this leads to a decrease in the teams season win percentage. Field goal percent has the largest effect on win percentage. When we control for the conference and every other stat, when field goal percent increases by 1%, the win percentage of the team increases by 4.25%. In addition, the Adjusted R-squared value is fairly high, so we know our model can explain 66% of the variability of win percentage from this set of variables.

```
plot(model1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(W_PCT ~ Conference + FGM + FG_PCT + FG3_PCT + DREB + BLKA + PTS)
38
28
51


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(W_PCT ~ Conference + FGM + FG_PCT + FG3_PCT + DREB + BLKA + PTS)
28
38
11






```
pred <- predict(model1,test_data_2123)
error <- test_data_2123$W_PCT - pred
MSE <- mean(error^2)
RMSE <- sqrt(MSE)
MAE <- mean(abs(error))
MAPE <- mean(abs(error/test_data_2123$W_PCT))*100
```

```
print("MSE          RMSE         MAE         MAPE")
```

```
## [1] "MSE          RMSE         MAE         MAPE"
```

```
c(MSE,RMSE,MAE,MAPE)
```

```
## [1]  0.009421229  0.097063018  0.073377855 16.596979334
```

The MSE and RMSE for our testing data is very close to 0. This means that our model fits the data very well.

```
win_percent <- test_data_2123 %>% select('W_PCT')
win_percent$pred <- predict(model1,test_data_2123)
win_percent$error <- win_percent$W_PCT - win_percent$pred
win_percent$less_than_1 <- ifelse(abs(win_percent$error) < 0.01, 1, 0)
win_percent$less_than_5 <- ifelse(abs(win_percent$error) < 0.05, 1, 0)
head(win_percent)
```

```
##    W_PCT      pred        error less_than_1 less_than_5
## 91 0.524 0.5850698 -0.061069754           0           0
## 92 0.622 0.5696871  0.052312852           0           0
## 93 0.537 0.5291479  0.007852142           1           1
## 94 0.524 0.5171823  0.006817662           1           1
## 95 0.561 0.5441299  0.016870091           0           1
## 96 0.537 0.5346157  0.002384280           1           1
```

```
## [1] "The number of teams with a win percentage error less than 1%: 9"
```

```
## [1] "The number of teams with a win percentage error less than 5%: 26"
```

```
## [1] "Max percentage error: 29.8472230843676"
```

```
## [1] "Max number of games error: 24.4747229291814"
```

```
## [1] "Mean percentage error: 7.33778546698325"
```

```
## [1] "Mean number of games error: 6.01698408292626"
```

26 out of 60 teams had a predicted win percentage that was less than 5% off from their true win percentage and 9 of those teams had a predicted win percentage that was less than 1% off from the actual. The highest error was about 30%, which for a 82 game season is about 24 games off. However, the average error for our test data is about 7%, which for a 82 game season is about 6 games off, which is not too bad. And this can be expected as there are many games throughout the season that are very close games and are sometimes only decided by a point or two, and can come down to luck in some situations. Unfortunately our model does not take luck into factor.

Model 2- Predict Winner of Game and Score

The goal of this model is to create a model that predicts the winner of a match up between two given teams. We combined the data from the 17-22 seasons to be the train set and the data from 22-23 season will be the test set.

```
NBA_train <- rbind(nba_17_18,nba_18_19 ,nba_20_21,nba_21_22)
NBA_train <- NBA_train[, (colnames(NBA_train)) %in% c('Conference', 'Division', 'W', 'L', 'W_PCT', 'MIN
```

We take the predictors that have a high correlation to win percentage and points scored and create a model with them. The mutual predictors are PLUS_MINUS, FG_PCT, FG3_PCT, FGM, FG3M, DREB, REB, FG3M, and AST.

```
Model <- glm(PTS ~ PLUS_MINUS + FG_PCT + FG3_PCT + FGM + FG3M + DREB + REB + FG3M + AST , data = NBA_tra
summary(Model)
```

```
##
## Call:
## glm(formula = PTS ~ PLUS_MINUS + FG_PCT + FG3_PCT + FGM + FG3M +
##     DREB + REB + FG3M + AST, data = NBA_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7335  -0.8891  -0.1110   0.8275   5.7497
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.38792    9.88682   0.242  0.80959
## PLUS_MINUS    0.02160    0.05085   0.425  0.67176
## FG_PCT       62.08736   25.02870   2.481  0.01462 *
## FG3_PCT     -27.15160   11.74417  -2.312  0.02263 *
## FGM           1.83835    0.18038  10.192  < 2e-16 ***
## FG3M          1.09317    0.09402  11.628  < 2e-16 ***
## DREB         -0.17969    0.18946  -0.948  0.34497
## REB           0.31486    0.18075   1.742  0.08428 .
## AST          -0.27142    0.10015  -2.710  0.00779 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.072279)
##
##     Null deviance: 2334.87  on 119  degrees of freedom
## Residual deviance:  230.02  on 111  degrees of freedom
## AIC: 438.63
##
## Number of Fisher Scoring iterations: 2
```

```
Score_Predict <- function(model, Team1, Team2){
  Team1_index <- which(nba_22_23$TEAM_NAME == Team1)
  Team2_index <- which(nba_22_23$TEAM_NAME == Team2)

  Team1_score <- predict(Model, nba_22_23)[Team1_index]
  Team2_score <- predict(Model, nba_22_23)[Team2_index]

  if(round(Team1_score,0) > round(Team2_score,0)){
    winner <- Team1
    return(paste('Winner =',winner,'| Scores:',Team1,'-',round(Team1_score,0),',',Team2,'-',round(Team2
  } else if (round(Team1_score,0) < round(Team2_score,0)){
    winner <- Team2
    return(paste('Winner =',winner,'| Scores:',Team1,'-',round(Team1_score,0),',',Team2,'-',round(Team2
  } else {if (Team1_score > Team2_score){
      return(paste(Team1,'win in overtime'))
  } else {return(paste(Team2,'win in overtime'))}
}
}
```

```
Score_Predict(Model, 'Atlanta Hawks', 'Miami Heat')
```

```
## [1] "Winner = Atlanta Hawks | Scores: Atlanta Hawks - 118 , Miami Heat - 108"
```

```
Score_Predict(Model, 'Minnesota Timberwolves', 'Los Angeles Lakers')
```

```
## [1] "Minnesota Timberwolves win in overtime"
```

```
Score_Predict(Model, 'Oklahoma City Thunder', 'New Orleans Pelicans')
```

```
## [1] "Winner = Oklahoma City Thunder | Scores: Oklahoma City Thunder - 115 , New Orleans Pelicans - 1
```

```
Score_Predict(Model, 'Chicago Bulls', 'Toronto Raptors')
```

```
## [1] "Winner = Chicago Bulls | Scores: Chicago Bulls - 113 , Toronto Raptors - 112"
```

```
Score_Predict(Model, 'Philadelphia 76ers','Miami Heat')
```

```
## [1] "Winner = Philadelphia 76ers | Scores: Philadelphia 76ers - 111 , Miami Heat - 108"
```

```
Score_Predict(Model, 'Orlando Magic','Cleveland Cavaliers')
```

```
## [1] "Winner = Cleveland Cavaliers | Scores: Orlando Magic - 110 , Cleveland Cavaliers - 113"
```

```
Score_Predict(Model, 'San Antonio Spurs','Portland Trail Blazers')
```

```
## [1] "Winner = San Antonio Spurs | Scores: San Antonio Spurs - 114 , Portland Trail Blazers - 111"
```

```
Score_Predict(Model, 'Oklahoma City Thunder','Utah Jazz')
```

```
## [1] "Winner = Utah Jazz | Scores: Oklahoma City Thunder - 115 , Utah Jazz - 116"
```

```
Score_Predict(Model, 'Denver Nuggets','Phoenix Suns')
```

```
## [1] "Winner = Denver Nuggets | Scores: Denver Nuggets - 116 , Phoenix Suns - 113"
```

Based on this model, we can learn what factors are most important in predicting how many points a team will score in a game. These scores can be used to predict winners of games. Even though this model is not very accurate we can address some concerns of why it is not accurate.

1. Data from previous years may not be useful in training a model (Players moving teams/retiring, coaches moving, injuries, etc.)
2. Games tested on may not be true representations of how a team would normally play (Since the regular season just ended, teams that secured a playoff spot may have been resting star players, whereas teams fighting for a spot may be playing harder in order to win.)
3. Team stats update after every game, although it is possible for us to update our data after every game, it would be very laborious with minimal changes to the actual data.

Even though this model is not very accurate, it is very interpretable. The prediction returns a score for each team based on their stats at this point in the season. Like I addressed above, the predictions will not update unless we update the data from the NBA website. If I were to go back and have more time to improve the model, that would be something I would most likely add.

Model 3- Predict number of wins

The goal of this model is to predict how many wins a team will have in an individual season.
We first created a function so we can repeat our findings across many different seasons. In this function we train the model using 6 different stats- field goals made, field goal attempts, free throw attempts, three point percentage, and plus minus. After training the model we predict how many wins each team will have and then return the predicted and actual values.

```
predict_wins <- function(nba) {
  # Create linear regression model
  lm_model <- lm(W ~ FGM + FGA + FTM + FTA + FG3_PCT + PLUS_MINUS, data = nba)
```

```
  # Predict wins using linear regression
  nba$predicted_wins_lm <- predict(lm_model, newdata = nba)
  # Return data frame with predicted wins for each team
  return_val <- nba %>% select(TEAM_NAME, predicted_wins_lm, W)
  return_val$Difference <- round(return_val$predicted_wins_lm,0) - return_val$W
  return_val$Under3 <- ifelse(abs(return_val$Difference) <= 3, 1,0)
  return_val$Under5 <- ifelse(abs(return_val$Difference) <= 5, 1,0)
  return(return_val)
}
```

```
lm_model <- lm(W ~ FGM + FGA + FTM + FTA + FG3_PCT + PLUS_MINUS, data = nba_17_18)
summary(lm_model)
```

```
##
## Call:
## lm(formula = W ~ FGM + FGA + FTM + FTA + FG3_PCT + PLUS_MINUS,
##     data = nba_17_18)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.013 -2.100 -0.400  2.547  5.810
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.539864  60.236199   0.192    0.850
## FGM          1.100539   0.705611   1.560    0.132
## FGA         -0.399574   0.631128  -0.633    0.533
## FTM          0.082605   1.640965   0.050    0.960
## FTA         -0.001184   1.191264  -0.001    0.999
## FG3_PCT     52.242060  80.394661   0.650    0.522
## PLUS_MINUS   2.364125   0.223817  10.563 2.69e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.731 on 23 degrees of freedom
## Multiple R-squared:  0.9261, Adjusted R-squared:  0.9068
## F-statistic: 48.05 on 6 and 23 DF,  p-value: 7.142e-12
```

```
nba1718_predicted_wins <- predict_wins(nba_17_18)
head(nba1718_predicted_wins)
```

```
##             TEAM_NAME predicted_wins_lm  W Difference Under3 Under5
## 1       Atlanta Hawks          26.50256 24          3      1      1
## 2      Boston Celtics          49.19002 55         -6      0      0
## 3       Brooklyn Nets          30.15890 28          2      1      1
## 4   Charlotte Hornets          41.44102 36          5      0      1
## 5       Chicago Bulls          21.82890 27         -5      0      1
## 6 Cleveland Cavaliers          45.14707 50         -5      0      1
```

```
## [1] "Teams with under +/- 3 game difference: 23"
```

```
## [1] "Teams with under +/- 5 game difference: 28"
```

```
lm_model <- lm(W ~ FGM + FGA + FTM + FTA + FG3_PCT + PLUS_MINUS, data = nba_18_19)
summary(lm_model)
```

```
##
```

```
## Call:
## lm(formula = W ~ FGM + FGA + FTM + FTA + FG3_PCT + PLUS_MINUS,
##     data = nba_18_19)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2205 -1.2182  0.1997  1.1932  4.7982
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.3890    23.5487  -0.144   0.8868
## FGM          -0.1358     0.4768  -0.285   0.7783
## FGA           0.1810     0.3259   0.555   0.5840
## FTM           0.6645     0.6571   1.011   0.3224
## FTA          -0.3026     0.5134  -0.589   0.5614
## FG3_PCT      81.6582    35.8256   2.279   0.0322 *
## PLUS_MINUS    2.2898     0.1097  20.871   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.241 on 23 degrees of freedom
## Multiple R-squared:  0.9725, Adjusted R-squared:  0.9653
## F-statistic: 135.5 on 6 and 23 DF,  p-value: < 2.2e-16
```

```
nba1819_predicted_wins <- predict_wins(nba_18_19)
head(nba1819_predicted_wins)
```

```
##             TEAM_NAME predicted_wins_lm  W Difference Under3 Under5
## 1       Atlanta Hawks          27.21895 29         -2      1      1
## 2      Boston Celtics          51.61496 49          3      1      1
## 3       Brooklyn Nets          40.87452 42         -1      1      1
## 4   Charlotte Hornets          38.78062 39          0      1      1
## 5       Chicago Bulls          21.04013 22         -1      1      1
## 6 Cleveland Cavaliers          18.81989 19          0      1      1
```

```
## [1] "Teams with under +/- 3 game difference: 27"
```

```
## [1] "Teams with under +/- 5 game difference: 30"
```

```
lm_model <- lm(W ~ FGM + FGA + FTM + FTA + FG3_PCT + PLUS_MINUS, data = nba_20_21)
summary(lm_model)
```

```
##
## Call:
## lm(formula = W ~ FGM + FGA + FTM + FTA + FG3_PCT + PLUS_MINUS,
##     data = nba_20_21)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7688 -1.8443  0.9102  1.4831  6.3558
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.0359    43.4860   1.358    0.188
## FGM           0.8375     0.7326   1.143    0.265
## FGA          -0.6480     0.4938  -1.312    0.202
```

```
## FTM           0.7595     1.4883   0.510    0.615
## FTA          -0.6333     1.0905  -0.581    0.567
## FG3_PCT       1.7804    73.8179   0.024    0.981
## PLUS_MINUS    1.6746     0.3102   5.399 1.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.511 on 23 degrees of freedom
## Multiple R-squared:  0.9025, Adjusted R-squared:  0.8771
## F-statistic: 35.49 on 6 and 23 DF,  p-value: 1.649e-10
```

```
nba2021_predicted_wins <- predict_wins(nba_20_21)
head(nba2021_predicted_wins)
```

```
##             TEAM_NAME predicted_wins_lm  W Difference Under3 Under5
## 1       Atlanta Hawks          40.85348 41          0      1      1
## 2      Boston Celtics          38.41930 36          2      1      1
## 3       Brooklyn Nets          46.29417 48         -2      1      1
## 4   Charlotte Hornets          31.87440 33         -1      1      1
## 5       Chicago Bulls          35.51681 31          5      0      1
## 6 Cleveland Cavaliers          20.79571 22         -1      1      1
```

```
mean(nba2021_predicted_wins$Diff)
```

```
## [1] -0.03333333
```

```
## [1] "Teams with under +/- 3 game difference: 23"
```

```
## [1] "Teams with under +/- 5 game difference: 27"
```

```
lm_model <- lm(W ~ FGM + FGA + FTM + FTA + FG3_PCT + PLUS_MINUS, data = nba_21_22)
summary(lm_model)
```

```
##
## Call:
## lm(formula = W ~ FGM + FGA + FTM + FTA + FG3_PCT + PLUS_MINUS,
##     data = nba_21_22)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1142 -1.3966 -0.0818  2.2432  5.5938
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.96430   46.25010   0.432    0.670
## FGM          0.17689    0.78130   0.226    0.823
## FGA         -0.08346    0.47972  -0.174    0.863
## FTM          1.15945    1.44023   0.805    0.429
## FTA         -1.13966    1.05400  -1.081    0.291
## FG3_PCT     74.90494   77.02165   0.973    0.341
## PLUS_MINUS   2.02794    0.22104   9.175 3.79e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.888 on 23 degrees of freedom
## Multiple R-squared:  0.9104, Adjusted R-squared:  0.887
## F-statistic: 38.94 on 6 and 23 DF,  p-value: 6.369e-11
```

```r
nba2122_predicted_wins <- predict_wins(nba_21_22)
head(nba2122_predicted_wins)
```

```
##             TEAM_NAME predicted_wins_lm  W Difference Under3 Under5
## 1       Atlanta Hawks          46.76702 43          4      0      1
## 2      Boston Celtics          57.23173 51          6      0      0
## 3       Brooklyn Nets          44.23913 44          0      1      1
## 4   Charlotte Hornets          41.98955 43         -1      1      1
## 5       Chicago Bulls          42.70488 46         -3      1      1
## 6 Cleveland Cavaliers          45.06882 44          1      1      1
```

```r
mean(nba2122_predicted_wins$Diff)
```

```
## [1] -0.03333333
```

```
## [1] "Teams with under +/- 3 game difference: 21"
```

```
## [1] "Teams with under +/- 5 game difference: 26"
```

```r
lm_model <- lm(W ~ FGM + FGA + FTM + FTA + FG3_PCT + PLUS_MINUS, data = nba_22_23)
summary(lm_model)
```

```
##
## Call:
## lm(formula = W ~ FGM + FGA + FTM + FTA + FG3_PCT + PLUS_MINUS,
##     data = nba_22_23)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3451 -2.3911  0.2609  1.2283  6.9768
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.36453   45.22651   0.561    0.580
## FGM          0.09246    0.77169   0.120    0.906
## FGA         -0.01228    0.40728  -0.030    0.976
## FTM         -1.16918    1.04916  -1.114    0.277
## FTA          0.67211    0.88159   0.762    0.454
## FG3_PCT     51.42910   57.98208   0.887    0.384
## PLUS_MINUS   2.29104    0.19855  11.539 4.81e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.159 on 23 degrees of freedom
## Multiple R-squared:  0.9212, Adjusted R-squared:  0.9007
## F-statistic: 44.84 on 6 and 23 DF,  p-value: 1.473e-11
```

```r
nba2223_predicted_wins <- predict_wins(nba_22_23)
head(nba2223_predicted_wins)
```

```
##             TEAM_NAME predicted_wins_lm  W Difference Under3 Under5
## 1       Atlanta Hawks          40.70404 41          0      1      1
## 2      Boston Celtics          56.51358 57          0      1      1
## 3       Brooklyn Nets          43.81810 45         -1      1      1
## 4   Charlotte Hornets          26.12465 27         -1      1      1
## 5       Chicago Bulls          43.84710 40          4      0      1
## 6 Cleveland Cavaliers          54.07276 51          3      1      1
```

```
## [1] "Teams with under +/- 3 game difference: 24"
```

```
## [1] "Teams with under +/- 5 game difference: 28"
```

For each year we calculated the difference in projected wins vs. actual wins. From there we looked at the difference and we grouped the teams by if the difference was +/- 3 and +/- 5. This model turned out to be very accurate. The lowest accuracy we had was in 21-22 with 21 teams having a difference of +/- 3 and 26 teams +/- 5. In 18-19 we were able to correctly predict all 30 teams within +/- 5 games.

Model 4- Predicting Playoffs

For this model we are trying to predict if a team will make the playoffs or not based on stats from throughout the season. We are omitting games won from the model because ultimately games won is what determines the playoff teams. We trained two different logistic models with different stats to see if one model would give us a better outcome then the other. We will use the same training data for method 1 and method 2. We are using the data from 2017-2022 to then predict the playoffs for the 2022-23 season.

```
train_data_1722 <- subset(nba_17_23, Season != '22-23')
```

Method 1

This model predicts if a team will make the playoffs based on six different statistics- field goals made, field goals attempted, free throws made, free throws attempted, three point percentage, and plus minus. These statistics are used to predict a yes or no outcome deciding whether or not a team will make the playoffs. For past seasons, a dummy variable was also added to determine if teams did end up making the playoffs regardless of the teams outcome in the playoffs. It was only used to test the accuracy of the model's predictions.

```
model_1722 <- glm(Playoffs_dummy~FGM + FGA + FTM + FTA + FG3_PCT + PLUS_MINUS, data = train_data_1722 ,
summary(model_1722)
```

```
##
## Call:
## glm(formula = Playoffs_dummy ~ FGM + FGA + FTM + FTA + FG3_PCT +
##     PLUS_MINUS, family = binomial(link = "logit"), data = train_data_1722)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.26277  -0.01662   0.00240   0.08389   2.15282
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  34.9657    25.3947   1.377  0.16855
## FGM           0.2629     0.3781   0.695  0.48686
## FGA          -0.1467     0.2541  -0.577  0.56380
## FTM           1.6020     0.7950   2.015  0.04391 *
## FTA          -1.2125     0.5908  -2.052  0.04014 *
## FG3_PCT     -91.2421    49.6777  -1.837  0.06626 .
## PLUS_MINUS    2.0066     0.6120   3.279  0.00104 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 164.216  on 119  degrees of freedom
## Residual deviance:  39.567  on 113  degrees of freedom
## AIC: 53.567
##
```

```
## Number of Fisher Scoring iterations: 9
```

```
probabilities <- model_1722 %>% predict(nba_22_23, type = "response")
nba_22_23$Predicted_Playoffs <- ifelse(probabilities > 0.5, 1, 0)
```

```
## [1] "The accuracy is: 0.966666666666667"
```

```
## [1] "Number of teams predicted to make the playoffs: 21"
```

```
## [1] "Number of teams to actually make the playoffs: 20"
```

```
confusion <- table(actual = nba_22_23$Playoffs_dummy, predicted = nba_22_23$Predicted_Playoffs)
confusion
```

```
##       predicted
## actual  0  1
##      0  9  1
##      1  0 20
```

```
nba_22_23[nba_22_23$Predicted_Playoffs != nba_22_23$Playoffs_dummy, ]$TEAM_NAME
```

```
## [1] Utah Jazz
## 30 Levels: Atlanta Hawks Boston Celtics ... Washington Wizards
```

This model was very accurate for the 22-23 season. It predicted one team wrong. It predicted that the Utah Jazz would make the playoffs when they did not.

Method 2

We created a logistic regression model with predictors that had high correlation in the correllagram created in the early visuals. These predictors are plus minus, points scored, assists, and three point percentage. These statistics are used to predict a yes or no outcome deciding whether or not a team will make the playoffs.

```
Model_4 <- glm(Playoffs_dummy ~ PLUS_MINUS+PTS+AST+FG3_PCT, data = train_data_1722 , family = binomial(
summary(Model_4)
```

```
##
## Call:
## glm(formula = Playoffs_dummy ~ PLUS_MINUS + PTS + AST + FG3_PCT,
##     family = binomial(link = "logit"), data = train_data_1722)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.23425  -0.02522   0.00575   0.13394   2.13173
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   9.73504   16.32316   0.596 0.550912
## PLUS_MINUS    1.77409    0.49123   3.611 0.000304 ***
## PTS           0.06766    0.13531   0.500 0.617081
## AST          -0.05981    0.25880  -0.231 0.817245
## FG3_PCT     -42.93067   35.92287  -1.195 0.232056
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 164.22  on 119  degrees of freedom
## Residual deviance:  45.78  on 115  degrees of freedom
```

```
## AIC: 55.78
##
## Number of Fisher Scoring iterations: 8
```

After training the model using our training data, we are now predicting this years outcomes with this seasons data.

```
PredictedProbability <- predict(Model_4, nba_22_23, type="response")
PredictedProbability
```

```
##            1            2            3            4            5
## 8.417076e-01 9.999895e-01 7.774833e-01 8.085847e-05 9.387753e-01
##            6            7            8            9           10
## 9.999367e-01 5.846585e-01 9.956141e-01 1.021527e-06 9.347568e-01
##           11           12           13           14           15
## 5.190355e-06 4.303725e-03 6.275814e-01 9.138573e-01 9.996437e-01
##           16           17           18           19           20
## 6.033903e-01 9.987573e-01 5.824871e-01 9.751542e-01 9.978942e-01
##           21           22           23           24           25
## 9.475256e-01 2.706129e-02 9.991214e-01 9.694560e-01 1.105762e-03
##           26           27           28           29           30
## 9.935655e-01 5.064140e-08 9.855626e-01 3.433043e-01 1.768850e-01
```

```
Predicted_Playoffs <- ifelse(PredictedProbability > 0.5, 1, 0)
```

```
## [1] "The accuracy is: 0.966666666666667"
```

```
## [1] "Number of teams predicted to make the playoffs: 21"
```

```
## [1] "Number of teams to actually make the playoffs: 20"
```

```
confusion <- table(actual=nba_22_23$Playoffs_dummy, predicted=Predicted_Playoffs)
confusion
```

```
##       predicted
## actual  0  1
##      0  9  1
##      1  0 20
```

```
nba_22_23[Predicted_Playoffs != nba_22_23$Playoffs_dummy, ]$TEAM_NAME
```

```
## [1] Dallas Mavericks
## 30 Levels: Atlanta Hawks Boston Celtics ... Washington Wizards
```

This model was very accurate for the 22-23 season. It predicted one team wrong. It predicted that the Dallas Mavericks would make the playoffs when they did not.

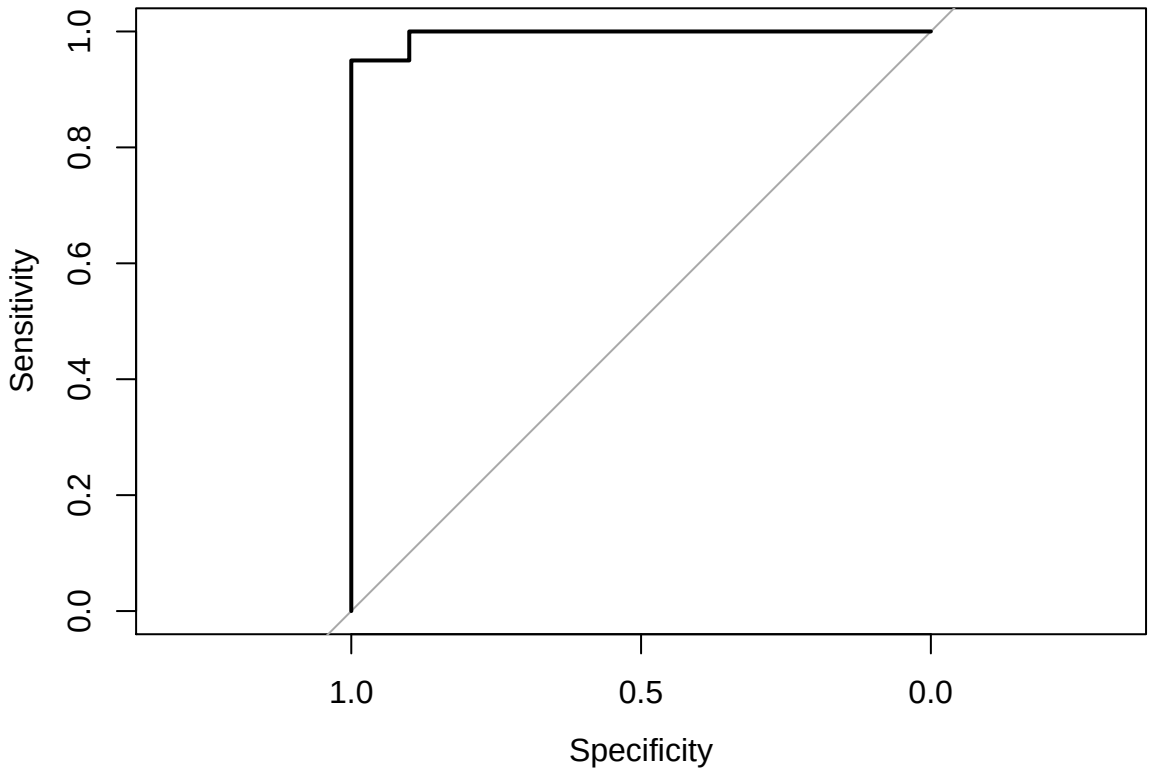
Below is an ROC graph for the predictions for our model.

```
MyROC <- roc(nba_22_23$Playoffs_dummy, PredictedProbability)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(MyROC)
```



Model 4 Method Comparisons

When comparing our two models to predict if a team would make the playoffs, we can see they had very similar outcomes. The only common predictors between the two models were FG3_PCT and PLUS_MINUS. Both models had the same accuracy when tested on the 22-23 season. They both also predicted 1 more team would make the playoffs than actually should. It was just a different team from both models. Method 2 had less stats then method 1 did, which means we don't necessarily need more stats to make it more accurate. One thing that we could do to improv the model is we know how many teams are going to make the playoffs- it is 10 teams from each conference (that includes the play in teams). Instead of saying if the probability is above 0.5 the team makes the playoffs, we could set it so that the top 10 teams from each conference with the highest probability make the playoffs (so 20 teams total), noting that of the top 10, the bottom 4 would be in the play ins (so 8 total teams for the play ins).