STA 141A - Project Report

Ryandeep Chawla, Justin Luong, Yuting Qiu, Roissom Myel Ringor

due December 10

Contents

List of Figures

i. Contributions

Group 18

Ryandeep Chawla

Home-Game Model/Away-Game Model/Model Analysis

Justin Luong

Interpretation and Conclusion of Results/Full Model Summary/Google Colab Formatting/-Support for Home-Game and Away Game Models

Yuting Qiu

Introduction/Full Model Analysis/Full Model Summary/Correlation Analysis

Roissom Myel Ringor

Introduction/OverLeaf Formatting/NBA Background Info/Debugging/Model Analysis

ii. Introduction

For this paper, we will use the data set from Kaggle that includes all NBA seasonal games from 2003 to 2020. This data set specifically includes field goal percentage, free throws percentage, three-point percentage, number of assists, number of rebounds, number of winning game for both home team and away team. Although this data set provides a lot of useful information, it is too broad to analyze all games,so we decided to set the time range from 2015 to 2020. Therefore, we will investigate the relationship between winning game and the predictor variables and analyze the factor that help the NBA team to increase their chances of winning based on data from 2015 to 2020.

iii. Data Background and Questions of Interest

Data source: https://www.kaggle.com/nathanlauga/nba-games/version/7?select=games.csv

For our project, we will be using data uploaded to Kaggle to analyze NBA games. Specifically, we will be using the "games.csv" data set that spans the 2003 to 2020 NBA seasons, which consists of 21 variables. These are the following columns in which we believe will be relevant information for data analysis for this project:

- GAME_DATA_EST - (year-month-day) date of a given game as a character
- SEASON - (year) the season that a game occurred; 2003 to 2020
- PTS_home - number of points scored by the home team
- FG_PCT_home - (FG%) field goal percentage for the home team
- FT_PCT_home - (FT%) free throw percentage for the home team
- FG3_PCT_home - (3P%) three-point percentage for the home team
- AST_home - number of assists by the home team
- REB_home - number of rebounds by the home team
- PTS_away - number of points scored by the away team
- FG_PCT_away - (FG%) field goal percentage for the away team
- FT_PCT_away - (FT%) free throw percentage for the away team
- FG3_PCT_away - (3P%) three-point percentage for the away team
- AST_away - number of assists by the away team
- REB_away - number of rebounds by the away team
- HOME_TEAM_WINS - 1 if the home team won, 0 if the home team lost

The teams are differentiated by a TEAM_ID variable, but the corresponding ID of teams with their actual names are indicated in a separate "teams.csv" file. While this is important to easily distinguish which teams played in each game, we can easily add our own column to the data set to

match each TEAM_ID with its corresponding NBA team name as one of our functions. In total, there are 24,677 objects of 21 variables in this data set. However, there are some rows in which NAs are present, so we will most likely omit these rows and/or simply narrow down the seasons we analyze to form our conclusions. It seemed that the 2003 season was the only season in which any NAs were present since omitting this season yielded no NAs left in the data set.

Our goal, we will compare the home and away teams to see which team, in general, performs better in the NBA seasonal games from the 2015 to 2020 NBA seasons. Then, we will predict the winning teams for each game, and analyze the factors that will increase the chances of winning. We will consider the following factors:

- Do more assists by a team increase the chances of winning a game?
- Does a higher team free throw percentage lead to more wins?
- How important are rebounds when it comes to wins?
- Do teams with a high three-point percentage also have a high field goal percentage?

We will decide on which statistical learning technique to use to analyze our data set once we are more comfortable with the material. For our plots, we will try to help visualize our predictions and show which are the most important factors for winning teams. Finally, our functions could include cleaning up the data set since the entire set is not fully usable (hence the specific time frame via NBA seasons).

iv. Function

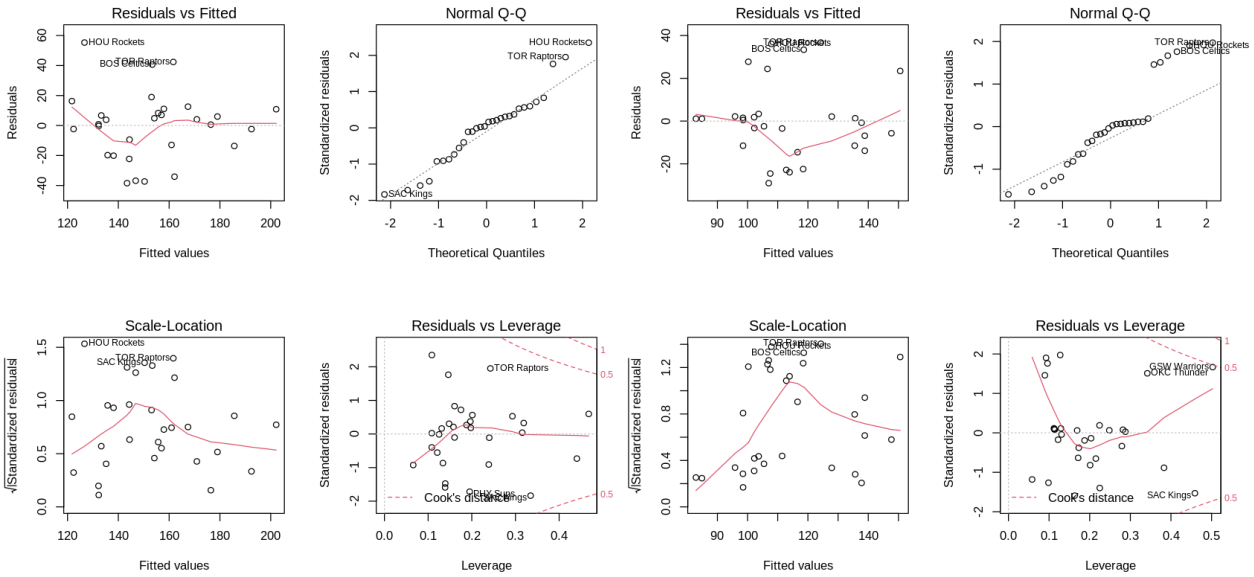
```
# This function was created to switch home team wins to
# away team wins because there was not an away team wins column

# x is the original column number you want to change
# y is the new column number

home2away <- function(x,y,NBA) {
  NBA$AWAY_TEAM_WINS<- "value"
  # adding a new column to the data frame,
  # which shows 1 if the away team won and 0 if the away team lost
  for (i in 1:nrow(NBA)) {
    if (NBA[i,x]==0){
      NBA[i,y]<-1
    }else{
      NBA[i,y]<-0
    }
  }
  return(NBA)
}

NBA <- home2away(21,22,NBA)
```

Figure 1: Function for NBA Data Set

Figure 1 was used to create a new column in the data set for away team wins. This allowed us to determine if there were any differences for how home and away teams won games. Later in this report, it does seem like how home teams get wins have some differences between how away teams get wins.

v. Setting up Models

1. Analysis of Full Models



(a) home team (b) away team

Figure 2: Full Model Linear Plots

From observing the Normal Q-Q plots, we can see that most points fall apart from the regression lines and curve off in the extremities for both the home team model and away team model. Normal Q-Q plots usually show the data having more extreme values than expected, so we can conclude that both the home team model and away team model don't hold the assumption of normality.

From observing our Residual vs. Fitted plots, we see that linearity is violated for both the home team model and away team model. The residuals appear to be randomly spread. There are three outliers close to residual 40: HOU Rockets, BOS Celtics, and the TOR Raptors.

From the Scale-Location plot, we can see that the red line is approximately a quadratic relationship for both the home team model and away team model. It suggests that the residual variables appear randomly, which means the average of the standardized residuals is not approximately constant.

From the Residual vs. Leverage plot, we can see that there is no outlier fall outside the red dash line for both home and away team models, which indicates a low leverage and small potential influence of a single observation on the regression model.

2. Summary of Full Models

For this project's home and away team wins statistics, we manipulated the data into total amount of wins for each team from the 2015 to 2020 NBA seasons so there would be 30 rows. Therefore, we used the mean of each statistic to be consistent with the number of NBA teams in our regression models and correlation analysis. Another reason why we used means was that it would not have been plausible to sum up the percentage statistics among others. We created one regression model each for the home team statistics, away team statistics, and field goal and three-point percentage.

The model of home team has R-squared = 0.4466, which represents the weak positive correlation between the response variable (home team wins) and the predictor variables (mean of field goal percentage, mean of free throw percentage, mean of three-point percentage, mean of assists, and mean of rebounds). The regression model predicts the home wins based on the predictor variables with a residual standard error of about 24.93 on 24 degrees of freedom. There are large coefficients and standard error for the percentage stats (mean of field goal percentage, mean of free throw percentage, and mean of three-point percentage), which hint that the shooting percentages play a big part in determining wins. The model also has a low p-value (0.01027), which means that there is a significant relationship between these predictor variables and the response variable. However, all of the predictor variables have p-values that are higher than the significance level, which indicate statistical insignificance.

The model of away team has R-squared = 0.4927, which represents the weak positive correlation between the response variable (away team wins) and the predictor variables (mean of field goal percentage, mean of free throw percentage, mean of three-point percentage, mean of assists, and mean of rebounds). The regression model predicts the away wins based on the predictor variables with a residual standard error of about 19.92 on 24 degrees of freedom. There are large coefficients and standard error for the percentage stats (mean of field goal percentage, mean of free throw percentage, mean of three-point percentage), which hint that the shooting percentages play a big part in determining wins. The model also has a low p-value (0.004096), which means that there is a significant relationship between these predictor variables and the response variable. However, all of the predictor variables have p-values that are higher than the significance level, which indicate statistical insignificance.

The model of three-point percentage regressed on field percentage R-squared = 0.4843, which represent the weak positive correlation between response variable (field goal percentage) and the predictor variable (three-point percentage). The regression model predicts the field goal percentage based on three-point percentage with a residual standard error of 0.007667 on 28 degrees of freedom. The three-point percentage coefficient is 0.60384 with a standard error of 0.11775 and p-value of 1.95e-05. This indicates that the sample means are not widely spread across the population mean and that the sample likely represents the population closely. The p-value is lower than 0.01 so that means the data is statistically significant.

vi. Data Analysis

1. Statistical Analysis

```
> summary(home_team_wins_lm)

Call:
lm(formula = HW ~ FG_mean + FT_mean + FG3_mean + AST_mean + REB_mean)

Residuals:
    Min      1Q  Median      3Q     Max
-38.440 -13.485   2.305  10.206  55.323

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1182.371    433.811  -2.726   0.0118 *
FG_mean      1326.085    822.621   1.612   0.1200
FT_mean       285.822    328.816   0.869   0.3933
FG3_mean      573.346    646.821   0.886   0.3842
AST_mean       -1.671      3.858  -0.433   0.6688
REB_mean        7.498      5.433   1.380   0.1803
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.93 on 24 degrees of freedom
Multiple R-squared:  0.4466,    Adjusted R-squared:  0.3313
F-statistic: 3.873 on 5 and 24 DF,  p-value: 0.01027
```

(a) Home Teams

```
> summary(away_team_wins_lm)

Call:
lm(formula = AW ~ FG_mean_away + FT_mean_away + FG3_mean_away +
    AST_mean_away + REB_mean_away)

Residuals:
    Min      1Q  Median      3Q     Max
-29.014 -11.508  -0.157   2.139  36.716

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1080.239    306.516  -3.524  0.00174 **
FG_mean_away   1050.544    661.884   1.587  0.12556
FT_mean_away    238.690    212.242   1.125  0.27188
FG3_mean_away  1169.048    548.924   2.130  0.04365 *
AST_mean_away    -2.665      3.613  -0.738  0.46787
REB_mean_away     4.222      3.680   1.147  0.26258
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.92 on 24 degrees of freedom
Multiple R-squared:  0.4927,    Adjusted R-squared:  0.387
F-statistic: 4.661 on 5 and 24 DF,  p-value: 0.004096
```

(b) Away Teams

Figure 3: Stat Averages Summary

Home Team Model: From the regression of all the predictor variables related to the home games, we can see that most of the variables have a high p-value, which results in failing to reject the null hypothesis, as all the p-values of each variable is higher than 0.05 level of significance. Furthermore, there is high standard error for almost all predictor variables except AST_mean and REB_mean comparatively. This may be due to the use of the mean of these predictor variables.

Away Team Model: From the regression of all the predictor variables related to the away games, similar to the home game models, we can see that most of the variables have a high p-value, which results in failing to reject the null hypothesis, as all the p-values of each variable is higher than 0.05 level of significance. We fail to reject the null for all predictor variables except for the predictor variable FG3_mean, as it has a p-value of 0.04365 which is less than the 0.05 level of significance so we can conclude to reject the null. Furthermore, similar to the home game model, there is high standard error for almost all predictor variables except AST_mean and REB_mean comparatively. This may be due to the use of the mean of these predictor variables.

```
Call:
lm(formula = HW ~ FG_mean + FT_mean + FG3_mean + AST_mean + REB_mean +
    FG_mean_away + FT_mean_away + FG3_mean_away + AST_mean_away +
    REB_mean_away)

Residuals:
    Min      1Q  Median      3Q     Max
-33.266 -16.738  -1.134  14.842  43.562

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1139.396    473.086  -2.408   0.0263 *
FG_mean       1462.519   1531.196   0.955   0.3515
FT_mean       -241.689    634.887  -0.381   0.7077
FG3_mean        15.747    918.997   0.017   0.9865
AST_mean         4.946      6.754   0.732   0.4729
REB_mean         8.344     15.556   0.536   0.5979
FG_mean_away    -8.714   1484.399  -0.006   0.9954
FT_mean_away   502.893    558.176   0.901   0.3789
FG3_mean_away  852.638   1006.143   0.847   0.4073
AST_mean_away   -9.680      8.371  -1.156   0.2619
REB_mean_away   -3.589     12.995  -0.276   0.7854
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.64 on 19 degrees of freedom
Multiple R-squared:  0.5368,    Adjusted R-squared:  0.2931
F-statistic: 2.202 on 10 and 19 DF,  p-value: 0.06654
```

(a) Home Teams

```
Call:
lm(formula = AW ~ FG_mean + FT_mean + FG3_mean + AST_mean + REB_mean +
    FG_mean_away + FT_mean_away + FG3_mean_away + AST_mean_away +
    REB_mean_away)

Residuals:
    Min      1Q  Median      3Q     Max
-26.721 -11.591  -4.694   7.664  38.102

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1125.1345   392.6122  -2.866  0.00989 **
FG_mean       1376.1442  1270.7340   1.083  0.29238
FT_mean        106.7607   526.8906   0.203  0.84158
FG3_mean      -643.0776   762.6722  -0.843  0.40961
AST_mean         1.4494     5.6054   0.259  0.79874
REB_mean         3.8803    12.9099   0.301  0.76701
FG_mean_away   -23.8900  1231.8972  -0.019  0.98473
FT_mean_away   244.7582   463.2285   0.528  0.60336
FG3_mean_away 1452.0508   834.9949   1.739  0.09821 .
AST_mean_away   -5.4065     6.9473  -0.778  0.44603
REB_mean_away   -0.4754    10.7842  -0.044  0.96530
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.28 on 19 degrees of freedom
Multiple R-squared:  0.5419,    Adjusted R-squared:  0.3008
F-statistic: 2.248 on 10 and 19 DF,  p-value: 0.06177
```

(b) Away Teams

Figure 4: Full Models Summary

Home Team Model: From the regression model with setting home team wins as a responsible variable and predictor variables related to both home games and away games, we can see that all variables have a high p-value, which results in failing to reject the null hypothesis, as all p-values of each variables is higher than 0.05 of significance.

Away Team Model: From the regression model with setting away team wins as a responsible variable and predictor variables related to both home games and away games, we can see that all variables have a high p-value, which results in failing to reject the null hypothesis, as all p-values of each variables is higher than 0.05 of significance.

Home Team model is similar to Away Team model. Both suggests that all predictor variables play an insignificant role for team to win the game and it also suggests that where a team plays does not have a statistically significant factor for wins the game.

2. Correlation Plots

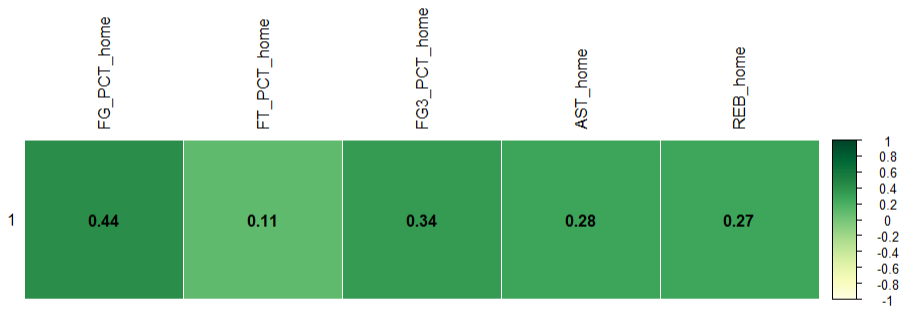


Figure 5: Correlation Plot for Home Teams

- The correlation between home team wins and field goal percentage for the home team is 0.44
- The correlation between home team wins and free throw percentage for the home team is 0.11
- The correlation between home team wins and three-point percentage for the home team is 0.34
- The correlation between home team wins and number of assists by the home team is 0.28
- The correlation between home team wins and number of rebounds by the home team is 0.27

Since the correlation between home team wins and free throw percentage for the home team (0.11), the number of assists (0.28), and number of rebounds (0.27) are relatively low, we can say that the free throw percentage, the number of assists, and numbers of rebounds for the home team have a weak correlation when it comes to increasing the chance to winning games.

Figure 6: Correlation Plot for Away Teams

- The correlation between away team wins and field goal percentage for the away team is 0.44
- The correlation between away team wins and free throw percentage for the away team is 0.11
- The correlation between away team wins and three-point percentage for the away team is 0.34
- The correlation between away team wins and number of assists by the away team is 0.29
- The correlation between away team wins and number of rebounds by the away team is 0.29

Since the correlation between home team wins and free throw percentage for the home team (0.11), the number of assists (0.29), and number of rebounds (0.29) are relatively low, we can say that the free throw percentage, the number of assists, and numbers of rebounds for the home team have a weak correlation when it comes to increasing the chance to winning games.

Concerning the chance to win a game, home teams and away teams both have a correlation of 0.44 with field goal percentage and a correlation of 0.34 with three point percentage, which means that where a team plays does not have a statistically significant factor on the outcome of a game.

vii. Interpretation and Conclusions

1. Do more assists by a team increase the chances of winning a game?

For home games, the wins column increases by 4.946 on average when there is an additional assist. The statistic is insignificant since the p-value of 0.29238 is greater than the significance level of 0.05. From this, we can conclude that assists do not significantly contribute by itself to whether or not a team will win a game or not. The small coefficient of 4.946 further supports this finding that more assists do not increase the chance of winning the game individually. There is a low standard error of 6.754, but it is big in comparison to the coefficient of 4.946. This means that the sample means are widely spread across the population mean and that the sample likely does not represent the population well. The correlation between home team wins and number of assists by the home team is 0.28, which indicates very weak correlation and supports our findings.

The results are similar but even less significant for away games, as the wins column increases by only 1.4494 on average when there is an additional assist. The statistic is insignificant since the p-value of 0.79874 is greater than the significance level of 0.05. There is a low standard error of 5.6054, but it is big in comparison to the coefficient of 1.4494. This means that the sample means are widely spread across the population mean and that the the sample likely does not represent the population well. The correlation between away team wins and number of rebounds by the away team is 0.29, which indicates very weak correlation and supports our findings.

2. Does a higher team free throw percentage lead to more wins?

For home games, the wins column decreases by 241.689 on average when there is an additional increase in free throw percentage. The statistic is insignificant since the p-value of 0.7077 is very high and is greater than the significance level of 0.05. There is a very high standard error of 634.887 which indicates that the sample means are widely spread across the population mean and that the sample likely does not represent the population well. The correlation between home team wins and free throw percentage is 0.44, which indicates weak correlation and supports our findings. Since the free throw percentage coefficient was negative and the statistics were insignificant, we conclude with these findings that free throw percentage do not increase the chance of winning the game individually.

The results are a bit different for away games, as the wins columns increase by 106.7607 on average when there is an additional increase in free throw percentage. The statistic is insignificant since the p-value of 0.84158 is very high and is greater than the significance level of 0.05. There is a very high standard error of 526.8906 which indicates that the sample means are widely spread across the population mean and that the sample likely does not represent the population well. The correlation between away team wins and free throw percentage is 0.44, which indicates weak correlation and supports our findings. With this, we can conclude that free throw percentage for away teams is insignificant when it comes to increasing the chance of winning the game individually.

3. How important are rebounds when it comes to wins?

For home games, the wins column increases by 8.344 when there is an additional increase in rebounds per game. The statistic is insignificant since the p-value of 0.5979 is very high and is greater than the significance level of 0.05. There is a low standard error of 15.556, but it is big in comparison to the coefficient of 8.344, which indicates that the sample means are widely spread

across the population mean and that the sample likely does not represent the population well. The correlation between away team wins and free throw percentage is 0.27, which indicates very weak correlation and supports our findings. With this, we can conclude that rebounds for home teams is insignificant when it comes to increasing the chance of winning the game individually.

For away games, the wins column increases by 3.8803 when there is an additional increase in rebounds per game. The statistic is insignificant since the p-value of 0.76701 is very high and is greater than the significance level of 0.05. There is a low standard error of 12.9099, but it is big in comparison to the coefficient of 3.8803, which indicates that the sample means are widely spread across the population mean and that the sample likely does not represent the population well. The correlation between away team wins and free throw percentage is 0.29, which indicates very weak correlation and supports our findings. With this, we can conclude that rebounds for away teams is insignificant when it comes to increasing the chance of winning the game individually.

4. Do teams with a high three-point percentage also have a high field goal percentage?

When regressing the average field goal percentage on the average three-point percentage, there is an increase of 0.60384 to the field goal percentage when there is an additional increase of 1 in three-point percentage. The statistic is significant since the p-value of 1.95e-05 is very low and is less than 0.001. There is a low standard error of 0.11775 which indicates that the sample means are not widely spread across the population mean and that the sample likely represents the population closely. The correlation between field goal percentage and three point percentage for home games is 0.47890109708297, which indicates weak correlation. The correlation between field goal percentage and three point percentage for away games is 0.428102773945153, which also indicates weak correlation. With this, we can conclude that three-point percentage significantly influences the field goal percentage, so teams with a high three-point percentage will likely have a high field goal percentage as well. This makes sense because there has been trend in recent years of basketball teams shooting more three points to emulate the success of the Golden State Warriors. Since, they have been shooting more three pointers, they must have worked on their success rate, so the three-point percentage should also be connected to a high field goal percentage.

5. Reasoning for statistical significance when determining wins:

From our analysis, it is obvious to see that the stat variables from the data set do not show much statistical significance towards a win for either home or away teams. However, our models show that including all stats without leaving a single stat out as variables to determine wins was the best way to predict the outcomes of games. From basketball, this is especially true since the ability to score points is not the only variable that determines the outcome of games. Defensive plays against the opposing team, intangibles for individual players, matchups between teams, and a lot of other factors help determine which team wins. This data set is great, but it only gave the total stats after the end of each recorded game. There are also stats that are tracked per quarter, and since there are four, twelve-minute quarters to play, it is entirely possible and even quite common for one team to play well in the first half of a game, but then lose the entire game in the last quarter. Some teams also play better than their usual selves when playing against certain teams and whether or not they play at home, which is another factor that is quite hard to assess with this particular data set. Overall, this data set simply proved the importance of the super in-depth statistical analysis that sports networks conduct for games, which range from brief overviews of total stats produced by each team to per-minute statistics of each individual player.

Appendices

a. Project Report code (ProjectReport.R)

note: Google Colab cell codes were compiled into one ProjectReport.R file

```r
NBA <- read.csv("games.csv")
NBA <- NBA[which(NBA$SEASON >= 2015),]

home_team_total_wins <- lapply(split(NBA$HOME_TEAM_WINS, NBA$HOME_TEAM_ID),sum)
home_team_total_wins <- lapply(home_team_total_wins, sort) # teams in order of ID
team_names <- c("ATL Hawks",
                "BOS Celtics",
                "CLE Cavaliers",
                "NOP Pelicans",
                "CHI Bulls",
                "DAL Mavericks",
                "DEN Nuggets",
                "GSW Warriors",
                "HOU Rockets",
                "LAC Clippers",
                "LAL Lakers",
                "MIA Heat",
                "MIL Bucks",
                "MIN Timberwolves",
                "BKN Nets",
                "NYK Knicks",
                "ORL Magic",
                "IND Pacers",
                "PHI 76ers",
                "PHX Suns",
                "POR Trail Blazers",
                "SAC Kings",
                "SAS Spurs",
                "OKC Thunder",
                "TOR Raptors",
                "UTA Jazz",
                "MEM Grizzlies",
                "WAS Wizards",
                "DET Pistons",
                "CHA Hornets")
names(home_team_total_wins) <- team_names
home_team_total_wins[which.max(home_team_total_wins)] # team with most wins

home_team_total_wins

NBA <- NBA[order(NBA$HOME_TEAM_ID),]
unique(NBA$HOME_TEAM_ID)
```

```r


FG_mean<- sapply(split(NBA$FG_PCT_home, NBA$HOME_TEAM_ID),mean)
FG_mean<- sapply(FG_mean, sort)
names(FG_mean) <- team_names
#FG_mean

FT_mean<- sapply(split(NBA$FT_PCT_home, NBA$HOME_TEAM_ID),mean)
FT_mean<- sapply(FT_mean, sort)
names(FT_mean) <- team_names
#FT_mean

FG3_mean<- sapply(split(NBA$FG3_PCT_home, NBA$HOME_TEAM_ID),mean)
FG3_mean<- sapply(FG3_mean, sort)
names(FG3_mean) <- team_names
#FG3_mean

AST_mean<- sapply(split(NBA$AST_home, NBA$HOME_TEAM_ID),mean)
AST_mean<- sapply(AST_mean, sort)
names(AST_mean) <- team_names
#AST_mean

RED_mean<- sapply(split(NBA$REB_home, NBA$HOME_TEAM_ID),mean)
RED_mean<- sapply(RED_mean, sort)
names(RED_mean) <- team_names
#RED_mean


HW<- unlist(home_team_total_wins)


home_team_wins_lm <- lm(HW~FG_mean+FT_mean+FG3_mean+AST_mean+RED_mean)
summary(home_team_wins_lm)



# This function was created to switch home team wins to
# away team wins because there was not an away team wins column

# x is the original column number you want to change
# y is the new column number

home2away <- function(x,y,NBA) {
  NBA$AWAY_TEAM_WINS<- "value" #adding a new column to the data frame, which
  ↪  shows 1 if the away team won and 0 if the away team lost
  for (i in 1:nrow(NBA)) {
    if (NBA[i,x]==0){
      NBA[i,y]<-1
```

```
    }else{
      NBA[i,y]<-0
    }
  }
  return(NBA)
}

NBA <- home2away(21,22,NBA)



away_team_total_wins <- sapply(split(as.numeric(NBA$AWAY_TEAM_WINS),
↪  NBA$TEAM_ID_away),sum)
away_team_total_wins <- sapply(away_team_total_wins, sort) # teams in order of ID
names(away_team_total_wins) <- team_names

FG_mean_away<- sapply(split(NBA$FG_PCT_away, NBA$TEAM_ID_away),mean)
FG_mean_away<- sapply(FG_mean_away, sort)
names(FG_mean_away) <- team_names
#FG_mean_away

FT_mean_away<- sapply(split(NBA$FT_PCT_away, NBA$TEAM_ID_away),mean)
FT_mean_away<- sapply(FT_mean_away, sort)
names(FT_mean_away) <- team_names
#FT_mean_away

FG3_mean_away<- sapply(split(NBA$FG3_PCT_away, NBA$TEAM_ID_away),mean)
FG3_mean_away<- sapply(FG3_mean_away, sort)
names(FG3_mean_away) <- team_names
#FG3_mean_away

AST_mean_away<- sapply(split(NBA$AST_away, NBA$TEAM_ID_away),mean)
AST_mean_away<- sapply(AST_mean_away, sort)
names(AST_mean_away) <- team_names
#AST_mean_away

RED_mean_away<- sapply(split(NBA$REB_away, NBA$TEAM_ID_away),mean)
RED_mean_away<- sapply(RED_mean_away, sort)
names(RED_mean_away) <- team_names
#RED_mean_away

AW<- unlist(away_team_total_wins)

away_team_wins_lm <-
↪  lm(AW~FG_mean_away+FT_mean_away+FG3_mean_away+AST_mean_away+RED_mean_away)
summary(away_team_wins_lm)


par(mfrow=c(2,2))
```

```r
plot(home_team_wins_lm)
par(mfrow=c(2,2))
plot(away_team_wins_lm)
par(mfrow=c(1,1))


homewins_vs_homeawaystats_lm <- lm(HW~FG_mean+FT_mean+FG3_mean+AST_mean+RED_mean+FG_mean_away+FT_mean_away+FG3_mean_away+AS
summary(homewins_vs_homeawaystats_lm)

awaywins_vs_homeawaystats_lm <- lm(AW~FG_mean+FT_mean+FG3_mean+AST_mean+RED_mean+FG_mean_away+FT_mean_away+FG3_mean_away+AS
summary(awaywins_vs_homeawaystats_lm)


require(corrplot)
M=cor(NBA$HOME_TEAM_WINS,NBA[9:13])
corrplot(M,method = "color",addgrid.col = "white",col.lim = c(-1,1),col = COL1("YlGn"),col.pos="b",addCoef.col = "Black",tl.col="Black")

N=cor(as.integer(NBA$AWAY_TEAM_WINS),NBA[16:20])
corrplot(N,method = "color",addgrid.col = "white",col.lim = c(-1,1),col = COL1("YlGn"),col.pos="b",addCoef.col = "Black",tl.col="Black")


homewins_vs_homeawaystats_lm <- lm(HW~FG_mean+FT_mean+FG3_mean+AST_mean+RED_mean+FG_mean_away+FT_mean_away+FG3_mean_away+AS
summary(homewins_vs_homeawaystats_lm)


awaywins_vs_homeawaystats_lm <- lm(AW~FG_mean+FT_mean+FG3_mean+AST_mean+RED_mean+FG_mean_away+FT_mean_away+FG3_mean_away+AS
summary(awaywins_vs_homeawaystats_lm)


threept_fg_lm <- lm(FG_mean~FG3_mean)
summary(threept_fg_lm)


home_wins=subset(NBA,HOME_TEAM_WINS==1)
cor(home_wins$FG_PCT_home,home_wins$FG3_PCT_home)

awaywins=subset(NBA,AWAY_TEAM_WINS==1)
cor(awaywins$FG_PCT_away,awaywins$FG3_PCT_away)
```