

Final Project: An Analysis on NBA Team Stats

Team Members

(in alphabetical order by last name)

Tim Hill, John Ingresoll, Jason Min-Liang Kang, Rajalakshmi Venkateswaran

Introduction

In May 2018, Supreme Court stroke down a 1992 federal law that had prohibited states from authorizing sports betting. This change allows each state to decide on its own whether to open up their sports betting sector. This is a huge deal to sports and entertainment industry. The general consensus is that most states will lean toward allowing sports betting since doing so will generate huge tax revenues. As more states adapt to allowing sports betting, the demands for sports data and sports data analytics will rise. In this project, we want to explore some opportunities in analysis of NBA player stats.

Questions and Hypothesis

We came up with many ideas about analyzing NBA player stats. For example, we can see if there is a potential relationship between NBA schedule, players playing minutes, and player injuries. We can also analyze home-games and away-games, and see how they impact player performances. For this project, however, we decided to keep things simple — we only want to make some predictions using team stats.

We say this is simpler because, for real NBA fans like us, we can already tell there is definitely a positive correlation between team stats and winning (i.e. teams with higher stats tend to win more). The point is not to see whether there exists a correlation; the point is to figure out what variables attribute to that correlation.

We divided all possible variables into tow categories — intangible variables and tangible variables. Intangible variables are things like injury, fatigue, team chemistry, and off-court distractions. There are plenty of examples through NBA history where intangible variables altered the outcome of NBA championship. For example, in 2003, Los Angeles Lakers had four hall-of-frame players in Shaquille O'Neal, Kobe Bryant, Karl Malone, and Gary Payton. On the stats sheet, they had everything it takes to win championship. But in reality, they lost to the underdog Detroit Piston in an embarrassing final series. After the season, Laker's head coach Phil Jackson revealed in his memoir for the first time that the team was under achieving because they were having team chemistry issues all alone. Even though, intangible variables can be significant, they are not captured in official NBA stats. On the other hand, tangible variables are the quantitative variables such as points per game and turn over per game. These variables are well captured in official NBA stats. In this project, we will overlook the intangible variables, and only consider the tangible ones.

Using the tangibles variables from NBA stats data, we want to answer the following questions:

- What are the variables that will affect winning games?
- Can we predict which team will be in playoff?
- Can we predict how many games a team will win?

Data Acquisition, Munging, and Dictionary

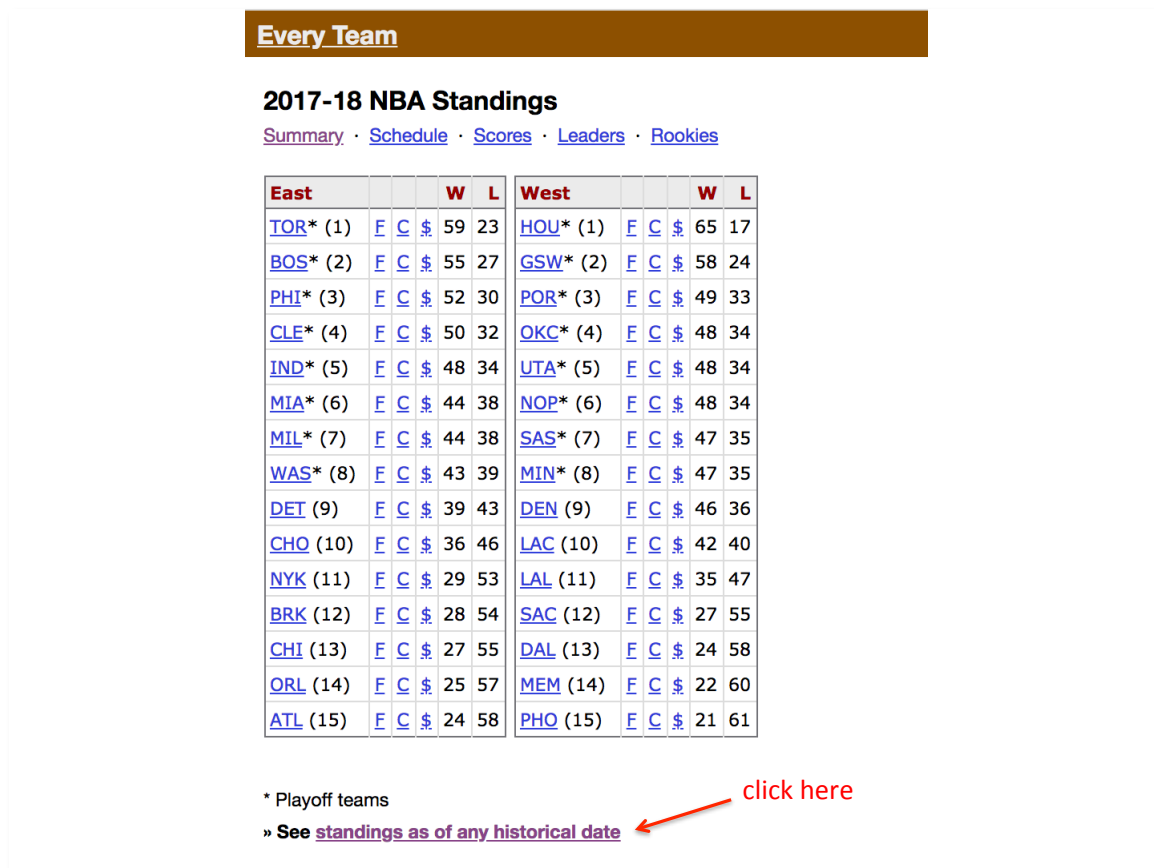
One thing we learned from this project is that getting good data is a very time consuming process. We had gotten a few datasets from Kaggle but they didn't work for our purpose. So we decided to collect our own dataset. In the following section, we would provide detail documentation on how to reproduce that dataset we ended up using.

Data Acquisition

We got our data from two sources — Basketball Reference and Wikipedia. Both sources require manual web scraping. The process is long, manual, and tedious. We wished there was a tool that can help us to automate this process.

Getting Team Stats

We got all team stats from Basketball Reference (basketball-reference.com). First, go to their homepage and click on the link to historical data page (Figure 1).



Every Team

2017-18 NBA Standings

[Summary](#) · [Schedule](#) · [Scores](#) · [Leaders](#) · [Rookies](#)

East					West				
			W	L				W	L
TOR * (1)	F	C	S	59 23	HOU * (1)	F	C	S	65 17
BOS * (2)	F	C	S	55 27	GSW * (2)	F	C	S	58 24
PHI * (3)	F	C	S	52 30	POR * (3)	F	C	S	49 33
CLE * (4)	F	C	S	50 32	OKC * (4)	F	C	S	48 34
IND * (5)	F	C	S	48 34	UTA * (5)	F	C	S	48 34
MIA * (6)	F	C	S	44 38	NOP * (6)	F	C	S	48 34
MIL * (7)	F	C	S	44 38	SAS * (7)	F	C	S	47 35
WAS * (8)	F	C	S	43 39	MIN * (8)	F	C	S	47 35
DET (9)	F	C	S	39 43	DEN (9)	F	C	S	46 36
CHO (10)	F	C	S	36 46	LAC (10)	F	C	S	42 40
NYK (11)	F	C	S	29 53	LAL (11)	F	C	S	35 47
BRK (12)	F	C	S	28 54	SAC (12)	F	C	S	27 55
CHI (13)	F	C	S	27 55	DAL (13)	F	C	S	24 58
ORL (14)	F	C	S	25 57	MEM (14)	F	C	S	22 60
ATL (15)	F	C	S	24 58	PHO (15)	F	C	S	21 61

* Playoff teams


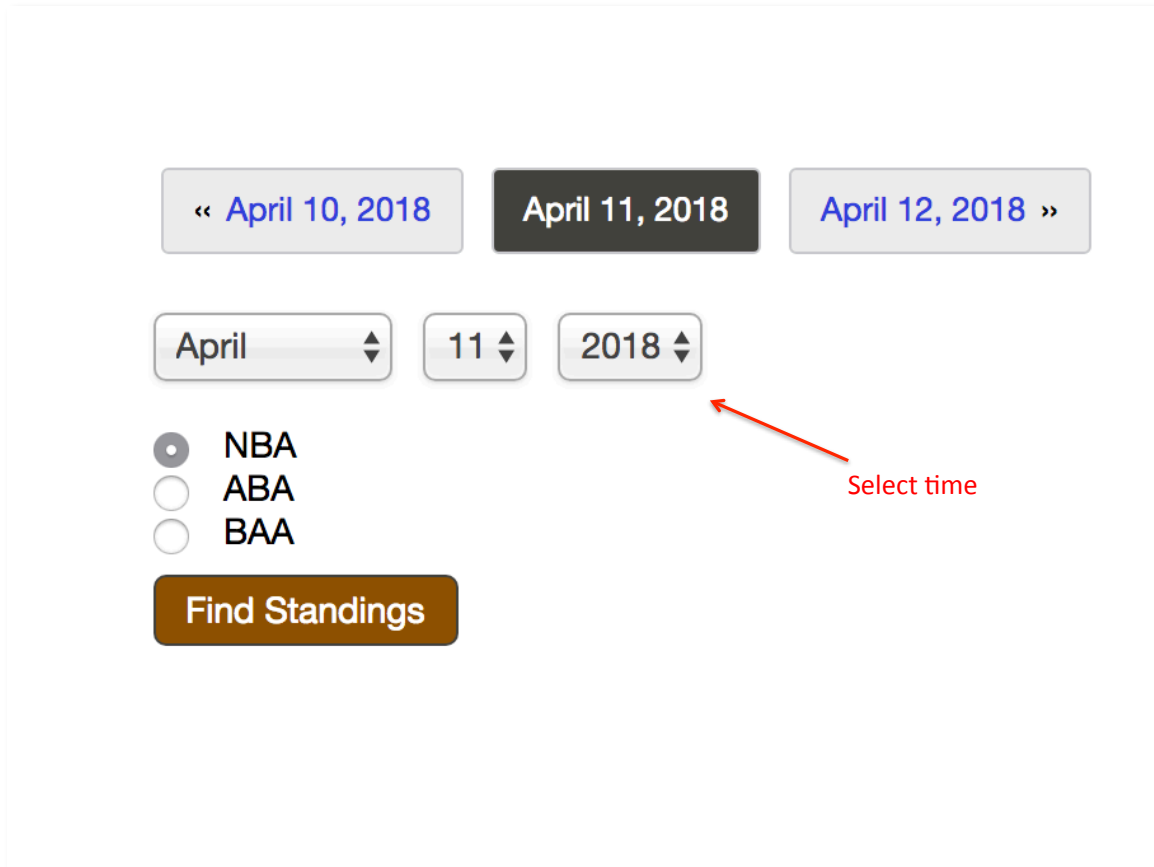
» See [standings as of any historical date](#)  **click here**

Figure 1. Link to Historical Data

On the Historical League Standing page, we use the timeline selection button (figure 2) at the top to select the year we want. Since we were collecting 20 years worth of data (from 1998 to 2018), except for 1999 and 2012 because those are short seasons. The page will generate teams stats at the bottom (figure 3). We then copy the team stats to a spreadsheet.



The image shows a web interface for selecting a date and league. At the top, there are three buttons: « April 10, 2018, April 11, 2018, and April 12, 2018 ». The middle button is highlighted. Below these are three dropdown menus for month (April), day (11), and year (2018). To the left of these are three radio buttons for NBA (selected), ABA, and BAA. A red arrow points to the year dropdown with the text "Select time". At the bottom is a brown button labeled "Find Standings".

Figure 2. Timeline Selection

Team Stats

Share & more ▼

[Glossary](#)

Rk	Team	G	MP	FG	FGA	FG%	3P	3PA
1	Golden State Warriors	82	19730	3509	6981	.503	926	2370
2	Houston Rockets	81	19515	3143	6824	.461	1243	3424
3	Toronto Raptors	81	19565	3341	7077	.472	959	2675
4	New Orleans Pelicans	81	19715	3448	7153	.482	832	2293
5	Cleveland Cavaliers	81	19490	3274	6856	.478	975	2612
6	Denver Nuggets	81	19615	3297	7012	.470	927	2504
7	Philadelphia 76ers	81	19540	3297	6993	.471	886	2409
8	Minnesota Timberwolves	81	19540	3323	6974	.476	649	1822
9	Los Angeles Clippers	81	19465	3261	6921	.471	771	2178
10	Charlotte Hornets	82	19780	3197	7106	.450	824	2233

Figure 3. Team Stats

Getting Team Standing

On the same Historical League Standing page, there is team standing data (figure 4) placed slightly above team stats. We would manually match the team name on team standing data to team stats, and copy the games won to the appropriate row on spreadsheet.

Standings

only need
this column

Eastern Conference	W	L	W/L%	GB	PW	PL	PS/G	PA/G
Toronto Raptors *	59	22	.728	—	60	21	111.7	103.7
Boston Celtics *	54	27	.667	5	50	31	103.9	100.5
Philadelphia 76ers *	51	30	.630	8	51	30	109.6	105.4
Cleveland Cavaliers *	50	31	.617	9	43	38	111.0	109.9
Indiana Pacers *	48	34	.585	11.5	45	37	105.6	104.2
Milwaukee Bucks *	44	37	.543	15	41	40	106.6	106.5

Figure 4. Team Standing

Getting Playoffs Data

We get payoffs data from Wikipedia (wikipedia.org). Unfortunately, playoff data on Wikipedia are stored in diagram format (figure 5). It adds extra works for us to translate it into tabular format. If a team is a playoff, we put a “Y” for final-16, else we put “N”. Final-8 means 2nd round of playoff and finaly-4 mean 3rd round, etc.

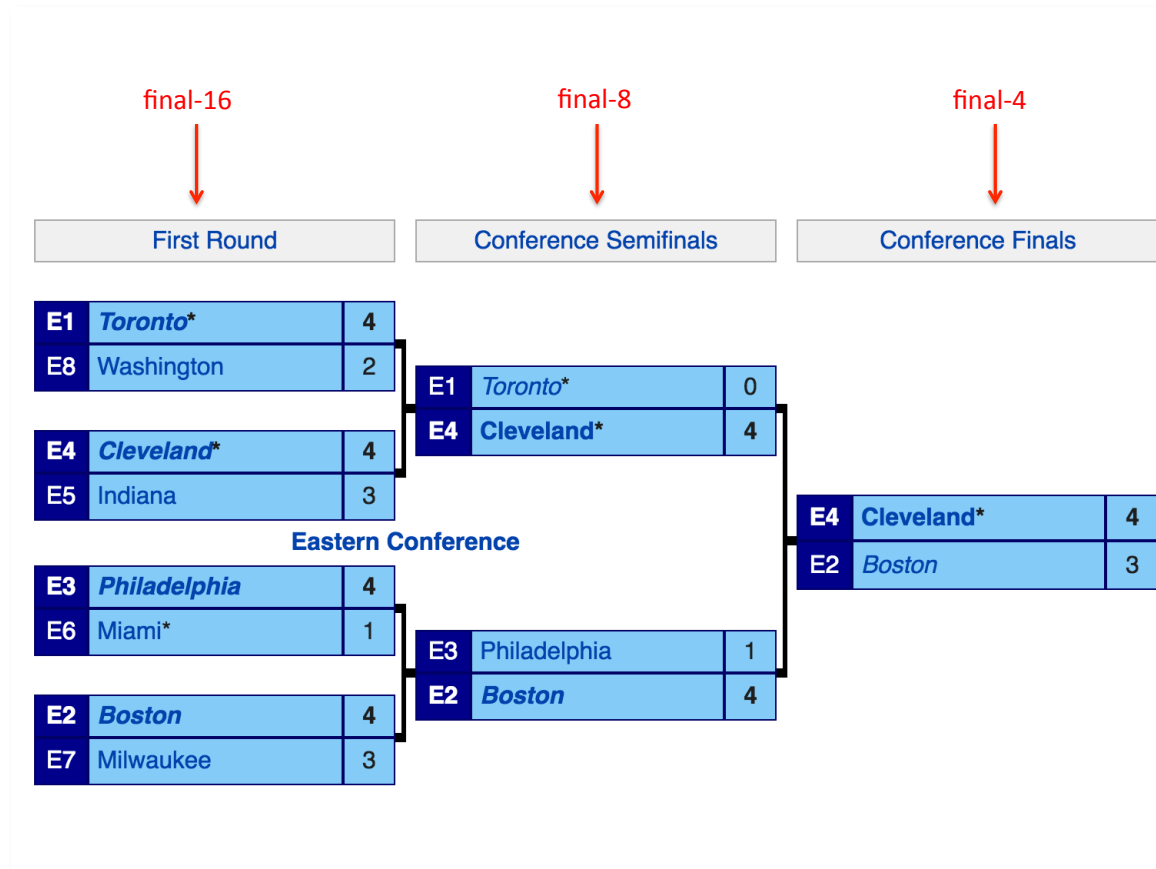


Figure 5. Playoffs Data

Data Munging

After we got all data in a spreadsheet, we exported the data into a CSV file and then loaded that into R. We also upload a copy of the CSV file on google drive. It is accessible using a sharing [URL](#) (CSV).

```
# Data is loaded from my local machine. It can be downloaded from
# https://drive.google.com/open?id=1A4CR2CGAe80u26xT4rk6U6kSAq16BswX

nba.team_stats <-
read.csv("/Users/kang/Documents/syracuse_datascience/IST_687_applied_data_science/project/
Nba_Players_Stats/team-stats.csv")
```

Inspecting the data frame reviews 564 rows and 31 variables, for a total of 17,484 data points. All data translated over to R nicely. Most variables are numbers and integers. Team name and playoff data are in factors.

```
str(nba.team_stats)
## 'data.frame': 564 obs. of 31 variables:
## $ Team : Factor w/ 36 levels "Atlanta Hawks",...: 15 32 34 19 12 28 20 36 6 5 ...
## $ year : int 1998 1998 1998 1998 1998 1998 1998 1998 1998 1998 ...
## $ G : int 77 78 77 77 78 77 76 77 77 77 ...
```

```
## $ MP      : int 18630 18795 18580 18730 18870 18755 18440 18605 18655 18580 ...
## $ FG      : int 2944 2899 2797 2959 2807 2930 2803 2888 2894 2770 ...
## $ FGA     : int 6156 6128 5724 6432 6198 6312 6405 6432 6408 5936 ...
## $ FG.     : num 0.478 0.473 0.489 0.46 0.453 0.464 0.438 0.449 0.452 0.467 ...
## $ X3P     : int 471 592 231 266 543 394 297 295 286 322 ...
## $ X3PA    : int 1341 1485 627 797 1575 1132 885 877 882 851 ...
## $ X3P.    : num 0.351 0.399 0.368 0.334 0.345 0.348 0.336 0.336 0.324 0.378 ...
## $ X2P     : int 2473 2307 2566 2693 2264 2536 2506 2593 2608 2448 ...
## $ X2PA    : int 4815 4643 5097 5635 4623 5180 5520 5555 5526 5085 ...
## $ X2P.    : num 0.514 0.497 0.503 0.478 0.49 0.49 0.454 0.467 0.472 0.481 ...
## $ FT      : int 1761 1466 1926 1567 1565 1360 1609 1403 1388 1567 ...
## $ FTA     : int 2589 2023 2493 2123 2029 1821 2175 2018 1867 2086 ...
## $ FT.     : num 0.68 0.725 0.773 0.738 0.771 0.747 0.74 0.695 0.743 0.751 ...
## $ ORB     : int 1010 872 899 1007 946 949 1246 1080 1185 928 ...
## $ DRB     : int 2298 2125 2265 2300 2240 2315 1989 2188 2309 2209 ...
## $ TRB     : int 3308 2997 3164 3307 3186 3264 3235 3268 3494 3137 ...
## $ AST     : int 1886 1897 1935 1935 1705 1966 1563 1781 1844 1810 ...
## $ STL     : int 704 754 606 587 652 702 715 649 650 655 ...
## $ BLK     : int 525 367 387 398 279 404 301 349 334 294 ...
## $ TOV     : int 1150 1061 1144 1043 1186 1123 1040 1037 1037 1132 ...
## $ PF      : int 1755 1714 1839 1774 1578 1661 1771 1719 1582 1653 ...
## $ PTS     : int 8120 7856 7751 7751 7722 7614 7512 7474 7462 7429 ...
## $ finaly.16: Factor w/ 3 levels "N","T","Y": 3 3 3 3 3 3 3 1 3 3 ...
## $ final.8  : Factor w/ 3 levels "N","T","Y": 3 3 3 1 1 1 1 1 3 3 ...
## $ final.4  : Factor w/ 3 levels "N","T","Y": 3 1 3 1 1 1 1 1 3 1 ...
## $ final.2  : Factor w/ 3 levels "N","T","Y": 1 1 3 1 1 1 1 1 3 1 ...
## $ champion : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 2 1 ...
## $ win      : int 57 59 59 42 40 53 42 38 60 48 ...
```

Some variables names were translated incorrectly, so we fix them by running the following script.

```
colnames(nba.team_stats)[which(names(nba.team_stats) == "FG.") <- "FGPer"
colnames(nba.team_stats)[which(names(nba.team_stats) == "X3P.") <- "3P"
colnames(nba.team_stats)[which(names(nba.team_stats) == "X3PA.") <- "3PA"
colnames(nba.team_stats)[which(names(nba.team_stats) == "X3P.") <- "3PPer"
colnames(nba.team_stats)[which(names(nba.team_stats) == "X2P.") <- "2P"
colnames(nba.team_stats)[which(names(nba.team_stats) == "X2PA.") <- "2PA"
colnames(nba.team_stats)[which(names(nba.team_stats) == "X2P.") <- "2PPer"
colnames(nba.team_stats)[which(names(nba.team_stats) == "FT.") <- "FTPer"
colnames(nba.team_stats)[which(names(nba.team_stats) == "finaly.16")] <- "final16"
colnames(nba.team_stats)[which(names(nba.team_stats) == "final.8")] <- "final8"
colnames(nba.team_stats)[which(names(nba.team_stats) == "final.4")] <- "final4"
colnames(nba.team_stats)[which(names(nba.team_stats) == "final.2")] <- "final2"
```

Miscellaneous Variables

```
head(nba.team_stats[,c(1,2:4, 20:25)])
## Team year G MP AST STL BLK TOV PF PTS
## 1 Los Angeles Lakers 1998 77 18630 1886 704 525 1150 1755 8120
## 2 Seattle SuperSonics 1998 78 18795 1897 754 367 1061 1714 7856
## 3 Utah Jazz 1998 77 18580 1935 606 387 1144 1839 7751
## 4 Minnesota Timberwolves 1998 77 18730 1935 587 398 1043 1774 7751
## 5 Houston Rockets 1998 78 18870 1705 652 279 1186 1578 7722
## 6 Phoenix Suns 1998 77 18755 1966 702 404 1123 1661 7614
```


Field Goals

(FGA = 3PA + 2PA)

```
head(nba.team_stats[,c(1,5:13)])
## Team FG FGA FG. 3P 3PA 3PPer X2P 2PA 2PPer
## 1 Los Angeles Lakers 2944 6156 0.478 471 1341 0.351 2473 4815 0.514
## 2 Seattle SuperSonics 2899 6128 0.473 592 1485 0.399 2307 4643 0.497
## 3 Utah Jazz 2797 5724 0.489 231 627 0.368 2566 5097 0.503
## 4 Minnesota Timberwolves 2959 6432 0.460 266 797 0.334 2693 5635 0.478
## 5 Houston Rockets 2807 6198 0.453 543 1575 0.345 2264 4623 0.490
## 6 Phoenix Suns 2930 6312 0.464 394 1132 0.348 2536 5180 0.490
```

Free Throw

```
head(nba.team_stats[,c(1,14:16)])
## Team FT FTA FTPer
## 1 Los Angeles Lakers 1761 2589 0.680
## 2 Seattle SuperSonics 1466 2023 0.725
## 3 Utah Jazz 1926 2493 0.773
## 4 Minnesota Timberwolves 1567 2123 0.738
## 5 Houston Rockets 1565 2029 0.771
## 6 Phoenix Suns 1360 1821 0.747
```

Rebounding

```
head(nba.team_stats[,c(1,17:19)])
## Team ORB DRB TRB
## 1 Los Angeles Lakers 1010 2298 3308
## 2 Seattle SuperSonics 872 2125 2997
## 3 Utah Jazz 899 2265 3164
## 4 Minnesota Timberwolves 1007 2300 3307
## 5 Houston Rockets 946 2240 3186
## 6 Phoenix Suns 949 2315 3264
```

Playoffs and Games Won

```
head(nba.team_stats[,c(1,26:31)])
## Team final16 final8 final4 final2 champion win
## 1 Los Angeles Lakers Y Y Y N N 57
## 2 Seattle SuperSonics Y Y N N N 59
## 3 Utah Jazz Y Y Y Y N 59
## 4 Minnesota Timberwolves Y N N N N 42
## 5 Houston Rockets Y N N N N 40
## 6 Phoenix Suns Y N N N N 53
```

Data Dictionary

Term	Definition
Team	Team name
year	Which
G	Number of games played

MP	Number of minutes played
FG	Number of field goals made; 2P+3P
FGA	Number of field goals attempted; 2PA+3PA
FGPer	Field goal percentage; FG/FGA
3P	Number of 3 points made
3PA	Number of 3 points attempted
3PPer	3 point percentage; 3P/3PA
2P	Number of 2 points made
2PA	Number of 2 points attempted
2PPer	2 point percentage; 2P/2PA
FT	Number of Free Throws made
FTA	Number of Free Throws attempted
FTPer	Free throw percentage; FT/FTA
ORB	Offensive rebound
DRB	Defensive rebound
TRB	Total rebound; ORB+DRB
AST	Assits
STL	Steal
BLK	Block
TOV	Turnover the ball
PF	Personal fouls
PTS	Total number of points
final16	In 1 st round of playoff
final8	In 2 nd round of playoff
final4	In 3 rd round of playoff
final2	In 4 th round of playoff; NBA final
champion	Wins NBA final
win	Games won

Interesting Finding

As we were preparing the dataset, we noticed that team with good winning records do not necessary have good stats. That means winning does not necessary have a strong correlation with good stats.

Descriptive Statistics

Our goal is to find correlation between team stats and winning. In this section, we plot out key variable against final-16 and wins.

AST

Strong correlation between assists and games won. However, assist does not predict which team will end up in playoff.

```
ggplot(nba.team_stats) +
  geom_point(aes(x=AST, y=win, color=final16)) +
  labs(title="Assists vs Playoff Brackets 16 and Wins")
```

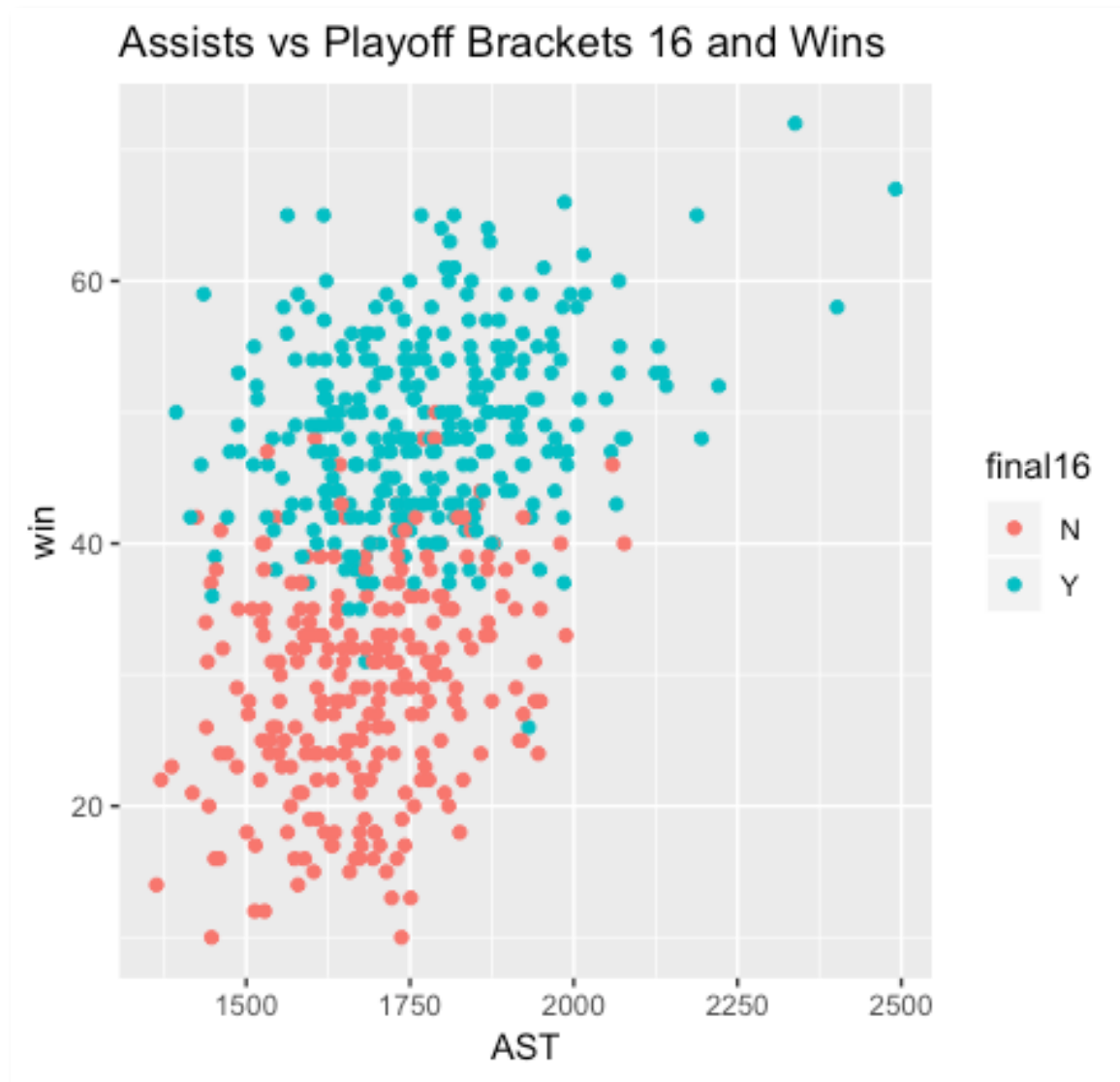


Figure 6

STL

Strong correlation between steals and games won. However, steal does not predict which team will end up in playoff.

```
ggplot(nba.team_stats) +
  geom_point(aes(x=STL, y=win, color=final16)) +
  labs(title="Steals vs Playoff Brackets 16 and Wins")
```

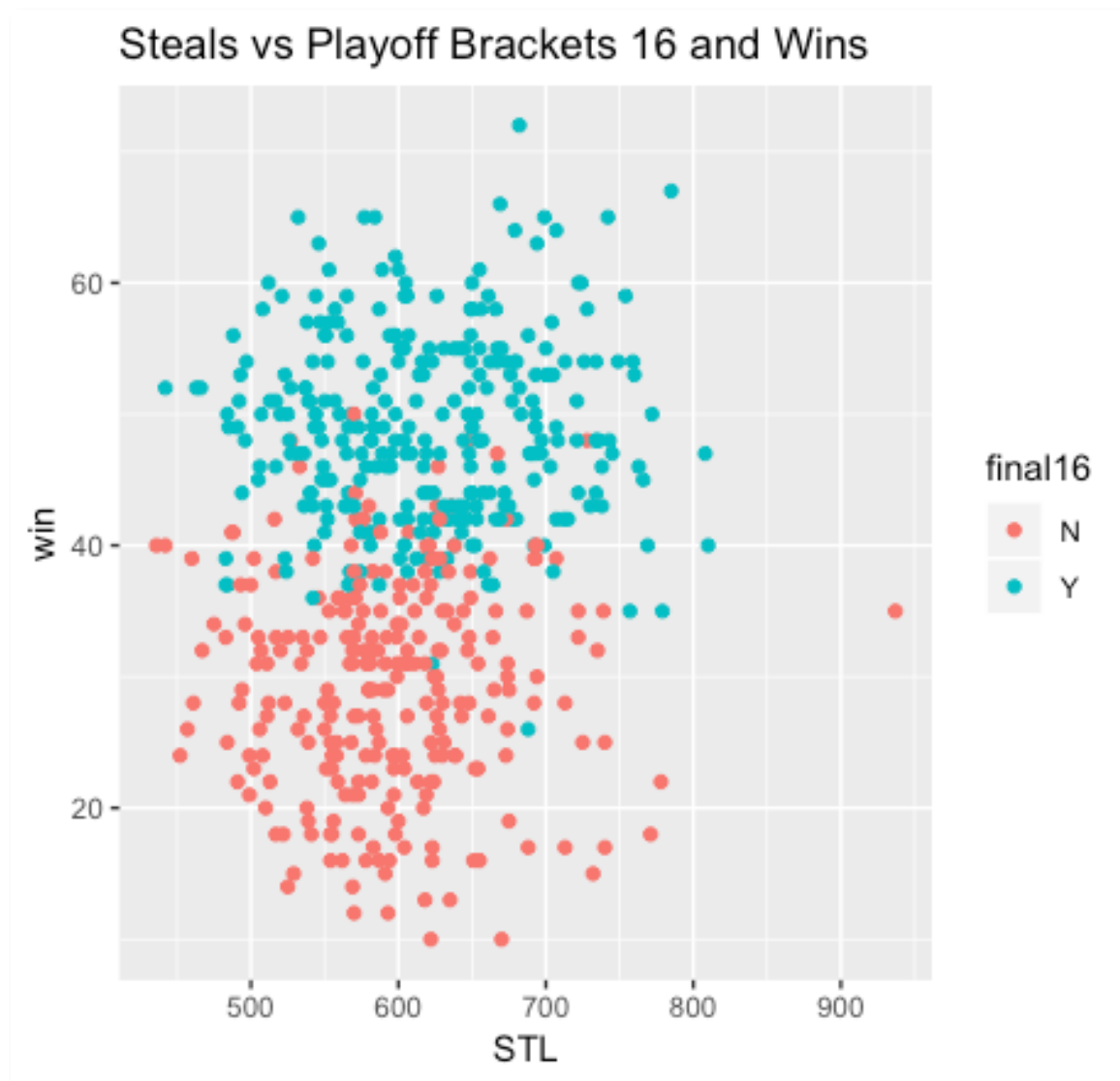


Figure 7

BLK

Strong correlation between block shoots and games won. However, block shoot does not predict which team will end up in playoff.

```
ggplot(nba.team_stats) +  
  geom_point(aes(x=BLK, y=win, color=final16)) +  
  labs(title="Blocks vs Playoff Brackets 16 and Wins")
```

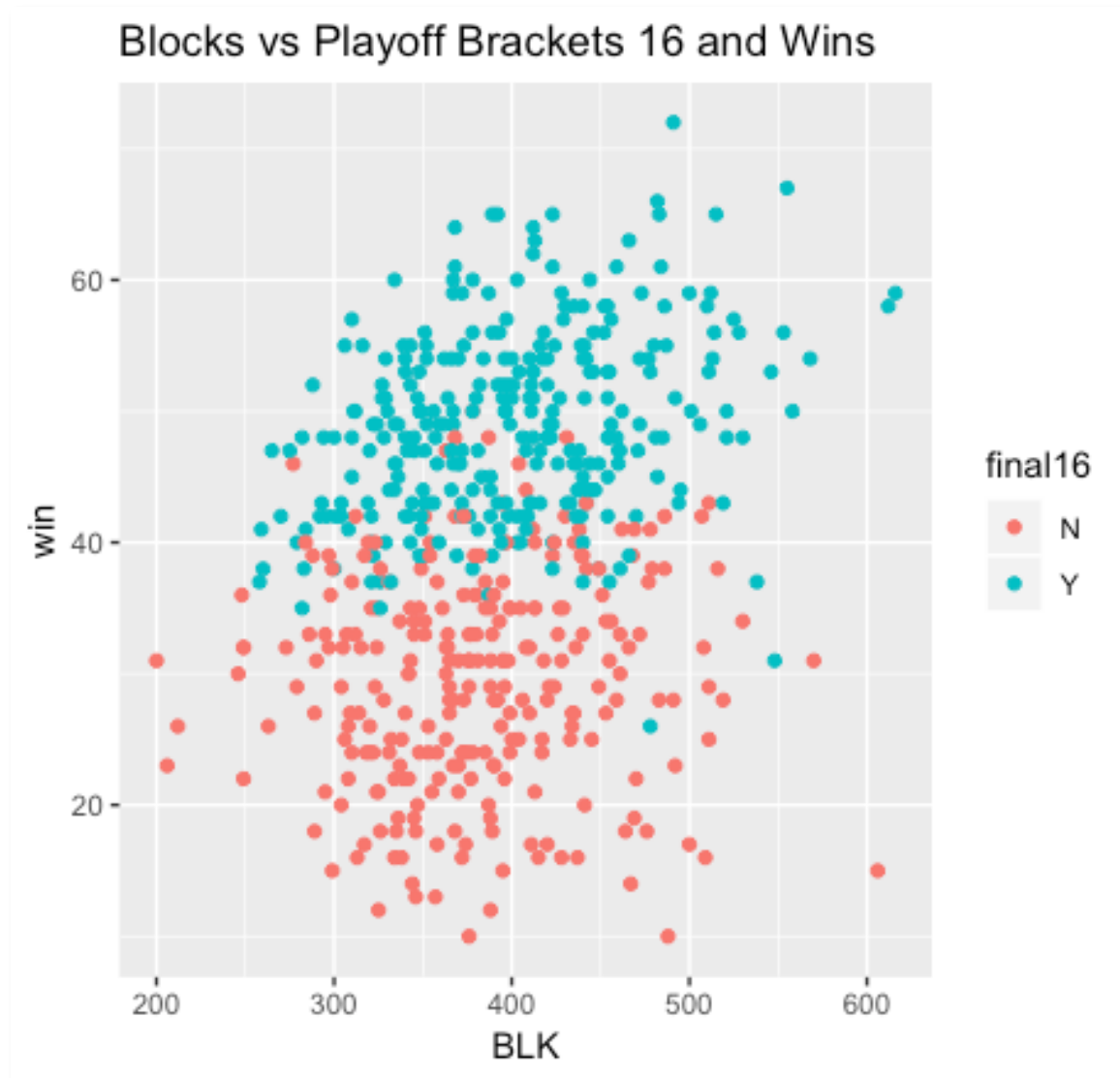


Figure 8

FG

Field goals percentage has a strong correlation to games won. As field goals percentage goes up, there is visually more green dots than red dot. This is a good indicator that there exists some sort of correlation between field goals percentage and playoffs as well.

```
ggplot(nba.team_stats) +  
  geom_point(aes(x=FGPer, y=win, color=final16)) +  
  labs(title="Field Goals Percentages vs Playoff Brackets 16 and Wins")
```

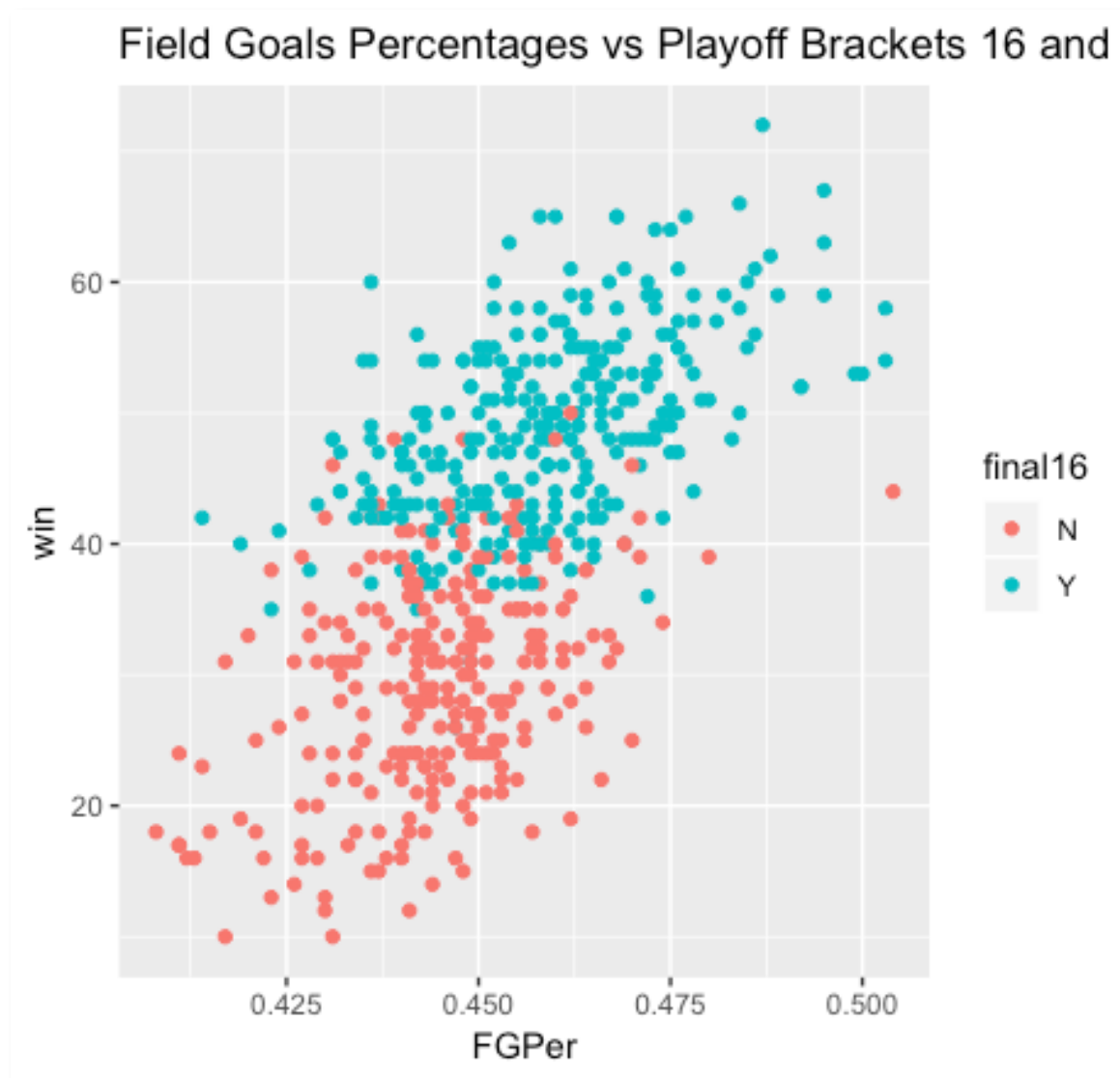


Figure 9

FT

Teams that win more games tend to shoot better at free throw line. However, free throw percentage is not an indicator of playoff team.

```
ggplot(nba.team_stats) +  
  geom_point(aes(x=FTPer, y=win, color=final16)) +  
  labs(title="Free Throw Percentages vs Playoff Brackets 16 and Wins")
```

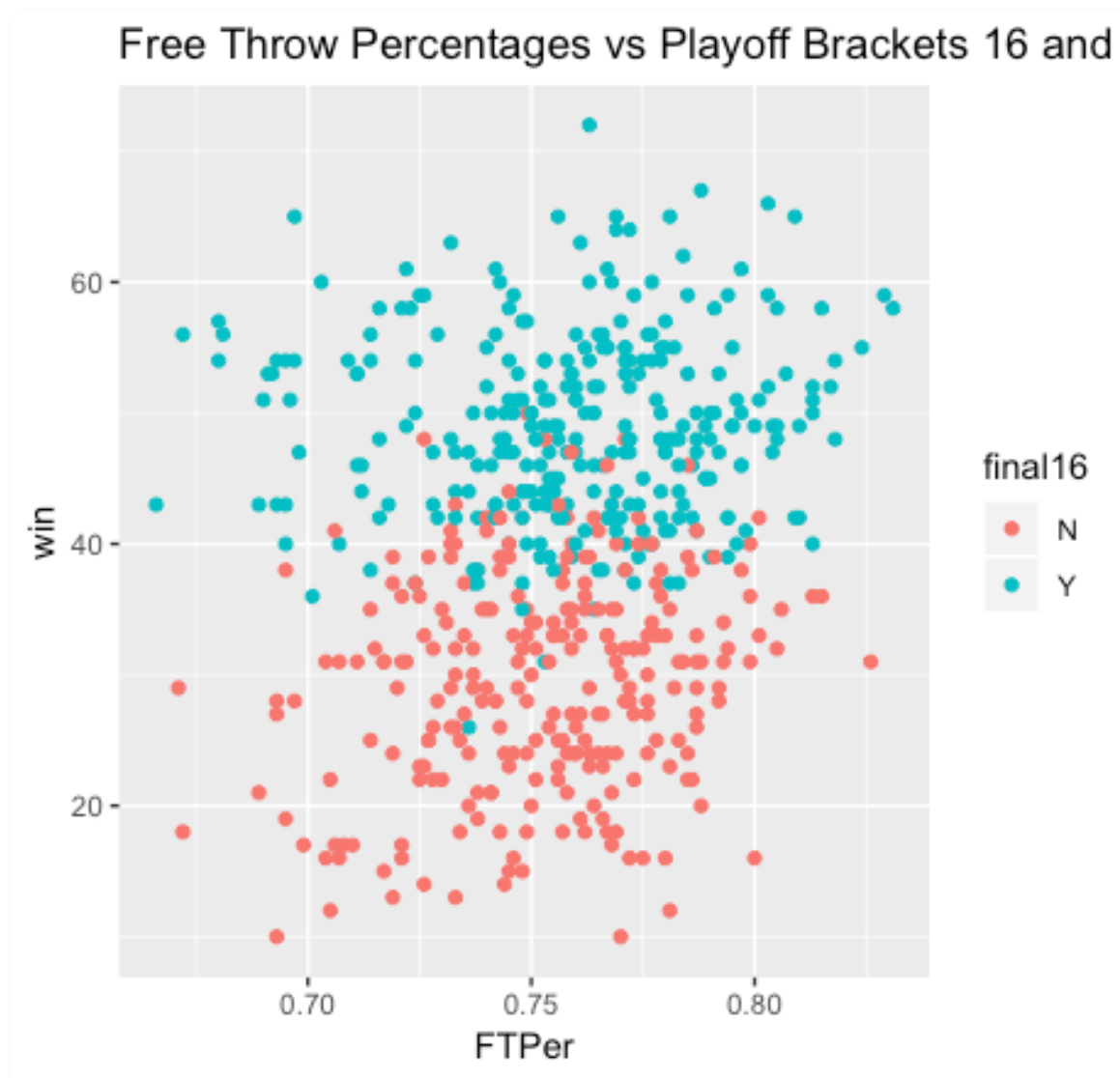


Figure 10

RB

Rebound has a strong correlation to winning more games, but it is not a good indicator of playoff team.

```
ggplot(nba.team_stats) +  
  geom_point(aes(x=TRB, y=win, color=final16)) +  
  labs(title="Rebounds vs Playoff Brackets 16 and Wins")
```

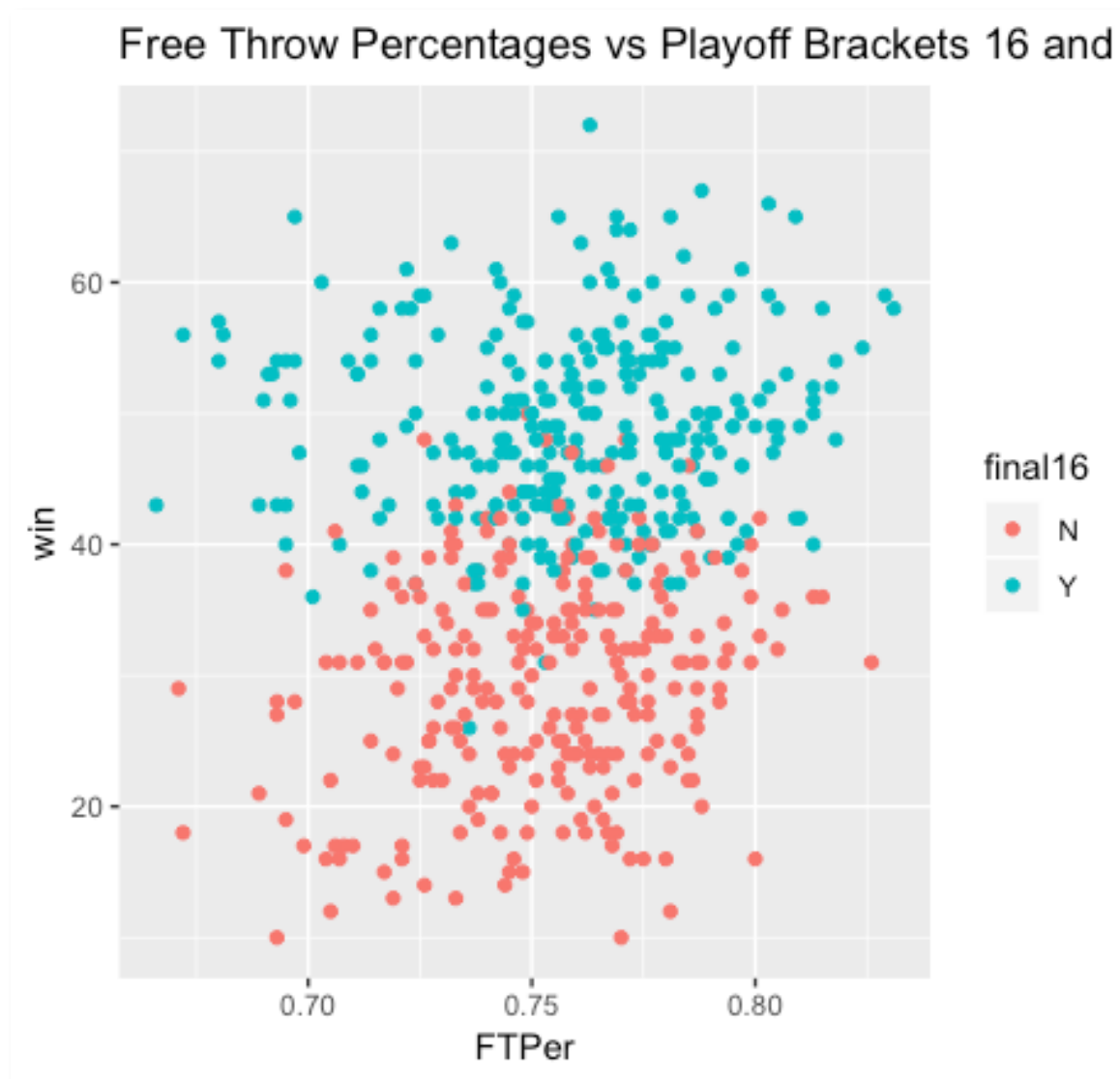


Figure 11

Use Of Modeling Techniques

In this section, we will be using modeling techniques to predict games won and playoff team using team stats. First, we need to divide our dataset into training dataset and test dataset.

```
# Create Training and Test Datasets
randIndex <- sample(1:dim(nba.team_stats)[1])
cutePoint2_3 <- floor(2 * dim(nba.team_stats)[1]/3)
nba.trainData <- nba.team_stats[randIndex[1:cutePoint2_3],]
nba.testData <- nba.team_stats[randIndex[(cutePoint2_3+1):dim(nba.team_stats)[1]],]
```


Predict Wins

For predicting games won, we use linear regression on field goal percentage and wins because field goal percentage is the best indicator of wins among all available variables. The result turns out really well, with a standard deviation of 10 games. That is pretty good for 80 games season. On hindsight, defense will certainly contribute to winning games. Unfortunately, defensive data was not available when we were creating our dataset.

```
wins_lm <- lm(data=nba.trainData, formula=win ~ FGPer)
nba.testData$predicted_win <- predict(wins_lm, nba.testData, type="response")
nba.testData$prediction_error <- nba.testData$win - nba.testData$predicted_win
sd(nba.testData$prediction_error) #=> 9.590283
ggplot(nba.testData) +
  geom_point(aes(x=FGPer, y=win, color=prediction_error)) +
  labs(title="Field Goal Percentages vs Wins Prediction Error")
```

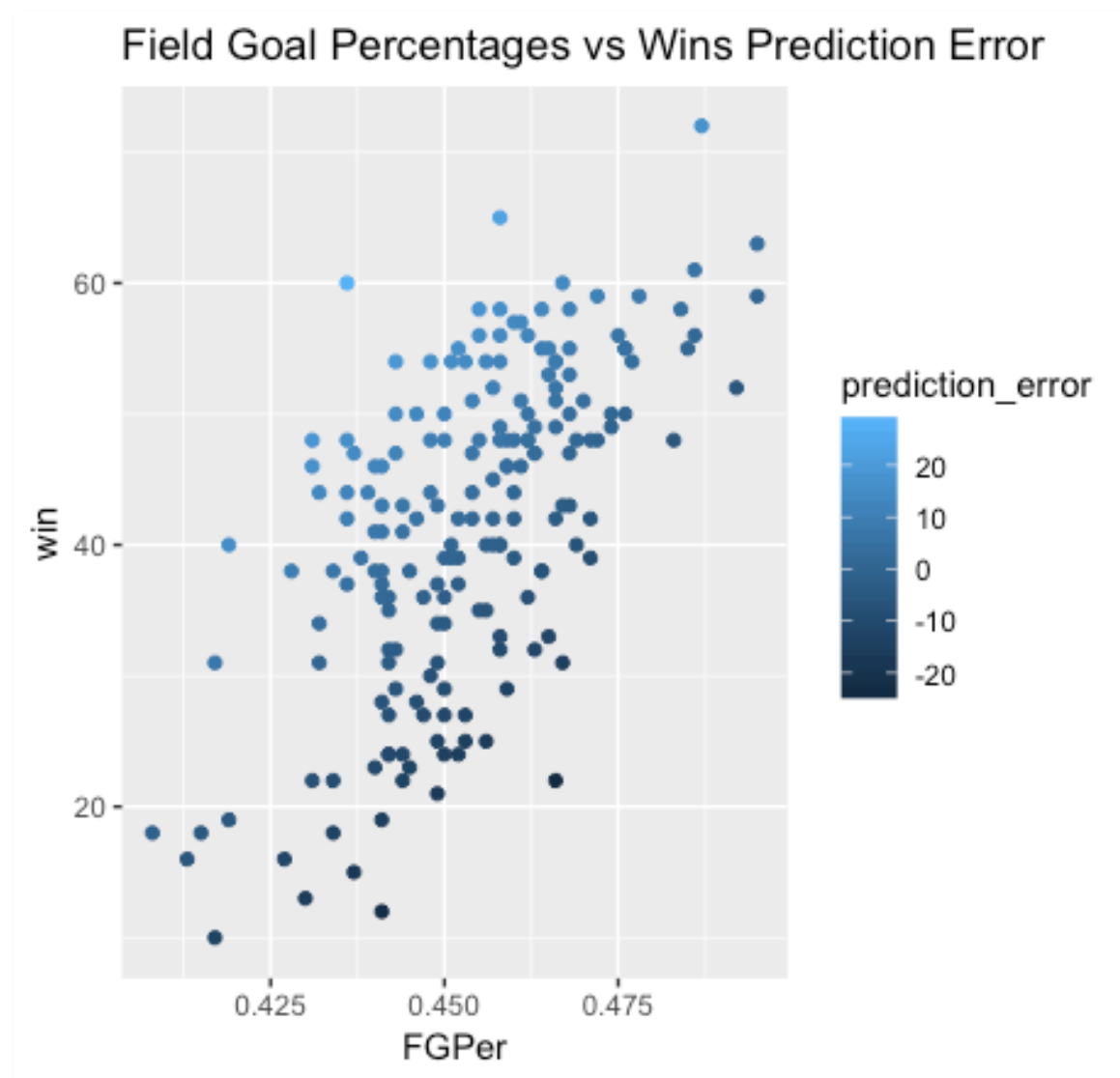


Figure 12

Predict Playoff Brackets

Since our descriptive analysis shows that all key variables have strong correlation with being a playoff team, we use all key variables to construct a support vector model. The result turns out really good as expected. The green triangles and red dots indicate correct prediction; the red triangles and green dots indicate wrong predictions. There are clearly more right predictions than wrong.

```
library(kernlab)
final16_ksvm <- ksvm(final16 ~ AST+STL+BLK+FGPer+FTPer+TRB, data=nba.trainData,
kernel="rbfdot", kpar="automatic", C=5, cross=3, prob.model=TRUE)
nba.testData$predicted_final16 <- predict(final16_ksvm, nba.testData, type="votes")[2,]
nba.testData$predicted_final16 <- factor(nba.testData$predicted_final16)
ggplot(nba.testData) +
  geom_point(aes(x=FGPer, y=win, shape=predicted_final16, color=final16))
```

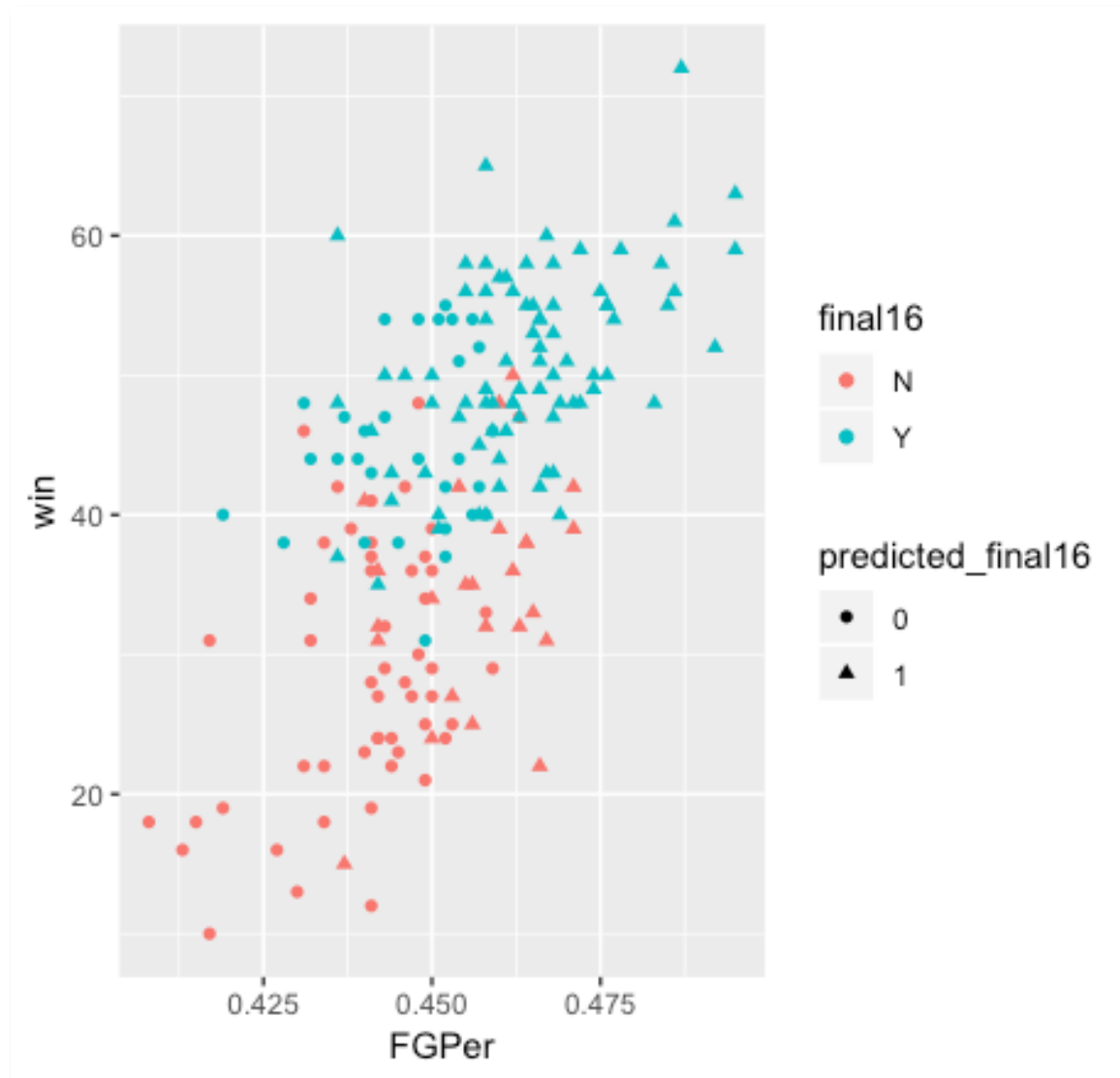


Figure 13

Summary and Interpretation

As we were gathering the dataset, we were worrying about there might not be a strong correlation between team stats and winning. However, the results turn out really well. We found strong correlation between team stats and being in playoffs, and we found a good predictor of winning games in field goal percentage. Looking at these results, we wonder why there is a discrepancy between being in playoff and winning games. Are they supposed to win a lot of games in order to be in playoff? We think the discrepancy lies in the way NBA season is structured. During regular season, eastern teams compete with western teams. During playoff, however, eastern teams and western teams compete within their own realm before meeting in final. As a result, teams with less wins may get to be in playoff as well.