EDA and Basketball

Theo Dimitrasopoulos

2019-12-29

Introduction

Basketball (and professional sports in general) is currently undergoing a data analytics revolution. Previously, players in team sports were evaluated qualitatively (e.g. "Shaquille O'Neal is an all-time great because if he gets the ball close to the basket, he almost always scores"). Now, NBA teams (and spectators like us) have access to a great deal of data about the game. Analysts and teams can answer questions about players and teams in a data-driven way, and while oftentimes these quantitative results corroborate the qualitative observations, teams believe that there are advantages to be gained from analyzing the data well.

I am not an analyst for NBA teams, so I won't be developing game plans or trying to identify underrated players. Instead, I use NBA data to practice plotting. I've scraped all the data from the NBA stats webpage. For the current season data, the data was last built on the morning of Feburary 23, 2016.

If you have no familiarity with basketball, I recommend you watch a game! Nationally televised games are on ESPN on Wednesdays and on TNT on Thursdays. And if you don't have time for games, you could watch an ESPN "30-for-30" mini-documentary about basketball. Two popular NBA basketball ones are "Winning Time: Reggie Miller vs. the New York Knicks" and "Bad Boys" — they're on Netflix and various video sites (just Google the title). Or, talk with a friend who knows the game!

Some background reading:

- New York Times article by Michael Lewis '82 on the now retired Shane Battier. The ending of the Lakers-Rockets game described the article can be seen at this YouTube link. **Highly recommended read!**. Some credit this article for bringing NBA analytics to the attention of the average spectator.
- Article about the general increase in three point shooting
- Detailed article about how the Toronto Raptors are using data analytics.

Tidy chunk for LaTeX cleanup:

```r
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Load libraries:

```r
library(reshape2)
library(dplyr)
library(ggplot2)
library(babynames)
library(stringr)
library(hexbin)
```

Exploratory Data Analysis

League wide statistics

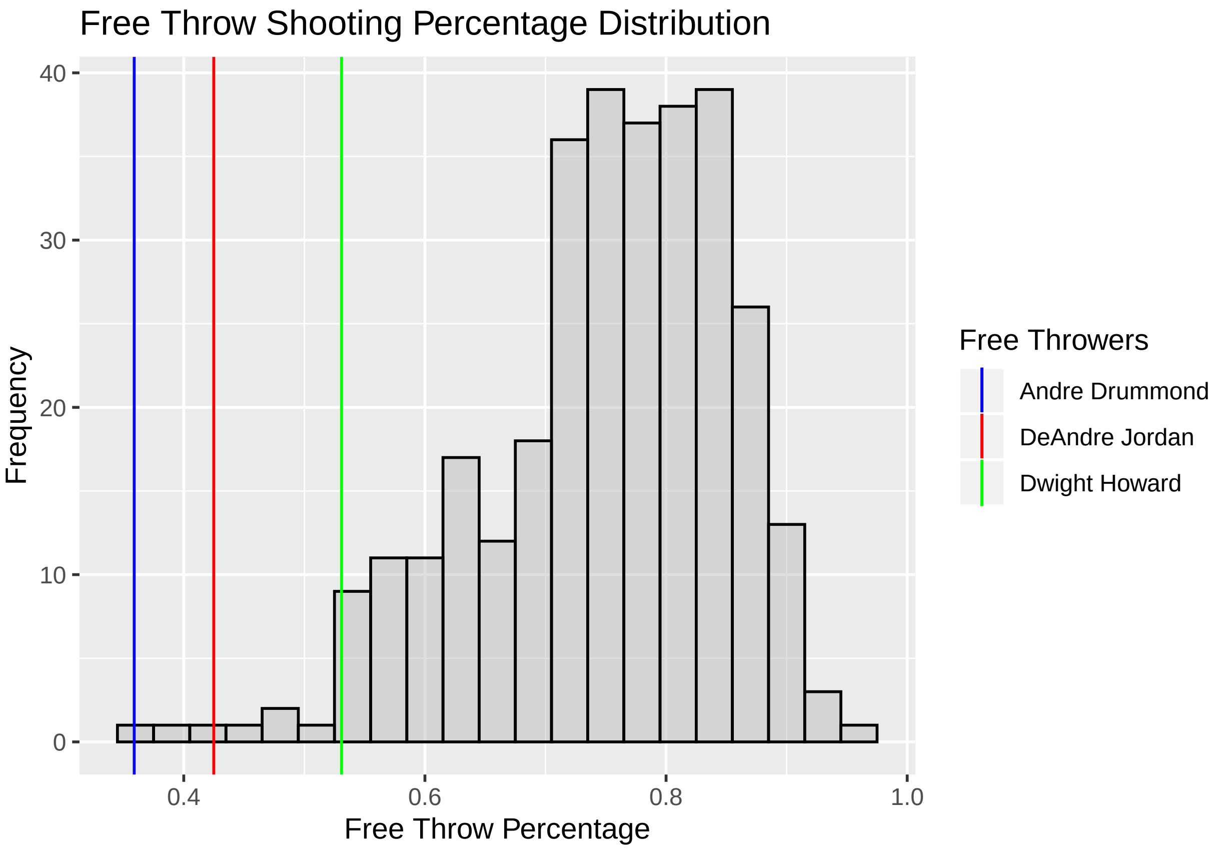
The file league_stats.txt (in the data folder) consists of statistics for players across the league. The data comes from this website and column header information can be found by hovering over the columns. Not all of the columns are in the text file.

```
league_stats <- read.table(
  "/Users/Theo/Desktop/SML_201/Projects/project_2/data/league_stats.txt",
  header=TRUE, sep="")
```

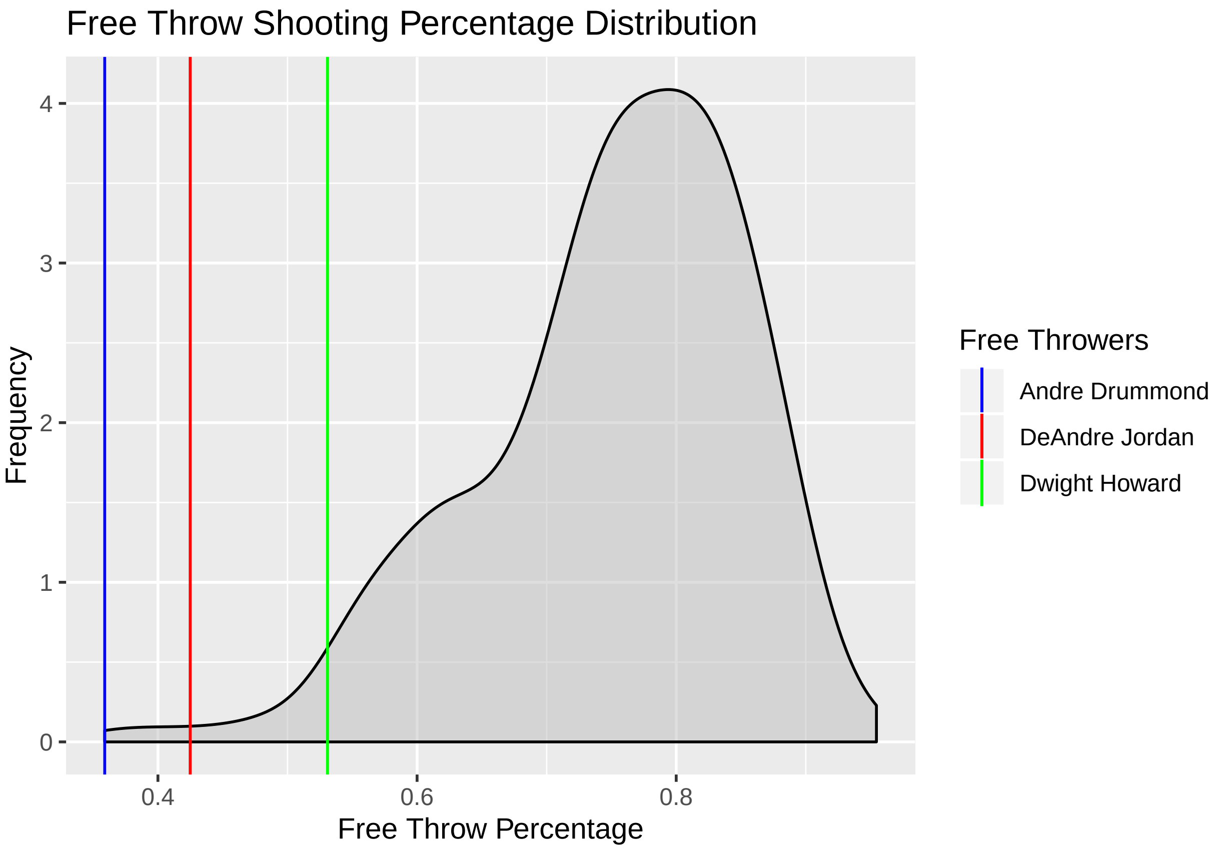
Shooting percentages in basketball are calculated by the number of shots of that type made divided by the number of shots of that type attempted. *Free throws* are penalty shots awarded when a player on the opposing team commits a foul. Three players often criticized for their poor free throw shooting percentage are Dwight Howard, DeAndre Jordan, and Andre Drummond. Consider the .**Here, I plot the distribution of the free throw shooting percentages across the league for all players who have taken an appreciable number of shots (at least 20) in two different ways, and where the three players mentioned earlier fall on the distribution.**

```
league_stats_sig <- mutate(filter(league_stats, FREE_THROWS_MADE >= 20), FREE_THROW_PERCENTAGE = FREE_T

#Method 1: Plot distribution using a histogram
ggplot(data = league_stats_sig, aes(x = FREE_THROW_PERCENTAGE)) +
  geom_histogram(binwidth = 0.03, col = "black", fill = "grey", alpha = 0.5)  +
  labs(title="Free Throw Shooting Percentage Distribution") +
  labs(x="Free Throw Percentage", y="Frequency") +
  geom_vline(aes(xintercept = 0.4248826, color = "DeAndre Jordan")) +
  geom_vline(aes(xintercept = 0.5308642, color = "Dwight Howard")) +
  geom_vline(aes(xintercept = 0.3589165, color = "Andre Drummond")) +
  scale_colour_manual(name='Free Throwers', values=c('DeAndre Jordan'='red', 'Dwight Howard'='green', '
```



```
#Method 2: Plot distribution using a density plot
ggplot(data = league_stats_sig, aes(x = FREE_THROW_PERCENTAGE)) +
  geom_density(col = "black", fill = "grey", alpha = 0.5)  +
  labs(title="Free Throw Shooting Percentage Distribution") +
  labs(x="Free Throw Percentage", y="Frequency") +
  geom_vline(aes(xintercept = 0.4248826, color = "DeAndre Jordan")) +
  geom_vline(aes(xintercept = 0.5308642, color = "Dwight Howard")) +
  geom_vline(aes(xintercept = 0.3589165, color = "Andre Drummond")) +
  scale_colour_manual(name='Free Throwers', values=c('DeAndre Jordan'='red', 'Dwight Howard'='green', '
```



Three point shots are shots taken outside of the outer most arc on the court. They award three points instead of the usual two. In recent years, coaches in the league have begun to consider three point shooting to be a critical part of a successful offensive strategy (see Background Reading). Some players are considered volume shooters, i.e. they take a ton of shots. **Here, I consider the players who have played at least 100 minutes this season; I compute the mean, median, and mode for the number of three point shot attempts. Further, I identify players that are outliers.**

```
#filter out players with <100 minutes of playtime
league_stats_100_min <- filter(league_stats, MIN >= 100)

#Mean calculation:
mean(league_stats_100_min[,13])
```

```
## [1] 99.09975
```

```
#Median calculation:
median(league_stats_100_min[,13])
```

```
## [1] 70
```

```
#Mode function No1:
league_mode <- function(i) {
  sep <- unique(i)
  sep[which.max(tabulate(match(i, sep)))]
}

#Mode function No2 (given in notes):
sample_mode <- function(x) {
  as.numeric(names(which(table(x) == max(table(x)))))
```

```r
}


#Mode calculation:
league_mode(league_stats_100_min[,13])
```

```
## [1] 0
```

```r
sample_mode(league_stats_100_min[,13])
```

```
## [1] 0
```

```r
#store the IQR
iqr_value <- 1.5*IQR(league_stats_100_min[,13])

#Method 1 for outliers:
outliers <- league_stats_100_min[which((league_stats_100_min[,13] - mean(league_stats_100_min[,13]) > (

outliers_1 <- subset(outliers, select = c(PLAYER_NAME))

outliers_1
```

```
##        PLAYER_NAME
## 72  Damian Lillard
## 152   James Harden
## 217  Klay Thompson
## 304    Paul George
## 351  Stephen Curry
```

```r
#Method 2 for outliers:
quantiles <- quantile(league_stats_100_min[,13])

quantiles
```

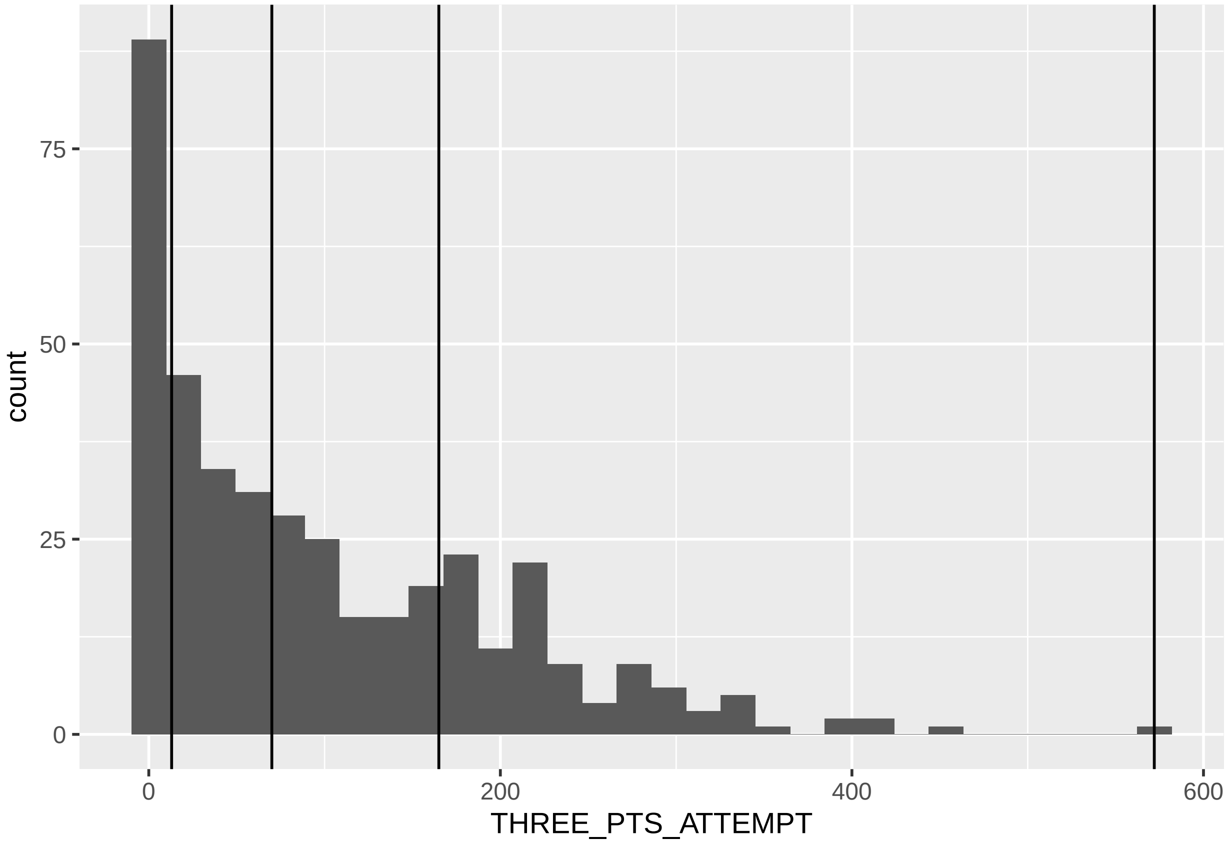
```
##   0%  25%  50%  75% 100%
##    0   13   70  165  572
```

```r
#could not remove the titles with row.names <- NULL so I stored them manually
quantiles <- c(0, 13, 70, 165, 572)

outliers_2 <- league_stats_100_min[which(league_stats_100_min[,13] - 13 <= (1.5*iqr_value) | league_sta

#Plot:
ggplot() +
  geom_histogram(data = league_stats_100_min, mapping = aes(x = THREE_PTS_ATTEMPT)) +
  geom_vline(xintercept = 13) +
  geom_vline(xintercept = 70) +
  geom_vline(xintercept = 165) +
  geom_vline(xintercept = 572)
```

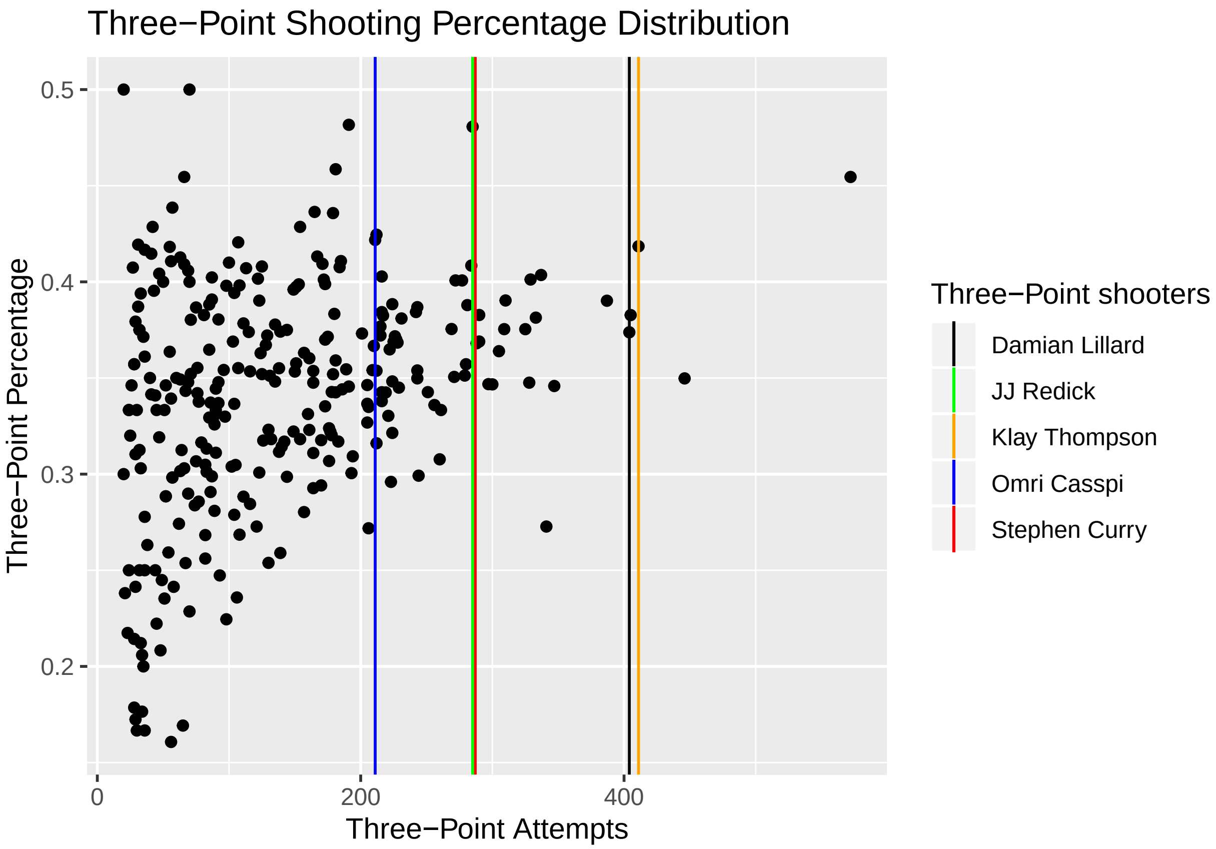
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



What makes a three pointer shooter good? Looking at three point shooting percentage won't tell the whole story. A player could shoot a high percentage, but only take shots where there aren't any defenders nearby to contest. It's more impressive to maintain a high percentage while also attempting a lot of three point shots. To tease this apart, **I plot three point shooting percentage versus number of three point shot attempts for players in the league.** I filter out players who have attempted too few three pointers (less than 20) to be worth including. Five notable three point shooters are: Stephen Curry, JJ Redick, Omri Casspi, Damian Lillard, and Klay Thompson.

```
league_stats_3pt <- mutate(filter(league_stats, THREE_PTS_ATTEMPT >= 20), THREE_PT_PERCENTAGE = THREE_P

ggplot(data = league_stats_3pt, aes(x = THREE_PTS_ATTEMPT, y = THREE_PT_PERCENTAGE)) +
  geom_point() +
  labs(title="Three-Point Shooting Percentage Distribution") +
  labs(x="Three-Point Attempts", y="Three-Point Percentage") +
  geom_vline(aes(xintercept = 287, color = "Stephen Curry")) +
  geom_vline(aes(xintercept = 285, color = "JJ Redick")) +
  geom_vline(aes(xintercept = 211, color = "Omri Casspi")) +
  geom_vline(aes(xintercept = 404, color = "Damian Lillard")) +
  geom_vline(aes(xintercept = 411, color = "Klay Thompson")) +
  scale_colour_manual(name='Three-Point shooters', values=c('Stephen Curry'='red', 'JJ Redick'='green',
```



Stephen Curry's historic and ongoing season

In the previous section, it is clear that Stephen Curry, a guard for the Golden State Warriors, shoots three pointers at a very high percentage and attempts many shots. In fact, he's having a historically good season. Let's compare his current season with some of the previous best three point shooting seasons for any player. ("Best" means making the most three pointers in a single season.) The data (`best_three_pt_season.txt`) consists of game-by-game three point shooting numbers for each season in tidy format. Each row corresponds to a game. The columns are:

- `NAME`: player name
- `SEASON`: season
- `GAME_NUM`: numbering of the games in the `SEASON` that `NAME` played, always consecutive starting at 1
- `DATE`: game date
- `TEAM`: `NAME`'s team
- `OPP`: opposing team for this game.
- `THREE_PT`: number of three pointers made
- `THREE_PT_ATTEMPT`: number of three pointers attempted
- `THREE_PT_PCT`: percentage for this game, i.e. `THREE_PT/THREE_PT_ATTEMPT`

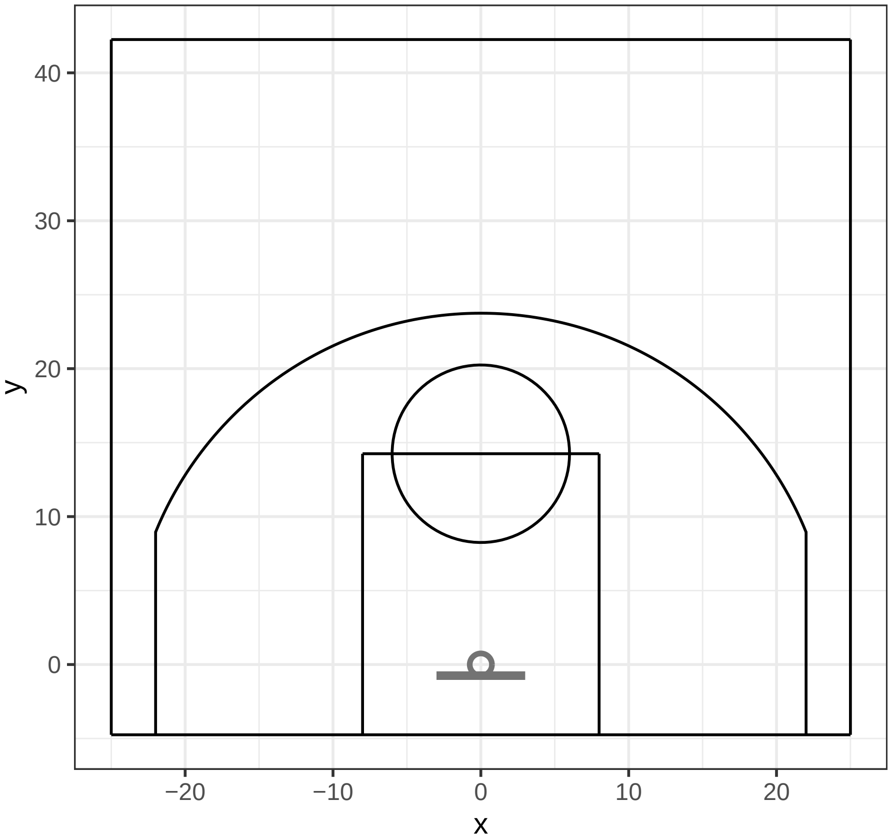
Here, I plot the number of three point shots made in the season as a function of the number of games played over the course of a season. I also compare Curry's 2015-2016 current rate to rates in previous historic seasons.

(The 94-95, 95-96, and 96-97 seasons are excluded from the list of top single season performances because the three-point line was closer to the basket for those three seasons.)

Stephen Curry's historic season (cont.)

The function `draw_court()` is called within a `ggplot2` object to draw a part of the NBA court.

```r
draw_court <- function() {
  #the y max of 9 is fudged a little so that the three pt arc is prettier.
  three_pt_straight1 <- data.frame(x=c(-22, -22), y=c(-4.75, 9))
  three_pt_straight2 <- data.frame(x=c(22, 22), y=c(-4.75, 9))
  theta <- seq(-1.184513, 1.184513, length.out=500)
  three_pt_arc <- data.frame(x=23.75*sin(theta), y=23.75*cos(theta))
  key_straight1 <- data.frame(x=c(-8, -8), y=c(-4.75, 14.25))
  key_straight2 <- data.frame(x=c(8, 8), y=c(-4.75, 14.25))
  key_straight3 <- data.frame(x=c(8, -8), y=c(14.25, 14.25))
  backboard <- data.frame(x=c(-3,3), y=c(-0.75, -0.75))
  circle_theta <- seq(0, 2*pi, length.out=100)
  ft_circle <- data.frame(x=6*cos(circle_theta), y=6*sin(circle_theta)+14.25)
  baseline <- data.frame(y=c(-4.75,-4.75), x=c(-25,25))
  sideline1 <- data.frame(x=c(25, 25), y=c(-4.75, 42.25))
  sideline2 <- data.frame(x=c(-25, -25), y=c(-4.75, 42.25))
  half <- data.frame(x=c(-25, 25), y=c(42.25, 42.25))
  basket <- data.frame(x=0.75*cos(circle_theta), y=0.75*sin(circle_theta))
  list(geom_line(aes(x=x, y=y), data=three_pt_straight1),
    geom_line(aes(x=x, y=y), data=three_pt_straight2),
    geom_line(aes(x=x, y=y), data=three_pt_arc),
    geom_line(aes(x=x, y=y), data=baseline),
    geom_line(aes(x=x, y=y), data=key_straight1),
    geom_line(aes(x=x, y=y), data=key_straight2),
    geom_line(aes(x=x, y=y), data=key_straight3),
    geom_path(aes(x=x, y=y), data=ft_circle),
    geom_line(aes(x=x, y=y), data=backboard, size=1.5, color="grey45"),
    geom_path(aes(x=x, y=y), data=basket, size=1, color="grey45"),
    geom_line(aes(x=x, y=y), data=sideline1),
    geom_line(aes(x=x, y=y), data=sideline2),
    geom_line(aes(x=x, y=y), data=half))
}
ggplot() + draw_court() + coord_fixed(ratio=1) + theme_bw()
```



The file `curry.csv` contains data on all the shots Curry has taken this season. The columns`LOC_X` and `LOC_Y` are the positions of each shot in feet, with the center of the basket as the origin. **Here, I plot all of his shots for this season on the court and color the shots by whether or not he made them.**

Sometimes Curry takes shots from extemely challenging locations, for example, this shot (YouTube link). This type of shot is called a *pullup jump shot* (the shooter stops and takes the shot after dribbling forward). **On a new figure, I plot the shots that are categorized as pullup shots on the court.**

Regular season LeBron versus playoff LeBron

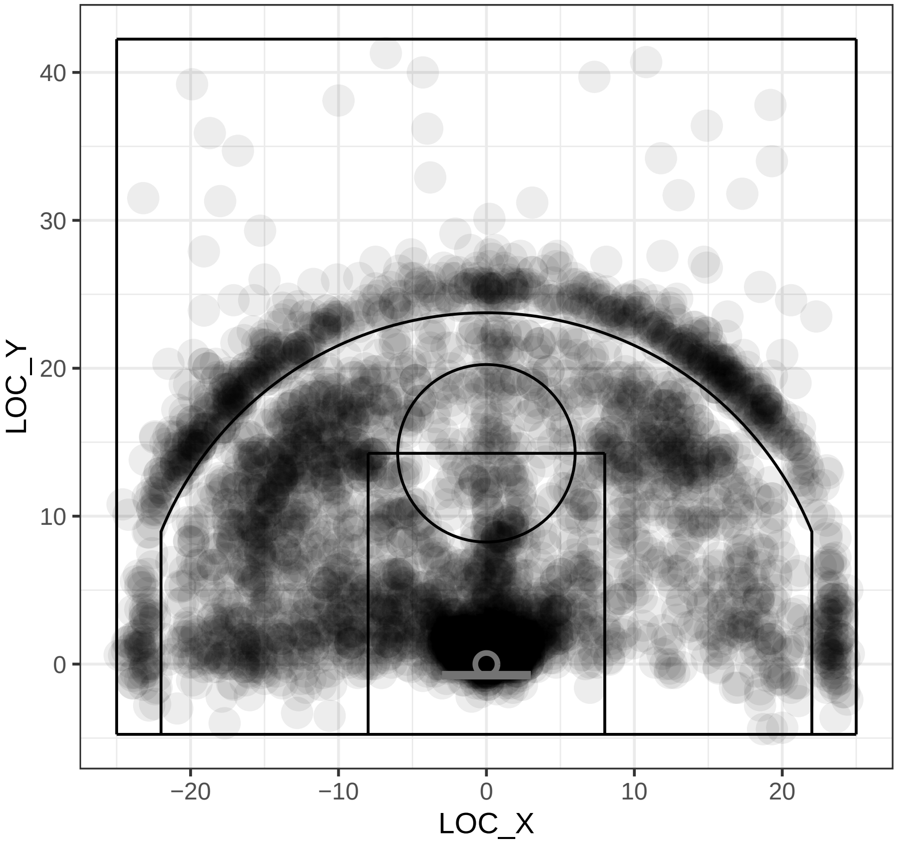
LeBron James is one of the most dominant players of the modern era. He's talented, smart, and athletic, and spectators have come to expect greatness from him. He's been involved in a number of spectacular playoff moments, e.g. YouTube link 1, YouTube link 2, and YouTube link 3. It's of interest to compare his regular season and playoff performances.

The `lebron_regular_season.txt` and `lebron_playoffs.txt`. files contain his shooting data between 2010 and 2014 for the regular season and playoffs, respectively.

```
lebron_regular <- read.table("/Users/Theo/Desktop/SML_201/Projects/project_2/data/lebron_regular_season
                             header = TRUE, sep = "\t")
lebron_regular_loc <- subset(lebron_regular, select = c(LOC_X,LOC_Y))
lebron_playoffs <- read.table("/Users/Theo/Desktop/SML_201/Projects/project_2/data/lebron_playoffs.txt"
                             header = TRUE, sep = "\t")
lebron_playoffs_loc <- subset(lebron_playoffs, select = c(LOC_X,LOC_Y))
```

Here, I plot a 2-D distribution on the court of where he attempts shots in the regular season.

```
ggplot(data = lebron_regular_loc, mapping = aes(x = LOC_X, y = LOC_Y)) +
  geom_point(alpha = 0.07, size = 5) +
  draw_court() +
  coord_fixed(ratio=1) +
  theme_bw()
```



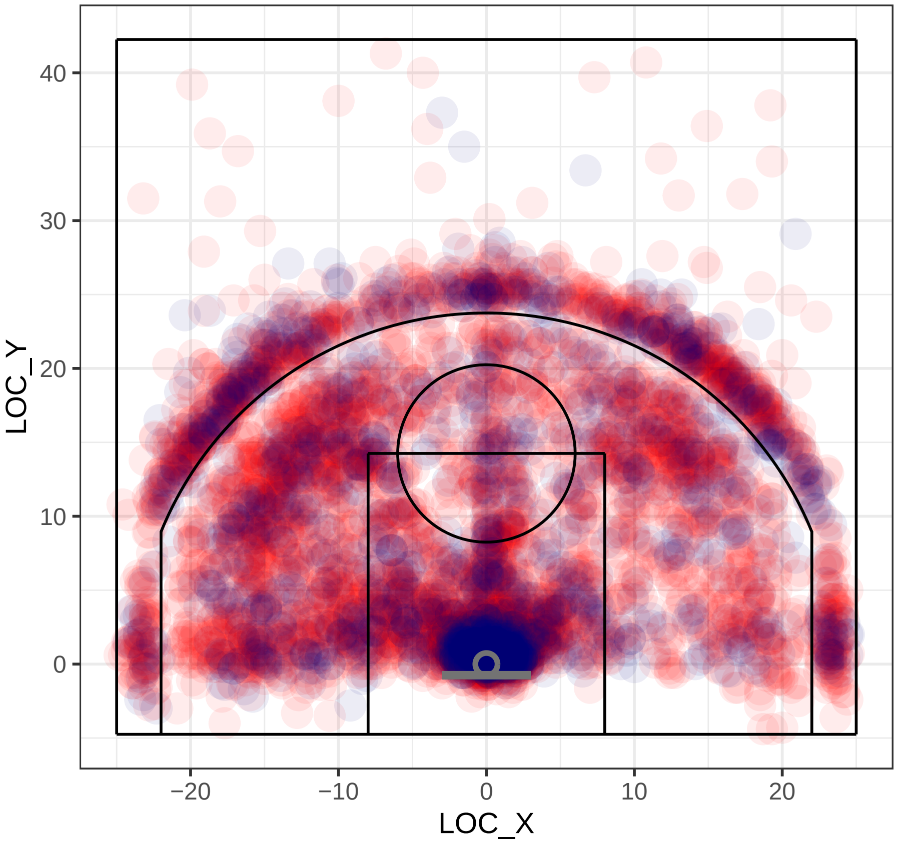
I chose a scatterplot because it is an effective way to create a heatmap of coordinates.

Finally, **I use plotting to compare LeBron's regular season and playoff performances in terms of where he takes his shots.**

```
#Overlaid to show the higher concentration of the blue (playoffs) close to the hoop. On the contrary, t

lebron_regular_smooth <-filter(lebron_regular, LOC_X <= -22 | LOC_X >= 22, LOC_Y >= 10)


ggplot() +
  geom_point(data = lebron_regular_loc, mapping = aes(x = LOC_X, y = LOC_Y), color = "red", alpha = 0.0
  geom_point(data = lebron_playoffs_loc, mapping = aes(x = LOC_X, y = LOC_Y), color = "navy", alpha = 0
  draw_court() +
  coord_fixed(ratio=1) +
  theme_bw()
```



Session Information

Session information always included for reproducibility!

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Mojave 10.14.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] hexbin_1.28.0   stringr_1.4.0   babynames_1.0.0 ggplot2_3.2.1
## [5] dplyr_0.8.3     reshape2_1.4.3  knitr_1.26
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3       magrittr_1.5     munsell_0.5.0    tidyselect_0.2.5
##  [5] lattice_0.20-38  colorspace_1.4-1 R6_2.4.1         rlang_0.4.2
##  [9] plyr_1.8.5       tools_3.6.2      grid_3.6.2       gtable_0.3.0
## [13] xfun_0.11        withr_2.1.2      htmltools_0.4.0  lazyeval_0.2.2
## [17] yaml_2.2.0       digest_0.6.23    assertthat_0.2.1 lifecycle_0.1.0
## [21] tibble_2.1.3     crayon_1.3.4     farver_2.0.1     purrr_0.3.3
## [25] glue_1.3.1       evaluate_0.14    rmarkdown_2.0    labeling_0.3
## [29] stringi_1.4.3    compiler_3.6.2   pillar_1.4.3     scales_1.1.0
## [33] pkgconfig_2.0.3
```