NBA 3 Pointer

Josh Kong

6/17/2020

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------- t

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------------------------- tidyve
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
nba_shot_log <- read_csv("shot_logs.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   MATCHUP = col_character(),
##   LOCATION = col_character(),
##   W = col_character(),
##   GAME_CLOCK = col_time(format = ""),
##   SHOT_RESULT = col_character(),
##   CLOSEST_DEFENDER = col_character(),
##   player_name = col_character()
## )

## See spec(...) for full column specifications.
```

About the dataset

```
head(nba_shot_log)
```

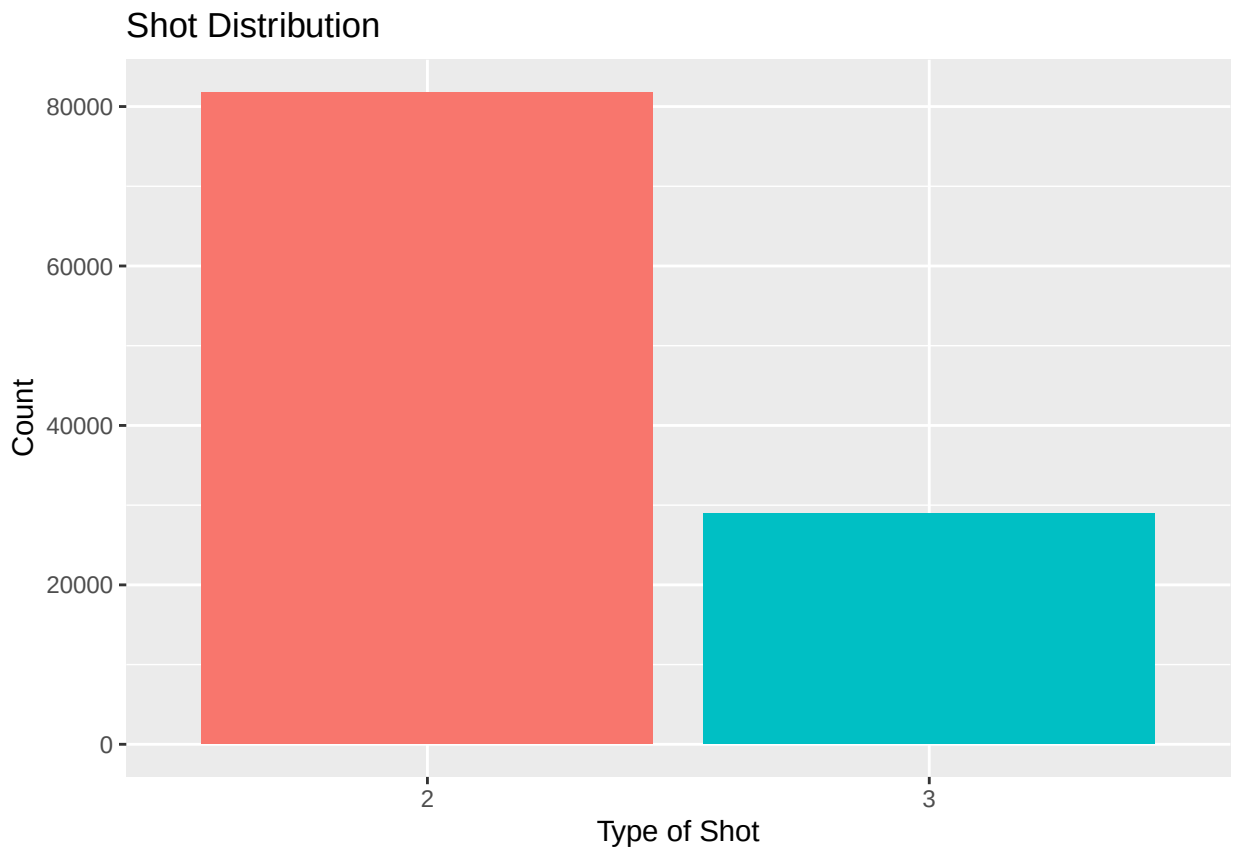
```
## # A tibble: 6 x 21
##   GAME_ID MATCHUP LOCATION W     FINAL_MARGIN SHOT_NUMBER PERIOD GAME_CLOCK
##     <dbl> <chr>   <chr>    <chr>        <dbl>       <dbl>  <dbl> <time>
## 1  2.14e7 MAR 04~ A        W               24           1      1 01:09
```

```
## 2  2.14e7 MAR 04~ A        W               24           2      1 00:14
## 3  2.14e7 MAR 04~ A        W               24           3      1 00:00
## 4  2.14e7 MAR 04~ A        W               24           4      2 11:47
## 5  2.14e7 MAR 04~ A        W               24           5      2 10:34
## 6  2.14e7 MAR 04~ A        W               24           6      2 08:15
## # ... with 13 more variables: SHOT_CLOCK <dbl>, DRIBBLES <dbl>,
## #   TOUCH_TIME <dbl>, SHOT_DIST <dbl>, PTS_TYPE <dbl>, SHOT_RESULT <chr>,
## #   CLOSEST_DEFENDER <chr>, CLOSEST_DEFENDER_PLAYER_ID <dbl>,
## #   CLOSE_DEF_DIST <dbl>, FGM <dbl>, PTS <dbl>, player_name <chr>,
## #   player_id <dbl>
```

This dataset shows all the shots that were taken in the 2014-2015 nba season. It includes many different variables such as the point type, shot distance, closest defender, location, and more.

Distribution of two point shots and three point shots.

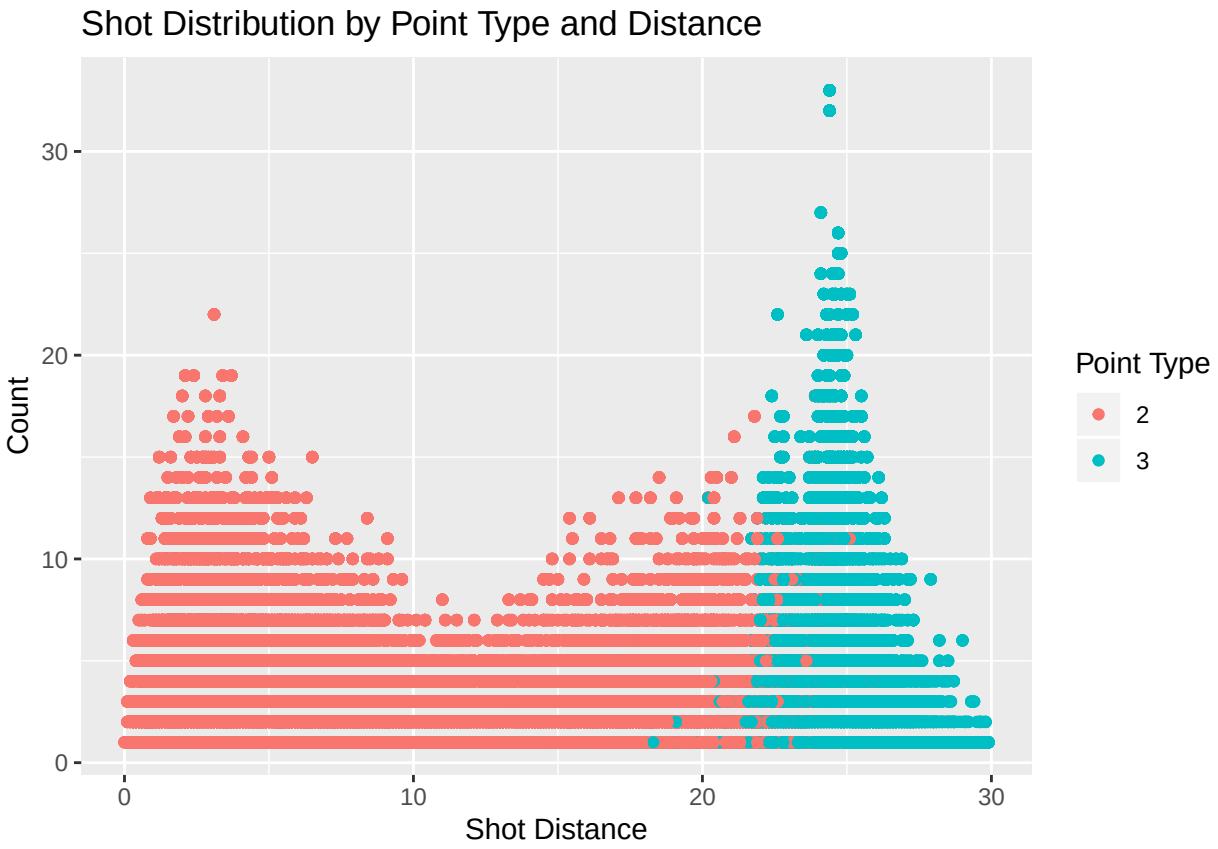
```
#filter out players who have shot less than 300 shots
nba_shot_log <- nba_shot_log %>%
  group_by(player_name) %>%
  filter(n() >= 300)

#looking at the distribution of shots between 2 pointers and 3 pointers
nba_shot_log %>%
  group_by(PTS_TYPE) %>%
  count(PTS_TYPE) %>%
  ggplot(aes(as.factor(PTS_TYPE),n))+ geom_col(aes(fill = as.factor(PTS_TYPE))) + labs(x = "Type of Sho
  theme(legend.position = "none")
```



This is the distribution of the number of 2 point shots and 3 point shots taken.

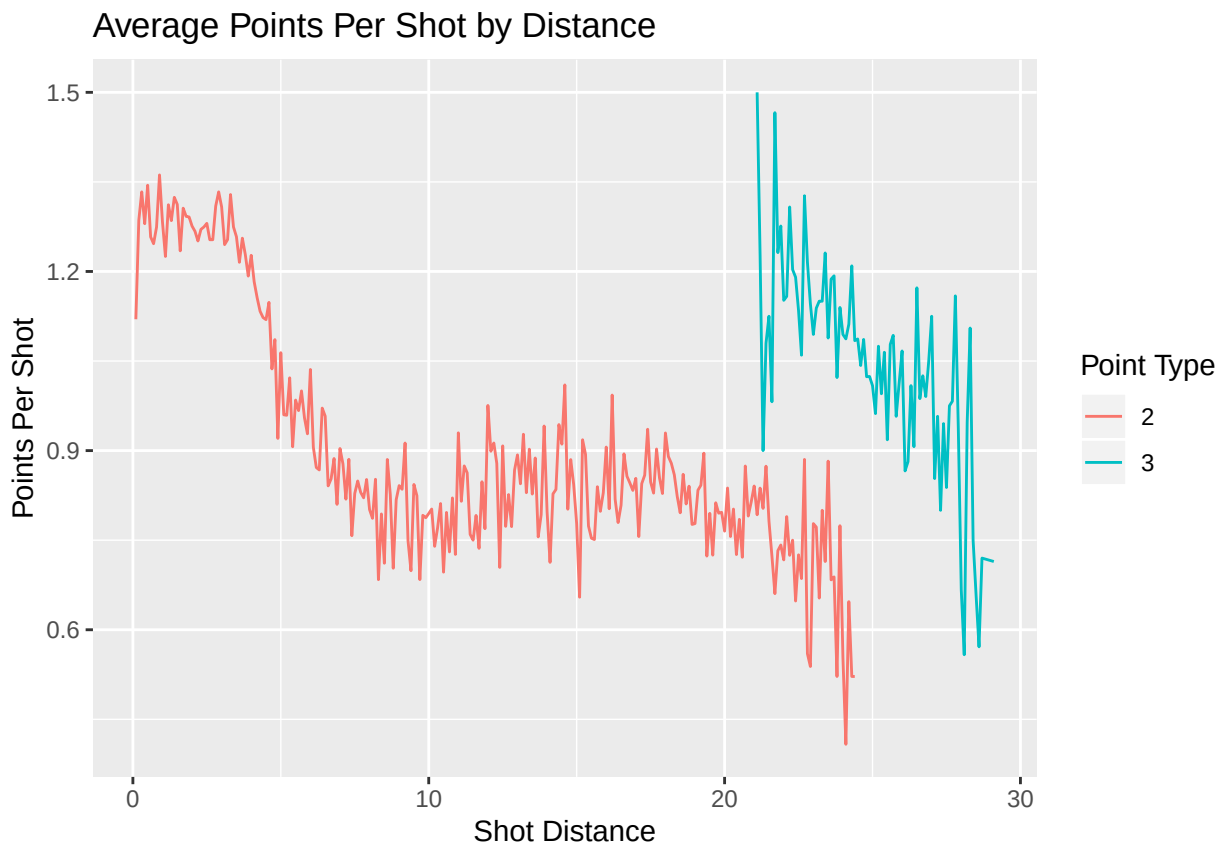
NOTE: I am going to refer to any shot that is more than 5 feet away from the basket and not a three point shot as a midrange shot.

```
# We want to look at the distribution of shots in the NBA according to the distance
nba_shot_log %>%
  add_count(SHOT_DIST) %>%
  filter(SHOT_DIST < 30) %>%
  ggplot(aes(SHOT_DIST,n)) +geom_point(aes(color = as.factor(PTS_TYPE))) +
  labs(x = "Shot Distance", y= "Count", color = "Point Type", title = "Shot Distribution by Point Type a
```



Looking at the shot distribution, it appears that the shots from 10 feet to 20 feet are not very frequent. There is a noticeable difference in the number of midrange shots compared to 3 point shot. It shows that the midrange shot in the NBA is dying. The NBA is slowly becoming a game where the player either shoots a three pointer, or shoots a shot in the paint.

```
#we want to look compare the average points per distance.
#begin by making the shot result into 1 if it's a make, and 0 if it's a miss
shot_res <- rep(NA,nrow(nba_shot_log))
for (i in 1: nrow(nba_shot_log))
{
  if(nba_shot_log$SHOT_RESULT[i] == "made"){
    shot_res[i] <- 1
  }
  else
  {
    shot_res[i] <- 0
  }
}
nba_shot_log$shot_res <- shot_res

nba_shot_log %>%
  group_by(SHOT_DIST,PTS_TYPE) %>%
  filter(n() > 20) %>%
  summarize(points_per_shot = mean(PTS_TYPE*shot_res)) %>%
  ggplot(aes(x = SHOT_DIST,y = points_per_shot)) + geom_line(aes(color = factor(PTS_TYPE))) +
  labs(x = "Shot Distance", y = "Points Per Shot", color = "Point Type",
       title = "Average Points Per Shot by Distance")
```




```
nba_shot_log %>%
  group_by(PTS_TYPE,shot_res) %>%
  summarise(total = n()) %>%
  ungroup() %>%
  group_by(PTS_TYPE) %>%
  summarise(pts_per_shot = sum(PTS_TYPE * shot_res * total)/sum(total))
```

```
## # A tibble: 2 x 2
##   PTS_TYPE pts_per_shot
##      <dbl>        <dbl>
## 1        2        0.977
## 2        3        1.06
```

In our analysis, I am going to be looking at the points per shot. Points per shot formula is:

(make % at the distance) * (Point type) / (total shots taken).

Looking at the Points Per Shot based on the shot distance, it is clear why many teams prefer either a shot in the paint or a three pointer, rather than shooting a midrange shot. While a shot within 5 feet of the basket yields similar amounts of points as a three pointer, 2 point shots within 5 to 20 feet from the basket yields significantly less points than a three pointer and a shot in the paint.

Looking at the points per shot of a three pointer vs a two pointer, a three point shot yields 0.083 more points per shot than a 2 pointer. Next, I am going to compare the points per shot of a midrange shot (2 point shots that are at least 5 feet away from the basket) and a three pointer.

```
nba_shot_log %>%
  filter(SHOT_DIST > 5) %>%
  group_by(PTS_TYPE,shot_res) %>%
  summarise(total = n()) %>%
  ungroup() %>%
  group_by(PTS_TYPE) %>%
  summarise(pts_per_shot = sum(PTS_TYPE * shot_res * total)/sum(total))
```

```
## # A tibble: 2 x 2
##   PTS_TYPE pts_per_shot
##      <dbl>        <dbl>
## 1        2        0.835
## 2        3        1.06
```

Comparing the points per shot of a midrange shot and a three pointer, it is clear why teams steer away from the midrange shot, and look to shoot a three pointer instead. A three point shot yields a significantly more points than a midrange shot.

Moving the three point line back to create a more balanced game.

I believe that we should move the three point line back so that the difference in expected point per shot of a midrange shot and a three point shot is 0.1. This way, the midrange shot will not be as undervalued as it currently is, and the game will begin to be more balanced where players shoot equally from all three parts of the floor.

The first step I did was to make all the different areas of the three point shot equidistant from the basket. Originally, the corner shot is only 22 feet away from the basket, while the three point shot at the top of the key is 23.75 feet away from the basket. My first adjustment was to make all three point shots 23.75 feet away from the basket, and any shots closer than that will be considered a two point shot. Doing so eliminates the corner three point shot, which is arguably the most valuable shot in the game (due to it being the 3 point shot that is closest to the basket).

```
new_pt_type <- rep(NA,nrow(nba_shot_log))
for(i in 1:nrow(nba_shot_log))
{
  if(nba_shot_log$SHOT_DIST[i] > 23.75)
  {
    new_pt_type[i] <- 3
  }
  else
  {
    new_pt_type[i] <- 2
  }
}

nba_shot_log$new_pt_type <- new_pt_type
  nba_shot_log %>%
  filter(SHOT_DIST > 5) %>%
  group_by(new_pt_type,shot_res) %>%
  summarise(total = n()) %>%
  ungroup() %>%
  group_by(new_pt_type) %>%
  summarise(pts_per_shot = sum(new_pt_type * shot_res * total)/sum(total))
```

```
## # A tibble: 2 x 2
##   new_pt_type pts_per_shot
##         <dbl>        <dbl>
## 1           2        0.829
## 2           3        1.02
```

After getting making all the three point shots equidistant, the difference in the points per shot of a midrange and a three point shot still remains around 0.2. My next step is to keep moving the three point line back in increments of 0.05 feet, until the difference in the points per shot of a midrange shot and a three pointer is only 0.1.

```
counter <- 1
while(TRUE)
{
  nba_three_diff <- nba_shot_log %>%
  filter(SHOT_DIST > 5) %>%
  group_by(new_pt_type,shot_res) %>%
  summarise(total = n()) %>%
  ungroup() %>%
  group_by(new_pt_type) %>%
  summarise(pts_per_shot = sum(new_pt_type * shot_res * total)/sum(total))
  if(diff(nba_three_diff$pts_per_shot) < 0.1)
  {
    break
  }
  else
  {
    new_pt_type <- rep(NA,nrow(nba_shot_log))
    for(i in 1:nrow(nba_shot_log))
    {
     if(nba_shot_log$SHOT_DIST[i] > 23.75 + counter*.05)
      {
       new_pt_type[i] <- 3
     }
      else
      {
        new_pt_type[i] <- 2
      }
    }

    nba_shot_log$new_pt_type <- new_pt_type
  }

  if(counter > 100)
  {
    break
  }
  counter <- counter + 1
}
new_three_point_line <- 23.75 + counter*.05; new_three_point_line
```

```
## [1] 25.45
```

```
nba_shot_log %>%
  filter(SHOT_DIST > 5) %>%
  group_by(new_pt_type,shot_res) %>%
  summarise(total = n()) %>%
  ungroup() %>%
  group_by(new_pt_type) %>%
  summarise(pts_per_shot = sum(new_pt_type * shot_res * total)/sum(total))
```

```
## # A tibble: 2 x 2
##   new_pt_type pts_per_shot
##         <dbl>        <dbl>
## 1           2        0.804
## 2           3        0.903
```

Moving the three point line back from 23.75 feet to 25.45 feet decreases the difference between the points per shot of a midrange shot and a three pointer to 0.1. Although there is still a significant difference between the points per shot between these two shots, moving the three point line back has decreased the difference by a lot, which in turn will make the midrange shot more viable.

How does moving the three point line affect the top ten shooters in the league??

I determined the top ten shooters by looking at the ten players with the highest three point percentage that have taken a minimum of 200 three point shots. This means they're taking a minimum of 2.5 three pointers a game.

```
#we are now going to see how those same ten players shoot with the new three point line
nba_shooters_filtered <- nba_shot_log %>%
  filter(PTS_TYPE == 3) %>%
  add_count(player_name) %>%
  filter(n > 200)

top_shooters <- nba_shooters_filtered %>%
  group_by(player_name) %>%
  count(shot_res) %>%
  summarise(three_point_perc = sum(shot_res * n) / sum(n),
            num_of_shots = sum(n)) %>%
  arrange(desc(three_point_perc)) %>%
  mutate(player_name = fct_reorder(player_name, three_point_perc)) %>%
  head(10)

#Top Ten Shooters with regulation three point line
top_shooters
```

```
## # A tibble: 10 x 3
##    player_name     three_point_perc num_of_shots
##    <fct>                      <dbl>        <int>
##  1 kyle korver                0.496          355
##  2 klay thompson              0.435          407
##  3 jj redick                  0.426          329
##  4 stephen curry              0.417          456
##  5 aaron brooks               0.408          218
```

```
##  6 kyrie irving               0.404          297
##  7 demarre carroll            0.399          223
##  8 kyle singler               0.399          208
##  9 brandon knight             0.397          300
## 10 marcus morris              0.395          220
```

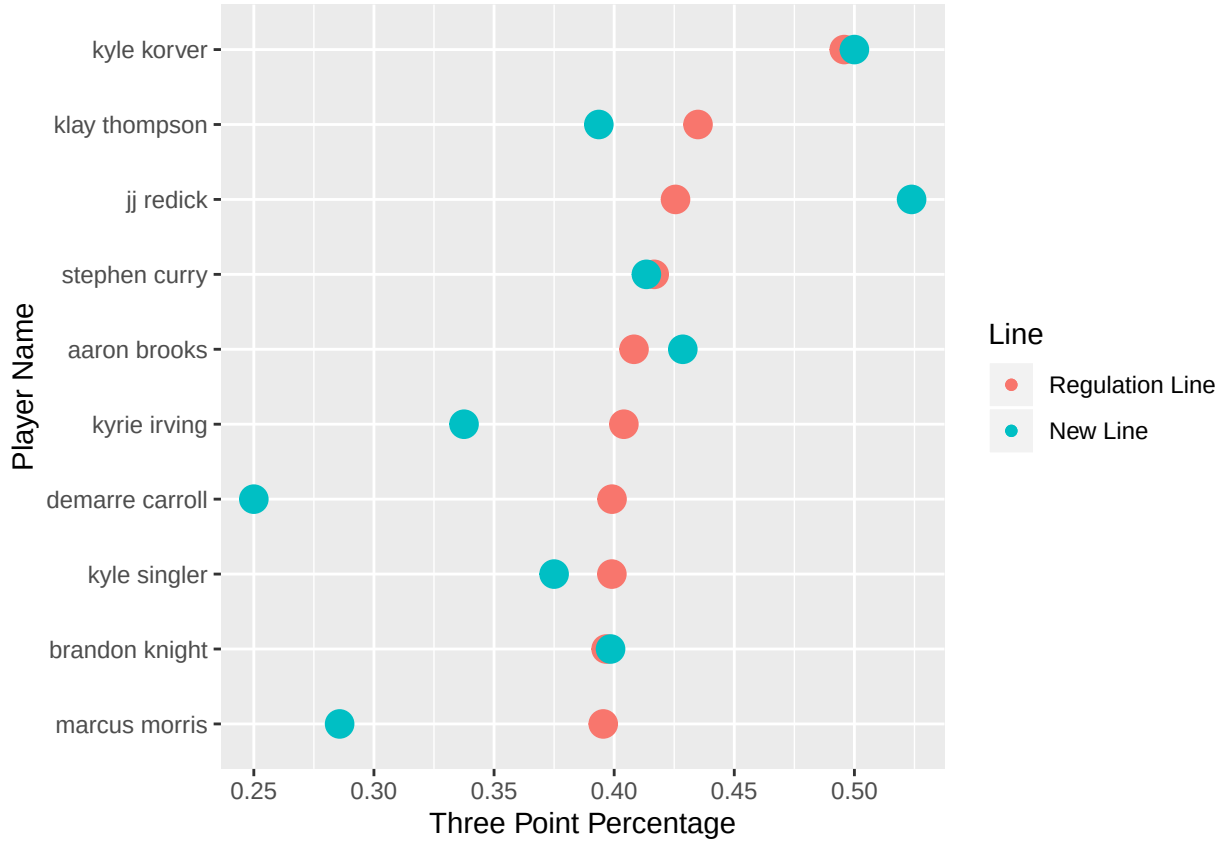
```
top_ten <- top_shooters$player_name
top_ten_shooters <- nba_shot_log[which(nba_shot_log$player_name %in% top_ten),]
top_ten_shooters <- top_ten_shooters %>%
  filter(new_pt_type == 3) %>%
  group_by(player_name) %>%
  count(shot_res) %>%
  summarise(three_point_perc = sum(shot_res * n) / sum(n),
            num_of_shots = sum(n)) %>%
  arrange(desc(three_point_perc)) %>%
  mutate(player_name = fct_reorder(player_name, three_point_perc))

#The shooting statistics of the top ten shooters with the new three point line
top_ten_shooters
```

```
## # A tibble: 10 x 3
##    player_name     three_point_perc num_of_shots
##    <fct>                      <dbl>        <int>
##  1 jj redick                  0.524           42
##  2 kyle korver                0.5             72
##  3 aaron brooks               0.429          112
##  4 stephen curry              0.413          179
##  5 brandon knight             0.398          133
##  6 klay thompson              0.394           94
##  7 kyle singler               0.375           16
##  8 kyrie irving               0.338           80
##  9 marcus morris              0.286           21
## 10 demarre carroll            0.25             4
```

```
top_ten_shooters <- top_ten_shooters %>%
  inner_join(top_shooters, by ="player_name") %>%
  arrange(desc(three_point_perc.y)) %>%
  mutate(player_name = fct_reorder(player_name, three_point_perc.y))
```

```
## Warning: Column `player_name` joining factors with different levels, coercing to
## character vector
```

```
top_ten_shooters %>%
  ggplot() +
  geom_point(aes(x = player_name, y = three_point_perc.y, color = "blue",size = 2))+
  geom_point(aes(x = player_name, y = three_point_perc.x, color = "red",size = 2)) + coord_flip() +
  labs(x = "Player Name", y = "Three Point Percentage")+
  scale_color_discrete(name = "Line", labels = c("Regulation Line", "New Line")) +
  guides(size = FALSE)
```




```
top_ten_shooters
```

```
## # A tibble: 10 x 5
##    player_name  three_point_perc~ num_of_shots.x three_point_per~ num_of_shots.y
##    <fct>                    <dbl>          <int>            <dbl>          <int>
##  1 kyle korver              0.5               72            0.496            355
##  2 klay thomps~             0.394             94            0.435            407
##  3 jj redick                0.524             42            0.426            329
##  4 stephen cur~             0.413            179            0.417            456
##  5 aaron brooks             0.429            112            0.408            218
##  6 kyrie irving             0.338             80            0.404            297
##  7 demarre car~             0.25               4            0.399            223
##  8 kyle singler             0.375             16            0.399            208
##  9 brandon kni~             0.398            133            0.397            300
## 10 marcus morr~             0.286             21            0.395            220
```

```
mean(top_ten_shooters$three_point_perc.x)
```

```
## [1] 0.3906116
```

```
mean(top_ten_shooters$three_point_perc.y)
```

```
## [1] 0.4175423
```

Looking at the top ten three point shooters, on average most shooters see a dip in their three point percentage as the three point line gets moved back (as expected). However, there are a some players whose three point percentage remains the same, or even goes up as the three point line gets extended.

The average three point percentage amongst the top ten shooters goes from 41.7% to 39%.

There does appear to be a significant drop in the three point percentage as the three point line gets moved back.