Data Story

Curtis Higa

January 22, 2018

Introduction

In 1979, the National Basketball Association introduced a new component to their game that would take points how points are scored away from the rim, the three point shot. Since then we've seen a plethora of players score a majority of their points from behind the arc like Reggie Miller and Ray Allen. But does this shift change the way the game is played? Do you have a better chance of winning a game if you outscore your opponent in three-point shots? What about for the season? With teams like the Golden State Warriors and Houston Rockets, it may seem that your teams ability to convert three-point shots would equate to more wins but to verify this, data was taken from the 2016-2017 NBA regular season.

Data Wrangling

This data was taken off a website called MySportsFeeds where they house countless sports statistics for baseball, hockey, football, and basketball. For this project, I've pulled their game log data on the 2016-2017 NBA regular season. From their data set, I've removed columns such as rebounds and assists along with adding columns for three point percentage, total points, and percentage of points contributed by three-pointers. The total points column was particularly important to identifying who was the winner of each game and indicating it within the data set. It was done by mutating a column that compared the total points of the team in question with the winning teams score. Depending if the total points equaled, a "1" or "0" was added to signify if the team won. Once those statistics were added to the data set, I appended columns that coincided with their opponents statistics for each of their games and the difference between their statistics to their opponents.

Modeling

Over the 1,230 games played throughout the season, the difference between numerous three-point statistics of the winning and losing teams were plotted to visualize any patterns. They were as follows:

```
game_filter <- complete_gamelogs %>% filter(win_team == 1)
geom_hist <- geom_histogram(aes(y = ..count..), position = position_dodge(), binwidth = 1)

game_filter %>%
  ggplot(aes(x = diff_tpm)) +
  geom_hist +
  scale_x_continuous(breaks = round(seq(-12,16, by = 1),1)) +
  labs(x = "Difference in Three-Pointers Made", y = "Count")
```
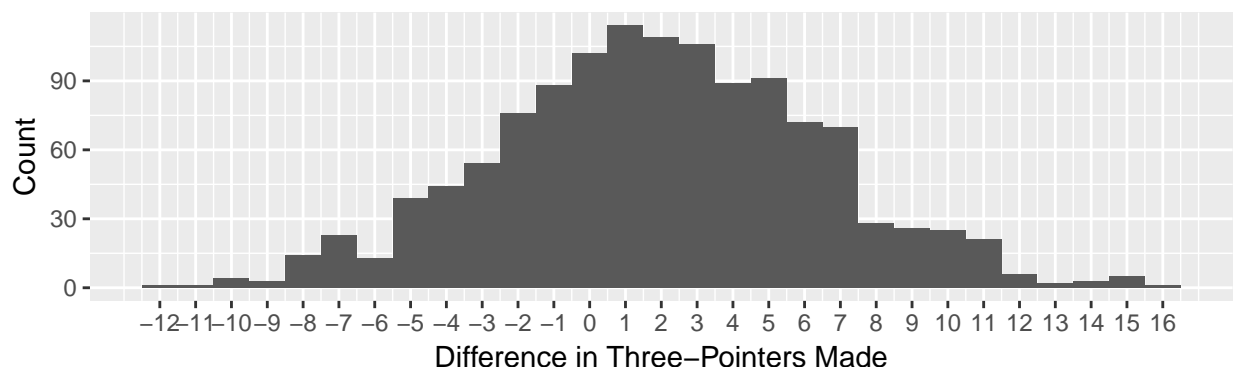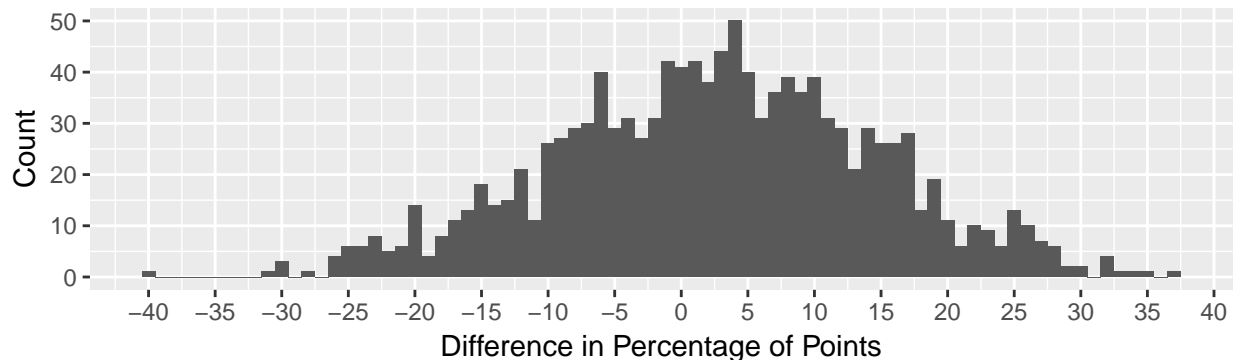
```
game_filter %>%
  ggplot(aes(x = diff_tpp)) +
  geom_hist +
  scale_x_continuous(breaks = round(seq(-70,70, by = 5),1)) +
  labs(x = "Difference in Three-Point Percentage", y = "Count")
```
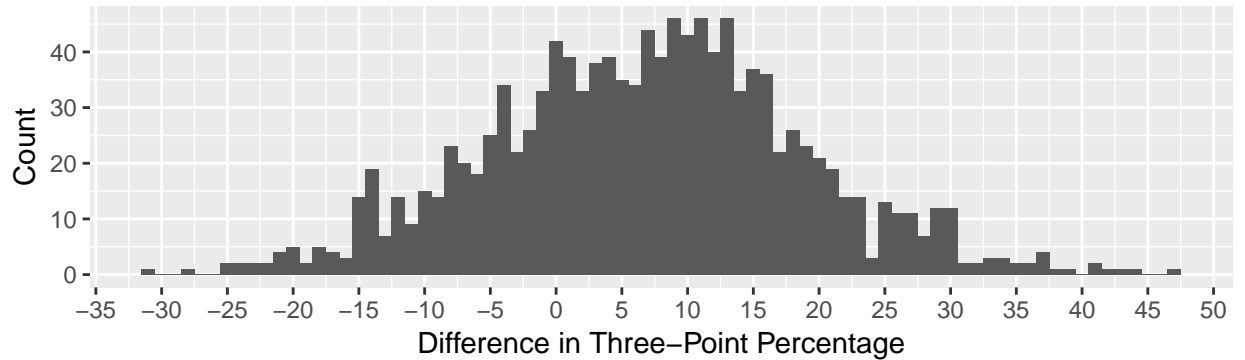


```
game_filter %>%
  ggplot(aes(x = diff_pop)) +
  geom_hist +
  scale_x_continuous(breaks = round(seq(-40,40, by = 5),1)) +
  labs(x = "Difference in Percentage of Points", y = "Count")
```


As the histograms illustrate, the winning team typically out performs their opponents is in all the three-point statistics looked. On average, the winning team makes 2 more three's, shoots threes better by around 7%, and has roughly 2.5% more of their points coming from threes. But looking at the amount of spread of these histograms there's a lot more to say about these distributions. Taking a look at the number of three-pointers made by winning teams compared to losing teams, numerous teams were out shot in three-pointers by 4 or more and were still able to win games. The same is true for three-point percentage and percentage of points. So when looking at a single game, although it could be a factor, being able to produce three's more often or with better efficiency doesn't signify winning a particular game.

However, if a team is able to excel at a couple three-point statistics, it could correlate to an above average winning percentage throughout the season. Splitting up the same data of the 2016-2017 NBA regular season into teams, a different pattern emerges when we plot these statistics with respect to the number of wins/win percentage. Using the *corrplot* in the *corrplot* library, we're able to visually see the correlation between different statistics.

```
team_statistics %>% ggplot(aes(x = win_team_sum, y = three_pt_Percentage_mean)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Wins", y = "Three Point Percentage(Avg)")
```

2
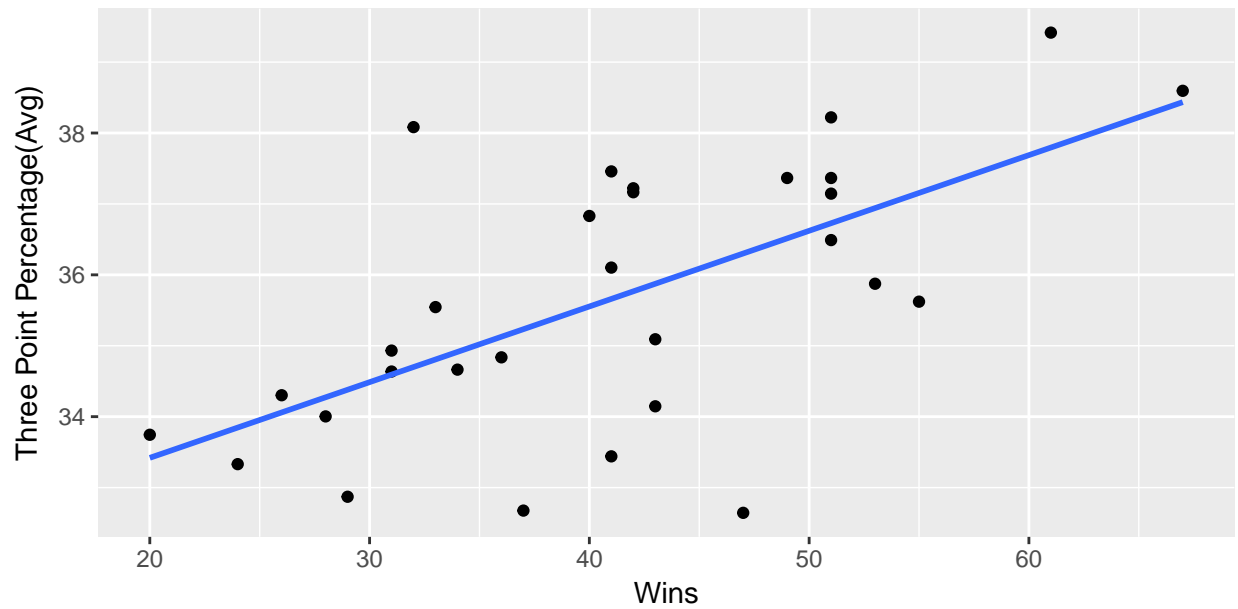
```
team_statistics %>% ggplot(aes(x = win_team_sum, y = opp_three_pt_Percentage_mean)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Wins", y = "Opponent Three Point Percentage(Avg)")
```
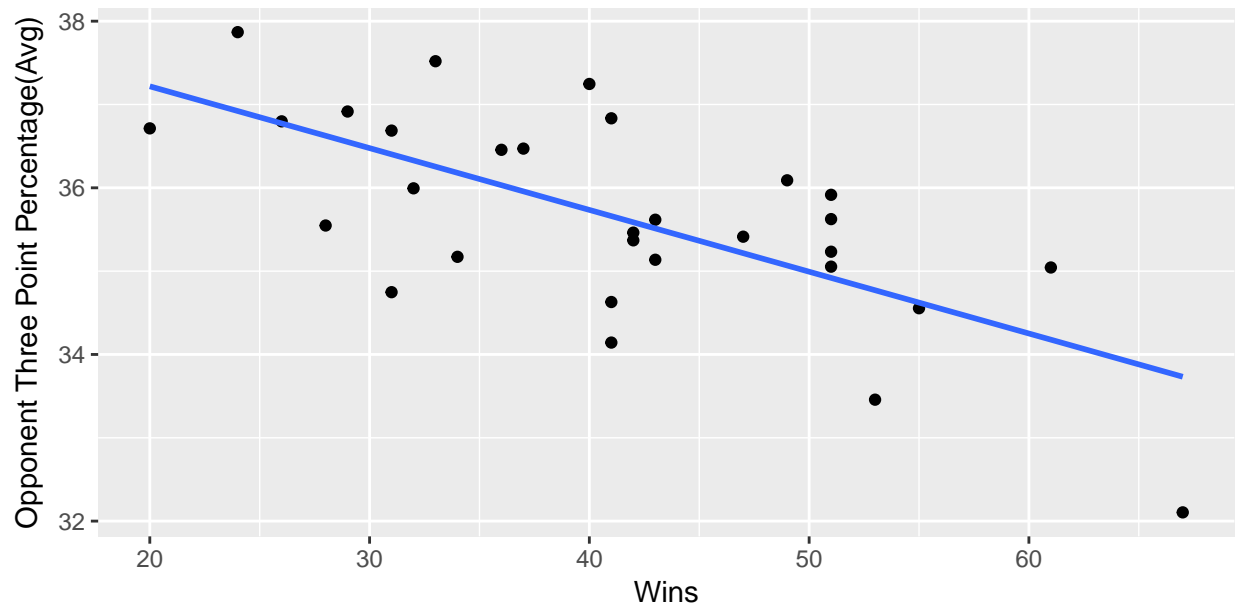


```
cor <- cor(team_statistics[,c("win_team_mean",
                              "three_pt_Made_mean",
                              "opp_three_pt_Made_mean",
                              "three_pt_Percentage_mean",
                              "opp_three_pt_Percentage_mean",
                              "Percent_of_Points_mean",
                              "opp_Percent_of_Points_mean")])

corrplot(cor, method="circle")
```
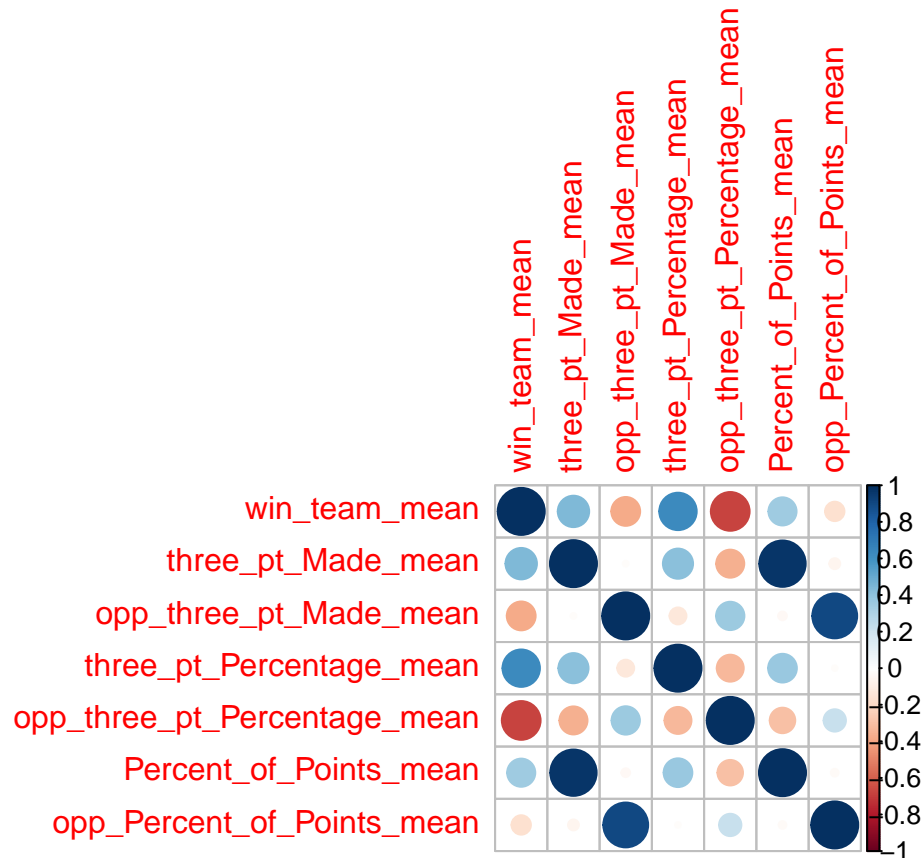
As shown in these graphs, there seems to be a positive correlation to the number of wins and a teams three-point percentage in addition to being a negative correlation to their opponents three point percentage. The last figure confirms that of the statistics, three point percentage and opponents three point percentage hold the strongest linear relationship to your teams win percentage by 0.632 and -0.678 respectively.

Summary

With that said, of the three-point statistics analyzed here, being able to defend and shoot behind the three-point line is a key factor if you want to win basketball games. However, this doesn't take into consideration your offensive and defensive philosophies and isn't the only factor that matters in basketball. On offense, being able to run a system of plays that get players open looks and easy shots will help your efficiency and outscore your opponents. The opposite is true for defense. Being able to disrupt and guard shots effectively will lower your opponents efficiency so they can't score as much. This doesn't just go for three point offense and defense but to all of basketball, which is why it makes sense that overall offensive and defensive efficiency factors into how often you win over the course of an entire season.