Using Data Science Methods to Characterize Trends in the NBA

Anh-Minh Nguyen, Benjamin Stan

Introduction

Though the core concept of basketball has remained the same over the 75 years of the NBA, the style of gameplay has shifted both organically and as a result of rule changes. For example, in the 2004-2005 season, the NBA implemented a rule banning hand checks, which forbade defensive players from placing one or both of their hands on an offensive player with the ball. In recent years, wider adoption of the three point shot has changed the way in which offenses operate. Three pointers increased from 17% of field goal attempts in the 1999-2000 season to 38% in 2019-2020.[1] With this, offenses can generate more points in fewer positions. One measure of this efficiency is true shooting percentage (TSP), which is defined using points (PTS) and true shooting attempts (TSA). TSA is calculated as follows: (field goal attempts)+0.44*(free throw attempts). TSP is then calculated as: PTS/(2*TSA).[2] As players are scoring more points on fewer shots due to three pointers, it would be expected that this value would increase along with the increase in three pointers attempted. We were interested in modeling the average true shooting percentage of players over the time span 1997 to 2021, with a particular emphasis on years corresponding to rule changes. These years and the changes that occurred are below:

- 2001: Rule changes preceding the 2001-2002 NBA season allowed the utilization of zone defenses, in which players are no longer required to defend an individual on the opposing team and can instead defend a space on the floor. To accompany this, a rule for defensive three seconds was added to prevent defensive players from occupying the area near the basket for an extended period.[3]
- 2004: Changes forbade the practice known as “hand-checking,” reducing the allowable contact by defenders on the perimeter and enabling freer offensive movement.[4]
- 2011: A rule change led to the “rip-through” move, in which an offensive player initiates contact and continues into a shooting motion, no longer resulting in a shooting foul.[5]
- 2018: Rule emphasis called the “freedom of movement” further reduced contact by defenders.[6]

A separate interest was in understanding the composition of successful teams following these changes in the game. Conceptually, championship teams are driven by superstars who score among the highest rates in the league. However, with the increase in three point attempts, it is possible that shooting specialists could play a larger role in the scoring output of their team. Our goal was to understand if greater scoring depth is a common characteristic among championship teams. We compared the points scored by the non-leading scorers on championship teams to those on non-championship teams. From this analysis, we hoped to glean a stronger understanding of the trends that surround scoring and roster construction.

In summary, our scientific questions of interest were:

- Model the trend in true shooting percentage over the timeframe 1997-2021 to better understand the observed pattern. We also hoped to identify the impact of certain change points, which correspond to changes in rules, to see if these led to measurable differences in efficiency between the year preceding and the year following.
- Determine whether the average points per game (PPG) of the 4th to 9th highest scorers on championship teams (those who played in the NBA Finals) equaled the average PPG of similarly ranked players on non-championship teams over the timeframe 2015-2021.

Data

Our primary dataset was the NBA Players dataset from Kaggle, which has season-long stats for all players from 1996 to 2021. It includes details such as basic stats (points, rebounds, assists) and player draft details (college, draft position), with each row corresponding to one player's data for a given season. Originally, the data was pulled from the NBA Stats API and any gaps in the data were supplemented by Basketball Reference.[2] The data appropriately addressed the problems of interest due to the thorough record of player performance and their associated team. For the question regarding true shooting percentage, the data does not contain each of the component variables for calculation, but it does contain the computed metric for each player. Thus, we considered the average TSP, weighing each player in the analysis equally.

Methods

To identify the trend in scoring efficiency, as measured through true shooting percentage, we considered the average TSP for each year from 1997 to 2021. This is the average of all players with a usage percentage above 10%. This metric estimates the percentage of team plays used by a player while they are on the floor, and this filtering removes 3.3% of the records.[2] We considered both polynomials (degrees 1-4) and spline functions (degrees 1-3) with knots defined by known change points in rules (2001, 2004, 2011, 2018). The best-fitting model was identified by evaluating mean squared error on a validation set. Training and validation sets were randomly defined by a 60/40 split stratified by year. We also used permutation testing to examine changes in mean TSP from the year immediately preceding a change to the year immediately following it. This process involved generating a distribution of the change in TSP from year to year by shuffling the year associated with players' TSP.

Identifying the depth of championship teams took a separate approach. For each season from 2015-2021, and for each team, we ordered players by PPG. We then computed the average PPG for players ranked 4-9 (which we denote as non-leading scorers) on all teams, labeling the players by whether or not their team played in the championship series for that given season. Following this, we evaluated the difference in means between the two groups. The next step was to determine whether this observed difference was due to random chance. Permutation test was appropriate, as we shuffled the label of whether or not a player was on a championship team. Note that we only shuffled labels within the same season. We calculated 10,000 permutation differences in means to use as our sampling distribution to test the null hypothesis that the mean PPG of non-leading scorers on championship teams equals the mean PPG of non-leading scorers on non-championship teams.

Results

For the true shooting percentage question, an initial analysis plotted the data points over the timespan of interest and visualized the various fits on the data. These are shown in Figure 1 and Figure 2.

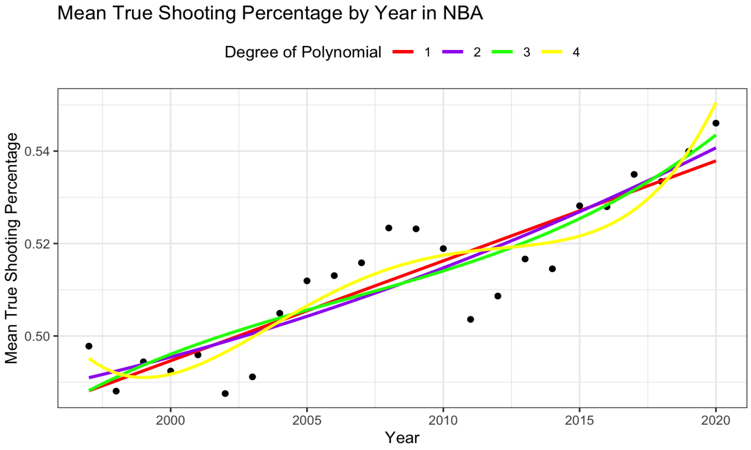

Figure 1. Polynomial fits on average TSP by year.

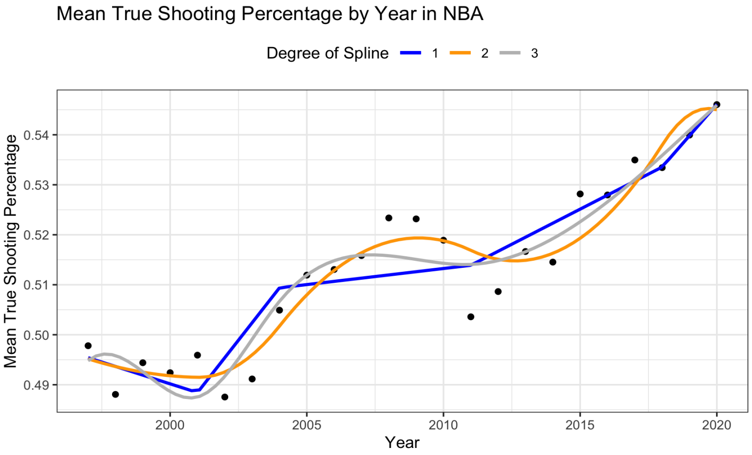

Figure 2. Spline fits with knots at 2001, 2004, 2011, and 2018 on average TSP by year.

The identification of the best-fitting model utilized training and validation sets stratified by year. The results in Figure 3 are displayed using degrees of freedom (df). For the polynomial fits, the df corresponds to the degree of the polynomial plus one. For the spline fits, the df corresponds to the degree of the spline plus five, as the model includes four knots and an intercept term.

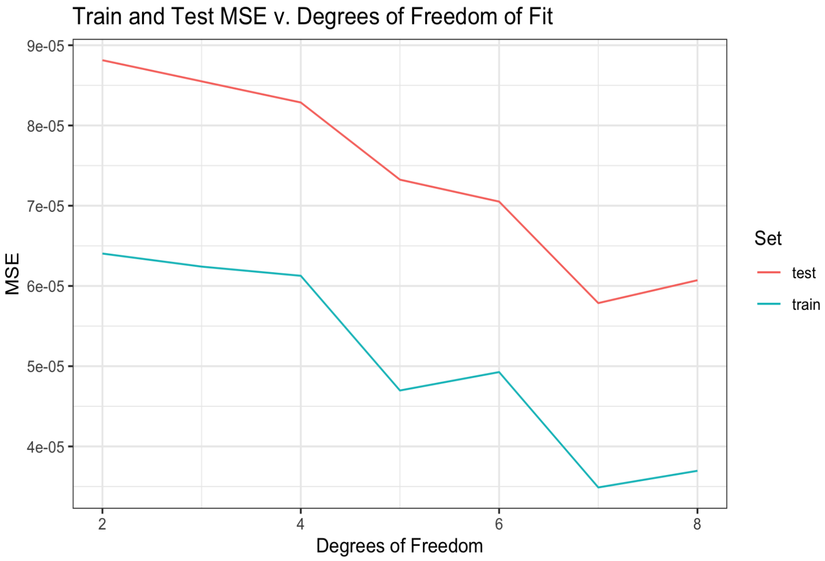

Figure 3. Train and test mean squared error for all model fits.

Viewing the trends in the test set error, the mean squared error (MSE) decreases with increasing degrees of freedom from 1-7. The test error only increases from 7 to 8 degrees of freedom, indicating that the quadratic spline fit with 7 degrees of freedom provides the minimal error. This fit is shown in Figure 4. Note that the train error increases from df 5 to df 6, which can be explained by the change in model from polynomial to spline at this point.

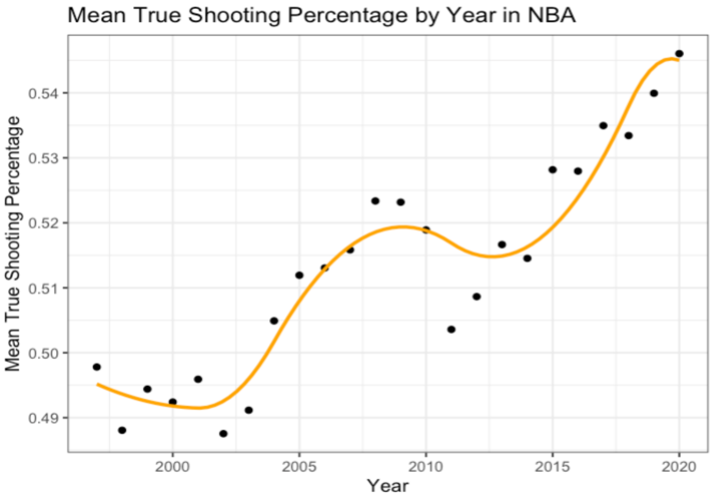

Figure 4. Average TSP by year and visualization of quadratic spline fit.

The second analysis within this scientific question evaluated the change in TSP corresponding to rule changes. The results of the permutation tests concerning the impact of our specified changepoints are shown in Table 1.

Year	Rule Change	Point Estimate (after-before)	p-value
2001	Zone defense	+0.35%	0.27
2004	Hand checking	+1.38%	0.0085
2011	Rip-through move	-1.53%	0.0011
2018	Freedom of movement	-0.15%	0.41

Table 1. Point estimates and p-values from permutation tests evaluating the effect of changepoints corresponding to rule changes on TSP.

From our table, we observe reasonable support that there were two significant change points. We observed a 1.38% increase of mean TSP between the seasons that saw the hand checking rule change (2004). Additionally, we observe a 1.53% decrease in mean TSP between the seasons that bordered the rip-through move rule change (2011).

For the results of our championship team depth analysis, we performed hypothesis testing using permutation, stratified by season. Our original difference of mean PPG of non-leading scorers on championship teams and mean PPG of non-leading scorers on non-championship teams is -0.46. The distribution of the permuted differences in PPG is shown in Figure 5 along with the point estimate.

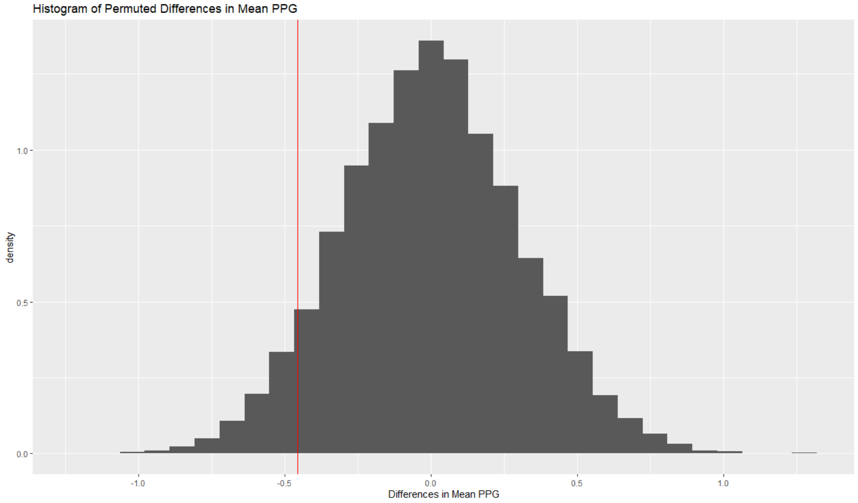

Figure 5. Permuted differences in mean PPG of non-leading scorers between championship and non-championship teams; red line corresponds to observed value in the data.

We did not have sufficient evidence to reject the null hypothesis that the average PPG of the non-leading scorers on championship teams equals the average PPG of non-leading scorers on non-championship teams. We failed to conclude whether championship teams have greater or lesser depth in terms of PPG than other teams.

Discussion

Through modeling the trend in mean TSP, we observed that shooting efficiency is increasing over the time period of interest. This confirms intuition based on the increase in three point shot attempts over the same time span. Despite the positive first-order linear trend over time, there remained instances of year-to-year decreases in mean TSP. For this reason, a more flexible model that utilizes quadratic spline functions was shown to have the lowest error when fit to the data.

The variability in year-to-year change in TSP also informed our decision to evaluate certain change points. The first point considered, comparing the 2000-2001 season to the 2001-2002 season to evaluate the effect of zone defenses and the defensive three seconds rule, did not show a significant change in TSP. A similar inconclusive result was reached when evaluating the effect of freedom of movement, which compared the TSP in 2017-2018 to 2018-2019. A significant positive change of 1.38% was found when comparing TSP from the 2003-2004 season to the 2004-2005 season. This corresponded to the implementation of the hand check rule, but the change in the metric from year to year may be attributable to other factors. This difficulty in attribution is also relevant when considering the other change point with a significant difference in TSP, which was comparing the 2010-2011 season to 2011-2012. This decrease of 1.53% may be due to the change in enforcement of rip-through fouls, which results in fewer free throws and a theoretically lower TSP. However, it may also be due to the player strike that caused a delay in the start of the NBA season that year. The attribution of metric changes to potential sources is outside the scope of this analysis.

In our second scientific question, we were unable to make strong conclusions regarding the depth of championship teams. There could be several limitations in this analysis, especially concerning our cutoffs on which seasons to include and which players to consider on each team. To reduce variability and have more relevant findings, we used a span of recent seasons. However, the decision to start with the 2015-2016 seasons was rather arbitrary. It could have been more helpful to compare seasons within our specified change points for consistency. We defined non-leading scorers as 4th-9th because we did not want to include players that were leading the team in points. The results suggest that championship and non-championship teams have comparable depth and may be differentiated by the abilities of their top scorers. A follow-up analysis should consider a one-sided hypothesis and redesign the comparisons. For example, instead of using only the championship teams, we could have considered playoff teams versus non-playoff teams.

In conclusion, we observed measurable changes in the style of gameplay. There are a multitude of factors that affect TSP, but our analysis showed that there is some evidence that the trend in mean TSP follows our specified change points. Though we did not make conclusions about the depth of scoring comparing championship teams to non-championship teams, this analysis could be used to inform further work in the area.

References

1. Dator, J. (2021, March 10). The NBA is at a breaking point with three-point shooting. SBNation.com. Retrieved November 13, 2021, from https://www.sbnation.com/nba/2021/3/10/22323023/nba-three-point-shooting-breaking-point.
2. Glossary. Basketball. (n.d.). Retrieved November 13, 2021, from https://www.basketball-reference.com/about/glossary.html.
3. *Strategically driven rule changes in NBA: Causes and consequences*. The Sport Journal. (2020, June 2). Retrieved December 3, 2021, from https://thesportjournal.org/article/strategically-driven-rule-changes-in-nba-causes-and-consequences/.
4. Johnson, D. (2018, November 21). NBA: How hand-check penalty changed basketball forever. Sportskeeda. Retrieved December 3, 2021, from https://www.sportskeeda.com/basketball/how-hand-checking-foul-changed-the-nba-forever
5. Bucher, R. (2011, December 8). *NBA set to implement rule changes*. ESPN. Retrieved December 3, 2021, from https://www.espn.com/nba/story/_/id/7329584/nba-alters-emphasis-shooting-fouls-2011-12.
6. Rapp, T. (2018, November 2). *Adam Silver: NBA rule changes having intended effect of increasing scoring*. Bleacher Report. Retrieved December 3, 2021, from https://syndication.bleacherreport.com/amp/2804046-adam-silver-nba-rule-changes-having-intended-effect-of-increasing-scoring.amp.html.

Code Appendix

See attached file

BIOST 544 Project Code

Anh-Minh Nguyen and Benjamin Stan

```r
## Load libraries
rm(list = ls())
library(dplyr)
library(ggplot2)
library(readr)
library(readxl)
library(splines)
library(splitTools)
library(purrr)
options(digits = 3)
setwd("/Users/bstan/Documents/UW/Courses/BIOST 544")
data = read_csv("data/all_seasons.csv",show_col_types = FALSE)
champs = read_csv("data/all_champions.csv",show_col_types = FALSE)
```

True Shooting Percentage

```r
## Define threshold for usage rate
data$year = as.numeric(substring(data$season,1,4))
data = data %>% filter(year >= 1997)
ggplot(data,aes(x=usg_pct,y=..density..)) +
    geom_histogram() +
  geom_vline(xintercept=0.1, colour = "red") +
    ggtitle('Histogram of Usage Percentage') +
    xlab('Usage Percentage') +
    ylab('Density')
```

```r
summary(data$usg_pct)
mean(data$usg_pct<0.1)
## Filter out players with low usage rates (< 10%)
data_high_usg = data %>% filter(usg_pct>0.1)
```

```r
## Calculate avg TSP
ts_time = data_high_usg %>%
  group_by(year) %>%
  summarise(mean_tsp = mean(ts_pct))
## Plot avg TSP by year
ggplot(aes(x=year,y=mean_tsp),data=ts_time) +
  geom_point() +
  xlab("Year") +
  ylab("Mean True Shooting Percentage") +
  ggtitle("Mean True Shooting Percentage by Year in NBA") +
  theme_bw()
```

```r
## Plot polynomial fits
ggplot(aes(x=year,y=mean_tsp),data=ts_time) +
  geom_point() +
  xlab("Year") +
  ylab("Mean True Shooting Percentage") +
  ggtitle("Mean True Shooting Percentage by Year in NBA") +
  geom_smooth(aes(col="1"), method='lm', formula=y~x, se=FALSE) +
  geom_smooth(aes(col="2"), method='lm', formula=y~poly(x,2), se=FALSE) +
  geom_smooth(aes(col="3"), method='lm', formula=y~poly(x,3), se=FALSE) +
  geom_smooth(aes(col="4"), method='lm', formula=y~poly(x,4), se=FALSE) +
  scale_colour_manual("Degree of Polynomial",
                      values = c("red", "purple", "green","yellow")) +
  theme_bw() +
  theme(legend.position="top")
```

```r
## Plot spline fits
ggplot(aes(x=year,y=mean_tsp),data=ts_time) +
  geom_point() +
  xlab("Year") +
  ylab("Mean True Shooting Percentage") +
  ggtitle("Mean True Shooting Percentage by Year in NBA") +
  geom_smooth(aes(col="1"),
            method='lm',
            formula=y~bs(x, degree = 1, knots = c(2001,2004,2011,2018), intercept = TRUE),
            se=FALSE) +
  geom_smooth(aes(col="2"),
            method='lm',
            formula=y~bs(x, degree = 2, knots = c(2001,2004,2011,2018), intercept = TRUE),
            se=FALSE) +
  geom_smooth(aes(col="3"),
          method='lm',
          formula=y~bs(x, degree = 3, knots = c(2001,2004,2011,2018), intercept = TRUE),
          se=FALSE) +
  scale_colour_manual("Degree of Spline", values = c("blue","orange","grey")) +
  theme_bw() +
  theme(legend.position="top")
```

```r
set.seed(40)
## Partition data into train and test sets
inds = partition(as.factor(data_high_usg$year),p=c(train=0.6,test=0.4))
train_df = data_high_usg[inds$train,] %>%
  group_by(year) %>%
  summarise(ts_pct = mean(ts_pct))
test_df = data_high_usg[inds$test,] %>%
  group_by(year) %>%
  summarise(ts_pct = mean(ts_pct))
train_error = NULL
test_error = NULL
## Fit all models and calculate train and test error
for (i in 1:7) {
  ## Fit model
  if (i <= 4) {
    lm_fit = lm_fit = lm(ts_pct~poly(year,i),data=train_df) }
  else {
        lm_fit = lm(ts_pct~bs(year,
                              degree = i-4,
                              knots = c(2001,2004,2011,2018),
                              intercept = TRUE),
                    data=train_df)
        ## bs uses truncated power basis with K+degree+1 df
  }
  ## Calculate train error
  ls_train_error = mean(lm_fit$residuals^2)
  train_error = c(train_error,ls_train_error)
  ## Calculate test error
  predicted_vals = predict(lm_fit,newdata=test_df)
  actuals_preds = data.frame(cbind(actual=test_df$ts_pct, predicted=predicted_vals))
  actuals_preds$error = actuals_preds$actual - actuals_preds$predicted
  ls_test_error = mean(actuals_preds$error^2) # MSE for test data
  test_error = c(test_error,ls_test_error)
}
## Plot error values
ggplot() +
  geom_line(aes(x=2:8, y = train_error, colour = "train")) +
  geom_line(aes(x=2:8, y = test_error, colour = "test")) +
  labs(x='Degrees of Freedom', y='MSE', colour="Set") +
  ggtitle("Train and Test MSE v. Degrees of Freedom of Fit") +
  theme_bw()
```

```
## Plot final model fit
ggplot(aes(x=year,y=mean_tsp),data=ts_time) +
  geom_point() +
  xlab("Year") +
  ylab("Mean True Shooting Percentage") +
  ggtitle("Mean True Shooting Percentage by Year in NBA") +
  geom_smooth(aes(col="2"),
              method='lm',
              formula=y~bs(x, degree = 2, knots = c(2001,2004,2011,2018), intercept = TRUE),
              se=FALSE) +
  scale_colour_manual("Degree of Spline", values = c("orange")) +
  theme_bw()
```

```
## Create function to permute year and calculate difference in TSP
set.seed(40)
perm_season = function(data, year_of_change) {
  df = data %>% filter(year %in% c(year_of_change,year_of_change-1))
  perm = sample(1:nrow(df), replace = FALSE)
  perm_df = df
  perm_df$year = df$year[perm]
  perm_mean_diff = with(perm_df,
                        (mean(ts_pct[year == year_of_change],na.rm=TRUE) -
                           mean(ts_pct[year == year_of_change-1],na.rm=TRUE))
                        )
  return(perm_mean_diff)
}
## Create function to perform set number of permutations for input year
```

```
season_comparison = function(data,year_of_change,nsim=10000) {
  point_est = mean((data %>% filter(year==year_of_change) %>% select(ts_pct))$ts_pct) -
    mean((data %>% filter(year==year_of_change-1) %>% select(ts_pct))$ts_pct)
  permuted_diff = replicate(nsim,perm_season(data,year_of_change))
  print(point_est)
  print(min(mean(point_est <= permuted_diff),mean(point_est >= permuted_diff)))
  ggplot(data.frame(permuted_diff),aes(x=permuted_diff,y=..density..)) +
    geom_histogram() +
    geom_vline(xintercept=point_est, colour = "red") +
    ggtitle(paste('Difference in TS % Between',
                  toString(year_of_change-1),
                  'and',
                  toString(year_of_change))) +
    xlab('Difference in TS %') +
    ylab('Density')
}
## Calculate permuted difference for 2001
season_comparison(data_high_usg,2001)
```

```
## Calculate permuted difference for 2004
season_comparison(data_high_usg,2004)
```

```
## Calculate permuted difference for 2011
season_comparison(data_high_usg,2011)
```

```
## Calculate permuted difference for 2018
season_comparison(data_high_usg,2018)
```

Evaluating Depth of Championship Teams

```
## Identify the championship players by merging the datasets
bc = left_join(data, champs, by = 'season')

## Wrangle data for only necessary columns, and past season 2010
bc = bc %>%
  dplyr::select(player_name, team_abbreviation, pts, season, champions, runner_up) %>%
  dplyr::filter(season %in% c('2015-16', '2016-17',
                              '2017-18', '2018-19', '2019-20', '2020-21'))

## Create a new column identifying whether a player was on a championship team
bc = bc %>% mutate(is.champ = ifelse(team_abbreviation == champions | team_abbreviation == runner_up, 1

## Find the player rankings in terms of ppg on each team
bc = bc %>%
  group_by(season, team_abbreviation) %>%
  arrange(-pts) %>%
  mutate(team_ranking = row_number())

## Make sure every team has at least nine players
```

```r
## Filter for only the 4th-9th ranked ppg players
bc2 = bc %>% filter(team_ranking >= 4, team_ranking <= 9) %>%
  select(-champions, -runner_up)

## Our original statistic
orig.diff = mean(bc2$pts[bc2$is.champ == 1]) - mean(bc2$pts[bc2$is.champ == 0])
```

```r
## Visualization
bc2$Status = factor(ifelse(bc2$is.champ == 1, 'Championship', "Non-championship"))
bc2$Status = relevel(bc2$Status, ref = 'Non-championship')

ggplot(bc2) +
  geom_histogram(aes(x = pts, fill = Status, color = Status),
                 position = 'stack', binwidth = 0.5,) +
  xlab('PPG') +
  ggtitle('Histograms of PPG by championship status')
```

```r
## Create a permutation function to shuffle players within the same season
perm.season = function(data, year) {
  df = data[which(data$season == year),]
  shuffle.champ = sample(df$is.champ, replace = FALSE)
  return(data.frame(
    season = year,
    is.champ = shuffle.champ,
    pts = df$pts
  ))
}
```

```r
## Create a function to calculate permuted stat
perm.all.seasons = function(data) {
  shuffled.data = c('2015-16', '2016-17', '2017-18', '2018-19', '2019-20', '2020-21') %>%
    map_dfr(~ perm.season(data = bc2, year = .x))
  mean(shuffled.data$pts[shuffled.data$is.champ == 1]) -
    mean(shuffled.data$pts[shuffled.data$is.champ == 0])
}

## Perfom 10000 permutations, to get 10000 simulated values
set.seed(1)
nperms = 10000
diff.perms = replicate(nperms, perm.all.seasons(bc2))

## Compare to original statistic and test
ggplot(data.frame(x = diff.perms), aes(x = x, y=..density..)) +
  geom_histogram() +
  geom_vline(xintercept=orig.diff, color="red") +
  xlab('Differences in Mean PPG') +
  ggtitle('Histogram of Permuted Differences in Mean PPG')
```

```r
2 * (sum(diff.perms <= orig.diff)/nperms)
```