The Evolution of Shot Taking in the NBA

2023-06-17

Abbreviations

NBA : National Basketball Association

WNBA : Woman's National Basketball Association

NCAA : National Collegiate Athletics Association

EV : Expected Value

1. Motivation

The tree point shot is one the most interesting phenomena in sporting history. When playing basketball, if a shot is made from within the three point line, it counts as 2 points, however, baskets made from beyond the line are worth three. Its earliest introduction to a professional basketball league was in 1961, when the American Basketball League deemed that shots made from a distance of 25 feet and greater were to be counted as three points [1]. The distance of the three point line from the net has varied over time, and still varies across different leagues. The official NBA three point line is 23 feet and 9 inches from the center of the basket [2], where as in the NCAA and WNBA, the line is only 22 feet and 1.75 inches from the basket's center [3, 4].

However, its not only the three point line itself that has evolved, but the way it influences the game. What was once considered a desperation play, the three pointer has now become a quintessential part of a team's preparation. There are now teams whose coaches build teams around elite three point shooters, and are willing to pay these players tens of millions of dollars each year [5]. In 1979-80, an average of 2.8 shots per game were taken from beyond the line, whereas in the 2022-23 NBA season, an average of 32.4 attempts were taken every game [6].

With three point shooting becoming such a prevalent part of the NBA, fans and commentators have taken notice. Some have even gone as far as to say three point shooting has ruined the game, but do these claims have any weight? With this analysis, I aim to explore the evolution of three point shooting over the years, and examine how effective three point shooting really is.

2. Background

For my analysis, I'll be using a combination of NBA shot log data pulled from the NBA Stat API (now discontinued), which contains spatial data for all shots taken in the NBA from 1997 to 2020, the NBA Database data set (found on Kaggle), which contains record off all NBA games played since the 1946-47 NBA season, and the NBA Stats data set which was scraped from Stathead, and contains seasonal information regarding the stats of all NBA players since 1947.

3. Methods and Analysis

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
## Warning: package 'ggpubr' was built under R version 4.3.1
```

```
## Warning: package 'viridis' was built under R version 4.3.2
```

```
## Loading required package: viridisLite
```

3.1 Data Import and Cleaning

The first data set we're going to import is the one keeping track of game data. Very little needs to be done in terms of preparing the data set, we specify that the id columns ought to be parsed as integers (as opposed to doubles), and in place of the "W"s and "L"s used to denoted wins and losses respectively in the win loss column, we'll parse them as logical TRUE and FALSE values instead.

```
game_data
```

```
## # A tibble: 62,367 x 54
##    season_id team_id_home team_abbreviation_home team_name_home          game_id
##        <int>        <int> <chr>                  <chr>                    <int>
##  1     21946   1610610035 HUS                    Toronto Huskies         2.46e7
##  2     21946   1610610034 BOM                    St. Louis Bombers       2.46e7
##  3     21946   1610610032 PRO                    Providence Steamrollers 2.46e7
##  4     21946   1610610025 CHS                    Chicago Stags           2.46e7
##  5     21946   1610610028 DEF                    Detroit Falcons         2.46e7
##  6     21946   1610610026 CLR                    Cleveland Rebels        2.46e7
##  7     21946   1610610031 PIT                    Pittsburgh Ironmen      2.46e7
##  8     21946   1610612738 BOS                    Boston Celtics          2.46e7
##  9     21946   1610610028 DEF                    Detroit Falcons         2.46e7
## 10     21946   1610610032 PRO                    Providence Steamrollers 2.46e7
## # i 62,357 more rows
## # i 49 more variables: game_date <dttm>, matchup_home <chr>, min <dbl>,
## #   fgm_home <dbl>, fga_home <dbl>, fg_pct_home <dbl>, fg3m_home <dbl>,
## #   fg3a_home <dbl>, fg3_pct_home <dbl>, ftm_home <dbl>, fta_home <dbl>,
## #   ft_pct_home <dbl>, oreb_home <dbl>, dreb_home <dbl>, reb_home <dbl>,
## #   ast_home <dbl>, stl_home <dbl>, blk_home <dbl>, tov_home <dbl>,
## #   pf_home <dbl>, pts_home <dbl>, plus_minus_home <dbl>, ...
```

Next, we'll perform some manipulations to transform our game data set into a summary of seasonal team statistics. Firstly, note that for each game, there are two team's statistics we have to record, the home teams'

stats, and the away team's stats. We'll start by creating two data frames, one that groups all games by home team, by year and another that groups all games by away team, by year. We can sum each group's numeric results to get aggregated statistics on how many shots, fouls, assists, etc, each team took by year. Using these aggregate statistics, we can compute summary statistics such as averages, counts and percentages.

```
## `summarise()` has grouped output by 'team_id'. You can override using the
## `.groups` argument.
```

```
team_data
```

```
## # A tibble: 1,032 x 48
## # Groups:   team_id [30]
##       team_id season_id season team_name     team_abbreviation games_played
##         <int>     <int>  <dbl> <chr>         <chr>                     <int>
##  1 1610612737     21985   1985 Atlanta Hawks ATL                          52
##  2 1610612737     21986   1986 Atlanta Hawks ATL                          82
##  3 1610612737     21987   1987 Atlanta Hawks ATL                          82
##  4 1610612737     21989   1989 Atlanta Hawks ATL                          82
##  5 1610612737     21990   1990 Atlanta Hawks ATL                          82
##  6 1610612737     21991   1991 Atlanta Hawks ATL                          82
##  7 1610612737     21992   1992 Atlanta Hawks ATL                          82
##  8 1610612737     21993   1993 Atlanta Hawks ATL                          82
##  9 1610612737     21994   1994 Atlanta Hawks ATL                          82
## 10 1610612737     21995   1995 Atlanta Hawks ATL                          82
## # i 1,022 more rows
## # i 42 more variables: feild_goals_made <dbl>, feild_goals_attempted <dbl>,
## #   three_pointers_made <dbl>, three_pointers_attempted <dbl>,
## #   two_pointers_attempted <dbl>, two_pointers_made <dbl>,
## #   free_throws_made <dbl>, free_throws_attempted <dbl>,
## #   offensive_rebounds <dbl>, deffensive_rebounds <dbl>, rebounds <dbl>,
## #   assits <dbl>, steals <dbl>, blocks <dbl>, turnovers <dbl>, fouls <dbl>, ...
```

By grouping the team data by year, we can also generate a data frame for league wide seasonal data.

```
league_data
```

```
## # A tibble: 36 x 46
##    season  team_id season_id games_played feild_goals_made feild_goals_attempted
##     <dbl>    <dbl>     <int>        <int>            <dbl>                 <dbl>
##  1   1985  3.70e10    505655         1158            50198                102410
##  2   1986  3.70e10    505678         1886            80422                167461
##  3   1987  3.70e10    505701         1886            79473                165441
##  4   1989  4.35e10    593703         2214            91914                192951
##  5   1990  4.35e10    593730         2214            91551                193059
##  6   1991  4.35e10    593757         2214            91371                193391
##  7   1992  4.35e10    593784         2214            90056                190294
##  8   1993  4.35e10    593811         2214            87064                186951
##  9   1994  4.35e10    593838         2214            84105                180423
## 10   1995  4.67e10    637855         2378            88096                190675
## # i 26 more rows
## # i 40 more variables: three_pointers_made <dbl>,
## #   three_pointers_attempted <dbl>, two_pointers_attempted <dbl>,
```

```
## #   two_pointers_made <dbl>, free_throws_made <dbl>,
## #   free_throws_attempted <dbl>, offensive_rebounds <dbl>,
## #   deffensive_rebounds <dbl>, rebounds <dbl>, assits <dbl>, steals <dbl>,
## #   blocks <dbl>, turnovers <dbl>, fouls <dbl>, points <dbl>, ...
```

Finally, we'll import the player data. The only cleaning that needs to be done is the renaming of columns into a human readable format.

```
## Rows: 31135 Columns: 35
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): player, pos, lg, tm
## dbl (31): seas_id, season, player_id, birth_year, age, experience, g, gs, mp...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
player_data
```

```
## # A tibble: 31,135 x 35
##    seas_id season player_id player   birth_year position   age experience league
##      <dbl>  <dbl>     <dbl> <chr>         <dbl> <chr>    <dbl>      <dbl> <chr>
##  1   30458   2023      5025 A.J. Gr~         NA SG          23          1 NBA
##  2   30459   2023      5026 A.J. La~         NA SG          22          1 NBA
##  3   30460   2023      5026 A.J. La~         NA SG          22          1 NBA
##  4   30461   2023      5026 A.J. La~         NA SG          22          1 NBA
##  5   30462   2023      4219 Aaron G~         NA PF          27          9 NBA
##  6   30463   2023      4582 Aaron H~         NA PG          26          5 NBA
##  7   30464   2023      4805 Aaron N~         NA SF          23          3 NBA
##  8   30465   2023      4900 Aaron W~         NA SG          24          2 NBA
##  9   30466   2023      4688 Admiral~         NA PF          25          3 NBA
## 10   30467   2023      5027 AJ Grif~         NA SF          19          1 NBA
## # i 31,125 more rows
## # i 26 more variables: team_name <chr>, games <dbl>, games_started <dbl>,
## #   minutes_played <dbl>, feild_goals <dbl>, feild_goals_attempted <dbl>,
## #   feild_goal_percentage <dbl>, three_pointers_made <dbl>,
## #   three_pointers_attempted <dbl>, three_point_percentage <dbl>,
## #   two_pointers_made <dbl>, two_pointers_attempted <dbl>,
## #   two_point_percentage <dbl>, effective_feild_goal_percentage <dbl>, ...
```

We're also going to import some spatial data. This data set includes every single shot taken in the NBA from 1997 to 2020. So important features include the X and Y coordinates, the distance from which the shot was taken, the player who took the shot and the date on which the shot was taken.

```
## Rows: 4729512 Columns: 22
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): Game ID, Player Name, Team Name, Action Type, Shot Type, Shot Zone...
## dbl (11): Game Event ID, Player ID, Team ID, Period, Minutes Remaining, Seco...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
spatial_data
```

```
## # A tibble: 4,729,512 x 23
##    game_id    game_event_id player_id player_name    team_id team_name    period
##    <chr>              <dbl>     <dbl> <chr>            <dbl> <chr>         <dbl>
##  1 0029700427           389       100 Tim Legler  1610612764 Washington ~      4
##  2 0029700427           406       100 Tim Legler  1610612764 Washington ~      4
##  3 0029700427           475       100 Tim Legler  1610612764 Washington ~      4
##  4 0029700427           487       100 Tim Legler  1610612764 Washington ~      4
##  5 0029700427           497       100 Tim Legler  1610612764 Washington ~      4
##  6 0029700449            79       100 Tim Legler  1610612764 Washington ~      1
##  7 0029700449           152       100 Tim Legler  1610612764 Washington ~      2
##  8 0029700449           336       100 Tim Legler  1610612764 Washington ~      3
##  9 0029700453           141       100 Tim Legler  1610612764 Washington ~      2
## 10 0029700482           116       100 Tim Legler  1610612764 Washington ~      2
## # i 4,729,502 more rows
## # i 16 more variables: minutes_remaining <dbl>, seconds_remaining <dbl>,
## #   action_type <chr>, shot_type <chr>, shot_zone_basic <chr>,
## #   shot_zone_area <chr>, shot_zone_range <chr>, shot_distance <dbl>,
## #   x_location <dbl>, y_location <dbl>, shot_made_flag <dbl>, game_date <date>,
## #   home_team <chr>, away_team <chr>, season_type <chr>, season <dbl>
```

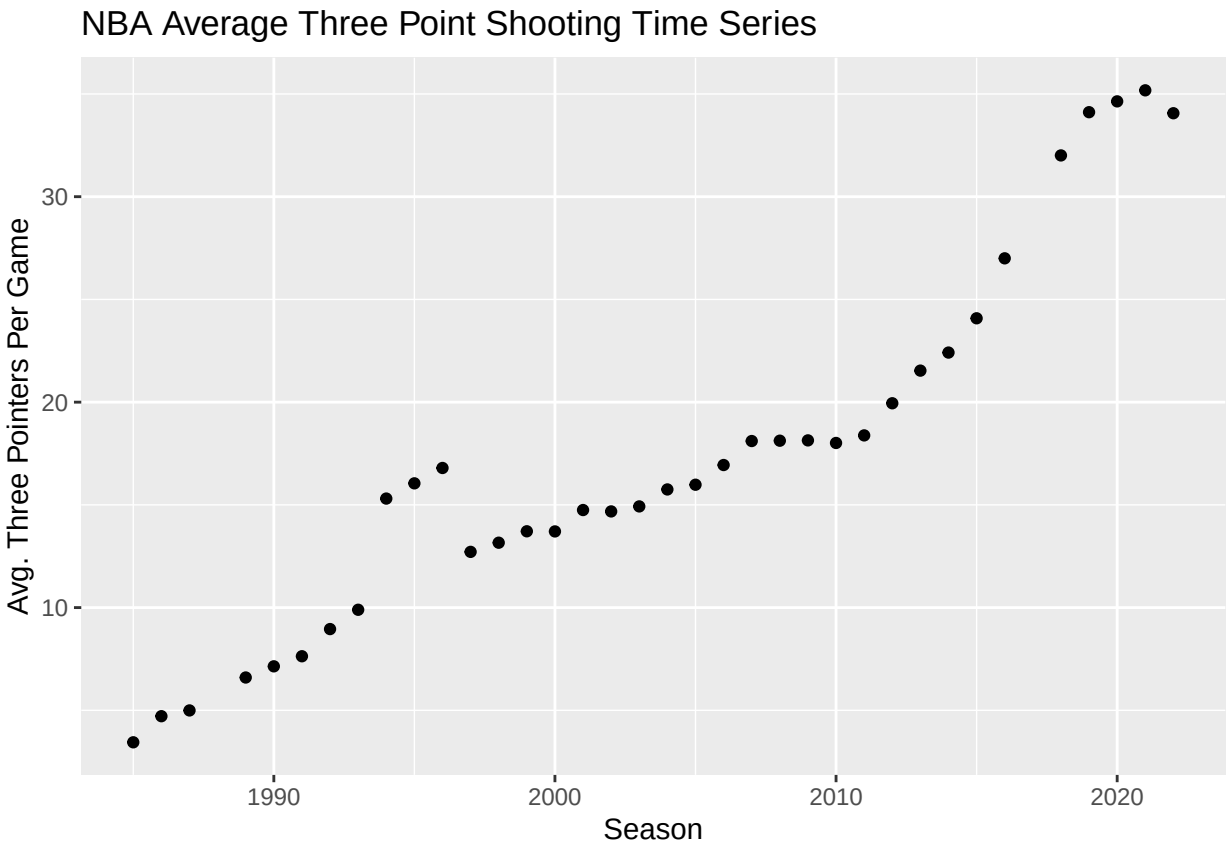
3.2 Exploring the Prevelance of Three Point Shooting Over Time

The first step in our analysis will consist exploring how common the three point shot really is. We explore time series data, as well as the distribution of three point shooting across the league and different players. While we can't draw any causal conclusions from our analysis, we hope to provide insight into how the three point shot has evolved over time, possible explanations as to why it has become so prevalent in modern basketball, and points of interest in the history of the three point shot.

3.2.1 Increase in Three Point Shooting Over the Years

First, let's examine the average number of three pointers taken per game.

```
league_data %>%
  ggplot(aes(x = season, y = three_pointers_attempted / games_played)) +
  geom_point() +
  ylab("Avg. Three Pointers Per Game") +
  xlab("Season") +
  ggtitle("NBA Average Three Point Shooting Time Series")
```



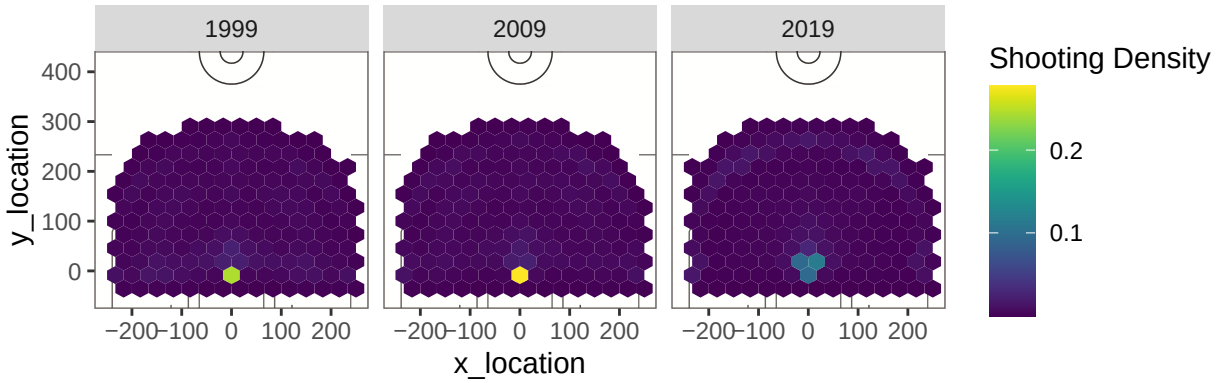
Clearly, there's been a steady increase in the amount of three point shooting per game across the NBA. We can observe a relatively steady increase in the average number of three point shots taken per game, up until about 2010, when the average number starts to sky rocket, and eventually plateau around 2019.

Even when viewing this data from a spatial perspective, can see that a league that was once dominated by layups has slowly started to shift. Likely due to improved defensive strategies and an over increase in the skill of players, teams have started taking shots from the three point line more often.

```
img <- readJPEG("./data/nba_court.jpg", native = TRUE)

games_played_by_season <- as.data.frame.table(table(spatial_data$season))
games_played_by_season$Var1 <- as.numeric(games_played_by_season$Var1) + 1996
games_played_by_season$Freq <- as.double(games_played_by_season$Freq)
games_played_by_season <- games_played_by_season %>%
  rename("games_played_in_season" = "Freq")

augmented_spatial_data <- spatial_data %>%
  left_join(games_played_by_season, c("season" = "Var1")) %>%
  filter(season %in% c(2019, 2009, 1999)) %>%
  filter(shot_distance < 30)

augmented_spatial_data %>%
  ggplot(aes(x = x_location, y = y_location, fill=after_stat(density))) +
  background_image(img) +
  geom_hex(bins = 15) +
  scale_fill_viridis() +
  coord_fixed() +
```

```
  xlim(-250,250) +
  ylim(-52, 418) +
  facet_wrap(~ season) +
  labs(fill = "Shooting Density")
```

```
## Warning: Removed 26 rows containing missing values (`geom_hex()`).
```

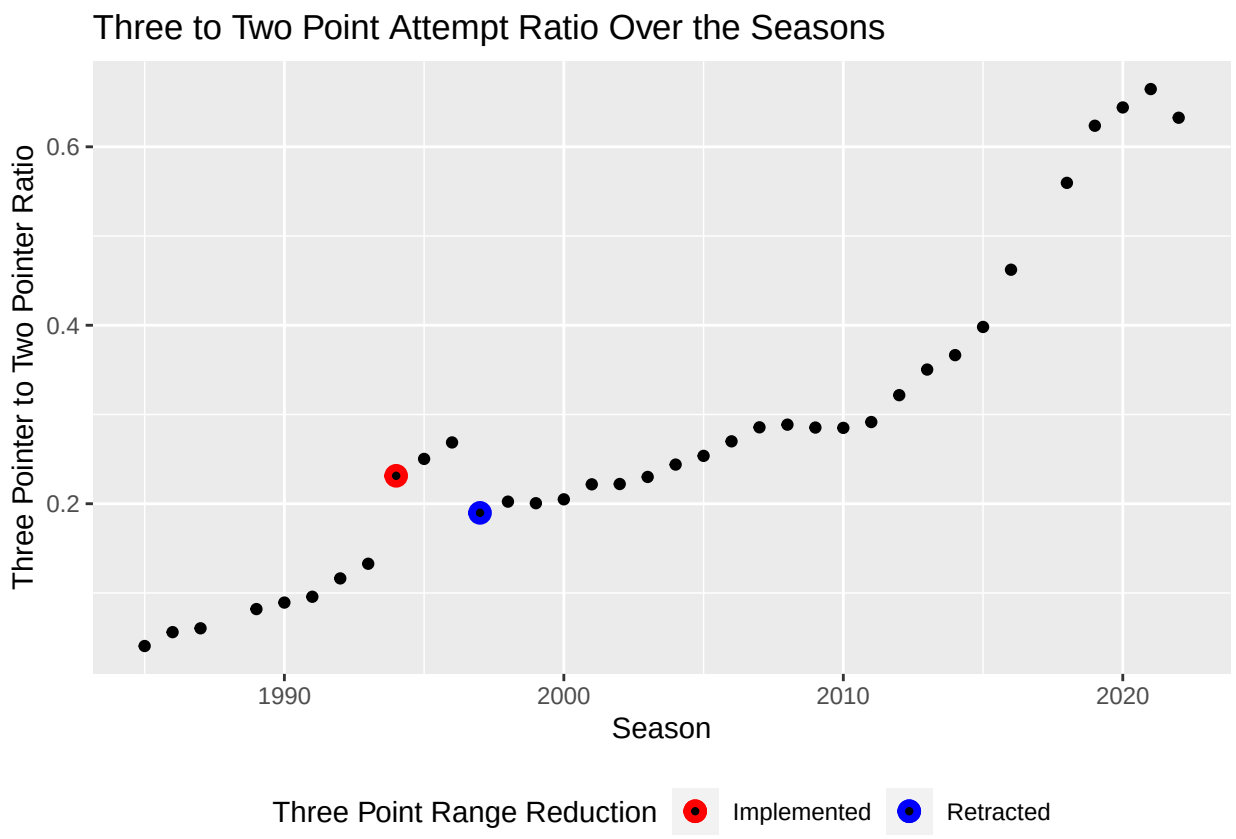


3.2.2 Reducing The Three Point Range

The plot below is a time series graph that depicts how the ratio of three pointers to two pointers has evolved over the seasons.

```
league_data %>%
  ggplot(aes(x = season, y = three_pointers_attempted / two_pointers_attempted)) +
  geom_point() +
  geom_point(aes(color = "Implemented"),
             fill = "black",
             shape = 21,
             stroke = 2,
             data = filter(league_data, season == 1994)) +
  geom_point(aes(color = "Retracted"),
             fill = "black",
             shape = 21,
             stroke = 2,
```

```
            data = filter(league_data, season == 1997)) +
  scale_color_manual(name = "Three Point Range Reduction",
                     breaks = c("Implemented", "Retracted"),
                     values = c("Implemented" = "red", "Retracted" = "blue")) +
  xlab("Season") +
  ylab("Three Pointer to Two Pointer Ratio") +
  ggtitle("Three to Two Point Attempt Ratio Over the Seasons") +
  theme(legend.position = "bottom")
```



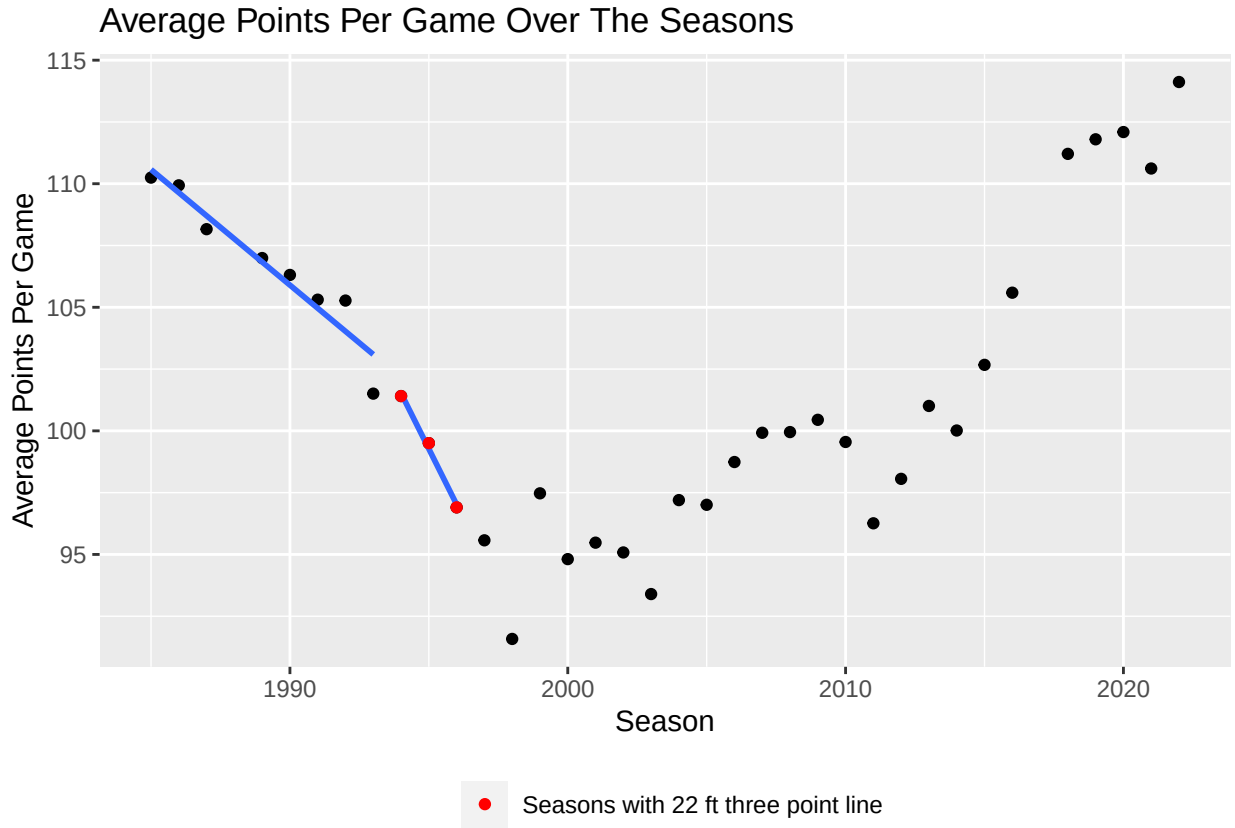
There's clearly been a steady increase in three point shooting over the history of the NBA, however, there are some points of interest in the graph above.

Firstly, note the large jump in the three to two pointer ratio between the 1993 and 1994 season, it went from 0.13 to 0.23. After some investigation, it turns out that in 1994, the NBA implemented an interesting rule change. If we create plot of the league's average points per game, we can see that preceding 1994, the average number of points per game had been declining steadily year over year.

```
league_data %>%
  ggplot(aes(x = season, y = points / games_played)) +
  geom_point() +
  geom_smooth(data = filter(league_data, season < 1994),
              method = "lm",
              se = FALSE) +
  geom_smooth(data = filter(league_data, 1994 <= season & season <= 1996),
              method = "lm", se = FALSE) +
  geom_point(data = filter(league_data, 1994 <= season & season <= 1996),
```

```
            mapping = aes(color = "Seasons with 22 ft three point line")) +
  ylab("Average Points Per Game") +
  xlab("Season") +
  scale_color_manual(name = "",
                     breaks = c("Seasons with 22 ft three point line"),
                     values = c("red")) +
  theme(legend.position = "bottom") +
  ggtitle("Average Points Per Game Over The Seasons")
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



During the 1994 season, the NBA implemented a rule change that would reduce the distance of the three point point line from 23 feet and 9 inches, down to an even 22 feet. This was their attempt to increase the amount of scoring that happened per game. While the policy did increase the rate at which teams shot three pointers, as shown in the figure above, the average number of points per game continued to decline. In fact, if we fit a linear regression to the points prior to the policy change, and another one while the policy change was in place, we see that the average points per game were declining even faster than they had been without the rule change.

```
lm(points / games_played ~ season,
   filter(league_data, season < 1994))
```

```
##
```

```
## Call:
## lm(formula = points/games_played ~ season, data = filter(league_data,
##     season < 1994))
##
## Coefficients:
## (Intercept)       season
##   1967.0862      -0.9353
```

```
lm(points / games_played ~ season,
   filter(league_data, 1994 <= season & season <= 1996))
```

```
##
## Call:
## lm(formula = points/games_played ~ season, data = filter(league_data,
##     1994 <= season & season <= 1996))
##
## Coefficients:
## (Intercept)       season
##    4591.442       -2.252
```

$$Y_{\text{Pre Policy}} = 1967.0862 - 0.9353X$$
$$Y_{\text{During Policy}} = 4591.442 - 2.252X$$

While the intercept is rather meaningless, these regressions tell us that prior to the policy change, the average number of points per game was falling at a rate of about 0.94 points per season, where as after the during the period in which the policy was active, it was falling at a rate of 2.25 points per season.

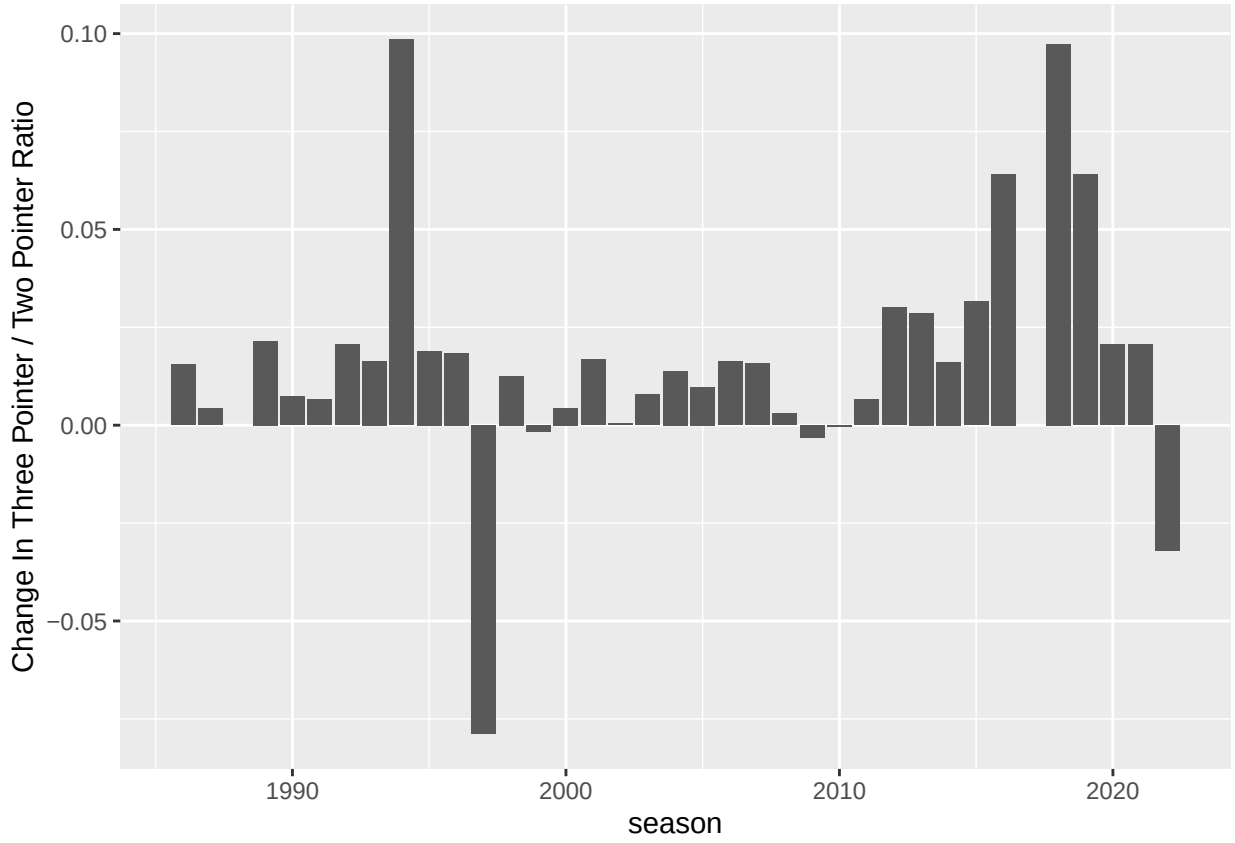
After three years, the league reverted their decision, which coincides with a sudden drop in three point shooting in 1997.

3.2.3 Rise in the Growth Rate of Three Point Shooting

Turning our attention back to the original time series, there is another noteworthy feature. After 2012, the rate at which three point shoots have been growing year over year explodes. This is made clear when we plot a time series showing the change in the three to two pointer ratio.

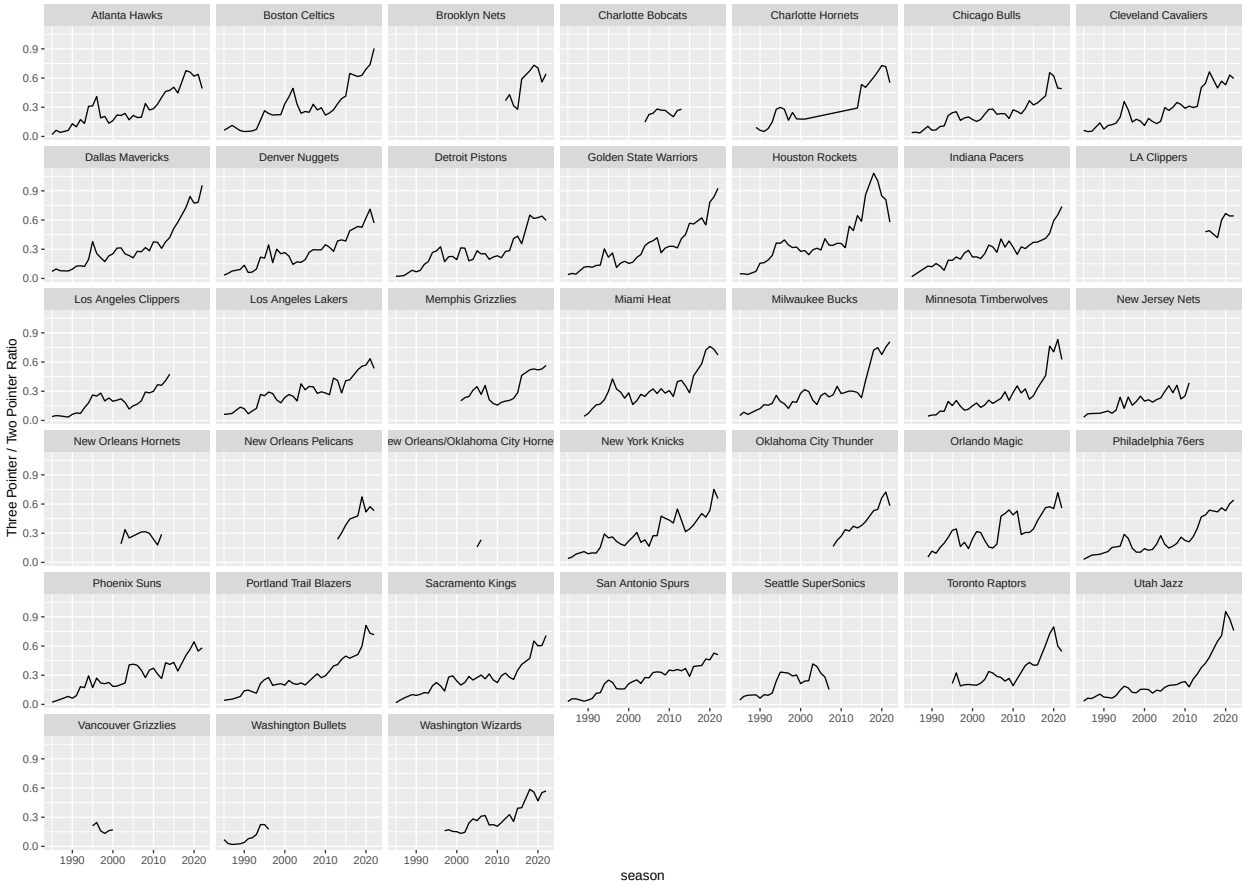
```
league_data %>%
  mutate(delta_tpr = three_pointers_attempted / two_pointers_attempted -
           lag(three_pointers_attempted) / lag(two_pointers_attempted)) %>%
  ggplot(aes(season, delta_tpr)) +
  geom_bar(stat = "identity") +
  ylab("Change In Three Pointer / Two Pointer Ratio")
```

```
## Warning: Removed 1 rows containing missing values (`position_stack()`).
```




```
team_data %>%
  group_by(team_name, season) %>%
  summarise(tpr = three_pointers_attempted / two_pointers_attempted) %>%
  ggplot(aes(season, tpr)) +
  geom_line() +
  facet_wrap(~ team_name) +
  ylab("Three Pointer / Two Pointer Ratio")
```

```
## `summarise()` has grouped output by 'team_name'. You can override using the
## `.groups` argument.
```



As shown in the graph above, every team in the league seems to have simultaneously increased their growth rate of three point shooting in 2010. As of now, I've been unable to find a particular event that catalyzed this increase in the growth rate of three point shooting. It's possible that the league as a whole started to analyze the data more closely, and began to build teams that focused more heavily on three point shooting. It could also have been the case that new talent were more prone to specializing in three point shooting, and the changes to the league were a result of players rather than management. Deciphering a particular cause extends for beyond the scope of this analysis, so I suggest that anyone interested do their own digging.

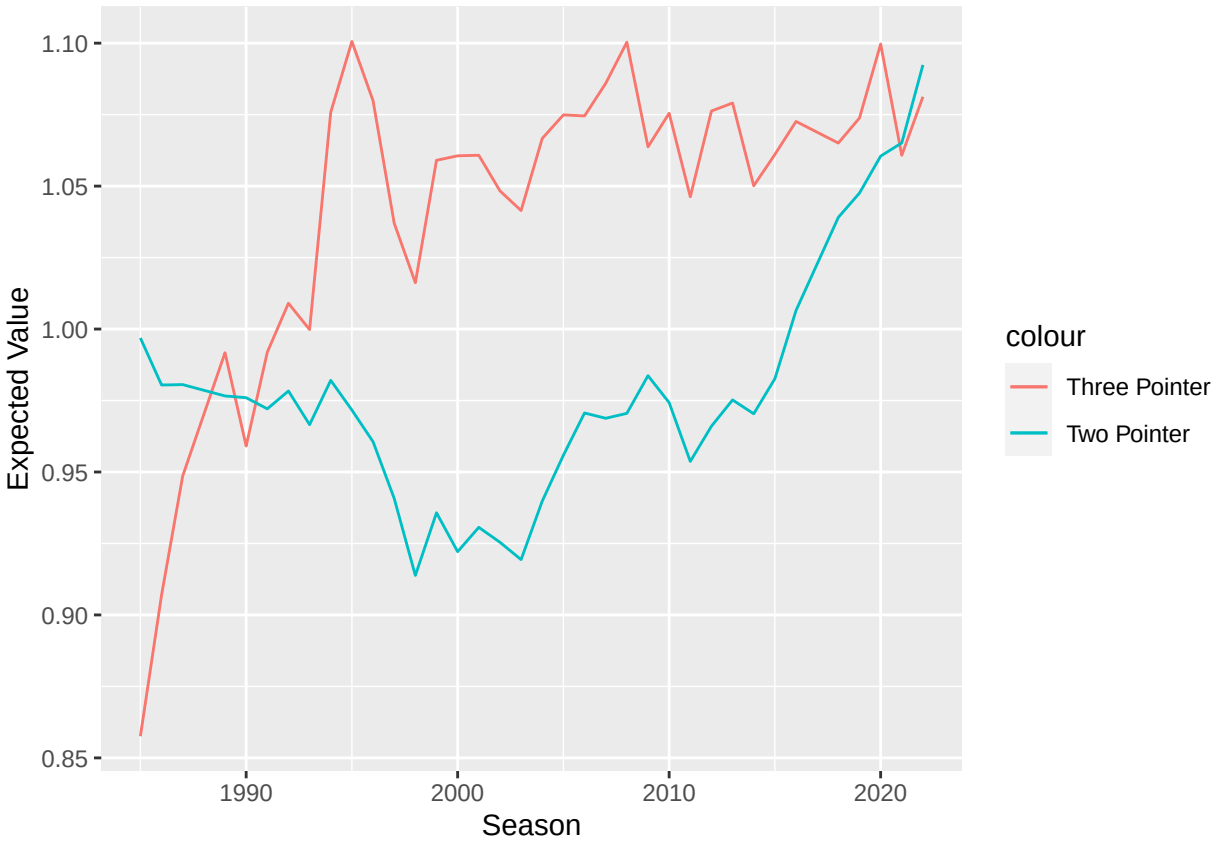
3.3 The Efficacy Of Three Point Shooting

In the previous section, we clearly identified that there's been a league wide trend in which teams have increased their proportion of three point shooting year over year. However, whether or not this has "destroyed the game" depends on how good teams are at shooting a three pointer. In order to determine how the efficacy of three point shooting has evolved over the years, a good first step is to contrast its expected value against that of a two point shot.

3.3.1 Expected Value

```
league_data %>%
  ggplot(aes(season)) +
  geom_line(aes(y = three_point_percentage * 3, color = "Three Pointer")) +
  geom_line(aes(y = two_point_percentage * 2, colour = "Two Pointer")) +
```

```
  xlab("Season") +
  ylab("Expected Value")
```



Notice that the expected value (hereinafter EV) of a three point shot had steadily increased until 1995, at which point it plateaued. In fact, the EV of a three point shot has exceeded the value of a two pointer since 1991, and the efficiency of two point shooting has lagged behind. It even experienced a slum between 1994 and 2006, where its EV dipped by about 0.07 points at its trough. However, starting in 2011, it experienced a resurgence, and it has now exceeded the EV of a three point shot.

3.3.2 Changing Times

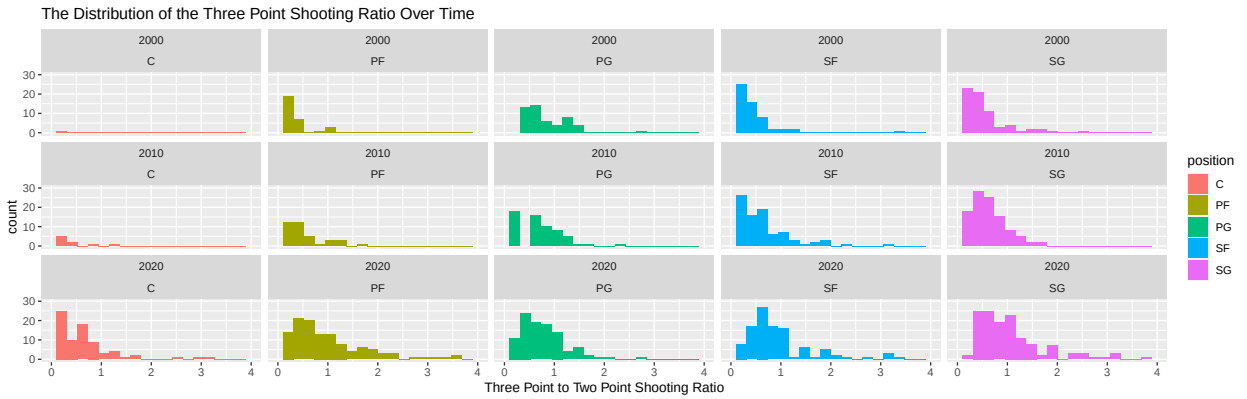
Despite the expected value of a three point shot having exceeded that of a two point from 1991 to 2022, the two point shot was, and still is, the more popular option. Moreover, the expected value of a three pointer has plateaued since 1995, yet it's popularity relative to the two point shot continued to increase. One plausible explanation is that due to the difficult nature of the three point shot, most players didn't bother shooting from a distance, and only as more players started to specialize in three point shooting did its popularity rise.

```
player_data %>%
  filter(season %in% c(2000, 2010, 2020),
         position %in% c("PG", "SF", "SG", "PF", "C")) %>%
  ggplot(aes(x = three_pointers_attempted / two_pointers_attempted,
             fill = position)) +
  geom_histogram(bins = 20) +
  xlim(c(0,4)) +
  ylim(c(0,30)) +
```

```
  facet_wrap(season ~ position, ncol = 5) +
  xlab("Three Point to Two Point Shooting Ratio") +
  ggtitle("The Distribution of the Three Point Shooting Ratio Over Time")
```

```
## Warning: Removed 27 rows containing non-finite values (`stat_bin()`).
```

```
## Warning: Removed 32 rows containing missing values (`geom_bar()`).
```



Notice that as the years have gone on, shooting guards and small forwards have steadily increased the rate at which the shoot three pointer relative to two pointers. While their roles were once mainly in place to feed more dominant players and maneuver the ball around the paint, as three point shooting has become more relevant, they taken on a new roll. Notice that while less noticeable, even power forwards and center players have begun to put more emphasis on their three point shooting.

I will be interesting to see how the future data plays out. Now that the EV of a two point shot has reached parity with the three pointer, teams will start to reevaluate their strategies, and I speculate that this could result in a plateau in the increase in three point shooting.

4. Conclusion

It's clear that the NBA has been evolving over the years. As we can see, three point shooting has completely changed the game's landscape as it has affect how teams score, how money is allocated, and how roles are defined. We've explored how rule changes can lead to drastic changes in how shots are taken, and how a shot's popularity isn't determined solely by its expected value. There are some questions we've left unanswered, such as the unprecedented rise in the growth rate of three point shooting seen in 2010; however, these extend far beyond the reach of this analysis.

It will be interesting to see how this data evolves over the next 10 to 20 years. In particular, will we see three point shooting to continue to explode, could we see a plateau, or possibly even a reversal of the trend? Ultimately, the future of this data will depend on a variety of factors such how coaches and managers adapt their strategies, the distribution on new talent among various positions, and changes made by the NBA. While many have accused the three point shot of ruining the game, its progression through the years has made for an interesting case study into how a single rule can leave an outsized impact on the rest of the game.