NBA Season Statistics - Exploratory Data Analysis

Albert Li

Loading necessary libraries

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## Warning: package 'tibble' was built under R version 4.1.3
```

```
## Warning: package 'tidyr' was built under R version 4.1.3
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
## Warning: package 'purrr' was built under R version 4.1.3
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
## Warning: package 'stringr' was built under R version 4.1.3
```

```
## Warning: package 'forcats' was built under R version 4.1.3
```

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

Loading in tables downloaded from the internet

```
#read in basketball statistics for players
Seasons_Stats <- read_csv("Data/Processed/Seasons_Stats.csv")
```

```
## New names:
## * `` -> ...1
```

```
## Rows: 24691 Columns: 53
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): Player, Pos, Tm
## dbl (48): ...1, Year, Age, G, GS, MP, PER, TS%, 3PAr, FTr, ORB%, DRB%, TRB%,...
## lgl  (2): blanl, blank2
##
## 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.
```

```
#read in player info
Players <- read_csv("Data/Processed/Players.csv")
```

```
## New names:
## * `` -> ...1
## Rows: 3922 Columns: 8-- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): Player, collage, birth_city, birth_state
## dbl (4): ...1, height, weight, born
## 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.
```

The above data frames were downloaded at https://www.kaggle.com/drgilermo/nba-players-stats/data .

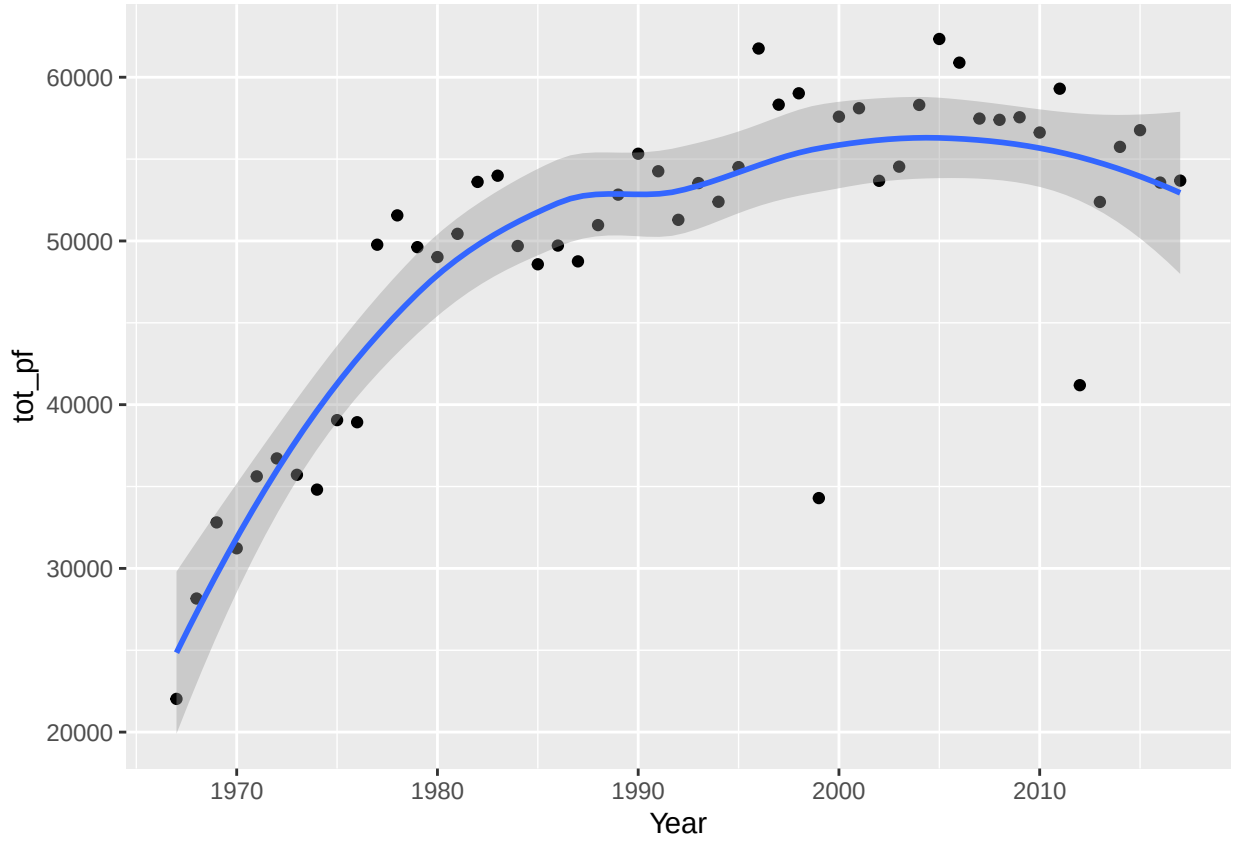
Seasons_Stats includes all players who have played since 1950 and their statistics from each season played. Players gives us more background information about all the different players who have played in the NBA.

Free throws attempted and personal fouls committed

The National Basketball Association (NBA) has seen its changes throughout the years. Many spectators believe that the current era is much less physical compared to before. To measure this, we will be taking a look at free throws attempted and personal fouls committed. First, we took a look at how many personal fouls were called and how many free throws were attempted each NBA season.

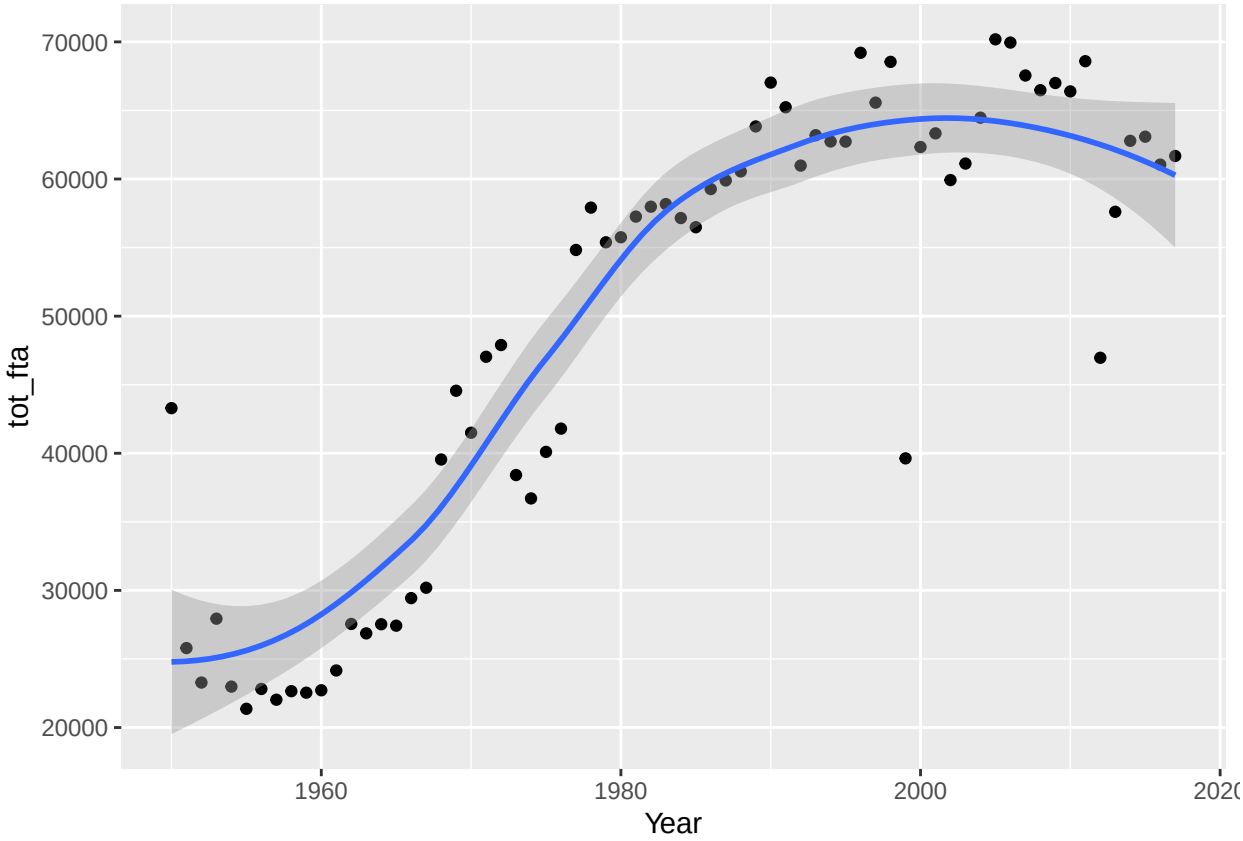
```
#total fouls each year
#teams played 82 games starting in 1966
Seasons_Stats %>%
  group_by(Year) %>%
  filter(Year > 1966) %>%
  summarize(tot_pf = sum(PF)) %>%
  ggplot(mapping = aes(x = Year, y = tot_pf)) +
  geom_point() +
  geom_smooth()
```

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




```
#total fta each year
Seasons_Stats %>%
  group_by(Year) %>%
  summarize(tot_fta = sum(FTA)) %>%
  ggplot(mapping = aes(x = Year, y = tot_fta)) +
  geom_point() +
  geom_smooth()
```

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

## Warning: Removed 1 rows containing non-finite values (stat_smooth).

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



What we find is the total number of free throws attempted has increased and the total number of personal fouls has also increased. What's surprising is that there is a dramatic jump in free throws attempted and personal fouls between the 1950s and 1980s. After doing more research, we found that early on, there were less NBA teams and less games being played. To account for this, we decided to take the total number of fouls and free throws in a season and divide it by the total number of games played that season. This would give us the the average number of personal fouls called and free throws attempted in a typical game that season. Using Excel, we created the csv file called Games. By using information online, we calculated the total number of NBA games played in each season. After joining the Games dataframe to Seasons_Stats, we can show the average fouls called per game and free throws attempted.

```
#read in total games played each season
Games <- read_csv("Data/Processed/Games.csv")
```

```
## Rows: 68 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): Year, GP_per_team, num_teams, tot_games
##
## 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.
```

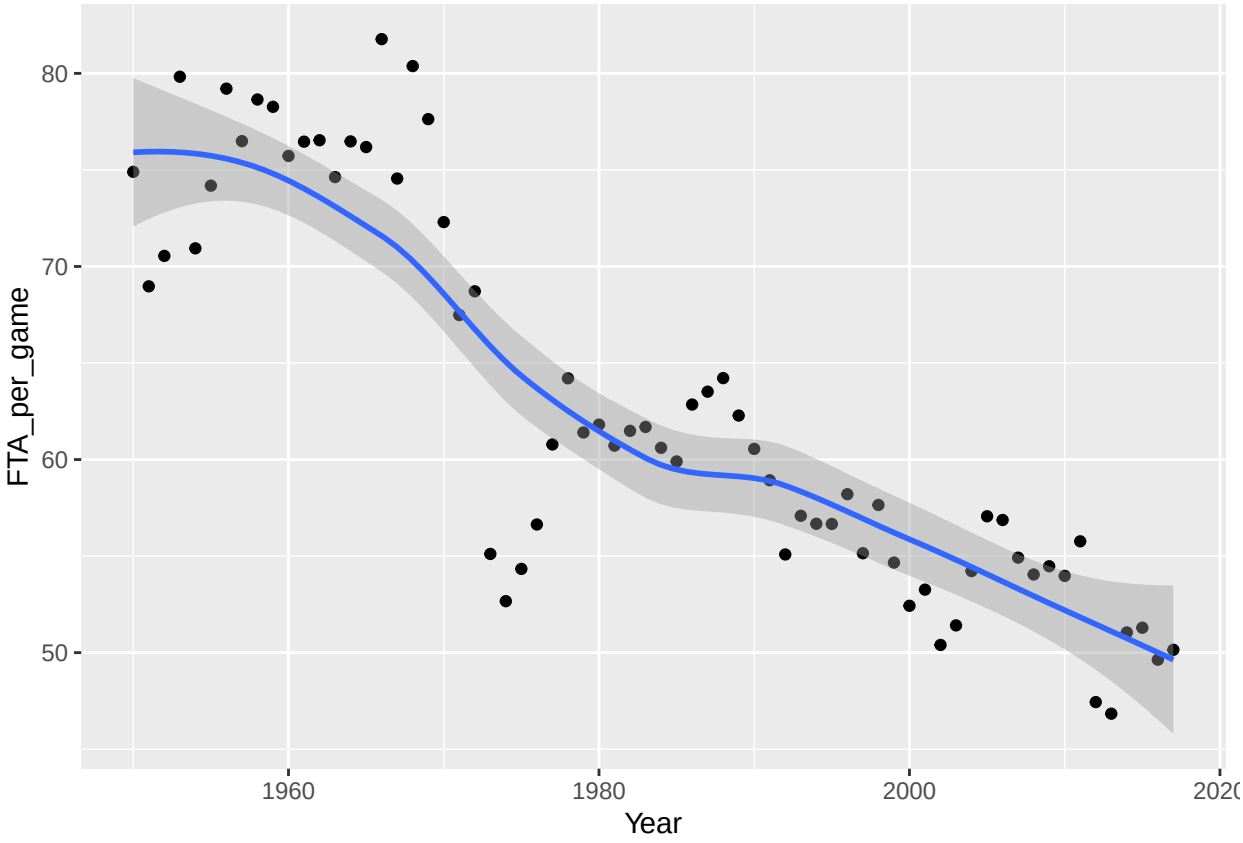
```
#average free throws attempted per game (total number of games played in a season)
Seasons_Stats %>%
  group_by(Year) %>%
  summarize(tot_FTA = sum(FTA)) %>%
  left_join(Games, by = "Year") %>%
```

```
  mutate(FTA_per_game = tot_FTA / tot_games) %>%
  ggplot(mapping = aes(x = Year, y = FTA_per_game)) +
  geom_point() +
  geom_smooth()
```

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

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

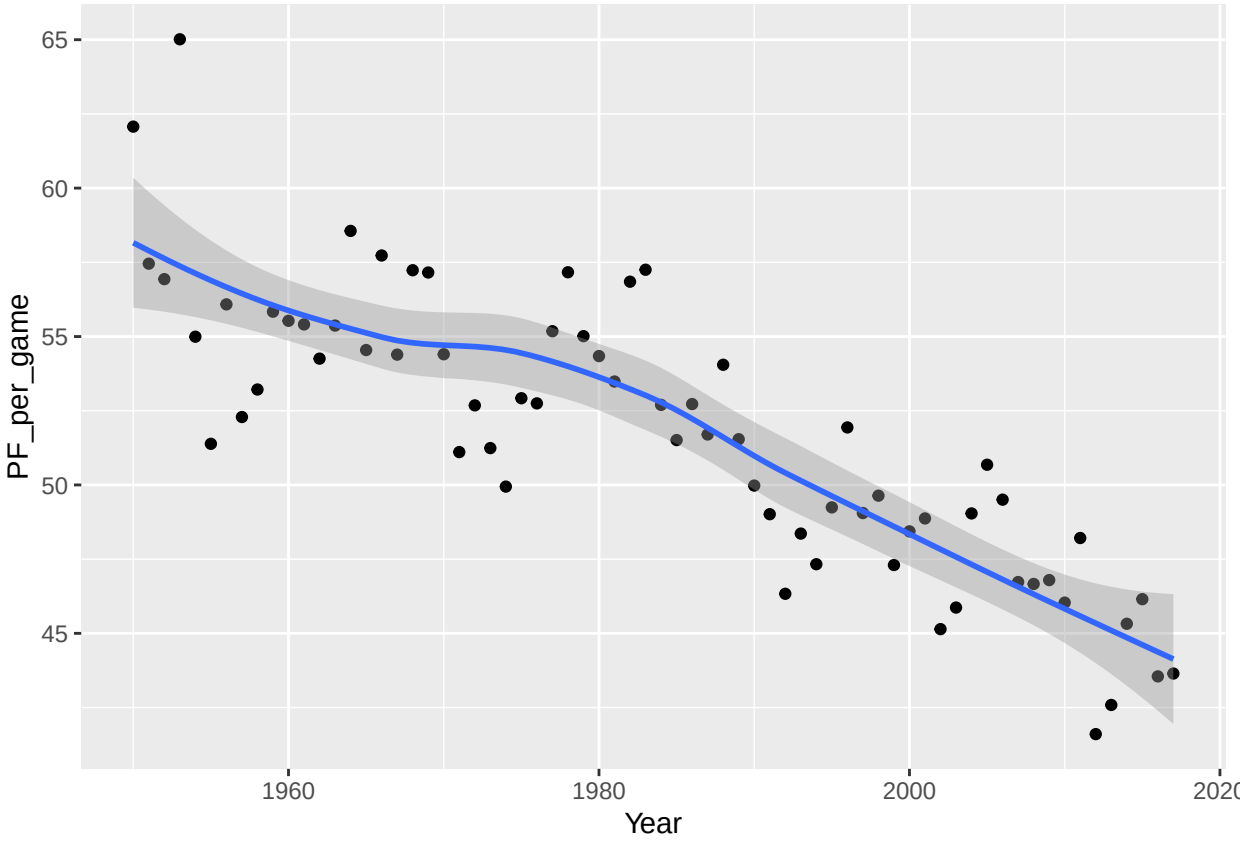
```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
#average personal fouls called per game (total number of games played in a season)
Seasons_Stats %>%
  group_by(Year) %>%
  summarize(tot_PF = sum(PF)) %>%
  left_join(Games, by = "Year") %>%
  mutate(PF_per_game = tot_PF / tot_games) %>%
  ggplot(mapping = aes(x = Year, y = PF_per_game)) +
  geom_point() +
  geom_smooth()
```

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

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
## Removed 1 rows containing missing values (geom_point).
```



By looking at the averages, we actually observe that the average number of personal fouls and free throws attempted each game in a season decreases as time progresses. This is particularly interesting because some may claim that back then, more physicality led to less fouls being called and the referees allowing the players to play with less restrictions. This certainly is not the case with average number of fouls being called per game and free throws attempted per game both trending downwards. One possible explanation could be that the NBA could be implementing rule changes (banning hand checking) or calling more flagrant fouls, preventing defenders from playing aggressively. As the defenders play less aggressively, less fouls would be called, and as a result, less free throws would be attempted.

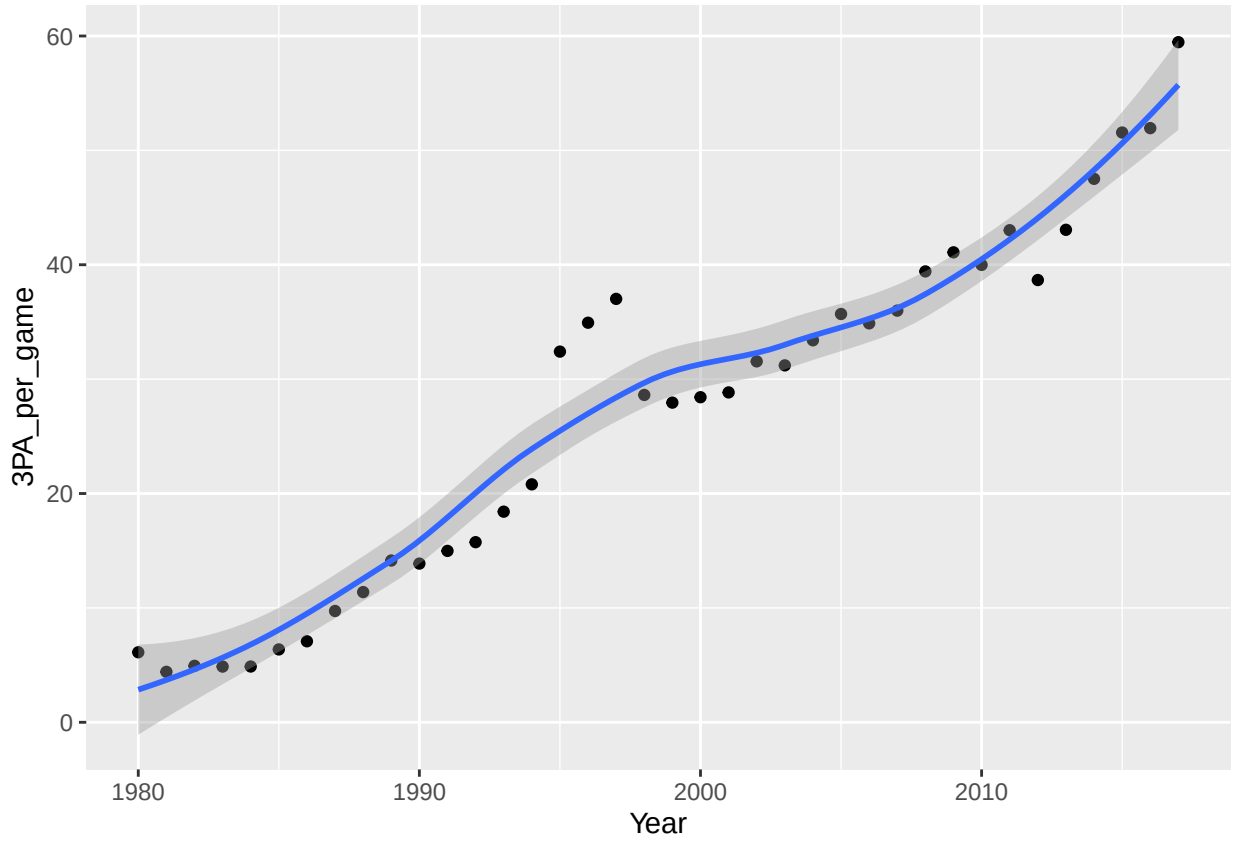
Implementation of 3-Pt Line

Spectators and analysts have also noted that NBA players are now shooting more three pointers than ever, not just the shooting guards. But the three point line has not always been in existence. It wasn't until the 1979-1980 season when the NBA first implemented the three point line. We wanted to see how the NBA players adjusted to this change. First we took a look at the average number of three point shots attempted in a typical game.

```
#converting character column to numeric column
Seasons_Stats$`3PA` <- as.numeric(as.character(Seasons_Stats$`3PA`))
Seasons_Stats$`3P` <- as.numeric(as.character(Seasons_Stats$`3P`))

#three pointers attempted, for all players
Seasons_Stats %>%
```

```
  filter(Year > 1979) %>%
  group_by(Year) %>%
  summarize(tot_3PA = sum(`3PA`)) %>%
  left_join(Games, by = "Year") %>%
  mutate(`3PA_per_game` = tot_3PA / tot_games) %>%
  ggplot(mapping = aes(x = Year, y = `3PA_per_game`)) +
  geom_point() +
  geom_smooth()
```

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

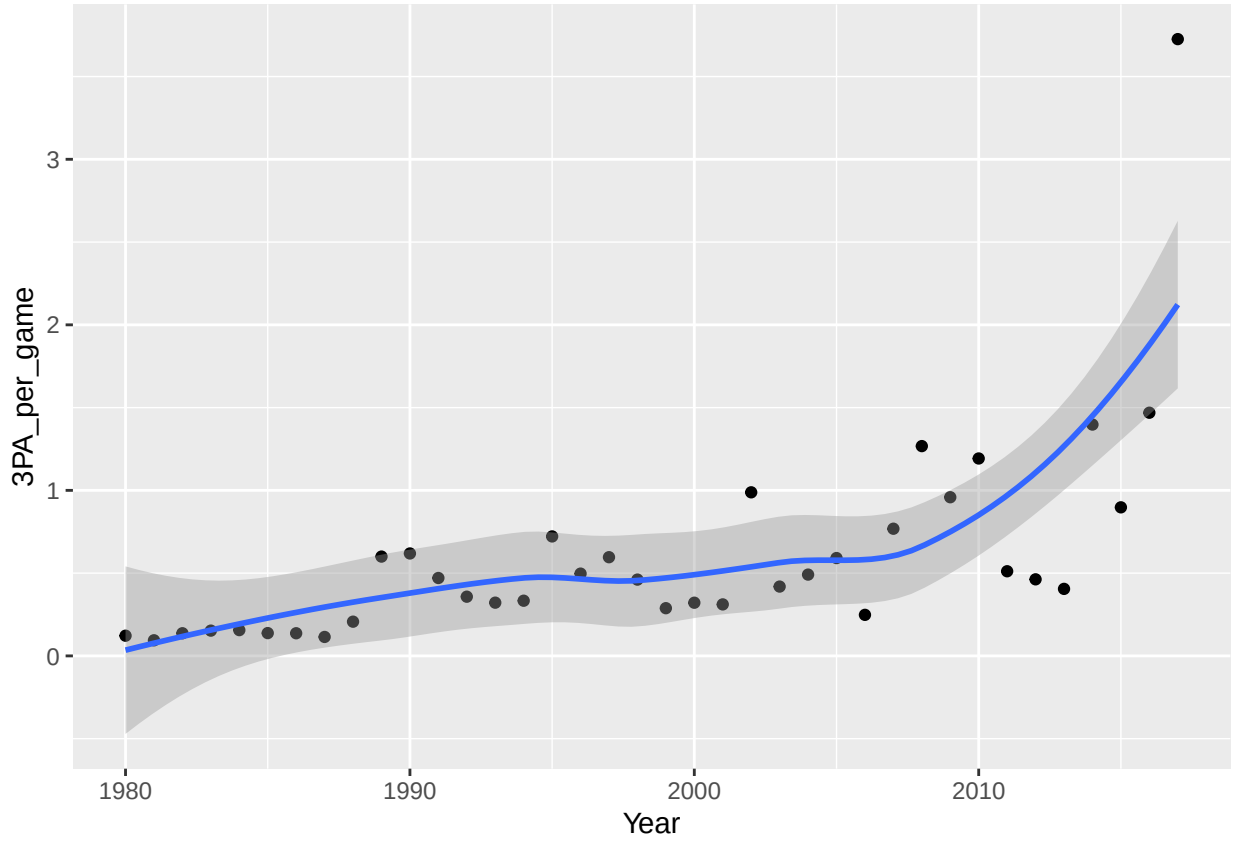


We see a clear upward trend in the number of three pointers attempted in an NBA game over the years. There is also no sign of the number of attempts plateauing, suggesting that the number of attempts will likely continue to increase the next several seasons. Now, let's take a look at the number of attempts by player position, using a function.

```
#by position
avg3PA_pos <- function(position) {
  Seasons_Stats %>%
    filter(Year > 1979) %>%
    filter(Pos == position) %>%
    group_by(Year) %>%
    summarize(tot_3PA = sum(`3PA`)) %>%
    left_join(Games, by = "Year") %>%
    mutate(`3PA_per_game` = tot_3PA / tot_games) %>%
```

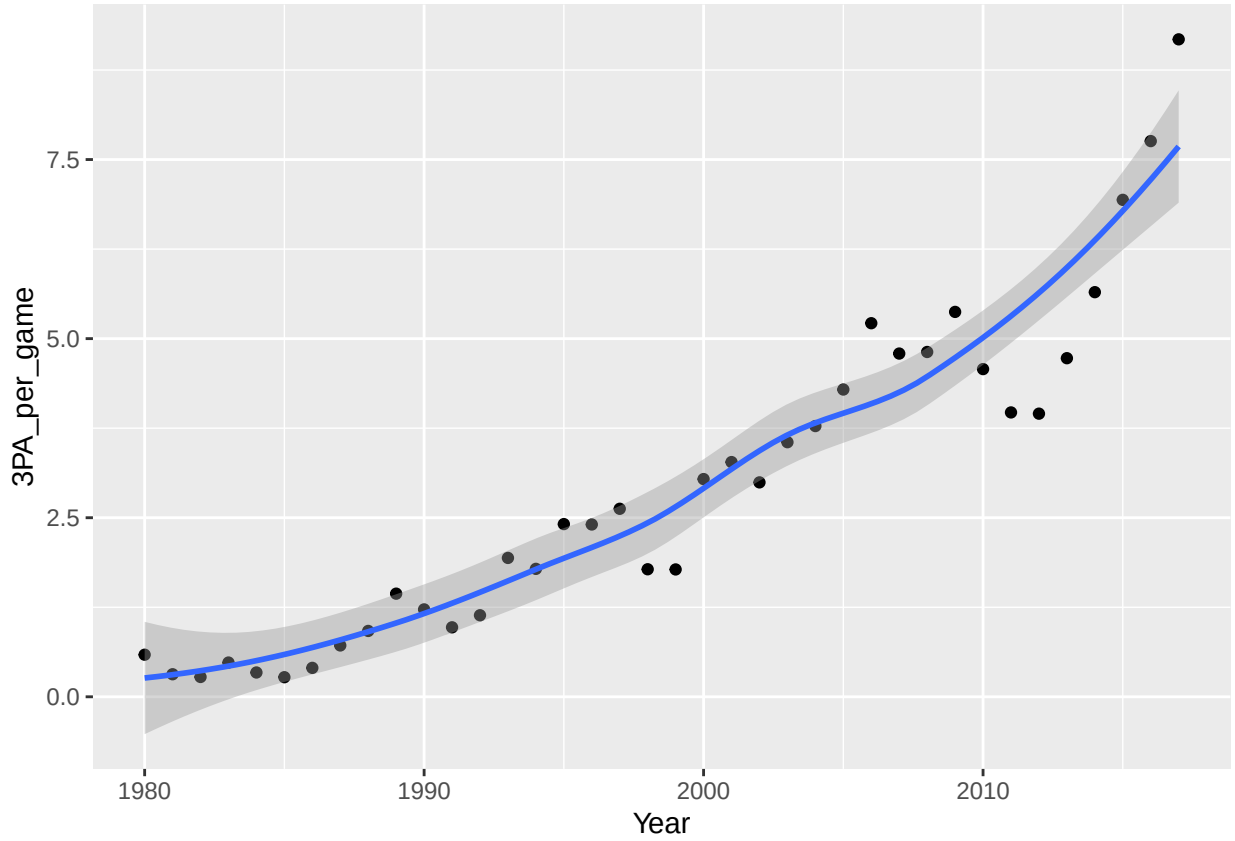
```
    ggplot(mapping = aes(x = Year, y = `3PA_per_game`)) +
    geom_point() +
    geom_smooth()
}

avg3PA_pos("C")
```

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



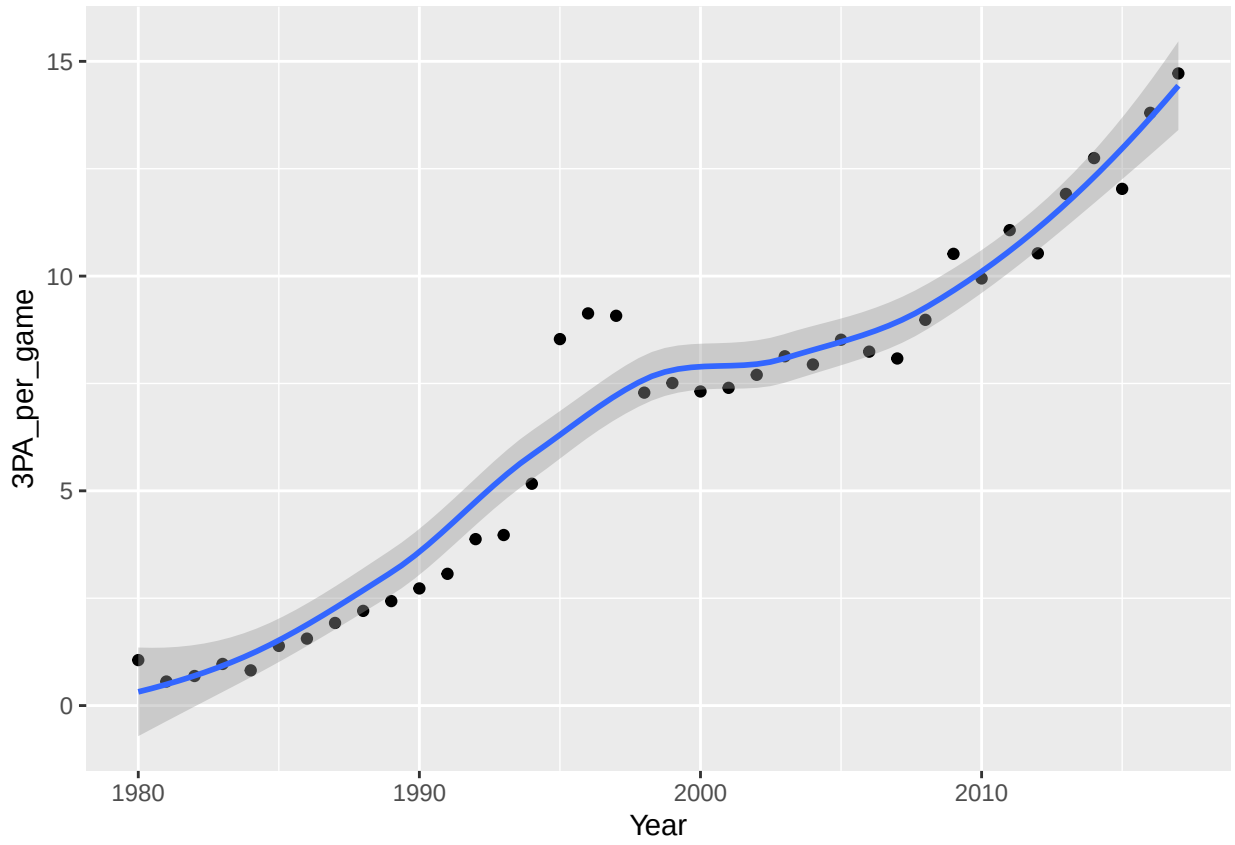

```
avg3PA_pos("PF")
```

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



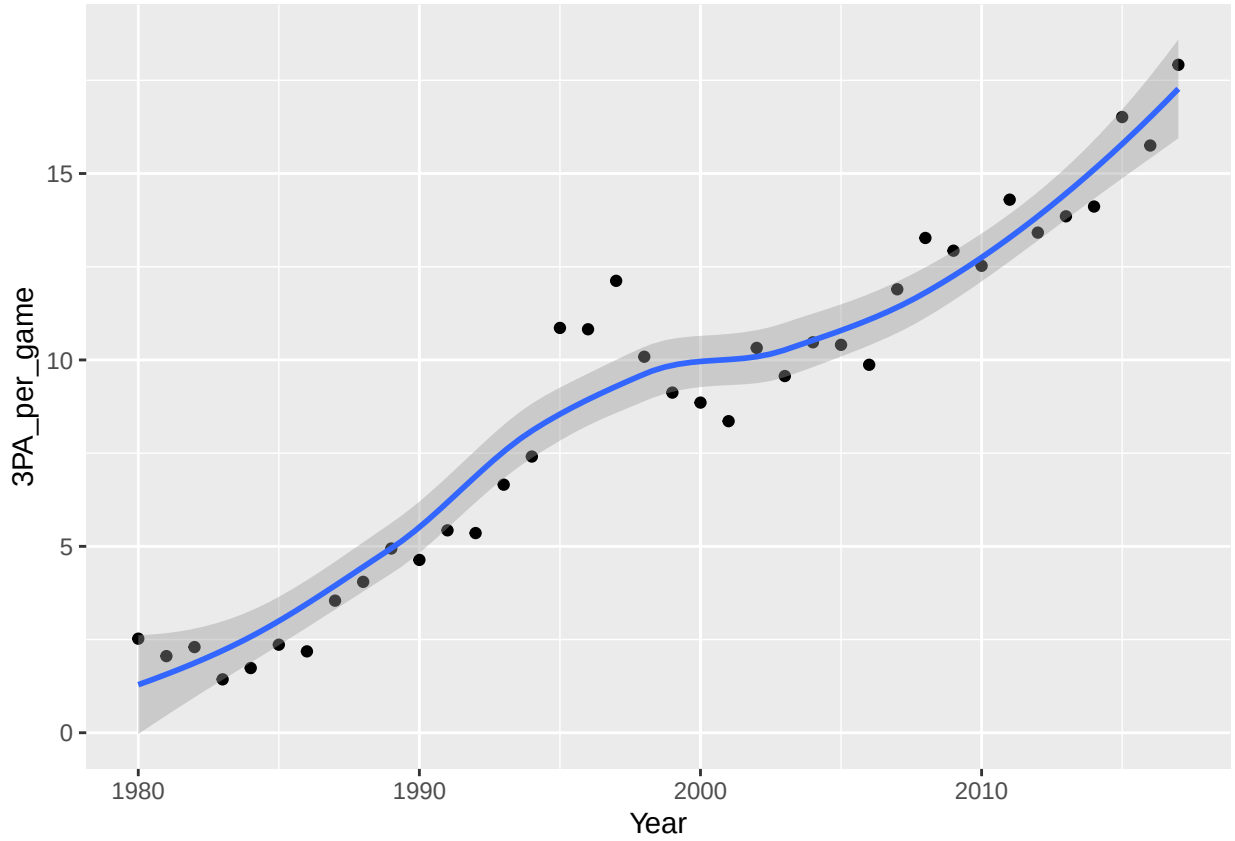

```
avg3PA_pos("SF")
```

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



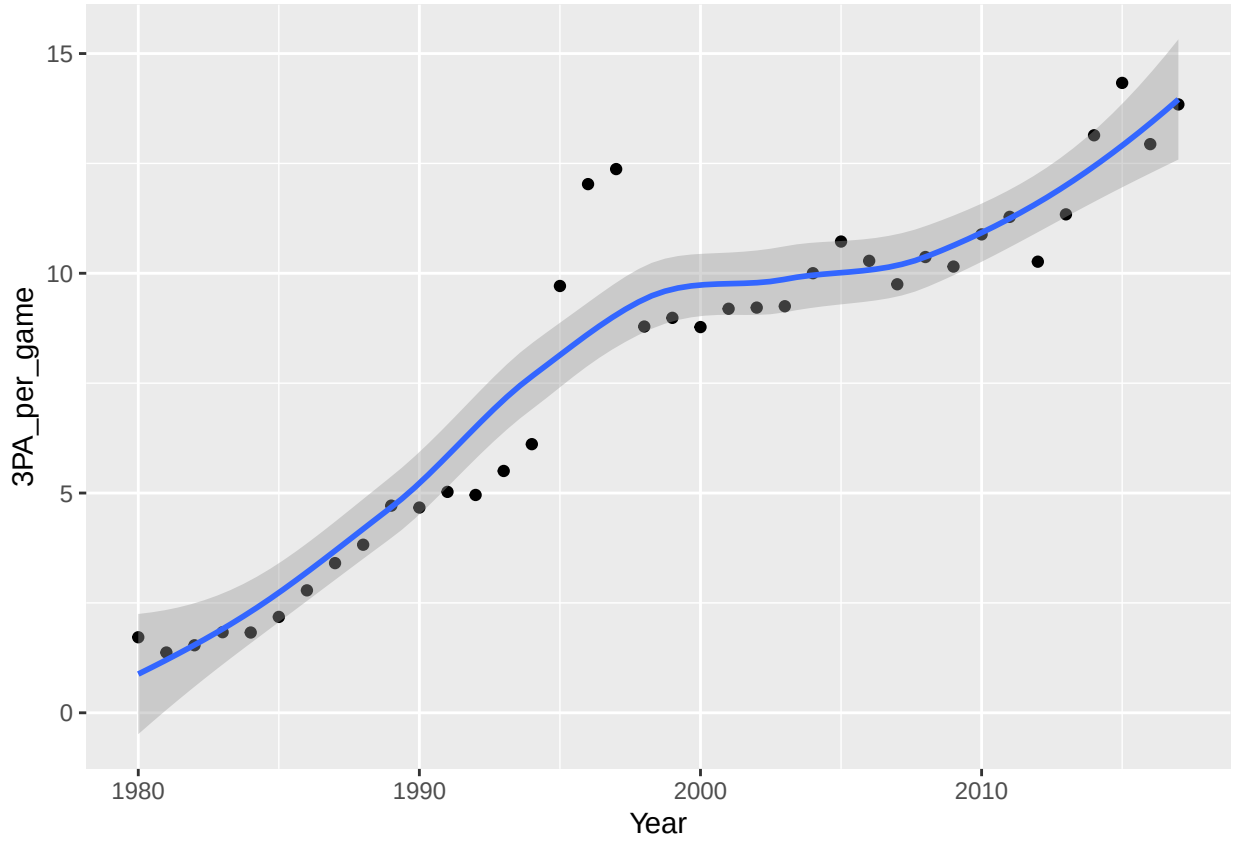

```
avg3PA_pos("SG")
```

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




```
avg3PA_pos("PG")
```

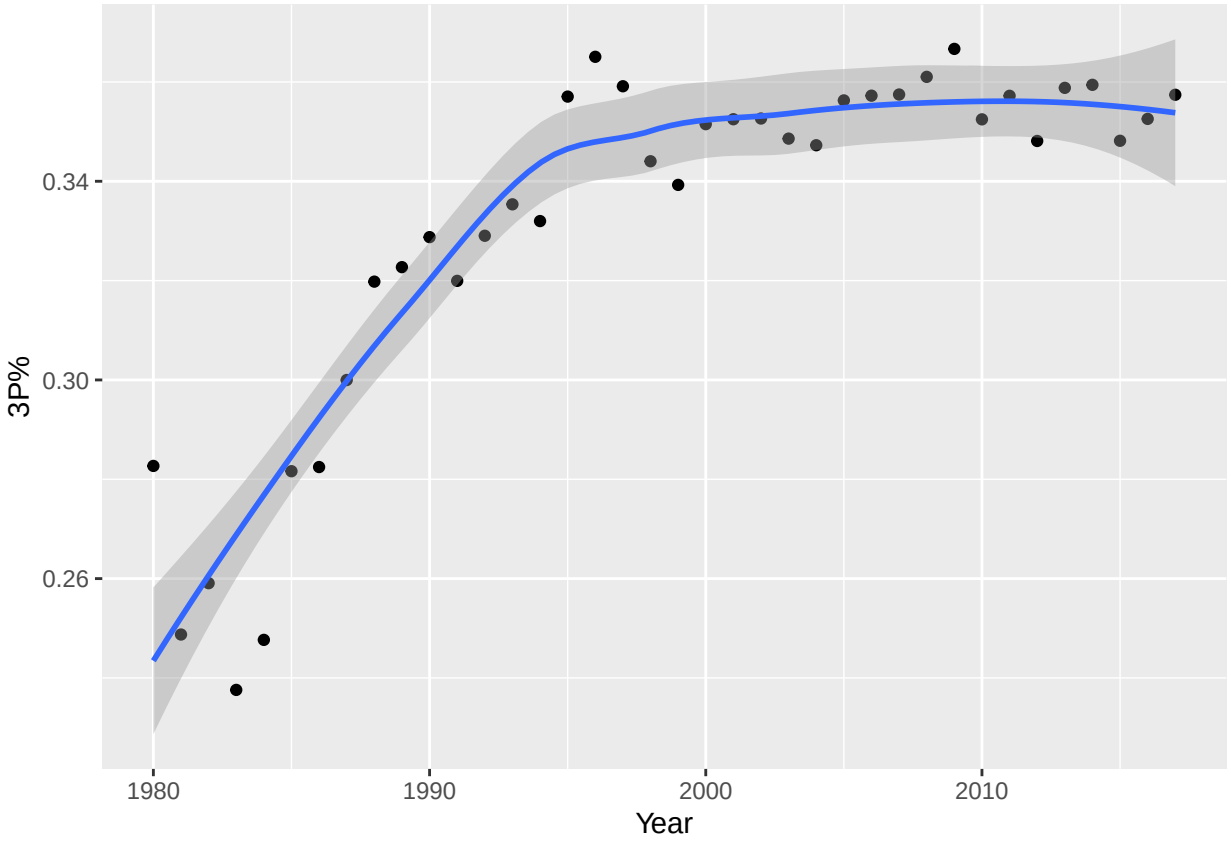
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Traditionally, Centers and Power Forwards tend to do less shooting as they were mostly situated around the paint area. But by looking at the number of three point attempts they are taking over the years, it seems that they are moving outwards and more willing to shoot from beyond the arc. What's especially interesting is that Centers have only started shooting more threes in recent years, with a large jump between the 2016 and 2017 season. Next, we wanted to see how accurate players were shooting the three point shot. We first took a look at how the players' three point field goal percentage has changed over the seasons. To calculate this, we divided the total number of three point makes by the total number of three point attempts each season.

```
#all players
Seasons_Stats %>%
  filter(Year > 1979) %>%
  group_by(Year) %>%
  summarize(tot_3PA = sum(`3PA`),
            tot_3PM = sum(`3P`)) %>%
  left_join(Games, by = "Year") %>%
  mutate(`3P%` = tot_3PM / tot_3PA) %>%
  ggplot(mapping = aes(x = Year, y = `3P%`)) +
  geom_point() +
  geom_smooth()
```

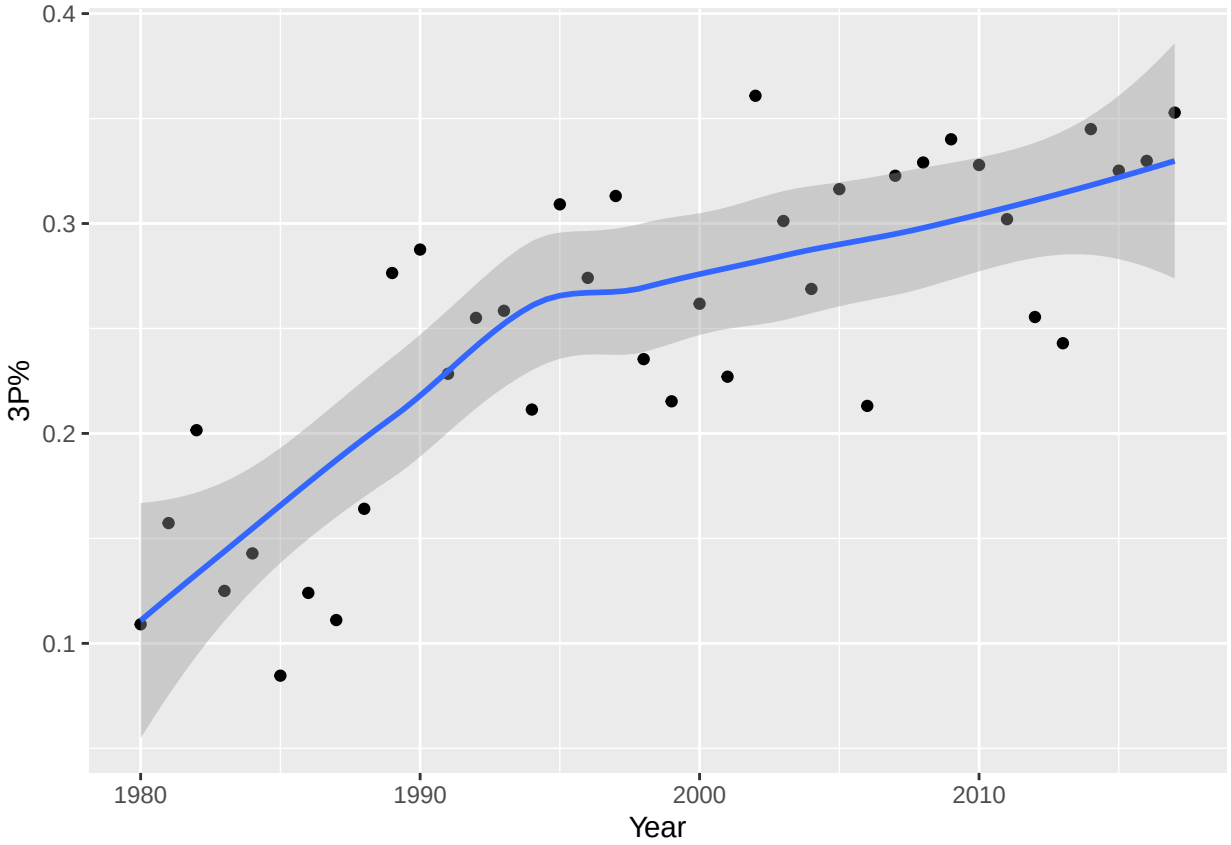
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



From the plot above, we can see that initially, players shot the three point shot at a relatively inaccurate percentage. By as time went on, players began to adjust and their three point percentage increased and plateaued by the mid 1990's. Ever since then, the league average hovers between 34% and 36%. Despite the players shooting the three ball more than ever before, their three point field goal percentage has been relatively the same since the mid 1990's. Now, let's take a look at how accurately players from each position shoots the three point shot.

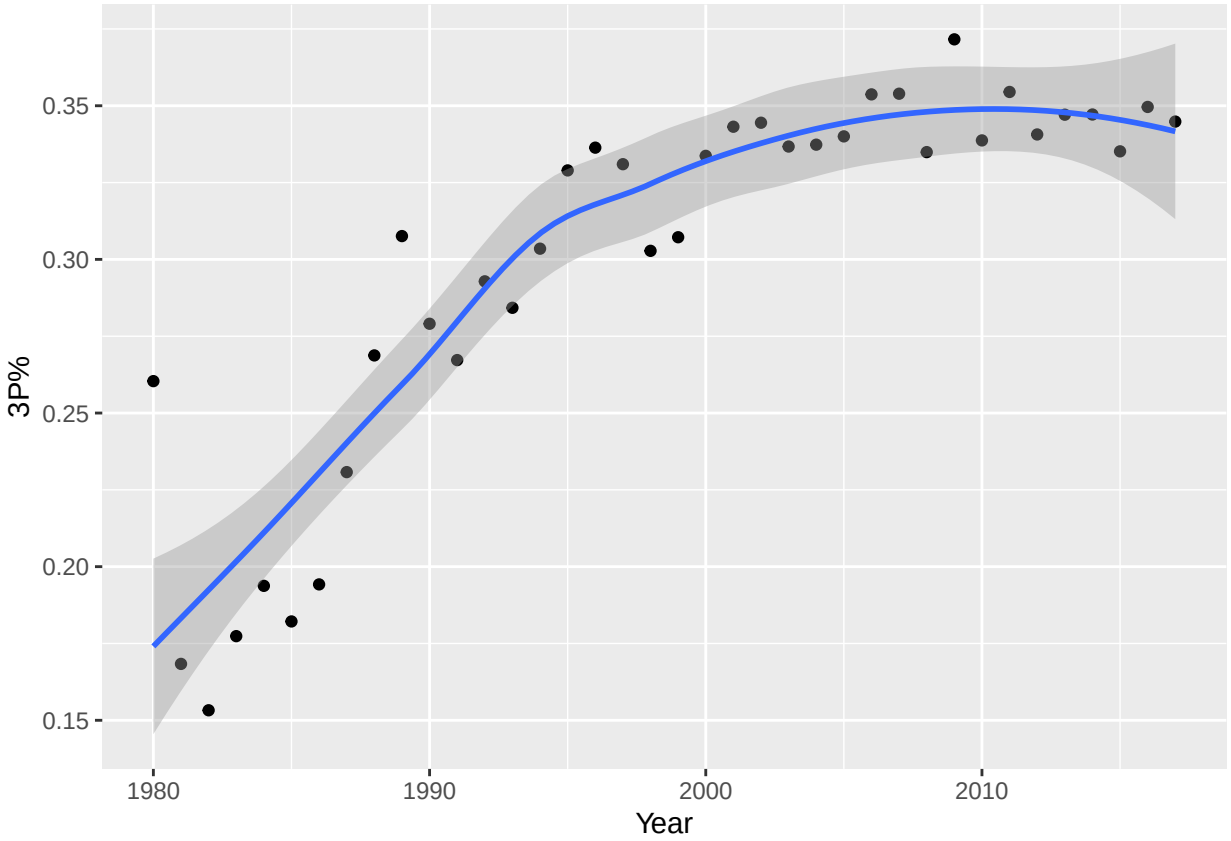
```
#by position
threepointperc <- function(position) {
  Seasons_Stats %>%
    filter(Year > 1979) %>%
    filter(Pos == position) %>%
    group_by(Year) %>%
    summarize(tot_3PA = sum(`3PA`),
              tot_3PM = sum(`3P`)) %>%
    left_join(Games, by = "Year") %>%
    mutate(`3P%` = tot_3PM / tot_3PA) %>%
    ggplot(mapping = aes(x = Year, y = `3P%`)) +
    geom_point() +
    geom_smooth()
}

threepointperc("C")
```

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



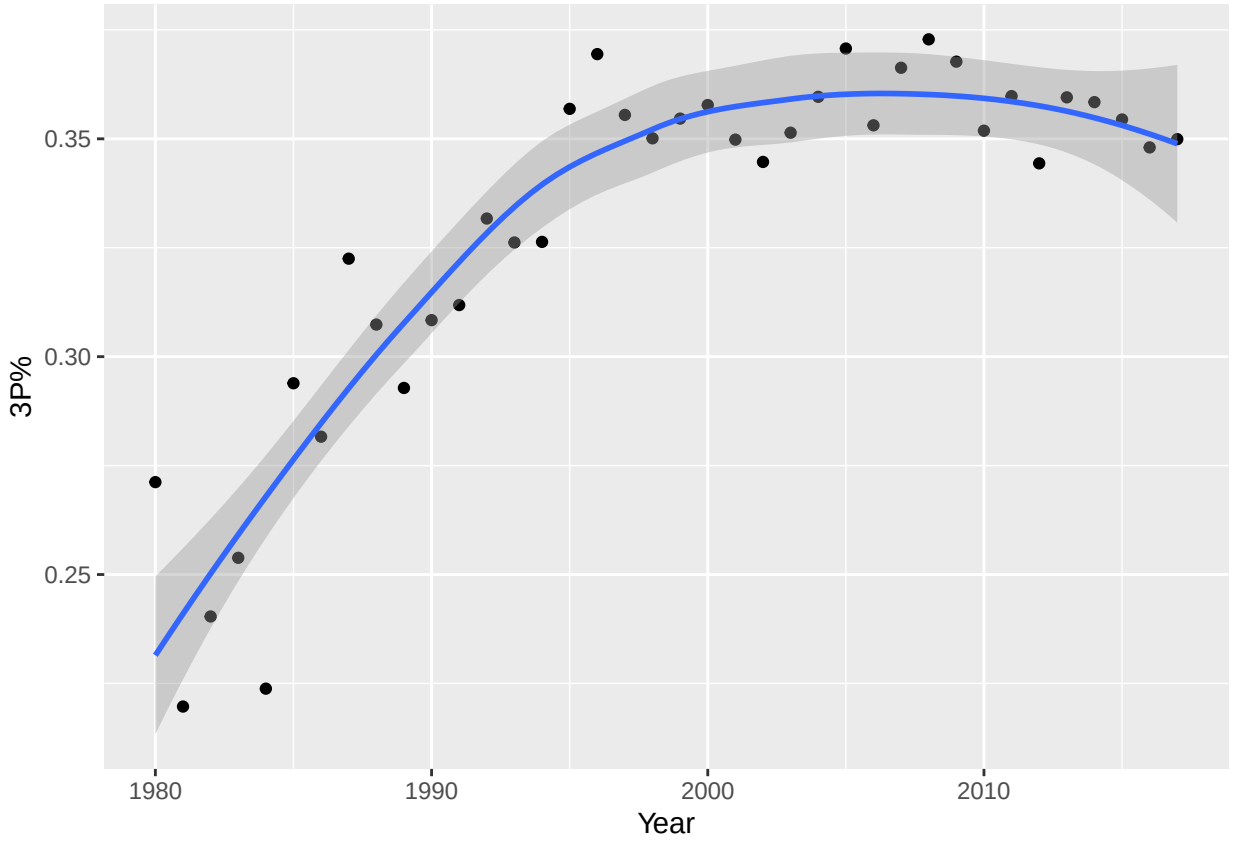

```
threepointperc("PF")
```

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



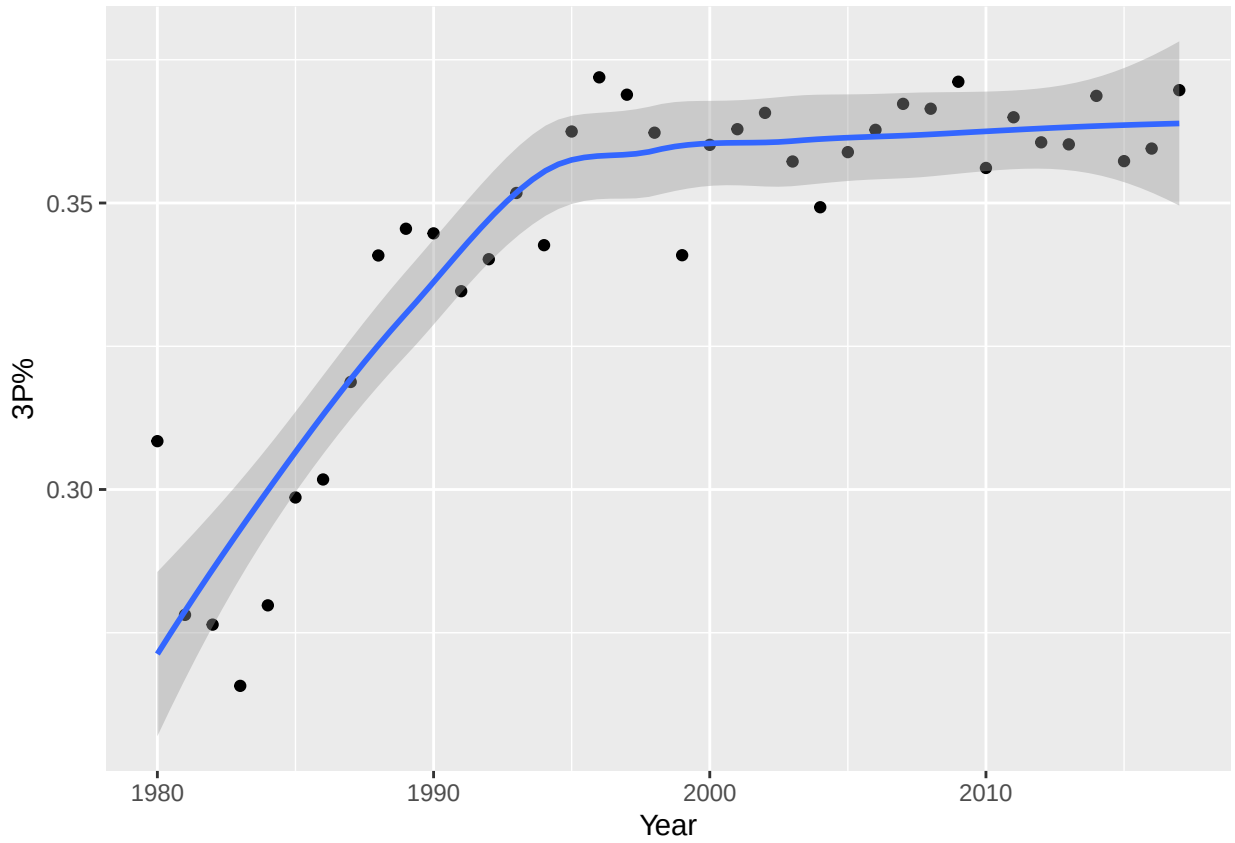

```
threepointperc("SF")
```

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



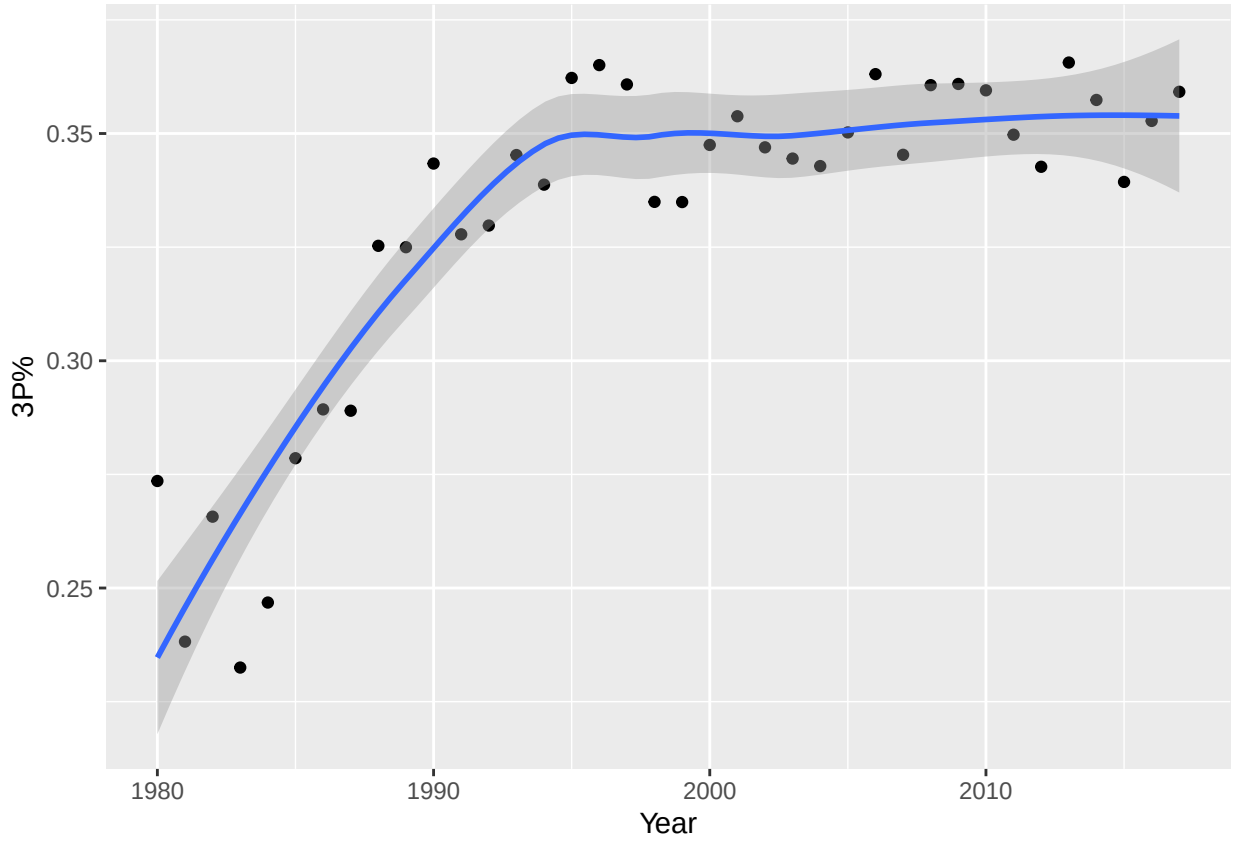

```
threepointperc("SG")
```

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




```
threepointperc("PG")
```

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



We observe a similar pattern from each of the positions, with the exception of the Center position. Point Guards, Shooting Guards, Small Forwards, and Power Forwards have all initially increased their three point shooting percentage and have plateaued around the mid 1990's. They all have plateaued at around the 35% to 37% mark. The Center position is the only position whose three point percentage is still trending upwards towards the 35% mark.

Which colleges produced the best basketball players?

Certain colleges dominate the scene in the NCAA on a yearly basis, such as Duke and Kentucky. But do these schools hone players' talent that translate to the NBA level? To measure how good a NBA players is, we first decided to measure their Player Efficiency Rating (PER). We decided to take the median PER for players who went to each college that was recorded in the Players dataframe.

```
#PER

#median PER
PER_stats <- Seasons_Stats %>%
  left_join(Players, by = "Player") %>%
  group_by(collage) %>%
  summarize(med_PER = median(PER, na.rm = TRUE)) %>%
  arrange(desc(med_PER))

PER_stats
```

```
## # A tibble: 423 x 2
##    collage                          med_PER
```

```
##    <chr>                                <dbl>
##  1 United States Naval Academy           26.8
##  2 Belmont Abbey College                 22.2
##  3 Truman State University               22
##  4 Master's College                      21.2
##  5 University of Wisconsin-River Falls   21.1
##  6 Wheaton College                       20.8
##  7 Gardner-Webb University               20.2
##  8 Indiana State University              20.2
##  9 Trinity Valley Community College      19
## 10 Lawrence Technological University     18.9
## # ... with 413 more rows
```

What we find is a bit troubling, as many of these top colleges are not very well-known. A possible cause for this phenomenon is that only one or two really good NBA players have come from these colleges. To counter this effect, we decided to weight the median PER by the proportion of NBA seasons accounted for by a college. This means that colleges that produce more NBA talent will have their weighted PER to be more significant.

```
#total number of player seasons
count(Seasons_Stats)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 24691
```

```
#24691

#find proportion of seasons from each college
prop_seasons <- Seasons_Stats %>%
  left_join(Players, by = "Player") %>%
  group_by(collage) %>%
  count() %>%
  mutate(prop_seasons = n /24691)

#join back to average PER, weighted top colleges
PER_stats %>%
  left_join(prop_seasons, by = "collage") %>%
  mutate(wPER = med_PER * prop_seasons) %>%
  arrange(desc(wPER))
```

```
## # A tibble: 423 x 5
##    collage                              med_PER     n prop_seasons  wPER
##    <chr>                                  <dbl> <int>        <dbl> <dbl>
##  1 <NA>                                    13.4  2290       0.0927 1.24
##  2 University of California, Los Angeles   13.4   627       0.0254 0.339
##  3 University of North Carolina            13.7   556       0.0225 0.309
##  4 University of Kentucky                  12.8   533       0.0216 0.277
##  5 University of Kansas                    12.5   410       0.0166 0.208
##  6 Duke University                         14     354       0.0143 0.201
##  7 University of Michigan                  13.8   316       0.0128 0.177
##  8 Michigan State University               13.2   315       0.0128 0.168
```

```
##  9 University of Arizona                    12.6   324       0.0131 0.165
## 10 St. John's University                    13.2   300       0.0122 0.160
## # ... with 413 more rows
```

After weighting the PER, now we see a more accurate representation of the colleges that produced the best NBA talent. Now let's see what are the best colleges measured by the players' Win Shares.

```
#win shares

#median win shares
WS_stats <- Seasons_Stats %>%
  left_join(Players, by = "Player") %>%
  group_by(collage) %>%
  summarize(med_WS = median(WS, na.rm = TRUE)) %>%
  arrange(desc(med_WS))

#weighted win shares
WS_stats %>%
  left_join(prop_seasons, by = "collage") %>%
  mutate(wWS = med_WS * prop_seasons) %>%
  arrange(desc(wWS))
```

```
## # A tibble: 423 x 5
##    collage                              med_WS     n prop_seasons    wWS
##    <chr>                                 <dbl> <int>        <dbl>  <dbl>
##  1 <NA>                                    1.8  2290       0.0927 0.167
##  2 University of California, Los Angeles   1.8   627       0.0254 0.0457
##  3 University of North Carolina            2     556       0.0225 0.0450
##  4 University of Kentucky                  1.6   533       0.0216 0.0345
##  5 Duke University                         2.2   354       0.0143 0.0315
##  6 Ohio State University                   2.3   265       0.0107 0.0247
##  7 Georgia Institute of Technology         2     289       0.0117 0.0234
##  8 University of Michigan                  1.8   316       0.0128 0.0230
##  9 University of Arizona                   1.65  324       0.0131 0.0217
## 10 University of Notre Dame                1.7   306       0.0124 0.0211
## # ... with 413 more rows
```

When ranking by weighted win shares, we see almost the same colleges when ranking by weighted PER. One thing to note is that many players either came to the NBA straight out of high school or did not have their college recorded in the Players dataframe. This is why we see NA as the top result.