# STA 486C NOT Project 1

Jacob Berlin

2025-02-07

## Introduction

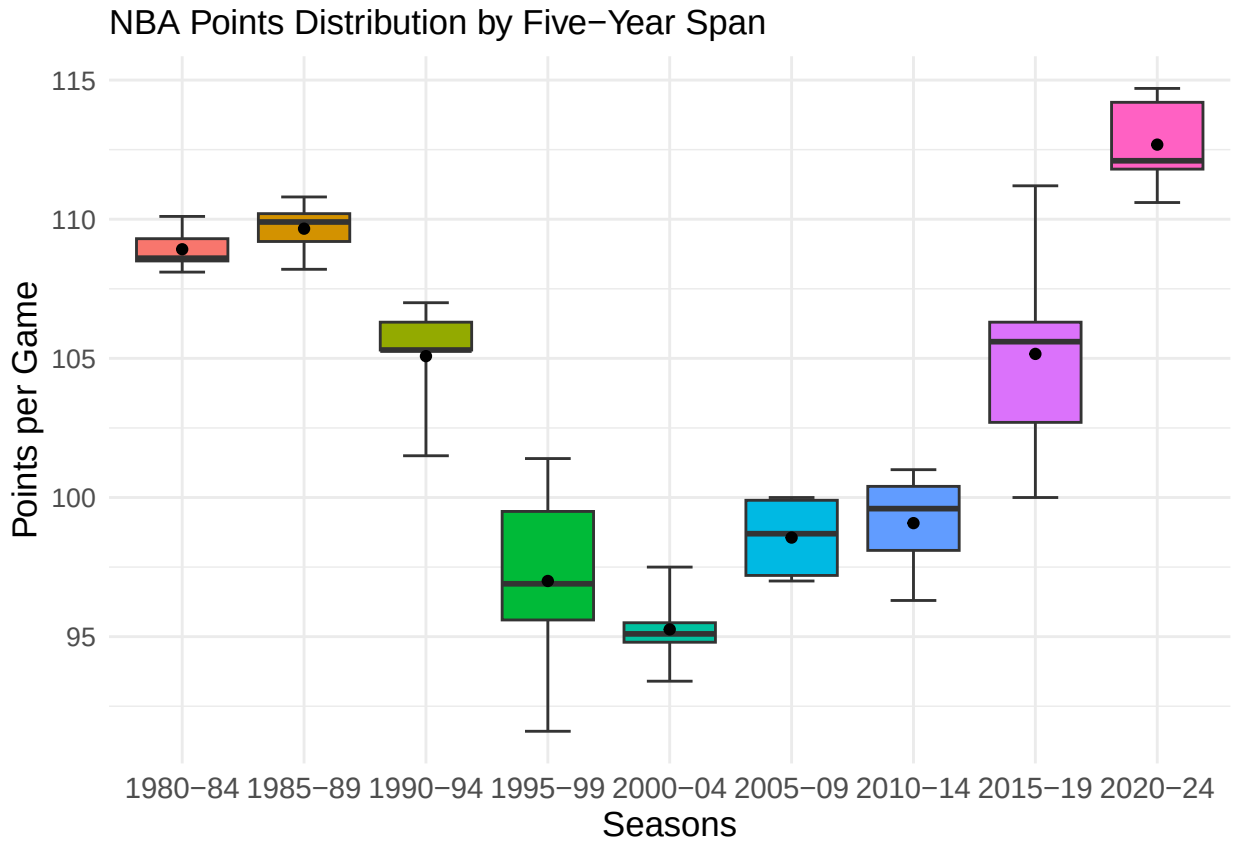
I chose a data set including every NBA team's basic box score stats from each season. The seasons ranged from 1947-2025. I chose this data set simply because I am a big fan of basketball and was curious to see how the statistics of the NBA has changed since the 3-point line was introduced. I was also hoping to learn if the introduction of the 3-point line directly affected other stats tracked like personal fouls, points scored, or free throws attempted. In regards to the accuracy and legitimacy of the data, I am very confident that it is correct and real. Although the data was downloaded directly from kaggle.com, the data was said to be taken directly from basketball-reference.com which is arguably the most accurate data site for basketball statistics (besides the official NBA website). I made sure to double check and compare the data downloaded from Kaggle to the basketball reference data and they matched up, so I can confidently say that this data is accurate.

## Data Cleaning

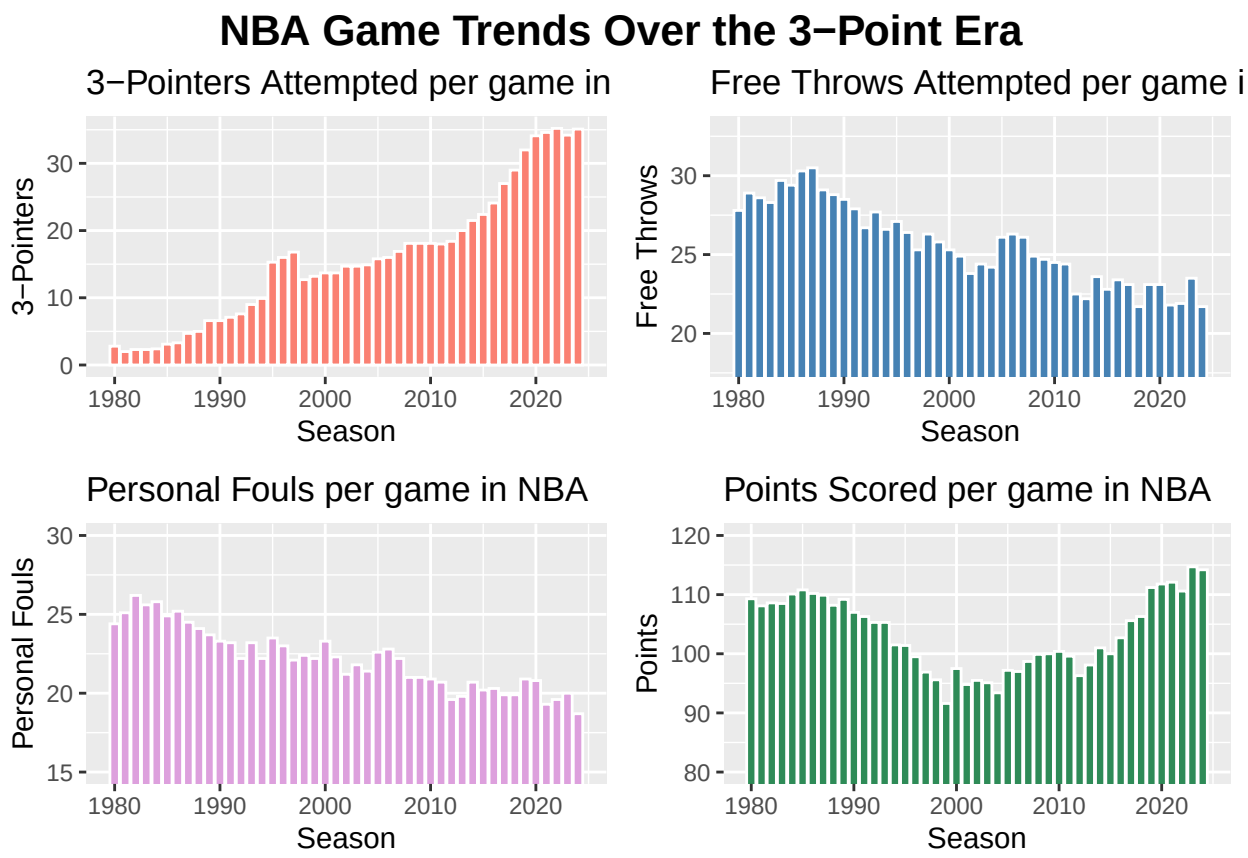
The primary cleaning that was necessary for the data was trimming down the amount of seasons and variables I would need for the analysis. The original data ranged from 1947 to 2025, but I cut down the range to 1980 to 2024. I did this because as of writing this the 2025 NBA season is still going on and I wanted to use data from complete seasons. I cut off everything before 1980 because 1980 was the year the 3-point line was introduced. The original data set also had 28 variables and I only kept the 6 that suited my needs. Because the NBA is so good at recording data there weren't any NA's or major problems I had to deal with, so cleaning (or more accurately wrangling in this case) just included me taking out anything I found unnecessary as well as adding in one column. The one column I added in assigns a five-year span classification for each season (this column isn't initially important but it was necessary fo the development of a boxplot). The key variables in the data set include season, x3pa_per_game, fta_per_game, pf_per_game, and pts_per_game. The season column just indicates what year that specific season and stats took place in while the other variables are as follows: x3pa_per_game is 3-pointers attempted per game, fta_per_game is free throws attempted per game, pf_per_game is personal fouls committed per game, and pts_per_game is points scored per game. Below is a look at the first row of the data set for the purpose of understanding the data's format.

| | season | team | x3pa_per_game | fta_per_game | pf_per_game | pts_per_game | FiveSpan |
|---|---|---|---|---|---|---|---|
| 62 | 2024 | League Average | 35.1 | 21.7 | 18.7 | 114.2 | 2020-24 |

## Data Visualizations




(Fig. 1) A box plot showing the league average for points scored per game in five-year spans. Black dot represents the mean while black line represents median. Whisker caps represents maximum/minimum points scored for each time span.




(Fig. 2) A figure containing four different graphs. Each graph demonstrates how different statistics recorded in the NBA have changed since 1980.

Figure 1 is a box plot that displays the average amount of points scored per game in five-year segments from 1980 to 2024. We can see a parabola-esque shape formed by the trend of the plots, with the average amount of points scored being around 110 in the 1980's and gradually declining when entering the the 1990's and early 2000's. A gradual increase begins during the 2005-2009 span and shoots up significantly with some variance in between 2015 and 2019. The plot ends with the 2020-2024 time span where scoring is consistently the highest. It is interesting to see how consistent the scoring averages were in the 1980's compared to how inconsistent they are in the 1990's and 2010's. Also, considering how the skill level among players and teams are constantly increasing, I find it interesting that the points scored took a massive dip in the early 200's. I was curious to see why this may have been the case so I looked at how other stats including points scored per game developed through time in Figure 2.

Figure 2 shows how different parts of basketball have changed throughout time, not just points scored. The top left graph shows on average how many 3-pointers have been taken since 1980. For the most part, the amount of threes being shot has increased each season with a few spikes. The graphs following the rate of free throws attempted and personal fouls committed are very similar looking as they are highly correlated. Interestingly enough they have both consistently decreased through time which is both surprising and not surprising. It is somewhat surprising because a popular narrative in today's era of basketball is that more "light" fouls are being called meaning that referees are calling fouls on only somewhat physical plays that don't meet the criteria of being a foul. A reason it isn't surprising though is that due to the higher rate of threes being shot, players are pushing their way to the hoop and drawing contact from their opponents. Naturally, more outside shots are going to lead to less free throws and personal fouls committed. The bottom right graph is similar to Figure 1 in regards to it tracking points scored per game but in this format it is easier to see how the other stats affect its own progression. Considering that more threes were being taken during the 2000's as well as still a fair amount of free throws being attempted it is interesting to see that the 2000's suffers from such a decline in points.

# Conclusions

If I had more time with this data, I would like to take a deeper dive into why there is such a significant dip in scoring in the late 90's and early 2000's. Maybe overall field goal percentage took a dive due to the increase in threes or maybe there were more turnovers being forced with more intense defense. A challenge I could see in further analysis would be the inclusion of the early years of the NBA. Some statistics weren't actively tracked like blocks, steals, and turnovers so it would be hard to see how those progressed throughout all of NBA history instead of just from 1980 to present day.

## GitHub Repo

https://github.com/jgberlin2003/NBAStatistics