NBA: A Look into the Three-Pointer (Part 1)

Quang Nguyen

Snippet of the Dataset

```
##   Year          Player Pos Age  Tm  G GS MP  FG FGA   FG. X3P X3PA X3P.
## 1 1950 Curly Armstrong G-F  31 FTW 63 NA NA 144 516 0.279  NA   NA   NA
## 2 1950    Cliff Barker  SG  29 INO 49 NA NA 102 274 0.372  NA   NA   NA
## 3 1950   Leo Barnhorst  SF  25 CHS 67 NA NA 174 499 0.349  NA   NA   NA
## 4 1950      Ed Bartels   F  24 TOT 15 NA NA  22  86 0.256  NA   NA   NA
## 5 1950      Ed Bartels   F  24 DNN 13 NA NA  21  82 0.256  NA   NA   NA
## 6 1950      Ed Bartels   F  24 NYK  2 NA NA   1   4 0.250  NA   NA   NA
##   X2P X2PA  X2P.  FT FTA   FT. ORB DRB TRB AST STL BLK TOV  PF PTS
## 1 144  516 0.279 170 241 0.705  NA  NA  NA 176  NA  NA  NA 217 458
## 2 102  274 0.372  75 106 0.708  NA  NA  NA 109  NA  NA  NA  99 279
## 3 174  499 0.349  90 129 0.698  NA  NA  NA 140  NA  NA  NA 192 438
## 4  22   86 0.256  19  34 0.559  NA  NA  NA  20  NA  NA  NA  29  63
## 5  21   82 0.256  17  31 0.548  NA  NA  NA  20  NA  NA  NA  27  59
## 6   1    4 0.250   2   3 0.667  NA  NA  NA   0  NA  NA  NA   2   4
```

Description of the Dataset - NBA Statistics from 1950 to 2017

Background Information

This dataset is an extensive collection of individual statistics for basketball players in the NBA from 1950 to 2017. These individual statistics were compiled from gamelogs, box-scores, and with access to the NBA developer tools (*1). The variables and relevant metrics that are included in this dataset for use in this project are included down below. Generally, the metrics that are being used measure basic statistical data concerning basketball such as points, free-throws, field-goals, attempts, minutes played, and more that can be used to find trends and correlations.

The overall project will focus on demonstrating how the game of basketball has changed as a whole over time specifically focusing on the introduction of the three-point shot as well as how games have become faster and higher scoring. Other contexts that will be explored

include the physicality of the game (using personal fouls as the indicator) and how the game has evolved to either become more offensive or defensive (based off of several metrics such as offensive rebounds versus defensive rebounds). For this portion of the project, the dataset will be used to lay the foundation for identifying correlating trends between the introduction of the three-point shot and the gradual increase of points scored through the NBA. More details will be provided about the dataset and certain methods in later sections.

Rows and Variables

Each row within the raw dataset represents an individual player within the respective season. Each column within the raw dataset for the most part represents a unique statistic tracked by basketball analysts to help generate performance metrics of how well individual basketball players play. Of course, there are other factors that contribute to how effective a player is in the game, but these numbers provide a general overview of the performance of a player. Each column variable is described below (*2):
"Year" - Season (recorded as the year in which the season ended so for 1949-1950 season, 1950 is reported) "Player" - Name of the player
"Pos" - Position of the player (could have more than one position)
"Age" - Age of the player during that season
"Tm" - Team that the player is on
"G" - Total amount of games that the player played in
"GS" - Total amount of games that the player started
"MP" - Total amount of minutes that the player played throughout the whole season "FG" - Total field goals made "FGA" - Total field goals attempted
"FG." - Field goal percentage
"X3P" - Total three-pointers made "X3PA" - Total three-pointers attempted
"X3P." - Three-point shot percentage
"X2P" - Total two-pointers made
"X2PA" - Total two-pointers attempted
"X2P." - Two-point shot percentage
"FT" - Total free throw shots made "FTA" - Total free throw shots attempted
"FT." - Free throw percentage
"ORB" - Total amount of offensive rebounds
"DRB" - Total amount of defensive rebounds
"TRB" - Total amount of rebounds
"AST" - Total amount of assists

"STL" - Total amount of steals
"BLK" - Total amount of blocks
"TOV" - Total amount of turnovers
"PF" - Total amount of personal fouls
"PTS" - Total amount of points

Collection of the Data

There are inconsistencies within the collection of data that exist that will be discussed in a later section. The data when it was collected was considered a population and contains statistics from all NBA players from 1950 to 2017. The time range chosen begins when the teams from the Basketball Association of America and the National Basketball League consolidated into the National Basketball Association prior to the 1950 season, although the NBA recognizes 1947 as the inaugural season (*3). The time range ends when the data was collected and this potential issue will be raised in a later section as well. Regarding how the data was specifically collected, the NBA's methods of maintaining statistical records have largely remained the same despite the advancement of technology. Teams of people are hired specifically to watch the basketball game and make records that can be uploaded to their database. These people are consistent in their approach as other members of the team are constantly cross-referencing and ensuring that the data is accurate. As a result, although certain metrics have been added and made more advanced which results in some null values for the earlier years, the basic information needed has been thoroughly collected in this population.

For the purpose of my project, not all statistical recordings will be used. In the pre-subsetted dataframe, there were variables that were advanced basketball metrics such as the player efficiency rating and the offensive box plus/minus. If needed, I will calculate them based on the basic statistics that remain in the subsetted dataframe by using established formulas accepted by professional analysts. For the purpose of this investigation however, all that is needed is identifying information such as year, name, and team as well as the statistic counts for individual contributions throughout the duration of their season. Within the dataset, there are some NA values but these can also be of use to this investigation as rule changes were introduced that led to the creation of new statistic measures.

Potential Issues

This dataset is fairly complete and exhaustive including metrics that will not be needed for this project. However, there is some missing data from the early stages of the NBA when data being stored was either not electronic, full records were not kept, or certain metrics had yet to be introduced. This can be combatted by taking a smaller sample from the population with conditions that ensure for a more complete representation. Some data is missing not because of recording errors but because of rule changes such as the addition of a three-point shot. Another minor issue is that this is not a fully updated dataset as it does not include data from the 2018 season that ended in June however, missing one year will not be a major detraction from the validity of the dataset. Overall, the dataset is comprehensive and will provide a good foundation for further analysis.

Numerical Representation

Five-Number Summary

```
library(data.table) # (*5)
byYear = setDT(nba.raw[,c(1,6,8:10,12:13,15:16,18:19,21:29)])[,lapply(.SD, sum),
                                                              by=Year]
summary(byYear[,c(3:20)])
```

```
##        MP                FG              FGA               X3P
##  Min.   :143290   Min.   : 19490   Min.   : 51522   Min.   : 1035
##  1st Qu.:267720   1st Qu.: 37514   1st Qu.:101608   1st Qu.: 4771
##  Median :484591   Median : 84997   Median :178448   Median :12448
##  Mean   :442428   Mean   : 70731   Mean   :155944   Mean   :11065
##  3rd Qu.:620318   3rd Qu.: 94076   3rd Qu.:205998   3rd Qu.:15820
##  Max.   :686746   Max.   :104956   Max.   :234206   Max.   :26140
##  NA's   :3        NA's   :1        NA's   :1        NA's   :31
##       X3PA             X2P              X2PA               FT
##  Min.   : 4161   Min.   :19490   Min.   : 51522   Min.   :15741
##  1st Qu.:14712   1st Qu.:37514   1st Qu.: 96274   1st Qu.:26718
##  Median :35082   Median :78274   Median :162927   Median :43050
##  Mean   :31680   Mean   :64548   Mean   :138241   Mean   :37077
##  3rd Qu.:44204   3rd Qu.:82597   3rd Qu.:172480   3rd Qu.:47343
##  Max.   :73136   Max.   :92455   Max.   :190093   Max.   :53015
```

```
##  NA's   :31      NA's   :1       NA's   :1        NA's   :1
##       FTA             ORB             DRB             TRB
##  Min.   :21365   Min.   :19317   Min.   :44474   Min.   : 31429
##  1st Qu.:35079   1st Qu.:28699   1st Qu.:59958   1st Qu.: 55653
##  Median :57206   Median :29997   Median :68595   Median : 88438
##  Mean   :49529   Mean   :29394   Mean   :69575   Mean   : 81513
##  3rd Qu.:62860   3rd Qu.:31322   3rd Qu.:79378   3rd Qu.:106054
##  Max.   :70185   Max.   :34164   Max.   :89757   Max.   :119597
##  NA's   :1       NA's   :25      NA's   :25      NA's   :2
##       AST             STL             BLK             TOV
##  Min.   :11310   Min.   :12617   Min.   : 6663   Min.   :22463
##  1st Qu.:22744   1st Qu.:17869   1st Qu.:10488   1st Qu.:35330
##  Median :50525   Median :19664   Median :12098   Median :36766
##  Mean   :41590   Mean   :18858   Mean   :11566   Mean   :36314
##  3rd Qu.:56649   3rd Qu.:20453   3rd Qu.:12859   3rd Qu.:37892
##  Max.   :62470   Max.   :22080   Max.   :13608   Max.   :40542
##  NA's   :1       NA's   :25      NA's   :25      NA's   :29
##        PF              PTS
##  Min.   :14799   Min.   : 55252
##  1st Qu.:26808   1st Qu.:102126
##  Median :49706   Median :215544
##  Mean   :42128   Mean   :184722
##  3rd Qu.:54519   3rd Qu.:249888
##  Max.   :62337   Max.   :282466
##  NA's   :1       NA's   :1
```

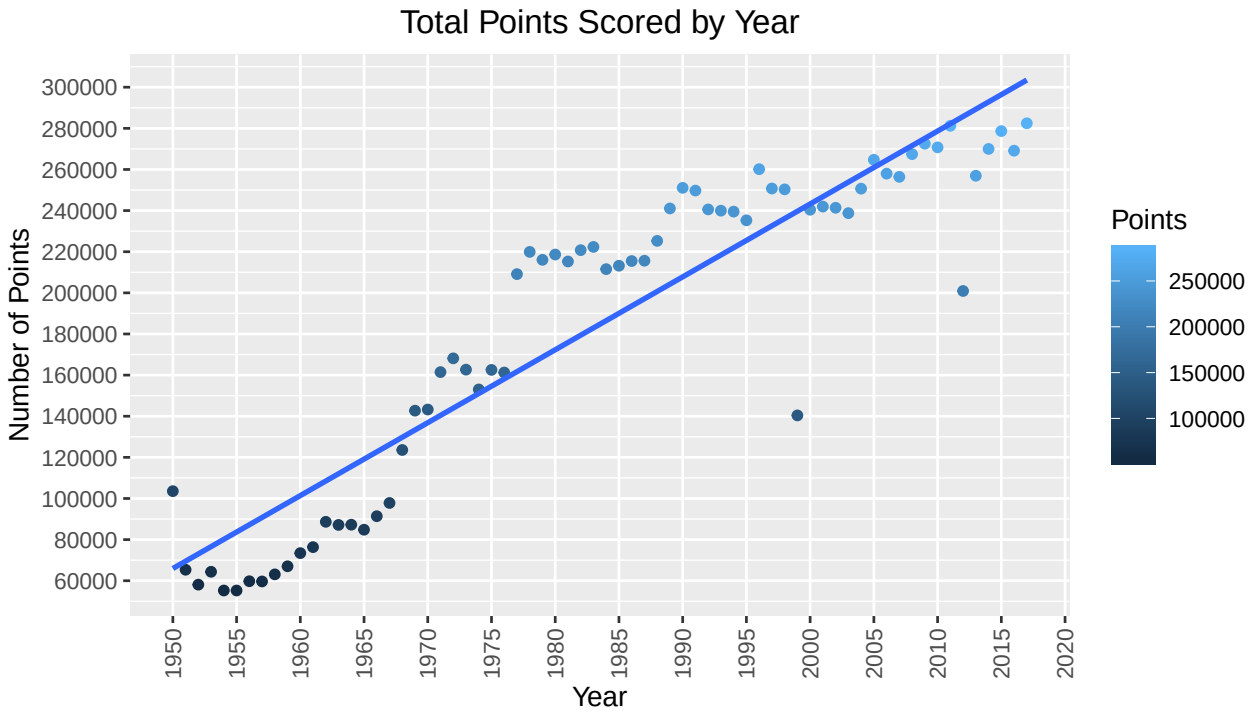
Here, I present the data using the five number summary considering the totals for each year. This provides a more accurate representation of how the NBA has changed as the ranges of each variable (that are not percentages) are shown. In order to this, I first summed up every metric by year and then grouped them together using the data table function. This summary also details provides a clear picture of the mean versus the median for comparison which will be beneficial for further exploratory analysis.

Graphical Representations

Total Points Graph

```
install.packages("ggplot2", repos = "http://cran.us.r-project.org")
library(ggplot2)
year.list = split(nba.raw, nba.raw$Year)

totpoints = sapply(year.list, function(x) sum(x$PTS))
threepointattempts = sapply(year.list, function(x) sum(x$X3PA))
dfnba = data.frame(Year=as.numeric(names(totpoints)),Points=totpoints,
                   ThreePointAttempts=threepointattempts)
ggplot(dfnba, aes(x=Year,y=Points)) + geom_point(aes(colour=Points)) +
  theme(axis.text.x = element_text(angle = 90, hjust=0.95,vjust=0.5)) +
  geom_smooth(method="lm", se=F) + scale_x_continuous(breaks = seq(1950, 2020, 5)) +
  scale_y_continuous(breaks = seq(0, 300000, 20000)) +
  labs(title="Total Points Scored by Year", y="Number of Points") +
  theme(axis.text.x = element_text(angle = 90, hjust=0.95,vjust=0.5),
        plot.title = element_text(hjust=0.5), panel.grid.minor.x = element_blank())
```

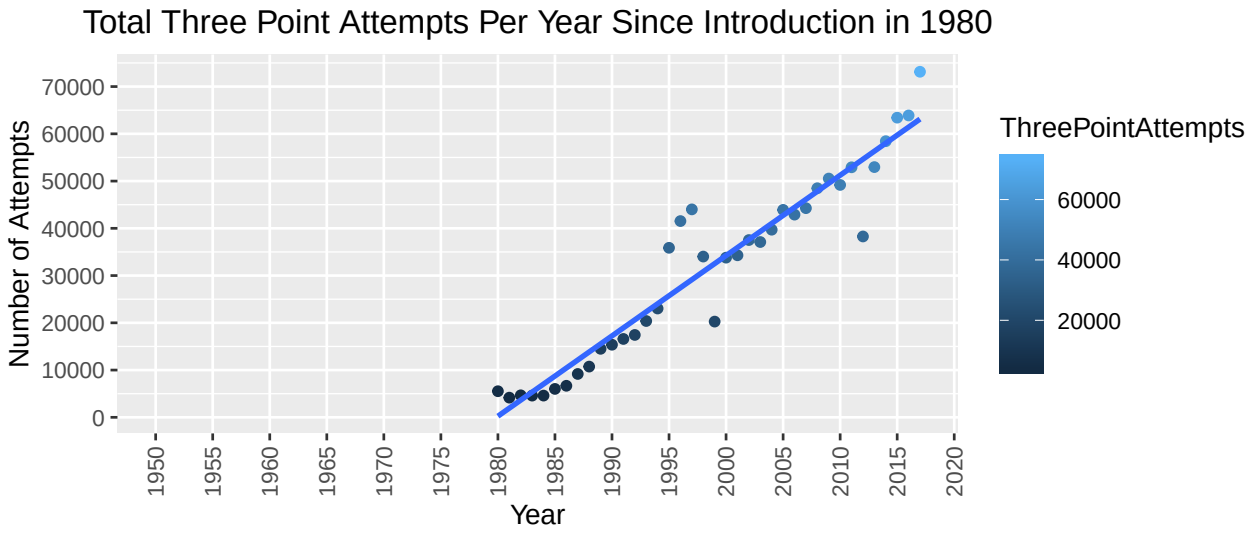


This graphical representation implicitly demonstrates how the NBA has grown. As the initial graphical representation of the data, this serves as a plot showing the total number of points

that have been scored by all players throughout the years. For further exploration, certain conditions would need to be placed on the data being passed in such as how many minutes were played or the number of games during that season. That would provide a more accurate model of how the pace of the game has changed. Looking at the graph, there are certain outliers which are caused either by lockouts, player strikes, and other extraneous situations (*6). Additionally, schedules were made longer as more teams joined the league leading to more opportunities to score for an additional number of players.

Three-Point Attempts Graph

```
ggplot(dfnba, aes(x=Year,y=ThreePointAttempts)) +
  geom_point(aes(colour=ThreePointAttempts)) +
  theme(axis.text.x = element_text(angle = 90, hjust=0.95,vjust=0.5)) +
  geom_smooth(method="lm", se=F) + scale_x_continuous(breaks = seq(1950, 2020, 5)) +
  scale_y_continuous(breaks = seq(0, 80000, 10000)) +
  labs(title="Total Three Point Attempts Per Year Since Introduction in 1980",
       y="Number of Attempts") +
  theme(axis.text.x = element_text(angle = 90,hjust=0.95,vjust=0.5),
        plot.title = element_text(hjust=0.5), panel.grid.minor.x = element_blank())
```



Since the three-point shots introduction into the NBA in the 1980 season, it has quickly grown in popularity among players. Today, some players have made the shot their specialty such as 2x league wide most-valuable player Steph Curry. From initial observations, it can be seen that the positive linear trend of the number three-point attempts follows the increase of points scored the graphical representation shown first.

Initial Conclusions and Further Exploration

The numerical summaries and graphical representations provided in this portion of the investigation lays the foundation for deeper dives into the statistical world of basketball. General trends across most of the metrics provided in the raw dataset have seen large increases and the positive linear trend from my graphical representation leads me to make the initial conclusion that the game has become faster lending to more score opportunities, the size of the league has grown, and the introduction of the three-point shot has generated additional points per play. Additionally, the structure and how the game is played has almost certainly changed. The large ranges that exist in each of the five-number summaries are key indicators that rather than the players becoming better, stronger, and faster; there are also more opportunities for them to accumulate these statistical metrics. However, also using the numerical summary, these initial conclusion will be challenged as I explore further since the large range of total minutes played by all players does create some uncertainty as to whether the increase in points was simply caused by more players playing more games. For the future, I will focus on understanding whether the game has become less physical as many broadcasters have stated (*4), whether there is actual causality between the three-point shot and total points scored, and shifts between offensive/defensive mindsets throughout the years.

References - Citations In-Text Marked by (*Reference Number)

1. Dataset retrieved from Kaggle and was mined from Basketball-Reference (*1):
   https://www.kaggle.com/drgilermo/nba-players-stats/home
   https://www.basketball-reference.com/
2. Information about basketball specific metrics (*2):
   https://www.basketball-reference.com/about/glossary.html
3. Information about NBA rule changes throughout history (*3):
   http://www.nba.com/analysis/rules_history.html
4. Article from Jabari Davis concerning physicality of the NBA (*4):
   http://www.basketballinsiders.com/is-the-nba-becoming-too-soft/
5. Referenced for help with data tables (*5):
   https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html
6. Article from CNN archives about lockouts and strikes in sports (*6):
   https://www.cnn.com/2013/09/03/us/pro-sports-lockouts-and-strikes-fast-facts/index.html