ST595_TechnicalReport2

Michael Jones

4/30/2021

Introduction

I began watching the NBA in the early 1990's. Recently, I have become interested in advanced NBA statistics. These statistics are supposed to more accurately reflect player performance in different aspects of the game than more traditional statistics. The purpose of this report is to do an exploratory analysis of both traditional and advanced statistics to see how they have changed over time. In addition, I wanted to explore how these statistics differed by position.

The dataset I used contains statistics for every NBA player from 1950 to 2017. However, since some statistics that I am interested in are missing prior to 1960, I just considered the players and statistics starting from 1960. There are 20,287 observations in the data set and 53 variables. Since this set of data contains every member of the population (every NBA player), it should be considered a census rather than a sample. I started out by exploring how the average age of NBA players has changed over time.

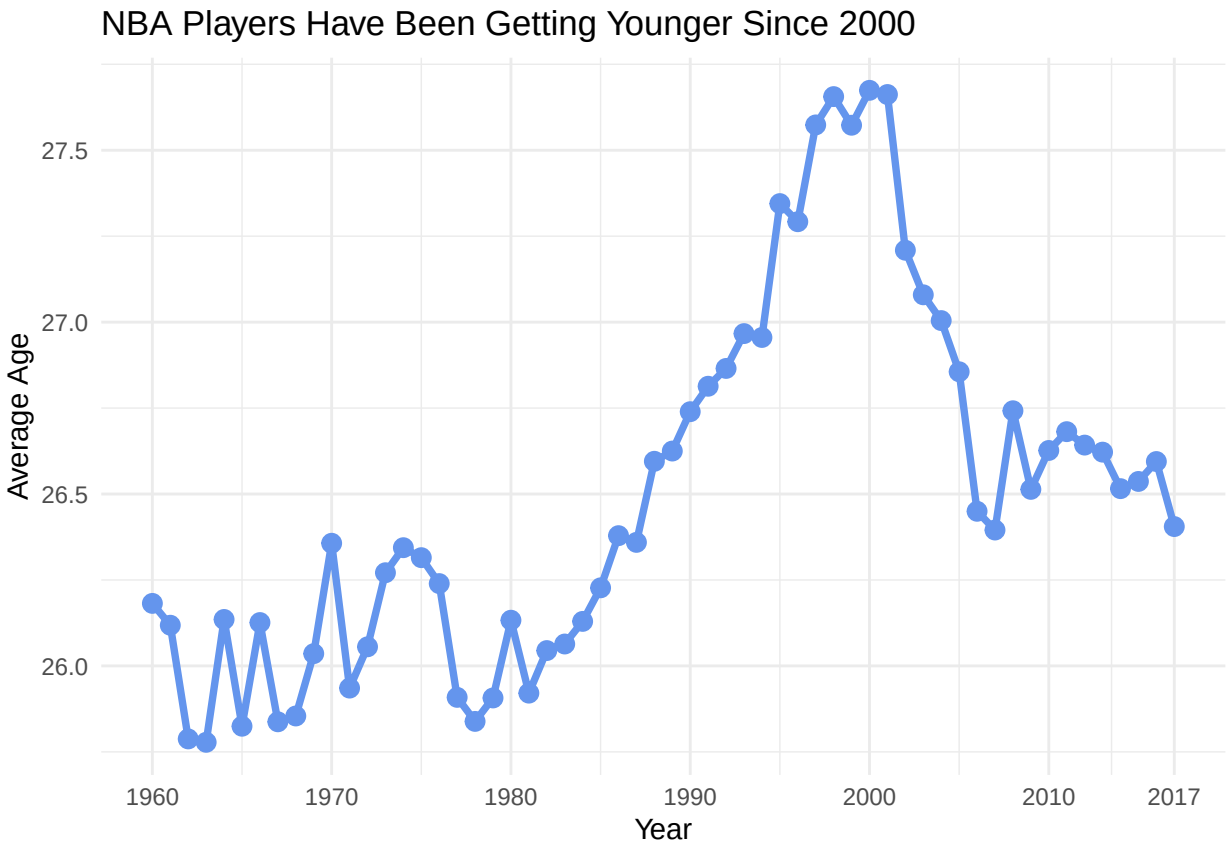
Exploratory Analysis: Age

I started out by getting the summary statistics for the *age* variable. The average age of an NBA player since 1960 is 26.65 years old. The median is slighly less than the mean (26.0 years old) suggesting that the ages are somewhat right-skewed. There is quite a bit of variance in *ages* as the youngest player is 18 years old, while the oldest player is 44 years old.

Table 1: Age

Min	18.00
1st Quartile	24.00
Median	26.00
Mean	26.65
3rd Quartile	29.00
Max	44.00

But what I am really intersted in is how the average age of NBA players has changed over time. Therefore, I grouped the players by year, got the average age by year, and created a line graph to show the trend.



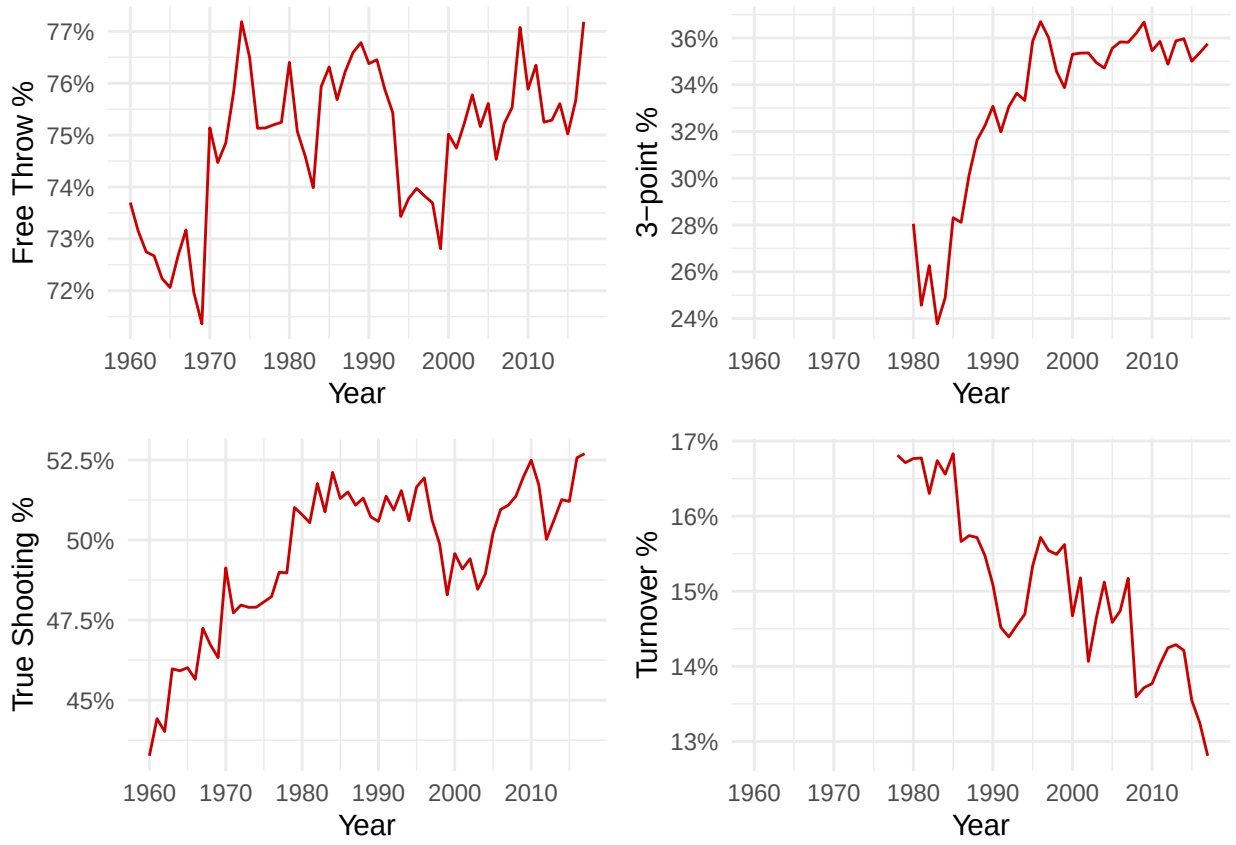
This plot shows some interesting trends. First, we can see that the average age of NBA players hovered around 26 years old from 1960 until the early 1980's. Then, the average increased on an almost yearly basis from the early 1980's until about the year 2000, culminating in an average age of nearly 28 in 2000. The average age declined to about 26.5 years old by the mid-2000's where it stayed until 2017.

It would be interesting to know the reasons behind these trends. My initial guess was that the average age of NBA players had probably been increasing over the past 10-20 years due to improvements in physical training and health knowledge (especially about nutrition). But this is not the case. In fact, there was a steady decline in the average age of NBA players from 2000 until about 2005, and no noticeable increase since then. As far as potential statistical analysis is concerned, one interesting idea would be to use a time series prediction model (such as exponential smoothing) to forecast the average age of NBA players from 2017 until 2025 or 2030.

Exploratory Analysis: Efficiency

Next, I examined how four efficiency metrics (free throw %, 3-pt. %, true shooting %, and turnover %) have changed over time. The summary statistics didn't reveal anything that interesting this time. For one thing, the minimum values and maximum values can be discounted as they are always 0% and 100% respectively due to outlier players who did not shoot the ball or play much. Skewness was evident for some of the variables, espcieally 3-pt%. In this metric, the mean is just 24.8%, but the median is 29.2%, indicating a strong leftward skew.

The trends I explored here was how each of these statistics has changed over time. The 3-pt. shot and turnover statistics have just been kept since 1980, but the other statistics go back to 1960. Here are the plots of these four variables.

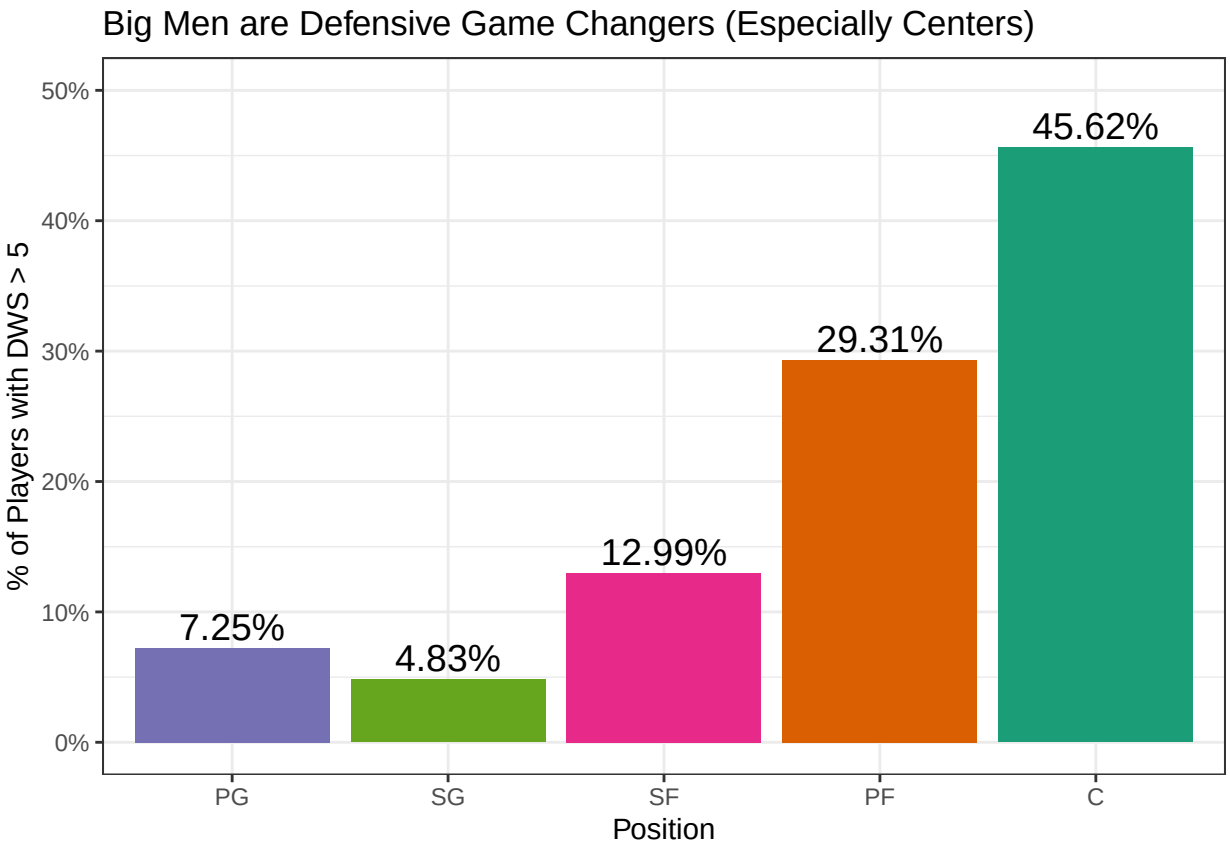


These plots reveal some unexpected results. A common assumption among NBA fans, and even some NBA analysts, is that players nowadays are less fundamentally sound, meaning they don't shoot and take care of the ball as well as players in the past. The statistics, however, show otherwise. All of the shooting metrics (free throw %, 3-pt. %, true shooting %) show an upward trend, while turnover percentage shows a downward trend. An objective look at the data, then, seems to debunk this long-held assumption.

These trends seem pretty clear, but in order to be sure they are statistically significant, formal statistical tests could be done to verify the trends. The easiest way to do this would simply be to do a proportions test. This would give a p-value that would help in determining whether these percentages are truly different in later years compared to earlier years. I went ahead and performed one proportions test myself. I compared the free throw percentage in 2017 to 1960, by looking at the total number of free throws made versus attempted in each year. This proportions test gave convincing evidence that the difference in free throw percentage is statistically significant (p-value < 0.001), meaning that NBA players were better free throw shooters in 2017 than in 1960. Similar tests could be done for other years and other metrics to verify these trends.

Exploratory Analysis: Defensive Metrics

The final bit of exploratory analysis involved defensive metrics. I analyzed the statistic Defensive Win Shares (DWS), which attempts to quantify the number of wins a player's defensive ability gave his team per year. The summary statistics shows that an NBA player's mean DWS since 1960 is 1.389, and the median DWS is 1.0. This indicates leftward skewed data, which means there are an inordinate number of players with fairly low DWS's. Rather than analyzing how this statistic has changed over time, I examined DWS values by position. I chose a fairly high value of 5 and looked into the percentages of players with DWS over 5 at each position. The goal was to try to identify which positions were more likely to have players who were defensive game changers, as defined by a high DWS of above 5.



As the title to this bar graph suggests, the vast majority of defenseive game changers are power forwards or centers. These two positions account for nearly 75% of all players with a DWS greater than 5. Guards are much less likely to be elite defenders as they account for only about 12% of all playes with a DWS greater than 5. There are not that many observations in this data set, since there are few players with a DWS above 5. So to know if there is enough evidence to conclude that there is a difference among these positions, we can do a proportions test.

I decided to do a proportions test comparing the guard (point guard and shooting guard) positions to each other and the big men (power forwards and centers) positions to each other. That is, I wanted to know if there was a statistically significant difference between the number of point guards who had a DWS above 5 and the number of shooting guards that did. Likewise, for power forwards and centers. The proportions test showed that there was no evidence of a difference among the positions of point guard and shooting guard (p-value = 0.2535). In other words, we have no reason to claim that point guards are more likely to have a DWS of greater than 5 than shooting guards. However, the proportions test gave convincing evidence that centers are more likely to have a DWS above 5 than power forwards (p-value $< .0001$).

Conclusion

This analysis explored NBA statistics, most notably advanced statistics. I started out, however, by looking at how the average age of an NBA player has changed over time. Interesting trends emerged from this analysis that showed the mean age of NBA player's increasing during the 1980's and 1990's, decreasing during the early 2000's, then remaining steady since the mid-2000's. After this, I explored efficiency metrics, which gave surprising results. Contrary to popular opinion, NBA players are better shooters and turn the ball over less frequently than in the past. Finally, I looked at the defensive statistic DWS. A proportions test gave convincing evidence that the center position is more likely to contain players with a DWS above 5 than the power forward position. But evidence was lacking that there is a difference between the guard positions.

This exploratory analysis has yielded several ideas for further research. While I did a few proportions tests, many more could be done. For instance, a proportions test was done which gave convincing evidence that players in 2017 were better free throw shooters than players in 1960. Other proportions tests could be done for other years or for other efficiency statistics. Proportions tests could even be done for individual players. Perhaps the most promising area of further investigation, though, is time series analysis. Time series analysis could be used to try to predict the mean player age in the NBA in the future, or to predict efficiency metrics over the coming years. In fact, there appears to be many potential areas of further analysis with this set of data.

Technical Appendix (R Code)

```
knitr::opts_chunk$set(echo = FALSE)
library(ggplot2)
library(tidyverse)
library(gridExtra)
library(knitr) # for five number summary table

nba <- read.csv("Seasons_Stats.csv")
nba <- nba %>% distinct(Player, Year, .keep_all = TRUE) # gets rid of duplicates
nba_1960 <- nba %>% filter(Year >= 1960)
nba_age <- nba_1960 %>% group_by(Year) %>%
  summarise(Avg_Age = mean(Age))
summary(nba_1960$Age)

age_table <- data.frame(Quartile = c("Min", "1st Quartile", "Median", "Mean", "3rd Quartile",
                                     "Max"), Age = c(18.00, 24.00, 26.00, 26.65, 29.00,
                                                     44.00))
age_table2 <- kable(age_table, col.names = c("", ""), caption = "Age")
age_table2

# Creating a line graph
ggplot(data = nba_age, aes(x = Year, y = Avg_Age)) +
  geom_point(size = 3, color  = "cornflowerblue", fill = "black") +
  geom_line(color = "cornflowerblue", size = 1.3) +
  labs(title = "NBA Players Have Been Getting Younger Since 2000",
       y = "Average Age") +
  scale_x_continuous(breaks = c(1960, 1970, 1980, 1990, 2000, 2010, 2017)) +
  theme_minimal()

# Getting summary statistics for efficiency metrics
summary(nba_1960$FT.)
summary(nba_1960$FG.)
summary(nba_1960$X3P.)
summary(nba_1960$TOV.)

# Grouping statistics by year
nba_ft <- nba_1960 %>% group_by(Year) %>%
  summarise(FT._YR_AVG = sum(FT)/sum(FTA))

nba_3p <- nba_1960 %>% group_by(Year) %>%
  summarise(X3P._YR_AVG = sum(X3P)/sum(X3PA))
```

```r
nba_ts <- nba_1960 %>% group_by(Year) %>%
  summarise(AVG_TS = mean(TS., na.rm = TRUE))

nba_tov_perc <- nba_1960 %>% group_by(Year) %>%
  summarise(AVG_TOV_PERC = mean(TOV., na.rm = TRUE))

# Creating side by side line graphs
plot_ft <- ggplot(data = nba_ft, aes(x = Year, y = FT._YR_AVG * 100)) +
  geom_line(color = "red3",na.rm = TRUE) +
  labs(y = "Free Throw %") +
  scale_x_continuous(breaks = c(1960, 1970, 1980, 1990, 2000, 2010, 2017)) +
  scale_y_continuous(breaks = c(72, 73, 74, 75, 76, 77),
                     labels = c ("72%", "73%", "74%", "75%", "76%", "77%")) +
  theme_minimal()

plot_3p <- ggplot(data = nba_3p, aes(x = Year, y = X3P._YR_AVG)) +
  geom_line(color = "red3", na.rm = TRUE) +
  labs(y = "3-point %") +
  scale_x_continuous(breaks = c(1960, 1970, 1980, 1990, 2000, 2010, 2017)) +
  scale_y_continuous(breaks = c(0.24, 0.26, 0.28, 0.30, 0.32, 0.34, 0.36),
                     labels = c ("24%", "26%", "28%", "30%", "32%", "34%", "36%")) +
  theme_minimal()

plot_ts <- ggplot(data = nba_ts, aes(x = Year, y = AVG_TS)) +
  geom_line(color = "red3", na.rm = TRUE) +
  labs(y = "True Shooting %") +
  scale_x_continuous(breaks = c(1960, 1970, 1980, 1990, 2000, 2010, 2017)) +
  scale_y_continuous(breaks = c(0.450, 0.475, 0.500, 0.525),
                     labels = c ("45%", "47.5%", "50%", "52.5%")) +
  theme_minimal()

plot_tov_perc <- ggplot(data = nba_tov_perc, aes(x = Year, y = AVG_TOV_PERC)) +
  geom_line(color = "red3", na.rm = TRUE) +
  labs(y = "Turnover %") +
  scale_x_continuous(breaks = c(1960, 1970, 1980, 1990, 2000, 2010, 2017)) +
  scale_y_continuous(breaks = c(13, 14, 15, 16, 17),
                     labels = c ("13%", "14%", "15%", "16%", "17%")) +
  theme_minimal()

grid.arrange(plot_ft, plot_3p, plot_ts, plot_tov_perc)

# Setting minimum games for this stat
nba_1960_mod <- nba_1960 %>% filter (G >= 50) %>%
  filter(Pos %in% c("PG", "SG", "SF", "PF", "C"))
# This is needed to remove outliers due to low games played and simplify positions

nba_1960_dws5 <- nba_1960_mod %>% filter(DWS > 5)
dws_counts <- xtabs(~ Pos, nba_1960_dws5) # Get table of counts
dws_prop <- prop.table(dws_counts) %>% round(4) * 100 # Turn to prop then perc
dws_df <- data.frame(dws_prop)
dws_df <- dws_df %>% filter(Pos %in% c("PG", "SG", "SF", "PF", "C"))

# Creating a Bar Graph
```

```r
Percentages <- c("45.62%", "29.31%", "7.25%", "12.99%", "4.83%")
ggplot(data = dws_df, aes(x = Pos, y = Freq)) +
  geom_bar(stat = "identity", aes(fill = Pos), show.legend = FALSE) +
  geom_text(aes(label = Percentages), vjust = -0.3, size = 5) +
  labs(title = "Big Men are Defensive Game Changers (Especially Centers)",
       y = "% of Players with DWS > 5",
       x = "Position") +
  scale_y_continuous(limits = c(0,50),
                     breaks = c(0, 10, 20, 30, 40, 50),
                     labels = c ("0%", "10%", "20%", "30%", "40%", "50%")) +
  scale_x_discrete(limits = c("PG", "SG", "SF", "PF", "C")) +
  scale_fill_brewer(type = "qual", palette = "Dark2") +
  theme_bw()

# Proportions test for PG/SG
pg_suc = sum(nba_1960_dws5$Pos == "PG")
sg_suc = sum(nba_1960_dws5$Pos == "SG")
total = 331

prop.test(x = c(pg_suc, sg_suc), n = c(total, total))

# Proportions test for PG/SG
pg_suc = sum(nba_1960_dws5$Pos == "PG")
sg_suc = sum(nba_1960_dws5$Pos == "SG")
total = 331

prop.test(x = c(pg_suc, sg_suc), n = c(total, total))
```