# *Statistics of Athletes in The National Basketball Association Between 2012-2018*

*Team L.L.M*
*Denny Mathew, Austin Luong, Anthony Lam*

November 7, 2019

## 1 Introduction

The purpose of this project was to find a dataset and identify interesting sightings based off any questions that we may have as a group. The dataset that we found and explored was from the data collected from the NBA Official API. There was two excel files from the dataset. The first dataset had the matches and statistics from all games as well as individual player statistics for each game back from 2012-2018 and the second dataset was an average player statistic of all seasons between 2012-2018 based off from the first dataset. We decided as a group to use the second dataset for our project because we wanted to gather most of our information based on the average amounts of points and statistics of each individual athlete to compare with one another. The attributes include the player name (*PLAYER_NAME*), average minutes played (*MIN*), average points scored (*PTS*), field-goals made (*FGM*), field-goals attempted (*FGA*), 3-point field-goals made (*FGM3*), etc.

Every player has a range of statistics. In this dataset, we were interested to see what makes a player great in terms of points because the rules follow that the team who scores the most points win the game. Using the data, we can possibly find correlations and common characteristics among "great" players. On the other hand, part of our motivation and curiosity to pick an NBA dataset was that we wanted to know more about players who do not necessarily have the top tier of points, but we recognize and pay attention to other value they provide such as assists, rebounds, blocks, etc. The game of basketball is followed by a set of rules. The goal is to have one team score more points than the other team within a four 15-minute quarters with breaks and timeouts in between. For a player to move from one side of the court with the basketball, they must dribble their way to their desired location to shoot the ball. Every shot made outside the large 3-point arc is worth 3 points while the rest are worth 2 points, except for free throws which is for players that are fouled. Of course, there are restrictions to the players but the important background information to understand is the statistics each player has and its impact on the team. Figure 1 below shows the layout of a basketball court.

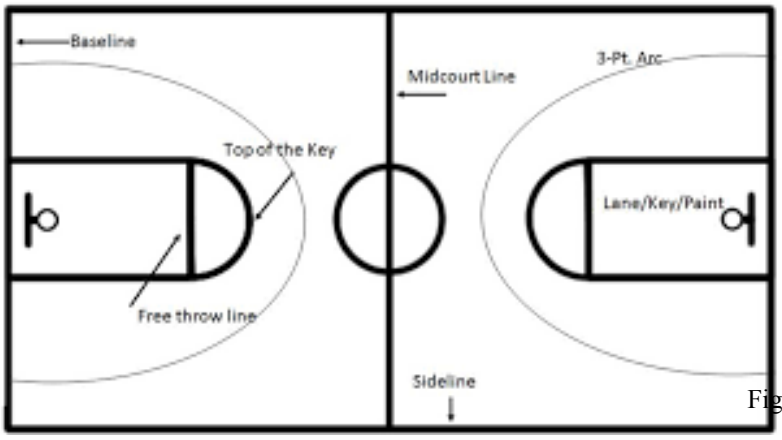


Figure 1: Layout of a Basketball Court

As a group, we wanted to look at if a player that does not make a significant amount of points makes them an inefficient player to the team. Following this, we would look at any statistics that these top players would have in common and see if there could be another correlation instead of points that makes a great player.

## 2 Analysis of the Data

In this section, we will be creating graphs based on the information given to us in the dataset and trying our best to analyze any findings. Taking into thought about our first question, we must decide the cutoff for what we would consider the amount of points a above average player will score. We understand that this is a completely subjective argument but for our project, we have chosen that the cutoff for a player to be above average is that the player must score an average of at least 20 points. The first thing that we can create is a graph that shows each individual player and their average amount of points. However, doing this will cause the graph to look very disorganized and it doesn't really say anything that will help us answer our questions. To organize our graph better, we separated each player based on which conference they were from. In the NBA, there are two conferences, Eastern and Western, that has 15 teams on each side. Teams try to battle it out and become the top team of their conference to play against the top team of the other conference. Figure 2 below shows us the Points vs Conference of the dataset.

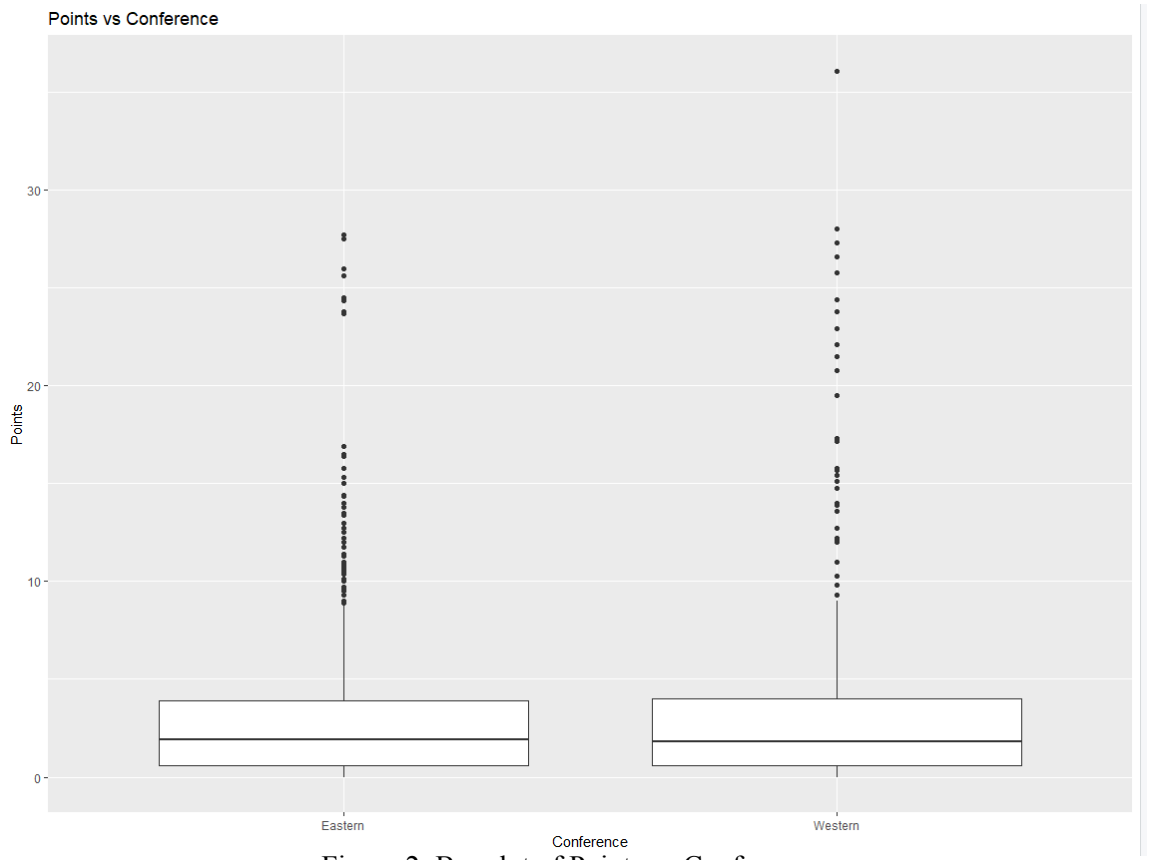

Figure 2: Boxplot of Points vs Conference

As we can see, the graph of the data shows a better representation of the top players that have scored on average at least 20 points based on their conferences. However, this graph doesn't give anymore information other than that. If these players are averaging high amounts of points, we can see the types of shots that they are shooting. We can use only the top players we see above because they are the ones we care about in our project.

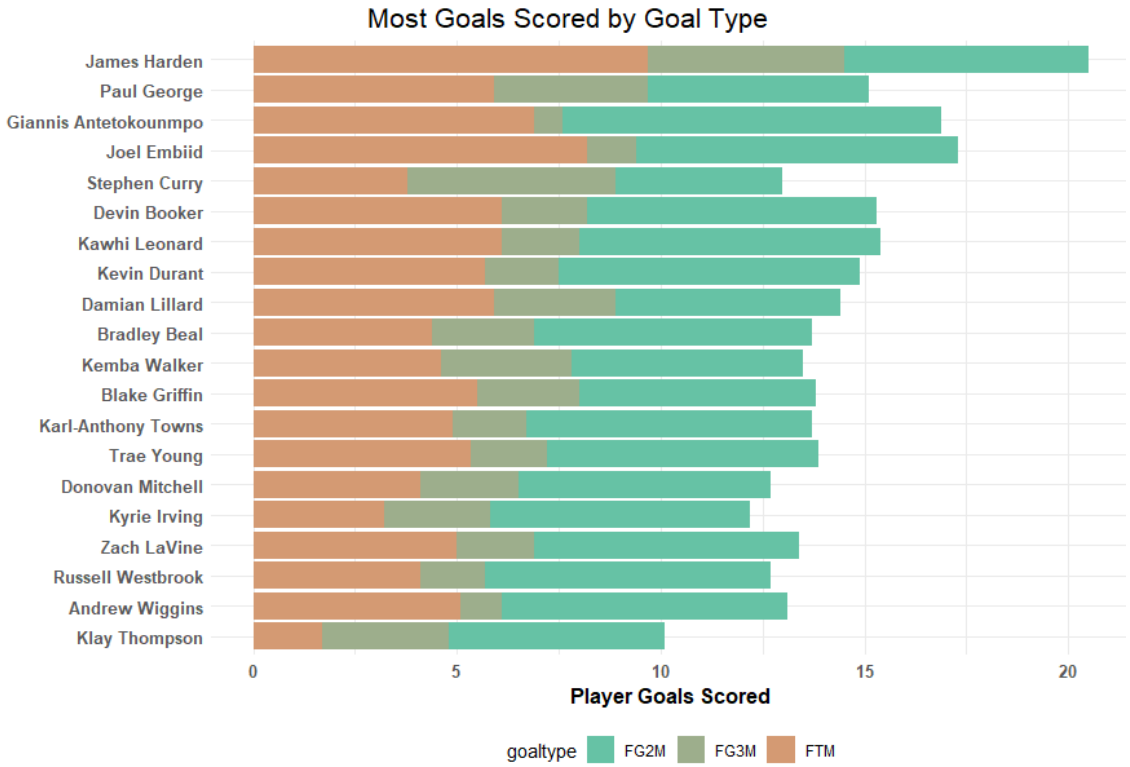


Figure 3: Stacked Plot of Different Field Goals Made

This graph alone gives us a more specific detail of the different types of shots that these players are shooting. We see that James Harden can score much more compared to any other player in the graph. Yet, we see that the amount of 3-point shots made (*FG3M*) is approximately the same as Stephen Curry and the number of free-throws made (*FTM*) and 2-point shots made (*FG2M*) are roughly the same as Giannis Antetokounmpo and Joel Embiid. So how is it possible that James Harden can score more points in total but have close to the same amount or even less than scored when comparing each field-goal separately? This not only goes for James Harden but for every top player we see here in the graph.

While looking at these graphs, we became curious as to how these players that average over 20 points a game are able to score that much. We thought about it and realize that if a player can play more than others, then that means the players have a higher chance of scoring more. In order to find out, we created a graph that will show us any correlation or not about whether the number of minutes a player plays will determine the amount of points scored.

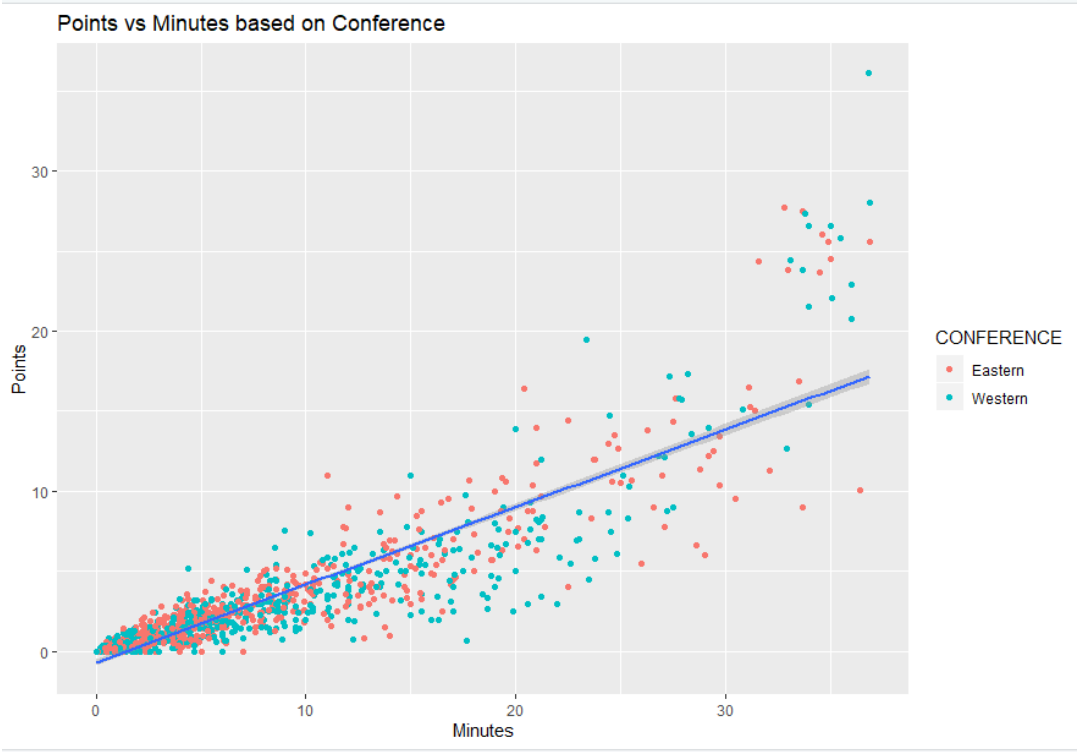

Figure 4: Points vs Minutes based on Conference

The graph tells us that there is a positive correlation between the number of minutes played and the amount of points scored based on the trendline we added. We can see that the top players from the previous figure are being distinguish with the number of minutes they are able to play. All the players that score at least 20 points and above have all at least 30 minutes of gameplay.

Can we say that the more minutes you have the more points you will score? Yes, but also no. We can see that the top scoring players do have many minutes being played, but we can also find that there are a decent number of players that are getting at least 25 minutes of gameplay but are not scoring above 20 or even 15 points. Possibly we see that rather than just scoring points, the more minutes being played correlates to how great of a player they are. We understand that there are many other beneficial factors when it comes to winning such as: blocking the ball when the opponent is trying to score, stealing the ball so the opponent doesn't have the opportunity to score, having less turnovers meaning not losing or giving up the ball before you score, the position that the player plays, etc. All of which are high factors that can determine an outcome of a game.

We can narrow it down to specific factors and identify common statistics that top players have that would potentially make them overall great players. We wanted to get the list of the top players only and compare more of their statistics than just points and minutes. The goal is the find more things that these top players could potentially have in common. To double check our graph, we cross-referenced it with Microsoft Excel tools.

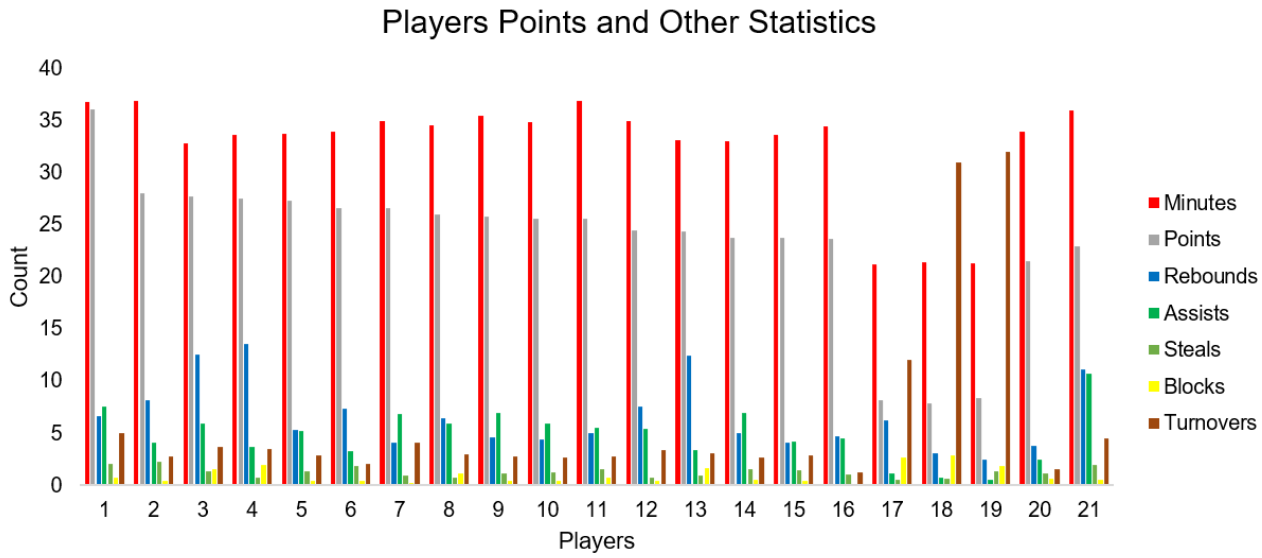


Figure 5: Comparing Other Statistics of Top Players

The players are in order from left to right: 1. James Harden, 2. Paul George, 3. Giannis Antetokounmpo, 4. Joel Embiid, 5. Stephen Curry, 6. Kawhi Leonard, 7. Devin Booker, 8. Kevin Durant, 9. Damian Lillard, 10. Kemba Walker, 11. Bradley Beal, 12. Blake Griffin, 13. Karl-Anthony Towns, 14. Kyrie Irving, 15. Donovan Mitchell, 16. Zach LaVine, 17. DeMar DeRozan, 18. Julius Randle, 19. LaMarcus Aldridge, 20. Klay Thompson, 21. Russell Westbrook. We can see that the graph has a mix of several players that have a common high attribute than others, but it also is vice versa. This has us believing that these players not only score many points, but they all have a specific role when they are playing on the court. For example, we can see James Harden, Devin Booker, and Kyrie Irving have a high count of assists but a low count of rebounds. Assists are when the player passes the ball to his teammate for his teammate to score immediately and rebounds are when the player gets the ball after his teammate or the opposing player doesn't make the shot. But we can also see that Giannis Antetokounmpo, Joel Embiid, and Karl-Anthony Towns have a high count of rebounds but low count of assists. This finding has led us to determine that depending on the position the player plays, they each have a specific role that they must do instead of just scoring many points for the team.

## 3 Conclusion

After performing and analyzing the dataset, we have found many interesting facts. When analyzing the graph of finding players who average at least 20 points based on conferences, we discovered that there were more players on the Western side with at least 20 points than the Eastern side. However, below the cutoff there was a higher density area with more players on the Eastern side than Western. Is it a possibility that because this dataset includes years up to 2018, the new players on the Eastern conference are slowly rising to become top players? We notice when comparing the amount of points and number of minutes has a weak positive correlation due to the high number of minutes that the players averaging below 20 points. We can say that not only does minutes and points matter but other factors like assists, steals, rebounds, etc. This led us to the question if these high scoring players have common statistics as their competitors.

Our conclusion has made us recognize that basketball is not just a sport that requires top players to score, but to have a team that everyone knows their specific role in order to win. Being able to work on this dataset and see what differentiates player to player would be a huge step in the world of basketball. Future work that we would consider is really being able to get specific details like instead of looking at

conference of Western or Eastern, we can look at the specific 15 teams in each conference and determine how well a team can perform at its best. Another future work that I believe would be a very important is the chemistry within a team. If a team can't get along with each other personally, does that influence the team chemistry to work together to win or not.

# 4 References

https://www.kaggle.com/yalcinberkay/nba-matches-dataset-w-player-stats

http://sportstalkmei.blogspot.com/

# A Code for Plots

```
install.packages("ggplot2")
install.packages("tidyr")
install.packages("dplyr")
install.packages("RColorBrewer")
library(ggplot2)
library(tidyr)
library(dplyr)
library("RColorBrewer")

#Created variable for dataset to be included in R
nba_stats <- read.csv(file = "nba_stats.csv", stringsAsFactors = FALSE)


#FIGURE 2 Boxplot for graphing Points vs Conference
ggplot(data = nba_stats) +
  geom_boxplot(mapping =aes(x = CONFERENCE, y = PTS)) +
  labs(x = "Conference", y = "Points", title = "Points vs Conference")


#FIGURE 4 Scatter plot for graphing Points vs Minutes based on Conference
ggplot(data = nba_stats) +
  geom_point(mapping =aes(x = MIN, y = PTS, color = CONFERENCE)) +
  geom_smooth(mapping = aes(x = MIN, y = PTS), method = "lm") +
  labs(x = "Minutes", y = "Points", title = "Points vs Minutes based on Conference")


#FIGURE 3 Stacked Plot of Different Field Goals Made
top20 <- nba %>% arrange(desc(PTS)) %>% head(., 20)
top20
# Each player is in the dataset once
table(nba$PLAYER_NAME)
levels(top20$PLAYER_NAME)
top20$PLAYER_NAME <- factor(top20$PLAYER_NAME, levels =
unique(top20$PLAYER_NAME[order(-top20$PTS)]))
levels(top20$PLAYER_NAME)
# STACKED BAR CHART#
# Each player - (top20 bottom20 for clarity)
# How many shots made?
# Each type of shots?
# Field goals FGM
# Free throws made FTM
# 3 point field goals made (FG3M)
# Number of field goals made total minus number 3 point field goals made
top20$FG2M <- top20$FGM - top20$FG3M
top20$totgoals <- top20$FTM + top20$FG2M + top20$FG3M
```

```
calcpts <- (top20$FTM*1) + (top20$FG2M*2) + (top20$FG3M*3)

# Checking that the points were calculated correctly
head(cbind(top20$PTS, calcpts))
names(top20)
top20 <- as_tibble(top20)
str(top20)
subset <- top20 %>%
  select(PLAYER_NAME, PTS, FTM, FG2M, FG3M) %>%
  gather(key = goaltype, value = numgoals, FTM, FG2M, FG3M) %>%
  arrange(PLAYER_NAME)
subset

#FINAL STACKED BARPLOT
colourCount <- length(unique(top20$PLAYER_NAME))
getPalette <- colorRampPalette(brewer.pal(9, "Set2"))
playerpts <- subset %>%
  group_by(PLAYER_NAME) %>%
  summarize(pts = min(PTS)) %>%
  ungroup()

stackedplot <- ggplot(data=subset, aes(x=PLAYER_NAME, y=numgoals, fill=goaltype)) +
  geom_col() +
  scale_fill_manual(values = getPalette(colourCount)) +
  ggtitle("Most Goals Scored by Goal Type") +
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.3, size = 16, family = "Palatino")) +
  coord_flip()+
  scale_x_discrete(limits = rev(levels(top20$PLAYER_NAME))) +
  labs(y = "Player Goals Scored", x = NULL) +
  theme(axis.text = element_text(family ="GillSansMT", face = "bold",size = 10, color =
"grey35"),
      axis.title=element_text(family ="GillSansMT", size = 12, face = "bold")) +
  theme(legend.position = "bottom")
stackedplot
```

**#FIGURE 5 Comparing Other Statistics of Top Players**

```
#Select the file and organize the observations as false strings
#Run the file to check
nba_stats <- read.csv(file = "nba_stats.csv", stringsAsFactors = FALSE)
str(nba_stats)
#Filter through the data set with the players above 20 points and facet their stats
ggplot(data = nba_stats) + geom_bar(mapping = aes(x = nba_stats[which(nba_stats$PTS >= "20")],
                                    fill = categories), position = "dodge") +
  labs(x = "Players", y = "Count", title = "Players Points and Other Statistics")
```