NBA Scoring Research

Collin Lu, Justin Spencer, Brien Wildermuth

2023-12-13

Exploring NBA Data

Basketball is a complex game that, like all other major sports, involves large scales of data. Some might look at all the data being recorded during an NBA game and ask the question, "Is that really necessary?" One of the numerous purposes of recording all these different statistics during games is so that the people behind the scenes can make more confident decisions when trying to build a championship winning team. A couple of positions who end up making these corporate decisions can be seen as the coaches, general managers, and owners. People tend to forget that professional basketball is a business, just as much as it is a sport.

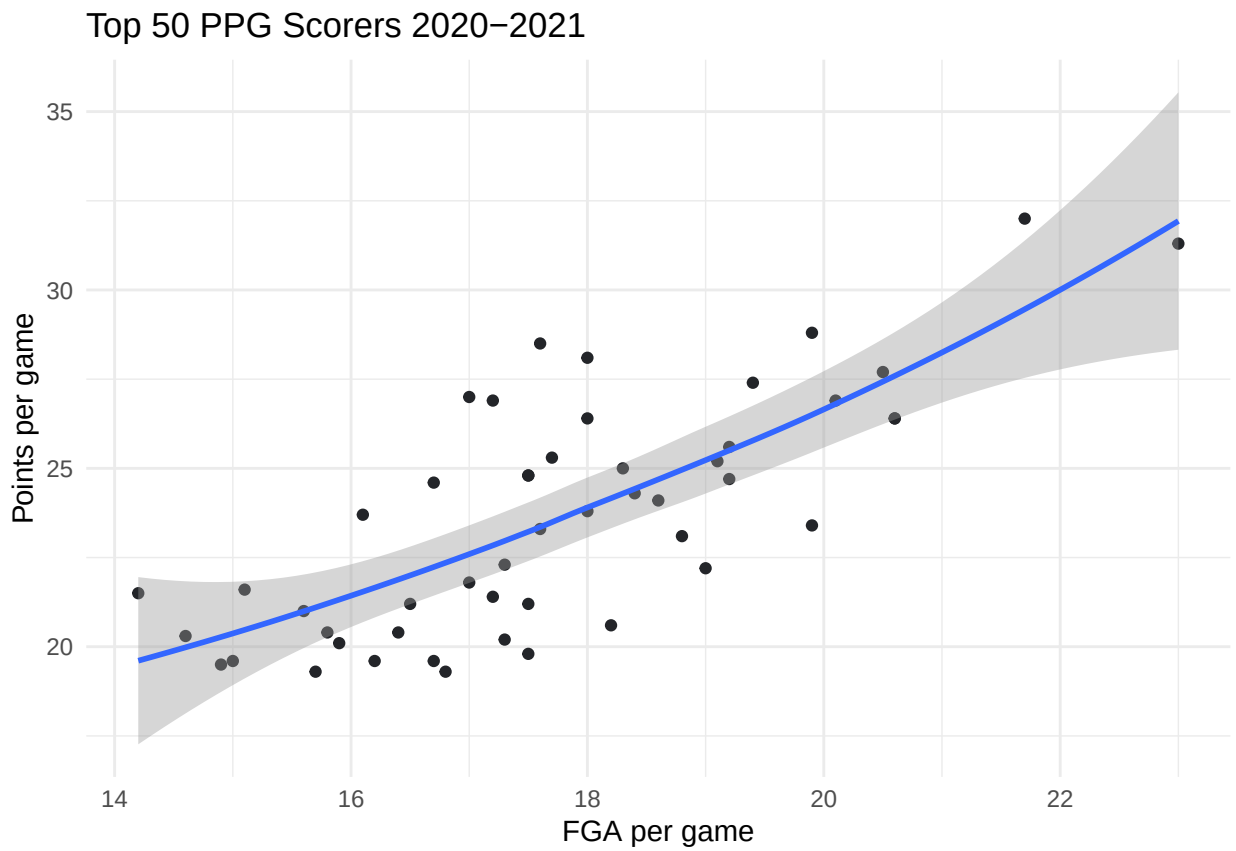
There are plenty of statistics that can allow managers, coaches and fans to determine how "good" or "talented" basketball players are. A few of these statistics include rebounds, assists, win percentage, and even plus/minus. Although some of these averages and totals are often debated more important than others, when determining how valuable a player is, the most commonly referred to statistic is how many individual points a player scores.

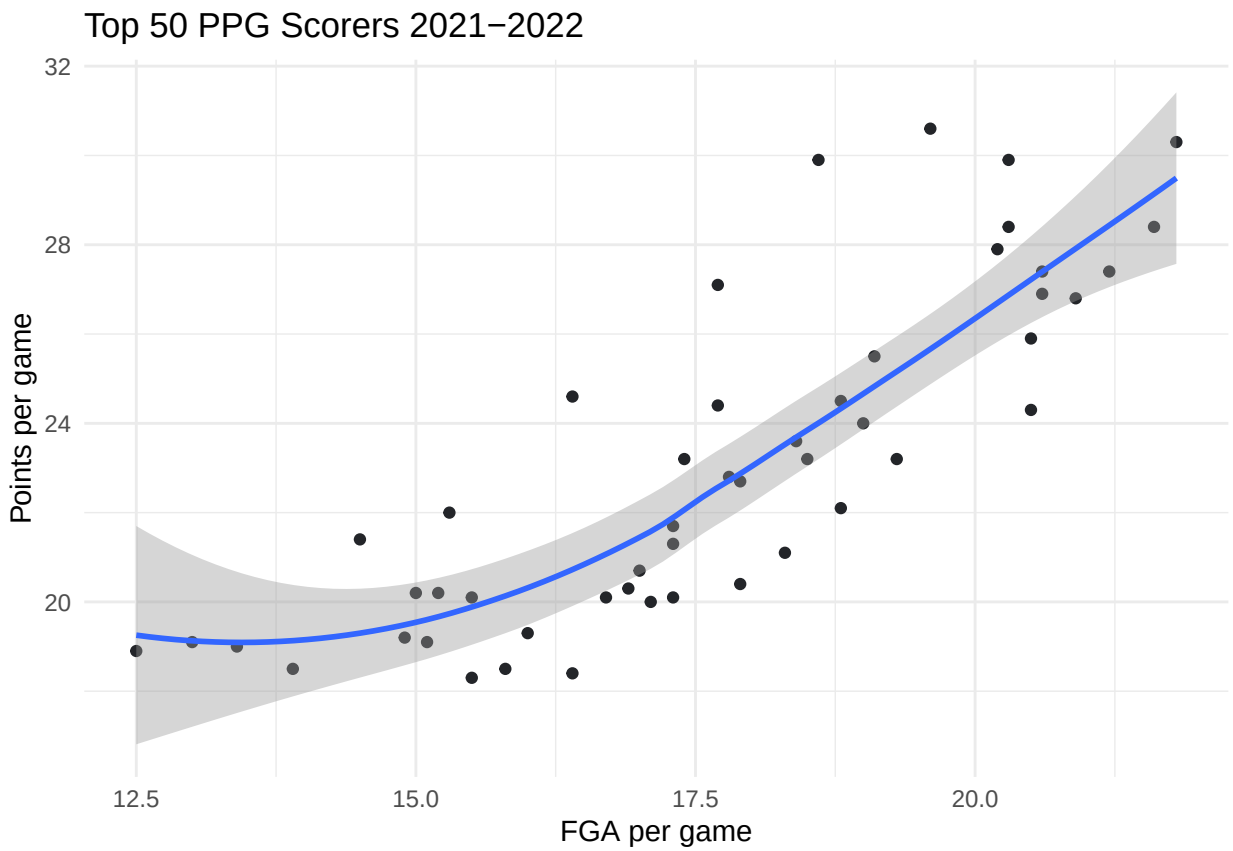
In recent years, many people have stated that a large reason the best point scorers in the National Basketball Association (NBA) are at the top of the league list is because they shoot the most shots. Although that statement could very well be true, we want to explore how strong that correlation really is. In our exploration, we will attempt to analyze the correlation between an individual player's field goal attempts per game and their overall points per game. We decided to research the data sets of the NBA leading scorers for the past three seasons. We used the top fifteen scorers in points per game (PPG) for our data. We wanted to find out if a greater number of field goals attempted per game (FGA) would lead to more points scored per game.

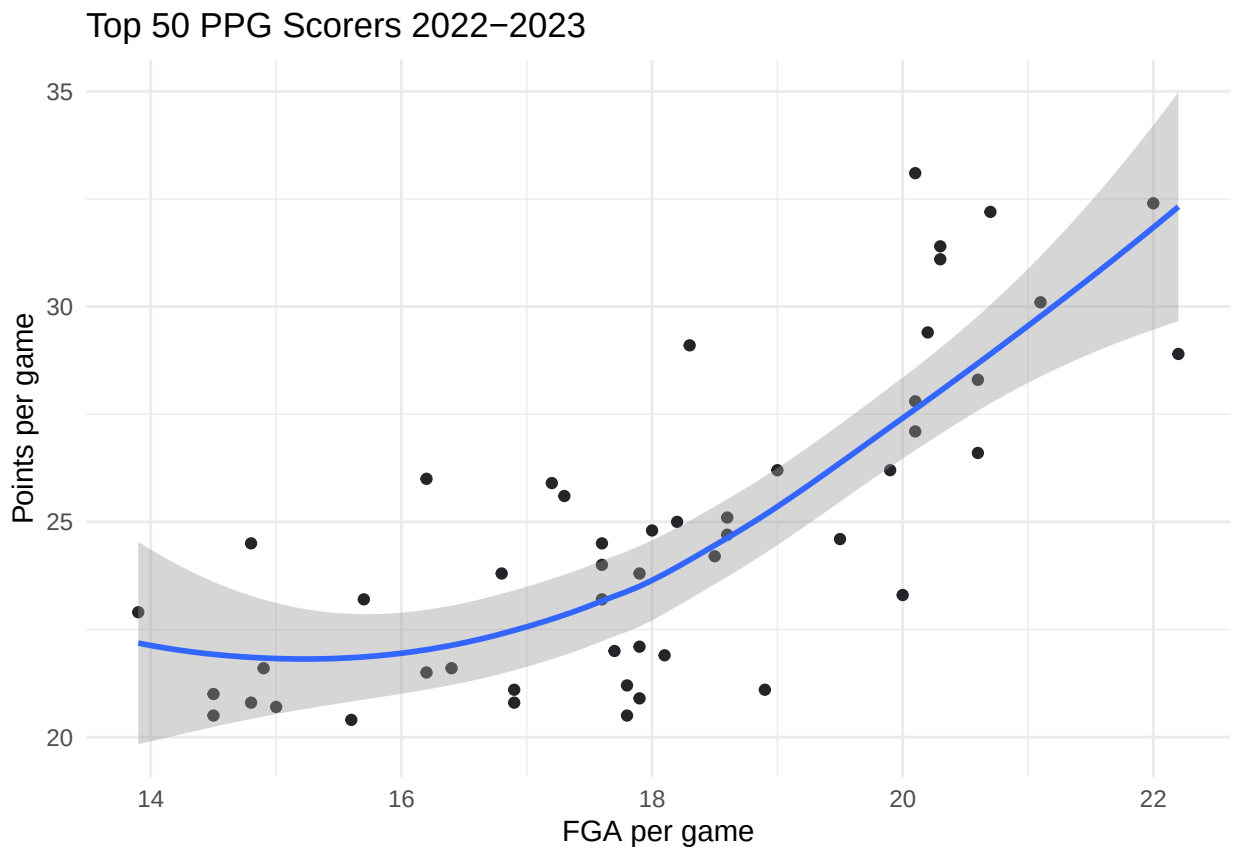
In addition to exploring the amount of shots taken by the top scorers in the league, we are curious to see if there is a certain position that tends to score the most points or if positions score equally. The use of box plots and group summary statistics in our research will be important to come to a confident conclusion.

Discovering Correlations

The visuals below show the correlation between points per game (PPG) and (FGA) for for the top 50 leaders in points in the NBA for their respective season. We were able to collect our data by using the NBA statistics tab on the ESPN website. The purpose of these larger scatter plots is to hopefully set a foundation that the more shots a single player takes, the more individual points they score.


Top 50 PPG Scorers 2020−2021
Points per game
35
30
25
20
14
16
18
20
22
FGA per game


Top 50 PPG Scorers 2021−2022
Points per game
32
28
24
20
12.5
15.0
17.5
20.0
FGA per game



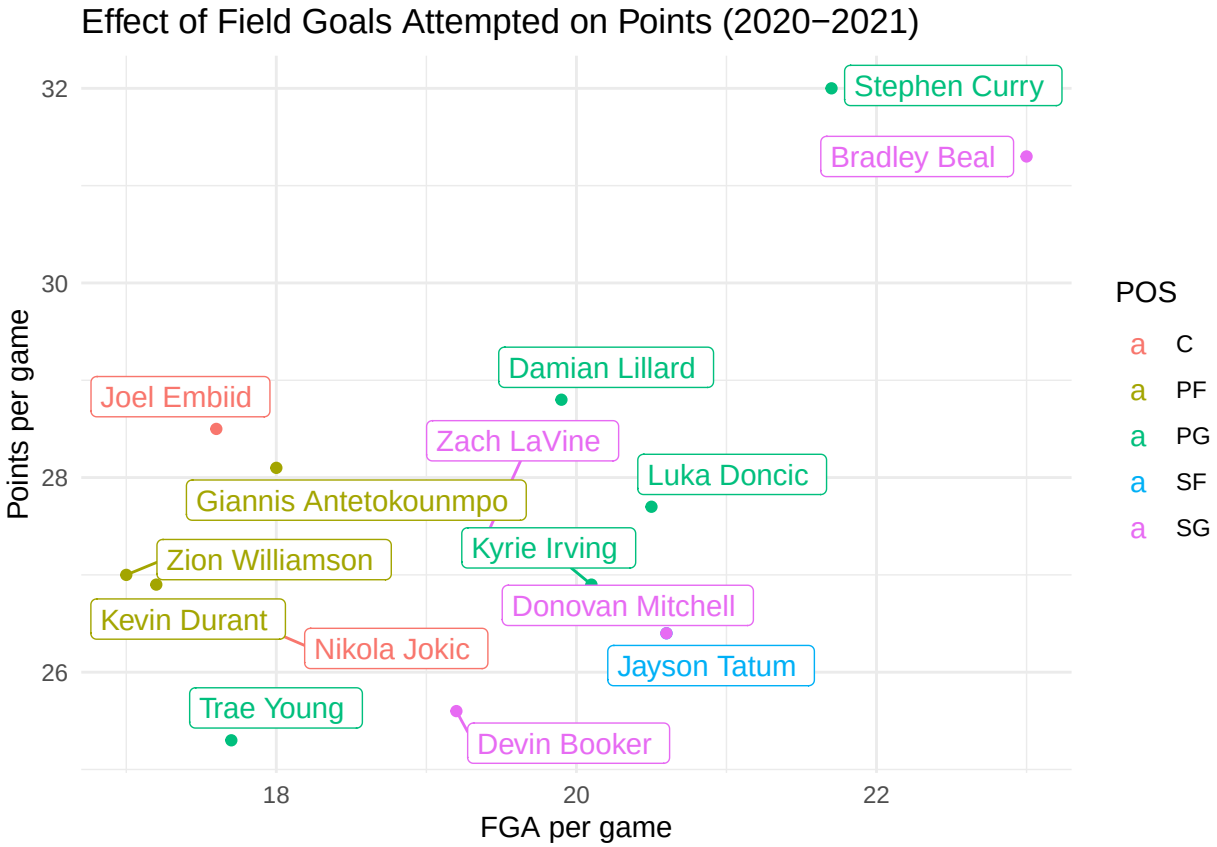
It is very obvious that all graphs depict a positive correlation between the variables, however the three seasons have somewhat different slopes regarding the best fit line. While the scales aren't a perfect match, they are extremely similar for each graph which allows us to compare them.

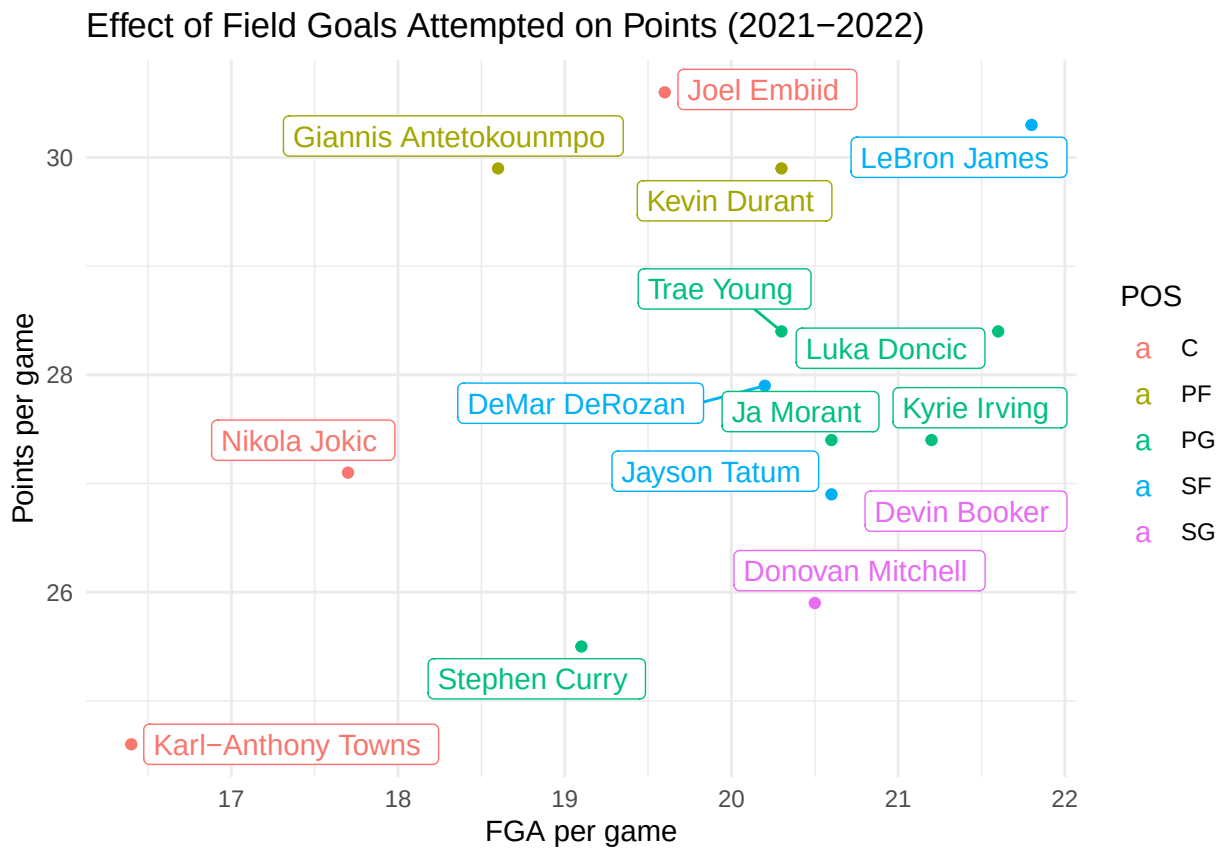
A couple key takeaways from these visualizations are the median of the data sets and potential outliers. The median for each visualization seems to remain constant at around 17.5 field goals attempted and 25 points, roughly speaking. This is expected because we are analyzing the 3 most recent NBA seasons, so heavy change regarding the top scorers is unlikely. Hypothetically if we used the top 50 scorers from 3 different decades of the NBA, the medians and line of best fits would be much more different. Usually outliers on a visualization can be seen at the most left or right point of the picture. But since we are using a line of best fit, the biggest outliers in our sample will be the points furthest away from the shaded area. Overall, outliers don't seem to play a huge role in our sample data. The 2020-2021 season does appear to have the most points significantly away from the line of best fit, both present above and below.

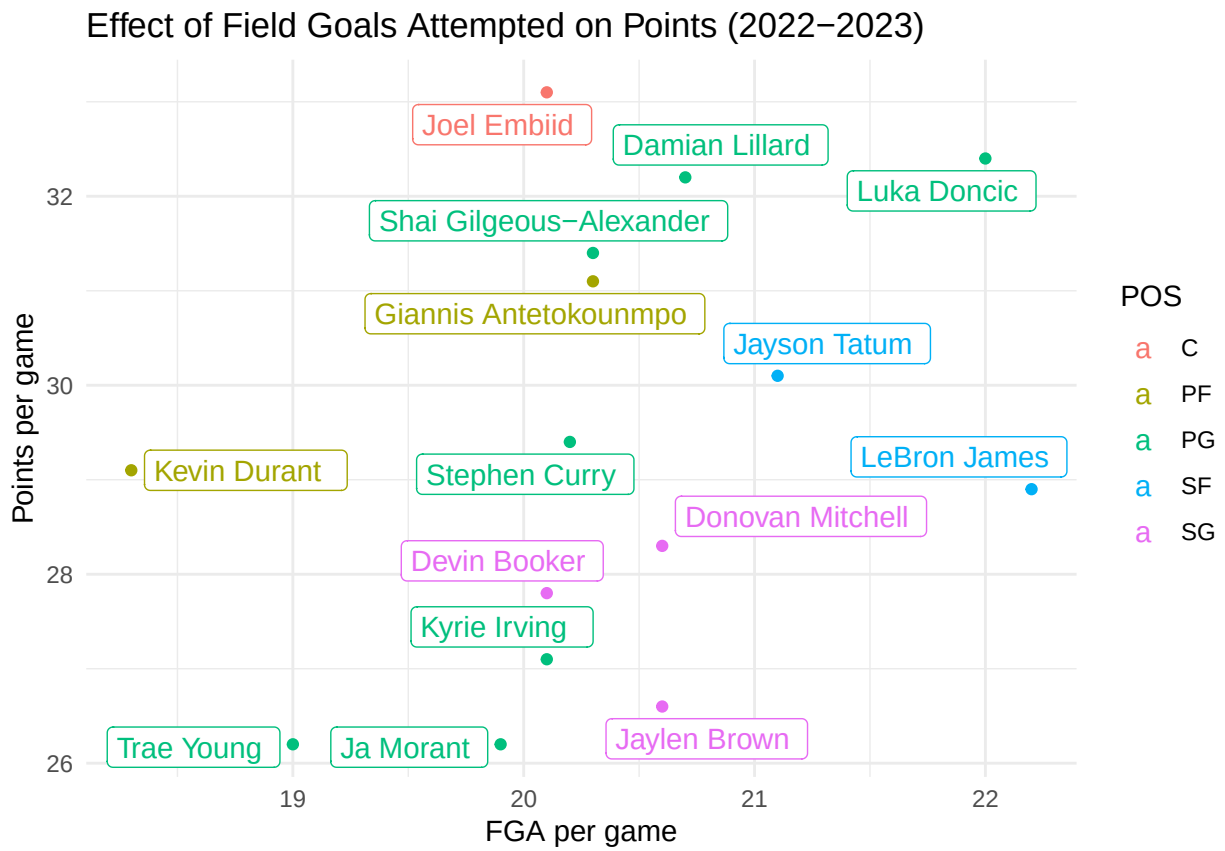
To conclude, the 50 player scatter plot displayed what we had already expected to be true: The more points a player scores, the more shots they tend to shoot.

Narrowing the Samples

Next, we wanted to isolate the top 15 scorers into one single scatter plot for each of the last three years. So, we slimmed the samples for each year down to the top 15 leaders points per season and focus on position to see if that plays a role in either variable. A main reason why these scatter plots differ from the ones we previously analyzed, is because now we are also looking to see how a player's position, those including, point guard, shooting guard, small forward, power forward, and center, could affect how many points players are scoring per game. We created these scatter plots from the tidied data isolating the top 15 scorers from the past three years, where the color of the three scatter plots correlates to the position they play.


Effect of Field Goals Attempted on Points (2020−2021)
Points per game
FGA per game
32
30
28
26
18
20
22
Stephen Curry
Bradley Beal
Damian Lillard
Joel Embiid
Zach LaVine
Luka Doncic
Giannis Antetokounmpo
Kyrie Irving
Zion Williamson
Donovan Mitchell
Kevin Durant
Nikola Jokic
Jayson Tatum
Trae Young
Devin Booker
POS
C
PF
PG
SF
SG


Effect of Field Goals Attempted on Points (2021−2022)
Joel Embiid
Giannis Antetokounmpo
LeBron James
Kevin Durant
Trae Young
Luka Doncic
DeMar DeRozan
Ja Morant
Kyrie Irving
Nikola Jokic
Jayson Tatum
Devin Booker
Donovan Mitchell
Stephen Curry
Karl−Anthony Towns
Points per game
30
28
26
17
18
19
20
21
22
FGA per game
POS
C
PF
PG
SF
SG



Again, a positive correlation between field goals attempted and points in all three scatter plots is present, with the strongest correlation being from 2021-2022. The weakest correlation between points per game and FGA per game is from 2022-2023. There seems to be a similar trend in all three scatter plots and that is referring to point guards being on the higher end of FGA per game. Other than that, there seems to be little indication that position has a factor on players FGA per game. Each of the thirty teams in the NBA have different styles of play and are stronger in some positions while weaker in others. For example, Karl Anthony-Towns, Nikola Jokic, and Joel Embiid, 76ers, Timberwolves, and Nuggets, are the only three centers to make the top 15 in points in any of the three years. The 76ers, Timberwolves, and Nuggets also don't have any players in any other positions on the graph. On the other hand, there are at least 5 point guards in the top 15 all three years. Many teams like to have a star point guard like Steph Curry or Damian Lillard. Every team typically has a predominant scorer and it could be any one of the five positions based on what style of basketball the team wants to play. All of this reestablishes the point that the best players on the team take the most shots.

There appears to be little, to no, relationship between position and FGA when looking at the scatter plots. The reasoning for this could be that every team is unique and every season is different from the last. Another main reason why we chose FGA is because the result would also tell us how much free throw attempts per game (FTA) affect PPG. The positive correlation between FGA and PPG shows that while some players have higher FTA per game, perhaps because they are only worth one point it does not drastically affect PPG.

Box Plots and Summary Statistics Tables by Season

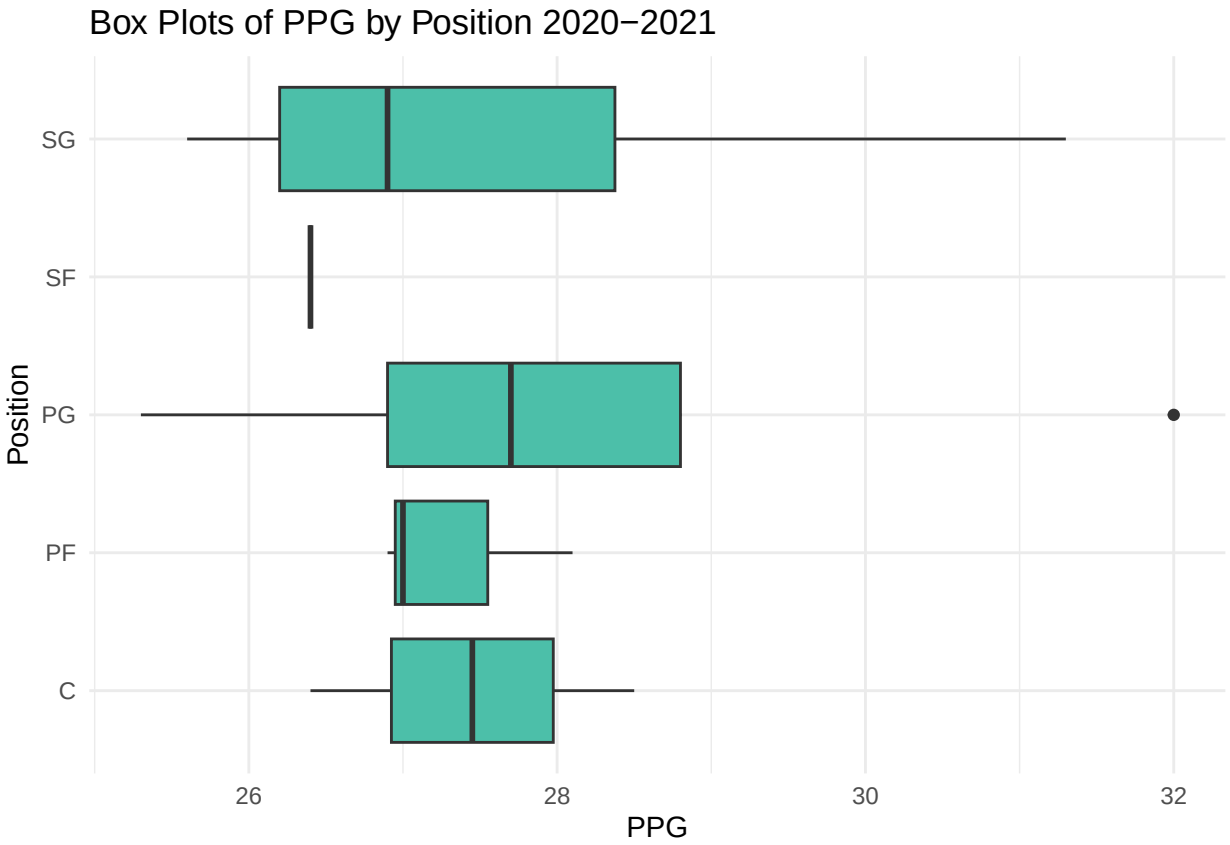


Table 1: Summary Statistics for Points by Position 2021-2022

	n	Min	Q1	Median	Q3	Max	MAD	Mean	SD
C	2	26	27	27	28	28	2	27	1
PF	3	27	27	27	28	28	0	27	1
PG	5	25	27	28	29	32	2	28	3
SF	1	26	26	26	26	26	0	26	NA
SG	4	26	26	27	28	31	1	28	3

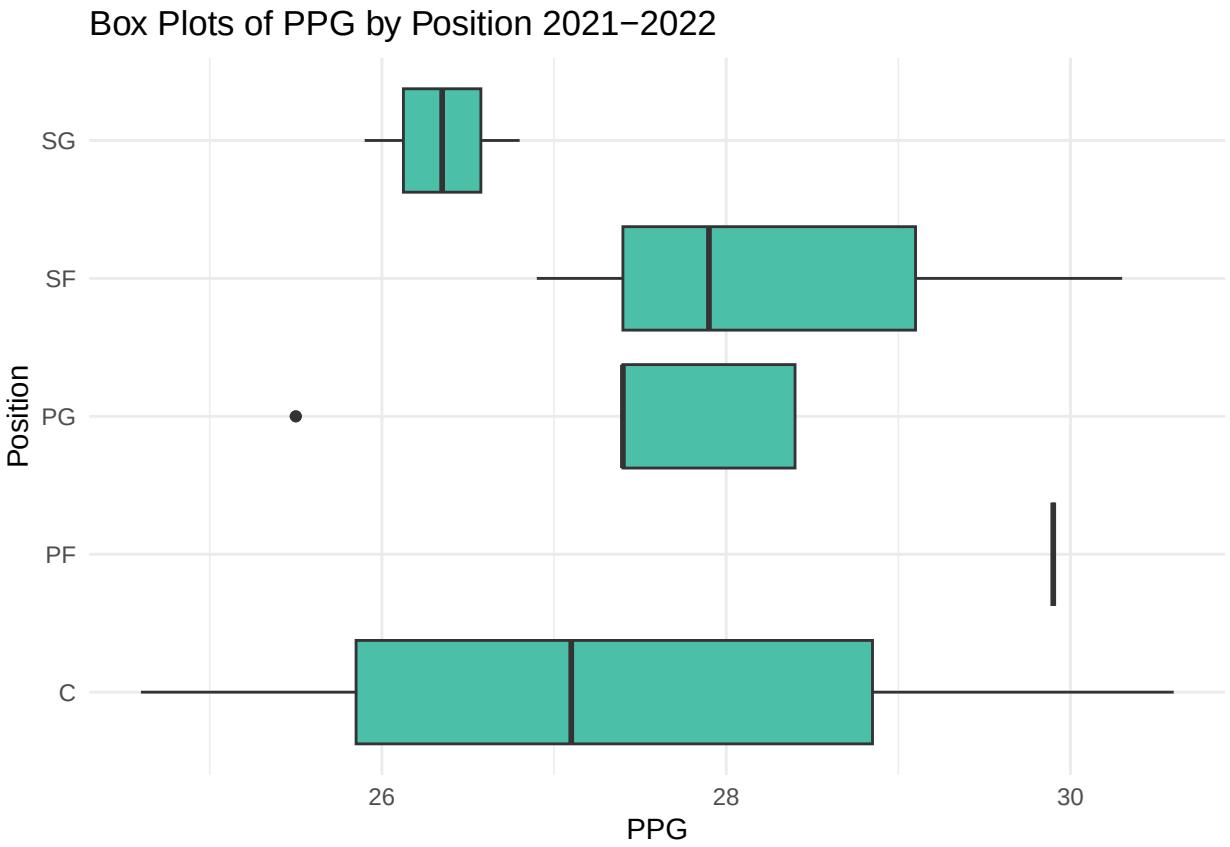


Table 2: Summary Statistics for Points by Position 2021-2022

	n	Min	Q1	Median	Q3	Max	MAD	Mean	SD
C	3	25	26	27	29	31	4	27	3
PF	2	30	30	30	30	30	0	30	0
PG	5	26	27	27	28	28	1	27	1
SF	3	27	27	28	29	30	1	28	2
SG	2	26	26	26	27	27	1	26	1

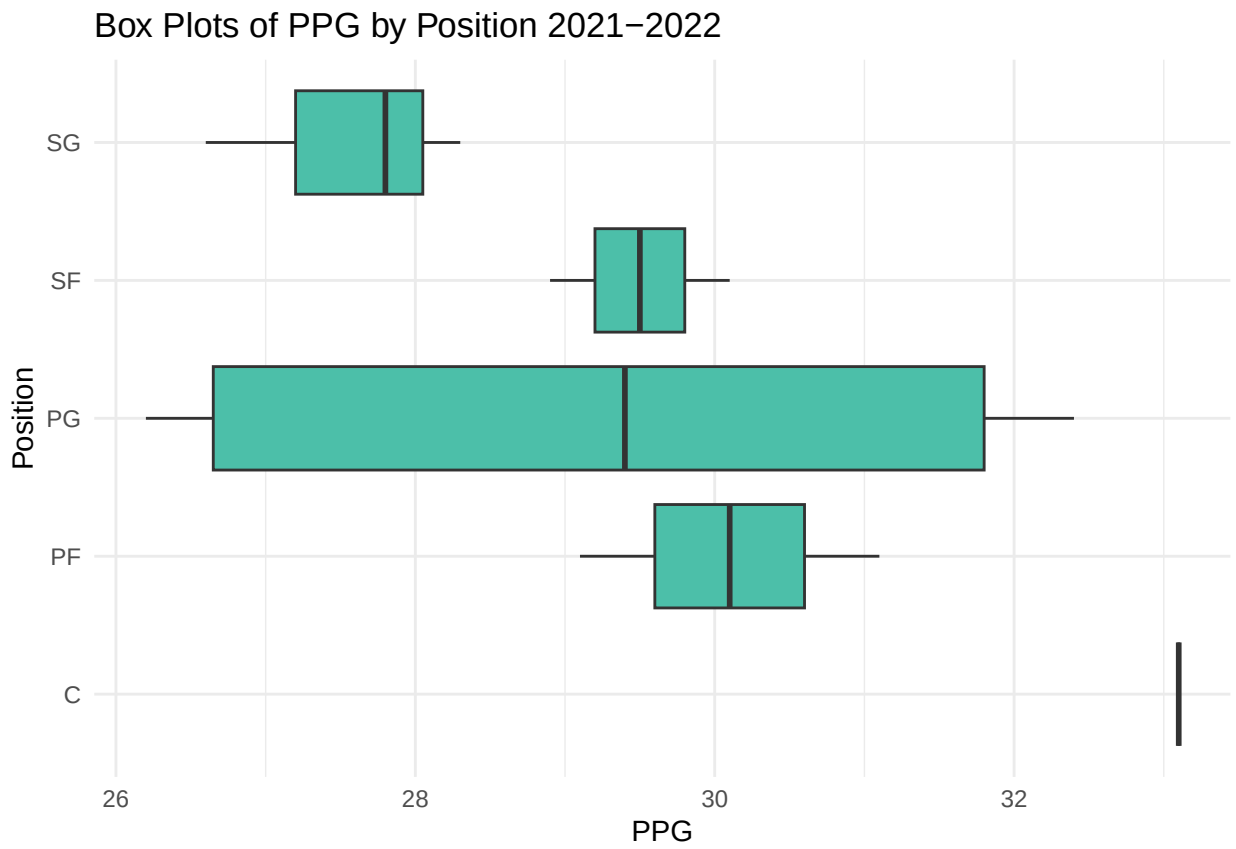


Table 3: Summary Statistics for Points by Position 2021-2022

	n	Min	Q1	Median	Q3	Max	MAD	Mean	SD
C	1	33	33	33	33	33	0	33	NA
PF	2	29	30	30	31	31	1	30	1
PG	7	26	27	29	32	32	4	29	3
SF	2	29	29	30	30	30	1	30	1
SG	3	27	27	28	28	28	1	28	1

There is a lot to take in when looking at the boxplots and numerical data from each year. The boxplots were created to provide another form of visual insight on the relationship between points per game and positions. When initially looking at each visual, an immediate reaction might be to say that there is no correlation because the boxplots by position all differ. But it's important to consider that the number of observations by position each season is different, which leads to the misleading skew in box plot shapes.

Disregarding position, the sample minimum for each season hovered around 25 ppg. The point guard position (PG) seems to dominate the minimum value slot out of all the other positions for all 3 seasons, especially in 2022-2023 where the number of observations is almost half of the entire sample. Seeing that PGs are consistently in the back half of the sample and not towards the top for the mean and max supports an idea already known, that the point guard isn't the most scoring dominant position in the NBA. This isn't only

true according to the sample data, but this is also true according to general basketball logic. One of the many responsibilities of the point guard on the basketball court is to bring the ball up the court and get the offense situated. For those who are unfamiliar with the sport, getting the offense to run smoothly includes lots of passing the ball. Lots of passing the ball translates to less shots being taken and less individual points being scored. On the other side, the sample maximum for each season was around 32 ppg. Most contributed by the power forward (PF) and center positions (C), also known as the front court. Unlike the point guard, the taller players are closer to the basket and have less offensive responsibilities. After calculating the cumulative sample means for each position (not shown on table) the means came out to 28.1 ppg (PG), 27.5 ppg (SG) and 28.3 ppg (SF), 28.7 ppg (PF), 28.0 ppg (C). Why are the means relatively equal when the minimums and maximums are mainly single positions? This could be because the same players appear in all 3 box plots, given the fact that the data used is from the 3 most recent NBA seasons. In other words, the best scorers in the NBA at a single period are likely to be in the top percentile for more than a single year. For example, Joel Embiid averaged 28.5 ppg in 20-21, 30.6 ppg in 21-22, and 33.1 ppg in 22-23.

After analyzing the summary statistics for each position from each individual season, we've come to a confident conclusion. When considering the top scorers in the NBA in recent years, the back court (guards) tend to score less on average than the players in the front court (forwards). We were able to come to this conclusion by using the combination of the visuals and numerical tables with background knowledge of the position responsibilities.

Conclusion

So in the end, we've come to the conclusion that FGA and PPG are positively correlated, but position does not heavily affect FGA. After analyzing the box plots and statistics tables from each season, it does also appear that the best of the best when it comes to scoring points are the taller players. The power forwards and centers seem to consistently be at the top of league scoring in each of the last 3 NBA seasons. It's very interesting to realize that although FGA is positively correlated with PPG, it doesn't automatically mean position and FGA are positively correlated as well. The coaching schemes determine who the main scorer for each team will be, giving them better looks to score and more chances per game. Every position can score and every position needs more attempts to score more.

```
knitr::opts_chunk$set(
  echo = FALSE,
  warning = FALSE,
  message = FALSE,
  fig.align = "center"
)

#Creating a 50 player scoring scatter plot 2020-2021
library(dplyr)
library(readr)
library(ggplot2)
library(tidyr)
library(rvest)
library(stringr)

rawScoringDataTwo <- read_html(
  x = "https://www.espn.com/nba/stats/player/_/season/2021/seasontype/2"
) %>%
  html_elements(css = "table") %>%
  html_table()

rawDataTwo <- bind_cols(rawScoringDataTwo[[1]], rawScoringDataTwo[[2]])
```

```r
ggplot(rawDataTwo) +
  aes(x = FGA, y = PTS) +
  geom_point(shape = "circle", size = 1.5, colour = "#232529") +
  geom_smooth(span = 1L) +
  labs(
    x = "FGA per game",
    y = "Points per game",
    title = "Top 50 PPG Scorers 2020-2021"
  ) +
  theme_minimal()

#Creating a 50 player scoring scatter plot 2021-2022
library(dplyr)
library(readr)
library(ggplot2)
library(tidyr)
library(rvest)
library(stringr)


rawScoringData <- read_html(
  x = "https://www.espn.com/nba/stats/player/_/season/2022/seasontype/2"
) %>%
  html_elements(css = "table") %>%
  html_table()

rawData <- bind_cols(rawScoringData[[1]], rawScoringData[[2]])

ggplot(rawData) +
  aes(x = FGA, y = PTS) +
  geom_point(shape = "circle", size = 1.5, colour = "#232529") +
  geom_smooth(span = 1L) +
  labs(
    x = "FGA per game",
    y = "Points per game",
    title = "Top 50 PPG Scorers 2021-2022"
  ) +
  theme_minimal()

#Creating a 50 player scoring scatter plot 2022-2023
library(dplyr)
library(readr)
library(ggplot2)
library(tidyr)
library(rvest)
library(stringr)

rawScoringDataThree <- read_html(
  x = "https://www.espn.com/nba/stats/player/_/season/2023/seasontype/2"
) %>%
  html_elements(css = "table") %>%
  html_table()
```

```
rawDataThree <- bind_cols(rawScoringDataThree[[1]], rawScoringDataThree[[2]])

ggplot(rawDataThree) +
  aes(x = FGA, y = PTS) +
  geom_point(shape = "circle", size = 1.5, colour = "#232529") +
  geom_smooth(span = 1L) +
  labs(
    x = "FGA per game",
    y = "Points per game",
    title = "Top 50 PPG Scorers 2022-2023"
  ) +
  theme_minimal()
#Creating a 15 player scatter plot for the 2020-2021 top scorers
scoringDataTwo <- subset(rawDataTwo, RK <= 15)

scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "WSH", " ")
scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "PHI", " ")
scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "DEN", " ")
scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "POR", " ")
scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "CHI", " ")
scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "BKN", " ")
scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "DAL", " ")
scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "MIL", " ")
scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "ATL", " ")
scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "PHX", " ")
scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "GS", " ")
scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "NO", " ")
scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "BOS", " ")
scoringDataTwo$Name <- str_replace(scoringDataTwo$Name, "UTAH", " ")

removeColumns <- c("RK", "MIN", "FGM", "3PA", "FG%", "3PM", "3P%", "FTM", "FTA", "FT%", "REB", "AST", "S
scoringDataTwo <- select(scoringDataTwo, -one_of(removeColumns))

library(ggrepel)

ggplot(scoringDataTwo) +
 aes(x = FGA, y = PTS, colour = POS, label = Name) +
 geom_point(shape = "circle", size = 1.5) +
  geom_label_repel() +
 scale_color_hue(direction = 1) +
 labs(x = "FGA per game", y = "Points per game", title = "Effect of Field Goals Attempted on Points (20
 theme_minimal()
#Creating a 15 player scatter plot for the 2021-2022 top scorers
scoringData <- subset(rawData, RK <= 15)

scoringData$Name <- str_replace(scoringData$Name, "BKN", " ")
scoringData$Name <- str_replace(scoringData$Name, "PHI", " ")
scoringData$Name <- str_replace(scoringData$Name, "DEN", " ")
scoringData$Name <- str_replace(scoringData$Name, "MEM", " ")
scoringData$Name <- str_replace(scoringData$Name, "CHI", " ")
scoringData$Name <- str_replace(scoringData$Name, "ATL", " ")
scoringData$Name <- str_replace(scoringData$Name, "DAL", " ")
scoringData$Name <- str_replace(scoringData$Name, "MIL", " ")
```

```
scoringData$Name <- str_replace(scoringData$Name, "LAL", " ")
scoringData$Name <- str_replace(scoringData$Name, "PHX", " ")
scoringData$Name <- str_replace(scoringData$Name, "GS", " ")
scoringData$Name <- str_replace(scoringData$Name, "MIN", " ")
scoringData$Name <- str_replace(scoringData$Name, "BOS", " ")
scoringData$Name <- str_replace(scoringData$Name, "UTAH", " ")

scoringData <- select(scoringData, -one_of(removeColumns))

ggplot(scoringData) +
 aes(x = FGA, y = PTS, colour = POS, label = Name) +
 geom_point(shape = "circle", size = 1.5) +
  geom_label_repel() +
 scale_color_hue(direction = 1) +
 labs(x = "FGA per game", y = "Points per game", title = "Effect of Field Goals Attempted on Points (20
 theme_minimal()
#Creating a 15 player scatter plot for the 2022-2023 top scorers
scoringDataThree <- subset(rawDataThree, RK <= 15)

scoringDataThree$Name <- str_replace(scoringDataThree$Name, "OKC", " ")
scoringDataThree$Name <- str_replace(scoringDataThree$Name, "PHI", " ")
scoringDataThree$Name <- str_replace(scoringDataThree$Name, "LAL", " ")
scoringDataThree$Name <- str_replace(scoringDataThree$Name, "POR", " ")
scoringDataThree$Name <- str_replace(scoringDataThree$Name, "BKN/", " ")
scoringDataThree$Name <- str_replace(scoringDataThree$Name, "BKN/PHX", " ")
scoringDataThree$Name <- str_replace(scoringDataThree$Name, "DAL", " ")
scoringDataThree$Name <- str_replace(scoringDataThree$Name, "MIL", " ")
scoringDataThree$Name <- str_replace(scoringDataThree$Name, "ATL", " ")
scoringDataThree$Name <- str_replace(scoringDataThree$Name, "PHX", " ")
scoringDataThree$Name <- str_replace(scoringDataThree$Name, "GS", " ")
scoringDataThree$Name <- str_replace(scoringDataThree$Name, "MEM", " ")
scoringDataThree$Name <- str_replace(scoringDataThree$Name, "BOS", " ")
scoringDataThree$Name <- str_replace(scoringDataThree$Name, "CLE", " ")

removeColumns <- c("RK", "MIN", "FGM", "3PA", "FG%", "3PM", "3P%", "FTM", "FTA", "FT%", "REB", "AST", "S
scoringDataThree <- select(scoringDataThree, -one_of(removeColumns))

library(ggrepel)

ggplot(scoringDataThree) +
 aes(x = FGA, y = PTS, colour = POS, label = Name) +
 geom_point(shape = "circle", size = 1.5) +
  geom_label_repel() +
 scale_color_hue(direction = 1) +
 labs(x = "FGA per game", y = "Points per game", title = "Effect of Field Goals Attempted on Points (20
 theme_minimal()
#Creating a box plot for the 2020-2021 season top 15 scorers
ggplot(scoringDataTwo) +
  aes(x = PTS, y = POS) +
  geom_boxplot(fill = "#4CBFA9") +
  labs(
    x = "PPG",
    y = "Position",
```

```
    title = "Box Plots of PPG by Position 2020-2021"
  ) +
  theme_minimal()
#Creating a summary statistic table for the 2020-2021 season top 15 scorers
library(psych)
nbaStatsTwo <- psych::describeBy(
  x = scoringDataTwo$PTS,
  group = scoringDataTwo$POS,
  na.rm = TRUE,
  ranges = TRUE,
  quant = c(0.25, 0.75),
  IQR = TRUE,
  mat = TRUE,
  digits = 0.2
)

library(tibble)
positionStatsTwo <- nbaStatsTwo %>%
  remove_rownames() %>%
  column_to_rownames(
    var = "group1"
  ) %>%
  select(
    n, min, Q0.25, median, Q0.75, max, mad, mean, sd
  )

library(knitr)
library(kableExtra)
positionStatsTwo %>%
  kable(
    caption = "Summary Statistics for Points by Position 2021-2022",
    digits = 3,
    format.args = list(big.mark = ","),
    align = rep('c', 11),
    col.names = c("n", "Min", "Q1", "Median", "Q3", "Max", "MAD", "Mean", "SD"),
    booktabs = TRUE
  ) %>%
  kable_styling(
    font_size = 12,
    latex_options = c("scale_down", "HOLD_position")
  )
#Creating a box plot for the 2021-2022 season top 15 scorers
ggplot(scoringData) +
  aes(x = PTS, y = POS) +
  geom_boxplot(fill = "#4CBFA9") +
  labs(
    x = "PPG",
    y = "Position",
    title = "Box Plots of PPG by Position 2021-2022"
  ) +
  theme_minimal()
#Creating a summary statistic table for the 2021-2022 season top 15 scorers
library(psych)
```

```r
nbaStats <- psych::describeBy(
  x = scoringData$PTS,
  group = scoringData$POS,
  na.rm = TRUE,
  ranges = TRUE,
  quant = c(0.25, 0.75),
  IQR = TRUE,
  mat = TRUE,
  digits = 0.2
)

library(tibble)
positionStats <- nbaStats %>%
  remove_rownames() %>%
  column_to_rownames(
    var = "group1"
  ) %>%
  select(
    n, min, Q0.25, median, Q0.75, max, mad, mean, sd
  )

library(knitr)
library(kableExtra)
positionStats %>%
  kable(
    caption = "Summary Statistics for Points by Position 2021-2022",
    digits = 3,
    format.args = list(big.mark = ","),
    align = rep('c', 11),
    col.names = c("n", "Min", "Q1", "Median", "Q3", "Max", "MAD", "Mean", "SD"),
    booktabs = TRUE
  ) %>%
  kable_styling(
    font_size = 12,
    latex_options = c("scale_down", "HOLD_position")
  )
#Creating a box plot for the 2022-2023 season top 15 scorers
ggplot(scoringDataThree) +
  aes(x = PTS, y = POS) +
  geom_boxplot(fill = "#4CBFA9") +
  labs(
    x = "PPG",
    y = "Position",
    title = "Box Plots of PPG by Position 2021-2022"
  ) +
  theme_minimal()

#Creating a summary statistic table for the 2022-2023 season top 15 scorers
library(psych)
nbaStatsThree <- psych::describeBy(
  x = scoringDataThree$PTS,
  group = scoringDataThree$POS,
  na.rm = TRUE,
```

```
  ranges = TRUE,
  quant = c(0.25, 0.75),
  IQR = TRUE,
  mat = TRUE,
  digits = 0.2
)

library(tibble)
positionStatsThree <- nbaStatsThree %>%
  remove_rownames() %>%
  column_to_rownames(
    var = "group1"
  ) %>%
  select(
    n, min, Q0.25, median, Q0.75, max, mad, mean, sd
  )

library(knitr)
library(kableExtra)
positionStatsThree %>%
  kable(
    caption = "Summary Statistics for Points by Position 2021-2022",
    digits = 3,
    format.args = list(big.mark = ","),
    align = rep('c', 11),
    col.names = c("n", "Min", "Q1", "Median", "Q3", "Max", "MAD", "Mean", "SD"),
    booktabs = TRUE
  ) %>%
  kable_styling(
    font_size = 12,
    latex_options = c("scale_down", "HOLD_position")
  )
```