# STATS 15 Final Project: NBA All-Stars

Joana Fang, Tony He, Evan Garcia, and Julia Tran

2022-11-26

## Section 1 - Introduction

### 1.1 - Research Question and Scope of Analysis

The National Basketball Association (NBA) is by far the most well-recognized basketball league in North America, and likely the world. As with any other sport, the athletes that make up the NBA spend long hours both during the season and off-season training to ensure that their bodies and skills remain in top condition. However, it is inevitable that the natural processes of time and countless hours on the court lead to slower reaction times, higher rates of injuries, and ancillary to that, a lower likelihood of recovering from such injuries. For this project, our group aims to investigate the trajectory of the careers of the players at the top of the top: All-Stars. Some of the guiding questions that lead to our research question include: When in their career does an All-Star hit their prime? How do they compare to their rookie-selves or other All-Stars? What parts of the game do players improve the most on and how are specific positions affected? Ultimately, we are led to our overarching research question: **How does a 3-time All-Star's performance in the NBA change over time?**

### 1.2 - Background on the NBA

***Structure of the NBA:*** The NBA is presently comprised of 30 teams, with an Eastern and Western Conference, each consisting of 15 teams. Players primarily join an NBA team through the NBA Draft. Additionally, the conferences are important in the All-Star selection process as 12 players from each Conference are selected to form the 24 player All-Star roster for each year.

***All-Star Voting Process:*** The starting 5 for each Conference are voted for by a combination of fan, current player, and media vote. Prior to 2017, these 10 players were determined using a purely fan vote, but a rule change that year made it so that fan vote accounted for 50% and player and media vote counted for 25% each. The 14 reserves are determined by NBA coaches, who cannot vote for players on their own team. If there are selections that are later unavailable due to injury, the NBA commissioner chooses a replacement.

***Positions:*** For the purposes of our project, we chose to focus on 5 positions of the basketball team. These 5 positions include Center (C), Power Forward (PF), Small Forward (SF), Point Guard (PG), and Shooting Guard (SG). Centers and Power Forwards are the more defensive heavy positions whereas Small Forwards, Point Guards, and Shooting Guards are all more focused on offense.

- ***Centers:*** This player is expected to block shots and and get rebounds while on defense, and while on offense they are expected to get offensive rebounds and and attempt short-distance shots.
- ***Power Forwards:*** These players are very similar to Centers with the exception that they take some longer-distance shots.
- ***Small Forwards:*** Small forwards are all over the court and are expected to be able to shoot from both long and short distances.

- ***Point Guards:*** Point guards should be the best dribbler and passer on the team and they are typically in charge of offense as a result. They also defend against the opposing team's point guard and try to steal the ball from them.
- ***Shooting Guards:*** These players are typically the best shooter on the team, so they should be good at shooting long distance and are also usually good dribblers.

## 1.3 - Variable & Data Explanation

Our player statistics data is scraped directly from basketballreference.com, a well-known and reputable website that has collected data from each NBA season stretching back several decades. Additionally, we used the website, basketball.realgm.com, in order to scrape the All-Star rosters from all of the All-Star games from 2004 to present. We chose these years because beginning in 2004 till now, the NBA had 30 teams,so players have the same likelihood of being chosen.

It is important for us to note that although we are using All-Star rosters from 2004 to present, many of the players on these rosters began their careers before 2004. We are choosing to include this data because we believe that the data from the beginnings of these players' careers is integral to investigating our research question, which involves how a player's career changes over time. Despite this discrepancy, we are still maintaining standards in our data, as by choosing All-Stars from 2004 and beyond, we are ensuring that all of our selected players have at least one season after 82 game season change.

Our Player Statistics Data Frame has 1000 observations, each representing an individual All-Star during a single regular season, and 31 total variables. Our All-Star Data Frame contains the All-Star rosters from our selected time frame, and has 495 observations, each representing an All-Star from the All-Star team of that year, and 9 total variables.

Considering that our group is attempting to measure a variable that can be as vague as "performance", we are planning to explore the different facets that constitute an NBA player's performance so that we can create a holistic picture of an All-Star's game. We will be using the response variables below to measure performance. These response variables are all for individual players, and are averaged per game across an individual season.

#### Explanatory Variables

1. ***Age:*** The age of a player in the data frame is the age that they were during the given season. For example, LeBron James was 20 during the 2004-2005 season and 29 during the 2013-2014 season.
2. ***POS:*** This variable denotes the position of the player. There are 5 possible positions in our data set: Center (C), Power Forward (PF), Small Forward (SF), Point Guard (PG), Shooting Guard (SG).

#### Response Variables

1. ***PTS:*** The amount of points scored by an individual player. Points consist of three-pointers, two-pointers, and any other shot that makes it into the basket. This is an integral part of the offensive side of a player's performance; the more points a player scores the better we can assume their offensive game is.
2. ***AST:*** The number of assists made by an individual player. An assist is counted when a player passes the ball to another player who scores after that pass.
3. ***3PA:*** The number of three-pointers attempted, including shots that miss the basket and those that make it. It is considered a three-point attempt when a player shoots the ball from behind the three-point line on a basketball court.

4. ***3P%:*** Three-point percentage is the number of three-pointers made divided by the total amount of three-pointers attempted (3PA). The higher this percentage is, the more accurate and efficient we can assume the player is at shooting the ball.

5. ***2PA:*** The amount of two-pointers attempted, including shots that miss the basket and those that make it. It is considered a two-point attempt when a player shoots the ball from anywhere inside the three-point line or with their foot touching the line on either side.

6. ***2P%:*** Two-point percentage is the number of two-pointers made divided by the total amount of two-pointers attempted (2PA). A higher two-point percentage indicates similar implications as a higher three-point percentage, where we can assume that a player is better at scoring and making the most out of their attempts.
7. ***BLK:*** The number of blocks made by an individual player. A block is counted as when a player makes a deflection of a shot from an opposing player that is legal within the rules of the game, thereby preventing the other player from scoring. The more blocks an individual player has recorded, the better we can assume the defensive aspect of their performance is.
8. ***FGA:*** The number of field goals attempted. Field goals attempted are the total of three-pointers and two-pointers attempted.
9. ***FG%:*** The field goal percentage is the number of field goals made divided by the amount of field goals attempted (FGA).
10. ***ORB:*** The amount of offensive rebounds made by an individual player. An offensive rebound occurs when a shot by the attacking team is missed and comes back as a loose ball, and it is then recovered by a member of that same team, thereby maintaining possession.
11. ***DRB:*** The amount of defensive rebounds made by an individual player. A defensive rebound occurs when a shot by the attacking team is missed and comes back as a loose ball, and it is then recovered by a member of the defending team, thereby switching possession.

# Section 2 - Creating The Data Sets

## 2.1 - The All-Star Data Set

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.1.2
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

```
## Warning: package 'purrr' was built under R version 4.1.2
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
## Warning: package 'stringr' was built under R version 4.1.2
```

```
## Warning: package 'forcats' was built under R version 4.1.2
```

```
library(dplyr)
library(rvest) # package used to extract information from websites
```

```
## Warning: package 'rvest' was built under R version 4.1.2
```

```
library(stringr) # package used to process text and strings
library(readr)
library(ggrepel)
library(reshape2)
```

### Scraping the Data

Here, we have a function made using the help from the Professor that takes the website containing the All-Star rosters and returns a data frame containing the names of the All-Star players and other information of the given year.

```
make_tables <- function(year){
  url <- paste0("https://basketball.realgm.com/nba/allstar/game/rosters/", year)
  list_of_tables <-
    read_html(url) %>%
    html_nodes("table") %>%
    html_table()
  n <- length(list_of_tables)
  all_star_west <- list_of_tables[[n-1]]
  all_star_east <- list_of_tables[[n]]
  return(bind_rows(all_star_west, all_star_east)%>% mutate(year = year))
}
```

This is where we used the function above to create tables for each of our selected years (2004-Present), and underneath is where we then used bind_rows() in order to create our master all_star data frame containing all of the All-Stars from 2004 to present.

```
all_star_04 <- make_tables(2004)
all_star_05 <- make_tables(2005)
all_star_06 <- make_tables(2006)
all_star_07 <- make_tables(2007)
all_star_08 <- make_tables(2008)
all_star_09 <- make_tables(2009)
all_star_10 <- make_tables(2010)
all_star_11 <- make_tables(2011)
all_star_12 <- make_tables(2012)
all_star_13 <- make_tables(2013)
all_star_14 <- make_tables(2014)
all_star_15 <- make_tables(2015)
all_star_16 <- make_tables(2016)
all_star_17 <- make_tables(2017)
```

```
all_star_18 <- make_tables(2018)
all_star_19 <- make_tables(2019)
all_star_20 <- make_tables(2020)
all_star_21 <- make_tables(2021)
all_star_22 <- make_tables(2022)

all_star <- bind_rows(all_star_04, all_star_05, all_star_06, all_star_07,
                      all_star_08, all_star_09, all_star_10, all_star_11,
                      all_star_12, all_star_13, all_star_14, all_star_15,
                      all_star_16,all_star_17, all_star_18, all_star_19,
                      all_star_20, all_star_21, all_star_22)
```

This code below was to observe how many All-Star selections these players have, and to see what players have the most selections. Additionally, we chose to filter out All-Stars that had fewer than 3 selections because we wanted to focus on the "top" players of the NBA, and by virtue of fan, player, and media votes, we thought that those with as few as 1 or 2 selections could have been flukes, so we chose to eliminate them from our data.

Furthermore, we added a filter to the pipe to get rid of replacement All-Stars, which are chosen by the NBA Commissioner in the case of an original All-Star selection getting injured or being otherwise unfit to play in the game. We chose to do this because the data set that we are working with here already has the original All-Star selections, and we wanted to focus only on them.

```
suppressWarnings(all_star_ranked <- all_star %>%
                   filter(`Selection Type` != c("Western All-Star Replacement Selection", "Eastern All-S
  count(Player) %>%
  filter(n >= 3) %>%
  arrange(desc(n)))
```

## Section 2.2 - The Player Statistics Data Set

To get the data for each player who were selected as an All-Star, we looked at the per game stats for each of them on basketballreference.com, manually scraping and pasting the stats into one table shown below.

```
df <- read_csv('Player_Stats.csv', show_col_types = FALSE)
view(df)
```

We then realized that many of the numeric variables were not numeric in the .csv, so we converted them so that we could actually manipulate the data for our analysis.

```
suppressWarnings(df$G <- as.numeric(df$G))
suppressWarnings(df$GS <- as.numeric(df$GS))
suppressWarnings(df$PTS <- as.numeric(df$PTS))
suppressWarnings(df$`3P%` <- as.numeric(df$`3P%`))
suppressWarnings(df$`3PA` <- as.numeric(df$`3PA`))
suppressWarnings(df$`2P%` <- as.numeric(df$`2P%`))
suppressWarnings(df$`2PA` <- as.numeric(df$`2PA`))
suppressWarnings(df$FG <- as.numeric(df$FG))
suppressWarnings(df$`FG%` <- as.numeric(df$`FG%`))
suppressWarnings(df$`FGA` <- as.numeric(df$`FGA`))
suppressWarnings(df$`MP` <- as.numeric(df$`MP`))
suppressWarnings(df$`AST` <- as.numeric(df$`AST`))
```

```
suppressWarnings(df$`BLK` <- as.numeric(df$`BLK`))
suppressWarnings(df$`ORB` <- as.numeric(df$`ORB`))
suppressWarnings(df$`DRB` <- as.numeric(df$`DRB`))
suppressWarnings(df$`TRB` <- as.numeric(df$`TRB`))
suppressWarnings(df$`STL` <- as.numeric(df$`STL`))
suppressWarnings(df$`eFG%` <- as.numeric(df$`eFG%`))
suppressWarnings(df$PF <- as.numeric(df$`PF`))
suppressWarnings(df$TOV <- as.numeric(df$`TOV`))
suppressWarnings(df$`2P` <- as.numeric(df$`2P`))
suppressWarnings(df$`3P` <- as.numeric(df$`3P`))
suppressWarnings(df$`FG` <- as.numeric(df$`FG`))
```

## Section 2.3 - Cleaning Up the Data

### Outliers

Throughout our exploration of the data, we encountered several outliers. In order to account for outliers such as this, we chose to use median in our summarise() functions which is more effective for our purposes because median is not as easily influenced by outliers as mean is.

```
age_count <- df %>%
  group_by(Age) %>%
  summarise(age_count = n())

age_count %>%
  arrange(age_count) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##      Age age_count
##    <dbl>     <int>
##  1    41         1
##  2    42         1
##  3    43         1
##  4    18         3
##  5    40         3
##  6    39         7
##  7    38        14
##  8    19        16
##  9    37        16
## 10    36        30
```

```
age_count %>%
  arrange(desc(age_count)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##      Age age_count
##    <dbl>     <int>
##  1    26        69
##  2    22        68
```

```
##  3    23        68
##  4    28        65
##  5    24        64
##  6    25        64
##  7    27        64
##  8    29        60
##  9    30        60
## 10    21        58
```

Taking a look at the amount of players at different ages, we saw that there were few players who were 40 or older, or 18 or younger in our dataset. If we kept them, they would hold much more weight in the graph, compared to players in another age group. For example, there are 67 twenty-three year olds, but only 1 player at 42: Vince Carter. Therefore, when looking at trends over age, we decided to look at players who were between 19 and 39 years old. In this, we also eliminated an outlier of 40-year-old Joe Johnson, who played one game with 100 FG%. Last, we also filtered, so the league was the NBA, eliminating players like Marc Gasol in the year he played in Spain.

```
filtered_data <- df %>%
  filter(Lg == "NBA", G >= 2, Age >= 19, Age <= 39)
```

### TOT

Next, we realized that there were repeats in some of the rows for players that played on multiple teams in one season. We went through the data, and saw that for these players, they had an additional row called "TOT" that totaled the player's stats for the season from each of the teams that they played on during that year. Seeing this, we wanted to remove the two additional rows that represented the player's time with the two teams during a single season, while keeping their total ("TOT") stats for that year. Effectively, the data for the year they played for two teams would be unnecessarily emphasized in modeling.

In order to do so, we first used the which() function to find all of the rows in which "TOT" appears.

```
which(filtered_data$Tm == "TOT")
```

```
##  [1]  15  21  34  58  97 133 136 156 167 175 233 294 297 360 411 488 501 525 530
## [20] 533 579 590 598 601 619 650 658 669 674 692 725 729 732 747 758 763 793 799
## [39] 815 819 824 858 916 929 933 979
```

```
filtered_data <- filtered_data[-c(16,17, 22,23, 35,36, 60,61, 99,100, 135,136,
                                  138,139, 159,160, 170,171, 178,179, 240,241,
                                  302,303, 305,306, 369,370, 420,421, 497,498,
                                  510,511, 534,535, 539,540, 542,543, 588,589,
                                  599,600, 607,608, 610,611, 628,629, 659,660,
                                  667,668, 678,679, 683,684, 701,702, 734,735,
                                  738,739, 741,742, 756,757, 767,768, 772,773,
                                  802,803, 808,809, 824,825, 828,829, 833,834,
                                  867,868, 925,926, 938,939, 942,943, 988,989), ]
```

### Players Who Play Multiple Positions

In the graphs below, we faceted by position and plotted Age against the median of Points (PTS). After seeing all of the graphs for each position, we chose to omit players who played multiple positions, listed as

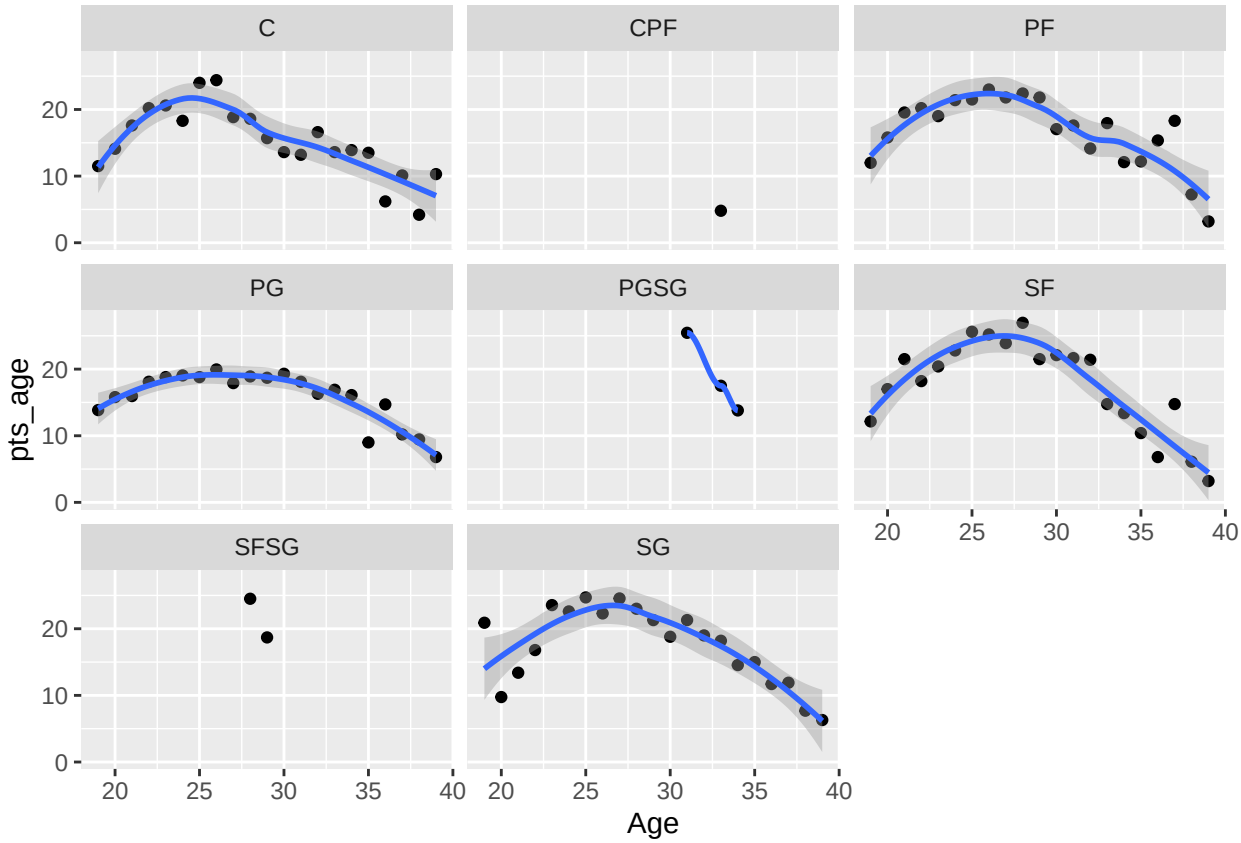
SFSG, PGSG and CPF, because there are so few of them and the resulting graphs for each of those variables did not yield enough data points to create an observable trend.

This graph below exemplifies why we had to omit those positions for when we facet our graphs by position. For other graphs where we do not facet by position, we chose to keep those players because they are not categorized into their positions and are a part of the larger data set.

```
suppressWarnings(filtered_data %>% # Points
  group_by(Age, Pos) %>%
  summarize(pts_age=median(PTS,na.rm = TRUE)) %>%
  arrange(desc(pts_age)) %>% ggplot(aes(x=Age, y=pts_age))
  + geom_point() + facet_wrap(~Pos) + geom_smooth())
```

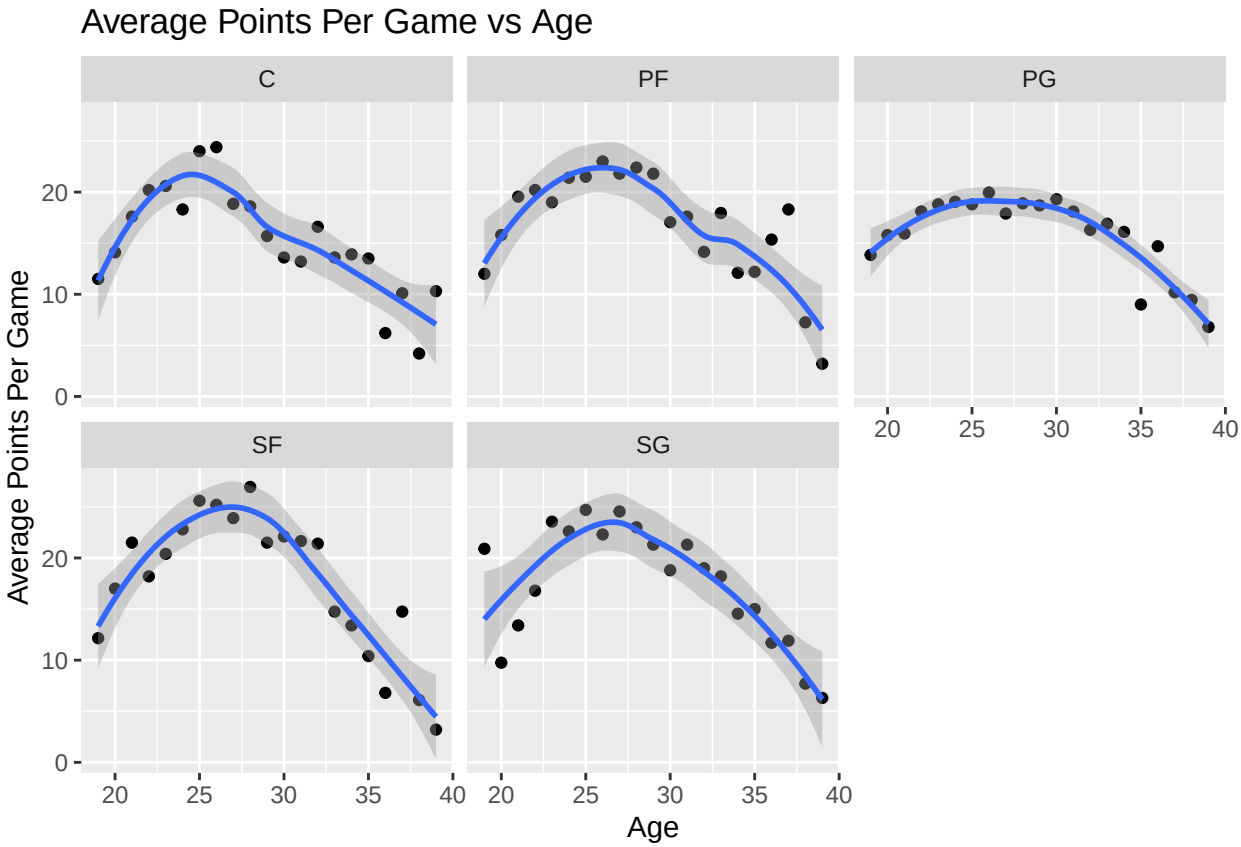
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 30.985

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 2.015

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1.0302

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : span too small. fewer
## data values than degrees of freedom.

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## 30.985

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 2.015

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 0

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 1.0302
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : at 27.995

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : radius 2.5e-05

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 27.995

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 0.005

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : at 29.005

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : radius 2.5e-05

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 2.5e-05

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : zero-width neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : zero-width neighborhood. make span bigger

## Warning: Computation failed in `stat_smooth()`:
## NA/NaN/Inf in foreign function call (arg 5)

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```




This graph is after we filter out the players that play multiple positions, and we can see that each position has an observable trend as we intended.

```
filtered_data %>% # Points
  filter(nchar(Pos) <= 2) %>%
  group_by(Age, Pos) %>%
  summarize(pts_age=median(PTS,na.rm = TRUE)) %>%
  arrange(desc(pts_age)) %>% ggplot(aes(x=Age, y=pts_age)) + geom_point() +
  facet_wrap(~Pos) + geom_smooth() + ylab("Average Points Per Game") +
  ggtitle("Average Points Per Game vs Age")
```




# Section 4 - Data Analysis

## Section 4.1 - Exploratory Data Analysis

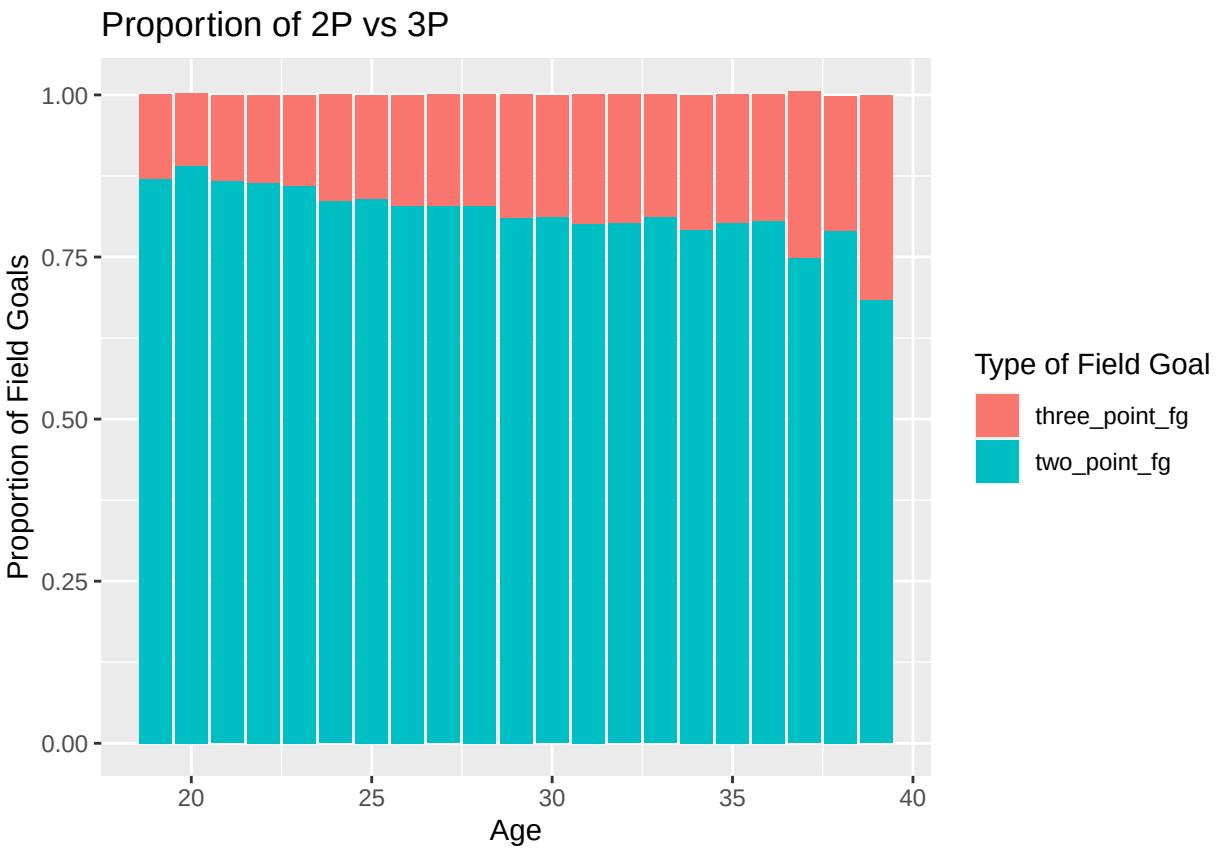
### Function for Labeling Faceted Position Graphs

```
Pos_names <- list(
  'C'="Center",
  'PF'="Power Forward",
  'PG'="Point Guard",
  'SF'="Small Forward",
  'SG'="Shooting Guard"
)

Pos_labeller <- function(variable,value){
  return(Pos_names[value])
}
```

###Shooting Graphs

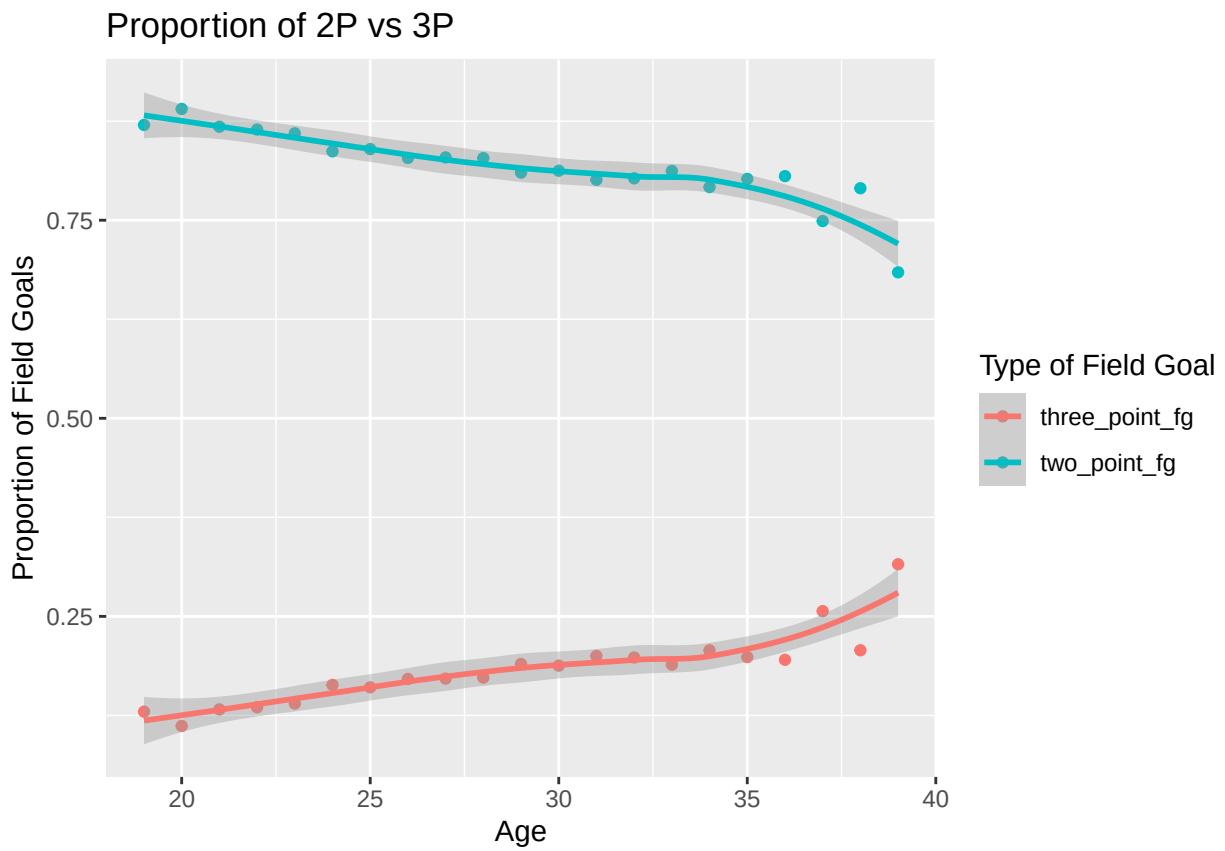
```
goals_two <- filtered_data %>%
  group_by(Age) %>%
```

```
  summarize(two_point_fg = sum(`2P`)/sum(FG))


goals_three <- filtered_data %>%
  group_by(Age) %>%
  summarize(three_point_fg = sum(`3P`)/sum(FG))

goals_combined <- goals_two %>%
  inner_join(goals_three, by = "Age")


goals_long <- goals_combined %>%
  pivot_longer(-Age, names_to = "Type of Field Goal", values_to = "Proportion of Field Goals")

goals_long %>%
  ggplot(aes(x=Age, y= `Proportion of Field Goals`, fill = `Type of Field Goal`)) + geom_col() + ggtitl
```




```
goals_long %>%
  ggplot(aes(x=Age, y=`Proportion of Field Goals`, color =`Type of Field Goal`)) + geom_point() + geom_s
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




Taking a look at the proportion of goals that players made as they age, we can see there is an increasing trend of the proportion of three points sunk and a decreasing proportion of two points sunk. To further understand this trend we use a linear regression model.

```
model15 <- lm(three_point_fg ~ Age, data = goals_combined)
summary(model15)
```

```
##
## Call:
## lm(formula = three_point_fg ~ Age, data = goals_combined)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.033433 -0.007740 -0.001238  0.007332  0.068673
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0046341  0.0232628  -0.199    0.844
## Age          0.0064551  0.0007852   8.221 1.12e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02179 on 19 degrees of freedom
## Multiple R-squared:  0.7805, Adjusted R-squared:  0.769
## F-statistic: 67.58 on 1 and 19 DF,  p-value: 1.116e-07
```

```r
  model16 <- lm(two_point_fg ~ Age, data = goals_combined)
  summary(model16)
```

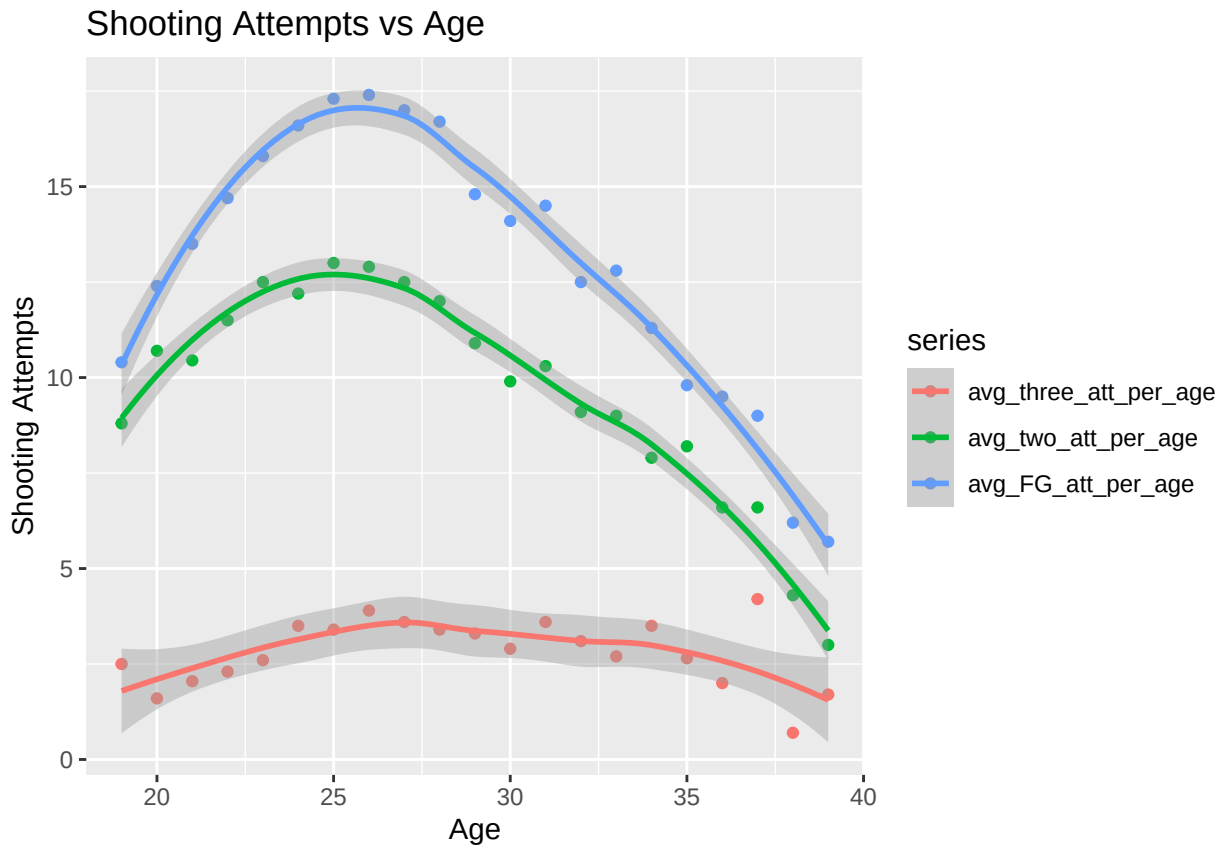
```
##
## Call:
## lm(formula = two_point_fg ~ Age, data = goals_combined)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.069414 -0.007877  0.000956  0.006021  0.032693
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0046102  0.0229983   43.68  < 2e-16 ***
## Age         -0.0064355  0.0007763   -8.29 9.84e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02154 on 19 degrees of freedom
## Multiple R-squared:  0.7834, Adjusted R-squared:  0.772
## F-statistic: 68.72 on 1 and 19 DF,  p-value: 9.836e-08
```

From this test we can see that the trend is significant as far as the decrease and increase of type of point that is made. The amount of 3p made by a player tends to increase by about 0.6% each year that they age and the amount of 2p decrease by about 0.6%. The P value is very low so this cannot be attributed to randomness, as well as the r squared value is fairly close to 1 meaning that there is not much varying between the positive slope and negative slope established. We can conclusively say that as players get older the amount of 3p made increases relative to the total shots they made. We analyze the attempts per age to see how that varies with respect to the shots they actually made.

```r
three_att <- filtered_data %>%
  select(Age, `3PA`) %>%
  group_by(Age) %>%
  summarize(avg_three_att_per_age = median(`3PA`))

FG_att <- filtered_data %>%
  select(Age, `FGA`) %>%
  group_by(Age) %>%
  summarize(avg_FG_att_per_age = median(`FGA`))

two_att <- filtered_data %>%
  select(Age, `2PA`) %>%
  group_by(Age) %>%
  summarize(avg_two_att_per_age = median(`2PA`))

shooting_att <- bind_rows(three_att, two_att, FG_att)

shooting_att <- melt(shooting_att ,  id.vars = 'Age', variable.name = 'series')

#create line plot for each column in data frame
ggplot(shooting_att, aes(x =Age, y =value, color = series)) +
  geom_point() + geom_smooth() + ggtitle("Shooting Attempts vs Age") + ylab("Shooting Attempts")
```

```
## Warning: Removed 126 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 126 rows containing missing values (geom_point).
```




As a player gets older, the graph shows that in general, two-pointer attempts and field goal attempts increase before decreasing tremendously, while three point attempts do not show this trend. This corroborates the proportions shown in the previous graphs, two point attempts overall decrease as they age while three point attempts do not vary as much, thus the proportion of three pointers made would increase while two points made proportions decrease.Indirectly the three points proportion increases as a result of two point attempts and shots made decreasing, the reason for this could be because there are fewer opportunities for players to make two point shots and or the three point shots ability remains consisten over time.

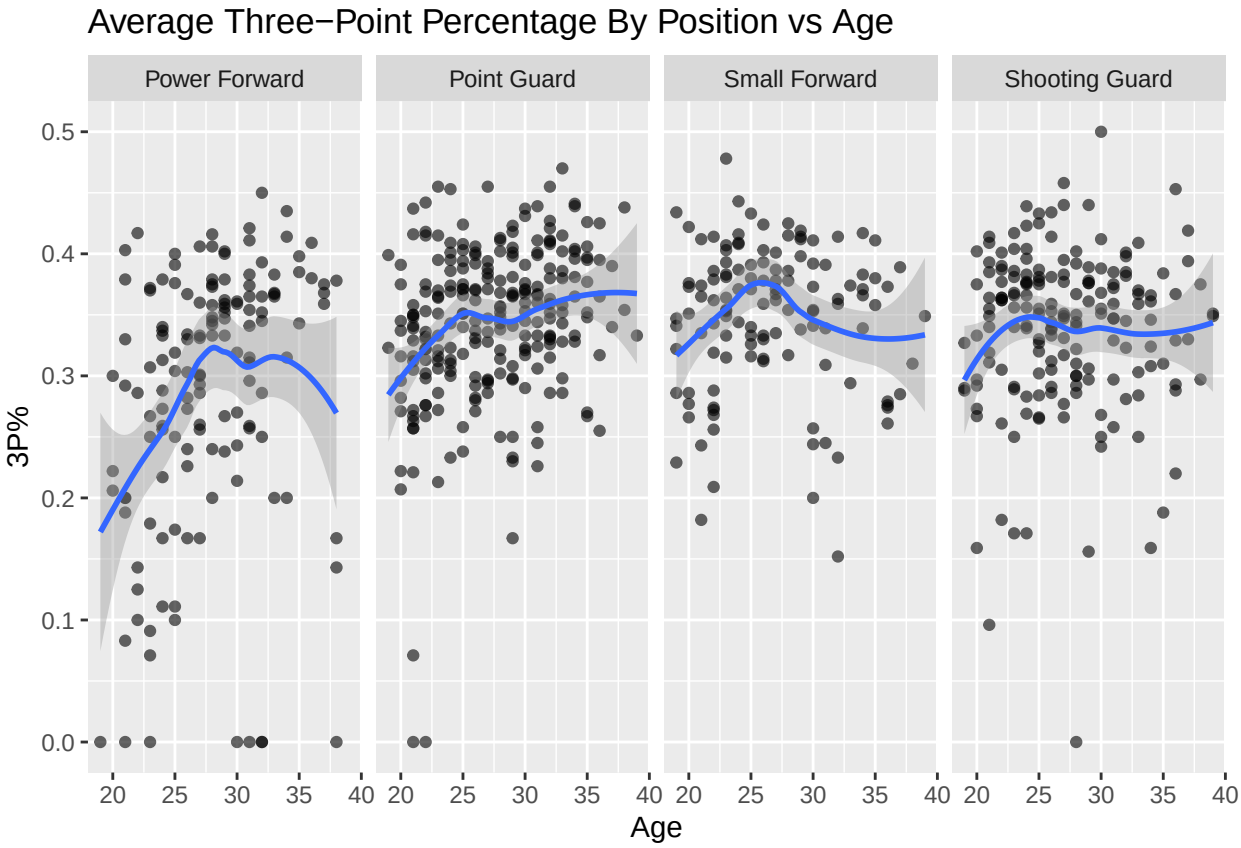
### 3P% Analysis

We wanted to look at three-point percentage in the lens of different positions. We chose to exclude Centers because three-pointers are not a key part of their game. In this graph and the next, we filtered out attempts that were 0 as it should lead to a null 3P% instead of 0 which brings down the average.

```
filtered_data %>% # 3 point %
filter(nchar(Pos) <= 2, Pos != "C", `3PA` != 0, `3P%` != 1) %>%
ggplot(aes(x=Age, y=`3P%`))  + geom_point(alpha = 0.6) +
  geom_smooth() + facet_grid(~Pos, labeller=Pos_labeller) + ggtitle("Average Three-Point Percentage By
```

```
## Warning: The labeller API has been updated. Labellers taking ‘variable‘ and
## ‘value‘ arguments are now deprecated. See labellers documentation.
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




**Sample T test for power forward's 3P% overtime**

In the graphs above, we can see that Power Forwards showed the most variation in that they improved the most in comparison to the other positions. Thus, we decided to take a look at PFs in particular below.

Mean of 3P% at age <= 27 is u1 Mean of 3P% at age >= 30 is u2

Null hypothesis: u1 = u2 Alternative hypothesis: u1 < u2

```
x <- filtered_data %>%
  filter(Pos == "PF", Age <= 27) %>%
  summarize(`3P%` = `3P%`)

y <- filtered_data %>%
  filter(Pos == "PF", Age >= 30) %>%
  summarize(`3P%` = `3P%`)

t.test(x, y, alternative = "less", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
```

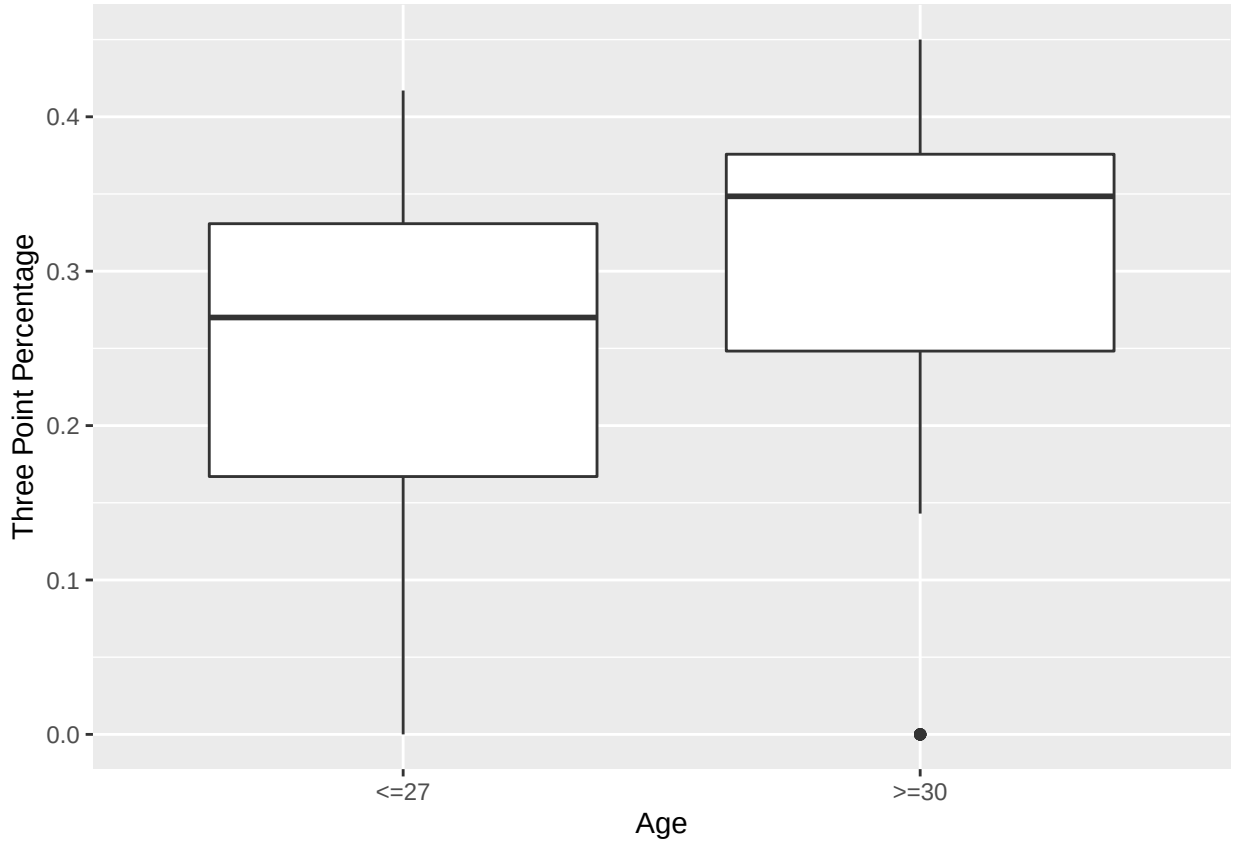
```
## t = -2.2065, df = 102.31, p-value = 0.01479
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.01266933
## sample estimates:
## mean of x mean of y
## 0.2382027 0.2893462
```

Since the P-value is less than 0.05, we reject the null-hypothesis. The increase in the three point percentage of power forwards as they age is statistically significant.

Below is the visualization of the T-test:

```
t_data_x <- cbind(Age='<=27', x)
t_data_y <- cbind(Age='>=30', y)
t_data_combined <- bind_rows(t_data_x, t_data_y)
ggplot(t_data_combined, aes(x = Age, y = `3P%`)) +
  geom_boxplot() + ylab("Three Point Percentage")
```

```
## Warning: Removed 6 rows containing non-finite values (stat_boxplot).
```

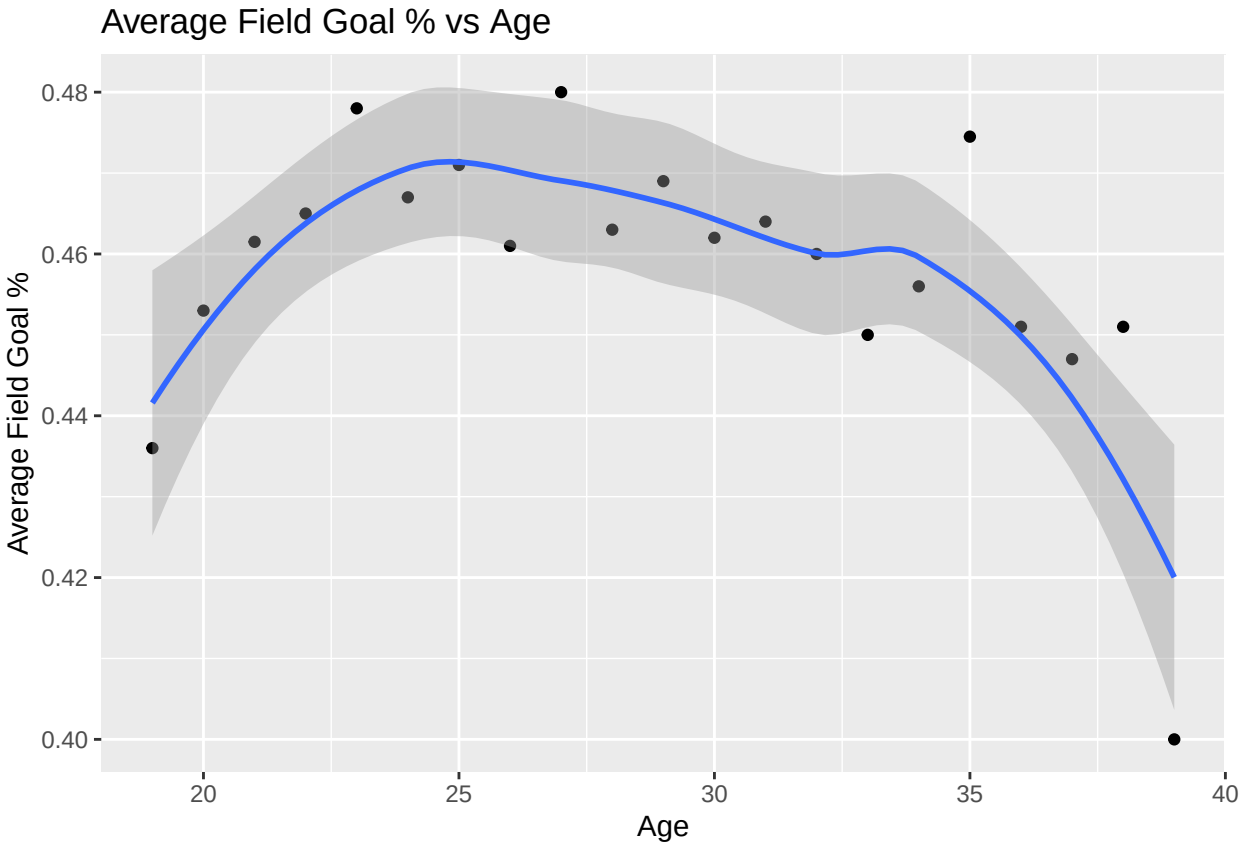



This result is statistically significant because we have a p-value of $< 0.05$. We were originally surprised that power forwards increased in their 3-point accuracy. However, we realized that the role of power forwards is similar to a center, but also works on shooting. Because shooting is not their primary focus as power forwards, when first drafted, it makes sense their 3-point accuracy would not be great. By adding a 3-pointer to their skill set over time, a player is more sought after by teams and likely earn a better contract.

**FG% Analysis**

Below, we plotted the relationship between Age and Field Goal Percentage (FG%) and once again saw a general peak around 25 years old, and then a gradual decline following that. Because this is a general trend that we have seen before, we decided to investigate further into how specific positions would vary in terms of average field goal percentage.

```
filtered_data %>% # field goal % (decreased)
  group_by(Age) %>%
  summarize(fg_perc_age=median(`FG%`,na.rm = TRUE))%>%
  ggplot(aes(x=Age, y=fg_perc_age)) + geom_point() +
  geom_smooth() + ylab("Average Field Goal %") +
  ggtitle("Average Field Goal % vs Age")
```

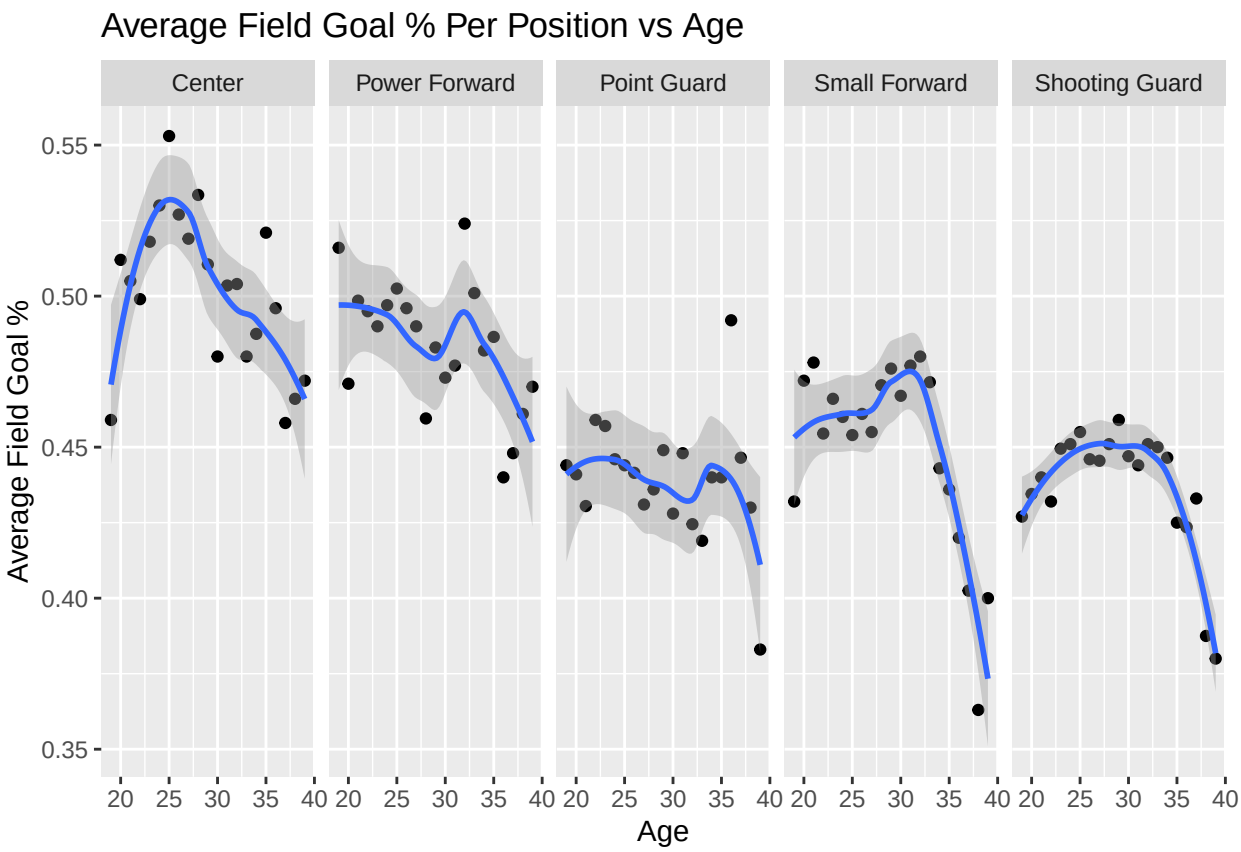



The graph below is the same data from the graph above, but is faceted by position.

```
filtered_data %>% # field goal % (decreased)
  filter(nchar(Pos) <= 2) %>%
  group_by(Age, Pos) %>%
  summarize(fg_perc_age=median(`FG%`,na.rm = TRUE))%>%
  ggplot(aes(x=Age, y=fg_perc_age)) + geom_point() +
  facet_grid(~Pos, labeller=Pos_labeller) +
  geom_smooth() + ylab("Average Field Goal %") +
  ggtitle("Average Field Goal % Per Position vs Age")
```

```
## Warning: The labeller API has been updated. Labellers taking ‘variable‘ and
```

```
## `value` arguments are now deprecated. See labellers documentation.
```




Here, the graph for Centers is the most interesting because they exhibit not only a much higher peak than the other positions, but they also seem to have a higher field goal percentage overall with over half of their data points residing higher than 0.50, while the other positions are almost entirely beneath that number. We can make sense of this by noting that Centers typically make a lot of their shots from within the 3-point line, while other positions tend to shoot from farther away, which is possibly why both Centers and Power Forwards have higher FG%: because they are shooting shots that are easier to make.

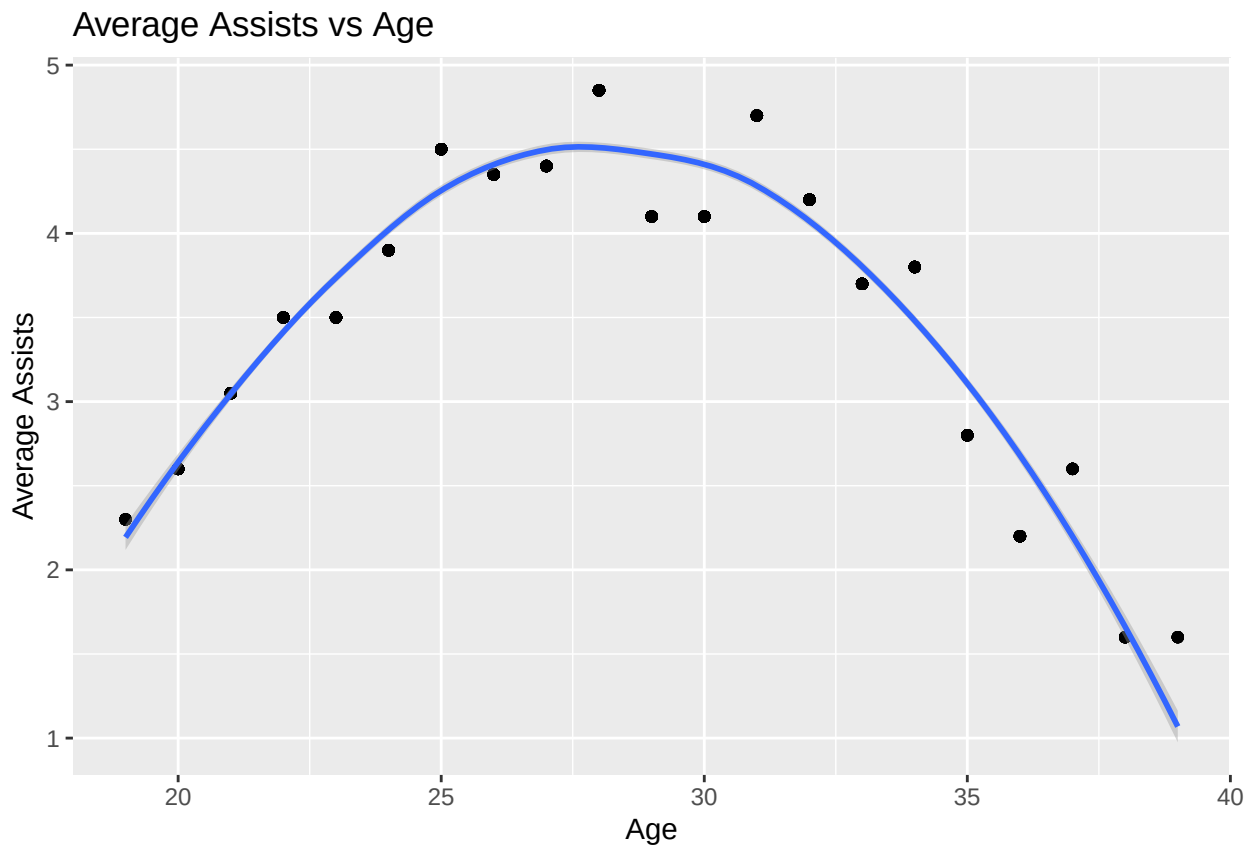
# Analysis of Assists

## Assists Over Time

Something that we were also curious about is whether the drop in performance in the player stats following their peaks was due to those players actually making less assists, or if their lower number of assists is due to the fact that coaches are giving them less time as they age, and consequently older players have less time to make assists. In order to investigate this, we decided to use a new response variable to test our inquiry by looking at Assists against Age as well as Assists per Minute against Age. These graphs are displayed below.

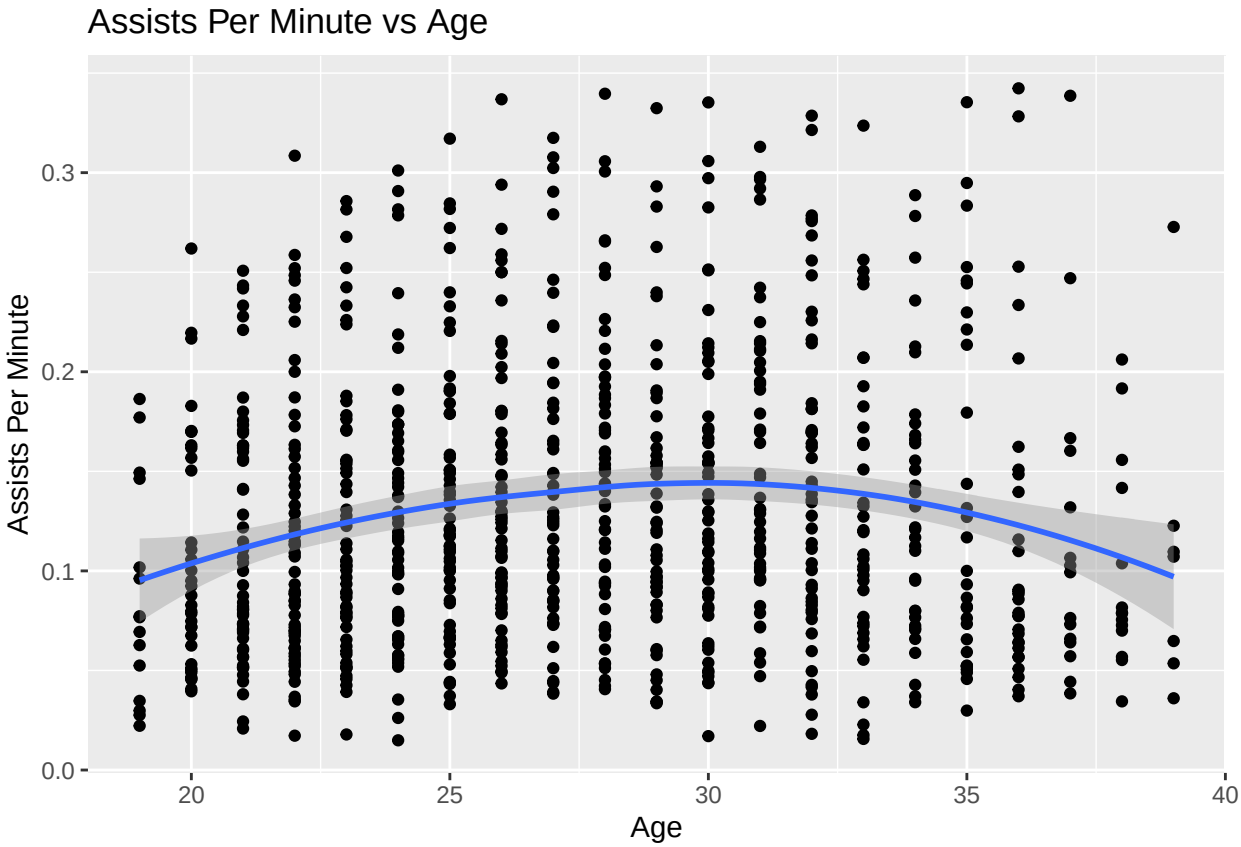
```
filtered_data %>% # Assists
  filter(nchar(Pos) <= 2) %>%
  group_by(Age) %>%
  mutate(avg_ast_age = median(AST, na.rm = TRUE)) %>%
  ggplot(aes(x = Age, y = avg_ast_age)) + geom_point() +
  geom_smooth() + ylab("Average Assists") + ggtitle("Average Assists vs Age")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
filtered_data %>%
  filter(nchar(Pos) <= 2) %>%
  mutate(ASTPerMin = AST/MP) %>%
  ggplot(aes(x = Age, y = ASTPerMin)) + geom_point() +
  geom_smooth() + ylab("Assists Per Minute") +
  ggtitle("Assists Per Minute vs Age")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




From these two graphs, we can see that although the rise to a peak and the ensuing decline in performance for Assists per Minute is not nearly as drastic as it is in the graph for solely Assists against Age, there is still a notable n-shaped trend. From this, we can conclude that although it may be the case that some of the decrease in assists following their peak can be attributed to less time allotted on the court, we cannot completely disprove that older players are simply getting worse at assists as they age. Because of this ambiguity, we chose to take a closer look at Assists per Minute and perform some statistical analysis to see how significant our findings are.

### Assists Per Minute

```
rise <- filtered_data %>%
    filter(Age <= 29) %>%
  mutate(ASTPerMin = AST/MP)

  model1 <- lm(ASTPerMin ~ Age, data = rise)
  summary(model1)
```

```
##
## Call:
## lm(formula = ASTPerMin ~ Age, data = rise)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.11282 -0.05098 -0.01348  0.04342  0.20281
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0271866  0.0242542   1.121    0.263
## Age         0.0041093  0.0009839   4.176 3.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06713 on 569 degrees of freedom
## Multiple R-squared:  0.02974,    Adjusted R-squared:  0.02804
## F-statistic: 17.44 on 1 and 569 DF,  p-value: 3.425e-05
```

```
  decline <- filtered_data %>%
    filter(Age >= 29) %>%
  mutate(ASTPerMin = AST/MP)

  model2 <- lm(ASTPerMin ~ Age, data = decline)
  summary(model2)
```

```
##
## Call:
## lm(formula = ASTPerMin ~ Age, data = decline)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12694 -0.05544 -0.01493  0.03988  0.21856
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.244857   0.046538   5.261 2.41e-07 ***
## Age         -0.003363   0.001425  -2.360   0.0188 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07507 on 375 degrees of freedom
## Multiple R-squared:  0.01463,    Adjusted R-squared:  0.012
## F-statistic: 5.567 on 1 and 375 DF,  p-value: 0.01881
```

Above, we called a linear regression on both the rise preceding and the decline following a player's peak in the frequency of assists per minute. There seems to be a 0.5% increase every year in assists per minute as they age to 28. Conversely, there is an approximate decrease of 0.4% each year in assists per minute following the age of 28. Most importantly, we found that both the increase and the decrease have p-values of $< 0.05$, meaning that an All-Star's performance in terms of assists per minute *does* in fact change as they age, and thus we can conclude that for AST, the decline in an older player's minutes can account for their worsening performance in this aspect of their game.

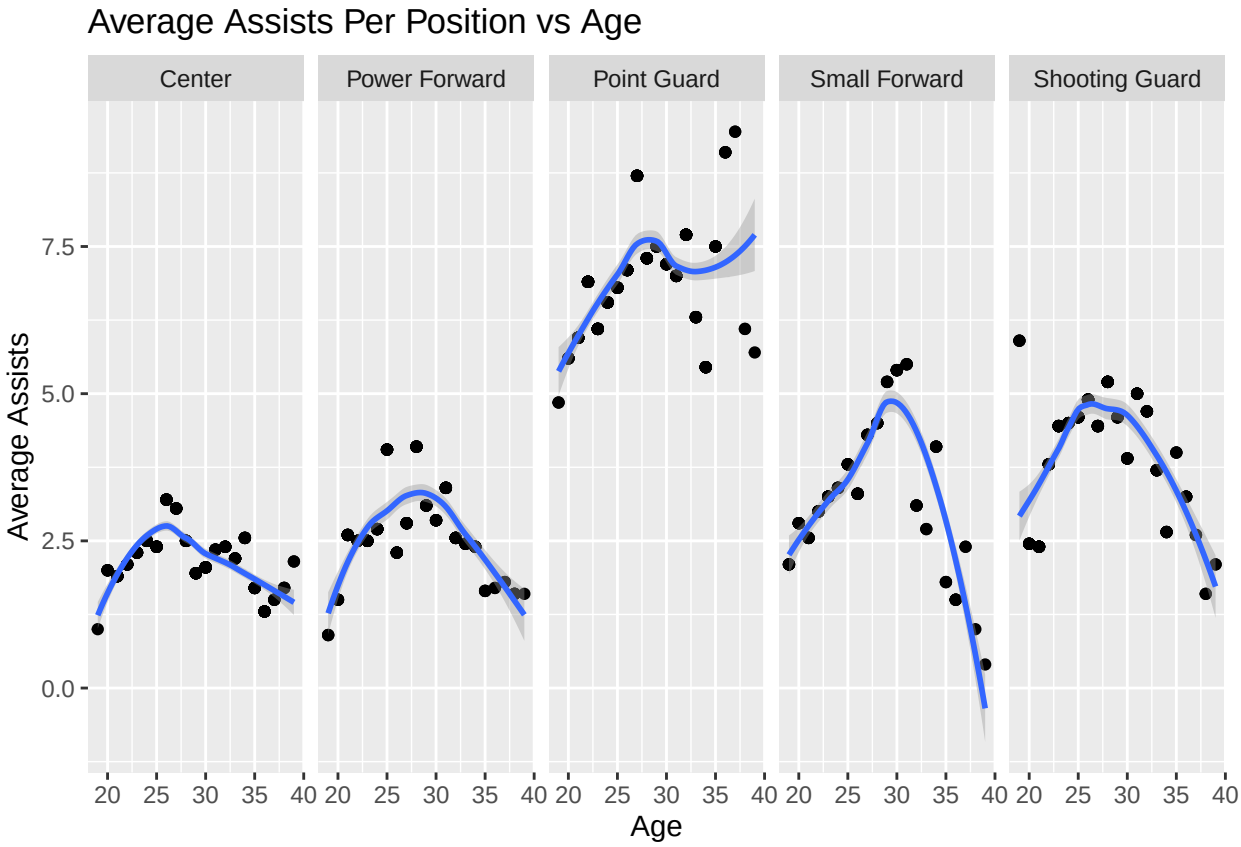
## Assists by Position

```
filtered_data %>% # Assists
  filter(nchar(Pos) <= 2) %>%
  group_by(Age, Pos) %>%
```

```
  mutate(avg_ast_age = median(AST, na.rm = TRUE)) %>%
  ggplot(aes(x = Age, y = avg_ast_age)) + geom_point() + facet_grid(~Pos, labeller=Pos_labeller) + geom_
  ggtitle("Average Assists Per Position vs Age")
```

```
## Warning: The labeller API has been updated. Labellers taking ‘variable‘ and
## ‘value‘ arguments are now deprecated. See labellers documentation.
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```




**Sample T-Test for Point Guard Assists**

Mean of assists at age <= 27 is u1 Mean of assists at age >= 30 is u2

Null hypothesis: u1 = u2 Alternative hypothesis: u1 < u2

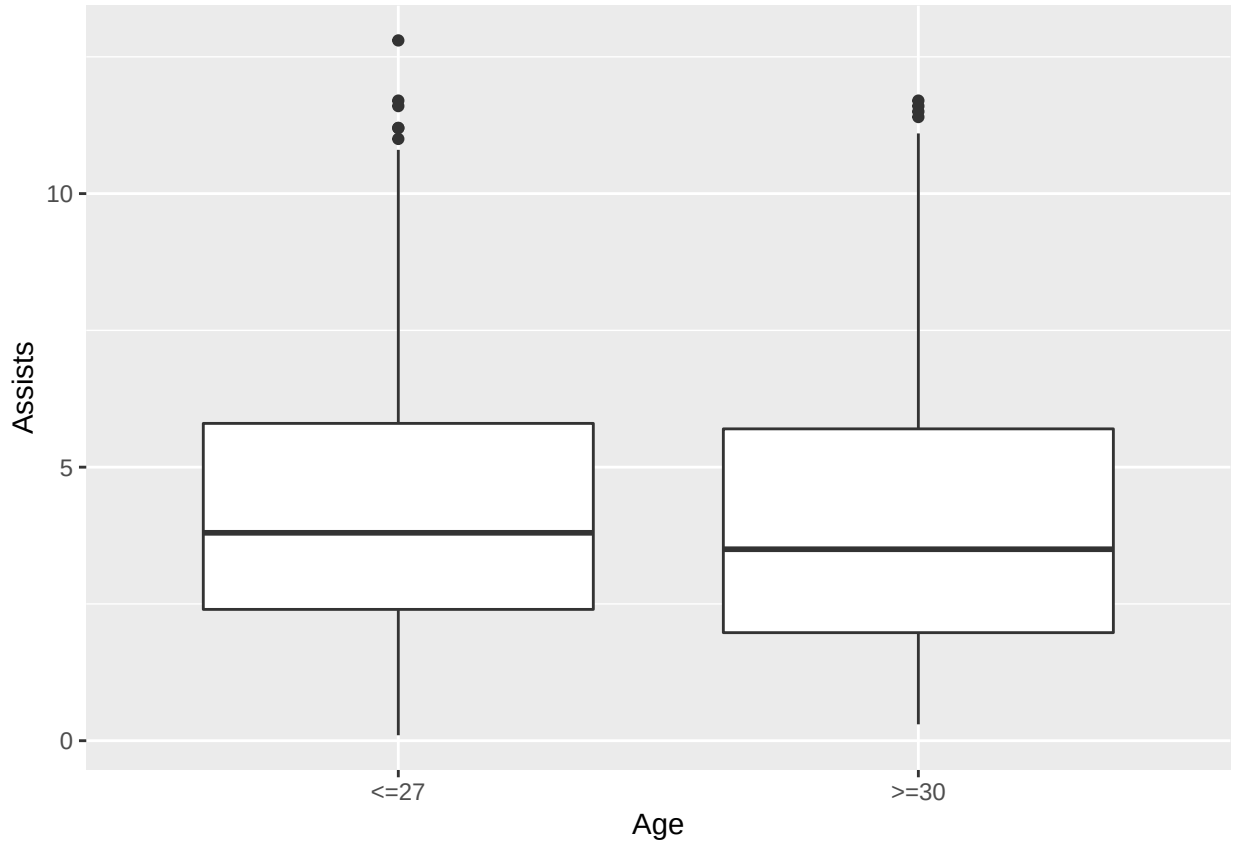
```
x <- filtered_data %>%
  filter(Age <= 27) %>%
  summarize(AST = AST)

y <- filtered_data %>%
  filter(Age >= 30) %>%
  summarize(AST = AST)

t.test(x, y, alternative = "less", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 0.67405, df = 650.9, p-value = 0.7497
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.4472849
## sample estimates:
## mean of x mean of y
##  4.323094  4.193210
```

Since the P-value is higher than 0.05, we fail to reject the null-hypothesis. The increase in assists shown by the linear regression graph is not statistically significant enough to show that as point guards age, they tend to get more assists. This can be explained by the decrease in playing time as they age. The less time they have on the field, the less chance they have to get assists.

Below is the visualization of the T-test:

```
t_data_x <- cbind(Age='<=27', x)
t_data_y <- cbind(Age='>=30', y)
t_data_combined <- bind_rows(t_data_x, t_data_y)
ggplot(t_data_combined, aes(x = Age, y = AST)) +
  geom_boxplot() + ylab("Assists")
```

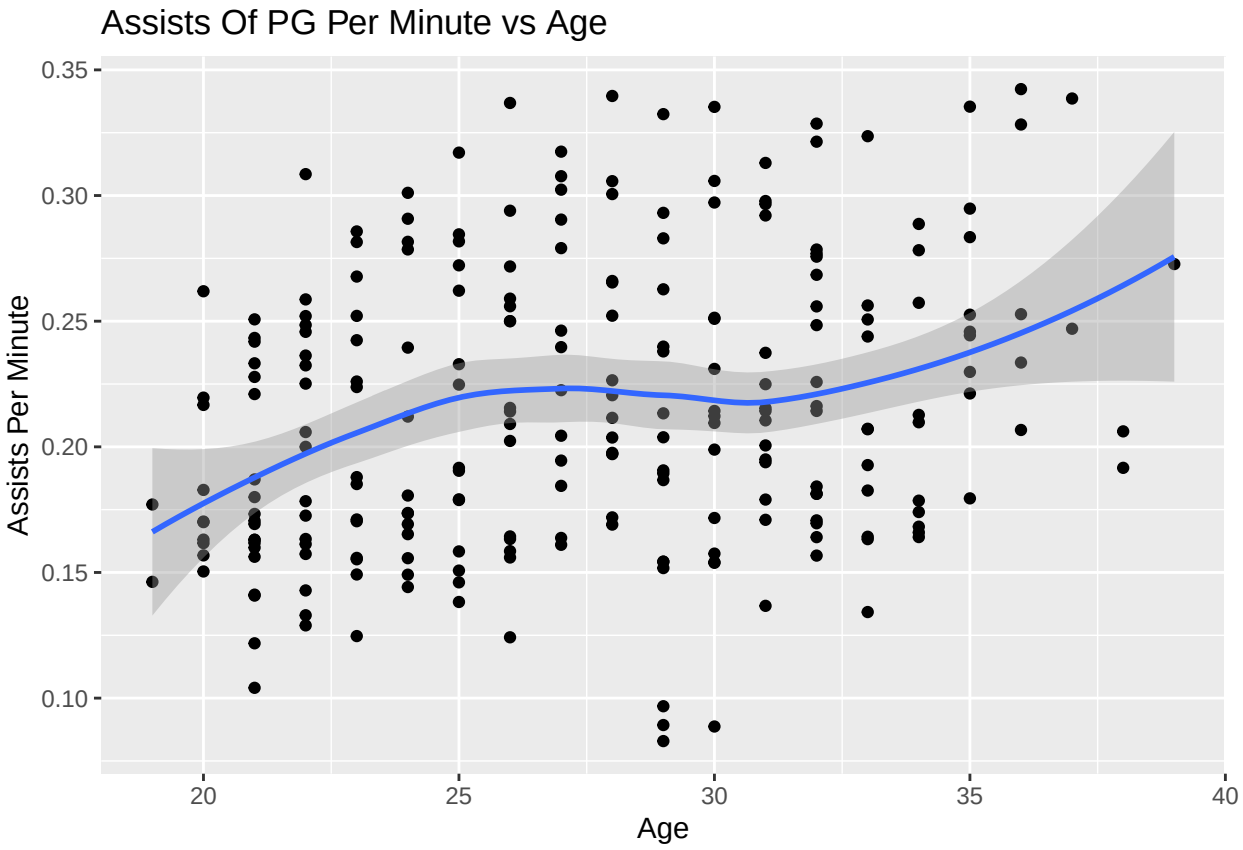



This got us wondering, even though the total number of assists for players decrease, does that mean their

frequency of assists decrease also? Their frequency could have increased but because overall they have less assists, we failed to see a trend on the test above.

```
filtered_data %>% # Assists per minute
  filter(Pos == "PG") %>%
  mutate(ass_per_min = AST / MP ) %>%
  ggplot(aes(x = Age, y = ass_per_min)) + geom_point() +
  geom_smooth() + ylab("Assists Per Minute") +
  ggtitle("Assists Of PG Per Minute vs Age")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




**Sample T test for point guard's frequency of assists overtime**

Mean of assists per minute at age <= 27 is u1 Mean of assists per minute at age >= 30 is u2

Null hypothesis: u1 = u2 Alternative hypothesis: u1 < u2

```
x <- filtered_data %>%
  filter(Age <= 27) %>%
  summarize(AST_per_min = AST / MP)

y <- filtered_data %>%
  filter(Age >= 30) %>%
```

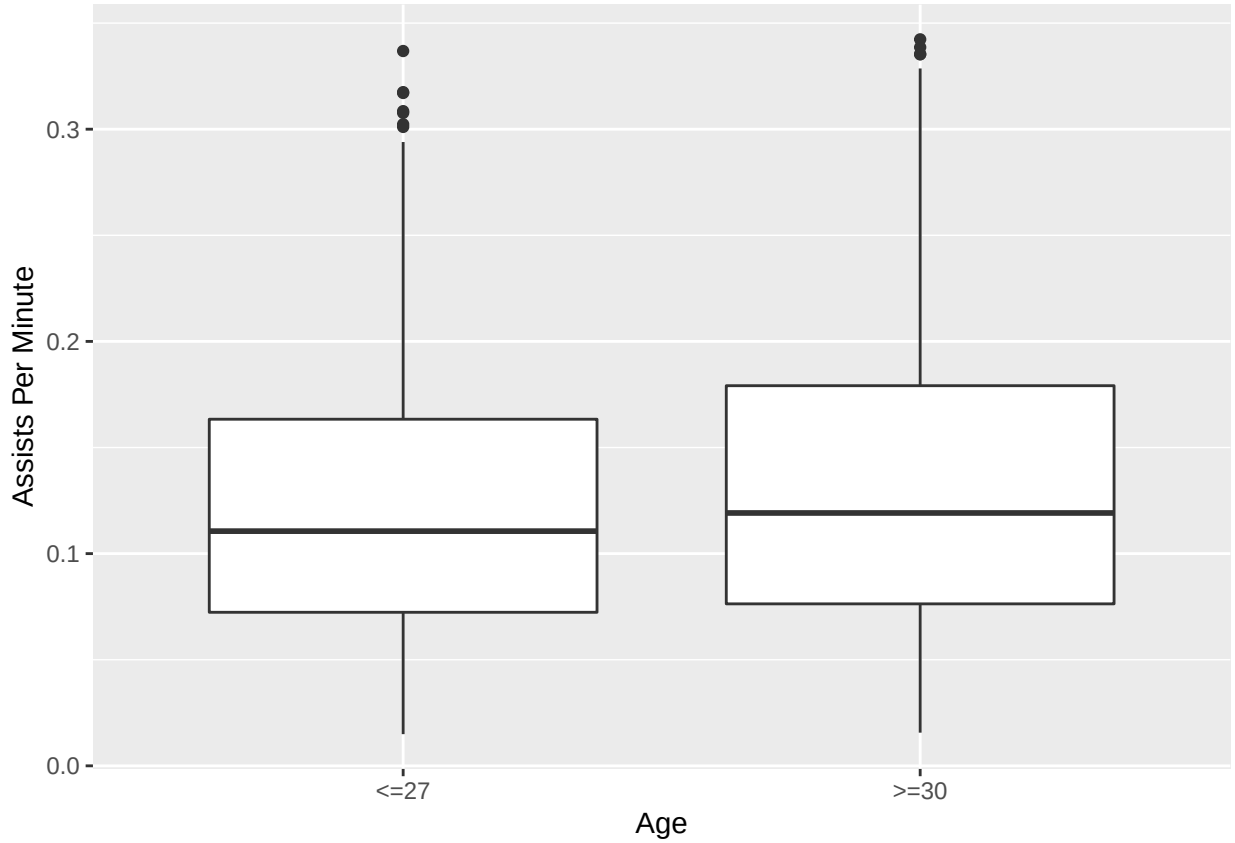
```
  summarize(AST_per_min = AST/ MP)

t.test(x, y, alternative = "less", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -2.0903, df = 636.96, p-value = 0.01849
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##          -Inf -0.002342244
## sample estimates:
## mean of x mean of y
## 0.1248041 0.1358542
```

Since the P-value is less than 0.05, we reject the null-hypothesis. The increase in the frequency of assists shown by the linear regression is statistically significant enough to show that as all-star point guards age, their frequency of assists increases.

Below is the visualization of the T-test:

```
t_data_x <- cbind(Age='<=27', x)
t_data_y <- cbind(Age='>=30', y)
t_data_combined <- bind_rows(t_data_x, t_data_y)
ggplot(t_data_combined, aes(x = Age, y = AST_per_min)) +
  geom_boxplot() + ylab("Assists Per Minute")
```



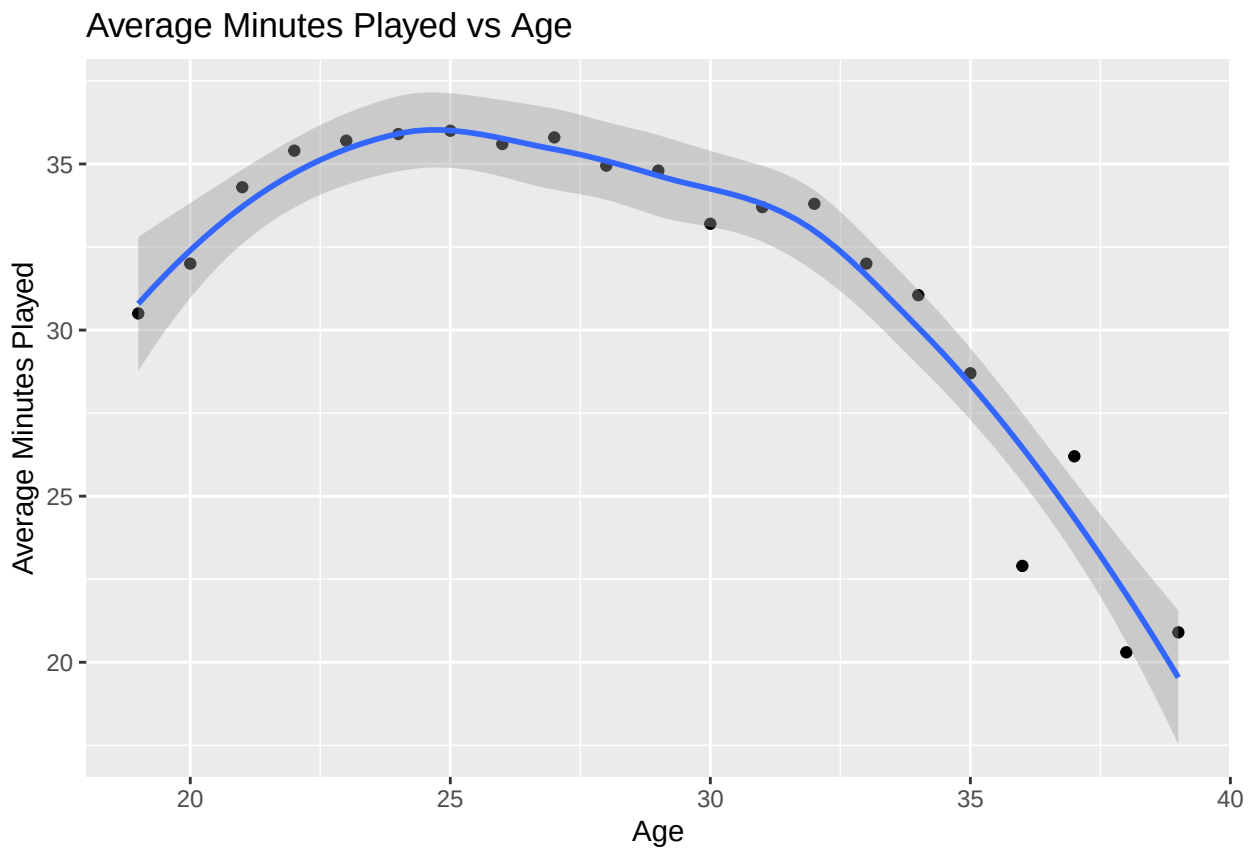

To summarize, we found that the upwards trend we observed for Assists for Point Guards was not statistically significant because the p-value was $> 0.05$, and so despite it making sense that PGs would have higher assist stats than other positions, our data could not support this claim. However, when we performed a t-test for PG assists per minute which still exhibited an upward trend, our p-value was $< 0.05$, meaning that our result here was statistically significant and that PG assist performance *do* change (improve) as they age.

Due to this, we hypothesize that it is not the Point Guards getting worse at assists, but that they are getting fewer minutes to get a high number of assists. This makes sense because assists is less about being physically fit and more about having a high basketball IQ. Point guards know how to read the court and create shots for teammates, and this would not get lost over age.

**Minutes**

```
filtered_data %>% # mins played
  filter(nchar(Pos) <= 2) %>%
  group_by(Age) %>%
  summarize(min_by_age=median(MP,na.rm = TRUE)) %>%
  ggplot(aes(x=Age, y=min_by_age)) + geom_point() +
  geom_smooth() + ylab("Average Minutes Played") +
  ggtitle("Average Minutes Played vs Age")
```

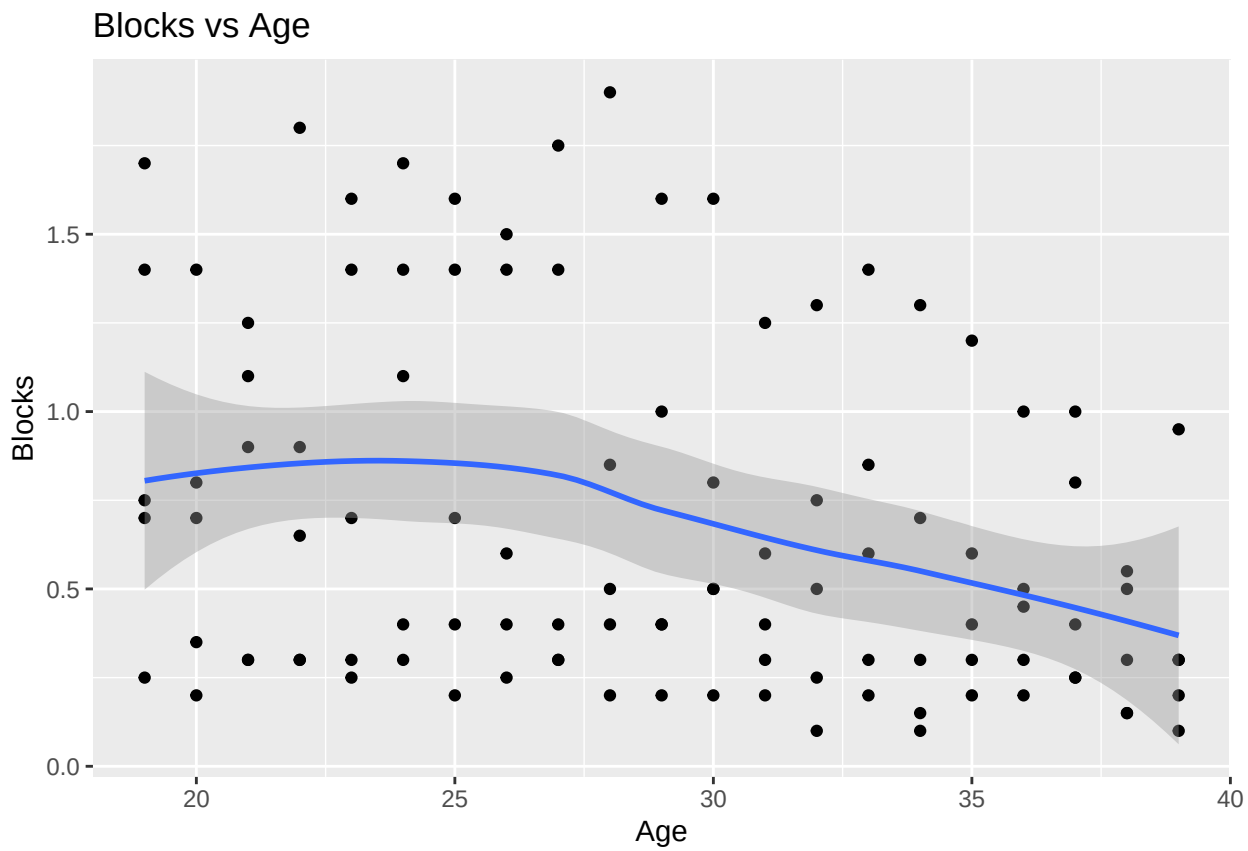
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



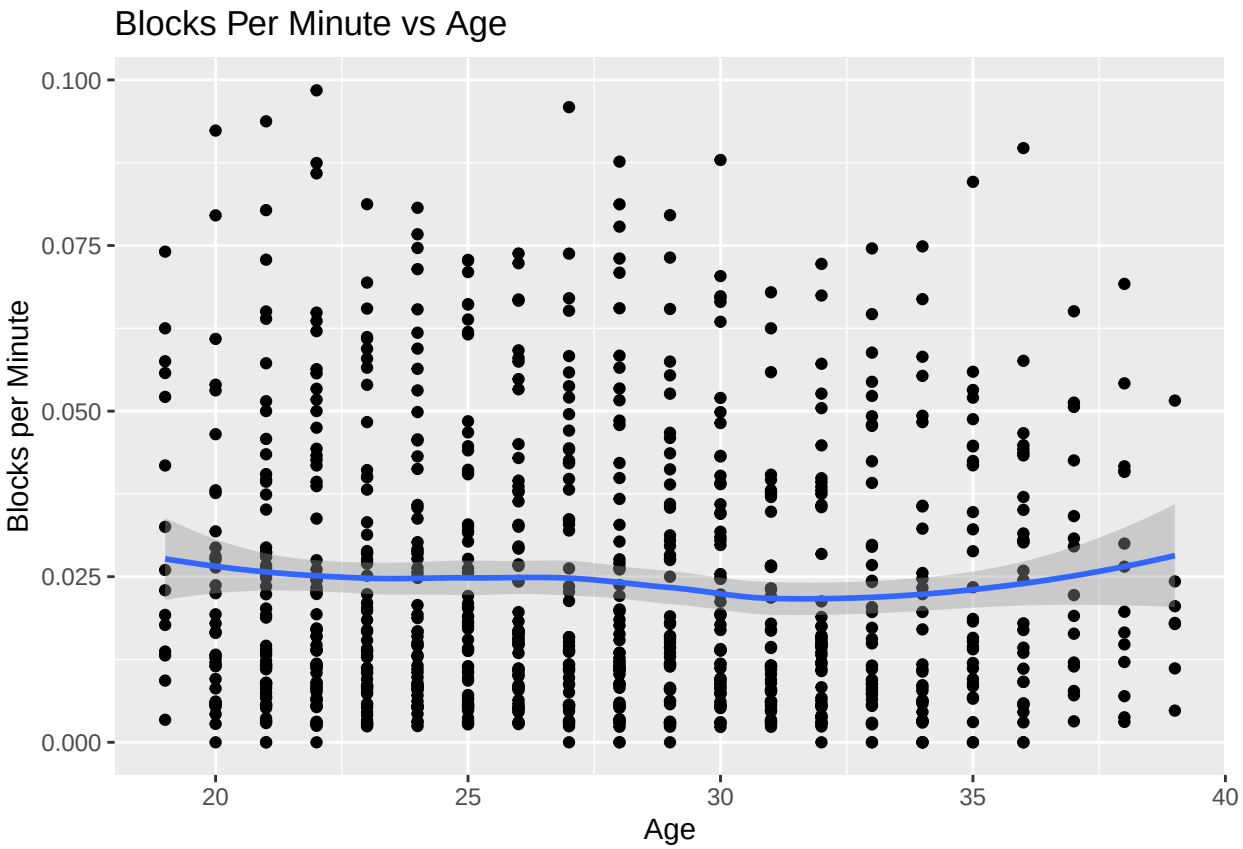

Looking at the trends from our other response variables, we can begin to conjecture that many NBA All-Stars reach their peak at somewhere between 25 and 30 years old. With this in mind, it would make sense that players who are at their peak performance are given more minutes, as their coaches want to make the most out of their time on the court. This is reflected in the mapping of Avg. Minutes against Age, which supports the claim that there is a point at which All-Stars play their best.

**Blocks**

```
filtered_data %>% # Blocks (Decrease)
  filter(nchar(Pos) <= 2) %>%
  group_by(Age, Pos) %>%
  summarize(block_by_age=median(`BLK`,na.rm = TRUE)) %>%
  ggplot(aes(x=Age, y=block_by_age)) + geom_point() +
  geom_smooth() + ylab("Blocks") + ggtitle("Blocks vs Age")
```



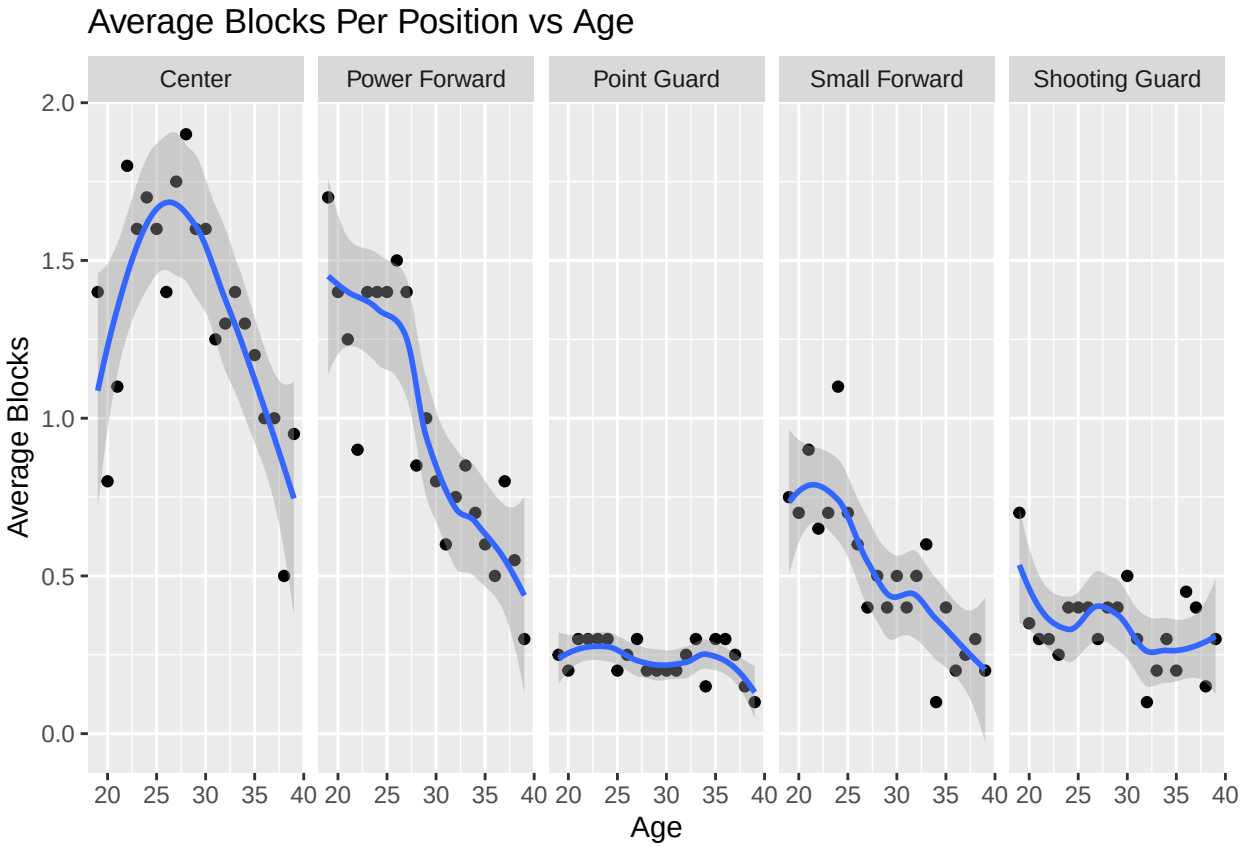

```
filtered_data %>%
  mutate(BLKPerMin = BLK/MP) %>%
  ggplot(aes(x = Age, y = BLKPerMin)) + geom_point() +
  geom_smooth() + ylab("Blocks per Minute") +
  ggtitle("Blocks Per Minute vs Age")
```




Overall, the total number of blocks decreases following the start of their career, which stands in contrast to other response variables that reach a peak *and then* decrease. In addition to looking BLKs, we also chose to analyze BLKs per minute, which indicates that although there is a slight decrease, the trend seems to show that blocks are not as impacted by age. Because of this, we chose to look at how the different positions were specifically impacted by this.

```
filtered_data %>% # Blocks
  filter(nchar(Pos) <= 2) %>%
  group_by(Age, Pos) %>%
  summarize(block_by_age=median(`BLK`,na.rm = TRUE)) %>%
  ggplot(aes(x=Age, y=block_by_age)) + geom_point() + facet_grid(~Pos, labeller=Pos_labeller) + geom_sm
  ggtitle("Average Blocks Per Position vs Age")
```

```
## Warning: The labeller API has been updated. Labellers taking ‘variable‘ and
## ‘value‘ arguments are now deprecated. See labellers documentation.
```



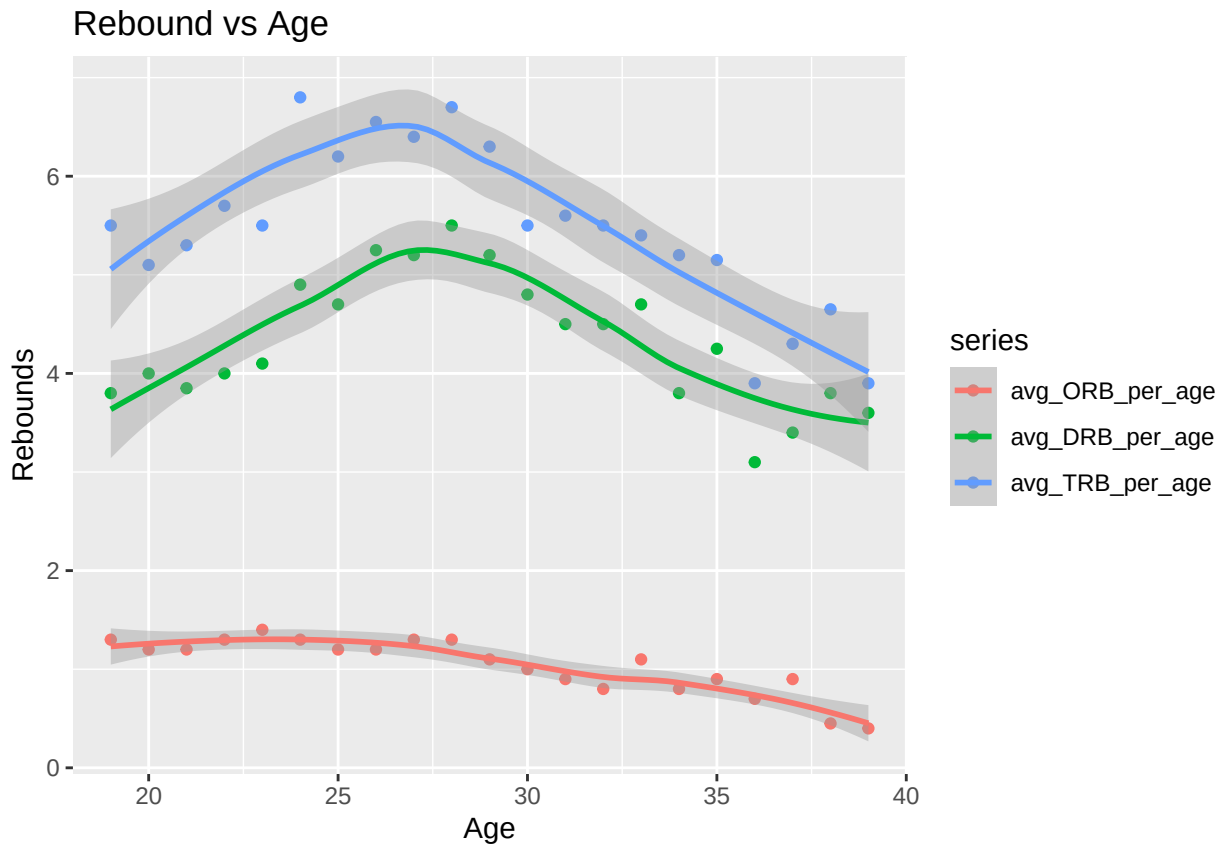

**Rebounds**

Below, we've mapped Offensive, Defensive, and total rebounds (Offensive + Defensive) on the same line graph against Age. What this graph shows us is that Total Rebounds (TRB) has the highest values, which makes sense as it is the sum of both ORB and DRB. What is most notable is that DRB are much more common than ORB, and this too makes sense in the context of the game as it is much more often the case that the defensive team is in a better position to regain the ball after it is shot since they are guarding the basket, while the offensive team usually begins to run back after shooting the ball.

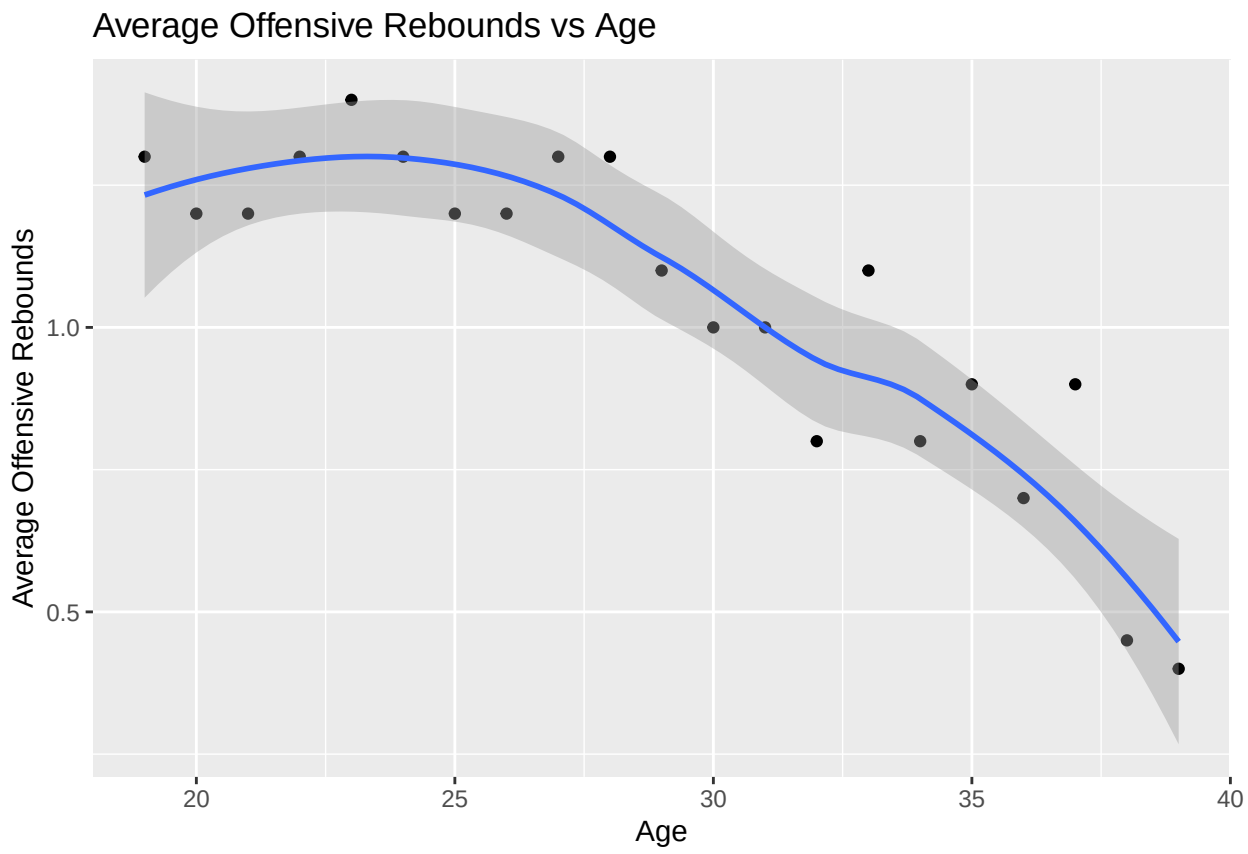
```
ORB <-filtered_data %>%
  select(Age, `ORB`) %>%
  group_by(Age) %>%
  summarize(avg_ORB_per_age = median(`ORB`))

DRB <- filtered_data %>%
  select(Age, `DRB`) %>%
  group_by(Age) %>%
  summarize(avg_DRB_per_age = median(`DRB`))

TRB <- filtered_data %>%
  select(Age, `TRB`) %>%
  group_by(Age) %>%
  summarize(avg_TRB_per_age = median(`TRB`))

RB <- bind_rows(ORB, DRB, TRB)
```

```
RB <- melt(RB ,  id.vars = 'Age', variable.name = 'series')

#create line plot for each column in data frame
ggplot(RB, aes(x = Age, y =value, color = series)) +
  geom_point() + geom_smooth() + ggtitle("Rebound vs Age") + ylab("Rebounds")
```

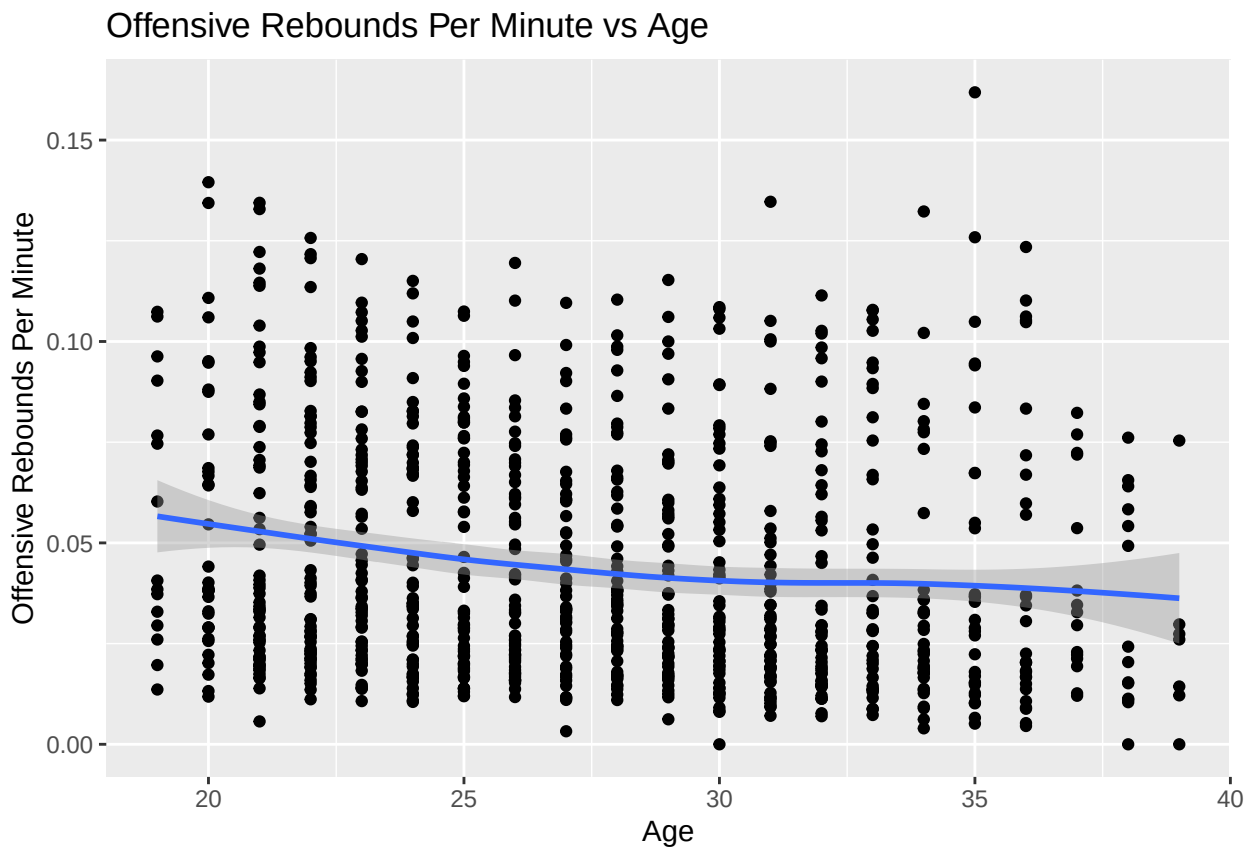
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

## Warning: Removed 126 rows containing non-finite values (stat_smooth).

## Warning: Removed 126 rows containing missing values (geom_point).
```




```
filtered_data %>% # Offensive rebounds
  filter(nchar(Pos) <= 2) %>%
  group_by(Age) %>%
  summarize(orb_by_age=median(`ORB`, na.rm = TRUE)) %>%
  ggplot(aes(x=Age, y=orb_by_age)) + geom_point()  + geom_smooth() +
  ylab("Average Offensive Rebounds") +
  ggtitle("Average Offensive Rebounds vs Age")
```




```
filtered_data %>% # Offensive rebounds
  mutate(ORBPerMin = ORB/MP) %>%
  ggplot(aes(x=Age, y=ORBPerMin)) + geom_point()  + geom_smooth() +
  ylab("Offensive Rebounds Per Minute") +
  ggtitle("Offensive Rebounds Per Minute vs Age")
```




Here, we've mapped average ORB against Age as well as average ORB per minute against Age in order to compare the decline we see in the total ORB vs. the frequency of ORB in order to determine if the amount of playing time a player is allotted has any impact on the decline in ORB as an All-Star gets older. As we can see, the total ORB has a much more obvious decline following a player's peak than the per minute graph, which exhibits a very gradual and small decline in contrast. Before making any conclusions, we chose to run a t-test to see if these findings were statistically significant.

**Sample T test for frequency of offensive rebounds per minute**

Mean of ORB per minute at age <= 27 is u1 Mean of ORB per minute at age >= 32 is u2

Null hypothesis: u1 = u2 Alternative hypothesis: u1 > u2

```
x <- filtered_data %>%
  filter(Age <= 27) %>%
  summarize(ORB_Per_Min = ORB/MP)

y <- filtered_data %>%
  filter(Age >= 32) %>%
  summarize(ORB_Per_Min = ORB/MP)

t.test(x, y, alternative = "greater", var.equal = FALSE)
```

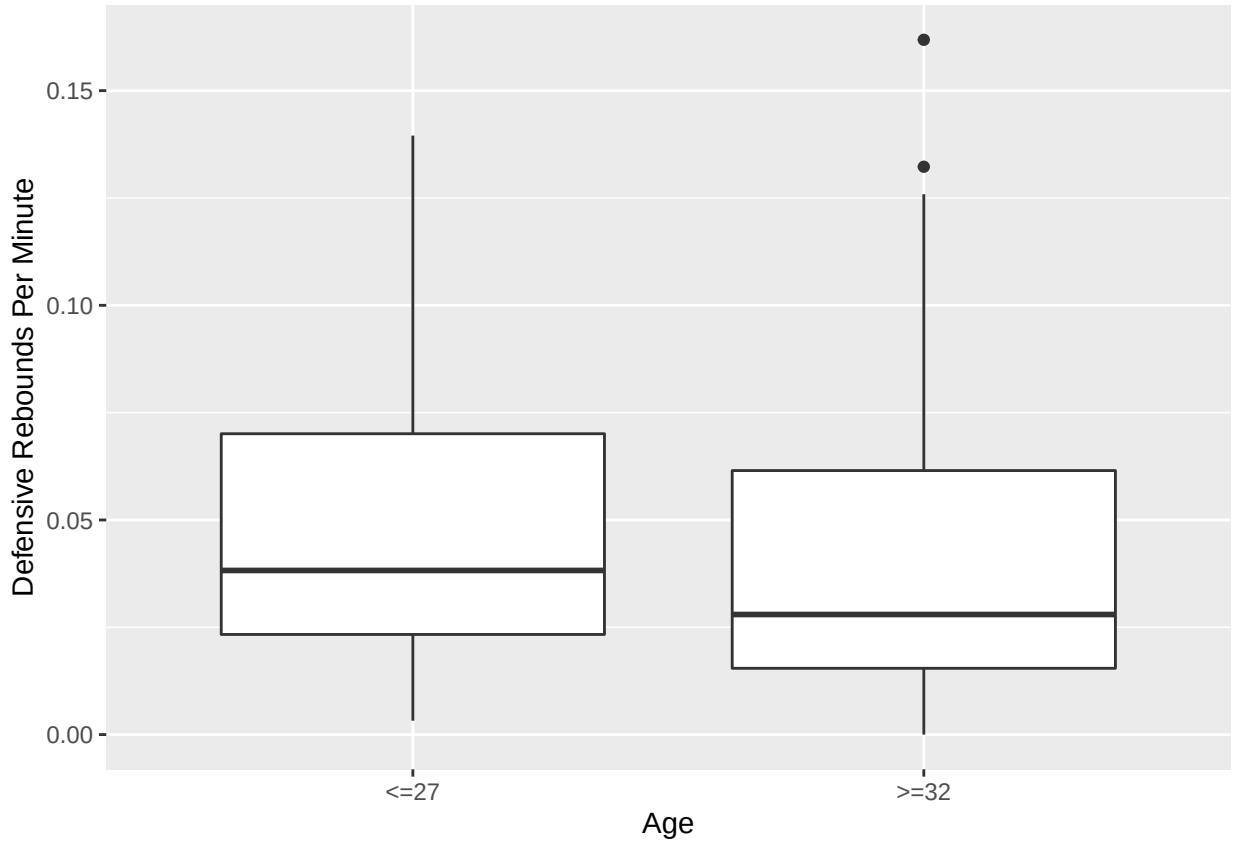
```
##
##  Welch Two Sample t-test
##
```

```
## data:  x and y
## t = 3.2508, df = 402.86, p-value = 0.0006238
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.004169607         Inf
## sample estimates:
##  mean of x  mean of y
## 0.04855035 0.04009011
```

Since the p-value is $< 0.05$, we are able to reject the null-hypothesis. We take this to mean that our results are statistically significant, and that we can claim that the performance of an All-Star in terms of average ORB per minute decreases as they age. This means that the small decline we see in average ORB per minute is indicative of All-Stars getting worse regardless of how many minutes they are allotted as they get older.

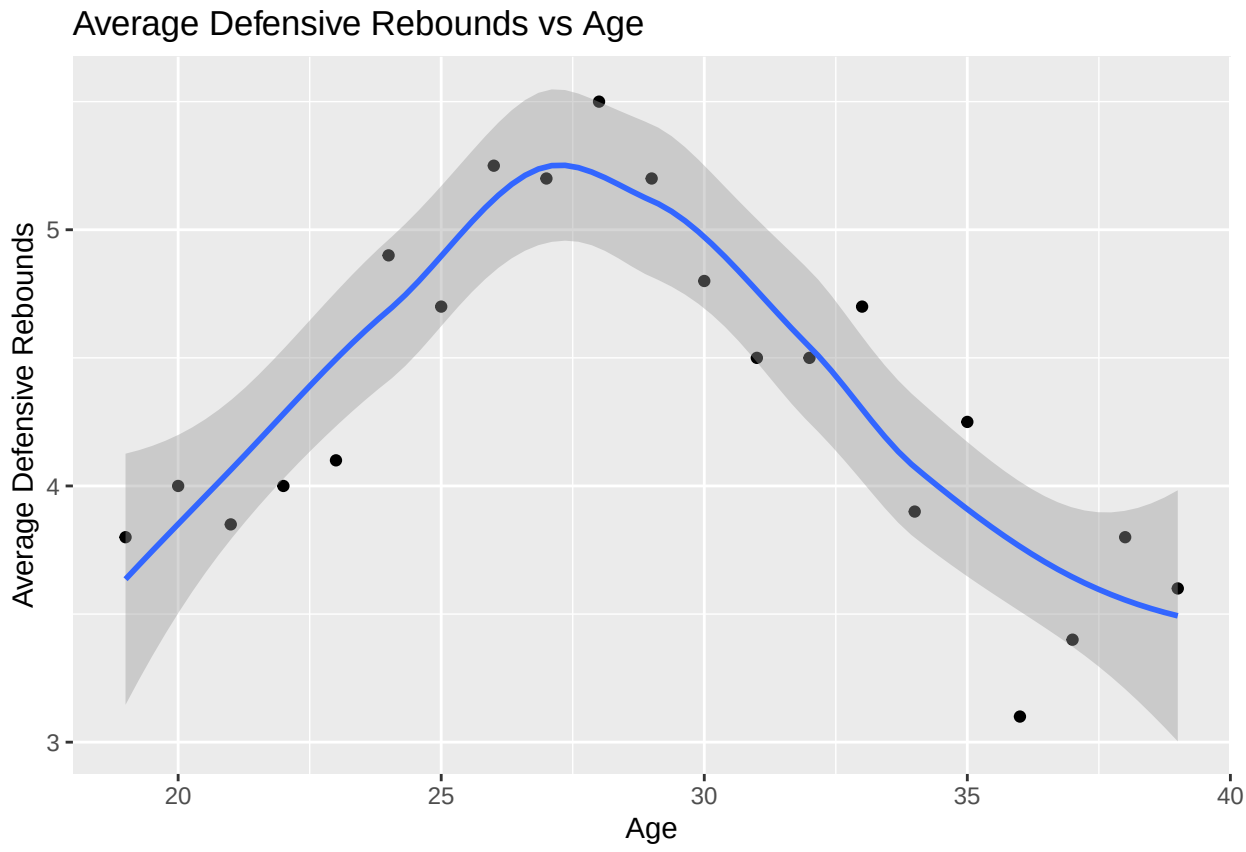
Below is the visualization of the T-test:

```
t_data_x <- cbind(Age='<=27', x)
t_data_y <- cbind(Age='>=32', y)
t_data_combined <- bind_rows(t_data_x, t_data_y)
ggplot(t_data_combined, aes(x = Age, y = ORB_Per_Min)) + geom_boxplot() +
  ylab("Defensive Rebounds Per Minute")
```

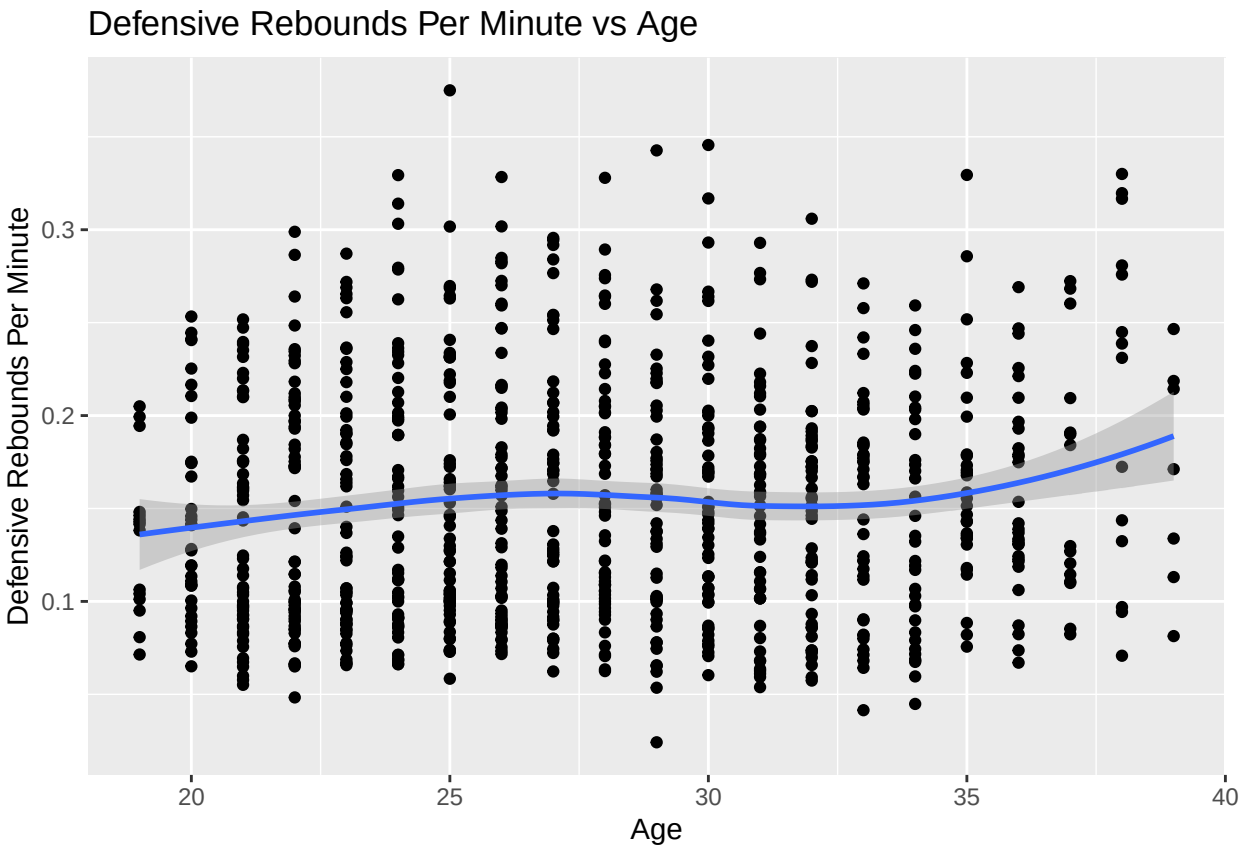



In these two graphs above, we first mapped average ORB against Age as well as average ORB per minute against Age. Similar to the trend that we saw with average AST and average AST per minute, the decline we see in average ORB against Age becomes noticeably less drastic when we look at the per minute stat.

```
filtered_data %>% # Defensive rebounds
  filter(nchar(Pos) <= 2) %>%
  group_by(Age) %>%
  summarize(drb_by_age=median(`DRB`,na.rm = TRUE)) %>%
  ggplot(aes(x=Age, y=drb_by_age)) + geom_point()  +
  geom_smooth() + ylab("Average Defensive Rebounds") +
  ggtitle("Average Defensive Rebounds vs Age")
```




```
filtered_data %>% # Defensive rebounds
  mutate(DRBPerMin = DRB/MP) %>%
  ggplot(aes(x=Age, y=DRBPerMin)) + geom_point()  + geom_smooth() +
  ylab("Defensive Rebounds Per Minute") +
  ggtitle("Defensive Rebounds Per Minute vs Age")
```




Although the total amounts of defensive rebounds certainly decreases with the age of the players, the per minute statistic which actually shows a small amount of improvement once again shows us that it could be less to do with an All-Star's performance actually degrading as they age, and more to do with being allotted less time as they get older, subsequently decreasing the amount of DRB they can get in their time. Below, we performed a t-test in order to see if these results were statistically significant.

**Sample T test for frequency of defensive rebounds per minute**

Mean of DRB per minute at age <= 27 is u1 Mean of DRB per minute at age >= 32 is u2

Null hypothesis: u1 = u2 Alternative hypothesis: u1 < u2

```
x <- filtered_data %>%
  filter(Age <= 27) %>%
  summarize(DRB_Per_Min = DRB/MP)

y <- filtered_data %>%
  filter(Age >= 32) %>%
  summarize(DRB_Per_Min = DRB/MP)

t.test(x, y, alternative = "less", var.equal = FALSE)
```

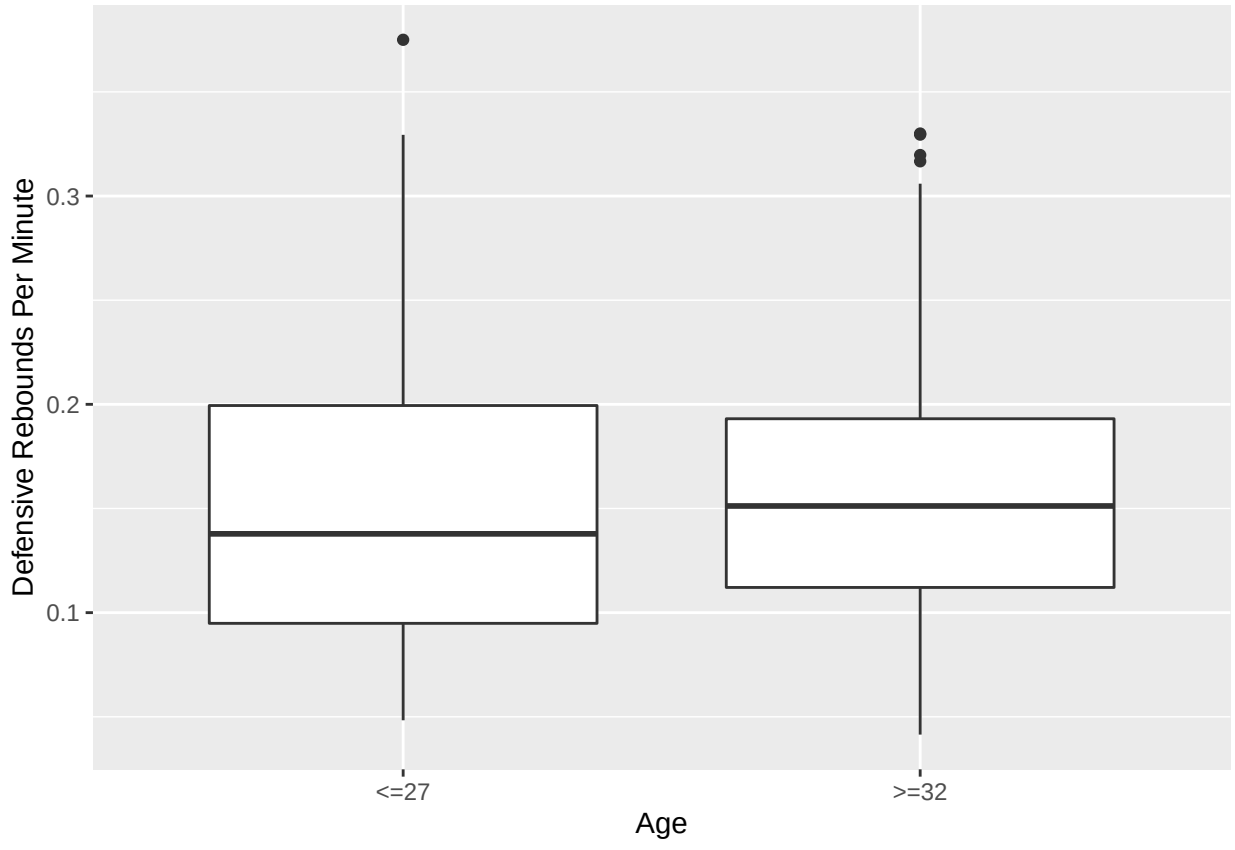
```
##
##  Welch Two Sample t-test
##
## data:  x and y
```

```
## t = -1.2122, df = 443.67, p-value = 0.113
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf 0.002274697
## sample estimates:
## mean of x mean of y
## 0.1501852 0.1565076
```

Since the p-value is $> 0.05$, we fail to reject the null-hypothesis. We take this to mean that our results are not statistically significant, and that the performance of an All-Star in terms of average DRB per minute stays the same as they age. This is actually not too far off from our claim that older players being given less time has more bearing on their DRB decreasing, thus yielding a trend that is much closer to a flat line which the null hypothesis supports.

Below is the visualization of the T-test:

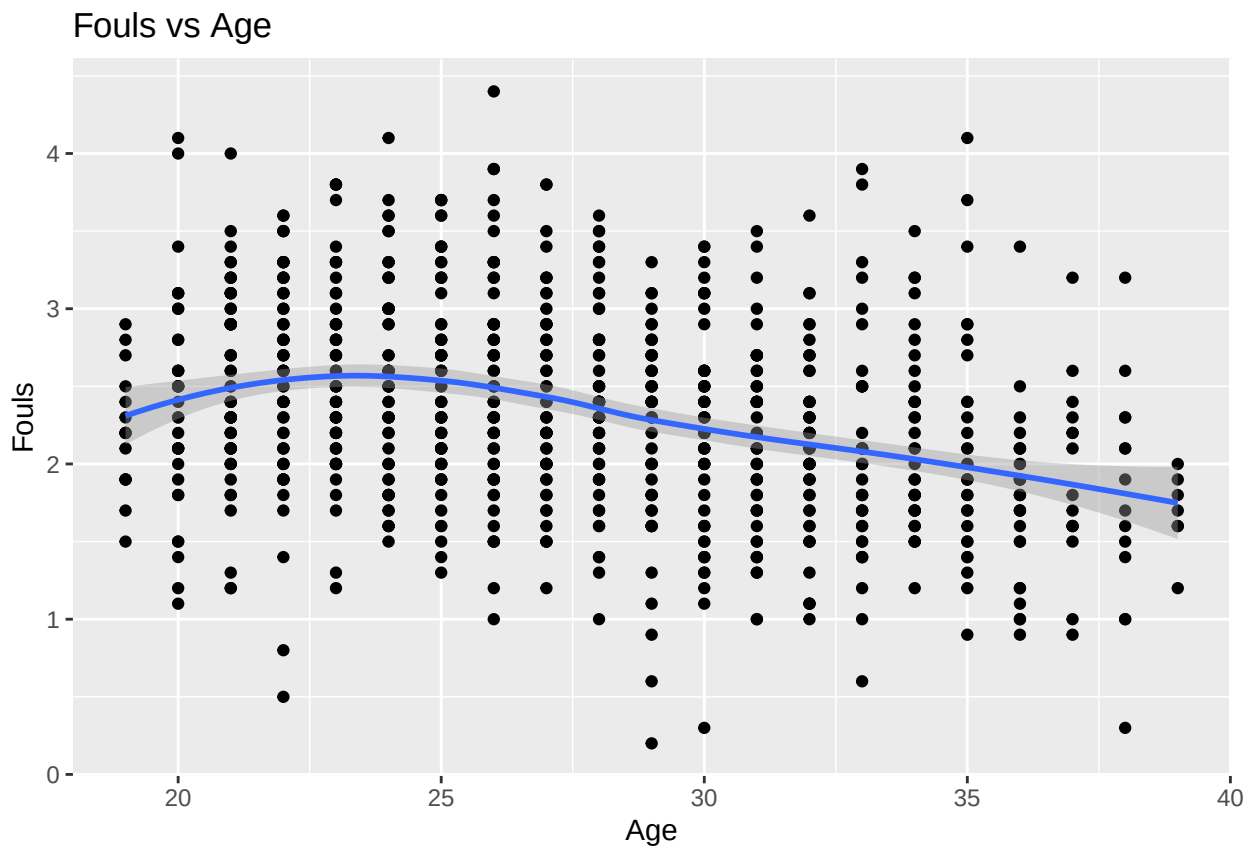
```
t_data_x <- cbind(Age='<=27', x)
t_data_y <- cbind(Age='>=32', y)
t_data_combined <- bind_rows(t_data_x, t_data_y)
ggplot(t_data_combined, aes(x = Age, y = DRB_Per_Min)) + geom_boxplot() +
  ylab("Defensive Rebounds Per Minute")
```




## Fouls

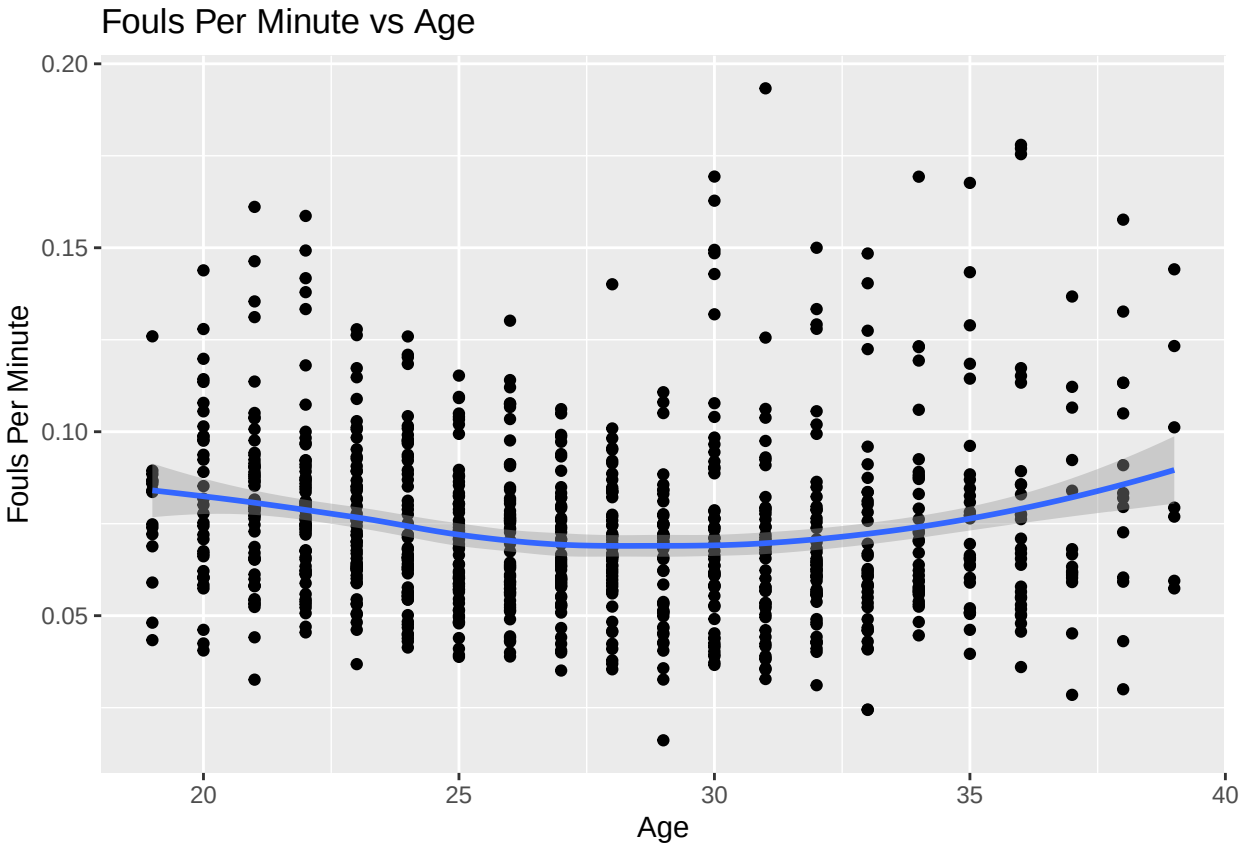
```
filtered_data %>%
  filter(nchar(Pos) <=2) %>%
  ggplot(aes(x = Age, y = PF)) + geom_point() + geom_smooth() + ylab("Fouls") + ggtitle("Fouls vs Age")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
filtered_data %>%
  filter(nchar(Pos) <= 2) %>%
  mutate(FoulsPerMin = PF/MP) %>%
  ggplot(aes(x = Age, y = FoulsPerMin)) + geom_point() + geom_smooth() +
  ylab("Fouls Per Minute") + ggtitle("Fouls Per Minute vs Age")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




From the Fouls vs Age graph above, we see that as all-stars age, their total number of fouls decrease. However, when we graphed out player's frequency of fouls (fouls per minute), we noticed that as all-stars got older, they actually tend to foul more frequently.

Do all-stars actually become more aggressive as they age? Is the trend listed on the regression line enough to show this claim?

We decided to delve in deeper.

## 2 sample T test for fouls and fouls per minute

### Number of Fouls

In order to investigate this further, we chose to perform some statistical tests in order to see if our findings above were statistically signficant.

Mean of fouls at age <= 27 is u1 Mean of fouls at age >= 30 is u2

Null hypothesis: u1 = u2 Alternative hypothesis: u1 > u2

```
x <- filtered_data %>%
  filter(Age <= 27) %>%
  summarize(PF = PF)

y <- filtered_data %>%
  filter(Age >= 30) %>%
  summarize(PF = PF)
```

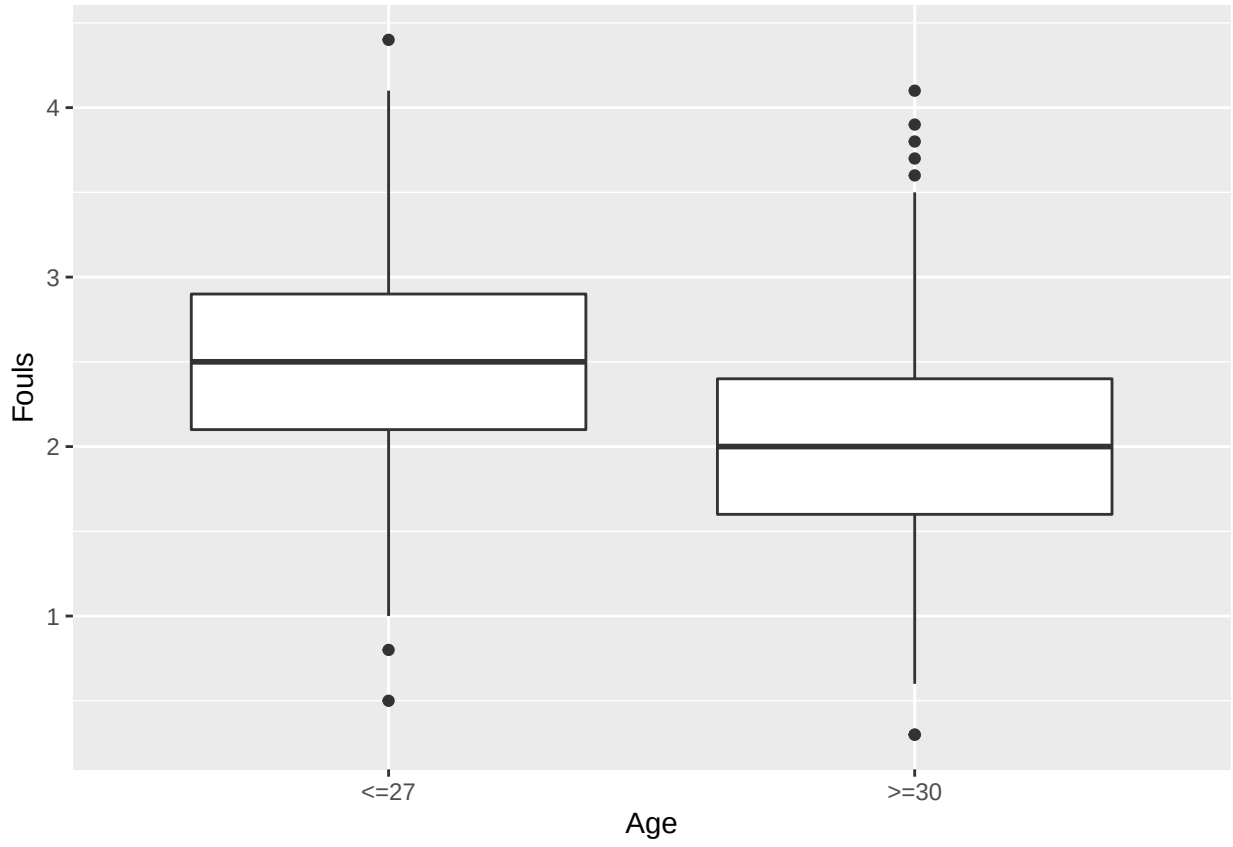
```
t.test(x, y, alternative = "greater", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 9.5839, df = 691.95, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.3621099       Inf
## sample estimates:
## mean of x mean of y
##  2.507625  2.070370
```

Since the p-value is less than 0.05, we are able to reject the null-hypothesis. As expected, All-Stars indeed commit less fouls as they age. This makes sense because as they age, they get less playing time which means they have less time to commit fouls.

Below is the visualization of the T-test:

```
t_data_x <- cbind(Age='<=27', x)
t_data_y <- cbind(Age='>=30', y)
t_data_combined <- bind_rows(t_data_x, t_data_y)
ggplot(t_data_combined, aes(x = Age, y = PF)) + geom_boxplot() + ylab("Fouls")
```



### Frequency of Fouls

Below, we performed the same statistical analysis as above except on the average number of fouls per minute, as opposed to total fouls. In performing this test, we are attempting to test the statistical significance of our earlier finding which showed that All-Stars actually increase the frequency of their fouls as they get older.

Mean of average foul per minute at age <= 27 is u1 Mean of average foul per minute at age >= 30 is u2

Null hypothesis: u1 = u2 Alternative hypothesis: u1 < u2

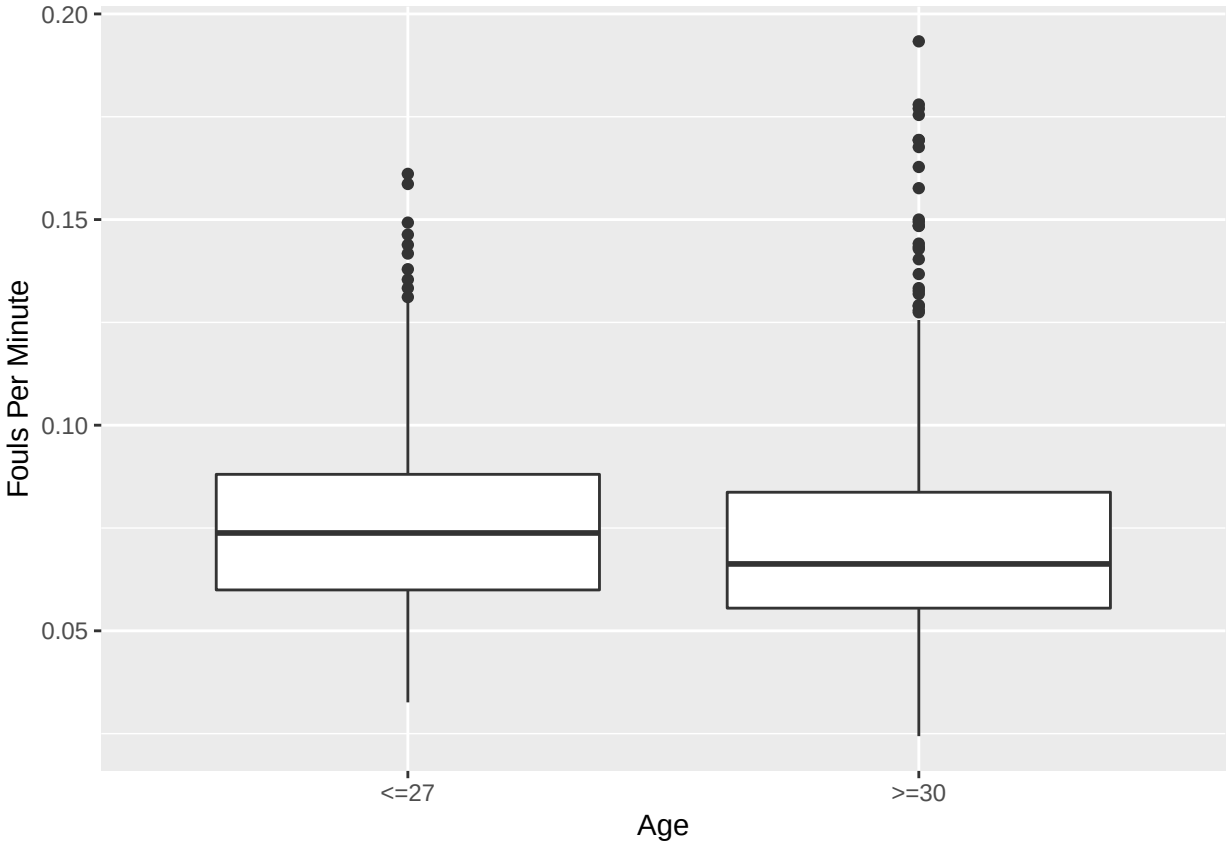
```
x <- filtered_data %>% #players younger than or equal to 25
  filter(Age <= 27) %>%
  summarize(FoulsPerMin = PF/MP)

y <- filtered_data %>% #players older than or equal to 35
  filter(Age >= 30) %>%
  summarize(FoulsPerMin = PF/MP)

t.test(x, y, alternative = "less", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 0.92945, df = 553.93, p-value = 0.8235
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf 0.005044993
## sample estimates:
##  mean of x  mean of y
## 0.07570714 0.07388760
```

Since the p-value is greater than 0.05, we fail to reject the null-hypothesis. It is not statistically significant enough to support our claim that All-Stars tend to foul more frequently as they age.

Below is the visualization of the T-test:

```
t_data_x <- cbind(Age='<=27', x)
t_data_y <- cbind(Age='>=30', y)
t_data_combined <- bind_rows(t_data_x, t_data_y)
ggplot(t_data_combined, aes(x = Age, y = FoulsPerMin)) + geom_boxplot() +
  ylab("Fouls Per Minute")
```

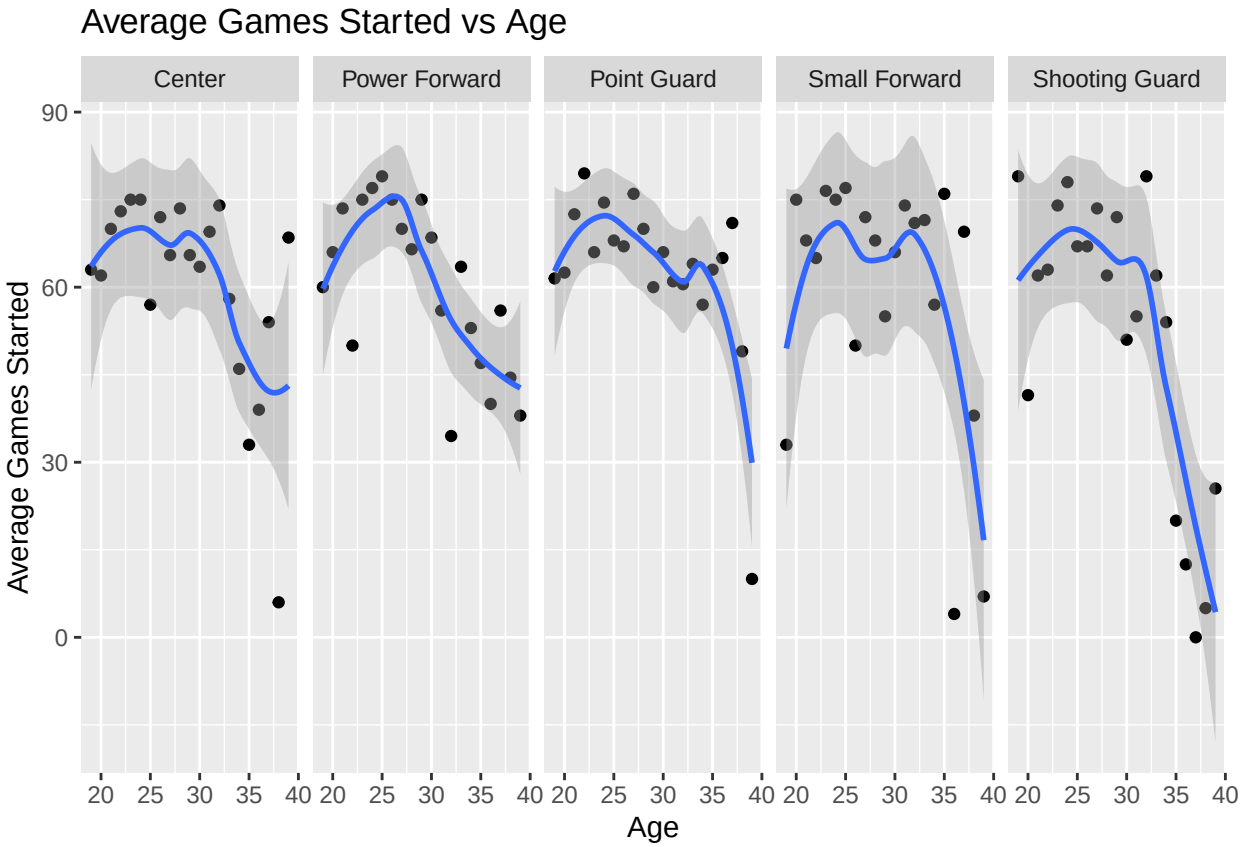



A failure in proving that there is a trend of All-Stars fouling more frequently as they age could be due to the fact that our sample is not big enough. For example. the amount of observations we have for players older than 35 is 90 observations. However, for players younger than 25, we have 369 observations. Since we only have 90 observations for players older than 35, we could say that our data isn't representative of the entire All-Star population, particularly for those who are older. Consequently, we will defer to our statistically significant finding, which is that All-Stars commit less fouls in total as they get older, indicating that All-Stars become less aggressive as age takes a physical toll on their bodies.

**Games Started**

```
filtered_data %>% # games started per season (Decrease)
  filter(nchar(Pos) <= 2) %>%
  group_by(Age, Pos) %>%
  summarize(gs_by_age=median(`GS`,na.rm = TRUE)) %>%
  ggplot(aes(x=Age, y=gs_by_age)) + geom_point() +
  facet_grid(~Pos, labeller=Pos_labeller) + geom_smooth() +
  ylab("Average Games Started") + ggtitle("Average Games Started vs Age")
```

```
## Warning: The labeller API has been updated. Labellers taking ‘variable‘ and
## ‘value‘ arguments are now deprecated. See labellers documentation.
```




As expected, there is a decrease in the number of games that players are starters from the beginning to the end of one's career. From the beginning, there is an initial increase which makes sense because as players get better performance-wise, as shown by our previous graphs on field goals and others, coaches will depend on them to start the games. As they get older and their overall performance declines, they will start fewer games and gain a role on their team where they likely support up and coming players who are now leading the team.

### 4.3 - Possible Improvements

One possible improvement that we an immediately observe in our research project is that we would have liked to take the time to better optimize how we scraped and cleaned up our data. Many of our processes were manual, particularly for the Player Stats as well as for our removal of players that played multiple positions. In the future, we would hope to take advantage of functions or more code to make this process more efficient and reproducable. Furthermore, we ended up with more than a few statistically insignificant findings because many p-values were $> 0.05$. This is likely because of the nature of our dataset, as we chose to focus on a small subset of NBA players, the All-Stars, which limits the amount of data points that we can have in order to create a holistic and statistically significant understanding of our research question. In the future, we may remedy this by not only looking at All-Stars and instead all players in the NBA, but limited to a smaller range of years so as to not oversaturate our analysis. In addition to this, in the future we would hope to look into more relationships between the variables that are a part of a basketball game, including more defensive statistics such as blocks and steals in order to form a more rounded view of an All-Star's "performance".

## Section 5 - Conclusion

Overall, we expected NBA All-Stars to decline in several aspects of their performance as they age, as we know that not even the top players of the NBA can be immune to time's impact on physical strength. From our exploratory data analysis, we saw many graphs mapping our response variables against age that substantiated this belief, in addition to statistical tests that we ran in order to determine the statistical significance of the results that we saw. These analyses varied in terms of their adherence to our expectations, with some variables such as average PTS having more of a drastic decline than average AST. Although our expectations and findings do show an overall decline in various aspects of performance as an All-Star ages, it is important for us to note that we cannot expect for these declines to be so drastic to the extent that they can no longer viably compete in the NBA; after all, these are supposed to be the best of the best in an already competitive league. Some of our most surprising findings came from our analysis on how specific positions would affect our response variables. For example, it was interesting to see that Power Forwards improved in their three-point percentage as they got older, particularly because they are not nearly as relied upon for shooting long-distance as other offensive positions. For situations such as this, however, we showed that context was integral in making sense of some of our more surprising results. Additionally, our investigation of how the decline in an older player's amount of playing time may be a contributing factor to some of the aspects of their performance, such as AST or rebounds, showed that we cannot attribute *all* of the decline in the player statistics to an All-Star simply degrading in performance.