Aging Curve of Strikers in the EPL

Junwoo Kim

2023-12-11

Introduction

I am a huge fan of football. Throughout the years of being a football fan, I've seen countless players reach their primes and fall which was very interesting. So I am using this final project as an opportunity to study the aging curve of professional football players by using functions and data visualization that I've learned in Stat 184.

Brief Description of Striker(Forward) Position.

The striker(Forward) is one of the most important football positions there is. Generally the most advanced players on the pitch, their primary job is to score goals and assist other attacking players around them. The most exciting position in my opinion.

Why Striker(Forward) position?

To analyze the aging curve, I had to go with the Striker position since they are the ones that clash the most with the opponent's defense. Winger was definitely another viable option but it had so many variables to consider which could be misleading.

Why the English Premier League?

The EPL is the highest level of the English Football league system. It is the most watched sports league in the world. Moreover, EPL generated the most revenue among all football leagues. Therefore, the best players in the world tend to join the EPL.

Elements to consider when analyzing the Aging Curve.

I have gathered data sets of 10 legendary players that played in the EPL. I have made a Google sheet by myself that contains the Age(From 19-33) and corresponding player stats of Goals, Assists, and Shot Volume. Then I will come up with the result of 'When does the players tend to escalate?', 'When does the players tend to reach their primes?', and last but not least "When does the players start to descend from their primes?'.

Implementing Google Sheet to RStudio

```
library(googlesheets4)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
gs4_auth()
```

```
## ! Using an auto-discovered, cached token.
##   To suppress this message, modify your code or options to clearly consent to
##   the use of a cached token.
##   See gargle's "Non-interactive auth" vignette for more details:
##   <https://gargle.r-lib.org/articles/non-interactive-auth.html>
## i The googlesheets4 package is using a cached token for 'jjk6181@psu.edu'.
```

```
sheet_url <- "https://docs.google.com/spreadsheets/d/1AC1jvmwmiC1Ro4ToWubHtrndgT1Thk96VTXIsPGUly4/edit#gid=0"
My_data <- read_sheet(sheet_url)
```

```
## v Reading from "mydata".
## v Range ' 1'.
```

```
View(My_data)
```

Initial attempt of Visualizing My_data

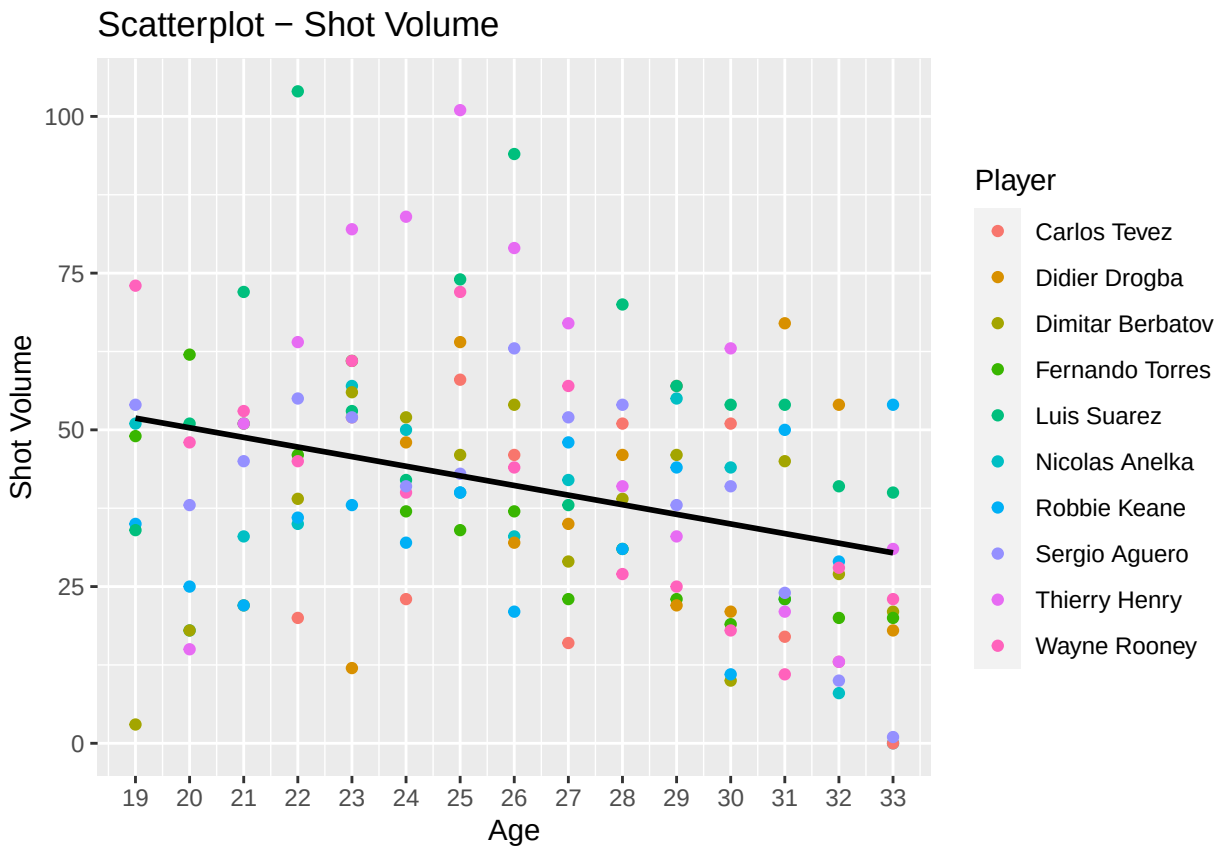
I initially planned to make scatter plots for Age* Each Elements. However the results of the scatter plot were extremely unclear even with Trend lines and polished. Here is the initial code and the result of my visualization

Initial Scatter plots of Age*Shot_Volume

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

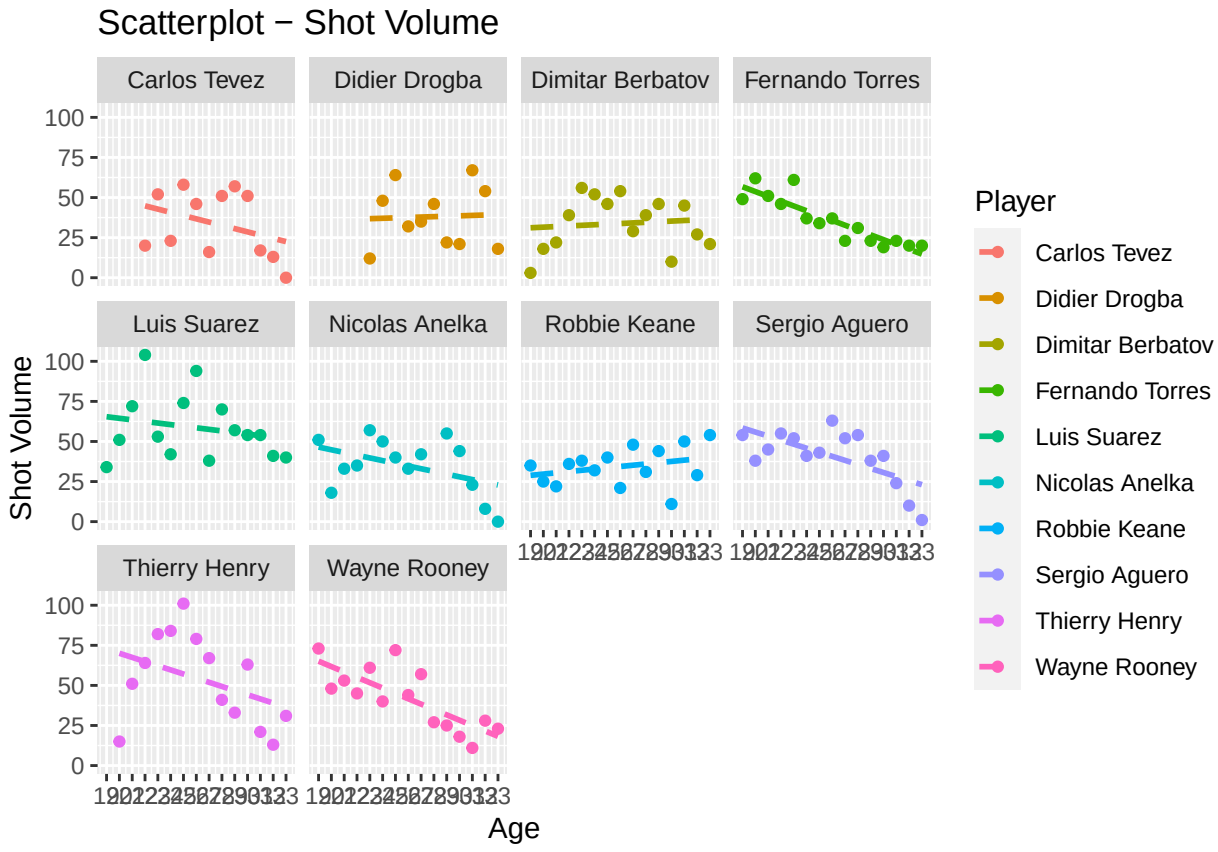
```
## Warning: Removed 9 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 9 rows containing missing values (`geom_point()`).
```




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

## Warning: Removed 9 rows containing non-finite values (`stat_smooth()`).
## Removed 9 rows containing missing values (`geom_point()`).
```

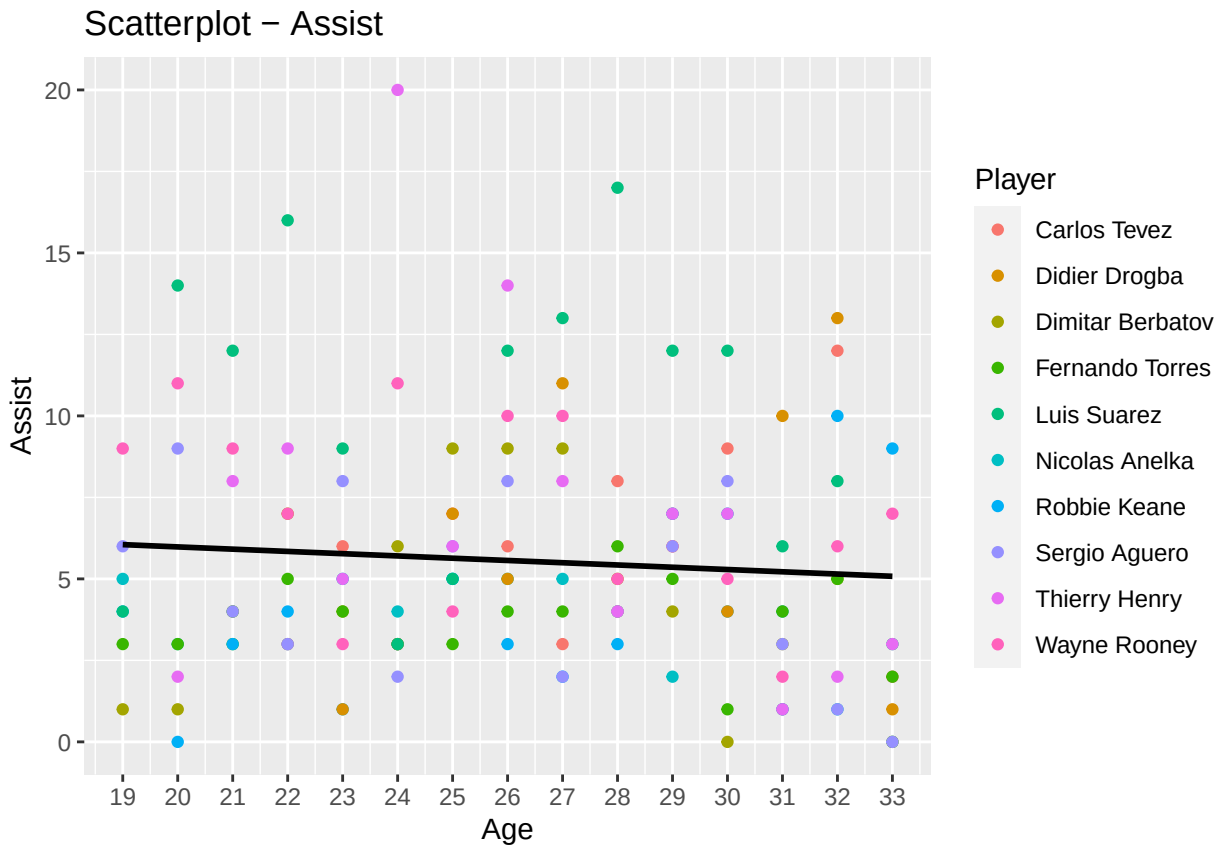


Initial Scatter plots of Age*Assist

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

## Warning: Removed 8 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 8 rows containing missing values (`geom_point()`).
```

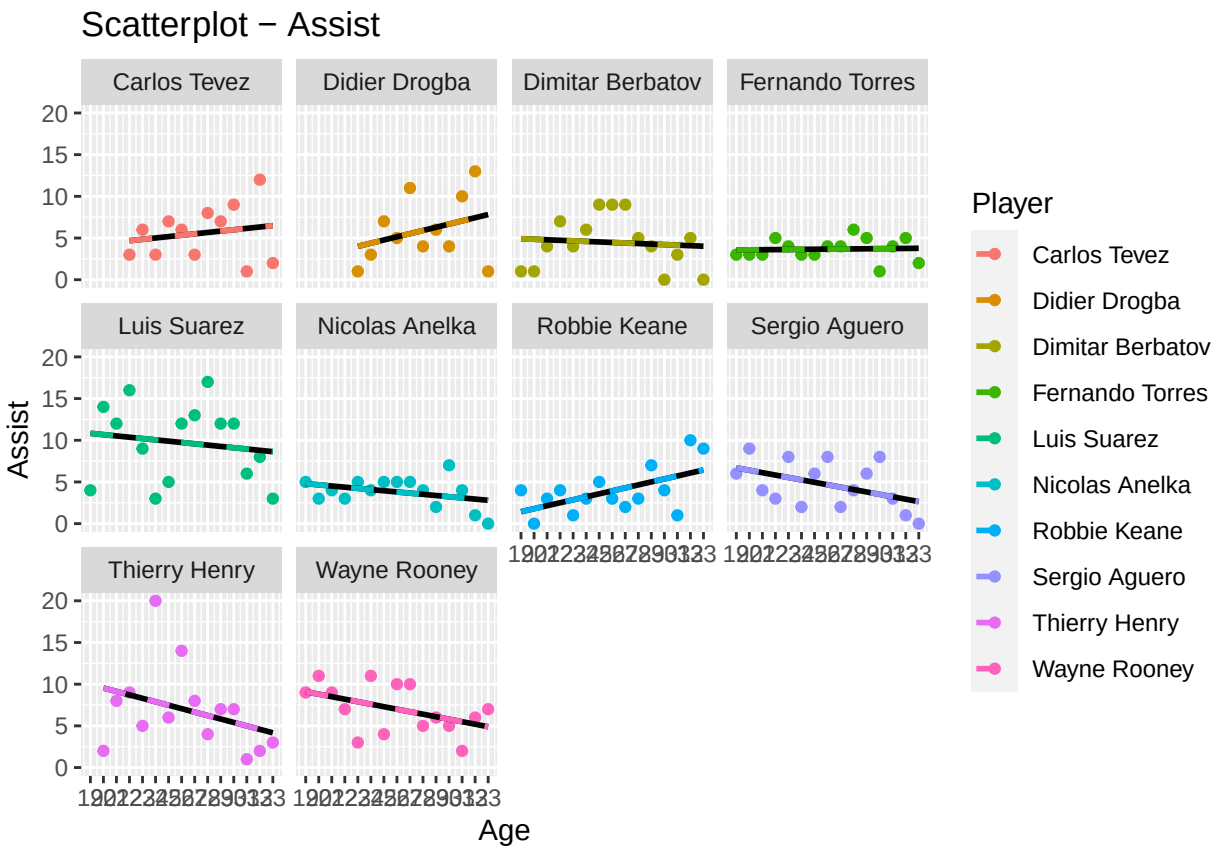



```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 8 rows containing non-finite values (`stat_smooth()`).

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

## Warning: Removed 8 rows containing non-finite values (`stat_smooth()`).
## Removed 8 rows containing missing values (`geom_point()`).
```

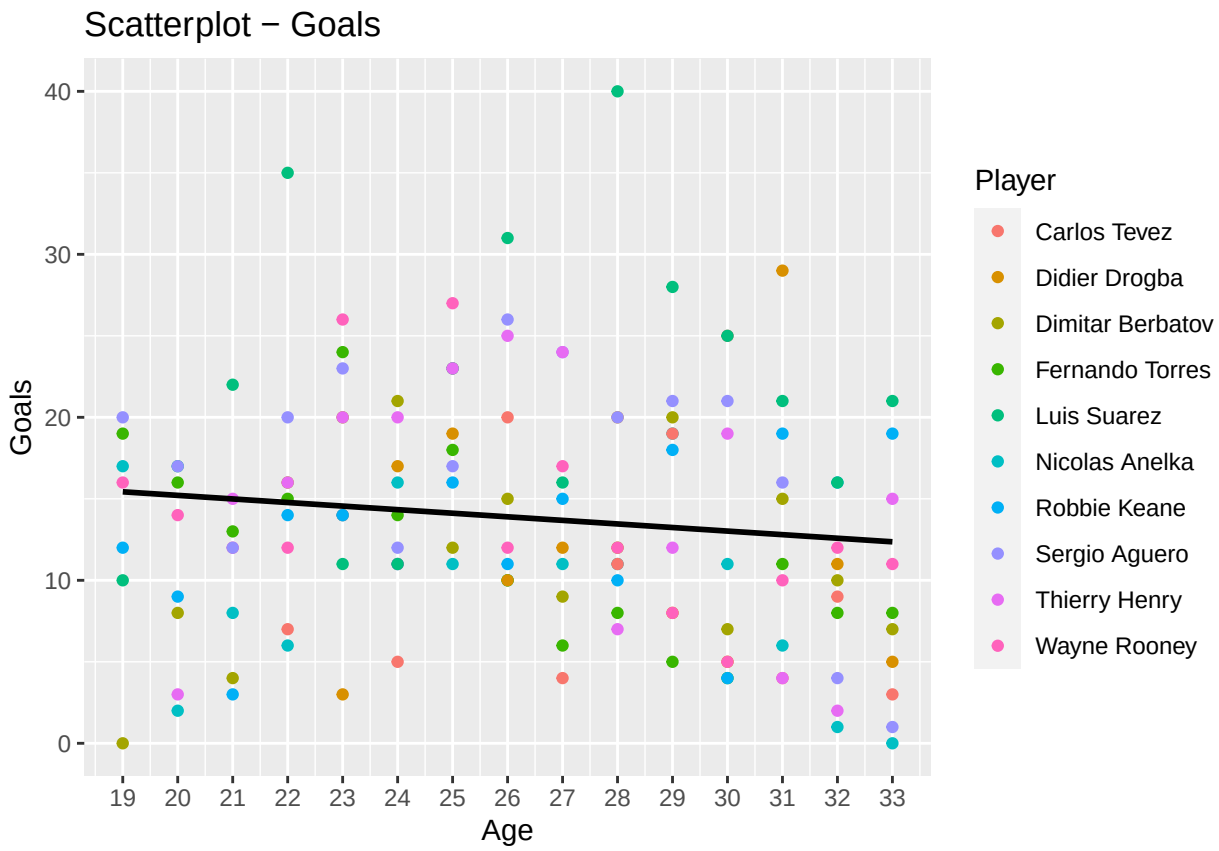


Initial Scatter plots of Age*Goals

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

```
## Warning: Removed 8 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 8 rows containing missing values (`geom_point()`).
```

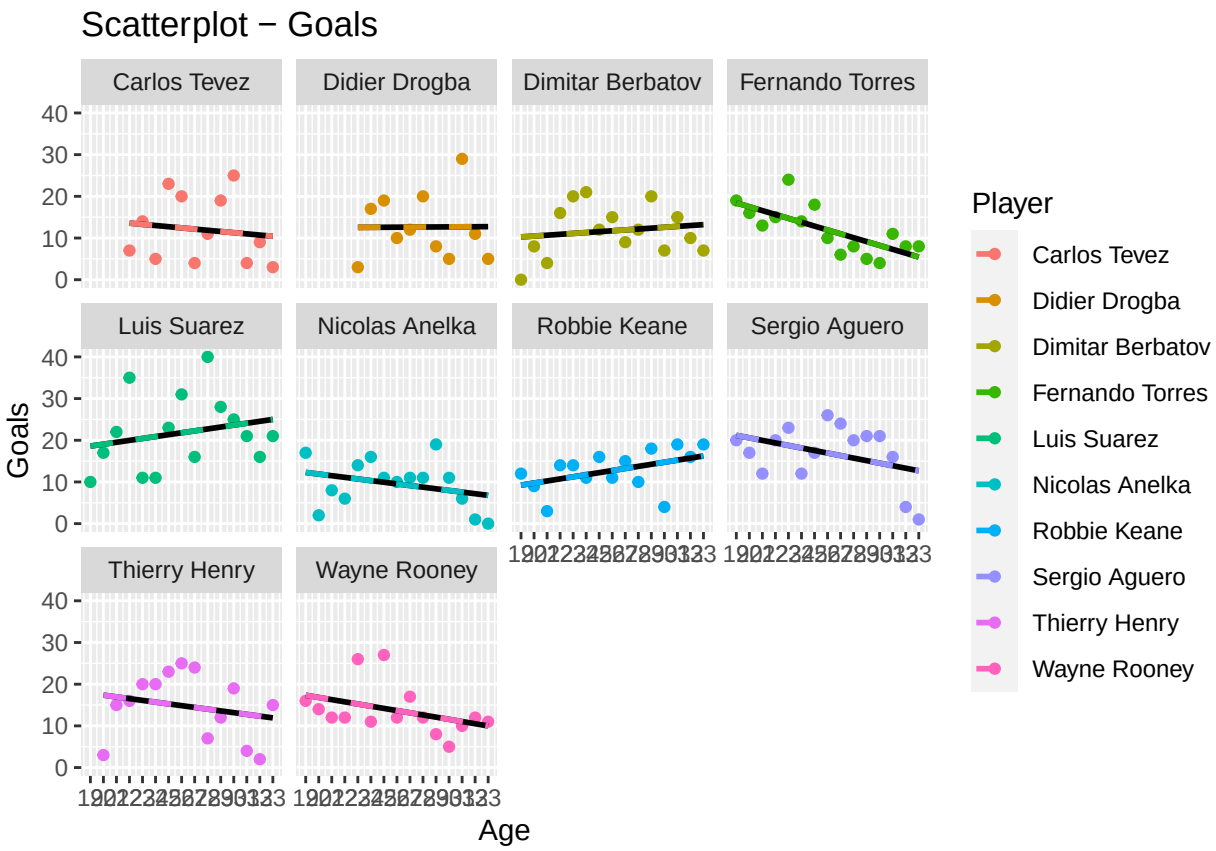



```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 8 rows containing non-finite values (`stat_smooth()`).

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

## Warning: Removed 8 rows containing non-finite values (`stat_smooth()`).
## Removed 8 rows containing missing values (`geom_point()`).
```



As you can see from the created Scatter plots it is extremely hard to analyze the Aging Curve of the Overall trend lines for each elements.The Scatter plots for individual trend lines are bit simpler and better in terms of visualization but analysis for each player is not my goal.Studying the overall trend is my goal.

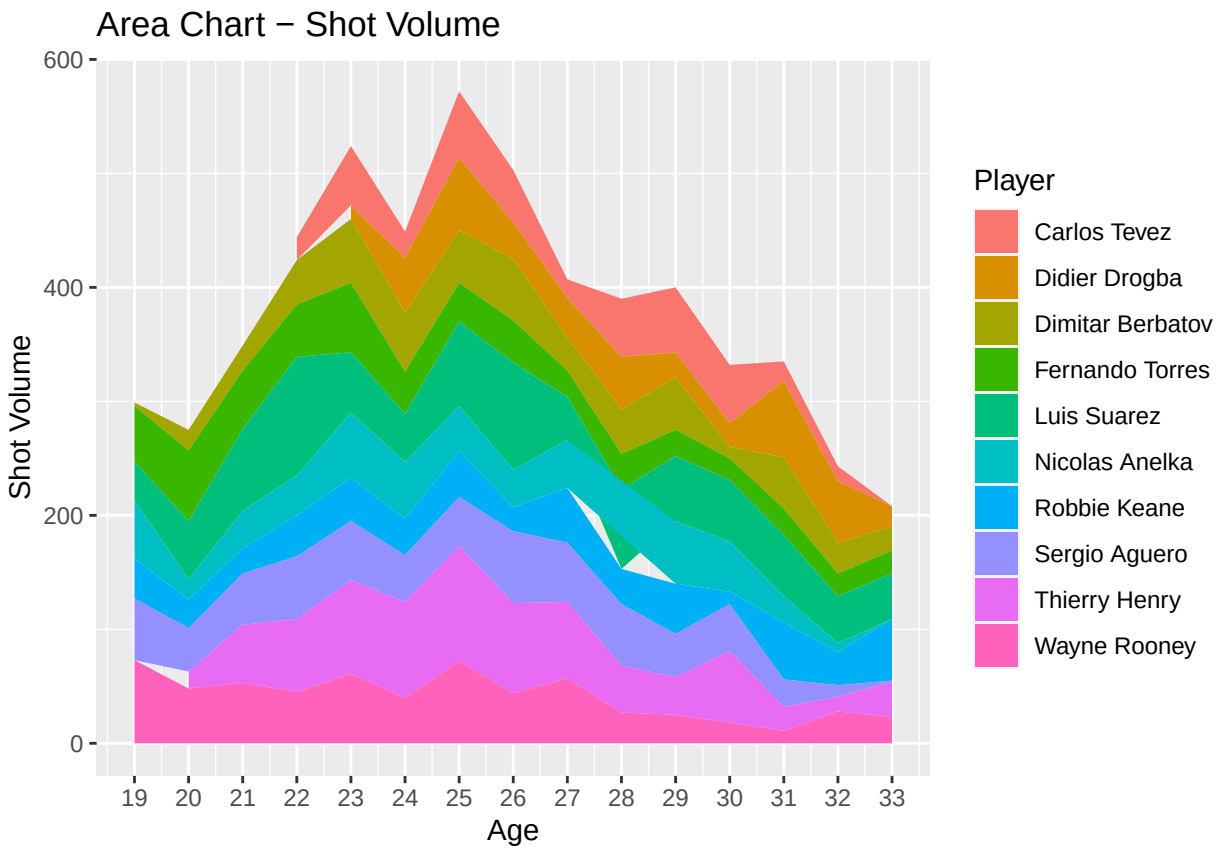
Area Chart as my polished Data Visualization

After a few trials and research, I've realized that the Area Chart is the best way to analyze my data set. Here are the codes and the results of my Area Charts.Moreover, to find when the players tend to reach their primes, I will use the code that calculates the mean of Elements for each age. Arranges the results in descending order, and then retrieve the top row, which represents the age with the highest mean of Elements. The resulting data frame will show the age with the highest mean of Eleme.

Area Chart of Age*Shot_Volume

According to this Chart and the code for peak value, When does the players tend to escalate: Age of 20 When does the players tend to reach their primes: Age of 25 When does the players start to descend from their primes: Age of 27

```
## Warning: Removed 9 rows containing missing values (`position_stack()`).
```

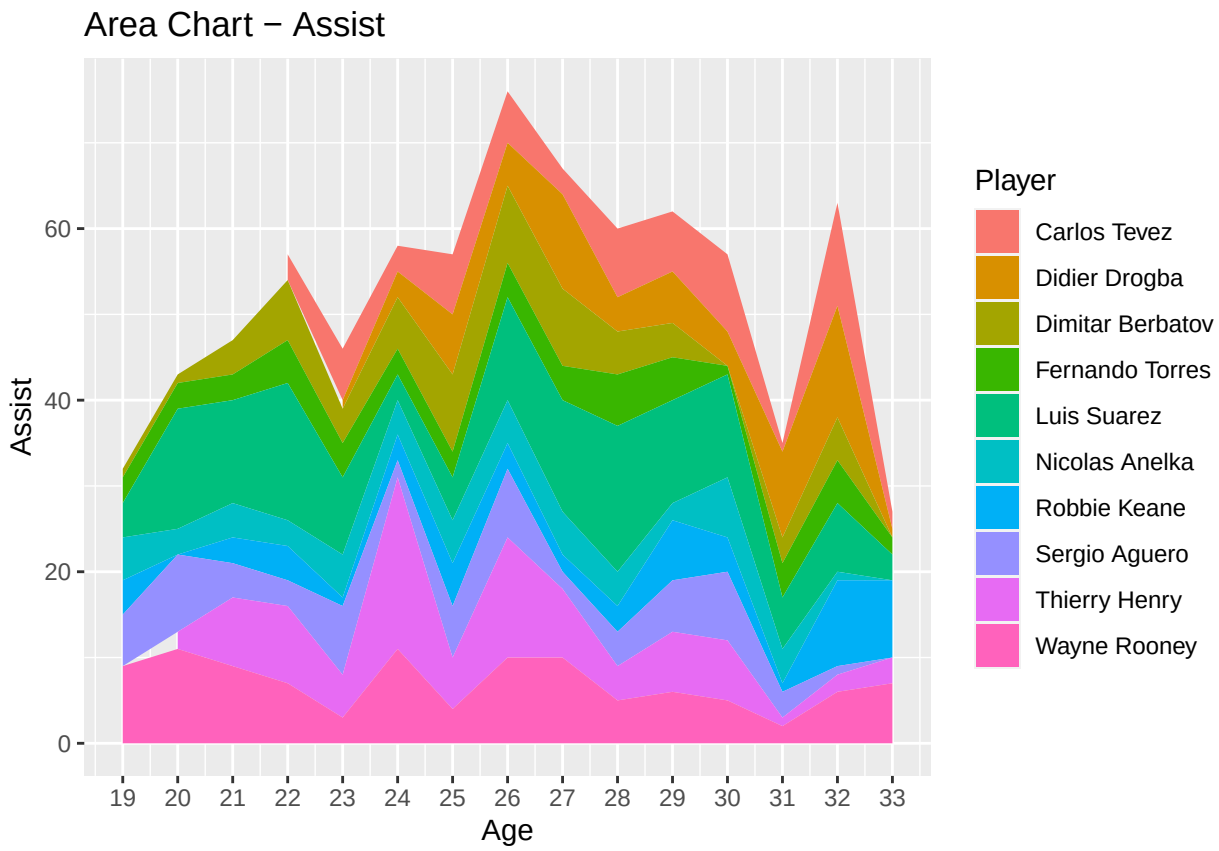



```
## # A tibble: 1 x 2
##     Age mean_shot_volume
##   <dbl>            <dbl>
## 1    25             57.2
```

Area Chart of Age*Assist

According to this Chart and the code for peak value, When does the players tend to escalate: Age of 19 When does the players tend to reach their primes: Age of 26 When does the players start to descend from their primes: Age of 30

```
## Warning: Removed 8 rows containing missing values (`position_stack()`).
```

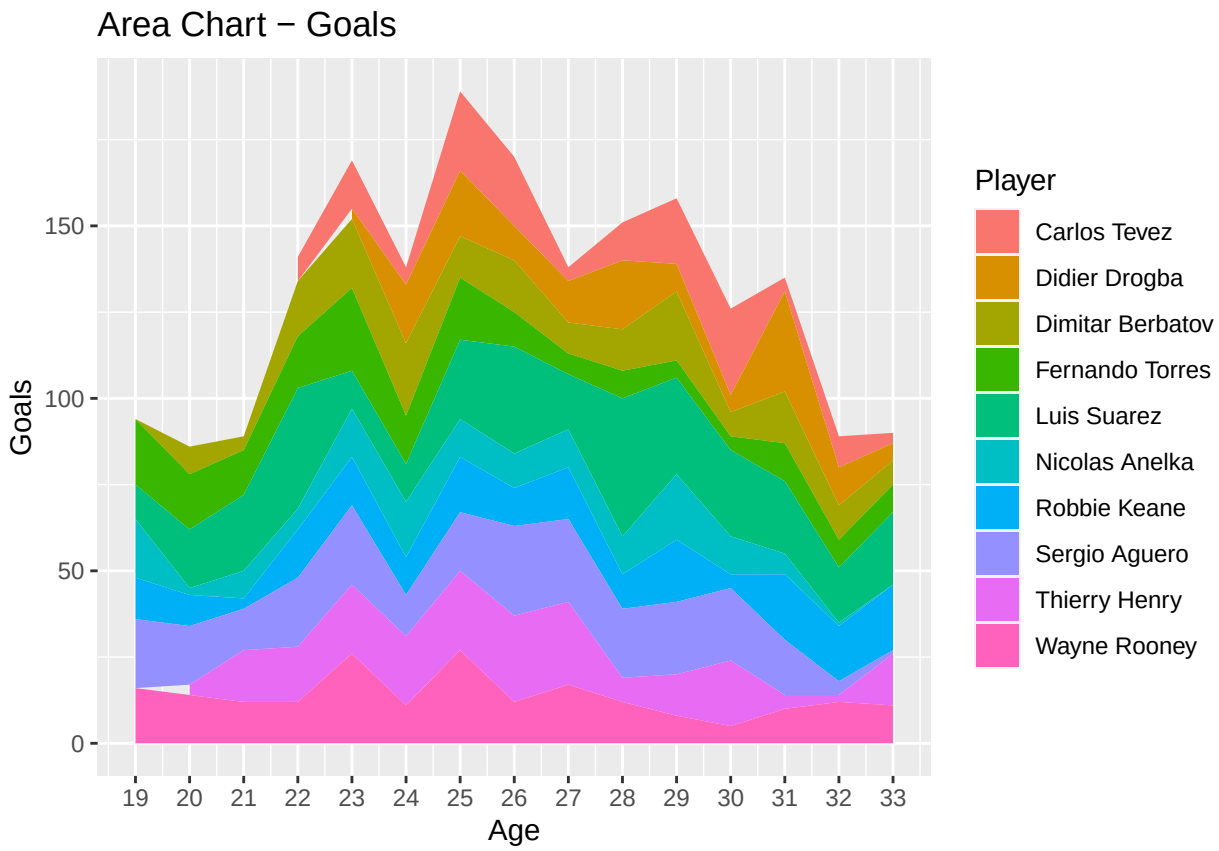



```
## # A tibble: 1 x 2
##     Age mean_shot_volume
##   <dbl>            <dbl>
## 1    26              7.6
```

Area Chart of Age*Goals

According to this Chart and the code for peak value, When does the players tend to escalate: Age of 20 When does the players tend to reach their primes: Age of 25 When does the players start to descend from their primes: Age of 30

```
## Warning: Removed 8 rows containing missing values (`position_stack()`).
```




```
## # A tibble: 1 x 2
##     Age mean_shot_volume
##   <dbl>            <dbl>
## 1    25             18.9
```

Conclusion

After exploring through my data and trying out various ways of data visualization, I have come to a conclusion that Striker position tends to escalate at the age of 20, peaks at the age of 25, and descends at the age of 29. Since my data set is extremely small, I know there are a lot of errors and is not ultimate. However, there was a limit to my data set since I made them by myself on the Google Sheet. Overall, It was a valuable time to review through what I've learned in Stat 184, utilizing them in the field that I'm interested.

Sources of my Data and Research

https://www.fourfourtwo.com/features/ranked-30-best-strikers-premier-league-history https://fbref.com/en/players/ https://statsbomb.com/articles/soccer/player-aging-attacking-players/ https://www.transfermarkt.co.uk/beliebtheit/spieler