Math 268 Final Project

Jake Reynolds

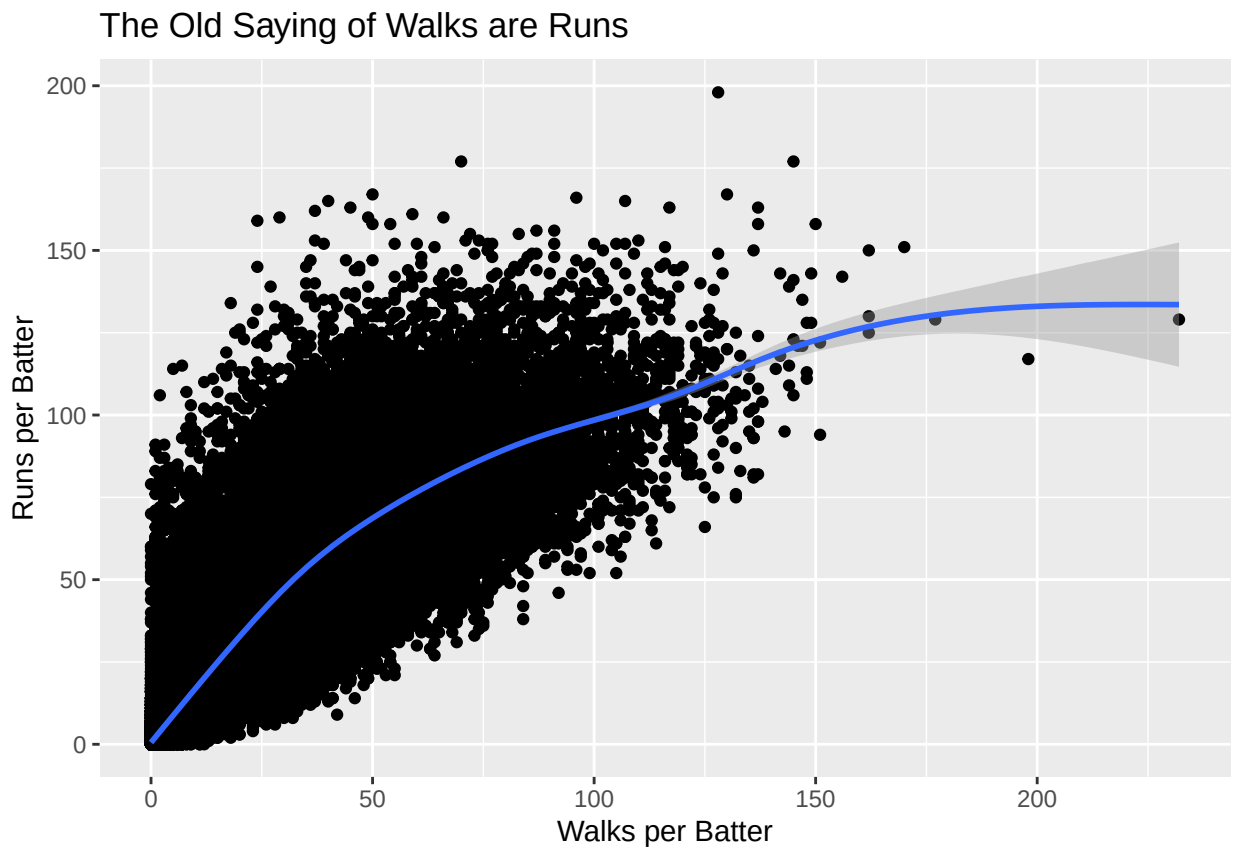
4/21/2021

```
library(tidyverse)
library(corrplot)
library(dplyr)

batting = read_csv('C:/Users/Jake Reynolds/Documents/Batting.csv')
postseason_batting = read_csv('C:/Users/Jake Reynolds/Documents/BattingPost.csv')
fielding = read_csv('C:/Users/Jake Reynolds/Documents/Fielding.csv', na = ' ')
postseason_fielding = read_csv('C:/Users/Jake Reynolds/Documents/FieldingPost.csv')
pitching = read_csv('C:/Users/Jake Reynolds/Documents/Pitching.csv', na = ' ')
postseason_pitching = read_csv('C:/Users/Jake Reynolds/Documents/PitchingPost.csv')
```

There is an old saying, from when I grew up playing baseball. I was always told as a pitcher and a batter that "Walks are runs," so lets test out that saying. First, I graphed walks taken per Batter, versus the amount of runs they scored in that season.

```
ggplot(data = batting,
                 aes(x = BB,
                 y = R)) +
  geom_point()+
  labs(title = 'The Old Saying of Walks are Runs',
       x = 'Walks per Batter',
       y = 'Runs per Batter')+
  geom_smooth()
```



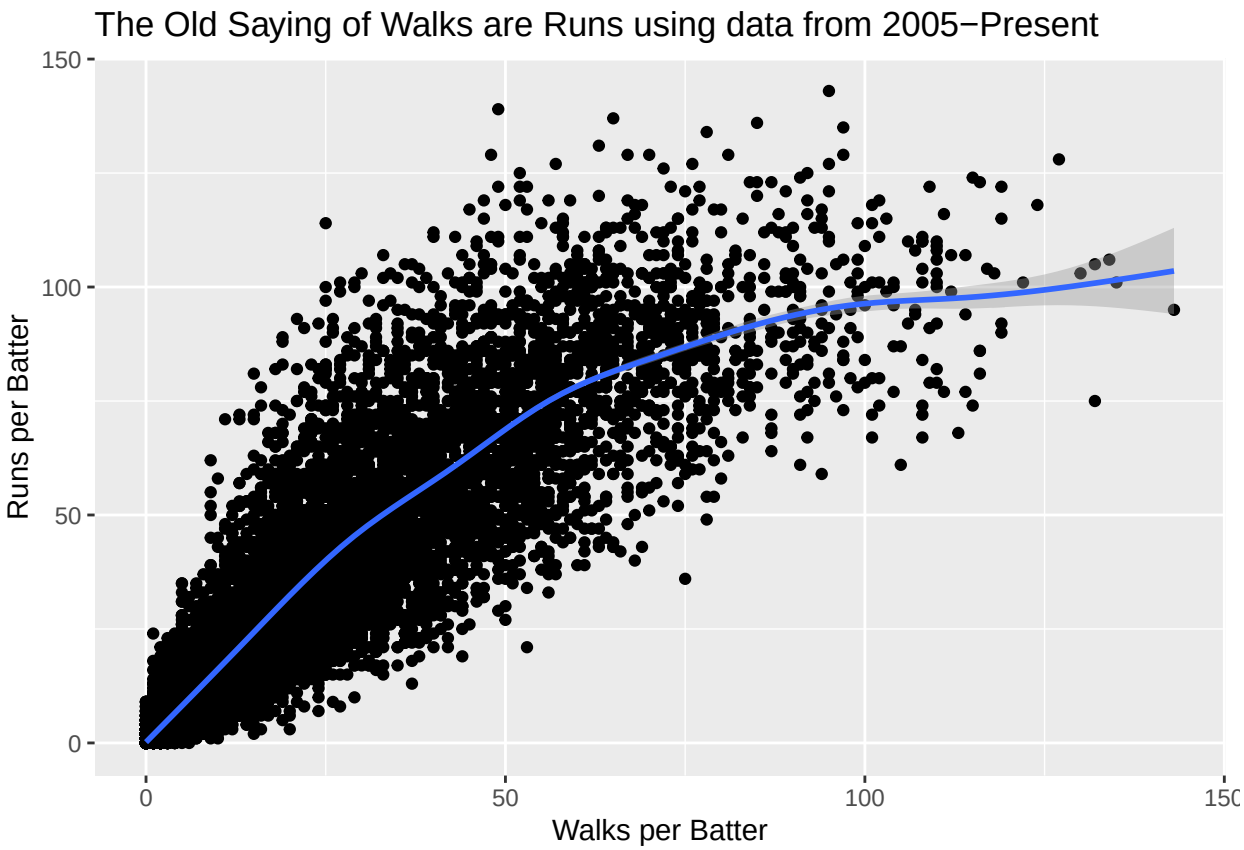
This turned out kind of ugly since there is just a big spot of black at the bottom of the graph, so even with the line of best fit, it doesn't really help. But, it does establish some form of a relationship between the two. Since this set contains data dating back to the 1800s I figured I could trim this data down a bit to what I would call "Modern" baseball. Which in my own definition is when all 30 of the teams became what they are today. The first year that all 30 teams were all in the league together was the 2005 season, which is the first season the Washington Nationals were an MLB team, due to the relocation of the Montreal Expos to Washington. Needless to say, I trimmed the batting, fielding, and pitching datasets to only include data from 2005 - present day.

```
batting_after_2005 = batting[-(1:86000),]
postseason_batting_after_2005 = postseason_batting[-(1:9353),]
fielding_after_2005 = fielding[-(1:115111),]
postseason_fielding_after_2005 = postseason_fielding[-(1:8642),]
pitching_after_2005 = pitching[-(1:36248),]
postseason_pitching_after_2005 = postseason_pitching[-(1:3469),]
```

Now, that that is done, instead of having 108789 observations in one graph I only have 22789 observations, which could make the graphs less heavy at the bottom.

```
ggplot(data = batting_after_2005,
                  aes(x = BB,
                  y = R)) +
  geom_point()+
  labs(title = 'The Old Saying of Walks are Runs using data from 2005-Present',
       x = 'Walks per Batter',
```

```
      y = 'Runs per Batter')+
  geom_smooth()
```

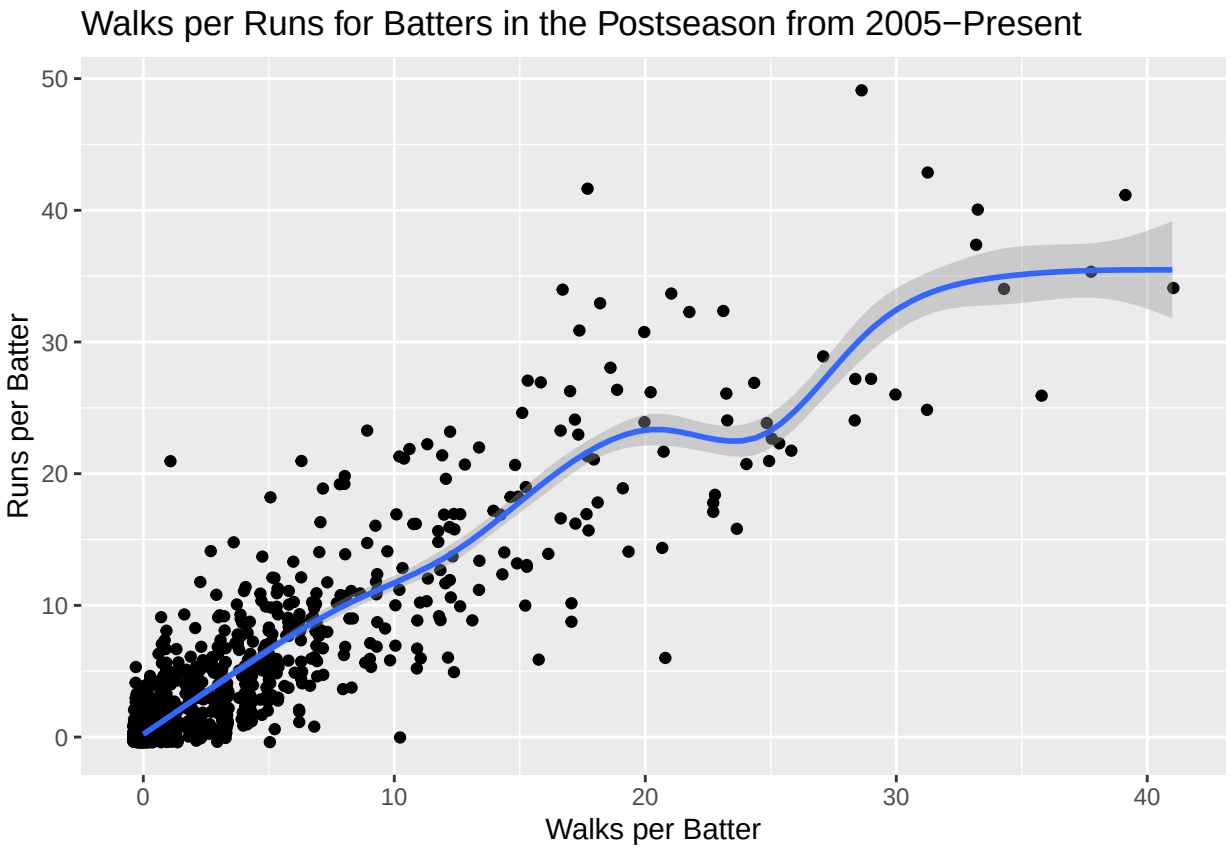


These are a lot of data points but it is a lot less heavy towards the bottom, and basically the same line of best fit is output. I can conclude from this visualization that a relationship exists between walks per batter and the amount of runs that they scored in a season, which can give credence to the saying I was taught as a kid.

Next, I want to see if the relationship gets stronger, weaker, or no change in the postseason. I will again only use data from 2005 on just to avoid having too much data in the bottom left corner of the graph.

```
postseason_batting_after_2005_3 = postseason_batting_after_2005 %>%
  select(-c(teamID, lgID, round, yearID))
postseason_batting_after_2005_4 = aggregate(. ~playerID, data=postseason_batting_after_2005_3, sum, na.

ggplot(data = postseason_batting_after_2005_4,
                aes(x = BB,
                y = R)) +
  geom_jitter()+
  labs(title = 'Walks per Runs for Batters in the Postseason from 2005-Present',
       x = 'Walks per Batter',
       y = 'Runs per Batter')+
  geom_smooth()
```

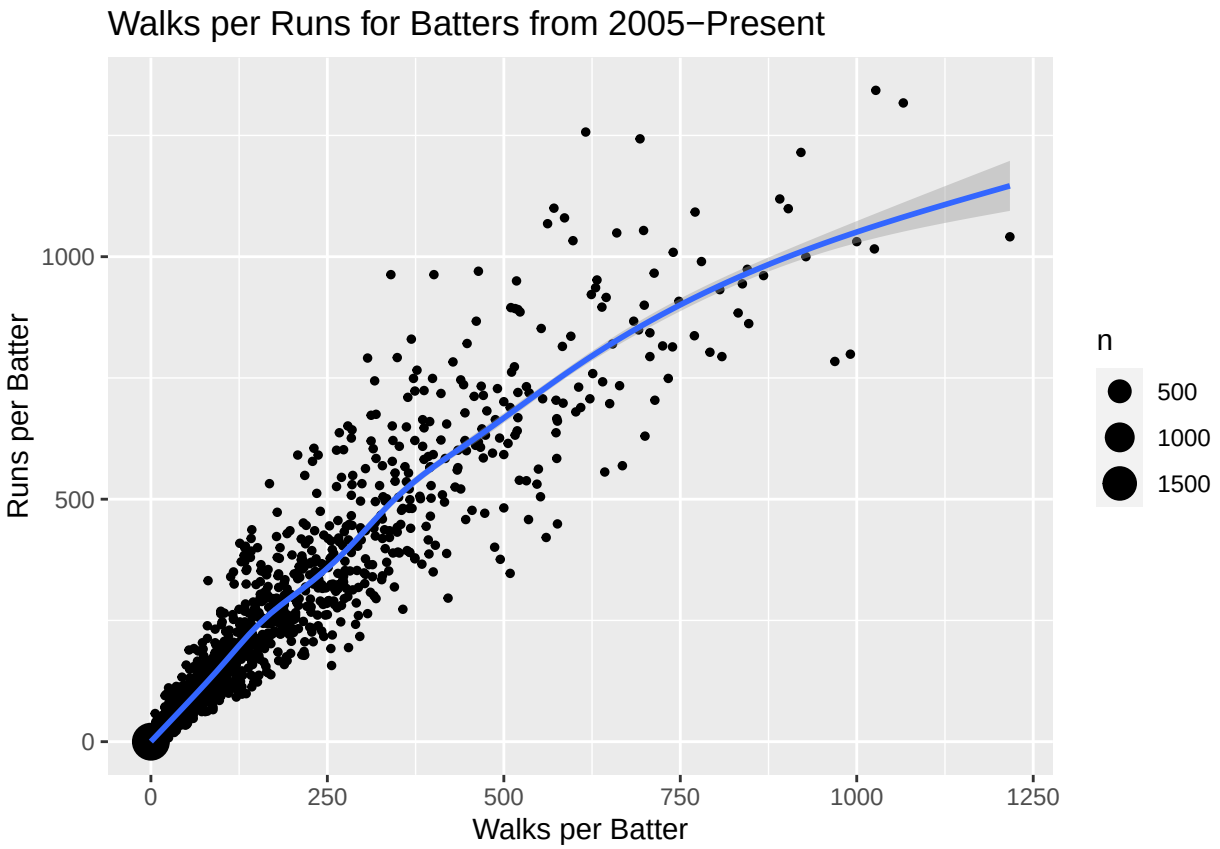


I grouped all the columns by ID so these are career postseason batting stats, and we can see a relationship between walks and runs. The relationship isn't as clear as the regular season but there appears to at least be one. I'm going to try the grouping for regular season data as well just to see what I find.

```
batting_after_2005_2 = batting_after_2005 %>%
  select(-c(yearID, teamID, IBB, HBP, SH, SF, lgID, stint)) %>%
  group_by(playerID)

batting_after_2005_2 = aggregate(. ~playerID, data=batting_after_2005_2, sum, na.rm=TRUE)

ggplot(data = batting_after_2005_2,
                aes(x = BB,
                y = R)) +
  geom_count()+
  labs(title = 'Walks per Runs for Batters from 2005-Present',
       x = 'Walks per Batter',
       y = 'Runs per Batter')+
  geom_smooth()
```



This is probably the best graph so far because I used combines stats for players instead of numbers from each individual season.

Next, I'm going to see which team has the best batting average since 2005 and who has the most runs scored.

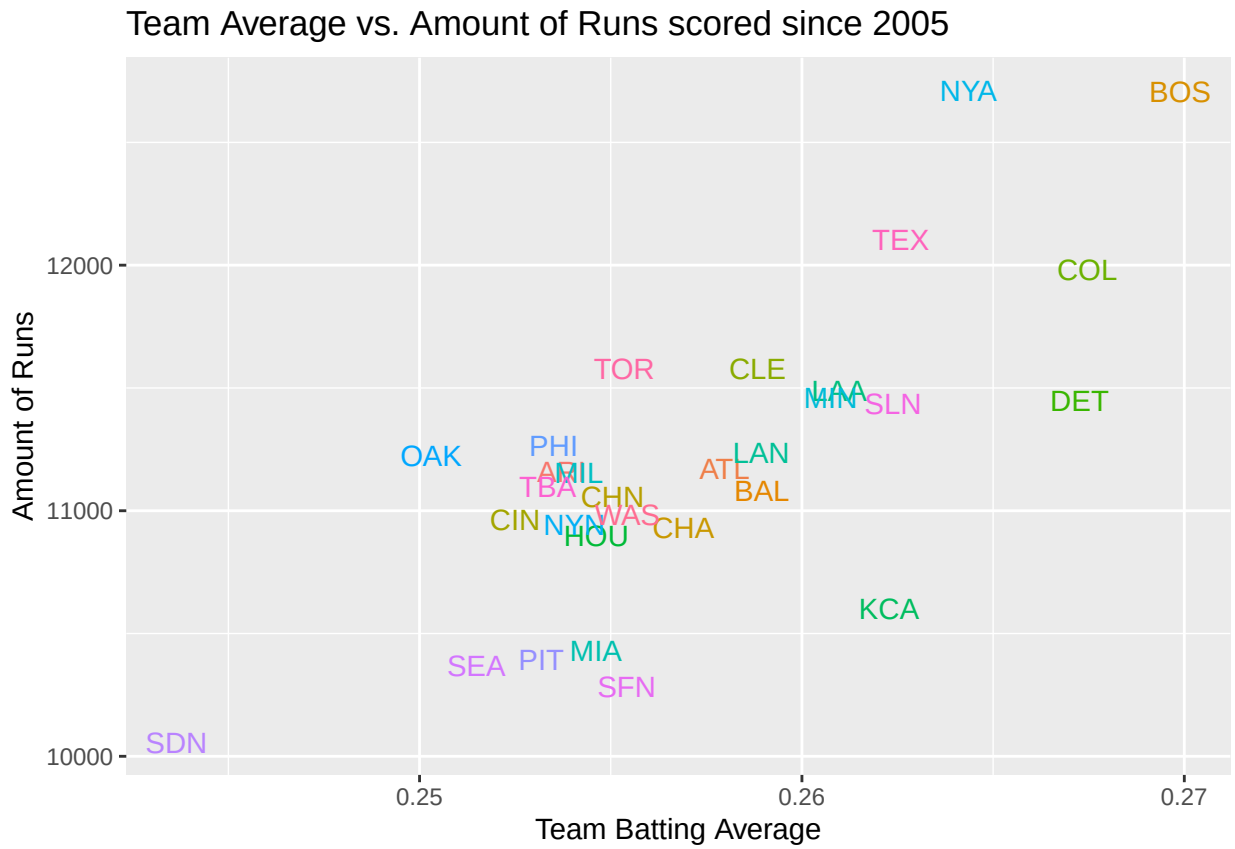
```
batting_after_2005_3 = batting_after_2005 %>%
  select(-c(yearID, playerID, IBB, HBP, SH, SF, lgID, stint)) %>%
  group_by(teamID)

team_batting_data = aggregate(. ~teamID, data=batting_after_2005_3, sum, na.rm=TRUE)

team_batting_data[16,2:14] = team_batting_data[11,2:14]+team_batting_data[16,2:14]
team_batting_data = team_batting_data[-11,] #I'm combining the Miami and Florida rows here since they a

team_batting_data = team_batting_data %>%
  mutate(AVG = H/AB)

ggplot(data = team_batting_data,
                aes(x = AVG,
                y = R), label = teamID) +
  labs(title = 'Team Average vs. Amount of Runs scored since 2005',
       x = 'Team Batting Average',
       y = 'Amount of Runs')+
  geom_text(aes(label = teamID, color = teamID), show.legend = FALSE)
```



Historically over the past 16 seasons, the Red Sox and Yankees have been some of the best teams in the league. Also, the Padres being at the bottom isn't surprising at all too since they have not been good up until these past few years.

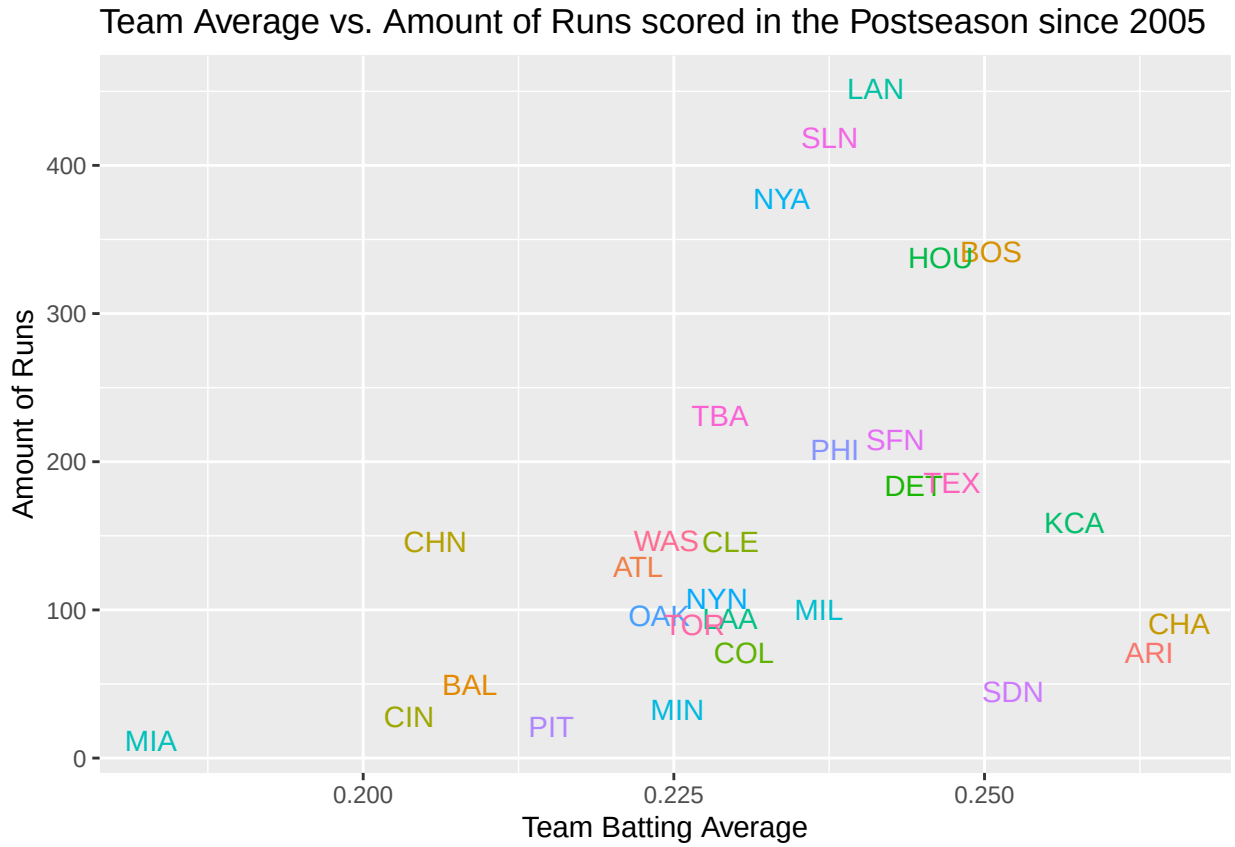
Lets try out these stats for the postseason to see who performed the best and worst.

```
postseason_team_batting_data = postseason_batting_after_2005 %>%
  select(-c(yearID, playerID, lgID, round)) %>%
  group_by(teamID)

postseason_team_batting_data = aggregate(. ~teamID, data=postseason_team_batting_data, sum, na.rm=TRUE)

postseason_team_batting_data = postseason_team_batting_data %>%
  mutate(AVG = H/AB)

ggplot(data = postseason_team_batting_data,
                aes(x = AVG,
                y = R), label = teamID) +
  labs(title = 'Team Average vs. Amount of Runs scored in the Postseason since 2005',
       x = 'Team Batting Average',
       y = 'Amount of Runs')+
  geom_text(aes(label = teamID, color = teamID), show.legend = FALSE)
```



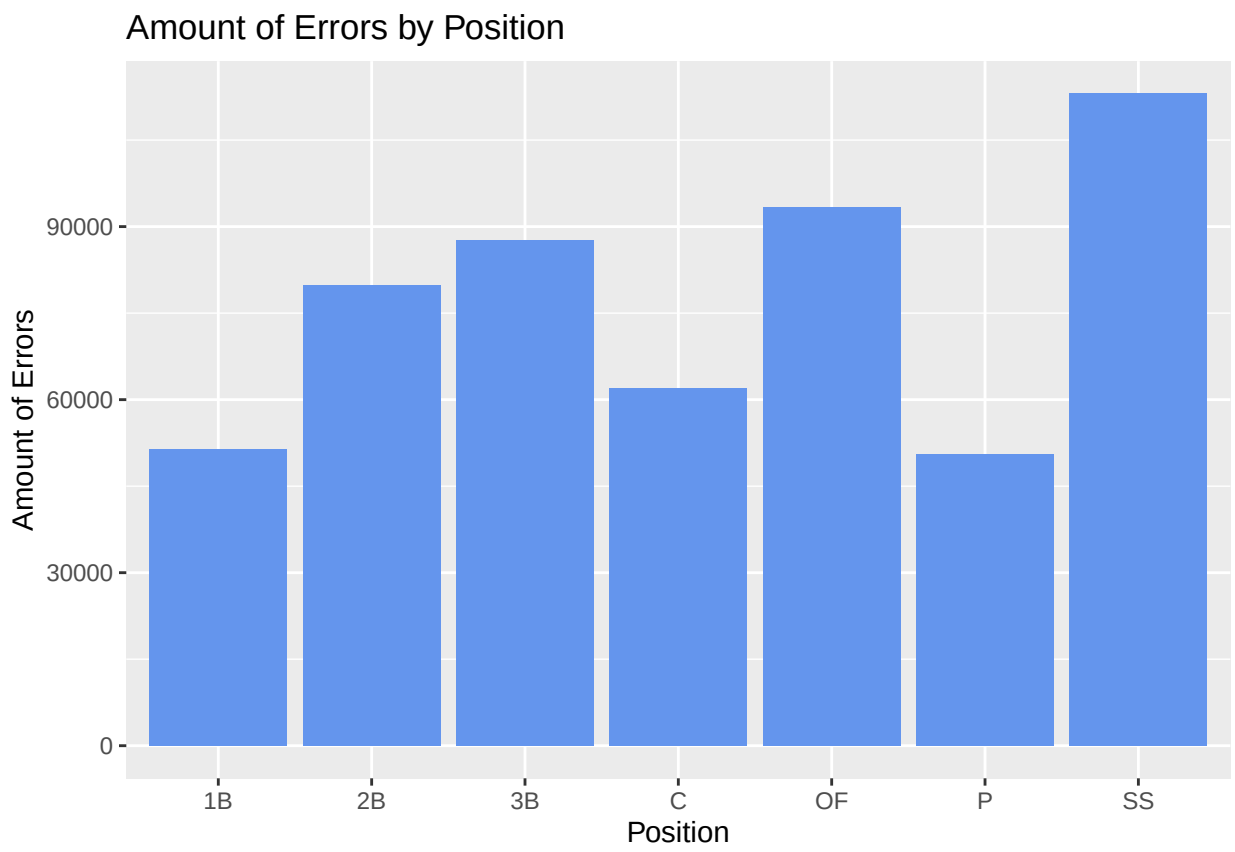
Yet again, the most successful playoff teams in this category are ones that usually do very well in the post season. Miami has only played five playoff games since 2005 so then having a low amount of runs isn't surprising. Also note that only 29 teams are shown since the Mariners haven't made the playoffs since 2001. This might be scary to read but I'm only about 1/3 of the way done because I still need to explore pitching and fielding data, but I'm having a ton of fun exploring so it won't be an issue for me to finish. Another interesting note here is that Arizona and the White Sox both have outstanding playoff batting averages but don't score a lot of runs, which is not good because that means they leave a lot of people on base.

Now lets move on to fielding, one interesting question I have is which position historically has the most errors. The other question I will try to answer is which team has the most errors in the postseason. The first chunk here is just me cleaning the data of columns with NAs since they aren't needed for what I am graphing.

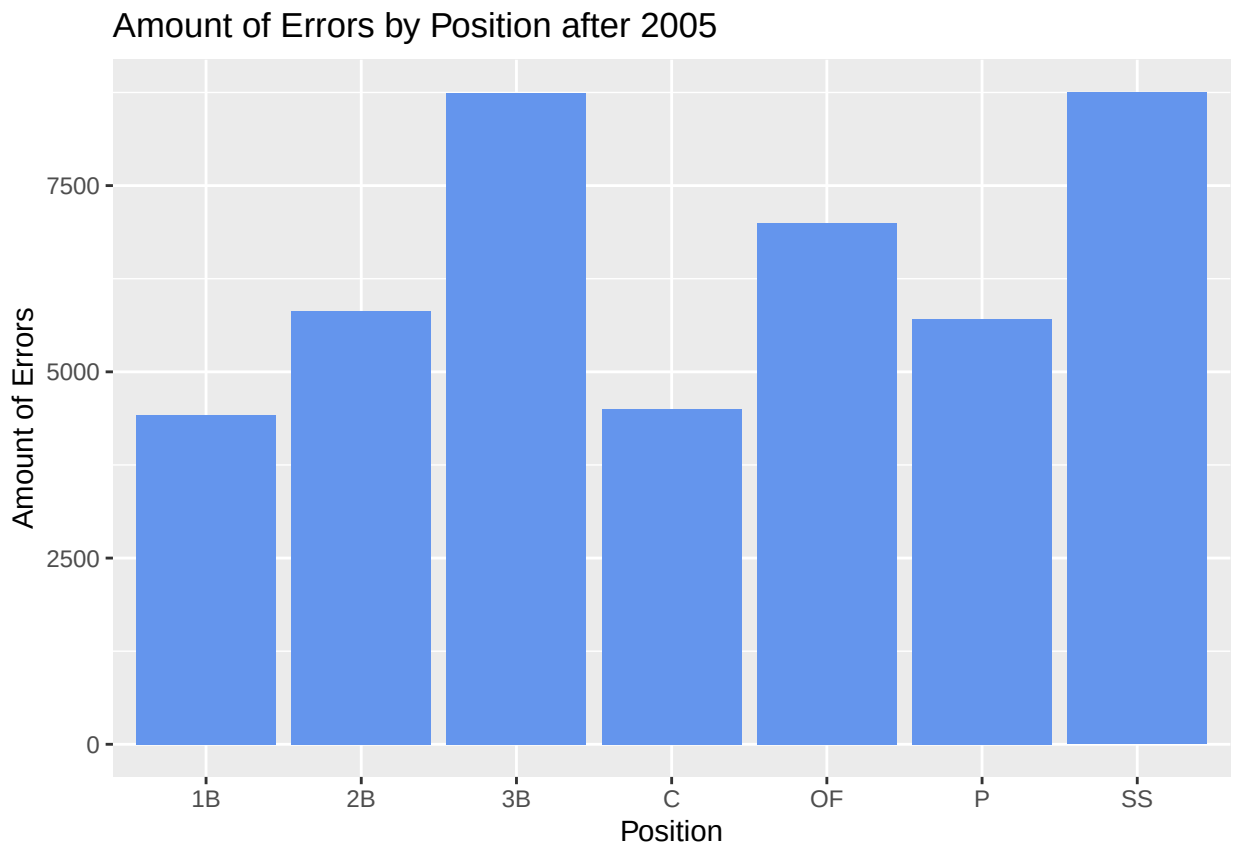
```
fielding = fielding %>%
  select(-c(GS, InnOuts, PB, WP, SB, CS, ZR)) %>%
  mutate(E = ifelse(is.na(E), 0, E))

fielding_after_2005 = fielding_after_2005 %>%
  select(-c(GS, InnOuts, PB, WP, SB, CS, ZR)) %>%
  mutate(E = ifelse(is.na(E), 0, E))
```

```
ggplot(fielding,
       aes(x=POS, y = E))+
  geom_bar(stat = 'identity', fill = 'cornflowerblue')+
  labs(title = 'Amount of Errors by Position',
            x = 'Position',
            y = 'Amount of Errors')
```




```
ggplot(fielding_after_2005,
       aes(x=POS, y = E))+
  geom_bar(stat = 'identity', fill = 'cornflowerblue')+
  labs(title = 'Amount of Errors by Position after 2005',
            x = 'Position',
            y = 'Amount of Errors')
```



Historically the top 3 positions for errors are Shortstop, Outfield, then Third Base, with Pitchers committing the least. In "Modern" Baseball", it appears that Third Base and Shortstop are right around the same amount of errors committed at the top, and then third is Outfield. It also looks like Catcher and First Base have the least amount of errors. For some reason when I try to display the counts onto the graph they just show up as a black square so I'll just display the counts real quick.

```
fielding3 = fielding %>%
  group_by(POS) %>%
  summarise(E = sum(E))
fielding3
```

```
## # A tibble: 7 x 2
##   POS        E
## * <chr>  <dbl>
## 1 1B     51340
## 2 2B     79846
## 3 3B     87677
## 4 C      61984
## 5 OF     93379
## 6 P      50405
## 7 SS    113033
```

```
fielding_after_2005_3 = fielding_after_2005 %>%
  group_by(POS) %>%
  summarise(E = sum(E))
fielding_after_2005_3
```

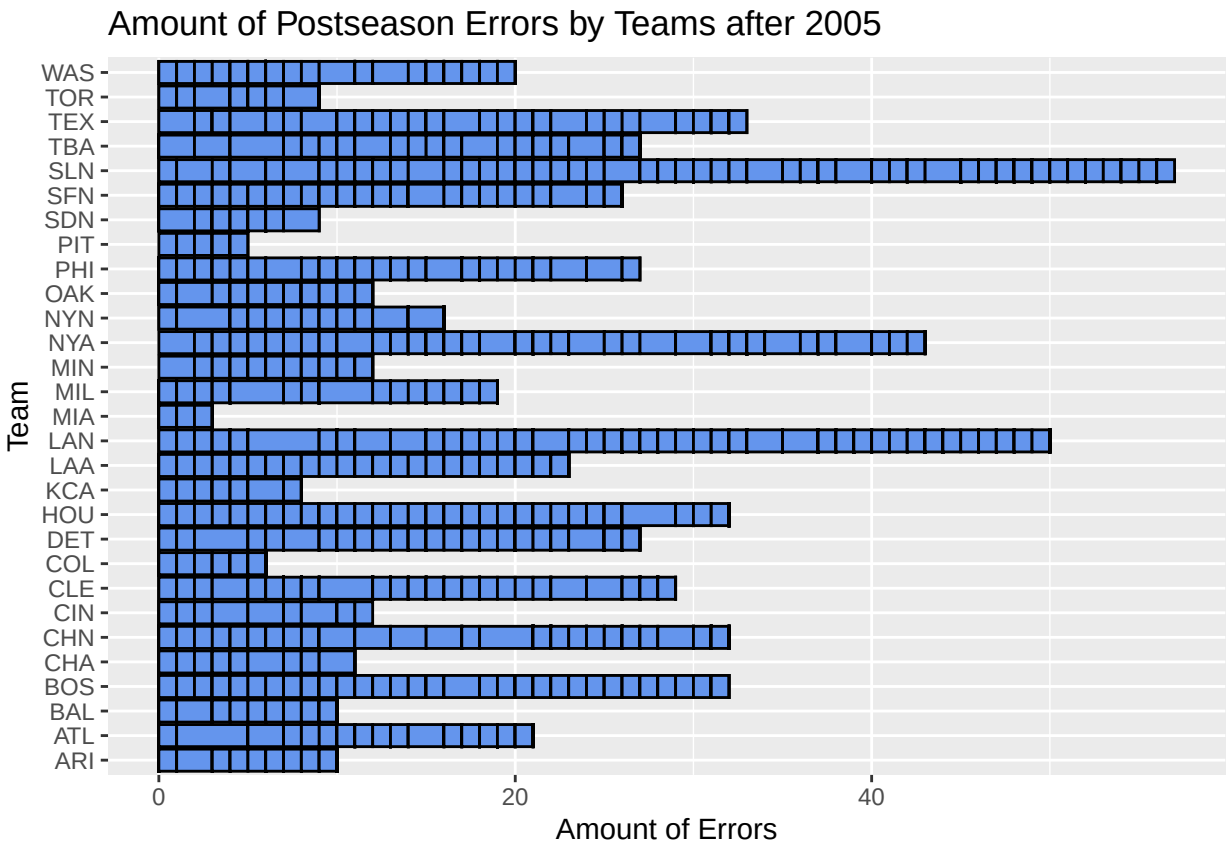
```
## # A tibble: 7 x 2
##   POS       E
## * <chr> <dbl>
## 1 1B     4413
## 2 2B     5812
## 3 3B     8744
## 4 C      4500
## 5 OF     6993
## 6 P      5699
## 7 SS     8752
```

Pitchers do in fact have the all time least amount of errors. In "modern" baseball, the most goes to Shortstop, then third base, then Outfield. For the least, first base wins with 87 less than catchers.

Now lets try to see which teams have the most errors in the postseason. I'm just going to use the data after 2005 so I don't have to combine teams (ex. Montreal Expos and Washington Nationals).

```
postseason_fielding_after_2005_2 = postseason_fielding_after_2005 %>%
  select(c(teamID, E)) %>%
  mutate(E = ifelse(is.na(E), 0, E))

ggplot(postseason_fielding_after_2005_2,
       aes(x=teamID, y = E))+
  geom_bar(stat = 'identity', fill = 'cornflowerblue', color = 'black')+
  labs(title = 'Amount of Postseason Errors by Teams after 2005',
            x = 'Team',
            y = 'Amount of Errors')+
  coord_flip()
```

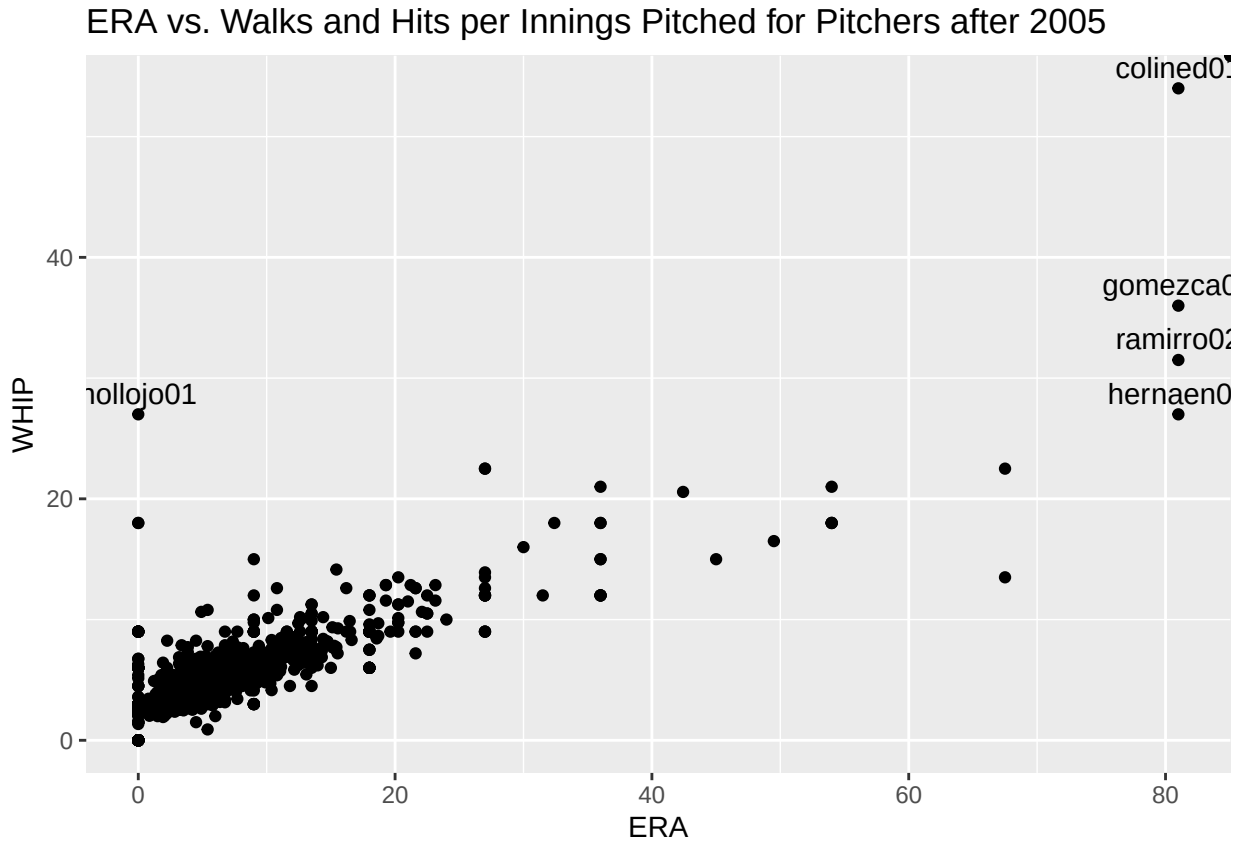


When I filled in the outlines of these I realized that the mini boxes inside are individual players recording an error, so you can actually count how many players recorded an out for that team. After reading the plot you'll realize that St. Louis has the most errors followed by the Dodgers, and third being the Yankees.

Lastly I will plot Earned Run Average vs. Walks and Hits per Inning Pitched to see if there is some relationship. In my opinion these are the best stats for a pitcher and you would want both of them to be very low.

```
pitching_after_2005_2 = pitching_after_2005 %>%
  select(c(playerID, ER, IPouts, H, BB))
pitching_after_2005_2 = aggregate(. ~playerID, data=pitching_after_2005_2, sum, na.rm=TRUE)
pitching_after_2005_2 = pitching_after_2005_2 %>%
  mutate(ERA = ER * 27 / IPouts,
         Innings = IPouts / 9 ,
         WHIP = (H+BB)/Innings)

ggplot(data = pitching_after_2005_2,
                aes(x = ERA,
                y = WHIP)) +
  geom_point()+
  labs(title = 'ERA vs. Walks and Hits per Innings Pitched for Pitchers after 2005',
       x = 'ERA',
       y = 'WHIP')+
  geom_text(aes(label = ifelse(WHIP>25, as.character(playerID), ' ')), hjust = 0.5, vjust = -0.5)
```

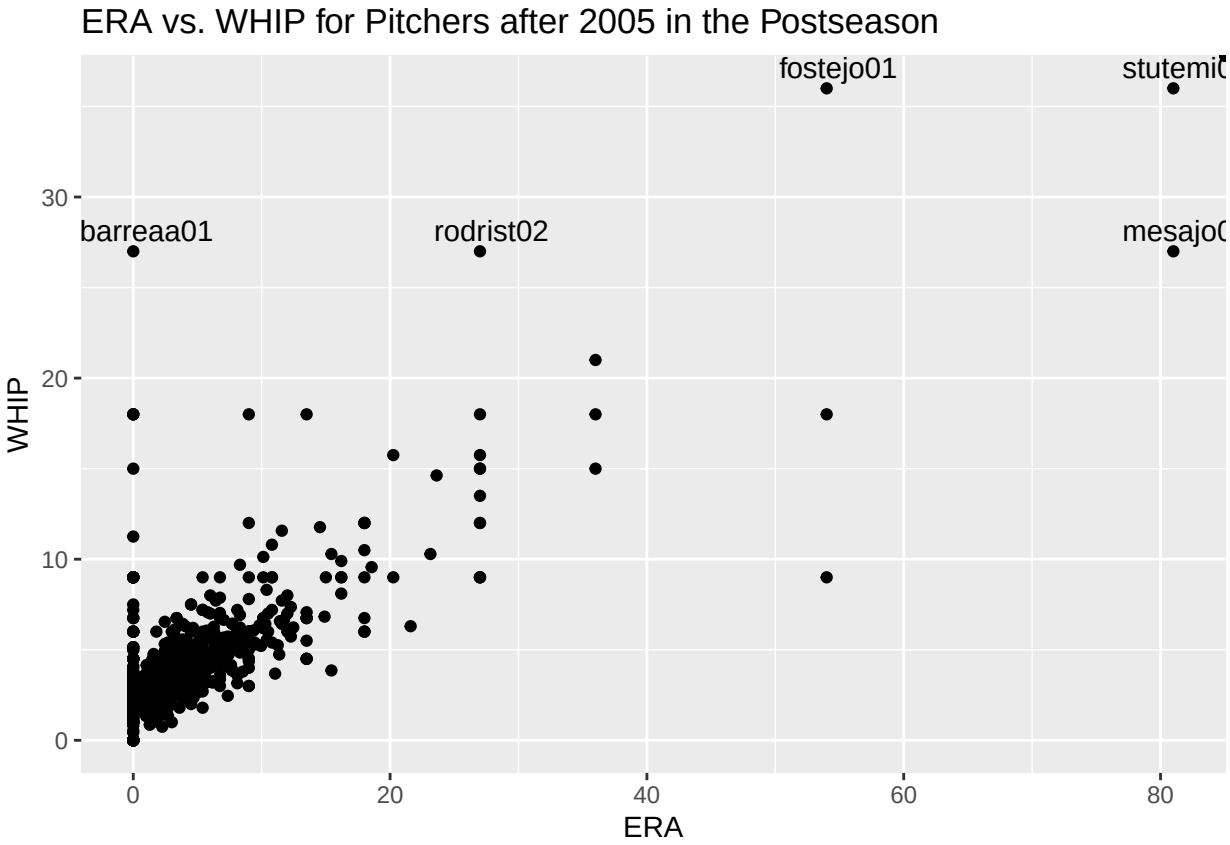


Normally, you look for the best in every category but here we found the five worst. Jordon Holloway is the pitcher on the far left who miraculously has a super low ERA but manages to give up walks and hits like it is his day job. Starting from the bottom of the listed guys on the right we have Enrique Hernandez, Roel Ramirez, Carlos Gomez, and lastly, the worst of them all, Edwar Colina

Now lets look in the postseason data to expose the worst in the postseason.

```
postseason_pitching_after_2005_2 = postseason_pitching_after_2005 %>%
  select(c(playerID, ER, IPouts, H, BB))
postseason_pitching_after_2005_2 = aggregate(. ~playerID, data=postseason_pitching_after_2005_2, sum, na
postseason_pitching_after_2005_2 = postseason_pitching_after_2005_2 %>%
  mutate(ERA = ER * 27 / IPouts,
         Innings = IPouts / 9 ,
         WHIP = (H+BB)/Innings)

ggplot(data = postseason_pitching_after_2005_2,
                aes(x = ERA,
                y = WHIP)) +
  geom_point()+
  labs(title = 'ERA vs. WHIP for Pitchers after 2005 in the Postseason',
       x = 'ERA',
       y = 'WHIP')+
  geom_text(aes(label = ifelse(WHIP>25, as.character(playerID), ' ')), hjust = 0.4, vjust = -0.5)
```



So the five worst postseason pitchers are as follows: Aaron Barrett, Paco Rodriguez, John Foster, Jose Mesa, and Michael Stutes. Sorry I made a lot of graphs but this is the first project I've ever enjoyed doing so thank you for the opportunity to do this!

Baseball Dataset source: http://www.seanlahman.com/baseball-archive/statistics/