Final_Draft

Trent Meyer

2022-05-14

Introduction

What

My project is about MLB hitting stats, specifically home runs and exit velocity. Average exit velocity means on average, how fast did the ball leave the players bat if they put a ball in play. The idea of Exit Velocity has existed ever since the founding of physics, and it has always been recognized in the sport of baseball. The general idea is that the harder you hit a baseball, the farther it may go, but this is also dependent on the angle it is hit at, which many may have experimented when looking at projectile motion in physics class. Recent technology (starting in 2015) has made it possible for mlb teams to track the exit velocity of every single ball put in play during every baseball game throughout the season. This has caused baseball statisticians to dig more into the stat and recognize its importance in the game, and its predictive capabilities. I am going to look at both home runs and average exit velocity separately. Then I am going to look at their relationship between each other.

Why

My goal throughout this project is to explore and use many different functions in RStudio learned throughout the semester. I am very interested in sports, especially baseball, since I played in high school and statistics are a very important aspect of analyzing the game. I want to see how home runs have varied over different seasons. I also want to see how exit velocity has changed over multiple seasons. I want to recognize if there are any trends here. Finally, I want to see what type of relationship exists between home runs and exit velocity.

How

I downloaded season long hitting statistics for certain players during each of the 2015 - 2021 seasons. This data included a players name, as well as multiple different hitting statistics that I found of interest. I had a very ambitious vision at the beginning and wanted to analyze more than just home runs and exit velocity, but due to time constraint and the difficulty of the project, I just stuck to those two. The criteria for the players included in the list are players that have 2.1 plate appearances per game for their team. This means these players played in a lot of games and recorded a lot of plate apperances throughout the season. I thought selecting players based on this criteria would give me more reliable numbers, especially since the statistics are aggregate. If you are looking at season long statistics, the data could potentially be skewed if there were players in the list that did not record many plate appearances.

note: The numbers in the 2020 season are far lower than all of the other datasets, and this is because of the COVID-19 pandemic. During this season, the teams only played 60 games each, in comparison to 162 in a normal season.

Body: Data Analysis

Data Load and Sample

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

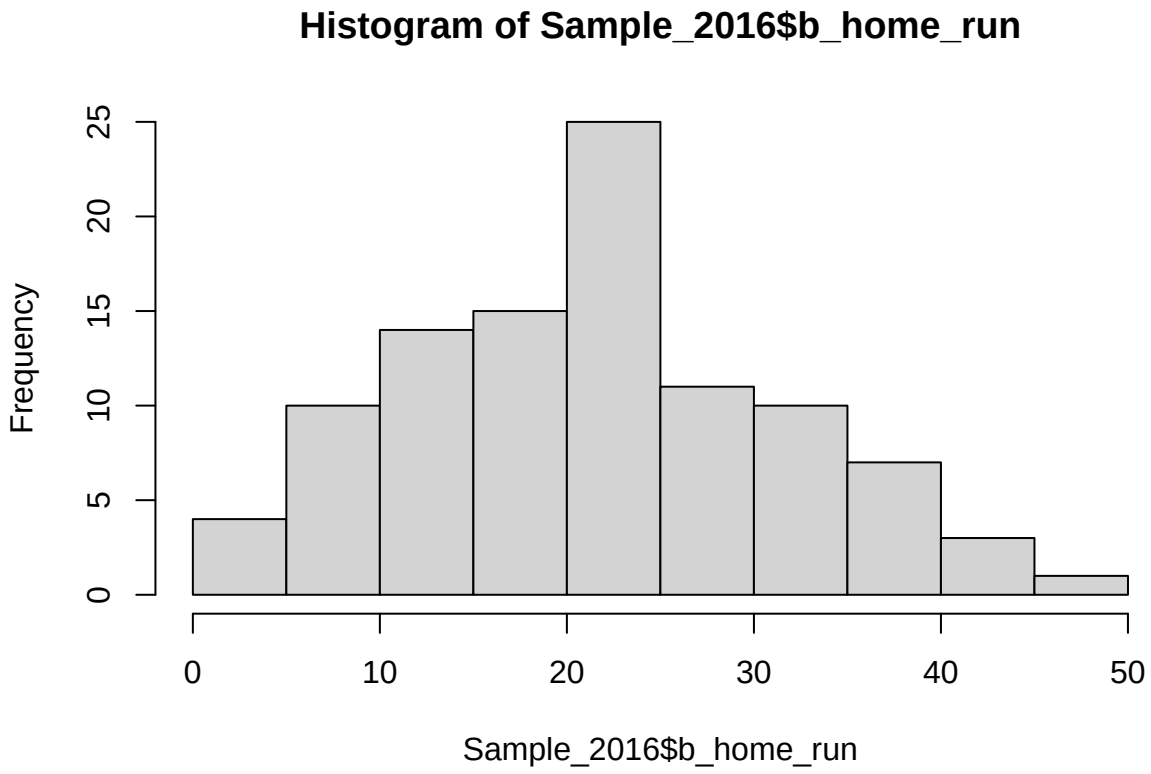
```
library(readxl)

Hitting_Stats_2015 <- read_excel("~/Desktop/Hitting_Stats_2015.xlsx", skip = 1)

Hitting_Stats_2016 <- read_excel("~/Desktop/Hitting_Stats_2016.xlsx", skip = 1)

Hitting_Stats_2017 <- read_excel("~/Desktop/Hitting_Stats_2017.xlsx", skip = 1)

Hitting_Stats_2018 <- read_excel("~/Desktop/Hitting_Stats_2018.xlsx", skip = 1)

Hitting_Stats_2019 <- read_excel("~/Desktop/Hitting_Stats_2019.xlsx", skip = 1)

Hitting_Stats_2020 <- read_excel("~/Desktop/HittingStats2020.xlsx", skip = 1)

Hitting_Stats_2021 <- read_excel("~/Desktop/HittingStats2021.xlsx", skip = 1)

set.seed(12345)
Sample_2015 <- Hitting_Stats_2015[sample(nrow(Hitting_Stats_2015), size=100), ]

set.seed(12345)
Sample_2016 <- Hitting_Stats_2016[sample(nrow(Hitting_Stats_2016), size=100), ]

set.seed(12345)
Sample_2017 <- Hitting_Stats_2017[sample(nrow(Hitting_Stats_2017), size=100), ]

set.seed(12345)
Sample_2018 <- Hitting_Stats_2018[sample(nrow(Hitting_Stats_2018), size=100), ]

set.seed(12345)
Sample_2019 <- Hitting_Stats_2019[sample(nrow(Hitting_Stats_2019), size=100), ]

set.seed(12345)
Sample_2020 <- Hitting_Stats_2020[sample(nrow(Hitting_Stats_2020), size=100), ]
```

```
set.seed(12345)
Sample_2021 <- Hitting_Stats_2021[sample(nrow(Hitting_Stats_2021), size=100), ]
```

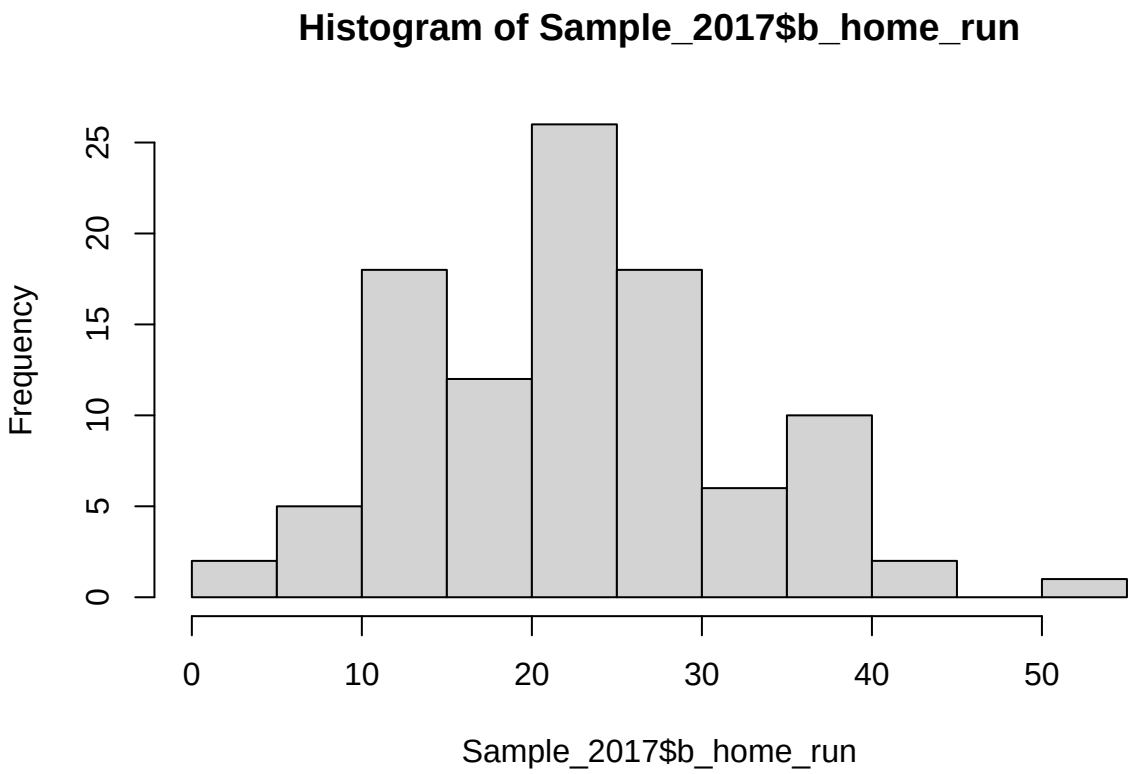
Graphs and Statistics

Home Runs

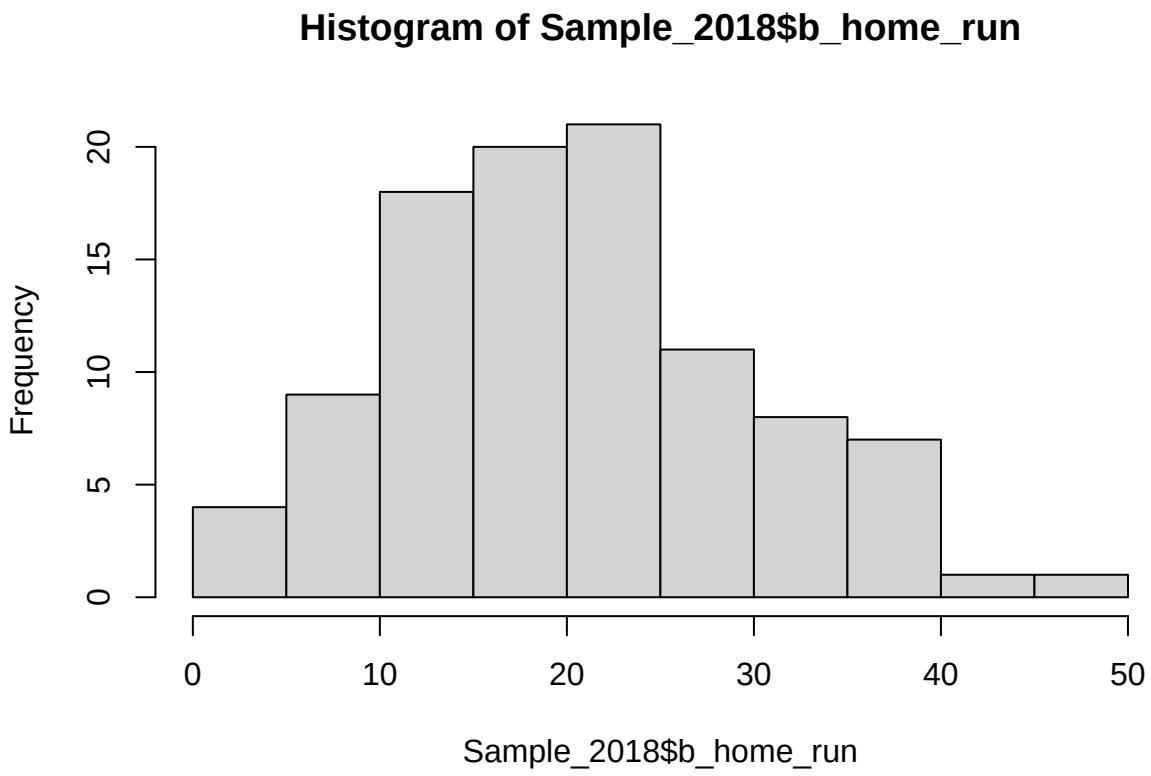
```
hist(Sample_2016$b_home_run)
```



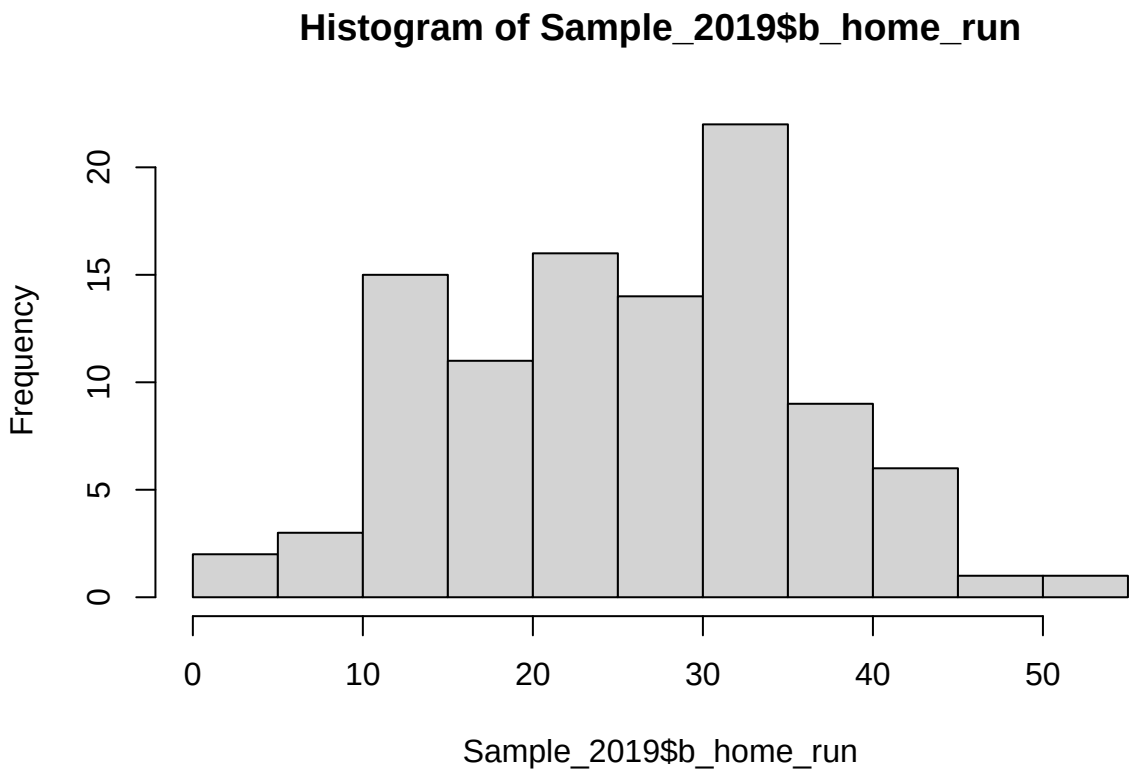

```
hist(Sample_2017$b_home_run)
```



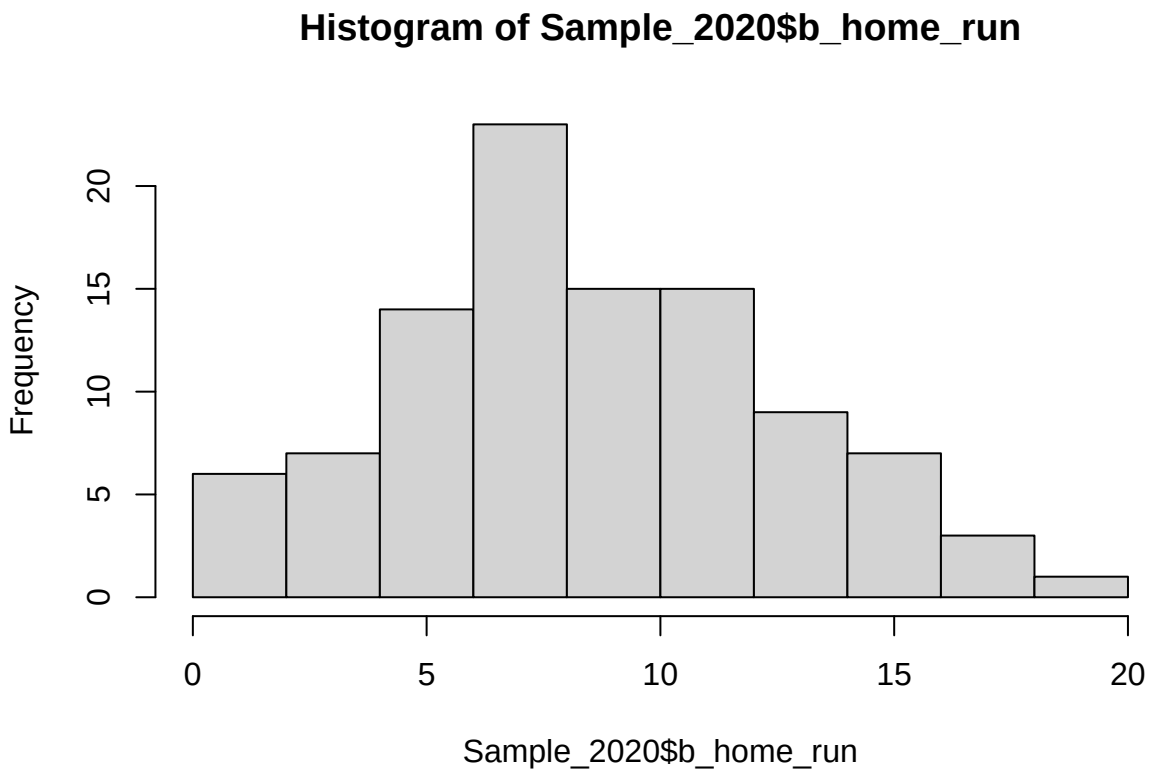

```
hist(Sample_2018$b_home_run)
```



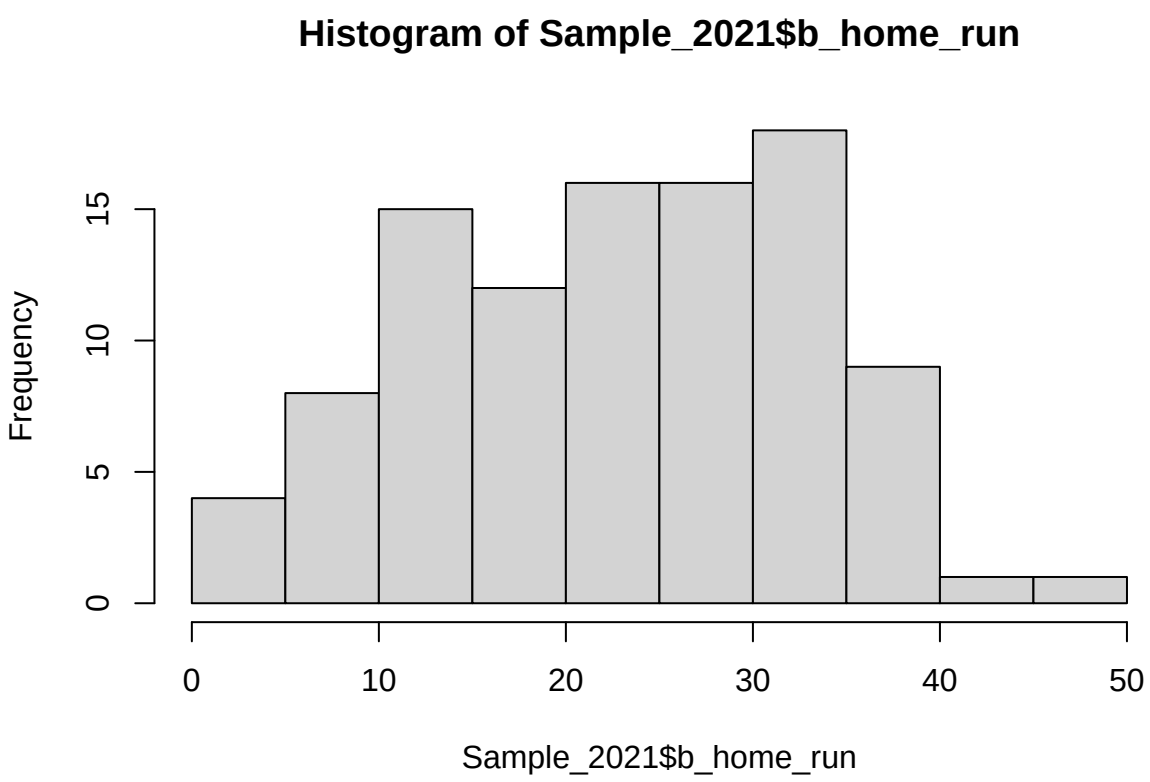

```
hist(Sample_2019$b_home_run)
```




```
hist(Sample_2020$b_home_run)
```




```
hist(Sample_2021$b_home_run)
```



Here, I visualized histograms of the number of home runs that each player hit in that season. This histogram shows bins for home runs and the frequency of players that hit the number of home runs for that specific bin. My goal here was to see if the distributions of home runs hit changed from year to year. From what I can see, all of the distributions here are unimodal, and fairly normal. It is difficult to see if there are significant shifts in the distributions.

Note: I took random samples of these, and set a seed, so these samples would be repeatable instead of random each time I used the command. I took samples in order to make the data sets the same size. Some of the data sets included data for 132 players, while another one may have had 146 players. I took a random sample of 100 players each year to make the data sets the same size. This way it would be easier to compare certain visualizations, such as histograms.

```
summary(Hitting_Stats_2015$b_home_run)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   11.00   17.00   18.39   23.00   47.00
```

```
summary(Hitting_Stats_2016$b_home_run)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0    14.0    22.0    22.1    29.0    47.0
```

```
summary(Hitting_Stats_2017$b_home_run)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   15.00   23.00   22.88   30.00   59.00
```

```
summary(Hitting_Stats_2018$b_home_run)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   14.00   20.00   20.76   26.00   48.00
```

```
summary(Hitting_Stats_2019$b_home_run)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   18.00   25.00   25.63   33.00   53.00
```

```
summary(Hitting_Stats_2020$b_home_run)
```

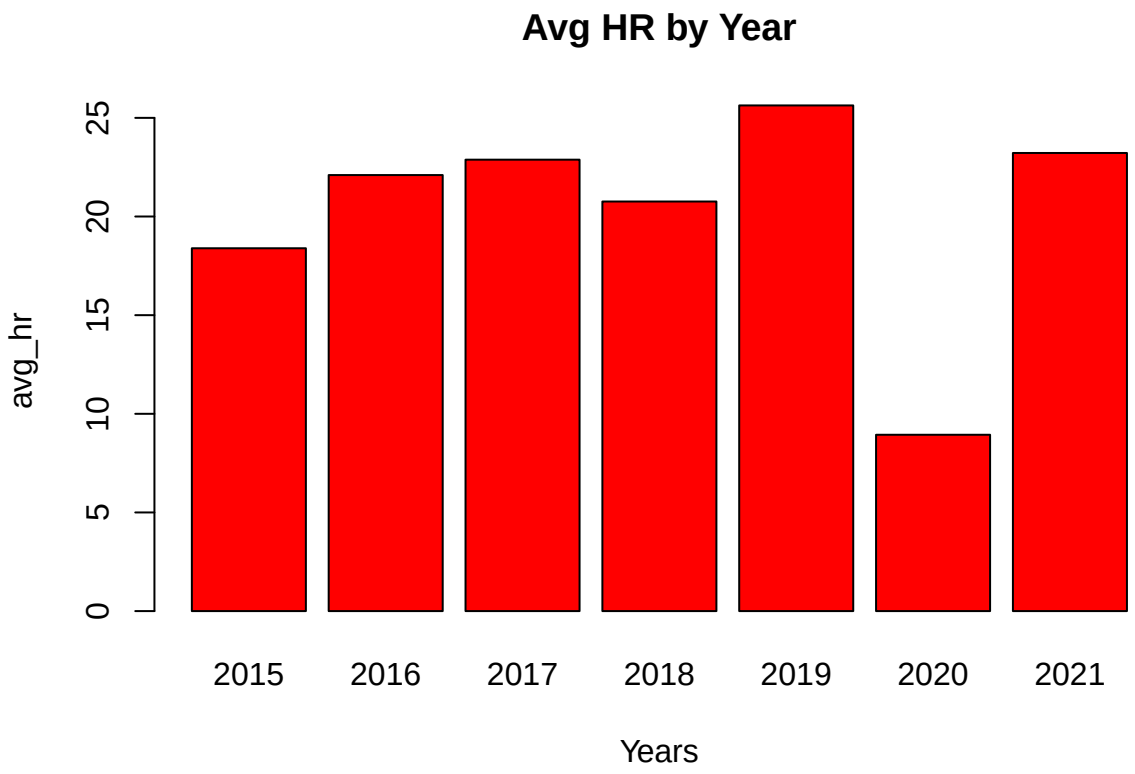
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.000   8.000   8.937  12.000  22.000
```

```
summary(Hitting_Stats_2021$b_home_run)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   15.00   23.00   23.22   31.00   48.00
```

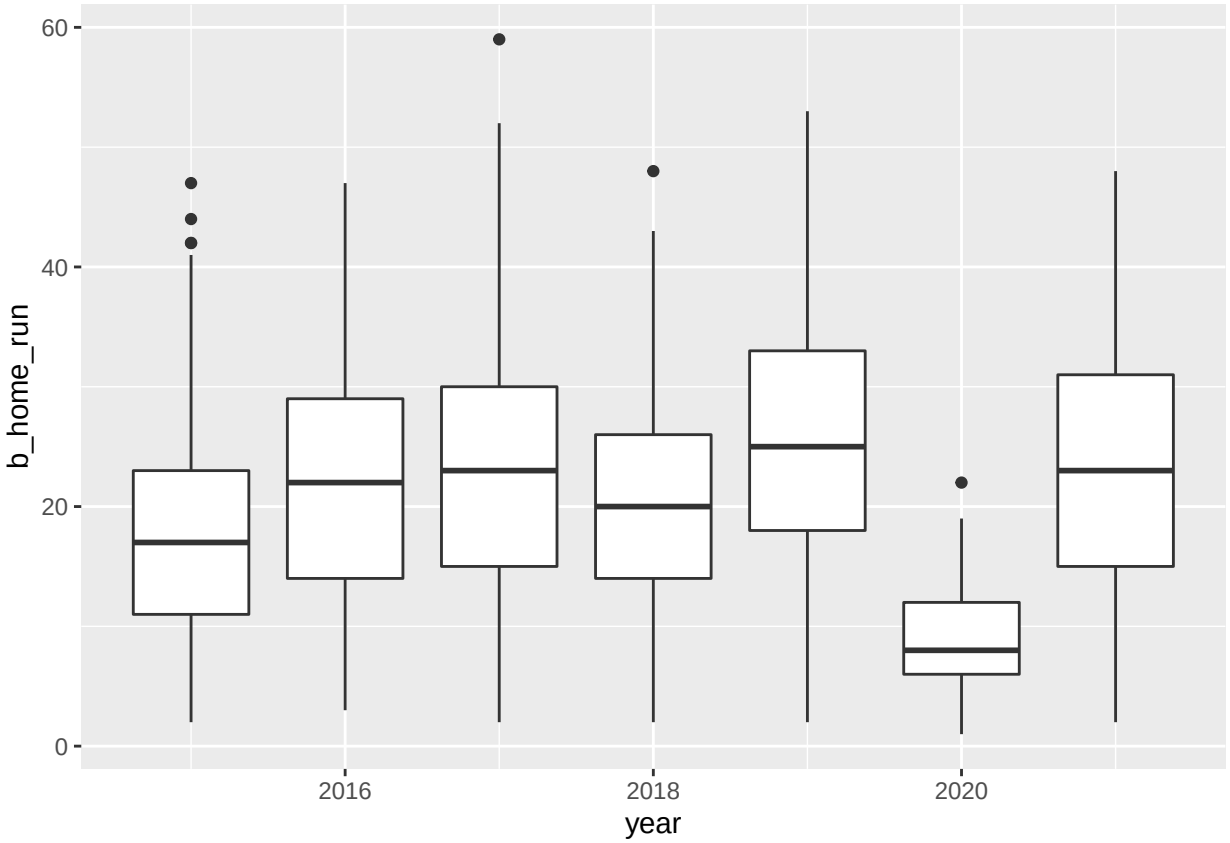
I found the summary statistics for the number of home runs hit each season by player. This allowed me to get a better picture if there were shifts from season to season.

```
avg_hr <- c(18.39, 22.1, 22.88, 20.76, 25.63, 8.937, 23.22)
Years <- c(2015, 2016, 2017, 2018, 2019, 2020, 2021)
barplot(avg_hr,names.arg=Years,xlab="Years",ylab="avg_hr",col="Red",main="Avg HR by Year",border="black
```



I used my summary statistics found above, specifically mean, to recognize changes in the data over the different seasons. For example, the highest mean # of home runs occured in the year 2019 with an average 25.63 home runs for all of the players included in the data set (Average 2.1 plate appearances per game). The home run numbers in 2020 are down due to COVID-19, which caused a lack of games. 2015 really sticks out to me as a year that didn't have many home runs in comparison. It seems that the home run numbers have trended upward over these 7 years.

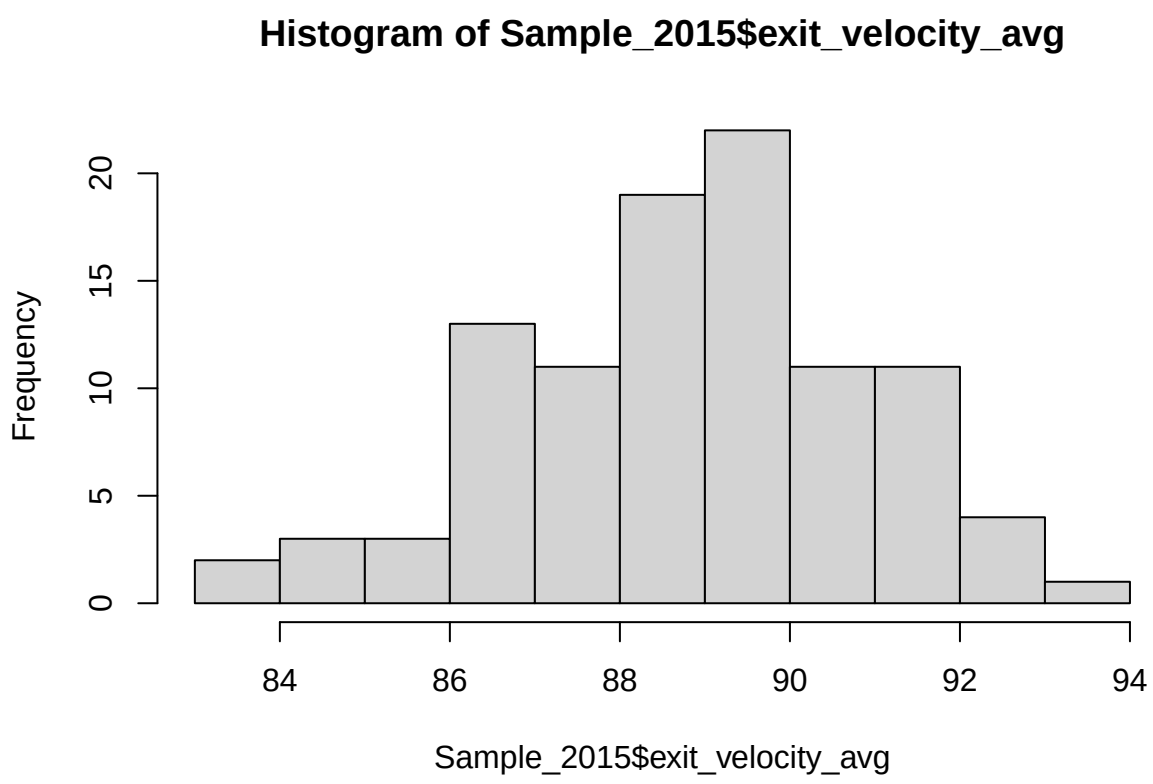
```
DataBind <- rbind(Hitting_Stats_2015, Hitting_Stats_2016, Hitting_Stats_2017, Hitting_Stats_2018, Hitti

ggplot(DataBind, aes(x=year, y=b_home_run)) + geom_boxplot(aes(group = year))
```



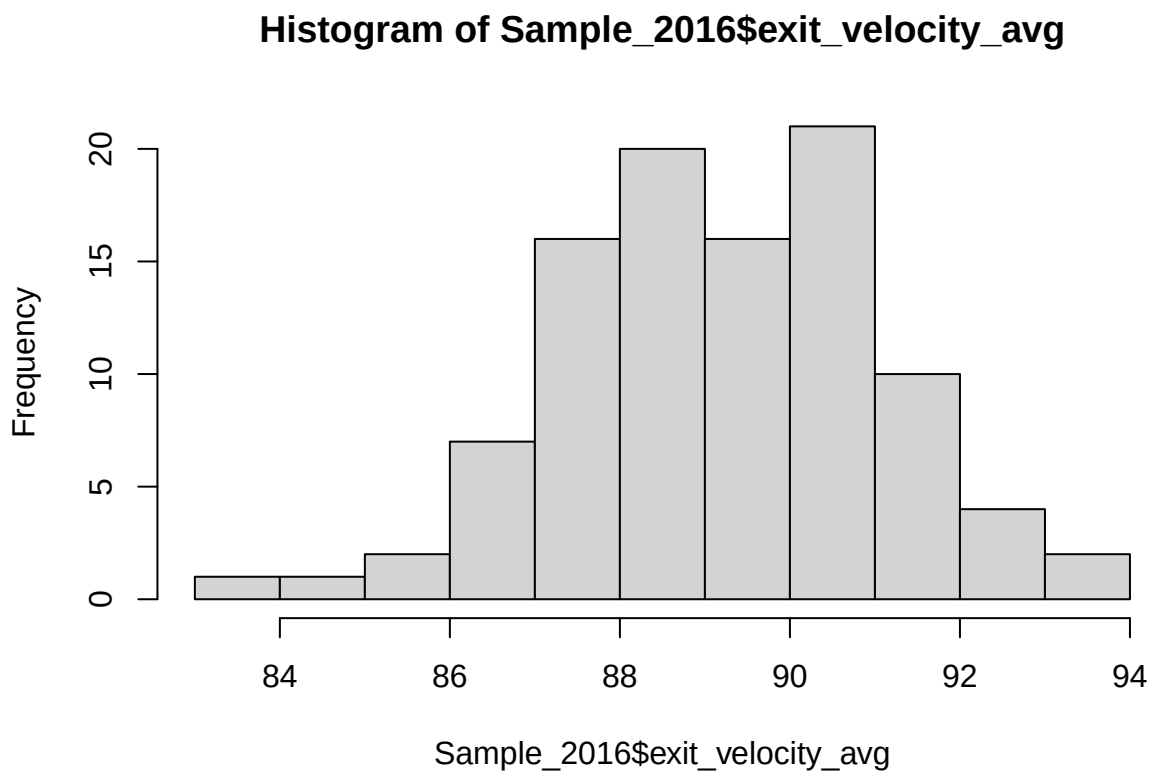
This plot helps visualize the summary statistics, such as the median, the first and third quartiles, as well as outliers that exist. This confirms the analysis from the bar plot that the most home runs occured in 2019 and the fewest occured in 2015. Also, there is asignificant outlier in the year 2017.In this year, a player had almost 60 home runs, which is very unusual.

Exit Velocity

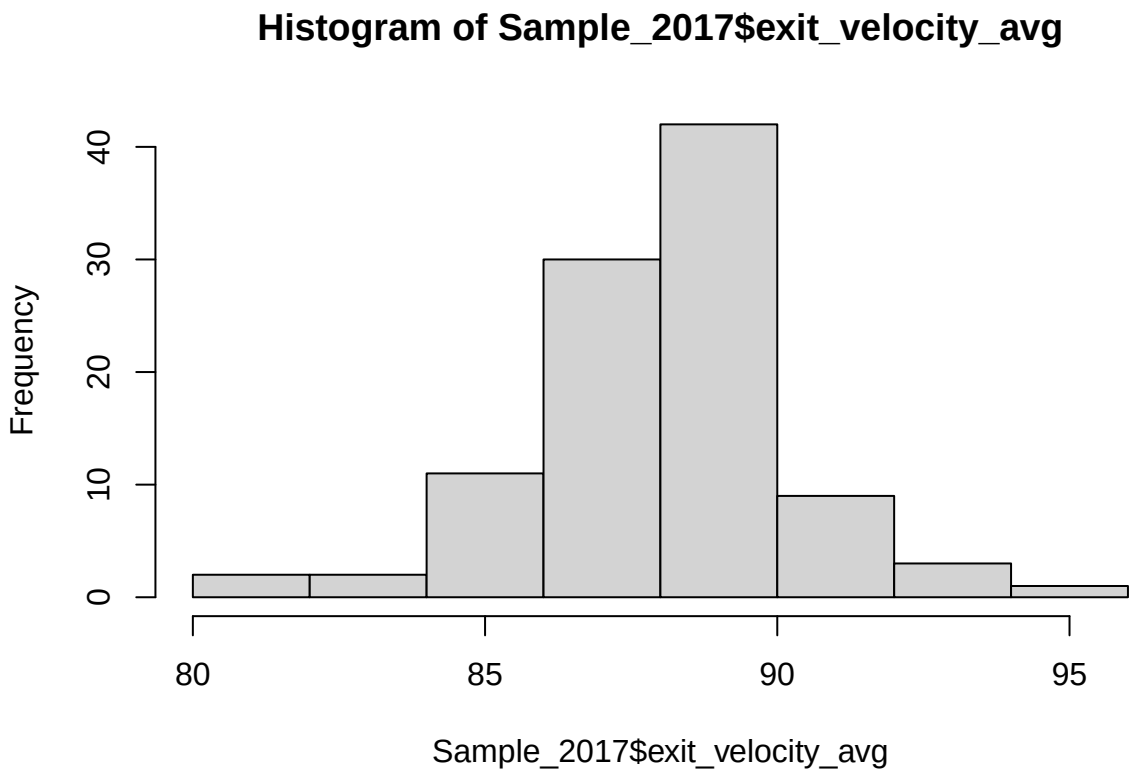
```
hist(Sample_2015$exit_velocity_avg)
```



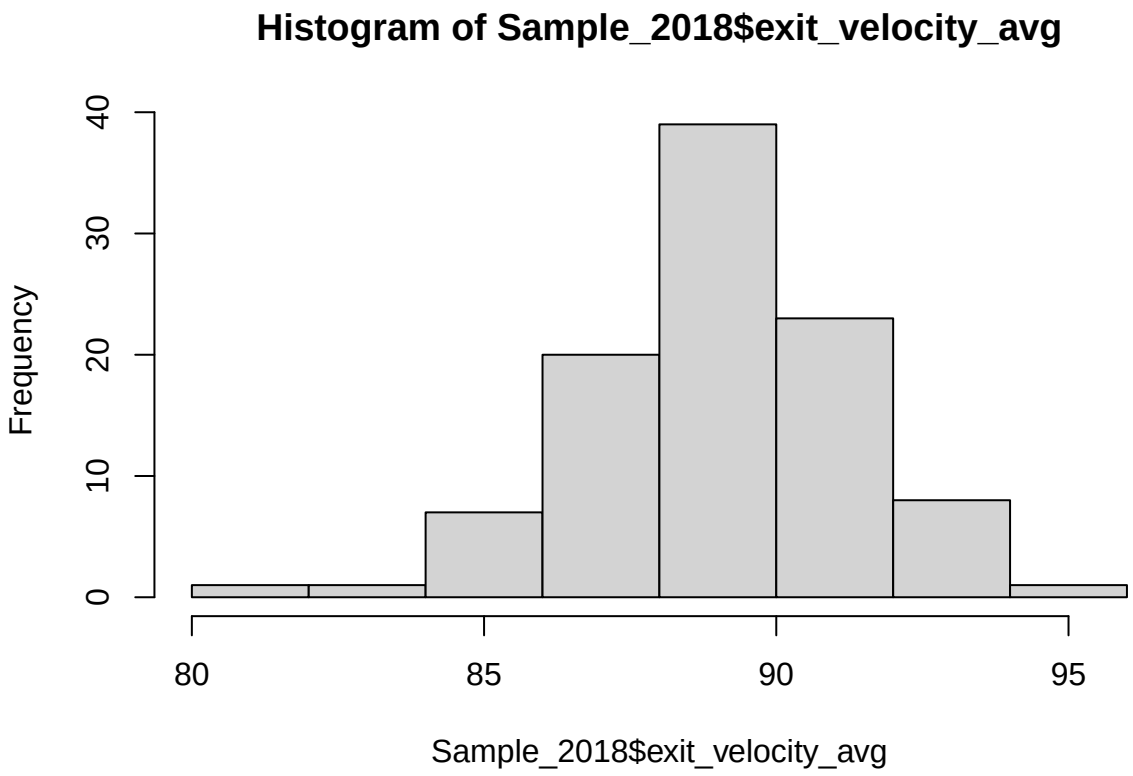

```
hist(Sample_2016$exit_velocity_avg)
```



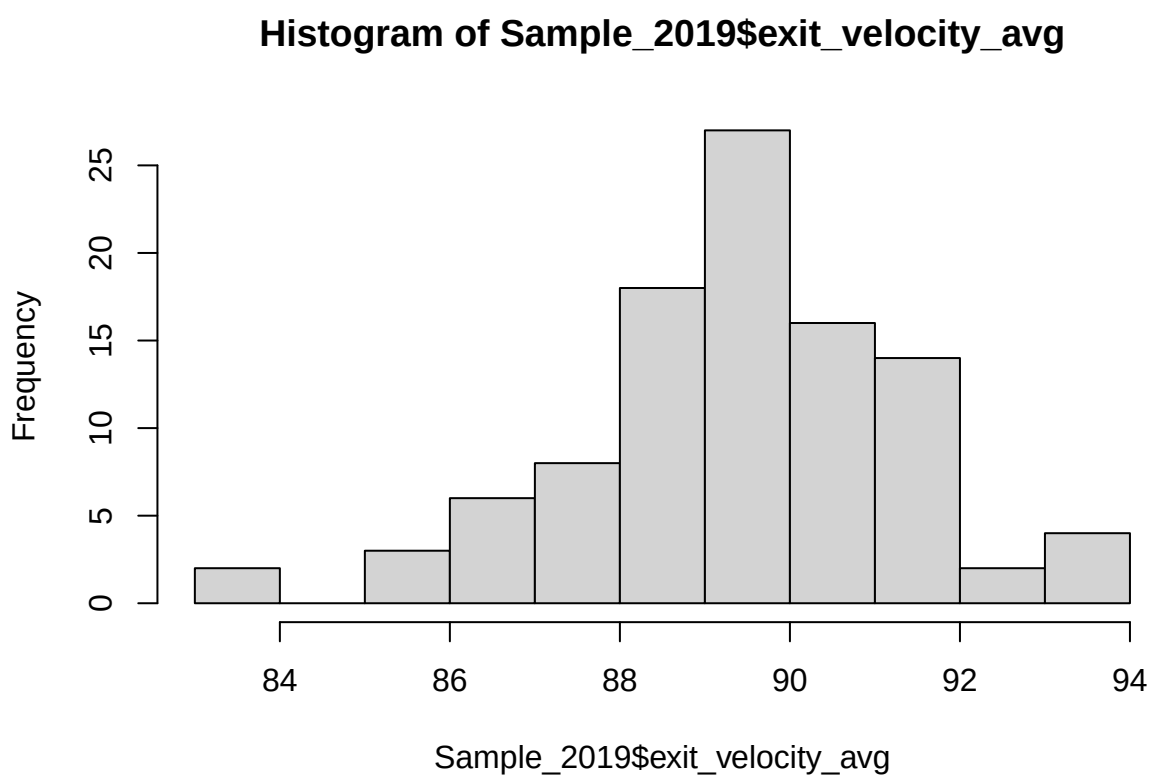

```
hist(Sample_2017$exit_velocity_avg)
```



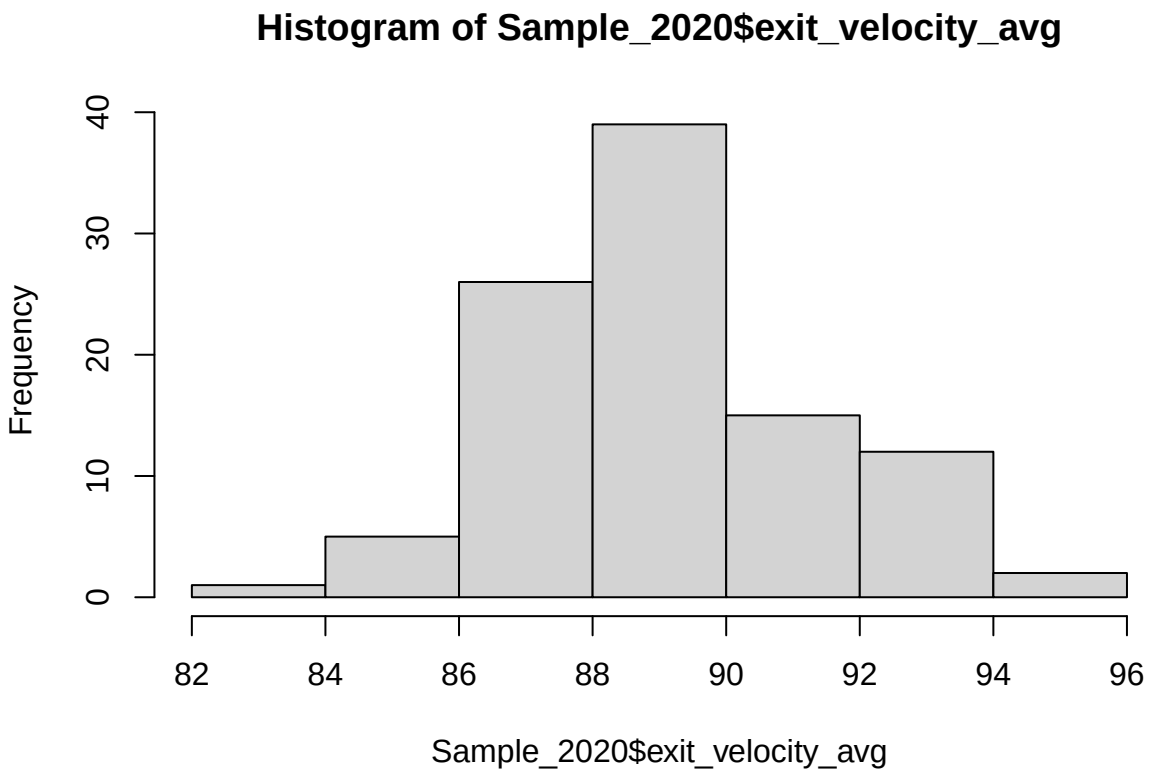

```
hist(Sample_2018$exit_velocity_avg)
```



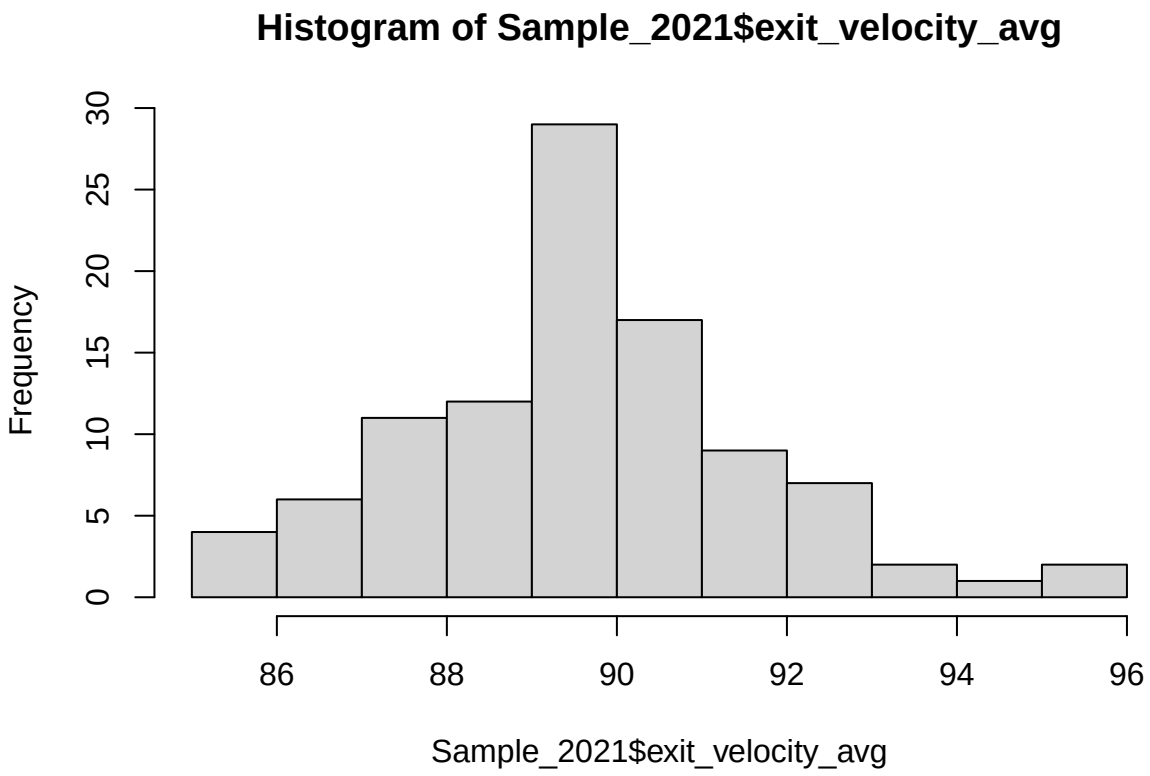

```
hist(Sample_2019$exit_velocity_avg)
```




```
hist(Sample_2020$exit_velocity_avg)
```




```
hist(Sample_2021$exit_velocity_avg)
```



These histograms look very similar to the histograms of homeruns mentioned earlier. They seem to be unimodal and normally distributed. It is hard to notice the differences in the distributions of exit velocity by year. It is hard to recognize significant shifts in the distributions between years.

```
summary(Hitting_Stats_2015$exit_velocity_avg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   83.10   87.30   89.00   88.86   90.28   93.80
```

```
summary(Hitting_Stats_2016$exit_velocity_avg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   83.60   87.92   89.30   89.21   90.60   94.60
```

```
summary(Hitting_Stats_2017$exit_velocity_avg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    80.5    87.2    88.3    88.1    89.2    94.9
```

```
summary(Hitting_Stats_2018$exit_velocity_avg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   80.50   87.80   89.10   89.09   90.60   94.40
```

```
summary(Hitting_Stats_2019$exit_velocity_avg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   83.30   88.45   89.50   89.45   90.90   93.70
```

```
summary(Hitting_Stats_2020$exit_velocity_avg)
```

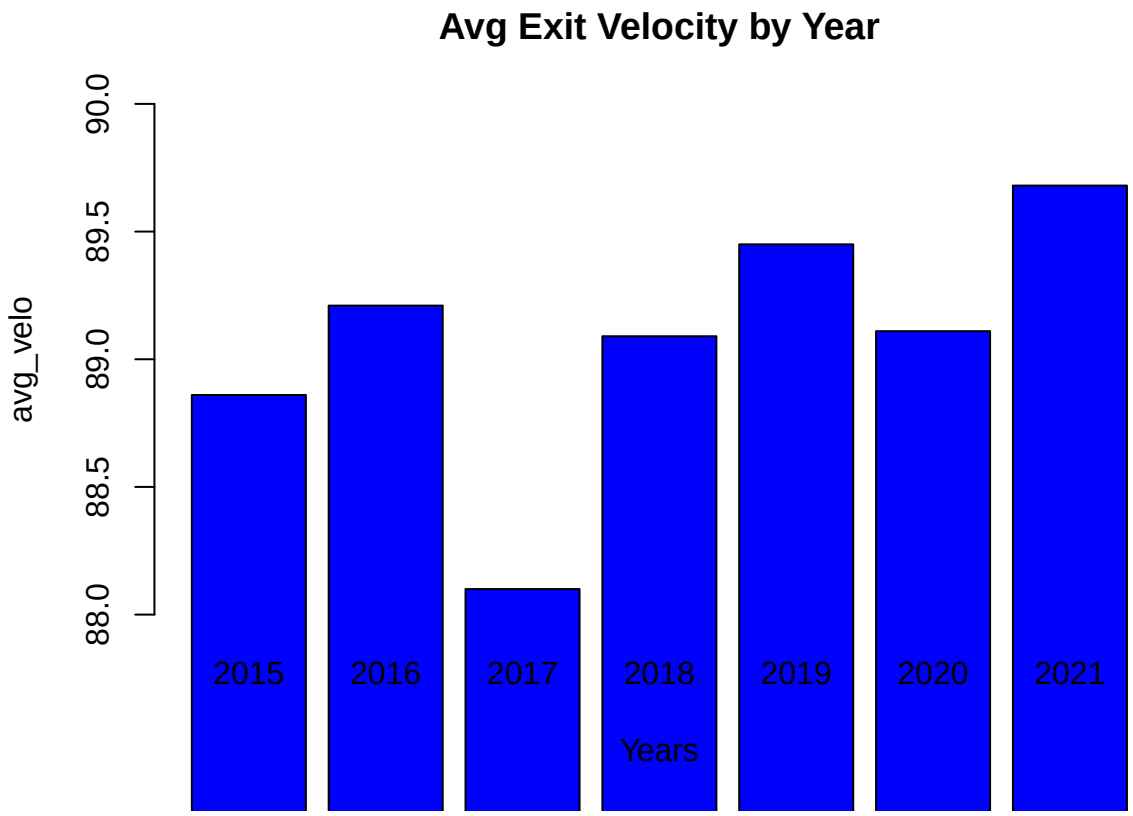
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   82.20   87.55   89.00   89.11   90.38   95.90
```

```
summary(Hitting_Stats_2021$exit_velocity_avg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   82.30   88.20   89.60   89.68   90.80   95.80
```

I found the summary statistics for the averave exit velocity by each player for every season. This allowed me to get a better picture if there were shifts from season to season.

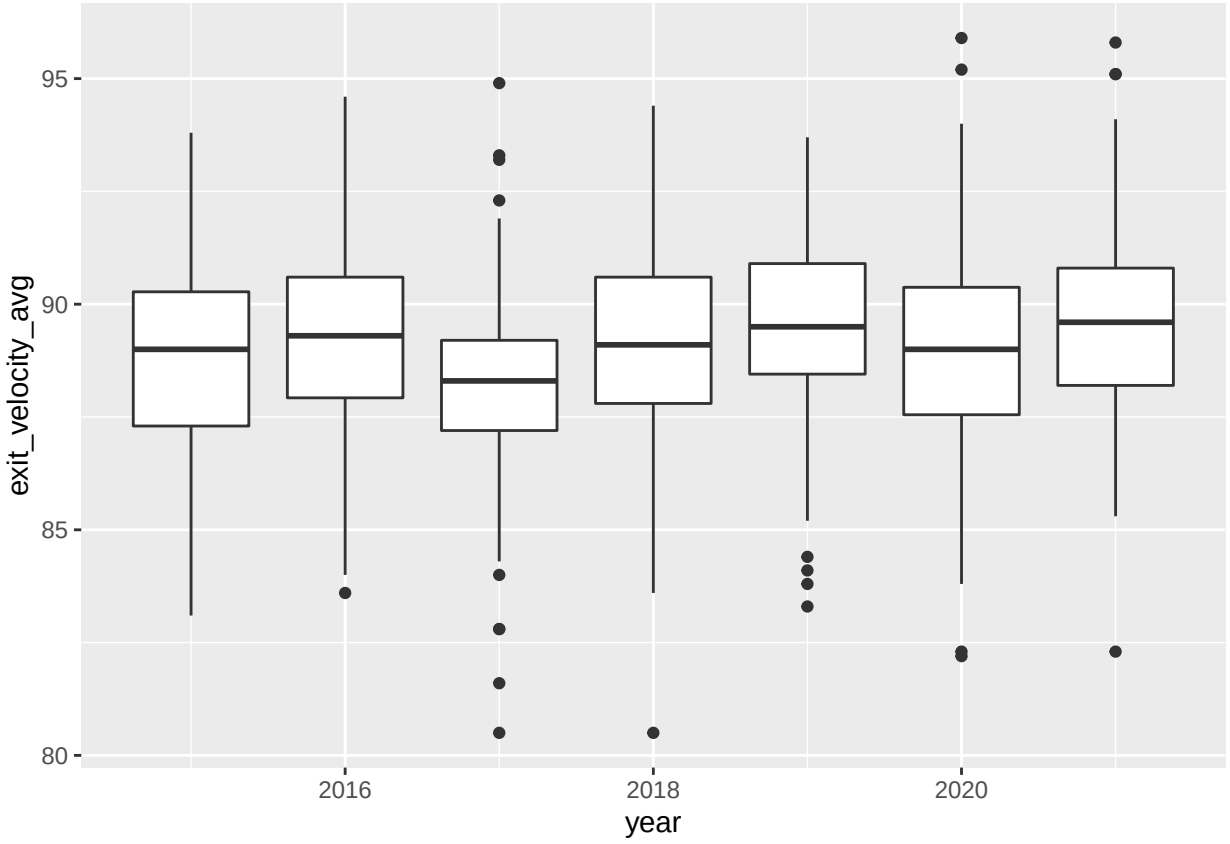
```
avg_velo <- c(88.86, 89.21, 88.1, 89.09, 89.45, 89.11, 89.68)
barplot(avg_velo,names.arg=Years,xlab="Years",ylab="avg_velo",col="Blue",main="Avg Exit Velocity by Yea
```



This shows the average exit velocity for all players in each season. It seems to have a general trend upward from 2015 to 2021. I changed the y-intercept so it was easier to note the difference between the years. It is

interesting to note that 2015 is not the lowest on average for exit velocity, even though they had the least average home runs. Also, the highest average exit velocity occcurs in 2021, and this was not the year with the highest average home run value.

```
ggplot(DataBind, aes(x=year, y=exit_velocity_avg)) + geom_boxplot(aes(group = year))
```



This plot helps visualize the summary statistics for averag exit velocity, such as the median, the first and third quartiles, as well as outliers that exist.

Exit Velocity vs Home Runs

```
plot(Hitting_Stats_2015$exit_velocity_avg, Hitting_Stats_2015$b_home_run)
```


```
cor(Hitting_Stats_2015$b_home_run, Hitting_Stats_2015$exit_velocity_avg)
```

```
## [1] 0.6673589
```

```
HR_Velo_2015.lm <- lm(b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2015)
summary(HR_Velo_2015.lm)
```

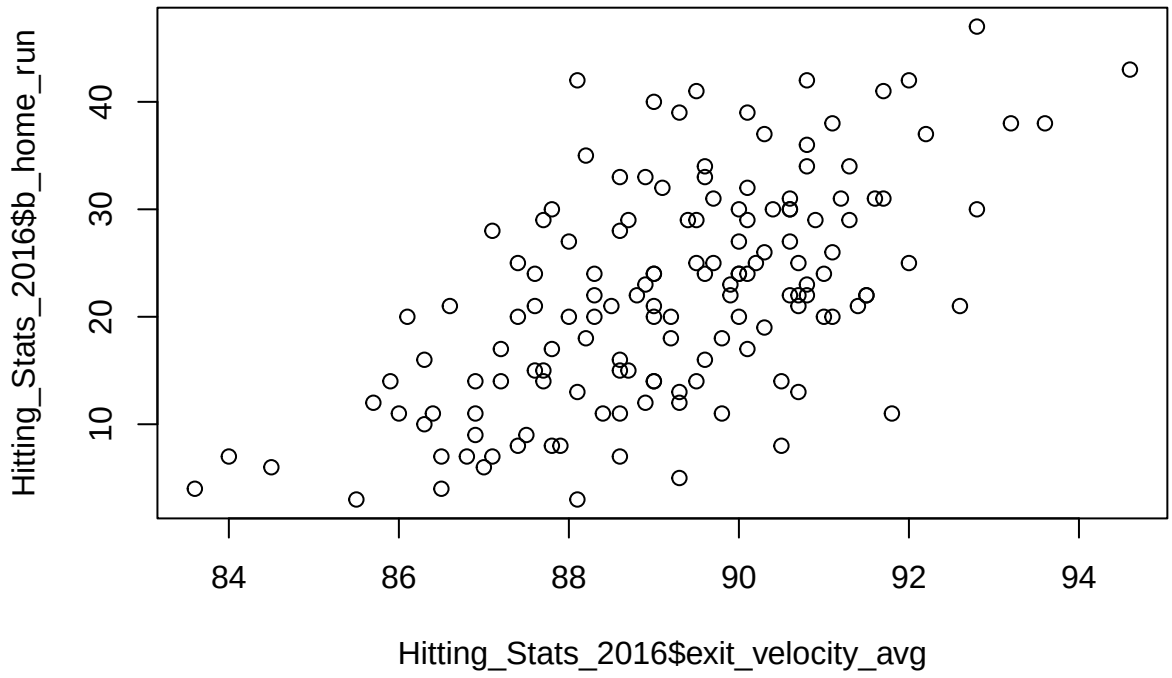
```
##
## Call:
## lm(formula = b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2015)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.6791  -5.1556  -0.6756   4.8544  18.3409
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -271.8807    27.3842  -9.928   <2e-16 ***
## exit_velocity_avg    3.2667     0.3081  10.603   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.743 on 140 degrees of freedom
## Multiple R-squared:  0.4454, Adjusted R-squared:  0.4414
## F-statistic: 112.4 on 1 and 140 DF,  p-value: < 2.2e-16
```

linear regression equation:

home runs = -271.88 + (3.27) * exit_velocity_avg

Interpretation: This means that if a player increased their season long exit velocity average by 1 mph, then they would hit 3.27 more home runs throughout the course of the season.

```
plot(Hitting_Stats_2016$exit_velocity_avg, Hitting_Stats_2016$b_home_run)
```




```
cor(Hitting_Stats_2016$b_home_run, Hitting_Stats_2016$exit_velocity_avg)
```

```
## [1] 0.6096707
```

```
HR_Velo_2016.lm <- lm(b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2016)
summary(HR_Velo_2016.lm)
```

```
##
## Call:
## lm(formula = b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2016)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.4573  -5.9765  -0.3544   4.7364  23.4885
##
## Coefficients:
```

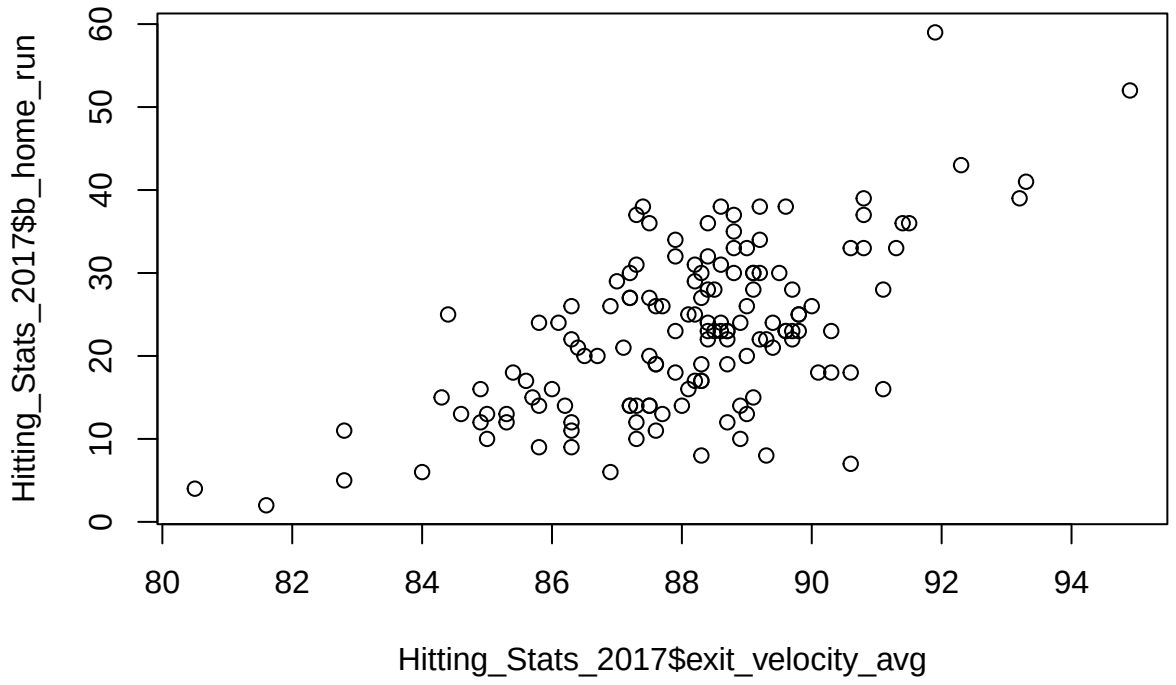
```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -265.9301    31.2139   -8.52 1.95e-14 ***
## exit_velocity_avg    3.2286     0.3498    9.23 3.16e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.04 on 144 degrees of freedom
## Multiple R-squared:  0.3717, Adjusted R-squared:  0.3673
## F-statistic: 85.19 on 1 and 144 DF,  p-value: 3.163e-16
```

linear regression equation:

home runs = -265.93 + (3.23) * exit_velocity_avg

Interpretation: This means that if a player increased their season long exit velocity average by 1 mph, then they would hit 3.23 more home runs throughout the course of the season.

```
plot(Hitting_Stats_2017$exit_velocity_avg, Hitting_Stats_2017$b_home_run)
```




```
cor(Hitting_Stats_2017$b_home_run, Hitting_Stats_2017$exit_velocity_avg)
```

```
## [1] 0.6186679
```

```
HR_Velo_2017.lm <- lm(b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2017)
summary(HR_Velo_2017.lm)
```

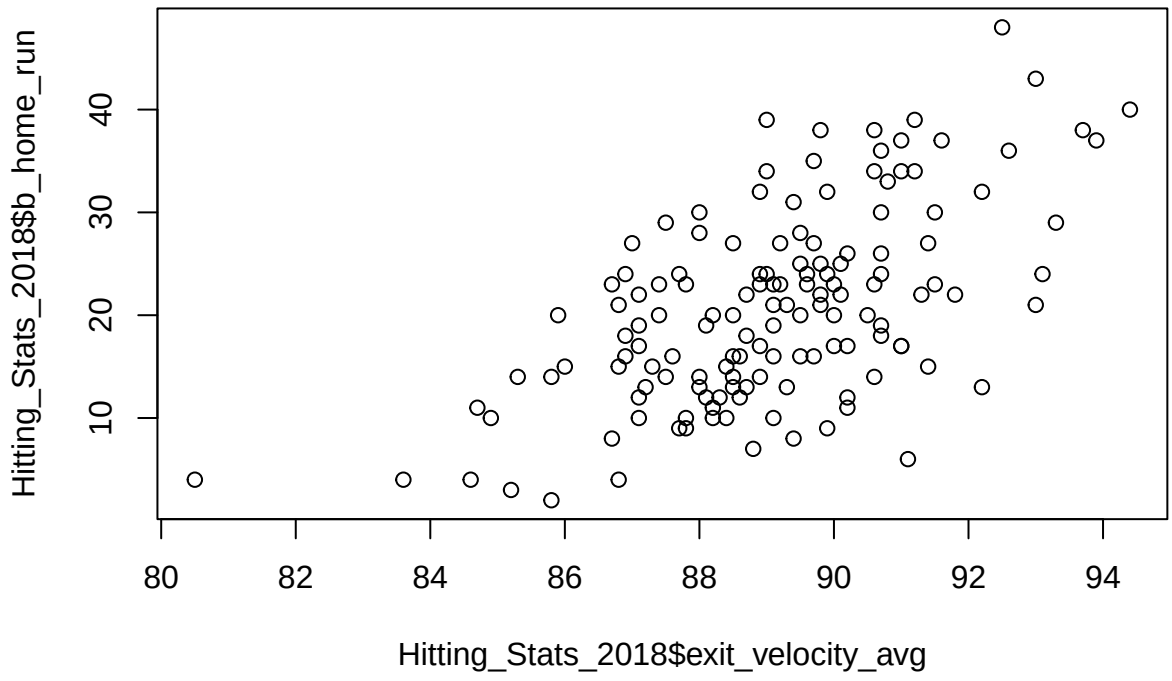
```
##
## Call:
## lm(formula = b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2017)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.0380  -5.2453  -0.5352   4.6798  25.2460
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -228.9376    26.8438  -8.529    2e-14 ***
## exit_velocity_avg    2.8585     0.3046   9.384   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.748 on 142 degrees of freedom
## Multiple R-squared:  0.3828, Adjusted R-squared:  0.3784
## F-statistic: 88.05 on 1 and 142 DF,  p-value: < 2.2e-16
```

linear regression equation:

home runs = -228.94 + (2.86) * exit_velocity_avg

Interpretation: This means that if a player increased their season long exit velocity average by 1 mph, then they would hit 2.86 more home runs throughout the course of the season.

```
plot(Hitting_Stats_2018$exit_velocity_avg, Hitting_Stats_2018$b_home_run)
```



```
cor(Hitting_Stats_2018$b_home_run, Hitting_Stats_2018$exit_velocity_avg)
```

```
## [1] 0.5989107
```

```
HR_Velo_2018.lm <- lm(b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2018)
summary(HR_Velo_2018.lm)
```

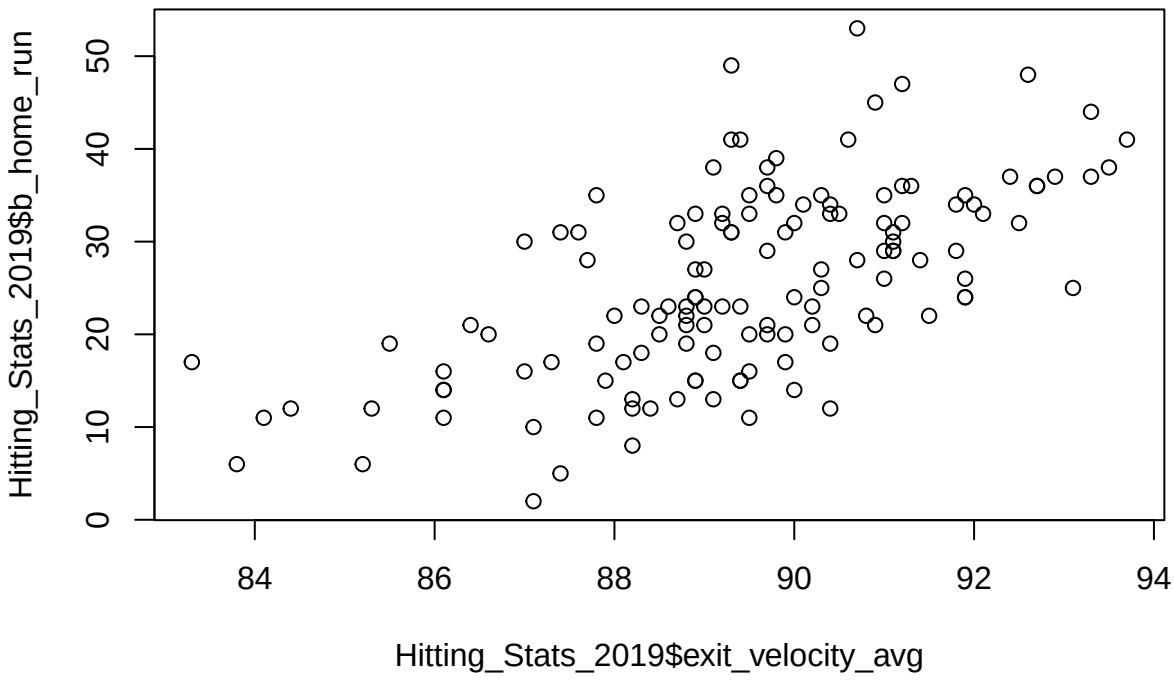
```
##
## Call:
## lm(formula = b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2018)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.0758  -6.0191   0.2074   5.2069  18.4716
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -214.5739    26.6974  -8.037 3.57e-13 ***
## exit_velocity_avg    2.6416     0.2996   8.817 4.34e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.544 on 139 degrees of freedom
## Multiple R-squared:  0.3587, Adjusted R-squared:  0.3541
## F-statistic: 77.75 on 1 and 139 DF,  p-value: 4.345e-15
```

linear regression equation:

home runs = -214.57 + (2.64) * exit_velocity_avg

Interpretation: This means that if a player increased their season long exit velocity average by 1 mph, then they would hit 2.64 more home runs throughout the course of the season.

```
plot(Hitting_Stats_2019$exit_velocity_avg, Hitting_Stats_2019$b_home_run)
```




```
cor(Hitting_Stats_2019$b_home_run, Hitting_Stats_2019$exit_velocity_avg)
```

```
## [1] 0.62123
```

```
HR_Velo_2019.lm <- lm(b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2019)
summary(HR_Velo_2019.lm)
```

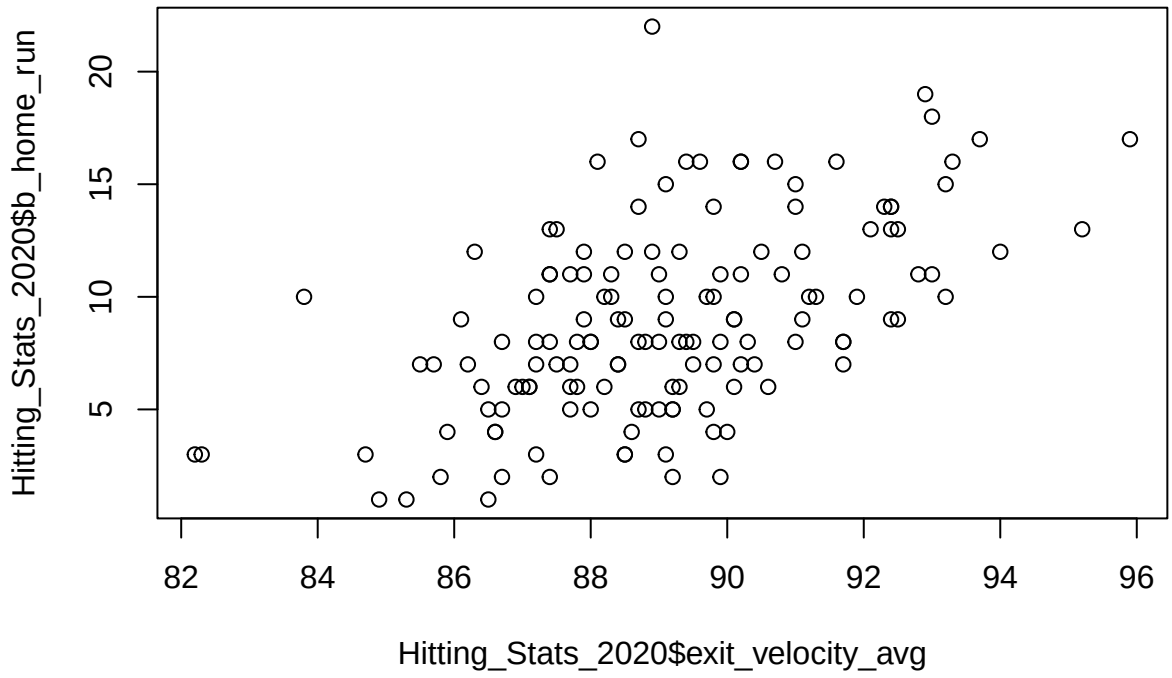
```
##
## Call:
## lm(formula = b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2019)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5653  -5.5825  -0.6687   4.9003  23.8485
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -251.9861    30.3731  -8.296 1.05e-13 ***
## exit_velocity_avg    3.1034     0.3395   9.143 9.07e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.036 on 133 degrees of freedom
## Multiple R-squared:  0.3859, Adjusted R-squared:  0.3813
## F-statistic: 83.59 on 1 and 133 DF,  p-value: 9.07e-16
```

linear regression equation:

home runs = -251.99 + (3.10) * exit_velocity_avg

Interpretation: This means that if a player increased their season long exit velocity average by 1 mph, then they would hit 3.10 more home runs throughout the course of the season.

```
plot(Hitting_Stats_2020$exit_velocity_avg, Hitting_Stats_2020$b_home_run)
```




```
cor(Hitting_Stats_2020$b_home_run, Hitting_Stats_2020$exit_velocity_avg)
```

```
## [1] 0.5454801
```

```
HR_Velo_2020.lm <- lm(b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2020)
summary(HR_Velo_2020.lm)
```

```
##
## Call:
## lm(formula = b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2020)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.721 -2.603 -0.539  1.947 13.268
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -79.2455    11.4551  -6.918 1.50e-10 ***
## exit_velocity_avg   0.9896     0.1285   7.701 2.23e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.605 on 140 degrees of freedom
## Multiple R-squared:  0.2975, Adjusted R-squared:  0.2925
## F-statistic:  59.3 on 1 and 140 DF,  p-value: 2.226e-12
```

linear regression equation:

home runs = -79.25 + (0.99) * exit_velocity_avg

Interpretation: This means that if a player increased their season long exit velocity average by 1 mph, then they would hit 0.99 more home runs throughout the course of the season.

```
plot(Hitting_Stats_2021$exit_velocity_avg, Hitting_Stats_2021$b_home_run)
```


```
cor(Hitting_Stats_2021$b_home_run, Hitting_Stats_2021$exit_velocity_avg)
```

```
## [1] 0.6859323
```

```
HR_Velo_2021.lm <- lm(b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2021)
summary(HR_Velo_2021.lm)
```

```
##
## Call:
## lm(formula = b_home_run ~ exit_velocity_avg, data = Hitting_Stats_2021)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2141  -5.1342  -0.2803   5.0441  21.7307
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -270.2331    27.3116  -9.894   <2e-16 ***
## exit_velocity_avg    3.2720     0.3044  10.748   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.676 on 130 degrees of freedom
## Multiple R-squared:  0.4705, Adjusted R-squared:  0.4664
## F-statistic: 115.5 on 1 and 130 DF,  p-value: < 2.2e-16
```

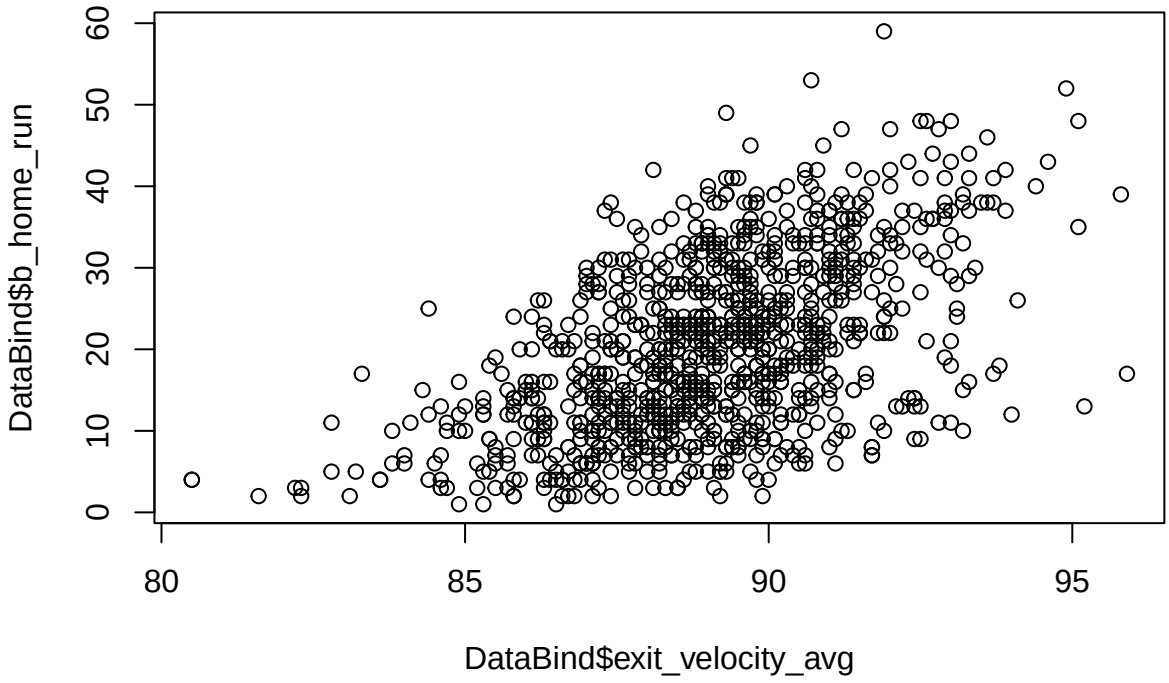
linear regression equation:

home runs = -270.23 + (3.27) * exit_velocity_avg

Interpretation: This means that if a player increased their season long exit velocity average by 1 mph, then they would hit 3.27 more home runs throughout the course of the season.

Here are plots of average exit velocity vs home runs. All of these correlations have are positive and moderately strong. I did not plot the the summary of the linear regression because it ultimately made the document extremely long. This would have included more information about the residuals of each model.

```
plot(DataBind$exit_velocity_avg, DataBind$b_home_run)
```




```
cor(DataBind$exit_velocity_avg, DataBind$b_home_run)
```

```
## [1] 0.5288894
```

```
HR_Velo.lm <- lm(b_home_run ~ exit_velocity_avg, data = DataBind)
summary(HR_Velo.lm)
```

```
##
## Call:
## lm(formula = b_home_run ~ exit_velocity_avg, data = DataBind)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.1999  -6.1107  -0.1219   6.1806  31.3907
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -211.6264    11.8877  -17.80   <2e-16 ***
## exit_velocity_avg    2.6032     0.1334   19.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.089 on 980 degrees of freedom
## Multiple R-squared:  0.2797, Adjusted R-squared:  0.279
## F-statistic: 380.6 on 1 and 980 DF,  p-value: < 2.2e-16
```

Here is a scatter plot of all the years combined into one visualization, as well as a linear regression model.

linear regression equation:

home runs = -211.63 + (2.60) * exit_velocity_avg

Interpretation: This means that if a player increased their season long exit velocity average by 1 mph, then they would hit 2.60 more home runs throughout the course of the season.

Topics From Class

R Markdown

In this course, we used a ton of R Markdown in order to export our results from RStudio. We learned how to add R Chunks into Markdown as well as formatting titles and adding comments. This allowed us to generate good looking PDF documents that would summarize our findings in homework assignments and projects. I think I would find much use out of this software beyond class.

GitHub

During the creation of this project, I was exposed to GitHub as well. This was a place where we could store and share our project with the rest of the class, as well as the public. I do not have much experience with GitHub, so it was very helpful that we were able to walk through the instructions of how to navigate the site and upload projects. I also think an extremely neat feature of GitHub is the collaboration aspect of it. If someone is helping review your project, they can send you suggestions to change certain things about your code, and you can accept these changes and merge them into your project. We did not spend much time on GitHub, but I plan to explore it more outside of class.

Summarizing Data With Graphs

Throughout this class I learned many different ways to visualize data of importance. Some of these graphing methods included histograms, bar plots, scatter plots, etc. I used histograms in this project to analyze the distribution of data. I used bar plots in order to compare different metrics of interest and highlight values that were higher or lower than others. I used scatter plots in order to visualize the relationship between two variables to see what type of correlation existed. This feature is important to understanding relationships between explanatory and response variables.

Sampling

During this class, we learned many different examples of sampling a population of data. We learned when and why you would use certain samples. In this project, I used sampling in order to make sure I understood how to use the command. I also used it in order to make the number of players in the dataset the same for all of the different years. Another feature we learned about sampling was the ability to set a seed. This allows you the ability to repeat random sampling and recieve the same random sample each time you run the code. Otherwise, it would be a different random sample each time you run the command.

Linear Regression

The last topic that we learned about during class was linear regression. This topic is a continuation of the prior discussion of relationships and scatter plots. Linear regression is extremely important to assess

a relationship between two or more variables. In this case, I only analyzed the relationship between two variables: Home Runs and Average Exit Velocity. The response variable here was home runs, while the explanatory variable was average exit velocity. I wanted to see how an increase in average exit velocity could potentially increase the number of home runs that a player hits throughout a season. We also learned about multiple regression, which includes more variables than just the two.

Conclusion

Overall, I explored and gained a further understanding of multiple different topics we discussed during this course. I think this project was a great learning experience because it forced me to explore these topics without much guidance. It allowed me to look at a data set and try to figure out how I could visualize it and summarize important findings. Ultimately, I did not discover as much about mlb hitting stats as I set out to, but I think the project was a success in terms of my experience in RStudio, GitHub, and R Markdown. I think that I could have done some more creative visualizations if i had access to a dataset that included the hitting statistics of every single at bat throughout the mlb season (specifically, exit velocity, launch angle, and the outcome of the at bat). Unfortunately, I only could find access to the aggregate data for these newer statistics.

References

Baseball Savant