Exploratory data analysis

Andres Izquierdo

2/16/2022

```
# Data cleaning and set up.
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(tsibble)
```

```
##
## Attaching package: 'tsibble'

## The following object is masked from 'package:lubridate':
##
##     interval

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, union
```

```r
library(stats)

path <- here::here("Lahman-master", "data", "Batting.RData")

load(path)

Start_Year = 1955

# removing missing statistics
Batting.clean <- na.omit(Batting)

# Combining stats for players that were on different teams during the same year.
Batting.merge <- group_by(Batting.clean, playerID, yearID) %>% mutate(G = sum(G), AB = sum(AB), R = sum
  ungroup()

# Setting UP Batting Average Statistic
Batting.merge <- Batting.merge %>% mutate(Bavg = H/AB)

# Replacing NaN for all the times a batter got 0 hits and 0 at bats, can't divide 0/0, so replacing wit
Batting.merge[is.na(Batting.merge)] = 0

# Setting UP Slugging Percentage Statistic
Batting.merge <- Batting.merge %>% mutate(SLG = (H+X2B*2+X3B*3+HR*4)/AB)

# Setting UP On Base Percentage Statistic
Batting.merge <- Batting.merge %>% mutate(OBP = (H+BB+HBP)/(AB+BB+HBP+SF))

# Setting UP Slugging Percentage Statistic
Batting.merge <- Batting.merge %>% mutate(OPS = (OBP+SLG))

# removing missing statistics after OPS, SLG, and OPS calculations
Batting.merge <- na.omit(Batting.merge)

# Getting rid of duplicates
Batting.merge <- Batting.merge [!duplicated(Batting.merge[c(1,2)]),]

# keeping statistics from 1955 on as that is when all statistics started to be tracked.
Batting.merge <- Batting.merge %>% filter(yearID >= Start_Year)
Batting.merge %>%
  mutate(yearID = lubridate::as_date(yearID)) %>%
  mutate(playerID = as.factor(playerID)) %>%
  mutate(yearID   = as_date(yearID)) %>%
  as_tsibble(key = playerID, index = yearID)
```

```
## # A tsibble: 51,417 x 26 [1D]
## # Key:       playerID [9,685]
##    playerID  yearID     stint teamID lgID      G    AB     R     H   X2B   X3B
##    <fct>     <date>     <int> <fct>  <fct> <int> <int> <int> <int> <int> <int>
##  1 aardsda01 1975-06-30     1 CHN    NL       45     2     0     0     0     0
##  2 aardsda01 1975-07-02     1 BOS    AL       47     1     0     0     0     0
##  3 aardsda01 1975-07-09     1 ATL    NL       33     1     0     0     0     0
##  4 aaronha01 1975-05-10     1 ML1    NL      153   602   105   189    37     9
##  5 aaronha01 1975-05-11     1 ML1    NL      153   609   106   200    34    14
```

```
##  6 aaronha01 1975-05-12     1 ML1    NL      151   615   118   198    27     6
##  7 aaronha01 1975-05-13     1 ML1    NL      153   601   109   196    34     4
##  8 aaronha01 1975-05-14     1 ML1    NL      154   629   116   223    46     7
##  9 aaronha01 1975-05-15     1 ML1    NL      153   590   102   172    20    11
## 10 aaronha01 1975-05-16     1 ML1    NL      155   603   115   197    39    10
## # ... with 51,407 more rows, and 15 more variables: HR <int>, RBI <int>,
## #   SB <int>, CS <int>, BB <int>, SO <int>, IBB <int>, HBP <dbl>, SH <int>,
## #   SF <int>, GIDP <int>, Bavg <dbl>, SLG <dbl>, OBP <dbl>, OPS <dbl>
```

```
Batting_tsbl <- as_tsibble(Batting.merge, key = playerID, index = yearID)
Batting_tsbl
```

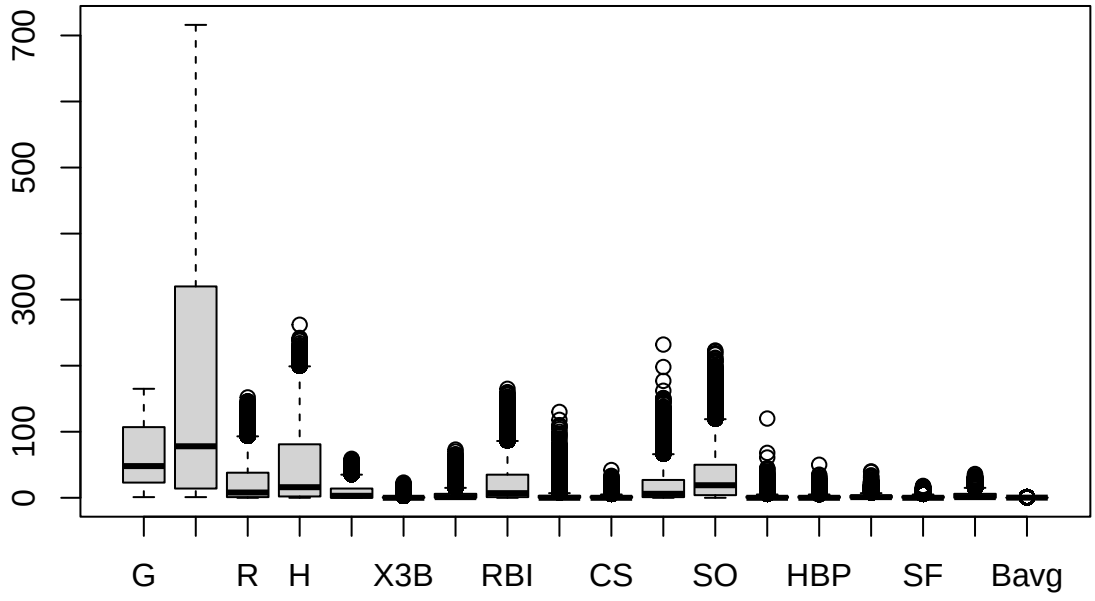
```
## # A tsibble: 51,417 x 26 [1Y]
## # Key:       playerID [9,685]
##    playerID  yearID stint teamID lgID      G    AB     R     H   X2B   X3B    HR
##    <chr>      <int> <int> <fct>  <fct> <int> <int> <int> <int> <int> <int> <int>
##  1 aardsda01   2006     1 CHN    NL       45     2     0     0     0     0     0
##  2 aardsda01   2008     1 BOS    AL       47     1     0     0     0     0     0
##  3 aardsda01   2015     1 ATL    NL       33     1     0     0     0     0     0
##  4 aaronha01   1955     1 ML1    NL      153   602   105   189    37     9    27
##  5 aaronha01   1956     1 ML1    NL      153   609   106   200    34    14    26
##  6 aaronha01   1957     1 ML1    NL      151   615   118   198    27     6    44
##  7 aaronha01   1958     1 ML1    NL      153   601   109   196    34     4    30
##  8 aaronha01   1959     1 ML1    NL      154   629   116   223    46     7    39
##  9 aaronha01   1960     1 ML1    NL      153   590   102   172    20    11    40
## 10 aaronha01   1961     1 ML1    NL      155   603   115   197    39    10    34
## # ... with 51,407 more rows, and 14 more variables: RBI <int>, SB <int>,
## #   CS <int>, BB <int>, SO <int>, IBB <int>, HBP <dbl>, SH <int>, SF <int>,
## #   GIDP <int>, Bavg <dbl>, SLG <dbl>, OBP <dbl>, OPS <dbl>
```

```
Batting.num <- Batting.merge[,c(6:23)]
summary(Batting.num)
```

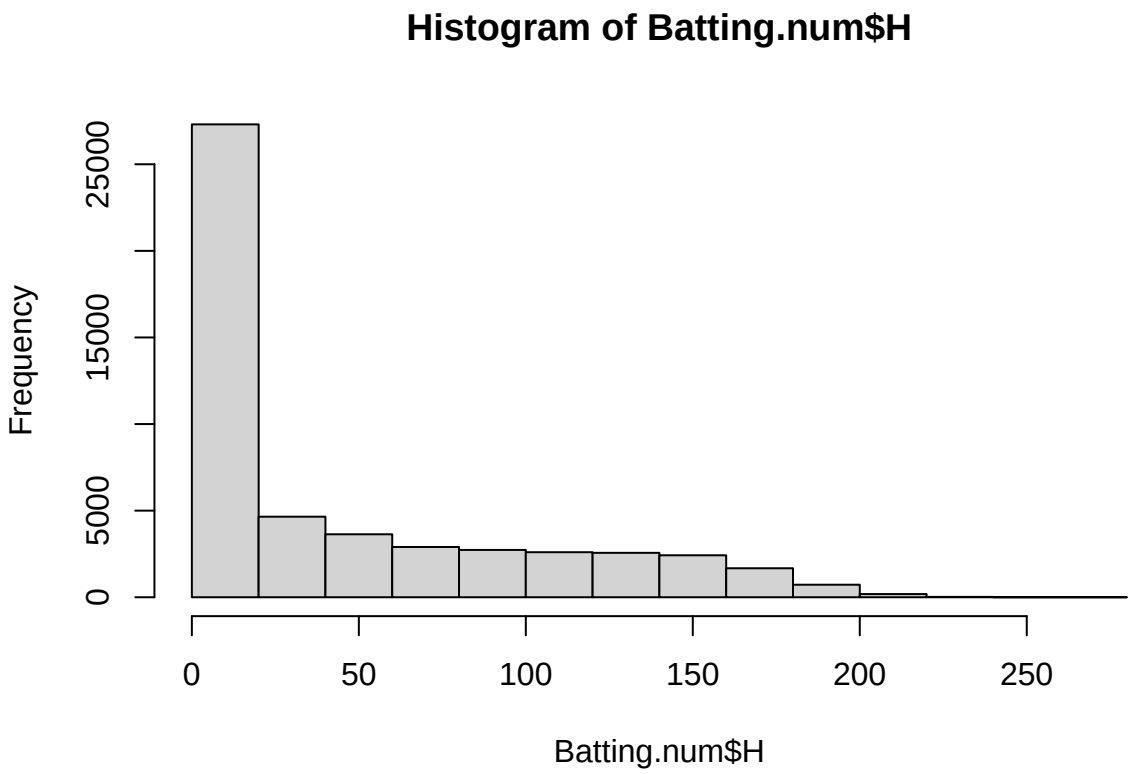
```
##        G                AB               R                H         
##  Min.   :  1.00   Min.   :  1.0   Min.   :  0.00   Min.   :  0.00  
##  1st Qu.: 23.00   1st Qu.: 14.0   1st Qu.:  1.00   1st Qu.:  2.00  
##  Median : 48.00   Median : 78.0   Median :  8.00   Median : 16.00  
##  Mean   : 64.13   Mean   :176.8   Mean   : 22.89   Mean   : 45.68  
##  3rd Qu.:107.00   3rd Qu.:320.0   3rd Qu.: 38.00   3rd Qu.: 81.00  
##  Max.   :165.00   Max.   :716.0   Max.   :152.00   Max.   :262.00  
##       X2B              X3B               HR              RBI        
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :  0.00  
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:  1.00  
##  Median : 3.000   Median : 0.000   Median : 1.000   Median :  7.00  
##  Mean   : 8.229   Mean   : 1.092   Mean   : 4.815   Mean   : 21.59  
##  3rd Qu.:14.000   3rd Qu.: 1.000   3rd Qu.: 6.000   3rd Qu.: 35.00  
##  Max.   :59.000   Max.   :23.000   Max.   :73.000   Max.   :165.00  
##        SB               CS               BB               SO        
##  Min.   :  0.00   Min.   : 0.000   Min.   :  0.00   Min.   :  0.00  
##  1st Qu.:  0.00   1st Qu.: 0.000   1st Qu.:  1.00   1st Qu.:  4.00  
##  Median :  0.00   Median : 0.000   Median :  6.00   Median : 19.00  
##  Mean   :  3.12   Mean   : 1.479   Mean   : 16.96   Mean   : 31.96  
##  3rd Qu.:  3.00   3rd Qu.: 2.000   3rd Qu.: 27.00   3rd Qu.: 50.00
```

```
##  Max.   :130.00   Max.   :42.000   Max.   :232.00   Max.   :223.00
##       IBB              HBP              SH               SF
##  Min.   :  0.00   Min.   : 0.000   Min.   : 0.000   Min.   : 0.00
##  1st Qu.:  0.00   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.00
##  Median :  0.00   Median : 0.000   Median : 1.000   Median : 0.00
##  Mean   :  1.48   Mean   : 1.342   Mean   : 1.899   Mean   : 1.43
##  3rd Qu.:  2.00   3rd Qu.: 2.000   3rd Qu.: 3.000   3rd Qu.: 2.00
##  Max.   :120.00   Max.   :50.000   Max.   :40.000   Max.   :18.00
##       GIDP             Bavg
##  Min.   : 0.00   Min.   :0.0000
##  1st Qu.: 0.00   1st Qu.:0.1373
##  Median : 1.00   Median :0.2286
##  Mean   : 3.96   Mean   :0.2024
##  3rd Qu.: 6.00   3rd Qu.:0.2700
##  Max.   :36.00   Max.   :1.0000
```

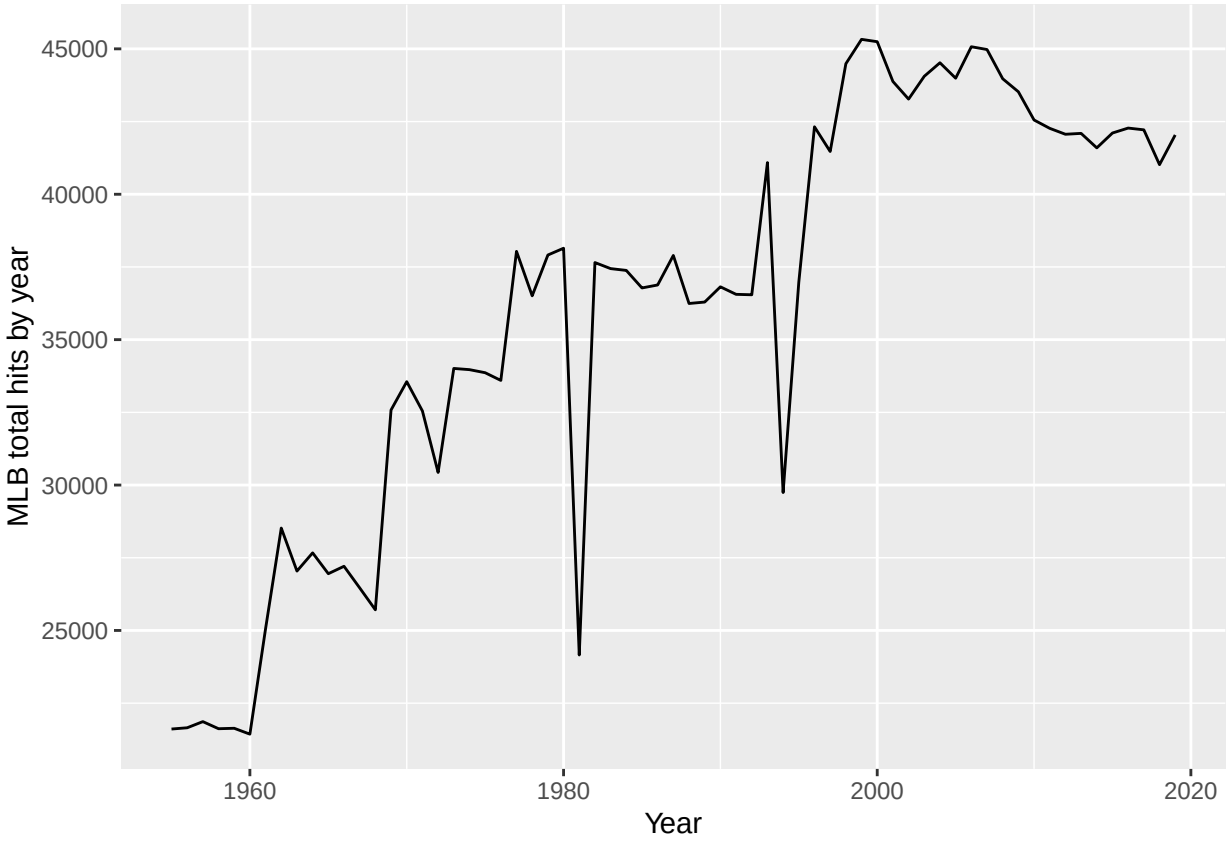
```
boxplot(Batting.num)
```




```
hist(Batting.num$H)
```



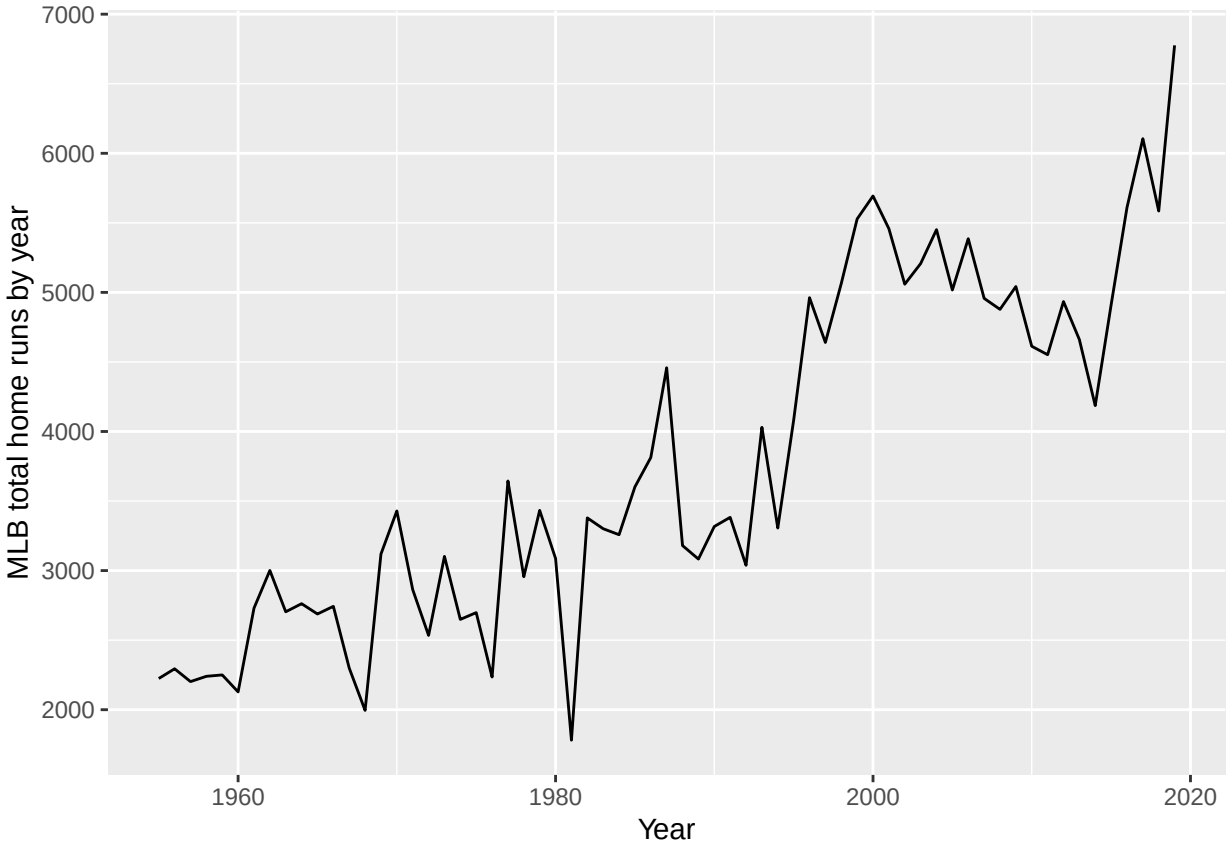

```
Batting.merge %>%
  filter(yearID <= 2019) %>%
  group_by(yearID) %>%
  summarise(H = sum(H)) %>%
  ggplot(aes(x=yearID,y=H)) +
    geom_line() + xlab("Year") + ylab("MLB total hits by year")
```



Looking at the number of hits over the years there does not seem to appear any seasonality really. The biggest things that stand out from the Hits graph is the 1981 year is significantly lower than the rest due to a players' strike that happened halfway through the season and same with the 1994 season which ended the season that year. Another one that stands out is the 2020 season which consisted of a 60 game season due an abbreviated season due to the Covid-19 pandemic.

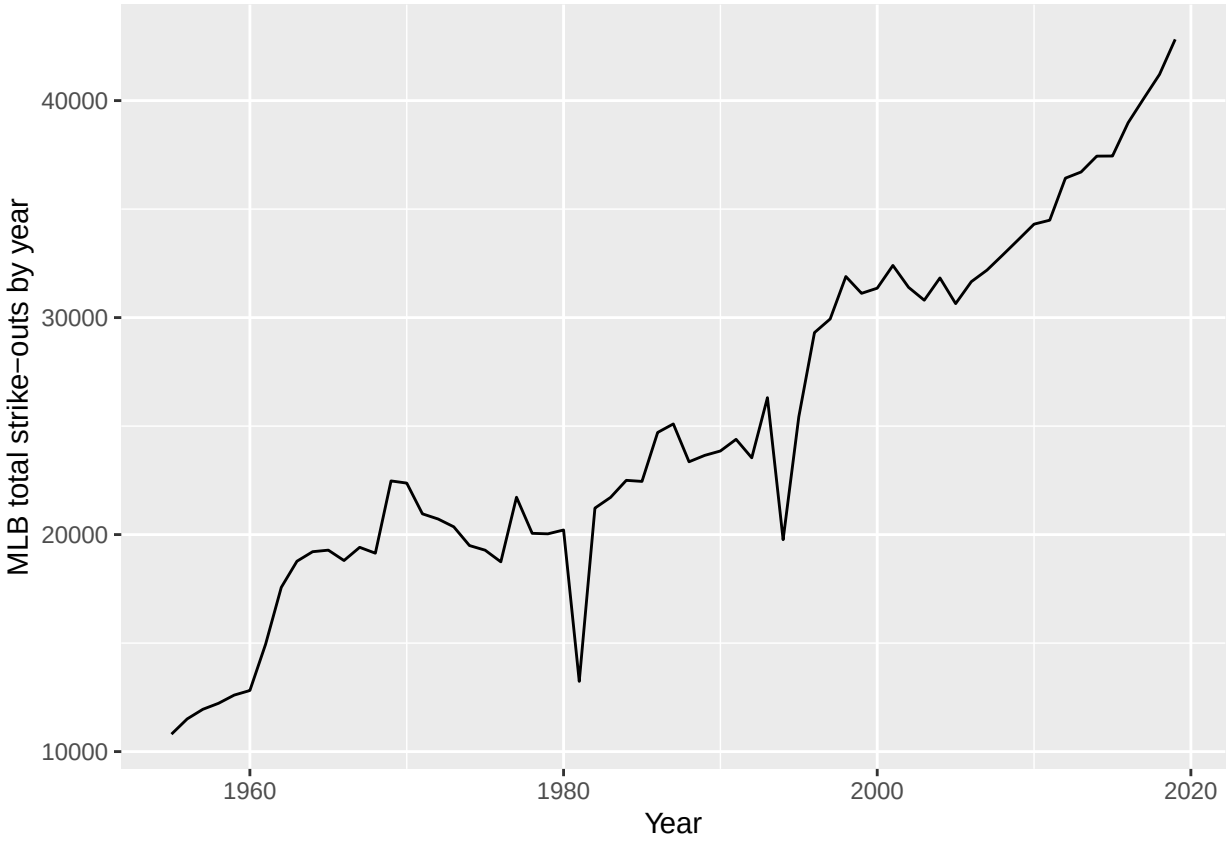
```
#ggplot(data=Batting.merge, aes(x=yearID,y=Bavg)) + geom_line() + xlab("Year") + ylab("Batting Average
```

```
Batting.merge %>%
  filter(yearID <= 2019) %>%
  group_by(yearID) %>%
  summarise(HR = sum(HR)) %>%
  ggplot(aes(x=yearID,y=HR)) +
    geom_line() + xlab("Year") + ylab("MLB total home runs by year")
```



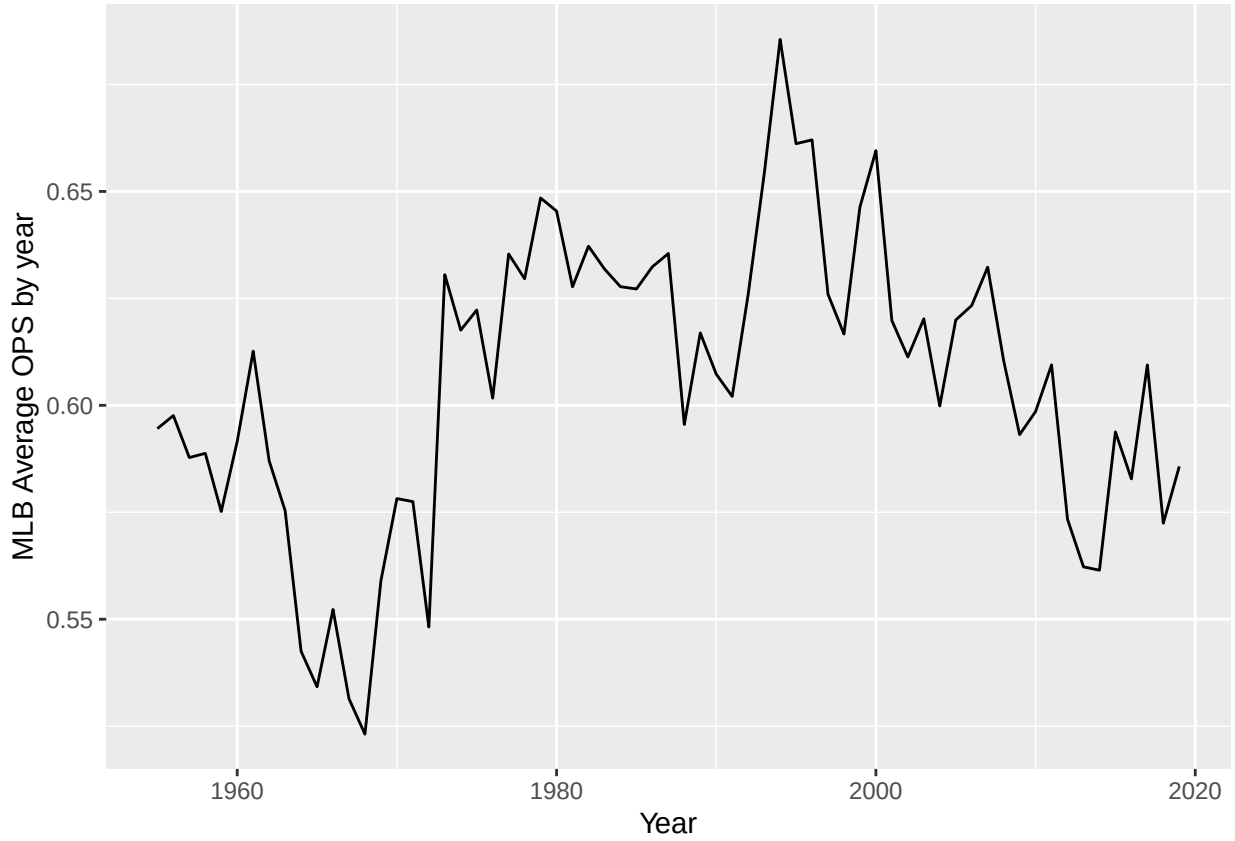
When we look at the HR graph we can't really see a trend or anything, we can tell that the peak of the graph near the late 90s and early 2000s is influenced by the peak of the steroids era in which the league did not implement PED testing until 2003 which we can see by the huge drop in HR for the season that year compared to the previous two.

```
Batting.merge %>%
  filter(yearID <= 2019) %>%
  group_by(yearID) %>%
  summarise(SO = sum(SO)) %>%
  ggplot(aes(x=yearID,y=SO)) +
    geom_line() + xlab("Year") + ylab("MLB total strike-outs by year")
```



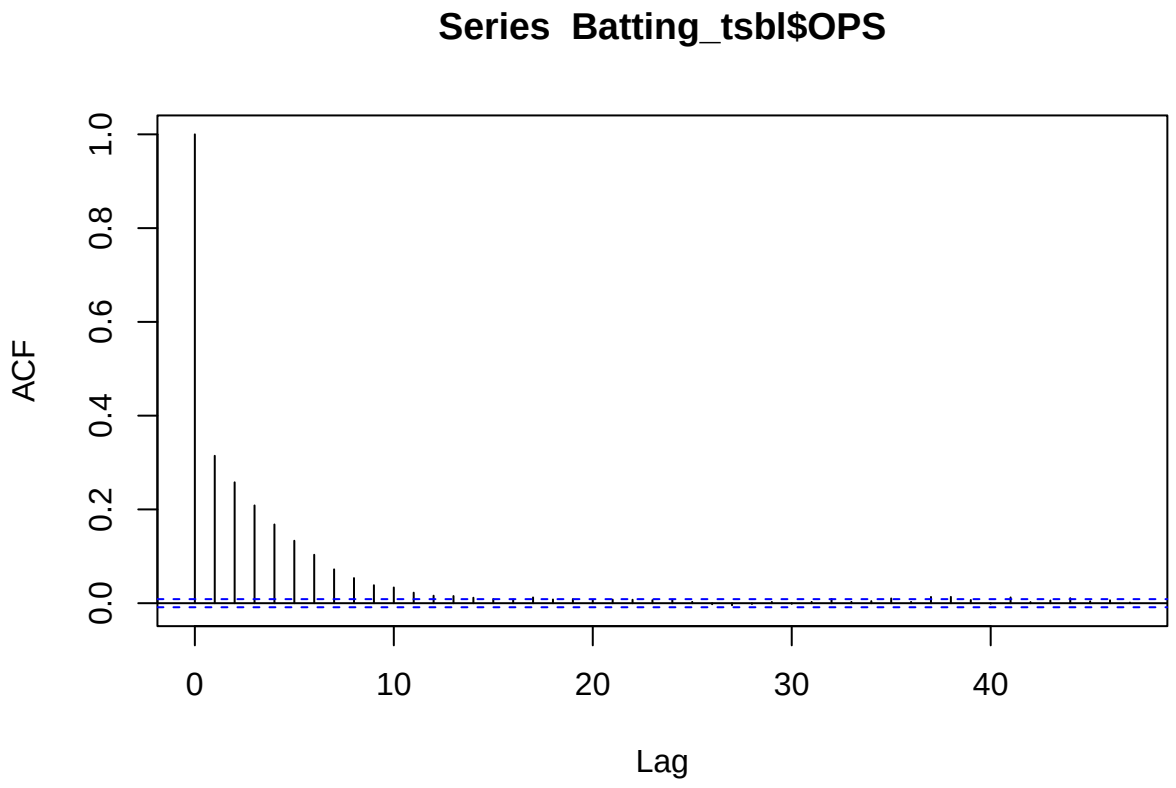
Now when we look at the strike out graph we do see an upward trend in the number of strikeouts throughout the years, we can see the peak coming in in the late 2000s and staying up in the area throughout the 2010s.

```
Batting.merge %>%
  filter(yearID <= 2019) %>%
  group_by(yearID) %>%
  summarise(OPS = mean(OPS)) %>%
  ggplot(aes(x=yearID,y=OPS)) +
    geom_line() + xlab("Year") + ylab("MLB Average OPS by year")
```

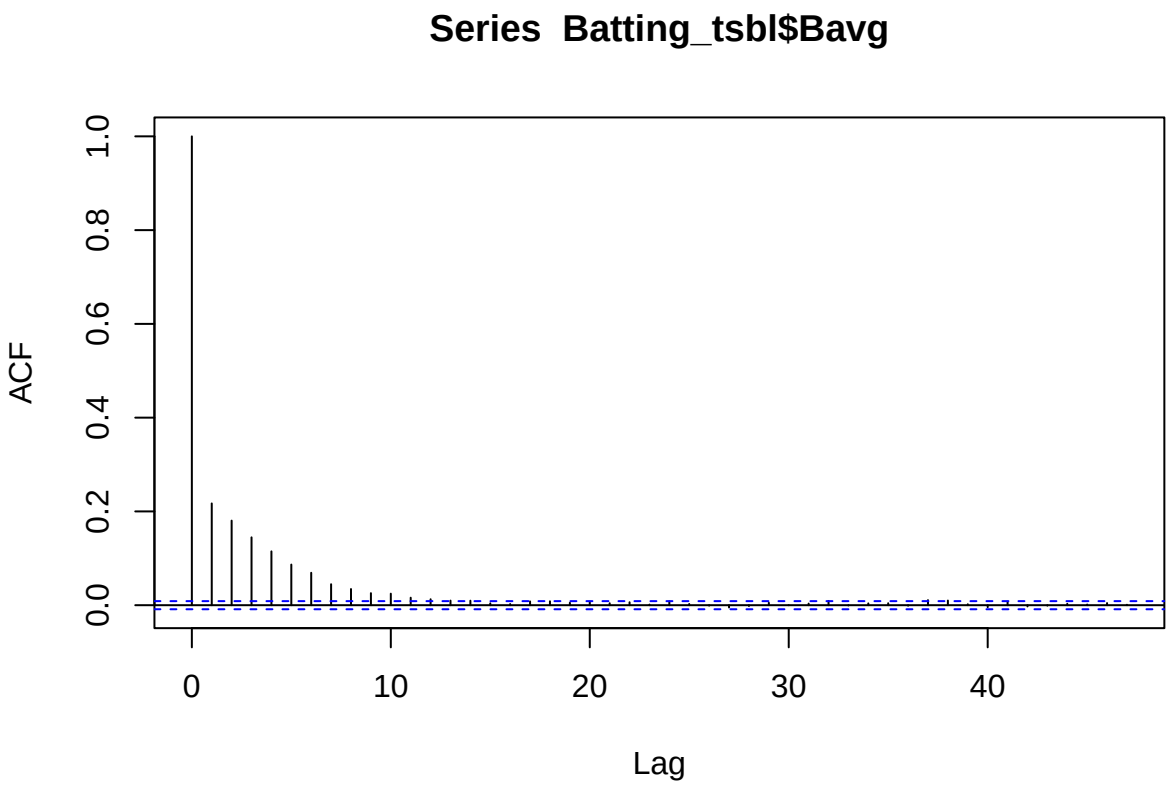


Now when we look at the average OPS graph we don't really see a trend, we can see a peak in the late 90s and early 2000s and a sudden drop off that seem to match the sudden rise of strikeouts in the previous graph.

```
acf(Batting_tsbl$OPS,  type = c("correlation"))
```




```
acf(Batting_tsbl$Bavg, type = c("correlation"))
```



Acf plots show Auto correlation after 12 time steps.