Baseball Homerun and Strikeout Data

*Chris Bailey PhD, CSCS*D, RSCC, University of North Texas*

8/21/2018

After hearing yet another sports reporter comment on the increase in homeruns and strikeouts along with them as if it were a new trend, I decided to take a look at the relationship between homeruns and strikeouts. Is this actually a new trend or relationship?

All data are from bbref.com.

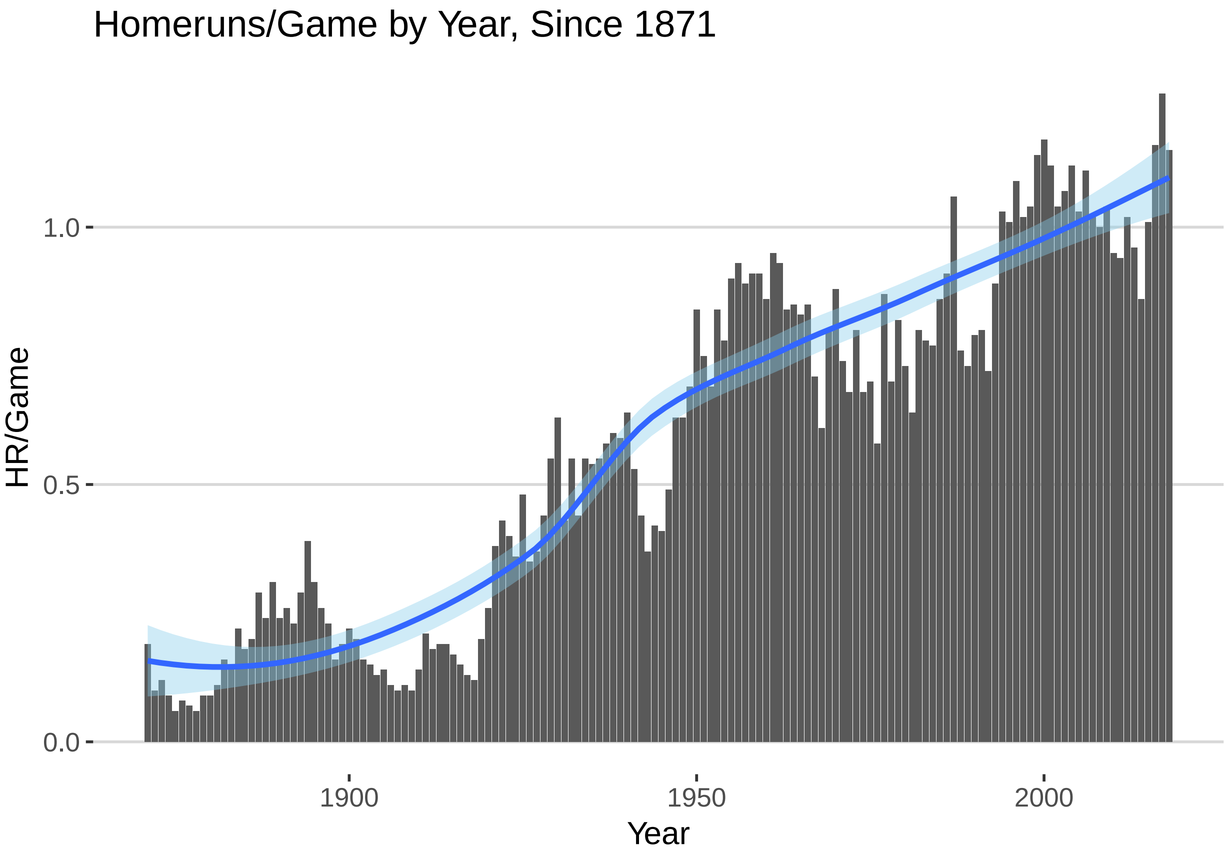
First, I took a look at the historical trends of homerun's and strikeouts. Our data start all the way back in 1871. Due to the difference in number of games being played as well as number of teams across the years, all stats are reported on a per game basis (i.e. HR/g, K/9).

```
library(ggplot2)
library(ggthemes)
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
```

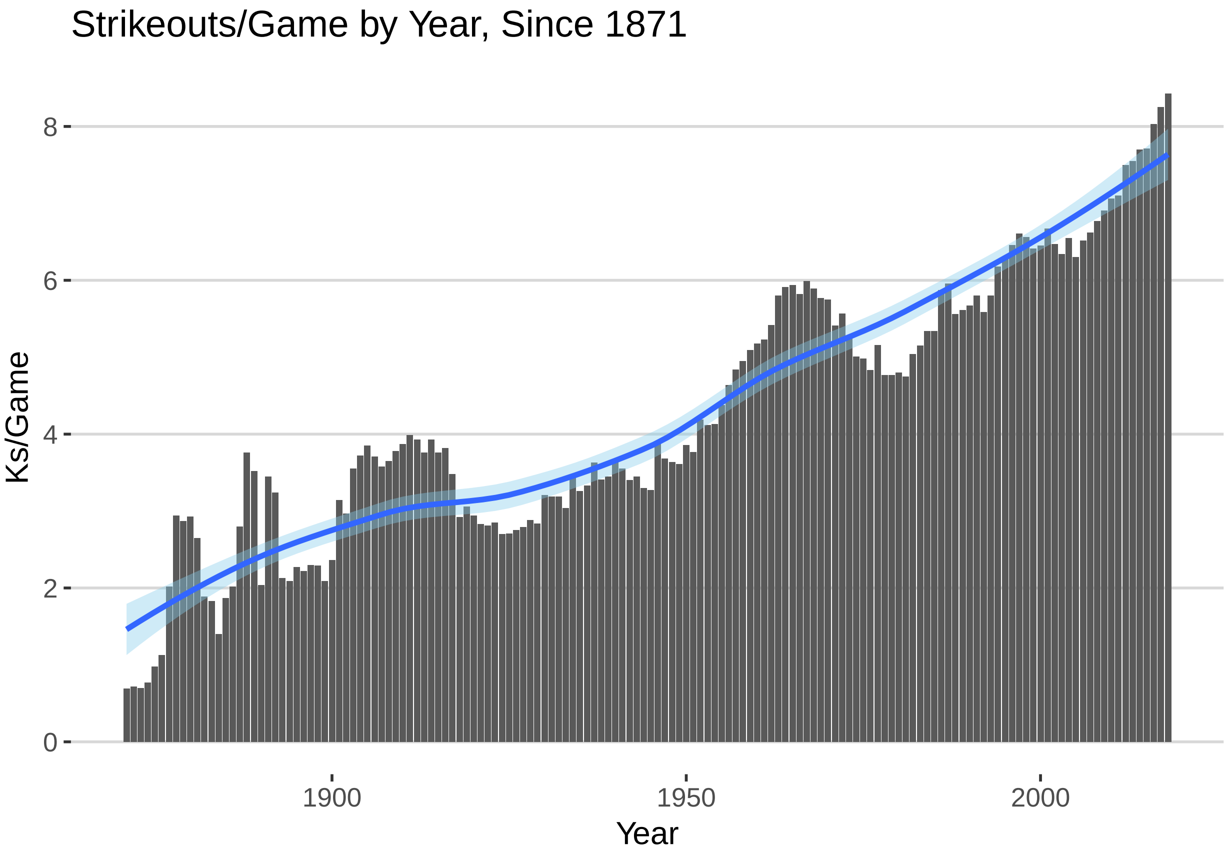
```
setwd("/Users/cbailey/Google Drive/GitHub/BaseballData")
data <- read.csv("/Users/cbailey/Google Drive/GitHub/BaseballData/Data.csv")
data$X1b<-(data$H-(data$X2B+data$X3B+data$HR))
df<-data.frame(Year=(data$Year),
               HR=(data$HR),
               Ks=(data$SO))

HR<-ggplot(df, aes(x=Year, y=HR)) +
  geom_bar(stat="identity")+
  geom_smooth(fill="skyblue")+ ylab("HR/Game")+
  theme_hc()+ggtitle("Homeruns/Game by Year, Since 1871")

HR
```

```
## `geom_smooth()` using method = 'loess'
```



```
Ks<-ggplot(df, aes(x=Year, y=Ks)) +
  geom_bar(stat="identity")+
  geom_smooth(fill="skyblue")+ ylab("Ks/Game")+
  theme_hc()+ggtitle("Strikeouts/Game by Year, Since 1871")

Ks
```

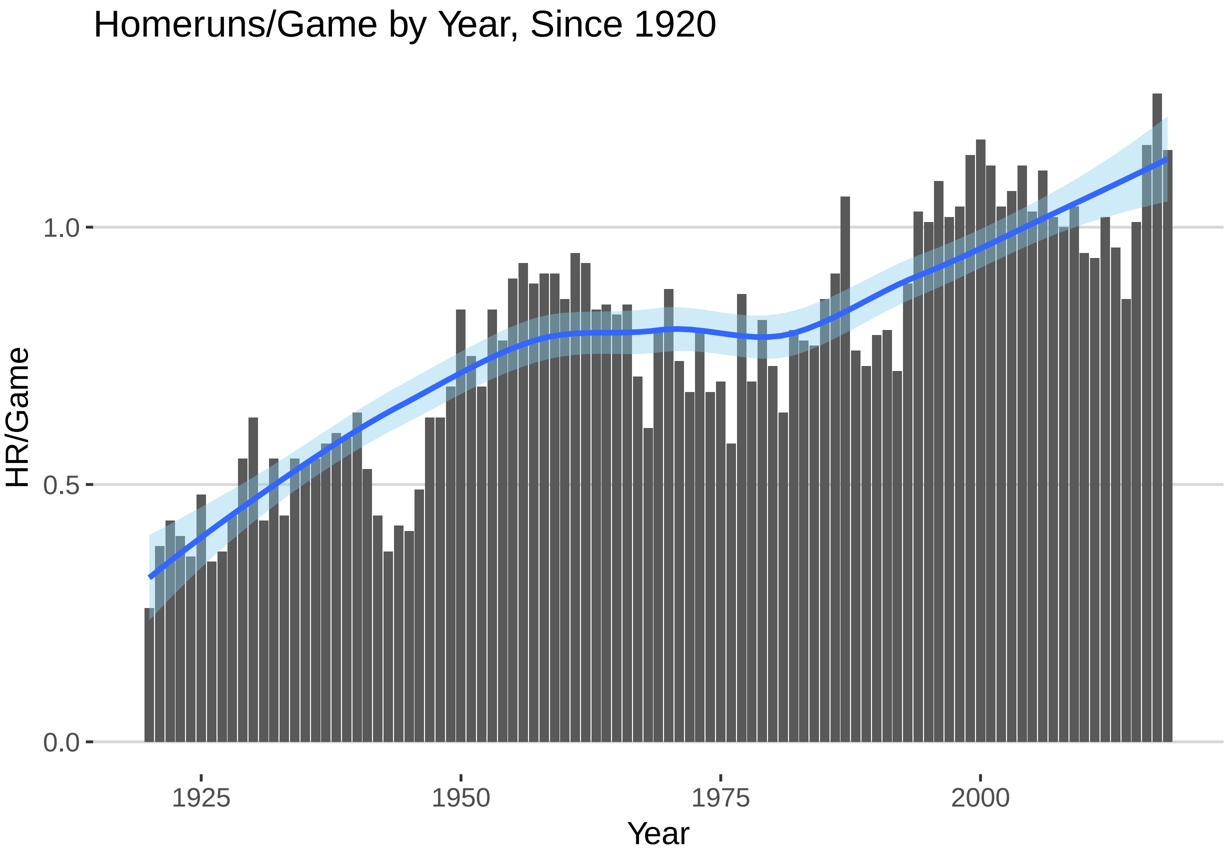
```
## `geom_smooth()` using method = 'loess'
```



That’s interesting, but it also includes the “deadball era.” Restricting the dataset to seasons after 1919, when the “live ball era” began, makes more sense.

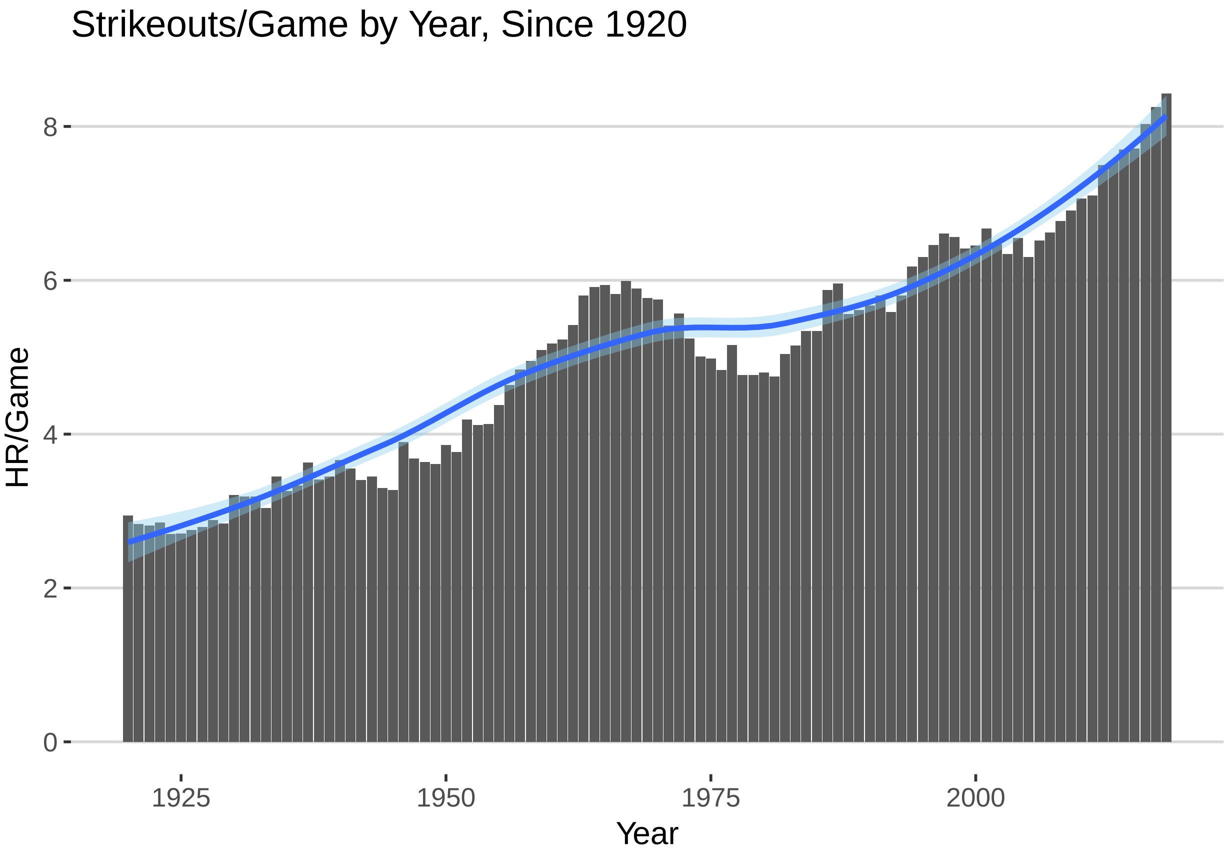
```
LBE<-dplyr::filter(df,Year>1919)


HR<-ggplot(LBE, aes(x=Year, y=HR)) +
  geom_bar(stat="identity")+
  geom_smooth(fill="skyblue")+ ylab("HR/Game")+
  theme_hc()+ggtitle("Homeruns/Game by Year, Since 1920")

HR
```

```
## `geom_smooth()` using method = 'loess'
```



```
Ks<-ggplot(LBE, aes(x=Year, y=Ks)) +
  geom_bar(stat="identity")+
  geom_smooth(fill="skyblue")+ ylab("HR/Game")+
  theme_hc()+ggtitle("Strikeouts/Game by Year, Since 1920")

Ks
```

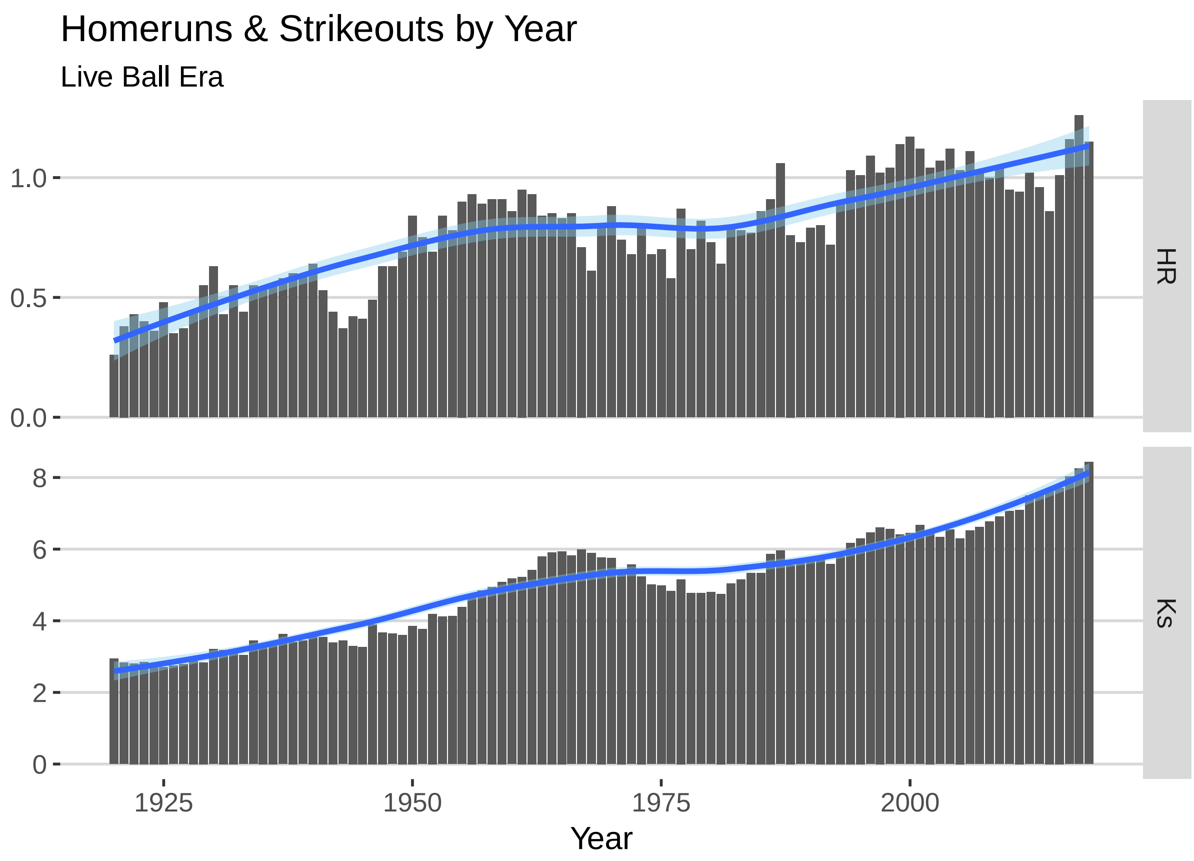
```
## `geom_smooth()` using method = 'loess'
```



Looking at the curves of the data, they appear very similar. It is more interesting to see them on the same plot. I am transformed the data from the wide into the long format with "Year" as an ID variable. Then I faceted them so they can be seen together. I set the y axis to free to further illustrate the similarity.

```
library(reshape2)
HRSOx<-melt(LBE, id.vars=c("Year"))

HRxSO<-ggplot(HRSOx, aes(x=Year,y=HRSOx$value))+
  geom_bar(stat="identity")+geom_smooth(fill="skyblue")+facet_grid(variable~., scales="free_y")+
  theme_hc()+labs(title="Homeruns & Strikeouts by Year", subtitle="Live Ball Era")+ylab("")

HRxSO
```

```
## `geom_smooth()` using method = 'loess'
```

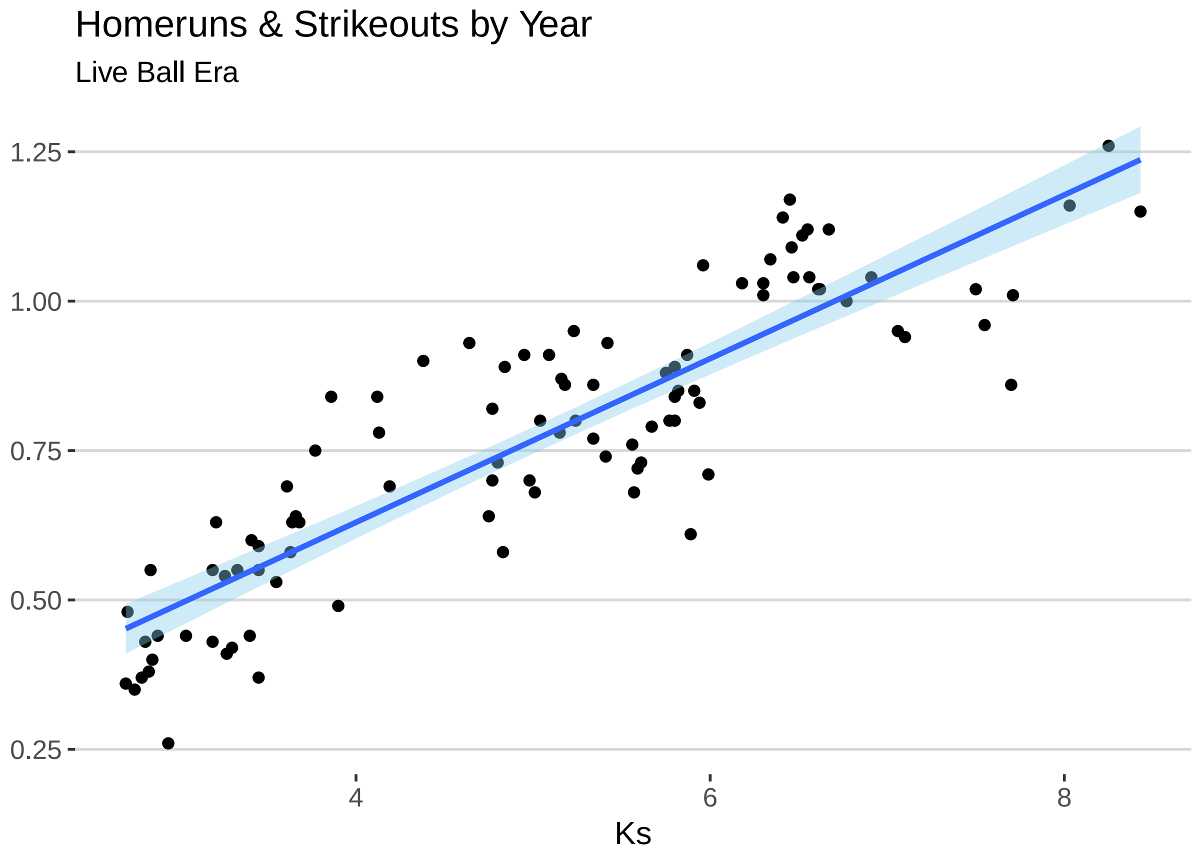


So it appears that the recent relationship between homeruns and strikeouts talked about in so much of the media, is not recent. In fact, the relationship is nearly perfect (r=0.879). It has been that way for some time. Our magnitudes are just larger now than they have been in the past. A correlation table and a plot are below for reference.

```
cor(LBE)
```

```
##           Year        HR        Ks
## Year 1.0000000 0.8539941 0.9527332
## HR   0.8539941 1.0000000 0.8785184
## Ks   0.9527332 0.8785184 1.0000000
```

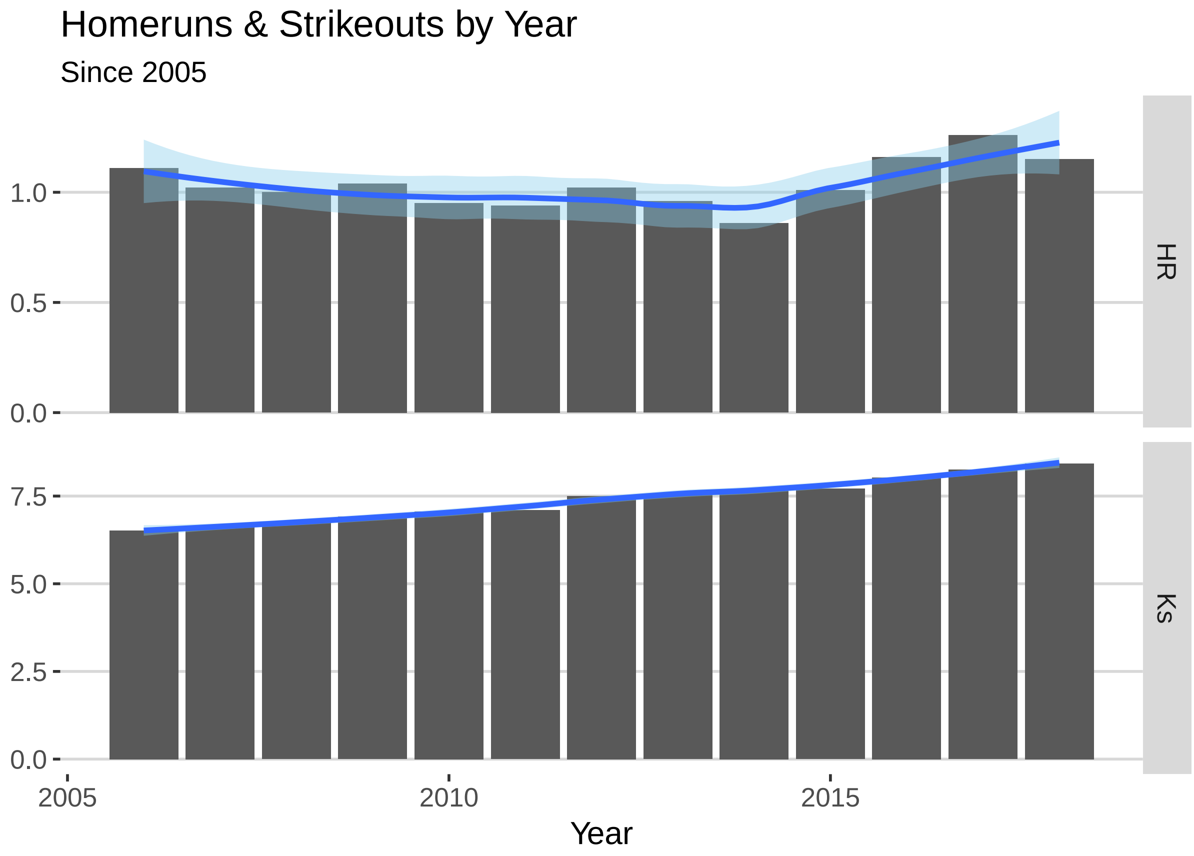
```
HRxSOscatterplot<-ggplot(LBE, aes(Ks,HR))+geom_point()+geom_smooth(method=lm,fill="skyblue")+
  theme_hc()+labs(title="Homeruns & Strikeouts by Year", subtitle="Live Ball Era")+ylab("")
HRxSOscatterplot
```



The trend is that both are up right now, but this trend for increased K/9 has been on the rise for a long time (pretty much the entire live ball era). HR/g has seen more fluctuation over the years, but the smoothed trend has also increased pretty steadily over time. If we look more recently, we can see that HR/g have increased at a different rate than K/9.

```
HR2k<-dplyr::filter(LBE,Year>2005)
HRSO2k<-melt(HR2k, id.vars=c("Year"))

HRxSO2k<-ggplot(HRSO2k, aes(x=Year,y=HRSO2k$value))+
  geom_bar(stat="identity")+geom_smooth(fill="skyblue")+facet_grid(variable~., scales="free_y")+
  theme_hc()+labs(title="Homeruns & Strikeouts by Year", subtitle="Since 2005")+ylab("")

HRxSO2k
```

```
## `geom_smooth()` using method = 'loess'
```



Fairly recently, homeruns per game were actually decreasing until 2015. So, if strikeouts have increased at a fairly steady rate, but the homeruns per game have only very recently begun to increase again, it seems like people are actaully mad at the longball.

There is the concept that both of these increase the length of games. It will be interesting to see if rule changes are on the way to speed up the game, since rules usually favor someone (pitchers or hitters, but not both). Maybe there will be some separation in these variables coming along with speed-up rules.