Deriving Insight From Lahman's Baseball Database

Michael Calabro

7/31/2020

Welcome to my journey_to_insight.Rmd file! In this markdown, I will be going step by step through my process of using the Lahman Baseball Database to seek answers to my question regarding Strikeouts in Major League Baseball.

What is the Lahman Baseball Database? And what is your question about MLB strikeouts?

The Lahman Baseball Database contains MLB statistics and data from 1871-2019…it contains nearly 250 years worth of MLB data to explore! Documentation for the database and all of its tables and columns can be found using this link.
I downloaded the SQLite database into my RStudio Project, and can access all of its information with the following commands.

```
# RSQLite library allows me to use SQLite
# DBI library allows me to connect to the database
# tidyverse contains many libraries, most notably ggplot2 for visualizations
# kableExtra allows me to make my tables more presentable
library(RSQLite)
library(DBI)
library(tidyverse)
library(kableExtra)

# Now I need to establish my connection to the Lahman database
con <- dbConnect(SQLite(),
                 dbname = "lahmans_baseball_db.sqlite")
```

This "con" keyword now allows me to connect to the database. To view the list of tables in the database, I simply run the code below.

```
dbListTables(con)
```

```
##  [1] "allstarfull"        "appearances"         "awardsmanagers"
##  [4] "awardsplayers"      "awardssharemanagers" "awardsshareplayers"
##  [7] "batting"            "battingpost"         "collegeplaying"
## [10] "divisions"          "fielding"            "fieldingof"
## [13] "fieldingofsplit"    "fieldingpost"        "halloffame"
## [16] "homegames"          "leagues"             "managers"
## [19] "managershalf"       "parks"               "people"
## [22] "pitching"           "pitchingpost"        "salaries"
## [25] "schools"            "seriespost"          "teams"
## [28] "teamsfranchises"    "teamshalf"
```

To view data from a specific table, I use the dbGetQuery function. The “batting” table will be key to my analysis, let’s check it out.

```r
# First I write out the query and assign it to "query"
query <-  "
          SELECT *
          FROM batting
          LIMIT 5
          "

# Then I send it through the DBGetQuery function,
# along with the "con" connection, to run the query
dbGetQuery(con, query)
```

```
##   ID  playerID yearID stint teamID team_ID lgID  G G_batting  AB  R  H 2B 3B HR
## 1  1 abercda01   1871     1    TRO       8   NA  1        NA   4  0  0  0  0  0
## 2  2  addybo01   1871     1    RC1       7   NA 25        NA 118 30 32  6  0  0
## 3  3 allisar01   1871     1    CL1       3   NA 29        NA 137 28 40  4  5  0
## 4  4 allisdo01   1871     1    WS3       9   NA 27        NA 133 28 44 10  2  2
## 5  5 ansonca01   1871     1    RC1       7   NA 25        NA 120 29 39 11  3  0
##   RBI SB CS BB SO IBB HBP SH SF GIDP
## 1   0  0  0  0  0  NA  NA NA NA    0
## 2  13  8  1  4  0  NA  NA NA NA    0
## 3  19  3  1  2  5  NA  NA NA NA    1
## 4  27  1  1  0  2  NA  NA NA NA    0
## 5  16  6  2  2  1  NA  NA NA NA    0
```

```r
# And for tables that aren't so wide,
# the kable funtion with kable_styling makes the table very presentable
dbGetQuery(con, query) %>%
  kable() %>%
  kable_styling(full_width = FALSE, bootstrap_options = "bordered")
```

ID	playerID	yearID	stint	teamID	team_ID	lgID	G	G_batting	AB	R	H	2B	3B	HR	RBI
1	abercda01	1871	1	TRO	8	NA	1	NA	4	0	0	0	0	0	0
2	addybo01	1871	1	RC1	7	NA	25	NA	118	30	32	6	0	0	13
3	allisar01	1871	1	CL1	3	NA	29	NA	137	28	40	4	5	0	19
4	allisdo01	1871	1	WS3	9	NA	27	NA	133	28	44	10	2	2	27
5	ansonca01	1871	1	RC1	7	NA	25	NA	120	29	39	11	3	0	16

```r
# I like the kable and think I'll use it a lot,
# so I am going to make a function kable_query for efficiency
kable_query <- function(con, query){
  return(
    dbGetQuery(con, query) %>%
      kable() %>%
      kable_styling(full_width = FALSE, bootstrap_options = "bordered")
  )
}
```

As we can see, every row in the batter table consists of a player, a year, and all of the player’s batting statistics in that year.

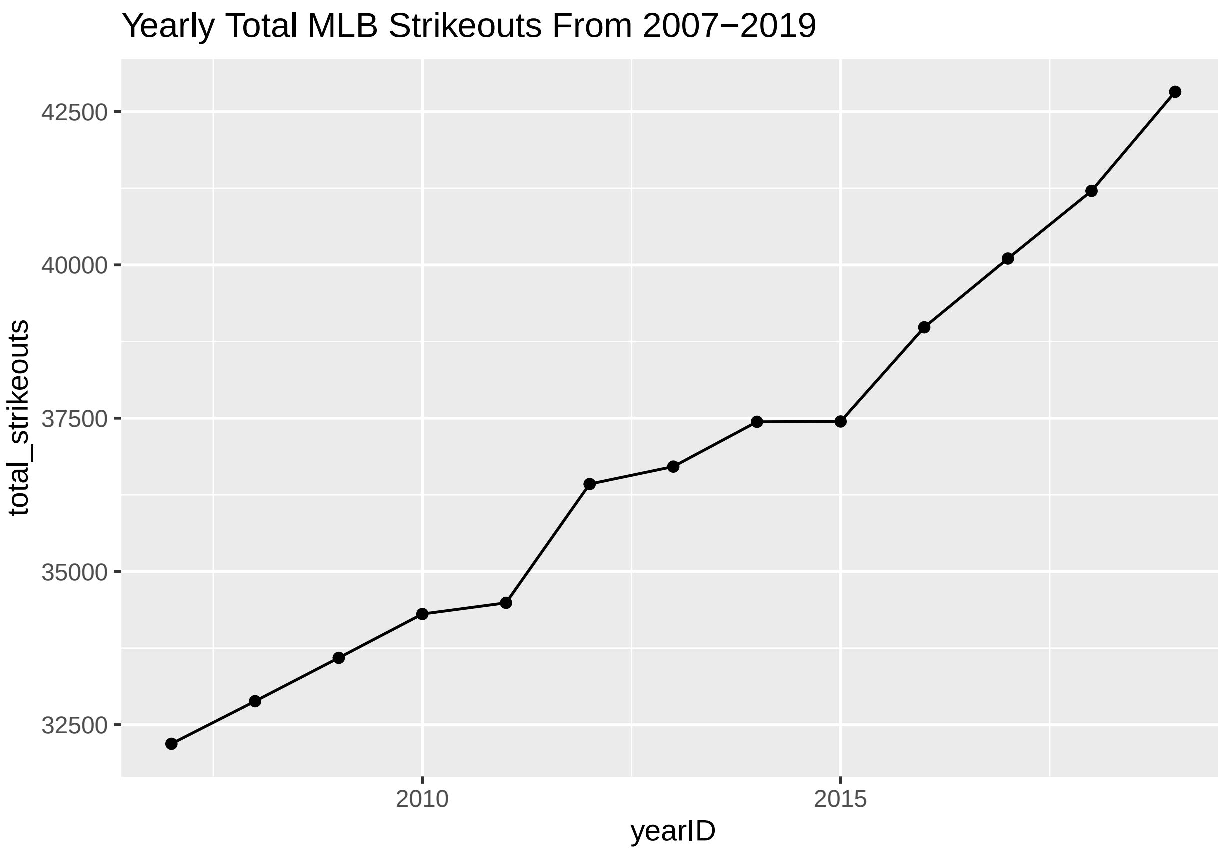
So What is your question about Strikeouts in the MLB?

A few years back (in 2018 I believe) I read an article which mentioned the fact that MLB players are striking out more and more every year, always breaking the single season record for total strikeouts. A more updated article, linked here, notes that this increase in strikeouts has been happening every year since 2008! Let's see if our Lahman data shows the same trend.

```
query <-  "
          SELECT
            yearID,
            SUM(SO) AS total_strikeouts
          FROM batting
          WHERE yearID > 2006
          GROUP BY yearID
          "

kable_query(con, query)
```

yearID	total_strikeouts
2007	32189
2008	32884
2009	33591
2010	34306
2011	34488
2012	36426
2013	36710
2014	37441
2015	37446
2016	38982
2017	40104
2018	41207
2019	42823

```r
# Now I can use ggplot, with geom_point to make a scatterplot,
# geom_line to make a line graph, or both!
#
# Also, using %>% allows me to "pipe" my data into certain functions, like ggplot
dbGetQuery(con, query) %>%
  ggplot(aes(x = yearID, y = total_strikeouts)) +
  geom_point() +
  geom_line() +
  ggtitle("Yearly Total MLB Strikeouts From 2007-2019")
```



So what is causing this increase in strikeouts every year?? That is what I wish to investigate throughout the rest of this document/project.

HYPOTHOSES

- Perhaps there are just more at bats every year, while strikeouts per at bat is remaining constant
- Perhaps hitters are willing to strike out more often, in exchange for an increase in another statistic
- Perhaps pitching skill is improving faster than batting skill in the MLB

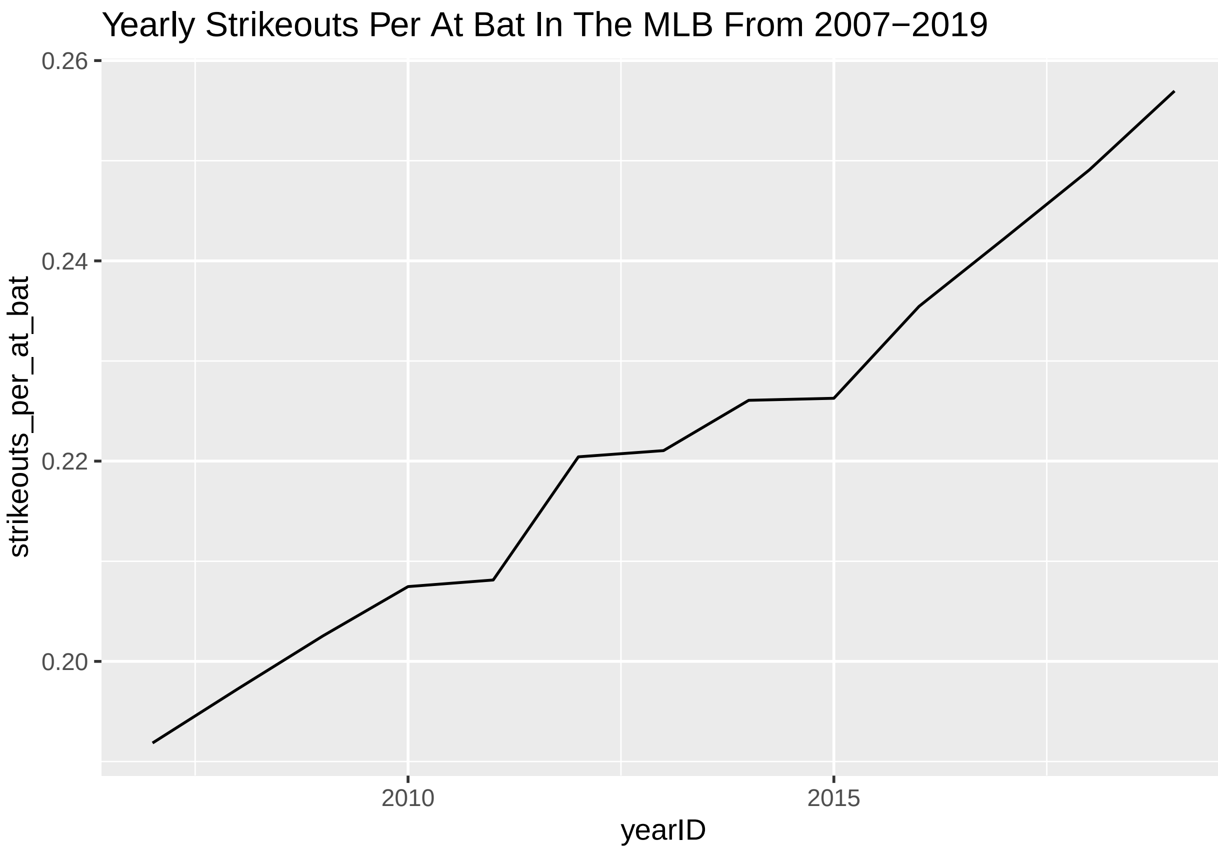
Strikeouts Per At Bat

This hypothesis is pretty easy to check. We just need to divide total strikeouts by total at bats.

```
query <-    "
            SELECT
              yearID,
              SUM(AB) AS total_at_bats,
              SUM(SO) AS total_strikeouts,
              CAST(SUM(SO) AS FLOAT) / SUM(AB) AS strikeouts_per_at_bat
            FROM batting
            WHERE yearID > 2006
            GROUP BY yearID
                  "

kable_query(con, query)
```

yearID	total_at_bats	total_strikeouts	strikeouts_per_at_bat
2007	167783	32189	0.1918490
2008	166714	32884	0.1972480
2009	165849	33591	0.2025397
2010	165353	34306	0.2074713
2011	165705	34488	0.2081289
2012	165251	36426	0.2204283
2013	166070	36710	0.2210514
2014	165614	37441	0.2260739
2015	165488	37446	0.2262762
2016	165561	38982	0.2354540
2017	165567	40104	0.2422222
2018	165432	41207	0.2490872
2019	166651	42823	0.2569622

```
dbGetQuery(con, query) %>%
  ggplot(aes(x = yearID, y = strikeouts_per_at_bat)) +
  geom_line() +
  ggtitle("Yearly Strikeouts Per At Bat In The MLB From 2007-2019")
```



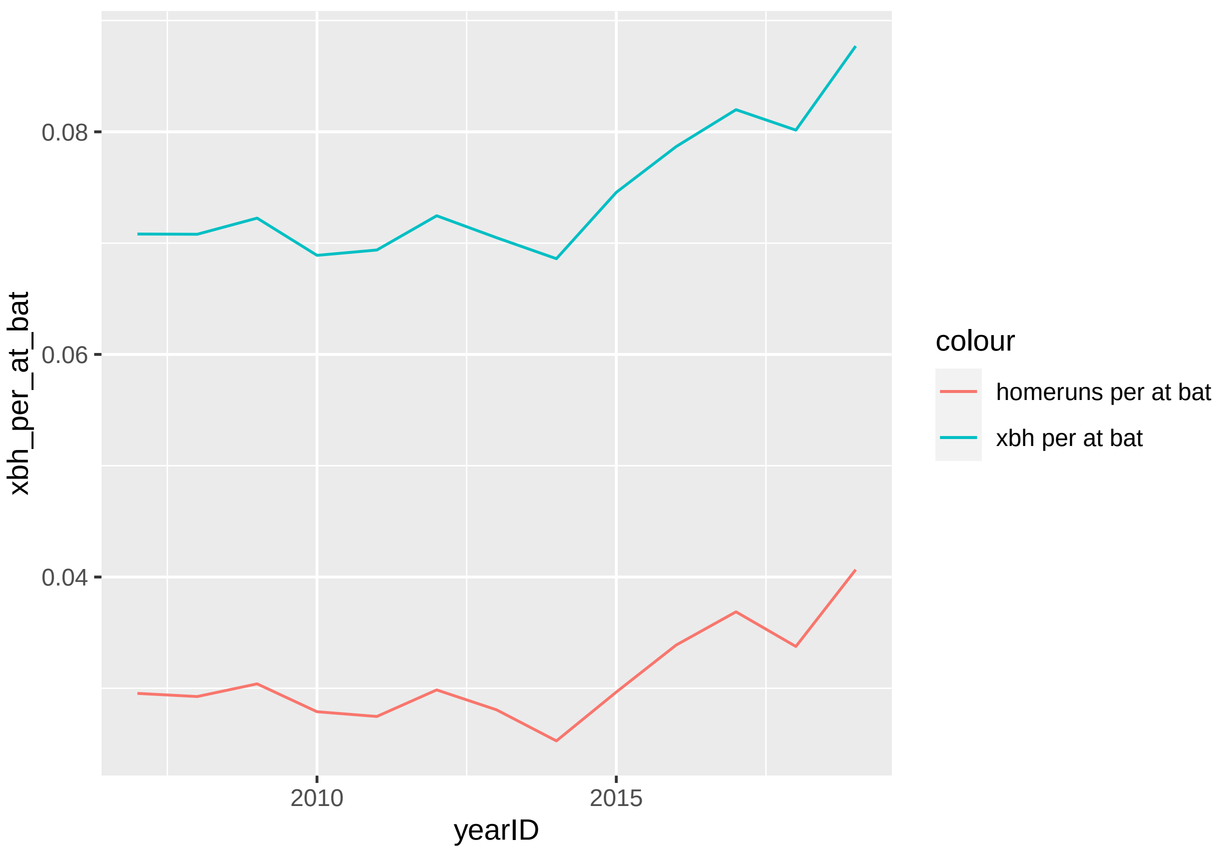
Now the obvious question is: Have any other hitting stats been rising or falling in lockstep with the strikeouts stat?

```
query <-    "
            SELECT
              yearID,
              CAST(SUM(SO) AS FLOAT) / SUM(AB) AS strikeouts_per_at_bat,
              CAST(SUM(H) AS FLOAT) / SUM(AB) AS hits_per_at_bat,
              CAST(SUM(HR) AS FLOAT) / SUM(AB) AS homeruns_per_at_bat,
              CAST(SUM('2B') AS FLOAT) / SUM(AB) AS doubles_per_at_bat,
              CAST(SUM('3B') AS FLOAT) / SUM(AB) AS triples_per_at_bat,
              CAST(SUM('3B') + SUM('2B') + SUM(HR) AS FLOAT) / SUM(AB) AS xbh_per_at_bat,
              CAST(SUM(BB) AS FLOAT) / SUM(AB) AS walks_per_at_bat
            FROM batting
            WHERE yearID > 2006
            GROUP BY yearID
            "

kable_query(con, query)
```

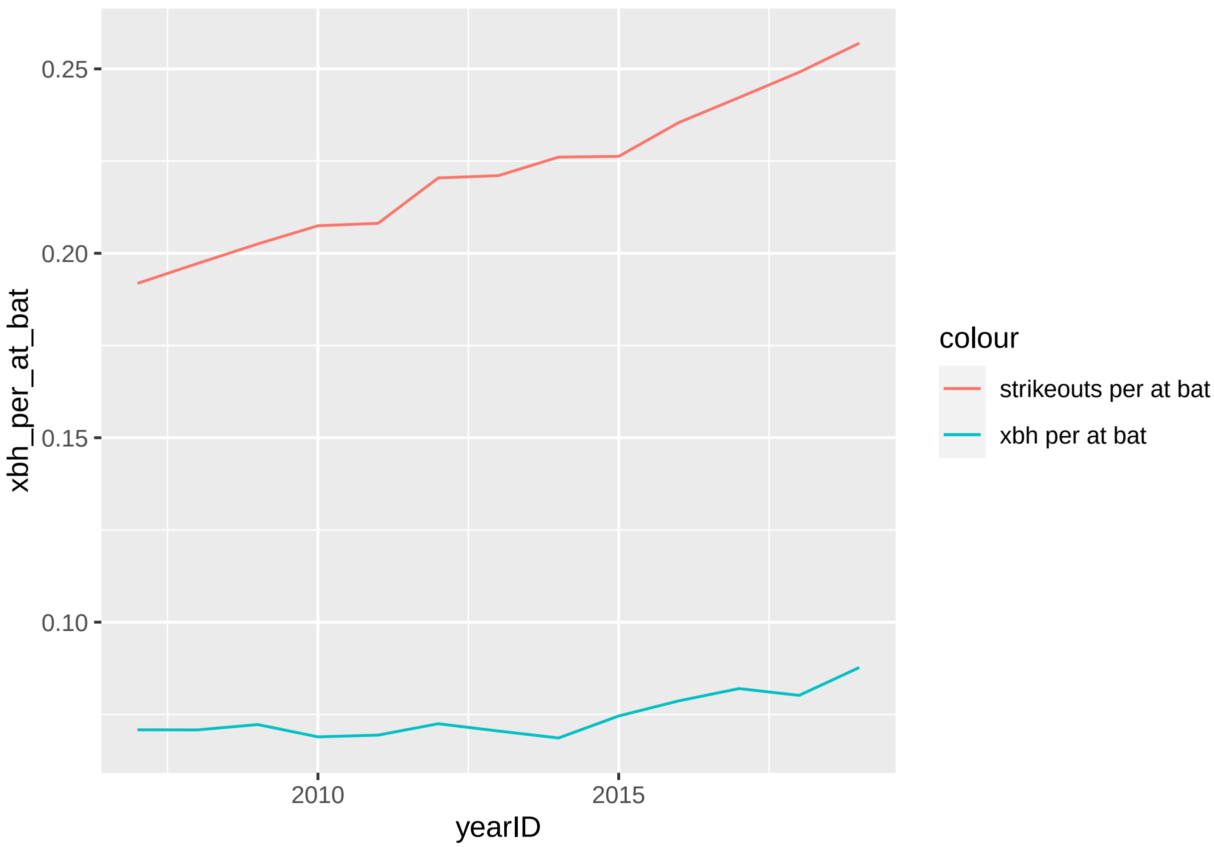
yearID	strikeouts_per_at_bat	hits_per_at_bat	homeruns_per_at_bat	doubles_per_at_bat	triples_per_at_
2007	0.1918490	0.2680665	0.0295441	0.0165094	0.0247
2008	0.1972480	0.2637571	0.0292597	0.0166153	0.0249
2009	0.2025397	0.2624315	0.0304011	0.0167381	0.0251
2010	0.2074713	0.2573525	0.0278979	0.0164013	0.0246
2011	0.2081289	0.2550738	0.0274705	0.0167647	0.0251
2012	0.2204283	0.2545401	0.0298576	0.0170407	0.0255
2013	0.2210514	0.2534654	0.0280665	0.0169687	0.0254
2014	0.2260739	0.2511563	0.0252756	0.0173295	0.0259
2015	0.2262762	0.2544354	0.0296638	0.0179590	0.0269
2016	0.2354540	0.2553500	0.0338848	0.0179148	0.0268
2017	0.2422222	0.2549723	0.0368733	0.0180471	0.0270
2018	0.2490872	0.2479448	0.0337601	0.0185575	0.0278
2019	0.2569622	0.2522577	0.0406598	0.0188178	0.0282

Homeruns and extra base hits (xbh) are also much higher in 2019 than they were in 2007, but the jump was not as consistent as the gradual rise of strikeouts.

```r
dbGetQuery(con, query) %>%
  ggplot(aes(x = yearID)) +
  geom_line(aes(y = xbh_per_at_bat, color = "xbh per at bat")) +
  geom_line(aes(y = homeruns_per_at_bat, color = "homeruns per at bat"))
```



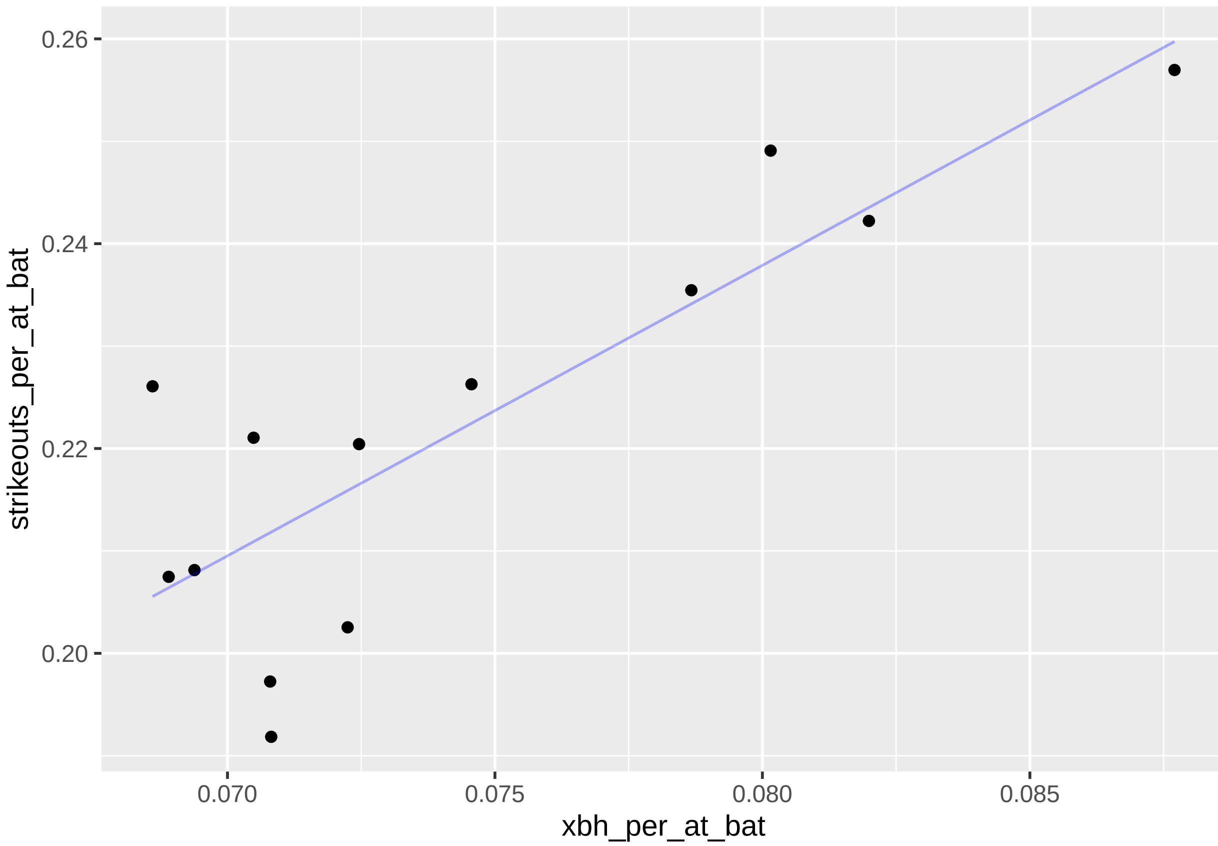

```r
dbGetQuery(con, query) %>%
  ggplot(aes(x = yearID)) +
  geom_line(aes(y = xbh_per_at_bat, color = "xbh per at bat")) +
  geom_line(aes(y = strikeouts_per_at_bat, color = "strikeouts per at bat"))
```



They don't appear to be too correlated. Let's check it out with a scatterplot.

```
dbGetQuery(con, query) %>%
  ggplot(aes(x = xbh_per_at_bat, y = strikeouts_per_at_bat)) +
  geom_point() +
  # Stat_smooth, with geom=line and method=lm,
  # adds a regression line of best fit to the graph
  stat_smooth(geom='line', method = "lm", alpha=0.3, se=FALSE, color = "blue")
```

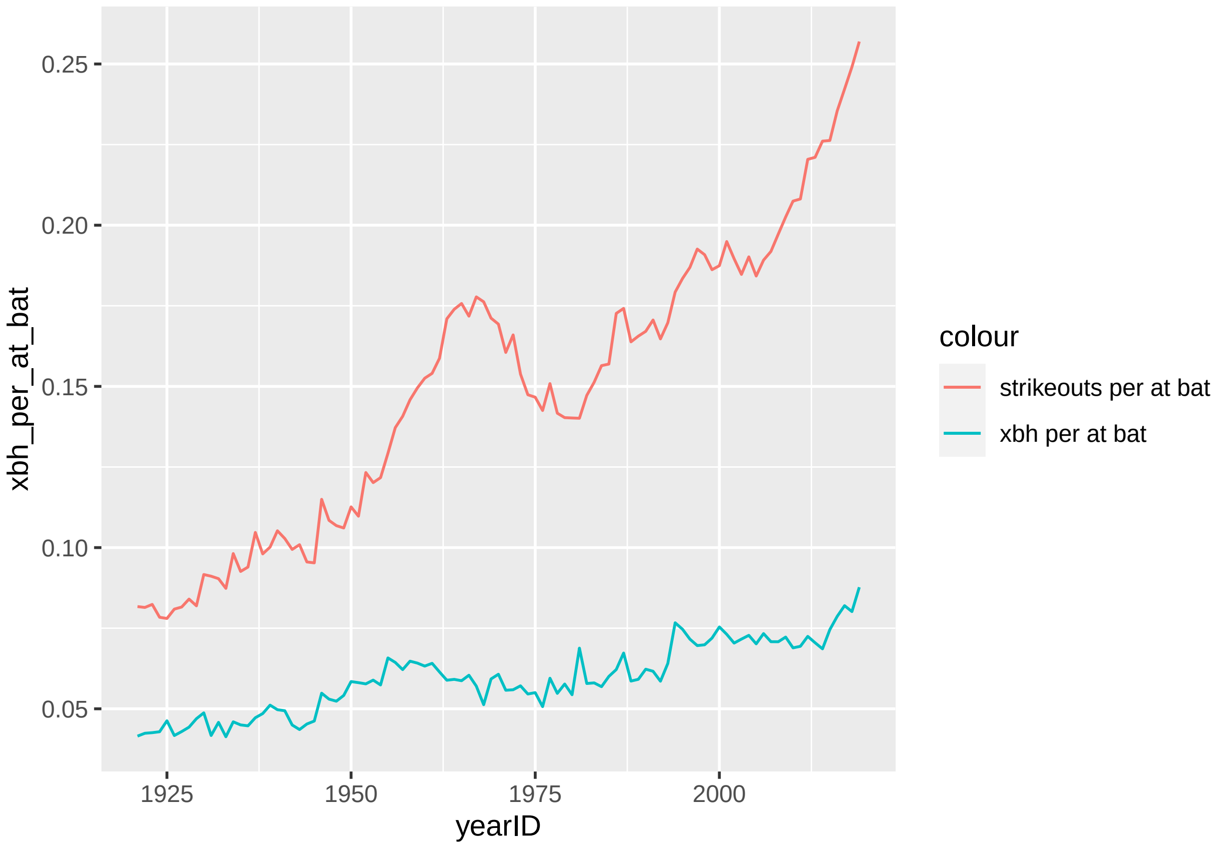
```
## `geom_smooth()` using formula 'y ~ x'
```



There appears to be some correlation between strikeouts and extra base hits, but it is not a very strong one, and there is some heteroskedasticity.

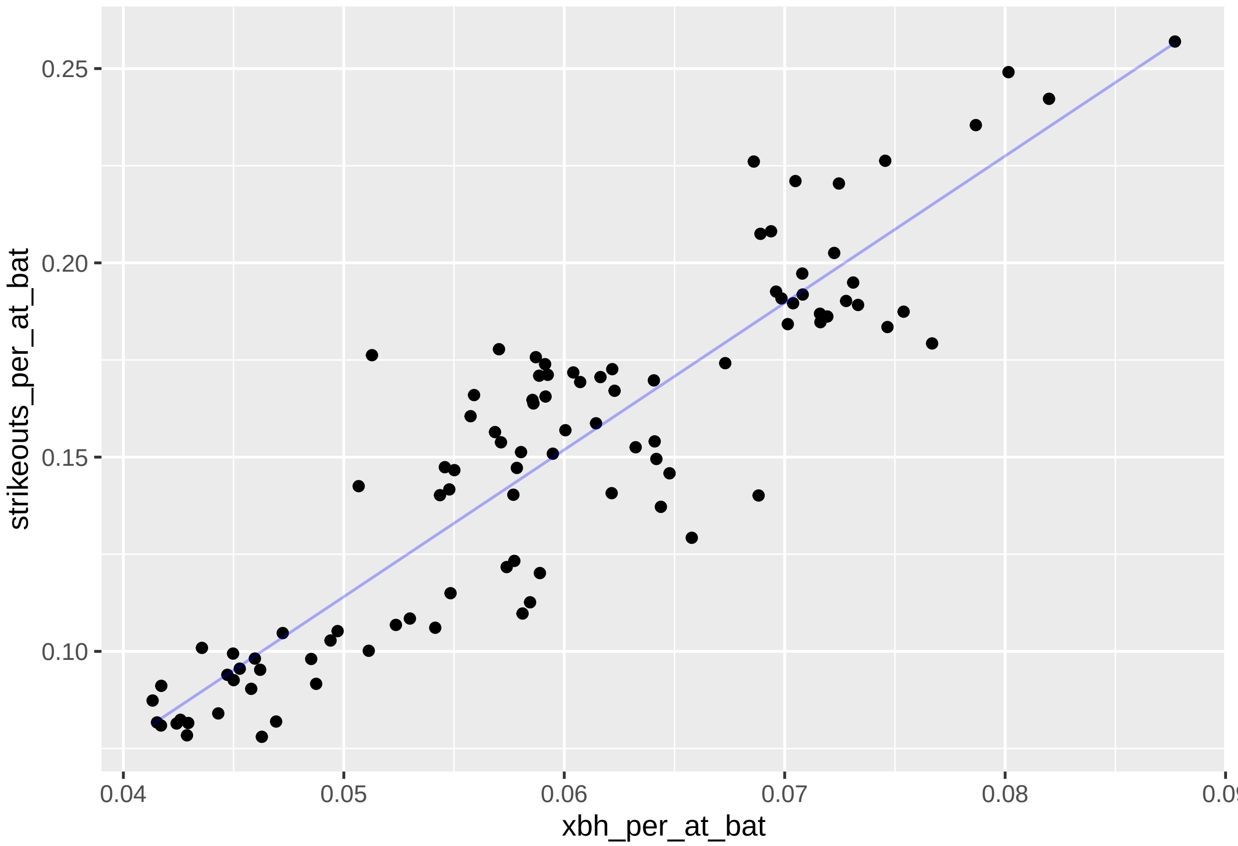
What if we go further back in time?

```
query <-    "
            SELECT
              yearID,
              CAST(SUM(SO) AS FLOAT) / SUM(AB) AS strikeouts_per_at_bat,
              CAST(SUM(H) AS FLOAT) / SUM(AB) AS hits_per_at_bat,
              CAST(SUM(HR) AS FLOAT) / SUM(AB) AS homeruns_per_at_bat,
              CAST(SUM('2B') AS FLOAT) / SUM(AB) AS doubles_per_at_bat,
              CAST(SUM('3B') AS FLOAT) / SUM(AB) AS triples_per_at_bat,
              CAST(SUM('3B') + SUM('2B') + SUM(HR) AS FLOAT) / SUM(AB) AS xbh_per_at_bat,
              CAST(SUM(BB) AS FLOAT) / SUM(AB) AS walks_per_at_bat
            FROM batting
            WHERE yearID > 1920
            GROUP BY yearID
            "

dbGetQuery(con, query) %>%
  ggplot(aes(x = yearID)) +
  geom_line(aes(y = xbh_per_at_bat, color = "xbh per at bat")) +
  geom_line(aes(y = strikeouts_per_at_bat, color = "strikeouts per at bat"))
```




```
dbGetQuery(con, query) %>%
  ggplot(aes(x = xbh_per_at_bat, y = strikeouts_per_at_bat)) +
  geom_point() +
  stat_smooth(geom='line', method = "lm", alpha=0.3, se=FALSE, color = "blue")
```

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




```
y <- dbGetQuery(con, query)$strikeouts_per_at_bat
x <- dbGetQuery(con, query)$xbh_per_at_bat
summary(lm(y ~ x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.045094 -0.011341 -0.000317  0.012877  0.057340
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.07501    0.01059  -7.083 2.25e-10 ***
## x            3.78145    0.17571  21.520  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01892 on 97 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.825
## F-statistic: 463.1 on 1 and 97 DF,  p-value: < 2.2e-16
```

There is certainly *some* correlation here - extra base hits and strikeouts have both been increasing in the MLB over time. But this doesn't confirm any specific hypothesis. What would we also expect to see as a trend?

- Intuitively, you would expect a player who strikes out more to have a lower salary, and someone who hits more extra base hits to have a higher salary. Perhaps extra base hits have a larger effect on salary than strikeouts, so players are incentivized to hit extra base hits at the expense of striking out more for the sake of making more money in contract negotiations.

- I also want to look at what is more correlated with scoring runs - strikeouts or extra base hits. Perhaps extra base hits have a larger effect on scoring runs than strikeouts, so players are incentivized to hit extra base hits at the expense of striking out more for the sake of winning more games.

Salary and Strikeouts

```
query <- "
         SELECT *
         FROM salaries
         LIMIT 5
        "

kable_query(con, query)
```

ID	yearID	teamID	team_ID	lgID	playerID	salary
1	1985	ATL	1918	NL	barkele01	870000
2	1985	ATL	1918	NL	bedrost01	550000
3	1985	ATL	1918	NL	benedbr01	545000
4	1985	ATL	1918	NL	campri01	633333
5	1985	ATL	1918	NL	ceronri01	625000

```
query <-  "
         SELECT
           playerID || ' ' || yearID AS player_and_year,
           salary
         FROM salaries
         LIMIT 5
         "

kable_query(con, query)
```

player_and_year	salary
barkele01 1985	870000
bedrost01 1985	550000
benedbr01 1985	545000
campri01 1985	633333
ceronri01 1985	625000

```r
query <-  "
          SELECT
            playerID || ' ' || yearID AS player_and_year,
            playerID,
            yearID,
            teamID,
            AB,
            R,
            H,
            batting.'2B',
            batting.'3B',
            HR,
            RBI,
            BB,
            SO
          FROM batting
          LIMIT 5
          "

kable_query(con, query)
```

player_and_year	playerID	yearID	teamID	AB	R	H	2B	3B	HR	RBI	BB	SO
abercda01 1871	abercda01	1871	TRO	4	0	0	0	0	0	0	0	0
addybo01 1871	addybo01	1871	RC1	118	30	32	6	0	0	13	4	0
allisar01 1871	allisar01	1871	CL1	137	28	40	4	5	0	19	2	5
allisdo01 1871	allisdo01	1871	WS3	133	28	44	10	2	2	27	0	2
ansonca01 1871	ansonca01	1871	RC1	120	29	39	11	3	0	16	2	1

Okay we've hit a roadblock. I am trying to combine these two tables together, but when I add certain "Where" conditions, it gets super slow. So I am going to try using dplyr to help me "filter".
That did not work either. Fixing it by limitting my rows and columns to

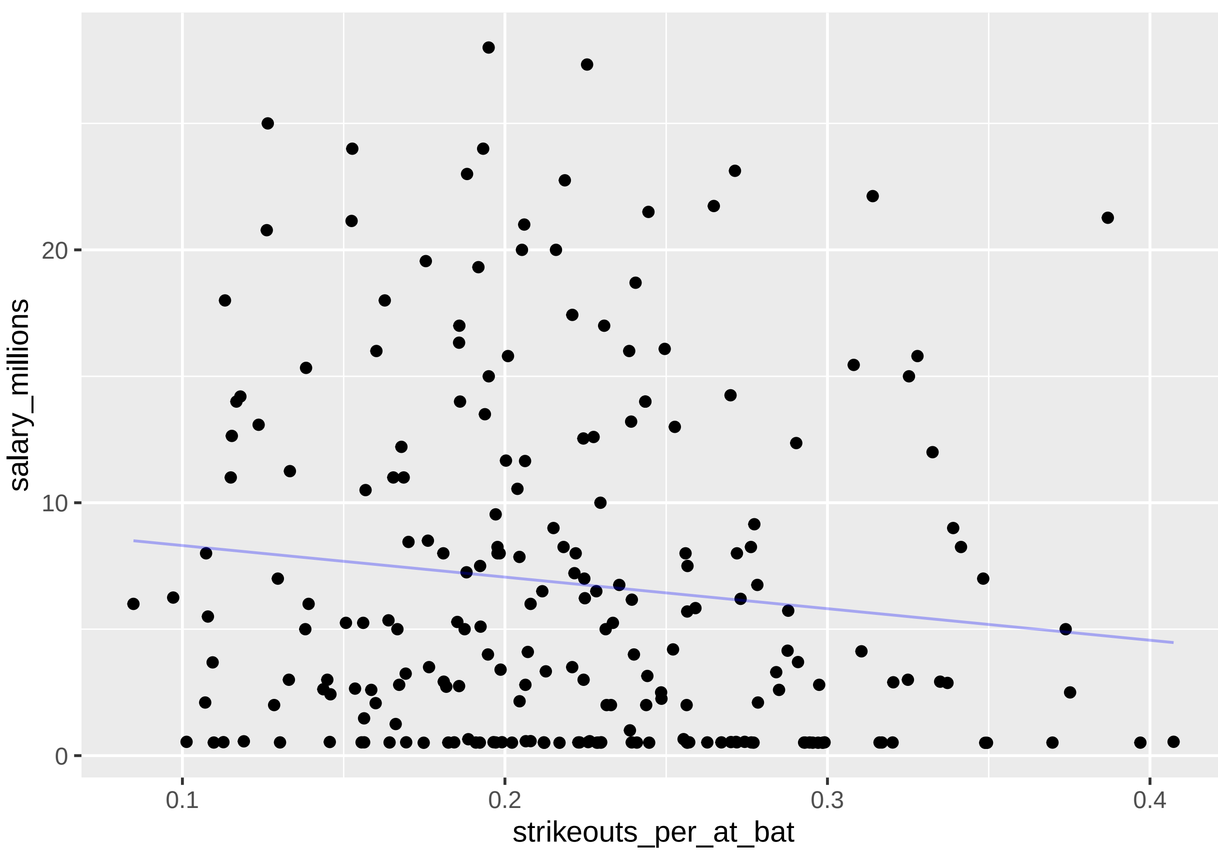
```r
query <-  "
      WITH select_batting AS (
          SELECT
            playerID || ' ' || yearID AS player_and_year,
            CAST(SO AS FLOAT) / AB AS strikeouts_per_at_bat,
            CAST(b.'3B' + b.'2B' + HR AS FLOAT) / AB AS xbh_per_at_bat,
            CAST(HR AS FLOAT) / AB AS hr_per_at_bat
          FROM batting AS b
          WHERE (AB > 300) AND (yearID IS 2016)
      ), select_salary AS (
          SELECT
            playerID || ' ' || yearID AS player_and_year,
            salary
          FROM salaries
          WHERE yearID IS 2016
      )
      SELECT
        sb.*,
        ss.salary /CAST(1000000 AS FLOAT) AS salary_millions
      FROM select_batting AS sb
      INNER JOIN select_salary AS ss ON ss.player_and_year = sb.player_and_year
```

```
      WHERE salary_millions > 0
          "

head(dbGetQuery(con, query))
```

```
##   player_and_year strikeouts_per_at_bat xbh_per_at_bat hr_per_at_bat
## 1  abreujo02 2016             0.2003205     0.09294872   0.040064103
## 2  alonsyo01 2016             0.1535270     0.08506224   0.014522822
## 3  altuvjo01 2016             0.1093750     0.11093750   0.037500000
## 4  alvarpe01 2016             0.2878338     0.12462908   0.065281899
## 5  andruel01 2016             0.1383399     0.09090909   0.015810277
## 6   aokino01 2016             0.1079137     0.07673861   0.009592326
##   salary_millions
## 1       11.666667
## 2        2.650000
## 3        3.687500
## 4        5.731704
## 5       15.333333
## 6        5.500000
```

```
dbGetQuery(con, query) %>%
  ggplot(aes(x = strikeouts_per_at_bat, y = salary_millions)) +
  geom_point() +
  stat_smooth(geom='line', method = "lm", alpha=0.3, se=FALSE, color = "blue")
```

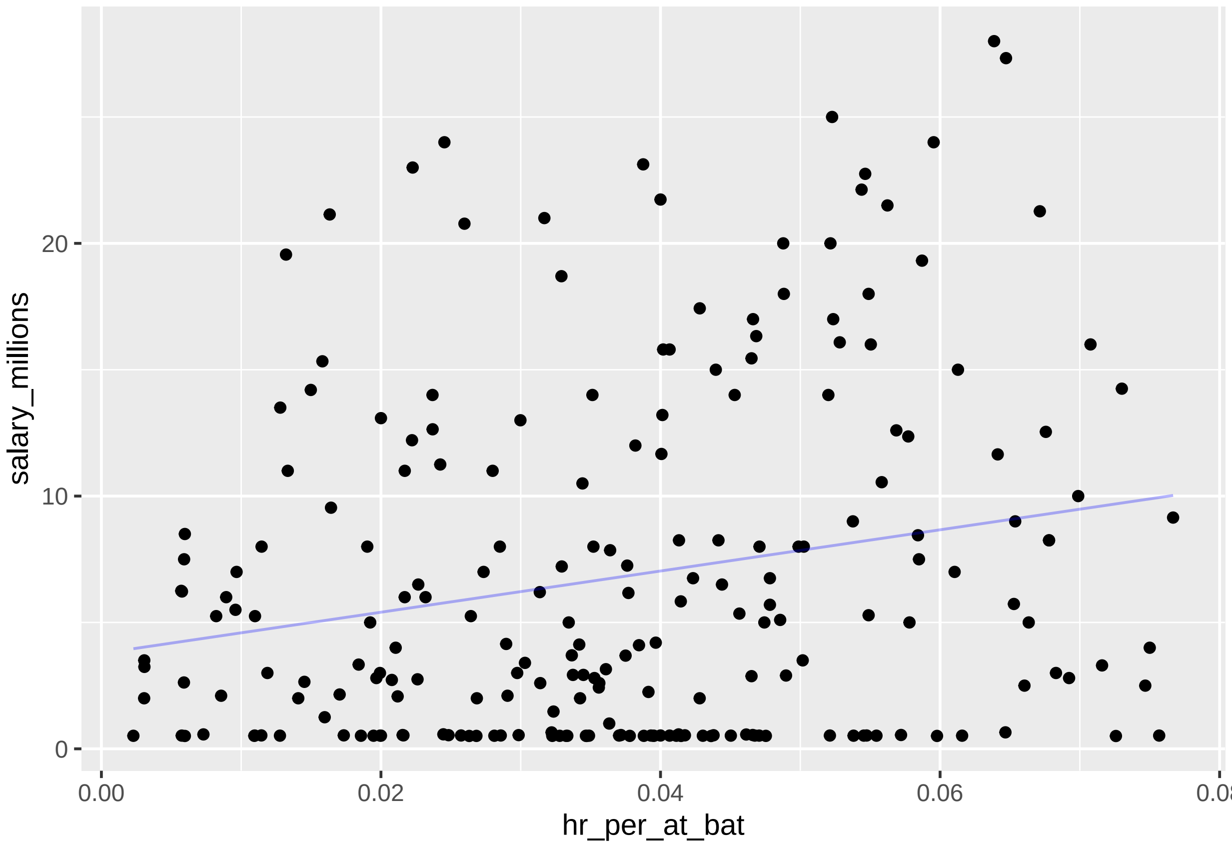
```
## `geom_smooth()` using formula 'y ~ x'
```



Obviously little to no correlation there. Interesting. To do this better I would want to know stats for a year before a contract is signed, but for now this is a fine approximation.

```
dbGetQuery(con, query) %>%
  ggplot(aes(x = hr_per_at_bat, y = salary_millions)) +
  geom_point() +
  stat_smooth(geom='line', method = "lm", alpha=0.3, se=FALSE, color = "blue")
```

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



Not much there either.

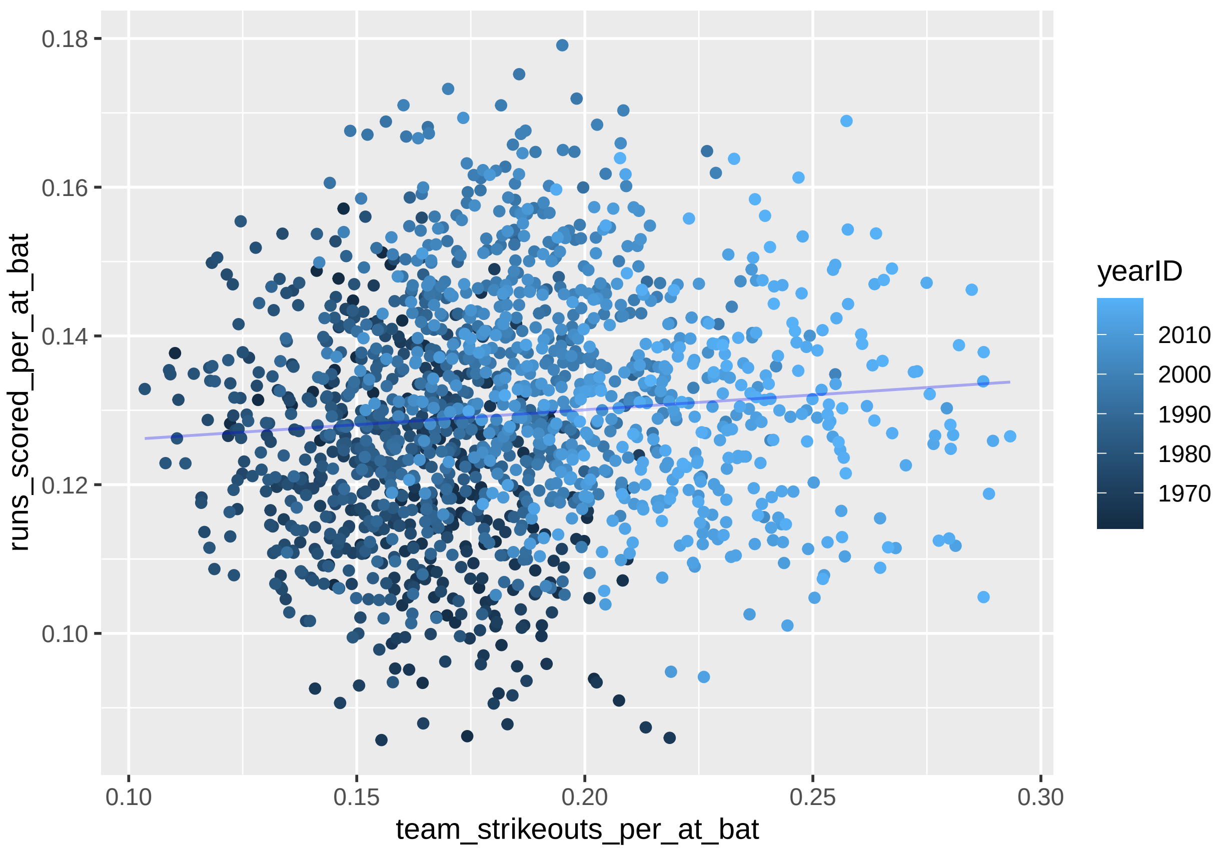
Now I'm going to see the correlation of strikeouts with runs scored, and the correlation of xbh with runs scroed for teams in certain years.

```
query <-  "
          SELECT
            yearID,
            teamID,
            CAST(SUM(R) AS FLOAT) / SUM(AB)  AS runs_scored_per_at_bat,
            CAST(SUM(SO) AS FLOAT) / SUM(AB) AS team_strikeouts_per_at_bat,
            SUM(b.'2B') + SUM(b.'3B') + SUM(HR) / CAST(SUM(AB) AS FLOAT) AS team_xbh_per_at_bat
          FROM batting AS b
          WHERE yearID > 1960
          GROUP BY yearID, teamID
          "

head(dbGetQuery(con, query))
```

```
##   yearID teamID runs_scored_per_at_bat team_strikeouts_per_at_bat
## 1   1961    BAL              0.1260719                  0.1645685
## 2   1961    BOS              0.1323529                  0.1535948
## 3   1961    CHA              0.1376890                  0.1101512
## 4   1961    CHN              0.1289296                  0.1921781
## 5   1961    CIN              0.1354187                  0.1451459
## 6   1961    CLE              0.1313960                  0.1283651
##   team_xbh_per_at_bat
## 1            263.0272
## 2            288.0203
## 3            262.0248
## 4            289.0329
## 5            282.0301
## 6            296.0267
```

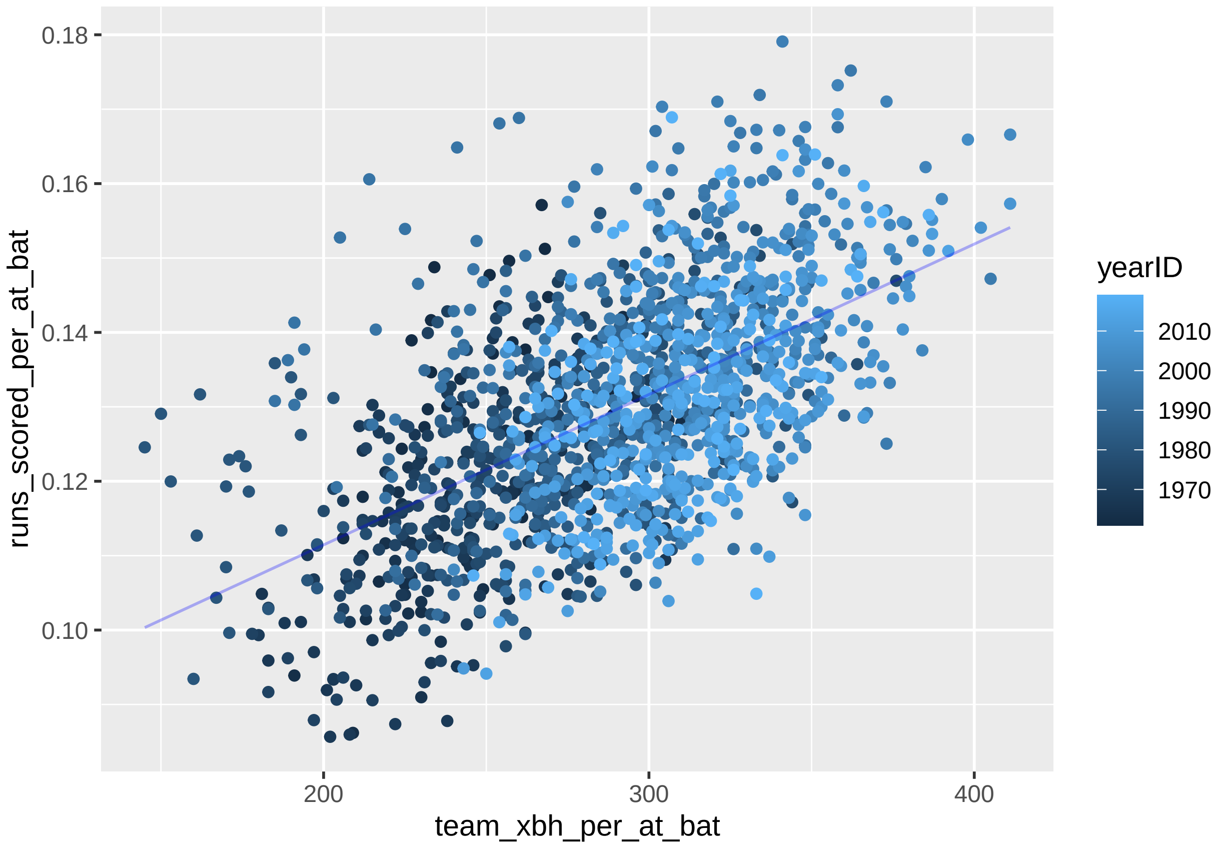
```
dbGetQuery(con, query) %>%
  ggplot(aes(x = team_strikeouts_per_at_bat, y = runs_scored_per_at_bat)) +
  geom_point(aes(color = yearID)) +
  stat_smooth(geom='line', method = "lm", alpha=0.3, se=FALSE, color = "blue")
```

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




```
dbGetQuery(con, query) %>%
  ggplot(aes(x = team_xbh_per_at_bat, y = runs_scored_per_at_bat)) +
  geom_point(aes(color = yearID)) +
  stat_smooth(geom='line', method = "lm", alpha=0.3, se=FALSE, color = "blue")
```

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



SO I think this is interesting. It seems as though strikeouts per at bat has nearly no correlation with the amount of runs a team scores, while extra base hits has a decent, clear correlation.

Let's see if this relationship holds in, say, the 2000s

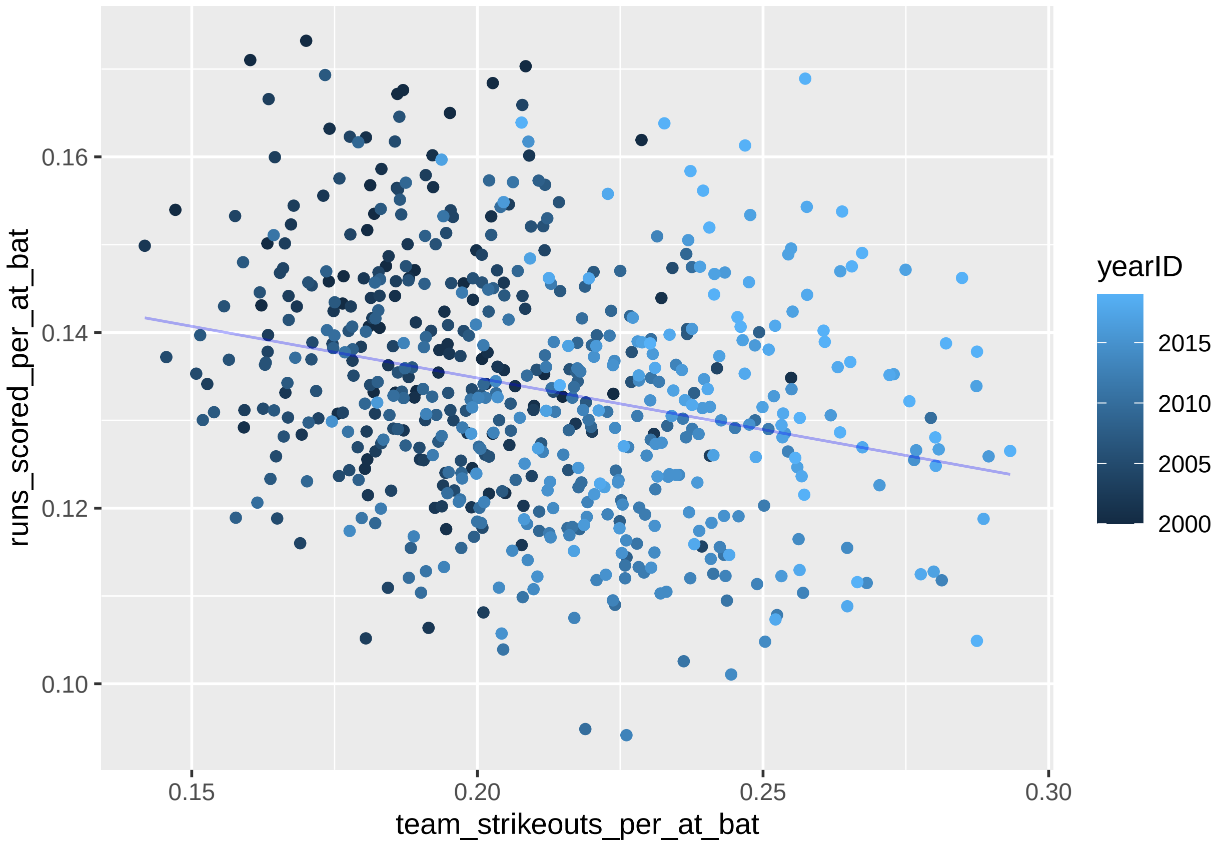
```
query <-  "
          SELECT
            yearID,
            teamID,
            CAST(SUM(R) AS FLOAT) / SUM(AB)  AS runs_scored_per_at_bat,
            CAST(SUM(SO) AS FLOAT) / SUM(AB) AS team_strikeouts_per_at_bat,
            SUM(b.'2B') + SUM(b.'3B') + SUM(HR) / CAST(SUM(AB) AS FLOAT) AS team_xbh_per_at_bat
          FROM batting AS b
          WHERE yearID > 1999
          GROUP BY yearID, teamID
          "

head(dbGetQuery(con, query))
```

```
##   yearID teamID runs_scored_per_at_bat team_strikeouts_per_at_bat
## 1   2000    ANA              0.1535181                  0.1819474
## 2   2000    ARI              0.1432965                  0.1764067
## 3   2000    ATL              0.1475679                  0.1840044
## 4   2000    BAL              0.1430888                  0.1621914
## 5   2000    BOS              0.1406750                  0.1809947
## 6   2000    CHA              0.1732200                  0.1700319
##   team_xbh_per_at_bat
## 1            343.0419
```

```
## 2             326.0324
## 3             300.0326
## 4             332.0332
## 5             348.0297
## 6             358.0383
```

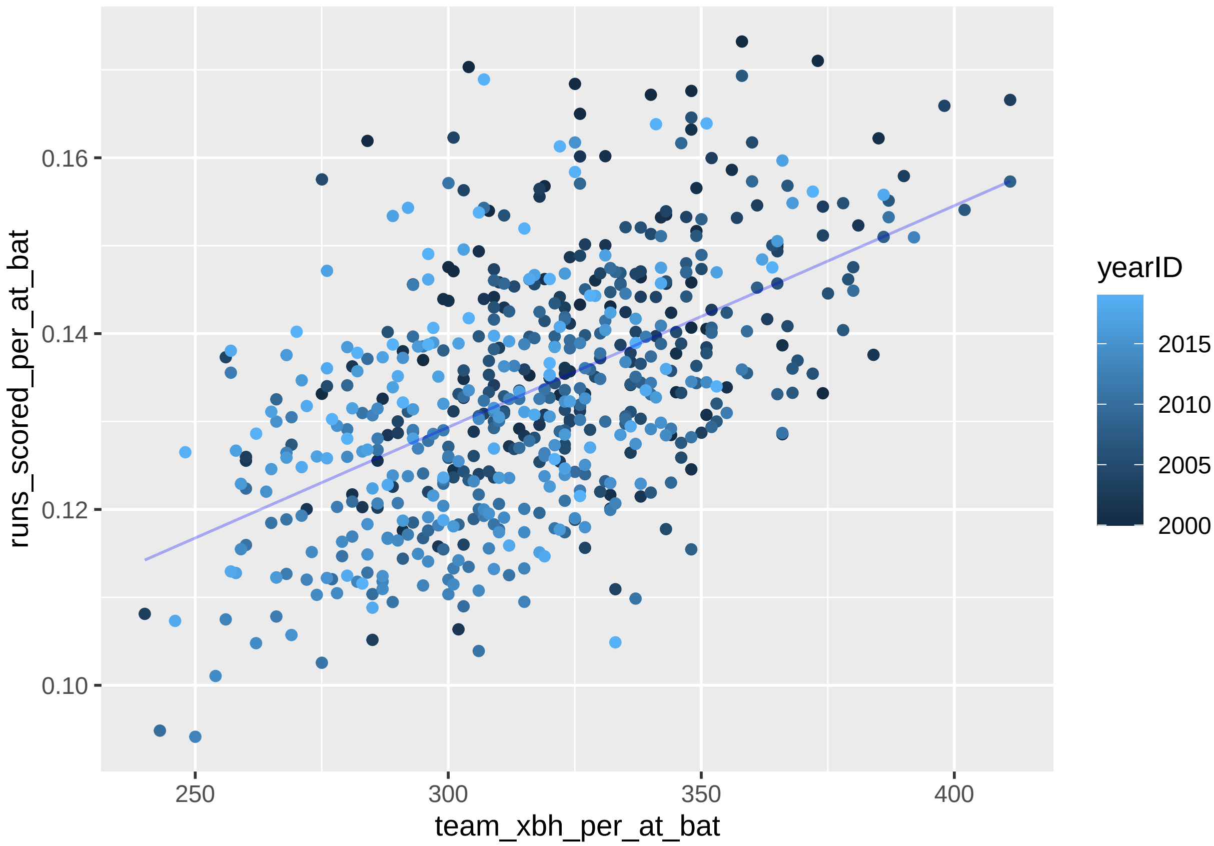
```r
dbGetQuery(con, query) %>%
  ggplot(aes(x = team_strikeouts_per_at_bat, y = runs_scored_per_at_bat)) +
  geom_point(aes(color = yearID)) +
  stat_smooth(geom='line', method = "lm", alpha=0.3, se=FALSE, color = "blue")
```

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




```r
dbGetQuery(con, query) %>%
  ggplot(aes(x = team_xbh_per_at_bat, y = runs_scored_per_at_bat)) +
  geom_point(aes(color = yearID)) +
  stat_smooth(geom='line', method = "lm", alpha=0.3, se=FALSE, color = "blue")
```

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



The correlation of strikeouts to runs is still weaker than the correlation between extra base hits and runs. Which points to the idea that teams/players would be willing to risk getting x more strikeouts if it came with getting x more extra base hits in a season.

```
y <- dbGetQuery(con, query)$runs_scored_per_at_bat
x <- dbGetQuery(con, query)$team_strikeouts_per_at_bat

#Prints the R-squared of the regression of runs scored on strikeouts
summary(lm(y ~ x))$r.squared
```

```
## [1] 0.06459704
```

```
y <- dbGetQuery(con, query)$runs_scored_per_at_bat
x <- dbGetQuery(con, query)$team_xbh_per_at_bat

#Prints the R-squared of the regression of runs scored on extra base hits
summary(lm(y ~ x))$r.squared
```

```
## [1] 0.2934591
```

There seems to be some legs here. Another trend I would expect to see is that batters who have increased the average exit velocities of their hits have gotten more extra base hits and also strike out more. And I would also expect to see the average exit velocity for hitters in the MLB to be increasing over time. So I need to see if I can find some new data to investigate these hypotheses.

I am pretty sure the MLB started keeping stats like exit velocity around 2000, so I am going to do a little hunting on the internet.

I found some data from Fangraphs from 2008-2019 which includes a Hard Hit %, which I think will act similarly to exit velocity in analysis.

```r
fangraphs_batting <- read_csv("fangraphs_batting.csv")

head(fangraphs_batting)
```

```
## # A tibble: 6 x 11
##   Season Name    Team     HR     R    AB `2B` `3B`    SO Hard_Percent playerid
##    <dbl> <chr>   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>        <dbl>    <dbl>
## 1   2015 Bryce ~ Nati~    42   118   521    38     1   131        0.409    11579
## 2   2008 Albert~ Card~    37   100   524    44     0    54        0.429     1177
## 3   2013 Miguel~ Tige~    44   103   555    26     1    94        0.451     1744
## 4   2018 Mookie~ Red ~    32   129   520    47     5    91        0.445    13611
## 5   2009 Albert~ Card~    47   124   568    45     1    64        0.406     1177
## 6   2018 Mike T~ Ange~    39   101   471    24     4   124        0.444    10155
```

```r
fangraphs_batting %>%
  group_by(Season) %>%
  summarise(count = n())
```

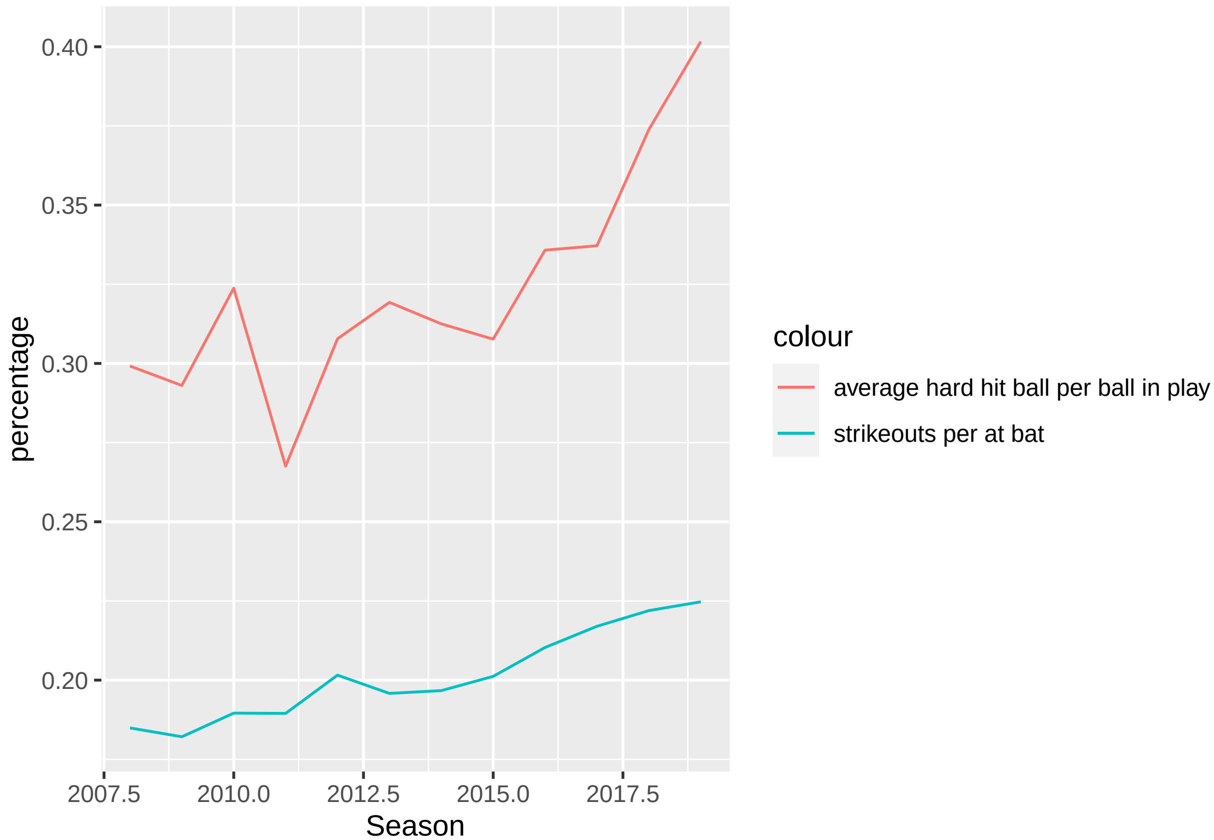
```
## # A tibble: 12 x 2
##    Season count
##     <dbl> <int>
##  1   2008   147
##  2   2009   154
##  3   2010   149
##  4   2011   145
##  5   2012   143
##  6   2013   140
##  7   2014   146
##  8   2015   141
##  9   2016   146
## 10   2017   144
## 11   2018   140
## 12   2019   135
```

There are a similar number of observations for every season. Let's see if strikeouts/hard hit % have been increasing for this sample of players.

```r
advanced_batting <- fangraphs_batting %>%
  group_by(Season) %>%
  summarise(count = n(),
            so_per_ab = sum(SO) / sum(AB),
            avg_hard_percentage = mean(Hard_Percent))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

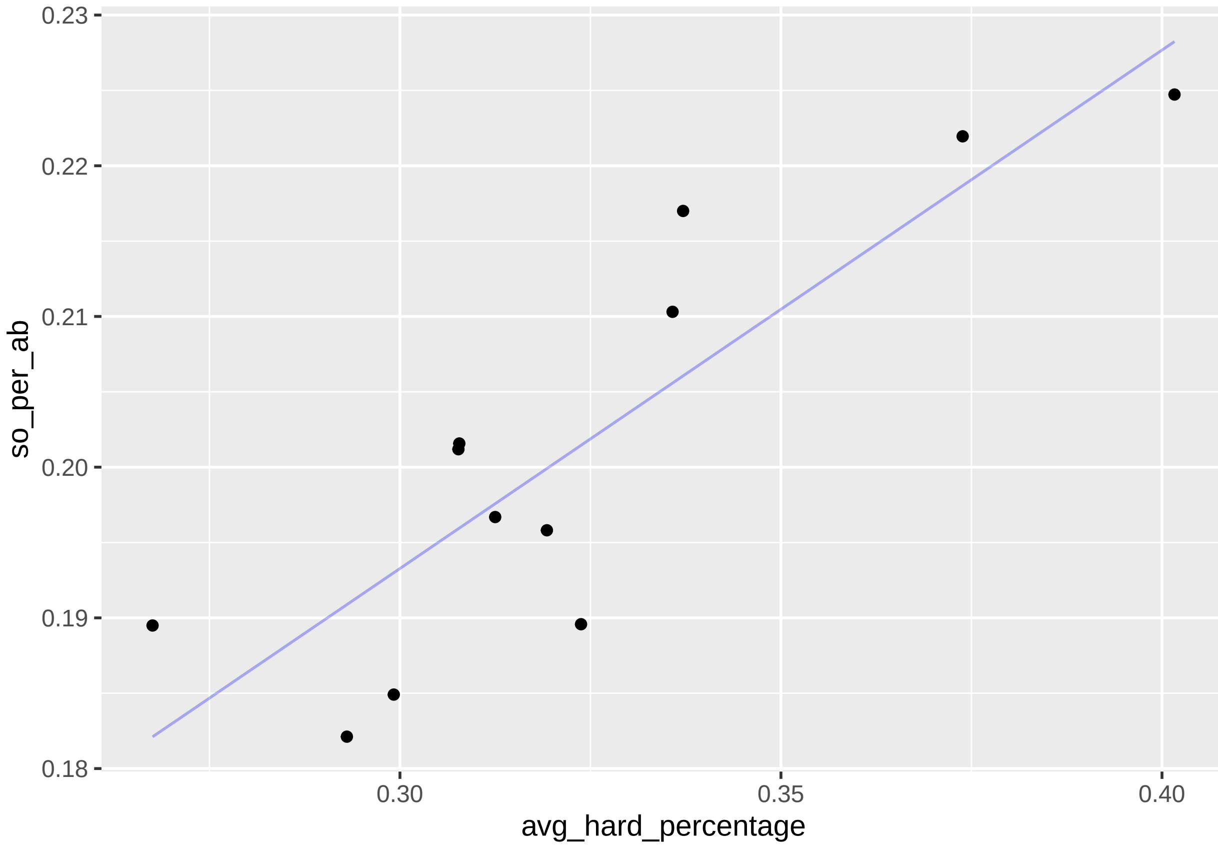
```r
advanced_batting %>%
  ggplot(aes(x = Season)) +
  geom_line(aes(y = so_per_ab, color = "strikeouts per at bat")) +
  geom_line(aes(y = avg_hard_percentage, color = "average hard hit ball per ball in play")) +
  ylab("percentage")
```



The strikeout pattern is not the same as the one for all players, but it is quite similar. Also, the hard hit percentage has risen dramatically, especially in the last few years, as the strikeout percentage has similarly been rising dramatically.

```
advanced_batting %>%
  ggplot(aes(x = avg_hard_percentage, y = so_per_ab)) +
  geom_point() +
  stat_smooth(geom='line', method = "lm", alpha=0.3, se=FALSE, color = "blue")
```

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




```
y <- advanced_batting$so_per_ab
x <- advanced_batting$avg_hard_percentage

#Prints the R-squared of the regression of strikeouts on hard hit balls
summary(lm(y ~ x))$r.squared
```

```
## [1] 0.7443516
```

Can we see this pattern with individuals?
This is going to take a bit of data manipulation.

```
select_years_list <- fangraphs_batting %>%
  group_by(Name) %>%
  summarise(count = n()) %>%
  filter(count %in% c(8, 9))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
select_years_list
```

```
## # A tibble: 30 x 2
##    Name               count
##    <chr>              <int>
##  1 Adrian Beltre          8
##  2 Adrian Gonzalez        9
##  3 Alcides Escobar        9
##  4 Alex Gordon            9
##  5 Alexei Ramirez         9
##  6 Andrew McCutchen       9
```

```
##  7 Ben Zobrist          9
##  8 Brandon Phillips     9
##  9 Brett Gardner        9
## 10 Carlos Santana       9
## # ... with 20 more rows
```

I am going to use this group of players, some who are reaching the tail end of their career, and some who are reaching the peak years of their career. I would expect the correlation between changes in hard hit % and changes in strikeouts ber at bat to be fairly strong for most of these players, but stronger for those reaching the tail end of their career, as the younger players are still gaining strength, and are potentially able to hit more hard balls without changing their swing/approach.

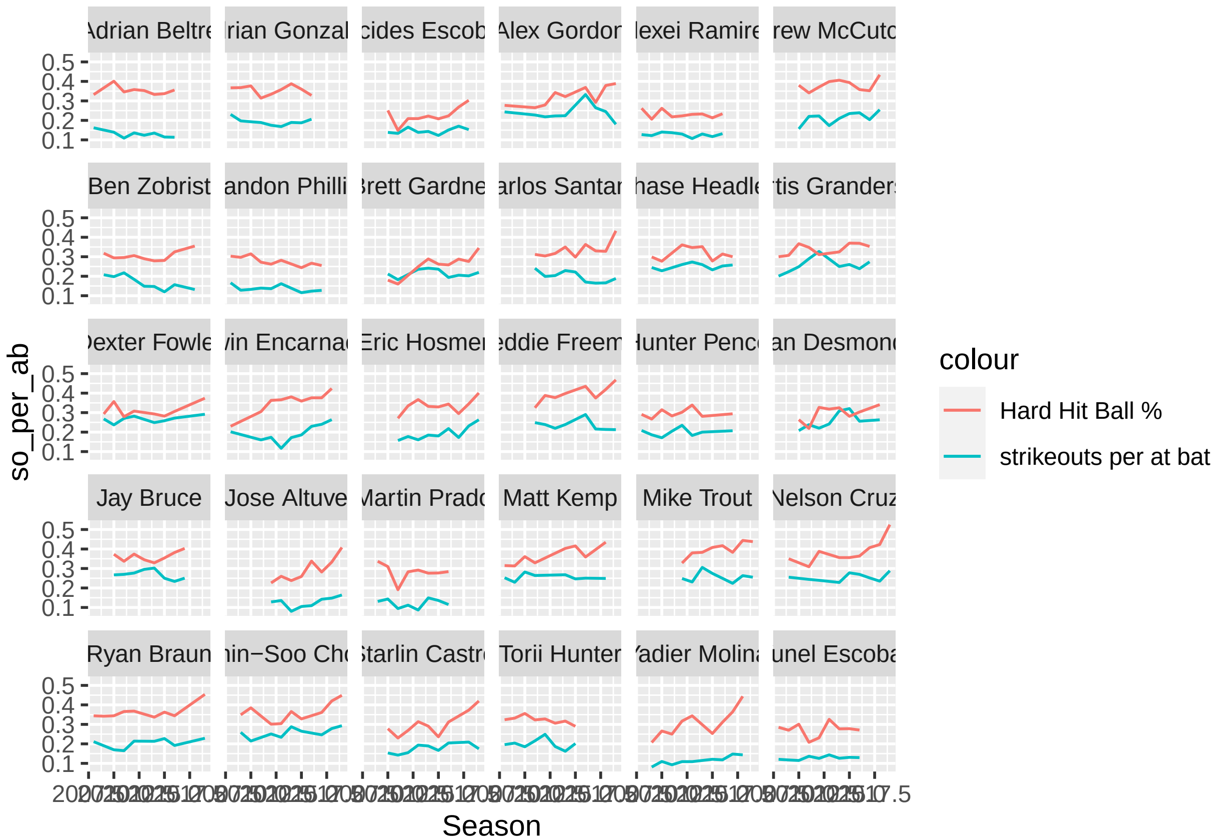
```
select_years_list$Name
```

```
##  [1] "Adrian Beltre"     "Adrian Gonzalez"   "Alcides Escobar"
##  [4] "Alex Gordon"       "Alexei Ramirez"    "Andrew McCutchen"
##  [7] "Ben Zobrist"       "Brandon Phillips"  "Brett Gardner"
## [10] "Carlos Santana"    "Chase Headley"     "Curtis Granderson"
## [13] "Dexter Fowler"     "Edwin Encarnacion" "Eric Hosmer"
## [16] "Freddie Freeman"   "Hunter Pence"      "Ian Desmond"
## [19] "Jay Bruce"         "Jose Altuve"       "Martin Prado"
## [22] "Matt Kemp"         "Mike Trout"        "Nelson Cruz"
## [25] "Ryan Braun"        "Shin-Soo Choo"     "Starlin Castro"
## [28] "Torii Hunter"      "Yadier Molina"     "Yunel Escobar"
```

```
fangraphs_batting %>%
  filter(Name %in% select_years_list$Name) %>%
  mutate(so_per_ab = SO/AB) %>%
  select(Name, Season, so_per_ab, Hard_Percent)
```

```
## # A tibble: 258 x 4
##    Name        Season so_per_ab Hard_Percent
##    <chr>        <dbl>     <dbl>        <dbl>
##  1 Mike Trout    2018     0.263        0.444
##  2 Mike Trout    2017     0.224        0.383
##  3 Mike Trout    2019     0.255        0.438
##  4 Ryan Braun    2011     0.165        0.366
##  5 Mike Trout    2013     0.231        0.38
##  6 Mike Trout    2016     0.250        0.417
##  7 Nelson Cruz   2019     0.289        0.525
##  8 Mike Trout    2015     0.275        0.408
##  9 Ryan Braun    2012     0.214        0.368
## 10 Matt Kemp     2011     0.264        0.329
## # ... with 248 more rows
```

```
fangraphs_batting %>%
  filter(Name %in% select_years_list$Name) %>%
  mutate(so_per_ab = SO/AB) %>%
  select(Name, Season, so_per_ab, Hard_Percent) %>%
  ggplot(aes(x = Season)) +
  geom_line(aes(y = so_per_ab, color = "strikeouts per at bat")) +
  geom_line(aes(y = Hard_Percent, color = "Hard Hit Ball %")) +
  facet_wrap(~ Name)
```



This visualization did not help much, but it look pretty cool.

I still think my hypothesis has legs. Just to make it crystal clear, my hypothesis is as follows:

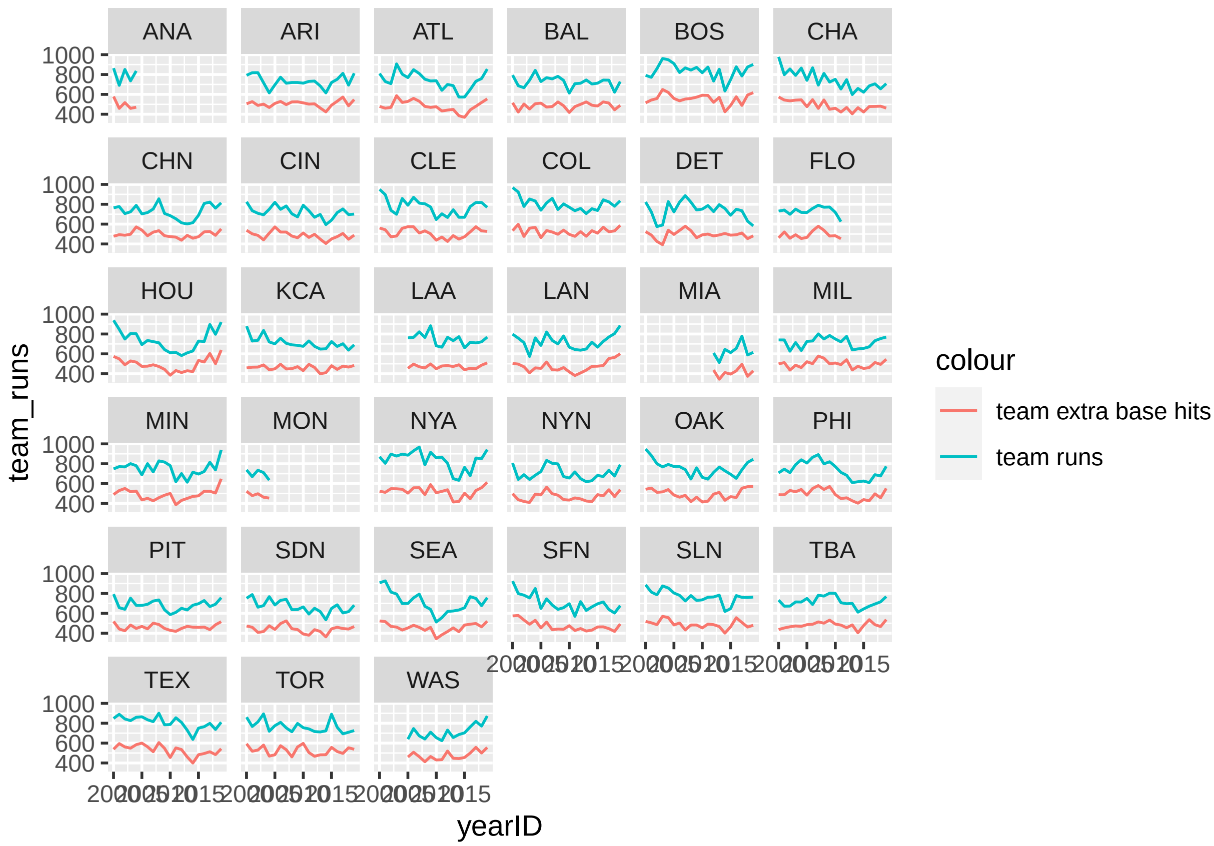
A FUNDAMENTAL CHANGE IN HITTING MINDSET IN THE MLB IS RESPONSIBLE FOR THE HUGE INCREASE IN STRIKEOUTS IN RECENT YEARS. SPECIFICALLY, PLAYERS ARE CONSCIOUSLY SWINGING HARDER, KNOWING THAT THEY MAY WHIFF MORE OFTEN, BUT ALSO KNOWING THAT THEY WILL GET MORE EXTRA BASE HITS. THIS CHANGE IN MINDSET IS MOTIVATED BY GOAL OF SCORING AS MANY RUNS AS POSSIBLE, BECAUSE, ALL ELSE EQUAL, MORE RUNS MEANS MORE WINS.

I want to make a graph that shows team strikeouts, extra base hits, and runs scored since 2000. Let's see if I can get that.

```
query <-  "
          SELECT
            yearID,
            teamID,
            SUM(SO) AS team_strikeouts,
            SUM(R) AS team_runs,
            SUM(b.'2B') + SUM(b.'3B') + SUM(HR) AS team_xbh
          FROM batting AS b
          WHERE yearID > 1999
          GROUP BY yearID, teamID
          "

head(dbGetQuery(con, query))
```

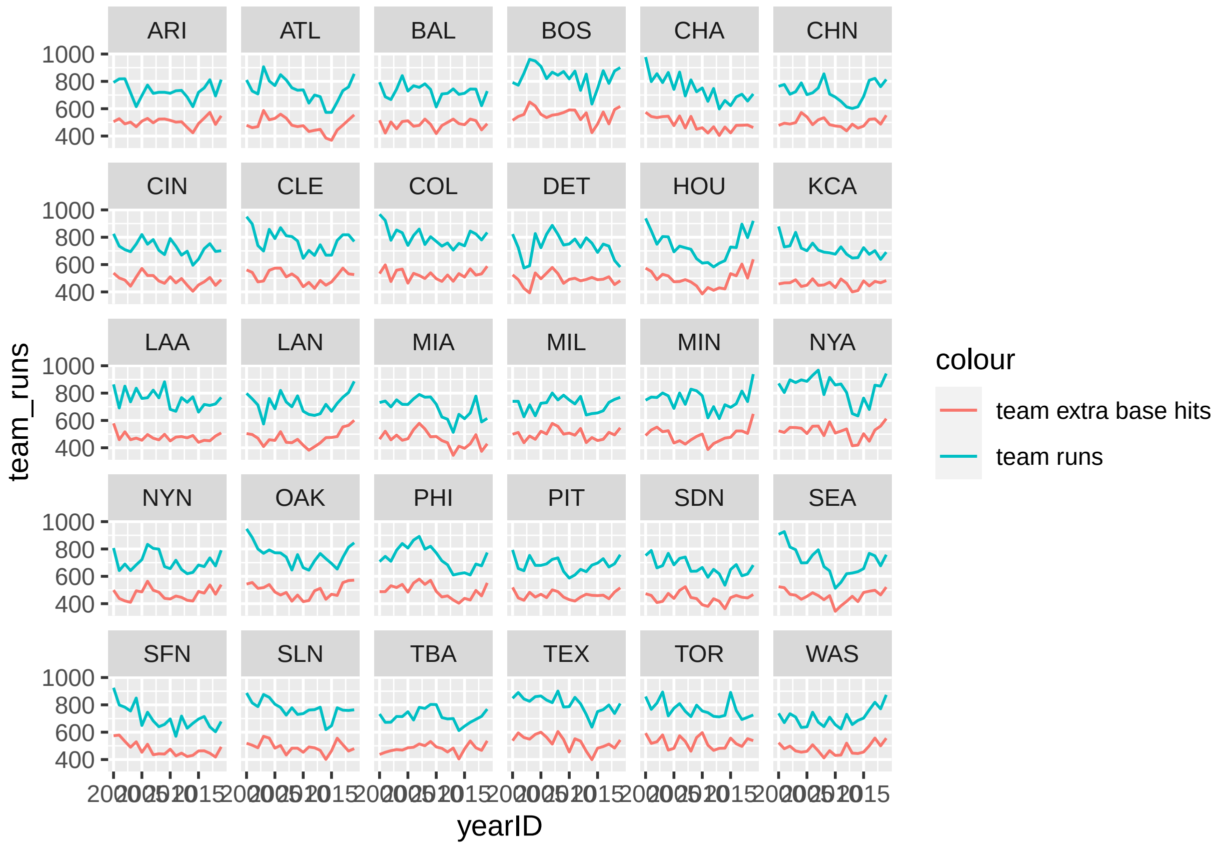
```
##   yearID teamID team_strikeouts team_runs team_xbh
## 1   2000    ANA            1024       864      579
## 2   2000    ARI             975       792      505
## 3   2000    ATL            1010       810      479
## 4   2000    BAL             900       794      516
## 5   2000    BOS            1019       792      515
## 6   2000    CHA             960       978      574
```

```
dbGetQuery(con, query) %>%
  ggplot(aes(x = yearID)) +
  geom_line(aes(y = team_runs, color = "team runs")) +
  geom_line(aes(y = team_xbh, color = "team extra base hits")) +
  facet_wrap(~ teamID)
```



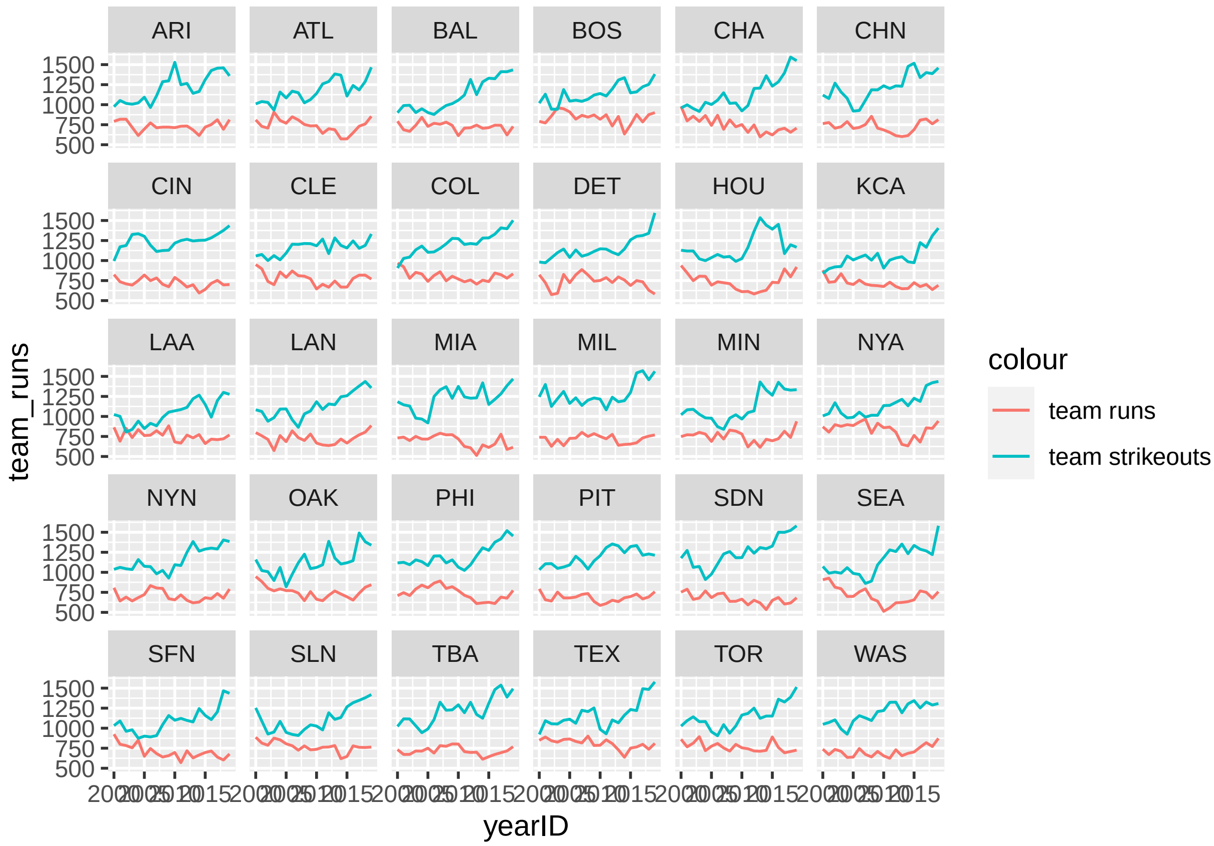
OK this plot is unreal. I am going to add the teams that changed cities together in order to make it more uniform and even easier to visualize.

```
dbGetQuery(con, query) %>%
  mutate(teamID = ifelse(teamID == "ANA", "LAA",
                         ifelse(teamID == "MON", "WAS",
                                ifelse(teamID =="FLO", "MIA", teamID)))) %>%
  ggplot(aes(x = yearID)) +
  geom_line(aes(y = team_runs, color = "team runs")) +
  geom_line(aes(y = team_xbh, color = "team extra base hits")) +
  facet_wrap(~ teamID)
```



Okay awesome. Now I am going to look at the same graph, but see how correlated strikeouts are with runs scored. As we saw in an earlier scatterplot, there should be little to no correlation evident.

```
dbGetQuery(con, query) %>%
  mutate(teamID = ifelse(teamID == "ANA", "LAA",
                         ifelse(teamID == "MON", "WAS",
                                ifelse(teamID =="FLO", "MIA", teamID)))) %>%
  ggplot(aes(x = yearID)) +
  geom_line(aes(y = team_runs, color = "team runs")) +
  geom_line(aes(y = team_strikeouts, color = "team strikeouts")) +
  facet_wrap(~ teamID)
```



Pretty much every teams' strikeouts are increasing season over season, but some teams are not seeing the tradeoff expected with an increase in xbh and runs. For that reason I suspect that there may be another variable adding to the increase in strikeouts - pitcher skill.

To me, it would make sense that if pitchers are outpacing batters in improving their talent year over year, the result would be more strikeouts for everyone. Let's see if we can determine what makes a pitcher induce more strikeouts using our Lahman pitching data.

```
query <-  "
          SELECT *
          FROM pitching
          LIMIT 5
          "

dbGetQuery(con, query)
```

```
##   ID  playerID yearID stint teamID team_ID lgID  W  L  G GS CG SHO SV IPouts
## 1  1 bechtge01   1871     1    PH1       6   NA  1  2  3  3  2   0  0     78
## 2  2 brainas01   1871     1    WS3       9   NA 12 15 30 30 30   0  0    792
## 3  3 fergubo01   1871     1    NY2       5   NA  0  0  1  0  0   0  0      3
## 4  4 fishech01   1871     1    RC1       7   NA  4 16 24 24 22   1  0    639
## 5  5 fleetfr01   1871     1    NY2       5   NA  0  1  1  1  1   0  0     27
##     H  ER HR BB SO BAOpp   ERA IBB WP HBP BK  BFP GF   R SH SF GIDP
## 1  43  23  0 11  1    NA  7.96  NA  7  NA  0  146  0  42 NA NA   NA
## 2 361 132  4 37 13    NA  4.50  NA  7  NA  0 1291  0 292 NA NA   NA
## 3   8   3  0  0  0    NA 27.00  NA  2  NA  0   14  0   9 NA NA   NA
## 4 295 103  3 31 15    NA  4.35  NA 20  NA  0 1080  1 257 NA NA   NA
## 5  20  10  0  3  0    NA 10.00  NA  0  NA  0   57  0  21 NA NA   NA
```

IPOuts appears to be a stat showing the number of outs that pitcher got that season. Therefore, SO/IPOuts would be the percentage of outs that were strikeouts. So I can run a fairly simple query that will give me the top strikeout pitchers of all time.

```
query <-  "
          SELECT
            pe.nameFirst || ' ' || pe.nameLast AS name,
            pe.debut AS first_game_date,
            SUM(GS) AS career_starts,
            SUM(IPouts) AS career_outs,
            SUM(SO) AS career_strikeouts,
            ROUND(CAST(SUM(SO) AS FLOAT) / SUM(IPouts), 3) AS strikeout_percentage
          FROM pitching AS pi
          INNER JOIN people AS pe ON pe.playerID = pi.playerID
          WHERE IPouts > 100
          GROUP BY pi.playerID
          ORDER BY strikeout_percentage DESC
          LIMIT 20
          "

kable_query(con, query)
```

name	first_game_date	career_starts	career_outs	career_strikeouts	strikeout_percentage
Josh Hader	2017-06-10	0	614	349	0.568
Aroldis Chapman	2010-08-31	0	1393	774	0.556
Josh James	2018-09-01	1	184	100	0.543
Dellin Betances	2011-09-22	0	1120	607	0.542
Craig Kimbrel	2010-05-07	0	1536	828	0.539
Edwin Diaz	2016-06-06	0	747	400	0.535
Nick Anderson	2019-03-28	0	131	69	0.527
Tanner Rainey	2018-04-10	0	145	74	0.510
Corey Knebel	2014-05-24	0	545	272	0.499
Kenley Jansen	2010-07-24	0	1754	862	0.491
Kirby Yates	2014-06-07	0	770	370	0.481
Jose Leclerc	2016-07-06	3	516	245	0.475
Tanner Scott	2017-09-20	0	160	76	0.475
Trey Wingenter	2018-08-07	1	153	72	0.471
Ken Giles	2014-06-12	0	891	419	0.470
Colin Poche	2019-06-08	0	155	72	0.465
Carl Edwards	2015-09-07	0	463	213	0.460
Brad Boxberger	2012-06-10	0	543	249	0.459
Rob Dibble	1988-06-29	0	1352	619	0.458
Ernesto Frieri	2009-09-26	0	558	254	0.455

Some clear observations here:

- These guys are pretty much all relief pitchers
- The list is heavily populated with current MLB pitchers
- Of the names I recognize, these pitchers throw HEAT

I have downloaded some pitching data from Fangraphs. Now I can look at the correlation between fastball speed and strikeouts more precisely.

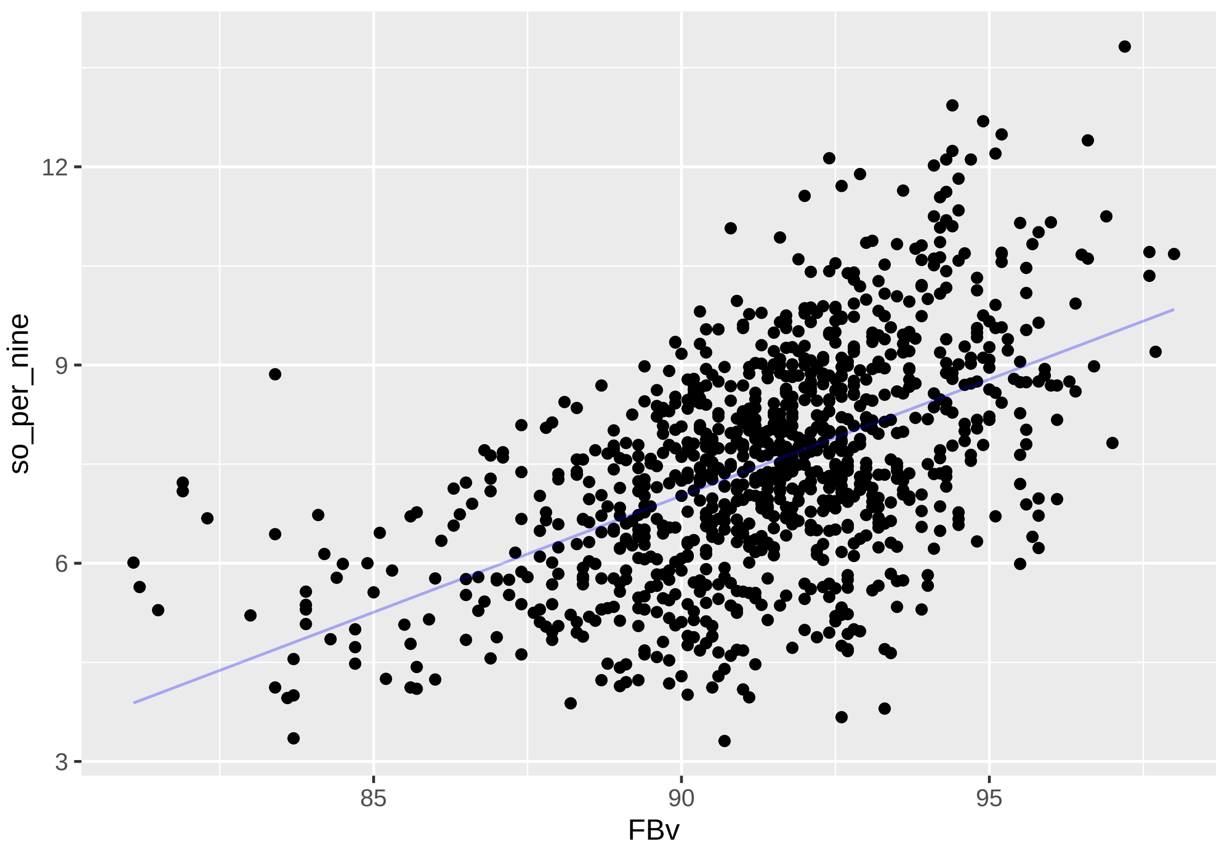
```r
fangraphs_pitching <- read_csv("fangraphs_pitching.csv")

head(fangraphs_pitching)
```

```
## # A tibble: 6 x 10
##   Season Name       Team      G    GS    IP so_per_nine FB_percent   FBv playerid
##    <dbl> <chr>      <chr> <dbl> <dbl> <dbl>       <dbl>      <dbl> <dbl>    <dbl>
## 1   2015 Zack Gre~ Dodg~    32    32  222.         8.08      0.507  91.8     1943
## 2   2018 Jacob de~ Mets     32    32  217         11.2       0.521  96      10954
## 3   2015 Jake Arr~ Cubs     33    33  229          9.28      0.507  94.6     4153
## 4   2014 Clayton ~ Dodg~    27    27  198.        10.8       0.554  93       2036
## 5   2013 Clayton ~ Dodg~    33    33  236          8.85      0.607  92.6     2036
## 6   2018 Blake Sn~ Rays     31    31  180.        11.0       0.515  95.8    13543
```

All I have to do is see if FBv (average fastball velocity) is correlated to so_per_nine (strikeouts per 9 innings/27 outs). I am also going to remove knuckle ball pitchers from the list (fastballs slower than 80mph)

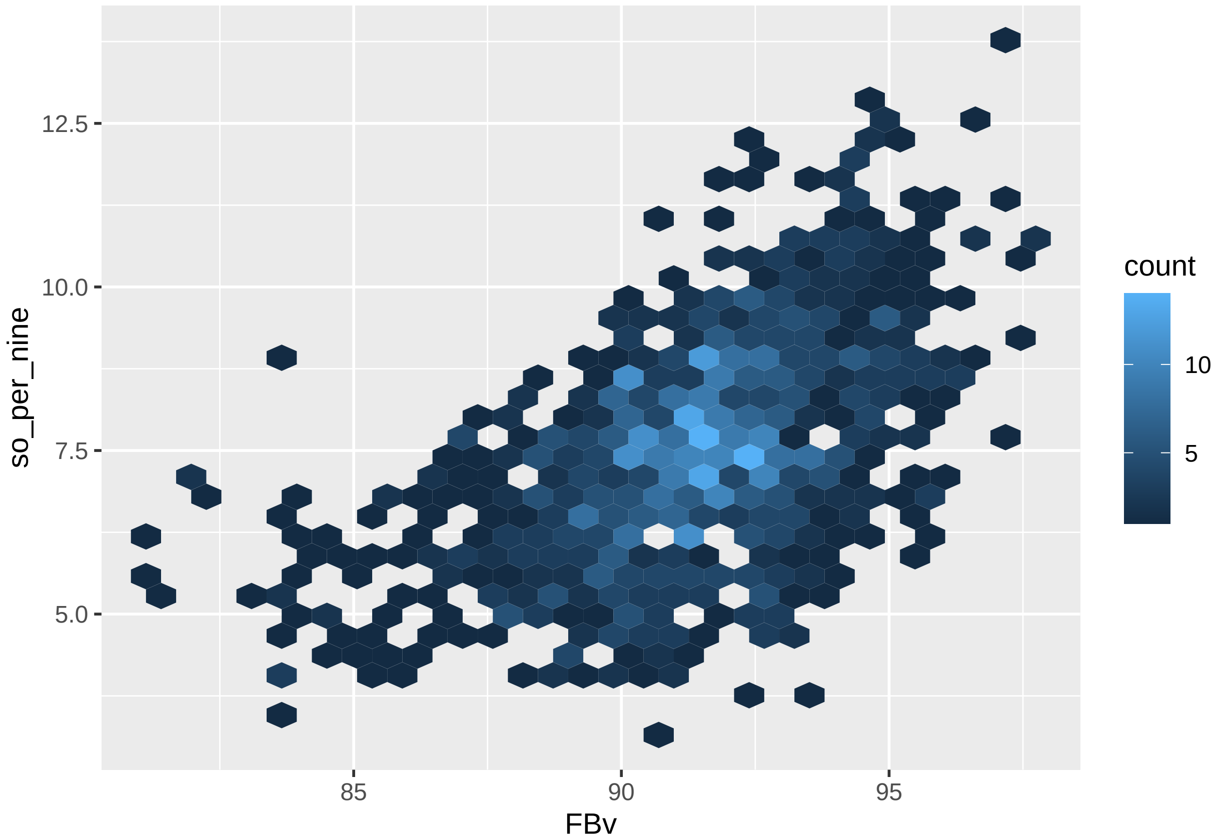
```r
fangraphs_pitching %>%
  filter(FBv > 80) %>%
  ggplot(aes(x = FBv, y = so_per_nine)) +
  geom_point() +
  stat_smooth(geom='line', method = "lm", alpha=0.3, se=FALSE, color = "blue")
```

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




```r
# Hexbin library allows me to use geom_hex
library(hexbin)
fangraphs_pitching %>%
```

```
  filter(FBv > 80) %>%
  ggplot(aes(x = FBv, y = so_per_nine)) +
  geom_hex()
```



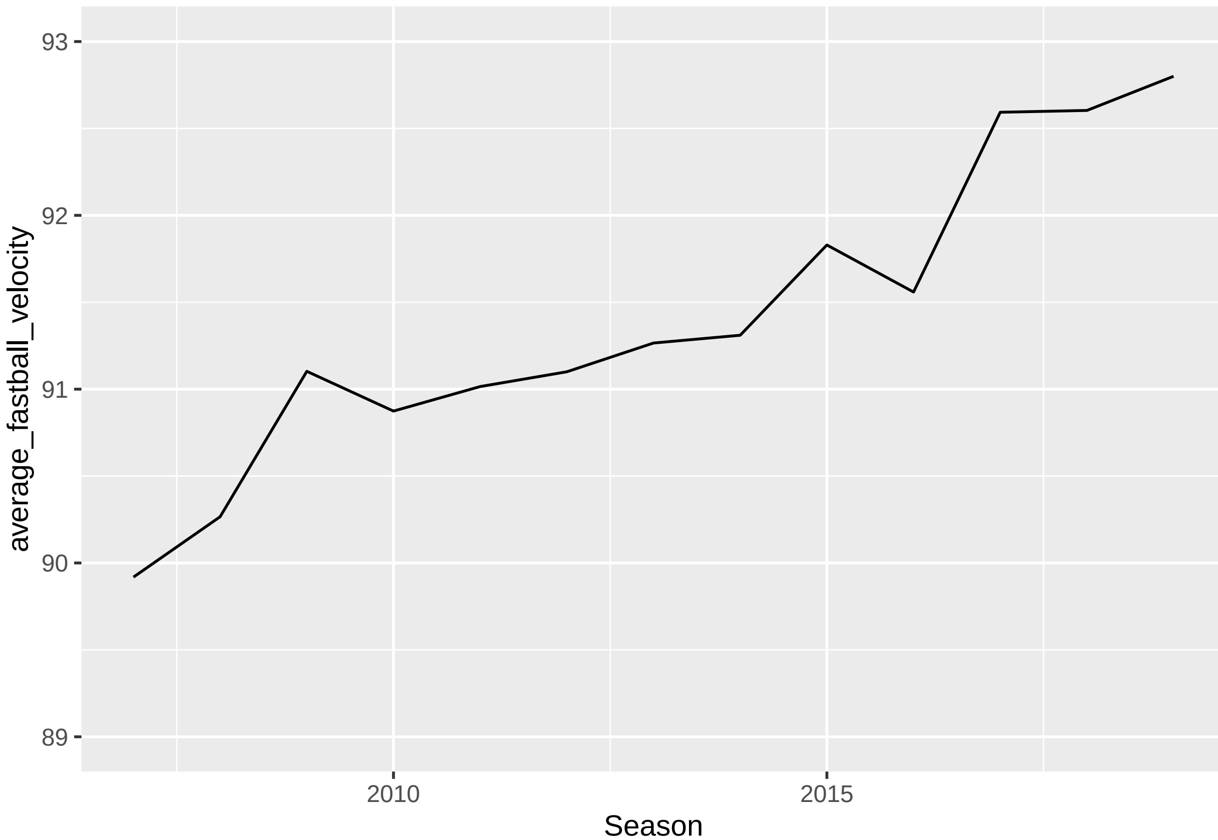

```
y <- fangraphs_pitching$so_per_nine
x <- fangraphs_pitching$FBv

#Prints the R-squared of the regression of strikeouts on fastball velocity
summary(lm(y ~ x))$r.squared
```

```
## [1] 0.2776654
```

There is certainly some relationship there. Now lets see if we can show that fastball velocity has been increasing year over year.

```
fangraphs_pitching %>%
  group_by(Season) %>%
  summarise(average_fastball_velocity = mean(FBv)) %>%
  ggplot(aes(x = Season, y = average_fastball_velocity)) +
  geom_line() +
  ylim(89, 93)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

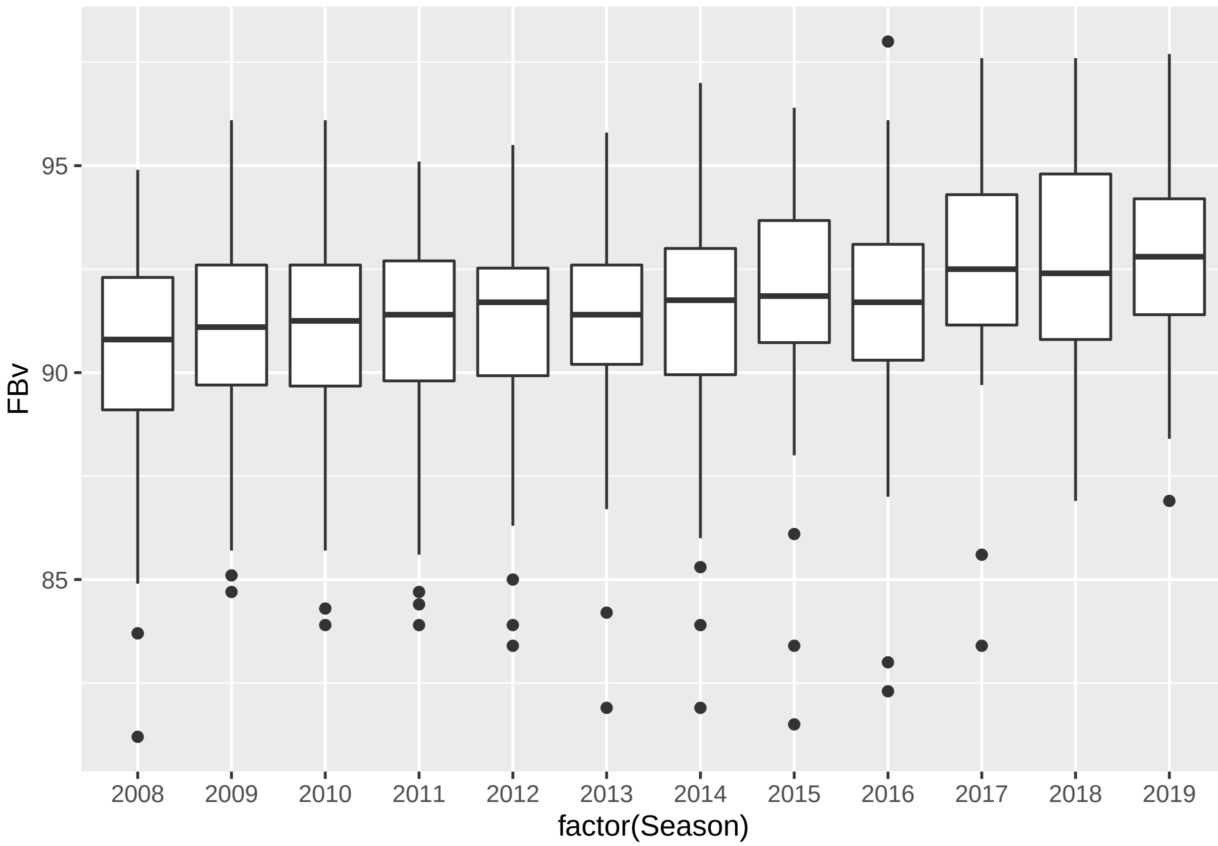



```
fangraphs_pitching %>%
  filter(Season > 2007) %>%
  filter(FBv > 80) %>%
  ggplot(aes(x = FBv)) +
  geom_histogram(aes(color = Season), fill = "white", binwidth = 1, show.legend = FALSE) +
  geom_vline(data = fangraphs_pitching %>%
                   filter(Season > 2007) %>%
                   filter(FBv > 80) %>%
                   group_by(Season) %>%
                   summarise(avg = mean(FBv)),
             aes(xintercept = avg), color = "red") +
  facet_wrap(~ Season)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Think this could be better shown as a box/violin plot.

```
fangraphs_pitching %>%
  filter(Season > 2007) %>%
  filter(FBv > 80) %>%
  ggplot(aes(x = factor(Season))) +
  geom_boxplot(aes(y = FBv))
```



Better. So we see that pitchers who throw harder, all else equal, get more strikeouts. And we also know that pitchers have thrown harder fastablls on average ofer the past 12 years. This seems to be a good explanation as to potentially why more batters are striking out.