Exploring Data and Statistics

Brian Hallberg

6/11/2018

Initialize everything

load any needed libraries

Open Data Files

```
eligible = readRDS("eligible.rds")
```

Explore data

Add fields

Added the following fields

- pas - Number of possible all star years for each player based on start of all star games
- pcas - Percent of possible all star games the player appeared in (note: can be >100% since 4 years had 2 all-star games)
- lba - Lifetime Batting Average

Converted fields to integers

- pas

```
eligible <- eligible %>%
  mutate(pas = case_when (
    debut > '1933-01-01'       ~ timein,
    final_game <= '1933-01-01' ~ 0,
    TRUE                       ~ timein - floor(difftime('1933-01-01', debut, units="days")/365)
  )
  ) %>%
  mutate(pas = as.integer(pas)) %>%
  mutate(pcas = allstar/pas)
```

Summarize some of what we know now

As we proceed a little bit of summary data is provided. The set of data for players covers the years 1871-2015 and includes **18,846** names.

There are **1741** players that were selected to at least 1 All-star games.

There are **247** players in the Baseball Hall of Fame in Cooperstown New York. So **14.2%** of the All-stars and just **1.3%** of all players that have played baseball in over 100 years have become Hall of Famers.

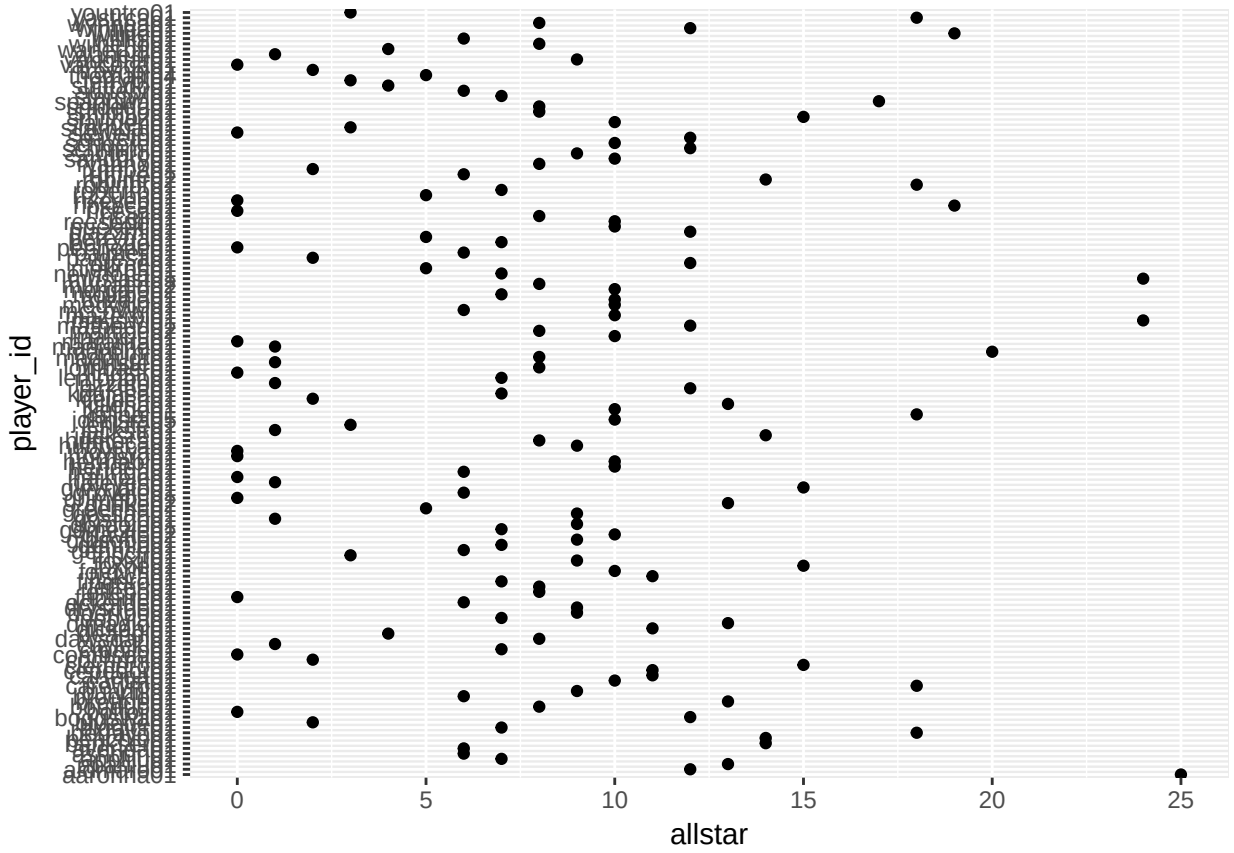
Looking at eligible players of the **18,846** there are **2385** eligible for the hall of fame which means that only **12.7%** of players even have a chance to be inducted.

Including Plots

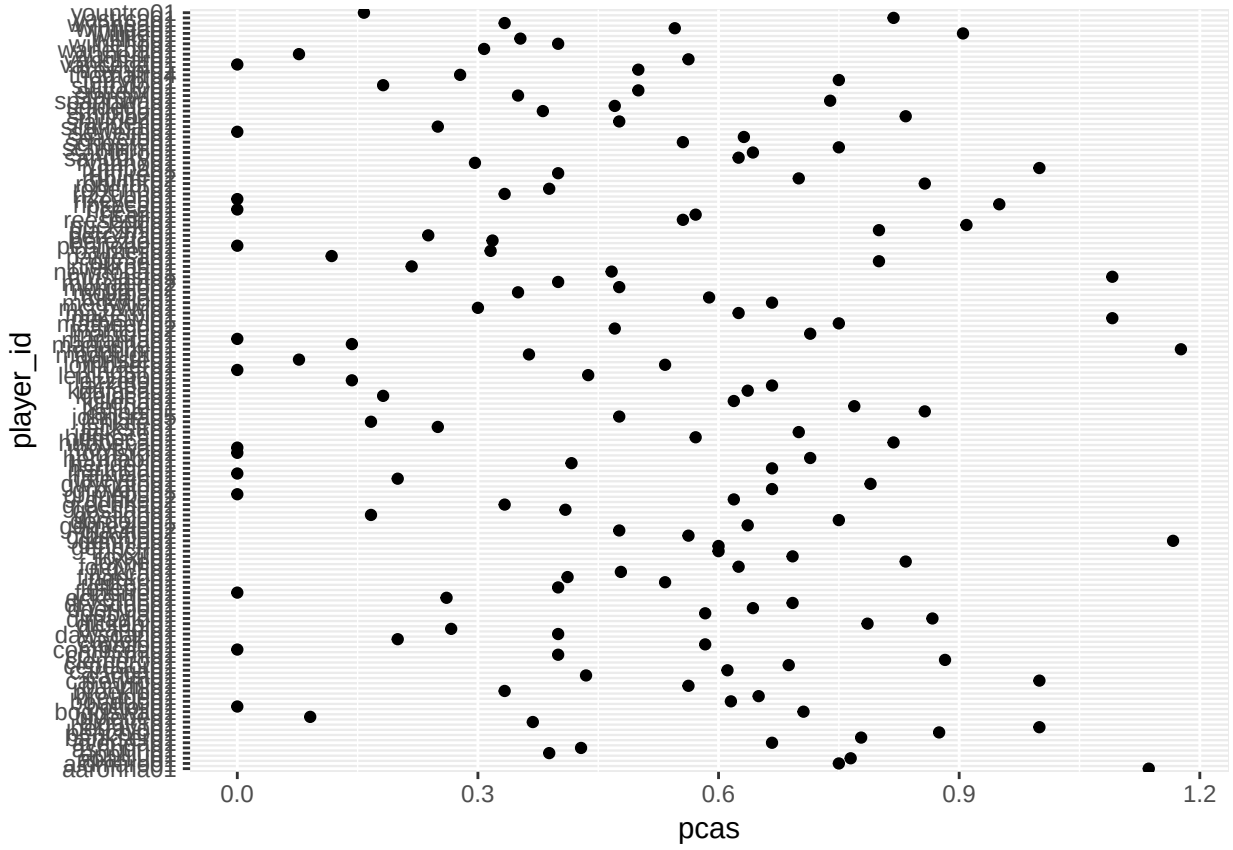
Scatter plot of the hall of fame players and the number of all star games they participated in

The first thought was that the number of times a player appears in the all-star game may be a good indicator being elected to the hall of fame. Make some plots to see if we can find some sort of threshold that is important. First step is to create a file with just hall of fame players and their appearances in the all-star game. The first all-star game (at least the data for it) is from 1933 so anyone elected before what will not appear in a game, so we dropped those players.

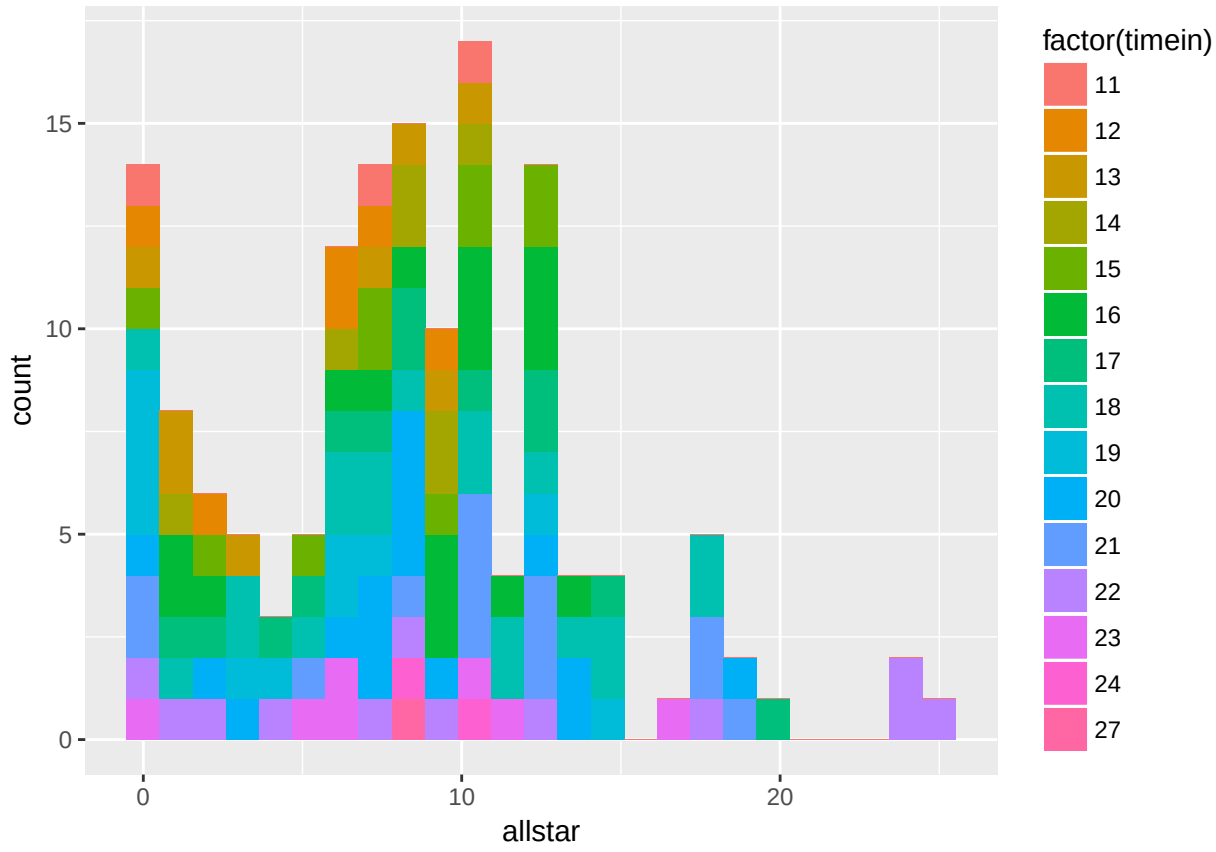
```
hofa <- eligible %>%
  filter(inducted == 'Y') %>%
  filter(final_game > '1933-01-01') %>%
  select(player_id, inducted, allstar, timein, final_game, pcas, pas)

ggplot(hofa, aes(x=allstar, y=player_id)) +
  geom_point()
```




```
ggplot(hofa, aes(x=pcas, y=player_id)) +
  geom_point()
```

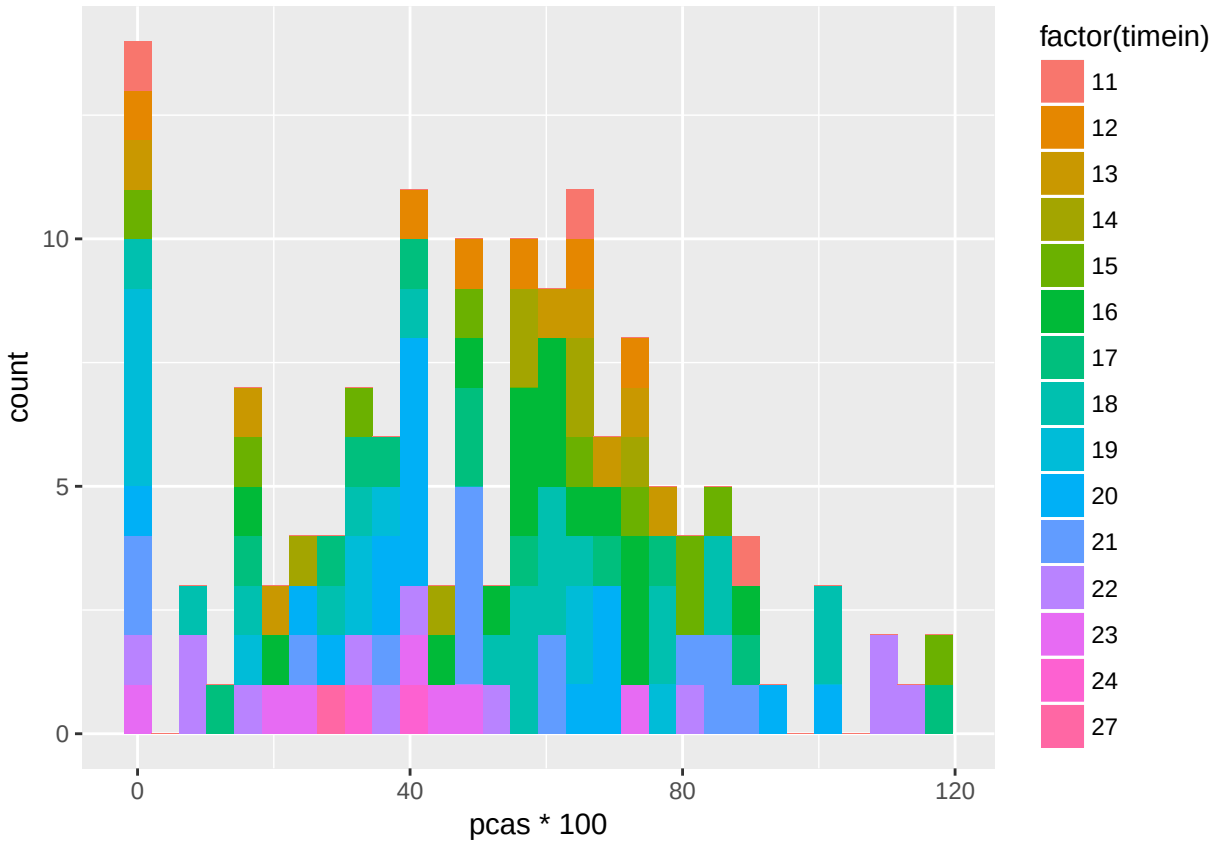



```
ggplot(hofa, aes(x=allstar, fill=factor(timein))) +
  geom_histogram(bins=25)
```




```
ggplot(hofa, aes(x=pcas*100, fill=factor(timein))) +
  geom_histogram()
```

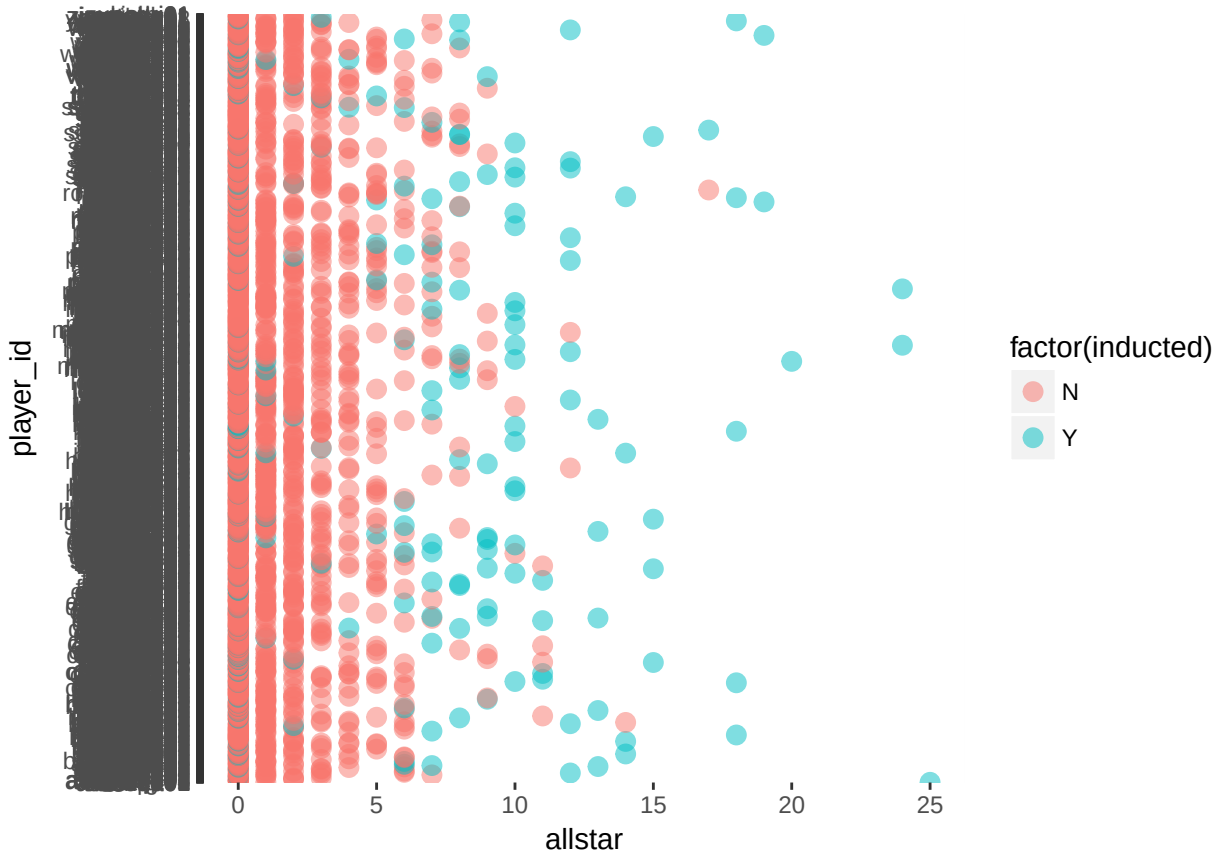
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Scatter Plot of all players eligible

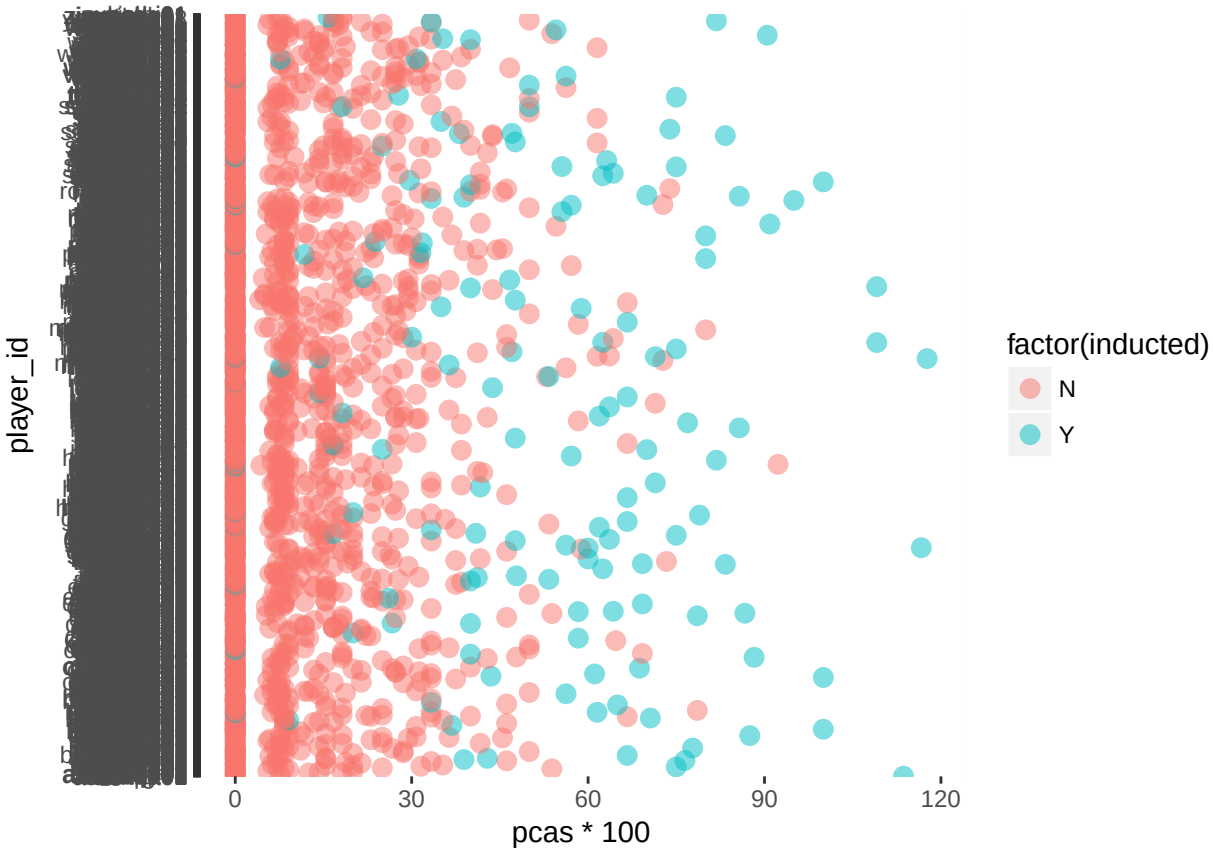
This two plots include all eligible players. First plot is just based on the raw number of times a player was in the all star game. Plot 2 is the percent of all start appearances in during their baseball career.

```
ggplot(eligible, aes(x=allstar, y=player_id, color=factor(inducted))) +
  geom_point(size = 3, alpha = 0.5)
```




```
ggplot(eligible, aes(x=pcas*100, y=player_id, color=factor(inducted))) +
  geom_point(size = 3, alpha = 0.5)
```

```
## Warning: Removed 515 rows containing missing values (geom_point).
```



A quick look at the data indicates that at about 10 appearances, the player seems to be more likely to make the Hall of Fame. It is not the only indicator as many with fewer than 10 made it as well. Of the 10 individuals that were not inducted they break down as follows:

- Two are still eligible, but not yet elected
- Seven were not elected in the 10 years they were eligible to be elected, now in the hands of the ERA Committee
- One is Pete Rose and he is banded from the Hall of Fame by Baseball

Now we should look at some data other than all-star

First thing we do is get summary data for the current hall of fame players

- Mean data in all numeric fields
- Min data in all numberic fields

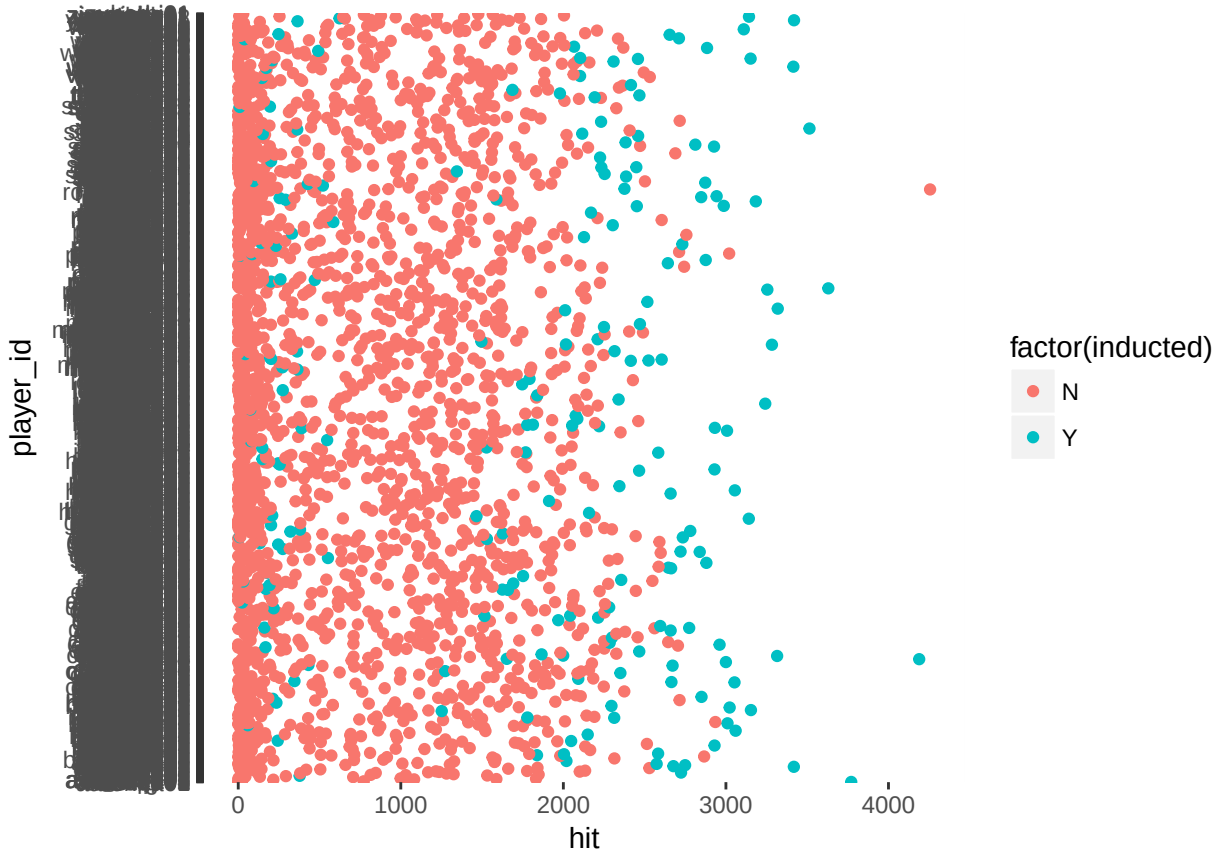
Good Possible options are: * HR > 500 * Hits > 2000 * Lifetime Batting Average > .300 * RBI > xxx * Runs scored > xxx * Wins > xxx * Pitched Strikeouts > xxx * Saves > xxx * Games Pitched > xxx

```
hof <- eligible %>%
  filter(inducted == 'Y')
meanhithof <- hof %>%
  filter(is.na(gpitch) == FALSE | gpitch < 100) %>%
  summarise_if(is.numeric, mean, na.rm = TRUE) %>%
  mutate(lba = hit/ab) %>%
  select(-starts_with("birth_"), -starts_with("death_"), -weight, -height, -ba)
minhithof <- hof %>%
```

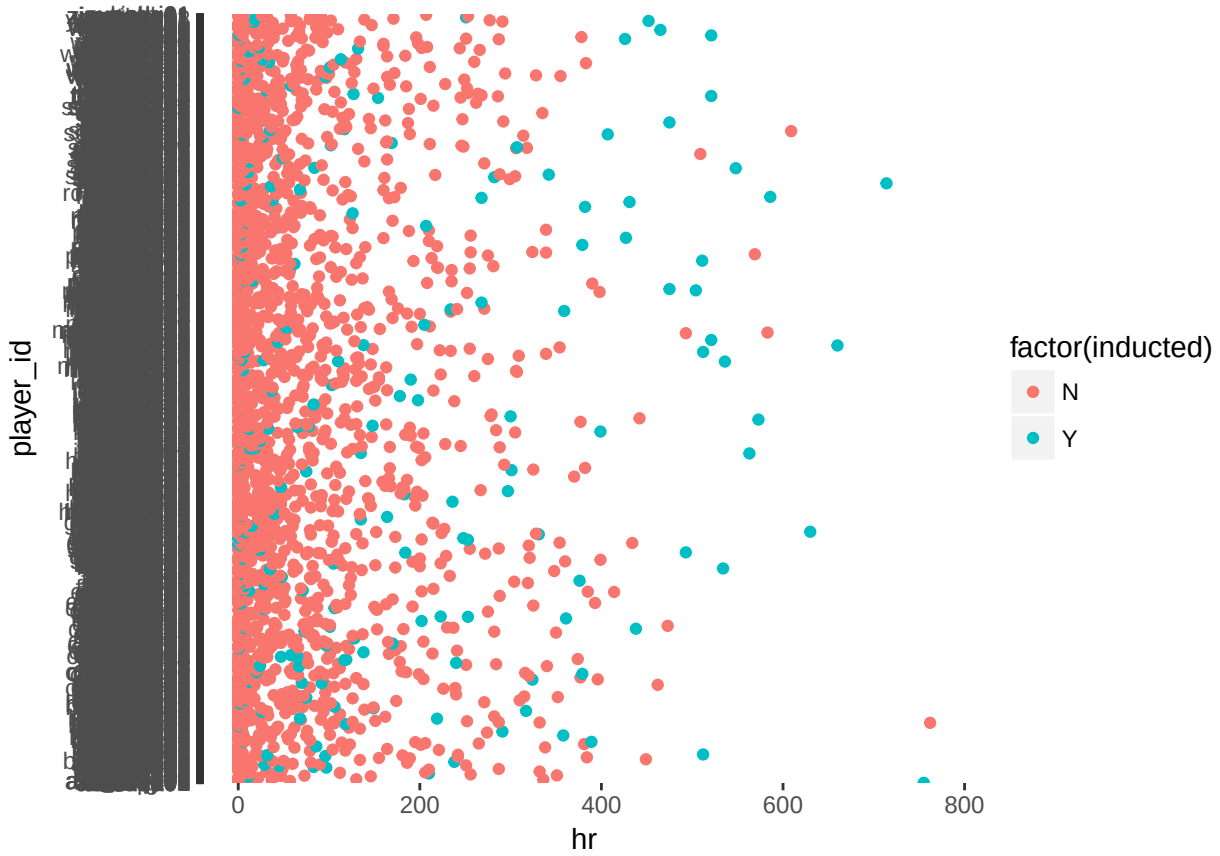
```
  filter(is.na(gpitch) == FALSE | gpitch < 100) %>%
  summarise_if(is.numeric, min, na.rm = TRUE) %>%
  mutate(lba = hit/ab) %>%
  select(-starts_with("birth_"), -starts_with("death_"), -weight, -height, -ba)
```

Do some plots of hall of fame player data

```
p <- ggplot(eligible, aes(y=player_id, color=factor(inducted)), size=1, alpha=0.5)
p + geom_point(aes(x=hit))
```

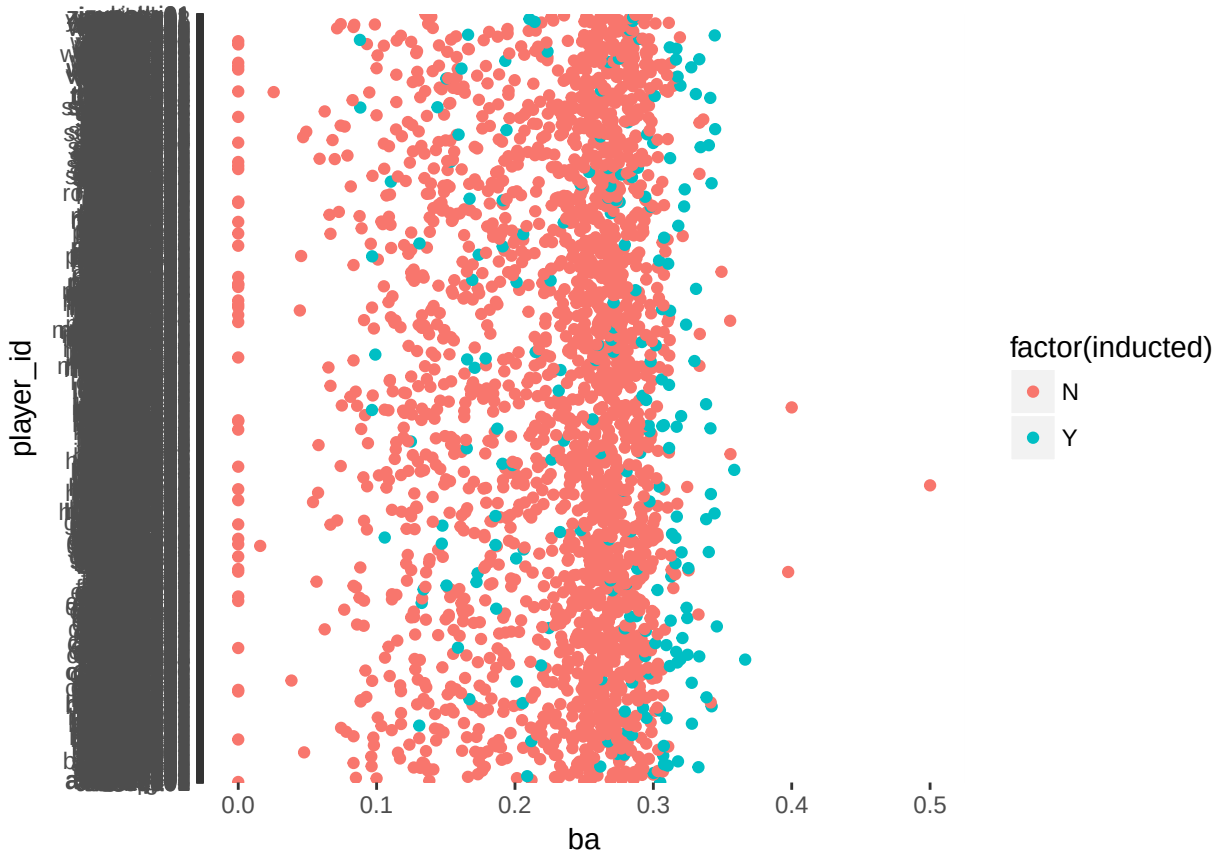



```
p + geom_point(aes(x=hr))
```

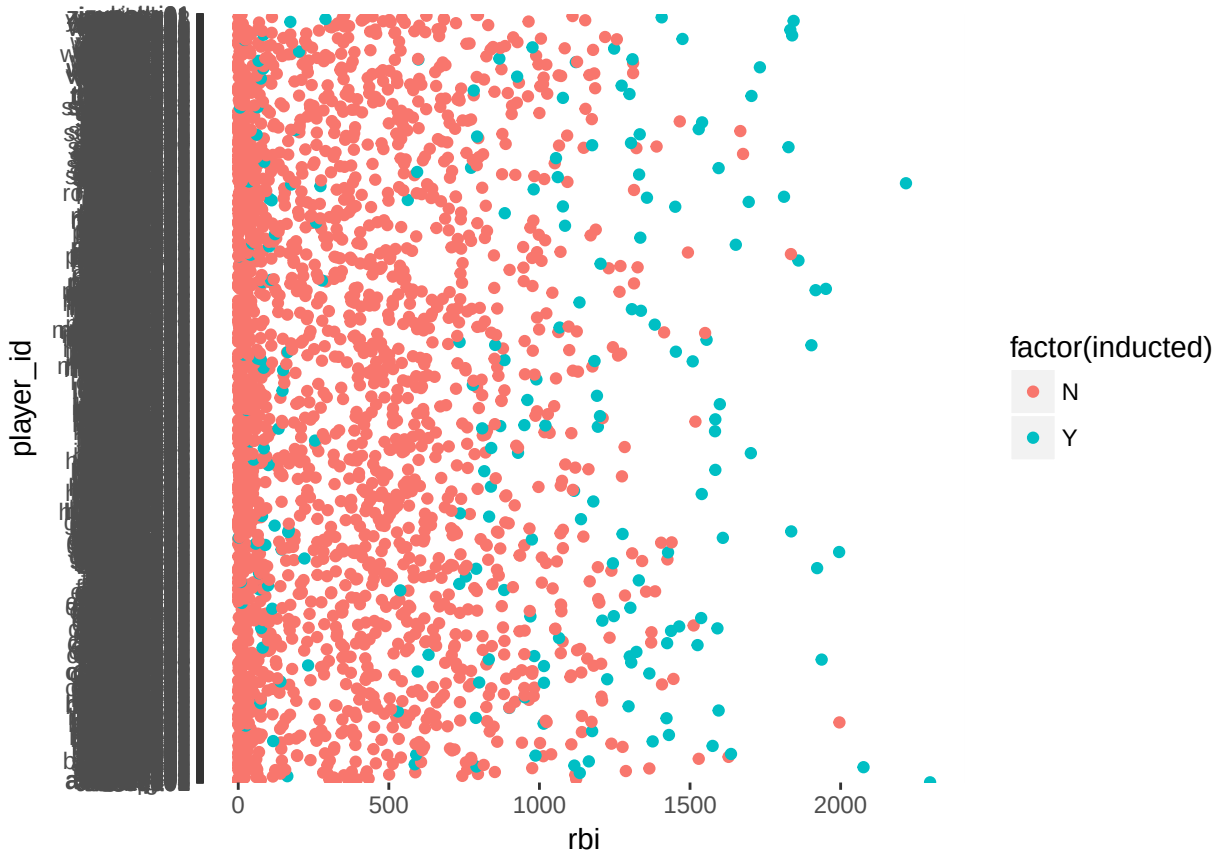



```
p + geom_point(aes(x=ba))
```

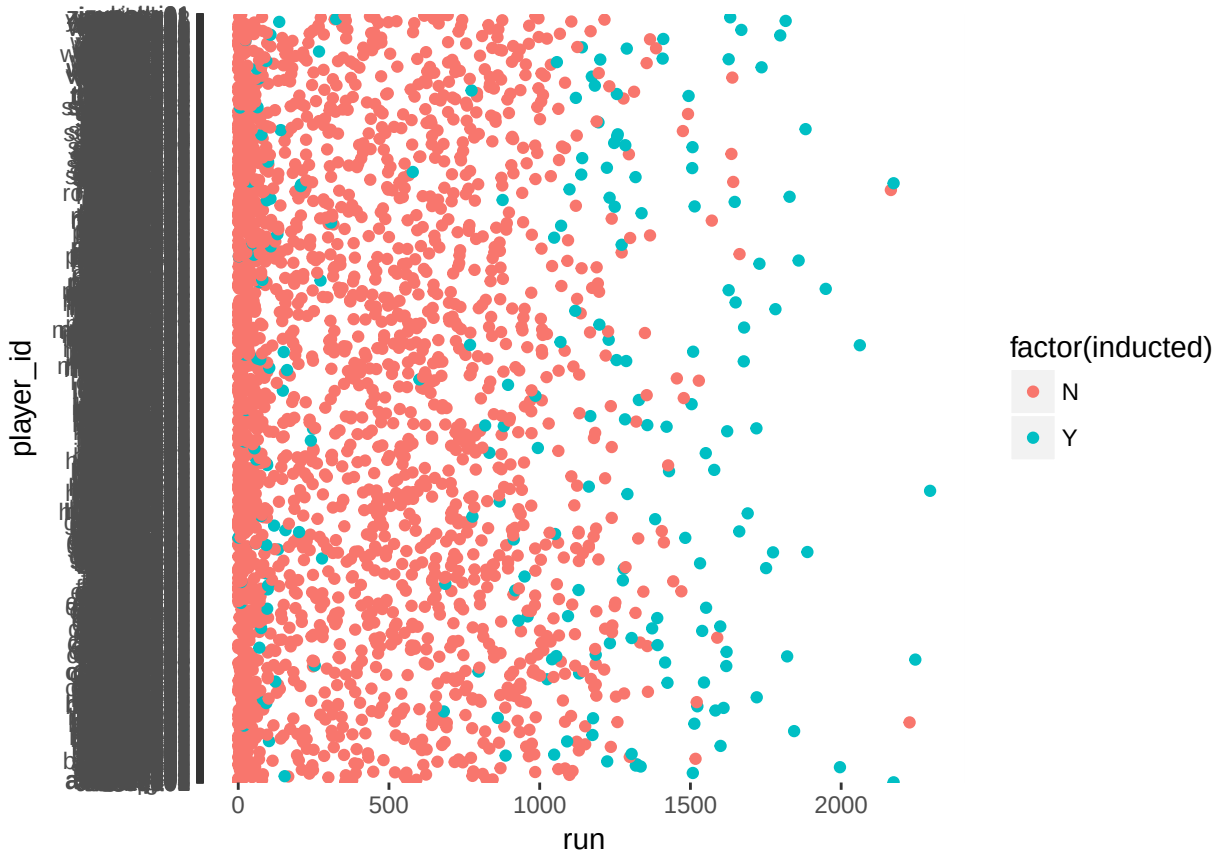
```
## Warning: Removed 23 rows containing missing values (geom_point).
```




```
p + geom_point(aes(x=rbi))
```

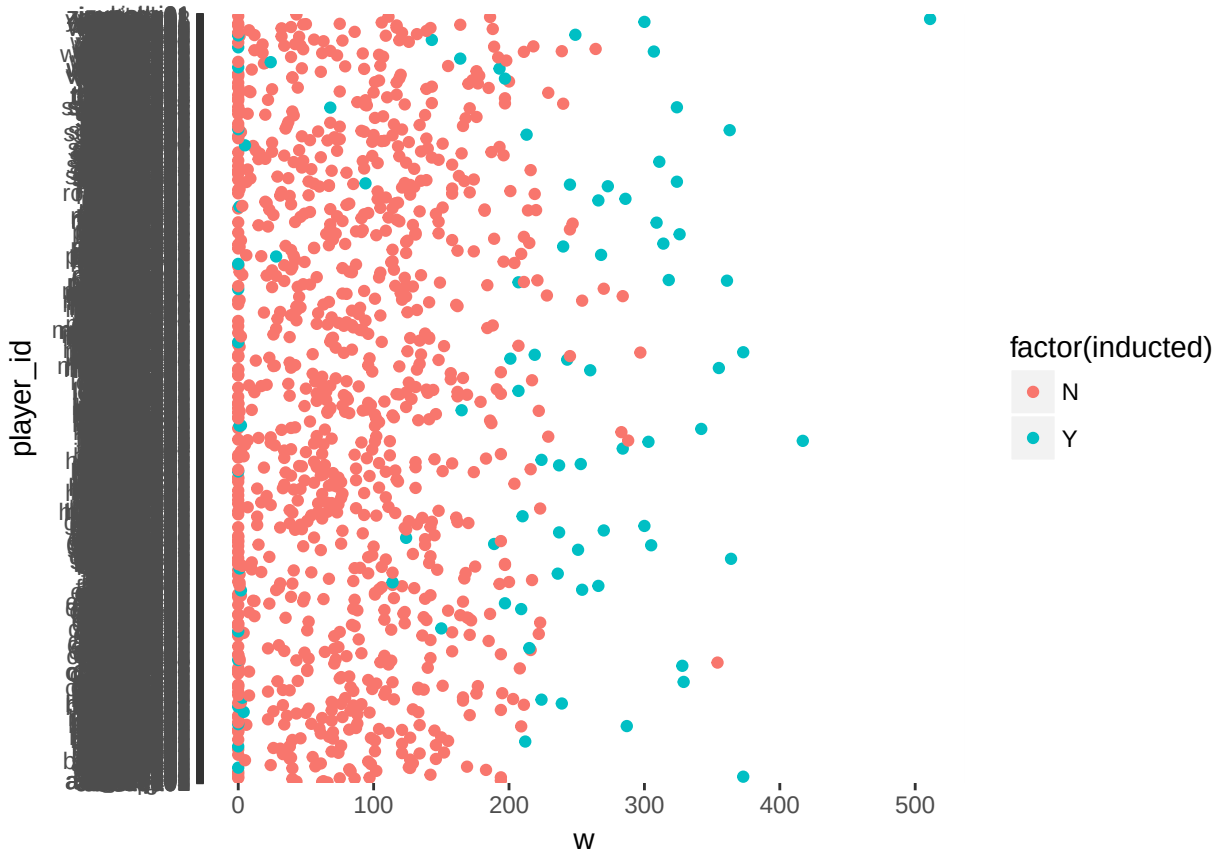



```
p + geom_point(aes(x=run))
```



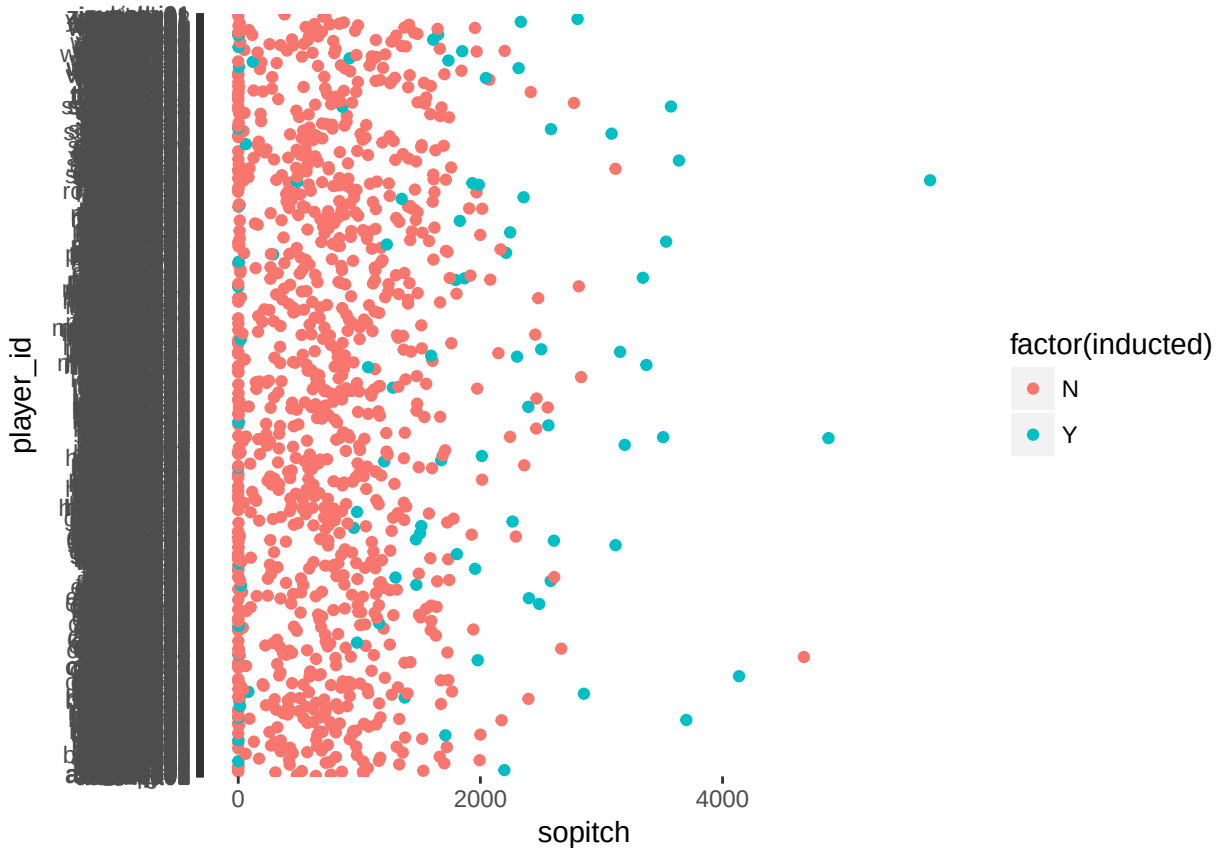

```
p + geom_point(aes(x=w))
```

```
## Warning: Removed 1318 rows containing missing values (geom_point).
```



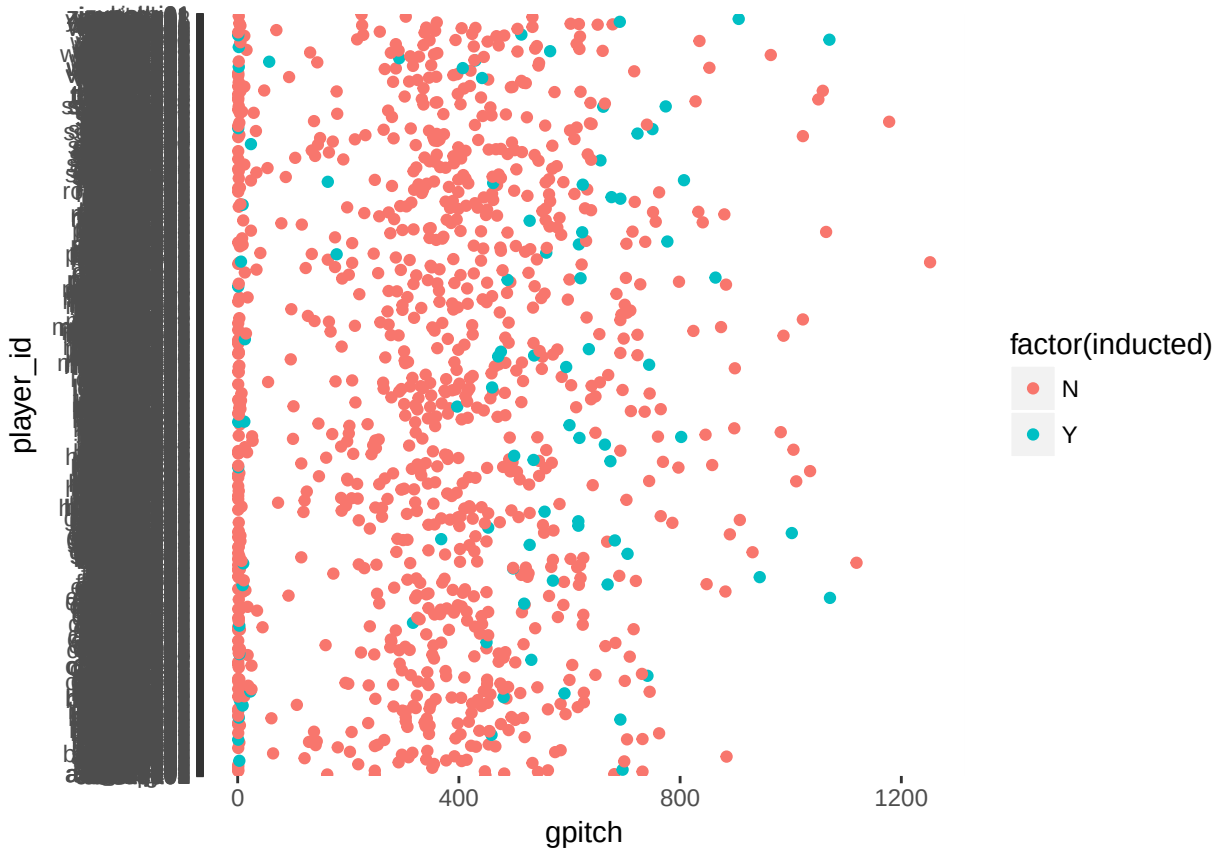

```
p + geom_point(aes(x=sopitch))
```

```
## Warning: Removed 1318 rows containing missing values (geom_point).
```




```
p + geom_point(aes(x=gpitch))
```

```
## Warning: Removed 1318 rows containing missing values (geom_point).
```



Next is do Hall of Fame players have data above the mean on a number of statistics?