Milestone Report

Brian Hallberg

June 14, 2018

Project Introduction

This project will attempt to create a predictive model for identifying players that will become memebrs of the Baseball Hall of Fame using career statistics including All-Star Game appearances as well as possibly hits, home runs, runs, career batting average, runs batted in for non pitchers and wins, strike outs, saves, and games pitched for pitchers.

Data

The project makes use of a data set called the History of Baseball that contains data from professional baseball covering the period of 1871 to 2015. The data set includes tables for general player information, All-Star Game appearences, Hall of Fame selections, as well as pitching and hitting yearly statistics.

There are additional tables that are avaialble as part of the data set, but not used in this project. The data is available on the Kraggle web side and can be obtained at this here:

- https://www.kaggle.com/seanlahman/the-history-of-baseball

The hall of fame players are identified using a inducted field in the hall_of_fame table. The project will attempt to use player data and all-start appearances to predict which players were inducted. Significant tables and fields containing the data this project will use are described below.

Player Table

The player table contains unique code assigned to each player (player_id), as well as the dates of their debut and final game. Also used is the first and last name of the player.

All_star Table

The all_star table contains fields for player_id, year, game for those selected for the game.

Batting Table

The batting table contains fields for player_id, at bats, runs, hits, home runs, and runs batted in all of which are evaluated.

Hall_of_fame Table

The hall_of_fame table contains fields for player_id and if they were inducted into the hall that year. Other fields are not used.

Pitching Table

The pitching table contains fields for player_id, games pitched, wins, and strikeouts which are evaluated.

Limitations

Although there are fields in tables for batting average and era, those cannot be averaged over multiple seasons. There is sufficient data to calculate that for players if needed. There have been exceptions made to the rules for eligibility for the hall of fame and some of the rules have changed over the years, this process ignores exceptions and uses the current rules.

Data Cleaning

The raw data was contained in Comma Separated Values (CSV) files, one for each table. Each file was loaded into a separate data frame using read_csv(). Relevant fields were selected using select(). The example below shows the player and batting table files being read into data frames.

Read in the data

```
# read player table into data frame.
player <- read_csv("player.csv")

# read batting table into data frame.
batting <- read_csv("batting.csv")
```

Summarize data

THe batting and pitching tables are broken down beyond the player so that a row of the table represents the team and year since a player could have played for more that one team in a year. I only care about the totals for the players career so summaries of these two tables were created with these statements.

```
#  Summarize the batting data first
sumhit <- batting %>%
  group_by(player_id) %>%
  summarise_if(is.numeric, sum, na.rm = TRUE) %>%
  select(-year, -stint)
#  Now get the pitching data
sumpitch <- pitch %>%
  group_by(player_id) %>%
  summarise_if(is.numeric, sum, na.rm = TRUE) %>%
  select(-year, -stint, -baopp, -era)
```

The all_star table has a row that represents a player, year and which all-star game they were voted to play (yes, in 4 years they played 2 all-star games in the year). I only wanted the number of all-star games each played was voted to play in so another summary was done with this statement.

```
sumallstar <- all_star %>%
  select(player_id, league_id, year) %>%
  group_by(player_id) %>%
  count(player_id)
```

Finally the hall_of_fame table shows each year a player was voted on by the baseball writers including the number of votes they received. I only care if the player was inducted so the following statements were used to reduce all the data to a unique set of players inducted.

```
sumhof <- hof %>%
  filter(category =="Player") %>%
```

```
  select(player_id, inducted) %>%
  filter(inducted == "Y")
```

Build a final data set of all information by player

Since data was spread across different tables, the corresponding data frames needed to be joined in order to pull the data together into a single data frame. The statements below shows the joining of the four tables into one data frame used.

```
#  Add player and summarized Hall of Fame data
fullplay <- full_join(player, sumhof, by = "player_id")
#  Add the summarized All-Star data
fullplay <- full_join(fullplay, sumallstar, by = "player_id")
#  Add the summarized batting data
fullplay <- full_join(fullplay, sumhit, by = "player_id")
#  And finally add the summarized pitching data
fullplay <- full_join(fullplay, sumpitch, by = "player_id")
```

Add fields and update some missing data values

Some fields needed to be derived. For example, the batting average is calculated. There are also some missing data that is replaced with values to help with plotting. Examples are included below.

```
#  Calculate the batting average
fullplay <- fullplay %>% mutate(ba = h/ab)
#  Replace missing data in specific columns
fullplay <- fullplay %>% mutate_at(vars(inducted), funs(replace(., is.na(.), 'N')))
fullplay <- fullplay %>% mutate_at(vars(allstar), funs(replace(., is.na(.), 0)))
```

Cleanup some bad column names

The joins of the batting and pitching tables, created some conflicting column names and so they were updated to make some plots more readable and also to help simplify some future code. These are included below

```
fullplay <- rename(fullplay, allstar = n, rallow = r.y, hrallow = hr.y, bballow = bb.y, hallow = h.y)
fullplay <- rename(fullplay, gpitch = g.y, sopitch = so.y, batterhit = hbp.y)
fullplay <- rename(fullplay, game = g.x, run = r.x, hit = h.x, hr = hr.x, bb = bb.x, so = so.x, hbp = hb
```

Reduce to Hall of Fame Eligible Players

In this section, reduce the data set of players to those eligible for induction into the hall of fame. The rules say that the player must have played major league baseball for 10 years and they must be out of the game for 5 years. So we'll reduce the set of players to those that meet those requirements.

```
#  Add a column that calulates the years a player was in baseball
eligible <- fullplay %>%
  mutate(timein = floor(difftime(final_game, debut, units="days")/365)) %>%
  filter(timein > 10) %>%
  filter(final_game <= "2011-01-01")
```

```
##
## 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
```

Some fields were added to determine how many possible all star games a player could have appeared in based on when he played. Using that and the number of all star games they did play we can get a percent of years in the league that a player was an all-star. This seemed like it might be a good data value to have.

```
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)
```

Initial Findings

In order to identify which data might be useful in predicting hall of fame inductions, some exploratory analysis was performed to compare the data between players that are and are not in the hall of fame. The results of this analysis which indicate a potentially useful data field are shown below.

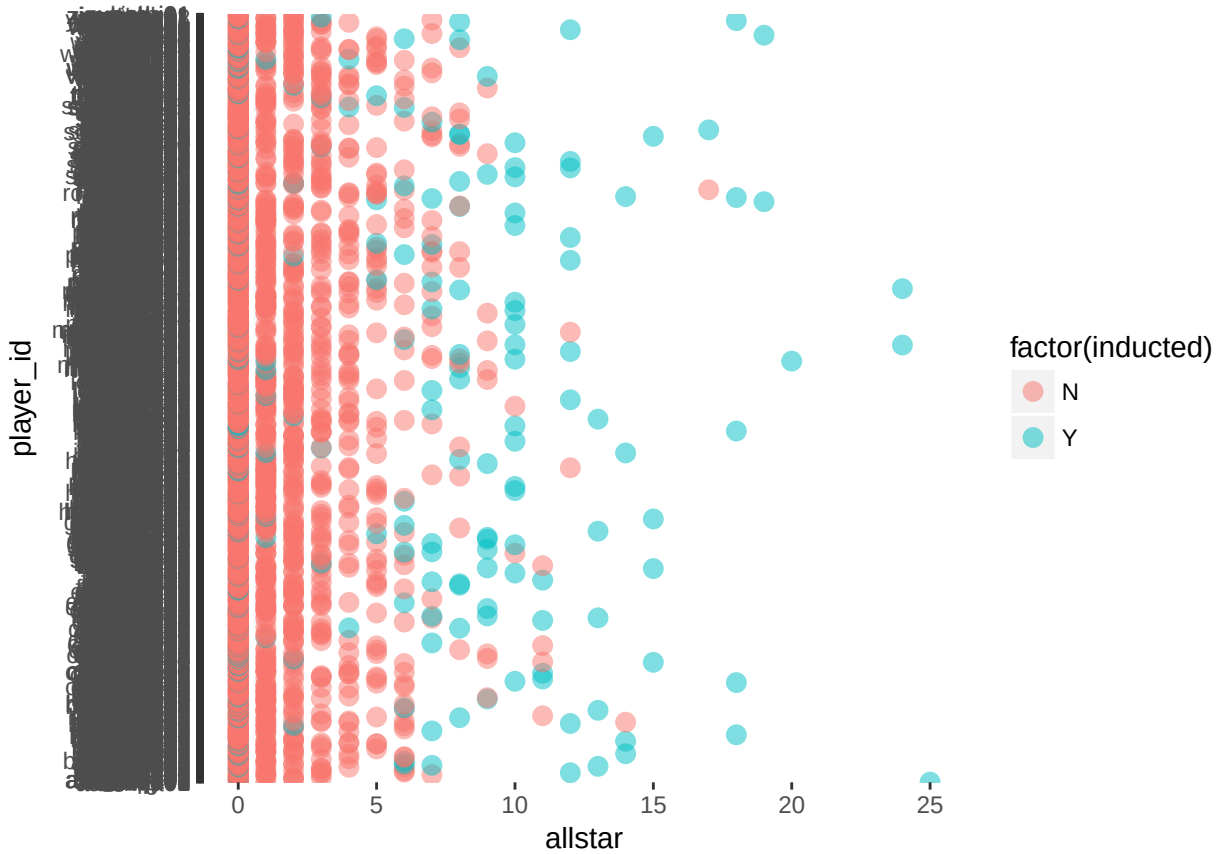
All-Star Game Appearances

So one idea is that the number of game apperances would be a good indicator of a future hall of famer. This scatter plot shows players and whether they were inducted and their all star appearances. The second one shows the players and the percentage of their career they played it the all-star game.

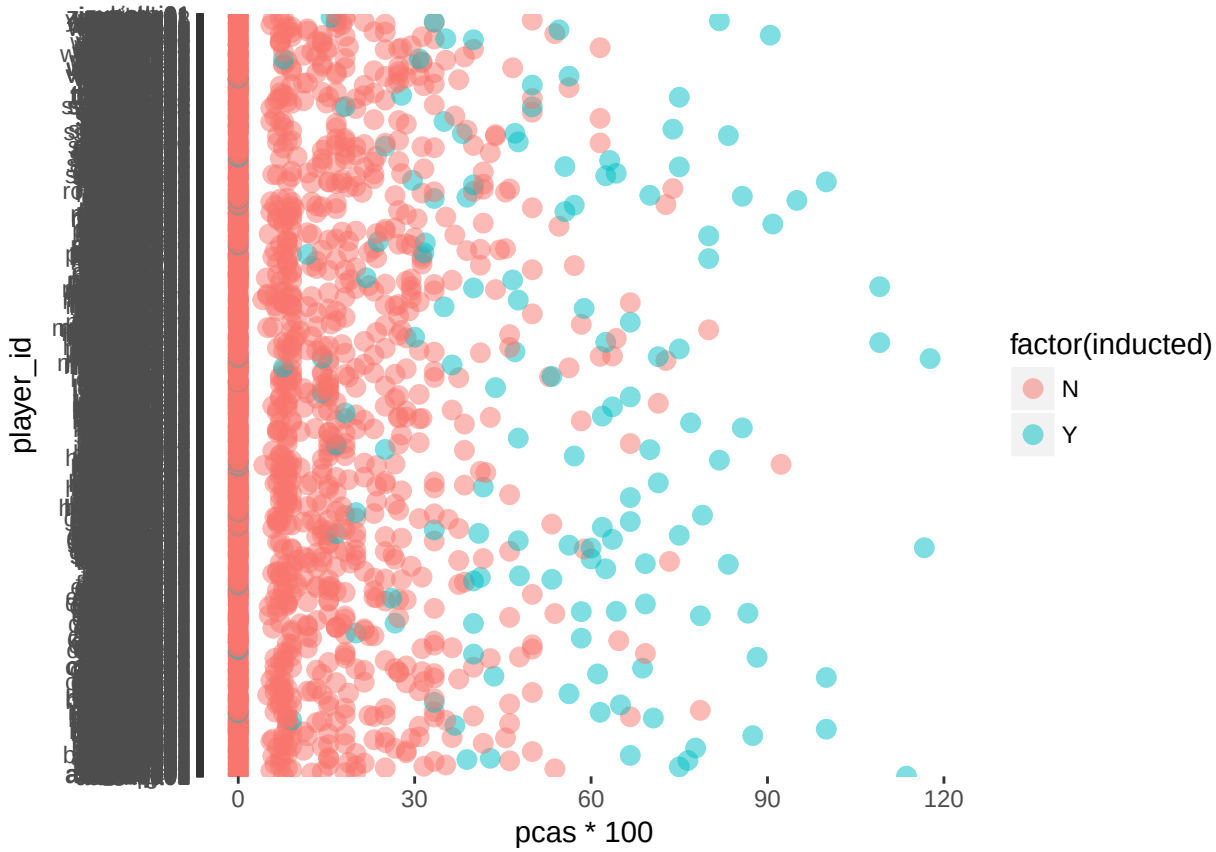
In a sense this is a crowd source approach since the all-star voting tends to be along the lines of popular players. In most cases, this also follows with the best players, but there are times where the population may like a player more, but there is more talent in the other player. The hope here is that over time, the best players make the team more than just a popular player.

A quick review of these two plots indicates that above 10 appearances or above 60% of your career making an all-star game may be a good indicator of hall of fame induction. More needs to be done.

```
library(ggplot2)
ggplot(eligible, aes(x=allstar, y=player_id, color=factor(inducted))) +
  geom_point(size = 3, alpha = 0.5, na.rm=TRUE)
```




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

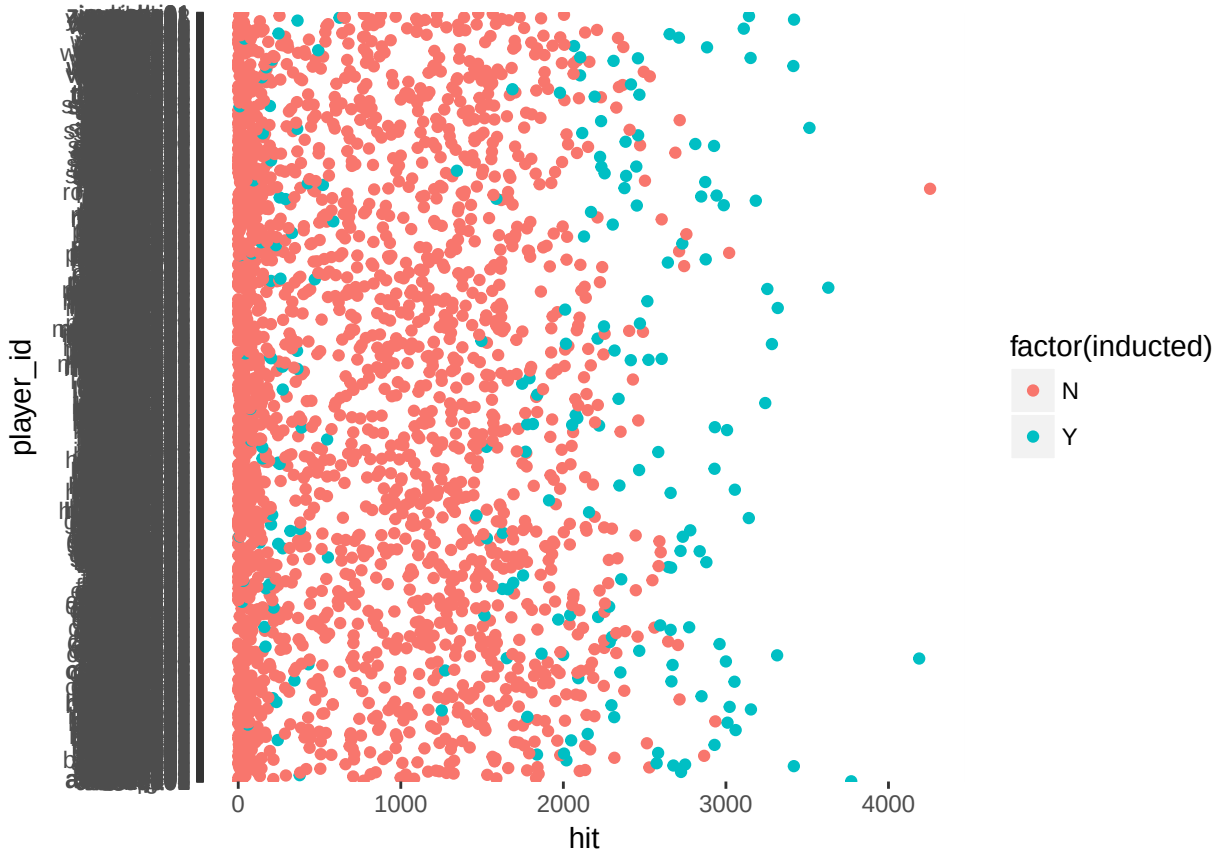

player_id
0
30
60
90
120
pcas * 100
factor(inducted)
N
Y

Data for Batters

Here we do some scatter plots of data for hitters. We remove the data for pitchers since they were not selected to the hall for their hitting abilities.

Hits

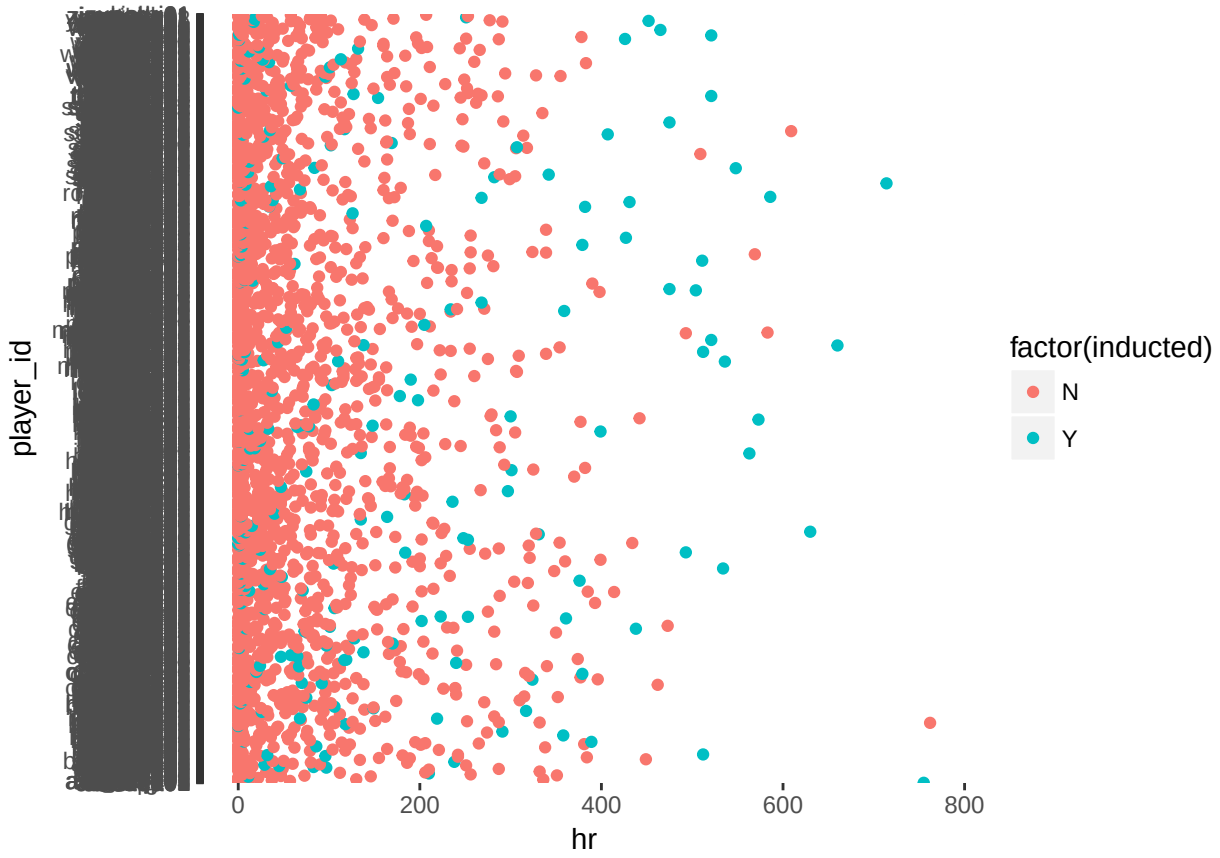
Scatter plot of Hits versus player color coded by induction status. Higher numbers of these do seem to indicated a better chance of induction.



Home Runs

Scatter plot of home runs versus player color coded by induction status. This is another where the data seems to indicate higher numbers of home runs may predict induction.

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

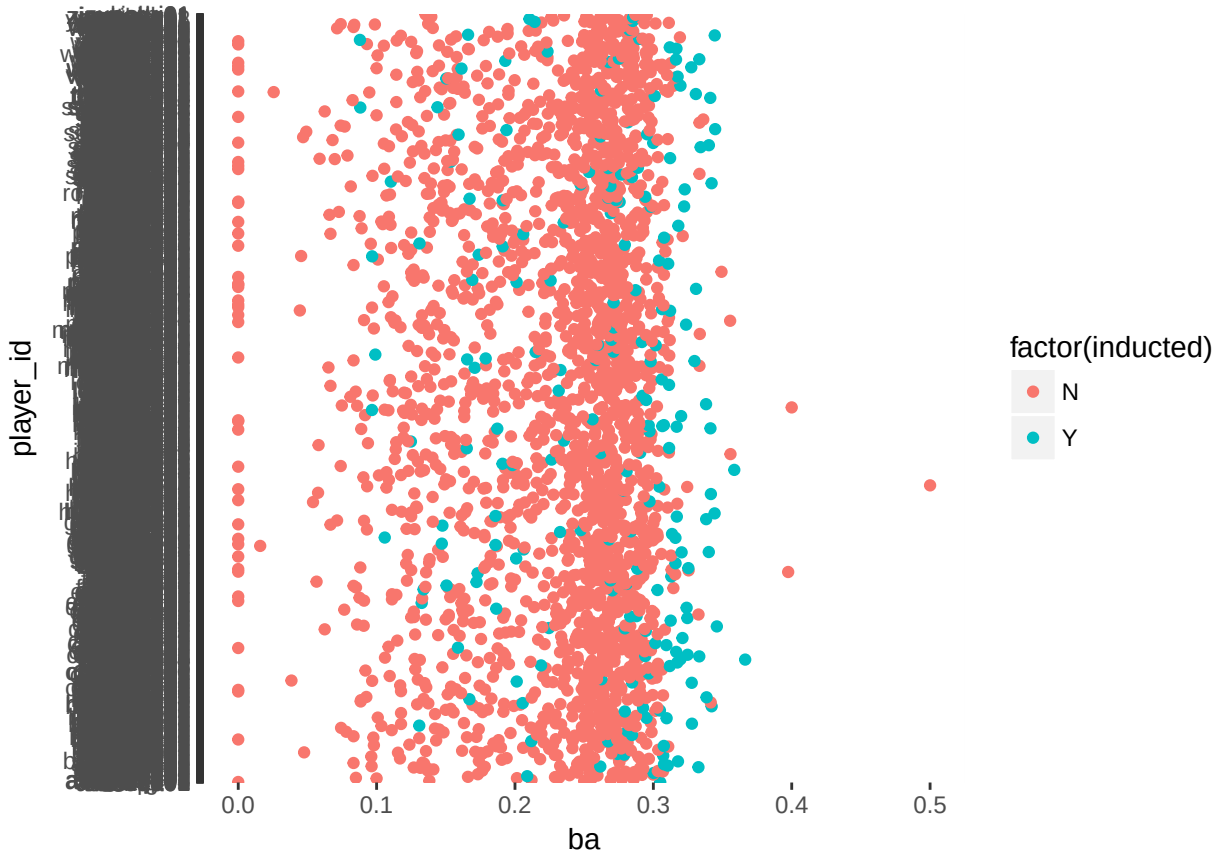


Batting Average

Scatter plot of batting average versus player color coded by induction status. This one needs more work. The majority of players are in the .200 to .300 and very few are above a .300. If that .200 and above are enhanced it may show something but this initial plot is not very telling.

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

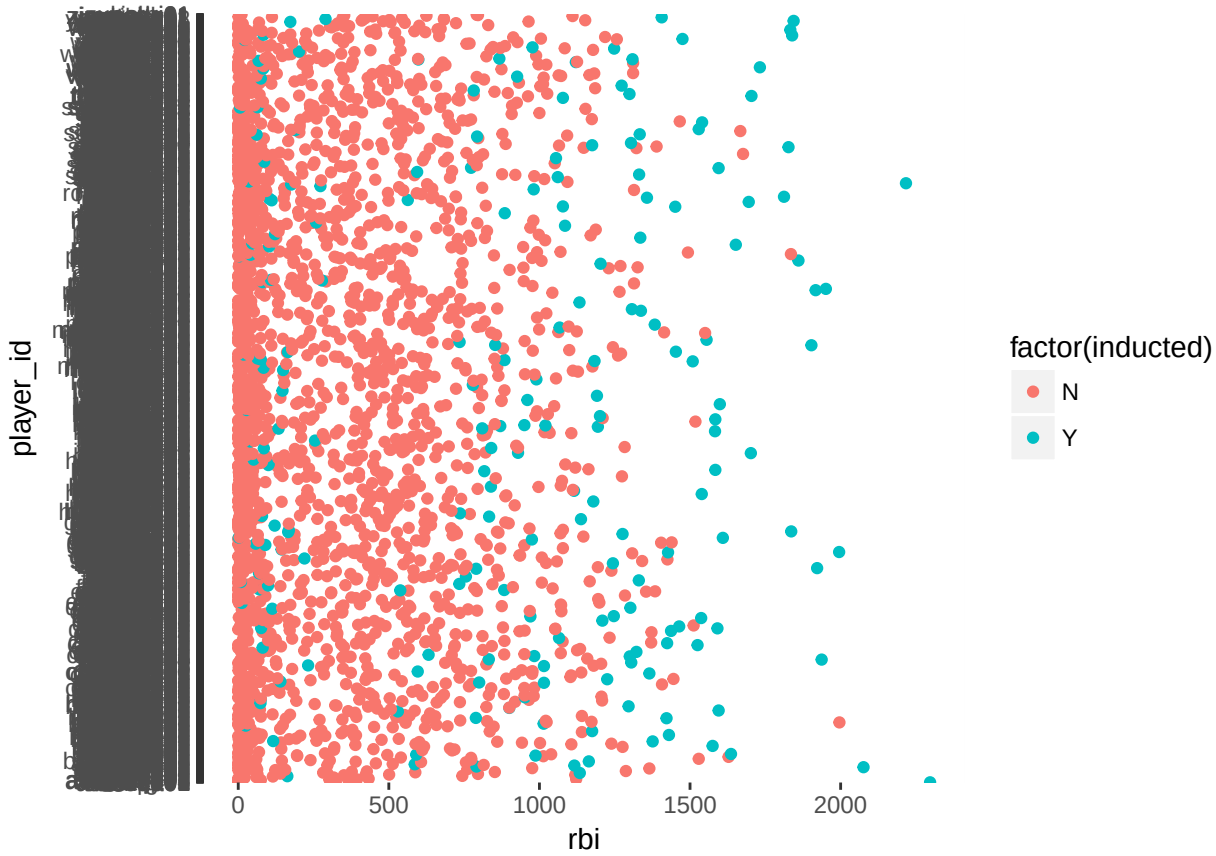
```
## Warning: Removed 23 rows containing missing values (geom_point).
```



Runs Batted In

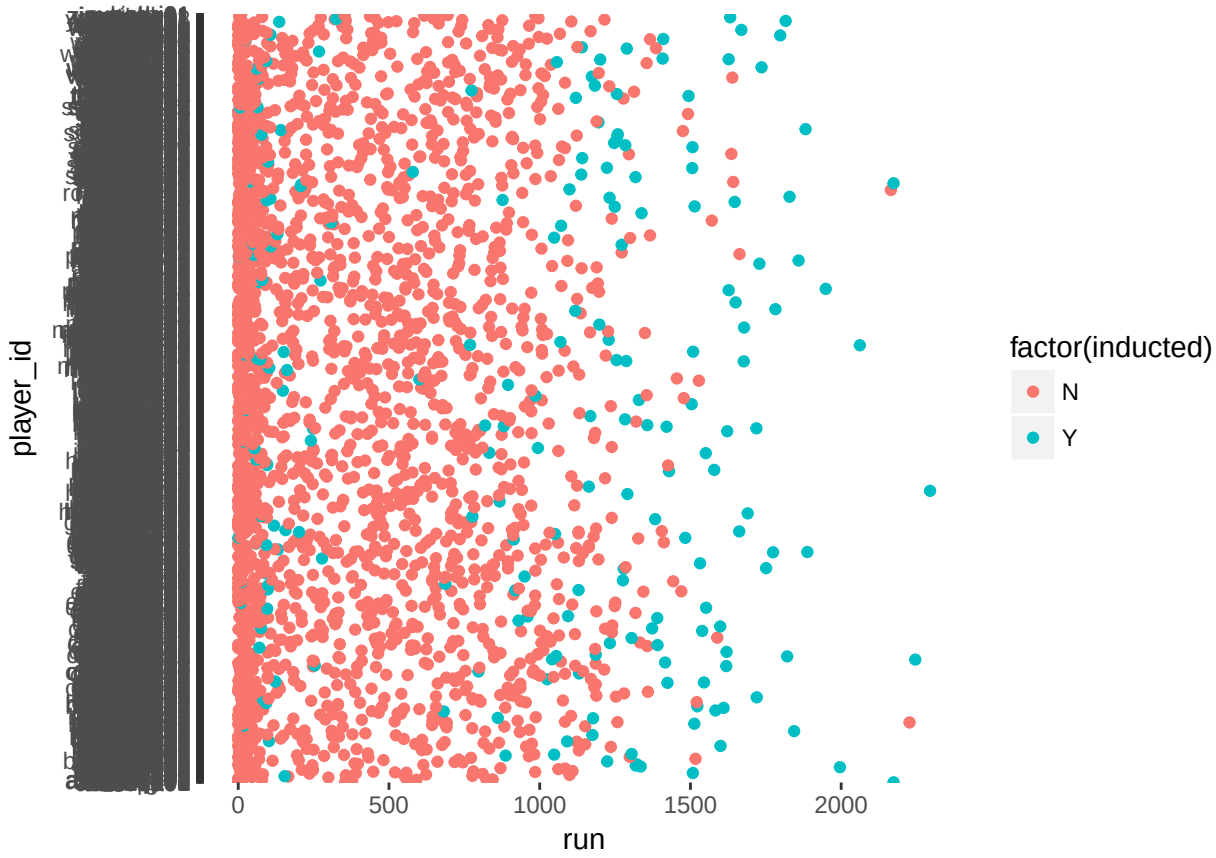
Scatter plot of runs batted in versus player color coded by induction status. Could be a good field to add to the model.

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



Runs Scored Scatter plot of runs versus player color coded by induction status. Another promising field to include in the final model.

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


player_id
0
500
1000
1500
2000
run
factor(inducted)
N
Y

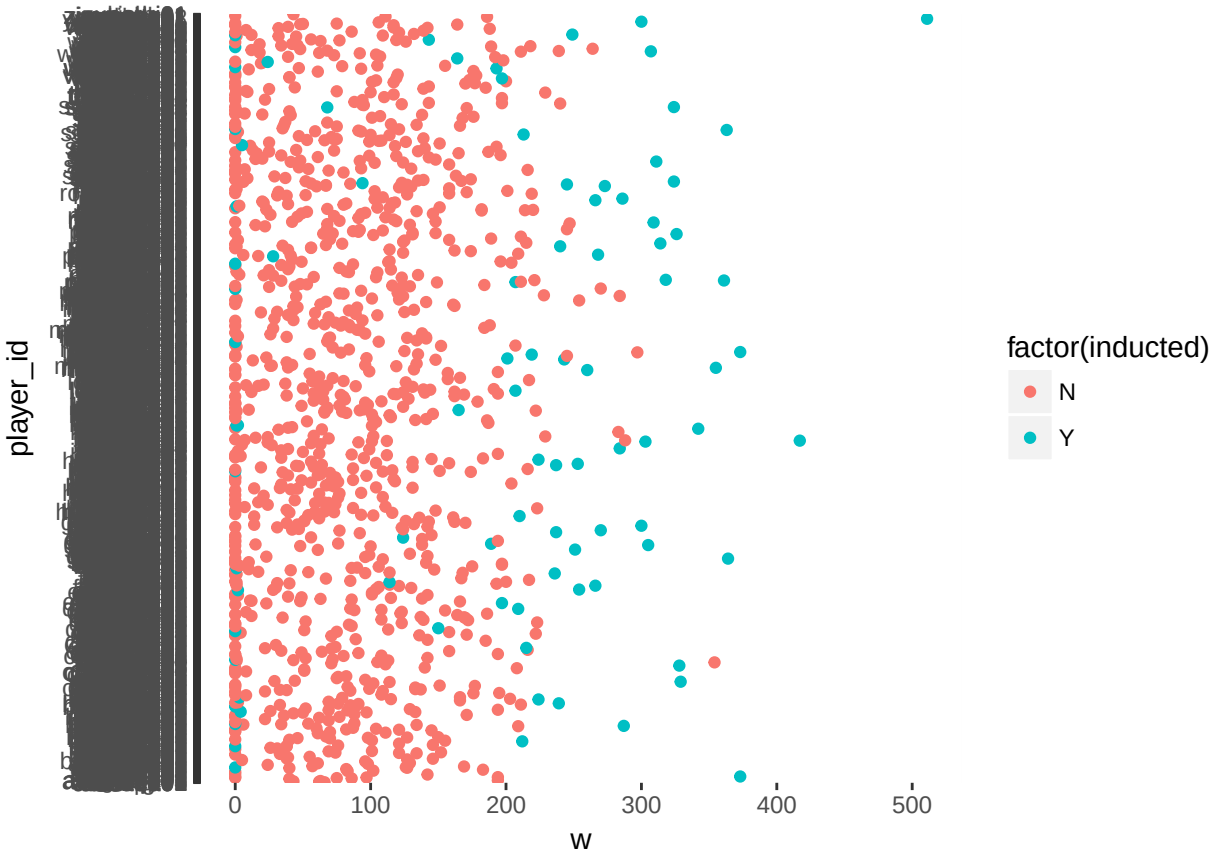
Data for Pitchers

Wins

Scatter plot of wins versus player color coded by induction status. Another promising field to include in the final model.

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

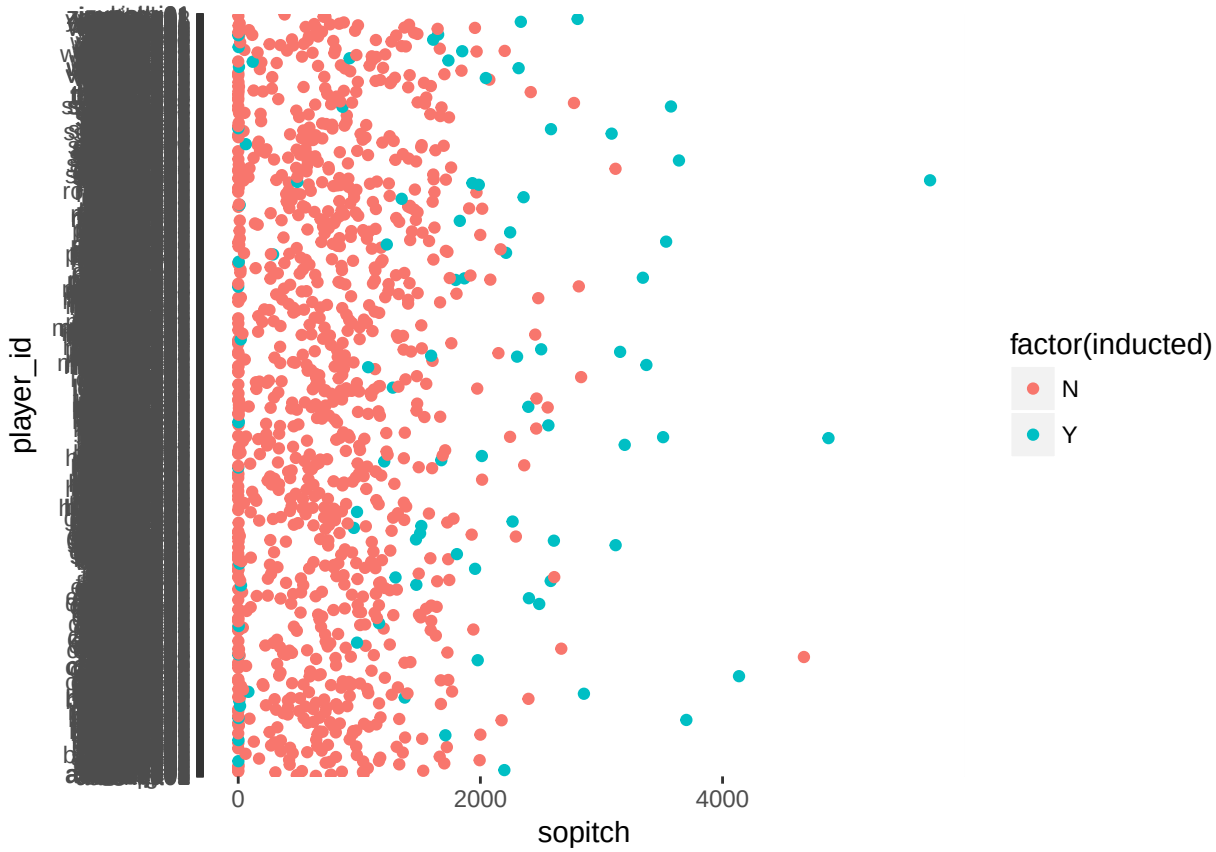
```
## Warning: Removed 1318 rows containing missing values (geom_point).
```



Strike Outs Scatter plot of strikeouts versus player color coded by induction status. Another promising field to include in the final model.

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

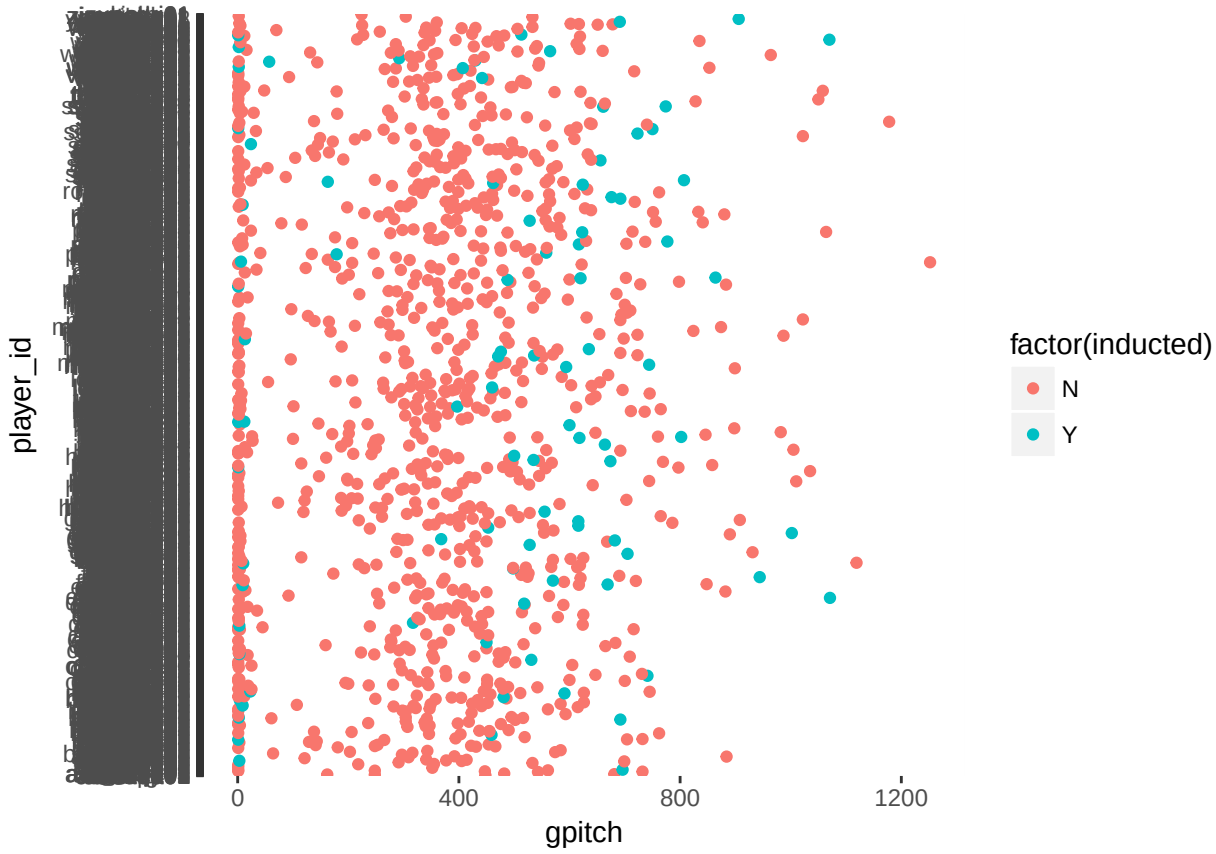
```
## Warning: Removed 1318 rows containing missing values (geom_point).
```



Game Appearances Scatter plot of games pitched versus player color coded by induction status. This one does not seem to show a pattern that would predict induction.

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

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


player_id
0
400
800
1200
gpitch
factor(inducted)
N
Y

Approach

My intenet is to use logistic regression and use results from that to help identify the best independent variables. I'll use the confusion matrix to see how well the model works. If this does not provide good results, I'll look at the random forest model to see if it works better.

I will be testing the model both using a train/test split in the original data.

The dependent variable will be inducted, and the independent variables will be some combination of All Star Game Appearances, Career Hits, Career Home Runs, Career Batting Average, Career Runs Batted In, Career Runs Scored, Career Wins, Career Games Pitched, and Career Strike Outs Thrown.

The final report will outline the process of how this model is developed, any further data cleaning required, and evaluation.