Homework 2

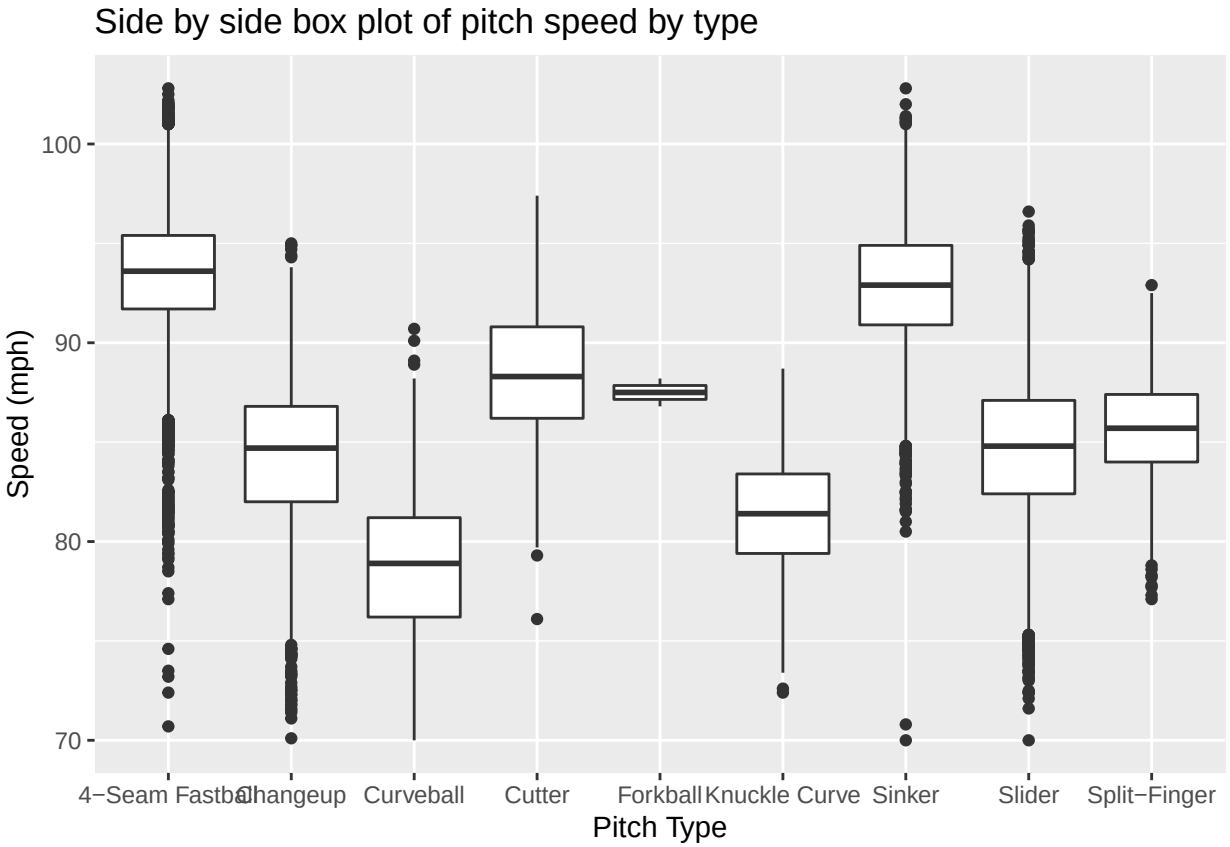
2022-09-08

The data set I would like to use is from Kaggle. It comes from a project where someone wanted to find the best predictors of a home run using a log loss model. We obviously haven't gotten that far yet into modeling, so I simplified my question a little bit which you will see in my research question. This is a fantastic dataset as it has information from every at bat in the 2020 Major League Baseball season. There are so many things that could be done with this data set. Some of the key variables that I will look at include pitch speed and pitch type. I do realize that this model I will be creating is a logistic regression model, but I feel that I have simplified it enough that it shouldn't be too much to handle.

Research question: Is there a relationship between pitch speed and home run rate? OR Is there a relationship between pitch type and home run rate? OR Is there a relationship between pitch speed and exit velocity off the bat? (I will use this last option if you don't want us to use logistic regression. Let me know which option you think is best for this stage of the class.)

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## Rows: 46244 Columns: 25
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (9): game_date, home_team, away_team, batter_team, batter_name, pitcher...
## dbl (16): bip_id, batter_id, pitcher_id, is_batter_lefty, is_pitcher_lefty, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## # A tibble: 9 x 2
##   pitch_name       name
##   <chr>           <dbl>
## 1 4-Seam Fastball  93.5
## 2 Changeup         84.4
## 3 Curveball        78.6
## 4 Cutter           88.5
## 5 Forkball         87.5
## 6 Knuckle Curve    81.3
## 7 Sinker           92.8
## 8 Slider           84.6
## 9 Split-Finger     85.7
```



Data set link: https://www.kaggle.com/datasets/jcraggy/baseball