baseballr Presentation

Ian Curtis

2022-11-06

Load Packages

I first load the necessary packages.

```
library(baseballr)
library(tidyverse)
```

Introduction

`baseballr` is a package used to collect data on various baseball (MLB) statistics from multiple sources on the internet. It can also provide some interesting data on pre-selected trends and certain calculations.

As of October 2022, `baseballr` is capable of fetching data from the following sources:

- The MLB API
- The MLB Statcast database (Baseball Savant)
- Retrosheet
- NCAA
- Baseball Reference
- FanGraphs

This package is incredibly useful for searching for MLB data, especially when dataset joining is needed (such as combining statistics from Baseball Reference and FanGraph, for instance).

Demonstration

Below are a few ways in which the `baseballr` package might be used to grab data.

Statcast (Baseball Savant)

The Savant database is a large, searchable repository of MLB data extending back to 2008. The database can be searched on the web and contains a large number of custom filters to apply. The database will automatically create aggregate summaries according to selections, but the raw data is pitch-by-pitch and gives the researcher much freedom when using the data.

The package here will grab raw data based on the query which can either specify a specific batter or pitcher or request all of the raw data between a certain time frame.

```
# Search for all data for Max Scherzer in June 2021

scherzer <- statcast_search_pitchers(
  start_date = '2021-06-01',
  end_date = '2021-06-30',
  pitcherid = 453286)
head(scherzer)
```

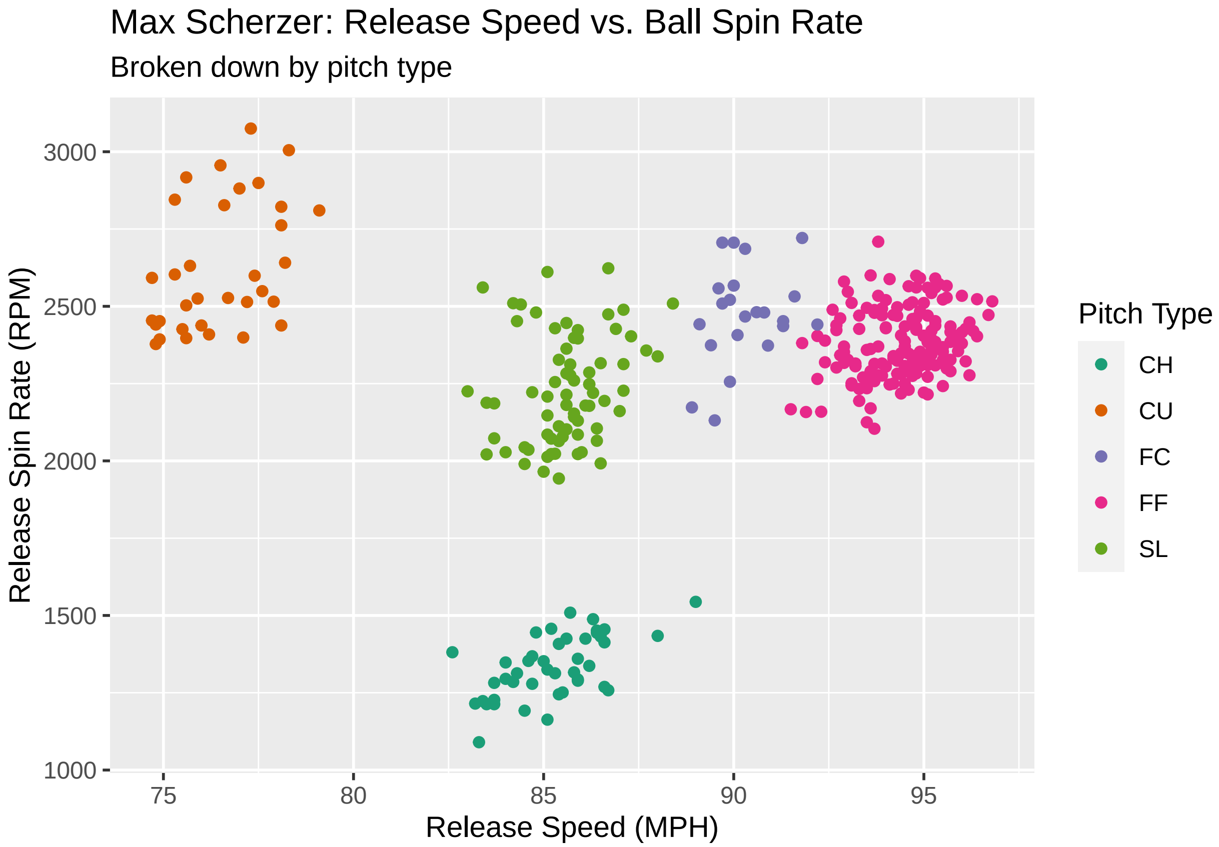
```
## # A tibble: 6 x 92
##   pitch_type game_date  release_~1 relea~2 relea~3 playe~4 batter pitcher events
##   <chr>      <date>          <dbl>   <dbl>   <dbl> <chr>    <dbl>   <dbl> <chr>
## 1 FF         2021-06-27       95     -3.26    5.59 Scherz~ 542932  453286 "stri~
## 2 CH         2021-06-27       84.3   -3.38    5.22 Scherz~ 542932  453286 ""
## 3 FF         2021-06-27       94.6   -3.16    5.52 Scherz~ 542932  453286 ""
## 4 SL         2021-06-27       85.8   -3.39    5.19 Scherz~ 542932  453286 ""
## 5 FF         2021-06-27       95.5   -3.18    5.5  Scherz~ 542932  453286 ""
## 6 CH         2021-06-27       83.2   -3.29    5.41 Scherz~ 542932  453286 ""
## # ... with 83 more variables: description <chr>, spin_dir <lgl>,
## #   spin_rate_deprecated <lgl>, break_angle_deprecated <lgl>,
## #   break_length_deprecated <lgl>, zone <dbl>, des <chr>, game_type <chr>,
## #   stand <chr>, p_throws <chr>, home_team <chr>, away_team <chr>, type <chr>,
## #   hit_location <int>, bb_type <chr>, balls <int>, strikes <int>,
## #   game_year <int>, pfx_x <dbl>, pfx_z <dbl>, plate_x <dbl>, plate_z <dbl>,
## #   on_3b <dbl>, on_2b <dbl>, on_1b <dbl>, outs_when_up <int>, ...
```

The above chunk searches for all of the pitch-by-pitch data for Max Scherzer in June 2021. The result is a large data frame containing attributes that can be pulled out for use.

From here, we might plot some data!

```
scherzer_plot <- scherzer %>%
  ggplot(aes(x = release_speed, y = release_spin_rate, color = pitch_type)) +
  geom_point() +
  labs(title = 'Max Scherzer: Release Speed vs. Ball Spin Rate',
       subtitle = 'Broken down by pitch type',
       x = 'Release Speed (MPH)',
       y = 'Release Spin Rate (RPM)') +
  guides(color = guide_legend(title = "Pitch Type")) +
  scale_color_brewer(palette = "Dark2")

scherzer_plot
```



Baseball Reference

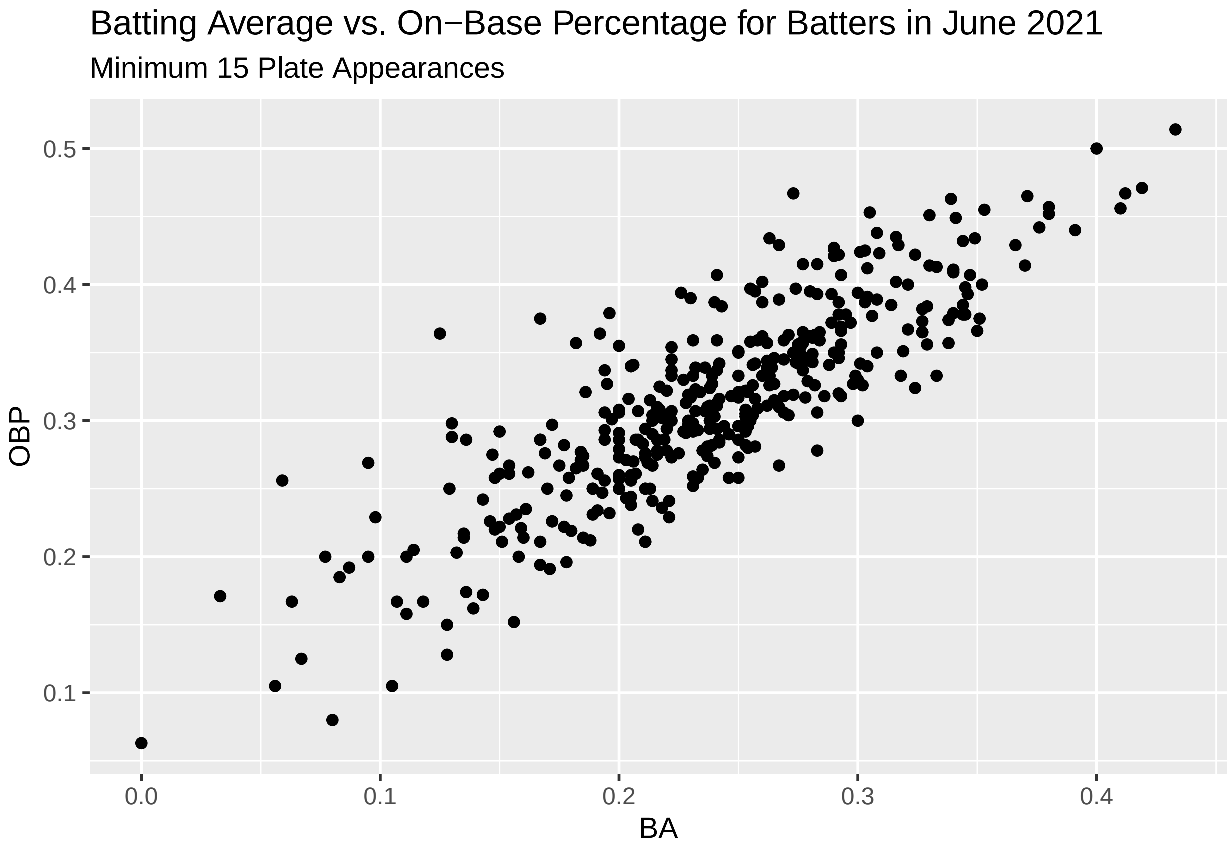
Baseball Reference is another source of baseball data. The package `baseballr` allows for aggregate player performance data to be scraped as well as historical standings at any date. There is also a function to calculate "team consistency". Baseball Reference might be used more for getting "typical" statistics such as batting average, ERA, and number of home runs.

```
bref_batter <- bref_daily_batter("2021-06-01", "2021-06-30")

head(bref_batter)
```

```
## # A tibble: 6 x 30
##   bbref_id season Name       Age Level Team      G    PA    AB     R     H   X1B
##   <chr>     <int> <chr>    <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 547989     2021 Jose Al~    31 Maj-~ Hous~    27   129   101    24    26    13
## 2 660670     2021 Trea Tu~    28 Maj-~ Wash~    28   123   113    24    39    27
## 3 642715     2021 DJ LeMa~    32 Maj-~ New ~    26   123   113    12    33    24
## 4 571431     2021 Freddie~    31 Maj-~ Atla~    28   122   108    20    33    23
## 5 656180     2021 Marcus ~    30 Maj-~ Toro~    26   122   110    24    29    15
## 6 501303     2021 Jonatha~    24 Maj-~ Cinc~    27   121    99    24    30    21
## # ... with 18 more variables: X2B <dbl>, X3B <dbl>, HR <dbl>, RBI <dbl>,
## #   BB <dbl>, IBB <dbl>, uBB <dbl>, SO <dbl>, HBP <dbl>, SH <dbl>, SF <dbl>,
## #   GDP <dbl>, SB <dbl>, CS <dbl>, BA <dbl>, OBP <dbl>, SLG <dbl>, OPS <dbl>
```

```
bref_batter %>%
  filter(PA >= 15) %>%
  ggplot(aes(x = BA, y = OBP)) +
  geom_point() +
  labs(title = 'Batting Average vs. On-Base Percentage for Batters in June 2021',
       subtitle = 'Minimum 15 Plate Appearances')
```



References

- https://baseballsavant.mlb.com/
- https://razzball.com/mlbamids/
- https://www.baseball-reference.com/
- https://billpetti.github.io/baseballr/index.html