Data Expedition - Statcast

Jake Coleman

September 22, 2016

Data Description

This dataset is pitch-level data for all Major League Baseball teams in the first week of the 2016 season, recorded by Statcast tracking technology. Using high-resolution cameras and radar equipment, Statcast can calculate variables describing the trajectory of the pitch (e.g. speed, movement, pitch type), as well as batted-ball information (e.g. distance/angle/speed the ball was hit). The data comes from baseballsavant.com, which nightly downloads and organizes Statcast data from MLB Advanced Media, L.P. This dataset was put together by a search performed on baseballsavant.com for all pitches thrown from 4/3/2016 to 4/7/2016, and then cleaned and trimmed for ease of exploration and analysis.

Exercises

Use ggplot2 for any plotting questions, and dplyr for all other questions.

1. **Basic exploration**
    a. Who threw the fastest pitch, and what was its speed?
    b. Who threw the slowest pitch, and what was its speed?
    c. How many different pitchers are in the dataset?
    d. Arrange the data by the pitches with the largest hit speed, descending. What was the pitch result of the top five?
    e. What was the farthest hit? What was its pitch type?
2. **More challenging exploration**
    a. Which pitcher had the highest number of called strikes? Which pitcher had the highest ratio of strikes (swinging or called) to balls?
    b. What percent of pitches thrown by Clayton Kershaw were curveballs? What about Madison Bumgarner?
    c. Which pitcher(s) gave up the most home runs, and how many did they give up? Note: this question is particularly challenging because we are looking for a summary on *outcome*-level data, when the data that we have is *pitch*-level data. You may assume an outcome here is defined by each unique combination of *pitcher, batter, ab_outcome, inning,* and *outs*. (*Hint: try as an intermediary step to get an event-level dataset.*)
3. **Pitch Location**
    a. The "strike zone" is defined to be the width of home plate (17 inches) and height from the batter's knees to the middle of his torso. This changes depending on the batter, which is why Statcast has *strike_zone_top* and *strike_zone_bottom*. Using dplyr's mutate() function, construct a variable titled "*in_zone*" which is TRUE for pitches in the strike zone and FALSE for pitches outside of the strike zone. What is the percentage of all pitches in the strike zone? (*Hint: in the data,* x_location *and* z_location *are measured in feet, and* x_location *is 0 in the very middle of the strike zone.*)
    b. What are the percentages of pitches in the strike zone grouped by pitch type? What changes do you notice? Does this change for right-handed pitchers versus left-handed pitchers?

c. Plot z-location versus x-location for right-handed pitchers, coloring pitches by whether or not they were in the strike zone. (I would suggest separate plots for each pitch type for clarity). What differences do you notice in the distribution of pitches?

4. **Pitch Type**

   a. What was the average speed of all pitches thrown? X-movement? Z-movement?
   b. Aggegating over everything ignores crucial structure of the data. What are the results when you group by pitch type, and how do they differ from above?
   c. Plot Z-movement against X-movement for all pitches. Then, plot it again but this time coloring pitches by their pitch type.
   d. You should see some clustering but also symmetry across the x-axis (X-movement). This suggests another aspect of the data - we will see this is the handedness of the pitcher. Plot Z-movement against X-movement for right-handed pitchers, coloring the pitches by their pitch type. Do the same for left-handed pitchers.
   e. Find the average speed, X-movement, and Z-movement for right-handed pitchers, grouped by pitch type. Which results changed substantially from when we looked at all pitchers? Do you have an explanation?

Potential Research Questions

Here are some example questions which may be of interest to MLB organizations.

- Which pitchers had the most effective curveball?
- What aspects of a pitch correspond with the most positive results?
- Which pitchers have been the most “unlucky" (in the sense that they have thrown quality pitches but have had poor results)?

Data Source

MLB Advanced Media, L.P. (2016). *Statcast* [Data file]. Available from https://baseballsavant.mlb.com/statcast_search