Moneyball with Pig and Hive Russ Lankenau rlankenau@maprtech.com

The provided data contains various information about Major League Baseball between 1921 and 2011. The data is freely available on the web from retrosheet.org¹. The majority of the data is in the form of event records, i.e. plays. Each play describes an event that occurs when a batter steps up to the plate. This could be an out, a hit, or some other event such as a runner stealing a base. In addition, there are other types of data such as team rosters, game information, player lists, team lists, and reconstructed data for games where no clear record exists.

info Information about a single game (e.g. time of day, weather, location)

start A starting player, including name, ID, position, home/away, and batting order

sub A player substituted in during play

play A single event during a game

com A comment inserted by the recorder

id The game ID, consisting of year, month, day, and game number for the day

version The file version

data Additional data about a player (e.g. number of errors during the game)

Table 1: Record Types

The records are comma-separated, and are intermixed within the event files. Each event file contains the home games for a single team for a single year. American League teams are listed in .EVA files, and National League teams are listed in .EVN files.

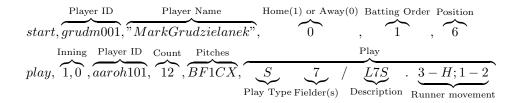


Figure 1: Sample Records

The provided sample code² extracts one very basic metric, a count of plays for each player, grouped by ball location and result. Using the sample as a guide, can we extract other interesting data from this data set?

- What is the relationship between player names and IDs? (e.g. one-to-one?)
- Which player has the most career RBIs?
- Can we tell which players' performance drops off most in later innings?
- Which players are the most consistent between home games and away games?
- •
- •

 $^{^1 \}mathrm{http://www.retrosheet.org/game.htm}$

 $^{^2} https://github.com/rlankenau/mapr-moneyball\\$

²Rarely, a pitchout will result in a strike or foul ball. These are recorded as Q and R respectively.

³n is 1, 2, or 3, corresponding to base thrown to. The base number is preceded by a '+' if the pickoff was thrown by the pitcher.

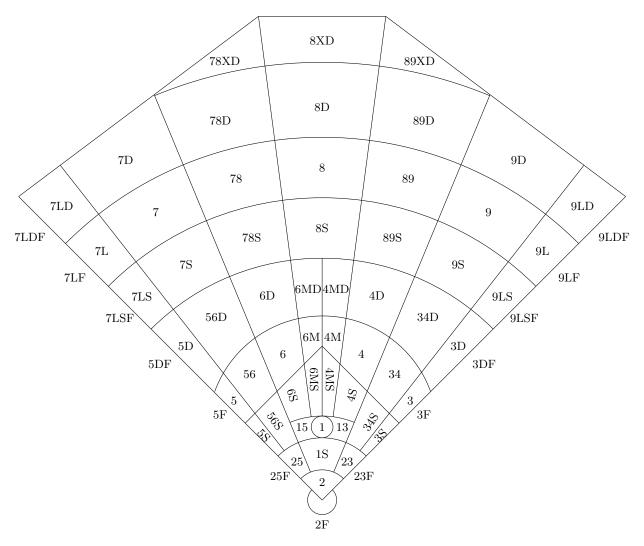


Figure 2: Field Locations

| | | | | | | C S | Called Strike Swinging Strike |
|---------------|--------------|---------------|--------------|--------------|------------------------|-----------------|----------------------------------|
| | | | | | | В | Ball |
| 1 | Catcher | (none) | Out | | | \mathbf{F} | Foul Ball |
| 2 | Pitcher | \mathbf{S} | Single | | | $_{\rm L}$ | Foul Bunt |
| 3 | First Base | D | Double | | | \mathbf{M} | Missed Bunt |
| 4 | Second Base | ${ m T}$ | Triple | | | P^3 | Pitchout |
| 5 | Third Base | $_{ m HR}$ | Home Run | G | Ground Ball | I | Int. Ball |
| 6 | Shortstop | HP | Hit by pitch | L | Line Drive | $_{\mathrm{H}}$ | Hit by pitch |
| 7 | Left Field | W | Walk | Ρ | Pop Fly | K | Strike (unknown) |
| 8 | Center Field | IW | Int. Walk | \mathbf{F} | Fly Ball | U | Unknown |
| 9 | Right Field | WP | Wild Pitch | В | Prefix indicating bunt | n^4 | Pickoff |
| (a) Positions | | (b) Play Type | | | (c) Description | | (d) Pitches |

Table 2: Play Codes