Baseball Value

Matthew Helbig

2019-09-28

Introduction

The goal of this project is to identify the attributes of the players with the most value and the least value, as well as predicting which players will have the most and least value in the future.

Getting started

The first step we'll take is to identify which datasets we'll be working with. The Lahman Database has a lot of good information that we'll need, namely Player Value data, salary, and dates of birth.

We'll download that from http://www.seanlahman.com/baseball-archive/statistics/

It's possible to download the database as a .CSV file, and you could technically do everything needed for this project without using SQL, but it's a good practice to get used to working with databases, especially since a lot of the datasets you'd see in the real world will be bigger than the one we're currently working with.

Sean Lahman has posted the entire .sql file on his website, so importing it into MySQL is as simple as hitting 'Data Import' and loading that database into its own schema.

Once the Lahman database is all settled in its own schema, we want to start to work with that data. It's definitely possible to run all your SQL commands inside the MySQL environment itself, but RStudio has packages available for installation that make it easy to work with your database right from R. This has a few advantages, most notably the fact that you don't have to keep hopping back and forth between programs.

Talking to our Database

The first thing you'll do is make sure that the RMySQL package is installed. This is as simple as calling "install.packages("RMySQL")" once.

Next we'll load up our RMySQL library and get to work.

```
library(RMySQL)
```

```
## Loading required package: DBI
```

```
dbListTables(mydb)
```

```
##  [1] "AllstarFull"         "Appearances"         "AwardsManagers"
##  [4] "AwardsPlayers"       "AwardsShareManagers" "AwardsSharePlayers"
##  [7] "Batting"             "BattingPost"         "CollegePlaying"
## [10] "Fielding"            "FieldingOF"          "FieldingOFsplit"
## [13] "FieldingPost"        "HallOfFame"          "HomeGames"
## [16] "Managers"            "ManagersHalf"        "Master"
```

```
## [19] "Parks"              "Pitching"            "PitchingPost"
## [22] "Salaries"           "Schools"             "SeriesPost"
## [25] "Teams"              "TeamsFranchises"     "TeamsHalf"
```