ECON 128: 181: Data Science

Claremont McKenna College

Profs. A. Vossmeyer & M. Salloum August 29th, 2017

Problem Set 1

1 Description

In this assignment we will practice answering queries posed over a dataset,

In this assignment you will apply your data wrangling and exploratory data analysis skills to baseball data. In particular, we want to know how well did Moneyball work for the Oakland A's. Was it worthy of a movie?

1.1 A bit of background

Well be looking at data about teams in Major League Baseball. A couple of important points: Major League Baseball is a professional baseball league, where each team aims to win as many games out of a 162 game season as possible. In principle, better players are costlier, so teams that want good players need to spend more money. Teams that spend the most, frequently win the most. So, the question is, how can a team that cant spend so much win? The basic idea that Oakland (and other teams) used is to redefine what makes a player good. I.e., figure out what player characteristics translated into wins. Once they realized that teams were not really pricing players using these characteristics, they could exploit this to pay for undervalued players, players that were good according to their metrics, but were not recognized as such by other teams, and therefore not as expensive. You can get more information about this period in baseball history from `https://en.wikipedia.org/wiki/Moneyball`.

You will be using data from a very useful database on baseball teams, players and seasons curated by Sean Lahman available at `http://www.seanlahman.com/baseball-archive/statistics/`. The database has been made available as a sqlite database `https://github.com/jknecht/baseball-archive-sql` sqlite is a light-weight, file-based database management system that is well suited for small projects and prototypes.

You can read more about the dataset here: `http://seanlahman.com/files/database/readme2014.txt`. You can download the sqlite file directly from github at `https://github.com/jknecht/baseball-archive-sqlite/raw/master/lahman2014.sqlite`.

You will be accessing the sqlite database in python using the sqlite package. This package provides a straightforward interface to extract data from sqlite databases using standard SQL commands.

Once you establish a connection with the sqlite database, you can store query results directly in a pandas dataframe using the $read_sql function. For example, here's how you would tabulate total league payrolls for each

The homework assignment is found on the course github page: `https://github.com/msalloum/econ128/tree/master/Homeworks/HW2`. Starter code is provided to ingest the data file, and the problems/questions are given within the iPython notebook.

2 Submission

Please submit the homework files via Sakai. Alternatively, for 5 extra points, create a github account and a repository called ECON128, and push your homework solution to that ECON128/HW2. You must still submit the repository location via SAKAI is you choose to do the extra credit.