S&DS 173
Ydata: Analysis of Baseball Data

Ethan Meyers

Overview

Course overview

- Introductions
- Syllabus and logistics

The history of statistics in baseball

Baseball statistics and structured data

Python basics and lab 0

Contact Information

Email: ethan.meyers@yale.edu

~~Office: 24 Hillhouse Ave, Room 206~~ Zoom

Planned office hours:

- Monday 11am
- Friday 2pm

About me

Visiting assistant professor at Yale

Assistant professor of Statistics Hampshire College

Research Fellow at the Center for Brains, Minds and Machines at MIT

Research: Machine learning to analyze neural data

Introductions

About you:

- Name
- A bit about your background
  - E.g. experience with baseball, programming and Statistics
- Favorite baseball team if you have one
- Any topic you are particularly interested related to class
- Anything else you want to say

Teaching Assistants

Neel Malhotra: neel.malhotra@yale.edu

Neel’s office hours will be posted soon

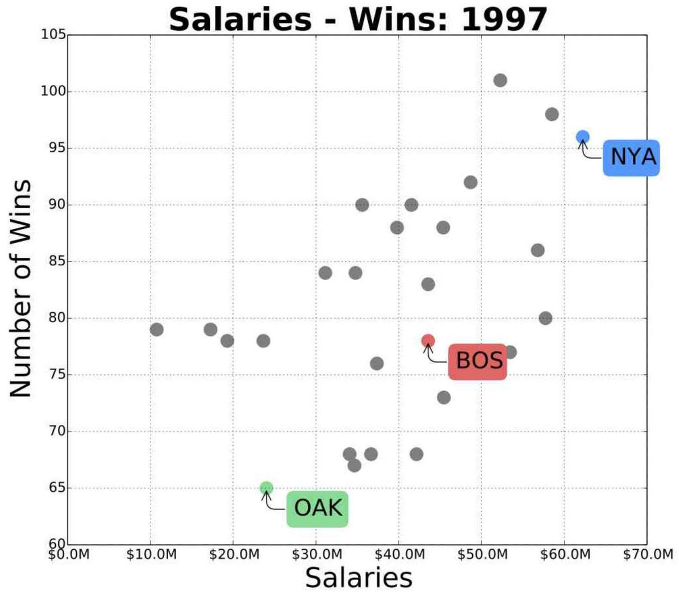
Learning goals

1. To explore statistical/data science concepts by analyzing real and simulated baseball data

2. To learn how to use Python to analyze, visualize and wrangle data

3. To understand how Major League Baseball teams are using Data Science to improve their chances of winning

Additional goal: to have fun/socialize!



Why use baseball to study Statistics and Data Science?

High degree of randomness

- Very good players hit safely 3 out of 10 times (ave = .300)
- Bad players hit safely 2 out of 10 times (ave = .200)

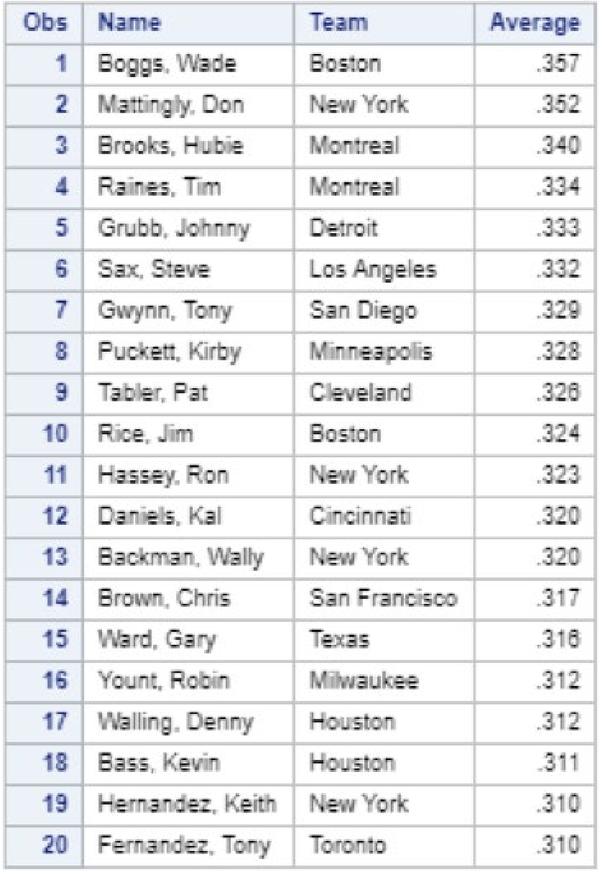
Contains a rich structure that repeats, which makes it possible to isolate components and analyze them

- Discrete events makes it relatively easy to analyze:
    - Pitches -> plate appearances -> innings -> games -> seasons

Lots of data available

- Data going back to 1871

Overall: Excellent system to practice using data to answer real questions



Obs	Name	Team	Average
1	Boggs, Wade	Boston	.357
2	Mattingly, Don	New York	.352
3	Brooks, Hubie	Montreal	.340
4	Raines, Tim	Montreal	.334
5	Grubb, Johnny	Detroit	.333
6	Sax, Steve	Los Angeles	.332
7	Gwynn, Tony	San Diego	.329
8	Puckett, Kirby	Minneapolis	.328
9	Tabler, Pat	Cleveland	.326
10	Rice, Jim	Boston	.324
11	Hassey, Ron	New York	.323
12	Daniels, Kal	Cincinnati	.320
13	Backman, Wally	New York	.320
14	Brown, Chris	San Francisco	.317
15	Ward, Gary	Texas	.316
16	Yount, Robin	Milwaukee	.312
17	Walling, Denny	Houston	.312
18	Bass, Kevin	Houston	.311
19	Hernandez, Keith	New York	.310
20	Fernandez, Tony	Toronto	.310

Prerequisites

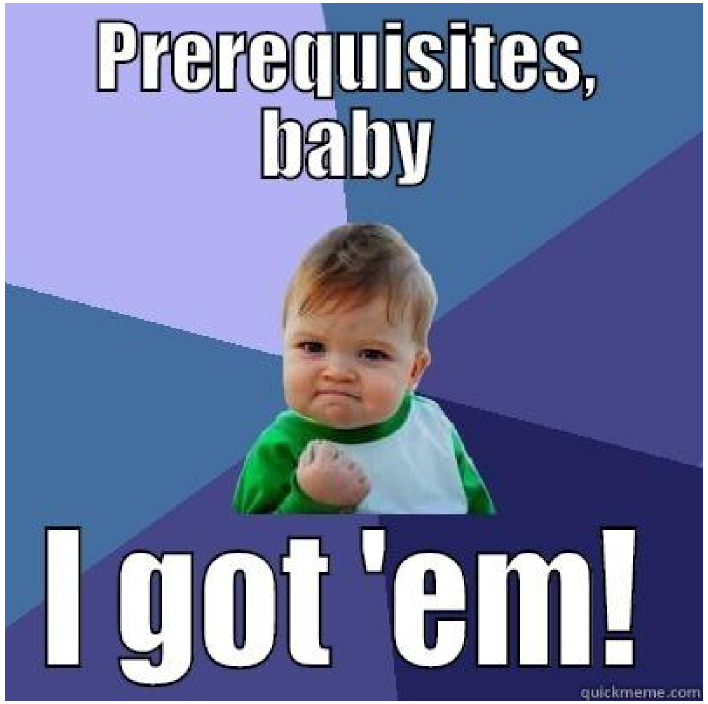
Taking the main YData class (S&DS 123) either previously or concurrently

Some familiarity with baseball

- If you are not familiar at all with baseball talk to me and we will figure out a time to go over the basics

No other background is assumed

- We will be learning Python starting from the basics



Textbook: Y123 online textbook from data8

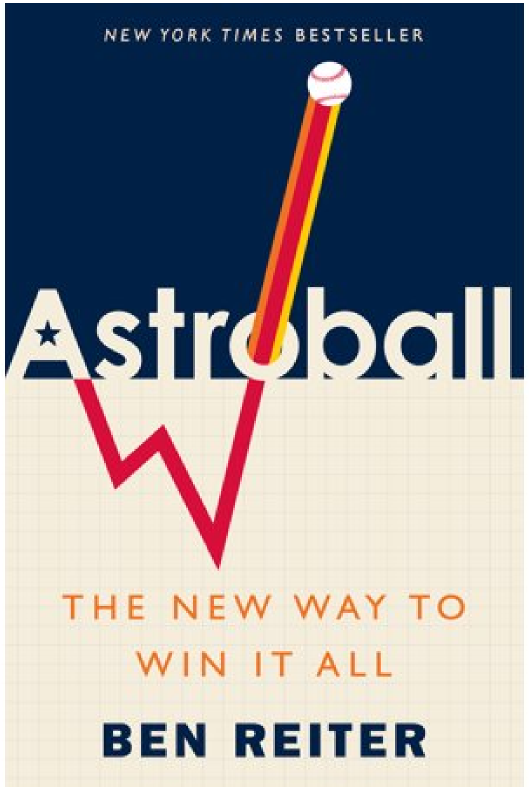
Online textbook for Y123

- https://www.inferentialthinking.com/chapters/intro

If you are not taking Y123, looking over the class material will be helpful

- http://ydata123.org/sp21/calendar.html

Reading: Astroball by Reiter

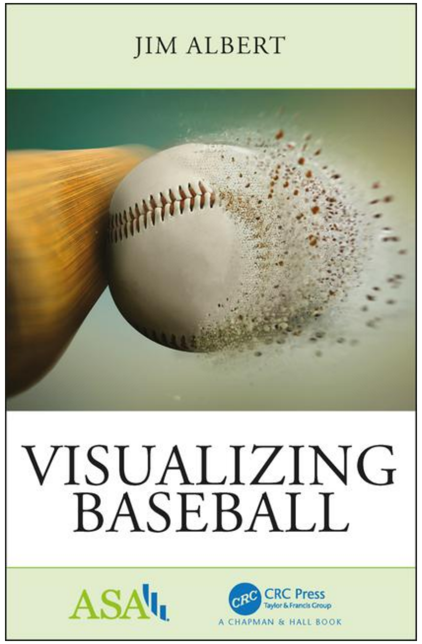


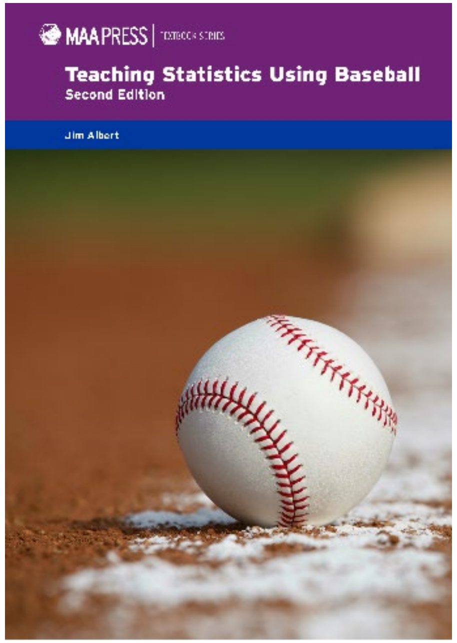
Addition reading and other resources will be posted to Canvas:

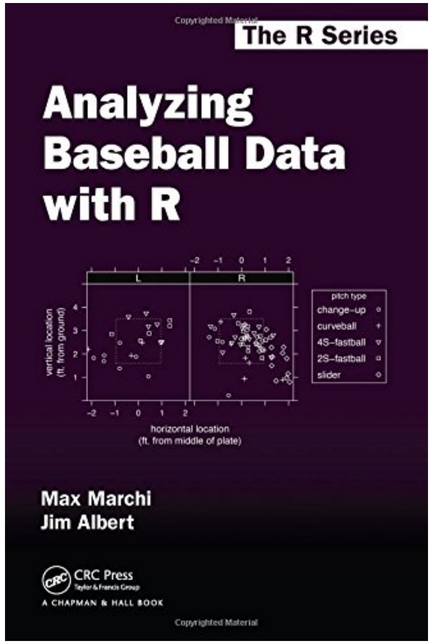
https://yale.instructure.com/courses/56241

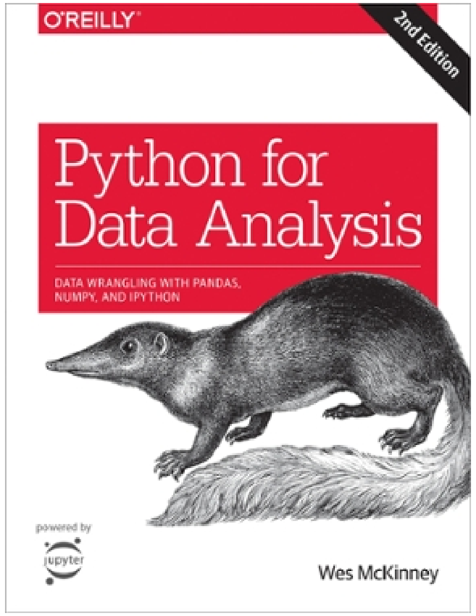
Other relevant books

- Visualizing Baseball
- Teaching Statistics Using Baseball
- Analyzing Baseball Data with R
- Python for Data Analysis




MAA PRESS | TEXTBOOK SERIES
Teaching Statistics Using Baseball
Second Edition
Jim Albert





Class structure

We will start with a discussion of Astroball

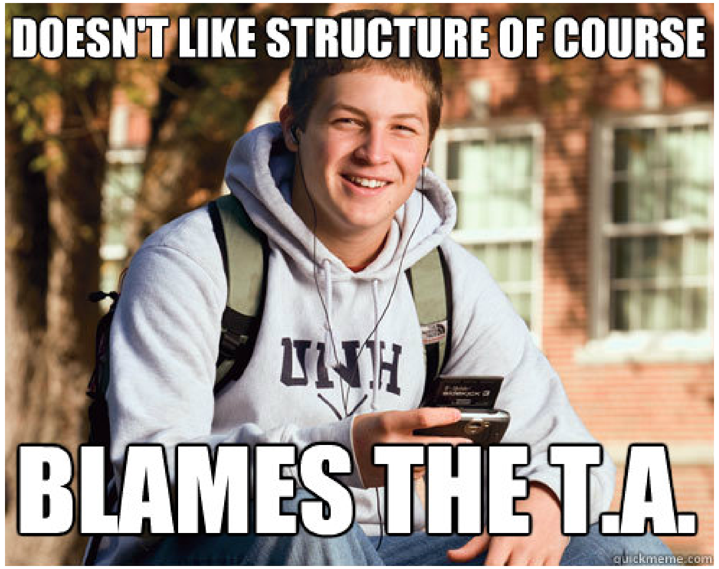
We will then cover material on a particular topic related to analyzing baseball data

- The class will loosely follow topics covered in Y123

The last part of the class will be an opportunity to work on a Jupyter notebook “lab”

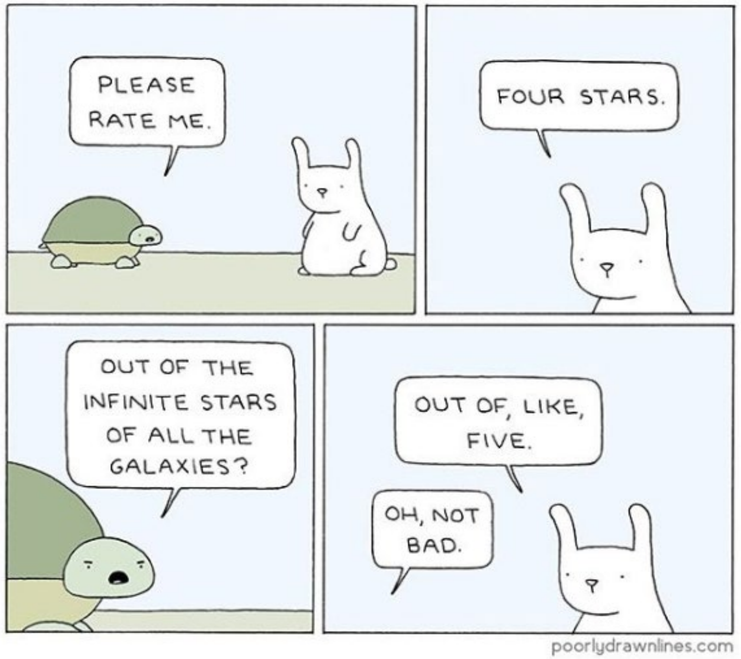
- You will finish the lab as the homework for the week

There might be some prerecorded videos to watch before class



Assignments and grades

1. Lab homework (55%)
   - Exploring questions in baseball using Python
   - Weekly: 10 total, lowest score will be dropped



Homework lab policies:

- You may discuss questions with other but the work you turn in must be your own
- Homework will be started in class on Wednesdays and are due at 11:30pm on Mondays
- Late worksheets (90%) credit if turned in by 11:30pm on Tuesday

Assignments and grades

2. Final project (20%)

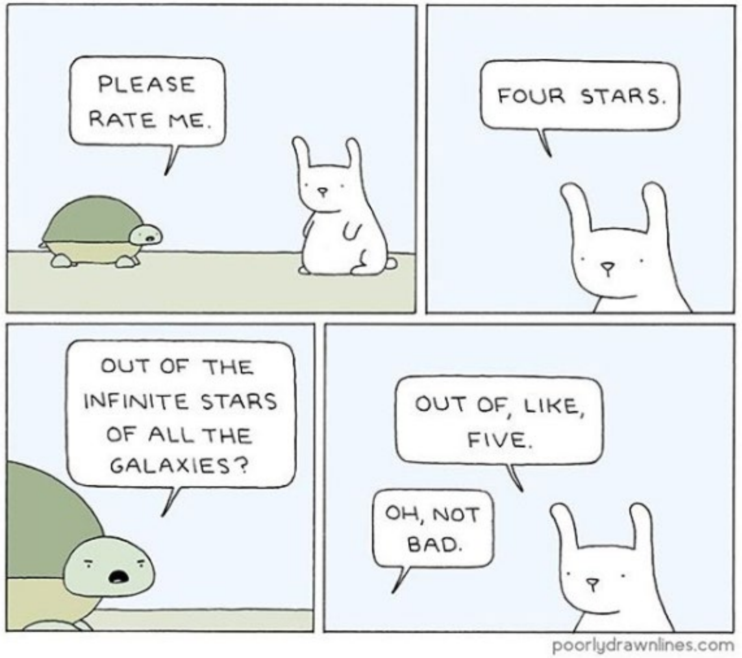
- This will be an opportunity to explore a question related to baseball in more depth. ~10 page paper.
- You will give a ~5 minute presentation on your project.

3. Exam (20%)

- Midterm: March 31st

4. Participation (5%)

- Asking and answering questions in class and on Ed Discussions



Policies

Accommodation: please let me know if you have accommodations for homework and/or exams

Academic dishonesty: Don’t do it!

- You work with others on the homework but the work you turn in needs to be your own (i.e., you need to understand the concepts)
- You can’t talk with others on exam, etc.

Class schedule

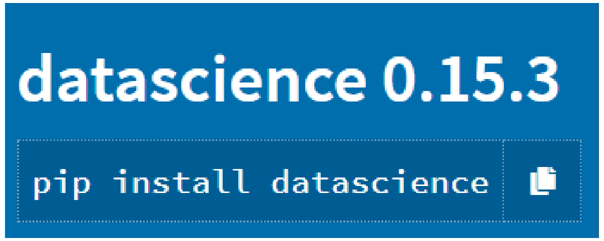
WEEK	DATE	TOPIC	HOMEWORK	DUE
1	Feb 3	Introduction to baseball and Python programming	0	
2	Feb 10	Summary statistics and plots	1	14-Feb
3	Feb 17	Data wrangling I	2	21-Feb
4	Feb 24	Data wrangling II	3	28-Feb
5	Mar 3	Probability and simulations with games I	4	7-Mar
6	Mar 10	Probability and simulations with games II	5	14-Mar
7	Mar 17	Hypothesis tests	6	21-Mar
	Mar 24	Break day		
8	Mar 31	Midterm exam		
9	Apr 7	Parametric hypothesis tests and confidence intervals	7	11-Apr
10	Apr 14	Calculating confidence intervals with the bootstrap and relationships between measures	8	18-Apr
11	Apr 21	Linear regression	9	25-Apr
12	Apr 28	Multiple regression	10	2-May
13	May 5	Transformations in regression and cross-validation		

Tentative plan for the semester and topics covered

Weeks 1-5: Descriptive statistics, data wrangling and visualization

matplotlib

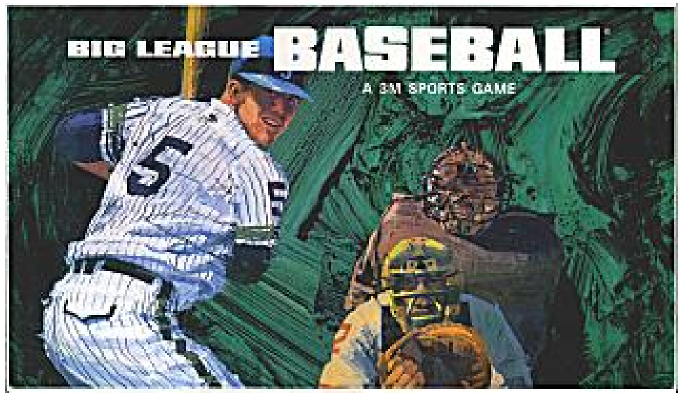
NumPy

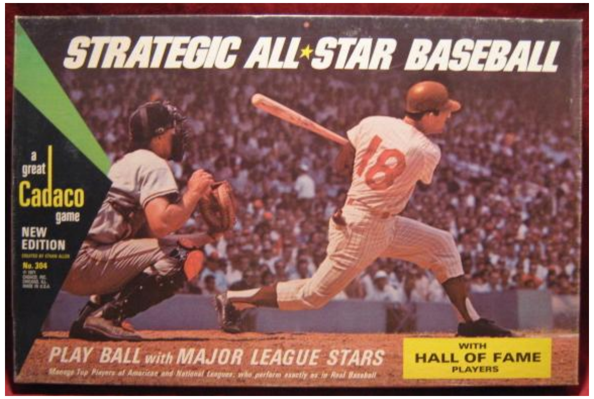


datascience 0.15.3

```
pip install datascience
```

Weeks 6-7: Probability models and simulations using table top games



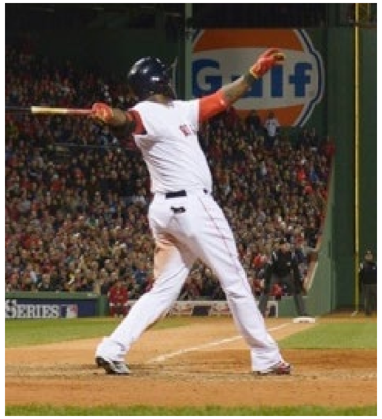

STRATEGIC ALL★STAR BASEBALL
a great Cadaco game
NEW EDITION
PLAY BALL with MAJOR LEAGUE STARS
WITH HALL OF FAME PLAYERS

Tentative plan for the semester and topics covered

Weeks 8-10: Inferential Statistics

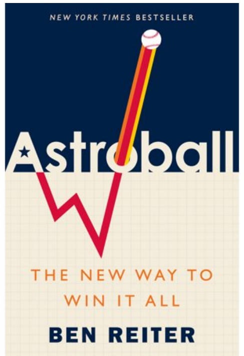
- Hypothesis tests and confidence intervals
- Who is better?

Derek Jeter



David Ortiz

Weeks 11-13: Linear regression classification and ethics



Types of questions we will be trying to answer

How much more valuable is a home run compared to a single?

Who is the best baseball hitter of all time?

Which statistics best capture a baseball players ability?

- i.e., is on base percentage a better measure than batter average?

Are certain baseball players streaky or clutch hitters?

etc.

Class survey

In order for me to get to know you and to better adjust the class to your interests, please fill out the class survey on canvas

- Under the Quizzes link on the left

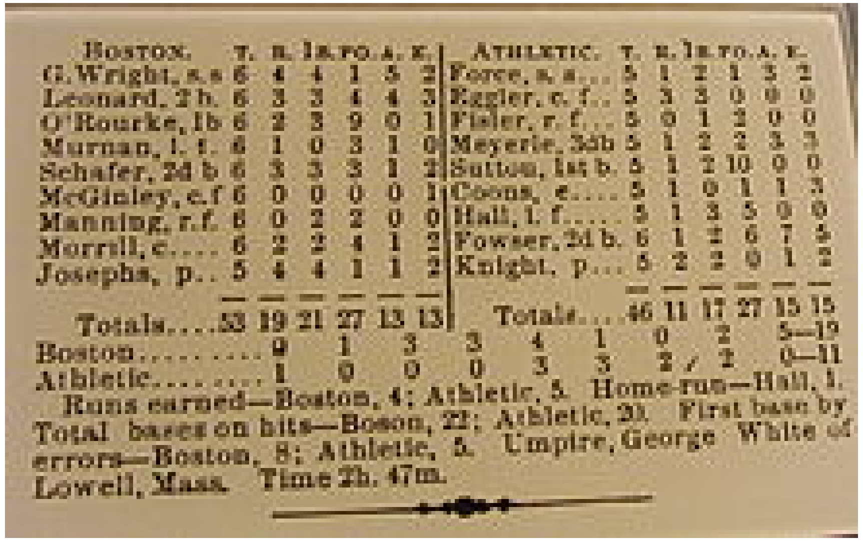
Any questions about the class logistics???

The history of baseball statistics

Early statistics

Henry Chadwick (1824-1908) created the first box score in the 1859 issue of Clipper.

- First to use K for strike outs, said to have invented batting average and earned run average
- Did not record walks because he did not feel they reflected a batters skill



BOSTON.	T.	R.	1B.	PO.	A.	E.	ATHLETIC.	T.	R.	1B.	PO.	A.	E.
G. Wright, s. s	6	4	4	1	5	2	Force, s. s....	5	1	2	1	3	2
Leonard, 2 b.	6	3	3	4	4	3	Eggler, c. f...	5	3	3	0	0	0
O'Rourke, 1b	6	2	3	9	0	1	Fisler, r. f...	5	0	1	2	0	0
Murnan, l. f.	6	1	0	3	1	0	Meyerle, 3db	5	1	2	2	3	3
Schafer, 3d b	6	3	3	3	1	2	Sutton, 1st b.	5	1	2	10	0	0
McGinley, c. f	6	0	0	0	0	1	Coons, c....	5	1	0	1	1	3
Manning, r. f.	6	0	2	2	0	0	Hall, l. f....	5	1	3	5	0	0
Morrill, c....	6	2	2	4	1	2	Fowser, 2d b.	6	1	2	6	7	4
Josephs, p..	5	4	4	1	1	2	Knight, p....	5	2	2	0	1	2
Totals....	53	19	21	27	13	13	Totals...	46	11	17	27	15	15

Boston..........	0	1	3	3	4	1	0	2	5—19
Athletic..........	1	0	0	0	3	3	2	2	0—11

Runs earned—Boston, 4; Athletic, 5. Home-run—Hall, 1. Total bases on hits—Boson, 22; Athletic, 20. First base by errors—Boston, 8; Athletic, 5. Umpire, George White of Lowell, Mass. Time 2h. 47m.

Early statistics

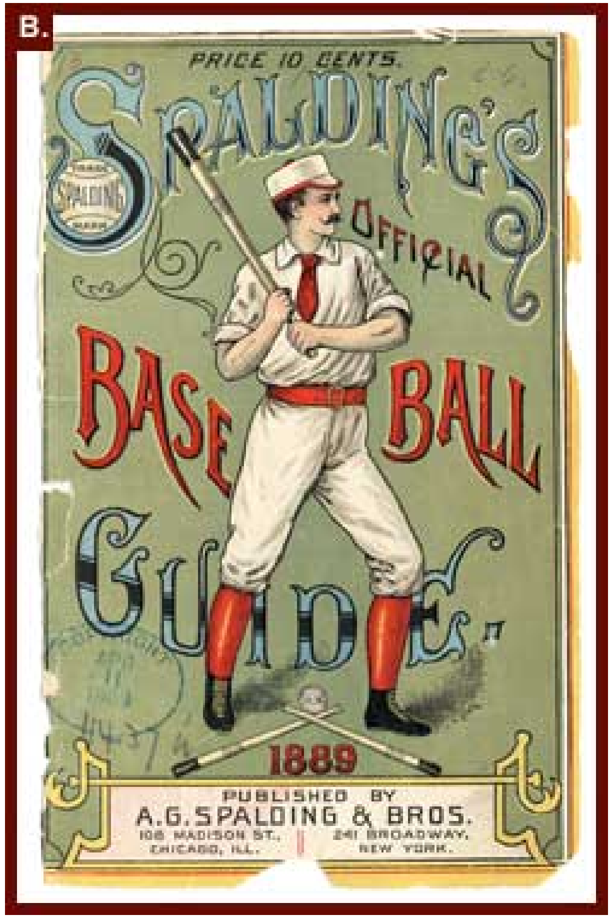
Classic statistics

Most prominent hitting statistics:

- Batting average, RBIs, and home runs

Most prominent pitching statistics:

- Wins, earned run average, strike outs



Sabermetrics

Around 1970 Bill James, and others began to question how useful traditional measures of performance

- i.e., are batting average, pitcher wins, etc. the best ways to tell how good a player is?

Sabermetrics definitions:

1. The empirical or mathematical/Statistical study of baseball
2. "the search for objective knowledge about baseball"

   - Bill James

Name comes from 'Society for American Baseball Research' (SABR), a group started in 1971

- Pre-computers, so they **had to compile all information from box scores by hand since there was no encyclopedia that had pitch-by-pitch data**

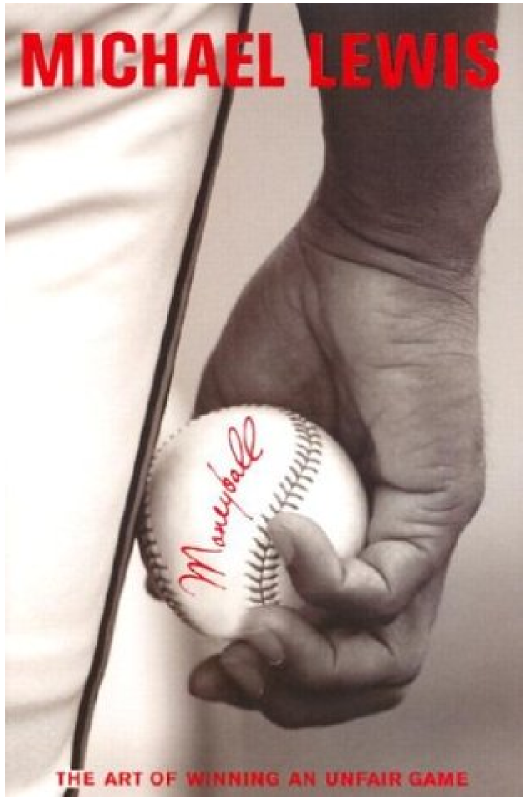
Sabermetrics first widely introduce to the public in 1982 with the publication of *Bill James Baseball Abstract*

Bill James, 1981

Moneyball

Story about how the Billy Bean, the general manager of the A’s, was able to put together a top ranked team in 2002 on a tight budget by finding undervalued players using advanced statistics

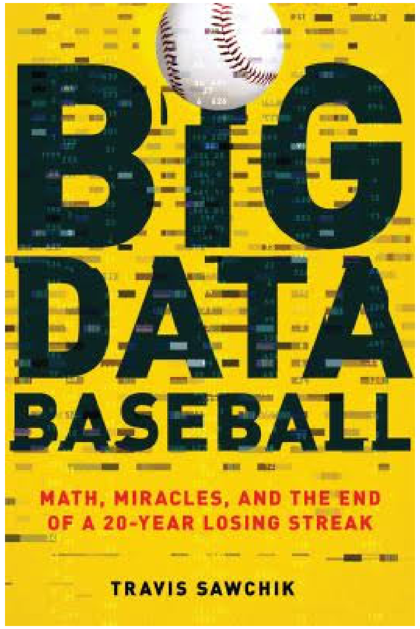
Some of the claims in the book might be exaggerated but it had a big impact on the expansion of major league clubs doing advanced data analyses

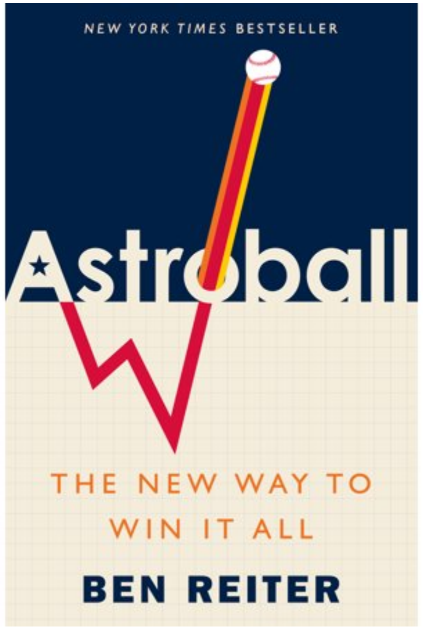


Sabermetrics continues to advance

Several books have been written about more recent sabermetric advances

- 2013 Pittsburgh Pirates
- 2017 Astros





Bill James Online
14.3K Tweets

Bill James Online
@billjamesonline

billjamesonline.com Joined September 2009

571 Following **50.9K** Followers

A few prominent Sabermetric publications/websites

Society for American Baseball Research

Bill James Online

Baseball Analysts

Baseball Prospectus

Beyond the Box Score

Fan Graphs

The Hardball Times

Tango Tiger

Statcast: MLBAM @ 2014 MIT Sloan Analytics Conference

Analyzing a catch by Jason Heyward

Baseball data sets

Lahman Database: Season-by-season data

Retrosheet Game-by-Game data

Retrosheet Play-by-Play data

PITCHf/x: Pitch-by-Pitch location, pitch type data (2006)

Statcast: high-accuracy tracking of player movements (2015)

Using cognitive neuroscience to improve player performance

https://www.wsj.com/articles/baseballs-science-experiment-1411135882

Python and the datascience package

Python basics and using Jupyter notebooks are reviewed in Lab 0

- my_name = "Ethan"   # an assignment statement
- 3**2   # math
- Etc.

Python also has packages that you can import that make available additional functions and objects

- import numpy as np

A package we will use extensively in this class is Berkeley's datascience package which is part of Berkeley's Data8 and the YData classes

- Looking at the documentation will be very useful: http://data8.org/datascience/

Python and the datascience package

To import the datascience package we will use:

- from datascience import *

We will extensively use Table objects from the datascience package in this class to process structured data

- See the documentation at: http://data8.org/datascience/tables.html

We can create a Table object by reading in data from a .csv file

- batting = Table.read_table('Batting.csv')

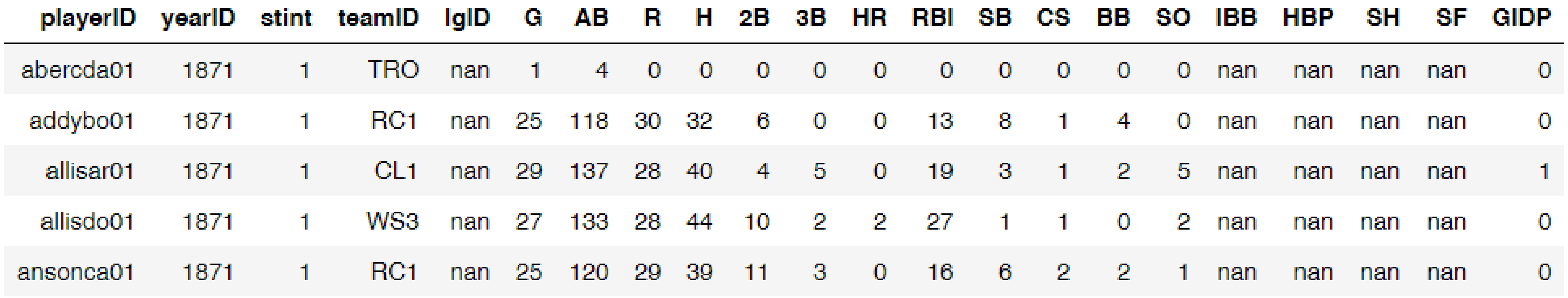
Python and the datascience package

An object in Python is a combination of data and functions that operate on the data

- These functions that operate on the data are called *methods*

Some Table methods you will use on the *batting* table in lab 0 are:

- batting.show(5)   # shows the first 5 rows of a Table



playerID	yearID	stint	teamID	lgID	G	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	IBB	HBP	SH	SF	GIDP
abercda01	1871	1	TRO	nan	1	4	0	0	0	0	0	0	0	0	0	0	nan	nan	nan	nan	0
addybo01	1871	1	RC1	nan	25	118	30	32	6	0	0	13	8	1	4	0	nan	nan	nan	nan	0
allisar01	1871	1	CL1	nan	29	137	28	40	4	5	0	19	3	1	2	5	nan	nan	nan	nan	1
allisdo01	1871	1	WS3	nan	27	133	28	44	10	2	2	27	1	1	0	2	nan	nan	nan	nan	0
ansonca01	1871	1	RC1	nan	25	120	29	39	11	3	0	16	6	2	2	1	nan	nan	nan	nan	0

Python and the datascience package

An object in Python is a combination of data and functions that operate on the data

- These functions that operate on the data are called methods

Some Table methods you will use on the *batting* table in lab 0 are:

- batting.show(5)   # shows the first 5 rows of a Table
- batting.select()   # select a subset of columns from a Table
- batting.take()   # get a subset of rows from a Table
- batting.sum()   # sums the values in a column
- batting.sort()   # arrange the rows in a table based on the values in a column

Lab 0

Let’s start on Python and exploring data!

https://github.com/emeyers/SDS173/

If you need help installing Python let me know

- The instructions are on Canvas

Ask questions as they come up

For next class

Read the preface and prologue to Astroball

- I will post scanned copies to the class Canvas site

Complete lab 0

- This will not be turned in but good practice to make sure you are ready for the class