

CS 149: Understanding Statistics Using Baseball



Outline for today

Introductions

Go over syllabus

The history of statistics in baseball

Baseball statistics and structured data

Introductions

About you:

- Name
- Preferred gender pronouns
- Year at Hampshire: Div I, II-first year, II-second year, III
- A bit about your background
 - E.g. experience with baseball, programming and Statistics
- Anything else you want to say

Is there any topic you are particularly interested in?
i.e., why are you interested in this class?

General Information

My contact info:

Email: emeyers@hampshire.edu

Phone: x5500

Office: ASH 133

Drop-in office hours:

- Monday 2:30-3:30pm
- Tuesday 12:30-1:30
- By appointment

Website:

<https://moodle.hampshire.edu/course/view.php?id=5532>

Course objectives

1. To understand how to use **statistical methods and thinking** to make sense of baseball **and other systems** that involve random processes
2. To be able to use the R programming language to analyze data

Note: statistics vs. Statistics

statistics: are numerical summaries of data

Statistics: is the mathematics of collecting, organizing and interpreting data

Why use baseball to study Statistics?

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

Background knowledge

No prerequisites

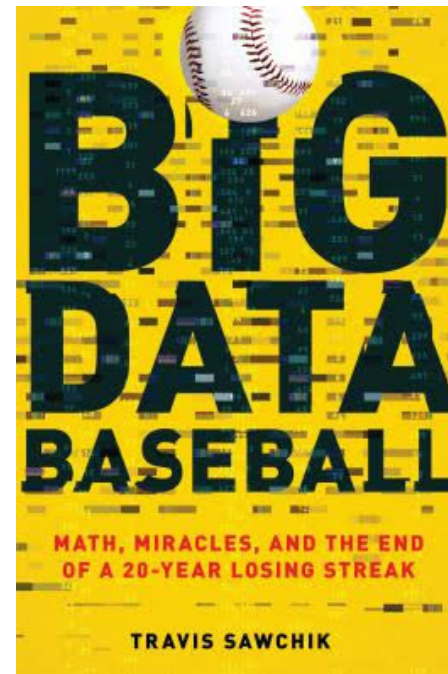
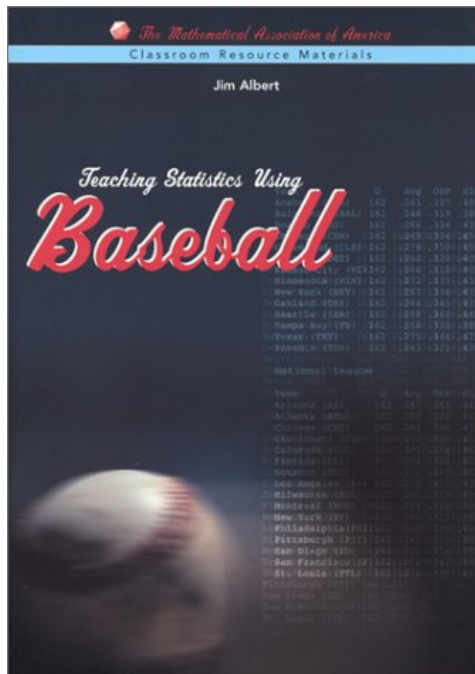
Familiarity with the basic rules of baseball will be useful

- We will briefly go over the rules and watch an innings of baseball next class
 - We can also arrange additional time for students who want

Required Textbooks

Teaching Statistics Using Baseball (TSUB)

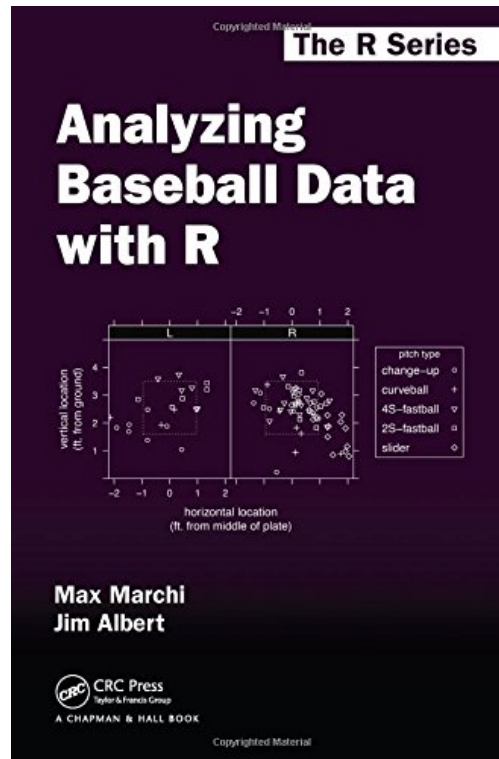
Big Data Baseball (BDB)



Highly recommended textbook

Analyzing Baseball Data with R

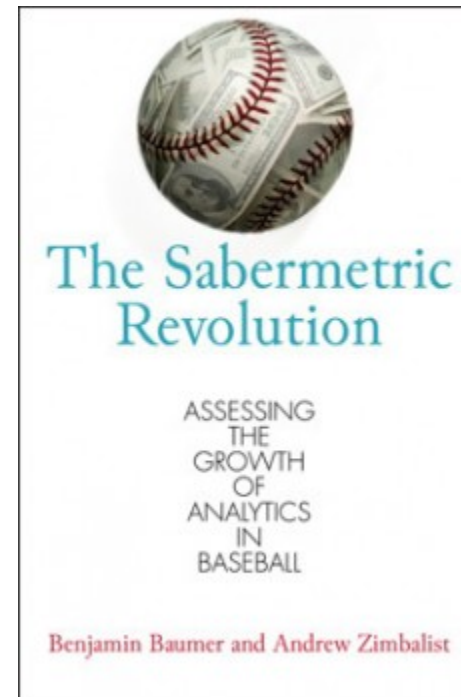
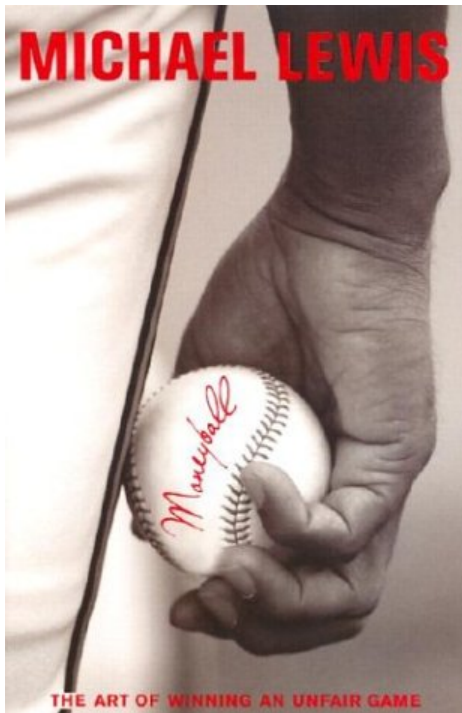
- Available as an e-book through Hampshire



Other recommended books

Moneyball

The Sabermetric revolution



Topics covered

Visualizing and exploring a single batch of data

Comparing batches of data and standardization

Relationships between measured variables

Introduction to probability

Introduction to statistical inference

- Hypothesis tests and confidence intervals

If there is time:

- Modeling baseball games
- Other sabermetric topics

Types of questions we will be trying to answer using statistics

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

Which statistics best capture a baseball players ability

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

Who is the best baseball hitter of all time?

Are certain baseball players streaky or clutch hitters?

Did the Tigers make a good decision signing Miguel Cabrera for \$292 million over the next 8 years?

Class assignments

1. Weekly worksheets that involve R and other problems
 - Assigned on Thursdays and they are due the following Sunday at midnight
 - Also class readings from Teaching Statistics Using Baseball
2. Weekly 'quote and reactions' to Big Data Baseball chapter
3. One 'better know a player presentations' (on Mondays)
 - 5-10 minute presentation about a baseball/softball player (or another topic related to the class)
4. A final project where you analyze related to baseball (or the class more generally) and write up a 5-10 page paper about the results
5. A class presentation about your final project

Policies

Very important to turn your work in on time!

- Three strikes late policy
- If you turn in three assignments late you don't get an evaluation

Attend class

You can use a laptop for notes, but obviously do not use it for anything not related to the class

Check your Hampshire email and/or setup mail forwarding. I will send announcements to your Hampshire email account

Policies

Incompletes will only be given out in very exceptional circumstances

Special needs: please let me know if you have a disability or special need. Also it would be a good to talk to Aaron Ferguson – x5498

Academic dishonesty: you can work on the worksheets with others but the work you turn in needs to be your own (i.e., you need to understand the concepts).

To make this class a success...

Actively engaged in the class

- Do the readings and join in the discussions

Community learning environment

- Share interesting things you've learned
- I want to learn from you too

Following your interests

- I can adapt the class to people's interests

Class survey

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

The history of baseball statistics



Early statistics

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

- First to use K for strike out, 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.							ATHLETIC.								
G.	L.	O.	M.	S.	M.	J.	F.	E.	S.	L.	O.	M.	S.	M.	J.
G. Wright, s.s.	6	4	4	1	5	2	Force, s.s.	5	1	2	1	3	2		
Leonard, 2b.	6	3	3	4	4	3	Egler, 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, i. f.	6	1	0	3	1	0	Meyerle, 3sb.	5	1	2	2	3	3		
Schafer, 2d 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, i. f.	5	1	3	5	0	0		
Morrill, c.	6	2	2	4	1	2	Fowler, 2d b.	6	1	2	6	7	5		
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.....9 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—Boston, 22; Athletic, 20. First base by															
errors—Boston, 8; Athletic, 5. Umpire, George White of															
Lowell, Mass. Time 2h. 47m.															

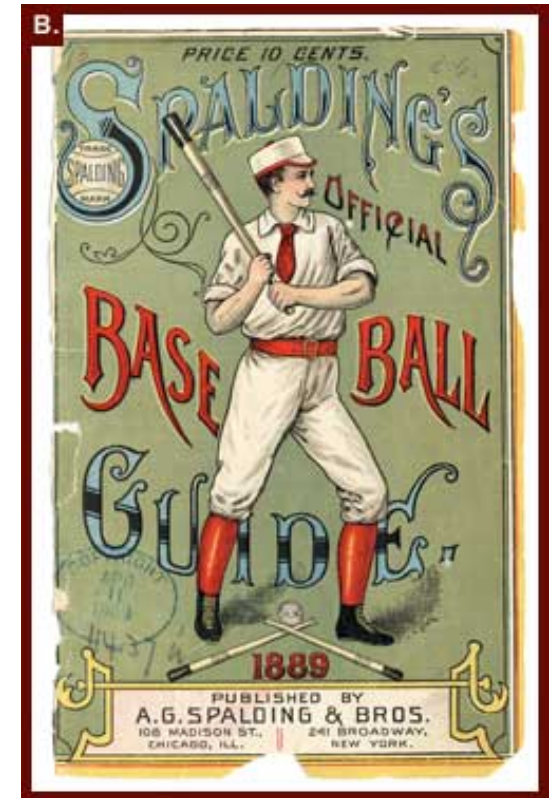
Classic statistics

Most prominent hitting statistics:

- Batting average, RBIs, and home runs

Most prominent pitching statistics:

- Wins, earned run average, strike outs



Sabermetrics

Questioned how useful traditional measures of performance, such as batting average or pitcher wins

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, **had to compile all information from box scores by hand since there was no encyclopedia that had game by game data**

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

Baseball data sets

Lahman Database: Season-by-season data

Retrosheet Game-by-Game data

Retrosheet Play-by-Play data

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

A few prominent Sabermetric publications/websites

[Society for American Baseball Research](#)

[Bill James Online](#)

[Baseball Analysts](#)

[Baseball Prospectus](#)

[Beyond the Box Score](#)

[Fan Graphs](#)

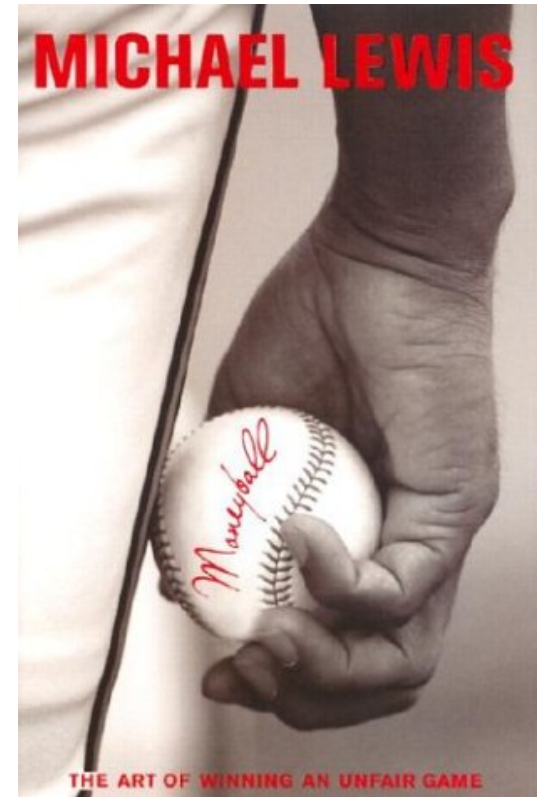
[The Hardball Times](#)

[Tango Tiger](#)

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

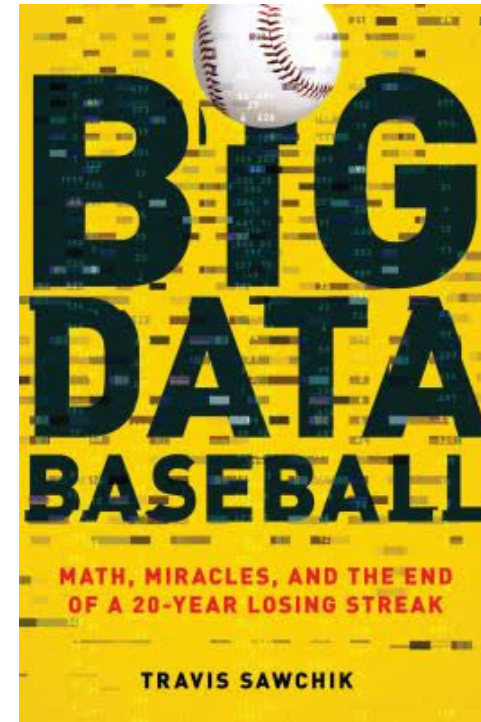
Some of the claims of the book might be exaggerated by the book had a bit impact on the expansion of major league clubs doing advanced data analyses.



Big Data Baseball

More recent sabermetric advances

- 2013 Pittsburgh Pirates



Increasing analytics to gain an edge

http://online.wsj.com/news/article_email/baseballs-science-experiment-1411135882-1MyQjAxMTA0NzE3OTIxNDkwWj

Common baseball statistics

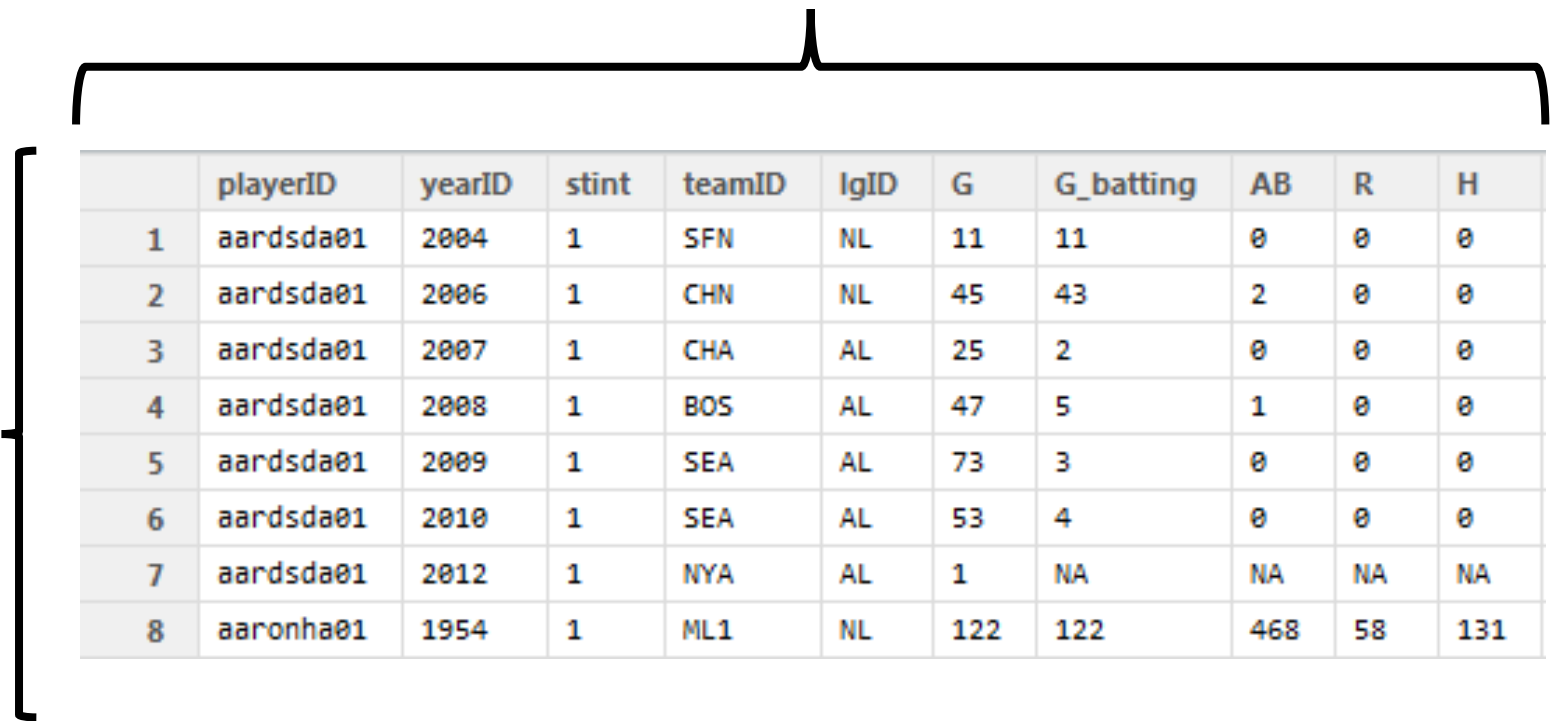
Let's look at some baseball cards...

Structured data

Lahman Database – Individual player yearly batting statistics

Variables

Cases



	playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H
1	aardsda01	2004	1	SFN	NL	11	11	0	0	0
2	aardsda01	2006	1	CHN	NL	45	43	2	0	0
3	aardsda01	2007	1	CHA	AL	25	2	0	0	0
4	aardsda01	2008	1	BOS	AL	47	5	1	0	0
5	aardsda01	2009	1	SEA	AL	73	3	0	0	0
6	aardsda01	2010	1	SEA	AL	53	4	0	0	0
7	aardsda01	2012	1	NYA	AL	1	NA	NA	NA	NA
8	aaronha01	1954	1	ML1	NL	122	122	468	58	131

Data taken from the Lahman Batting dataset

Example Dataset – Individual player yearly statistics

Variables

Cases

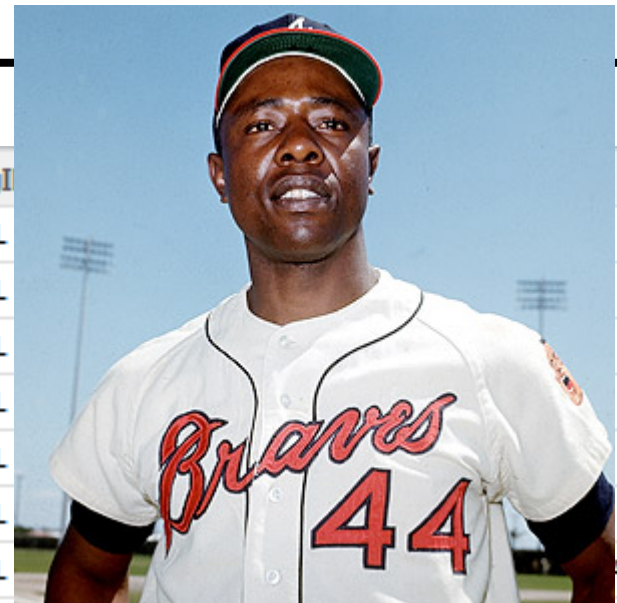
	playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H
1	aardsda01	2004	1					0	0	0
2	aardsda01	2006	1					2	0	0
3	aardsda01	2007	1					0	0	0
4	aardsda01	2008	1					1	0	0
5	aardsda01	2009	1					0	0	0
6	aardsda01	2010	1					0	0	0
7	aardsda01	2012	1					NA	NA	NA
8	aaronha01	1954	1	ML1	NL	122	122	468	58	131

Example Dataset – Individual player yearly statistics

Variables

Cases

	playerID	yearID	stint	teamID	lgID					
1	aardsda01	2004	1	SFN	NL					
2	aardsda01	2006	1	CHN	NL					
3	aardsda01	2007	1	CHA	AL					
4	aardsda01	2008	1	BOS	AL					
5	aardsda01	2009	1	SEA	AL					
6	aardsda01	2010	1	SEA	AL					
7	aardsda01	2012	1	NYA	AL					
8	aaronha01	1954	1	ML1	NL	122	122	468	58	131



Categorical and Quantitative Variables

Categorical Variable

Quantitative Variable

Cases

	playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H
1	aardsda01	2004	1	SFN	NL	11	11	0	0	0
2	aardsda01	2006	1	CHN	NL	45	43	2	0	0
3	aardsda01	2007	1	CHA	AL	25	2	0	0	0
4	aardsda01	2008	1	BOS	AL	47	5	1	0	0
5	aardsda01	2009	1	SEA	AL	73	3	0	0	0
6	aardsda01	2010	1	SEA	AL	53	4	0	0	0
7	aardsda01	2012	1	NYA	AL	1	NA	NA	NA	NA
8	aaronha01	1954	1	ML1	NL	122	122	468	58	131

Explanatory and Response Variables

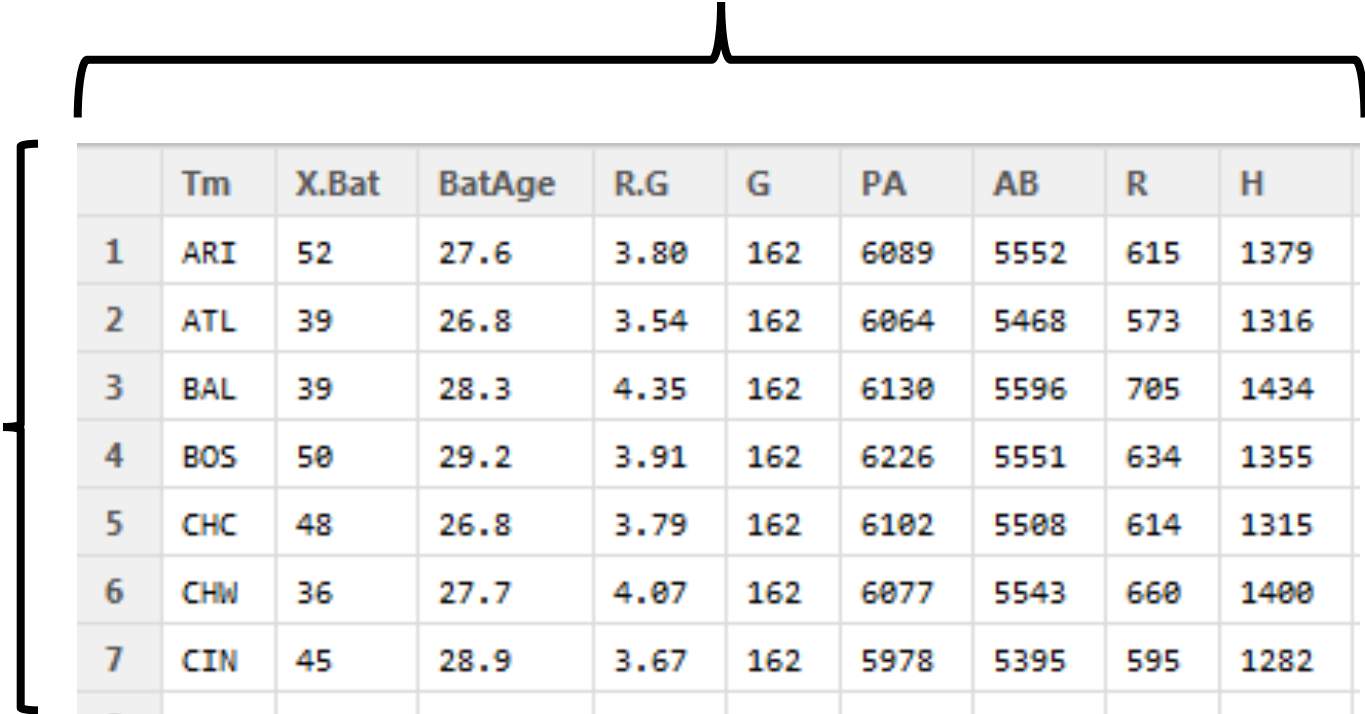
Sometimes we use one variable (**the explanatory variable**) to understand/predict another variable (**the response variable**)

	playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H
1	aardsda01	2004	1	SFN	NL	11	11	0	0	0
2	aardsda01	2006	1	CHN	NL	45	43	2	0	0
3	aardsda01	2007	1	CHA	AL	25	2	0	0	0
4	aardsda01	2008	1	BOS	AL	47	5	1	0	0
5	aardsda01	2009	1	SEA	AL	73	3	0	0	0
6	aardsda01	2010	1	SEA	AL	53	4	0	0	0
7	aardsda01	2012	1	NYA	AL	1	NA	NA	NA	NA
8	aaronha01	1954	1	ML1	NL	122	122	468	58	131

Another Dataset – 2014 Team statistics

Variables

Cases



	Tm	X.Bat	BatAge	R.G	G	PA	AB	R	H
1	ARI	52	27.6	3.80	162	6089	5552	615	1379
2	ATL	39	26.8	3.54	162	6064	5468	573	1316
3	BAL	39	28.3	4.35	162	6130	5596	705	1434
4	BOS	50	29.2	3.91	162	6226	5551	634	1355
5	CHC	48	26.8	3.79	162	6102	5508	614	1315
6	CHW	36	27.7	4.07	162	6077	5543	660	1400
7	CIN	45	28.9	3.67	162	5978	5395	595	1282
8	CLE	43	28.5	4.13	162	6222	5575	669	1411

Next class – examining a single batch of data

Fill out survey online! (link is on Moodle)

Read chapter 1 of Big Data Baseball and post a quote and reaction to the Moodle forum by midnight on Sunday

Read chapter 1 of Teaching Statistics Using Baseball