

Baseball statistics and an introduction to R

Overview

Discussion of Big Data Baseball

Watch half an inning of the 2014 All-star game

Review of structured data and classic baseball statistics

Introduction to R!

Discussion of Big Data Baseball chapter 1

Statistics can get us beyond what we can “see” if we trust them (Phillip)

- Should we just trust the analyses, what about players who have “heart”? (James)
- How do we maximize our decisions with both analysis and human decisions (Campbell)

How to find and quantify the relevant data? (on-base percentage, etc.) (Henne)

- New/different statistics and analyzes can give powerful new insights (Aodhan)
- New computational systems can shed new insights (Kefentse)
- What is the value of different hit types, e.g., singles vs. home runs? (Julia)

Why didn't anyone realize that equidistant spacing of defensive players was suboptimal? (Helen)

- Yes, the defensive changes will be explained more in future chapters (Ian)

How can we make changes that are within our reach? (Maddie)

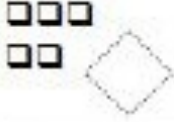
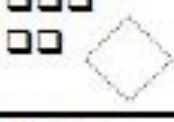
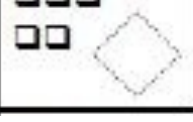
- With only a \$15 million budget (Sheyla)
- And taking on challenging situations (Matt)

Rules of the game are the same, but the way players are acquired has changed (Christopher)

2014 All-star game

National				American		
Order	Player	Position		Order	Player	Position
1	Andrew McCutchen	CF		1	Derek Jeter	SS
2	Yasiel Puig	RF		2	Mike Trout	LF
3	Troy Tulowitzki	SS		3	Robinson Canó	2B
4	Paul Goldschmidt	1B		4	Miguel Cabrera	1B
5	Giancarlo Stanton	DH		5	José Bautista	RF
6	Aramis Ramírez	3B		6	Nelson Cruz	DH
7	Chase Utley	2B		7	Adam Jones	CF
8	Jonathan Lucroy	C		8	Josh Donaldson	3B
9	Carlos Gómez	LF		9	Salvador Pérez	C
	Adam Wainwright	P			Félix Hernández	P

Score card

#	Player	Pos	1	2
	Henry			
	Clarke			
	Navi			
	Terra			
	Gabe			
	Jake			

sstatistics and structured data

statistics: a numerical summary of data

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

Describing and summarizing data

statistics that are used to summarize a data set (sample of data) are called **descriptive statistics**

Examples:

- Maximum value in the data set
- Minimum value in the data set
- Mean value of the data set

Common baseball descriptive **s**tatistics

G = games

- Number of games a player participated in (out of 162 games in a season)

AB = at bats

- Number of times a batter was hitting and either got a hit or got out (does not include walks or reaching base on an error)

R = runs

- Number of runs the player scored

H = hit

- Number of times a player hit the ball on got on base or hit a home run (sum of 1B, 2B, 3B, HR)

Common baseball **s**tatistics

BB = base on balls (walks)

- Number of times a player got on base do to the pitcher throwing 4 balls

RBI = Runs batted in

- How many runs scored as a result of a player getting a hit

SB = stolen bases

- Number of times a runner advanced by 'stealing a base'

Common derived baseball statistics

AVG= batting average

- $\text{Hits}/(\text{At bats}) = H/AB = (1B + 2B + 3B + HR)/AB$

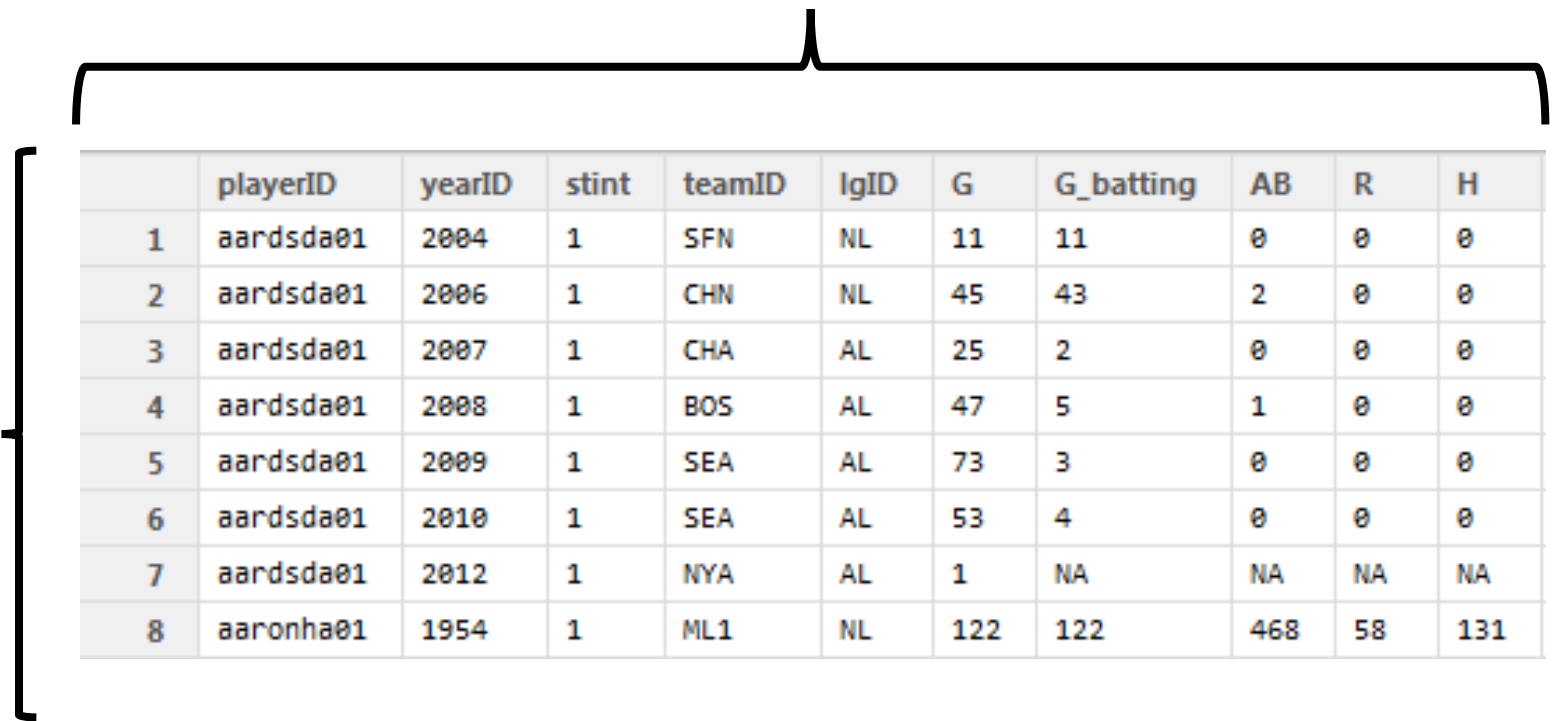
SLG = slugging percentage

- $(1 * 1B + 2 * 2B + 3 * 3B + 4 * 4B) / AB$

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

Categorical and Quantitative Variables

Categorical Variable

Quantitative Variable

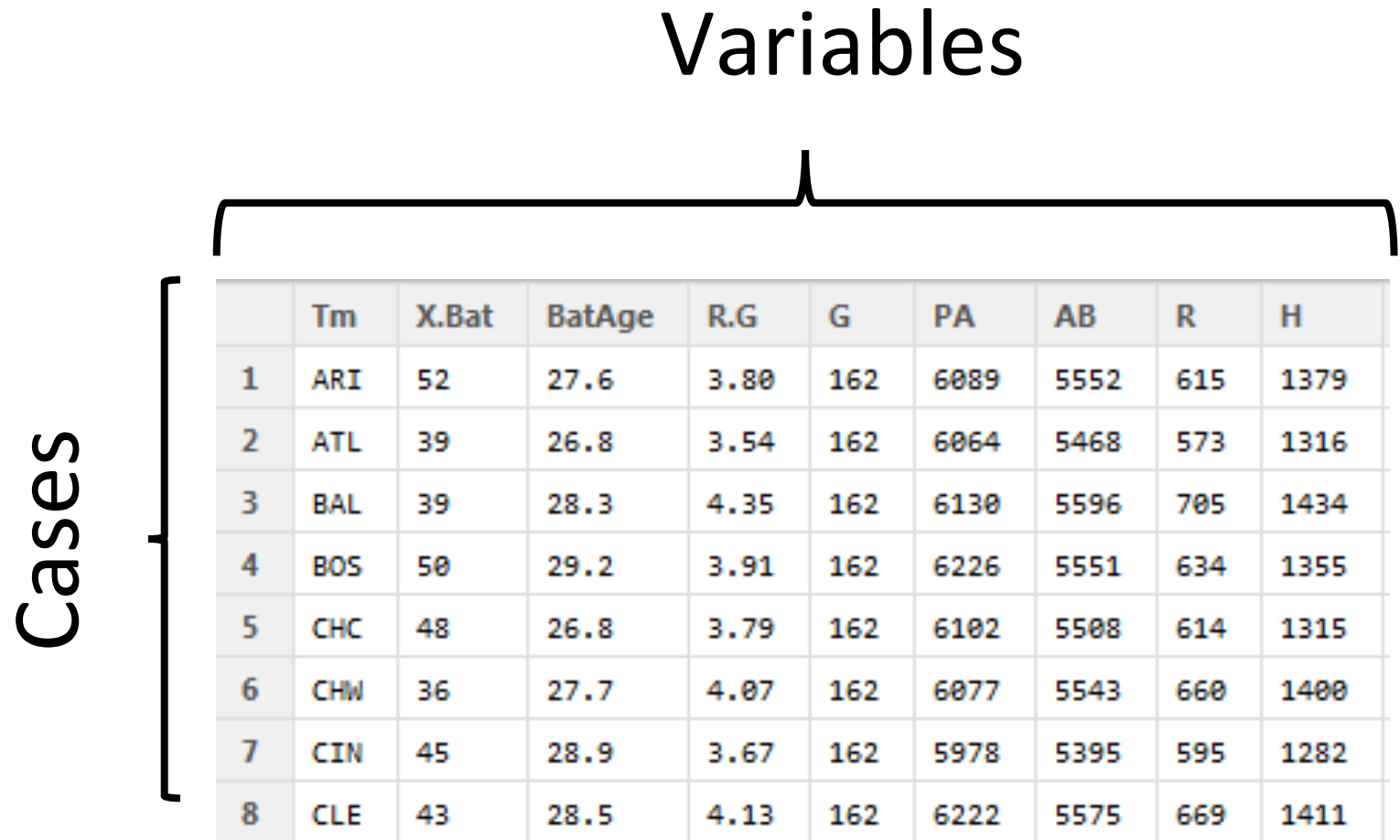
Cases

	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

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

A Question



Q: What programming language do the pirates use?

A: Arrrr

Q: Worst joke of the semester?

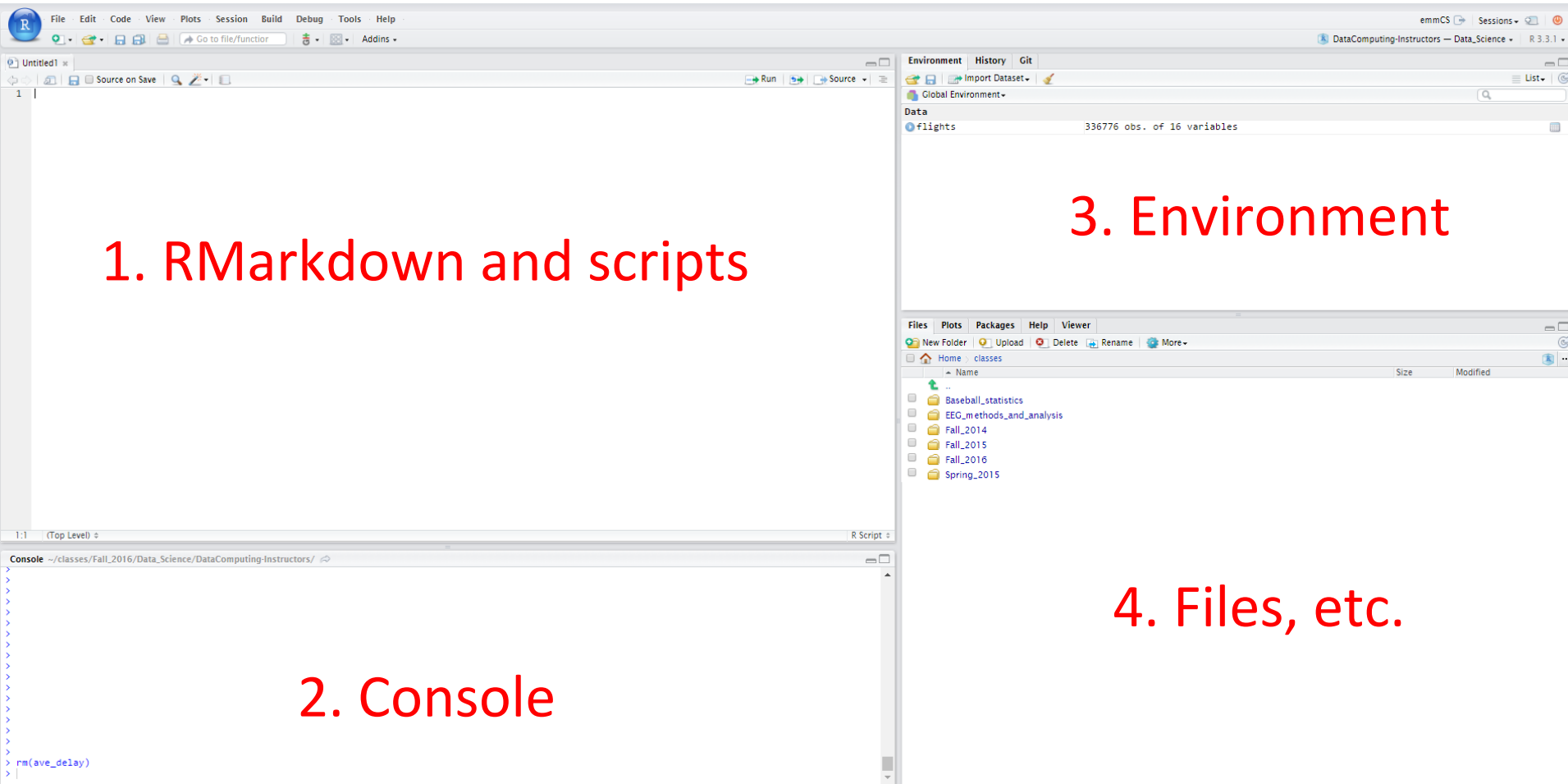
A: Wait and see...

Basics of R

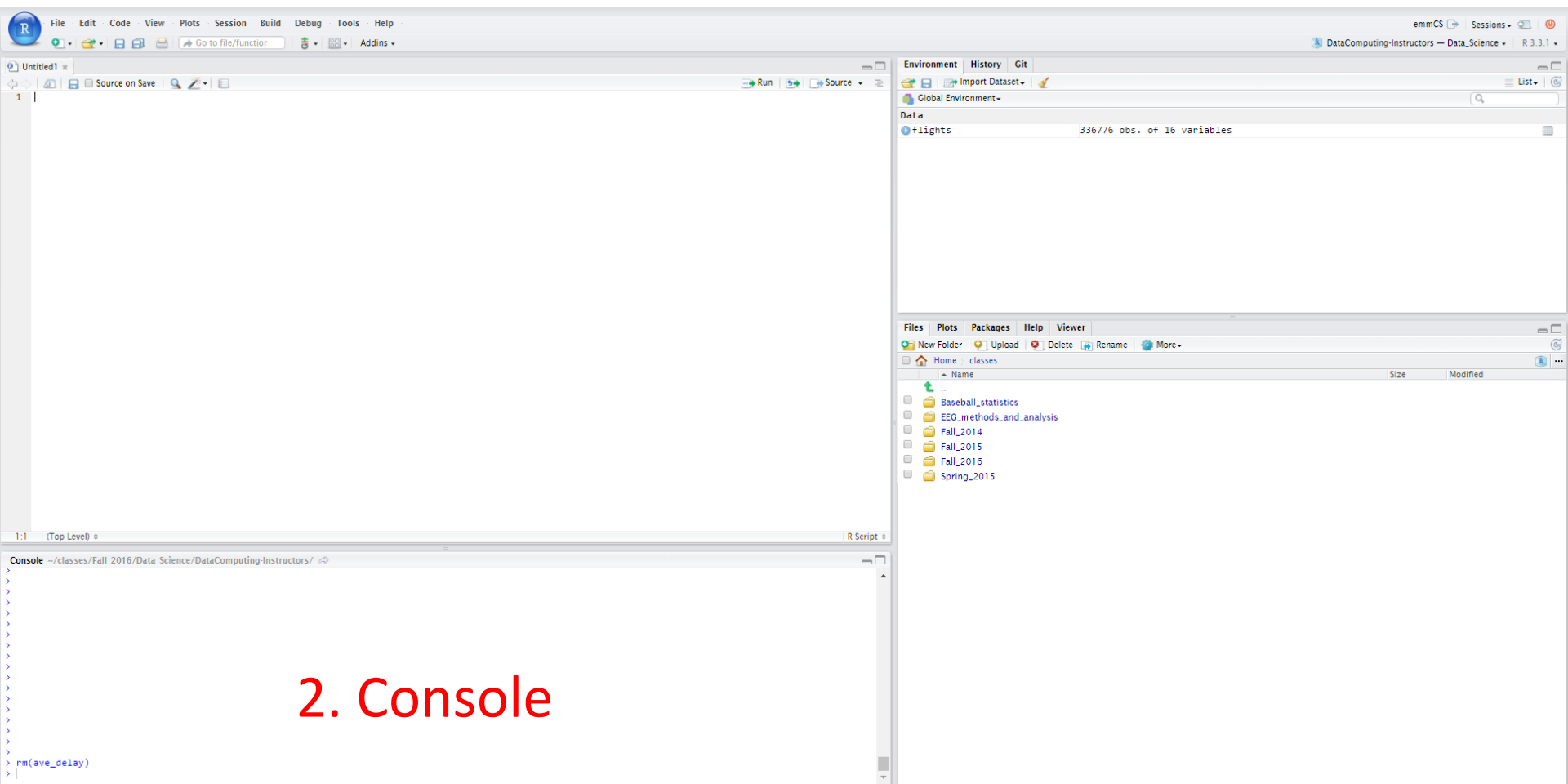
Everyone log on to: <https://asterius.hampshire.edu/>

Create a new script to keep notes about your work

RStudio layout



RStudio layout



2. Console

R as a calculator

> 2 + 2

> 7 * 5

R Basics

Arithmetic:

```
> 2 + 2
```

```
> 7 * 5
```

Assignment:

```
> a <- 4
```

```
> b <- 7
```

```
> D <- a + b
```

```
> D
```

```
[1] 11
```

Number journey...

Number journey

```
> a <- 7
```

```
> b <- 52
```

```
> d <- a * b
```

```
> d
```

```
[1] 364
```

Character strings and booleans

```
> a <- 7
```

```
> s <- "hello everyone"
```

```
> b <- TRUE
```

```
> class(a)
```

```
[1] numeric
```

```
> class(s)
```

```
[1] character
```

Functions

Functions use parenthesis: functionName(x)

```
> sqrt(49)
```

```
> tolower("HELLO everyone")
```

To get help

```
> ? sqrt
```

One can add comments to your code

```
> sqrt(49) # this takes the square root of 49
```

Getting help

You can get help about a function in R using the ? command.

```
> ? sqrt
```

Vectors

Vectors are ordered sequences of numbers or letters

The `c()` function is used to create vectors

```
> v <- c(5, 232, 5, 543)
```

One can access elements of a vector using square brackets `[]`

```
> v[3]      # what will the answer be?
```

Works with strings too

```
> z <- c("a", "b", "c", "d")
```

```
> z[3]
```

Can add names to vector elements

```
> names(v) <- c("first", "second", "third", "fourth")
```

Question?



Q: What kind of grades did the Pirates get in Statistics class?

A: High Seas

Q: Worst joke of the semester?

A: Stay tuned...

Data types: data frames

Data Frames are collections of vectors of that same length.

- Each vector can have a different type of data

	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	5064	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	5077	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

Let's look at a data frame

Load a function I wrote into R by typing:

```
source('/home/shared/baseball_stats_2017/  
baseball_class_functions.R')
```

If you load this correctly you should have a function in your Global Environment called `get.Lahman.batting.data()`

Let's look at a data frame

	playerID	nameFirst	nameLast	yearID	stint	teamID	lgID	G	AB
1	shoppke01	Kelly	Shoppach	2005	1	BOS	AL	9	15
2	shoppke01	Kelly	Shoppach	2006	1	CLE	AL	41	110
3	shoppke01	Kelly	Shoppach	2007	1	CLE	AL	59	161
4	shoppke01	Kelly	Shoppach	2008	1	CLE	AL	112	352
5	shoppke01	Kelly	Shoppach	2009	1	CLE	AL	89	271
6	shoppke01	Kelly	Shoppach	2010	1	TBA	AL	63	158
7	shoppke01	Kelly	Shoppach	2011	1	TBA	AL	87	221
8	shoppke01	Kelly	Shoppach	2012	2	NYN	NL	28	79
9	shoppke01	Kelly	Shoppach	2012	1	BOS	AL	48	140
10	shoppke01	Kelly	Shoppach	2013	2	CLE	AL	1	2
11	shoppke01	Kelly	Shoppach	2013	1	SEA	AL	35	107



Use this function to get batting data on a specific player:

```
> card.data <- get.Lahman.batting.data("Kelly", "Shoppach")  
> View(card.data)
```

Let's look at a data frame

	playerID	nameFirst	nameLast	yearID	stint	teamID	lgID	G	AB
1	shoppke01	Kelly	Shoppach	2005	1	BOS	AL	9	15
2	shoppke01	Kelly	Shoppach	2006	1	CLE	AL	41	110
3	shoppke01	Kelly	Shoppach	2007	1	CLE	AL	59	161
4	shoppke01	Kelly	Shoppach	2008	1	CLE	AL	112	352
5	shoppke01	Kelly	Shoppach	2009	1	CLE	AL	89	271
6	shoppke01	Kelly	Shoppach	2010	1	TBA	AL	63	158
7	shoppke01	Kelly	Shoppach	2011	1	TBA	AL	87	221
8	shoppke01	Kelly	Shoppach	2012	2	NYN	NL	28	79
9	shoppke01	Kelly	Shoppach	2012	1	BOS	AL	48	140
10	shoppke01	Kelly	Shoppach	2013	2	CLE	AL	1	2
11	shoppke01	Kelly	Shoppach	2013	1	SEA	AL	35	107



Getting number of games (G) Kelly played each season:

```
> card.data$G
```

```
[1] 9 41 59 112 89 63 87 28 48 35 1
```

Computing statistics

One compute statistics on vectors (columns of a data frame)

```
> sum(card.data$G)  
[1] 572
```

Or we can assign vectors in a data frame to an object

```
> games <- card.data$G  
> games
```

Practice R with DataCamp!

Try chapters 1 and 2 on the introduction to R DataCamp tutorial

<https://www.datacamp.com/courses/free-introduction-to-r>

Read chapter 2 of Big Data Baseball and post a quote and reaction by midnight on Wednesday