

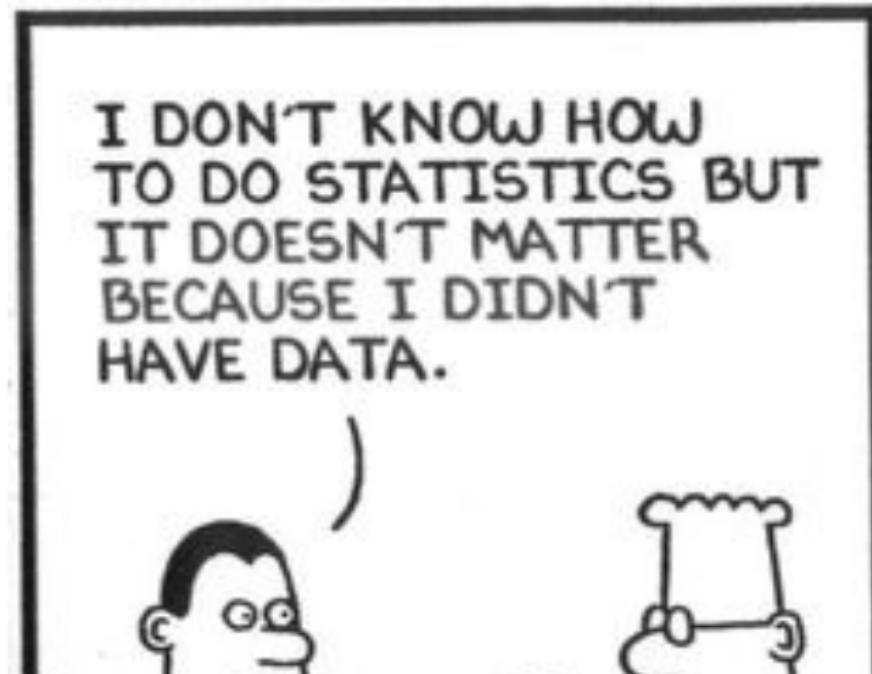
431 Class 02

Thomas E. Love

2017-08-31

Today's Agenda

- ① Administration
- ② The Class 1 Survey and How To Ask Questions
- ③ Using R, R Studio and R Markdown



TA Office Hours start Tuesday 2017-09-05

Tuesdays and Thursdays

- 11:30 - 12:30
- 2:30 - 4:30
- 5:45 - 6:45

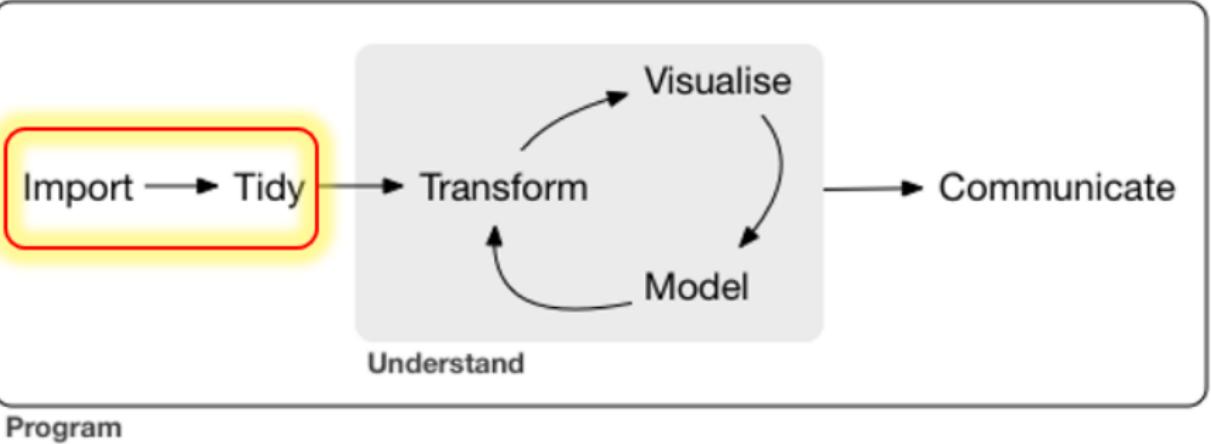
Fridays

- 4:30 - 5:30

TA office hours are held in Wood WG-56 (Computing Lab) or WG-67 (Student Lounge), so be sure to look in both places.

Contact us at 431-help@case.edu

Our web site: <https://github.com/thomaselove/431>



Types of Data (and see the Course Notes, section 4.3)

Data can be **quantitative (numerical)** or **qualitative (categorical)**

- **Quantitative**

- Variables recorded in numbers that we use as numbers.
- All quantitative variables must have units of measurement.
- Can break into *continuous* (may take any value in a range) or *discrete* (limited set of potential values.)
 - Height is certainly continuous as a concept, but how precise is our ruler?
 - Piano vs. Violin
- (less common) *interval* (equal distances between values, but zero point is arbitrary) as compared to *ratio* variables (a meaningful zero point.)
 - Is *weight* an interval or ratio variable? How about *IQ*?
- Taking a mean or median is a reasonable idea.

Types of Data (and see Part A Notes, section 2)

Data can be **quantitative (numerical)** or **qualitative (categorical)**

- Qualitative

- Variables consisting of names of categories.
- Each possible value is a code for a category (could use numerical or non-numerical codes.)
 - *Binary* categorical variables (two categories, often labeled 1 or 0)
 - *Multi-categorical* variables (usually taken to be 3+ categories)
- Also, *nominal* (no underlying order) or *ordinal* (categories are ordered.)
 - How is your overall health? (Excellent, Very Good, Good, Fair, Poor)
 - Which candidate would you vote for if the election were held today?
 - Did this patient receive this procedure?

Day 1 Survey Handout

431 First Day Survey (15 Questions)

Please introduce yourself to someone you do not know, ask them these 15 questions, and record their answers on this sheet. At the same time, provide your partner with your answers so they can record your responses on their sheet. Do not place any names on this sheet so that the responses will remain anonymous. Thank you!

1. What is your sex? (Male or Female) _____

2. Is English your most comfortable language? (Yes or No) _____

3. Fill in the number that best describes your answer to this question:

| Has statistical thinking been important in your life so far? | | | | | | |
|--|-----------------------|-----------------------|------------------------|---|---|---|
| Not at all important | Slightly important | Somewhat important | Extremely important | | | |
| ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ |

4. How old (in years) do you think Professor Love is? _____ years.

5. Do you smoke? Fill in the appropriate circle:

| No | I used to. | Yes. |
|------------|---------------|--------|
| Non-Smoker | Former Smoker | Smoker |

6. Please indicate which hand you use for each of the following activities by putting a + in the appropriate column, or ++ if you would never use the other hand for that activity. If, in any case, you really are indifferent, put + in both columns.

| Task | Left | Right |
|--|------|-------|
| Writing | | |
| Drawing | | |
| Throwing | | |
| Scissors | | |
| Toothbrush | | |
| Knife (without fork) | | |
| Spoon | | |
| Broom (upper hand) | | |
| Striking match (hand that holds the match) | | |
| Opening box (hand that holds the lid) | | |
| Total Count of +s: | | |

$$\text{Right} - \text{Left} = \underline{\hspace{2cm}} \quad \text{Right} + \text{Left} = \underline{\hspace{2cm}} \quad \frac{\text{Right} - \text{Left}}{\text{Right} + \text{Left}} = \underline{\hspace{2cm}}$$

August 30, 2016

431 First Day Survey (15 Questions)

7. How important do you think statistics will be in your *future* career?

| Not at all important | Slightly important | Somewhat important | Extremely important |
|-------------------------|-----------------------|-----------------------|------------------------|
| ① | ② | ③ | ④ |

8. How much did you pay for your most recent haircut? (in \$): _____

Please indicate your agreement with the following statements:

| | Strongly Disagree | 2 | 3 | 4 | 5 |
|---|----------------------|---|---|---|---|
| 9. I prefer to learn from lectures than to learn from activities. | 1 | | | | |
| 10. I prefer to work on projects alone than in a team. | 1 | 2 | 3 | 4 | 5 |

11. What is your height (indicate units of measurement): _____

12. Use the ruler provided on the side of this page to measure the span of your right hand (distance from the thumb to the little finger when your fingers are spread apart): _____ cm.

13. What is your favorite color? _____

14. How many hours did you sleep last night? _____ hours.

15. Record your pulse by counting the beats of your heart for 30 seconds, then doubling the result: _____ beats/minute.

Evaluating some Day 1 Survey variables

- ① Do you **smoke**? (1 = Non-Smoker, 2 = Former Smoker, 3 = Smoker)
- ② How much did you pay for your most recent **haircut**? (in \$)
- ③ What is your favorite **color**?
- ④ How many hours did you **sleep** last night?
- ⑤ Has statistical thinking been important in your life? (1 = Not at all important to 7 = Extremely important)

Are these quantitative or qualitative?

- If quantitative, are they *discrete* or *continuous*? Do they have a meaningful *zero point*?
- If qualitative, how many categories? *Nominal* or *ordinal*?

Day 1 Survey

01

| | A | B | C | D | E | F | G | H | I | J |
|-----|---------|-----|---------|-----------|----------|-------|--------|---------|----------|-------------|
| 1 | student | sex | english | statsofar | ageguess | smoke | h.left | h.right | handedne | statfutureh |
| 157 | 201701 | | | | | | | | | |
| 158 | 201702 | | | | | | | | | |
| 159 | 201703 | | | | | | | | | |
| 160 | 201704 | | | | | | | | | |
| 161 | 201705 | | | | | | | | | |
| 162 | 201706 | | | | | | | | | |
| 163 | 201707 | | | | | | | | | |
| 164 | 201708 | | | | | | | | | |

gle cm to in

All Shopping Books News Images More

About 939,000,000 results (0.88 seconds)

Length

168 Centimeter = 66.1417 Inch

Day 1 Survey

- 48 people completed it Tuesday
- Another 64 did it in 2016, 49 did in 2015, and another 42 in 2014.

Question 1

About how many of those 203 surveys caused *no problems* in recording responses?

Day 1 Survey Handout

431 First Day Survey (15 Questions)

Please introduce yourself to someone you do not know, ask them these 15 questions, and record their answers on this sheet. At the same time, provide your partner with your answers so they can record your responses on their sheet. Do not place any names on this sheet so that the responses will remain anonymous. Thank you!

1. What is your sex? (Male or Female) _____

2. Is English your most comfortable language? (Yes or No) _____

3. Fill in the number that best describes your answer to this question:

| Has statistical thinking been important in your life so far? | | | | | | |
|--|-----------------------|-----------------------|------------------------|---|---|---|
| Not at all important | Slightly important | Somewhat important | Extremely important | | | |
| ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ |

4. How old (in years) do you think Professor Love is? _____ years.

5. Do you smoke? Fill in the appropriate circle:

| No | I used to. | Yes. |
|------------|---------------|--------|
| Non-Smoker | Former Smoker | Smoker |

6. Please indicate which hand you use for each of the following activities by putting a + in the appropriate column, or ++ if you would never use the other hand for that activity. If, in any case, you really are indifferent, put + in both columns.

| Task | Left | Right |
|--|------|-------|
| Writing | | |
| Drawing | | |
| Throwing | | |
| Scissors | | |
| Toothbrush | | |
| Knife (without fork) | | |
| Spoon | | |
| Broom (upper hand) | | |
| Striking match (hand that holds the match) | | |
| Opening box (hand that holds the lid) | | |
| Total Count of +s: | | |

$$\text{Right} - \text{Left} = \underline{\hspace{2cm}} \quad \text{Right} + \text{Left} = \underline{\hspace{2cm}} \quad \frac{\text{Right} - \text{Left}}{\text{Right} + \text{Left}} = \underline{\hspace{2cm}}$$

August 30, 2016

431 First Day Survey (15 Questions)

7. How important do you think statistics will be in your *future* career?

| Not at all important | Slightly important | Somewhat important | Extremely important |
|-------------------------|-----------------------|-----------------------|------------------------|
| ① | ② | ③ | ④ |

8. How much did you pay for your most recent haircut? (in \$): _____

Please indicate your agreement with the following statements:

| | Strongly Disagree | 2 | 3 | 4 | 5 |
|---|----------------------|---|---|---|---|
| 9. I prefer to learn from lectures than to learn from activities. | 1 | | | | |
| 10. I prefer to work on projects alone than in a team. | 1 | 2 | 3 | 4 | 5 |

11. What is your height (indicate units of measurement): _____

12. Use the ruler provided on the side of this page to measure the span of your right hand (distance from the thumb to the little finger when your fingers are spread apart): _____ cm.

13. What is your favorite color?: _____

14. How many hours did you sleep last night? _____ hours.

15. Record your pulse by counting the beats of your heart for 30 seconds, then doubling the result: _____ beats/minute.



August 30, 2016

The 15 Survey Items

| # | Topic | # | Topic |
|----|--------------|-----|-----------------------|
| Q1 | sex | Q9 | lectures v activities |
| Q2 | english | Q10 | projects alone |
| Q3 | stats so far | Q11 | height |
| Q4 | guess TL age | Q12 | hand span |
| Q5 | smoke | Q13 | color |
| Q6 | handedness | Q14 | sleep |
| Q7 | stats future | Q15 | pulse rate |
| Q8 | haircut | - | - |

Question 1

About how many of those 203 surveys caused *no problems* in recording responses?

- Guesses?

Question 1

About how many of those 203 surveys caused *no problems* in recording responses?

- Guesses?
- 74/203 (38%)

Question 1

About how many of those 203 surveys caused *no problems* in recording responses?

- Guesses?
- 74/203 (38%)
- 20 of the 48 surveys turned in Tuesday had **no** problems (42%)

Guess My Age

4. How old (in years) do you think Professor Love is?

early fifties years

4. How old (in years) do you think Professor Love is?

late 50's years.

English best language?

2. Is English your *most comfortable* language? (Yes or No)

English

TEL Decision: Yes

1. What is your *gender*? (Male or Female)

(Male or Female)

2. Is English your *most comfortable* language? (Yes or No)

(Yes or No)

TEL Decision: NA

Is English your *most comfortable* language? (Yes or No)

may be

TEL decision: NA

Favorite color

13. What is your favorite color? depends

NA

13. What is your favorite color? Brown

orange

13. What is your favorite color? Blue / Brown

13. What is your favorite color? none

Following the Rules?

15. Record your pulse by counting the beats of your heart for 30 seconds, then doubling the result:

75

beats/minute.

2017 pulse responses, sorted ($n = 48$)

| | | | | | |
|----|----|----|----|-----|-----|
| 50 | 54 | 54 | 57 | 58 | 60 |
| 62 | 64 | 64 | 68 | 68 | 68 |
| 70 | 70 | 72 | 72 | 74 | 74 |
| 74 | 75 | 75 | 75 | 76 | 78 |
| 78 | 80 | 80 | 80 | 80 | 80 |
| 80 | 82 | 82 | 84 | 84 | 86 |
| 86 | 86 | 88 | 88 | 91 | 94 |
| 96 | 98 | 98 | 98 | 100 | 100 |

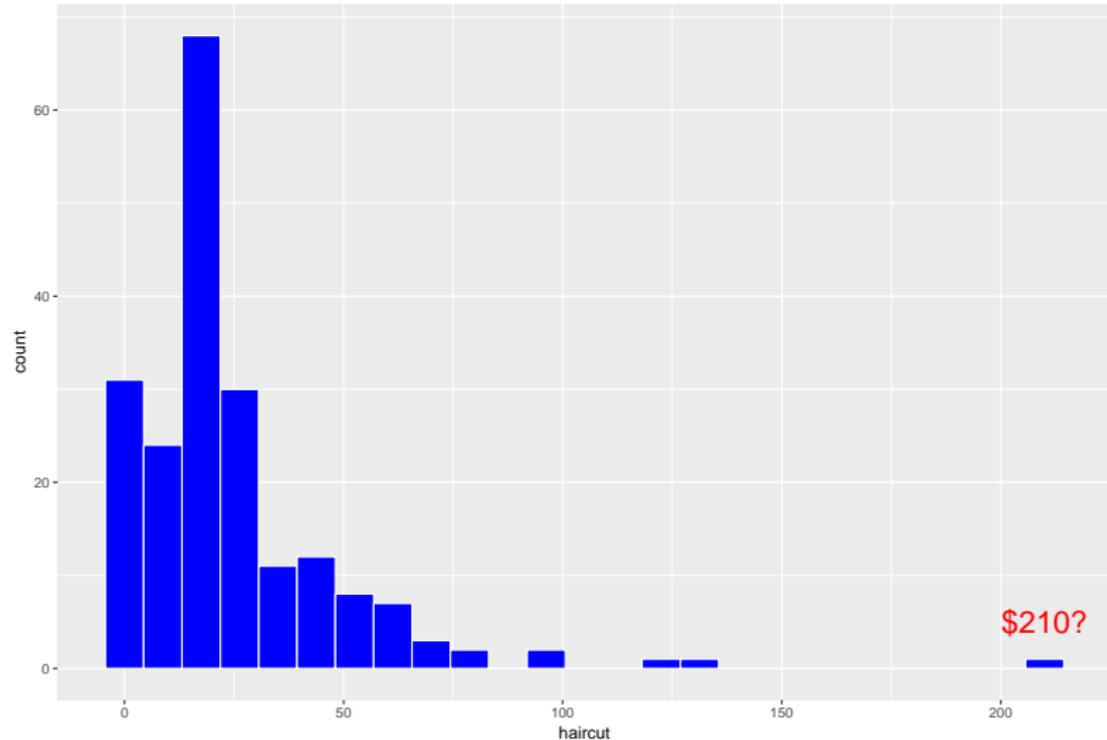
Stem and Leaf: Haircut \$ (Thanks, John Tukey)

The decimal point is 1 digit(s) to the right of the |

| | | |
|----|--|--|
| 0 | | 000000000000000000000000000000003555555677899 |
| 1 | | 000022222224444445555555555555555677888889 |
| 2 | | 000000000000000000000000000000012235555555556688 |
| 3 | | 0000000000022222555555 |
| 4 | | 000000000558 |
| 5 | | 00000555 |
| 6 | | 0000000 |
| 7 | | 0005 |
| 8 | | 0 |
| 9 | | |
| 10 | | 00 |
| 11 | | |
| 12 | | 0 |
| 13 | | 0 |
| 14 | | |
| 15 | | |
| 16 | | |
| 17 | | |
| 18 | | |
| 19 | | |
| 20 | | |
| 21 | | 0 |

Haircut Histogram

Histogram of 155 Haircut Prices from Survey 1 data



Hand Span (in cm)

12. Use the ruler provided on the side of this page to measure the span of your right hand (distance from the thumb to the little finger when your fingers are spread apart): 26 cm.

Hand Span Stem-and-Leaf Plot (HT: John Tukey)

The decimal point is at the |

| | | |
|----|--|--|
| 8 | | 0 |
| 9 | | 55 |
| 10 | | 00 |
| 11 | | 0 |
| 12 | | |
| 13 | | 5 |
| 14 | | |
| 15 | | |
| 16 | | 00008 |
| 17 | | 0000055 |
| 18 | | 000000000000055555555555 |
| 19 | | 0000000000000002235555555557 |
| 20 | | 00000000000000000000000000002345555555578 |
| 21 | | 000000000000000000000000000045555555555577 |
| 22 | | 00000000000000000000000000002555568 |
| 23 | | 000000000055 |
| 24 | | 000001 |
| 25 | | 000 |

Eight Items had just a few problems (< 15 each)

| # | Topic | # | Topic |
|----|---------------------|-----|-----------------------|
| - | sex | - | lectures v activities |
| Q2 | <i>english</i> | Q10 | <i>projects alone</i> |
| - | stats so far | - | height |
| Q4 | <i>guess TL age</i> | Q12 | <i>hand span</i> |
| - | smoke | Q13 | <i>color</i> |
| - | handedness | Q14 | <i>sleep</i> |
| - | stats future | Q15 | <i>pulse rate</i> |
| Q8 | <i>haircut</i> | - | - |

Question 2

Of the remaining 7 (sex, stats so far, smoke, handedness, stats future, lectures vs activities, height), 5 had no real problems, and two were messy. Which two?

Height

- ii. What is your height (indicate units of measurement): 5'4 (inches)
- ii. What is your height (indicate units of measurement): 6'0
- ii. What is your height (indicate units of measurement): 5'2
- ii. What is your height (indicate units of measurement): 5'7"
- ii. What is your height (indicate units of measurement): 5'5

Handedness Scale (2014-15 version)

6. Please indicate which hand you use for each of the following activities by putting a + in the appropriate column, or ++ if you would never use the other hand for that activity. If in any case you really are indifferent, put + in both columns.

| Task | Left | Right |
|--|------|-------|
| Writing | | ✓ |
| Drawing | | ✓ |
| Throwing | | ✓ |
| Scissors | | ✓ |
| Toothbrush | ✓ | |
| Knife (without fork) | ✓ | |
| Spoon | ✓ | ✓ |
| Broom (upper hand) | | ✓ |
| Striking match (hand that holds the match) | | ✓ |
| Opening box (hand that holds the lid) | | ✓ |
| Total Count of +s: | 3 | 8 |

Handedness Scale (2016-17 version)

6. Please indicate which hand you use for each of the following activities by putting a + in the appropriate column, or ++ if you would *never* use the other hand for that activity. If, in any case, you really are indifferent, put + in both columns.

| Task | Left | Right |
|--|------|-------|
| Writing | ++ | + |
| Drawing | ++ | + |
| Throwing | ++ | + |
| Scissors | ++ | + |
| Toothbrush | ++ | + |
| Knife (without fork) | ++ | + |
| Spoon | ++ | + |
| Broom (upper hand) | ++ | ++ |
| Striking match (hand that holds the match) | ++ | + |
| Opening box (hand that holds the lid) | ++ | + |
| Total Count of +s: | 20 | 11 |

Handedness Scale (2016-17 version)

6. Please indicate which hand you use for each of the following activities by putting a + in the appropriate column, or ++ if you would *never* use the other hand for that activity. If, in any case, you really are indifferent, put + in both columns.

| Task | Left | Right |
|--|------|-------|
| Writing | ++ | + |
| Drawing | ++ | + |
| Throwing | ++ | + |
| Scissors | ++ | + |
| Toothbrush | ++ | + |
| Knife (without fork) | ++ | + |
| Spoon | ++ | + |
| Broom (upper hand) | ++ | + |
| Striking match (hand that holds the match) | ++ | + |
| Opening box (hand that holds the lid) | ++ | + |
| Total Count of +s: | | (10) |

$$\text{Right} - \text{Left} = 10$$

$$\text{Right} + \text{Left} = 10$$

$$\frac{\text{Right} - \text{Left}}{\text{Right} + \text{Left}} = \frac{10}{10} = 1$$

Garbage in, garbage out . . .

www.VADLO.com



“Data don’t make any sense,
we will have to resort to statistics.”

Today's Steps

We assume you were able to follow the software installation instructions.

- ① Get data from our web site and store it in a new directory on your machine.
- ② Open R Studio and start a Project, selecting the new directory for the Project.
- ③ Open and set up an R Markdown file to do the work.
- ④ Write code in your R Markdown file to
 - load the data in R
 - load up the tidyverse suite of R packages you'll need
 - look at the data
 - visualize the data
 - summarize the data numerically
 - compare the data
 - build a model for the data
- ⑤ Compile your R Markdown file to generate an HTML or Word document.

Analyzing the Index Card Guesses of My Age

47 students turned in an index card, meant to contain both a first and a second guess of my age.

Step 1: Get the Data

- I've stored the data in a .csv file on our web site, for instance, at

https://github.com/THOMASELOVE/431slides/tree/master/class_02

- We'll grab just that data file, for now, by clicking on it, selecting Raw, and saving the resulting **age-love-2017.csv** file to our computer.
- Specifically, we'll save it to a new directory called **431class2**.

Class 2 Github page - select the age-love-2017.csv file

The screenshot shows a web browser window with the GitHub URL https://github.com/THOMASELOVE/431slides/tree/master/class_02. The page displays a repository named "431slides / class_02". The "Code" tab is selected. A commit message from "THOMASELOVE" is visible, stating "Delete age-love.csv". Below the commit, there is a file named "age-love-2017.csv". The browser's address bar shows the full URL, and the top navigation bar includes links for Pull requests, Issues, Marketplace, and Explore.

Right-click Raw to download (just) this file, into a 431class2 directory, please.

THOMASELOVE / 431slides

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Settings Insights

Branch: master 431slides / class_02 / age-love-2017.csv Find file Copy path

THOMASELOVE Add files via upload 656d3a1 23 minutes ago

1 contributor

49 lines (48 sloc) 583 Bytes Raw Blame History

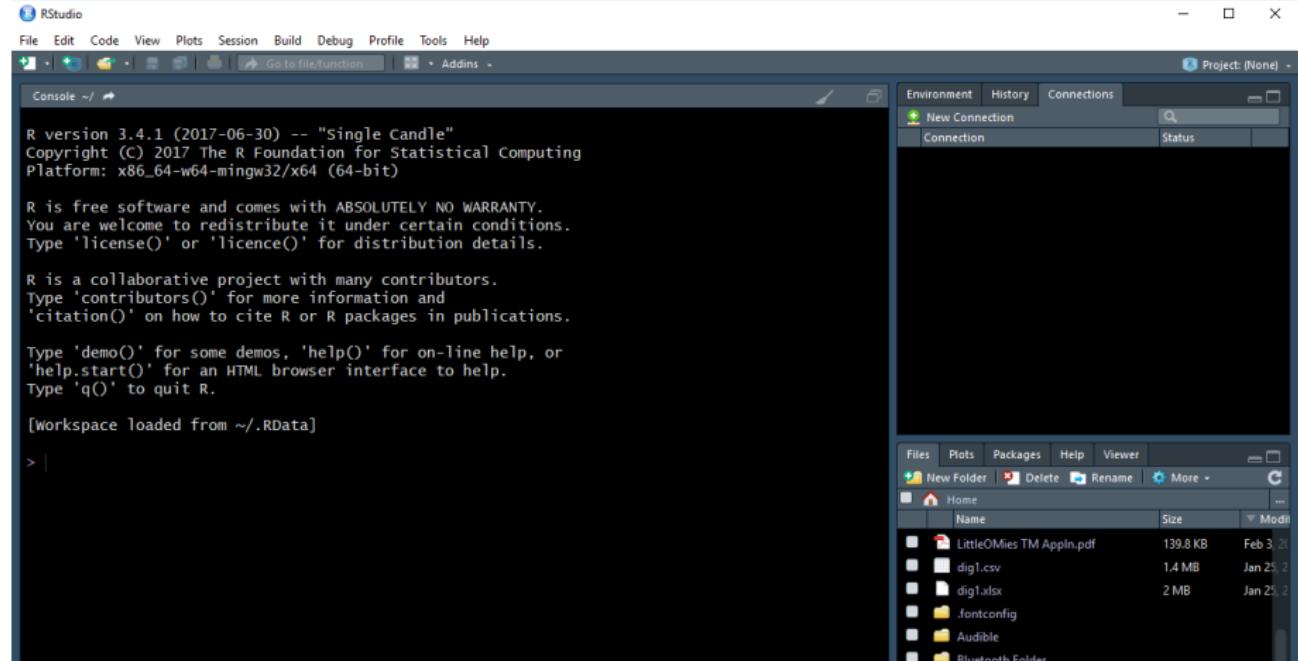
Search this file...

| | subject | age1 | age2 |
|---|---------|------|------|
| 1 | S-01 | 47 | 42 |
| 2 | S-02 | 52 | NA |
| 3 | S-03 | 55 | 55 |
| 4 | S-04 | 48 | 48 |
| 5 | etc | 60 | 40 |

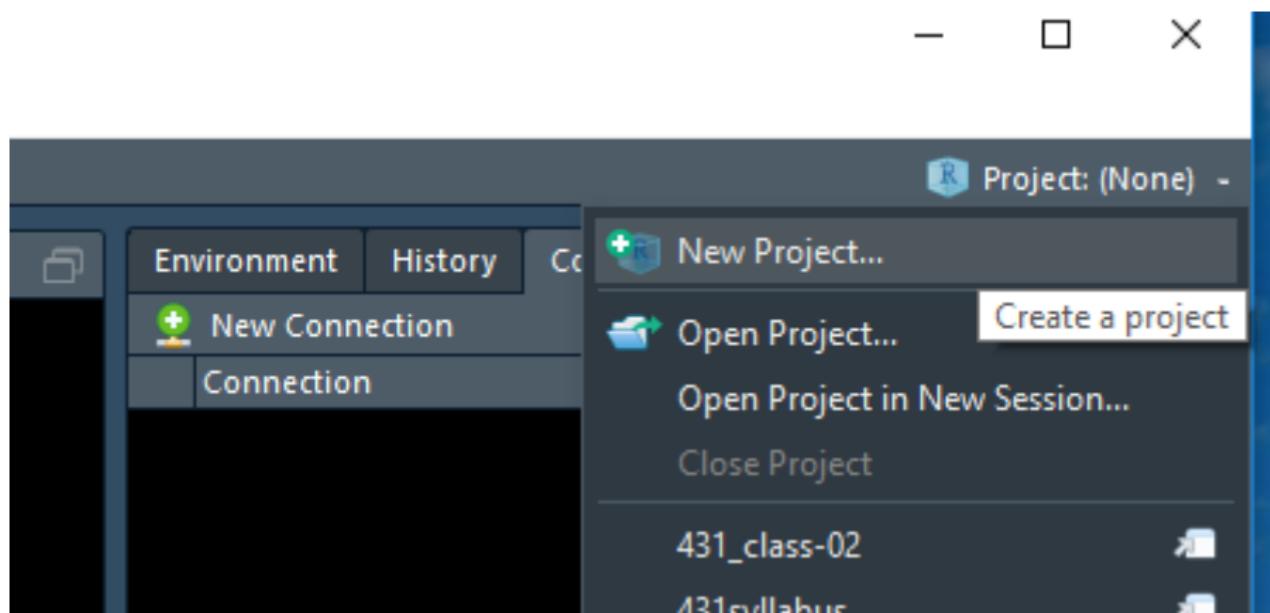
What the file looks like

| | A | B | C |
|----|---------|------|------|
| 1 | subject | age1 | age2 |
| 2 | S-01 | 47 | 42 |
| 3 | S-02 | 52 | NA |
| 4 | S-03 | 55 | 55 |
| 5 | S-04 | 48 | 48 |
| 6 | S-05 | 50 | 48 |
| 7 | S-06 | 51 | 50 |
| 8 | S-07 | 47 | 49 |
| 9 | S-08 | 45 | 48 |
| 10 | S-09 | 55 | 55 |
| 11 | S-10 | 40 | 48 |
| 12 | S-11 | 50 | 54 |

Open R Studio and start a New Project



We'll select our existing 431class2 directory for this project.



New Project

Create Project



New Directory

Start a project in a brand new working directory



Existing Directory

Associate a project with an existing working directory



Version Control

Checkout a project from a version control repository



Cancel

Choose Directory

Dropbox > 431class2

Search 431class2

Organize New folder

Quick access

| Name | Date modified | Type | Size |
|-----------------------------|---------------|------|------|
| No items match your search. | | | |

Desktop Downloads Documents 431_class-02 Better Health C images Survey Day 1 Sca

Desktop Dropbox OneDrive Thomas This PC Libraries

Folder: 431class2

Open Cancel

New Project

Back

Create Project from Existing Directory



Project working directory:

C:/Users/Thomas/Dropbox/431class2

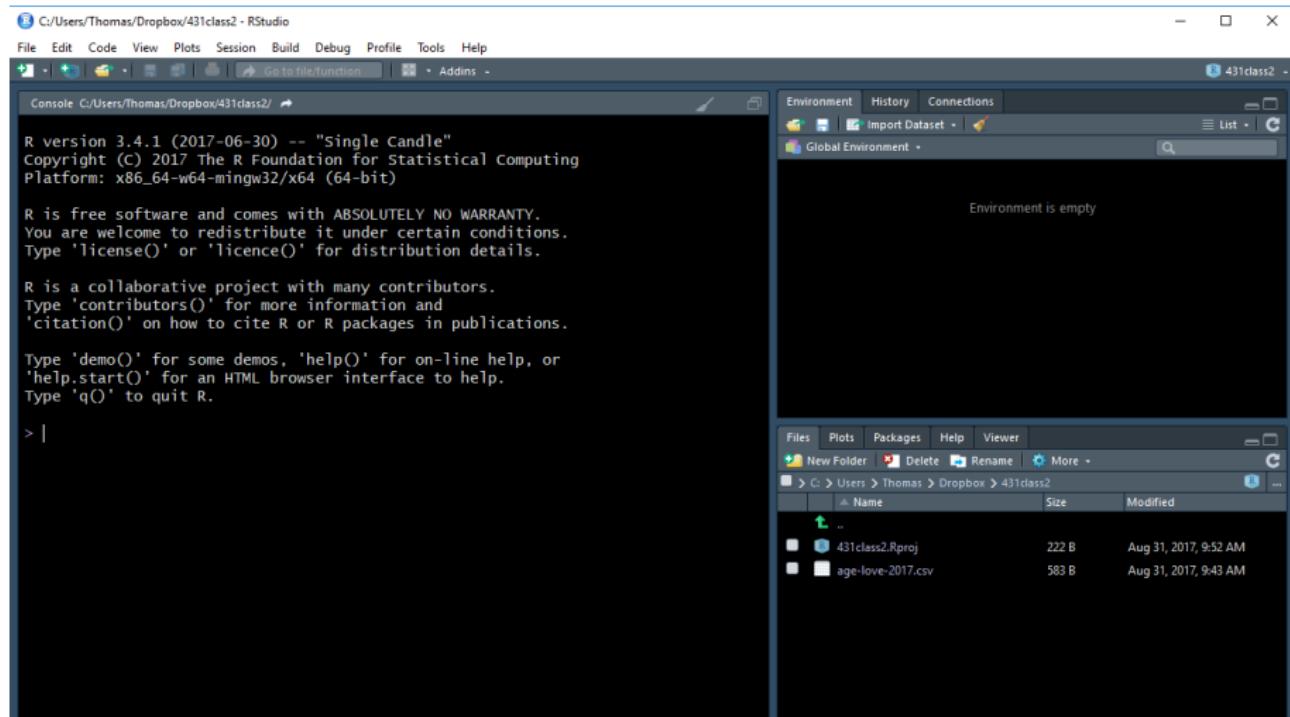
Browse...

Open in new session

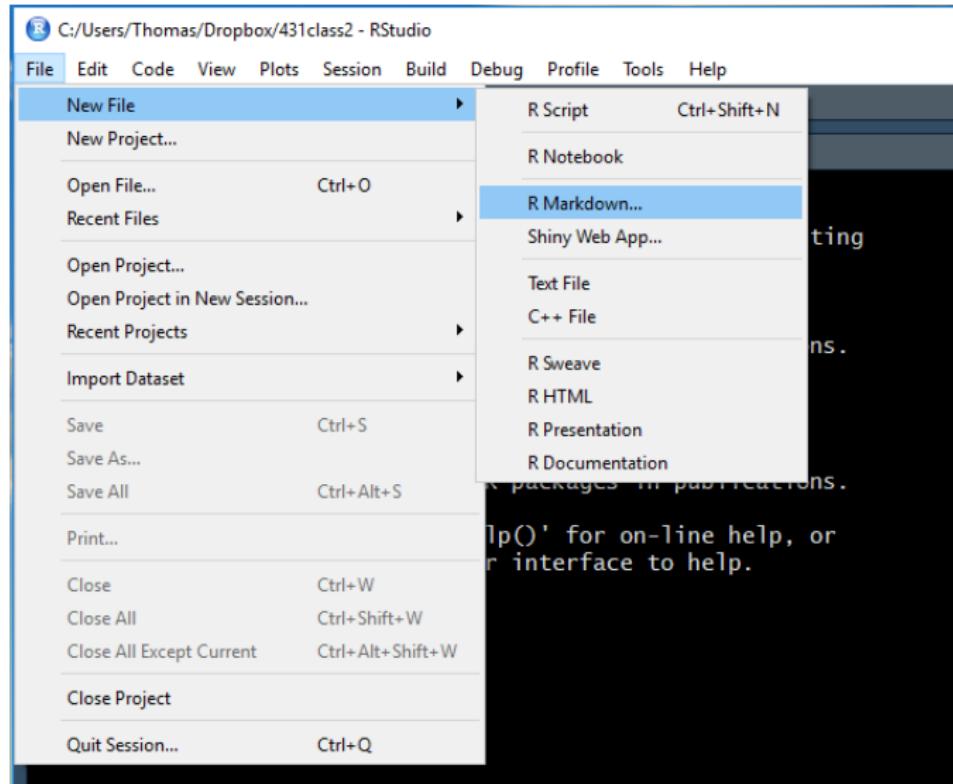
Create Project

Cancel

Now we are in our new Project space.



Start a new R Markdown file



Fill in your name and title for the analysis

New R Markdown

Document Class 2 Age Guess Analysis

Presentation Thomas E. Love

Shiny

From Template

Default Output Format:

HTML
Recommended format for authoring (you can switch to PDF or Word output anytime).

PDF
PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

Word
Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).

OK Cancel

A sample Markdown file is generated. Let's knit it into an HTML file.

The screenshot shows the RStudio interface with an R Markdown script open in the main editor pane. The code includes YAML front matter, R code chunks, and a summary of what R Markdown is. The right sidebar displays the Global Environment and File Browser panes.

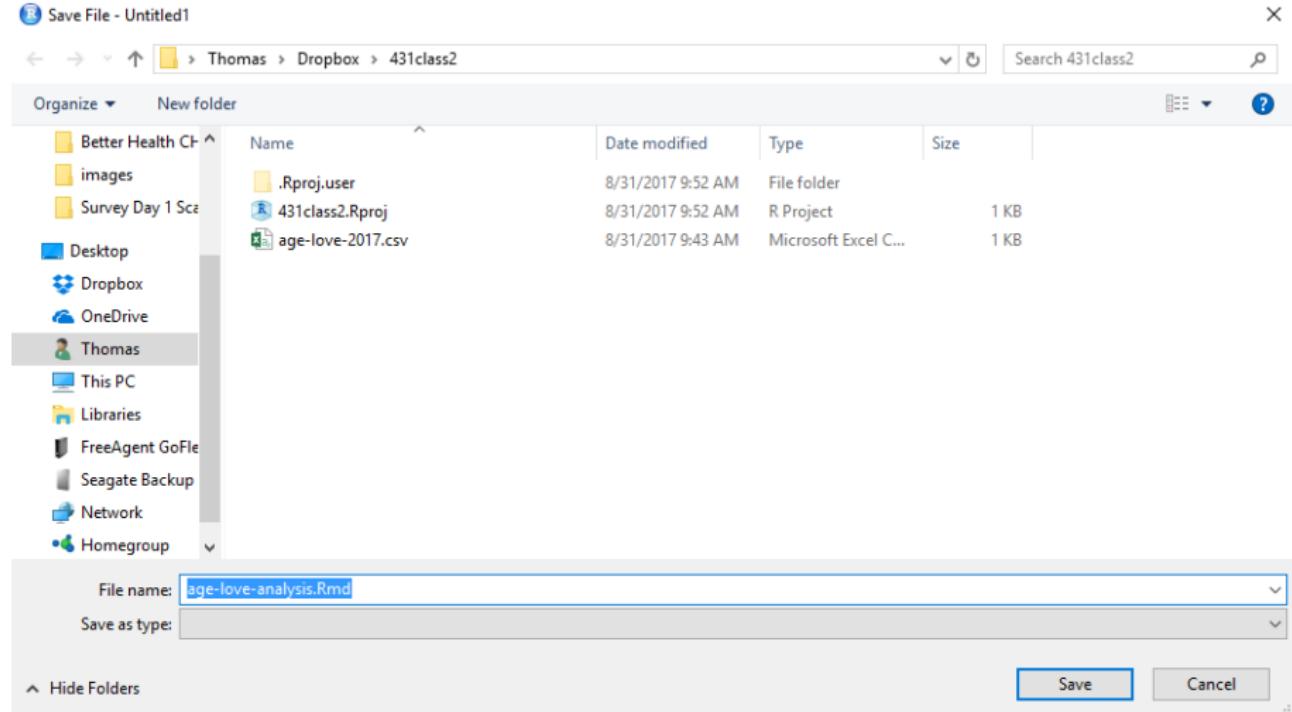
```
1+ ---
2 title: "Class 2 Age Guess Analysis"
3 author: "Thomas E. Love"
4 date: "August 31, 2017"
5 output: html_document
6 ---
7
8 ````{r setup, include=FALSE}
9 knitr:::opts_chunk$set(echo = TRUE)
10 ...
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
15
16 When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
17
18 ````{r cars}
19 summary(cars)
20 ...
21
22 ## Including Plots
23
24 You can also embed plots, for example:
25
```

2:1 Class 2 Age Guess Analysis : R Markdown

Console C:/Users/Thomas/Dropbox/431class2/ ↵

```
R version 3.4.1 (2017-06-30) -- "Single Candle"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
```

Save your work as age-love-analysis.Rmd



The result - a web file!

A screenshot of a web browser window displaying an R Markdown document titled "Class 2 Age Guess Analysis". The browser interface includes a title bar with the file path "C:/Users/Thomas/Dropbox/431class2/age-love-analysis.html", a toolbar with "Open in Browser" and "Find" buttons, and a menu bar with "Publish". The main content area shows the document's title, author ("Thomas E. Love"), date ("August 31, 2017"), and a section titled "R Markdown". Below this, there is explanatory text about R Markdown and a code chunk demonstrating its use. A large rectangular box highlights the R code and its output.

Class 2 Age Guess Analysis

Thomas E. Love
August 31, 2017

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

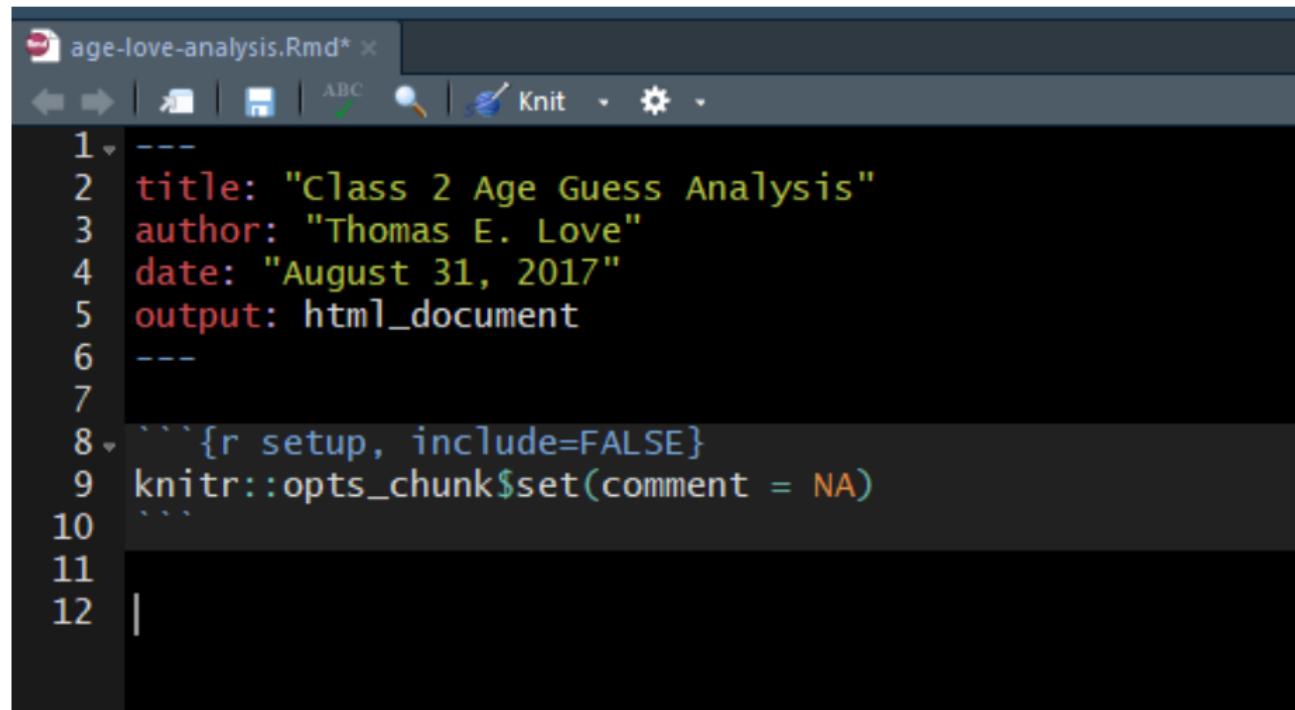
```
##      speed      dist
## Min.   :4.0   Min.   : 2.00
## 1st Qu.:12.0  1st Qu.: 26.00
## Median :15.0  Median : 36.00
## Mean   :15.4  Mean   : 42.98
## 3rd Qu.:19.0  3rd Qu.: 56.00
## Max.   :25.0  Max.   :120.00
```

Including Plots

You can also embed plots, for example:

A scatter plot showing the relationship between "age" (x-axis) and "love" (y-axis). The x-axis ranges from approximately 10 to 50, and the y-axis ranges from 600 to 800. There are three data points plotted: one at approximately (15, 650), one at approximately (35, 750), and one at approximately (45, 800).

Edit the file to change the setup materials



```
age-love-analysis.Rmd* ×
[File] [Edit] [View] [Knit] [Help]
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "August 31, 2017"  
5 output: html_document  
6 ---  
7  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(comment = NA)  
10```  
11  
12 |
```

Insert a new “chunk” of R code

The screenshot shows the RStudio interface with a dark theme. A file named "age-love-analysis.Rmd" is open. The code editor contains the following R Markdown setup code:

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "August 31, 2017"  
5 output: html_document  
6 ---  
7  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(comment = NA)
```

The "Insert" menu is open on the right side of the toolbar, showing options for R, Python, Rcpp, SQL, and Stan. The "R" option is highlighted, and a tooltip indicates "Insert a new R chunk".

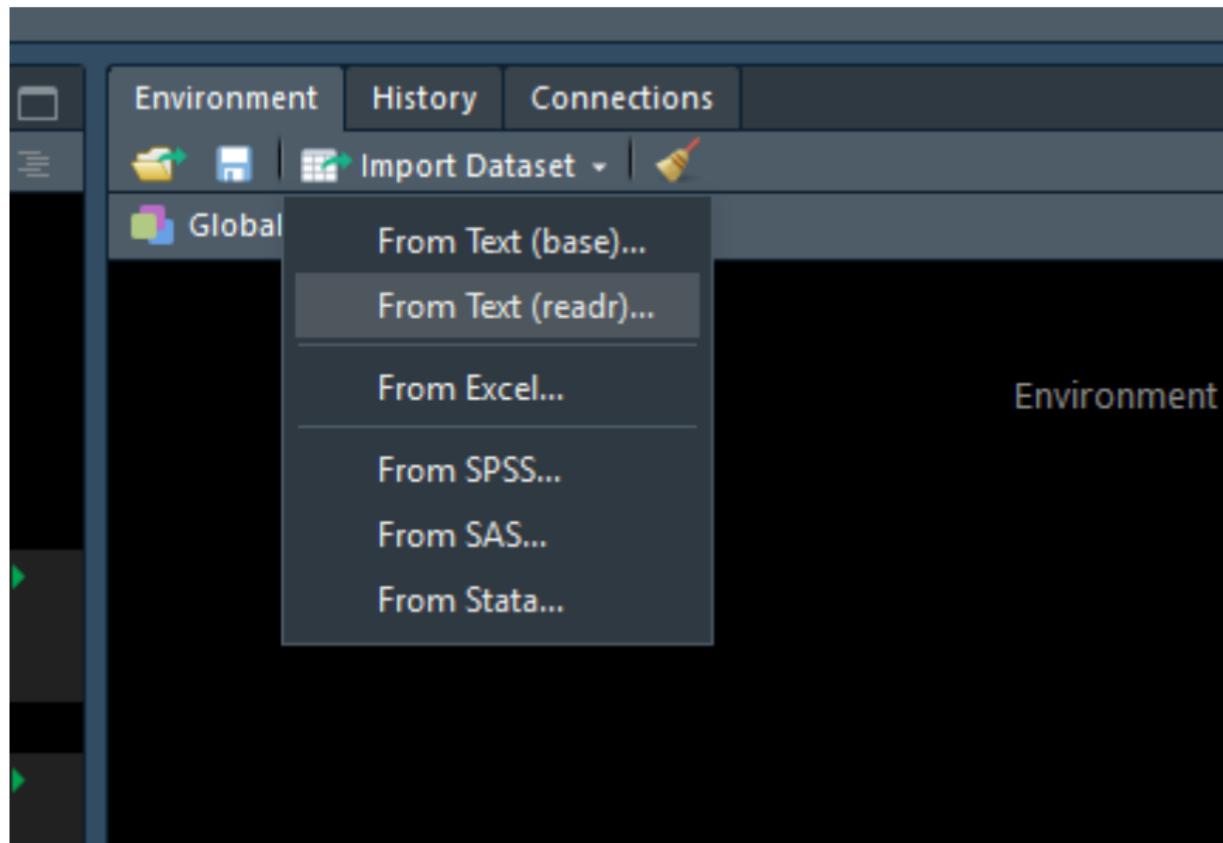
A blank “chunk”

```
8+ ````{r setup, include=FALSE}
9 knitr::opts_chunk$set(comment = NA)
10
11
12+ ````{r}
13 ...
14 ...
15
16
```

Load up two packages in R (should be installed already)

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "August 31, 2017"  
5 output:  
6     html_document:  
7         code_folding: show  
8 ---  
9  
10 ````{r setup, include=FALSE}  
11 knitr::opts_chunk$set(comment = NA)  
12 ````  
13  
14 ````{r load-packages}  
15 library(magrittr); library(tidyverse)  
16 ````
```

Import the .csv data set



Import window

Import Text Data

File/Uri: [Browse...](#)

Data Preview:

Import Options:

| | | | |
|--|--|---|---|
| Name: <input type="text" value="dataset"/> | <input checked="" type="checkbox"/> First Row as Names | Delimiter: <input type="button" value="Comma"/> | Escape: <input type="button" value="None"/> |
| Skip: <input type="text" value="0"/> | <input checked="" type="checkbox"/> Trim Spaces | Quotes: <input type="button" value="Default"/> | Comment: <input type="button" value="Default"/> |
| <input checked="" type="checkbox"/> Open Data Viewer | | Locale: <input type="button" value="Configure..."/> | NA: <input type="button" value="Default"/> |

Code Preview:

```
library(readr)
dataset <- read_csv(NULL)
View(dataset)
```

Reading rectangular data using readr

[Import](#) [Cancel](#)

After we make our choices...

Import Text Data

File/Url
C:/Users/Thomas/Dropbox/431/class2/age-love-2017.csv

Data Preview:

| subject (character) | age1 (integer) | age2 (integer) |
|------------------------|-------------------|-------------------|
| S-01 | 47 | 42 |
| S-02 | 52 | 49 |
| S-03 | 55 | 55 |
| S-04 | 48 | 48 |
| S-05 | 50 | 48 |
| S-06 | 51 | 50 |
| S-07 | 47 | 49 |
| S-08 | 45 | 48 |
| S-09 | 55 | 55 |
| S-10 | 40 | 48 |
| S-11 | 50 | 54 |
| S-12 | 54 | 50 |
| S-13 | 50 | 55 |
| S-14 | 45 | 52 |
| S-15 | 58 | 51 |
| S-16 | 45 | 49 |
| S-17 | 42 | 51 |
| S-18 | 45 | 48 |
| S-19 | 52 | 49 |
| S-20 | 40 | 49 |
| S-21 | 45 | 52 |
| S-22 | 52 | 52 |

Previewing first 50 entries.

Import Options:

| | | | |
|--|--|---|---|
| Name: <input type="text" value="age_love_2017"/> | <input checked="" type="checkbox"/> First Row as Names | Delimiter: <input type="button" value="Comma"/> | Escape: <input type="button" value="None"/> |
| Skip: <input type="text" value="0"/> | <input checked="" type="checkbox"/> Trim Spaces | Quotes: <input type="button" value="Default"/> | Comment: <input type="button" value="Default"/> |
| <input checked="" type="checkbox"/> Open Data Viewer | | Locale: <input type="button" value="Configure..."/> | NA: <input type="button" value="Default"/> |

Code Preview:

```
library(readr)
age_love_2017 <- read_csv("age-love-2017.csv")
View(age_love_2017)
```

Reading rectangular data using readr

Result (note code in Console)

The screenshot shows the RStudio interface with the following components:

- Environment pane:** Shows the global environment with one dataset: "age_love_2017" containing 47 observations and 3 variables.
- File browser pane:** Shows the project directory structure under "C:\Users\Thomas\Dropbox\431class2".
- Console pane:** Displays the R session history, including the command to read the CSV file and the resulting data frame "age_love_2017".

```
## C:\Users\Thomas\Dropbox\431class2 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
age-love-analysis.Rmd age_love_2017.R
[1] subjid age1 age2
[1] 5-01 47 42
[2] 5-02 52 NA
[3] 5-03 55 55
[4] 5-04 48 48
[5] 5-05 50 48
[6] 5-06 51 50
[7] 5-07 47 49
[8] 5-08 45 48
[9] 5-09 55 55
[10] 5-10 40 48
[11] 5-11 50 54
[12] 5-12 54 50
[13] 5-13 50 55
[14] 5-14 45 52
[15] 5-15 58 51
[16] 5-16 45 49
[17] 5-17 42 51
[18] 5-18 45 48
[19] 5-19 52 49
[20] 5-20 40 NA
Showing 1 to 20 of 47 entries
Console R Markdown
C:\Users\Thomas\Dropbox\431class2
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(readr)
> age_love_2017 <- read_csv("age-love-2017.csv")
Parsed with column specification:
cols(
  subject = col_character(),
  age1 = col_integer(),
  age2 = col_integer()
)
> View(age_love_2017)
> |
```

Add data load code to Markdown and also look at the data

```
age-love-analysis.Rmd* | age_love_2017* | Knit | Insert | Run | Help
```

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "August 31, 2017"  
5 output: html_document  
6 ---  
7  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(comment = NA)  
10 ...  
11  
12 ```{r load-packages}  
13 library(tidyverse)  
14 ...  
15  
16 ```{r load-dataset}  
17 age_love_2017 <- read_csv("age-love-2017.csv")  
18 ...  
19  
20 ```{r look-at-data}  
21 age_love_2017  
22 ...  
23
```

Running the code so we can see results

The screenshot shows the RStudio interface. On the left, there are two tabs: "age-love-analysis.Rmd*" and "age_love_2017". The main area displays an R Markdown code block:

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "August 31, 2017"  
5 output: html_document  
6 ---  
7  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(comment = NA)  
10  
11  
12 ```{r load-packages}  
13 library(tidyverse)  
14  
15  
16 ```{r load-dataset}  
17 age_love_2017 <- read_csv("age-love-2017.csv")
```

A context menu is open on the right side of the screen, specifically over the line "Run All". The menu contains the following items:

- Run Selected Line(s) Ctrl+Enter
- Run Current Chunk Ctrl+Shift+Enter
- Run Next Chunk Ctrl+Alt+N
- Run Setup Chunk
- Run Setup Chunk Automatically
- Run All Chunks Above Ctrl+Alt+P
- Run All Chunks Below
- Restart R and Run All Chunks
- Restart R and Clear Output
- Run All Ctrl+Alt+R

Or run all of the “chunks” up to a particular point

The screenshot shows the RStudio interface with the following details:

- File Tab:** age-love-analysis.Rmd*
- Panel Tab:** age_love_2017
- Toolbar:** Includes back, forward, search, Knit, Run, and other document-related icons.
- Code Area:** Displays R Markdown code. Lines 1 through 25 are shown, with lines 12 and 20 highlighted in yellow. The code includes setup, package loading, dataset reading, and data examination chunks.
- Context Menu (Visible on the right):** A context menu is open at the bottom right of the code area, with the option "Run All Chunks Above" highlighted.
- Environment Panel:** Shows the global environment with objects like age_love_2017.
- File Explorer:** Shows files like New Fo...
- Plots:** Shows plots like 4...

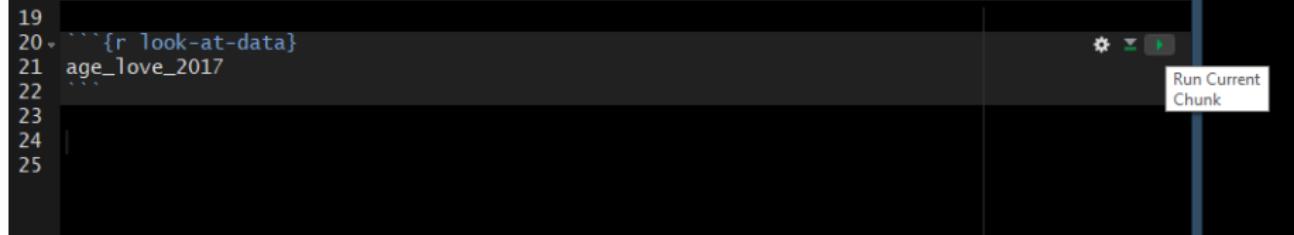
Running the first three chunks of code

The screenshot shows the RStudio interface with the following code in the script pane:

```
age-love-analysis.Rmd* age_love_2017.x
Insert Run Knit
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "August 31, 2017"  
5 output: html_document  
6 ---  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(comment = NA)  
10 ...  
11  
12 ```{r load-packages}  
13 library(tidyverse)  
14 ...  
15  
16 Loading tidyverse: ggplot2  
17 Loading tidyverse: tibble  
18 Loading tidyverse: tidyverse  
19 Loading tidyverse: purrr  
20 Loading tidyverse: dplyr  
21 Conflicts with tidy packages -----  
22 filter(): dplyr, stats  
23 lag(): dplyr, stats  
24  
15  
16 ...  
17 age_love_2017 <- read_csv("age-love-2017.csv")  
18 ...  
19  
20 Parsed with column specification:  
21 cols(  
22   subject = col_character(),  
23   age1 = col_integer(),  
24   age2 = col_integer()  
25 )  
26  
27  
28  
29  
30  
31  
32  
33  
34
```

The code consists of three chunks. The first chunk sets up the document metadata and initializes knitr options. The second chunk loads the tidyverse package. The third chunk reads a CSV file named "age-love-2017.csv" into a dataset called "age_love_2017". The RStudio interface shows the code in the script pane, the rendered output in the preview pane, and various status indicators like "Loading tidyverse" and "Parsed with column specification".

Now, let's look at the data



A screenshot of a Jupyter Notebook interface. On the left, there are five numbered code cells (19 to 23). Cell 19 contains the command `!ls`. Cells 20 through 23 contain the command `%%r look-at-data`. Cell 21 also includes the command `age_love_2017`. The right side of the interface shows a toolbar with several icons, and a tooltip box is open over the "Run Current Chunk" button.

```
19 !ls
20 %%r look-at-data
21 age_love_2017
22
23
24
25
```

The age_love_2017 tibble

```
```{r look-at-data}
age_love_2017
```

| subject | age1 | age2 |
|---------|------|------|
| S-01    | 47   | 42   |
| S-02    | 52   | NA   |
| S-03    | 55   | 55   |
| S-04    | 48   | 48   |
| S-05    | 50   | 48   |
| S-06    | 51   | 50   |
| S-07    | 47   | 49   |
| S-08    | 45   | 48   |
| S-09    | 55   | 55   |
| S-10    | 40   | 48   |

1-10 of 47 rows

Previous  2 3 4 5 Next

# Typing in some documentation, mixing text with R code.

```
20 ````{r look-at-data}
21 age_love_2017
22
```

| subject<br><chr> | age1<br><int> | age2<br><int> |
|------------------|---------------|---------------|
| S-01             | 47            | 42            |
| S-02             | 52            | NA            |
| S-03             | 55            | 55            |
| S-04             | 48            | 48            |
| S-05             | 50            | 48            |
| S-06             | 51            | 50            |
| S-07             | 47            | 49            |
| S-08             | 45            | 48            |
| S-09             | 55            | 55            |
| S-10             | 40            | 48            |

1-10 of 47 rows

Previous  2 3 4 5 Next

```
23
24
25
```

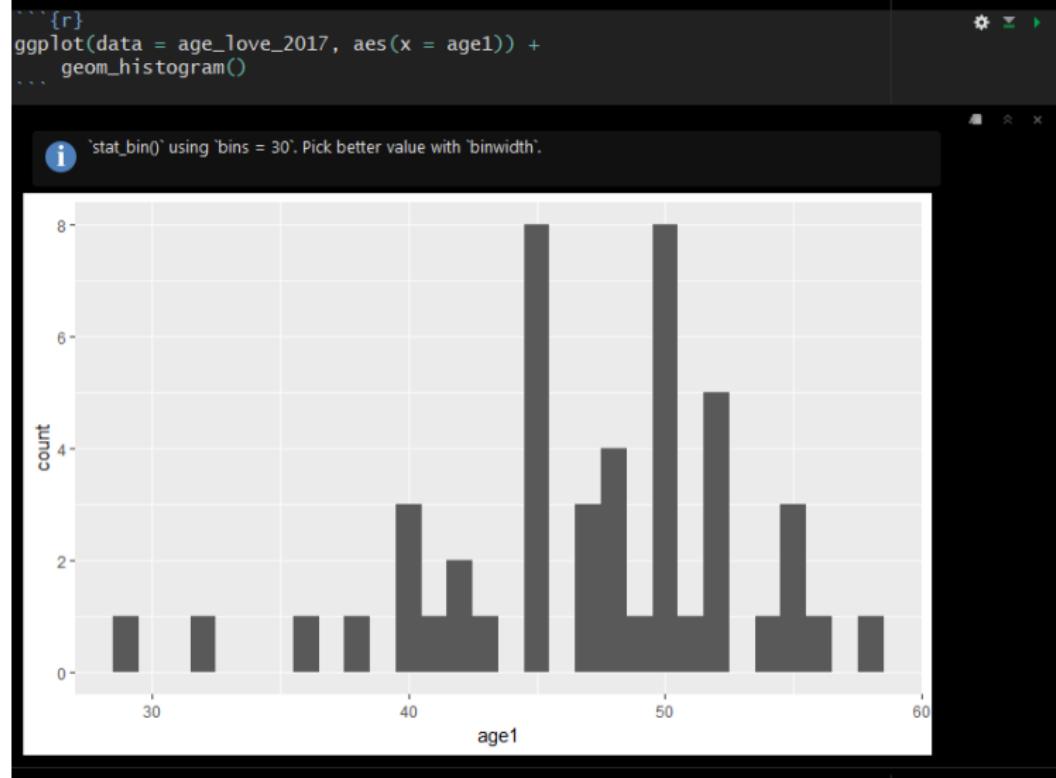
Our data set has `r nrow(age\_love\_2017)` rows, and `r ncol(age\_love\_2017)` columns.

# Let's make a histogram

```
25
26 # Build a Picture of the First Guesses
27
28 ````{r}
29 ggplot(data = age_love_2017, aes(x = age1)) +
30 geom_histogram()
31
32
```

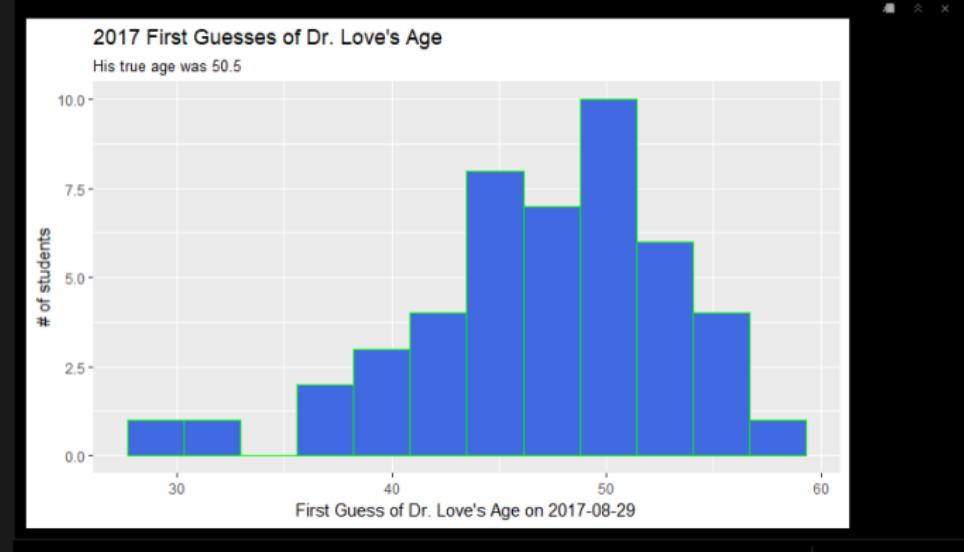


# The result. Can we do better?



# A second attempt at the histogram

```
33 ## A nicer picture
34
35 ## [r second-histogram]
36 ggplot(data = age_love_2017, aes(x = age1)) +
37 geom_histogram(bins = 12, fill = "royalblue", col = "green") +
38 labs(x = "First Guess of Dr. Love's Age on 2017-08-29",
39 y = "# of students",
40 title = "2017 First Guesses of Dr. Love's Age",
41 subtitle = "His true age was 50.5")
42 ...
```



# A numerical summary of the data

```
43
44 # A Numerical Summary
45
46 ``{r numerical-summary}
47 summary(age_love_2017)
48 ```

 subject age1 age2
Length:47 Min. :29.0 Min. :29.00
Class :character 1st Qu.:45.0 1st Qu.:48.00
Mode :character Median :48.0 Median :50.00
 Mean :47.0 Mean :48.73
 3rd Qu.:50.5 3rd Qu.:52.00
 Max. :58.0 Max. :55.00
 NA's :6
```

49  
50 |

# Calculating and Summarizing the Errors

```
49
50 # Better First or Second Guess?
51
52 ```{r calculate-errors}
53 age_love_2017 <- age_love_2017 %>%
54 mutate(error1 = abs(age1 - 50.5),
55 error2 = abs(age2 - 50.5))
56
57 summary(age_love_2017)
58 ````
```



# What do these summaries suggest?

```
19
50 # Better First or Second Guess?
51
52 ````{r calculate-errors}
53 age_love_2017 <- age_love_2017 %>%
54 mutate(error1 = abs(age1 - 50.5),
55 error2 = abs(age2 - 50.5))
56
57 summary(age_love_2017)
58 ````
```

| subject          | age1         | age2          | error1         | error2         |
|------------------|--------------|---------------|----------------|----------------|
| Length:47        | Min. :29.0   | Min. :29.00   | Min. : 0.500   | Min. : 0.500   |
| Class :character | 1st Qu.:45.0 | 1st Qu.:48.00 | 1st Qu.: 1.500 | 1st Qu.: 0.500 |
| Mode :character  | Median :48.0 | Median :50.00 | Median : 4.500 | Median : 1.500 |
|                  | Mean :47.0   | Mean :48.73   | Mean : 5.117   | Mean : 3.354   |
|                  | 3rd Qu.:50.5 | 3rd Qu.:52.00 | 3rd Qu.: 6.500 | 3rd Qu.: 3.500 |
|                  | Max. :58.0   | Max. :55.00   | Max. :21.500   | Max. :21.500   |
|                  | NA's :6      |               | NA's :6        |                |

# Build a scatterplot to compare the errors

```
59
60 # Compare the Guesses
61
62 ````{r guess-1-vs-2}
63 ggplot(data = age_love_2017, aes(x = error1, y = error2)) +
64 geom_point()
65 ...
```

# A new scatterplot, with a model for the relationship of age1 to age2

```
66
67 ## Add a Prediction Model?
68
69 ````{r guess1-vs-guess2-with-loess-smooth}
70 ggplot(data = age_love_2017, aes(x = age1, y = age2)) +
71 geom_point(size = 3) +
72 geom_smooth(method = "loess")
73 ...
```

## Plot the age1 - age2 differences

```
Plot the (matched) differences
```{r histogram-of-differences}
ggplot(age_love_2017, aes(x = age1 - age2)) +
  geom_histogram(binwidth = 2,
                 col = "green",
                 fill = "royalblue")
```
```

# Numerical summary of the age1 - age2 differences

```
83
84 # Numerical Summary of the Difference in Ages
85
86 ```{r comparing-guess1-to-guess2}
87 age_love_2017 %$%
88 summary(age1 - age2)
89 ...
90
```

# How many people thought I looked younger the second time?

```
92
93 # How many people thought I looked younger in guess 2?
94
95 `r count-of-younger-guesses}
96 age_love_2017 %>%
97 count(age1-age2 < 0)
98 ...
```

| age1 - age2 < 0<br><lg > | n<br><int> |
|--------------------------|------------|
| FALSE                    | 18         |
| TRUE                     | 23         |
| NA                       | 6          |

3 rows

```
99
100 I think this is good news! Maybe ...
101
```

## T tests - making a statistical inference

```
97
98 # The Much-Dreaded t test
99
100 ````{r t-test-comparing-guess1-to-guess2}
101 age_love_2017 %$%
102 t.test(age1 - age2)
103 ````
```

# Knit the file into an HTML document

C:/Users/Thomas/Dropbox/431class2/age-love-analysis.html

age-love-analysis.html | Open in Browser | Find

## Class 2 Age Guess Analysis

Thomas E. Love

August 31, 2017

```
library(magrittr); library(tidyverse)
```

```
Loading tidyverse: ggplot2
Loading tidyverse: tibble
Loading tidyverse: tidyr
Loading tidyverse: readr
Loading tidyverse: purrr
Loading tidyverse: dplyr
```

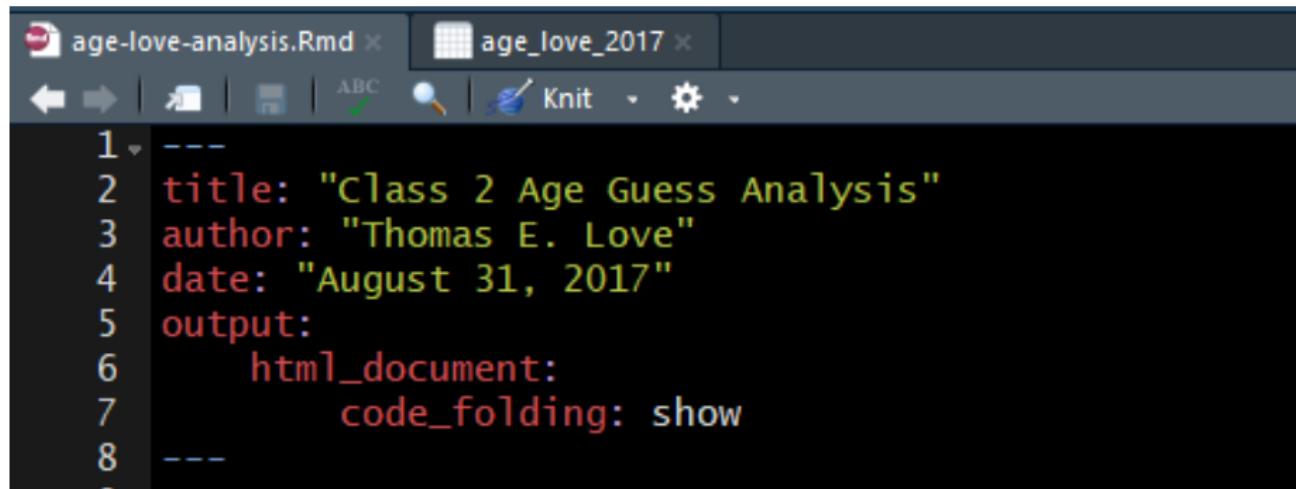
```
Conflicts with tidy packages -----
```

```
filter(): dplyr, stats
lag(): dplyr, stats
```

```
age_love_2017 <- read_csv("age-love-2017.csv")
```

```
Parsed with column specification:
```

## Adjust the YAML to fold the code on demand



The screenshot shows the RStudio interface with two tabs open: 'age-love-analysis.Rmd' and 'age\_love\_2017'. The 'age\_love\_2017' tab is active, displaying the following YAML code:

```
1 ---
2 title: "Class 2 Age Guess Analysis"
3 author: "Thomas E. Love"
4 date: "August 31, 2017"
5 output:
6 html_document:
7 code_folding: show
8 ---
```

# New, more final, report

The screenshot shows a RStudio interface with the following details:

- Title Bar:** C:/Users/Thomas/Dropbox/431class2/age-love-analysis.html
- Toolbar:** age-love-analysis.html, Open in Browser, Find, Publish
- Section Header:** Class 2 Age Guess Analysis
- Text:** Thomas E. Love, August 31, 2017
- Code Block:** library(magrittr); library(tidyverse)
- Output Block:** Loading tidyverse: ggplot2, Loading tidyverse: tibble, Loading tidyverse: tidyr, Loading tidyverse: readr

# Analyzing the Survey Data - A little challenge

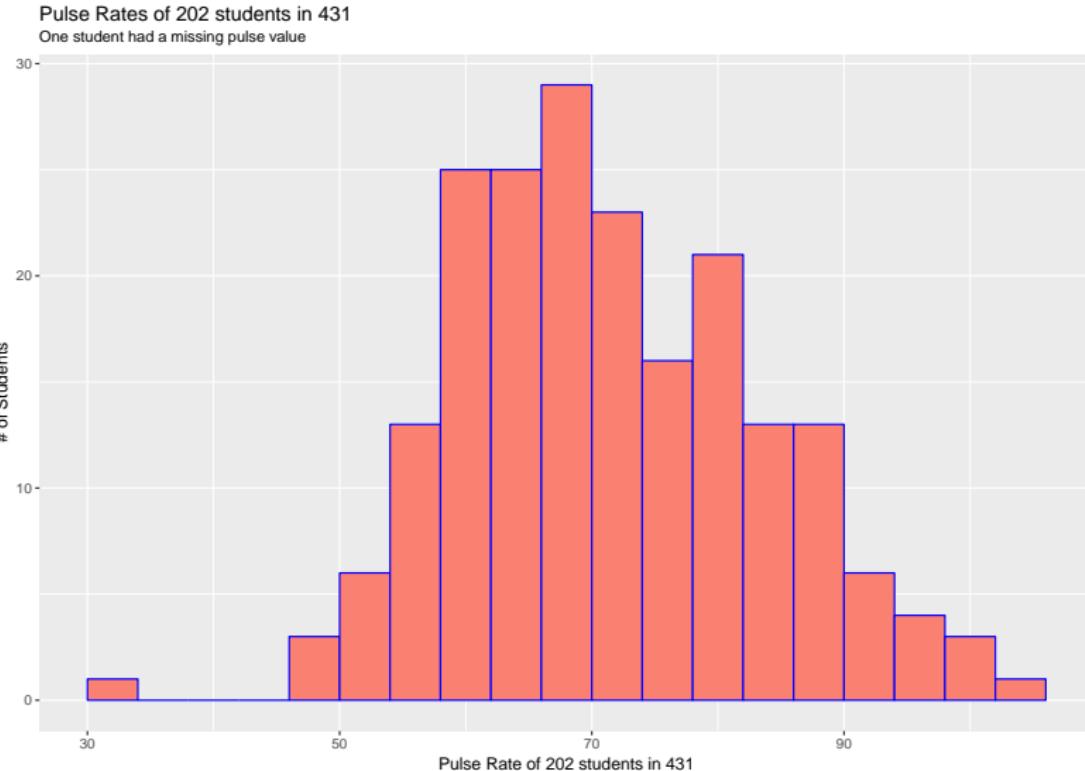
We have data on the site in a file called `surveyday1_2017.csv`. Build a project to study those data.

Put the data in a file called `surv1` in R.

- I'd call my R Markdown file `day1surveyanalysis`

Can you reproduce the following...

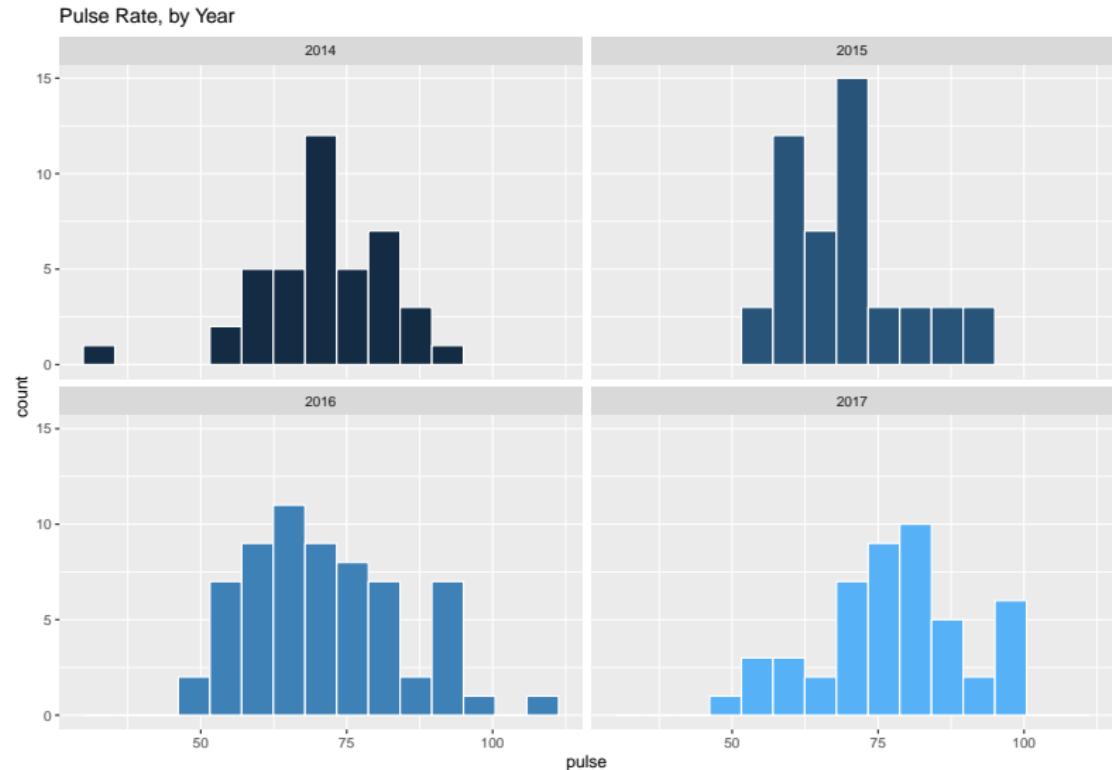
# A. That fill color is called *salmon*, I used 20 bins.



## Code for Plot A.

```
ggplot(surv1, aes(x = pulse)) +
 geom_histogram(bins = 20,
 col = "blue", fill = "salmon") +
 labs(x = "Pulse Rate of 202 students in 431",
 y = "# of Students",
 title = "Pulse Rates of 202 students in 431",
 subtitle = "One student had a missing pulse value")
```

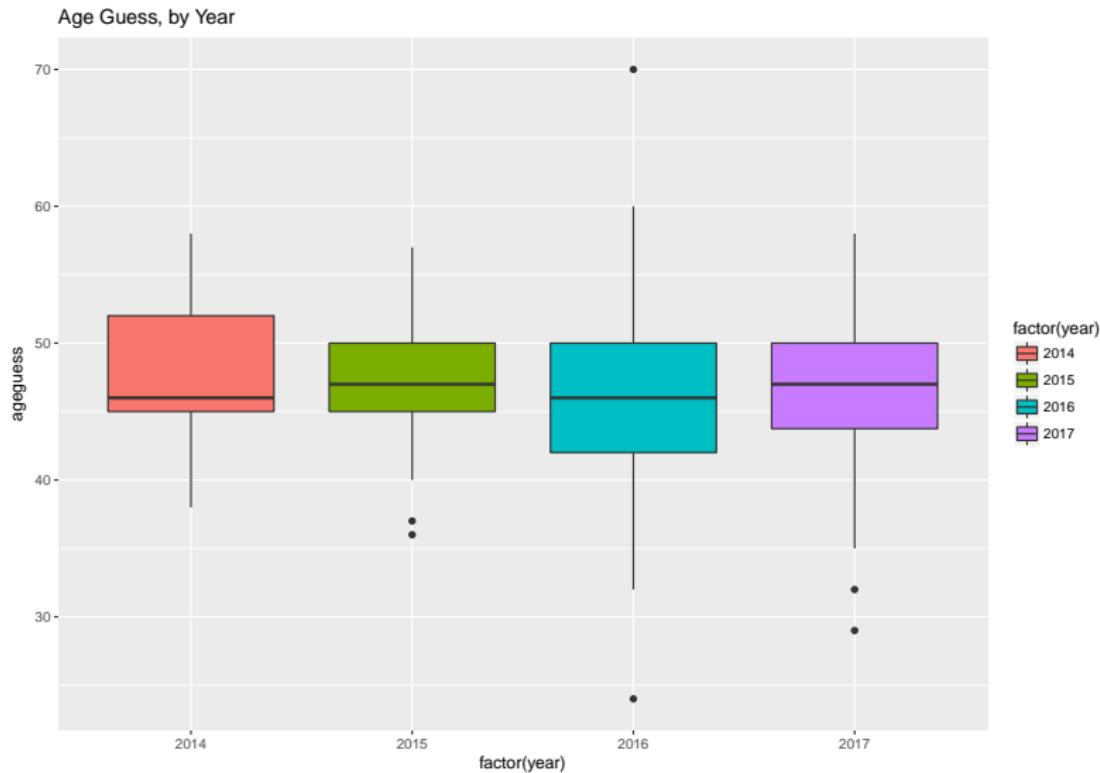
## B. Histograms of Pulse Rates, by Year



## Code for Plot B.

```
ggplot(surv1, aes(x = pulse, fill = year)) +
 geom_histogram(bins = 15, col = "white") +
 facet_wrap(~ year) +
 guides(fill = FALSE) +
 labs(title = "Pulse Rate, by Year")
```

# C.Boxplot of Age Guesses, by Year



## Code for Plot C

```
ggplot(surv1, aes(x = factor(year), y = ageguess,
 fill = factor(year))) +
 geom_boxplot() +
 labs(title = "Age Guess, by Year")
```

## Summary Table of Age Guesses, by Year

```
A tibble: 4 x 5
 year n mean sd median
 <int> <int> <dbl> <dbl> <dbl>
1 2014 42 47.34146 5.213491 46
2 2015 49 47.12245 4.617145 47
3 2016 64 45.96721 6.999922 46
4 2017 48 46.54167 6.146988 47
```

## Code for Summary Table

```
surv1 %>%
 group_by(year) %>%
 summarize(n = n(),
 mean = mean(ageguess, na.rm=TRUE),
 sd = sd(ageguess, na.rm=TRUE),
 median = median(ageguess, na.rm=TRUE)
)
```

# What's coming up?

- Running more involved analyses in R and R Studio
- More on exploratory data analysis for distributions and associations
- Discussion of the project requirements is coming next week
- Never too early to get started
  - Read Leek Chapters 5, 9, 10 and 13 (about 30 pages in total) by 2017-09-07 (Class 4)
  - Read Silver Introduction and Chapter 1 (about 50 denser pages) by 2017-09-12 (Class 5)
  - Assignment 1 is due 2017-09-15 at noon