

431 Class 02

Thomas E. Love

2018-08-30

I DON'T KNOW HOW
TO DO STATISTICS BUT
IT DOESN'T MATTER
BECAUSE I DIDN'T
HAVE DATA.



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 2018-09-04

This schedule is also part of the Course Calendar.

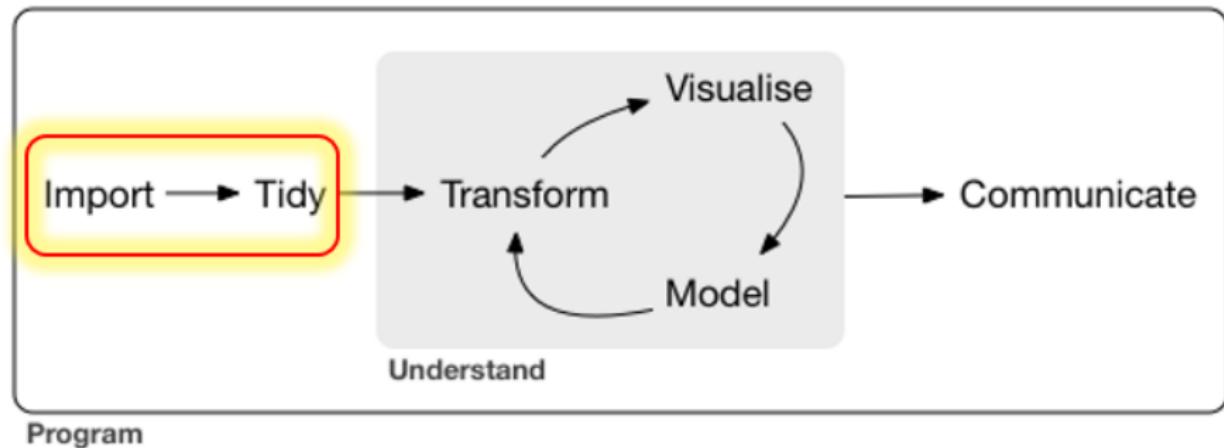
- Tuesdays 11:30 AM to 12:45 PM
- Wednesdays 12 noon to 1:30 PM
- Thursdays 11:30 AM to 12:45 PM, 2:30 - 4 PM, 5:30 - 7 PM
- Fridays 10:30 AM to 12 noon

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/431-2018>

Data Science



Types of Data (see 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

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 ① | Slightly ② | Somewhat ③ | Extremely ④ | important ⑤ | important ⑥ | important ⑦ |

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

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

| | | |
|---------|-----------------|---------------------|
| No ① | I used to. ② | Yes. Smoker ③ |
|---------|-----------------|---------------------|

6. Please indicate which hand you use for each of the following activities by putting + 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 ① | Slightly important ② | Somewhat important ③ | Extremely important ④ |
|------------------------------|----------------------------|----------------------------|-----------------------------|
|------------------------------|----------------------------|----------------------------|-----------------------------|

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

Please indicate your agreement with the following statements:

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

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

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

- 51 people completed it Tuesday. Prior counts:

| Fall | 2018 | 2017 | 2016 | 2015 | 2014 | Total |
|----------|------|------|------|------|------|------------|
| <i>n</i> | 51 | 48 | 64 | 49 | 42 | 254 |

Question 1

About how many of those 254 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 ① | Slightly ② | Somewhat ③ | Extremely ④ | important ⑤ | important ⑥ | important ⑦ |

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

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

| | | |
|------------|-----------------|-----------|
| No ① | I used to. ② | Yes. ③ |
| Non-Smoker | Former Smoker | Smoker |

6. Please indicate which hand you use for each of the following activities by putting + 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 ① | Slightly important ② | Somewhat important ③ | Extremely important ④ |
|------------------------------|----------------------------|----------------------------|-----------------------------|
| ⑤ | ⑥ | ⑦ | ⑧ |

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

Please indicate your agreement with the following statements:

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

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 254 surveys caused *no problems* in recording responses?

- Guesses?

Question 1

About how many of those 254 surveys caused *no problems* in recording responses?

- Guesses?
- $90/254$ (34%)

Question 1

About how many of those 254 surveys caused *no problems* in recording responses?

- Guesses?
- $90/254$ (34%)
- 16 of the 51 surveys turned in Tuesday had **no** problems (31%)

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.

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

50ish years.

What should we do in these cases?

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) maybe

TEL decision: NA

Favorite color

13. What is your favorite color? depends

NA

13. What is your favorite color? orange

orange

13. What is your favorite color? Blue, Brown

13. What is your favorite color? N/A

Following the Rules?

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

75 beats/minute.

2018 pulse responses, sorted ($n = 51$)

| | |
|------------------------------|---------------------|
| 48 52 54 56 56 58 60 | 4 8 |
| 62 62 64 64 64 64 64 | 5 24668 |
| 66 66 66 66 68 68 68 | 6 022444446666888 |
| 70 70 70 70 72 74 74 | 7 0000244446688 |
| 74 74 76 76 78 78 80 | 8 0000024456 |
| 80 80 80 80 82 84 84 | 9 22258 |
| 85 86 92 92 92 95 98 100 104 | 10 04 |

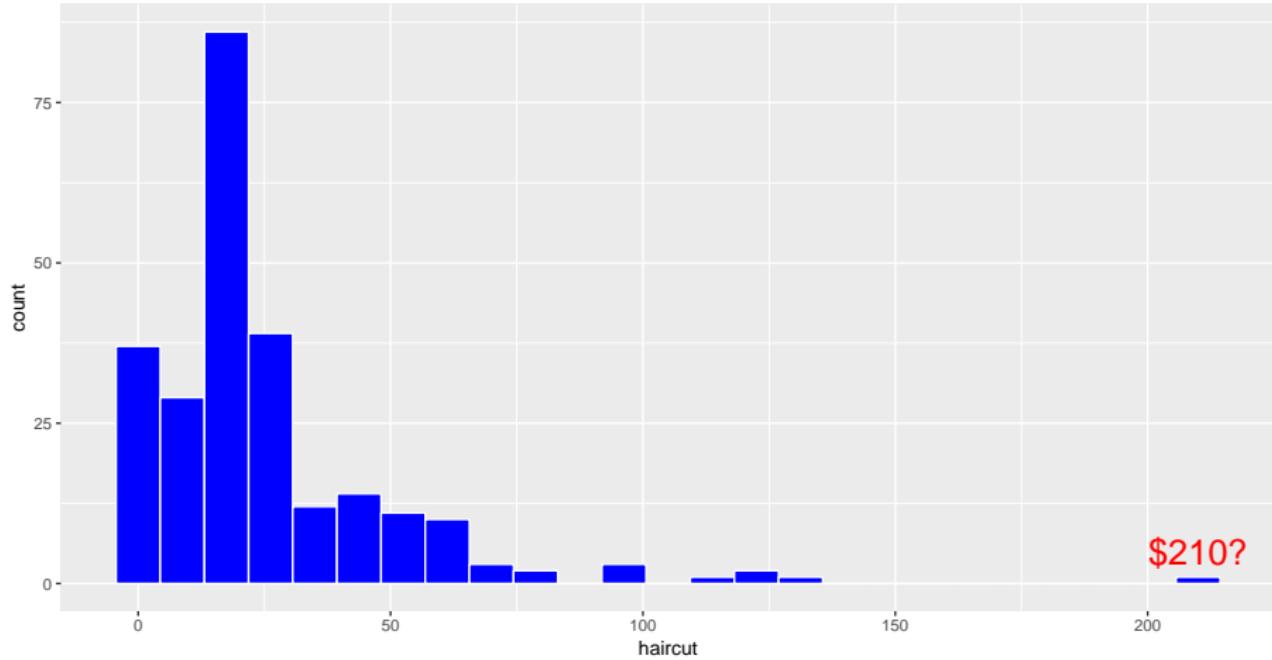
Stem and Leaf: Haircut \$ (Thanks, John Tukey)

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

| | |
|----|--|
| 0 | 00035555556778999 |
| 1 | 00000022222222244444455555555555555555555556778888889 |
| 2 | 00012234555555555555556688 |
| 3 | 00000000000000002222555556 |
| 4 | 00000000003558 |
| 5 | 00000000555 |
| 6 | 0000000005 |
| 7 | 0005 |
| 8 | 0 |
| 9 | |
| 10 | 000 |
| 11 | 0 |
| 12 | 00 |
| 13 | 0 |
| 14 | |
| 15 | |
| 16 | |
| 17 | |
| 18 | |
| 19 | |
| 20 | |
| 21 | 0 |

Haircut Histogram

Histogram of 251 Haircut Prices from Day 1 Survey



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 Numerical Summaries

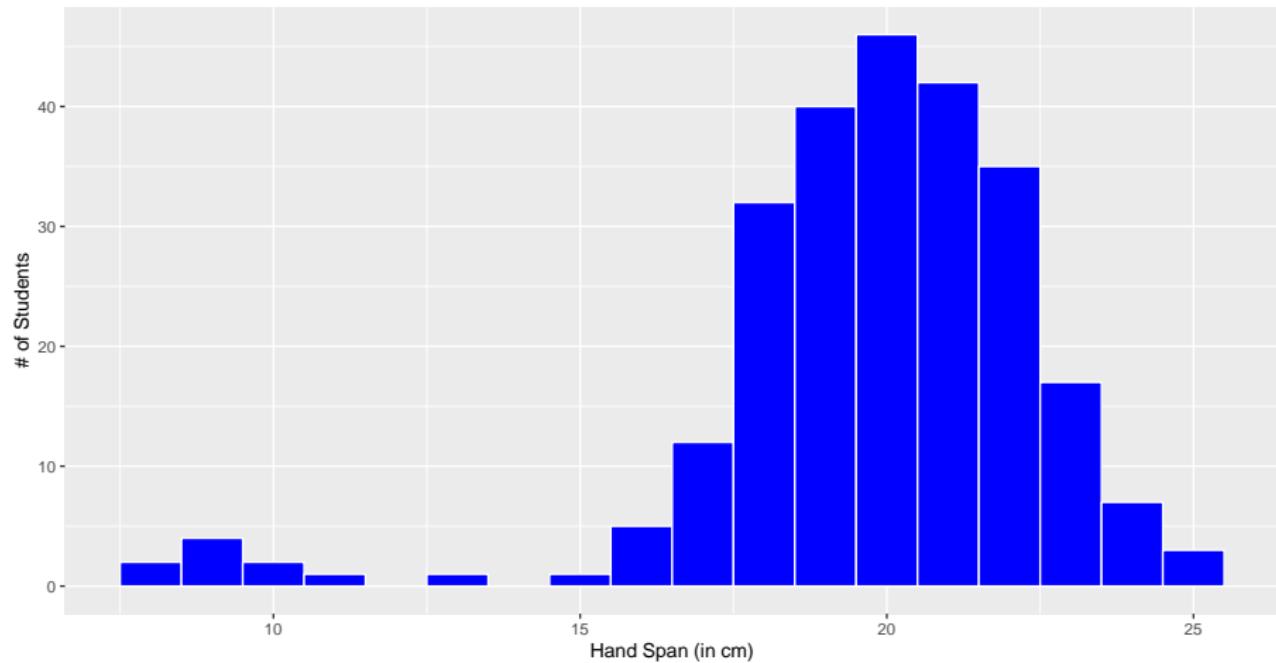
```
summary(surv1$hand.span)
```

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|------|------|---------|--------|-------|---------|-------|
| | 8.00 | 19.00 | 20.00 | 19.88 | 21.50 | 25.00 |
| NA's | | | | | | |
| | 4 | | | | | |

Hand Span (cm) Histogram

Warning: Removed 4 rows containing non-finite values (stat_bin).

2014–2018 431 student hand span measurements



Hand Span (cm) Histogram (Code)

```
ggplot(data = surv1, aes(x = hand.span)) +
  geom_histogram(bins = 18, col = "white", fill = "blue") +
  labs(x = "Hand Span (in cm)",
       y = "# of Students",
       title = "2014-2018 431 student hand span measurements")
```

Hand Span Stem-and-Leaf, (Two digits per stem)

The decimal point is at the |

| | | |
|----|--|---|
| 8 | | 050055 |
| 10 | | 000 |
| 12 | | 5 |
| 14 | | 5 |
| 16 | | 00005800000000055 |
| 18 | | 00000000000000000555555555555560000000000000000+14 |
| 20 | | 00000000000000000000000000000000002345555555557800+29 |
| 22 | | 0000000000000000000000000000000024555556800000000000555 |
| 24 | | 0000001000 |

Eight Items had just a few problems

| # | 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.

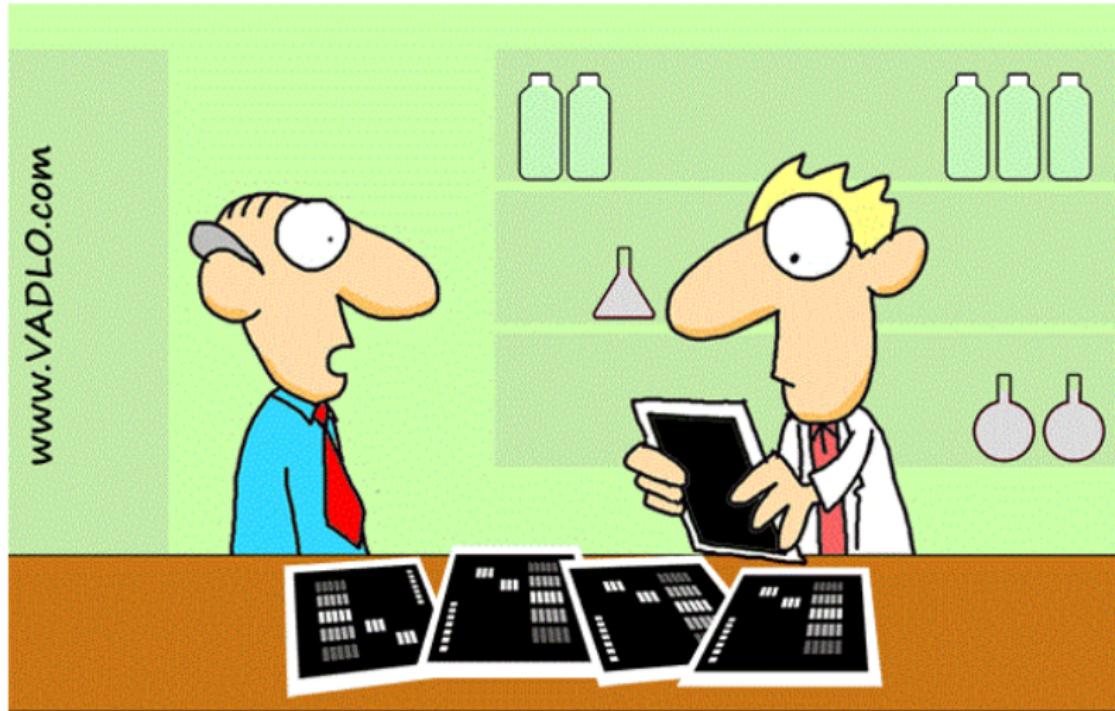
| Task | Left | Right |
|--|------|-------|
| Writing | | ✓ |
| Drawing | | ✓ |
| Throwing | | ✓ |
| Scissors | | ✓ |
| Toothbrush | ✓ | |
| Knife (without fork) | ✓ | |
| Spoon | ✓ | ✓ |
| Broom (upper hand) | | ✓ |
| Striking match (hand that holds the match) | | ✓ |
| Opening box (hand that holds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

Handedness Scale (2016-18 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 |

Garbage in, garbage out . . .



“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 site to a new directory on your machine.
- ② Open R Studio and start a new Project, in the new directory.
- 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
 - explore and 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 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/431-2018-data>

and also at

<https://github.com/THOMASELOVE/431-2018/tree/master/slides/class02>

- We'll grab just that data file, for now, by clicking on it, selecting Raw, and saving the resulting **love-age-guess-2018.csv** file to our computer.
- Specifically, we'll save it to a new directory called **431class02**.

Class 2 Github page - love-age-guess-2018.csv file

The screenshot shows a web browser window with the URL <https://github.com/THOMASELOVE/431-2018/tree/master/slides/class02>. The GitHub interface includes a search bar, navigation links for Pull requests, Issues, Marketplace, and Explore, and a sidebar with links for Code, Issues (0), Pull requests (0), Projects (0), Wiki, Insights, and Settings. The main content area displays a commit from 'THOMASELOVE' adding data sets and a template for demonstrations. Below this, there are four files listed: '431-r-template.Rmd', 'README.md', 'love-age-guess-2018.csv', and 'surveyday1_2018.csv', each with a brief description indicating they were added for demonstrations.

431-2018/slides/class02

GitHub, Inc. [US] | https://github.com/THOMASELOVE/431-2018/tree/master/slides/class02

Apps Bookmarks NY Times 538 Slate The Athletic sAm MyFitnessPal Ringer DS Bill James BP Fangraphs

Search or jump to... / Pull requests Issues Marketplace Explore

THOMASELOVE / 431-2018

Unw...

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

Branch: master 431-2018 / slides / class02 / Create

THOMASELOVE Added data sets and template for demonstrations

..

431-r-template.Rmd Added data sets and template for demonstrations

README.md Update README.md

love-age-guess-2018.csv Added data sets and template for demonstrations

surveyday1_2018.csv Added data sets and template for demonstrations

Right-click Raw to download (just) this file, into a 431class02 directory, please.

THOMASELOVE / 431-2018

Unwatch 2 Star 2 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

Branch: master 431-2018 / slides / class02 / love-age-guess-2018.csv Find file Copy path

THOMASELOVE Added data sets and template for demonstrations 0f17e45 3 minutes ago

1 contributor

53 lines (52 sloc) 631 Bytes Raw Blame History

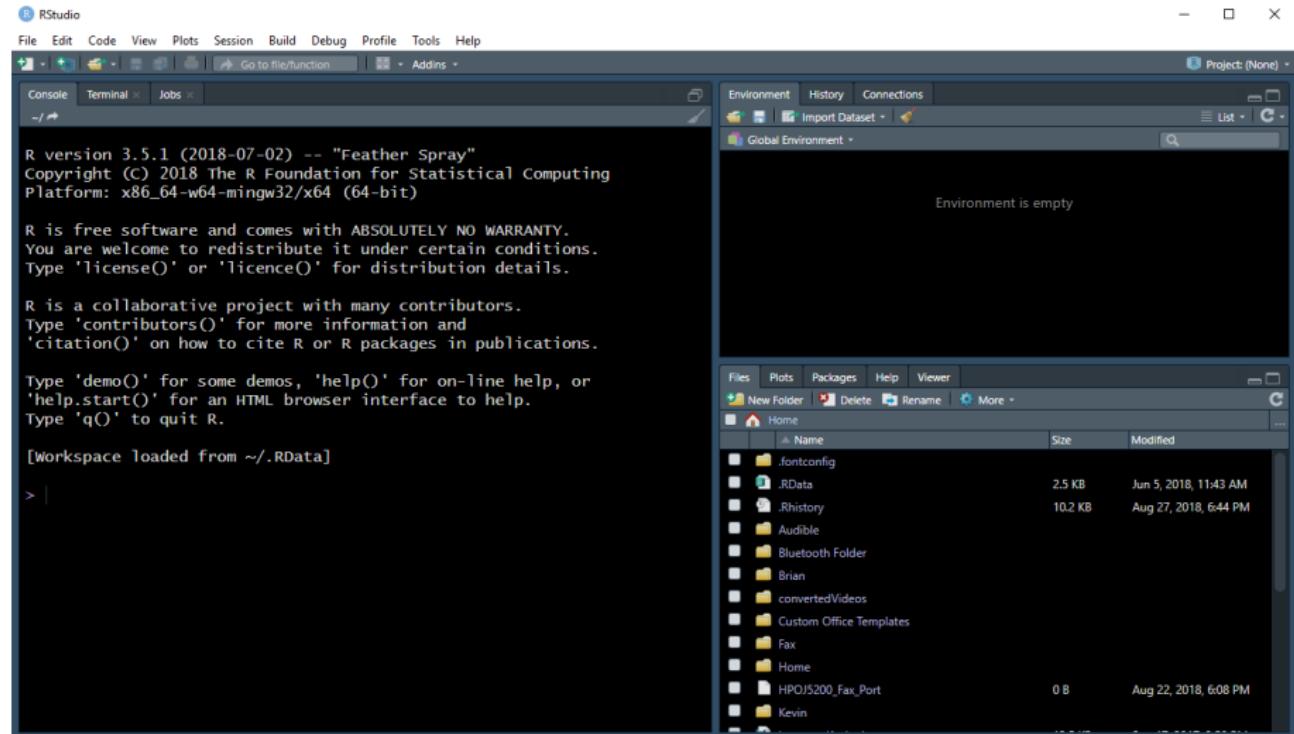
Search this file...

| | subject | age1 | age2 |
|---|---------|------|------|
| 1 | S-01 | 52 | 52 |
| 2 | S-02 | 50 | 49 |
| 3 | S-03 | 50 | NA |
| 4 | S-04 | 48 | 48 |

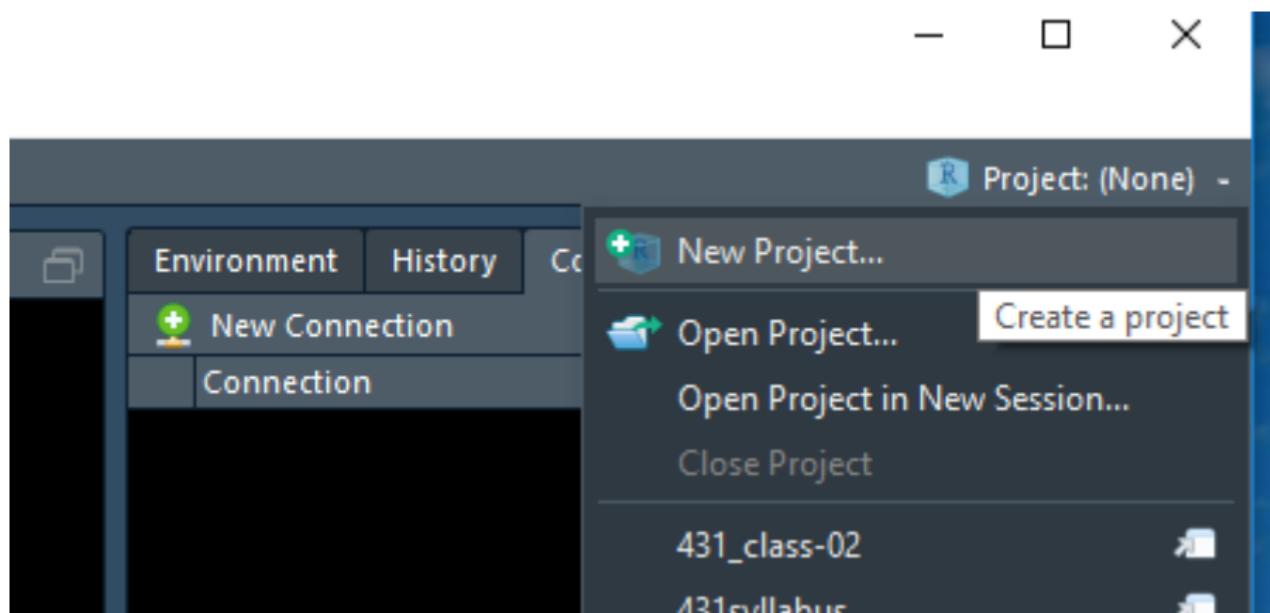
What the file looks like

| | A | B | C |
|---|---------|------|------|
| 1 | subject | age1 | age2 |
| 2 | S-01 | 52 | 52 |
| 3 | S-02 | 50 | 49 |
| 4 | S-03 | 50 | NA |
| 5 | S-04 | 48 | 48 |
| 6 | S-05 | 52 | 61 |
| 7 | S-06 | 54 | 54 |
| 8 | S-07 | 45 | 40 |
| 9 | S-08 | 51 | 49 |

Open R Studio and start a New Project



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



Create Project



New Directory

Start a project in a brand new working directory



Existing Directory

Associate a project with an existing working directory



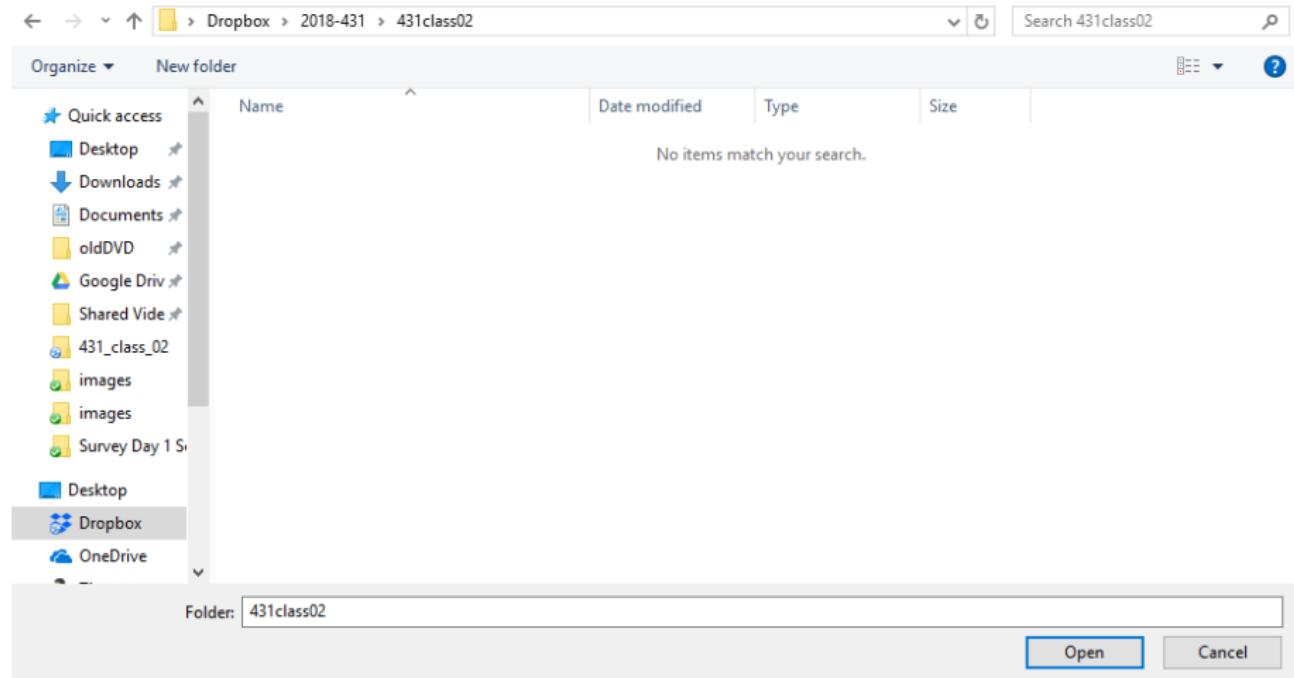
Version Control

Checkout a project from a version control repository



Cancel

R Choose Directory



 Back

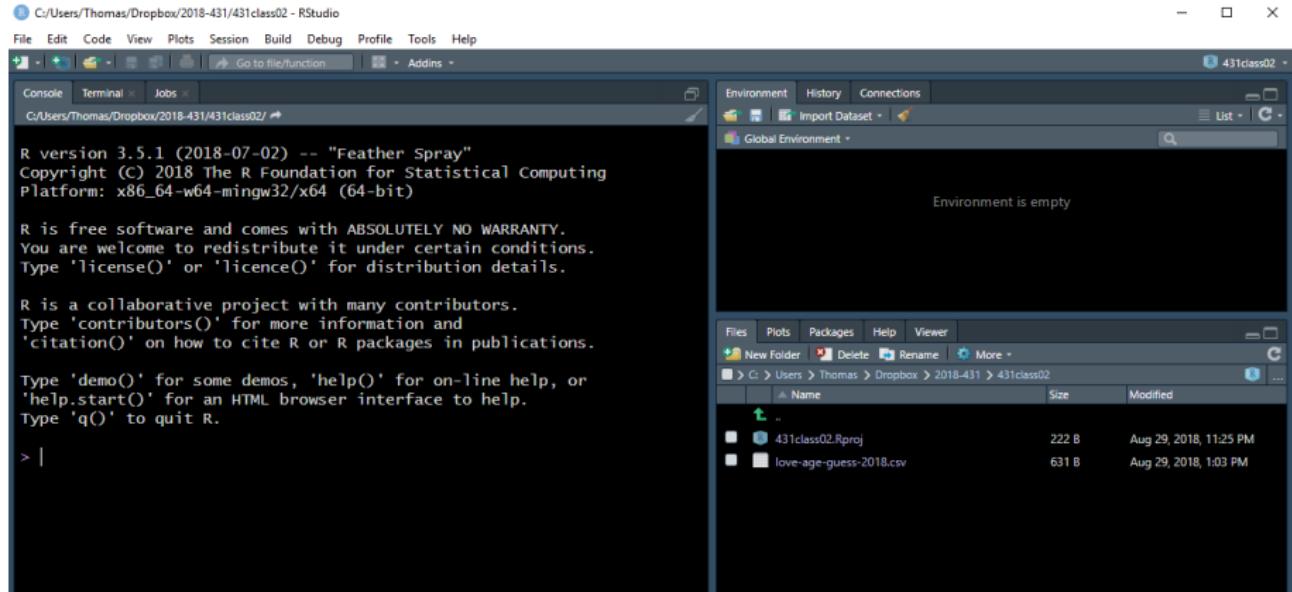
Create Project from Existing Directory



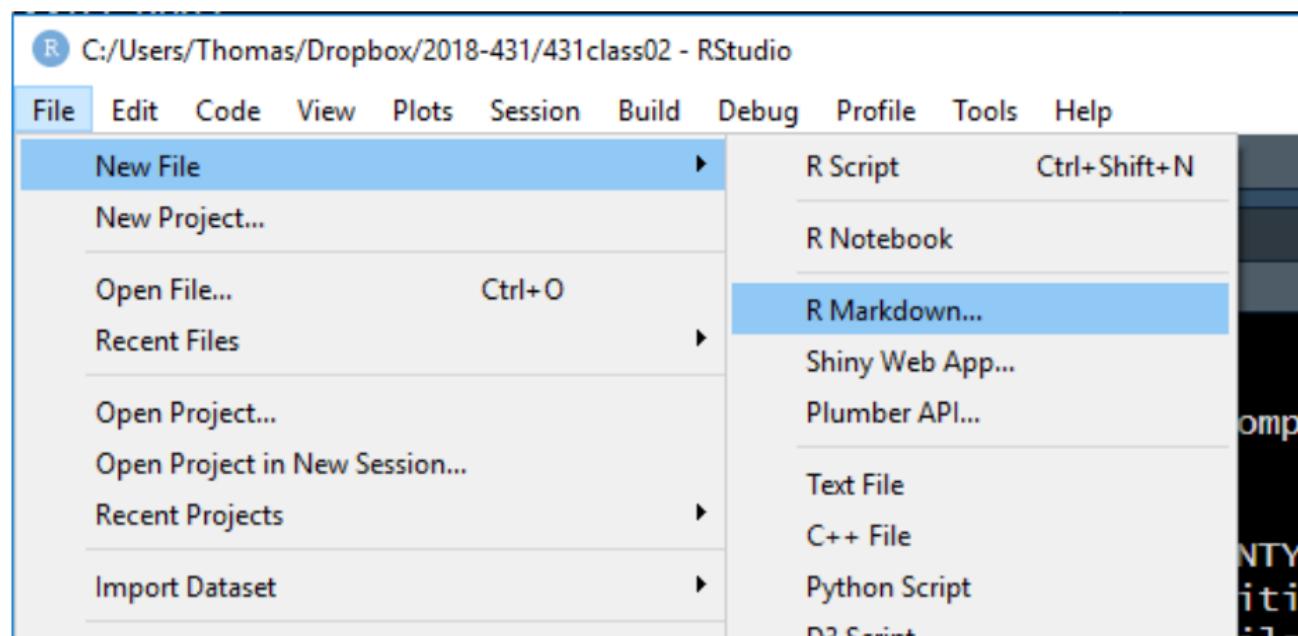
Project working directory:

[Browse...](#) Open in new session[Create Project](#)[Cancel](#)

Now we are in our new Project space.

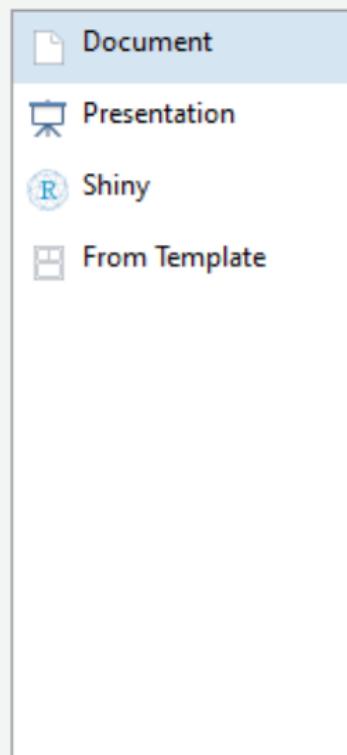


Start a new R Markdown file



Fill in your name and title for the analysis

New R Markdown



Title: Class 2 Age Guess Analysis

Author: Thomas E. Love

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).

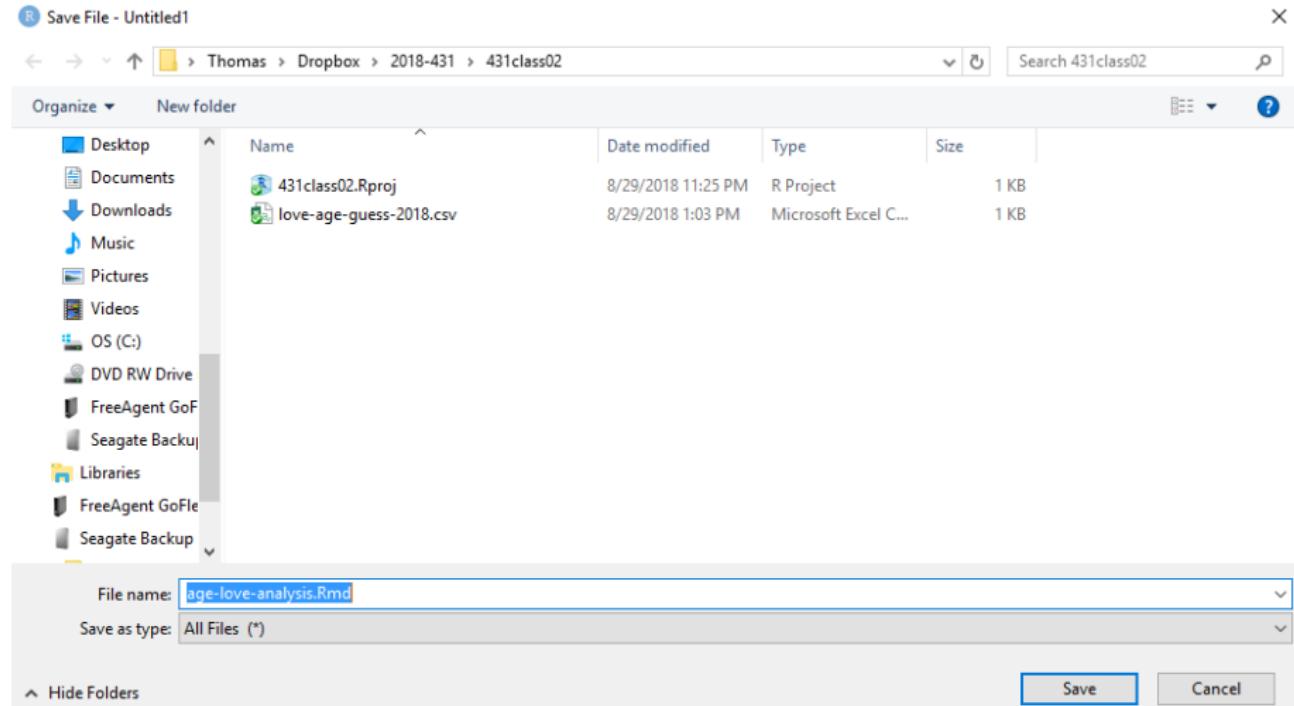
A sample Markdown file is generated. Let's knit it into an HTML file.

The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays the R Markdown code for "Class 2 Age Guess Analysis".
- Global Environment:** Shows that the environment is empty.
- File Browser:** Lists files in the project directory: "431class02.Rproj" (222 B, modified Aug 29, 2018, 11:25 PM) and "love-age-guess-2018.csv" (631 B, modified Aug 29, 2018, 1:03 PM).
- R Console:** Displays the R startup message and license information.

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "8/29/2018"  
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  
2M 1 Class 2 Age Guess Analysis : R Markdown  
Console Terminal Jobs  
C:/Users/Thomas/Dropbox/2018-431/431class02/  
  
R version 3.5.1 (2018-07-02) -- "Feather Spray"  
Copyright (C) 2018 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.
```

Save your work as age-love-analysis.Rmd



Now knit the R Markdown file.

The screenshot shows the RStudio interface with the following details:

- Title Bar:** C:/Users/Thomas/Dropbox/2018-431/431class02 - RStudio
- Menu Bar:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help
- Toolbar:** Includes icons for New, Open, Save, Print, Go to file/function, and Addins.
- Code Editor:** Displays the content of 'age-love-analysis.Rmd'. The code includes YAML front matter, R code chunks, and a Markdown section.

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "8/29/2018"  
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>.  
2:1 # Class 2 Age Guess Analysis
```
- Console:** Shows the R command 'C:/Users/Thomas/Dropbox/2018-431/431class02/' followed by the R version information.
- Status Bar:** R Markdown

The result - a web file!

C:/Users/Thomas/Dropbox/2018-431/431class02/age-love-analysis.html

age-love-analysis.html

Open in Browser

Find

Class 2 Age Guess Analysis

Thomas E. Love

8/29/2018

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:

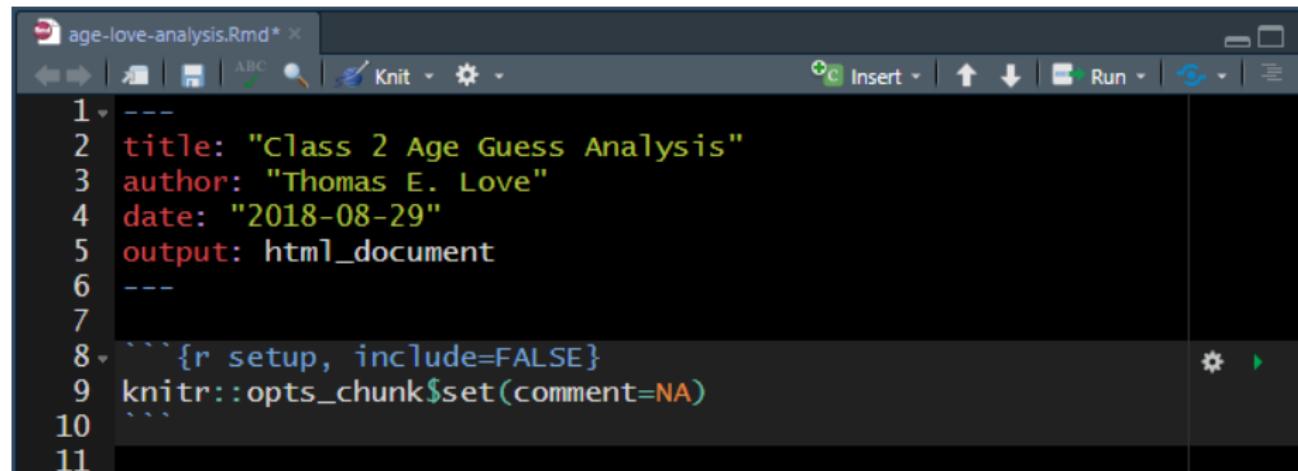
```
summary(cars)
```

```
##      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:

Edit the file to change the date and setup code



The screenshot shows the RStudio interface with an R Markdown file open. The title bar says "age-love-analysis.Rmd*". The code editor contains the following R Markdown code:

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "2018-08-29"  
5 output: html_document  
6 ---  
7  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(comment=NA)  
10```  
11
```

Insert a new “chunk” of R code

The screenshot shows the RStudio interface. The menu bar at the top includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu bar is a toolbar with various icons. The main workspace shows an R Markdown file named "age-love-analysis.Rmd". The code in the file is:

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "2018-08-29"  
5 output: html_document  
---  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(comment=NA)  
10  
11  
12
```

A context menu is open over the line of code starting with "8", specifically over the "```{r setup, include=FALSE}" line. The menu is titled "Insert" and contains options for inserting code in R, Bash, D3, Python, Rcpp, SQL, and Stan. The "R" option is highlighted, and a tooltip indicates "Insert a new R chunk".

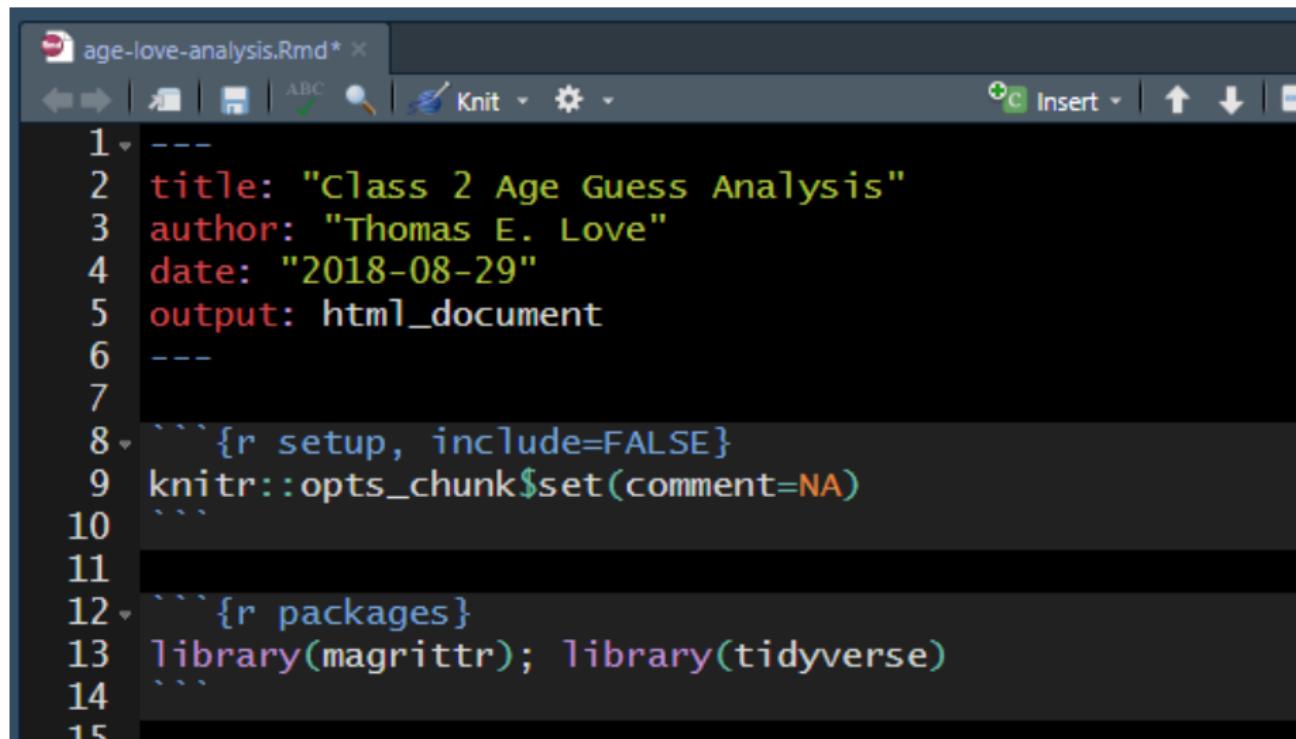
A blank “chunk”

The screenshot shows the RStudio interface with the following details:

- Title Bar:** The file is titled "age-love-analysis.Rmd*".
- Toolbar:** Includes icons for back, forward, search, and various document operations.
- Text Editor:** Displays the R Markdown code. Lines 1 through 6 define the document's metadata. Lines 8 and 9 show a blank R chunk setup. Lines 12 and 13 show another blank R chunk.

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "2018-08-29"  
5 output: html_document  
6 ---  
7  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(comment=NA)  
10  
11  
12 ```{r}  
13  
14  
15
```

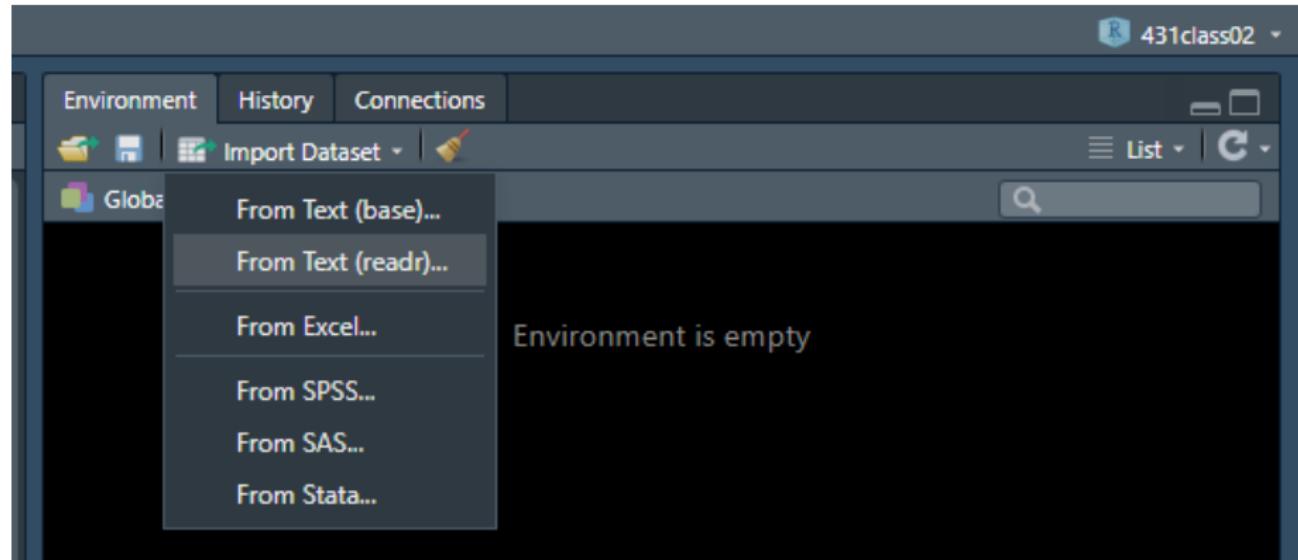
Load up two packages in R (should be installed already)



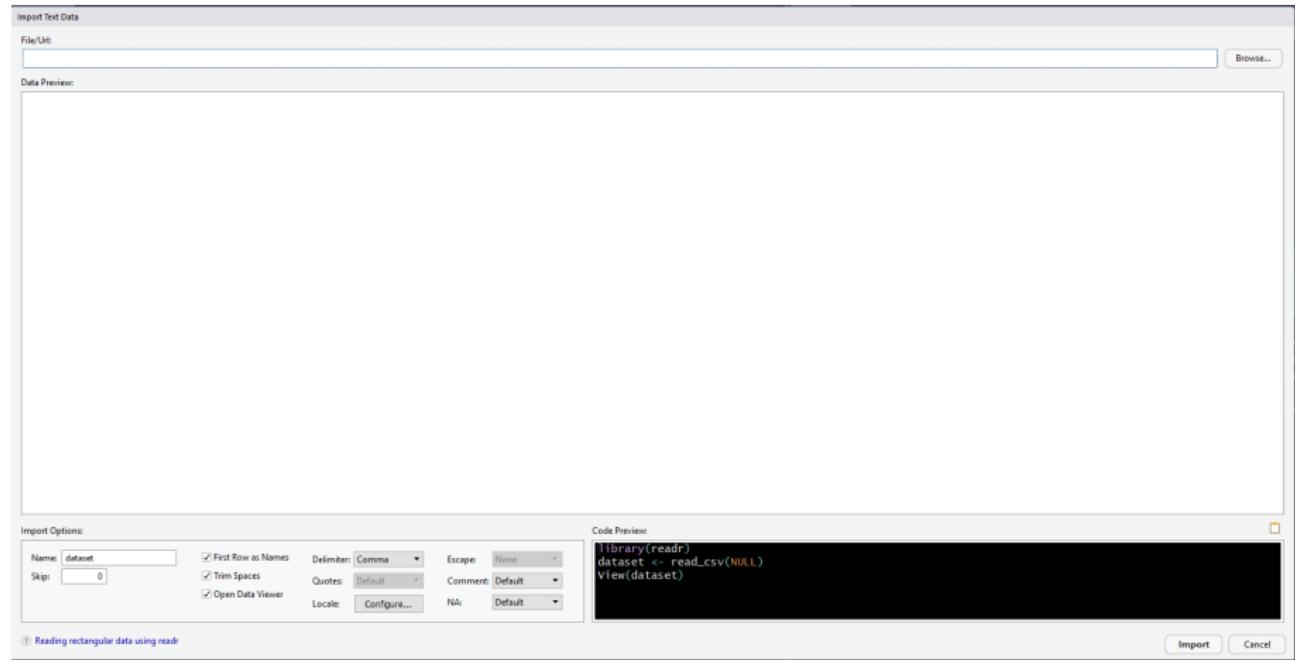
The screenshot shows the RStudio interface with an R Markdown file open. The title bar says "age-love-analysis.Rmd*". The toolbar includes icons for back, forward, search, and knit. The main code editor area contains the following R Markdown code:

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "2018-08-29"  
5 output: html_document  
6 ---  
7  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(comment=NA)  
10  
11  
12 ```{r packages}  
13 library(magrittr); library(tidyverse)  
14  
15
```

Import the .csv data set



Import window



After we make our choices...

Import Text Data

File/Url:

C:/Users/Thomas/Dropbox/2018-431/431class02/love-age-guess-2018.csv

Data Preview:

| subject (character) | age1 (integer) | age2 (integer) |
|------------------------|-------------------|-------------------|
| S-01 | 52 | 52 |
| S-02 | 50 | 49 |
| S-03 | 50 | NA |
| S-04 | 48 | 48 |
| S-05 | 52 | 61 |
| S-06 | 54 | 54 |
| S-07 | 45 | 40 |
| S-08 | 51 | 49 |

Previewing first 50 entries.

Import Options:

Name: First Row as Names Trim Spaces Open Data Viewer Delimiter: Escape: Quotes: Comment: Locale: NA:

Skip:

Code Preview:

```
library(readr)
love_age_guess_2018 <- read_csv("love-age-guess-2018.csv")
View(love_age_guess_2018)
```

⑦ Reading rectangular data using readr

Result (note code in Console)

The screenshot shows the RStudio interface with the following components:

- File**, **Edit**, **Code**, **View**, **Plots**, **Session**, **Build**, **Debug**, **Profile**, **Tools**, **Help** menu bar.
- Environment** pane: Shows a data frame named "love_age_guess_2018" with 51 observations and 3 variables: subject, age1, and age2. The data is as follows:

| subject | age1 | age2 |
|---------|------|------|
| 1 S-01 | 52 | 52 |
| 2 S-02 | 50 | 49 |
| 3 S-03 | 50 | NA |
| 4 S-04 | 48 | 48 |
| 5 S-05 | 52 | 61 |
| 6 S-06 | 54 | 54 |
| 7 S-07 | 45 | 40 |
| 8 S-08 | 51 | 49 |
| 9 S-09 | 53 | 55 |
| 10 S-10 | 55 | 49 |
| 11 S-11 | 42 | 45 |
| 12 S-12 | 52 | 50 |

- Data** pane: Shows the same data frame "love_age_guess_2018" with the description "51 obs. of 3 variables".
- Files** pane: Shows the project structure: 431class02.Rproj, love-age-guess-2018.csv, age-love-analysis.Rmd, age-love-analysis_files, and age-love-analysis.html. The files were modified on Aug 29, 2018, at various times.
- Console** pane: Displays the R code used to load and view the data.

```
> library(readr)
> love_age_guess_2018 <- read_csv("love-age-guess-2018.csv")
Parsed with column specification:
cols(
  subject = col_character(),
  age1 = col_integer(),
  age2 = col_integer()
)
> View(love_age_guess_2018)
```

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

The screenshot shows the RStudio interface with the following details:

- Title Bar:** Shows two tabs: "age-love-analysis.Rmd*" and "love_age_guess_2018".
- Toolbar:** Includes icons for back, forward, search, Knit, and settings.
- Code Editor:** Displays an R Markdown script with numbered lines from 1 to 20. Lines 1 through 6 are YAML front matter. Lines 8 through 10 are setup code for knitr. Lines 12 through 14 are for loading packages. Line 16 starts a section for loading data. Line 17 reads a CSV file named "love-age-guess-2018.csv" into a variable "love_2018". Lines 19 and 20 start sections for viewing the data.

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "2018-08-29"  
5 output: html_document  
6 ---  
7  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(comment=NA)  
10```  
11  
12 ```{r packages}  
13 library(magrittr); library(tidyverse)  
14```  
15  
16 ```{r load_data}  
17 love_2018 <- read_csv("love-age-guess-2018.csv")  
18```  
19  
20 ```{r view data}
```

Running the code so we can see results

The screenshot shows the RStudio interface with an R Markdown file named "age-love-analysis.Rmd" open. The code editor displays the following R Markdown code:

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "2018-08-29"  
5 output: html_document  
6 ---  
7  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(comment=NA)  
10```  
11  
12 ```{r packages}  
13 library(magrittr); library(tidyverse)  
14```  
15  
16 ```{r load_data}  
17 love_2018 <- read_csv("love-age-guess-2018.csv")  
18```  
19  
20 ```{r view data}  
21 love_2018  
22```
```

A context menu is open over the code editor, specifically over the line `love_2018 <- read_csv("love-age-guess-2018.csv")`. The menu is titled "Run" and contains the following options:

- Run Selected Line(s) Ctrl+Enter
- Run Current Chunk Ctrl+Shift+R Run the current line or selection
- Run Next Chunk Ctrl+Alt+Down
- Run Setup Chunk
- Run Setup Chunk Automatically (checked)
- 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 an R Markdown file open. The code editor displays the following R code:

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "2018-08-29"  
5 output: html_document  
6 ---  
7  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(comment=NA)  
10  
11  
12 ```{r packages}  
13 library(magrittr); library(tidyverse)  
14  
15  
16 ```{r load_data}  
17 love_2018 <- read_csv("love-age-guess-2018.csv")  
18  
19  
20 ```{r view data}  
21 love_2018  
22  
23
```

A context menu is open at the bottom right of the code editor, with the option "Run All Chunks Above" highlighted.

Running the first three chunks of code

The screenshot shows the RStudio interface with the following details:

- Title Bar:** Shows "age-love-analysis.Rmd" and "love_age_guess_2018".
- Toolbar:** Includes "Insert", "Run", and other standard RStudio icons.
- Code Editor:** Displays the R Markdown code. Lines 1 through 18 are visible, with line 15 being the current active line.
- Output Panel:** Located below the code editor, it displays the results of the executed code, including package loading information and a column specification for a CSV file.

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "2018-08-29"  
5 output: html_document  
6 ---  
7  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(comment=NA)  
10 ...  
11  
12 ```{r packages}  
13 library(magrittr); library(tidyverse)  
14 ...  
15 -- Attaching packages --> tidyverse 1.2.1 --  
16 ✓ ggplot2 3.0.0    ✓ purrr   0.2.5  
17 ✓ tibble  1.4.2    ✓ dplyr   0.7.6  
18 ✓ tidyr   0.8.1    ✓ stringr 1.3.1  
19 ✓ ggplot2 3.0.0    ✓ forcats 0.3.0  
-- Conflicts --> tidyverse_conflicts()  
20 ✘ tidyr::extract() masks magrittr::extract()  
21 ✘ dplyr::filter()  masks stats::filter()  
22 ✘ dplyr::lag()     masks stats::lag()  
23 ✘ purrr::set_names() masks magrittr::set_names()  
24  
25 ...  
26  
27 love_2018 <- read_csv("love-age-guess-2018.csv")  
28 ...  
29  
30 Parsed with column specification:  
31 cols(  
32   subject = col_character(),  
33   age1 = col_integer(),  
34   ...)
```

Now, let's look at the data

```
```{r view data}
love_2018
```

```

The love_2018 tibble

```
19  
20 - ````{r view data}  
21 love_2018  
22 ````
```

| subject | age1 | age2 |
|---------|------|------|
| S-01 | 52 | 52 |
| S-02 | 50 | 49 |
| S-03 | 50 | NA |
| S-04 | 48 | 48 |
| S-05 | 52 | 61 |
| S-06 | 54 | 54 |
| S-07 | 45 | 40 |
| S-08 | 51 | 49 |
| S-09 | 53 | 55 |
| S-10 | 55 | 49 |

1-10 of 51 rows

Previous 2 3 4 5 6 Next

Add some documentation, mixing text with code

```
19  
20 ``{r view data}  
21 love_2018  
22 ````
```

| subject | age1 | age2 |
|---------|------|------|
| S-01 | 52 | 52 |
| S-02 | 50 | 49 |
| S-03 | 50 | NA |
| S-04 | 48 | 48 |
| S-05 | 52 | 61 |
| S-06 | 54 | 54 |
| S-07 | 45 | 40 |
| S-08 | 51 | 49 |
| S-09 | 53 | 55 |
| S-10 | 55 | 49 |

1-10 of 51 rows

Previous 1 2 3 4 5 6 Next

23

24 Our data set has `r nrow(love_2018)` rows, and `r

Let's make a histogram

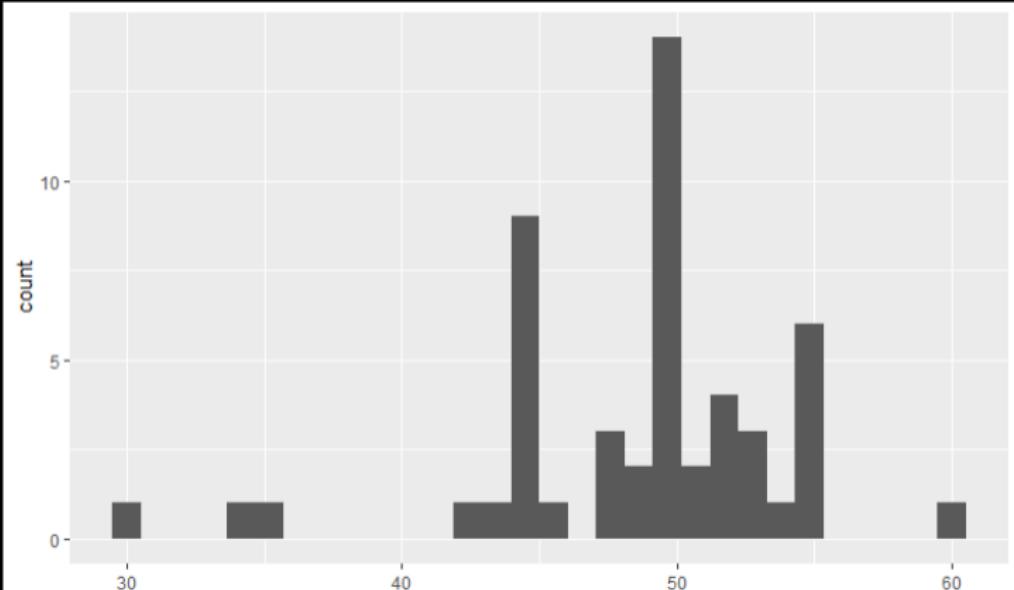
```
25  
26 # Build a Histogram of the First Guesses  
27  
28 ````{r}  
29 ggplot(data = love_2018, aes(x = age1)) +  
30   geom_histogram()  
31 ````  
32
```

The result. Can we do better?

```
25  
26 # Build a Histogram of the First Guesses  
27  
28 ```{r}  
29 ggplot(data = love_2018, aes(x = age1)) +  
30 ...  
31
```

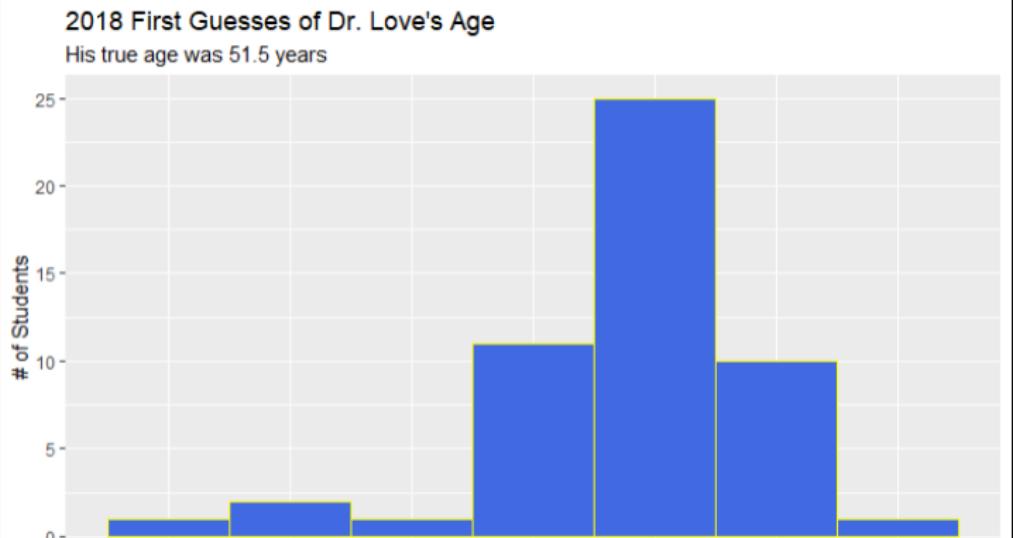


`stat_bin` using `bins = 30`. Pick better value with `binwidth`.



A second attempt at the histogram

```
32  
33 ## A Nicer Histogram  
34  
35 ````{r histogram2}  
36 ggplot(data = love_2018, aes(x = age1)) +  
37   geom_histogram(bins = 7, fill = "royalblue", col = "yellow") +  
38   labs(x = "First Guess of Dr. Love's Age on 2018-08-28",  
39         y = "# of Students",  
40         title = "2018 First Guesses of Dr. Love's Age",  
41         subtitle = "His true age was 51.5 years")  
42 ...
```



A numerical summary of the data

```
44 # A numerical summary
45
46 ````{r}
47 summary(love_2018)
48

      subject           age1          age2
Length:51      Min.   :30.00    Min.   :38.00
Class :character 1st Qu.:45.00  1st Qu.:48.00
Mode  :character Median :50.00  Median :50.00
                  Mean   :48.78  Mean   :50.37
                  3rd Qu.:52.00  3rd Qu.:53.00
                  Max.   :60.00  Max.   :65.00
                  NA's    :2
```

49
50

Calculating and Summarizing the Errors

```
49  
50 # Better First or Second Guess?  
51  
52 ````{r calculate-errors}  
53 love_2018 <- love_2018 %>%  
54     mutate(error1 = abs(age1 - 51.5),  
55             error2 = abs(age2 - 51.5))  
56  
57 summary(love_2018)  
58 ````
```

What do these summaries suggest?

```
49  
50 # Better First or Second Guess?  
51  
52 ``-{r calculate-errors}  
53 love_2018 <- love_2018 %>%  
54     mutate(error1 = abs(age1 - 51.5),  
55         error2 = abs(age2 - 51.5))  
56  
57 summary(love_2018)  
58  
59
```

| subject | age1 | age2 | error1 | error2 |
|------------------|---------------|---------------|----------------|----------------|
| Length:51 | Min. :30.00 | Min. :38.00 | Min. : 0.500 | Min. : 0.500 |
| Class :character | 1st Qu.:45.00 | 1st Qu.:48.00 | 1st Qu.: 1.500 | 1st Qu.: 1.500 |
| Mode :character | Median :50.00 | Median :50.00 | Median : 2.500 | Median : 3.500 |
| | Mean :48.78 | Mean :50.37 | Mean : 4.225 | Mean : 4.194 |
| | 3rd Qu.:52.00 | 3rd Qu.:53.00 | 3rd Qu.: 6.500 | 3rd Qu.: 5.500 |
| | Max. :60.00 | Max. :65.00 | Max. :21.500 | Max. :13.500 |
| | NA's :2 | | | NA's :2 |

Build a scatterplot to compare the errors

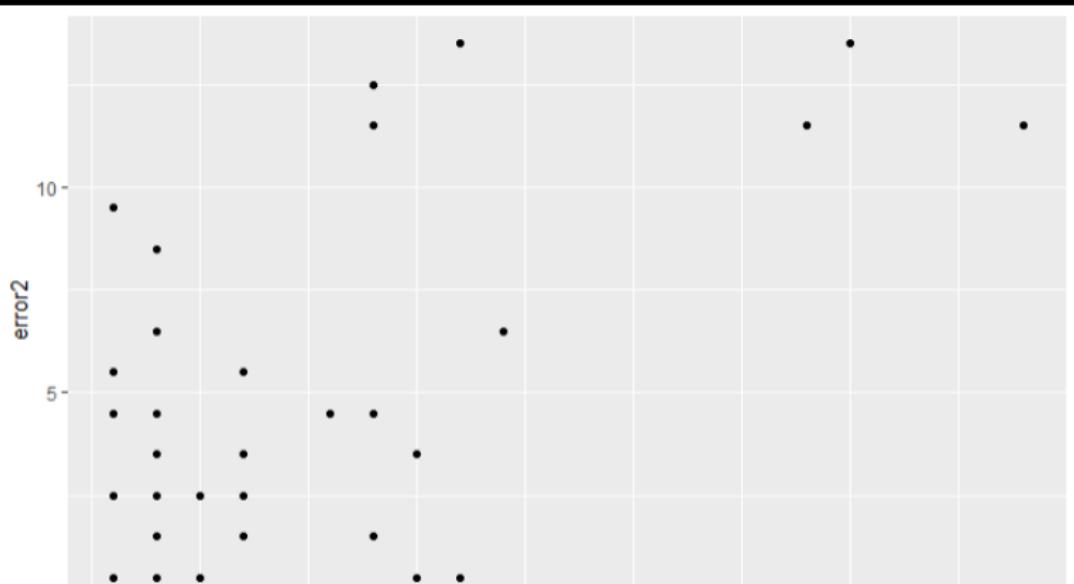
```
59  
60 # Compare the Guesses  
61  
62 ``{r guess1vs2}  
63 ggplot(data = love_2018, aes(x = error1, y = error2)) +  
64   geom_point()  
65 ...  
66
```

The Result

```
60 # Compare the Guesses  
61  
62 ``{r guess1vs2}  
63 ggplot(data = love_2018, aes(x = error1, y = error2)) +  
64   geom_point()  
65
```



Removed 2 rows containing missing values (geom_point).



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

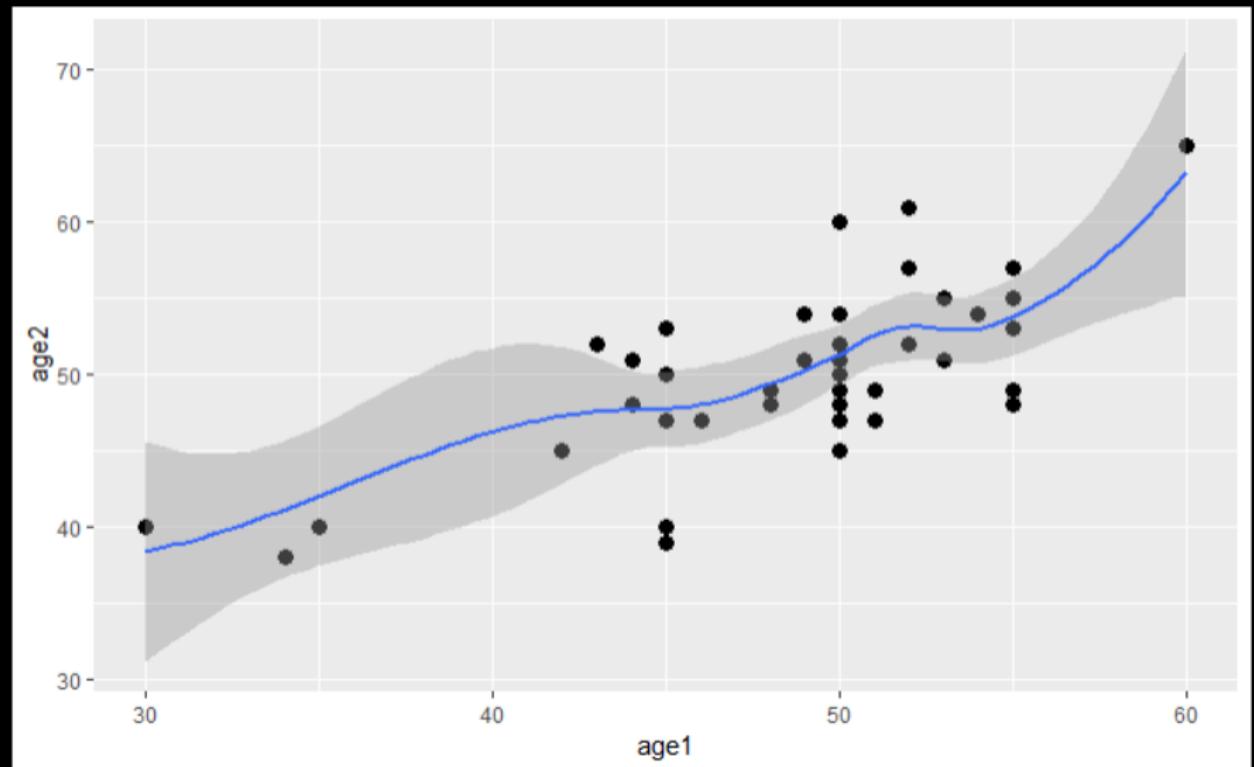
```
66  
67 ## Add a Prediction Model?  
68  
69 ````{r}  
70 ggplot(data = love_2018, aes(x = age1, y = age2)) +  
71   geom_point(size = 3) +  
72   geom_smooth(method = "loess")  
73 ...  
74  
75
```

The Result



Removed 2 rows containing non-finite values (stat_smooth).

Removed 2 rows containing missing values (geom_point).



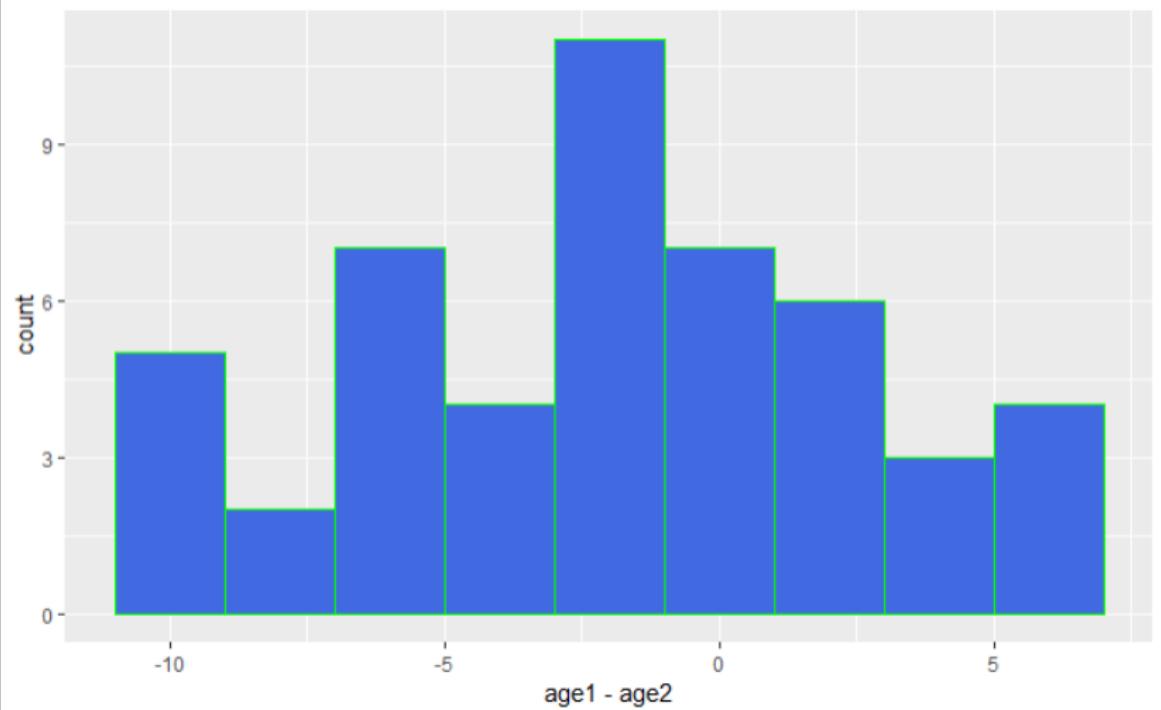
Plot the age1 - age2 differences

```
74  
75 # Plot the (matched) differences  
76  
77 ``-{r}  
78 ggplot(love_2018, aes(x = age1 - age2)) +  
79   geom_histogram(binwidth = 2,  
80                   col = "green", fill = "royalblue")  
81 ...
```

The Result



Removed 2 rows containing non-finite values (stat_bin).



Numerical summary of the age1 - age2 differences

```
82
83 # Numerical Summary of Difference in Age Guesses
84
85 ````{r}
86 love_2018 %$%
87   summary(age1 - age2)
88 ````
```

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|--|---------|---------|--------|--------|---------|-------|------|
| | -10.000 | -5.000 | -2.000 | -1.592 | 2.000 | 7.000 | 2 |

How many people thought I looked younger the second time?

```
90 # How many people thought I looked younger in Guess 2?  
91  
92 `~`{r}  
93 love_2018 %>%  
94   count(age1 - age2 < 0)  
95 ...
```

| age1 - age2 < 0 | n |
|-----------------|----|
| FALSE | 20 |
| TRUE | 29 |
| NA | 2 |

3 rows

T tests - making a statistical inference

```
96  
97 # The Much-Dreaded t test  
98  
99 ````{r}  
100 love_2018 %$%  
101 t.test(age1 - age2)  
102 ````
```

One Sample t-test

```
data: age1 - age2  
t = -2.4612, df = 48, p-value = 0.01749  
alternative hypothesis: true mean is not equal to 0  
95 percent confidence interval:  
-2.8922340 -0.2914395  
sample estimates:  
mean of x  
-1.591837
```

Knit the file into an HTML document

C:/Users/Thomas/Dropbox/2018-431/431class02/age-love-analysis.html

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

Class 2 Age Guess Analysis

Thomas E. Love

2018-08-29

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

```
-- Attaching packages ----- tidyverse 1.2.1 --
```

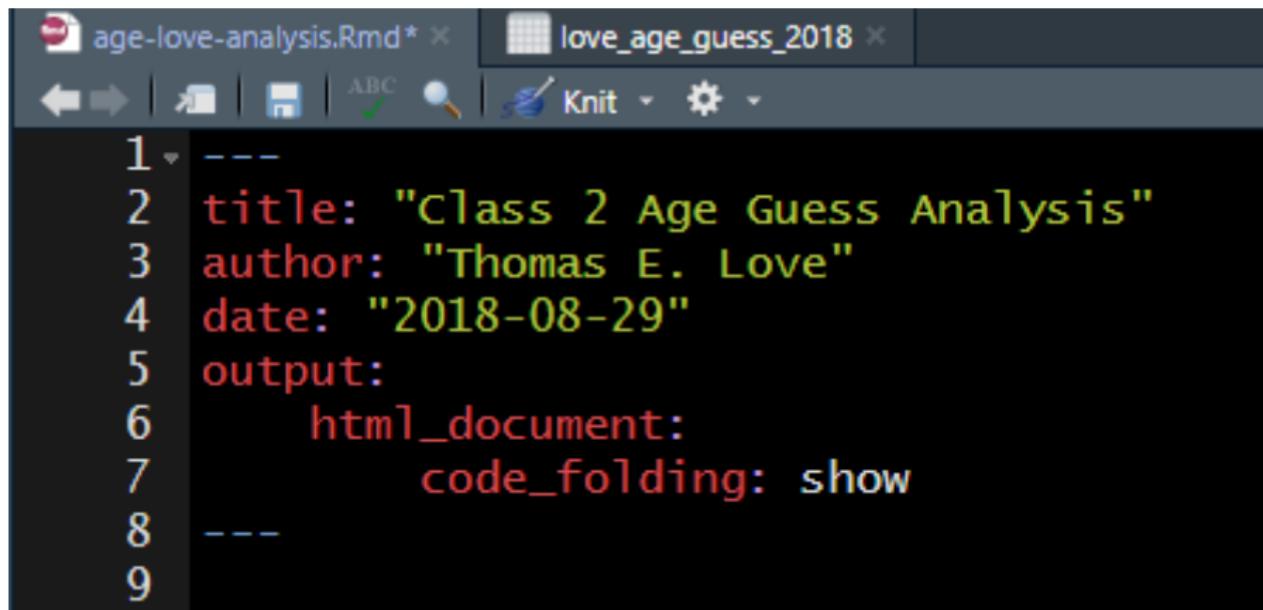
```
v ggplot2 3.0.0      v purrr   0.2.5
v tibble  1.4.2      v dplyr   0.7.6
v tidyr   0.8.1      v stringr 1.3.1
v readr   1.1.1      vforcats 0.3.0
```

```
-- Conflicts ----- tidyverse_conflicts() --
x tidyrr::extract()  masks magrittr::extract()
x dplyr::filter()   masks stats::filter()
x dplyr::lag()       masks stats::lag()
x purrr::set_names() masks magrittr::set_names()
```

```
love_2018 <- read_csv("love-age-guess-2018.csv")
```

```
Parsed with column specification:
cols(
```

Adjust the YAML to fold the code on demand



The screenshot shows the RStudio interface with two tabs open: 'age-love-analysis.Rmd*' and 'love_age_guess_2018'. The 'love_age_guess_2018' tab is active, displaying the following YAML code:

```
1 ---  
2 title: "Class 2 Age Guess Analysis"  
3 author: "Thomas E. Love"  
4 date: "2018-08-29"  
5 output:  
6   html_document:  
7     code_folding: show  
8 ---  
9
```

New, more final, report

C:/Users/Thomas/Dropbox/2018-431/431class02/age-love-analysis.html

age-love-analysis.htm

Open in Browser

Find

Code ▾

Class 2 Age Guess Analysis

Thomas E. Love

2018-08-29

Hide

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

```
-- Attaching packages ----- tidyverse 1.2.1 --
```

```
v ggplot2 3.0.0      v purrr   0.2.5
v tibble  1.4.2      v dplyr    0.7.6
v tidyrr   0.8.1      v stringr 1.3.1
v readr    1.1.1      v forcats 0.3.0
```

```
-- Conflicts ----- tidyverse_conflicts() --
```

```
x tidyrr::extract()  masks magrittr::extract()
x dplyr::filter()    masks stats::filter()
x dplyr::lag()       masks stats::lag()
x purrr::set_names() masks magrittr::set_names()
```

Hide

```
love_2018 <- read_csv("love-age-guess-2018.csv")
```

```
Parsed with column specification:
```

Analyzing the Survey Data - A little challenge

We have data on the site in a file called `surveyday1_2018.csv`. Build a project to study those data.

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

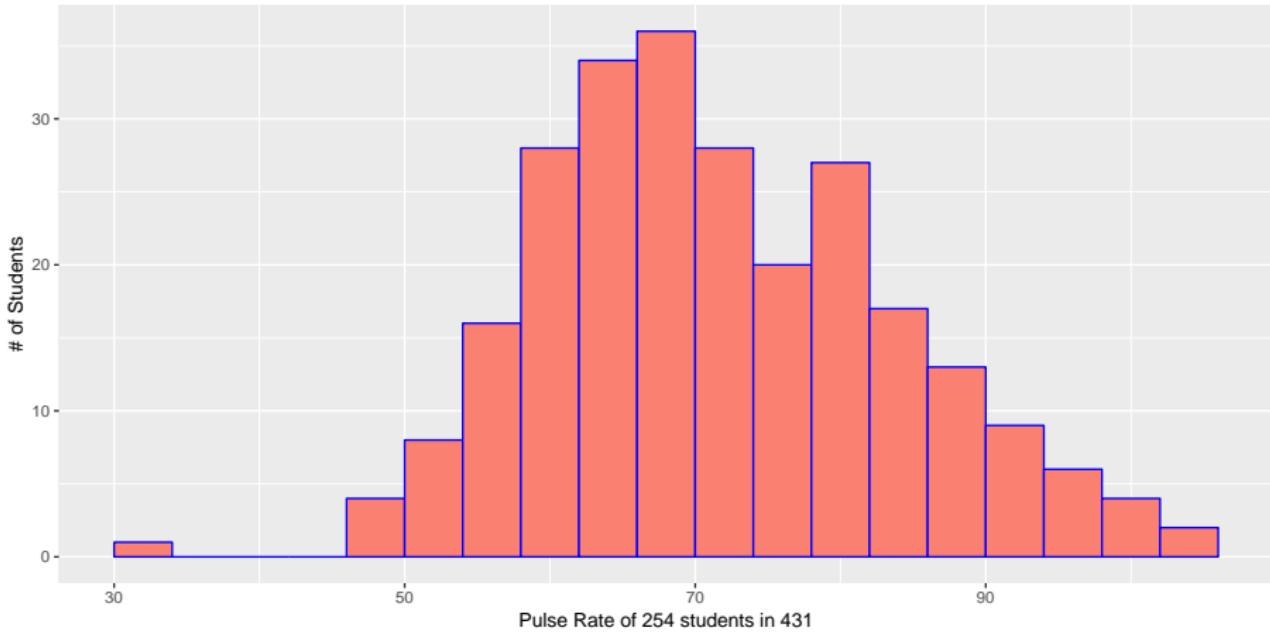
- I'd call my R Markdown file `day1surveyanalysis`

Can you reproduce the following...

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

Pulse Rates of 254 students in 431

One student had a missing pulse value

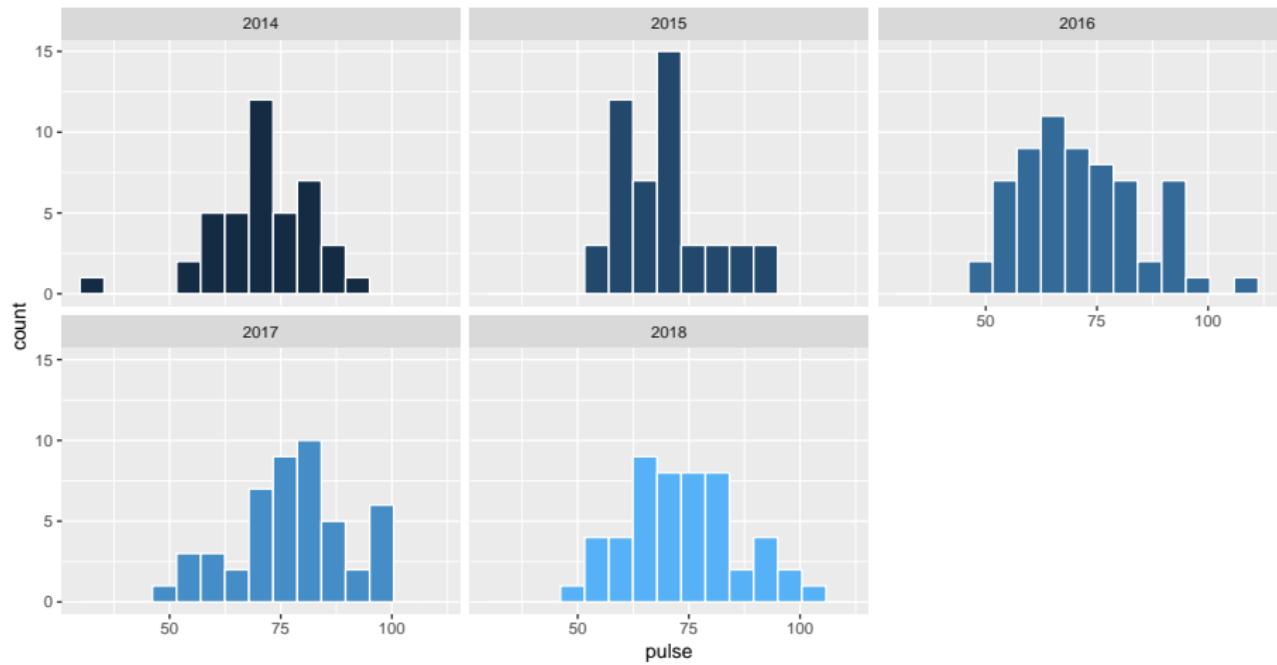


Code for Plot A.

```
ggplot(surv1, aes(x = pulse)) +
  geom_histogram(bins = 20,
                 col = "blue", fill = "salmon") +
  labs(x = "Pulse Rate of 254 students in 431",
       y = "# of Students",
       title = "Pulse Rates of 254 students in 431",
       subtitle = "One student had a missing pulse value")
```

B. Histograms of Pulse Rates, Faceted by Year

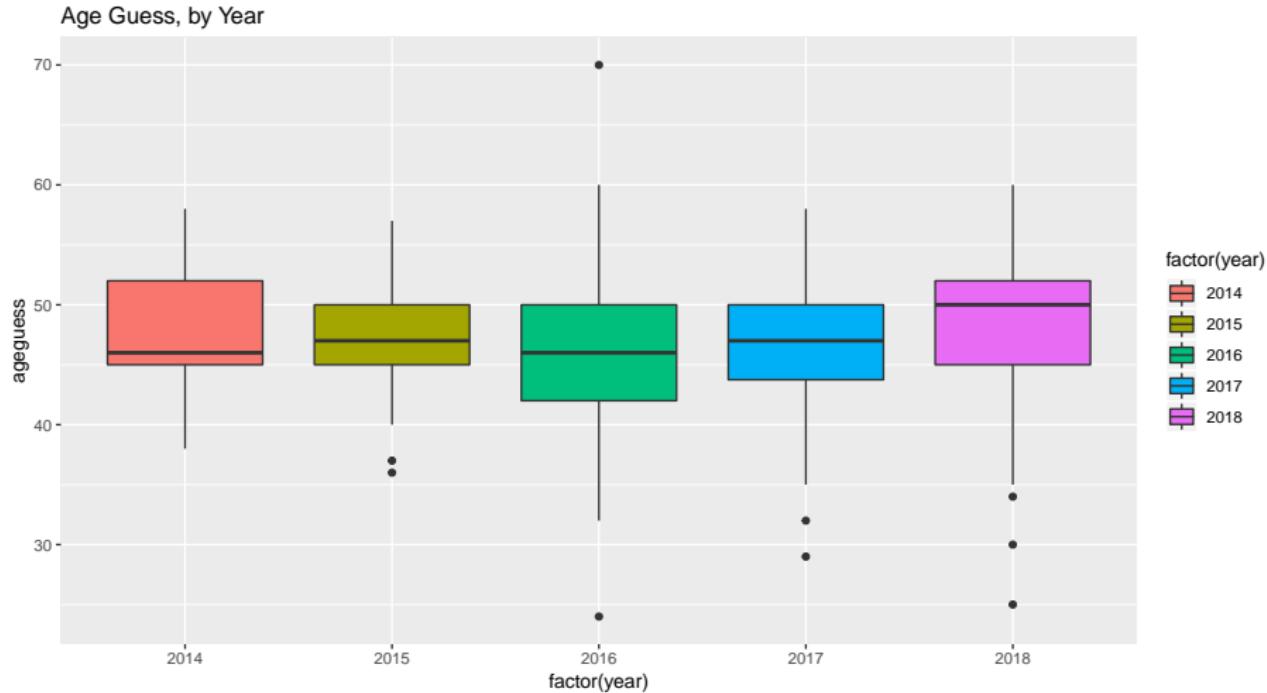
Pulse Rate, by Year



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. Boxplots 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: 5 x 5
  year     n   mean    sd median
  <int> <int> <dbl> <dbl>  <dbl>
1 2014     42  47.3  5.21    46
2 2015     49  47.1  4.62    47
3 2016     64  46.0  7.00    46
4 2017     48  46.5  6.15    47
5 2018     51  48.2  6.47    50
```

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)
  )
```

HELP!!!

Visit our Software page for:

- complete installation instructions
- a document I've written called Getting Started with R, which demonstrates more of these tools.

Don't forget about **431-help at case dot edu**.

TA office hours start Tuesday at 11:30 AM. See the Course Calendar for all the details.

Datacamp has several courses that will help you learn R. Watch for the email invitation.

What's coming up?

- Running more involved analyses in R and R Studio, live
- 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 2017-09-06 (Class 4)
 - Read Silver Introduction and Chapter 1 (about 50 denser pages) by 2017-09-18 (Class 7)
 - Homework 1 is due 2017-09-07 at noon