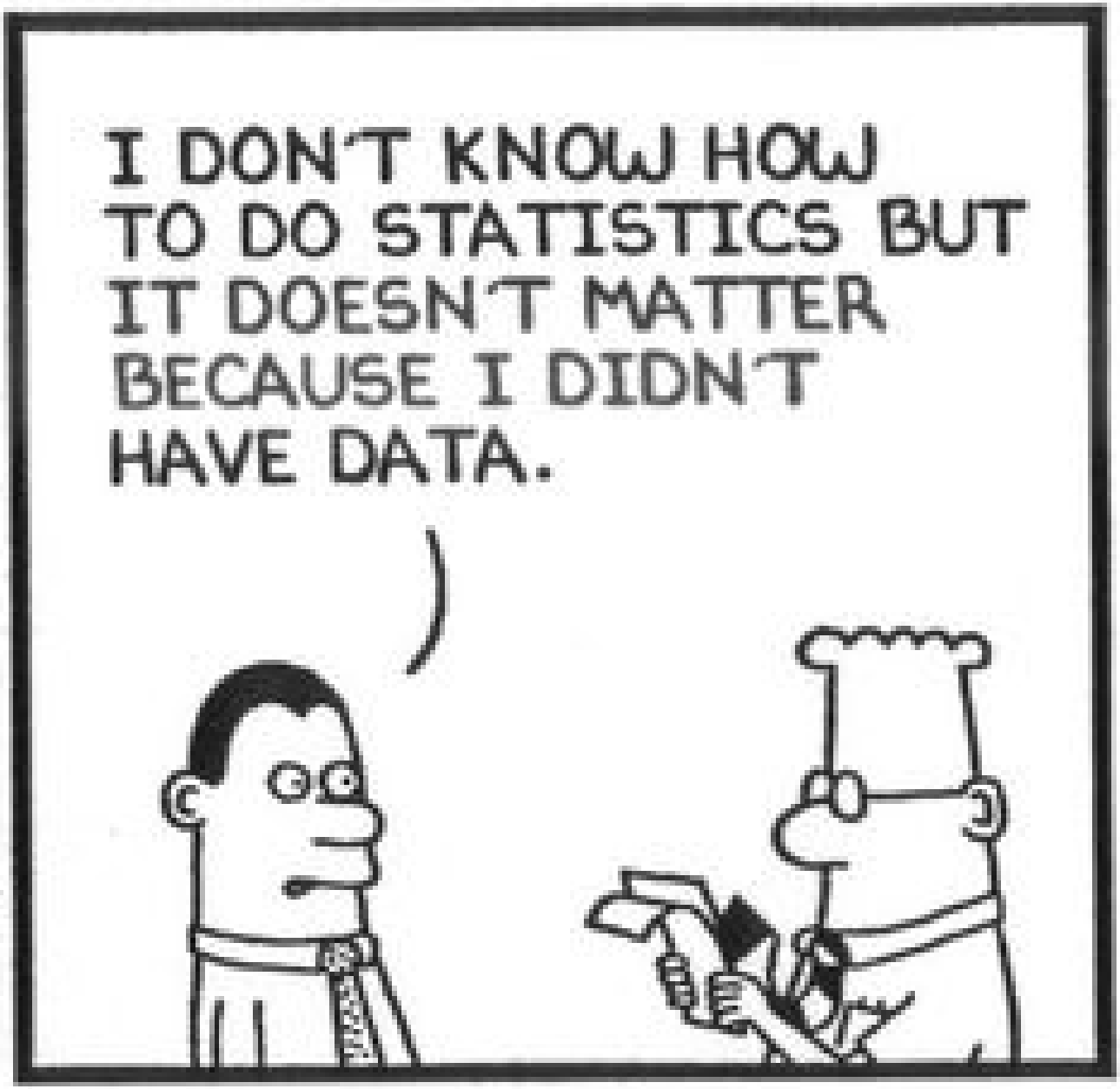
431 Class 02

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.

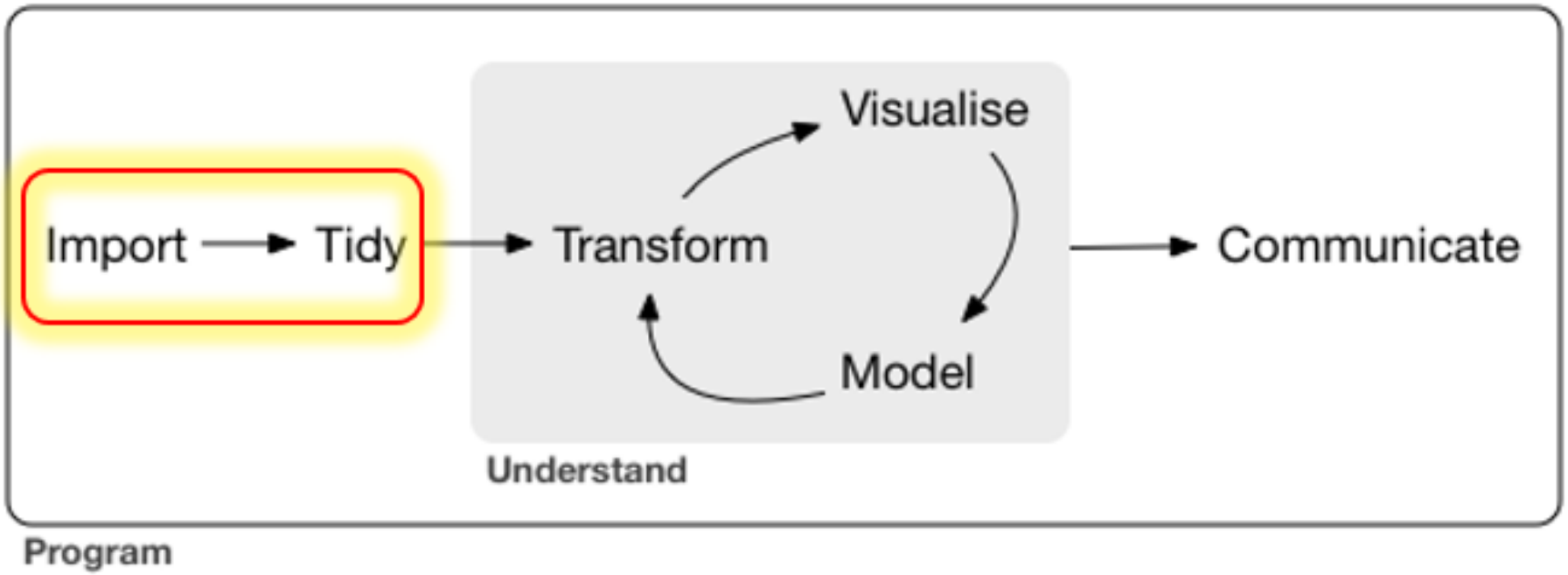
- Tuesdays 11:30 AM to 12:45 PM
- Wednesdays 12 noon to 1:30 PM
- Thursdays 11:30 AM to 12:45 PM, 2:30 - 4 PM, 5:30 - 7 PM
- Fridays 10:30 AM to 12 noon

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 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 Non-Smoker	I used to. Former Smoker	Yes. Smoker
①	②	③

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:		

Right − Left = _____ Right + Left = _____ $\frac{Right-Left}{Right+Left}$ = _____

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	⑤	⑥	⑦

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

Please indicate your agreement with the following statements:

	Strongly Disagree				Strongly Agree
9. I prefer to learn from lectures than to learn from activities.	1	2	3	4	5
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): _____

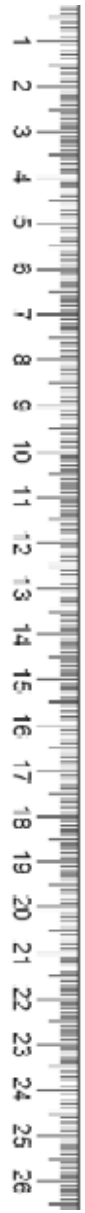
12. Use the ruler provided on the side of this page to measure the span of your right hand (distance from the thumb to the little finger when your fingers are spread apart: _____ cm.

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



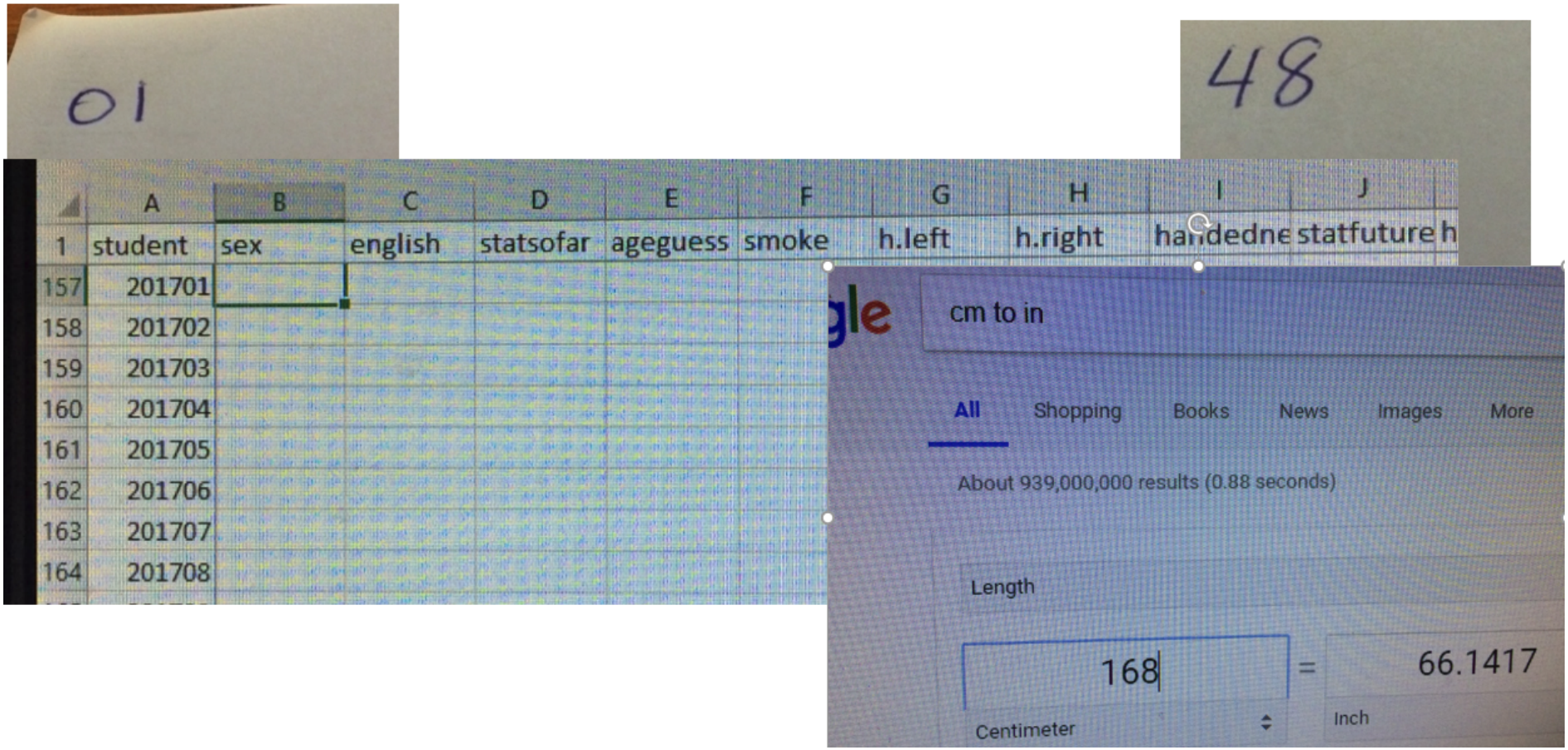
Evaluating some Day 1 Survey variables

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
48
A B C D E F G H I J
1 student sex english statsofar ageguess smoke h.left h.right handedne statfuture h
157 201701
158 201702
159 201703
160 201704
161 201705
162 201706
163 201707
164 201708
cm to in
All Shopping Books News Images More
About 939,000,000 results (0.88 seconds)
Length
168
=
66.1417
Centimeter
Inch

Day 1 Survey

- 51 people completed it Tuesday. Prior counts:

Fall	2018	2017	2016	2015	2014	Total
n	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 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 Non-Smoker	I used to. Former Smoker	Yes. Smoker
①	②	③

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:		

Right − Left = _____ Right + Left = _____ $\frac{Right-Left}{Right+Left}$ = _____

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	⑤	⑥	⑦

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

Please indicate your agreement with the following statements:

	Strongly Disagree				Strongly Agree
9. I prefer to learn from lectures than to learn from activities.	1	2	3	4	5
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 254 surveys caused *no problems* in recording responses?

- Guesses?

Question 1

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

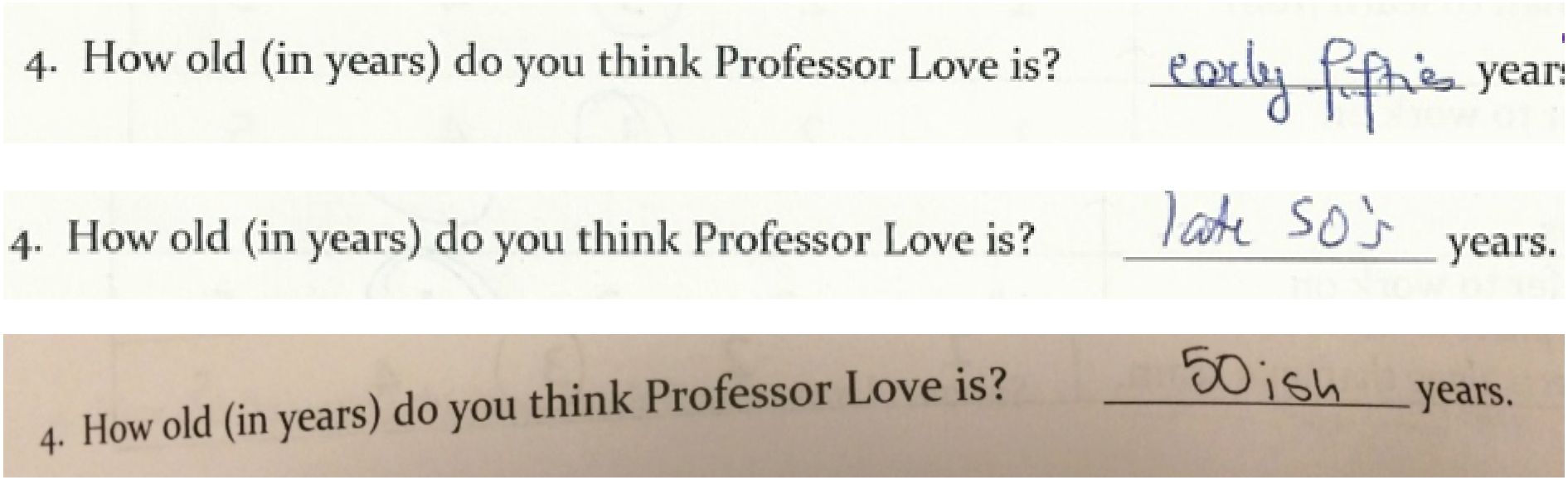
- Guesses?
- 90/254 (34%)

Question 1

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

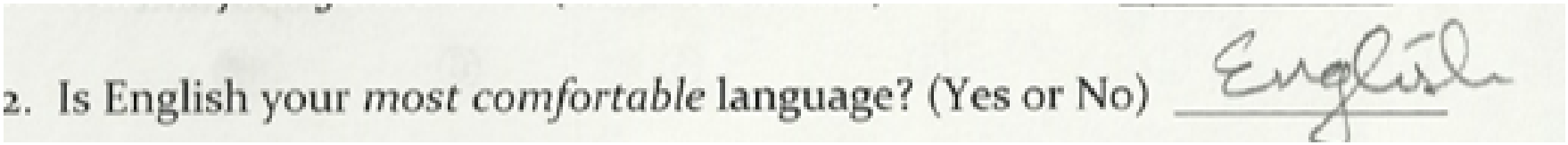
- Guesses?
- 90/254 (34%)
- 16 of the 51 surveys turned in Tuesday had **no** problems (31%)

Guess My Age

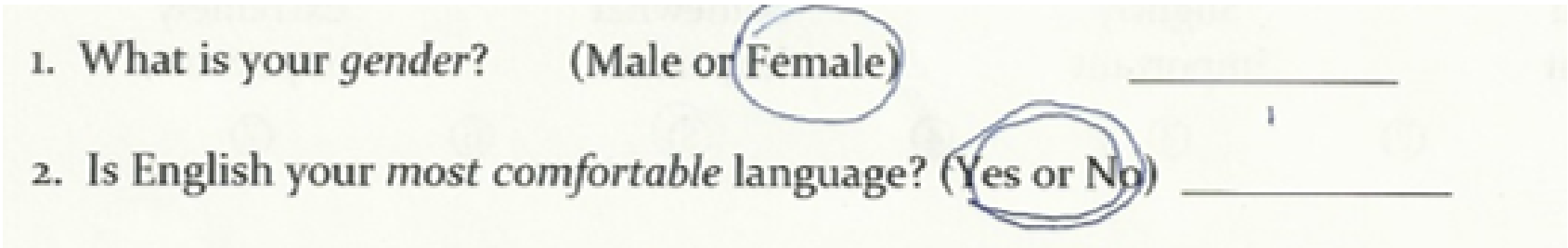


What should we do in these cases?

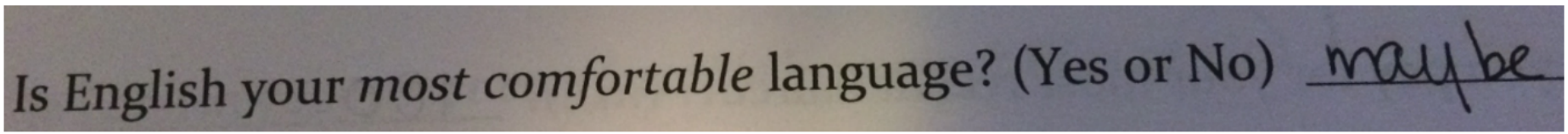
English best language?



TEL Decision: **Yes**



TEL Decision: **NA**



TEL decision: NA

Favorite color

13. What is your favorite color? depends

NA

13. What is your favorite color? Orun

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 ($n = 51$)

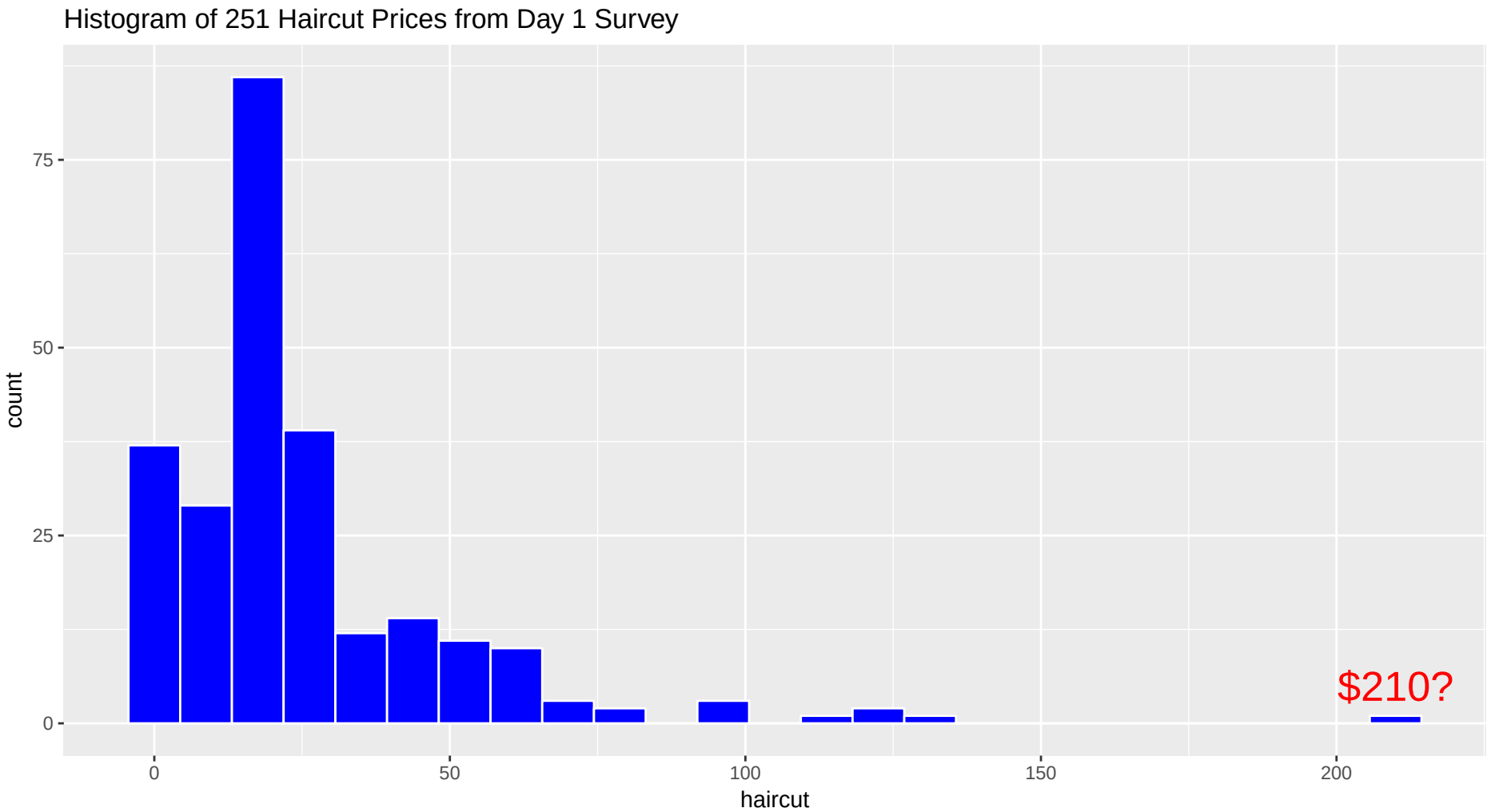
```
48 52 54 56 56 58 60                    4 | 8
62 62 64 64 64 64 64                    5 | 24668
66 66 66 66 68 68 68                    6 | 022444446666888
70 70 70 70 72 74 74                    7 | 0000244446688
74 74 76 76 78 78 80                    8 | 0000024456
80 80 80 80 82 84 84                    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 | 00000000000000000000000000000000035555556778999
 1 | 00000022222222224444444555555555555555555555556778888889
 2 | 00000000000000000000000000000000000012223455555555555555556688
 3 | 000000000000002222255555556
 4 | 000000000003558
 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



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.

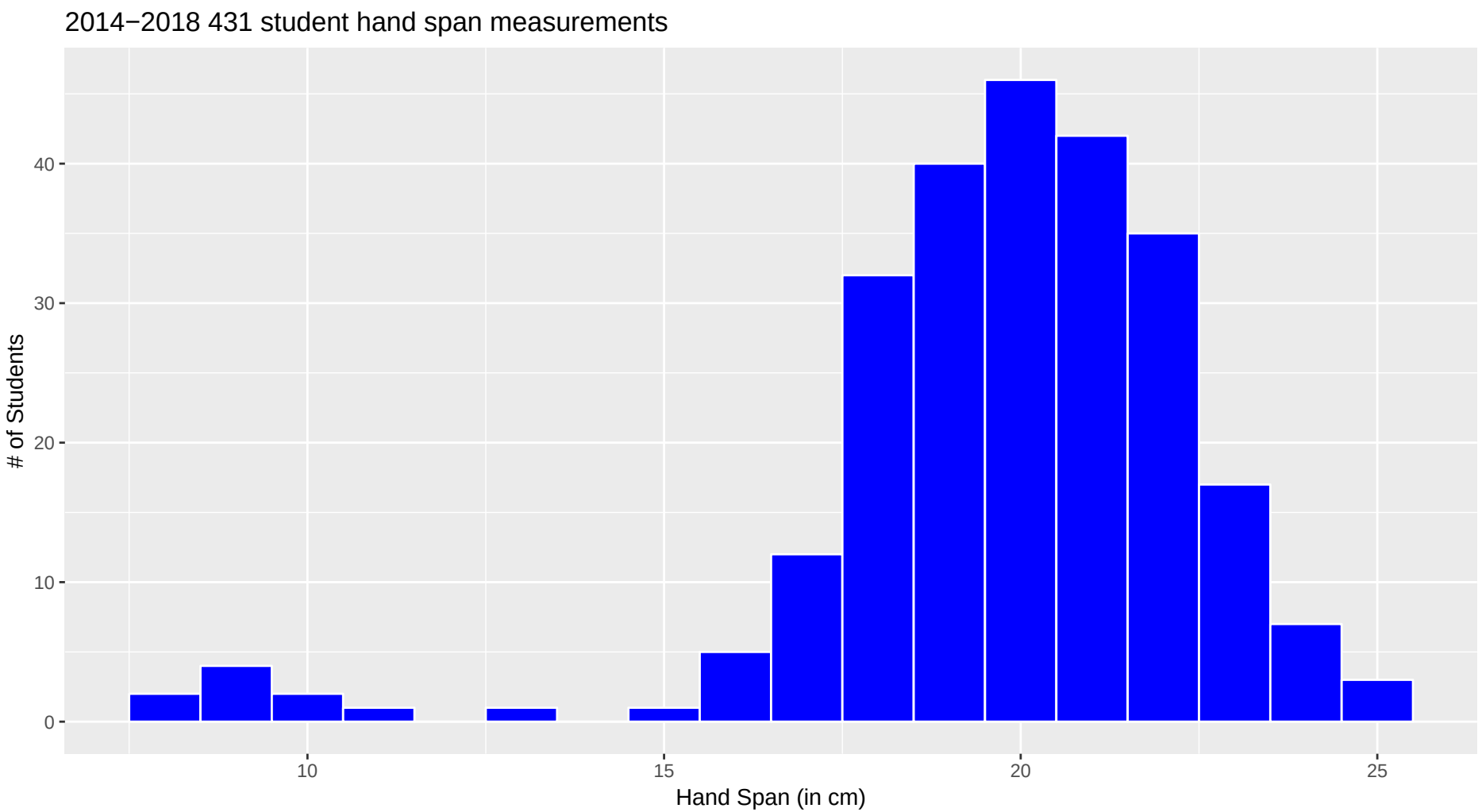
Hand Span Numerical Summaries

```
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).
```



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 | 00000000000000000555555555555556000000000000000+14
20 | 00000000000000000000000000000234555555555557800+29
22 | 00000000000000000000000245555568000000000000555
24 | 0000001000
```

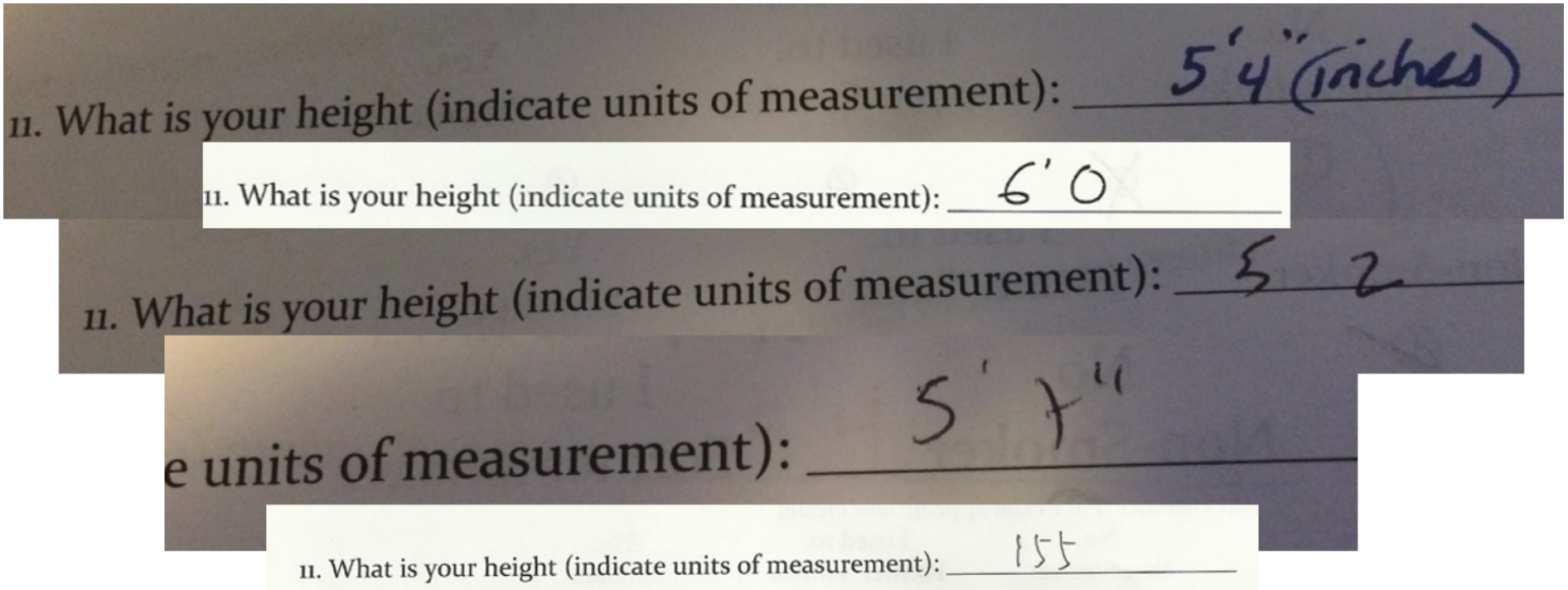
Eight Items had just a few problems

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

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


11. What is your height (indicate units of measurement): 5'4" (inches)
11. What is your height (indicate units of measurement): 6'0
11. What is your height (indicate units of measurement): 5 2
e units of measurement): 5' 7"
11. What is your height (indicate units of measurement): 155

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

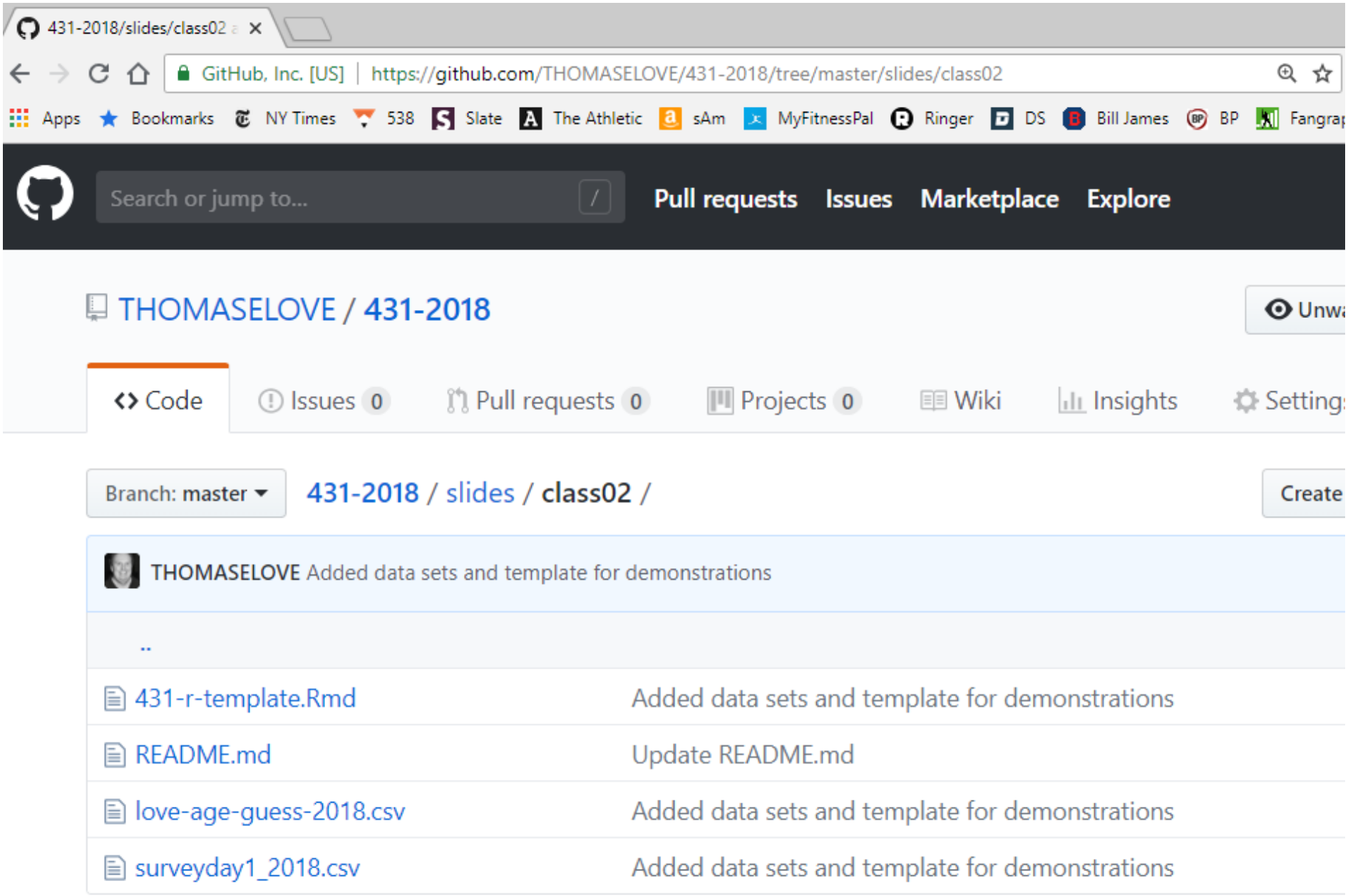
https://github.com/THOMASELOVE/431-2018-data

and also at

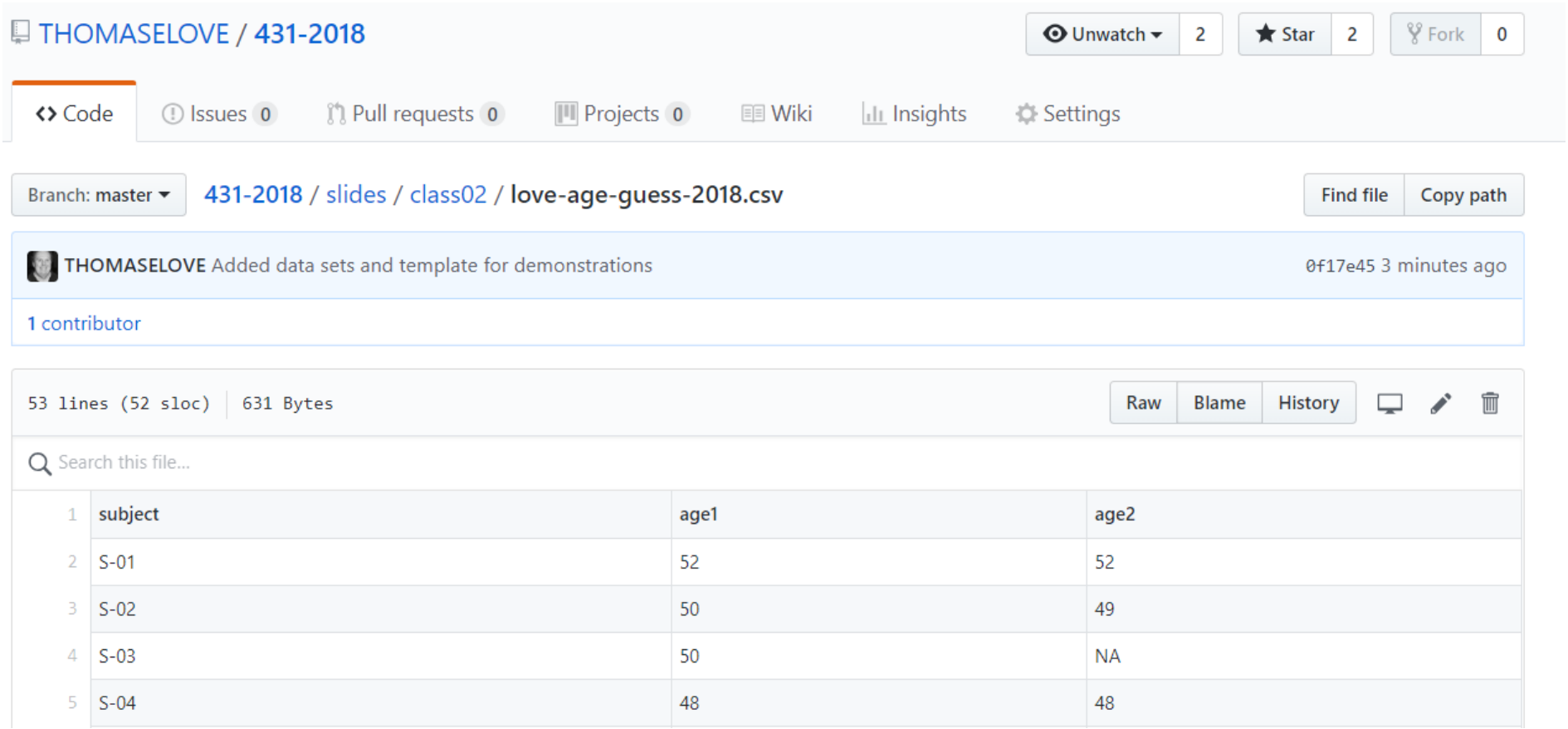
https://github.com/THOMASELOVE/431-2018/tree/master/slides/class02

- 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



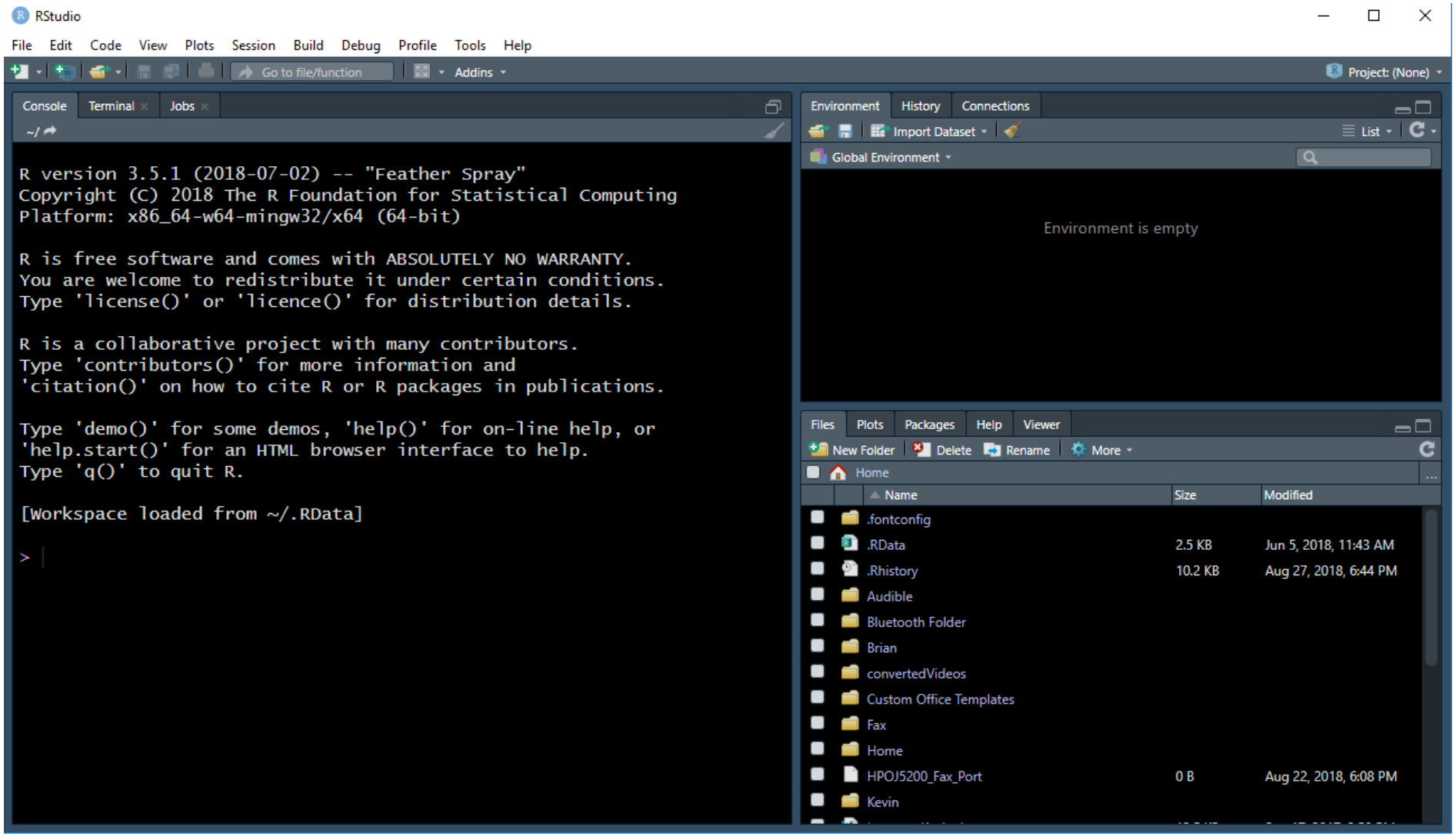
Right-click Raw to download (just) this file, into a 431class02 directory, please.



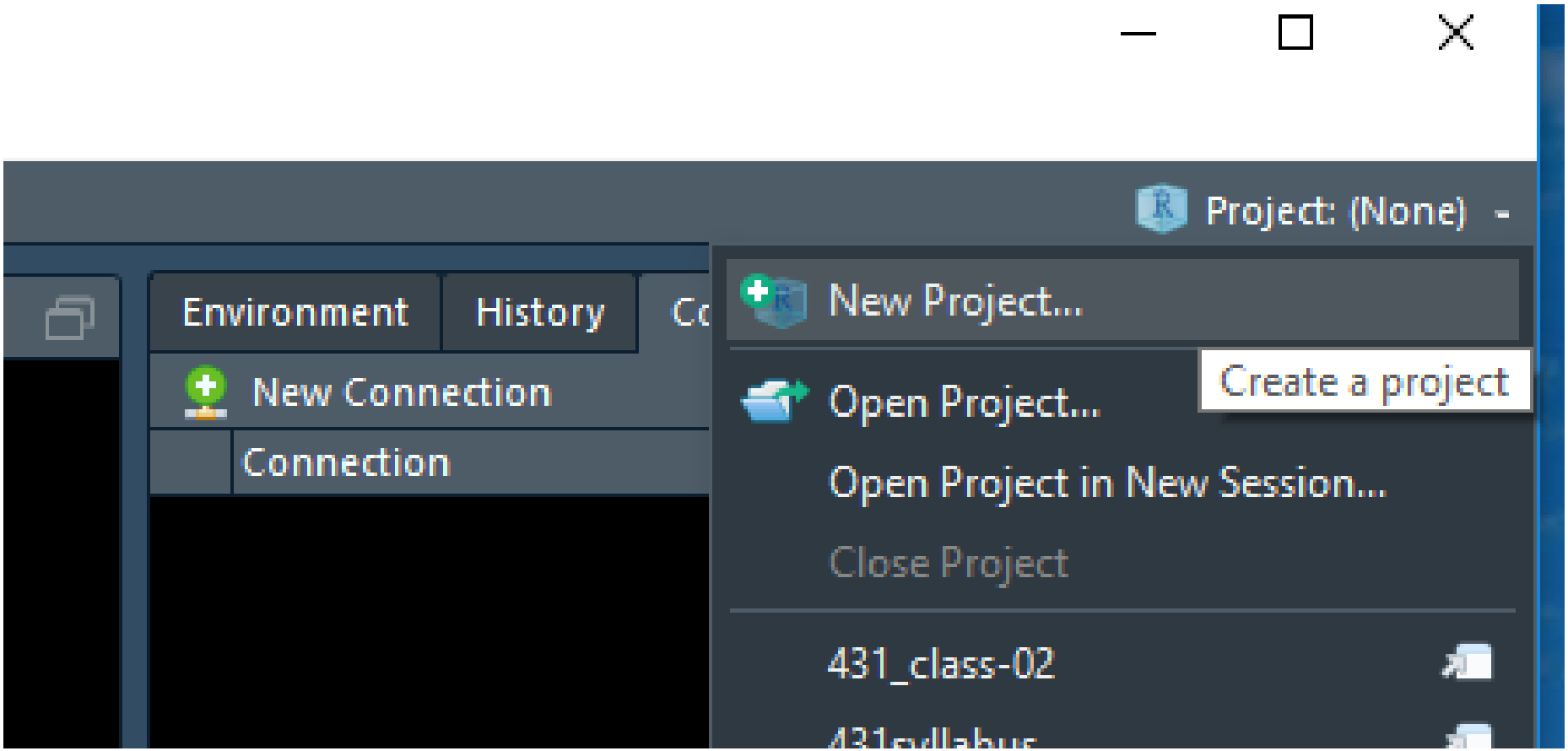
What the file looks like

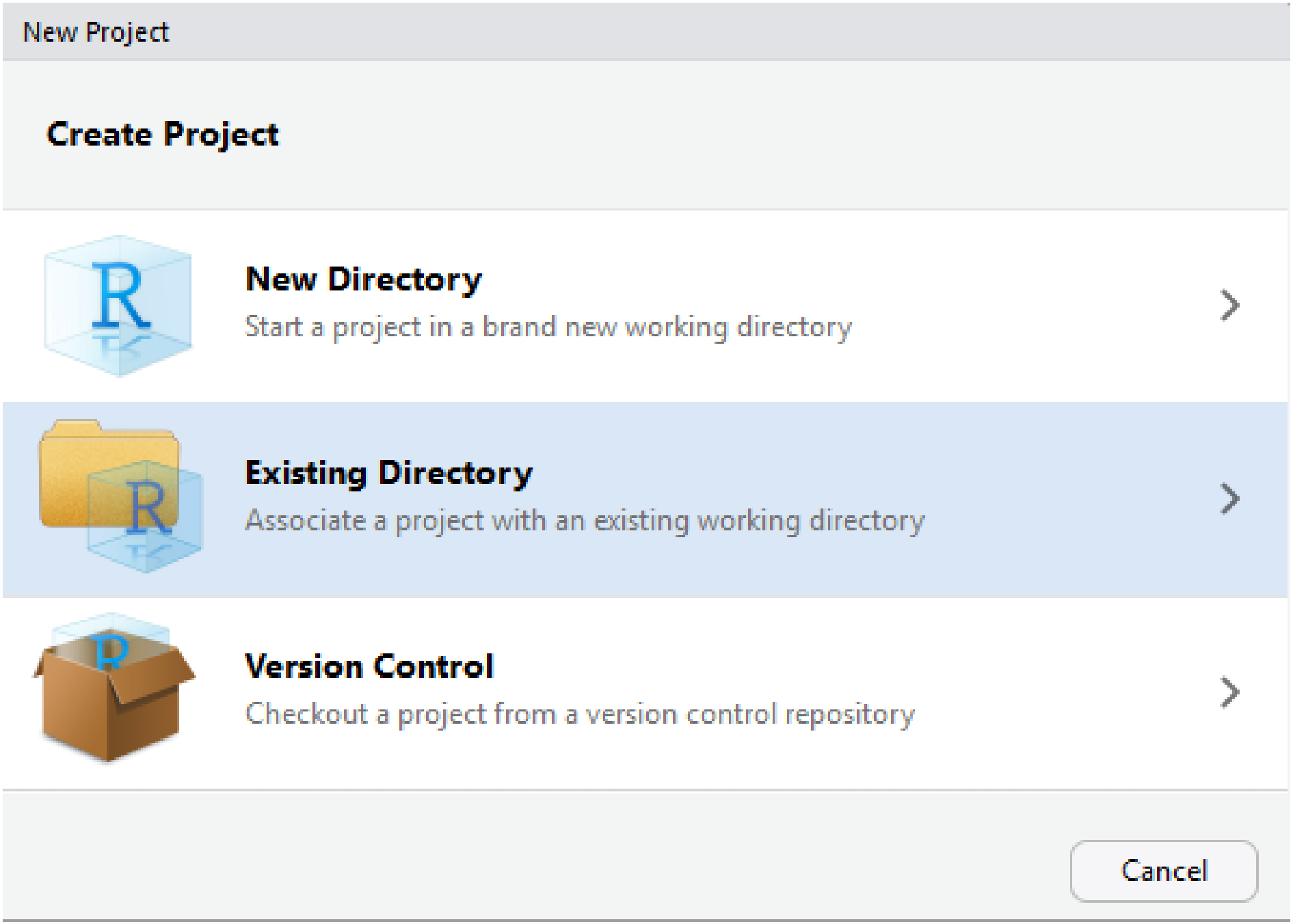
	A	B	C
1	subject	age1	age2
2	S-01	52	52
3	S-02	50	49
4	S-03	50	NA
5	S-04	48	48
6	S-05	52	61
7	S-06	54	54
8	S-07	45	40
9	S-08	51	49

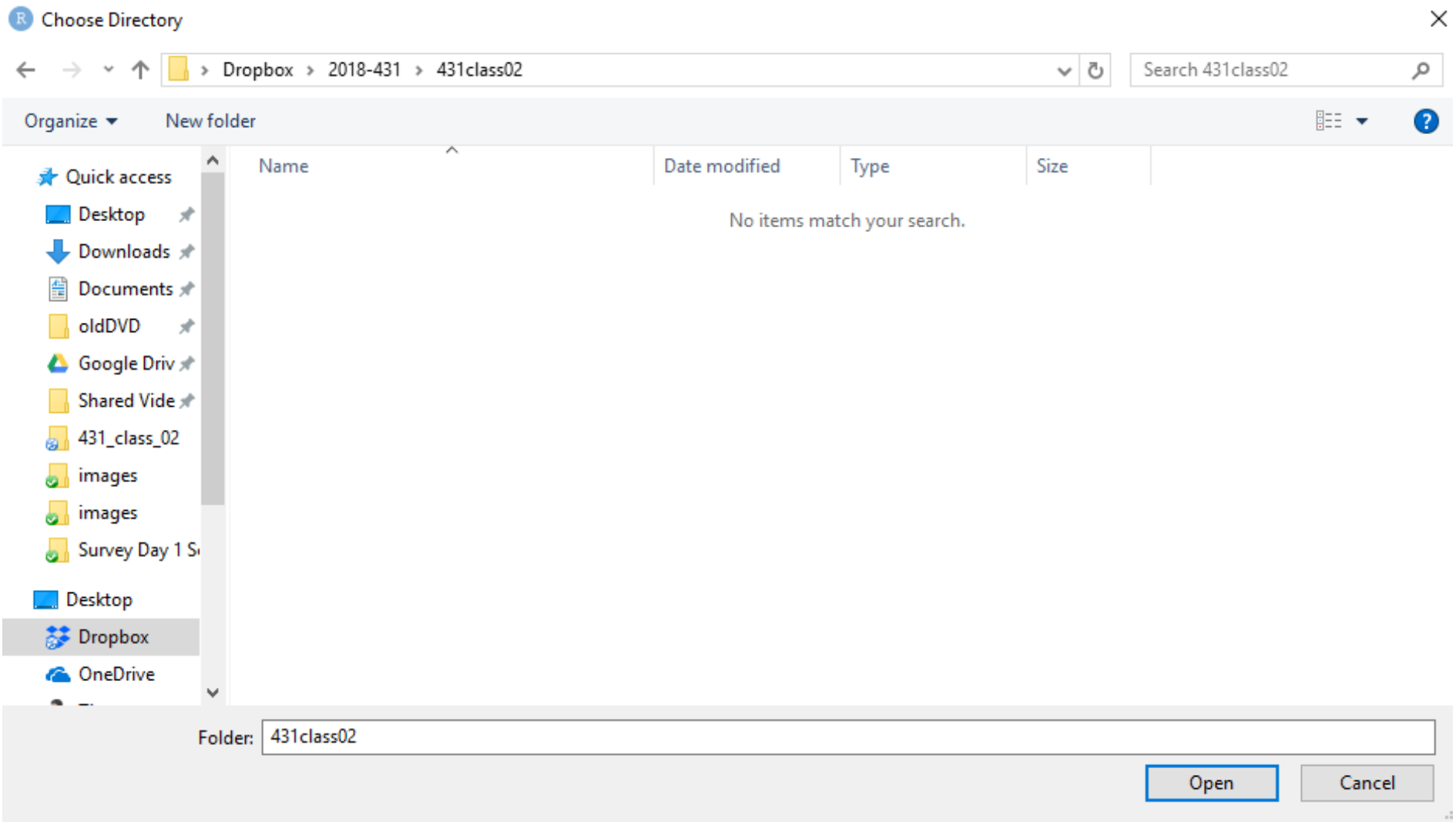
Open R Studio and start a New Project

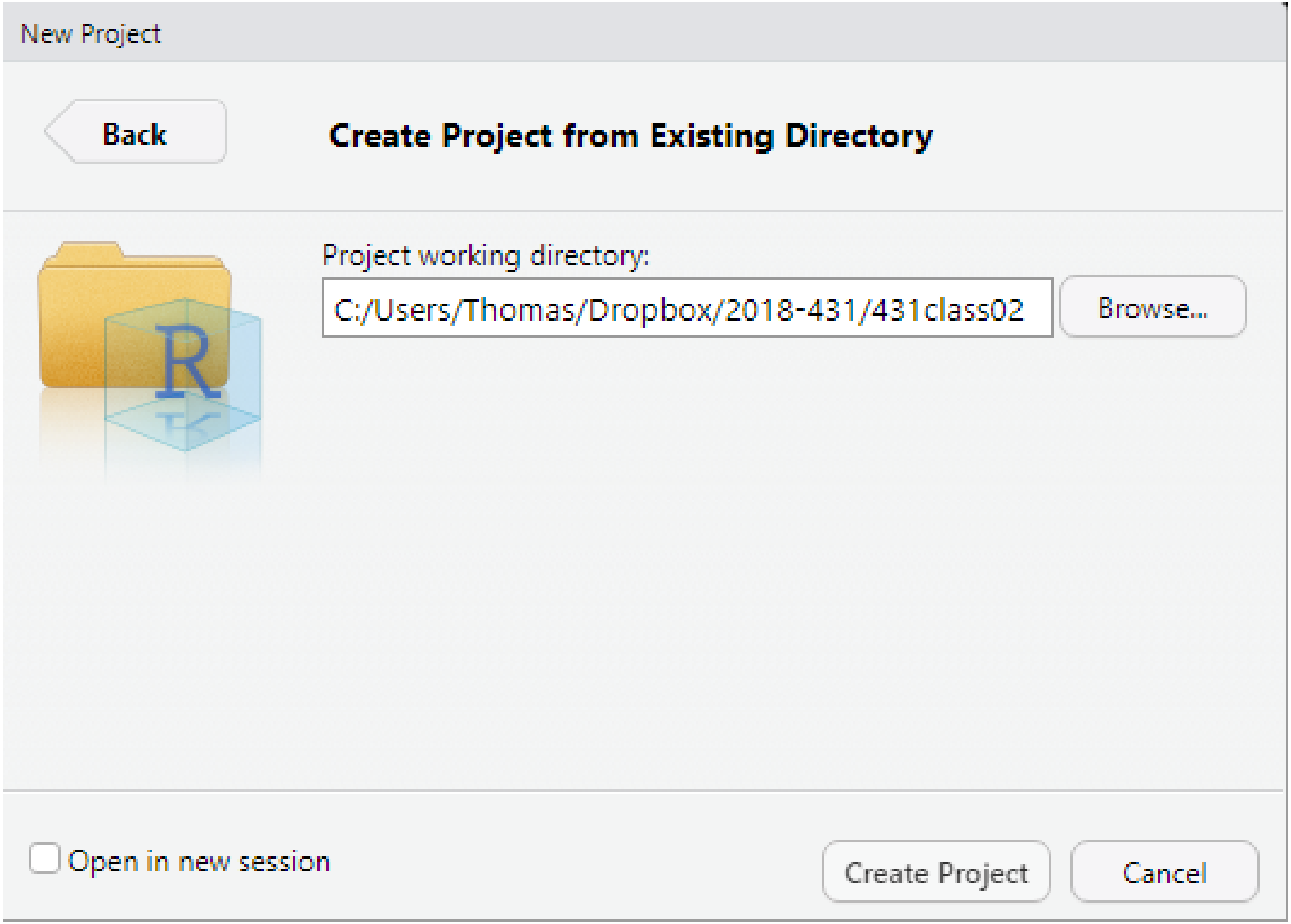


We'll select our existing 431class2 directory for this project.

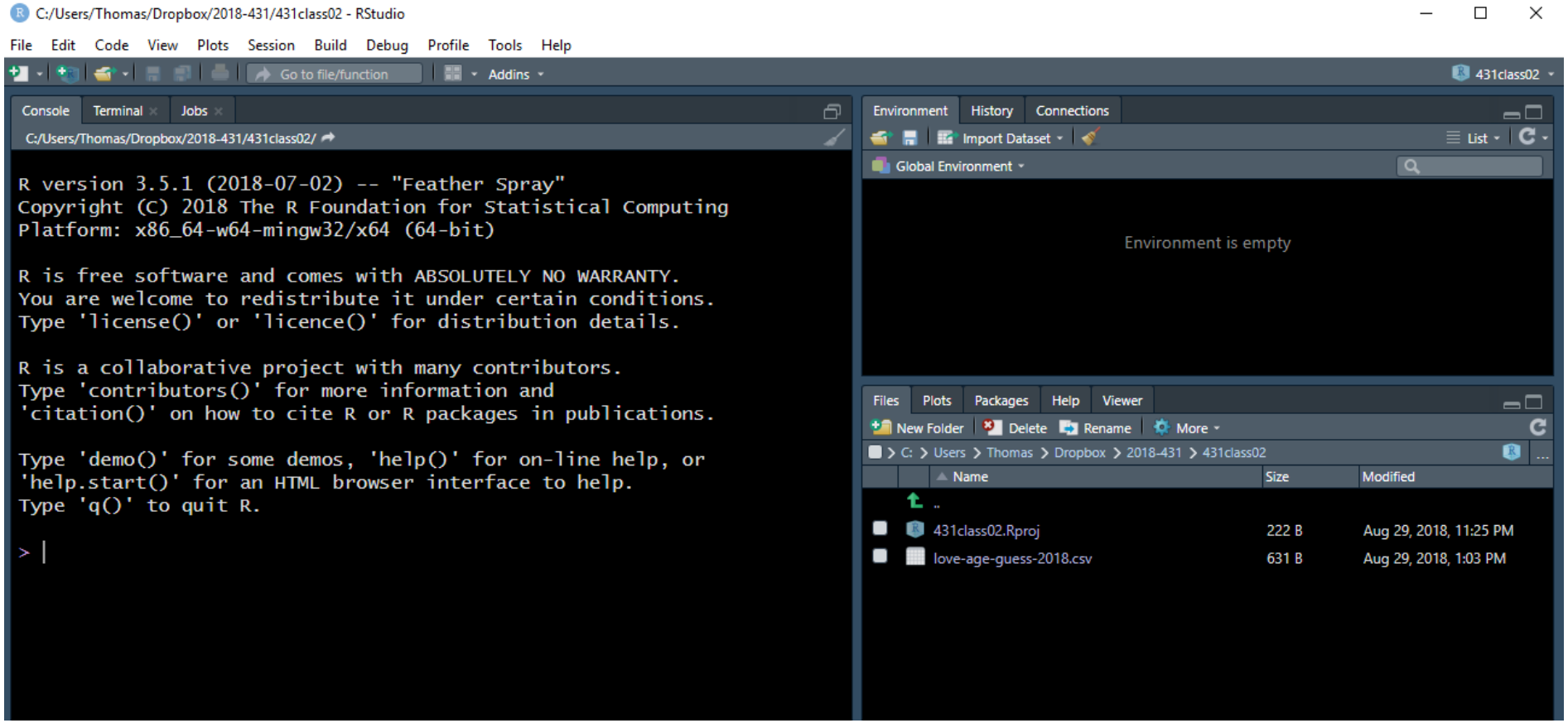



New Project
Create Project
New Directory
Start a project in a brand new working directory
Existing Directory
Associate a project with an existing working directory
Version Control
Checkout a project from a version control repository
Cancel

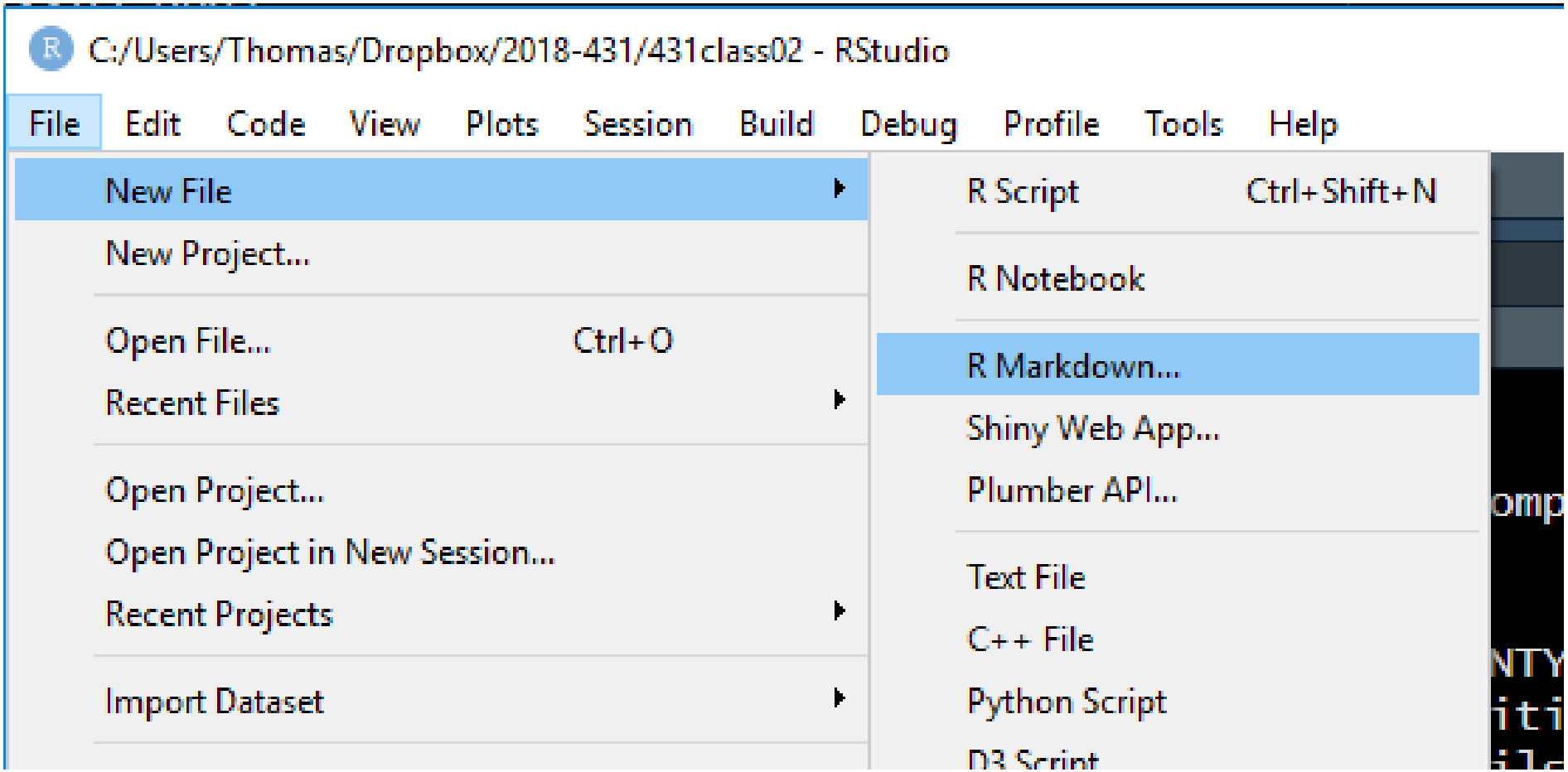

Choose Directory
Dropbox › 2018-431 › 431class02
Search 431class02
Organize
New folder
Quick access
Desktop
Downloads
Documents
oldDVD
Google Driv
Shared Vide
431_class_02
images
images
Survey Day 1 S
Desktop
Dropbox
OneDrive
Name
Date modified
Type
Size
No items match your search.
Folder:
431class02
Open
Cancel


New Project
Back
Create Project from Existing Directory
Project working directory:
C:/Users/Thomas/Dropbox/2018-431/431class02
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
- Presentation
- Shiny
- From Template

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).

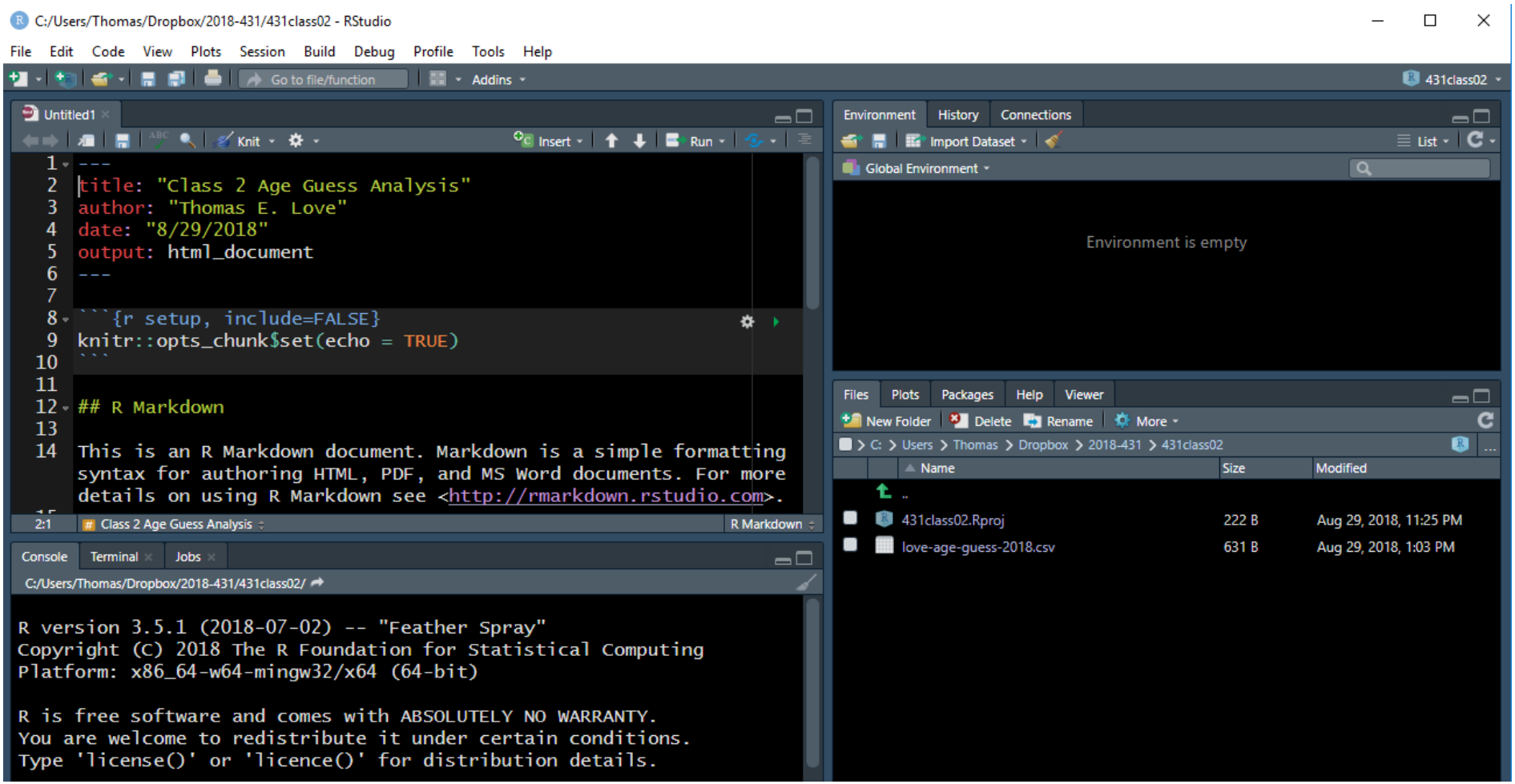
( ) PDF

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

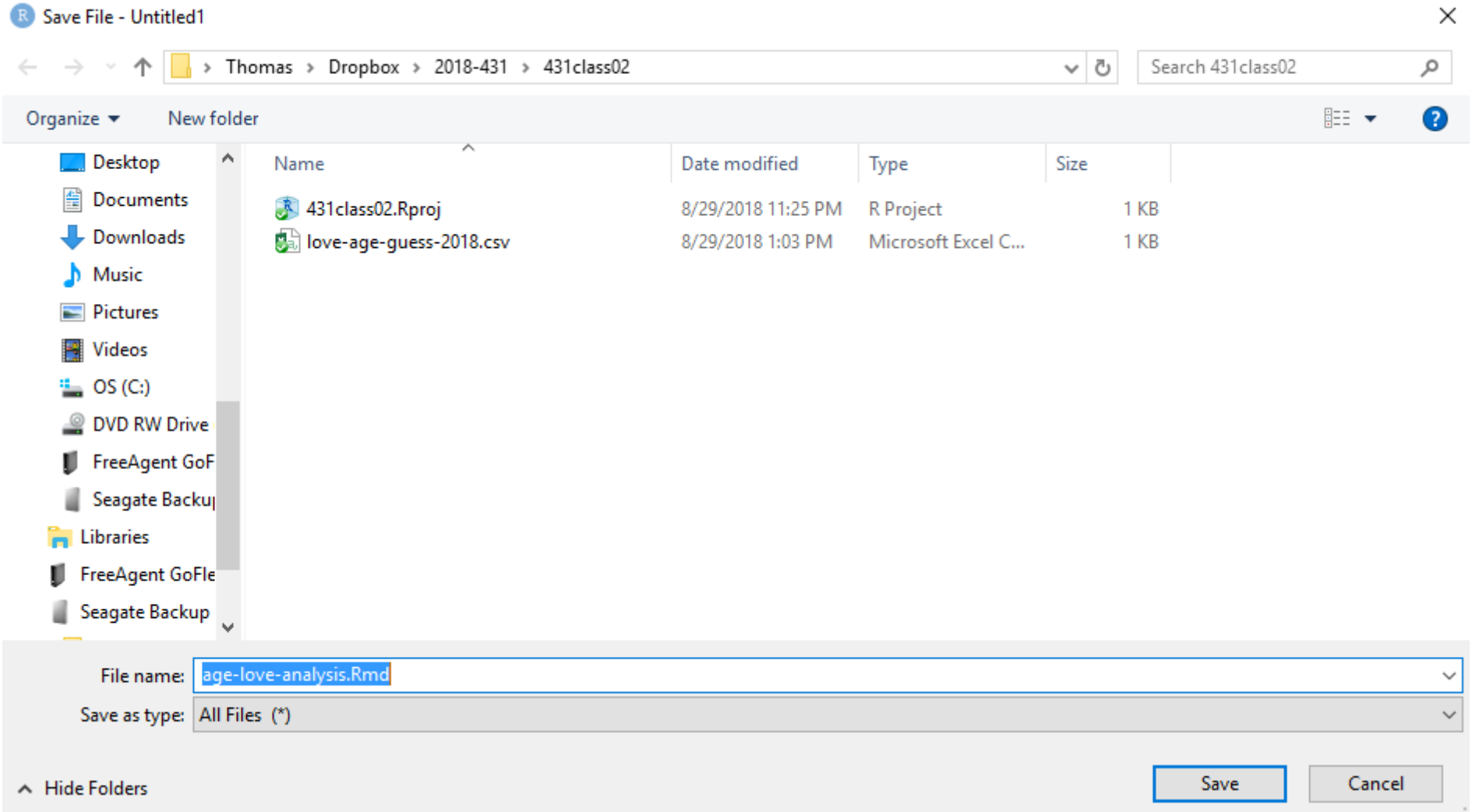
( ) Word

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

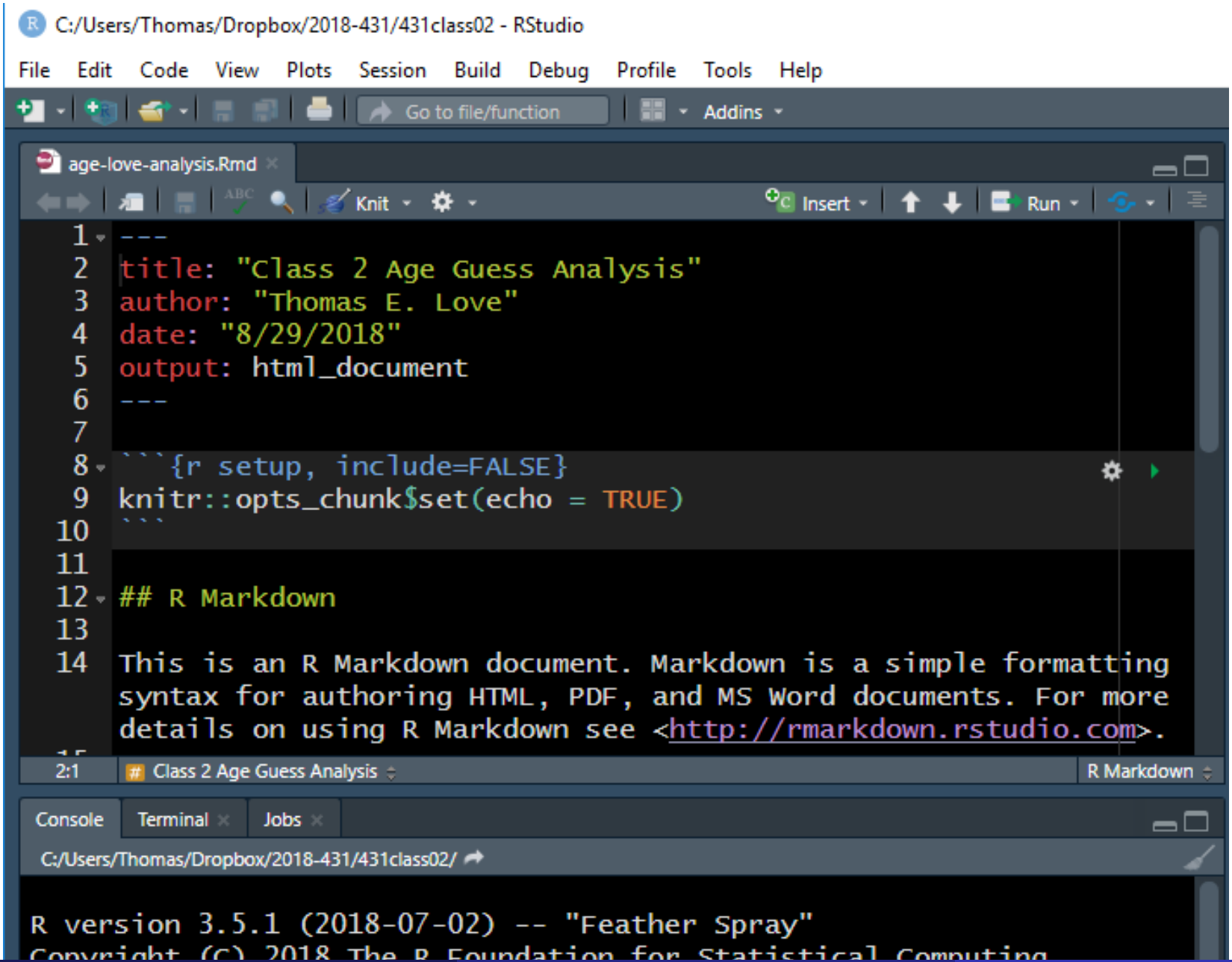
A sample Markdown file is generated. Let's knit it into an HTML file.



Save your work as age-love-analysis.Rmd



Now knit the R Markdown file.



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:

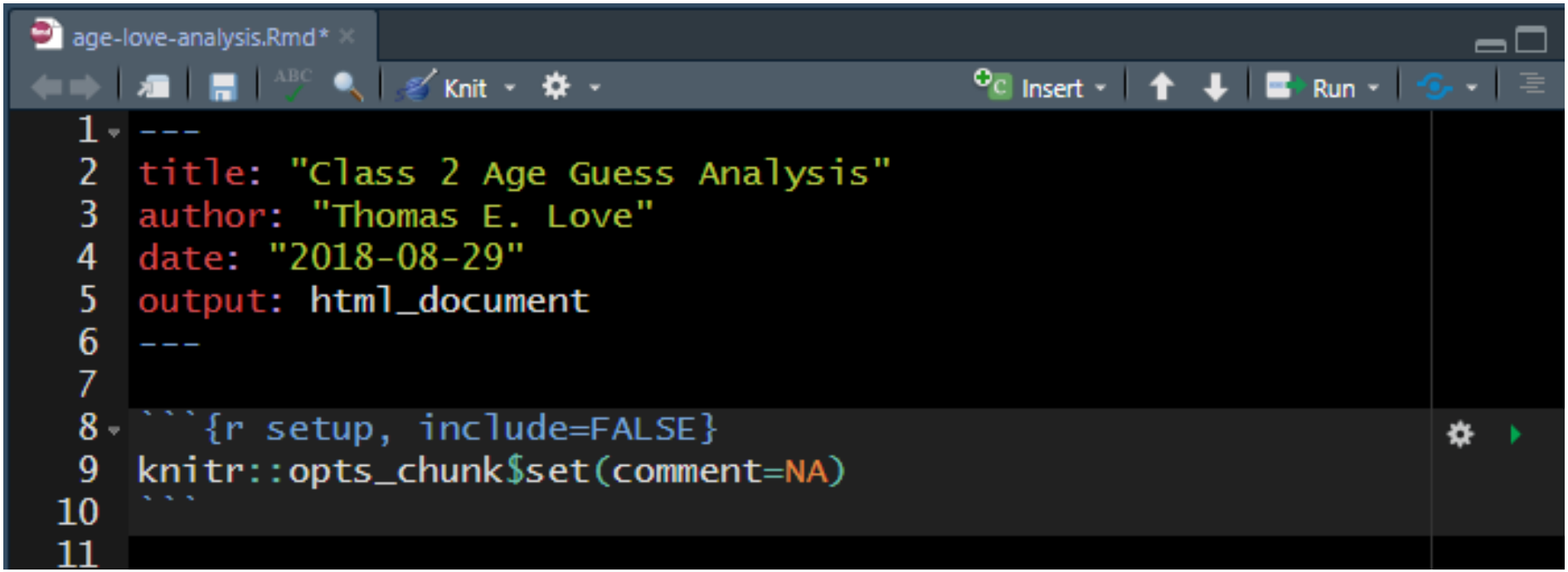
```
summary(cars)
```

```
##      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
```

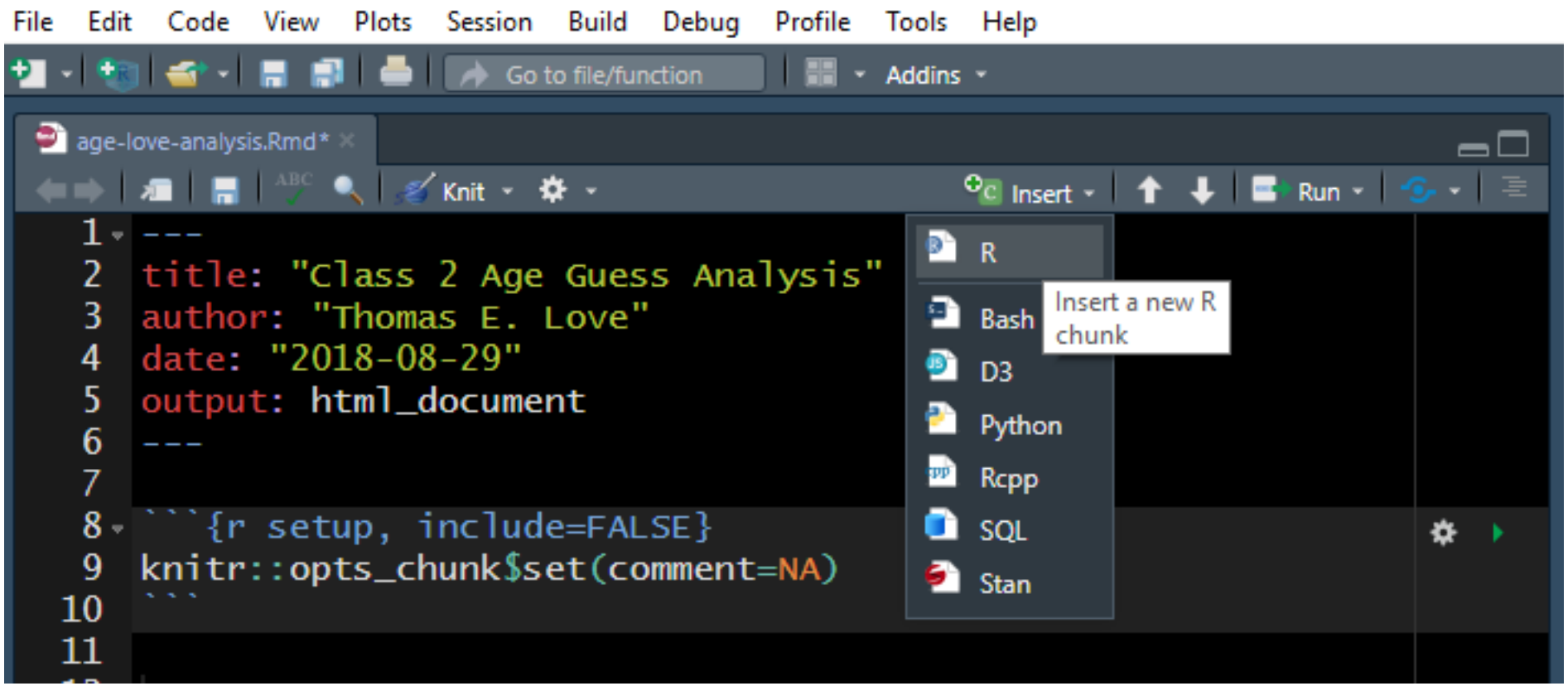
Including Plots

You can also embed plots, for example:

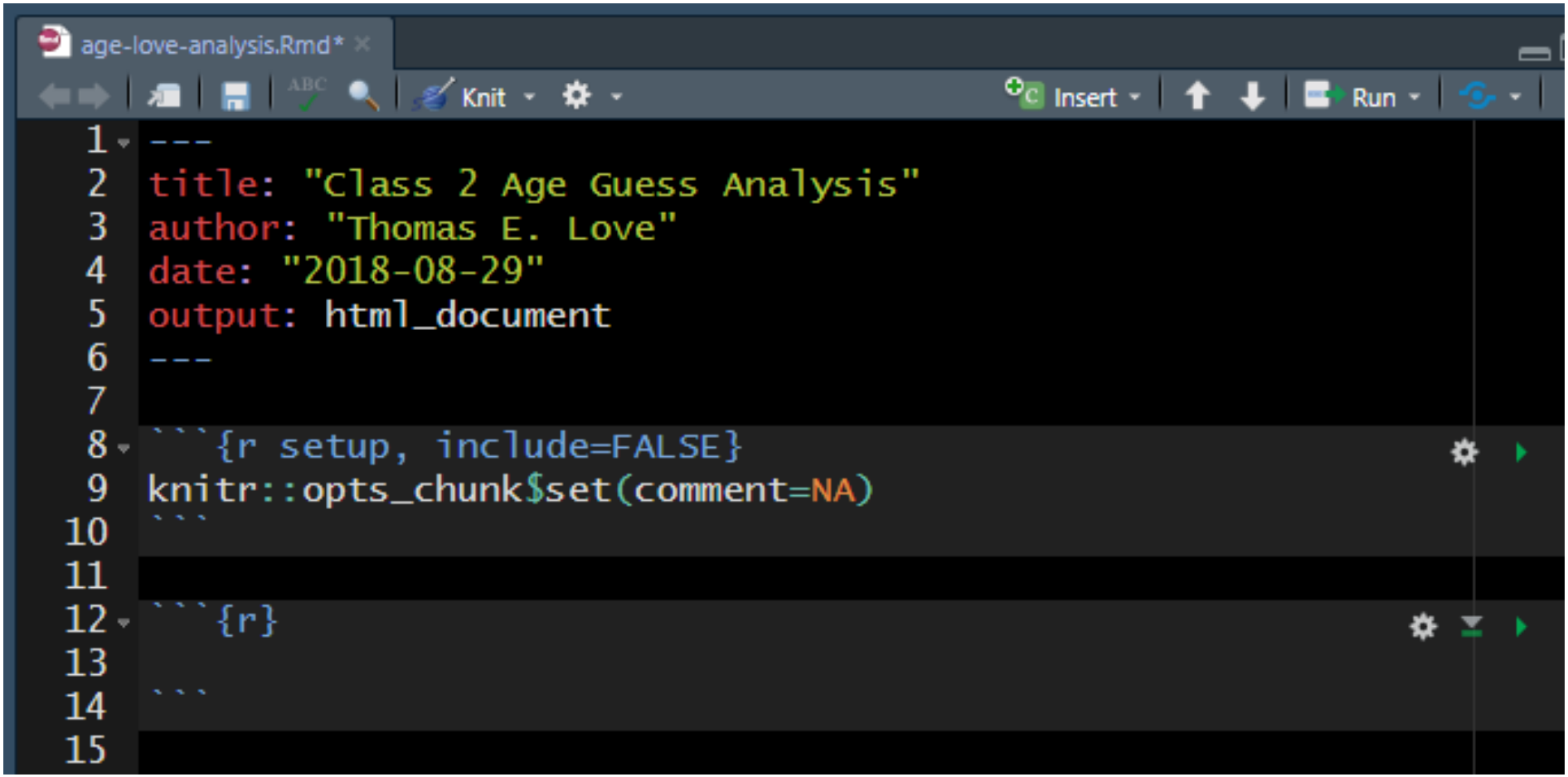
Edit the file to change the date and setup code



Insert a new "chunk" of R code



A blank "chunk"



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

age-love-analysis.Rmd*

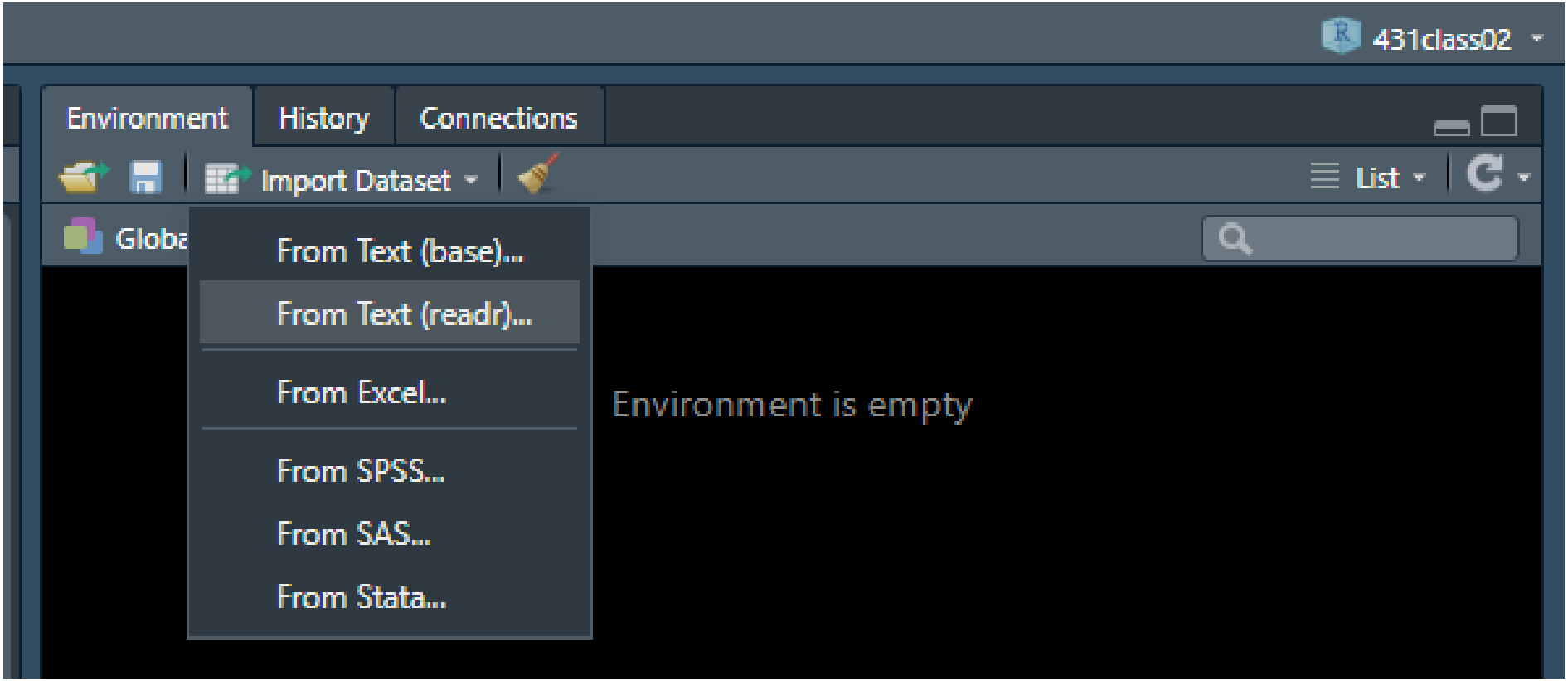
Knit | Insert

```
---
title: "Class 2 Age Guess Analysis"
author: "Thomas E. Love"
date: "2018-08-29"
output: html_document
---

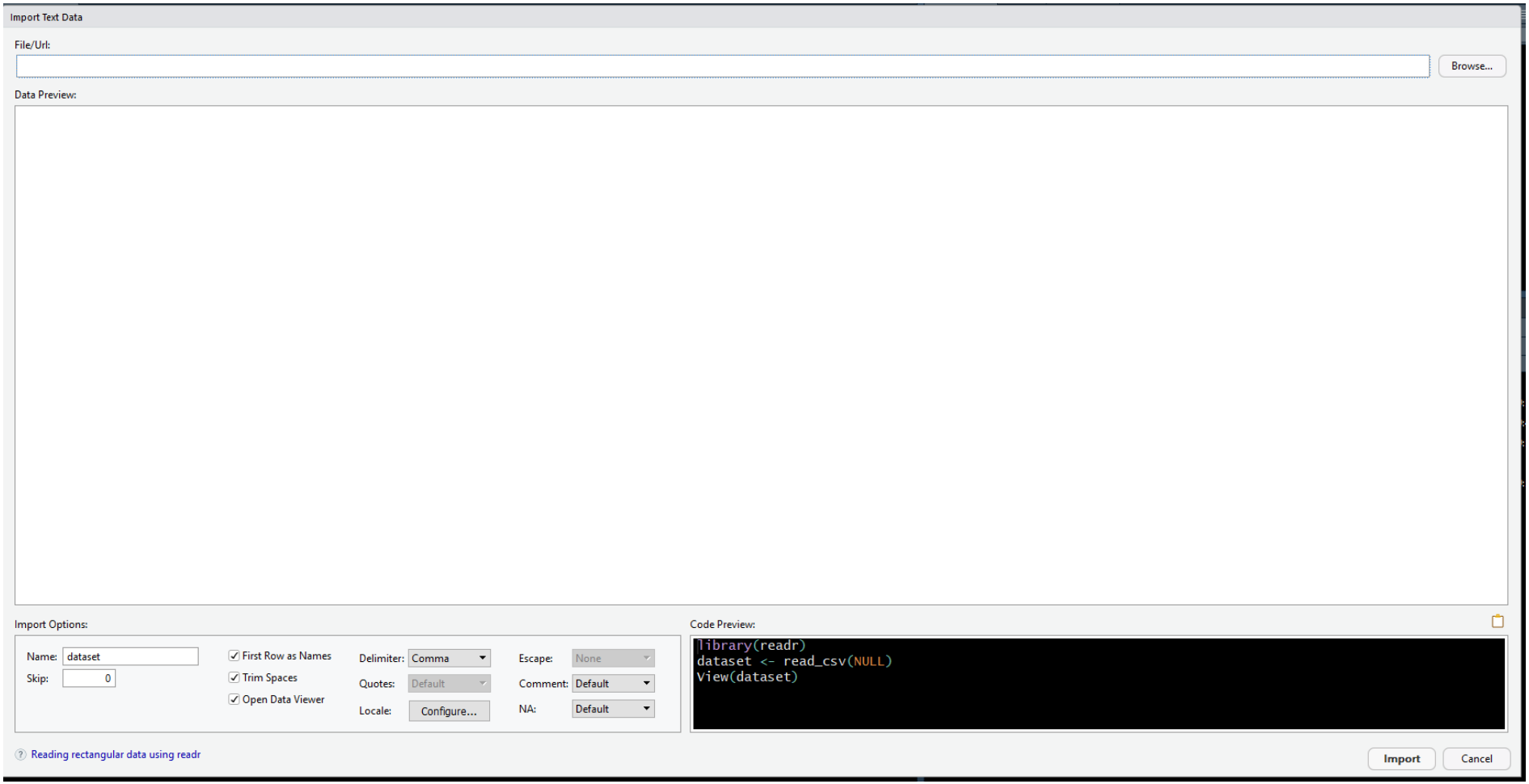
```{r setup, include=FALSE}
knitr::opts_chunk$set(comment=NA)
```

```{r packages}
library(magrittr); library(tidyverse)
```
```

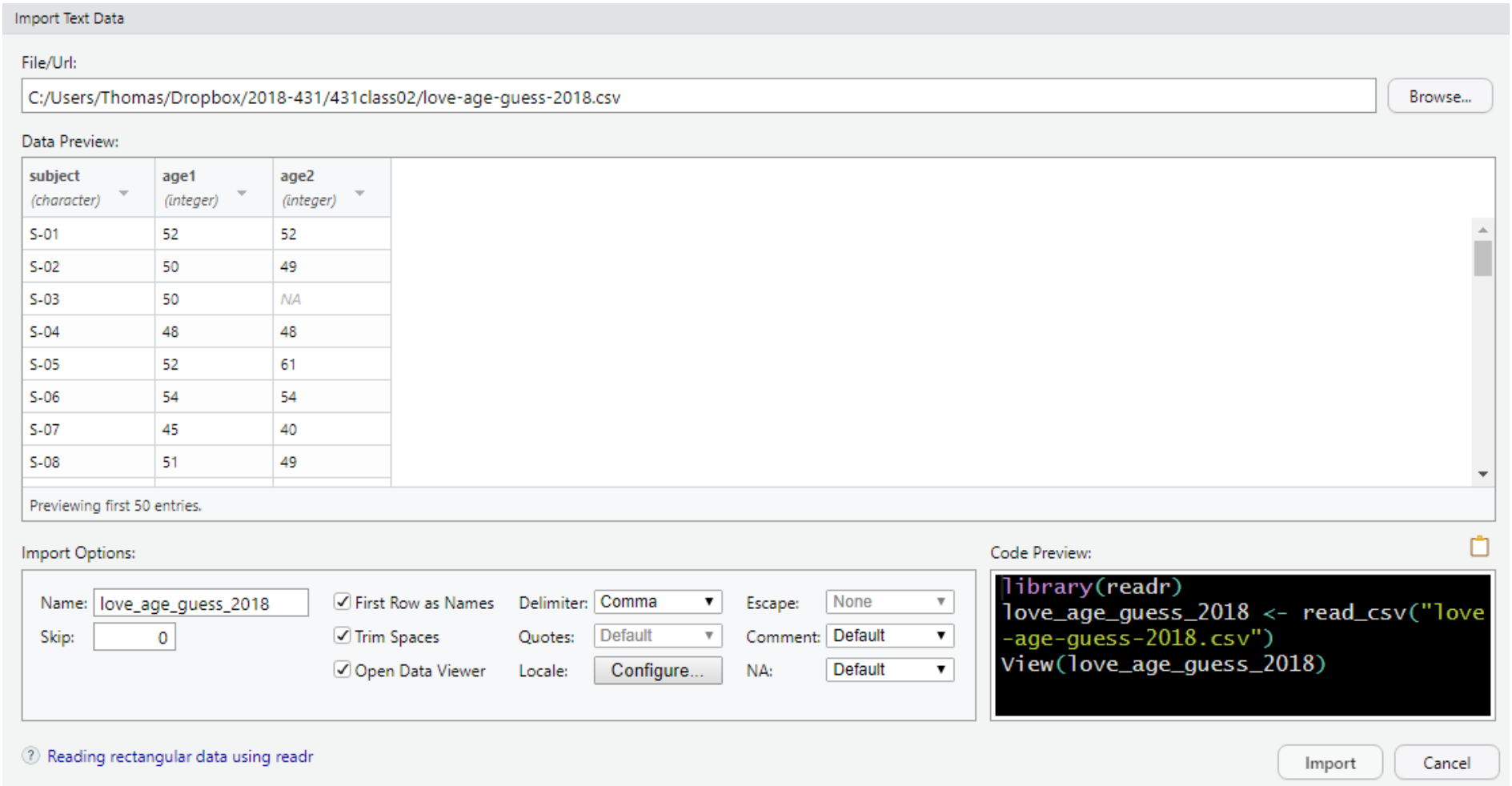
Import the .csv data set



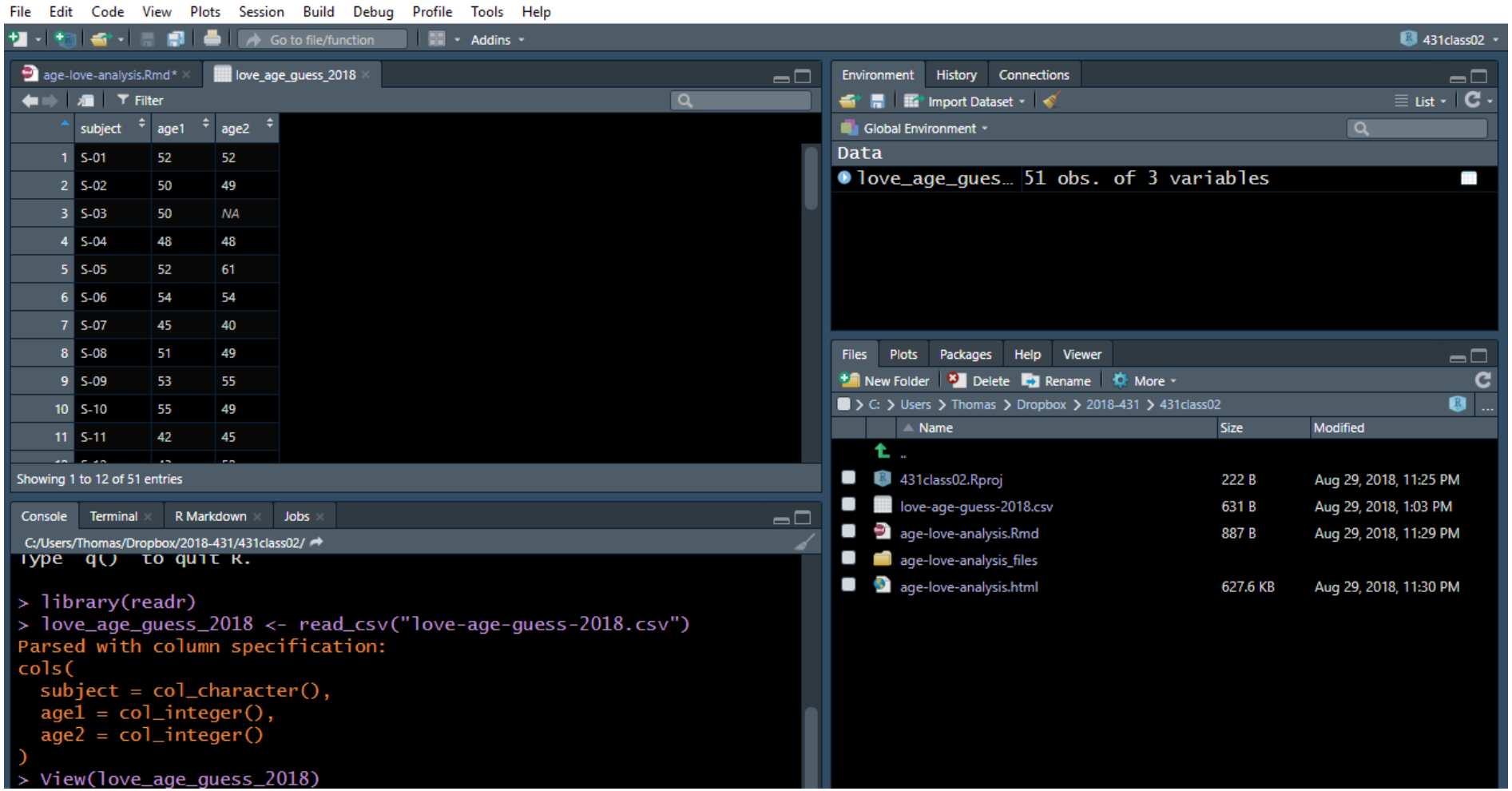
Import window



After we make our choices. . .



Result (note code in Console)



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

age-love-analysis.Rmd* | love_age_guess_2018

Knit | Insert | Run

```
---
title: "Class 2 Age Guess Analysis"
author: "Thomas E. Love"
date: "2018-08-29"
output: html_document
---

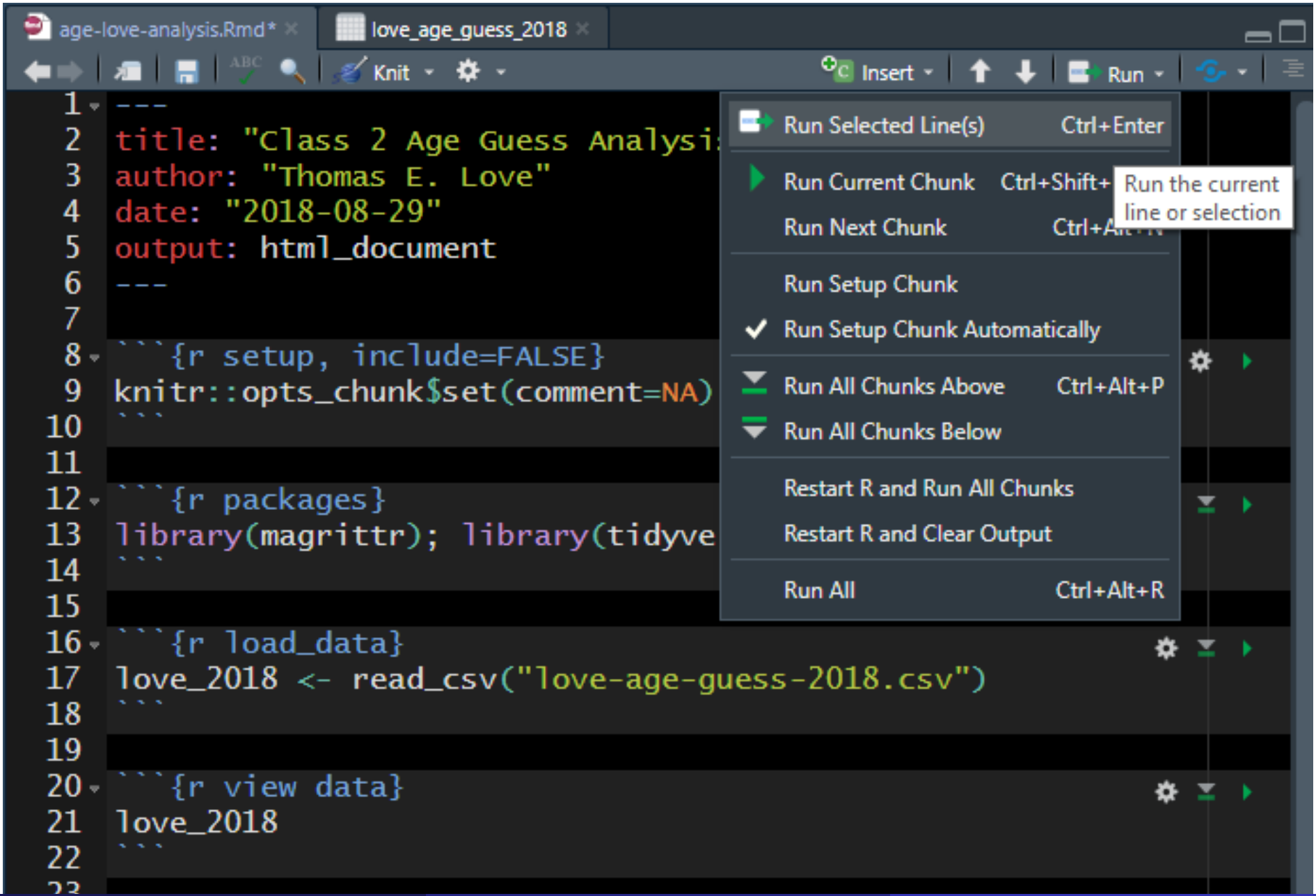
```{r setup, include=FALSE}
knitr::opts_chunk$set(comment=NA)
```

```{r packages}
library(magrittr); library(tidyverse)
```

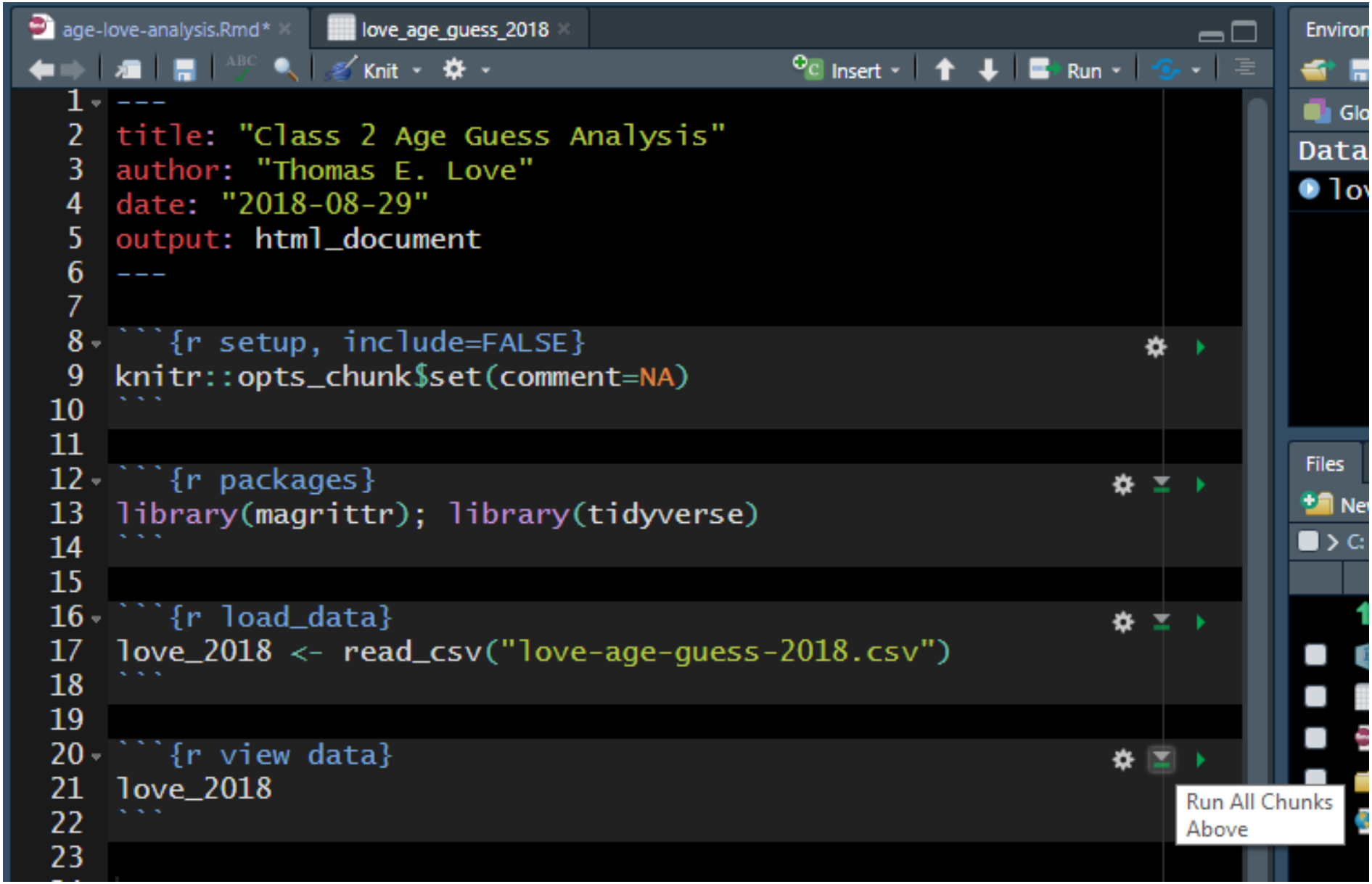
```{r load_data}
love_2018 <- read_csv("love-age-guess-2018.csv")
```

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

Running the code so we can see results



Or run all of the "chunks" up to a particular point



Running the first three chunks of code

age-love-analysis.Rmd* | love_age_guess_2018

Knit | Insert | Run

```
---
title: "Class 2 Age Guess Analysis"
author: "Thomas E. Love"
date: "2018-08-29"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(comment=NA)
```

```{r packages}
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 ggplot2 3.0.0     v forcats 0.3.0
-- Conflicts ---------------------------------------- tidyverse_conflicts() --
x tidyr::extract()   masks magrittr::extract()
x dplyr::filter()    masks stats::filter()
x dplyr::lag()       masks stats::lag()
x purrr::set_names() masks magrittr::set_names()
```

```{r load_data}
love_2018 <- read_csv("love-age-guess-2018.csv")
```
```

```
Parsed with column specification:
cols(
  subject = col_character(),
  age1 = col_integer(),
```
```

Now, let's look at the data

The love_2018 tibble

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

subject <chr>	age1 <int>	age2 <int>
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

Add some documentation, mixing text with code

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

subject <chr>	age1 <int>	age2 <int>
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

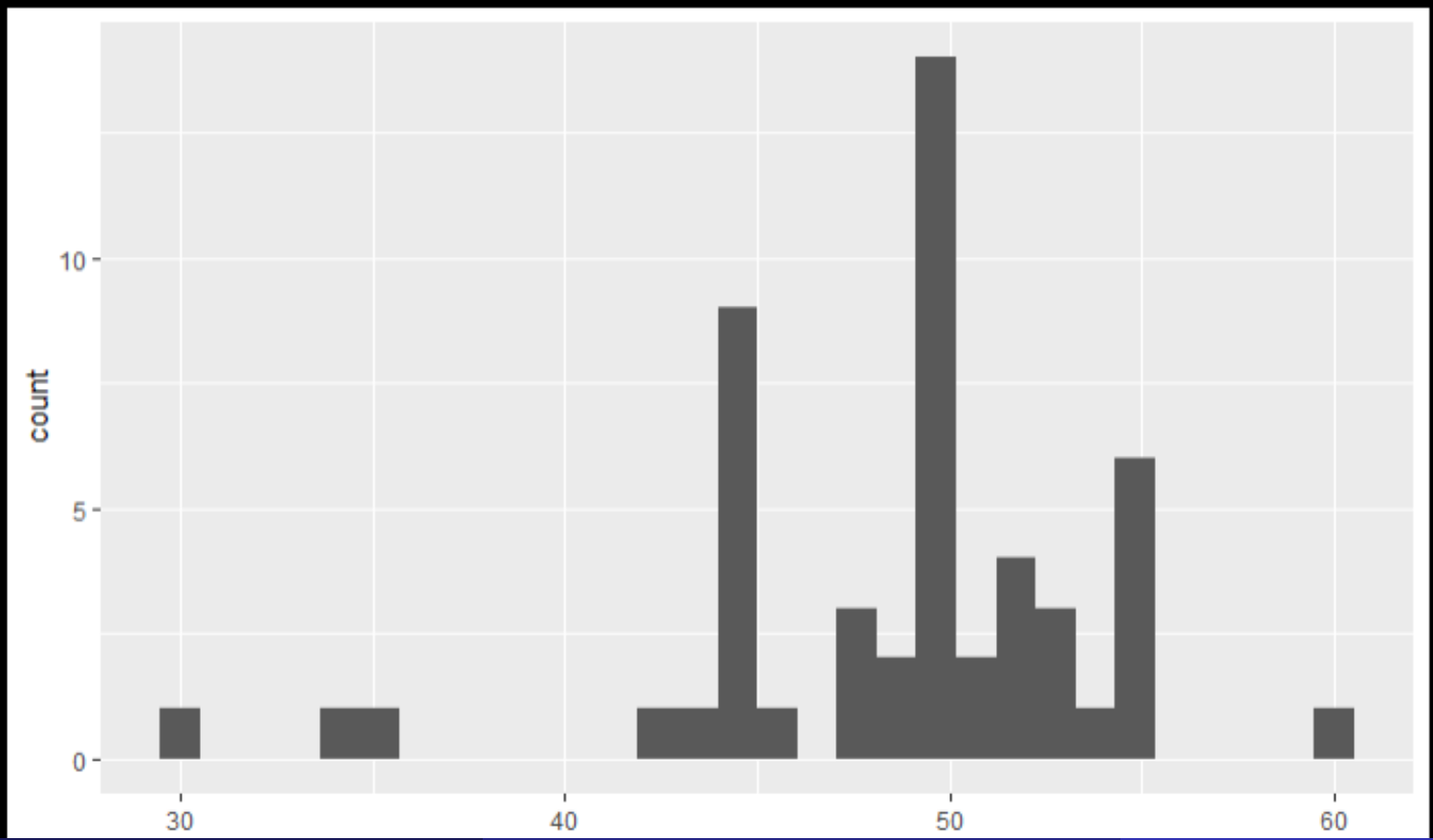
```
25
26 # Build a Histogram of the First Guesses
27
28 ```{r}
29 ggplot(data = love_2018, aes(x = age1)) +
30     geom_histogram()
31 ```
32
```

The result. Can we do better?

```
# Build a Histogram of the First Guesses

```{r}
ggplot(data = love_2018, aes(x = age1)) +
    geom_histogram()
```
```

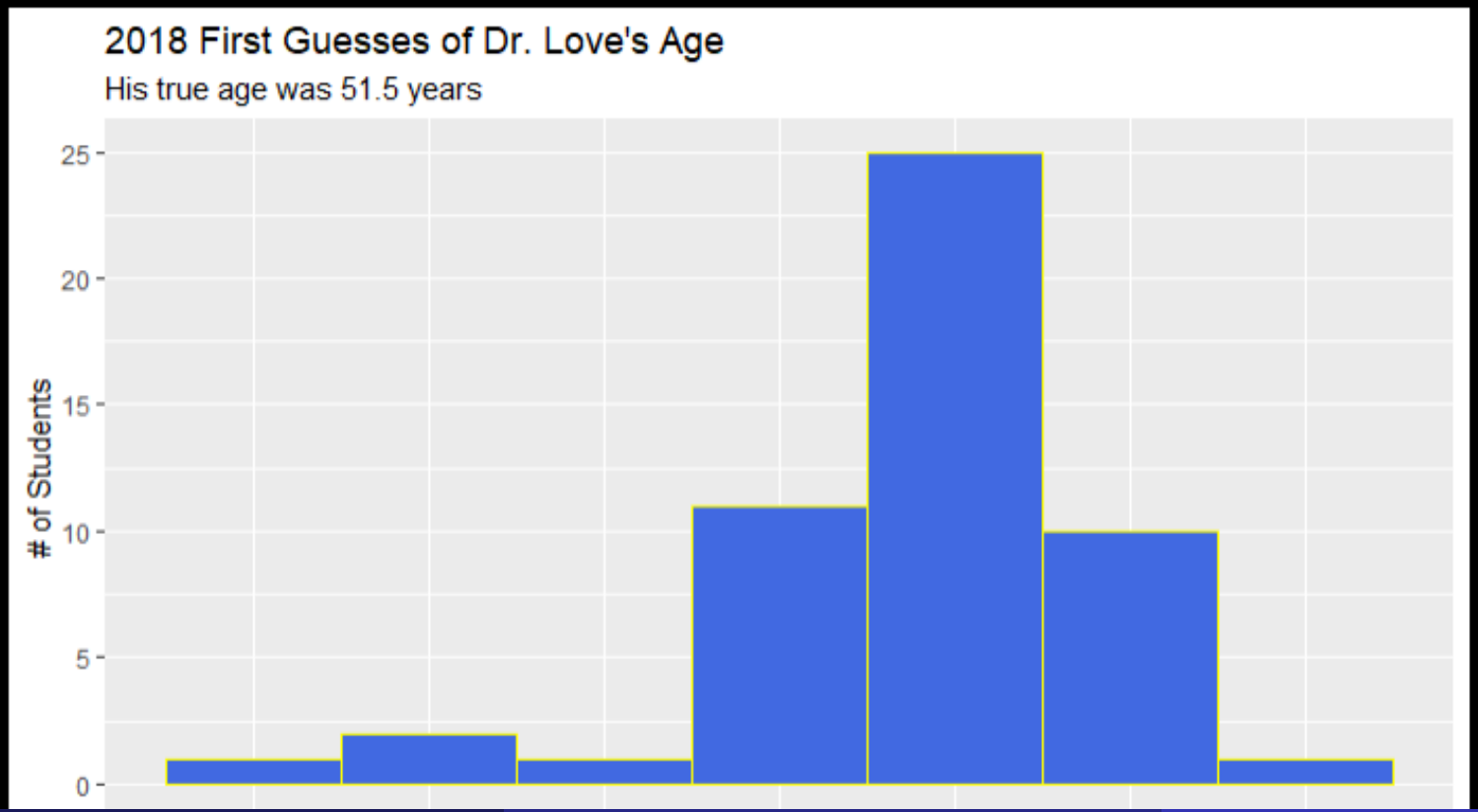
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



A second attempt at the histogram

```
## A Nicer Histogram

```{r histogram2}
ggplot(data = love_2018, aes(x = age1)) +
    geom_histogram(bins = 7, fill = "royalblue", col = "yellow") +
    labs(x = "First Guess of Dr. Love's Age on 2018-08-28",
         y = "# of Students",
         title = "2018 First Guesses of Dr. Love's Age",
         subtitle = "His true age was 51.5 years")
```
```



A numerical summary of the data

```
# A numerical summary

```{r}
summary(love_2018)
```
```

```
   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      
```

Calculating and Summarizing the Errors

```
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?

```
# Better First or Second Guess?

```{r calculate-errors}
love_2018 <- love_2018 %>%
    mutate(error1 = abs(age1 - 51.5),
           error2 = abs(age2 - 51.5))

summary(love_2018)
```
```

```
   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
```

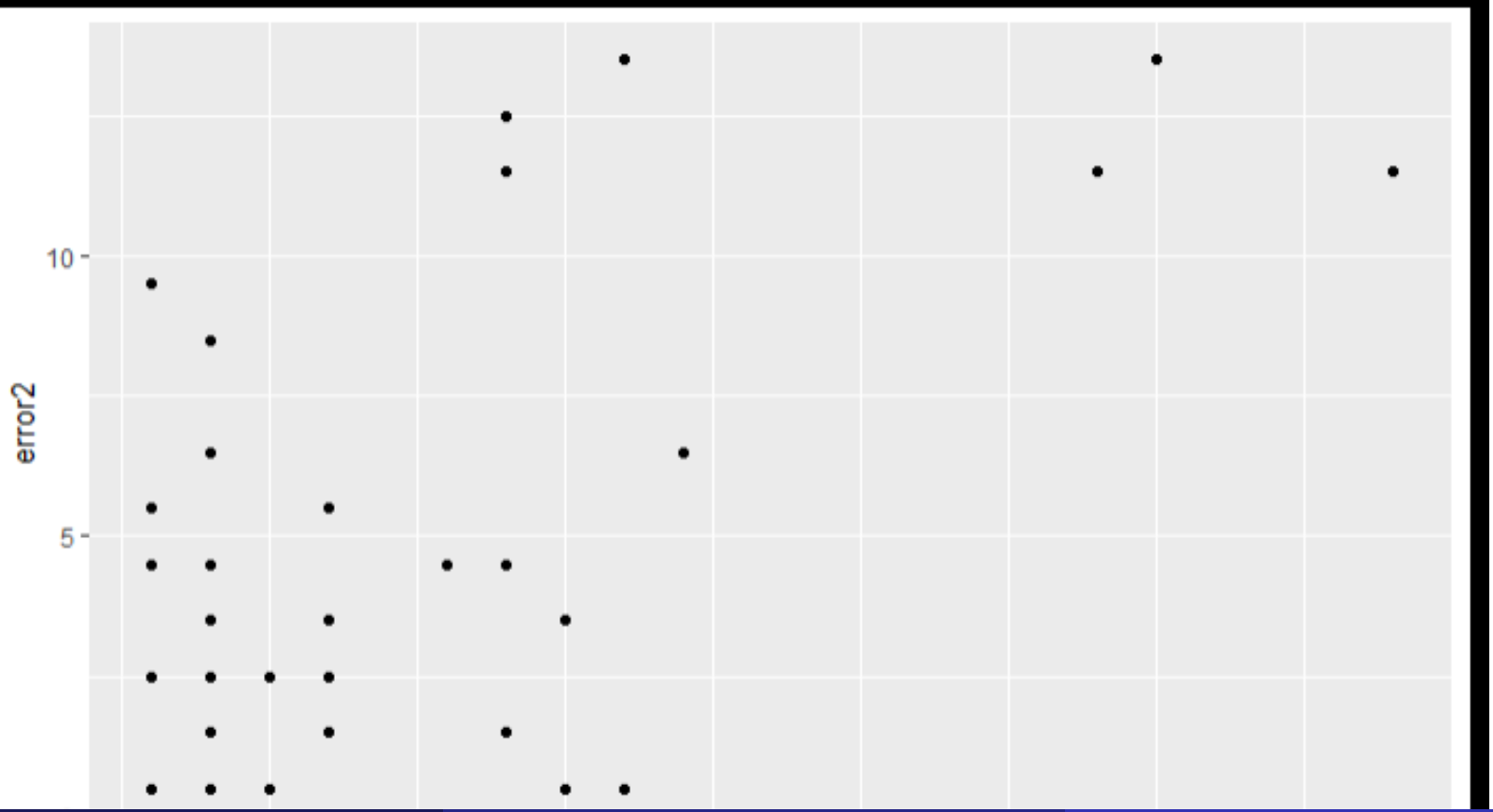
Build a scatterplot to compare the errors

```
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 ```
```

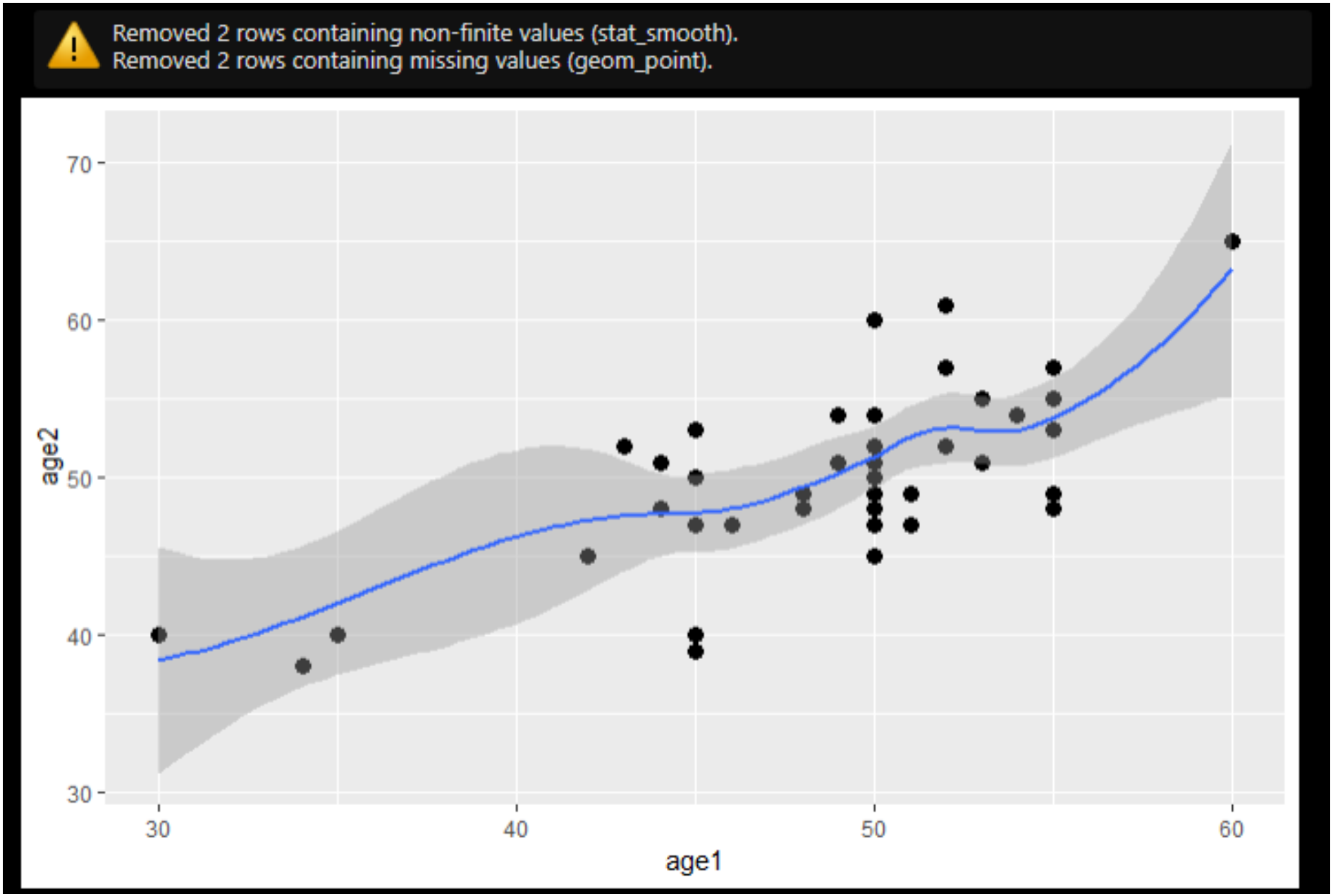
Removed 2 rows containing missing values (geom_point).



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
```

The Result

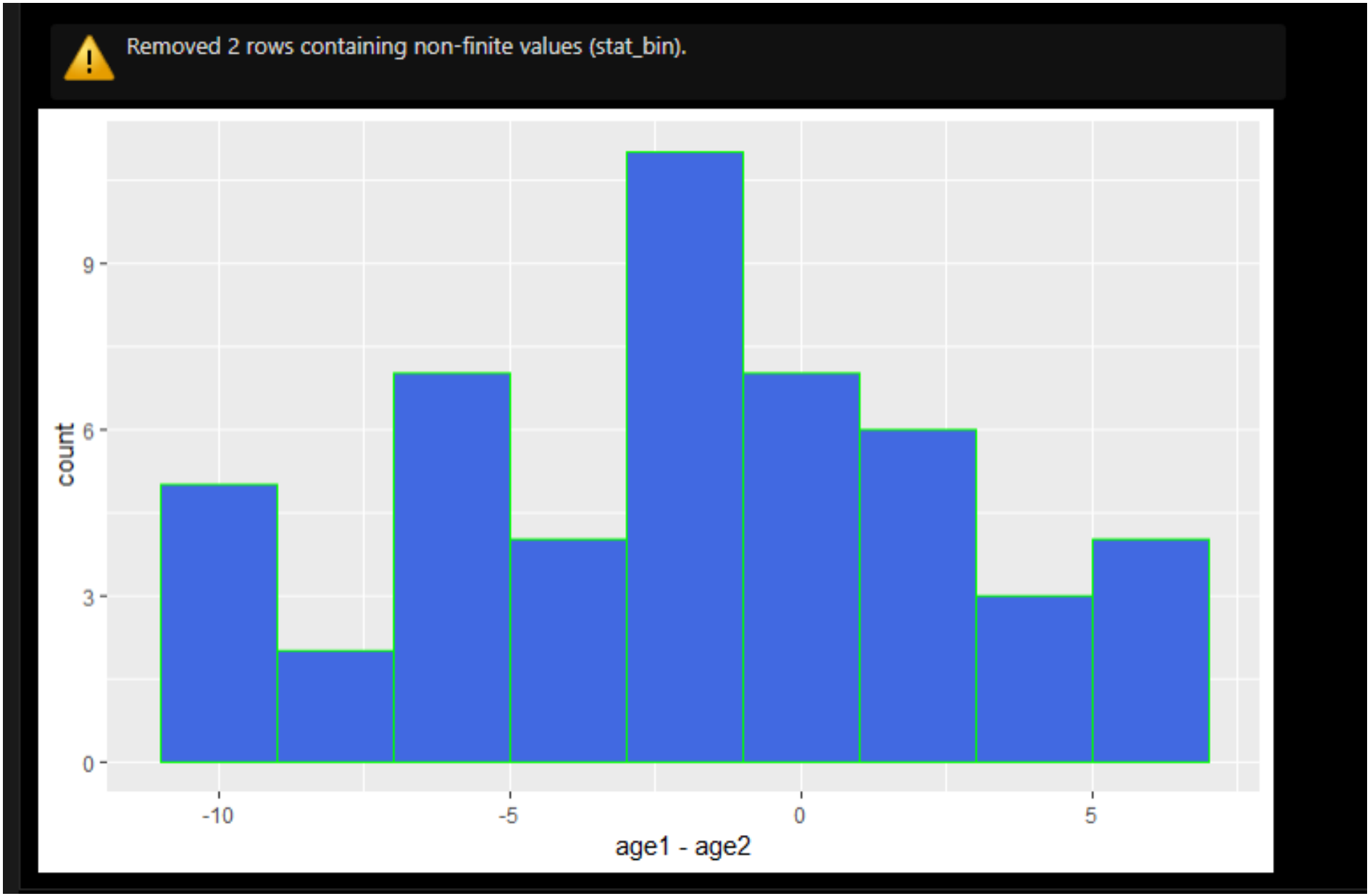

Removed 2 rows containing non-finite values (stat_smooth).
Removed 2 rows containing missing values (geom_point).
70
60
50
40
30
age2
30
40
50
60
age1

Plot the age1 - age2 differences

```
# Plot the (matched) differences

```{r}
ggplot(love_2018, aes(x = age1 - age2)) +
    geom_histogram(binwidth = 2,
                   col = "green", fill = "royalblue")
```
```

The Result


Removed 2 rows containing non-finite values (stat_bin).
count
9
6
3
0
-10
-5
0
5
age1 - age2

Numerical summary of the age1 - age2 differences

```
# Numerical Summary of Difference in Age Guesses

```{r}
love_2018 %$%
    summary(age1 - age2)
```
```

```
   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?

```
# How many people thought I looked younger in Guess 2?

```{r}
love_2018 %>%
    count(age1 - age2 < 0)
```
```

age1 - age2 < 0 <lgl>	n <int>
FALSE	20
TRUE	29
NA	2

3 rows

T tests - making a statistical inference

```
# The Much-Dreaded t test

```{r}
love_2018 %$%
    t.test(age1 - age2)
```
```

```
        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 --
```

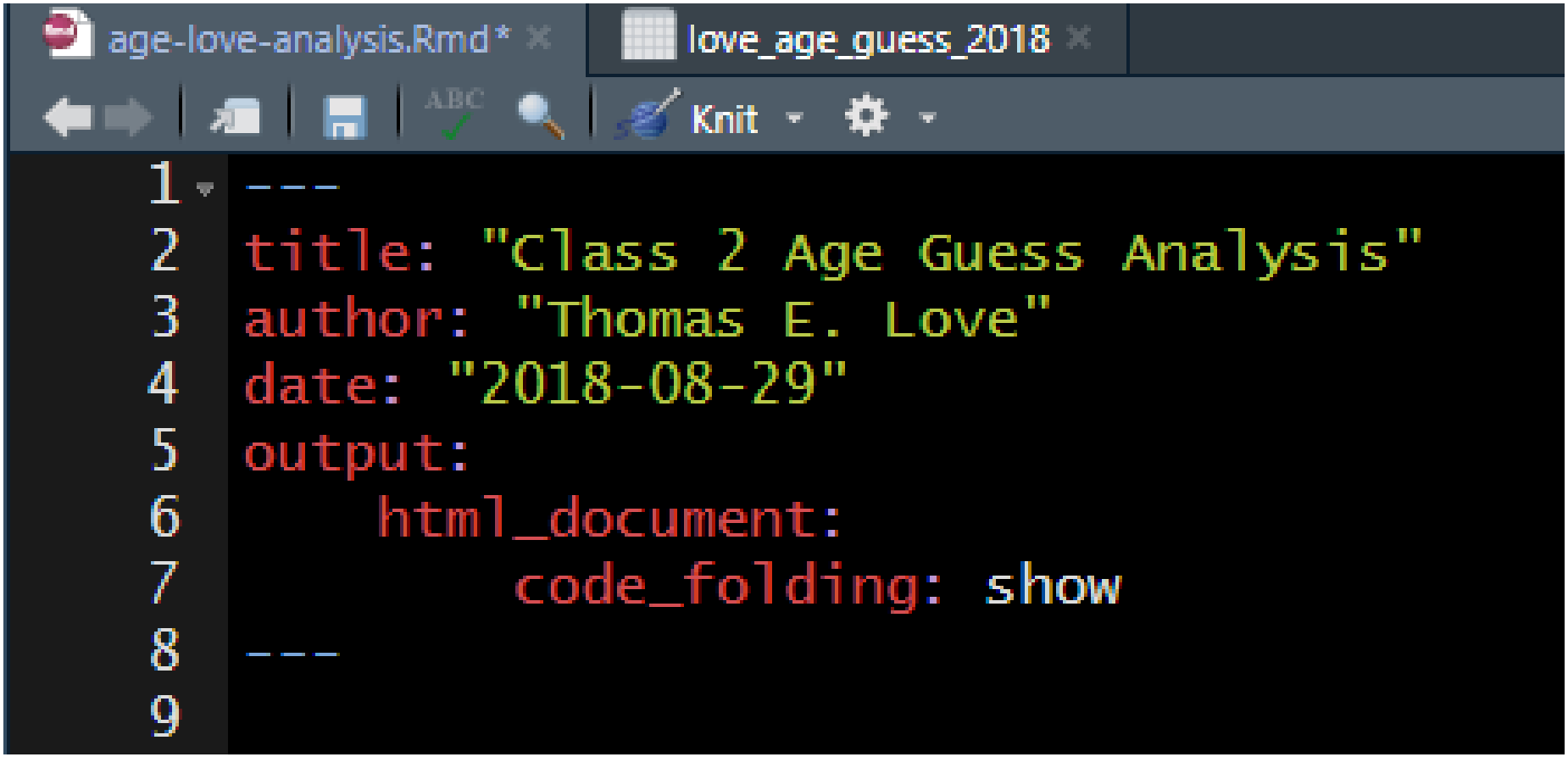
```
v ggplot2 3.0.0     v purrr   0.2.5
v tibble  1.4.2     v dplyr   0.7.6
v tidyr   0.8.1     v stringr 1.3.1
v readr   1.1.1     v forcats 0.3.0
```

```
-- Conflicts ---------------------------- tidyverse_conflicts() --
x tidyr::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



New, more final, report

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

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

Class 2 Age Guess Analysis

Code

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 tidyr   0.8.1     v stringr 1.3.1
v readr   1.1.1     v forcats 0.3.0
```

```
-- Conflicts ---------------------------- tidyverse_conflicts() --
x tidyr::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

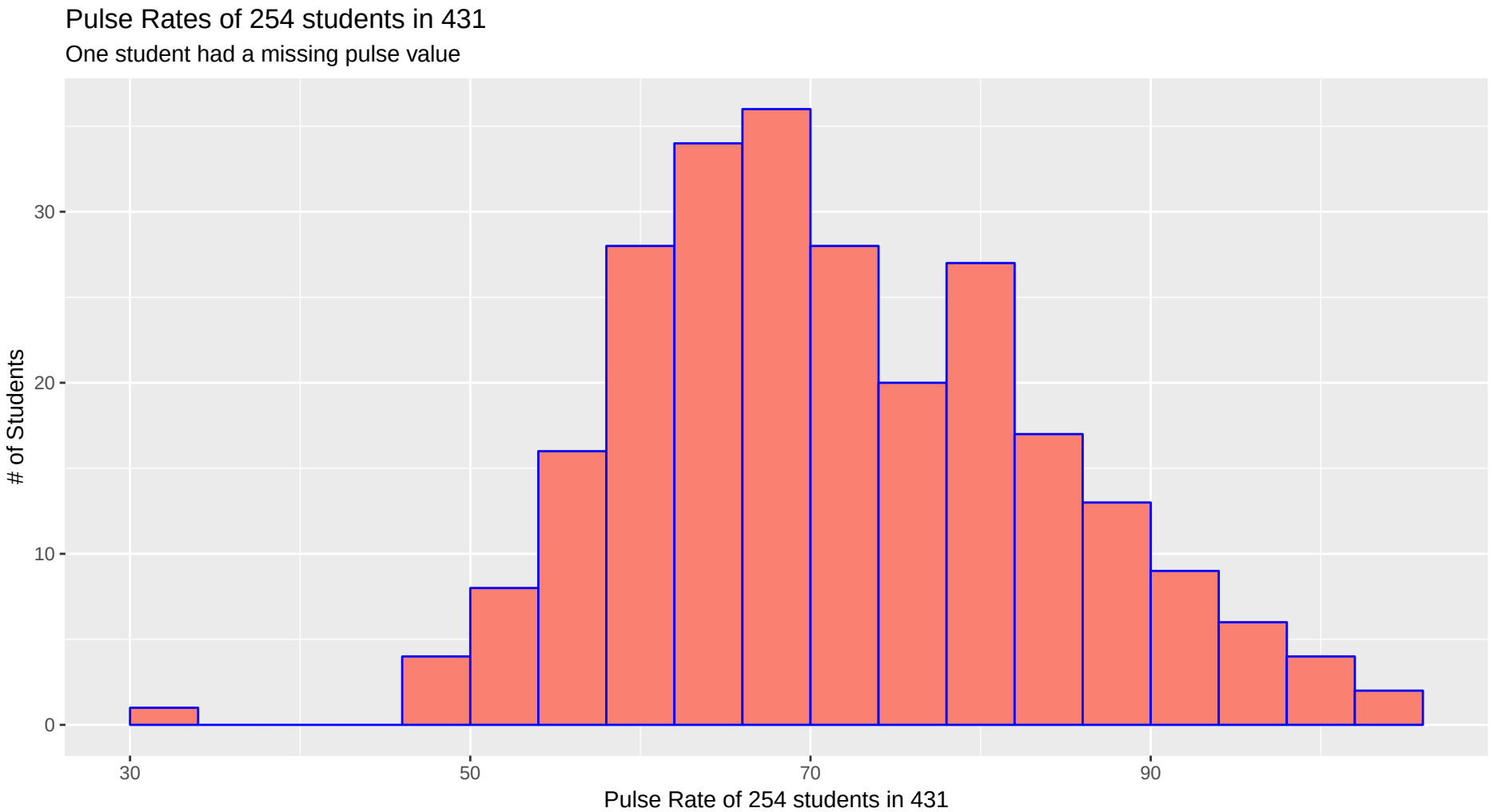
We have data on the site in a file called `surveyday1_2018.csv`. Build a project to study those data.

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

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

Can you reproduce the following. . .

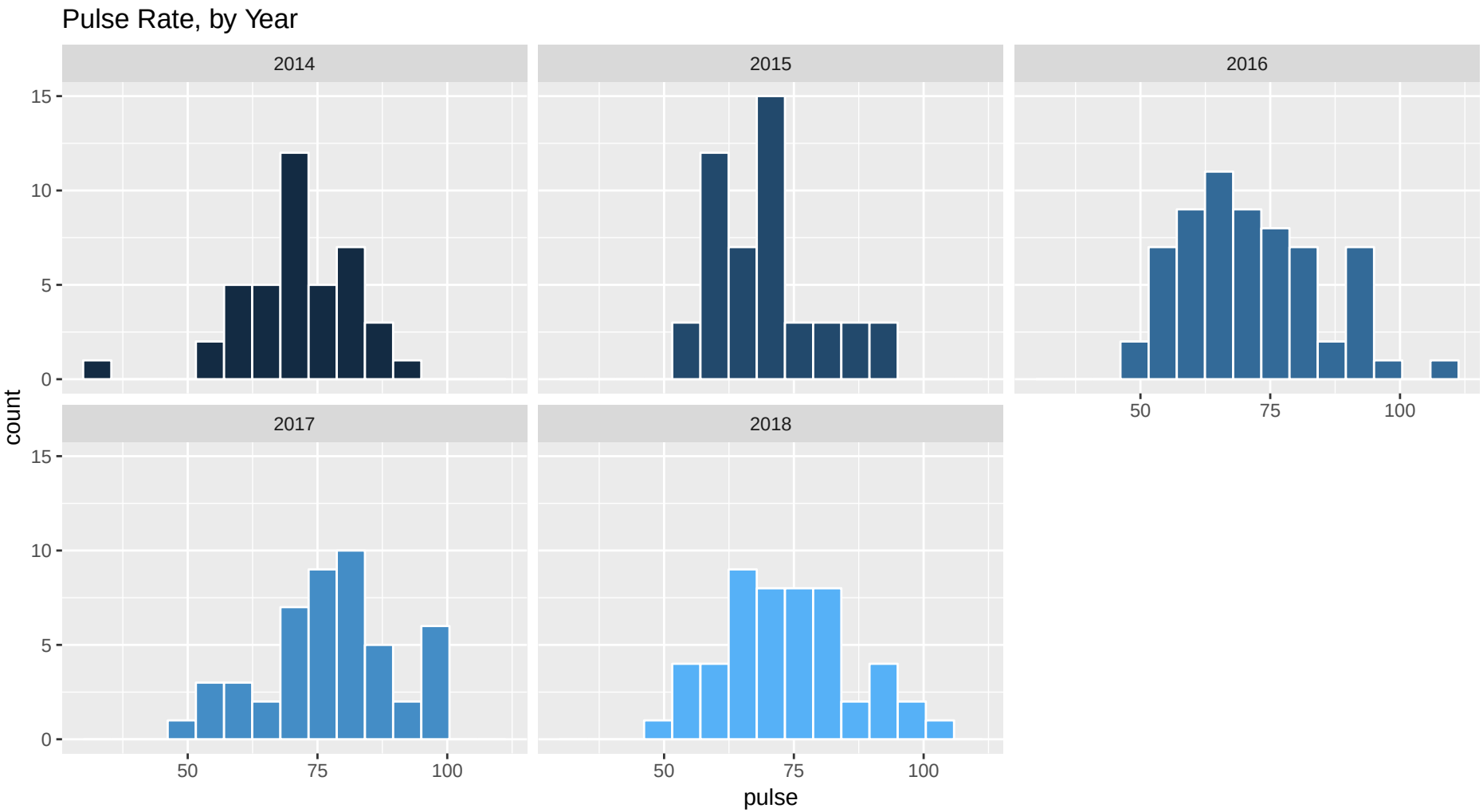
A. That fill color is called *salmon*, I used 20 bins.



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")
```

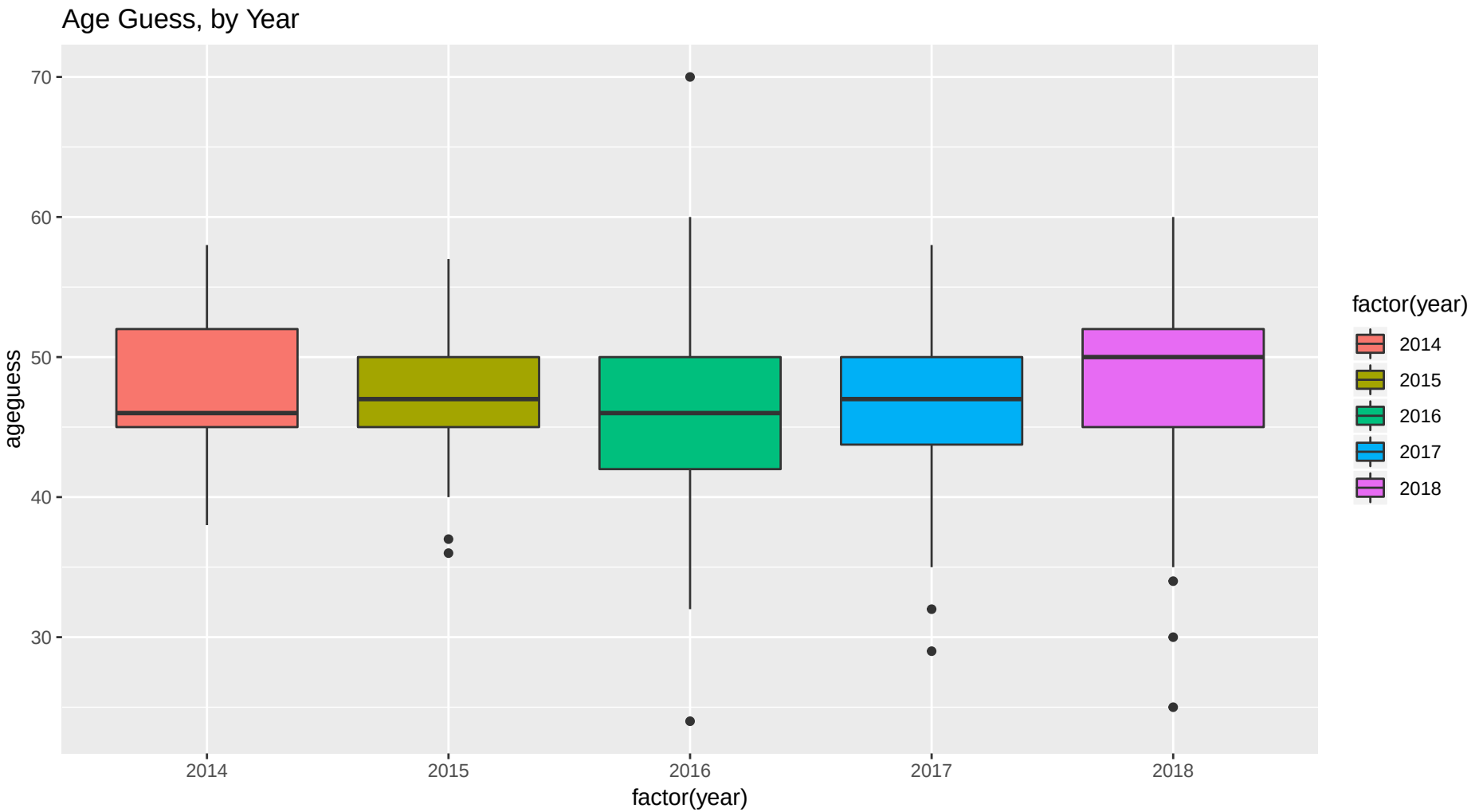
B. Histograms of Pulse Rates, Faceted by Year



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