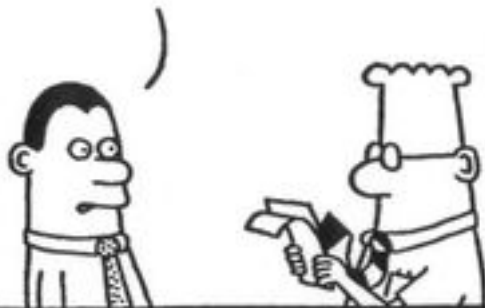


# 431 Class 02

Thomas E. Love

2019-08-29

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

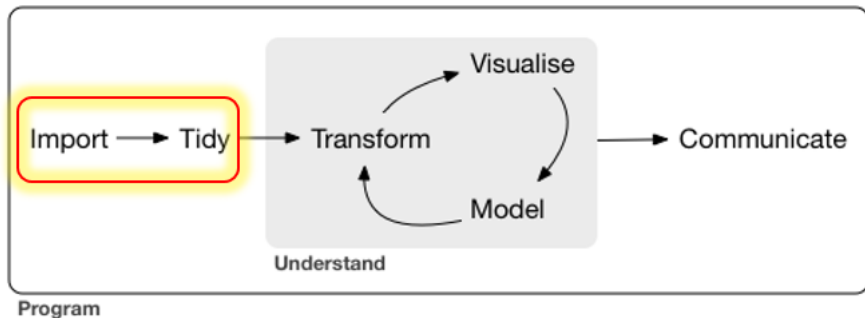


# Today's Agenda

- 1 The Class 1 Survey, Asking Questions
- 2 Some Administration
- 3 Using R, RStudio and R Markdown and the 431 RStudio Cloud

Contact us at [431-help@case.edu](mailto:431-help@case.edu)

Our web site: <https://github.com/THOMASELOVE/431-2018>



# Types of Data (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!

- Do you wear corrective lenses (contacts or glasses)? (Yes or No) \_\_\_\_\_
- Is English your *most comfortable* language? (Yes or No) \_\_\_\_\_
- 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

- How old (in years) do you think Professor Love is? \_\_\_\_\_ years.
- Do you smoke? Fill in the appropriate circle:  
 No I used to. Yes.  
 Non-Smoker Former Smoker 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 = \_\_\_\_\_  $\frac{\text{Right} - \text{Left}}{\text{Right} + \text{Left}}$  = \_\_\_\_\_

## 431 First Day Survey (15 Questions)

- 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

- 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

- What is your height (indicate units of measurement): \_\_\_\_\_

- 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.

- What is your favorite color? \_\_\_\_\_

- How many hours did you sleep last night? \_\_\_\_\_ hours.

- 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

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	student	sex	glasses	english	statsofar	ageguess	smoke	h.left	h.right	handedness	statfuture	haircut	lecture	alone	height.in	hand.s
2	201901	NA	y	y	6	42	1	1	19							
3	201902	NA	y	y	7	53	1	19	10							
4	201903	NA	y	y	4	45	1	0	10							
5	201904	NA	y	y	7	45	1	16	10							
6	201905	NA	y	y	6	42	1	2	16							
7	201906	NA	y	y	7	50	1	10	0							
8	201907	NA	y	y	5	56	1	1	13							
9	201908	NA	n	n	6	50	1	0	10							
10	201909	NA	n	y	6	52	1	0	17							
11	201910	NA	n	y	4	42	1	18	10							
12	201911	NA	n	n	5	43	1	5	13							
13	201912	NA	y	y	5	52	1	1	13							
14	201913	NA	y	y	7	50	2	1	19							
15	201914	NA	y	y	4	50	1	1	9							

165 cm to inches - Google Search

google.com/search?q=165+cm+to+inches&oq=165+cm+&aqs=chrome.2.0j69j57j0l4...

Apps ★ Bookmarks NY Times 538 NYT Crossword Slate The Athletic sAm

Google 165 cm to inches

Q All Shopping News Images Videos More Setting

About 20,900,000 results (0.56 seconds)

Length

165 = 64.9606

Centimeter Inch

Formula divide the length value by 2.54

# Day 1 Survey

- 61 people completed it Tuesday. Prior counts:

Fall	2019	2018	2017	2016	2015	2014	Total
$n$	61	51	48	64	49	42	<b>315</b>

## Question 1

About how many of those 315 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. Do you wear corrective lenses (contacts or glasses)? (Yes or No) \_\_\_\_\_
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  
 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 = \_\_\_\_\_  $\frac{\text{Right} - \text{Left}}{\text{Right} + \text{Left}}$  = \_\_\_\_\_

## 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.

# The 15 Survey Items

#	Topic	#	Topic
Q1	glasses	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 315 surveys caused *no problems* in recording responses?

- Guesses?

# Question 1

About how many of those 315 surveys caused *no problems* in recording responses?

- Guesses?
- 110/315 (35%)

# Question 1

About how many of those 315 surveys caused *no problems* in recording responses?

- Guesses?
- 110/315 (35%)
- 20 of the 61 surveys turned in Tuesday had **no** problems (33%)

# 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) \_\_\_\_\_
2. Is English your *most comfortable* language? (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

# Most Popular Colors in 2019

```
survey1 %>%  
  filter(year == 2019) %>%  
  count(favcolor)
```

```
# A tibble: 13 x 2
```

	favcolor	n
	<chr>	<int>
1	black	1
2	blue	23
3	dark green	1
4	gray	1
5	green	9
6	light blue	1
7	light purple	1
8	pink	3
9	purple	10
10	red	7

# Most Popular Colors in 2019

```
survey1 %>%  
  filter(year == 2019) %>%  
  count(favcolor, sort = TRUE)
```

```
# A tibble: 13 x 2
```

	favcolor	n
	<chr>	<int>
1	blue	23
2	purple	10
3	green	9
4	red	7
5	pink	3
6	teal	2
7	black	1
8	dark green	1
9	gray	1
10	light blue	1

# Following the Rules?

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

75 beats/minute.

## 2019 pulse responses, sorted ( $n = 61$ , 1 NA)

33	46	48	56	60	60	3		3
62	63	65	65	66	66	4		68
68	68	68	69	70	70	5		6
70	70	70	70	70	70	6		002355668889
71	72	72	74	74	74	7		00000000122444445666888
74	74	75	76	76	76	8		000012445668
78	78	78	80	80	80	9		000046
80	81	82	84	84	85	10		44
86	86	88	90	90	90	11		0
90	94	96	104	104	110			

# Stem and Leaf: Pulse Rates, 2014-2019

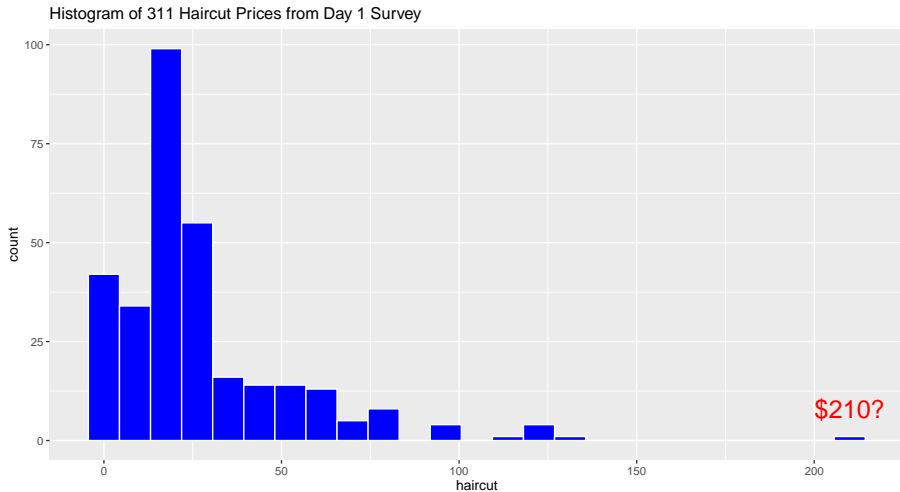
```
> stem(survey1$pulse)
```

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

```
3 | 03
3 |
4 |
4 | 688
5 | 00022244444
5 | 566666666667888889
6 | 0000000000000000000022222222223344444444444444
6 | 555555666666666666666666668888888888888888888899
7 | 00000000000000000000000000001222222222222222244444444444444444
7 | 555556666666667888888888888888
8 | 00000000000000000000000000001222222444444444444
8 | 55666666666668888888
9 | 0000000000001222224444
9 | 5668888
10 | 0000444
10 | 6
11 | 0
```

(Thanks, John **Tukey** )

# 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): 27 cm.

## Hand Span Numerical Summaries

```
summary(survey1$hand.span)
```

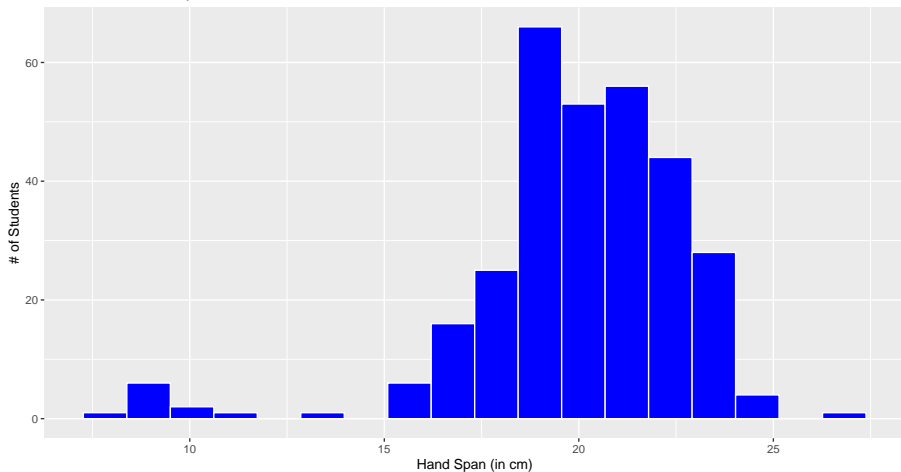
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8.00	19.00	20.00	19.94	21.70	27.00
NA's					
5					



# Hand Span (cm) Histogram

Warning: Removed 5 rows containing non-finite values (stat\_bin).

2014–2019 Hand Span measurements



# Hand Span (cm) Histogram (Code)

```
ggplot(data = survey1, aes(x = hand.span)) +  
  geom_histogram(bins = 18, col = "white", fill = "blue") +  
  labs(x = "Hand Span (in cm)",  
        y = "# of Students",  
        title = "2014-2019 Hand Span measurements")
```

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

The decimal point is at the |

```
 8 | 0500055
10 | 000
12 | 5
14 | 5
16 | 0000058000000000000000555
18 | 000000000000000000000055555555555555556900000000+30
20 | 000000000000000000000000000000000000234455555555+48
22 | 000000000000000000000000000000000000245555556800000+4
24 | 00000001000
26 | 0
```

## Eight Items had just a few problems

#	Topic	#	Topic
-	glasses	-	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>	-	-

Of the remaining seven items (glasses, stats so far, smoke, handedness, stats future, lectures vs activities, height), 5 had only minimal 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
11. What is your height (indicate 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 (2016-19 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.”



# Administrative Details

# TA Office Hours start Tuesday 2019-09-03

This schedule is found at the bottom of the Course Calendar, as well as on the Class 02 README.

- Mondays 11:30 - 12:45
- Tuesdays 11:30 - 12:45, 2:30 - 3:45 and 5:30 - 6:45
- Wednesdays 12:45 - 2:00
- Thursdays 11:30 - 12:45, 2:30 - 3:45 and 5:30 - 6:45
- Fridays 11:30 - 2:00

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

Find me (office: Wood WG-82J) T Th 12:30 - 1 and immediately after class, or email to make an appointment.

Contact us at [431-help@case.edu](mailto:431-help@case.edu)

Our web site: <https://github.com/THOMASELOVE/431-2018>

# Some Course Policies (1 of 2)

- Course Project (details after Labor Day)
  - Quizzes (2 or 3)
  - Class Participation (including TA office hours, surveys, corrections)
  - Regular Deliverables (Homework)
- 1 Each deliverable (except A) is worth 100 points.
  - 2 Dr. Love will throw out your lowest score during the semester on those 100-point deliverables.
  - 3 **Things happen.** If you have to miss a deliverable, you need to email Dr. Love as soon as possible. He will excuse the first missed deliverable (no questions asked, no details needed) but will do so only in dire circumstances otherwise.

## Some Course Policies (2 of 2)

- ④ We **do not accept** deliverables that are more than an hour late, because we post answer sketches to deliverables an hour after they're due. Late = not done.
- ⑤ **Attendance** Sometimes you may need to miss a class. We don't pay attention to attendance until after Class 04, but after that, we'd like to hear from you (no need for any details - just let us know) if you're going to miss **more than one class in a row**. Missing more than three classes over the term is a problem, and you should also email Dr. Love if that becomes necessary. You're responsible for anything you miss.
- ⑥ Want to complain about a grade? Read the last section of the Syllabus first.

# Deliverable A due Tomorrow at 2 PM

- 1 Review a plot I've prepared for you, using the DNase data set available in the automatically loaded `datasets` package in R. (You'll want to look at the Help window, and search for DNase to learn more about what's involved.) Submit a paragraph, as described, in Word or PDF format to Canvas.
- 2 Complete a (Google Form - must sign in through CWRU) survey about your attitudes toward statistics, and your thoughts on the first couple of classes. Make sure you hit the button to submit the form. (You'll receive an email confirmation.)

Deadline: 2 PM Friday. Worth 30 points. Graded exceedingly lightly.

[Deliverables page on our web site](#) for details.

# Using R and RStudio

# RStudio Cloud In-Class Early Project

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

If so, you'd want to:

- ➊ Get data from our site to a new directory on your machine.
- ➋ Open RStudio and start a new Project, in the new directory.
- ➌ Open and set up an R Markdown file to do the work.

But, perhaps you haven't gotten to that yet. So we have RStudio Cloud.

Link to join is: <http://bit.ly/431-2019-join-cloud>

# Analyzing the Index Card Guesses of My Age

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

For the slides, I have this information in a subfolder called data in my R Project.

```
love_2019 <- read_csv("data/love-age-guess-2019.csv")
```

Parsed with column specification:

```
cols(  
  subject = col_character(),  
  age1 = col_double(),  
  age2 = col_double()  
)
```



# The love\_2019 tibble

```
love_2019
```

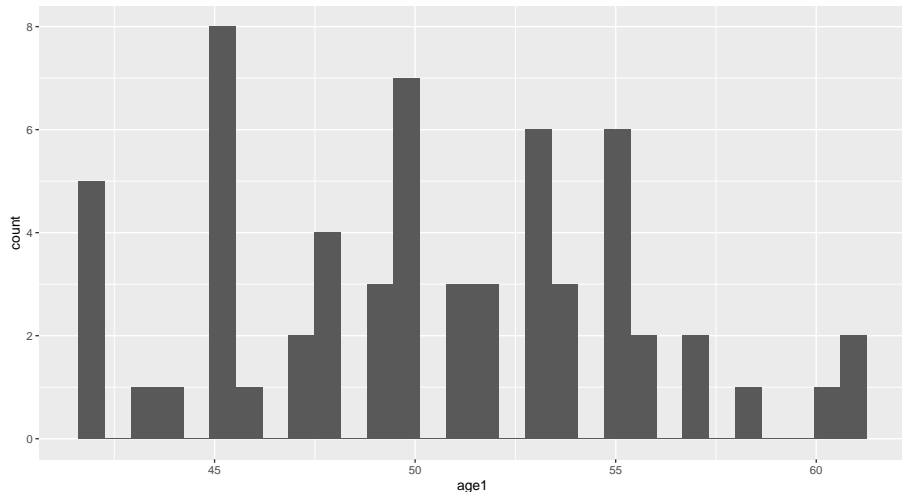
```
# A tibble: 61 x 3
  subject age1 age2
  <chr>   <dbl> <dbl>
1 S19-01     47    52
2 S19-02     55    59
3 S19-03     55    NA
4 S19-04     45    45
5 S19-05     45    48
6 S19-06     42    49
7 S19-07     43    55
8 S19-08     50    46
9 S19-09     54    50
10 S19-10     61    57
# ... with 51 more rows
```

# Histogram of initial guesses?

```
ggplot(data = love_2019, aes(x = age1)) +  
  geom_histogram()
```

# Histogram of initial guesses?

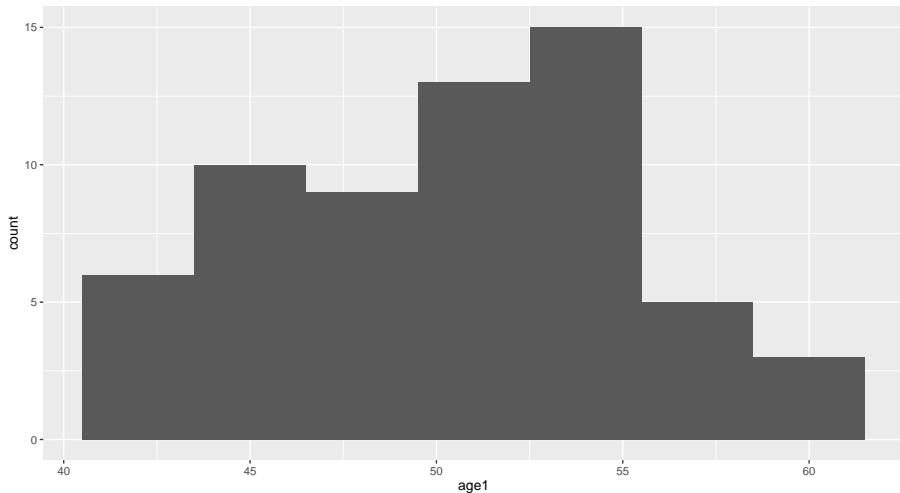
``stat_bin()`` using ``bins = 30``. Pick better value with ``binwidth``.



# Make the width of the bins 3 years?

```
ggplot(data = love_2019, aes(x = age1)) +  
  geom_histogram(binwidth = 3)
```

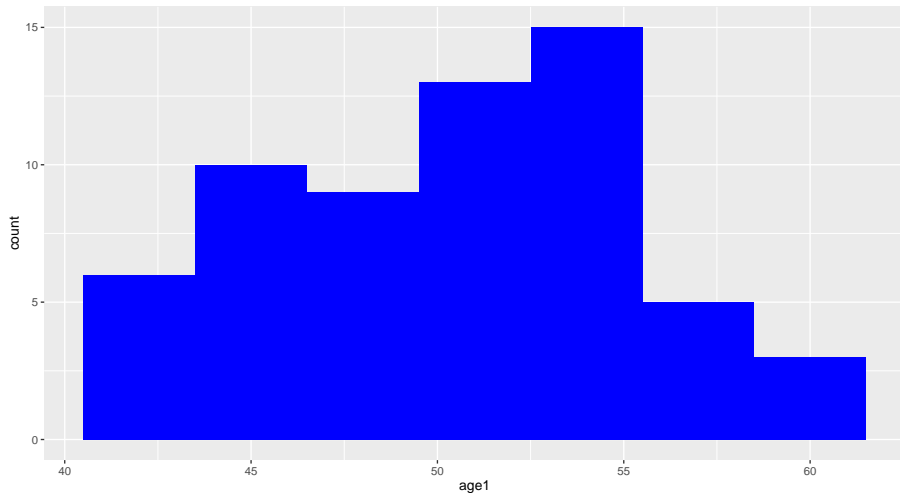
# Make the width of the bins 3 years?



# Fill in the bars with a better color?

```
ggplot(data = love_2019, aes(x = age1)) +  
  geom_histogram(binwidth = 3,  
                 fill = "blue")
```

# Fill in the bars with a better color?

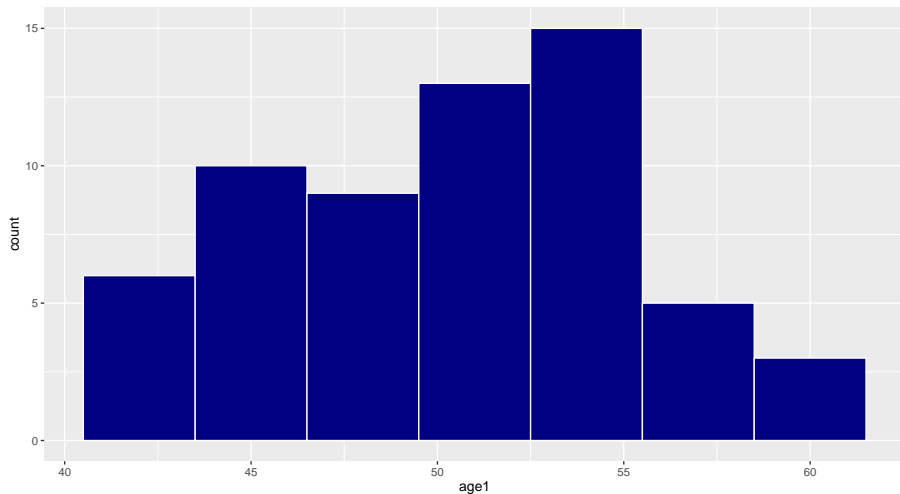


# Make it a little prettier?

```
ggplot(data = love_2019, aes(x = age1)) +  
  geom_histogram(binwidth = 3,  
                 fill = "navy", color = "white")
```



# Make it a little prettier?



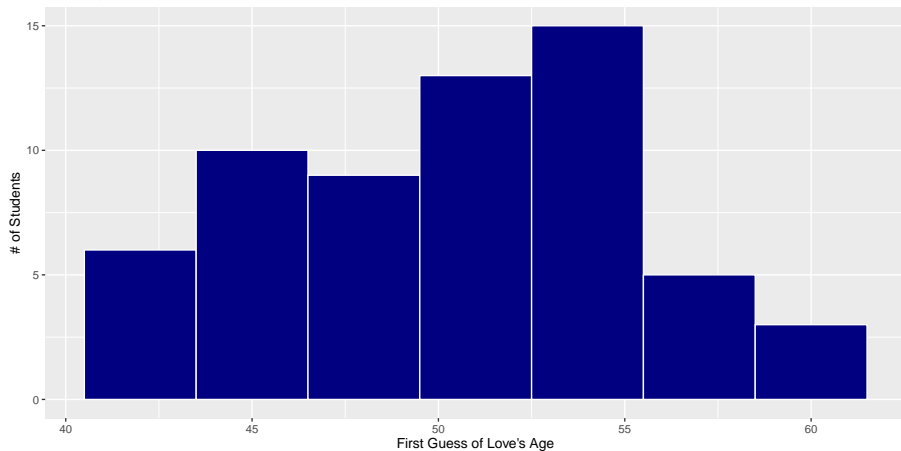
# Add more meaningful labels?

```
ggplot(data = love_2019, aes(x = age1)) +  
  geom_histogram(binwidth = 3,  
                 fill = "navy", color = "white") +  
  labs(x = "First Guess of Love's Age",  
       y = "# of Students",  
       title = "2019 Guesses of Professor Love's Age",  
       subtitle = "Actual Age was 52.5")
```

# Add more meaningful labels?

2019 Guesses of Professor Love's Age

Actual Age was 52.5



# Numerical Summaries of Age Guesses

```
summary(love_2019)
```

subject	age1	age2
Length:61	Min. :42.00	Min. :42.00
Class :character	1st Qu.:46.00	1st Qu.:48.75
Mode :character	Median :50.00	Median :52.00
	Mean :50.34	Mean :51.82
	3rd Qu.:54.00	3rd Qu.:55.00
	Max. :61.00	Max. :62.00
		NA's :1

# Some Additional Summaries

```
mosaic::favstats(~ age1, data = love_2019)
```

min	Q1	median	Q3	max	mean	sd	n	missing
42	46	50	54	61	50.34426	4.989607	61	0

```
mosaic::favstats(~ age2, data = love_2019)
```

min	Q1	median	Q3	max	mean	sd	n	missing
42	48.75	52	55	62	51.81667	4.545408	60	1

# Another Approach

```
love_2019 %>%  
  skimr::skim()
```

Skim summary statistics

n obs: 61

n variables: 3

-- Variable type:character -----

variable	missing	complete	n	min	max	empty	n_unique
subject	0	61	61	6	6	0	61

-- Variable type:numeric -----



variable	missing	complete	n	mean	sd	p0	p25	p50
age1	0	61	61	50.34	4.99	42	46	50
age2	1	60	61	51.82	4.55	42	48.75	52

p75 p100 hist

54 61 <U+2585><U+2586><U+2586><U+2587><U+2586><U+2587><U+2

# A Better Look

```
love_2019 %>%  
  skimr::skim()
```

```
Skim summary statistics  
n obs: 61  
n variables: 3  
  
-- Variable type:character -----  
variable missing complete  n min max empty n_unique  
subject         0         61 61  6  6      0         61  
  
-- Variable type:numeric -----  
variable missing complete  n  mean   sd p0    p25  p50  p75  p100  hist  
age1           0         61 61 50.34 4.99 42 46    50  54   61    
age2           1         60 61 51.82 4.55 42 48.75 52 55 62  
```

# What about the second guess?

```
ggplot(data = love_2019, aes(x = age2)) +  
  geom_histogram(binwidth = 3,  
                 fill = "forestgreen", color = "white") +  
  labs(x = "Second Guess of Love's Age",  
       y = "# of Students",  
       title = "2019 Guesses of Professor Love's Age",  
       subtitle = "Actual Age was 52.5")
```

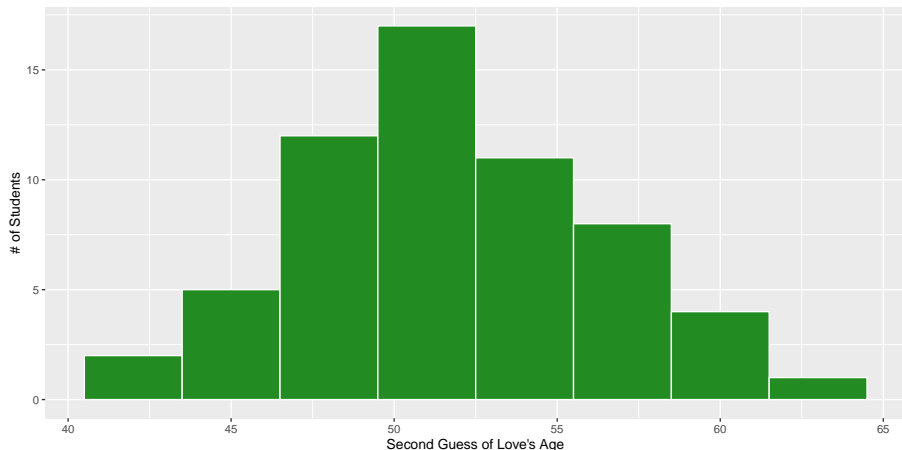


# What about the second guess?

Warning: Removed 1 rows containing non-finite values  
(stat\_bin).

2019 Guesses of Professor Love's Age

Actual Age was 52.5



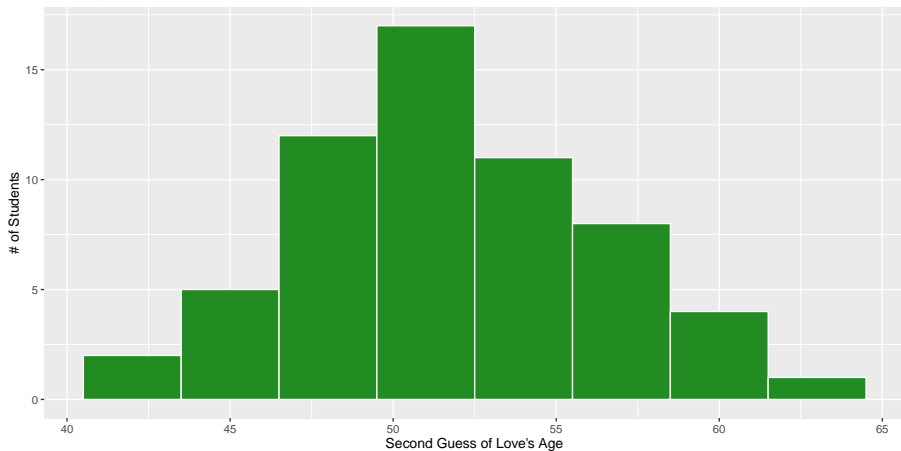
# Filter to complete cases only

```
love_2019 %>%  
  filter(complete.cases(age2)) %>%  
  ggplot(data = ., aes(x = age2)) +  
  geom_histogram(binwidth = 3,  
                 fill = "forestgreen", color = "white") +  
  labs(x = "Second Guess of Love's Age",  
       y = "# of Students",  
       title = "2019 Guesses of Professor Love's Age",  
       subtitle = "Actual Age was 52.5")
```

# Filter to complete cases only

2019 Guesses of Professor Love's Age

Actual Age was 52.5

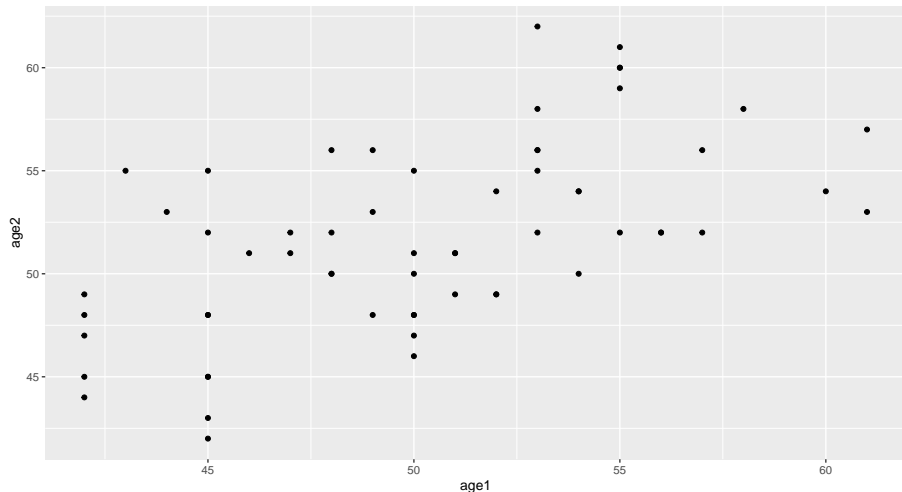


# Comparing First Guess to Second Guess

```
ggplot(data = love_2019, aes(x = age1, y = age2)) +  
  geom_point()
```

# Comparing First Guess to Second Guess

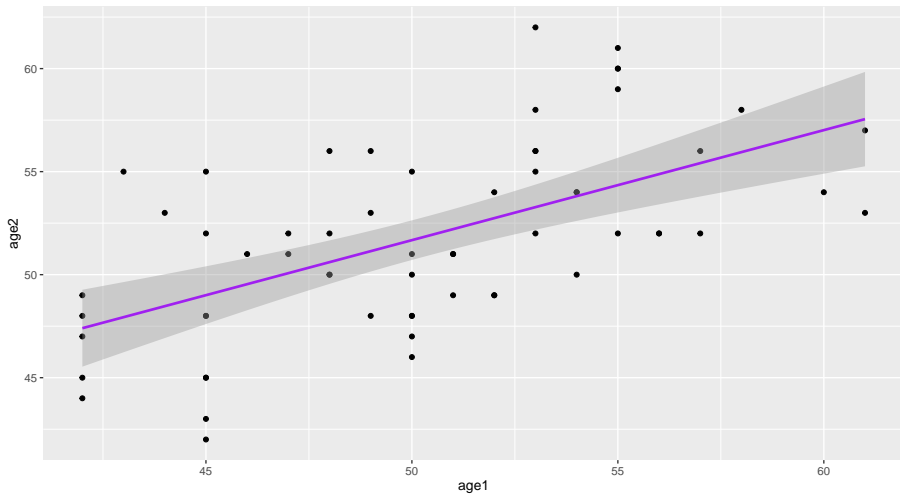
Warning: Removed 1 rows containing missing values  
(geom\_point).



# Filter to complete cases, add regression line

```
love_2019 %>%  
  filter(complete.cases(age1, age2)) %>%  
  ggplot(data = ., aes(x = age1, y = age2)) +  
  geom_point() +  
  geom_smooth(method = "lm", col = "purple")
```

# Filter to complete cases, add regression line



# What's that regression line?

```
lm(age2 ~ age1, data = love_2019)
```

Call:

```
lm(formula = age2 ~ age1, data = love_2019)
```

Coefficients:

(Intercept)	age1
24.973	0.534

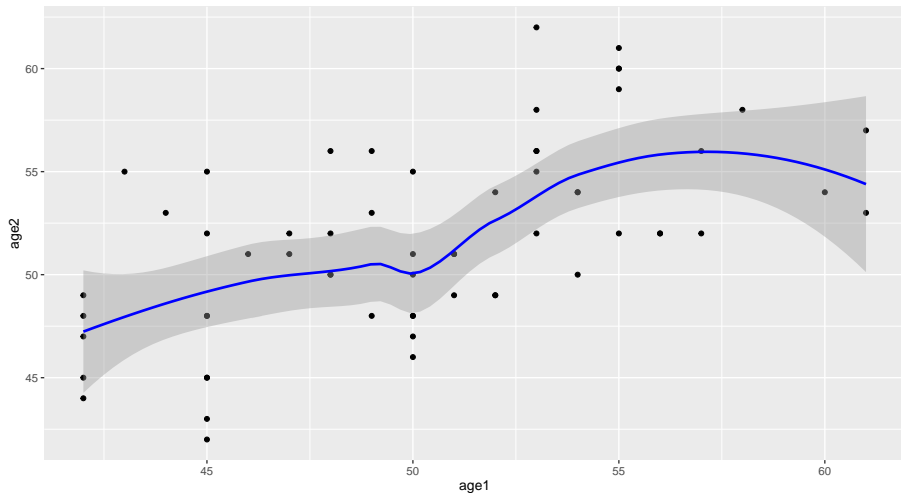
- `lm` (by default) filters to complete cases.



# How about a loess smooth curve, instead?

```
love_2019 %>%  
  filter(complete.cases(age1, age2)) %>%  
  ggplot(data = ., aes(x = age1, y = age2)) +  
  geom_point() +  
  geom_smooth(method = "loess", col = "blue")
```

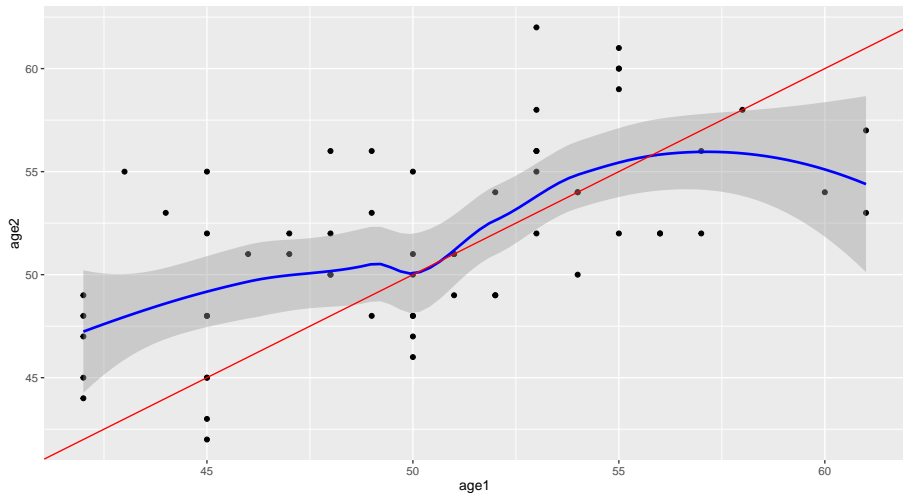
# How about a loess smooth curve, instead?



## Add a $y = x$ line (no change in guess)?

```
love_2019 %>%  
  filter(complete.cases(age1, age2)) %>%  
  ggplot(data = ., aes(x = age1, y = age2)) +  
  geom_point() +  
  geom_smooth(method = "loess", col = "blue") +  
  geom_abline(intercept = 0, slope = 1, col = "red")
```

# Add a $y = x$ line (no change in guess)?



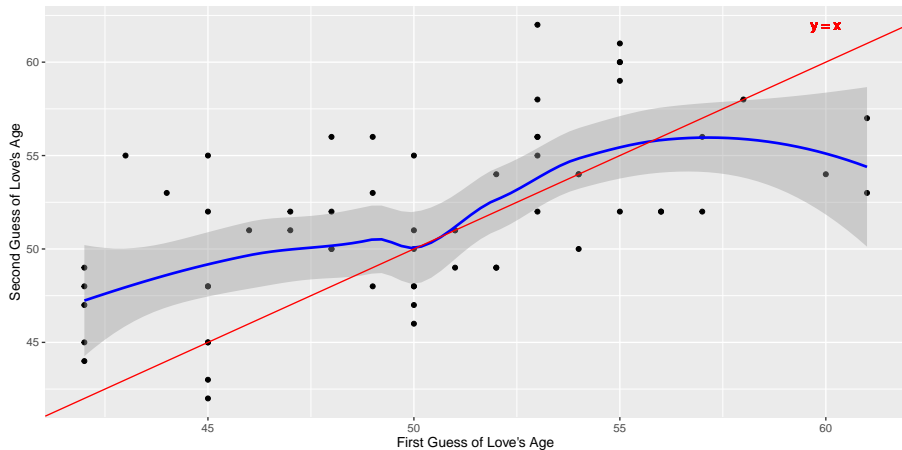
# Add more meaningful labels

```
love_2019 %>%  
  filter(complete.cases(age1, age2)) %>%  
  ggplot(data = ., aes(x = age1, y = age2)) +  
  geom_point() +  
  geom_smooth(method = "loess", col = "blue") +  
  geom_abline(intercept = 0, slope = 1, col = "red") +  
  geom_text(x = 60, y = 62,  
            label = "y = x", col = "red") +  
  labs(x = "First Guess of Love's Age",  
       y = "Second Guess of Love's Age",  
       title = "Comparing 2019 Age Guesses",  
       subtitle = "Love's actual age = 52.5")
```

# Add more meaningful labels

Comparing 2019 Age Guesses

Love's actual age = 52.5



# age1 - age2 difference in guesses?

```
love_2019 %>%  
  mutate(diff = age1 - age2) %>%  
  skimr::skim()
```

Skim summary statistics

n obs: 61

n variables: 4

-- Variable type:character -----

variable	missing	complete	n	min	max	empty	n_unique
subject	0	61	61	6	6	0	61

-- Variable type:numeric -----

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100
age1	0	61	61	50.34	4.99	42	46	50	54	61
age2	1	60	61	51.82	4.55	42	48.75	52	55	62
diff	1	60	61	-1.55	4.35	-12	-5	-2	2	8



# How Many Guesses Increased?

```
love_2019 %>%  
  mutate(diff = age1 - age2) %>%  
  count(diff < 0)
```

```
# A tibble: 3 x 2  
  `diff < 0`      n  
  <lgl>         <int>  
1 FALSE         28  
2 TRUE          32  
3 NA            1
```



# Increased / Stayed the Same / Decreased

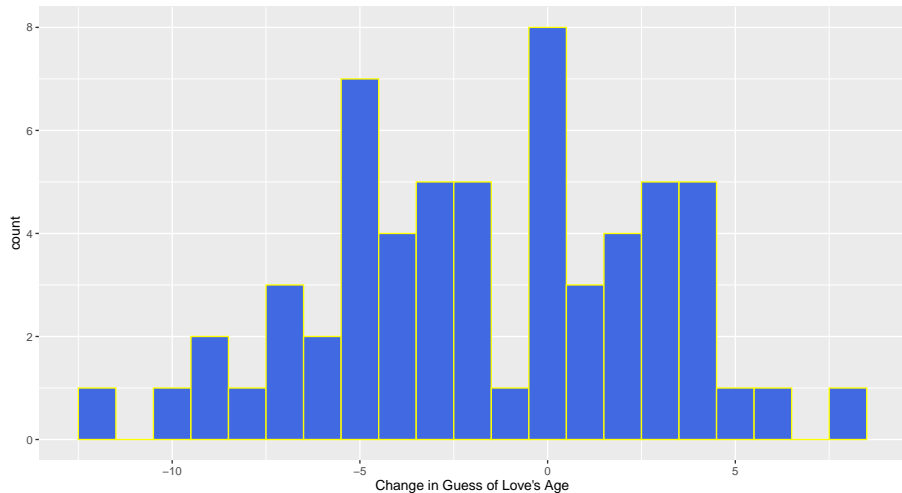
```
love_2019 %>%  
  mutate(diff = age1 - age2) %>%  
  count(sign(diff))
```

```
# A tibble: 4 x 2  
  `sign(diff)`      n  
    <dbl> <int>  
1      -1     32  
2       0      8  
3       1     20  
4      NA      1
```

# Histogram of difference in guesses

```
love_2019 %>%  
  mutate(diff = age1 - age2) %>%  
  filter(complete.cases(diff)) %>%  
  ggplot(data = ., aes(x = diff)) +  
  geom_histogram(binwidth = 1,  
                 fill = "royalblue", color = "yellow") +  
  labs(x = "Change in Guess of Love's Age")
```

# Histogram of difference in guesses



## Next Steps

# Analyzing the Survey Data - A little challenge

We have data on the site in a file called `surveyday1_2019.csv`. Build a project to study those data.

Put the data in a file called `survey1` in R.

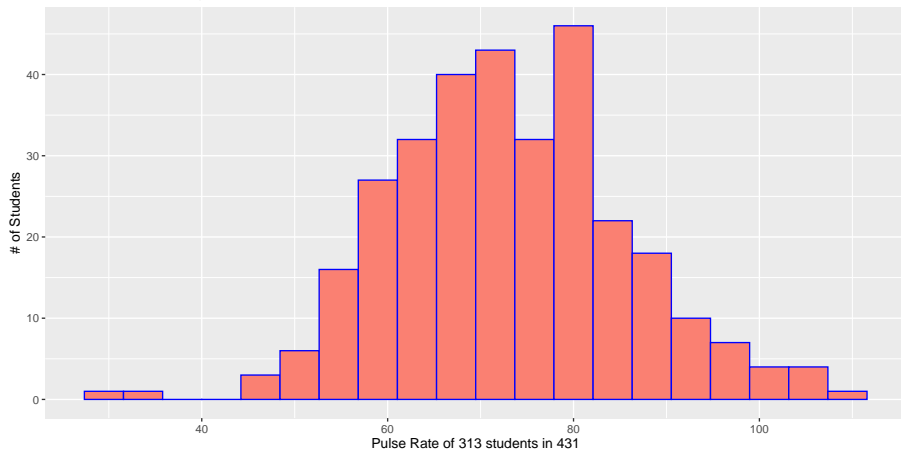
- I'd call my R Markdown file `day1surveyanalysis`

Can you reproduce the following...

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

Pulse Rates of 313 students in 431

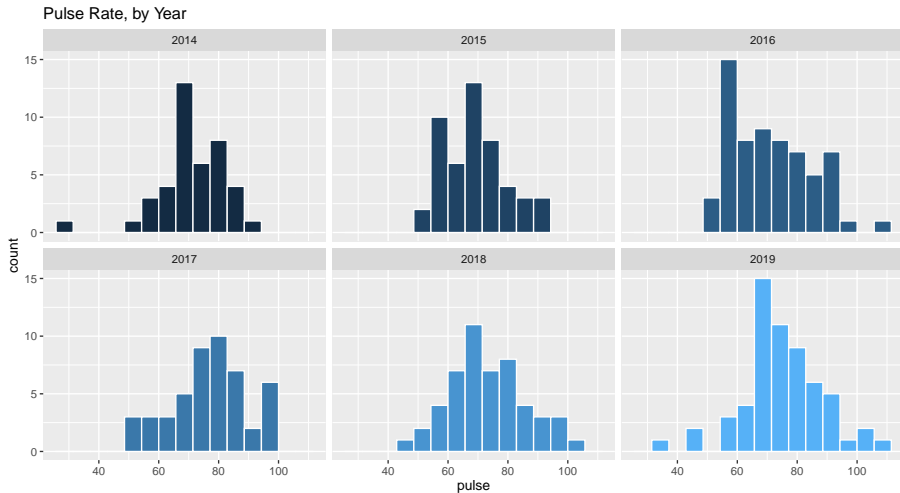
Two students had missing pulse values



# Code for Part A.

```
ggplot(survey1, aes(x = pulse)) +  
  geom_histogram(bins = 20, col = "blue", fill = "salmon") +  
  labs(x = "Pulse Rate of 313 students in 431",  
       y = "# of Students",  
       title = "Pulse Rates of 313 students in 431",  
       subtitle = "Two students had missing pulse values")
```

## B. Histograms of Pulse Rates, Faceted by Year

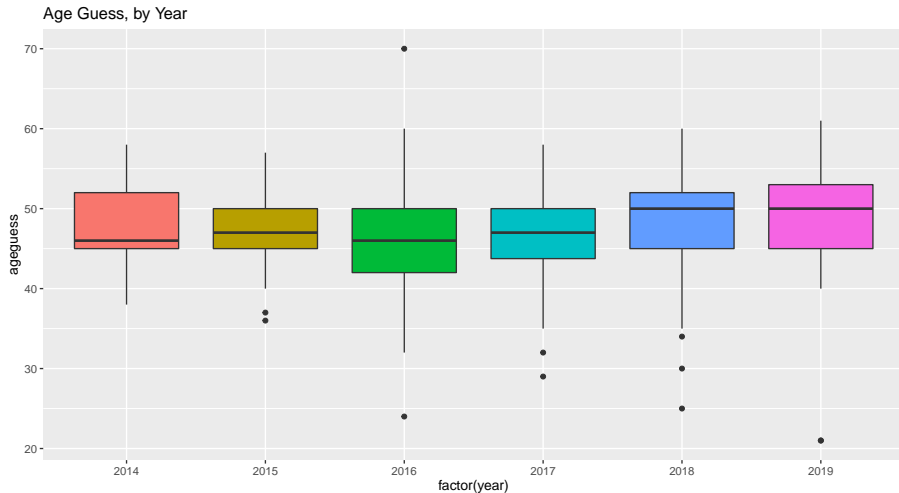




## Code for Plot B.

```
ggplot(survey1, 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(survey1, aes(x = factor(year), y = ageguess,  
                    fill = factor(year))) +  
  geom_boxplot() +  
  guides(fill = FALSE) +  
  labs(title = "Age Guess, by Year")
```

# Summary Table of Age Guesses, by Year

```
# A tibble: 6 x 5
  year      n mean    sd median
<dbl> <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
6  2019     61  48.6  7.09     50
```

# Code for Summary Table

```
survey1 %>%  
  group_by(year) %>%  
  summarize(n = n(),  
            mean = mean(ageguess, na.rm=TRUE),  
            sd = sd(ageguess, na.rm=TRUE),  
            median = median(ageguess, na.rm=TRUE)  
            )
```

# Reminders

- 1 Deliverable A due Friday at 2 PM.
- 2 Get R, RStudio, etc. installed on a machine you can use.
- 3 Sign up for RStudio Cloud at <http://bit.ly/431-2019-join-cloud> .
- 4 You might want to get started reading Jeff Leek's *Elements of Data Analytic Style*.