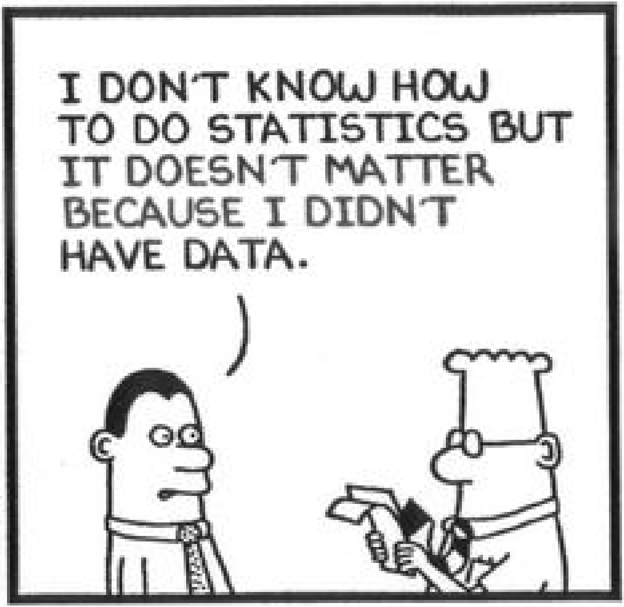
431 Class 02

thomaselove.github.io/431

2020-08-27


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

Today's Agenda

1. Asking Questions: The "Short" Survey
2. Doing Data Analysis and Understanding Limitations
3. Using R to manage and visualize some data

Chatfield's Six Rules for Data Analysis

1. Do not attempt to analyze the data until you understand what is being measured and why.
2. Find out how the data were collected.
3. Look at the structure of the data.
4. Carefully examine the data in an exploratory way, before attempting a more sophisticated analysis.
5. Use your common sense at all times.
6. Report the results in a clear, self-explanatory way.

Chatfield, Chris (1996) *Problem Solving: A Statistician's Guide*, 2nd ed.

Breakout: The "Short" Survey

- Goal: mimic the process for a telephone or in-person survey.

Each breakout group will have 3 or 4 people.

- Within your group, each of you will respond to the questions in the survey in turn, but **don't fill out the form for yourself**.
- Instead, one of you should share their screen and type in the responses spoken by the subject, so that you both know what the response is. Then switch roles, until everyone's responses to the questions have been recorded.
- The data are collected anonymously in this Google Form, and if you are uncomfortable answering any questions, leave the response blank.
- When you finish recording one person's results and submit the form, the system will give you a link to fill out the form again for another person.

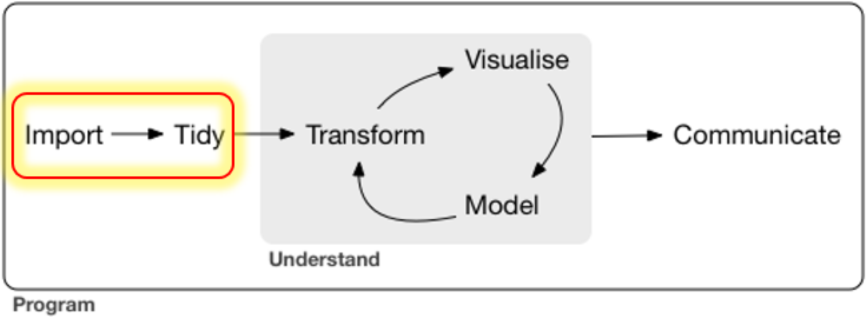
Breakout Session Now Underway

The survey is at **http://bit.ly/431-2020-class02-breakout**

Make sure everyone in your breakout session has a submitted set of responses. We hope this will take at most ten minutes.

- If you have some extra time, make sure you get to know one another a little bit, trying to ensure that everyone knows everyone else's name, and what they are studying or what they do professionally.

Data Science



Types of Data

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?

Evaluating some “Short” 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. Statistical thinking in your future **career**? (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*?

What was different in 2020?

- In the past, I've done this in Class 01, in person and using a paper form, gathering data in pairs (each person writes down the other's responses)

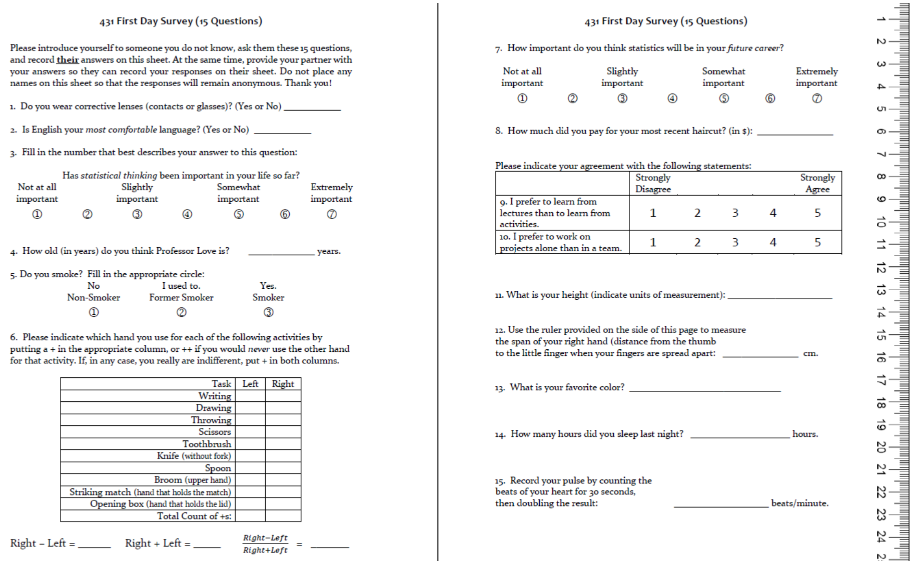
Items asked in 2019 (and earlier) but not 2020:

- `Q03` Has statistical thinking been important in your life **so far**? (1-7 on importance)
- `Q04` **How old** (in years) do you think Professor Love is?
- `Q12` Included ruler and asked for a **hand span** measurement in cm
- `Q15` Record your **pulse** by counting the beats of your heart for 30 seconds, then doubling the result.

Other differences:

- `Q06` 10-item handedness scale with alternate measurement scale
- `Q09-10` Changed wording of learning / projects stems

“Short” Survey Handout from 2019



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

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): ___________________

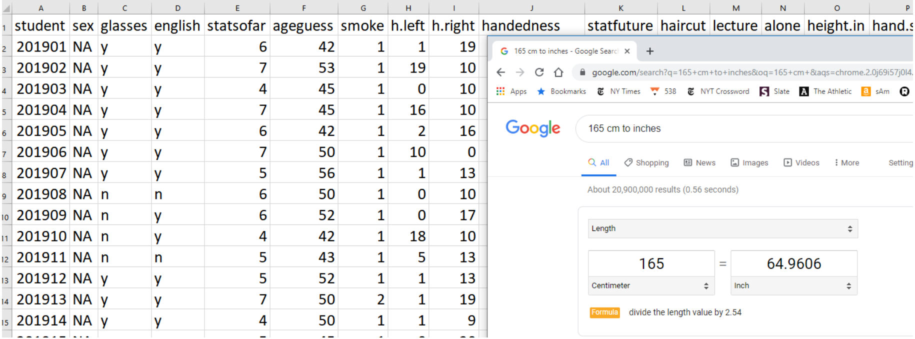
12. Use the ruler provided on the side of this page to measure the span of your right hand (distance from the thumb to the little finger when your fingers are spread apart: ______________ cm.

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.

Ingesting the Paper “Short” Surveys



	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 Searc
google.com/search?q=165+cm+to+inches&oq=165+cm+&aqs=chrome.2.0j69i57j0l4.
Apps
Bookmarks
NY Times
538
NYT Crossword
Slate
The Athletic
sAm
Google
All
Shopping
News
Images
Videos
More
Setting

165 cm to inches

About 20,900,000 results (0.56 seconds)

Length

165 = 64.9606

Centimeter

Inch

Formula divide the length value by 2.54

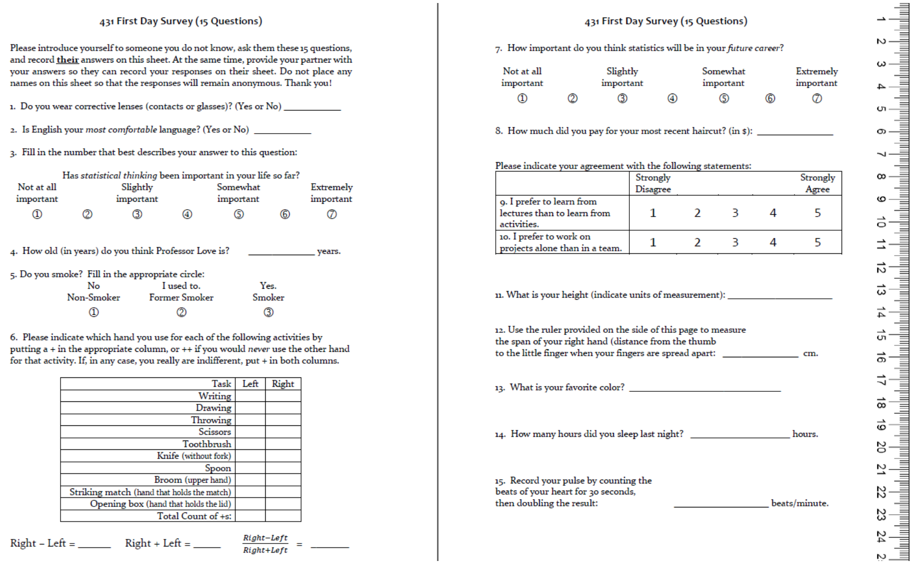
“Short” Survey

Fall	2019	2018	2017	2016	2015	2014	Total
n	61	51	48	64	49	42	**315**

Poll Question

What percentage of those 315 paper 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 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 = _____ $\frac{Right-Left}{Right+Left}$ = _______

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
Q01	`glasses`	**Q09**	`lectures v activities`
Q02	`english`	**Q10**	`projects alone`
Q03	`stats so far`	**Q11**	`height`
Q04	`guess TL age`	Q12	`hand span`
Q05	`smoke`	**Q13**	`color`
Q06	`handedness`	**Q14**	`sleep`
Q07	`stats future`	Q15	`pulse rate`
Q08	`haircut`	-	-

(Bolded items were asked in the 2020 Google Form version.)

Question 1

What percentage of those 315 paper surveys caused *no problems* in recording responses?

- OK. Take the poll now.

Question 1

What percentage of those 315 paper surveys caused *no problems* in recording responses?

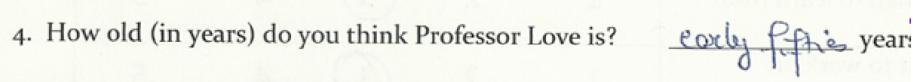
- OK. Take the poll now.
- First, we'll get the poll results.

Question 1

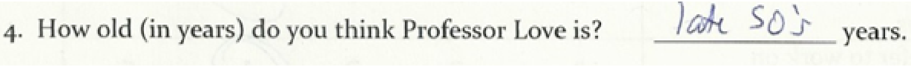
What percentage of those 315 paper surveys caused *no problems* in recording responses?

- OK. Take the poll now.
- First, we'll get the poll results.
- 110/315 were clean and caused no problems, or **35**%.

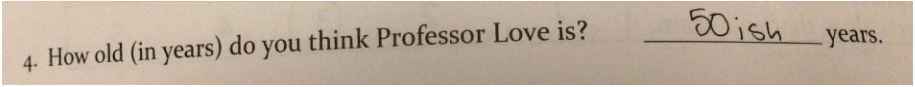
Guess My Age



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



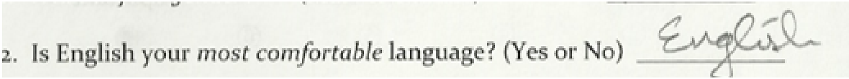
4. How old (in years) do you think Professor Love is? late 50's years.



4. How old (in years) do you think Professor Love is? 50ish years.

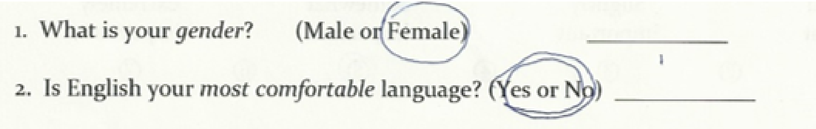
What should we do in these cases?

English best language?



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

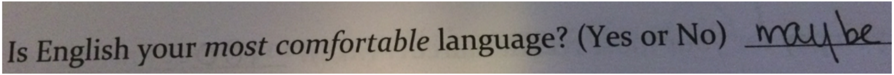
TEL Decision: **Yes**



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

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

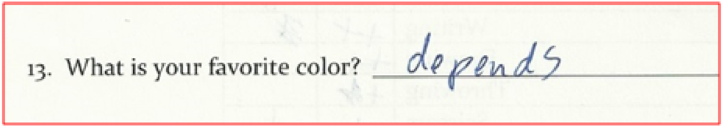
TEL Decision: **NA**



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

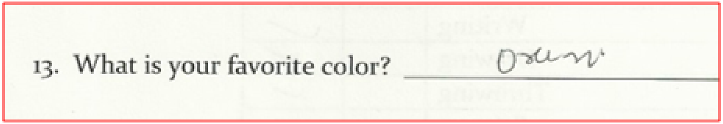
TEL decision: NA

Favorite color



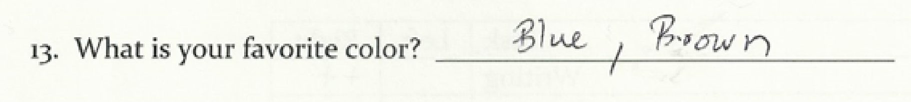
13. What is your favorite color? depends

NA

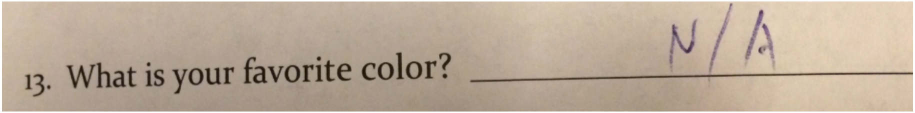


13. What is your favorite color? Orange

orange

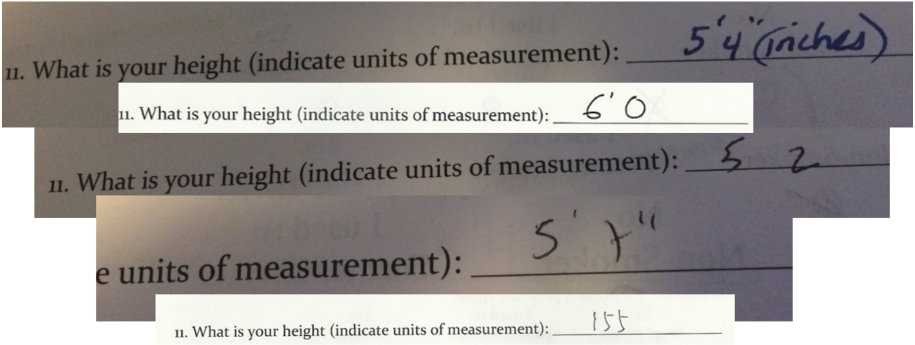


13. What is your favorite color? Blue, Brown



13. What is your favorite color? N/A

Height



11. What is your height (indicate units of measurement): 5'4" (inches)

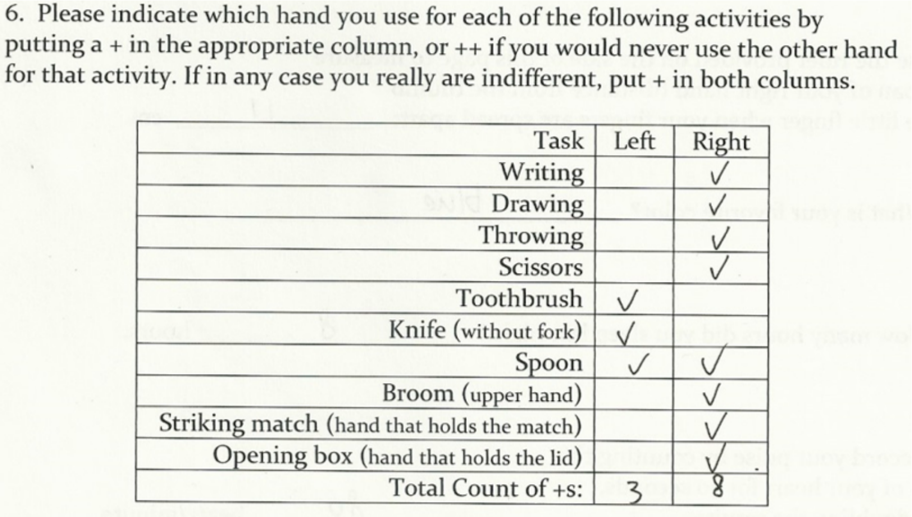
11. What is your height (indicate units of measurement): 6'0

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

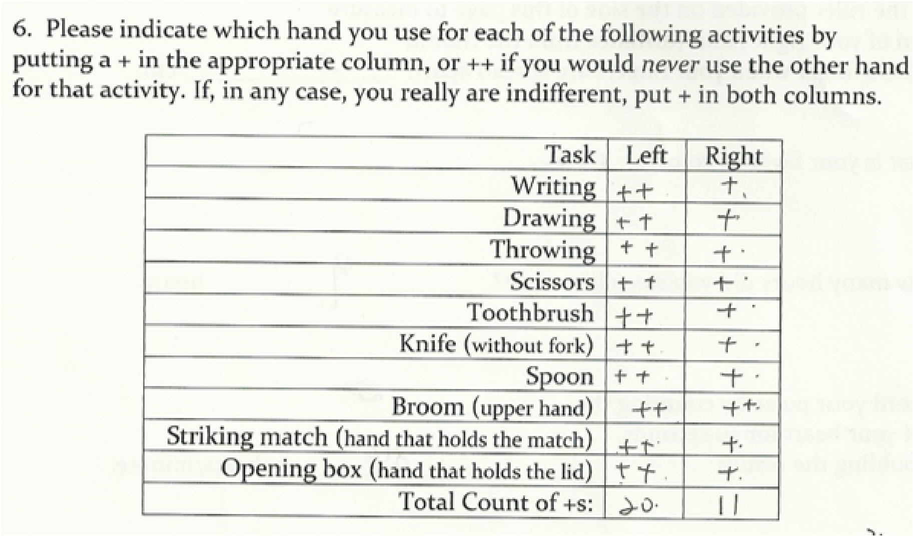
Handedness Scale (2014-15 version)



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

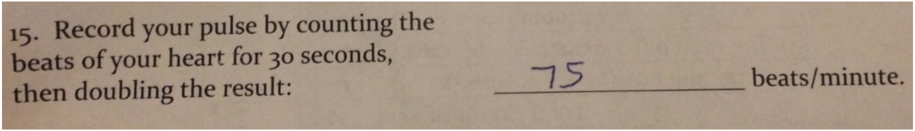
Handedness Scale (2016-19 version)



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

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        Stem-and-Leaf display
 62 63 65  65  66  66          3 | 3
 68 68 68  69  70  70          4 | 68
 70 70 70  70  70  70          5 | 6
 71 72 72  74  74  74          6 | 002355668889
 74 74 75  76  76  76          7 | 00000001224444456668888
 78 78 78  80  80  80          8 | 000012445668
 80 81 82  84  84  85          9 | 000046
 86 86 88  90  90  90         10 | 44
 90 94 96 104 104 110         11 | 0
```

Thanks, John **Tukey**

Garbage in, garbage out . . .

"Data don't make any sense,
we will have to resort to statistics."

Working with R and the “Day 1” survey data

- The `surveyday1_2019.csv` file is available to you as part of the Data download for the course.
- It’s a comma-separated version text file, which is pretty future-proof and can be read easily into R.
- We’ll first load the tidyverse set of R packages, which will let us do a lot of things very cleanly. Learn more about the tidyverse in the Course Notes and in *R for Data Science*.
- Then we’ll read the data into R, so we can look it over more closely.
- This won’t be the last time we do this sort of thing in this class.

Loading the tidyverse of R packages

```r
library(tidyverse)
```

```
-- Attaching packages ---------------------------------------

v ggplot2 3.3.2     v purrr   0.3.4
v tibble  3.0.3     v dplyr   1.0.2
v tidyr   1.1.1     v stringr 1.4.0
v readr   1.3.1     v forcats 0.5.0

-- Conflicts ------------------------------------------ tidy
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

Reading in (ingesting) the data

We'll place the data in a specialized data frame (called a **tibble**) named survey1.

```
survey1 <- read_csv("data/surveyday1_2019.csv")
```

```
Parsed with column specification:
cols(
  .default = col_double(),
  sex = col_character(),
  glasses = col_character(),
  english = col_character(),
  favcolor = col_character()
)

See spec(...) for full column specifications.
```

The survey1 data

```
survey1
```

```
# A tibble: 315 x 21
   student sex   glasses english statsofar ageguess
     <dbl> <chr> <chr>   <chr>       <dbl>    <dbl>
 1  201901 <NA>  y       y               6       42
 2  201902 <NA>  y       y               7       53
 3  201903 <NA>  y       y               4       45
 4  201904 <NA>  y       y               7       45
 5  201905 <NA>  y       y               6       42
 6  201906 <NA>  y       y               7       50
 7  201907 <NA>  y       y               5       56
 8  201908 <NA>  n       n               6       50
 9  201909 <NA>  n       y               6       52
10  201910 <NA>  n       y               4       42
# ... with 305 more rows, and 15 more variables:
#   smoke <dbl>, h.left <dbl>, h.right <dbl>,
```

Most Popular Colors in 2019

```r
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 (code)

Counting and sorting are under-rated parts of exploring data.

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

Most Popular Colors in 2019 (result)

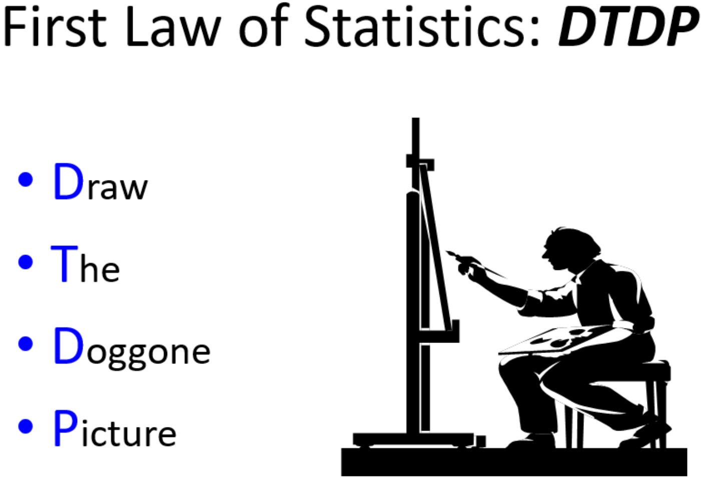
```
# A tibble: 13 x 2
   favcolor         n
   <chr>        <int>
 1 blue            23
 2 purple          10
 3 green            9
 4 red              7
 5 pink             3
 6 teal             2
 7 black            1
 8 dark green       1
 9 gray             1
10 light blue       1
11 light purple     1
12 white            1
13 <NA>             1
```

What about Haircut Prices?

```r
survey1$haircut
```

```
  [1] 120.00  20.00  20.00   0.00   6.99     NA  25.00
  [8]  80.00  16.00  12.50   1.00  25.00  20.00  30.00
 [15] 100.00   3.50  30.00  30.00  20.00  15.00  30.00
 [22]   0.00  50.00  60.00  80.00  20.00  50.00  35.00
 [29]  29.00  80.00  25.00   7.00  35.00  35.00  25.00
 [36]  70.00  16.00   0.00  60.00  35.00  70.00  23.00
 [43]  30.00  15.00  80.00  18.00  60.00  50.00  25.00
 [50]  25.00   8.00  30.00  25.00  20.00  15.00  27.00
 [57]  12.00  80.00  80.00  20.00 120.00  15.00  25.00
 [64]  22.00  20.00   0.00  20.00  40.00  50.00  20.00
 [71]  30.00  50.00 120.00  25.00   0.00  50.00  20.00
 [78]  20.00  20.00   0.00   0.00  43.00  36.00  65.00
 [85] 100.00  15.00  12.00 110.00  20.00   0.00  12.00
 [92]  20.00   9.00  10.00  24.00  30.00  30.00  20.00
 [99]  20.00  18.00  25.00  25.00  20.00  20.00  60.00
```

DTDP: The first rule of working with data



First Law of Statistics: ***DTDP***

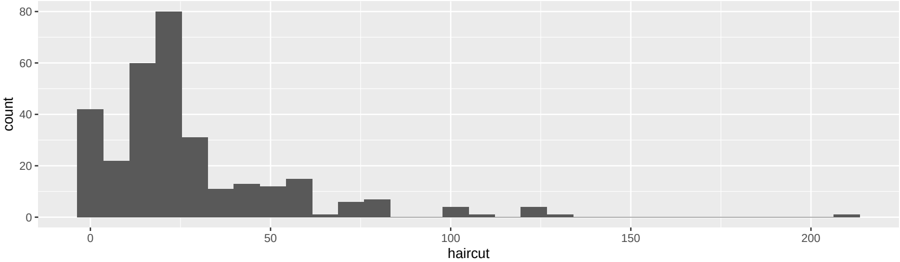
- **D**raw
- **T**he
- **D**oggone
- **P**icture

Histogram of Haircut Prices (First Attempt)

```
ggplot(survey1, aes(haircut)) +
  geom_histogram()
```

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

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



Uh, oh. What happened here?

Numerical Summary of Haircut Prices

```r
survey1 %>% select(haircut) %>% summary
```

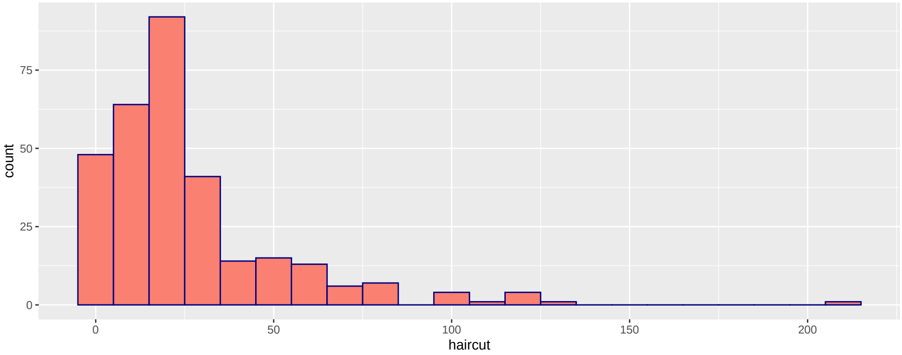
```
    haircut      
 Min.   :  0.00  
 1st Qu.: 14.00  
 Median : 20.00  
 Mean   : 27.32  
 3rd Qu.: 32.00  
 Max.   :210.00  
 NA's   :4       
```

```r
mosaic::favstats(~ haircut, data = survey1)
```

```
 min Q1 median Q3 max    mean       sd   n missing
   0 14     20 32 210 27.3199 26.35565 311       4
```

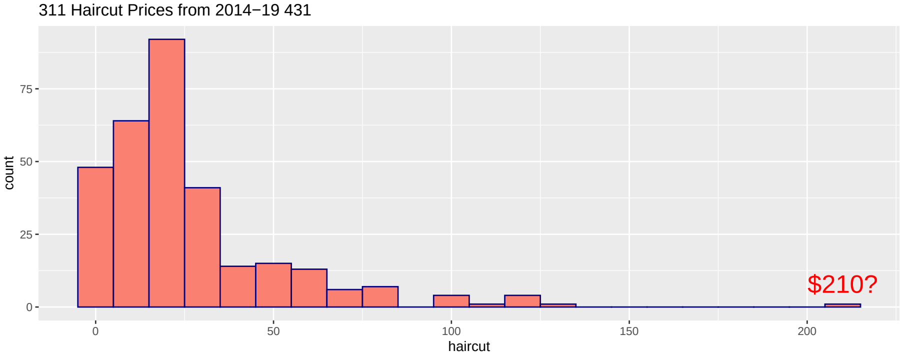
Revising the Histogram

```
survey1 %>%
  filter(complete.cases(haircut)) %>%
  ggplot(., aes(x = haircut)) +
  geom_histogram(binwidth = 10, fill = "salmon", col = "navy")
```



Adding a Title and an Annotation

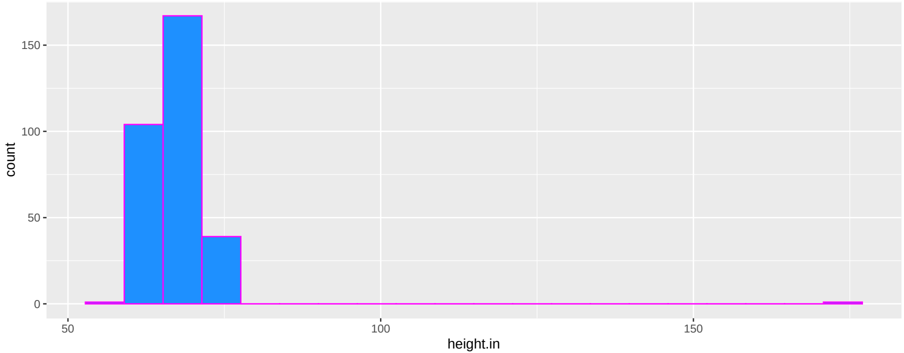
```
survey1 %>% filter(complete.cases(haircut)) %>%
  ggplot(., aes(x = haircut)) +
  geom_histogram(binwidth = 10, fill = "salmon", col = "navy")
  annotate("text", x = 210, y = 8, label = "$210?",
           col="red", size = 7) +
  labs(title = "311 Haircut Prices from 2014-19 431")
```



What about Height?

```
survey1 %>%
  ggplot(., aes(x = height.in)) +
  geom_histogram(bins = 20, fill = "dodgerblue", col = "magent
```

```
Warning: Removed 3 rows containing non-finite values
(stat_bin).
```



Numerical Summaries

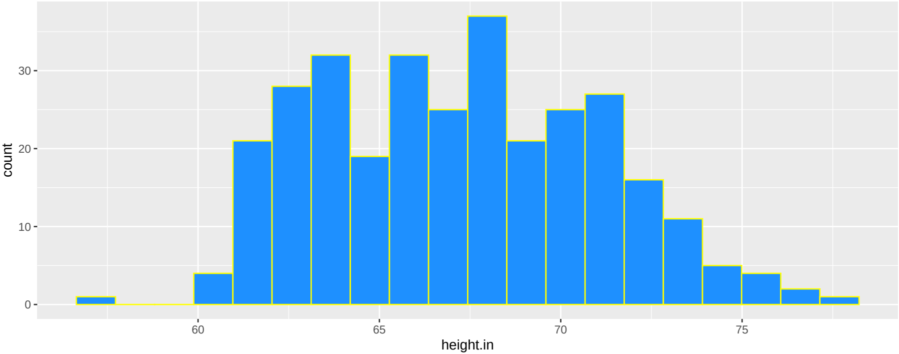
```
mosaic::favstats(height.in ~ year, data = survey1)
```

```
  year min    Q1 median Q3   max     mean        sd  n
1 2014  60 64.75     68 71  73.0 67.78750  3.462042 40
2 2015  61 65.00     68 70  74.0 67.34694  3.321653 49
3 2016  60 64.00     67 70  76.0 67.22656  3.864706 64
4 2017  62 65.00     67 69 175.0 69.60417 15.915321 48
5 2018  60 63.00     66 70  73.0 66.49020  3.807217 51
6 2019  57 65.00     68 70  77.5 67.43333  3.829487 60
  missing
1       2
2       0
3       0
4       0
5       0
6       1
```

What should we do?

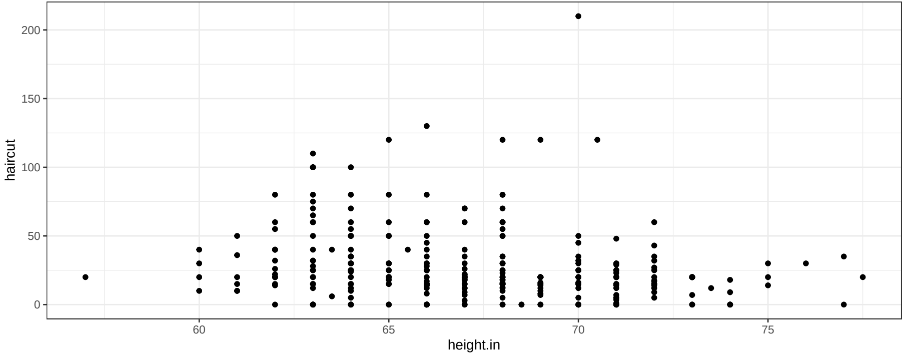
Distribution of Heights, without the outlier

```
survey1 %>%
  filter(height.in < 80) %>%
  ggplot(., aes(x = height.in)) +
  geom_histogram(bins = 20,
                 fill = "dodgerblue", col = "yellow")
```



Association of Height with Haircut Price

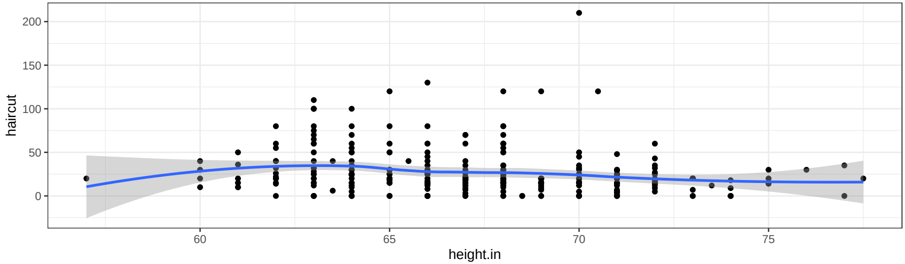
```
survey1 %>%
  filter(complete.cases(height.in, haircut)) %>%
  filter(height.in < 84) %>%
  ggplot(aes(x = height.in, y = haircut)) +
  geom_point() +
  theme_bw()
```



Does the relationship look linear?

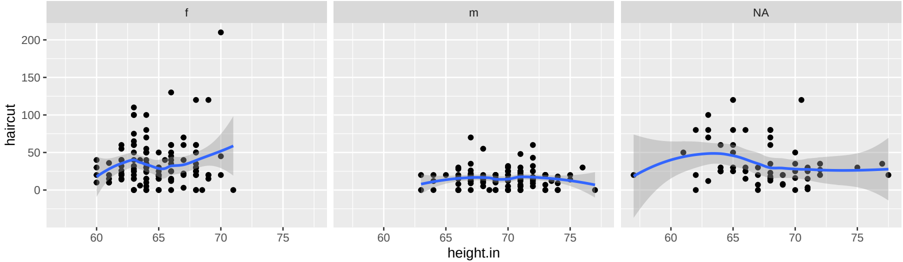
```r
survey1 %>%
  filter(complete.cases(height.in, haircut)) %>%
  filter(height.in < 84) %>%
  ggplot(aes(x = height.in, y = haircut)) +
  geom_point() +
  geom_smooth(method = "loess") +
  theme_bw()
```

```
`geom_smooth()` using formula 'y ~ x'
```



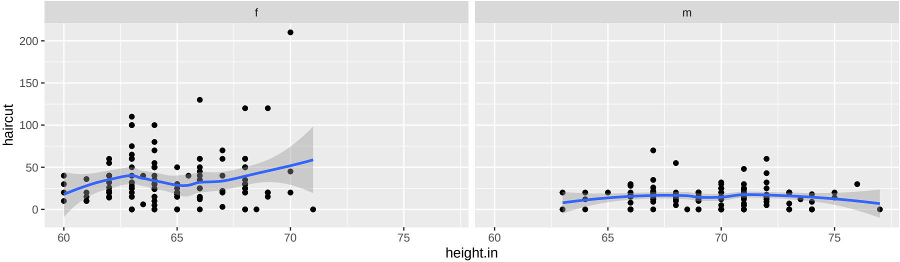
What if we stratify (facet) the plot by sex?

```
survey1 %>%
  filter(complete.cases(height.in, haircut)) %>%
  filter(height.in < 84) %>%
  ggplot(aes(x = height.in, y = haircut)) +
  geom_point() +
  geom_smooth(method = "loess") +
  facet_wrap(~ sex)
```



Eliminate the subjects where we didn't collect sex

```
survey1 %>%
  filter(complete.cases(height.in, haircut, sex)) %>%
  filter(height.in < 84) %>%
  ggplot(aes(x = height.in, y = haircut)) +
  geom_point() +
  geom_smooth(method = "loess") +
  facet_wrap(~ sex)
```

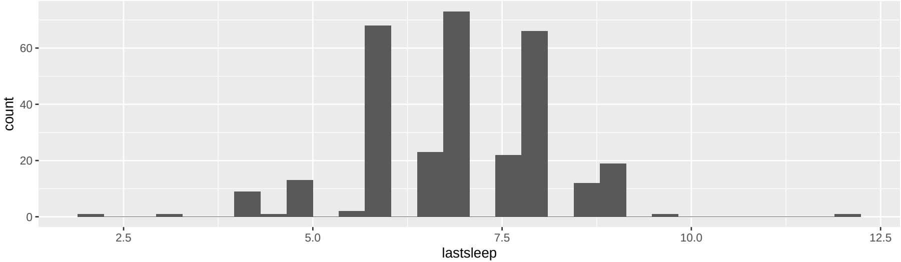


Looking at Hours of Sleep Last Night

```
ggplot(data = survey1, aes(x = lastsleep)) +
  geom_histogram()
```

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

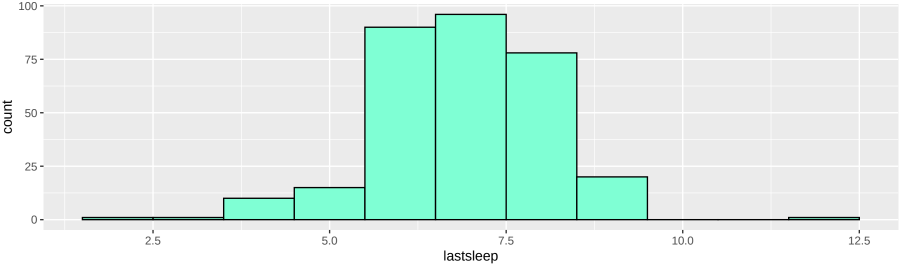
```
Warning: Removed 3 rows containing non-finite values
(stat_bin).
```



What should we do?

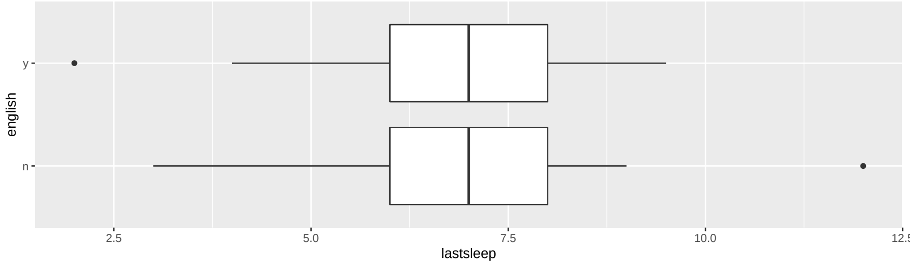
Looking at Hours of Sleep Last Night

```
survey1 %>% filter(complete.cases(lastsleep)) %>%
ggplot(data = ., aes(x = lastsleep)) +
  geom_histogram(binwidth = 1, fill = "aquamarine",
                 col = "black")
```



Hours of Sleep by Prefers English?

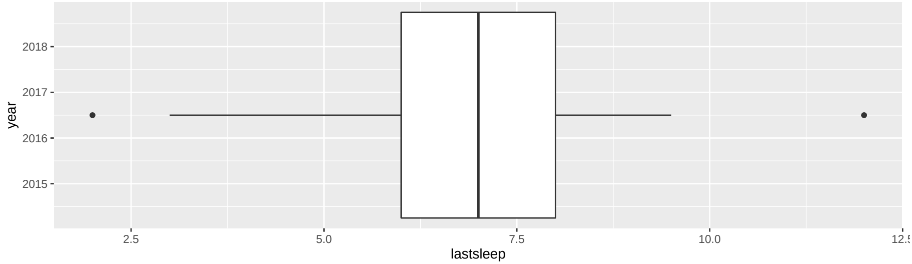
```
survey1 %>% filter(complete.cases(english, lastsleep)) %>%
ggplot(data = ., aes(x = english, y = lastsleep)) +
  geom_boxplot() +
  coord_flip()
```



Hours of Sleep by Survey Year

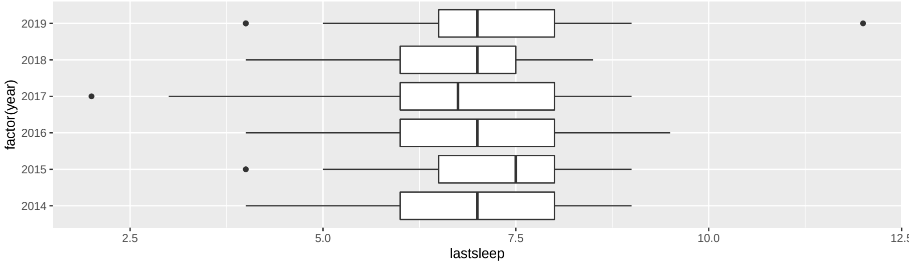
```
survey1 %>% filter(complete.cases(year, lastsleep)) %>%
ggplot(data = ., aes(x = year, y = lastsleep)) +
  geom_boxplot() +
  coord_flip()
```

```
Warning: Continuous x aesthetic -- did you forget
aes(group=...)?
```



Get R to recognize year as categorical here

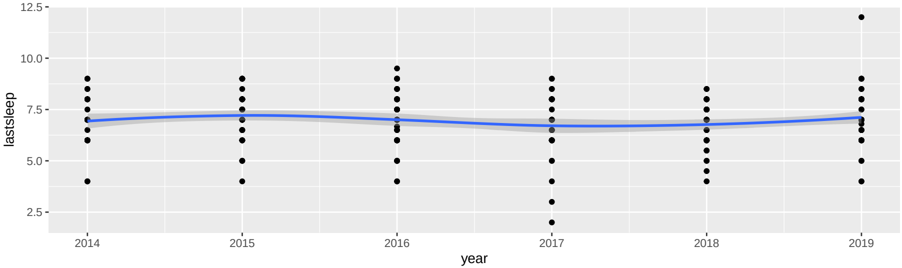
```
survey1 %>% filter(complete.cases(year, lastsleep)) %>%
ggplot(data = ., aes(x = factor(year), y = lastsleep)) +
  geom_boxplot() +
  coord_flip()
```



Or present in a scatterplot

```
survey1 %>% filter(complete.cases(year, lastsleep)) %>%
ggplot(data = ., aes(x = year, y = lastsleep)) +
  geom_point() +
  geom_smooth(method = "loess")
```

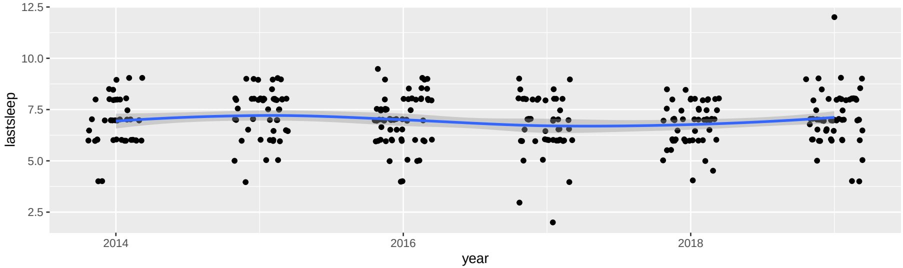
```
`geom_smooth()` using formula 'y ~ x'
```



Maybe jitter the points horizontally?

```
survey1 %>% filter(complete.cases(year, lastsleep)) %>%
ggplot(data = ., aes(x = year, y = lastsleep)) +
  geom_jitter(width = 0.2) +
  geom_smooth(method = "loess")
```

```
`geom_smooth()` using formula 'y ~ x'
```



Chatfield's Six Rules for Data Analysis

1. Do not attempt to analyze the data until you understand what is being measured and why.
2. Find out how the data were collected.
3. Look at the structure of the data.
4. Carefully examine the data in an exploratory way, before attempting a more sophisticated analysis.
5. Use your common sense at all times.
6. Report the results in a clear, self-explanatory way.

Chatfield, Chris (1996) *Problem Solving: A Statistician's Guide*, 2nd ed.

Another example that we won't discuss in class today

Analyzing Guesses of My Age

61 students turned in an index card in 2019, 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.

```r
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
```

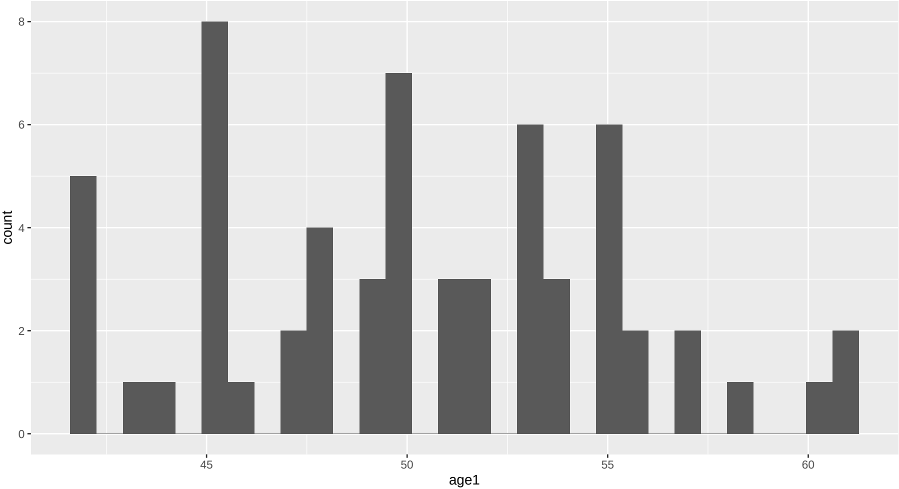
```
# A tibble: 61 x 3
   subject  age1  age2
   <chr>   <dbl> <dbl>
 1 S19-01     47    52
 2 S19-02     55    59
 3 S19-03     55    NA
 4 S19-04     45    45
 5 S19-05     45    48
 6 S19-06     42    49
 7 S19-07     43    55
 8 S19-08     50    46
 9 S19-09     54    50
10 S19-10     61    57
#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

Histogram of initial guesses?

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

Histogram of initial guesses?

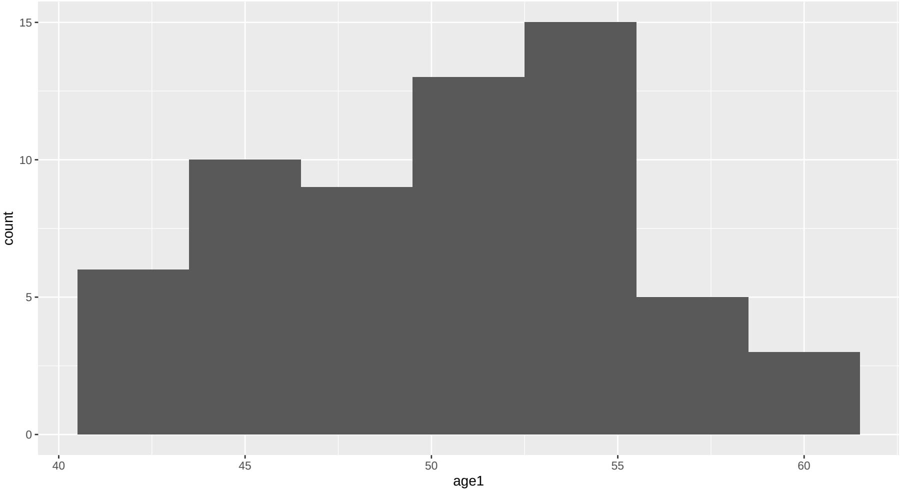
```
`stat_bin()` using `bins = 30`. Pick better value
with `binwidth`.
```



Make the width of the bins 3 years?

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

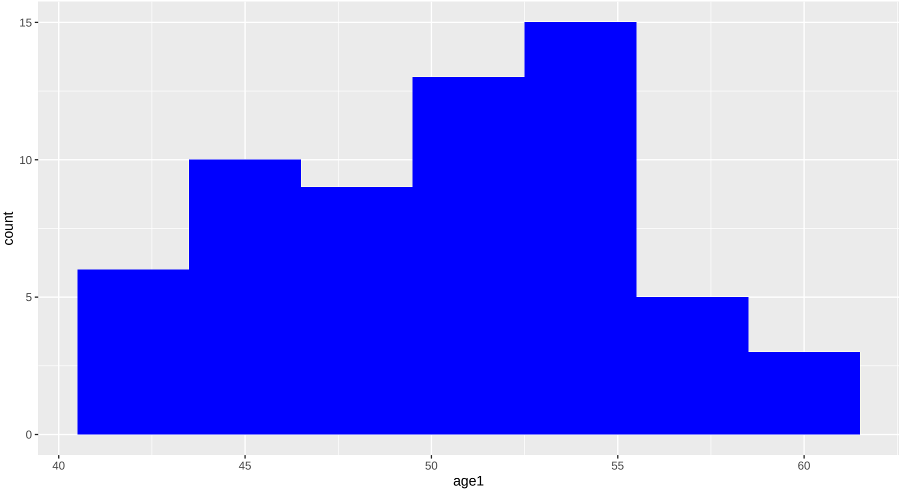
Make the width of the bins 3 years?



Fill in the bars with a better color?

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

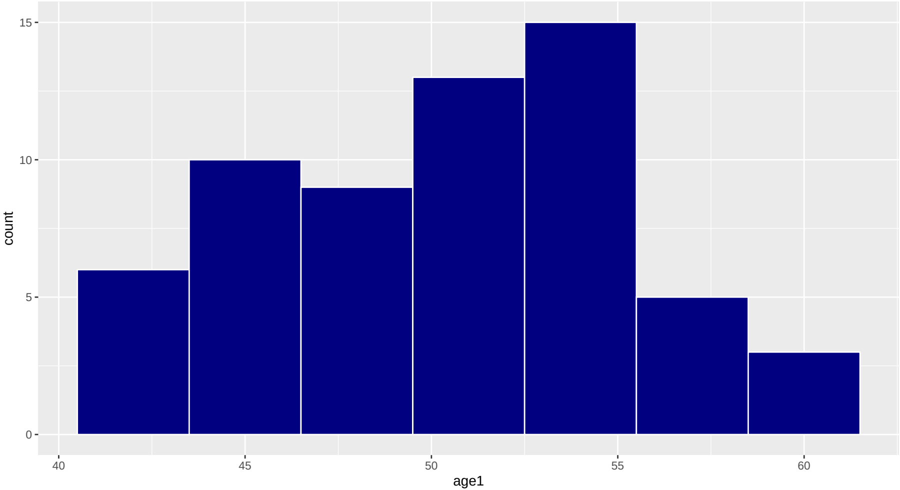
Fill in the bars with a better color?



Make it a little prettier?

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

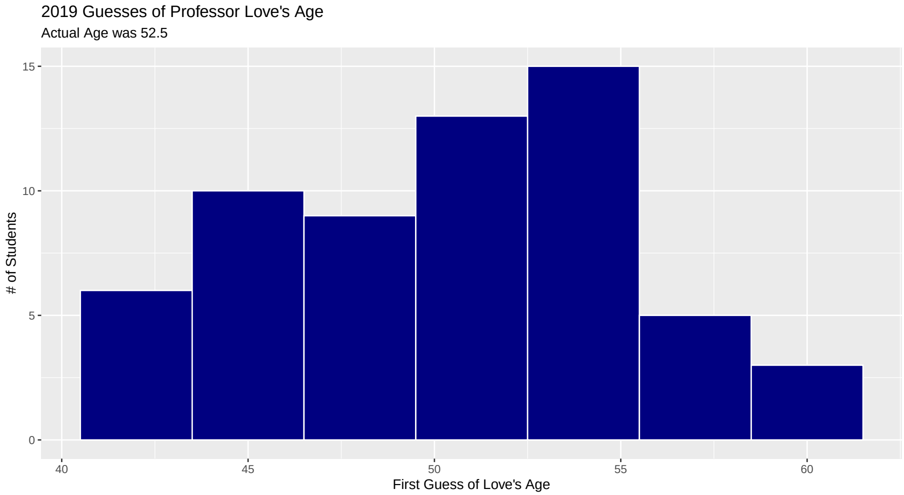
Make it a little prettier?



Add more meaningful labels?

```r
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?



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

```r
mosaic::inspect(love_2019)
```

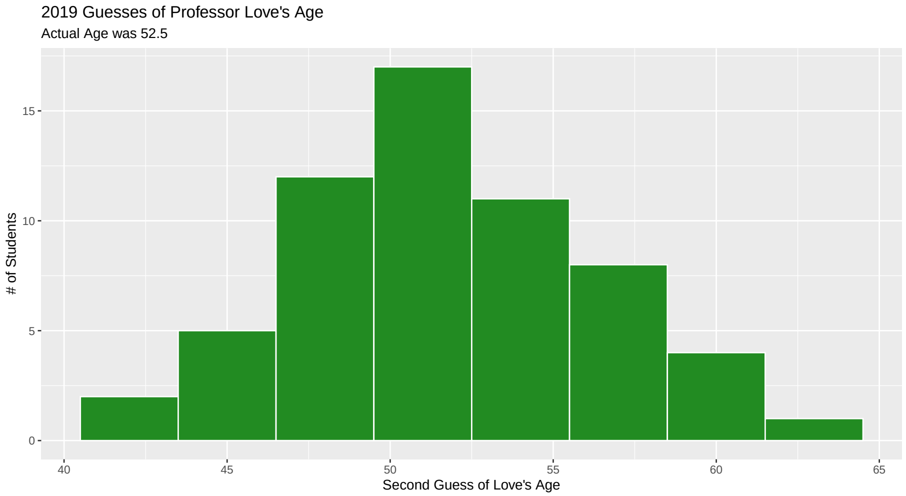
```

categorical variables:  
     name     class levels  n missing
1 subject character     61 61       0
                                   distribution
1 S19-01 (1.6%), S19-02 (1.6%) ...

quantitative variables:  
     name   class min    Q1 median Q3 max     mean
...1 age1 numeric  42 46.00     50 54  61 50.34426
...2 age2 numeric  42 48.75     52 55  62 51.81667
           sd  n missing
...1 4.989607 61       0
...2 4.545408 60       1
```

What about the second guess?

```r
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?

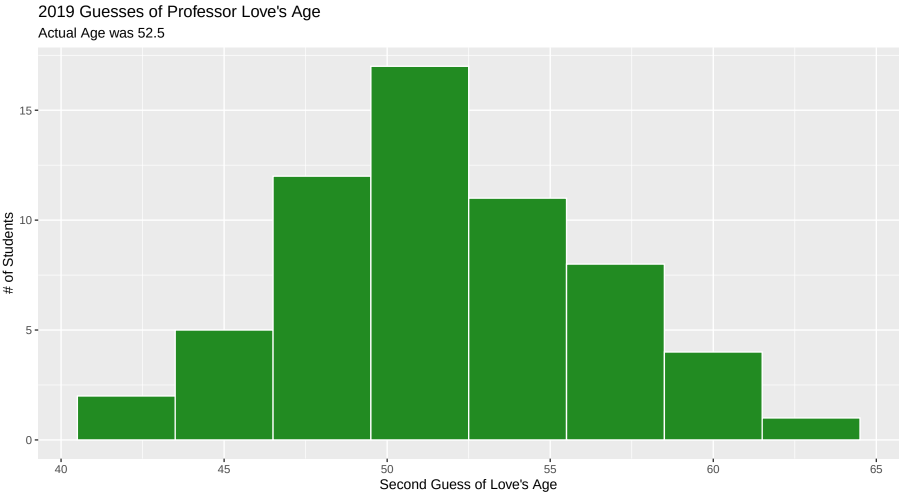
```
Warning: Removed 1 rows containing non-finite values
(stat_bin).
```



Filter to complete cases only

```r
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

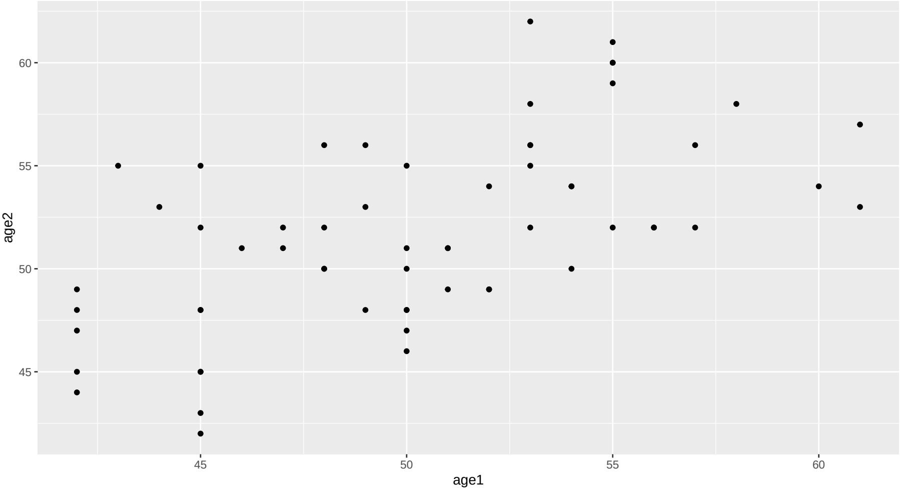


Comparing First Guess to Second Guess

```
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).
```

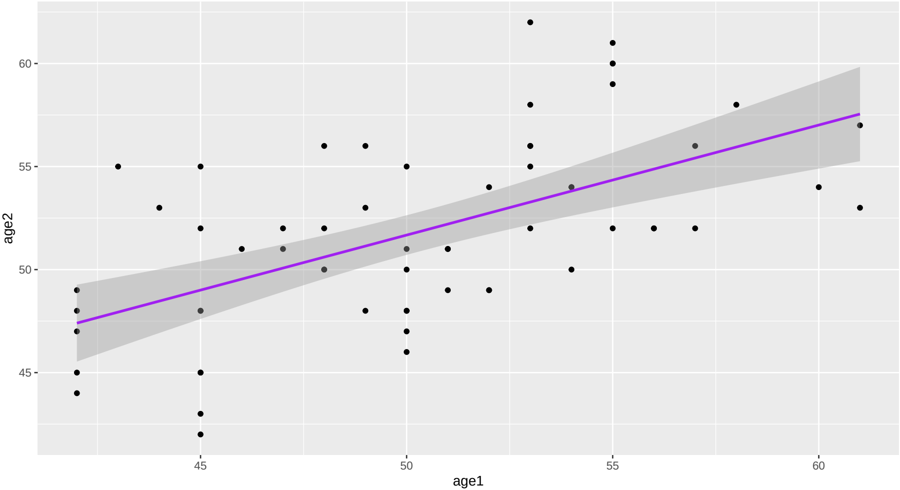


Filter to complete cases, add regression line

```
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

```
`geom_smooth()` using formula 'y ~ x'
```



What's that regression line?

```r
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.

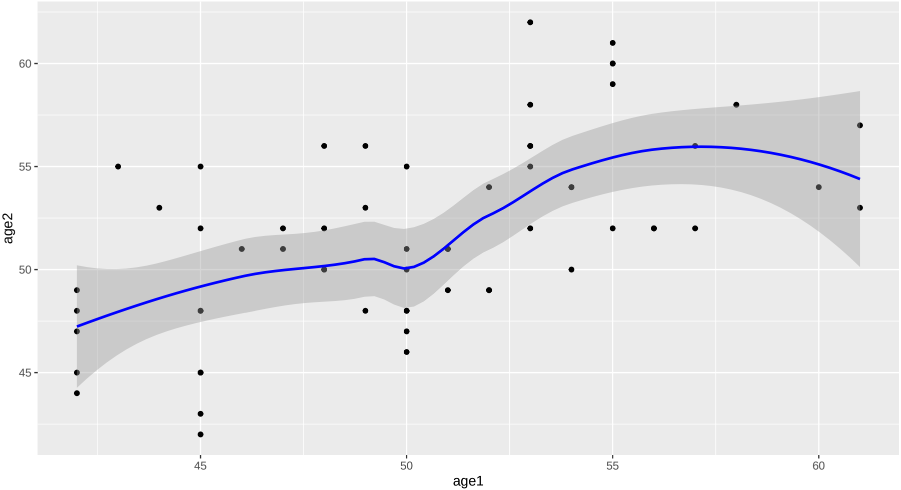
We'll have several alternative approaches to fit regressions coming up.

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?

```
`geom_smooth()` using formula 'y ~ x'
```

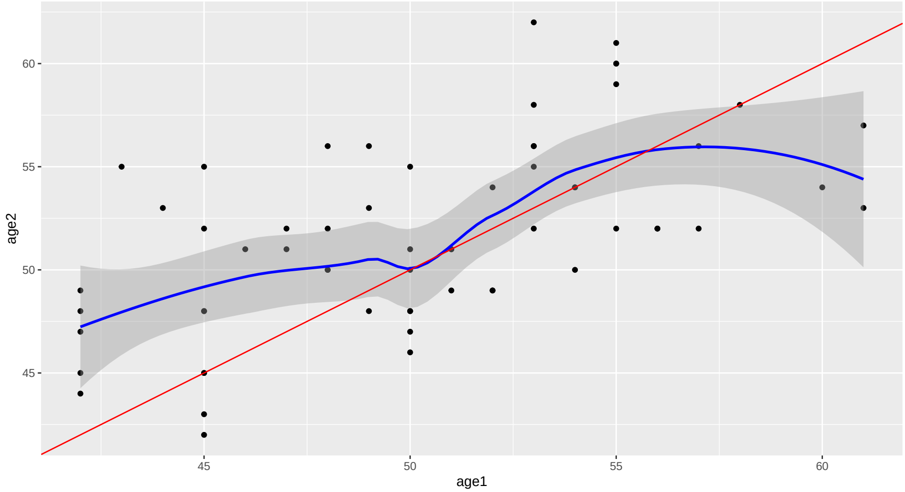


Add a y = x line (no change in guess)?

```
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)?

```
`geom_smooth()` using formula 'y ~ x'
```

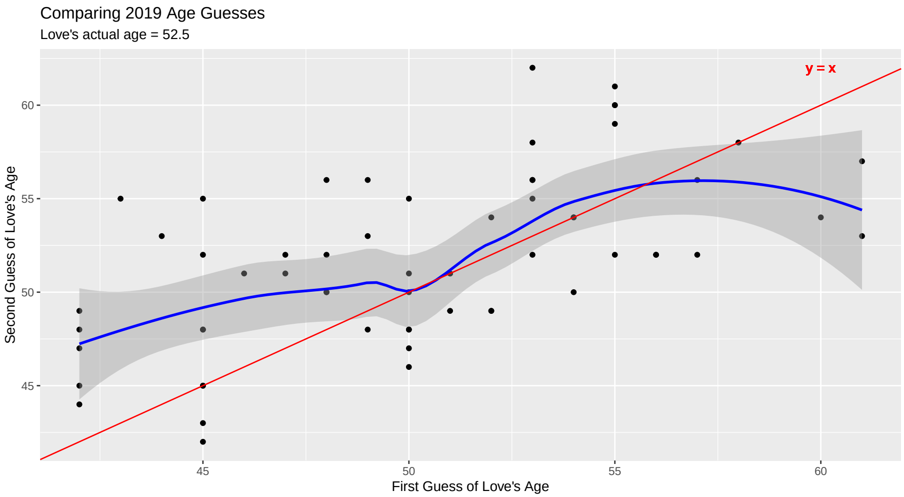


Add more meaningful labels

```r
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

```
`geom_smooth()` using formula 'y ~ x'
```



age1 - age2 difference in guesses?

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

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

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

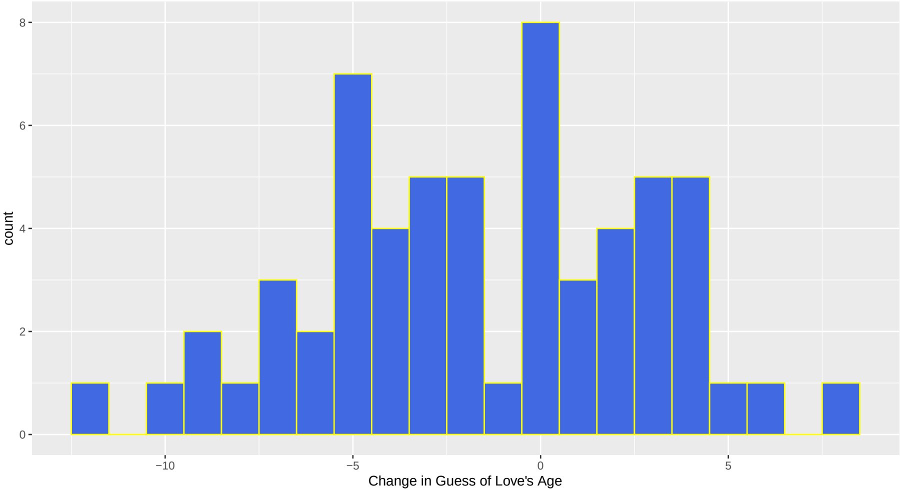
```
love_2019 %>%
  mutate(diff = age1 - age2) %>%
  count(sign(diff))
```

```
# 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

```r
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


count
Change in Guess of Love's Age