431 Class 02

https://thomaselove.github.io/431-2024/

2024-08-29

## Instructions for the Quick Survey

Please read these instructions **before** writing.

1. Introduce yourself to someone that you don’t know.
2. Record the survey answers **for that other person**, while they record your responses.
3. Be sure to complete all 15 questions (both sides.)
4. When you are finished, thank your partner and raise your hand. Someone will come to collect your survey.

Regarding Question 4, Professor Love is the large fellow standing in the front of the room.

## Today’s Agenda

* Data Structures and Variables
  + Evaluating some of the Quick Survey variables
* Looking at some of the data collected in Class 01
  + Group Guessing of Ages from 10 Photographs
  + Guessing Dr. Love’s Age (twice)
* Welcome to 431 Survey Report
* What to work on this weekend

## The R Packages I’ll Load Today

library(janitor)  
library(rstanarm)  
library(easystats)  
library(tidyverse)  
  
source("c02/data/Love-431.R")  
  
knitr::opts\_chunk$set(comment = NA)

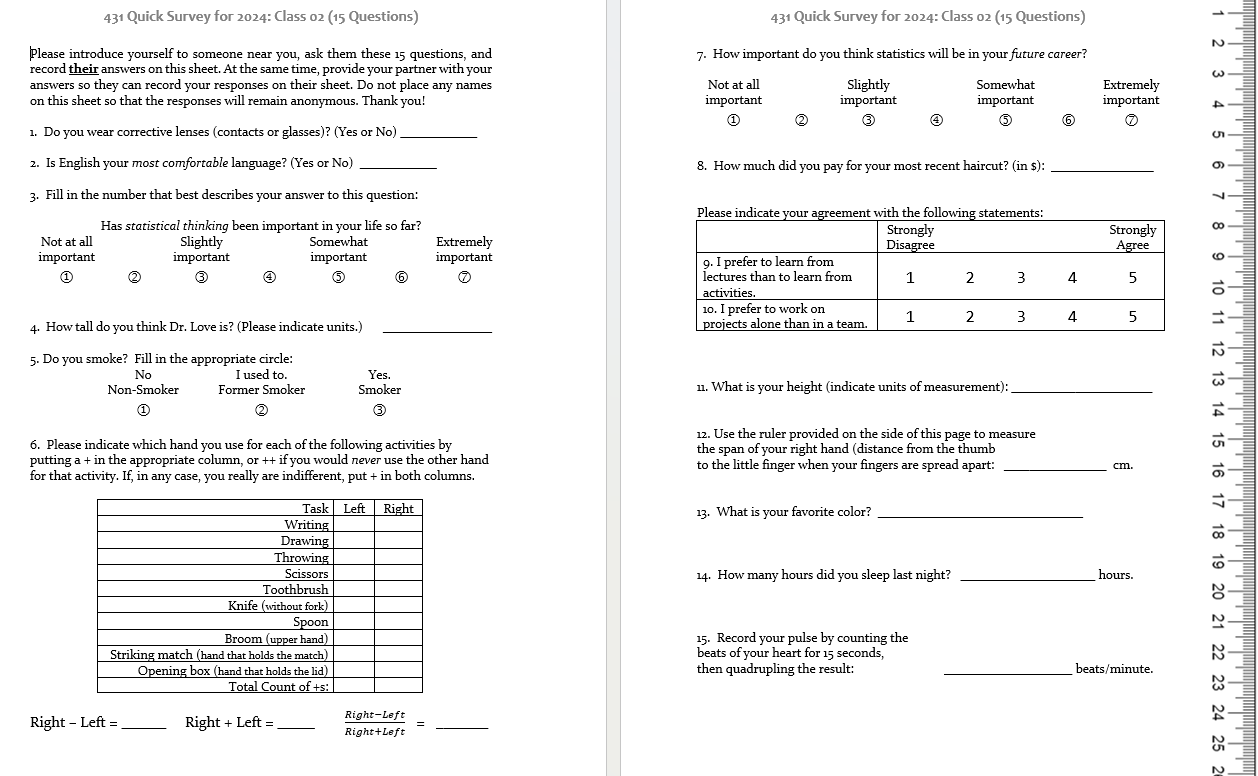
* If you actually run this in R, you will get some messages which we will suppress and ignore today.

## 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.

## Our Quick Survey



## Types of Data

The key distinction we’ll make is between

* **quantitative** (numerical) and
* **categorical** (qualitative) information.

Information that is quantitative describes a **quantity**.

* All quantitative variables have units of measurement.
* Quantitative variables are recorded in numbers, and we use them as numbers (for instance, taking a mean of the variable makes some sense.)

## Continuous vs. Discrete Quantities

**Continuous** variables (can take any value in a range) vs. **Discrete** variables (limited set of potential values)

* Is Height a continuous or a discrete variable?
* Height is certainly continuous as a concept, but how precise is our ruler?
* Piano vs. Violin

## Quantitative Variable Subtypes

We can also distinguish **interval** (equal distance between values, but zero point is arbitrary) from **ratio** variables (meaningful zero point.)

* Is Weight an interval or ratio variable?
* How about IQ?

## Qualitative (Categorical) Data

Qualitative variables consist 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 (three or more categories)
* Can distinguish *nominal* (no underlying order) vs. *ordinal* (categories are ordered.)

## Some Categorical Variables

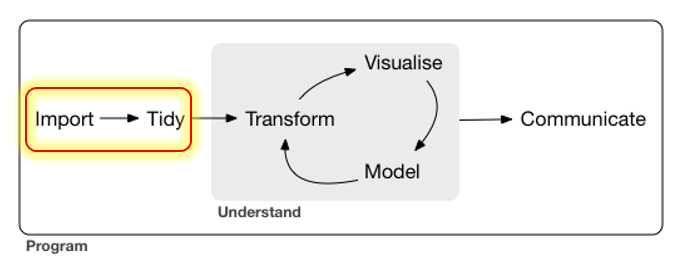
* How is your overall health? (Excellent, Very Good, Good, Fair, Poor)
* Which candidate would you vote for if the election were held today?
* Did this patient receive this procedure?
* If you needed to analyze a small data set right away, which of the following software tools would you be comfortable using to accomplish that task?

## Are these quantitative or categorical?

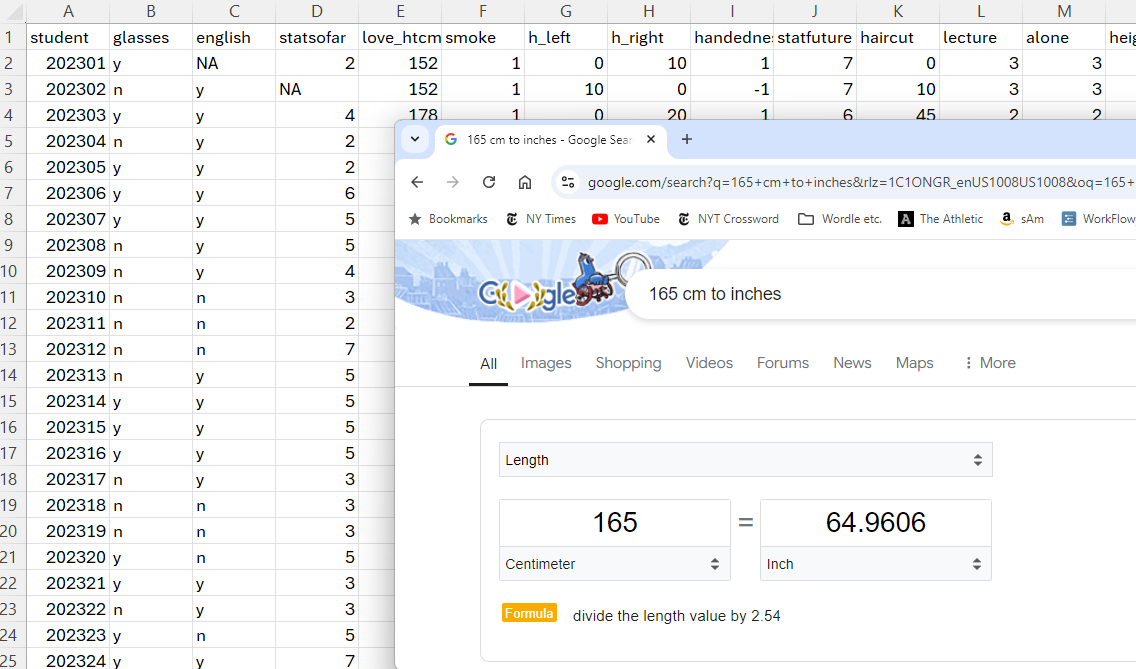
1. Do you **smoke**? (1 = Non-, 2 = Former, 3 = Smoker)
2. How much did you pay for your most recent **haircut**? (in $)
3. What is your favorite **color**?
4. How many hours did you **sleep** last night?
5. Statistical thinking in your future **career**? (1 = Not at all important to 7 = Extremely important)

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

## Importing and Tidying Data



## Ingesting the Quick Surveys



## The Quick Survey

Over 10 years, 547 people took (essentially) the same survey in the same way.

| Fall | 2023 | 2022 | 2021 | 2020 | 2019 |
| --- | --- | --- | --- | --- | --- |
| *n* | 53 | 54 | 58 | 67 | 61 |

| Fall | 2018 | 2017 | 2016 | 2015 | 2014 | Total |
| --- | --- | --- | --- | --- | --- | --- |
| *n* | 51 | 48 | 64 | 49 | 42 | **547** |

### Question

About how many of those 547 surveys caused *no problems* in recording responses?

## The 15 Survey Items

| # | Topic | # | Topic |
| --- | --- | --- | --- |
| Q1 | glasses | Q9 | lectures\_vs\_activities |
| Q2 | english | Q10 | projects\_alone |
| Q3 | stats\_so\_far | Q11 | height |
| Q4 | guess\_TL\_ht | Q12 | hand\_span |
| Q5 | smoke | Q13 | color |
| Q6 | handedness | Q14 | sleep |
| Q7 | stats\_future | Q15 | pulse\_rate |
| Q8 | haircut | - | - |

* At one time, I asked about sex rather than glasses.
* In prior years, people guessed my age, rather than height here.
* Sometimes, I’ve asked for a 30-second pulse check, then doubled.

## Response to the Question I asked

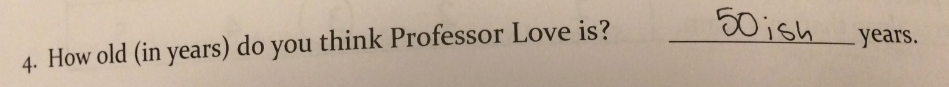
About how many of those 547 surveys caused *no problems* in recording responses?

* Guesses?
* 196/547 (36%) caused no problems.

## Guess My Age

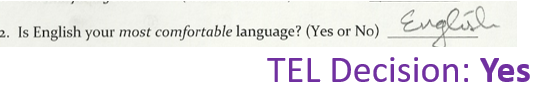


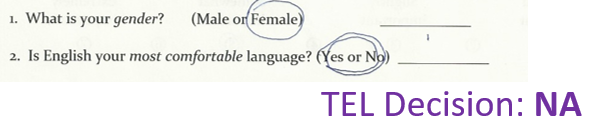


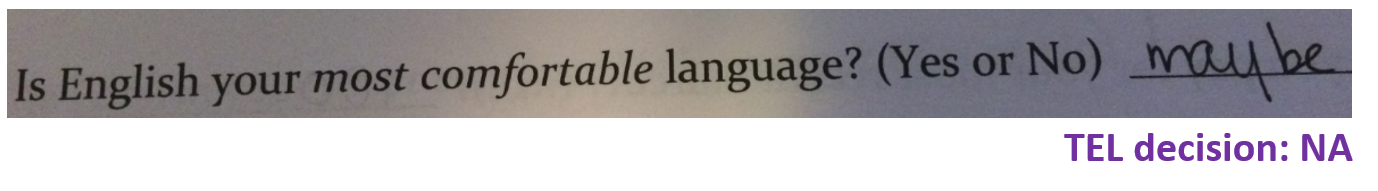


What should we do in these cases?

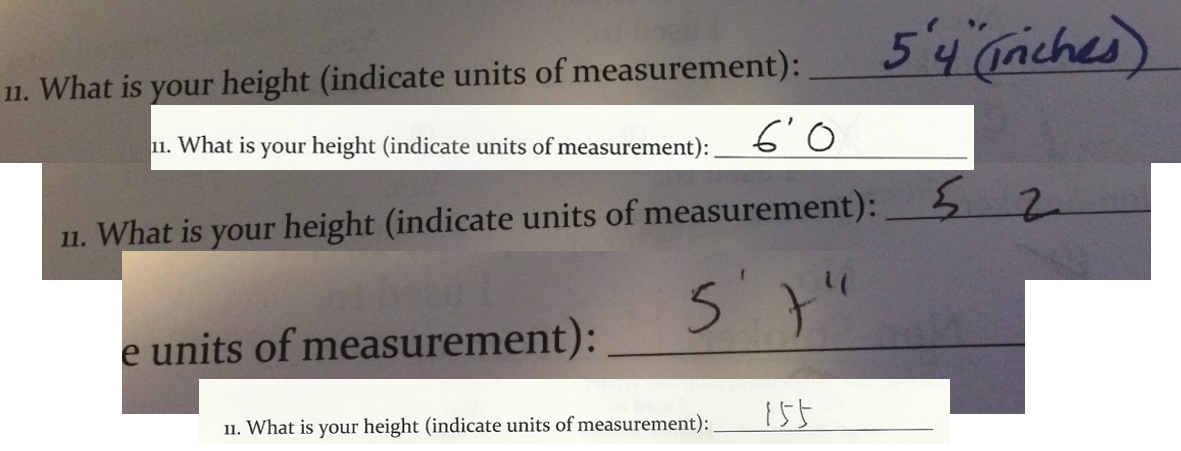
## English best language?







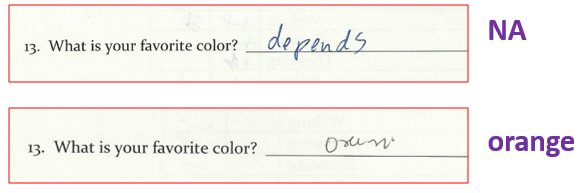
## Height



## Handedness Scale (2016-21 version)



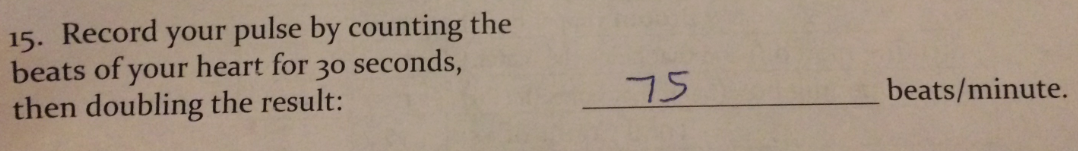
## Favorite color







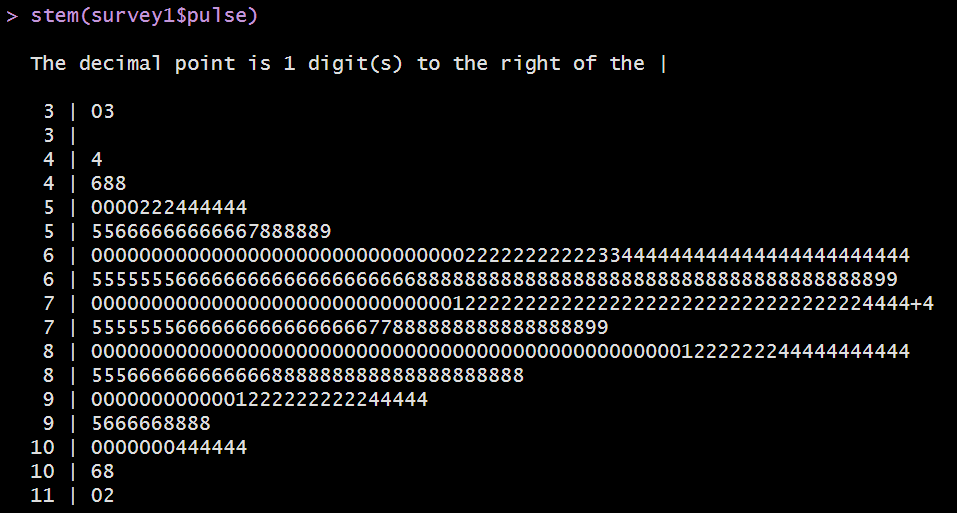
## Following the Rules? (2019 version)



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

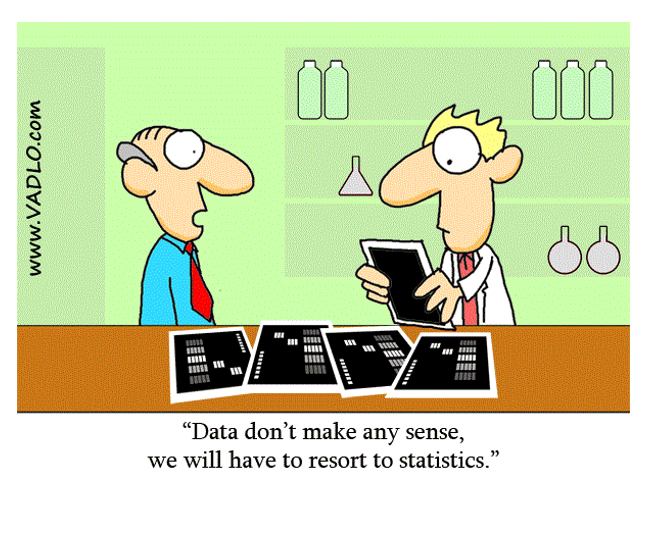
33 46 48 56 60 60 3 | 3  
 62 63 65 65 66 66 4 | 68  
 68 68 68 69 70 70 5 | 6  
 70 70 70 70 70 70 6 | 002355668889   
 71 72 72 74 74 74 7 | 00000000122444445666888  
 74 74 75 76 76 76 8 | 000012445668  
 78 78 78 80 80 80 9 | 000046  
 80 81 82 84 84 85 10 | 44  
 86 86 88 90 90 90 11 | 0  
 90 94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

## Stem and Leaf: Pulse Rates 2014-2023



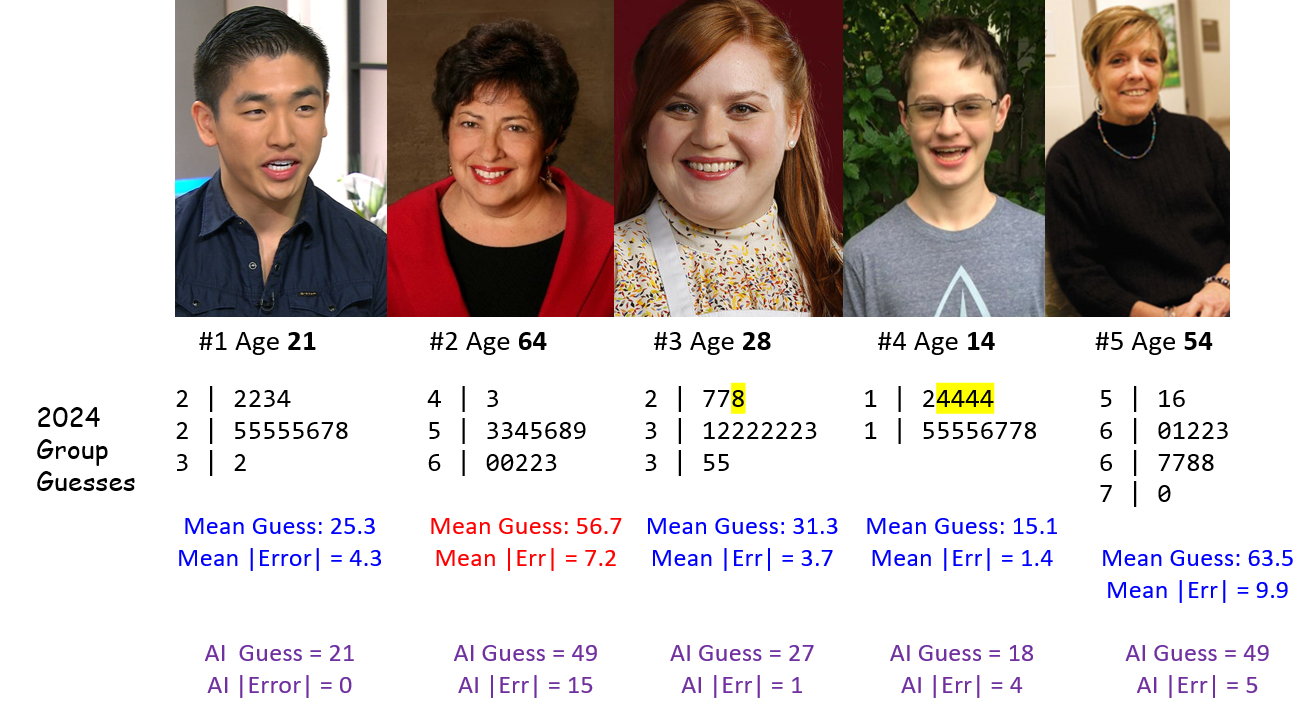
(Thanks, John **Tukey** )

## Garbage in, garbage out …

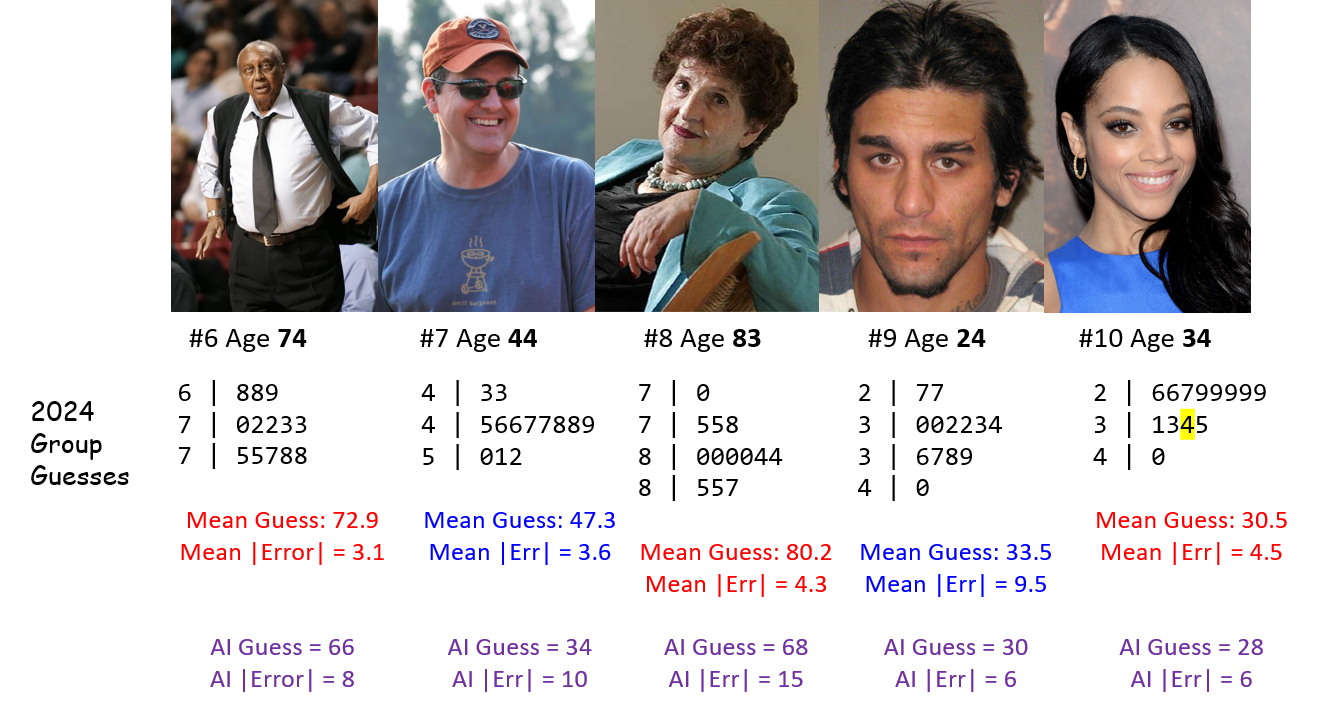


# Group Age Guessing from Photos (13 groups, 10 Photos)

## Photos 1-5



## Photos 6-10



## 2024 Groups 1-6: Guessing Ten Photos

| Group | Within 2 | Within 5 | Too Low | Correct | Too High | Beat AI |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| The Confident Interval | 3 | 5 | 1 | 1 | 8 | 6 |  |
| MAWC | 3 | 7 | 4 | 1 | 5 | 7 |  |
| The Renaissance Coders | 1 | 7 | 2 | 1 | 7 | 6 |  |
| R-rational | 2 | 6 | 5 | 1 | 4 | 7 |  |
| TVMB | 6 | 9 | 1 | 1 | 8 | 7 |  |
| Something Creative & Original | 2 | 5 | 5 | 1 | 4 | 5 |  |
| **AI** | 2 | 4 | 7 | 1 | 2 | – |  |

*These six groups (and the AI at* [*https://howolddoyoulook.com/*](https://howolddoyoulook.com/)*) each guessed one age correctly. The other seven groups are shown on the next slide.*

## 2024 Groups 7-13: Ten Photos

| Group | Within 2 | Within 5 | Too Low | Correct | Too High | Beat AI |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Baked Split | 4 | 8 | 3 | 0 | 7 | 6 |  |
| Pineapple Pizza | 4 | 8 | 5 | 0 | 5 | 7 |  |
| CWRU Crew | 4 | 4 | 5 | 0 | 5 | 6 |  |
| Statasaurous rex | 2 | 6 | 3 | 0 | 7 | 5 |  |
| Tukey 60 | 3 | 8 | 4 | 0 | 6 | 5 |  |
| Beat the Curve | 0 | 5 | 4 | 0 | 6 | 5 |  |
| Stats Avengers | 4 | 6 | 3 | 0 | 7 | 5 |  |

* So … who wins?
* What other summaries might be helpful?

## Error Distribution, Groups 1-6

| Group | Mean Error | SD (Errors) | Median Error | (Min, Max) Error |
| --- | --- | --- | --- | --- |
| The Confident Interval | 5 | 7.3 | 3 | -8, 16 |
| MAWC | 1 | 7 | 1 | -11, 14 |
| The Renaissance Coders | 2.6 | 6.1 | 4 | -9, 13 |
| R-rational | 0.8 | 5.6 | -1 | -6, 10 |
| TVMB | 2 | 2.5 | 1 | -2, 7 |
| Something Creative and Original | -0.2 | 6.2 | -0.5 | -8, 8 |
| **AI** | -5 | 7.3 | -5.5 | -15, 6 |

## Error Distribution, Groups 7/13

| Group | Mean Error | SD (Errors) | Median Error | (Min, Max) Error |
| --- | --- | --- | --- | --- |
| Baked Split | 3.4 | 5.7 | 2 | -3, 14 |
| Pineapple Pizza | 1 | 6.3 | 0.5 | -6, 16 |
| CWRU Crew | -2 | 7.3 | 0 | -13, 9 |
| Statasaurous rex | 1.8 | 5.7 | 4 | -10, 8 |
| Tukey 60 | 2.1 | 6.3 | 1 | -5, 16 |
| Beat the Curve | 2 | 7.5 | 3.5 | -8, 13 |
| Stats Avengers | 1.9 | 9.4 | 3 | -21, 15 |

* How helpful are these summaries in this setting?
* Should we be looking at |error| or maybe squared error?

## Absolute and Squared Errors (first 6)

* **AE** = Absolute Value of Error = |guess - actual|
* **RMSE** = square Root of Mean Squared Error

| Group | Mean AE | Range (AE) | Median AE | RMSE |
| --- | --- | --- | --- | --- |
| The Confident Interval | 6.6 | 0, 16 | 5.5 | 8.5 |
| MAWC | 5.2 | 0, 14 | 3.5 | 6.7 |
| The Renaissance Coders | 5.4 | 0, 13 | 4.5 | 6.3 |
| R-rational | 4.6 | 0, 10 | 4.5 | 5.3 |
| TVMB | 2.4 | 0, 7 | 1.5 | 3.1 |
| Something Creative and Original | 5.2 | 0, 8 | 5.5 | 5.9 |
| **AI** | 7 | 0, 15 | 6 | 8.5 |

## Absolute and Squared Errors (7-13)

* **AE** = Absolute Value of Error = |guess - actual|
* **RMSE** = square Root of Mean Squared Error

| Group | Mean AE | Range (AE) | Median AE | RMSE |
| --- | --- | --- | --- | --- |
| Baked Split | 4.4 | 1, 14 | 3 | 6.4 |
| Pineapple Pizza | 4.4 | 1, 16 | 3 | 6 |
| CWRU Crew | 5.8 | 1, 13 | 6 | 7.2 |
| Statasaurous rex | 5.2 | 2, 10 | 5 | 5.7 |
| Tukey 60 | 4.7 | 1, 16 | 4 | 6.4 |
| Beat the Curve | 6.6 | 3, 13 | 5.5 | 7.4 |
| Stats Avengers | 6.5 | 1, 21 | 4 | 9.1 |
| **AI** | 7 | 0, 15 | 6 | 8.5 |

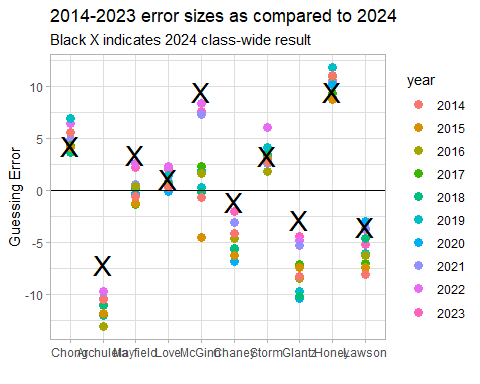
* So … now who wins?

## Importing guesses from 2014-2024

photos <-   
 read\_csv("c02/data/ten-photo-age-history-2024.csv",  
 show\_col\_types = F)  
  
photos <- photos |>  
 mutate(label = fct\_reorder(label, card))  
  
head(photos)

# A tibble: 6 × 13  
 order card label age sex facing year mean\_guess error abs\_error sq\_error  
 <dbl> <dbl> <fct> <dbl> <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl>  
1 1 1 Chong 21 M R 2024 25.3 4.3 4.3 18.5   
2 2 2 Arch… 64 F L 2024 56.8 -7.2 7.2 51.8   
3 3 3 Mayf… 28 F L 2024 31.4 3.4 3.4 11.6   
4 4 4 Love 14 M L 2024 15.1 1.1 1.1 1.21  
5 5 5 McGi… 54 F R 2024 63.5 9.5 9.5 90.2   
6 6 6 Chan… 74 M L 2024 72.9 -1.1 1.1 1.21  
# ℹ 2 more variables: `detailed description` <chr>, jpeg <chr>

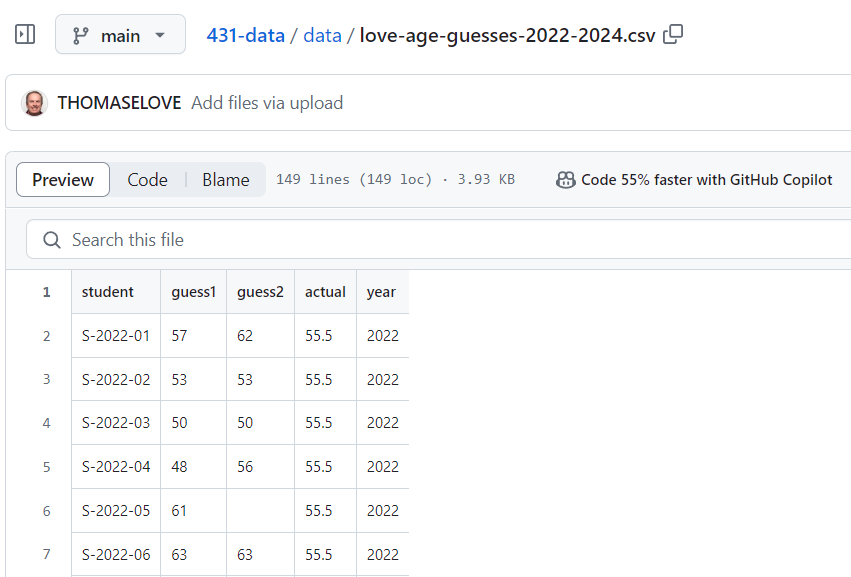
## 2014-2024 Errors



# Guessing My Age (Twice) from Class 01

## From our 431-Data Page: A .csv file

I’ve placed love-age-guesses-2022-2024.csv on our 431-data page. This includes guesses from 2022-2024.



## Creating the age\_guess Tibble

Clicking on RAW in the 431-data presentation takes us to a (long) URL that contains the raw data in this sheet.

I’ll read in the sheet’s data to a new tibble (a special kind of R data frame) called age\_guess using the read\_csv() function.

url\_age <-   
 "https://raw.githubusercontent.com/THOMASELOVE/431-data/main/data/love-age-guesses-2022-2024.csv"  
  
age\_guess <- read\_csv(url\_age, show\_col\_types = FALSE)

## The age\_guess tibble

What do we get?

age\_guess

# A tibble: 148 × 5  
 student guess1 guess2 actual year  
 <chr> <dbl> <dbl> <dbl> <dbl>  
 1 S-2022-01 57 62 55.5 2022  
 2 S-2022-02 53 53 55.5 2022  
 3 S-2022-03 50 50 55.5 2022  
 4 S-2022-04 48 56 55.5 2022  
 5 S-2022-05 61 NA 55.5 2022  
 6 S-2022-06 63 63 55.5 2022  
 7 S-2022-07 67 58 55.5 2022  
 8 S-2022-08 50 57 55.5 2022  
 9 S-2022-09 50 50 55.5 2022  
10 S-2022-10 43 56 55.5 2022  
# ℹ 138 more rows

## How many guesses in each year?

age\_guess |> count(year)

# A tibble: 3 × 2  
 year n  
 <dbl> <int>  
1 2022 53  
2 2023 39  
3 2024 56

How many first guesses in each year were less than 57.5?

age\_guess |> count(year, guess1 < 57.5)

# A tibble: 6 × 3  
 year `guess1 < 57.5` n  
 <dbl> <lgl> <int>  
1 2022 FALSE 14  
2 2022 TRUE 39  
3 2023 FALSE 11  
4 2023 TRUE 28  
5 2024 FALSE 26  
6 2024 TRUE 30

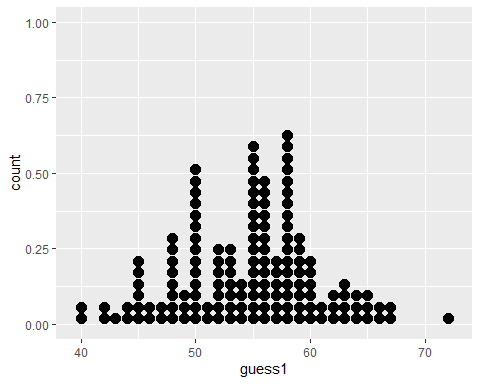
## What do the guess1 values look like?

age\_guess |>   
 select(guess1) |>   
 arrange(guess1)

# A tibble: 148 × 1  
 guess1  
 <dbl>  
 1 40  
 2 40  
 3 42  
 4 42  
 5 43  
 6 44  
 7 44  
 8 45  
 9 45  
10 45  
# ℹ 138 more rows

## Plot the guess1 values?

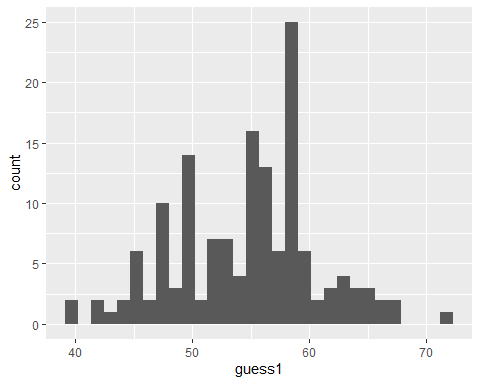
ggplot(data = age\_guess,   
 aes(x = guess1)) +  
 geom\_dotplot(binwidth = 1)



## Can we make a histogram?

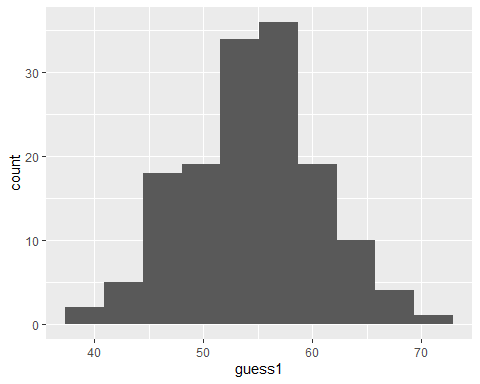
ggplot(age\_guess,   
 aes(x = guess1)) +  
 geom\_histogram()

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



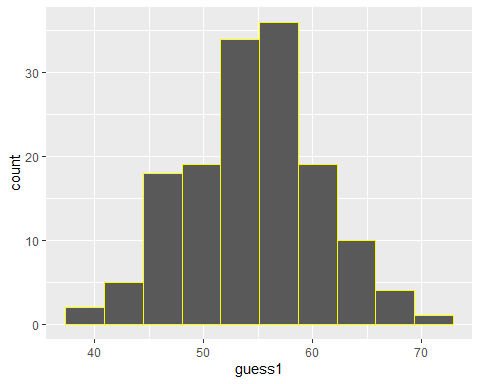
## Improving the Histogram, 1

ggplot(age\_guess,   
 aes(x = guess1)) +  
 geom\_histogram(bins = 10)



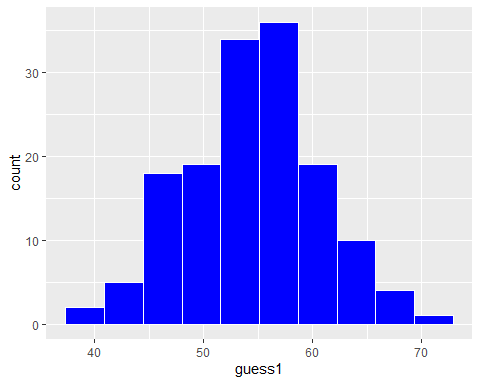
## Improving the Histogram, 2

ggplot(age\_guess,   
 aes(x = guess1)) +  
 geom\_histogram(bins = 10,   
 col = "yellow")



## Improving the Histogram, 3

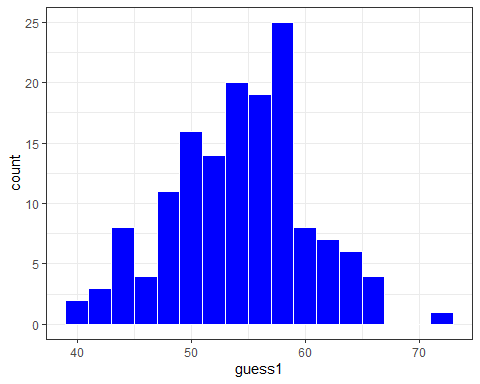
ggplot(age\_guess,   
 aes(x = guess1)) +  
 geom\_histogram(bins = 10,   
 col = "white",   
 fill = "blue")



## Improving the Histogram, 4

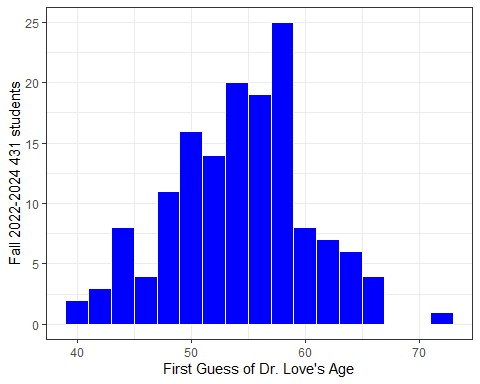
Change theme, specify bin width rather than number of bins

ggplot(age\_guess,   
 aes(x = guess1)) +  
 geom\_histogram(binwidth = 2,   
 col = "white", fill = "blue") +  
 theme\_bw()



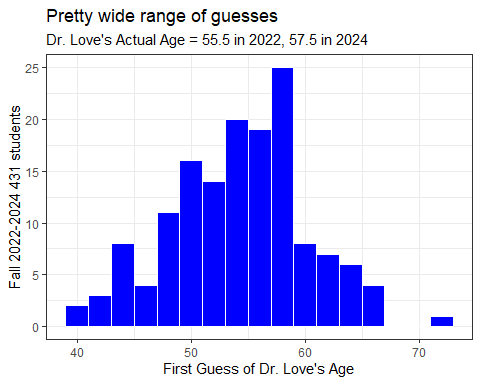
## Improving the Histogram, 5

ggplot(age\_guess,   
 aes(x = guess1)) +  
 geom\_histogram(binwidth = 2,   
 col = "white", fill = "blue") +  
 theme\_bw() +  
 labs(  
 x = "First Guess of Dr. Love's Age",  
 y = "Fall 2022-2024 431 students")



## Add title and subtitle (ver. 6)

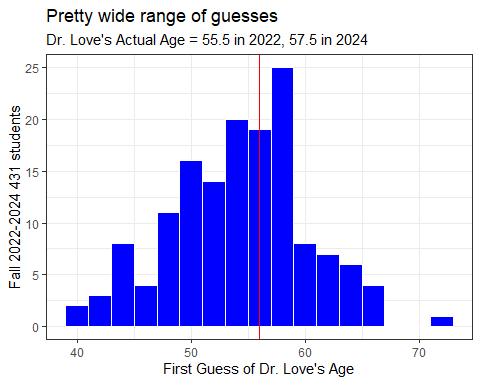
ggplot(age\_guess,   
 aes(x = guess1)) +  
 geom\_histogram(binwidth = 2,   
 col = "white", fill = "blue") +  
 theme\_bw() +  
 labs(  
 x = "First Guess of Dr. Love's Age",  
 y = "Fall 2022-2024 431 students",  
 title = "Pretty wide range of guesses",  
 subtitle = "Dr. Love's Actual Age = 55.5 in 2022, 57.5 in 2024")



## Improving the Histogram, 7

Add a vertical line at 57.5 years to show my actual age.

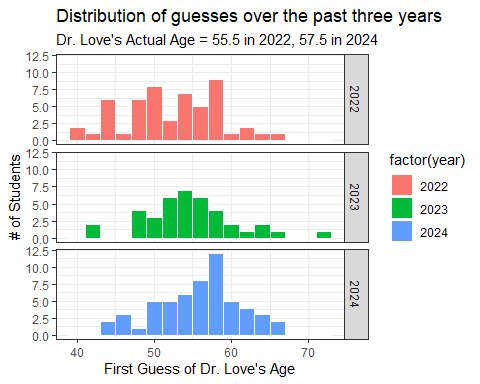
ggplot(age\_guess,   
 aes(x = guess1)) +  
 geom\_histogram(binwidth = 2,   
 col = "white", fill = "blue") +  
 geom\_vline(aes(xintercept = 56), col = "red") +  
 theme\_bw() +  
 labs(  
 x = "First Guess of Dr. Love's Age",  
 y = "Fall 2022-2024 431 students",  
 title = "Pretty wide range of guesses",  
 subtitle = "Dr. Love's Actual Age = 55.5 in 2022, 57.5 in 2024")



## In which year did I look older?

Create three *facets*, for 2022, 2023 and 2024 guesses…

ggplot(age\_guess,   
 aes(x = guess1, fill = factor(year))) +  
 geom\_histogram(binwidth = 2, col = "white") +  
 theme\_bw() +  
 facet\_grid(year ~ .) +  
 labs(  
 x = "First Guess of Dr. Love's Age",  
 y = "# of Students",  
 title = "Distribution of guesses over the past three years",  
 subtitle = "Dr. Love's Actual Age = 55.5 in 2022, 57.5 in 2024")



## Numerical Summary

age\_guess |> select(student, guess1, guess2, year) |> summary()

student guess1 guess2 year   
 Length:148 Min. :40.00 Min. :40.00 Min. :2022   
 Class :character 1st Qu.:50.00 1st Qu.:53.00 1st Qu.:2022   
 Mode :character Median :55.00 Median :56.00 Median :2023   
 Mean :54.65 Mean :56.08 Mean :2023   
 3rd Qu.:58.00 3rd Qu.:59.00 3rd Qu.:2024   
 Max. :72.00 Max. :70.00 Max. :2024   
 NA's :3

* Was the average guess closer on guess 1 or 2?
* What was the range of first guesses? Second guesses?
* What does the NA's : 3 mean in guess2?
* Why is student not summarized any further?

## Let’s Focus on 2024 guesses

age\_24 <- age\_guess |>  
 filter(year == "2024")  
  
age\_24

# A tibble: 56 × 5  
 student guess1 guess2 actual year  
 <chr> <dbl> <dbl> <dbl> <dbl>  
 1 S-2024-01 45 53 57.5 2024  
 2 S-2024-02 50 55 57.5 2024  
 3 S-2024-03 65 62 57.5 2024  
 4 S-2024-04 55 60 57.5 2024  
 5 S-2024-05 65 67 57.5 2024  
 6 S-2024-06 58 56 57.5 2024  
 7 S-2024-07 58 56 57.5 2024  
 8 S-2024-08 60 55 57.5 2024  
 9 S-2024-09 56 53 57.5 2024  
10 S-2024-10 62 58 57.5 2024  
# ℹ 46 more rows

## First Guesses in 2024

age\_24 |> select(guess1) |> table()

guess1  
44 45 46 47 48 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67   
 1 1 1 2 1 4 1 2 3 1 5 5 3 6 6 4 1 3 1 1 2 1 1

### Simple Stem-and-Leaf

4 | 4  
4 | 56778  
5 | 00001223334  
5 | 5555566666777888888999999  
6 | 0000122234  
6 | 5567

age\_24 |> select(guess1) |> summary()

guess1   
 Min. :44.00   
 1st Qu.:53.00   
 Median :57.00   
 Mean :56.25   
 3rd Qu.:59.25   
 Max. :67.00

## Summarizing 2024 Guesses

describe\_distribution(age\_24 |> select(guess1, guess2), ci = 0.90)

Variable | Mean | SD | IQR | 90% CI | Range | Skewness | Kurtosis | n | n\_Missing  
-------------------------------------------------------------------------------------------------------  
guess1 | 56.25 | 5.39 | 6.75 | [54.84, 57.43] | [44.00, 67.00] | -0.31 | -0.25 | 56 | 0  
guess2 | 57.61 | 4.55 | 5.75 | [56.70, 58.68] | [47.00, 67.00] | -0.07 | -0.06 | 56 | 0

* Mean = sum of values divided by number of values
* Standard Deviation = square root of variance, measure of variation
* IQR = difference between 75th and 25th percentiles
* 90% confidence interval for mean estimated via bootstrap
* Range = minimum and maximum values
* n = sample size
* n\_Missing = # of missing values

## Summarizing 2024 Guesses

describe\_distribution(age\_24 |> select(guess1, guess2),   
 centrality = "median", ci = 0.90,   
 range = FALSE, quartiles = TRUE)

Variable | Median | MAD | IQR | 90% CI | Quartiles | Skewness | Kurtosis | n | n\_Missing  
------------------------------------------------------------------------------------------------------  
guess1 | 57.00 | 4.45 | 6.75 | [55.50, 58.00] | 53.00, 59.25 | -0.31 | -0.25 | 56 | 0  
guess2 | 57.50 | 3.71 | 5.75 | [57.00, 59.00] | 55.00, 60.25 | -0.07 | -0.06 | 56 | 0

* Median = 50th percentile (middle value when data are sorted)
* 90% CI here is a bootstrap 90% confidence interval for the median
* MAD = median absolute deviation (scaled to take the same value as the standard deviation when the data are Normally distributed)
* Quartiles = 25th and 75th percentiles

## Summarizing 2024 Guesses

* Using the lovedist() function from the Love-431.R script

age\_24 |>  
 reframe(lovedist(guess1))

# A tibble: 1 × 10  
 n miss mean sd med mad min q25 q75 max  
 <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 56 0 56.2 5.39 57 4.45 44 53 59.2 67

age\_24 |>  
 reframe(lovedist(guess2))

# A tibble: 1 × 10  
 n miss mean sd med mad min q25 q75 max  
 <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 56 0 57.6 4.55 57.5 3.71 47 55 60.2 67

## How did guesses change in 2024?

* Did your guesses decrease / stay the same / increase?
* Calculate guess2 - guess1 and examine its sign.

age\_guess |>   
 filter(year == "2024") |>  
 count(sign(guess2 - guess1))

# A tibble: 3 × 2  
 `sign(guess2 - guess1)` n  
 <dbl> <int>  
1 -1 14  
2 0 14  
3 1 28

## How much did guesses change in 2024?

Create new variable (change = guess2 - guess1)

age\_guess <- age\_guess |>  
 mutate(change = guess2 - guess1)  
  
age\_guess |> filter(year == "2024") |> select(change) |> summary()

change   
 Min. :-8.000   
 1st Qu.:-0.250   
 Median : 0.500   
 Mean : 1.357   
 3rd Qu.: 4.000   
 Max. :17.000

## Histogram of Guess Changes

What will this look like?

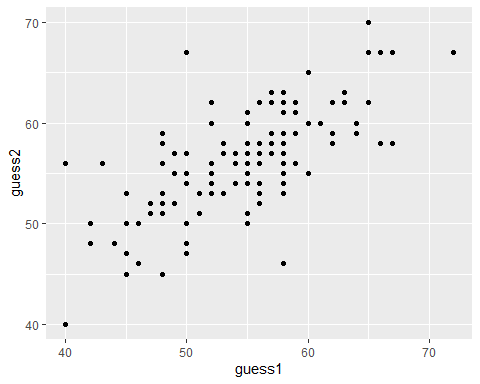
ggplot(data = age\_guess, aes(x = change)) +  
 geom\_histogram(binwidth = 2, fill = "royalblue", col = "yellow") +   
 theme\_bw() +  
 labs(x = "Change from first to second guess",  
 y = "Students in 431 for Fall 2022-2024",  
 title = "Most stayed close to their first guess.")



## Guess 1 vs. Guess 2 Scatterplot

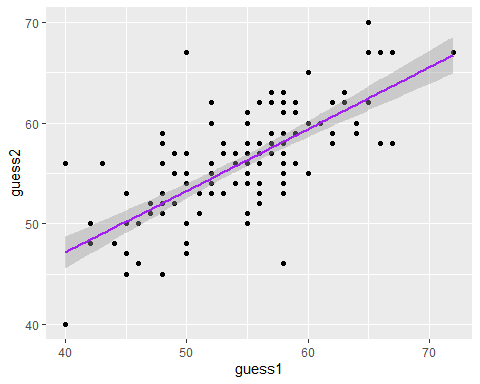
ggplot(data = age\_guess, aes(x = guess1, y = guess2)) +  
 geom\_point()

Warning: Removed 3 rows containing missing values or values outside the scale range  
(`geom\_point()`).



## Filter to complete cases, and add regression line

temp <- age\_guess |>  
 filter(complete.cases(guess1, guess2))  
  
ggplot(data = temp, aes(x = guess1, y = guess2)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", formula = y ~ x, col = "purple")



## What is that regression line?

lm(guess2 ~ guess1, data = age\_guess)

Call:  
lm(formula = guess2 ~ guess1, data = age\_guess)  
  
Coefficients:  
(Intercept) guess1   
 22.6760 0.6118

* Note that lm filters to complete cases by default.

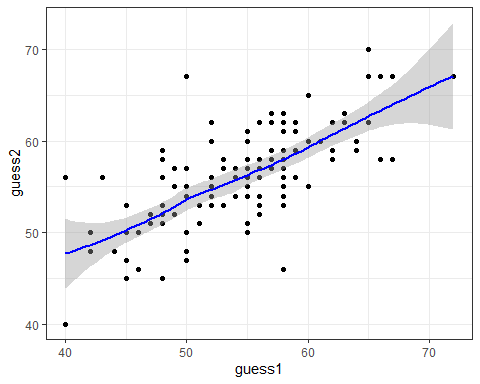
## Bayesian linear regression instead?

set.seed(431)  
stan\_glm(guess2 ~ guess1, data = age\_guess, refresh = 0)

stan\_glm  
 family: gaussian [identity]  
 formula: guess2 ~ guess1  
 observations: 145  
 predictors: 2  
------  
 Median MAD\_SD  
(Intercept) 22.8 2.7   
guess1 0.6 0.0   
  
Auxiliary parameter(s):  
 Median MAD\_SD  
sigma 3.7 0.2   
  
------  
\* For help interpreting the printed output see ?print.stanreg  
\* For info on the priors used see ?prior\_summary.stanreg

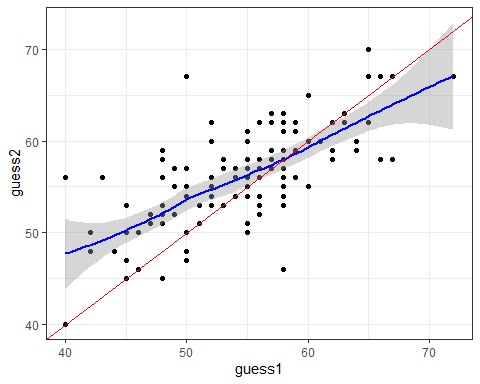
## How about a loess smooth curve?

temp <- age\_guess |>  
 filter(complete.cases(guess1, guess2))  
  
ggplot(data = temp, aes(x = guess1, y = guess2)) +  
 geom\_point() +  
 geom\_smooth(method = "loess", formula = y ~ x, col = "blue") +  
 theme\_bw()



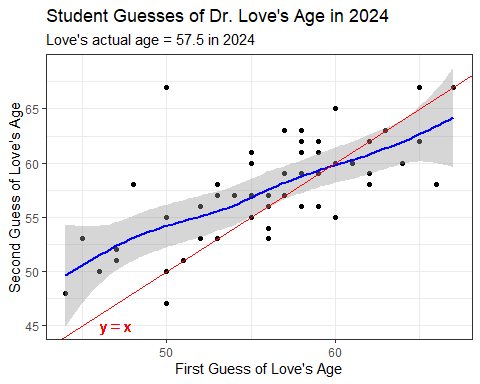
## Add y = x line (no change in guess)?

temp <- age\_guess |>  
 filter(complete.cases(guess1, guess2))  
  
ggplot(data = temp, aes(x = guess1, y = guess2)) +  
 geom\_point() +  
 geom\_smooth(method = "loess", formula = y ~ x, col = "blue") +  
 geom\_abline(intercept = 0, slope = 1, col = "red") +  
 theme\_bw()



## 2024 Data, With Better Labels

ggplot(data = temp |> filter(year == "2024"), aes(x = guess1, y = guess2)) +  
 geom\_point() +  
 geom\_smooth(method = "loess", formula = y ~ x, col = "blue") +  
 geom\_abline(intercept = 0, slope = 1, col = "red") +  
 geom\_text(x = 47, y = 45, label = "y = x", col = "red") +  
 labs(x = "First Guess of Love's Age",  
 y = "Second Guess of Love's Age",  
 title = "Student Guesses of Dr. Love's Age in 2024",  
 subtitle = "Love's actual age = 57.5 in 2024") +  
 theme\_bw()



# OK. That’s it for the slides. Back to [the Class 02 README](https://github.com/THOMASELOVE/431-classes-2024/tree/main/class02).