Course Reminders

Due Dates:

- A1 - due Sunday 4/14 (11:59 PM)
- Project Proposal due *next* Sunday

Notes:

- Group assignments were sent out by email. If you come off of the waitlist, you'll be assigned a group within the week.
- If you are enrolled and have not been assigned a group, I want to know that (email please)

Data Wrangling

getting the data you have in the format you need

- Reading files (CSV, JSON, XML) into Python
- `pandas`
  - `pd.read_csv()`
  - Indexing & subsetting
  - Exploring the data: `describe()`
  - Dropping columns: `drop()`
  - Missing Data

Think about the data you have & the data you want...and where things could have gone wrong during this process

Data Wrangling Warm-up

Getting to know you

Answers here will assist in group formation for the class project.

First Name *

Short answer text

Last Name *

Short answer text

What is your PID?

Short answer text

When would you be available to meet and work with your group for the course group project? SELECT all that apply.

- [ ] weekdays - morning (9 AM - noon)
- [ ] weekdays - afternoon (noon - 6 PM)
- [ ] weekdays - evening (after 6 PM)
- [ ] Friday nights (after 6 PM)
- [ ] Saturday - morning (9 AM - noon)
- [ ] Saturday - afternoon (noon - 6 PM)
- [ ] Saturday - evenings (after 6 PM)
- [ ] Sunday - morning (9 AM - noon)
- [ ] Sunday - afternoon (noon - 6 PM)
- [ ] Sunday - evenings (after 6 PM)

Which of the following do you enjoy thinking/learning about MOST? *

- ○ government & politics
- ○ movies
- ○ music
- ○ sports (watching)
- ○ sports (participating)
- ○ public health
- ○ climate change
- ○ machine learning
- ○ ethics
- ○ tech industry
- ○ economics
- ○ education
- ○ Other...

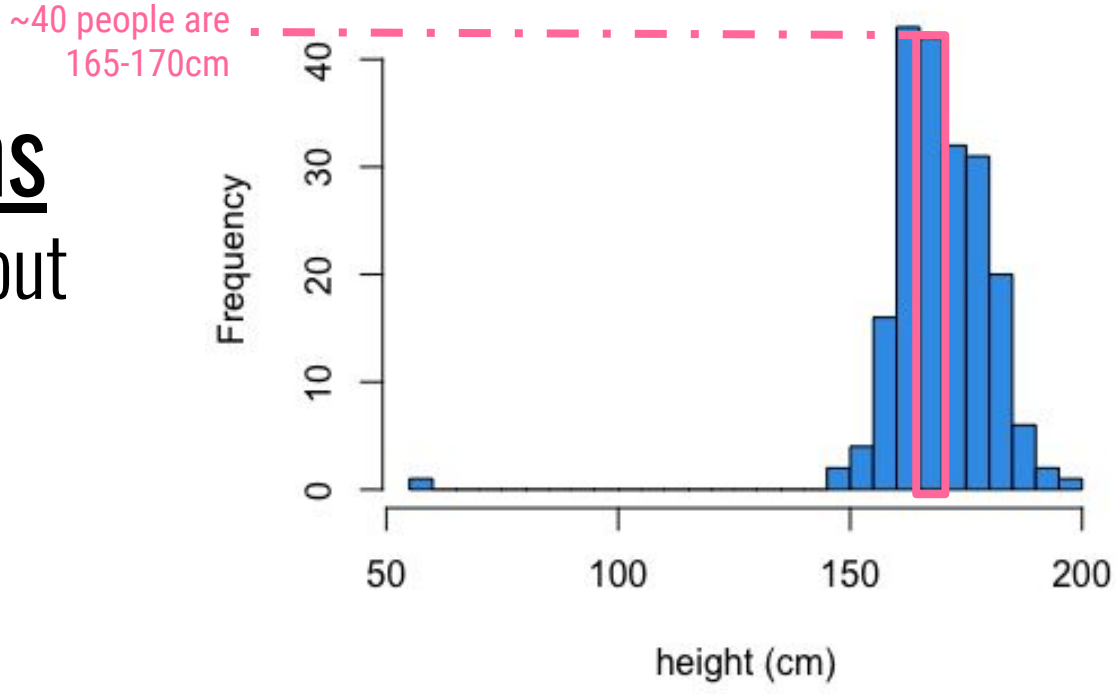
Data Visualization

Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

Histograms

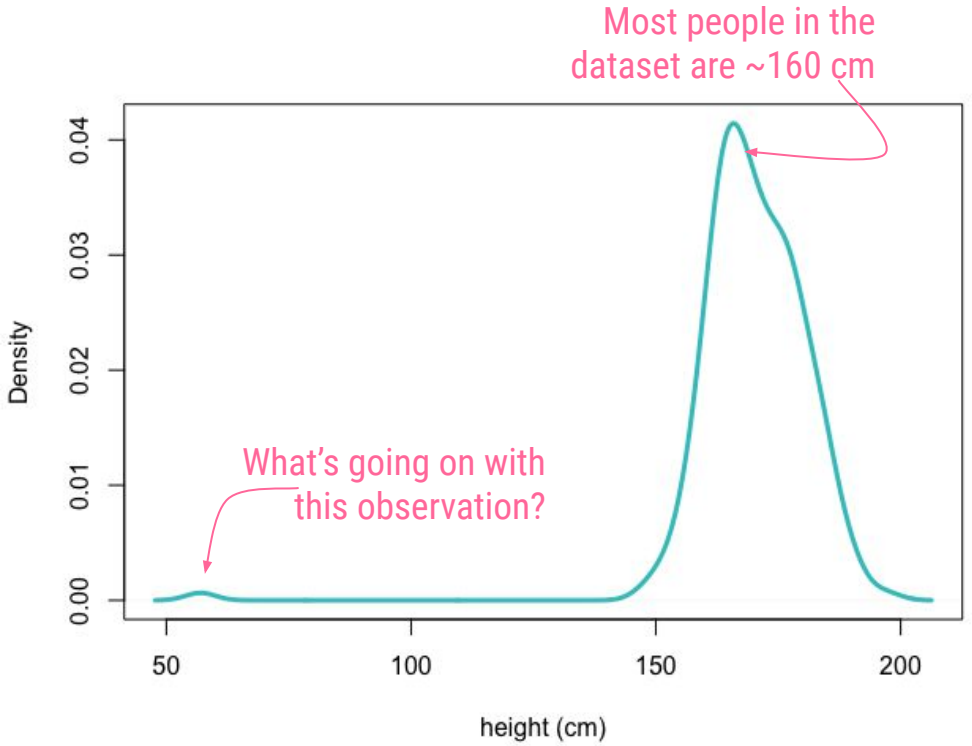
Information about a single quantitative variable



Range of possible height values is easily visualized

Densityplot

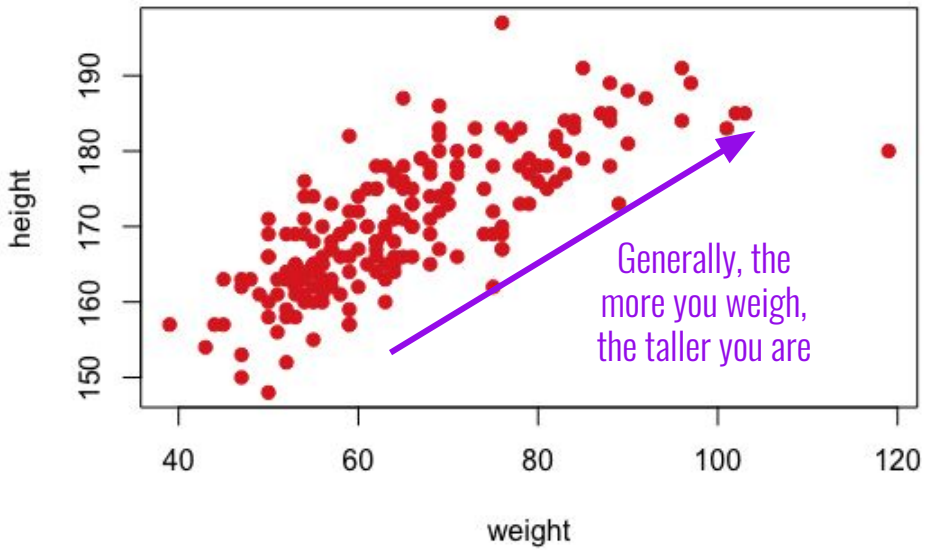
Information about a single quantitative variable



A smoothed version of a histogram - demonstrates the *distribution* of the data; helps to identify extreme values

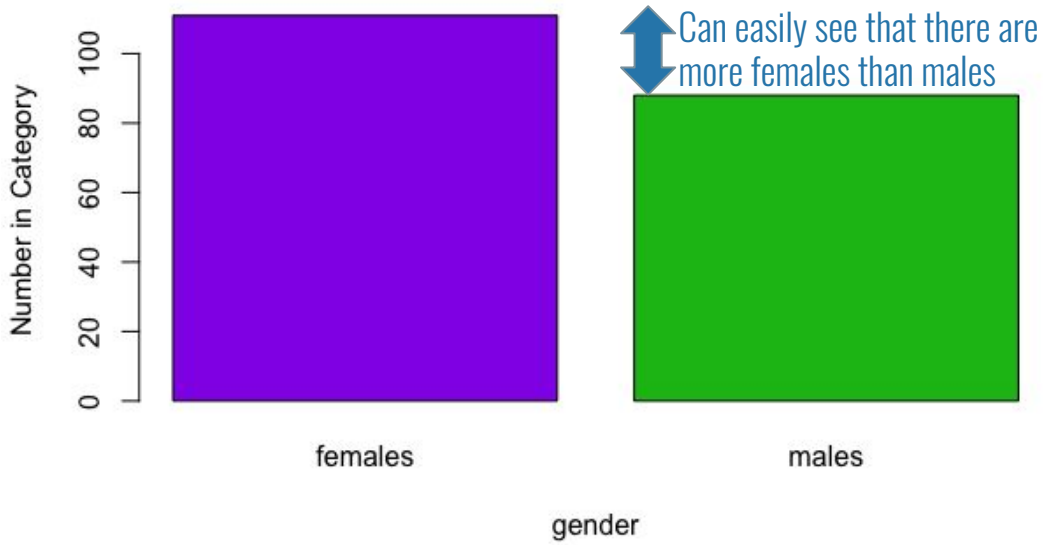
Scatterplot

Relationship between two quantitative variables



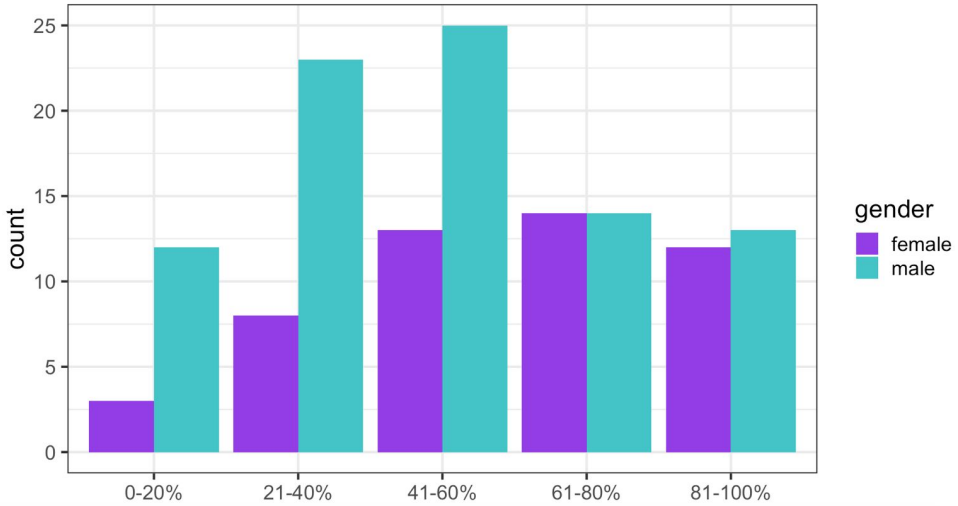
Barplot

Count of values within a single categorical variable



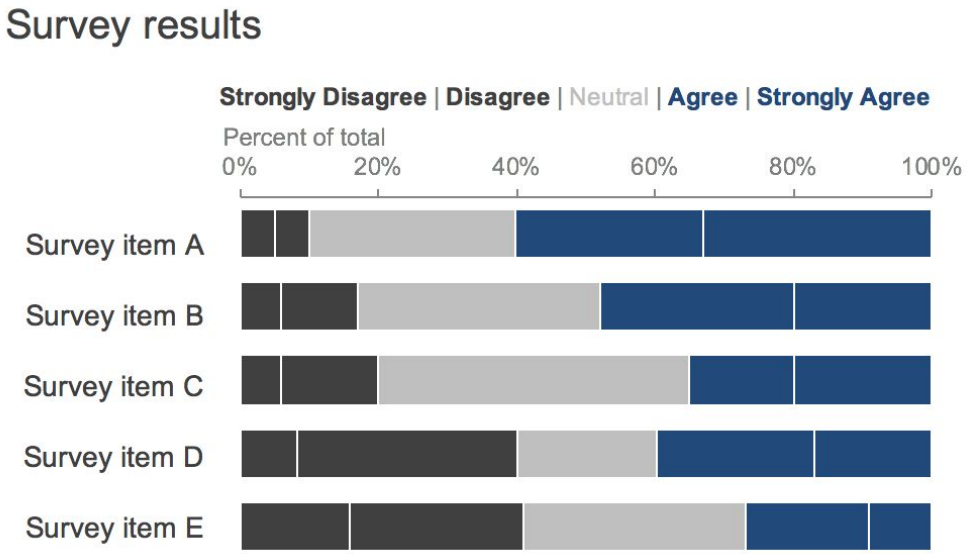
Grouped Barplot

Count of values broken down across two categorical variables



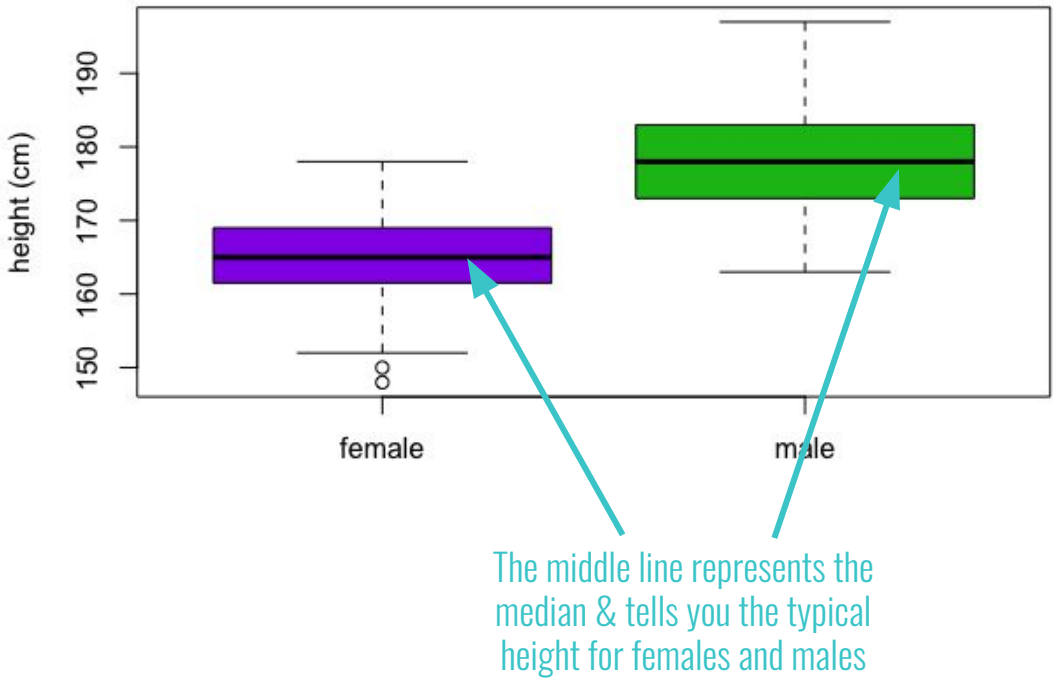
Stacked Barplot

Count/proportion of values broken down across two categorical variables



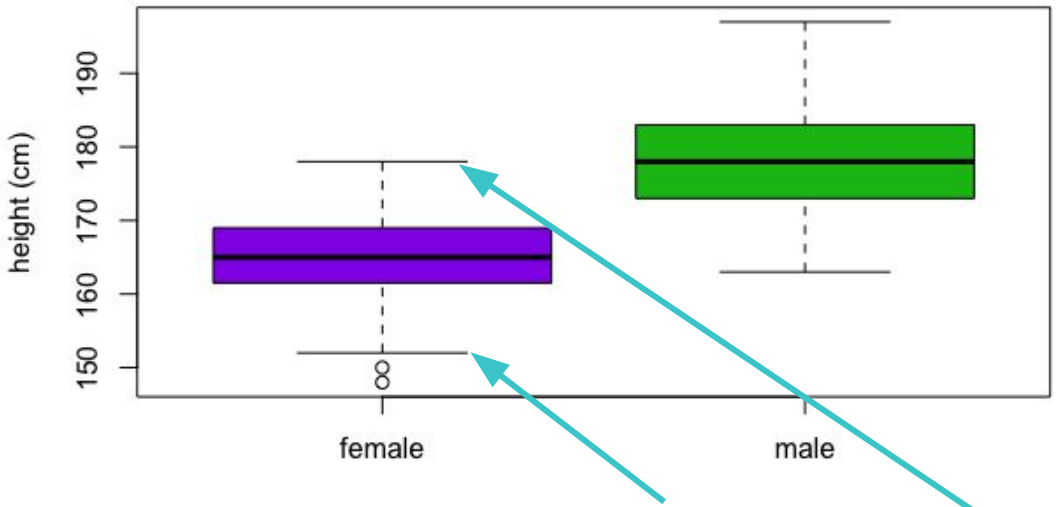
Boxplot

Summary of a quantitative variable broken down by a categorical variable



Boxplot

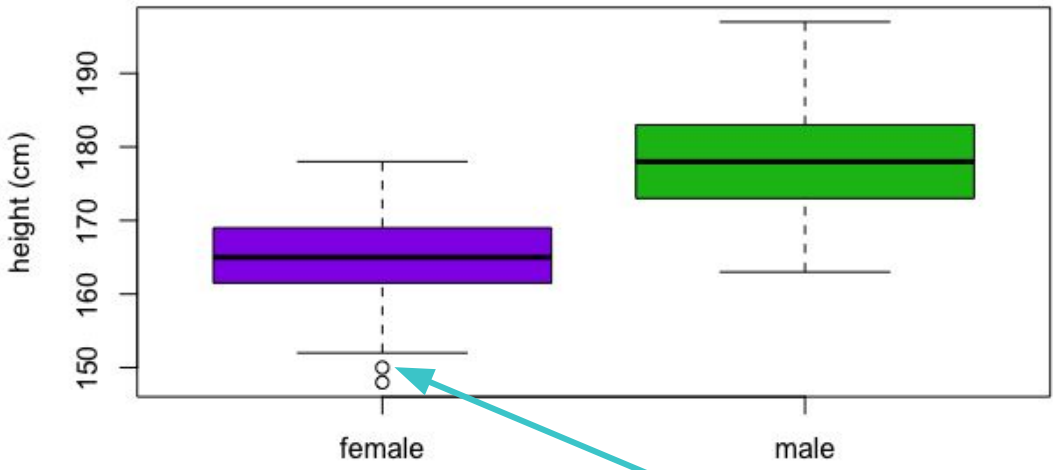
Summary of a quantitative variable broken down by a categorical variable



The lines give you an idea of the typical range of values for each category

Boxplot

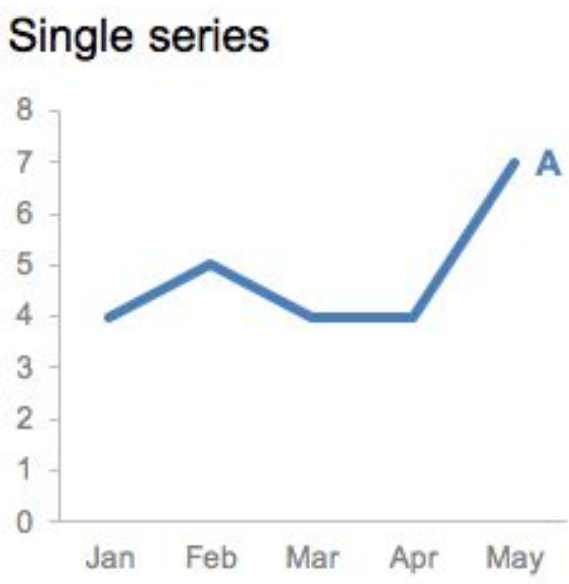
Summary of a quantitative variable broken down by a categorical variable

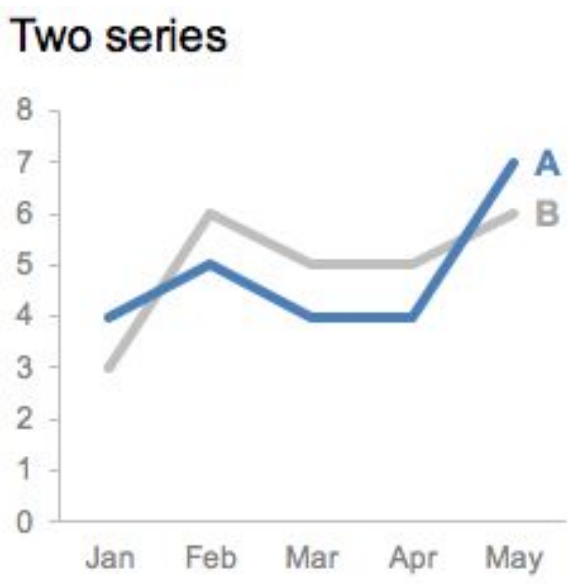


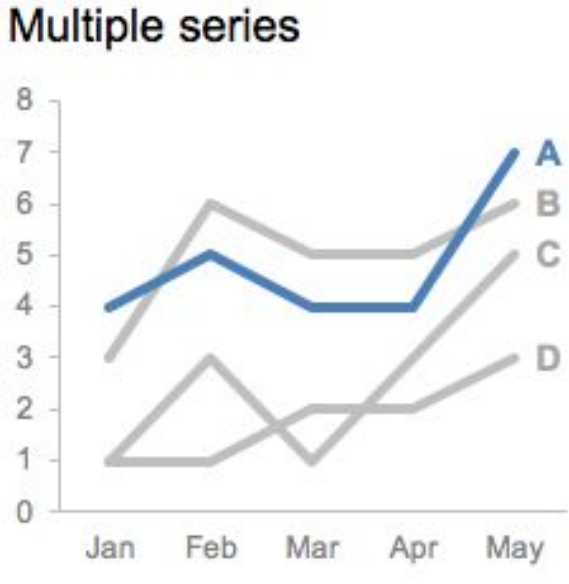
Values outside the typical range are shown as circles. These are known as **outliers**.

Line plot

quantitative trend over time



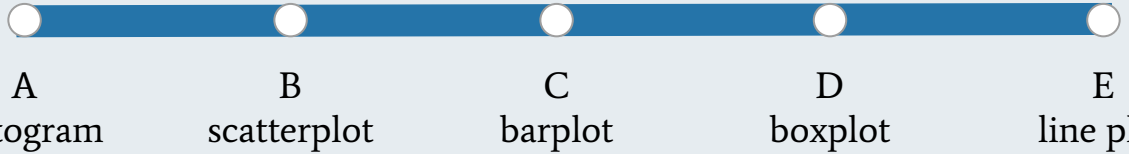




Graphical Choices

You want to visualize how many people in your dataset prefer chocolate chip cookies and how many prefer oatmeal raisin cookies.

What type of visualization would be most appropriate?



Graphical Choices

Your interested in visualizing how many servings of milk an individual drinks each day among those who prefer chocolate chip cookies and those who prefer oatmeal raisin cookies.

What type of visualization would be most appropriate?

A
histogram

B
scatterplot

C
barplot

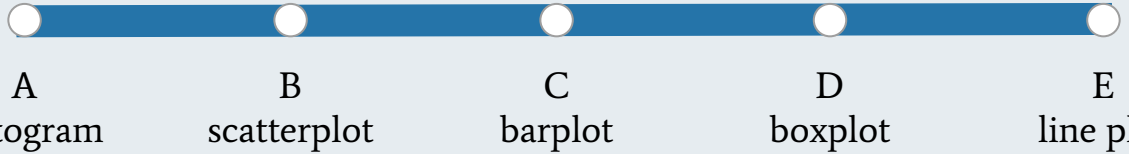
D
boxplot

E
line plot

Graphical Choices

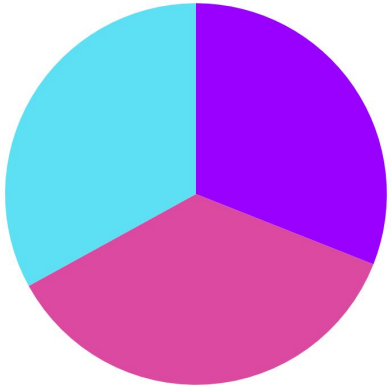
Your interested in visualizing how many servings of milk an individual drinks each year over the course of their life.

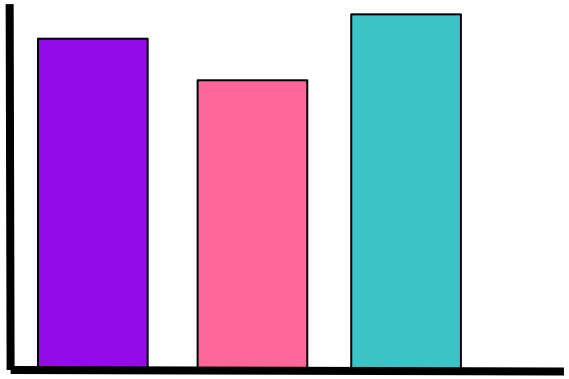
What type of visualization would be most appropriate?



Visualization Best Practices

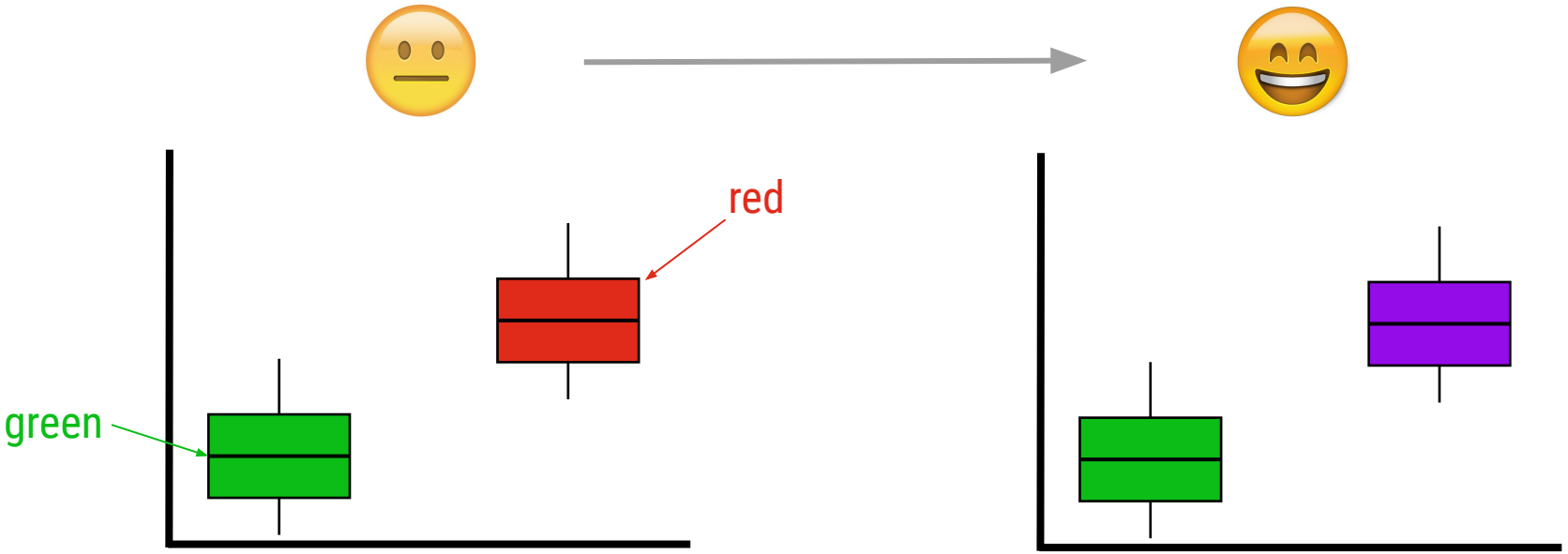
Choose the right type of visualization





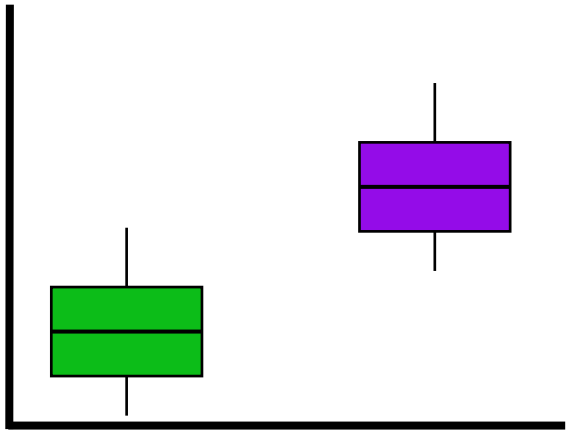
When looking at values, bar charts make it much easier to see the difference between groups!

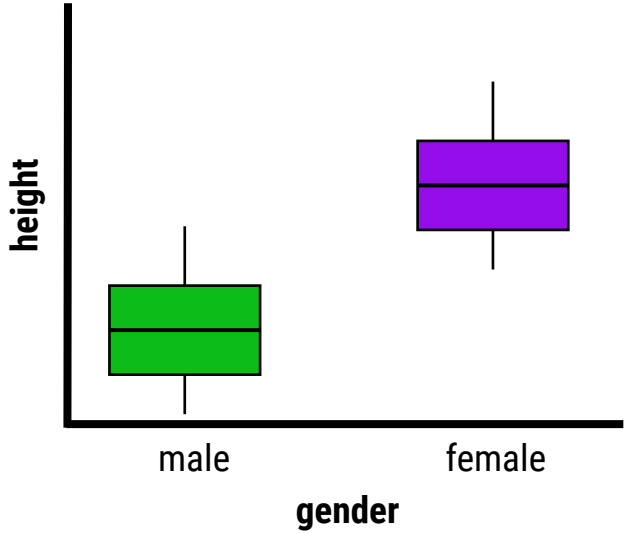
Be mindful when choosing colors



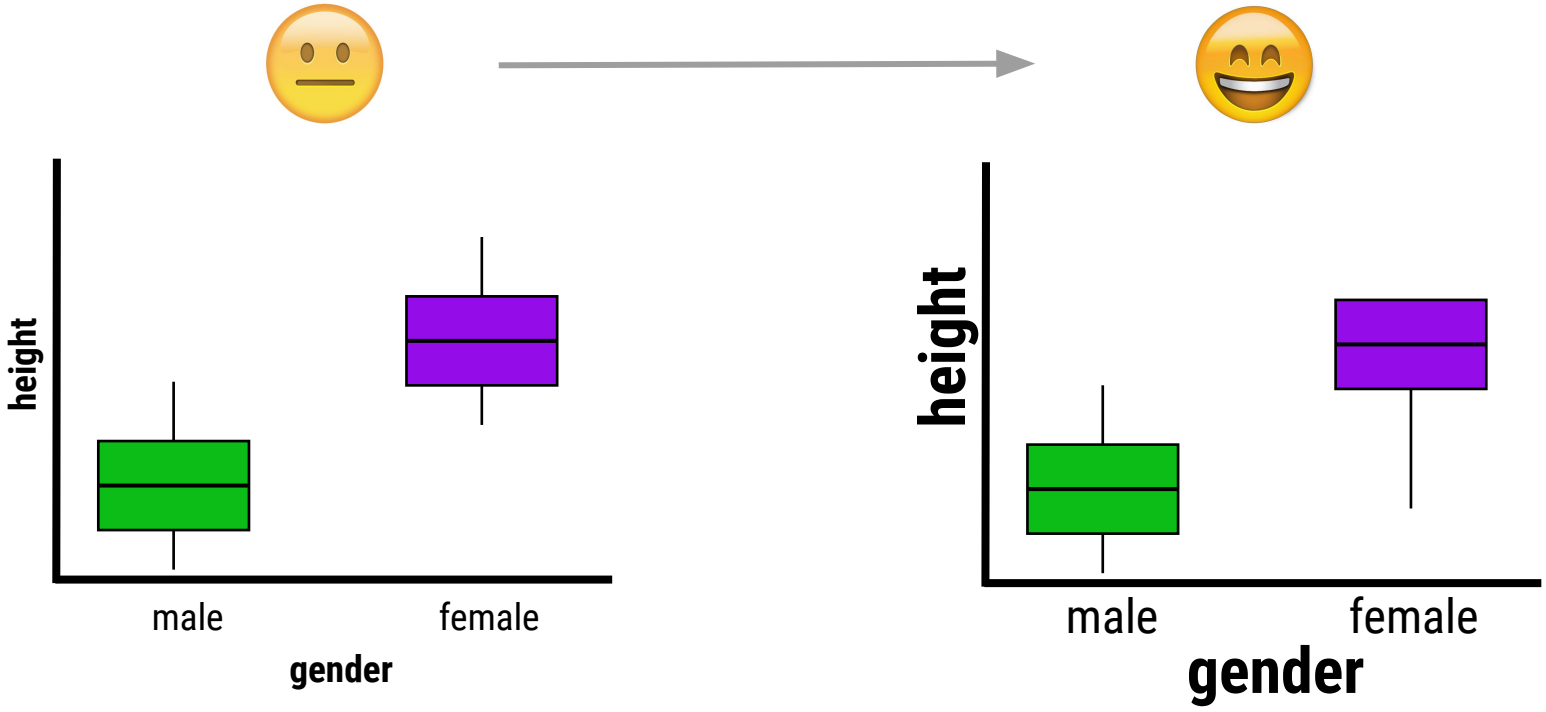
Many color-blind individuals cannot see the difference between red and green.

Label your axes!

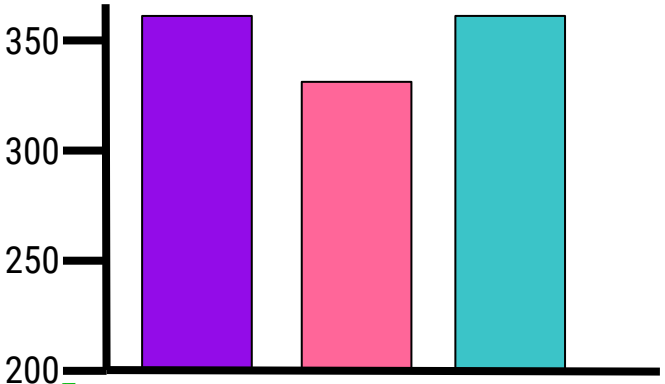


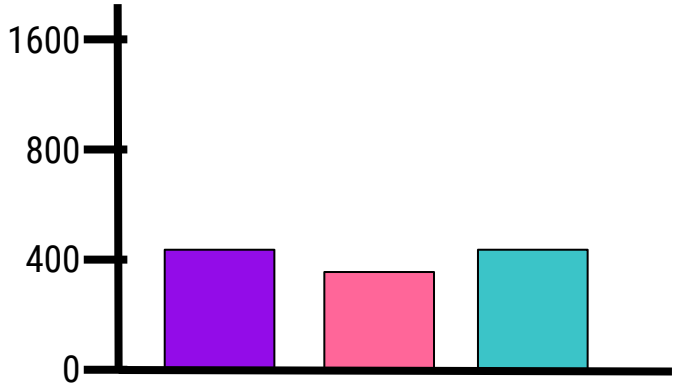


Make sure the text size is big enough!



Use y-axes that start at 0 for barplots

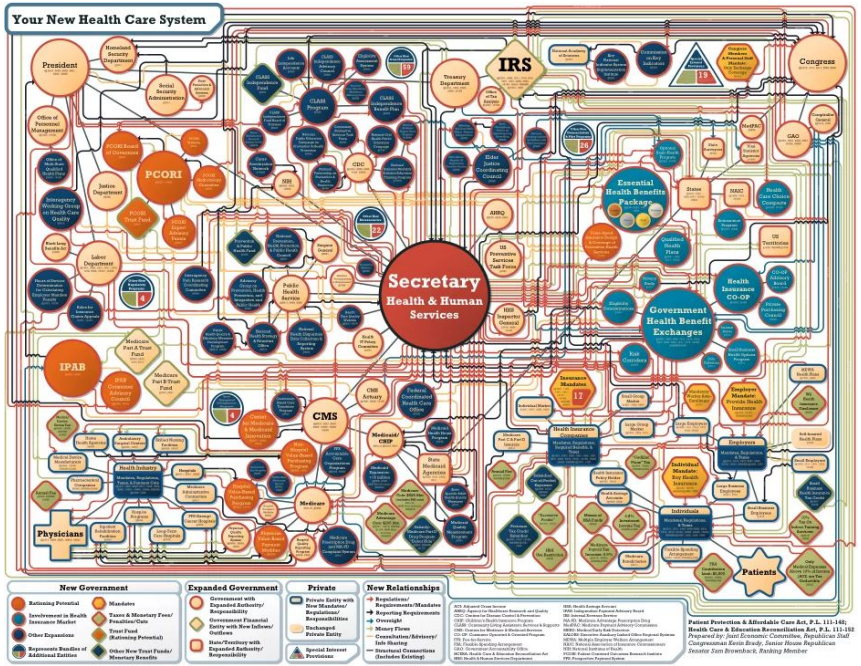




The y-axis starts at 200.
This is misleading & makes differences look larger than they actually are

Keep it Simple

😐



"...detailed organization chart displays a bewildering array of new government agencies, regulations and mandates"

https://www.jec.senate.gov/public/index.cfm/republicans/committeenews?ID=bb302d88-3d0d-4424-8e33-3c5d2578c2b0

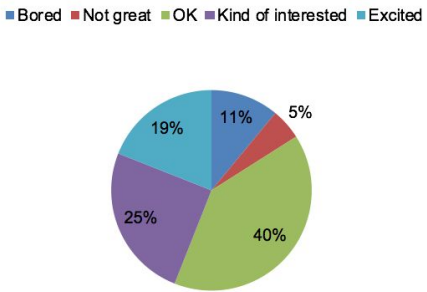
Everything on the page should serve a purpose. If it doesn't, remove it or edit it (declutter!).

AIM TO IMPROVE YOUR:

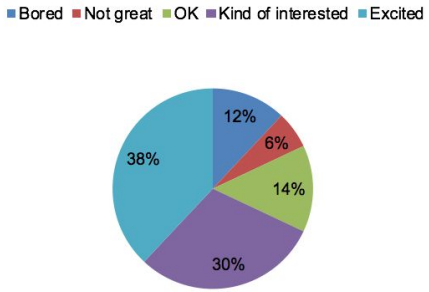
data:ink ratio

Survey Results

PRE: How do you feel about doing science?



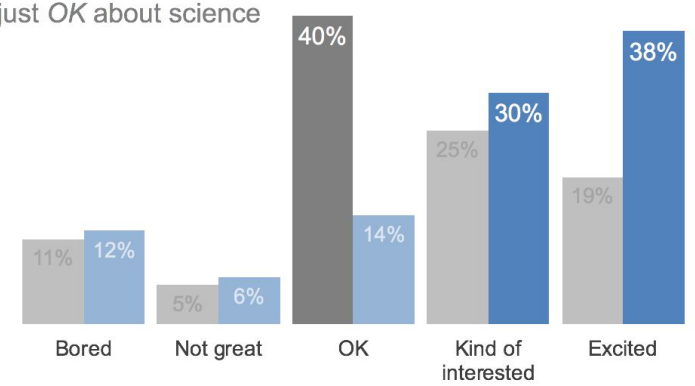
POST: How do you feel about doing science?



Pilot program was a success

How do you feel about science?

BEFORE program, the majority of children felt just *OK* about science

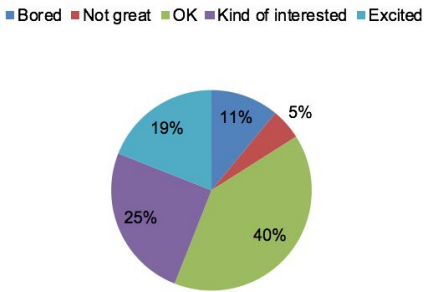


AFTER program, more children were *Kind of interested* & *Excited* about science.

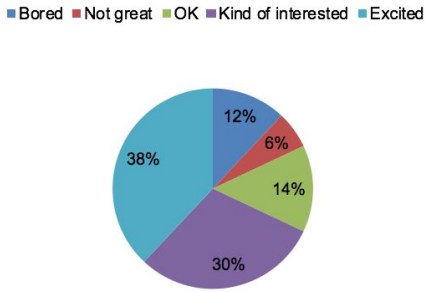
Based on survey of 100 students conducted before and after pilot program (100% response rate on both surveys).

Survey Results

PRE: How do you feel about doing science?



POST: How do you feel about doing science?



Pilot program was a success

After the pilot program,

68%

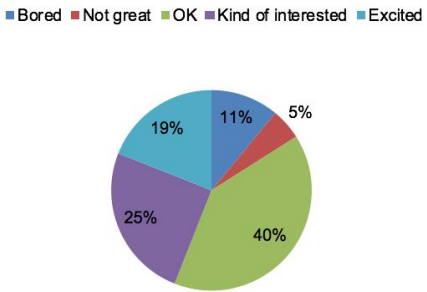
of kids expressed interest towards science,
compared to 44% going into the program.

Based on survey of 100 students conducted before and after pilot program (100% response rate on both surveys).

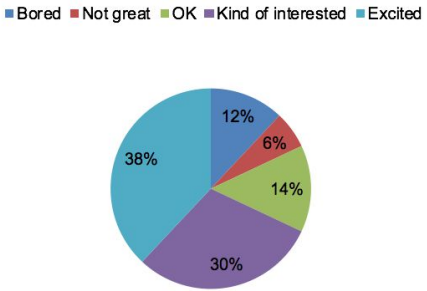
Pilot program was a success

Survey Results

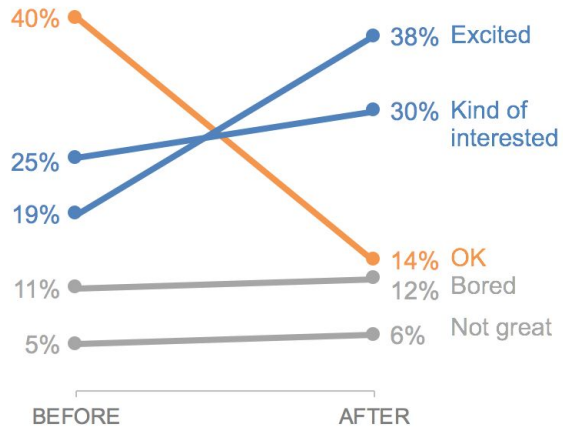
PRE: How do you feel about doing science?



POST: How do you feel about doing science?



How do you feel about science?



BEFORE program, the majority of children felt just *OK* about science

AFTER program, more children were *Kind of interested* & *Excited* about science.

Based on survey of 100 students conducted before and after pilot program (100% response rate on both surveys).

less is more

(effective)
(attractive)
(impactive)