Introduction to Data Science

Data and Visualization

Joanna Bieri DATA101

Welcome

Important Information

- Email: joanna_bieri@redlands.edu
- Office Hours: Duke 209 Click Here for Joanna's Schedule

Today's Big Ideas

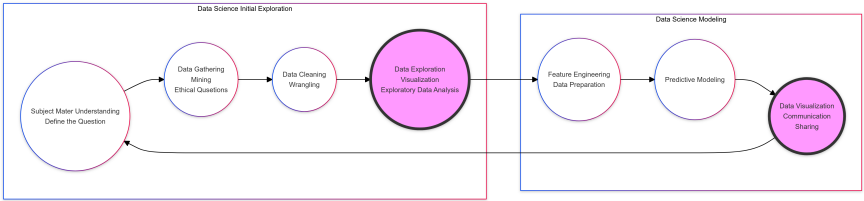
- The data science lifecycle: where visualization fits in
- Getting to know a data set before you analyze it
- Exploratory Data Analysis (EDA) with graphs
- Why summary statistics are not enough

The Data Science Lifecycle

Big Idea: Data Science is a Cycle

- Data science is not one step, it is a full cycle: ask a question, gather data, clean it, explore it, model it, share what you found
- Visualization shows up twice in that cycle, once while you are exploring and once when you are communicating
- Two jobs are always yours: make your work **reproducible**, and really **get to know your data**

The Lifecycle



Check Your Understanding

- **Q1.** Which part of the cycle are we in today when we make a graph to see what the data looks like?
- **Q2.** What does it mean for your work to be reproducible?
- **Q3.** Why does the cycle loop back to the beginning instead of just ending?

Getting to Know Your Data

Big Idea: Observations and Variables

- Each row of a data set is an **observation**, one thing you measured
- Each column is a **variable**, one piece of information about every observation
- `NaN` means the value is **missing**, which is different from the value being zero
- Before analyzing anything, find out how big the data is and what the columns actually mean

Check Your Understanding

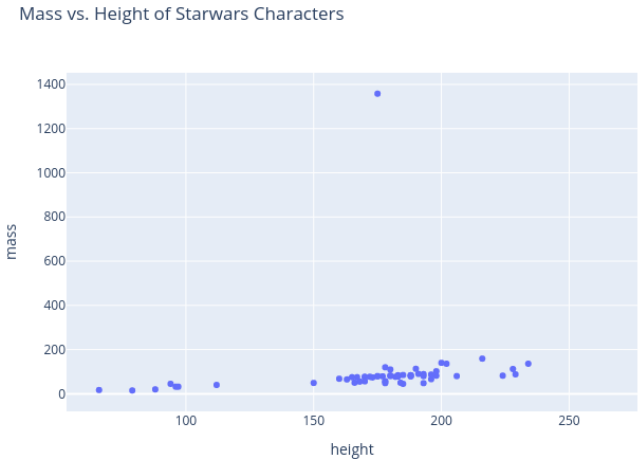
- **Q4.** In the Star Wars data, is Luke Skywalker a row or a column? What about height?
- **Q5.** Which command tells you the number of rows and columns at the same time?
- **Q6.** C-3PO has `NaN` for hair color. Is that a mistake in the data, or does it make sense?
- **Q7.** Why does capitalization matter when you search for a character by name?

Exploring with Graphs

Big Idea: Exploratory Data Analysis

- **EDA** is how you get a feel for a data set: summary statistics, visualization, and wrangling
- A scatter plot puts one dot per observation, so you can see the whole data set at once
- Graphs make the odd points jump out. One Star Wars character has a mass that is way off from everyone else
- Adding `hover_data` lets you point at a dot and find out which observation it is

Example: Mass vs. Height



Check Your Understanding

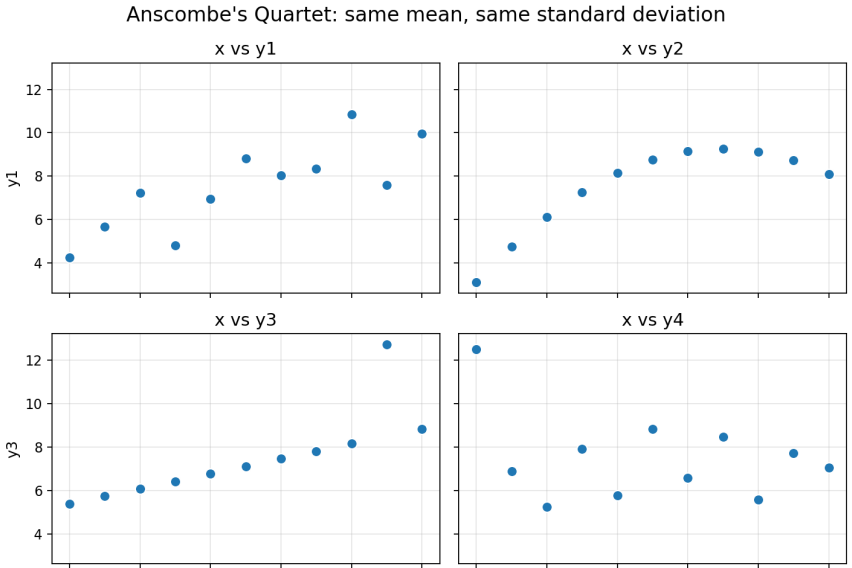
- **Q8.** Looking at the graph, would you say taller characters generally have more mass?
- **Q9.** There is a point way up at the top. How would you figure out who it is?
- **Q10.** What would you change in the code to plot birth year against mass instead?

Why Visualize?!?!

Big Idea: Statistics Alone Can Fool You

- Anscombe's Quartet is four small data sets, y1 through y4, that share the same x values
- All four have the **same mean** and about the **same standard deviation**
- If you only look at the summary statistics you would guess the four graphs look the same
- Graph them and they are completely different. This is why we always look at the data

Example: Anscombe's Quartet



Check Your Understanding

- **Q11.** Before seeing the graphs, what would you have predicted from the summary statistics alone?
- **Q12.** Which of the four looks like a straight line relationship? Which one has a single point pulling everything around?
- **Q13.** What does this tell you about handing someone a table of averages and calling it analysis?

Wrap-Up

Go Explore!

Now it is your turn to explore the data!

Questions?