

Introduction to Data Visualization D1 - 2023-10

October 12, 2023

1 Introduction to Data Visualization

By Ruben D. Canlas Jr. rubencanlas@gmail.com

1.1 Outline

1. Why is it important?
2. What is it?
3. How to use data visualization?

1.2 1. Why is Data Visualization important?

Let's start with a historical mystery...

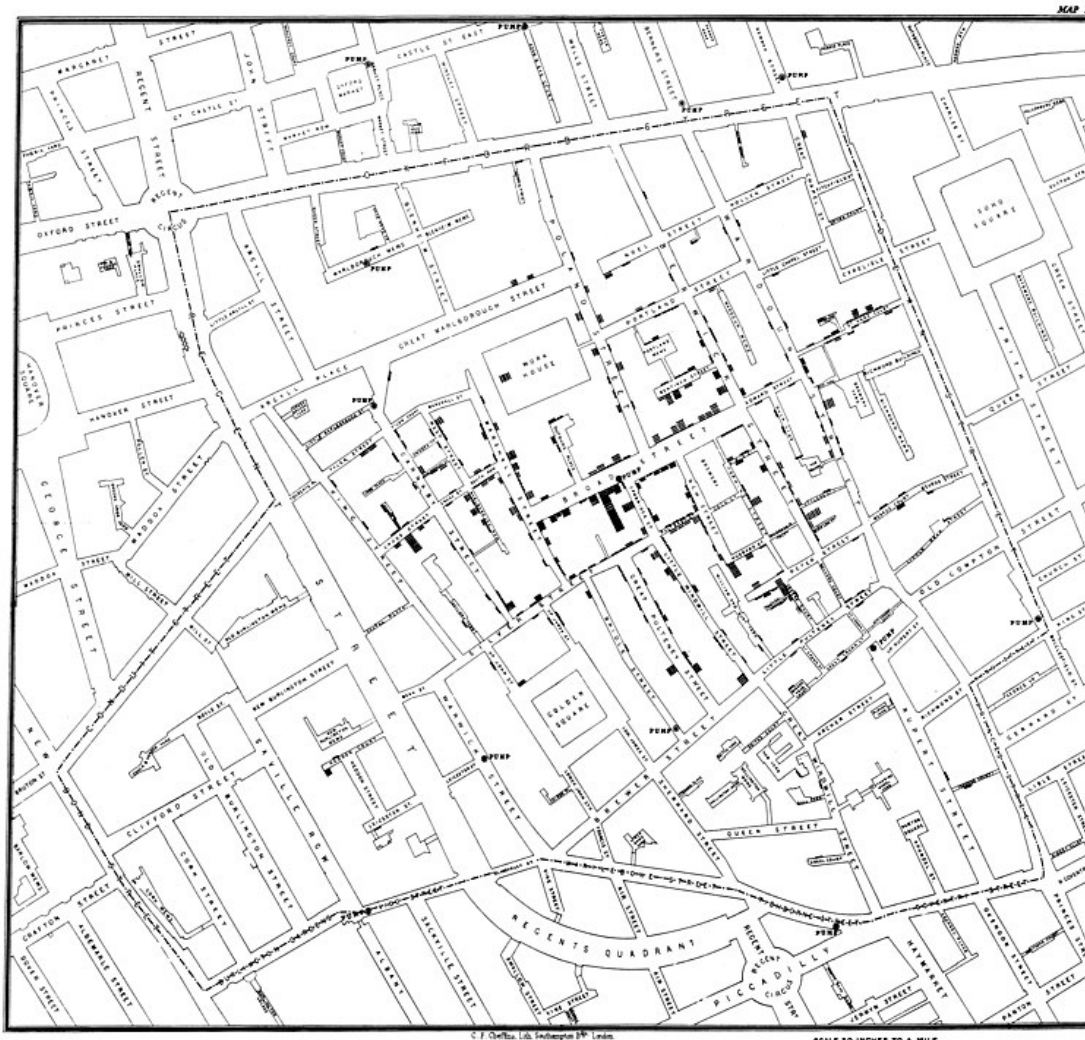
1.2.1 The London Cholera Epidemic

- 1854 cholera outbreak
- Killed 616 people
- Source of infection traced by **Dr. John Snow**
- Detective work with simple (but ingenious) visualization

What did John Snow do?

1. Got the addresses of the victims
2. Got a map of London
3. Plotted one death in the vicinity as one rectangle

This is the map that he made. ...



1.3 Notice the following:

- At the center of the map, see the tallest black stack of bars.
- Dr. Snow suspected that water was the source of the outbreak.
- Victims frequented a pub in the area.
- Pub drew water from a nearby pump. You can still see the location of the pump near the tall stack of bars.

1.4 Breakout: Visual Inspection

7 min

1. As a group, come up with one sentence summarizing how visualization over a map helped solve the cholera outbreak.
2. Discuss how plotting numbers against a map can help your team or organization make sense of data. Write down one example.
3. Write down your answers in Jamboard. One idea: one sticky note.

1.5 Sample Insights

1. Plotting morbidity numbers against the map added a spatial dimension that enriched the information.
2. Visualizing the numbers against the map made it easier to see patterns from the data.

1.5.1 Consider this sample dataset:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | 7 | 7 | 8 | 4 | 7 | 5 | 9 | 4 | 3 | 2 | 7 |
| 5 | 7 | 3 | 1 | 5 | 9 | 2 | 5 | 5 | 5 | 6 | 7 |
| 9 | 5 | 1 | 8 | 7 | 2 | 5 | 1 | 1 | 6 | 1 | 2 |
| 2 | 7 | 8 | 7 | 6 | 1 | 6 | 6 | 7 | 1 | 2 | 1 |
| 6 | 9 | 1 | 3 | 4 | 4 | 2 | 5 | 7 | 5 | 2 | 9 |
| 3 | 5 | 4 | 6 | 5 | 2 | 2 | 9 | 9 | 2 | 3 | 8 |
| 4 | 5 | 8 | 3 | 1 | 8 | 1 | 3 | 2 | 1 | 1 | 1 |
| 4 | 5 | 7 | 1 | 4 | 9 | 9 | 9 | 2 | 8 | 4 | 4 |
| 4 | 2 | 3 | 7 | 2 | 5 | 7 | 9 | 3 | 8 | 6 | 5 |
| 6 | 4 | 6 | 1 | 5 | 9 | 4 | 6 | 8 | 4 | 2 | 5 |
| 1 | 9 | 8 | 2 | 6 | 4 | 6 | 9 | 5 | 7 | 6 | 1 |
| 7 | 9 | 9 | 8 | 1 | 7 | 3 | 1 | 5 | 5 | 1 | 8 |

1.5.2 Now answer these questions:

1. How many times does the number 5 occur?
2. Which number occurs the most?

1.5.3 Let's tweak the formatting a bit...

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 9 | 7 | 7 | 8 | 4 | 7 | 5 | 9 | 4 | 3 | 2 | 7 |
| 5 | 7 | 3 | 1 | 5 | 9 | 2 | 5 | 5 | 5 | 6 | 7 |
| 9 | 5 | 1 | 8 | 7 | 2 | 5 | 1 | 1 | 6 | 1 | 2 |
| 2 | 7 | 8 | 7 | 6 | 1 | 6 | 6 | 7 | 1 | 2 | 1 |
| 6 | 9 | 1 | 3 | 4 | 4 | 2 | 5 | 7 | 5 | 2 | 9 |
| 3 | 5 | 4 | 6 | 5 | 2 | 2 | 9 | 9 | 2 | 3 | 8 |
| 4 | 5 | 8 | 3 | 1 | 8 | 1 | 3 | 2 | 1 | 1 | 1 |
| 4 | 5 | 7 | 1 | 4 | 9 | 9 | 9 | 2 | 8 | 4 | 4 |
| 4 | 2 | 3 | 7 | 2 | 5 | 7 | 9 | 3 | 8 | 6 | 5 |
| 6 | 4 | 6 | 1 | 5 | 9 | 4 | 6 | 8 | 4 | 2 | 5 |
| 1 | 9 | 8 | 2 | 6 | 4 | 6 | 9 | 5 | 7 | 6 | 1 |
| 7 | 9 | 9 | 8 | 1 | 7 | 3 | 1 | 5 | 5 | 1 | 8 |

Now, can you answer how many times **5** occurs?

Which number occurs most frequently?

1.5.4 Insights

1. Our eyes and brain are programmed to notice differences quickly
2. Coloring will make values stand out
3. Font formatting can also highlight differences (eg, bold, italics, size, capitalization etc)

Let's make it more interesting...

The same dataset, but now presented as bar shapes.

1. What do you think do the x- and y-axes represent?
2. Which numbers occur the most?
3. How many instances of the most frequently appearing numbers?
4. Which numbers occur the least frequently?

1.6 Insights

1. Use of color to make numbers stand out
2. Font formatting (eg, bold, italics, etc)
3. Use of shape (bar) adds a new dimension: length to represent count

2 2.What is Data Visualization?

Now let's define Data Visualization

Data Visualization. The use of charts, diagrams, pictures to represent information.

Below is the same dataset but with a slight change. *What changed?*

One more time:

1. Which numbers occur the most?
2. How many instances of the most frequently appearing numbers?
3. Which numbers occur the least frequently?

How easy is it now to answer the questions? Why?

2.1 Insights

1. Use of color to make numbers stand out.
2. Font formatting (eg, bold, italics, etc).
3. Use of shape (bar) adds a new dimension: length to represent count.
4. Sorting facilitates comparison.

2.2 Stretch and Summarize! (2 min)

Grab pen and paper and write down a summary of important concepts you've learned.

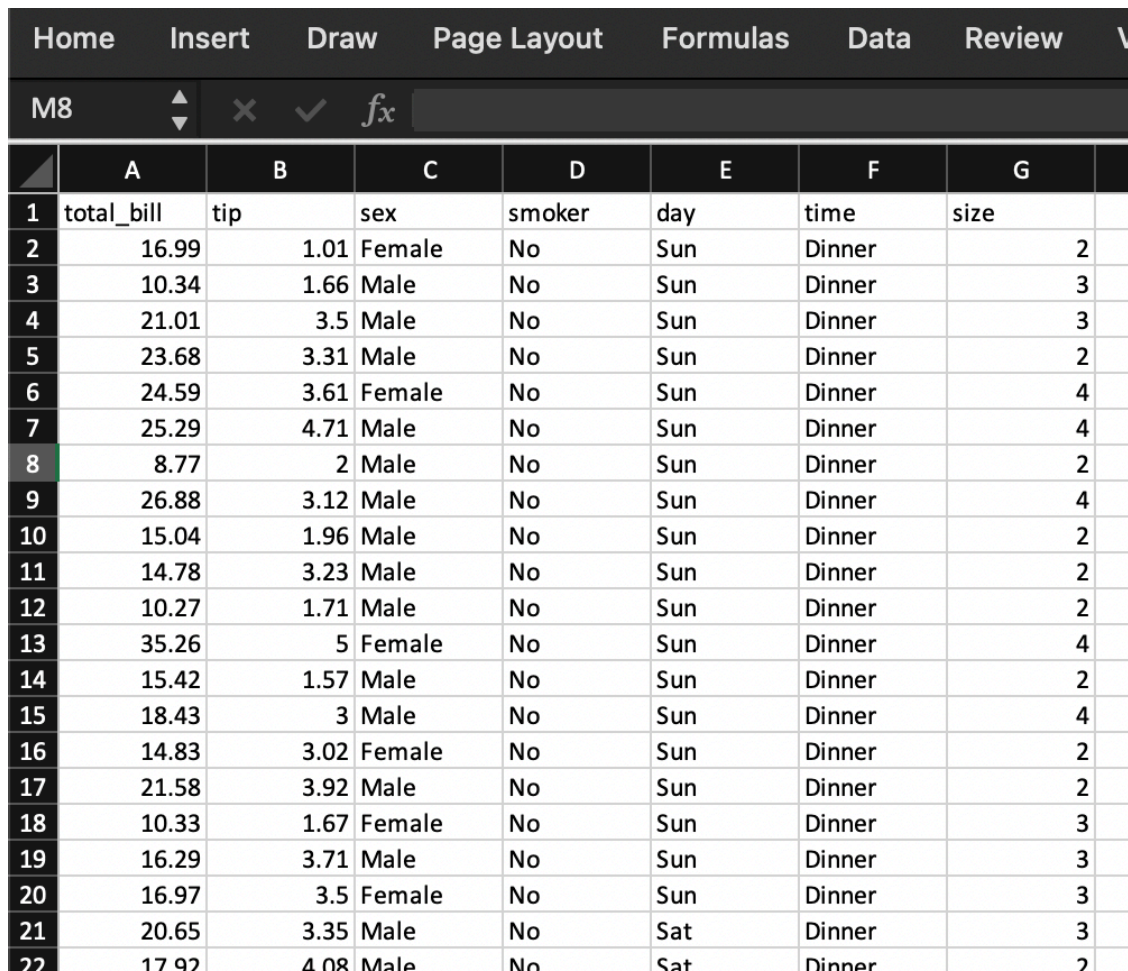
2.2.1 Let's do more exploration

The technique I will demonstrate today is a subset of Exploratory Data Analysis (EDA).

We call this VEDA: Visual Exploratory Data Analysis.

3. How to use data visualization?

3.1 Let's start with your usual table of data, aka the dataset:



| | A | B | C | D | E | F | G | |
|----|------------|------|--------|--------|-----|--------|------|--|
| 1 | total_bill | tip | sex | smoker | day | time | size | |
| 2 | 16.99 | 1.01 | Female | No | Sun | Dinner | 2 | |
| 3 | 10.34 | 1.66 | Male | No | Sun | Dinner | 3 | |
| 4 | 21.01 | 3.5 | Male | No | Sun | Dinner | 3 | |
| 5 | 23.68 | 3.31 | Male | No | Sun | Dinner | 2 | |
| 6 | 24.59 | 3.61 | Female | No | Sun | Dinner | 4 | |
| 7 | 25.29 | 4.71 | Male | No | Sun | Dinner | 4 | |
| 8 | 8.77 | 2 | Male | No | Sun | Dinner | 2 | |
| 9 | 26.88 | 3.12 | Male | No | Sun | Dinner | 4 | |
| 10 | 15.04 | 1.96 | Male | No | Sun | Dinner | 2 | |
| 11 | 14.78 | 3.23 | Male | No | Sun | Dinner | 2 | |
| 12 | 10.27 | 1.71 | Male | No | Sun | Dinner | 2 | |
| 13 | 35.26 | 5 | Female | No | Sun | Dinner | 4 | |
| 14 | 15.42 | 1.57 | Male | No | Sun | Dinner | 2 | |
| 15 | 18.43 | 3 | Male | No | Sun | Dinner | 4 | |
| 16 | 14.83 | 3.02 | Female | No | Sun | Dinner | 2 | |
| 17 | 21.58 | 3.92 | Male | No | Sun | Dinner | 2 | |
| 18 | 10.33 | 1.67 | Female | No | Sun | Dinner | 3 | |
| 19 | 16.29 | 3.71 | Male | No | Sun | Dinner | 3 | |
| 20 | 16.97 | 3.5 | Female | No | Sun | Dinner | 3 | |
| 21 | 20.65 | 3.35 | Male | No | Sat | Dinner | 3 | |
| 22 | 17.92 | 4.08 | Male | No | Sat | Dinner | 2 | |

3.1.1 Note the following:

1. It's in CSV format.
2. It's an image of how the dataset will look like, using Excel.
3. I could load this same CSV file into Jupyter Notebook using a tool called **pandas**.

...

3.2 Exploring the Tips Dataset

Tips Dataset available here: [Download tips.csv in Github](#)

```
[1]: # Code 1.1
import pandas as pd
tips = pd.read_csv("tips.csv")
display(tips)
```

| | total_bill | tip | sex | smoker | day | time | size |
|-----|------------|------|--------|--------|------|--------|------|
| 0 | 16.99 | 1.01 | Female | No | Sun | Dinner | 2 |
| 1 | 10.34 | 1.66 | Male | No | Sun | Dinner | 3 |
| 2 | 21.01 | 3.50 | Male | No | Sun | Dinner | 3 |
| 3 | 23.68 | 3.31 | Male | No | Sun | Dinner | 2 |
| 4 | 24.59 | 3.61 | Female | No | Sun | Dinner | 4 |
| .. | ... | ... | ... | ... | ... | ... | ... |
| 239 | 29.03 | 5.92 | Male | No | Sat | Dinner | 3 |
| 240 | 27.18 | 2.00 | Female | Yes | Sat | Dinner | 2 |
| 241 | 22.67 | 2.00 | Male | Yes | Sat | Dinner | 2 |
| 242 | 17.82 | 1.75 | Male | No | Sat | Dinner | 2 |
| 243 | 18.78 | 3.00 | Female | No | Thur | Dinner | 2 |

[244 rows x 7 columns]

3.3 Stand and stretch

1. At first glance, what information can we get from the dataset?
2. Guess what each type of information means?

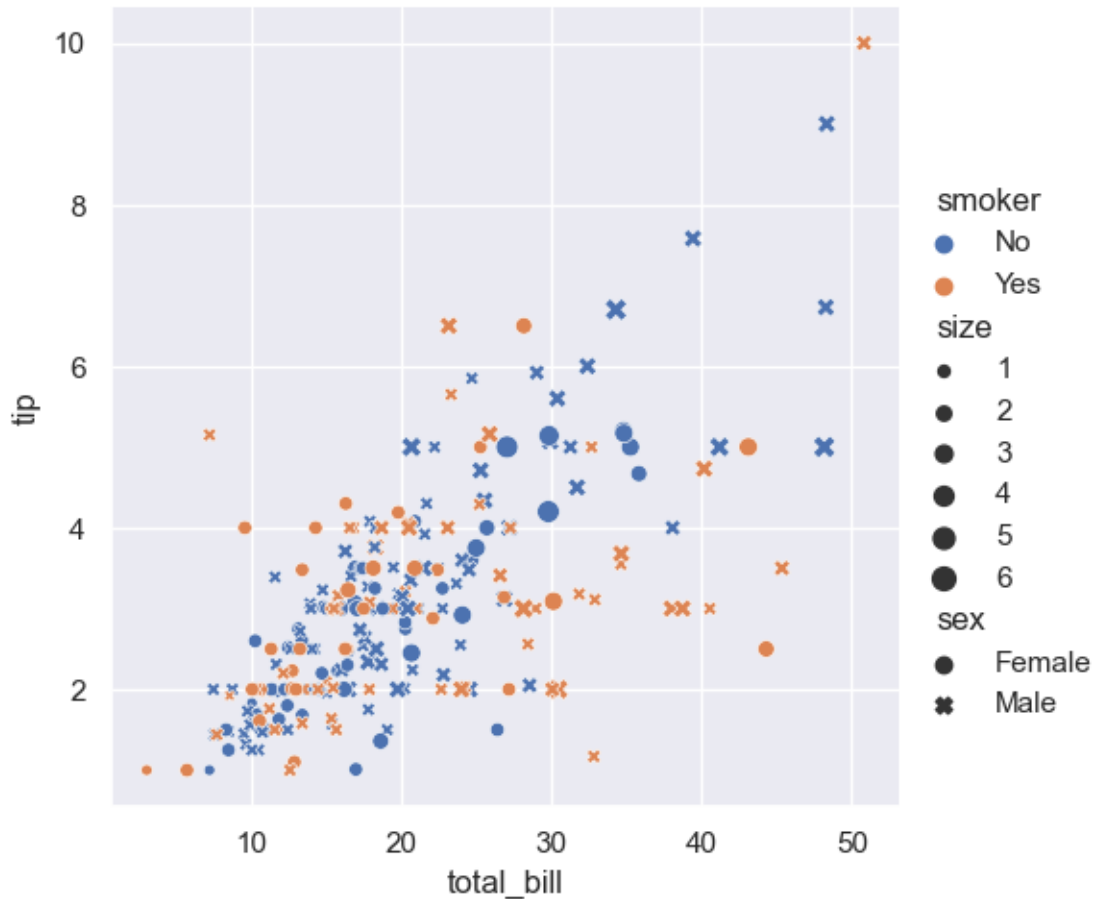
3.4 Visualizing the tips dataset

Let's try a common visual from high school: the scatter plot

```
[2]: # Code 1.2
# Use the Seaborn graphing library
import seaborn as sns

# Hide deprecated warning
import warnings
warnings.filterwarnings('ignore')

# Create the x-y plot
sns.set()
sns.relplot(x='total_bill', y='tip', hue='smoker', style='sex', size='size',
            data=tips);
```



3.5 *But: imagine a large dataset. It would be hard to figure out a dense chart*

3.6 *Let's try other ways of exploring the dataset*

3.7 Exploratory Data Analysis (EDA)

Key points: * John Tukey's book, *Exploratory Data Analysis* (1977) * As a balance to confirmatory data analysis (start from a hypothesis and test it) * Contrast: EDA means no hypothesis, just inspect the data * Primary method used in data mining

Goals of EDA:

- Discover unexpected features, patterns, and anomalies in the dataset
- Provide clues on what or where to investigate further
- Start the seeds of a hypothesis that could lead to major discoveries

See [Wikipedia](#)

The way we explore data is by trying to *grasp the "shape" of the data*. The method to do this may involve looking at the following: * Centrality of data versus the spread of values * How data tend to form clusters * Frequently occurring values * Outliers and how they affect the dataset

The pandas library contains commands that ingest the data and output basic statistical measurements of centrality.

```
[3]: # Code 1.3
tips.describe()
```

```
[3]:      total_bill      tip      size
count  244.000000  244.000000  244.000000
mean    19.785943    2.998279    2.569672
std      8.902412    1.383638    0.951100
min      3.070000    1.000000    1.000000
25%     13.347500    2.000000    2.000000
50%     17.795000    2.900000    2.000000
75%     24.127500    3.562500    3.000000
max     50.810000   10.000000    6.000000
```

4 Pop quiz:

1. What is the typical tip amount?
2. How many records are in the dataset?
3. What was the highest tip given?
4. The lowest tip?
5. What do the values in 25%, 50%, 75% mean?

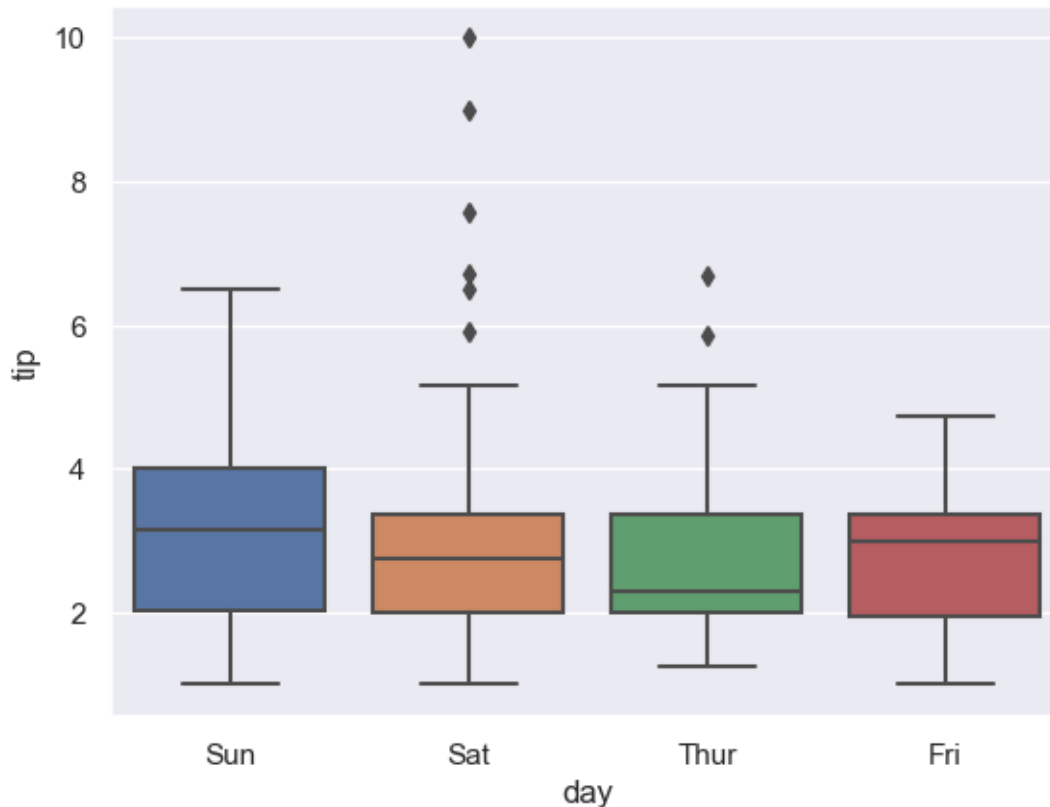
5 Exploring through IQR

The 25%, 50%, and 75% values are known as the **Interquartile Range (IQR)**. * How data is spread out * Where the middle 50% of the data is, in relation to the dataset * Best visualized as a **box plot**

5.1 Visualize the IQR through boxplots

```
[4]: # Code 1.4
# Use the Seaborn graphing library
import seaborn as sns

# Create the visualization
sns.boxplot(x="day", y="tip", data=tips);
```

5.2 Breakout: Analyzing the Boxplots

5 min

1. Which day has the most outliers?
2. Where is the median for each day?
3. What is interesting about the spread of values in Thur and Fri?
4. What is the highest tip amount?
5. What is the median tip amount?

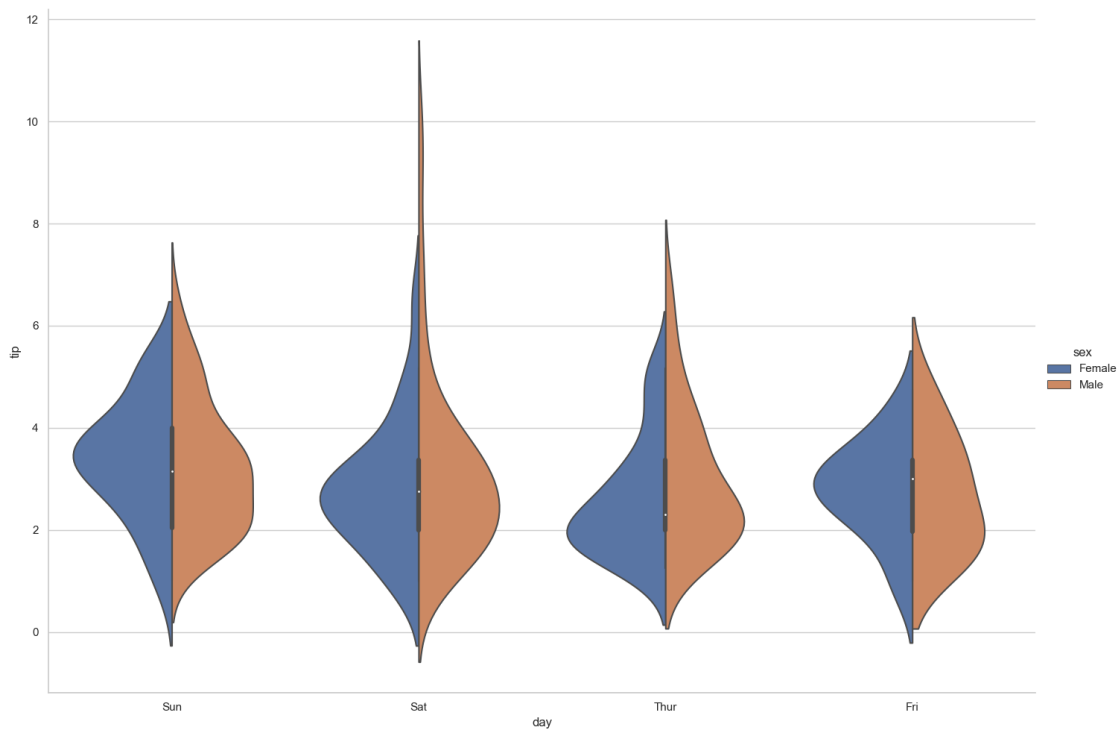
5.3 Insights

1. Use of color to make numbers stand out
2. Font formatting (eg, bold, italics, etc)
3. Use of shape (bar) adds a new dimension: length to represent count
4. Sorting facilitates comparison
5. Dimensions increase the richness or depth of information (eg., time, gender, location)

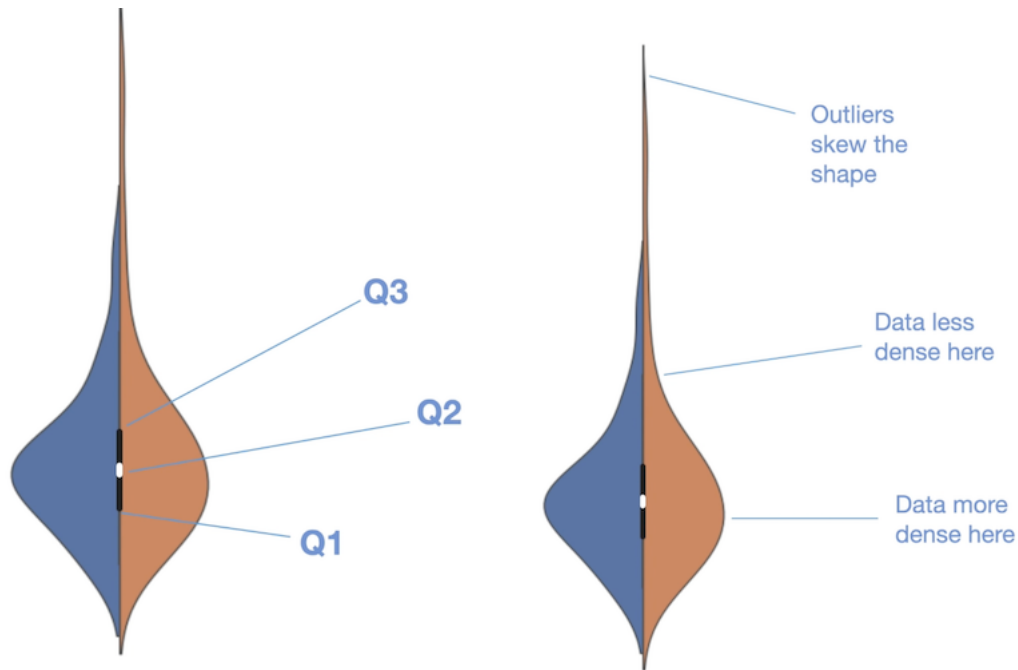
5.4 Violin plots and more shapeful datasets

The graph below is the same dataset, but now using violin plots.

```
[6]: # Code 1.5
# Set up the graph in Seaborn
sns.set_style("whitegrid")
# Graph the violin plot with the specs in the paranthesis
sns.catplot(x="day", y="tip", hue="sex", kind="violin", split=True, height=10,
↪ aspect=11.7/8.27, data=tips);
```



Violin plots, explained:



5.5 Breakout: Visual Investigation

(5 min)

Take screenshots of the images above and answer the following in your group:

1. What dimensions do you see in this graph?
2. When do waiters get the most tips?
3. Which gender tips the most?
4. What is the median tip amount for Fri compared to Thur?
5. What information can you extract from the shape densities?

5.6 Breakout: Pair Share

5 min

Pick one question below and take turns sharing about the question you picked. **Time limit: 1 minute per speaker.**

QUESTIONS (Pick only one)

1. What is 1 important idea that you will use at work?
2. What is 1 idea you would like to investigate further?
3. What is 1 question I still have about data visualization?

6 Conclusion

1. Caveat: EDA is exploratory, to get an initial sense of the dataset.
2. Needs deeper investigation: validate.
3. GIGO.