Introduction to Data Visualization

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

Outline

1. What is it?
2. Why is it important?
3. How to apply it at work?

Why is Data Visualization important?

Let's start with a historical mystery...

The London Cholera Epidemic

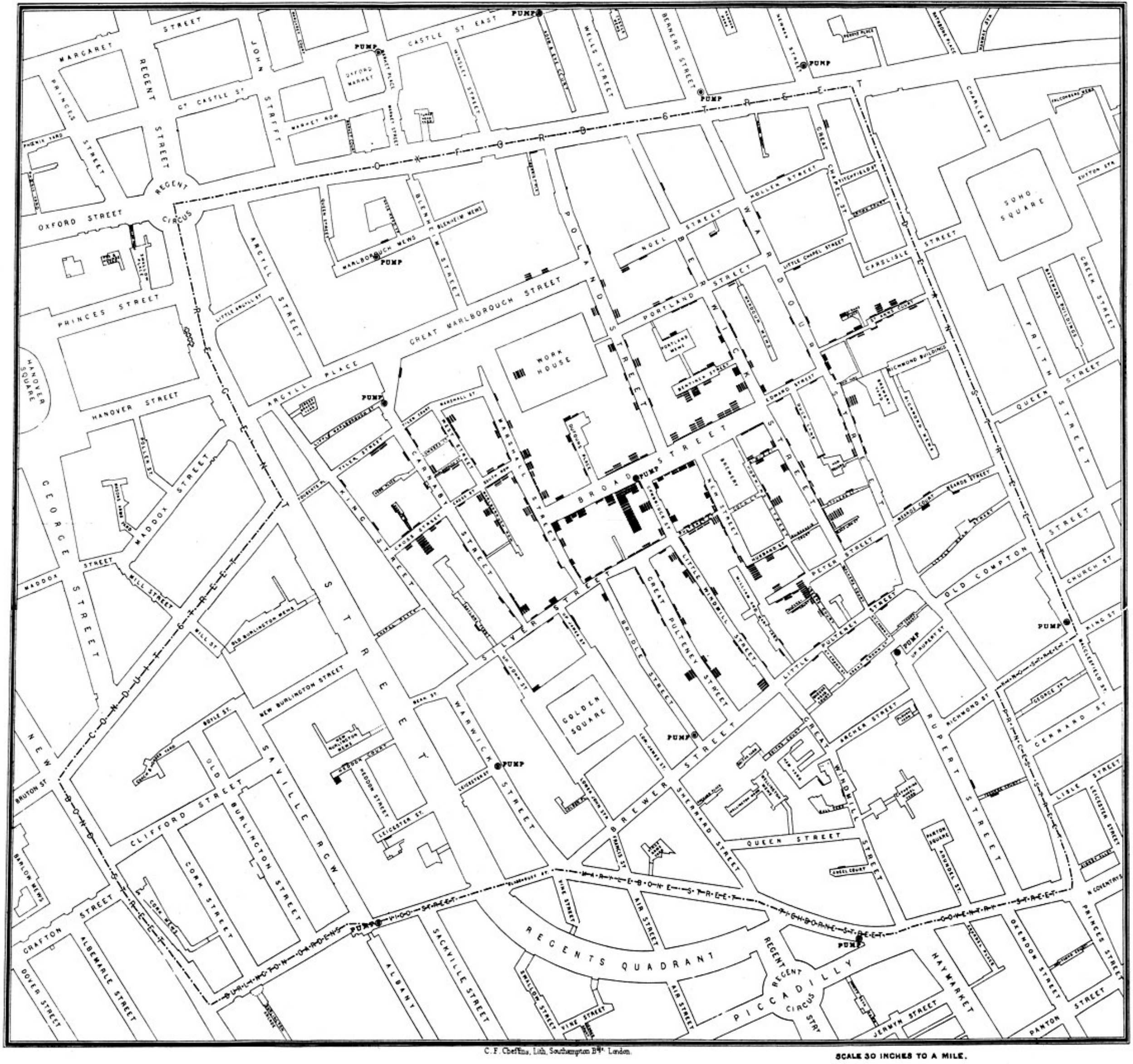
- 1854 cholera outbreak
- Killed 616 people
- Source of infection traced by **Dr. John Snow**
- Detective work with simple (but ingenious) visualization

What did John Snow do?

- Got the addresses of victims
- Got a map and drew one bar per person who died in a location

This is the map that he made...

MAP 1.



C.F. Cheffins, Lith, Southampton Bgs. London.

SCALE 30 INCHES TO A MILE.

- 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.

Breakout Session 1 (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.

Some Sample Insights

Plotting morbidity numbers against the map added a spatial dimension that enriched the information.

Visualizing the numbers against the map made it easier to see patterns from the data.

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

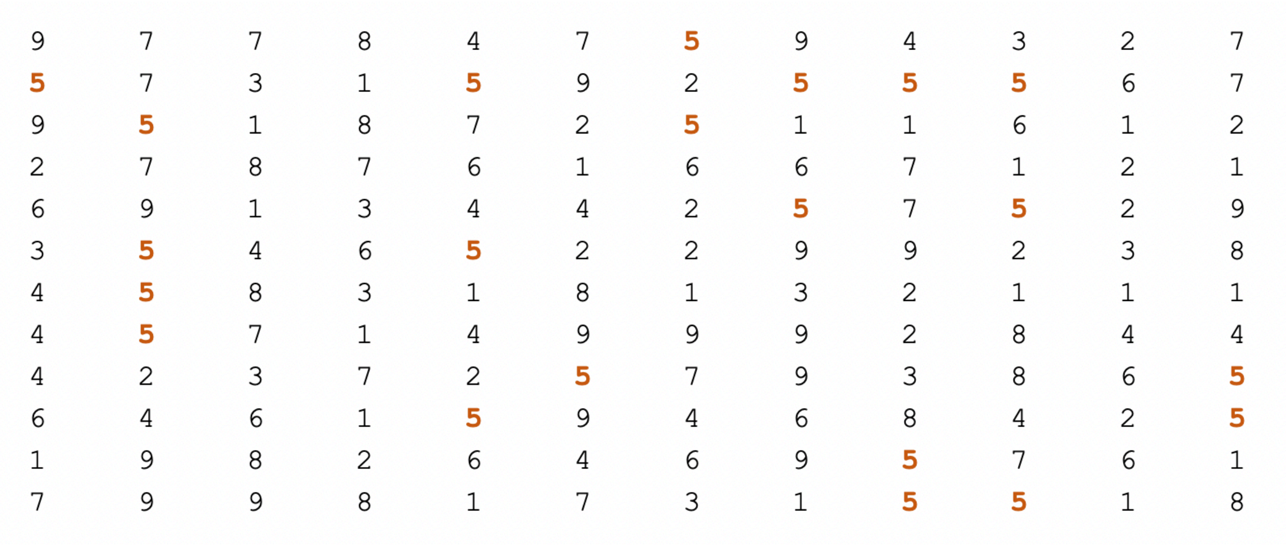
Now answer these questions:

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

Let's tweak the formatting a bit...

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

And which number occurs most frequently?



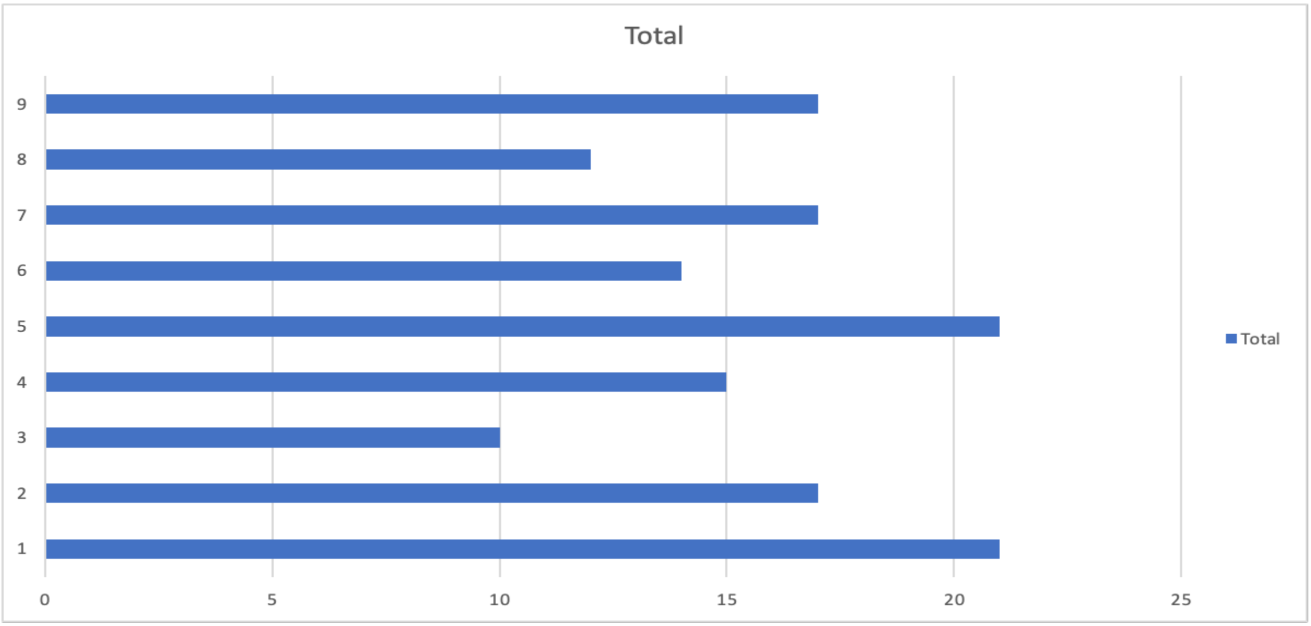
9	7	7	8	4	7	**5**	9	4	3	2	7
5	7	3	1	**5**	9	2	**5**	**5**	**5**	6	7
9	**5**	1	8	7	2	**5**	1	1	6	1	2
2	7	8	7	6	1	6	6	7	1	2	1
6	9	1	3	4	4	2	**5**	7	**5**	2	9
3	**5**	4	6	**5**	2	2	9	9	2	3	8
4	**5**	8	3	1	8	1	3	2	1	1	1
4	**5**	7	1	4	9	9	9	2	8	4	4
4	2	3	7	2	**5**	7	9	3	8	6	**5**
6	4	6	1	**5**	9	4	6	8	4	2	**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

Insights

1. Use of color to make numbers stand out
2. Font formatting (eg, bold, itals, 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?

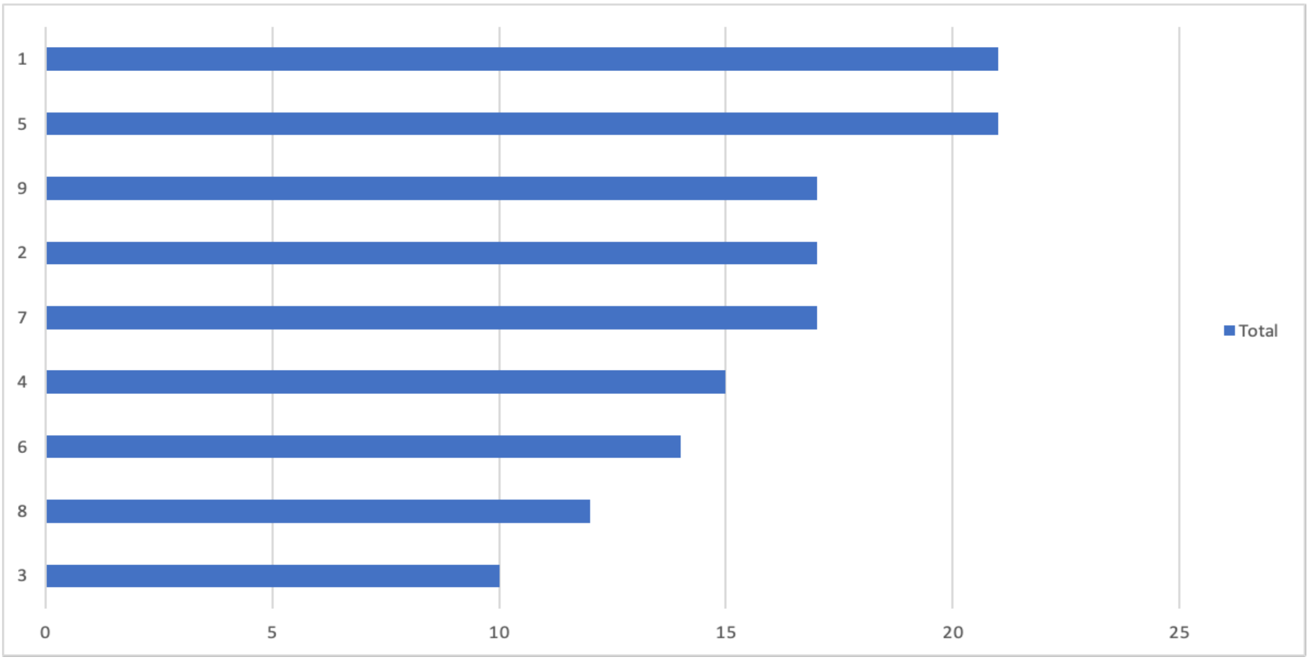
Insights

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

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?

Insights

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

Stretch and Summarize! (2 min)

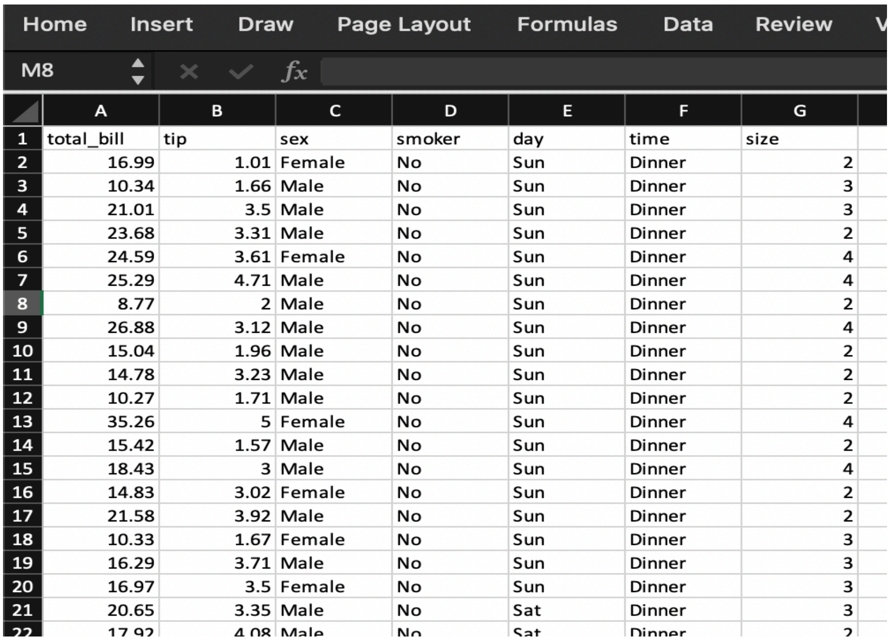
Grab pen and paper and write down a summary of important concepts you've learned.

Let's do more exploration

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

We may call this VEDA: Visual Exploratory Data Analysis.

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

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**.

...

1.1 Exploring the Tips Dataset

Tips Dataset available here: Download tips.csv in Github

```
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 × 7 columns

Stand and stretch

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

1.2 Visualizing the tips dataset

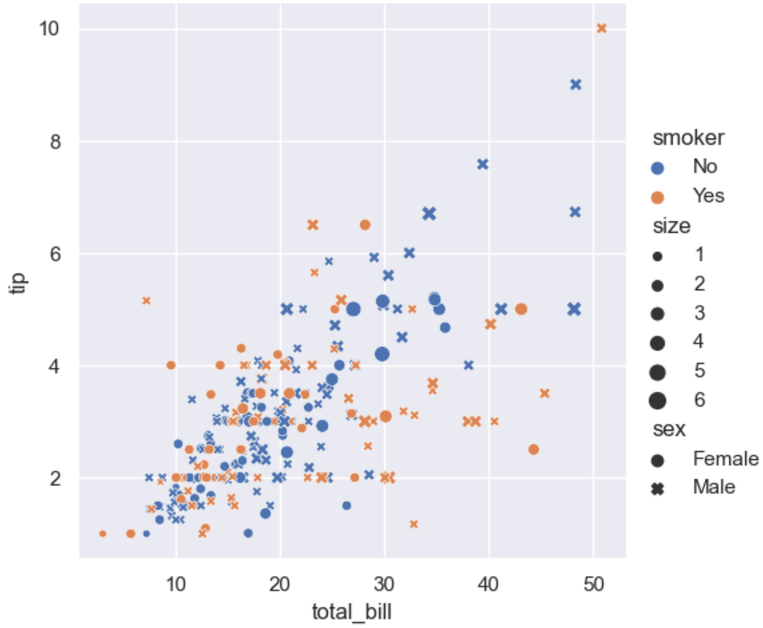
Let's try a common visual from high school: the scatter plot

```
# Use the Seaborn graphing library
import seaborn as sns

# Hide depcrecated 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);
```



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

Let's try other ways of exploring the dataset

1.3 Exploratory Data Analysis (EDA)

- The goal of EDA is to get an idea about the "shape" of the data.
- Shape of data:
  - Centrality and spread
  - How data tends to cluster
  - Outliers

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

```
tips.describe()
```

	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

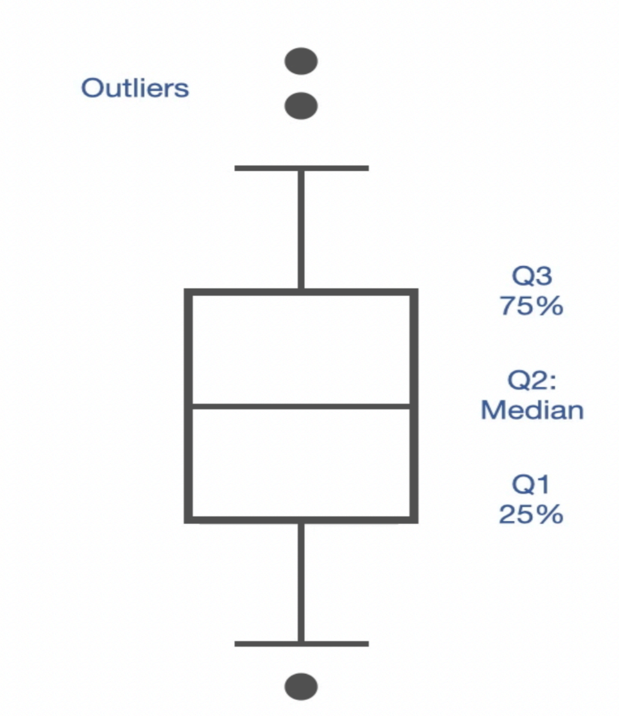
Pop quiz:

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?

Exploring through IQR

The 25%, 50%, and 75% values are known as the **Interquartile Range (IQR)**.

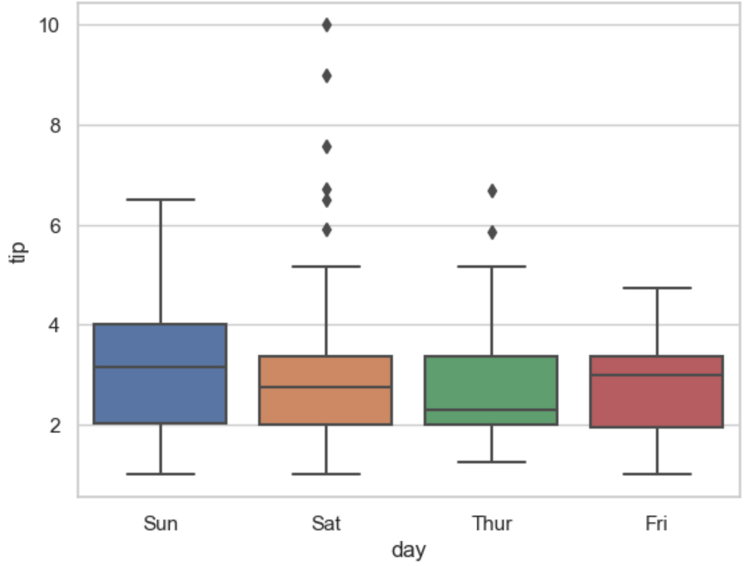
□ How data is spread out
□ Where the middle 50% of the data is, in relation to the dataset
□ Best visualized as a **box plot**



1.4 Visualize the IQR through boxplots

```
# Use the Seaborn graphing library
import seaborn as sns

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



Breakout Session 2 (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?

Insights

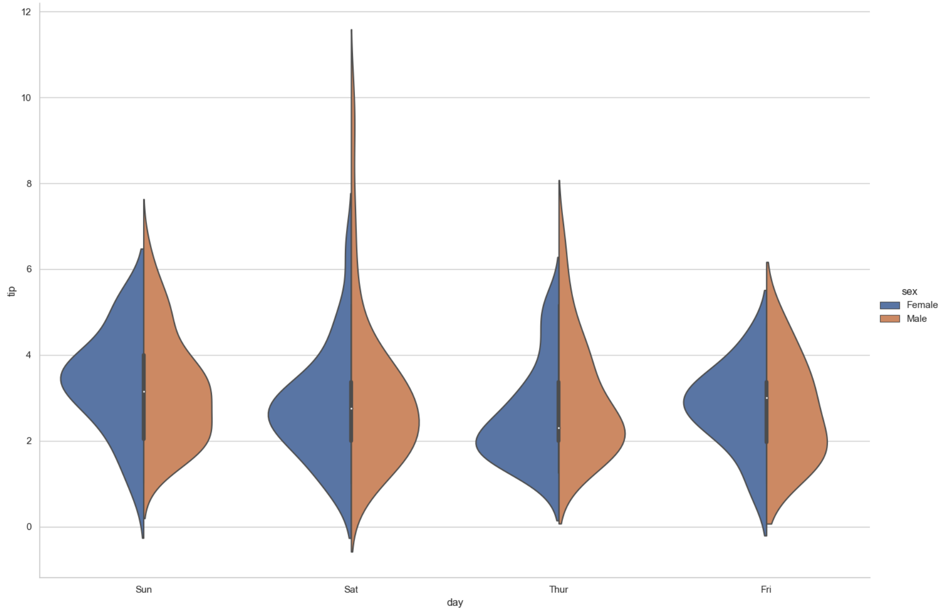
1. Use of color to make numbers stand out
2. Font formatting (eg, bold, itals, etc)
3. Use of shape (bar) adds a new dimension: length to represent count
4. Sorting facilitates comparison
5. Dimensions increase the richness or depth of information (eg., time, gender, location)

1.5 Violin plots and more shapely datasets

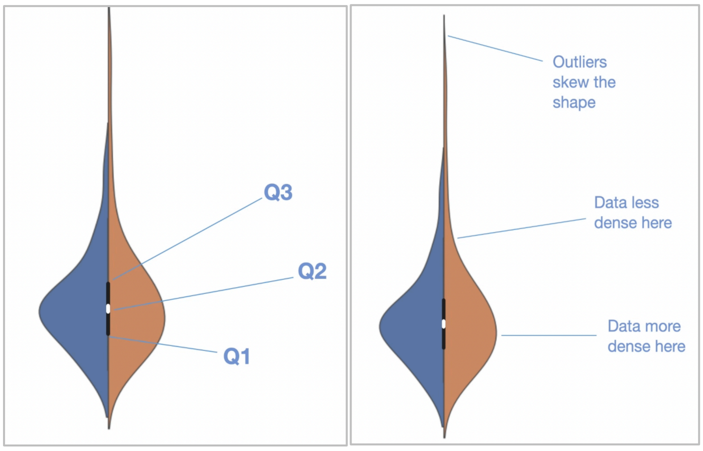
The graph below is the same dataset, but now using violin plots.

In [9]:

```
# 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:



Breakout Session 3 (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?

Breakout Session 4: Pair Share (5 min)

1. Pick one question below.
2. Take turns talking about the question you picked.
3. Time limit: 1 minute per speaker.

QUESTIONS (Pick only one)

1. 1 important idea you will use at work
2. 1 idea you would like to investigate further

Conclusion

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