Introduction to Data Visualization Day 1 - 2023-10

October 14, 2023

By Ruben D. Canlas Jr.

https://www.linkedin.com/in/rubencanlas/

COURSE OUTLINE

- Day 1: Data Visualization Basics
- Day 2: Storytelling with Data

Today: Data Visualization Basics

1. Why is data visualization important?
2. What is it?
3. How do we do data visualization?

1 Part 1: Why is data visualization important?

1.1 Warm Up: What do you know about any of these topics?

Pick one and type your answers in chat.

1. Exploratory Data Analysis (EDA)
2. How visualization helps create informed decisions
3. Mean or median
4. The interquartile range (IQR)
5. Box plots or violin plots

Type your answer in the chat box.

1.2 The London Cholera Outbreak of 1854

- Killed 616 people
- Source of infection found and stopped by **Dr. John Snow**
- Using detective work and visualization
- Started the discipline of epidemiology

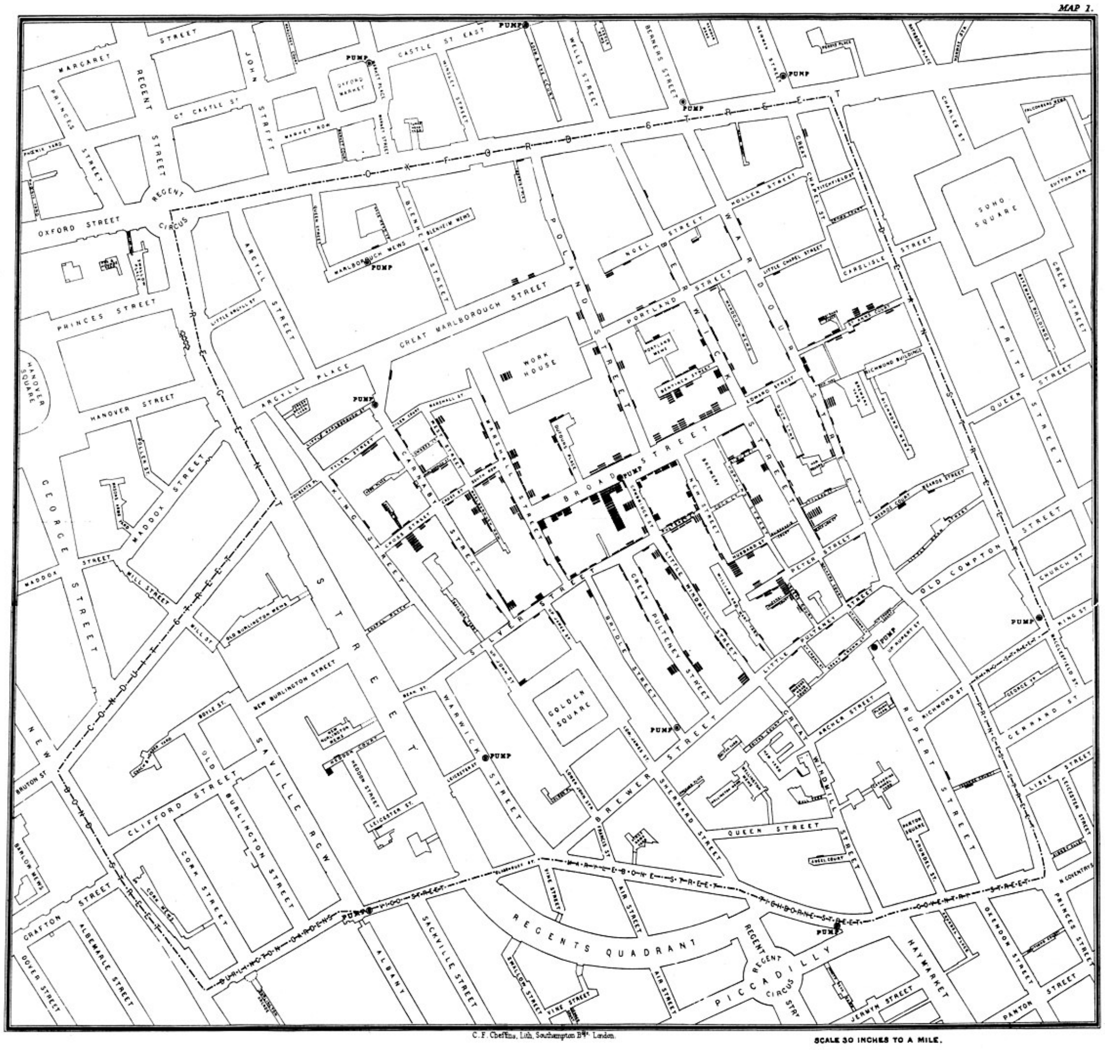
What did Dr. John Snow do?

1. Got the addresses of the victims
2. Got a map of London

3. Marked deaths in a location using stacks of rectangles

This is the map that he made.

...



1.3 Notice these interesting features:

- At the center of the map, see the tallest black stack of bars (Broad Street).
- 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.
- Dr. Snow suspected that water was the source of the outbreak. The plot supported this hypothesis.
- Actionable insight: try shutting down the pump. (It worked.)

1.4 Chat Storm: Write 1 sentence summarizing how visualization helped solve the cholera outbreak.

Type your answer in the chat box.

1.5 Some Insights

1. Clustering morbidity data based on the address of the victim and visualizing this made the patterns stand out.
2. Placing morbidity numbers on the map added a spatial dimension that further enriched the information.
3. The visualization created an insight (potential source of the epidemic) that was actionable (seal off the water pump).

---

2 Part 2: What is data visualization?

A simple definition of data visualization: The use of charts, diagrams, pictures and any visual element to represent information.

Visual elements include:

- Shapes, lines, dots
- Color and shading
- Size
- Symbols or icons
- Maps
- Images

…

2.1 Consider this sample dataset:

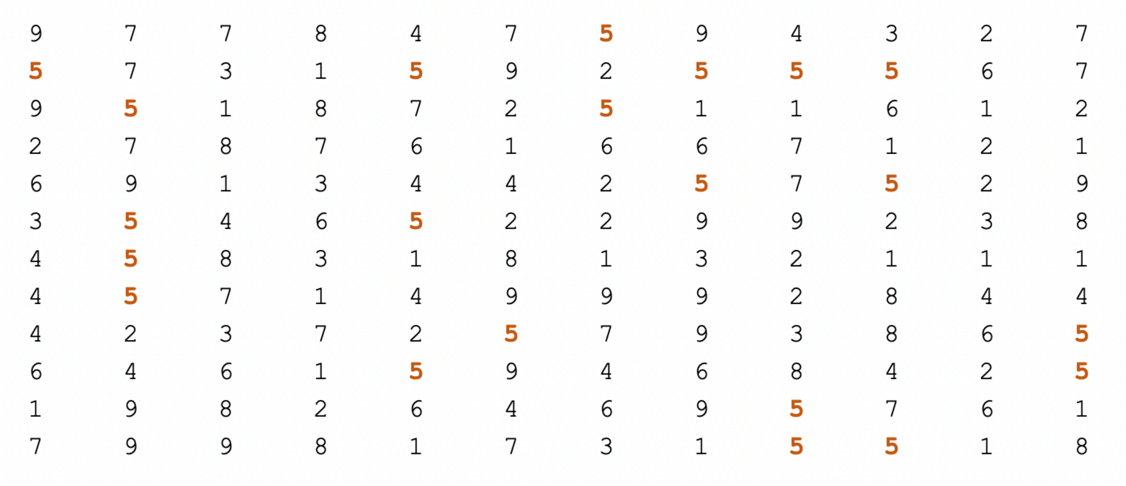
.

```
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
```

2.2 How easy will it be to find the answers to these questions?

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

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

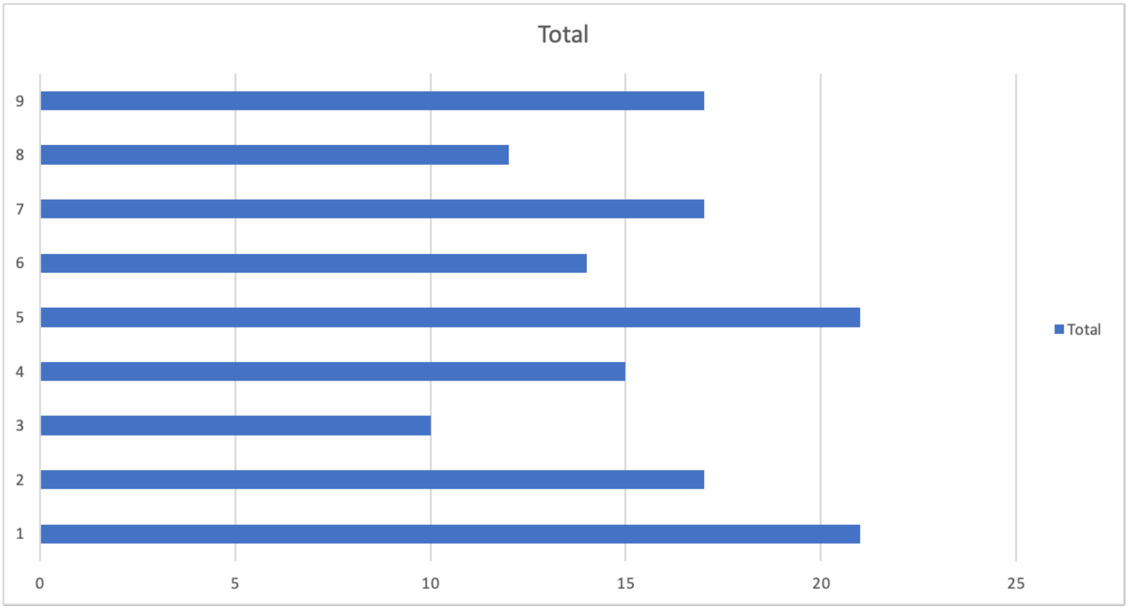
Compared with the previous one, which display makes it easier for you to count the occurence of number 5 or find out which number occurs most frequently?

2.3 Insights

1. Our eyes and brain are programmed to notice differences quickly (survival instinct).
2. Changing the color, for example, will make values stand out.
3. Changing the font formatting can also highlight differences: eg, bold, itals, size, capitalization etc.

Let's make it more interesting.

Same dataset, using bars.



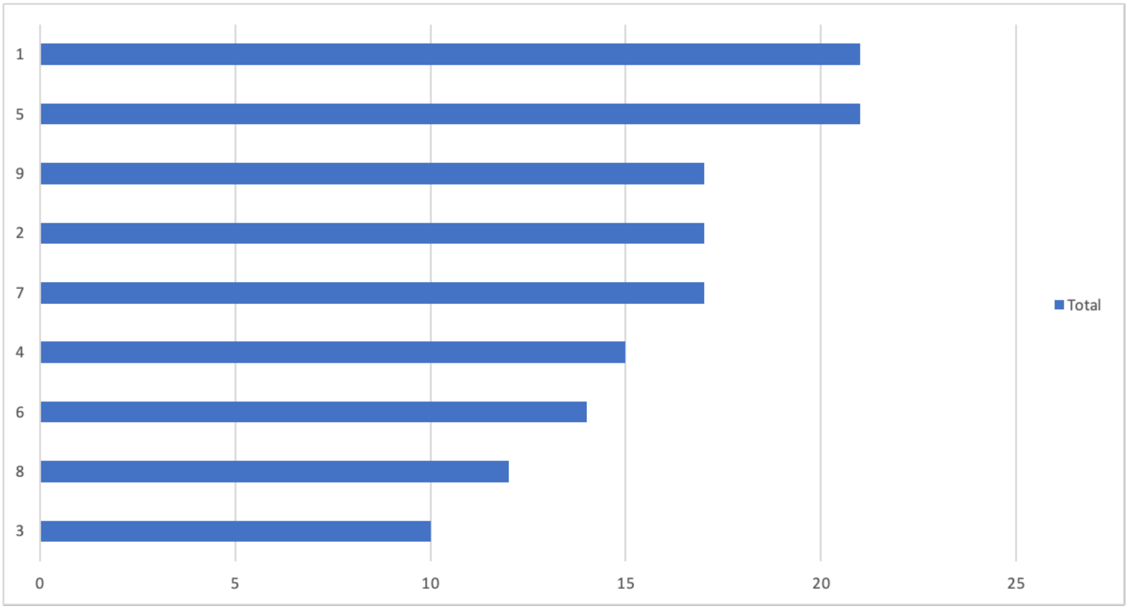
2.4 Questions

1. What information do the x- and y-axes show?
2. How easy is it to find patterns like the most or least frequent numbers?

2.4.1 Insights

- Using a shape (bar) adds a new dimension, length
- Length represents frequency count
- Bars make it easier to spot trends.

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



Now, it's easier to answer these questions:

1. Which number occurs the most?
2. What is the count of the most frequently appearing number?
3. Which number occurs the least?
4. What is the count of the least frequent number?

What made it easier to answer the questions using the second graph?

2.4.2 Insight

- Using a shape (bar) adds a new dimension, length
- Length represents frequency count
- Bars make it easier to spot trends.
- **Sorting from most to least helps make comparisons easier.**

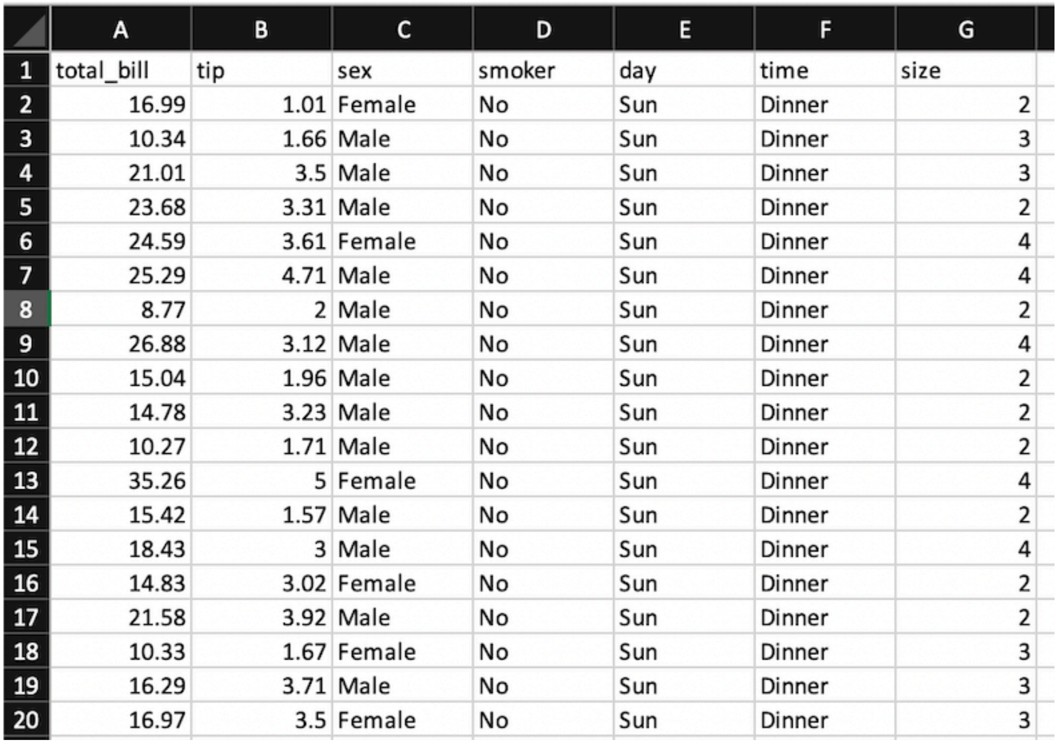
2.4.3 Stretch and Summarize!

2 min

Grab your notebook and summarize the important concepts you've learned in 1-3 sentences.

2.4.4 Introducing the Tips dataset

We will be using the Tips dataset to understand how to do data visualization. Let's get familiar with it first.



	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

2.4.5 Breakout: Exploring the Tips Dataset

5 min

INSTRUCTIONS: * Take a screenshot of the dataset, these instructions, and the questions below. * Join your breakout room. * Stand up, do some stretches and quickly introduce yourselves. * Stay standing while you try to answer the questions, below.

QUESTIONS:

1. Just by looking at the image above, what information can we get from the dataset?
2. Take a guess about what each of the column headers mean.

Note the following: 1. The data is in CSV format and you can find it here: tips.csv in Github 4. Above is a snaphsot of how the CSV file will look like, if opened as a spreadsheet. 5. Jupyter Notebook can also load the CSV, using the `pandas.read_csv` command (demo in Part 3). Read more

---

3 Part 3: How do we do data visualization?

This is a demo of how to do data visualization and visual exploratory analysis, step by step.

- Import the dataset in csv format, using `pandas`.
- Explore the data with `pandas`.
- Create some visuals using `seaborn` visualization library.

3.1 Importing and displaying a dataset using Pandas

Below is a demo of how to read the tips.csv dataset and show a snapshot of it.

[16]:
```
# Code 1.1
# Import the pandas library and name it pd
import pandas as pd

# Using the read_csv() command, read the tips dataset and name it "tips"
tips = pd.read_csv("tips.csv")

# Show the first 5 and last 5 records of the tips dataset
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]
```

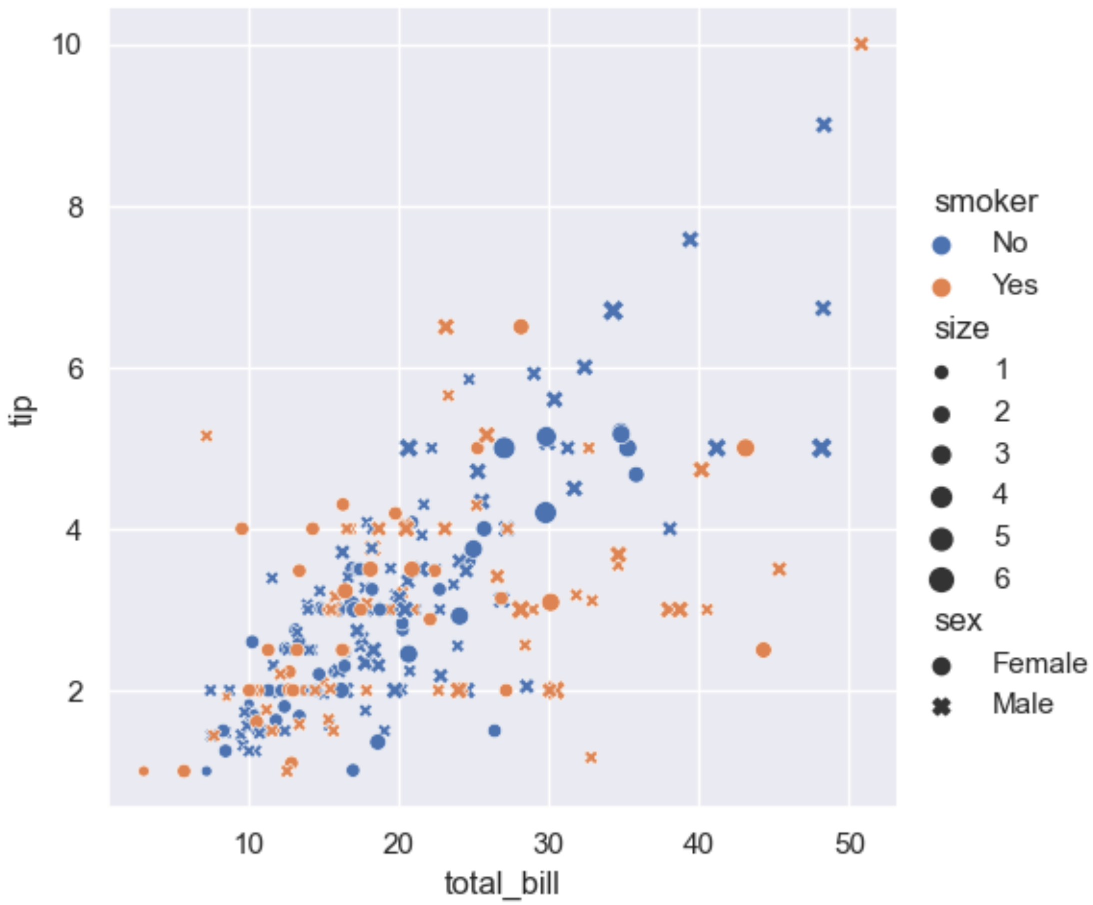
3.1.1 Visualizing the tips dataset

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

[17]:
```
# Code 1.2
# 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);
```



3.2 Insights from the scatter plot

- Shows multiple dimensions in a 2-D plot (aka the columns)
  - Total bill amount
  - Tip amount
  - Gender
  - Smoker or not
  - Number of diners in the table
- By using these elements
  - Color: orange, blue, black
  - Symbols: x, circle
  - Size: varying sizes of x and circle

But: imagine if this were a large dataset. * Overlapping symbols * Bigger shapes will cover others

We need other ways of visualizing the information.

Let's try other ways of exploring the dataset

3.3 Exploratory Data Analysis (EDA)

3.3.1 Key points

- Described in 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

3.3.2 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

3.3.3 How we do the exploration

The way we explore data is by trying to feel around for the "shape" of the data. The method to do this includes investigating these: * Central tendency versus the dispersion of values across the dataset * How data tend to form clusters * Frequently occuring values * Outliers and how they affect the dataset

Let's start exploring some data using statistical summaries.

Below, we create a sample dataset. Let's make a hypothetical subset of the Tips data, for a few minutes of restaurant operations during a Friday.

[18]:

```
# You need to install the pandas library to be able to do this.
# The first command imports the pandas library into an object called "pd".
# pd now contains the functions available to pandas. We invoke the functions␣
 ↪using the . command
# hence, in the example below, pd.DataFrame() invokes the DataFrame function.

import pandas as pd
# BTW, pd is lousy name but it's the convention and you'll find it a lot in␣
 ↪pandas documentation

# Generate sample data. This is just made up data.
data = [1,1,1,2,2,2,2,2,8,9,12]

# Create a pandas DataFrame, store the dataset as "fridaytips", label the␣
 ↪column as "Amounts".
fridaytips = pd.DataFrame(data, columns=['Amounts'])

# Print the sample data we created
fridaytips
```

[18]:

```
   Amounts
0        1
1        1
2        1
3        2
4        2
5        2
6        2
7        8
8        9
9       12
```

A few notes:

- Our dataset is called `fridaytips`.
- Technically, it is no longer just a dataset. It is a pandas `DataFrame`.
- That is, `fridaytips` now possesses useful analytical commands like `.describe()` demonstrated below.

[19]:

```
# The `.describe()` function gives out descriptive statistics to show␣
 ↪centrality and dispersion
# Find out more: https://pandas.pydata.org/pandas-docs/stable/reference/api/
 ↪pandas.DataFrame.describe.html

fridaytips.describe()
```

[19]:

```
         Amounts
count  10.000000
mean    4.000000
std     4.055175
min     1.000000
25%     1.250000
50%     2.000000
75%     6.500000
max    12.000000
```

3.3.4 Visual Exploratory Data Analysis

Bar graphs, scatter plots etc: subset of EDA called **Visual Exploratory Data Analysis (VEDA).**

VEDA uses visuals to facilitate summarizing and reporting of datasets. Dashboards are an example of VEDA.

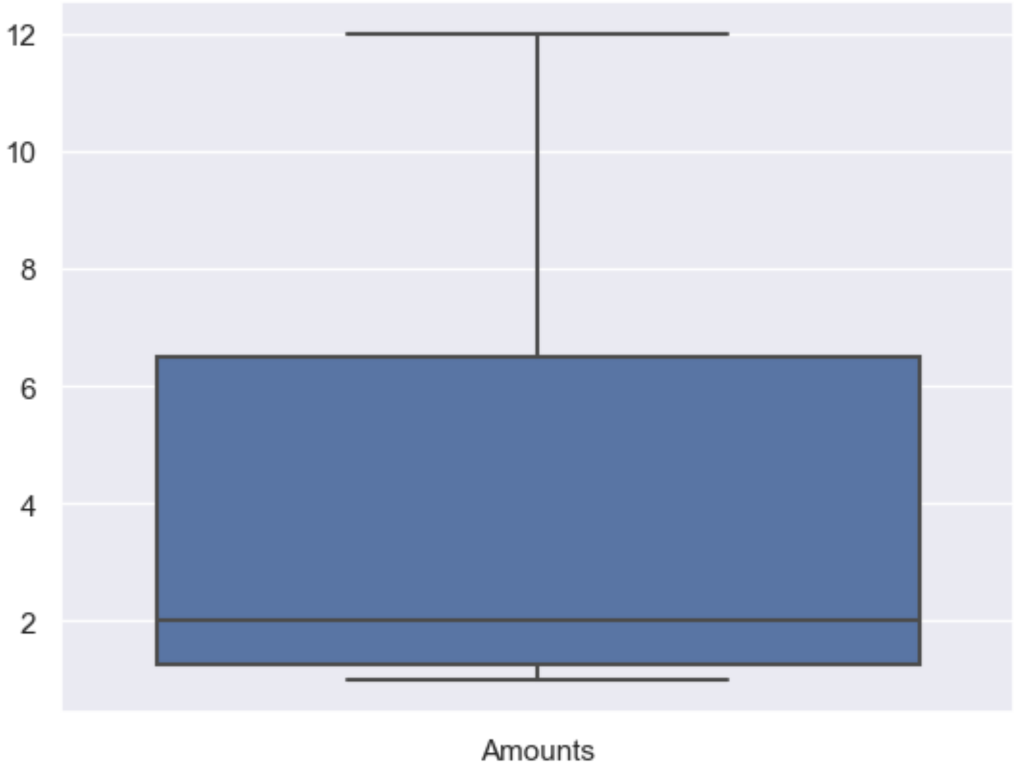
Let's do more VEDA using a visual called box plot.

[20]:

```
# Use the Seaborn graphing library
# Store the library in an object called 'sns'
# BTW, sns is lousy name but it's the convention and you'll find it a lot in␣
 ↪seaborn documentation
```

```
import seaborn as sns

# Create the visualization
sns.boxplot(data=fridaytips);
```



3.4 Breakout

10 min

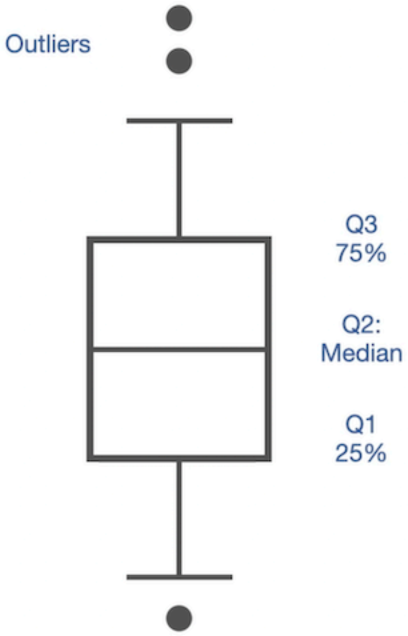
Questions about the Friday tips dataset:

Investigate the box plot by comparing it to the statistical summary generated by `fridaytips.describe()`.

1. How much is the median tip value?
2. Where is the median located in the box plot?
3. What values do the horizontal lines on 1 and 12 indicate?
4. What does the blue box represent?
5. What value does the *upper* edge of the blue box show?
6. What value does the *lower* edge of the blue box show?
7. What causes the blue box to "sink" towards the bottom of the vertical axis?

3.5 Exploring data through IQR

The 25%, 50%, and 75% values are known as the **Interquartile Range (IQR)**. The IQR shows: * How data is spread out * The **median** - if you listed all values in ascending order, find the mid-point of the list. The value at the mid-point is the median. * Best visualized as a **box plot**



3.5.1 Learn more about box plots

Read this: Learn more about box plots, outliers, and the max and min values

[21]:
```
# Code 1.3
tips.describe()
```

[21]:
```
       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
```

3.6 Pop quiz

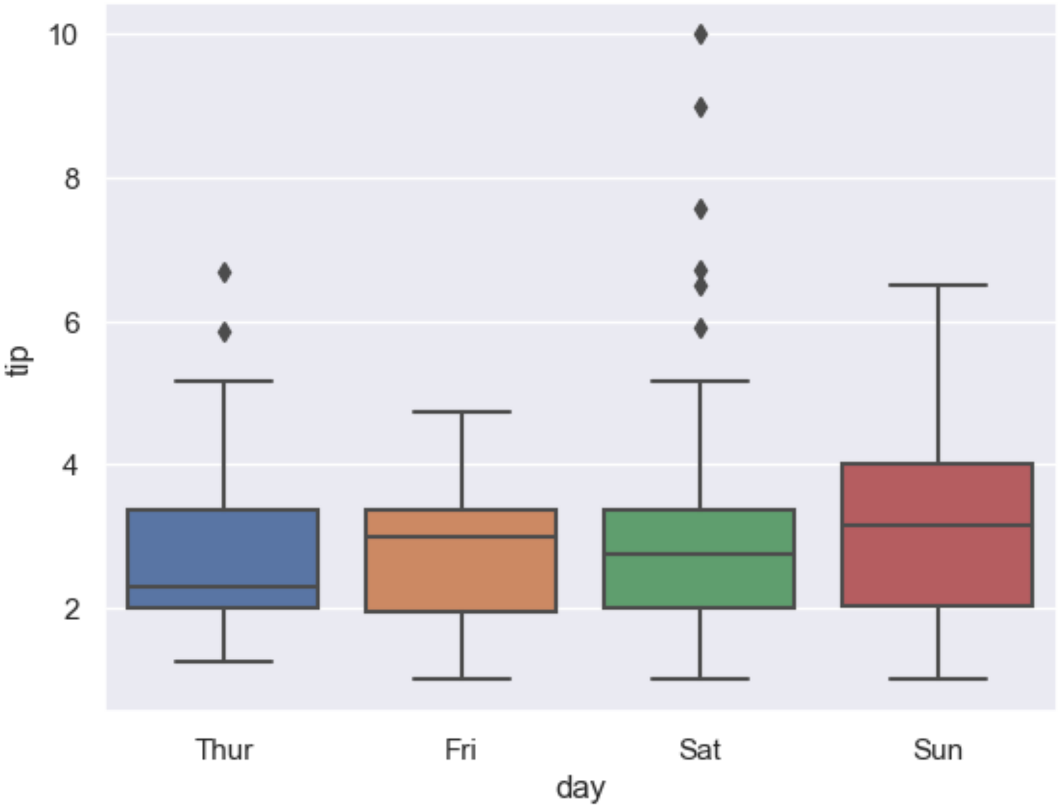
1. What is the typical tip amount?
2. How many records are in the dataset?
3. What was the highest tip given?

4. The lowest tip?
5. What do the values in 25%, 50%, 75% mean?

[22]:

```
# Code 1.4
# Use the Seaborn graphing library
import seaborn as sns

# Create the visualization
sns.boxplot(x="day", y="tip",
            data=tips,
            order=['Thur','Fri', 'Sat', 'Sun']);
```



3.7 Breakout: Analyzing the Boxplots

10 min

1. What data are shown in the the x- and y- axes?
2. What are the highest and lowest tip amounts in a week?
3. Which day/s have the highest median tip value?
4. What is interesting about Thur and Fri?
5. What is different about Sun tips? How might you explain the difference?

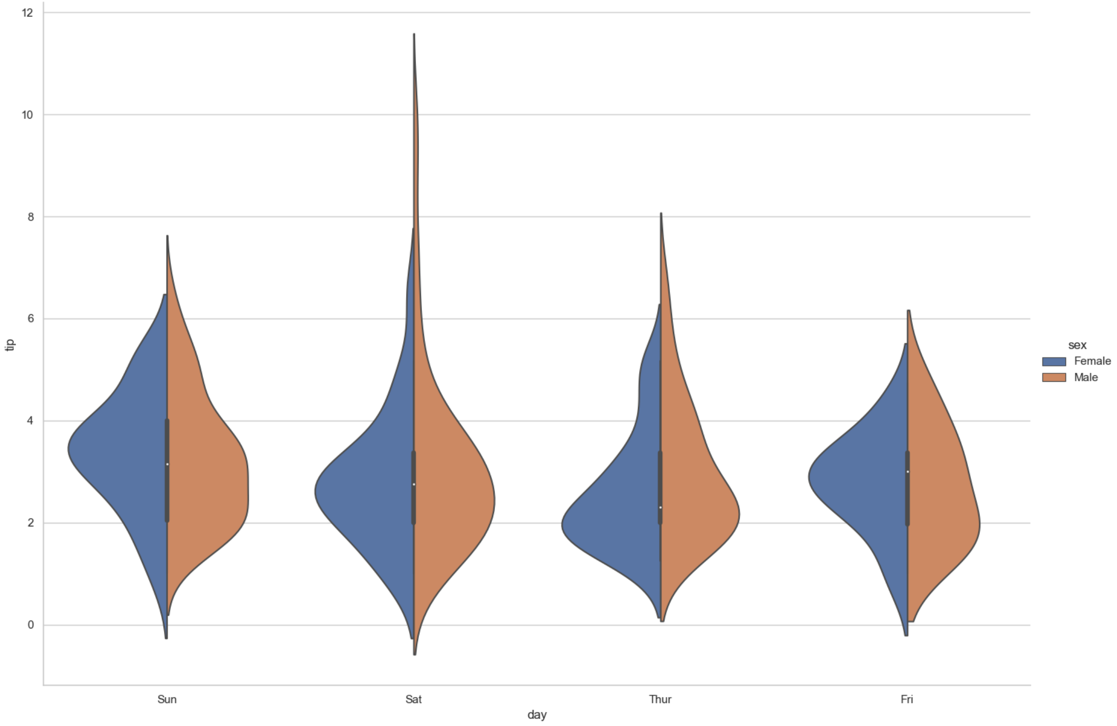
3.8 Violin plots

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

[23]:

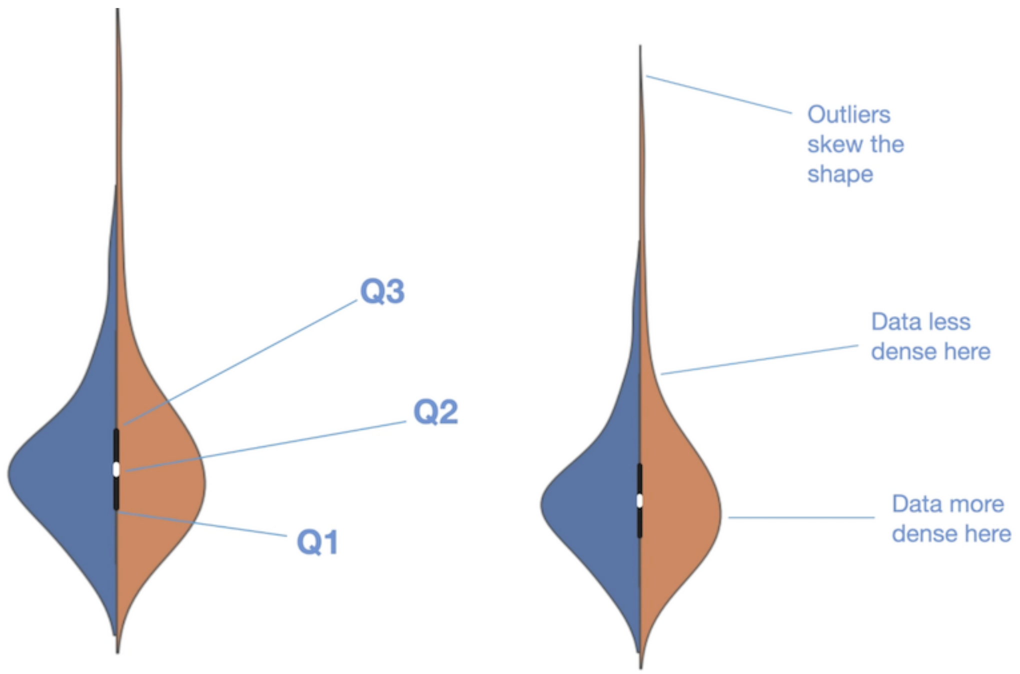
```
# Code 1.5
# Set up the graph in Seaborn
sns.set_style("whitegrid")

# 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);
```



3.9 Violin plots, explained

.



3.10 Breakout: Visual Investigation

5 min

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

1. What dimensions can you get from the violin plot?
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?

Conclusion

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

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?