Data Visualization (Reading: 6.1-6.3)

Learning goals:

- Motivate the importance of data visualization
- Introduce the seaborn and matplotlib libraries
- Understand common chart types

UC Berkeley Data 100 Summer 2019
Sam Lau

(Slides adapted from Deb Nolan, Sandrine Dudoit, & Fernando Perez)

Announcements

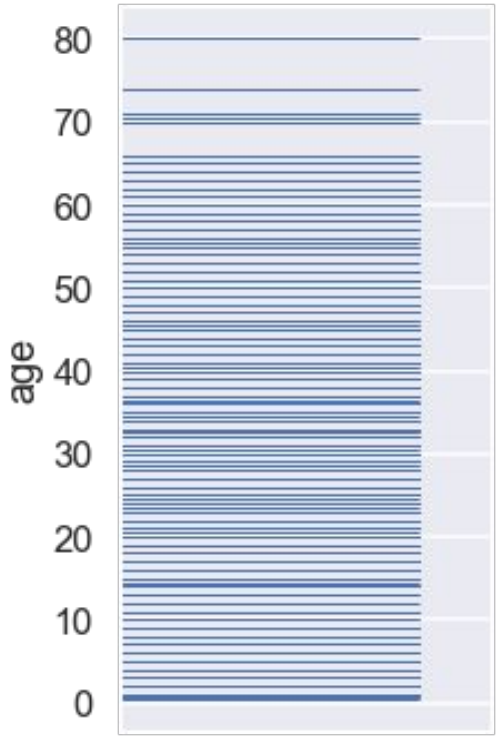
- Small group tutoring is starting this week
  - Sign up: http://bit.ly/d100-tutor
- Project 1 due Tues!
- HW2 out Wednesday
  - Will be “officially” due Friday but we will take submissions without penalty until Tuesday (July 9)
- HW3 out Friday
  - Due the following Friday (July 12)
- Starting this week, I have OH:
  - Mondays 11-12pm and 1-2pm in 355 Evans

What is Data Visualization?

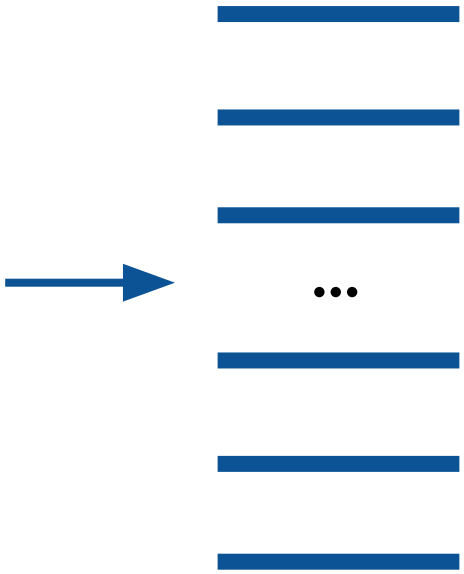
Computer Readable

	age
0	22.0
1	38.0
2	26.0
...	...
888	NaN
889	26.0
890	32.0

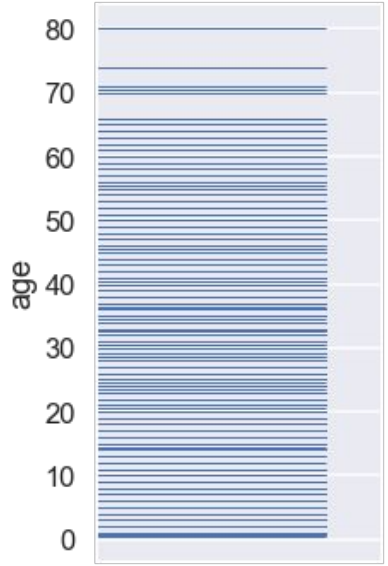
Human Readable



	age
0	22.0
1	38.0
2	26.0
...	...
888	NaN
889	26.0
890	32.0



10px
16px
11px
...
0px
11px
15px



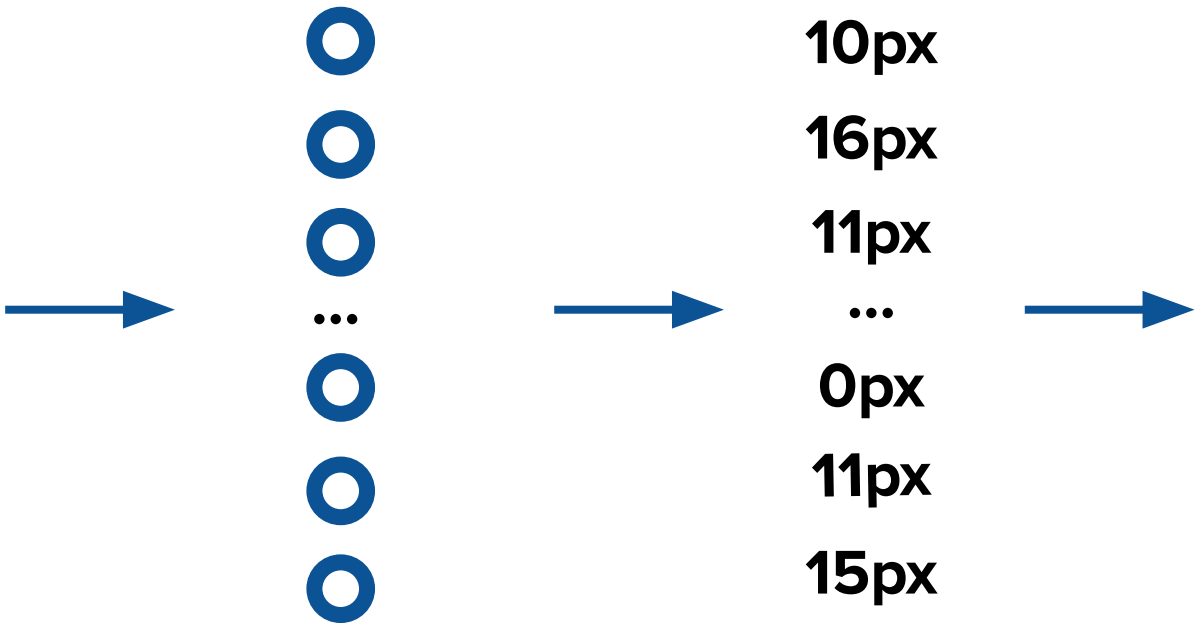
Mark

(Represents a datum)

Encoding

(Maps datum to visual position)

	age
0	22.0
1	38.0
2	26.0
...	...
888	NaN
889	26.0
890	32.0

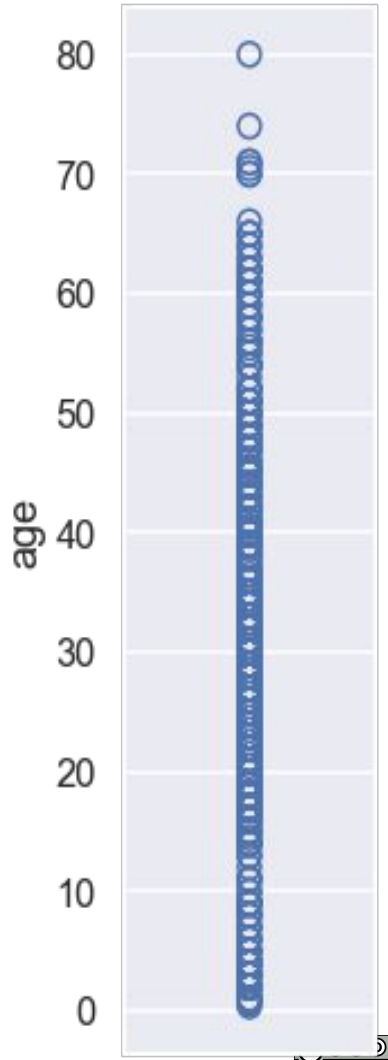


Mark

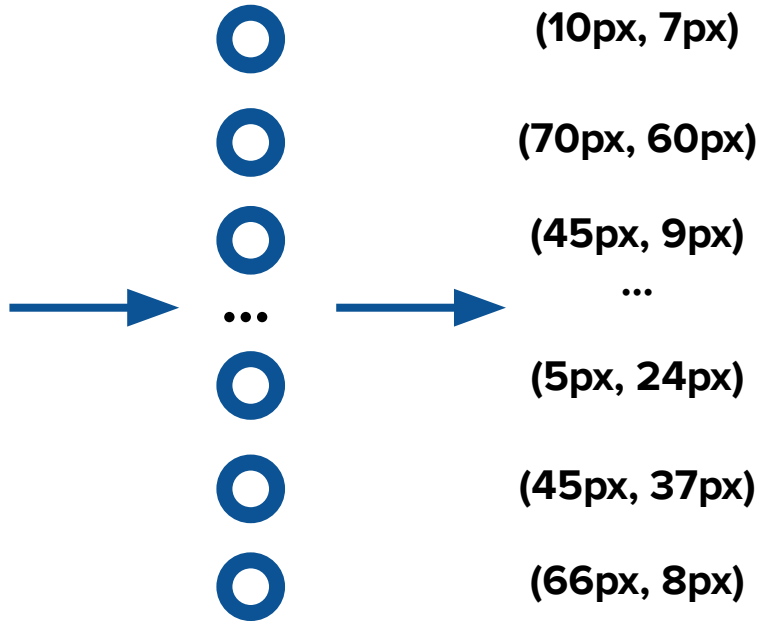
(Represents a datum)

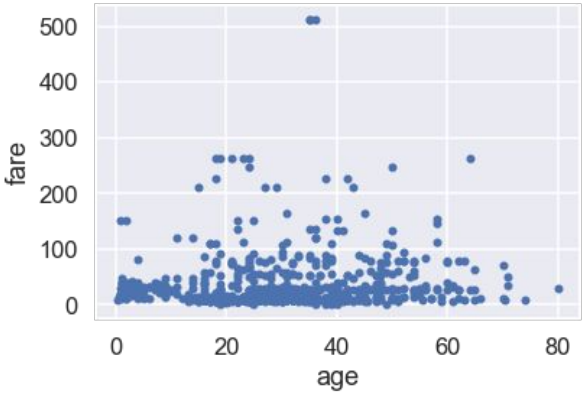
Encoding

(Maps datum to visual position)

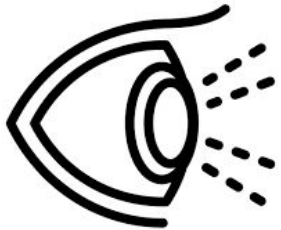


	age	fare
0	22.0	7.25
1	38.0	71.28
2	26.0	7.92
...	...	...
888	NaN	23.45
889	26.0	30.00
890	32.0	7.75

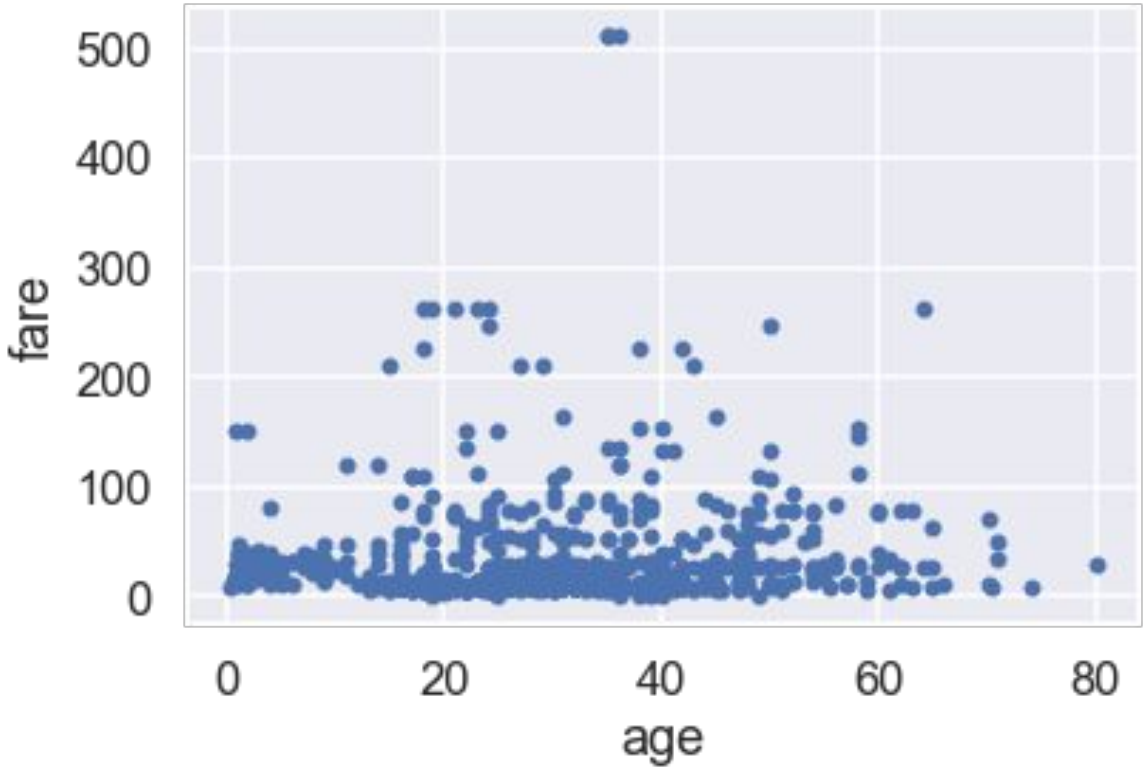




Visualizations are for Humans



“Looks like older people didn’t spend more than younger people.”



Visualizations are for Humans

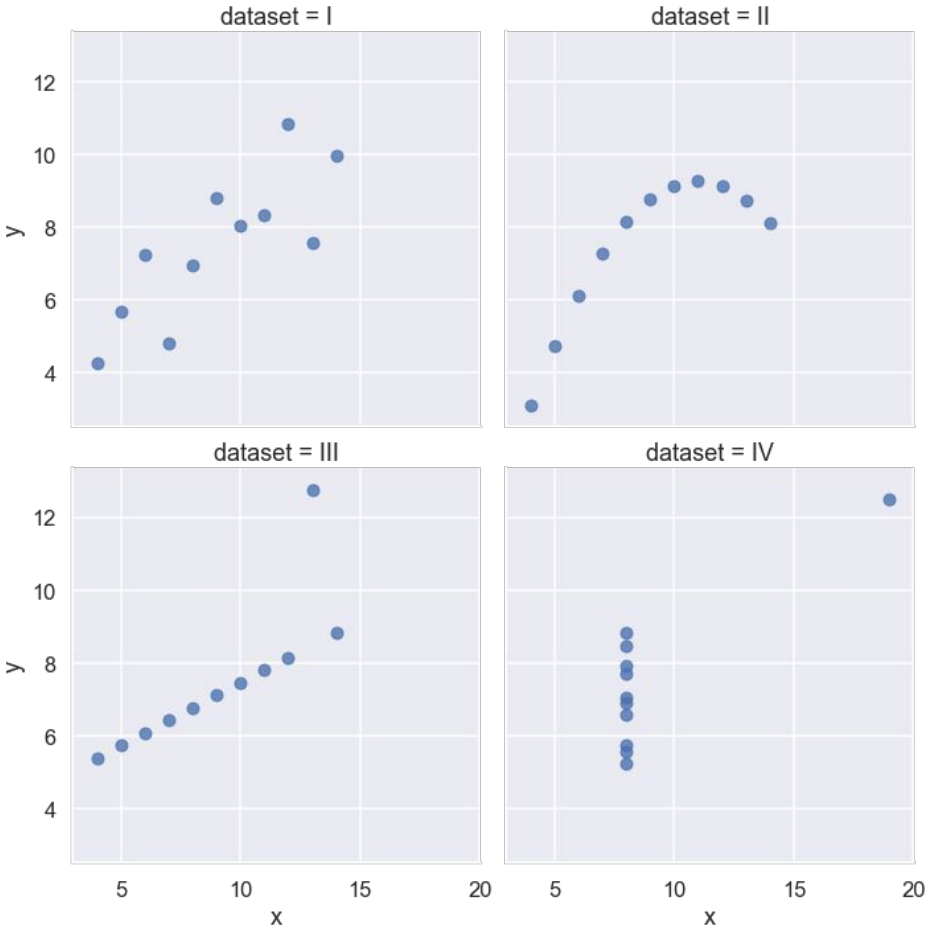
x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

x	y
10.0	9.14
8.0	8.14
13.0	8.74
9.0	8.77
11.0	9.26
14.0	8.10
6.0	6.13
4.0	3.10
12.0	9.13
7.0	7.26
5.0	4.74

x	y
10.0	7.46
8.0	6.77
13.0	12.74
9.0	7.11
11.0	7.81
14.0	8.84
6.0	6.08
4.0	5.39
12.0	8.15
7.0	6.42
5.0	5.73

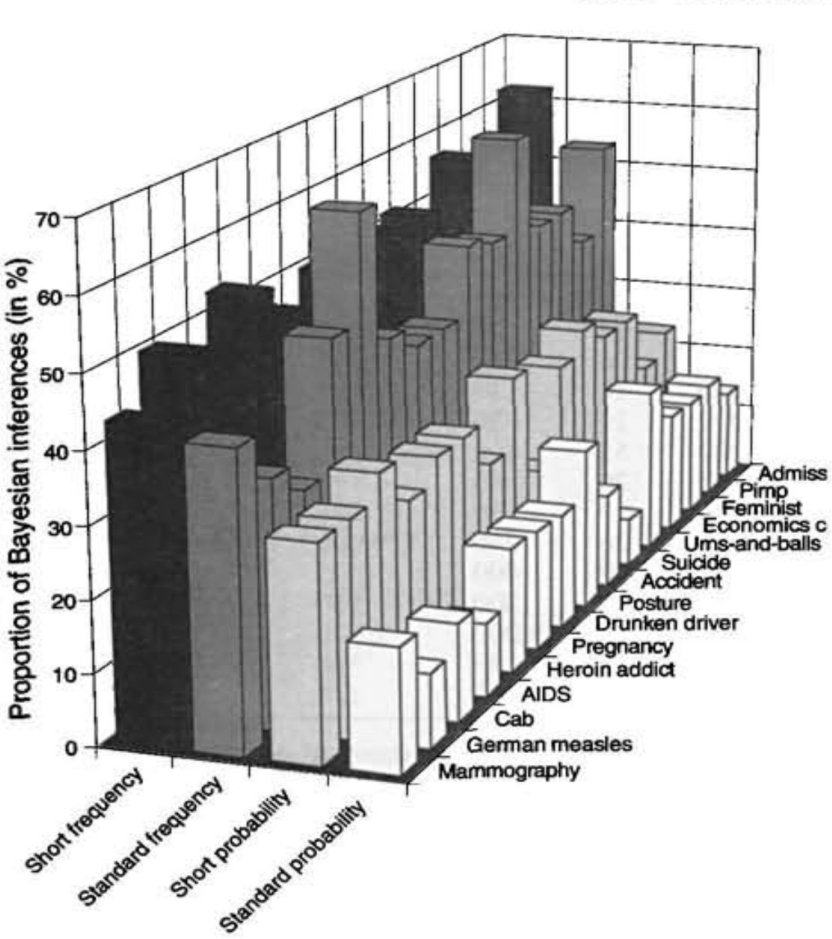
x	y
8.0	6.58
8.0	5.76
8.0	7.71
8.0	8.84
8.0	8.47
8.0	7.04
8.0	5.25
19.0	12.50
8.0	5.56
8.0	7.91
8.0	6.89

Visualizations are for Humans



Human eyes good at seeing visual patterns!

Visualizations are for Humans



Human eyes good at seeing visual patterns!...

Sometimes.

Why Data Visualization?

- One goal of data science is to inform human decisions
  - Excellent plots **directly** address this goal
  - Sometimes the most useful results from data analysis are the visualizations!
- Data viz is not as simple as calling plot()
  - Many plots possible, but only a few are useful
  - Every visualization has tradeoffs

seaborn

(Demo)

seaborn

Best used with tidy (aka long-form) data.

- Seaborn will perform groupby automatically

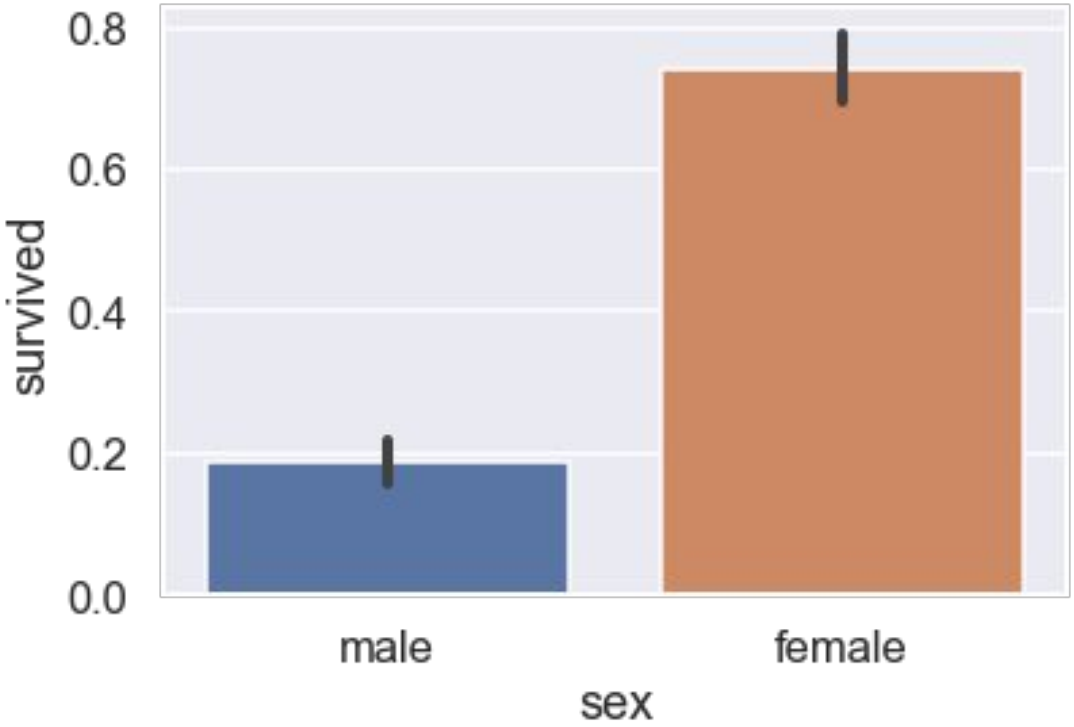
Typical usage:

```
sns.someplot(x='...', y='...', data=...)
```

seaborn

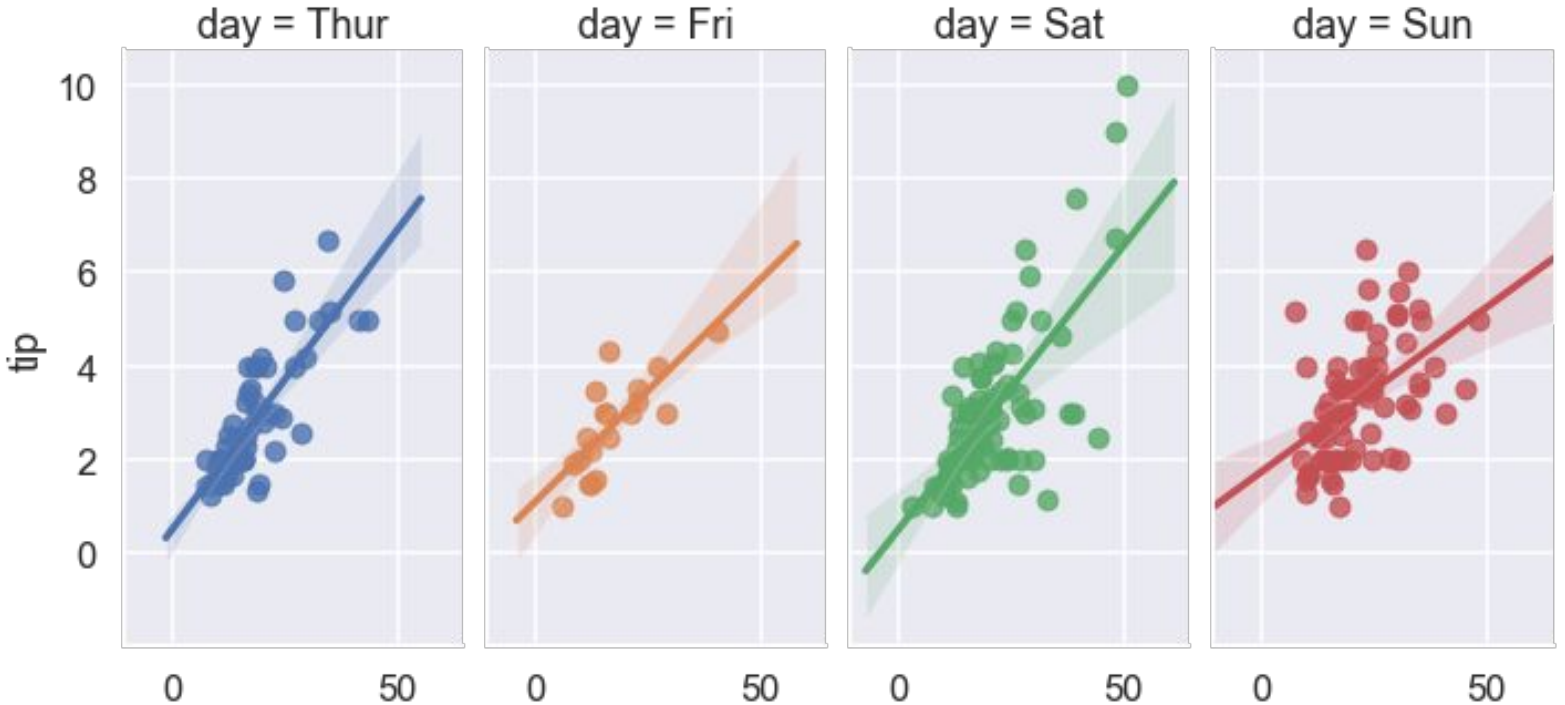
```
sns.barplot(x='sex', y='survived', data=ti)
```

	survived	class	sex	age	fare
0	0	Third	male	22.0	7.25
1	1	First	female	38.0	71.28
2	1	Third	female	26.0	7.92
...	...	...	...	...	...
888	0	Third	female	NaN	23.45
889	1	First	male	26.0	30.00
890	0	Third	male	32.0	7.75



seaborn

```
sns.lmplot(x="total_bill", y="tip",
           col="day", hue="day", data=tips)
```



Break!
Fill out Attendance:
http://bit.ly/at-d100

Customizing Plots using matplotlib

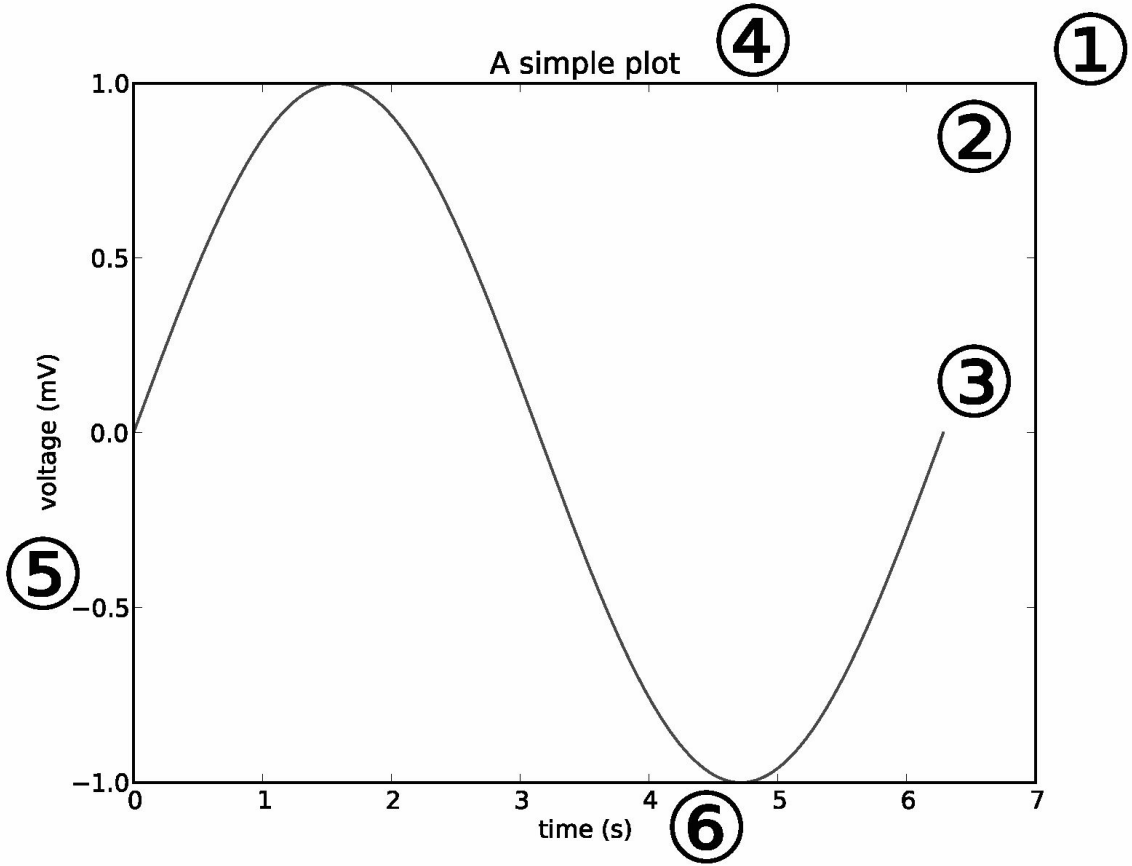
(Demo)

matplotlib

- Underlying library for seaborn, pandas, and most other Python plotting libraries
- A Figure contains several Axes. Each Axes contains a plot.
- When creating a plot, a new figure + axes is created if not already initialized.
  - Matplotlib remembers that axes for the duration of the cell (hidden state!)
- Note: Axes = one chart within a larger Figure
  - Axis = x or y-axis within a chart (sorry!)

matplotlib

1. Figure
2. Axes
3. Line
4. Title
5. YAxis
6. XAxis



Typical Workflow

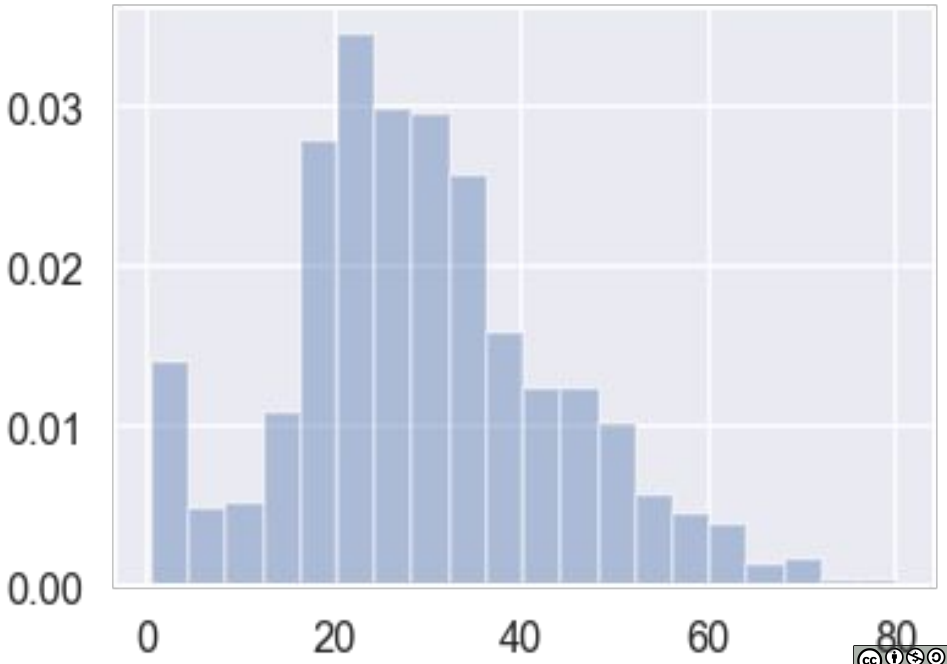
- Start with seaborn plot
  - Get as close to desired result as possible
- Fine-tune with matplotlib, e.g:
  - Changing title, axis labels
  - Annotating interesting points
- Publication-ready plots take lots of fine-tuning!

Common Visualizations for One Quantitative Variable

Histograms

Always have proportion per unit on y-axis

- Same as in Data 8
- Total area = 1
- Deciding on number of bins is hard!
  Trial-and-error process.



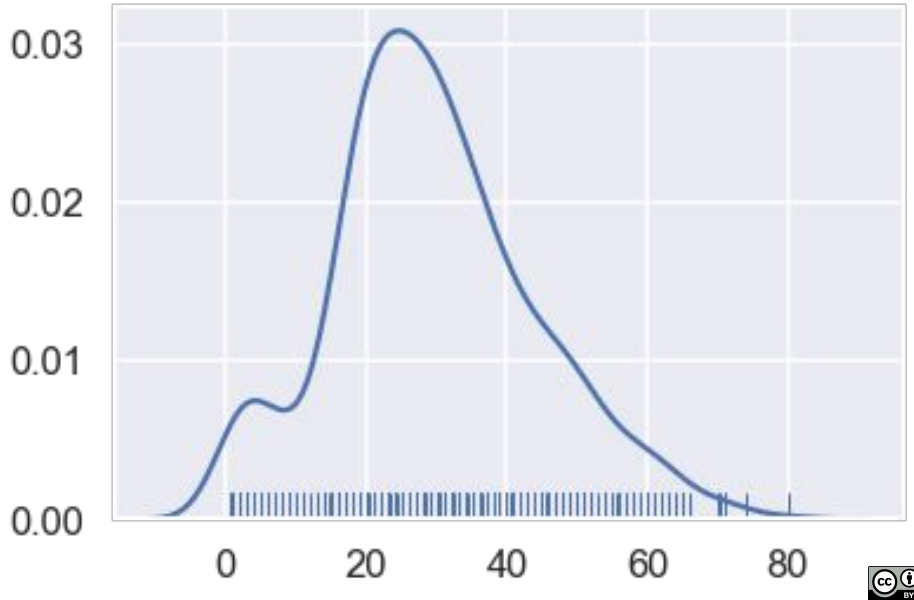
Density Plots

Density plots similar to a “smoothed” histogram

- More on smoothing tomorrow

Rug plots put a tick at each data point

- Used to show all points

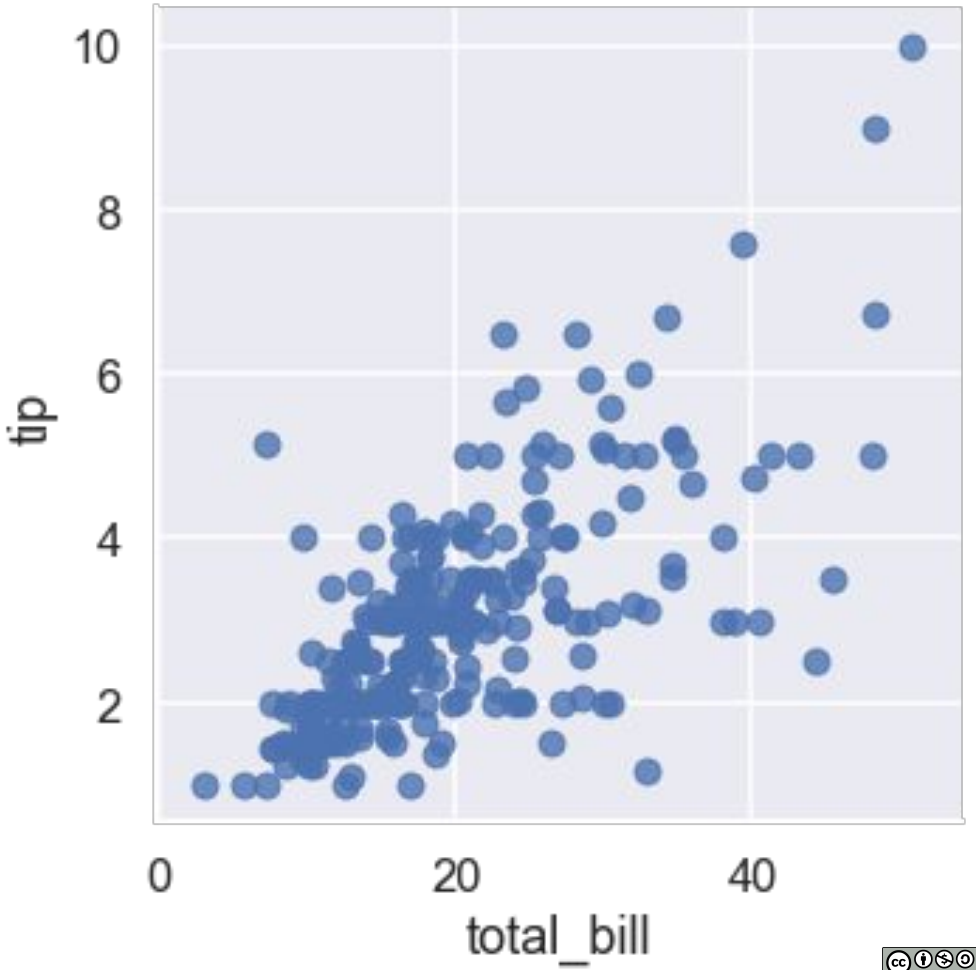


Common Visualizations for Two Quantitative Variables

Scatter Plots

Used to reveal relationships between pair of variables

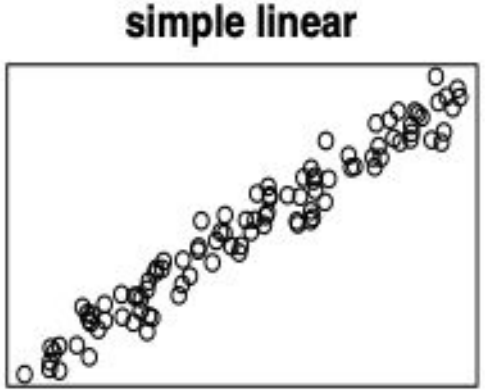
- Susceptible to *overplotting*
  - Points overlap!
- More discussion tomorrow

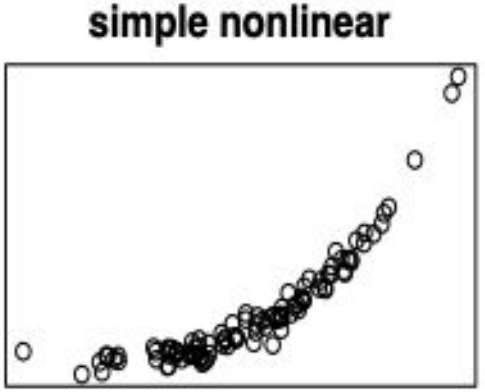


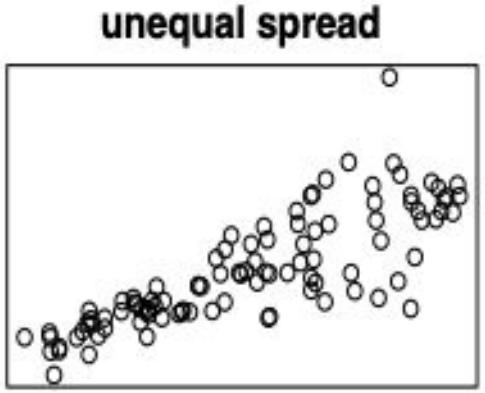
Scatter Plots

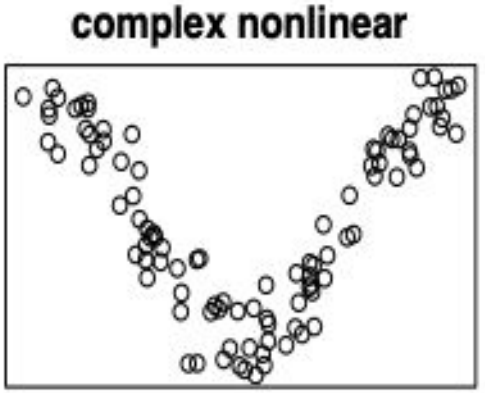
Used to inform model choices

- E.g. simple linear model requires linear trend and equal spread.



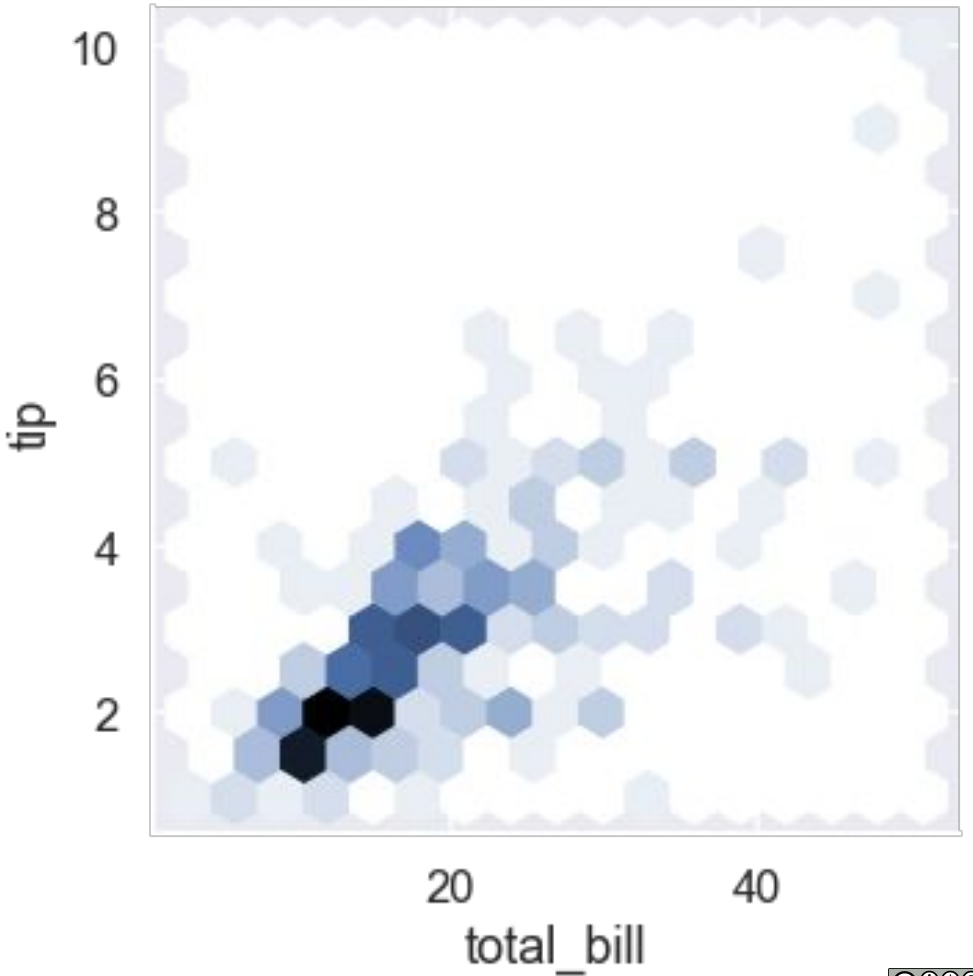






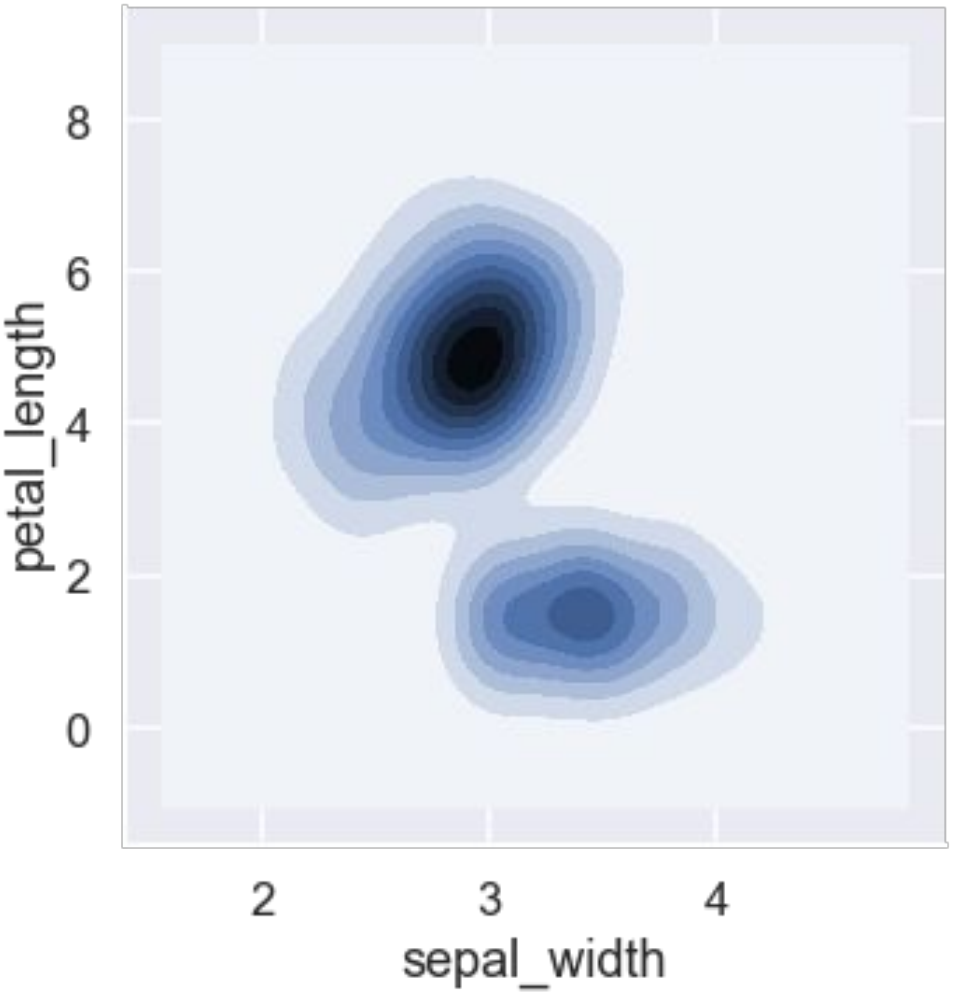
Hex Plots

- Equivalent of histogram in two dimensions
- Shaded hexagons usually correspond to more points



2D Density Plots

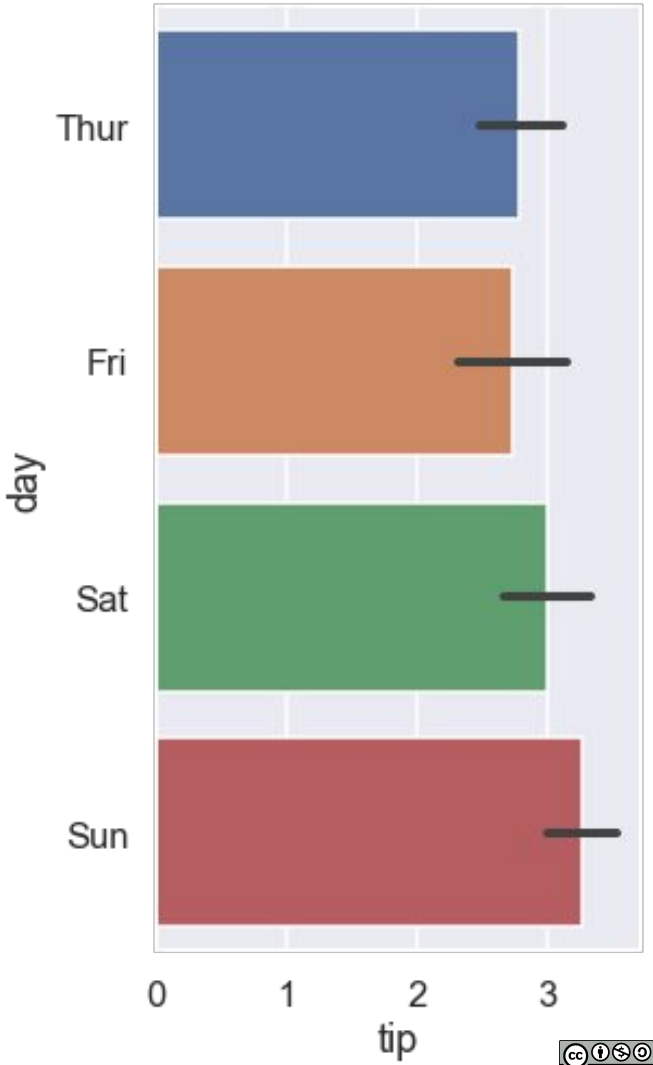
- Density plots also work in two dimensions!



Common Visualizations for Qualitative + Quantitative Variable

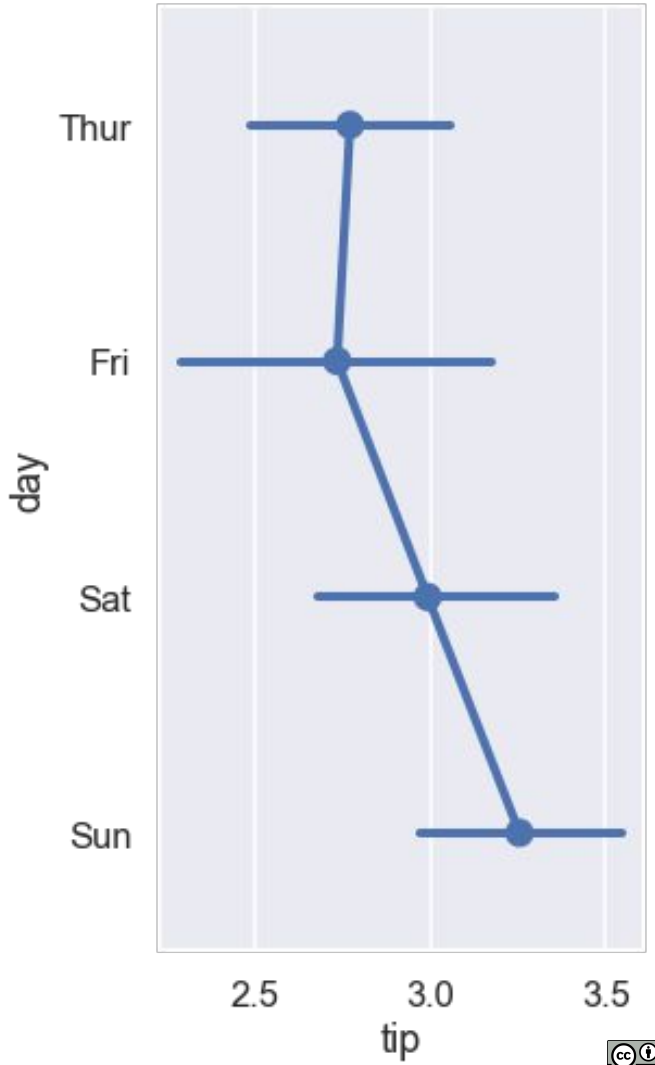
Bar Plots

- Typically use horizontal bars to avoid label overlap
- Can also plot confidence intervals on bars if appropriate



Point Plots / Dot Plots

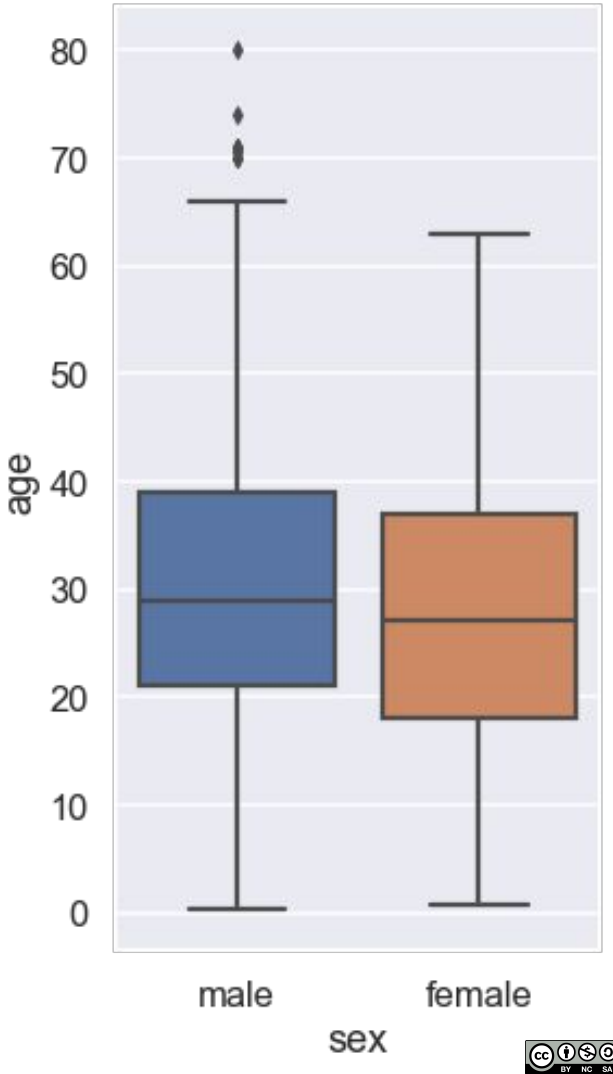
- Minimal cousin of the bar plot
- Some prefer point plots since the bar widths in a bar plot have no meaning



Box Plots

Used to compare distributions

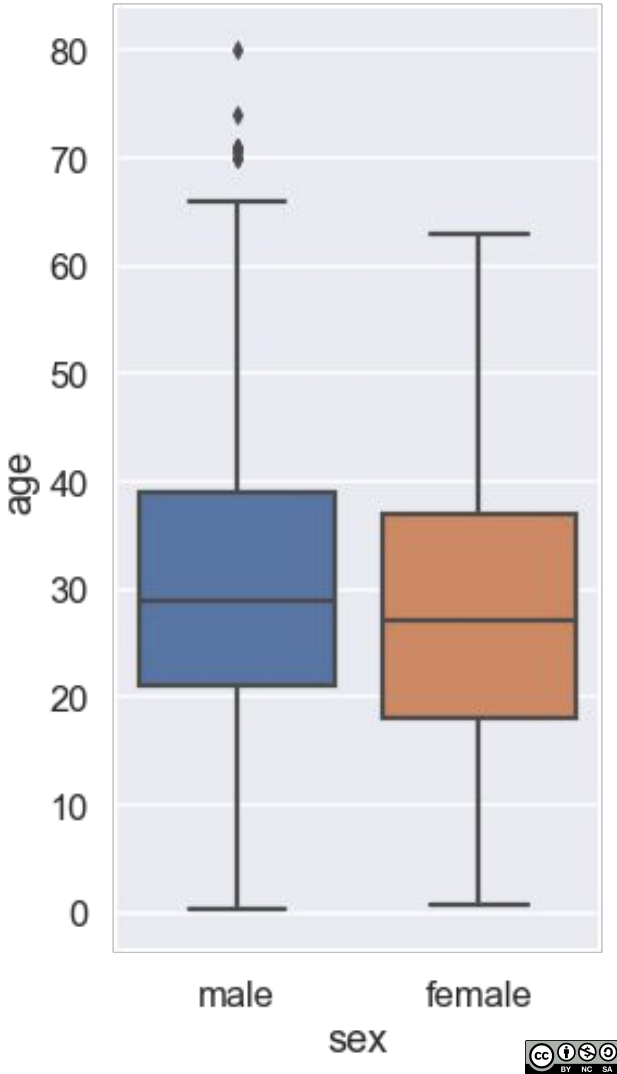
- Uses quartiles
  - Q1: 25th percentile
  - Q2 (median): 50th
  - Q3: 75th
- Middle line = median
- Box shows 1st and 3rd quartile
- Whiskers show rest of data
- Outliers = 1.5 * (Q3 - Q1) past Q1 or Q3



Box Plots

Outliers plotted beyond whiskers

- Interquartile range IQR = Q3 - Q1
- Outliers are defined as:
  - 1.5 * IQR beyond Q1 or Q3
- Example for male ages:
  - Q1 = 21; Q2 = 29; Q3 = 39
  - IQR = 18; 1.5*IQR = 27
  - Outliers are:
    - Above Q3 + 1.5*IQR = 66
    - Below Q1 - 1.5*IQR = -6



Summary

- Data visualization is underappreciated!
- This class mainly uses seaborn + matplotlib
  - Pandas also has basic built-in plotting methods
- Types of variables constrain the charts you can make
  - Single quantitative: histogram, density plot
  - 2+ quantitative: scatter plot, 2D density plot
  - Quantitative + qualitative: bar plot, point plot, box plot
- Tomorrow: Four case studies to illustrate visualization principles!