A3 Recap, Plotting, A4

Learning goals:

- Go over commonly asked questions for A3.
- Understand how common Python plotting libraries relate with each other.
- Walk through questions on A4

COGS 108 Winter 2020

Sam Lau

Discussion 6

bit.ly/sam-wi20

lau@ucsd.edu

OH: Thurs 11a-12p in SSRB 100

A3 Recap

Question 1f: merging DataFrames

- df_steps has 11k rows, df_income has 12k rows, but merging the two gets 9k rows. Why?
- Goal: Get you to understand how merging works in pandas.
- Default in pandas is to drop rows without matching values!
  - This is a very easy way to mess up your data.
- Answer: Some id values were missing in the other DF.

Inner join vs. Left join

Inner joins drop all rows without a matching value.

Left joins keep all rows in the left table, even if values do not have a match.

Email	Name
sam@ucsd.edu	Sam
jen@ucsd.edu	Jen
kay@ucsd.edu	Kay
min@ucsd.edu	Min

Email	Order
jen@ucsd.edu	Keyboard
sam@ucsd.edu	Mouse
kay@ucsd.edu	Cable
wade@ucsd.ed	Lamp

Inner join:

Email	Name	Order
sam@ucsd.edu	Sam	Mouse
jen@ucsd.edu	Jen	Keyboard
kay@ucsd.edu	Kay	Cable

Left join:

Email	Name	Order
sam@ucsd.edu	Sam	Mouse
jen@ucsd.edu	Jen	Keyboard
kay@ucsd.edu	Kay	Cable
min@ucsd.edu	Min	NULL

Question 4a: counting -1

- How to count number of rows that have -1 in steps column?
- Simplest method: keep only rows that have -1 in steps, then count how many rows:

```
len(df[df['steps'] == -1])
```

- Or, create boolean Series and count number of Trues:

```
sum(df['steps'] == -1)
```

Question 5c: Correlations

- **Values in correlation table are correlations between pairs of variables.**
- **Most correlated = correlation furthest away from 0. Not always the most positive value!**
- **Most correlated with age? Steps**
- **Most correlated with income? Age**

	id	age	steps	income	income10
id	1.00e+00	-6.85e-03	5.56e-03	-0.03	-7.75e-03
age	-6.85e-03	1.00e+00	-2.82e-01	0.27	1.03e-01
steps	5.56e-03	-2.82e-01	1.00e+00	0.05	2.78e-02
income	-2.57e-02	2.67e-01	5.11e-02	1.00	4.70e-01
income10	-7.75e-03	1.03e-01	2.78e-02	0.47	1.00e+00

Plotting

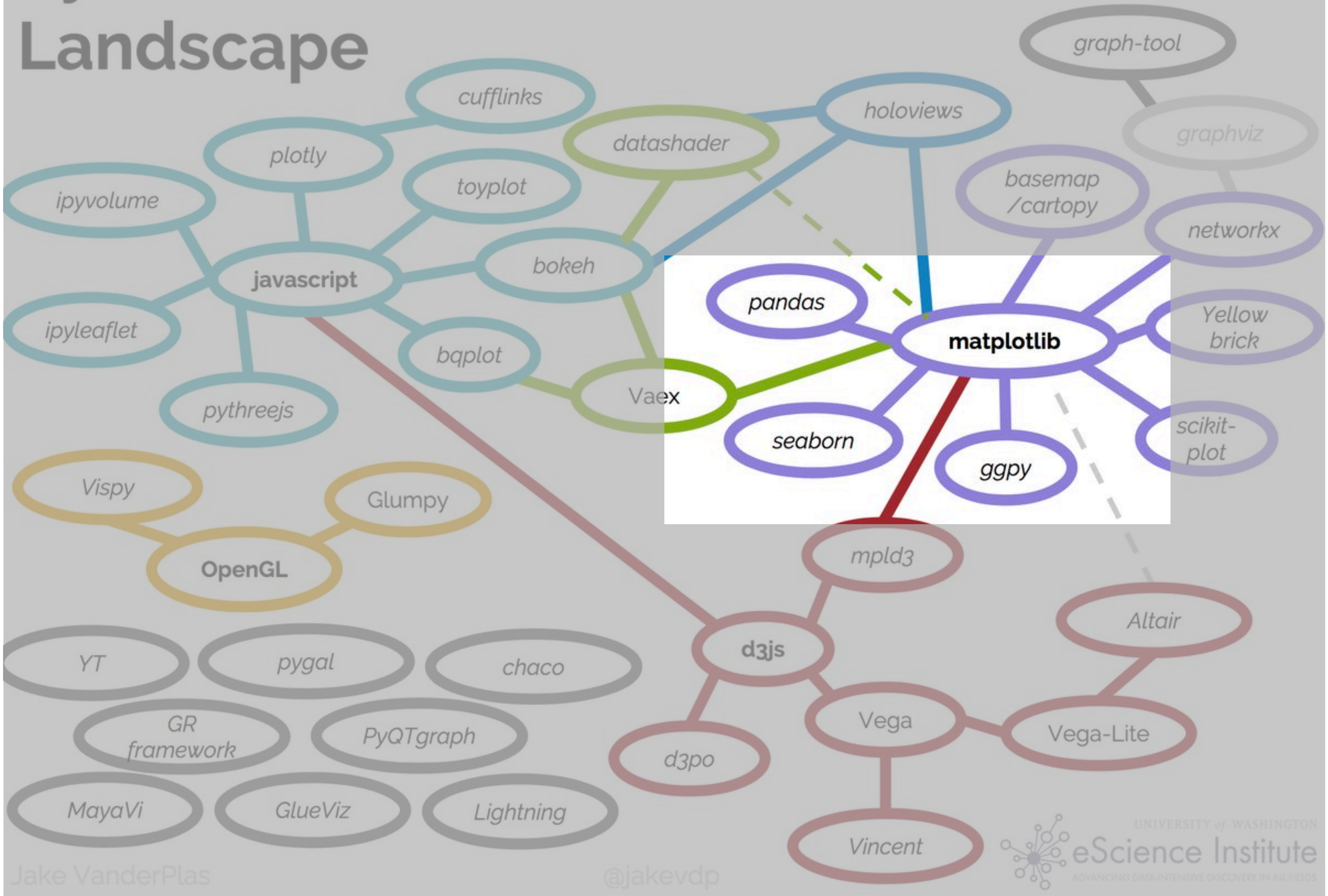
Why are there so many ways to make the same plot?

- **All of these do the same thing:**

```
plt.hist(df['income10'], 25)
df['income10'].hist(bins=25)
df.hist('income10', bins=25)
```

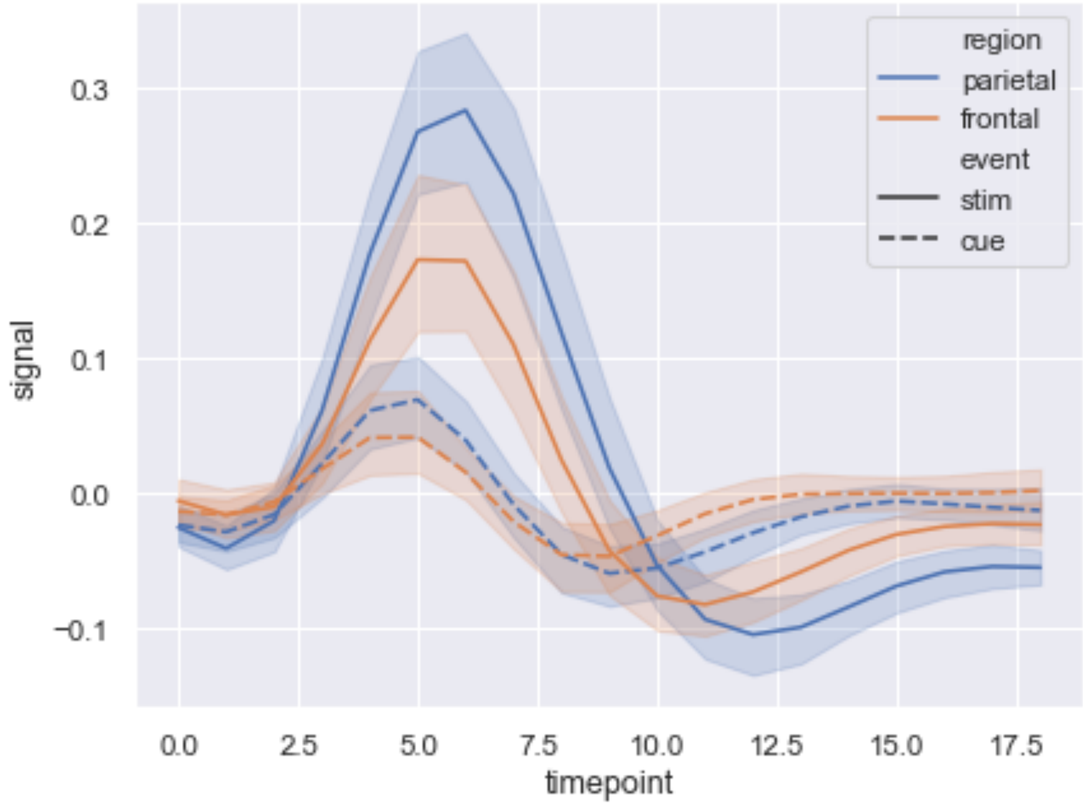
- **In Python, most image-based plots created using Matplotlib.**
  - `plt.hist` `plt.bar` `plt.plot` **etc.**
- **Pandas gives shortcuts for matplotlib plots. Lines 2 and 3 are shortcuts for line 1.**

Python's Visualization Landscape



Seaborn

- **My personal favorite is the seaborn library.**
- **Makes common statistical charts easy to create, like bar plots with confidence intervals.**
- **Again, seaborn is really just a bunch of shortcuts for matplotlib.**



https://seaborn.pydata.org/tutorial.html

For more details

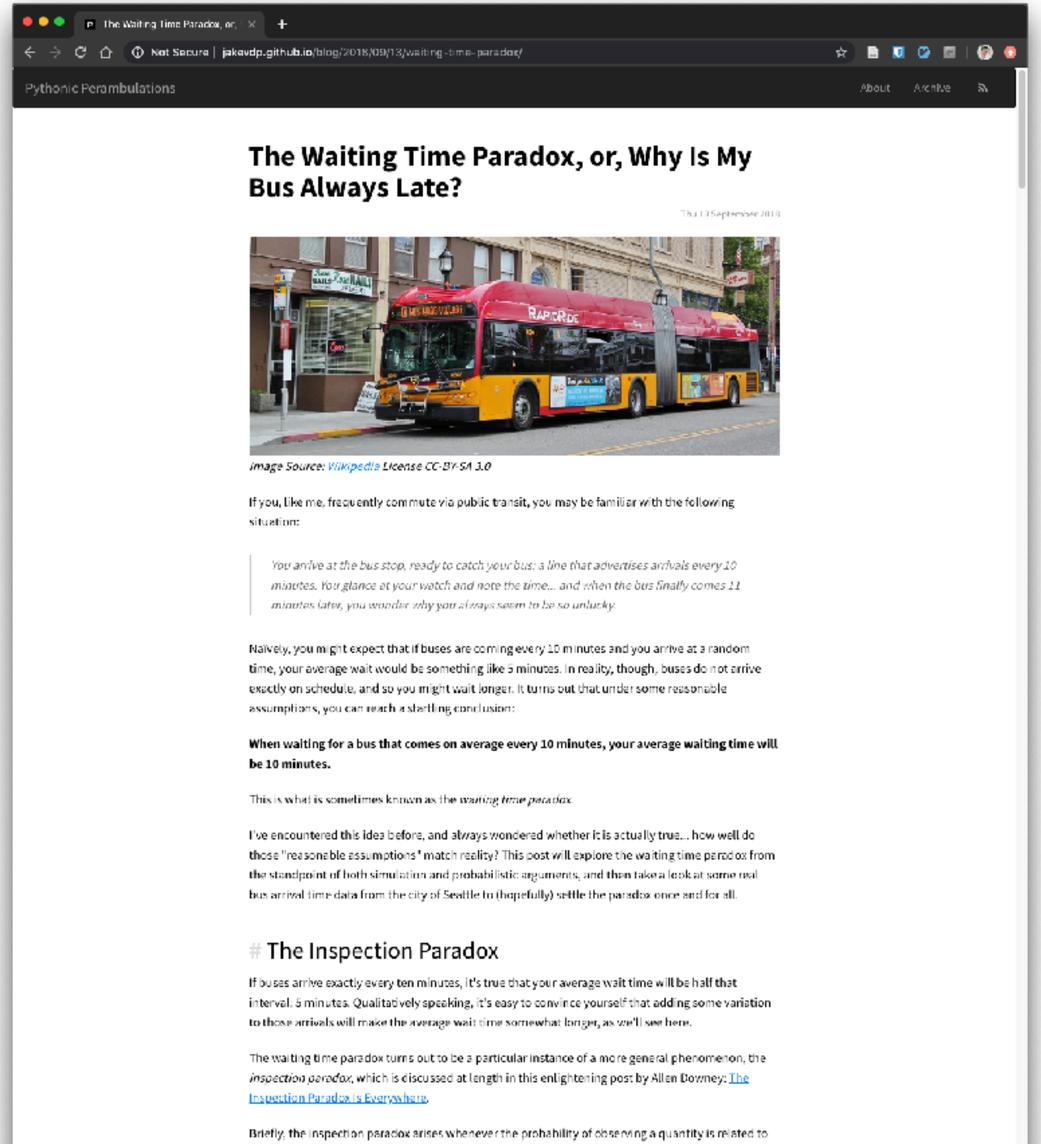
- **Making good plots is a key skill! This just scratches surface.**
- **You can get many great jobs just by being able to make informative data visualizations.**
- **For more, see Ch 6 of textbook.ds100.org.**

A4 Walkthroughs

Preview of next week

An easy way to set up a personal website using Jupyter notebooks and GitHub.

A5 question walkthroughs



A4 quick tips

- 1d: Your DF cells should have `\n` at the end (same for 1e)
- 2b: Don’t manually make a new Series – slice a column out of a DF
- 2e: Use a slice with multiple boolean expressions
- 2f: Your for loop should loop through the index of a DF
- 3b: Don’t do anything to the zip column
- 3f: Loop through the DF’s index again. Your 3-digit zip codes should be stored as strings, not ints.