Working with Datasets in Python

Lecture and In-class Activities

	Lecture	Activity
Day 1	Access and explore the data	Worksheet
Day 2 half day	Explore the data and perform statistical analysis	Lab
Day 2-3	Use libraries and documentation to create data visualizations	Gallery Walk

Working with Datasets in Python

- This week's lecture and in-class activities use U.S. COVID-19 data to introduce how to access and use datasets in Python programs. The data was downloaded from Kaggle, and was originally collected by the COVID-19 Tracking Project.
- Throughout the week, students will also reflect on the limits of data, data interpretation, data, and the responsibility that comes with using real-world datasets.

- This document includes notes for each lecture and the corresponding in-class activity.

Day 1: Access and Explore the Data

Opening files

- In Python, the `open(file, mode)` function opens a file, and returns it as a file object.
- You can specify a mode (read, append, write, create, etc.), otherwise it defaults to read.
- When you use `'with open(file, mode)'`, the file automatically closes.
- After reading a file, you'll need to parse the data so it becomes usable.
- The `split()` method divides a string into a list of substrings based on a specified delimiter, often a comma when working with .csv (comma separated values) files.
- The `strip()` method removes leading and trailing characters from a string, such as whitespace.

- You may see other methods for reading .csv files.
    - Example: Using Python's built-in csv module or the pandas library.
- There are also different ways to parse the data.
    - Example: Organizing lists by rows or columns.

File structure

- File structure is important when accessing data from another file
- If the data is not in the same folder as the file you are running, Python will return a FileNotFoundError.
- Folder structure:

```
cs111
hello.csv
|-----     world
      |----- run_hello.py
```

- To use `hello.csv` in `run_hello.py`, we need to use a relative path: `../hello.csv`
    - Folders up: `../hello.csv` → `../../hello.csv`
    - In a subfolder: `subfolder/hello.csv`
    - Same folder: `hello.csv`

Rows, columns, etc.

- How your data was imported will affect how you access it.

- ```
  data =
   [['20201206', 'MN', '350862.0']
   ['20201206', 'MO', '322298.0']
   ['20201206', 'MP', '106.0']]
  ```

date	state	positive_cases
20201206	MN	350862.0
20201206	MO	322298.0
20201206	MP	106.0

- The dataset is stored as a list of lists, where each inner list is a row
- Indices start at 0
- Indexing the outer list selects a row.
    - `data[0]` returns the first row: `['20201206', 'MN', '350862.0']`
- Indexing the inner list selects a column within that row.
- `data[1][0]` returns the first element of the second row: `'20201206'`
- You can also access multiple rows or columns using slices: `data[0:2]` returns the first three rows.
- If you go beyond the number of rows or columns, you will get an IndexError
  ```

Activity:

- Worksheet on their own practicing ideas taught in class. Encouraged to share ideas and help each other.

Day 2.1: Continue exploring the data and perform statistical analysis

- Access data from the `covid.csv` file and go through different statistical analyses.
- In both lecture and lab, we will use the full COVID-19 dataset.
    - If students want to follow along, prompt them to download the dataset and open a new Python file.

- Go through the following examples (`day2_lecture.py`):

    - **Technical Reflection**: We want to find the state with the most positive cases of COVID-19? Discuss with your partner how you might do this.
    - **Code**: Most cases.
    - **Code**: Least cases.
    - **Technical Reflection**: How could we check if our code is printing the correct data without looking at the CSV? Discuss with your partner.
    - **Code**: Checking least cases.
    - **Responsibility Reflection**: Why is it important to think carefully about logic and check results when working with data? Discuss with your partner.
    - **Responsibility Reflection**: What do you think accounts for the difference in positive cases between the states with the least and most? Discuss with your partner.
    - **Code**: Get most/least cases with max/min. Explain what is different about how data is stored with this logic.

Activity:

- Lab on their own practicing ideas taught in class. Encouraged to share ideas and help each other.

Day 2.2-3: Use libraries and documentation to create data visualizations

Libraries

- Libraries are a group of files (modules) that contain functions, classes and methods to perform tasks
- Examples:
    - Matplot lib for data visualization
    - Datetime for working with dates and times (built-in)
    - PyTorch for machine learning
- Instead of writing your own code, you can use a library that already does what you need
- Import the library to use it
    - `import x`
    - `import x as`
    - `from x import y`

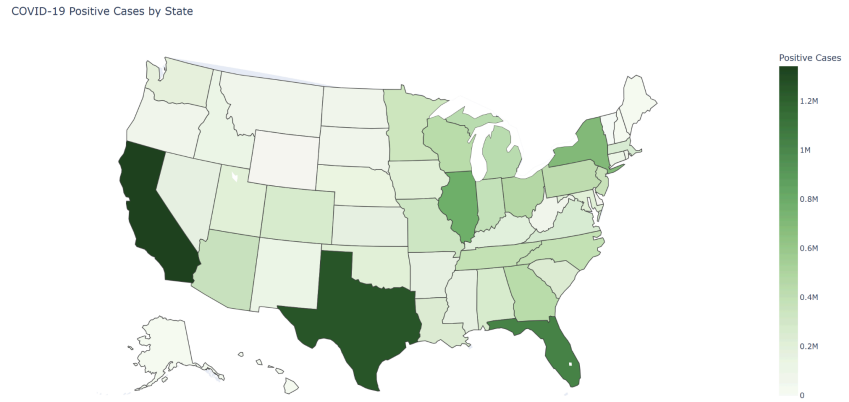
Documentation

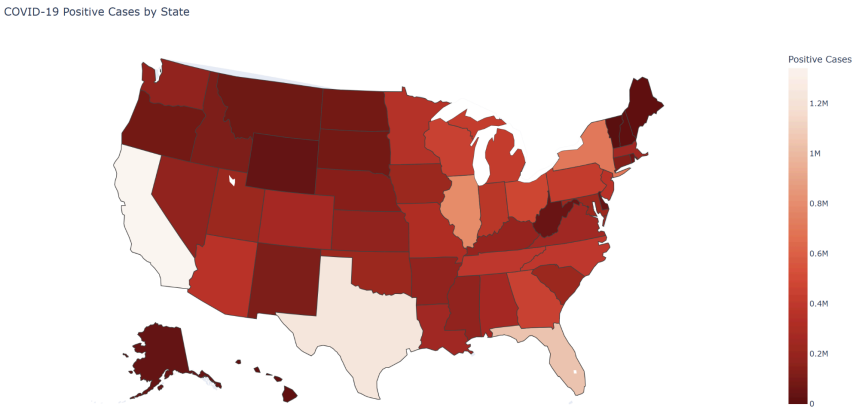
- Look at matplotlib documentation as a class:
    - Quick start guide
    - API reference
    - Examples

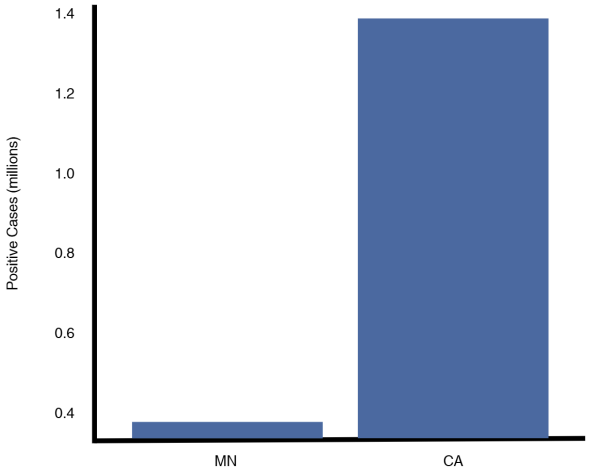
Activity (Day 2):

- Next class you will be making a graph using the COVID-19 data. With your group (2-4), think about how you might want to do this. We will look at some examples:
    - Have students reflect:
        - What might be misleading?
        - What about a graph can influence perception?
        - What may have been the creators' intentions?
    - Emphasize how understanding of data can easily be skewed and the responsibilities we have when presenting data.
    - Give students code from example, `map.py`.

Examples of misleading graphs:







Activity (Day 3):

- 35 minutes: With a group (2-4), students use a library and read documentation to create a graph of the `covid.csv` data. Students should use their `day2_covid_lab.py`file.
- 25 minutes: Walk around and look at other groups' code and graphs. After looking at all graphs, have groups stop at one that is not their groups. Write on the board:
    - Does the graph clearly communicate the data?
    - Could there be any misinterpretations?
    - If you were presenting this to the public, would you change anything? (Hoping to connect to ideas from earlier in the week. For example, accounting for population size, missing data, etc.)