Lab 05: Introduction to (or refresher of) data science tools in Python

Read the instructions COMPLETELY before starting the lab

This lab is the first exercise in the second half of the class, which focuses on spatial data science using open source Python tools. This lab is designed to help you gain some basic familiarity with data structures and functions, in particular the DataFrame.

This following tasks use the CSV file: *lab05_climate_data.csv*, which can be found in the /labs/lab05/ directory of the Course GitHub repository (https://github.com/pjbitterman/UNL_geog432/tree/main/labs/lab05). This file includes a limited set of *daily* climate variables for Lincoln, NE in 2020.

- month: numeric representation month
- day: numeric day of month
- year: 2020
- day_of_week: String for the day of week of the observation
- t_max: maximum daily temperature
- t_min: minimum daily temperature
- t_mean: arithmetic mean daily temperature
- precip: daily precipitation (a "T" value means "trace precipitation")

Using this file, complete the following tasks below. Show your work. All results must be "printed" to the Jupyter notebook or to the console. The in-class exercises and examples here: https://darribas.org/gds_course/content/bB/lab_B.html will aid you in your task. As they discuss in the preceeding link, real world data is often messy. You WILL need to engage in data munging to clean up some of the fields to make them useful. You may need to consult the help documentation for pandas or matplotlib (or whichever packages you use) to achieve your tasks. Note, you will need to *programmatically* deal with trace precipitation values and any other clean up tasks.

Tasks:

1. On what day of the year (yyyy-mm-dd) did the coldest minimum temperature (t_min) occur? The second coldest?

2. How many days had measurable precipitation in July 2020?

3. What is the month with the highest *average* maximum temperature (t_max)?

4. What day of the week had the greatest *total* precipitation in 2020? The lowest *total* precipitation?

5. Print the 7 days with the highest minimum temperature (t_min)

6. Make a histogram of maximum daily temperature (t_max)

7. Make a histogram of daily precipitation. Bonus points if you add a smoothing function (e.g., kernel density estimator) to the plot

8. Make a scatter plot with minimum daily temperature (t_min) on the x-axis and daily precipitation on the y-axis. Make the points different colors by month. For example, January would plot as red, February as purple, etc.

Questions:

1. What previous experience (in Python or other languages) do/did you have with "data science" (broadly conceived)?

2. What new skills did you learn (or refresh) in this lab?

3. What did you find particularly challenging? How did you overcome this/these challenge(s)?

What to turn in

- Your Jupyter notebook (or Python script). I must be able to run your code - do not turn in a screenshot or code pasted into a Microsoft Word document

- The answers to the above questions