**DS2500**
**Fall 2023**
**Homework 2**
**Assigned:** September 29, 2023
**Deadline:** October 13, 2023 @ 9pm

You may not use Pandas or NumPy on this homework, as we're building/reinforcing our foundational programming skills.

Submit your solution on gradescope: https://www.gradescope.com/courses/559361 This homework is due at **9pm on Friday, October 13th.** If you miss the deadline, you can submit up until October 20 at 9pm -- look for the "HW2 - Late Submission" link on gradescope.

Homeworks are broken into three components:

1. **Accuracy**. You'll answer quantitative questions about the dataset on gradescope. These answers are auto-graded. Gradescope can be a little picky, so make sure you don't put extraneous characters or whitespace in your answers -- and double-check the "correct/incorrect" confirmation!
2. **Visualization**. You'll be asked to submit screenshots/downloads of at least one Python plot on every homework. We expect these plots to be labeled, easy to read and understand, and appropriate for the data.
3. **Code Quality.** You'll submit your code as well, which we will review and grade based on its modularity, readability, and reusability.

Please see the DS2500 Homework Grading Guidelines for more on grading.

You are welcome to work with classmates, but you may not share code with each other, and you may not post code on Ed Discussion. If you find a code snippet online, you're welcome to use it but you need to provide a citation.

**The 30-minute Guideline**
If you get stuck on a homework problem, come by office hours or post on Ed Discussion! We recommend you spend about 30 minutes trying to figure out a problem, and then ask for help. Enough time that you can try a few things to get unstuck, but not SO much time that you're banging your head against the wall. Try for 30 minutes, then ask us. :)

**Data for this Homework**

- marathon_data.zip

This is the same marathon data we've used in class, but with more years included.The data was collected from https://www.baa.org/, and every registered participant is required to sign a waiver which notifies them their results will be publicly available.

For every Boston Marathon from 2010 through 2023, we have data from the top 1000 runners, including:

- Name
- Gender[1]
- Country of Origin
- Official Time
- Rank

**Part 1 - Questions about the Data**

This part of your solution will be auto-graded. When you find this assignment on gradescope, the first part will ask you the following questions; type or select your answers. You must compute the answers to these questions programmatically. Gradescope will confirm your answers are correct/incorrect.

**Check the output!** Gradescope can be a little picky about formatting, and we don't want you to lose points for putting extra characters or whitespace in an answer. Make sure you've got the correct answer to each question for full credit.

Please refer to the "country of residence" columns for data about nationality (not "country of citizenship").

Answer these questions (make sure you compute these answers in your Python solution):

1. In 2013, what was the mean finish time of the top 1000 runners?
2. What is the median age of the top 1000 runners in 2010?
3. Apart from the US, which country had the most runners in 2023?
4. How many women finished in the top 1000 in 2021?
5. What is the correlation (r-value) of year vs. the mean finish time of women in the top 1000?
6. What is the correlation (r-value) of year vs. the mean finish time of American runners in the top 1000?
7. If the 2020 race had actually happened, what would you predict to be the mean finish time of Americans in the top 1000?

**Part 2 - Visualization**

Create two Python plots and upload them as screenshots/downloads.

[1] The BAA introduced a non-binary division in 2023, and runners in that division have "X" listed as their gender marker.

- **Plot #1**: A linear regression plot  modeling the relationship between year and mean finish times of American runners in the top 1000.

- **Plot #2**: A plot showing how median age and average finish times have changed over time. Because finish times and ages are on quite different scales, use the min/max normalization from class to scale the data -- we'll cover it on Tuesday 10/3!

**Part 3 - Code Quality**

Submit on gradescope the code you developed to compute your answers to the Part 1 questions, and to generate the plots for Part 2.

Your code will be graded on modularity, readability, and reusability.