DS2500
Spring 2022
Project #1 - Requirements

You'll complete two projects in DS2500 this semester. Project #1 is a solo project. You'll choose a dataset, identify and answer a question about it, and present visualizations that communicate your findings to an audience.

The dataset and topic are up to you. Find a new dataset from a source such as https://www.data.gov, https://data.boston.gov/, https://www.kaggle.com/datasets, https://toolbox.google.com/datasetsearch, https://data.noaa.gov/dataset/, https://healthdata.gov/, etc. Your chosen dataset can be related to something we've done together (e.g., it could be a dataset about the weather/snow, about bike traffic, etc.), but it must be a dataset that is new to DS2500, and your analysis and questions must be your own.

Important Dates

Milestone	**Date**	**Notes**
Sign up for Presentation	Tuesday, March 2nd	Sign up here: https://bit.ly/3GPQEjc
Presentations	In lab and lecture March 7th, 8th, and 11th.	Modify the slides linked in the presentation sign-up.
Presentation Peer Review	The day of your presentation	Complete a peer review on the same day you present: https://forms.gle/b3BTU4NqfhAF6pUJ8
Code submission	March 11th	Submit python files or a Jupyter notebook on gradescope; data files must be submitted as well.

Presentation

Everyone must present their work. Presentations will happen during lab and lecture the week of March 8th. We'll follow a lightning talk format; in each section we'll have one slidedeck that everyone presents.

Everyone presents the same three slides:

- *The dataset* (what data are we looking at, from what source and time span)
- *The question* (what do we want to know about the data, and why is it important)
- *The insight* (a visualization communicating what you learned with your analysis)

Your presentation will be evaluated on:

- **Content**. We expected substantive, meaningful work. You've gathered a dataset that has meaning for you, and constructed a question and answer that gleans some insight into the data.

- **Clarity**. We expect a clear, understandable speaking style. The dataset you used and code you wrote should be *described* but not included in the slides.

We expect that you will be in attendance for the presentation slot you signed up for. We do not accept late project submissions. In case of extenuating circumstances, fill out the Late Project form at least 24 hours before your presentation or the project deadline.

Peer Review

Part of your grade for this project will come from your completion of three peer reviews: one from your Monday lab, one from Tuesday lecture, and one from Friday lecture.

You'll attend your usual lab and lecture sections during the week of March 7th. You'll watch your classmates' presentations and choose one from each lab/lecture to evaluate. Fill out the peer review form (linked above) three separate times, no later than 9pm on March 11th.

Submission

You'll submit your code on Gradescope, either as Python files (.py) or as a Jupyter notebook (.ipynb). You will also submit your dataset(s). You do not need to submit your slides, because we'll have access to the slidedeck everyone used. You do not need to create a written report for this project.

Unlike your homeworks, projects are graded traditionally. You will receive a numeric score on the quality of your code. We'll run your submitted code against the dataset(s) that you provide on Gradescope. It will be evaluated along the same categories we use to give feedback on the homework:

Category	**What We Expect**
Program Correctness	Program meets specifications of the assignment and generates the expected output. No errors or warnings are produced when we run the code.
Readability + Reusability	Variable and function names are clear and concise. Code is modular and functions are used appropriately. Code is clean, understandable, and well-organized.
Documentation	Comments are clear and frequent; test cases or unit tests are provided for functions when appropriate.
Visualizations	Plots and other visuals created convey insights you learned about the data. They contain labels and legends that explain the content. The scale used makes sense for the data.