CMSC6950 — Fall 2025
Final Projects

The final project for this course is centred around writing data processing and analysis routines in Python.

Data set

You must first identify a data set that you would like to analyse. While you have freedom to choose a data set that interests you, your data set must have several features:

1. Your data set must have time-series (or similar) data with at least 100 data points, and multiple measurements for each data point.
2. It must either come from an original citable resource (website, open data set, or similar) or from a research code that you can run to produce the data. You may not use a site like Kaggle (or other dataset repositories) as a source for this project.
3. You must be able to identify meaningful extreme values in your data, either through statistical tests or appropriate interpretation of the data values.

As an example of a suitable data set, consider the historical weather data available from `https://climate.weather.gc.ca/historical_data/search_historic_data_e.html` , where you can download climate data (e.g., daily max and min temperatures) for a given year for multiple cities. This data has a reliable source, and extreme weather events (e.g., above average daily high temperatures or below average daily low temperatures) have a natural definition.

Data processing and analysis

Once you have identified and acquired your data set, you must

1. Plot your data in a series of clearly labelled plots with consistent and well-defined style.
2. Develop a meaningful function that you can compute from your data that you will write code and tests for and include in those plots.
3. Compute some meaningful statistics regarding extreme values in the data (such as days above/below historical mean temperatures in the above example) and present this data in a clear and concise way. Explore sensitivity of these results to the definition of "extreme values", again presenting data in a clear and concise way.
4. Identify and discuss trends (or lack thereof) in the data, using appropriate statistical or other tools.

Your grade for the project will be determined by the quality and thoroughness with which you approach the above tasks. All work must be committed at regular intervals to a git repository (that you create and "invite" me as a collaborator to), with proper unit tests for your code that computes functions and does data analysis (but not for scripts that produce figures).

Deliverables

There are four graded submissions required for this project.

1. You must submit a one-page project proposal by 5pm on October 21. This must clearly identify the data set that you intend to use in your project (what the data is and its source), as well as include a first plot of the data and a discussion of what function and statistics you intend to examine in detail. Feedback will be provided to help you improve your project.

2. You must initialize your git repository and invite me as a collaborator, also by 5pm on October 21st. Between then and the final project due date on December 4th, you must make regular, suitably incremental changes to the repository, gathering code and tests, and properly address any failing tests. The `README.md` file should include full instructions to reproduce every figure that appears in your project report from a Python script (not a Jupyter notebook).

3. You must submit an 8-10 page project report (with A4 or "letter" sized pages and legible fonts) by 5pm on December 4, including clear descriptions of your data set, methodology, and results. It is expected that this report include 6-8 distinct figures (highlighting both different aspects of the data and different visualization commands) that occupy about half of a page each (including a detailed caption). This report should **not** take the form of a Jupyter notebook (or similar), but should be written as a proper project report, with text integrated around the figures.

4. You must present a 5-minute "lightning talk" with up to 5 slides that summarize your data set, hypotheses, and results, for presentation in class on either December 2 or 4.

These deadlines are not eligible for the extensions allowed for regular homework assignments.