CMSC6950 — Fall 2023
Final Projects

The final project for this course is centred around writing data analysis routines in Python. 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 a citable resource (website, open data set, or similar) or from a research code that you can run to produce the data.

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 (above average daily high temperatures or below average daily low temperatures) have a natural definition.

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. 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.

3. 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 (to be shared with me), with proper unit tests for your code. A portion of your grade will be assigned based on your git commit history, including informative commit messages, suitably incremental changes to the repository, and properly addressing any failing tests. The `README.md` file (or similar, but clearly labelled file) should include full instructions to reproduce every figure that appears in your project report.

There are three graded submissions required for this project, in addition to the work in your git repository. First, you must submit a one-page project proposal by 5pm on October 17. This must clearly identify the data set that you intend to use in your project, as well as include a first plot of the data and a discussion of what statistics you intend to examine in detail. Feedback will be provided to help you improve your project. Secondly, you must submit an 8-10 page project report by 5pm on November 30, 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). Finally, you must present a 3-minute "lightning talk" with 1-3 slides that summarize your data set, hypotheses, and results, for presentation in class on either November 28 or 30. These deadlines are not eligible for the extensions allowed for regular homework assignments.