Homework 1 — EDA and Visualization

Due Date

This assignment is due at the end of the day February 1, 2026 (week 4). It is out of 40 points.

Learning goals

- Download, read, and get familiar with an external dataset.
- Step through the EDA checklist presented in class.
- Make exploratory plots of different types.

Assignment Description

We will work with Toronto Police Department Public Safety Data Portal, in particular, the major crime indicators open data.

Primary question:

What are the most numerous categories of crimes and offences in Toronto? Of these, when and where are they most likely to occur?

Download the data from here. Documentation are available here

Your assignment should be completed in **Quarto with Python**. Submit your **rendered html** and the **.qmd** to Quercus.

Steps

Given the formulated question above, conduct EDA checklist items 2–4.

1. **(3 points) Read in the data**

   Read the dataset into Python. We will focus on these key variables: `MCI_CATEGORY`, `OFFENCE`, `OCC_YEAR`, `OCC_MONTH`, `OCC_DAY`, `OCC_DOY`, `OCC_HOUR`, `LOCATION_TYPE`, `NEIGHBOURHOOD_158`, `LONG_WGS84`, `LAT_WGS84`.

Summarize the steps you took and include data summaries of the dataset dimensions, variable names and variable types, number of missing data.

If necessary, update missing data identifiers to `NaN`.

2. **(3 points) Preprocess the data**

   Since this is a very large dataset, let's simplify it for the rest of our analyses. Identify the most frequent **MCI category** and the most frequent **offence** within that category (summarize clearly what you chose and how you decided).

   Create an analytic dataset containing only:

   - observations in your selected category/offence, and
   - the necessary columns listed in Question 1.

   Rename key variables so they are easier to interpret and use in your analysis. Convert key string variables to pandas `category` dtype where appropriate. Identify any clear outliers or impossible values and justify how you handle them.

3. **(2 points) Check data completeness across years**

   Examine which years are present and cross-reference with the documentation (Appendix A / Open Data Summary Table). Subset to the years with the most complete data. Summarize what you did and why.

4. **(10 points) Examine annual change in crime**

   Aggregate the data to count of crimes per year. Create a table of annual counts and a plot of counts over time. Then fit a Poisson regression to quantify the annual trend (y is annual crime count, x is year treated as numeric). Report the estimated rate ratio per 1-year increase and a brief interpretation in plain language. Comment on the model fit and assumptions. *Is there evidence of change in crime over time across Toronto? If so, how large is the change?*

5. **(7 points) Examine seasonality in crime**

   Pick the most recent **complete** year in the dataset and subset the data to that year. Create a season variable (winter/spring/summer/fall). Produce a table of summary statistics by season and a plot of counts by season. Conduct a simple chi-square test and write a short interpretation that include relevant statistics to answer: *Which season(s) have higher/lower counts than expected?*

6. **(7 points) Examine seasonality, day vs night**

   Using the dataset from Q5, summarize crime counts by hour and create a day/night indicator (define your rule clearly). Using the season variable from Q5, create a season x day/night contingency table, visualize it, and conduct a chi-square test to assess whether the proportion of day vs night

crimes differs by season. Write a short interpretation: *Is there evidence that the day/night distribution of crimes changes across seasons? If so, which seasons look most different?*

7. **(4 points) Look at neighbourhood patterns**

   Using the same data from Q5 and Q6, summarize which neighbourhoods have more crimes. Present a table for the top 10 neighborhoods showing total count, day and night count, and night proportion. Provide a brief interpretation.

8. **(4 points) Create a map**

   Create an interactive **Folium** map showing where crimes occur (using latitude/longitude). Your map should clearly communicate spatial patterns (e.g., points, and/or a count-focused visualization if you choose to aggregate by neighbourhood).