Homework 01 - Exploratory Data Analysis

Due Date

This assignment is due at the end of the day January 29, 2025. It is out of 32 points.

Learning Goals

- Download, read, and get familiar with an external dataset.
- Step through the EDA "checklist" presented in class
- Practice making exploratory plots

Assignment Description

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

The primary question you will answer with these data is: *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 can be completed in RMarkdown or Jupyter Notebooks.

Steps

Given the formulated question from the assignment description, conduct EDA Checklist items 2-4.

1. (3 points) Read in the data. Update the missing data identifiers to `NA`. Check for import issues (dimensions, headers, footers, variable names and variable types). Summarize data and check for missing values, data errors particularly in the key variables we are analyzing (MCI_CATEGORY and OFFENCE, OCC_YEAR, OCC_MONTH, OCC_DAY, OCC_DOY, OCC_HOUR, LOCATION_TYPE, NEIGHBOURHOOD_158, LONG_WGS84, LAT_WGS84). Summarize the steps you took.
2. (3 points) Clean the data. Select the category of major crime and offence that is the most significant (largest counts). Create an analytic dataset with only these observations and the necessary columns listed above. Rename key variables so that they are easier to identify and summarize. Change the key variable type from string to factor as appropriate. Identify any outliers, and justify how you handle them.
3. (2 points) Examine the years of data, and cross reference with Appendix A Open Data Summary Table in the documentation. Further clean/subset the data to include years with the most complete data. Summarize your steps.
4. (5 points) Create a variable for season, and calculate summary statistics for each season and conduct basic analyses that enable you to compare across winter/spring/summer/fall (e.g. chi-square test, anova). Be sure to create a table of the results and write up explanations of what you observe in these data. Is there a difference in crime by season?

5. (5 points) Look at hour of the day and then create a variable for day/night. Is there a significant difference in the number of crimes by time of day? Again create a table and provide a written summary of your findings.

6. (5 points) Similar to above, look at and summarize which neighborhoods have more crimes.

7. (6 points) Create exploratory plots (e.g. boxplots, bar plots) for displaying what you summarized in questions 4-6.

8. (3 points) To visualize where the crimes are occurring, create a `leaflet()` map of the counts by neighborhood.