Lab: Analyzing Crime Data with Weather Information

Introduction

In this lab you will investigate how weather conditions may influence crime patterns. We will work with two real datasets:

- **crimes:** Records of crimes (see URL: Crimes.csv)
- **weather:** Daily weather observations (see URL: weather.csv)

Your tasks include:

1. Subsetting and cleaning the data.

2. Merging the datasets by date.

3. Engineering features to better categorize precipitation types.

4. Filtering the data to focus on key crime types and specific hours.

5. Creating plots that explore the relationship between weather and crime patterns.

Part 1: Data Import and Preliminary Cleaning

Step 1: Import Data

- Read the `crimes` dataset from its URL using `read_csv()`.
- Read the `weather` dataset from its URL using `read_csv()`.

Step 2: Subset the `crimes` Data

- **Filter** the `crimes` data to remove domestic incidents. (Hint: Use a condition that checks for "Domestic" in the appropriate column.)
- **Select** only the columns: `Date` and `Primary Type`.

Step 3: Subset the `weather` Data

- From the `weather` dataset, select only the columns: `datetime`, `temp`, and `preciptype`.

Consider:
- Do the column names appear consistently? If not, consider cleaning them (for example, with the **janitor** package).

- How might the date formats in `Date` and `datetime` affect your merging strategy?

Part 2: Merging the Datasets

Step 4: Prepare for Joining

- Extract or convert the date portion from `crimes$Date` and `weather$datetime` so that both datasets have a common key (e.g., using `as_date()` from **lubridate**).
- Think about any potential issues when joining data that have different timestamp granularities.

Step 5: Merging

- Merge the two datasets so that each crime record gains its corresponding weather information (temperature and precipitation type) for the date the crime occurred.
- **Question:** Considering that there are multiple crimes per day, what type of join (e.g., left join, inner join) should you use? Explain your reasoning.

Consider:
- If a crime record lacks a matching weather record, how does that affect your subsequent analyses?

- What happens if the weather dataset has only one record per day, while there are many crimes on that day?

Part 3: Further Data Cleaning and Feature Engineering

Step 6: Handling Missing Weather Information

- In the merged dataset, replace any `NA` values in the precipitation variable (`preciptype`) with `"None"`, indicating no recorded precipitation.

Step 7: Create a Consolidated Precipitation Variable

- Use string detection (e.g., `str_detect()` from **stringr**) to create a new variable, say `precip_group`, that categorizes precipitation into three groups: `"None"`, `"Rain"`, and `"Snow"`.

Consider:

- How will you distinguish between rain and snow events based on the `preciptype` variable?

- Do you need to account for case sensitivity or variations in naming?

Step 8: Limit Crime Types

- The `Primary Type` column may contain many unique values.
- Identify the 5 most frequent primary crime types.
- Filter the data to include only these top 5 crime types for subsequent analysis.

Consider:

- Which function (like `fct_infreq()`) can help you determine the most common crime types?

- Why might it be valuable to limit your analysis to these top categories?

Part 4: Data Aggregation and Plotting

Step 9: Crimes by day

- **Filter** these data to include only crimes that occurred between 5 PM and 1 AM. (Hint: Extract the hour from the `Date` column and filter accordingly.)
- **Aggregate** the crime data by day. For each day, count the number of crimes.

Step 10: Plotting Crimes by Precipitation

- Create a plot (using **ggplot2**) that displays the number of crimes for each category in your `precip_group` variable.
- Be sure to label your axes and title the plot clearly.

Step 11: Examining Crimes Versus Temperature

- Create a plot that demonstrates the relationship between daily temperature (from the `temp` column) and the number of crimes, while also reflecting the precipitation grouping (by using color, faceting, or another aesthetic).

Consider:

- How would you structure your aggregate data for this plot?

- What additional variables or grouping strategies could provide deeper insights into how temperature and precipitation relate to crime frequency?

Extra Reflection

- Discuss possible reasons why weather might impact crime rates.
- Reflect on any limitations of the data or analysis—what additional data might help refine your insights?
- Consider any further cleaning or transformation steps that might improve the quality of your analyses.

Good luck and happy coding!