Data Report

Debanjan Chakraborty

November 26,2024

1 Data Report: Analyzing the Impact of Weather and Crime Patterns in Chicago, Post Covid (2021-Present)

1.1 Main Question

1. What are the patterns in crime types and locations across different neighborhoods in Chicago?
2. Apply time-series models to predict crime rates based on weather trends.

1.2 Data Sources

Two data sources have been selected for this project.: Chicago Crime data which provides a comprehensive record of reported crimes in the city of Chicago, and Chicago Weather data which weather and climate data in Chicago

1.2.1 Data Source 1: Chicago Crime Dataset

- **Metadata URL:** Chicago Crime Data (2001-present
- **Data URL**: Chicago Crime Data (2001-present) Raw
- **Data Type:** CSV Directory
- **Description** The Chicago Crime Dataset from Chicago Data Portal is essential for analyzing crime patterns, trends, and correlations within the city and can aid in understanding the distribution and nature of criminal activity over time.
- **Reason for Selection:** I chose this dataset to analyze crime trends in Chicago, focusing on the alarming frequency of crime happening in the US.

- **Data Structure and Quality:** The dataset is structured as a CSV format. The data set contains all necessary information like date, community area, primary type, district. The data aligns with the need of the project. Additional filtering is applied in the preprocessing step to refine the dataset.
- **License and Obligation:** The dataset is freely available for both non-commercial and commercial use from Chicago Data Portal. It is licensed under City of Chicago, under its Terms of Use, allowing for sharing and adapting for various purposes.

1.2.2 Data Source 2: Chicago weather Dataset

- **Metadata URL:** Chicago Weather Data (2001-present)
- **Data URL:** Chicago Weather Data (2001-present) RAW
- **Data Type:** GZIP Directory
- **Description** This data source provide weather and climate data in Chicago, including average air temperature, daily minimum and maximum air temperature, monthly precipitation total, maximum snow depth, average wind direction and speed, peak wind gust, average sea-level air pressure, and monthly sunshine total.
- **Reason for Selection:** I chose this dataset to analyze the crime trends relation with the weather in Chicago.
- **Data Structure and Quality:** The dataset is structured as a gzip format. The data set contains all necessary information like date, temperature, precipitation, pressure. The presentation of data fits the project's objective. Additional filtering is applied in the preprocessing step to refine the dataset.
- **License and Obligation:** The dataset is freely available for both non-commercial and commercial use from Meteostat. Weather data provided by NOAA , Deutscher Wetterdienst and Environment Canada. The website allows data for sharing and adapting for various purposes.

1.3 Data Pipeline

1.3.1 High-Level Overview

The data pipeline follows a traditional ETL (Extract, Transform, Load) process to handle datasets, including crime and weather data. It automates the tasks of fetching, cleaning, transforming, and saving data into SQLite databases for further analysis. Technologies Used: Python: For data processing and pipeline implementation, Pandas: To handle data transformation and filtering, Requests: For fetching data from external APIs or URLs, SQLite: For storing transformed data in a structured format, Gzip: To handle compressed data sets.

1.3.2 Transformation and Cleaning Steps

he pipeline transforms and cleans data by parsing dates, dropping missing values, and filtering records within a specified date range (e.g., 2021-01-01 to 2024-11-11). It ensures column consistency, handles compressed datasets (e.g., gzip files), and validates the schema for accuracy. The cleaned data is then saved into SQLite databases for further analysis.

1.3.3 Challenges and Solutions

The pipeline faced challenges such as missing values, inconsistent date formats, and handling compressed files. Missing values were resolved using pandas.dropna(), while inconsistent date parsing was managed with error-handling blocks. For compressed datasets, the pipeline used gzip and io.BytesIO to extract and process data seamlessly.

1.3.4 Meta-Quality Measures

The pipeline includes meta-quality measures like validating column names to ensure schema consistency and using error-handling blocks to manage unexpected data issues. It dynamically handles various file formats, such as CSV and gzip, and filters data based on relevant criteria, such as date ranges, to maintain data accuracy and reliability.

1.4 Results and Limitations

1.4.1 Output Data:

The pipeline outputs the transformed data in SQLite database format. This format was chosen because it is lightweight, easy to query, and supports structured data storage, making it ideal for handling multiple datasets and performing efficient data analysis. SQLite also ensures data persistence and compatibility with various data analysis tools.

1.4.2 Data Structure and Quality:

Transformed data is clean and consistent, with invalid rows removed and columns standardized. Stored in relational databases to support efficient querying.

1.4.3 Limitations:

Temporal Granularity: Crime data is aggregated by date, which may lose information at a finer temporal scale (e.g., time of day). Weather Data Lag: Monthly weather summaries may not align perfectly with daily crime data. Static Schema: Any changes in source datasets (e.g., new columns) require manual updates to the pipeline.