data-report

November 25, 2024

0.1 Data Report: Analysing the Relationship between Crime Data Analysis and Weather in Los Angeles (2020–2023)

0.1.1 Main Question

How have crime patterns (type and location) varied across different neighborhoods in Los Angeles from 2020 to 2023, and how do temporal factors (time of day, day of the week, and season) and weather conditions (temperature, precipitation, and wind speed) influence crime rates during this period?

0.1.2 Data Sources

To answer the question, two data sources have been selected for this project: the Crime in Los Angeles Dataset, which provides detailed records of crime incidents across neighborhoods in Los Angeles, and the Weather in Los Angeles Dataset, which offers monthly weather metrics for the same region. These datasets are complementary, enabling spatial and temporal analysis of crime patterns and their potential correlation with weather trends.

Data Source 1: Crime in Los Angeles Dataset

- **Metadata URL:** Crime Metadata
- **Data URL:** Crime Data
- **Data Type:** CSV

Description: This dataset includes detailed records of reported crimes in Los Angeles from 2020 to the present. It provides attributes like crime type, occurrence location, date, and time, enabling spatial and temporal analysis. Additional columns such as `AREA NAME` and `Crm Cd Desc` offer insights into neighborhood-specific crime distribution and classifications. The dataset also includes unique identifiers for each incident, ensuring robust data exploration capabilities.

Data Structure & Quality: This dataset is structured as a CSV file, featuring consistent tabular data suitable for exploratory analysis. After preprocessing, no major missing values or duplicates were found, ensuring high data reliability for further analysis.

However, some inconsistencies in formatting and column names were corrected during the pipeline's transformation phase.

License and Obligations: This dataset is openly available under the California Open Data License, allowing free use for analysis and research purposes. Proper attribution to the City of Los Angeles Open Data Portal will be included in all outputs.

Data Source 2: Weather in Los Angeles Dataset

- **Data URL:** Weather Data
- **Data Type:** CSV

Description: This dataset provides monthly weather data for Los Angeles, including average, minimum, and maximum temperatures (`TAVG`, `TMIN`, `TMAX`), total precipitation (`PRCP`), and average wind speeds (`WSPD`). The data is ideal for exploring potential correlations between weather conditions and crime patterns over time.

Data Structure & Quality: The weather data is provided as a clean, tabular CSV file, with rows corresponding to monthly records. Key columns include `Date`, `TAVG`, `PRCP`, and `WSPD`. Data quality is high, with minimal missing or inconsistent entries. During preprocessing, any missing values are handled using interpolation techniques.

License and Obligations: Published by Meteostat, this dataset is freely available for research purposes under an open-data policy. Proper credit to Meteostat and link to the License will be included to comply with usage terms.

0.1.3 Data Pipeline

The ETL (Extract [`extraction.py`], Transform [`transformation.py`], Load [`saving.py`]) pipeline is implemented using Python to handle both data sources, each downloaded as a CSV file. This process involves extracting the raw data, transforming it into a clean and consistent format, and saving it as CSV files ready for analysis.

Crime in Los Angeles Dataset ETL:

- **Extraction:** The dataset is fetched as a raw CSV file from the provided URL. A function (`extract_crime_data`) reads the CSV directly from the source. No special parsing was required as the dataset structure is straightforward.
- **Transformation:** The data is loaded into a Pandas DataFrame. A function (`transform_crime_data`) filters the data for the years 2020–2023. Columns irrelevant to the analysis are dropped, and the remaining columns are renamed for clarity. Duplicate and inconsistent entries are removed. Temporal features (e.g., "Month" and "Year") are created for aggregation.
- **Loading:** The cleaned DataFrame is saved as a new CSV file in the `/data/crime_cleaned.csv` directory using a function (`save_dataframe_to_csv`).

Weather in Los Angeles Dataset ETL:

- **Extraction:** The weather dataset is fetched as a gzip file from the provided URL. A function (`extract_weather_data`) decompresses the file and reads the content into a Pandas DataFrame.

- **Transformation:** The function (`transform_weather_data`) processes the data by filtering for the years 2020–2023, retaining only necessary columns like average temperature and precipitation. Missing values are imputed using forward-filling techniques. Column names are standardized for consistency.

- **Loading:** The transformed DataFrame is saved as a CSV file in the `/data/weather_cleaned.csv` directory using the same `save_dataframe_to_csv` function.

Problems Encountered, Solutions, and Error Handling:

- **Dynamic file names:** The Crime dataset filenames sometimes included inconsistent naming conventions, making it difficult to automate file identification. Regular expressions (regex) were employed to match file patterns and dynamically extract relevant files.

- **Incorrect or inconsistent data:** During the transformation process, several inconsistencies were observed, including incorrect date formats, missing weather data, and redundant columns in the crime dataset. Data cleaning steps such as standardizing date formats, forward-filling missing weather values, and dropping unnecessary columns ensured data consistency. Additional validation checks were implemented using Pandas to verify data types and detect anomalies.

- **Error Handling:** Exception handling was implemented throughout the ETL process to manage issues during file extraction, data reading, and transformation. Comprehensive logging mechanisms provided error diagnostics and ensured smooth pipeline execution.

0.1.4 Result and Limitations

Data Output: The output of the data pipeline consists of transformed datasets in CSV format. The Crime in Los Angeles dataset is processed to yield comprehensive monthly aggregates of crime incidents by type and neighborhood, covering the data from 2020 to 2023. For the Weather in Los Angeles dataset, the output includes a CSV file detailing monthly summaries of weather conditions such as temperature, precipitation, and wind speeds, aligned with the same timeframe.

Data Structure, Quality, and Format: The transformed datasets maintain a tabular format with data types for each attribute stringently validated to ensure consistency and reliability. The quality of the output is further enhanced through meticulous data cleaning steps, which include normalization of crime descriptions, alignment of temporal data points, and handling of any missing or anomalous data entries. CSV was selected as the final output format due to its broad compatibility with data analysis tools such as Python's Pandas library, facilitating easy access and manipulation for subsequent analytical processes.

Reflection and Potential Issues: While the data pipeline efficiently integrates and cleanses the data, several potential issues could impact the analysis. The aggregation of crime data into monthly intervals may dilute more nuanced daily or hourly trends that are crucial for understanding peak crime periods. The spatial resolution, restricted to

neighborhood-level aggregates, might mask more detailed spatial variations within larger neighborhoods. Additionally, the reliance on historical weather data may not accurately reflect short-term weather variations that could influence crime patterns. Future iterations of the project might benefit from incorporating higher-resolution crime data, real-time weather updates, and advanced analytical techniques to mitigate these limitations and refine the insights derived from the data.