Data Report

Abantika Bose

26th November, 2024

1 Data-Report: Analysis of Shooting Incidents and Weather Data: A Pipeline-Driven Report

1.1 Main Question

1. What are the spatial and temporal patterns of shooting incidents, and can these trends help identify high-risk areas and times for potential future incidents? 2. What are the trends in shooting incidents across different locations and times? 3. How do factors like time, day, or location influence the frequency of shootings? 4. Can we identify hotspots or predict shooting occurrences based on historical data? 5. How do weather conditions (such as temperature, humidity, and precipitation) correlate with the frequency and distribution of shooting incidents across different locations and times, and can certain weather patterns be linked to higher shooting occurrences?

1.2 Data Sources

To answer these questions, two data sources have been selected for this project: Shooting Incidents Dataset, which contains records of shooting incidents, including details such as location, date, and time and Weather Dataset (United States), which contains records of weather conditions across various U.S. locations.

1.2.1 Data Source 1: Shooting Incidents Dataset

- **Data URL**:Shooting Incidents Dataset
- **Data Type**: CSV Directory
- **Description**: A comprehensive dataset detailing police-involved shootings in the U.S. It includes demographics (age, gender, race), incident specifics (date, city, state), and factors like armed status and mental illness.
- **Reason for Selection**: I chose this dataset to analyze crime trends in the U.S., focusing on the alarming frequency of shootings and uncovering factors contributing to these incidents.
- **Structure and Quality**:

- **Format**: CSV, 30,000 records.
- **Issues**: Missing values, inconsistent date formatting, and minor redundancies.
  Despite these issues, the dataset is well-suited for analysis after preprocessing.

1.2.2 Data Source 2: Weather Dataset for U.S. Cities

- **Data URL**: Weather Dataset
- **Description**: Daily weather data for U.S. cities, including temperature, precipitation, humidity, and wind speed, covering 10 years.
- **Reason for Selection**: I incorporated weather data to examine how environmental factors, such as temperature or rainfall, might influence the occurrence of shooting incidents.
- **Structure and Quality**:
  - **Format**: CSV, 2 GB.
  - **Challenges**: Aligning weather data with shooting dates and standardizing geographic identifiers.
    Clean and structured, suitable for temporal and environmental analysis.

License, Obligations and Proof: Licensed under **Creative Commons Attribution 4.0 International (CC BY 4.0)**. Attribution is mandatory in reports or publications. Modifications must acknowledge changes. The dataset's licensing terms are explicitly stated on its source page, confirming its CC BY 4.0 license.

1.3 Data Pipeline

1.3.1 Overview: The data pipeline automates the ETL (Extract, Transform, Load) process to streamline data preparation for analysis. The steps include:

- **Extraction:** Datasets are downloaded from Kaggle using the Kaggle CLI.
- **Transformation:** Cleaning and preprocessing steps ensure the datasets are consistent and ready for integration.
- **Loading:** Cleaned datasets are stored in an SQLite database for analysis.

1.3.2 Technologies:

- **Python:** Libraries like `pandas` for data manipulation and `sqlite3` for database management.
- **Kaggle CLI:** Automates dataset downloads.
- **SQLite:** Chosen for lightweight, structured data storage.

1.3.3 Key Transformation and Cleaning Steps:

- **Weather Dataset:**
  - Filtered data to cover 2015–2020, aligning with the shooting dataset timeframe.

  - Converted date strings to datetime objects for consistency.
  - Addressed missing values and removed invalid entries.
- **Shooting Dataset:**
  - Dropped rows with missing location or time data to ensure completeness.
  - Standardized date-time formats for seamless integration with the weather data.

1.3.4 Challenges and Solutions:

- 1. Large weather dataset size (2 GB) and Filtered data during ingestion using city and date constraints to reduce memory usage.
- 2. Mismatched date formats in shooting data and Implemented a universal date parser using Python's datetime library.
- 3. Missing values in shooting data and Imputed using logical rules (e.g., replacing unknown race with NULL).

1.3.5 Error Handling and Meta-quality Measures:

- **Error Logging:** The pipeline logs all errors during extraction and transformation, ensuring traceability.
- **Dynamic Input Handling:** The pipeline reprocesses new or updated input data, overwriting outdated entries in the SQLite database.

1.4 Results and Limitations

1.4.1 Output Data:

- **Structure:** Cleaned datasets stored in SQLite (`processed_data.db`) with separate tables:
  - `shootings`: Shooting incidents.
  - `weather_2015_2020`: Filtered weather data.
- **Format:** SQLite database enables efficient querying and integration with analysis tools.

1.4.2 Data Quality:

- Shooting data is complete, with no missing values for critical fields (date, time, location).
- Weather data for 2015–2020 is consistent and free of outliers after cleaning.

1.4.3 Limitations:

- **Spatial Granularity:** Shooting data locations may not always match the city-level granularity of weather data.
- **Temporal Mismatch:** Aligning weather observations with exact shooting times is challenging.
- **Source Bias:** Reporting mechanisms in both datasets may introduce biases.

1.4.4 Reflections: The datasets provide a solid foundation for analyzing spatial and temporal crime patterns. However, improving data alignment and further validation are necessary for reliable predictive modeling.