Data Report: Geographic Hotspots and Incident Repetition Analysis of Gun Violence in the U.S.

By Mirza Shaheen Iqubal (22998316)

Question

Research Question:
This project aims to answer the following question:
"Which geographic locations in the U.S. experience the highest severity of gun violence incidents, and are there repeated incidents over time that show patterns of escalation?"

Objectives:

- **Identify** geographic hotspots for gun violence in the U.S.
- **Analyze** trends in repeated incidents and any escalation patterns over time.
- **Provide insights** to inform policies and preventive measures.

Data Sources

Data Source 1: FiveThirtyEight Gun Deaths Dataset

- **Why Chosen**: This dataset was chosen because it provides valuable demographic information on gun-related deaths, including details such as age, race, sex, and the intent behind the death (homicide, suicide, or accidental). The data allows us to analyze trends in gun violence based on demographic factors.
- **Source**: FiveThirtyEight GitHub Repository
  - **Direct Link to Data**: Full Data CSV File
- **Data Content**: The dataset includes columns like year, month, intent, sex, age, race, place, and education.
- **Data Structure and Quality**:
  - The dataset contains 16 columns with demographic data. Missing values are handled by filling the age column with the median age and the place and education columns with 'unknown'.
  - The structure allows for easy analysis of the relationship between gun deaths and demographic factors.
  - **Quality**: The dataset is generally high-quality, but it contains missing values, especially in non-critical columns (e.g., place, education).
- **License**: The dataset is available under the **CC0 1.0 Universal Public Domain Dedication**. This license allows for unrestricted use, modification, and distribution of the data.
  - **License Information**: CC0 1.0 Universal License
  - **Fulfilling Obligations**: Although no attribution is required by the **CC0 license**, it is good practice to credit the dataset's source. We will provide attribution in our project by crediting the original authors and providing a direct link to the dataset.

Data Source 2: Jamesqo Gun Violence Incident Data

- **Why Chosen**: The Jamesqo dataset offers incident-level data that includes geographic and temporal details, such as the number of fatalities, injuries, and locations (state, city). It allows us to perform fine-grained analysis of gun violence incidents across the U.S.
- **Source**: Jamesqo GitHub Repository
  - **Direct Link to Data**: Jamesqo Raw Data
- **Data Content**: The dataset contains columns such as incident_id, state, city_or_county, n_killed, n_injured, latitude, longitude, and date.
- **Data Structure and Quality**:
  - The dataset includes incident-level data with geographic and temporal attributes. Some columns, like latitude and longitude, had missing values, which were filled with zeros.
  - **Quality**: The dataset is comprehensive but has missing or incomplete values in less critical fields (e.g., latitude, longitude).
- **License**: The dataset is available under the **Open Data Commons Public Domain Dedication and License (PDDL) 1.0**, allowing unrestricted use, modification, and redistribution.
  - **License Information**: Open Data Commons PDDL License
  - **Fulfilling Obligations**: Though the **PDDL 1.0** license does not require attribution, we will **credit** the dataset source and include a **link** to the data in our final report.

Data Pipeline

High-Level Overview:

The **ETL pipeline** extracts data from both the **FiveThirtyEight** and **Jamesqo** datasets, transforms it (by cleaning and reformatting), and loads it into a structured format (CSV files and SQLite databases) for further analysis.

- **Technologies Used**:
  - **Python**: For scripting the ETL pipeline.
  - **Pandas**: For data manipulation and transformation.
  - **SQLite**: For storing transformed data in a structured format.
  - **Requests**: For downloading data files.
  - **Tarfile**: For extracting compressed data files.

Transformation and Cleaning Steps:

- **For FiveThirtyEight**:
  - Dropped rows with missing values in critical columns (intent, age, sex, race, place, education).
  - Filled missing age with the median and standardized text fields (intent, race, place, education).
- **For Jamesqo**:
  - Dropped rows with missing critical columns (date, state, city_or_county, n_killed, n_injured).
  - Converted the date column to datetime format and extracted year and month.
  - Filled missing latitude, longitude, and gun_stolen with default values (0 and 'unknown').

Challenges and Solutions:

- **Missing Data**: Missing values in non-critical columns were handled by filling with default placeholders (e.g., unknown for text fields and 0 for numeric fields).
- **File Format Handling**: Extracting the .tar.gz file for the Jamesqo dataset required using Python's **Tarfile** module, which was successfully implemented to extract and load the data.

Meta-Quality Measures:

- **Error Handling**: The pipeline is designed to handle errors such as missing files or incorrect formats, ensuring robustness.

- **Data Integrity**: All transformations ensure that the integrity of the original data is maintained, especially when filling missing values or converting formats.

Result and Limitations

Output Data:

- The cleaned data is stored in both **CSV files** and **SQLite databases** for easy querying and analysis.
- **Data Structure**: The final datasets contain columns like state, city_or_county, year, month, n_killed, n_injured, providing the necessary structure for geographic and temporal analysis.

Data Quality:

- The data quality is high after cleaning, but there may still be some residual missing data in non-critical fields.
- **Limitations**:
    - **Geographic Bias**: Areas with more media attention or higher populations may have more reported incidents, leading to potential bias in hotspot analysis.
    - **Temporal Completeness**: Some data may be outdated, and reporting delays might affect the analysis of recent incidents.

Data Format:

- **CSV** format was chosen for its simplicity and portability.
- **SQLite** format is used for efficient querying, especially for large datasets.

Critically Reflecting on Data:

- **Data Completeness**: Some underreporting may affect the completeness of gun violence data in certain areas.
- **Bias**: Areas with more frequent news coverage of gun violence may report more incidents, leading to potential geographic bias.

Conclusion

This project successfully implements an **ETL pipeline** to process and clean gun violence datasets from **FiveThirtyEight** and **Jamesqo**. The analysis will help identify geographic hotspots for gun violence and provide insights into the temporal escalation of incidents. The data pipeline ensures that the data is cleaned, structured, and ready for analysis to provide meaningful insights for policy and intervention planning.