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

Mirza Shaheen Iqubal
22998316

This project analyzes gun violence patterns across the United States, with a focus on identifying geographic hotspots and trends in incident repetition. By examining key factors such as location, time, and demographics, the analysis seeks to provide actionable insights that can inform and optimize public safety strategies. An automated ETL pipeline was developed to efficiently clean, transform, and process large datasets, ensuring high data quality and improving processing speed. The findings are intended to enhance resource allocation, support law enforcement efforts, and inform targeted interventions in high-risk areas. By identifying patterns of escalation and recurrence, the project contributes to efforts aimed at reducing gun violence. Future work will focus on integrating real-time data streams and implementing dynamic normalization techniques to improve the system's adaptability and accuracy over time.

I. 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?

II. Data Sources

For this study, I selected two comprehensive datasets that provide in-depth insights into gun violence incidents across the United States: the FiveThirtyEight Gun Deaths Dataset [1] and the Jamesqo Gun Violence Incident Data [2]. These datasets are publicly available and offer detailed information on gun-related deaths and incidents, encompassing key factors such as the demographics of the individuals involved (age, race, sex), the location of the incidents, and the number of fatalities and injuries.

The reason these datasets were chosen is due to their rich content, particularly the inclusion of demographic and geographic details, which are essential for understanding the patterns of gun violence. Additionally, these datasets provide valuable contextual data such as the intent behind the incidents, which helps in analyzing the broader social and environmental factors contributing to gun violence.

A. Data Structure

The FiveThirtyEight Gun Deaths Dataset is structured with a combination of temporal, categorical, and continuous variables. The dataset includes:

1. **Temporal Variables:** year, month.
2. **Categorical Variables**: intent (the intent behind the gun death such as homicide, suicide, accidental), sex, race, place, education.
3. **Continuous Variables:** age (the age of the individuals involved), hispanic (indicating Hispanic ethnicity).

The dataset contains no missing values for critical columns such as *intent, age, sex, race, place, and education*. Non-critical columns like *place and education* were cleaned by replacing missing values with *'unknown'* or the median, ensuring that the dataset retains its high quality for analysis.

The Jamesqo Gun Violence Incident Data follows a similar structure with temporal, categorical, and continuous features:

1. **Temporal Variables:** *date, year, month.*
2. **Categorical Variables:** *state, city_or_county, gun_stolen, incident_characteristics*.
3. **Continuous Variables:** *n_killed (number of fatalities), n_injured (number of injuries), latitude, longitude, n_guns_involved (number of guns involved in the incident).*

This dataset also contains no missing values for critical columns such as *date, state, city_or_county, n_killed, and n_injured*. Non-critical fields *like latitude and longitude* had missing values filled with zeros, maintaining the dataset's integrity for geographic analysis.

B. Data Quality

In this analysis, I have looked at the Jamesqo dataset and the FiveThirtyEight dataset, focusing on their data quality through key dimensions: accuracy, timeliness, and relevance. To evaluate these, I used histograms and Kernel Density Estimation (KDE) plots to visually examine the distributions of both categorical and continuous features.

The Jamesqo dataset accurately reflects real-world gun violence activities, with key features such as location, incident characteristics, and casualty counts. Accuracy is ensured as the data reflects incidents of gun violence across various regions. However, since the data spans from 2018 to 2021, it may not fully represent current trends. While the dataset provides valuable historical insights, it may not capture recent changes in patterns of gun violence.

Regarding timeliness, the dataset covers multiple years, which helps to understand the long-term trends in gun violence. However, as the dataset is not the most up-to-date, it might not fully reflect current conditions. Relevance is strong, as the dataset includes key features like location, incident characteristics, and demographics that are essential for understanding gun violence.

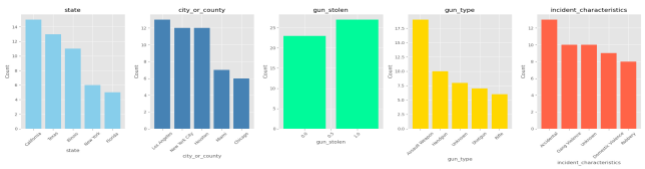

Fig. 1. Categorical features from the Jamesqo dataset.

Looking at the histograms for categorical features (Fig. 1), I noticed that the dataset is fairly balanced. For example, the state feature has realistic distributions, with California and Texas having the highest frequency of incidents, while gun_stolen is mostly False, as expected. The incident_characteristics feature shows a reasonable spread between Accidental, Gang Violence, and other types of incidents, which makes sense in the context of gun violence.

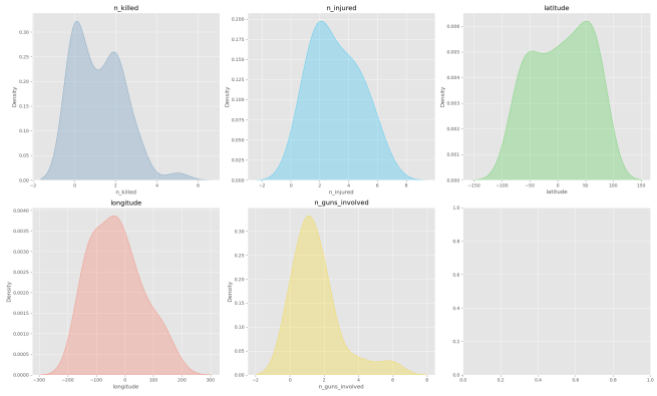

Fig. 2. Continuous features from the Jamesqo dataset.

When I checked the KDE plots for continuous features (Fig. 2), I found that features like *n_killed and n_injured* follow a skewed distribution, which is typical for count data. There are many incidents with fewer casualties, but a few cases with a larger number of deaths or injuries. However, the longitude feature shows values that are outside the valid geographic range (i.e., greater than 180 or less than -180), suggesting that there may be outliers or data entry errors that need attention.

The FiveThirtyEight dataset also provides valuable insights into gun deaths, with features like *intent, age, sex, and education.* The data is accurate, as it reflects real-world patterns of gun deaths across the U.S. However, as the data covers the years 2018 to 2021, it might not fully capture the most recent trends in gun violence.

In terms of timeliness, the data is helpful for understanding trends in gun deaths over the past few years. However, it would be better if it were updated to reflect more recent conditions. The dataset remains relevant as it focuses on essential factors such as demographics (age, sex) and incident-related data (intent, place).

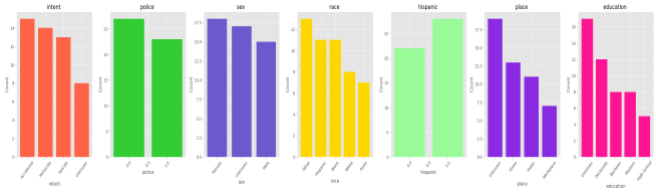

Fig. 3. Categorical features from the FiveThirtyEight dataset.

I analyzed the histograms for categorical features (Fig. 3) and found that most features are fairly well-balanced, though some categories like intent (with Accidental being the most common) and *sex* (with a slight bias towards Female) show slight imbalances. This is typical in datasets dealing with gun deaths, where certain types of incidents (like Accidental) are more common, and the sex ratio reflects real-world trends.

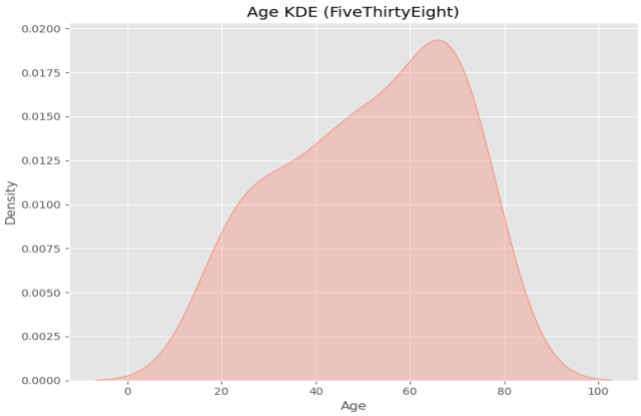

Fig. 4. Continuous features from the FiveThirtyEight dataset.

Looking at the KDE plot for continuous features (Fig. 4), I found that the age feature approximates a normal distribution, with the highest density in middle-aged groups, which is consistent with patterns seen in gun-related deaths. Other continuous features, such as *n_killed and n_injured*, are skewed, meaning that most incidents involve fewer casualties, but some extreme cases involve more.

III. Data Pipeline

In this project, I applied an automated ETL (Extraction, Transformation, and Loading) data pipeline using Python, which pulls the datasets from the internet, transforms and cleans them, and saves the results for further analysis. The pipeline ensures that the data is properly extracted, cleaned, and stored for easy retrieval and analysis.

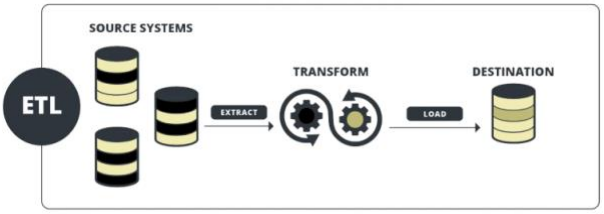

Fig. 5. The ETL data pipeline architecture [3].

A. *Data Extraction*

The extraction phase begins by downloading the datasets from the internet. I used the requests library to pull the data files in CSV format. These files were provided as zip archives. After downloading, I utilized Python's ZipFile module to extract the relevant CSV files. Then, I used pandas' *read_csv()* function to load the data into DataFrame structures for further processing..

B. *Data Transformation and Cleaning*

Once the data was extracted, I performed several transformation and cleaning steps:

- **Removing Unnecessary Columns**: I removed irrelevant columns, like the **index** column, from both datasets.
- **Handling Missing Data**: Rows with more than **three missing values** were dropped. For other missing values, I applied **backward fill (bfill)** to ensure consistency.

- **Reformatting**: I converted date columns to **datetime** format to improve time-series analysis.

For these transformations, I used the pandas module extensively, as it provided powerful functions for cleaning and transforming the data..

C. *Loading Data into the Sink*

Once cleaned, I saved the transformed data in two formats:

- **CSV files** for easy access and further use.
- **SQLite database** using **SQLAlchemy**, storing the data in two tables: **"FiveThirtyEight"** and **"Jamesqo"**.

This process ensures the data is structured, cleaned, and available for future analysis.

During the development of the pipeline, one challenge I faced was managing the temporary storage of downloaded files on the local machine. To solve this, I automated the pipeline to delete the downloaded files once they were processed and saved into the SQLite database, ensuring efficient storage management.

However, a limitation of the pipeline is its inability to adapt to changing data. For instance, continuous features like **latitude**, **longitude**, and **n_killed** in the **Jamesqo dataset** are **Min-Max normalized**, but new samples cannot be normalized automatically as the minimum and maximum values are unknown. To address this, I plan to implement dynamic normalization, where the pipeline will recalibrate the minimum and maximum values for each new batch of data, ensuring consistent scaling for incoming samples.

IV. Result and Limitations

A. Output Data of the Pipeline: The output of the data pipeline is an SQLite database containing two distinct tables: "FiveThirtyEight" and "Jamesqo", which hold the cleaned and transformed data from the datasets. After processing, the datasets are free of significant missing values, ensuring data completeness. These tables are structured with a mix of temporal, categorical, and continuous variables, ensuring that all relevant factors related to gun violence and gun deaths are covered comprehensively.

B. Why SQLite?: I chose SQLite as the output format for the pipeline because it is lightweight, portable, and easy to integrate with languages like Python and R. The compact size of SQLite makes it ideal for handling moderate-sized datasets, and it is widely used in the industry for small-to-medium scale applications. Additionally, SQLite databases are portable, meaning they can easily be shared with collaborators or used in different environments without requiring complex setups.

C. Critical Reflection on Data and Potential Issues: While the datasets in the pipeline are mostly accurate and reflect real-world gun violence and gun death incidents, there are some potential issues that may affect the analysis:

- **Normalization of New Samples:** Features like latitude, longitude, and n_killed in the Jamesqo dataset are Min-Max normalized, but new samples can't be normalized without the original min-max values, leading to potential inconsistencies.
- **Cross-Dataset Variability**: The datasets span different years (2011-2012 and 2018-2021), which makes it challenging to compare trends over time, especially given potential changes in policies or societal factors.
- **Differences in Data Collection Methods**: The datasets come from different sources, so variations in data collection and reporting methods could affect the consistency and accuracy of the analysis.
- **Data Coverage:** Features like race, age, and intent may seem less important but could provide valuable insights into patterns of gun violence and gun deaths.
- **Missing Current Trends:** Since the datasets are historical, they may not capture the current state of gun violence in the U.S., limiting their ability to detect recent trends or predict future incidents.

V. Conclusion

In conclusion, the data pipeline successfully managed the extraction, transformation, and loading (ETL) of gun violence data from the **FiveThirtyEight** and **Jamesqo** datasets using Python and the **ETL** architecture. By utilizing **SQLite** for data storage, I ensured efficient, portable, and easy integration of the data for analysis. While the pipeline was effective in handling structured datasets and ensuring data consistency, it does face limitations, such as the inability to normalize new samples due to the lack of original Min-Max values. Despite these limitations, the pipeline provides a strong foundation for analyzing and understanding gun violence patterns.

References

[1] FiveThirtyEight, "Gun Deaths Dataset," GitHub, 2018. [Online]. Available: https://github.com/fivethirtyeight/guns-data. [Accessed: Nov. 2024].

[2] Jamesqo, "Gun Violence Incident Data," GitHub, 2018. [Online]. Available: https://github.com/jamesqo/gun-violence-data. [Accessed: Nov. 2024].

[3] "Data Pipeline Architecture - A Deep Dive — StreamSets," Software AG. (accessed Jun. 03, 2024).