Analysis of the Impact of Hate Crimes on Police Shootings in the United States

MD BADHON MIAH - 23078804

1 Introduction

The police shootings are increasing day by day in the USA. The main aim of this research is to explore if the cities with high hate crime rates in the USA experience more police shooting incidents. Additionally, factors such as the racial identity of a person, gender, or age have any influence on hate crimes and police shootings. To do this, two datasets have been selected - one contains the hate crimes data in the states of the USA, and the other dataset contains the police shooting records.

2 Question

Do US states with high hate crime rates experience higher rates of fatal police shootings?

3 Data Sources

3.1 U.S Police Shootings 2013-2020 [1]

- Description: This dataset contains us police shooting data from 2013 to 2020.
- Data Structure: This dataset contains total 9536 number of records in tabular format. There are several significant features such as victim race, victim age, victim gender, date of incidence, city and state. The dataset is complete and there are only a few null values.
- License: This dataset is under cc0: public domain license [3], which is open to use.

3.2 FBI Hate Crimes in the USA (1991–2020) [2]

- Description: This dataset contains hate crime information across US states from 1999 to 2020. However, the dataset is filtered to 2013 to 2020 for this research to sync with the first dataset.
- Data Structure:This dataset is also in tabular format, it contains total 426968 number of records with many potential features for our analysis such as OFFENDER RACE, REGION NAME, INCIDENT DATE etc.
- License: This dataset is under U.S. Government Works [4]license, which is open to use.

3.3 Data Quality

- Accuracy: The datasets contain real-world police shootings and hate crimes incidents.
- Completeness: Both datasets contain all necessary information to conduct this research.
- Consistency: The data format is consistent across all records and columns.
- Timeliness:The datasets cover incidents from 2013 to 2020.
- Relevancy: The datasets focus on hate crimes and police shootings, relevant to the research question.

4 Data Pipeline

In this research, python has been used to build a data pipeline. ETA(Extraction, Transformation and Loading) is one of the popular data pipeline architecture, that has been used to build the data pipeline. The data has been pulled kaggle, then performed transformation step to clean the dataset and provide a shape that can be usable. Finally, cleaned data have been stored to the data folder in SQLite database.

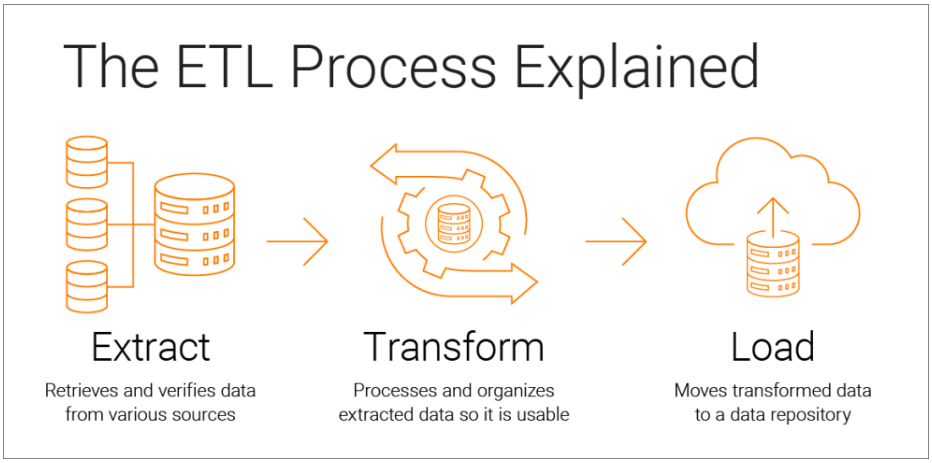

Figure 1: ETA Data Pipeline

4.1 Data Extraction

Both datasets have been fetched from Kaggle using the KaggleApi library. This library is responsible for authenticating the valid Kaggle user and downloading the static data in zip format. The datasets have been downloaded to a temporary directory using the tempfile library. Finally, CSV files are read through the pandas library and stored in a dictionary as DataFrames for further processing.

4.2 Data Transformation

The extracted datasets are loaded into a Python dictionary, and these DataFrames have been preprocessed to give a usable shape. Firstly, the second dataset (FBI Hate Crimes in the USA (1991–2020)) contains separate datasets for each state's crime data; these datasets have been merged into a single DataFrame. As mentioned in the data sources section, the second DataFrame is filtered to include data from 2013 to 2020 (108836 instances) for further analysis to align with the first DataFrame. Then, the features from both DataFrames that don't contribute, such as IDs and names,

have been dropped to avoid any biased correlation. Some columns, such as "ADULT VICTIM COUNT," had most of their values as null; these types of columns have been removed. On the other hand, some features contained only a few null values, which have been replaced by the mean or median for numerical and categorical features, respectively. Additionally, the instances containing two or more null feature values have been removed from the DataFrame.

4.3 Data Loading

The data pipeline ends with storing the transformed data into the SQLite database. There are a total of 2 DataFrames for this research from two different sources as mentioned earlier. Both DataFrames are stored as tables inside the database in the data repository.

4.4 Quality

This data pipeline handles any future errors related to updating the dataset by the author. For instance, attention has been provided while dealing with features and file names. The file names and the number of datasets from each data source are dynamically handled based on the names retrieved from the sources. Additionally, while dropping the columns, attention has been provided to ensure that if any features are deleted or renamed, no errors occur while running the pipeline. To achieve this, the 'errors' parameter of the 'drop' function is set to ''ignore''.

4.5 Challenges and Solution

The main challenges during building this pipeline were not getting any direct dataset link that could be used to fetch the data. To overcome this, brief research was conducted on how to fetch data from Kaggle. There are many techniques available to fetch data from Kaggle. The KaggleApi has been used in this pipeline, which has already been discussed briefly in the data extraction section.

5 Conclusion

The outcome of this data pipeline is storing the usable data into the SQLite database. As there are two data sources for this analysis, the data has been stored in two tables as two different data frames inside the database in the data repository.

References

[1] James Vandenberg. (2013–2020). US Police Shootings Dataset. Retrieved from `https://www.kaggle.com/datasets/jamesvandenberg/us-police-shootings-20132020`

[2] Jonathan Revere. (1991–2020). FBI Hate Crimes in USA Dataset. Retrieved from `https://www.kaggle.com/datasets/jonathanrevere/fbi-hate-crimes-in-usa-19912020`

[3] Creative Commons. CC0 1.0 Universal Public Domain Dedication. Retrieved from `https://creativecommons.org/publicdomain/zero/1.0/`

[4] USA Government. Government Copyright Information. Retrieved from `https://www.usa.gov/government-copyright`