Unveiling Connections: US Crimes and Mass Shootings Analysis

Introduction

This study aims to find the relationship between the US Crimes and the mass shootings analysis patterns in the different regions of the Americas. By the analyzing the historical data on US crimes and mass shootings analysis over a specific time period, the project will tell us or will show us how the mass shootings can greatly affect the US crimes ration in overall. Understanding this relationship is very crucial for assessing the crime ration of the desired city or state in the specific region of the America

Question:

What correlations exist between overall crime trends in the USA and mass shootings. Along with that how these insights guide law enforcement and policymakers in crafting targeted preventative measures for the desired cities or states.

Data Sources:

1. US Crimes Dataset:
    - Data URL : https://www.kaggle.com/datasets/johnybhiduri/us-crime-data?select=US_Crime_Data.csv
    - Description : This dataset contains the details of the all US crimes from the year 2017 to onward and I used it because it give me the macro level understanding of it.
    - Data Structure: Tabular format with the columns for title , organization , city , state, URL , summary.
    - Data Quality : The data is consistent , cleaned and contains all necessary information.
2. History of Mass Shootings in the USA Dataset:
    - Data URL : https://www.kaggle.com/datasets/rprkh15/history-of-mass-shootings-in-the-usa?select=History_of_Mass_Shootings_in_the_USA.csv
    - Description : This dataset contains the details of the all US shooting crimes from the year 1924 and this complements with the first data set to pluck out the information
    - Data Structure: Tabular format with the columns for date , city , state, dead , injured , total and description.
    - Data Quality : The data is consistent , cleaned and contains all necessary information with no null values.

Reasons for Choosing these Data Sources:

- Relevance : Both datasets are from the USA details , which made them highly relevant for this project needs.
- Coverage Period : As both data do contains the relevant time frame which I need to pluck out the information and going to use for the analysis (2017 - 2022).
- Open Data : Both of the datasets are the publicly available.

Licenses and Permissions

Both of the datasets are under open-data licenses, allowing the free usage of them.

Obligations:

To comply with this license, the report and any publications from this study will include the proper attribution to these open datasets, clearly indicating any modifications made to the original data.

- US crime Dataset: https://opendatacommons.org/licenses/odbl/1-0/
- History of mass shootings Dataset: https://creativecommons.org/publicdomain/zero/1.0/

Data Pipeline:

The data pipeline is implemented using Python leveraging the panda's library for data manipulation and saved the data in csv file.

Load Datasets: Load USA crime dataset and History of Mass Shootings in the USA Dataset from the Kaggle with the help of the Kaggle API.

Preprocess and Cleaning Data:

Dropped the irrelevant columns from the datasets which I don't need in the pipeline and not in the analysis of it. Also removed the null values and negative values if there are any.

Merging Data Frames: Merged the both preprocessed datasets on 'City' column which results in the consolidated dataset and remove the duplication of the data as well.

Export the Data: Saved the data in the form of the csv file for the future analysis.

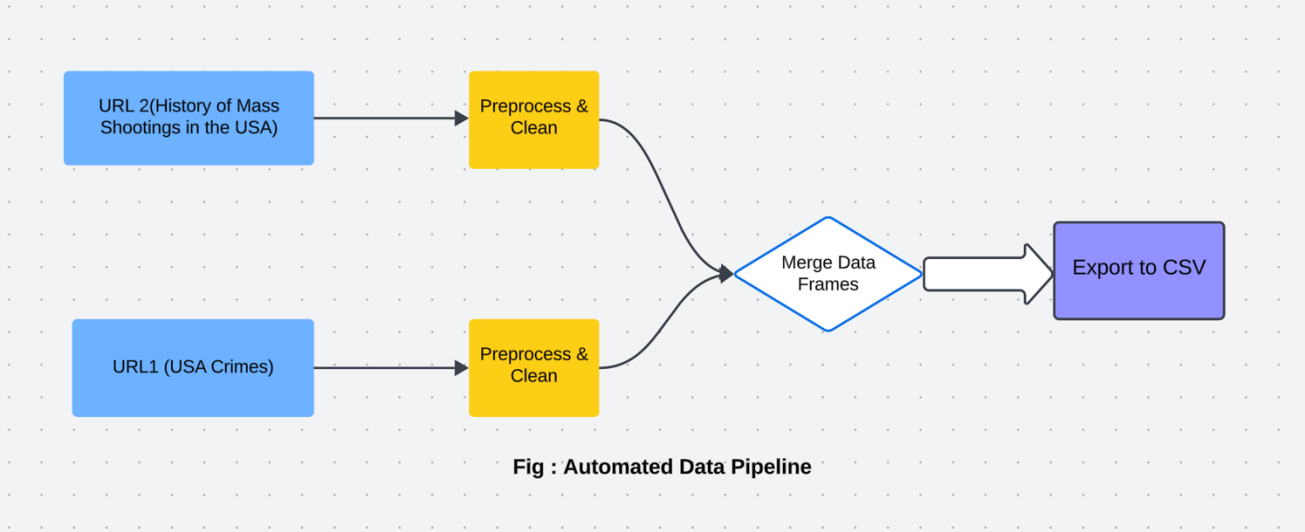

Fig : Automated Data Pipeline

Problems Encountered and Solved:

Data Cleaning: The USA crimes dataset does have the missing values inside of the data frame require careful clearing and I have made functions for it to handle them.

Data Merging: Merging the dataset on the basis of the 'City' column has given me the duplicates values which can affect in two ways, the very first increase the number of rows and second it effects badly for the data analysis aspect too.

Pipeline Smoothness:

- **API Handling:** My Pipeline uses the Tenacity for the retrying operations which I used to handle potential issues like network errors or timeouts when calling APIs.
- **Data Validation:** Ensuring that data meets the expected standards.
- **Error Handling:** Gracefully managing errors with logging and recovery mechanism.
- **Version Control:** Track changes and enabling rollbacks for seamless troubleshooting.
- **Feedback Mechanism:** Monitoring performance and adjusting processing logic for continuous improvement.

Data Quality:

- Accuracy: data reflects the real world and is correct.
- Completeness: contains all necessary information
- Consistency: is consistent in its format.
- Timeliness: emission of data is possible.

Data Storage:

The dataset is saved as the .csv file (final_merged_data.csv) for the analysis in the future.

Date	City	State	Dead	Injured	Total	Description
31-07-2022	Decatur	Illinois	1	3	4	Four teenagers, one of whom died, were shot at...
31-07-2022	Hartford	Connecticut	0	4	4	Four people, including a teenager, were shot i...
31-07-2022	Indianapolis	Indiana	0	4	4	Four people were shot after a fight in the Bro...
31-07-2022	Detroit	Michigan	1	7	8	One person was killed, and seven others were i...
31-07-2022	Orlando	Florida	0	7	7	Seven people were shot when someone opened fir...

Finally, the output of my data pipeline comprises processed data sets, insights or values. I ensure that it is a quality data through the validation mechanism. I choose CSV format for interoperability. Potential issues are data bias, drift and ethical concerns. In final report, I'll transparently represent limitations.