James Lee

09-06-2022

Data Engineering Project Write-up

Abstract

The goal of this project was to create an app that produces graphs and maps of gun violence incidents in USA. It started with development of data engineering pipeline, from data ingestion to data storage in SQLite database and deployment in Streamlit. Working with data from Kaggle, an app that shows overall death/injury of each state and city/county across different years in bar graphs and the locations of gun violence casualties in mapbox is generated. Users are able to select the state, city/county, and the year of interest. Ultimately, the app's objective is to inform the users who are traveling and/or looking to move into a new area more information so they can make more sound decisions.

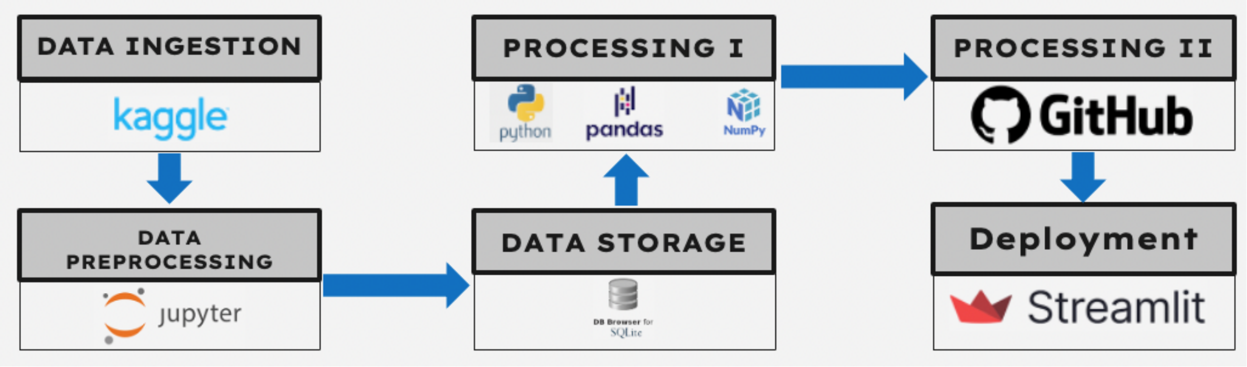
Design

Recently there have been more negative media coverages on gun violence. With safety of citizens in mind, this project was created to provide more information about gun violence to people who are traveling or looking to move into locations. While in search of the data, I found gun violence data in Kaggle. The author of Kaggle data had a similar interest of using the gun violence data to predict and prevent future incidents. Although the original intents of the author were to implement machine learning algorithm, the design of the project is simpler, at least to begin with. With the app, citizens will be informed of areas of high gun violence and be able to avoid and make plans accordingly.

Data

The original dataset from Kaggle contains 239,677 rows and 29 columns. After cleaning, the actual test dataset reduced down to 231,754 rows and 8 columns. Data are both numeric and string data types. Features include year, state, city/county, no. of killed, no. of injury, latitude, and longitude.

Data Engineering Pipeline.



1. Data Ingestion:
    a. Raw data is downloaded from Kaggle and saved into a project folder.
2. Data Preprocessing:
    a. Jupyter Notebook is utilized to:
        i. Call downloaded data (csv) from the project folder.
        ii. Clean data.
        iii. Export dataframe data to SQLite.
        iv. Import database from SQLite.
        v. Export the data to csv to be used for processing in Step 4.
3. Data Storage:
    a. SQLite is used as storage tool to receive dataframe, store, and export out to Jupyter Notebook for conversion to csv.
4. Processing I:
    a. Python script is utilized to:
        i. Manipulate data.
        ii. Analyze data.
        iii. Aggregate data.
        iv. Generate visualization.
5. Processing II:
    a. Github is utilized to:
        i. Store all the files and data.
        ii. Work as a cloud tool for Streamlit connection.
6. Deployment:
    a. Streamlit is utilized to deploy web application from Github.
    b. Streamlit app is open for public use.

Tools

- Pandas and numpy for data cleaning, manipulation, and analysis.
- SQLite for data storage.
- Text Editor for Streamlit web application python script.
- GitHub for cloud storage/processing.
- Streamlit for web application deployment in cloud.

Communication

- Write-up
- Powerpoint
- Presentation
- GitHub
- Streamlit