**EXTRACTION, TRANSFORMATION AND LOADING PROJECT**

Project title: Gun Violence in USA

Group members: Ikran Askar and Salieu Barrie

**Preprocessing**

Every year, tens of thousands of people are killed or injured as a result of gun violence in the United States. Gun violence has far-reaching consequences in America, affecting the lives of millions of Americans who witness it, know someone who has been shot, or live in fear of the next massacre. To aid decision-making, we want to provide analyzed data gun violence data for a sustainable solution to the current situation.

For the pre-processing exercise, we identified that we needed a sufficiently large data set that is also up-to-date. We were able to get three data sets from the Gun violence archives and Kaggle. The data were preprocessed to get only for 2022.

We also ensure that the that data sets have a common unique identifier to facilitate joining and analyses.

We confirmed that the dataset contains the information required to further analyses.

**Extraction**

We used 3 different datasets from the Gun Violence Archive website and Kaggle. The three files covers the period 2022 and contains the following information:

The fields of interest include the following:

1. Incident date
2. State
3. City/county
4. Number killed
5. Number injured

The datasets were extracted in in CSV format but needed cleaning (transformation).

The following sources for our datasets used:

<https://www.gunviolencearchive.org/mass-shooting>

<https://www.gunviolencearchive.org/accidental-deaths>

<https://www.kaggle.com/datasets/ecodan/school-shootings-us-1990present?select=pah_wikp_combo.csv>

**Transformation**

After extracting the data we performed the following steps to make the data useful/relevant for our purpose:

1. Used Pandas functions in Jupyter Notebook to load all three CSV files.
2. Inspected the columns to determine the ones of interest to the analyses team
3. Removed the operator’s column and the address column due to missing information on the accidental death and mass shootings csv files
4. Removed the incident ID and incident date because they are not relevant for further analyses.

For the school shooting data we dropped the columns not required for analysis ('Date','City','AreaType', 'Dupe', 'School','Source', 'Desc', 'Wounded').

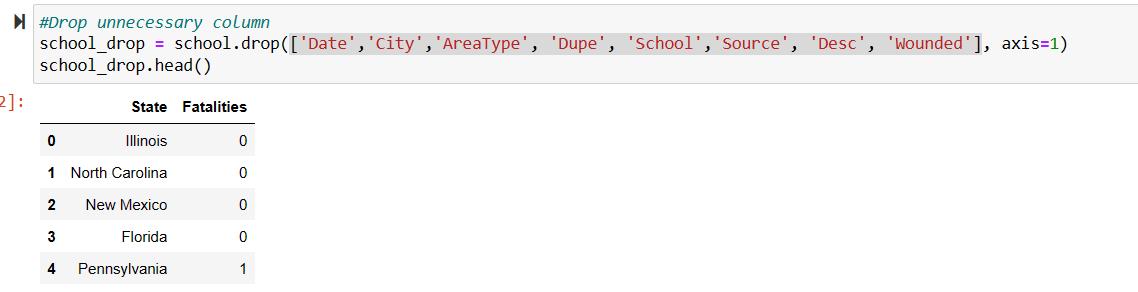
1. Changed the date format from object to datetime on the mass shootings and accidental death files
2. Renamed the columns to simplify the analysis process
3. We confirmed the data types to especially those related to numbers to confirm that they are integers.
4. Group the data by state and sorted in descending order so that we can preview to confirm that the data is set for loading to the database

**Graphical user interface, application

Description automatically generated**

**Graphical user interface

Description automatically generated with medium confidence**

****

**Load**

Following the transformation of the three CSV files and loaded them into the data frames, we did an initial connection to the Postgres database using PG admin to store our original clean data sets. We used the quick database website to create the initial table schema that got loaded into the Postgres database that generated the first set of tables. After running the queries and created the new tables with only the relevant information we reconnected to the database and generated additional tables for the data frames.

After uploading the three cleaned datasets to the database, we merged the tables to one document using ‘state’ as the primary key for each table. We now have a cleaned and merged data ready for further analysis

The QuickDBD table schema is shown below: Graphical user interface, text, application, chat or text message, Teams

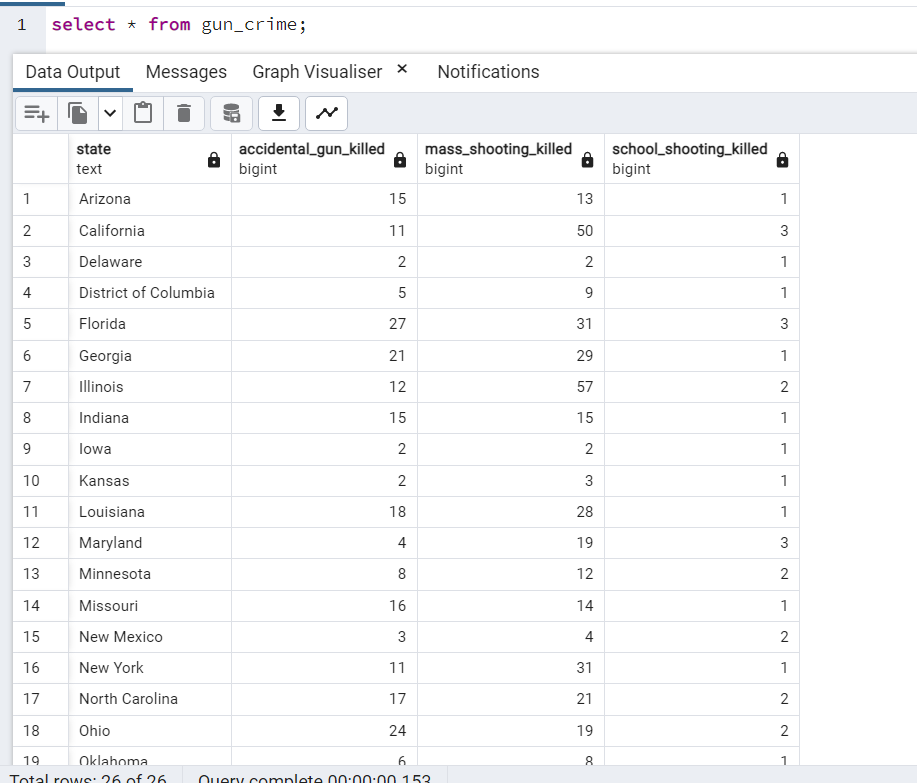
Description automatically generated

**The merged data in python**

Table

Description automatically generated

The loaded data in Postgres



**Summary**

We successfully extracted, transformed and loaded the data to Postgres database. The areas of focus of the mini-project were addressed. It was challenging to get all the datasets to match the duration. We found that 2022 is the only year common across all three datasets. We therefore preprocessed the data to 2022 to be able to do a precise analysis.