**Technical Report**

**Project Two**- Extract, Transform, and Load (ETL): Data on Shootings and Gun Violence in America

**Group Members:** Lynne Freilich, Ronnie Phillips, Corban Doran

**Conceptualizing the Development of our Project:**

1. Our team was required to create a case study that encompasses multiple datasets and their sources, then perform the Extract, Transform, and Load (ETL) process on the datasets.
2. The following requirements/goals were considered when determining which datasets to use:
   1. Ensure datasets have some commonality and are sufficiently large, each containing at least 1000+ rows of data.
   2. Collect at least 2 different file types (CSV, pdf, API, Spreadsheets, etc.) across datasets.
   3. Determine which, if any, unique identifier(s) we could be used as ‘Primary Keys’, or if we would use ‘Foreign Keys’ to assemble the data.
3. After preliminary research, and considering recent events such as the school shooting in Uvalde, TX, we decided to use the topical subject of gun violence in America for our case study.

**Project Outline:**

Tragically, mass shootings have plagued the United States of America for decades. The recent, horrible, school shooting at Robb Elementary School in Uvalde, Texas, in addition to the hundreds of other shootings that have occurred at schools, at work, in homes, or simply out in the streets, prompted our analysis of gun violence in America.

A picture containing diagram

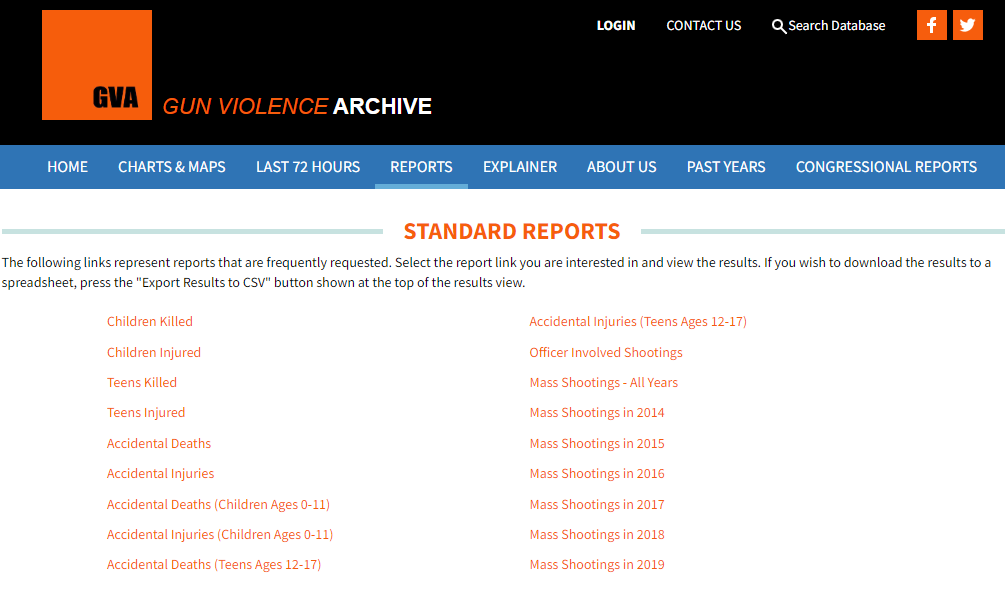
Description automatically generatedWe wanted to find out if gun control laws should in fact be girded more tightly and examine which locations could benefit the most from reform. Future analysis could provide insight into states that may need better gun safety and training due to frequency of accidental discharges resulting in injuries and/or death. It could also examine what correlation exists, if any, of wrongful transfers and gun-show loopholes, with mass shootings and other gun violence. It could surface states with higher instances of mass and/or criminal shootings. We wanted to compare the FBI’s monthly NICs background check data with gun violence records and examine how deliberate, malicious shootings stack up against accidental discharges and subsequent injuries or death. Datasets were selected, extracted, transformed, and loaded to enable future analysis.

Figure 1. Gun Violence ETL Project Overview

**Extracting the Data:**

First, we read and verified formats of multiple sets of data and ensured each row of data contained usable data. Next, we used selected datasets taken from multiple websites including the FBIs National Instant Criminal Background Check Division which includes data from 1998-2022 for all background checks made in the United States. Then, we juxtaposed the FBI’s information with multiple CSV documents from “Gun Violence Archive.Org”. Finally, we extracted an ‘Accidental Death’ dataset as a result from firearms, an ‘Accidental Injuries’ dataset as a result from accidental discharge or misuse of firearms, and a ‘Mass Shooting’ dataset that included information from every mass shooting in the United States since November 14th, 2018.

Because our data came from disparate sources (pdf, CSV, Spreadsheets, HTML tables), we used Pandas and Jupyter Notebook to extract the data, read in our CSV, and convert the pdf to CSV. Then we put them all in a Pandas dataframe to prepare for the next stage of our project: transformation.

****Figure 2. Gun Violence Archive.Org Report Website

**Sources for Extraction:**

<https://www.fbi.gov/file-repository/nics_firearm_checks_-_month_year_by_state_type.pdf/view>

<https://www.gunviolencearchive.org/reports>

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

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

**Transforming the Data:**

We utilized Pandas and Python to clean and structure our data. We read in the CSV files, then stored them as data frames. We then transformed the datasets by copying only the columns we needed in each respective CSV. For example, in the FBI NICs CSV file we looked only at: ‘State’, ‘Month’, ‘Handgun’, ‘Long gun’, ‘Multiples and Totals’.

****

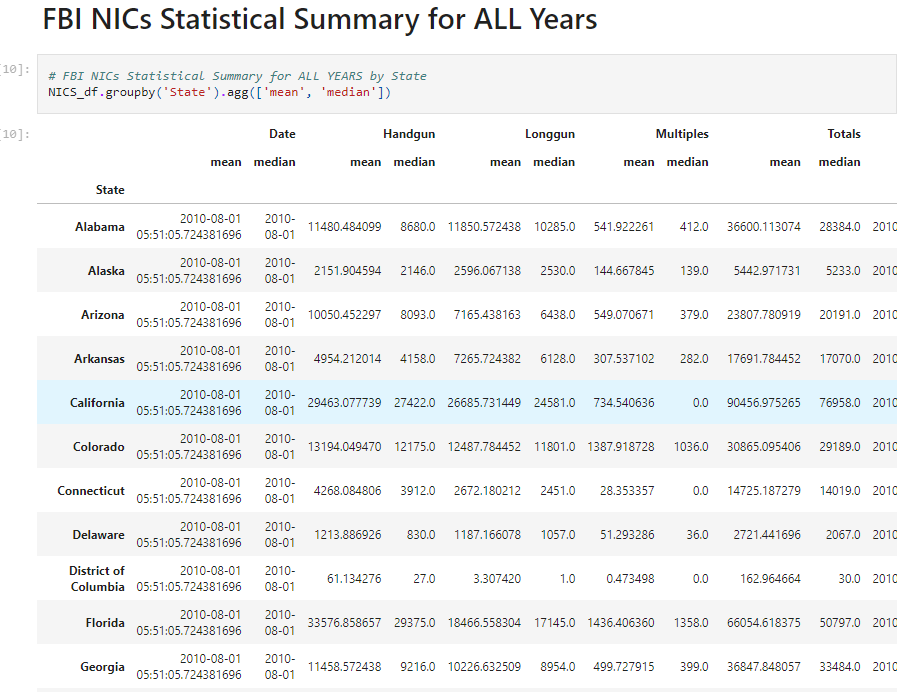
Figure 3. Pandas Renaming of Columns

**Table

Description automatically generated**After selecting our preferred columns, we removed any null value columns, renamed all unique columns, selected and performed queries on exclusive years, states, totals, incidents, deaths, etc. to identify any duplicates across all four data sets and then perform a statistical summary on each. We imported and used datetime series to look and individual Months and Years. We performed data wrangling, cleaning, filtering, and aggregation on all CSV files.

Figure 4. Datetime Series Transformation

We then sorted our data by year and by month so we could visually see which state had the highest number of accidental deaths, mass shootings, and accidental injuries, in addition to the FBI NIC’s background checks ran during the same period. Lastly, we set the index to either a previously created ‘Primary Key,’ or reset the index if no Primary Key was available.

****Figure 5. Statistical Summary by State

**Loading the Data:**

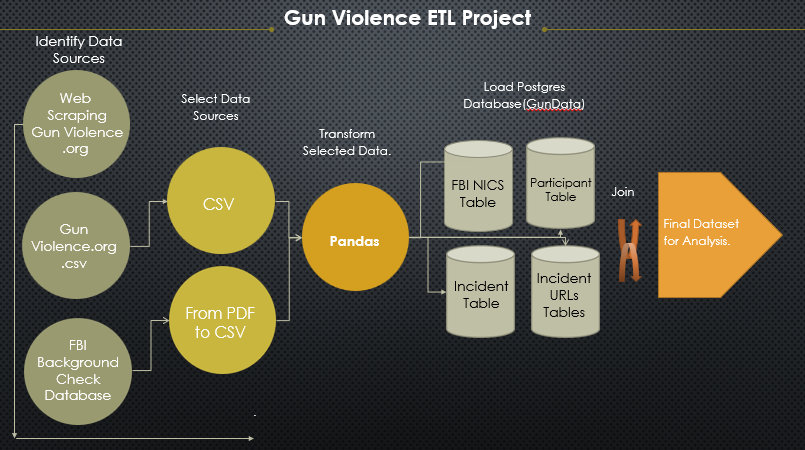
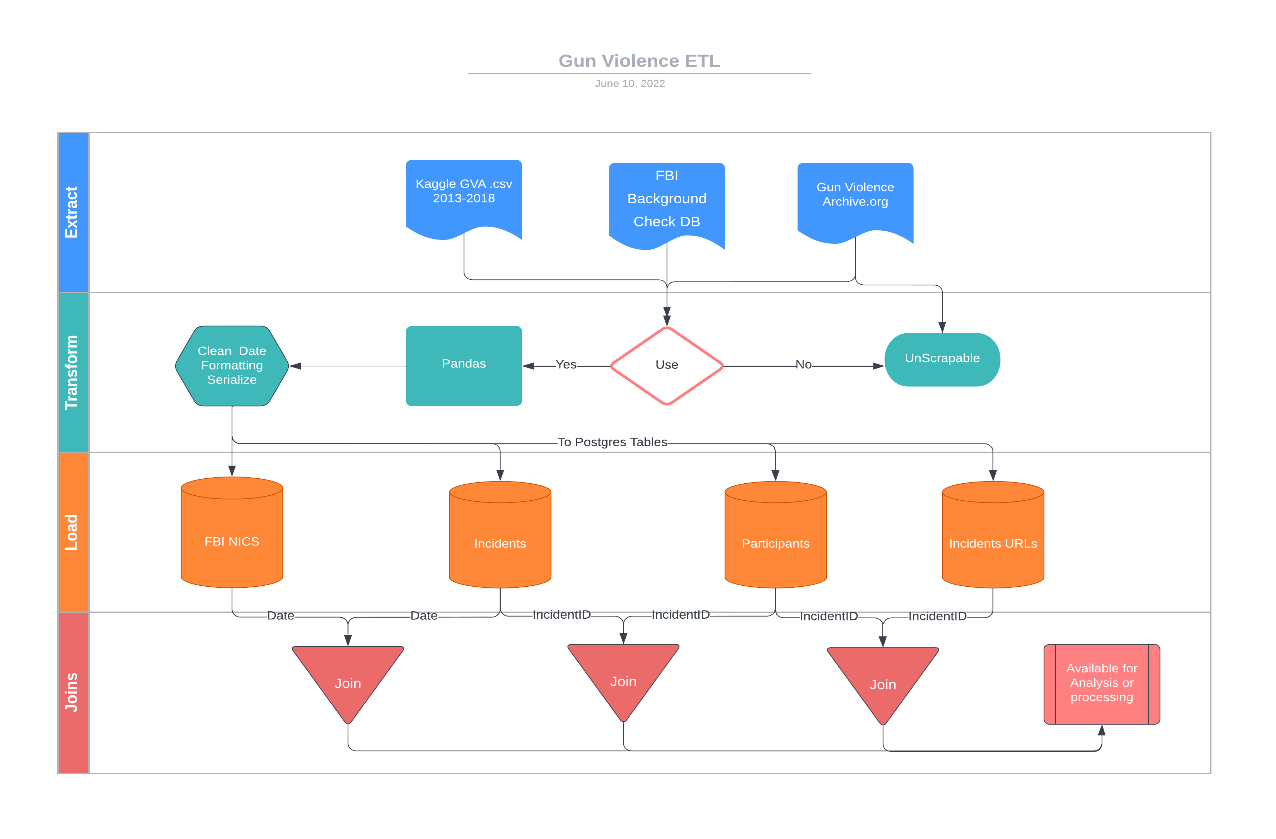
After the transformation was complete, we used SQLAlchemy to create the database, tables, and load the data. We used a Postgres database to load Incident, Participant, IncidentURL, and FBI NICs data. We used SQL to join all our database so that the final dataset can be used for further analysis.

Figure 6. Gun Violence ETL Project Flow Chart

**Summary:**

We encountered a limitation during the extract process. The website GunViolenceArchive.Org has implemented several very effective safeguards against web scraping. Several methods were attempted, including multiple python libraries both with and without browsers, rolling IP’s, and a variety of header changes, but did not find a solution. However, we believe our data is complete, and that future analysis efforts will not be adversely affected by the exclusion of that data.

Additionally, we took some outside-of-scope steps on this project, choosing to include all column names (after renaming and cleaning) to increase the usability of this dataset in future projects, as it contains a variety of useful information for analysis. Though exploratory data analysis is outside the scope of this ETL project, the information provided with our data extraction, transformation, and subsequent joining could constitute an interesting project on gun control. This could provide insight into states that need to improve gun safety, education, and training to decrease the occurrence of accidental deaths and/or injuries. More importantly, we hope to identify states that are in desperate need of gun reform by examining different factors. We hope to then analyze policy in those select states and provide gun reform recommendations, to decrease the occurrence gun-related injury and death.

****

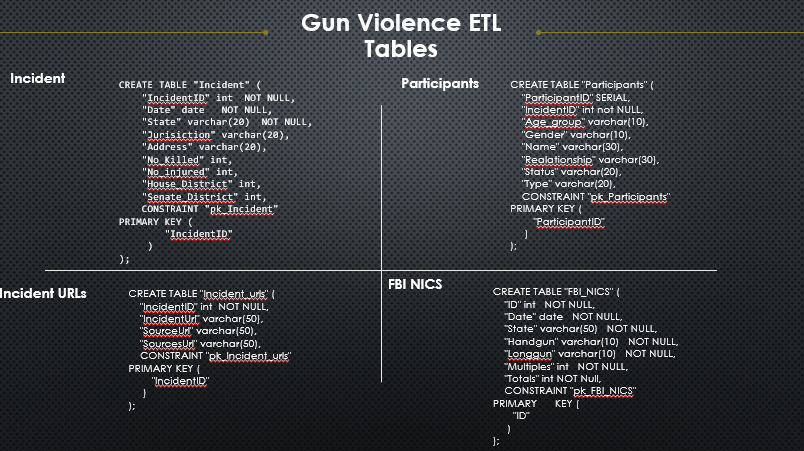
****Figure 7. Gun Violence Summary Chart

Figure 8. Code Summary