Project 2 - ETL

By: Nikunj Patel, Robert Gonzalez, Solomon Rotich.

**Project Proposal**

**Project Description**  
In the recent past, we have witnessed an increase in gun violence and mass shootings. For this project, we were interested in available data to aid in the hypothesis and analysis of this unfortunate emerging trend.

Below are the hypotheses that could be analyze from the processed data in this project.

H0: There is no direct correlation between the amount of gun permits and the amount of mass shootings per state.

Ha: There is a direct correlation between the amount of gun permits and the amount of mass shootings per state.

**Pre-Processing**

Our three objectives in this project:

1. Extracting the data from their existing locations.
2. Transforming the data (i.e. cleaning, joining, filtering, aggregating, etc).
3. Loading the data to a database (relational or non-relational).

The following table illustrates the observations and actions taken by the group to ensure a clean data set.

| **Pre-process Step** | **Data Need** | **Observation** | **Action** |
| --- | --- | --- | --- |
|  | 1000+ rows of data needed | 501 rows for each set (2 sets to be used) | None taken. |
|  | Unique identifier | Incident ID is the unique identifier for two of the data sets. However, the two dataset has different identifier | Resolved to use state as a unique identifier. |
|  | Verified formats of alphanumeric data and availability of all rows | Date, City, State and Address data | Inspect the data for null  Identified null and replace with (0)  Clean DF – by copying only relevant columns |
|  | Data for mass shooting per State | # Injured  # Killed | Read CSV  Create a clean DF. |
|  | Data for Background | # Permit per State | Read & create clean DF |

**Extraction**

We used 2 different datasets from the public platforms Kaggle and NCIS for a background checks dataset by state. The data in the two files included some of the following information:

* Mass Shootings data set

| **Incident ID** | **Incident Date** | **State** | **City Or County** | **Address** | **# Killed** | **# Injured** | **Operations** |
| --- | --- | --- | --- | --- | --- | --- | --- |

* Background checks data set

| **month** | **State** | **permit** | **permit\_recheck** | **handgun** | **long\_gun** | **other** | **multiple** | **admin** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |

The fields of interest include the following:

* Year
* State
* Number killed
* Number injured

\*Including data from 2016/2019 years.

The following sources for our datasets used:

<https://www.kaggle.com/gunviolencearchive/gun-violence-database>

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

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

<https://www.kaggle.com/datasets/pedropereira94/nics-firearm-background-checks>

**Transformation**

In order to transform the public data and use it in our study we performed the following:

* Used Pandas functions in Jupyter Notebook to load all CSV files.
* Removed the operator’s column and the address column due to missing information which was not relevant to the focus of this study.
* Merge the data frames and remove the duplicates.
* Created queries to capture the data we intent to use for analysis at the next level, as this project is only design for extracting data, transform then loading to the database.

Below is a preview of the two original dataset we used and the final merged dataframe.

Dataset for US Mass Shootings

Table

Description automatically generated

**Dataset for US Background Checks by NCIS**

Table

Description automatically generated

Combined dataset for mass shooting per state and total number of permits issued

Table

Description automatically generated

**Load**

After all the clean-up, filtering, and merging of the relevant data from the two separate datasets, we loaded them into a final clean data frame and then did an initial connection to the Postgres database using PG admin to store our final clean data set.

After running the queries and creating the new tables with only the relevant information, we reconnected to the database, and generated additional tables for the data frames.

**Postgres Database:**

**Table

Description automatically generated**

* *Reference: Jupyter notebook for merge\_mass\_shootings\_final.ipynb for more information on data clean-up.*

**Summary**

In this project, the goal was to focus on ETL and understand the process of preparing data prior to performing analysis. In extracting the data, we learned that it is possible to pull data from multiple sources of different formats to be used together. This is made possible by the transformation step. The transform step was the most time consuming in this project, as it involved writing several lines of code in Python to format and merge the different sources of data. Finally, the load step was crucial to prepare for analysis, as we know that it is essential for data to be part of an SQL database to perform analysis.

Due to the limited nature of the project, no analysis on the data was conducted at this time. We do, however, recommend that further analysis be done to determine if there is a trend to mass shooting and is there any relation to background checks.