**ETL-Extract-Load-Transform**

By: Mohamed Riahi

3 datasets were used in this project and were extracted from the public platform Kaggle that contained Gun Violence reports of deaths, injuries and mass shooting events among different community populations from March 2017 – July 2021

The final data set included

* Case ID
* Case date
* State
* City/county
* Number of Killed Victims
* Number of injured Victims

The following sources for our datasets used:

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

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

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

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

**Transformation**

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

The following steps were performed for data transformation process for public dataset.

* Used Pandas functions in Jupyter Notebook to load all three CSV files.
* Converted data into data frames
* Deleted few columns due to missing information; operator’s and address columns
* Used inner merge function to identify duplicated data on the Incident id among the three data sets
* Performed queries by grouping all data by State to get the sum of killed victims and injured victims accordingly to each state. The data frame was sorted in a descending order to get top states.

**Load**

In final step, a table was inserted using Postgres PG to match the final Data Frame and to store the the original clean data.

**Summary**

Data set selected for this project contained certain limitations with data collection process, however it was relevant to answer up to date general questions as the analytical process indicated