ETL Project

By: *Daneille Belliveau, Fatima Carrillo, Jeriel Tenorio, Hossein Esteaneh*.

**Pre-Processing**

Do states with high rates of police shooting have any relation to the high percentage of household gun ownership?

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

| **Pre-process Step** | **Data Need** | **Observation** | **Sources Links** |
| --- | --- | --- | --- |
|  | Fatal-Police-Shooting-data.csv | 2020 data that shows shooting | https://github.com/washingtonpost/data-police-shootings |
|  | GUN OWNERSHIP BY STATE 2020 | Gun ownership by state | https://worldpopulationreview.com/state-rankings/gun-ownership-by-state |

**Extraction**

We used 2 different datasets from the Washington Post and World population review. The data in the two files included the following information:

* Fatal Police officer shootings state
* Gun Ownerships by state

The fields of interest include the following:

* State
* Number killed

Including data from 2020.

**Final Technical Report**

After having identified our datasets, we performed ETL on the data and documented the following:

Data Transformation:

* We used Pandas package in Jupyter Notebook to load our CSV files.
* Reviewed files and transformed into Dataframes.
* Performed clean up on the database by removing columns. For the fatal-police-shootings-data we found 3 out of 17 columns relevant to the analysis. We removed the following columns: id, name, armed, age, gender, race, city, signs\_of\_mental\_illness, threat\_level, flee, body\_camera, longitude, latitude, is\_geocoding\_exact. We used all the columns in the gunowners\_data file.

Final Production:

* Utilizing PgAdmin we joined the two databases into one.
* We used the quickdatabasediagrams.com to create an entity diagram and exported schema data.
* Due to the data sources being structured (CSV files), utilizing a relational database (PostgreSQL) was optimal.