ETL Project

Final Report

Team Members: Michelle Davis, Marie Prosper, Novac Radovic

Project Objectives

The aim of this explore, transform and load (ETL) project is to provide a database for data analysts to analyze various relationship of gun violence and other socio economic factors.

Extract

For the given ETL project, we used three sets of data. The first dataset was on gun violence from 2013 - 2018. This data was taken from Kaggle at the following link. [Comprehensive record of over 260k US gun violence incidents from 2013-2018](file:///C:\Users\MDavi23\Documents\Git%20hub\Project-ETL2\Comprehensive%20record%20of%20over%20260k%20US%20gun%20violence%20incidents%20from%202013-2018). This data had over 260,000 rows and 29 columns. The data included information such as, gun type, participant age, state, and description of incident, etc.

The second data set was on 2016 and 2017 Poverty by Race/Ethnicity by the Kaiser Family Foundation at following link. ([Poverty by Race/Ethnicity KFF](https://www.kff.org/other/state-indicator/poverty-rate-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D). The dataset was too larges so we filtered on website for two years, 2016 and 2017. We decided on those years because more recent data on gun violence would be more relevant for the analysis required. The data was taken from US Census data.

The third data set was from population data for United States by State. This data was taken from following link. ([US Population](https://worldpopulationreview.com/countries/united-states-population) ). We wanted population to see how density affect gun violence.

Transform

The transformation process began by working on the csv data files mentioned in extract phase. First we imported csv file for gun violence and read into Dataframe in Jupyter Notebook. We filtered the pandas dataframe to select only columns needed. The dataset was narrowed from 30 columns to 4 columns, (date, state, number killed & number injured). Since this dataset had 2013-2018 data and we decided to study 2016-2017 were subset the dataframe. For the date field we split the column into 3 columns (month, day, year). From this we decided we only needed by year so we did a groupby (year and state) and aggregate by sum of number killed and number injured. After aggregation we reset the index to create new column id. Additionally, we created two data sets, 2016 &2017, gun violence incidents. These data sets were then exported to csv files.

Second we imported the poverty csv file and filtered out Footnotes column and Puerto Rico as a state row since we were not including Puerto Rico as a state. We then changed the NAN to zeros then save csv files for 2016 Poverty and 2017 Poverty.

Third, were imported the population csv and selected the columns needed (rank, state, population2018 and density). We chose population for 2018 because the dataset had every ten years of data and 2018 was closest. We did not feel there would be significant change in population so 2018 population data would be representative. We then created Id as index so we can have as a primary key in Postgres.

We then join all 5 tables by year and state.

Load

Finally, for the load phase we created a database using PgAdmin. We created five tables with matching column names and data types and created primary id. We created SQL Alchemy to connect to the database. We used pandas to import from pandas to the database. Check PgAdmin to make sure data was properly imported into the schema tables.

Next Steps

This database will now be handed off to data analysts. Some of the questions that can be answered from this database are as follows:

* How does population density affect gun violence?
* Is there a correlation between poverty and gun violence?
* Does the demographic percentage play a factor in gun violence?
* Is gun violence consistent between 2016 and 2017?
* Which states have the most gun violence per capita?