Project Title: Police Violence Relationship with Socioeconomic Conditions

Group Members: Robert Peron

ETL Project Technical Report

September 9, 2020

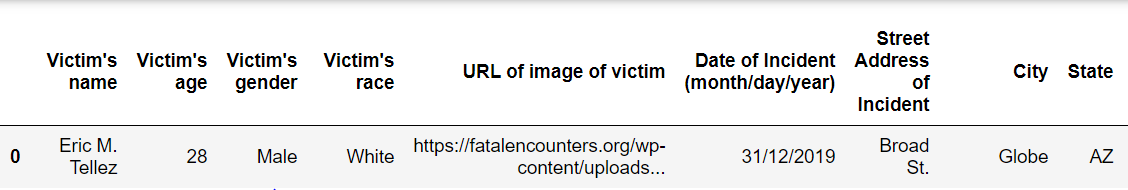
**Extraction**

I began my data extraction by researching two datasets from Kaggle provided in 2 csv files from the site. All data used is based on information and reporting across cities in the United States. The police killings dataset shows individual case events from 2013 to 2019. The census information is provided from the latest U.S. census. The sources for the datasets are as follows:

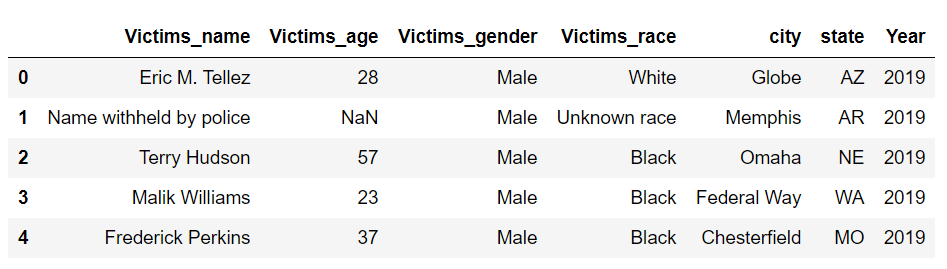
* US Poverty Census from Kaggle
* US Police Killings from Kaggle

**Transformation**

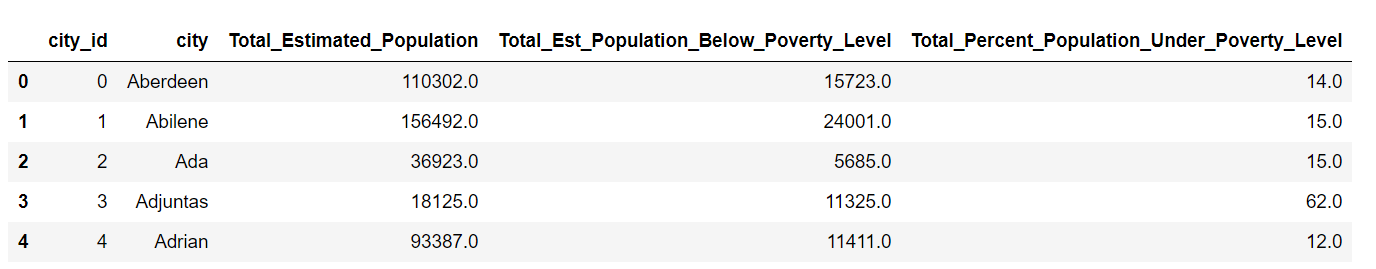
To begin cleaning both datasets, I loaded the csv files into Jupyter Notebook and using the panda’s library, I explored the datasets separately to identify any unnecessary information provided. For the police\_killings dataset, I first dropped the unnecessary columns from the dataset. I noted that a lot of the column headers would be hard to reference in code, so I renamed the column headers I pulled to simpler titles. As well, I dropped all rows that had NaN values in them as to not cause any issues later in merging. The final step was stripping the column titled “Date of Incident” of the Year, and creating a column titled “Year” to replace the column “Date of Incident” as seen in Figure 1 below:

**Figure 1.**

The cleaned dataset can be seen in Figure 2 below:

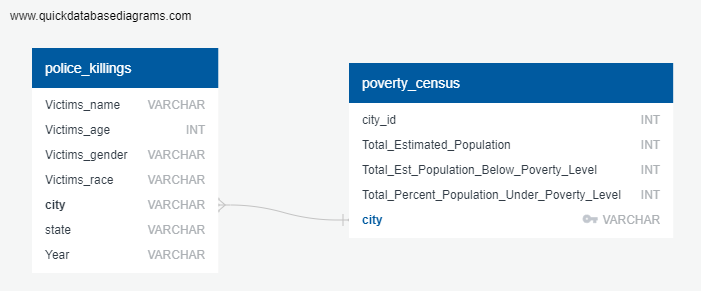
**Figure 2.**

For the poverty\_census dataset, I first had to rename the columns for ease of use. Once this was completed, I reviewed the columns and used a string split function in pandas to extract the City and State from the column “Geographic Area Name” and added those results as columns in the dataframe. After that, I noted that the cities still had small additional information that would affect the columns “Total Estimated Population”, “Total Estimated Population Below Poverty Level” and “Total Percentage Population Under Poverty Level”. To resolve this, I used a groupby function to group the rows by city to get the sum of “Total Estimated Population” and “Total Estimated Population Below Poverty Level”. From there, I calculated the new “Total Percentage Population Under Poverty Level” calculation and added the results to a column to replace the percentage column. The final dataframe can be seen in Figure 2 below:

**Figure 3.**

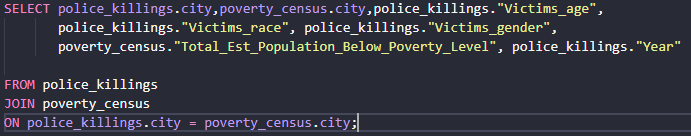
The final stage was merging the two datasets to ensure that there were no cities that were not common between both datasets. To ensure the quality of the data, I merged the two datasets on the “city” column, trimming the police\_killings data to only cities I had census information for.

**Load**  **Figure 4**.

The last step was to transfer our final output to a Database. I first created the GRD on QuickDatabaseDiagrams (seen in Figure 4 beside) to create tables that match the respective datasets.

I then used SQL Alchemy and Psycopg2 to connect to the database and load the tables. I performed queries as well to create an additional table that can reference a City’s census information with each police killing event in the city with age, sex and racial demographic information included (seen in Figure 5 below).

**Figure 5.**



**Summary**

I built this dataset to not only include the records of each cities police killings with demographic information on each case, but a table and structure where additional census information for each city can be added on to (education spending, median income, per capita income) and can be used to:

* Track Police Violence/Killings trends over time by city
* Research Proportion of what demographic groups face Police Violence
* Understand How Socioeconomic Measures relate to Police Violence
* Use as Reference for potential areas where Police Reform measures should be explored