ETL Final Report - Group 1

Aja Ould

Althea McMillian

Brittaney Marshall

Our group was tasked with sourcing data and loading it to a production database. We used the Pandas library for our data manipulation. We used a Jupyter notebook to import pandas and create SQL engine to connect to the final destination in postgres.

Created DataFrames using CSV files, US Gun Violence data from Kaggle and US Places of worship from DataWorld.

<https://www.kaggle.com/jameslko/gun-violence-data?select=gun-violence-data_01-2013_03-2018.csv>

<https://data.world/awram/us-places-of-worship>

We began looking at the data to understand points of overlap where we may be able to extract new values adding related data. Using a file found at Kaggle with US State data, saved it as a CSV so we could correlate abbreviations to states and create a US State Abbreviations DataFrame.

<https://www.kaggle.com/stansilas/us-state-county-name-codes/data>

Created a new gun table using US Gun Violence imported data and kept only the columns we were using state, the number of people killed, and the number injured by using the copy function.

Created a new worship table using US Places of Worship data and kept only the columns we were using by using the copy function, State Abbreviations, and ID.

Used groupby and count functions to gather abbreviations together.

Joined gun table with states DataFrame on the state column and produced our Full DataFrame and set the index on Abbreviation.

Joined using merge the worship table to the Full DataFrame on Abbreviation. Renamed ID from US Places of Worship DataFrame to Worship Numbers and set index on Abbreviation.

Created new columns, Total Average, Average Injured, Average Killed with new values by combining data from the existing datasets. Adding number of people killed and number injured then dividing them by Worship Numbers creating a Total Average.

Adding number injured then dividing by Worship Numbers created a Average Injured.

Adding number killed then dividing by Worship Numbers created an Average Killed.

Using our new DataFrame to find max values by max function of Average Injured and Average Killed.

Load

Created a CSV of our Full DataFrame with combined and relevant values.

We used the Abbreviation column from the US State data. We used the state, n\_killed, n\_injured from the US Gun Violence, and the worship numbers from the US Places of Worship CSV. We joined them and calculated the total average impacted by gun violence per state (both injured and killed) followed by averages of injured and killed. This data is what we loaded into the database.

Created a connection with the database and confirmed our table. We loaded our CSV into Postgres as our production destination system for user-friendliness.

Due to Github file size restriction, we deleted unused columns from the Gun Violence CSV.