DSC 540 Term Project Summary

Teresa Ferrill

Bellevue University

Abstract

This project is about cleaning, transforming, merging, and storing data from multiple sources then visualizing that data. Using a comma separated value (CSV) file, a website located table, and an application program interface (API) data source, like data comes together to complete an analysis of number of those killed and injured through gun violence in the United States between the years of 2013 and 2020. The analysis will provide multiple visualizations of various data brought together for presentation.

*Keywords:* CSV, API, clean, transform, merge, store, data analysis, visualization

DSC 540 – Data Preparation Term Project Summary

**Topic: Gun Violence in the United States**

This project sets out to review and analyze the number of those killed and injured by gun violence in the United States between the years of 2013 and 2020. Through a series of milestones that explore a comma separated value (CSV) file, and website table, and an application program interface (API) accessed file, this project conducts analysis of the counts of people killed and injured in the United States due to gun violence between the years of 2013 and 2020.

**What I Learned and Actions Taken**

The first milestone proved challenging. Not only did I have to identify a topic, the topic had to have multiple data resources that included multiple sources. I settled on the topic of gun violence and located a CSV, website, and API data source. Trough the use of these data sources, I was able to compile data covering a period of 8 years.

The second milestone of the project included obtaining data from a CSV file. The CSV file I used came from the Kaggle website and covered the years of 2013 through 2018 (Kaggle, n.d.). While completing the first milestone, I learned how to download a CSV file as a Pandas DataFrame, determine the data types ov the elements of the DataFrame, identify and drop rows containing columns that have no values in relevant data elements, and find duplicate rows within the DataFrame.

For the third milestone of the project, I obtained data from a table within a website. Using the Gun Violence Archive site, I acquired gun violence data covering the years of 2019 and 2020 (GVA, 2020). Through this milestone, I learned how to setup and use convert HREF information for inclusion in a table, break-up dates within a date column and separate into a day, month, and year column, and fix casing and inconsistent values within the dataset.

For the fourth milestone, I obtained access to an API data source from the Kaggle website (Kaggle, 2020). While executing this milestone, I learned how to leverage Beautiful Soup to access and acquire data through the use of a json file containing a personalized username and access key, determine unique values within a column in a DataFrame, and determine the existance of any NaN values within a DataFrame then delete them if found.

During the completion of the fifth milestone, I learned how to create an SQLite database then convert DataFrames into database tables then create SQL queries that join tables through the union and join commands. With the results of these queries, I then created multiple visualizations leveraging the Matplotlib and NumPy libraries.

Through the execution of each milestone, I was able to apply skills learned in one milestone to the next. In creating the final milestone, I prepared each of the dataset by importing, cleaning, and transforming them within the notebook. This ensured inclusion of the most current and up-to-date data for analysis. I was able to determine that Illinois had the highest number of those killed and injured across all years examined. I analyzed the data for the state and determined the city with the highest numbers was Chicago and its surrounding area. I then determined the next highest twenty cities and plotted their killed and injured values. Using pie charts, I was able to determine that, from year to year, the total number killed and injured in the United States vary equally from year to year.

References:

GVA. (2020). *Mass Shootings*. Retrieved from Gun Violence Archive: https://www.gunviolencearchive.org/

Kaggle. (2020). *Gun Violence Database*. Retrieved from Kaggle: https://kaggle.com/duttadbebadri/gun-violence-in-usa-insights-forecast

Kaggle. (n.d.). *Gun Violence*. Retrieved from Kaggle: https://www.kaggle.com/jameslko/gun-violence-data/data