Mariana Brinkley

5/18/2019

**EXTRACT:**

The data used for this project was sourced from three different websites. We gathered data from the FBI.gov, data.world, and Kaggle. After choosing our sources we downloaded, converted, and saved the data as CSV files, where we then used Excel to filter through the data and see how our data should be formatted. [[1]](#footnote-1)

**TRANSFORM:**

For cleaning and transforming our data, we altered the files and made data frames in Jupyter Notebook using Pandas in Python. Our largest CSV file from Kaggle had several years of data. We filtered the years to only show data from 2011 to match our other two sources, which displayed data from the same year. Our file from data.world had extra/unnecessary rows at the top. We removed these to make our data better to read. We also decided what columns to use and not use in our data frames to create a cleaner display of information. Once we picked our columns, we renamed all of them to transfer over to SQL easily.

**LOAD:**

The final database we created in SQL consisted of all three data sources as tables. We joined our FBI.gov table and our data.world table on a similar column to compare who committed the crimes and with what weapon.

1. https://data.world/ucr/expanded-homicide-data-tables

   https://www.kaggle.com/murderaccountability/homicide-reports

   https://ucr.fbi.gov/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/offenses-known-to-law-enforcement/expanded/expanded-homicide-data [↑](#footnote-ref-1)