Crime Data ETL Project Report

-Jack Stromberg

-Bridgette Sanders

Extraction:

We extracted the crime data from the City of Orlando (<https://data.cityoforlando.net/Orlando-Police/OPD-Crimes/4y9m-jbmz>) and the City of Baltimore (<https://data.baltimorecity.gov/Public-Safety/BPD-Part-1-Victim-Based-Crime-Data/wsfq-mvij>) open data sites and exported the data to two csv files, one for each city. The raw data set for Baltimore contained crime data dating back to 1963 and up until 2019, while the raw data for Orlando contained data from 2010-2017. Due to incompleteness in the two data sets, we decided to use the data from 2014-2017 for both cities. Each dataset contained varying columns and data types which needed to be made more uniform, in addition to cutting down the data to years 2014-2017.

Transformation:

Due to the variation between the two data sets, we needed to do some cleaning and transforming of our data. Our first step in this process was determining which variables were relevant between the datasets to standardize the data and to better facilitate analysis. The verbiage tended to be different between the data sets in addition to the different columns so we decided to cut down the data into the categories of crime date, location, crime description, and premise type. These variables were basis on which we cleaned and transformed our data and designed the tables of our database. We used jupyter notebooks and the Pandas library for the transformation process. Due to the differences between the data sets, different cleanups and transformations were needed for each. We imported our csv files and began manipulating our data, removing columns we decided were unnecessary, leaving the columns that fit under our categories of crime date, location, crime description, and premise type. We then added columns for City and State, which were missing from each individual dataset. We then reordered and renamed our columns to ensure both cities were organized the same and to correspond to the model of our database tables. For Baltimore, we needed to filter down the years which required changing the data type of the date column from object to datetime. This allowed for easier manipulation of that column and filtering of the data down to years 2014-2017 using conditional statements and .loc. For Orlando, the crime\_date column contained both date and time in the same column. So removal of the time portion was necessary to standardize the data. This was done using str.split to split out and remove the time portion of that column.

Load:

We used the sqlalchemy module for the loading process to an SQL relational database. We chose the use Postgres to load and house our data.