**ETL Project**

By Kasia Gorska

Nedal Swehli

Chandra Kent

GWU Data Analytics Bootcamp

1. **Purpose of the Project.**

Extract-Transform-Load (ETL) is the process by which data is extracted from data sources (that are not optimized for analytics), transformed to make it comprehensible and loaded into a target system (a database or a data warehouse). The purpose of our project is to perform and document the ETL process of the crime data in Arlington, VA, Washington, D.C., and Bethesda, MD.

1. **Data Extraction.**

The first step of the ETL process involved connecting to the source systems, and both selecting and collecting the necessary data required for analytical processing. We used 3 datasets from the counties websites:

1. Arlington, VA: <https://data.arlingtonva.us/dataviews/225891/police-incident-log/>
2. Washington D.C : <https://opendata.dc.gov/datasets/crime-incidents-in-2019>
3. Bethesda, MD: <https://data.montgomerycountymd.gov/Public-Safety/Crime/icn6-v9z3>

All of our crime data was based in Arlington, VA, Washington, D.C., and Bethesda, MD, ranging over various years from 1985 to 2019.

1. **Data Cleanup/Transformation.**

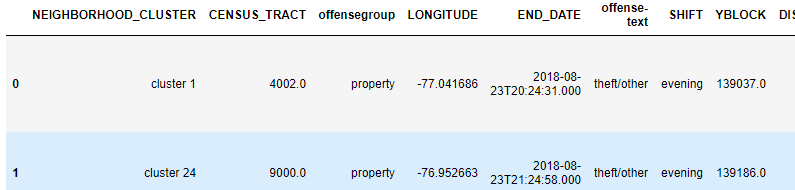
The second step in the ETL process involved data clean-up/transformation to convert it to a standard format.

Our steps in cleaning up the datasets involved analyzing them and determining which variables were not relevant. For all three datasets, we followed the following steps:

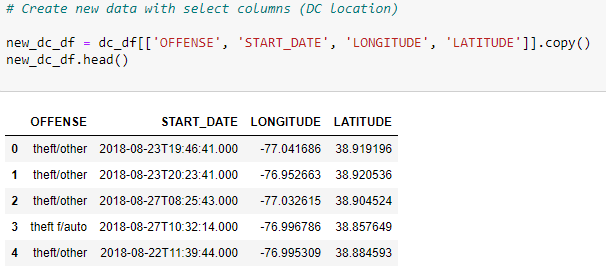
* Step 1 was to select relevant columns only;
* Step 2 involved trimming the data;
* Step 3 involved renaming the columns for better readability;
* Step 4 involved merging all three datasets into one file.

Since all three datasets involved the same cleanup/transformation steps, we only documented the cleanup process for crime data located in Washington, D.C. (see below).

**Figure 1. Crime Data in Washington, D.C. (original dataset).**



**Figure 2. Crime Data in Washington, D.C. (dataset with relevant data only).**



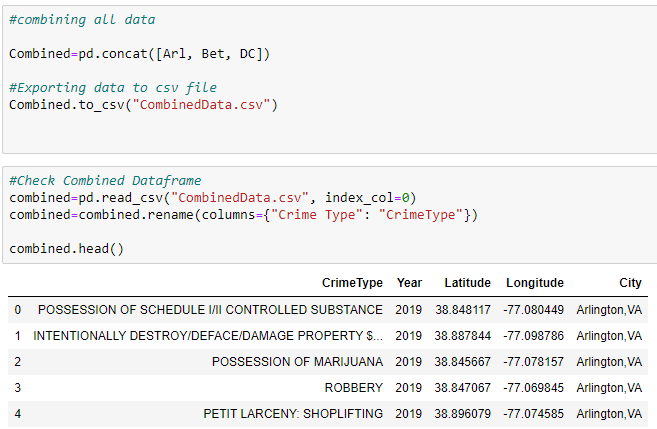
After selecting relevant columns, we trimmed the date data and renamed the columns for better readability (see Figure 3).

**Figure 3. Crime Data in Washington, D.C. (dataset with renamed columns).**



The last step of data transformation involved merging all three datasets into one csv file (see Figure 4), so it could be used for uploading it into SalesForce and PowerBi platforms.

**Figure 4. Final output (combined crime data in Washington, D.C, Arlington, VA, and Bethesda, MD).**



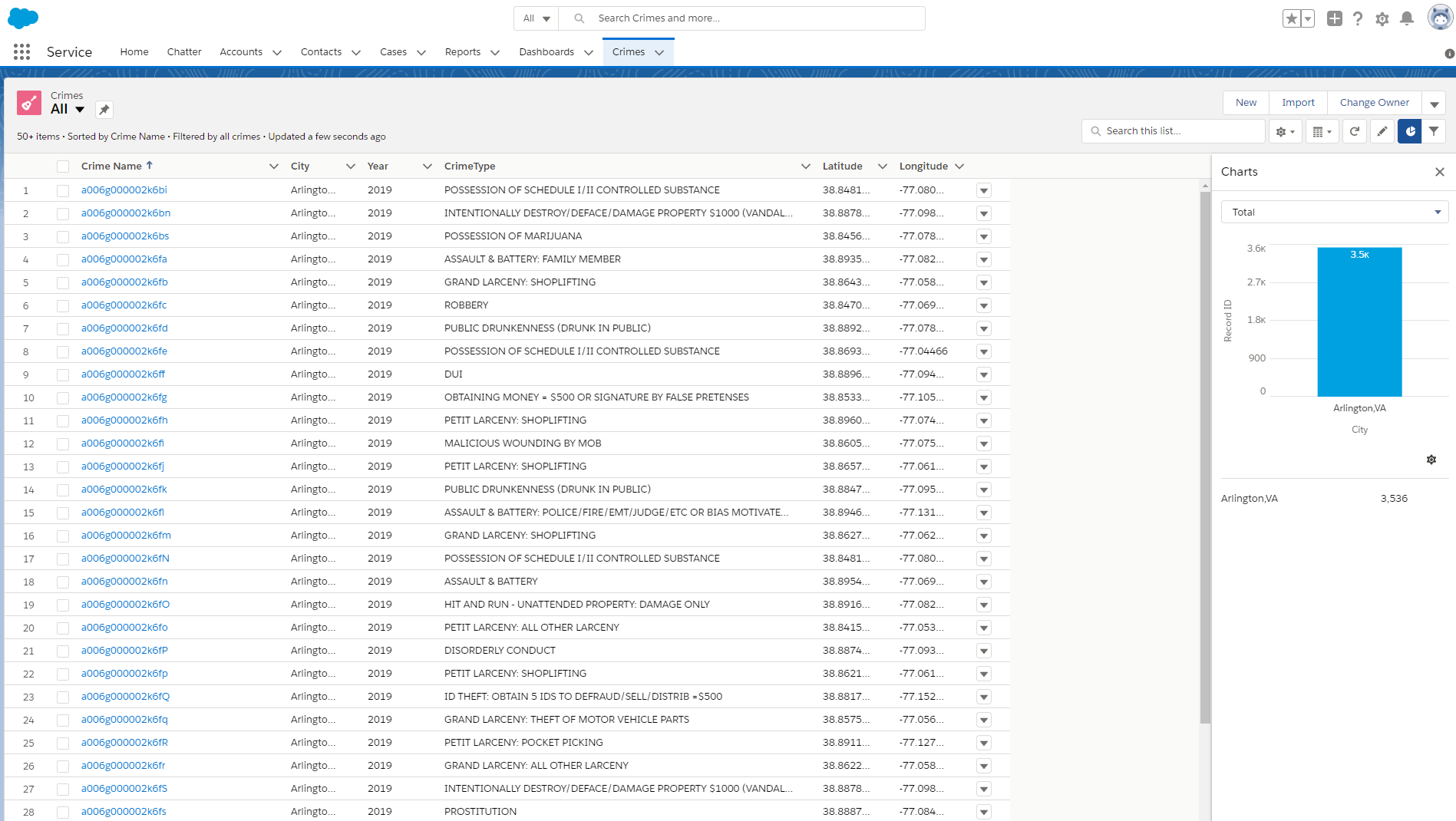
1. **Data Storage into a Database.**

The third, last step in the ETL process involved storing the data into the SalesForce (see Figure 5) and PowerBi (see Figure 6) platforms.

Using Simple SalesForce module in Python, we uploaded the records from the CSV file to SalesForce. The issues faced were as follows:

1. Limits of 10,000,000 character is set for Bulk upload method which the data exceeded.
2. Limited storage allowed in our SalesForce account which limited the uploaded data to 3500 records.
3. Difficulty in processing Latitude and Longitude variables as geolocation data and the uploaded data had to be defined as Text.

**Figure 5. Final output stored in SalesForce (combined crime data in Washington, D.C, Arlington, VA, and Bethesda, MD).**



Also, a sample of the data was uploaded on Google Maps by using data import wizard which limited the uploaded data to 2000 records:

<https://www.google.com/maps/d/u/0/edit?mid=1pd0piZ58pY0t4gpRhaoU2yHt4FAWG3pe>

All data was loaded to PowerBI and easily accessible:

<https://app.powerbi.com/view?r=eyJrIjoiYzc1YjZjMGUtM2JjMS00ZjE2LTk5YmYtMzUwNjEyNDJlYWY2IiwidCI6IjgxNWE4NGQ4LTc0NWEtNGFiNC04MzIwLTI2ZGM1MTU1MjM1YiIsImMiOjJ9>

**Figure 5. Final output uploaded into the PowerBi (combined crime data in Washington, D.C, Arlington, VA, and Bethesda, MD).**

