

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.

2. 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.

3. 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).

	NEIGHBORHOOD_CLUSTER	CENSUS_TRACT	offensegroup	LONGITUDE	END_DATE	offense-text	SHIFT	YBLOCK	DIS
0	cluster 1	4002.0	property	-77.041686	2018-08-23T20:24:31.000	theft/other	evening	139037.0	
1	cluster 24	9000.0	property	-76.952663	2018-08-23T21:24:58.000	theft/other	evening	139186.0	

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

```
# Create new data with select columns (DC Location)

new_dc_df = dc_df[['OFFENSE', 'START_DATE', 'LONGITUDE', 'LATITUDE']].copy()
new_dc_df.head()
```

	OFFENSE	START_DATE	LONGITUDE	LATITUDE
0	theft/other	2018-08-23T19:46:41.000	-77.041686	38.919196
1	theft/other	2018-08-23T20:23:41.000	-76.952663	38.920536
2	theft/other	2018-08-27T08:25:43.000	-77.032615	38.904524
3	theft f/auto	2018-08-27T10:32:14.000	-76.996786	38.857649
4	theft/other	2018-08-22T11:39:44.000	-76.995309	38.884593

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).

	OFFENSE	START_DATE	LONGITUDE	LATITUDE	Year
0	theft/other	2018-08-23T19:46:41.000	-77.041686	38.919196	2018
1	theft/other	2018-08-23T20:23:41.000	-76.952663	38.920536	2018
2	theft/other	2018-08-27T08:25:43.000	-77.032615	38.904524	2018
3	theft f/auto	2018-08-27T10:32:14.000	-76.996786	38.857649	2018
4	theft/other	2018-08-22T11:39:44.000	-76.995309	38.884593	2018

```
# Copy relevant columns (DC Location)
new2_dc_df = new_dc_df[['OFFENSE', 'Year', 'LATITUDE', 'LONGITUDE']].copy()

# Rename the columns (DC Location)
new2_dc_df.rename(columns = {"OFFENSE": "Crime Type", "Year": "Year", "LATITUDE": "Latitude (D.C.)",
```

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).

```
#combining all data
Combined=pd.concat([Arl, Bet, DC])

#Exporting data to csv file
Combined.to_csv("CombinedData.csv")

#Check Combined Dataframe
combined=pd.read_csv("CombinedData.csv", index_col=0)
combined=combined.rename(columns={"Crime Type": "CrimeType"})
combined.head()
```

	CrimeType	Year	Latitude	Longitude	City
0	POSSESSION OF SCHEDULE I/II CONTROLLED SUBSTANCE	2019	38.848117	-77.080449	Arlington, VA
1	INTENTIONALLY DESTROY/DEFACE/DAMAGE PROPERTY \$...	2019	38.887844	-77.098786	Arlington, VA
2	POSSESSION OF MARIJUANA	2019	38.845667	-77.078157	Arlington, VA
3	ROBBERY	2019	38.847067	-77.069845	Arlington, VA
4	PETIT LARCENY: SHOPLIFTING	2019	38.896079	-77.074585	Arlington, VA

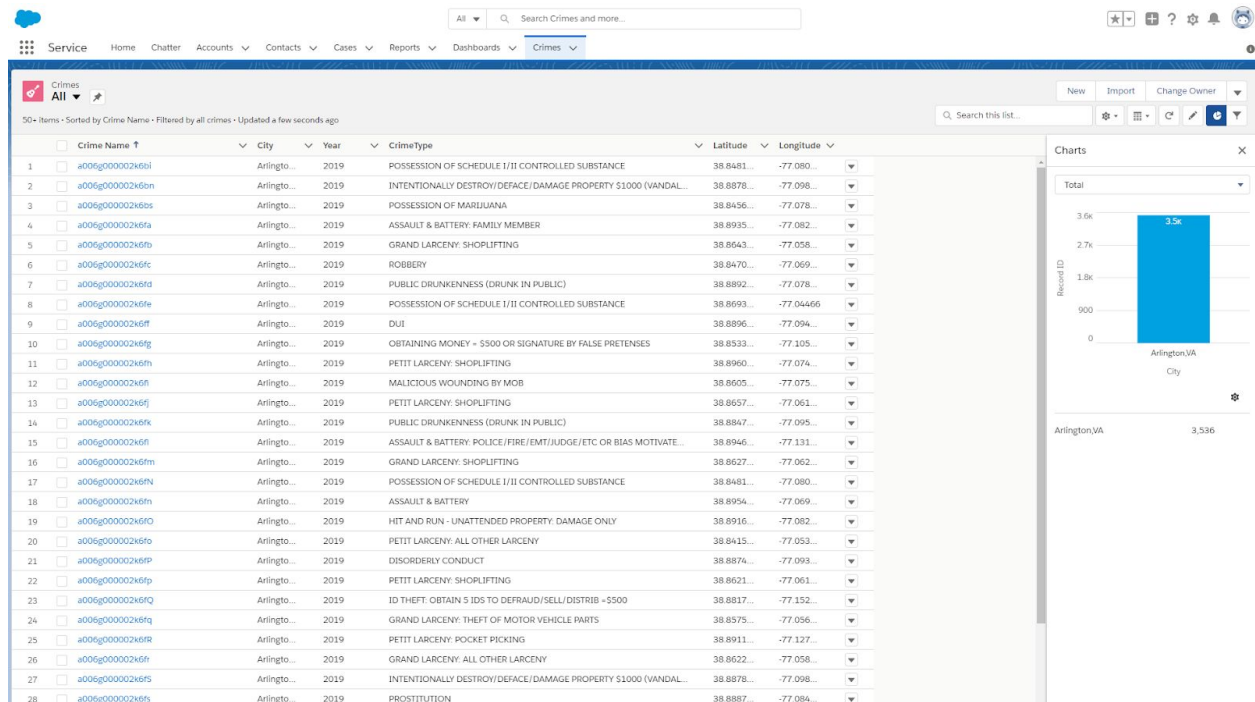
4. Data Storage into a Database.

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=eyJrIjoiYzc1YjZjMGUtM2JjMS00ZjE2LTk5YmYtMzUwNjEyNDJlYWY2IiwidCI6IjgxNWE4NGQ4LTc0NWEtNGFiNC04MzIwLTl2ZGM1MTU1MjM1YiIsImMiOiJ9>

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

