## Data Cleanup & Analysis

**EXTRACT:** The sources of data that you will extract from.

We extracted our government-provided crime statistics-datasets from the District of Columbia Metropolitan Police Department. We had intended on using a .json and .csv file for the ETL process but the datasets were only available as geojson files. After discussing it with Reed - we ended up downloading the datasets as two .csv spreadsheets because we were unable to manipulate the .geojson format as .json so it was giving us errors in python. Our original data sources can be found here:

<https://opendata.dc.gov/datasets/crime-incidents-in-2018>

<https://opendata.dc.gov/datasets/crime-incidents-in-2019>

**TRANSFORM:** The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).

First we stored our two downloaded .csv files into a DataFrame using Pandas. Then, we created a new DataFrame by selecting only the relevant columns we desired for our final tables to be used in Postgres. Using SQLAlchemy was required to import the engine allowing us to load our .csv converted DataFrames into our local Postgres database. We also needed to rename two table columns within Postgres in order for Pandas to successfully detect and load the DataFrames into the database.

**LOAD:** The type of final production database to load the data into (relational or non-relational).

The datasets were loaded from pandas into a PostgreSQL relational database called ‘crime\_db’. We used a relational database because each of the datasets (and corresponding tables) would represent a different year. While the table elements would be unique from year to year, the crimes they’d represent could be related by one or more features (e.g. shift, ward, time). Having the data stored in this relational format would facilitate the comparison of crimes by those features of interest.

The final tables or collections that will be used in the production database.

Through the transformation/cleanup process we ended up with two final new tables for crime in the District of Columbia during 2018 & 2019. We only want to keep pertinent info in the final production database so these new tables include columns Report Date, Shift, Offense, Block, Ward, Bid. These new tables were loaded into our final crime production database (crime\_db).

Once you have identified your datasets, perform ETL on the data. Make sure to plan and document the following: You will be required to submit a final technical report with the above information and steps required to reproduce your ETL process.

## Project Report

At the end of the week, your team will submit a Final Report that describes the following:

* **E**xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).
* **T**ransform: what data cleaning or transformation was required.
* **L**oad: the final database, tables/collections, and why this was chosen.

Please upload the report to Github and submit a link to Bootcampspot.