ETL PROJECT – NYC Parking Tickets

The NYC Department of Finance collects data on every parking ticket issued in NYC (~10M per year!).

There are four files, covering Aug 2013-June 2017. The files are roughly organized by fiscal year (July 1 - June 30) with the exception of the initial dataset. The initial dataset also lacks 8 columns that are included in the other three datasets (although be warned that these additional data columns are used sparingly). See the dataset descriptions for exact details. Columns include information about the vehicle ticketed, the ticket issued, location, and time.

Using the ETL processes, the following tasks were done:

Extract:

1. Extracted 2017 Parking Violations Issued in New York.

Source:

<https://www.kaggle.com/new-york-city/nyc-parking-tickets>

Transform:

1. Issued Parking Tickets Data
   1. Used Python Pandas library to load and read the three .csv files.
   2. Used pd.concat function to combine the three DataFrame results
   3. Stored addresses by selecting location 2 column and putting it in a list
2. Addresses of Issued Parking Tickets must be transformed to be identical.
   1. Converted address list to upper case
   2. Removed dots from address
   3. Stored cleaned up data to a list called streets
   4. Verified if the counts of common addresses on both datasets match

Load:

1. Created SQL connection
2. Exported to MySQL.