ETL - Final Report

Boston Crimes & Boston Public Schools

BJ Lim

Randolf

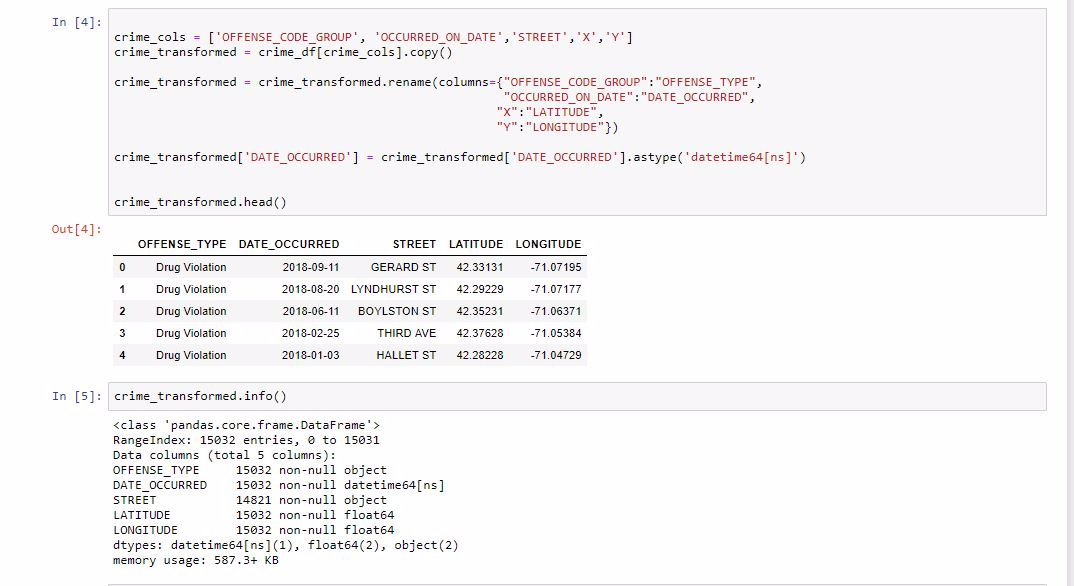
Extract:

For the extraction process, we looked at different datasets from Kaggle and decided to extract two datasets focusing on Boston. The two datasets we found were boston-crime-data and boston-public-schools. Both datasets were in CSV format so we downloaded them and studied them to see what kind of tables we could make.

Transform:

After looking at the data, we quickly realized there was a huge difference in the number of observations in the two datasets. There were more than 327,000 crimes recorded in this dataset in Boston and only 131 public schools in the dataset. Because it would take a lot of time manipulating such a large set of data in MySQL, we decided to focus on only the drug related crimes by looking at the OFFENSE\_CODE\_GROUP. After significantly trimming down the data we were able to look at five specific columns that we wanted to import from each dataset. We also did data validation by checking what kind of data types that the columns were stored as and changed them as needed to be the correct data types.

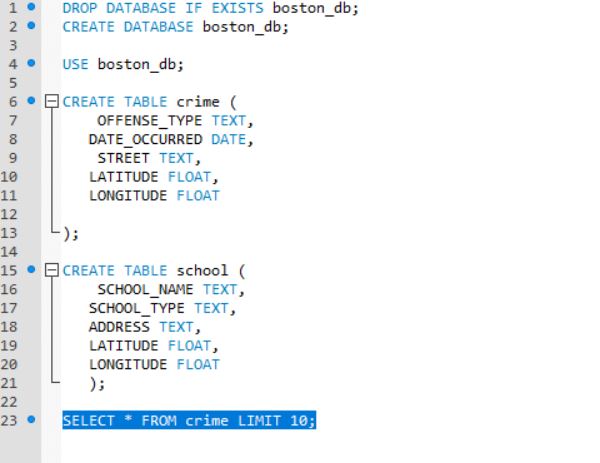
As shown below, we also transformed the data by renaming the columns to a more appropriate name that we thought. We made the dataframe based on the five columns that we wanted and also labeled them.



Load:

The final step in the ETL process involves loading the transformed data into targeted destinations.

We created a database on MySQL and called it boston\_db to load in our data as shown below.



We proceeded with a full load from our already processed Boston-crime-data and Boston-public-schools data into boston\_db.

Using SQLAlchemy, we created a database connection and loaded the data frames into the database(boston\_db). Appending the data into the respective tables found on the database. 