**Motivation**

Reports of high rates of crime are ever present in the media today. We were particularly interested to what extend youth crime contributes to crime rates, and the factors associated with youth crime. We investigated crime datasets and came upon the 2006-2019 NYC Crime Dataset1, originally accessed from New York Open Data, and available in CSV format from Kaggle. Our motivations for using this dataset were:

* Use the crime data from a well known, large urban city to model potential factors contributing to youth crime
* It would allow analysis of demographic factors (e.g. race, location, gender) contributing to crime
* CSV format allowing easy integration with Python Pandas

To investigate this problem further we sought to see what influence educational factors might have on crime rates. The New York Federal Reserve Bank provides data on school spending per student2, which we accessed via a web scrape. Our motivations for utilising this data were:

* Allow analysis of the influence of school spending on crime rates
* Allow analysis of how school spending differs between NYC districts and if this affects the crime rate

The Intention of the project was to create a database of both NYC crime data for youths (those aged <18y.o) which would link to data for school spending in the NYC area. Such a database could be used by organisations such as the NYPD, NYC Education department, as well as NGO’s to reduce crime and improve outcomes in communities.

**Extract**

NYC Crime Dataset

* The dataset was download locally from Kaggle1 and extracted into a Jupyter Notebook using the Pandas read\_csv function.

New York Federal Reserve Bank

* A web scrape was used to access the values from the interactive table of available on the NY Federal Reserve Website.
* Splinter library was initiated and used to click on the data tabs for the table
* Beautiful soup was used to scrape the data and the html collected was then passed to Pandas read\_html function to convert to a useable form

**Transform**

NYC Crime Dataset

* Firstly the relevant columns were selected from the larger total dataframe
* Columns not thought to be relevant to the subject matter (factors affecting youth crime) were excluded (for example, latitude, longitude, law code, offense sub-category).
* The dataframe then filtered to keep only those offenders aged <18y.o., and the age column removed from the final table

New York Federal Reserve Bank

* The schools data was first converted to a data frame
* Column titles required re-naming
* Columns containing NaN values were present in the HTML converted table and these were removed with Pandas dropna function
* As the NYC Crime was only divided into the 5 Boroughs, not 31 districts, the summarised data for each Borough had to be selected for the table
* Finally, as a different naming system was used for the Boroughs between datasets this had to be updated to allow future users of the database to perform joins when querying the database

**Load**

A new database was created (youth\_crime\_db) in PostGRESQL and table schemata defined for crime\_table and school\_table

* Both tables have primary keys (arrest\_key for crime\_table, nyc\_district for school\_table)
* Crime table also has a foreign key (arrest\_boro)

SQL was chosen as the data was in a relational structure which lends itself well to SQL, as well as making it easy to access and query the dataset in the future.

Python SQLAlchemy was used to load the data into the database

* A connection was established to the database with the create\_engine function.
* Each table was sequentially loaded into the database with the to\_sql function.

**References**

1. <https://www.kaggle.com/ajkarella/nyc-crime-stats>
2. <https://www.newyorkfed.org/data-and-statistics/data-visualization/nyc-school-spending>