Extraction/Sources of Data

1. Chicago Data Portal: [Crimes 2001 to present](https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data)
2. New York City Open Data: [NYPD Data History](https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i)

Data was extracted in CSV format.

Transformation of Data

1. Cleaned columns
2. Deleted extraneous columns that would not be useful for the purposes of this project. This made both datasets uniform.
3. Filtered dates to be within same range for both datasets.
4. Renamed columns to be user-friendly and help associate between cities.
5. Combined datasets up-down on ‘Date’ after aggregating on ‘Primary Type’ of violation
6. Saved new dataset as .CSV

Loading of new dataset

1. Uploaded in to SQL Database using PGAdmin/PostGres
   1. Created a database “etl-project”
   2. Created a table with corresponding column names: id, date, primary type, and description
   3. Pulled a table output so that the data is viewable
   4. Saved code schema and table in .sql format

GitHub

1. Uploaded all documents related to etl-project in to github repository
2. Zipped data files since they were too big to upload without doing so

In the end, we combined the Chicago and New York crime data sets on Date and Primary Type (of violation/crime). We are hoping this would provide anyone interested in crime-type rates or a very general look at what type of crimes tend to happen on what month/day a simplified view in to a complicated database; this will help researchers formulate questions and then dive deeper in to a more robust database.