**ETL Report**

**Extract**

Data files related to Chicago Population, Crime, and Public Health Indicators were identified. CSV files were accessed and downloaded from the Chicago Data Portal and the Act for Children websites. Below is a list of datafiles used for this project:

* Chicago Public Health Data 2015
  + Data Source: <https://data.cityofchicago.org>
* Chicago Crime Data 2018
  + Data Source: <https://data.cityofchicago.org>
* Chicago Population Data 2017
  + Data Source: <https://www.actforchildren.org/>

**Transform**

Several transformations were performed to prepare the data for analysis. Each transformation per data source is described below:

* Chicago Public Health Data 2015
  + Select Columns
    - Community Area \* (PK)
    - Community Area Name
    - Unemployment
    - Dependency
    - No High School Diploma
  + Drop NULLS
  + Community Area – change type to INT
* Chicago Crime Data 2018
  + Select Columns
    - ID
    - Community Area \* (PK)
    - Primary Type
  + Drop NULLS
  + Drop rows where Community Area = 0
  + Community Area – change type to INT
  + Group by Community Area
* Chicago Population Data 2017
  + Select Columns
    - Community Area \* (PK)
    - Population
* Merge – Crime & Population Data
  + Merged on ‘Community Area’
  + Create variable – Crime rate (Calculation)
    - Calculate crime rate
      * # crimes / population \* 1,000
  + Create variable – Drug Related Crime
    - Primary Type – Narcotics & Other Narcotics Violation
    - Combine these two offenses into new variable

**Load**

TBD

The team decided to use PostgreSQL.

* PostgreSQL is an open source object-relational database system that uses SQL language
* PostgreSQL is a free and open-source relational database management system (RDBMS)

Two tables were loaded

* Public Health
* Merged DF – Crime & Population

**Analysis**

* Correlation
  + Crime Rate and Unemployment
  + Unemployment and Education (No HS diploma)
  + Drug Related Crime and Dependency