ETL Project

By John Rafferty, Olga Taranukha & Abdullah Sher

**Extraction**

Our extraction process started by searching for data sets in Kaggle. We came across 2 datasets that we thought could be interesting from an analysis perspective. We were sent to the city of Chicago’s public data website where were used a JSON API and loaded them into Jupyter Lab. The first file contained a selection of six socioeconomic indicators of public health significance and a “hardship index,” for Chicago wards (neighborhoods) from the years 2008-2012. The second file reported incidents of crime (with the exception of murders) that occurred in the City of Chicago from 2001 to present.

Data #1: <https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2>

Data #2: <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>

**Transformation**

We were working with two very large data sets. In order to transform these data sets, we decided to eliminate any unnecessary columns. We wanted to only work with the cleanest and most efficient data, and the first step was to remove unnecessary columns. Secondly, we renamed the remaining columns to be more intuitive. For example, in Data #1, the initial column name was “ca” (community area). However, Data #2’s same column name was “ward\_no”. We changed the column “ca” to “ward\_no” so we could eventually merge and perform analysis.

**Loading**

The last step was to transfer our final output into a DataBase. We created a database and respective tables to match the columns from the final Panda’s Data Frame using MYSQL and then connected to the database using SQLAlchemy and loaded the result. Here we were able to perform multiple queries to suit a desired criterion.

**Sample Query**

A screenshot of a cell phone

Description automatically generated