For our project we had data from two sources. The first source of data was Chicago crime data pulled from Chicago.gov which is a state government run repository for public data. Our second source of data was Chicago school data pulled from dataworld.com which was publicly available.

The crime dataset had longitude and latitude information in two separate columns so we combined the latitude and longitude information into one column called “lat_long”. We then did a groupby of the “BEAT” column and “lat_lang” column to give an arbitrary latitude and longitude for every BEAT. Next, we utilized geocoders and geopy to transform the latitude and longitude with their corresponding BEATS to output a zip code for each BEAT. After extracting and merging the zip code column relative to each beat to our data, we created a new data frame with only columns we cared about. The school dataset had many columns therefore we created a new data frame only containing the columns we were interested in.

Next, we used pandas to convert  both our data sets (crime and school) into database and then confirmed that the data had been added by querying the crime and school table. Once gaining confirmation that our data had been added into our database, within our database, we created our two tables and then joined the data together on zip code information from both the crime table and the school table.

Resources

https://www.chicago.gov/city/en/dataset/crime.html

https://data.world/cityofchicago/chicago-public-schools-progress-report-cards-2011-2012/workspace/file?filename=chicago-public-schools-progress-report-cards-2011-2012-1.csv