Yuyan Wang

DATA 23700 Exercise 5 Project Check-in

Motivation

Imagine you have never been to Chicago and are planning to move there soon. Concerns about the city's safety issues might prompt you to seek out more information on the topic. This was precisely my situation a few months ago. I searched for online reviews regarding safety in Chicago but wasn't sure which sources—be they webpages or other users' posts—were trustworthy. Moreover, the information I found online was so varied that it seemed everyone had a different opinion on the level of danger in various parts of Chicago. After hours of reading, I still did not have a clear understanding.

This project, therefore, zeroes in on "Chicago Crimes" to tackle questions such as "which areas of Chicago are more prone to crime?", "is the crime rate increasing over the years?", "what types of crimes are most common?" among others. I believe this is an excellent opportunity to find the answer myself using a dataset from the Chicago Data Portal (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2).

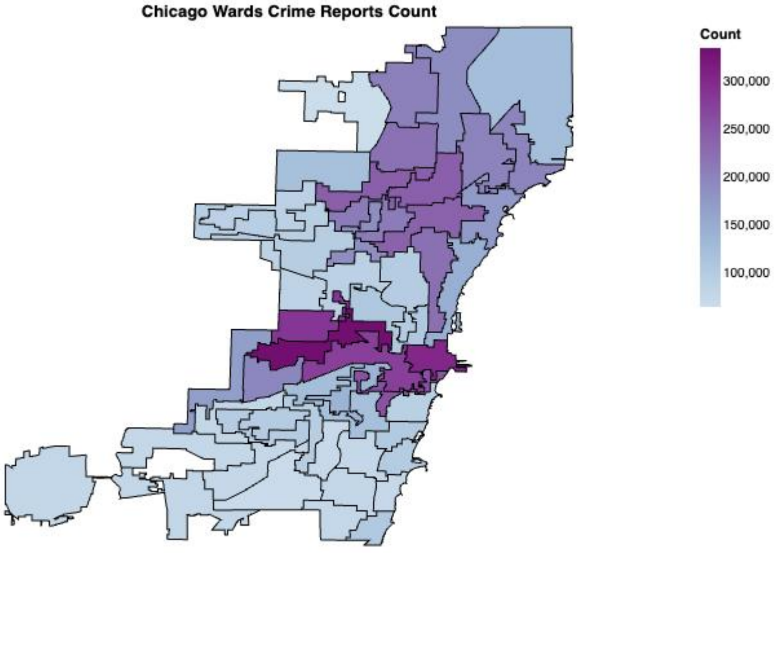
I am aware that the topic is somewhat general, as the dataset is rich with information. This is really the essence of planning a good visualization! So I'll try to analyze the dataset from different angles and give some visualizations regarding "Crime in Chicago".

Dataset

As stated, the dataset is obtained from the "Chicago Data Portal" (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2). It originally has 7.93M rows and 22 columns, where each row is a reported crime. I definitely will not work with all the columns since some of them are keys/identifiers and some are not relevant for the main focus of this project. And maybe will consider dropping some rows (still need to be decided) since 7.93M rows may be too much (more than 1GB) and will eventually make the project's code run slower.

The columns that I plan to use are: "ID" the unique identifier; "Date" the date when the incident occurred; "Primary Type" the primary description of the crime; "Description" the more detail desciption of the crime; "Location Description" the description of the location where the incident occurred (residence, retail store, street, apartment...); "Arrest" which indicates whether an arrest was made; "Domestic" which indicates whether the incident was domestic-related; "District" the police district where the incident occurred; "Ward" the chicago ward where the incident occurred; "Community Area" the incident community area; "Year" the year that the incident occurred; and other geographical columns to make map plots.

It is easy to see that this dataset is suitable for the project's purpose since it records Chicago crimes from 2001 to the present. I am planning to create user-interactive plots for this project (after having learned how to make them) along with more sophisticated designs that are yet to be decided. Since this is just a project checkpoint, I will only include a draft visualization that displays the Chicago wards and colors them according to the number of crimes reported in each ward.


Chicago Wards Crime Reports Count
Count
300,000
250,000
200,000
150,000
100,000