# Datasets

In this project, we will be using the technique of machine learning and data science for crime prediction of Chicago crime data set. The crime data is extracted from the official portal of Chicago police. The data is publicly available data from the City of Chicago’s data portal linked below to explore crime in Chicago from January 2001 to February 2018

Data Sources:

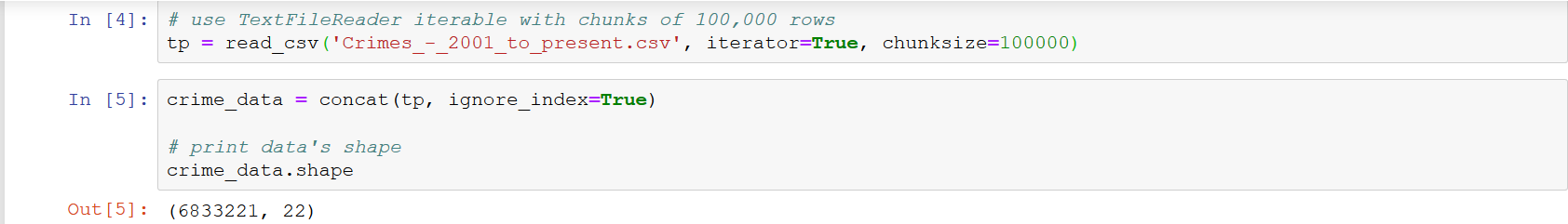
* <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>
* <https://data.cityofchicago.org/Public-Safety/Boundaries-Police-Beats-current-/aerh-rz74>

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days.

## **Columns in this Dataset:**

| **Column** **Name** | **Description** | **Type** |
| --- | --- | --- |
| ID | Unique identifier for the record. | Number |  |
| Case Number | The Chicago Police Department RD Number (Records Division Number), which is unique to the incident. | Plain Text |  |
| Date | Date when the incident occurred. this is sometimes a best estimate. | Date & Time |  |
| Block | The partially redacted address where the incident occurred, placing it on the same block as the actual address. | Plain Text |  |
| IUCR | The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at <https://data.cityofchicago.org/d/c7ck-438e>. | Plain Text |  |
| Primary Type | The primary description of the IUCR code. | Plain Text |  |
| Description | The secondary description of the IUCR code, a subcategory of the primary description. | Plain Text |  |
| Location Description | Description of the location where the incident occurred. | Plain Text |  |
| Arrest | Indicates whether an arrest was made. | Checkbox |  |
| Domestic | Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act. | Checkbox |  |
| Beat | Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts. See the beats at <https://data.cityofchicago.org/d/aerh-rz74>. | Plain Text |  |
| District | Indicates the police district where the incident occurred. See the districts at <https://data.cityofchicago.org/d/fthy-xz3r>. | Plain Text |  |
| Ward | The ward (City Council district) where the incident occurred. See the wards at <https://data.cityofchicago.org/d/sp34-6z76>. | Number |  |
| Community Area | Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at <https://data.cityofchicago.org/d/cauq-8yn6>. | Plain Text |  |
| FBI Code | Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at <http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html>. | Plain Text |  |
| X Coordinate | The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block. | Number |  |
| Y Coordinate | The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block. | Number |  |
| Year | Year the incident occurred. | Number |  |
| Updated On | Date and time the record was last updated. | Date & Time |  |
| Latitude | The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block. | Number |  |
| Longitude | The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block. | Number |  |
| Location | The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block. | Location |  |

The original dataset from the city data portal was in .CSV format but was too large (at 1.5 GB) for my laptop’s resources. I could not fit the whole file in memory so I used pandas' *TextFileReader* function which allowed me to load the large file in chunks of 100,000 rows and then concatenate the chunks back together into a new data frame.



Here is a snippet of the first few rows of the dataset:

