Data Wrangling

Below are steps for cleaning up the dataset on Chicago Crimes:

1. General examination

   With the methods and attributes of *pandas*, overall the data is quite clean. With further exploration, some cleanups are conducted.

2. Remove rows containing at least one null value

   The total number of cases is over 6.7 million, and the number of cases that has at least one NULL value is about 0.7 mil. Majority of the NULL values happen for the fields 'Ward' and 'Community Area'. Since the latter is only about 10% of the total cases, they are removed from the whole dataset.

3. Format date

   a. Some column names such as 'Community Area' include a space. I replace space with '_'.
   b. The fields of 'District', 'Ward', and 'Community_Area' are float64 data. Actually, they are just numeric codes contain no decimal point. I convert them to the int64 format.
   c. I changed the fields of 'IUCR', 'Primarhy_Type', and 'FBI_Code' to the categorical variables.
   d. I converted the 'Date' from the string format to the date-time format. Since the column 'Year' comes from the 'Date', at this moment, I remove it.

4. Remove duplicates
   The field 'ID' is unique, but the field 'Case_Number' included 378 duplicates. I remove them from the dataset.

5. Remove outliers
   I use the scatter plot to examine 'X_Coordinate' and 'Y_Coordinate', and find 107 outliers. I remove them from the dataset.

6. Save the clean file
   The dataset cleaned up gets saved to a file for data analysis.

The Jupyter Notebook is available in Github:
https://github.com/DavidFlanders/Capstone/blob/master/ChicagoCrimeData%20-%20Data%20Wrangling.ipynb

Appendix: Description of Chicago Crimes 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 Uniform 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.	