Datasets Used

- Chicago crime reports from 2001 into 2018 provided by the City of Chicago (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2)
- Chicago train ('L') stops provided by the City of Chicago (https://data.cityofchicago.org/Transportation/CTA-System-Information-List-of-L-Stops/8pix-ypme)
- Chicago bus stops provided by the City of Chicago (https://data.cityofchicago.org/Transportation/CTA-Bus-Stops-kml/84eu-buny)
- Chicago business licenses provided by the City of Chicago (https://data.cityofchicago.org/Community-Economic-Development/Business-Licenses/r5kz-chrr)
- Chicago police stations provided by the City of Chicago (https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e)
- U.S. federal holidays provided by Kaggle (https://www.kaggle.com/gsnehaa21/federal-holidays-usa-19662020)
- Chicago community area boundaries provided by the City of Chicago (https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6)
- Chicago ward boundaries provided by the City of Chicago (https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Wards-2015-/sp34-6z76)

All of the above datasets, with the exception of the bus stop data, were csv files and were imported directly as Pandas data frames. The bus stops were provided in kmz format. I used MyGeodata Converter (https://mygeodata.cloud/converter/) to convert this file to csv format and then imported it as a Pandas data frame.

Data Wrangling

Chicago Crime Reports

This dataset consisted of 22 columns and 6,726,718 rows with each row being a reported crime. The columns that were used during the data cleaning are as follows:

- ID: unique identifier for each report
- Case Number: Chicago Police Records Division Number, which is unique to the incident
- Date: approximate date and time when the incident occurred
- Year: year when the incident occurred
- Block: block where the incident occurred in the format of a partially obscured street address
- Primary Type: primary type of the crime (what I will be trying to predict)
- Location Description: description of the location where the incident took place
- Domestic: indicates if the crime was domestic
- Beat: smallest police geographic area where the incident occurred
- District: police district where the incident occurred

- Ward: ward where the incident occurred
- Community Area: community where the incident occurred
- X Coordinate: X coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection (shifted from the actual location but falls on the same block)
- Y Coordinate: Y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection (shifted from the actual location but falls on the same block)
- Year: year the incident occurred
- Latitude: latitude in degrees of the location where the incident occurred (shifted from the actual location but falls on the same block)
- Longitude: longitude in degrees of the location where the incident occurred (shifted from the actual location, but falls on the same block)

ID:

The 'ID' column was kept as it would be useful to have a unique identifier for each report in case I had to split and merge my data. I verified that there was a unique ID for each report by finding the length of the set of the 'ID' column. It was the same as the number of rows.

I then dropped the duplicate reports while excluding the 'ID' column using drop_duplicates. This removed 206 reports.

Case Number:

I used the case number to check for any more duplicate reports. There were 175 case numbers that had more than one report. I looked up a few of the cases online and saw that in general, duplicate case numbers meant there were multiple victims in the same location on the same day. There was one case I saw where the same case number was used for a different day but the same location. No further information was found online regarding this incident, so it is unknown if this was a case of a case number being mismarked or if the incidents were connected. I therefore left the reports with duplicate case numbers in the dataset.

Date:

Per dropna(), there were no apparent null values in this column. I created a regular expression matching the syntax of the date and checked that all the dates followed that format. There were no irregular dates. Using pandas.to_datetime, I converted the date to a datetime object.

Year:

Per dropna(), there were no apparent null values in this column. The column was properly imported as an integer type, so it appeared to be clean.

Block:

Per dropna(), there were no apparent null values in this column. I capitalized all of the blocks and then created a regular expression matching the syntax of the blocks and checked that all blocks followed that format. There

were 2 entries that were listed as 'XX UNKNOWN'. In both cases the latitude and longitude were missing so I wasn't able to estimate these blocks. Several blocks showed up with single quotes, accents, and random characters. To simplify this column, all of these characters were stripped.

Primary Type:

Per dropna(), there were no apparent null values in this column. I capitalized all of the types of crime and performed value_counts() to get the unique values. I saw that 'NON-CRIMINAL' showed up 3 times: 'NON-CRIMINAL', 'NON - CRIMINAL', and 'NON-CRIMINAL' (SUBJECT SPECIFIED). I removed the spaces around the hyphen and the '(SUBJECT SPECIFIED)'.

Location Description

Per dropna(), there were 3974 null values in this column. One way I used to figure out the location description was to use the 'Domestic' column which said if the crime was domestic related. Using value_counts() for reports which were domestic related, I found that domestic related crimes occurred more in residences. For reports that were domestic related, I therefore filled in 'RESIDENCE' for the missing location description. Unfortunately, this only removed 1 null value and there were no other features I could use to accurately estimate the location description. To further clean this column, I capitalized all entries and removed extra spaces around hyphens and backslashes.

Beat:

Per dropna(), there were no apparent null values in this column. This column was properly imported as an integer type, so it appeared to be clean.

District:

Per dropna(), there were 47 null values. Looking at the value counts, there were very low report counts for districts 21 and 31 (4 and 147 reports, respectively). A look at the Chicago Police website (https://home.chicagopolice.org/community/districts/) showed that these districts no longer existed. I therefore made these districts null.

Examining the police beat boundaries and police district boundaries (Figure 1), it appeared that the police beats generally fell nicely within the boundaries of the police districts. In order to find the missing districts, I created a dictionary with beat as the key and district as a value. I then used the dictionary to fill in the missing districts.

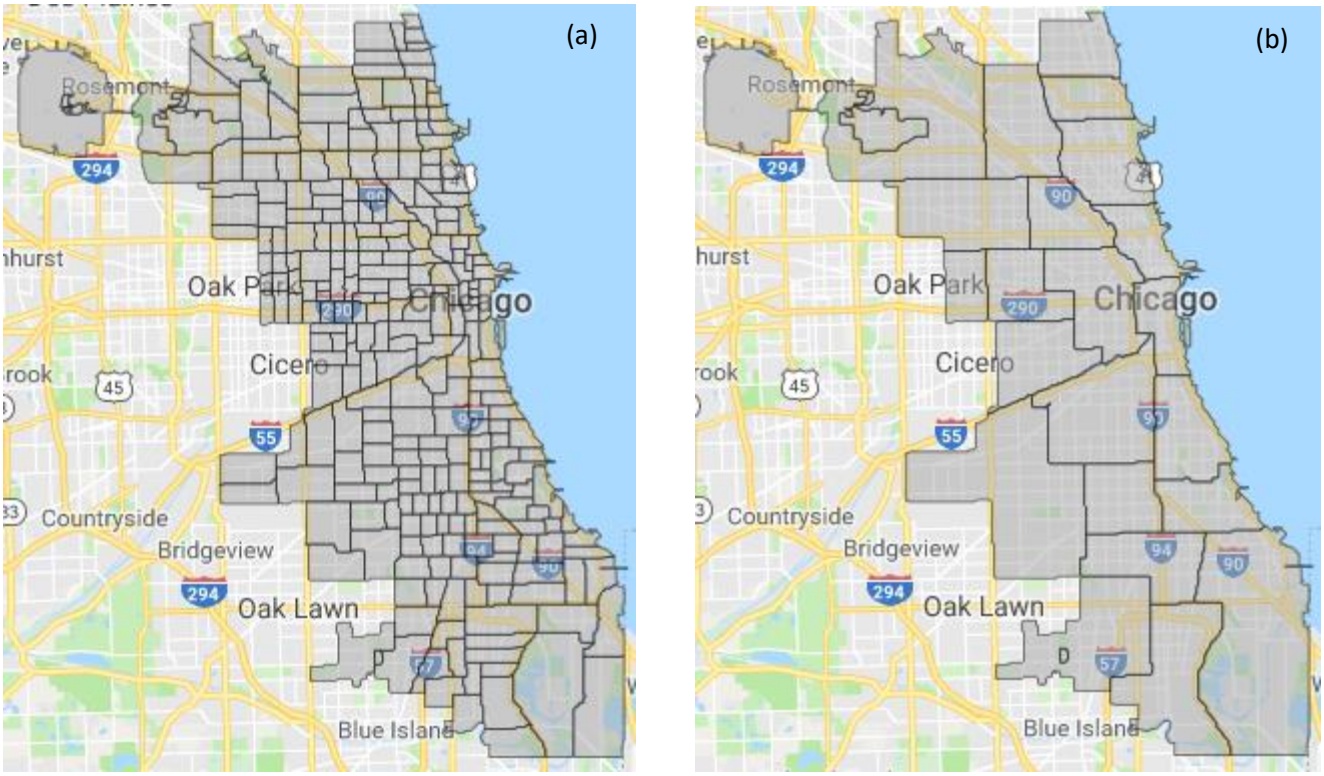

Figure 1: Boundaries for (a) police beats (City of Chicago, 2018b) and (b) police districts (City of Chicago, 2018c).

Ward:

Per dropna(), there were 614,846 null values in this column. I looked at the number of reports missing the ward per year and saw that 2001 and 2002 had the highest numbers (481,619 and 133,121 reports, respectively). It is possible that the ward wasn't recorded until sometime in 2002.

Examining the police beat boundaries and ward (Figure 2), it appeared that the police beats did not fit nicely within the wards. So the technique used for the police district could not be accurately used here. The latitude/longitude had to be used to figure out the ward. I revisited this later on.

Community Area:

Per dropna(), there were 616,022 null values in this column. I looked at the number of reports missing the community per year and saw that 2001 and 2002 had the highest numbers (481,630 and 133,156 reports, respectively). It is possible that the community wasn't recorded until sometime in 2001.

Examining the police beat boundaries and community boundaries (Figure 3), it appeared that the police beats did not fit nicely within the communities. So the technique used for the police district could not be accurately used here. The latitude/longitude had to be used to figure out the ward. I revisited this later on.

Looking at the value counts for each community, there were 91 reports for community 0. There is no community 0, so I changed this community to null.

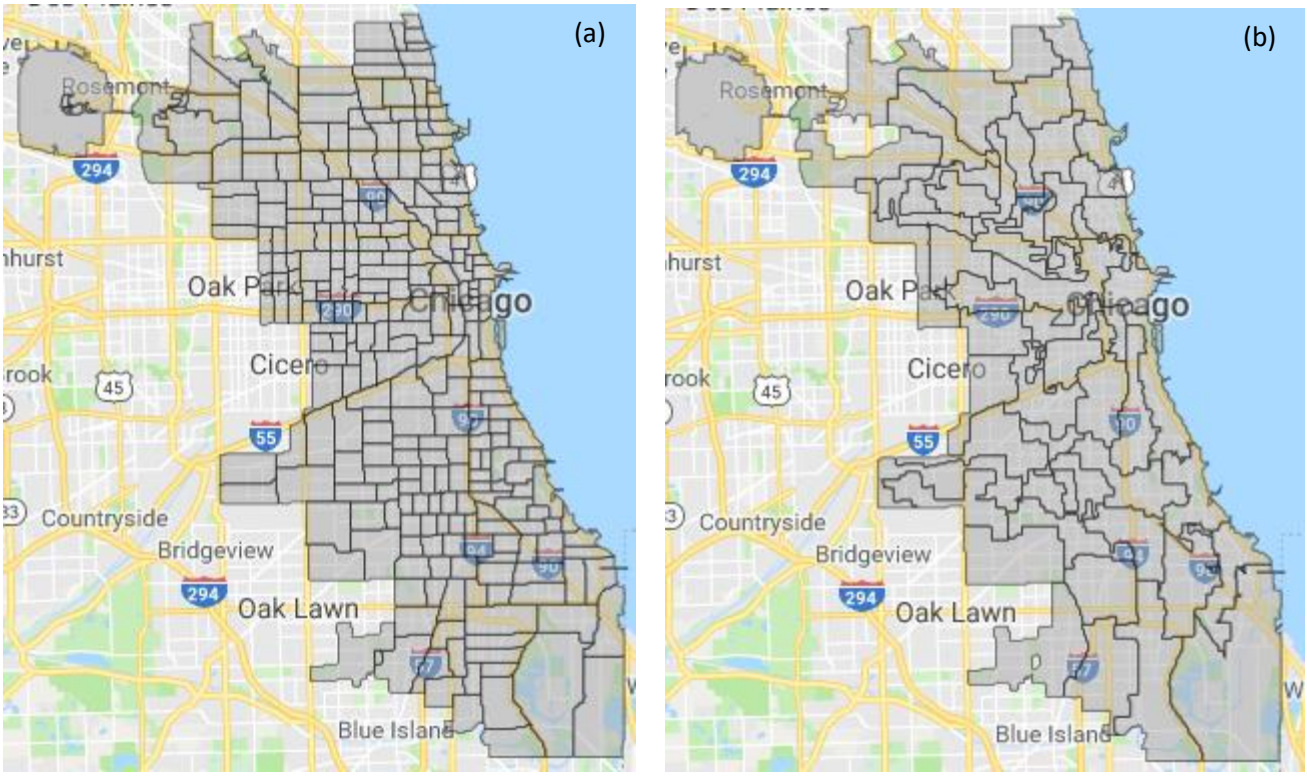

Figure 2: Boundaries for (a) police beats (City of Chicago, 2018b) and (b) wards (City of Chicago, 2016).

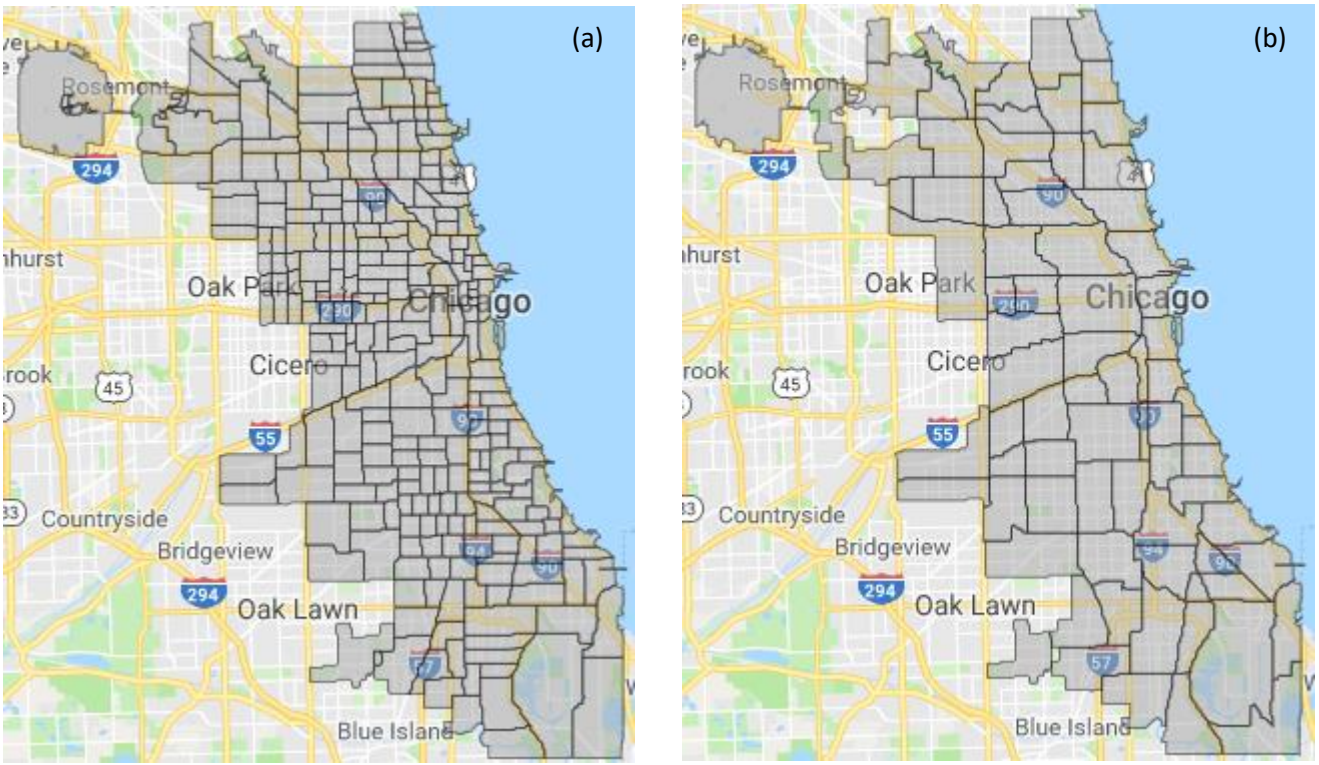

Figure 3: Boundaries for (a) police beats (City of Chicago, 2018b) and (c) communities (City of Chicago, 2018a).

Latitude/Longitude:

Per dropna(), there were 60,175 null values for latitude and longitude. I created a regular expression matching the syntax of the latitude/longitude and checked that all entries followed that format. No irregular locations were found this way. I then plotted the latitude and longitude and found points well southwest of the Chicago Area (Figure 4).

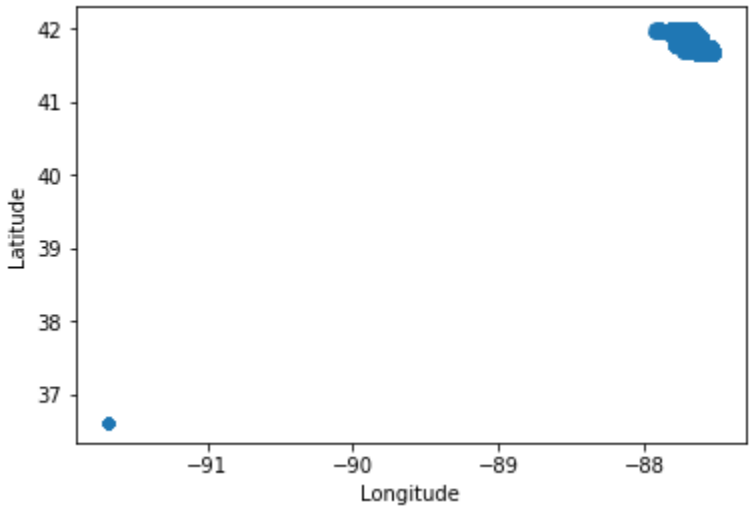

Figure 4: Plot of latitude vs. longitude in degrees.

After examining some of these points, it appeared that the latitude/longitude of 36.619446/-91.686566 degrees was given to several crime reports. It may be the case that this position was used when the latitude/longitude was not noted. I made these positions null values and revisited them later. In total, there were 60,336 reports missing latitude/longitude.

X Coordinate/Y Coordinate:

Per dropna() there were 60,175 null values for the coordinates. I created a regular expression matching the syntax of the coordinates and checked that all entries followed that format. No irregular coordinates were found this way. I then plotted the X coordinate and Y coordinate and again found points well southwest of the Chicago Area (Figure 5).

It appeared that when the latitude/longitude were not known, the X/Y coordinates were made 0. I made these null values and revisited them later. In total, there were 60,336 reports with missing X/Y coordinates.

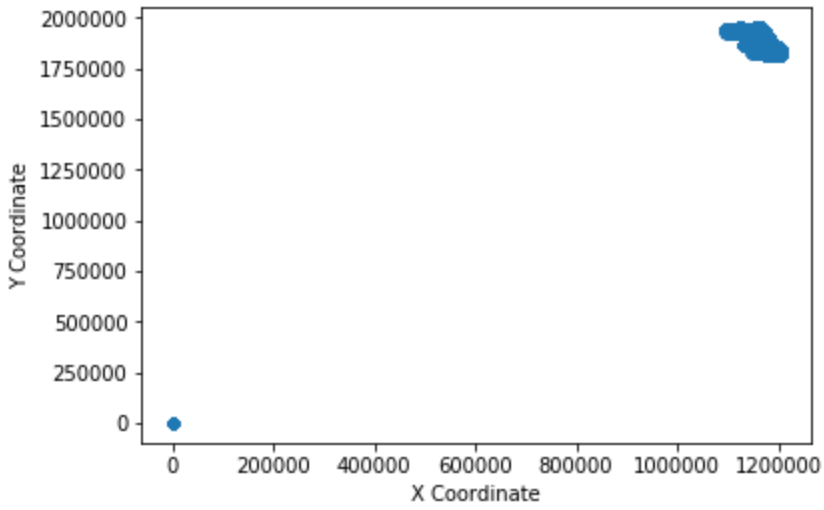

Figure 5: Plot of Y Coordinate vs. X Coordinate.

Missing Latitude/Longitude/X Coordinate/Y Coordinate:

There were 60,336 reports missing location. I decided that the block would be a small enough feature that would likely estimate the location fairly well. First I created a dataframe with reports missing location and then made a list of unique blocks. For each block, I looked up all non-null reports with that same block, picked a report at random, and linked its location details to the block. I then iterated through these blocks and filled in the missing locations in the dataframe I created. By doing this, I reduced the number of reports missing the location to 2464.

In future research, perhaps geocoding could be used on the blocks to get most of the missing locations.

Missing Ward and Community:

There were 614,846 reports missing ward. I used the latitude/longitude to figure out the ward. I read in a csv file with coordinates of the boundaries of each ward as a dataframe. For each ward, the coordinates were in the form of a long string. In order to make them usable, I converted them to a list of tuples. I created a dataframe with reports missing ward and iterated through it. For each known latitude/longitude in the dataframe, I used the Shapely package to figure out which ward polygon it was located in. By doing this, I reduced the number of reports missing ward to 2952.

One issue for finding the ward this way is the ward boundaries shift after each federal census (Knox, 2005). I used ward boundaries that were effective from 2015 onwards. Perhaps in future research, all of the wards in the dataset could be redone so that they all use the same boundary system.

There were 616,113 reports missing community. I used the same technique as above to figure out the community. By doing this, I reduced the number of reports missing community to 2710.

I examined the police district counts for reports that had a location but were missing ward or community and found that police districts 16 and 24 had the most reports missing ward or community. They are the north/northwestern most districts and district 16 (the northwestern most district) is made up of two areas that

have a gap between them. Therefore it makes sense that the number of reports missing ward or community was higher than those missing location since there likely were several positions that fell outside of these boundaries.

Adding Columns:

Using the 'Date' column, columns for the month, day of the month, day of the week, day type (weekday/weekend), and hour were found. Using month, a column for seasons was created with winter (December, January, February), spring (March, April, May), summer (June, July, August), and fall (September, October, November). Using month, a column for quarter of the year was also created with Q1 (January, February, March), Q2 (April, May, June), Q3 (July, August, September), Q4 (October, November, December). Using the day of the month, a column for the third of the month was created with T1 (days 1-10), T2 (days 11-20), and T3 (days 21-31). Using the hour, a column for time of day was created with overnight (hours 0-5), morning (hours 6-11), afternoon (hours 12-17), and evening (hours 18-23).

Using the 'Block' column, a column for street was created by splitting up the block into 2 pieces and taking the second piece with the direction and street name.

I used the following coordinates for the Chicago city center: 41.881832, -87.623177. These coordinates are from https://www.latlong.net/place/chicago-il-usa-1855.html. Using the haversine formula, I created a new column with the distance between each crime report and the Chicago city center.

Bus Stops

This dataset consisted of 21 columns and 10,916 rows with each row being a bus stop. There were no apparent null values for latitude/longitude and the bus stop name. There was a 'Status' column which showed if the bus stop was still in service. There were several stops that were flagged, but an online search of a few of these stops showed that they were still in service. I therefore kept all of the bus stops. I then plotted all of the locations of the bus stops (Figure 6). There were a couple of far western bus stops west of -87.85. Further investigation showed that they were valid UPS stops.

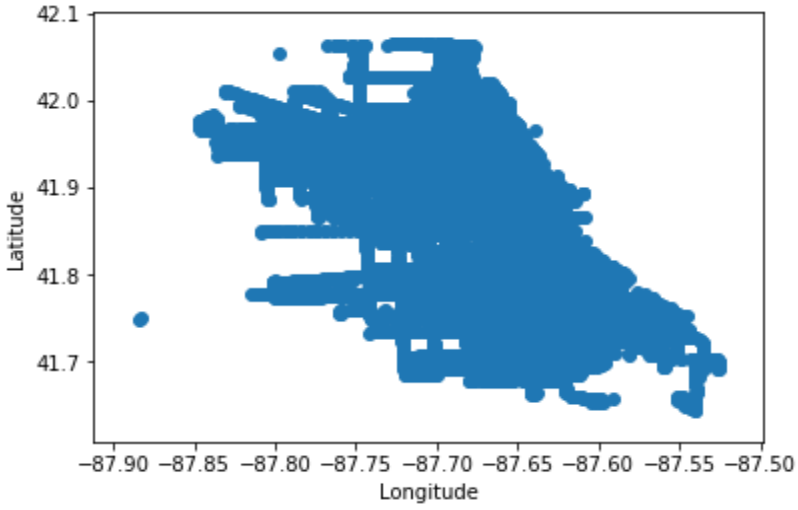

Figure 6: Plot of latitude and longitude in degrees of bus stops.

I used a ball tree query from Sklearn to find the closest bus stop and the distance from the closest bus stop for each report.

Train Stops

This dataset consisted of 17 columns and 300 rows with each row being a train stop. There were no apparent null values for the station names and locations. There was a column with a descriptive name of the station which included the train line in parentheses. I extracted the train line and created a new column for it. The location column contained a tuple of the latitude and longitude. I extracted the latitude and longitude into separate columns. I then plotted all of the locations of the train stops (Figure 7) and all looked good.

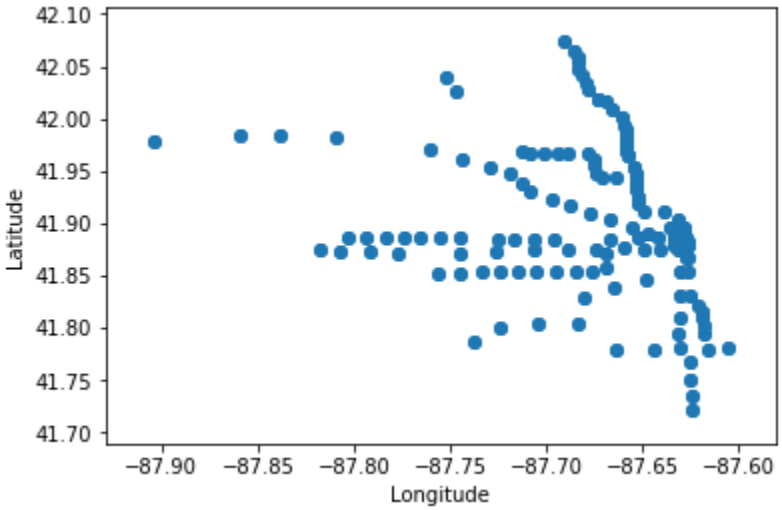

Figure 7: Plot of latitude and longitude in degrees of train stops.

I used a ball tree query to find the closest train stop, the associated train line, and the distance from the closest train stop for each report.

Liquor Stores

There were 34 columns and 954,489 rows with each row being a business license. The columns I used were the legal name, latitude/longitude, and address. There was a column that showed the license status, but I decided to look at all of the businesses as they were likely operating at some point during the period of 2001-2018.

There were 4 businesses with missing legal names. Fortunately, these businesses had nothing to do with liquor, so I removed them. In order to find liquor stores, I searched for businesses that contained one of the following terms in their legal names: liquor, spirits, wine, or alcohol. I sorted the resulting data frame by legal name and saw that there were businesses that were duplicates as they had to regularly apply for licenses. The duplicates were dropped.

There were 4 liquor stores that did not have a latitude/longitude. I used the website https://www.latlong.net/convert-address-to-lat-long.html to convert their addresses to latitude/longitude.

One store was located outside of the Chicago Area so I deleted it. I then plotted all of the locations of the liquor stores (Figure 8) and all looked good.

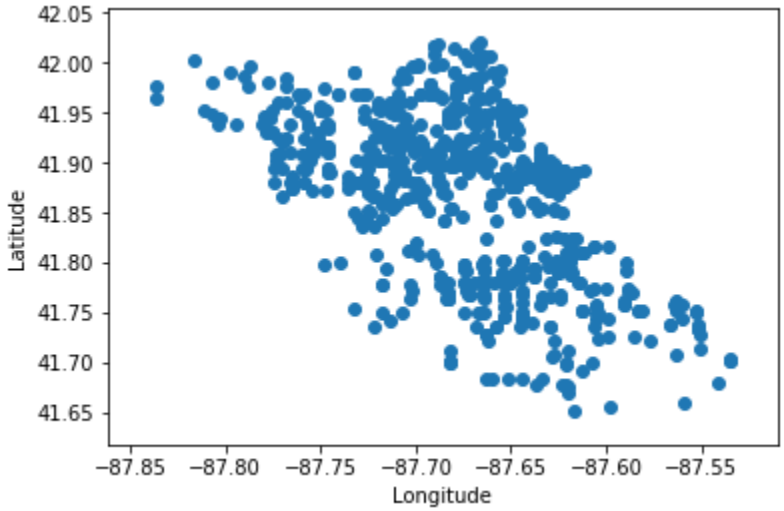

Figure 8: Plot of latitude and longitude in degrees of liquor stores.

I used a ball tree query to find the closest liquor store and the distance from the closest liquor store for each crime report.

Police Stations

This dataset contained 9 columns and 22 rows with each row being a police station (district). The columns I used were district and location. The police headquarters was listed as a district in this dataset. As the headquarters was not used anywhere in my Chicago crime dataset, I removed the headquarters. The location column contained a tuple of the latitude and longitude. I split this column using .split and created new columns for the latitude and longitude. I then plotted all of the locations of police stations (Figure 9) and all looked good.

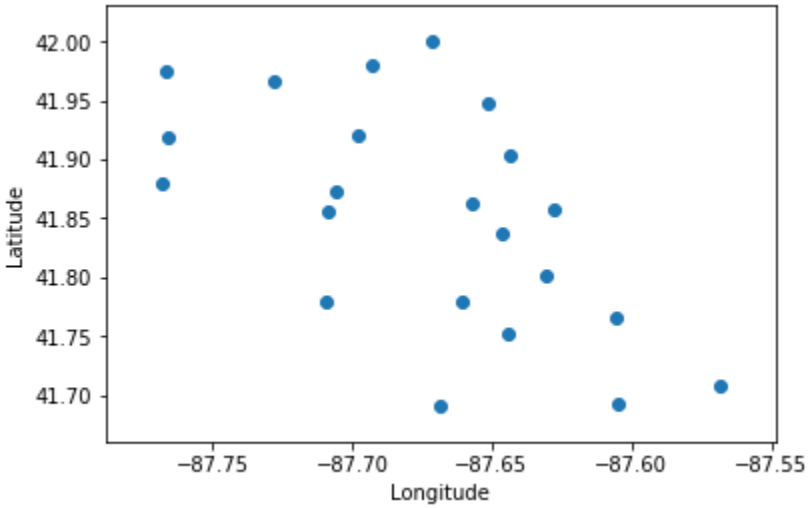

Figure 9: Plot of latitude and longitude in degrees of police stations.

I used a ball tree query to find the closest police station and the distance from the closest police station for each crime report.

Federal Holidays

This dataset contained 2 columns and 485 rows with each row being a federal holiday. The 2 columns were date and holiday. I first created a new dataset with just the holidays occurring during 2001 through 2018. I then counted how many holidays occurred each year. All years except for 2010 and 2011 had 10 holidays. It turned out that 2010 had an extra occurrence of New Year's Day and 2011 was missing New Year's Day. After fixing this, I checked out a list of the unique holidays and saw that Martin Luther King Jr. Day, New Year's Day, and Washington's Birthday had different notations. These issues where fixed.

I then merged the dataframe of holidays with the dataframe of crime reports and filled the null values in the holiday column with 'No Holiday'. I also created a new column called 'Is Holiday' which said if a specific day was a holiday or not.

References

City of Chicago. (2018a). Boundaries – Community Areas (current) [map]. Retrieved from https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6

City of Chicago. (2018b). Boundaries – Police Beats (current) [map]. Retrieved from https://data.cityofchicago.org/Public-Safety/Boundaries-Police-Beats-current-/aerh-rz74

City of Chicago. (2018c). Boundaries – Police Districts (current) [map]. Retrieved from https://data.cityofchicago.org/Public-Safety/Boundaries-Police-Districts-current-/fthy-xz3r

City of Chicago. (2016). Boundaries – Wards (2015-) [map]. Retrieved from https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Wards-2015-/sp34-6z76

Knox, D. (2005). *Encyclopedia of Chicago.* Retrieved from http://www.encyclopedia.chicagohistory.org/pages/1316.html