Introduction

Per Figure 1, for the period of 2002-2016, Chicago has had, for the most part, crime rates above the U.S. average. Judging by how much above average Chicago is, it is likely that Chicago has had above average crime rates for a much longer period than 2002-2016. In an effort to mitigate crime rates in Chicago, I would like to figure out any spatial/temporal factors that would be useful in predicting the type of crime committed.

The Chicago police department would by my main client. If they were aware of connections between spatial/temporal factors and the type of crime committed, perhaps officers may be better able to anticipate the type of crime that may occur at a given area/time and either prevent the crime from happening or more easily catch the perpetrator. Depending on what type of crime would be most prevalent, the police department could act accordingly to better prepare themselves; for example, deciding how many officers are required to patrol in a certain area and what specific skills they would need to deal with the prevalent crime in that area. In being prepared, this could also help cut down on police injuries/fatalities.

In addition, the Department of Transportation could use data found in this study. For example, if there was an area that had a significant amount of theft, perhaps the installation of more streetlights could help with the problem.

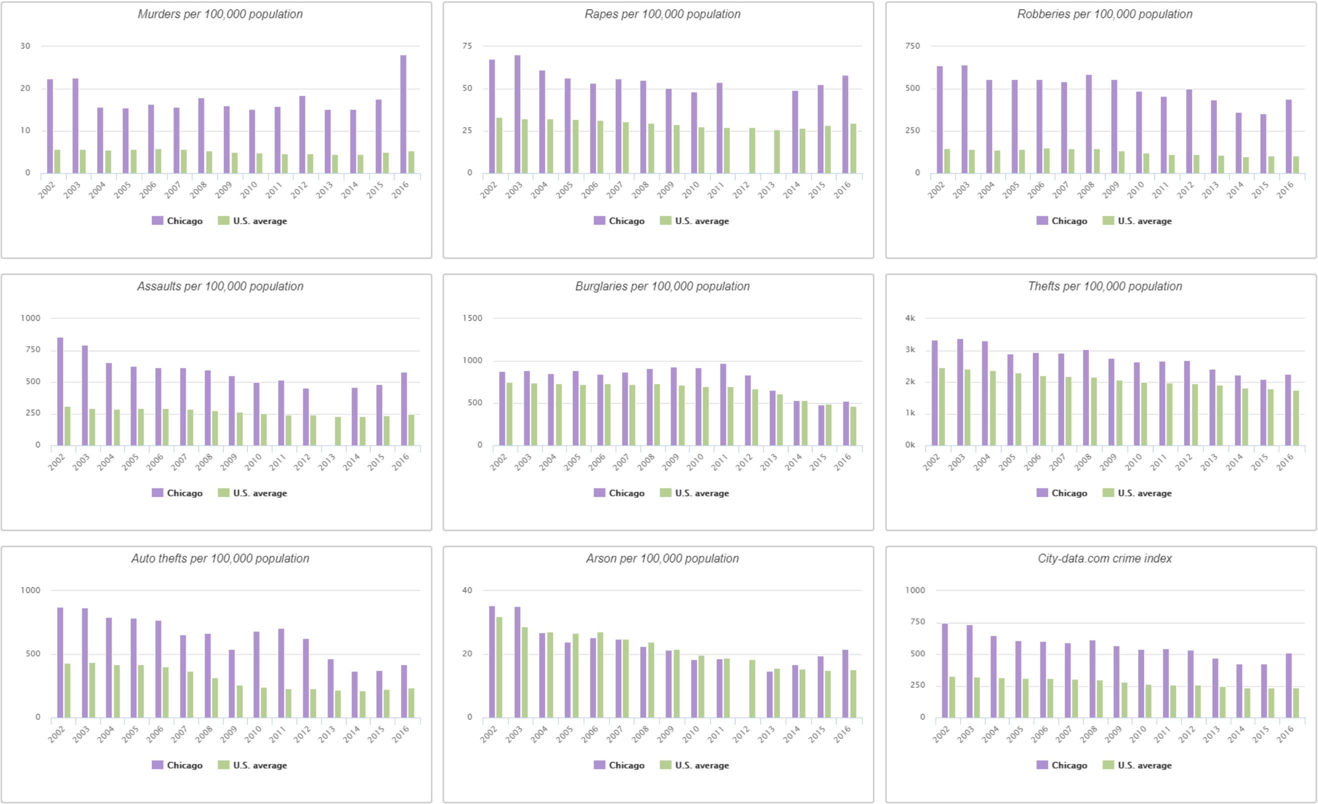

Figure 1: Crime in Chicago compared to the U.S. average (City-Data, 2018).

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 (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
- 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
- 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 2), 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.

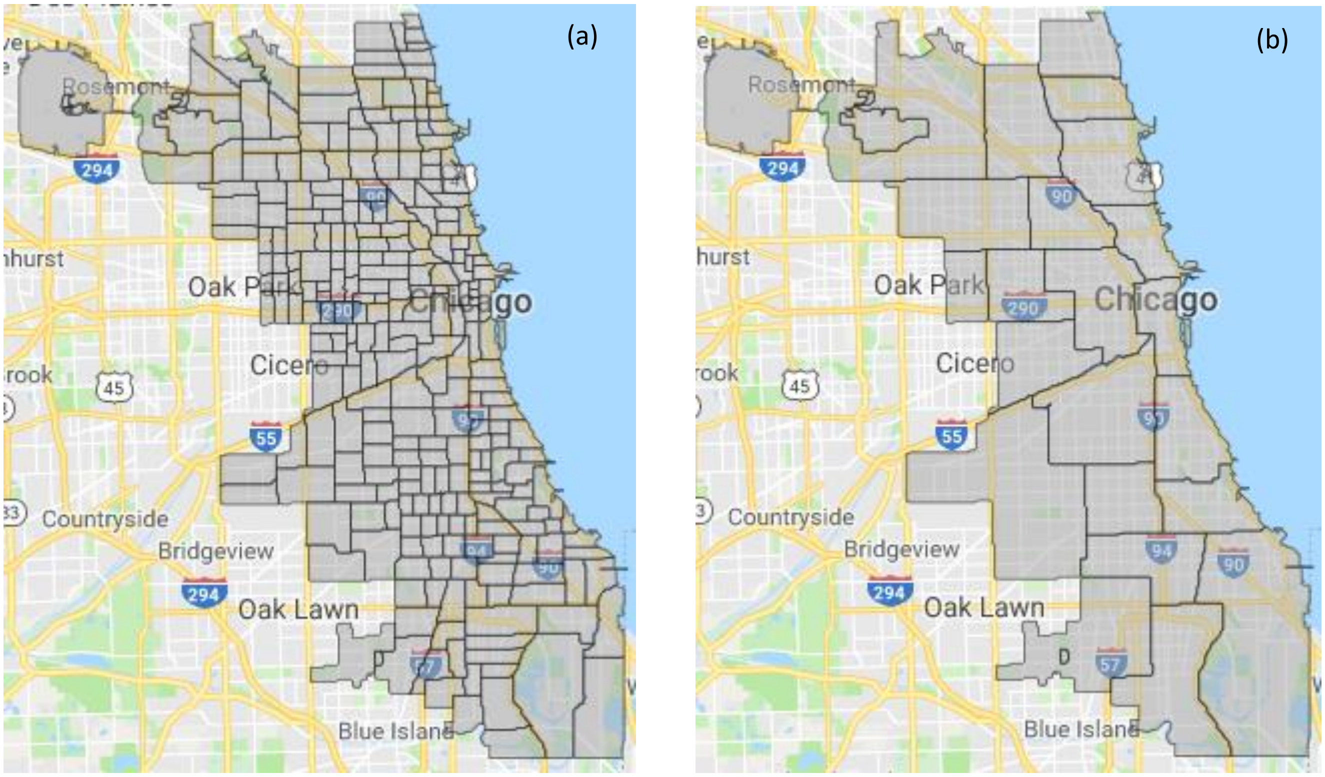

Figure 2: Boundaries for (a) police beats (City of Chicago, 2018b) and (b) police districts (City of Chicago, 2018c).

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 3), 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 2002.

Examining the police beat boundaries and community boundaries (Figure 4), 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.

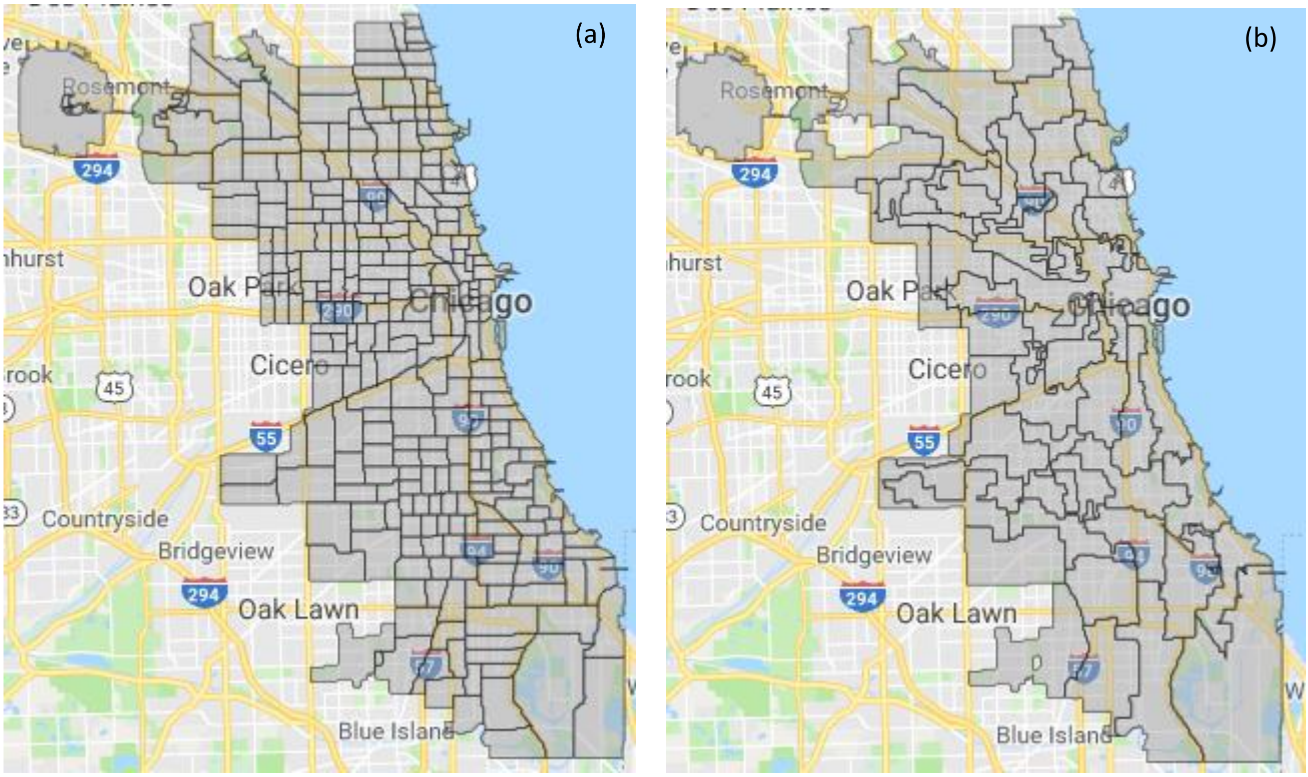

Figure 3: Boundaries for (a) police beats (City of Chicago, 2018b) and (b) wards (City of Chicago, 2016).

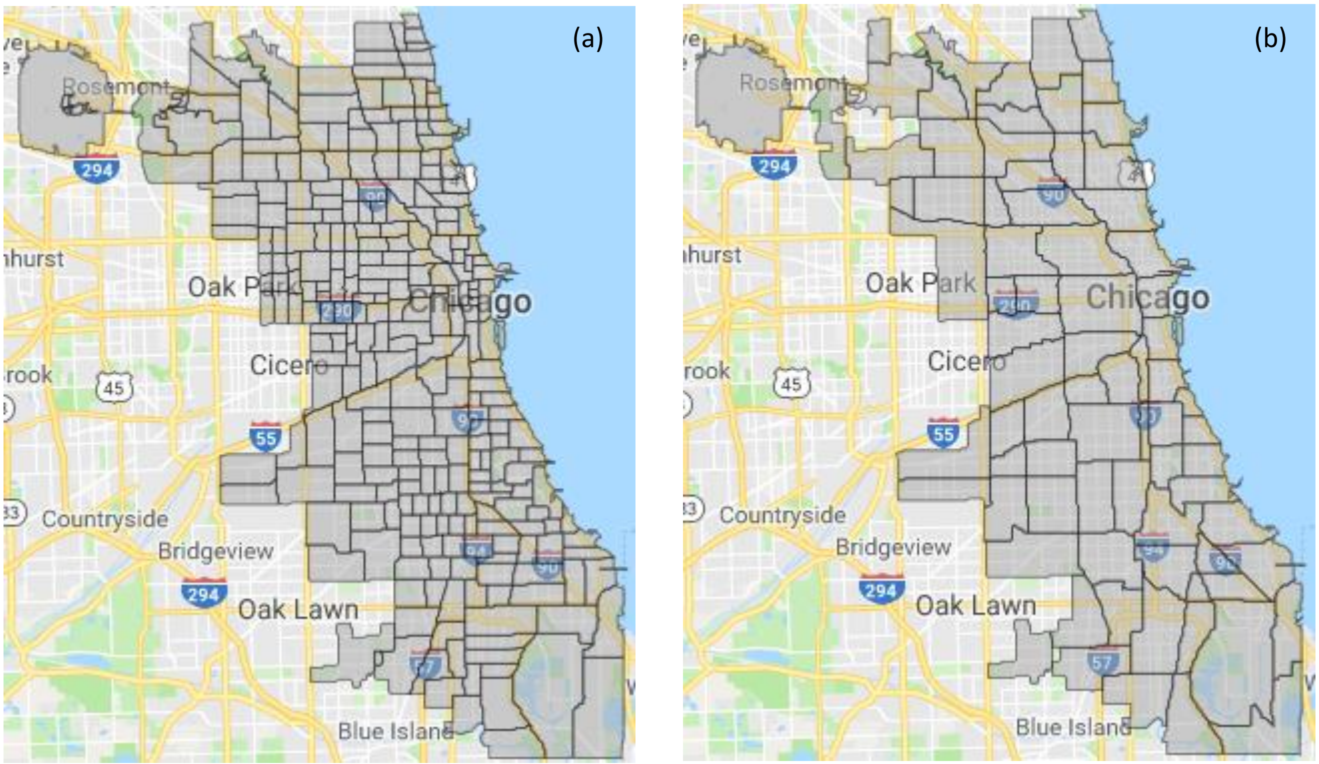

Figure 4: Boundaries for (a) police beats (City of Chicago, 2018b) and (c) communities (City of Chicago, 2018a).

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 5).

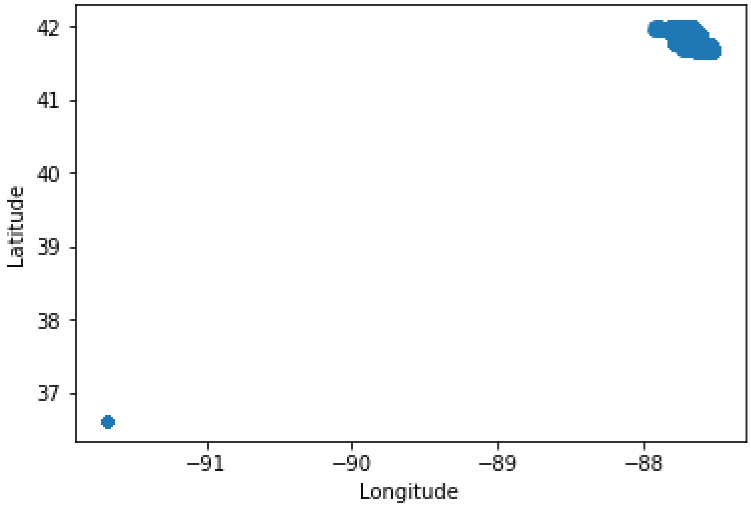

Figure 5: Plot of latitude vs. longitude in degrees.

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 6).

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.

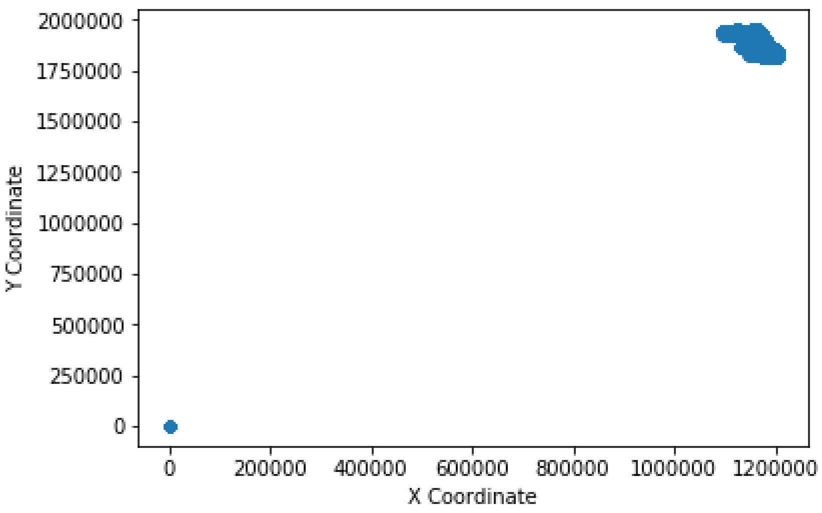

Figure 6: Plot of Y Coordinate vs. X Coordinate.

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 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 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 7). There were a couple of far western bus stops west of -87.85. Further investigation showed that they were valid UPS stops.

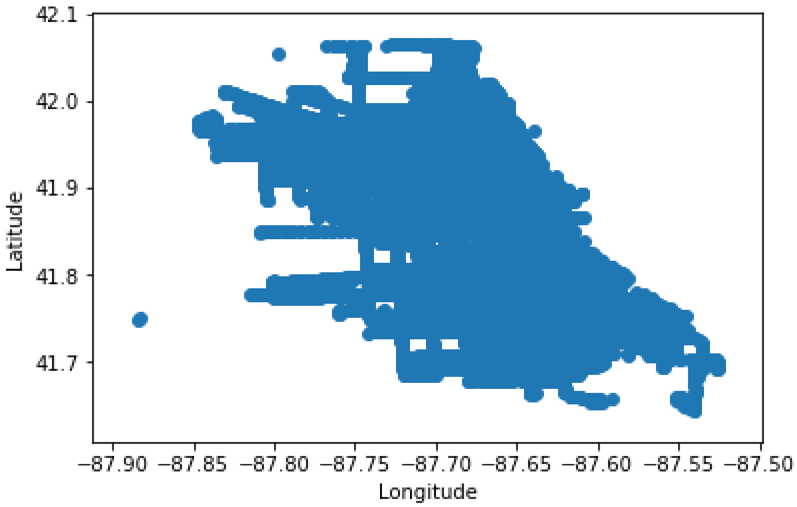

Figure 7: Plot of latitude and longitude in degrees of bus stops.

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 8) and all looked good.

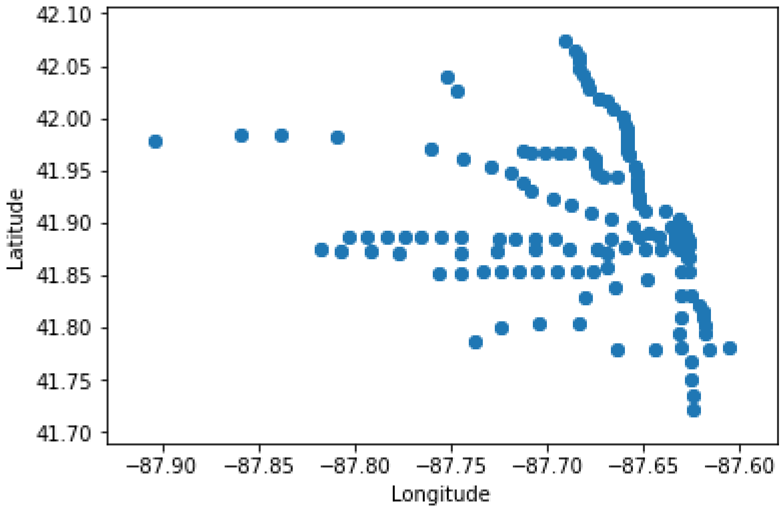

Figure 8: Plot of latitude and longitude in degrees of train stops.

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 9) and all looked good.

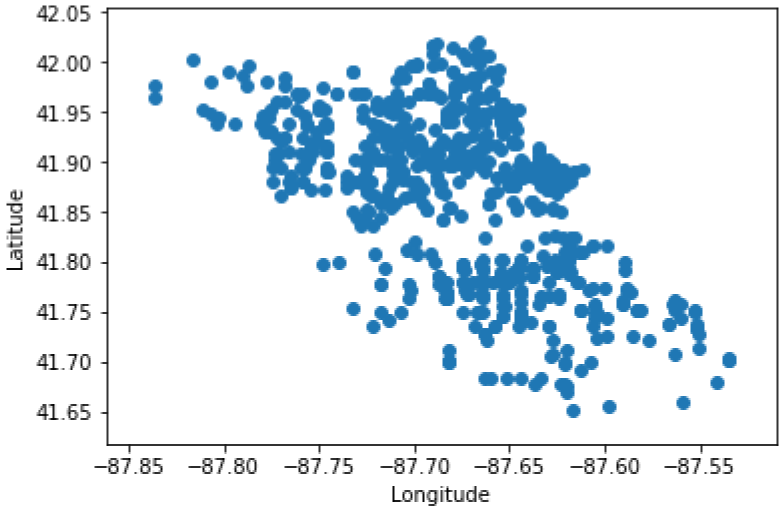

Figure 9: Plot of latitude and longitude in degrees of liquor stores.

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 10) and all looked good.

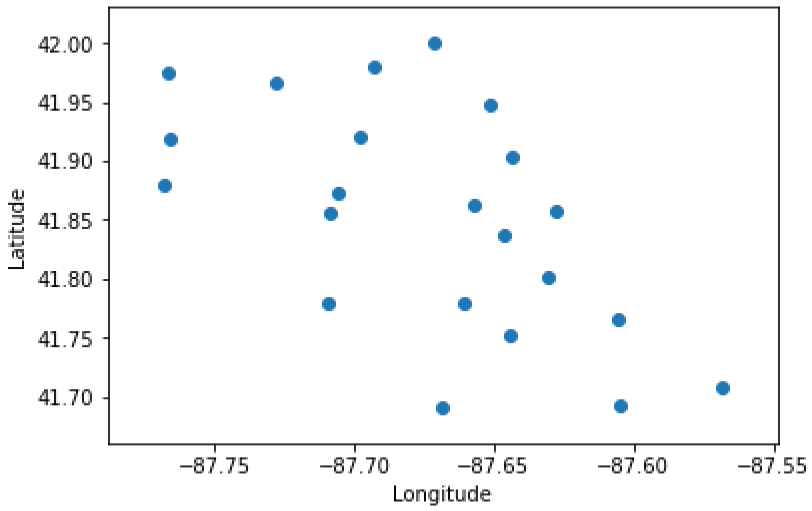

Figure 10: Plot of latitude and longitude in degrees of police stations.

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.

Exploratory Data Visualization and Analysis

With over 1.2 million crimes each, theft and battery were the most frequent crimes (Figure 11). I limited my study to the top 11 primary types of crime (theft to criminal trespassing) as there are sufficient samples for them.

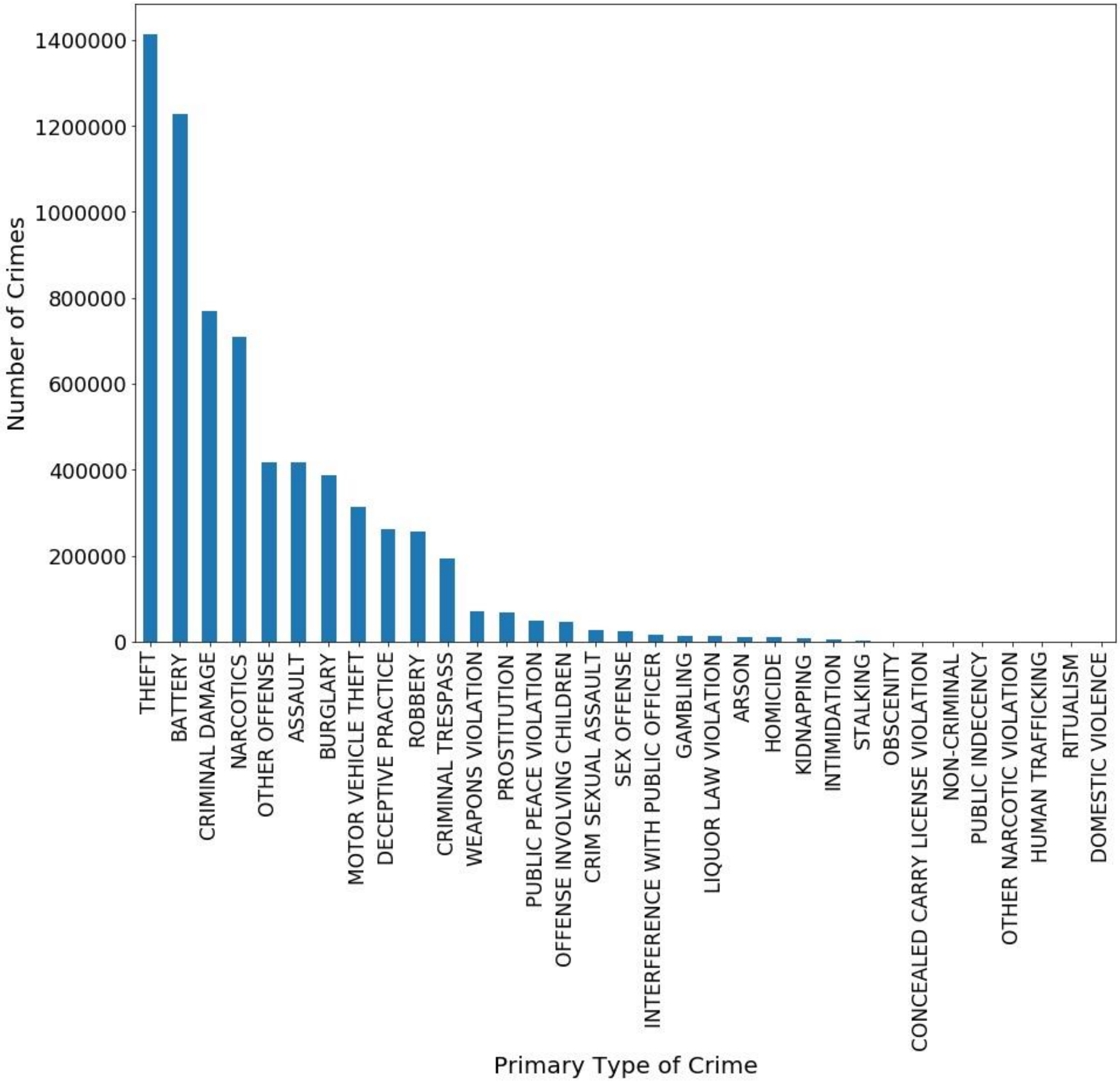

Figure 11: Number of crimes for each primary type of crime.

Looking at crimes involving theft, battery, and narcotics, there is quite some variation in their proportions within each ward (Figure 12), police district (Figure 13), police beat (Figure 14), and community (Figure 15). Wards 42, 43, and 32 have the highest proportions of theft. In several wards, approximately 20-25% of the total number of crimes involve battery. Wards 28 and 24 have the highest proportions of crimes involving narcotics.

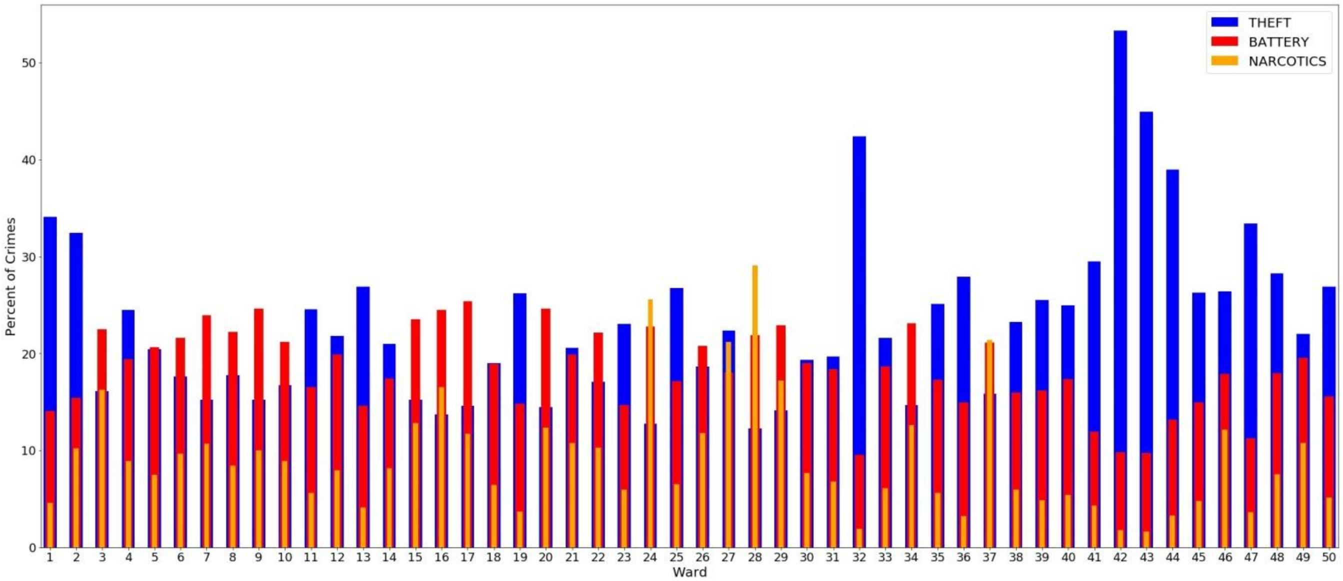

Figure 12: Percentage of 3 primary types of crime per ward.

Police districts 1 and 18 have the highest proportions of theft. Districts 7 and 5 have the highest proportions of battery. Districts 11 and 15 have the highest proportions of crimes involving narcotics.

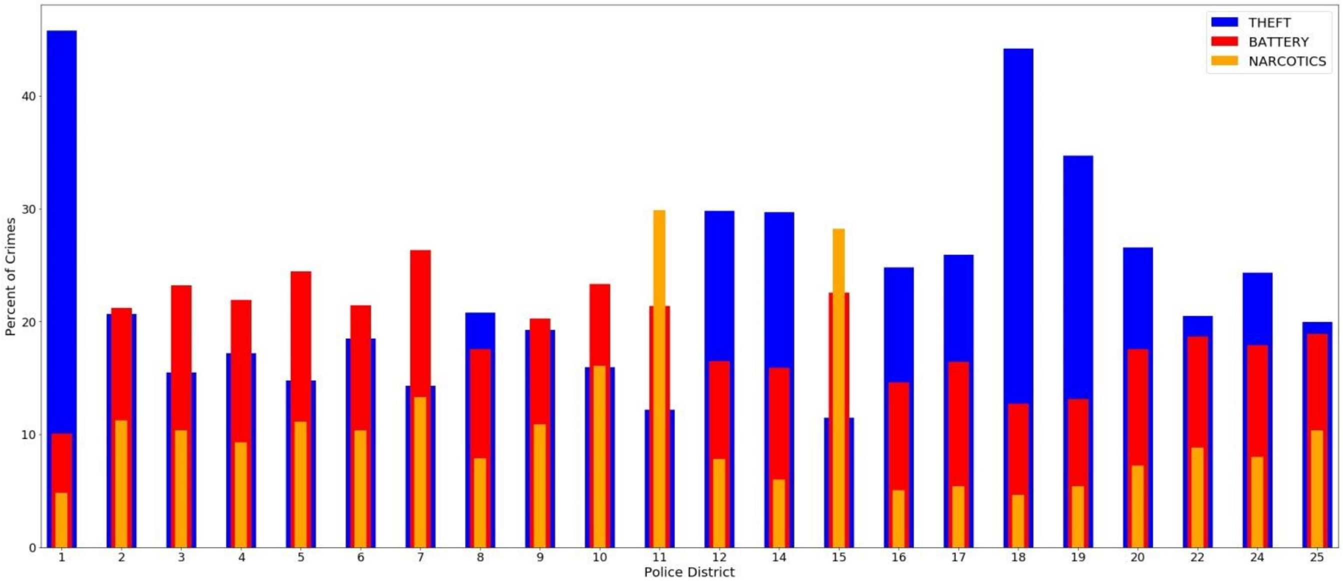

Figure 13: Percentage of 3 primary types of crime per police district.

From the first one or two number of the police beat number, you can tell which police district it is in. For example, beats 111 through 134 are all a part of district 1. In Figure 14, in all but one beat in district 1, close to or over 40% of the total number of crimes involve theft. For beats within district 2 (beats 211-235), crimes involving battery or narcotics tend to dominate except for in beat 235 where theft is responsible for a significant portion of the crimes. So in addition to the police district, it is important to look at the police beat.

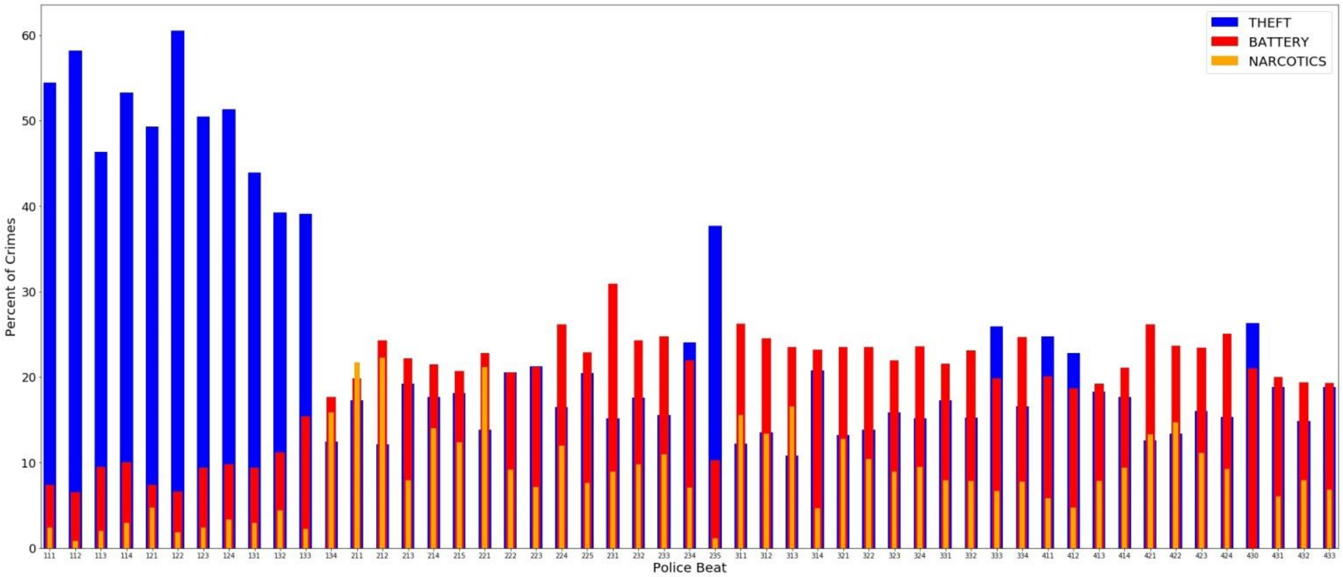

Figure 14: Percentage of 3 primary types of crime for a sample of police beats.

Community 32 has the highest proportion of theft while community 54 has the highest proportion of battery. Communities 23, 25-27, and 29 have the highest proportions of crimes involving narcotics.

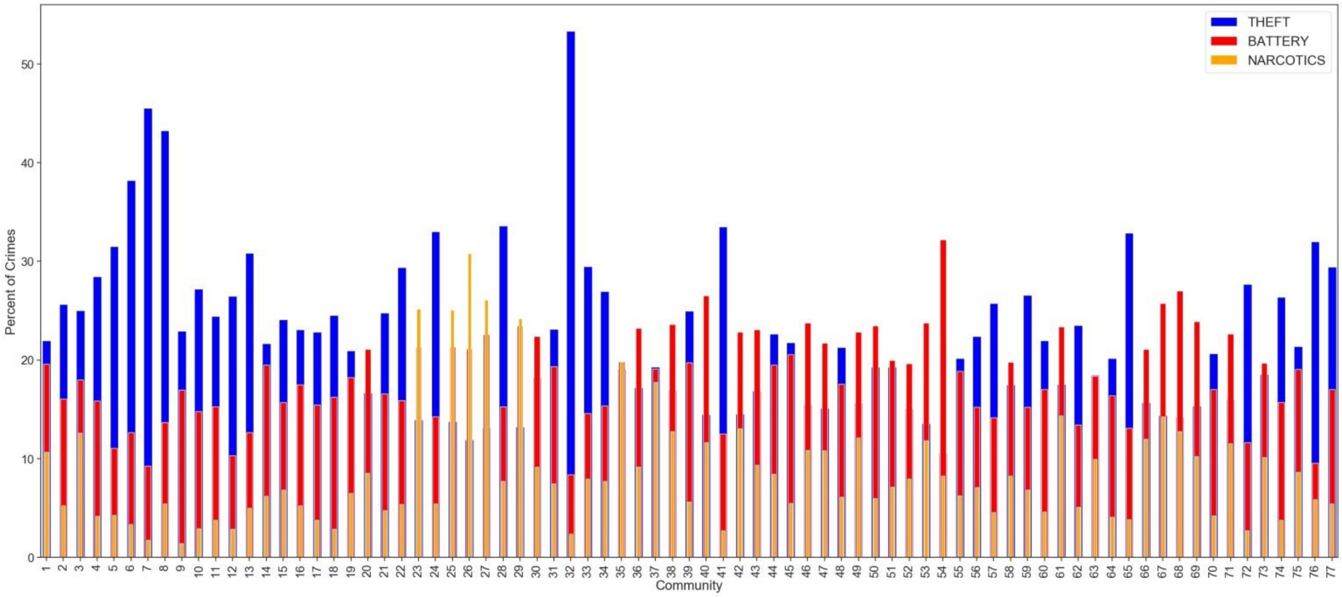

Figure 15: Percentage of 3 primary types of crime within each community.

Per Figure 16, the 4 locations with the most crime are street, residence, apartment, and sidewalk. The breakdown of crimes for these locations is shown in Figure 17. Apartments have the highest proportion of battery and very low proportions of motor vehicle theft and robbery. Residences have a high proportion of battery and very low proportions of motor vehicle theft and robbery. Sidewalks have the highest proportion of crimes involving narcotics and very low proportions of burglary, criminal trespassing, deceptive practice, and motor vehicle theft. Streets have the highest proportion of theft and very low proportions of burglary, criminal trespassing, and deceptive practice.

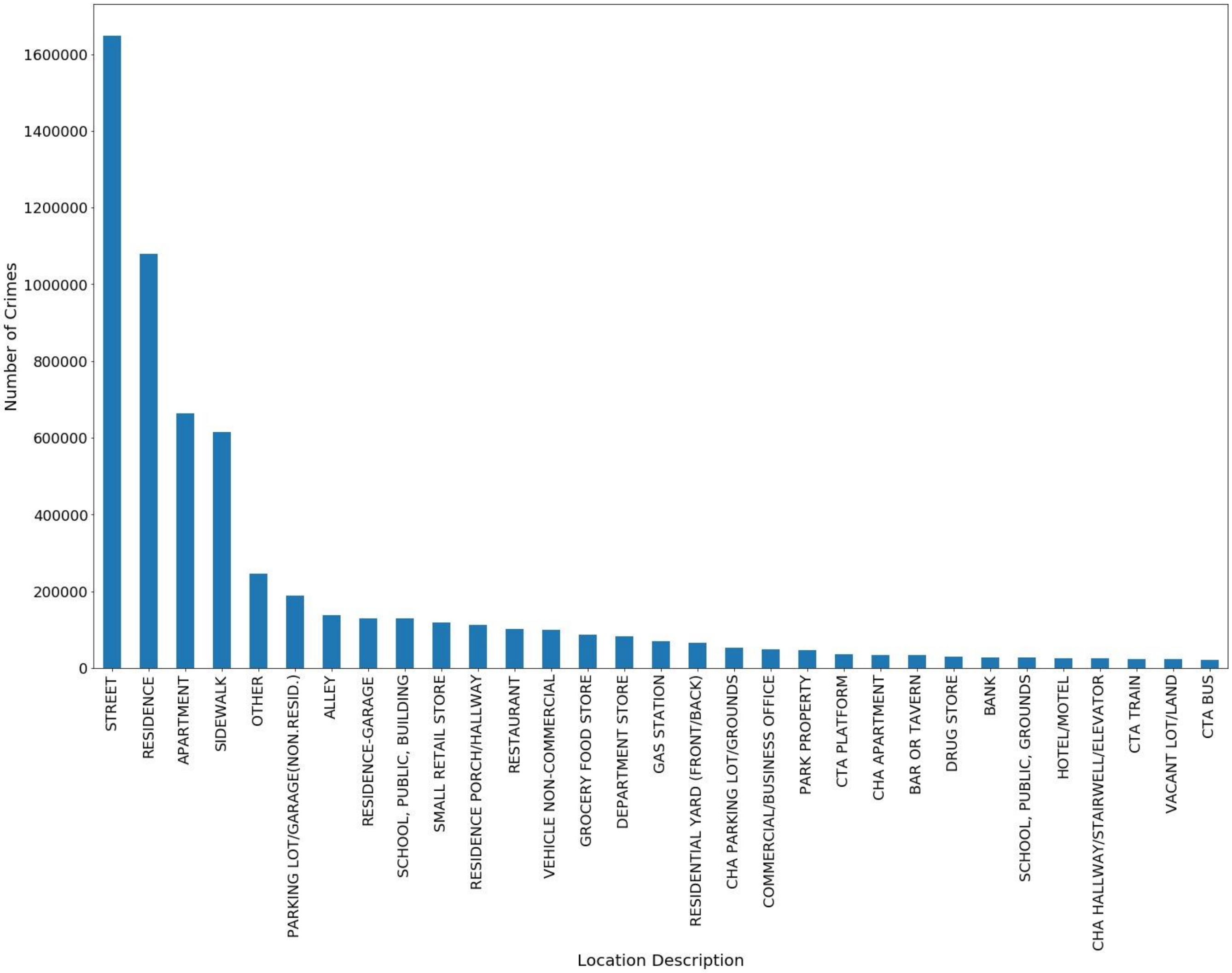

Figure 16: Number of crimes for each location description.

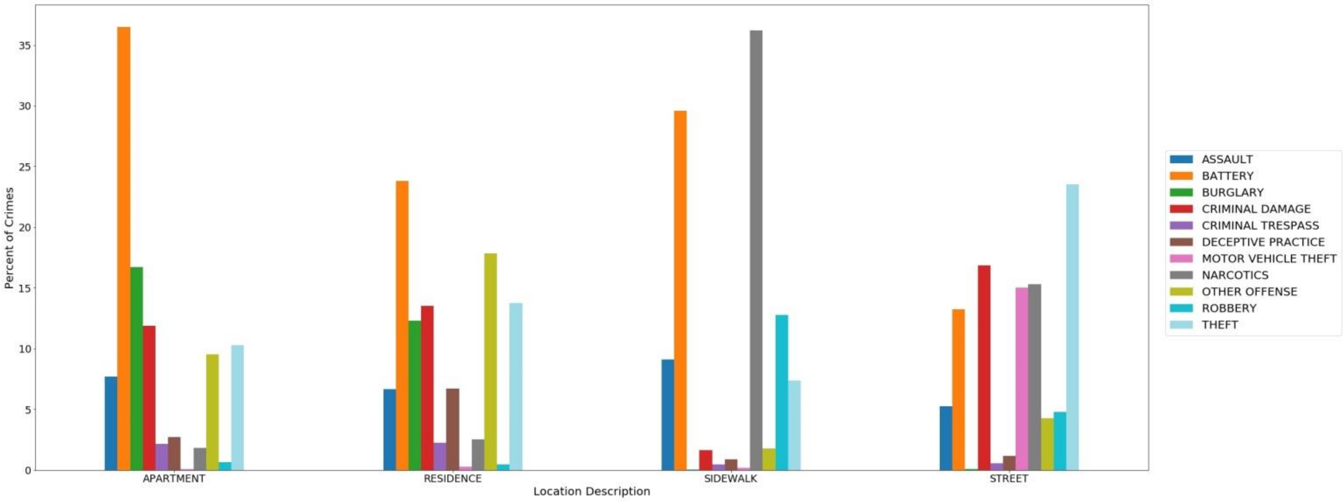

Figure 17: Percentage of each primary type of crime for each location description.

There could possibly be some relationship between the block or street and the type of crime; however, it would not be feasible to use these features as they have 57,758 and 3814 unique blocks and streets, respectively.

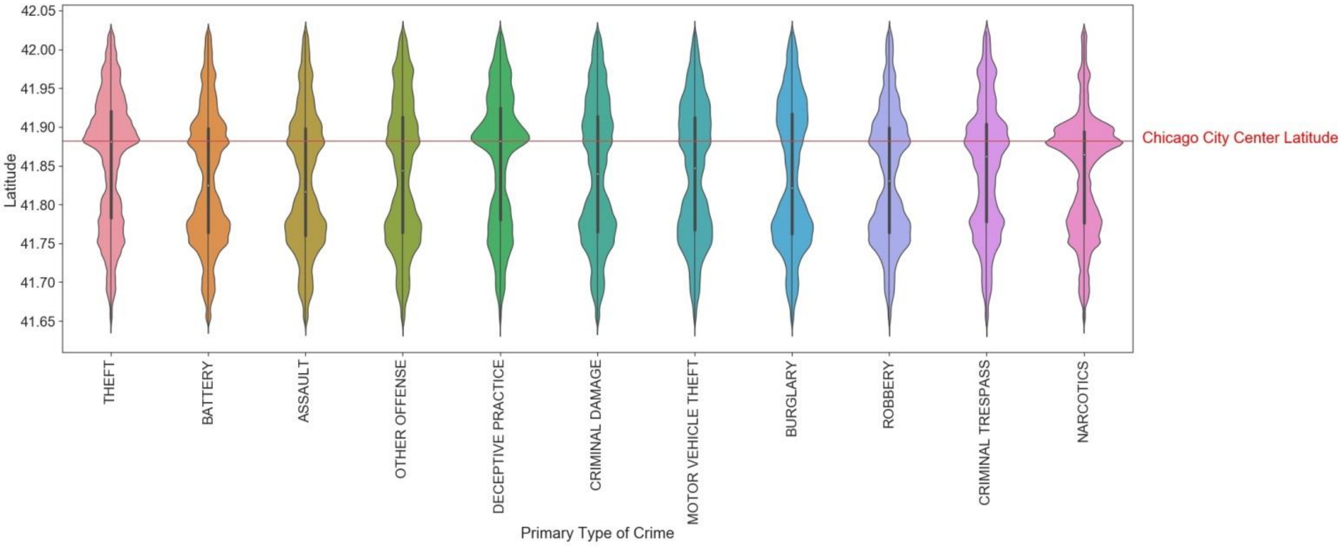

Figure 18 shows that while each primary type of crime generally has a bimodal distribution of latitudes, there are still differences between them. Crimes involving theft, deceptive practice, and criminal trespassing, and narcotics have higher numbers of crime towards the northern part of the city. For theft, deceptive practice, and narcotics especially, there is a pronounced increase in the number of crimes near the latitude of the city center of Chicago.

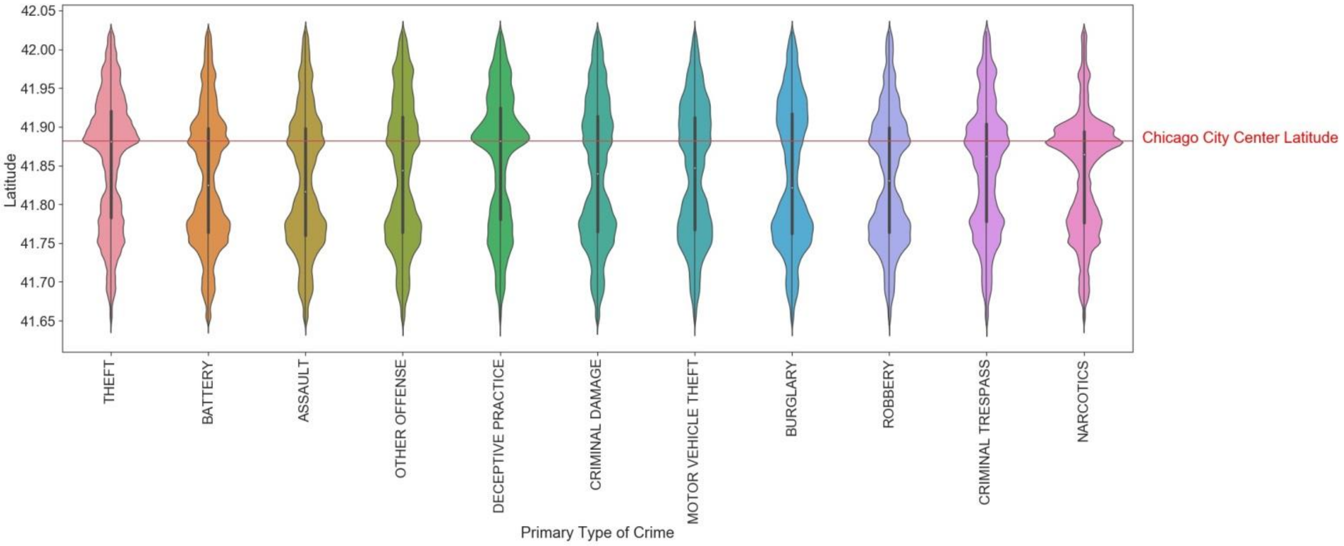

Figure 18: Distribution of crimes across latitude for each primary type of crime.

In Figure 19, the distributions have fat tails to the left. Crimes involving theft, deceptive practice, and criminal trespassing have a pronounced increase in the number of crimes near the longitude of the city center of Chicago. A slight increase in the number of crimes can be seen near -87.9° for crimes involving theft, other offenses, deceptive practice, and criminal trespassing. This may be due to these types of crimes occurring in a far northwestern community (community 76). For crimes involving narcotics, there is a pronounced increase in the number of crimes just west of -87.7°.

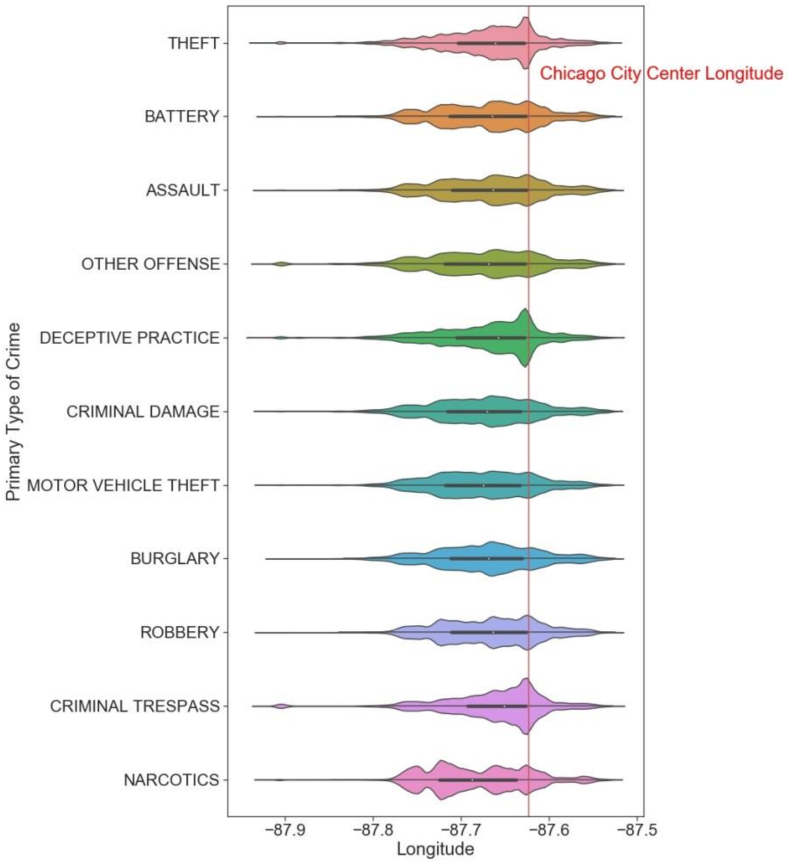

Figure 19: Distribution of crimes across longitude for each primary type of crime.

Figure 20 generally shows different spatial distributions of crimes for each primary type of crime. All crime types have a high concentration of crimes along Lake Michigan (the eastern edge of the map). For crimes involving narcotics, there are some gaps in the concentration of crimes along the lake and slightly farther inland on the north side. Crimes involving theft, battery, criminal damage, and other offenses have higher concentrations of crime in the northwest area of Chicago (between -87.7° and -87.8°).

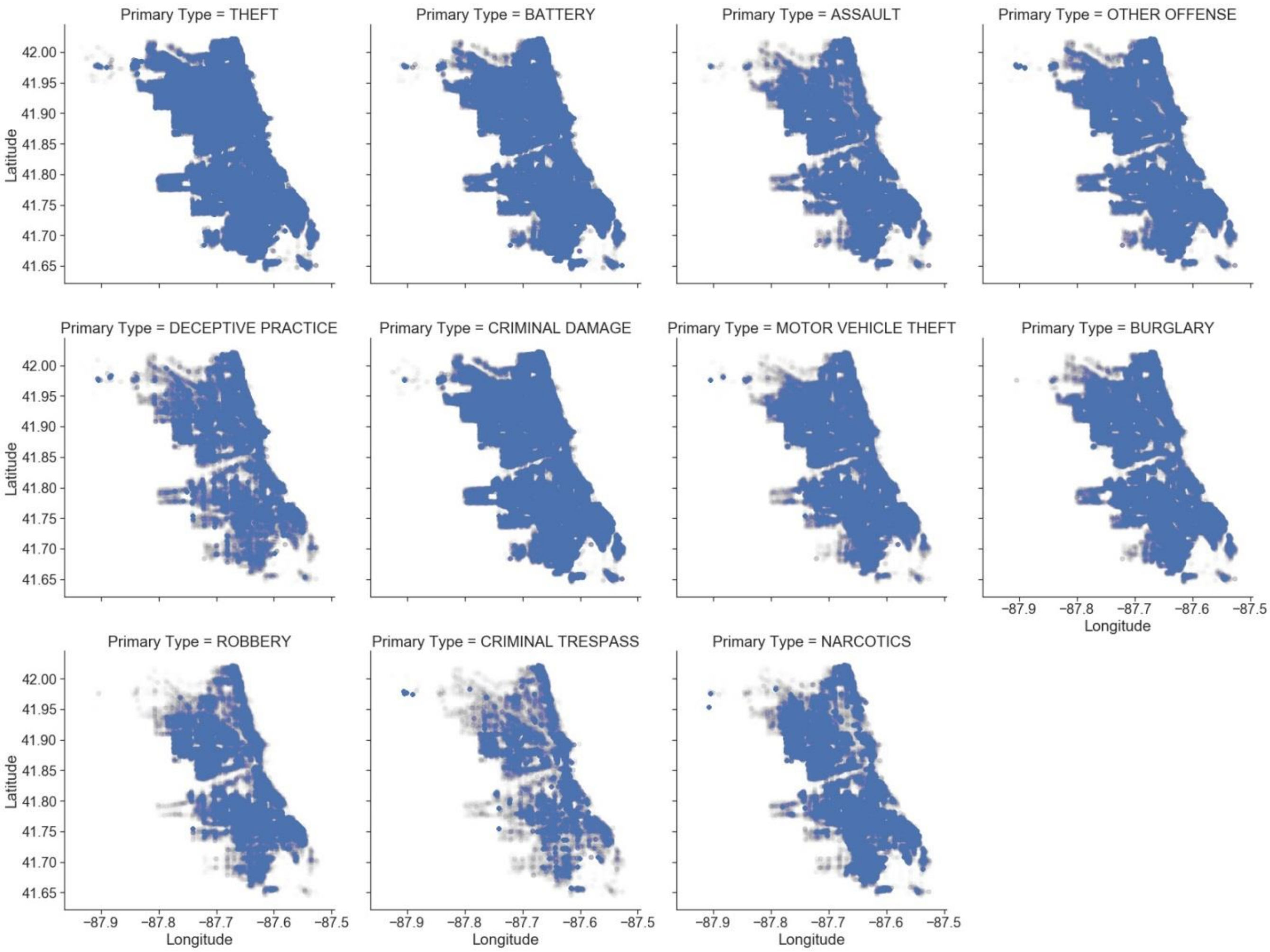

Figure 20: Spatial distribution of crimes by primary type of crime.

Figure 21 shows that the distribution of distance from Chicago varies for each primary type of crime. Theft, deceptive practice, and criminal trespassing have higher concentrations of crime closer to the city center (within 5km). Theft and deceptive practice have bimodal distributions with one maximum within 3km of the center of Chicago. The number of crimes then plateaus from 5 to 15km away from the city center before decreasing. For criminal trespassing, the number of crimes does not vary significantly from 0 to 15km and then decreases. The remainder of the crime types have somewhat normal distributions with a slight skew to the right except for narcotics, which appears to have a multimodal distribution.

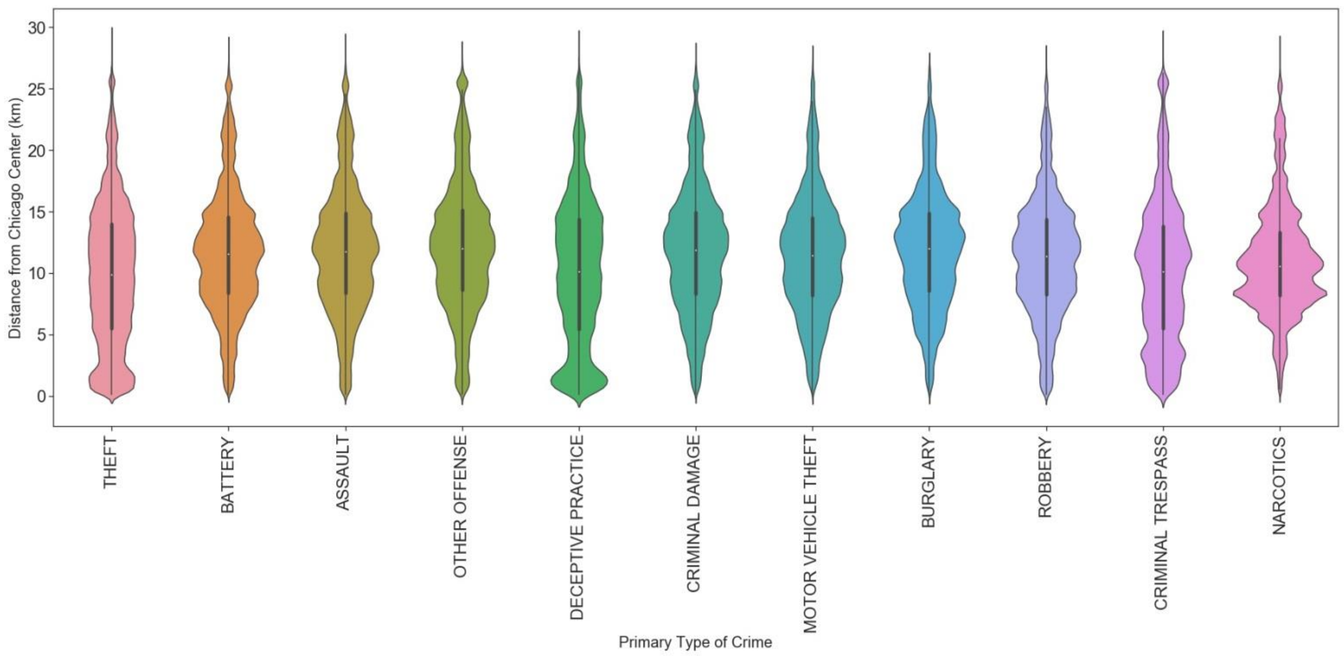

Figure 21: Distribution of crimes based on distance from Chicago city center for each primary type of crime.

In Figure 22, there isn't a significant amount of difference between the distribution of crime and the distance from the closest police station for each type of crime. The average distance from the closest police station for each crime is approximately 2km. All of the distributions are fat tailed with the bulk of crimes occurring at around 0 to 4km away from a police station. An examination of the average distance from the police stations by community (Figure 23) shows that crimes with the highest distances may have mainly occurred in community 76, which is a bit more removed to the northwest from the remainder of Chicago.

Figure 24 shows that there is no significant improvement in the tails of the distributions after excluding crimes within community 76. It is likely that more communities would have to be removed in order for there to be a considerable improvement in the distributions.

Taking the square root of the distance from the closest police station for crimes in all communities helped reduce the tails (Figure 25). All of the distributions are skewed to the right, but the variations between them are slightly more apparent when using the square root of the distance.

As I am already using the police district where the crime occurred, it would be redundant to use the closest police district as there are no significant differences between the two.

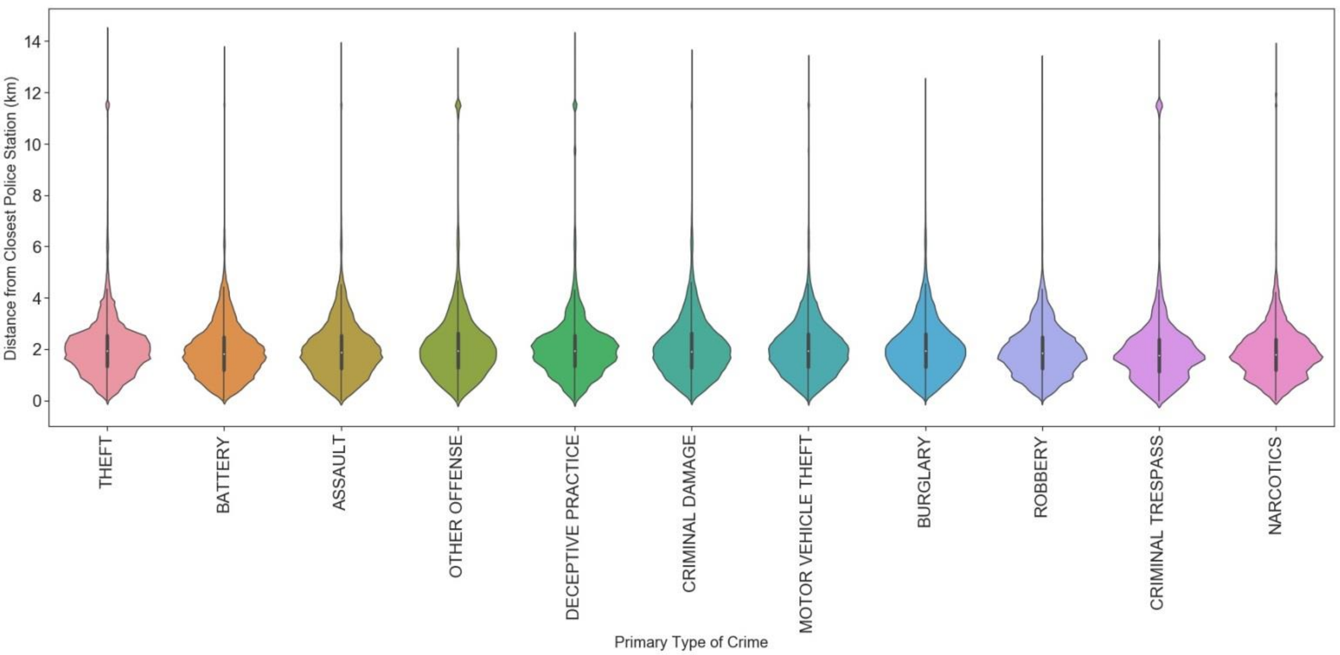

Figure 22: Distribution of crimes based on distance from closest police station for each primary type of crime.

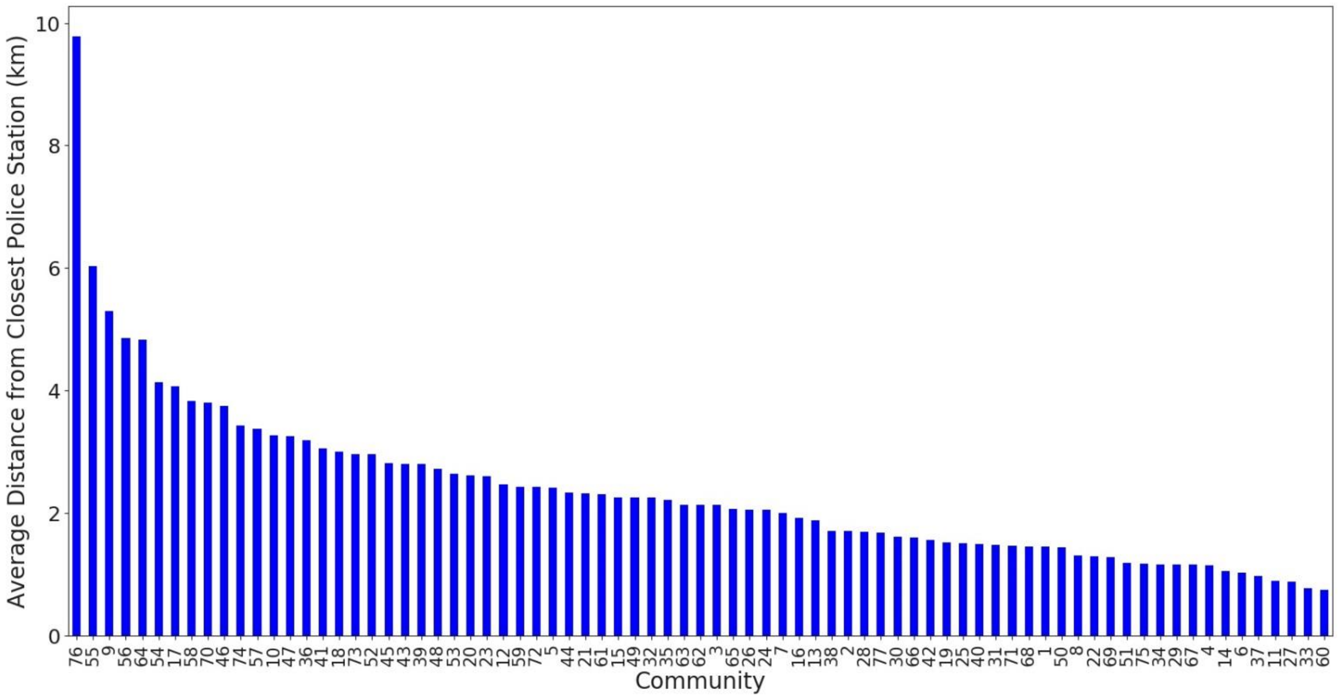

Figure 23: Average distance from closest police station for each community.

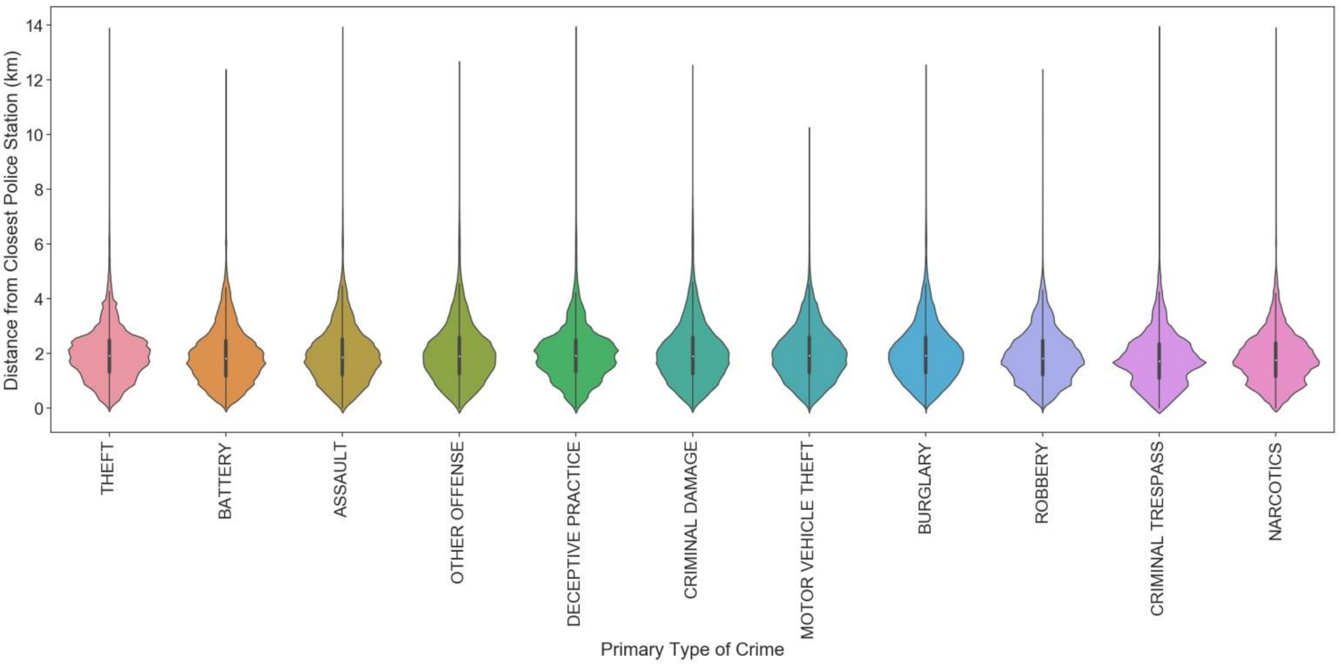

Figure 24: Distribution of crimes based on distance from closest police station for each primary type of crime (excluding community 76).

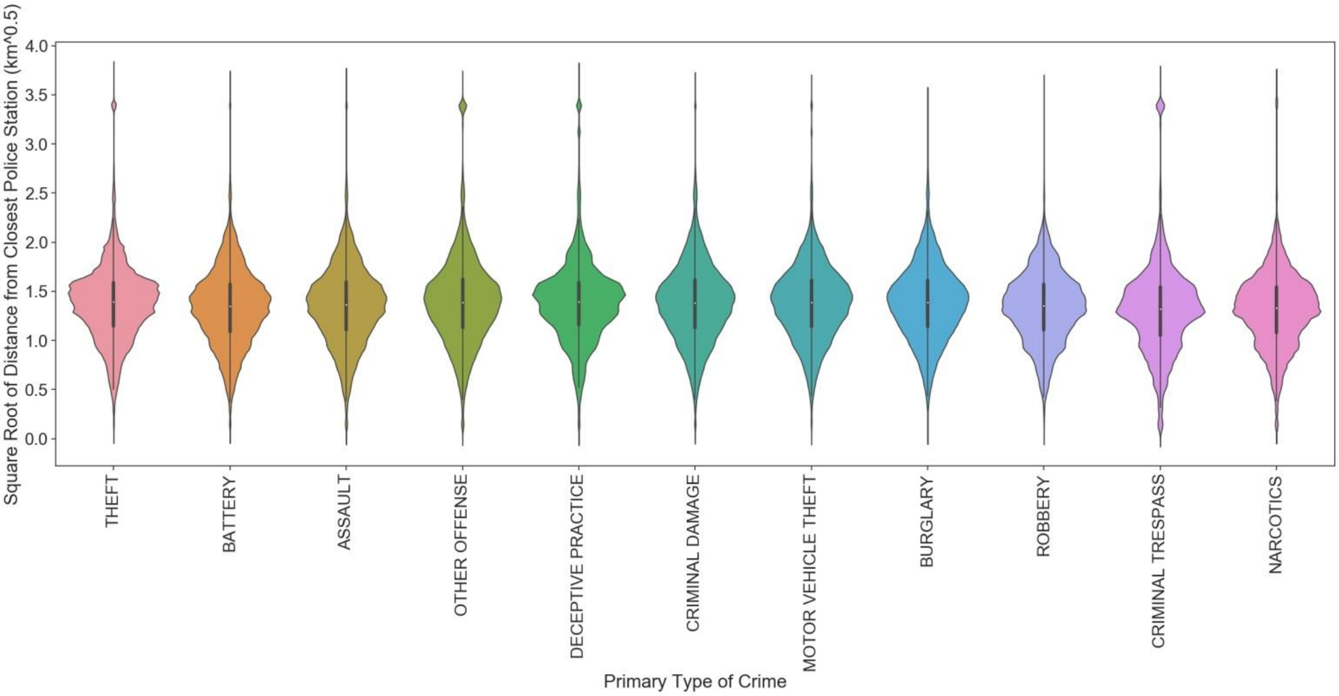

Figure 25: Distribution of crimes based on square root of distance from closest police station for each primary type of crime.

Figure 26 shows that all of the distributions of crimes have fat tails and on average, crimes occur approximately 1km away from train stops. The bulk of crimes occur within 2km of train stops. There are higher concentrations of theft, deceptive practice, and criminal trespassing closer to train stops.

In order to reduce the tails of the distributions, the square root of the distance from the closest train stop was taken and then plotted in Figure 27. There are some variations in the distributions, but not as much as in the original plot.

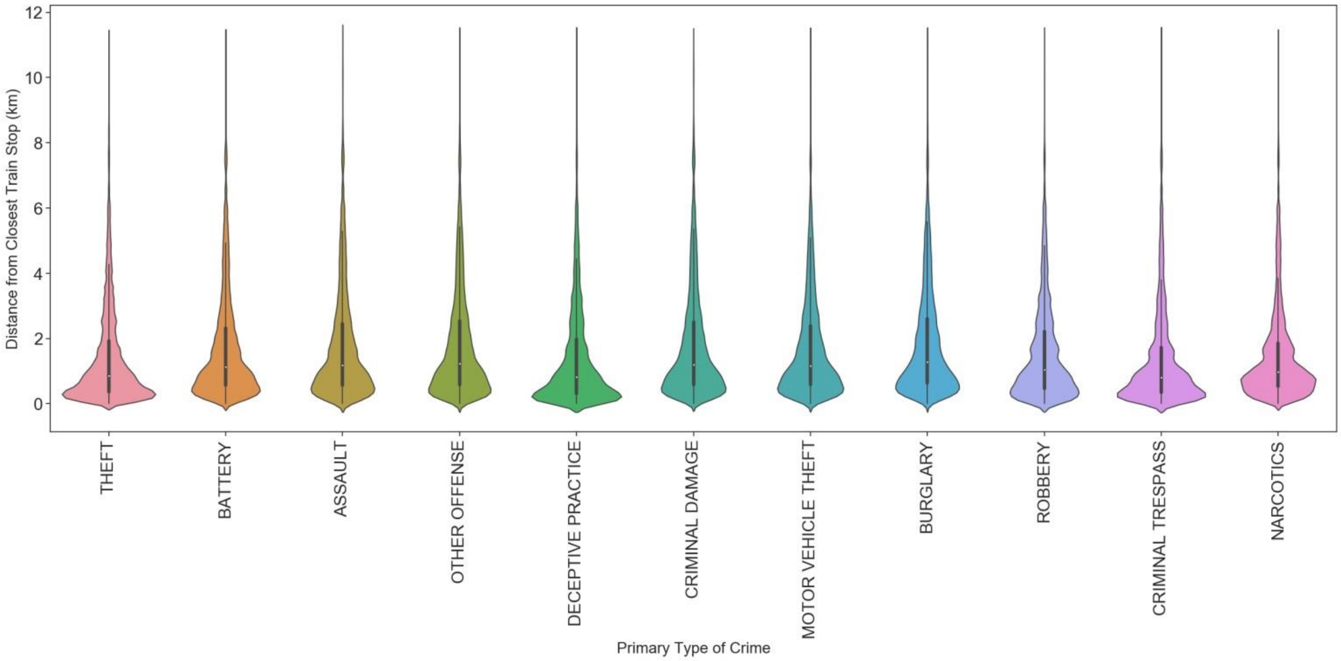

Figure 26: Distribution of crimes based on distance from closest train stop for each primary type of crime.

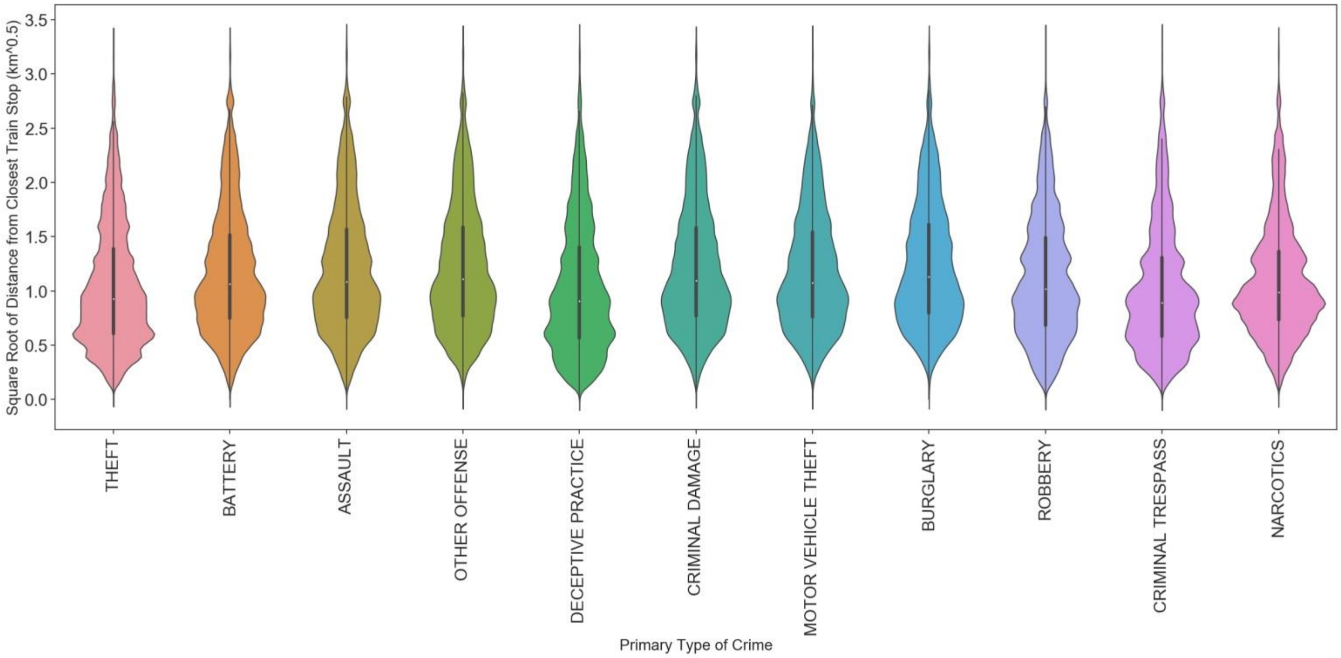

Figure 27: Distribution of crimes based on square root of distance from closest train stop for each primary type of crime.

According to Figure 28, the most crimes occurred near stops associated with the Blue, Green, Orange, and Red Lines. Figure 29 breaks down the proportion of each crime type for these 4 lines. Stops associated with the Blue Line have a high proportion of theft and then battery. Stops associated with the Green Line have a high proportion of battery and then theft/narcotics. Stops associated with the Orange Line have a high proportion of theft and then battery. Stops associated with the Red Line have a high proportion of theft and then battery.

It is possible that there is some relationship between the actual train stop and type of crime. However, there are 100 unique train stops and it would be better to simplify this and just use the closest train line in order to reduce the number of features.

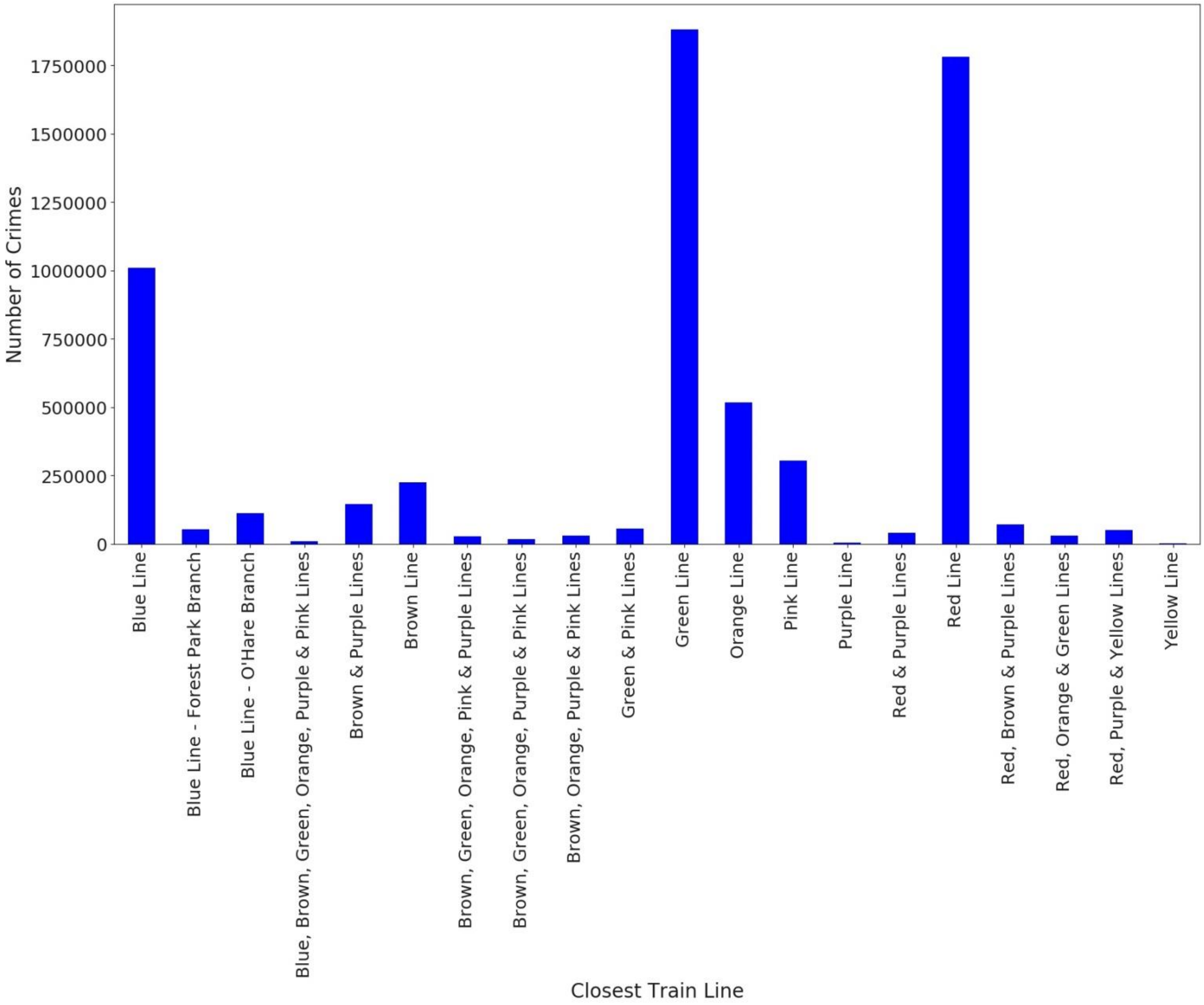

Figure 28: Number of crimes per closest train line.

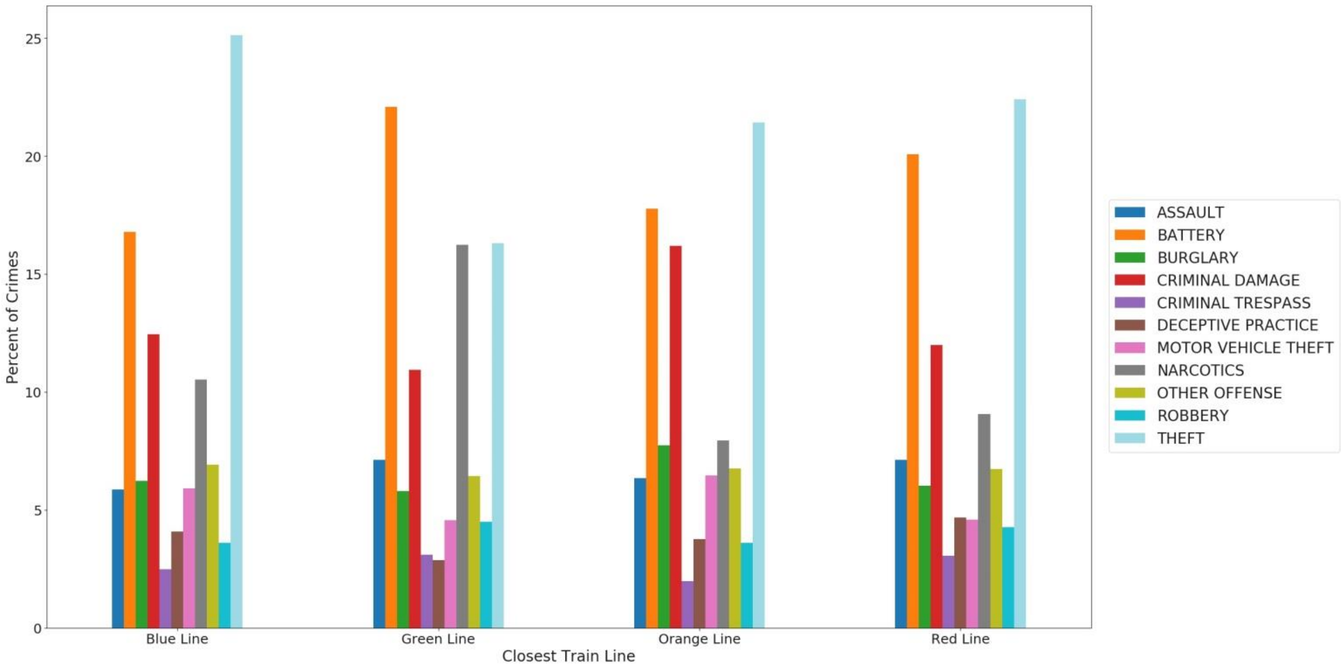

Figure 29: Percentage of each primary type of crime per closest train line.

Figure 30 shows that all of the distributions of crimes have fat tails and the bulk of them occur within 0.5km of a bus stop. The highest concentration of crimes is found extremely close to bus stops; this is especially true for robbery. In order to reduce the tails of the distributions, the square root of the distance from the closest bus stop was taken and plotted in Figure 31. Here we can see that the distributions for all crime types except for deceptive practice and criminal trespassing are multimodal. Though most of the distributions are multimodal, more differences can be noted between them than when comparing the original distributions.

It is possible that the closest bus stop could be somewhat useful to predict the type of crime. However, as there are 5,832 unique bus stops, this would not be feasible.

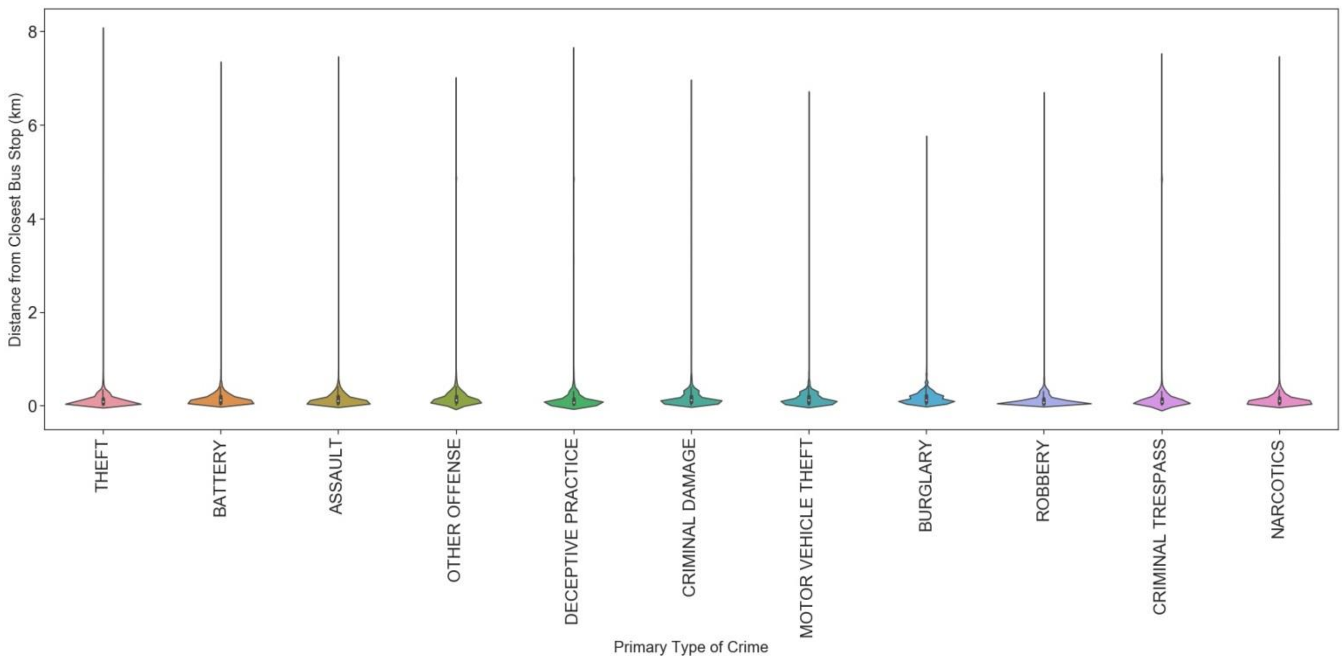

Figure 30: Distribution of crimes based on distance from closest bus stop for each primary type of crime.

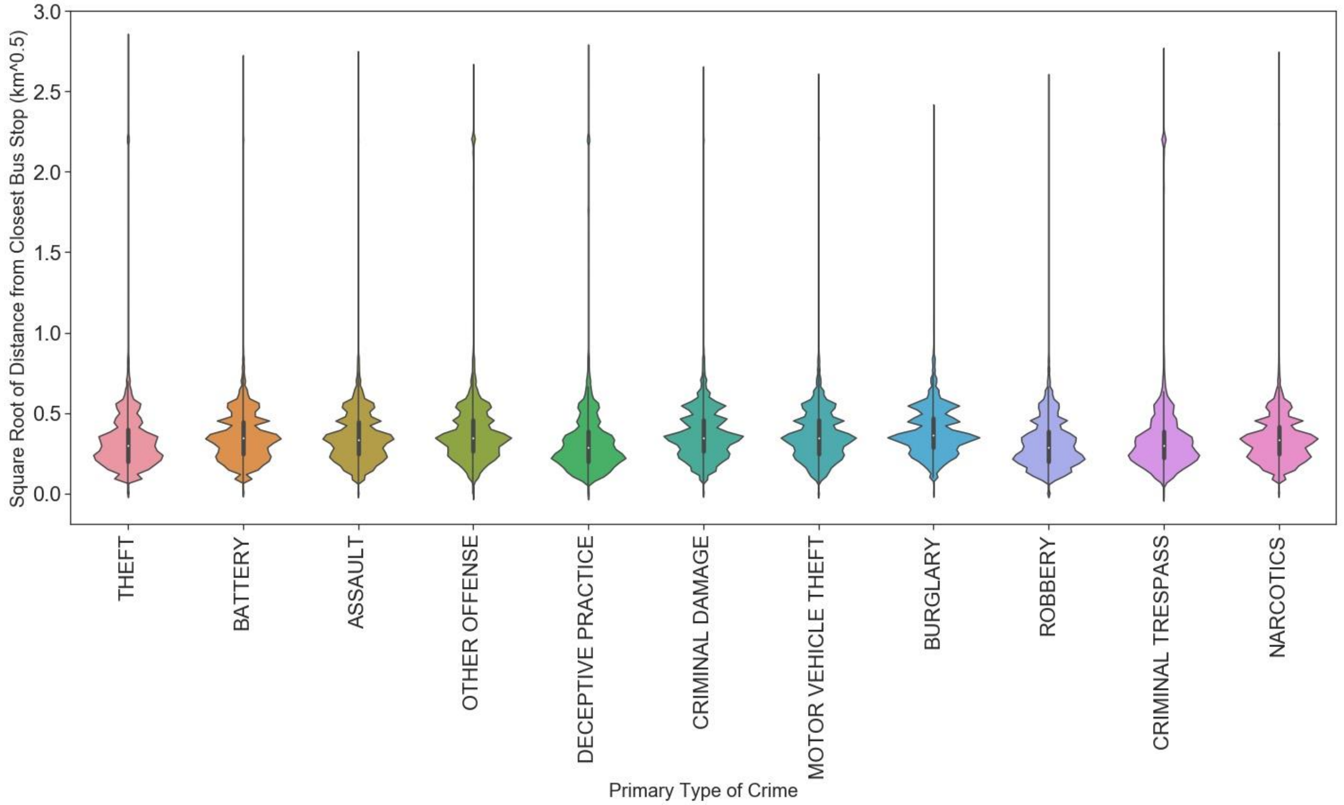

Figure 31: Distribution of square root of distance from closest bus stop for each primary type of crime.

Figure 32 shows that all of the distributions for the distance from the closest liquor store have fat tails and that on average, crimes occurs less than 0.5km from the closest liquor store. The bulk of crimes occur within 1km of a liquor store for each crime type. In order to reduce the tails, the square root of the distance from the closest liquor store was taken and plotted in

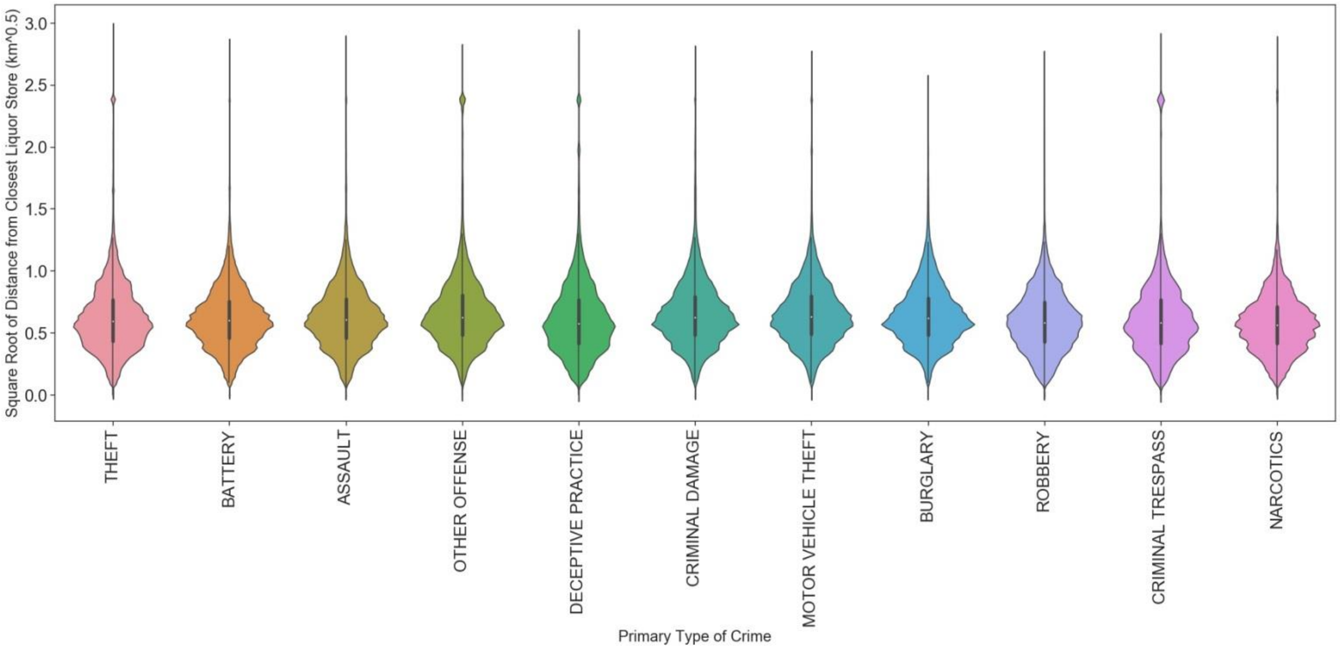

Figure 33 . Taking the square root does help uncover slight variations in the distributions between the crime types. For example, the concentration of crimes is slightly higher for deceptive practice, robbery, criminal trespassing, and narcotics very close to liquor stores.

It is possible that the closest liquor store could be somewhat useful to predict the type of crime. However, as there are 565 unique liquor stores, this would not be feasible.

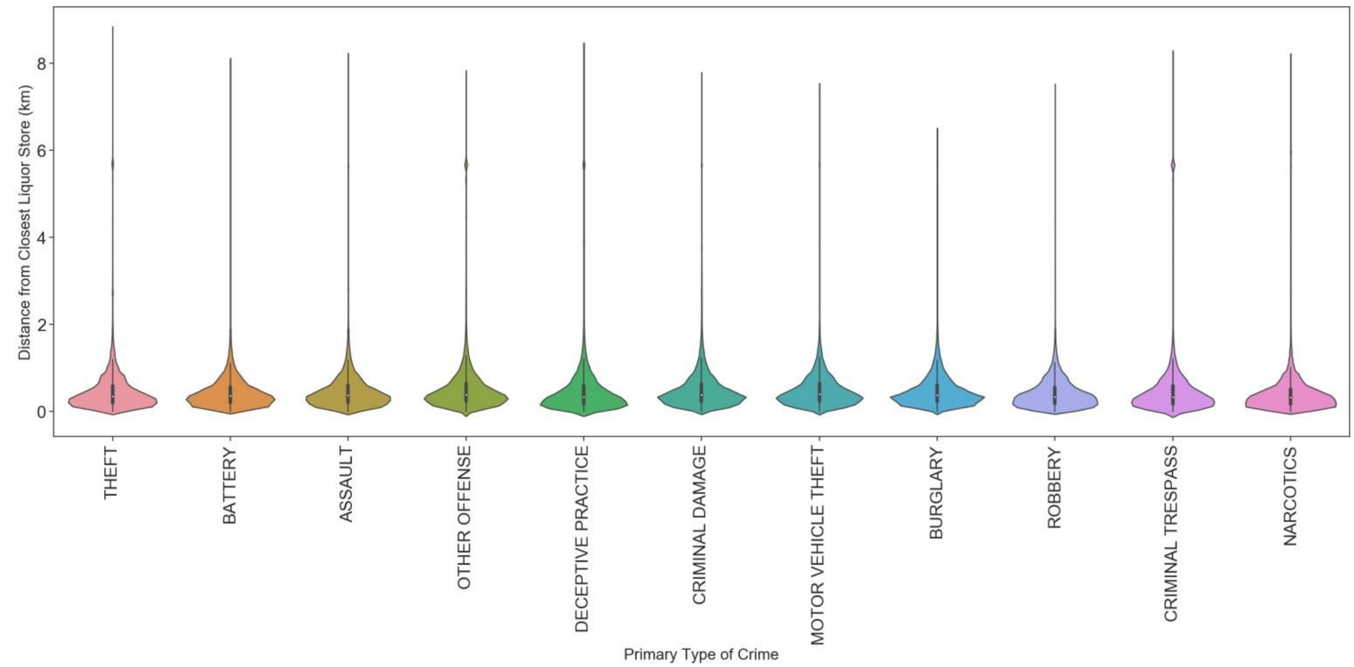

Figure 32: Distribution of crimes based on distance from closest liquor store for each primary type of crime.

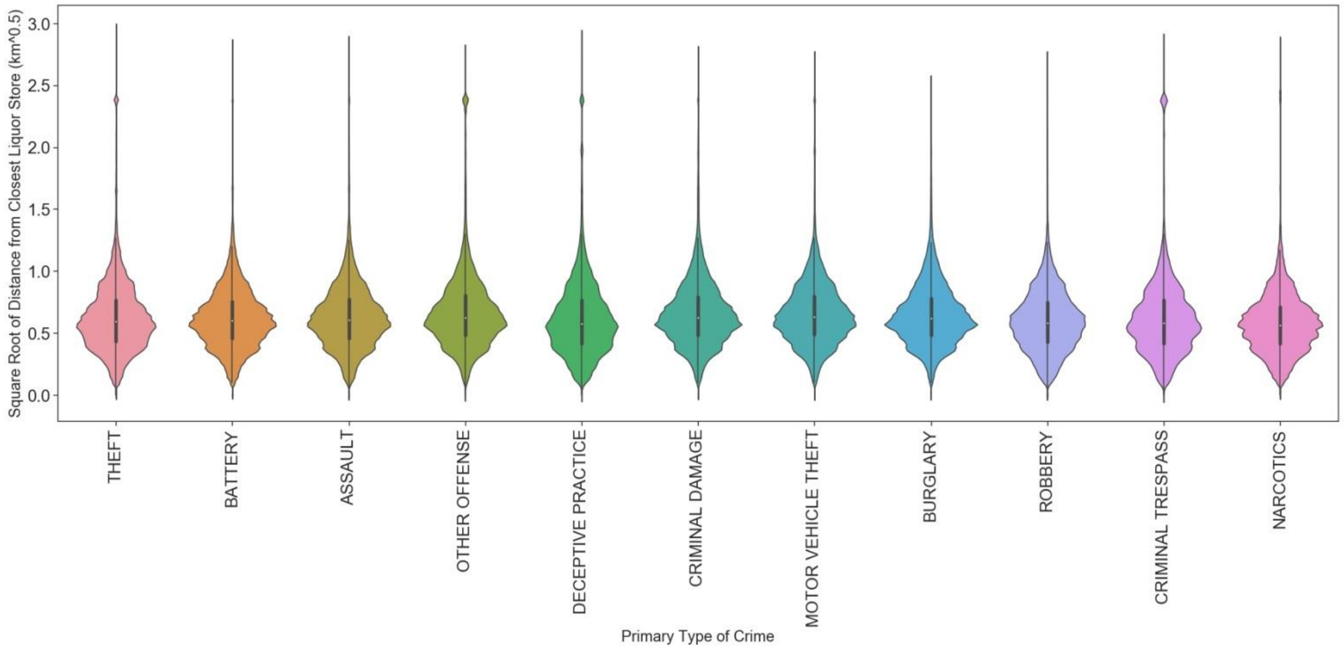

Figure 33: Distribution of crimes based on square root of distance from closest liquor store for each primary type of crime.

Figure 34 shows that during the third and fourth quarters of the year, there is a higher proportion of theft. The proportion of battery reaches a maximum in the second quarter and then gradually decreases through the fourth quarter. The proportion of narcotics is at a maximum during the first quarter, decreases during the second quarter and then remains stable through the end of the year.

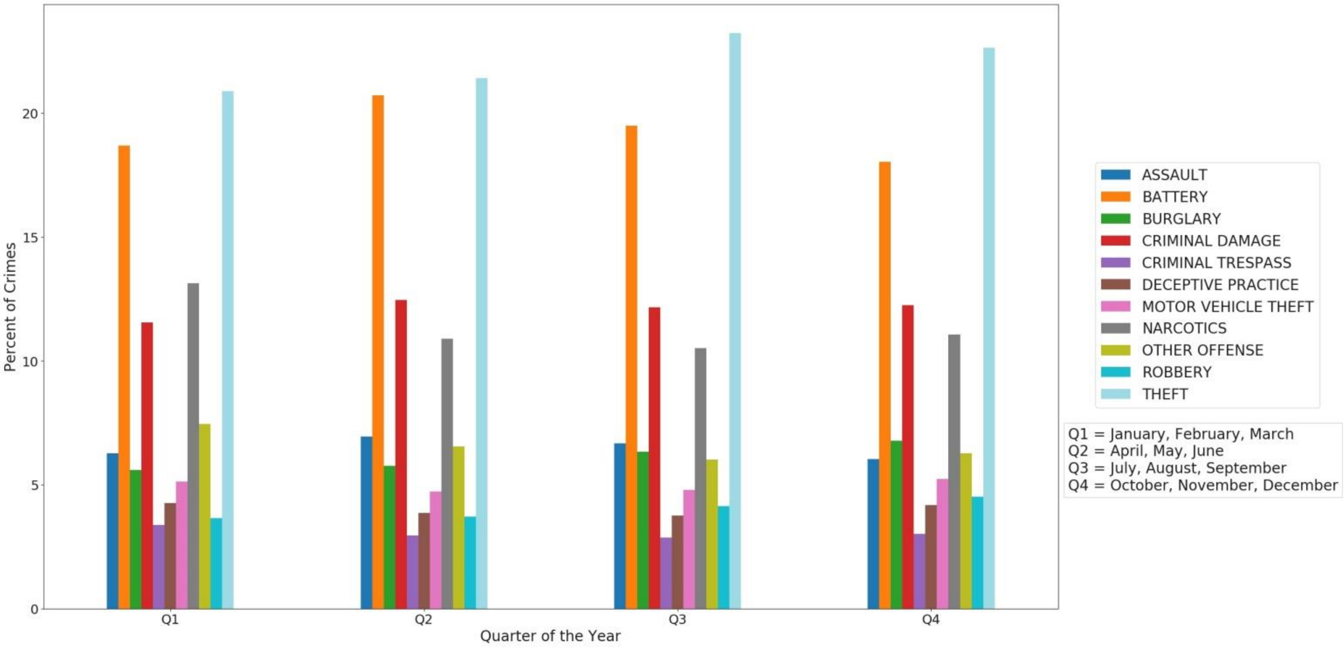

Figure 34: Percentage of each primary type of crime per quarter of the year.

Figure 35 shows a similar pattern as the quarter of the year. So there is no significant difference between the relationship of the quarter of the year and season with the type of crime. Therefore, either the quarter of the year or the season will be used as a feature.

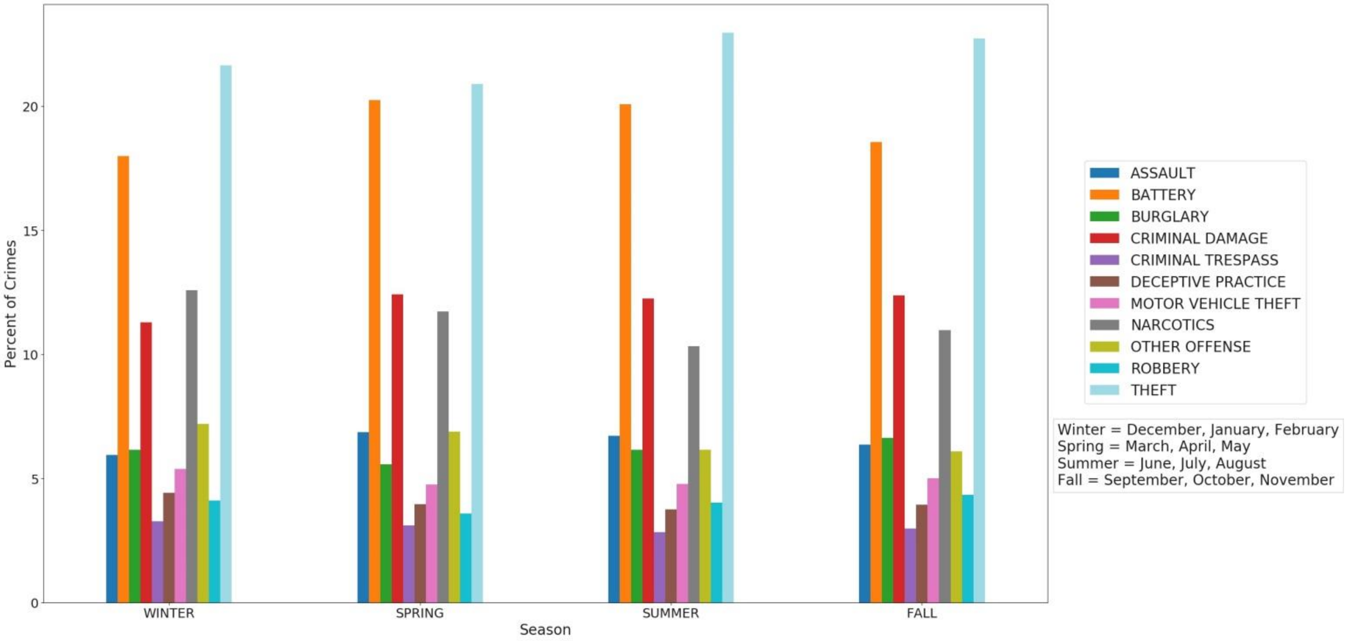

Figure 35: Percentage of each primary type of crime per season.

Figure 36 shows that the proportion of theft decreases slightly from January to March and then increases, reaching a maximum in August before gradually decreasing. The proportion of battery increases from January, reaching a maximum in May/June before decreasing. The proportion of narcotics reaches a maximum in February and then decreases into July. The proportion of criminal damage is at a minimum in December/January/February and reaches a maximum in April and the October/November.

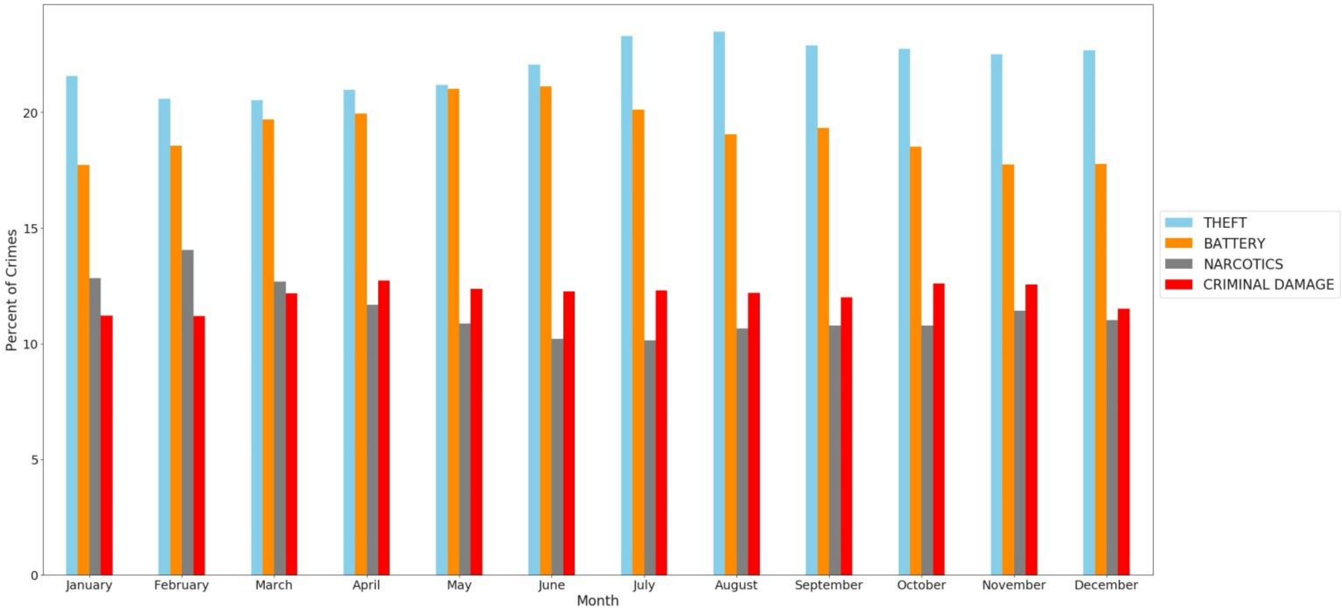

Figure 36: Percentage of 4 primary types of crime per month.

Figure 37 shows that there is a higher proportion of crimes involving battery and criminal damage on the weekend. There is a lower proportion of crimes involving theft, narcotics, burglary, and deceptive practice on the weekend.

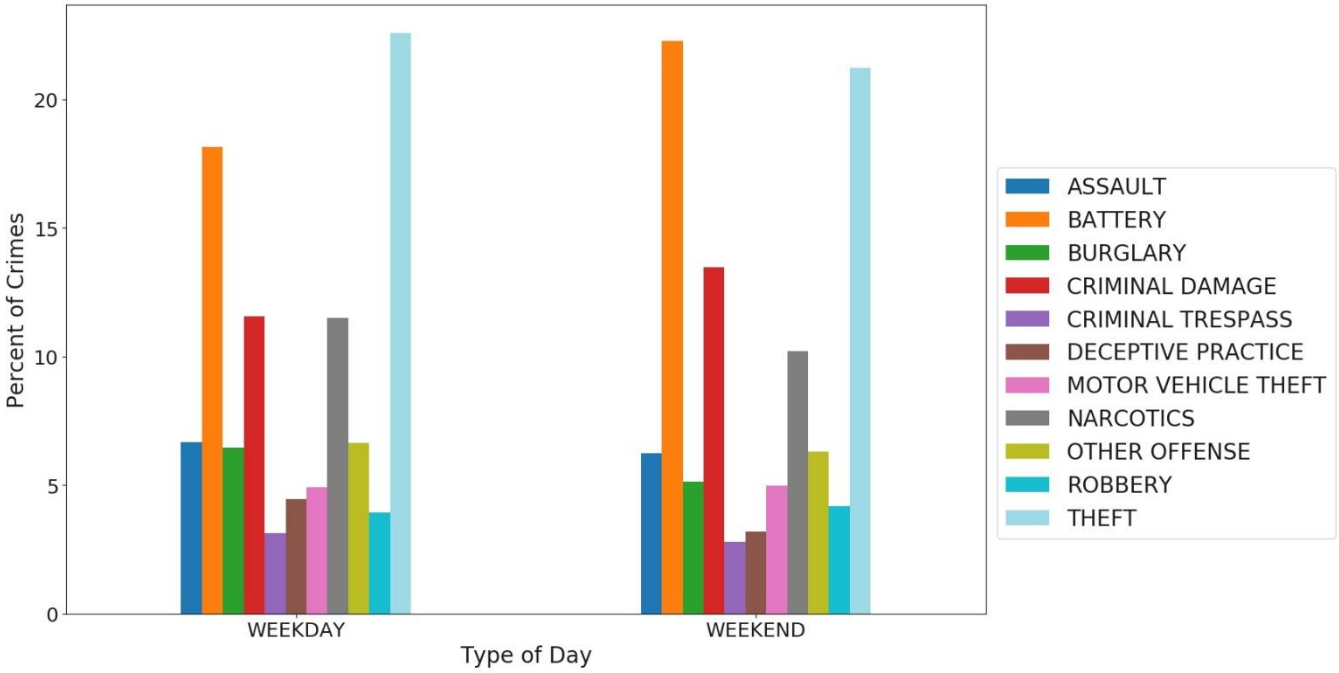

Figure 37: Percentage of each primary type of crime per type of day (weekday/weekend).

Figure 38 shows that there is a higher proportion of crimes involving battery and a lower proportion of crimes involving narcotics on federal holidays

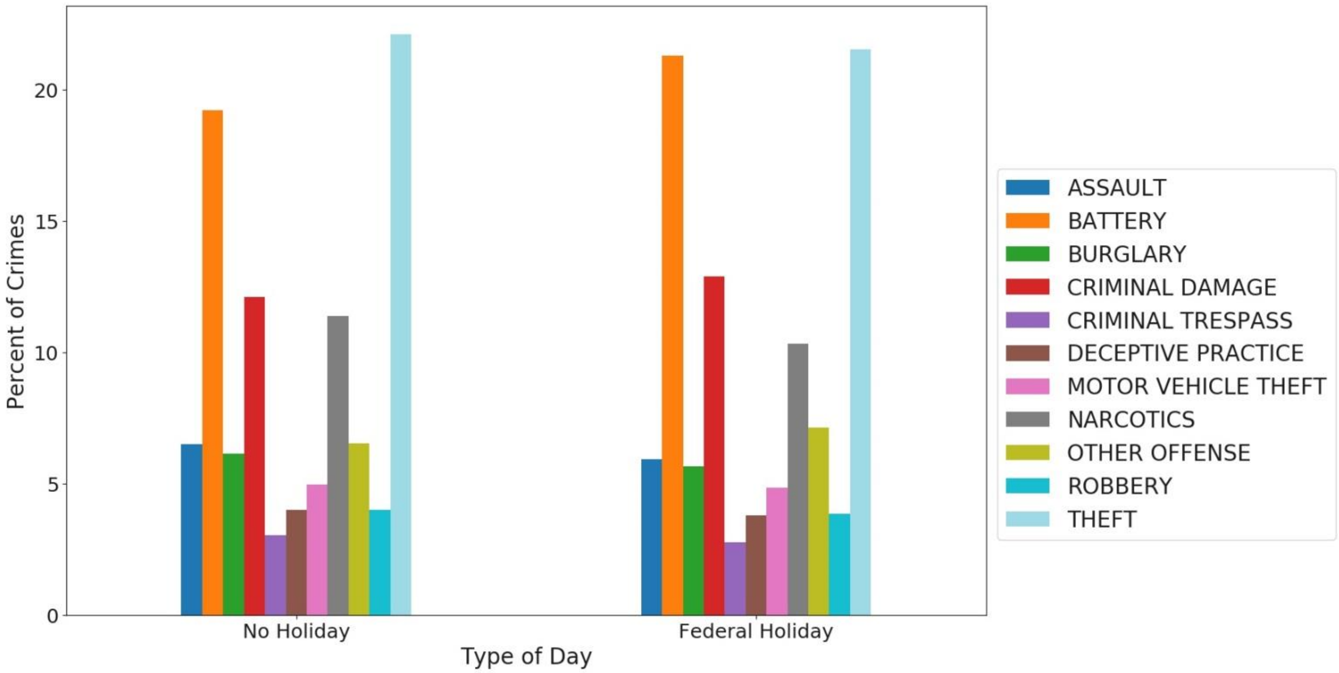

Figure 38: Percentage of each primary type of crime per type of day (no holiday/federal holiday).

Looking at the proportions of crime for each day of the week in Figure 39, Saturday and Sunday have the highest proportions of crimes involving battery and criminal damage. Sunday has the lowest proportion of crimes involving theft and narcotics.

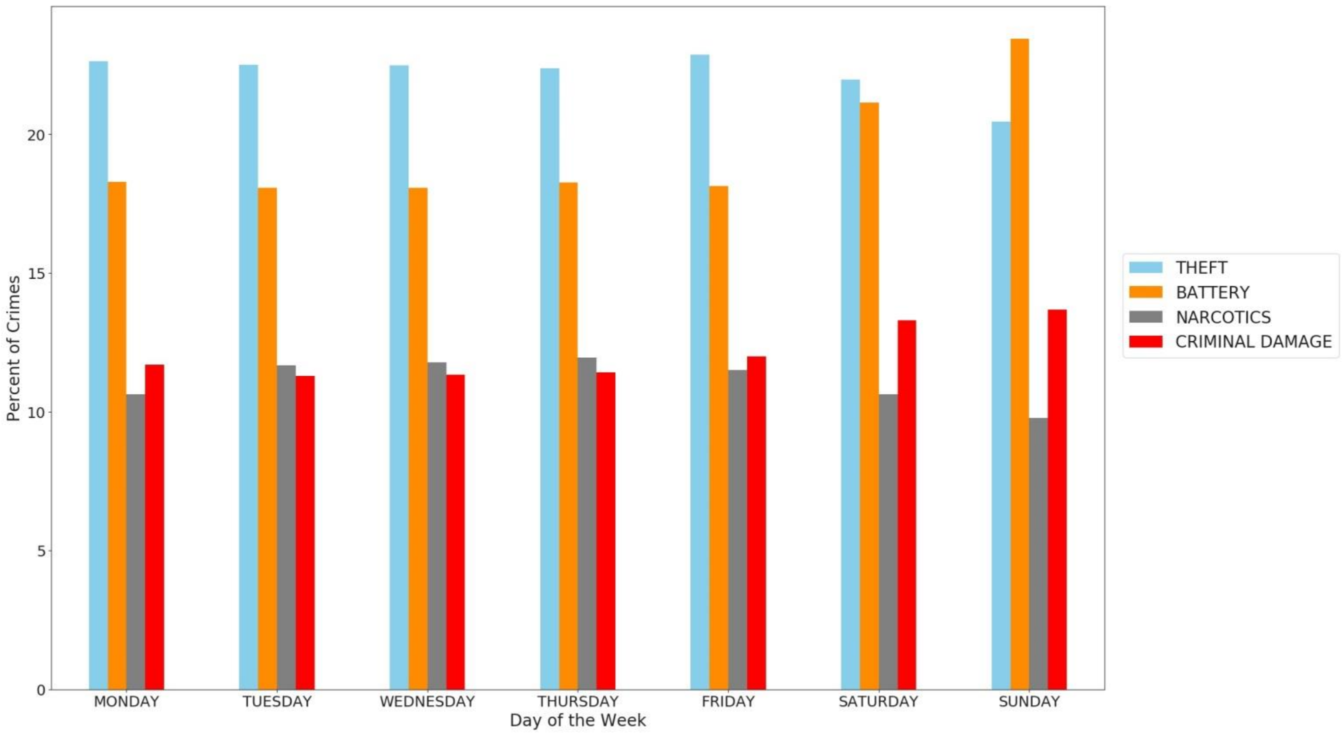

Figure 39: Percentage of 4 primary types of crime per day of the week.

Per Figure 40, Christmas and Independence Day have the highest proportions of crimes involving battery while Martin Luther King Jr. Day, Veterens Day, and Washington's Birthday have the lowest proportions. New Year's Day has the highest proportion of crimes involving theft while Christmas has the lowest proportion. Washington's Birthday has the highest proportion of crimes involving narcotics while New Year's Day and Thanksgiving have the lowest proportions.

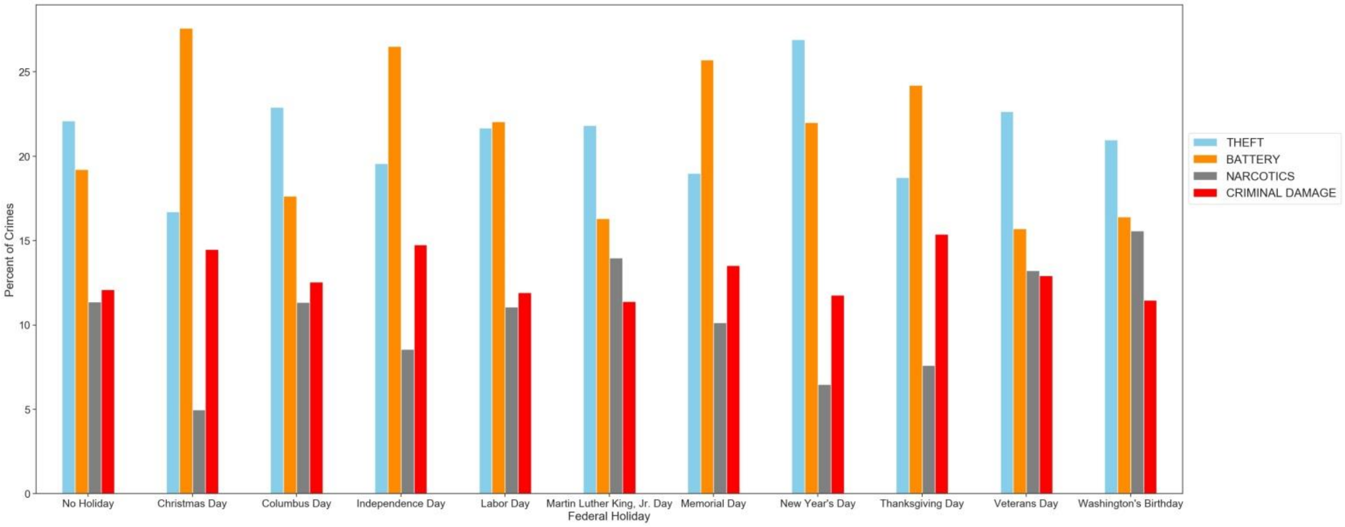

Figure 40: Percentage of 4 primary types of crime per federal holiday.

Figure 41 shows that there is no significant difference in the proportion of each crime for each third of the month. Additionally, looking at the proportion of crime for each day of the month individually (Figure 42), there is not much variation except for the first day of the month where there is a higher proportion of crimes involving theft and lower proportions of crimes involving battery and narcotics.

Both the third of the month and day of the month would therefore not make good predictors for the type of crime.

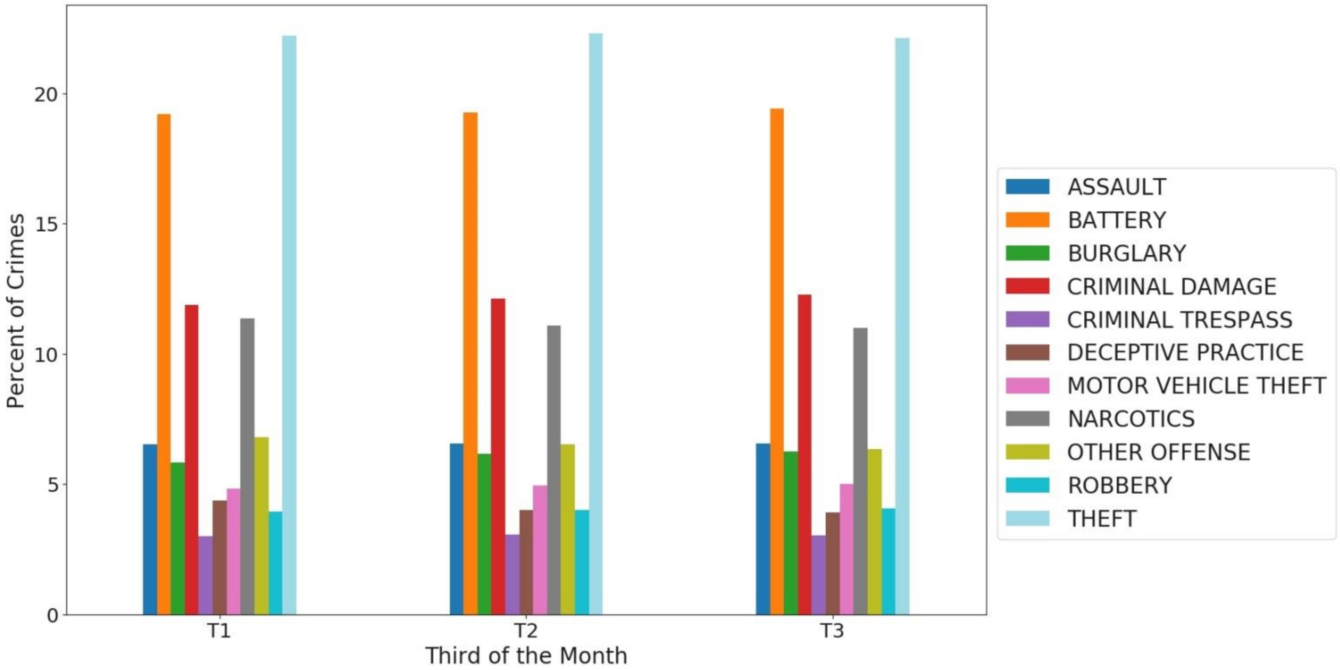

Figure 41: Percentage of each primary type of crime per third of the month.

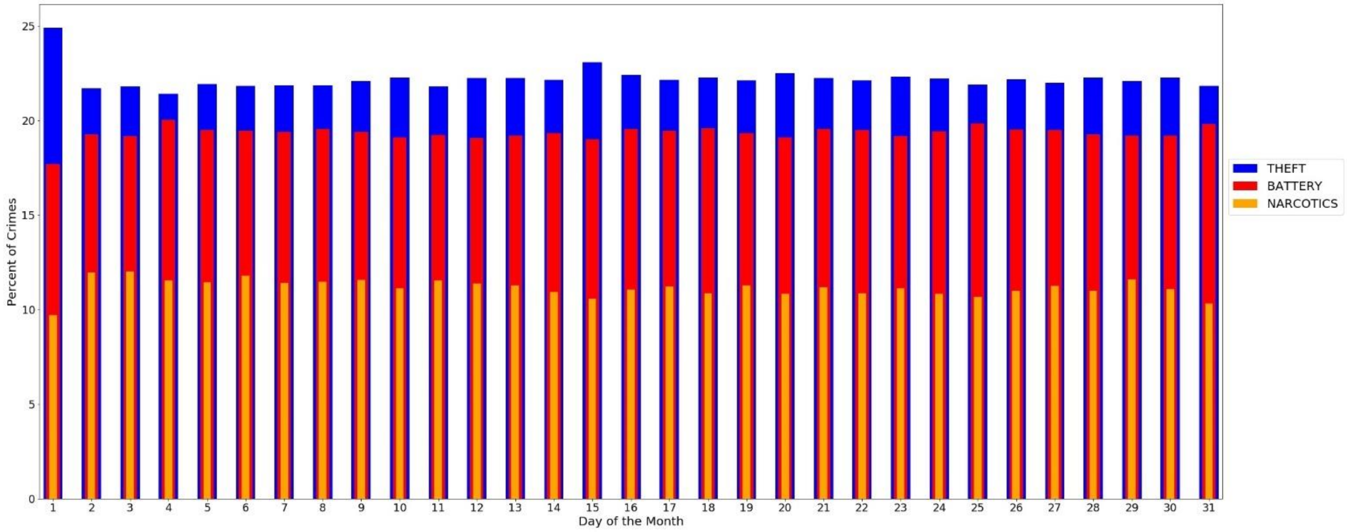

Figure 42: Percentage of 3 primary types of crime per day of the month.

Figure 43 shows the proportion of crimes involving theft is high in the morning and reaches a maximum during the afternoon and then drops off during the evening and overnight. The proportion of crimes involving battery is the least during the morning and then increases throughout the day, reaching a maximum overnight. The proportion of crimes involving narcotics is at a minimum overnight and then increases, reaching a maximum during the evening. The proportion of crimes involving criminal damage is at a minimum during the afternoon and reaches a maximum overnight.

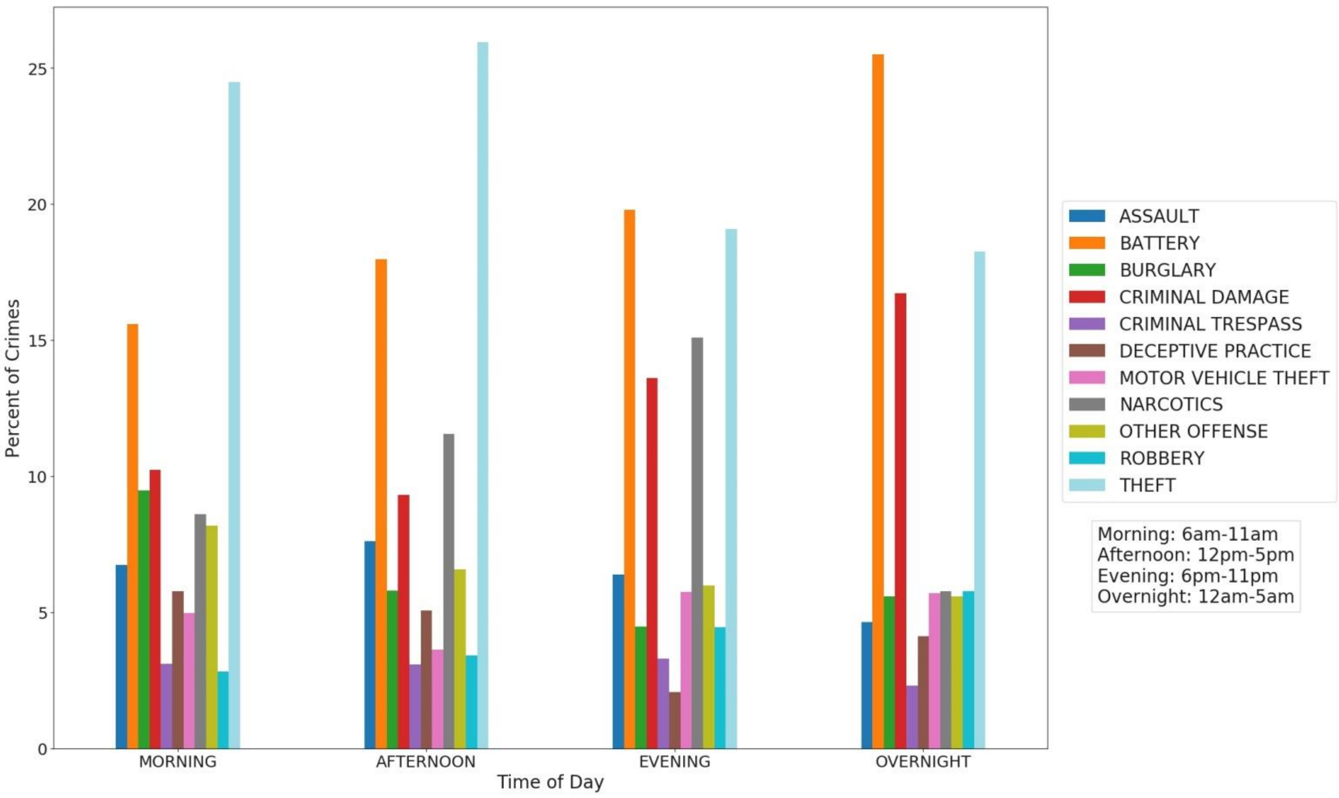

Figure 43: Percentage of each primary type of crime per time of day.

Per Figure 44, the proportion of crimes involving battery is at a maximum at 2/3am. The proportion of crimes involving theft is somewhat steady from 7am-6pm and then drop off, reaching a minimum at 1-4am. The proportion of crimes involving narcotics is the highest at 7-9pm but also reaches a maximum at 11am.

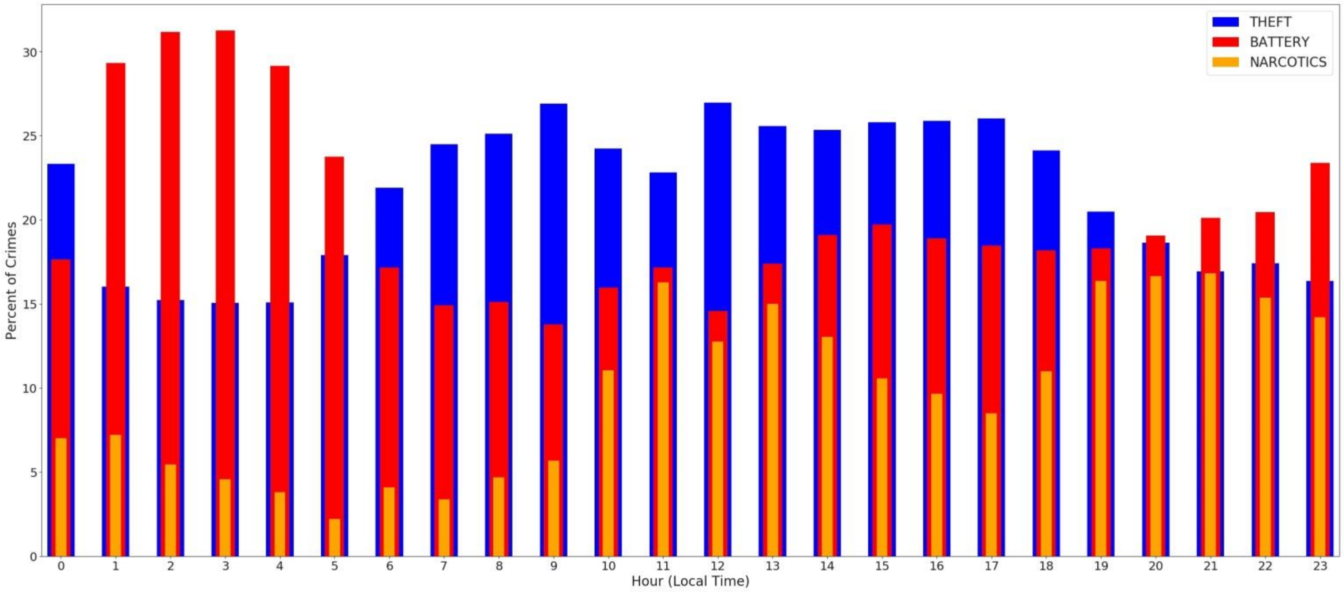

Figure 44: Percentage of 3 primary types of crime per hour.

In Figure 45, it is seen that there are no highly correlated numerical features in my dataset. There is a moderate, positive relationship between the distance from the closest liquor store and the distance from the closest bus stop. Figure 46a shows that for most of the reports, there isn't much of a relationship between the two features. But for distances greater than 3km, there is a strong positive linear relationship. Looking at reports when the distance from the closest bus stop is greater than 3km, I see that many of them occur in community 76, which is the community farthest to the northwest. Therefore crimes in this community would likely be farther removed from both bus stops and liquor stores, creating the positive correlation.

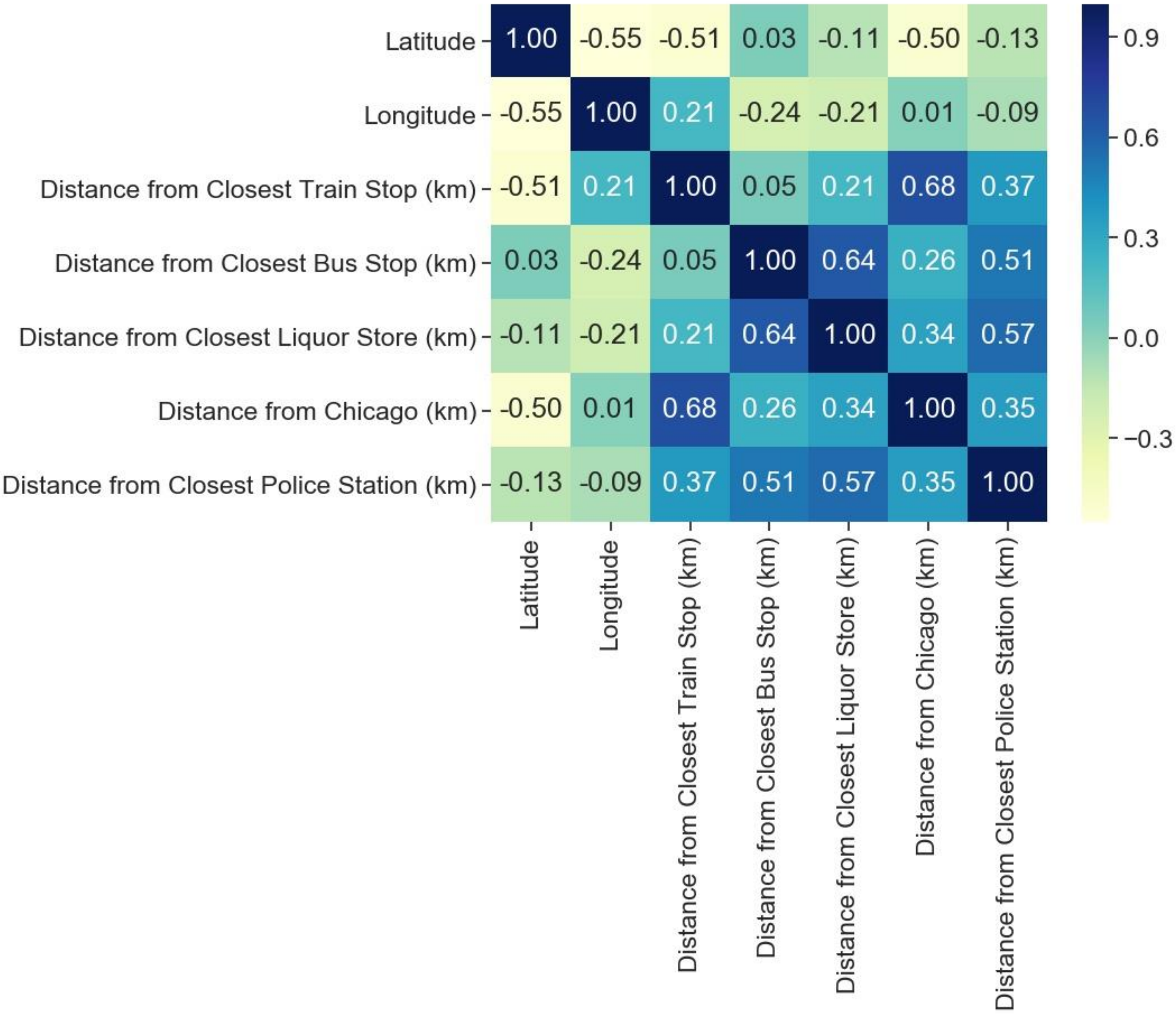

Figure 45: Pearson correlation coefficients between numerical features.

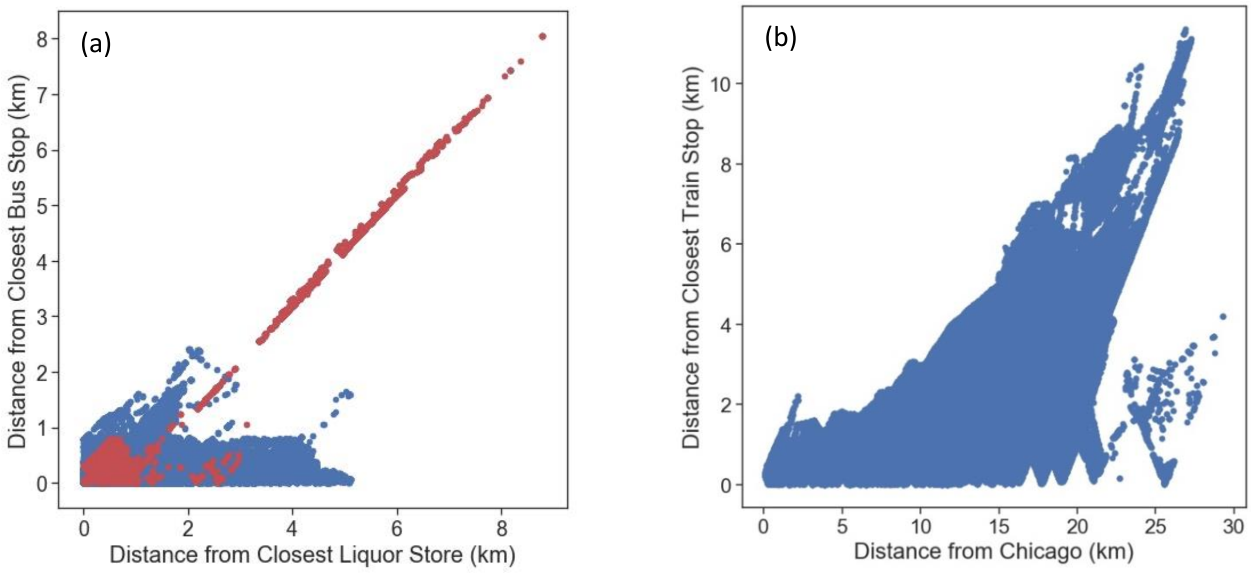

Figure 46: (a) Distance from closest bus stop vs. distance from closest liquor store (red points are crimes within community 76) and (b) distance from closest train stop vs. distance from Chicago.

There is also a moderate, positive relationship between the distance from Chicago (city center) and the distance from the closest train stop because generally, there are more train stops closer to the city center. Looking at Figure 46b, this relationship is not especially strong as the distance from the closest train stop varies significantly for larger distances from Chicago.

Based on the above data visualization/analysis, I decided to include the following features in my model evaluation as they appeared to have some relationship with the primary type of crime:

- Ward
- Police district
- Police beat
- Community
- Location description
- Latitude
- Longitude
- Distance from city center of Chicago
- Square root of distance from closest police station
- Distance from closest train stop
- Closest train line
- Square root of distance from closest bus stop
- Square root of distance from closest liquor store
- Season
- Month
- If it is a weekday or the weekend
- If it is a federal holiday
- What federal holiday it is
- Day of the week
- Time of day
- Hour