A Spatial and Temporal Analysis of Crimes in Chicago

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. Not only does this make the city more unsafe, it also could put more risk on the lives of police officers. In an effort to mitigate crime rates in Chicago and/or promoting officer safety, 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 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 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.

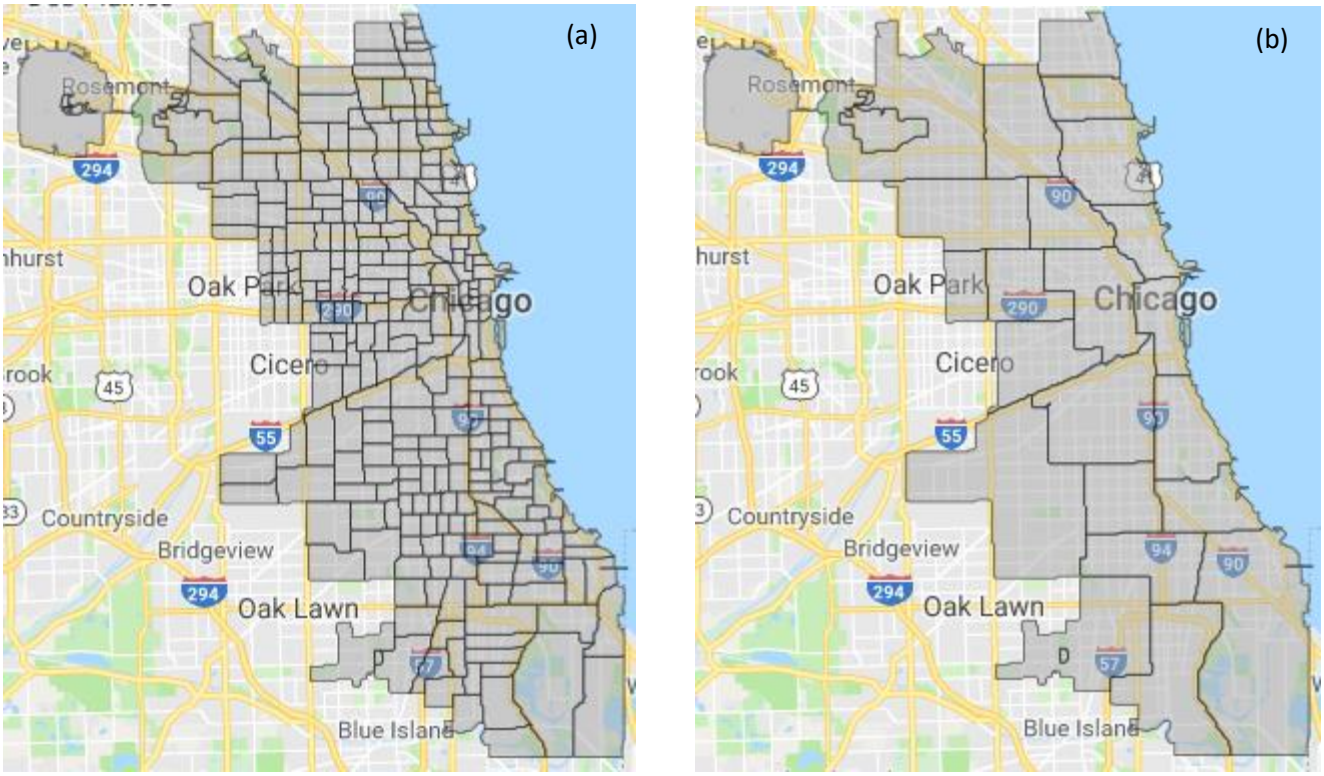

Figure 2: Boundaries for (a) police beats (City of Chicago, 2018b) and (b) police districts (City of Chicago, 2018c).

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 2001.

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.

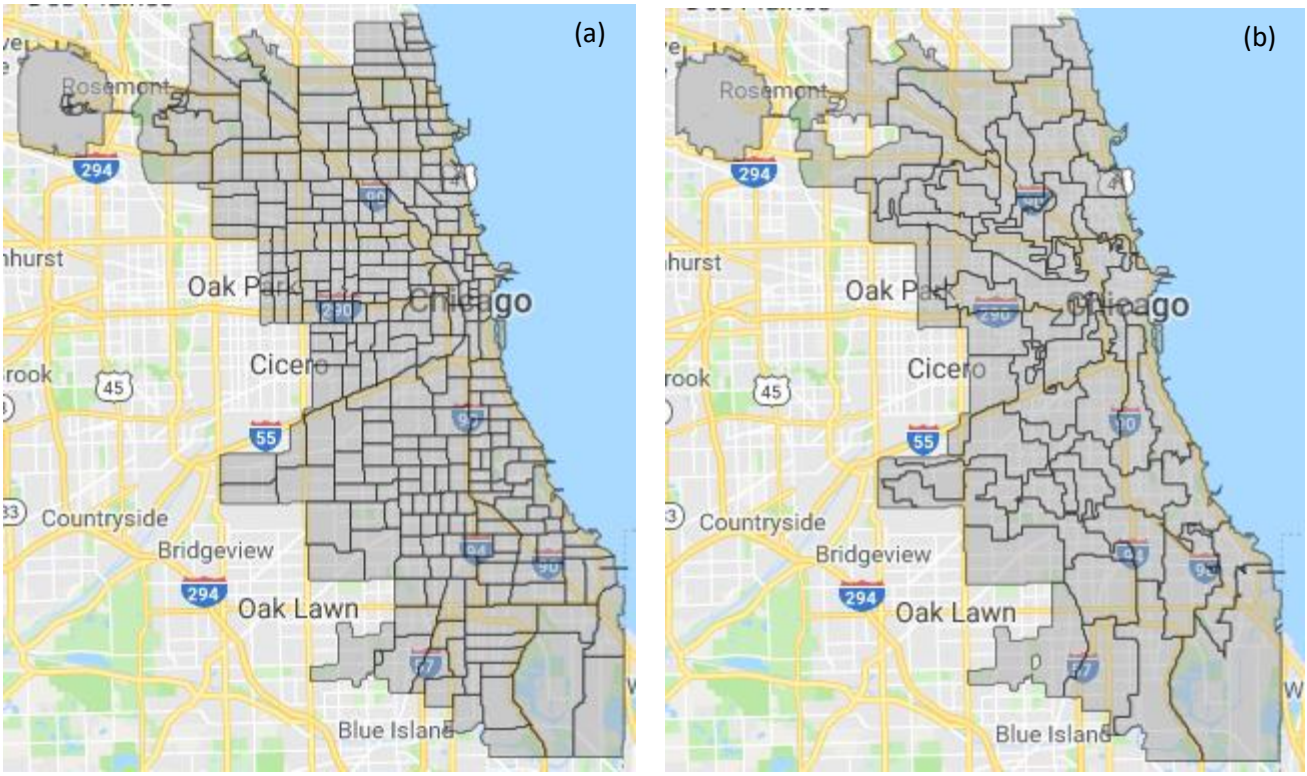

Figure 3: Boundaries for (a) police beats (City of Chicago, 2018b) and (b) wards (City of Chicago, 2016).

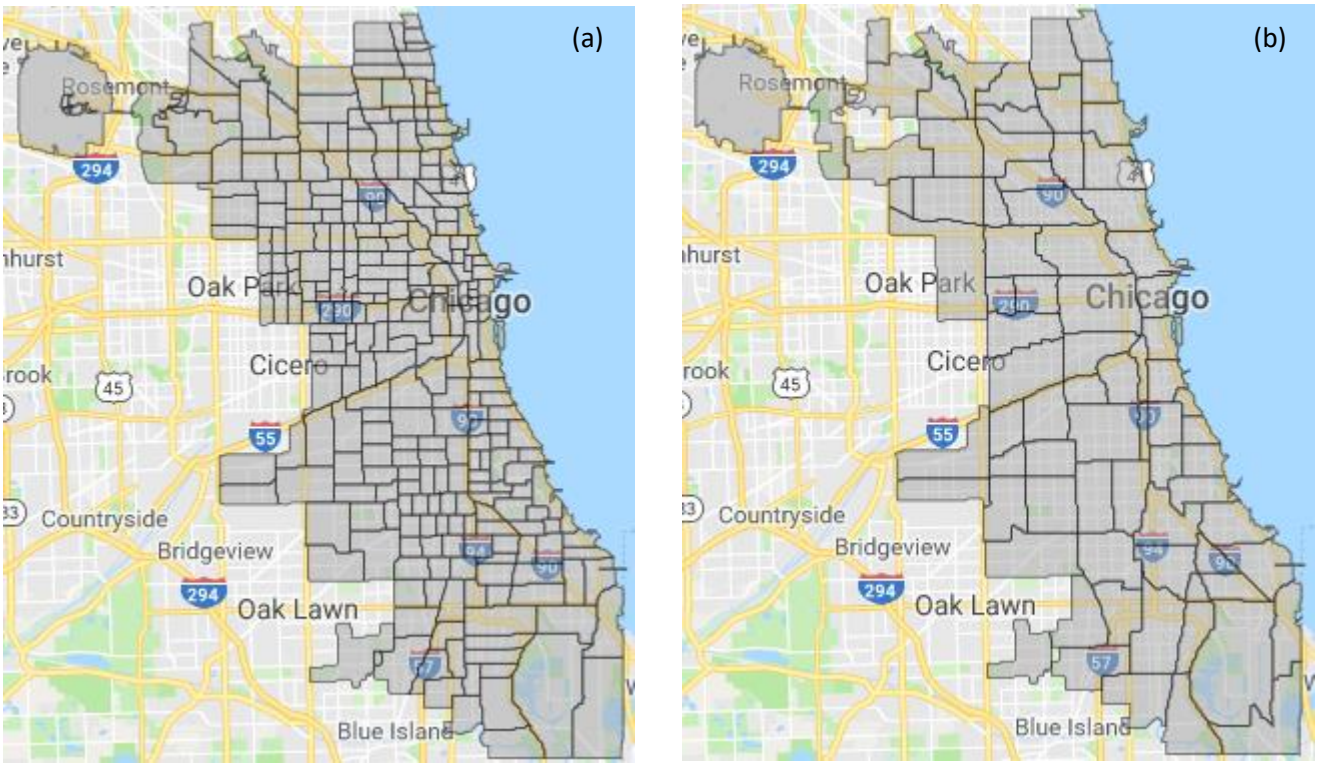

Figure 4: Boundaries for (a) police beats (City of Chicago, 2018b) and (c) communities (City of Chicago, 2018a).

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).

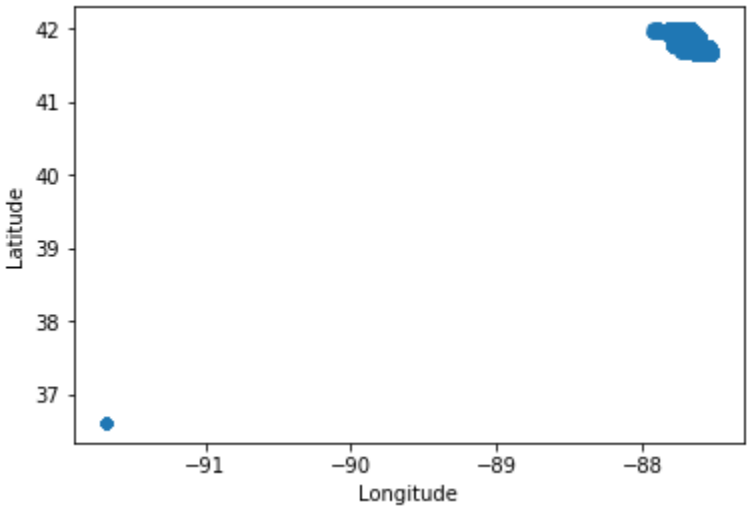

Figure 5: Plot of latitude vs. longitude in degrees.

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.

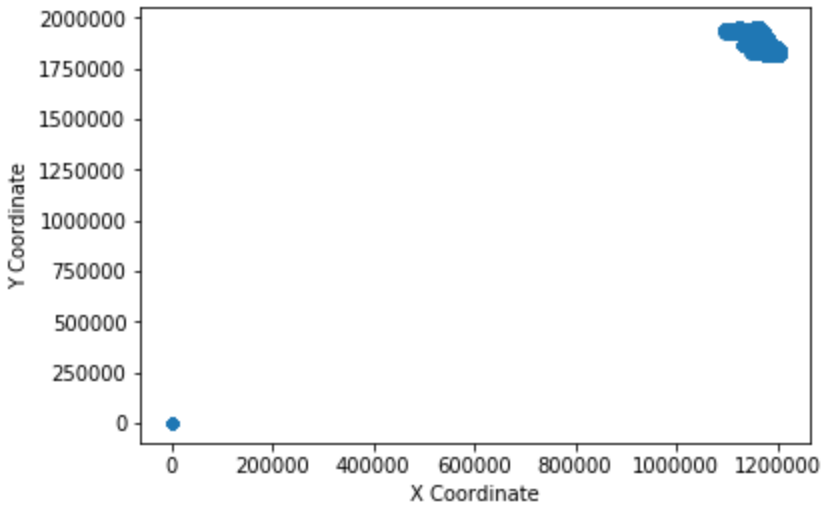

Figure 6: Plot of Y Coordinate vs. X Coordinate.

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

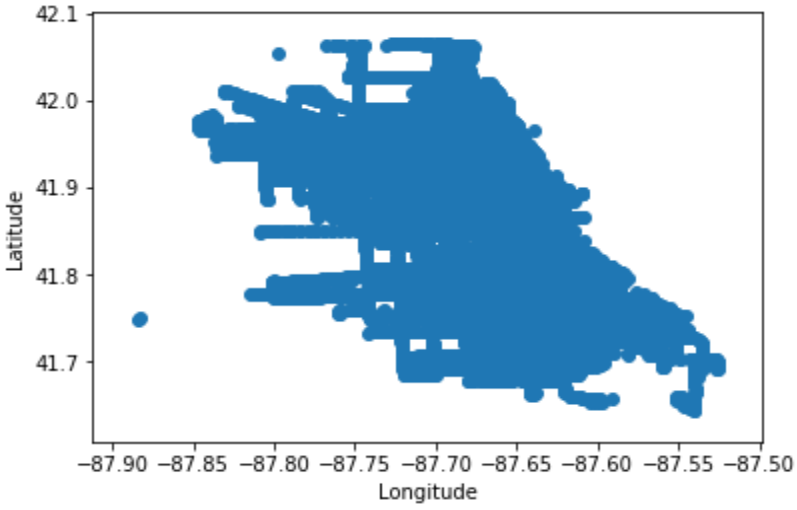

Figure 7: Plot of latitude and longitude in degrees of bus stops.

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.

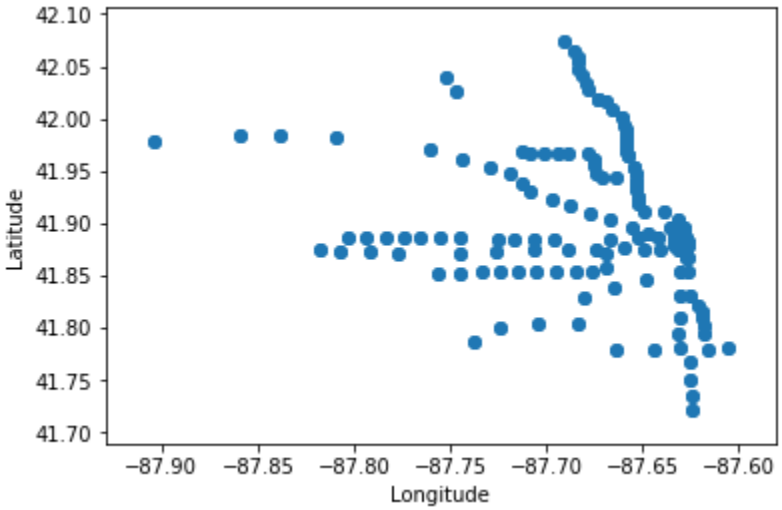

Figure 8: Plot of latitude and longitude in degrees of train stops.

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.

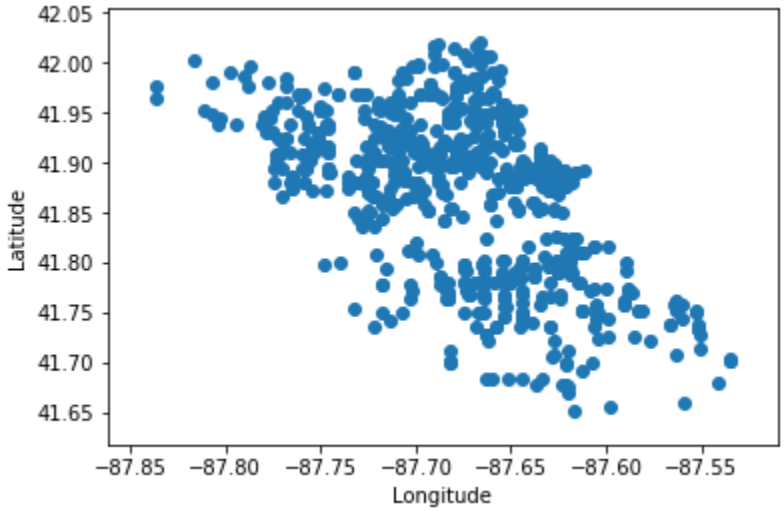

Figure 9: Plot of latitude and longitude in degrees of liquor stores.

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.

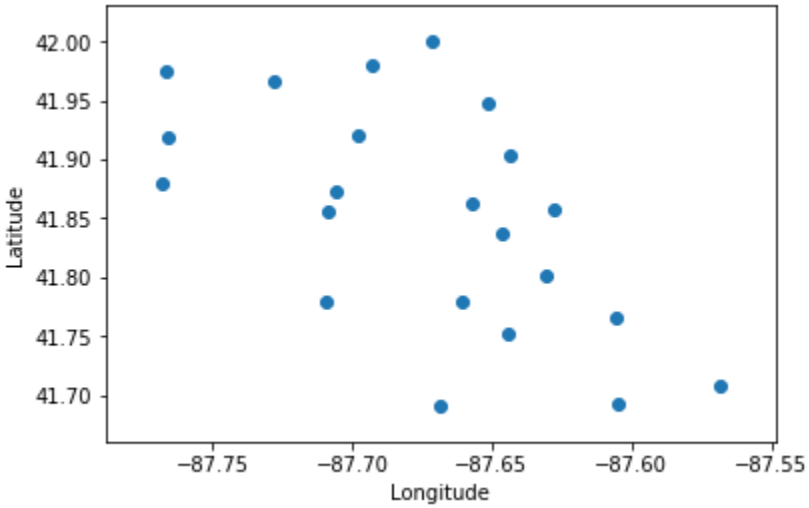

Figure 10: Plot of latitude and longitude in degrees of police stations.

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.

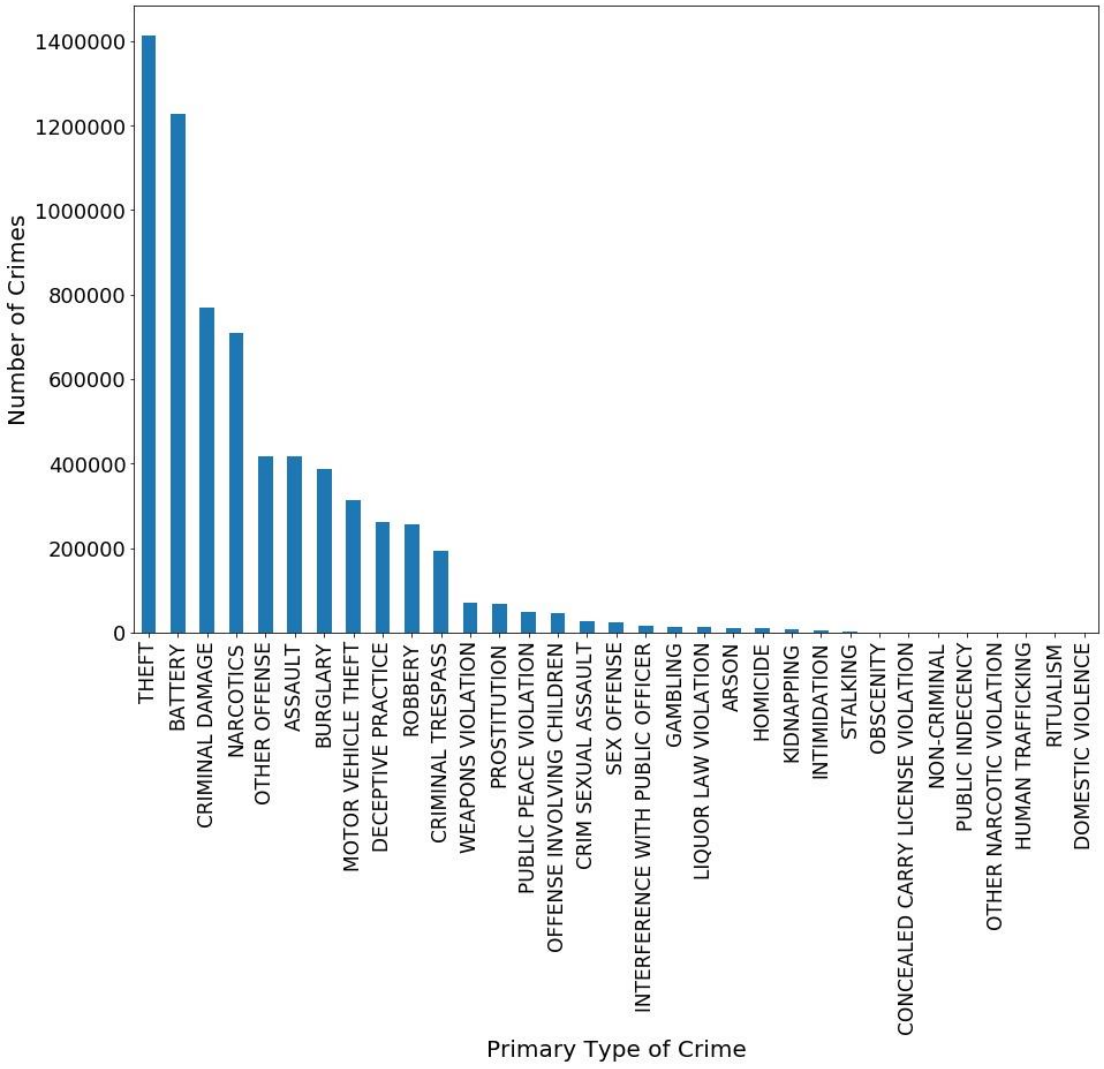

Figure 11: Number of crimes for each primary type of crime.

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.

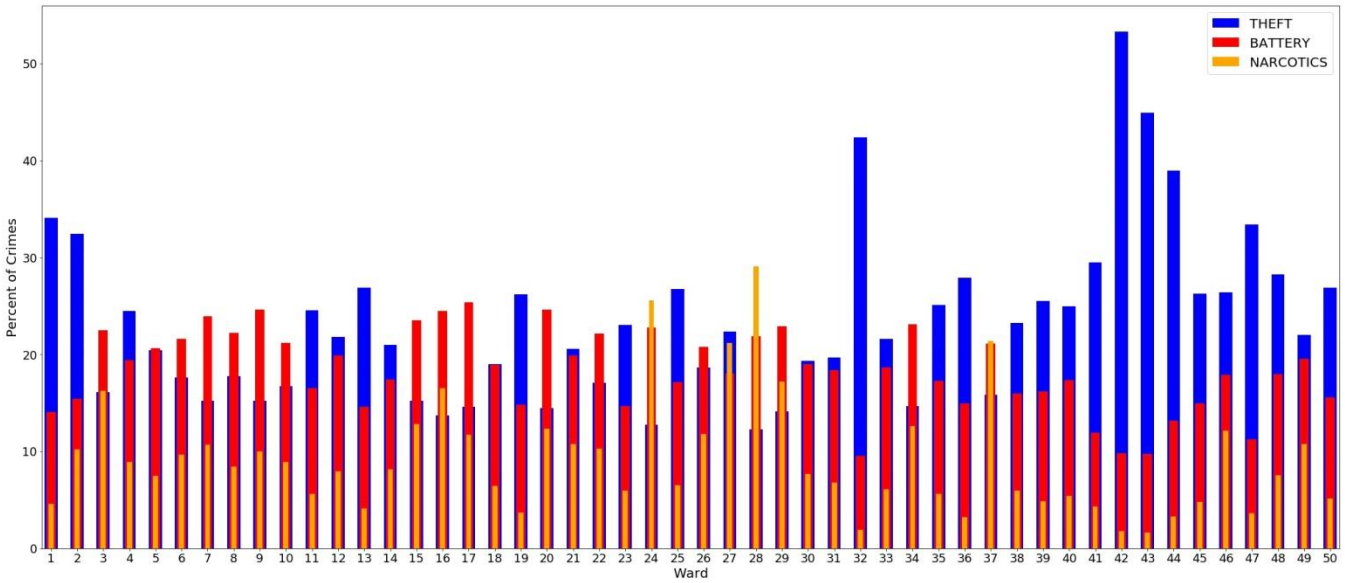

Figure 12: Percentage of 3 primary types of crime per ward.

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.

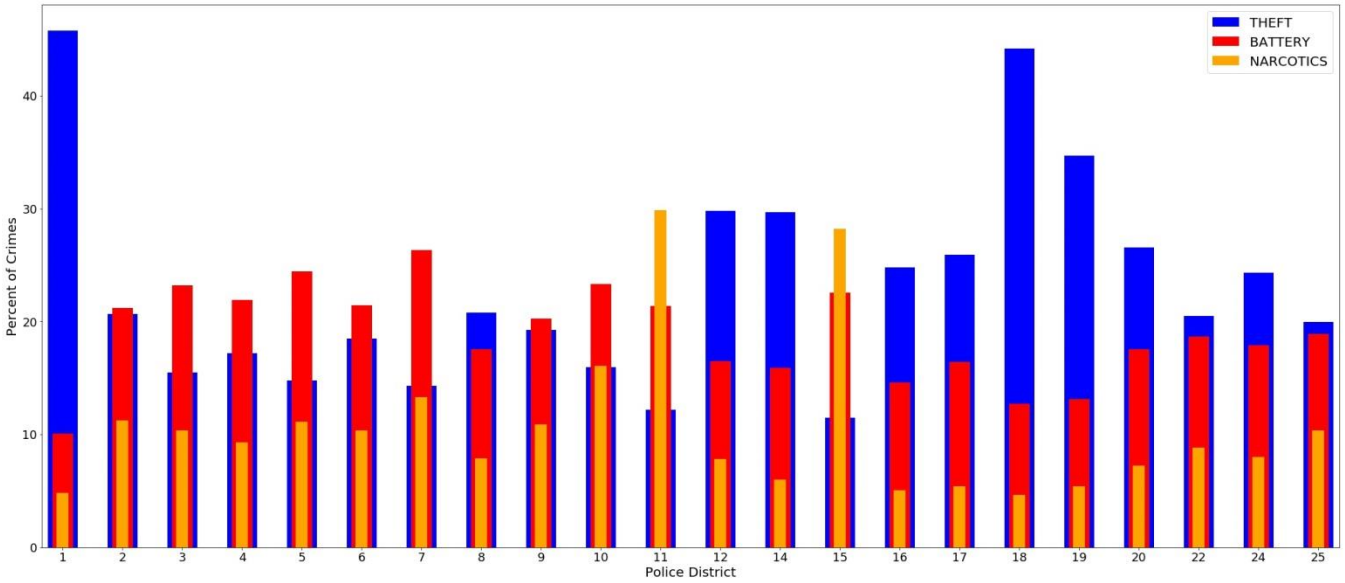

Figure 13: Percentage of 3 primary types of crime per police district.

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.

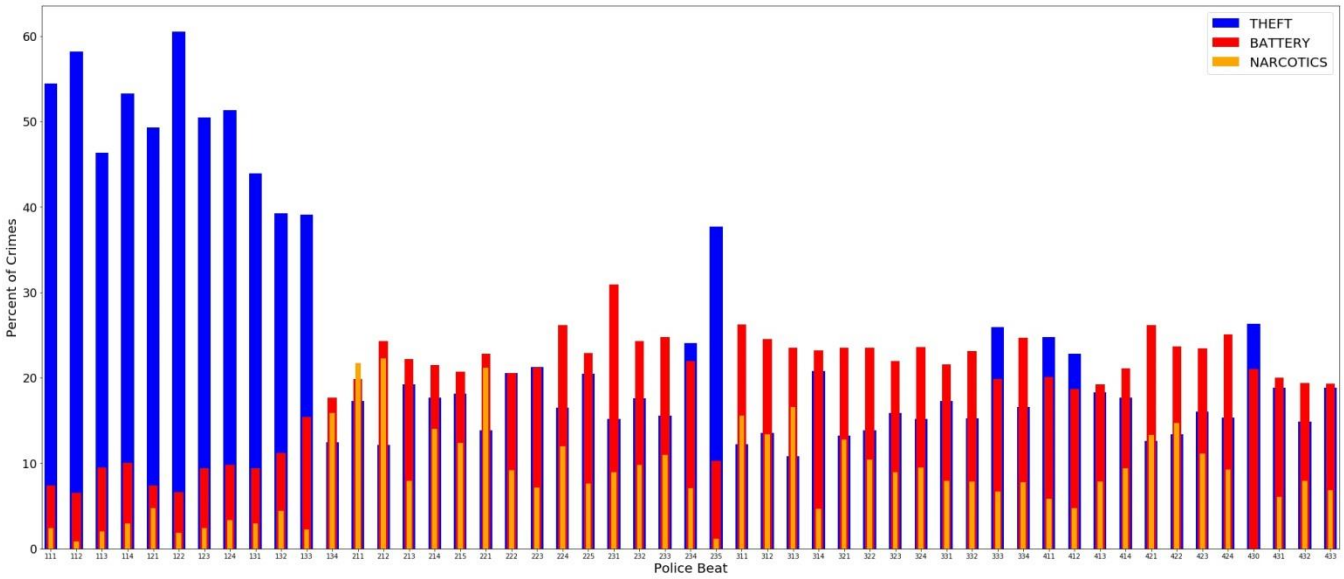

Figure 14: Percentage of 3 primary types of crime for a sample of police beats. There is a total of 304 police beats.

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.

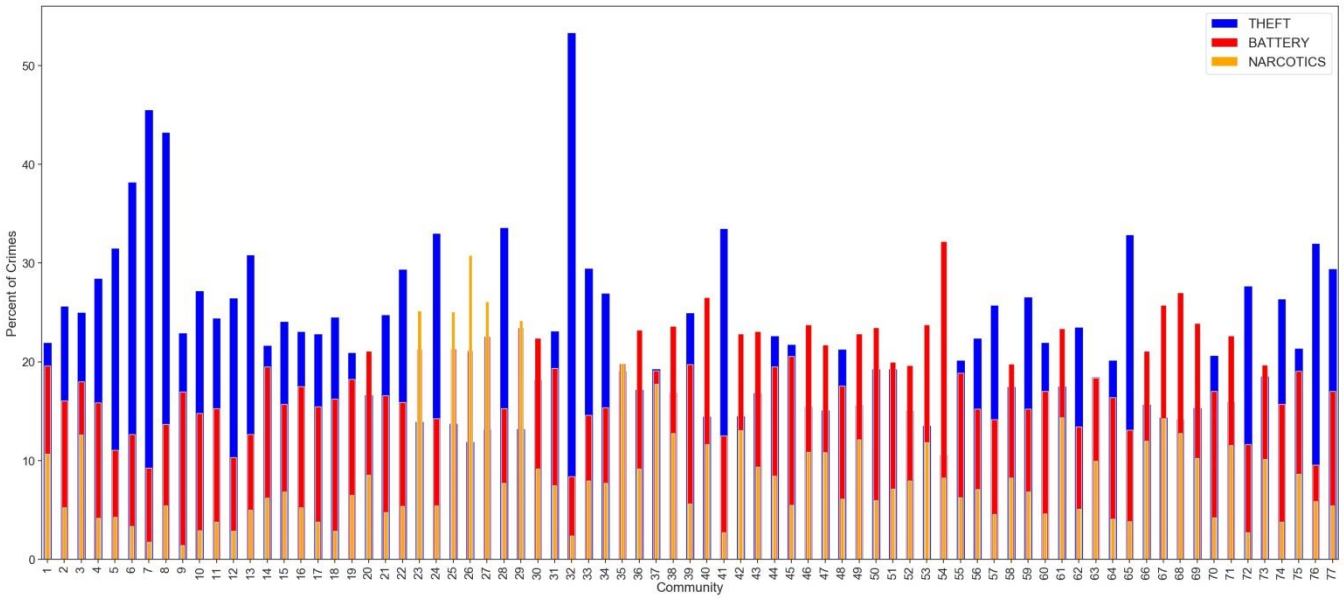

Figure 15: Percentage of 3 primary types of crime within each community.

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.

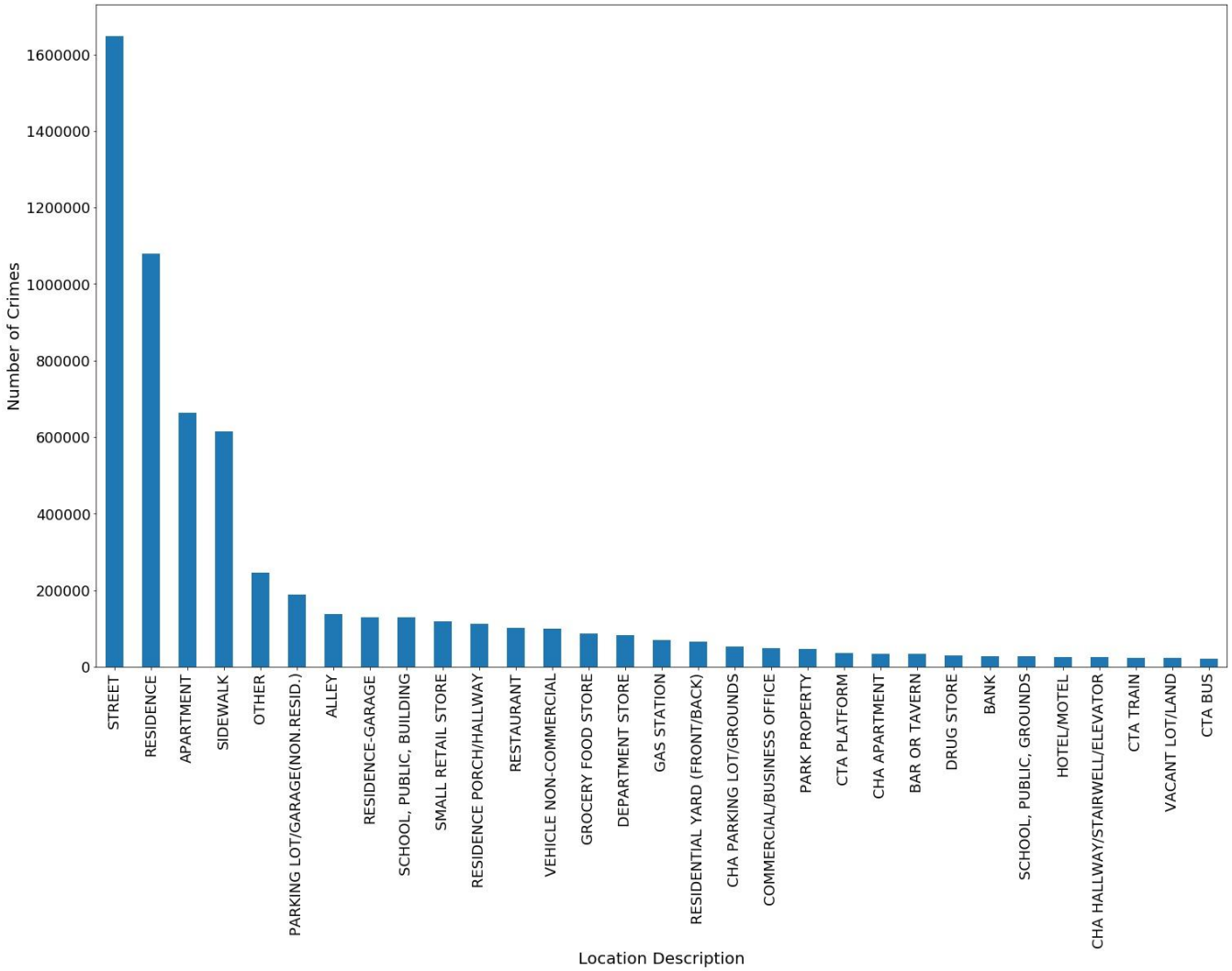

Figure 16: Number of crimes for a sample of location descriptions. There is a total of 109 location descriptions.

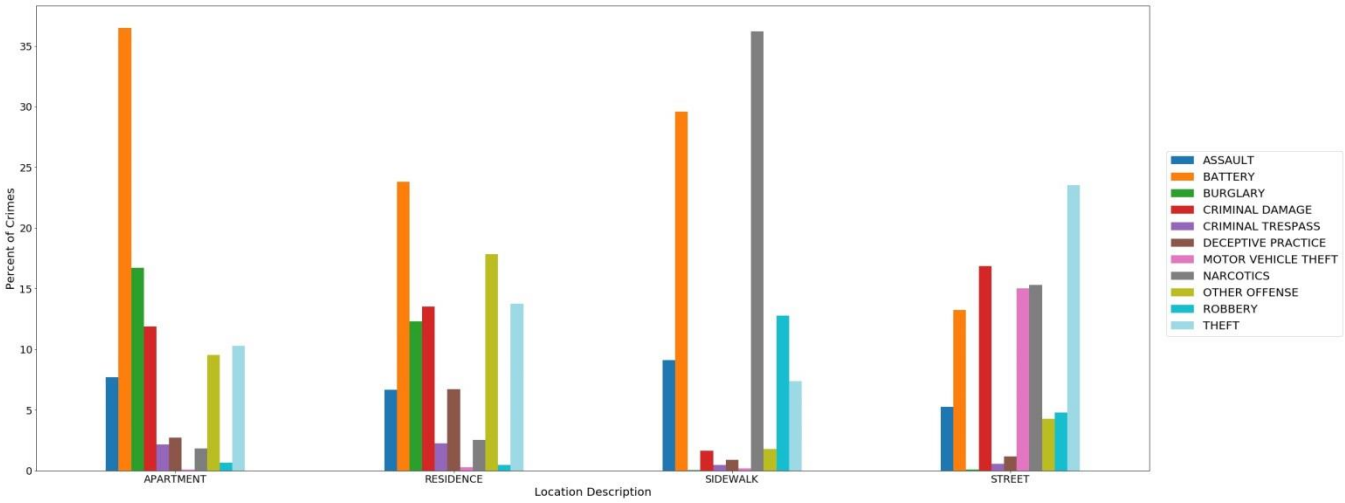

Figure 17: Percentage of each primary type of crime for top 4 location descriptions.

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.

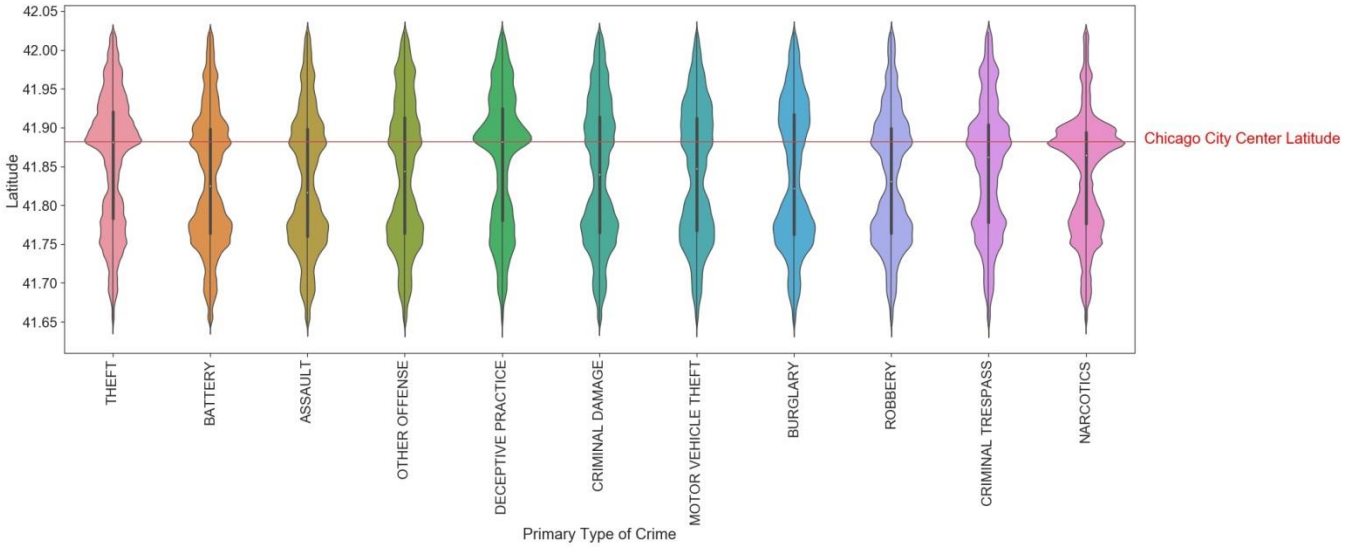

Figure 18: Distribution of crimes across latitude for each primary type of crime.

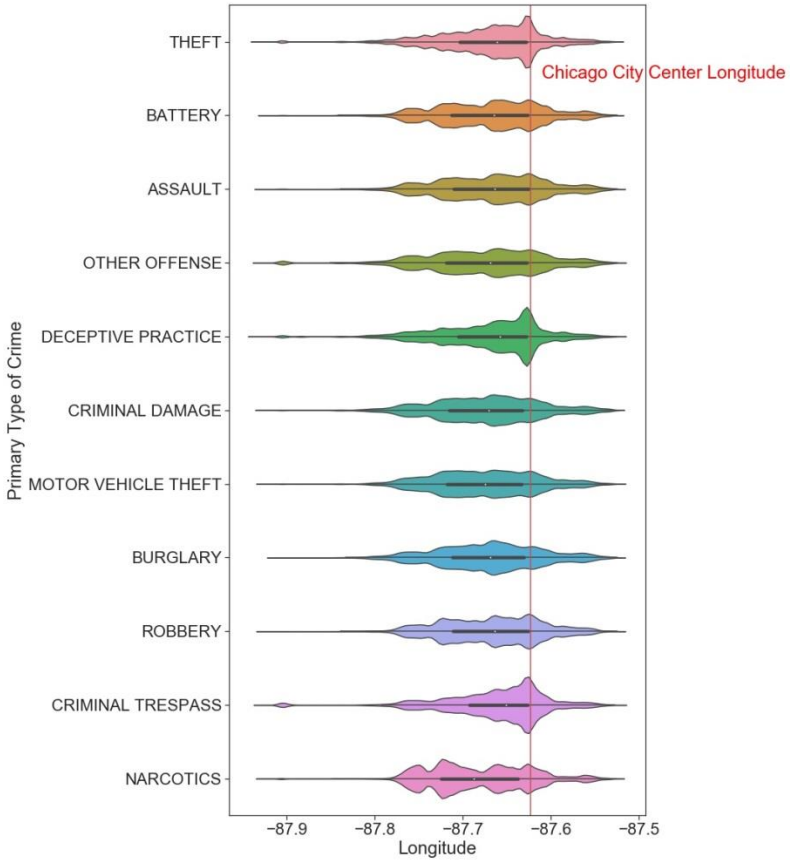

Figure 19: Distribution of crimes across longitude for each primary type of crime.

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°.

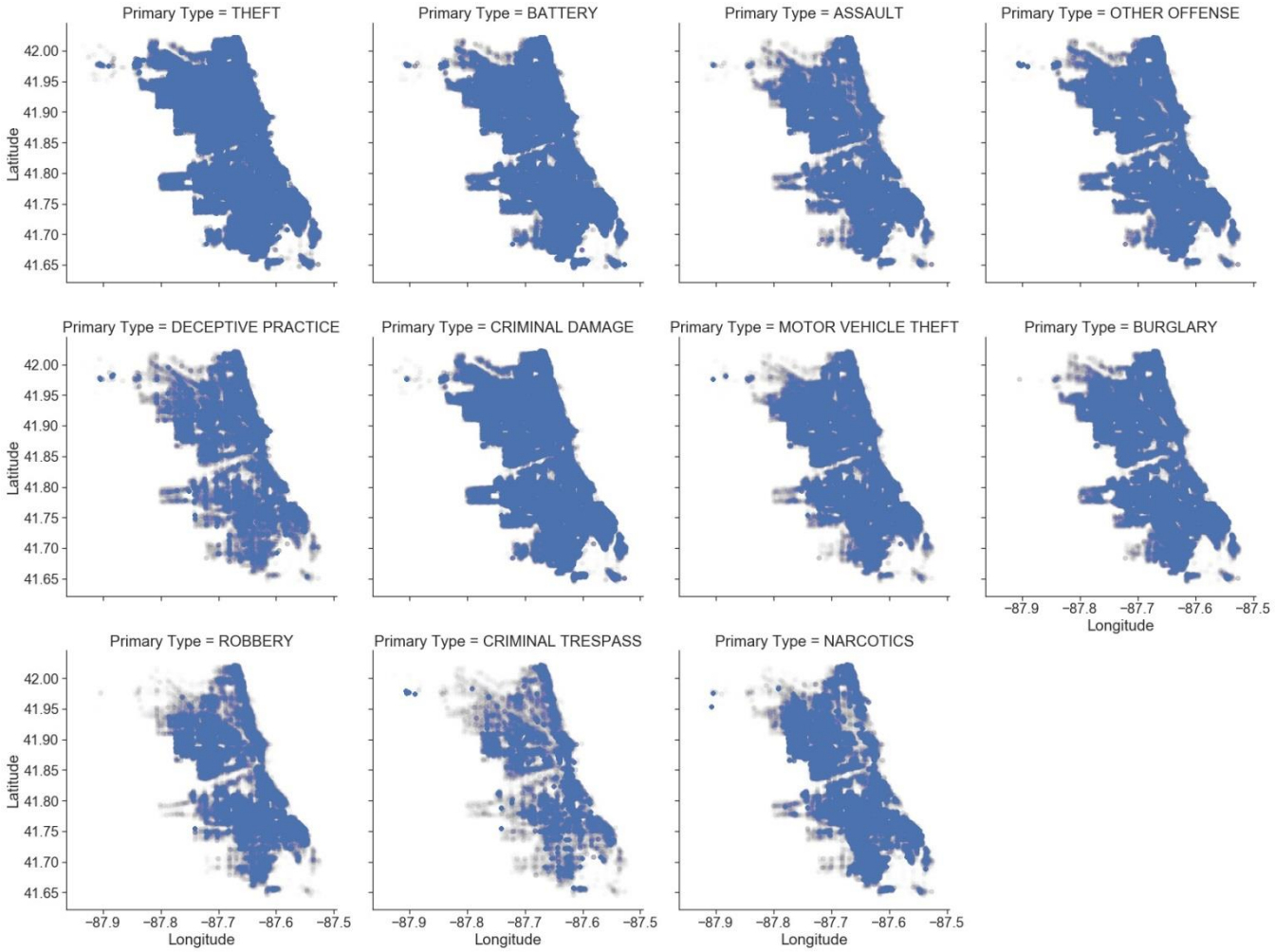

Figure 20: Spatial distribution of crimes by primary type of crime.

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°).

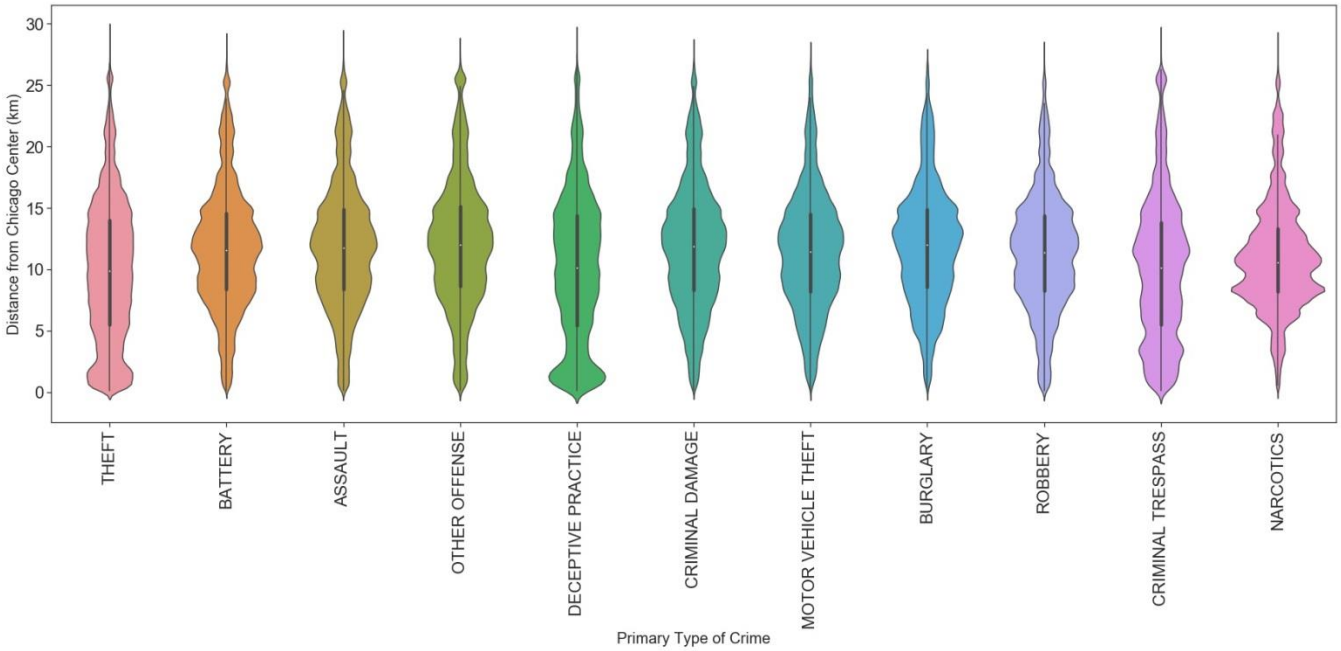

Figure 21: Distribution of crimes based on distance from Chicago city center for each primary type of crime.

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.

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.

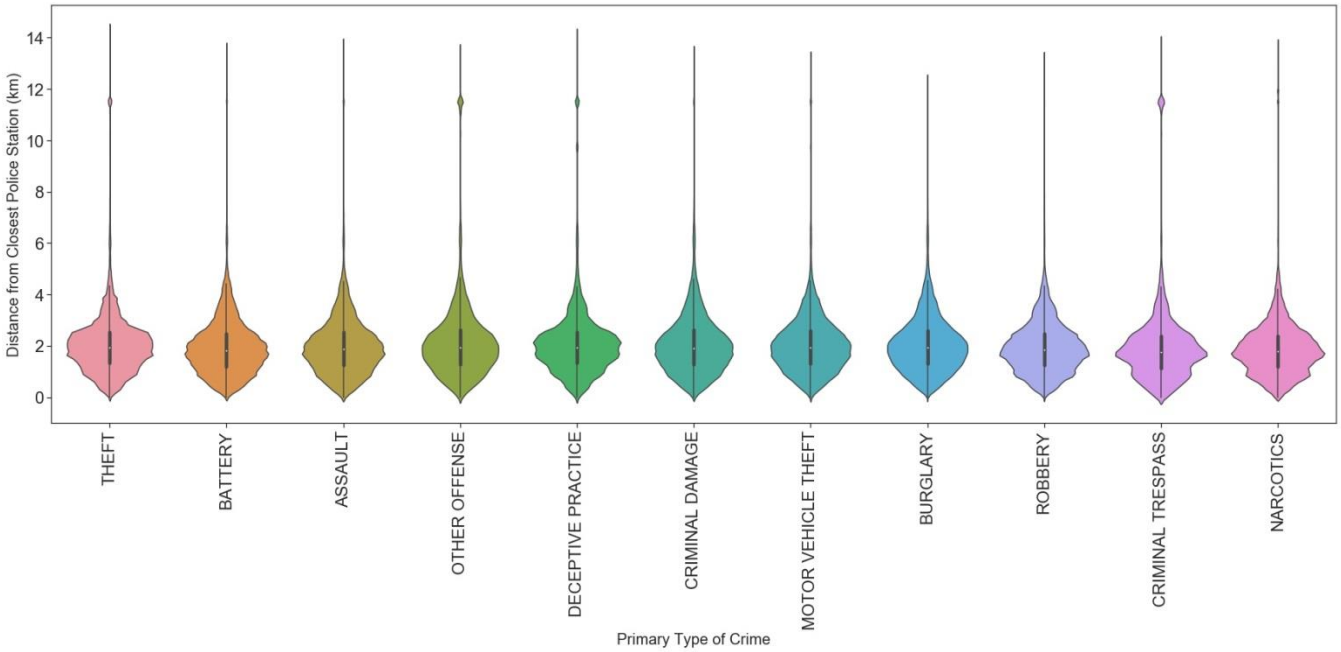

Figure 22: Distribution of crimes based on distance from closest police station for each primary type of crime.

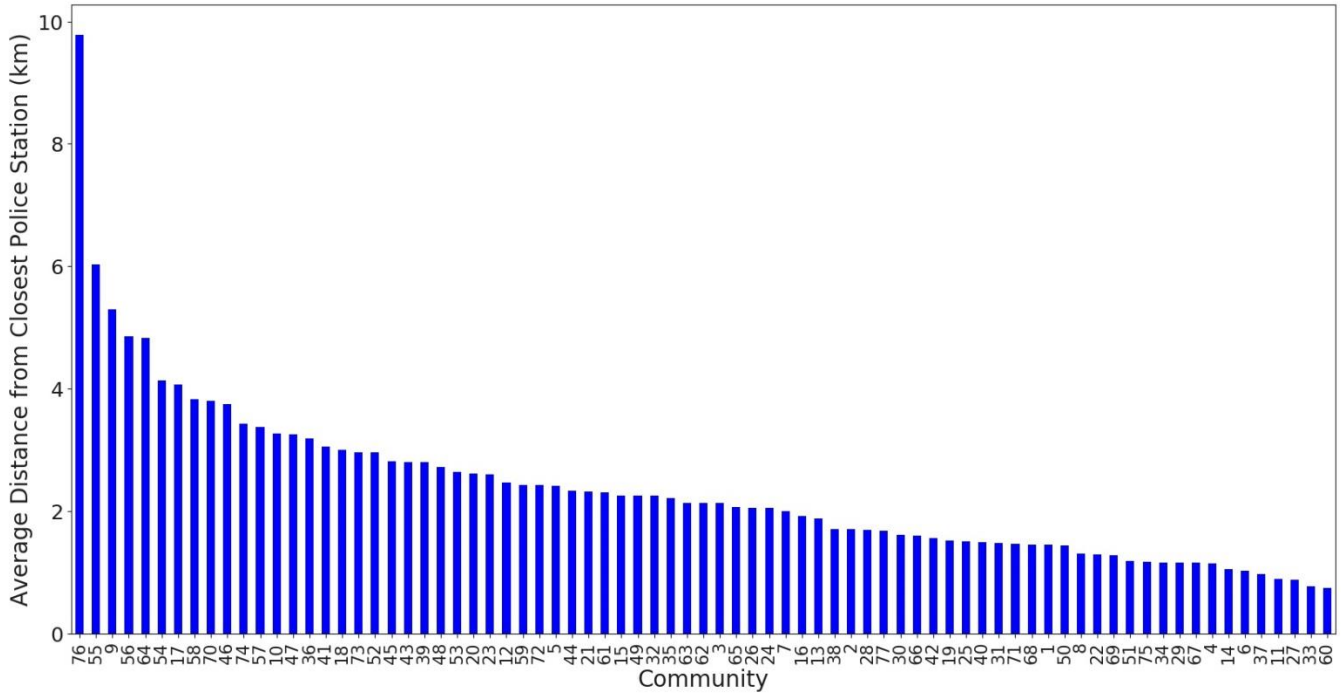

Figure 23: Average distance from closest police station for each community.

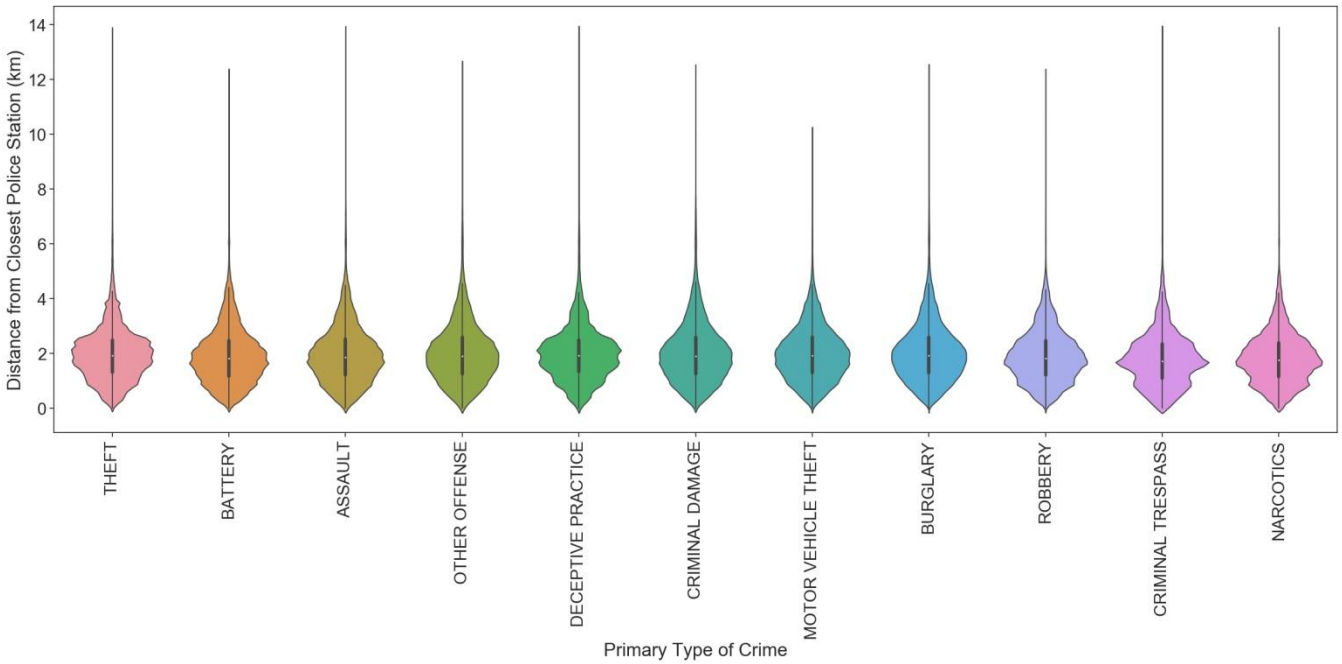

Figure 24: Distribution of crimes based on distance from closest police station for each primary type of crime (excluding community 76).

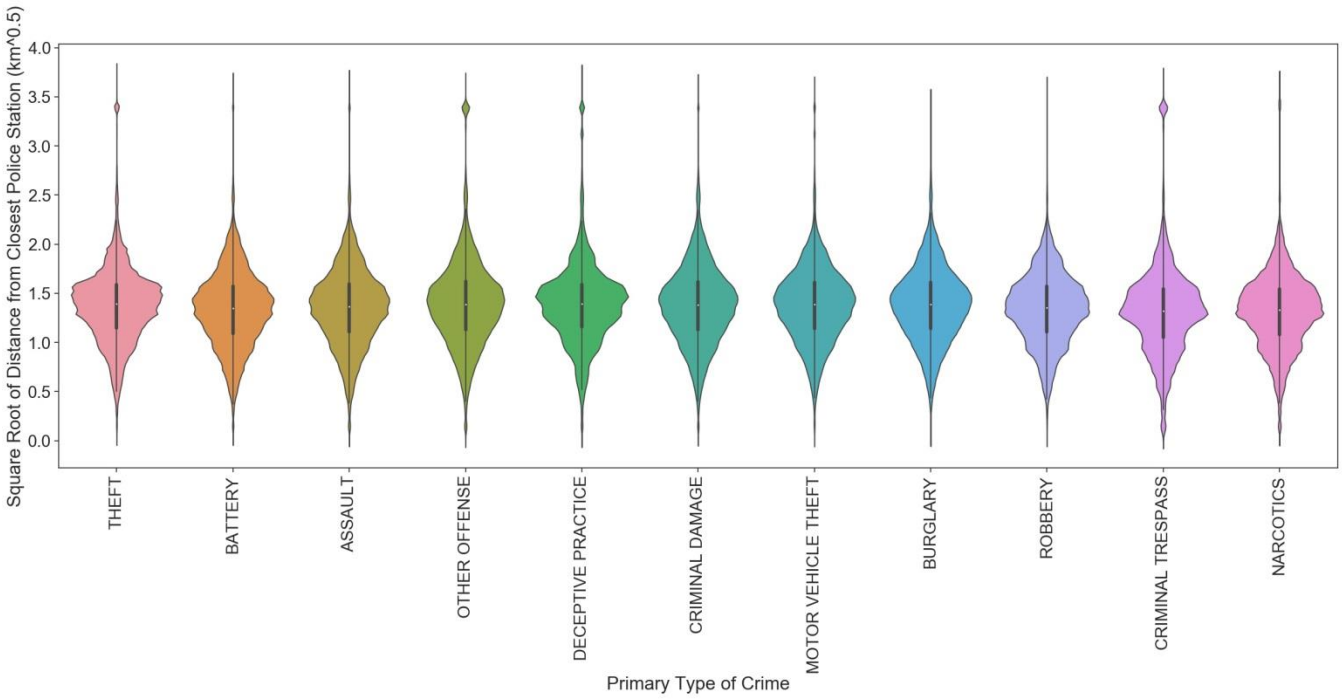

Figure 25: Distribution of crimes based on square root of distance from closest police station for each primary type of crime.

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.

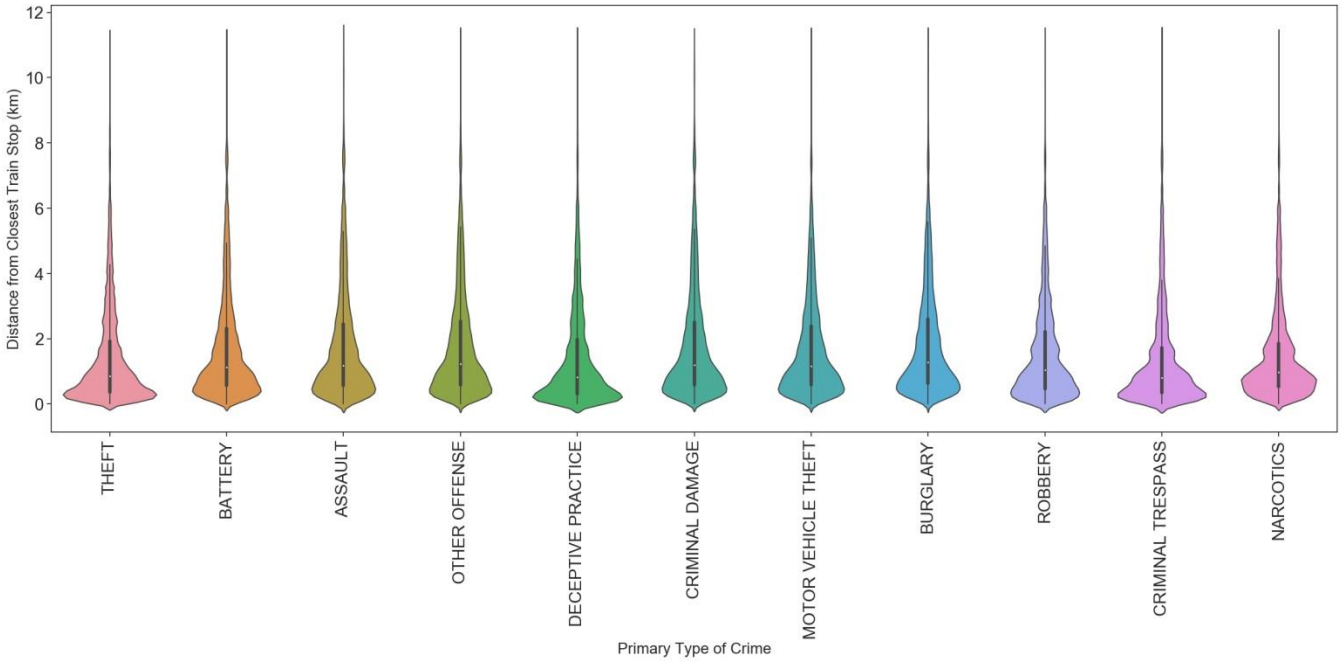

Figure 26: Distribution of crimes based on distance from closest train stop for each primary type of crime.

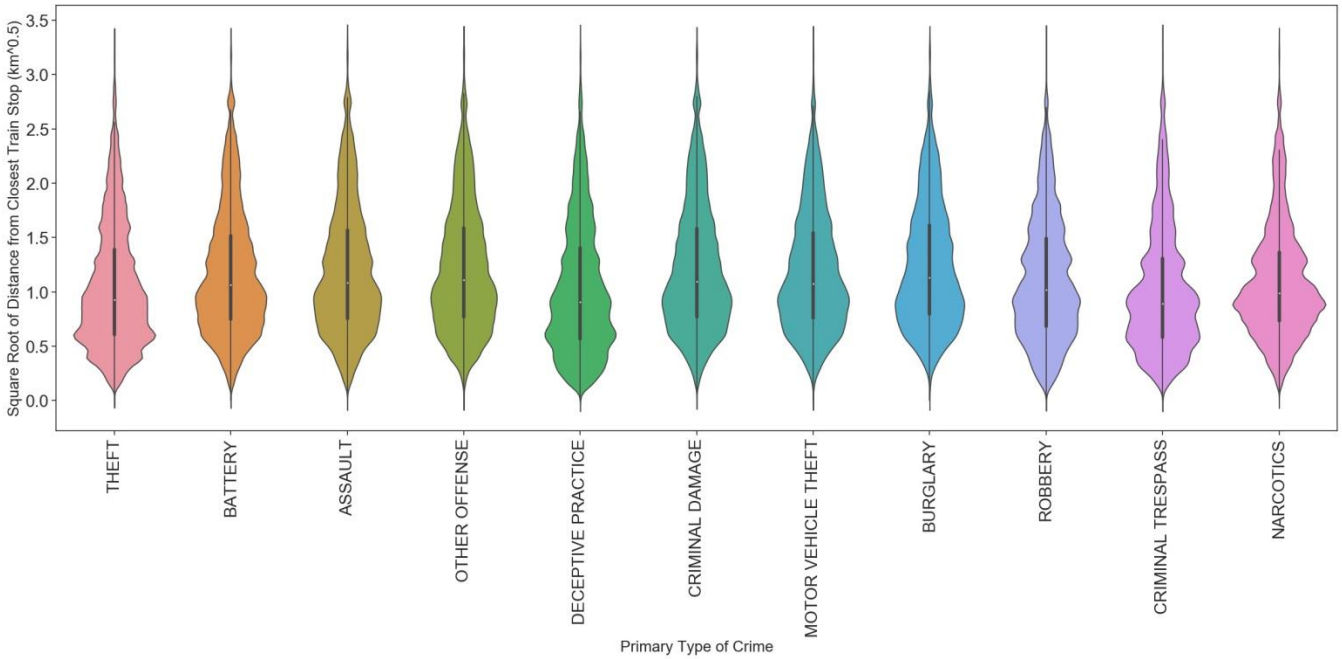

Figure 27: Distribution of crimes based on square root of distance from closest train stop for each primary type of crime.

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.

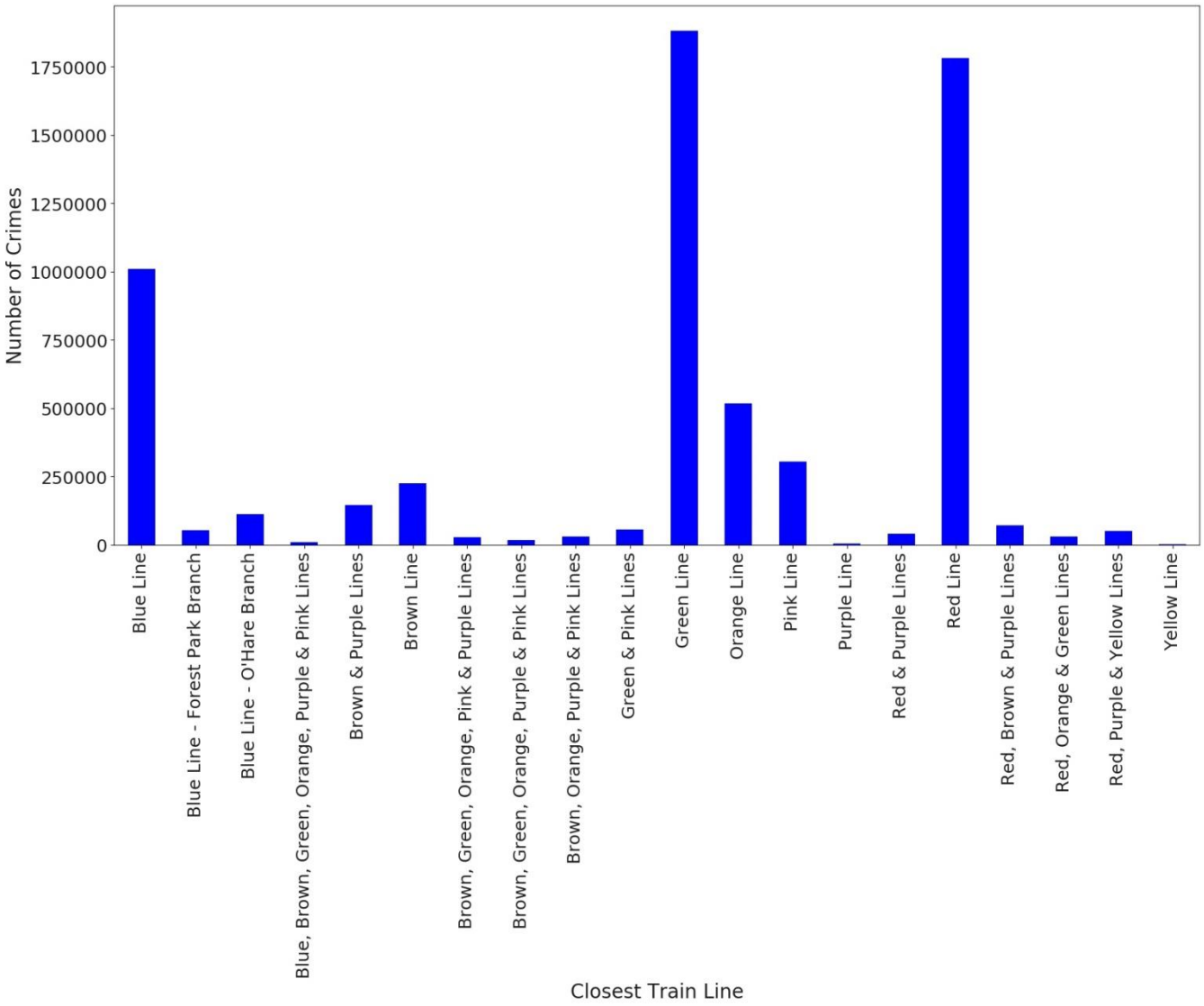

Figure 28: Number of crimes per closest train line.

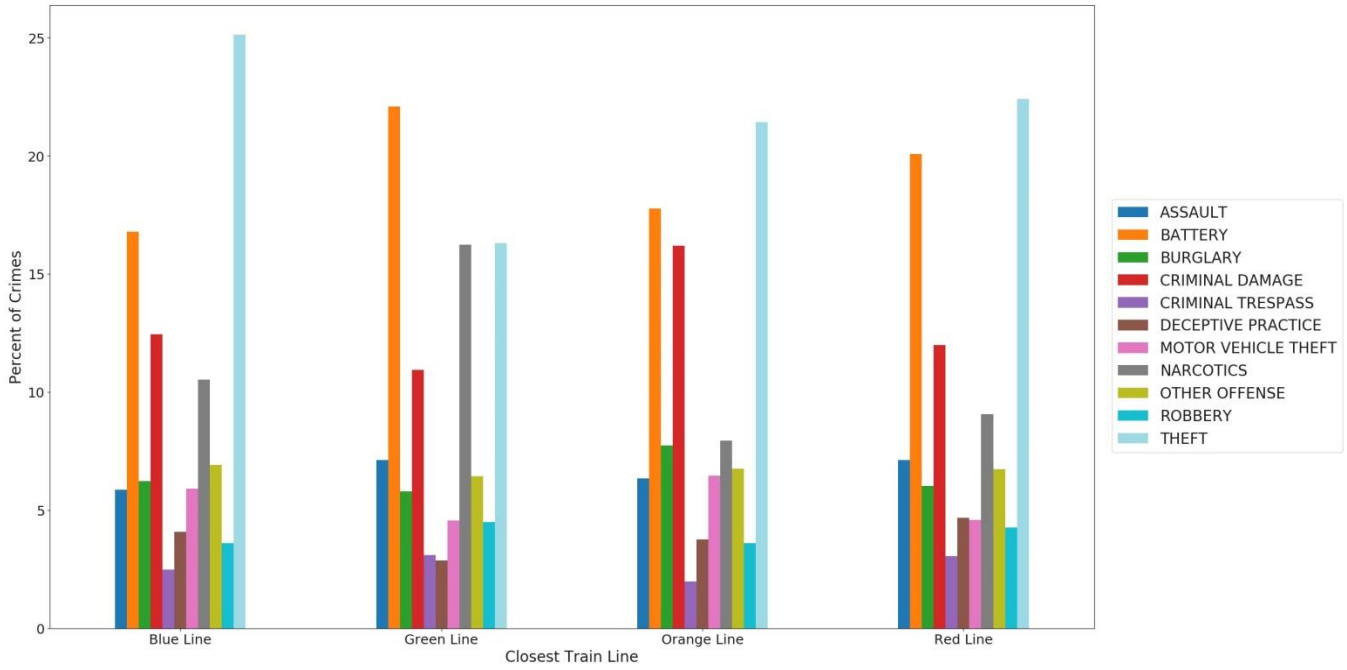

Figure 29: Percentage of each primary type of crime per closest train line.

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 most of the crime types 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.

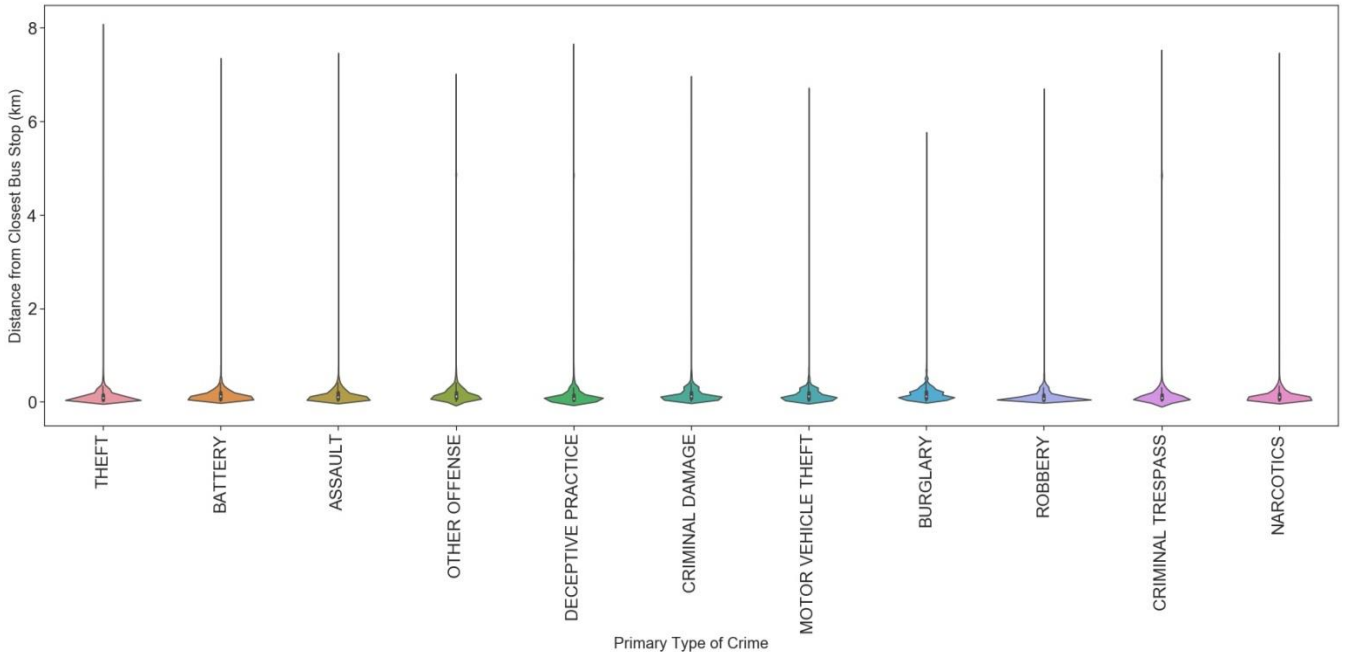

Figure 30: Distribution of crimes based on distance from closest bus stop for each primary type of crime.

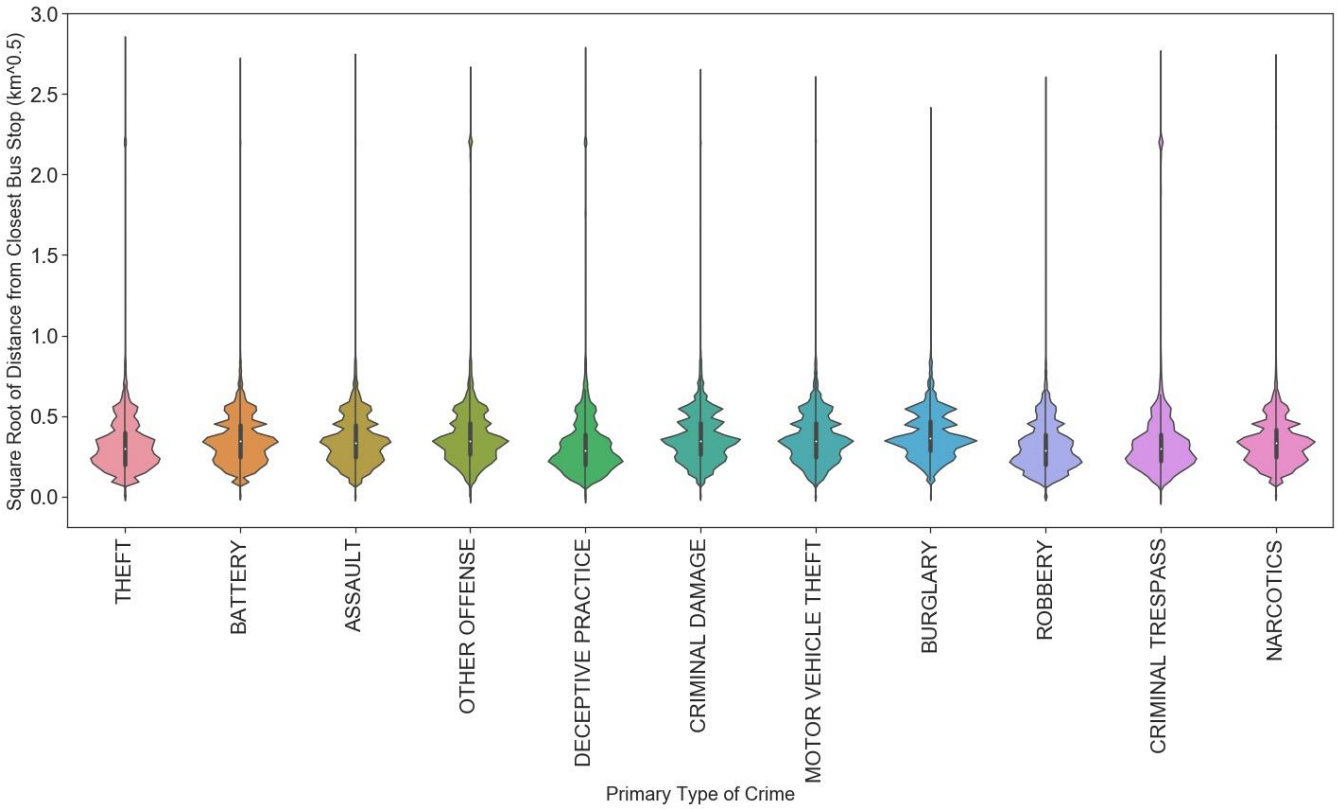

Figure 31: Distribution of square root of distance from closest bus stop for each primary type of crime.

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.

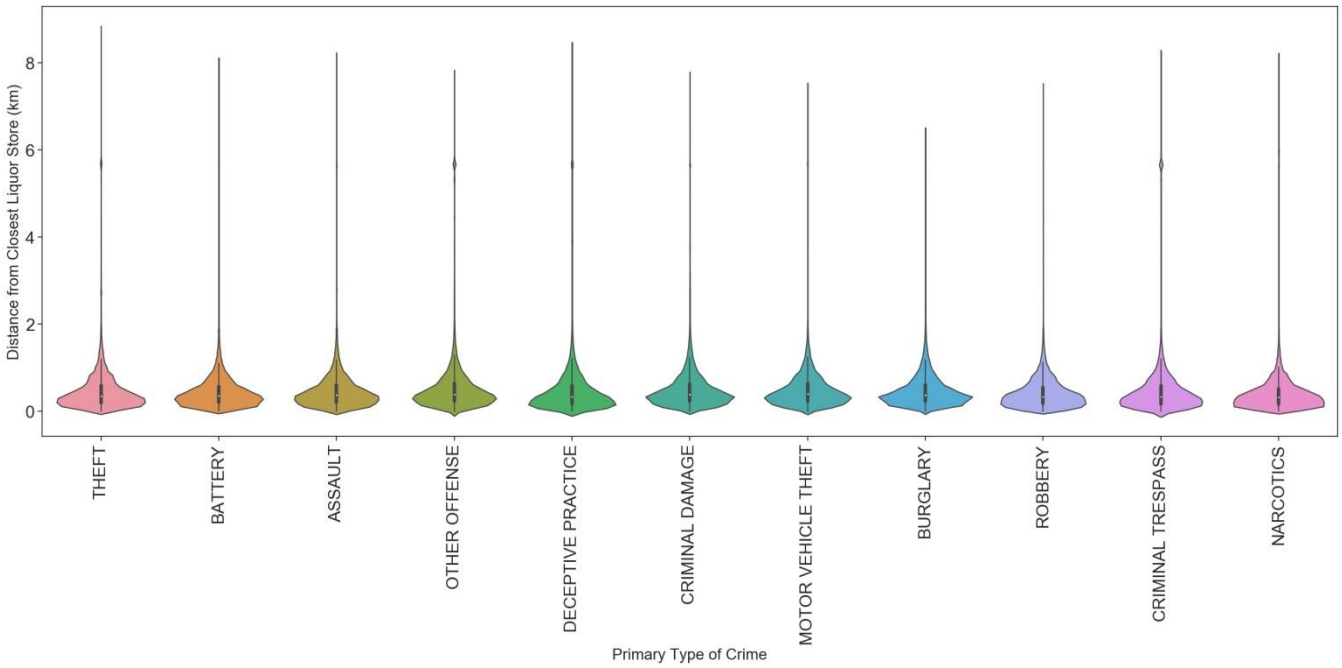

Figure 32: Distribution of crimes based on distance from closest liquor store for each primary type of crime.

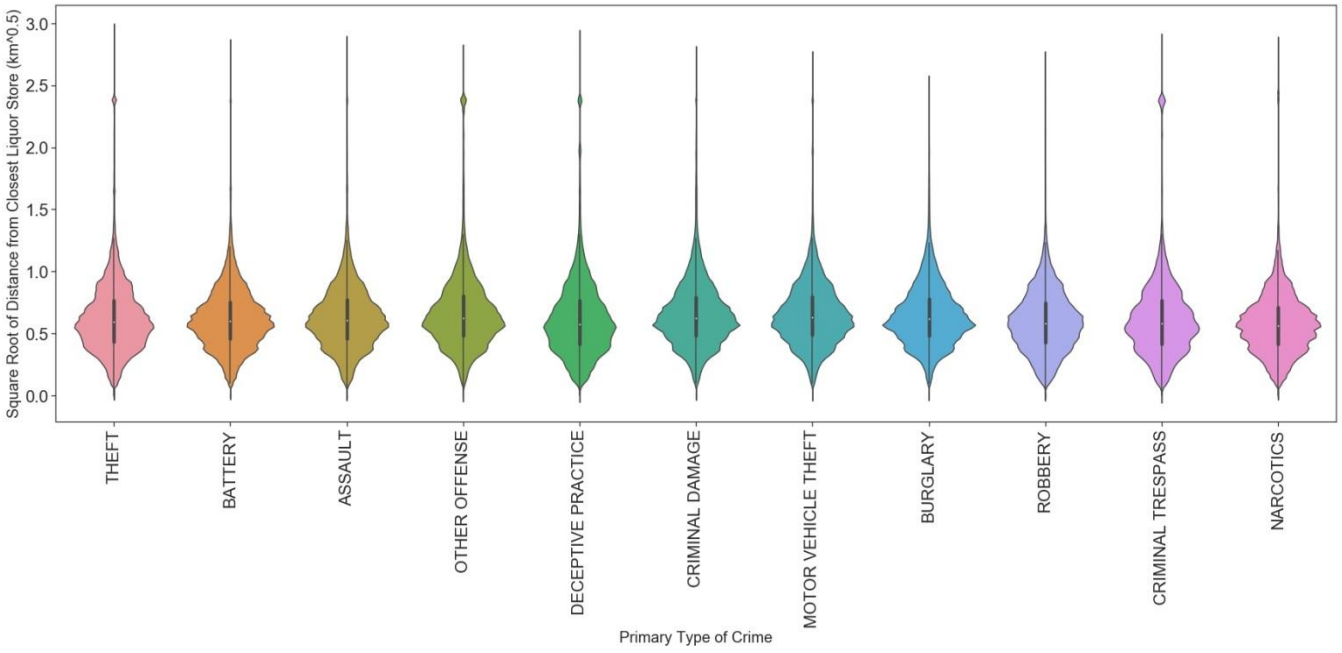

Figure 33: Distribution of crimes based on square root of distance from closest liquor store for each primary type of crime.

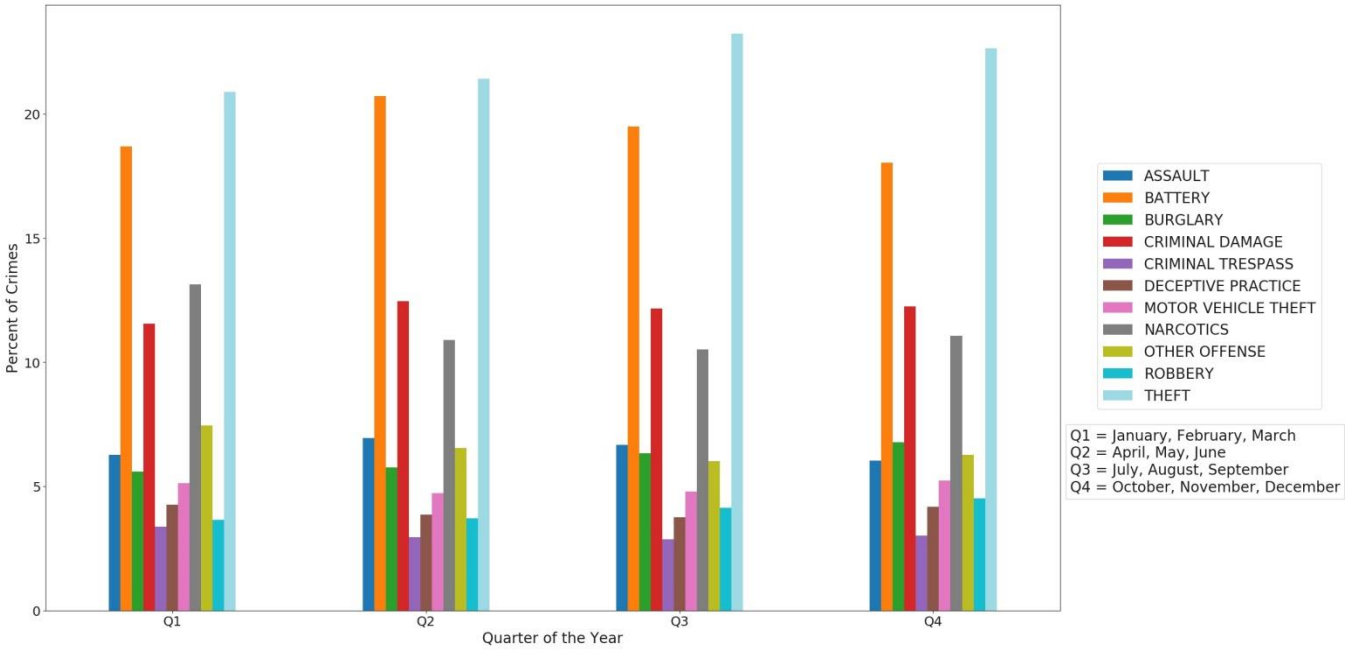

Figure 34: Percentage of each primary type of crime per quarter of the year.

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 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.

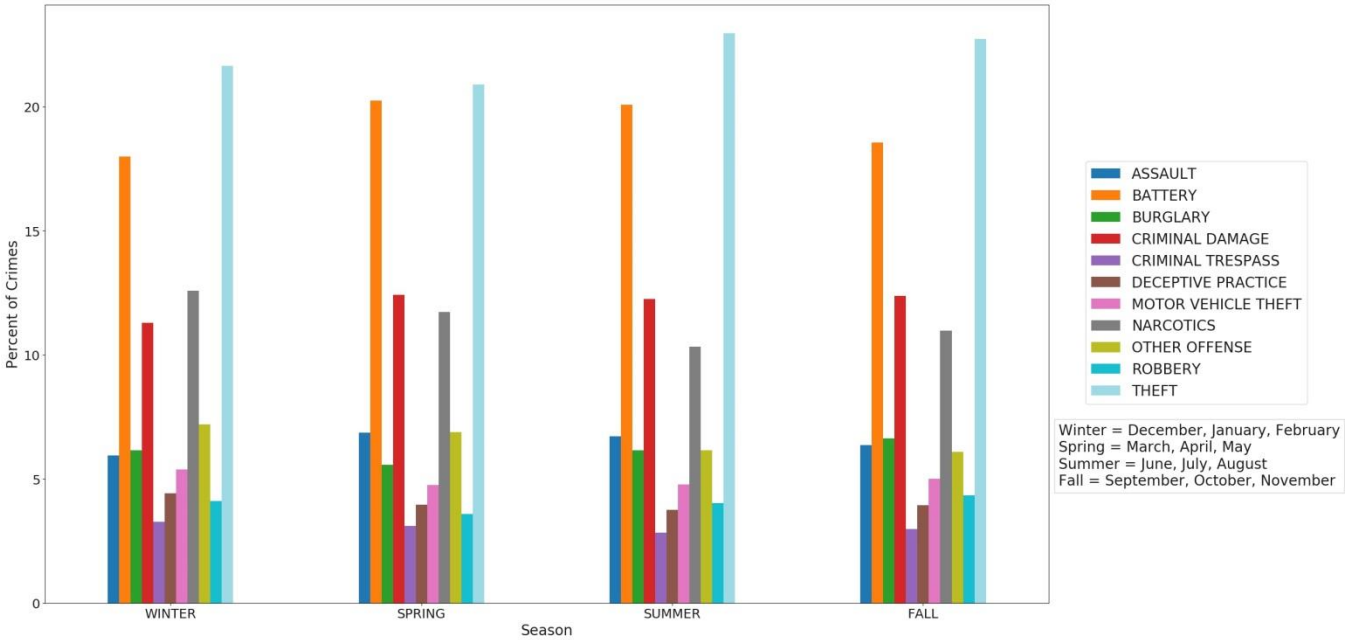

Figure 35: Percentage of each primary type of crime per season.

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.

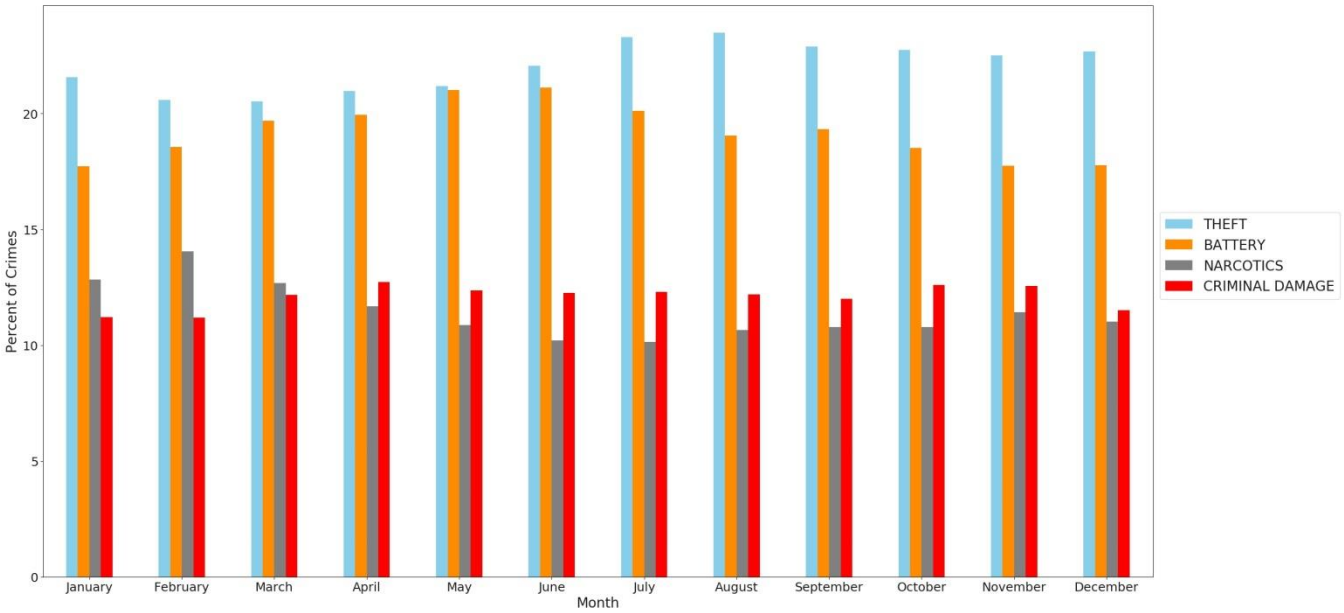

Figure 36: Percentage of 4 primary types of crime per month.

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

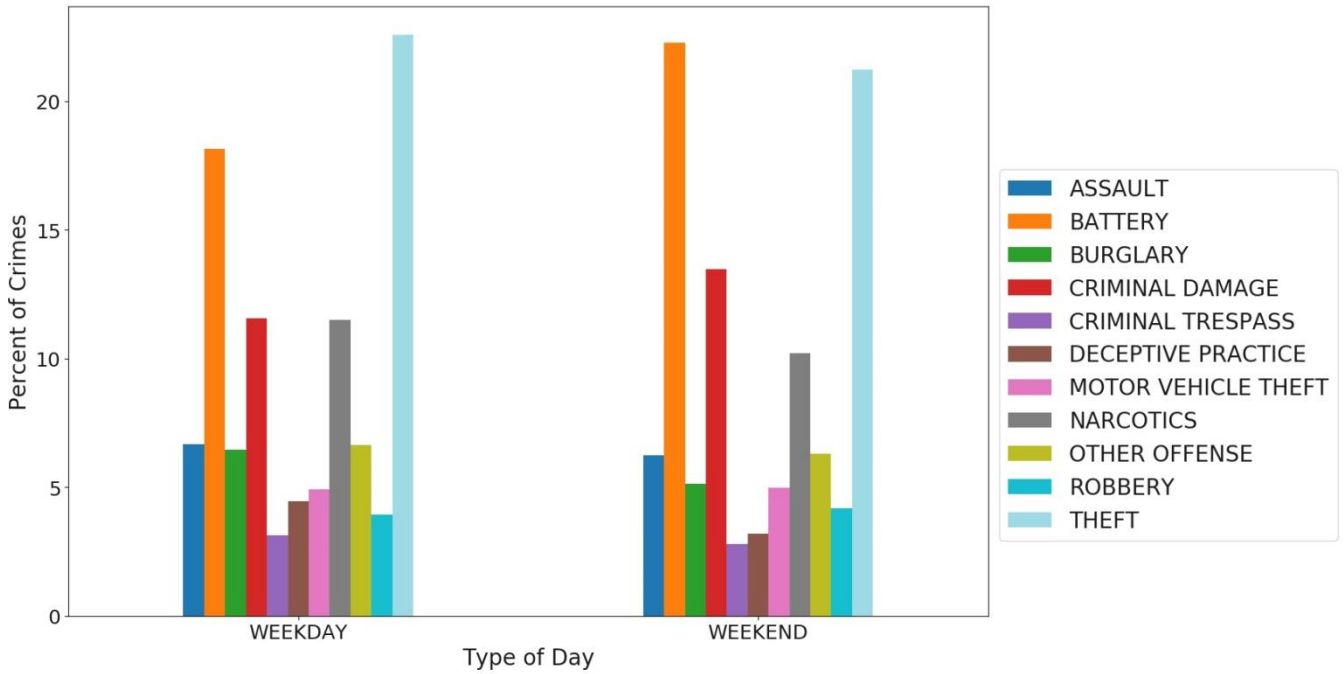

Figure 37: Percentage of each primary type of crime per type of day (weekday/weekend).

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

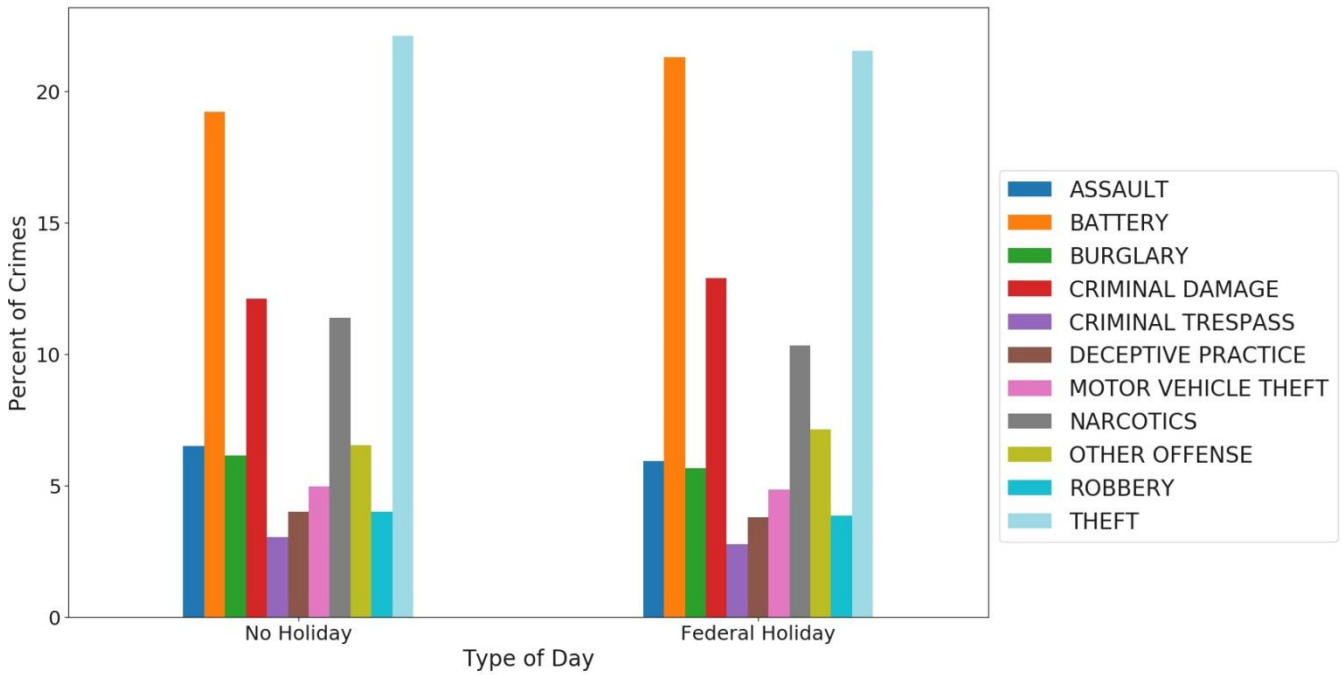

Figure 38: Percentage of each primary type of crime per type of day (no holiday/federal holiday).

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.

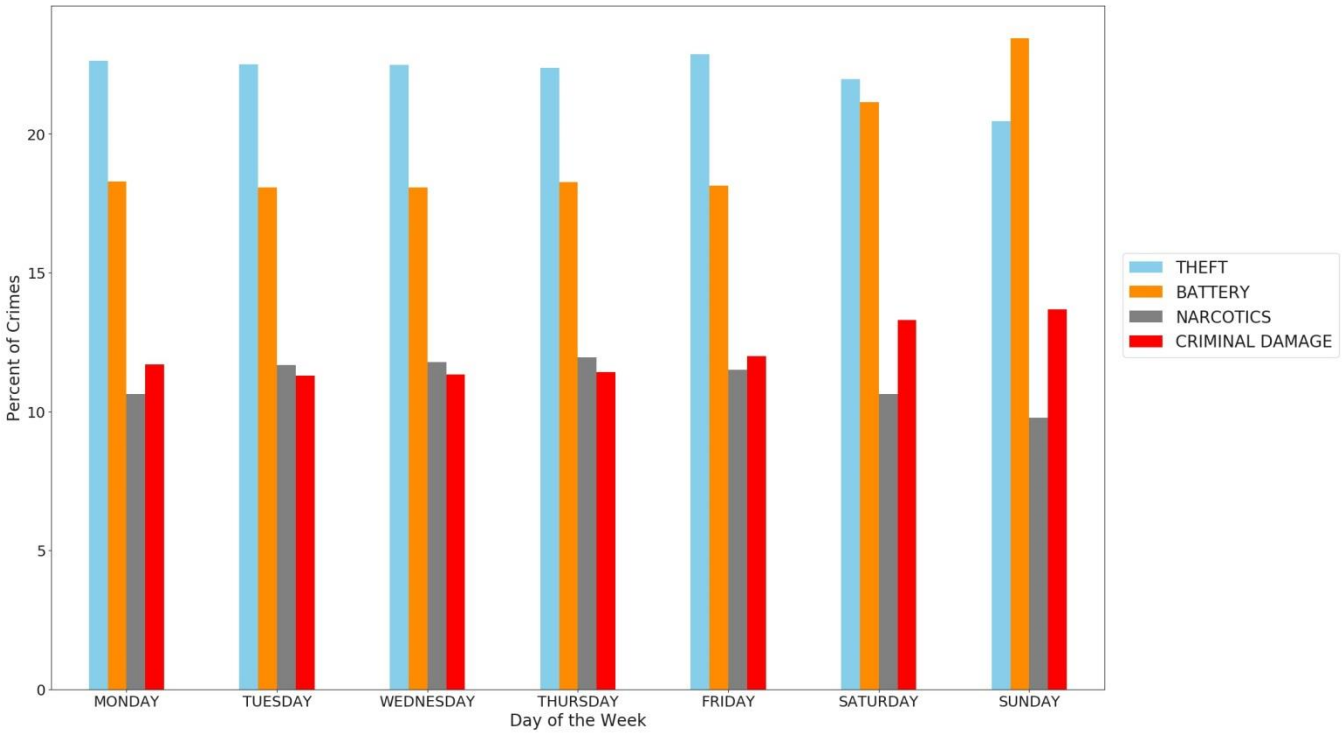

Figure 39: Percentage of 4 primary types of crime per day of the week.

Per Figure 40, Christmas and Independence Day have the highest proportions of crimes involving battery while Martin Luther King Jr. Day, Veterans 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.

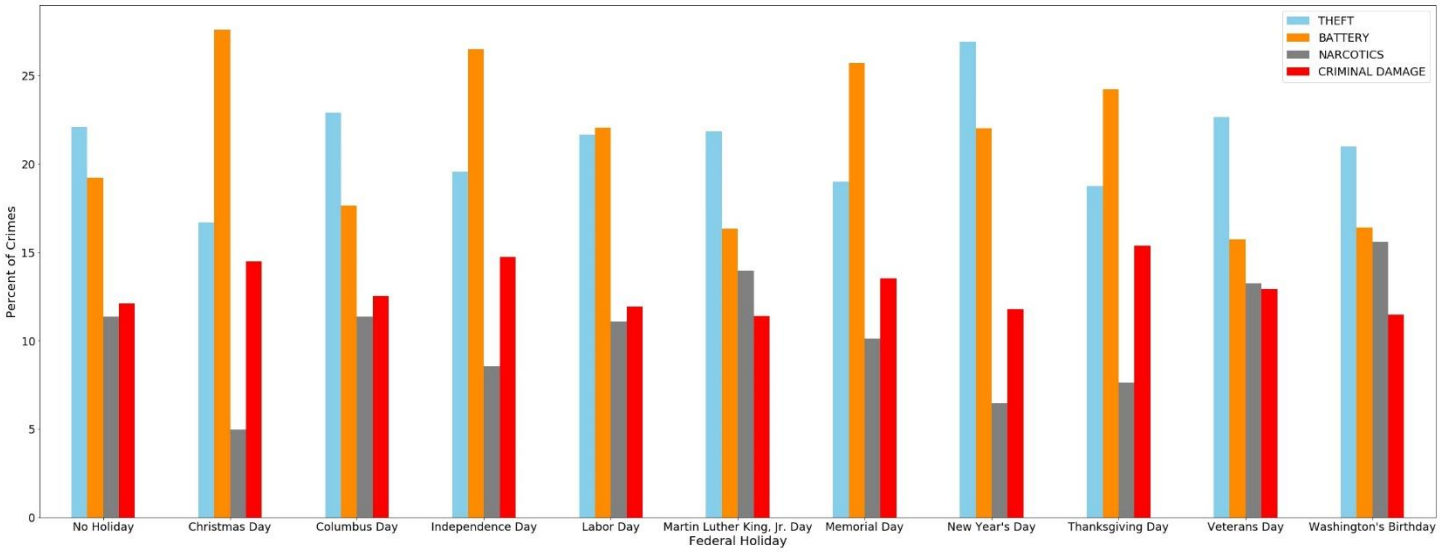

Figure 40: Percentage of 4 primary types of crime per federal holiday.

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 are higher proportions 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.

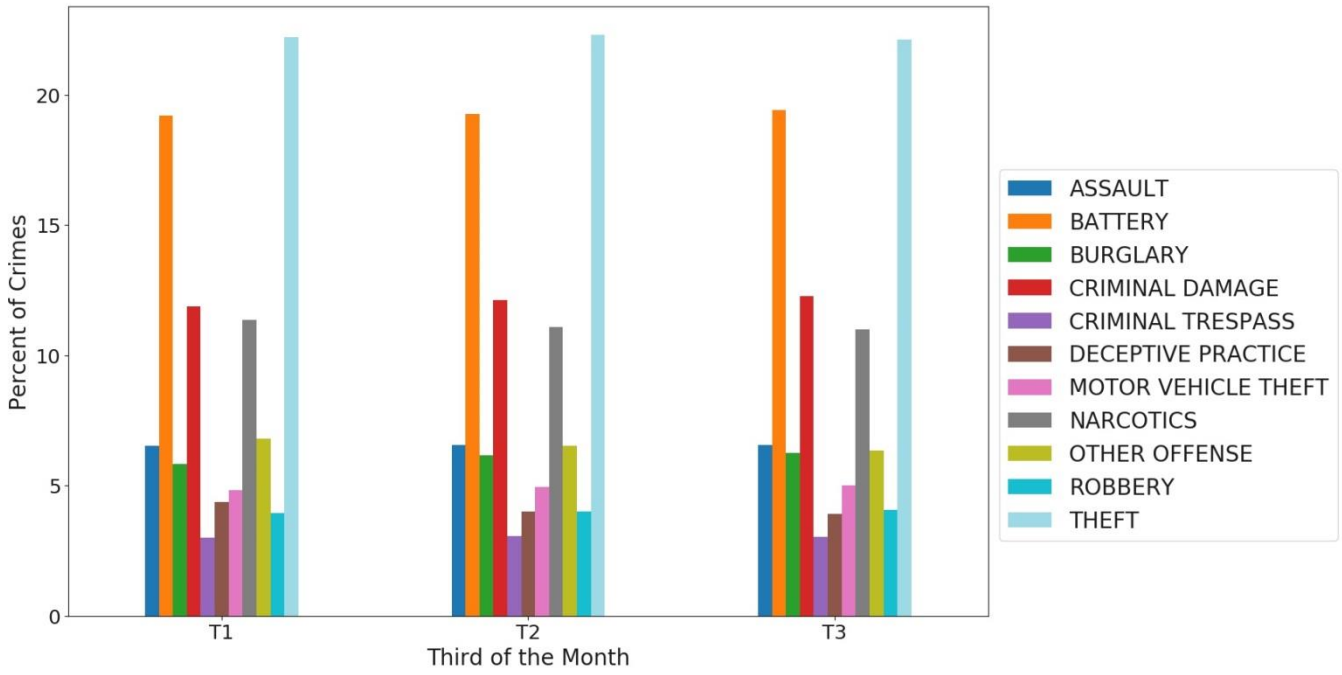

Figure 41: Percentage of each primary type of crime per third of the month.

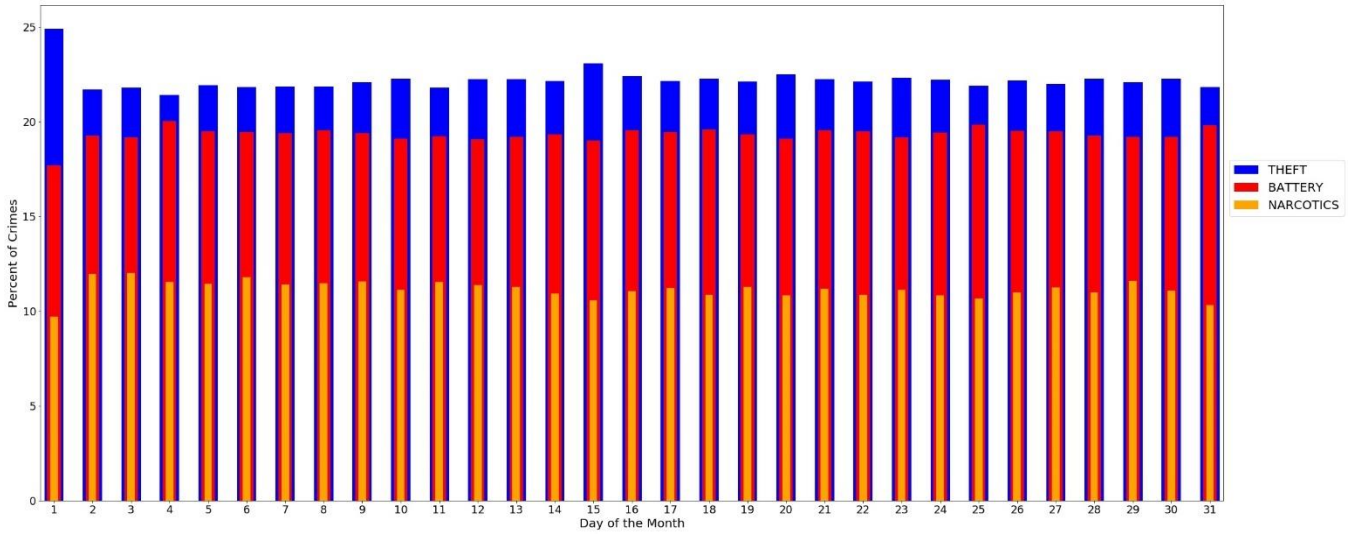

Figure 42: Percentage of 3 primary types of crime per day of the month.

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.

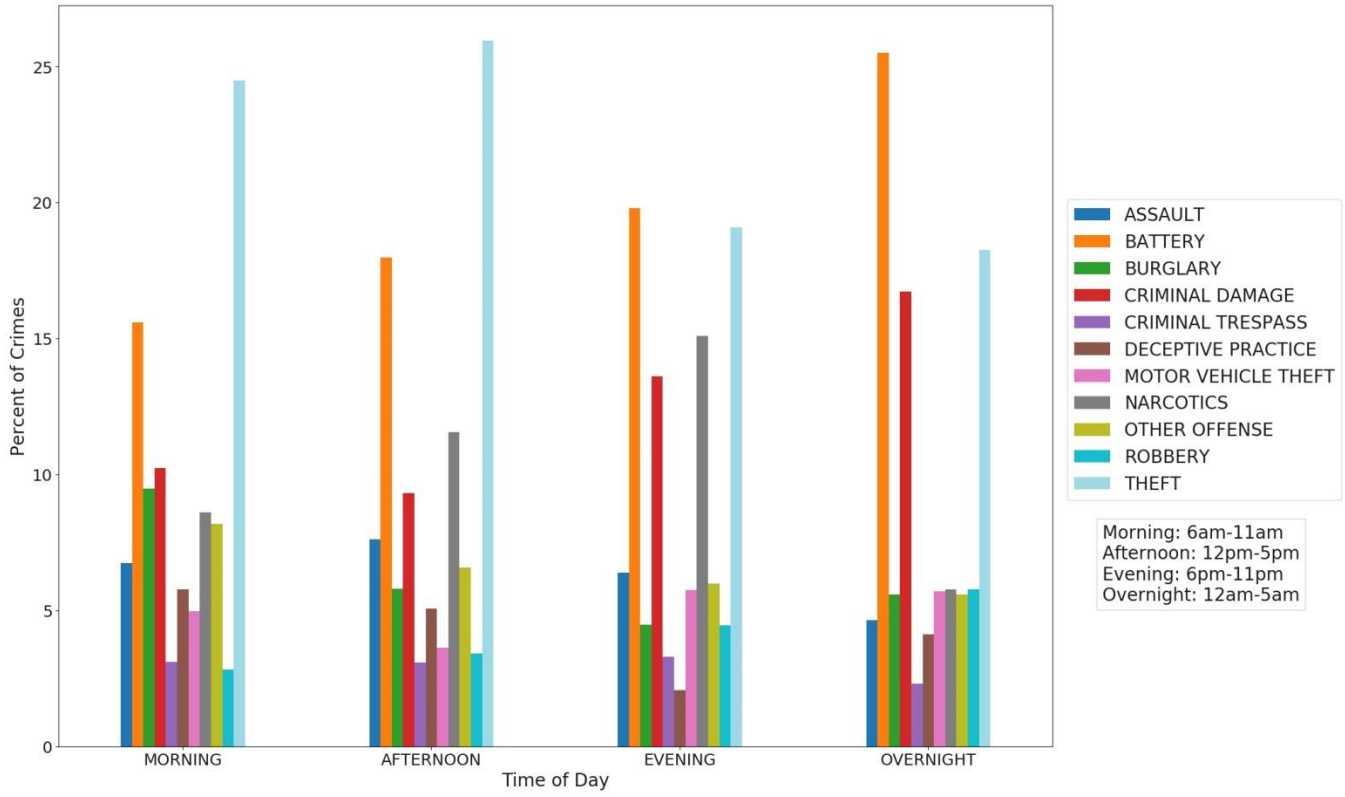

Figure 43: Percentage of each primary type of crime per time of day.

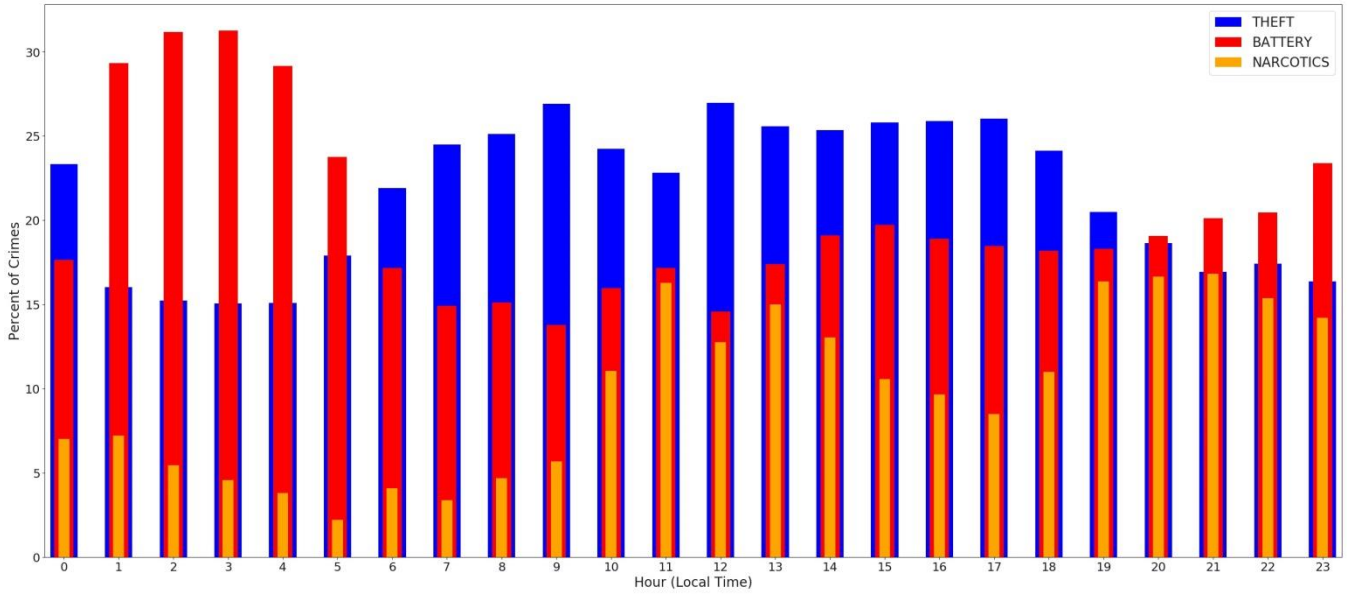

Figure 44: Percentage of 3 primary types of crime per hour.

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.

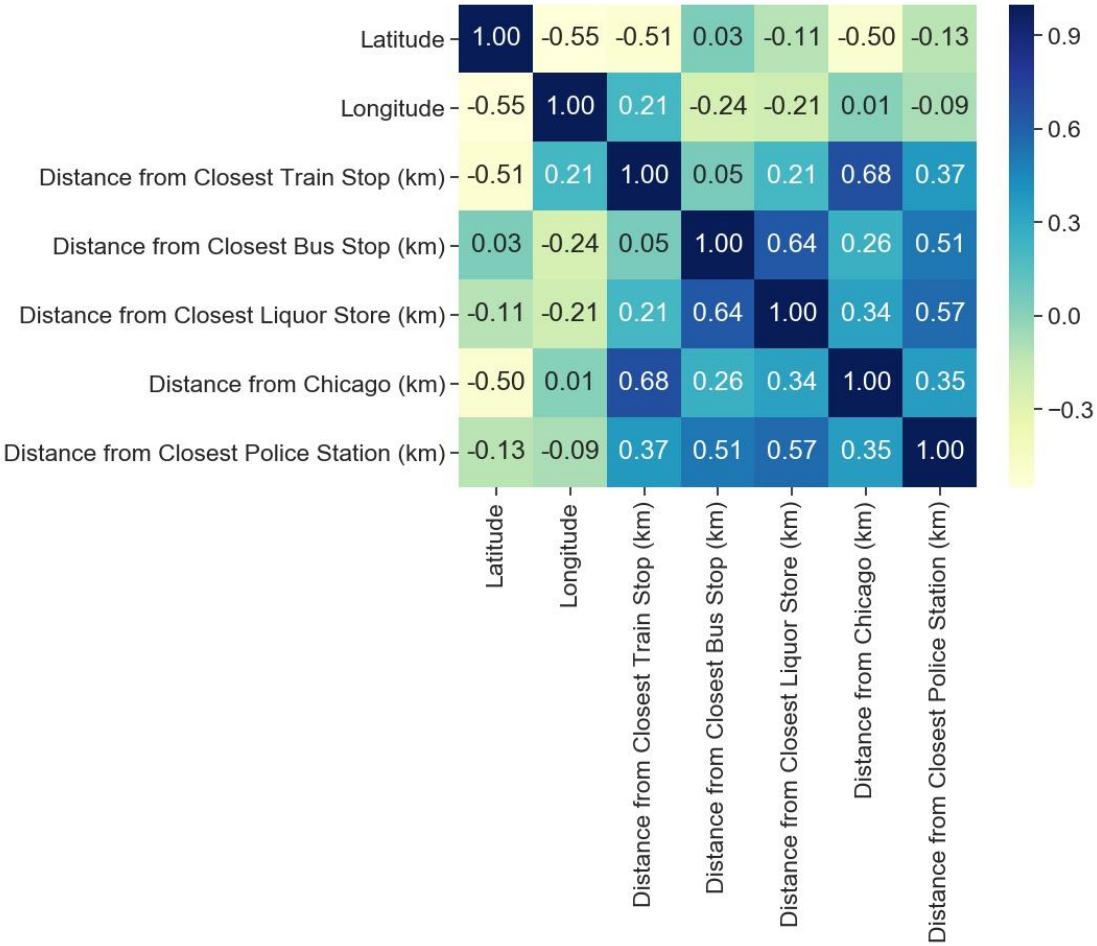

Figure 45: Pearson correlation coefficients between numerical features.

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.

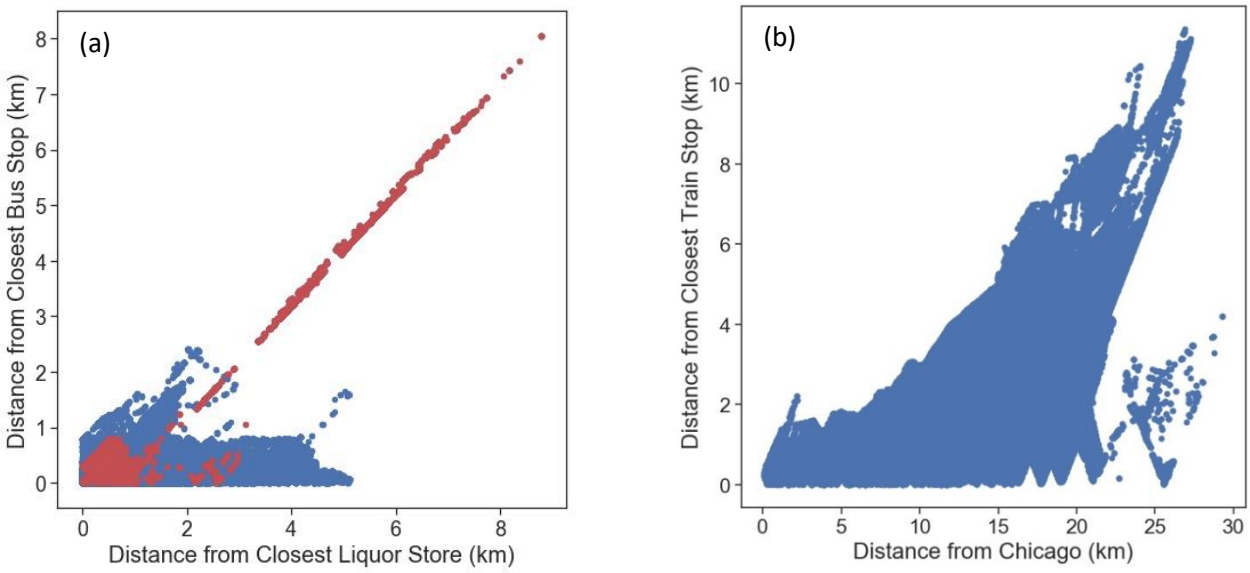

Figure 46: (a) Distance from closest bus stop vs. distance from closest liquor store (red points are crimes within community 76) and (b) distance from closest train stop vs. distance from Chicago.

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

Model Evaluation

Table 1 shows the models that were evaluated and the parameters that were adjusted. As the Multinomial NB could only accept positive values, I used the X/Y coordinates instead of the longitude/latitude.

Table 1: Adjusted parameters and shortened name for each evaluated model.

Model Name	Adjusted Model Parameters	Shortened Model Name
SGD Classifier		SGDa
SGD Classifier	class_weight='balanced'	SGDb
SGD Classifier	class_weight='balanced', loss='log'	SGDc
Multinomial NB		MNB
Decision Tree Classifier		DTCa
Decision Tree Classifier	class_weight='balanced'	DTCb
Extra Tree Classifier		ETCa
Extra Tree Classifier	class_weight='balanced'	ETCb
Random Forest Classifier		RFCa
Random Forest Classifier	class_weight='balanced'	RFCb
Gaussian NB		GNB

As I lacked sufficient computer memory to use all of the data (6,357,103 reports) to train/test the models, I randomly chose a sample of 3 million crime reports. I then converted the following features into dummy variables:

- Ward
- Police district
- Police beat
- Community
- Location description
- Closest train line
- 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

The data was then split, with 75% of the data used to train the models and the remainder of the data used to test them.

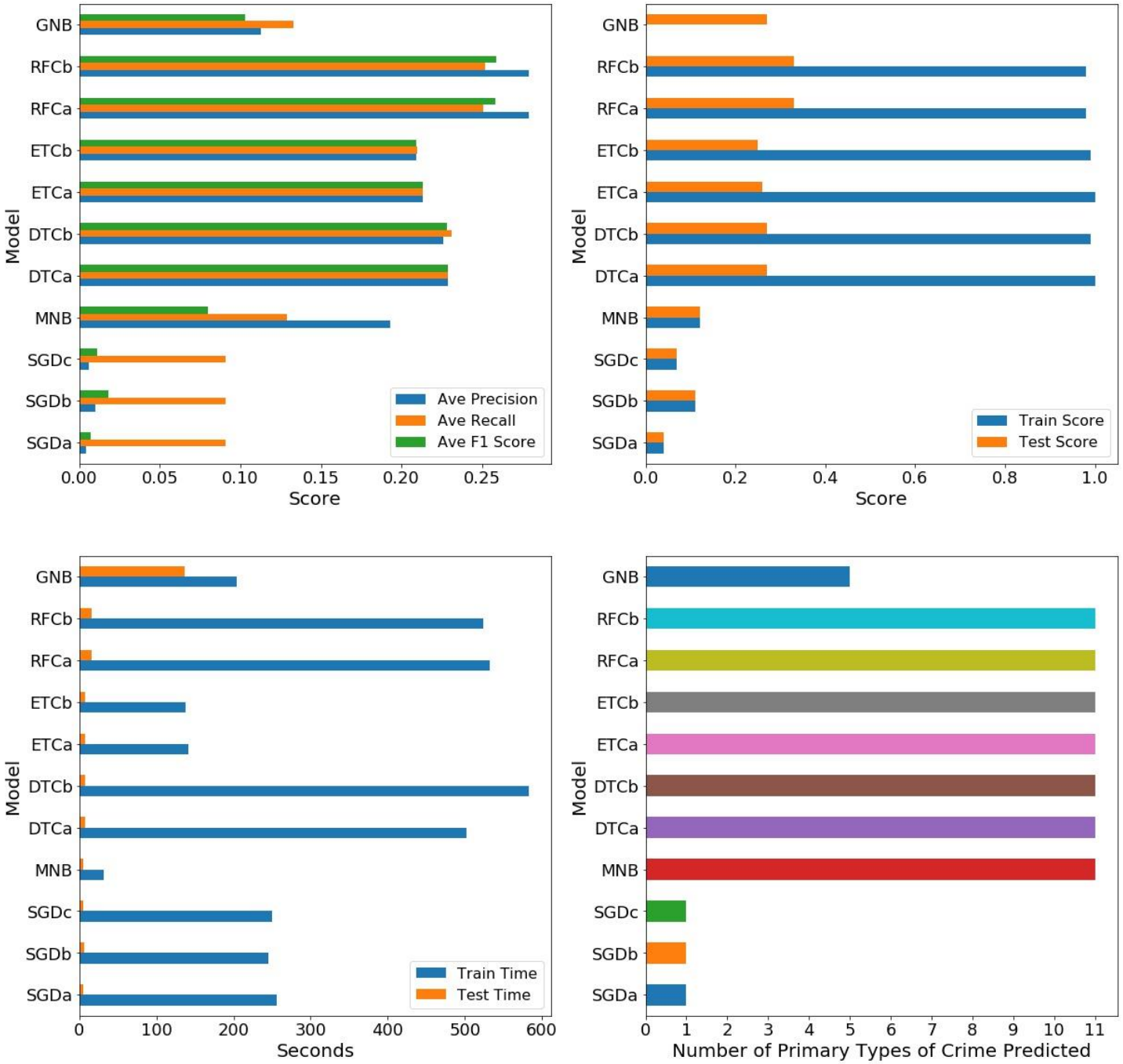

Figure 47: Model results for a sample of 3 million crime reports. * The training score for GNB is not reported as it took too long to retrieve.

Figure 47 shows the results for using the sample of crime reports with dummy variables as is. SGDa/b/c and GNB did not predict every primary type of crime and thus had the overall worst performances. It is interesting that setting the class weight to 'balanced' in SGDb/c did not increase the number of crimes predicted.

RFCa/b had the best average F1 score, recall, and precision; but according to its training score of approximately 1, it overfit the data. At 500-600s, DTCa/b and RFCa/b took the longest to train.

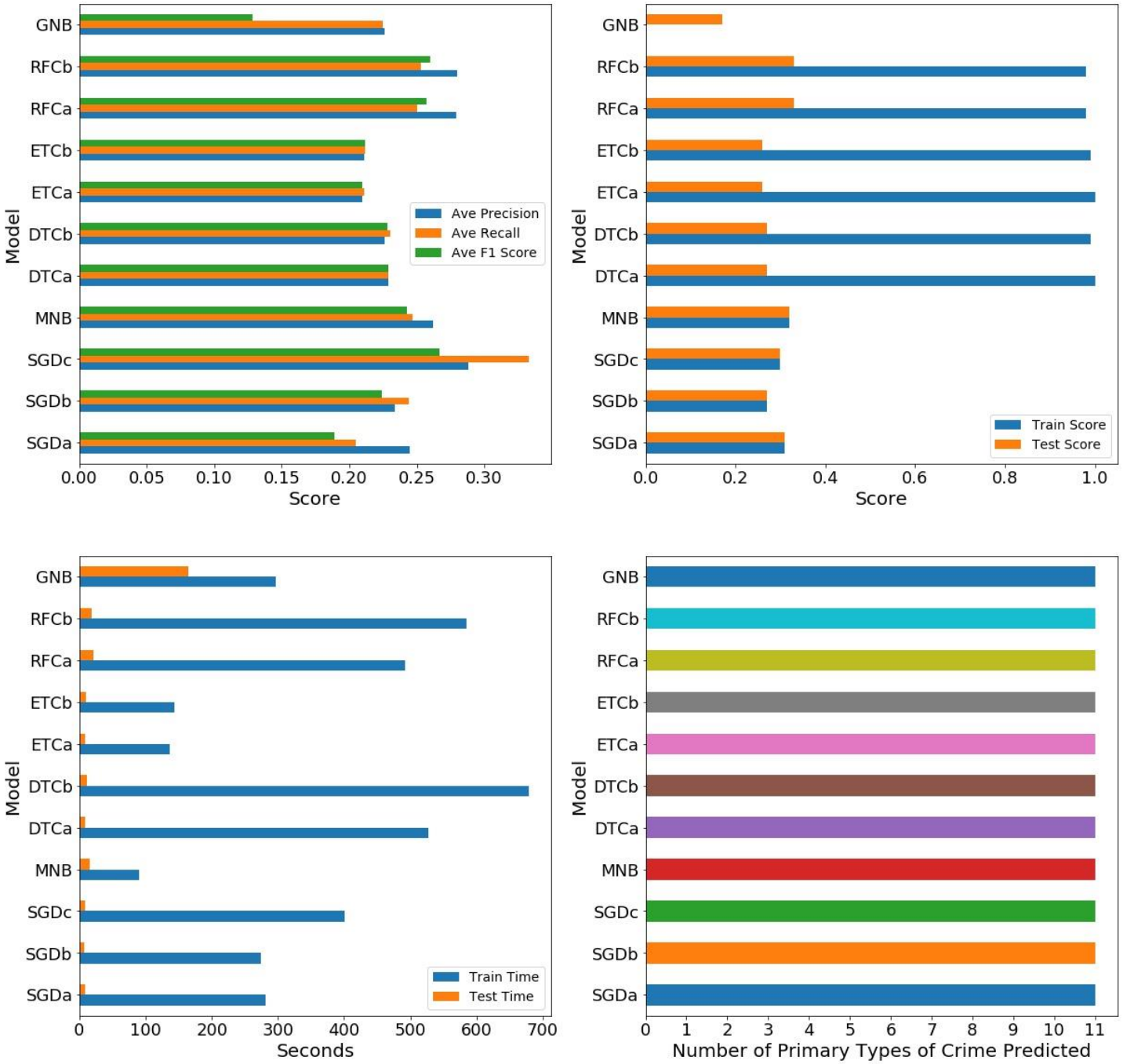

Figure 48: Model results for a sample of 3 million crime reports scaled with MinMaxScaler. * The training score for Gaussian NB is not reported as it took too long to retrieve.

The sample data with dummy variables was then scaled to a range between 0 and 1 using MinMaxScaler. I opted to use this transformation as it wouldn't affect the dummy variables which were either 0 or 1. Per Figure 48, scaling the data allowed for every model to predict all 11 primary types of crime. This was an unexpected outcome. As seen from running the models without scaling the data, setting the class weights to 'balanced' did not increase the number of crimes predicted. It could be that using the X/Y coordinates threw things off in the previous run for SGDa/b/c and GNB as these features are so much larger (on the order of 10^6). Therefore, scaling all features so that they have the same range may have allowed these models to perform better in this aspect.

SGDc performed the best when looking at the average F1 score, recall, and precision; RFCa/b was the second best, but it still overfit the training data. Looking at the test scores, SGDc did not do the best; MNB and RFCa/b both performed slightly better. It is possible that these scores could change with different test data. In

the future, cross validation should be performed so that we could be certain of which model performs the best. At 500-700s, DTCa/b and RFCa/b had the longest training times; SGDc was less at 400s.

The original sample data with dummy variables was then transformed using IncrementalPCA. Looking at the explained variance of each PCA feature (Figure 49), I chose to reduce the dimensionality of the data to 2 components. Since out of 638 features, only 2 features stood out, it is possible that there are a lot of correlated variables in my dataset. This makes sense as there are relationships between the police beat, police district, community, and ward. In addition, there are relationships between hour and time of day and month and season

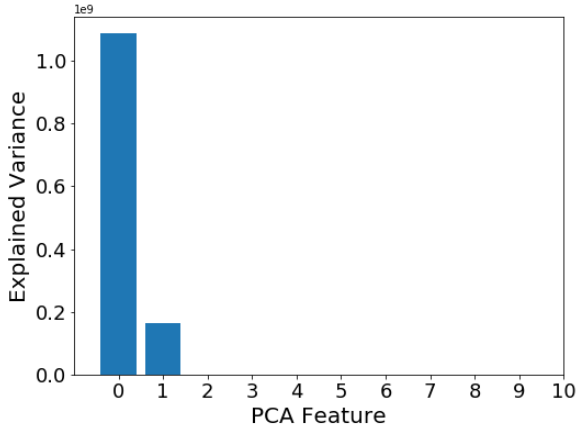

Figure 49: Explained variance vs PCA feature for a sample of 3 million crime reports. In total, there are 638 features.

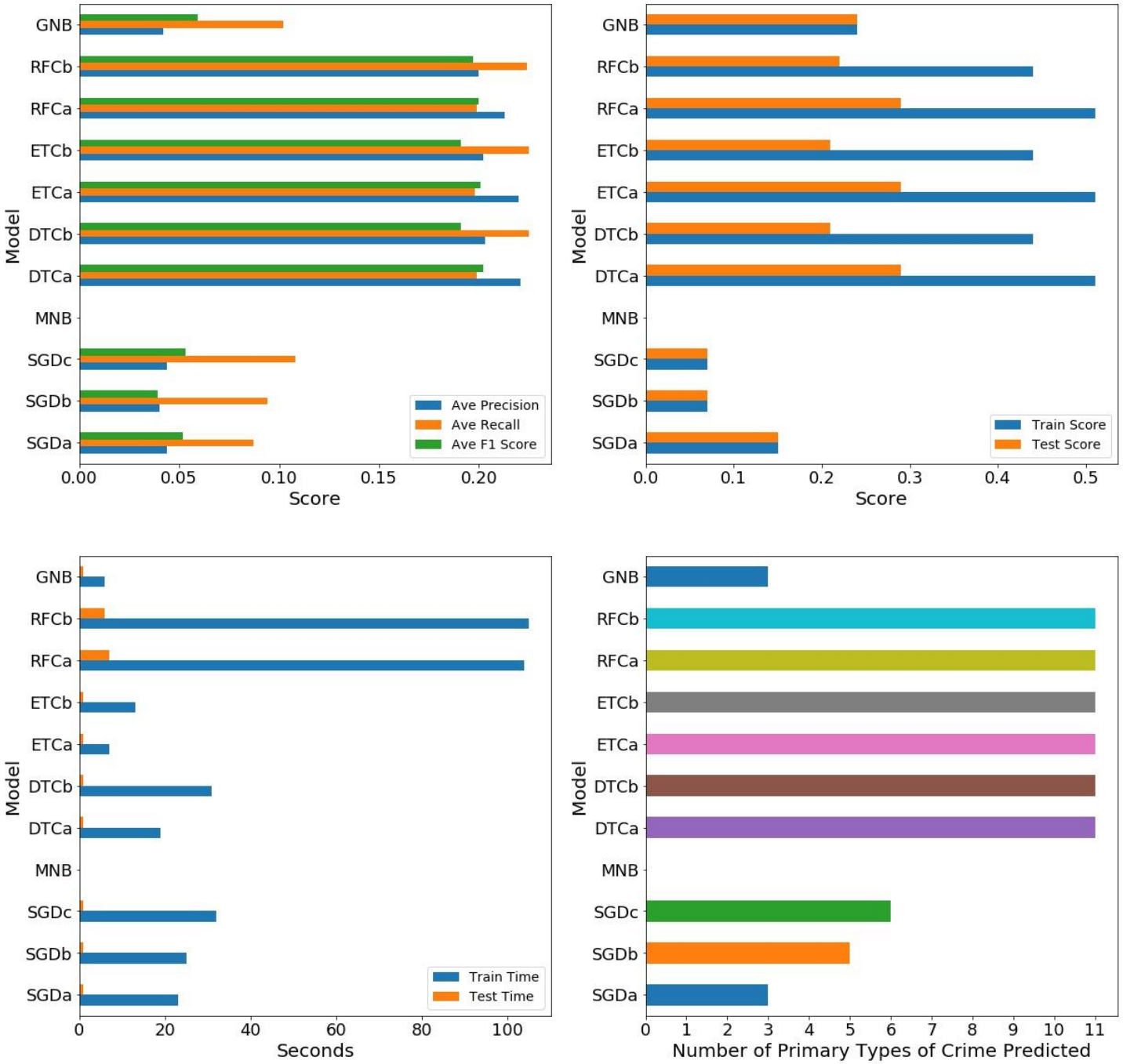

Figure 50: Model results for a sample of 3 million crime reports with the number of features reduced to 2 using IncrementalPCA. * MultinomialNB could not be used as there were negative features.

Per Figure 50, the amount of time it took to train each model was significantly reduced. However, this came at the cost of accuracy. Not all of the models predicted all 11 primary types of crime. All of the tree classifiers did the best when looking at the average F1 score, recall, and precision and they did not overfit the data. But overall, the scores obtained with just reducing the dimensionality were lower than those obtained from using scaled data.

The original sample data with dummy variables was first scaled using MinMaxScaler and then transformed using IncrementalPCA. Per Figure 51, scaling significantly reduced the difference in the explained variance for each feature. I chose to reduce the dimensionality of the data to 9 components.

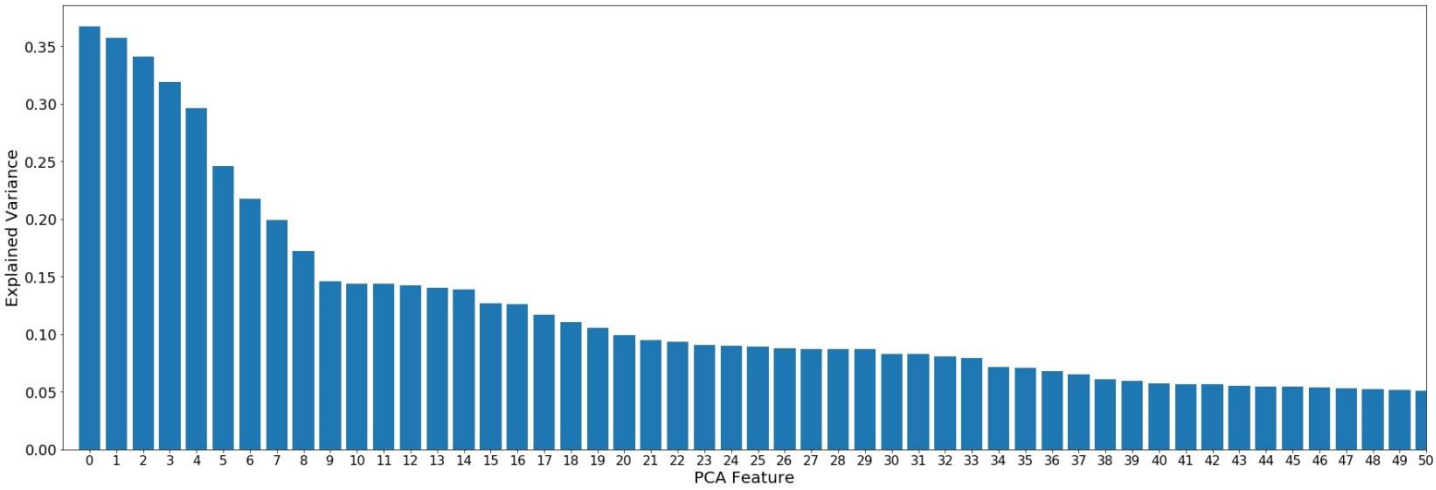

Figure 51: Explained variance vs PCA feature for a sample of 3 million crime reports first scaled with MinMaxScaler. In total, there are 638 features.

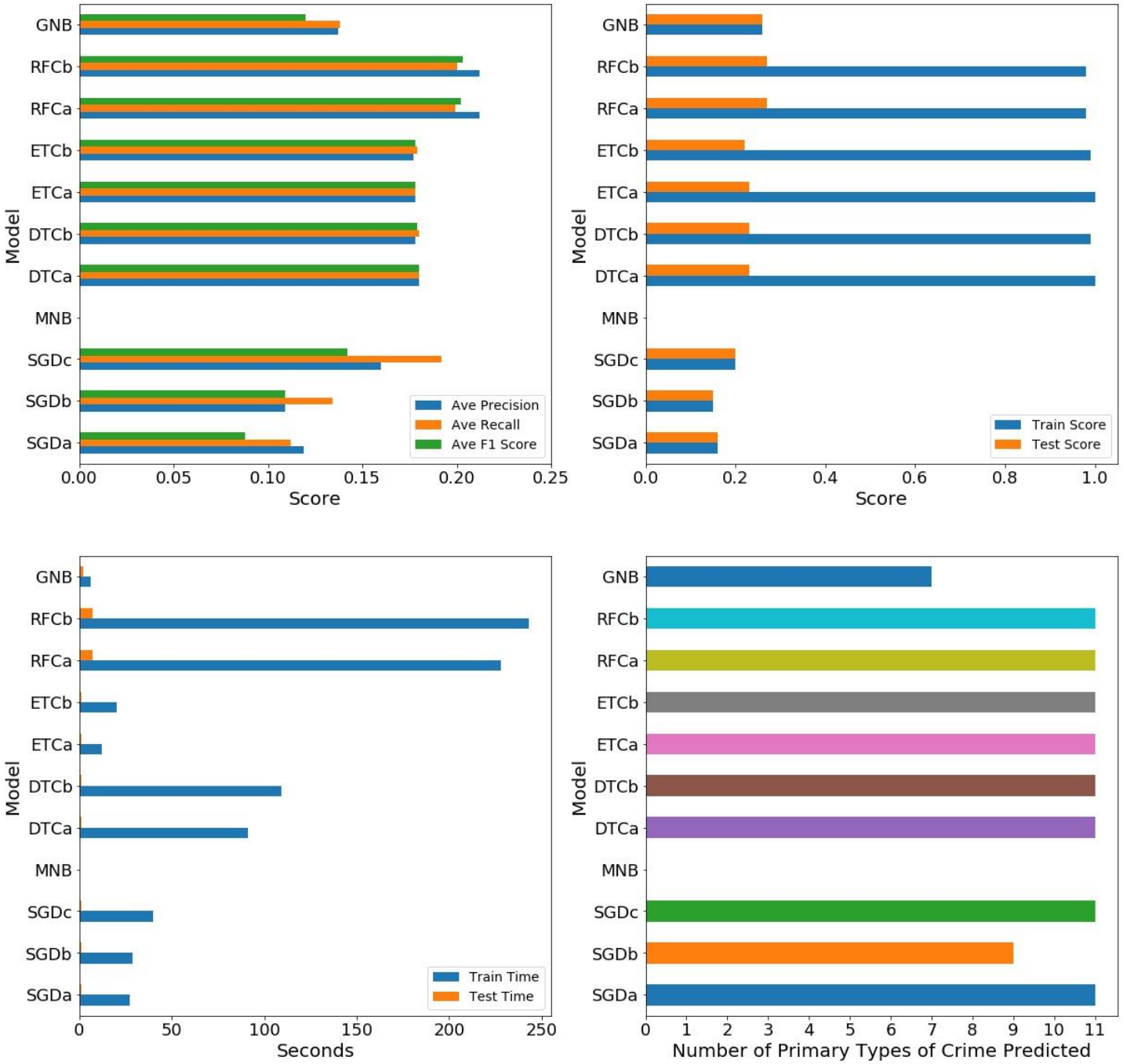

Figure 52: Model results for a sample of 3 million crime reports first scaled with MinMaxScaler and then reduced to 9 features using IncrementalPCA. * MultinomialNB could not be used as there were negative features.

Per Figure 52, the amount of time it took to train each model was reduced but the accuracy was reduced as well. GNB and SGDb did not predict all 11 primary types of crime. RFCa/b performed the best, but overfit the data. But overall, the scores obtained with scaling the data and reducing the dimensionality were lower than those obtained from using scaled data.

In the following section, I decided to explore output from the SGDc using scaled data as it had the highest average F1 score, recall, and precision. It also did not overfit the data and predicted all 11 primary types of crime.

Features of Importance

I used scaled training data on the SGD Classifier with the class weight set to 'balanced' and the loss set to 'log' and extracted the coefficients for each feature and primary type of crime. I then explored the feature with the largest, positive coefficient for the top 4 most frequent crimes: theft, battery, criminal damage, and narcotics.

In predicting crimes involving theft (Figure 53), a location on CHA (Chicago Housing Authority) grounds is the strongest negative indicator, meaning crimes involving theft are less likely to occur there. Residents on CHA grounds all having lower incomes could explain why theft isn't as prevalent; it would be uncommon that another resident would have something much nicer that another resident or someone who lived elsewhere would want.

A location in a department store is the strongest positive indicator, meaning crimes involving theft are more likely to occur there. It could also mean that department stores are more likely to report theft. Per Figure 54, theft is by far the most common reported crime in a department store.

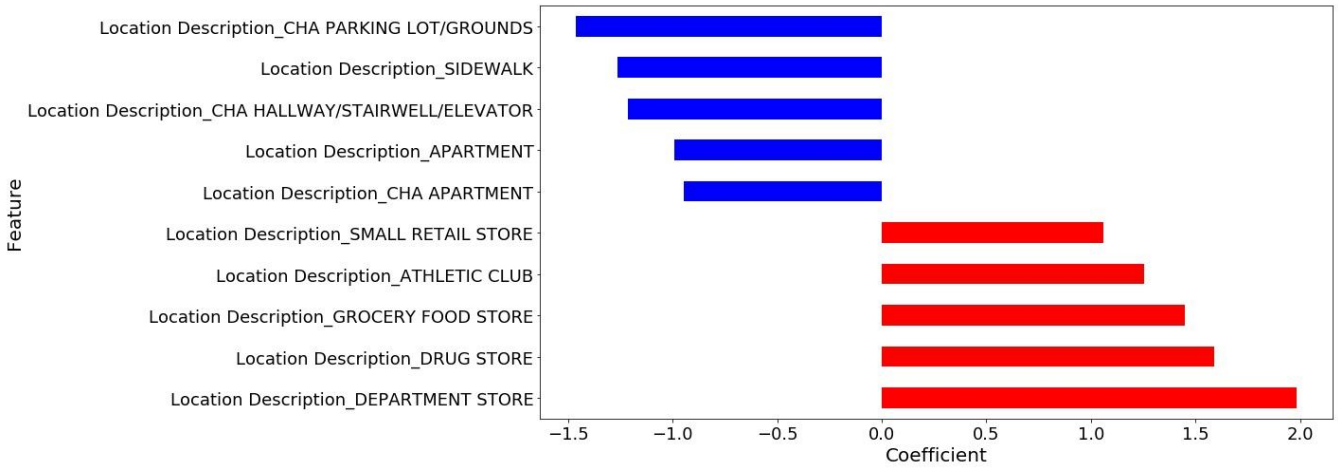

Figure 53: Coefficients of the top 10 features (in magnitude) for crimes involving theft.

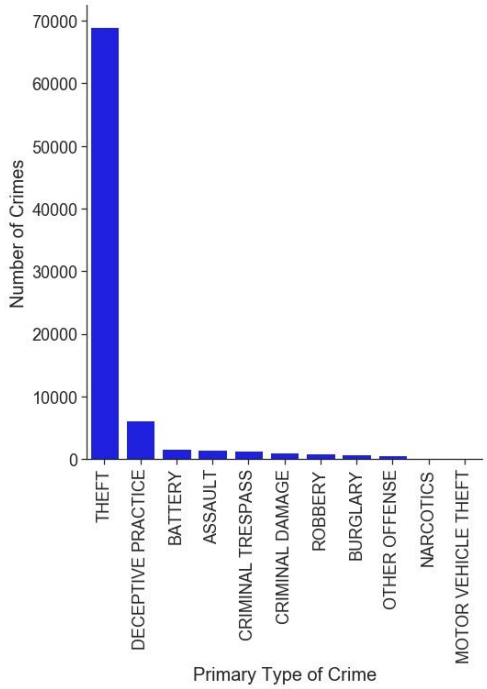

Figure 54: Breakdown of the number of crimes that occurred within a department store by primary type of crime.

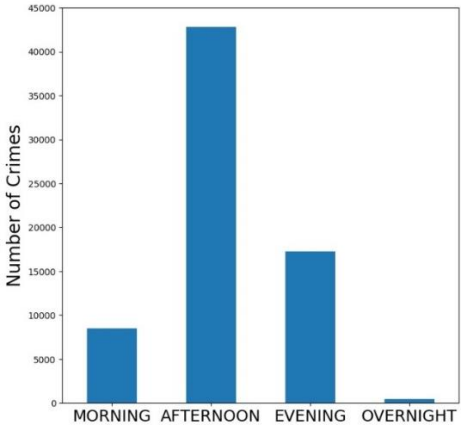

Figure 55: Number of crimes involving theft that occurred within a department store per time of day.

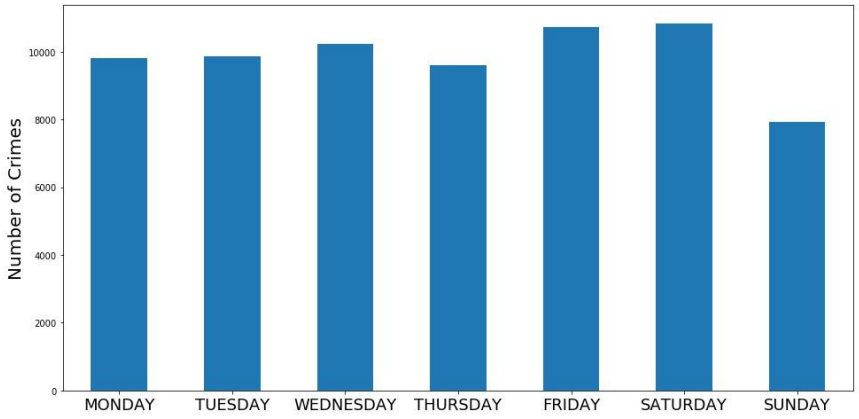

Figure 56: Number of crimes involving theft that occurred within a department store per day of the week.

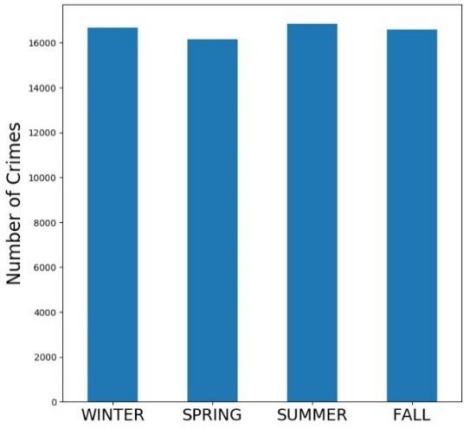

Figure 57: Number of crimes involving theft that occurred within a department store per season.

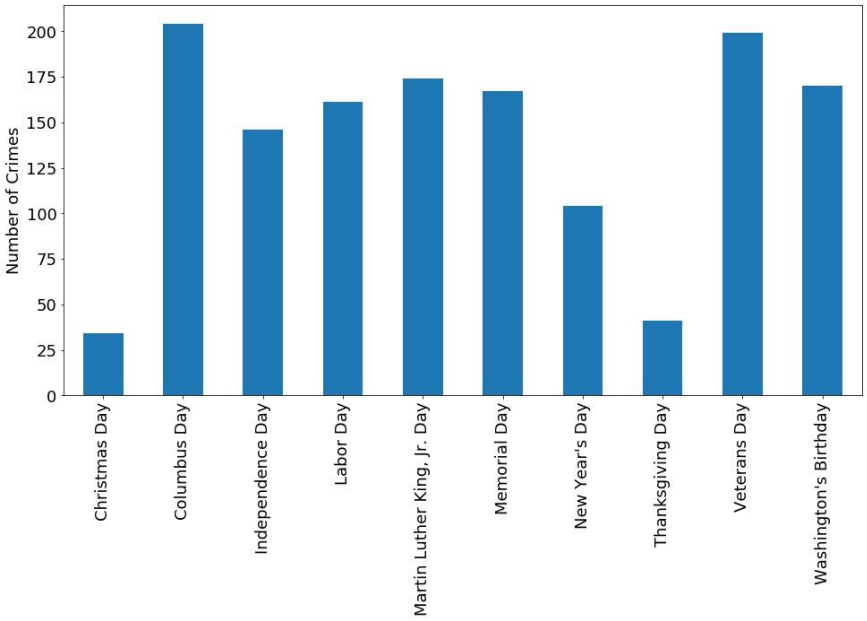

Figure 58: Number of crimes involving theft that occurred within a department store per federal holiday.

The number of crimes involving theft in a department store:

- peaks in the afternoon and is at a minimum overnight (Figure 55)
- is generally steady from Monday through Thursday and then increases slightly on Friday and Saturday before reaching a minimum on Sunday (Figure 56)
- does not change with the season (Figure 57)
- is at a minimum on Christmas and Thanksgiving Day and at a maximum on Columbus Day and Veterans Day (Figure 58)

It is possible that the number of thefts could be related to the number of people visiting the department store. More people are likely to visit the department store in the afternoon, on Fridays/Saturdays and on holidays where they have off from work/school and the stores are still open. Overnight, most department stores are closed, on Sundays they tend to close earlier, and on Christmas/Thanksgiving Day they are closed, thus greatly limiting the number of people who have access to the store and also limiting the number of people available to actually report theft.

Per Figure 59, there is a high concentration of crimes involving theft in department stores close to the center of Chicago. This could be because there is a higher concentration of department stores close to the city center. When comparing the north and south side of the city, the north side generally has a higher concentration of crimes. Communities 32 and 8 have the highest number of crimes involving theft in department stores (Figure 60). This makes sense as these communities are located close to the city center.

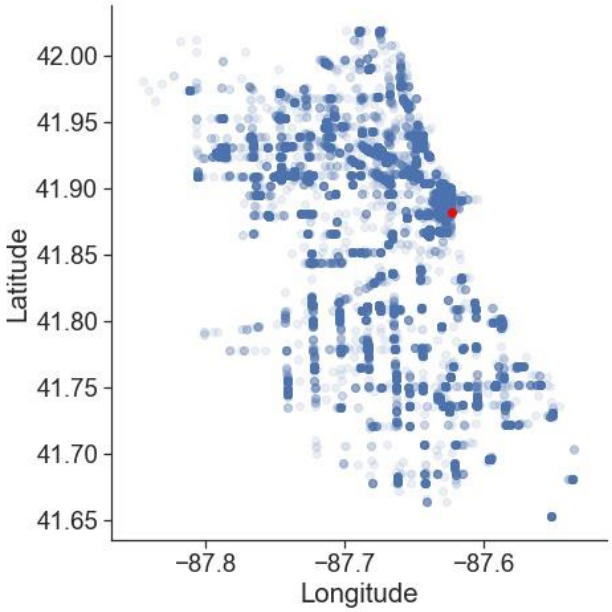

Figure 59: Spatial distribution of crimes that involved theft and occurred within a department store. The red dot indicates the center of Chicago.

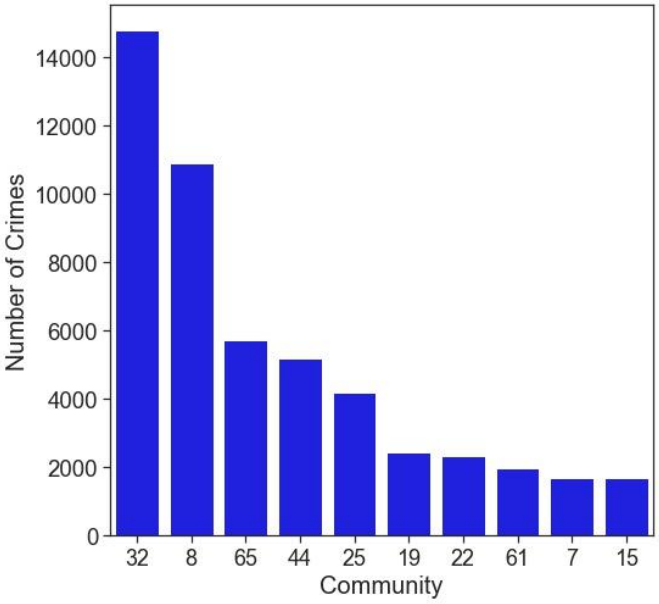

Figure 60: Top 10 communities with the highest number of crimes that involved theft and occurred within a department store.

In predicting crimes involving battery (Figure 61), a location in a residential garage is the strongest negative indicator, meaning crimes involving battery are less likely there. It would make sense that other crimes such as motor vehicle theft would more often occur in a residential garage.

A location in a public school building is the strongest positive indicator, meaning crimes involving battery are more likely there. Per Figure 62, battery is the most common crime in a public school building, followed by theft then assault. This is an interesting outcome as I would have expected a different location such as an apartment or the sidewalk to be the strongest indicator as more crimes involving battery would likely occur there. But it could be that apartments and sidewalks have large amounts of numerous crimes, so it would be hard to predict which crime would most likely occur.

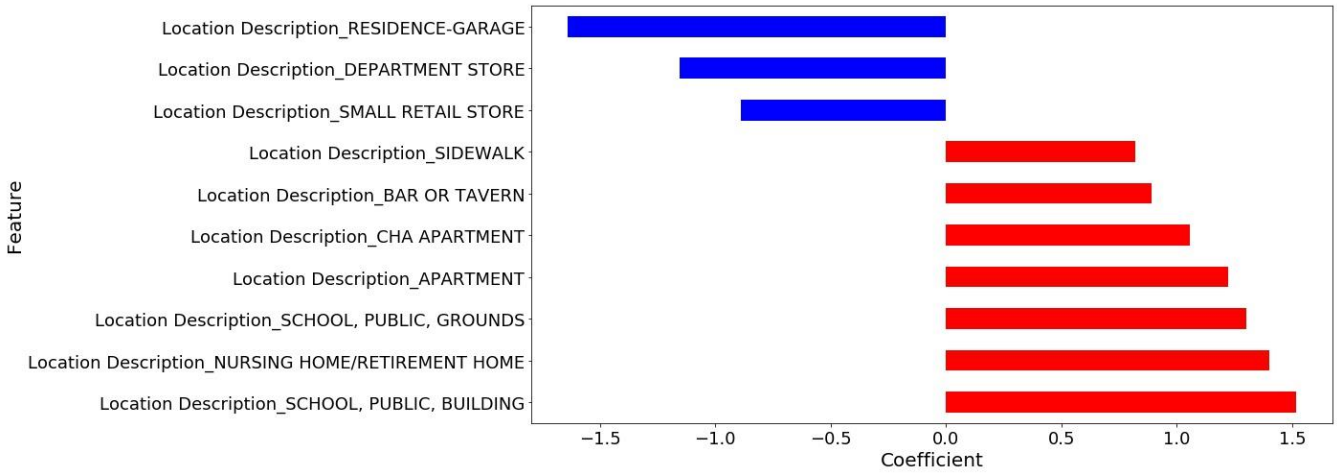

Figure 61: Coefficients of the top 10 features (in magnitude) for crimes involving battery.

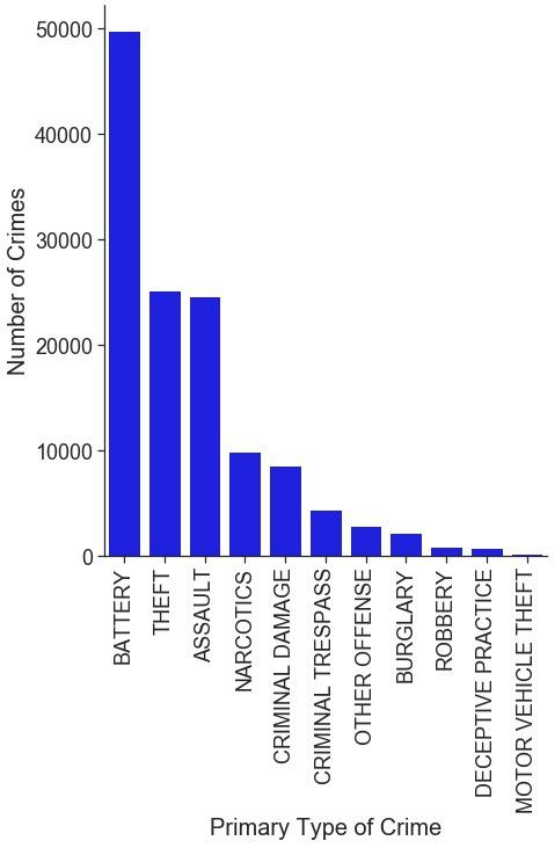

Figure 62: Breakdown of the number of crimes that occurred within a public school building by primary type of crime.

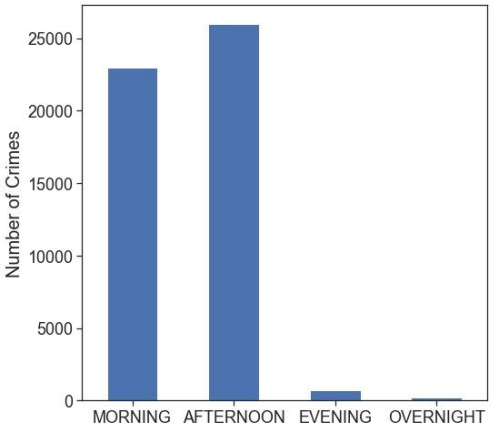

Figure 63: Number of crimes involving battery that occurred within a public school building per time of day.

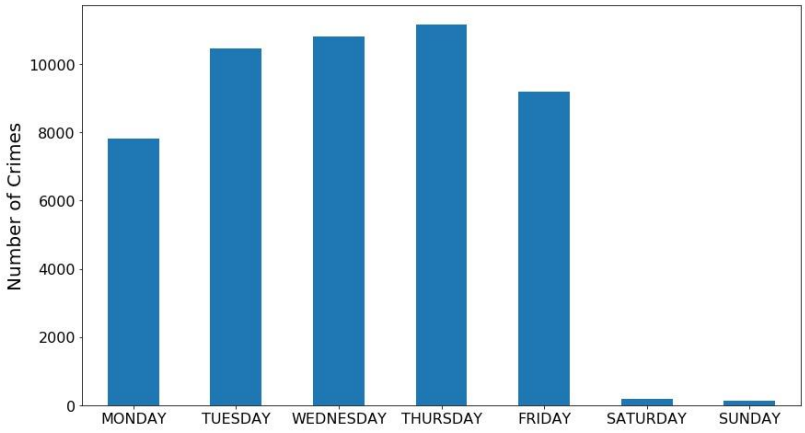

Figure 64: Number of crimes involving battery that occurred within a public school building per day of the week.

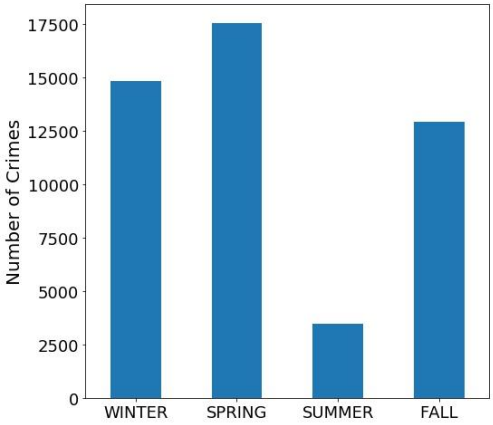

Figure 65: Number of crimes involving battery that occurred within a public school building per season.

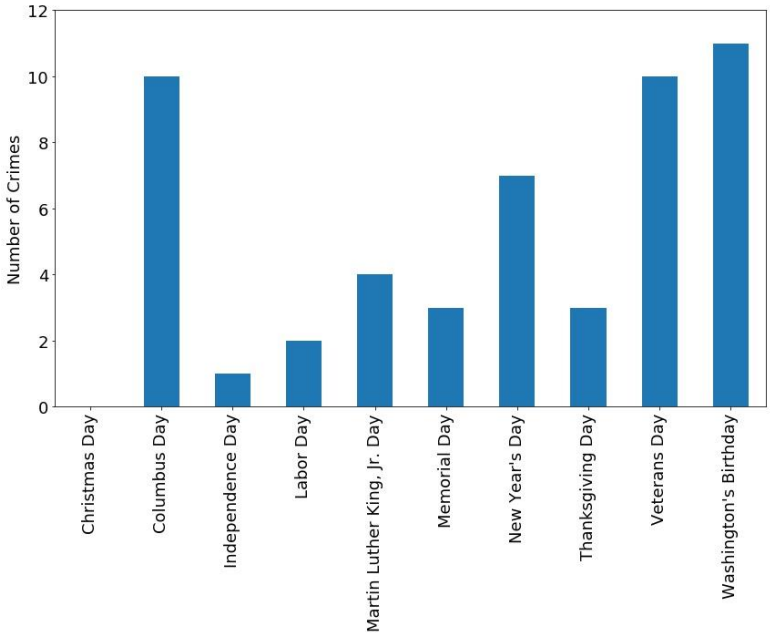

Figure 66: Number of crimes involving battery that occurred within a public school building per federal holiday.

The number of crimes involving battery in a public school building:

- is the highest during the morning and afternoon and lowest during the evening and overnight (Figure 63)
- is the highest from Tuesday through Thursday and the lowest over the weekend (Figure 64)
- is the highest during winter, spring, and fall and the lowest during the summer (Figure 65)
- is the highest on Columbus Day, Veterans Day, and Washington's Birthday and the lowest on Independence Day and Labor Day (Figure 66)

It appears that there could be a relationship between the number of students present at school and the occurrence of crimes involving battery. More students are present during the morning/afternoon, on weekdays, and in all seasons but summer. There may not be a significant relationship between the number of crimes involving battery and the federal holiday as relatively few crimes occurred (or were failed to be reported) on these days.

Per Figure 67, there is a low concentration of crimes involving battery in a public school building close to the center of Chicago. There is a higher concentration of crimes to the west of the city center near 41.85°N-41.9°N and 87.65°W-87.75°W. The crimes seem to be more widespread across the southern half of Chicago. Community 25 has the highest number of crimes involving battery in a public school building. This community is located along a similar latitude as the city center on the western edge of Chicago.

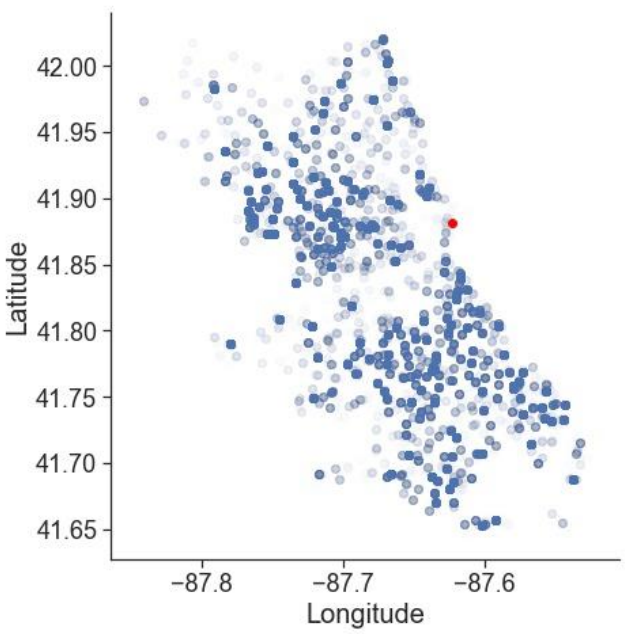

Figure 67: Spatial distribution of crimes that involved battery and occurred within a public school building. The red dot indicates the center of Chicago.

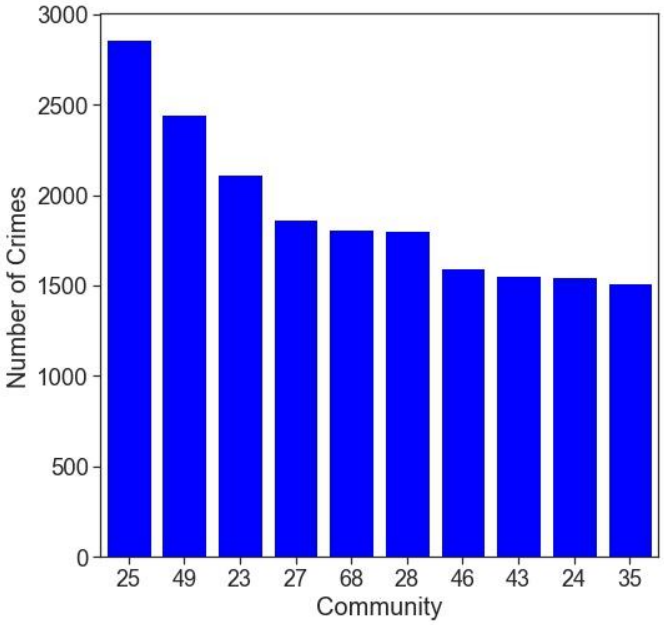

Figure 68: Top 10 communities with the highest number of crimes that involved battery and occurred within a public school building.

In predicting crimes involving criminal damage (Figure 69), a location on a sidewalk is the strongest negative indicator, meaning crimes involving criminal damage are less likely to occur there. This makes sense as there is not much damage one can do or even be motivated to do on a sidewalk.

A location on a residential driveway is the strongest positive indicator, meaning crimes involving criminal damage are more likely to occur there. Per Figure 70, criminal damage is the most common crime to occur on a residential driveway with theft being a close a second. This makes sense as an individual's driveway is easily accessible and could be vandalized.

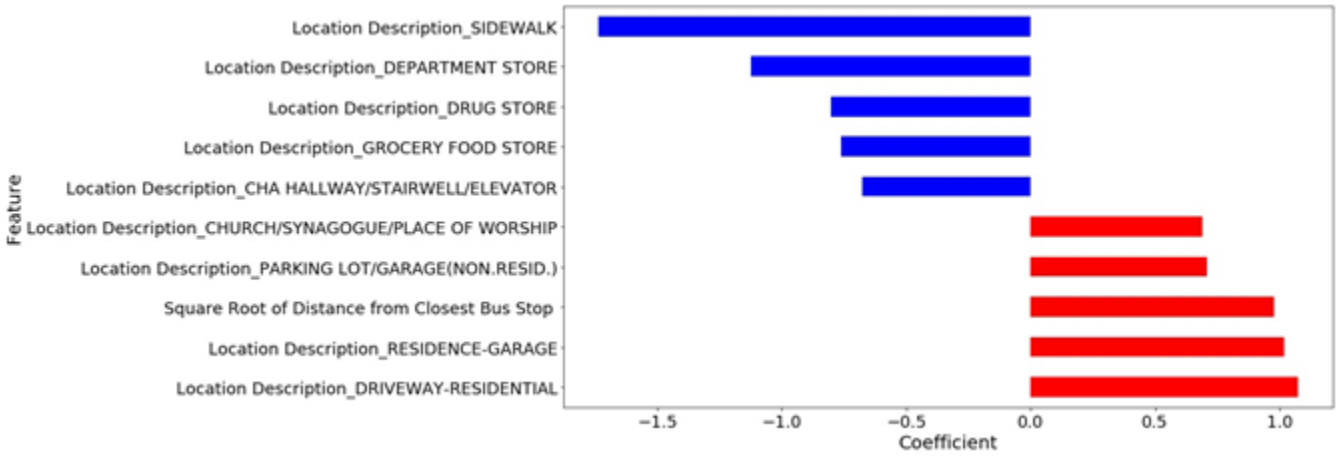

Figure 69: Coefficients of the top 10 features (in magnitude) for crimes involving criminal damage.

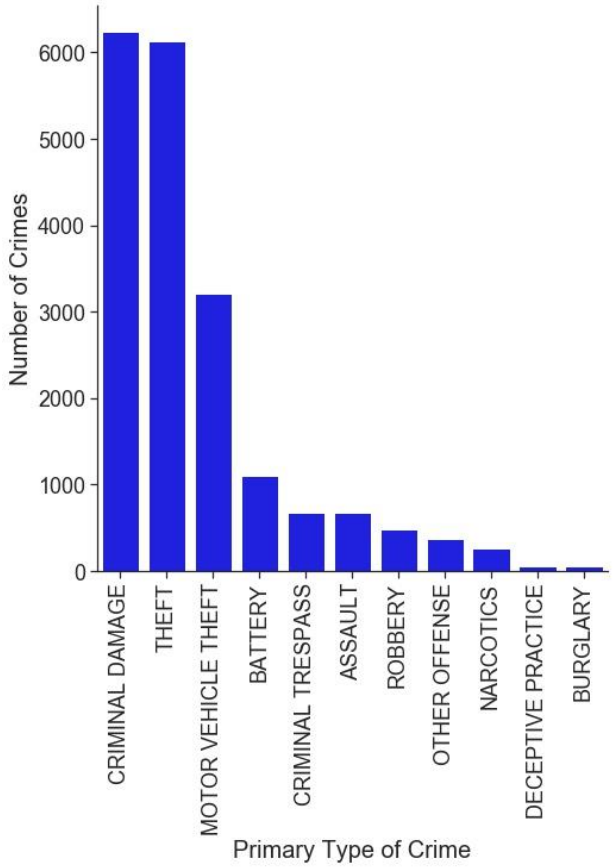

Figure 70: Breakdown of the number of crimes that occurred on a residential driveway by primary type of crime.

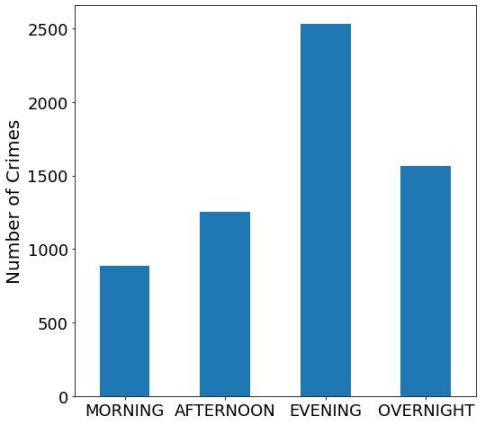

Figure 71: Number of crimes involving criminal damage that occurred on a residential driveway per time of day.

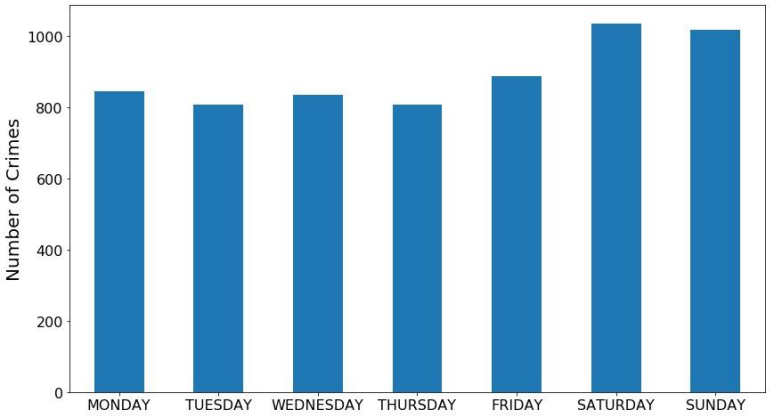

Figure 72: Number of crimes involving criminal damage that occurred on a residential driveway per day of the week.

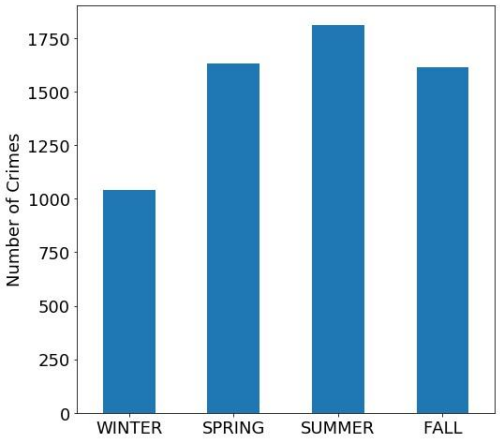

Figure 73: Number of crimes involving criminal damage that occurred on a residential driveway per season.

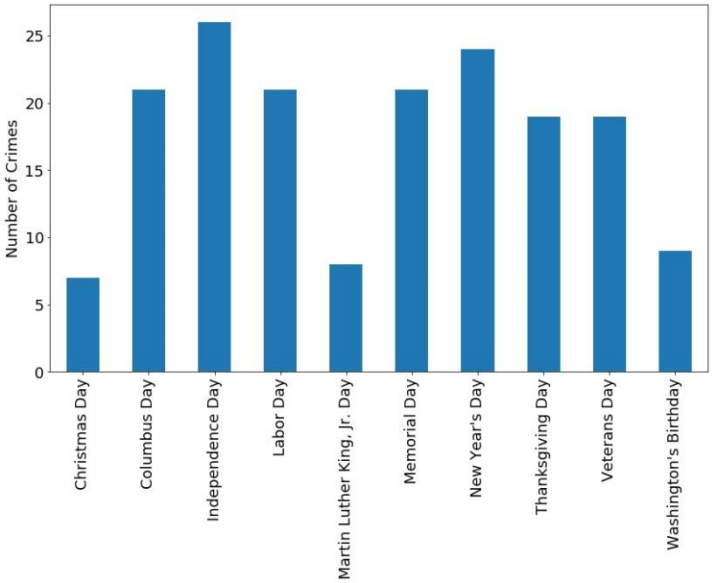

Figure 74: Number of crimes involving criminal damage that occurred on a residential driveway per federal holiday.

The number of crimes involving criminal damage on a residential driveway:

- peaks in the evening and is the lowest during the morning (Figure 71)
- increases slightly over the weekend (Figure 72)
- is the highest during the spring, summer, and fall (Figure 73)
- is the highest on Independence Day and New Year's Day and the lowest on Christmas Day, Martin Luther King Jr. Day, and Washington's Birthday (Figure 74)

It appears that there could be a relationship between the number of people outside their homes and the occurrence of crimes involving criminal damage. More people are likely to be outside after work in the evening, on the weekends, and when the weather isn't too cold. There may not be a significant relationship between the number of crimes involving criminal damage and federal holidays as relatively few crimes occurred on these days.

Per Figure 75, crimes involving criminal damage on residential driveways is widespread across Chicago, except for near the city center. This could be because there aren't many residences that have actual driveways near the city center. There are locally higher concentrations on the far southern edges of the city. Community 25 has the highest number of crimes involving criminal damage on residential driveways (Figure 76). This community is located along a similar latitude as the city center on the western edge of Chicago. Overall, the difference in the number of crimes per community is not significant.

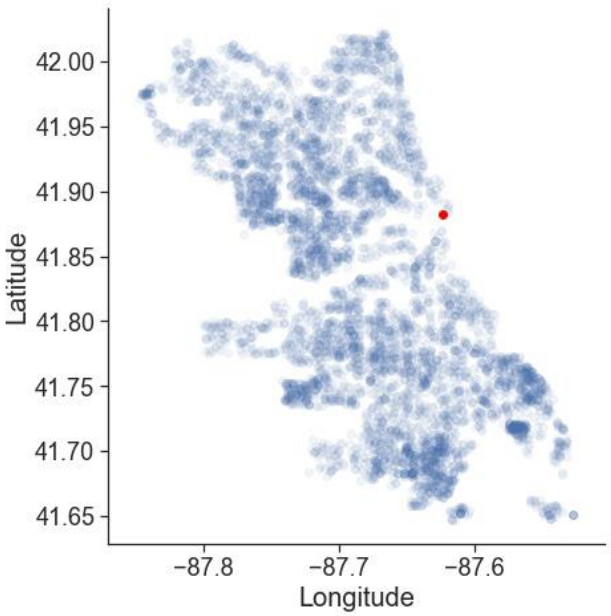

Figure 75: Spatial distribution of crimes that involved criminal damage and occurred on a residential driveway. The red dot indicates the center of Chicago.

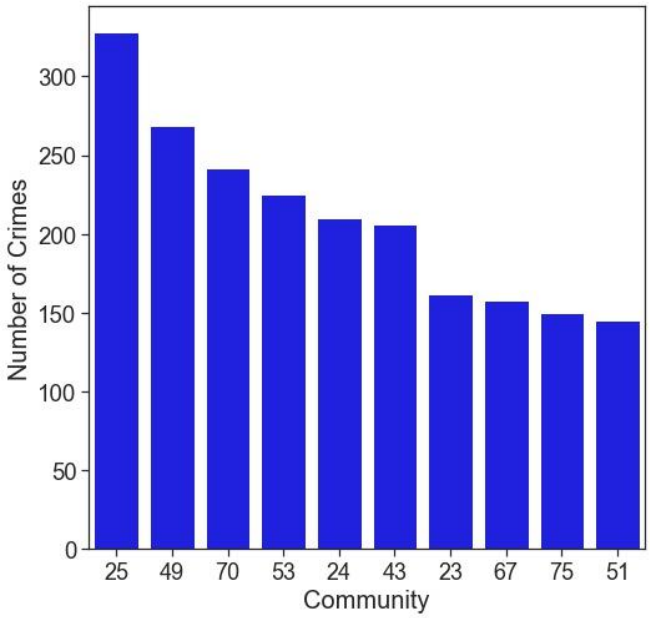

Figure 76: Top 10 communities with the highest number of crimes that involved criminal damage and occurred on a residential driveway.

In predicting crimes involving narcotics (Figure 77), a location in a small retail store is the strongest negative indicator, meaning crimes involving narcotics are less likely to occur there. This makes sense as there would be too many people around in a small retail store in order to attain the privacy needed to deal/use narcotics.

A location on CHA grounds is the strongest positive indicator, meaning crimes involving narcotics are more likely to occur there. Per Figure 78, narcotics is the most common crime to occur on CHA grounds followed by criminal trespassing then battery. As you tend to have more low-income people on CHA grounds, it makes sense that crimes involving narcotics would be more prevalent there.

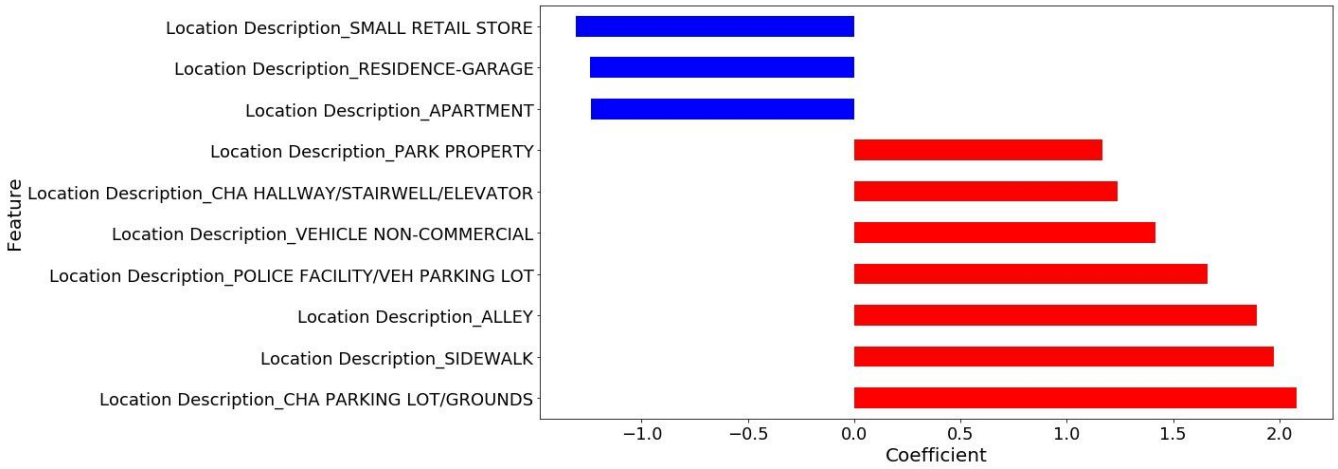

Figure 77: Coefficients of the top 10 features (in magnitude) for crimes involving narcotics.

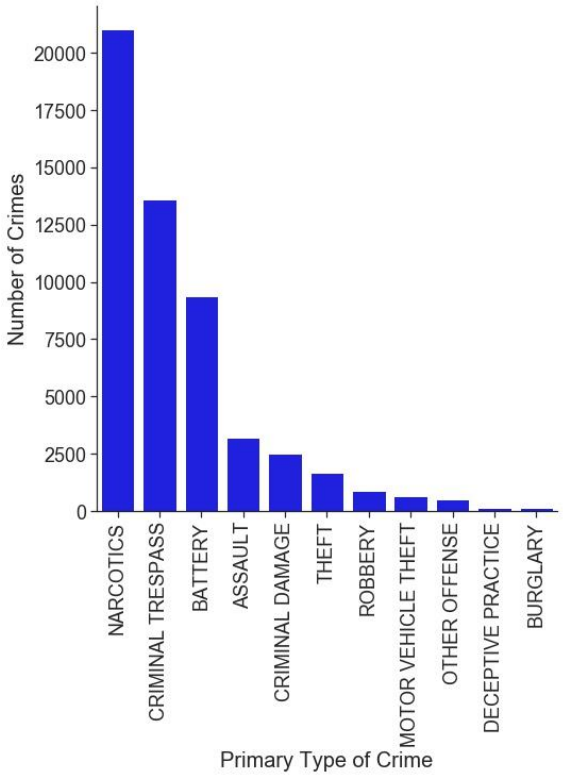

Figure 78: Breakdown of the number of crimes that occurred on CHA grounds by primary type of crime.

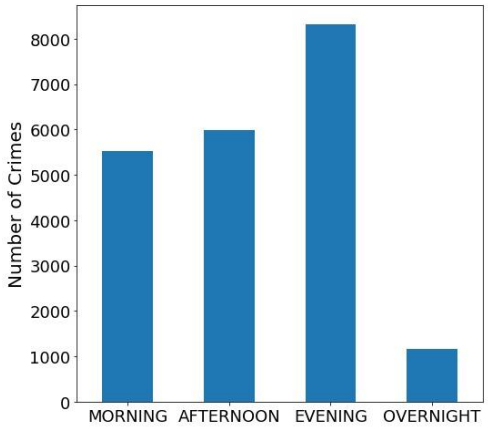

Figure 79: Number of crimes involving narcotics that occurred on CHA grounds per time of day.

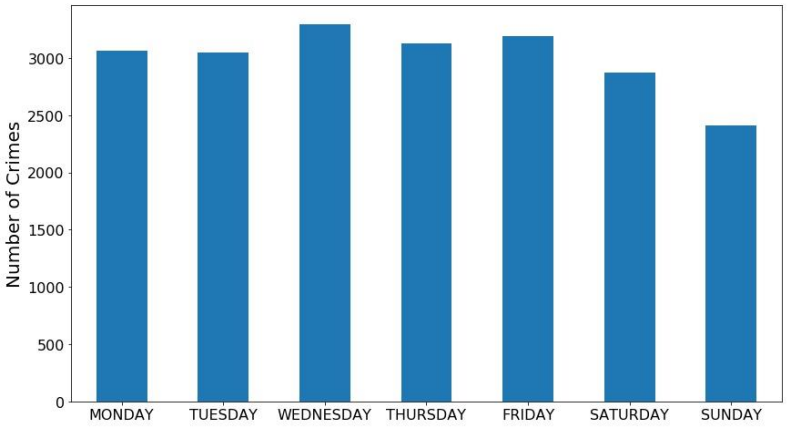

Figure 80: Number of crimes involving narcotics that occurred on CHA grounds per day of the week.

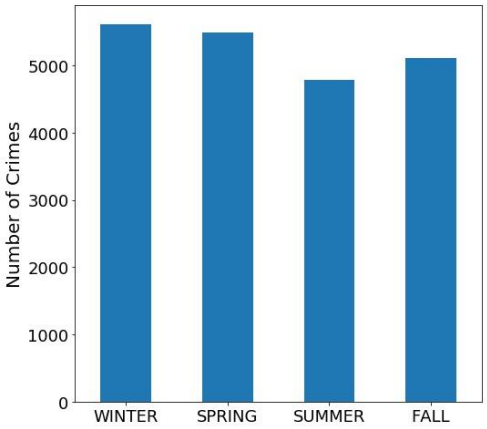

Figure 81: Number of crimes involving narcotics that occurred on CHA grounds per season.

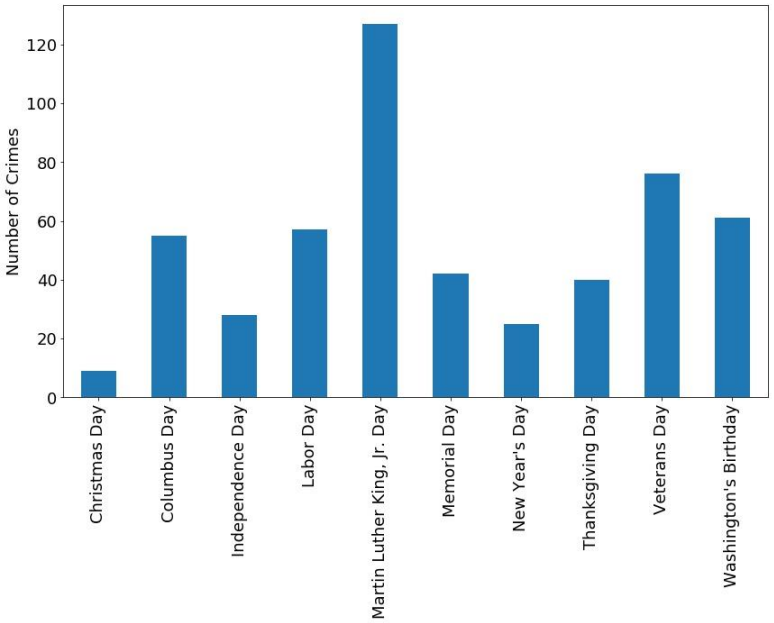

Figure 82: Number of crimes involving narcotics that occurred on CHA grounds per federal holiday.

The number of crimes involving narcotics on CHA grounds:

- peaks in the evening and is the lowest overnight (Figure 79)
- is steady on weekdays and decreases over the weekend, reaching a minimum on Sunday (Figure 80)
- decreases slightly during the summer (Figure 81)
- is at a maximum on Martin Luther King Jr. Day and a minimum on Christmas Day (Figure 82)

It could possibly be that the number of crimes involving narcotics on CHA grounds is higher during the evening when more people are back home from work. It is likely that there are more crimes involving narcotics occurring overnight than noted just because there would be fewer people around to actually report them.

It is interesting that the number of crimes decreases over the weekend, when more people are likely to be home. Perhaps it could be that spouses/kids are at home on the weekend, so some people are less apt to deal/use narcotics. That could explain why there is a dip in the number of crimes in the summer; kids are home more often due to summer break.

It is uncertain if the increase in the number of crimes on Martin Luther King Jr. Day is a coincidence because overall, there are much fewer crimes occurring on federal holidays. It could be possible that police are out in larger numbers on this day on CHA grounds to prevent riots or demonstrations that would disrupt public peace. Further investigation and hypothesis testing may be needed to confirm if there is a relationship.

Per Figure 83, there are high concentrations of crimes involving narcotics on CHA grounds to the north, west, and especially south of the city center. There is also a smaller cluster of crimes on the southern edge of Chicago. There are relatively low concentrations of crime near the city center, possibly because there may not be as many CHA locations. Community 35 has the highest number of crimes involving narcotics on CHA grounds (Figure 84). This community is located to the south of the city center.

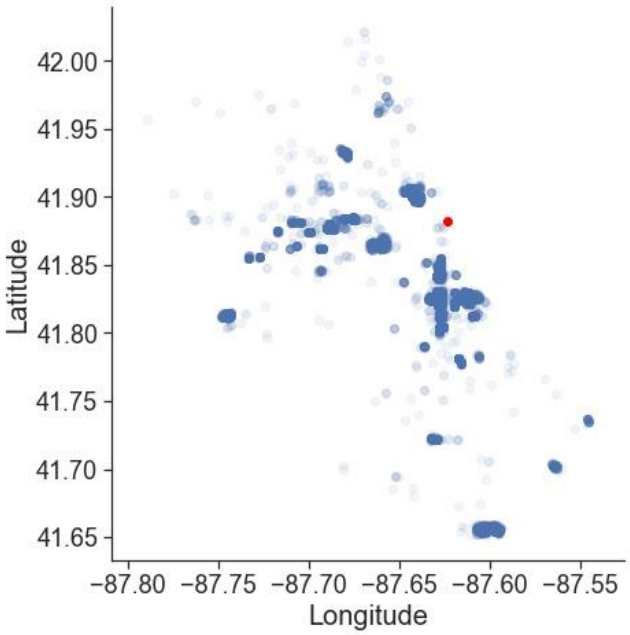

Figure 83: Spatial distribution of crimes that involved narcotics and occurred on CHA grounds. The red dot indicates the center of Chicago.

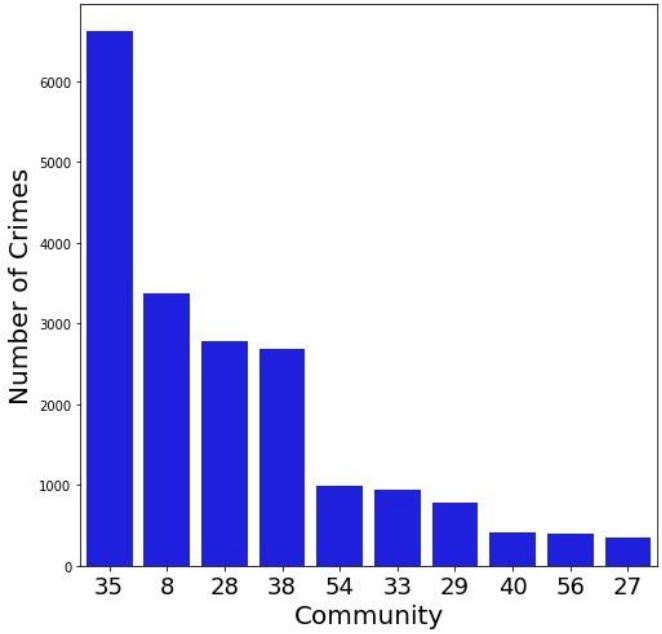

Figure 84: Top 10 communities with the highest number of crimes that involved narcotics and occurred on CHA grounds.

When looking at features individually for each type of crime, the location description was the most important indicator for the reported crime type. Per Figure 85, the 4 locations with the most crime are street, residence, apartment, and sidewalk.

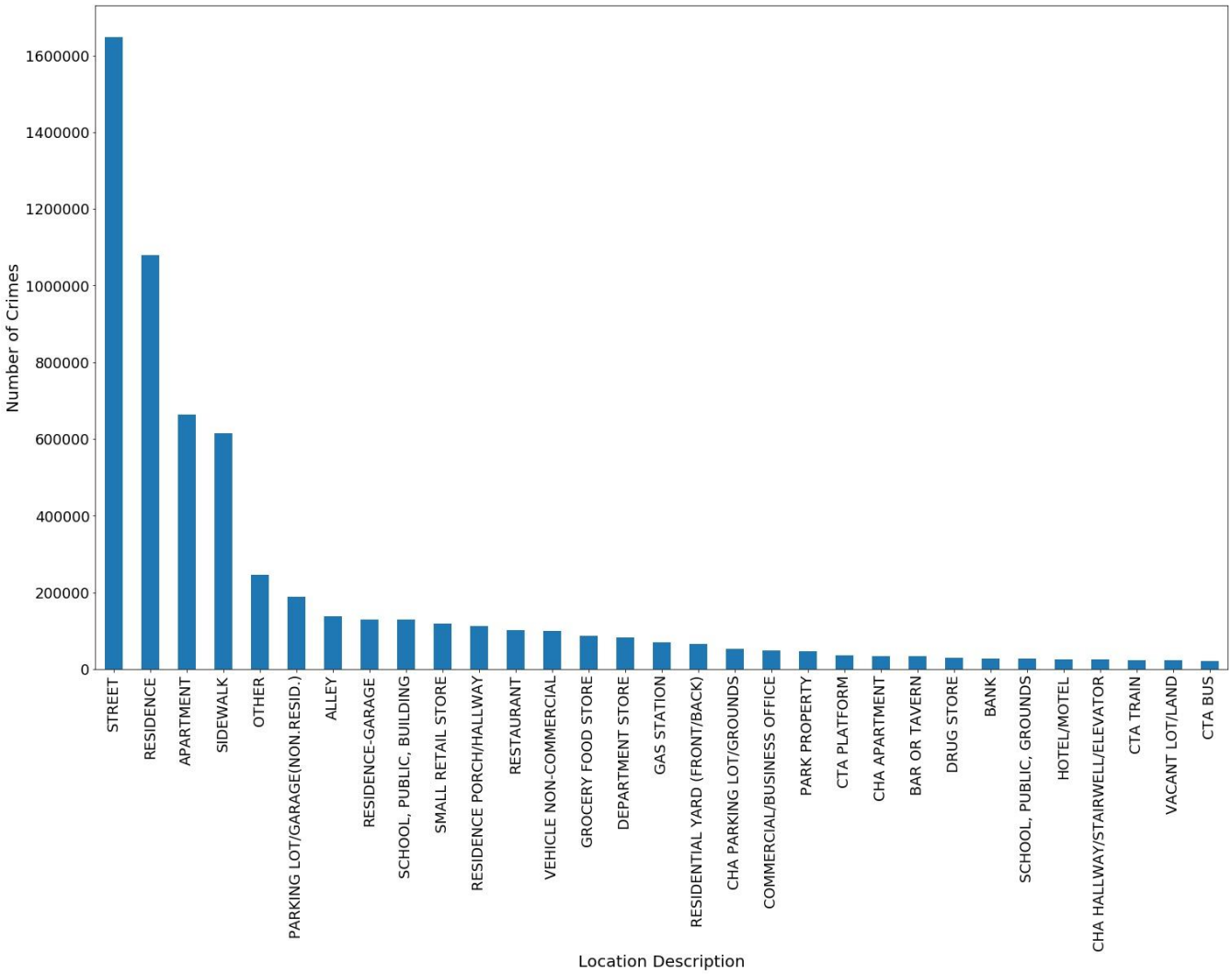

Figure 85: Number of crimes for a sample of location descriptions. There is a total of 109 location descriptions.

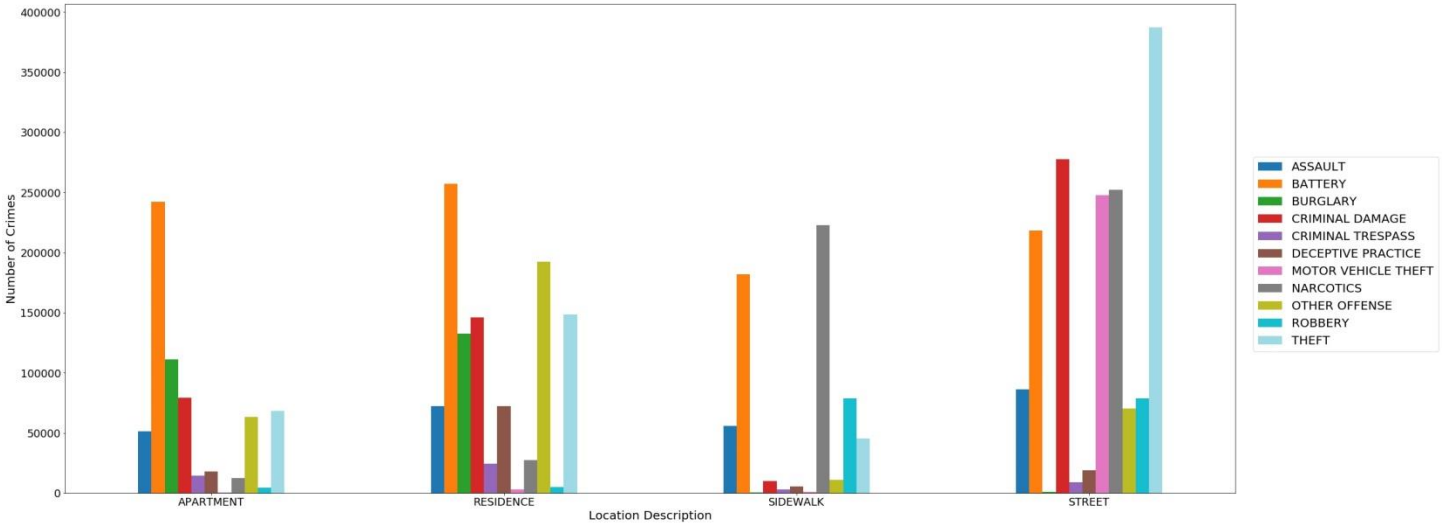

Figure 86: Number of each primary type of crime for the top 4 location descriptions.

The breakdown of crimes for these locations is shown in Figure 86. The street has the highest number of crimes involving theft, criminal damage, and motor vehicle theft. The number of crimes involving burglar is the highest in an apartment or residence. Crimes involving deceptive practice more likely occur in a residence. Crimes involving narcotics and robbery occur more on the sidewalk and street.

In order to see the overall influence of each feature on the occurrence of all 11 studied primary types of crime, I averaged the coefficients for each feature across all 11 primary types of crime and then averaged the coefficients for all of the dummy variables of a feature. Figure 87 shows the coefficients of the top 10 features (in magnitude). On average, the square root of the distance from the closest bus stop is the strongest negative indicator for the 11 primary types of crime. This means that this feature is generally less able to discriminate between the 11 primary types of crime. Per Figure 88, the distributions for most of the crime types are multimodal and there aren't significant variations for each one. For example, several of the crime types show a peak in the occurrence of crimes at the same square root of the distance from the closest bus stop. Per Figure 69, this feature was the third strongest indicator for criminal damage. It could be that this feature was only a good positive indicator for very few crime types, therefore overall it would have a negative coefficient.

Spatial features that are on average the strongest positive indicators are: X coordinate (longitude), distance from the city center of Chicago, and Y coordinate (latitude). Temporal features that are on average the strongest positive indicators are: day type (weekday/weekend), is holiday (no holiday/federal holiday), season, time of day, day of the week, and holiday (which federal holiday). All of these features are more strongly able to discriminate between the 11 primary types of crime; meaning they vary significantly across most of the crime types.

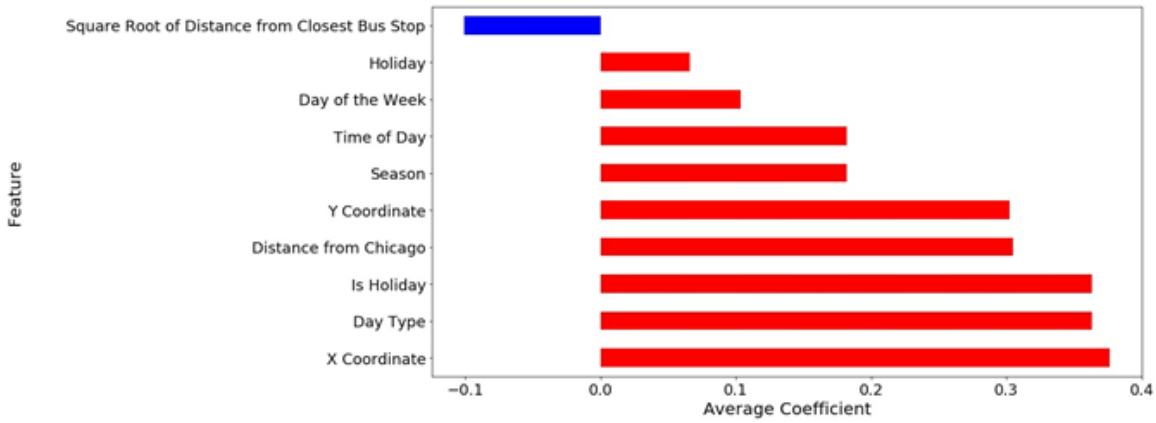

Figure 87: Average coefficients of the top 10 features (in magnitude) for all dummy features and crime types.

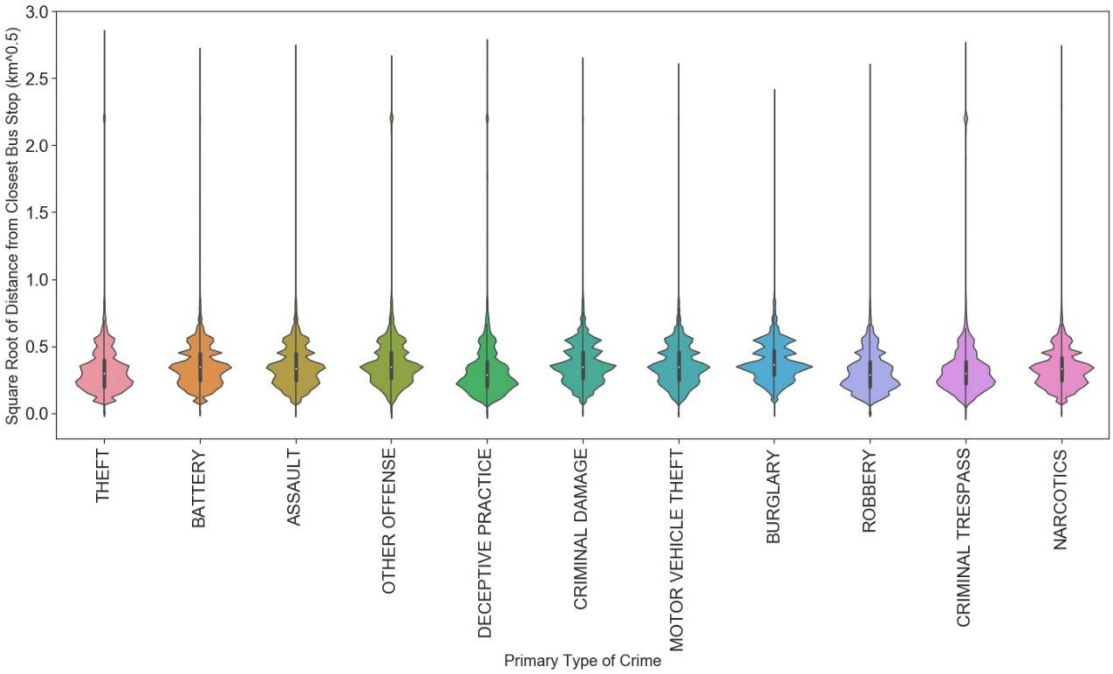

Figure 88: Distribution of square root of distance from closest bus stop for each primary type of crime.

In Figure 89, the distributions of the longitude for each crime type have fat tails to the left (west). 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). The distribution for narcotics appears to be somewhat multimodal with an especially pronounced increase in the number of crimes just west of -87.7°.

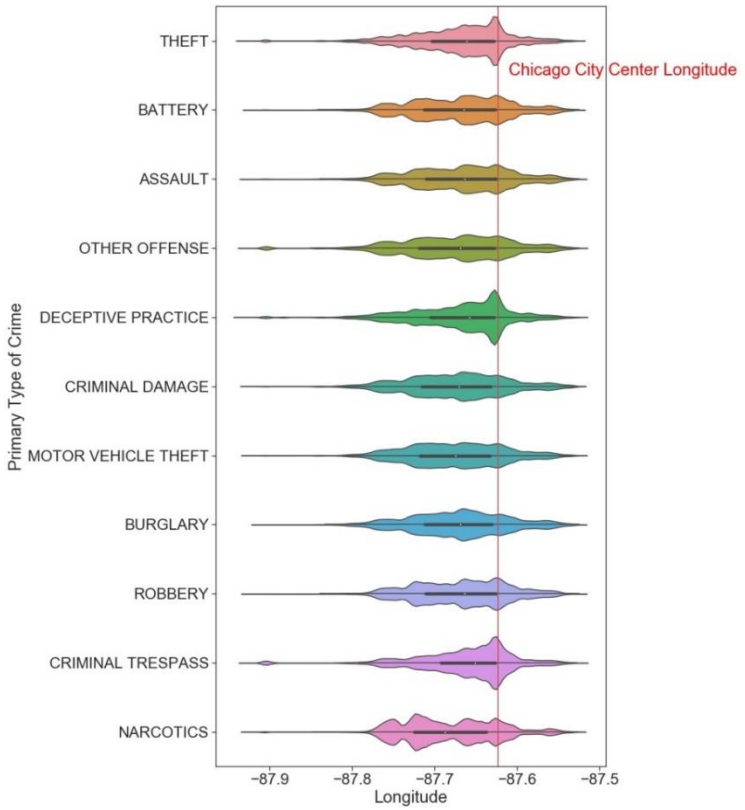

Figure 89: Distribution of crimes across longitude for each primary type of crime.

Figure 90 shows that each primary type of crime generally has a bimodal distribution of latitudes. Crimes involving theft, deceptive practice, 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.

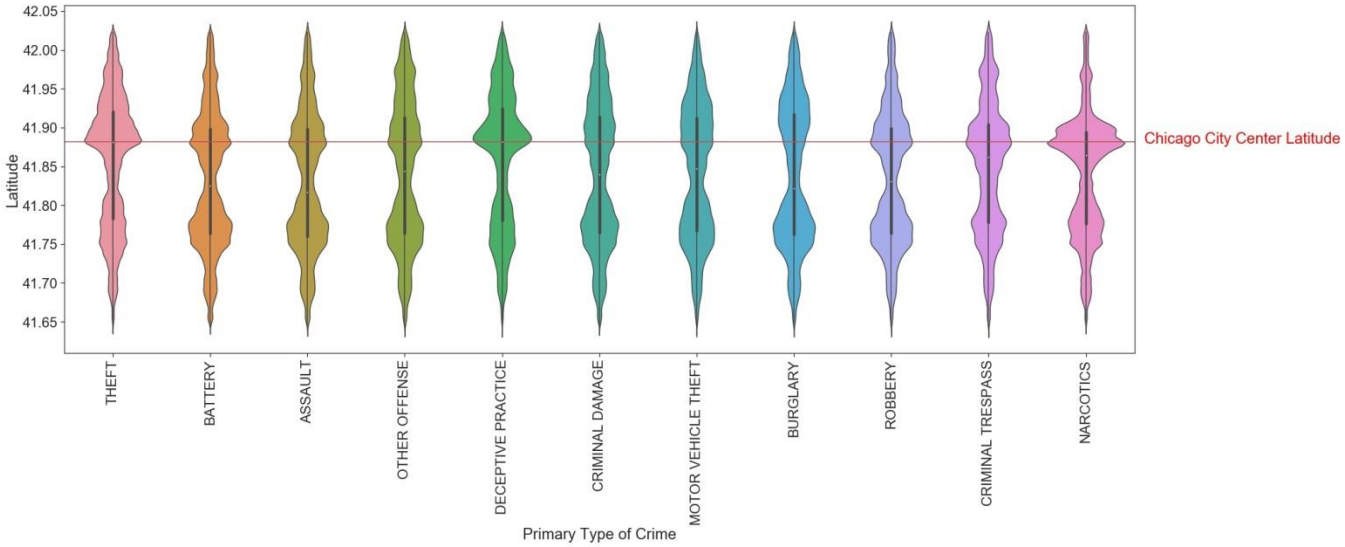

Figure 90: Distribution of crimes across latitude for each primary type of crime.

Figure 91 confirms that there is an increase in the number of crimes involving theft, deceptive practice, and criminal trespassing closer to the city center. For the distributions of latitude, crimes involving narcotics has an increase near the latitude of the city center, but there actually is no increase per Figure 91. It could be that there is an increase in crimes involving narcotics to the west of the city center.

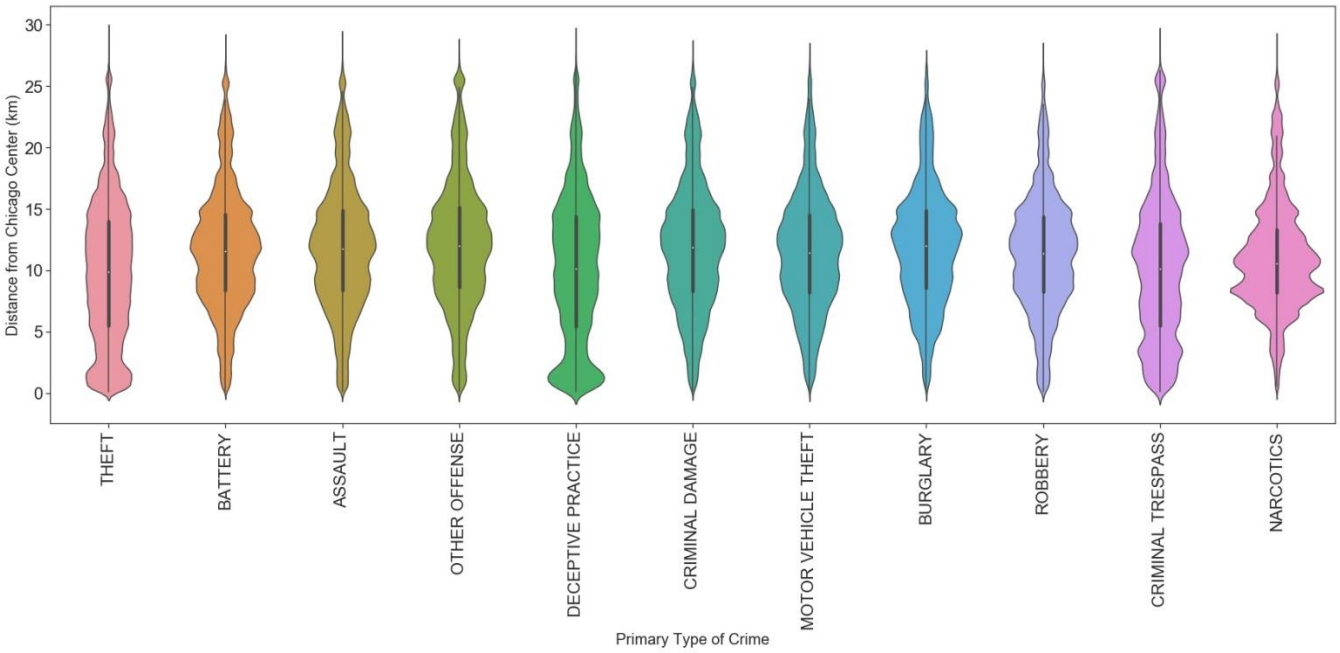

Figure 91: Distribution of crimes based on distance from Chicago city center for each primary type of crime.

Figure 92 shows that 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 crimes in the northwest area of Chicago between 87.7°W and 87.8°W.

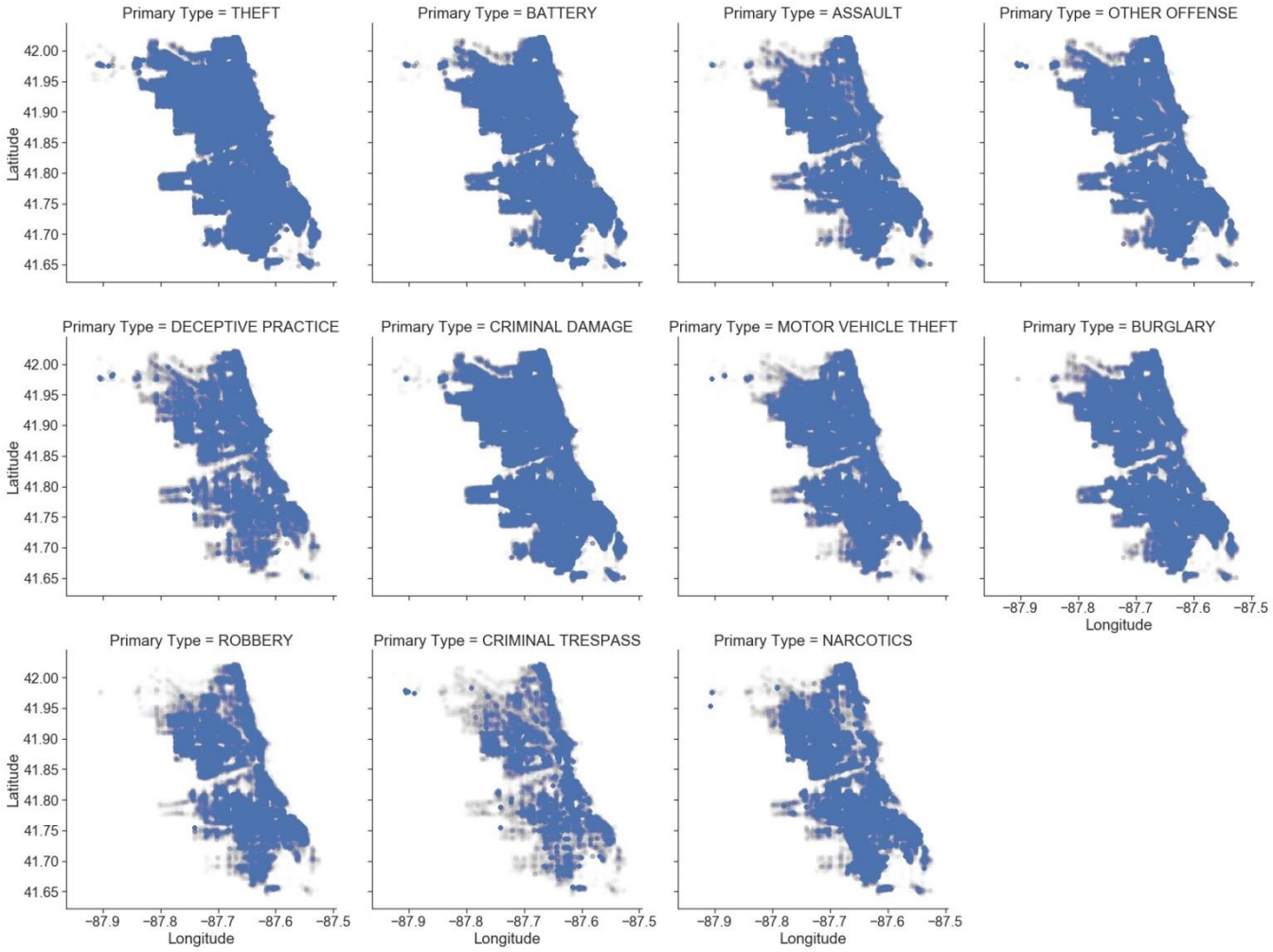

Figure 92: Spatial distribution of crimes by primary type of crime.

Per Figure 93, the chance of crimes involving battery and criminal damage increases over the weekend. This could be due to more people going out for leisure or in groups on weekends, so there would be a greater chance for these crimes. The chance of crimes involving theft and narcotics decreases slightly over the weekend. As a lot of people don't work on the weekend, they are likely to be at home for a larger part of the day, thus inhibiting theft from their residences or while they are out. Also, spouses/kids are more likely to be home on the weekend, so this could inhibit some individuals from dealing/using narcotics.

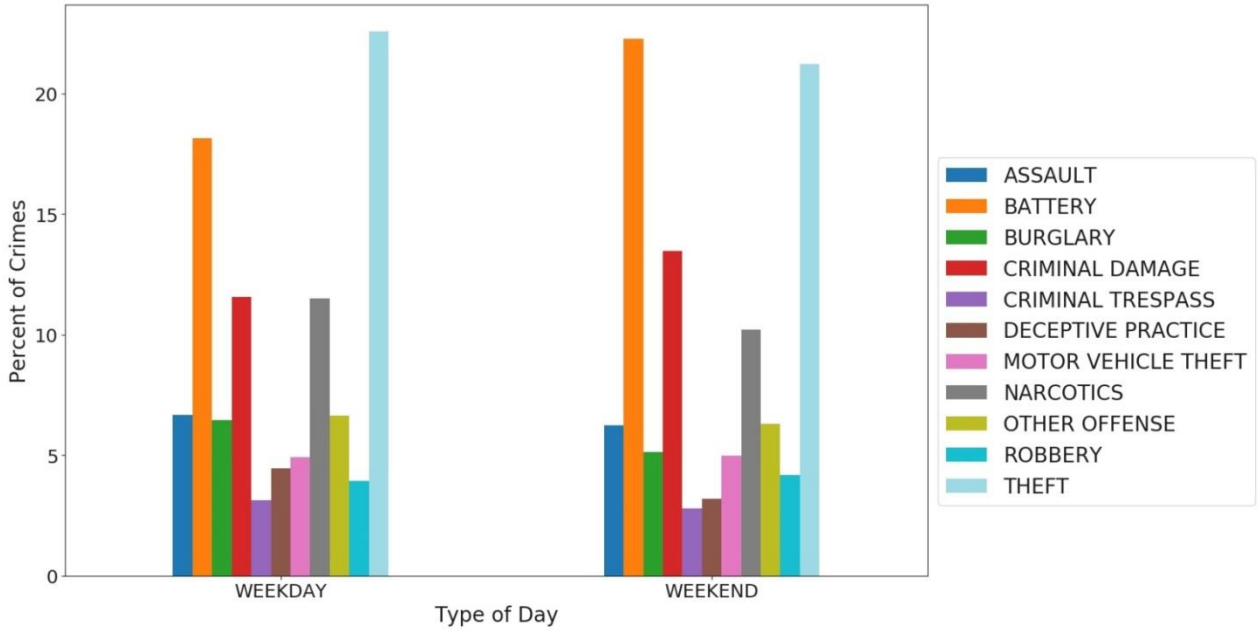

Figure 93: Percentage of each primary type of crime per type of day (weekday/weekend).

Figure 94 shows that battery increases on Saturday and peaks on Sunday while theft and narcotics decrease on Saturday and reach a minimum on Sunday. Criminal damage is slightly higher from Friday through Sunday.

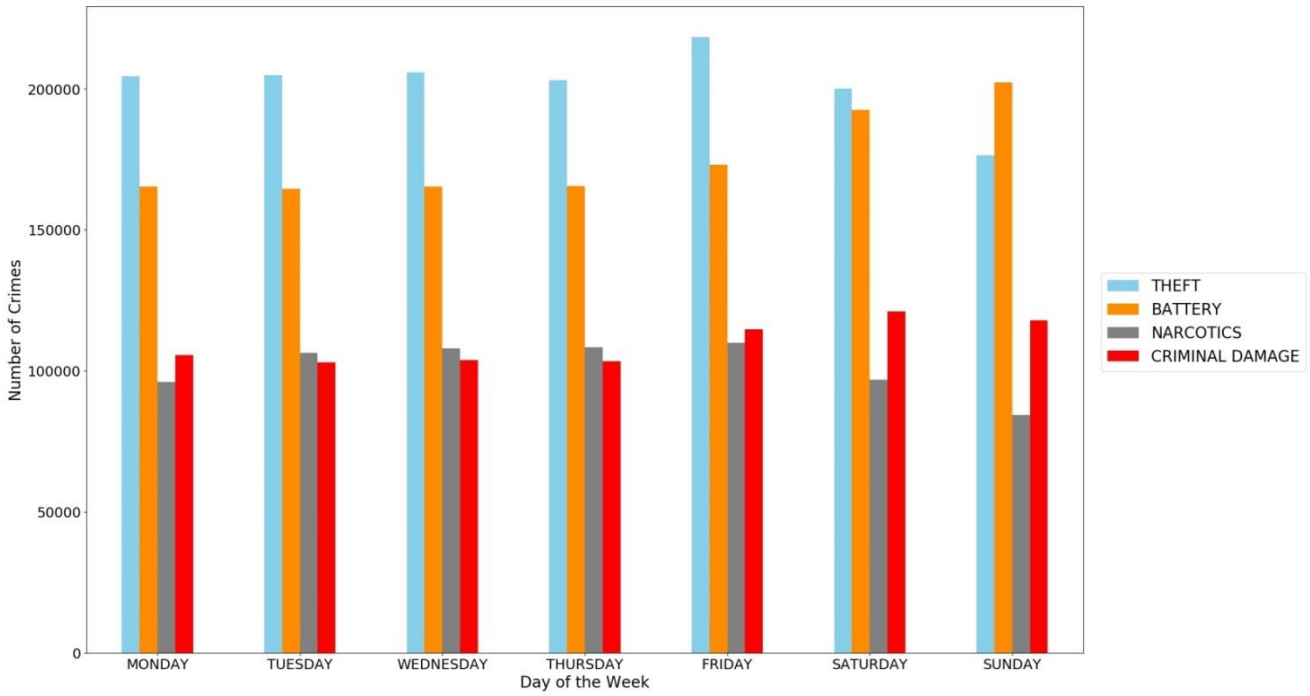

Figure 94: Number of each primary type of crime per day of the week.

On Federal holidays (Figure 95), there is a greater chance for crimes involving battery and a slightly lower chance for crimes involving narcotics. The increase in battery could be due to the higher possibility of people being out in larger groups. The decrease in crimes involving narcotics could again be due to the spouse/kids being home resulting in some individuals becoming less inclined to deal/use narcotics.

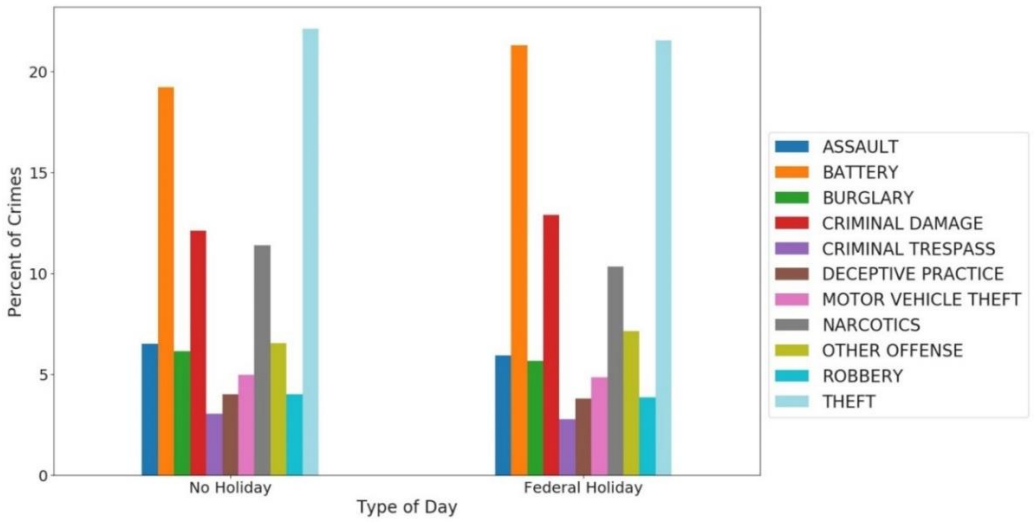

Figure 95: Percentage of each primary type of crime per type of day (no holiday/federal holiday).

Per Figure 96, Christmas day, Independence Day, Labor Day, Memorial Day, New Year's Day, and Thanksgiving Day all have higher proportions of crimes involving battery when compared to days with no holidays. It makes sense that there would be a higher proportion of crimes involving battery on Independence Day, Labor Day, Memorial Day, and New Year's Day as more people would be out in groups at parties or parades on these days. In addition, early on New Year's Day, more people are likely to be inebriated and therefore prone to cause trouble.

It is interesting that Christmas Day and Thanksgiving Day have higher proportions of crimes involving battery. I would have thought that the gatherings on Christmas and Thanksgiving would be smaller with mostly family. But perhaps because more estranged members of the family may be reluctantly gathering on these holidays, there could be more domestic related battery.

Compared to days with no holidays, Christmas Day, Independence Day, New Year's Day, and Thanksgiving Day have lower proportions of crimes involving narcotics. It makes sense to be lower on Christmas and Thanksgiving as more family would be around so people would be less apt to deal/use narcotics. Perhaps on Independence Day and New Year's Day, more people are consuming alcohol instead of using narcotics.

One more thing to note is that New Year's Day has a higher proportion of theft. This could be because more people are out of their residences on New Year's Eve and don't report theft until New Year's Day when they are back home. Also, as people are more apt to consume alcohol, more people may opt to walk to or from parties, thus putting themselves at risk of theft on the streets

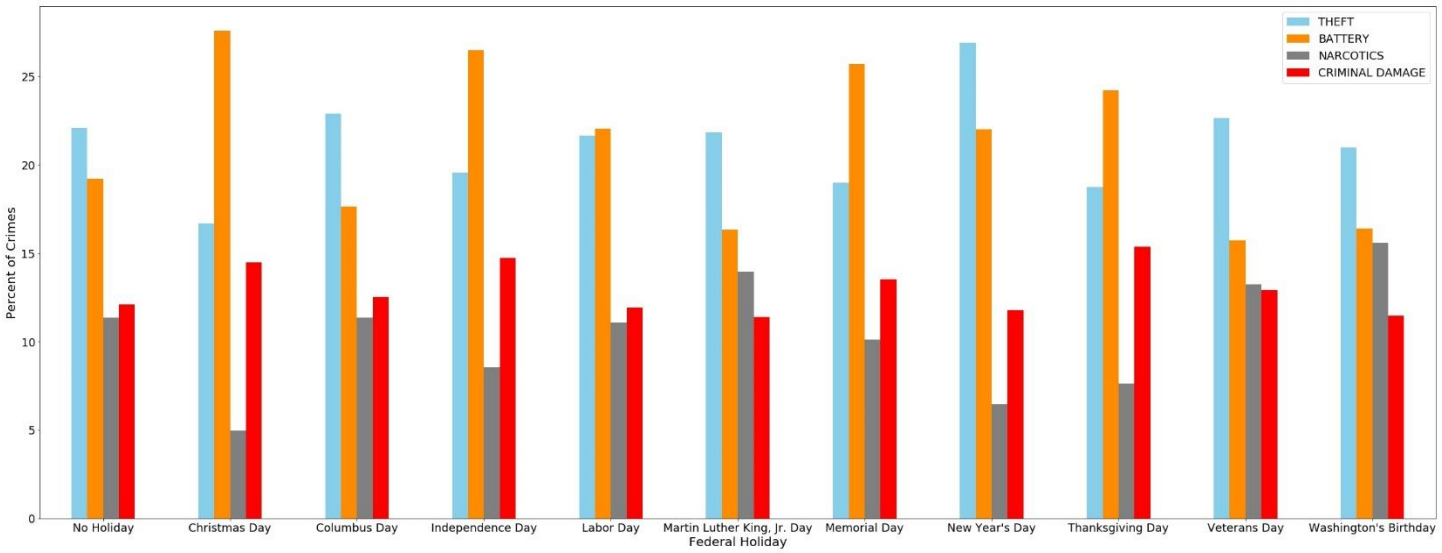

Figure 96: Percentage of 4 primary types of crime per federal holiday.

Per Figure 97, the number of crimes involving theft increases in the summer. This could be because more people are outside of their residences, leaving themselves and their residences vulnerable to theft. Crimes involving battery are more frequent during the spring and summer, possibly because the weather is nicer and it is likely that there are more events where a large number of people would gather. Crimes involving criminal damage are more frequent during spring, summer and fall. This could be due to the weather being nicer and more people being outside. There isn't too much of a change in the number of crimes involving narcotics across the seasons, but its occurrence in the winter is a little higher than the occurrence of criminal damage.

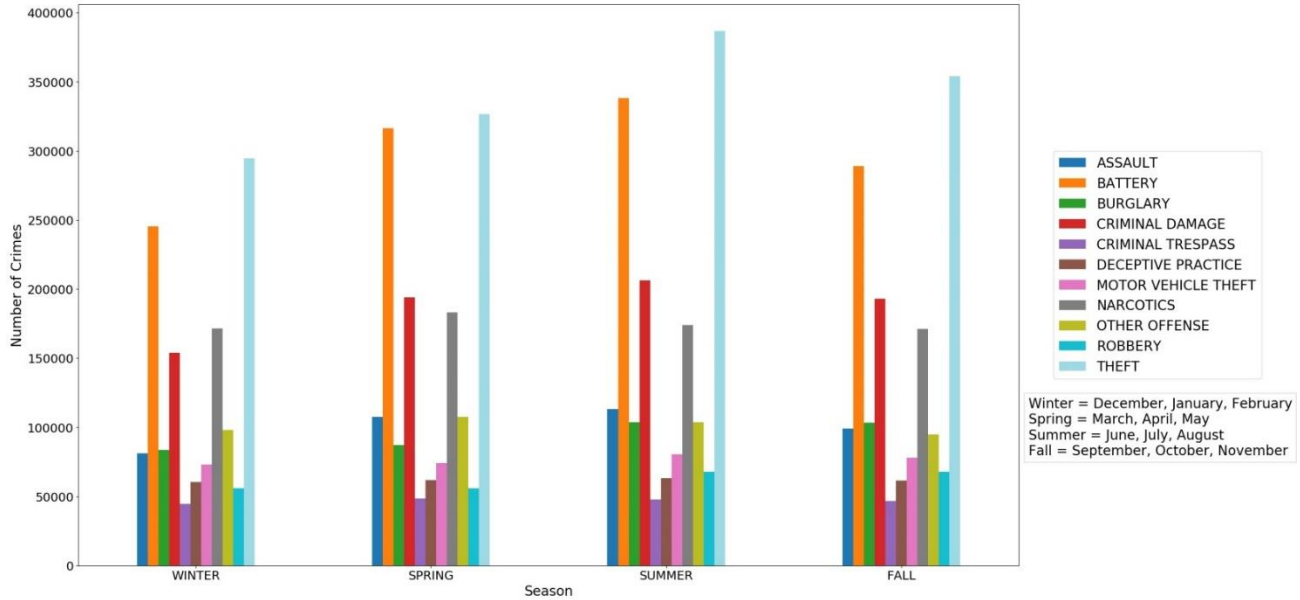

Figure 97: Number of each primary type of crime per season.

Per Figure 98, the occurrence of crimes involving theft is the highest during the afternoon, likely because more people are out of the house at this time. Crimes involving battery peak in the evening as it is usually the time when more people are out in groups. Also, during the overnight hours, it is the most frequent crime. Perhaps it could be more domestic related during the overnight hours or it could be that people who tend to go out in groups at such hours would be more apt to start trouble. Crimes involving narcotics peak in the evening, likely when more people are back home from work or school and are free to use/deal. Crimes involving criminal damage peak in the evening as people would wait for it to be dark before causing damage. I would have expected more criminal damage overnight, but perhaps many of the incidents weren't reported until the next morning when the victims woke up.

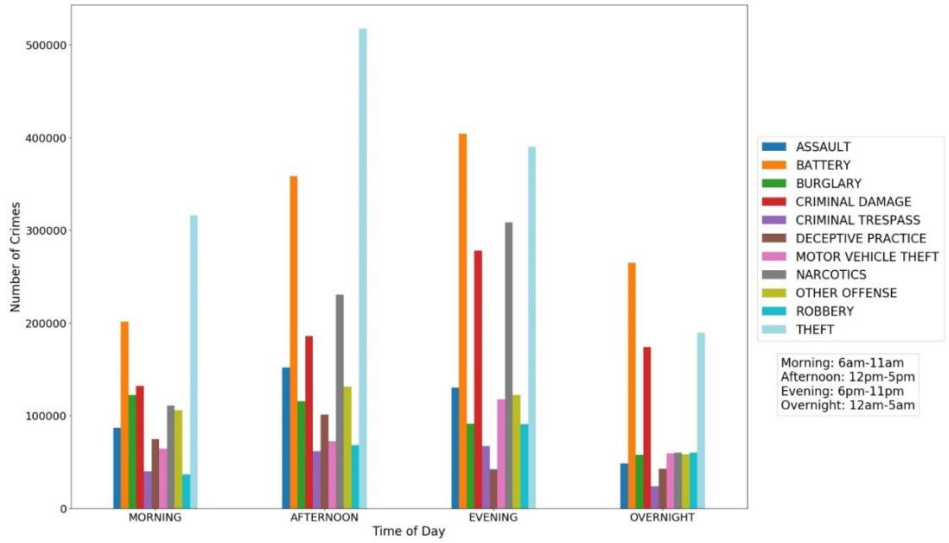

Figure 98: Number of each primary type of crime per time of day.

In Figure 99, there is a shift north in the average latitude for crimes involving deceptive practice, theft, criminal damage, and burglary into the evening. The average latitude for crimes involving motor vehicle theft shifts north during the afternoon and evening. The average latitude for crimes involving robbery shifts north into the overnight hours. The average latitude for crimes involving criminal trespassing shifts south during the afternoon.

Generally, the north side of a city is more affluent, so it could explain why there is a northward shift of crime into the evening. It is uncertain why criminal trespassing shifts south in the afternoon, though. For motor vehicle theft, it appears that perhaps the criminals start near their homes and then may move north, where there would be nicer cars that are unattended outside of places of work and then the criminals make their way back home late at night.

Another insight that can be gleaned from this figure is generally how far north or south certain crime types tend to occur. For example, deceptive practice and theft on average occur farther north while assault and battery, on average, occur farther south.

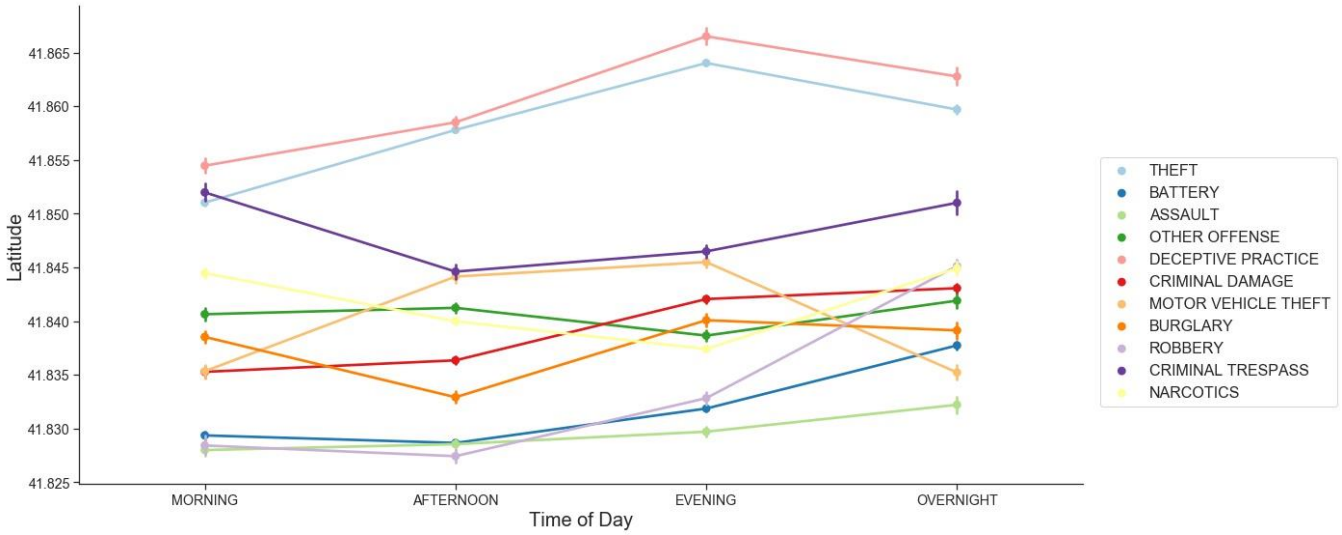

Figure 99: Average latitude for each primary type of crime per time of day.

In Figure 100, there is a shift east in the average longitude of crimes involving narcotics and criminal trespassing during the afternoon and evening. This possibly could be criminals starting near their homes and then moving east towards the city center where more affluent people would live and then moving back west during the overnight hours. In addition, it can be seen that crimes involving narcotics on average occur more to the west and crimes involving criminal trespassing on average occur more to the east.

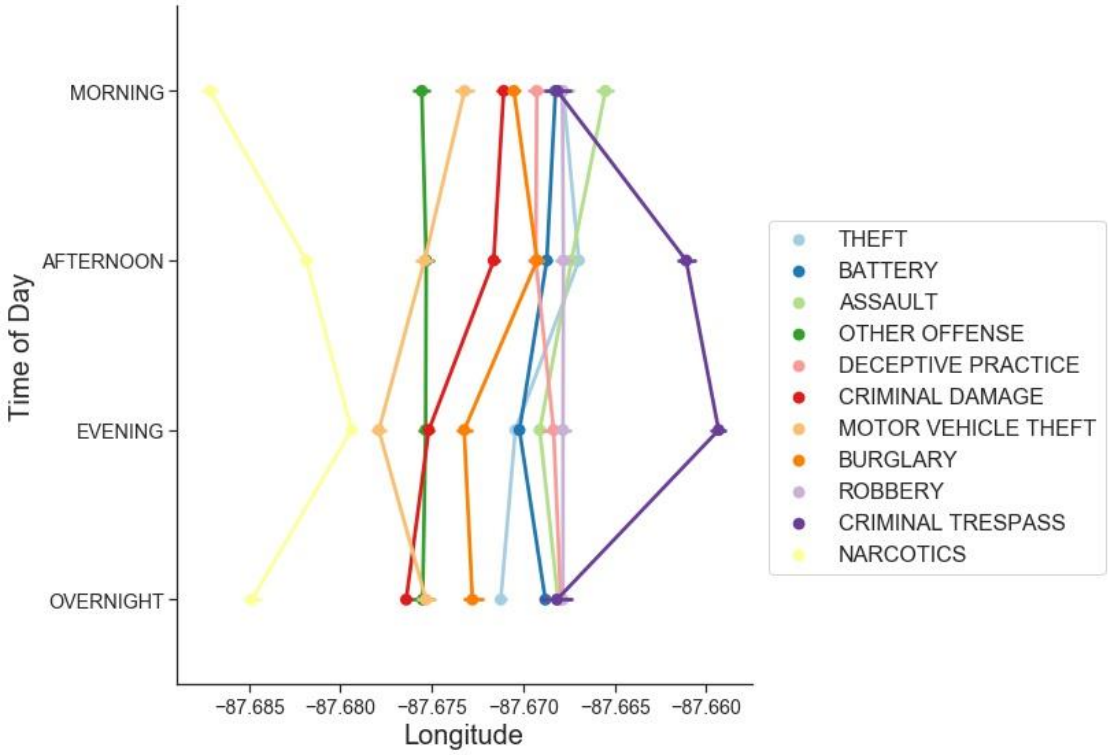

Figure 100: Average longitude for each primary type of crime per time of day.

In Figure 101, there is an increase in the average latitude of crimes involving deceptive practice, robbery, and battery during the weekend. This could be because many criminals are off from work and are free to go farther north to more affluent areas on the weekends. It is interesting that there is a shift south in the average latitude for crimes involving burglary. Further investigation would be needed to see why this is the case.

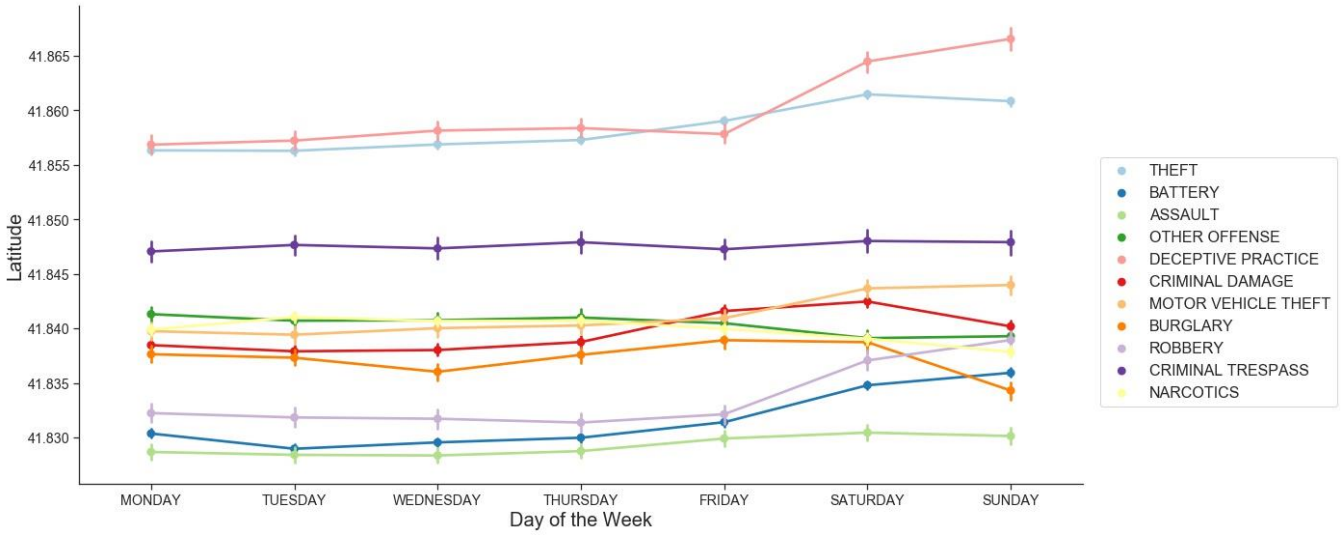

Figure 101: Average latitude for each primary type of crime per day of the week.

In Figure 102, the shifts in average longitude are subtler. In general, there appears to be a westward shift in crimes involving motor vehicle theft, criminal damage, battery, and criminal trespassing into the weekend. This is interesting as I would have thought that there would have been more of an eastward shift towards the city center.

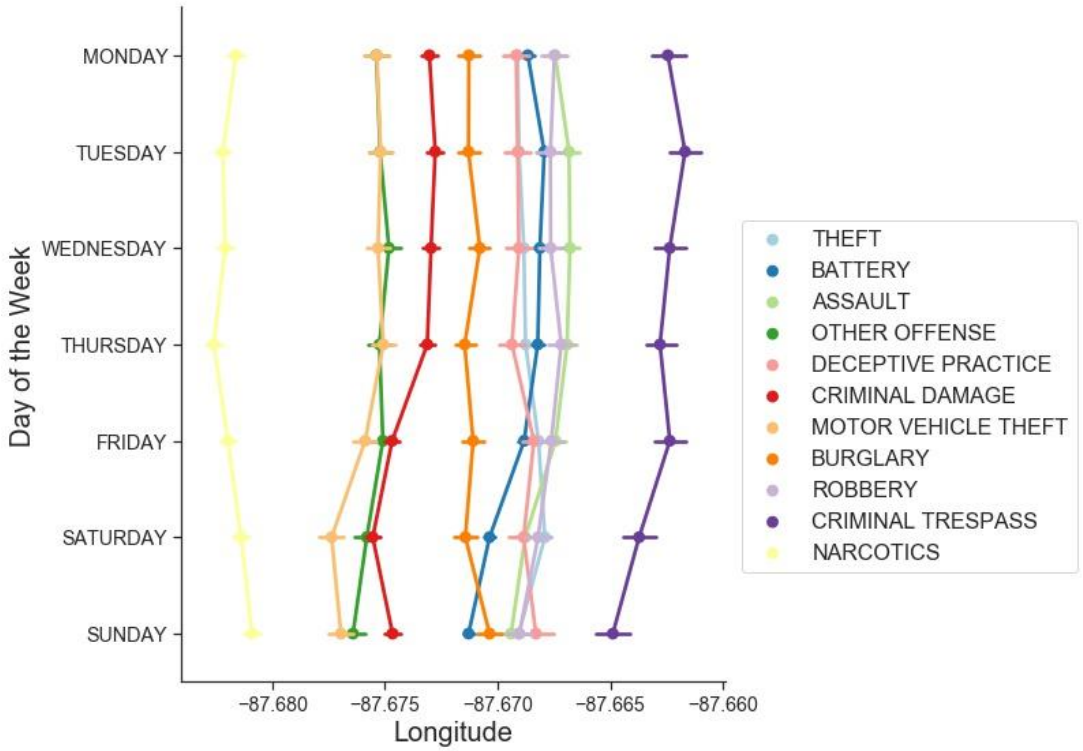

Figure 102: Average longitude for each primary type of crime per day of the week.

Conclusions, Further Research, and Recommendations

The intent of my research was to figure out what features would help predict the type of crime that would occur/be reported at a given place or time. I also explored some temporal and spatial attributes and how they affected the occurrence of certain crimes in Chicago. This research would be of great use to police officers as they may be able to better anticipate the type of crime that may occur or perhaps be able to prevent the crime from happening with their presence. Also, by knowing if there are hotspots of a certain crime, an adequate number of officers could be dispatched to patrol those areas.

I utilized crime reports in Chicago from 2001 into 2018 as well as data for train stops, bus stops, liquor stores, police stations, and federal holidays. To ensure I had enough samples of each type of crime, I focused my study on the top 11 most frequent crimes.

As I did not have adequate computer memory to train/test the entire dataset (6,357,103 reports), I randomly chose a sample of 3 million crime reports and used it to evaluate several models. Overall, scaling the data using MinMaxScaler and then running the SGD Classifier with the class weight set to 'balanced' and the loss set to 'log' produced the highest average F1 score, recall, and precision.

I then explored the individual coefficients for each feature and crime type produced by the SGD Classifier. The location description appeared to be the most important indicator for the reported crime type. The 4 locations with by far the most reported crime were street, residence, apartment, and sidewalk.

When I looked at the features' average coefficients across all crime types and dummy variables, the following were the features that most strongly discriminated between the 11 primary types of crime: X/Y coordinates (longitude/latitude), day type (weekday/weekend and no holiday/federal holiday), distance from the city center of Chicago, season, time of day, day of the week, and the federal holiday.

Crimes involving theft, deceptive practice, criminal trespassing, and narcotics had relatively higher number of crimes towards the northern part of Chicago. Crimes involving theft, deceptive practice, and criminal trespassing had higher concentrations closer to the city center.

Generally, all crime types had high concentrations of crimes along Lake Michigan. For crimes involving narcotics, there were some gaps in the concentration of crimes along the lake and a larger gap slightly farther inland on the north side.

The chance for crimes involving battery and criminal damage increased over the weekend while the chance for crimes involving narcotics decreased. On federal holidays, there was a greater chance for crimes involving battery and a slightly lower chance for crimes involving narcotics. Compared to days with no holidays, Christmas day, Independence Day, Labor Day, Memorial Day, New Year's Day, and Thanksgiving Day all had higher proportions of crimes involving battery. Christmas Day, Independence Day, New Year's Day, and Thanksgiving Day had lower proportions of crimes involving narcotics.

The occurrence of crimes involving theft was at the highest during summer and the lowest during the winter. It was also the highest during the afternoon and the lowest during the overnight hours. The occurrence of crimes involving battery was higher during spring and summer and during the evening. The occurrence of crimes involving criminal damage and narcotics was highest during the evening.

The average latitude of several crime types varied markedly with the time of day and the day of the week. There was also variability in the average longitude of a few crime types with time of day and subtle variability with the day of the week.

Based on the above conclusions, I would make the following recommendations:

- Increase police officer/security guard presence in public schools to help prevent the occurrence of battery. Be sure to properly train the officer/guard to safely handle battery related incidents. Start with schools within community 25 to see if it makes a difference.
- In areas surrounding the Chicago city center, especially in community 35, police officers should be on the lookout for individuals using/dealing narcotics. More police officers should patrol these areas in an effort to thwart narcotic related crimes.
- Increase police officer presence on the eastern edge of the city along Lake Michigan. Even criminal trespassing, which occurred the least frequently of the 11 studied primary types of crime had high concentrations of crime in this area.
- Police officers need to be prepared to deal with and be on the lookout for more instances of battery on weekends and on holidays.

- Increase the number of police officers patrolling near the city center or the number of security guards within department stores to thwart theft. Try not to reduce the number of officers/guards on duty in this area on holidays when the stores are still open but more people are off from work/school.
- Crimes involving criminal damage on residential driveways are quite widespread across Chicago and there are no significant hot spots for police officers to concentrate on. In order to mitigate the occurrence of criminal damage on residential property, it may be useful for the police to help residents set up neighborhood watches.

Further research possibilities include the following:

- In order to more accurately obtain the missing latitudes/longitudes, the blocks could be geocoded.
- The ward boundaries change after every federal census (Knox, 2005). The latitude/longitude could be used to redo all of the reported wards so that only one ward boundary system is used.
- Using all available crime reports for the top 11 most frequent types of crime, cross validation could be performed on the examined models to see if the SGD Classifier is indeed the most optimal model. After choosing the best model, it could be tweaked by adjusting the model parameters and removing unneeded features or adding new features to improve the accuracy.
- The location description was the most important indicator for the reported crime type when looking at feature coefficients individually. There were approximately 100 unique location descriptions in my dataset. Natural language processing could possibly be used to group similar locations together. The data could then be used to train/test the SGD Classifier to see if there is any improvement in the accuracy or if when looking at the most influential coefficients, any features other than location description show up.
- The top 4 location descriptions (street, residence, apartment, and sidewalk) could be analyzed to see if the number of certain crimes for each location changes spatially. For example, looking at how the concentration of crimes involving theft on streets changes across Chicago.
- For crimes involving narcotics, there was a large area with a lower concentration of crime slightly inland from the lake on the north side. A study should be done to see what it is about this area that makes it less prone to crimes involving narcotics; or if for some reason, it occurs here but isn't reported. Demographics, income, business density, and officer density are a few features that could be examined.
- As some variation in the frequency of certain crimes (especially theft and battery) occurred with season, the weather conditions preceding and during the crimes could be examined to see if it would be a good predictor for the type of crime.
- As several of the crime types varied throughout the day, foot traffic data could be studied to see if there really is a significant relationship between the number of people outside and the occurrence of certain crimes (e.g. battery).
- Variation in average latitude/longitude throughout the day and week was seen for several types of crime. In order to see where the biggest shifts in crime are occurring, the spatial distribution of each crime for each time of day and day of the week could be examined.

References

City-Data. (2018). Crime rate in Chicago, Illinois [online image]. Retrieved from http://www.city-data.com/crime/crime-Chicago-Illinois.html

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