STA 199: Final Project

Analyzing and finding patterns in Chicago crime during 2019

Team gitData: Blossom Mojekwu, Avanti Shah, Nathan Kim

Section 1: Introduction

Earlier this semester, we explored the crime in New York City by examining the "stop-question-and-frisk" program. Our group felt as though crime data similar to the New York dataset would be able to provide extensive and insightful trends and analysis about important factors and variables that affect crime rates; for example, how socio-economic factors (such as neighborhood of residence) might affect violence and misdemeanor.

Our selected dataset is from the Chicago Police Department, and contains reported incidents of crime from 2001 to the present. The data was extracted from the Chicago Police Department's "Citizen Law Enforcement Analysis and Reporting" or CLEAR system. In order to protect the privacy of crime victims, the specific addresses and locations were made confidential; only the block on which the crime occurred is provided. It is also important to note that this dataset does not include reported incidents of murder where data exists for each victim.

We narrowed the large dataset (which had approximately 7 million entries!) to only include data from the most recent annum (that is, 2019). We further narrowed down this dataset by *randomly* sampling 20,000 entries from approximately 200,000 total observations in 2019.

Our research question, then, is **"What are the trends in crime rates through different times of year, types of crimes committed, and most common location of crime in Chicago city in 2019?"** With these questions, we want to analyze the relationships between neighbourhood sides, weather, and date and time of day, and crime rates. Secondarily, we are interested in other variables that can shed light on crime, for example - the rates of arrest and domestic crimes and how these are correlated with the aforementioned variables.

Each case corresponds to a specific crime report, with each case having 22 variables. Each case has two unique identifiers: ID and case number. The case number specifically corresponds to the Chicago Police Department's Record Divisions number, which is unique to each incident. The approximate time and date are provided under "Date". This is sometimes a best estimate of when the crime occurred. The partially redacted address is given under the "Block" variable, giving a rough estimate as to where the crime occurred. The "IUCR" variable corresponds to the Illinois Uniform Crime Reporting code, which is directly tied to the "Primary Type" and "Description" variables. The "Primary Type" variable categorizes the crime under a general description and the "Description" variable places the crime into subcategories based on the "Primary Type" variable. The "Location Description" variable provides a general categorization for where the crime took place, such as "residence" or "sidewalk". The "Arrest" variable indicates whether an arrest was made and the "Domestic" variable indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act. The "Beat" variable provides the beat where the incident occurred. A beat is a more specific system used by the police department in order to split up police districts. The "District", "Ward", "Community Code", "X Coordinate", "Y Coordinate", "Latitude", "Longitude", and "Location" similarly provide more details on where the crime occurred. The "Location" variable is slightly significant as it is formatted to allow for the creation of maps and other geographical plots. The "FBI Code" provides the crime classification as outlined by the FBI's National Incident-Based Reporting System, or NIBRS. The "year"

variable provides the year in which the crime occurred and the "Updated On" variable provides the date and time the record was last updated.

Section 2: Data Analysis Plan

The outcome we are exploring is the highest proportion of a particular type of crime in 2019. The independent variables will be date, crime type, and community area. We want to analyze trends in the rates of crime through different parts of the year, the recorded crime most committed, and where the majority of crimes are committed.

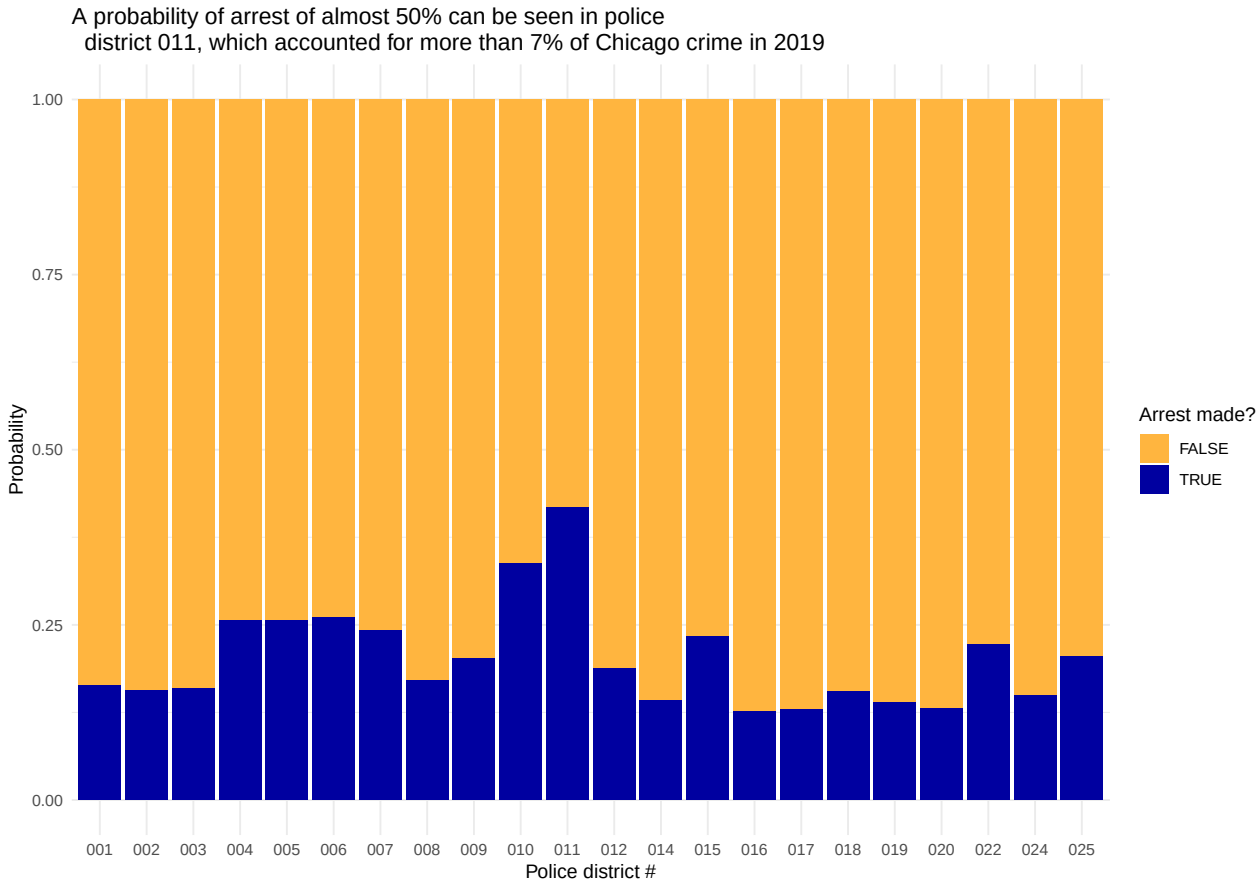
Our exploratory data analysis will include a line-scatterplot to illustrate how crimes over 2019 varied by month (and therefore by weather or by season; *See Section 4*), and a bar graph of crimes by ward and district, including whether an arrest was probable. We will create a summary table to count the number of crimes for each community area and what crime type is most prevalent in certain neighborhoods *(See Section 4).*

We will also analyze crime rates in different locations in Chicago, which is enabled by the abundance of location data present in the dataset. Here is a preliminary data analysis for different Chicago locations (general public locations, wards, districts):

```
# A tibble: 10 x 2
   location_description           num_crimes
   <chr>                               <int>
 1 STREET                               4425
 2 RESIDENCE                            3298
 3 APARTMENT                            2671
 4 SIDEWALK                             1543
 5 OTHER                                 841
 6 SMALL RETAIL STORE                    567
 7 PARKING LOT/GARAGE(NON.RESID.)        560
 8 RESTAURANT                            553
 9 ALLEY                                 401
10 RESIDENCE PORCH/HALLWAY               374
```

From this tibble, we can see that street, residences, and apartments are the most common locations where crime occurred in Chicago in 2019, which fits as predicted, since these areas are common places for interaction between individuals. These areas account for more than 50% of the crimes in our sample. Street crime is common in large, gentrified cities, while apartments and residences are common places for violent crime, sexual crime, and theft, such as home invasions *(See Section 4 for a detailed analysis and visualization.)*

```
# A tibble: 22 x 3
   district num_crimes perc_total_crimes
   <chr>         <int>             <dbl>
 1 011            1420              7.1
 2 006            1307              6.54
 3 008            1177              5.88
 4 001            1167              5.84
 5 018            1167              5.84
 6 004            1082              5.41
 7 007            1077              5.38
 8 012            1052              5.26
 9 025            1000              5
10 003             941              4.70
# ... with 12 more rows
```

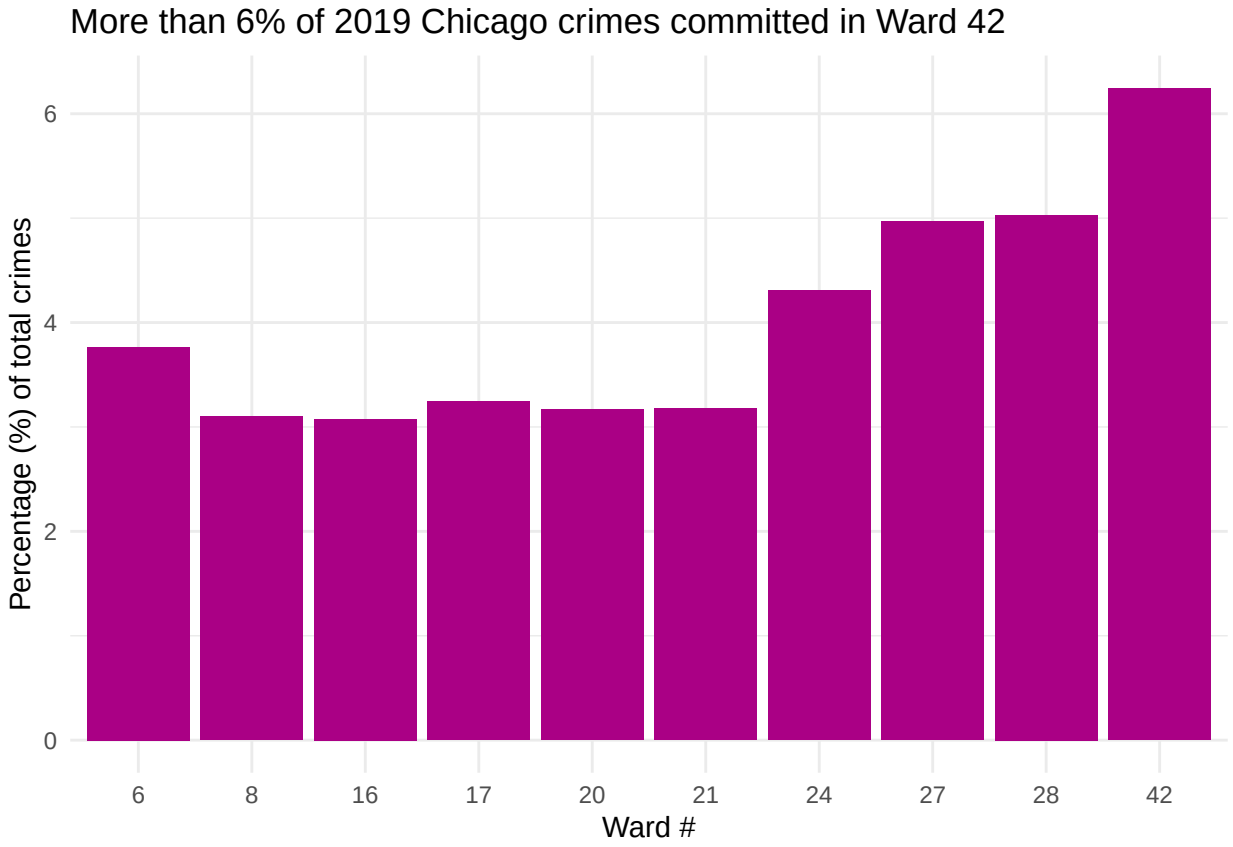


Tibble 2 in conjunction with this graph shows us that districts 011 and 006 totally accounted for more than 13% of total crime in Chicago in 2019. The Marshall Project called district 011 "the most dangerous neighborhood [with] the most inexperienced cops". This district is known to have an especially high crime rate, and has the highest murder rate in the city (not included in this dataset, but worth noting). As a result, rookie police personnel combined with an already high crime rate cause a high crime rate. This area also has a probability of arrest of nearly 50%, which means that of the crimes reported, about 50% of culprits are arrested. Numerous studies have shown that arrest rates and reported crime rates usually have a negative relationship (Levitt, 1995), so perhaps arrest rates are increased in hopes of bringing reduction to crime rates in a crime-ridden area.

District 006 is Gresham, which is located in the southside of Chicago, and has a reputation of poor safety and security, due to pre-existing racial underpinnings and discrimination. However, the southside is known less for violent crime and more for theft and robbery, perhaps due to the aforementioned wealth inequality. Hence, we can see that these areas have complex social factors affecting their crime rates.

```
# A tibble: 10 x 3
    ward num_crimes perc_crimes
   <dbl>      <int>       <dbl>
 1    42       1249        6.24
 2    28       1006        5.03
 3    27        993        4.96
 4    24        862        4.31
 5     6        753        3.76
 6    17        649        3.24
 7    21        635        3.18
 8    20        633        3.16
 9     8        620        3.1
```

```
10    16        615        3.08
```



Tibble 3 above and this graph convey that Ward 42 accounted for more than 6% of total crime in 2019, where Ward 42 contains many of Chicago's biggest tourist attractions, including Navy Pier, the Magnificent Mile, Millennium Park and the Art Institute. This is interesting, because a plethora of research shows that an increased number of people in a relatively small area (usually in tourist attraction areas in large cities) "contributes to the introduction of 'strangers' to local communities which can lead to higher crime rates" (Lisowska, 2017). As a result, these graphs and tibbles are important in highlighting geographic areas where crime tends to occur, and supports research being conducted in economics and public policy surrounding bases of crime.

These are our preliminary data analyses and visualizations. In order to answer our questions in more detail, using sophisticated statistical methods, we will make use of multinomial regression since we want to consider multiple categorical variables (such as time of year, date, location, and arrests made) and how they are correlated with classification of crime. We will analyze any confounding variables we encounter to see if the exclusion of one variable can affect the measure of correlation between two other variables; that is, if Simpson's paradox is demonstrated in our data. This is effective since we are handling data related to location, arrests, and domestic crimes, so we can use multiple variables to see their correlations. The correlation coefficients of these models can help us understand their relationships and offer a model under which we can extrapolate future values.

In terms of analysis and visualization, we will use data wrangling to manipulate variables and create new, composite variables of interest, as we have done above. Furthermore, since we have location and latitude-longitude data, we can use spatial data visualization to plot the observations on a Chicago map and plot these visually by neighborhood to display where the most crimes occur, and where a particular type of crime is more likely to occur (*See Section 4*). We will analyze the statistical significance of these coefficients against a pre-determined significance level.

We will also use relative risk calculations, since relative risk ratios will allow us to interpret the coefficients

from the multinomial regression we create, especially since we have several variables we are interested in considering.

Finally, we will try to analyze data collection and recording methods, and also try to account for any potential errors or assumptions that are important in analyzing these data. Moreover, we will outline any possible biases or extraneous variables that might skew our data, and attempt to analyze their impact.

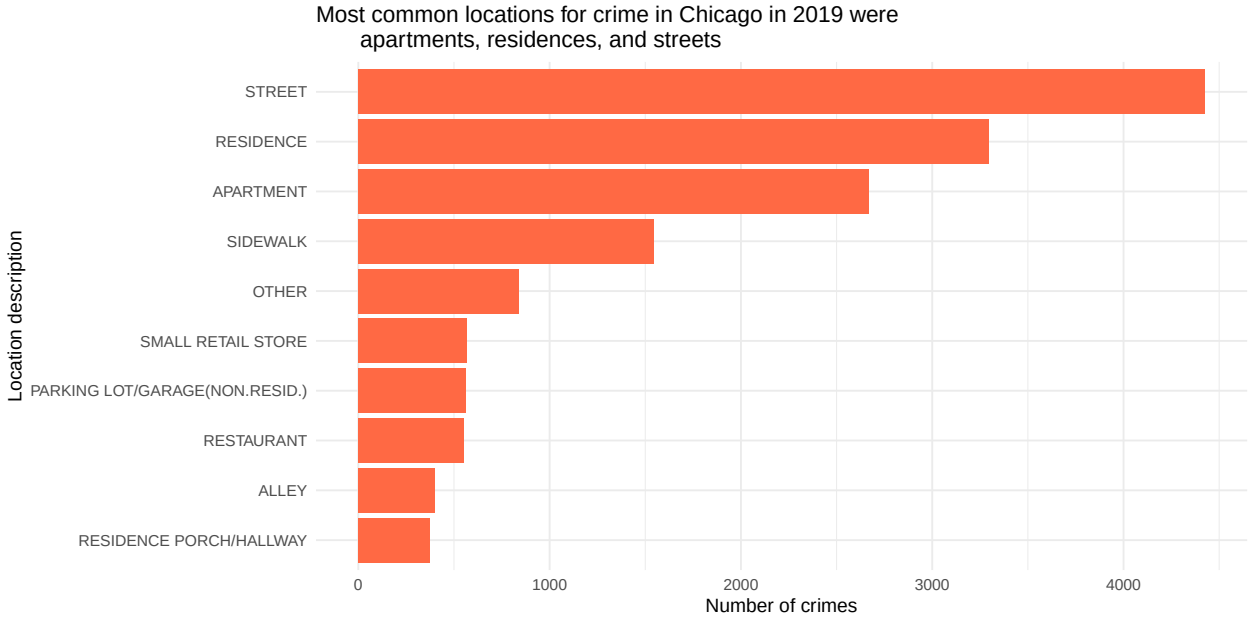
In terms of specific results, we will make use of results such as the mean number of crimes per month, the multinomial regression model's coefficients and results from spatial data visualization to answer our research question, as well as results from manipulating variables, data wrangling, and data visualization, which will be extremely helpful in displaying relationships.

Section 3: Data & Codebook

```
Observations: 20,000
Variables: 22
$ id                   <dbl> 11635151, 11724411, 11921816, 11797726, 119369...
$ case_no              <chr> "JC184260", "JC308572", "JC547082", "JC396741"...
$ date                 <chr> "03/12/2019 12:00:00 PM", "06/16/2019 08:10:00...
$ block                <chr> "012XX N LA SALLE DR", "064XX S ASHLAND AVE", ...
$ iucr                 <chr> "1153", "1320", "0460", "0430", "1152", "0610"...
$ crimetype            <chr> "DECEPTIVE PRACTICE", "CRIMINAL DAMAGE", "BATT...
$ description          <chr> "FINANCIAL IDENTITY THEFT OVER $ 300", "TO VEH...
$ location_description <chr> "APARTMENT", "PARKING LOT/GARAGE(NON.RESID.)",...
$ arrest               <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE...
$ domestic             <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE...
$ beat                 <chr> "1821", "0725", "1434", "1113", "1225", "2413"...
$ district             <chr> "018", "007", "014", "011", "012", "024", "008...
$ ward                 <dbl> 2, 15, 2, 28, 28, 50, 14, 37, 9, 30, 28, 20, 3...
$ community_area       <dbl> 8, 67, 24, 25, 28, 2, 57, 23, 53, 16, 25, 68, ...
$ fbi_code             <chr> "11", "14", "08B", "04B", "11", "05", "14", "2...
$ xco                  <dbl> 1174911, 1166779, 1163356, 1145207, 1160554, 1...
$ yco                  <dbl> 1908547, 1861916, 1910670, 1899610, 1894873, 1...
$ year                 <dbl> 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019...
$ updated              <chr> "03/29/2019 04:05:05 PM", "06/30/2019 03:56:27...
$ latitude             <dbl> 41.90444, 41.77666, 41.91052, 41.88053, 41.867...
$ longitude            <dbl> -87.63294, -87.66415, -87.67532, -87.74228, -8...
$ location             <chr> "(41.90444226, -87.63293969)", "(41.776660592,...
```

Section 4: Methods and Results

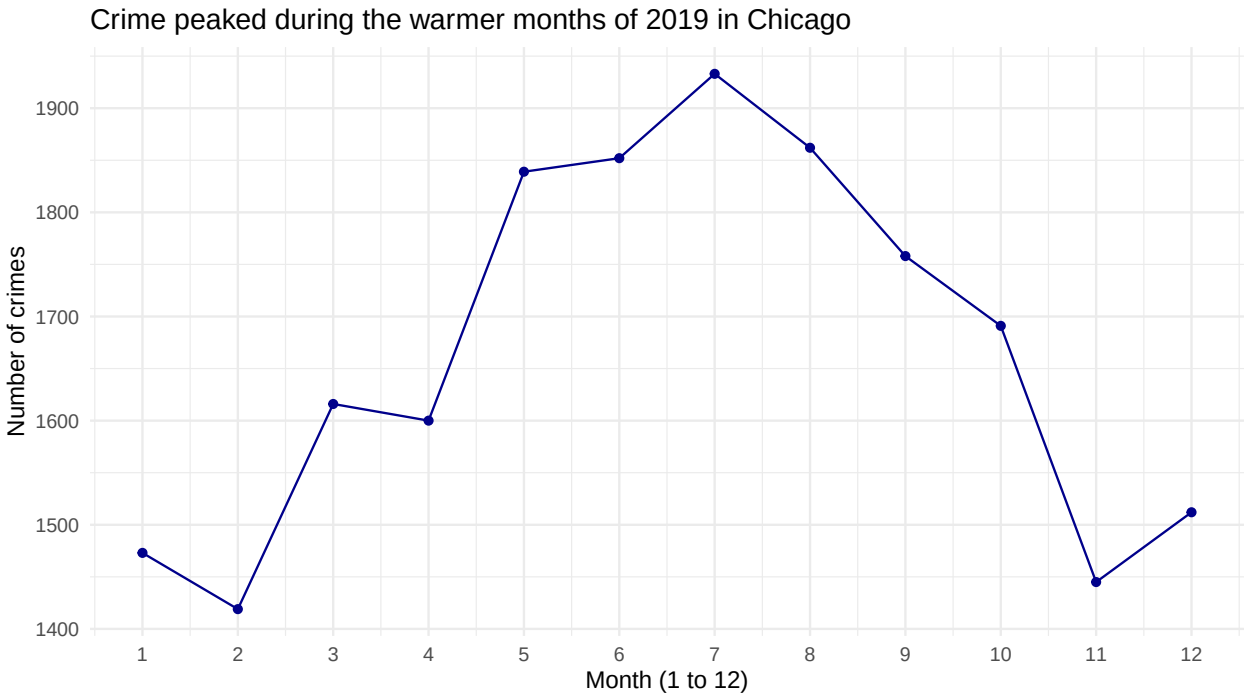
Most common crime locations



Methods and Results: The three most common locations for crimes committed in Chicago in 2019 were streets, residences, and apartments. The method used is the creation of a table from the crimes dataset that omits empty entries and counts the frequency of each `location_description`. The top 10 locations are sliced because they account for 15,240 observations out of the 20,000 sample observations; that is, the top 10 locations are representative of 76.2% of the sample data. We assume that the sample data from the raw dataset is representative of all Chicago crimes committed and recorded in 2019, since it was randomly sampled. A bar graph is created to visualize the 10 most common crime locations, with streets being the most common.

Streets accounted for about 22% of the crimes in the sample, which signifies a high rate of street crime, which is loosely defined as crimes that often take place in public areas, such as drug trade, pickpocketing, and mugging. Many studies have attempted to pinpoint the socio-economic factors that affect the rise of street crime, and often chalk it down to high levels of income inequality, poverty, unemployment, and parental neglect. A 2008 study by Iowa State University researchers find that offenders often commit street crime in search of "societal status and recognition and thrill and excitement" (Bennett and Brookman, 2008), and found a correlation between the prevalence of these crimes and large social inequities, especially in larger cities.

Date trends in crime



Methods and Results: We created a scatter-line plot to visualize the number of crimes plotted againt the months of the year. We found that crime correlates with weather in Chicago - warmer months span from June to September, and crime peaks during these months. Crime is at its lowest in Feburary, which is correlated with Chicago's coldest winter month. As a result, this visualization shows us that there is a correlation between weather (or temperature) and number of crimes. Moreover, these months tend to also be holiday months (usually June to August or September), so it is interesting that crime is on the upturn in December compared to November despite the winter season. This is, therefore, an interesting correlation of crime numbers with the two major seasons in the year (summer and winter).

An abundance of research in criminology and economics shows that in many places all over the globe, temperature does play a role in crime rates. A study in Finland (Tiihonen et al, 2017) showed a 1.7 percent increase in criminal activity for each degree centigrade rise in the temperature. More specifically, researchers found that high temperatures in the summertime often caused increased levels of serotonin (the "happy" neurotransmitter) contributed to increased impulsivity, human activity level, and social interaction, leading to a rise in the risk of violent incidents.

Incidence of each broad type of crime

```
# A tibble: 6 x 2
  crime_category crimetype_perc
  <chr>                   <dbl>
1 theft/robbery           34.6
2 violent crime           28.4
3 other                   26.4
4 drug-related             5.5
5 violation                3.64
6 sex offense              1.45
```

Methods and Results: We created groupings of the different types of crime, named `crime_category`. The groupings were constructed as follows:

`Theft/robbery`: Theft, burglary, motor vehicle theft, and robbery

`Violent crime`: Battery, assault, arson, kidnapping, human trafficking, offense involving children, and homicide

`Other`: Other offenses, intimidation, obscenity, gambling, deceptive practice, criminal trespass, stalking, and criminal damage

`Sex offense`: Criminal sexual assault, sex offense, and prostitution

`Violation`: Weapons violation, interference with public office, public peace violation, liquor law violation, and concealed carry license violation

We then calculated the crime percentage for each crime category. Evidently, more than half of the crimes are classified under theft or robbery, and violent crime, while sex offenses and drug-related crimes are significantly less likely. We chose to create sub-categories because the dataset initially contained approximately 30 different types; analyzing many different crime types would make it difficult to see any trends and create too many variables that could easily be grouped together and present a more coherent picture of the relationships.

Chicago sides and individual crime rates

```
# A tibble: 9 x 3
  nb_side       count crime_perc
  <chr>         <int>      <dbl>
1 west           5093      25.5
2 south          2687      13.4
3 southwest      2587      12.9
4 far southeast  2344      11.7
5 central        1893       9.46
6 far north      1700       8.5
7 north          1421       7.10
8 far southwest  1263       6.32
9 northwest      1012       5.06
```

Methods and Results: We created a table that categorized the community areas into groups of Chicago's informal city sides. We based the groupings on a "City of Chicago's Community Areas and Sides" map sourced by the City of Chicago Data Portal. Based on the count of crimes committed in each city side, we created a crime proportion of the representative sample of crimes. As a result, we see that the West side has both the highest number of crimes committed along with the highest crime percentage, account for about one-fourth of all crimes in our sample. We thought a table was the best visualization technique for these data because having few variables allows for easy comparison of the numbers between neighborhood sides, and arranging in descending order allows emphasis on the areas with the highest number and percentage of crime. Since it is difficult to put the raw numbers in context, the percentage column facilitates easy comprehension of the crime rates in each side.

Multinomial logistic regression for multiple variables

```
# weights:  78 (60 variable)
initial  value 35835.189385
iter  10 value 28012.320806
iter  20 value 27923.937138
iter  30 value 27797.125763
iter  40 value 27721.002124
iter  50 value 27692.096643
iter  60 value 27688.146873
final  value 27687.946915
```

converged

	Dependent variable:				
	other	sex offense	theft/robbery	violation	violent crime
	(1)	(2)	(3)	(4)	(5)
nb_sidecentral	0.400	0.828*	0.994***	0.509	0.045
	(0.261)	(0.468)	(0.258)	(0.393)	(0.262)
nb_sidefar north	0.699**	1.270***	0.822***	0.807**	0.555**
	(0.276)	(0.468)	(0.274)	(0.401)	(0.276)
nb_sidefar southeast	-0.450**	0.243	-0.524**	0.906***	-0.286
	(0.217)	(0.424)	(0.216)	(0.329)	(0.216)
nb_sidefar southwest	-0.278	-0.507	-0.468*	0.832**	-0.226
	(0.241)	(0.523)	(0.240)	(0.355)	(0.240)
nb_sidenorth	0.665**	0.889*	1.112***	0.380	0.224
	(0.293)	(0.508)	(0.290)	(0.445)	(0.295)
nb_sidesouth	-0.457**	-0.026	-0.435**	0.225	-0.233
	(0.214)	(0.427)	(0.212)	(0.335)	(0.213)
nb_sidesouthwest	-0.372*	0.262	-0.460**	0.858***	-0.268
	(0.215)	(0.421)	(0.214)	(0.328)	(0.215)
nb_sidewest	-1.648***	-0.622	-1.484***	-0.488	-1.406***
	(0.195)	(0.394)	(0.193)	(0.312)	(0.195)
seasonfall	0.198**	0.019	0.162*	0.235	0.131
	(0.099)	(0.201)	(0.097)	(0.146)	(0.098)
seasonspring	-0.019	0.016	-0.158*	0.149	0.002
	(0.095)	(0.190)	(0.093)	(0.140)	(0.094)
seasonsummer	0.170*	0.289	0.177*	0.436***	0.217**
	(0.097)	(0.186)	(0.094)	(0.139)	(0.095)
Constant	2.187***	-1.320***	2.373***	-0.815***	2.195***
	(0.198)	(0.398)	(0.197)	(0.316)	(0.198)
Akaike Inf. Crit.	55,495.890	55,495.890	55,495.890	55,495.890	55,495.890

Note: *p<0.1; **p<0.05; ***p<0.01

Methods and Results: We selected variables that would most likely have an effect on predicting the `crime_category` for the crime committed. The variables eliminated by nonselection include case number, description of crime, reporting code, specific location of crime and police location. These variables were eliminated because we reasonably assumed that they would have no bearing on the crime category of the crime committed. Some of these variables are accounted for through sub-categories previously created such

as `nb_side` that details the informal city side that the crime occurred. We created a multinomial regression model because the dependent variable, `crime_category`, is dependent. Our multinomial regression model explores the interaction of neighborhood side and weather season because we wanted to see how situational location can predict the crime category of the crime committed, answering our research question by including all the variables we are interested in.

Evaluation of Multinomial Logit Model: The table displays the `logit` coefficients relative to the reference category. For example, under theft/robbery, the 0.99 suggests that for a one unit increase in theft/robbery, the logit coefficient for `nb_side central` relative to `nb_side northwest` will go up by 0.99. In other words, if the theft/robbery number increase by 1, the chances of staying in the central side are higher compared to staying in the northwest side.

Thus, for every additional theft/robbery is committed, the chances of staying in the central, far north, north side are higher compared to staying in the northwest side and the chances of being in the fall and summer time are higher compared to those of the wintertime.

	Dependent variable:				
	other	sex offense	theft/robbery	violation	violent crime
	(1)	(2)	(3)	(4)	(5)
nb_sidecentral	1.492	2.290*	2.702***	1.664	1.046
	(0.261)	(0.468)	(0.258)	(0.393)	(0.262)
nb_sidefar north	2.012**	3.560***	2.275***	2.241**	1.742**
	(0.276)	(0.468)	(0.274)	(0.401)	(0.276)
nb_sidefar southeast	0.638**	1.275	0.592**	2.474***	0.752
	(0.217)	(0.424)	(0.216)	(0.329)	(0.216)
nb_sidefar southwest	0.757	0.602	0.626*	2.298**	0.798
	(0.241)	(0.523)	(0.240)	(0.355)	(0.240)
nb_sidenorth	1.944**	2.432*	3.040***	1.463	1.251
	(0.293)	(0.508)	(0.290)	(0.445)	(0.295)
nb_sidesouth	0.633**	0.974	0.647**	1.252	0.792
	(0.214)	(0.427)	(0.212)	(0.335)	(0.213)
nb_sidesouthwest	0.689*	1.300	0.631**	2.359***	0.765
	(0.215)	(0.421)	(0.214)	(0.328)	(0.215)
nb_sidewest	0.192***	0.537	0.227***	0.614	0.245***
	(0.195)	(0.394)	(0.193)	(0.312)	(0.195)
seasonfall	1.218**	1.019	1.176*	1.265	1.140
	(0.099)	(0.201)	(0.097)	(0.146)	(0.098)
seasonspring	0.982	1.016	0.853*	1.160	1.002
	(0.095)	(0.190)	(0.093)	(0.140)	(0.094)
seasonsummer	1.185*	1.335	1.194*	1.546***	1.243**
	(0.097)	(0.186)	(0.094)	(0.139)	(0.095)

```
Constant              8.912***   0.267***     10.724***    0.443***    8.977***
                      (0.198)     (0.398)      (0.197)     (0.316)      (0.198)

--------------------------------------------------------------------------------
Akaike Inf. Crit.    55,495.890 55,495.890   55,495.890   55,495.890  55,495.890
================================================================================
Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

```
# A tibble: 12 x 3
   y.level       term                  estimate
   <chr>         <chr>                    <dbl>
 1 theft/robbery (Intercept)             10.7
 2 theft/robbery nb_sidecentral           2.70
 3 theft/robbery nb_sidefar north         2.28
 4 theft/robbery nb_sidefar southeast     0.592
 5 theft/robbery nb_sidefar southwest     0.626
 6 theft/robbery nb_sidenorth             3.04
 7 theft/robbery nb_sidesouth             0.647
 8 theft/robbery nb_sidesouthwest         0.631
 9 theft/robbery nb_sidewest              0.227
10 theft/robbery seasonfall               1.18
11 theft/robbery seasonspring             0.854
12 theft/robbery seasonsummer             1.19
```

```
# A tibble: 12 x 3
   y.level       term                  estimate
   <chr>         <chr>                    <dbl>
 1 violent crime (Intercept)              8.98
 2 violent crime nb_sidecentral           1.05
 3 violent crime nb_sidefar north         1.74
 4 violent crime nb_sidefar southeast     0.752
 5 violent crime nb_sidefar southwest     0.798
 6 violent crime nb_sidenorth             1.25
 7 violent crime nb_sidesouth             0.792
 8 violent crime nb_sidesouthwest         0.765
 9 violent crime nb_sidewest              0.245
10 violent crime seasonfall               1.14
11 violent crime seasonspring             1.00
12 violent crime seasonsummer             1.24
```

Methods and Results: We created a relative risk ratios table because relative risk ratios allow an easier interpretation of the `logit` coefficients. They are the exponentiated value of the `logit` coefficients.

Theft/Robbery Relative Risk Ratios:

Because you can get the significance of the coefficients using the `stargazer()` function from the package `stargazer`, we utilized it to analyze the p-values of the different variables. For theft/robbery all of the nb_sides compared to the northwest side were statistically significant at the $\alpha = 0.1$ significance level. In addition, the seasons fall, spring, and summer compared to the winter baseline were also statistically significant at $\alpha = 0.1$ for theft/robbery. For violent crime, the coefficients for `nb_side` far north, `nb_side` west, and season summer were the only variables that were statistically significant.

Keeping all else constant, if `nb_side central` increases by one unit, theft/robbery is 2.7 times more likely as compared to drug-related crimes. This coefficient is statistically significant. Keeping all else constant, if `nb_side north` increases by one unit, theft/robbery is 3.04 times more likely and staying in the same category is 1.194 as likely if the season is summer, as compared to drug-related crimes. The drug-related

category is the baseline for relative risk ratios because it has the least cases of crimes such that all of the other crime categories are compared to the minimal baseline. All variables in the optimal relative risk ratio for theft/robbery are significant at the $\alpha = 0.1$ significance level. Hence, theft/robbery is most likely to be the crime category if the neighborhood side is in the north and the crime occurred in the summer compared to the drug-related crimes.

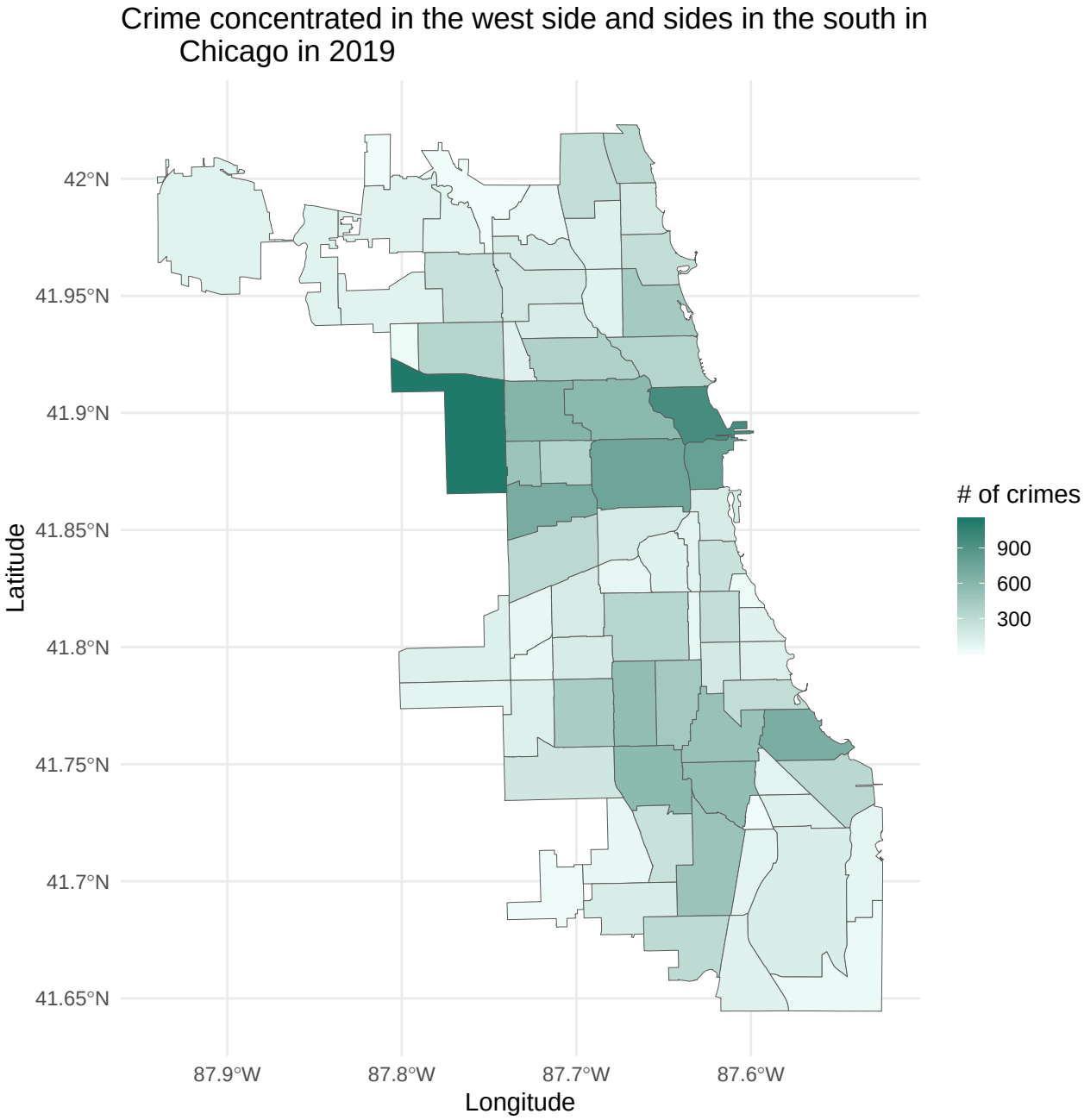
`Theft/Robbery: 10.724(intercept) + 3.040(nb_side north) + 1.194(season summer)`

Violent Crime Relative Risk Ratios:

Keeping all else constant, if `nb_side far north` increases by one unit, on average, violent crime is 1.74 times more likely as compared to the drug-related category. When all else is held constant, violent crime is 1.24 times more likely compared to drug-related crimes when the season is summer. The coefficients for `nb_side` far north and season summer are both significant at the $\alpha = 0.05$ significance level. Hence, violent crime is most likely to be the crime category if the neighborhood side is in the far north and the crime occurred in the summer compared to the drug-related crimes.

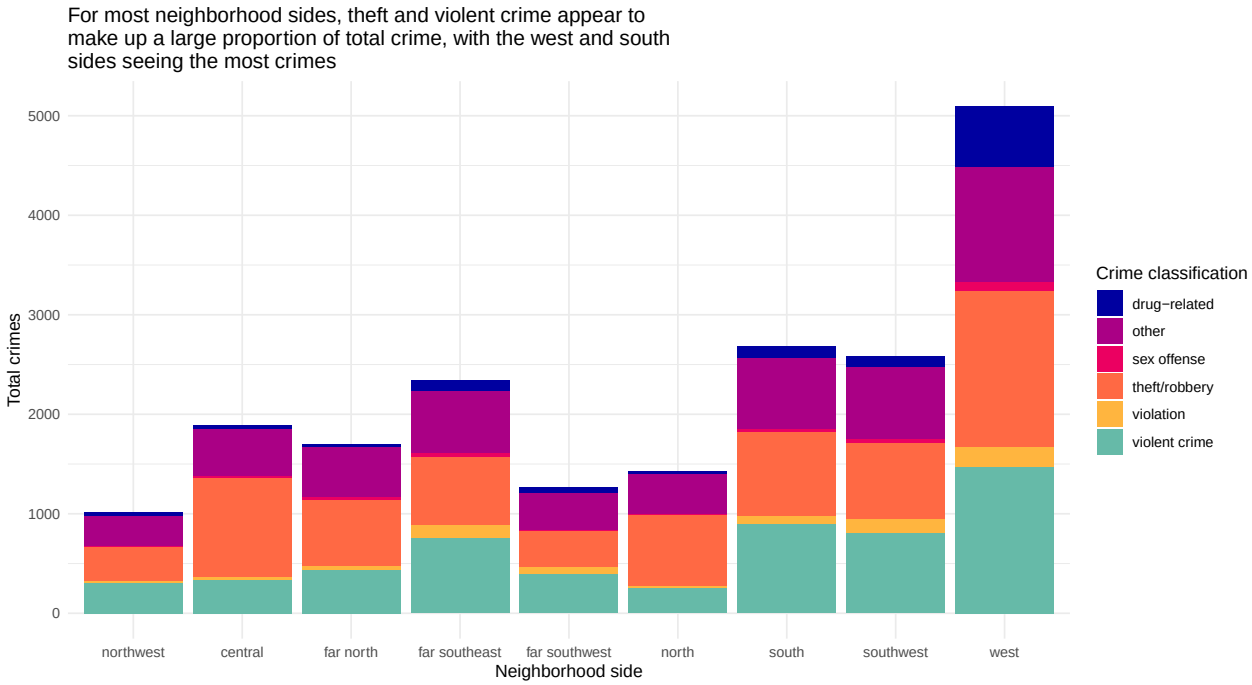
`Viol_crime: 8.977(intercept) + 1.742(nb_sidefar north) + 1.243(seasonsummer)`

Spatial analysis & visualization of the City of Chicago



Methods and Results: We created a spatial visualization using the Chicago City map sourced from the City of Chicago Crimes Data Portal which shows areas where crime occurred more frequently in darker shades of green, whereas less crime-ridden neighborhoods are in lighter green shades. We used latitude-longitude combinational values that were in the original dataset to create this spatial visualization. We thought this was the best method of analysis because it is helpful to geographically visualize Chicago and be able to isolate the areas with higher rates of crime, while being able to identify pockets of crime and possible socio-economic and political reasons for this. The highest number of crimes is concentrated in the community areas on the west side of Chicago, and patches of darker green can also be seen towards the north side and in parts of the southside.

Prevalence of each crime type in different Chicago neighborhood sides



Methods and Results: We created a filled bar graph to demonstrate how much each of the six crime classifications we created contributed to the total count of crime in a particular area. This graph's purpose is twofold - it demonstrates the total number of crime in each neighborhood side while displaying the breakdown of different crime types in that particular area. The colors and scale of the graph allow for easy comparison between the groups. From this visualization, we can see that for most neighborhood sides, theft (or robbery) and violent crime account for a large proportion of total crime in that area (equal to or greater than 50% of the total, which can be seen from the block sizes of these types from the graph). Moreover, drug-related crimes and sex offenses are seen to be much less prevalent.

Section 5: Discussion

When viewing the crime data and statistics in Chicago, there are many conclusions that can be drawn. Primarily, the two underlying narratives that can be found within our research is that the summer months experience significant increases in crime and that the West Side and South Chicago have the highest crime rates in Chicago.

Looking at the data and the month crimes occurred in, we can see that crime rates peaked in July and were the lowest in February. This is most likely due to more people being outside in the summer months, particularly for recreational or social reasons. A potential dangerous implication when analyzing this point is climate change. When crime is more prevalent during times where the average monthly temperature is warmer, even a city known to be quite cold in the winter like Chicago can have the possibility of increasing crime rates. As the status of climate change is ever changing, particularly with the current global pandemic COVID-19 leading to notable improvements in environments throughout the world, it is important to also take into account the lesser obvious consequences of our ecological footprint.

The other significant trend seen throughout our visualizations and calculations is where the crimes in Chicago are concetrated. It seems as though neighborhoods in the West Side and South Chicago are significantly higher in crime rate than other sides. Another trend that can be noted when looking at the geographical density of crime in Chicago is that the Sides in the southern half of Chicago tend to be higher than the sides in the north. A possible explanation for this is gang activity as gangs tend to be concentrated towards a

specific area, as one of our group members found in the New York City data that we explored earlier in the semester (Brooklyn and Bronx having significantly more Stop, Question, Frisk cases than the other boroughs). Another possible explanation is that neighborhoods like Wicker Park that are known to be trendy having more dense groups of people. This would certainly increase the potential for crime in these areas.

The results of our findings could potentially speak volumes on how crime could be dealt with in Chicago. This could aid or at least direct police departments to either hire more police officers in the months leading up to the summer months or to better equip or train the existing police officers in each city's police department. The specific graphs and conclusions we were able to draw can also potentially aid in investigating why crime is committed and recorded in certain community areas and know what specific parts of these community areas need monitoring. For example, we found that the most common location for crimes in Chicago were the streets and residential areas. This in conjunction with the specific map visualization of crime density for each community area can help the Chicago Police Department concentrate and prioritize officers in those locations.

Our analysis was quite in depth, requiring our group to do extensive external research in producing effective and narrative-driving visualizations and tables. As our data included so many different variables that were vital for extracting meaningful conclusions, we used linear models and multinomial regression. Using these methods, we were able to conclude that the data is not affected by Simpson's Paradox as location (both side/community area and location such as street) and month were large contributing factors to crime rates in Chicago. We were able to conclude this by taking into account different factors that could play into location being superficially viewed as being significant, in particular by viewing type of crime, the number of crimes, and percentage of overall crimes.

In addition to this, we used the stargazer package in order to analyze the p-values of the different variables. Through this, we found that for a crime, if the neighborhood is in the North Side and occurs in the summer, it is most likely a crime in the theft/robbery category compared to drug-related crimes. Similarly, a crime in the Far North Side and occurs in the summer, it is most likely a crime in the violent crime category compared to drug-related crimes.

There were several issues that we came across while trying to use this data set. Firstly, we had access to very few numerical variables in the data, as most of the variables were categorical. This created some difficulty in creating the most accurate numerical conclusions and could be a bit generalized as each crime case does not contain the full specific detail as a numerical dataset would. Secondly, we generalized the months into seasons in Chicago but the weather in Chicago can be a bit erratic. We also had limited hypothesis testing so we had no tests for statistical significance for some of our implicit hypotheses. Another possible problem in viewing this dataset is that this data set pertains to one specific year, so it might paint a particular picture that might have been specific to 2019 rather than Chicago as a whole. For example, it would be unfair to say that the hospital death rate in New York City during 2020 is high when contextually, it stemmed from the COVID-19 pandemic. Although we are able to generalize and make conclusions on the given data, as crime is still crime, the raw data just simply cannot explain all of the contextual background information pertaining to Chicago.

If we were to continue researching this topic, it would be interesting to view the data across multiple years to see greater trends and validify some of our claims. This would help in alleviating the potential issue of contextual trends explaining some of the data. Similarly, we thought perhaps corroborating this data with other data such as socioeconomic trends, gun legislation, or population density would aid in understanding Chicago crime as a whole. Another aspect that our group thought would be interesting to explore was if there was a correlation between the types of crime and month of the crime, as we could imagine that violent crimes, thefts, and violations would be higher in the summer months as they are inherently more involved with being outside and coming into contact with other people, causing crime in the form of as robbery, arson and public peace violation. Lastly, tracking a specific crime type, such as violence, over many years could be interesting to see as this would reduce the amount of variables and give more strength in concluding some trends in Chicago. Furthermore, we can compare crime rates and types in different major US cities with similar socio-economic factors (such as income inequality and gentrification), and see whether certain times of the year favor higher crime in more cities than just Chicago. Finally, it would be extremely interesting to conduct an analysis on the impact of tourism on crime rates, where we could combine tourism data and data on crimes to analyze relationships. As a result, this study can be taken further in many directions, each of

which would be very unique and compelling, especially since crime has numerous underlying social, economic, political, and historical factors at play.

Citations & References:

Torres-Reyna, Oscar. "Logit, Probit and Multinomial Logit Models in R," n.d., 27. https://www.princeton.edu/~otorres/LogitR101.pdf

"Community Areas in Chicago." In Wikipedia, April 23, 2020. https://en.wikipedia.org/w/index.php?title=Community_areas_in_Chicago&oldid= 952586757.

Bryan, Jenny, and The STAT 545 TAs. "Chapter 15 Join Two Tables | STAT 545." Accessed April 28, 2020. https://stat545.com/.

Grolemund, Garrett, and Hadley Wickham. "16 Dates and Times | R for Data Science." Accessed April 28, 2020. https://r4ds.had.co.nz/.

"Multinomial Regression." Accessed April 26, 2020. http://dwoll.de/rexrepos/posts/regressionMultinom.html#using-multinom-from- package-nnet.

Cross Validated. "Regression - Getting p-Values for 'Multinom' in R (Nnet Package)." Accessed April 26, 2020. https://stats.stackexchange.com/questions/63222/getting-p-values-for-multinom- in-r-nnet-package.

CRAN nnet Documentation: https://cran.r-project.org/web/packages/nnet/nnet.pdf

Bennett, Trevor; Brookman, Fiona (August 25, 2008). "The Role of Violence in Street Crime: A Qualitative Study of Violent Offenders". International Journal of Offender Therapy and Comparative Criminology.

Lisowska, Agnieszka (2017). "Crime in Tourism Destinations: Research Review." Tourism 2017, 27/1. Retrieved from http://dx.doi.org/10.18778/0867-5856.27.1.12

Tiihonen, J., Halonen, P., Tiihonen, L. et al. "The Association of Ambient Temperature and Violent Crime." Sci Rep 7, 6543 (2017). https://doi.org/10.1038/s41598-017-06720-z

Asher, Jeff. "A Rise in Murder? Let's Talk About the Weather." The New York Times, September 21, 2018, sec. The Upshot. https://www.nytimes.com/2018/09/21/upshot/a-rise-in-murder-lets-talk-about-the-weather.html

Authority, Independent Police Review. "The Most Dangerous Neighborhood, the Most Inexperienced Cops." The Marshall Project, September 21, 2016. https://www.themarshallproject.org/2016/09/20/the-most-dangerous-neighborhood- the-most-inexperienced-cops.

"Defining Seasons." Accessed April 24, 2020. https://www.timeanddate.com/calendar/aboutseasons.html.

"Multinomial Logistic Regression." UCLA Statistics Department. Accessed April 28, 2020. https://stats.idre.ucla.edu/stata/output/multinomial-logistic-regression-2/.