Analysis of Weapons Violation and Homicide Crimes in Chicago from 2010 to 2019

By: Bao Van and Aryan Kamath

I. Introduction and Problem

Chicago is one of the biggest cities in the U.S. and is one its top 3 most populated cities with a population of 2,716,450 in 2020. Because of this and various other socioeconomic factors, the number of crimes that occur in the city is significantly large. As the FBI 2019 Uniform Crime Report showed, the number of reported crimes is about 3,935 per 100,000 population. Which also means that the number of crimes is roughly around 106,892 in 2019. These large numbers cannot be ignored. It has been shown to cause damage to the community area and the neighborhoods. The data we looked at came from data.cityofchicago.org. It is an open data repository for the Chicago city area that everyone can access. By doing this project, we are hoping to find a pattern in crime occurrences over time that will help the Chicago Police Department.. More specifically, we wanted to look at crime which is most likely to result in death and gain an understanding as to how this criminal activity has changed over the course of a decade. To do this task, we have to do a lot of data cleaning and adding some new variables that will be important for analysis. We also wanted to see if we could predict the pattern for 2020 from what we already had seen from 2010 to 2019.

II. Chicago Crime Dataset and Cleaning Data

As we mentioned in the previous paragraph, our dataset came from data.cityofchicago.org. The data itself is a very well-documented dataset that almost has everything we want to analyze. Since the data in 2020 is not completed yet and there are multiple rows that are missing information, we decided to analyze from 2010 to 2019 to see if there are any interesting points during the 10-year bracket. The dataset contains 22 variables and 2,972,554 observations with detailed information such as Date, Type of Crime, and the locations of crime occurrence, etc. However, we do not end up using that many variables since our focus is on the number of crimes itself. We decided to delete columns that will not be useful to our analysis by using the DROP = command on all extraneous columns, and what we have left are Date, Primary Type, Location Description, Year, Longitude, Latitude, and datevar-which is a new variable that we filtered out from Date variable, since we are uninterested in the time that the Date variable contains. This is relevant as it greatly trims the data set and reduces the runtime of our code. (Figure 1.1)

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
1	Date	Num	8	DATETIME.	ANYDTDTM40.
4	Latitude	Num	8	BEST12.	BEST32.
5	Longitude	Num	8	BEST12.	BEST32.
2	Primary Type	Char	26	$26.	$26.
3	Year	Num	8	BEST12.	BEST32.
6	datevar	Num	8	DDMMYY10.	

Observations	2972554
Variables	6
Indexes	0
Observation Length	72
Deleted Observations	0
Compressed	NO
Sorted	NO

Figure 1.1: Dataset contents after cleaning

III. Analysis

Before doing anything, we were particularly interested in determining what type of crime was most common. In exploring the dataset, we discovered that the most common crime in Chicago was theft with 22.59 percent of all crime. After that is battery with 18.04 percent and criminal damage with 10.62 percent (Figure 2.1).

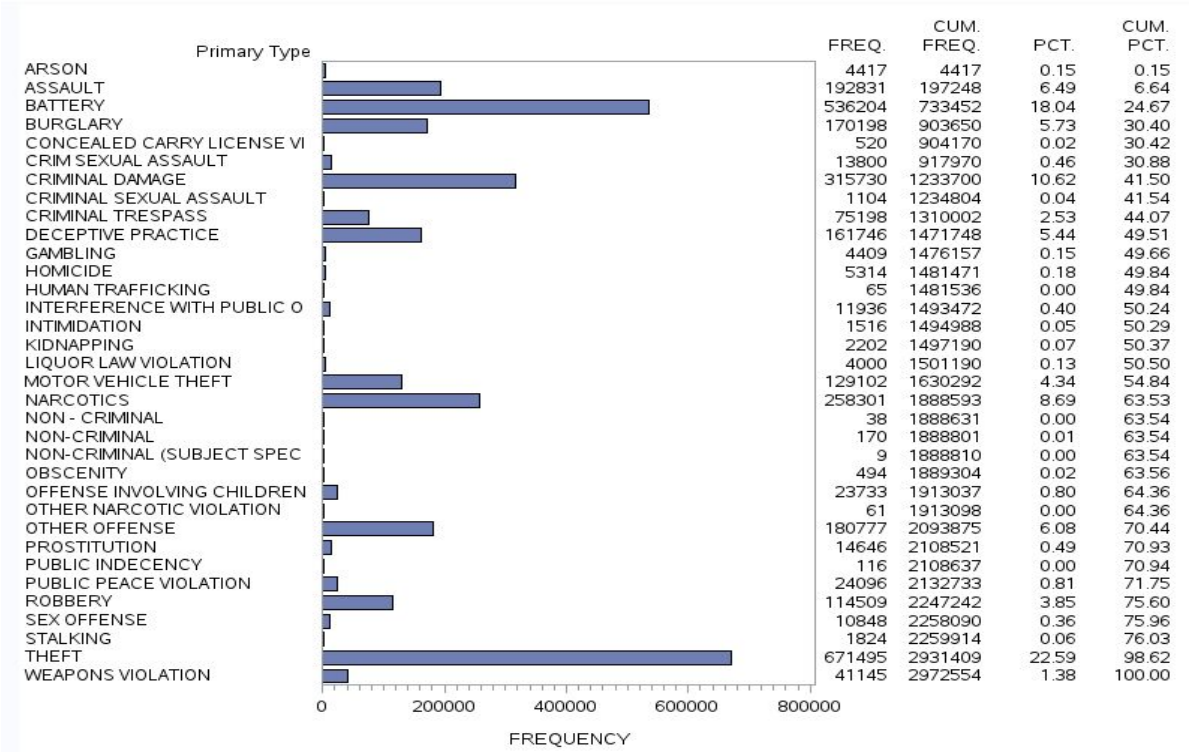
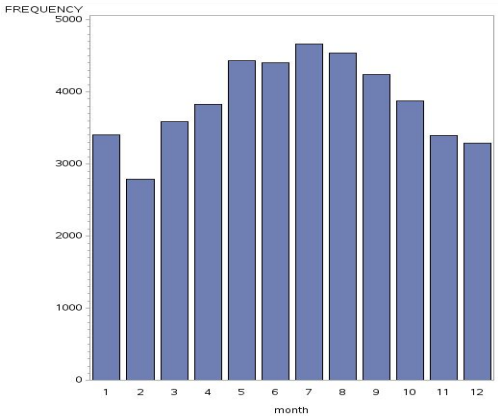


Figure 2.1: Frequency of Each Primary Type

We agreed that stealing and physical harm to others is bad and they appear to happen a lot, however, they are not as critical to determine how dangerous that area/neighborhood is when compared to crimes that result in death. In fact, weapons violation and homicide are the ones that concern us the most since they are the two most likely to result in death. We decided to filter out the dataset such that there are only weapons violation and homicide crime types left using. From this, we are interested where these two types happened the most. It turns out on the street is the most common with 36.11 percent and sidewalk is the second most common with 16.26 percent (Table 1). This result is not only showing the number itself but also showing that even streets and sidewalks have lots of people around but people still try to commit homicide and weapons violations.

We were also interested in when these crimes occurred. We wanted to know which month that crime occurs the most. We determined that the least amount of crime was in February while the most crimes were reported in July (Figure 2.2). This figure shows the total number of crimes that occured per month over the course of decade. So month 1 shows the total number of weapons violations and homicides that occurred in January 2010 plus January 2011 all the way to January 2019.



Obs	month	COUNT	PERCENT
1	1	3409	7.3377
2	2	2794	6.0139
3	3	3586	7.7186
4	4	3824	8.2309
5	5	4435	9.5461
6	6	4404	9.4793
7	7	4664	10.0390
8	8	4536	9.7634
9	9	4245	9.1371
10	10	3874	8.3385
11	11	3396	7.3097
12	12	3292	7.0858

Figure 2.2: Total Number of Crimes for each Month Throughout 2010-2019

From Figure 2.2 this result makes sense since people are more likely to be out in the summer so there are more opportunities for crime to occur. However, the only one month that acts irregularly is February. Besides the correlation between heat and criminal activities, is that December and January have higher totals than February, even though they have lower temperatures (rssweather.com). Or another reason could be that February has shorter days than December and January, adding with no holidays as big as New Year and Christmas.

We also want to know at which time crime occurs the most, or at which time is the most dangerous time to go outside. To do that, we used R code to generate a heatmap of week day vs. time (Figure 2.3)

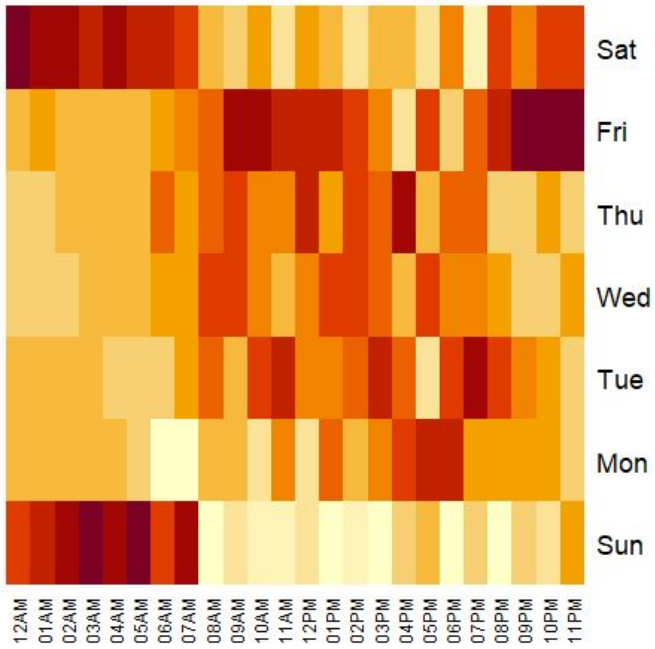


Figure 2.3: Heatmap of Week-day vs Time

As we can see the least amount of crime was reported on Sunday (8:00am - 8:00pm). In a mild level, there are three time slots: Monday - Friday (12:00am - 7:00am), Monday - Thursday (9:00pm - 11:00pm) and Saturday (8:00am - 7:00pm). The most dangerous time slots are: Friday 9:00am to Saturday 7:00am and Saturday 8:00pm - Sunday 7:00am. The rest of the time slots are relatively unpredictable, there is no clear pattern other than a spike in crime around 3:00PM to 7:00PM on workdays. This heat map, further shows the positive correlation between the amount of people on the street and the amount of crime.

In order to find the characteristics of these two types of crime, we need to look at them in more detail. First, we need to see the trends of all 10 years from 2010 to 2019. To do that we will use ANOVA test to analyze if the means of the number of crimes from 2010 to 2019 are all the same or not by inputting the total for each month for each year (Figure 2.4). This figure shows the average number of crimes (Weapons Violation and Homicide) per month in a year. So for 2010 each month had an average of 375 crimes occur.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	889969.575	98885.508	20.41	<.0001
Error	110	532962.417	4845.113		
Corrected Total	119	1422931.992			

R-Square	Coeff Var	Root MSE	crime Mean
0.625448	17.97891	69.60685	387.1583

Source	DF	Type I SS	Mean Square	F Value	Pr > F
year	9	889969.5750	98885.5083	20.41	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
year	9	889969.5750	98885.5083	20.41	<.0001

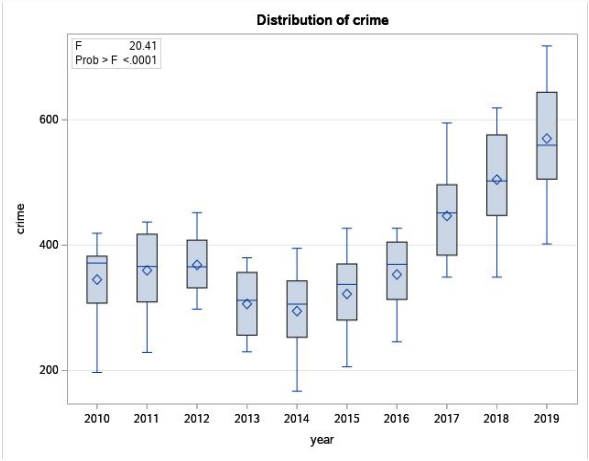


Figure 2.4: ANOVA Test for Means of Each Year

The ANOVA test results, shown in Figure 2.4, with 5% significance level and p-value less than 0.0001, there is at least one mean which is different between those 10 years. By looking at the "Distribution of crime" graph in Figure 2.4, we can see that the mean in 2018 is bigger than in 2010 but the mean in 2014 is smaller than in 2010. This proved that there are some fluctuations during those 10 years. Because of that, we are interested in looking at the big picture for each type of crime (Figure 2.5) to see if there is any recognized pattern or not. From Figure 2.5, we can see these points are randomly distributed and these points are apart from its mean value. Therefore, using a regular regression method to capture the trend for this dataset is not a good idea, there will be a lot of outlier points. Using time series analysis in this case will be a good idea because it will predict future values based on previously observed values.

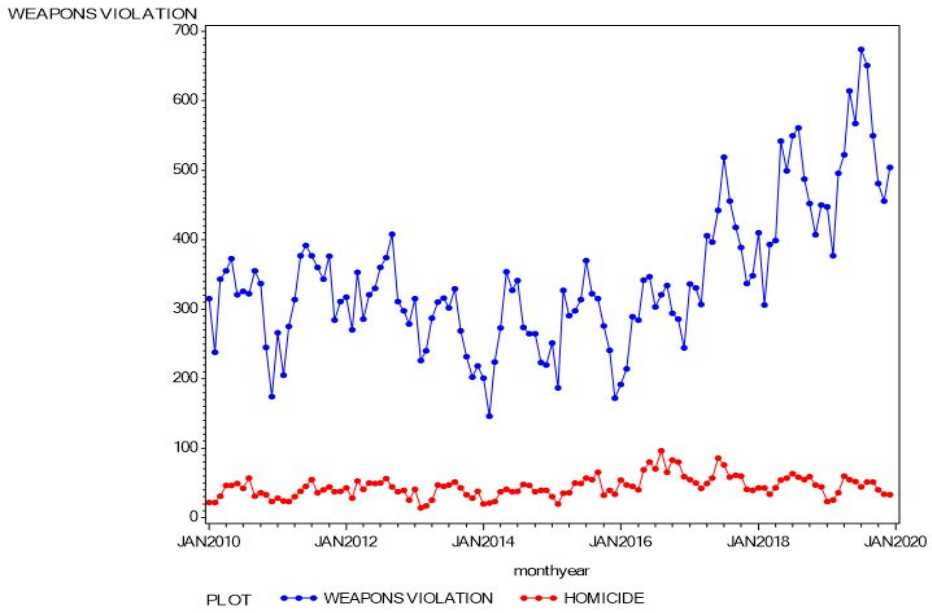


Figure 2.5: Plots of Total Weapons Violation and Homicide each month from 2010 to 2019

Figuring out which method to use when analyzing this data set was perhaps the hardest part of the project. Originally, we did a linear regression on Figure 2.5 to see whether the slope would be positive, negative or zero but that method of analysis was shoddy since crime is not best represented through a linear model. We then tried doing a linear regression on both the total number of crimes that happens per year as well as the average number of crimes that happens per month in a year. However even with this we run into a problem since there happened to be a crime spike after 2016 so the data is not truly linear as shown in Figure 2.6. So while this may have been a way to test for general trend lines in the data set we were looking for a more meaningful way to understand what was going on with our data.

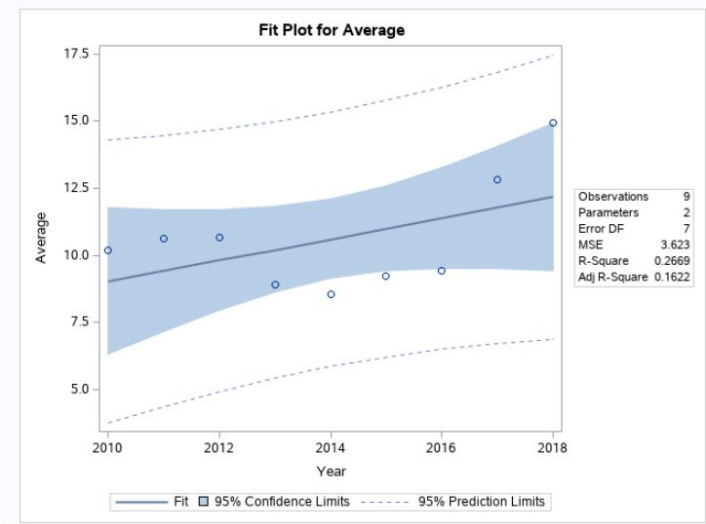


Figure 2.6: Average Weapons Violation per Month Across a Decade

Since using regular regression is not an optimal way to predict the dataset's characteristics. We need to use a time series model to help us to predict, and we want to see if the number of crimes will increase or decrease for each type of crime. We are going to use the ARIMA function in SAS to help us to create time series models. However, there are two challenges that we need to test whether or not the time series is reasonable. For Homicide crime, the first thing we need to test is if the data are white noise or predictable. With 5% in significance, and p-value is less than 0.0001 for each lag, we can say that the time series is not white noise (Figure 2.7a). The second thing we need to test is if there is any difference between fitted models and data. Again, the p-value is less than 0.0001 for each lag, so we can say there is no difference between fitted models and the data (Figure 2.7b). When we have all certified conditions, we can accept our time series model (Figure2.7c).

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	116.45	6	<.0001	0.689	0.534	0.339	0.224	0.113	0.039
12	212.85	12	<.0001	0.040	0.182	0.273	0.399	0.491	0.458
18	256.25	18	<.0001	0.413	0.279	0.131	-0.006	-0.107	-0.192
24	289.09	24	<.0001	-0.151	-0.092	0.071	0.171	0.277	0.276

Figure 2.7a: Autocorrelation Check for White Noise (Homicide)

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	6.73	3	0.0809	-0.048	0.153	-0.051	0.001	-0.075	-0.139
12	26.22	9	0.0019	-0.218	0.059	-0.022	0.112	0.260	0.125
18	42.14	15	0.0002	0.200	0.072	-0.005	-0.073	-0.075	-0.239
24	60.08	21	<.0001	-0.094	-0.180	0.087	0.037	0.210	0.161

Figure 2.7b: Autocorrelation Check of Residuals (Homicide)

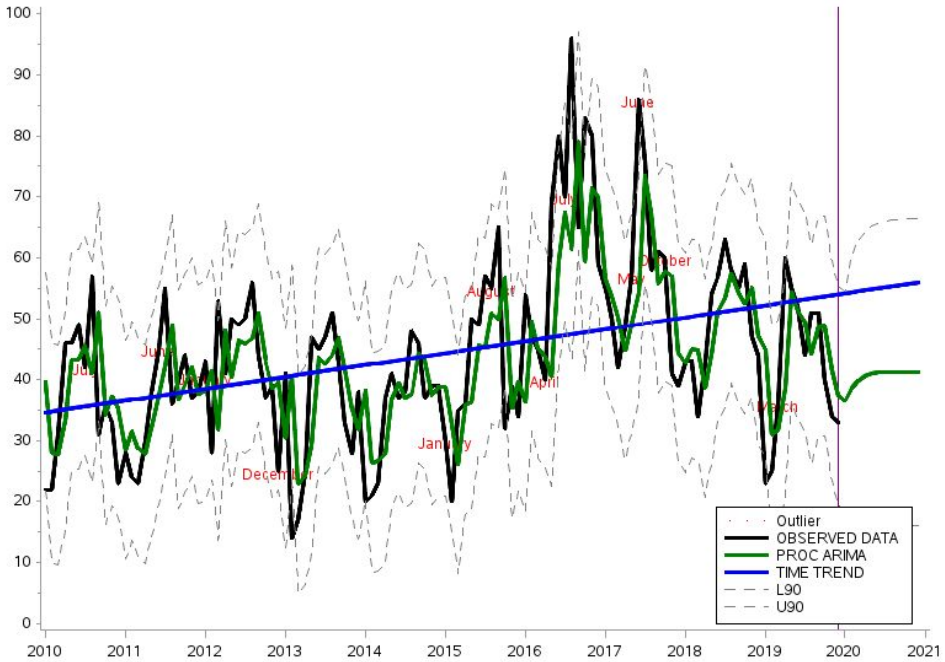


Figure 2.7c: Homicide from January 2010 to December 2019

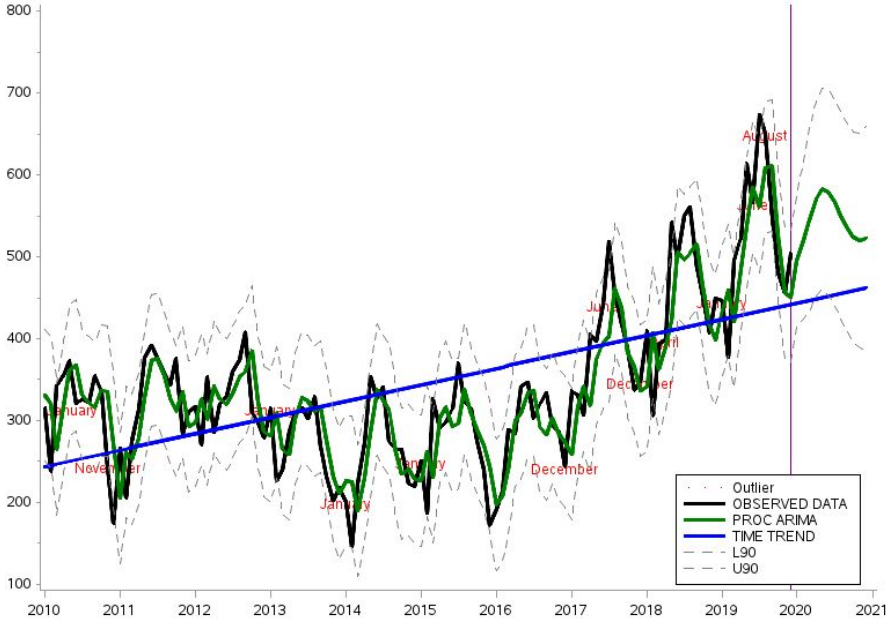
Forecasts for variable HOMICIDE				
Obs	Forecast	Std Error	90% Confidence Limits	
121	36.5293	10.9153	18.5752	54.4834
122	38.6247	13.0930	17.0887	60.1608
123	39.8570	14.0098	16.8129	62.9011
124	40.5699	14.4668	16.7741	64.3657
125	40.9708	14.7267	16.7475	65.1941
126	41.1846	14.8929	16.6880	65.6813
127	41.2868	15.0105	16.5967	65.9768
128	41.3225	15.1006	16.4842	66.1608
129	41.3192	15.1741	16.3601	66.2784
130	41.2933	15.2366	16.2314	66.3552
131	41.2547	15.2913	16.1027	66.4066
132	41.2092	15.3402	15.9768	66.4415

Figure 2.7d: Forecast for 2020 (Homicide)

As we can see from Figure 2.7d, it predicts that there will be around 485 Homicide crimes committed in 2020 total. We haven't really checked the dataset in 2020 because the year has not ended yet, but the result is interesting that we need to look at it in the future because it might be different by the COVID-19. For weapons violation crime, we can do the same process as above and here are the results:

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	285.37	6	<.0001	0.850	0.762	0.637	0.531	0.421	0.349
12	467.38	12	<.0001	0.348	0.385	0.438	0.508	0.560	0.577
18	554.22	18	<.0001	0.496	0.440	0.332	0.235	0.128	0.083
24	589.25	24	<.0001	0.074	0.097	0.150	0.226	0.265	0.272

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	.	0	.	-0.080	-0.034	0.002	0.088	-0.014	-0.044
12	14.95	3	0.0019	-0.035	0.054	-0.029	0.051	0.108	0.274
18	20.26	9	0.0164	-0.097	0.094	0.031	0.061	-0.119	-0.028
24	28.74	15	0.0174	-0.039	-0.018	-0.045	0.133	0.088	0.164



Forecasts for variable weapons				
Obs	Forecast	Std Error	90% Confidence Limits	
121	495.5851	48.7715	415.3632	575.8071
122	520.8001	58.7822	424.1121	617.4882
123	545.4009	67.8606	433.7801	657.0216
124	570.3429	72.3312	451.3686	689.3172
125	582.8887	74.7962	459.8600	705.9174
126	578.9734	76.0001	453.9643	703.9825
127	567.2766	76.3537	441.6859	692.8672
128	550.4183	76.6206	424.3887	676.4480
129	534.6871	77.0494	407.9520	661.4221
130	523.5345	78.0539	395.1473	651.9217
131	519.5440	79.9279	388.0743	651.0138
132	522.8788	82.6553	386.9230	658.8346

Figures 2.8: Results for Weapons Violations

With the prediction from Figure 2.8, we can see roughly around 6513 people will try to commit weapons violation crime in 2020, which is also a high number.

With the forecast values for 2020, the last thing we want to look at is we want to know what locations have a high rate of criminal activities or where are the dangerous areas in Chicago. We used R programming language to take all of the inputs from both homicide and weapons violation latitudes and longitudes from 2010 to 2019 (Figure 2.9)

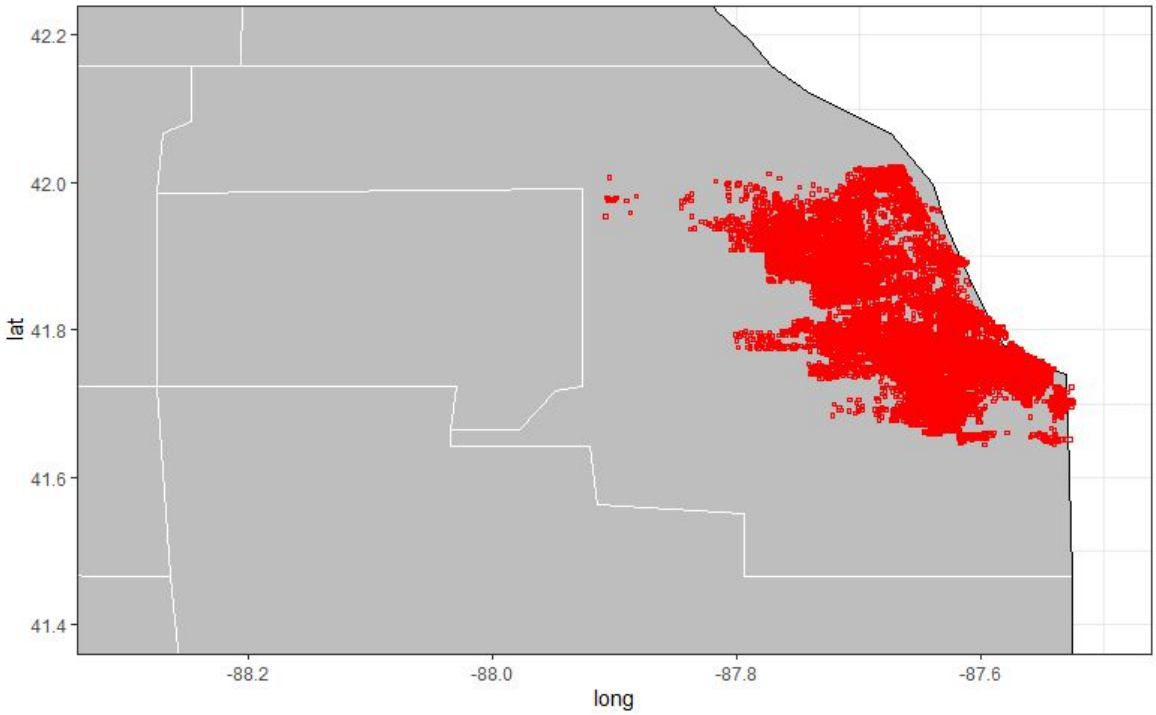


Figure 2.8: Homicide and Weapons Violation Locations From 2010 - 2019

As we expected, all of the red dots covered the whole Chicago map. This means that crimes have occurred everywhere in Chicago, but we can see a little scatter in the north part of Chicago. However, with this map, we cannot see much about the density so we decided to add a heatmap layer and use the Community Area variable instead of using Longitude and Latitude (Figure 2.9).

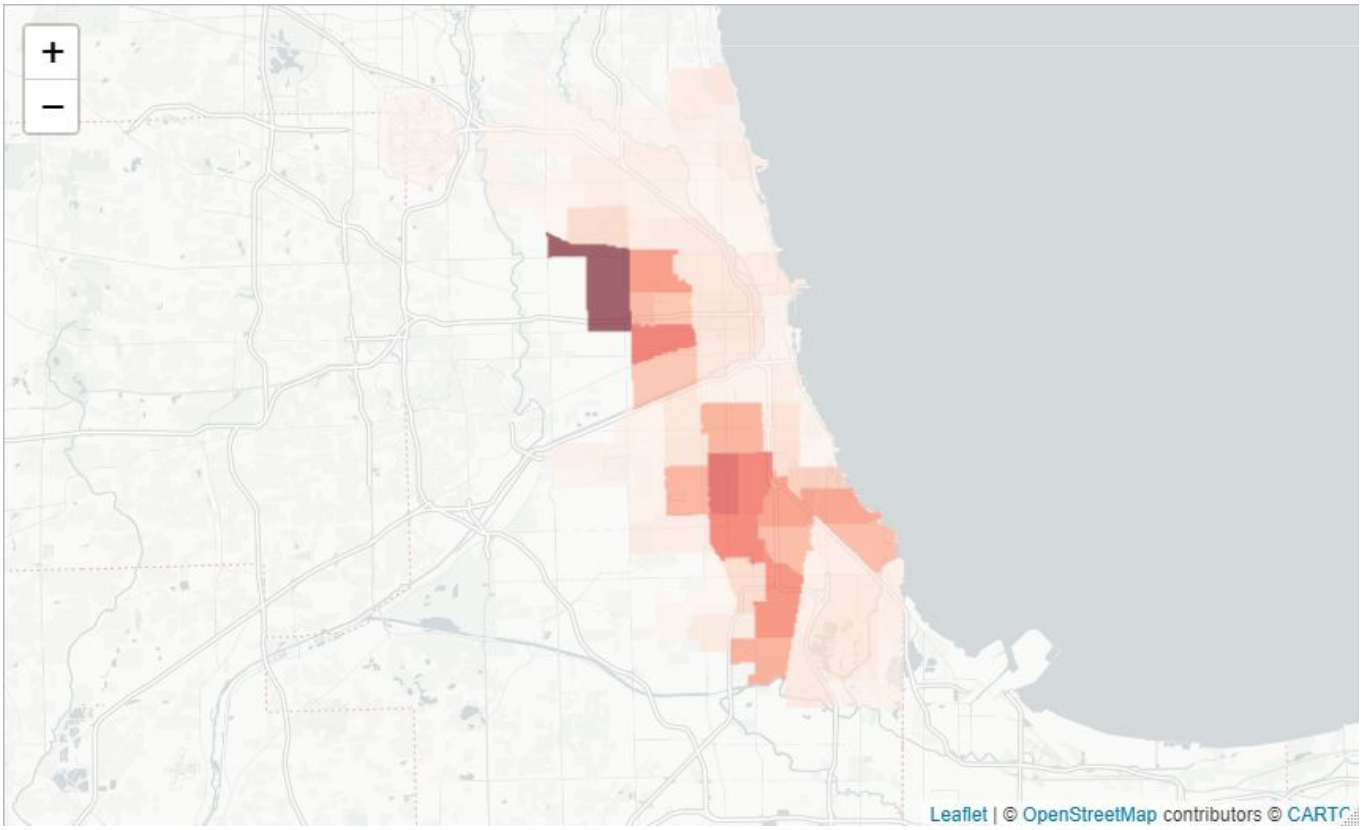
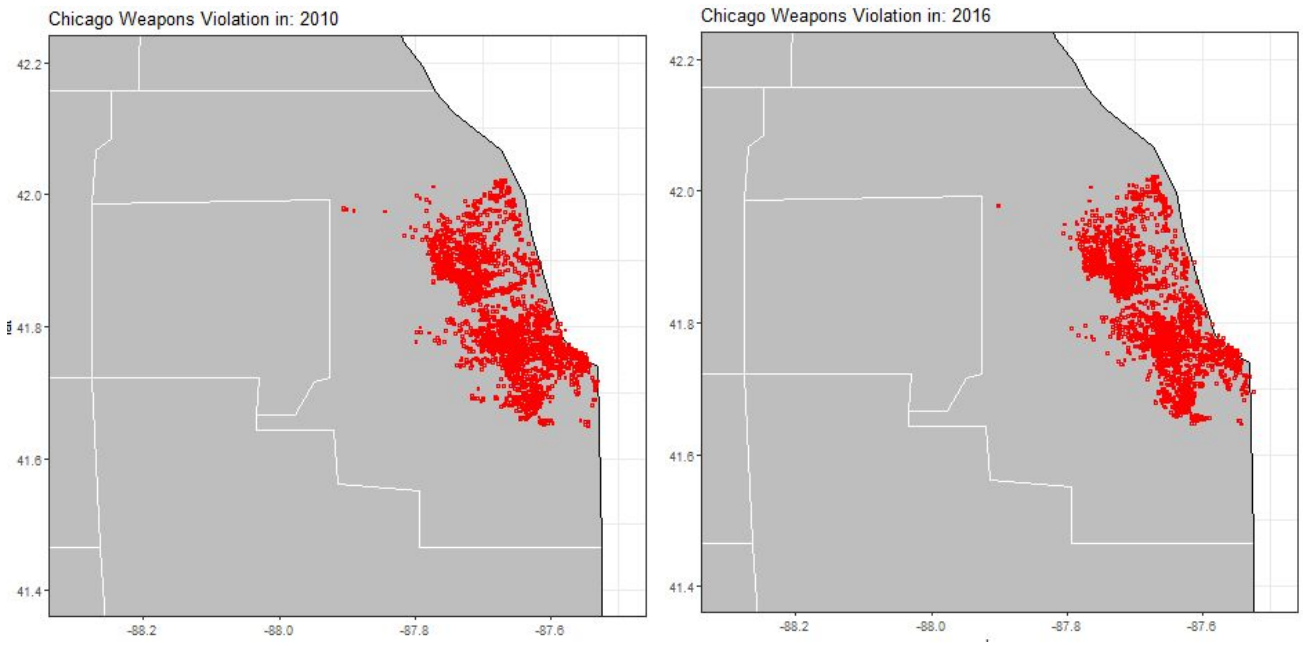


Figure 2.9: Chicago Heatmap of Homicide and Weapons Violation (2010-2019)

Based on Figure 2.9, there are two cluster areas: the most dangerous cluster is Community Area 25 (near Cicero area) and the other one is Community Area 67 (Southside Chicago area). These two areas are critical to Chicago public safety because years after years the population of Chicago will be larger, which means that more crimes will happen, they will start from these two clusters and might spread out to other areas. Because of that we want to see each crime for each year (Figure 2.10, Figure 2.11).



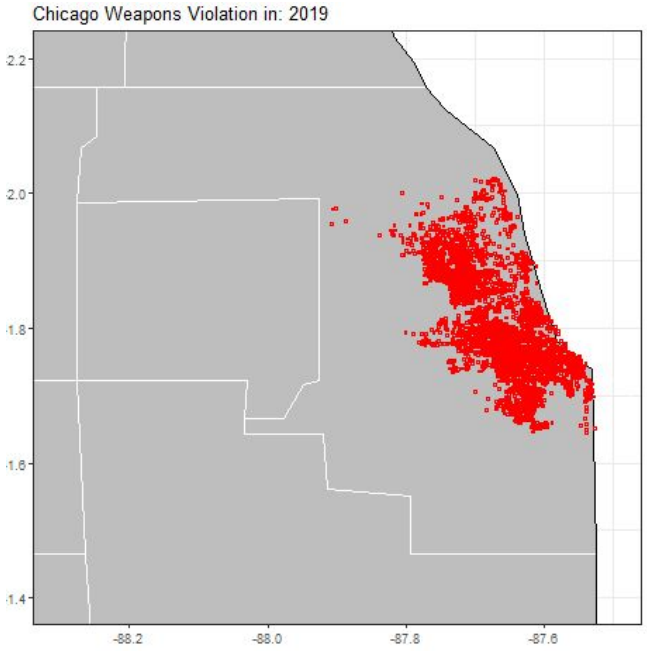
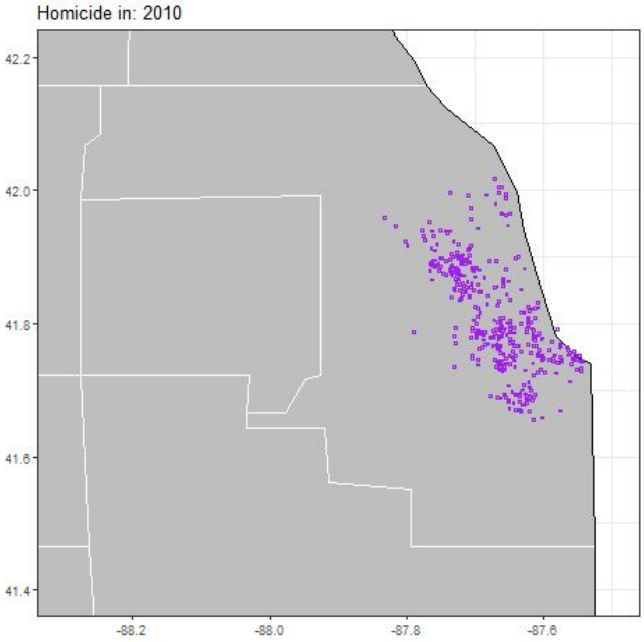
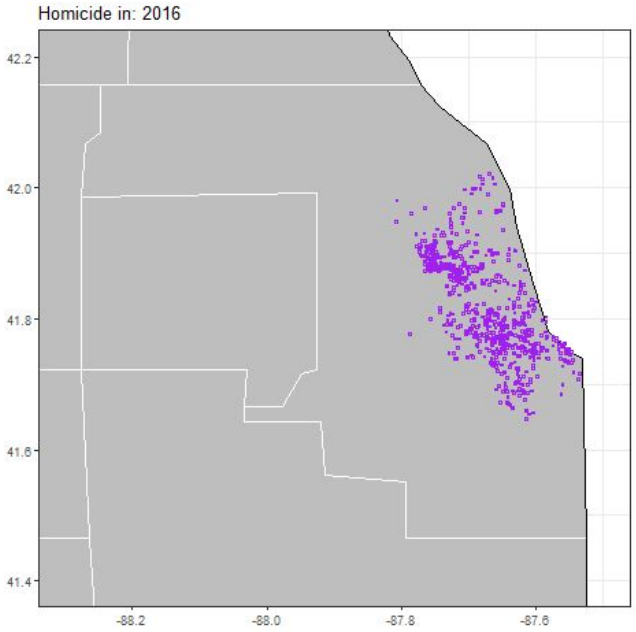


Figure 2.10: Elapsed Time of Weapons Violation (2010 - 2019)





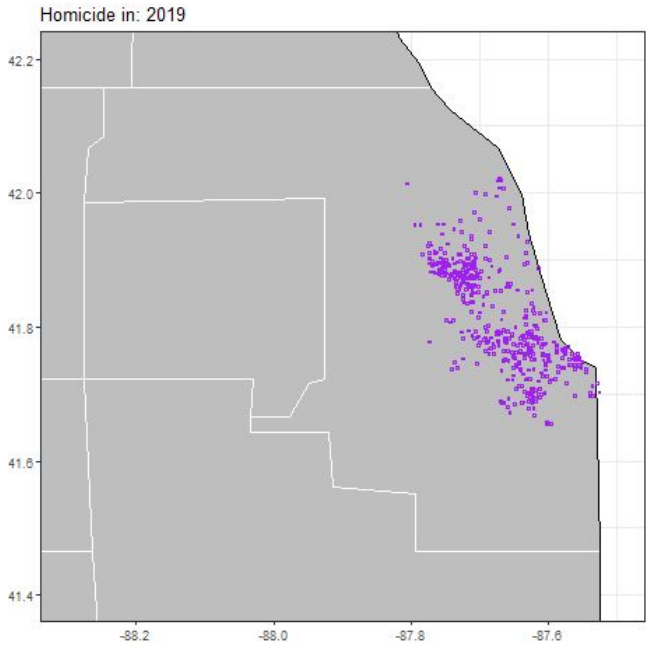


Figure 2.11: Elapsed Time of Homicide (2010 - 2019)

Based on these two figures, we can see that throughout the decade the number of crimes increased over time, and they are spreading (clearest to see in Figure 2.11). Even though these two types are not the majority crimes that are happening in Chicago, they do hold a critical type of crime that can bring down a whole community or a neighborhood.

IV. Conclusion

In this study, we found that crimes generally occurred in the Cicero and Southside of Chicago areas where there was more traffic, and especially were on the streets and sidewalks. Compared to a decade ago, the west side of Chicago has seen an increase in crime which we can see through the heatmaps. We also discovered there was less crime in the morning and more crime at night, especially those nights on Friday and Saturday. Another interesting thing we found out was that crimes spiked during the summer and tapered off during the fall and winter. So graphing crime throughout the decade lends itself to a sinusoidal like curve whose midlane changes based on if the average number of that crime is higher or lower than the year before it. We found that 2016 itself was an outlier in the sense that it was one of most violent years in Chicago history. From 2010-2014 a general trend of decreasing violent crime can be observed then it started to climb, peaking in 2016 and then decreasing again up until the present. So not only does violent crime have a sinusoidal tendency on a month by month basis but within the decade it shows the same pattern. That is not to say this would hold true over every decade, but it was an interesting fact to notice. This also made forecasting and prediction much more difficult than just doing a normal linear regression and led to us learning new techniques in SAS that helped us accomplish what we set out to do. Specifically the PROC ARIMA command and time series analysis in general were things that we learned along the way that go much farther in terms of analysis tools compared to a linear regression on averages or totals.

References

Zupko, W., & Smith, J. (2012). Creating and Displaying an Econometric Model Automatically. Retrieved from https://www.lexjansen.com/nesug/nesug12/sa/sa03.pdf