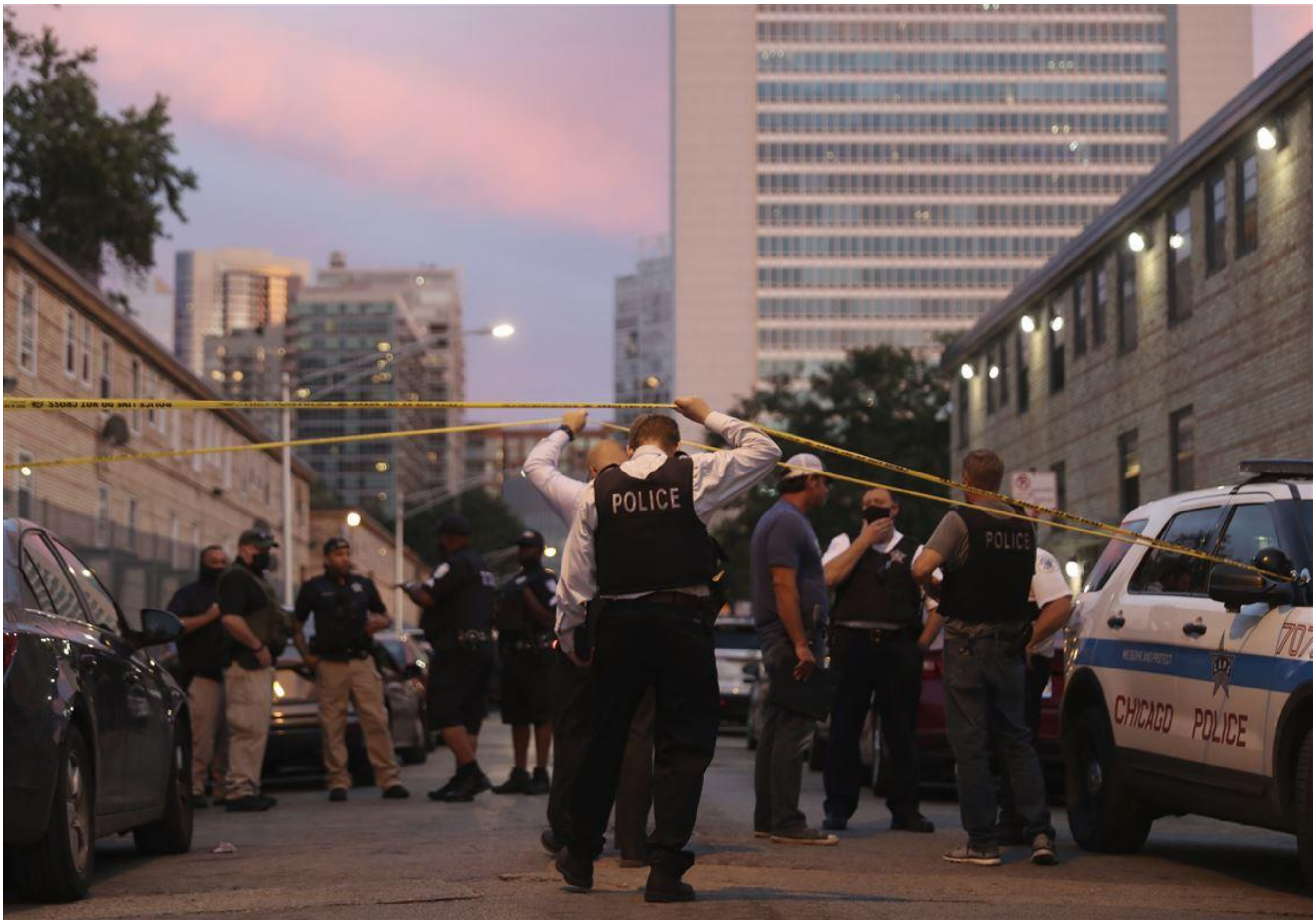
Chicago Crime Analysis



ISM6137 - Statistical Data Mining
Dr. Anol Bhattacherjee
MS-BAIS, Spring 2022

Group 3–
Amulya Alwala
Rajesh Makala
Adhithyan Rangarajan
Manish Talekar

Contents

Executive Summary

According to FBI, violent crimes are the offenses that involve force or threat of force/violence. Violent crime is the most serious form of crime as it has huge, sometimes unrecoverable implications for the victims and instills an overall sense of threat in the society. Usually, urban areas & popular tourist destinations like Seattle, Chicago, New Orleans etc., have seen a rise in violent crimes, particularly Chicago has seen a significant increase of 7.5% in violent crime from 2019 to 2021. Crimes also have a direct economic impact like medical needs for victims, police and correction costs, reduced home values and increased insurances etc. Hence preventing violent crime not only impacts the safety of the citizens, but also the overall development of a neighborhood/community. FBI also reports that less than 40% of violent crimes are reported in urban areas. Therefore, it is also practical to analyze possible factors and take necessary measures to prevent violent crime rather than being reactive to crime incidents. Analysis of crimes helped governments and law enforcement agencies with preventive measures such as predictive policing, which reduced crimes up to 12% in Manchester, London, and many other cities. Hence, the goal of this project is to identify factors that impact or drive crime in Chicago community areas and to recommend preventive measures that might help curb violent crime in the long term.

For this analysis, we chose data on community area level of Chicago city (which is divided into 77 community areas) from 2010 to 2020. The data has been sourced from Chicago government websites and used for GLMER modelling. Our key findings are increasing the number of public schools, health support agencies and encouraging high school education will help reduce crime in communities.

Problem Statement and Significance

In the United States, more than seven people per hour die a violent death. In 2019, more than 19,100 people were victims of homicide and over **47,500** people died by suicide. Coming to the areas in the U.S which are affected due to Violent Crime, it was observed that the average annual overall violent crime rate in urban areas was about **74% higher than the average rural rate and 37% higher than the average suburban rate**.

Police databases accumulate a large amount of data that could be analyzed to reduce crime rates. The analysis of criminal activity and the prediction of number of crimes remains one of the most interesting problems for the researchers. Police Departments & Governments across the world have been collecting data to analyze and understand how this massive problem could be cut down and how these violent crimes can be reduced to protect the lives of civilians, citizens and prevent loss to livelihood and the economy.

When it comes to Urban Crime, we performed analysis of the cities in the U.S which have the highest violent crimes and decided on City of Chicago for this project. The rate of violent crime in the Chicago metro area is 4.19 per 1,000 residents during a standard year. The Chicago total amount of daily crime is 1.98 times more than the Illinois average and 1.67 times more than that of the nation. As for violent crimes, the daily average in Chicago is 2.21 times more than the Illinois average, and it is 2.43 times more than the national average. The odds of being the victim of a violent crime in Chicago are 1 in 103. That's double the average chances for the state of Illinois. Crimes disrupt the development and the safety of the citizens. Murder, rape, and robbery are most common violent crimes in Chicago which have been reported by Chicago Police department. Chicago was responsible for nearly half of 2016's increase in homicides in the US, which shows the high levels of violent crimes in the city compared to other cities in the country.

Therefore, our analysis will be focused on understanding marginal effects on crime rate (community area-wise) of the metropolitan areas and we aim to suggest recommendations through this report to the Chicago PD and the state Government on how to reduce violent crimes in the City of Chicago.

Prior Literature

Table 1. Prior Literature Summary

Title	Important Predictors	Findings	Citation
Crime rate prediction in the urban environment using social factors	- GDP - Unemployment - Homelessness - Population Count - Police Stations/Precincts - Number of Schools, - Number of Bars, buildings	This study shows that unemployment, GDP, and homelessness were the biggest factors that increased urban crime in St. Petersburg in Russia	Varvara Ingilevich, Sergey Ivanov, Volume 136,2018
Crime prediction through urban metrics and statistical learning	- Unemployment - Illiteracy - Male population. - Child labor and homicides - Traffic accidents - Elderly population - Income	According to the findings in this paper, Unemployment, and illiteracy the biggest contributors to increase in crime at the urban level	Luiz G.A. Alves, Haroldo V. Ribeiro, Francisco A. Rodrigues, Volume 505,2018.
Using Geographically Weighted Regression to Explore Local Crime Patterns. Social Science Computer Review	- Heterogeneity index - ICE - Light-rail stop - Married families - Multiple land use - Population density - Residential stability - Single-person households	This study emphasizes the possible spatial variation in crime measures and their covariates by presenting a local analysis of crime using geographically weighted regression (GWR) and comparing the results to a global ordinary least squares (OLS) model	Cahill M, Mulligan G. 2007;25(2):174-193. doi:10.1177/0894439307298925
Property Crime Rates Forecasting with Economic Indicators	- Gross domestic product - Unemployment rate - Consumer price index	The purpose of this study is to introduce a hybrid model that combines support vector regression (SVR) and autoregressive integrated moving average (ARIMA) to be applied in crime rates forecasting.	(https://pdfs.semanticscholar.org/09d8/6d4d248ac9ca2d44e33077ece53
Alcohol abuse and crime: a fixed-effects regression analysis	- Alcohol abuse - Age - Deviant peer afflictions - Standard of living - Illicit drug abuse	This paper studies the effect of alcohol abuse on crime by considering both observed and non-observed confounding variables. Compares fixed effects regression model using fixed and time dynamic non observed confounders with Negative binomial regression model, controlling for observed confounders	David M. Fergusson, L. John Horwood, First published: 03 May 2002

Why Some Immigrant Neighborhoods Are Safer than Others	- Immigrant concentration - Neighborhood disadvantage - Residential instability - Percentage of young males	Studies the relationship between immigrant concentration and violent crime across neighborhoods, divergent findings of this correlation for Chicago (reduced crime rate), LA (uncreased crime rate).	Kubrin CE, Ishizawa H. The ANNALS of the American Academy of Political and Social Science. 2012;
Poisson-Based Regression Analysis of Aggregate Crime Rates	- Log population at risk - Residential instability - Ethnic heterogeneity - Female-headed households - Poverty Rate - Unemployment - Adjacent to metropolitan area	- This study talks about the use of Poisson based model for offense counts to analyze per capita offense rates and Poisson model will perform better than regression model.	Osgood, D.W. Journal of Quantitative Criminology 16, 21–43 (2000).
Race, economic inequality, and violent crime, Journal of Criminal Justice,	- Racial inequality - Income inequality - Unemployment - City disadvantage - Racial segregation	- Statistical research proves the relationship between racial inequality, economic inequality, and crime rates.	Lisa Stolzenberg, David Eitle, Stewart J. D'Alessio, Volume 34

Data Source and Preparation

Our main source of data is taken from the Chicago Data Portal Government website. This dataset contains the complete records of crimes in Chicago from 2001 to 2020. The link is as below - https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data

This raw dataset contains 7,530,492 records X 22 variables of crime in the City of Chicago. Only the useful variables were taken for our analysis as explained in the process below. Sample View of the Raw Dataset is as shown below in **Figure 1**.

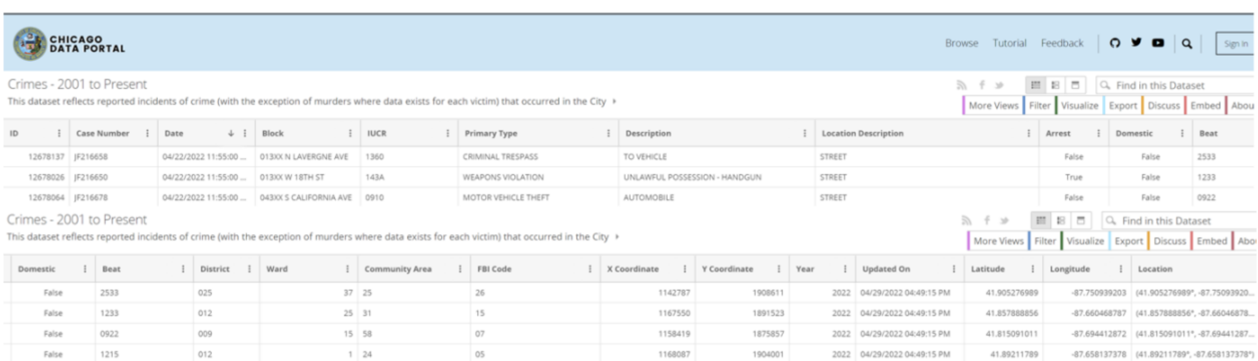


Figure 1. Preview of Raw Dataset from Chicago Data Portal

In order to make efficient use of available technology resources, we have utilized power of SQL Queries using Big Query on GCP's website which replicates this dataset at this link - https://console.cloud.google.com/marketplace/product/city-of-chicago-public-data/chicago-crime?q=search&referrer=search&project=elated-pathway-333103

This enabled us to get the aggregated and reduced dataset which pertains only to crimes between 2010 to 2020, violent crime types, aggregated and grouped by Year, Month, Ward, & Community Area. This was done to get the aggregated number of crimes for a specific crime type, at the drilldown level specific to the month, primary crime type, community area etc. We have merged data from different sources using communities, districts, and wards.

There are 77 Community Areas in the City of Chicago, and this is the most granular level possible for analysis, both from a data perspective and the recommendations suggested in this report for crime reduction. The number of crimes incidents at this granular level is our "Y Variable". Coming to dependent "X" variables, we performed an extensive literature review, selected the most important variables that are sensible for crime analysis and started looking for data that is pertinent to City of Chicago from various Government and Trustworthy sources. Unfortunately, the biggest challenge we had to encounter was that most of the data was not consistent across years, at community area levels and simply did not have the granularity needed for our analysis. Hence, we are working with only a few variables that are sensible, meaningful, offer insights useful for actionable recommendations or at least descriptive analysis. Although some of the other variables gained from literature review were extremely important, we had to leave out some due to inconsistency across Year and Community Area as it was not usable. We have sourced variables for analysis which also relate to multiple features out there which could have been used to ensure that we avoid multicollinearity and not miss out on analyzing the predictor variable's effect. For example, Illiteracy & Uneducated population was a feature mentioned as extremely important in many research papers, we have used "perc_nhsg" which indicates the percentage of people in each community area who do not have high school graduation. Similarly, Individual Income, GDP, Per Capita Income etc. Were mentioned as significant factors that could be analyzed against crime. Instead of these three variables, we have sourced the data for "Percentage of people below poverty level" at each community level over the years as it captures the essence of the previous variables and helps us to understand its effect on crime and by how much we can reduce it by creating better opportunities for the poor. Our goal was to have variables that can be manipulated to reduce crime. One trade-off that we need to live with is to have some variables which are major contributors to crime and help describe the marginal effect of increase/decrease in crime. Therefore, in our final dataset, we have variables of both types- 1. Used for understanding of marginal effects on crime incidents & 2. Manipulative variables on which we can build actionable recommendations for the Government.

Here are the other data from where we have sourced the data for our Variables.

1.Chicago Data Portal	4.Chicago Metropolitan Agency for Planning
2.DePaul University Library	5.Chicago Police Department
3.City of Chicago Office of Inspector General portal	6. Google Cloud Platform

Once all the variables were acquired, we merged this data with our reduced dataset that was extracted from GCP's Cloud Data Source using Big Query SQL as shown in **Figure 2**. The variables, predictor table and rationale are described in the section below. Below is the query used to extract the count of violent crimes for each month starting from 2011. Incidents are grouped by community, ward, and district.

```
SELECT *
FROM ( SELECT primary_type, district, community_area, ward, year, EXTRACT(MONTH FROM date) AS month, COUNT(1) AS incidents,
  FROM `bigquery-public-data.chicago_crime.crime`
  WHERE primary_type in ('CRIMINAL DAMAGE','SEX OFFENSE','CRIM SEXUAL ASSAULT','HOMICIDE','KIDNAPPING','HUMAN TRAFFICKING') AND year > 2010
  GROUP BY primary_type, district, community_area, ward, year, month)
ORDER BY month asc,year asc
```

Processing location: US

Query results

SAVE RESULTS EXPLORE DATA

JOB INFORMATION RESULTS JSON EXECUTION DETAILS

Row	primary_type	district	community_area	ward	year	month	incidents
1	HOMICIDE	4	45	8	2011	1	1
2	SEX OFFENSE	9	59	12	2011	1	1
3	CRIMINAL DAMAGE	4	52	10	2011	1	27

Figure 2. GCP Query to extract aggregated data for our analysis.

Our analysis is only for violent crimes in the City of Chicago. Although there are 36 types of crimes in the raw dataset, we have selected the below 6 types of crimes which fall under the Violent Category.

The final 6 types of crimes that we are working with are colored Yellow as shown in **Figure 3**. Our analysis, recommendations, and interpretations throughout this report are only for these 6 violent crime types. Some of the other crime types, although which may sound like violent crimes, fall under secondary detailed descriptions which indicate that these were not violent crimes. This can be explored in the main raw dataset link mentioned in the first section.

1	PRIMARY_TYPE	19	KIDNAPPING
2	CRIMINAL DAMAGE	20	WEAPONS VIOLATION
3	MOTOR VEHICLE THEFT	21	INTERFERENCE WITH PUBLIC OFFICER
4	PROSTITUTION	22	HOMICIDE
5	OTHER OFFENSE	23	LIQUOR LAW VIOLATION
6	DECEPTIVE PRACTICE	24	INTIMIDATION
7	ASSAULT	25	STALKING
8	CRIM SEXUAL ASSAULT	26	ARSON
9	BATTERY	27	PUBLIC INDECENCY
10	PUBLIC PEACE VIOLATION	28	OBSCENITY
11	THEFT	29	CONCEALED CARRY LICENSE VIOLATION
12	ROBBERY	30	OTHER NARCOTIC VIOLATION
13	BURGLARY	31	NON-CRIMINAL
14	SEX OFFENSE	32	RITUALISM
15	NARCOTICS	33	DOMESTIC VIOLENCE
16	CRIMINAL TRESPASS	34	HUMAN TRAFFICKING
17	OFFENSE INVOLVING CHILDREN	35	NON - CRIMINAL
18	GAMBLING	36	NON-CRIMINAL (SUBJECT SPECIFIED)

Figure 3. Highlighted Primary Crime Types Used for Analysis

Data Curation - Joins, Merging, Cleaning and Feature Engineering

We had to pull data for various features from multiple websites and sources. The biggest challenge was to merge them all logically. We did this part of Data Cleaning, segregation using Excel Pivot Tables and then the merging part using R Studio by joining the variables on Year, Community Area, districts, and wards. The R code for this part is also available in the appendix. Coming to the feature engineering part, we have used the population count of the community areas and normalized the metrics for every 10,000 population. This is done to overcome bias in the data as larger community areas will have more schools, more agencies, more facilities compared to smaller community areas.

The 47,459 rows in our dataset correspond to 2,95,848 incidents of Violent Crime and their distribution by Crime Type is as shown below in **Figure 4**.

Row Labels	Sum of incidents
CRIM SEXUAL ASSAULT	11896
CRIMINAL DAMAGE	267693
HOMICIDE	4669
HUMAN TRAFFICKING	59
KIDNAPPING	1843
SEX OFFENSE	9688
Grand Total	**295848**

Figure 4. Data Distribution by Violent Crime Type

Variable choice & Predictor Table

Instead of selecting all variables available and excluding them in the models, we only picked variables that have some relevance by considering several aspects like economic, demographic, safety related metrics, law enforcement agencies and other aspects indicating the quality of life of people in the community area. Some variables related to economic impact are percentage of population below poverty level, percentage of population with no high school education (school dropouts), percentage of owner-occupied households, employment support agencies, which can show any potential impact of the economic status of the community area on crime.

Similarly, potential impact of demographic factors, like race, ethnicity, percentage of households occupied by owners, number of public schools on crime-incidents in a community area can be analyzed through these variables. Safety related metrics like SSL score and sentiment score help analyze how safe or risk prone the community area is to crime in the future, so that relevant measures can be taken by the government to curb crime. Sentiment score indicates how safe the people in a community feel, it is calculated from survey responses given by citizens. SSL score on the other hand indicates the involvement in previous crimes-like involvement in battery, arrests in violent crimes, gang affiliation, narcotic involvement arrests etc., This score ranges from 0 to 500, 0 indicating extremely low risk and 500 indicating an extremely high risk. Following is the predictor table for each of these variables.

Table 2. Predictor Table

Y Variable – Number of Incidents (of violent crime)		
Predictor Variables	Effect	Rationale
Number of Public schools per 10k population	-	Access to state-funded education might result in better awareness, getting jobs, an improved standard of life and less inclination towards crime.
Sentiment Score	+/-	The higher the sentiment score, the safer the area and vice versa.
Percentage of people below poverty level	+	People below the poverty line may tend to be involved in criminal activities due to a lack of money/jobs/ facilities etc. and struggle to make ends meet.
Percentage of population by race and ethnicity-Asian, Hispanic, Black, White	?	Included to check if there is any racial impact on crime/Some races might feel oppressed or have cultural differences leading to a hostile environment due to being at a natural disadvantage etc.
SSL score (Strategic subject list)	+/-	A higher SSL score indicates more risk of committing crime, lower score indicates a lower risk.
Liquor moratorium	+/-	Community areas voted as “dry” by citizens do not issue certain types of liquor licenses, showing responsible citizens and less chances of alcohol abuse/crime and its chain-reaction effects.
Percentage of households occupied by owners	+/-	We are including this to understand bias and how it affects crime as rental properties have higher criminal activity compared to owner occupied dwellings which may be financially stable and less prone to crime.
Population without minimal high school education	+	Community areas with high illiteracy, leads to individuals with no means of income/jobs, which leads to an increase in crime.
Health support agencies per 10k population	-	More accessibility to healthcare and support services from the government may lead to better communities and less crime
Employment support agencies per 10k population	-	Employment and youth improvement agencies may lead to less unemployment and help the disadvantaged people to get jobs, which will reduce potential involvement in crime.
Homeless support agencies per 10k population	-	A smaller number of homeless or poor people may result in less crime and these agencies help them with necessities.
Community Area, Year, Month	+/-	Used to understand how crime incidents have increased/decreased over time in the community areas. (These were also used to join the datasets)
Excluded Columns		
Population	None	Used to normalize the features used for analysis per 10,000 people in each community area.
Ward, District	None	Used to join the variables.
Primary Type	None	Extracted from GCP for Violent Crime types

Descriptive Analysis & Data Visualizations

Before diving deep into the analysis, we performed basic Exploratory Data Analysis on our final dataset to understand the data better. Here are some of the visualizations built through R Code and the basic takeaways based on our understanding –

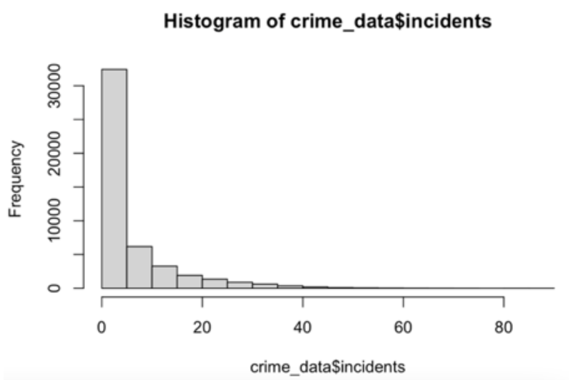


Figure 5.Histogram of Y Variable

Based on the histogram of the dependent variable we can observe that distribution seems to be exponential in nature and since the we have monthly count of the crimes in Chicago, we should use Poisson model for statistical analysis.

Performance Analytics Chart

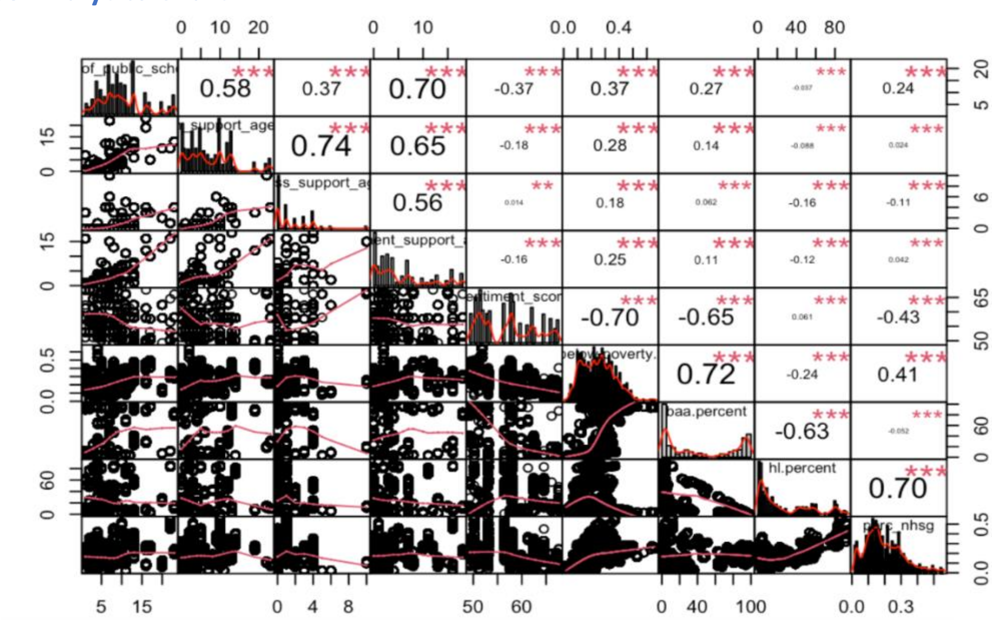


Figure 6. Auto-Correlation Chart

Based on the above correlation chart shows the highly correlated variables in the dataset which should be carefully considered before any further analysis. Therefore, the following variable would be used in different statistical models to negate the effect of correlation which will help us create robust model - Homeless Support Agencies, Employment Support Agencies, Sentiment Score, Percentage Below Poverty Line & Percentage White Population

Since Data Visualization at a geographical level is an important part when it comes to analyzing geo-spatial features of our raw dataset i.e., the latitude and longitude details of the crime incidents, we pulled this complete dataset into Tableau and created some interactive visualizations which can help any user irrespective of how much knowledge they have on this subject to go explore, play around with available filters and gain a good understanding of the high crime areas in the City of Chicago across its 77 community areas. The sample view of

our Tableau Heatmap is shown below. The use of this interactive heatmap is to allow users to find answers to specific questions regarding crimes in the city allowing them to apply those filters to view the geographic crime distribution for those specific crime types/ wards/ districts/ Timestamps. The interactive tableau dashboard can be accessed at -
https://public.tableau.com/app/profile/adhithyan.r/viz/ChicagoCrimesHeatMap-2001to2020/ChicagoCrimeHeatMap

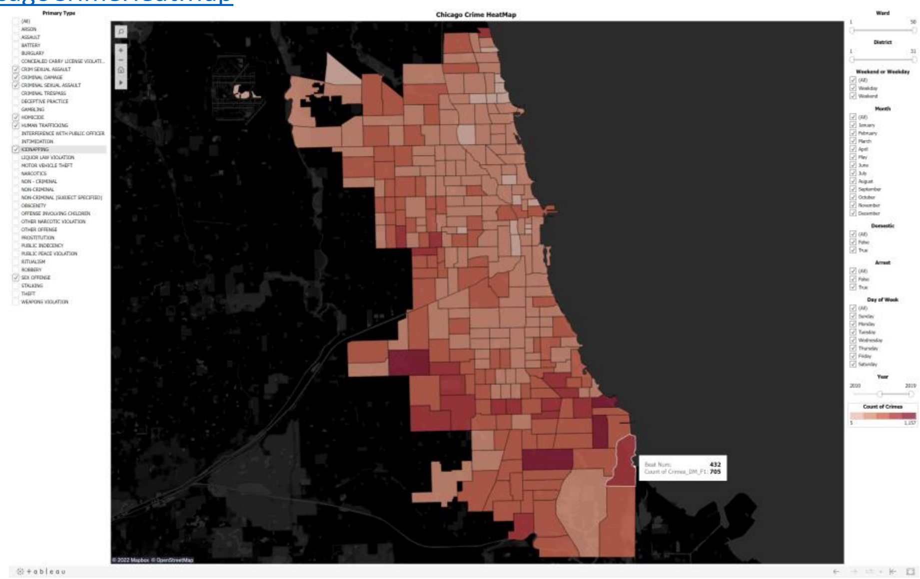


Figure 7. Violent Crime Heatmap

When we look at the violent crime increase/ decrease over the years from 2010 to 2020, we can clearly see that the highest number of crimes occurred from Q3 of 2015 to Q1 of 2019 as shown in the figures below. The one on the left is for crime over the years and the one on the right is the same broken down by Quarters for the peak period.

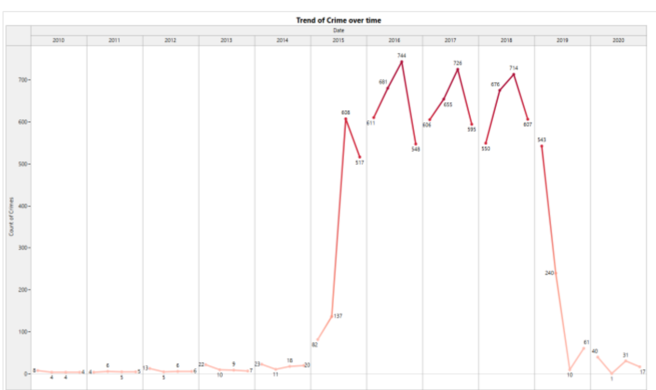


Figure 8. Crime over Years

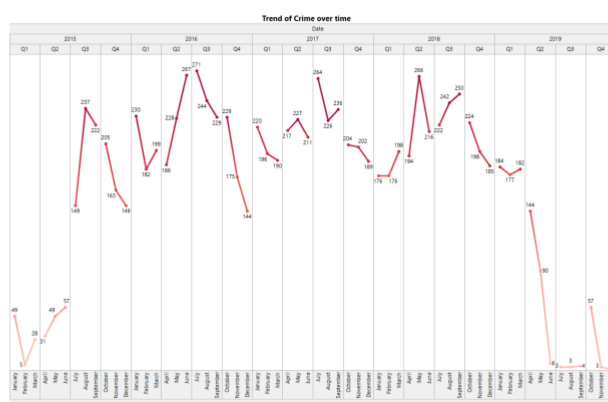


Figure 9.Crime during peak from 2015 to 2019

Models

Crime incidents are not normally distributed, and the data is at two different non-hierarchical levels ward and community. Hence, we must use multi-level analysis with Poisson distributions for this analysis. Therefore, we are using GLMER model for our Y variable which is “Number of Crimes” (Aggregated on monthly basis),

- Model 1 is the baseline Poisson model.
- Model 2 is a GLMER model that accounts for community and ward as random effects and the variables of highest importance as per our literature review, domain knowledge. Actionable recommendations were chosen from this model. This is our best model with lowest AIC value.

- Model 3 is used to analyze and understand the marginal effects of variables that are not included in model 2, because of correlations.

The model code, stargazer outputs are available in the appendix.

Quality Checks

1. The correlation plot and VIF test gives no indication for multicollinearity between the variables used in our model. As we can see in the **Figure 10**. values, all the VIF values are below 5. Therefore, we can conclude that the multicollinearity test is passed.
2. Our dataset does not have any zero values; therefore, excess zero assumption of Poisson model is passed.

```
> vif(m2)
                            GVIF Df GVIF^(1/(2*Df))
month                   1.000350 11        1.000016
no_of_public_schools    2.385695  1        1.544569
health_support_agencies 2.189957  1        1.479850
percent.owner.occupied  1.092149  1        1.045059
perc_nhsg               1.677817  1        1.295306
asian.percent           1.094385  1        1.046128
ssl_score               1.824462  1        1.350727
```

Figure 10. Multi-Collinearity

Interpretations

Findings

Based on our analysis and the beta coefficients for our dependent variables, we have the following observations regarding the reduction of crimes.

- An addition of 1 public school in a community area per 10000 population will reduce crime by 2.05 %.
- The addition of 1 health support agency in a community area per 10000 population will reduce crime by 8.29%.
- An increase of 1% of population with high school graduation decreases crime by 152.98%.

Insights into Violent Crimes Over the Years compared to 2011

Table 3. Violent Crimes over the years

YEAR	Crime reduction compared to 2011 (lesser by)	Crime reduction compared to previous year
2012	-0.7%	-
2013	-11.3%	10.6%
2014	-22.2%	11%
2015	-25.7%	3.5%
2016	-26.2%	0.5%
2017	-29.7%	3.5%
2018	-34.3%	6.6%
2019	-36.1%	1.8%

We can observe that the biggest annual reduction in violent crimes happened between 2013-14 by 11% followed by 2012-13 by 10.6%. The least annual crime reduction was in 2015-16 where crime reduced by just 0.5%.

Interpretation of Marginal Effects of Race & other factors on Violent Crimes from 2011-2019

Looking at the coefficients of the model we used to analyze the effect of race on crimes, we can draw the interpretations below-

m4 <- glmer (incidents ~ asian.percent + baa.percent + hl.percent + white.percent + (1|community_area) , family = poisson, data = df)

- An increase of the Asian population by 1% in a community area, leads to 1.6% increase in violent crimes.

- An increase of the Black population by 1% in a community area leads to a 6.06% increase in violent crimes.
- An increase of the Hispanic population by 1% in a community area, leads to a 5.31% increase in violent crimes.
- An increase of the White population by 1% in a community area leads to a 5.98% increase in violent crimes.

```
Fixed effects:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.719184   0.420809  -8.838  < 2e-16 ***
asian.percent  0.016994   0.005130   3.313 0.000925 ***
baa.percent    0.060667   0.004216  14.388  < 2e-16 ***
hl.percent     0.053194   0.004348  12.234  < 2e-16 ***
white.percent  0.059868   0.004440  13.484  < 2e-16 ***
---
```

Figure 11. Fixed Effects for model 4

- An increase of 1 unit in SSL Score (which indicates risk of possible crime, measures previous arrests in narcotics and violent crime) increases crime by 0.6%. Implying that communities that have high risk of violent crime involve previous involvement of citizens in violent crime, increasing the number of police stations/ patrolling in these areas might reduce the chances of further known involvements in violent crime.

Interpretation of effect of different types of Agencies on crime

- The addition of 1 employment support agency in a community area per 10000 population will reduce crime by 25.1%. This could be because employment agencies are doing good work in helping the disadvantaged to gain skills and jobs.
- The addition of 1 homeless support agency in a community area per 10000 population will increase crime by 69.4 %.

Interpretation of effect of Poverty on crime

- An increase of 1 percent of below poverty level population in a community area increases crime by 87.8%. This could be because these people are at a disadvantage due to the dearth of access to education, jobs, basic facilities, etc. Better conditions in helping people get out of poverty will reduce crime to a large extent but it is a challenging task.

Interpretation of Crime across the Chicago Community Areas

In this section, we are interpreting the coefficients of the community areas based on the random effects. The top 5 community areas shown in **green are the ones with lowest crime** compared to average whereas the **bottom 5 in red are the ones with highest crime.**

Table 4. Community Areas with Highest and Lowest Crimes

Community Area Name	Community Area	Coefficients	Percentage change compared to average
Avondale	21	-1.117730632	-111.7730632%
North Park	13	-0.984722224	-98.4722224%
Edison Park	9	-0.968214121	-96.8214121%
Forest Glen	12	-0.889901891	-88.9901891%
O'Hare	76	-0.855998667	-85.5998667%

Kenwood	39	0.626675779	62.6675779%
East Side	52	0.663486401	66.3486401%
Ashburn	70	0.757349083	75.7349083%
Bridgeport	60	0.81639812	81.639812%
Garfield Ridge	56	0.850059981	85.0059981%

Recommendations

According to our findings, following are some of the recommendations to the City of Chicago-Government, Chicago PD, State Govt. Of Illinois for the reduction of crime in the urban geographic boundaries of Chicago City. As there are no silver bullet actions to prevent violent crime in short-term, following recommendations can possibly prevent or reduce violent crime in the long term.

- Increasing the number of public schools in a community thereby making education more accessible in the community, could create better awareness among young individuals and potentially reduce their involvement in criminal activities
- Ensuring better healthcare, childcare, and family support, by increasing related agencies or implementing more welfare schemes for women, children and families could lead to a stable environment and a safe community area.
- Encouraging youngsters to complete high school education by creating awareness programs, employment opportunities and skill workshops, could lead to better job prospects and economic stability in the community.
- Creating more employment opportunities by enabling local businesses with tax benefits or implementing workshops about establishing and maintaining businesses, could reduce unemployment and poverty in the long run, there by curbing people to resort to criminal activities.
- For community areas with high crime, security measures like increasing law enforcement officers, patrolling, citizen-police interactions, community engagement can possibly prevent involvement in violence.

References:

Data Sources

1.Chicago Data Portal - https://data.cityofchicago.org/
2.DePaul University Library - https://libguides.depaul.edu/c.php?g=628313&p=4384736
3.City of Chicago Office of Inspector General portal- https://informationportal.igchicago.org/
4.Chicago Metropolitan Agency for Planning - https://www.cmap.illinois.gov/data/community-snapshots
5.Chicago Police Department - https://home.chicagopolice.org/statistics-data/

Links

https://crime-data-explorer.app.cloud.gov/pages/explorer/crime/crime-trend
https://www.cdc.gov/violenceprevention/datasources/nvdrs/index.html
https://law.jrank.org/pages/2222/Urban-Crime-Are-crime-rates-higher-in-urban-areas.html
https://www.ncjrs.gov/ovc_archives/ncvrw/2017/images/en_artwork/Fact_Sheets/
https://chicago.suntimes.com/crime/2022/1/3/22858995/chicago-violence-dangerous-murders-per-capita-2021-2020-surge-garfield-park-police-lori-lightfoot
https://bjs.ojp.gov/

Appendix: Stargazer Output, Community Area Analysis, Research Papers, and R Code

Stargazer Output

	Dependent variable:		
	incidents		
	Poisson	generalized linear mixed-effects	
	(1)	(2)	(3)
year2012	0.007 (0.007)	-0.007 (0.007)	-0.038*** (0.007)
year2013	-0.105*** (0.007)	-0.113*** (0.008)	-0.149*** (0.007)
year2014	-0.258*** (0.008)	-0.222*** (0.008)	-0.243*** (0.008)
year2015	-0.354*** (0.008)	-0.257*** (0.010)	-0.224*** (0.008)
year2016	-0.414*** (0.010)	-0.262*** (0.013)	-0.171*** (0.009)
year2017	-0.488*** (0.010)	-0.297*** (0.013)	-0.221*** (0.009)
year2018	-0.539*** (0.010)	-0.343*** (0.013)	-0.279*** (0.009)
year2019	-0.540*** (0.010)	-0.361*** (0.013)	-0.289*** (0.010)
month2	-0.047*** (0.010)	-0.052*** (0.010)	-0.049*** (0.010)
month3	0.137*** (0.009)	0.132*** (0.009)	0.134*** (0.009)
month4	0.185*** (0.009)	0.177*** (0.009)	0.182*** (0.009)
month5	0.187*** (0.009)	0.185*** (0.009)	0.188*** (0.009)
month6	0.174*** (0.009)	0.166*** (0.009)	0.170*** (0.009)
month7	0.210*** (0.009)	0.202*** (0.009)	0.207*** (0.009)
month8	0.185*** (0.009)	0.175*** (0.009)	0.181*** (0.009)
month9	0.130*** (0.009)	0.126*** (0.009)	0.131*** (0.009)
month10	0.158*** (0.009)	0.149*** (0.009)	0.155*** (0.009)
month11	0.105*** (0.009)	0.097*** (0.009)	0.103*** (0.009)
month12	0.105*** (0.010)	0.098*** (0.010)	0.103*** (0.010)
no_of_public_schools_per10k_pop	-0.050*** (0.002)	-0.021 (0.018)	
health_support_agencies_per10k_pop	0.030*** (0.002)	-0.083*** (0.025)	
percent.owner.occupied	-0.113*** (0.014)	-1.057*** (0.131)	
perc_nhsg	-0.293*** (0.019)	1.530*** (0.119)	
asian.percent	-0.008*** (0.0003)	-0.007*** (0.003)	
ssl_score	0.011*** (0.0003)	0.006*** (0.0004)	
employment_support_agencies_per10k_pop			-0.251*** (0.036)
homeless_support_agencies_per10k_pop			0.694*** (0.136)
baa.percent			-0.020*** (0.002)
hl.percent			-0.006*** (0.001)
per.below.poverty.level			0.878*** (0.122)
Constant	-1.049*** (0.069)	0.517*** (0.148)	2.641*** (0.150)
Observations	47,458	47,458	47,458
Log Likelihood	-266,319.100	-251,356.700	-256,282.500
Akaike Inf. Crit.	532,690.200	502,769.400	512,617.000
Bayesian Inf. Crit.		503,014.900	512,845.000

Figure 12. Stargazer Output

Comparison of all 77 community areas

Random effects of coefficients for all community areas, their relative increase/ decrease in crime compared to average are shown below. Green zones represent lower crime than average in the city and red zones represent higher crime. We have validated these findings with existing crime information from Chicago Neighborhoods and confirm that this analysis represents the reality of violent crime in the City of Chicago as shown in the below figure going from Green which represent the safest community areas to Red i.e., the most dangerous community areas.

Community_Area_Name	Community_Area	Coefficients	Percentage change compared to average
Avondale	21	-1.11773063	-111.7730632%
North Park	13	-0.98472222	-98.4722224%
Edison Park	9	-0.96821412	-96.8214121%
Forest Glen	12	-0.88990189	-88.9901891%
O'Hare	76	-0.85599867	-85.5998667%
Douglas	35	-0.62270846	-62.2708461%
Fuller Park	37	-0.61684279	-61.6842792%
Gage Park	63	-0.571243	-57.1243003%
Irving Park	16	-0.54675463	-54.6754634%
Hermosa	20	-0.51175242	-51.1752416%
Edgewater	77	-0.4637543	-46.3754295%
Humboldt Park	23	-0.45691686	-45.6916864%
Albany Park	14	-0.41785483	-41.7854828%
South Lawndale	30	-0.41183265	-41.1832647%
Englewood	68	-0.3825064	-38.25064%
Armour Square	34	-0.34549096	-34.5490956%
Washington Park	40	-0.33205709	-33.2057089%
New City	61	-0.32321485	-32.3214852%
Norwood Park	10	-0.31688881	-31.6888808%
Uptown	3	-0.3152481	-31.5248102%
Oakland	36	-0.27842149	-27.8421485%
Pullman	50	-0.23438358	-23.4383577%
Lake View	6	-0.18678692	-18.6786923%
West Ridge	2	-0.18601532	-18.6015317%
West Elsdon	62	-0.17726002	-17.7260016%
Riverdale	54	-0.17650931	-17.6509312%
Belmont Cragin	19	-0.173531	-17.3531%
Brighton Park	58	-0.1547232	-15.4723195%
Greater Grand Crossing	69	-0.13549421	-13.5494209%
Lower West Side	31	-0.12150583	-12.1505834%
Burnside	47	-0.1136689	-11.3668904%
Archer Heights	57	-0.11154026	-11.1540258%
Chicago Lawn	66	-0.09131419	-9.1314192%
Hyde Park	41	-0.05491818	-5.4918179%
Woodlawn	42	-0.05010842	-5.0108418%
Logan Square	22	-0.01123323	-1.123323%
McKinley Park	59	0.002272979	0.2272979%
East Garfield Park	27	0.035789411	3.5789411%
Montclaire	18	0.052853313	5.2853313%
Portage Park	15	0.10484342	10.484342%
West Garfield Park	26	0.108182899	10.8182899%
Austin	25	0.123855626	12.3855626%
Lincoln Square	4	0.136543844	13.6543844%
Roseland	49	0.155611652	15.5611652%
Clearing	64	0.164390697	16.4390697%
Chatham	44	0.166823604	16.6823604%
North Center	5	0.167111598	16.7111598%
West Englewood	67	0.194270573	19.4270573%
South Deering	51	0.209918041	20.9918041%
South Shore	43	0.212302486	21.2302486%
Lincoln Park	7	0.216385973	21.6385973%
Mount Greenwood	74	0.218262531	21.8262531%
West Town	24	0.218453235	21.8453235%
Morgan Park	75	0.222605639	22.2605639%
Near West Side	28	0.260552797	26.0552797%
Rogers Park	1	0.264180877	26.4180877%
Near North Side	8	0.269915641	26.9915641%
Auburn Gresham	71	0.270821698	27.0821698%
Near South Side	33	0.278543381	27.8543381%
Dunning	17	0.297318184	29.7318184%
Calumet Heights	48	0.315148636	31.5148636%
South Chicago	46	0.356288193	35.6288193%
Jefferson Park	11	0.43414073	43.414073%
Loop	32	0.436275198	43.6275198%
North Lawndale	29	0.449652968	44.9652968%
Grand Boulevard	38	0.463637204	46.3637204%
West Lawn	65	0.494700558	49.4700558%
Avalon Park	45	0.503386806	50.3386806%
Washington Heights	73	0.524011795	52.4011795%
Beverly	72	0.539806307	53.9806307%
West Pullman	53	0.554694349	55.4694349%
Hegewisch	55	0.586317604	58.6317604%
Kenwood	39	0.626675779	62.6675779%
East Side	52	0.663486401	66.3486401%
Ashburn	70	0.757349083	75.7349083%
Bridgeport	60	0.81639812	81.639812%
Garfield Ridge	56	0.850059981	85.0059981%

Figure 13. Community area wise crime effect

Research Papers

1. Varvara Ingilevich, Sergey Ivanov, Crime rate prediction in the urban environment using social factors, Procedia Computer Science, Volume 136,2018, Pages 472-478, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.08.261. (https://www.sciencedirect.com/science/article/pii/S1877050918315667)
2. Crime prediction through urban metrics and statistical learning, Luiz G.A. Alves, Haroldo V. Ribeiro, Francisco A. Rodrigues, Physica A: Statistical Mechanics and its Applications, Volume 505,2018, Pages 435-443, ISSN 0378-4371, https://doi.org/10.1016/j.physa.2018.03.084. (https://www.sciencedirect.com/science/article/pii/S0378437118304059)
3. Cahill M, Mulligan G. Using Geographically Weighted Regression to Explore Local Crime Patterns. Social Science Computer Review. 2007;25(2):174-193. doi:10.1177/0894439307298925. https://journals.sagepub.com/doi/10.1177/0894439307298925
4. Hybrid Support Vector Regression and Autoregressive Integrated Moving Average Models Improved by Particle Swarm Optimization for Property Crime Rates Forecasting with Economic Indicators (https://pdfs.semanticscholar.org/09d8/6d4d248ac9ca2d44e33077ece53775092057.pdf?_ga=2.140746673.2095185444.1647046415-1811409627.1647046415)
5. Alcohol abuse and crime: a fixed-effects regression analysis, David M. Fergusson,L. John Horwood , First published: 03 May 2002 https://doi.org/10.1046/j.1360-0443.2000.951015257.x
6. Kubrin CE, Ishizawa H. Why Some Immigrant Neighborhoods Are Safer than Others: Divergent Findings from Los Angeles and Chicago. The ANNALS of the American Academy of Political and Social Science. 2012;641(1):148-173. doi:10.1177/0002716211431688
7. Osgood, D.W. Poisson-Based Regression Analysis of Aggregate Crime Rates. Journal of Quantitative Criminology 16, 21–43 (2000). https://doi.org/10.1023/A:1007521427059
8. Lisa Stolzenberg, David Eitle, Stewart J. D'Alessio, Race, economic inequality, and violent crime, Journal of Criminal Justice, Volume 34, Issue 3,2006, Pages 303-316, ISSN 0047-2352, https://doi.org/10.1016/j.jcrimjus.2006.03.002. (https://www.sciencedirect.com/science/article/pii/S0047235206000250)

R Code

```
library(readxl)
library(dplyr)
getwd()
setwd("/Users/rajeshmakala/Desktop/SDM Project/Data")

#Importing Chicago crime data
Crime_data = read_excel("Crime_data.xlsx", sheet = "Crime_data")
str(Crime_data)

#number of public schools by community
public_schools = read_excel("Chicago_Public_Schools.xlsx", sheet = "Chicago_Public_Schools")
colnames(public_schools)=tolower(make.names(colnames(public_schools)))
public_schools <- public_schools %>% group_by(community.area.number) %>%
summarise(no_of_public_schools = n())
str(public_schools)

#number of health_support_agencies by community
Health_Support_Services = read_excel("Family_and_Support_Services_Delegate_Agencies.xlsx", sheet =
"Health")
```

```
colnames(Health_Support_Services)=tolower(make.names(colnames(Health_Support_Services)))
Health_Support_Services <- Health_Support_Services %>% group_by(community.area.number) %>%
summarise(health_support_agencies = n())
str(Health_Support_Services)

#number of homeless_support_agencies by community
Homeless_Support_Services = read_excel("Family_and_Support_Services_Delegate_Agencies.xlsx", sheet =
"Homeless")
colnames(Homeless_Support_Services)=tolower(make.names(colnames(Homeless_Support_Services)))
Homeless_Support_Services <- Homeless_Support_Services %>% group_by(community.area.number) %>%
summarise(homeless_support_agencies = n())
str(Homeless_Support_Services)

#number of employment_support_agencies by community
Employment_Support_Services = read_excel("Family_and_Support_Services_Delegate_Agencies.xlsx", sheet =
"employment")
colnames(Employment_Support_Services)=tolower(make.names(colnames(Employment_Support_Services)))
Employment_Support_Services <- Employment_Support_Services %>% group_by(community.area.number)
%>% summarise(employment_support_agencies = n())
str(Employment_Support_Services)


#police sentiment scores at district level
Sentiment_Scores = read_excel("Police_Sentiment_Scores.xlsx", sheet = "Police_Sentiment_Scores")
colnames(Sentiment_Scores)=tolower(make.names(colnames(Sentiment_Scores)))
Sentiment_Scores <- Sentiment_Scores %>% group_by(district) %>% summarise(sentiment_score =
mean(safety))
str(Sentiment_Scores)

#SSL scores at district level
SSL_Scores = read_excel("AA_ data_combined.xlsx", sheet = "SSL Score")
colnames(SSL_Scores)=tolower(make.names(colnames(SSL_Scores)))
SSL_Scores <- SSL_Scores %>% group_by(year,district) %>% summarise(ssl_score = mean(ssl.score))
str(SSL_Scores)

#liquor moratorium status of each ward
Liquor_Status = read_excel("AA_ data_combined.xlsx", sheet = "Liquor moratorium")
colnames(Liquor_Status)=tolower(make.names(colnames(Liquor_Status)))
str(Liquor_Status)

#percent of owner occupied at each community
owner_occupied = read_excel("percentowneroccupied.xlsx", sheet = "Sheet2")
colnames(owner_occupied)=tolower(make.names(colnames(owner_occupied)))
str(owner_occupied)

#percent of population below poverty at each community
below_poverty = read_excel("POVERTY_AR2.xlsx", sheet = "Sheet1")
colnames(below_poverty)=tolower(make.names(colnames(below_poverty)))
str(below_poverty)
```

```
#percent of population by race
per_race = read_excel("Races.xlsx", sheet = "Sheet1")
colnames(per_race)=tolower(make.names(colnames(per_race)))
str(per_race)

#Population by community
population = read_excel("POPULATION.xlsx", sheet = "Sheet1")
colnames(population)=tolower(make.names(colnames(population)))
str(population)

#Percentage of non high school graduates
nhsg = read_excel("perc_nhsg.xlsx", sheet = "Sheet1")
colnames(nhsg)=tolower(make.names(colnames(nhsg)))
str(nhsg)


Crime_data = merge(Crime_data,public_schools,by.x="community_area",by.y="community.area.number",
all.x=TRUE)
Crime_data =
merge(Crime_data,Health_Support_Services,by.x="community_area",by.y="community.area.number",
all.x=TRUE)
Crime_data =
merge(Crime_data,Homeless_Support_Services,by.x="community_area",by.y="community.area.number",
all.x=TRUE)
Crime_data =
merge(Crime_data,Employment_Support_Services,by.x="community_area",by.y="community.area.number",
all.x=TRUE)
Crime_data[is.na(Crime_data)] <- 0
Crime_data = merge(Crime_data,Sentiment_Scores,by.x="district",by.y="district", all.x=TRUE)
Crime_data = merge(Crime_data,SSL_Scores, by.x=c("district", "year"), by.y=c("district", "year"), all.x=TRUE)
Crime_data = merge(Crime_data,Liquor_Status, by.x=c("ward"), by.y=c("ward"), all.x=TRUE)
Crime_data = merge(Crime_data,owner_occupied, by.x=c("community_area", "year"), by.y=c("community",
"year"), all.x=TRUE)
Crime_data = merge(Crime_data,below_poverty, by.x=c("community_area", "year"),
by.y=c("community.area", "year"), all.x=TRUE)
Crime_data = merge(Crime_data,per_race, by.x=c("community_area", "year"), by.y=c("community", "year"),
all.x=TRUE)
Crime_data = merge(Crime_data,population, by.x=c("community_area", "year"), by.y=c("community.area",
"year"), all.x=TRUE)
Crime_data = merge(Crime_data,nhsg, by.x=c("community_area", "year"), by.y=c("community_area", "year"),
all.x=TRUE)

#normalizing data with population
Crime_data$no_of_public_schools_per10k_pop <-
(Crime_data$no_of_public_schools/Crime_data$population)*10000
Crime_data$health_support_agencies_per10k_pop <-
(Crime_data$health_support_agencies/Crime_data$population)*10000
Crime_data$homeless_support_agencies_per10k_pop <-
(Crime_data$homeless_support_agencies/Crime_data$population)*10000
```

```
Crime_data$employment_support_agencies_per10k_pop <-
(Crime_data$employment_support_agencies/Crime_data$population)*10000

# Crime_data$percent.owner.occupied[is.na(Crime_data$percent.owner.occupied )]<-
mean(Crime_data$percent.owner.occupied ,na.rm=TRUE)
# Crime_data$sentiment_score[is.na(Crime_data$sentiment_score )]<-mean(Crime_data$sentiment_score
,na.rm=TRUE)
# Crime_data$ssl_score[is.na(Crime_data$ssl_score )]<-mean(Crime_data$ssl_score ,na.rm=TRUE)
# Crime_data$per.below.poverty.level[is.na(Crime_data$per.below.poverty.level )]<-
mean(Crime_data$per.below.poverty.level ,na.rm=TRUE)
# Crime_data$asian.percent[is.na(Crime_data$asian.percent )]<-mean(Crime_data$asian.percent
,na.rm=TRUE)
# Crime_data$baa.percent[is.na(Crime_data$baa.percent )]<-mean(Crime_data$baa.percent ,na.rm=TRUE)
# Crime_data$hl.percent[is.na(Crime_data$hl.percent )]<-mean(Crime_data$hl.percent ,na.rm=TRUE)
# Crime_data$white.percent[is.na(Crime_data$white.percent )]<-mean(Crime_data$white.percent
,na.rm=TRUE)
# Crime_data$population[is.na(Crime_data$population )]<-mean(Crime_data$population ,na.rm=TRUE)
# Crime_data$perc_nhsg[is.na(Crime_data$perc_nhsg )]<-mean(Crime_data$perc_nhsg ,na.rm=TRUE)

str(Crime_data)
colSums(is.na(Crime_data))

col <- c("year", "month","ward","district", "community_area", "primary_type","liquor_moratorium_status")
Crime_data[col] <- lapply(Crime_data[col], factor)
str(Crime_data)

df <- Crime_data[complete.cases(Crime_data), ]
str(df)
#' Free up memory by removing unnecessary big data dataframes
rm(public_schools,population,Crime_data,nhsg,Health_Support_Services,Homeless_Support_Services,Employ
ment_Support_Services,Sentiment_Scores,SSL_Scores,Liquor_Status,owner_occupied,below_poverty,per_rac
e)

#write.csv(df,"/Users/rajeshmakala/Desktop/SDM Project/finaldata.csv", row.names = FALSE)


library(lattice)
bwplot(incidents ~ year, data=df)
histogram(df$incidents)


d <- df[, c(8,9,10,11,12,13,15,16,17,18,19,20,22)]
str(d)
library(PerformanceAnalytics)
#chart.Correlation(d)
cor(d)

library(lme4)
library(MASS)
```

```
m1 <- glm(incidents ~ year + month + no_of_public_schools_per10k_pop +
health_support_agencies_per10k_pop
          + percent.owner.occupied + perc_nhsg  + asian.percent  + ssl_score,
          family=poisson, data=df)

summary(m1)

m2 <- glmer(incidents ~  year + month + no_of_public_schools_per10k_pop +
health_support_agencies_per10k_pop
          + percent.owner.occupied + perc_nhsg  + asian.percent  + ssl_score
         + (1|community_area) + (1|ward), family=poisson, data=df)

summary(m2)

m3 <- glmer(incidents ~ year + month + employment_support_agencies_per10k_pop +
homeless_support_agencies_per10k_pop +
           baa.percent + hl.percent  + per.below.poverty.level +
           (1|community_area) , family=poisson, data=df)
summary(m3)

m4 <- glmer(incidents ~ asian.percent + baa.percent + hl.percent + white.percent +
           (1|community_area) , family=poisson, data=df)
summary(m4)

library(stargazer)
stargazer(m1, m2, m3, type="text", single.row=TRUE)

# library(AER)
# dispersiontest(m1)

#vif test
library("car")
vif(m2)

#fixed effects
fixef(m2)
#random effects
community_area <- data.frame(c(ranef(m2)$community_area))
community_area[order(community_area$X.Intercept),]

ranef(m2)$community_area
```