ANALYSIS OF CHICAGO'S CRIME DATA USING PYSPARK

(2001 - 2020)

Akshit Behera

ABSTRACT

The third most populous city in the entire United States, Chicago, is among the central hubs for international finance, culture, commerce, industry, education, technology, telecommunications, and transportation. The city has witnessed a decent average annual GDP growth rate of 0.73%*[1] over the last two decades.

However, the city has a dark side to its gleaming exterior. It boasts of one of the highest crime rates in America compared to all communities of all sizes - from the smallest towns to the very largest cities. The city's overall crime rate, especially the violent crime rate, is higher than the US average. Chicago was responsible for nearly half of 2016's increase in homicides in the US, though the nation's crime rates remain near historic lows. The reasons for the higher numbers in Chicago remain unclear

In recent times, however, crimes in general have dipped in the city due to a combination of policy changes and strict vigilance and action by the city's law enforcement authorities. Through this project we attempt to uncover some of the aspects of all criminal activity taking place in the city and explore potential reasons for the recent decline in crime. We will try and understand the nature of crimes being committed, identify the localities with violent streaks and the timings of when these crimes are being committed. From the law-enforcement's perspective we will try to identify which of the police districts have been successful in curbing crime in their localities, as well the arrest trends over the years. Based on our analysis, we will try to identify potential reasons for the high crime rate in the city by finding relationships between crime and other socio-economic factors and chart a way-forward.

ABOUT DATASET

Our dataset has been taken from the website Chicago Data Portal (https://data.cityofchicago.org/), which is a repository of all kinds of publicly available datasets on the socio-cultural and demographic aspects of the city of Chicago. From the website, we have chosen the data on all crimes committed in the city from 2001-till date. The data downloaded from the website is in CSV format and of size 1.5 GB.

[1] * *https://www.opendatanetwork.com*

The various column names and their data types are described below.

Column Name	Description	Data Type
ID	Unique identifier for the record.	Number
Case Number	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.	Plain Text
Date	Date when the incident occurred. This is sometimes a best estimate.	Date & Time
Block	The partially redacted address where the incident occurred, placing it on the same block as the actual address.	Plain Text
IUCR	The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description	Plain Text
Primary Type	The primary description of the IUCR code.	Plain Text
Description	The secondary description of the IUCR code, a subcategory of the primary description.	Plain Text
Location Description	Description of the location where the incident occurred.	Plain Text
Arrest	Indicates whether an arrest was made.	Boolean
Domestic	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.	Boolean

Column Name	Description	Data Type
Beat	Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts. See the beats at https://data.cityofchicago.org/d/aerh-rz74.	Plain Text
District	Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz3r.	Plain Text
Ward	The ward (City Council district) where the incident occurred. See the wards at https://data.cityofchicago.org/d/sp34-6z76.	Number
Community Area	Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at https://data.cityofchicago.org/d/cauq-8yn6.	Plain Text
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.	Plain Text
X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.	Number
Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.	Number

Column Name	Description	Data Type
Year	Year the incident occurred.	Number
Updated On	Date and time the record was last updated.	Date & Time
Latitude	The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	Number
Longitude	The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	Number
Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.	Location

Two additional datasets used on Chicago's population and unemployment rates from 2001-2020[2][3].

BUSINESS QUESTIONS & APPROACH

The business questions identified after multiple inspections of the dataset as well as the approach used to answer them are as follows:

1. **How are the Crimes and Crime Rates in Chicago trending over the years?**
   - *Loaded the Chicago crime and population datasets in DBFS (Databricks File System)*
   - *Saved the dataset as a temptable called 'Crimes'*

[2] https://ycharts.com/indicators/chicago_il_unemployment_rate

[3] https://datacommons.org/ranking/Count_Person/City/geoId/17043?h=geoId%2F1714000

- *Used the spark.sql function to select the year, count of crime IDs from the Crimes table, grouped by Year in ascending order. Saved the output in a temptable called Yearwise_Crimes*
- *Left joined the population table on Yearwise_Crime_Rates by Year and saved the output as Yearwise_Crime_Rates*
- *Calculated crime rate as crimes per 100,000 population*

2. What is the nature of crimes being committed?

a) Most common crimes over the years

Used the spark.sql function to select Primary Type as Crime_Type and count of the crimes by each type from the Crimes table, grouped by Crime_Type and ordered by the number of Crimes in descending order. Saved the output in a temptable called Top_Crimes. This output gives the most common crimes committed over the years

b) Crimes which have increase / decreased the most in the last 5 years

- *Created another temptable called Crime_by_year, which had year-wise count of crimes for each type of crime for the period (2015-2019). From this table created a pivot table view with crime types in rows and years in columns, and with the count as values*
- *Calculated the % difference in crime for each category from the previous year and took the average. Saved the results to a variable called 'Crime_Rate_Chane'*

c) Relationship between Prostitution and Sexual Assault

- *Created two temptables called 'Prostitution' and 'Sexual Assault' which had the year-wise count of each types of crime*
- *Joined the 'Sexual Assault' table on 'Prostitution' table by year*
- *Calculated correlation between the two columns – Prostitution and Sexual Assault*

d) Split of Domestic vs. Non-Domestic Crimes

Created a temptable called Domestic_Crimes, summarizing the count of crimes as 'Domestic' or 'Non-Domestic'. Used the same table to calculate the proportions as well

3. **Where exactly in the city are these crimes taking place?**

   a) **In terms of locality, blocks, community areas and police districts**

- Created a temptable called 'Crime_by_Location' and summarized the year by crimes by each location type
- For crimes by blocks, summarized the year by crimes for each block and sorted by descending to find out blocks with high number of crimes
- For crimes by community_areas, summarized the year by crimes for each community area and sorted by descending and ascending orders to find out the top and bottom 5 respectively in terms of crime
- For crimes by police districts, summarized the year by crimes for each police district and sorted by descending and ascending orders to find out the top and bottom 5 respectively in terms of crime

   b) **Community Areas with highest and lowest increase in crime rates in last 5 years**

- *Created another temptable called Crime_by_Community_Area, which had year-wise count of crimes for each community for the period (2015-2019). From this table created a pivot table view with community areas in rows and years in columns, and with the count as values*
- *Calculated the % difference in crime for each community area from the previous year and took the average*
- *Identified the top and bottom 5 from this list by sorting in ascending and descending orders*

4. **When do most crimes occur?**

   a) **By month, time of the day, time of month (starting, mid or ending of a month)**

- *Extracted the date from the 'Date' column and then the month from it. Summarized the number of crimes by each month across all 20 years and calculated the proportion*
- *For time of day, extracted the time component from the date field. Segmented the time slots as Dawn for 12 am – 6 am, Morning for 6 am – 11 am, and Afternoon from 12 pm – 4 pm, and Evening from 4 pm – 8 pm and finally Night from 8 pm – 12 am*

- *Summarized the total number of crimes by the above created groups and stored the results to a variable called Timewise_Crimes*
- *For time of month, segmented a month as Starting for 1-10th, Mid for 11-20th and Ending for 21-31st*
- *Summarized the total number of crimes by the above created groups and stored the results to a variable called Datewise_Crimes*

5. How is the response by law enforcement authorities?

a) Arrest trends over the years by year and crime types

- *Created a temptable called 'Arrests' and summarized the year-wise arrests*
- *Created another temptable called 'Arrest_Percent' and summarized the year-wise total crime cases*
- *Left joined the above two tables and calculated the arrest %, saved to a variable called 'Arrest_Percenta'*
- *For arrests by crime type, created two temptables called 'Arrest_by_Crime1' and 'Arrest_by_Crime2' having the count of total crimes and respective arrests. Joined the two tables and calculated the percentage*

b) Length of investigations by year and crime types

- *Converted the columns, Date and 'Updated On' into Date format and took difference between the two dates to find the length of investigation*
- *Summarized the average length of investigation by Year, Crime Type and Police District*

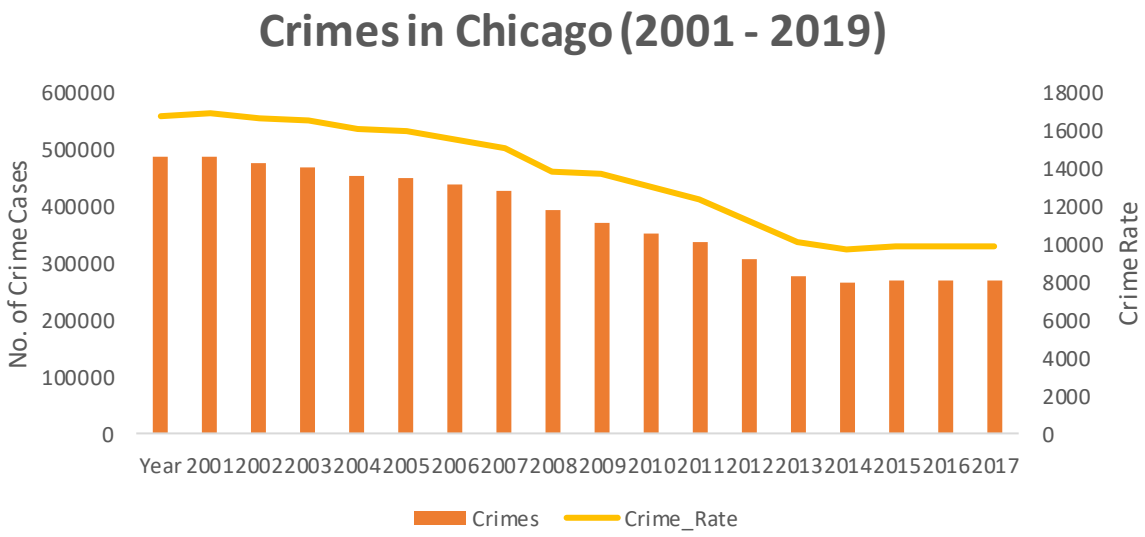
6. Are more crimes being caused due to higher unemployment levels?

- *Loaded the Unemployment dataset on DBFS*
- *Selected month-year from the Date column and count of crime from the Crimes temptable and got the month-year format similar to as in the Unemployment dataset*
- *Left joined the two tables and saved to a variable called 'Corr'. Calculated the correlation for both the variables*
- *Repeated the above steps for the top 5 crime types by quantum*

PLATFORM USED

We have used Databricks' Community platform to write all the queries on the dataset. Also, all the queries have been written using Spark SQL.

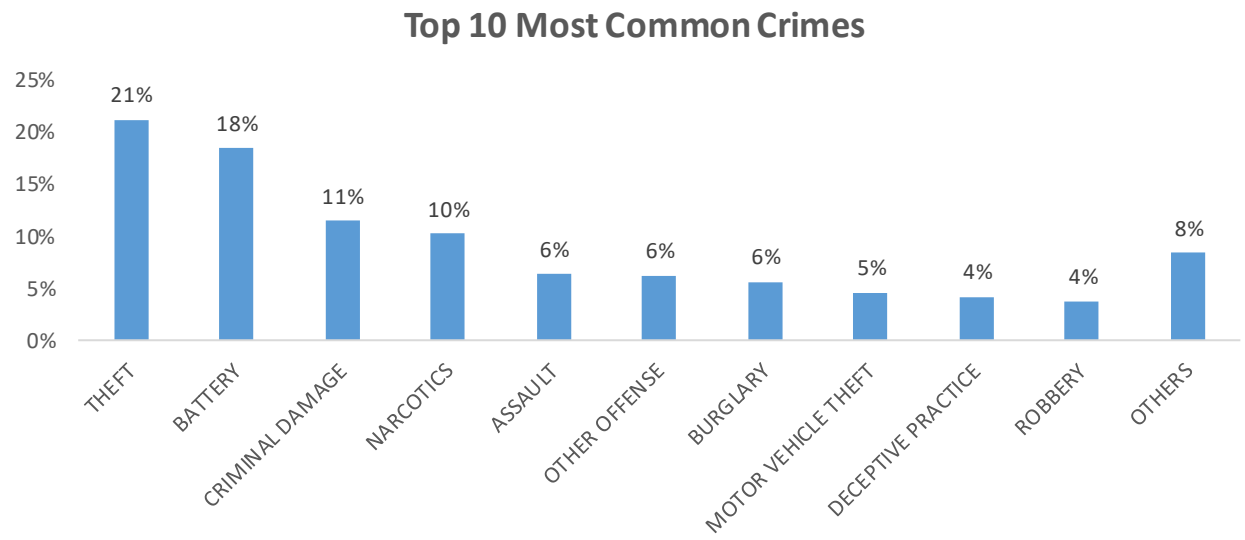
CRIME OVER THE YEARS



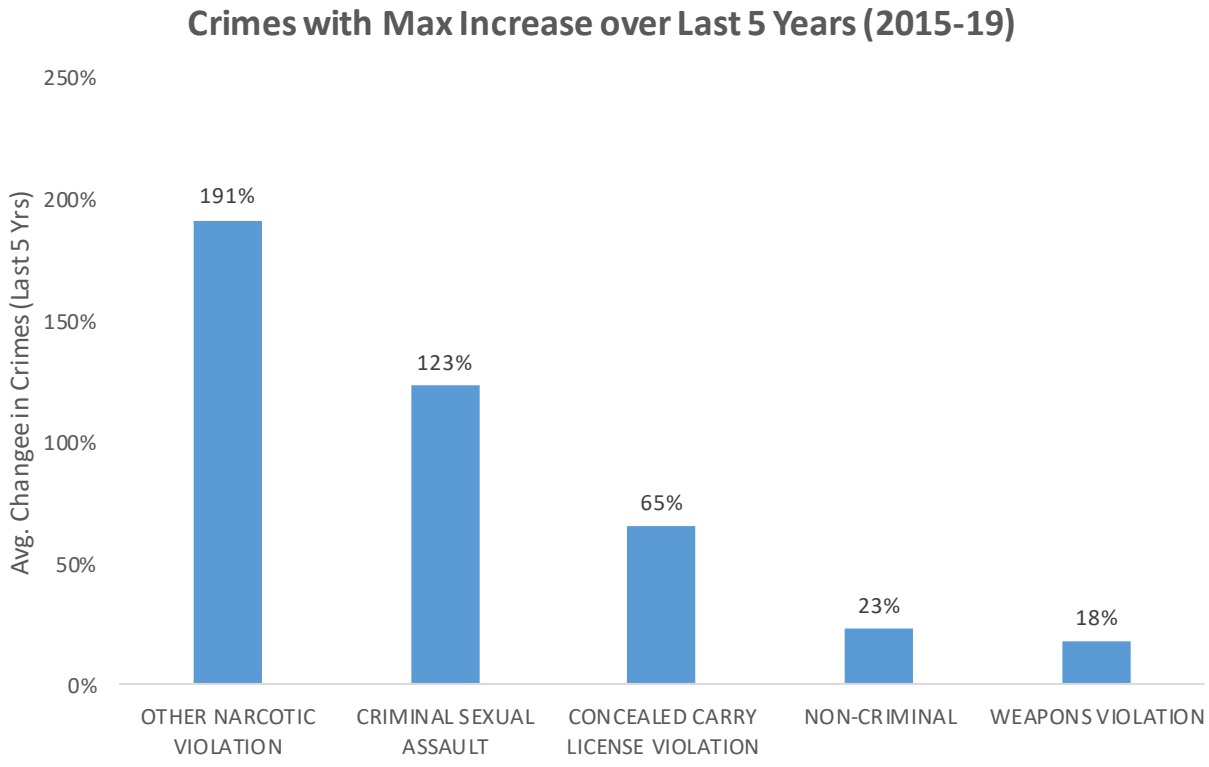
Despite its notorious image of being among the most crime prone cities, Chicago has witnessed a steady decline in crime as well crime rates over the last two decades. The crime rate, expressed as crimes per 100,000 general population, has seen a consistent average annual decline of 3.5% over the last 5 years.

NATURE OF CRIMES BEING COMMITTED

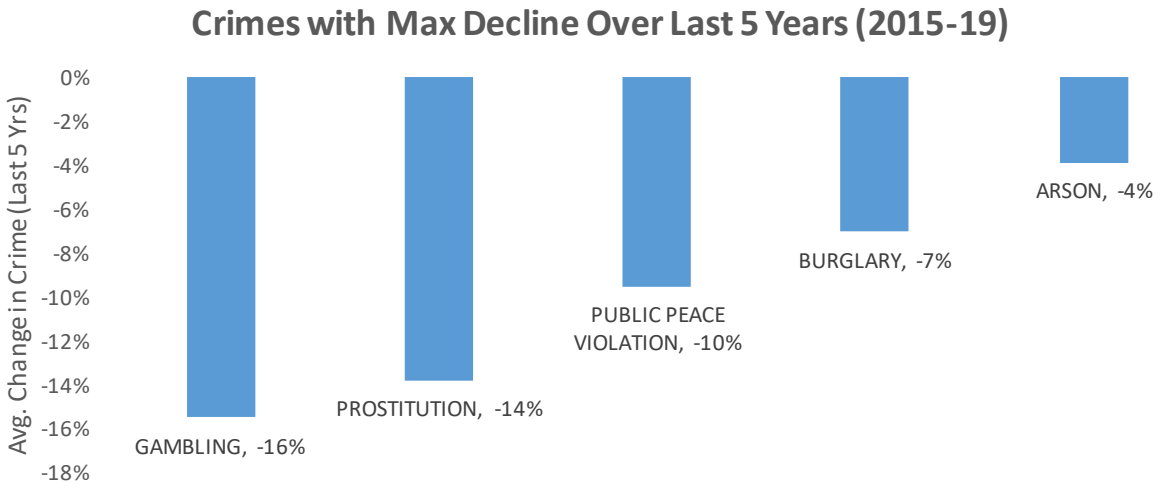
The decline in crime rate over the years gives an account of the commendable job that the law enforcement authorities have done over the years. However, in order to know the city and its ways better, it is important to understand the nature of crimes that are most prominent to the area.



Theft and battery turn out to be the most common forms of crimes being committed in Chicago. In fact, these two together constitute almost 40% of all the crimes that have been committed in the city in the last two decades. However, off late, narcotics related and sexual assaults have been on a dangerous rise, growing at an average of more than 100% over the last 5 years.

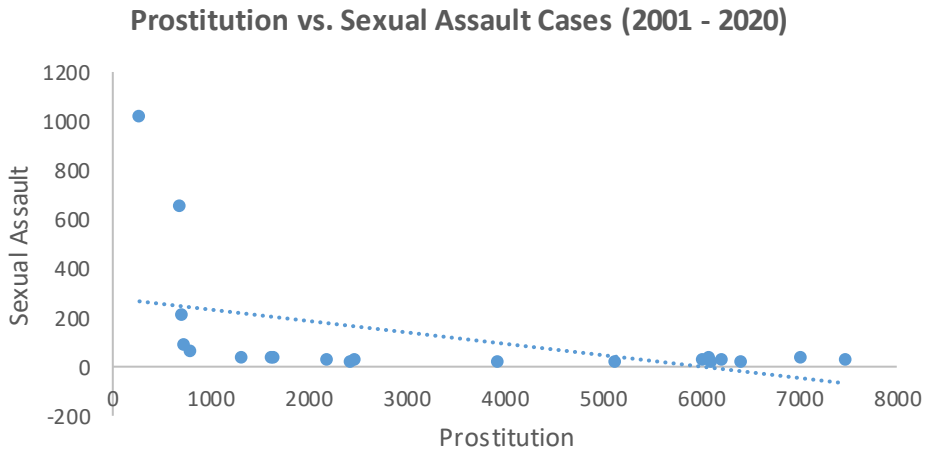


However, interestingly at the same time there are quite a few crimes that have shown consistent decline over the last 5 years. Some of the most prominent ones being gambling and prostitution, which have gone down by 15% on an average every year.



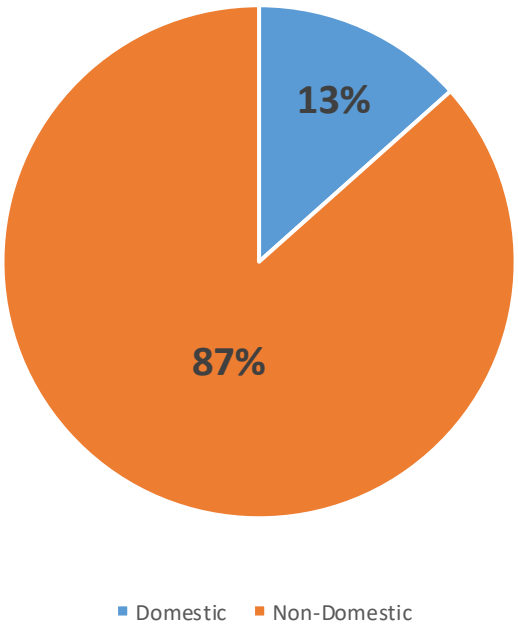
An interesting finding here that catches eye is the drastic rise in sexual assaults coinciding with consistent decline in prostitution. One possible reason for this could be varying degrees of strictness for both the crimes, where more than 99% prostitution cases has seen arrests, whereas sexual assaults have seen arrests in only 7% of cases *(we explore this further later)*.But is the authorities' attempt to curb prostitution in the city driving the increasing number of cases of sexual assaults? Are the two incidents related? We explore the relationship between both types of crimes over the years.

Year	Prostitution	Assault	Year	Prostitution	Assault
2001	6026	23	**2011**	2424	16
2002	6408	16	**2012**	2204	23
2003	6214	20	**2013**	1652	29
2004	7476	23	**2014**	1626	28
2005	6124	18	**2015**	1322	33
2006	7034	28	**2016**	800	62
2007	6087	34	**2017**	735	81
2008	5141	14	**2018**	718	203
2009	3940	17	**2019**	680	653
2010	2485	24	**2020**	272	1012



A negative relationship, along with a negative correlation of (-) 0.46 between the two types of crimes, justifies our observation that the rise in sexual assault related crimes could be a factor of the increased stringent laws around prostitution.

Domestic vs. Non - Domestic Crimes

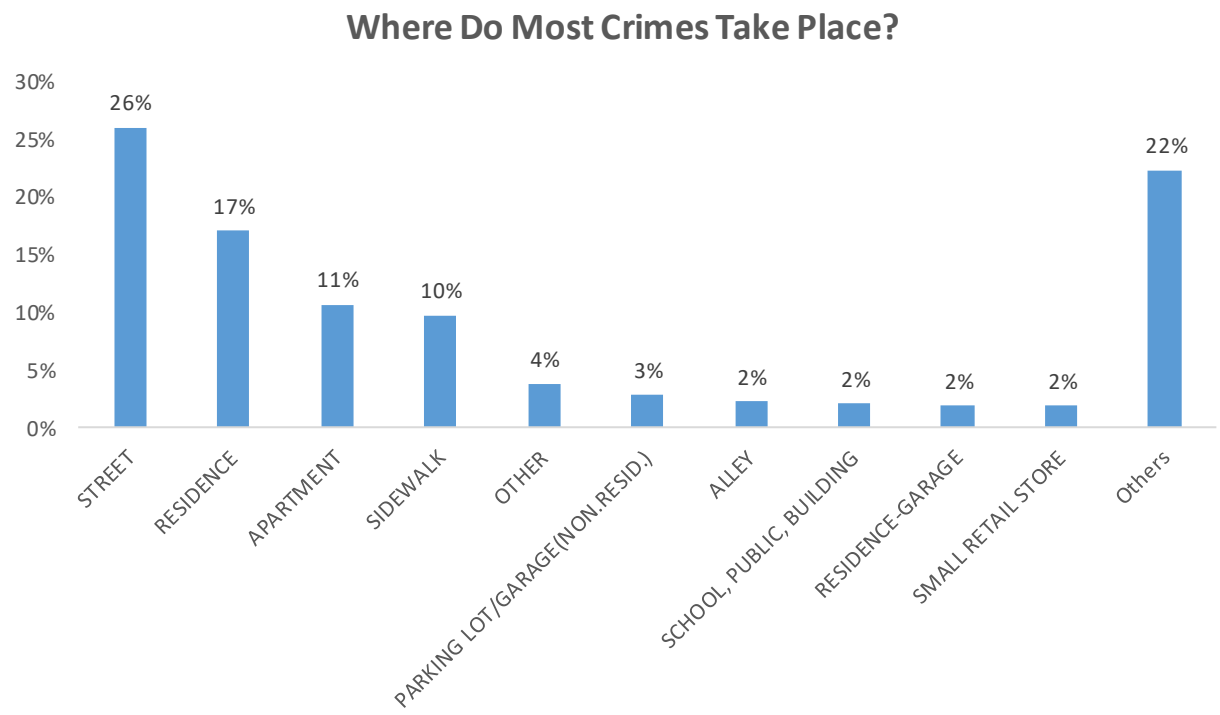


13% of the total crimes over the years have been domestic related

Domestic Crimes indicate whether an incident was domestic-related as defined by the Illinois Domestic Violence Act.

WHERE ARE THE CRIMES HAPPENING

Having gotten some idea of the recent trends in terms of the nature of crimes being committed in the city, the next step is to understand where exactly are these crimes being committed.



Most criminal activities are taking place outdoors, as can be seen in the graph above. 26% of the criminal activities take place on the streets and another 10% on the sidewalks, giving another strong indication of the notorious history of the city's criminal masterminds. In fact there are certain blocks within the city, infamous for being the hotspots for criminal activities over the years. These blocks record way more criminal cases every year as compared to the average across all other blocks.

Block	Crimes
100XX W OHARE ST	15782
001XX N STATE ST	13980
076XX S CICERO AVE	9684
008XX N MICHIGAN AVE	9112
0000X N STATE ST	8461

100XX W OHARE ST. sees approximately 800 crimes being committed every year

WHERE ARE THE CRIMES HAPPENING

Community_Area	Crime
25	419734
8	230172
43	217460
23	209518
28	197176

The suburb of Austin is the most crime prone area in the city of Chicago with an average of **20,986** crimes per year. Along with Austin, Near North Side, South shore, Humbolt Park and Near West Side suburbs complete the list of top 5 most crime prone areas in Chicago

Community_Area	Crime
9	6463
47	9939
12	12027
55	14377
36	14816

Similarly, Edison Park is the safest neighborhood to live in witnessing an average of just 323 crime per year in the last two decades. Burnside, Forest Glan, Hegewisch and Oakland complete the list of top 5 most peaceful neighborhoods in terms of crime

BEST AND WORST POLICE DISTRICTS[5]

District	Crime
8	489529
11	467296
7	425541
6	421269
25	415207

Chicago Lawn is the police district dealing with the maximum number of crimes at an average of around 25,000 crimes per year. Harrison, Englewood, Gresham and Grand Central are too among the top 5 police districts dealing with maximum number of crimes every year.

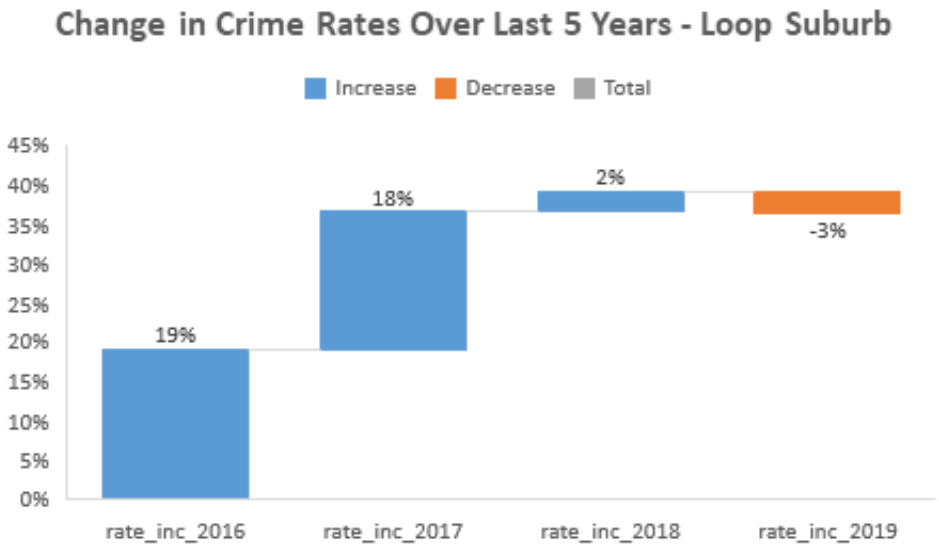
District	Crime
21	4
31	203
20	126522
17	209221
24	216708

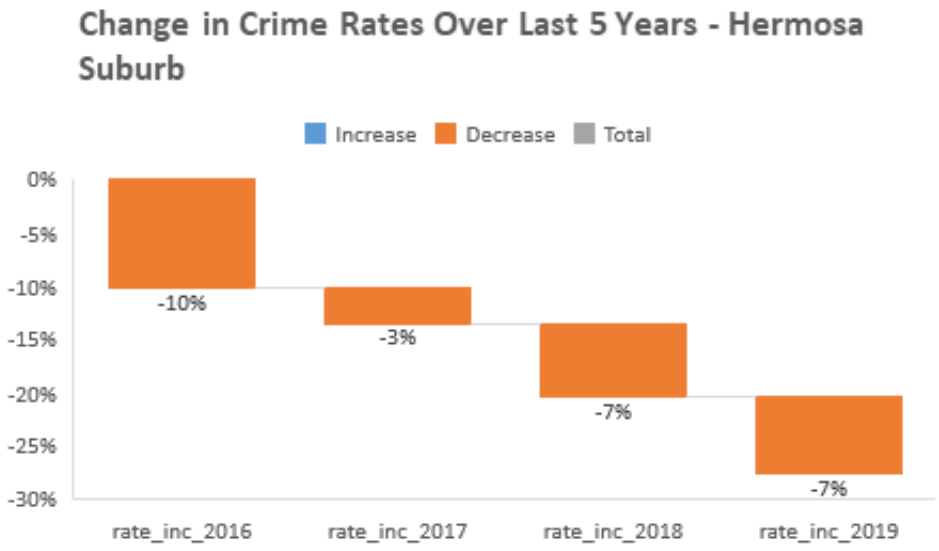
Districts 21 and 31 are newly added police districts and do not have much historical crime records. However, among the older ones, Lincoln is the district witnessing least crimes at approx. 6000 cases in a year. Albany Park and Rogers Park too feature among the districts with least crime

[4] https://en.wikipedia.org/wiki/Community_areas_in_Chicago
[5] https://home.chicagopolice.org/about/police-districts/

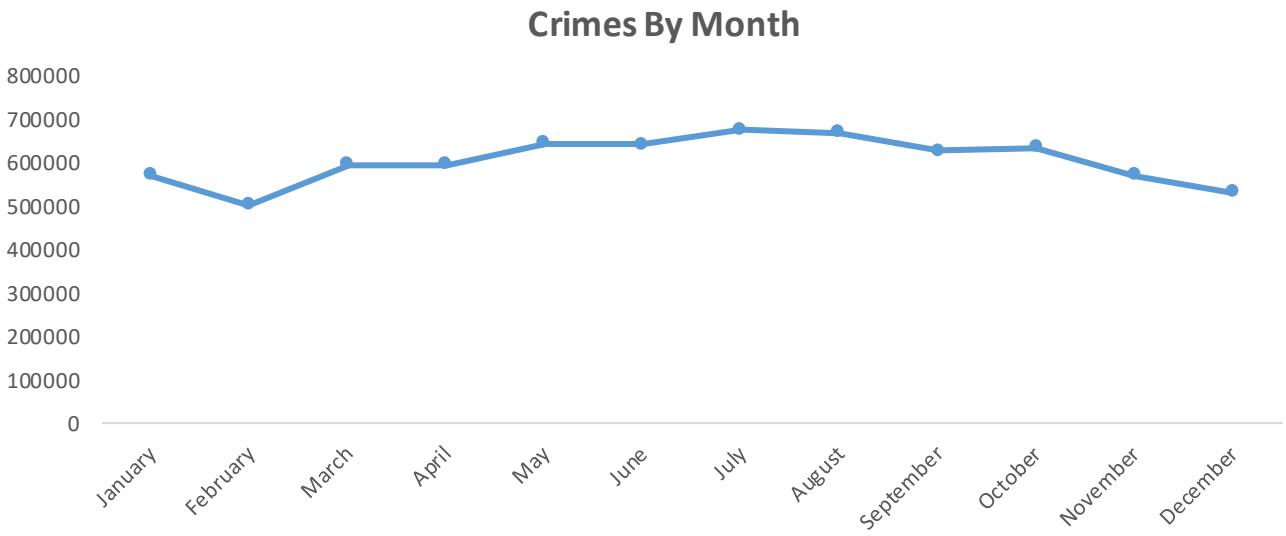
CRIME BY LOCALITY IN RECENT TIMES

28 community areas have witnessed an increase in crime rates over the last 5 years, whereas 49 community areas have witnessed a decline. Among the 28, The Loop suburb has witnessed the highest average increase (around 9%) in crimes over the last 5 years, whereas, Hermosa has seen the highest decline in terms of crime during the same time period.

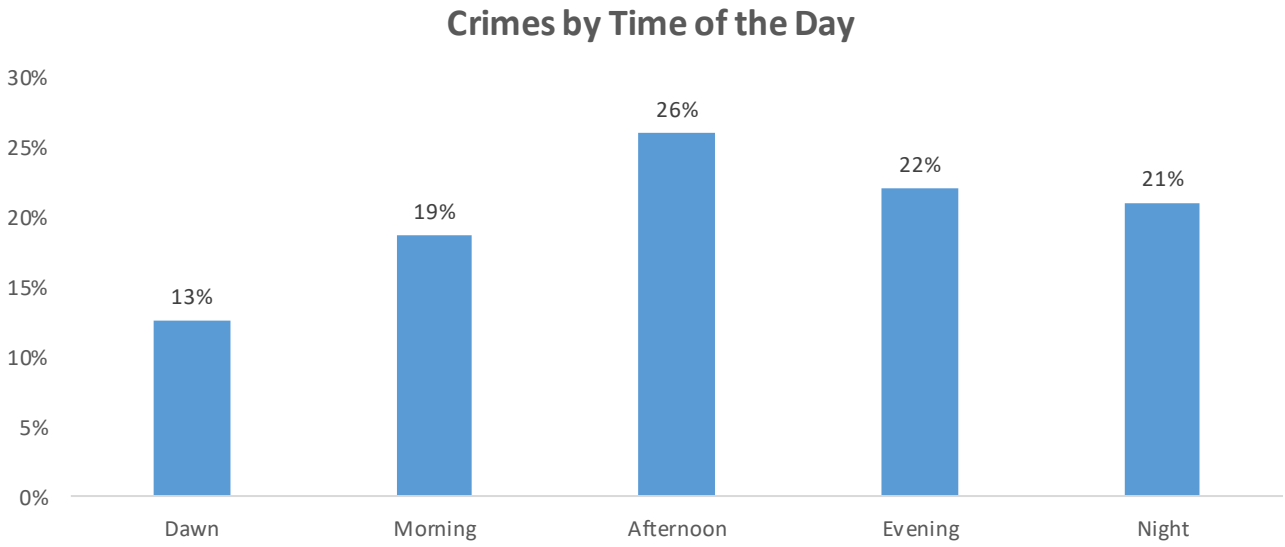




WHEN ARE THE CRIMES BEING COMMITTED?



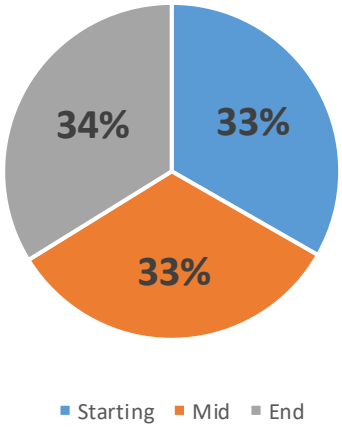
It is quite interesting to note that crime peak during the summers as months of May, June, July and August see maximum number of crimes being committed. At the same time, the winter months of December, January and February see a dip in crimes, probably due to the festive season. In terms of time of the day, to our extreme surprise, most of the crimes are committed in broad daylight in the afternoon and not night, contrary to the general



In an attempt to study further the patterns around when different kinds of crimes are being committed, we bifurcated the days of the month into three slots – the first 10 days as

'STARTING', the next 10 days as 'MID' and the remaining days as 'END' to see if any patterns come out. However, we witnessed an even spread across all the three blocks.

Breakup of Crimes by Block of Days in a Month

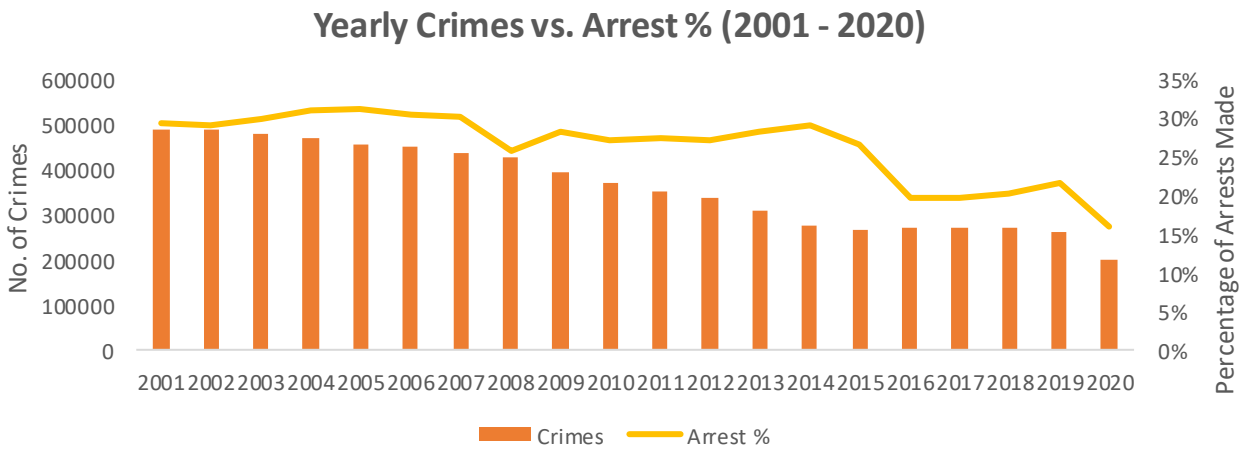


Crimes are evenly spread across starting, mid and end of every month

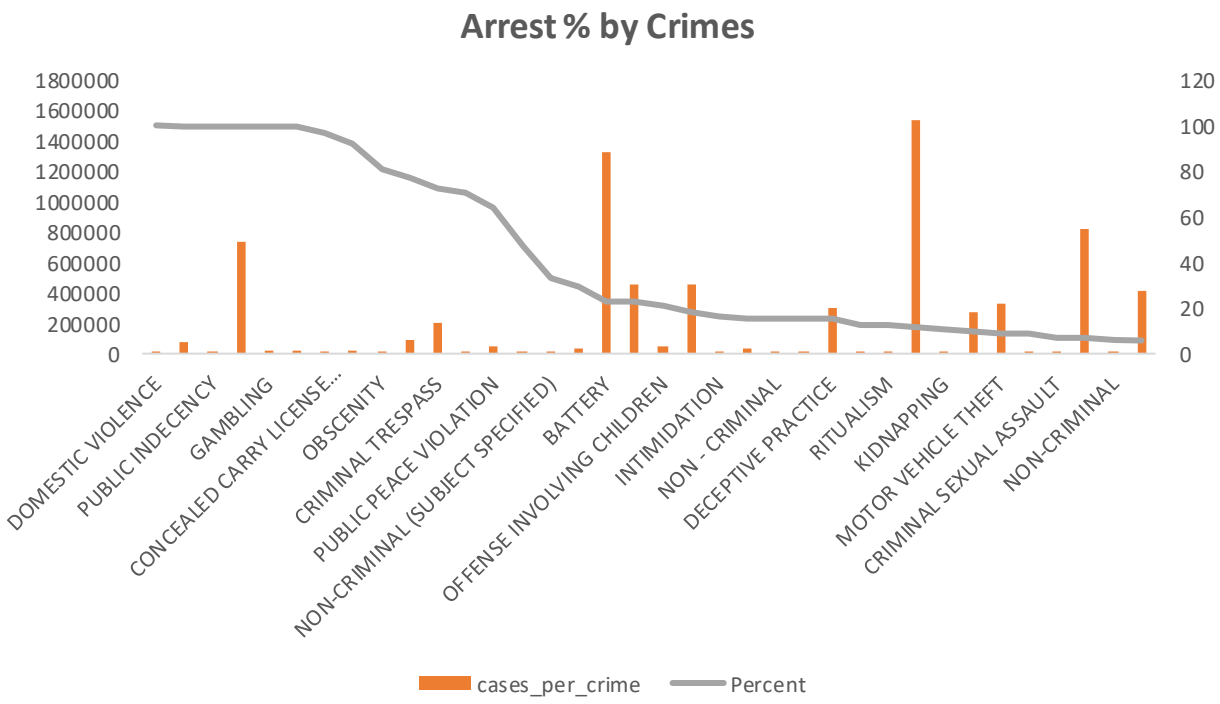
ACTION BY LAW ENFORCEMENT AUTHORITIES

The crimes being committed and the high historical crime rates in the city of Chicago are one part of the story. However, the fact that the quantum of crimes have gone down over the years through consistent vigilance and strict enforcement of laws by the authorities is a success story worth analysing. How has Chicago managed to bring down the crimes over the past two decades? Are there any patterns that point towards this? We will try and explore by analysing the arrests made over the years.

The average yearly arrests made over the last 20 years stands at around 26%, and we can see in the graph below that this number has more or less remained constant throughout, barring the last few years where it has hovered around the 20% mark. This signifies the consistency in vigilance on part of the law enforcement authorities of Chicago.



Looking at the arrest patterns by different types of crimes, we get to know that the enforcement authorities have approached different types of crime with varying degrees of strictness. There are some crimes like public indecency and narcotics, which have been low in quantum, but almost all cases have seen arrests being made. Similarly on the other side of the spectrum, there are crimes like Burglary and Criminal damage, which have seen a high number of cases over the years but only a small proportion of those having arrests.



CRIMES WITH HIGHEST ARREST PERCENTAGES

Primary Type	Overall_Arrest_Crime	cases_per_crime	Percent
DOMESTIC VIOLENCE	1	1	100.0
PROSTITUTION	69099	69368	99.61
PUBLIC INDECENCY	183	184	99.46
NARCOTICS	731173	735420	99.42
GAMBLING	14490	14594	99.29

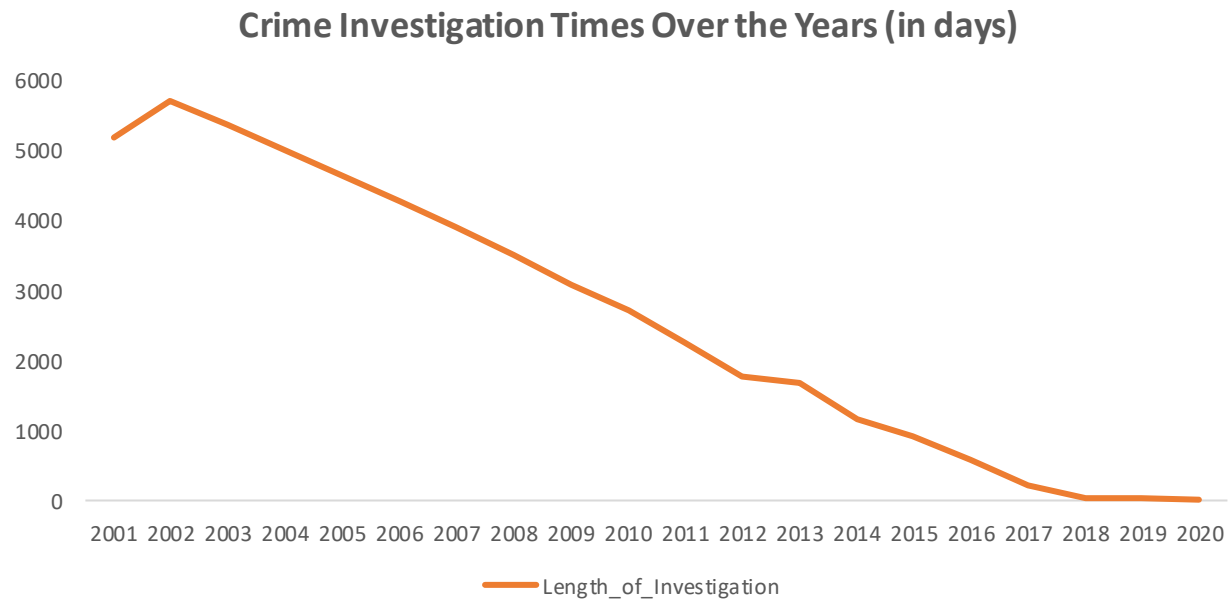
CRIMES WITH LOWEST ARREST PERCENTAGES

Primary Type	Overall_Arrest_Crime	cases_per_crime	Percent
BURGLARY	23406	407524	5.74
NON-CRIMINAL	11	173	6.36
CRIMINAL DAMAGE	57478	825269	6.96
CRIMINAL SEXUAL A...	171	2357	7.25
HUMAN TRAFFICKING	6	69	8.7

INVESTIGATION TIME

A critical factor to explore when it comes to actions by law enforcement authorities is the time being spent in investigating/closing a particular criminal case. An indicator of the same is the difference between the date when the crime was committed and the last date when the records for the particular crime were updated.

We get a graph with a declining trend in terms of average length of investigation by every year. Though, we get some idea of the expediting nature of the delivery of justice over the years, but this would not necessarily mean that the investigation times have gone down to such levels as there would be many records in recent years which would still be undergoing investigation and would be further updated in the coming years.



In terms of specific crimes, cases around domestic violence and ritualism go on for the maximum number of days, almost 14-15 years on average, and at the same time, crimes like human trafficking and concealed carry license violation are resolved a lot quicker.

Crimes with Maximum Investigation Period

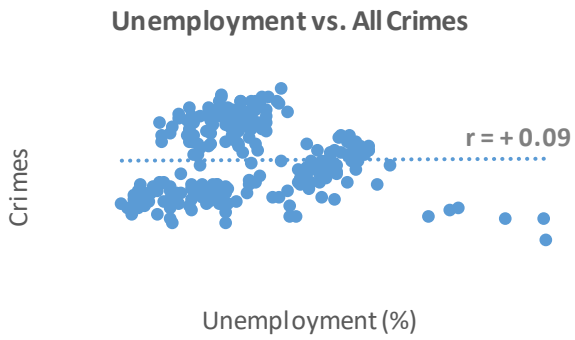
Crime_Type	Length_of_Investigation
DOMESTIC VIOLENCE	5331.0
RITUALISM	4590.0
PROSTITUTION	3933.0
LIQUOR LAW VIOLATION	3747.0
KIDNAPPING	3620.0

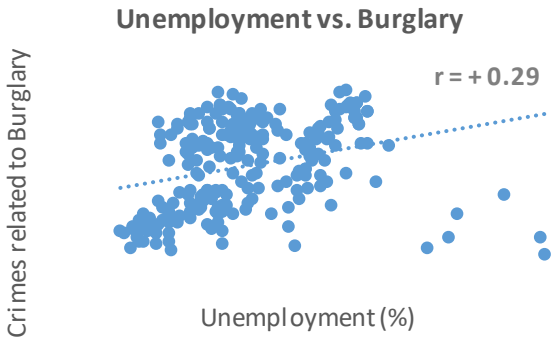
Crimes with Minimum Investigation Period

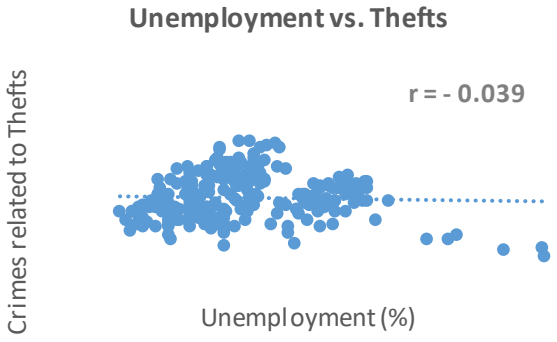
Crime_Type	Length_of_Investigation
CONCEALED CARRY LICENSE VIOLATION	134.0
HUMAN TRAFFICKING	509.0
NON-CRIMINAL	556.0
NON-CRIMINAL (SUBJECT SPECIFIED)	633.0
CRIMINAL SEXUAL ASSAULT	864.0

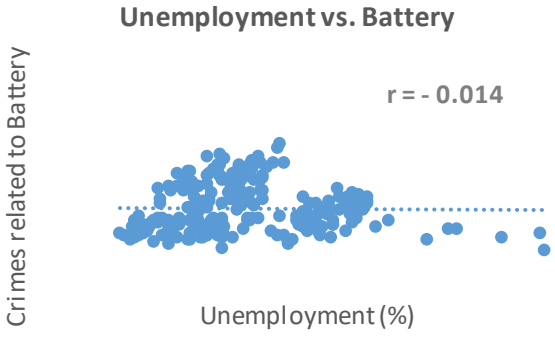
UNEMPLOYMENT CAUSING THE HIGH NUMBER OF CRIMES?

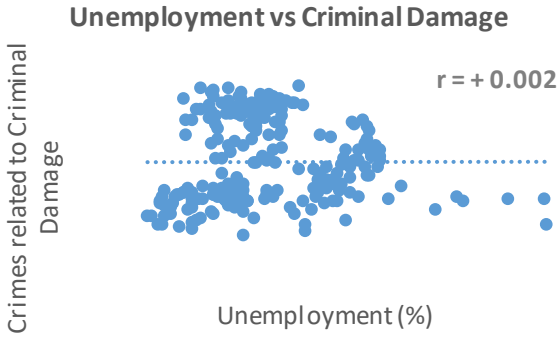
We next attempt to explore the potential reasons for the crimes being committed, especially, the high crime rates in the earlier parts of the 2000s. One of the major drivers for crime across worldwide is unemployment. A high proportion of the population being unemployed in most cases leads to a high number of crimes being committed. Thus, it would be meaningful to explore the relationship between unemployment rates and crime rates for Chicago, to establish if one of the major drivers for crime across the world is indeed driving crimes in Chicago as well.











We explore the relationship between monthly unemployment rates in the past 20 years with the overall and the top 4 most common crimes in the same time period. Contrary to our assumption, the unemployment rate and the number of crimes over the years do not seem

to have a strong correlation. Though some specific crimes do seem to show some sort of a positive relationship, however that too seems to be weak.

CONCLUSION

Despite its infamous reputation over the years as being among the topmost crime prone cities, the law enforcement authorities have done a great job of curbing crime in recent times. They have done this by maintaining a consistent arrest rate over the years and shortening the time taken to deliver justice for a particular crime.