Chicago Neighborhoods: Crime and Business Activities

1. Introduction

1.1. Background

In 2012, Chicago was called the “murder capital” of the USA and 2016 was the highest with 762 murders. The city often records the highest absolute total number of killings almost every year. There has been a significant drop in the past three years to about 240 murders per year. But if we look at murder rates - murders per 100,000 residents - Chicago regularly experiences much fewer killings than places like St. Louis, Baltimore, Kansas City, Missouri, Cleveland or Washington DC. The Chicago Police Department reported a 20% reduction of crime in May 2020 in the general category including murder, criminal sexual assault, robbery, aggravated battery, burglary, theft and vehicle theft from last May 2019. But discussions of crimes and business activities should not focus on cities but on neighborhoods. Just like income and education, crime and business activities vary even more in neighborhoods within cities than between cities.

Most business activities have been closed in Chicago since March 2020 due to Covid19. In this project, the focus will be to use Foursquare location services to explore venues in neighborhoods with interesting (high and low) results from the crime data.

1.2. Problem

What are the crime patterns and business activities in different neighborhoods in Chicago?

What is the relationship between crime and business activities in Chicago neighborhoods?

1.3. Stakeholders

Residents of the city would want to know the crime patterns and business activities in terms of location and time in their daily life. Prospective home buyers or renters need to know safe places and important business activities as part of their decision making process. Business investors would benefit from information about business activities in specific neighborhoods. Local government authorities would benefit from past crime activities and busy business areas to anticipate potential security needs.

2. Data

2.1. Data Sources

Chicago Data Portal is the main source of data for this project and it is provided by the City of Chicago. This data will be used to identify crime numbers and crime types across neighborhoods in different

periods of time. The data contains 22 columns and about 7 million incidents/rows. This includes crime incidents from 2001 to 2020. Not all columns are needed for this project and different restrictions - for example grouping by year, crime type, neighborhood, location - will be used for analysis.

Wikipedia webpage (table) will also be used for identifying official Chicago Area Numbers to identify their proper neighborhood names whenever needed. Unlike other cities Chicago's 77 official neighborhoods are designated by numbers instead of neighborhood names.

Foursquare will also be used to explore business activities and geographical locations. It will be used to get trending venues around a given location in some neighborhoods with high, medium and low crimes.

Finally, both the Chicago Data Portal and Foursquare data will be used to explore relationships between crime and business activities.

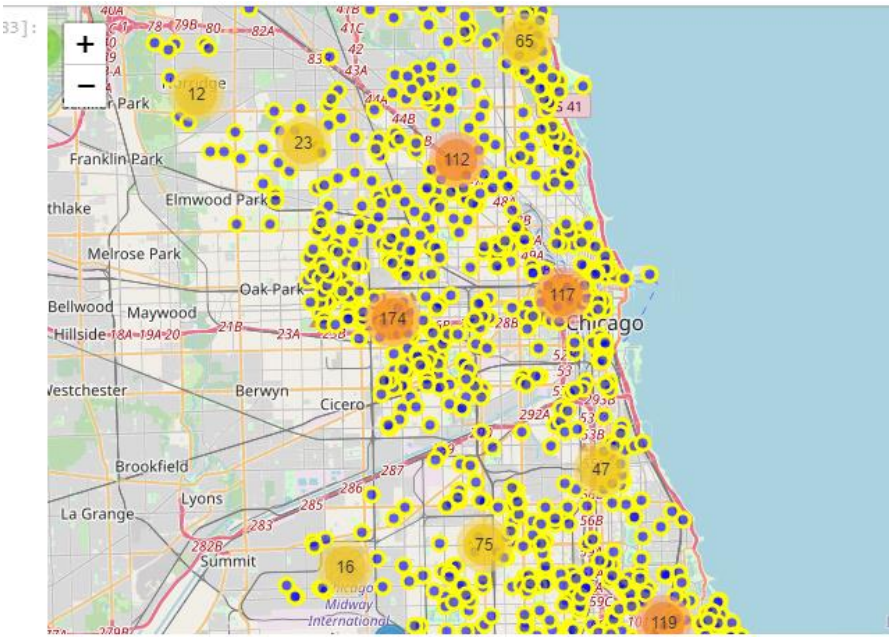
2.2. Data Cleaning

Due to the limited capacity of my computer to handle running codes on about 7 million rows(samples) and 22 columns(features), composed of crime incidents in the City of Chicago from 2001 to 2020, it became apparent that I should reduce the size of the data. After considering different options, I concluded that just the year 2020 data has all relevant and most recent information. Then all rows with missing values were discarded to further reduce the size and prepare the data for further analysis. The neighborhood information was provided as community area numbers. A column was added, from a Wikipedia page, to include the proper names of the neighborhoods. Some of the columns, such as FBI code and IUCR, were also discarded because they were not relevant to the study. This reduced the data size to about 98,000 rows and 10 columns.

3. Methodology

A quick look at the Dataframe table and map with markers of the incidents indicate that there were many crimes all over the city. The map was overcrowded and any meaningful inference was not possible. An exploratory data analysis was performed for a closer look by grouping the crime incidents by neighborhood, type-of-crime, and day-of-the week. The number of venues on four neighborhoods is also analyzed. As a remedy for the crowded markers, a *MarkerCluster* object was initiated to group the markers into different clusters. Each cluster is then represented by the number of crimes in different parts of the city.

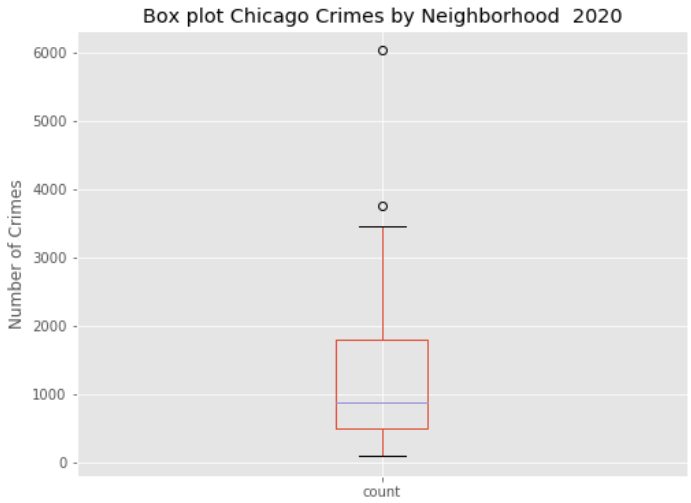
Crime incidents with different clusters of crime numbers. This is a screenshot of the live output generated by a set of codes.



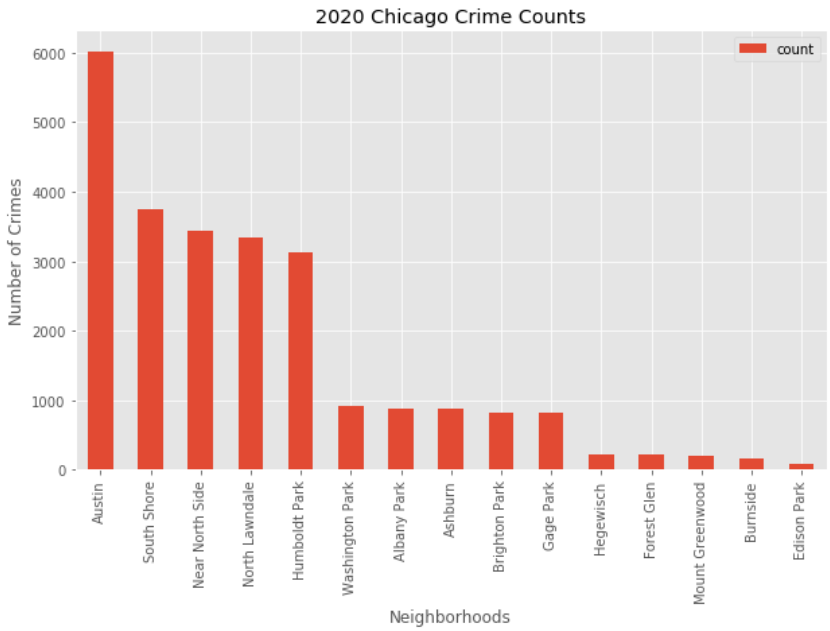
3.1. Number of incidents by neighborhood

A simple count of crime incidents by neighborhood is used to show the distribution of crimes across the 77 neighborhood areas. The number of crimes – frequencies – for each of the 77 neighborhoods were tabulated and graphed. A percentage of the crime counts for each neighborhood was calculated. A simple Chi-Square test of Goodness of fit is performed to check if the crimes are uniformly distributed across the neighborhoods.

The box-plot and summary statistics show that the values ranged from the smallest of 95 crimes in Edison Park neighborhood to the maximum of 6,021 crimes in Austin. The box-plot indicates, with small circles, that the top two – Austin with 6,021 and South Shore with 3,743 are outliers. The Chi-Square test of Goodness of Fit has a p-value of 0.0. This is another strong evidence that the crime incidents are not uniformly distributed across neighborhoods.



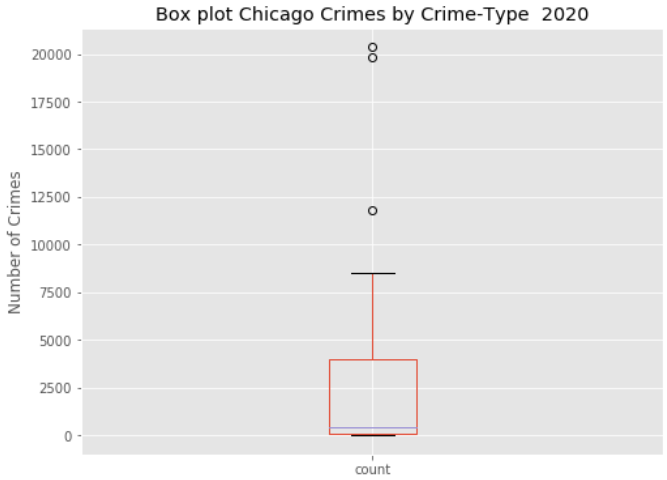
The bar graph below is based on the top 5, middle 5 and bottom 5 neighborhoods.



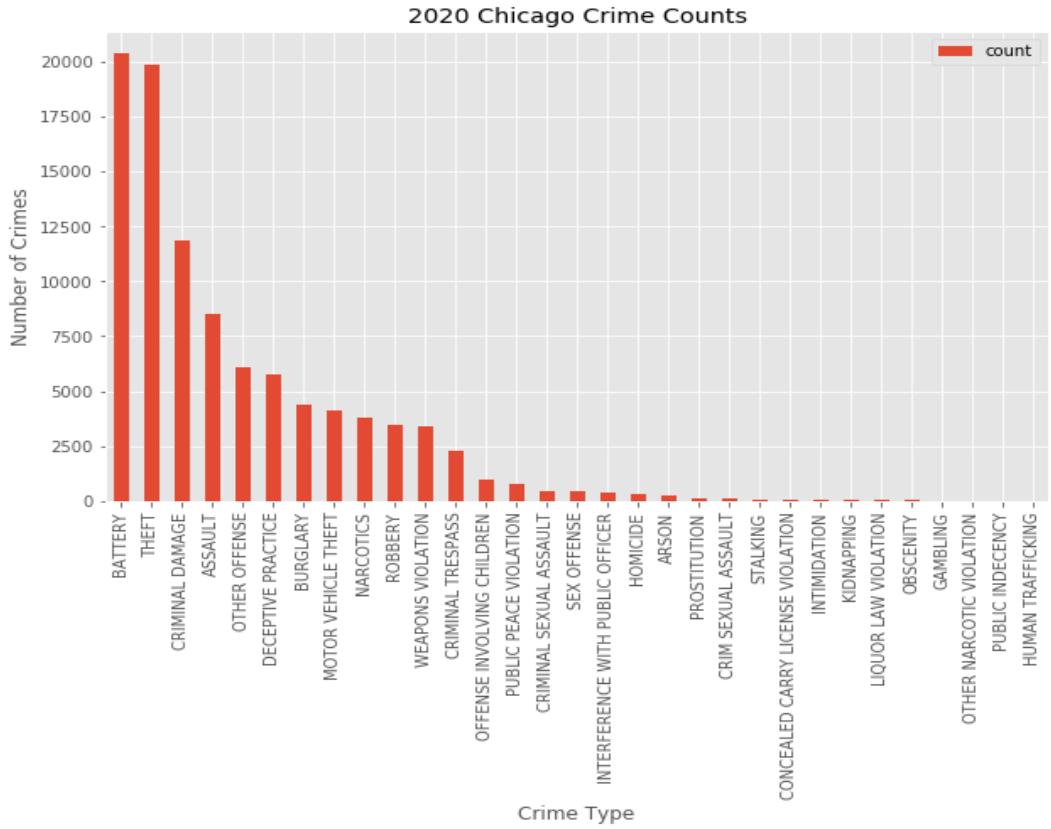
3.2.Number of incidents by type-of-crimes

A simple count of crime incidents by neighborhood is used to show the distribution of crimes across the 31 types of crimes reported. The number of crimes – frequencies – for each of the 31 types of crimes are tabulated and graphed. A simple Chi-Square test of Goodness of fit is performed to check if the crimes are uniformly distributed across the different types of crimes.

The box-plot and summary statistics show that the values ranged from the minimum of 3 crimes in Human Trafficking to the maximum of 20,356 in Battery. The box-plot indicates, with small circles, that the top three types of crimes – Battery, Theft and Criminal Damage - are outliers. The Chi-Square test of Goodness of Fit has a p-value of 0.0. This is another strong evidence that the crime Types are not uniformly distributed across all types of crimes.



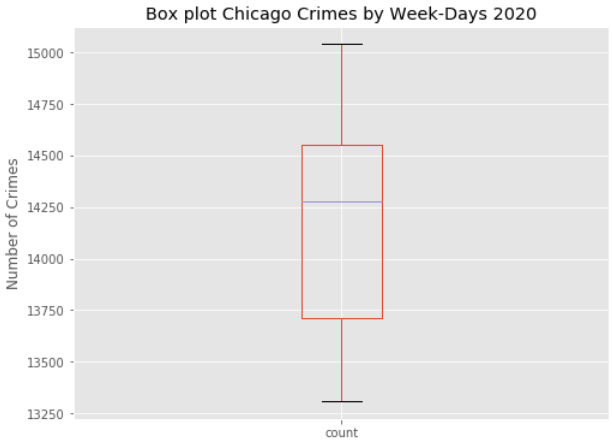
Below is a bar graph of the crime-types.



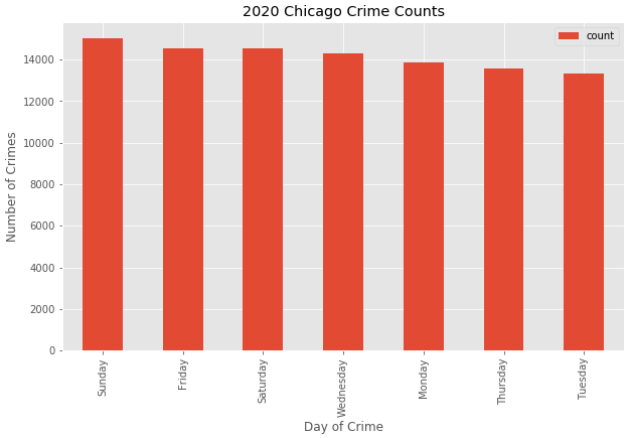
3.3. Number of incidents by day-of-week

A simple count of crime incidents by day-of-week, Sunday to Monday, is used to show the distribution of crimes across 7 days of the week. The number of crimes – frequencies – for each of the 7 days of the week are tabulated and graphed. A simple Chi-Square test of Goodness of fit is performed to check if the crime incidents are uniformly distributed across the seven days of the week.

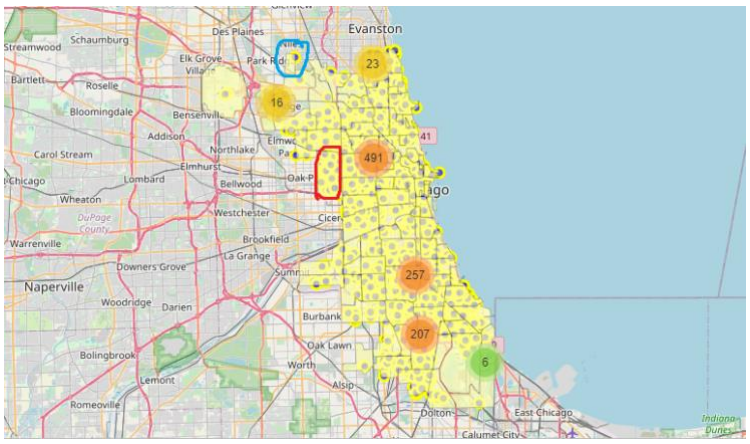
The box-plot and summary statistics show that the values ranged from the minimum of 13,309 crimes on Tuesdays to the maximum of 15,039 on Sundays. The box-plot does not indicate the presence of any outliers but the Chi-Square test of Goodness of Fit has a p-value of 7.25498977771133e-32. Thus there is evidence that the days of the week are not uniformly distributed across the days of the week.



Below is the bar graph for the day-of-the week.



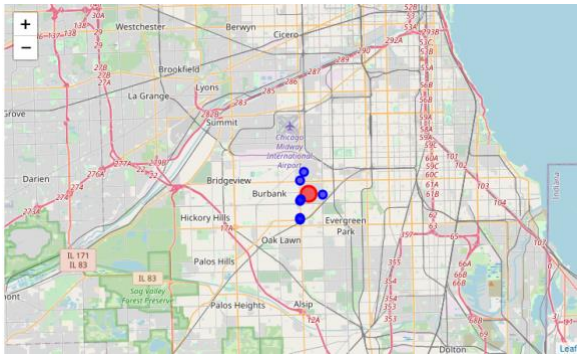
The picture below is a screenshot from a set of codes of Python on Notebook. Austin, the neighborhood with the highest crime circled red and Edison Park with the lowest crime is circled blue. The clustered numbers are grouped crime counts.



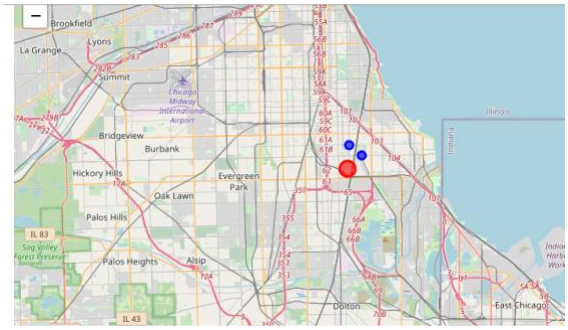
3.4. Number of venues on four neighborhoods

A location was selected from four neighborhoods with high, medium and low crime incident counts. A search and map for all restaurants was performed through Foursquare API as one measure of business activity in the neighborhood. Here, it is assumed that *more* restaurants – of any type – is associated with *more* business activities. Five neighborhoods with different crime numbers were selected to study business activities. Three neighborhoods were selected from high crime clusters - Ashburn, Chatham and North Lawndale. The Edison Park with the lowest crime number neighborhood and the Loop, downtown which is known for high business activity were included for comparison. From each of these neighborhoods a location with latitude and longitude values was selected. A search query for restaurant – any type of restaurant – was performed for this particular location. The result was transformed into dataframe and filtered with those with venue names associated with location. A visual map was created for all this filtered results. The screenshot for each neighborhood given below. All neighborhoods with the exception of The Loop show low restaurants.

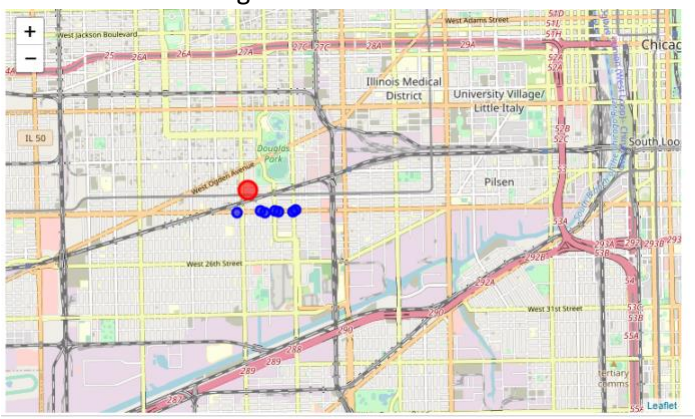
Ashburn Neighborhood



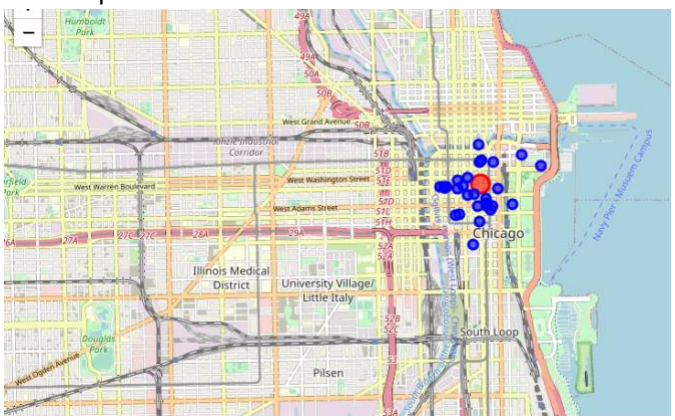
Chatham Neighborhood



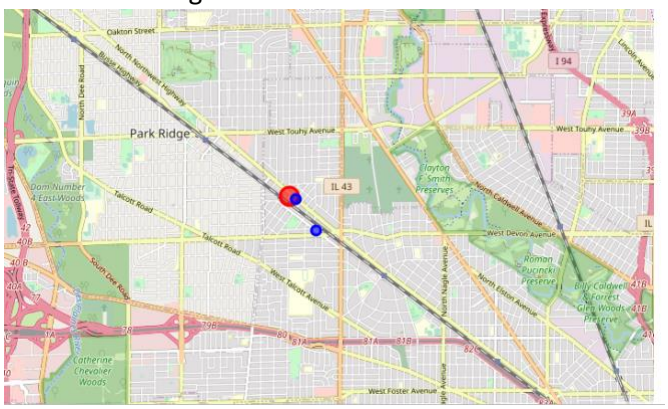
North Lawndale Neighborhood



The Loop



Edison Park Neighborhood



3.5. Machine learning used

Pandas, NumPy, SciPy, and Motplotlib libraries are used in this project. Pandas and NumPy are extensively used for data structure and data analysis. Motplotlib is used for visual representation of

data in different types of plotting. SciPy is used for hypothesis testing – Chi Square test of goodness of fit.

4. Results

The crime incident numbers in the City of Chicago are not evenly distributed throughout the City. For the current year of 2020 through the month of June, about 98,000 crime incidents were recorded by the City. From all the 77 neighborhoods in Chicago, Austin and South Shore have significantly high crime incidents. Edison Park has the lowest crime incidents. But the population and area of Austin is about eight times as that of Edison Park. There were 331 homicides in the City and Austin has the highest homicides.

The different type of crimes were not evenly distributed with Battery and theft as the most common types of crimes accounting for 40 percent of the crimes.

There were more crime incidents recorded on Sundays than any other days and Tuesdays the least.

Overall business activities were low in both high and low crime neighborhoods. Low business activities in low crime neighborhoods could probably be explained by the fact that this neighborhoods tend to be more residential and family type of homes. But The Loop – downtown Chicago ranked 12th on the list of number of crime incidents - has a good business activity. Overall the number of crimes was not generally associated with the number of business activities. The assumption that more restaurants implies more business activity may not have been valid.

5. Discussion

The amount of variation in terms of the number of crimes from neighborhood to neighborhood is vast. The average number of crimes for all 77 neighborhoods is about 1,270 with a big standard deviation of 1,090. Chicago may be known for crimes in the USA but it is important to note that most of the crimes occur in few part of the City. In addition, the biggest portion of the crimes reported are battery and theft. I recommend that city government and citizens spend more resources studying the causes of high crimes in Austin and South Shore and implement changes to address the causes.

I also recommend that more precaution be made by residents on Sundays as it has the highest crime numbers.

The study showed that there was no significant relationship between business activities and crime numbers in the four neighborhoods selected: Ashburn, Chatham, North Lawndale and The Loop.

The amount of data available was too much for my personal computer to handle. Bigger and faster machines would have been able to handle all the data and produce more comprehensive results.

6. Conclusion

This study was focused on identifying crime patterns and business activities in the neighborhoods of the City of Chicago. There was a big difference in the neighborhoods. Austin and South Shore had very high crime numbers while Edison Park and Burnside had the lowest. The types of crimes were also not distributed evenly with battery and theft being the most common. There were more crimes incidents on Sundays than any other day. Overall there was low business activities in both high and low crime neighborhoods.