Coursera Capstone Report

Ian Smith

6th February 2020

Introduction:

Chicago is one of the largest cities in the US. With a population of over 2.7 million as of 2017 it is a hugely important cultural and economic hub in the US. It also boasts one of the largest crime rates in the US. I will be attempting to perform analysis on Chicago crime data for the year 2017.

As the city grows and develops, it becomes increasingly important to examine and understand it quantitatively.

Some key questions that I would like to answer are:

1. What is the crime count by community area in Chicago?
2. What are the different areas within Chicago that have the highest number of aggregate crimes?

Developers and people looking to move to the city would be interested in the above questions. Naturally people would like to avoid areas of high crime and policy makers would like to know what areas have the highest amount of crime in order to implement policies to reduce crimes in those areas.

Data:

The data sources that I will be making use of for this project:

1. Chicago Crime Data - https://www.kaggle.com/currie32/crimes-in-chicago

Using the above datasets will allow us to answer the questions. The Chicago crime data will allow us to see what neighbourhoods have the most crime and the foursquare location data will allow us to visualise the venues that are most common in different locations around the city. The column names in the above dataset are as follows:

- ID - (ID number for the crime)
- Case Number
- Block - (The block where the crime took place)
- Primary Type - (The primary type of crime that took place)
- Battery
- Description - (A description of the type of crime that took place)
- Location Description - (Location where the crime took place - apartment, residence etc)

- Arrest - (States whether the crime resulted in an arrest or not)
- Domestic - (Whether the crime was domestic or not)
- Beat
- District
- Ward
- Community Area
- FBI Code
- Year
- Updated On
- Latitude - (Latitude coordinate)
- Longitude - (Longitude coordinate)
- Location - (This has latitude as the x-coordinate and longitude as the y-coordinate)

Methodology:

The dataset was obtained on Kaggle. The crime data from 2012-2017 was selected as this was the most recent data. I specifically wanted to examine data from 2016 as this was a pivotal year both in America and across the world.

Firstly I imported relevant libraries such as pandas and numpy. I then proceeded to load the dataset as a pandas dataframe which then looked like this.

	Unnamed: 0	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	...	Ward	C A
0	3	10508693	HZ250496	05/03/2016 11:40:00 PM	013XX S SAWYER AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	True	...	24.0	2
1	89	10508695	HZ250409	05/03/2016 09:40:00 PM	061XX S DREXEL AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	False	...	20.0	4
2	197	10508697	HZ250503	05/03/2016 11:31:00 PM	053XX W CHICAGO AVE	0470	PUBLIC PEACE VIOLATION	RECKLESS CONDUCT	STREET	False	...	37.0	2
3	673	10508698	HZ250424	05/03/2016 10:10:00 PM	049XX W FULTON ST	0460	BATTERY	SIMPLE	SIDEWALK	False	...	28.0	2
4	911	10508699	HZ250455	05/03/2016 10:00:00 PM	003XX N LOTUS AVE	0820	THEFT	$500 AND UNDER	RESIDENCE	False	...	28.0	2

After the dataset was loaded I proceeded to clean the dataset. I dropped a few irrelevant columns such as x-coordinate, y-coordinate and Unnamed: 0. I then proceeded to get the crimes from 2016 only as this was the year I was interested in.

I then selected the crimes by community area. All of the values in the community area column were numeric so I converted them to their proper names. I then obtained the number of crimes by community area. I also obtained descriptive statistics for the number of crimes. A snippet of the dataframe is given below.

	Community Area	Count
0	Albany Park	2387
1	Archer Heights	876
2	Armour Square	1013
3	Ashburn	2689
4	Auburn Gresham	7536
5	Austin	16462
6	Avalon Park	1298
7	Avondale	2263
8	Belmont Cragin	4716
9	Beverly	964

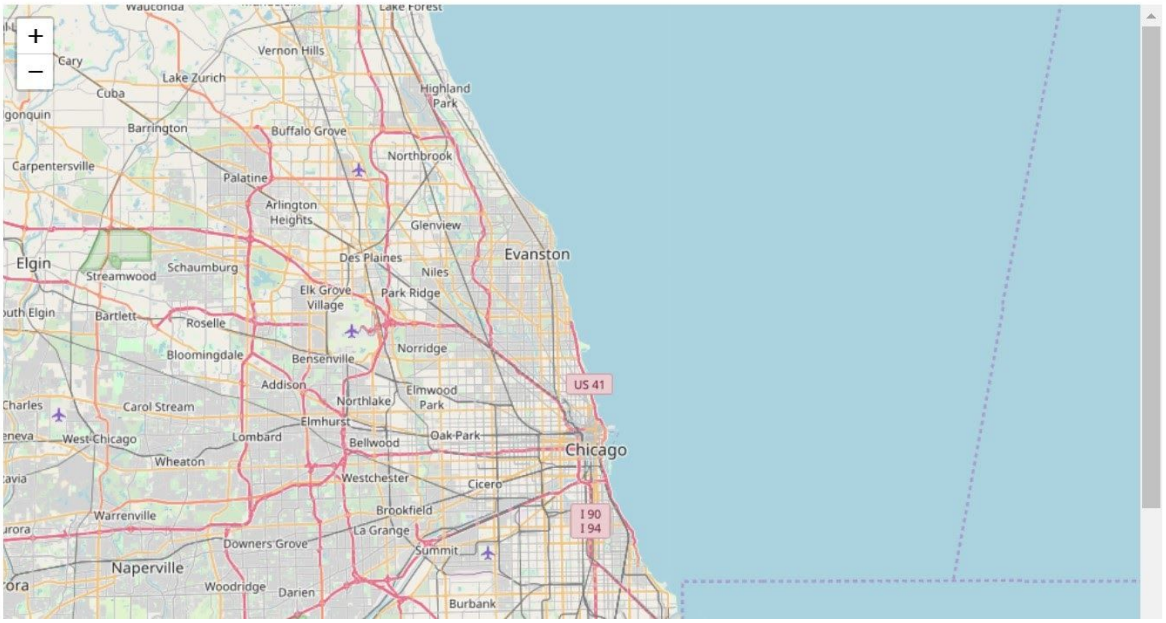
I also examined the different types of crimes.

	Description	Crime_Type_Count
0	$500 AND UNDER	24124
1	ABUSE/NEGLECT: CARE FACILITY	4
2	AGG CRIM SEX ABUSE FAM MEMBER	129
3	AGG CRIMINAL SEXUAL ABUSE	166
4	AGG PO HANDS ETC SERIOUS INJ	19
5	AGG PO HANDS NO/MIN INJURY	861
6	AGG PRO EMP HANDS SERIOUS INJ	34
7	AGG PRO.EMP: HANDGUN	30
8	AGG PRO.EMP: OTHER DANG WEAPON	101
9	AGG PRO.EMP: OTHER FIREARM	3
10	AGG PRO.EMP:KNIFE/CUTTING INST	47
11	AGG SEX ASSLT OF CHILD FAM MBR	49
12	AGG: HANDS/FIST/FEET NO/MINOR INJURY	138
13	AGG: HANDS/FIST/FEET SERIOUS INJURY	158
14	AGGRAVATED	769

I examined the number of aggregate crimes and obtained a similar dataframe to the one above which is again given below.

	Community Area	Count
0	Albany Park	6
1	Archer Heights	1
2	Armour Square	9
3	Ashburn	9
4	Auburn Gresham	24
5	Austin	57
6	Avalon Park	1
7	Avondale	6
8	Belmont Cragin	18
9	Beverly	1
10	Bridgeport	7
11	Brighton Park	7
12	Calumet Heights	3
13	Chatham	18
14	Chicago Lawn	19

I also developed a map of Chicago using folium which is given below.



Results:

As I mentioned above it is interesting to note the sheer number of crimes in Chicago! In the area of Ashburn there were 2689 crimes. Although not all of those crimes resulted in arrests and many of them may have been false calls it is still interesting to note the sheer number of crimes in Chicago.

With respect to the number of crimes it is important to look at the descriptive statistics which can be easily obtained using the .describe() function. The average (mean) number of crimes was 3447 which is quite large!

Discussion:

The analysis that was carried out shows that lawmakers and policy makers have their work cut out for them! The huge number of crimes in Chicago across a number of different community areas shows that a large amount of work is needed to fix the problem of crime in Chicago.

Also, for prospective buyers/renters of homes and apartments in Chicago, location is vitally important as ever. If a family decided to move to a certain community area they must be aware of the number of crimes that took place and what type of crimes that took place.

Also additional analysis of the Chicago crime data can take place. We can see if certain crimes are more likely to result in arrest. We can look at trends over a number of years in crime. All of these questions can be answered using a number of different types of ML algorithms and techniques. We can introduce additional datasets such as population to see if population density or economic status affect crime.

Conclusion:

This small project shows the power of data analysis in decision making for policy makers, developers and families to name just a few. I have managed to answer questions set out in the introduction of the report. Also, the project provides scope for improvement with additional questions as mentioned in the results section.