Analysis of criminal situation in Chicago from 2001-2019

Authors: Ken Pan, Sicong Chang

The research Questions:

1. What is the ranking of safety of different district of Chicago?

The safety is computed by the number of occurrence and weight of different types of crime.

Crime whose type is more severe should have higher weigh.

The result to this question can be a reference for people who want to live in Chicago.

2. Predict the possibility of the criminal being arrested at a given time of a day and a given criminal type.

Use machine learning model to predict the possibility of solving a crime.

Give the police an overview of the difficulty of the case and remind them to concentrate more on cases that are hard to be solved.

Can also be a reference to the victim to see how likely can their case can be solved and can provide an safety education for them.

3. Does the increase or decrease in rate of solved case affect the crime rate? Does the result vary from different criminal types?

We want to use the data from earlier years.

First compute the proportion of solved cases for each type of crime for each year.

Then see if the change of the police performance can affect the amount of crimes in later years.

It could be a motivation for police to keep improving their performance.

4. How does poverty rate, unemployment rate, educational level and age distribution among people affect the crime rate in a particular community area during a specific period of time? What is the most contributing factor and what is the least contributing factor?

We want to use other datasets to see the relationship between other social factors and crime rate as a guide to government decisions.

Motivation

Public safety has always been a major concern of people especially for those who live in large and crowded cities. So, cities are great objects to study on crime situation. As a result, we decided to analyze the crime situation with example of The City of Chicago, which is a city with a very complete public dataset of cases from 2001 to present.

The first question we ask is the ranking of safety of different community area of Chicago. The result to this question can remind the police of which areas they need to take special care with. It can also be a scientific reference to those who want to move to Chicago because crime situation is definitely an important aspect they need to consider when choosing a place to live and work.

The second question we ask is to predict the possibility of the criminal being arrested at a given time, a given location, and a given crime type.

The result of this can be a reference to both the police and victim. The result can be a general approximation of hardship of solving the case. This can guide police on how much force to put on a case and how to arrange their work. For the victim, this data can also make them be mentally prepared for the case result.

The third question we ask is whether the correlation between rate of case being solved with the changing of crime rate in the later years.

This question can be a reference to the police to motivate them to keep improving their performance. If we don’t know the answer to the question, the police would not know if their hardworking have long term effect on crime rate in the city.

The fourth question is that how does poverty rate, unemployment rate, educational level and age distribution among people affect the crime rate in a particular community area during a specific period of time? What is the most contributing factor and what is the least contributing factor?

The previous questions are all questions directly study crime situation. In this question, we can go behind the scene to see what might cause the variation of crime rate in different districts and in different times. Knowing this would give us an insight into the cause of crime rate changing and can help the government to take steps to work on some of the most important aspects to eliminate crime.

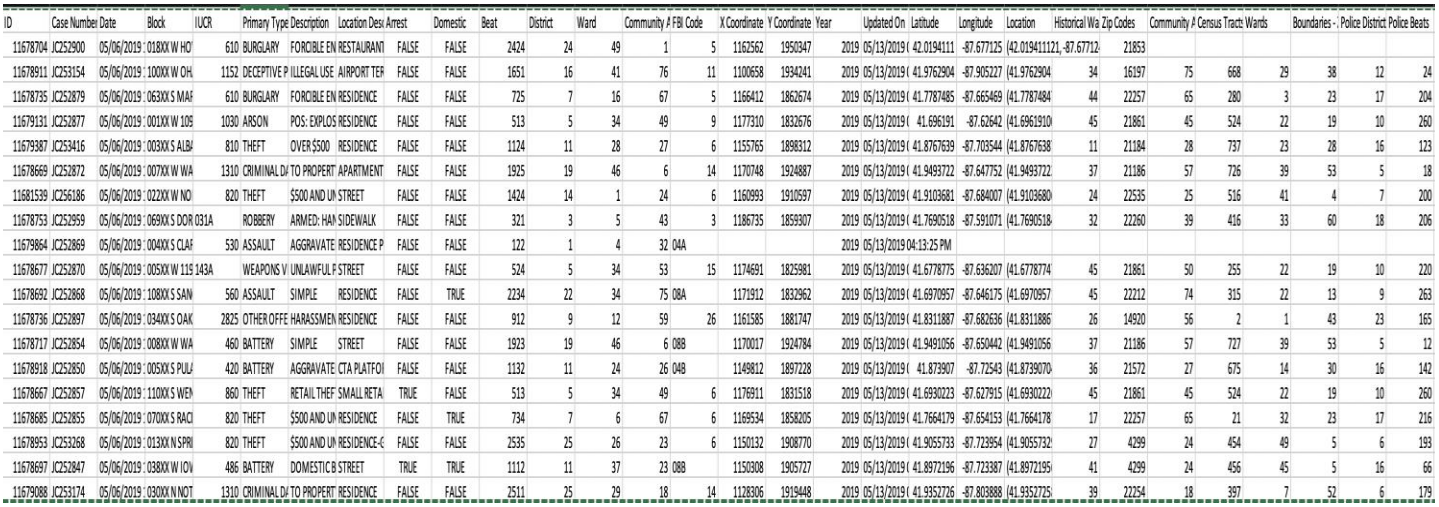
Data set:

The main data set we choose is crimes from 2001 to present in Chicago. It is published by the government of city of Chicago

The url is below

https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD

It’s a csv file almost without missing data. The frame size is 6866498 * 30. The index of the frame is ID for each case. The columns include some useless data to me such as case number and FBI number. However, most of the columns in the dataset are useful. The time of the case include hour and minutes. It allows us to analyze annually changing rate or the situation in a specific time period of a day. There are also lots of location information such as the block, position coordinate, district number and police beats, which allow us to graph map (only on block scale to protect privacy). The type of crimes is also well stated for most of the cases. There is also a column of Boolean which shows if the case is solved or not. The column is extremely useful for analyzing the performance of police on different type of cases.



ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Desc	Arrest	Domestic	Beat	District	Ward	Community A	FBI Code	X Coordinate	Y Coordinate	Year	Updated On	Latitude	Longitude	Location	Historical Wa	Zip Codes	Community A	Census Tracts	Wards	Boundaries -	Police District	Police Beats
11678704	JC252900	05/06/2019 :	018XX W HO	610	BURGLARY	FORCIBLE EN	RESTAURANT	FALSE	FALSE	2424	24	49	1	5	1162562	1950347	2019	05/13/2019 (	42.0194111	-87.677125	(42.019411121, -87.67712		21853						
11678911	JC253154	05/06/2019 :	100XX W OH	1152	DECEPTIVE P	ILLEGAL USE	AIRPORT TER	FALSE	FALSE	1651	16	41	76	11	1100658	1934241	2019	05/13/2019 (	41.9762904	-87.905227	(41.9762904	34	16197	75	668	29	38	12	24
11678735	JC252879	05/06/2019 :	063XX S MAF	610	BURGLARY	FORCIBLE EN	RESIDENCE	FALSE	FALSE	725	7	16	67	5	1166412	1862674	2019	05/13/2019 (	41.7787485	-87.665469	(41.7787484	44	22257	65	280	3	23	17	204
11679131	JC252877	05/06/2019 :	001XX W 109	1030	ARSON	POS: EXPLOS	RESIDENCE	FALSE	FALSE	513	5	34	49	9	1177310	1832676	2019	05/13/2019 (	41.696191	-87.62642	(41.6961910	45	21861	45	524	22	19	10	260
11679387	JC253416	05/06/2019 :	003XX S ALBA	810	THEFT	OVER $500	RESIDENCE	FALSE	FALSE	1124	11	28	27	6	1155765	1898312	2019	05/13/2019 (	41.8767639	-87.703544	(41.8767638	11	21184	28	737	23	28	16	123
11678669	JC252872	05/06/2019 :	007XX W WA	1310	CRIMINAL DA	TO PROPERT	APARTMENT	FALSE	FALSE	1925	19	46	6	14	1170748	1924887	2019	05/13/2019 (	41.9493722	-87.647752	(41.9493722	37	21186	57	726	39	53	5	18
11681539	JC256186	05/06/2019 :	022XX W NO	820	THEFT	$500 AND UN	STREET	FALSE	FALSE	1424	14	1	24	6	1160993	1910597	2019	05/13/2019 (	41.9103681	-87.684007	(41.9103680	24	22535	25	516	41	4	7	200
11678753	JC252959	05/06/2019 :	069XX S DOR	031A	ROBBERY	ARMED: HAN	SIDEWALK	FALSE	FALSE	321	3	5	43	3	1186735	1859307	2019	05/13/2019 (	41.7690518	-87.591071	(41.7690518	32	22260	39	416	33	60	18	206
11679864	JC252869	05/06/2019 :	004XX S CLAF	530	ASSAULT	AGGRAVATE	RESIDENCE P	FALSE	FALSE	122	1	4	32	04A			2019	05/13/2019 04:13:25 PM											
11678677	JC252870	05/06/2019 :	005XX W 119	143A	WEAPONS V	UNLAWFUL P	STREET	FALSE	FALSE	524	5	34	53	15	1174691	1825981	2019	05/13/2019 (	41.6778775	-87.636207	(41.6778774	45	21861	50	255	22	19	10	220
11678692	JC252868	05/06/2019 :	108XX S SAN	560	ASSAULT	SIMPLE	RESIDENCE	FALSE	TRUE	2234	22	34	75	08A	1171912	1832962	2019	05/13/2019 (	41.6970957	-87.646175	(41.6970957	45	22212	74	315	22	13	9	263
11678736	JC252897	05/06/2019 :	034XX S OAK	2825	OTHER OFFE	HARASSMEN	RESIDENCE	FALSE	FALSE	912	9	12	59	26	1161585	1881747	2019	05/13/2019 (	41.8311887	-87.682636	(41.8311886	26	14920	56	2	1	43	23	165
11678717	JC252854	05/06/2019 :	008XX W WA	460	BATTERY	SIMPLE	STREET	FALSE	FALSE	1923	19	46	6	08B	1170017	1924784	2019	05/13/2019 (	41.9491056	-87.650442	(41.9491056	37	21186	57	727	39	53	5	12
11678918	JC252850	05/06/2019 :	005XX S PULA	420	BATTERY	AGGRAVATE	CTA PLATFOI	FALSE	FALSE	1132	11	24	26	04B	1149812	1897228	2019	05/13/2019 (	41.873907	-87.72543	(41.8739070	36	21572	27	675	14	30	16	142
11678667	JC252857	05/06/2019 :	110XX S WEN	860	THEFT	RETAIL THEF	SMALL RETA	TRUE	FALSE	513	5	34	49	6	1176911	1831518	2019	05/13/2019 (	41.6930223	-87.627915	(41.6930222	45	21861	45	524	22	19	10	260
11678685	JC252855	05/06/2019 :	070XX S RACI	820	THEFT	$500 AND UN	RESIDENCE	FALSE	TRUE	734	7	6	67	6	1169534	1858205	2019	05/13/2019 (	41.7664179	-87.654153	(41.7664178	17	22257	65	21	32	23	17	216
11678953	JC253268	05/06/2019 :	013XX N SPRI	820	THEFT	$500 AND UN	RESIDENCE-G	FALSE	FALSE	2535	25	26	23	6	1150132	1908770	2019	05/13/2019 (	41.9055733	-87.723954	(41.9055732	27	4299	24	454	49	5	6	193
11678697	JC252847	05/06/2019 :	038XX W IOV	486	BATTERY	DOMESTIC B	STREET	TRUE	TRUE	1112	11	37	23	08B	1150308	1905727	2019	05/13/2019 (	41.8972196	-87.723387	(41.8972195	41	4299	24	456	45	5	16	66
11679088	JC253174	05/06/2019 :	030XX N NOT	1310	CRIMINAL DA	TO PROPERT	RESIDENCE	FALSE	FALSE	2511	25	29	18	14	1128306	1919448	2019	05/13/2019 (	41.9352726	-87.803888	(41.9352725	39	22254	18	397	7	52	6	179

We will also need the geo information of Chicago

url: https://data.cityofchicago.org/api/geospatial/5jrd-6zik?method=export&format=Shapefile

we also need the population for each census tract:

https://data.cityofchicago.org/api/views/5yjb-v3mj/rows.csv?accessType=DOWNLOAD

For Question 4, we need socioeconomic indicators data in Chicago

url: https://data.cityofchicago.org/api/views/kn9c-c2s2/rows.csv?accessType=DOWNLOAD

the data include income, employment, age, and education information of Chicago by each community.

Methodology

Preprocess:

Download crime data, census tract geo data, data of general type of crime and selected socioeconomic indicator data in Chicago. Unzip the census tract geo data.

Drop rows with missing data in the crime data.

Random pick 4 subsets of a thousand cases each as the set we can test our code on.

Get the useful information out from the crime data including:

ID, Date, Year, Block, Location, Description, Location, Police district, Census Tract, Community Area, IUCR, Primary type, Description, and Arrest

Combine the census tract geo data with crime data according to the census tract column. (inner join)

Question One:

Generate a csv file which has 3 columns. The first one is the crime type and the second one is the year of data and the third one is the corresponding sentence days in months of each type of crime.

The sentence information is found in the 12's page in the United States Sentencing Commission Statistical Information Packet State of Illinois

Create a new column in the crime data set which stores information of each case's more general type. The set of types in this column should be the same as the set of crime type in the data set of general type of crime.

Combine the 2 datasets according to crime type and year

Get the subset of data in 2017. Group the data by community area, and sum the sentence days of all cases to get the all sentences month. Compute the average sum of sentences of all the community around each community area and put the data into a new column. Then create a new column that sum the length of sentences in this district * 0.8 and the average length of sentences in the nearby communities*0.2.

The average sum of length of sentences indicates the safety rankings. The smaller the length of sentences, the safer the community is.

plot 2 images of chicago and show the safety level by color on the map.

Do the same processs for year period of 2001-2004, 2005-2008, 2009-2012, 2012-2016.

plot those four maps to show the changing in safety level during the past years.

Question Two:

Get a sub dataset of the cases in year 2001-2015 because the arrest situation might change in the recent years so only use data in that time range.

Get the hours when each case happened and add this data to a new column called Hour.

Group the data by hour, block, location type, and crime type, compute the rate of criminal being arrested in the given situation.

Use hour, block, location type, and crime type as feature to predict the rate of criminal being arrested with Sklearn regressor model. Split the data into 80% training data and 20% test data. Use training data to train the model and test set to test the model. The model gives the result which is the possibility of solving a case when given hour, block, location type, and crime type information. Provide a calculator which can compute how many cases the police still need to solve in 2019 by multiply the possibility given by the case and the number of cases then minus the number of cases they have already solved.

Question 3:

First, deal with data of population of each census block. Create a new column that contains the first 11 digits of the census block number of each census block called census tract. Group by the new column and sum all the populations. Join the data with crime data by the census tract column to add population information to the dataset.

Then compute the population in different wards. Compute crime rate which is case number divided by population for each year in different wards. Then compute the annually increasing/ decreasing rate of crime rate from 2001-2015. Compute the annually increasing / decreasing arresting rate for 2001-2015 too. Calculate the correlation between the 2 rates. Find the top 10 wards whose 2 rates has the greatest correlation and show the rates in a plot. Compute percentage of wards whose 2 rates has more than 0.5

Pearson product-moment correlation coefficient, which is considered to have strong relationship. This percentage is the result to the question. If it’s greater than 0.5, then we can make the conclusion that police performance can affect crime rate.

Question four:

Inner join the crime data sets with the selected socioeconomic indicator data sets on the attributes community area name. The filter out the table leaving only the attributes ID showing the crime, the column showing the population, the percentage of household below quality, the percentage of unemployed people aged over 16, the percentage of people aged over 25 without high school diploma and the percentage of people aged below 18 or above 24. Then group by the table using the attribute year and community area name to form a new table and count the number of crimes happened as well as the total population in any of the community area during a period of time. We can then use these two values to calculate the crime rate and create a new column. We then draw some maps of Chicago divided by community areas showing the poverty rate, unemployment rate, educational level, age distribution and crime rate respectively. We can then plot the graphs showing the relationships between crime rate and the other four factors given by the attributes listed above in order to figure out which factor has the strongest correlation, and which has the weakest correlation with the crime rate.

Useful libraries that may be used:

Pandas

Geopandas

Sci-kit learn

sea-born

matplotlib.pyplot

plotly

requests

Work Plan

We divided our work into four parts

The first part includes writing code to preprocess data together and each of us write codes for 2 questions separately. The code in this part should do all the computation works and roughly give a plot of the result. The plots in this part don't need to be very organized.

We use colab to write and run our code and will finish this part by 5.29. 15 hours for each person.

The second part involves summary the result we get from computation together and analyze the answers to our research questions. We will also debug and test codes for each other and figure out how to improve our code for example improve the efficacy of the code.

We will finish this part by 6.1. 9 hours each person

The third part involves modify our plot and write part 2. We will first modify the rough plots we get to more clear and readable plots. Then we will write our final report for part 2.

We expect to finish this part by 6.5. 9 hours each person

The last part is preparing for presentation. We need to make slides, organize the structure of our presentation, and rehearse for the final presentation.

We expect finish the majority part of presentation before final week comes by 6.8. 4 hours each person.