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?

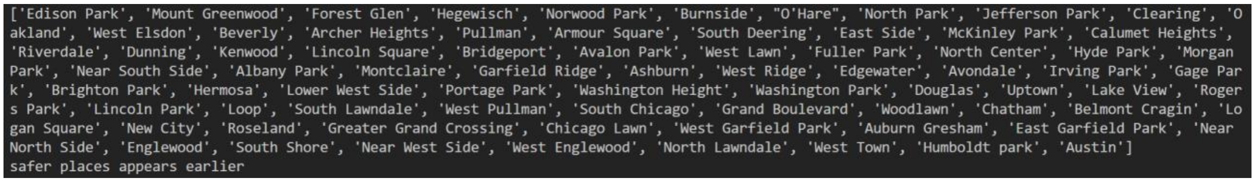
The safety is computed by the number of occurrence and weight of different types of crime.

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.

The result depends on the year you are asking.

For example, the ranking of year 2003 is:



```
['Edison Park', 'Mount Greenwood', 'Forest Glen', 'Hegewisch', 'Norwood Park', 'Burnside', "O'Hare", 'North Park', 'Jefferson Park', 'Clearing', 'O
akland', 'West Elsdon', 'Beverly', 'Archer Heights', 'Pullman', 'Armour Square', 'South Deering', 'East Side', 'McKinley Park', 'Calumet Heights',
'Riverdale', 'Dunning', 'Kenwood', 'Lincoln Square', 'Bridgeport', 'Avalon Park', 'West Lawn', 'Fuller Park', 'North Center', 'Hyde Park', 'Morgan
Park', 'Near South Side', 'Albany Park', 'Montclaire', 'Garfield Ridge', 'Ashburn', 'West Ridge', 'Edgewater', 'Avondale', 'Irving Park', 'Gage Par
k', 'Brighton Park', 'Hermosa', 'Lower West Side', 'Portage Park', 'Washington Height', 'Washington Park', 'Douglas', 'Uptown', 'Lake View', 'Roger
s Park', 'Lincoln Park', 'Loop', 'South Lawndale', 'West Pullman', 'South Chicago', 'Grand Boulevard', 'Woodlawn', 'Chatham', 'Belmont Cragin', 'Lo
gan Square', 'New City', 'Roseland', 'Greater Grand Crossing', 'Chicago Lawn', 'West Garfield Park', 'Auburn Gresham', 'East Garfield Park', 'Near
North Side', 'Englewood', 'South Shore', 'Near West Side', 'West Englewood', 'North Lawndale', 'West Town', 'Humboldt park', 'Austin']
safer places appears earlier
```

Safer places appears earlier in the ranking.

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.

The result depends on the conditions you want to check.

For example, the rate of solving the case predicted with community number 45, crime type of theft, and 5pm-6pm is about 18.7 percent.

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.

The results depend on the community area that you choose. There is about 38 percentage of all community area where the crime rate and arrest rate have a strong correlation with each other. Most of the community areas have positive correlation between crime rate and arrest rate, but there are still community areas that have negative correlation between crime rate and arrest rate.

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.

The result depends on the particular community area and the specific time period that you choose. We choose the year 2018 to further investigate the relationships between crime rate and possible factors. From the graphs plotted, we can see that all the factors have some positive or negative influences on the crime rate in a particular city during a specific period of time, with the unemployment rate being the most contributing factor and the age distribution being the least contributing factor.

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.

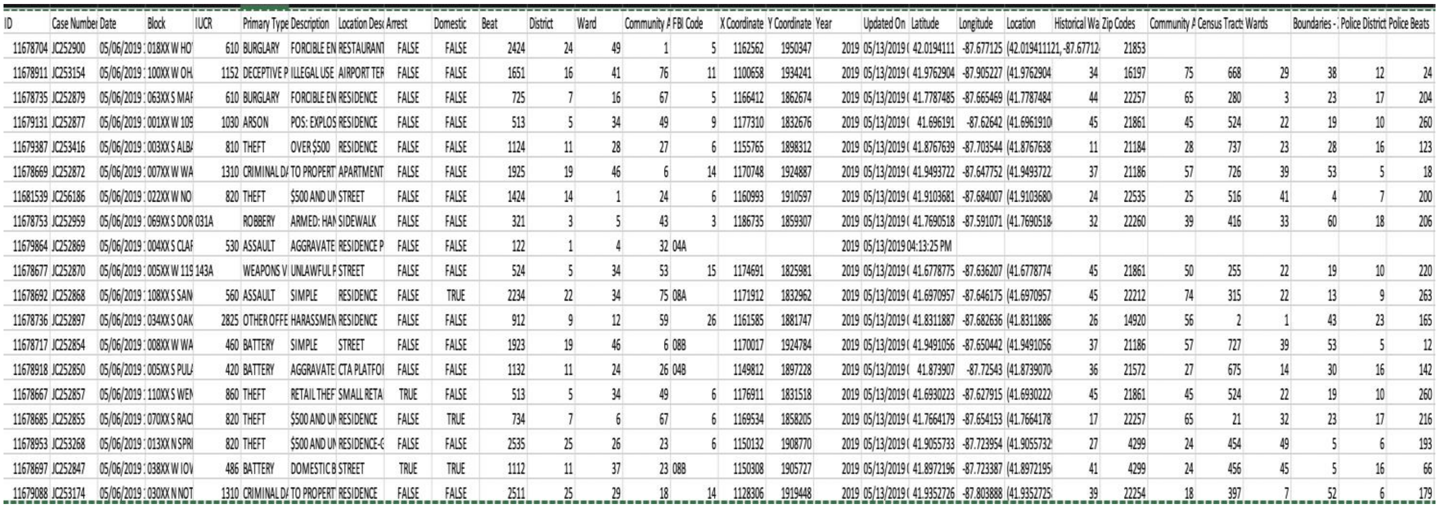
Data set:

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 Numbe	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 Tract	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 ALB	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 PUL	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-C	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.

We also create a csv file of average sentence length of each

Methodology

Preprocess:

Download crime data, census tract geo data, sentence length 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 2 columns. The first one is the crime type and 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

Combine crime data and sentence length datasets according to crime type

Get the subset of data in specific year. 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 image of Chicago and show the safety level by color on the map.

Plot the changing trend of safety level of several areas through years

Get the safety ranking of a specific year

Question Two:

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, community area 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, community area and crime type, and crime type information.

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 community area. Compute crime rate which is case number divided by population for each year in different community area. Then compute and estimate 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 community area 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. Then 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

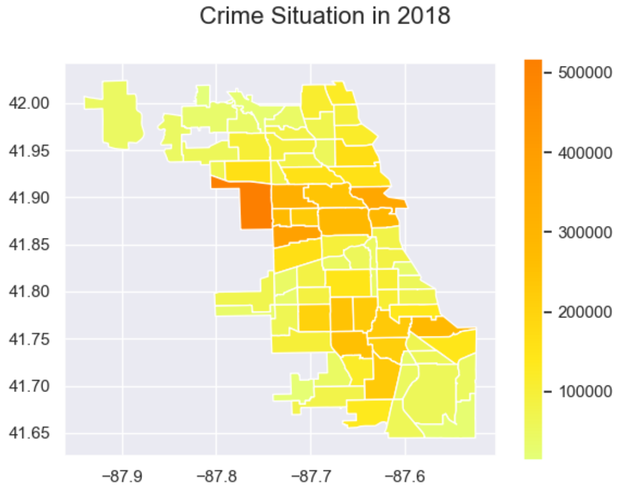
Results:

Question one:

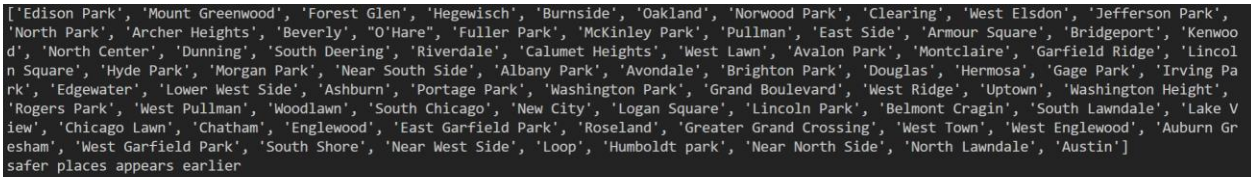
We provide a function that can show the safety ranking of a year you input.

For example, in the year of 2018, here is a plot which shows the overall situation of criminal.

The darker color implies higher crime harm in that community area. The plot shows a big difference in the largest harm and smallest harm.

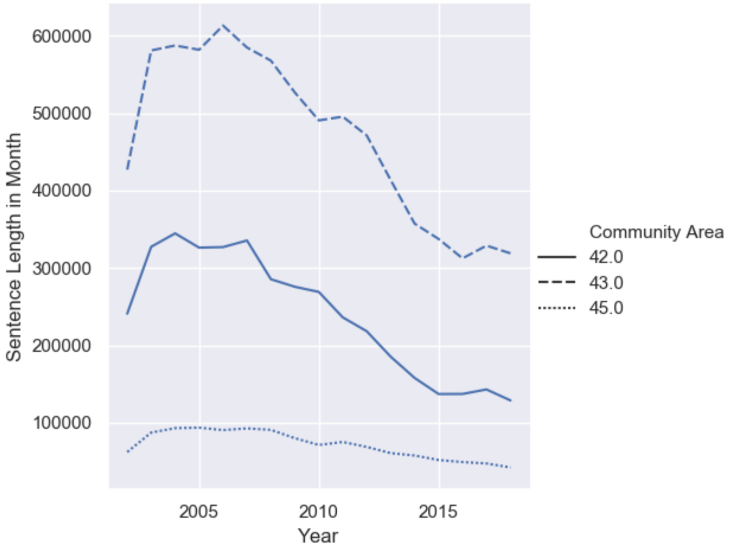


Below is the ranking of safety we get in 2018:



```
['Edison Park', 'Mount Greenwood', 'Forest Glen', 'Hegewisch', 'Burnside', 'Oakland', 'Norwood Park', 'Clearing', 'West Elsdon', 'Jefferson Park',
'North Park', 'Archer Heights', 'Beverly', "O'Hare", 'Fuller Park', 'McKinley Park', 'Pullman', 'East Side', 'Armour Square', 'Bridgeport', 'Kenwoo
d', 'North Center', 'Dunning', 'South Deering', 'Riverdale', 'Calumet Heights', 'West Lawn', 'Avalon Park', 'Montclaire', 'Garfield Ridge', 'Lincol
n Square', 'Hyde Park', 'Morgan Park', 'Near South Side', 'Albany Park', 'Avondale', 'Brighton Park', 'Douglas', 'Hermosa', 'Gage Park', 'Irving Pa
rk', 'Edgewater', 'Lower West Side', 'Ashburn', 'Portage Park', 'Washington Park', 'Grand Boulevard', 'West Ridge', 'Uptown', 'Washington Height',
'Rogers Park', 'West Pullman', 'Woodlawn', 'South Chicago', 'New City', 'Logan Square', 'Lincoln Park', 'Belmont Cragin', 'South Lawndale', 'Lake V
iew', 'Chicago Lawn', 'Chatham', 'Englewood', 'East Garfield Park', 'Roseland', 'Greater Grand Crossing', 'West Town', 'West Englewood', 'Auburn Gr
esham', 'West Garfield Park', 'South Shore', 'Near West Side', 'Loop', 'Humboldt park', 'Near North Side', 'North Lawndale', 'Austin']
safer places appears earlier
```

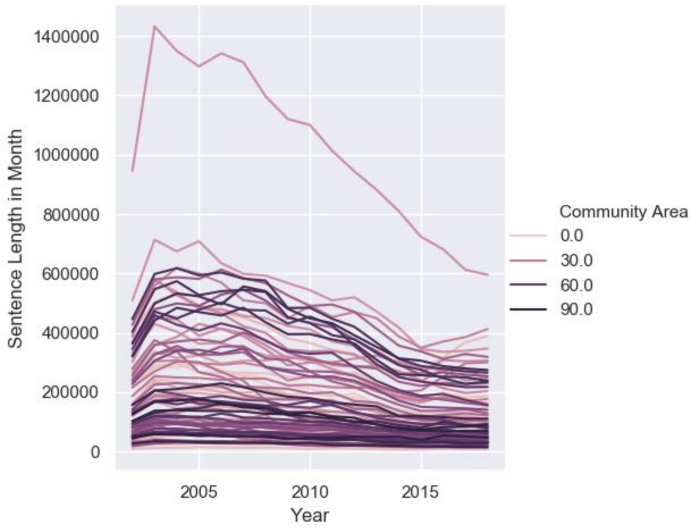
The interesting part of the question is that we also get plot to see the changing trend of harm through years, for example:



The interesting things is that although the harm we compute very from different areas, the trends are very similar. The harm first rises and riches its max after 2005, and then it constantly decreases until

2015 when it stays relatively steady. It strongly implies that the public safety situation is very much improved through 2005-2015.

Below is a plot showing curves of all communities to prove that:

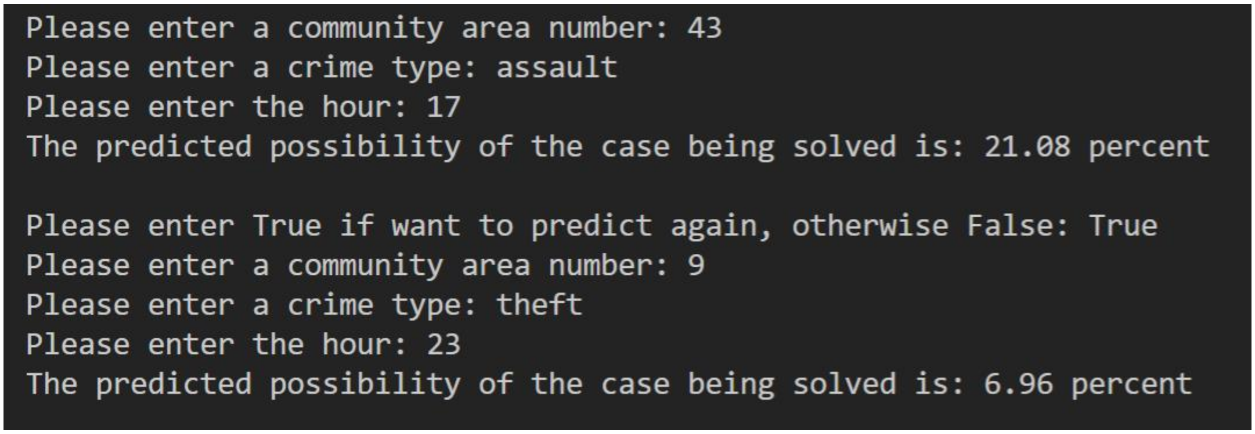


This is more obvious for those districts which have relatively higher crime rate.

A thing to mention here is that we are not computing sentence length on a very accurate level. We only use it to infer the harm of each time of crime. So, it's not a project on sentence length that kind of information can't be used to study sentence situation.

Question two:

The result of this question is a function we provide. The function behaves like this:

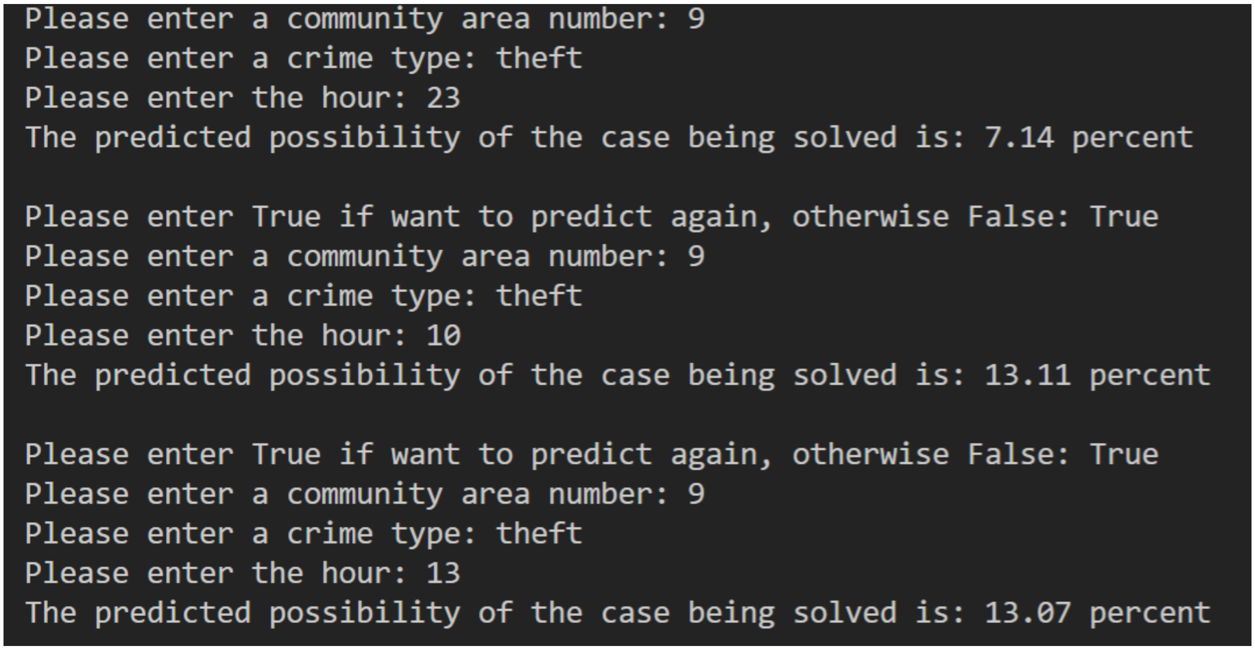


```
Please enter a community area number: 43
Please enter a crime type: assault
Please enter the hour: 17
The predicted possibility of the case being solved is: 21.08 percent

Please enter True if want to predict again, otherwise False: True
Please enter a community area number: 9
Please enter a crime type: theft
Please enter the hour: 23
The predicted possibility of the case being solved is: 6.96 percent
```

The function will ask for 3 inputs, which are community area number, a type of crime, and an hour in a day. It will then predict the possibility and print it out.

You can ask for result in several different conditions with a single model.

It's convenient to compare some the result. For example:

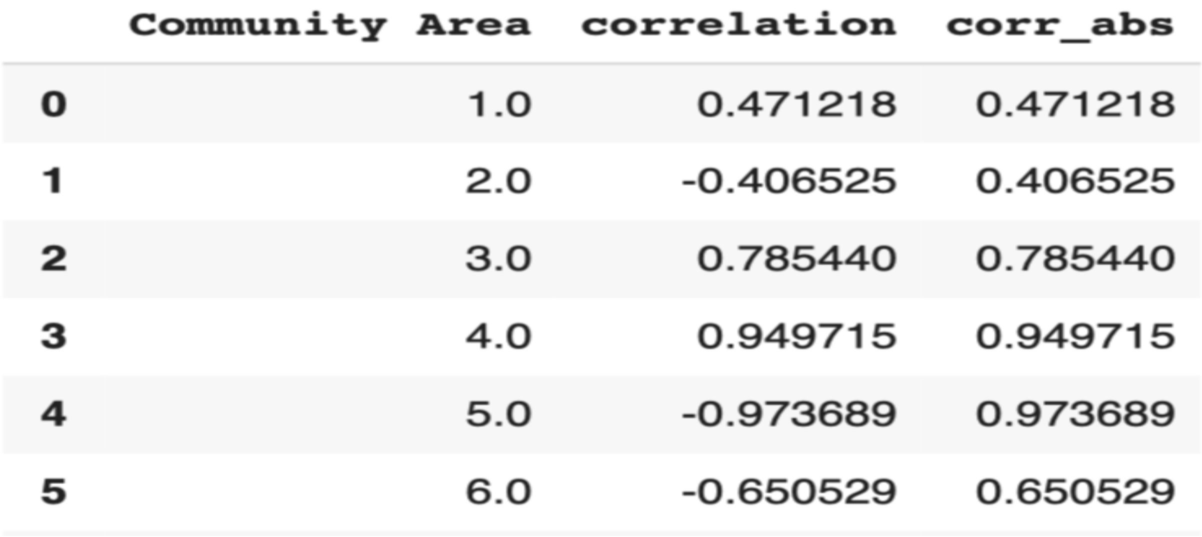


```
Please enter a community area number: 9
Please enter a crime type: theft
Please enter the hour: 23
The predicted possibility of the case being solved is: 7.14 percent

Please enter True if want to predict again, otherwise False: True
Please enter a community area number: 9
Please enter a crime type: theft
Please enter the hour: 10
The predicted possibility of the case being solved is: 13.11 percent

Please enter True if want to predict again, otherwise False: True
Please enter a community area number: 9
Please enter a crime type: theft
Please enter the hour: 13
The predicted possibility of the case being solved is: 13.07 percent
```

We can imply that the predicted rate is a lot smaller when theft at community 9 at night than at noon.

However, the machine learning model has randomness in it so this is a draw back of using it.

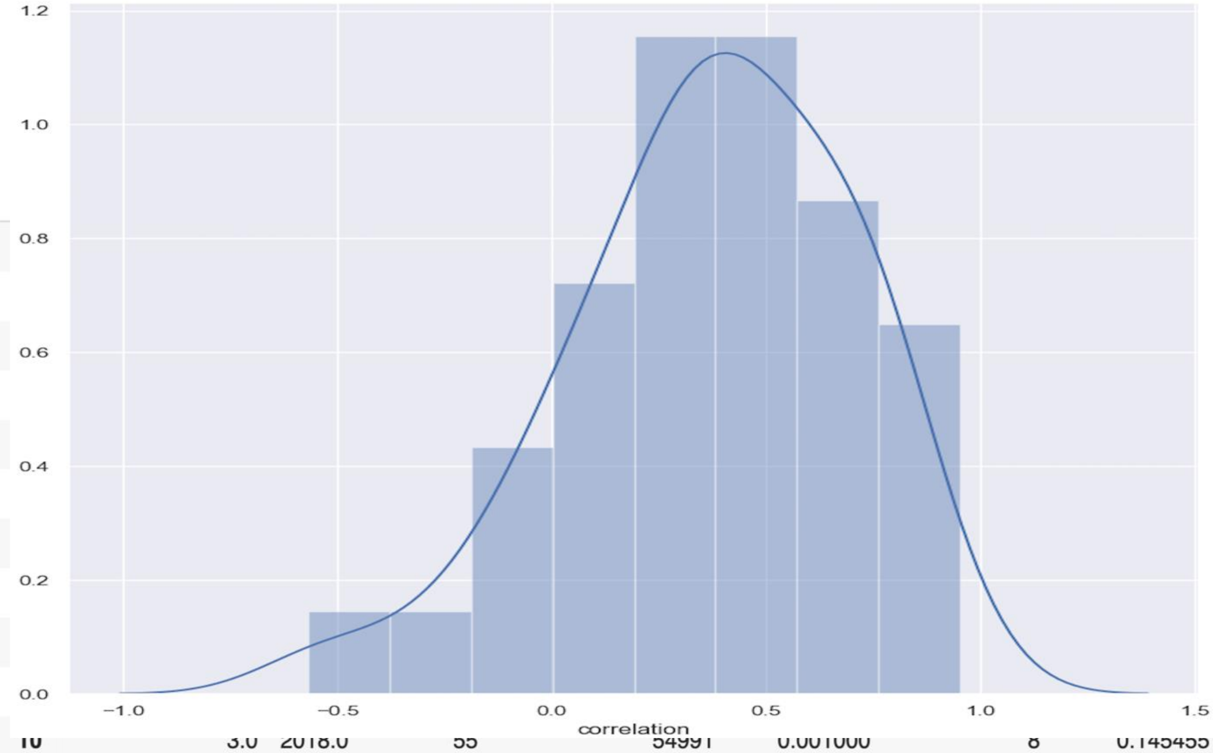
Question three:



	Community Area	correlation	corr_abs
0	1.0	0.471218	0.471218
1	2.0	-0.406525	0.406525
2	3.0	0.785440	0.785440
3	4.0	0.949715	0.949715
4	5.0	-0.973689	0.973689
5	6.0	-0.650529	0.650529

One of the research results for this question is the table(the one above) showing the correlation between the arrest rate and crime rate in a particular community Area as well as its magnitude, which can be derived from the table(the one below) that shows the crime rate and arrest rate for a specific community area in each year. What’s interesting is that some of the correlations are actually negative shown in the above table. Normally there should be a positive correlation between arrest rate and crime rate since increasing in the cases solved tends to be effective at frightening the potential criminals and keep them from committing crimes. However, in some community area this might not be true maybe due to the fact that some criminals that have been arrested are released immediately or very soon, so the cost of committing a crime is still pretty low.

Another important research result is the probability distribution of the correlation between arrest rate and crime rate(the graph is shown below). We have calculated the percentage of community area whose crime and arrest rates have a correlation of more than 0.5, which is around 0.38 round to two decimal places. That means that there is not enough evidence to show that the police performance is a primary factor to the change of crime rate, though it still exerts some impacts on the criminal behaviors.

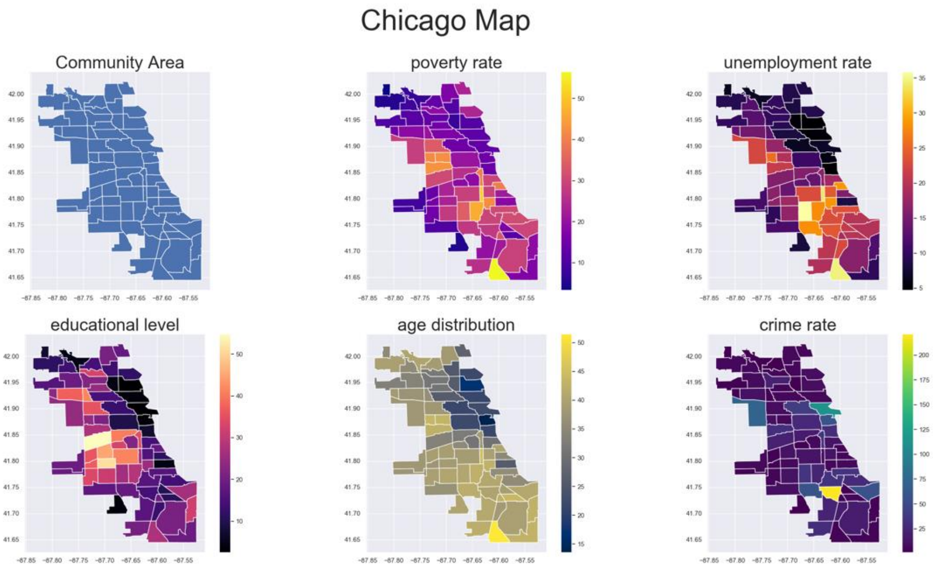


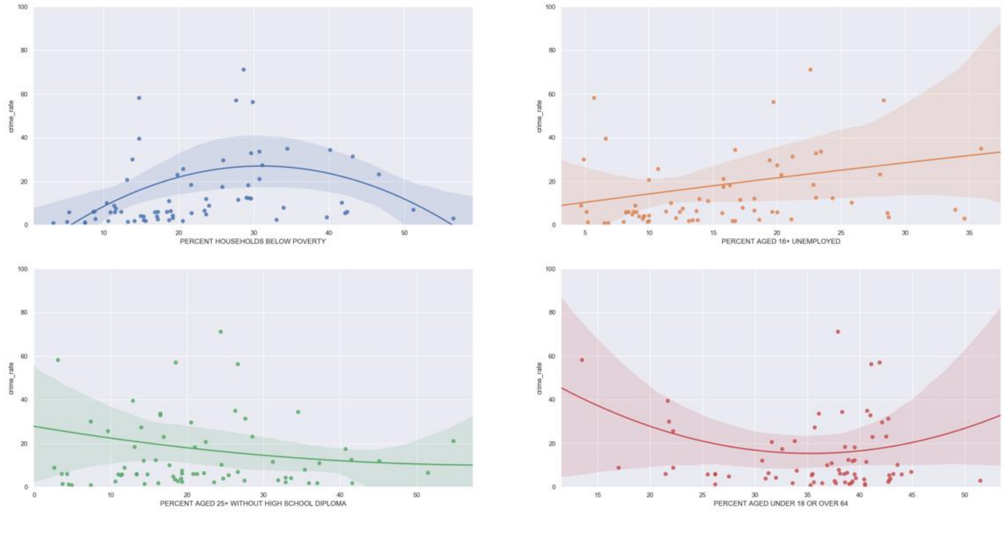
10	3.0	2018.0	55	54991	0.001000	8	0.145455

Question four:

The research results for this question are that the crime rate has strong correlation with any of the socio-economic data, which can be seen from the plotting of Chicago map on different standards and the regression plot of the crime rate with respect to socio-economic factors (both graphs are shown below). From the graphs, we can see that as the poverty rate increases, the crime rate gradually increases up until a point and then falls back. One of the possible reasons to explain this phenomenon is that taking a particular type of crime theft as an example, as the poverty continues to increase, more and more people are willing to become thefts and commit crime for survival. But then as more and more people’s wealth disappears, thefts may find it difficult to search for wealthy families and steal from them so the crime rate decreases. The crime rate has a positive relationship with the unemployment rate but a negative relationship with the education level. The reason for the former relationship is relatively easy since higher unemployment means more poverty and this can be furthered explained using the same reason mentioned above. The latter one is maybe because as more people become less educated, they are unlikely to be successful in committing the crime. The relationship between crime rate and age distribution can be a bit tricky. It may be because of the fact that children and elders are more vulnerable for crimes as well as the fact that adults are the majority of people who commit crimes, so the curve goes down at first and then rises up again.

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.





Reproducing your results:

Unzip project.zip. download a csv file from the url below:

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

put it into the same folder with other files

Run the main.py file in terminal. It should handle the process automatically. Use cse163 conda environment.

We also have an interaction main called interactive.py that can allow user interaction for question 1 and 2.

Question one :

If run main.py, it will automatically print the results.

The plots in the report are named: change_through_years_all_area.png, change_through_years.png, safety_ranking_2018.png. The ranking will be printed out

If run the alternate interaction version.

For question one, it will first ask for a year you want to plot. Enter a year between 2002-2018

If want to plot another one, Enter True then. Else enter false. The plot will be saved in the directory named safety_ranking_year.png. (year is the year you input)

It will then ask for community numbers. Enter several communities for example 3 with space between them for example: 43 44 45. The numbers should be between 1-77.

the plot will be saved to a file named change_through_years.png. It shows how the harm in these communities change through years.

After this, it will ask for a year again. Enter a year between 2002-2018 and it will print out the safety ranking.

Question Two:

If run main.py, it will automatically print the results.

If run the alternate interaction version.

For question two. It will ask for a community number (from 1-77), a criminal type(the types are in a file called criminal_types.txt we upload) case insensitive and an hour in a day(military time).

It will then print out the predicted possibility.

It will also ask for if you want to predict another condition too. Enter True if you want and False otherwise.

It will also print out the mean square error of the model it uses.

Question Three:

For this question, just run the main.py file in terminal with Question3 statement uncommented in the main method, it will automatically print out all the wanted data and tables in the console, which includes crime and arrest rates table, table regarding correlations of arrest and crime rates on different Community Area, table regarding top 10 correlations of arrest and crime rates on different Community Area and percentage of Community Area whose crime and arrest rates have a correlation of more than 0.5. It will also produce two files to the same directory, one named rate_change_with_year.png and the other named correlation.png. The first file shows the changes in arrest and crime rate with years in a particular community area, which in default is 1.0. The second file shows the distribution graph which is the probability distribution of the correlation between arrest rate and crime rate.

Question Four:

For this question, just run the main.py file in terminal with Question4s statement uncommented in the main method. It will produce two files in the same directory, one named geo_plot.png and the other named regress_plot.png. The first file plots the maps of Chicago divided by community areas showing the poverty rate, unemployment rate, educational level, age distribution and crime rate respectively. The second file plots the graphs showing the relationships between crime rate and the other four factors, poverty rate, unemployment rate, educational level, and age distribution.

Testing

Our code contains the process to generate smaller data size and save it to csv file to test and make our code fast. (named samplen.csv, n = 1, ,2, 3, 4)

We mainly test our code on colab.

The datasets we are using are provided by the government and are valid. We manually checked several points to see if the community area information is correct. We also sum up the population in the population csv and the sum is approximately the population in Chicago.

In problem I, when getting the areas around every community, We manually checked the information with https://en.wikipedia.org/wiki/Community_areas_in_Chicago

The _total_frame in question 1 also behaves correctly. set(self._total_frame['Year']) is from 2002-2018

In problem 2, We got 5 rows to see if we get the arrest count right.

assert_equals(1, sample1.head. groupby(['Community Area', 'Primary Type', 'Hour']).agg({'Arrest': ['count', arrest_count]}).reset_index())['Arrest']['arrest_count']

This is the main thing we compute in the question.

Question3:

# test the total population of Chicago in geo_pop_data table

assert_equals(2695598, geo_pop_data['TOTAL POPULATION'].sum())

```
# test the total number of unique community area in crime_arrest_rate table

assert_equals(77, len(crime_arrest_rate.groupby('Community Area')))

# test the crime rate calculated using crime_arrest_rate table

assert_equals(0.012274803388336728, crime_arrest_rate['case_num'].sum() /
crime_arrest_rate['TOTAL POPULATION'].unique().sum())

# test the arrest rate calculated using crime_arrest_rate table

assert_equals(0.2011, crime_arrest_rate['case_solved'].sum() / crime_arrest_rate['case_num'].sum())

# test the number of unique community area using the correlation column in corr_ca table

assert_equals(74, corr_ca['correlation'].count())

# test the number of unique years listed in the crime_arrest_rate table

assert_equals(4, len(crime_arrest_rate['Year'].unique()))
```

Question4:

```
# test the number of unique community area in crime_socioecon_data table

assert_equals(77, len(crime_socioecon_data['Community Area'].unique()))

# test the number of unique community area using the crime_rate column in crime_socioecon_data
table

assert_equals(76, len(crime_socioecon_data['crime_rate'].unique()))
```

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.

The time estimate for the first part of the work is accurate. However, for the second and third part of the work plan the time estimate is far from accurate. Our estimates were not very good since both of us waste some time programming in the wrong direction, which can make the code hard to debug and we also have to rewrite the code sometimes. All of the above consume time.

Live Presentation or Video

We will record a video.

Collaboration

No outsiders have helped us finish this assignment.