CS 398 Project Report

by Team Newbee
Team Member: yurenx2; zq2; penggu2; zehaoc2

Abstract

In this project, we want to explore the past crime situations in the city of Chicago and the objective factors that have relationships with the occurrences of crime cases in Chicago. We analyzed data that is relative to Chicago crimes in recent years. For the objective factors, we use the economic factor (GDP), temperatures and weather as our support data. We compared crime dataset with corresponding weather by time to verify whether there is a relationship between them. In this report, we would provide analysis over time, weather and location, and relative visualization.

To have a more readable coding style, we separated data processing and visualization into two files, and to let those two files interact with each other, we implemented OOP (Object Oriented Programming) inside of spark framework. Therefore, after we used spark for data processing, we can use the data for visualization immediately. Such reproductive implementation helps us to debug and make our process more reasonable. Also, we don't have to store additional files to store processed data as a transfer station of our dataset. What's more, our project could be a model, since we design such ports to get the input dataset and information. We can easily change the input dataset of other city crime data and easily get the report of other crime situations of such city.

To summarize this report, we would first introduce the dataset we use. Then, we would walk through the data processing component to briefly talk about various ways that we did with the provided dataset. After data processing, we will analyze graphs we obtained and additional comments relative to each topic. Then, we will introduce some challenges that we faced in this project. In the end, we would point out several future plans that may carry our project further.

Dataset and other Data Resource

We use *Crime in Chicago* dataset in Kaggle as our main database. It collects all criminal cases in Chicago from January 2001 to January 2017. The attributes of this table contain the date ('Date' and 'Year'), primary type ('Primary type') and locations ('Latitude', 'Longitude', and 'Community Area') and much other information about the criminal cases.

Based on analysis of the criminal cases with time, we find that the criminal cases came increasing badly since 2008. Then, the first term came to our mind was "financial crisis". Hence, we planned to find the relationship between the number of criminal cases and economy. We get the data from https://www.statista.com/. However, the data of the Gross domestic product (GDP) is a couple of numbers, so we hardcode them into our files.

We also use Weather dataset from National Center for Environmental Information as a support to help our analysis. To get the relationship between the criminal cases and the weather, we need a Chicago

weather database. We join two tables with the dates so that we can get the number of criminal cases with different weather type. With this data, we can do further analysis.

Data Processing

There are millions of tuples in our dataset, and the overall size of the file of the dataset is about 2.5 Gb, so local data processing is inapplicable on our personal computers. Therefore, we need to use cloud computing and AWS to help process our data. In this project, we basically use pyspark.sql model to prefetch our data; then we use pyspark.rdd model to process the data.

We upload the dataset to hdfs so that our cluster can access and help process it. We use pyspark.sql to get the total number, date, location and primary type of each criminal case. However, since there are various of types of crime. If there are lots of light crime cases, the analysis will be ambiguous. Hence, to make the analysis more specific and persuasive we artificially select the serious types of crime based on the primary type.
(SERIOUS = ['OFFENSE INVOLVING CHILDREN','PUBLIC PEACE VIOLATION','ARSON','CRIMINAL TRESPASS','ASSAULT','ROBBERY','HOMICIDE','CRIM SEXUAL ASSAULT','HUMAN TRAFFICKING','INTIMIDATION','CRIMINAL DAMAGE','KIDNAPPING','BURGLARY','WEAPONS VIOLATION'])

We download the weather data since 2001 to 2016 from National Center for Environmental Information. It contains the weather data collected from hundreds of stations. Hence, we use similar tool and technique of cloud computing to get the weather and temperature data.

Data processing on the cluster was slow. It usually costs about 3 to 5 minutes each time to get the result, if we don't have to wait for the cluster. However, one time, in order to get the all the locations of crime cases grouped by date, it spent 15 minutes to finish the job.

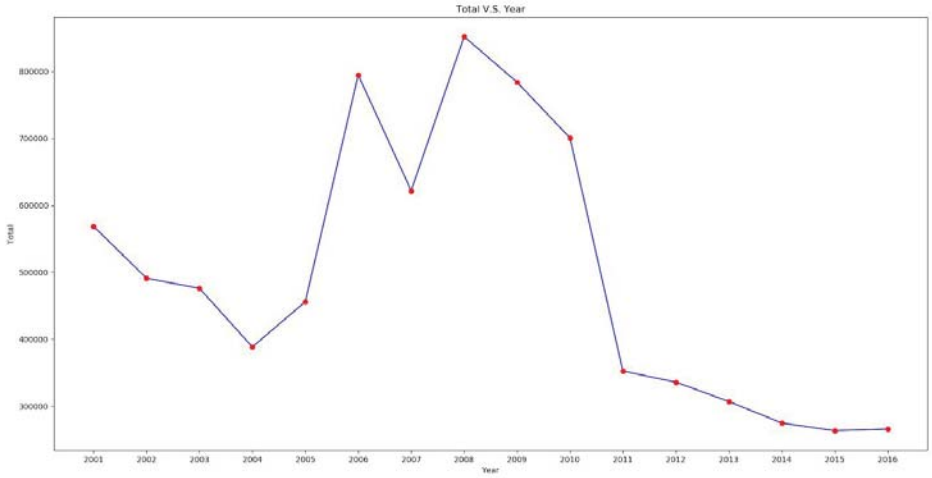
Data Analysis and Results

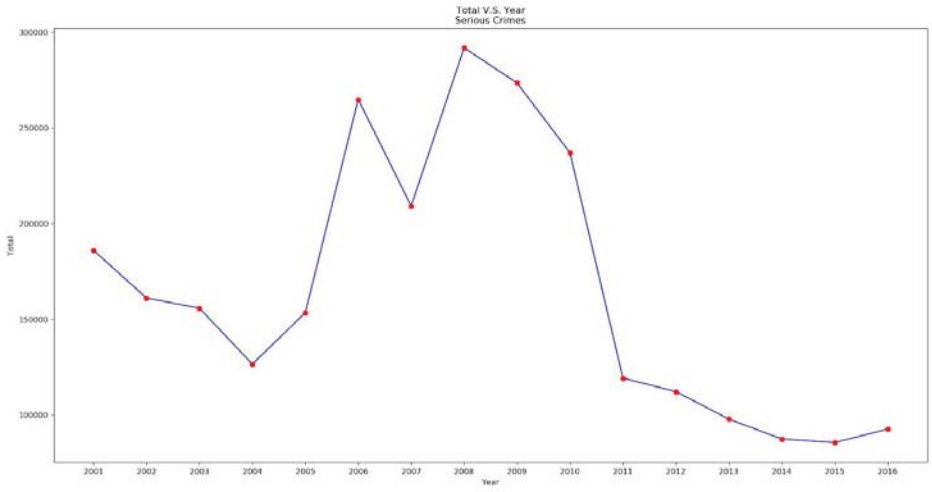
In general, we mainly used folium, matplotlib, scipy, and selenium for our data visualization. And following are specified analysis from different perspective.

Analysis by Time [dv_report.py, EvaluationByTime.py]

Every crime has a time label, and it seems reasonable to categorize crimes by time. Since time is a general term, we would mainly analyze crimes by year, by season, and by the hour. For each topic, there would be two graphs with the different y-axis, top image would be the general crime count and bottom one is the serious crime count. Since there is no major visual difference between them, we will

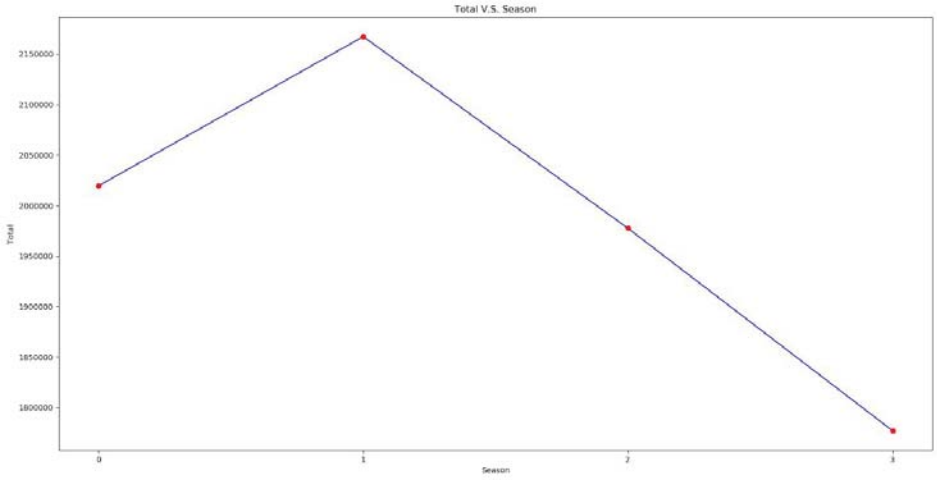
not analyze them separately. As a trend from general to specific, we will first look at the graph of "Year V.S. Total".

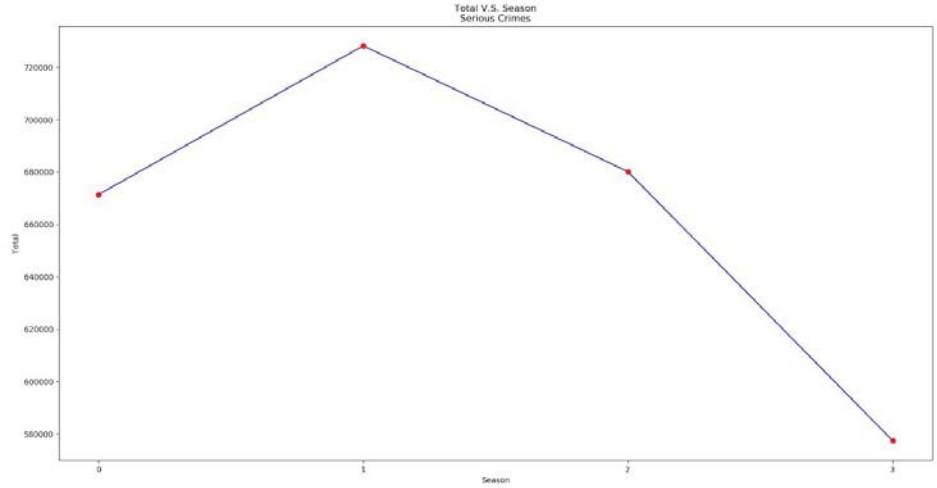




Above is the graph of crime counts by year, the crime number reaches its peak around 2006 - 2010. As the graph shown, the number of crime tend to decreases in recent years. Since 2017 only contains crime counts in January, the count is negligibly small.

Season matters because of special events and temperatures. For example, students may have spring break, summer break, winter break, but these breaks don't have equal length. However, since each school or company doesn't have identical calendars, we would analyze the dataset by season.

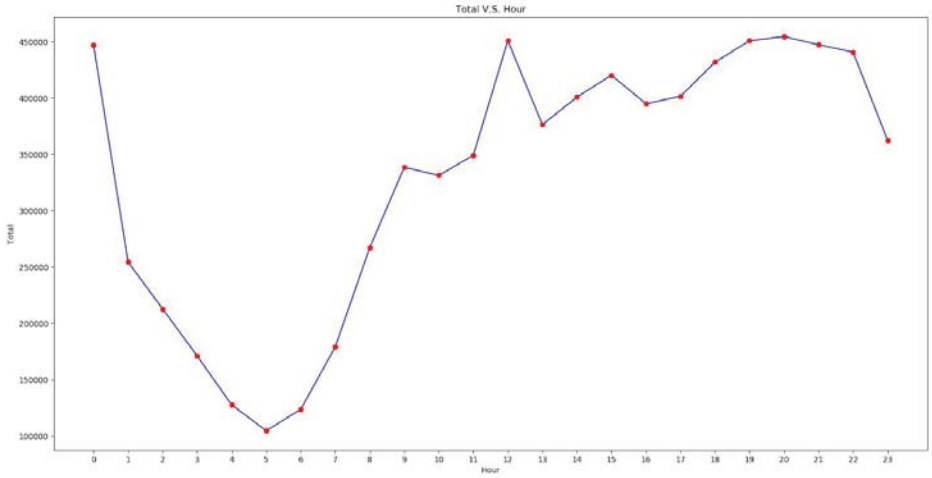


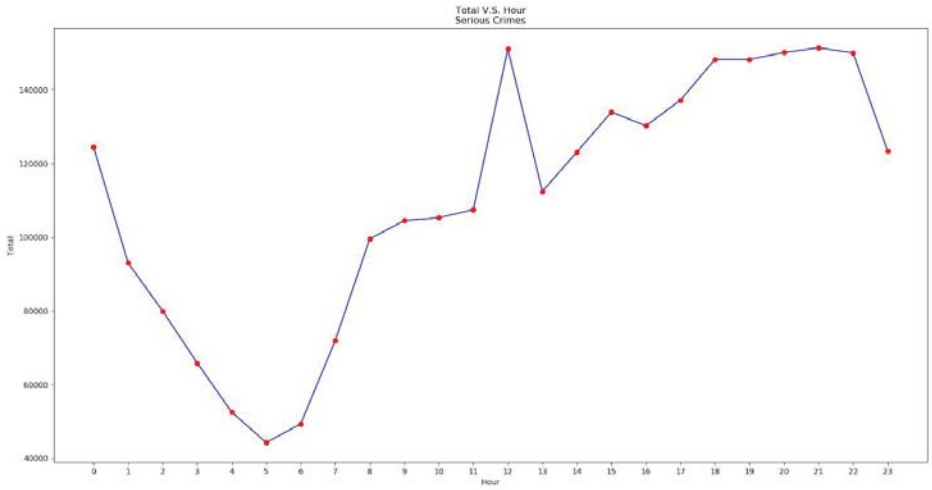


(0 is Spring, 1 is Summer, 2 is Fall, 3 is Winter)

Above is the graph of crime counts by season, the majority crime number happens in summer, and winter has the lowest crime number. There may be correlations between crimes and temperature, which leads to such diagram, we will further discuss it in the following analysis. But in general, winter is a safe time to visit Chicago.

Chicago's crime occurs daily. Therefore, an analysis over hours would provide a visual idea of Chicago crimes.

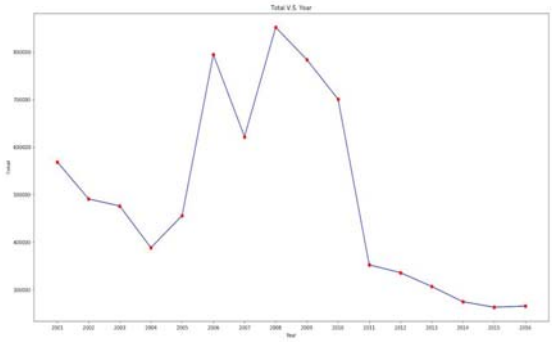


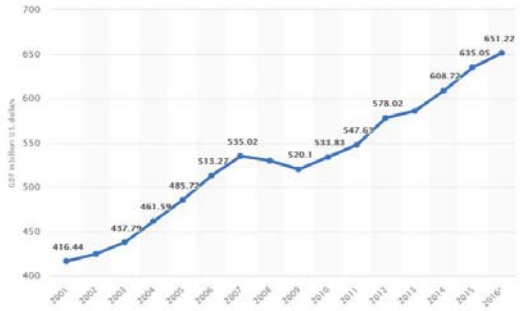


As we observe, the crime counts spread evenly from 10:00 - 24:00. However, the crime counts decrease dramatically at midnight. Thus, this graph suggests there is no real difference from the perspective by hours.

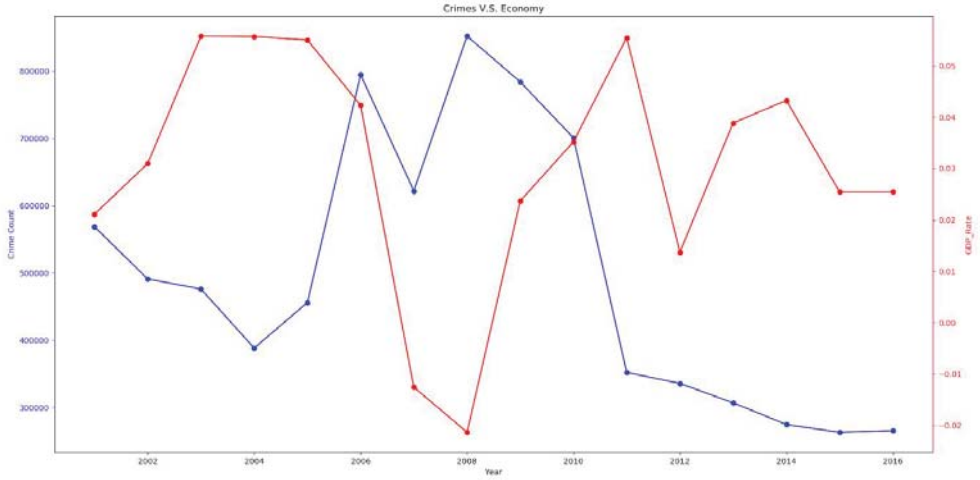
In all three analyses of the number of crimes cases and time, the diagrams of overall crime and serious crime don’t have an obvious difference. Both have a similar trend as time changes, which means that serious crime also follows that pattern of normal crimes in terms of time.

Analysis of Crime and Economy (GDP) [dv_report.py, EvaluationByEcon.py]



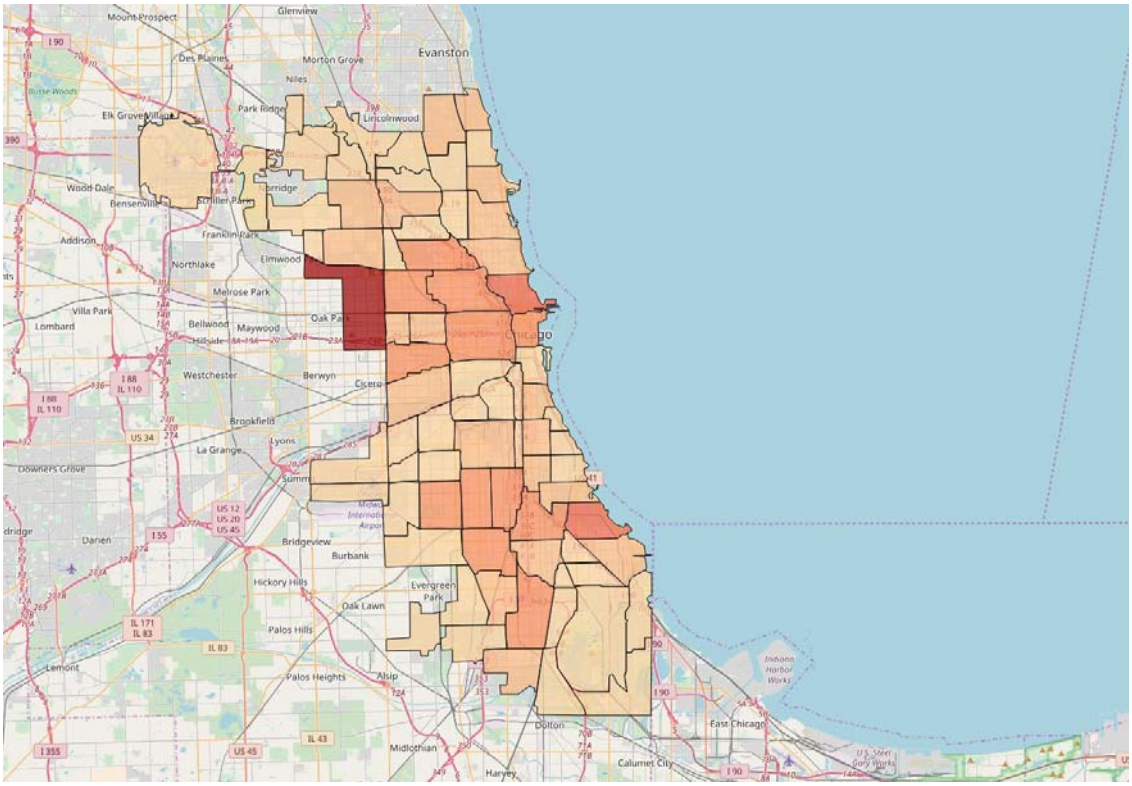


When we get the result of “number of criminal cases over years”, we find that the total number of criminal cases reach the peak in 2008. Then, we think the occurrence of criminal case may have a strong relationship between economy. Before the analysis, we make a hypothesis that the number of crime case is inversely proportional to the growth rate of GDP. Therefore, we found the data of GDP of Chicago metro area since 2001 to 2016 and plot the data through matplotlib. Furthermore, in order to see the relation clear, we put two plots of data together in one diagram. Based on this diagram, the result seems to prove our hypothesis.



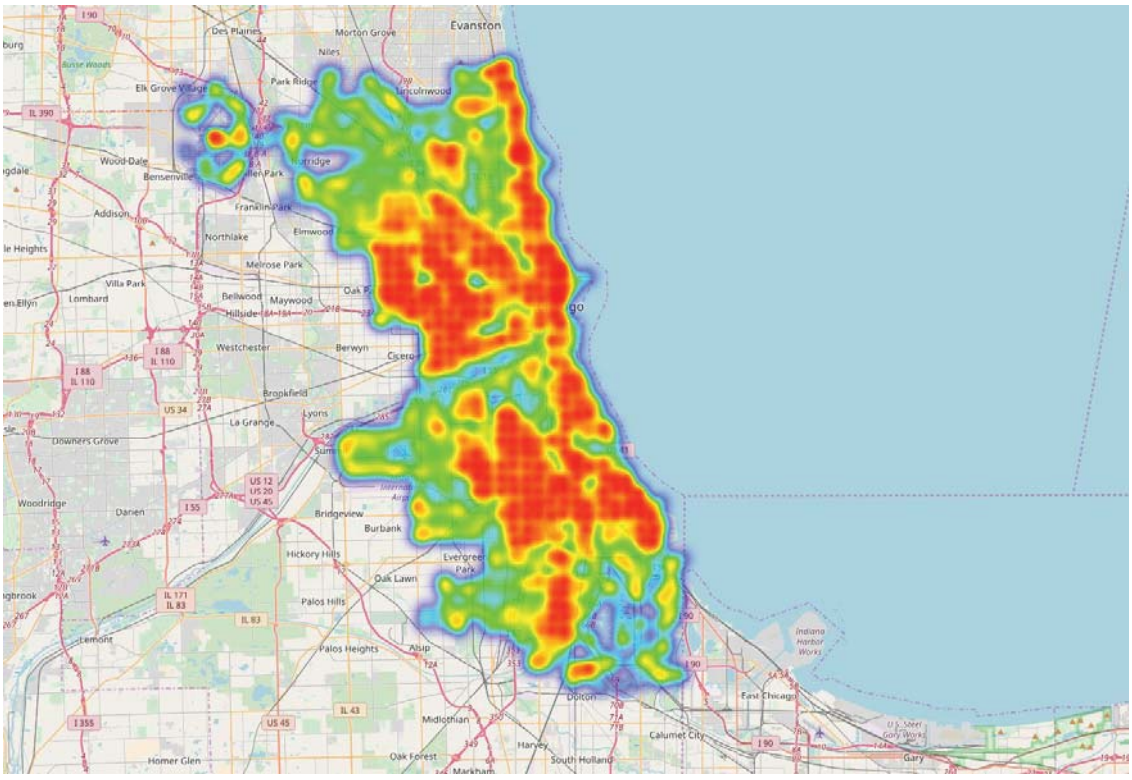
Analysis of Crime and Region
[heatmap*, visualRegion1.py, createSS.py]

We first made a Choropleth map on the number of crime incidents in different community areas in Chicago. This figure shows the crime between 2001 and 2017. To accomplish this, we first use Pyspark to aggregate and count the number of crimes by community area. Then we used folium package to create visualizations about the number of crimes in different regions. The darker the color is in the region, the more crimes took place there.



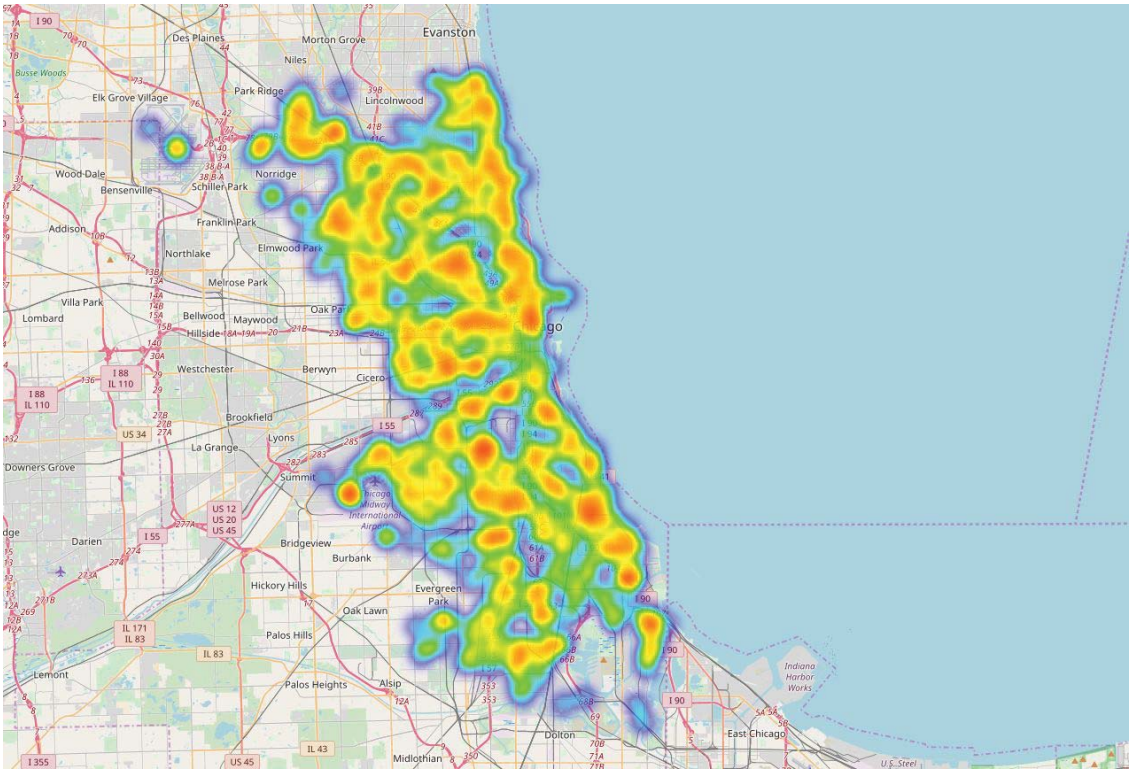
Then we create various heatmaps to compare the number and intensity of crimes occurred in different conditions.

We first created an overall heatmap to show the general distribution of crimes. It is obvious that the result of the overall heat map corresponds to the Choropleth map.



Then, we analyzed the serious crimes in Chicago.

A heat map visualization looks like this:



We next try to compare the difference between daytime and night. We manually define “Daytime” to be between 6 am to 6 pm and “night” to be 6 pm to 6 am next day. We first filtered data and get 2 lists of coordinates of crimes: daytime list and nighttime list. Then we pass those coordinates to folium and create the two visualizations. The contrast is quite easy to spot. The crime incidents that happen at night is much higher than in the daytime, a fact that complies with our common sense.

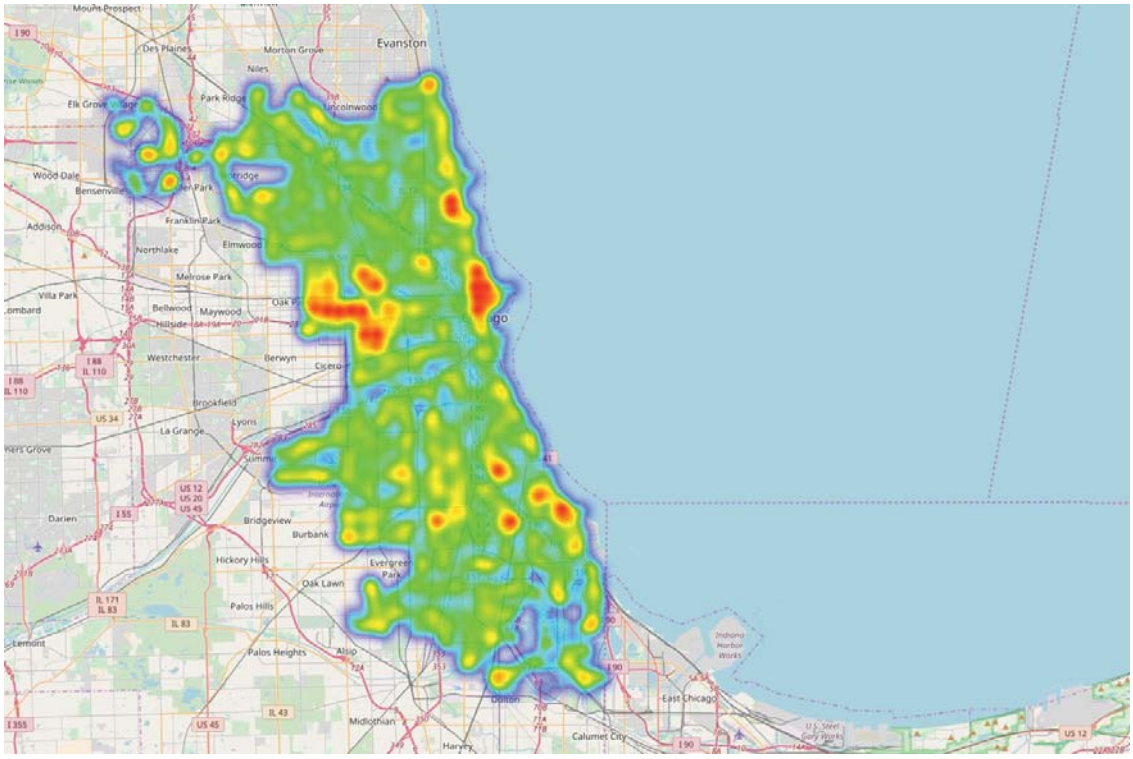


Fig.day

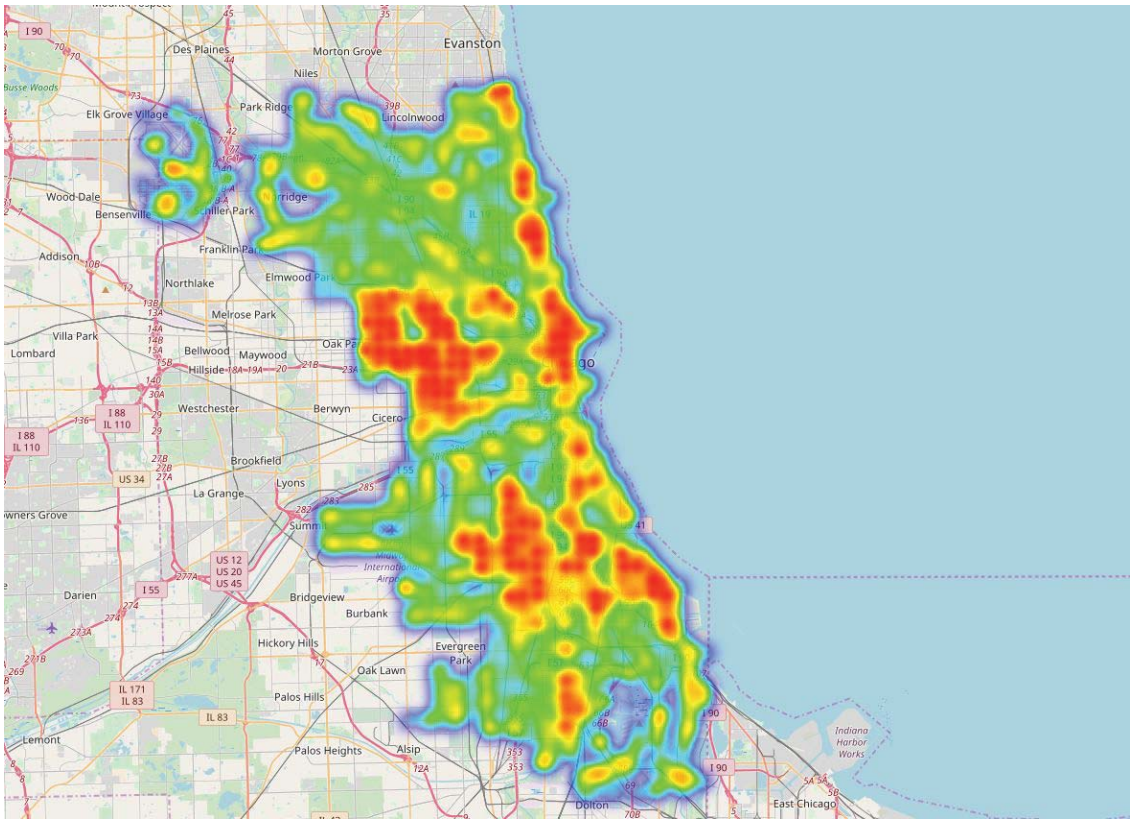
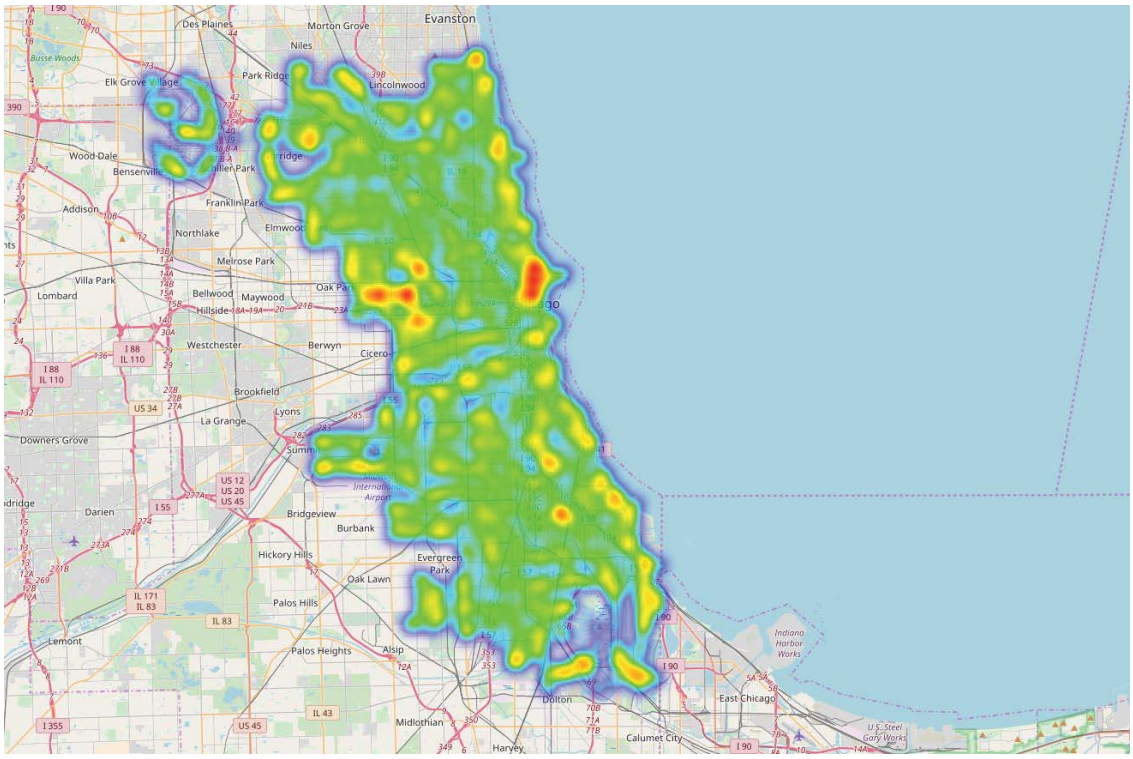
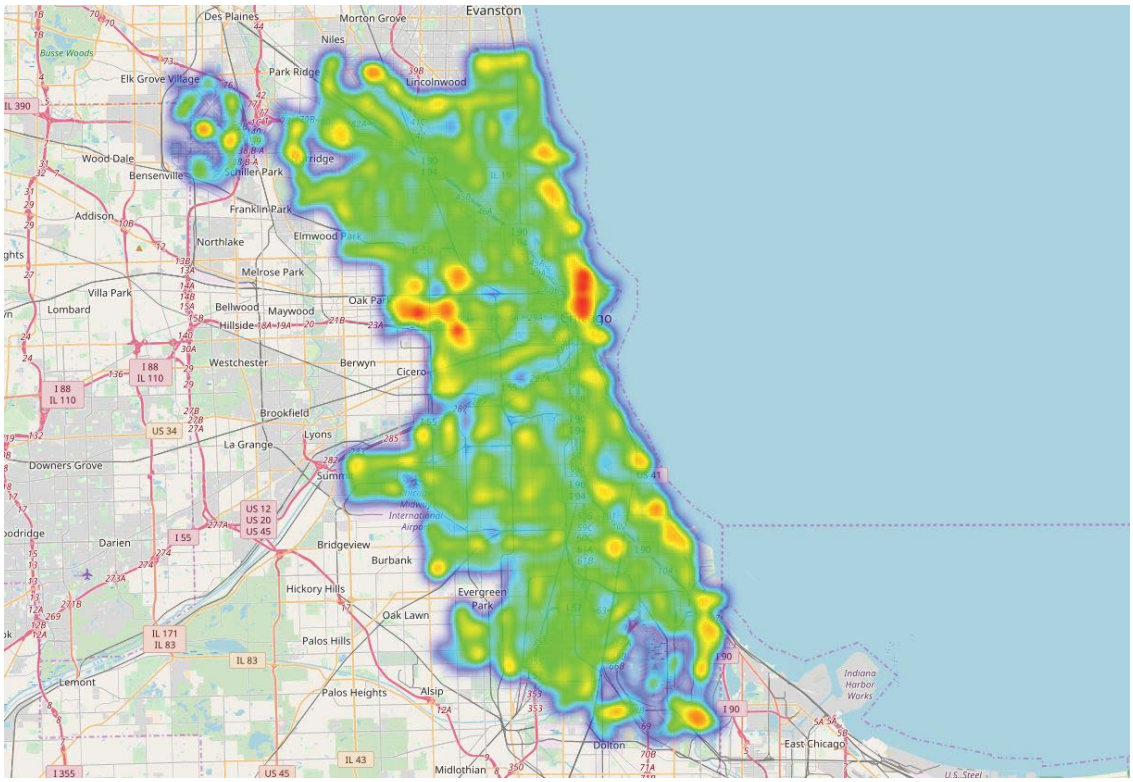


fig.night

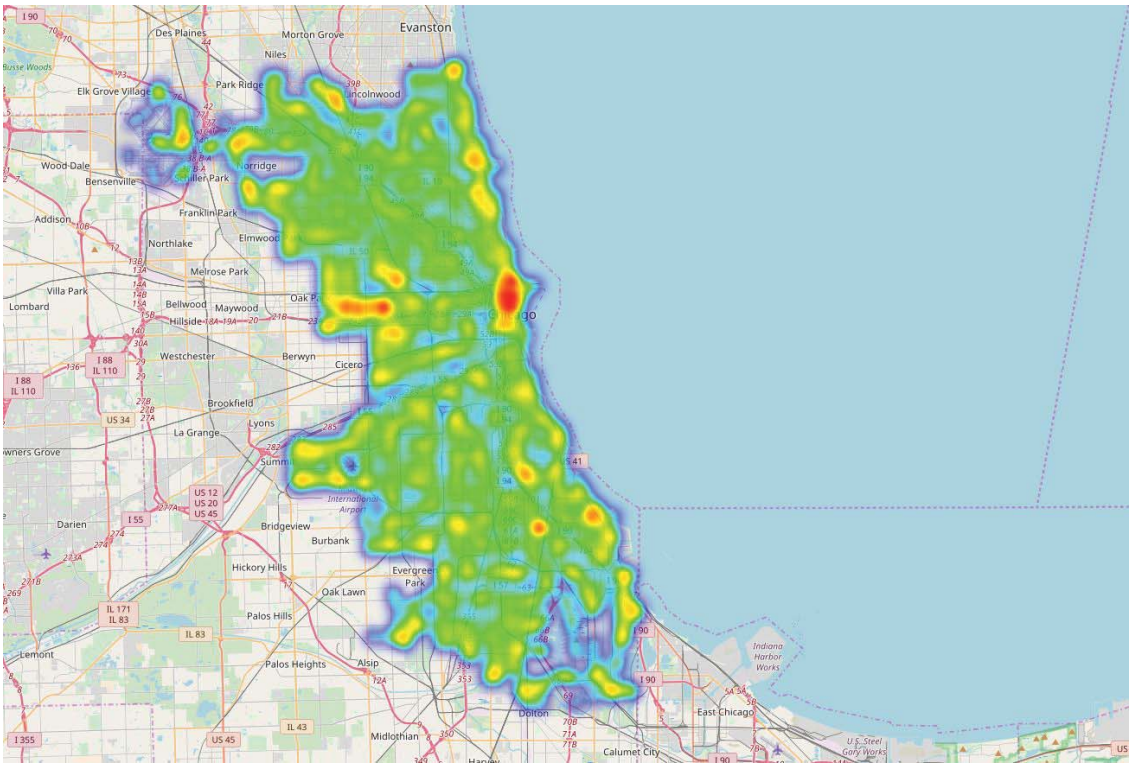
Then, we compare the heatmap over different seasons. We basically repeat what we did for daytime and night. In this case, we manually define “spring” to be from March to May, “summer” to be from June to August, “autumn” to be from September to November and “winter” to be from December to February.



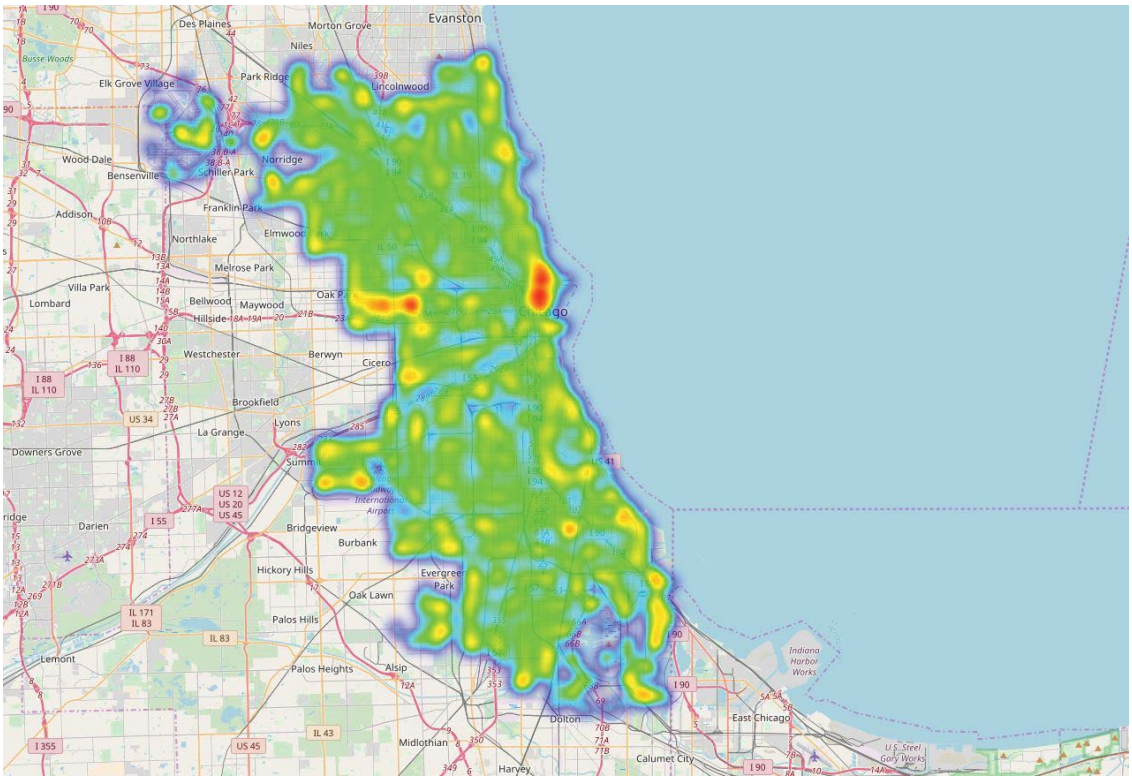
[spring.pic]



[summer.pic]



[fall.pic]

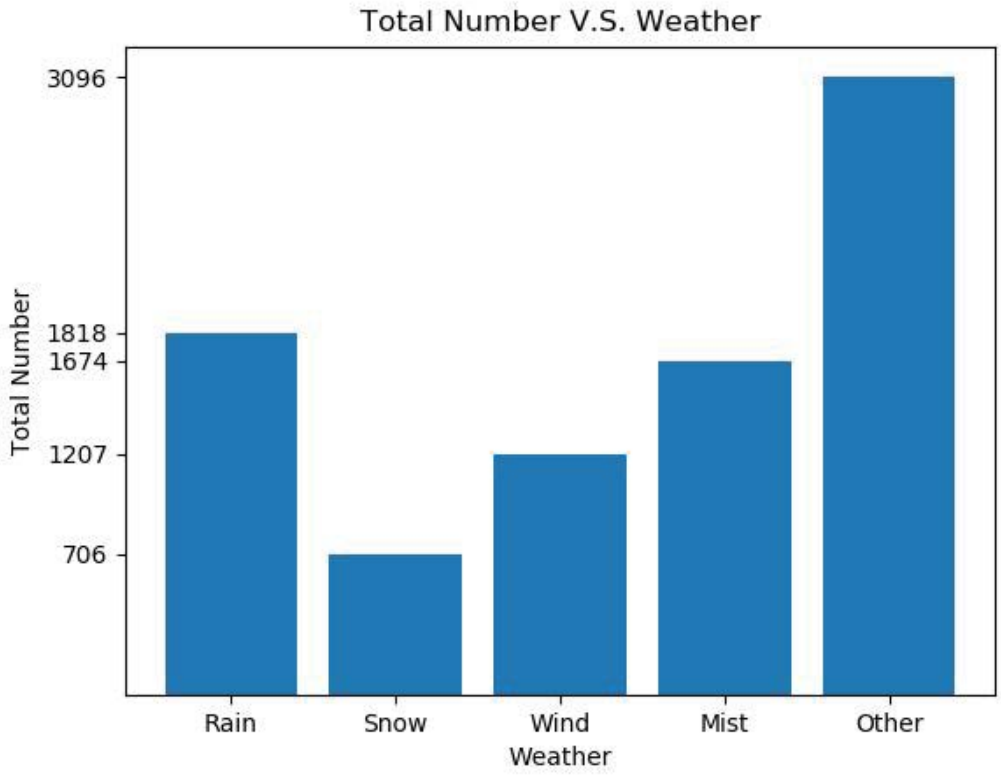


[winter.pic]

We also did time series analysis. We first grouped the coordinate by month and made heat map visualizations using those data. Then, we created screenshots of the heatmaps and combine them into a complete video showing the variation of crime data from January 2001 to December 2016.
The video link is: https://youtu.be/mM52D2GVNgU .
According to the video, there seems to be a burst of crime on 2006 and it becomes better in 2007. But when it comes to 2008, the crime rate comes up again. Finally, it seems that crime was effectively controlled by the year of 2011 and remains stable.

Analysis of Crime and Weather [dv_report.py, temp_crime.py, EvaluationByWeather.py]

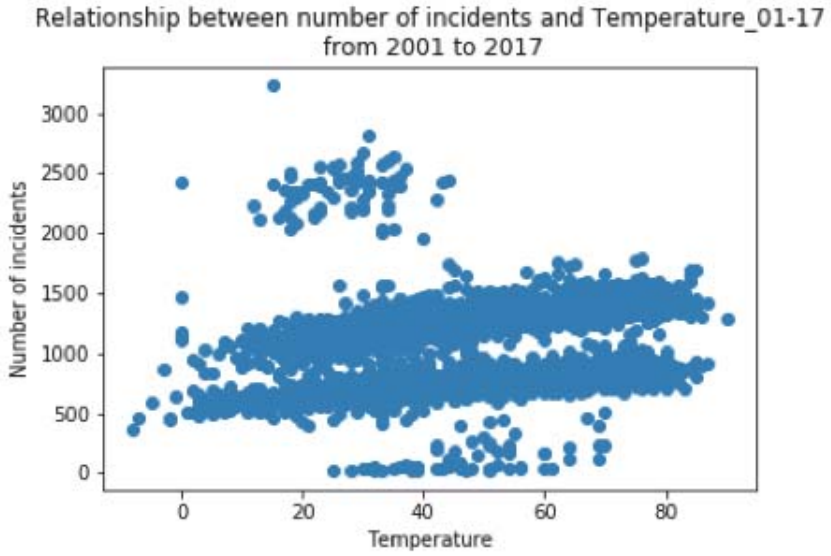
We assess the weather dataset and find that the most common weathers types except sunny day in Chicago are rain, snow, wind, and mist. Hence, the first diagram we get is the total number of criminal cases with the basic weather type. Based on the diagram below, we can know that rainy day has the highest crime cases among the most common weather types. However, this diagram doesn't have persuasion. Hence, we do further analysis based on precipitation, temperature, snow depth, wind speed and the average duration of sunshine.



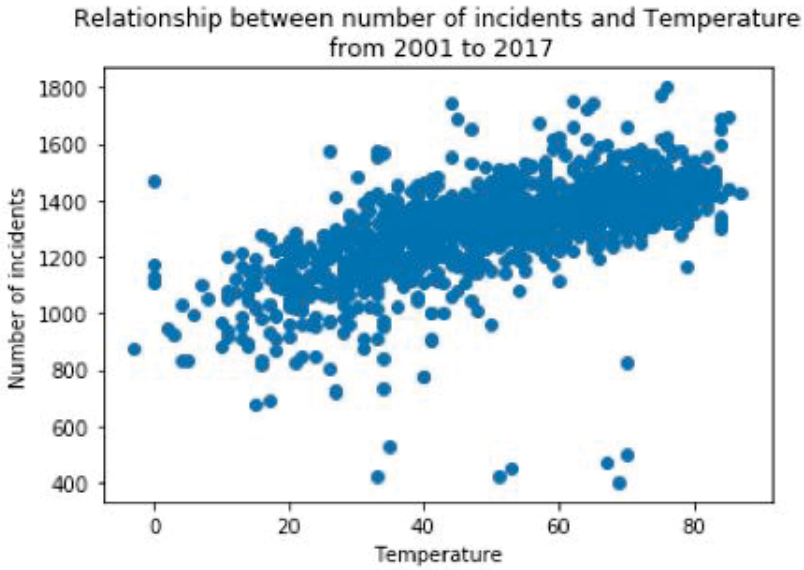
With these data, we tried to analyze whether crime rate and those weather factors are related and visualize the results. Since the dataset is huge, so we used PySpark to make it possible to read data and process data in a distributed and parallel way that could improve the efficiency of our work. After

obtaining the result of our data, we used numpy, matplotlib and scipy packages to further help me with data analysis and visualization. Specifically, we used matplotlib to draw scatter diagrams and scipy.stats.pearsonr function from scipy to calculate a Pearson correlation coefficient and the p-value for testing non-correlation.

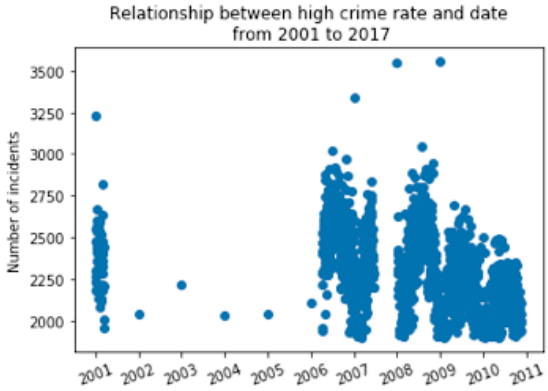
From the result, we could see there is a positive correlation between the number of incidents and temperature from 2001 to 2017 with Pearson correlation coefficient 0.13791, which seems there is no strong relationship between crime rates and temperature.

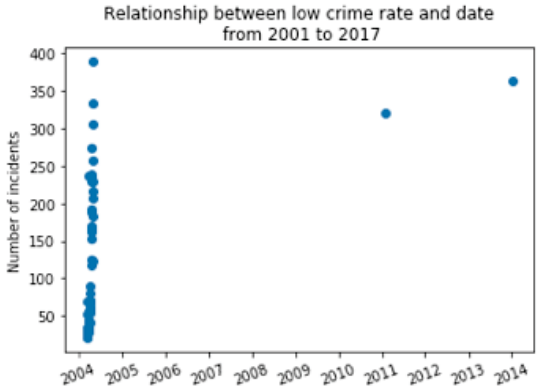


However, with a closer look at the graph, we found some interesting results that the crime rate could be relatively high on some days around 15 to 40 °F, whereas the crime rate could be relatively low on some days around 40 to 60 °F. If we remove those outlier data points, we could find a medium relation between temperature and crime rate with Pearson correlation coefficient 0.64581, which means a higher temperature will result in a higher crime rate in general, as shown in the graph below.

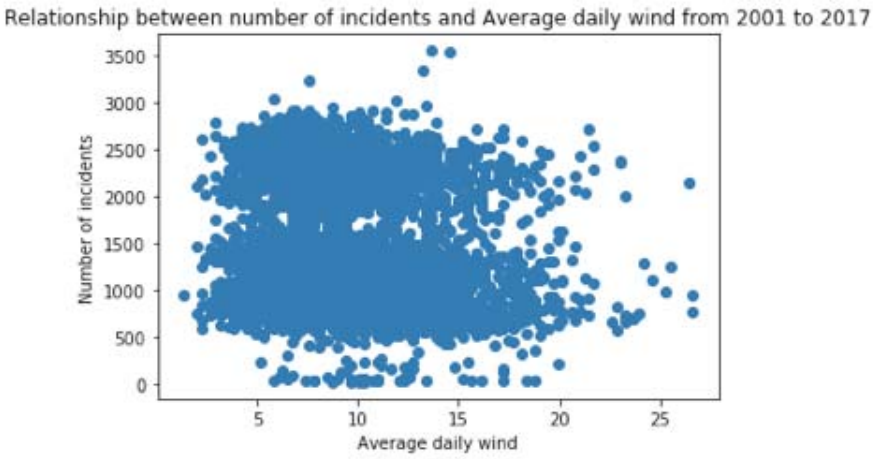


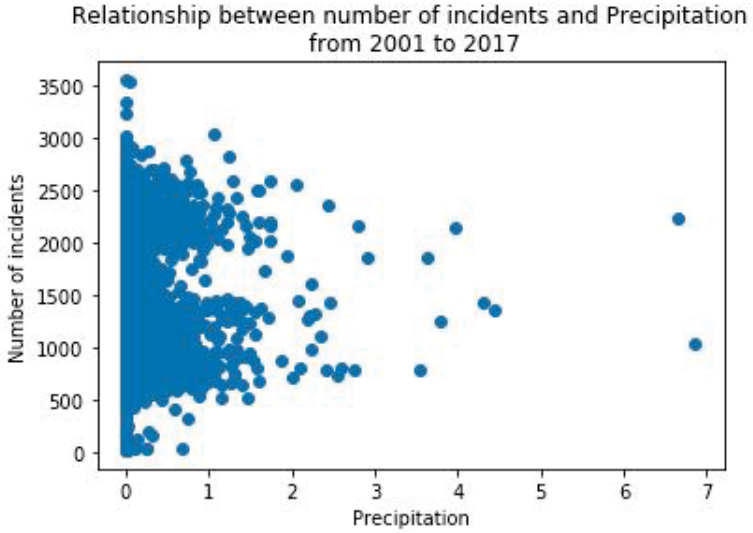
But what are those high crime rate and low crime rate data points? By plotting those extreme numbers, we got the following figures, where most high crime days happen between 2006 and 2011, with relatively few days of low crime days happen in 2004. Since the crime rate has sudden increase since 2016, although from other analysis we have before, we conclude this sudden change as the bias that is very likely to exist.

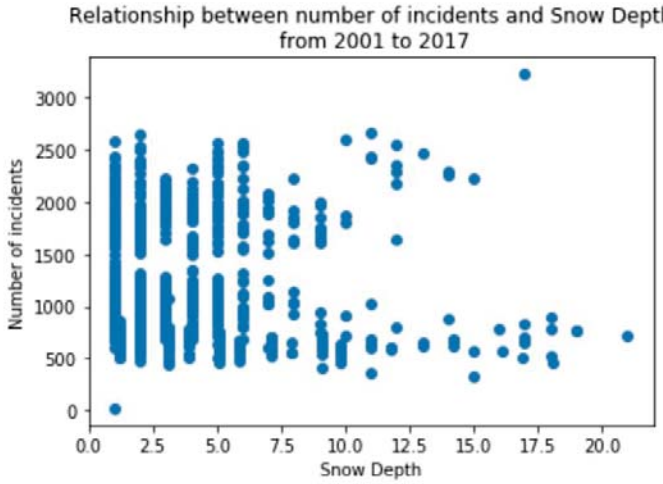




As for other data, there is a relative low Pearson correlation coefficient close to zero, so we could conclude there is no relation between crime rate and average daily wind, precipitation or snow depth.







Challenges We Meet

In order to modularize our codes and enhance the efficiency, we decide to implement Object-Oriented Programming (OOP). Since there is no resource online that introduces the concept of applying OOP over spark, we had a hard time of organizing script coding style into an OOP coding style. To resolve the challenge, we looked into the conceptual python OOP programming guide and twisted the arguments for each function to follow OOP programming rule.

There are also various challenges we met when we are doing data visualizations. The first challenge is to select the best way to represent data. It also involves a lot of document reading and parameter adjusting. We cannot freely and randomly adjust the parameters we want because the course cluster is too slow to allow us to infinitely try for the best parameters to pass. The second challenge is data cleaning. The data source we get from internet contains a lot of noises and missing values. The third challenge is EWS limitations. No sudo privilege in EWS has caused me a lot of problems and we spend a lot of time importing required packages. A lot of time is wasted debugging our correct code.

The challenge of finding the relationship between crime rate and weather is mostly to find proper ways to visualize the data and interpret the graph. As we mentioned above, the first graph that we plotted about criminal rate and the temperature is somewhat unexpected, since there is a cluster of an extremely high rate of crime and those data drives the whole correlation factor down. We solved those challenges by further carefully analyzed the data and we found out that the bias in data is highly likely to exist. Therefore, we decided to remove those points to make the approximations.

Reference Link:

https://www.kaggle.com/currie32/crimes-in-chicago

https://www.statista.com/

https://www.ncdc.noaa.gov/cdo-web/datasets#GHCND