Chicago Crime Data and Census Data Insight

Introduction

Crime rate is believed to be highly related to economy. The economic crisis that happened in 2008 is the turning point of the world's economy. Therefore, it is meaningful to study whether this turning point has affected the crime rate from the real data and how crime rate may affect the local economy.

This project has chosen one of the metropolises in the US, Chicago as the investigation object. The time span of interest is 9 years spanning the year of 2008. The objective of the project is to provide an insight of the crime data scraped from Chicago Data Portal and conclude the impact of 2008 crisis on Chicago's crime rate. Additionally, the census data of Chicago from 2008 to 2012 will be studied to find out the impact of crime on the households' financial situation of individual community.

The target audience of the project with its elaborate visualisation results will benefit any groups of audiences who are interested in economy-crime topic or those who are concerned about crime rate when doing investment or business such as real estate and small business.

The rest of the report consists of the following parts. Data section includes data collection and data preparation. In Methodology section, the detailed data analysis methods adopted in the project will be demonstrated. An analysis of the results obtained as well as some discussion and findings will be discussed in the Results and Discussion section followed by the Conclusion section.

Data

This section will introduce what the required data is, where and how the data is collected and how it has been cleaned and prepared for later use.

Understanding the datasets

This project will be using two datasets that are available on the city of Chicago's Data Portal:

- Chicago Crime Data

  This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.

  Chicago Crime Data will be used to obtain a comprehensive understanding of crime rate from 2004 to 2012 including the trend, types of crimes, crime rate of different communities and so on. Data visualization will help to achieve this goal. Census data of all the communities of

Chicago will be jointly considered to give a clear picture of the connection of community population and income to crime rate of that area.

The Chicago Crime Data contains 22 features as shown in the results of the following Python code. The primary features we are interested in are listed below:

- **YEAR:** Year the incident occurred.
- **PRIMARY_TYPE:** The primary description of the IUCR code.
- **COMMUNITY_AREA_NUMBER:** Indicates the community area where the incident occurred. Chicago has 77 community areas.
- **LOCATION_DESCRIPTION:** Description of the location where the incident occurred.
- **LATITUDE:** The latitude of the location where the incident occurred.
- **LONGITUDE:** The longitude of the location where the incident occurred.

➢ Census Data - Socioeconomic Indicators in Chicago

This dataset contains a selection of six socioeconomic indicators of public health significance and a "hardship index," by Chicago community area, for the years 2008 – 2012. The indicators are the percent of occupied housing units with more than one person per room (i.e., crowded housing); the percent of households living below the federal poverty level; the percent of persons in the labor force over the age of 16 years that are unemployed; the percent of persons over the age of 25 years without a high school diploma; the percent of the population under 18 or over 64 years of age (i.e., dependency); and per capita income. Indicators for Chicago as a whole are provided in the final row of the table.

The Census Data has 9 features and the following are essential to this project:

- **COMMUNITY_AREA_NUMBER:** Indicates the community area where the incident occurred. Chicago has 77 community areas.
- **COMMUNITY_AREA_NAME:** Community actual name
- **PERCENT HOUSEHOLDS BELOW POVERTY:** Percent of households living below the federal poverty level
- **PER_CAPITA_INCOME:** Community Area Per capita income is estimated as the sum of tract-level aggregate incomes divided by the total population

Data Collection and Data Preparation

Both datasets are available in Chicago Data Portal. To get the full dataset, one is able to extract it through Chicago API, powered by Socrata. As the project is interested in a certain period (9 years spanning 2008), crime data of year 2004 to 2012 has been extracted from the database, which contains 3686677 rows and 22 columns. In contrast, Census Data available officially is only from 2008 to 2012 and is relatively small.

The first step of the preparation of both datasets is to filter the interested features which has been introduced in the beginning of the section. After filtering, NaN rows in Chicago Crime Data have been dropped and proper data types have been redefined for features such as community area number, latitudes and longitudes. Similar cleaning work was done to Census Data. For convenience, the names of common features of both datasets are kept consistent.

```
In [7]: print(Chicago_Crime.shape)
        Chicago_Crime.head()
```

```
(3659533, 6)
```

Out[7]:

	Year	Community Area Number	Primary Type	Location Description	latitude	longitude
0	2012	1	DECEPTIVE PRACTICE	APARTMENT	42.001670	-87.673864
1	2011	70	DECEPTIVE PRACTICE	RESIDENCE	41.747362	-87.708424
2	2012	19	THEFT	OTHER	41.917863	-87.744601
3	2012	30	DECEPTIVE PRACTICE	RESIDENCE	41.841362	-87.729335
4	2011	25	DECEPTIVE PRACTICE	RESIDENCE	41.891766	-87.766554

Figure 1: Chicago Crime Data

```
print(CensusData.shape)
CensusData.head()
```

```
(77, 4)
```

	Community Area Number	COMMUNITY AREA NAME	PER CAPITA INCOME	PERCENT HOUSEHOLDS BELOW POVERTY
0	1	Rogers Park	23939	23.6
1	2	West Ridge	23040	17.2
2	3	Uptown	35787	24.0
3	4	Lincoln Square	37524	10.9
4	5	North Center	57123	7.5

Figure 2: Census Data

Methodology

This section covers the methodology used to analyze Chicago's crime data as well as the census data. The aim is to elaborate any correlations, trend and distribution by means of statistical analysis, visualization and machine learning tool. Specifically, the project first uses exploratory analysis to provide a complete picture of crime data over the 9 years. Methods include bar chart, line chart, box plot and heat map. Moreover, k-means clustering has been used to cluster communities according to the frequency of different crimes. In order to verify whether crime rate may have an impact on the community's economy, communities with different economic situations are plotted on the crime heat map.

Exploratory Analysis

- **Variation trend of total number of crimes**

The chart aims to provide an overview of the total number of crimes that happened in Chicago each year and demonstrate a trend over the 9 years being investigated.

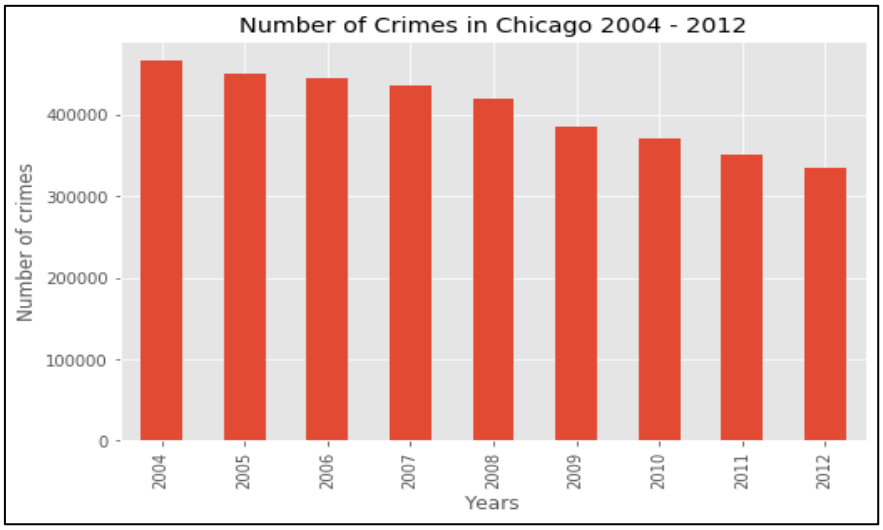

Figure 3: Total number of crimes of each year from 2004 to 2012

- **Total number of crimes of different types**

Overall view of the total number of crimes of different types over the 9 years. Only those types committed over 5000 times are plotted (Figure 4).

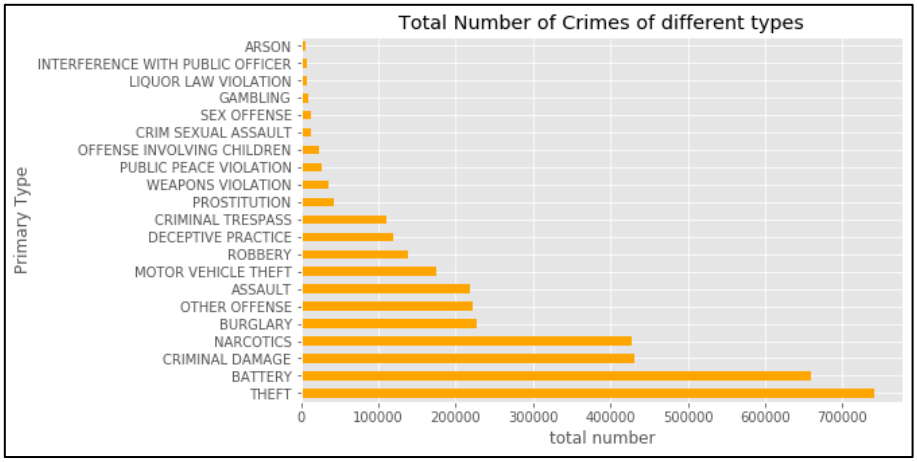

Figure 4: Total number of crimes of different types over the 9 years

- **Economy-related crimes**

This line chart aims to show a variation trend over the 9 years of crime types that are potentially related to economy such as theft, robbery, narcotics and so on. Meanwhile, a box plot is used to statistically demonstrate the distribution.

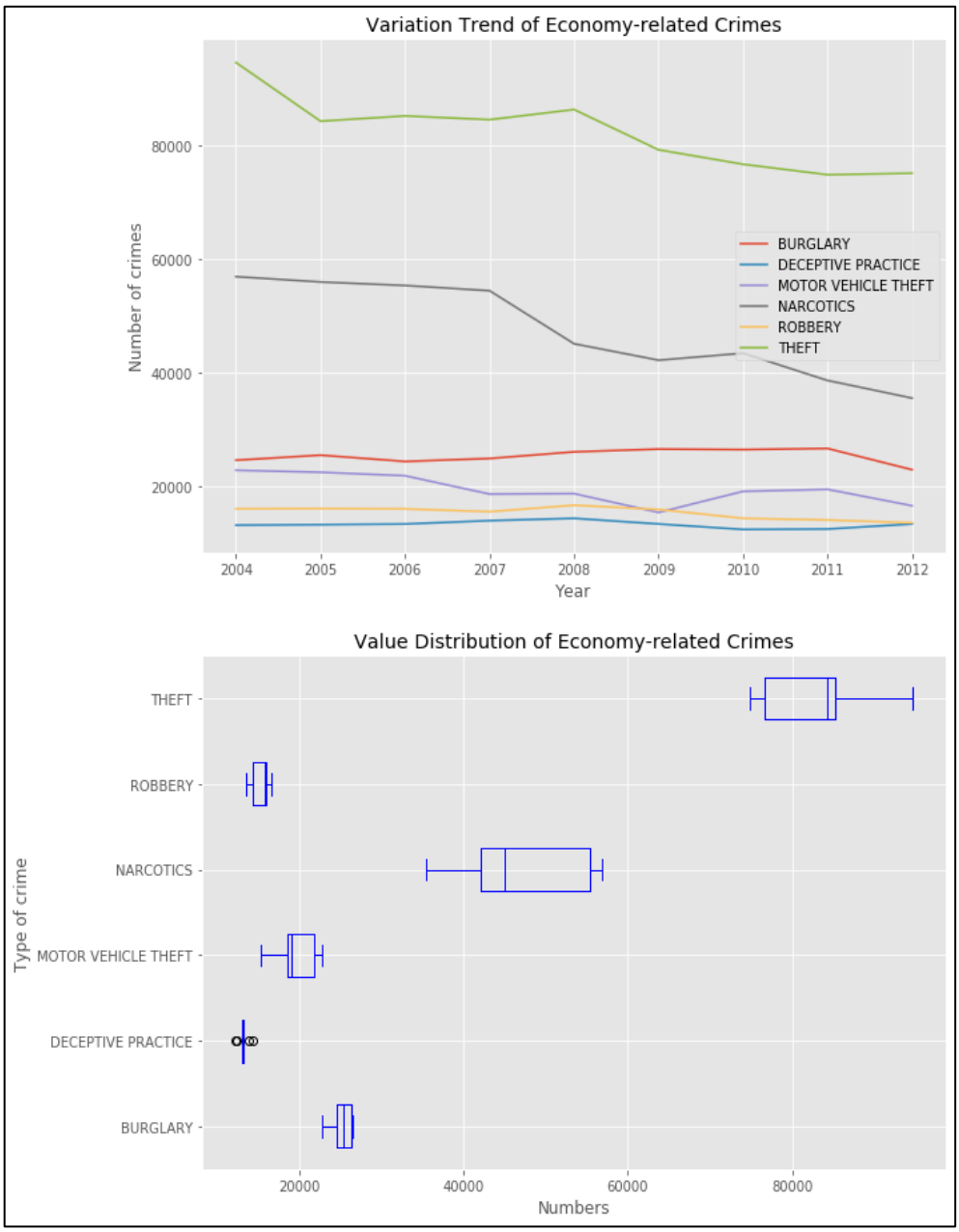

Figure 5: Economy-related crimes: trend and distribution

- **Heat map**

The heat map shows the locations of the crimes committed in Chicago. The aim is to show some areas with high crime rate.

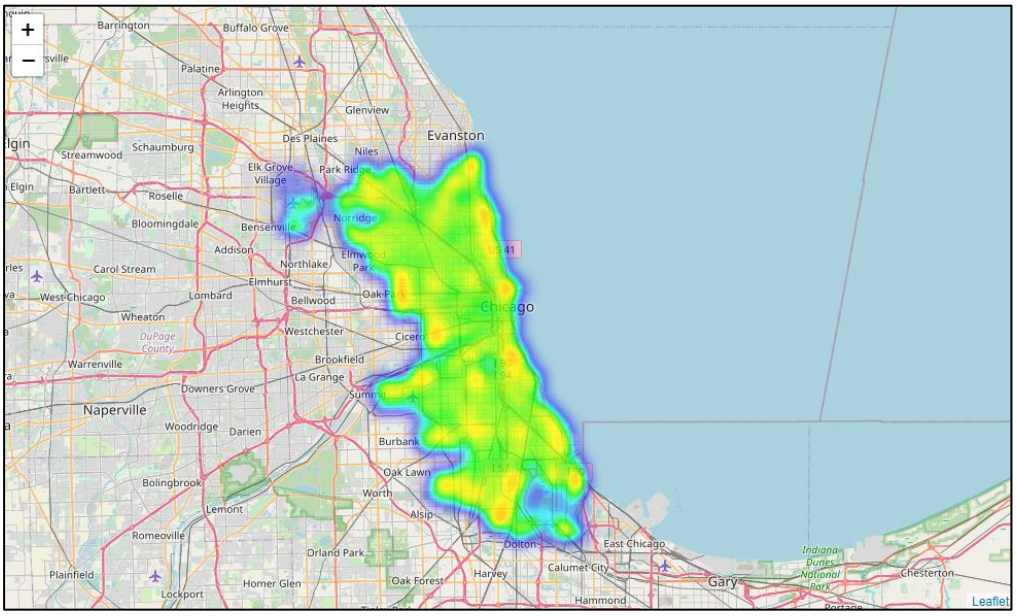

Figure 6: Heat Map of crimes in Chicago

- **K-means clustering**

With K-means clustering, the objective is to cluster 77 communities based on the frequency of crimes of different types. As the community coordinates are not available, average latitudes and longitudes of all the crimes are used (Figure 7).

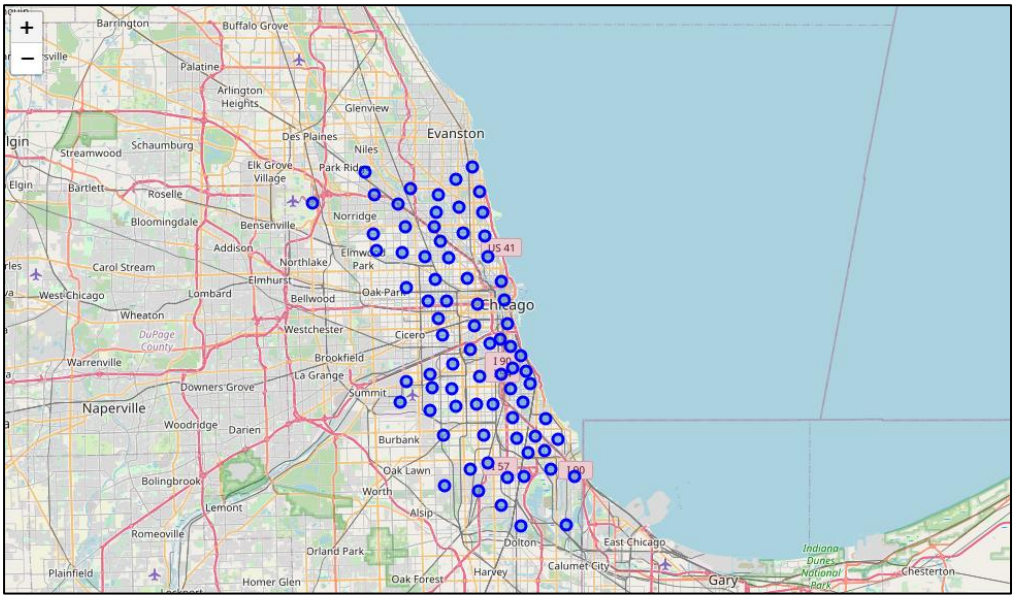

Figure 7: Community locations

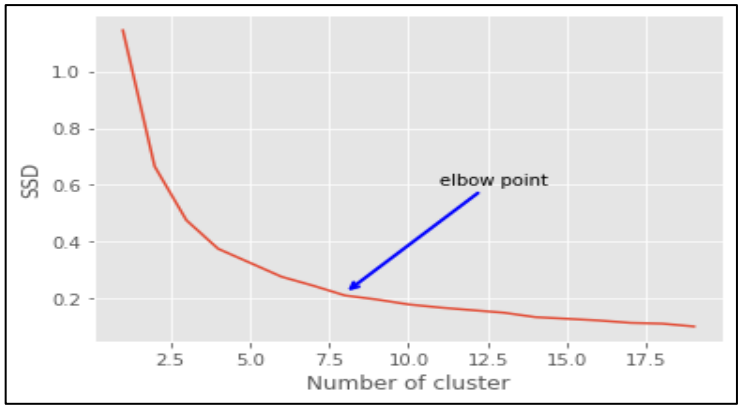

Figure 8: Figure to show the elbow point

In terms the best K, a range of K values have been tried to find the elbow point as shown in Figure 8[1] and the best K chosen is 8. With the optimal K, the 8 clusters of communities are shown in Figure 9.

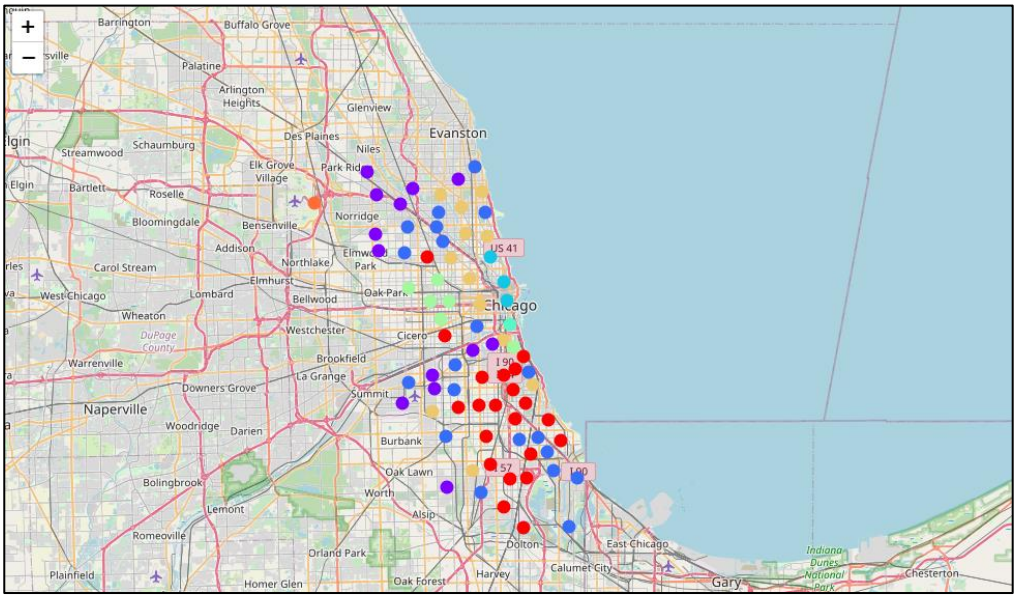

Figure 9: Clusters of Communities

- **Community economy vs crimes**

To show the potential correlation of community financial situation with crimes, the simplest way is plot relevant data points on the heat map. Therefore, the project selects top 10 poorest

[1] SSD: Sum of distances of samples to their closest cluster center

and top 10 richest communities in terms of poverty rate and income respectively and plot them on the heat map of crimes after 2008 (Figure 10 and Figure 11).

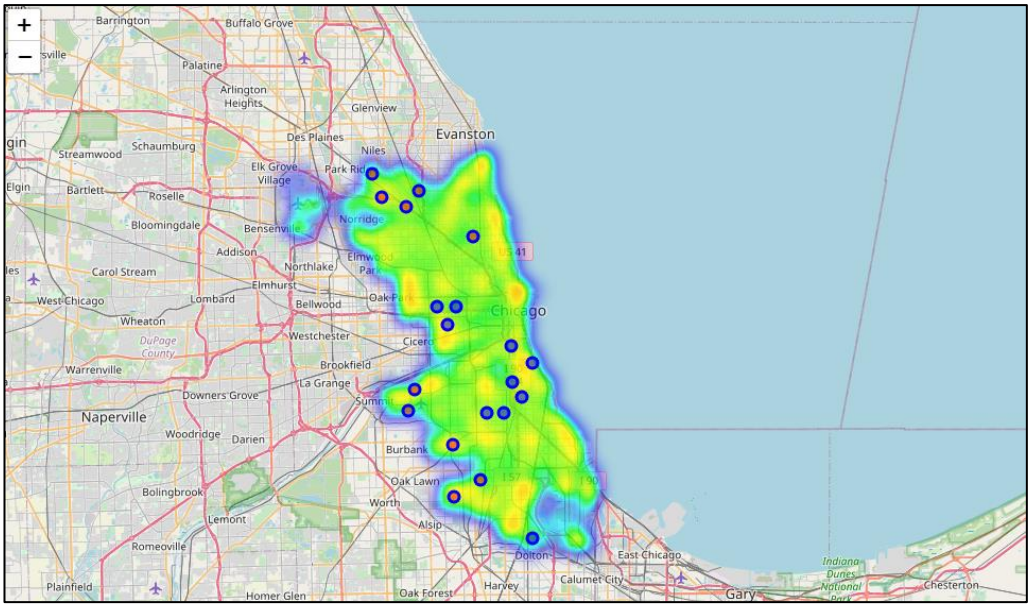

Figure 10: Community economic situation from poverty percentage aspect: blue points being the 10 communities with highest poverty rate; red points being the 10 with lowest poverty rate

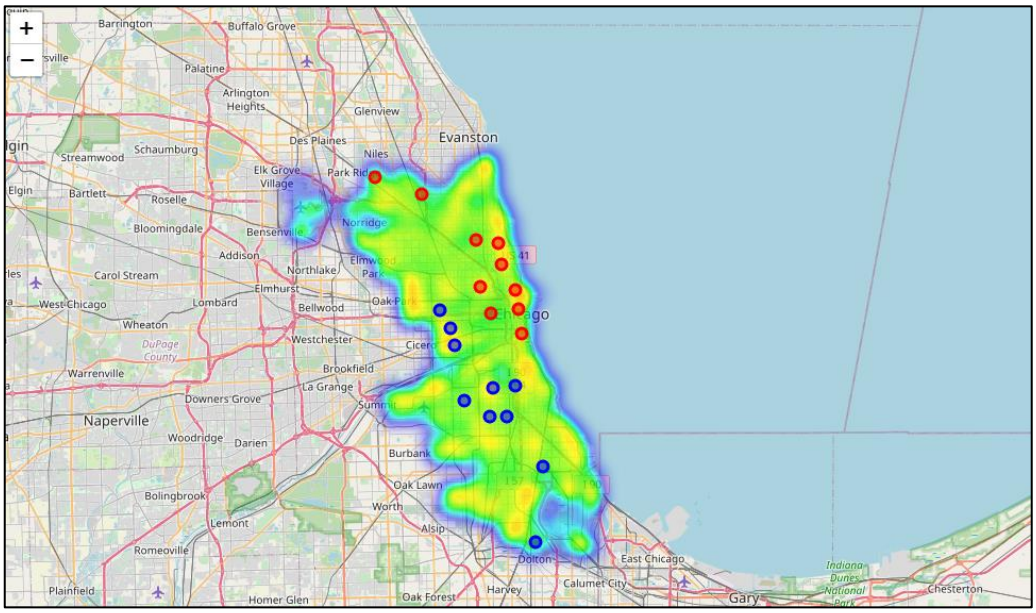

Figure 11: Community economic situation from capital income aspect: blue points being the 10 communities with lowest income; red points being the 10 with highest income

Results and Discussion

This section will analyze the results so far and discuss its potential reasons or any interesting observations.

Figure 3 and Figure 4 provide a general picture of the crime data from 2004 to 2012. In Figure 3, it is clear that the total number of crimes was declining gradually from nearly 500,000 in 2004 to approximately 350,000 in 2012. The observation is obvious that the year of 2008 when the world economic crisis happened did not bring a surge to Chicago's crime rate. On the contrary, the total number of crimes dropped dramatically in 2009 more than any year before. Figure 4 illustrates the number of crimes of different types. Theft and battery are the top 2 most frequent crime types and their quantities are nearly twice as the third and forth place, criminal damage and narcotics respectively.

Figure 5 aims to focus on crimes that are highly related to economic factors. For this purpose, the project has chosen 6 types of crimes i.e., theft, robbery, narcotics, motor vehicle theft, deceptive practice and burglary. The line chart shows the variation trend of these types. Burglary, robbery and deceptive practice remained in a stable level over the 9 years, though the former two rose slightly after 2008. Other three types all experienced an obvious decline especially for theft and narcotics. Although they are economy-related crime types, the figure has shown that the 2008 crisis did not negatively impact them. The box plot of figure 5 aims to give an idea of the data distribution of the 6 types.

Heat map helps to spot high crime areas. Figure 6 gives a comprehensive view of crimes that happened in Chicago spanning from 2004 to 2012. Essentially, eastern coastal areas of Chicago seem to see more frequent crimes forming a band. Besides, the southern and western Chicago also had areas that suffered from high crime rate.

	COMMUNITY AREA NAME	ARSON	ASSAULT	BATTERY	BURGLARY	CRIM SEXUAL ASSAULT	CRIMINAL DAMAGE	CRIMINAL SEXUAL ASSAULT	CRIMINAL TRESPASS	DECEPTIVE PRACTICE
0	Albany Park	0.001857	0.052711	0.182370	0.069456	0.003405	0.149189	0.000031	0.027331	0.029621
1	Archer Heights	0.002651	0.044179	0.127384	0.082910	0.003313	0.187100	0.000000	0.013254	0.033355
2	Armour Square	0.000688	0.043605	0.140453	0.047353	0.000994	0.130202	0.000000	0.046053	0.069997
3	Ashburn	0.001538	0.065941	0.160755	0.095968	0.003461	0.167293	0.000000	0.019199	0.036979
4	Auburn Gresham	0.001372	0.068055	0.207353	0.074339	0.003786	0.111437	0.000085	0.029329	0.023168

Figure 12: Example of the table of occurrence rate

One interesting thing to investigate is to find out how the communities can be grouped based on the occurrence rate of different crimes. To this end, the project adopted K-means clustering. Before clustering, it is necessary to create a dataset that contains the occurrence rate of crimes for each community such as Figure 12. To find the best K i.e., the number of clusters, the project has tried values ranging from 1 to 20 and by showing the sum of distances of samples (SSD) to their closest cluster centres vs the values of K in Figure 8, one can identify the elbow point (K=8) after which the declining trend becomes much smaller and stable. The actual clusters shown in Figure 9 demonstrates that for some clusters e.g., the red and green cluster, community locations can be agminate whereas the communities in purple and blue clusters, for example, look rather dispersive.

In terms of the relationship of community economic situation and its crime rate, Figure 10 and Figure 11 provide an insight. To be specific, consider the percent households below poverty of different communities in Figure 10. The top 10 communities with high poverty rate mostly gather

in middle part of Chicago as shown in blue points while the top 10 ones with low poverty rate shown as red points distribute in the northern and southwestern suburbs. It is also obvious to observe that basically all blue points locate on or are quite close to the areas of high crime rate whereas the red points have much safer neighbourhood. Figure 11 gives out a different picture in terms of the feature, per capita income. Communities of high income (red points) gather in the eastern coastal and downtown areas where the crime occurrence is also frequent. By contrast, low-income communities in blue points have a similar distribution with those in Figure 10.

Conclusion

The report summarizes the work of analysing Chicago crime data and census data. It has provided the readers with a comprehensive understanding of the crime data of Chicago from 2004 to 2012. It illustrates trend and classifications to conclude that 2008 economy crisis did not necessarily affect the crime rate in a negative way and heat maps as well as K-means clustering for a complete picture of this dataset. Combining the census data, the report has also shown the potential relationship between crime rate and community economic situations.