Safety Guide to Survive in Chicago

Name: Yiyang Lu

ID: 501403

Content

Executive Summary

This project aims to provide safety tips to newcomers to Chicago by analyzing the twenty-one-year crime data recorded by the Chicago Police Department. Being aware of the decreasing arrest rate of crimes, people shall be more vigilant in their surroundings in Chicago and be especially aware of potential common crimes such as theft and battery. To reduce property

damage and life threats, people should avoid potentially dangerous zones provided in our guide. Besides, gaining a general overview of the distribution of crime could also assist Chicago P.D. to have a more reasonable way to allocate the police force, saving the budget and making it more effective. The high crime rates should not tarnish the fame of Chicago. Providing a safety guide could prevent visitors from being harmed for not being familiarized with the circumstance of Chicago.

Another point is to find out whether Chicago P.D. is trying its best to close the criminal cases as soon as possible. Our findings indicate that the case stacking ratio of Chicago P.D. is decreasing as time flows. The arrest rate is also decreasing. I assume that COVID-19 lowers the police case processing rate. The most frequent primary crime type is theft and battery, which could be prevented by installing monitors or increasing the number of patrols per police beat. July and August are the most dangerous months for all travelers to Chicago. December and February are two ideal months to visit the city. Most crimes happen around O'Hare airport, so I recommend people be alerted all the time when they are around.

Description of dataset:

1.Overview

This dataset reflects the reported crime incidents in Chicago in twenty-one years from 2001 to 2021. The structured dataset was collected and updated each year on Chicago Data Portal. Data in the dataset was originally extracted from the CLEAR, short for Citizen Law Enforcement Analysis and Reporting System of Chicago Police Department.

2.Details

The released dataset is named Chicago_Crimes_2001_2021.csv. The size of the twenty-one-year dataset combined is 1.70 GB. This dataset contains 7448538 rows of crime reported in Chicago from 2001 to 2021. The data file includes the following 22 columns: case ID, case number, date, block (in terms of where the crime occurred), IUCR (The Illinois Uniform Crime Reporting code.), primary crime type, crime description, location description, arrest (whether the crime made an arrest yet), domestic, beat, district, ward, community area, FBI code, X coordinate, Y coordinate, year, updated on, latitude, longitude, and location.

Link to dataset: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

Why is this big data?

Although Chicago is a giant tourist destination for many good reasons, it also has the reputation of being one of the most violent cities. The higher-than-average crime rate concerns many people. Each month, hundreds of each type of crime are reported in Chicago. Over the years, the crime dataset can become overwhelmingly large for local police to analyze. Following the recent death of University of Chicago international students, I want to study crimes patterns in Chicago to bring people in Chicago alerts on crimes.

This dataset contains 7448538 rows of data, and each row includes 22 features. It is a giant dataset to run traditional local data processing analysis. As time goes on, the Chicago police department will have a growing volume in this crime dataset. In this case, this twenty-one-year dataset is big data.

Problem Statement

1.Problem statement:

People living in the city of Chicago are concerned about their safety. There are still many things to do to lower the city's crime rate, something I can do now to provide crime alerts based on analysis from historical crime records to the public. Hence, people get a chance to avoid some potential risks to personal and property safety. The purpose of our study is to point out the danger in Chicago and offer safety tips to those who are new to the city.

2.Research questions

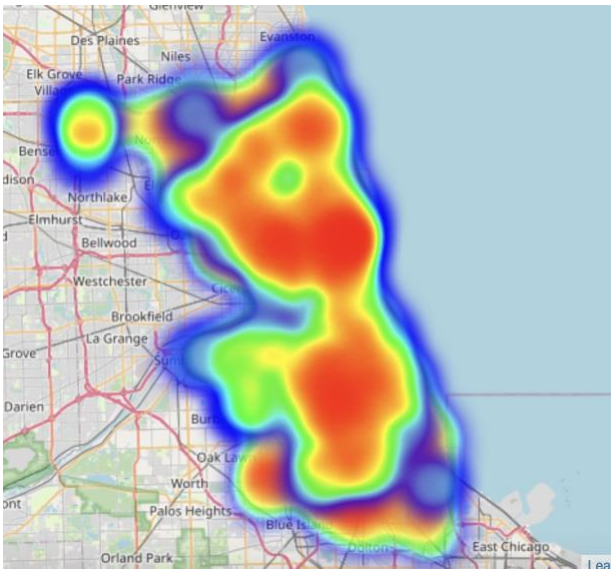
- Which times of the year do crimes occur the most?
- Which location suffers the most from crime?
- How has the arrest rate changed over the years?
- Crime locations and their crime type versus arrest rate?
- Average time of crimes occurred per month per year?
- Crime distribution in Chicago.
- Average time took to make an arrest on a case.

Method & Results:

1. Process Overview:

a. Download dataset to local and use python to remove missing values and duplicate values.
b. Upload to server
c. Analyze Data using SQL package in Python
d. Analyze Data using Linux & MapReduce
e. Visualize Data using Tableau and Python packages

2. Integral Analysis

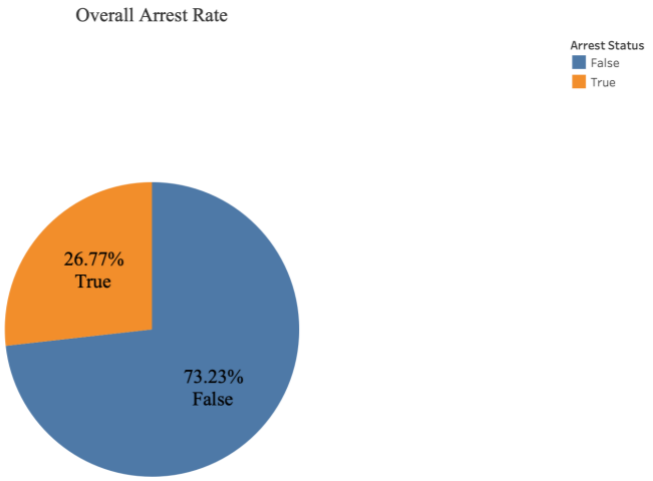


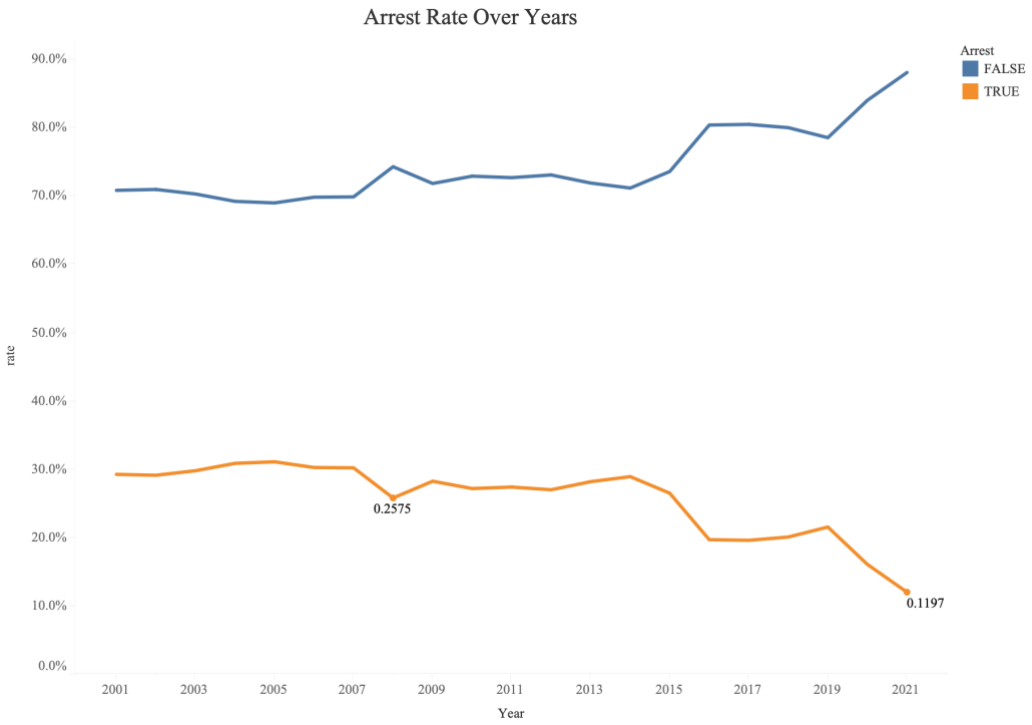
I plotted all the reported crimes cases that occurred in the past twenty-one years in the city of Chicago using the longitude and latitude coordinates in the dataset. The heatmap correctly reflects Chicago's reputation for high crime risks. Knowing Chicago's high violent crime rate, I want to offer some safety tips to students and travelers who temporarily visit Chicago.

a. Arrest Rate Comparison:

According to the data reported by the CLEAR, the pie chart below presents an overall rate that an arrest has been made on a crime and the rate that no arrest has yet been made for the past twenty-one years. The arrest rate on a crime is about 26.77% for the past twenty-one years. So

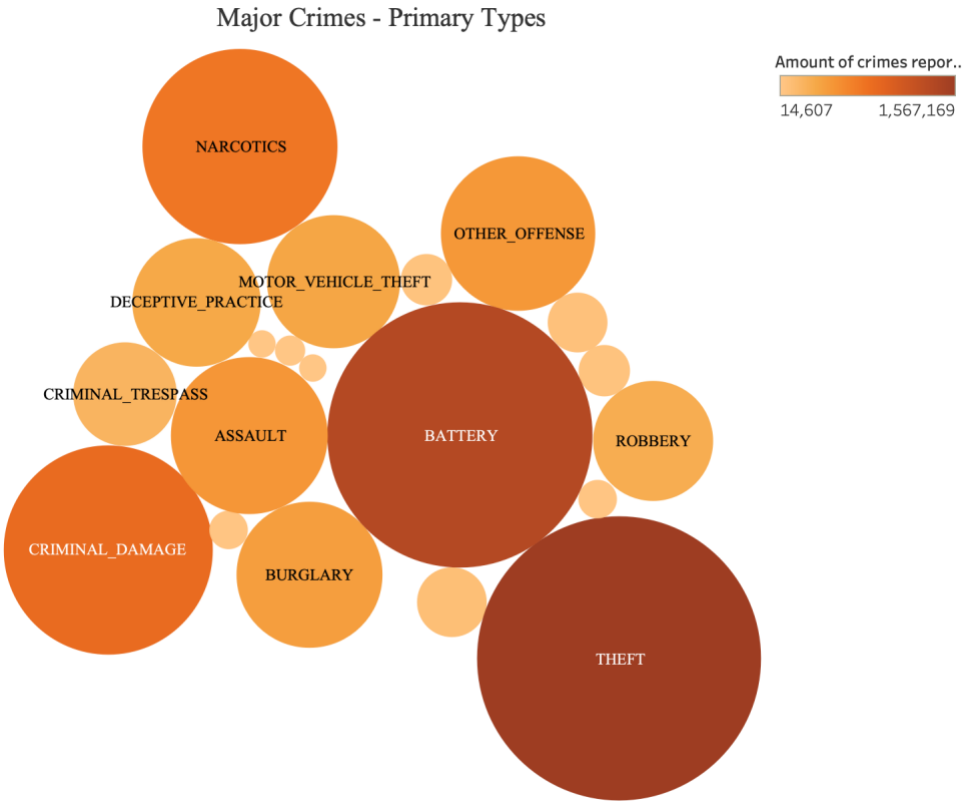
about one arrest is created for every four crimes. For the arrest rate over the years chart, the column ‘Arrest Rate’ is calculated using the number of arrests divided by the number of crime cases reported within the year. The overall arrest rates throughout the year show a steady trend until 2014. Starting in 2014, the arrest rate drops, with a slight recovery in 2019, then drops to the lowest level ever in 2021. People should not lose their vigilance about their surroundings when commuting or traveling in Chicago.





b. Primary Rates of Crimes

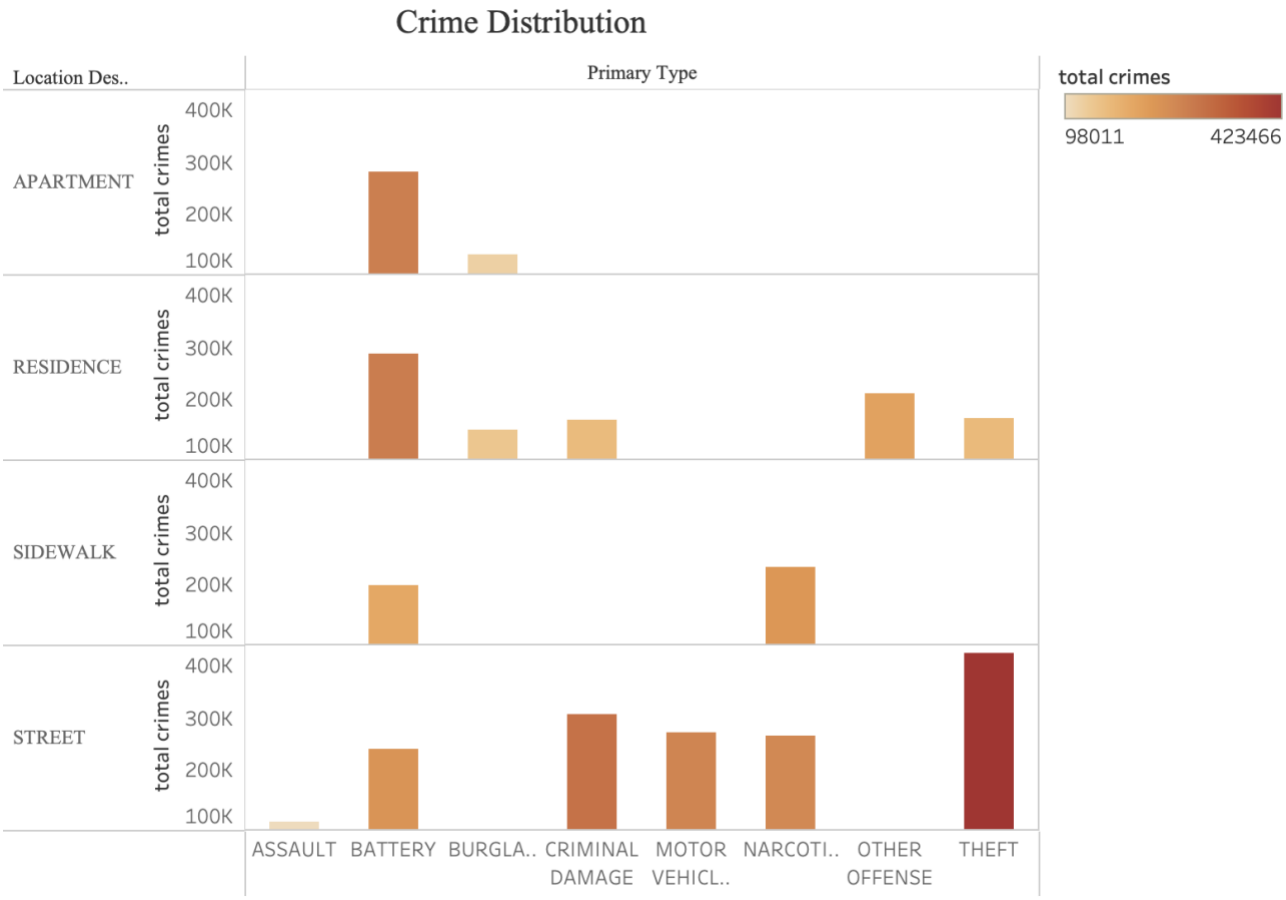
The chart below shows the primary types of crimes. Larger circle sizes and darker shades (shown on the scale on the right) indicate that more total crime occurred on the kind of crime. Theft and battery are the two most common crime types. Criminal damage, narcotics, assault, and burglary are also among the most concerning lists.



3. Detailed analysis:

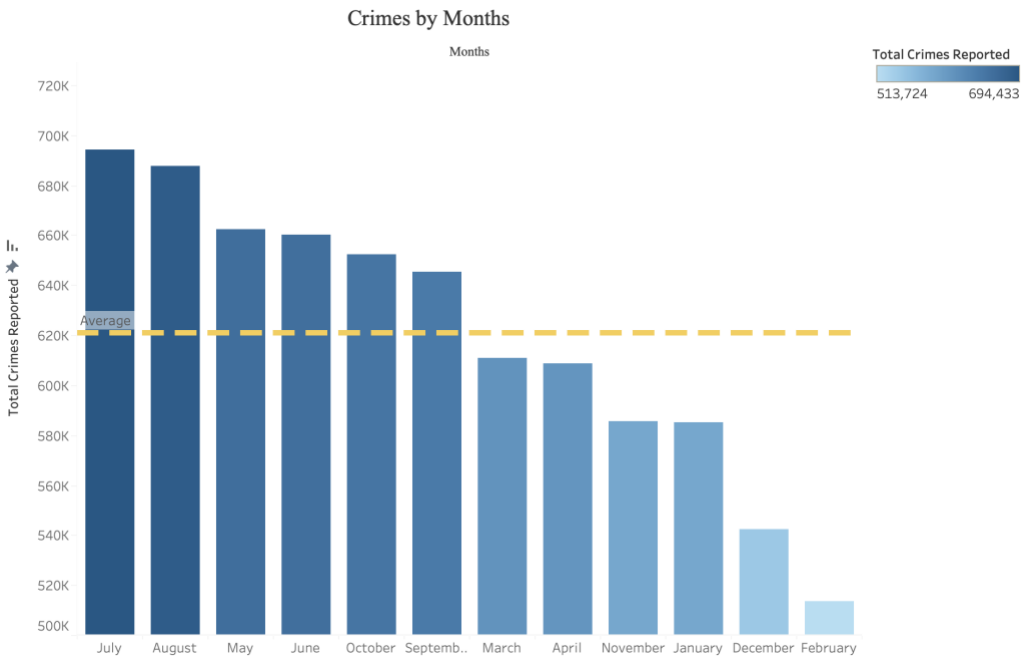
a. Crime distribution

It is helpful for the public to know where and what types of crimes frequently occur. To better analyze the primary-type crimes and the locations they take place, I regroup the total number of crimes reported by location (location description) and primary Type. I select the top 15 subgroups containing the highest number of crimes reported among all the subgroups.



All four locations that showed up have batteries of various severity reported in their top frequent primary crimes. Narcotics happen more in public areas like sidewalks and streets, while burglary primarily occurs in residential areas. The battery is more likely to occur in apartments and residential areas. Besides, more crimes about narcotics happened on the sidewalk than other primary types of crimes. More importantly, six out of the top fifteen most frequent crimes have happened on the street. The most severe one is theft, which is not surprising. The second most severe primary type is criminal damage.

b. Best time of year to travel



I utilize the twenty-one years' total crimes reported to determine the best month to travel for one year. The average number of actual crimes reported over the past twenty-one years is used as a reference line marked on the Crimes by Month chart above. The months between November and April are safer compared to other months. Thus, it is preferred that travelers plan their visit to Chicago as time allows for their safety. If possible, people should avoid travelling to the city in July and August, when crimes most frequently occur.

c. Dangerous blocks

The table below presents the top 20 blocks with the most crimes reported in Chicago. 100XX W Ohare St and 001XX N State St occurred almost doubled the number of crimes than the third most dangerous street, 076XX S Cicero Ave, on the list. Overall, Chicago people and travelers should avoid the top five blocks on the list. At the same time, they should be aware of the potential danger in other places around the city, especially other streets lower on the list below.

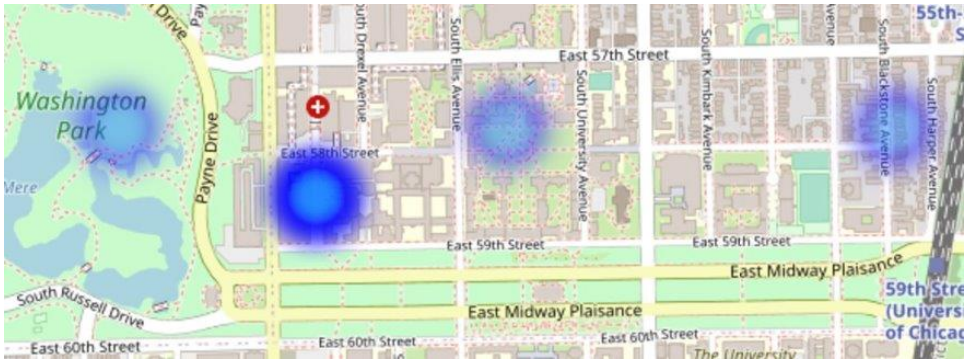
Most Dangerous Blocks in Chicago

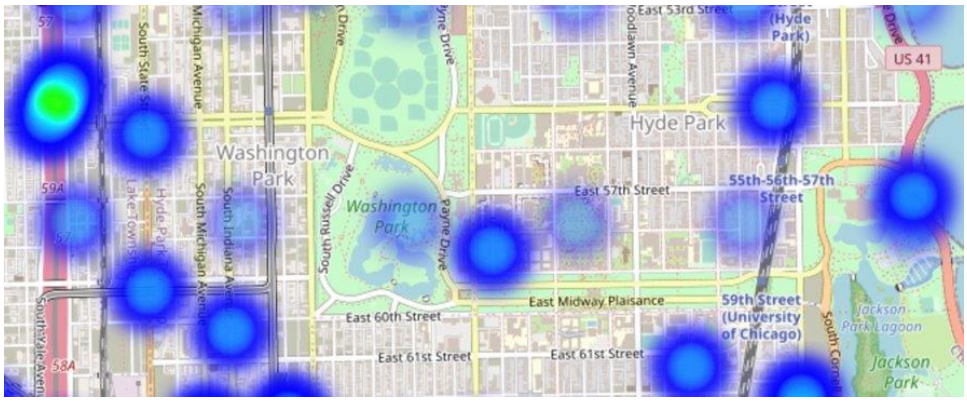
Block	Crimes reported
100XX W OHARE ST	16,009
001XX N STATE ST	14,303
076XX S CICERO AVE	9,821
008XX N MICHIGAN AVE	9,232
0000X N STATE ST	8,602
0000X W TERMINAL ST	5,744
064XX S DR MARTIN LUTH..	5,686
063XX S DR MARTIN LUTH..	5,403
023XX S STATE ST	5,227
001XX W 87TH ST	4,477
008XX N STATE ST	4,288
006XX N MICHIGAN AVE	4,131
0000X S STATE ST	4,080
012XX S WABASH AVE	4,079
022XX S STATE ST	4,005
009XX W BELMONT AVE	3,913
057XX S CICERO AVE	3,851
038XX W ROOSEVELT RD	3,717
075XX S STONY ISLAND A..	3,642
002XX W 87TH ST	3,627

Crimes reported: 3,627 – 16,009

d. Safer areas near University of Chicago campus

Due to the latest news of University of Chicago students being killed new campus, I want to provide some additional tips for students at the University of Chicago. Although the city is violent and full of crimes, students can find 'safer' spots near campus and avoid potential danger. I plotted the crimes on file that occurred near the U of Chicago. In the heatmap below, area locations with more than 500 crimes are plotted. The areas shaded in red to oranges have more crimes occurred, and the edge shades in green to blue represent fewer crimes. When looking for housing or places to go, students can use these maps, combined with our tips above, as a guide to try their best to stay in the safer zone.





e. Case Stacking Ratio Fluctuation

Year	Average Time taken to Close the Case	Case Stacking Ratio
2001	5451.89	0.71
2002	5668.3	0.78
2003	5351.43	0.77
...		
2018	39.25	0.03
2019	25.61	0.02
2020	22.4	0.03
2021	14.23	0.04

The table above represents the average length of getting the final update (the second column) and the case stacking ratio (the average length of reaching the last update divided by the total available days, the third column). The more recent a year is, the less time it takes on average to have the final update, and the case stacking ratio is lower, indicating that the efficiency of the police office improves as time flows.

Conclusion

Our Chicago crime project aims to offer safety tips to students and travelers who are not so familiar with Chicago's surroundings. Specific locations plotted on heat maps are more dangerous than other places. When planning their visits, travelers may consider this 'safety' map when picking their hotels or Airbnb for stay. Travelers should consider visiting Chicago between

November and April to avoid intensive crime activities. From May to October, visitors should pay more attention to their surroundings if they travel to Chicago. From November to April next year, people could travel to Chicago since those months are below the average. Areas near O'Hare airport have the most frequent crimes. One potential reason for the high crime frequency is the high volume of incoming and outcoming people. The decrease in case-stacking ratio indicates that the Chicago police may gradually increase their case processing ratio.

For our next steps, I will reach out to Chicago P.D. to further address our findings and concerns regarding this guide and modify this safety guide to offer to potential visitors. I also need more clarification about the 'Updated_On' column in the dataset, which is essential to determine whether Chicago P.D. has an increasing case processing efficiency.

Appendix :

Code to group dataset by primary type of crimes:

```
[jiang.y1@ip-172-31-95-86 ~]$ cut -d',' -f6 Chicago_Crimes_2001_2021.csv | tr " " "_" | sort | uniq -c | sort -rn | head -20
1567169 THEFT
1366439 BATTERY
 849461 CRIMINAL_DAMAGE
 739859 NARCOTICS
 478552 ASSAULT
 463053 OTHER_OFFENSE
 413850 BURGLARY
 343884 MOTOR_VEHICLE_THEFT
 320524 DECEPTIVE_PRACTICE
 279775 ROBBERY
 208332 CRIMINAL_TRESPASS
  94443 WEAPONS_VIOLATION
  69445 PROSTITUTION
  52509 OFFENSE_INVOLVING_CHILDREN
  51320 PUBLIC_PEACE_VIOLATION
  28829 SEX_OFFENSE
  27820 CRIM_SEXUAL_ASSAULT
  17783 INTERFERENCE_WITH_PUBLIC_OFFICER
  14621 LIQUOR_LAW_VIOLATION
  14607 GAMBLING
```

Code to group data by crime's primary type, location, and arrest:

```
[feite@ip-172-31-95-86 ~]$ cat crime_map.py
#!/usr/bin/env python

import sys

for line in sys.stdin:
        line = line.strip().split(",")
        crime, location, arrest = line[5], line[7], line[8]
        print '%s\t%s\t%s'%(crime, location, arrest)

[feite@ip-172-31-95-86 ~]$ cat crime_reduce.py
#!/usr/bin/env python

import sys

crime_c = {}

for line in sys.stdin:
        crime, location, arrest = line.strip().split("\t")
        try:
                crime_c[crime+"_"+location+"_"+arrest] += 1
        except:
                crime_c[crime+"_"+location+"_"+arrest] = 1

for key in crime_c.keys():
        print '%s,%d'%(key,crime_c[key])

[feite@ip-172-31-95-86 ~]$ cat bash_file.sh
#!/bin/bash
hadoop jar /opt/cloudera/parcels/CDH-7.1.7-1.cdh7.1.7.p0.15945976/jars/hadoop-streaming-3.1.1.7.1.7.0-551.jar \
        -Dmapred.reduce.tasks=1 \
        -input /user/feite/crime.csv \
        -output /user/feite/Crime1 \
        -file crime_map.py \
        -file crime_reduce.py \
        -mapper "python crime_map.py" \
        -reducer "python crime_reduce.py" \
```

```
[feite@ip-172-31-95-86 ~]$  hdfs dfs -cat Crime1/part-00000
CRIMINAL TRESPASS_MEDICAL/DENTAL OFFICE_true,90
INTIMIDATION_CLEANING STORE_false,1
LIQUOR LAW VIOLATION_CTA BUS_true,5
INTERFERENCE WITH PUBLIC OFFICER_CHURCH/SYNAGOGUE/PLACE OF WORSHIP_false,1
MOTOR VEHICLE THEFT_TAXICAB_true,5
WEAPONS VIOLATION_OTHER COMMERCIAL TRANSPORTATION_false,2
PUBLIC INDECENCY_ATHLETIC CLUB_true,1
HUMAN TRAFFICKING_BAR OR TAVERN_false,1
KIDNAPPING_TAXICAB_false,10
KIDNAPPING_APARTMENT_true,95
OFFENSE INVOLVING CHILDREN_OTHER RAILROAD PROPERTY / TRAIN DEPOT_false,4
CRIMINAL SEXUAL ASSAULT_CTA BUS STOP_false,4
ARSON_CHURCH/SYNAGOGUE/PLACE OF WORSHIP_true,8
DECEPTIVE PRACTICE_RESIDENCE_true,1417
SEX OFFENSE_SCHOOL - PRIVATE GROUNDS_false,2
INTERFERENCE WITH PUBLIC OFFICER_POLICE FACILITY / VEHICLE PARKING LOT_true,13
PUBLIC PEACE VIOLATION_BARBERSHOP_true,13
KIDNAPPING_POLICE FACILITY / VEHICLE PARKING LOT_false,2
MOTOR VEHICLE THEFT_STREET_false,224258
ARSON_APPLIANCE STORE_false,2
NARCOTICS_TAVERN/LIQUOR STORE_false,4
ROBBERY_ANIMAL HOSPITAL_false,9
CRIMINAL TRESPASS_AIRPORT TERMINAL UPPER LEVEL - SECURE AREA_true,51
WEAPONS VIOLATION_CEMETARY_true,8
STALKING_MEDICAL/DENTAL OFFICE_false,3
```

Code to find the total crimes in each month for past twenty-one years:

```
[jerry.lu@ip-172-31-93-164 ~]$ cut -d"," -f3 crimes_2001_2021.csv | cut -d" " -f
1 | cut -d"/" -f1 | sort | uniq -c | sort -rn | head -13
 694433 07
 687589 08
 662294 05
 660394 06
 652371 10
 645390 09
 610708 03
 608688 04
 585696 11
 584937 01
 542314 12
 513724 02
      1 Date
```

Code to group data by crime description:

```
[jerry.lu@ip-172-31-93-164 ~]$ cut -d"," -f7 crimes_2001_2021.csv | sort | uniq
-c | sort -rn | head -30
```

Code to find total number of crimes occurred on each block:

```
[jerry.lu@ip-172-31-93-164 ~]$ cut -d"," -f4 crimes_2001_2021.csv | sort | uniq
-c | sort -rn | head -20 > top_block.csv
```

Code to group data by location's latitude and longitude:

```
[jerry.lu@ip-172-31-93-164 ~]$ cut -d"," -f22,23 crimes_2001_2021.csv | sort | u
niq -c | sort -rn | tail -n+2 > location.csv
```

Code to generate smaller data frames for analysis and visualization:

df_arrest_eda = sqldf('SELECT Arrest, count(Arrest) as total from df group by Arrest')

df_arrest_eda.to_excel('df_arrest_eda.xlsx')

#MAP REDUCE

df_type_eda = sqldf('SELECT Year, Primary_Type, count(Primary_Type) from df group by Year, Primary Type')

df_type_eda.to_excel('df_type_eda.xlsx')

df_dist=sqldf('SELECT Year,Location_Description, Primary_Type, count(ID) from df group by Location_Description, Primary_Type')

df_dist.to_excel('df_dist.xlsx')

df_arrest_t = sqldf('SELECT Year, Arrest, count(ID) from df group by Year, Arrest')

df_arrest_t.to_excel('df_arrest_t.xlsx')

df_loc_prim = sqldf('SELECT Year, Location_Description, Primary_Type, count(ID) from df group by Year, Location_Description, Primary_Type')

df_loc_prim.to_excel('df_loc_prim.xlsx')

df_timeline = sqldf('SELECT Year, count(ID) as crimes_reported from df group by Year')

df_timeline.to_csv('df_timeline_total.csv')

Code used to generate Heatmap:

```
import pandas as pd
from folium.plugins import HeatMap
data = pd.read_csv('Crimes_2001_to_Present.csv')
data.head(10)
data['Date'] = pd.to_datetime(data.Date)
data['Updated On'] = pd.to_datetime(data['Updated On'])
data.Date = data.Date.dt.date
data['Updated On'] = data['Updated On'].dt.date
data['timeDelta'] = (data['Updated On'] - data['Date'])
data.head(10)
data = data.dropna(axis=0)
count = dict()
for i in data.index:
    try:
        count[eval(data.Location[i])] += 1
    except:
        count[eval(data.Location[i])] = 1
out = []
for key in count.keys():
    out.append([round(key[0],6), round(key[1],6), count[key]])
new = []
for i in range(len(out)):
    if out[i][2] > 500:
        new.append(out[i])
len(new)
import folium
world_map = folium.Map()
world_map
HeatMap(new).add_to(world_map)
world_map
```