Chicago's Crime Analysis Final Report

Excutive Summary

Purpose

Chicago is one of the most dangerous cities in the US and has a high frequency of crime occurrence. In this project, we aim to provide recommendations and suggestions to Chicago citizens and government on which parts are secure and better for living and on which parts should have stronger policing power by analyzing the data we found online.

Problem statement

In this project, we want to solve two problems:

1. How does the distribution of crimes in Chicago look like?
2. What are some deeper key statistics, such as arrest rates?

We could provide suggestions based on the analysis of these two problems.

Methods

We used python and pandas to do the data cleaning, Pyspark and Impala/Hive to do the data processing and analyzing, which do the counting, ranking and calculating, and Tableau to do the visialization.

Conclusion

From the data analysis, we suggest people in Chicago to live far away from the center because those areas are safer. We also find some areas of which the security could be improved by strengthening the policing power. In addition, we stated the further improvement of data and this project.

Introduction

Although there is no universally accepted definition of crime in modern criminal law, when it comes to crime, people always think of robbery, rape, murder and other acts that cause real harm to people or society. In order to avoid being hurt, people will always choose to live in a relatively safe area.

Chicago is one of the most dangerous cities in the United States, ranking 31th in the list provided by CBC news (2020, Feldstadt). This city has a crime rate higher than average crime rate of the United States, especially the violent crime rate. Shooting events occur frequently. However, Chicago is one of the largest cities in the US. It has a large population, and it is a very famous city among the world. For these reasons, it is necessary to figure out the relatively more secure and more suitable living areas in Chicago by, for example, analyzing the distribution of crimes in Chicago. Also, it will be helpful to find out the areas that have poor policing power that can be strengthened by calculating the arrest rates.

In this project, we aim to answer the following problems:

1. How does the distribution of crimes in Chicago look like? Including:
    - Crime Cases Count over years
    - Top 5 Types of Crimes by Police District
    - Top 5 Types of Crimes by Community Area
2. What are some deeper key statistics, such as arrest rates?

By solving these questions, we expect to provide living recommendations to people in Chicago and suggestions to the Chicago police department.

Data Description

The dataset is found on Data.gov, which has a title called “Crimes - 2001 to Present” and the link is https://catalog.data.gov/dataset/crimes-2001-to-present. It’s structured and we downloaded it as a CSV format. It collects reported incidents of crime from 2001 to Dec 2, 2022, collected by the Chicago Police Department and extracted by the Citizen Law Enforcement Analysis and Reporting system.

It is big data because:

- Volume: Its size is 1.81 GB with 22 columns and 7684983 rows,
- Variety: It consists of various types of data, including integer, string, boolean and timestamp.
- Velocity: We cannot use traditional tool to handle this dataset.
- Veracity: This dataset is collected by government agencies.

The 22 columns include ID, case number, date, block, IUCR, primary type, description, location description, arrest, domestic, beat, district, ward, community area, FBI code, X coordinate, Y coordinate, year, updated on, latitude, longitude and location.

Technically, for each crime recorded, it’s supposed to have its unique ID and case number in the dataset. The column of date marks when it occurred. The block records the crime happened in which block in Chicago. IUCR (Illinois Uniform Crime Reporting) means the corresponding code of the criminal incident. Primary type indicates what kind of crime it is, and description shows detail of crime. Arrest and domestic (boolean) is to note whether the criminal got arrested and whether it was a domestic case. For the criminal place in different uses, beat is the number of territory where police officers patrol, district is the unique code of police district in Chicago. Ward (City Council District) represents the number of Ward Zones. Community area means the unique number of each area defined by the government of Chicago. The division of community area shows in Figure 5 in Appendix. FBI code marks the number which the type of crime corresponded to in the Uniform Crime Reporting Handbook. Year just means which year the crime happened, and Updated on records when the data last updated in the system. X and Y coordinates portray the location where the incident occurred in the State Plane Illinois East NAD 1983 projection. The columns of latitude and longitude just mark the latitude and longitude of the criminal place. Location is the combination of latitude and longitude (2022, Chicago Police Department).

Due to its big size, traditional tools, such as Excel, can't handle it, and we have to use other big data tools like PySpark, Hive and Impala.

Data Preparing

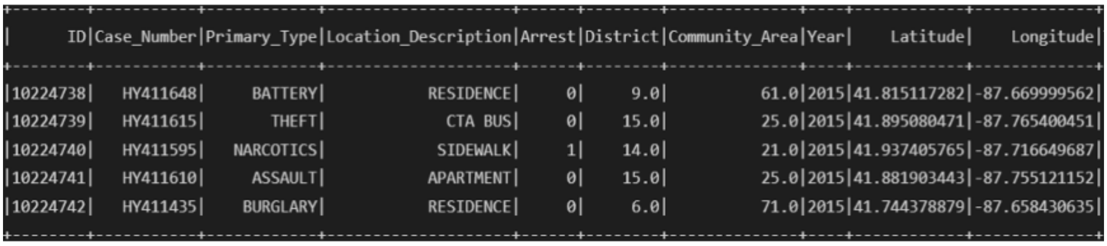
In order to make the data look cleaner and easier to analyze, we did a cleaning process before analyzing the data. The tool we use in this part is Python.

First of all, we used pandas to read the CSV data. Secondly, we dropped the column which are meaningless and will not be used in our analysis. The dropped columns include "Beat", "Ward", "IUCR", "FBI Code", "Domestic", "Description", "Y Coordinate", "X Coordinate", "Updated On", "Location", and "Block". After that, we manipulate dropna() code to drop NA values. After this, the total row number is now 6992709 rows.

To make it easier for us to do the following process, we replaced the True and False values in the Arrest column to Dummy values 1 and 0, and we replaced the spaces in the column names to underscores. Because we only need to use year rather than date, so we delete the Date column finally.

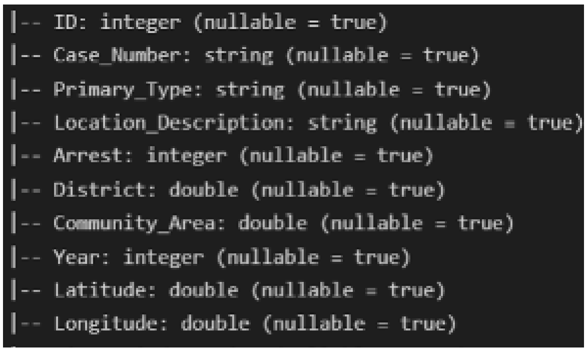
After cleaning the data, the data size reduced from 1.81GB to 625MB. The following figures are the overview and schema of the updated data:

Figure 1.



```
+--------+-----------+------------+--------------------+------+--------+--------------+----+------------+-------------+
|      ID|Case_Number|Primary_Type|Location_Description|Arrest|District|Community_Area|Year|    Latitude|    Longitude|
+--------+-----------+------------+--------------------+------+--------+--------------+----+------------+-------------+
|10224738|   HY411648|     BATTERY|           RESIDENCE|     0|     9.0|          61.0|2015|41.815117282|-87.669999562|
|10224739|   HY411615|       THEFT|             CTA BUS|     0|    15.0|          25.0|2015|41.895080471|-87.765400451|
|10224740|   HY411595|   NARCOTICS|            SIDEWALK|     1|    14.0|          21.0|2015|41.937405765|-87.716649687|
|10224741|   HY411610|     ASSAULT|           APARTMENT|     0|    15.0|          25.0|2015|41.881903443|-87.755121152|
|10224742|   HY411435|    BURGLARY|           RESIDENCE|     0|     6.0|          71.0|2015|41.744378879|-87.658430635|
+--------+-----------+------------+--------------------+------+--------+--------------+----+------------+-------------+
```

Figure 2.



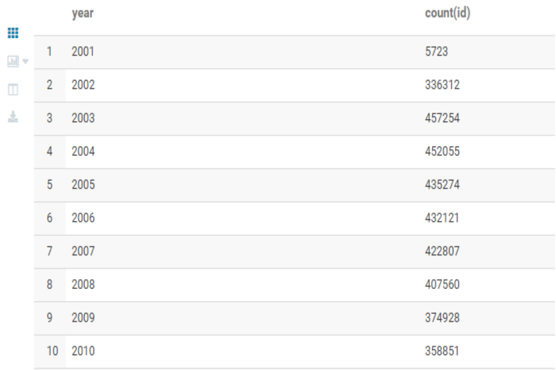
```
|-- ID: integer (nullable = true)
|-- Case_Number: string (nullable = true)
|-- Primary_Type: string (nullable = true)
|-- Location_Description: string (nullable = true)
|-- Arrest: integer (nullable = true)
|-- District: double (nullable = true)
|-- Community_Area: double (nullable = true)
|-- Year: integer (nullable = true)
|-- Latitude: double (nullable = true)
|-- Longitude: double (nullable = true)
```

Methods and Results

At the beginning, we uploaded the updated dataset to HDFS so that we can apply fast big data analysis to it in Hive and Impala on our server.

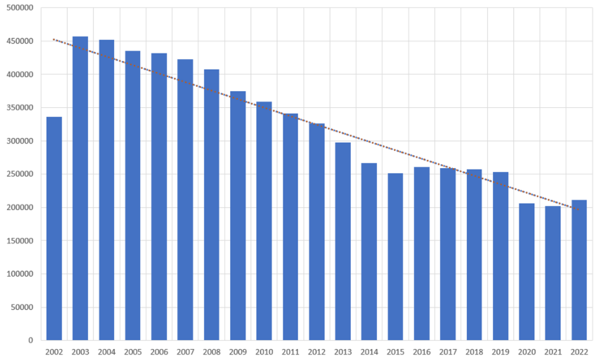
In Hive, we count the recorded crime events for each recorded year. Part of the results are shown in Figure 3. To show the results more visualized, we used Tableau to draw a histogram in Figure 4. According to the histogram, we can conclude that the number of cases is decreasing over years.

Figure 3.



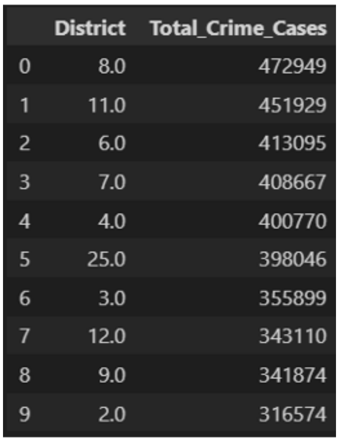
	year	count(id)
1	2001	5723
2	2002	336312
3	2003	457254
4	2004	452055
5	2005	435274
6	2006	432121
7	2007	422807
8	2008	407560
9	2009	374928
10	2010	358851

Figure 4.



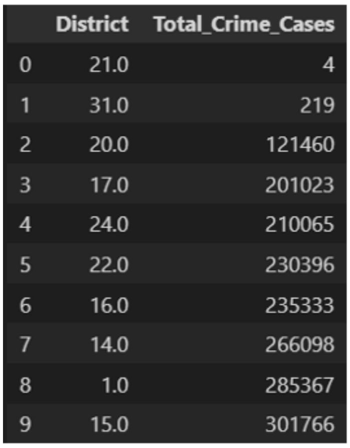
We want to see the ranking of the crime cases count by police district and by community area. We rank it in two ways because police district and community are two different partition criteria. We use district which divided by police department to analyze to policing power and use the community area to find safe place for living. To achieve these results, we used pyspark in jupyter notebook. For both partition criteria, we have both descending and ascending ranking orders. The results are shown below:

Police district descending ranking:



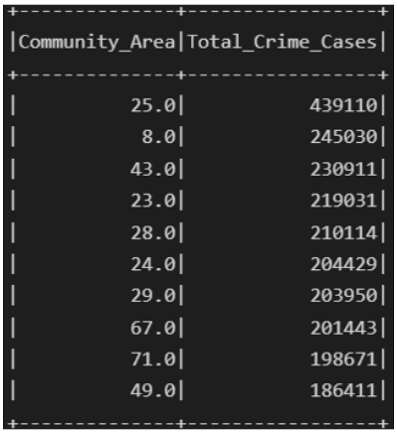
	District	Total_Crime_Cases
0	8.0	472949
1	11.0	451929
2	6.0	413095
3	7.0	408667
4	4.0	400770
5	25.0	398046
6	3.0	355899
7	12.0	343110
8	9.0	341874
9	2.0	316574

police district ascending ranking:



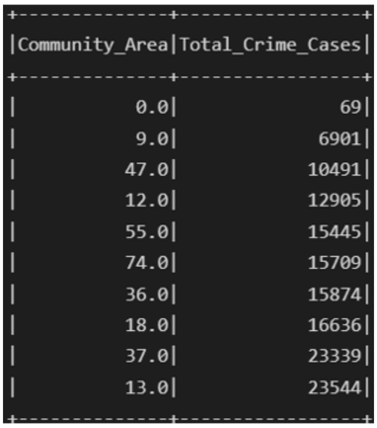
	District	Total_Crime_Cases
0	21.0	4
1	31.0	219
2	20.0	121460
3	17.0	201023
4	24.0	210065
5	22.0	230396
6	16.0	235333
7	14.0	266098
8	1.0	285367
9	15.0	301766

Community area descending ranking:



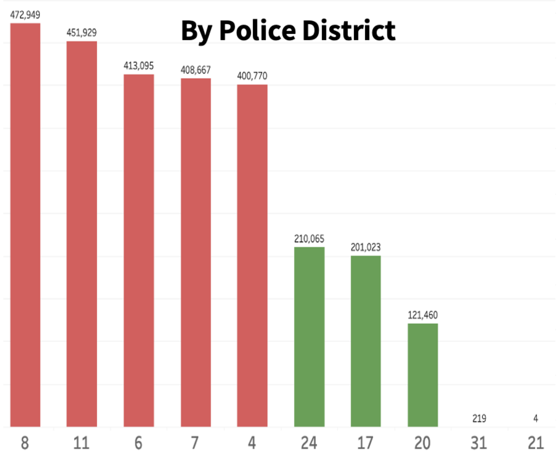
```
+--------------+-----------------+
|Community_Area|Total_Crime_Cases|
+--------------+-----------------+
|          25.0|           439110|
|           8.0|           245030|
|          43.0|           230911|
|          23.0|           219031|
|          28.0|           210114|
|          24.0|           204429|
|          29.0|           203950|
|          67.0|           201443|
|          71.0|           198671|
|          49.0|           186411|
+--------------+-----------------+
```

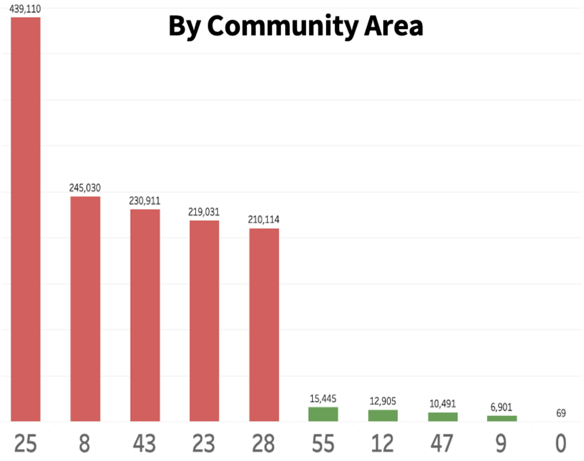
Community area ascending ranking:



```
+--------------+-----------------+
|Community_Area|Total_Crime_Cases|
+--------------+-----------------+
|           0.0|               69|
|           9.0|             6901|
|          47.0|            10491|
|          12.0|            12905|
|          55.0|            15445|
|          74.0|            15709|
|          36.0|            15874|
|          18.0|            16636|
|          37.0|            23339|
|          13.0|            23544|
+--------------+-----------------+
```

The visualization of the results are:

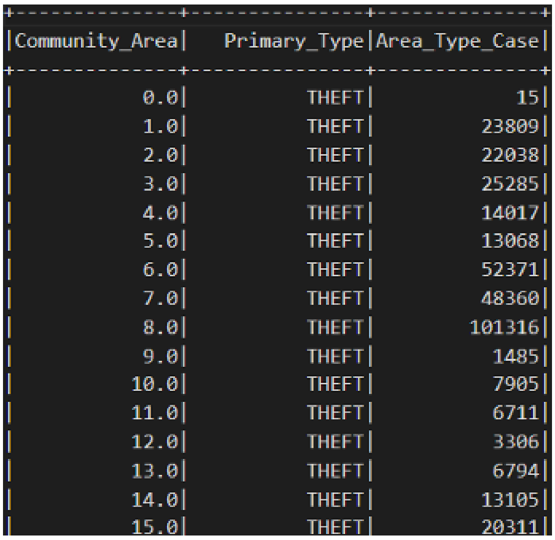




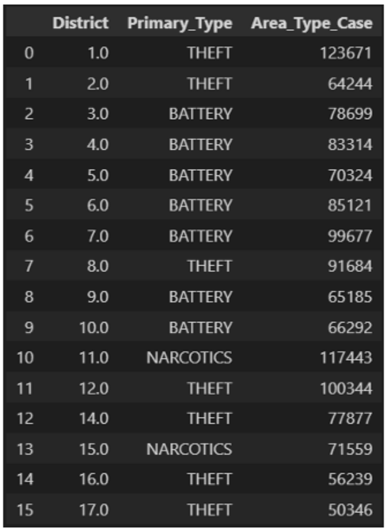
The red bars shows the five areas that have the most crime cases and green bars represent the five areas have the least crime cases.

We also care about the crime type in the whole Chicago. We used pyspark to find the most crime type for each police district and each community area.

Most crime type of police district: **Most crime type of community area:**

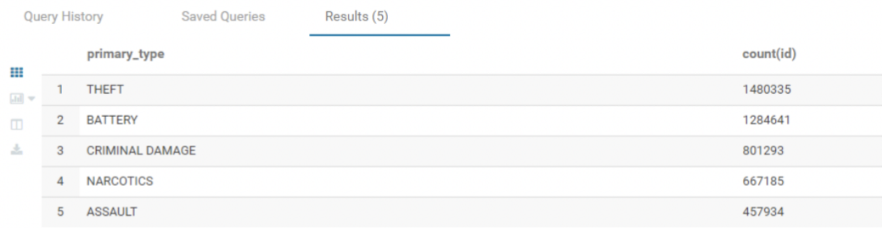


```
+--------------+---------------+-------------+
|Community_Area|   Primary_Type|Area_Type_Case|
+--------------+---------------+-------------+
|           0.0|          THEFT|           15|
|           1.0|          THEFT|        23809|
|           2.0|          THEFT|        22038|
|           3.0|          THEFT|        25285|
|           4.0|          THEFT|        14017|
|           5.0|          THEFT|        13068|
|           6.0|          THEFT|        52371|
|           7.0|          THEFT|        48360|
|           8.0|          THEFT|       101316|
|           9.0|          THEFT|         1485|
|          10.0|          THEFT|         7905|
|          11.0|          THEFT|         6711|
|          12.0|          THEFT|         3306|
|          13.0|          THEFT|         6794|
|          14.0|          THEFT|        13105|
|          15.0|          THEFT|        20311|
```



	District	Primary_Type	Area_Type_Case
0	1.0	THEFT	123671
1	2.0	THEFT	64244
2	3.0	BATTERY	78699
3	4.0	BATTERY	83314
4	5.0	BATTERY	70324
5	6.0	BATTERY	85121
6	7.0	BATTERY	99677
7	8.0	THEFT	91684
8	9.0	BATTERY	65185
9	10.0	BATTERY	66292
10	11.0	NARCOTICS	117443
11	12.0	THEFT	100344
12	14.0	THEFT	77877
13	15.0	NARCOTICS	71559
14	16.0	THEFT	56239
15	17.0	THEFT	50346

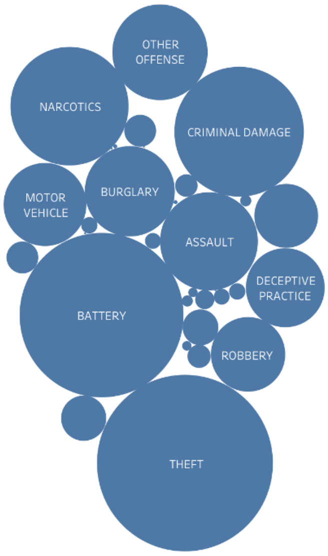
The top 5 types of crimes are found by using Impala. The results are:



	primary_type	count(id)
1	THEFT	1480335
2	BATTERY	1284641
3	CRIMINAL DAMAGE	801293
4	NARCOTICS	667185
5	ASSAULT	457934

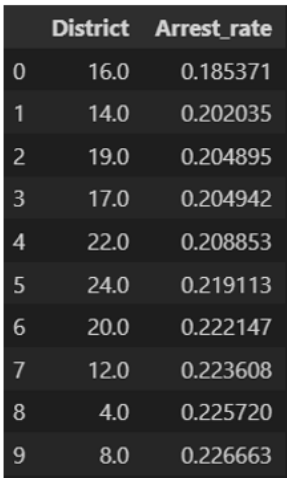
The visualization of the most common crime types is:

The most commen Primary Type in Chicago



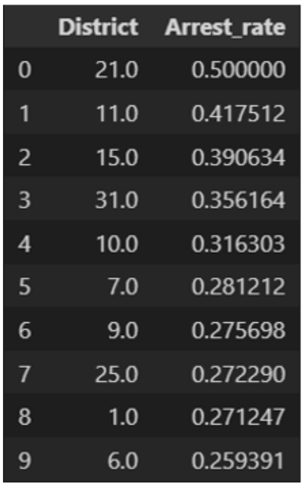
Moreover, we want to calculate the arrest rates for each police district and community area. The definition of arrest rate is dividing the amount of cases that criminal was arrested by total amount of cases. To do this, we used pyspark. The results are:

Police district descending ranking:



	District	Arrest_rate
0	16.0	0.185371
1	14.0	0.202035
2	19.0	0.204895
3	17.0	0.204942
4	22.0	0.208853
5	24.0	0.219113
6	20.0	0.222147
7	12.0	0.223608
8	4.0	0.225720
9	8.0	0.226663

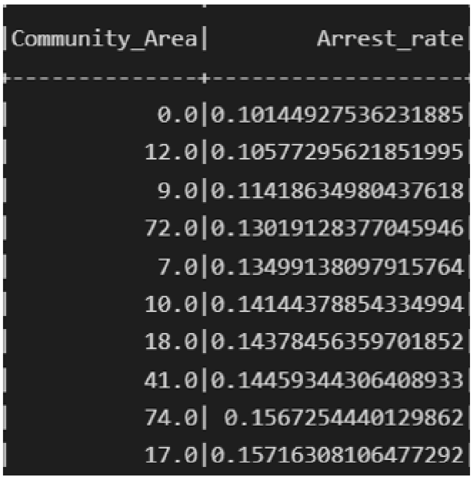
Police district ascending ranking:



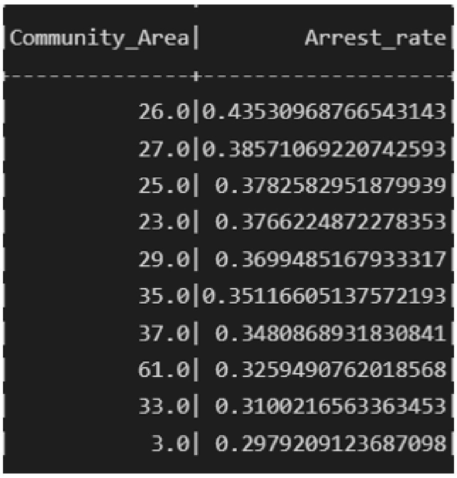
	District	Arrest_rate
0	21.0	0.500000
1	11.0	0.417512
2	15.0	0.390634
3	31.0	0.356164
4	10.0	0.316303
5	7.0	0.281212
6	9.0	0.275698
7	25.0	0.272290
8	1.0	0.271247
9	6.0	0.259391

Community area descending ranking:

Community area ascending ranking:

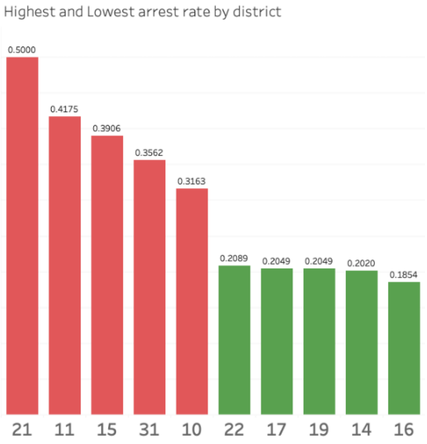


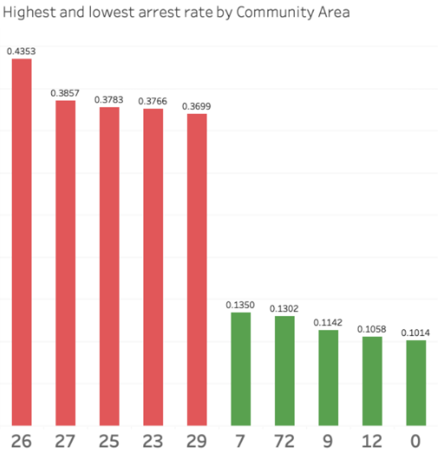
```
|Community_Area|        Arrest_rate|
+--------------+-------------------+
|           0.0|0.10144927536231885|
|          12.0|0.10577295621851995|
|           9.0|0.11418634980437618|
|          72.0|0.13019128377045946|
|           7.0|0.13499138097915764|
|          10.0|0.14144378854334994|
|          18.0|0.14378456359701852|
|          41.0|0.14459344306408933|
|          74.0| 0.1567254440129862|
|          17.0|0.15716308106477292|
```



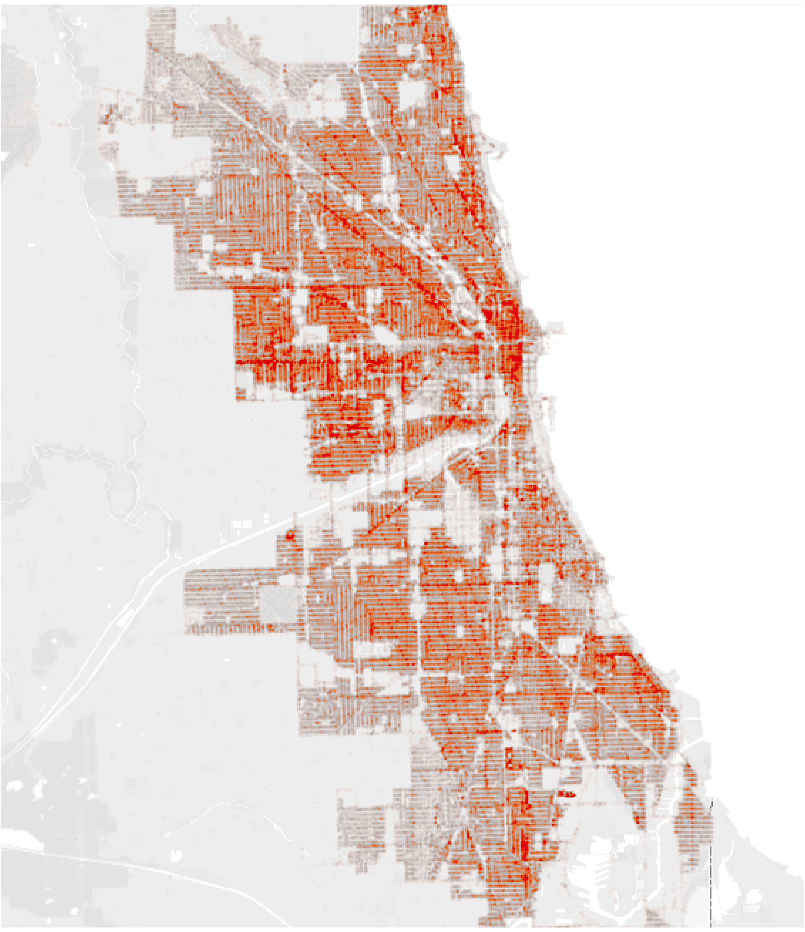
```
|Community_Area|        Arrest_rate|
+--------------+-------------------+
|          26.0|0.43530968766543143|
|          27.0|0.38571069220742593|
|          25.0| 0.3782582951879939|
|          23.0| 0.3766224872278353|
|          29.0| 0.3699485167933317|
|          35.0|0.35116605137572193|
|          37.0| 0.3480868931830841|
|          61.0| 0.3259490762018568|
|          33.0| 0.3100216563363453|
|           3.0| 0.2979209123687098|
```

The visualizations of the results are:





Finally, we used Tableau drawing a heat map to show a whole view of distribution of crime cases amount across Chicago. The heavier red shows the more crime cases the region has.



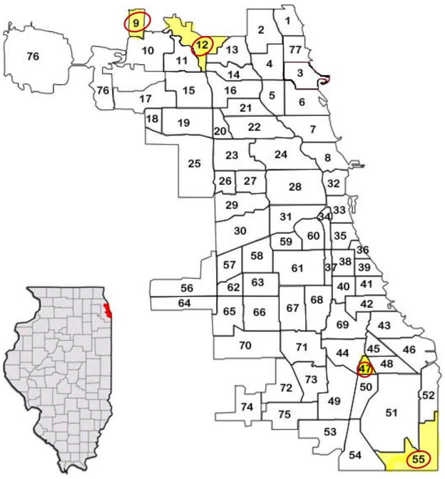
Conclusion

Most of them are in the margin area in Chicago. It indicates that the farther from the center of Chicago, the safer it is.

The center of Chicago, according to heat map and crime amount data, is much more dangerous than the marginal area. Therefore, we recommend that citizens live far away from the center.

What's more, we also find some district areas that can be better if the government enhances the armed police force in such areas. These are:

- 55: Hegewisch
- 12: Forest Glen
- 47: Burnside
- 9: Edison Park



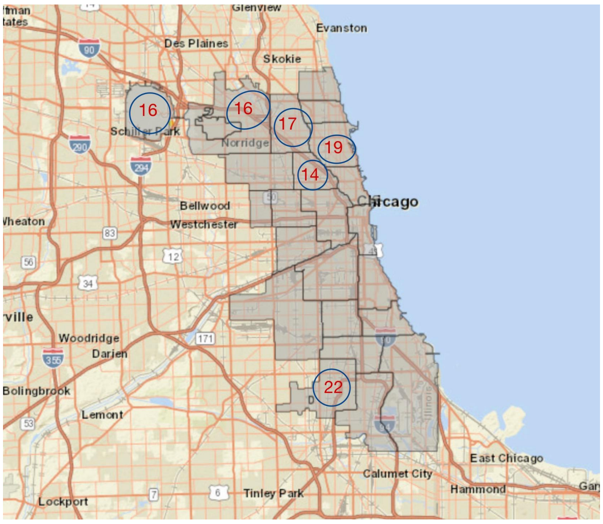
resource:https://commons.wikimedia.org/wiki/File:Blank_Chicago_Community_Area_Map.png

Most of them are in the margin area in Chicago. It indicates that the farther from the center of Chicago, the safer it is.

The center of Chicago, according to heat map and crime amount data, is much more dangerous than the marginal area. Therefore, we recommend that citizens live far away from the center.

What's more, we also find some district areas that can be better if the government enhances the armed police force in such areas. These are:

- 16
- 17
- 19
- 22



resource:
https://data.cityofchicago.org/Public-Safety/Boundaries-Police-Districts-current-/fthy-xz3r

The reason for choosing these district areas is, after we do the calculation, these five district areas have the lowest arrest rate. Hence, we hope that the police force can be enhanced in these areas to provide a safer environment for the citizens.

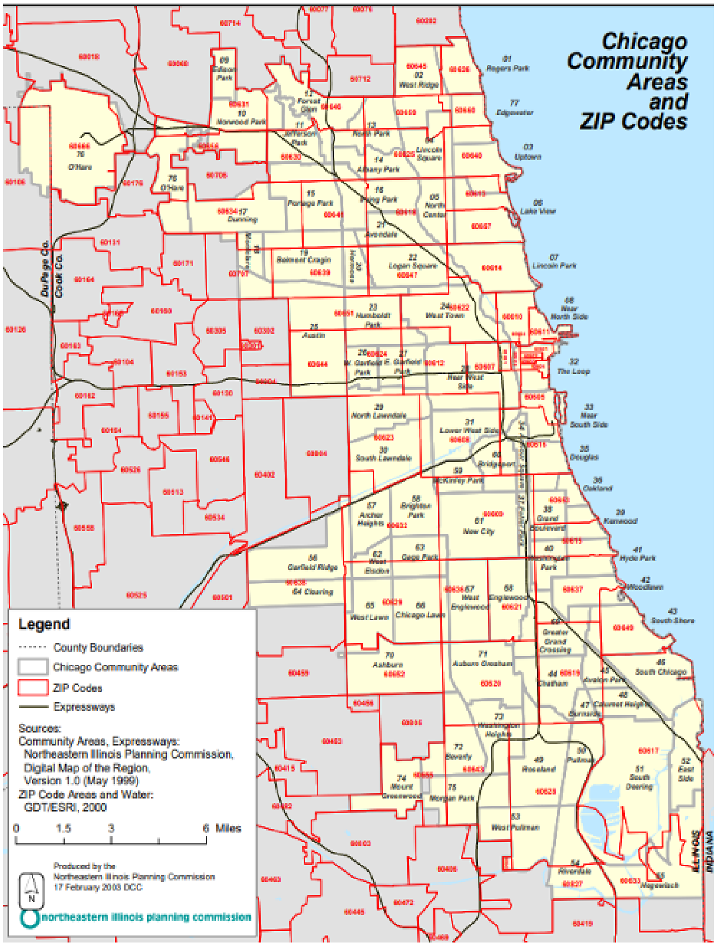
Overall, our team hoped that our findings could help more Chicago citizens live in a safer environment, and less suffering from kinds of crime, such as theft, robbery, and so on.

For the data part, although the size of district and community areas is divided by the government, the size of each area is not equal, therefore, this may cause data bias. For example, community area 17 is one of the community areas that has the least crime amount. However, it may be because it has a smaller size. If we do the calculation, for example, the criminal case amount divided by community area, the result could be different. We need to find a better solution for this data bias in the future.

What's more the website shows that the data are not guaranteed accuracy, which limits our findings. For example, there are many blank places in the crime amount contributed heat map. It may not indicate that there is no crime in this area at all. It is because of a lack of data in the system. In order to improve our study, the police department needs to provide more detailed and comprehensive data. In order to have a comprehensive and useful crime dataset. We need a team that can frequently check the data accuracy and inspect each department to update the data in the system. Crime data is always important for a city, especially for such a dangerous city. A perfect and comprehensive crime dataset can always do more than you think.

Appendix

Figure 5.



Legend

County Boundaries

Chicago Community Areas

ZIP Codes

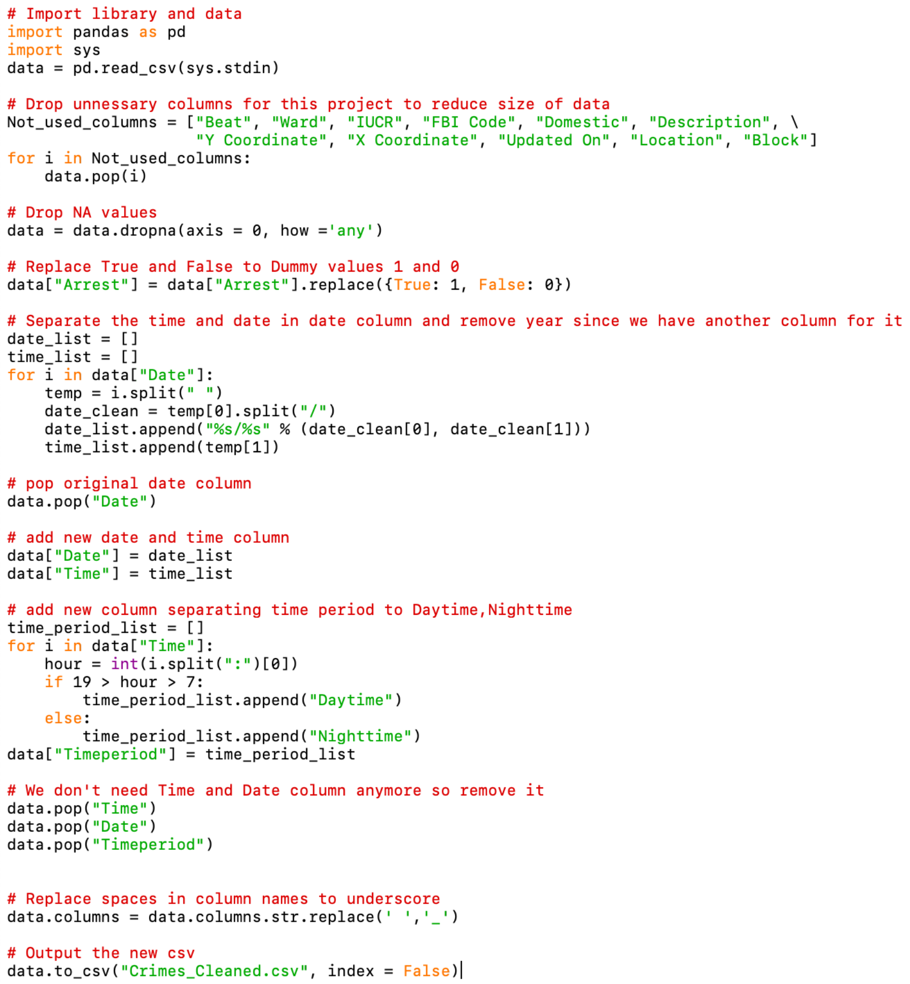
Expressways

Sources:
Community Areas, Expressways:
Northeastern Illinois Planning Commission,
Digital Map of the Region,
Version 1.0 (May 1999)
ZIP Code Areas and Water:
GDT/ESRI, 2000

Produced by the
Northeastern Illinois Planning Commission
17 February 2003 DCC

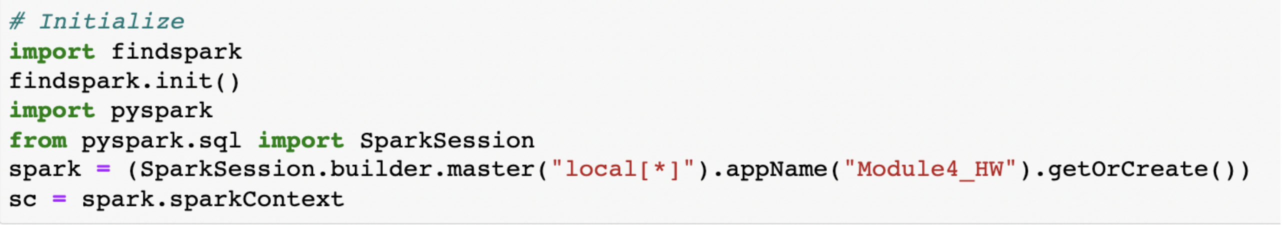
northeastern illinois planning commission

Python

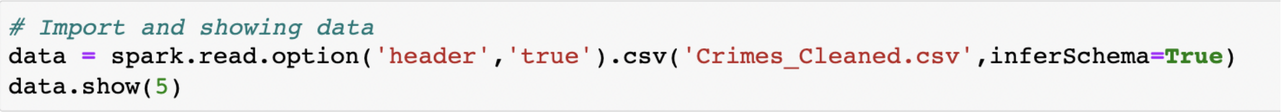


```
# Import library and data
import pandas as pd
import sys
data = pd.read_csv(sys.stdin)

# Drop unnessary columns for this project to reduce size of data
Not_used_columns = ["Beat", "Ward", "IUCR", "FBI Code", "Domestic", "Description", \
                    "Y Coordinate", "X Coordinate", "Updated On", "Location", "Block"]
for i in Not_used_columns:
    data.pop(i)

# Drop NA values
data = data.dropna(axis = 0, how ='any')

# Replace True and False to Dummy values 1 and 0
data["Arrest"] = data["Arrest"].replace({True: 1, False: 0})

# Separate the time and date in date column and remove year since we have another column for it
date_list = []
time_list = []
for i in data["Date"]:
    temp = i.split(" ")
    date_clean = temp[0].split("/")
    date_list.append("%s/%s" % (date_clean[0], date_clean[1]))
    time_list.append(temp[1])

# pop original date column
data.pop("Date")

# add new date and time column
data["Date"] = date_list
data["Time"] = time_list

# add new column separating time period to Daytime,Nighttime
time_period_list = []
for i in data["Time"]:
    hour = int(i.split(":")[0])
    if 19 > hour > 7:
        time_period_list.append("Daytime")
    else:
        time_period_list.append("Nighttime")
data["Timeperiod"] = time_period_list

# We don't need Time and Date column anymore so remove it
data.pop("Time")
data.pop("Date")
data.pop("Timeperiod")


# Replace spaces in column names to underscore
data.columns = data.columns.str.replace(' ','_')

# Output the new csv
data.to_csv("Crimes_Cleaned.csv", index = False)
```

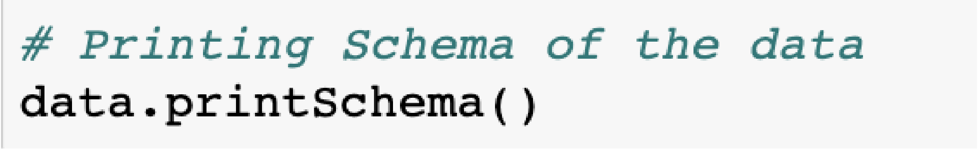
Pyspark



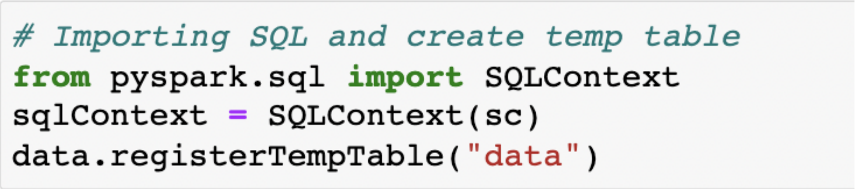
```
# Initialize
import findspark
findspark.init()
import pyspark
from pyspark.sql import SparkSession
spark = (SparkSession.builder.master("local[*]").appName("Module4_HW").getOrCreate())
sc = spark.sparkContext
```



```
# Import and showing data
data = spark.read.option('header','true').csv('Crimes_Cleaned.csv',inferSchema=True)
data.show(5)
```

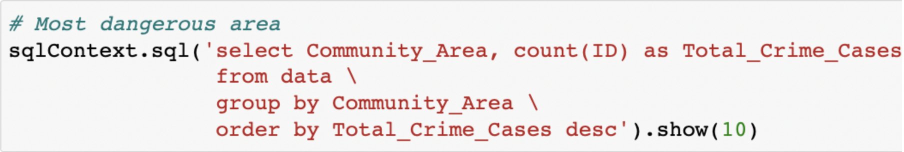


```
# Printing Schema of the data
data.printSchema()
```

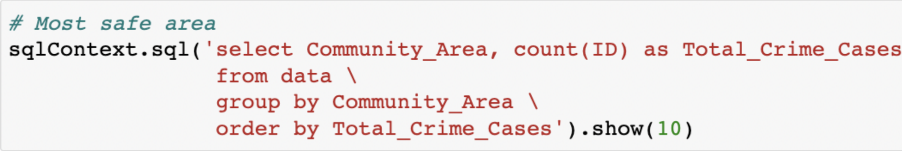


```
# Importing SQL and create temp table
from pyspark.sql import SQLContext
sqlContext = SQLContext(sc)
data.registerTempTable("data")
```

Ranking crime cases count by Community Area

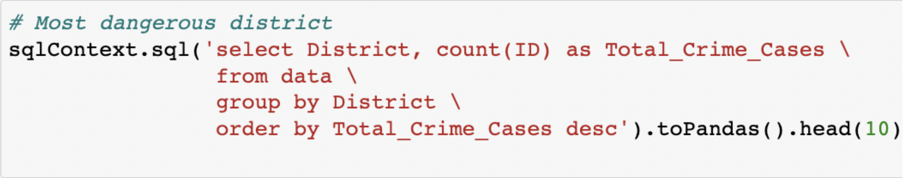


```
# Most dangerous area
sqlContext.sql('select Community_Area, count(ID) as Total_Crime_Cases
                from data \
                group by Community_Area \
                order by Total_Crime_Cases desc').show(10)
```

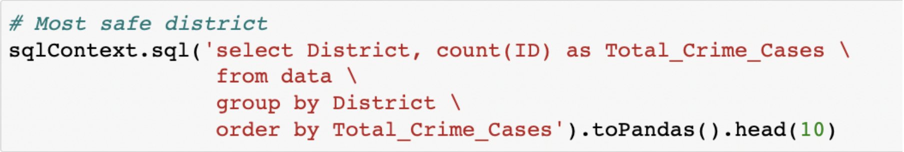


```
# Most safe area
sqlContext.sql('select Community_Area, count(ID) as Total_Crime_Cases
                from data \
                group by Community_Area \
                order by Total_Crime_Cases').show(10)
```

Ranking crime cases count by District ¶

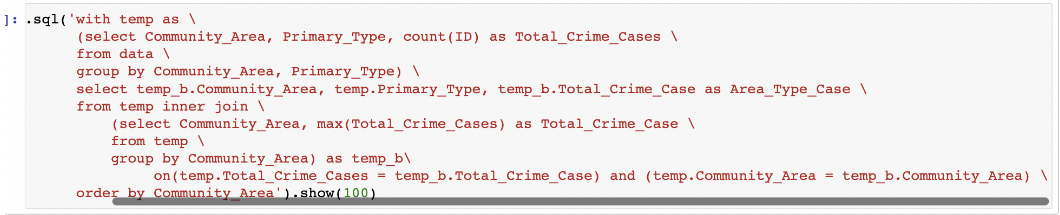


```
# Most dangerous district
sqlContext.sql('select District, count(ID) as Total_Crime_Cases \
                from data \
                group by District \
                order by Total_Crime_Cases desc').toPandas().head(10)
```



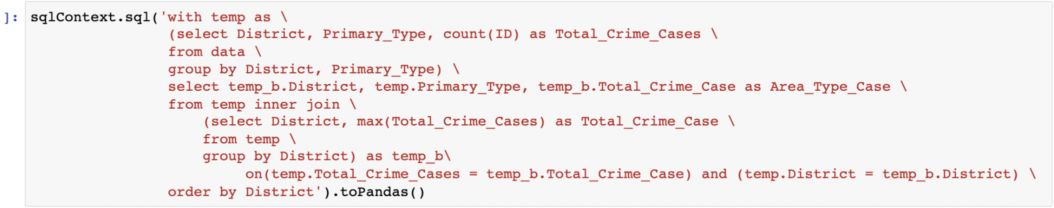
```
# Most safe district
sqlContext.sql('select District, count(ID) as Total_Crime_Cases \
                from data \
                group by District \
                order by Total_Crime_Cases').toPandas().head(10)
```

Ranking Crime Type by Community Area

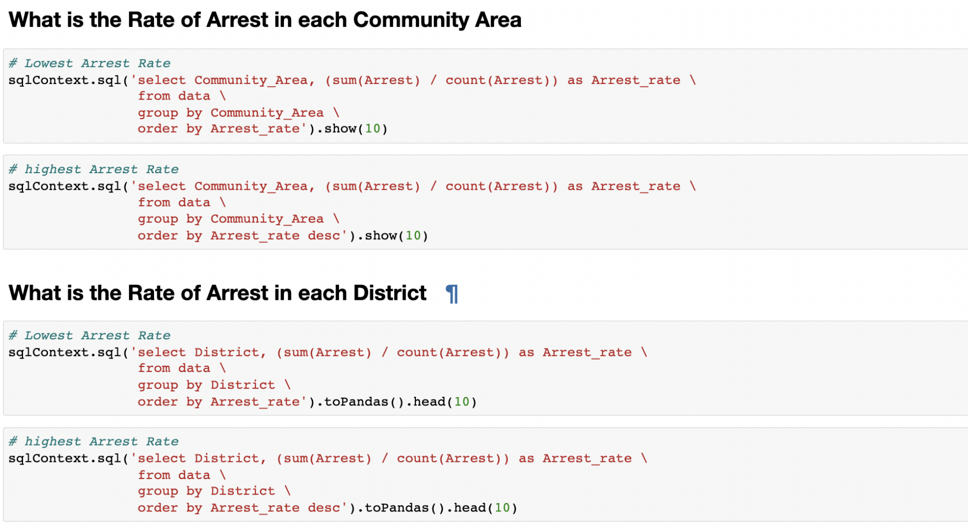


```
]: .sql('with temp as \
         (select Community_Area, Primary_Type, count(ID) as Total_Crime_Cases \
         from data \
         group by Community_Area, Primary_Type) \
         select temp_b.Community_Area, temp.Primary_Type, temp_b.Total_Crime_Case as Area_Type_Case \
         from temp inner join \
             (select Community_Area, max(Total_Crime_Cases) as Total_Crime_Case \
             from temp \
             group by Community_Area) as temp_b\
                  on(temp.Total_Crime_Cases = temp_b.Total_Crime_Case) and (temp.Community_Area = temp_b.Community_Area) \
         order by Community_Area').show(100)
```

Ranking Crime Type by District



```
]: sqlContext.sql('with temp as \
                   (select District, Primary_Type, count(ID) as Total_Crime_Cases \
                   from data \
                   group by District, Primary_Type) \
                   select temp_b.District, temp.Primary_Type, temp_b.Total_Crime_Case as Area_Type_Case \
                   from temp inner join \
                       (select District, max(Total_Crime_Cases) as Total_Crime_Case \
                       from temp \
                       group by District) as temp_b\
                            on(temp.Total_Crime_Cases = temp_b.Total_Crime_Case) and (temp.District = temp_b.District) \
                   order by District').toPandas()
```



What is the Rate of Arrest in each Community Area

```
# Lowest Arrest Rate
sqlContext.sql('select Community_Area, (sum(Arrest) / count(Arrest)) as Arrest_rate \
                from data \
                group by Community_Area \
                order by Arrest_rate').show(10)
```

```
# highest Arrest Rate
sqlContext.sql('select Community_Area, (sum(Arrest) / count(Arrest)) as Arrest_rate \
                from data \
                group by Community_Area \
                order by Arrest_rate desc').show(10)
```

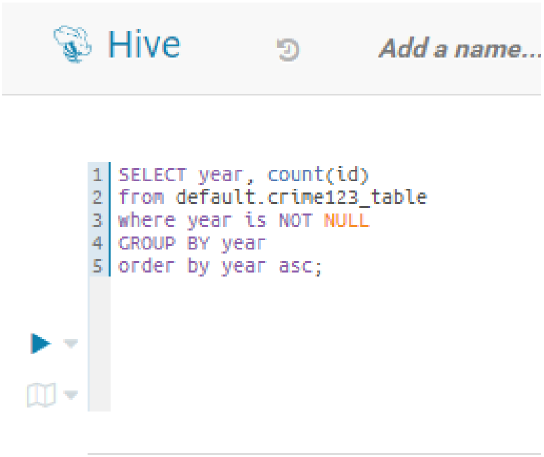
What is the Rate of Arrest in each District ¶

```
# Lowest Arrest Rate
sqlContext.sql('select District, (sum(Arrest) / count(Arrest)) as Arrest_rate \
                from data \
                group by District \
                order by Arrest_rate').toPandas().head(10)
```

```
# highest Arrest Rate
sqlContext.sql('select District, (sum(Arrest) / count(Arrest)) as Arrest_rate \
                from data \
                group by District \
                order by Arrest_rate desc').toPandas().head(10)
```

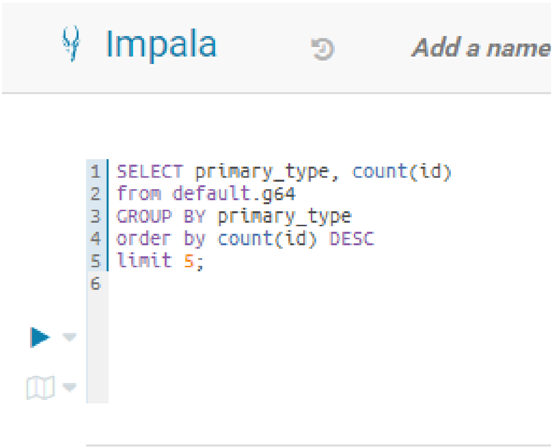
Hive/Impala

Total Cases by Year



```
SELECT year, count(id)
from default.crime123_table
where year is NOT NULL
GROUP BY year
order by year asc;
```

Top 5 Most types of Crimes



```
SELECT primary_type, count(id)
from default.g64
GROUP BY primary_type
order by count(id) DESC
limit 5;
```

Reference

Chicago Police Department. “Crimes - 2001 to Present.” *Chicago Data Portal*, 8 December 2022, https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2. Accessed 8 December 2022.

Fieldstadt, Elisha. “The most dangerous cities in America, ranked.” *CBS News*, 9 November 2020, https://www.cbsnews.com/pictures/the-most-dangerous-cities-in-america/5/. Accessed 8 December 2022.