# FACTORS BEHIND CHICAGO CRIMES

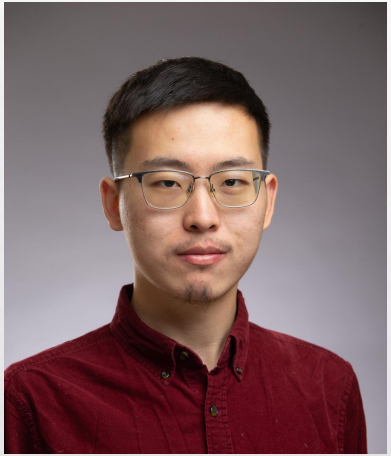
Jianan Fei, Yue Sun,
Ruobing Xue, Kaihao Fan
(Group 4: Future Splendid & X-Fabulous )

December 2020

# MEET THE TEAM

**Jianan Fei**

feij@uchicago.edu



**Kaihao Fan**

kaihaofan@uchicago.edu

**Ruobing Xue**

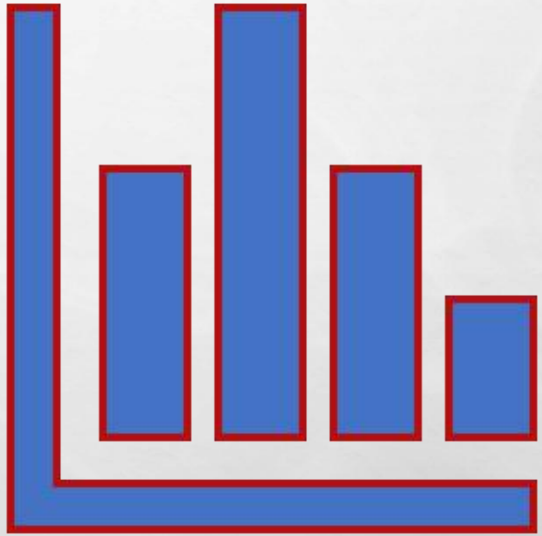
ruobing@uchicago.edu

**Yue Sun**

yus002@uchicago.edu

# CONTENT

- Executive Summary
- Business Use Case
- Sources & Tools
- Data ETL
- Relational Model EER & Dimensional Model
- Reporting
- Recommendations
- Lessons learned
- Appendix



# Executive Summary

Collecting data from various data source and merging to single data warehouse, then conducting analyzation and visualization by multiple data tools learned in Data Engineering Platform, nail down factors behind the crimes of Chicago City.

The purpose of this project is to provide valuable insights to citizens of Chicago and public service departments to relieve public safety pressure and help residents to understand trends of crimes at Chicago with tips.

# Business Use Case


**PROBLEM**
O Identify and analyze factors behind Chicago crimes

**MOTIVATION**
O Citizens of Chicago can understand which factors may influence crime rates at specific area
O Helpful to protect personal and property safety
O Public service departments can better understand hidden factors
O Implement better initiatives to enhance the security of Chicago

# Sources & Tools

## Sources

Portal: Chicago Public Safety-Crimes


External: Chicago Demographic


## Ingestion & Cleaning

Open Refine


Microsoft Excel

Python


## Warehousing

Google Cloud Platform


SQL


neo4j


## Visualization & Analytics

Tableau


Kepler


Mapbox


MongoDB


Python

# DATA Cleaning: Crime 2020

- Python:
- Drop columns that have missing values
- Rename ID to CrimeID to avoid confusion
- Convert all column names to uppercase
- Change DATE column to TimeStamp

```
crime.drop(['IUCR', 'Beat', 'Ward', 'FBI Code', 'X Coordinate', 'Y Coordinate', 'Year', 'Updated On'], axis = 1, inplace = True)
```

```
crime.shape
```

```
(163360, 14)
```

```
crime.rename(columns={"ID":"CrimeID"}, inplace = True)
crime.columns = crime.columns.str.upper()
crime.head()
```

| | CRIMEID | CASE NUMBER | DATE | BLOCK | PRIMARY TYPE | DESCRIPTION | LOCATION DESCRIPTION | ARREST | DOMESTIC | DISTRICT | COMMUNITY AREA | LATITUDE | LONGI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 167 | 25603 | JD423602 | 11/07/2020 07:40:00 PM | 054XX W CORTEZ ST | HOMICIDE | FIRST DEGREE MURDER | STREET | True | False | 15 | 25 | 41.899438 | -87.7 |
| 339 | 25598 | JD423611 | 11/07/2020 04:07:00 AM | 021XX E 72ND PL | HOMIC | | | | | | | | |
| 409 | 12178135 | JD381686 | 09/27/2020 11:00:00 PM | 068XX S RIDGELAND AVE | OTH OFFE | | | | | | | | |
| 508 | 25600 | JD424303 | 11/07/2020 06:20:00 PM | 074XX N OAKLEY AVE | HOMIC | | | | | | | | |
| 620 | 25599 | JD423568 | 11/07/2020 02:13:00 AM | 059XX W HURON ST | HOMIC | | | | | | | | |

```
crime['DATE'] = pd.to_datetime(crime['DATE'])
crime['DATE']
```

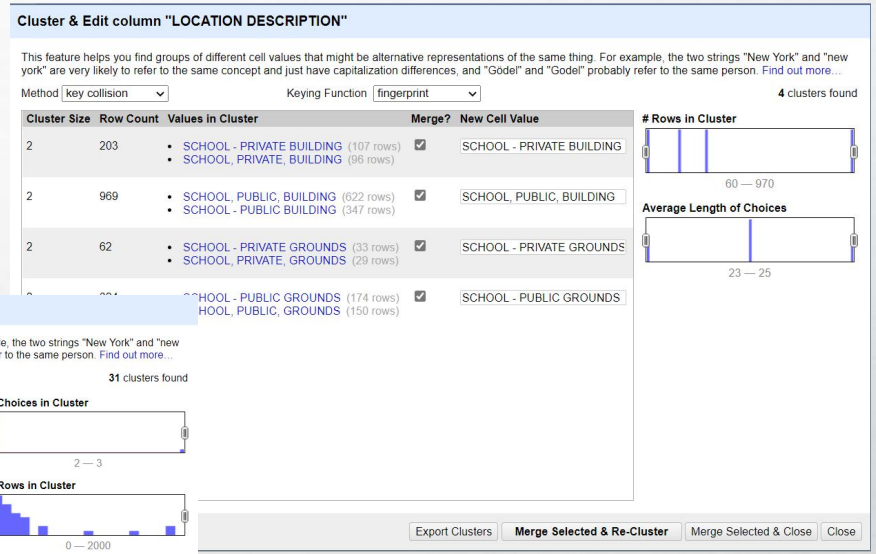
```
0         2020-11-07 19:40:00
1         2020-11-07 04:07:00
2         2020-09-27 23:00:00
3         2020-11-07 18:20:00
4         2020-11-07 02:13:00
                  ...
163355    2020-01-01 01:24:00
163356    2020-01-01 03:00:00
163357    2020-01-01 14:13:00
163358    2020-01-01 04:07:00
163359    2020-01-01 15:00:00
Name: DATE, Length: 163360, dtype: datetime64[ns]
```

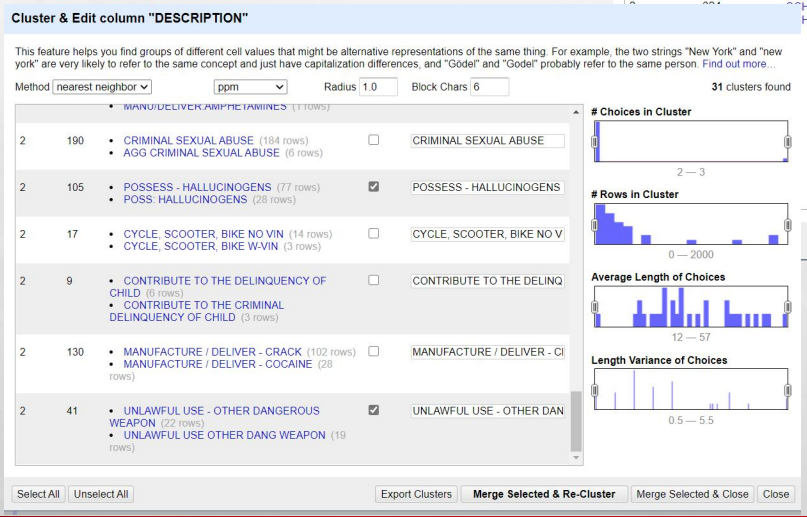
```
crime.head()
```

| | CRIMEID | CASE NUMBER | DATE | BLOCK | PRIMARY TYPE | DESCRIPTION | LOCATION DESCRIPTION | ARREST | DOMESTIC | DISTRICT | COMMUNITY AREA | LATITUDE | LONGITUDI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 25603 | JD423602 | 2020-11-07 19:40:00 | 054XX W CORTEZ ST | HOMICIDE | FIRST DEGREE MURDER | STREET | True | False | 15 | 25 | 41.899438 | -87.76251 |
| 1 | 25598 | JD423611 | 2020-11-07 04:07:00 | 021XX E 72ND PL | HOMICIDE | FIRST DEGREE MURDER | YARD | False | False | 3 | 43 | 41.763456 | -87.57333 |
| 2 | 12178135 | JD381686 | 2020-09-27 23:00:00 | 068XX S RIDGELAND AVE | OTHER OFFENSE | OTHER CRIME AGAINST PERSON | APARTMENT | False | False | 3 | 43 | 41.770931 | -87.58258 |
| 3 | 25600 | JD424303 | 2020-11-07 18:20:00 | 074XX N OAKLEY AVE | HOMICIDE | FIRST DEGREE MURDER | STREET | False | False | 24 | 2 | 42.016114 | -87.68751 |
| 4 | 25599 | JD423568 | 2020-11-07 02:13:00 | 059XX W HURON ST | HOMICIDE | FIRST DEGREE MURDER | STREET | False | False | 15 | 25 | 41.892831 | -87.77348 |

# DATA Cleaning: Crime 2020

- OpenRefine:
- Merge different cell values



# DATA Cleaning: Chicago Police Stations & Demographic

- Python:
- Change district ‘Headquarters’ to 31 (31 comes from another file with district code)
- Drop columns:['X COORDINATE', 'Y COORDINATE', 'CITY', 'STATE'] 4 columns in total
- Drop columns that have missing values
- Keep columns to be analyzed and merge some columns
- Drop other columns.

| | DISTRICT | DISTRICT NAME | ADDRESS | CITY | STATE | ZIP | WEBSITE | PHONE | FAX | TTY | X COORDINATE | Y COORDINATE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Headquarters | Headquarters | 3510 S Michigan Ave | Chicago | IL | 60653 | http://home.chicagopolice.org | NaN | NaN | NaN | 1177731.401 | 1881697.404 |
| 1 | 1 | Central | 1718 S State St | Chicago | IL | 60616 | http://home.chicagopolice.org/community/distri... | 312-745-4290 | 312-745-3694 | 312-745-3693 | 1176569.052 | 1891771.704 |
| 2 | 6 | Gresham | 7808 S Halsted St | Chicago | IL | 60620 | http://home.chi | | | | | |
| 3 | 11 | Harrison | 3151 W Harrison St | Chicago | IL | 60612 | http://home.chi | | | | | |
| 4 | 16 | Jefferson Park | 5151 N Milwaukee Ave | Chicago | IL | 60630 | http://home.chi | | | | | |

```
Poverty = poverty[['Community Number','2000 %Poverty Population','2010 %Poverty Population']]
```

```
Poverty.head(5)
```

| | Community Number | 2000 %Poverty Population | 2010 %Poverty Population |
|---|---|---|---|
| 0 | 1 | 0.21 | 0.26 |
| 1 | 2 | 0.14 | 0.17 |
| 2 | 3 | 0.25 | 0.26 |
| 3 | 4 | 0.11 | 0.12 |
| | | | 0.07 |

```
Education=education[['Community Number','%Educated with no degree','%Educated with degree']]
Education.head(5)
```

| | Community Number | %Educated with no degree | %Educated with degre |
|---|---|---|---|
| 0 | 1 | 0.39 | 0.3 |
| 1 | 2 | 0.39 | 0.2 |
| 2 | 3 | 0.34 | 0.4 |
| 3 | 4 | 0.32 | 0.4 |
| 4 | 5 | 0.21 | 0.5 |

```
Income=income[['Community Number','%Households Earning less than $49,999','%Households Earning 50,000_99,999','%Hous
```

```
Income.head(5)
```

| | Community Number | %Households Earning less than $49,999 | %Households Earning 50,000_99,999 | %Households Earning $100,000 or more |
|---|---|---|---|---|
| 0 | 1 | 0.59 | 0.25 | 0.16 |
| 1 | 2 | 0.54 | 0.28 | 0.18 |
| 2 | 3 | 0.54 | 0.25 | 0.21 |
| 3 | 4 | 0.41 | 0.35 | 0.24 |
| 4 | 5 | 0.27 | 0.29 | 0.44 |

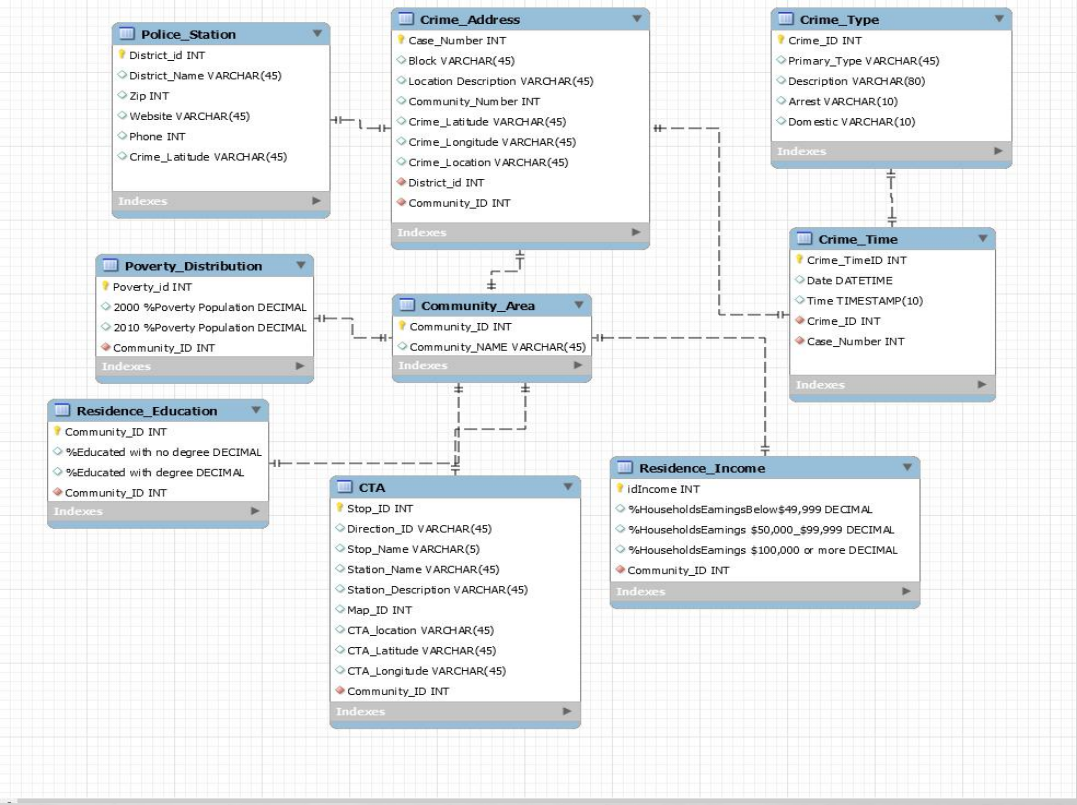
# Relational Model Normalization

- Normalize the cleaned Crime table

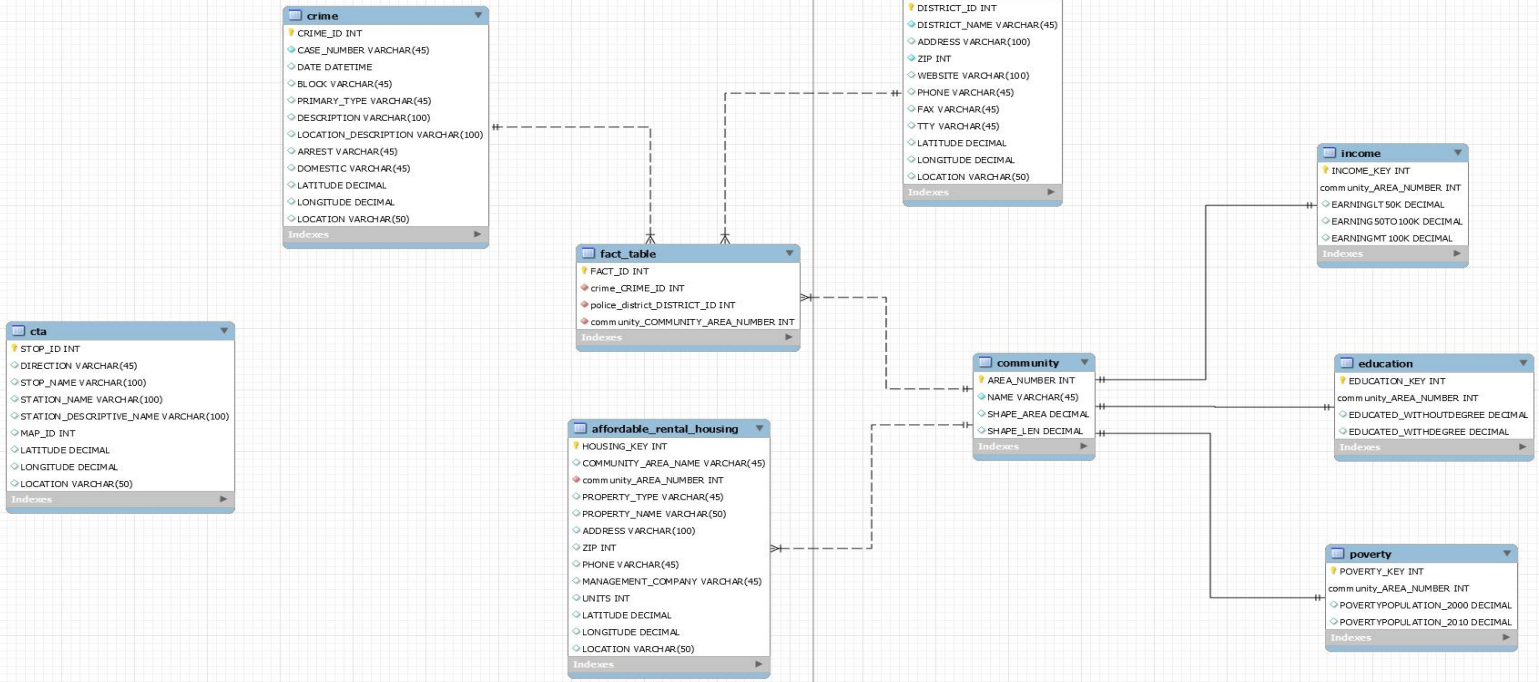
| | CRIME_ID | CASE_NUMBER | DATE | TIME |
|---|---|---|---|---|
| 2 | 25603 | JD423602 | 11/07/2020 | 19:40:00 |
| 3 | 25598 | JD423611 | 11/07/2020 | 4:07:00 |
| 4 | 12178135 | JD381686 | 9/27/2020 | 23:00:00 |
| 5 | 25600 | JD424303 | 11/07/2020 | 18:20:00 |
| 6 | 25599 | JD423568 | 11/07/2020 | 2:13:00 |
| 7 | 12186164 | JD390848 | 10/01/2020 | 7:30:00 |
| 8 | 12159178 | JD360100 | 9/07/2020 | 22:25:00 |
| 9 | 12149716 | JD346514 | 8/26/2020 | 23:53:00 |
| 10 | 12105734 | JD296828 | 7/14/2020 | 1:00:00 |
| 11 | 12065662 | JD249792 | 6/01/2020 | 12:35:00 |
| 12 | 12059214 | JD242399 | 5/17/2020 | 16:19:00 |
| 13 | 12054020 | JD236234 | 5/18/2020 | 22:40:00 |
| 14 | 11998737 | JD173249 | 3/02/2020 | 4:00:00 |
| 15 | 11986245 | JD157777 | 2/18/2020 | 20:00:00 |
| 16 | 25595 | JD422430 | 11/06/2020 | 2:54:00 |
| 17 | 25596 | JD422410 | 11/06/2020 | 2:22:00 |
| 18 | 25597 | JD422229 | 11/06/2020 | 2:58:00 |
| 19 | 11963964 | JD130474 | 1/23/2020 | 18:00:00 |
| 20 | 12192921 | JD398927 | 10/13/2020 | 22:00:00 |
| 21 | 12190098 | JD395600 | 10/11/2020 | 1:56:00 |
| 22 | 12178133 | JD371968 | 9/19/2020 | 0:01:00 |
| 23 | 12160276 | JD360992 | 9/08/2020 | 18:15:00 |
| 24 | 12152251 | JD351498 | 8/31/2020 | 11:17:00 |
| 25 | 12121583 | JD315568 | 7/30/2020 | 13:21:00 |
| 26 | 12121079 | JD314995 | 7/29/2020 | 19:00:00 |

Crime_Type | Crime_Time | Crime_Address +

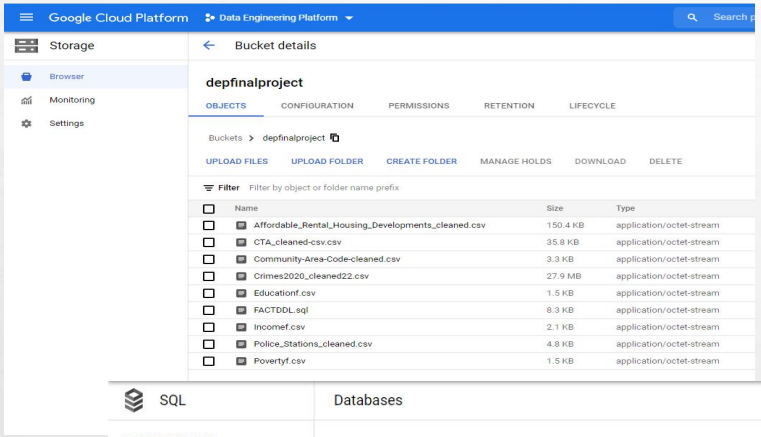
# Relational Model EER

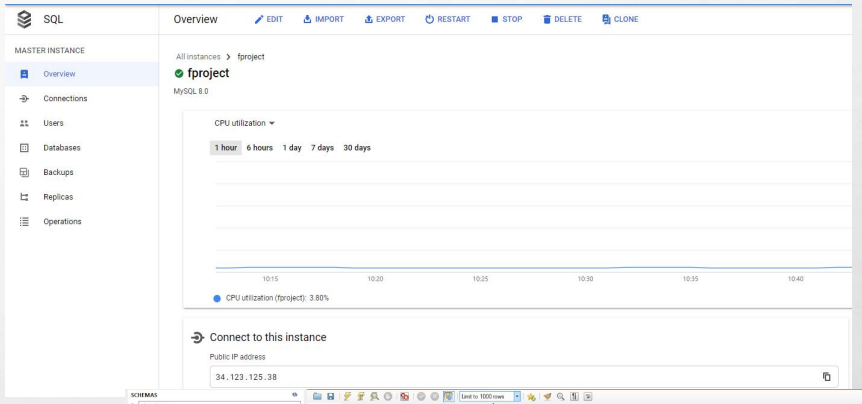


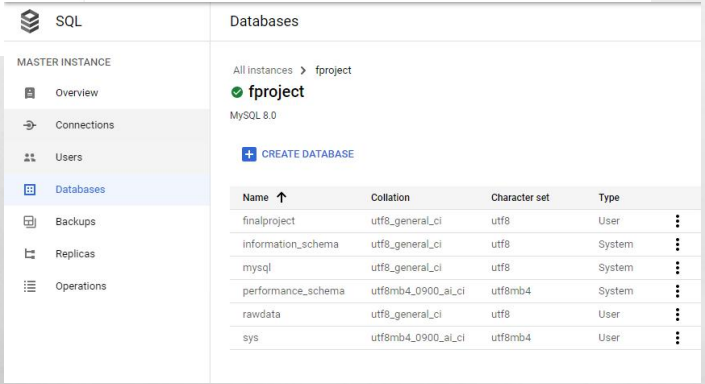
# Dimensional Model

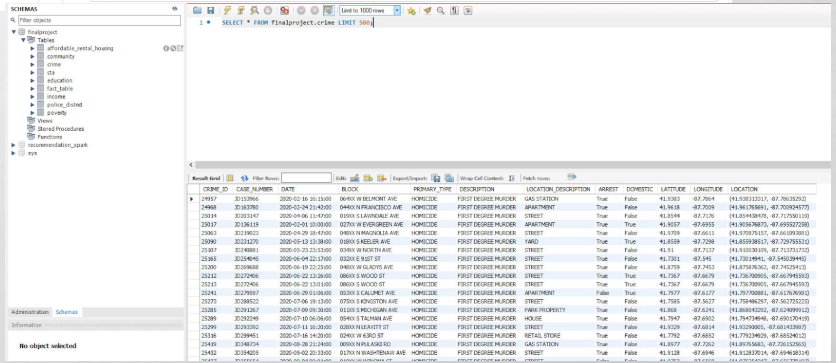


# DATA STORAGE & CLOUD

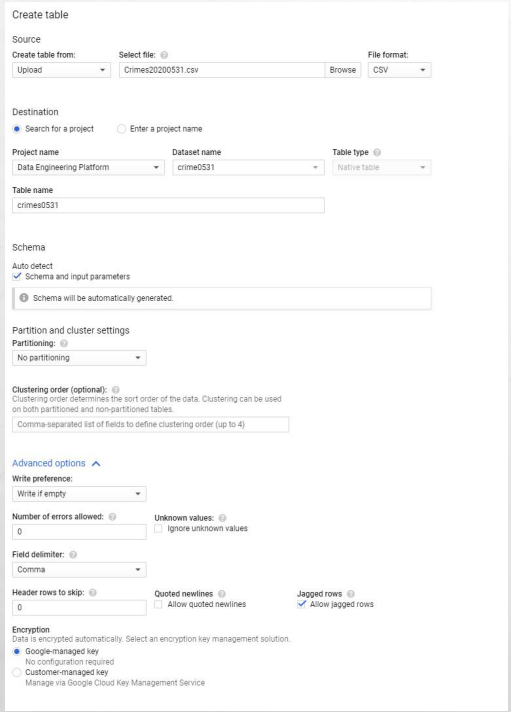


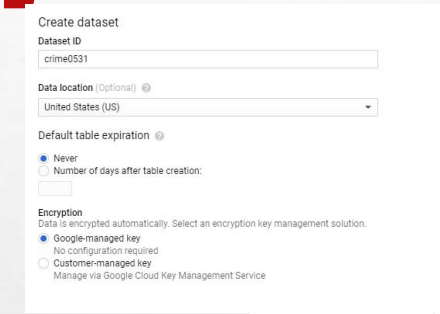



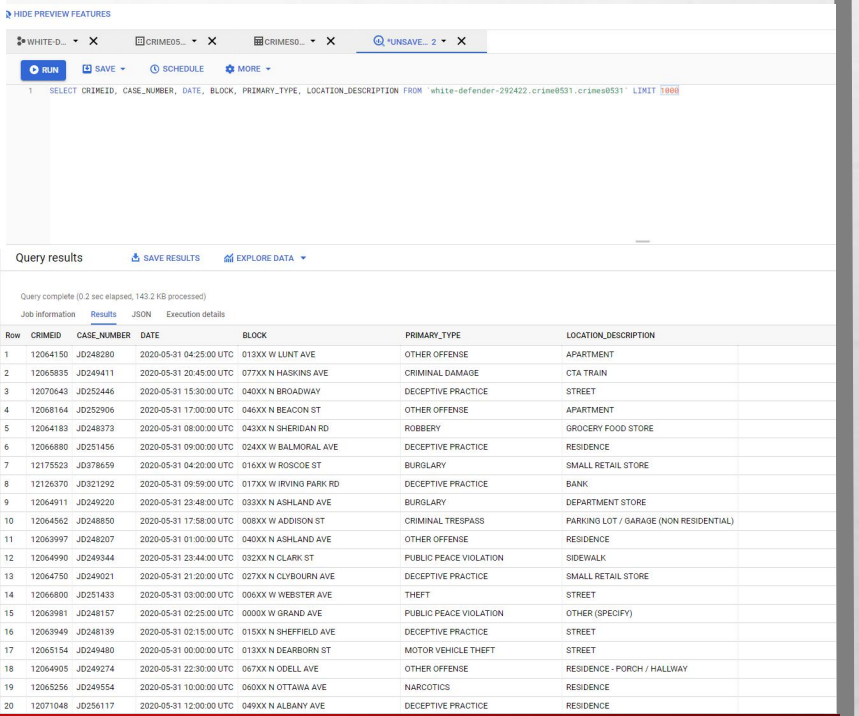



# DATA STORAGE & CLOUD

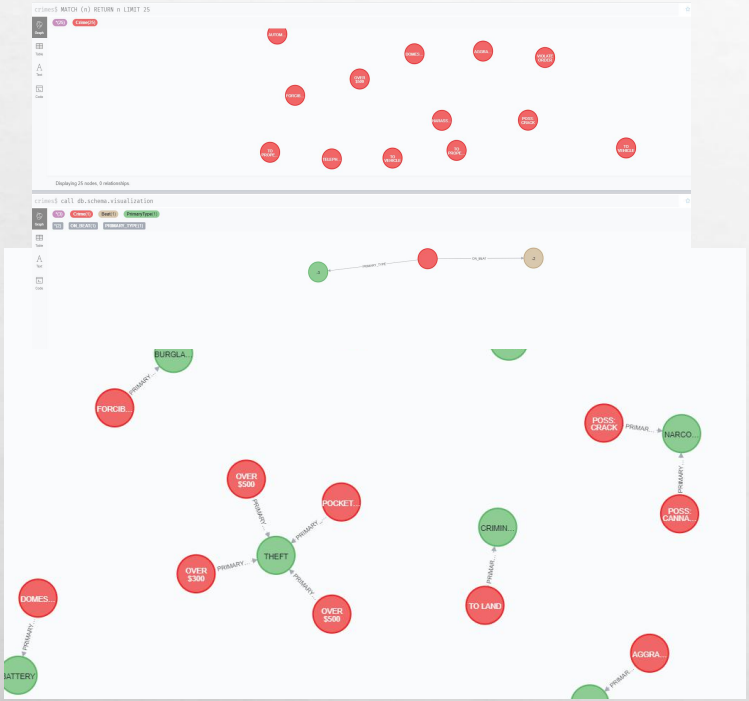


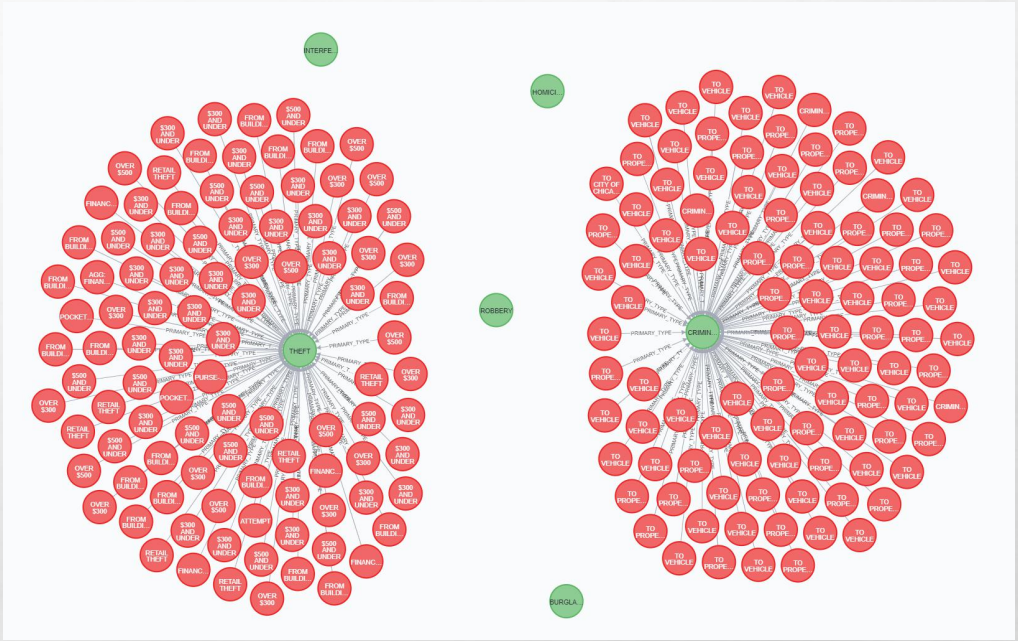




| Row | CRIMEID | CASE_NUMBER | DATE | BLOCK | PRIMARY_TYPE | LOCATION_DESCRIPTION |
|---|---|---|---|---|---|---|
| 1 | 12064150 | JD248280 | 2020-05-31 04:25:00 UTC | 013XX W LUNT AVE | OTHER OFFENSE | APARTMENT |
| 2 | 12065835 | JD249411 | 2020-05-31 20:45:00 UTC | 077XX N HASKINS AVE | CRIMINAL DAMAGE | CTA TRAIN |
| 3 | 12070643 | JD252446 | 2020-05-31 15:30:00 UTC | 040XX N BROADWAY | DECEPTIVE PRACTICE | STREET |
| 4 | 12068164 | JD252906 | 2020-05-31 17:00:00 UTC | 046XX N BEACON ST | OTHER OFFENSE | APARTMENT |
| 5 | 12064183 | JD248373 | 2020-05-31 08:00:00 UTC | 043XX N SHERIDAN RD | ROBBERY | GROCERY FOOD STORE |
| 6 | 12066880 | JD251456 | 2020-05-31 09:00:00 UTC | 024XX W BALMORAL AVE | DECEPTIVE PRACTICE | RESIDENCE |
| 7 | 12175523 | JD378659 | 2020-05-31 04:20:00 UTC | 016XX W ROSCOE ST | BURGLARY | SMALL RETAIL STORE |
| 8 | 12126370 | JD321292 | 2020-05-31 09:59:00 UTC | 017XX W IRVING PARK RD | DECEPTIVE PRACTICE | BANK |
| 9 | 12064911 | JD249220 | 2020-05-31 23:48:00 UTC | 033XX N ASHLAND AVE | BURGLARY | DEPARTMENT STORE |
| 10 | 12064562 | JD248850 | 2020-05-31 17:58:00 UTC | 008XX W ADDISON ST | CRIMINAL TRESPASS | PARKING LOT / GARAGE (NON RESIDENTIAL) |
| 11 | 12063997 | JD248207 | 2020-05-31 01:00:00 UTC | 040XX N ASHLAND AVE | OTHER OFFENSE | RESIDENCE |
| 12 | 12064990 | JD249344 | 2020-05-31 23:44:00 UTC | 032XX N CLARK ST | PUBLIC PEACE VIOLATION | SIDEWALK |
| 13 | 12064750 | JD249021 | 2020-05-31 21:20:00 UTC | 027XX N CLYBOURN AVE | DECEPTIVE PRACTICE | SMALL RETAIL STORE |
| 14 | 12066800 | JD251433 | 2020-05-31 03:00:00 UTC | 006XX W WEBSTER AVE | THEFT | STREET |
| 15 | 12063981 | JD248157 | 2020-05-31 02:25:00 UTC | 0000X W GRAND AVE | PUBLIC PEACE VIOLATION | OTHER (SPECIFY) |
| 16 | 12063949 | JD248139 | 2020-05-31 02:15:00 UTC | 015XX N SHEFFIELD AVE | DECEPTIVE PRACTICE | STREET |
| 17 | 12065154 | JD249480 | 2020-05-31 00:00:00 UTC | 013XX N DEARBORN ST | MOTOR VEHICLE THEFT | STREET |
| 18 | 12064905 | JD249274 | 2020-05-31 22:30:00 UTC | 067XX N ODELL AVE | OTHER OFFENSE | RESIDENCE - PORCH / HALLWAY |
| 19 | 12065256 | JD249554 | 2020-05-31 10:00:00 UTC | 060XX N OTTAWA AVE | NARCOTICS | RESIDENCE |
| 20 | 12071048 | JD256117 | 2020-05-31 12:00:00 UTC | 049XX N ALBANY AVE | DECEPTIVE PRACTICE | RESIDENCE |

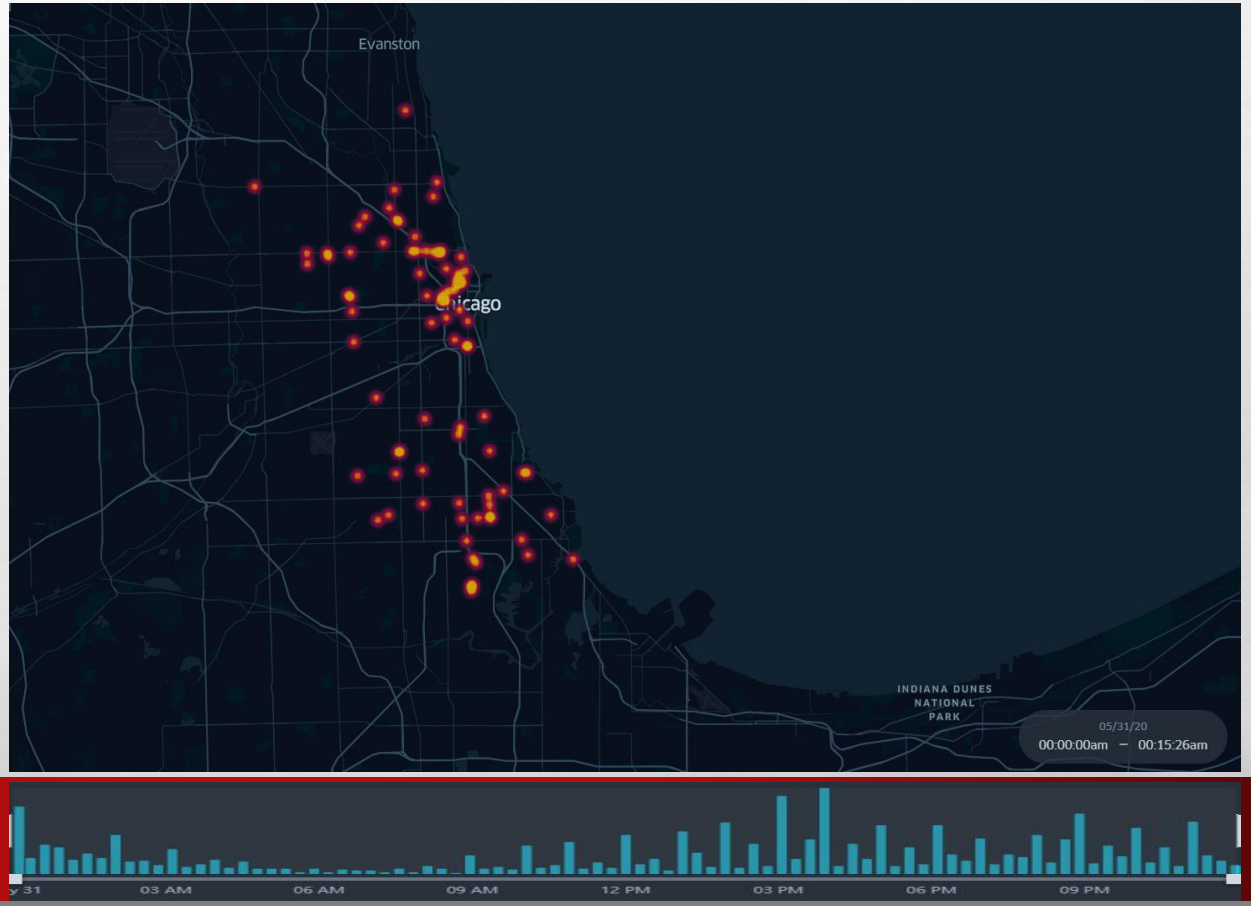
# GRAPH DATABASE: NEO4J



# ONE DAY IN CHICAGO

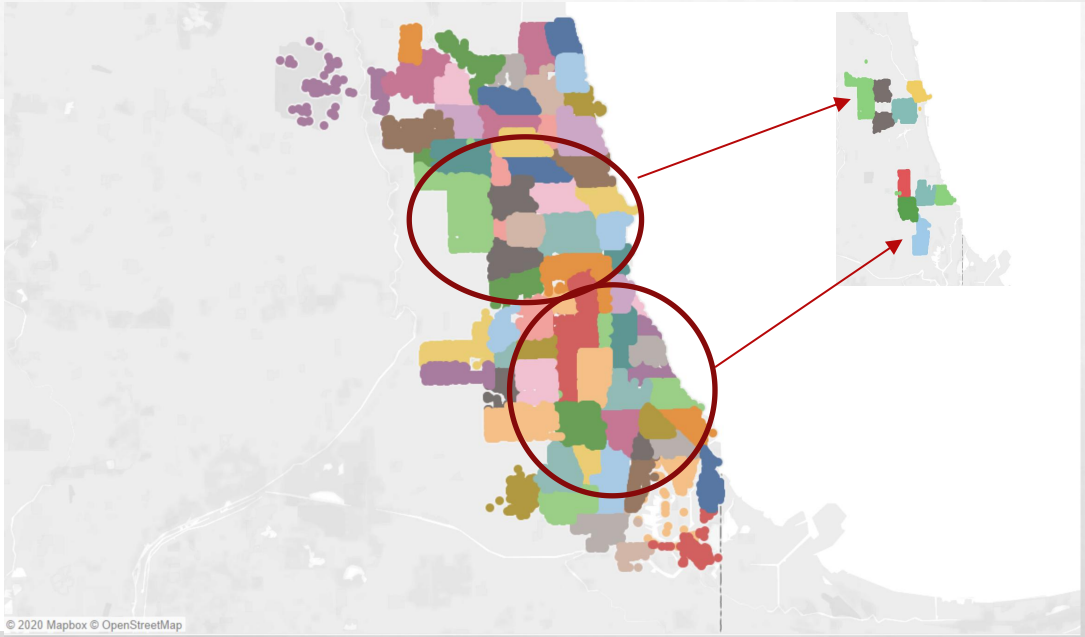
Heat for Every 15 Minutes



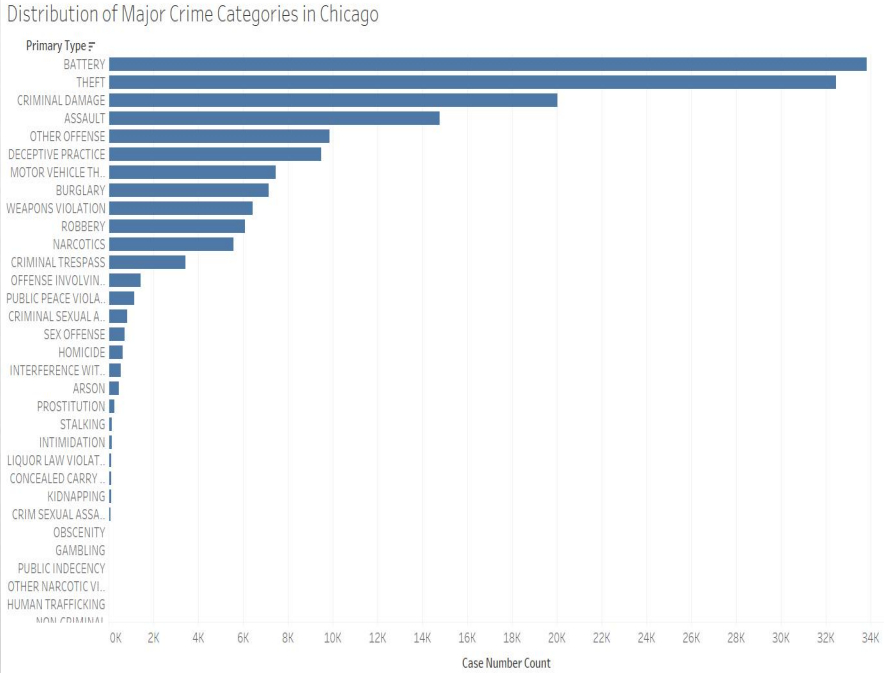
# Major Crime Prone Areas in Chicago

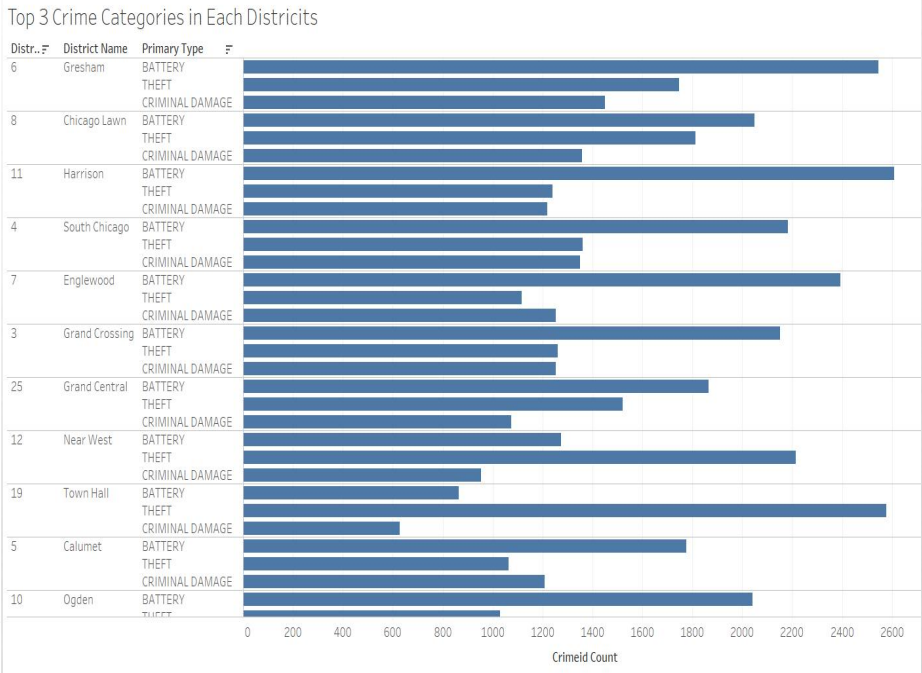
Major Crime Prone Areas in Chicago

| Community | |
|---|---|
| AUSTIN | 9,954 |
| SOUTH SHORE | 6,111 |
| NEAR NORTH SIDE | 5,736 |
| NORTH LAWNDALE | 5,567 |
| HUMBOLDT PARK | 5,303 |
| NEAR WEST SIDE | 4,992 |
| AUBURN GRESHAM | 4,876 |
| ROSELAND | 4,703 |
| GREATER GRAND CR.. | 4,637 |
| WEST ENGLEWOOD | 4,507 |



# Major Crime Issues in Chicago

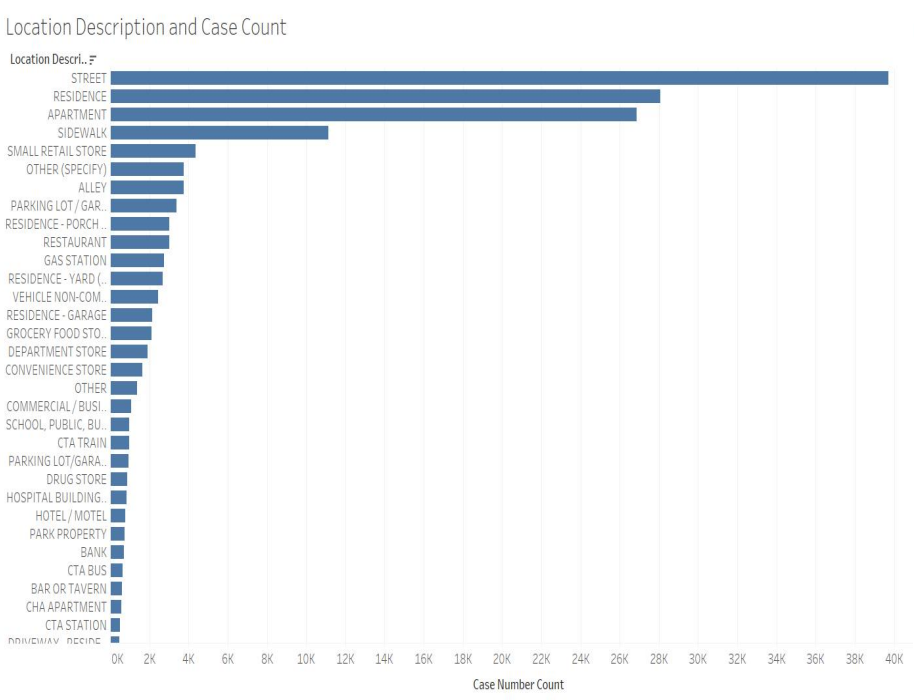



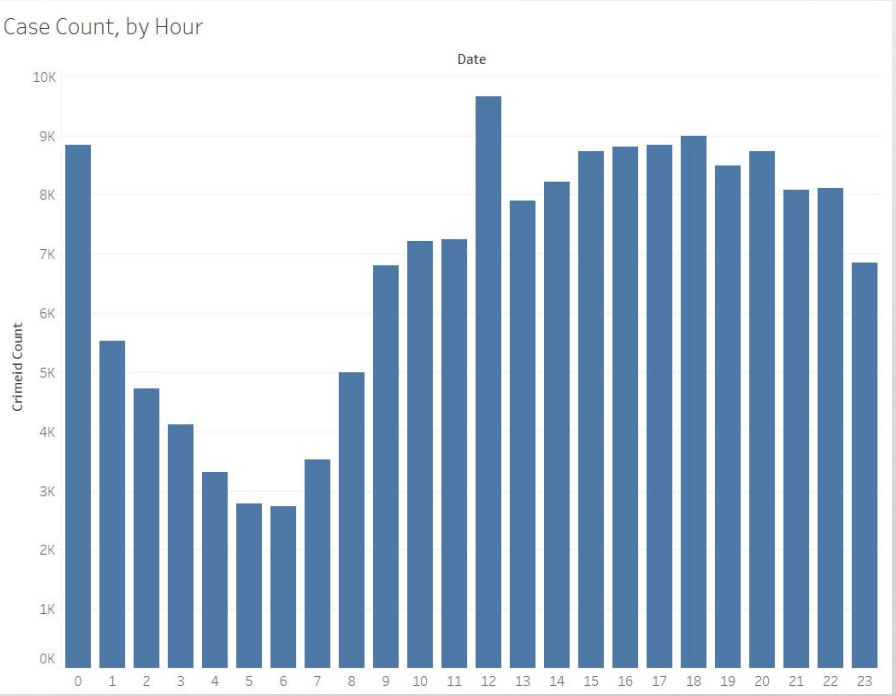




- Major Crime: 1. Battery 2. Theft 3. Criminal Damage 4. Assault

- "Battery" : the actual physical violence causing the physical harm.
- "Assault": the act which causes the victim to apprehend imminent physical harm
- "Criminal Damage": unauthorized destruction of public facilities or personal property

# Major Crime Prone Locations and Time in Chicago




- Crime prone locations: 1. residential & apartment, 2. street & sidewalk, 3. small retail stores
- Crime prone time: 12 noon, 3 to 6 pm, 0 am, 7 to 8 pm

# Top 20 Districts and Community locations in MongoDB

## Import cleaned Crime dataset

```
12  J;
13
14  mb.importContent({
15      connection: "localhost",
16      database: "inventory;",
17      fromType: "file",
18      batchSize: 2000,
19      contents
20  })
```

13.227 s  Show Timestamps

```
52  import into "inventory;.Crimes2020_dim"  62%, 101999 (+2000) docs inserted.
53  import into "inventory;.Crimes2020_dim"  64%, 103999 (+2000) docs inserted.
54  import into "inventory;.Crimes2020_dim"  65%, 105999 (+2000) docs inserted.
55  import into "inventory;.Crimes2020_dim"  66%, 107999 (+2000) docs inserted.
56  import into "inventory;.Crimes2020_dim"  67%, 109999 (+2000) docs inserted.
57  import into "inventory;.Crimes2020_dim"  69%, 111999 (+2000) docs inserted.
58  import into "inventory;.Crimes2020_dim"  70%, 113999 (+2000) docs inserted.
59  import into "inventory;.Crimes2020_dim"  71%, 115999 (+2000) docs inserted.
60  import into "inventory;.Crimes2020_dim"  72%, 117999 (+2000) docs inserted.
61  import into "inventory;.Crimes2020_dim"  73%, 119999 (+2000) docs inserted.
62  import into "inventory;.Crimes2020_dim"  75%, 121999 (+2000) docs inserted.
63  import into "inventory;.Crimes2020_dim"  76%, 123999 (+2000) docs inserted.
64  import into "inventory;.Crimes2020_dim"  77%, 125999 (+2000) docs inserted.
65  import into "inventory;.Crimes2020_dim"  78%, 127999 (+2000) docs inserted.
66  import into "inventory;.Crimes2020_dim"  80%, 129999 (+2000) docs inserted.
67  import into "inventory;.Crimes2020_dim"  81%, 131999 (+2000) docs inserted.
68  import into "inventory;.Crimes2020_dim"  82%, 133999 (+2000) docs inserted.
69  import into "inventory;.Crimes2020_dim"  83%, 135999 (+2000) docs inserted.
70  import into "inventory;.Crimes2020_dim"  84%, 137999 (+2000) docs inserted.
71  import into "inventory;.Crimes2020_dim"  86%, 139999 (+2000) docs inserted.
72  import into "inventory;.Crimes2020_dim"  87%, 141999 (+2000) docs inserted.
73  import into "inventory;.Crimes2020_dim"  88%, 143999 (+2000) docs inserted.
74  import into "inventory;.Crimes2020_dim"  89%, 145999 (+2000) docs inserted.
75  import into "inventory;.Crimes2020_dim"  91%, 147999 (+2000) docs inserted.
76  import into "inventory;.Crimes2020_dim"  92%, 149999 (+2000) docs inserted.
77  import into "inventory;.Crimes2020_dim"  93%, 151999 (+2000) docs inserted.
78  import into "inventory;.Crimes2020_dim"  94%, 153999 (+2000) docs inserted.
79  import into "inventory;.Crimes2020_dim"  96%, 155999 (+2000) docs inserted.
80  import into "inventory;.Crimes2020_dim"  97%, 157999 (+2000) docs inserted.
81  import into "inventory;.Crimes2020_dim"  98%, 159999 (+2000) docs inserted.
82  import into "inventory;.Crimes2020_dim"  99%, 161999 (+2000) docs inserted.
83  import into "inventory;.Crimes2020_dim" 100%, 163360 (+1361) docs inserted.
84  import into "inventory;.Crimes2020_dim" finished.
85
86  A total of 163360 document(s) have been imported into 1 collection(s).
87  {
88    "Crimes2020_dim": {
89      "nInserted": 163360,
90      "nModified": 0,
91      "nSkipped": 0,
92      "failed": 0
93    }
```

## Top 20 crimes by location description

```
db.Crimes2020_dim.aggregate([{$group: { _id: "$LOCATION DESCRIPTION", count: { $sum: 1 }}},{$sort:{"count":-1}}])
```

Crimes2020_dim  0.991 s  Fetch Count

| | _id | count |
|---|---|---|
| 1 | STREET | 39,747 (39.7K) |
| 2 | RESIDENCE | 28,071 (28.1K) |
| 3 | APARTMENT | 26,862 (26.9K) |
| 4 | SIDEWALK | 11,123 (11.1K) |
| 5 | SMALL RETAIL STORE | 4,348 (4.3K) |
| 6 | OTHER (SPECIFY) | 3,752 (3.8K) |
| 7 | ALLEY | 3,744 (3.7K) |
| 8 | PARKING LOT / GARAGE (NON RESIDENTIAL) | 3,394 (3.4K) |
| 9 | RESIDENCE - PORCH / HALLWAY | 3,028 (3.0K) |
| 10 | RESTAURANT | 3,005 (3.0K) |
| 11 | GAS STATION | 2,745 (2.7K) |
| 12 | RESIDENCE - YARD (FRONT / BACK) | 2,681 (2.7K) |
| 13 | VEHICLE NON-COMMERCIAL | 2,438 (2.4K) |
| 14 | RESIDENCE - GARAGE | 2,153 (2.2K) |
| 15 | GROCERY FOOD STORE | 2,121 (2.1K) |
| 16 | DEPARTMENT STORE | 1,899 (1.9K) |
| 17 | CONVENIENCE STORE | 1,648 (1.6K) |
| 18 | OTHER | 1,374 (1.4K) |
| 19 | COMMERCIAL / BUSINESS OFFICE | 1,072 (1.1K) |
| 20 | SCHOOL, PUBLIC, BUILDING | 969 |

# Top 20 Districts and Community locations in MongoDB

Input

```
db.Crimes2020_dim.aggregate([{$group: { _id: "$Community Number",
    count: { $sum: 1 }}},{$sort:{"count":-1}}])
```

Crimes2020_dim 0.730 s 23 Docs

| | _id | count |
|---|---|---|
| 1 | 11 | 11,619 (11.6K) |
| 2 | 6 | 11,083 (11.1K) |
| 3 | 8 | 10,153 (10.2K) |
| 4 | 4 | 9,501 (9.5K) |
| 5 | 7 | 9,387 (9.4K) |
| 6 | 25 | 8,681 (8.7K) |
| 7 | 3 | 8,601 (8.6K) |
| 8 | 10 | 8,054 (8.1K) |
| 9 | 5 | 8,000 (8.0K) |
| 10 | 12 | 7,755 (7.8K) |
| 11 | 9 | 7,234 (7.2K) |
| 12 | 2 | 7,190 (7.2K) |
| 13 | 15 | 7,069 (7.1K) |
| 14 | 19 | 7,019 (7.0K) |
| 15 | 18 | 6,880 (6.9K) |
| 16 | 1 | 6,613 (6.6K) |
| 17 | 22 | 5,333 (5.3K) |
| 18 | 16 | 5,232 (5.2K) |
| 19 | 24 | 5,172 (5.2K) |
| 20 | 14 | 5,087 (5.1K) |

Crimes2020_dim 0.922 s 23 Docs

```
/* 1 */
{
    "_id" : 11,
    "count" : 11619
},

/* 2 */
{
    "_id" : 6,
    "count" : 11083
},

/* 3 */
{
    "_id" : 8,
    "count" : 10153
},

/* 4 */
{
    "_id" : 4,
    "count" : 9501
},

/* 5 */
{
    "_id" : 7,
    "count" : 9387
},
```

Output

```
db.Crimes2020_dim.aggregate([{$group: { _id: "$DISTRICT",
count: { $sum: 1 }}},{$sort:{"count":-1}}])
```

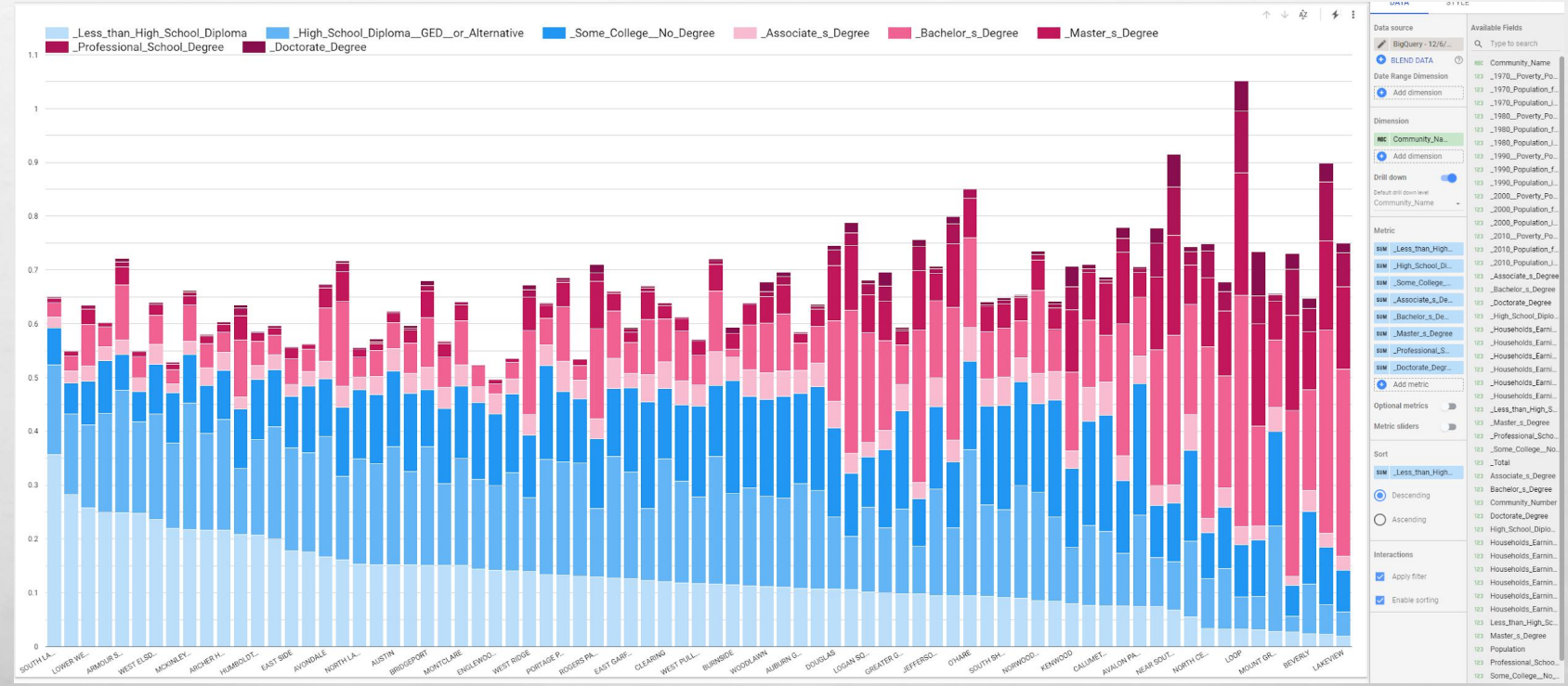
Crimes2020_dim 0.859 s 77 Docs

| | _id | count |
|---|---|---|
| 1 | 25 | 9,965 (10.0K) |
| 2 | 43 | 6,118 (6.1K) |
| 3 | 8 | 5,736 (5.7K) |
| 4 | 29 | 5,573 (5.6K) |
| 5 | 23 | 5,309 (5.3K) |
| 6 | 28 | 4,995 (5.0K) |
| 7 | 71 | 4,878 (4.9K) |
| 8 | 49 | 4,707 (4.7K) |
| 9 | 69 | 4,643 (4.6K) |
| 10 | 67 | 4,513 (4.5K) |
| 11 | 44 | 4,328 (4.3K) |
| 12 | 68 | 4,165 (4.2K) |
| 13 | 32 | 4,131 (4.1K) |
| 14 | 24 | 4,031 (4.0K) |
| 15 | 66 | 3,881 (3.9K) |
| 16 | 26 | 3,865 (3.9K) |
| 17 | 6 | 3,216 (3.2K) |
| 18 | 53 | 2,992 (3.0K) |
| 19 | 27 | 2,916 (2.9K) |
| 20 | 19 | 2,891 (2.9K) |

Crimes2020_dim 0.658 s 77 Docs

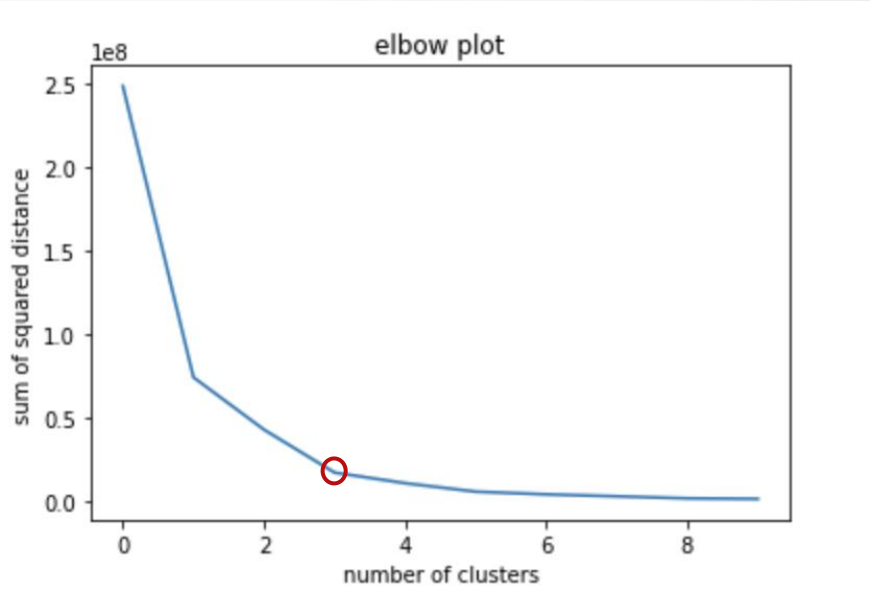
```
/* 1 */
{
    "_id" : 25,
    "count" : 9965
},

/* 2 */
{
    "_id" : 43,
    "count" : 6118
},

/* 3 */
{
    "_id" : 8,
    "count" : 5736
},

/* 4 */
{
    "_id" : 29,
    "count" : 5573
},

/* 5 */
{
    "_id" : 23,
    "count" : 5309
},
```

# GCP BigQuery & Data Visualization Studio

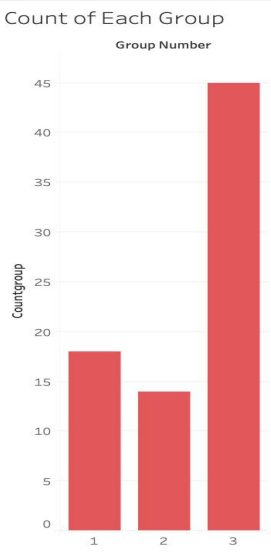


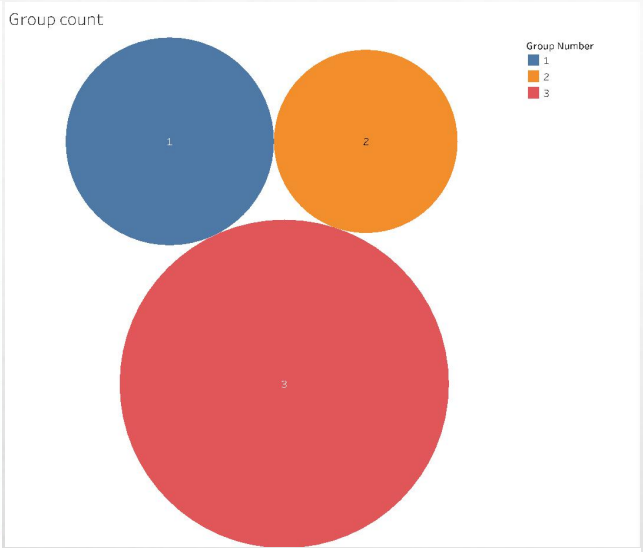
# Clustering: K-Means

- **Crime:**
  Community Area
  Crime Count
- **Poverty:**
  2010 % Poverty Population
- **Education:**
  % Educated Population with No Degree
  % Educated Population with No Degree
- **Income:**
  % Households Earning less than$49,999
  % Households Earning between $50,000 and $99,999
  % Households Earning more than 100,000



# Crime Demographics & Group Comparison

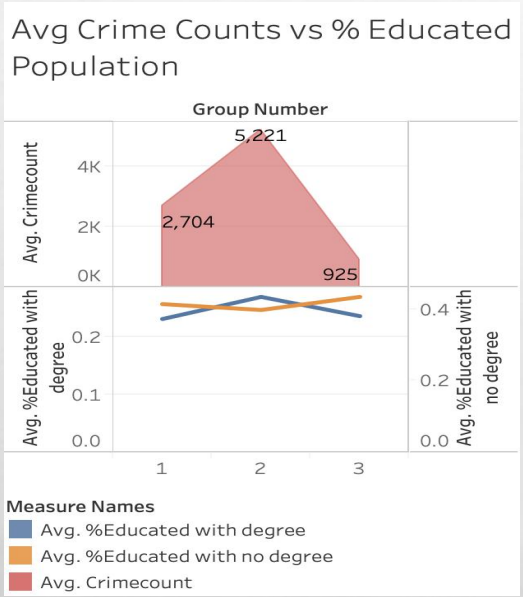


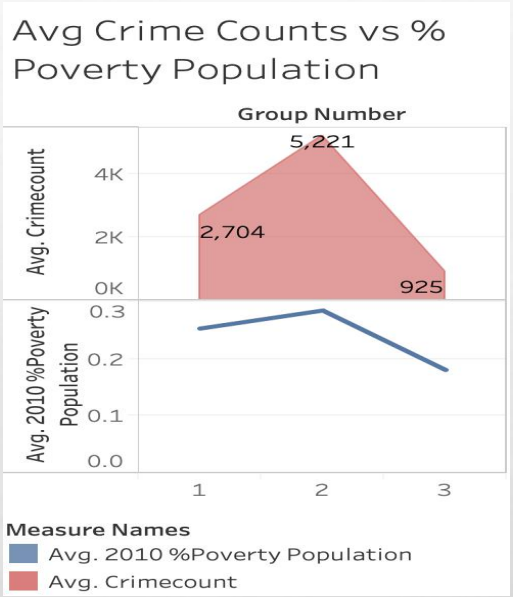



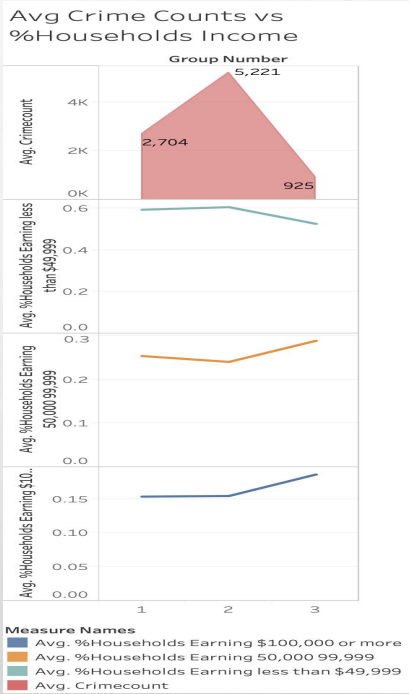

| Group Number | Avg. Crimecount | Median %Educated with degree | Median %Educated with no degree | Median 2010 %Poverty Population |
|---|---|---|---|---|
| 1 | 2,704 | 0.19 | 0.46 | 0.26 |
| 2 | 5,221 | 0.15 | 0.45 | 0.28 |
| 3 | 925 | 0.19 | 0.47 | 0.16 |

| | Group 2 | Group 3 |
|---|---|---|
| Population | 699,274 | 1,068,348 |
| Average Crime Count | 5,221 | 925 |
| Median % Educated with Degree | 15% | 19% |
| Median % Educated with No Degree | 45% | 47% |
| Median % 2010 Poverty Population | 28% | 16% |
| Top 3 Crime Type | Battery; Theft; Criminal Damage | Theft; Battery; Criminal Damage |

# Avg Crime Counts vs % Educated Population, % Poverty Population & Households Income



# RECOMMENDATIONS

Design Considerations

**Indexing for Performance :**

- Add additional index to support frequent queries to enhance performance (e.g. community areas)

**Data Integrity:**

- Add constraint foreign keys, auto increment primary keys in dimensional database to ensure data with references to related fields
- Decrease possibility for missing or duplicating data.

**Data Dimensions:**

- Apply snowflake schema for database, better support analysis.
- Thoroughly reviewed Chicago Data Portal to identify possible dimensions and related attributes to crime case analysis.
- Avoid data redundancies, inconsistencies and to ensure better maintenance of database.

**Enhancements for Future Consideration:**

- Using GCP Compute Engine where possible to improve efficiency.
- Getting more specific data of affordable rental housing and CTA and try to link with crime data to give a further analysis.
- Analyzing the crime data across years.

# LESSONS LEARNED

**Data Collecting and Cleaning:**

- Take down each step in the process, save files in chronological order.
- Import .csv tables into MySQL and GCP require to ensure consistent data types.
- Crime data is updated a few weeks later than they're first uploaded, thus the latest data should be analyzed carefully.

**Data Analysis Challenges due to Data Quality Issues :**

- Some data columns are frequently missing, we need to remove these lines before starting analyzing.
- Demographic data are macro level data which is updated in a much longer period.
- Some outliers may skew the average crime counts.

**Team Work:**

- Meeting regularly and make sure each other is on the same page.
- Follow the schedule and be open to share ideas.

# THANK YOU!

# APPENDIX

# DATA SOURCES

| DATA | DESCRIPTION |
|---|---|
| Chicago Public Safety – Crimes | Chicago city community level data of crimes such as crime time, type, location description and community area. |
| Chicago Public Safety - Police Stations | Chicago city community level data of police stations such as district, website and location. |
| Chicago Transportation - Metra Stations | Chicago city community level data of metra stations such as stop name and location. |
| Chicago Housing - Chicago Neighbourhoods | Chicago city community level data of housing such as property type, address, units and location. |
| Chicago Demographic-Chicago Neighbourhoods | Chicago city community level data of demographic information such as percentage of poverty population, percentage of population educated at different levels and percentage of population at different income levels. |

# References


**Crimes 2020:**
https://data.cityofchicago.org/Public-Safety/Crimes-2020/qzdf-xmn8
Data Size: 38.8 MB (Crimes_-_2020.csv)
Number of Observations/attributes: 22 variables

**Chicago Police Stations:**
https://www.chicago.gov/city/en/depts/cpd/dataset/police_stations.html
Data Size: 5.56 KB (Police_Stations.csv)
Number of Observations/attributes: 15 variables

**System Information of CTA:**
https://data.cityofchicago.org/Transportation/CTA-System-Information-List-of-L-StopsData/8pix-ypme/data
Data Size: 46.7 KB (CTA_System_Information_List_of_L_Stops.csv)
Number of Observations/attributes: 16 variables

**Affordable Rental Housing Developments**
https://data.cityofchicago.org/Community-Economic-Development/Affordable-Rental-Housing-Developments/s6ha-ppgi
Data Size: 174 KB (Affordable_Rental_Housing_Developments.csv)
Number of Observations/attributes: 14 variables

# References

**Chicago Demographic:**
http://cn2015.net/district/bronzeville-south-lakefront/bronzeville-south-lakefront-district-data-analysis/
Data Size: 29 KB (Demographic.csv)
Number of Observations/attributes: 49 variables

**Community Area Boundaries in Chicago:**
https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6
Data Size: 1.91 MB (CommAreas.csv)
Number of Observations/attributes: 77 variables

# DATA PROFILE - Crimes 2020

Rules:

1. “The unauthorized use of the words "Chicago Police Department," "Chicago Police," or any colorable imitation of these words or the unauthorized use of the Chicago Police Department logo is unlawful.”
2. All data visualizations on maps should be considered approximate and attempts to derive specific addresses are strictly prohibited

# DATA PROFILE - Crimes 2020

File Naming Conventions Used: IUCR - The Illinois Uniform Crime Report Code

Data Quality Metrics: the Chicago Police Department does not guarantee the accuracy, completeness, timeliness, or correct sequencing of the information and the information should not be used for comparison purposes over time

Types of data used: Structured Data

| | |
|---|---|
| Data Owner | Police |
| Time Period | 2001 to present, minus the most recent seven days |
| Frequency | Data is updated daily, Tuesday through Sunday |

# DATA PROFILE - Chicago Police Stations

Rules: 1. Data users should include the following disclaimer at the site where the software application, or other secondary or derivative application can be accessed or downloaded:

“This site provides applications using data that has been modified for use from its original source, www.cityofchicago.org, the official website of the City of Chicago. The City of Chicago makes no claims as to the content, accuracy, timeliness, or completeness of any of the data provided at this site. The data provided at this site is subject to change at any time. It is understood that the data provided at this site is being used at one’s own risk.”

Comply with any additional Terms of Use set forth by the City agency or department providing data used by the software application, or other secondary or derivative application, including, without limitation, requirements to include additional citations or disclaimers at the site where the application can be accessed or downloaded.

# DATA PROFILE - Chicago Police Stations

Rules: 2. To the fullest extent permitted by law, any user of the data provided at this website shall indemnify and hold harmless the City from any claim, loss, damage, injury, or liability of any kind (including, without limitation, incidental and consequential damages, court costs, attorney's fees and costs of investigation), that arises directly or indirectly, in whole or in part, from that user's use of this data, including any secondary or derivative use of the information provided herein. Every user of this data also specifically acknowledges and agrees to have an immediate and independent obligation to defend the City from any claim that may fall within this indemnification provision, even if the allegations are or may be groundless, false or fraudulent, which obligation arises at the time such claim is tendered to the user by the City and continues at all times thereafter.

# DATA PROFILE - Chicago Police Stations

File Naming Conventions Used: None.

Data Quality Metrics: The City of Chicago (“City”) voluntarily provides the data on this website as a service to the public.  The City makes no warranty, representation, or guaranty as to the content, accuracy, timeliness, or completeness of any of the data provided at this website.

Types of data used: Structured Data

# DATA PROFILE - System Information of CTA

File Naming Conventions Used: DIRECTION_ID - Normal Direction of train traffic at platform

Data Quality Metrics: The data provides location and basic service availability information. Updated at May 15, 2018, the latest data can be found through Chicago Data Portal for rail stations. The content of the data, such as Stop Name and Stop Location, does not change over time.

The license for this dataset is unspecified

Types of data used: Structured

Data Owner Chicago Transit Authority

# DATA PROFILE - Affordable Rental Housing Developments

Rules: The list is provided as a courtesy to the public. No rule is available.

Data Quality Metrics: The latest update is on July 23, 2020. Data are updated only as needed. The list does not include every City-assisted affordable housing unit that may be available for rent, nor does it include the hundreds of thousands of naturally occurring affordable housing units located throughout Chicago without City subsidies.

Types of data used: Structured

| | |
|---|---|
| Data Owner | Housing and Economic Development |
| Time Period | Current as of March 2013 |
| Frequency | Data are updated as needed |

# DATA PROFILE - Chicago Demographic

File Naming Conventions Used: Names of variables are easy to understand.

Data Quality Metrics:  The majority are historical data about information before 2016. In terms of accuracy, data in each indicator do not change over time and are correct. However, in terms of consistency, although the data provide certain insights, they are not very much consistent with the data of crimes in 2020

The data include each community's demographic information – poverty, education and income. Neighborhoods are in a constant state of change. Data for the built environment informs about existing conditions and allows for comparisons across neighborhoods and against citywide averages and helps to determine what backgrounds are in each community.

Types of data used: Metadata, Structured Data

# DATA PROFILE - Community Area Boundaries in Chicago

Rules: The permissions of using the data are granted to the public.

Data Quality Metrics:  The latest update is on Dec 18, 2018. The data content is fixed. Since the distribution of the Chicago area has not changed in the last two years, the data is not out of date. The data tends to be accurate because the difference between the region segmentation and the satellite images is not large.

Types of data used: Metadata

| | |
|---|---|
| Data Provided By | City of Chicago |
| Source Link | http://www.cityofchicago.org |
| Time Period | Current community area boundaries. |
| Frequency | Updated as needed. |

# SQL Queries

```sql
# Find out top 10 blocks which have the most crime records
SELECT
    BLOCK, COUNT(DISTINCT (CRIME_ID)) AS CASES
FROM
    finalproject.crime
GROUP BY BLOCK
ORDER BY CASES DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

| BLOCK | CASES |
|---|---|
| 001XX N STATE ST | 408 |
| 0000X W TERMINAL ST | 186 |
| 064XX S DR MARTIN LUTHER KING JR DR | 184 |
| 008XX N MICHIGAN AVE | 182 |
| 065XX S DR MARTIN LUTHER KING JR DR | 138 |
| 037XX W CHICAGO AVE | 126 |
| 100XX W OHARE ST | 124 |
| 012XX S WABASH AVE | 115 |
| 006XX N MICHIGAN AVE | 115 |
| 075XX S STONY ISLAND AVE | 113 |


```sql
# Find out top 10 blocks which have the most crime records
SELECT
    BLOCK, COUNT(DISTINCT (CRIME_ID)) AS CASES
FROM
    finalproject.crime
GROUP BY BLOCK
ORDER BY CASES DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

| BLOCK | CASES |
|---|---|
| 001XX N STATE ST | 408 |
| 0000X W TERMINAL ST | 186 |
| 064XX S DR MARTIN LUTHER KING JR DR | 184 |
| 008XX N MICHIGAN AVE | 182 |
| 065XX S DR MARTIN LUTHER KING JR DR | 138 |
| 037XX W CHICAGO AVE | 126 |
| 100XX W OHARE ST | 124 |
| 012XX S WABASH AVE | 115 |
| 006XX N MICHIGAN AVE | 115 |
| 075XX S STONY ISLAND AVE | 113 |


```sql
# Find out the most battery happened in which police district order by the amount of batteries
SELECT
    pd.DISTRICT_NAME,
    COUNT(DISTINCT (c.CRIME_ID)) AS BATTERY_Number
FROM
    finalproject.crime AS c
        JOIN
    finalproject.fact_table AS ft
        JOIN
    finalproject.police_district AS pd
WHERE
    c.CRIME_ID = ft.crime_CRIME_ID
        AND ft.police_district_DISTRICT_ID = pd.DISTRICT_ID
        AND c.PRIMARY_TYPE = 'BATTERY'
GROUP BY ft.police_district_DISTRICT_ID
ORDER BY BATTERY_Number DESC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

| DISTRICT_NAME | BATTERY_Number |
|---|---|
| Harrison | 2610 |
| Gresham | 2547 |
| Englewood | 2394 |
| South Chicago | 2184 |
| Grand Crossing | 2154 |
| Chicago Lawn | 2048 |
| Ogden | 2041 |
| Austin | 1923 |
| Grand Central | 1865 |
| Calumet | 1778 |
| Deering | 1624 |
| Wentworth | 1527 |
| Near West | 1274 |
| Morgan Park | 1008 |
| Rogers Park | 1000 |
| Central | 988 |
| Jefferson Park | 928 |
| Near North | 884 |
| Town Hall | 865 |
| Albany Park | 826 |
| Shakespeare | 751 |
| Lincoln | 607 |
| Headquarters | 1 |