

FACTORS BEHIND CHICAGO CRIMES

Jianan Fei, Yue Sun,



Ruobing Xue, Kaihao Fan

(Group 4: Future Splendid & X-Fabulous)

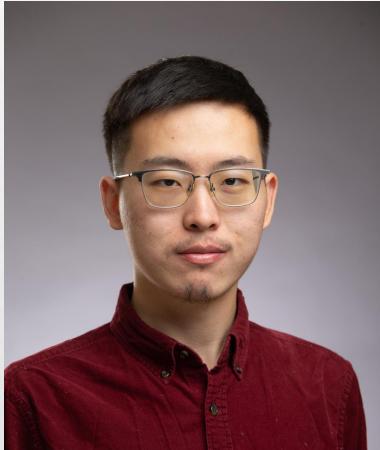
December 2020

MEET THE TEAM



Jianan Fei

feij@uchicago.edu



Kaihao Fan

kaihaofan@uchicago.edu



Ruobing Xue

ruobing@uchicago.edu

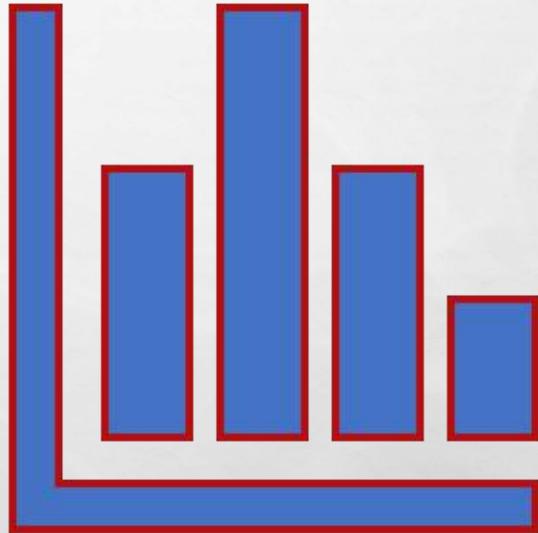


Yue Sun

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 toTimeStamp

```
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	LONGITUD	
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	-87.76251	
339	25598	JD423611	11/07/2020 04:07:00 AM	021XX E 72ND PL	HOMICIDE	crime['DATE'] = pd.to_datetime(crime['DATE']) crime['DATE']	0 1 2 3 4 ...	2020-11-07 19:40:00 2020-11-07 04:07:00 2020-09-27 23:00:00 2020-11-07 18:20:00 2020-11-07 02:13:00 ...	True	False	3	43	41.763456	-87.57333
409	12178135	JD381686	09/27/2020 11:00:00 PM	068XX S RIDGELAND AVE	OTHER OFFENSE	OT 1 2 3 4	APTMENT	2020-09-27 23:00:00 2020-11-07 18:20:00 2020-11-07 02:13:00 ...	False	False	3	43	41.770931	-87.58258
508	25600	JD424303	11/07/2020 06:20:00 PM	074XX N OAKLEY AVE	HOMICIDE	163356 163356 163357 163358 163359	STREET	2020-01-01 01:24:00 2020-01-01 03:00:00 2020-01-01 14:13:00 2020-01-01 04:07:00 2020-01-01 15:00:00	True	False	24	2	42.016114	-87.68751
620	25599	JD423568	11/07/2020 02:13:00 AM	059XX W HURON ST	HOMICIDE	163356 163356 163357 163358 163359	STREET	2020-01-01 01:24:00 2020-01-01 03:00:00 2020-01-01 14:13:00 2020-01-01 04:07:00 2020-01-01 15:00:00	False	False	15	25	41.892831	-87.77348

```
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	LONGITUD	
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	OT 1 2 3 4	APTMENT	2020-09-27 23:00:00 2020-11-07 18:20:00 2020-11-07 02:13:00 ...	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

Cluster & Edit column "DESCRIPTION"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method: nearest neighbor ▾ ppm ▾ Radius: 1.0 Block Chars: 6

31 clusters found

MANUFACTURE / DELIVER - AMPHETAMINES (11 rows)		
2	190	• CRIMINAL SEXUAL ABUSE (184 rows) • AGG CRIMINAL SEXUAL ABUSE (6 rows)
2	105	• POSSESS - HALLUCINOGENS (77 rows) • POSS - HALLUCINOGENS (28 rows)
2	17	• CYCLE, SCOOTER, BIKE NO VIN (14 rows) • CYCLE, SCOOTER, BIKE W-VIN (3 rows)
2	9	• CONTRIBUTE TO THE DELINQUENCY OF CHILD (6 rows) • CONTRIBUTE TO THE CRIMINAL DELINQUENCY OF CHILD (3 rows)
2	130	• MANUFACTURE / DELIVER - CRACK (102 rows) • MANUFACTURE / DELIVER - COCAINE (28 rows)
2	41	• UNLAWFUL USE - OTHER DANGEROUS WEAPON (22 rows) • UNLAWFUL USE OTHER DANG WEAPON (19 rows)

Select All Unselect All

Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Cluster & Edit column "LOCATION DESCRIPTION"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method: key collision ▾

Keying Function: fingerprint ▾

4 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	203	• SCHOOL - PRIVATE BUILDING (107 rows) • SCHOOL, PRIVATE, BUILDING (96 rows)	<input checked="" type="checkbox"/>	SCHOOL - PRIVATE BUILDING
2	969	• SCHOOL, PUBLIC, BUILDING (622 rows) • SCHOOL - PUBLIC BUILDING (347 rows)	<input checked="" type="checkbox"/>	SCHOOL, PUBLIC, BUILDING
2	62	• SCHOOL - PRIVATE GROUNDS (33 rows) • SCHOOL, PRIVATE, GROUNDS (29 rows)	<input checked="" type="checkbox"/>	SCHOOL - PRIVATE GROUNDS
		^HOOL - PUBLIC GROUNDS (174 rows) HOOL, PUBLIC, GROUNDS (150 rows)	<input checked="" type="checkbox"/>	SCHOOL - PUBLIC GROUNDS

Rows in Cluster

60 — 970

Average Length of Choices

23 — 25

Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Choices in Cluster

2 — 3

Rows in Cluster

0 — 2000

Average Length of Choices

12 — 57

Length Variance of Choices

0.5 — 5.5

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	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/district/1	312-745-4290	312-3694	312-3693	1176569.052	1891771.704
2	6	Gresham	7908 S Halsted St	Chicago	IL 60620	http://home.chi	Poverty = poverty[['Community Number','2000 %Poverty Population','2010 %Poverty Population']]				
3	11	Harrison	3151 W Harrison St	Chicago	IL 60612	http://home.chi	Poverty.head(5)				
4	16	Jefferson Park	5151 N Milwaukee Ave	Chicago	IL 60630	http://home.chi	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	
Education=education[['Community Number','%Educated with no degree','%Educated with degree']]											
Education.head(5)											
							Community Number	%Educated with no degree	%Educated with degree	Income=income[['Community Number','%Households Earning less than \$49,999','%Households Earning 50,000_99,999','%Households Earning 100,000 or more']]	
0	1						0	0.39	0.37	Income.head(5)	
1	2						1	0.39	0.21		
2	3						2	0.34	0.41	Community Number %Households Earning less than \$49,999 %Households Earning 50,000_99,999 %Households Earning \$100,000 or more	
3	4						3	0.32	0.41	0	0.59
4	5						4	0.21	0.55	1	0.25
										2	0.16
										3	0.18
										4	0.21
										5	0.24
											0.44

Relational Model Normalization

- Normalize the cleaned Crime table

1	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

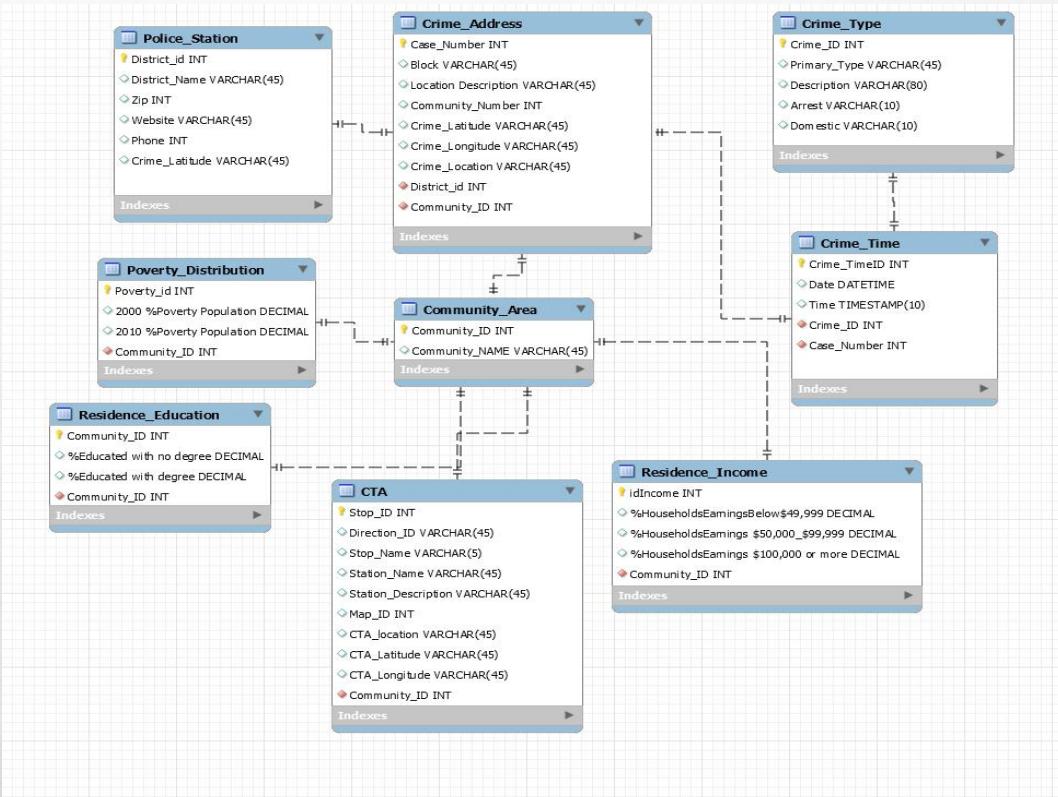
Crime_Time

Crime_Address

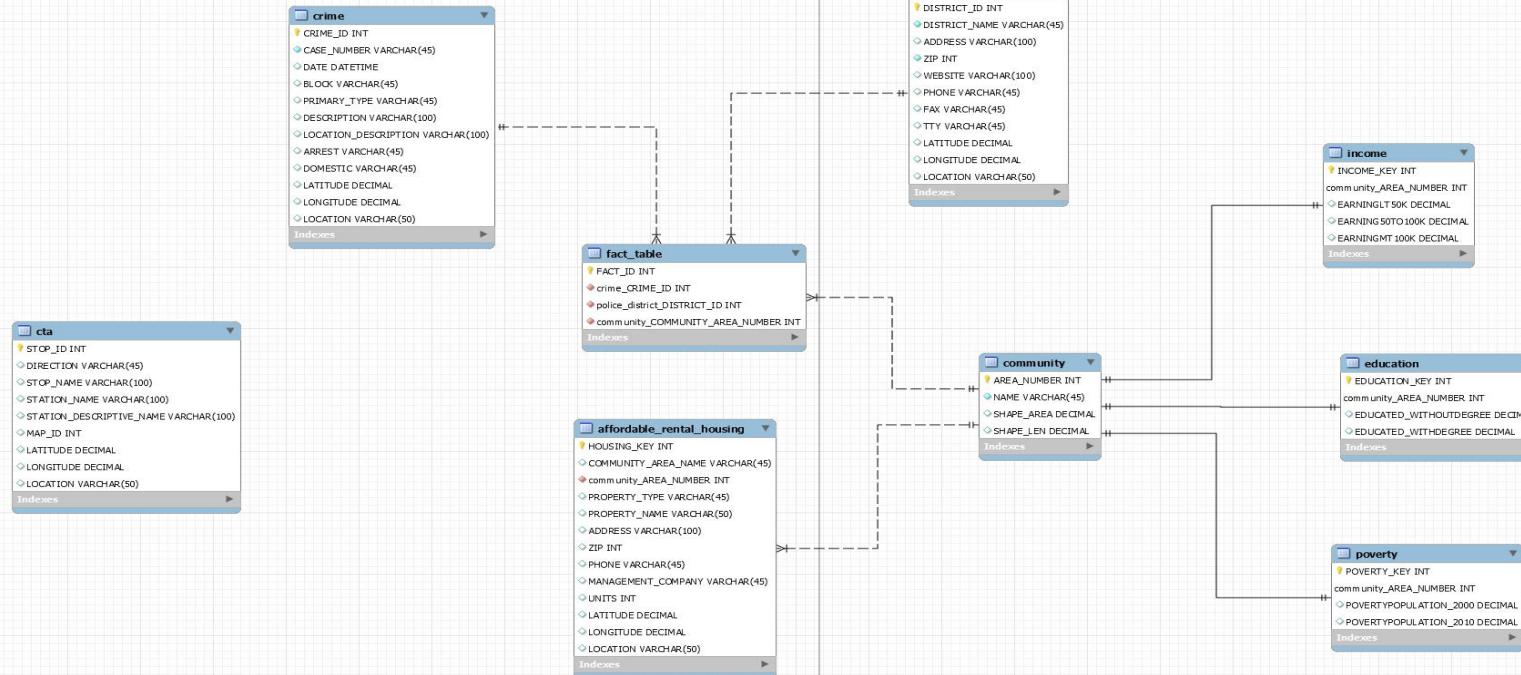
+

Relational Model

EER



Dimensional Model



DATA STORAGE & CLOUD

Google Cloud Platform Data Engineering Platform

Storage

Bucket details

definalproject

OBJECTS CONFIGURATION PERMISSIONS RETENTION LIFECYCLE

Buckets > definalproject

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS DOWNLOAD DELETE

Filter by object or folder name prefix

Name	Size	Type
Affordable_Rental_Housing_Developments_cleaned.csv	150.4 KB	application/octet-stream
CTA_cleaned.csv	35.8 KB	application/octet-stream
Community-Area-Code-cleaned.csv	3.3 KB	application/octet-stream
Crimes2020_cleaned22.csv	27.9 MB	application/octet-stream
Education.csv	1.5 KB	application/octet-stream
FACTDOL.sql	8.3 KB	application/octet-stream
Incomer.csv	2.1 KB	application/octet-stream
Police_Stations_cleaned.csv	4.9 KB	application/octet-stream
Poverty.csv	1.5 KB	application/octet-stream

SQL

MASTER INSTANCE

All instances > fproject

fproject

MySQL 8.0

CPU utilization

1 hour 6 hours 1 day 7 days 30 days

10:15 10:20 10:25 10:30 10:35 10:40

CPU utilization (fproject): 3.80%

Connect to this instance

Public IP address

34.123.125.38

Databases

MASTER INSTANCE

All instances > fproject

fproject

MySQL 8.0

CREATE DATABASE

Schemas

finalproject

Tables

affordable_rental_housing

community

da

education

income_cleaned

poverty

Stored Procedures

Functions

Triggers

Views

materialized_view

sys

Tables

finalproject

information_schema

mysql

performance_schema

rawdata

sys

Name Collation Character set Type

finalproject utf8_general_ci utf8 User

information_schema utf8_general_ci utf8 System

mysql utf8_general_ci utf8 System

performance_schema utf8mb4_0900_ai_ci utf8mb4 System

rawdata utf8_general_ci utf8 User

sys utf8mb4_0900_ai_ci utf8mb4 User

Core ID CASE NUMBER DATE INCIDENT PRIMARY_TYPE DESCRIPTION LOCATION_DESCRIPTION LATITUDE LONGITUDE ARREST DOMESTIC INCIDENT_ID

24957 3033766 2020-02-16 16:15:00 06/01 IN REEDWOOD AVE HOMICIDE FIRST DEGREE MURDER GAS STATION True False 41.5183 -87.7614 (41.51831121, -87.76142962)

24960 3033780 2020-02-24 21:45:00 04/01 IN FRANCISCO AVE HOMICIDE FIRST DEGREE MURDER APARTMENT True False 41.5183 -87.7058 (41.51831661, -87.70584775)

25913 3033803 2020-02-25 10:00:00 02/01 IN 11TH AVE HOMICIDE FIRST DEGREE MURDER STREET True False 41.5179 -87.6951 (41.51791101, -87.69512039)

25917 3033813 2020-02-29 10:00:00 02/01 IN EVERGREEN AVE HOMICIDE FIRST DEGREE MURDER APARTMENT True True 41.5157 -87.6951 (41.51570873, -87.69522388)

25920 3033823 2020-02-29 18:45:00 06/01 IN HAWAII AVE HOMICIDE FIRST DEGREE MURDER STREET True False 41.5159 -87.6911 (41.51591517, -87.69109813)

25923 3033833 2020-03-01 00:00:00 02/01 IN 11TH AVE HOMICIDE FIRST DEGREE MURDER STREET True False 41.5159 -87.6911 (41.51591517, -87.69109813)

25927 3033851 2020-03-23 21:30:00 06/01 IN NORTH AVE HOMICIDE FIRST DEGREE MURDER STREET True False 41.51 -87.7117 (41.51003010, -87.71371712)

25929 3033861 2020-03-24 00:00:00 06/01 IN 11TH AVE HOMICIDE FIRST DEGREE MURDER STREET True False 41.51 -87.7117 (41.51003010, -87.71371712)

25930 3033868 2020-03-19 21:30:00 06/01 IN ALICE AVE HOMICIDE FIRST DEGREE MURDER STREET True False 41.5179 -87.7403 (41.51790303, -87.74032449)

25942 3032966 2020-02-22 11:30:00 06/01 S WOOD ST HOMICIDE FIRST DEGREE MURDER STREET True True 41.5267 -87.6674 (41.52600955, 61.61790593)

25943 3032967 2020-02-22 11:30:00 06/01 S WOOD ST HOMICIDE FIRST DEGREE MURDER STREET True True 41.5267 -87.6674 (41.52600955, 61.61790593)

25941 3032957 2020-02-29 11:00:00 03/01 S CALUMET AVE HOMICIDE FIRST DEGREE MURDER APARTMENT True True 41.7077 -87.6177 (41.70790881, 61.61760981)

52579 3032958 2020-02-29 11:00:00 03/01 S CALUMET AVE HOMICIDE FIRST DEGREE MURDER APARTMENT True True 41.7077 -87.6177 (41.70790881, 61.61760981)

25945 3032967 2020-02-09 09:30:00 01/01 S MICHIGAN AVE HOMICIDE FIRST DEGREE MURDER PARK PROPERTY True False 41.5155 -87.6241 (41.51540257, -87.62409212)

15289 3032952 2020-02-07 06:30:00 05/01 S TAHANA AVE HOMICIDE FIRST DEGREE MURDER HOUSE True False 41.7047 -87.6903 (41.70470466, -87.69010119)

25346 3032953 2020-02-16 14:30:00 02/01 W 3RD ST HOMICIDE FIRST DEGREE MURDER RETAIL STORE True False 41.7092 -87.6951 (41.70920205, -87.69520412)

25499 3032954 2020-02-26 21:45:00 06/01 N PALMWOOD RD HOMICIDE FIRST DEGREE MURDER GAS STATION True False 41.5157 -87.7521 (41.51560465, 61.72032363)

25491 3032955 2020-02-23 21:30:00 02/01 N MICHIGAN AVE HOMICIDE FIRST DEGREE MURDER STREET True True 41.5123 -87.6494 (41.51230124, -87.64941514)

DATA STORAGE & CLOUD

Create table

Source

Create table from: Select file: [Crimes20200531.csv](#) File format: [CSV](#)

Destination

Search for a project Enter a project name

Project name Data Engineering Platform Dataset name crime0531 Table type Native table

Table name crime0531

Schema

Auto detect Schema and input parameters Schema will be automatically generated

Partition and cluster settings

Partitioning: No partitioning

Clustering order (optional): Clustering order determines the sort order of the data. Clustering can be used on both partitioned and non-partitioned tables. Comma-separated list of fields to define clustering order (up to 4)

Advanced options

Write preference: Write if empty Number of errors allowed: 0 Unknown values: Ignore unknown values Field delimiter: Comma Header rows to skip: 0 Quoted newlines: Allow quoted newlines Jagged rows: Allow jagged rows

Encryption

Data is encrypted automatically. Select an encryption key management solution.

- Google-managed key No configuration required
- Customer-managed key Manage via Google Cloud Key Management Service

Create dataset

Dataset ID crime0531

Data location (Optional) United States (US)

Default table expiration

Never Number of days after table creation:

Encryption

Data is encrypted automatically. Select an encryption key management solution.

- Google-managed key No configuration required
- Customer-managed key Manage via Google Cloud Key Management Service

HIDE PREVIEW FEATURES

RUN **SAVE** **SCHEDULE** **MORE**

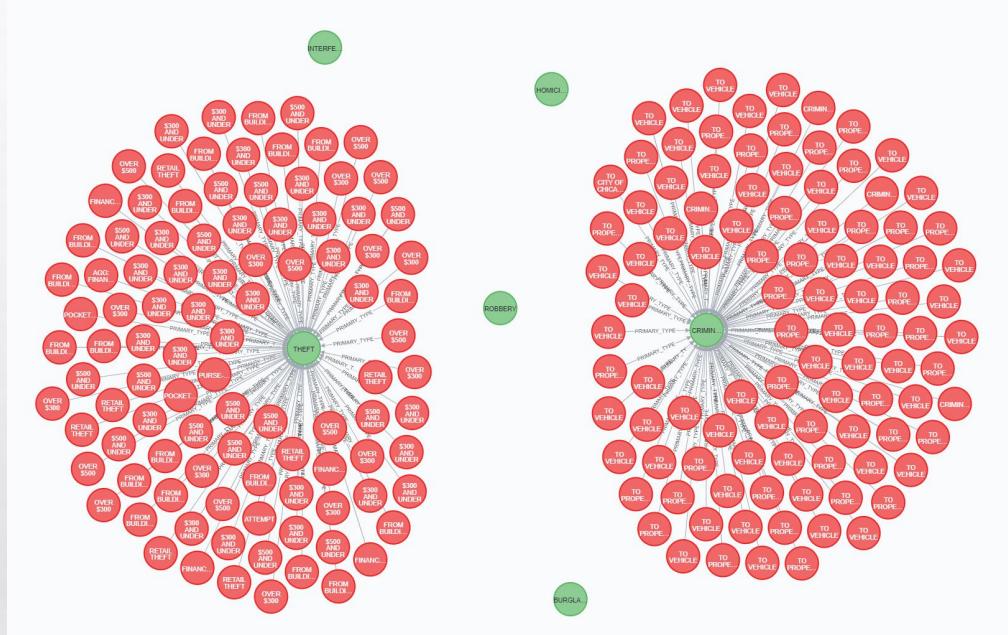
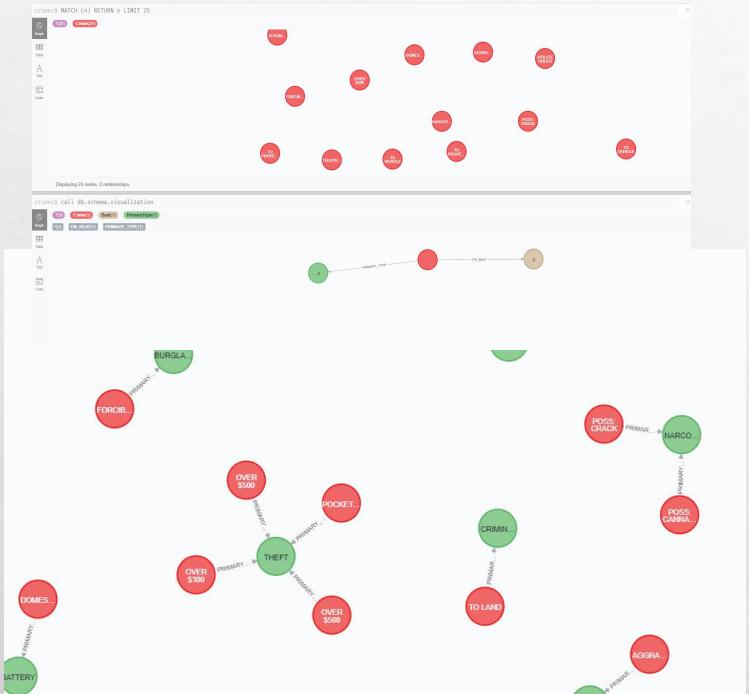
```
1 SELECT CRIMED, CASE_NUMBER, DATE, BLOCK, PRIMARY_TYPE, LOCATION_DESCRIPTION FROM `white-defender-292422.crime0531.crimes0531` LIMIT 1000
```

Query results **SAVE RESULTS** **EXPLORE DATA**

Query complete (0.2 sec elapsed, 143.2 KB processed)

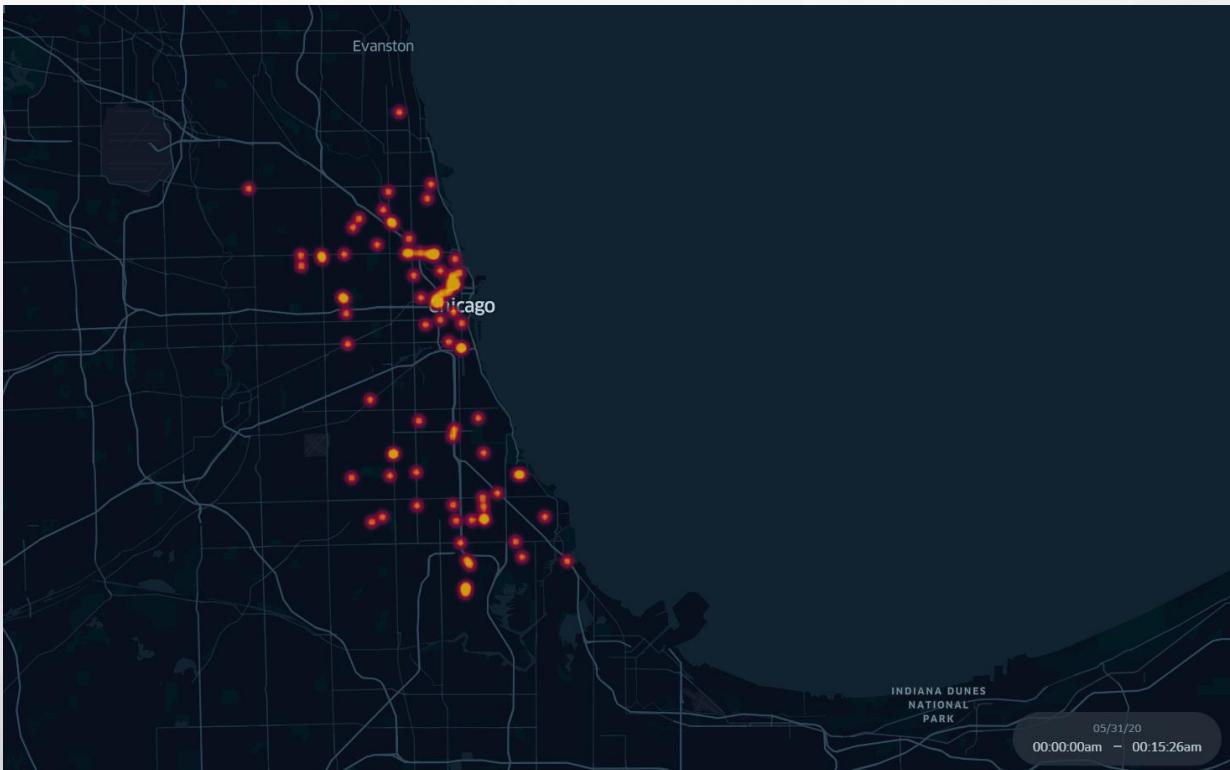
Row	CRIMED	CASE_NUMBER	DATE	BLOCK	PRIMARY_TYPE	LOCATION_DESCRIPTION
1	12064150	JD248280	2020-05-31 04:25:00 UTC	013XX W LUNT AVE	OTHER OFFENSE	APARTMENT
2	12065835	JD249411	2020-05-31 20:45:00 UTC	077XX N HASKINS AVE	CRIMINAL DAMAGE	CTA TRAIN
3	12070443	JD252446	2020-05-31 15:30:00 UTC	040XX N BROADWAY	DECEPTIVE PRACTICE	STREET
4	12068164	JD252901	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	039XX 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	12065997	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	12069981	JD248157	2020-05-31 02:25:00 UTC	000XX W GRAND AVE	PUBLIC PEACE VIOLATION	OTHER (SPECIFY)
16	12065949	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	057XX N ODELL AVE	OTHER OFFENSE	RESIDENCE - PORCH / HALLWAY
19	12065256	JD249559	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

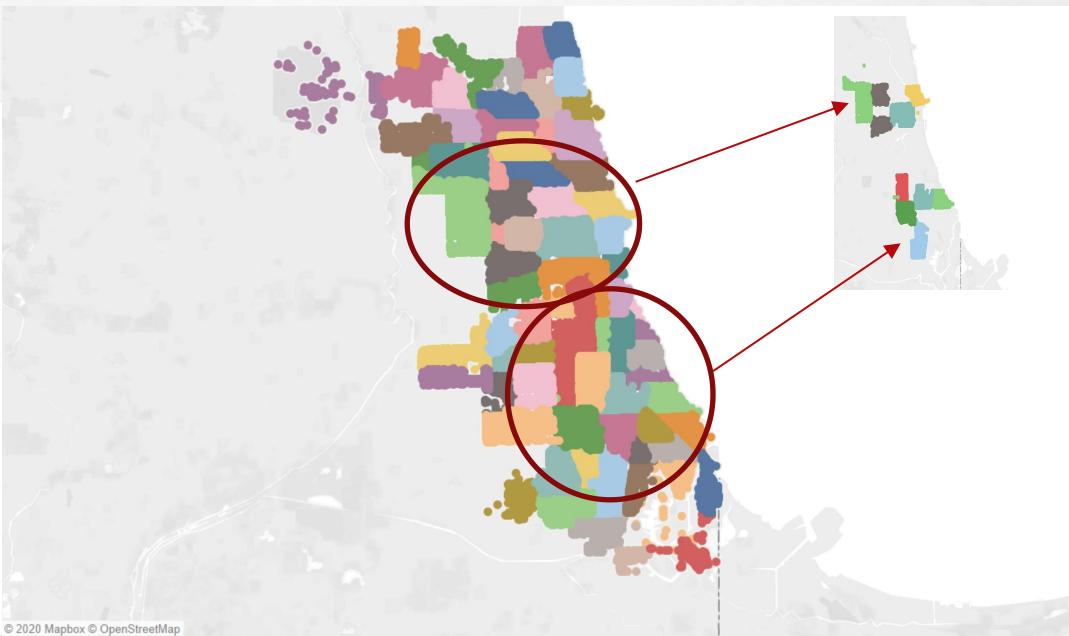
Heat for Every 15 Minutes



Major Crime Prone Areas in Chicago

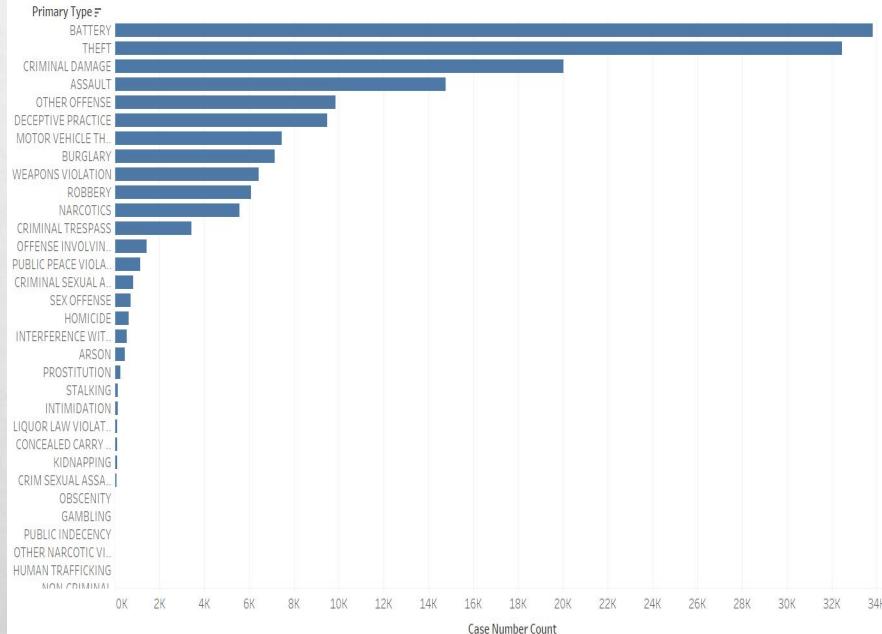
Major Crime Prone Areas in Chicago

Community	F
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

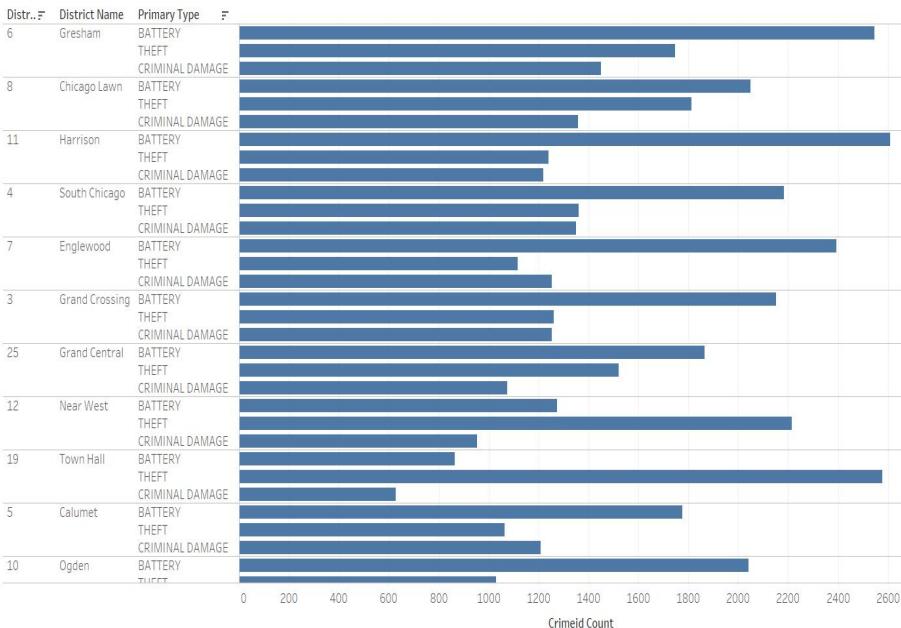


Major Crime Issues in Chicago

Distribution of Major Crime Categories in Chicago



Top 3 Crime Categories in Each Districts

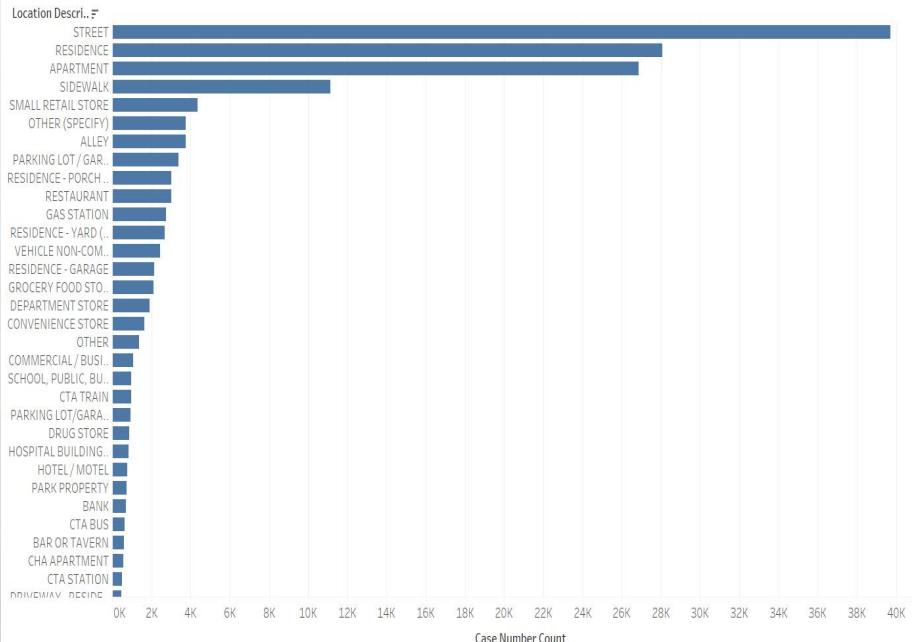


- Major Crime: 1. Battery 2. Theft
3. Criminal Damage 4. Assault

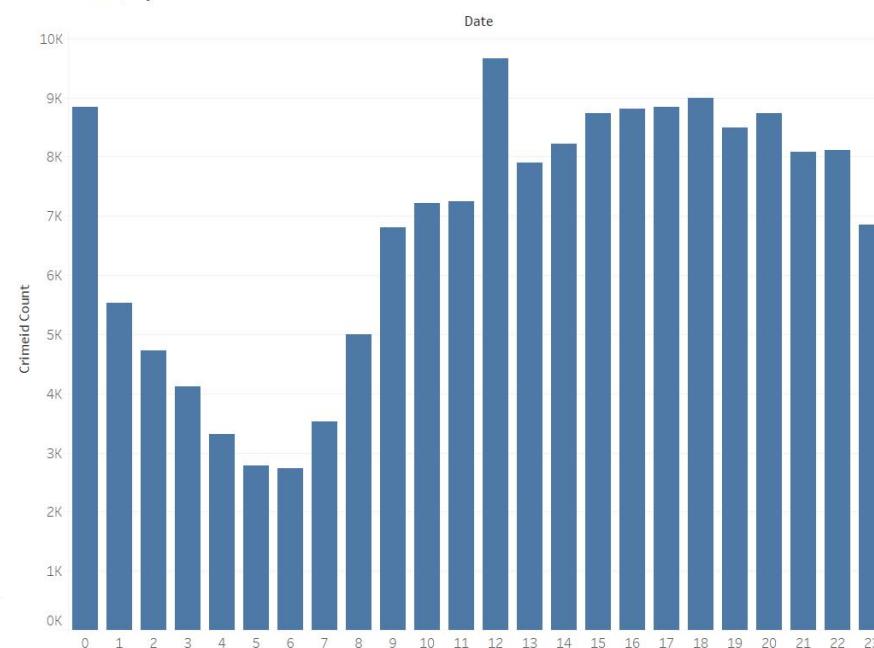
- “Battery” : the actual physical violence causing the physical harm.
- “Assault”: the act which causes the victim to apprehend imminent physical harm
- “Criminal Damage”: unauthorized destruction of public facilities or personal property

Major Crime Prone Locations and Time in Chicago

Location Description and Case Count



Case Count, by Hour



- 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  }
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 {
89   "Crime2020_dim": {
90     "nInserted": 163360,
91     "nModified": 0,
92     "nSkipped": 0,
93     "failed": 0
94   }
95 }
```

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 /* 1 */
2 {
3   "_id": 11,
4   "count": 11619
5 },
6
7 /* 2 */
8 {
9   "_id": 6,
10  "count": 11083
11 },
12
13 /* 3 */
14 {
15   "_id": 8,
16   "count": 10153
17 },
18
19 /* 4 */
20 {
21   "_id": 4,
22   "count": 9501
23 },
24
25 /* 5 */
26 {
27   "_id": 7,
28   "count": 9387
29 },
30
```

Output

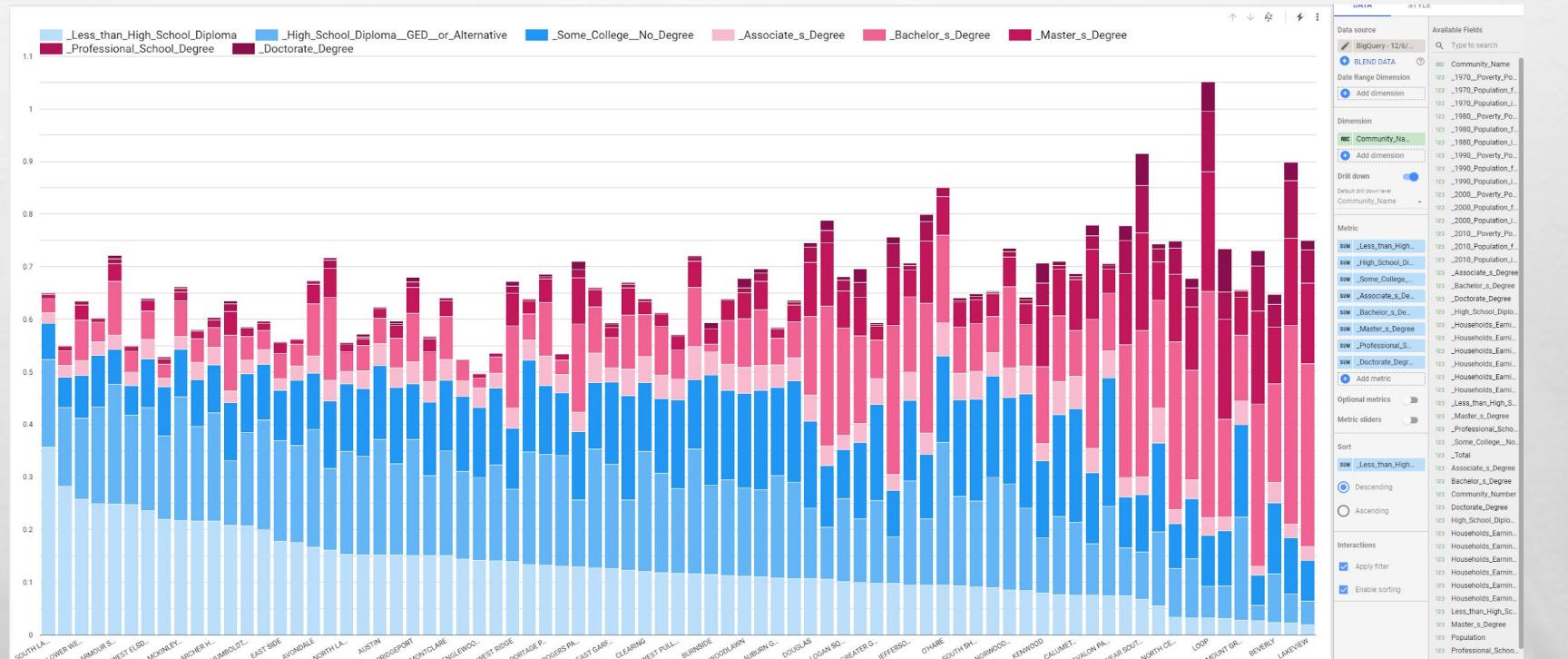
```
db.Crimes2020_dim.aggregate([{$group: { _id: "$DISTRICT",
    count: { $sum: 1 } }}, {$sort:{"count": -1}}])
```

Crimes2020_dim | 0.859 s | 77 Docs

```
1 /* 1 */
2 {
3   "_id": 25,
4   "count": 9965
5 },
6
7 /* 2 */
8 {
9   "_id": 43,
10  "count": 6118
11 },
12
13 /* 3 */
14 {
15   "_id": 8,
16   "count": 5736
17 },
18
19 /* 4 */
20 {
21   "_id": 29,
22   "count": 5573
23 },
24
25 /* 5 */
26 {
27   "_id": 23,
28   "count": 5309
29 },
30
```

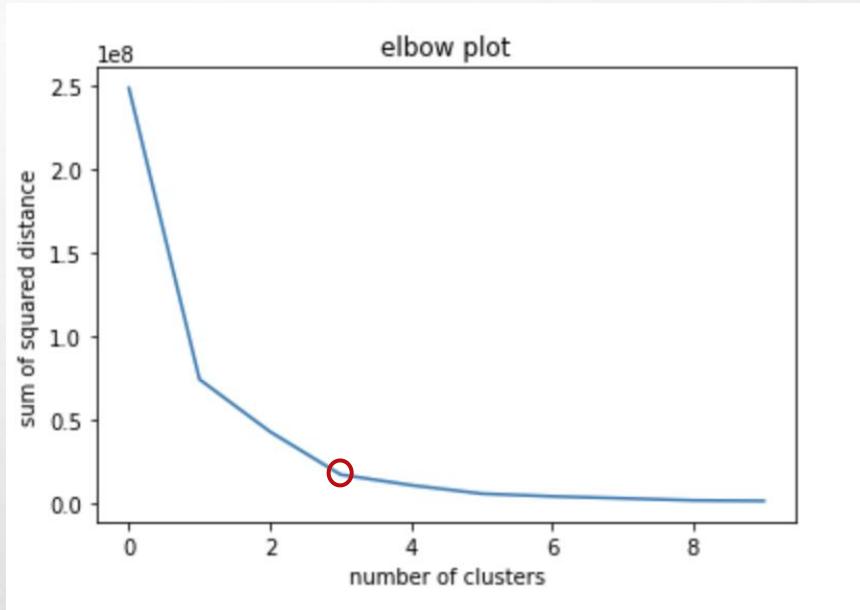
Crimes2020_dim | 0.658 s | 77 Docs

GCP BigQuery & Data Visualization Studio



Clustering: K-Means

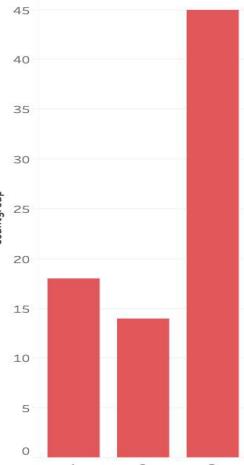
- **Crime:**
Community Area
Crime Count
- **Poverty:**
2010 % Poverty Population
- **Education:**
% Educated Population with No Degree
% Educated Population with No Degree
- **Income:**
% Households Earning less than \$49,999
% Households Earning between \$50,000 and \$99,999
% Households Earning more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



Crime Demographics & Group Comparison

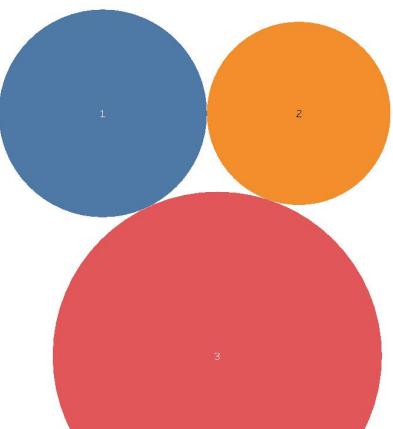
Count of Each Group

Group Number



Group count

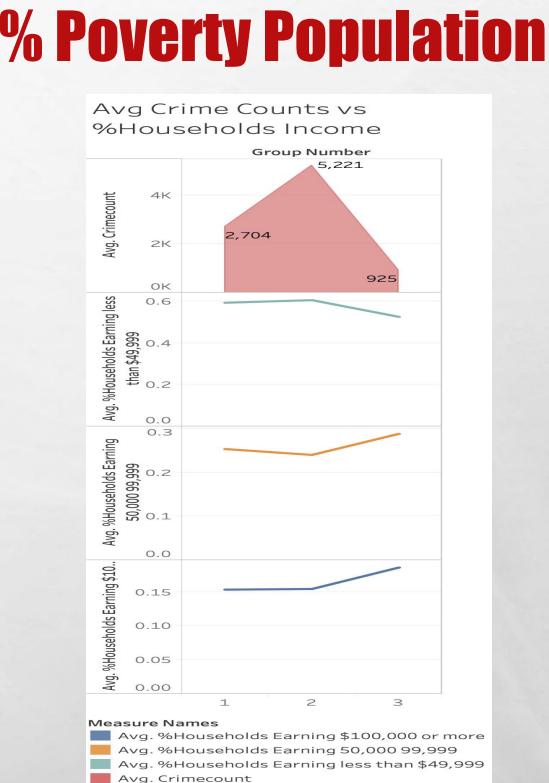
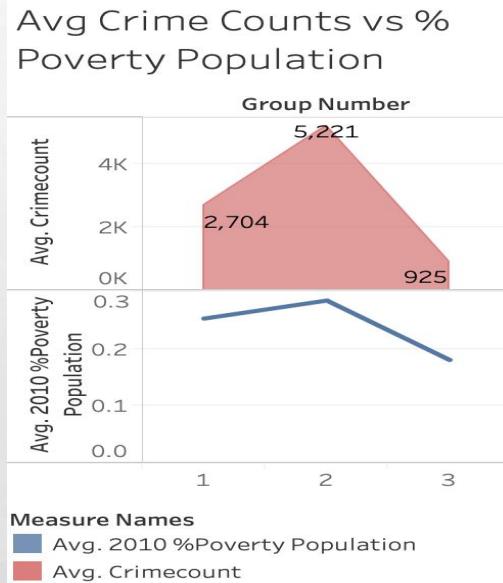
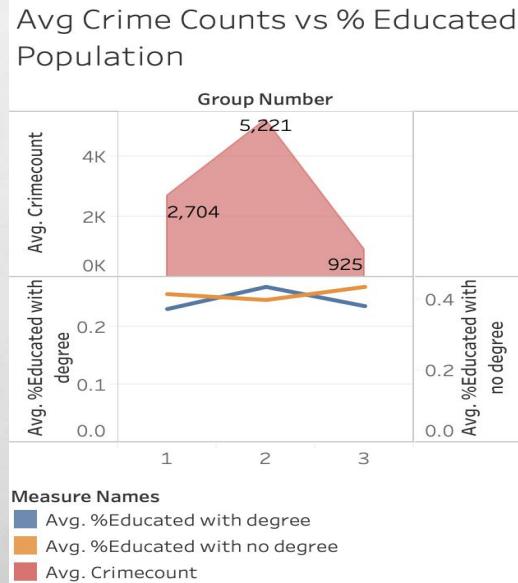
Group Number
■ 1
■ 2
■ 3



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, & 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!**



28



APPENDIX



DATA SOURCES

DATA	DESCRIPTION
<u>Chicago Public Safety – Crimes</u>	Chicago city community level data of crimes such as crime time, type, location description and community area.
<u>Chicago Public Safety - Police Stations</u>	Chicago city community level data of police stations such as district, website and location.
<u>Chicago Transportation - Metra Stations</u>	Chicago city community level data of metra stations such as stop name and location.
<u>Chicago Housing - Chicago Neighbourhoods</u>	Chicago city community level data of housing such as property type, address, units and location.
<u>Chicago Demographic-Chicago Neighbourhoods</u>	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

```
22 # Find out top 10 blocks which have the most crime records
23 • SELECT
24     BLOCK, COUNT(DISTINCT (CRIME_ID)) AS CASES
25 FROM
26     finalproject.crime
27 GROUP BY BLOCK
28 ORDER BY CASES DESC
29 LIMIT 10;
30
```

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

```
22 # Find out top 10 blocks which have the most crime records
23 • SELECT
24     BLOCK, COUNT(DISTINCT (CRIME_ID)) AS CASES
25 FROM
26     finalproject.crime
27 GROUP BY BLOCK
28 ORDER BY CASES DESC
29 LIMIT 10;
30
```

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

```
5 # Find out the most battery happened in which police district order by the amount of batteries
6 • SELECT
7     pd.DISTRICT_NAME,
8     COUNT(DISTINCT (c.CRIME_ID)) AS BATTERY_Number
9 FROM
10    finalproject.crime AS c
11    JOIN
12    finalproject.fact_table AS ft
13    JOIN
14    finalproject.police_district AS pd
15 WHERE
16    c.CRIME_ID = ft.crime_CRIME_ID
17    AND ft.police_district_DISTRICT_ID = pd.DISTRICT_ID
18    AND c.PRIMARY_TYPE = 'BATTERY'
19 GROUP BY ft.police_district_DISTRICT_ID
20 ORDER BY BATTERY_Number DESC;
```

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