CIS5200 Term Project Tutorial

Authors: **Hai Anh Le, Neha Shashidhara Guli, Jeevan Sai**

Instructor: **Jongwook Woo**

Date: 12/4/2017

Lab Tutorial

Hai Anh Le (hle55@calstatela.edu)
Neha Shashidhara Guli (sneha@calstatela.edu)
Jeevan Sai (jaddaga@calstatela.edu)
12/4/2017

Crime Data Analysis of Chicago & New York

OBJECTIVE:

In our project, we have analysed and visualized the crime rate in two major cities in USA – CHICAGO and NEW YORK.

Analysis of the data is majorly centred around the location of occurrence of crime (i.e. street, building, hotels, etc.), type of crime (Misdemeanour, felony or violation) and the locality where the crime occurred in both the cities.

The New York dataset challenge that this project was done on contains 11.9K views with 2765 downloads from 2006 to the end of last year (2016) with 5.58M rows and 24 columns and the Chicago dataset contains 349K views with 474K downloads with 6.48M rows and 22 columns.

Here we have considered 3 major crime types in both the cites and have done the data cleaning accordingly.
The analysis performed on New York and Chicago crime dataset are:

- Query to find out where crimes take place more in New York and Chicago- top 3 locations.
- Query to find the top 5 cities with highest crimes recorded
- Query to categorise the crimes in the top city
- Crime rate yearly in both the cites
- Trend of crime
- Geo map to show the crimes in different localities in both the cities

INTRODUCTION

This Project aims at performing data analysis and providing insights on Crime dataset of Chicago and New York using HIVE and presenting the visualization in Tableau.

In this tutorial, through each analysis we did you'll learn how to use BigInsights to:

- Load data from local desktop(windows) to Linux shell
- Extract CSV file using Hive
- Data cleaning using Hive
- Download and upload files to HDFS
- Create Hive tables to query the crime dataset for analysis
- Create Hive queries to analyze the sentiment of data
- Use Tableau for visualization of the analyzed data and also to present the forecast of crime data.

PREREQUISITES:

Everything you need to go through the scripts and queries is already provisioned with the cluster. To export the analysed data to Microsoft Excel and to do the visualization you must meet the following requirements:

- You must have Microsoft Excel 2010, 2013 or 2016 installed.
- Tableau 9.2 or 9.3 installed for visualization of the analysed data.
- IBM Bluemix Cluster version: IOP4.2

Crime Dataset Loaded into BigInsights

You need to remotely access your BigInsights that you executed in your Bluemix account using *ssh*. Below is the location of the Crime data that is used for this sample. You can download the crime data file of New York and Chicago:

```
wget -O Newyork_min_data.csv
https://data.cityofnewyork.us/api/views/qgea-
i56i/rows.csv?accessType=DOWNLOAD&bom=true&format=true/Newyork_min_data.csv

wget -O chicago_min_data.csv
https://data.cityofchicago.org/api/views/ijzp-
q8t2/rows.csv?accessType=DOWNLOAD&bom=true&format=true/chicago_min_data.csv
```

```
.ast login: Sat Dec  2 16:51:32 on ttys000
Nehas-MacBook-Air:~ nehasguli$ ssh nehaguli@bi-hadoop-prod-4025.bi.services.us-south.bluemix.net
nehaguli@bi-hadoop-prod-4025.bi.services.us-south.bluemix.net's password:
IBM's internal systems must only be used for conducting IBM's business or for purposes authorized by IBM management
Use is subject to audit at any time by IBM management
-bash-4.1$ wget -O Newyork_min_data.csv https://data.cityofnewyork.us/api/views/qgea-i56i/rows.csv?accessType=DOWNLOAD&bom=true&format=true/Newyo
[1] 15055
[2] 15056
-bash-4.1$ --2017-12-03 00:53:58--  https://data.cityofnewyork.us/api/views/qgea-i56i/rows.csv?accessType=DOWNLOAD
Resolving data.cityofnewyork.us... 52.206.140.199
Connecting to data.cityofnewyork.us|52.206.140.199|:443... connected.
HTTP request sent, awaiting response... 200 OK
.ength: unspecified [text/csv]
Saving to: "Newyork_min_data.csv"

    [                       <=>           ] 1,427,270,109 2.58M/s   in 8m 47s

.ast-modified header invalid -- time-stamp ignored.
2017-12-03 01:02:45 (2.58 MB/s) - "Newyork_min_data.csv" saved [1427270109]


bash-4.1$ wget -O chicago_min_data.csv https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD&bom=true&format=true/chicago_min_data.csv
1] 21257
2] 21258
bash-4.1$ --2017-12-03 01:25:07--  https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD
esolving data.cityofchicago.org... 52.206.140.205
onnecting to data.cityofchicago.org|52.206.140.205|:443... connected.
TTP request sent, awaiting response... 200 OK
ength: unspecified [text/csv]
aving to: "chicago_min_data.csv"

   [                                                                                <=>

ast-modified header invalid -- time-stamp ignored.
017-12-03 01:35:08 (2.42 MB/s) - "chicago_min_data.csv" saved [1526451956]
```

Before we upload these files to HDFS we have to create a directory in HDFS. Run the following

HDFS commands to create to store these files into a directory in HDFS.

```
hdfs dfs -mkdir NYPD
hdfs dfs -put Newyork_min_data.csv NYPD
hdfs dfs -ls

hdfs dfs -mkdir Chicago
hdfs dfs -put chicago_min_data.csv Chicago
hdfs dfs -ls
```

```
-bash-4.1$ hdfs dfs -put Newyork_min_data.csv NYPD
-bash-4.1$ hdfs dfs -ls
Found 4 items
drwx------   - nehaguli hdfs          0 2017-11-23 06:00 .Trash
drwx------   - nehaguli hdfs          0 2017-11-22 22:57 .staging
drwxr-xr-x   - nehaguli hdfs          0 2017-12-03 01:11 NYPD
drwxr-xr-x   - nehaguli hdfs          0 2017-11-22 22:45 dualcore
```

```
dfs dfs -mkdir Chicago
1]-  Done                    wget -O chicago_min_data.csv https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD
2]+  Done                    bom=true
bash-4.1$ hdfs dfs -put chicago_min_data.csv Chicago
bash-4.1$ hdfs dfs -ls
ound 5 items
rwx------   - nehaguli hdfs          0 2017-11-23 06:00 .Trash
rwx------   - nehaguli hdfs          0 2017-11-22 22:57 .staging
rwxr-xr-x   - nehaguli hdfs          0 2017-12-03 01:41 Chicago
rwxr-xr-x   - nehaguli hdfs          0 2017-12-03 01:11 NYPD
rwxr-xr-x   - nehaguli hdfs          0 2017-11-22 22:45 dualcore
```

Creating Hive table to Query Crime data

The following Hive statement creates an external table that allows Hive to query data stored in HDFS . External tables preserve the data in the original file format, while allowing Hive to perform queries against the data within the file.

The Hive statements below create two new tables, named **Chicago** and **NYPD** , by describing the fields within the files, the delimiter (comma) between fields, and the location of the file in Azure Blob Storage. This will allow you to create Hive queries over your data.

Open hive shell environment as follows:

```
$ HIVE
```

In the hive shell CLI, you need to copy and paste the following HiveQL code to create an external table ‘chicago’ and ‘nypd’

```
DROP TABLE IF EXISTS chicago;

CREATE EXTERNAL TABLE IF NOT EXISTS chicago (id BIGINT,
casenumber STRING, rgdate STRING, block STRING, IUCR STRING, primarytype STRING, lawcat
STRING, locationdescription STRING, year STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY
','
STORED AS TEXTFILE LOCATION '/tmp/chicago'
TBLPROPERTIES ('skip.header.line.count'='1');

DROP TABLE IF EXISTS nypd;

CREATE EXTERNAL TABLE IF NOT EXISTS nypd (cmplnt_num BIGINT,
cmplnt_fr_dt STRING, rpt_dt STRING, ky_cd BIGINT, ofns_desc STRING, law_cat_cd STRING, boro_nm
STRING, loc_of_occur_desc STRING, prem_typ_desc STRING) ROW FORMAT DELIMITED FIELDS
TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/tmp/nypd'
TBLPROPERTIES ('skip.header.line.count'='1');
```

Then, in the hive shell, you need to check if the table **“**chicago**” and “** nypd**”** is shown:

```
Hive> show tables;
```

```
hive> show tables;
OK
cart_items
cart_zipcodes
checkout_sessions
chicago
nypd
web_logs
Time taken: 0.051 seconds, Fetched: 6 row(s)
hive>
```

Creating Hive Queries to Analyze Data

The following Hive queries analyses the crime data in both the cities.

CHICAGO dataset

1. Query to categorise crimes into Felony, Violation and Misdemenor in Chicago

```
DROP TABLE IF EXISTS chicago1;

CREATE  TABLE chicago1 row format delimited fields
terminated by ','
stored as textfile location '/tmp/chicago/' as
select *,
if Primary Type == 'Arson' | 'Carry license violation' | 'Criminal trespass' | 'Gambling' |
'Public peace violation' | 'Liquor law violation' | 'intimidation'
| 'Public indecency', 'VIOLATION',
if Primary Type == 'non criminal Obscenity' | 'Other offense' | 'Ritualism' | 'Robbery' |
'Theft' | 'Burglary' | 'Weapons violation' | 'Other narcotic violation' | 'Deceptive practice'
|'Interference with public' | 'Non criminal', 'MISDEMENOR', 'FELONY')) AS lawcat from
Chicago;
```

2. Query to find out where crimes take place more in chicago top 3 locations

```
select locationdescription, count(iucr), rank() over (order by count(iucr)desc) AS
rank from chicago1
group by locationdescription limit 3;
```

```
Total MapReduce CPU Time Spent: 8 seconds 180 msec
OK
STREET  294157  1
RESIDENCE       188366  2
APARTMENT       90788   3
Time taken: 46.787 seconds, Fetched: 3 row(s)
hive>
```

3. Query to find the top 5 street with highest crimes recorded

```
SELECT block, count (iucr), rank () over (ORDER BY count (iucr) desc) AS rank
from chicago
GROUP BY block limit 5;
```

```
100XX W OHARE ST        2828    1
001XX N STATE ST        2268    2
076XX S CICERO AV       1275    3
035XX S FEDERAL ST      1122    4
0000X N STATE ST        1005    5
Time taken: 46.753 seconds, Fetched: 5 row(s)
```

4. Query to categorise the crims in the top city

```
SELECT lawcat, count (iucr) from chicago1 where block = "100XX W OHARE
ST" GROUP BY lawcat;
```

```
Total MapReduce CPU Time Spent: 5 seconds 710 msec
OK
Felony  795
Misdemeanor     1667
Violation       366
Time taken: 23.111 seconds, Fetched: 3 row(s)
hive>
```

5. Table to display top crime areas

```
create table hicrime row format delimited fields terminated by ',' stored as textfile
location '/tmp/chicago' as select block, count(iucr) as total_crimes, rank() over (order
by count(iucr) desc) as rank from chicago group by block;
```

```
Starting Job = job_1512687529490_0001, Tracking URL = http://bi-hadoop-prod-4154.bi.services.us
-south.bluemix.net:8088/proxy/application_1512687529490_0001/
Kill Command = /usr/iop/4.2.0.0/hadoop/bin/hadoop job  -kill job_1512687529490_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-12-07 23:26:22,408 Stage-1 map = 0%,  reduce = 0%
2017-12-07 23:26:29,282 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.56 sec
2017-12-07 23:26:34,956 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.98 sec
MapReduce Total cumulative CPU time: 8 seconds 980 msec
Ended Job = job_1512687529490_0001
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1512687529490_0002, Tracking URL = http://bi-hadoop-prod-4154.bi.services.us
-south.bluemix.net:8088/proxy/application_1512687529490_0002/
Kill Command = /usr/iop/4.2.0.0/hadoop/bin/hadoop job  -kill job_1512687529490_0002
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-12-07 23:26:48,408 Stage-2 map = 0%,  reduce = 0%
2017-12-07 23:26:52,232 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 3.61 sec
2017-12-07 23:26:58,701 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 7.71 sec
MapReduce Total cumulative CPU time: 7 seconds 710 msec
Ended Job = job_1512687529490_0002
Moving data to: /tmp/chicago
chgrp: changing ownership of '/tmp/chicago': User does not belong to hdfs
Table default.hicrime stats: [numFiles=1, numRows=54146, totalSize=1469023, rawDataSize=1414877
]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.98 sec   HDFS Read: 96228021 HDFS Write: 2
013869 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 7.71 sec   HDFS Read: 2020891 HDFS Write: 14
69102 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 690 msec
OK
Time taken: 68.868 seconds
```

6. Query for displaying the content of the above table

```
Select * from RCFile limit 3;
```

```
OK
100XX W OHARE ST        2828    1
001XX N STATE ST        2268    2
076XX S CICERO AV       1275    3
Time taken: 0.14 seconds, Fetched: 3 row(s)
```

NEW YORK Dataset

1. query to find out where crimes take place more in New York top 3 locations

```
Select PREM_TYP_DESC, count (KY_CD), rank() over (order by
count(KY_CD)desc) AS rank from nypd
group by PREM_TYP_DESC limit 3;
```

```
STREET  320256  1
RESIDENCE - APT. HOUSE  229485  2
RESIDENCE-HOUSE 97445   3
Time taken: 50.136 seconds, Fetched: 3 row(s)
```

2. query to find the top 5 cities with highest crimes recorded

```
SELECT BORO_NM, count (KY_CD), rank () over (ORDER BY count (KY_CD)
desc) AS rank from nypd
GROUP BY BORO  NM limit 5;
```

```
Total MapReduce CPU Time Spent: 9 seconds 60 msec
OK
BROOKLYN        315635  1
MANHATTAN       244739  2
BRONX   227473  3
QUEENS  211958  4
STATEN ISLAND   48743   5
Time taken: 45.539 seconds, Fetched: 5 row(s)
```

3. query to categorise the crimes in the top city

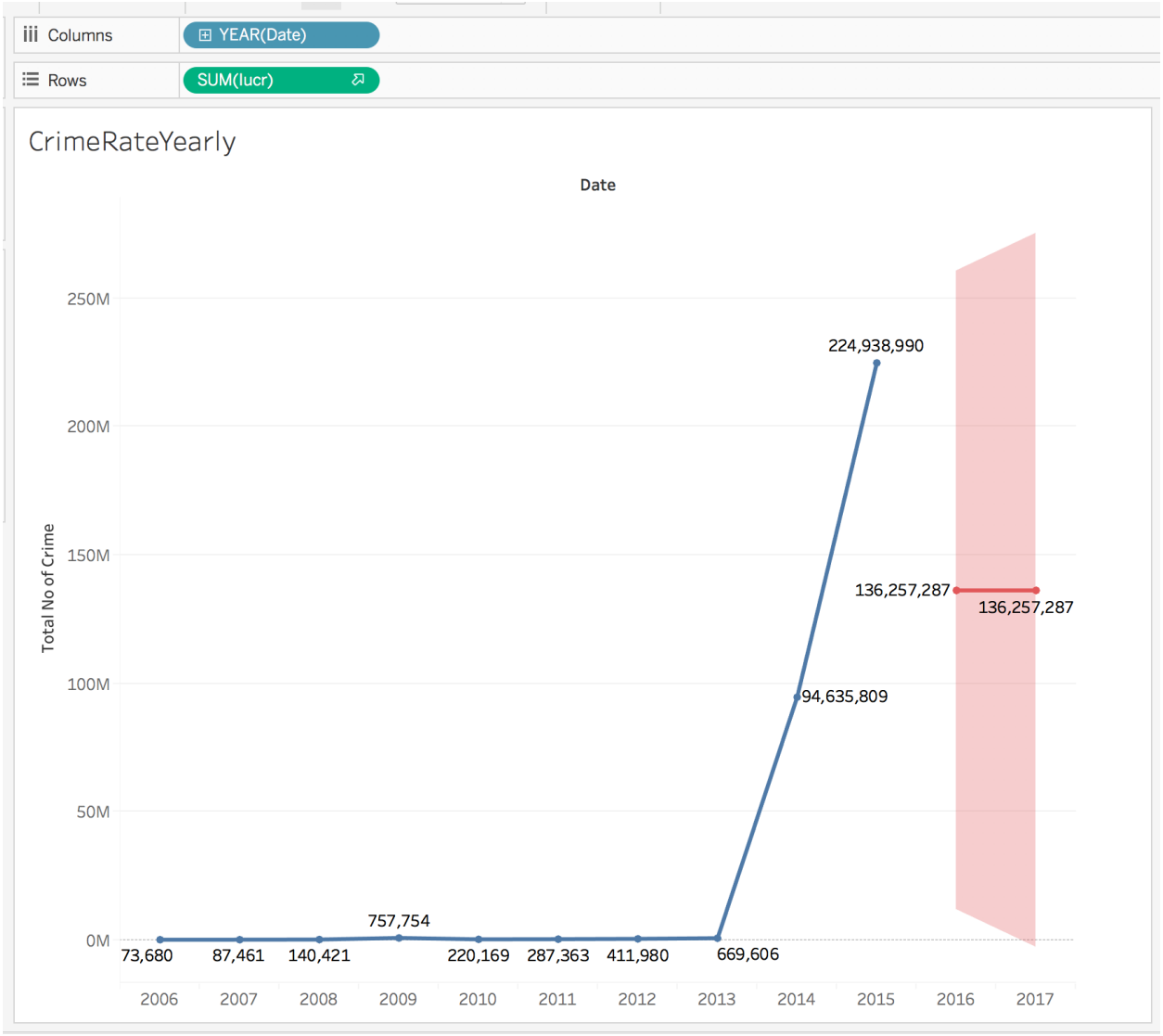
```
SELECT LAW_CAT_CD, count (KY_CD) from nypd where BORO_NM =
"MANHATTAN" GROUP BY LAW_CAT_CD;
```

```
Total MapReduce CPU Time Spent: 5 seconds 900 msec
OK
FELONY  76455
MISDEMEANOR     141374
VIOLATION       26910
Time taken: 24.228 seconds, Fetched: 3 row(s)
```

TABLEAU TO OPEN DATA FILE DIRECTLY FROM TABLEAU AND VISUALIZATION

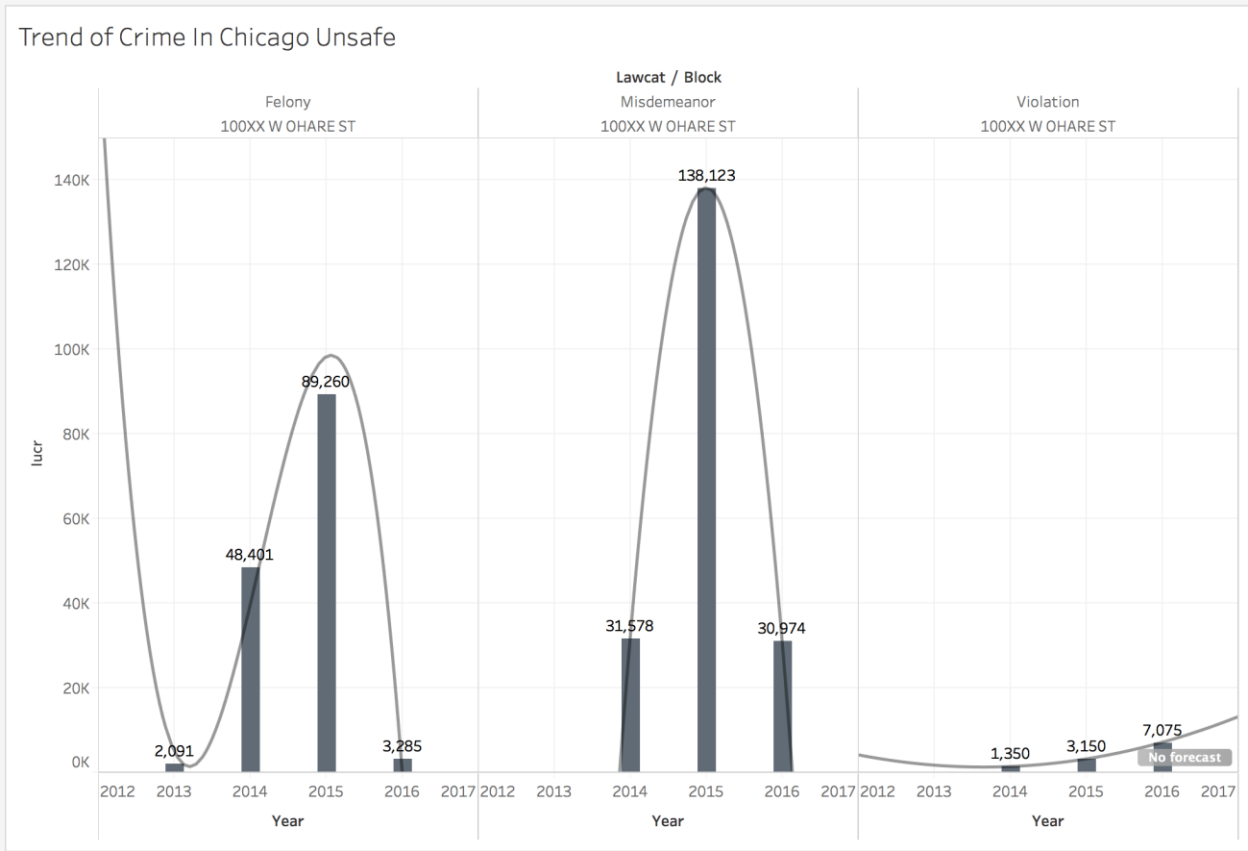
1. Open your Tableau to connect your server. You need to select **Text File** to open the file **chicago_min_data.csv and Newyork_min_data.csv**.

2. Chicago crime forecast



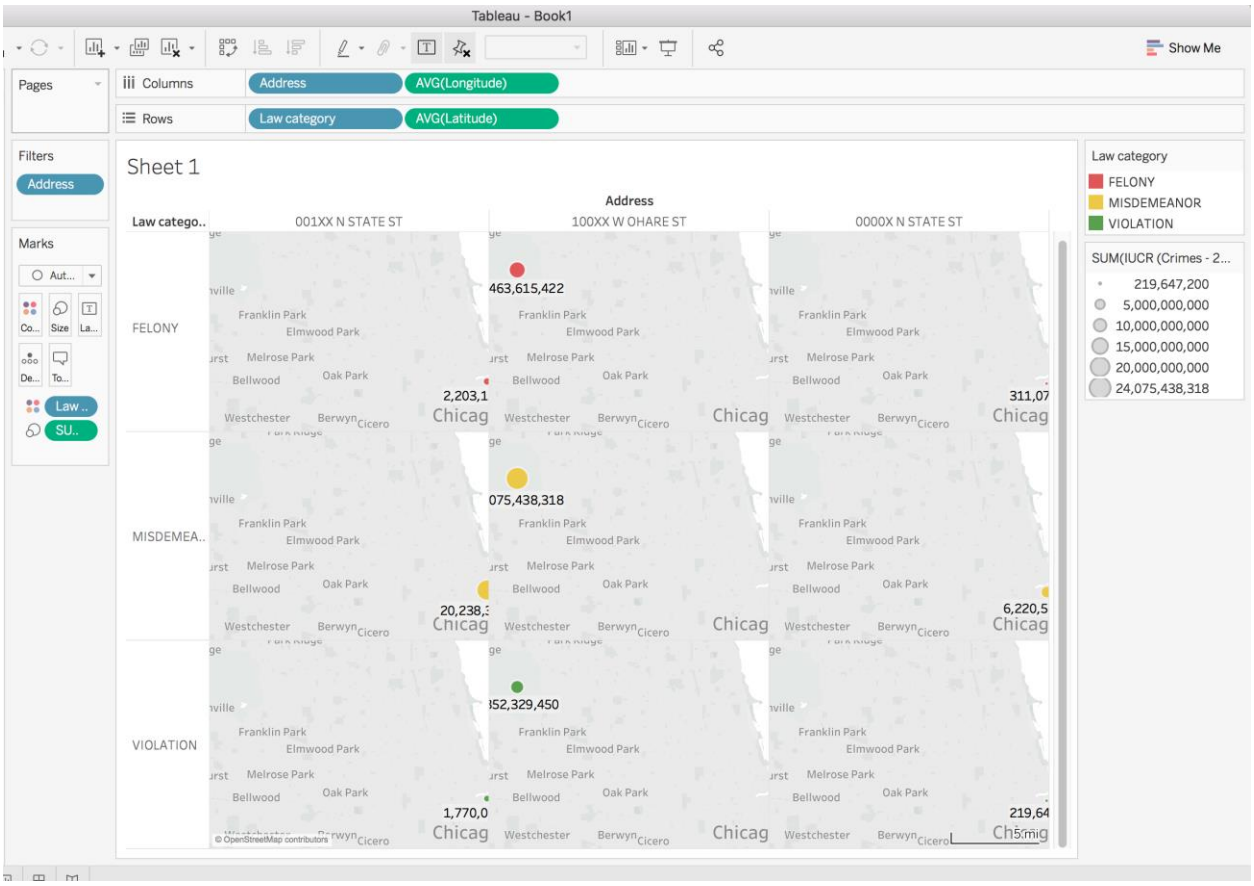
Select Years in columns of tableau and sum of the all the crimes in rows to forecast the crime for Chicago city. Using the analytics option and select the appropriate algorithm for calculating.

3. Chicago Crime trend



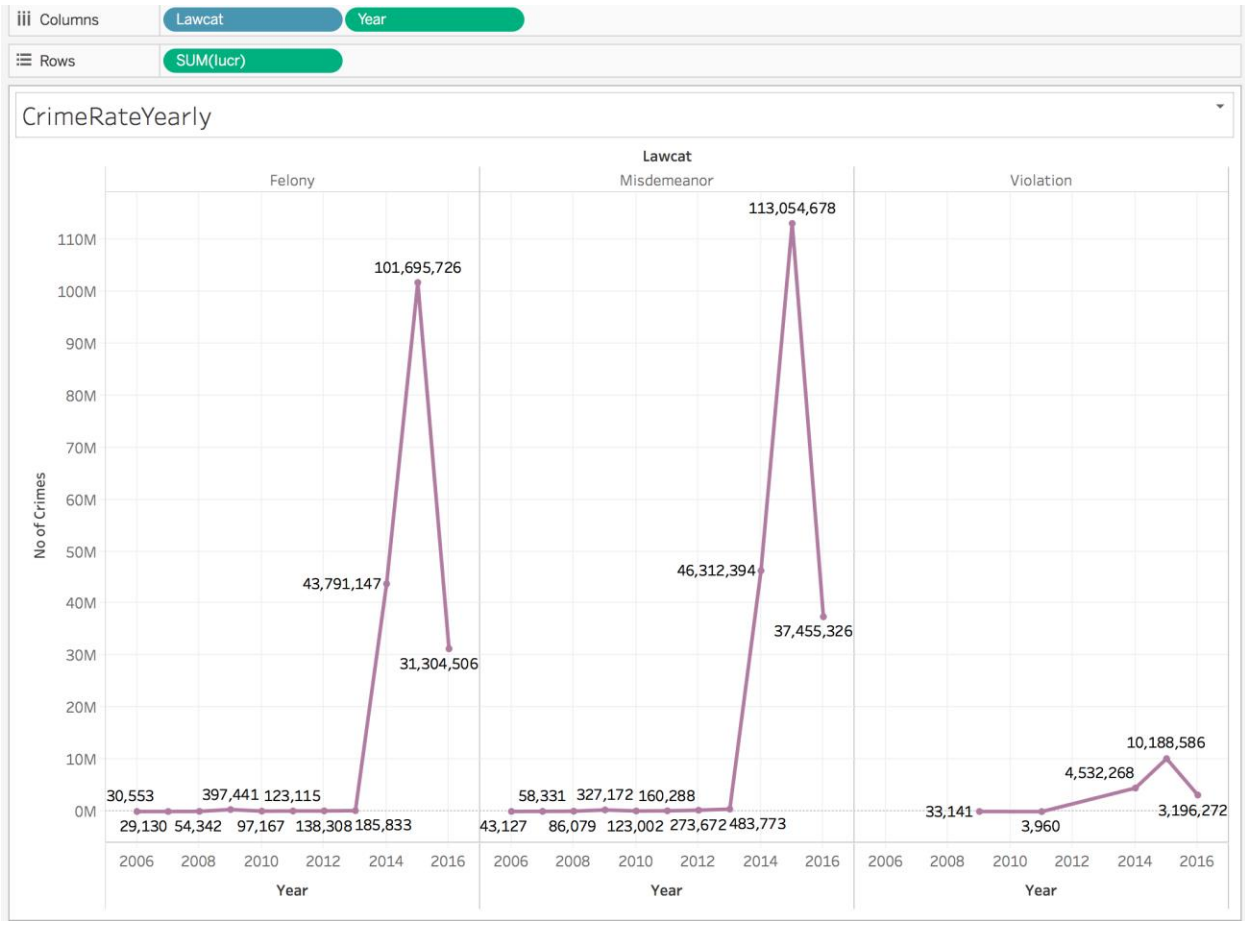
Select the top street which has the highest crime rate by differentiating with the types of crimes to find the trend of different crimes.

4. Chicago Geo map



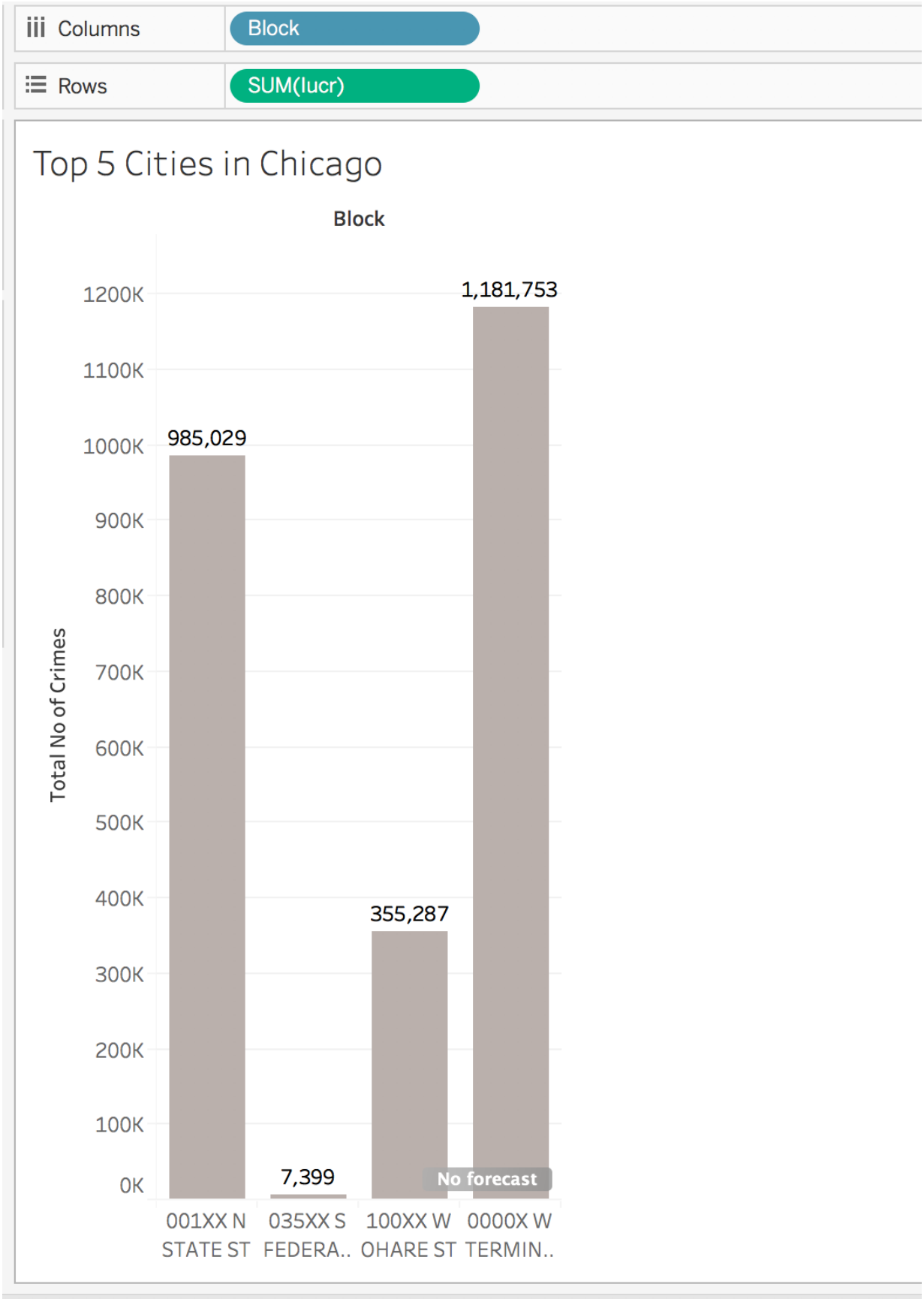
Select the top 3 streets with the count of the total crimes by adding the geo location specifications like latitude and longitude to locate the place in the Map. Different colors are given to different types of crime.

5. Number of crimes yearly



Put the types of crime and years in columns and Sum of the Crimes in rows so find out total number of crimes yearly in the Chicago city.

6. Top cities with highest crime

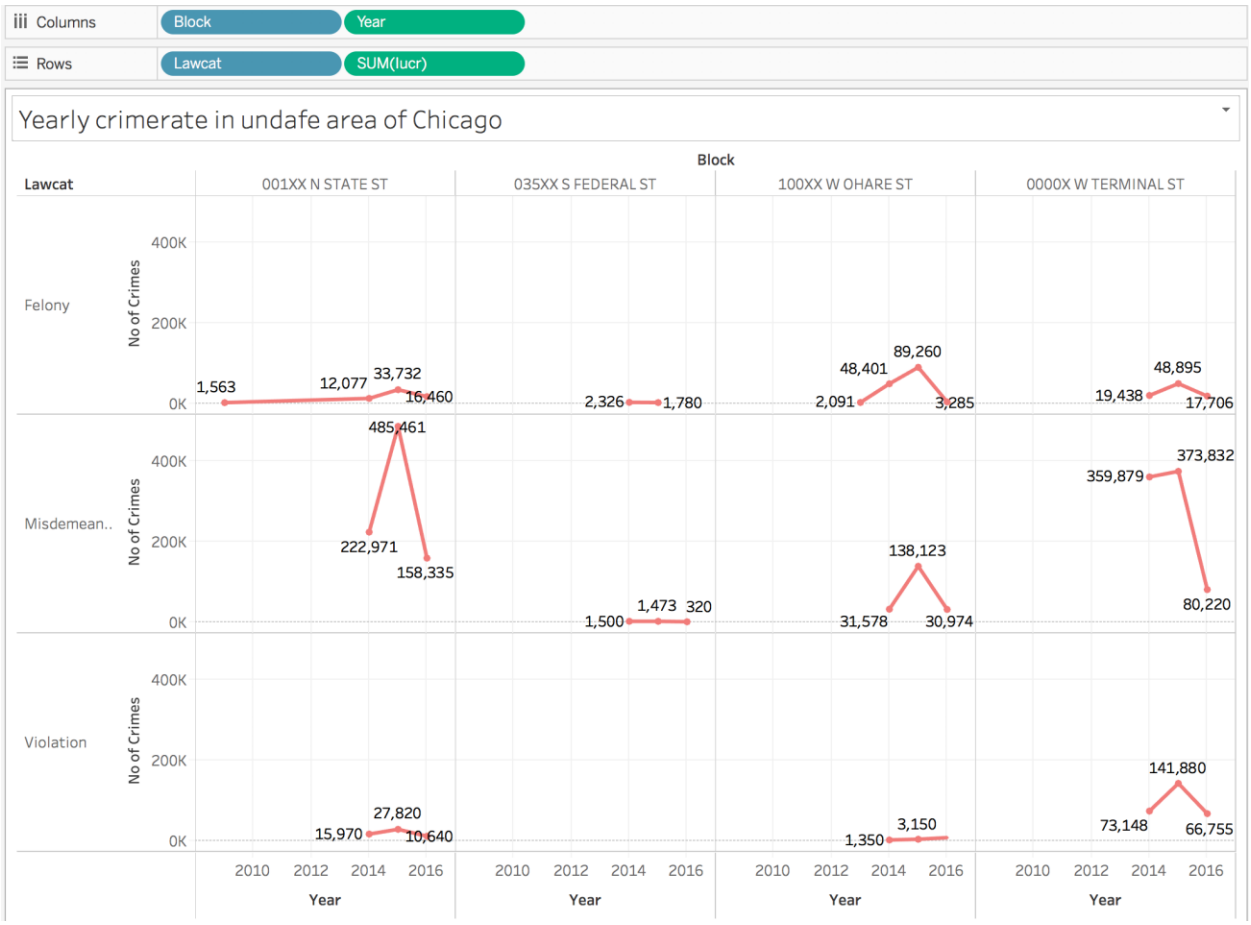

Columns
Block
Rows
SUM(Iucr)
Top 5 Cities in Chicago
Block
Total No of Crimes
1200K
1100K
1000K
900K
800K
700K
600K
500K
400K
300K
200K
100K
0K
985,029
7,399
355,287
1,181,753
No forecast
001XX N STATE ST
035XX S FEDERA..
100XX W OHARE ST
0000X W TERMIN..

Add the Block(which has streets location) in columns and sum of crimes to find the top 4 locations which highest crime rate.

7. Chicago Highest Crime area

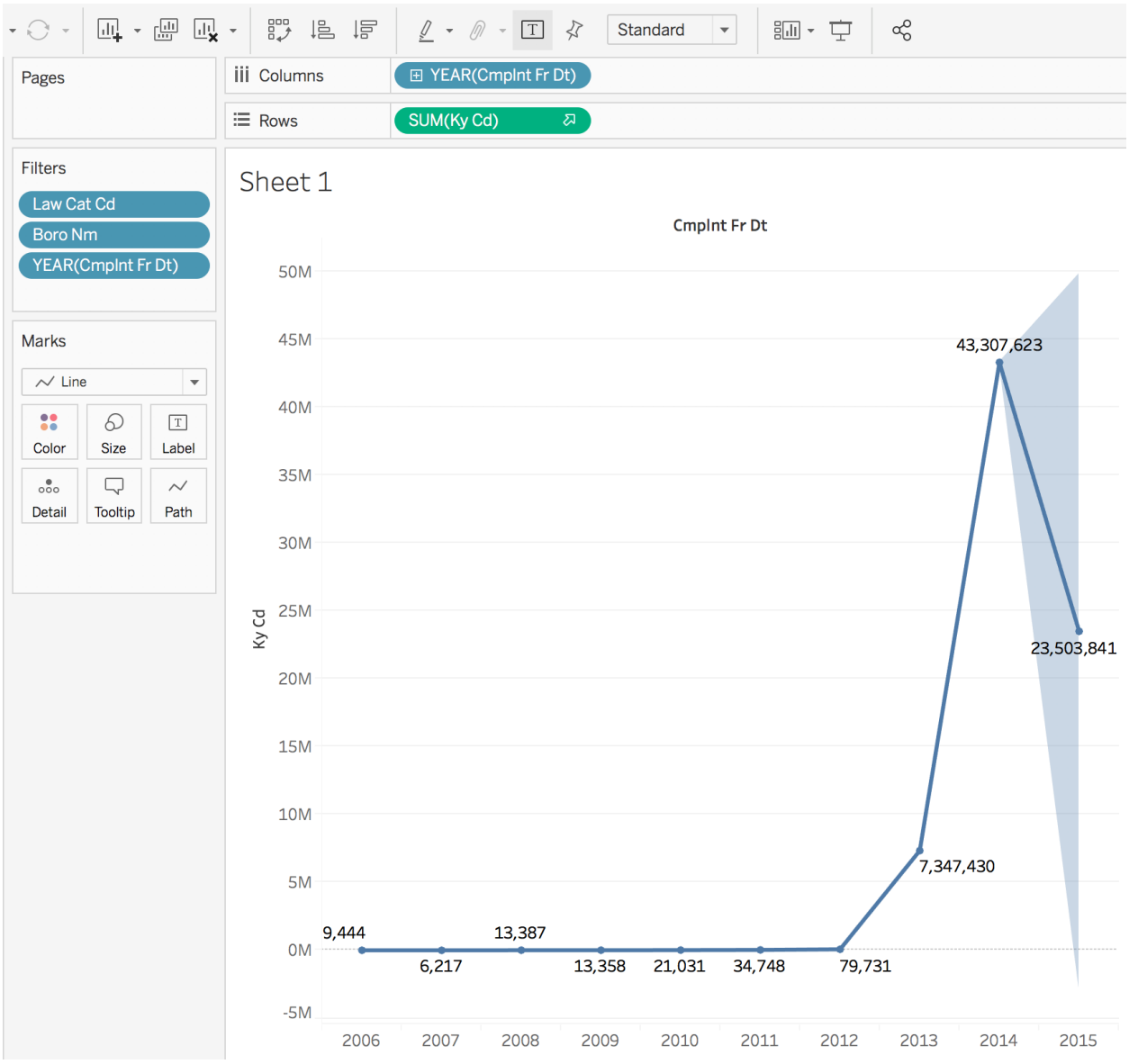
To find out the highest crime area in Chicago put lawcat,block and year in column then sum of the crimes in rows.

8. Yearly crime rate in unsafe area



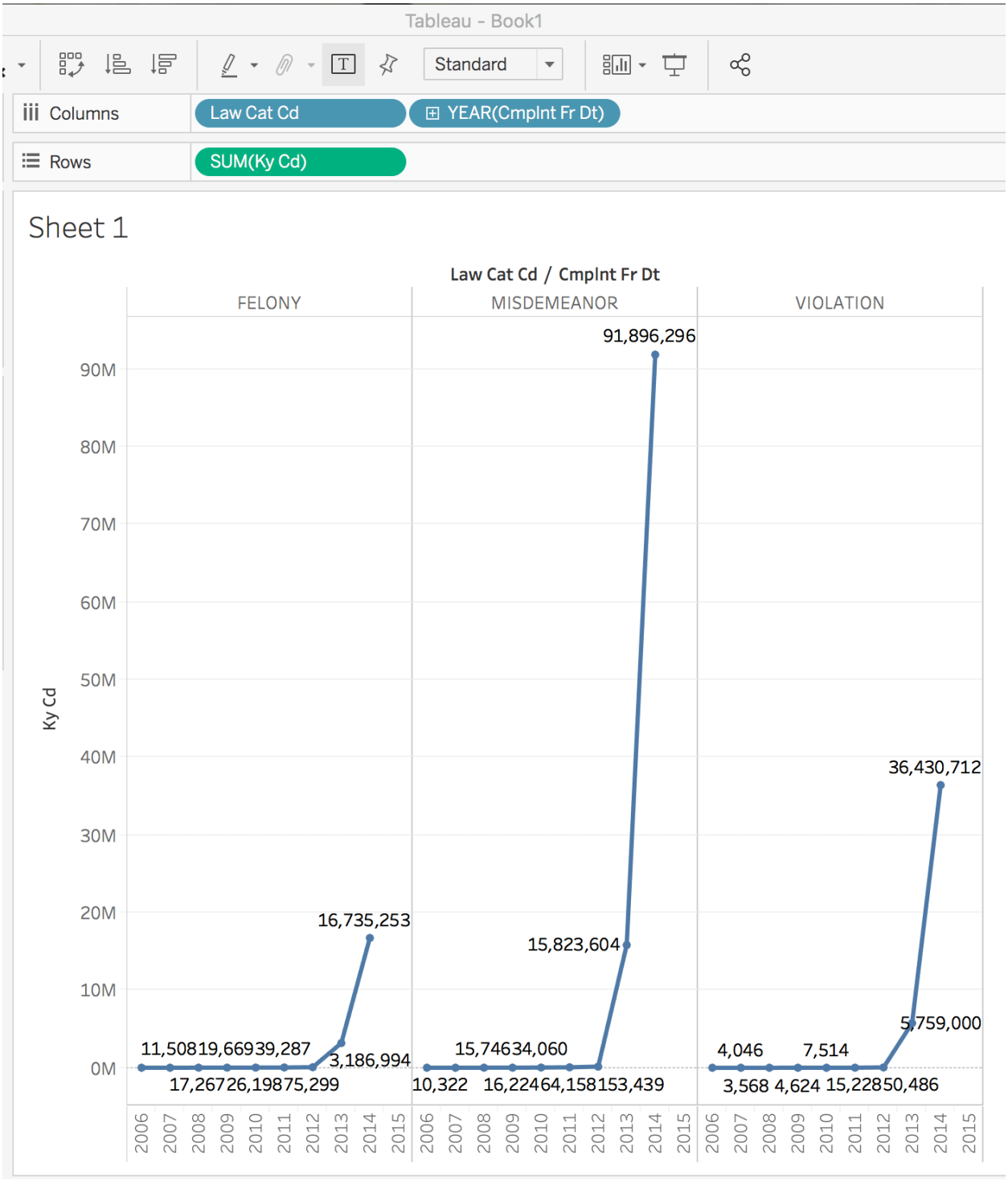
To visualize the yearly crime rate in top location in Chicago by differentiating it with the type of the crime. Drag and drop the block and year into columns then law cat and sum of iucr into rows.

9. New York crime forecast



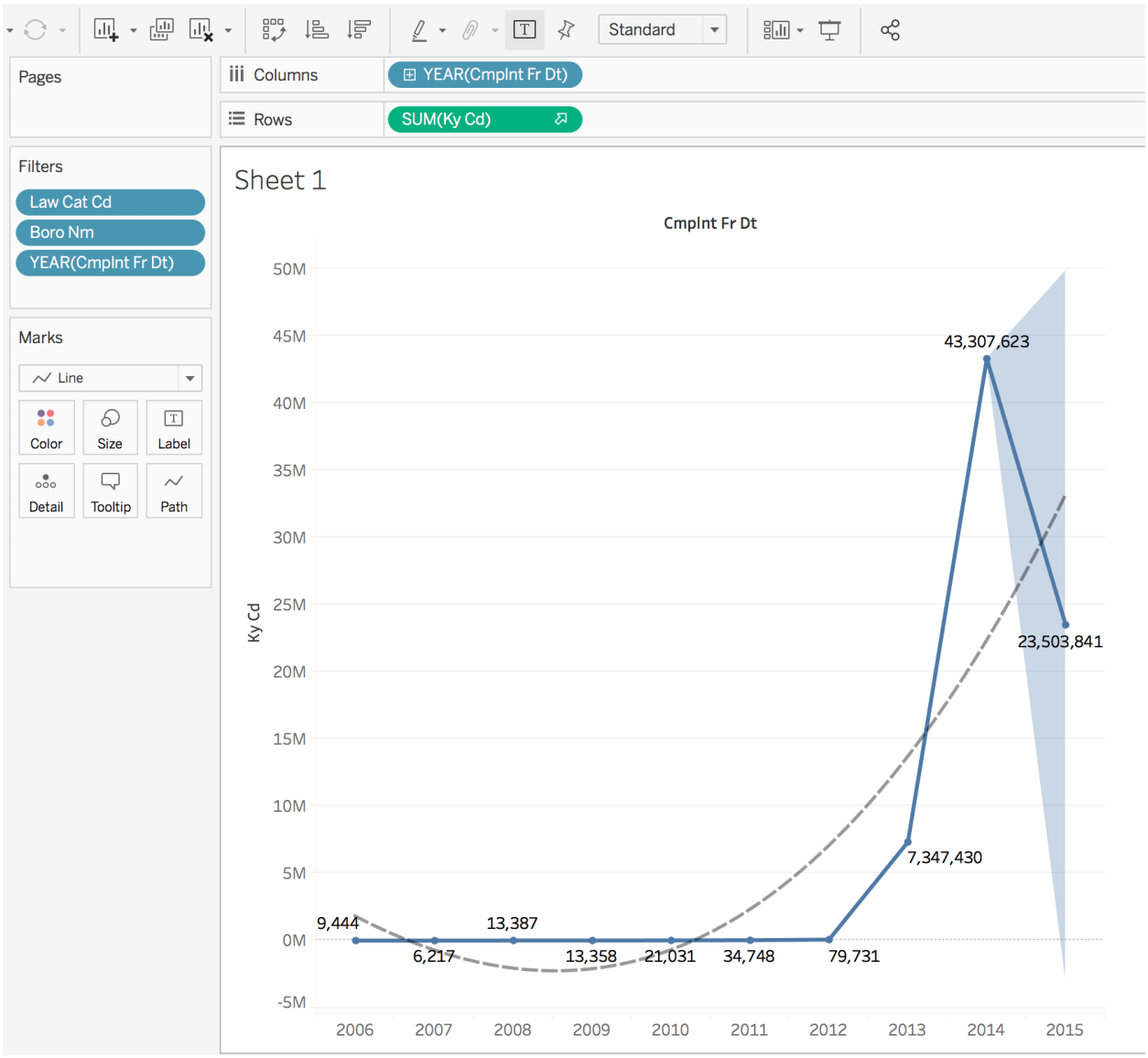
Select Years in columns of tableau and sum of the all the crimes in rows to forecast the crime for new york city. Using the analytics option and select the appropriate algorithm for calculating.

10. New York crime yearly



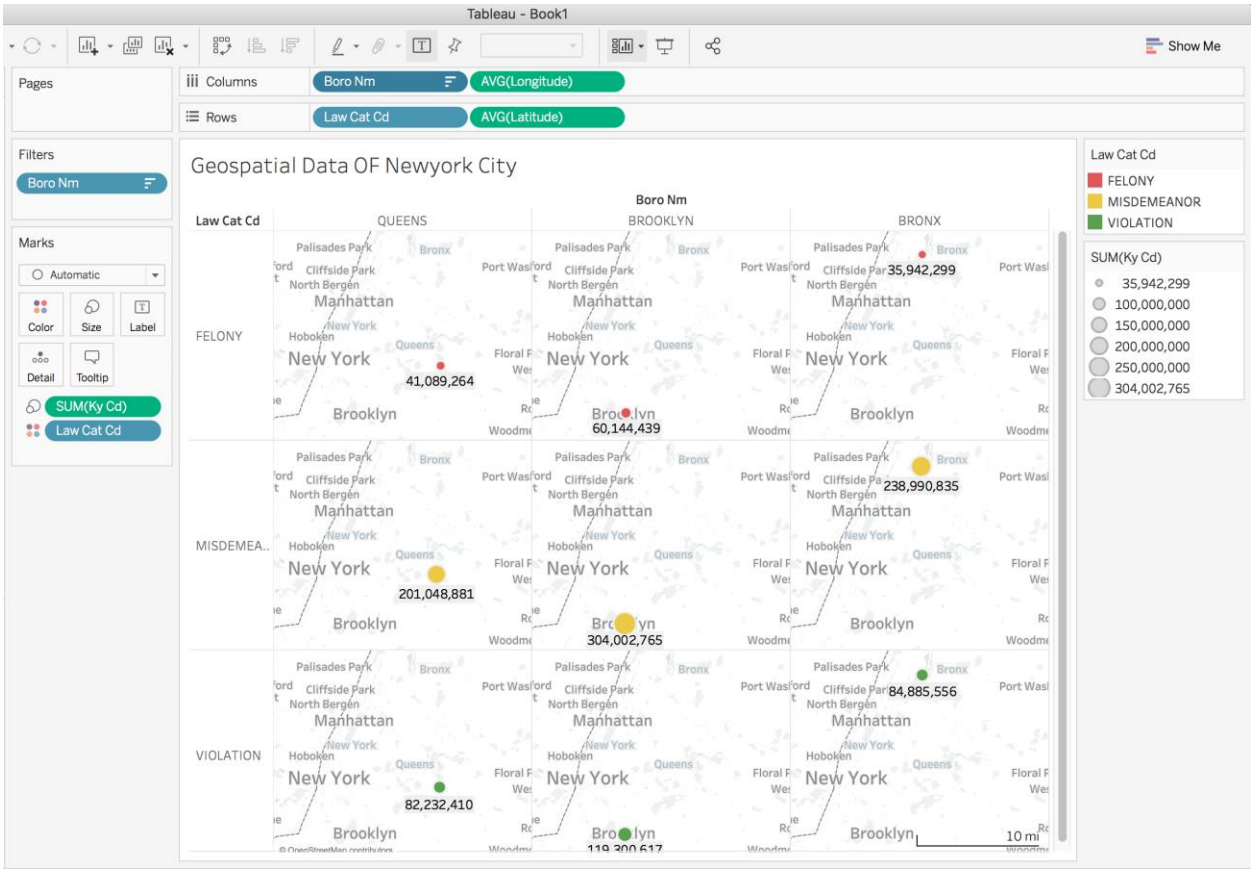
Yearly crime rate in NYC which are being differentiated by types of crime.

11. New York crime trending



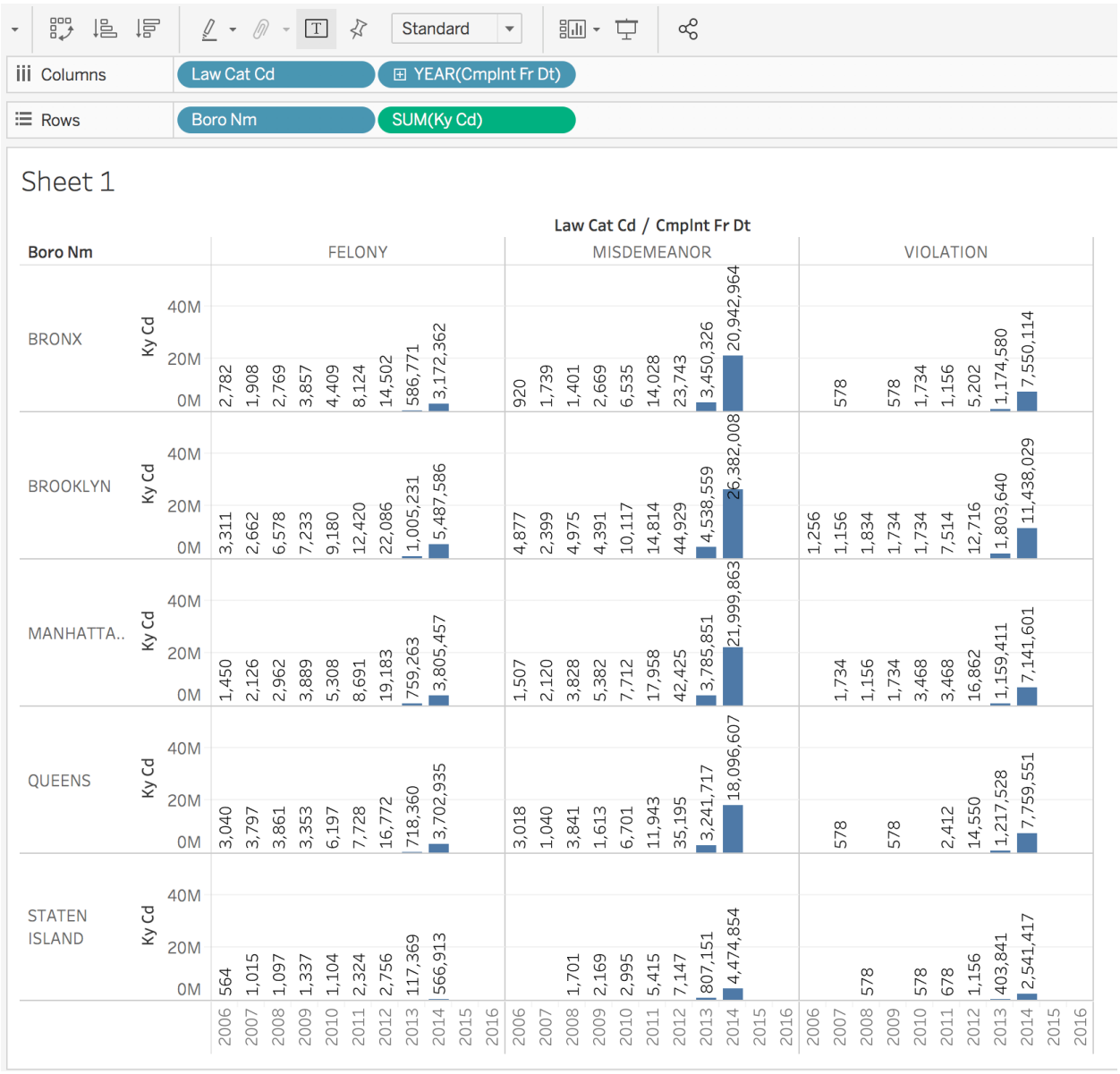
Select the top street which has the highest crime rate. find the trend of total crimes.

12. New York geo map



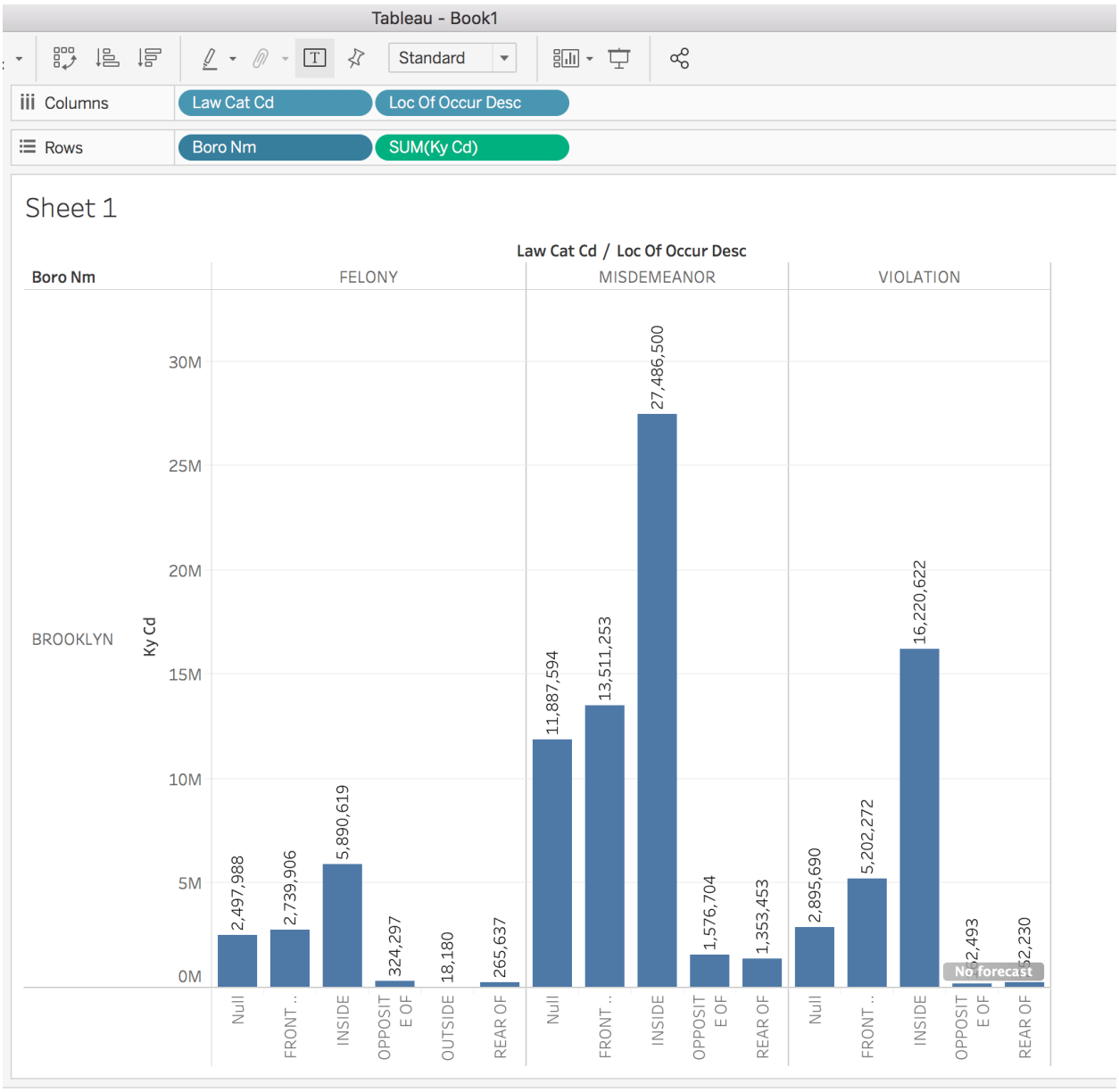
Select the top 3 streets with the count of the total crimes by adding the geo location specifications like latitude and longitude to locate the place in the Map. Different colors are given to different types of crime.

13. New York top 5 cities



Add the law_cat_cd and year in columns and sum of crimes(ky_cd) and boro nm to find the top 5 locations which highest crime rate.

14. New York top crime area



To find out the highest crime area in NYC put law_cat_cd and Loc of Occur Desc in column then sum of the crimes(ky_cd) and boro nm(Street name) in rows.

REFERENCES and GitHub link

https://github.com/annsummer94/mastercoder.git

http://www.calstatela.edu/centers/hipic/related-site

https://data.cityofnewyork.us/api/views/qgea-i56i/rows.csv?accessType=DOWNLOAD&bom=true&format=true

https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD&bom=true&format=true

https://console.bluemix.net/data/bic/

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Map/c4ep-ee5m

END OF THIS LAB