Big Data Crime Analysis Lab Tutorial

Authors: Jesus Perez, Abhinav Singh, Tien Cheng
Instructor: Jongwook Woo

Lab Tutorial: Analyzing Crime Data with HiveQL

Objectives: This tutorial guides you through downloading, uploading, processing, and analyzing crime datasets from Los Angeles (2010–Present) and Chicago (2001–Present) using Hadoop HDFS and HiveQL. You'll create tables, query the data, and export results as CSV files for visualization. Follow the steps carefully to ensure a smooth workflow.

Platform Spec: Oracle Linux Server v. 7.9, 5 Nodes (2 Master, 3 Worker), 31GBx5 RAM, 2.5GHz CPU for smooth data handling.

Prerequisites

- Access to Oracle Linux server, Hadoop cluster with HDFS tool and Hive installed to run beeline.
- SSH client (e.g., Git Bash, PuTTY) for file transfer and cluster access.
- Basic familiarity with SQL and command-line interfaces.
- SCP tool for file transfers (available in Git Bash or similar terminals).

Step 1: Download the Datasets

Download the following datasets and save them to your local machine (e.g., `C:\Users\YourName\Downloads`):

1. **LA Crime Data (2020–Present)**: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data
2. **LA Crime Data (2010–2019)**:

https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z/about_data

3. **Chicago Crime Data (2001–Present)**: https://data.cityofchicago.org/Public-Safety/Crimes-2025/t7ek-mgzi/about_data

Step 2: Upload Datasets to HDFS

2.1 Transfer Files to Cluster

Use SCP to transfer files to the `/tmp` directory of your Hadoop cluster:

```
scp "C:\Users\YourName\Downloads\Crime_Data_from_2010_to_2019.csv"
your_username@144.24.13.0:/tmp/
scp "C:\Users\YourName\Downloads\Crime_Data_from_2020_to_Present.csv"
your_username@144.24.13.0:/tmp/
scp "C:\Users\YourName\Downloads\Crimes_-_2001_to_Present.csv"
your_username@144.24.13.0:/tmp/
```

2.2 Verify Upload

```
ssh your_username@144.24.13.0
cd /tmp
ls
```

Ensure you see the uploaded files listed.

Step 3: Store Files in HDFS

3.1 Los Angeles Data

```
hdfs dfs -mkdir crime_data_LA
hdfs dfs -put /tmp/Crime_Data_from_2010_to_2019.csv crime_data_LA
hdfs dfs -put /tmp/Crime_Data_from_2020_to_Present.csv crime_data_LA
```

3.2 Chicago Data

```
hdfs dfs -mkdir crime_data_CH
hdfs dfs -put /tmp/Crimes_-_2001_to_Present.csv crime_data_CH
```

Step 4: Create Hive Tables

Start Beeline and Connect

```
> beeline
```

4.1 Create Los Angeles Table

```
USE your_username;
```

```
DROP TABLE IF EXISTS crime_data;

CREATE TABLE crime_data (
    DR_NO STRING,
    Date_Rptd STRING,
    DATE_OCC STRING,
    TIME_OCC INT,
    AREA INT,
    AREA_NAME STRING,
    Rpt_Dist_No INT,
    Part_1_2 INT,
    Crm_Cd INT,
    Crm_Cd_Desc STRING,
    Mocodes STRING,
    Vict_Age INT,
    Vict_Sex STRING,
    Vict_Descent STRING,
    Premis_Cd INT,
    Premis_Desc STRING,
    Weapon_Used_Cd STRING,
    Weapon_Desc STRING,
    Status STRING,
    Status_Desc STRING,
    Crm_Cd_1 INT,
    Crm_Cd_2 INT,
    Crm_Cd_3 INT,
    Crm_Cd_4 INT,
    LOCATION STRING,
    Cross_Street STRING,
    LAT DOUBLE,
    LON DOUBLE
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/your_username/crime_data_LA'
TBLPROPERTIES ('skip.header.line.count'='1');
```

```
SELECT * FROM crime_data LIMIT 10;
```

4.2 Create Chicago Table

```
DROP TABLE IF EXISTS crime_data_CH;

CREATE EXTERNAL TABLE crime_data_CH (
    ID BIGINT,
    Case_Number STRING,
    Dates STRING,
    Block STRING,
    IUCR STRING,
    Primary_Type STRING,
    Description STRING,
    Location_Description STRING,
    Arrest BOOLEAN,
    Domestic BOOLEAN,
    Beat INT,
    District INT,
    Ward INT,
    Community_Area INT,
    FBI_Code STRING,
    X_Coordinate DOUBLE,
    Y_Coordinate DOUBLE,
    Year INT,
    Updated_On STRING,
    Latitude DOUBLE,
    Longitude DOUBLE,
    Location STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/your_username/crime_data_CH'
TBLPROPERTIES ('skip.header.line.count'='1');
```

```
SELECT * FROM crime_data_CH LIMIT 10;
```

Step 5: Alternative Export via Hive Table

```
NOTE: CSV files don't have the headers when exported so give these headers
to the field in this sequence
```

6.1 Create Export Table

```
DROP TABLE IF EXISTS table_csv_export_data;

CREATE TABLE table_csv_export_data
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED AS TEXTFILE AS
<your entire Hive QL Query here>
```

```
INSERT INTO table_csv_export_data
<your entire Hive QL Query here>
```

6.2 Find Table Location

```
DESCRIBE FORMATTED table_csv_export_data;
```

Note the `Location` field in the output.

6.3 Extract CSV from HDFS (in a new terminal)

```
ssh your_username@144.24.13.0
hadoop fs -cat /path/to/table_csv_export_data/* > ~/name_your_csv_file.csv
```

6.4 Download to Local Machine (in a new terminal)

```
scp your_username@144.24.13.0:/home/your_username/name_your_csv_file.csv
~/name_your_csv_file.csv
```

Step 6: Analyze the Data (These visualizations are done in Excel 3D Map and Tableau, but you can also do them in Power BI)

5.1 Crime Location and Time

Chicago

```
SELECT
    Dates AS crime_datetime,
    Location,
    Latitude,
    Longitude,
    Primary_Type
FROM crime_data_ch
WHERE Dates IS NOT NULL
  AND Location IS NOT NULL
ORDER BY Dates
LIMIT 5;
```

Visualize on Excel 3D Map:

Open Excel and load your spreadsheet.

Click on **search** > type **3D Map** > **Open 3D Maps**.

A new window will open with a globe or map. Click **New Tour** if prompted.

Excel will auto-detect the location columns (e.g., latitude/longitude or addresses). If not:

- Use the right-side panel to manually assign fields (set `Latitude` and `Longitude` or `City/State` if that's what you have).

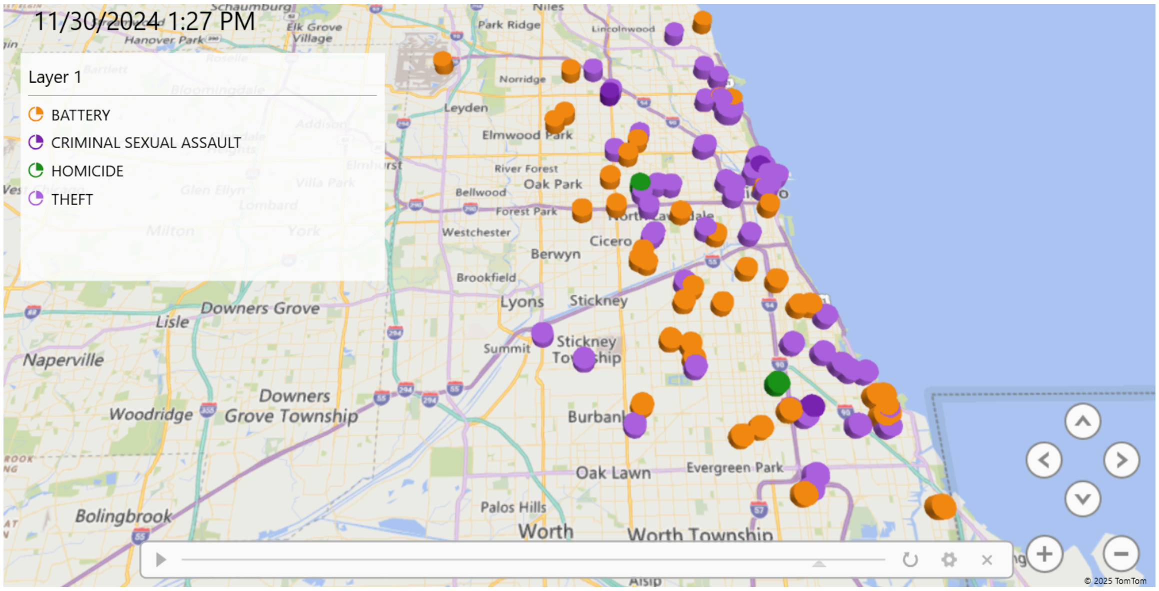
Choose "Primary Type" (count - Not Blank) under the Size selection.

Choose the "Primary Type" column to the Category section of the Layer Pane.

- This will allow different crimes to show in **distinct colors**, just like in our screenshot.

Drag the Date/Time column to the Time box (bottom of the Layer Pane) to activate the time slider.

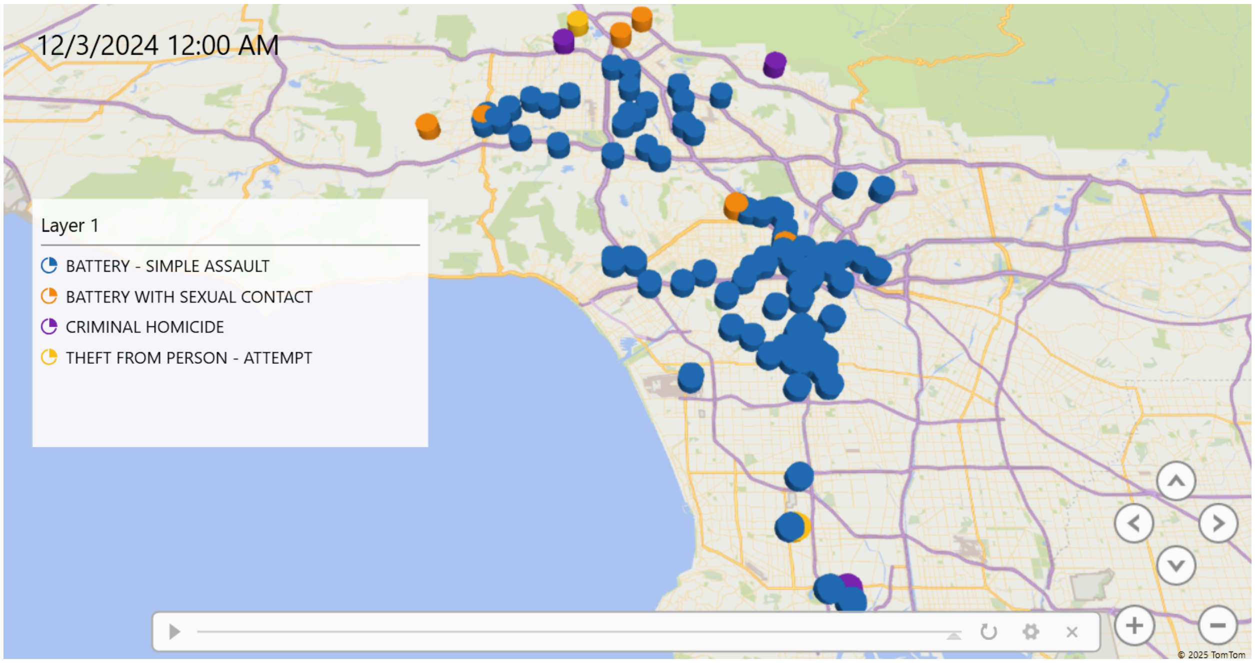
Note: We just visualized for particularly the 4 crimes shown in the screenshot, for both LA and Chicago cities, but one can analyze for whatever number of crimes one wants it to be.



Los Angeles

```
SELECT
    DATE_OCC AS crime_date,
    TIME_OCC AS crime_time,
    LOCATION,
    LAT AS latitude,
    LON AS longitude,
    Crm_Cd_Desc AS crime_type
```

```
FROM crime_data
WHERE DATE_OCC IS NOT NULL
  AND LOCATION IS NOT NULL
ORDER BY DATE_OCC, TIME_OCC
LIMIT 5;
```



5.2 Top Types of Crimes

Chicago

```
SELECT
    Primary_Type,
    COUNT(*) AS crime_count
FROM crime_data_ch
GROUP BY Primary_Type
ORDER BY crime_count DESC
LIMIT 10;
```

Los Angeles

```sql
SELECT
    Crm_Cd_Desc AS crime_type,
    COUNT(*) AS crime_count
FROM crime_data
GROUP BY Crm_Cd_Desc
ORDER BY crime_count DESC
LIMIT 10;
```

Visualize on Tableau for Top Types of Crimes:

Step 1: Load Your Data

1. Open **Tableau Desktop** or **Tableau Public**.
2. Click **"Connect" > "To a File" > "Text File"**.
3. Select your dataset: `crime_data_unified.csv` (or your own cleaned file).
4. Click **"Sheet 1"** at the bottom to start building your visualization.

Step 2: Drag and Drop Fields to Create the Tree Map

What You Want:

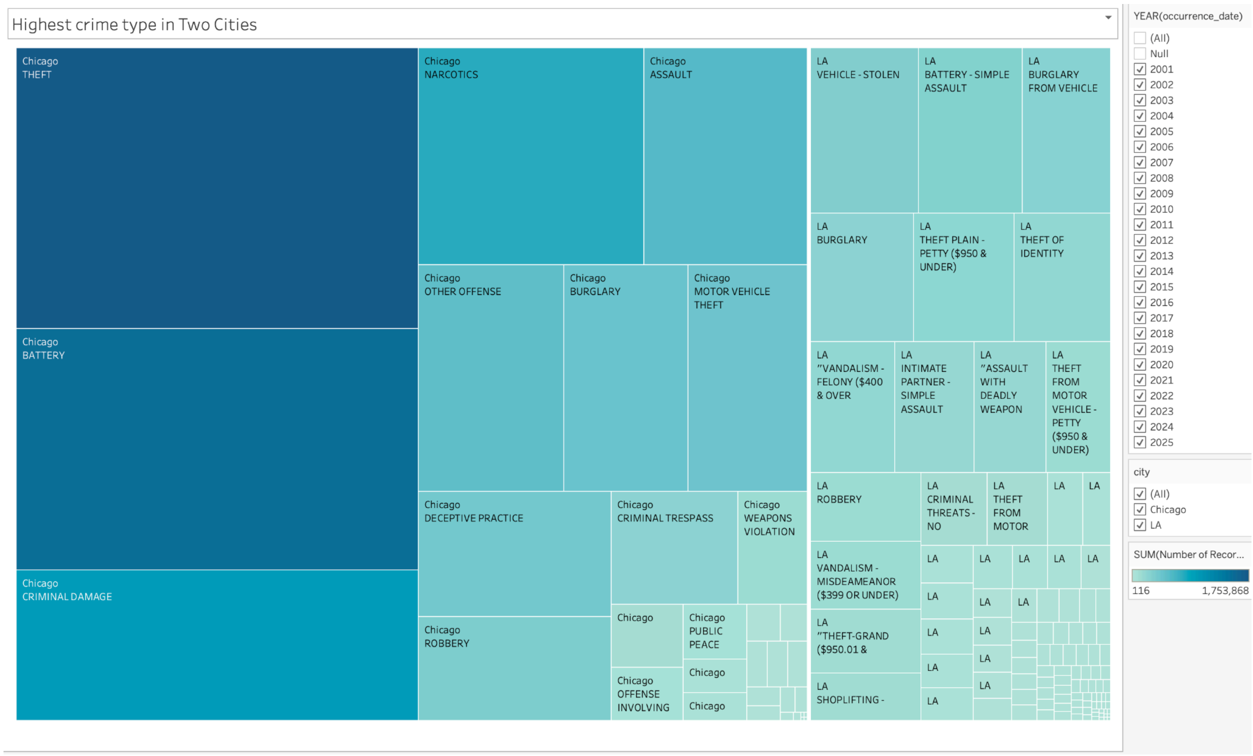
- **Size** of blocks = Number of crimes
- **Color** = City
- **Text** = Crime Type
- **Filters** = City + Year

Do the Following:

1. **Drag `crime_type` to Rows** or the **Label area** in the Marks card.
2. **Drag `Number of Records`** (or use `CNT(*)`) to **Size** on the Marks card.
3. **Drag `city` to Color** on the Marks card.
4. **Drag `crime_type` to Label** so the text appears inside the blocks.
5. Change the **Marks type** from "Automatic" to **Tree Map**.
6. **Drag `occurrence_date` to Filters** pane.
   - Choose "Years"
   - Select the years you want (or all)
7. **Drag `city` to Filters** to select between Chicago / LA.

Step 3: Final Formatting

1. Click on the **“Label”** on the Marks card:
   - ○ Check **“Show mark labels”**.
   - ○ Choose font size 10–12 for clarity.
2. Click on **“Color”** to adjust contrast (optional).
3. Add a **Title** like:
   `"Highest Crime Type in Two Cities"`



Highest crime type in Two Cities

YEAR(occurrence_date)
☐ (All)
☐ Null
☑ 2001
☑ 2002
☑ 2003
☑ 2004
☑ 2005
☑ 2006
☑ 2007
☑ 2008
☑ 2009
☑ 2010
☑ 2011
☑ 2012
☑ 2013
☑ 2014
☑ 2015
☑ 2016
☑ 2017
☑ 2018
☑ 2019
☑ 2020
☑ 2021
☑ 2022
☑ 2023
☑ 2024
☑ 2025

city
☑ (All)
☑ Chicago
☑ LA

Chicago THEFT
Chicago BATTERY
Chicago CRIMINAL DAMAGE
Chicago NARCOTICS
Chicago ASSAULT
Chicago OTHER OFFENSE
Chicago BURGLARY
Chicago MOTOR VEHICLE THEFT
Chicago DECEPTIVE PRACTICE
Chicago CRIMINAL TRESPASS
Chicago WEAPONS VIOLATION
Chicago ROBBERY
Chicago
Chicago PUBLIC PEACE
Chicago OFFENSE INVOLVING
Chicago
Chicago
LA VEHICLE - STOLEN
LA BATTERY - SIMPLE ASSAULT
LA BURGLARY FROM VEHICLE
LA BURGLARY
LA THEFT PLAIN - PETTY ($950 & UNDER)
LA THEFT OF IDENTITY
LA "VANDALISM - FELONY ($400 & OVER
LA INTIMATE PARTNER - SIMPLE ASSAULT
LA "ASSAULT WITH DEADLY WEAPON
LA THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)
LA ROBBERY
LA CRIMINAL THREATS - NO
LA THEFT FROM MOTOR
LA
LA
LA VANDALISM - MISDEAMEANOR ($399 OR UNDER)
LA "THEFT-GRAND ($950.01 &
LA SHOPLIFTING -
LA
LA
LA
LA
LA
LA
LA
LA
LA
LA
LA
LA
LA
LA
SUM(Number of Recor...
116
1,753,868

5.3 Police Surveillance Analysis

Chicago

```
SELECT
    District,
    COUNT(*) AS total_crimes,
    SUM(CAST(Arrest AS INT)) AS arrests,
    ROUND(SUM(CAST(Arrest AS INT)) * 100.0 / COUNT(*), 2) AS
arrest_percentage
FROM crime_data_CH
WHERE District IS NOT NULL
GROUP BY District
```

```
ORDER BY arrest_percentage DESC
LIMIT 5;
```

Los Angeles

```
SELECT
    AREA_NAME,
    COUNT(*) AS total_crimes,
    SUM(CASE WHEN Status IN ('AA', 'JA') THEN 1 ELSE 0 END) AS arrests,
    ROUND(SUM(CASE WHEN Status IN ('AA', 'JA') THEN 1 ELSE 0 END) * 100.0 /
COUNT(*), 2) AS arrest_percentage
FROM crime_data
WHERE AREA_NAME IS NOT NULL
GROUP BY AREA_NAME
ORDER BY arrest_percentage DESC
LIMIT 5;
```

Visualize on Tableau for Police Surveillance Analysis:

Step 1: Load the Data

1. Open **Tableau Desktop** or **Tableau Public**.
2. Click **Connect** → **Text File** and load your `crime_data_unified.csv`.
3. Go to **Sheet 1**.

Step 2: Build the Stacked Bar Chart

Drag and drop the following:

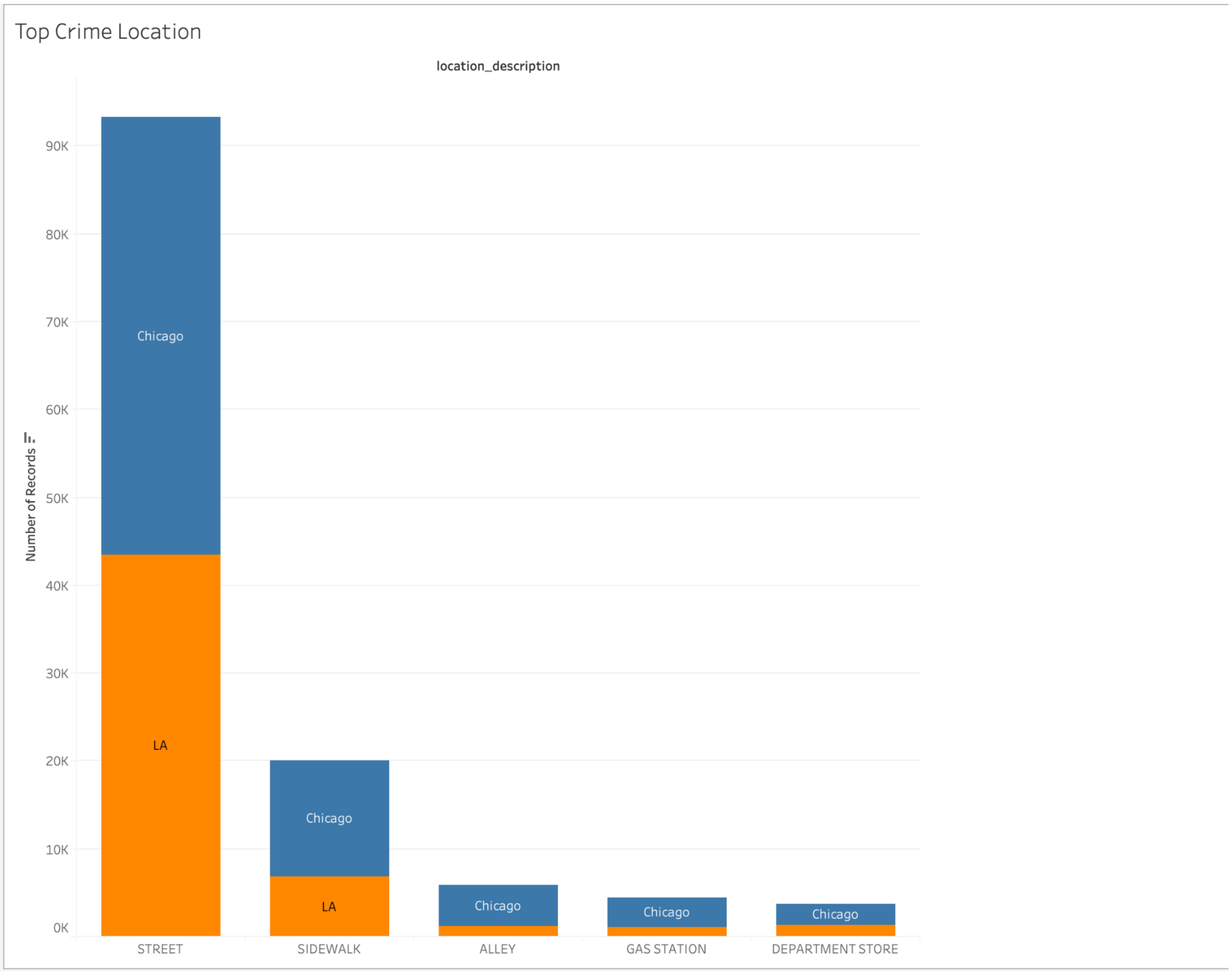
1. **Drag `location_description`** → to **Columns**
2. **Drag `Number of Records`** (or use `CNT(*)`) → to **Rows**
3. **Drag `city`** → to **Color** on the Marks card
4. **Drag `city`** again → to **Label** (so it shows inside the bar)
5. **Optional:** Drag `Number of Records` again → to **Label** to show the count

Step 3: Apply Filters

1. **Drag `city`** → to Filters → Select **Chicago and LA**
2. **Drag `occurrence_date`** → to Filters → Choose "Years" → Select years 2010–2025 or as needed
3. Optional: **Drag `location_description`** → to Filters to focus on top 5–10 locations

Step 4: Format the Chart

- Click on **Color** → adjust palette to blue/orange like your chart
- Click on **Label** → check **"Show mark labels"** to display numbers
- Edit **Title**:
  "Top Crime Locations in Chicago vs LA"



5.4 Arrests per Crime Type

Chicago

```
SELECT
    Primary_Type,
    COUNT(*) AS total_crimes,
    SUM(CAST(Arrest AS INT)) AS arrests,
    ROUND(SUM(CAST(Arrest AS INT)) * 100.0 / COUNT(*), 2) AS
arrest_percentage
FROM crime_data_CH
GROUP BY Primary_Type
ORDER BY arrest_percentage DESC
LIMIT 5;
```

Los Angeles

```
SELECT
    Crm_Cd_Desc AS crime_type,
    COUNT(*) AS total_crimes,
    SUM(CASE WHEN Status IN ('AA', 'JA') THEN 1 ELSE 0 END) AS arrests,
    ROUND(SUM(CASE WHEN Status IN ('AA', 'JA') THEN 1 ELSE 0 END) * 100.0 /
COUNT(*), 2) AS arrest_percentage
FROM crime_data
GROUP BY Crm_Cd_Desc
ORDER BY arrest_percentage DESC
LIMIT 5;
```

Visualize on Tableau for Arrests per Crime Type:

STEP 1: Connect to the Dataset

1. Open **Tableau Desktop** or **Tableau Public**.
2. Connect to `crime_data_unified.csv`.
3. Go to **Sheet 1**.

STEP 2: Build the Heat Map

Drag fields into the worksheet:

1. **Drag `crime_type`** → to **Rows**
2. **Drag `YEAR(occurrence_date)`** → to **Columns**
   - Right-click `occurrence_date` → Convert to **Discrete** → Choose **"Years"**
3. **Drag `city`** → to **Rows**, next to `crime_type`
4. **Drag `Number of Records`** → to **Color**
   - Use `SUM(Number of Records)`
5. **Drag `Number of Records`** → to **Label** so numbers appear inside each cell

STEP 3: Format the Heat Map

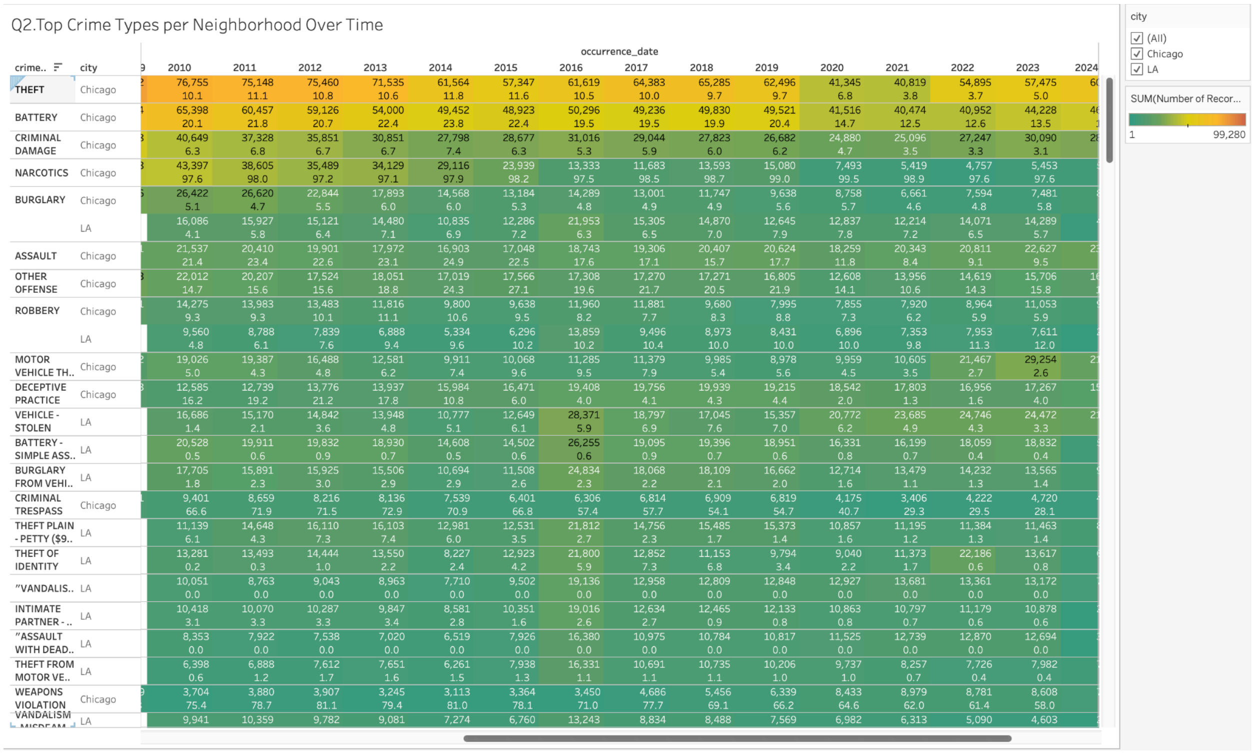
1. On the **Marks card**, change type to **Square**
2. Adjust **Color**:
   - Click on **Color** → Choose a **green-yellow color scale** or **Orange-Green** diverging scale
3. Click **Label**:
   - Check **"Show mark labels"**
   - Format font size to ~9–10 for readability

STEP 4: Filter and Focus

1. **Drag `city`** to **Filters** → select **Chicago** and **LA**
2. (Optional) **Drag `crime_type`** to Filters → show top 20 crime types only
   - Right-click `crime_type` → Filter → "Top" tab → "By field" → Top 20 by `SUM(Number of Records)`

STEP 5: Add Chart Title

- Double-click title and type:
  "Top Crime Types per Neighborhood Over Time"



Q2.Top Crime Types per Neighborhood Over Time

occurrence_date

crime..	city	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
THEFT	Chicago	76,755 10.1	75,148 11.1	75,460 10.8	71,535 10.6	61,564 11.8	57,347 11.6	61,619 10.5	64,383 10.0	65,285 9.7	62,496 9.7	41,345 6.8	40,819 3.8	54,895 3.7	57,475 5.0
BATTERY	Chicago	65,398 20.1	60,457 21.8	59,126 20.7	54,000 22.4	49,452 23.8	48,923 22.4	50,296 19.5	49,236 19.5	49,830 19.9	49,521 20.4	41,516 14.7	40,474 12.5	40,952 12.6	44,228 13.5
CRIMINAL DAMAGE	Chicago	40,649 6.3	37,328 6.8	35,851 6.7	30,851 6.7	27,798 7.4	28,677 6.3	31,016 5.3	29,044 5.9	27,823 6.0	26,682 6.2	24,880 4.7	25,096 3.5	27,247 3.3	30,090 3.1
NARCOTICS	Chicago	43,397 97.6	38,605 98.0	35,489 97.2	34,129 97.1	29,116 97.9	23,939 98.2	13,333 97.5	11,683 98.5	13,593 98.7	15,080 99.0	7,493 99.5	5,419 98.9	4,757 97.6	5,453 97.6
BURGLARY	Chicago	26,422 5.1	26,620 4.7	22,844 5.5	17,893 6.0	14,568 6.0	13,184 5.3	14,289 4.8	13,001 4.9	11,747 4.9	9,638 5.6	8,758 5.7	6,661 4.6	7,594 4.8	7,481 5.8
	LA	16,086 4.1	15,927 5.8	15,121 6.4	14,480 7.1	10,835 6.9	12,286 7.2	21,953 6.3	15,305 6.5	14,870 7.0	12,645 7.9	12,837 7.8	12,214 7.2	14,071 6.5	14,289 5.7
ASSAULT	Chicago	21,537 21.4	20,410 23.4	19,901 22.6	17,972 23.1	16,903 24.9	17,048 22.5	18,743 17.6	19,306 17.1	20,407 15.7	20,624 17.7	18,259 11.8	20,343 8.4	20,811 9.1	22,627 9.5
OTHER OFFENSE	Chicago	22,012 14.7	20,207 15.6	17,524 15.6	18,051 18.8	17,019 24.3	17,566 27.1	17,308 19.6	17,270 21.7	17,271 20.5	16,805 21.9	12,608 14.1	13,956 10.6	14,619 14.3	15,706 15.8
ROBBERY	Chicago	14,275 9.3	13,983 9.3	13,483 10.1	11,816 11.1	9,800 10.6	9,638 9.5	11,960 8.2	11,881 7.7	9,680 8.3	7,995 8.8	7,855 7.3	7,920 6.2	8,964 5.9	11,053 5.9
	LA	9,560 4.8	8,788 6.1	7,839 7.6	6,888 9.4	5,334 9.6	6,296 10.2	13,859 10.2	9,496 10.4	8,973 10.0	8,431 10.0	6,896 10.0	7,353 9.8	7,953 11.3	7,611 12.0
MOTOR VEHICLE TH..	Chicago	19,026 5.0	19,387 4.3	16,488 4.8	12,581 6.2	9,911 7.4	10,068 9.6	11,285 9.5	11,379 7.9	9,985 5.4	8,978 5.6	9,959 4.5	10,605 3.5	21,467 2.7	29,254 2.6
DECEPTIVE PRACTICE	Chicago	12,585 16.2	12,739 19.2	13,776 21.2	13,937 17.8	15,984 10.8	16,471 6.0	19,408 4.0	19,756 4.1	19,939 4.3	19,215 4.4	18,542 2.0	17,803 1.3	16,956 1.6	17,267 4.0
VEHICLE - STOLEN	LA	16,686 1.4	15,170 2.1	14,842 3.6	13,948 4.8	10,777 5.1	12,649 6.1	28,371 5.9	18,797 6.9	17,045 7.6	15,357 7.0	20,772 6.2	23,685 4.9	24,746 4.3	24,472 3.3
BATTERY - SIMPLE ASS..	LA	20,528 0.5	19,911 0.6	19,832 0.9	18,930 0.7	14,608 0.5	14,502 0.6	26,255 0.6	19,095 0.9	19,396 0.7	18,951 0.6	16,331 0.8	16,199 0.7	18,059 0.4	18,832 0.4
BURGLARY FROM VEHI..	LA	17,705 1.8	15,891 2.3	15,925 3.0	15,506 2.9	10,694 2.9	11,508 2.6	24,834 2.3	18,068 2.2	18,109 2.1	16,662 2.0	12,714 1.6	13,479 1.1	14,232 1.3	13,565 1.4
CRIMINAL TRESPASS	Chicago	9,401 66.6	8,659 71.9	8,216 71.5	8,136 72.9	7,539 70.9	6,401 66.8	6,306 57.4	6,814 57.7	6,909 54.1	6,819 54.7	4,175 40.7	3,406 29.3	4,222 29.5	4,720 28.1
THEFT PLAIN - PETTY ($9..	LA	11,139 6.1	14,648 4.3	16,110 7.3	16,103 7.4	12,981 6.0	12,531 3.5	21,812 2.7	14,756 2.3	15,485 1.7	15,373 1.4	10,857 1.6	11,195 1.2	11,384 1.3	11,463 1.4
THEFT OF IDENTITY	LA	13,281 0.2	13,493 0.3	14,444 1.0	13,550 2.2	8,227 2.4	12,923 4.2	21,800 5.9	12,852 7.3	11,153 6.8	9,794 3.4	9,040 2.2	11,373 1.7	22,186 0.6	13,617 0.8
"VANDALIS..	LA	10,051 0.0	8,763 0.0	9,043 0.0	8,963 0.0	7,710 0.0	9,502 0.0	19,136 0.0	12,958 0.0	12,809 0.0	12,848 0.0	12,927 0.0	13,681 0.0	13,361 0.0	13,172 0.0
INTIMATE PARTNER - ..	LA	10,418 3.1	10,070 3.3	10,287 3.3	9,847 3.4	8,581 2.8	10,351 1.6	19,016 2.6	12,634 2.7	12,465 0.9	12,133 0.8	10,863 0.8	10,797 0.7	11,179 0.6	10,878 0.6
"ASSAULT WITH DEAD..	LA	8,353 0.0	7,922 0.0	7,538 0.0	7,020 0.0	6,519 0.0	7,926 0.0	16,380 0.0	10,975 0.0	10,784 0.0	10,817 0.0	11,525 0.0	12,739 0.0	12,870 0.0	12,694 0.0
THEFT FROM MOTOR VE..	LA	6,398 0.6	6,888 1.2	7,612 1.7	7,651 1.6	6,261 1.5	7,938 1.3	16,331 1.1	10,691 1.1	10,735 1.1	10,206 1.0	9,737 1.0	8,257 0.7	7,726 0.4	7,982 0.4
WEAPONS VIOLATION	Chicago	3,704 75.4	3,880 78.7	3,907 81.1	3,245 79.4	3,113 81.0	3,364 78.1	3,450 71.0	4,686 77.7	5,456 69.1	6,339 66.2	8,433 64.6	8,979 62.0	8,781 61.4	8,608 58.0
VANDALISM - MISDEAM..	LA	9,941	10,359	9,782	9,081	7,274	6,760	13,243	8,834	8,488	7,569	6,982	6,313	5,090	4,603

5.5 Crimes Per Year

Chicago

```
SELECT
    Year,
    COUNT(*) AS crime_count
FROM crime_data_CH
WHERE Year IS NOT NULL AND Year BETWEEN 1900 AND 2100
GROUP BY Year
ORDER BY Year
LIMIT 5;
```

Los Angeles

```
SELECT
    SUBSTR(DATE_OCC, 7, 4) AS year, -- Extracts the 4-digit year (e.g., "2023") from DATE_OCC string, starting at position 7 (after "MM/DD/") for 4 characters

    COUNT(*) AS crime_count
FROM crime_data
WHERE DATE_OCC IS NOT NULL
GROUP BY SUBSTR(DATE_OCC, 7, 4)
ORDER BY year
LIMIT 5;
```

Visualize on Tableau for Crimes Per Year:

STEP 1: Load the Data

1. Open **Tableau Desktop** or **Tableau Public**.
2. Connect to your dataset: `crime_data_unified.csv`
3. Go to **Sheet 1**

STEP 2: Drag Fields to the View

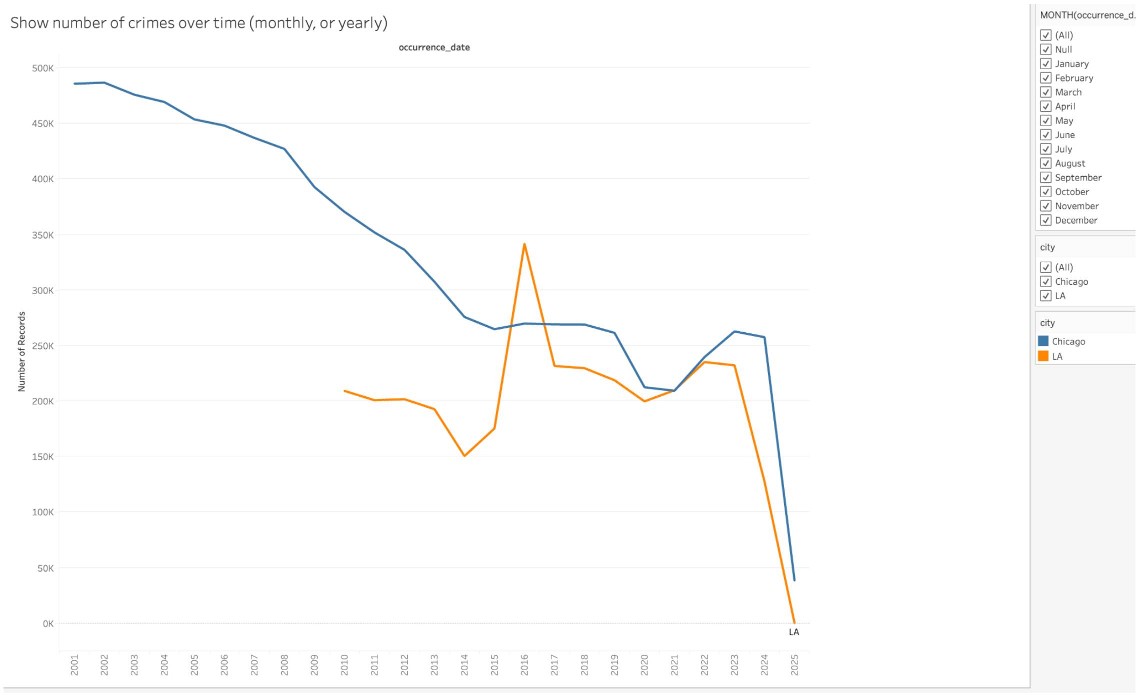
1. **Drag `YEAR(occurrence_date)`** → to **Columns**
    - Right-click `occurrence_date` → select **"Year"**
2. **Drag `Number of Records`** → to **Rows**
    - Use `SUM(Number of Records)`
3. **Drag `city`** → to **Color** on the **Marks** card
4. Optional: Drag `city` → to **Label** (to show name on line ends)

STEP 3: Filter the Data

1. **Drag `city`** → to Filters → select **Chicago and LA**
2. Optional: Drag `MONTH(occurrence_date)` → to Filters (as you did) if you want to filter by months
3. Right-click filters and select **"Show Filter"** to add interactivity

STEP 4: Format the Chart

1. Set chart type to **Line** (Tableau does this automatically)
2. Click **Label**:
   - Enable **Show mark labels** if you want point values
3. Add a title:
   "Number of Crimes Over Time (Yearly) – Chicago vs LA"



MONTH(occurrence_d..

☑ (All)
☑ Null
☑ January
☑ February
☑ March
☑ April
☑ May
☑ June
☑ July
☑ August
☑ September
☑ October
☑ November
☑ December

city

☑ (All)
☑ Chicago
☑ LA

References:

1. **URL of Data source:**

   LA Crime Data (2020–Present):
   https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data
   LA Crime Data (2010–2019):
   https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z/about_data
   Chicago Crime Data (2001–Present):
   https://data.cityofchicago.org/Public-Safety/Crimes-2025/t7ek-mgzi/about_data

2. **URL of GitHub: https://github.com/shjepz/BigDataCrimeAnalysis**