

# LOS ANGELES CRIME ANALYSIS

Business Intelligence & Big Data Analytics

KYPRITIDIS STEFANOS 8170050

# DATASET

- LOS ANGELES CRIME INCIDENTS  
FOR 2010 - 2019
- DATA FROM LA POLICE DEPARTMENT
- MORE THAN HALF A GIGABYTE  
INFORMATION
- 2,1 MILLION INCIDENTS



# CRIME INFORMATION

## THE MOST IMPORTANT INFORMATION PER INCIDENT

- DATE OCCURED AND DATE REPORTED OF INCIDENT
- SEX, AGE AND DESCENT OF THE VICTIM
- CRIME DESCRIPTION AND EXTRA CRIMES COMMITTED
- WEAPON USED AND PREMISES OF THE INCIDENT
- STATUS OF CASE
- SUSPECT ACTIVITIES
- GEOGRAPHIC AREA, LOCATION AND COORDINATES



# DATA EXTRACTION

DOWNLOADED FROM KAGGLE PROJECT: *POLICE VIOLENCE & RACIAL EQUITY*

The screenshot shows the Kaggle interface. On the left, there's a sidebar with links like Home, Compete, Data, Notebooks, Communities, Courses, and More. Under 'Recently Viewed', there are links to 'Police Violence & Raci...', 'Titanic with decision tr...', 'Titanic Data Science S...', 'Titanic - Machine Lear...', and 'Categorical Variables I...'. At the bottom of the sidebar is a 'View Active Events' button.

The main content area displays a dataset titled 'Police Violence & Racial Equity - Part 2 of 3'. It includes a search bar, a profile picture for 'JohnM', and a timestamp 'updated 3 days ago (Version 10)'. Below this are tabs for Data, Tasks (1), Notebooks (5), Discussion (3), Activity, and Metadata. There are also 'Download (773 MB)' and 'New Notebook' buttons.

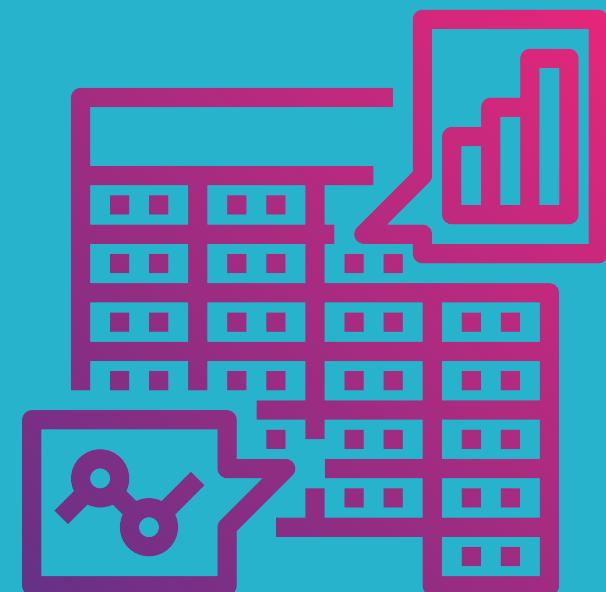
Below the tabs, there are sections for Usability (10.0), License (Other (specified in description)), and Tags (social science, social issues and advocacy, demographics, racial equity). The Description section contains a detailed description of the dataset, mentioning it's part of a three-part series and listing specific data types for each part. The Data Explorer section shows a file named 'LA Crime\_Data\_from\_2010\_to\_2019.csv' (510.09 MB) with a 'Download' button.

At the bottom, there's an 'About this file' section with a source link: <https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/2mrs-mtv>.

# DATA PREVIEW

HERE IS THE RAW DATA FROM KAGGLE

2,1 MILLION ROWS



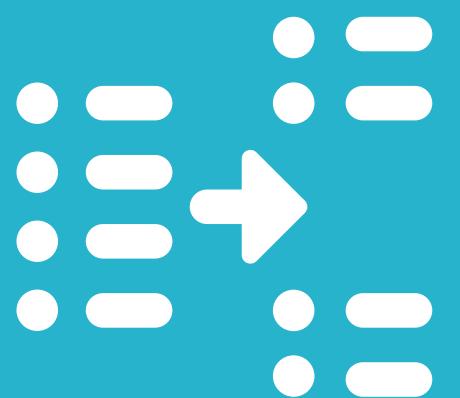
28 DIFFERENT COLUMNS

	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	...	Status	Status Desc	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4	LOC
0	1307355	02/20/2010 12:00:00 AM	02/20/2010 12:00:00 AM	1350	13	Newton	1385	2	900	VIOLATION OF COURT ORDER	...	AA	Adult Arrest	900.0	NaN	NaN	NaN	300 E
1	11401303	09/13/2010 12:00:00 AM	09/12/2010 12:00:00 AM	45	14	Pacific	1485	2	740	VANDALISM - FELONY (\$400 & OVER, ALL CHURCH VA...)	...	IC	Invest Cont	740.0	NaN	NaN	NaN	SEPU
2	70309629	08/09/2010 12:00:00 AM	08/09/2010 12:00:00 AM	1515	13	Newton	1324	2	946	OTHER MISCELLANEOUS CRIME	...	IC	Invest Cont	946.0	NaN	NaN	NaN	1300
3	90631215	01/05/2010 12:00:00 AM	01/05/2010 12:00:00 AM	150	6	Hollywood	646	2	900	VIOLATION OF COURT ORDER	...	IC	Invest Cont	900.0	998.0	NaN	NaN	CAHI
4	100100501	01/03/2010 12:00:00 AM	01/02/2010 12:00:00 AM	2100	1	Central	176	1	122	RAPE, ATTEMPTED	...	IC	Invest Cont	122.0	NaN	NaN	NaN	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2114694	190608903	03/28/2019 12:00:00 AM	03/28/2019 12:00:00 AM	400	6	Hollywood	644	1	648	ARSON	...	IC	Invest Cont	648.0	NaN	NaN	NaN	140 BI
2114695	190715222	08/15/2019 12:00:00 AM	08/14/2019 12:00:00 AM	1810	7	Wilshire	701	1	331	THEFT FROM MOTOR VEHICLE - GRAND (\$400 AND OVER)	...	IC	Invest Cont	331.0	NaN	NaN	NaN	WILLOU
2114696	192004409	01/06/2019 12:00:00 AM	01/06/2019 12:00:00 AM	2100	20	Olympic	2029	2	930	CRIMINAL THREATS - NO WEAPON DISPLAYED	...	IC	Invest Cont	930.0	NaN	NaN	NaN	
2114697	191716777	10/17/2019 12:00:00 AM	10/16/2019 12:00:00 AM	1800	17	Devonshire	1795	1	420	THEFT FROM MOTOR VEHICLE - PETTY (\$950 & UNDER)	...	IC	Invest Cont	420.0	NaN	NaN	NaN	1720
2114698	190805435	02/01/2019 12:00:00 AM	02/01/2019 12:00:00 AM	1615	8	West LA	852	1	330	BURGLARY FROM VEHICLE	...	IC	Invest Cont	330.0	NaN	NaN	NaN	1700 I

2114699 rows × 28 columns

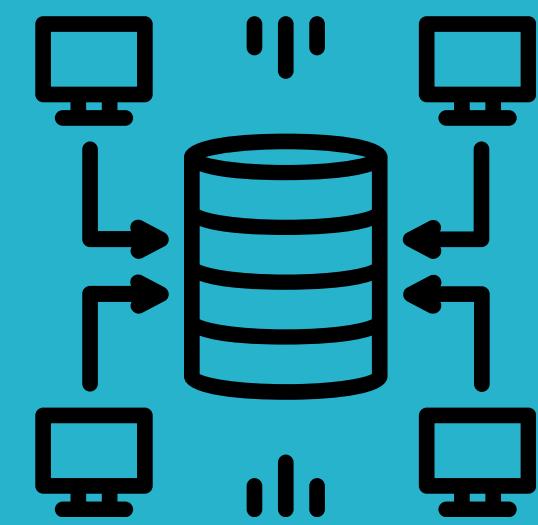
# DATA TRANSFORMATION

- ENCODED NULL VALUES
- FIXED FORMAT FOR DATES & TIME
- CORRECTED MISTAKEN ENTRIES E.G. NEGATIVE AGE
- CORRECTED MISTAKEN CORRESPONDENCES FROM CODES TO DESCRIPTIONS
- REMOVED UNNECESSARY COLUMNS
- CREATED & SEPARATED DIMENSIONS FROM DATA



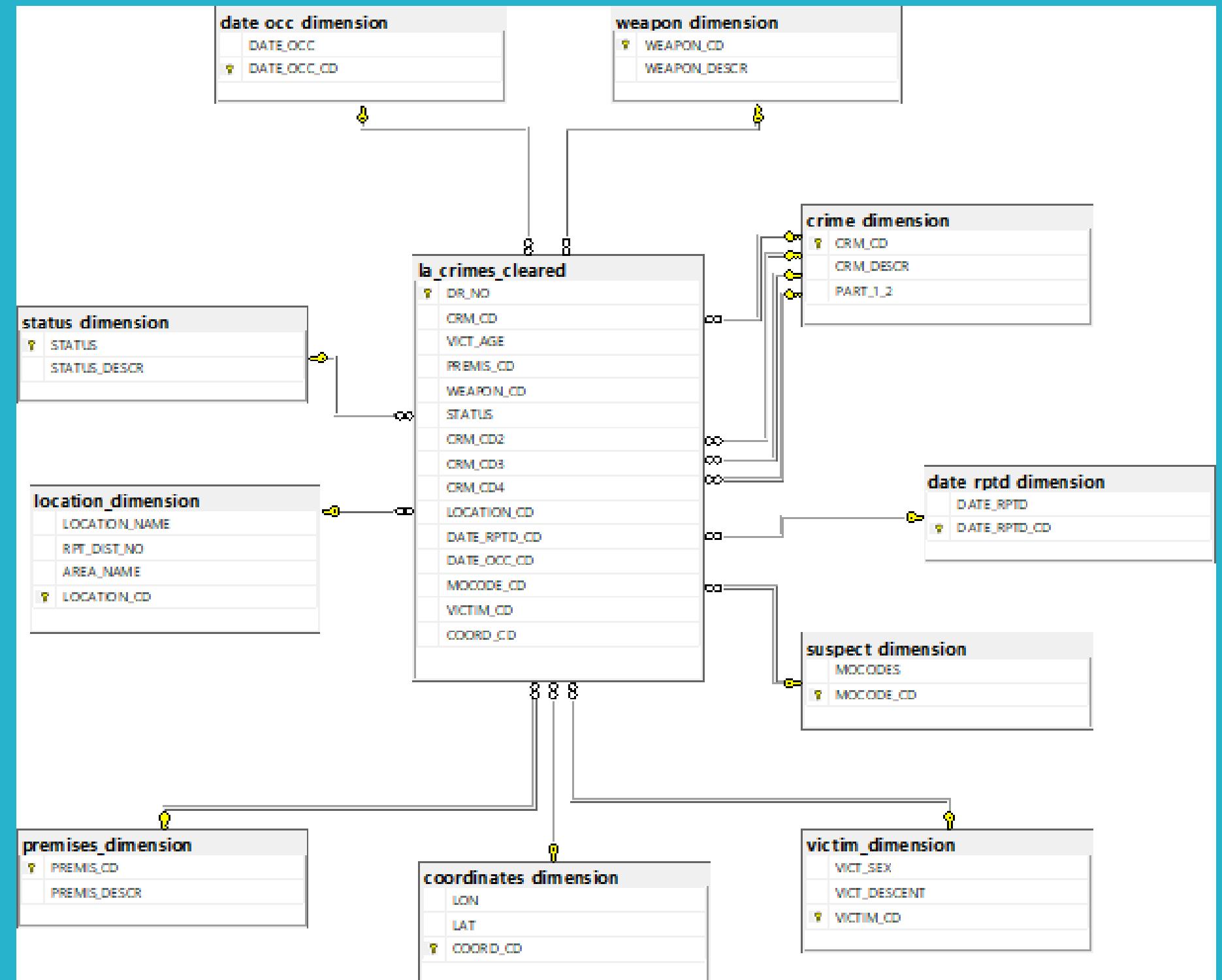
# DATA LOADING

LOADED DATA INTO SQL SERVER WAREHOUSE



STAR SCHEMA:

- ONE FACT TABLE WITH ONE METRIC
- 10 DIFFERENT DIMENSION TABLES

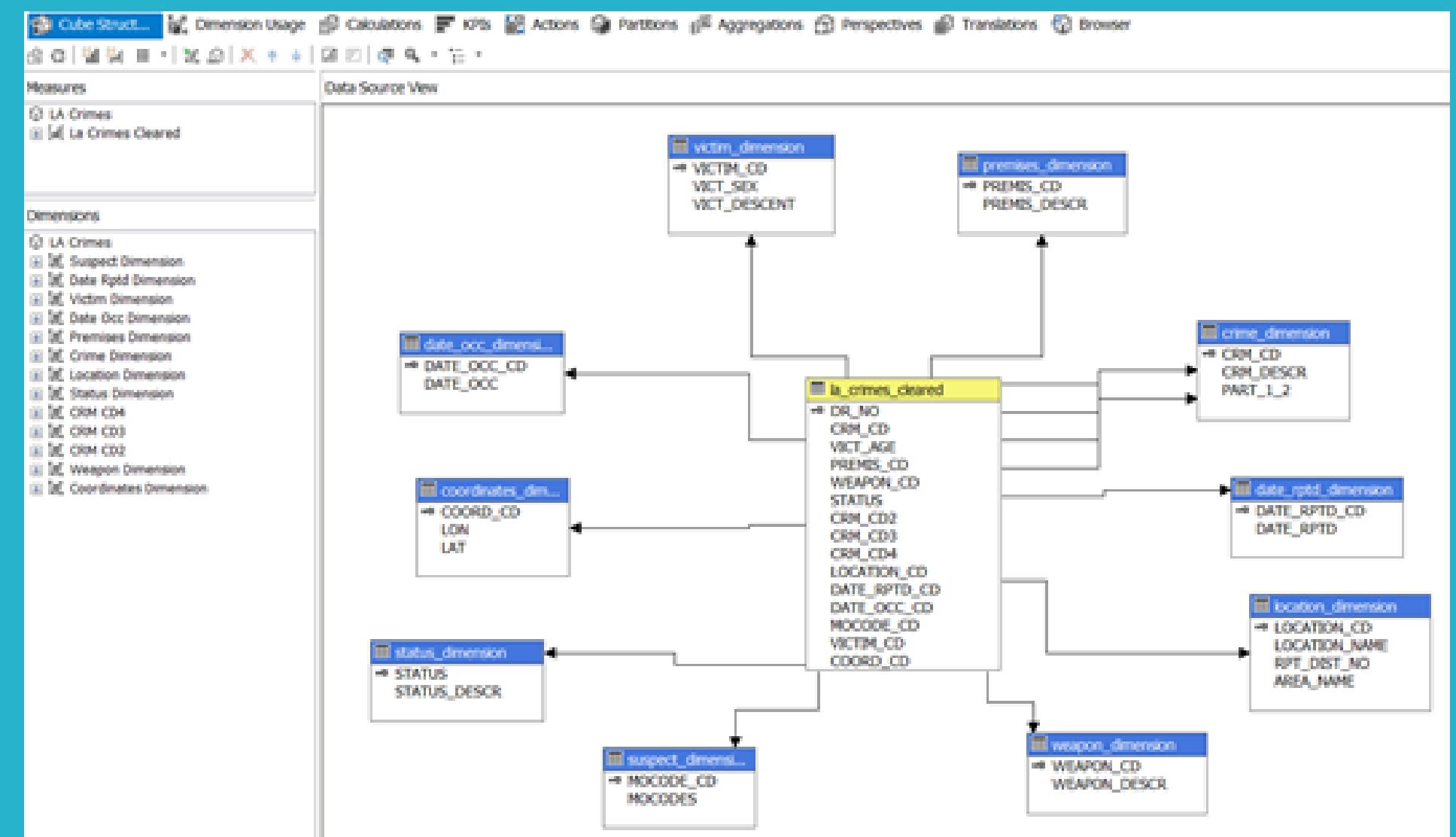
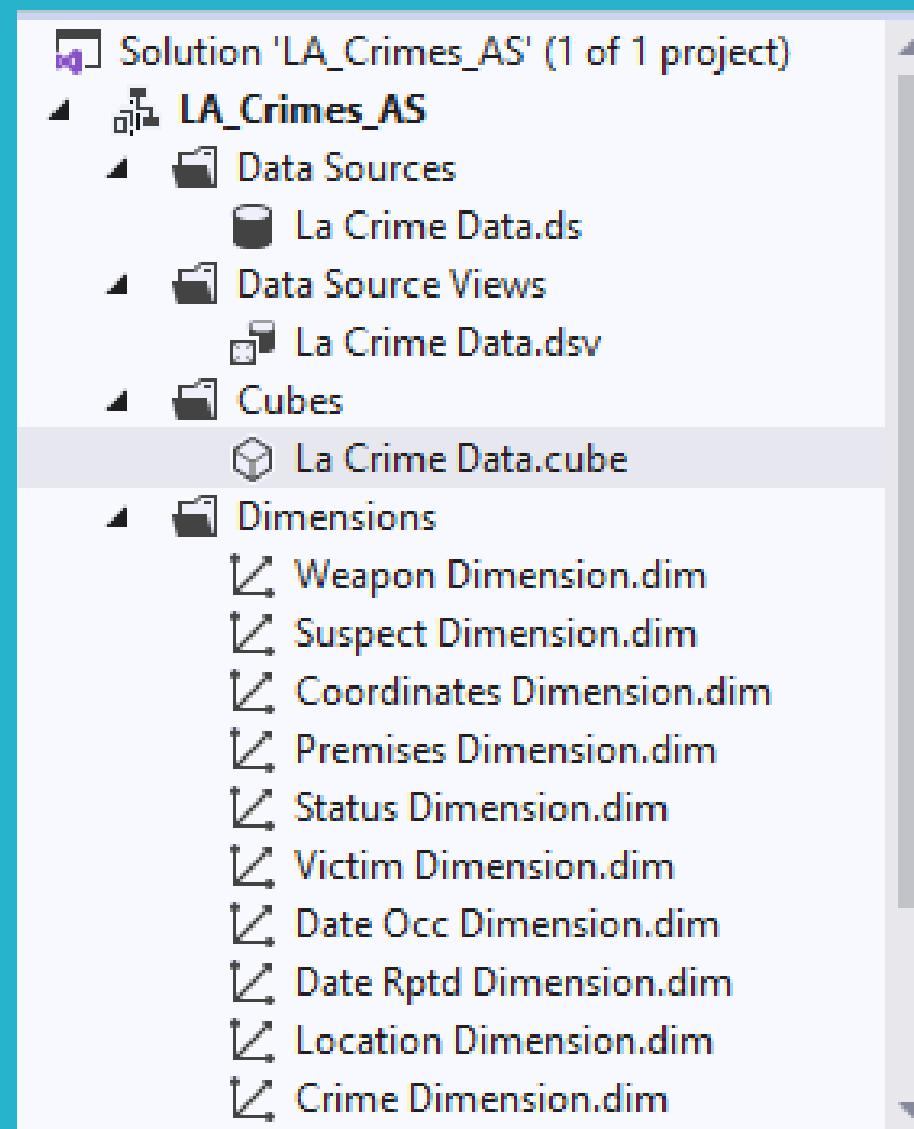


# CUBE

- USED LA CRIMES DATABASE FROM SQL SERVER WAREHOUSE AS A DATA SOURCE



## CUBE STRUCTURE



# CUBE BROWSING WITH EXCEL

CRIME STATUS PER PREMISE,  
GROUPED BY SEX

Row Labels	La Crimes Cleared Count
2 F	
3 77th Street	
4 Adult Arrest	8,81%
5 Adult Other	14,99%
6 Invest Cont	74,84%
7 Juv Arrest	1,01%
8 Juv Other	0,34%
9 UNK	0,01%
10 Central	4,13%
11 Devonshire	4,39%
12 Foothill	3,66%
13 Harbor	4,18%
14 Hollenbeck	3,52%
15 Hollywood	4,01%
16 Mission	4,92%
17 N Hollywood	5,08%
18 Newton	4,64%
19 Northeast	4,44%
20 Olympic	4,44%
21 Pacific	4,82%
22 Rampart	4,23%
23 Southeast	6,42%
24 Southwest	7,17%
25 Topanga	4,57%
26 Van Nuys	4,74%
27 West LA	4,21%
28 West Valley	4,14%
29 Wilshire	4,23%
30 M	
31 77th Street	
32 Adult Arrest	8,21%
33 Adult Other	9,17%
34 Invest Cont	81,15%
35 Juv Arrest	1,24%
36 Juv Other	0,21%
37 UNK	0,01%
38 Central	5,73%
39 Devonshire	4,54%
40 Foothill	3,83%
41 Harbor	4,05%

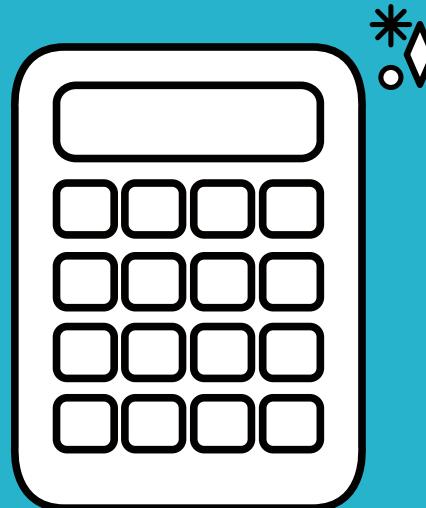
CRIME STATUS PER DESCENT

Row Labels	La Crimes Cleared Count
2 Adult Arrest	
3 B	2,09%
4 H	6,07%
5 W	2,87%
6 Adult Other	
7 B	3,59%
8 H	6,58%
9 W	3,52%
10 Invest Cont	
11 B	15,42%
12 H	32,83%
13 W	25,88%
14 Juv Arrest	
15 B	0,17%
16 H	0,51%
17 W	0,16%
18 Juv Other	0,31%
19 UNK	0,00%
20 Grand Total	100,00%

NUMBER OF INCIDENTS WHERE  
CERTAIN WEAPON WAS USED  
FOR EACH CRIME CATEGORY

36 BOW AND ARROW	
37 1.0	19
38 2.0	0
39 BOWIE KNIFE	
40 1.0	57
41 2.0	6
42 BRASS KNUCKLES	
43 1.0	383
44 2.0	35
45 CAUSTIC CHEMICAL/POISON	
46 1.0	328
47 2.0	151
48 CLEAVER	
49 1.0	74
50 2.0	7
51 CLUB/BAT	
52 1.0	3697
53 2.0	667
54 CONCRETE BLOCK/BRICK	
55 1.0	498
56 2.0	389
57 DEMAND NOTE	
58 1.0	288
59 2.0	27
60 DIRK/DAGGER	
61 1.0	108
62 2.0	16
63 DOG/ANIMAL (SIC ANIMAL ON)	
64 1.0	24
65 2.0	16
66 EXPLOSIVE DEVICE	
67 1.0	68
68 2.0	97
69 FIRE	
70 1.0	464
71 2.0	34
72 FIXED OBJECT	
73 1.0	446
74 2.0	361

# CUBE CALCULATIONS



## VICTIM'S AGE NORMALIZATION

- TRANSFORMING EVERY VICT\_AGE VALUE TO A NUMBER BETWEEN 0 AND 1
- MISSING VALUES (VICT\_AGE = 0) ARE CHANGED TO NEGATIVE VALUES

### Normalization Formula

$$X_{\text{normalized}} = \frac{(X - X_{\text{minimum}})}{(X_{\text{maximum}} - X_{\text{minimum}})}$$



# CUBE CALCULATIONS

## CREATING THE CALCULATION

The screenshot shows the 'LA Crimecube [Design]' window in the Analysis Services Management Studio. The 'Calculations' tab is selected. A new calculation is being created with the following details:

- Name:** NORMALIZED\_AGE
- Parent Properties:** Parent hierarchy: Measure; Parent member: (empty)
- Expression:**  $([Measures].[Victim Age] - 1) / (100 - 1)$
- Additional Properties:** Format string: ; Is total: True; Non-empty behavior: (empty); Associated measure group: (undefined); Display folder: (empty); Color Expressions: (empty); Font Expressions: (empty).

The left sidebar shows the cube structure with nodes like 'LA Crime', 'Measures', 'La Crimes Cleared', 'La Crimes Cleared Count', 'YEST AGE', 'Dimensions', 'Crime Dimension', 'Crime COO', 'Crime COO', 'CRM COO', and 'Data Oct Dimension'.

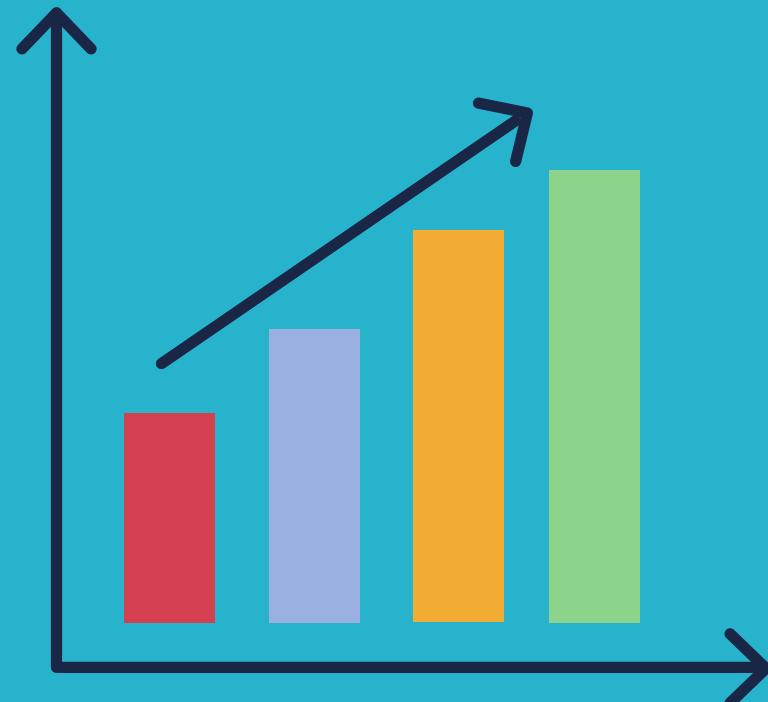
## SHOWING THE NORMALIZED AGE

Row Labels	La Crimes Cleared Count	NORMALIZED_AGE
2010-01-01 00:02:00.000	1	0,205128205
2010-01-01 00:07:00.000	1	0,213675214
2010-01-01 00:25:00.000	1	0,224788325
2010-01-01 00:33:00.000	1	0,213675214
2010-01-01 00:43:00.000	1	0,216239316
2010-01-01 00:53:00.000	1	0,258410256
2010-01-01 00:59:00.000	1	0,058119658
2010-01-01 01:35:00.000	1	0,196581197
2010-01-01 01:40:00.000	1	-0,008547009
2010-01-01 01:45:00.000	1	0,273504274
2010-01-01 01:53:00.000	1	-0,008547009
2010-01-01 01:55:00.000	1	0,258410256
2010-01-01 02:10:00.000	1	0,273504274
2010-01-01 02:15:00.000	1	0,282051282
2010-01-01 02:40:00.000	1	0,162393162
2010-01-01 02:45:00.000	1	0,555555556
2010-01-01 02:55:00.000	1	0,222222222
2010-01-01 03:10:00.000	1	0,247883248
2010-01-01 03:35:00.000	1	0,478632479
2010-01-01 03:40:00.000	1	0,504273504
2010-01-01 03:50:00.000	1	0,250788231
2010-01-01 04:20:00.000	1	0,188034188
2010-01-01 04:35:00.000	1	0,179487179
2010-01-01 04:40:00.000	1	-0,008547009
2010-01-01 04:45:00.000	1	0,196581197
2010-01-01 04:55:00.000	1	-0,008547009
2010-01-01 05:55:00.000	1	0,239316239

# VISUALIZATIONS



- POWER BI ENVIRONMENT,  
DATA LOADED FROM ANALYSIS  
SERVICES PROJECT



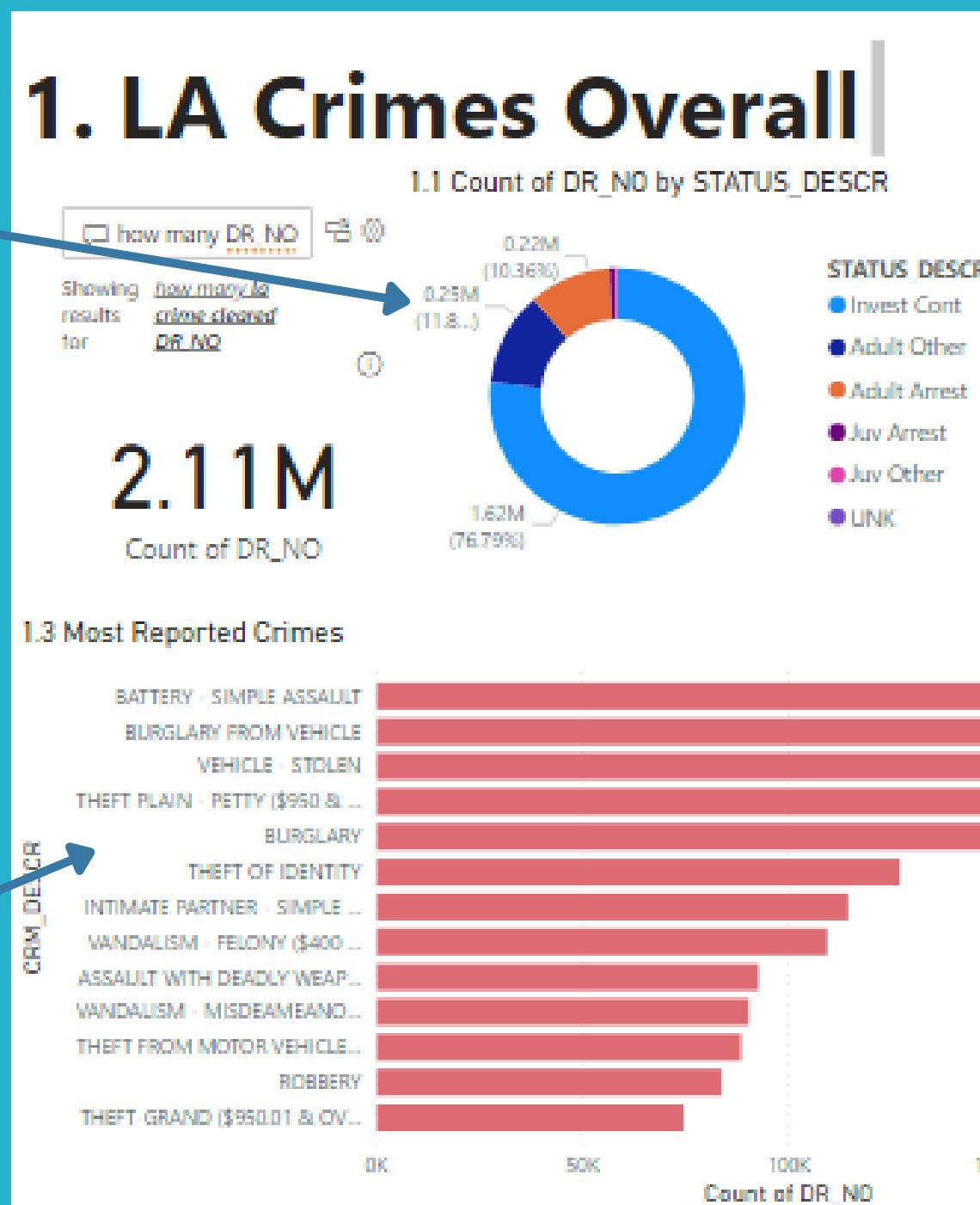
- JUPYTER NOTEBOOK  
ENVIRONMENT FOR MORE  
COMPLEX AND DEMANDING  
VISUALIZATIONS



## Total Crimes Timeline



Percentage  
of closed  
cases



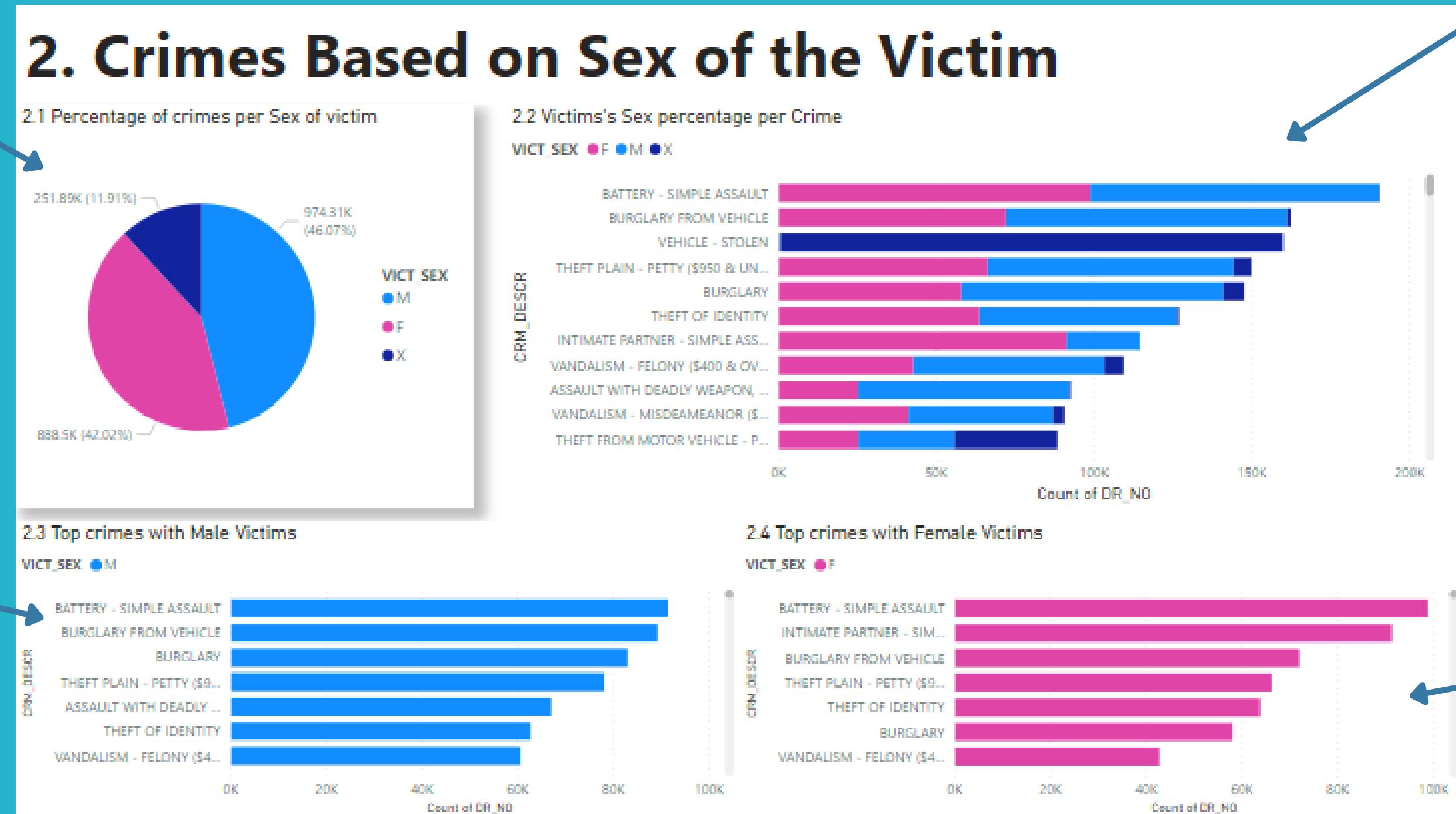
Most  
Reported  
Crimes

Areas with  
most  
Reported  
Crimes

Percentage  
of Victim's  
Sex for  
most  
Common  
Crimes

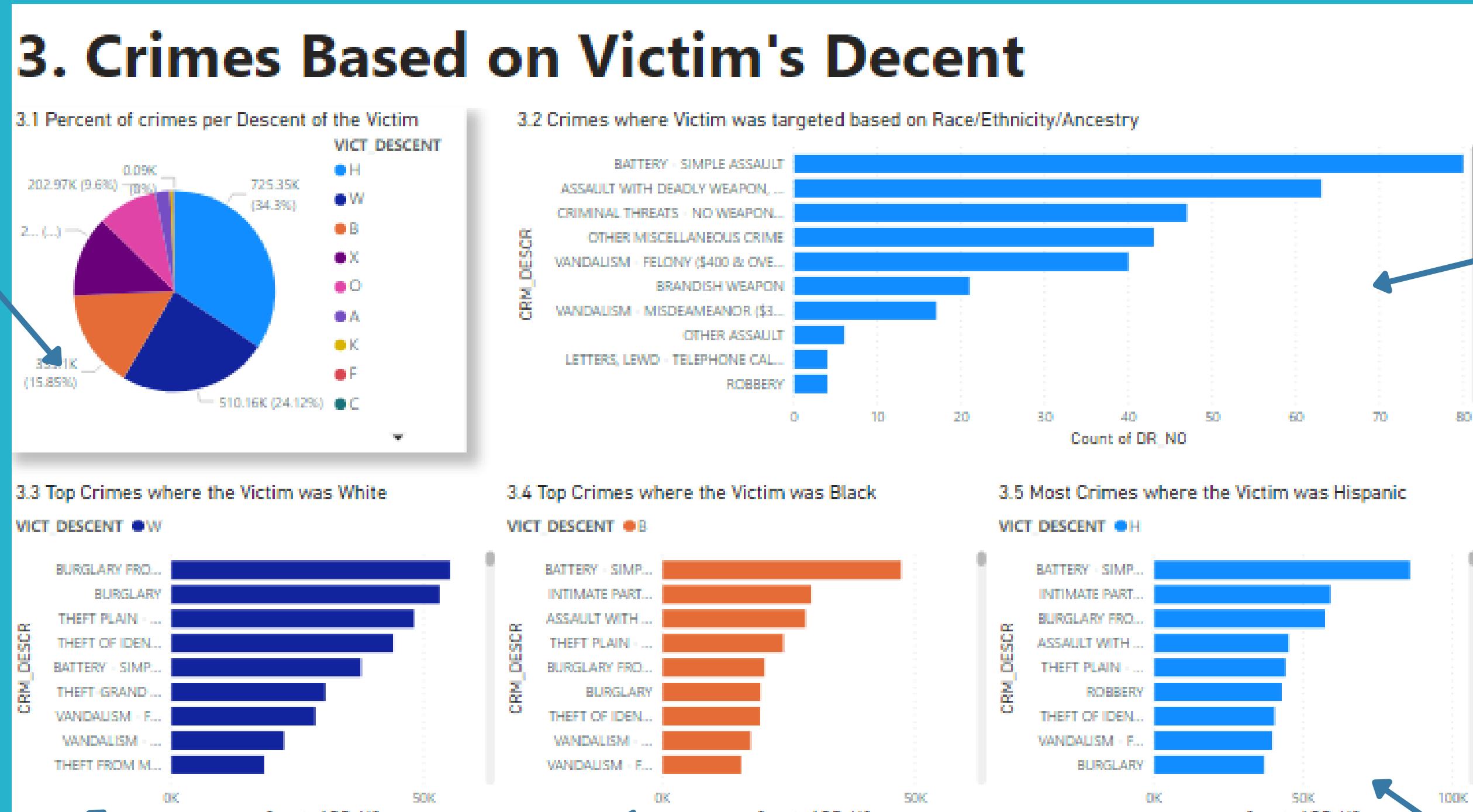
Percentage  
of Victim's  
Sex

Most usual  
Crimes with  
Male  
Victims



Most usual  
Crimes with  
Female  
Victims

Percentage  
of Victim's  
Descent



Most usual Crimes  
with Victims of  
White Descent

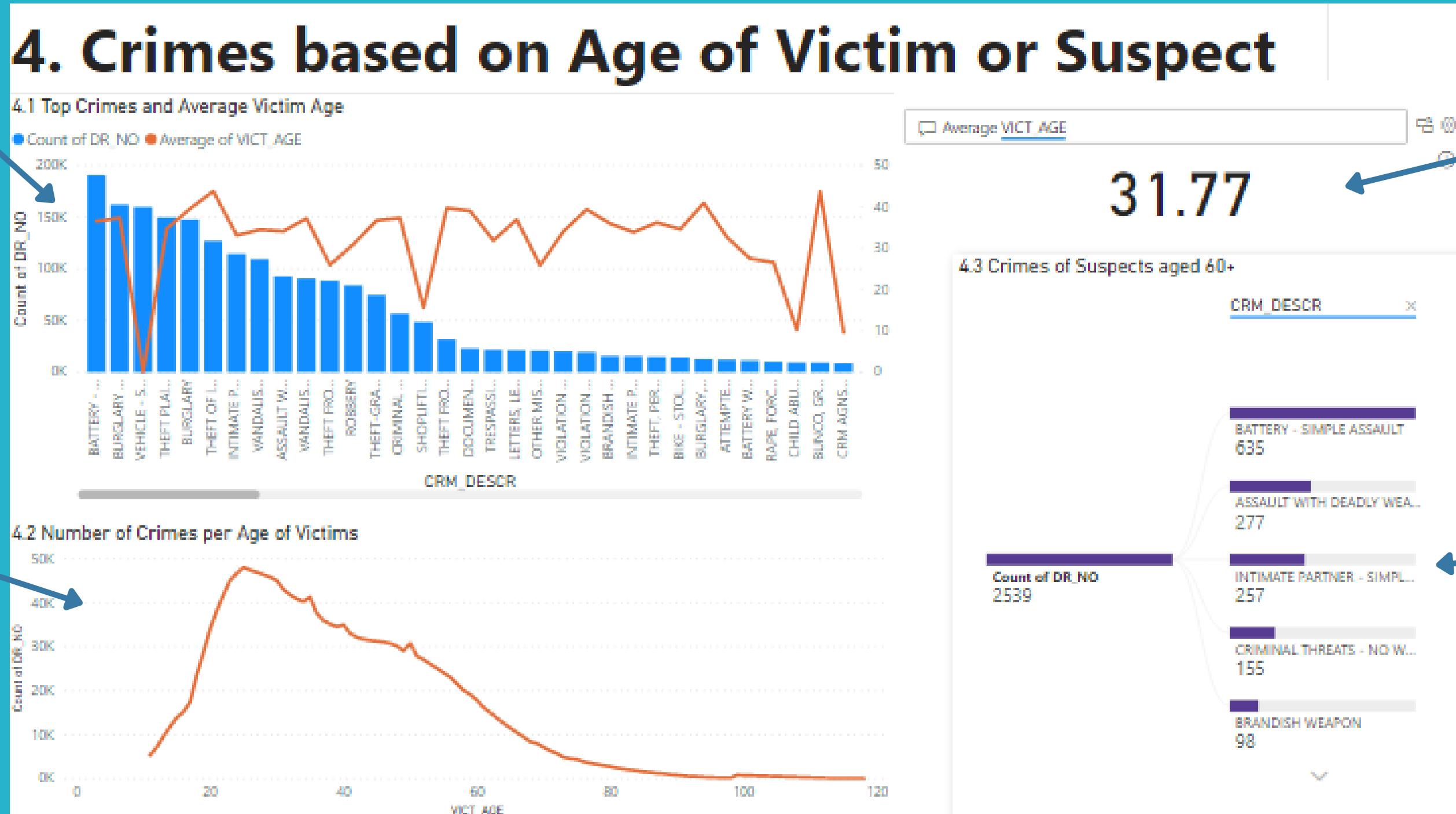
Most usual Crimes with  
Victims of Black Descent

Crimes  
where  
Victim  
Descent  
was  
Targeted

Most usual Crimes with  
Victims of Hispanic  
Descent

Average  
Victim Age  
per Crime

Number  
of Crimes  
per Age of  
Victims



Average  
Victim's  
Age

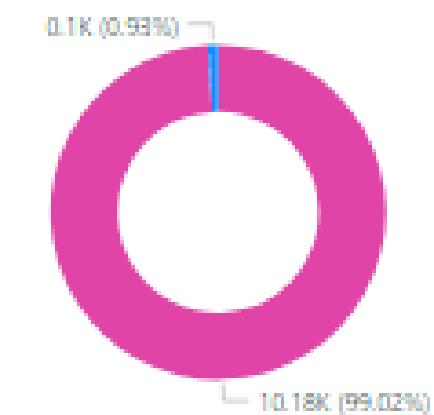
Crimes  
where  
Suspect  
was 60+  
years old

Percentage  
of  
Male/Female  
Victims

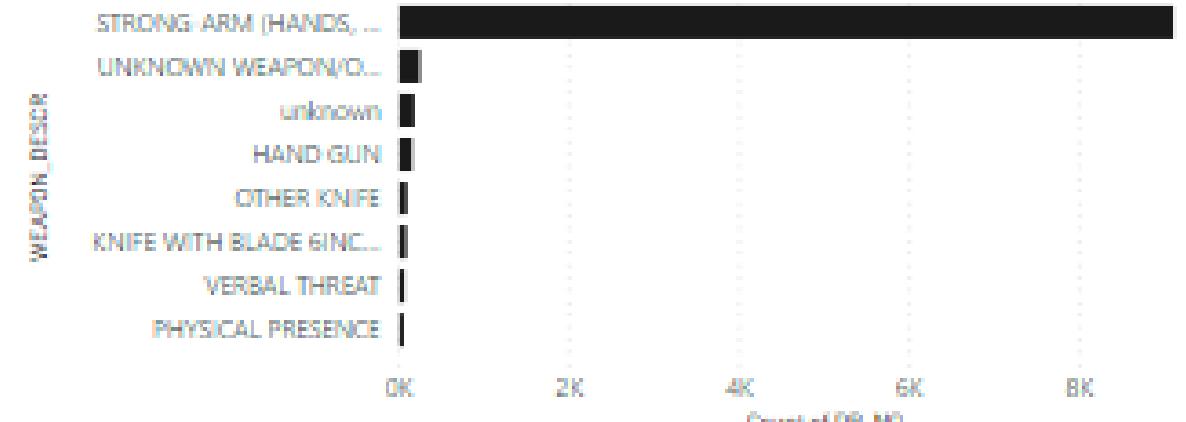
Percentage  
of  
closed/open  
Cases

## 5. Crimes of Rape

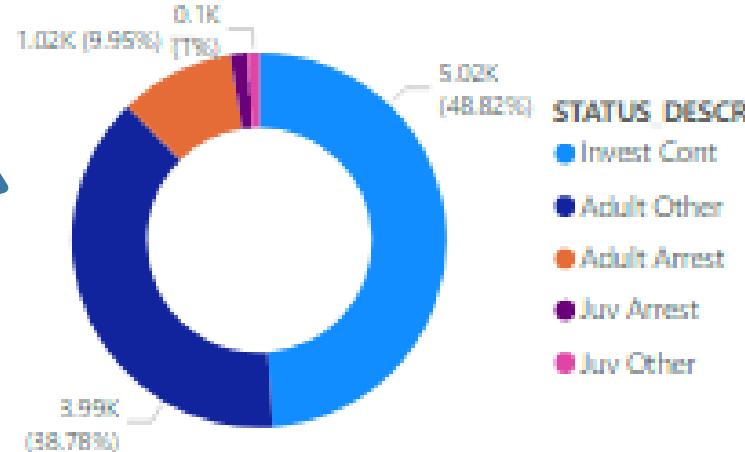
5.1 Percentage of Crimes for each Victim Sex



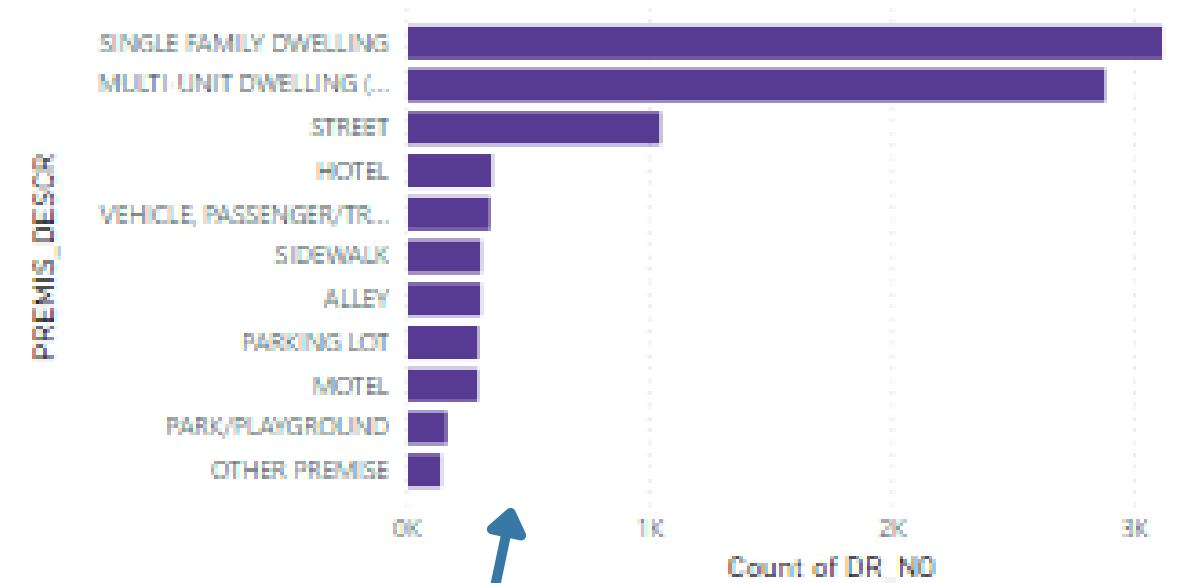
5.2 Most used Weapons



5.3 Percentage of Crimes per Status of Case



5.4 Most usual Premises where Crime was occurred



Most usual Premises

Most used Weapons

Number of Rape incidents

how many DR\_NO with CRM\_DESCR RAPE, FORCIBLE

Showing how many 1a crime cleared results DR\_NO with CRM\_DESCR for RAPE, FORCIBLE

10.28K

Count of DR\_NO

Number of  
Rape  
Cases

Average Age of Victim

average VICT\_AGE for CRM\_DESCR RAPE, FORCIBLE

26.57

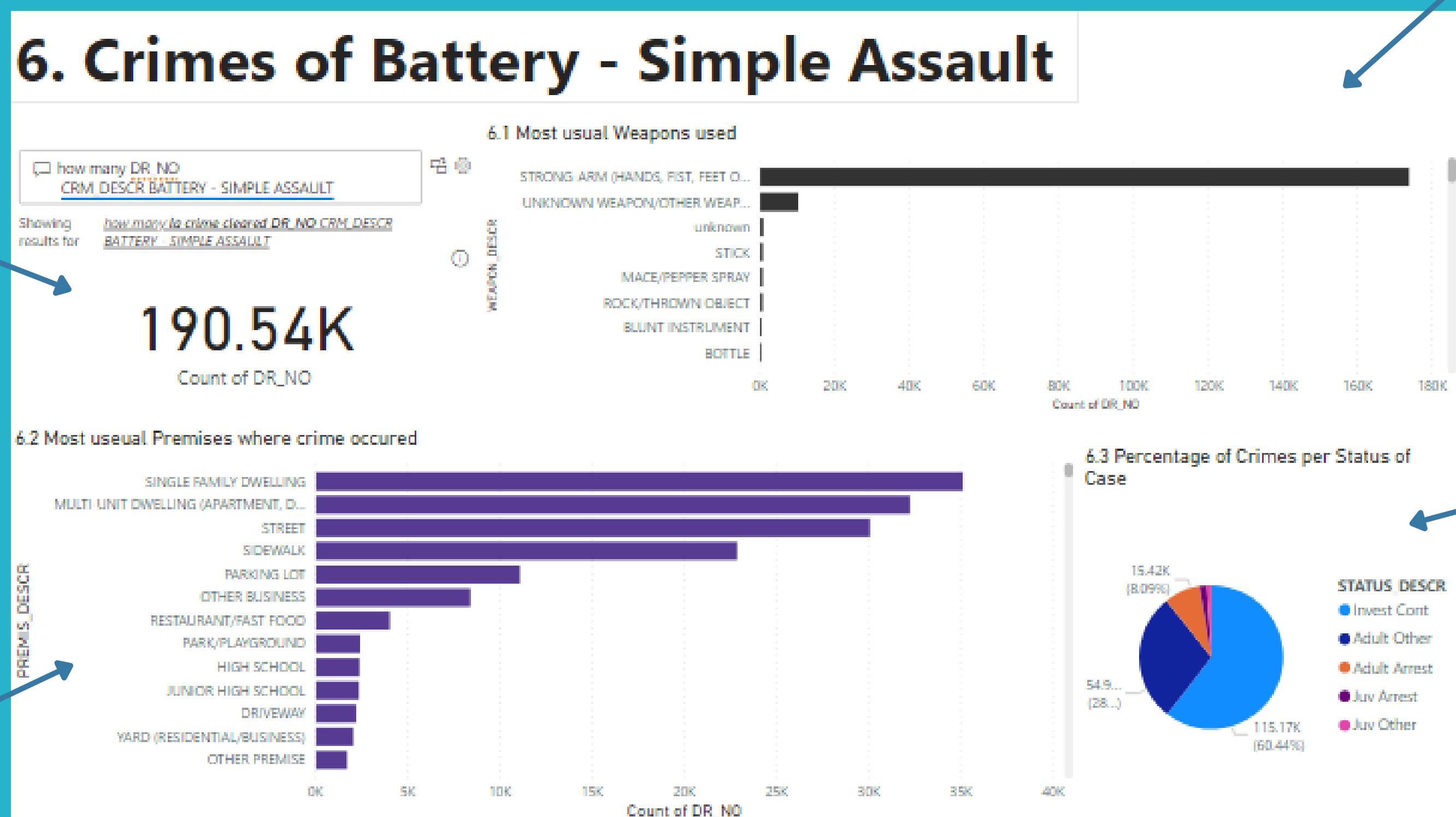
Average of VICT\_AGE

Average age  
of Victim

Number of  
Battery -  
Simple  
Assault

Incidents

Most  
usual  
Premises  
of Battery



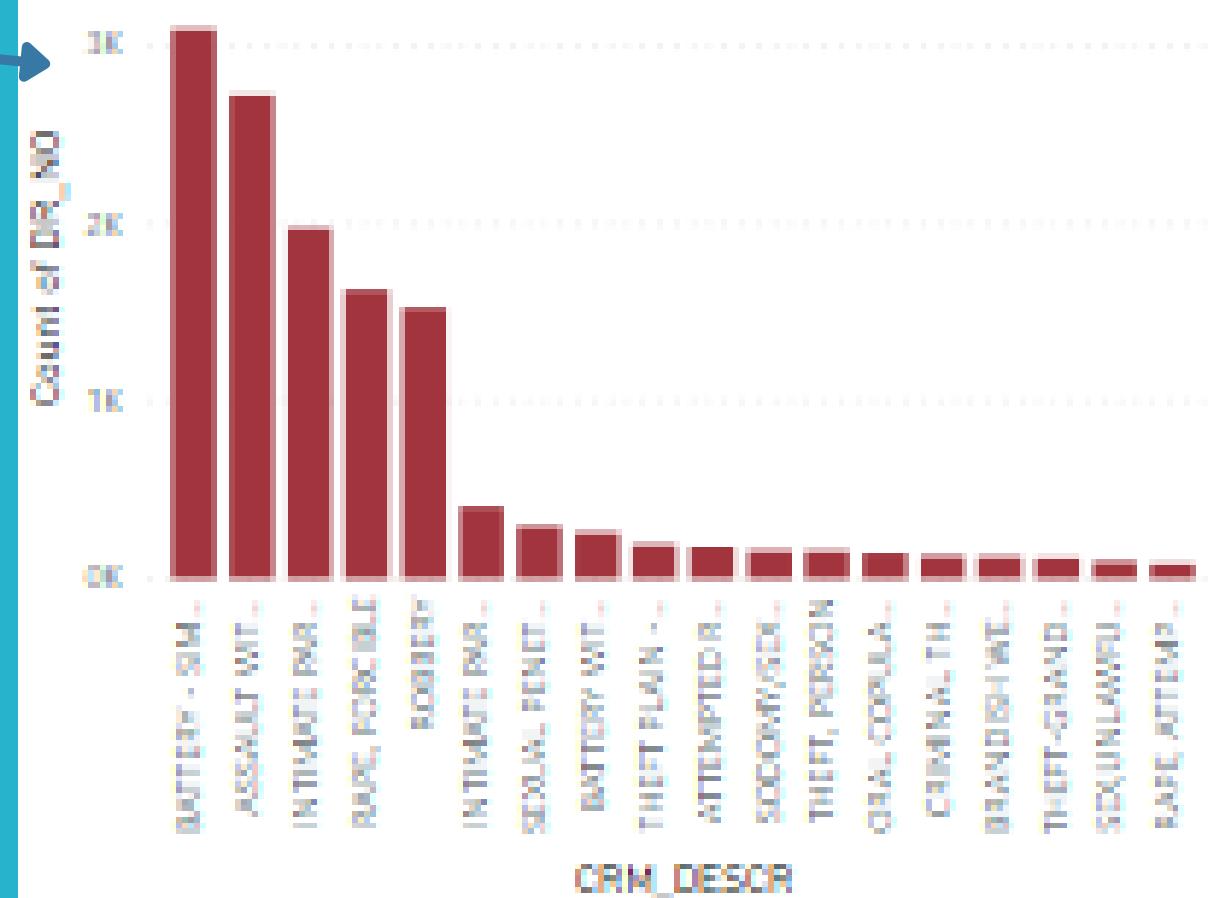
Most used  
weapons

Percentage  
of  
closed/open  
Cases

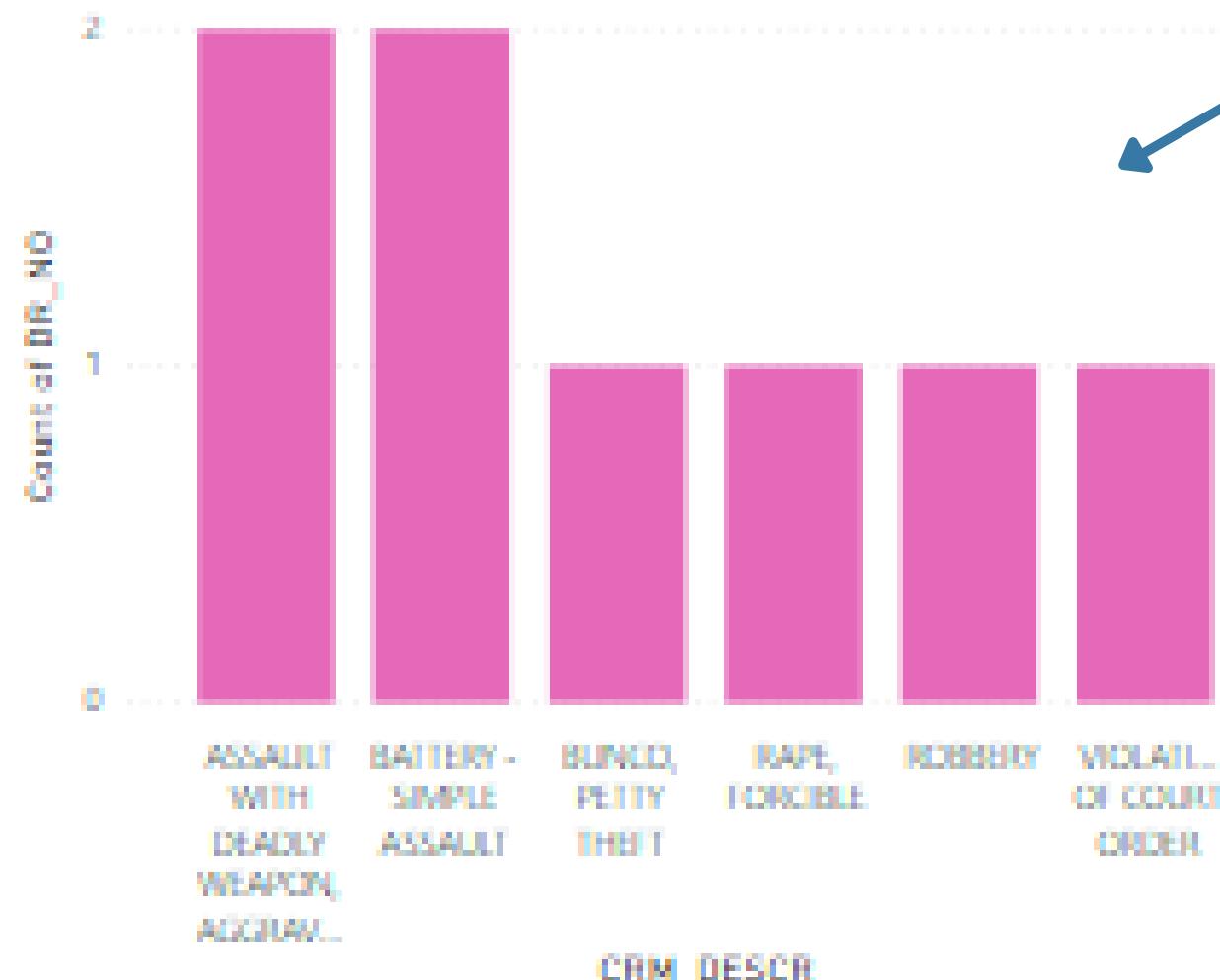
Crimes  
where  
Suspect  
was  
Undocume-  
nted Alien

## 8. Other

8.1 Suspect was intoxicated or drunk



8.2 Incidents where the suspect was dressed in a costume(Spiderman, Darth Vader etc)

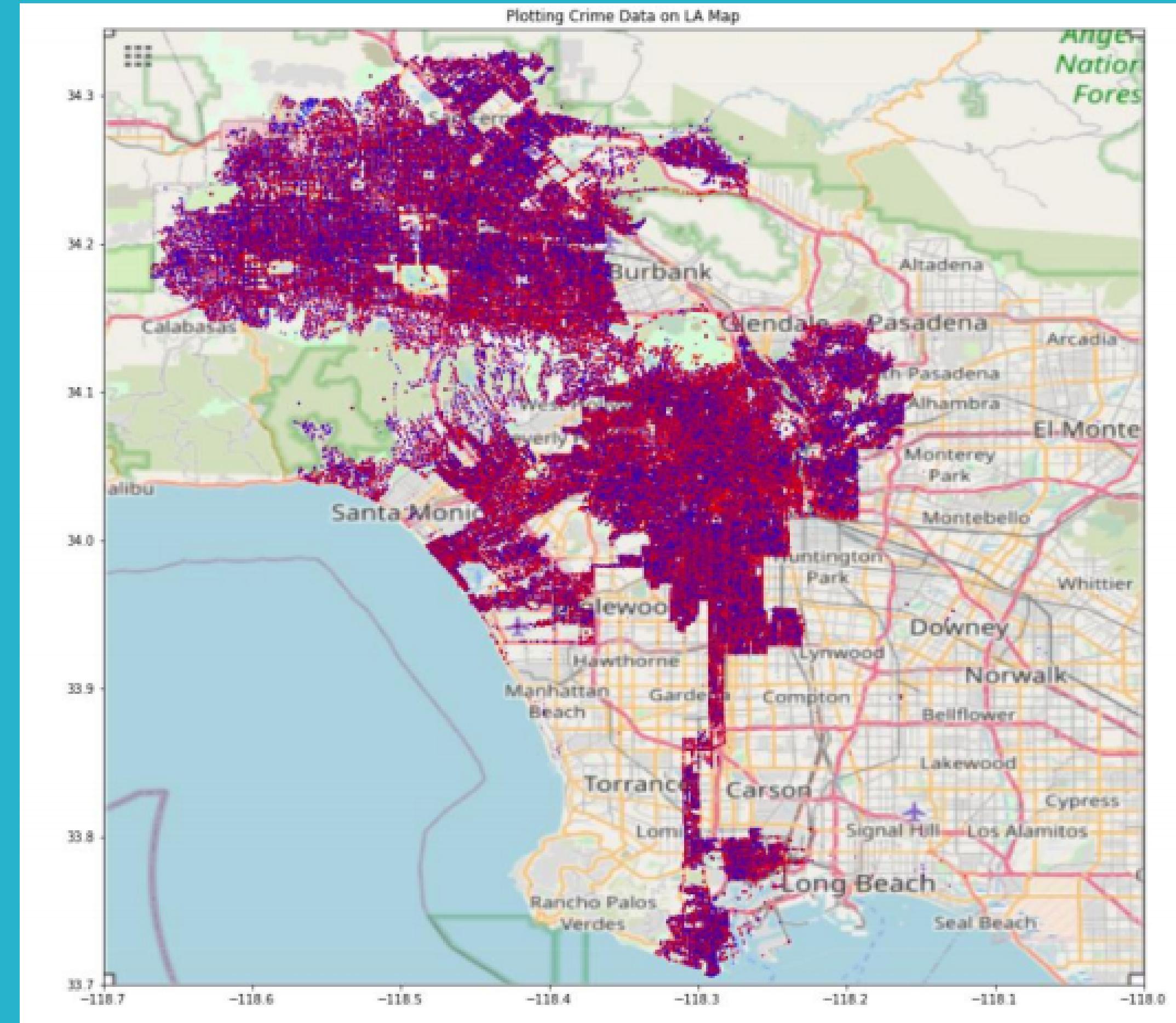


Crimes where  
Suspect was  
Dressed in a  
Character  
Costume



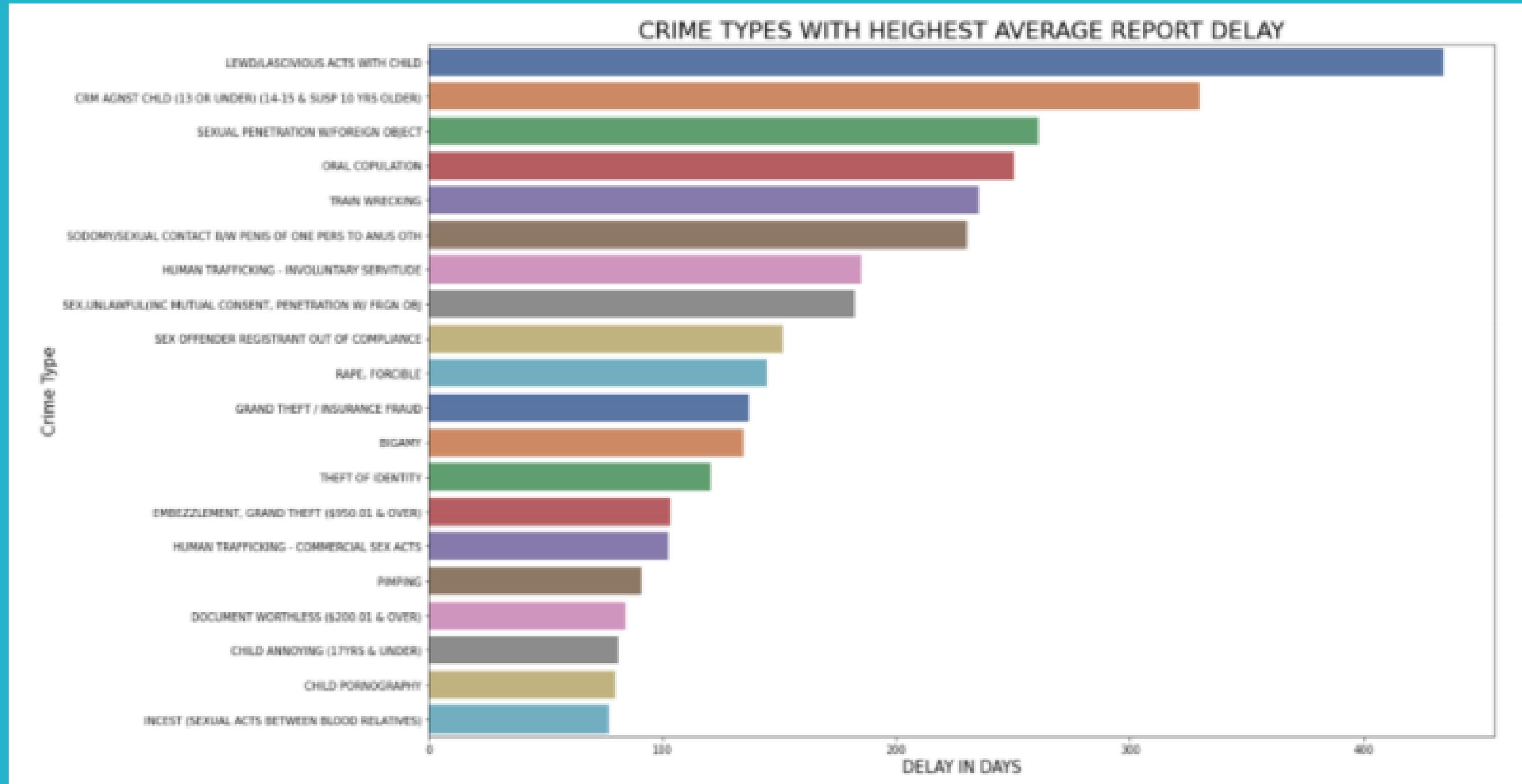
# CRIME INCIDENTS ON LA MAP

- Part 1 Offense (Robbery, Murder, Rape etc.)
- Part 2 Offense (Theft of Identity, Simple Assault etc.)



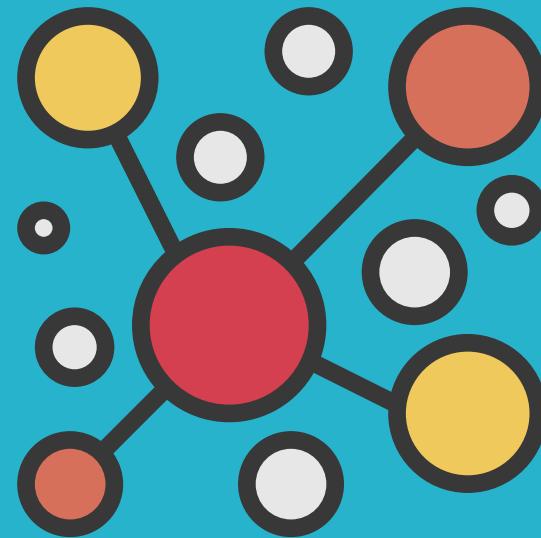
# CRIME INCIDENTS WITH HIGHEST REPORT DELAY

- Sexual Crimes
- Crime against Child
- Crime where Victim was not Aware of the Crime (e.g Theft of Identity)



# DATA MINING

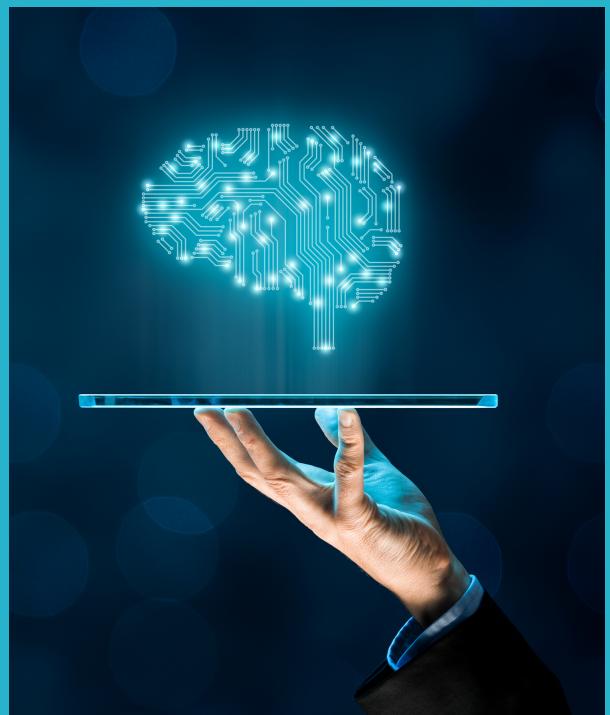
## CLUSTERING



- GEOGRAPHIC COORDINATE CLUSTERING
- DISTRICT CLUSTERING ON CRIMINALITY

## PREDICTION

- CRIME TYPE PREDICTION FROM REST OF INFORMATION



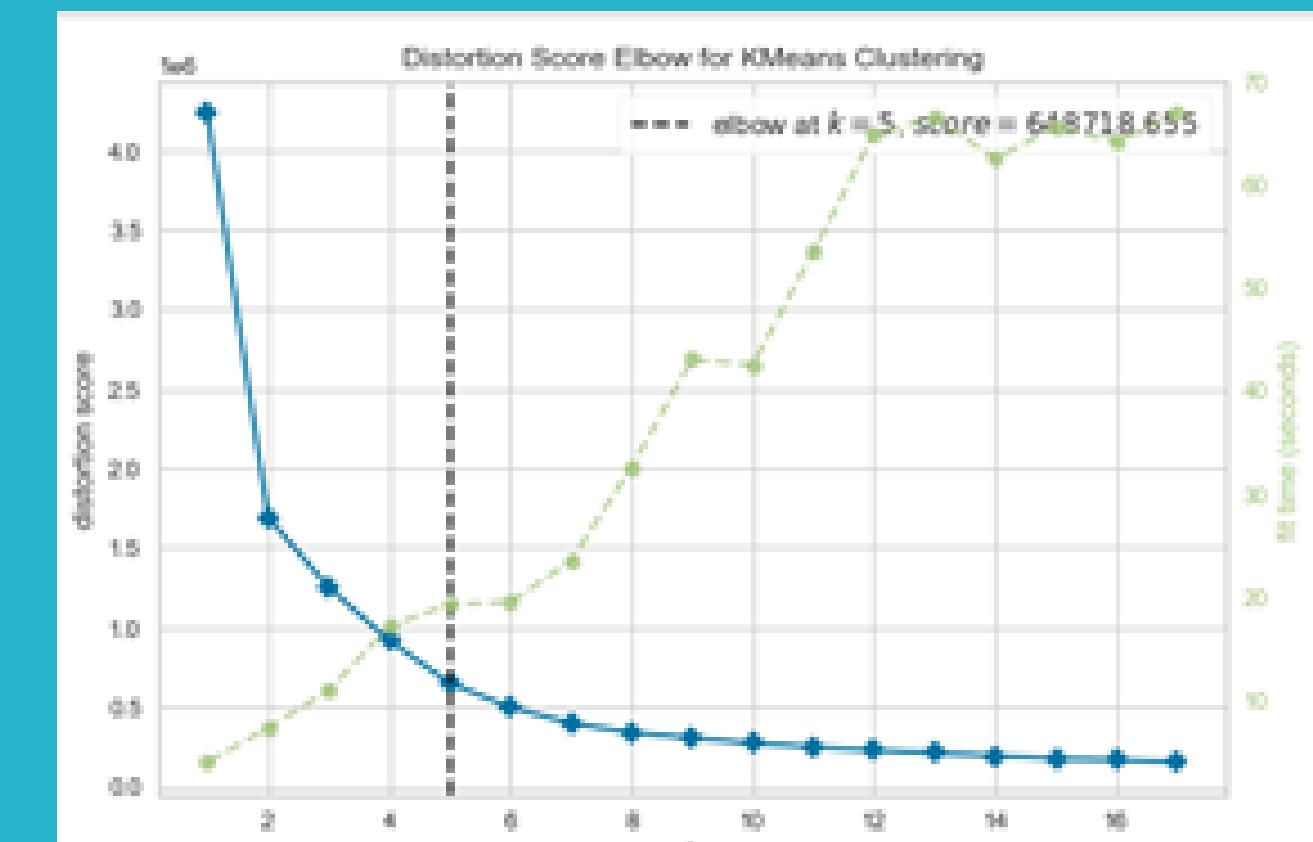
→ JUPYTER NOTEBOOK & PYTHON





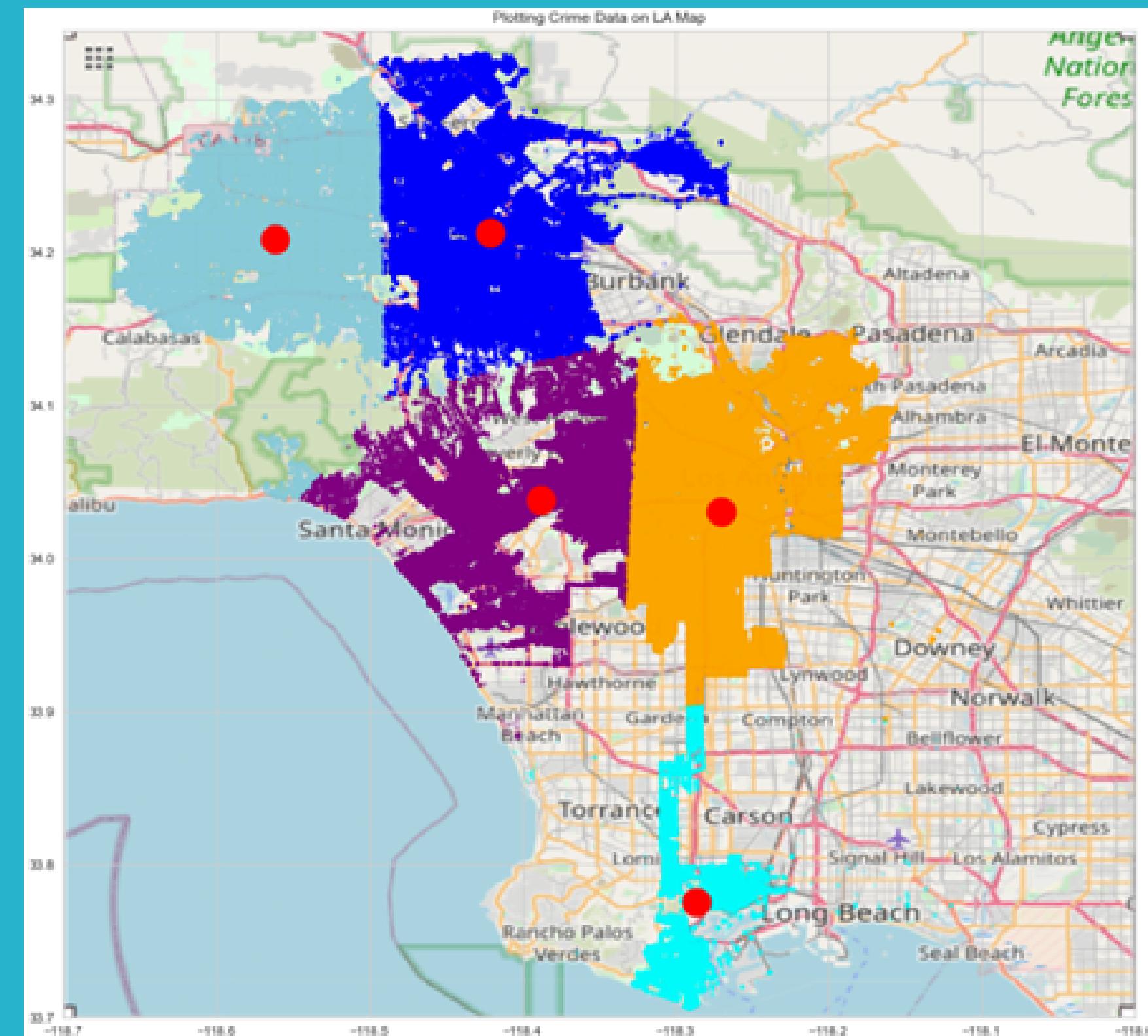
# GEOGRAPHIC COORDINATE CLUSTERING

- EACH ENTRY IS A DIFFERENT INCIDENT FROM 2010 - 2019
- FOR EACH INCIDENT ONLY USED LATITUDE AND LONGITUDE
- STANDARDIZED VALUES & USED K-MEANS ALGORITHM
- BEST NUMBER OF CLUSTERS ARE 5 ACCORDING TO ELBOW METHOD



# GEOGRAPHIC CLUSTERING RESULTS

- RED POINTS ARE CENTROIDS
- EACH OTHER COLOR DEFINES A CLUSTER
- EACH CLUSTER CONTAINS INCIDENTS THAT HAPPENED IN CLOSE (SIMILAR) AREAS
- IF THERE COULD ONLY BE 5 POLICE STATIONS IN LA, THE CENTROIDS COULD BE THEIR LOCATIONS



# DISTRICT CRIMINALITY CLUSTERING

- CREATED PIVOT TABLE WHICH INCLUDES FOR EACH DISTRICT THE SUM AMOUNT OF DIFFERENT CRIME CATEGORIES
- REMOVED CRIME CATEGORIES NOT CONNECTED TO DISTRICT E.G. INTIMATE PARTNER CRIMES
- STANDARDIZED VALUES TO HAVE A MEAN OF 0 AND STANDARD DEVIATION OF 1
- BEST NUMBER OF CLUSTERS ARE 3 USING ELBOW METHOD
- USED K-MEANS ALGORITHM

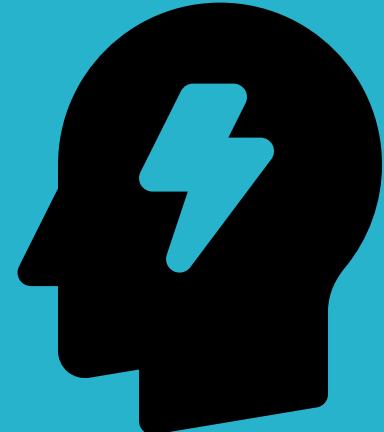


RESULTS

# DISTRICT CLUSTERING RESULTS

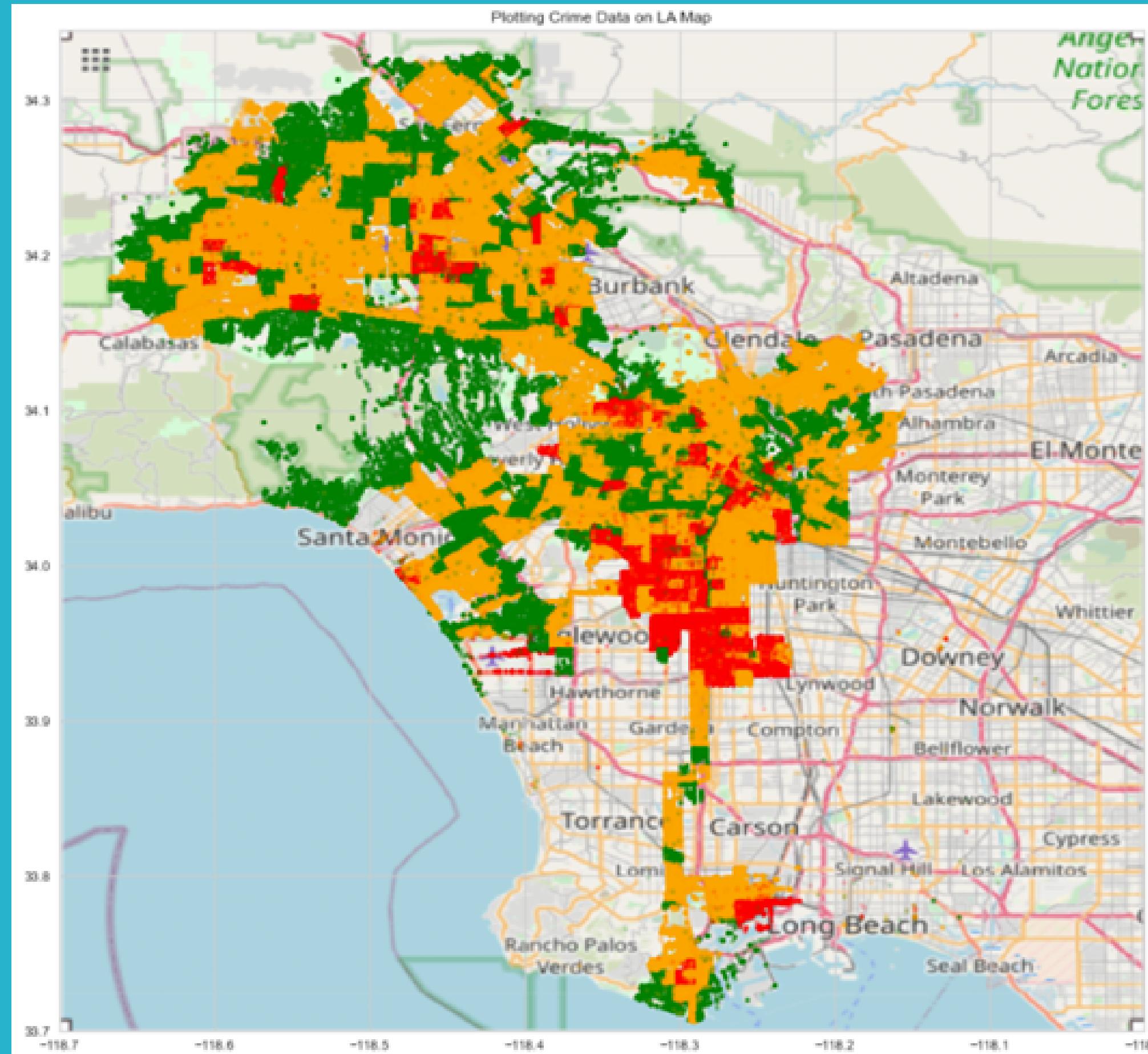
## CLUSTER CENTROID VALUES:

- FIRST CENTROID VALUES ARE MUCH BELOW 0 FOR ALL CRIME TYPES
- SECOND CENTROID VALUES ARE MORE THAN 1 IN MOST CRIME TYPES
- THIRD CENTROID VALUES ARE AROUND 0 FOR CRIME TYPES



STANDARDIZED VALUES HAVE MEAN 0 → VALUABLE CONCLUSIONS

# DISTRICT RESULTS VISUALIZATION



-  **SAFER ZONES**
-  **AVERAGE CRIMINALITY**
-  **DANGEROUS ZONES**

# CRIME TYPE PREDICTION

DATA FROM 2 DIFFERENT MONTHS OF 2019



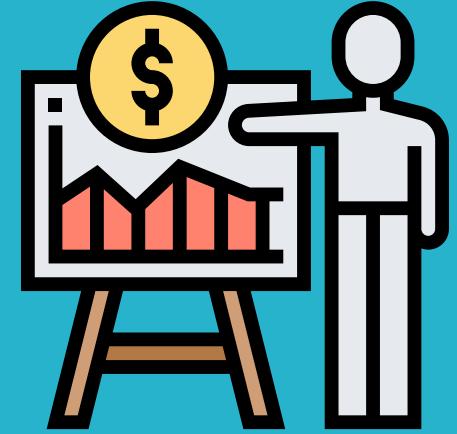
MAIN CRIME TYPE IS PREDICTED FROM THE REST  
INFORMATION

DATA PREPARATION:

- CREATED DUMMY VARIABLES FOR ALL CATEGORICAL DATA
- SHUFFLED DATA
- SEPARATED DATA INTO TRAINING (80%) AND TESTING (20%)



# PREDICTION RESULTS

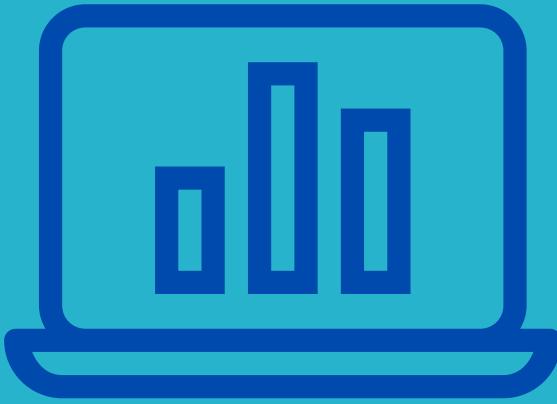


CLASSIFIER	AVG PRECISION	AVG RECALL	AVG F1 SCORE	ACCURACY
DECISION TREES	66%	67%	66%	67%
ADABOOST CLASSIFIER	68%	70%	68%	70%
RANDOM FOREST	67%	70%	65%	70%
BAGGING CLASSIFIER	69%	71%	68%	71%
EXTREMELY RANDOMIZED TREES	69%	71%	66%	71%



EXTREMELY RANDOMIZED TREES & BAGGING ARE THE MOST ACCURATE

# PREDICTION CONCLUSIONS



- USE OF RECALL AND PRECISION METRICS TO FIND THE BEST CLASSIFIER IN ANY CASE
  - E.G. CLASSIFIER THAT DOES NOT MISS ANY 'ROBBERY' → HIGHEST RECALL
  - E.G. CLASSIFIER THAT DOES NOT MISTAKENLY PREDICT 'MURDERS' → HIGHEST PRECISION
- THE MODELS ACTUALLY PREDICT THE PROBABILITIES OF THE INCIDENT TO BE ANY CRIME TYPE



KNOWING THE MOST PROBABLE CRIMES TYPES  
HELPS MAKING BETTER PREDICTIONS

WILLIAM