LOS ANGELES CRIME ANALYSIS

Business Intelligence & Big Data Analytics

KYPRITIDIS STEFANOS 8170050

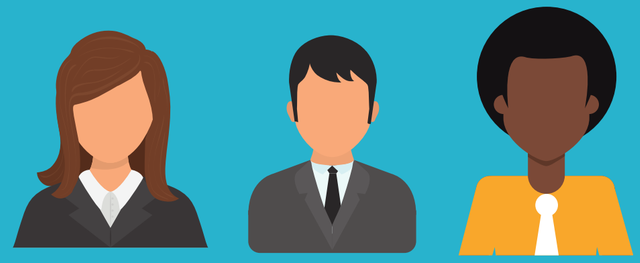
DATASET

- 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

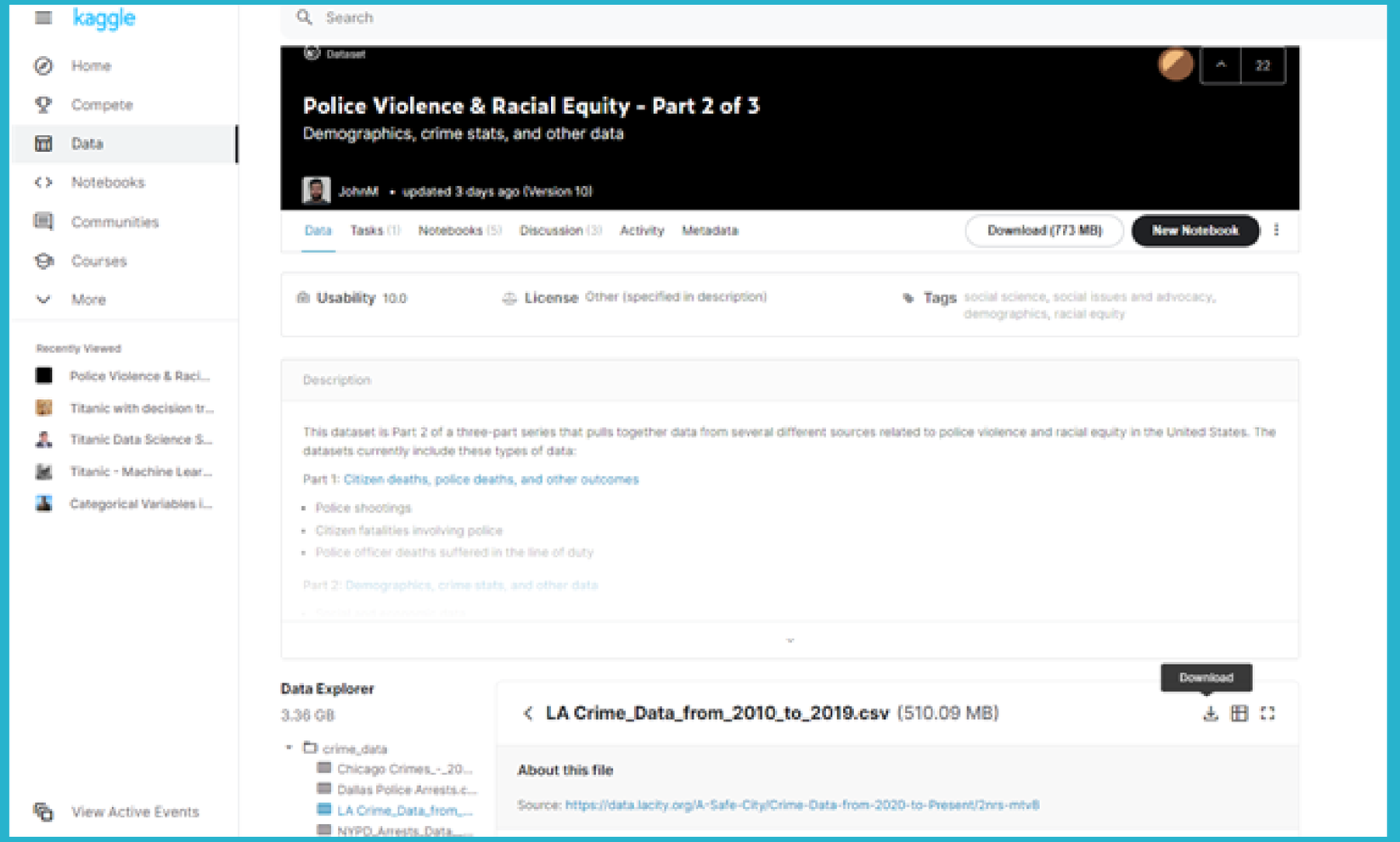


- CRIME DESCRIPTION AND EXTRA CRIMES COMMITED
- WEAPON USED AND PREMISES OF THE INCIDENT

- STATUS OF CASE
- SUSPECT ACTIVITIES
- GEOGRAPHIC AREA, LOCATION AND COORDINATES

DATA EXTRACTION

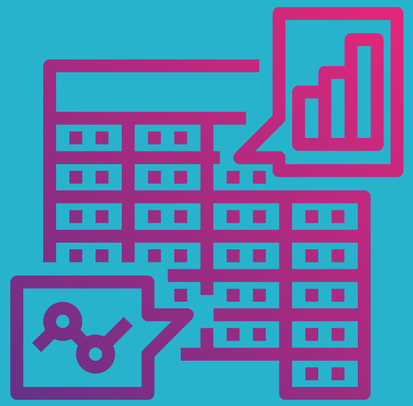
DOWNLOADED FROM KAGGLE PROJECT: *POLICE VIOLENCE &RACIAL EQUITY*



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	CAHU
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	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
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

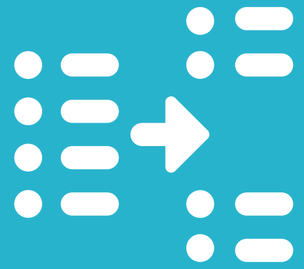
2114699 rows × 28 columns

DATA TRASNFORMATION

- 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

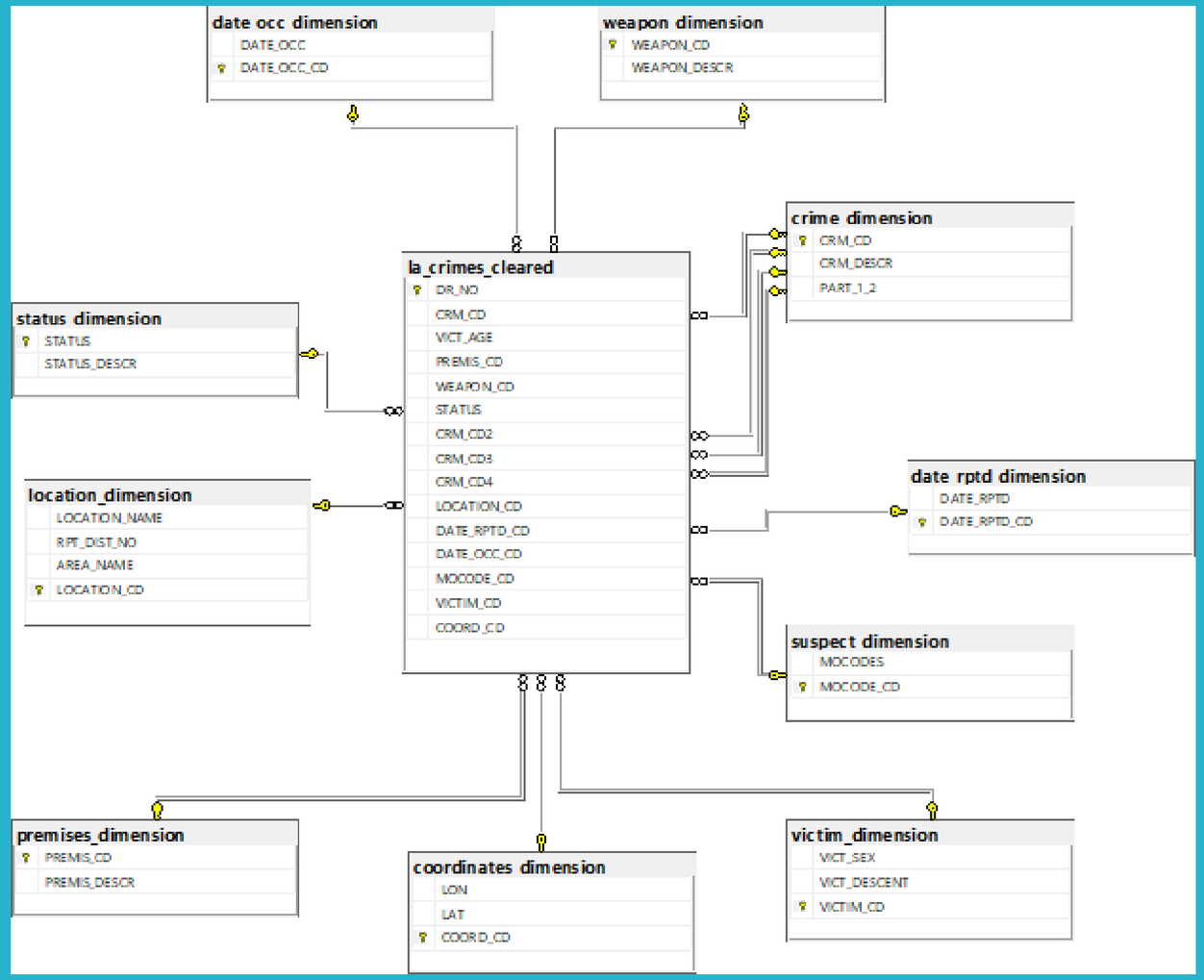


DATA LOADING

LOADED DATA INTO SQL SERVER WAREHOUSE

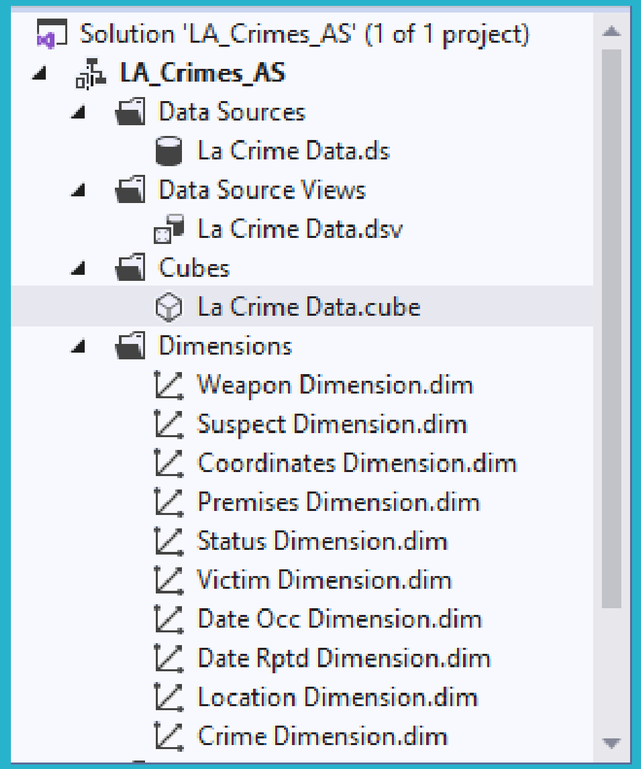
STAR SCHEMA:

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

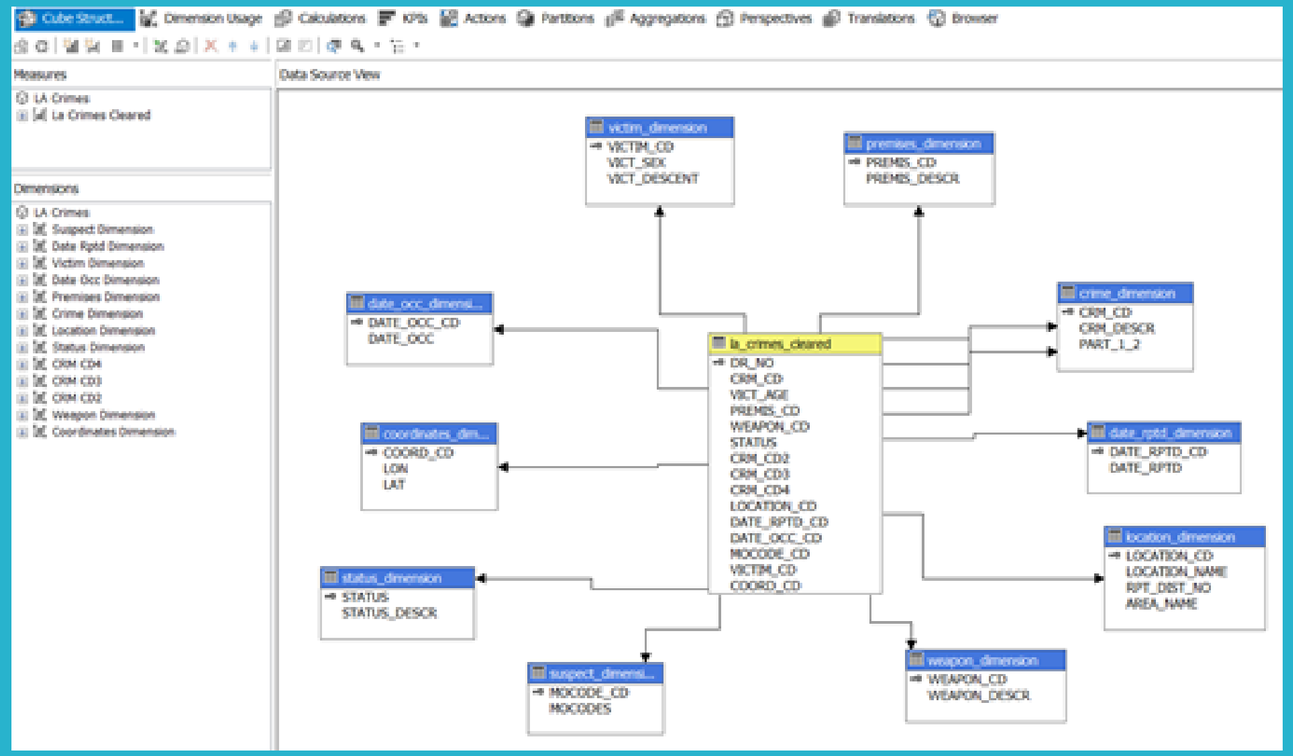


CUBE

- 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
F	
77th Street	
Adult Arrest	8,81%
Adult Other	14,99%
Invest Cont	74,84%
Juv Arrest	1,01%
Juv Other	0,34%
UNK	0,01%
Central	4,13%
Devonshire	4,39%
Foothill	3,66%
Harbor	4,18%
Hollenbeck	3,52%
Hollywood	4,01%
Mission	4,92%
N Hollywood	5,08%
Newton	4,64%
Northeast	4,44%
Olympic	4,44%
Pacific	4,82%
Rampart	4,23%
Southeast	6,42%
Southwest	7,17%
Topanga	4,57%
Van Nuys	4,74%
West LA	4,21%
West Valley	4,14%
Wilshire	4,23%
M	
77th Street	
Adult Arrest	8,21%
Adult Other	9,17%
Invest Cont	81,15%
Juv Arrest	1,24%
Juv Other	0,21%
UNK	0,01%
Central	5,73%
Devonshire	4,54%
Foothill	3,83%
Harbor	4,05%

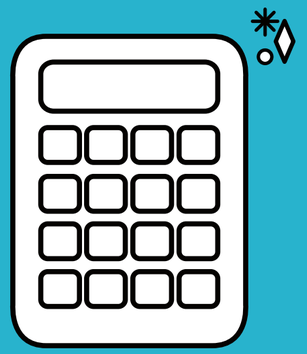
CRIME STATUS PER DESCENT

Row Labels	La Crimes Cleared Count
Adult Arrest	
B	2,09%
H	6,07%
W	2,87%
Adult Other	
B	3,59%
H	6,58%
W	3,52%
Invest Cont	
B	15,42%
H	32,83%
W	25,88%
Juv Arrest	
B	0,17%
H	0,51%
W	0,16%
Juv Other	0,31%
UNK	0,00%
Grand Total	100,00%

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

BOW AND ARROW	
1.0	19
2.0	6
BOWIE KNIFE	
1.0	57
2.0	6
BRASS KNUCKLES	
1.0	383
2.0	55
CAUSTIC CHEMICAL/POISON	
1.0	328
2.0	151
CLEAVER	
1.0	74
2.0	7
CLUB/BAT	
1.0	3697
2.0	667
CONCRETE BLOCK/BRICK	
1.0	498
2.0	389
DEMAND NOTE	
1.0	288
2.0	27
DIRK/DAGGER	
1.0	108
2.0	16
DOG/ANIMAL (SIC ANIMAL ON)	
1.0	24
2.0	16
EXPLOSIVE DEVICE	
1.0	68
2.0	97
FIRE	
1.0	464
2.0	34
FIXED OBJECT	
1.0	446
2.0	361

CUBE CALCULATIONS



VICTIM'S AGE NORMALIZATION

- TRANSFORMING EVERY VICT_AGE VALUE TO A NUMBER BETWEEN 0 AND 1

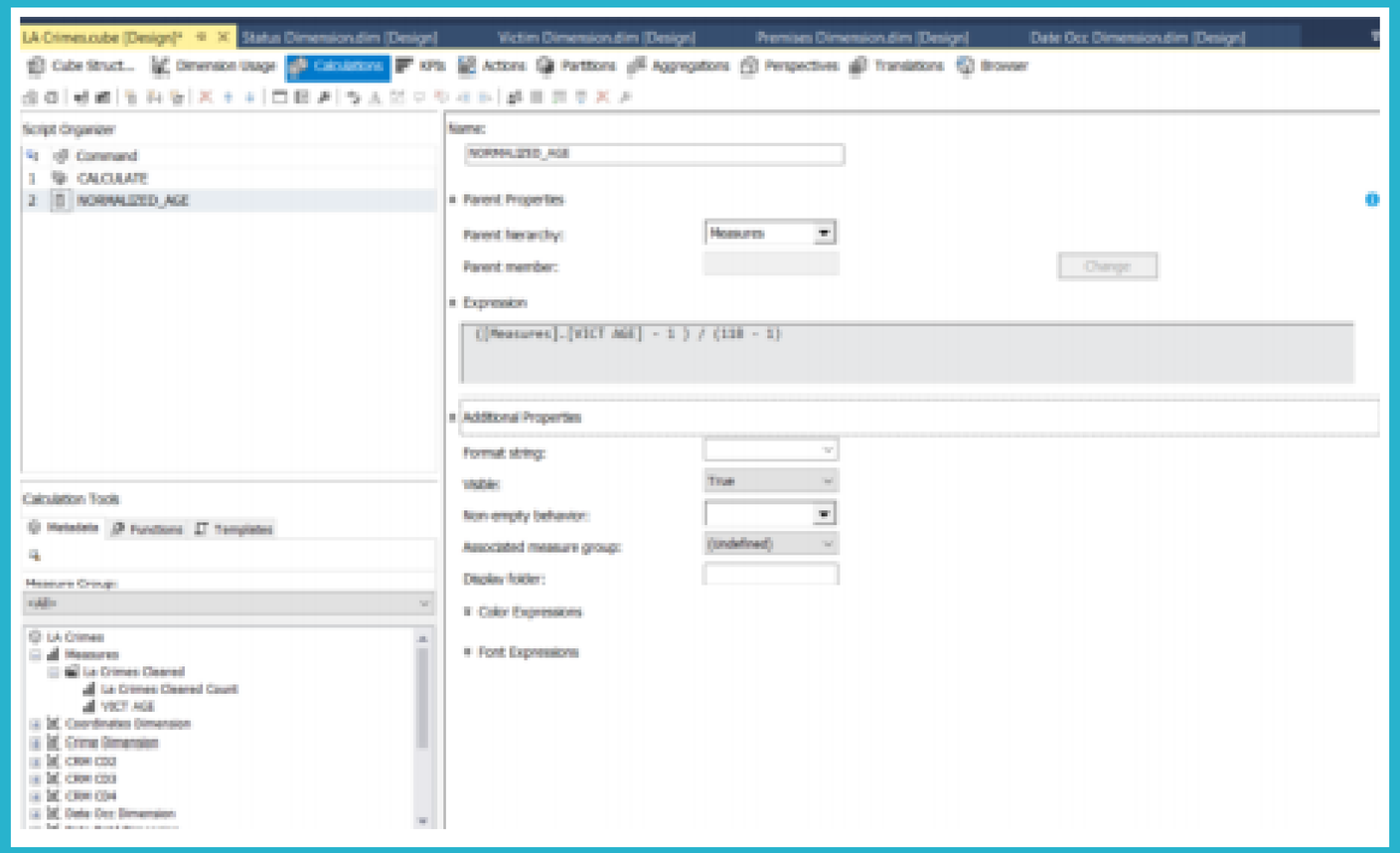
- MISSING VALUES (VICT_AGE = 0) ARE CHANGED TO NEGATIVE VALUES

Normalization Formula

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

CUBE CALCULATIONS

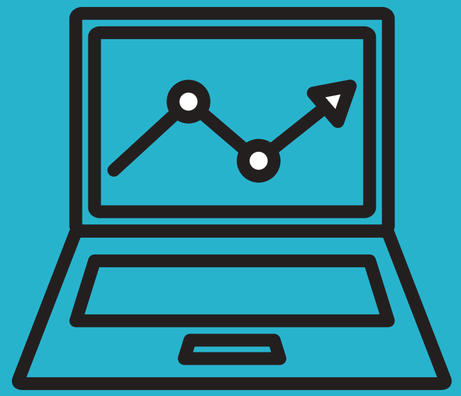
CREATING THE CALCULATION



SHOWING THE NORMALIZED AGE

	Row Labels	La Crimes Cleared Count	NORMALIZED_AGE
2	2010-01-01 00:02:00.000	1	0,205128205
3	2010-01-01 00:07:00.000	1	0,213675214
4	2010-01-01 00:25:00.000	1	0,324786325
5	2010-01-01 00:33:00.000	1	0,213675214
6	2010-01-01 00:43:00.000	1	0,316239316
7	2010-01-01 00:55:00.000	1	0,256410256
8	2010-01-01 00:59:00.000	1	0,058119658
9	2010-01-01 01:35:00.000	1	0,196581197
10	2010-01-01 01:40:00.000	1	-0,008547009
11	2010-01-01 01:45:00.000	1	0,273504274
12	2010-01-01 01:53:00.000	1	-0,008547009
13	2010-01-01 01:55:00.000	1	0,256410256
14	2010-01-01 02:10:00.000	1	0,273504274
15	2010-01-01 02:15:00.000	1	0,282051282
16	2010-01-01 02:40:00.000	1	0,162393162
17	2010-01-01 02:45:00.000	1	0,555555556
18	2010-01-01 02:55:00.000	1	0,222222222
19	2010-01-01 03:10:00.000	1	0,247863248
20	2010-01-01 03:35:00.000	1	0,478632479
21	2010-01-01 03:40:00.000	1	0,504273504
22	2010-01-01 03:50:00.000	1	0,230769231
23	2010-01-01 04:20:00.000	1	0,188034188
24	2010-01-01 04:35:00.000	1	0,179487179
25	2010-01-01 04:40:00.000	1	-0,008547009
26	2010-01-01 04:45:00.000	1	0,196581197
27	2010-01-01 04:55:00.000	1	-0,008547009
28	2010-01-01 05:55:00.000	1	0,239316239

VISUALIZATIONS



- POWER BI ENVIRONMENT, DATA LOADED FROM ANALYSIS SERVICES PROJECT

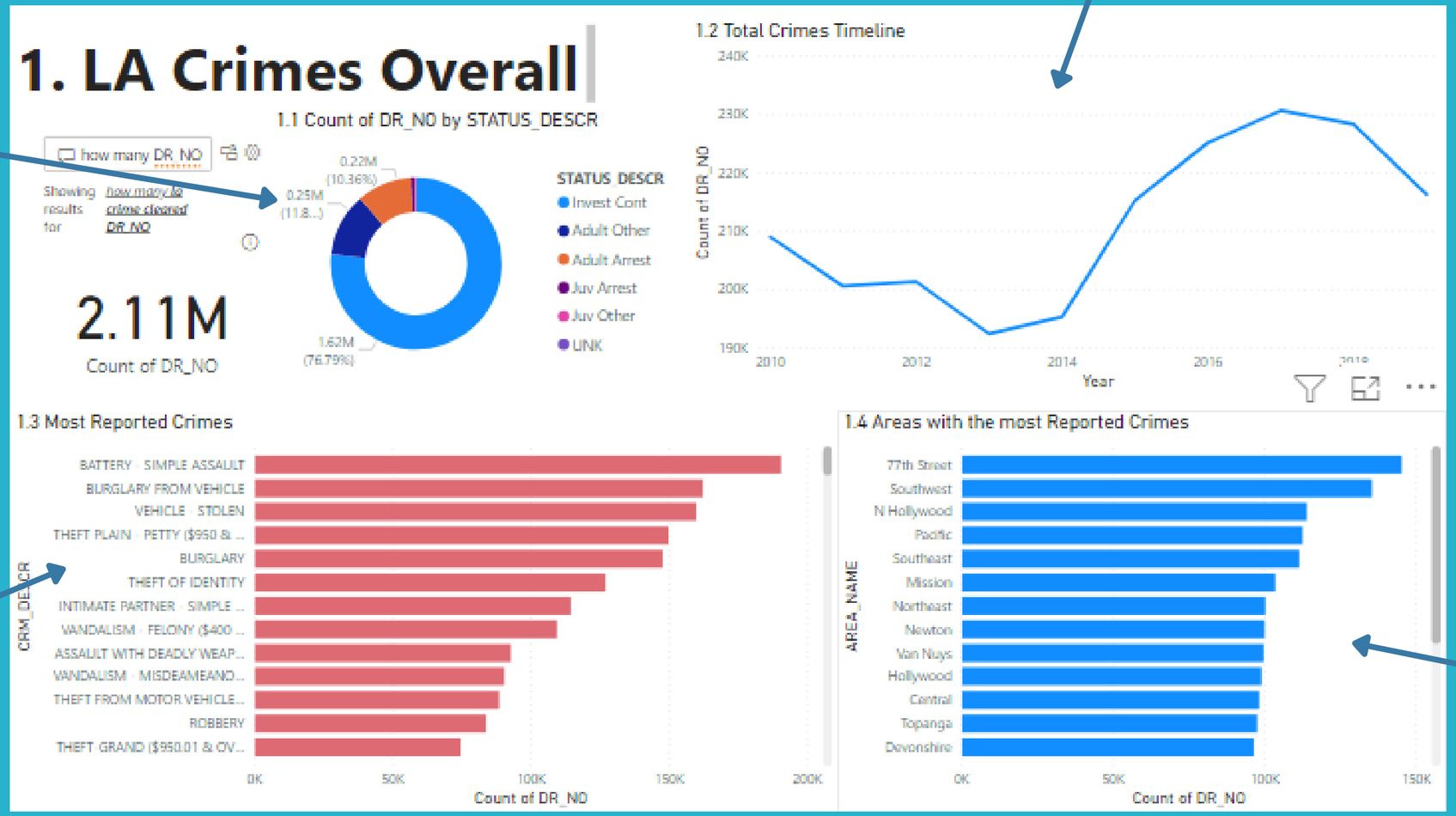
- JUPYTER NOTEBOOK ENVIRONMENT FOR MORE COMPLEX AND DEMANDING VISUALIZATIONS

Percentage of closed cases

Most Reported Crimes

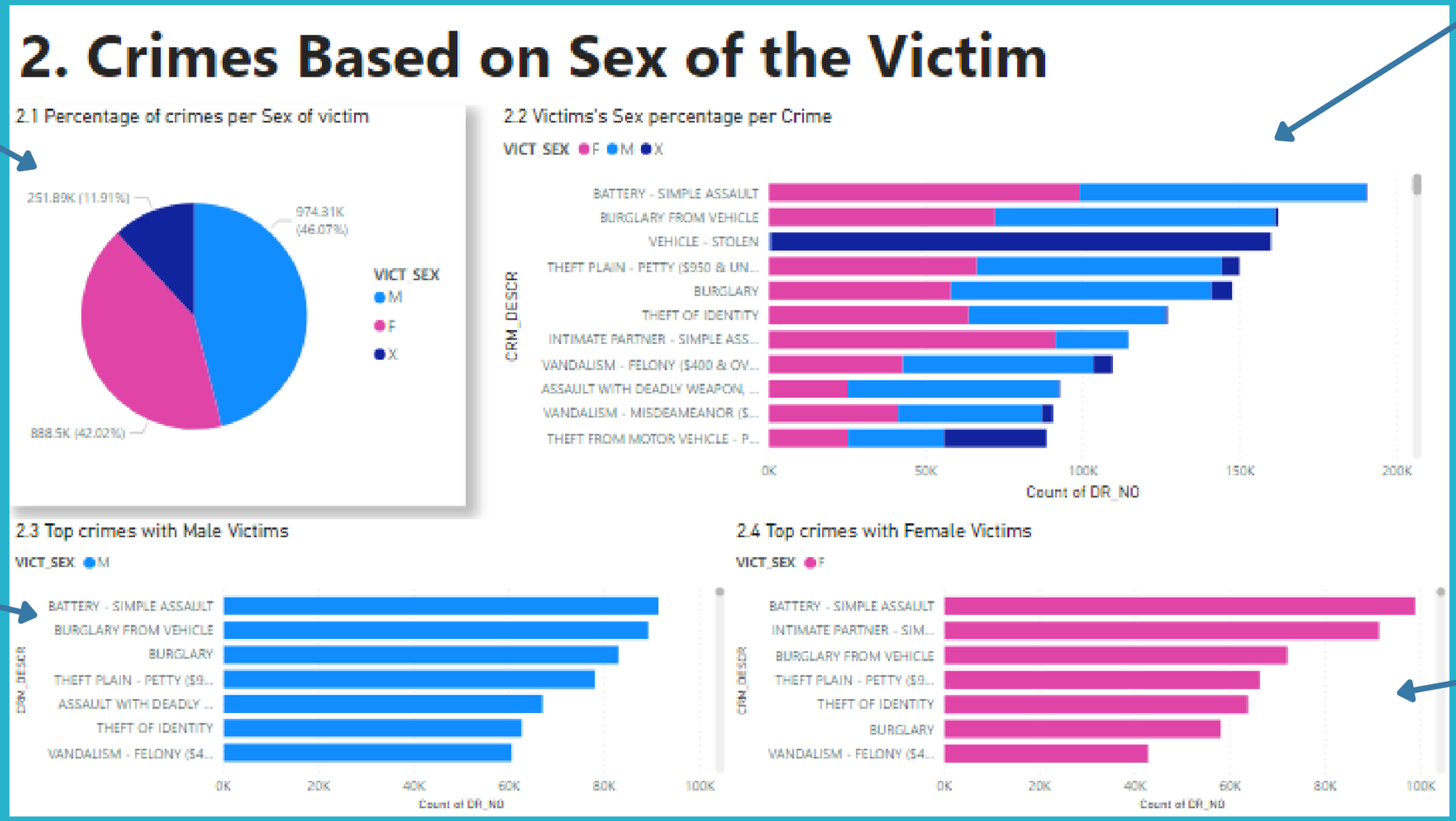
Total Crimes Timeline

Areas with most Reported Crimes



Percentage of Victim's Sex

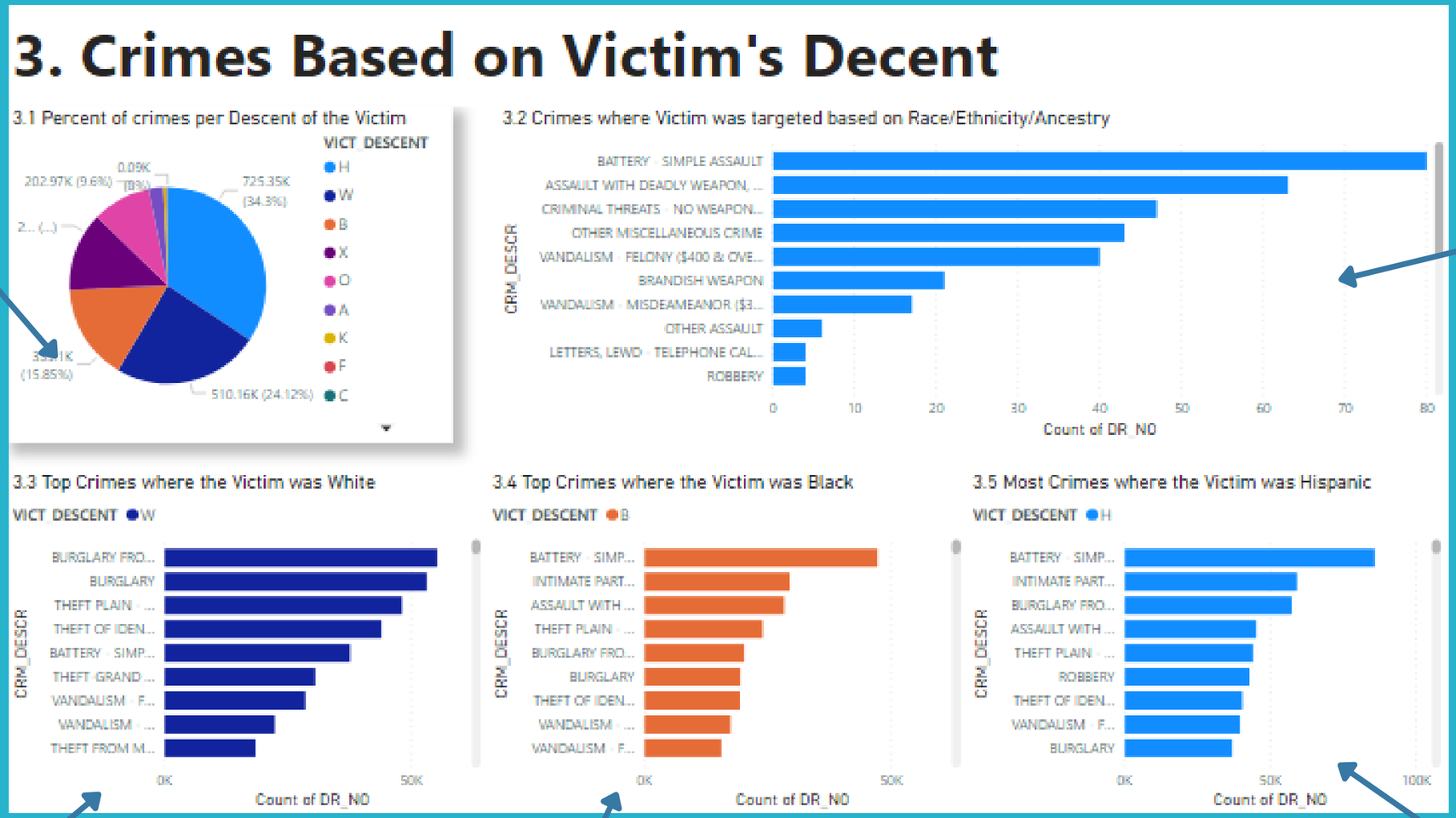
Percentage of Victim's Sex for most Common Crimes



Most usual Crimes with Male Victims

Most usual Crimes with Female Victims

Percentage of Victim's Descent



Crimes where Victim Descent was Targeted

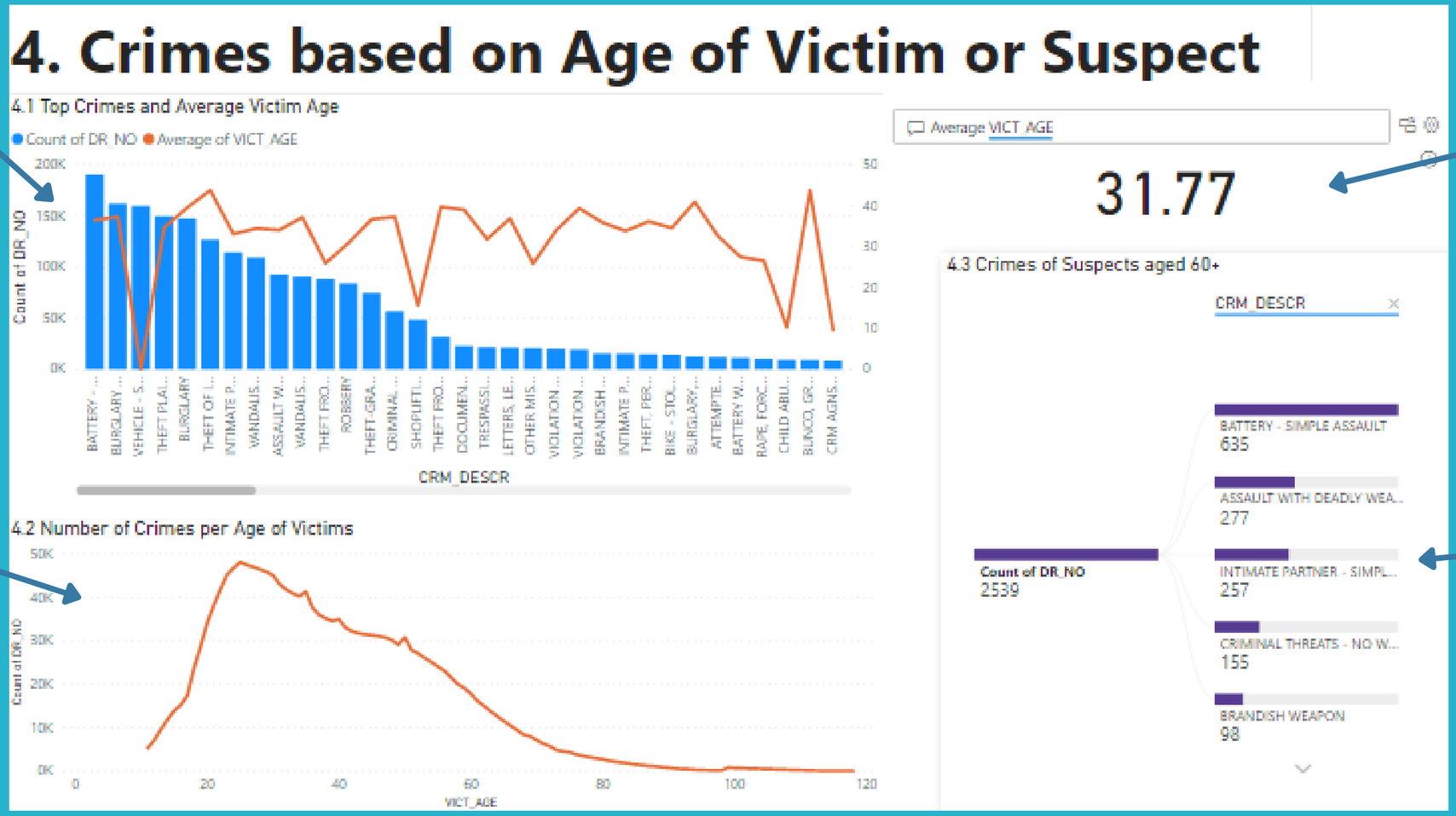
Most usual Crimes with Victims of White Descent

Most usual Crimes with Victims of Black Descent

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

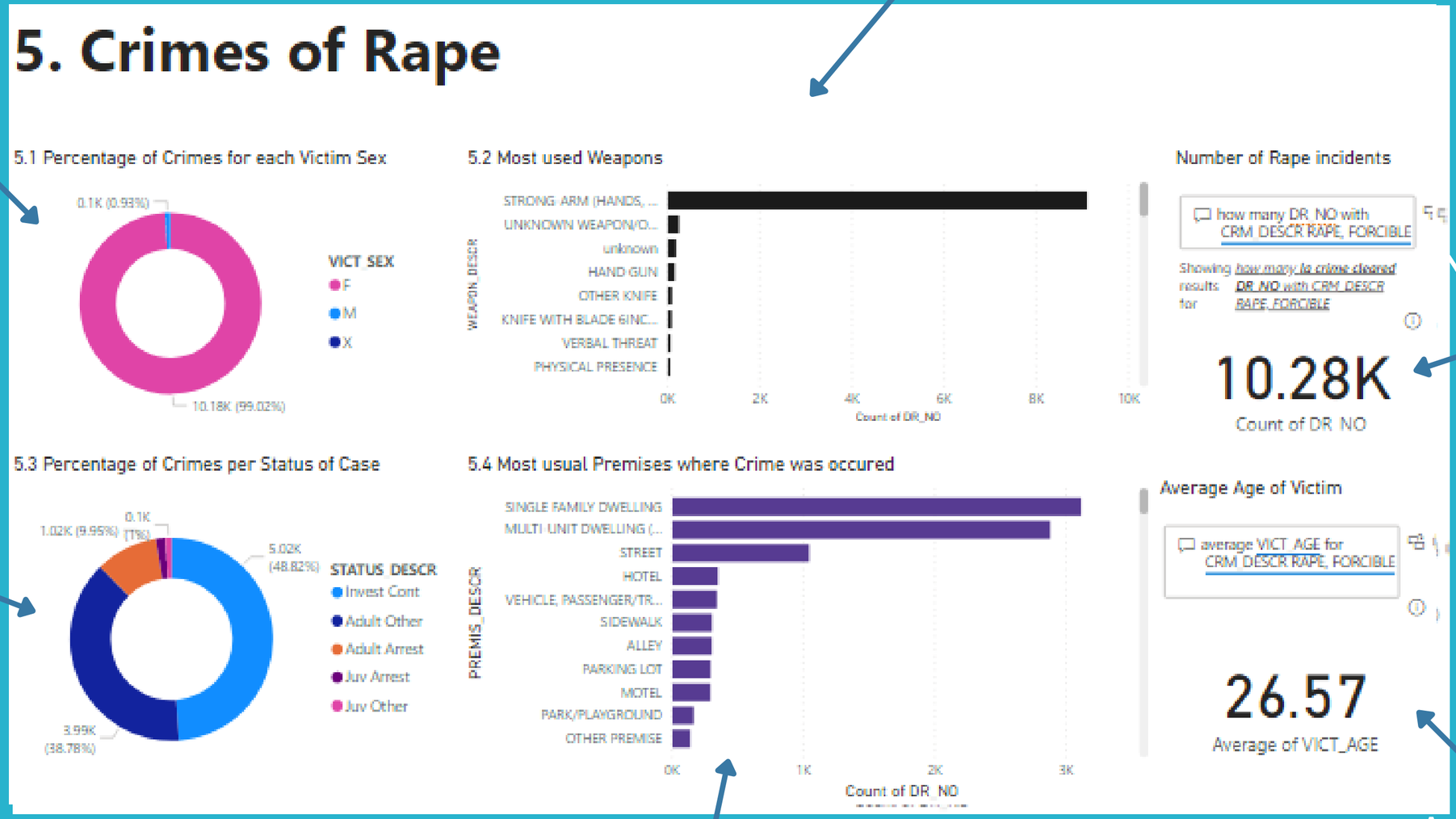
Most used Weapons

Percentage of closed/open Cases

Number of Rape Cases

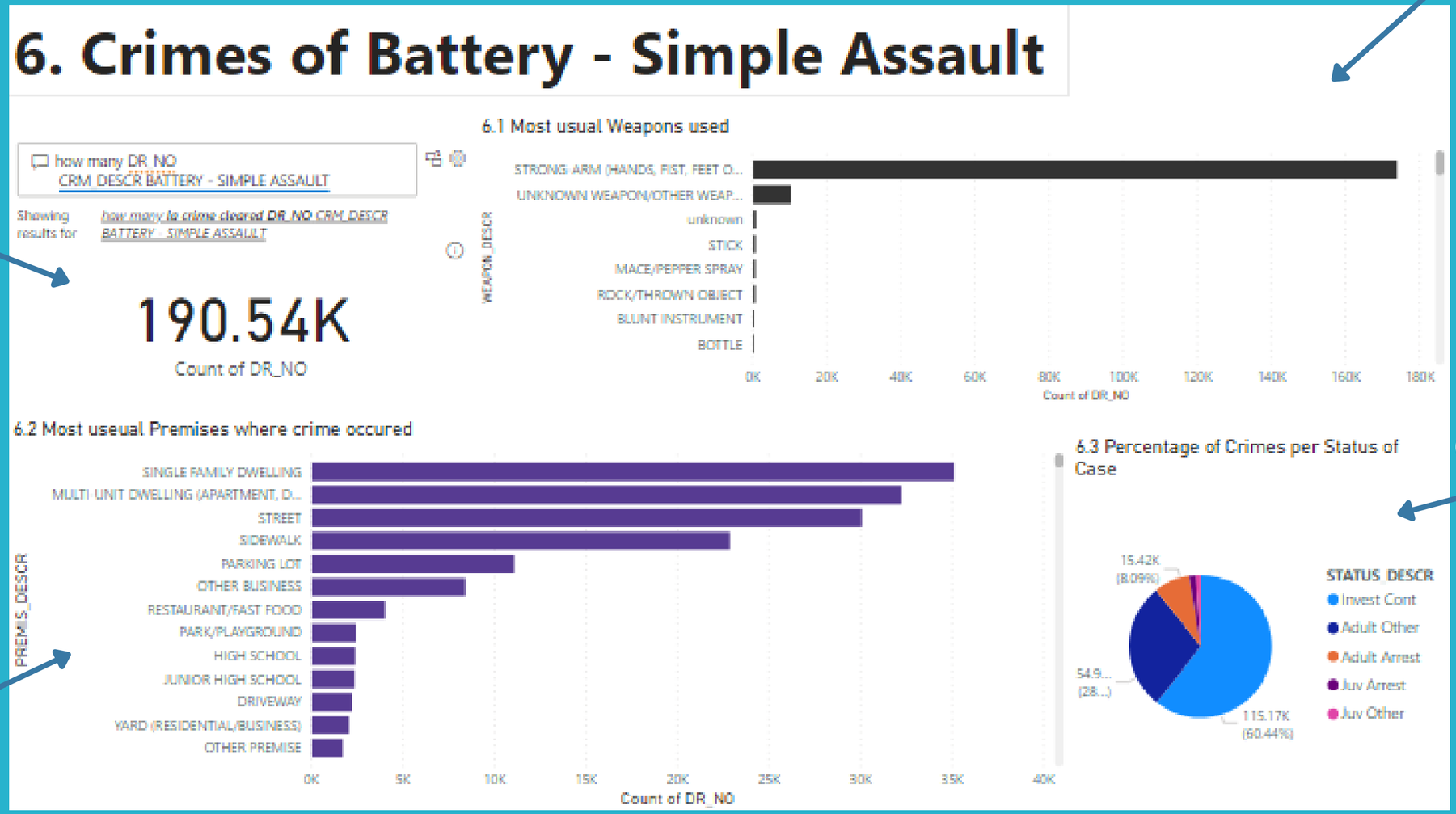
Most usual Premises

Average age of Victim



Number of Battery - Simple Assault Incidents

Most used weapons



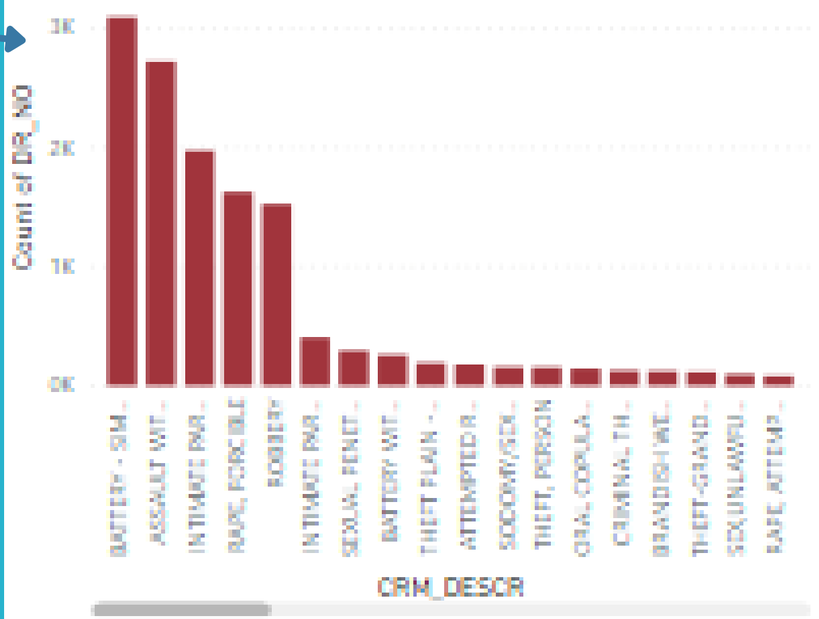
Percentage of closed/open Cases

Most usual Premises of Battery

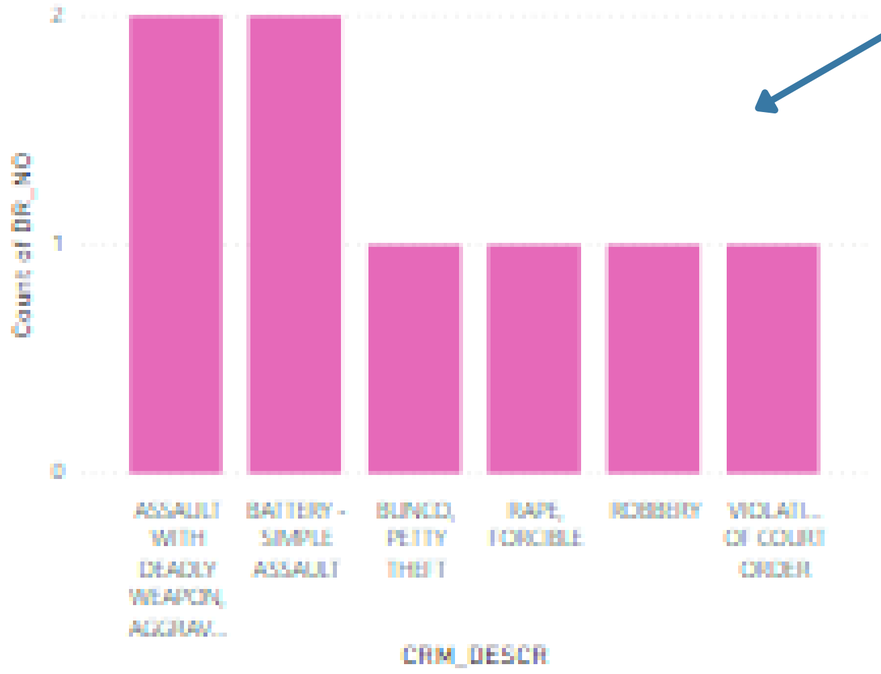
Crimes where Suspect was Undocumented Alien

8. Other

8.1 Suspect was intoxicated or drunk



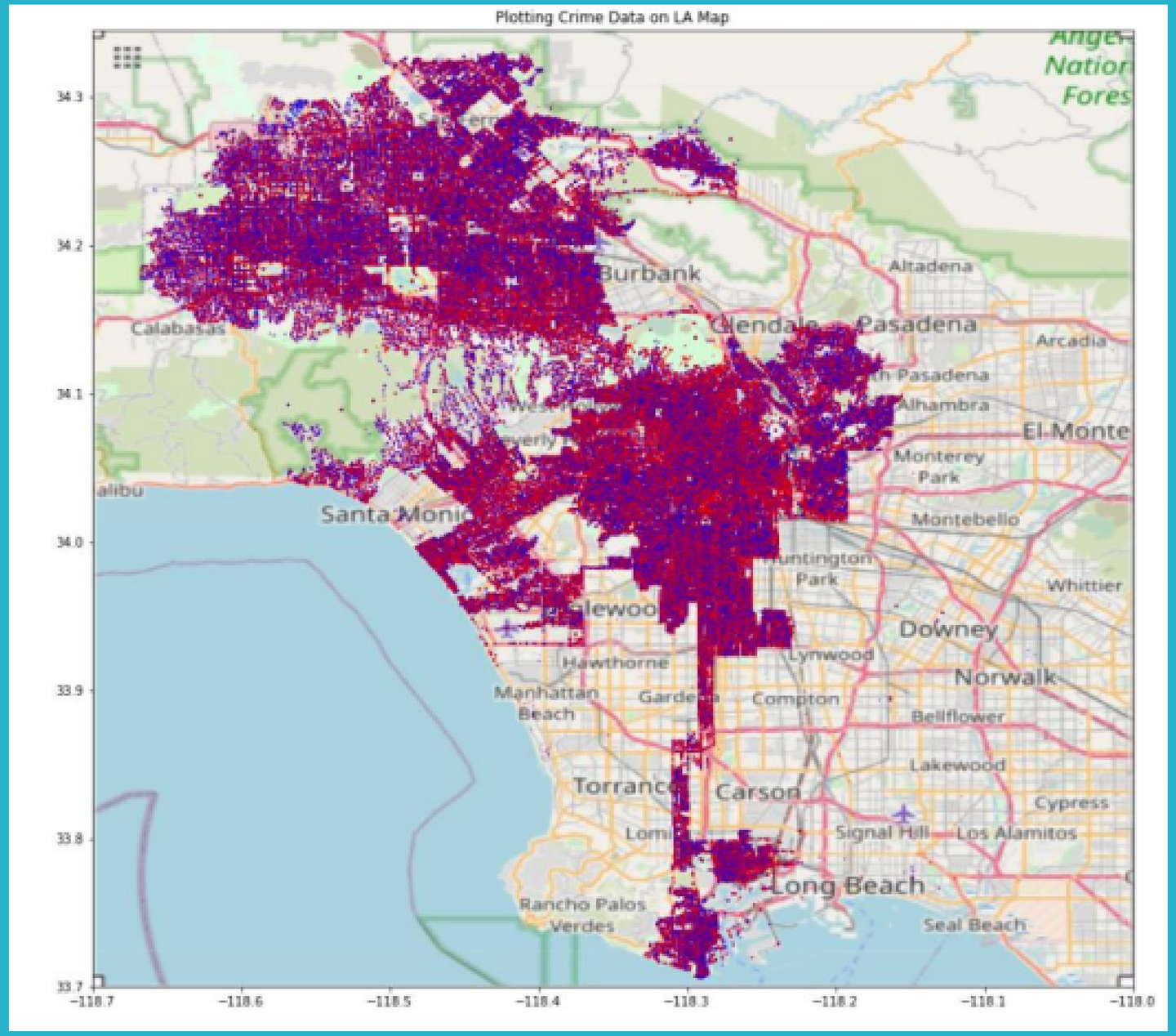
8.2 Incidents where the suspect was dressed in a costume(Spiderman, Darth Vader etc.)



Crimes where Suspect was Dressed in a Character Costume

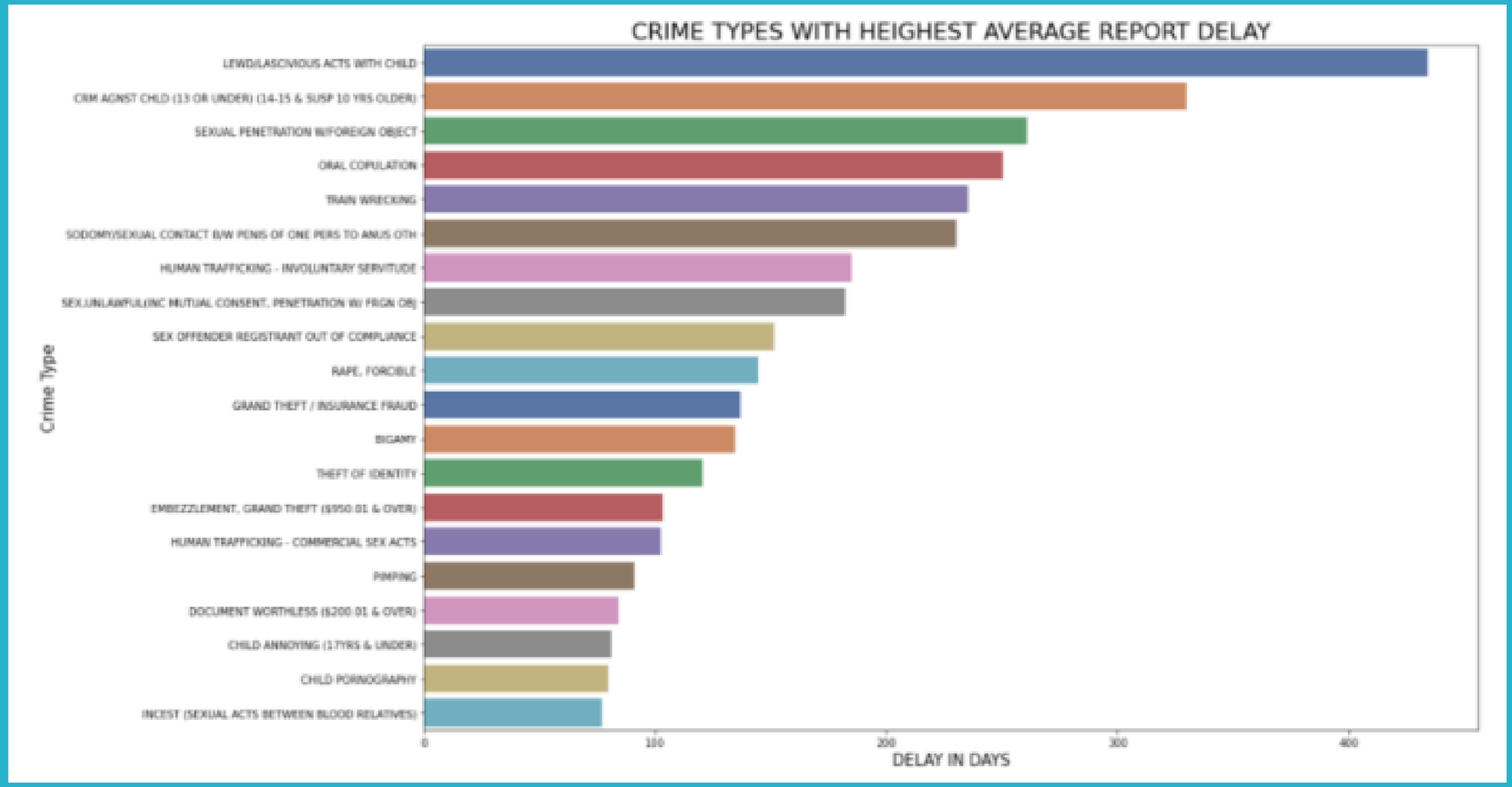
CRIME INCIDENTS ON LA MAP

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



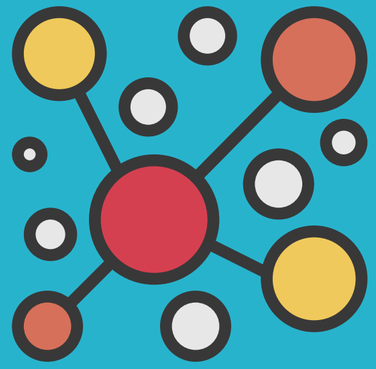
CRIME INCIDENTS WITH HIGHEST REPORT DELAY

- 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

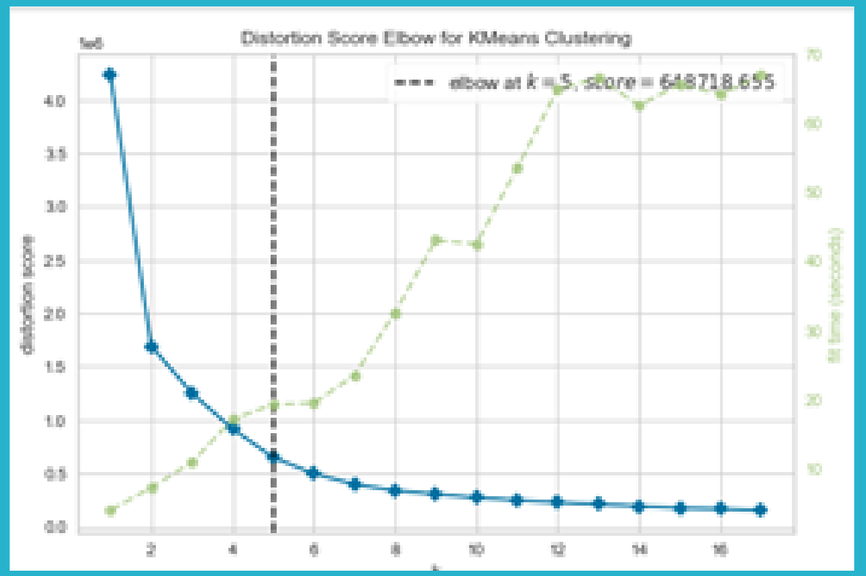
PREDICTION

- CRIME TYPE PREDICTION FROM REST OF INFORMATION

→ *JUPYTER NOTEBOOK* & *PYTHON*

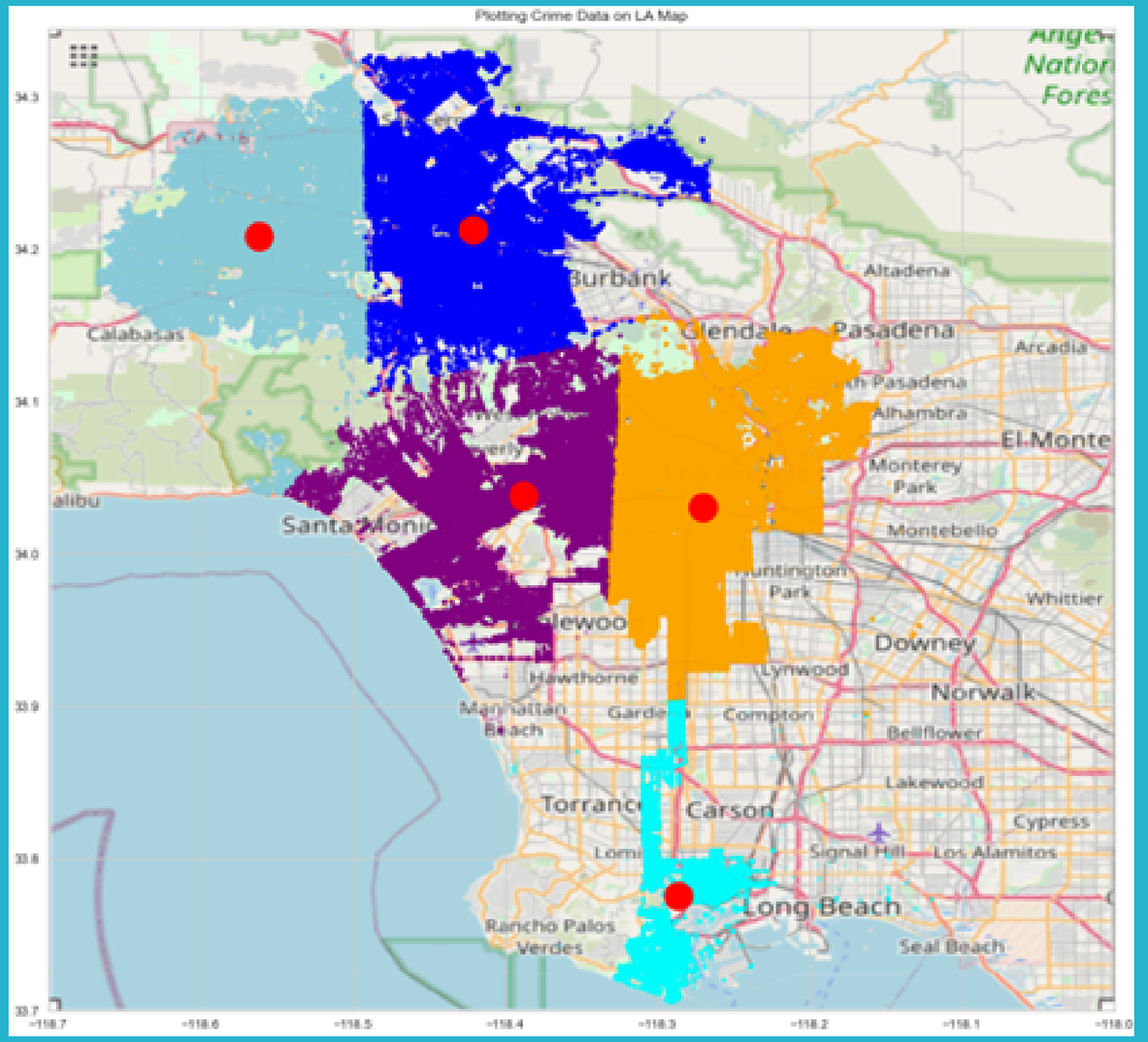
GEOGRAPHIC COORDINATE CLUSTERING

- 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

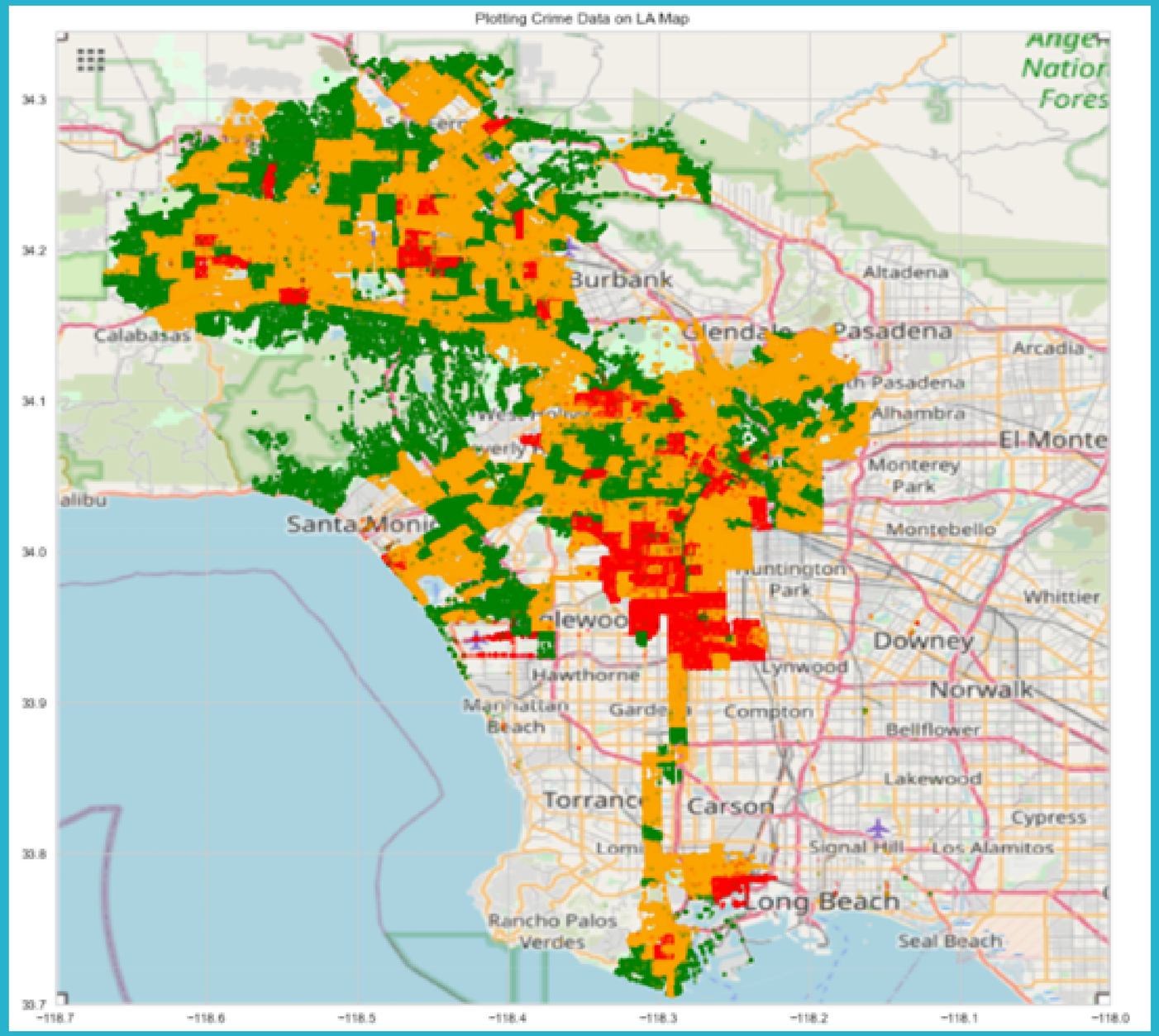
DISTRICT CLUSTERING RESULTS

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


THANK YOU