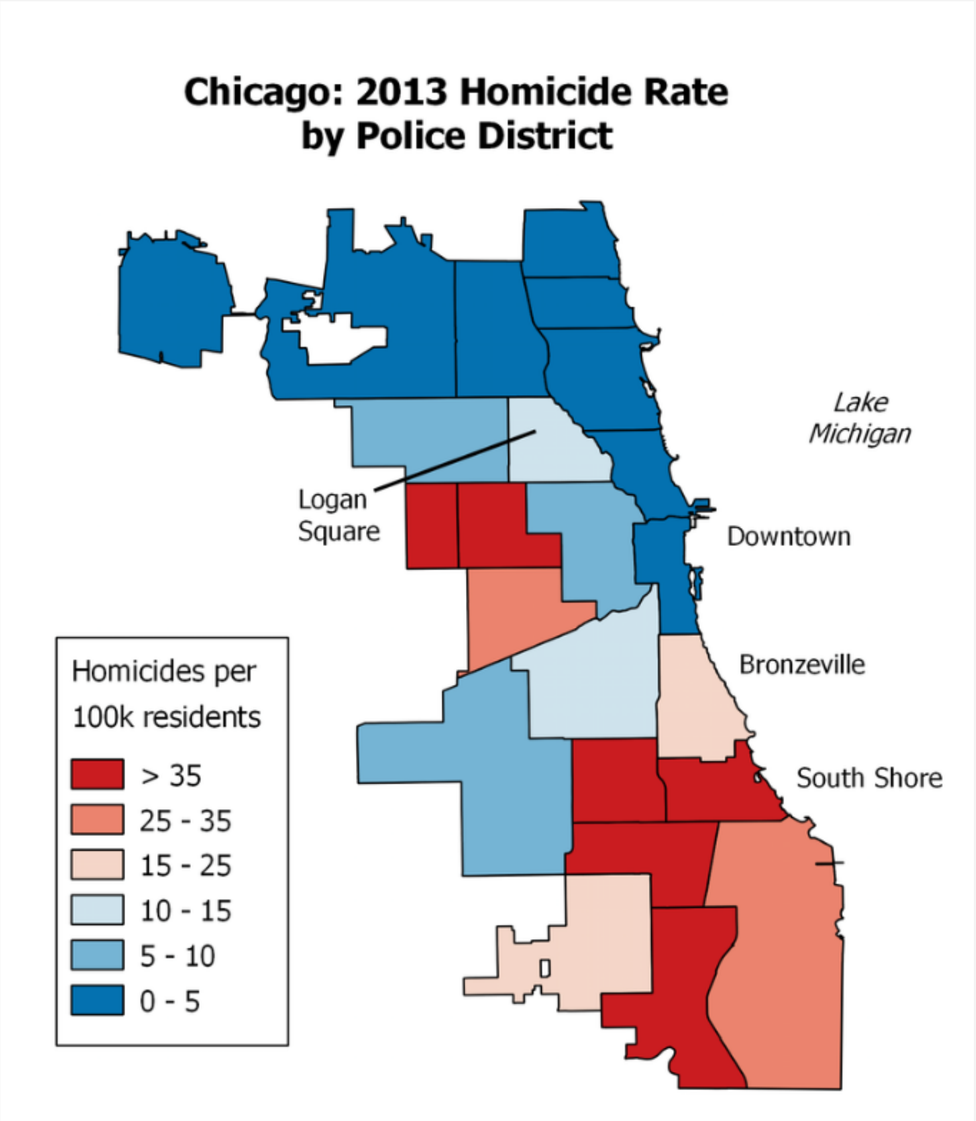
BUSINESS INTELIGENCE
AND BIG DATA
ANALYTICS

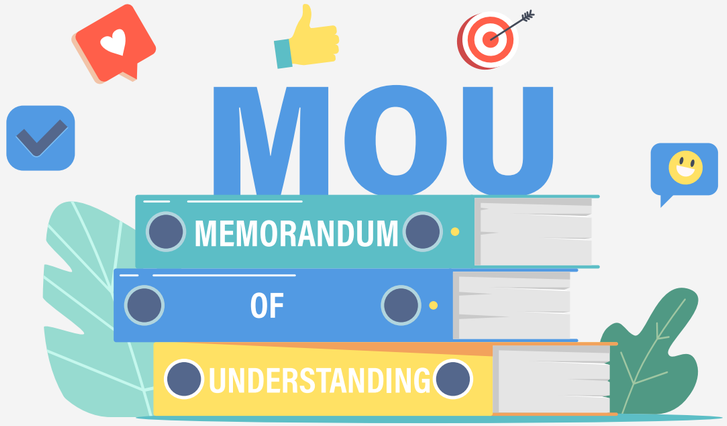
CRIMES IN CHICAGO

Κακωνάς Νικόλαος – 8190050
Μανιουδάκη Γεωργία – 8190097



Contents of the Report

Describe Dataset



Context

This dataset reflects reported incidents of crime that occurred in the City of Chicago from 2001 to 2021

Contents

7.2M rows
ID
Case Number
Date
Block
IUCR
Primary Type
Description
Location Description
Arrest
Domestic
Beat
District
Ward
Community Area
FBI Code
X Coordinate
Y Coordinate
Year
Updated On
Latitude
Longitude
Location

Source

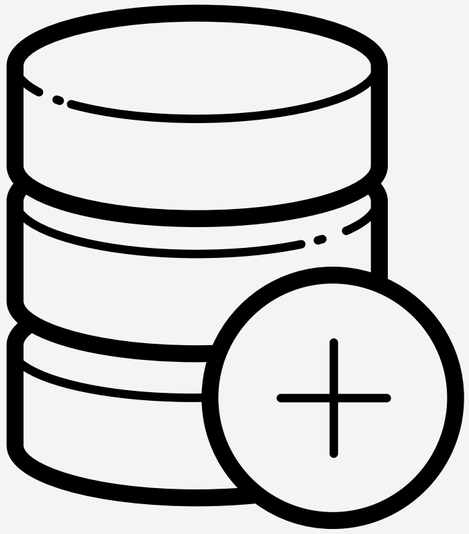
https://www.kaggle.com/datasets/mingyuouyang/chicago-crime-2001-to-2022

Data Cleaning and Processing

- Delete columns
- Longitude - Latitude
- Similar Values
- Delete of NON-CRIMIINAL cases
- Date & Time
- Remove commas (,)
- Boolean to Bit

Data Warehouse SQL Server

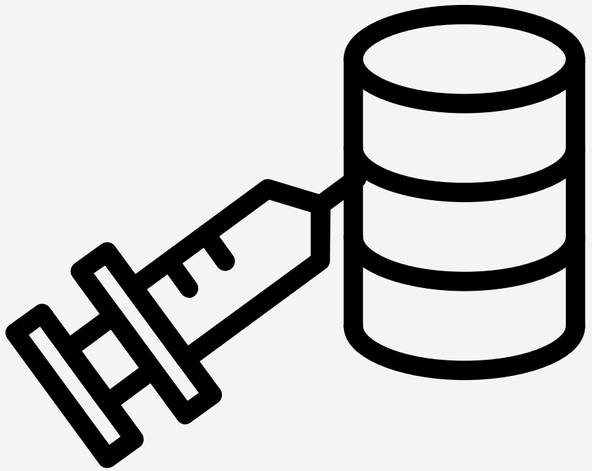
Create Table



Column Name	Data Type	Allow Nulls
ID	bigint	☑
[Case Number]	varchar(10)	☑
Date	date	☑
Time	time(7)	☑
Block	varchar(50)	☑
IUCR	varchar(4)	☑
[Primary Type]	varchar(50)	☑
Description	varchar(100)	☑
[Location Description]	varchar(100)	☑
Arrest	bit	☑
Domestic	bit	☑
Beat	bigint	☑
District	bigint	☑
Ward	bigint	☑
[Community Area]	bigint	☑
[FBI Code]	varchar(10)	☑
[X Coordinate]	bigint	☑
[Y Coordinate]	bigint	☑
Latitude	float	☑
Longitude	float	☑
Location	varchar(100)	☑
		☐

Data Warehouse SQL Server

Bulk Insert



```
BULK INSERT dbo.crimes
FROM 'C:\Users\***\Chicago-Crimes\Chicago_Crimes.csv'
WITH (FIRSTROW = 2,
      FORMAT = 'CSV',
      MAXERRORS = 0,
      FIELDQUOTE = '"',
      FIELDTERMINATOR = ',',
      ROWTERMINATOR = '\n'
     );
```

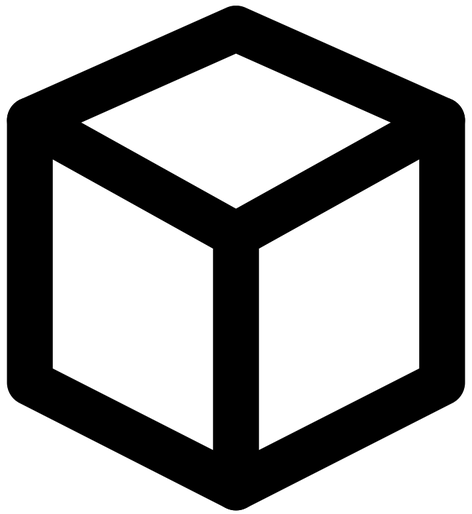
Data Warehouse SQL Server

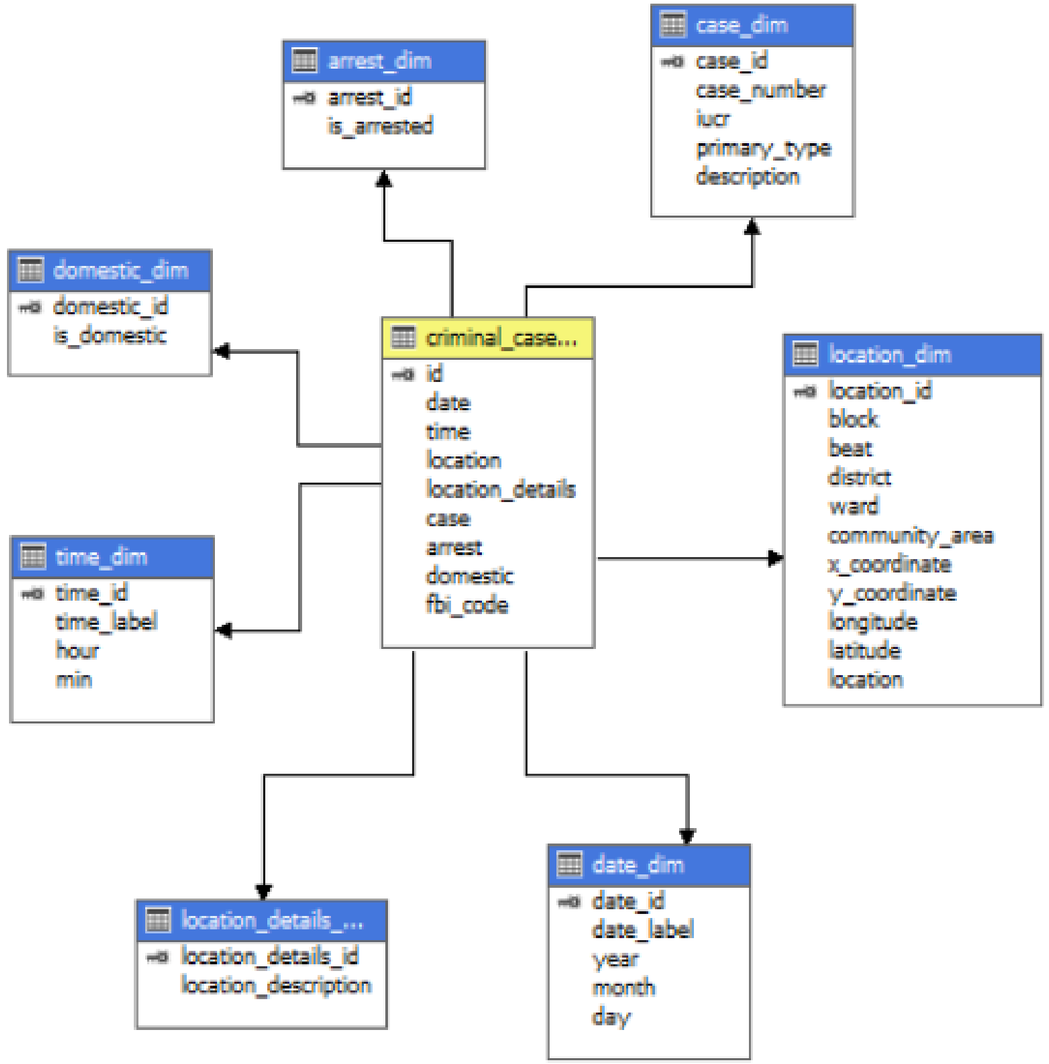
Dimensions

- Case Dimension
- Location Dimension
- Location Details Dimension
- Date Dimension
- Time Dimension
- Arrest Dimension
- Domestic Dimension

Visual Studio Datacube

Cube Schema





Visual Studio Datacube

Metrics

27% of offenders have been arrested

13% of total crimes involve relatives

Most crimes happen in July (9.4%)
The peak crime time is at 12:00 (5.7%)

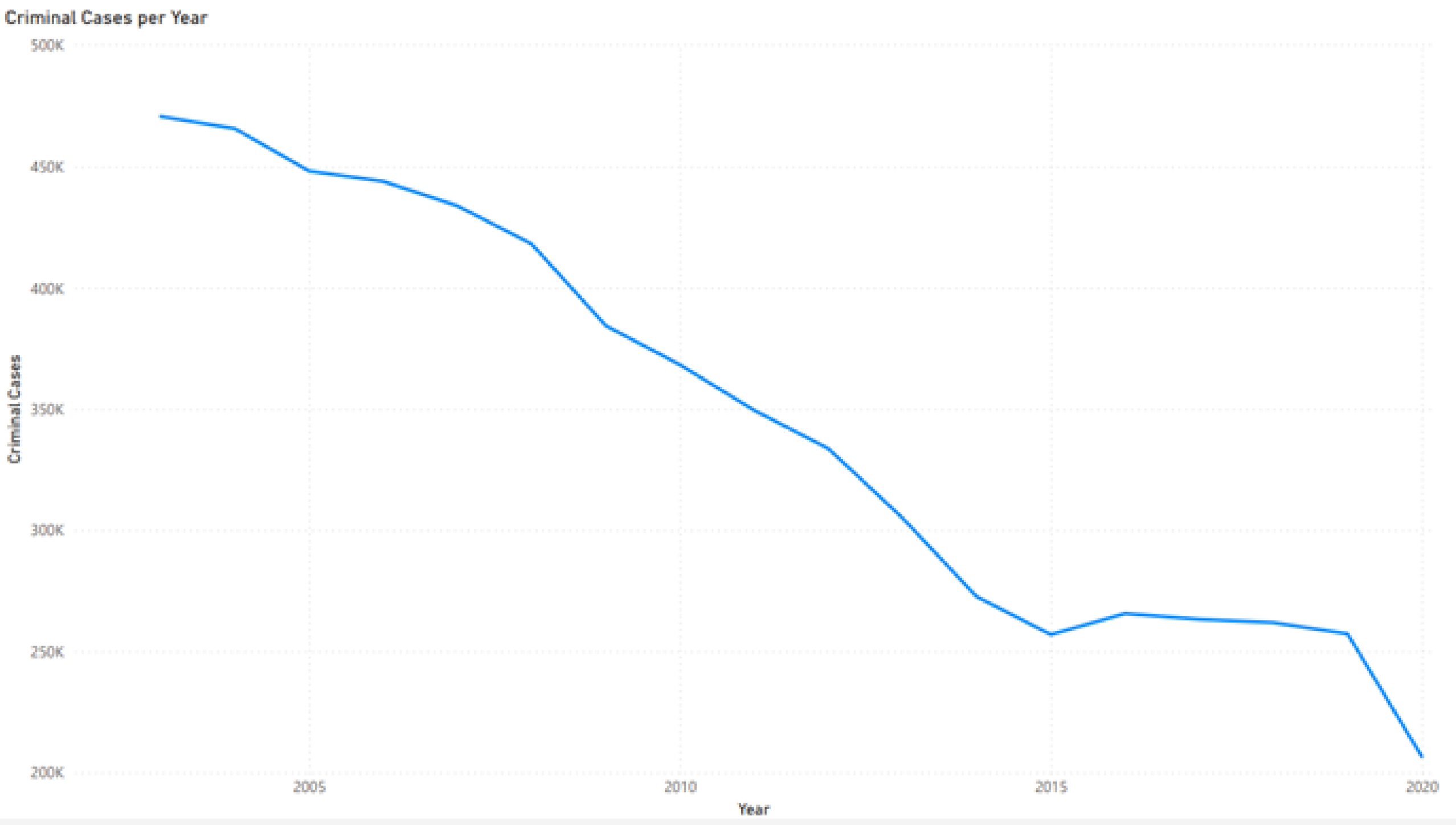
Less crimes happen in February (6.6%)
Time with the least crimes is at 05:00 (1.3%)

21% of crimes involve theft

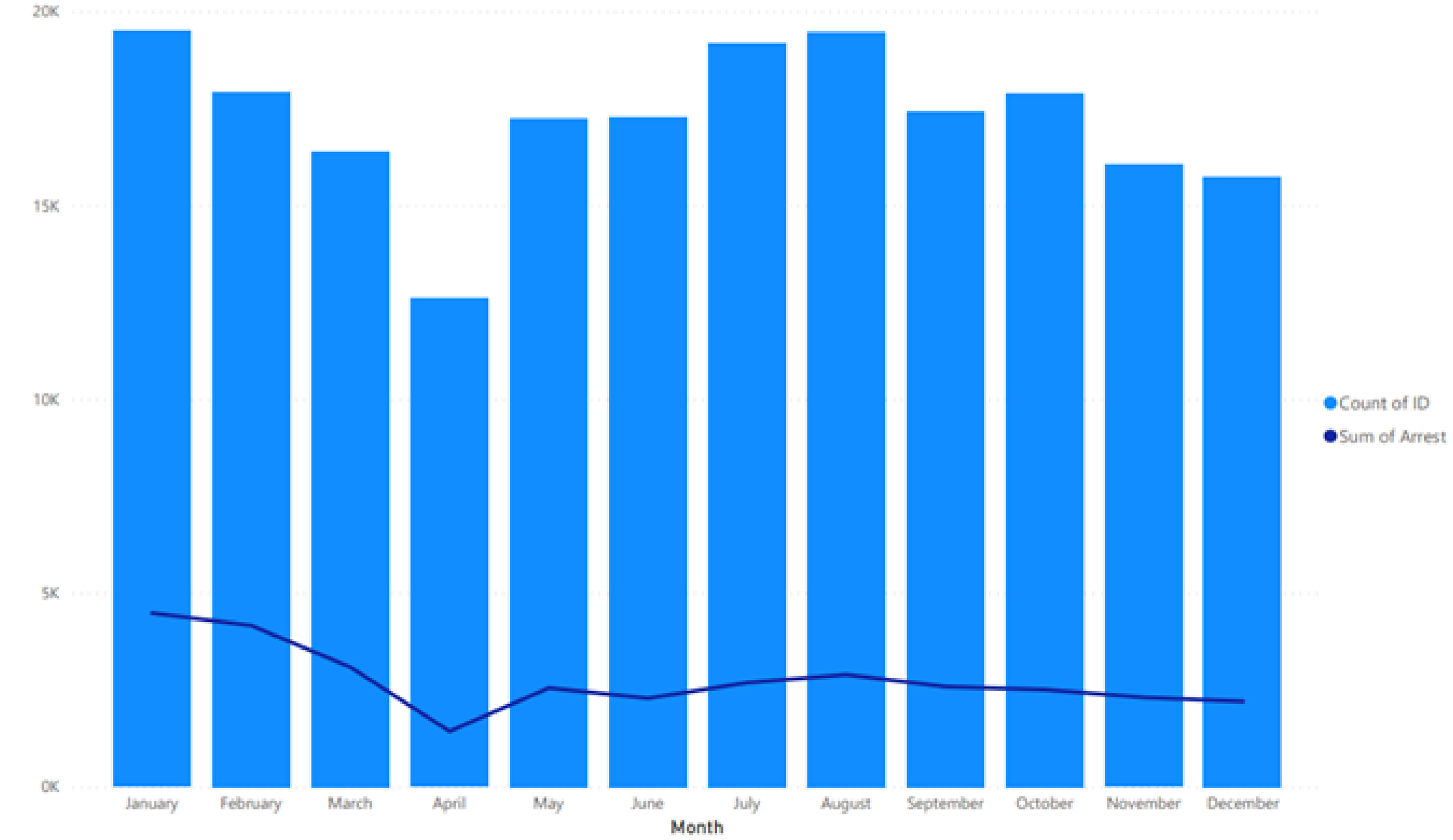
the average crimes per year are 313,252

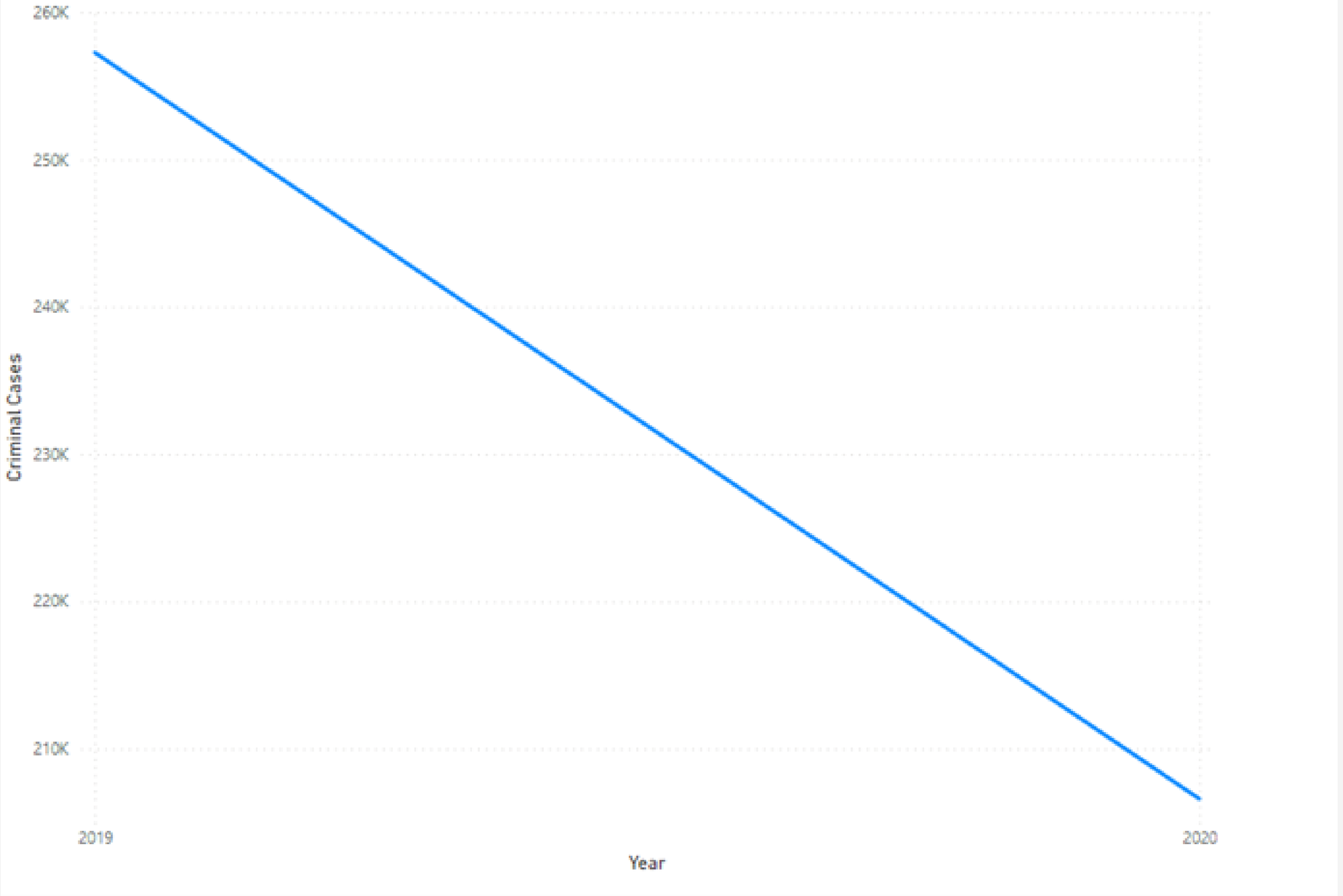
Visualisation

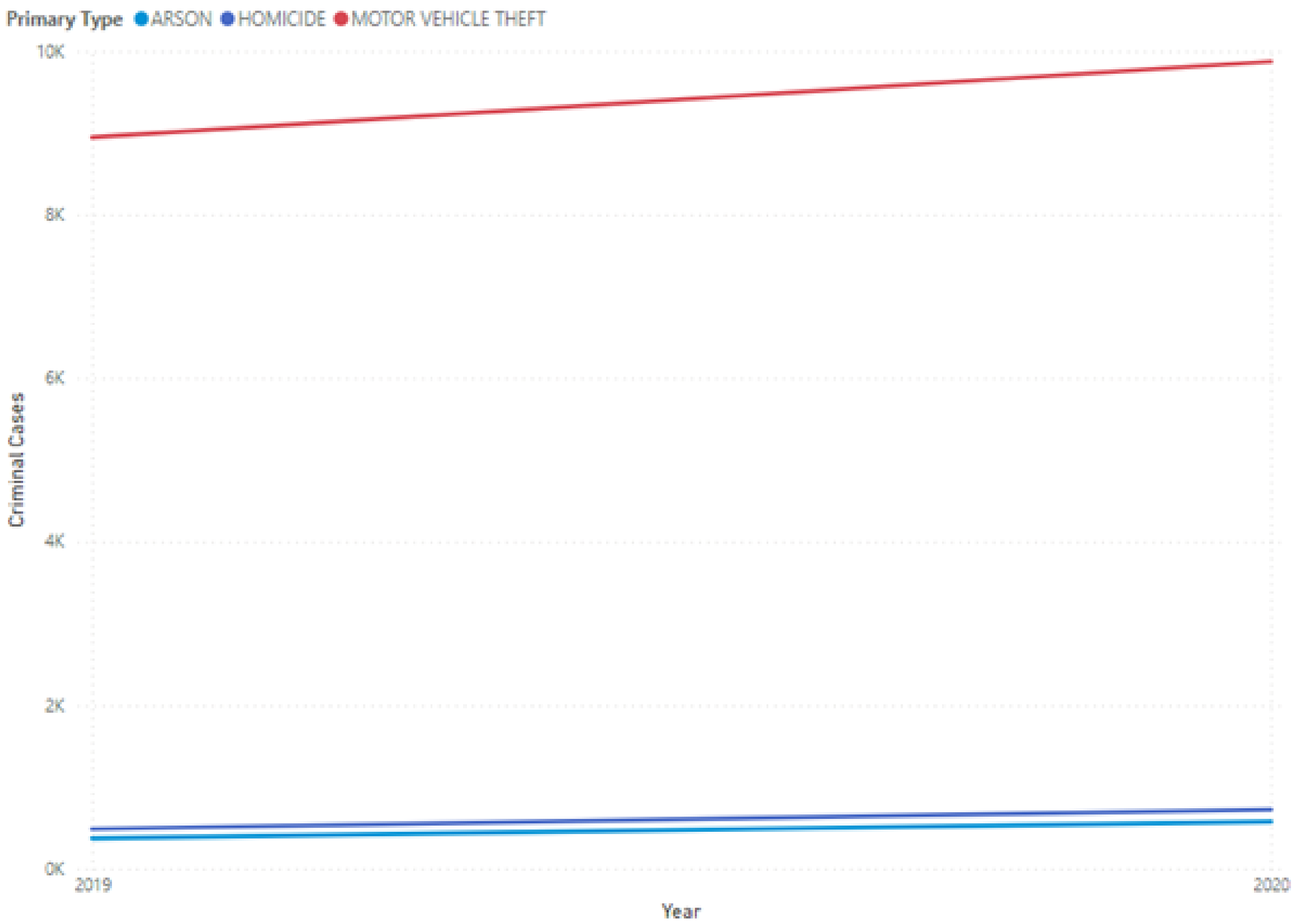
scenarios

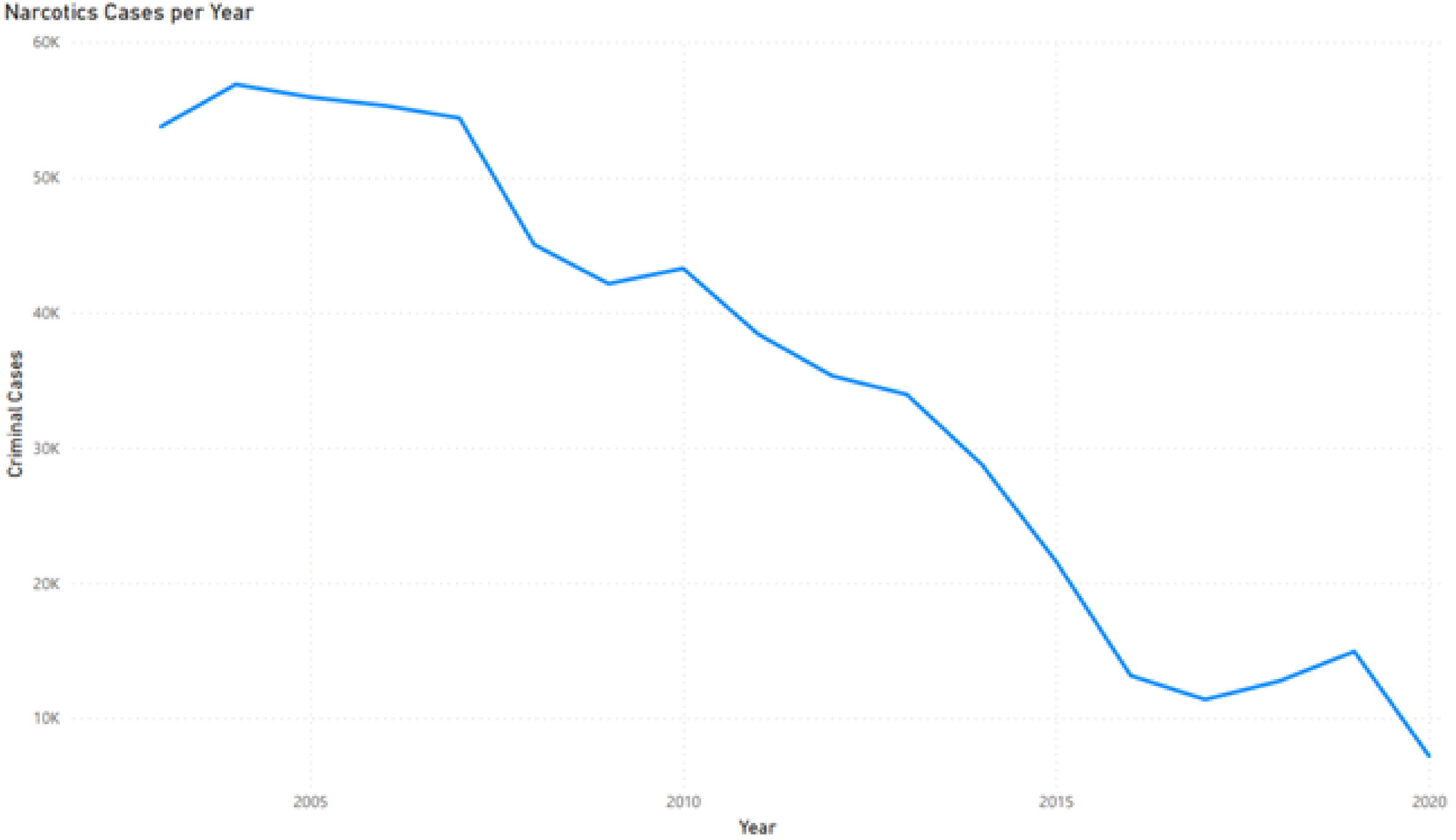

Criminal Cases per Year
500K
450K
400K
350K
300K
250K
200K
Criminal Cases
2005
2010
2015
2020
Year

Criminal Cases and Arrests per Month (2020)

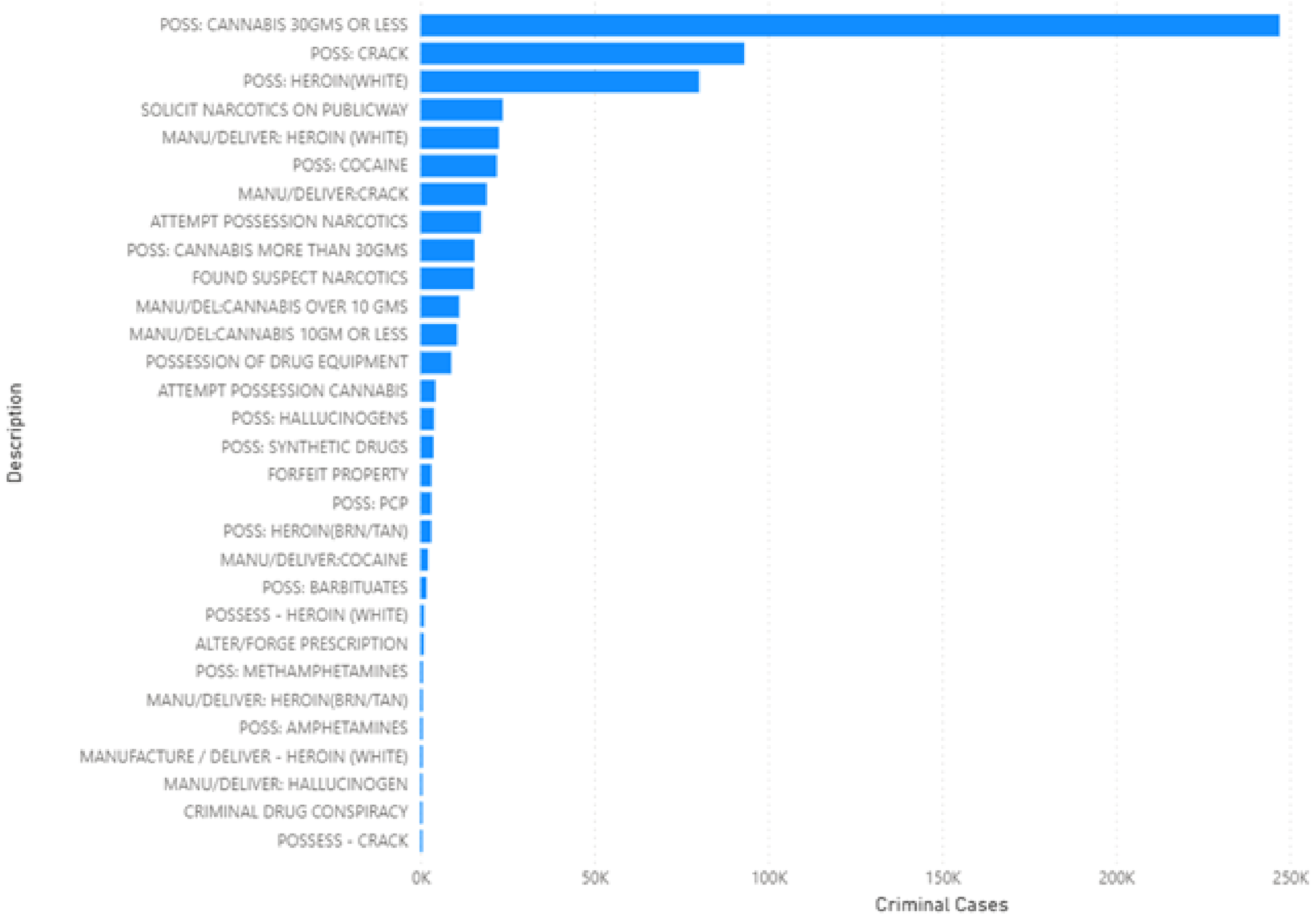


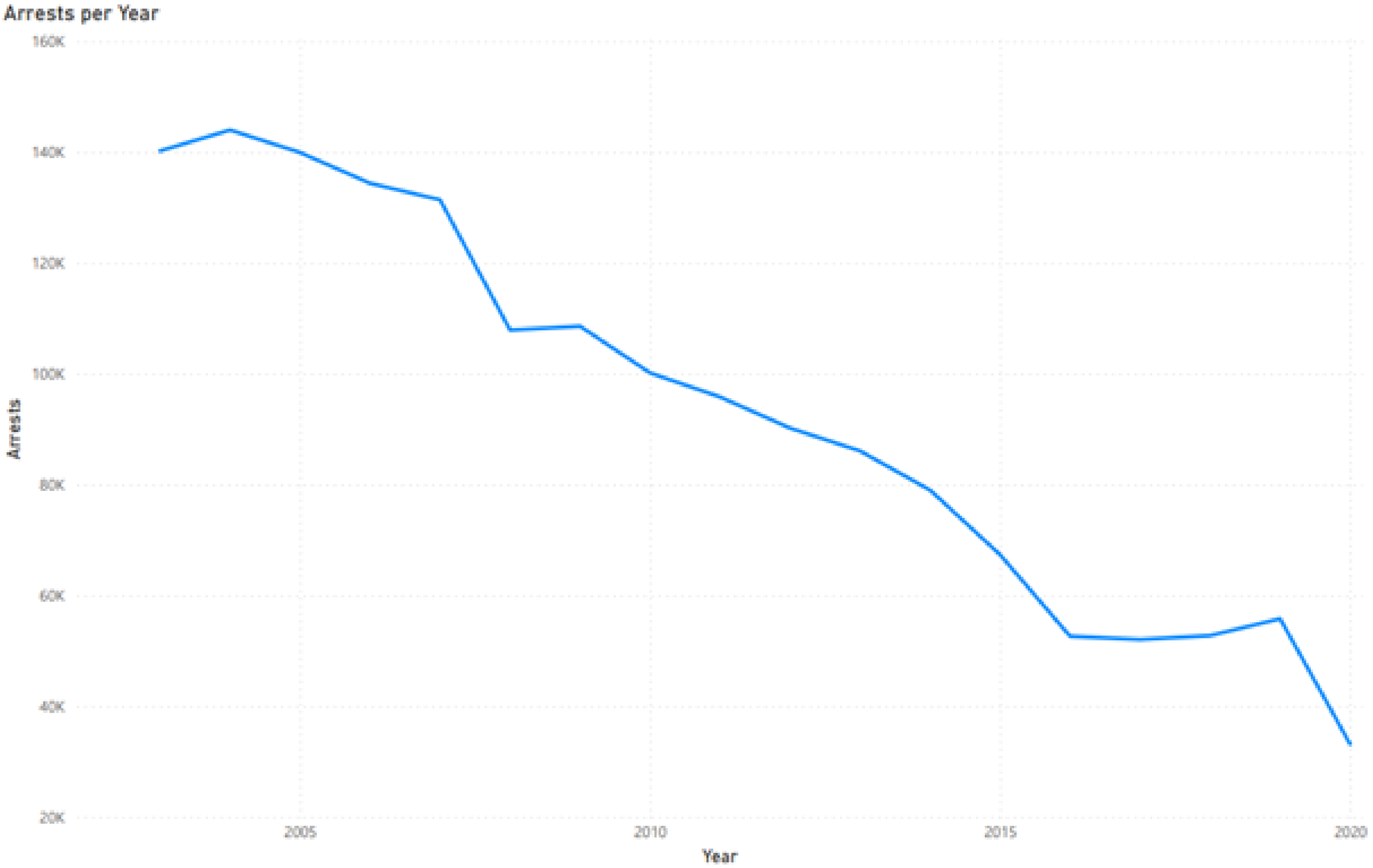

260K
250K
240K
230K
220K
210K
Criminal Cases
2019
2020
Year


Primary Type
ARSON
HOMICIDE
MOTOR VEHICLE THEFT
10K
8K
6K
4K
2K
0K
Criminal Cases
2019
2020
Year

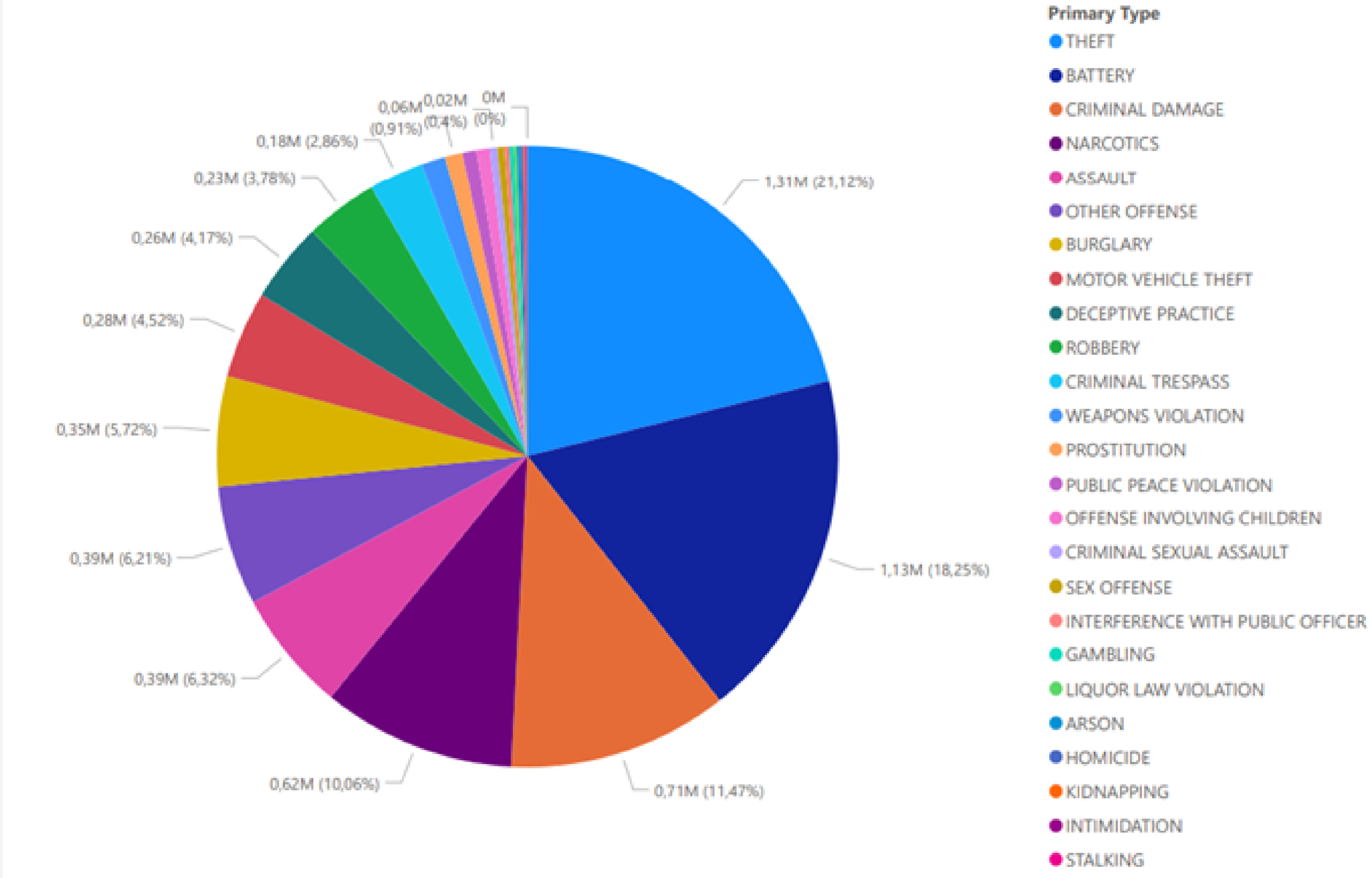

Narcotics Cases per Year
60K
50K
40K
30K
20K
10K
Criminal Cases
2005
2010
2015
2020
Year

Narcotics Cases by Description

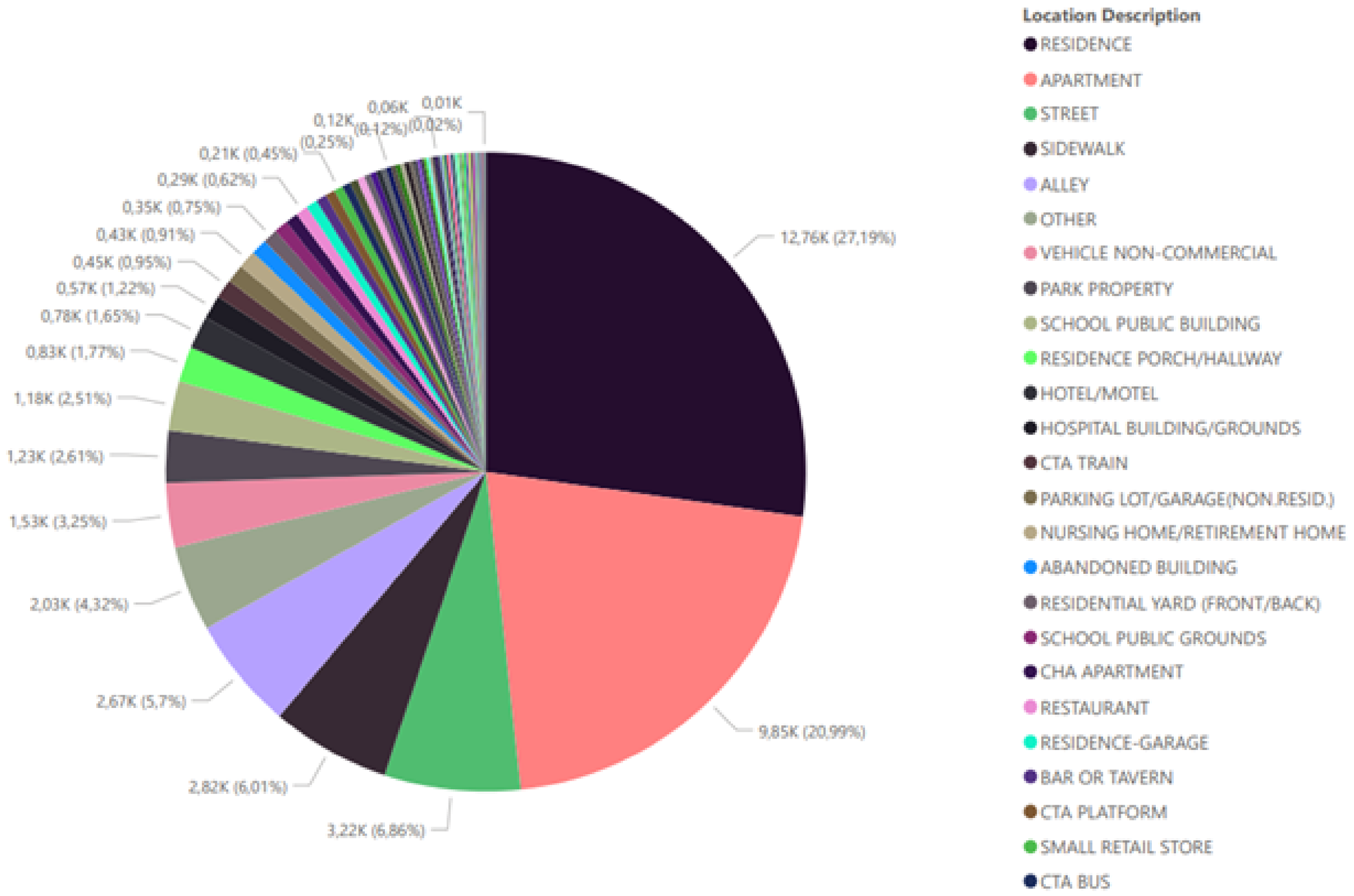



Arrests per Year
160K
140K
120K
100K
80K
60K
40K
20K
Arrests
2005
2010
2015
2020
Year

Number of Cases by Primary Type

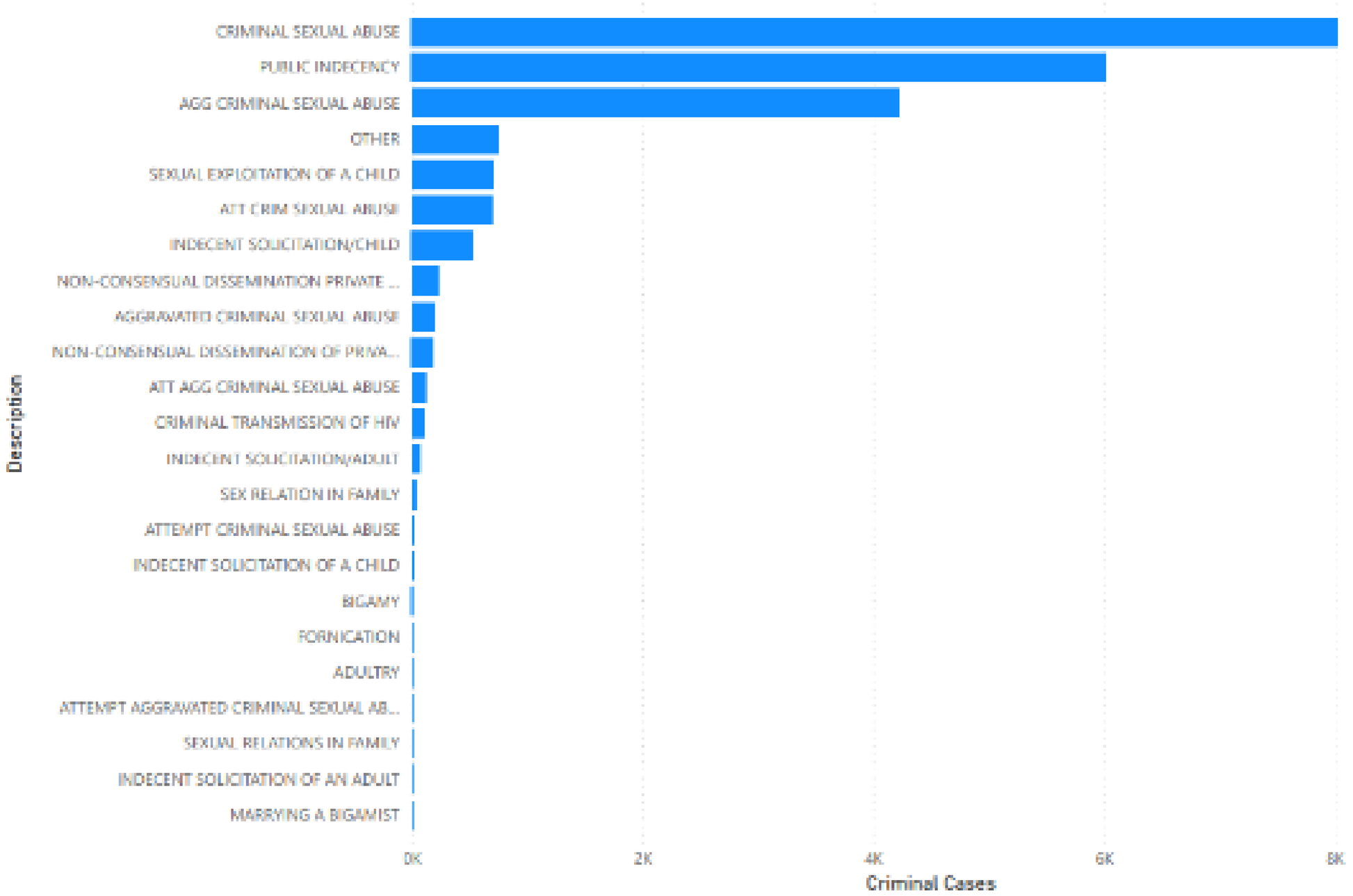


Sexcrime Cases by Location Description

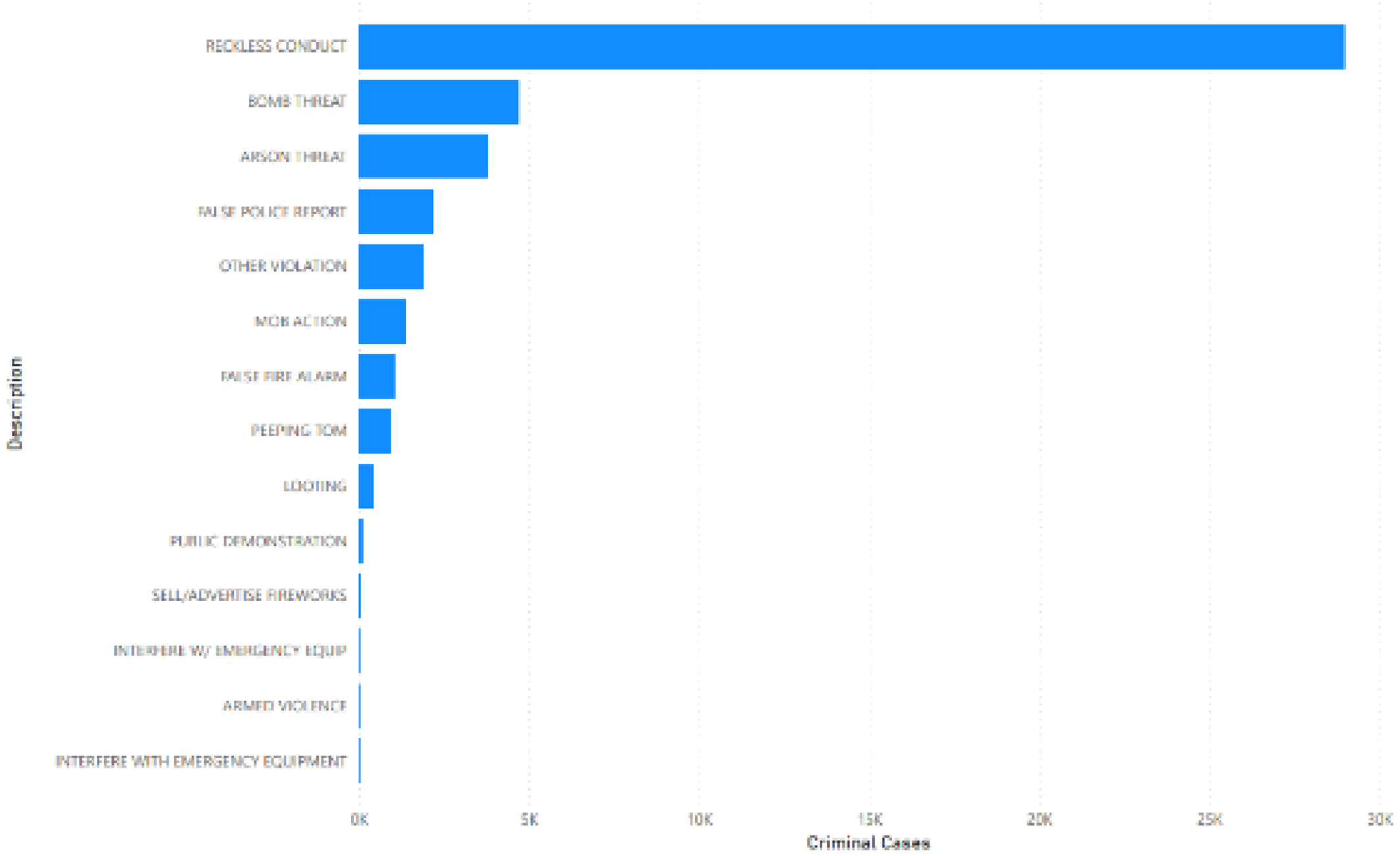



Sexcrime Cases per Year
3.500
3.000
2.500
2.000
Criminal Cases
2005
2010
2015
2020
Year

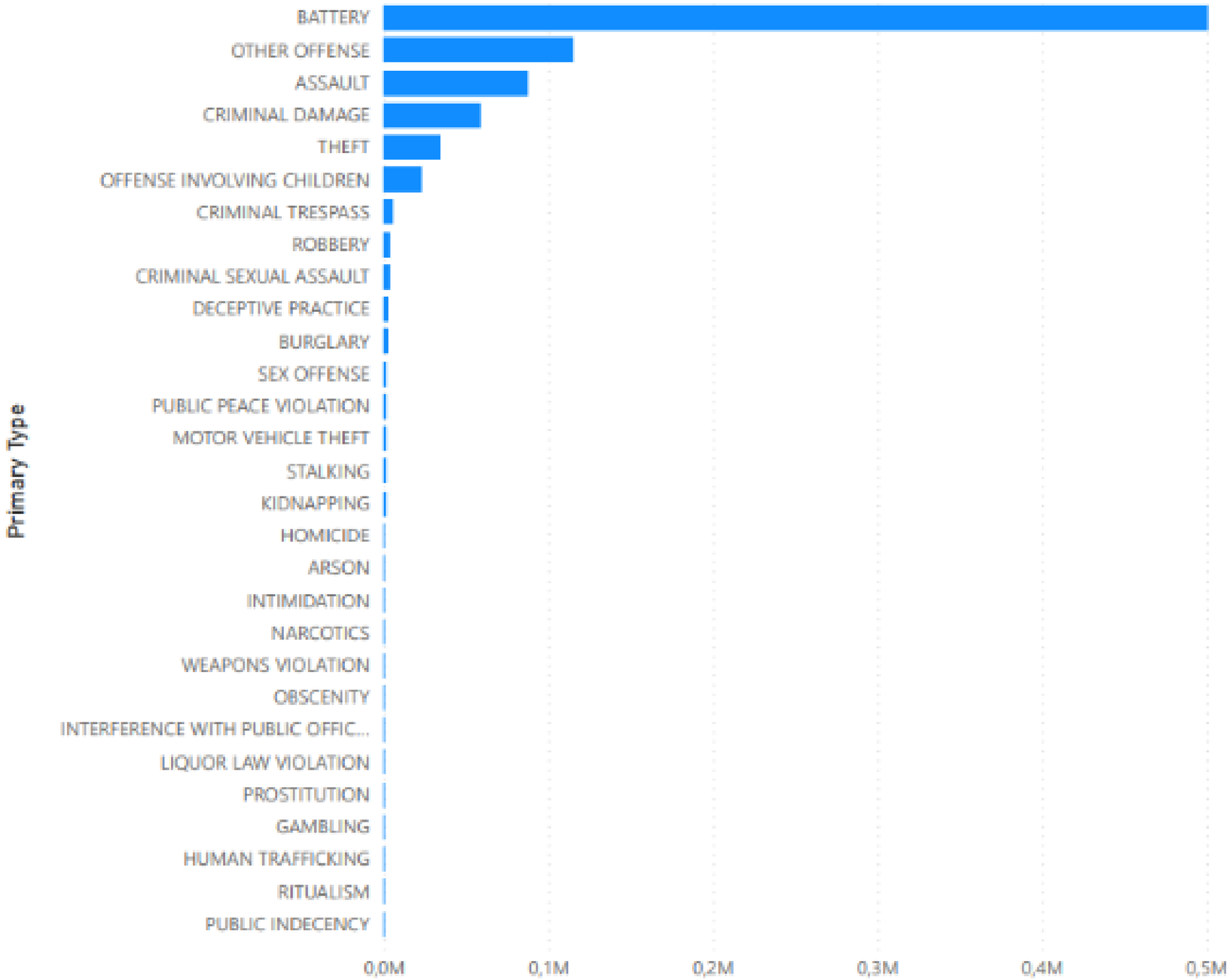
Sex Offense Cases by Description



Public Peace Violation Cases by Description



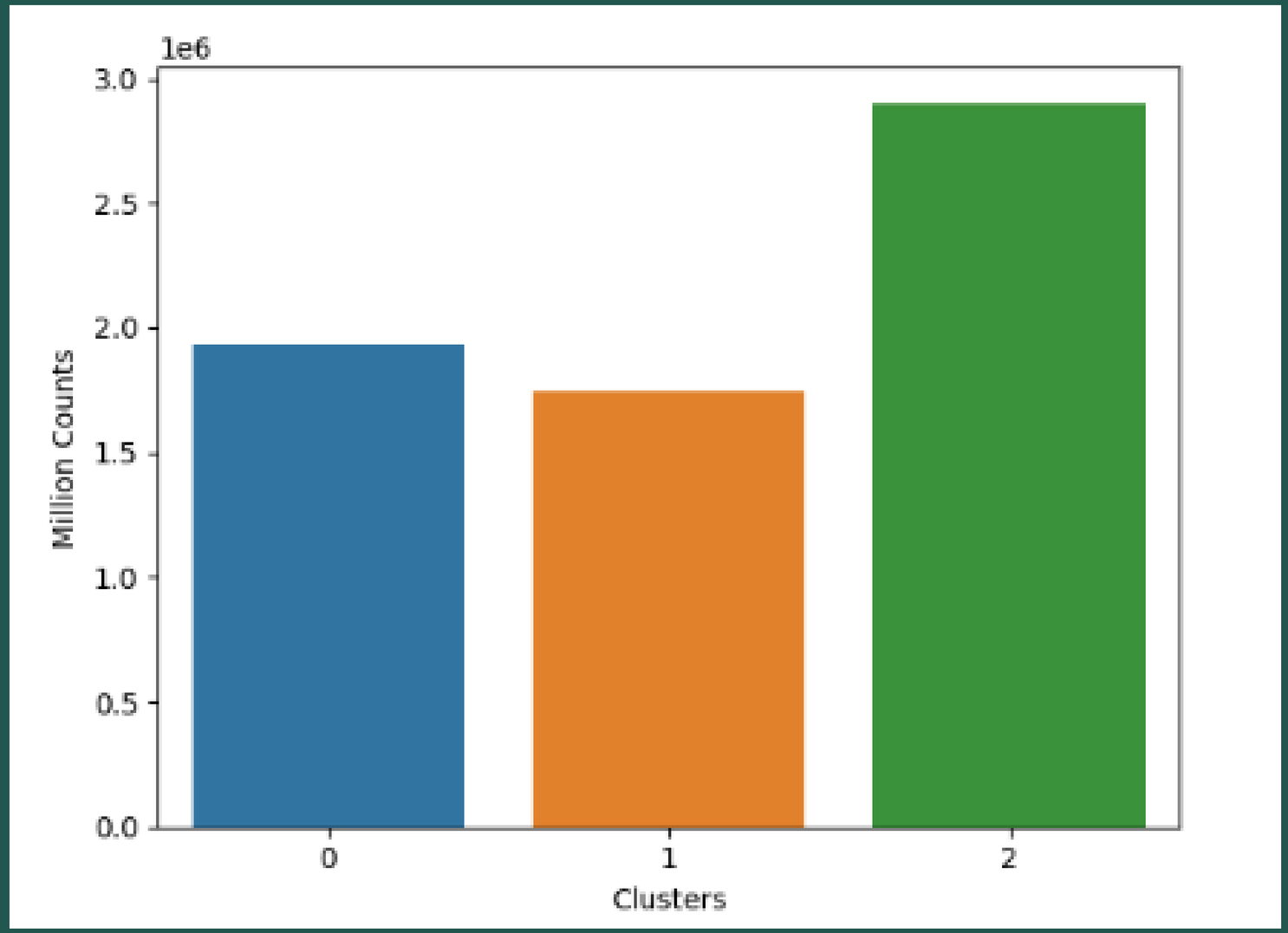
Domestic related Cases by Primary Type

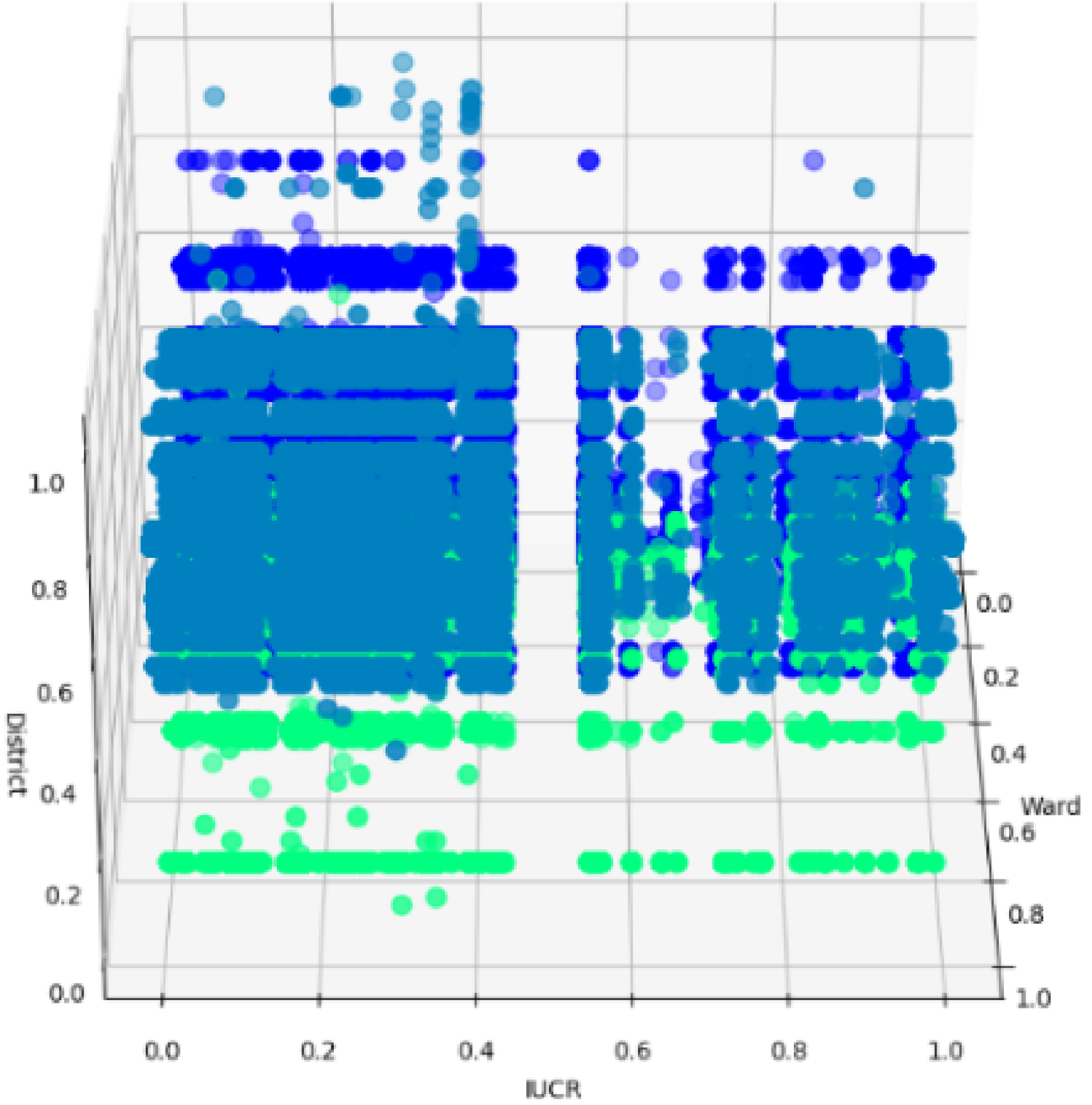


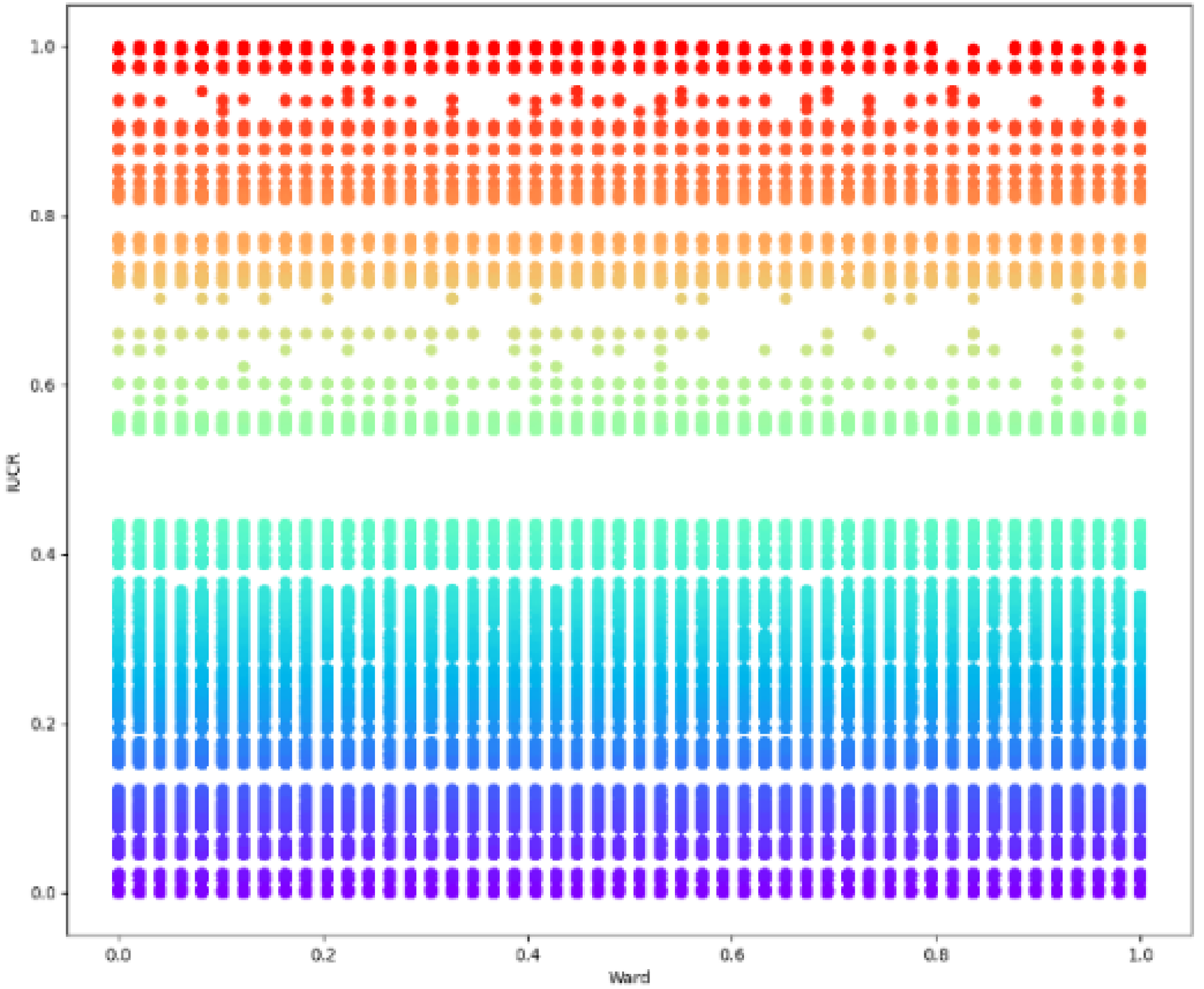
Data Mining

Clustering

For the clustering we used K-Means algorithm




District
IUCR
Ward
0.0
0.2
0.4
0.6
0.8
1.0


1.0
0.8
0.6
0.4
0.2
0.0
IUCR
0.0
0.2
0.4
0.6
0.8
1.0
Ward

Data Mining

Correlations Between Crimes

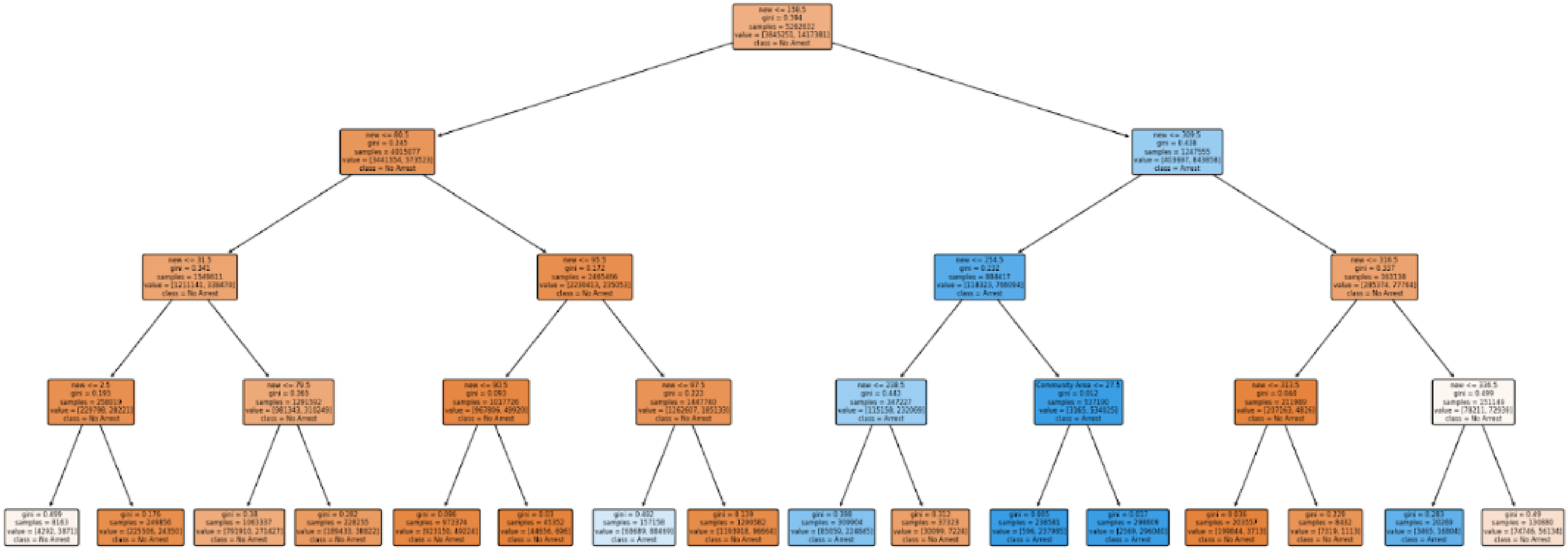
Apriori Algorithm

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(ASSAULT)	(BATTERY)	0.192308	0.538462	0.192308	1.000000	1.857143	0.088757	inf
1	(ASSAULT)	(BURGLARY)	0.192308	0.141026	0.115385	0.600000	4.254545	0.088264	2.147436
2	(BURGLARY)	(ASSAULT)	0.141026	0.192308	0.115385	0.818182	4.254545	0.088264	4.442308
3	(ASSAULT)	(CRIMINAL DAMAGE)	0.192308	0.410256	0.192308	1.000000	2.437500	0.113412	inf
4	(CRIMINAL TRESPASS)	(ASSAULT)	0.025641	0.192308	0.012821	0.500000	2.600000	0.007890	1.615385
5	(MOTOR VEHICLE THEFT)	(ASSAULT)	0.051282	0.192308	0.051282	1.000000	5.200000	0.041420	inf
6	(ASSAULT)	(NARCOTICS)	0.192308	0.282051	0.179487	0.933333	3.309091	0.125247	10.769231
7	(NARCOTICS)	(ASSAULT)	0.282051	0.192308	0.179487	0.636364	3.309091	0.125247	2.221154
8	(ASSAULT)	(OTHER OFFENSE)	0.192308	0.141026	0.141026	0.733333	5.200000	0.113905	3.221154
9	(OTHER OFFENSE)	(ASSAULT)	0.141026	0.192308	0.141026	1.000000	5.200000	0.113905	inf
10	(ROBBERY)	(ASSAULT)	0.025641	0.192308	0.025641	1.000000	5.200000	0.020710	inf
11	(ASSAULT)	(THEFT)	0.192308	0.641026	0.192308	1.000000	1.560000	0.069034	inf
12	(BURGLARY)	(BATTERY)	0.141026	0.538462	0.141026	1.000000	1.857143	0.065089	inf
13	(CRIMINAL DAMAGE)	(BATTERY)	0.410256	0.538462	0.397436	0.968750	1.799107	0.176529	14.769231
14	(BATTERY)	(CRIMINAL DAMAGE)	0.538462	0.410256	0.397436	0.738095	1.799107	0.176529	2.251748
15	(CRIMINAL TRESPASS)	(BATTERY)	0.025641	0.538462	0.025641	1.000000	1.857143	0.011834	inf
16	(DECEPTIVE PRACTICE)	(BATTERY)	0.064103	0.538462	0.064103	1.000000	1.857143	0.029586	inf
17	(MOTOR VEHICLE THEFT)	(BATTERY)	0.051282	0.538462	0.051282	1.000000	1.857143	0.023669	inf
18	(NARCOTICS)	(BATTERY)	0.282051	0.538462	0.282051	1.000000	1.857143	0.130178	inf
19	(BATTERY)	(NARCOTICS)	0.538462	0.282051	0.282051	0.523810	1.857143	0.130178	1.507692
20	(OTHER OFFENSE)	(BATTERY)	0.141026	0.538462	0.141026	1.000000	1.857143	0.065089	inf
21	(ROBBERY)	(BATTERY)	0.025641	0.538462	0.025641	1.000000	1.857143	0.011834	inf
22	(THEFT)	(BATTERY)	0.641026	0.538462	0.538462	0.840000	1.560000	0.193294	2.884615
23	(BATTERY)	(THEFT)	0.538462	0.641026	0.538462	1.000000	1.560000	0.193294	inf
24	(BURGLARY)	(CRIMINAL DAMAGE)	0.141026	0.410256	0.141026	1.000000	2.437500	0.083169	inf
25	(MOTOR VEHICLE THEFT)	(BURGLARY)	0.051282	0.141026	0.025641	0.500000	3.545455	0.018409	1.717949

Decision Tree

The purpose of this decision tree model is to predict whether an incident results in an arrest or not. It is using the features of the dataset 'Community Area' and 'Primary Type' to make predictions.

Accuracy= 86,47%



- **The Gini impurity measures the probability that if we pick an item at random this will be classified wrongly.**
- **The Gini impurity can be computed by summing the probability fifi of an item being of class ii times the probability 1-fi1-fi of a mistake in categorizing that item.**

THANK YOU!!!

Any questions?