

**ΟΙΚΟΝΟΜΙΚΟ  
ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΑΘΗΝΩΝ**



ATHENS UNIVERSITY  
OF ECONOMICS  
AND BUSINESS

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΕΡΓΑΣΙΑ 7<sup>ου</sup> ΕΞΑΜΗΝΟΥ

ΜΑΘΗΜΑ: Επιχειρηματική Ευφυΐα και Ανάλυση Μεγάλων  
Δεδομένων

Υποβληθείσα στον Καθηγητή:

Χατζηαντωνίου Δαμιανό

Οι σπουδαστές:

Κακωνάς Νικόλαος - ΑΜ: 8190050, t8190050@aueb.gr

Μανιουδάκη Γεωργία - ΑΜ: 8190097 – t8190097@aueb.gr

ΑΘΗΝΑ 2023

# Περιεχόμενα

<b>Εισαγωγή</b>	<b>4</b>
<b>1. Εύρεση και Δημιουργία του Dataset</b>	<b>4</b>
1.1 Περιγραφή	4
1.1.1 Πλαίσιο	4
1.1.2 Περιεχόμενο	4
1.1.3 Πηγή	5
1.2 Καθαρισμός και Επεξεργασία Δεδομένων	5
1.2.1 Διαγραφή Στηλών	5
1.2.2 Longitude - Latitude	6
1.2.3 Παρόμοιες τιμές	6
1.2.4 Διαγραφή NON-CRIMINAL	7
1.2.5 Date - Time	7
1.2.6 Remove commas (,)	7
1.2.7 Boolean σε Bit	7
<b>2. Data Warehouse – SQL Server</b>	<b>7</b>
2.1 Δημιουργία SQL Server	8
2.2 Καταχώρηση Δεδομένων στον SQL Server	8
2.3 Δημιουργία Dimensions	9
2.3.1 Case Dimension	9
2.3.2 Location Dimension	11
2.3.3 Location Details Dimension	12
2.3.4 Date Dimension	13
2.3.5 Time Dimension	14
2.3.6 Arrest Dimension	14
2.3.7 Domestic Dimension	15
2.4 Δημιουργία Fact Table	16
2.5 Star Schema	18
<b>3. Visual Studio</b>	<b>18</b>
3.1 Δημιουργία Project	18
3.2 Δημιουργία Datacube	20
3.3 Υπολογισμός Μετρικών	20

4. Οπτικοποίηση	21
5. Εξόρυξη Δεδομένων	29
5.1 Clustering	29
5.2 Συσχετίσεις εγκλημάτων	32
5.3 Δέντρο Απόφασης	34

## Εισαγωγή

Στην παρούσα εργασία θα εξετάσουμε ένα μεγάλο dataset, το οποίο θα καθαριστεί και θα εισαχθεί σε μία αποθήκη δεδομένων. Στην συνέχεια, θα δημιουργηθεί ένας κύβος δεδομένων και θα παραχθούν μετρικές. Έπειτα, θα χρησιμοποιηθεί ένα εργαλείο οπτικοποίησης (Power BI) για να δημιουργηθούν διάφορες περιπτώσεις οπτικοποίησης δεδομένων. Τέλος, τα δεδομένα της αποθήκης θα χρησιμοποιηθούν για κάποιες λειτουργίες εξόρυξης δεδομένων, όπως δέντρα αποφάσεων για πρόβλεψη τιμών, clustering και συσχετίσεις μεταξύ διαφόρων τιμών.

Το dataset που θα αναλυθεί αφορά τα δεδομένα για τις καταχωρήσεις εγκλημάτων στο Σικάγο. Επιλέξαμε το συγκεκριμένο dataset καθώς μας φάνηκαν αρκετά ενδιαφέρουσες οι αναλύσεις και τα συμπεράσματα που μπορεί να προκύψουν από αυτό, καθώς θα μπορούσαμε να παράξουμε μετα-δεδομένα όπως:

- Clustering βάση τριών διαστάσεων
- Συσχετίσεις μεταξύ των τύπων των εγκλημάτων αναλόγως με την περιοχή
- Δέντρο απόφασης για την πρόβλεψη σύλληψης βάσει τύπου εγκλήματος και κοινότητας

## 1. Εύρεση και Δημιουργία του Dataset

### 1.1 Περιγραφή

#### 1.1.1 Πλαίσιο

Το Dataset αφορά τις ποινικές υποθέσεις που έλαβαν χώρα στο Chicago από το 2001 μέχρι και τους δύο πρώτους μήνες του 2021. Τα δεδομένα εξάγονται από το σύστημα CLEAR (Citizen Law Enforcement Analysis and Reporting) του Αστυνομικού Τμήματος του Σικάγου. Προκειμένου να προστατευθεί το απόρρητο των θυμάτων εγκλημάτων, οι διευθύνσεις εμφανίζονται μόνο σε επίπεδο μπλοκ και δεν προσδιορίζονται συγκεκριμένες τοποθεσίες. Αυτά τα στοιχεία περιλαμβάνουν μη επαληθευμένες αναφορές που παρασχέθηκαν στο Αστυνομικό Τμήμα. Οι προκαταρκτικές ταξινομήσεις εγκλημάτων μπορούν να αλλάξουν σε μεταγενέστερη ημερομηνία βάσει πρόσθετης έρευνας και υπάρχει πάντα η πιθανότητα μηχανικού ή ανθρώπινου λάθους. Ως εκ τούτου, το Αστυνομικό Τμήμα του Σικάγο δεν εγγυάται την ακρίβεια, πληρότητα, επικαιρότητα ή τη σωστή αλληλουχία των πληροφοριών.

#### 1.1.2 Περιεχόμενο

Το Dataset περιέχει περίπου 7.2 εκατομμύρια εγγραφές και 22 στήλες. Αυτές οι στήλες περιγράφουν τα παρακάτω για κάθε μοναδικό έγκλημα, που ορίζεται ανά σειρά:

- ID: Μοναδικό αναγνωριστικό για την εγγραφή

- Case Number: Ο αριθμός RD της Αστυνομίας του Σικάγο (Αριθμός Records Division), ο οποίος είναι μοναδικός για το περιστατικό
- Date: Ημερομηνία που συνέβη το περιστατικό (ή μερικές φορές η καλύτερη εκτίμηση)
- Block: Η μερικώς διορθωμένη διεύθυνση όπου συνέβη το περιστατικό, τοποθετώντας την στο ίδιο μπλοκ με την πραγματική διεύθυνση
- IUCR: Ο κώδικας αναφοράς εγκλήματος του Illinois Unifrom. Αυτό συνδέεται άμεσα με τον κύριο τύπο και την περιγραφή
- Primary Type: Η κύρια περιγραφή του κώδικα IUCR
- Description: Η δευτερεύουσα περιγραφή του κώδικα IUCR, μια υποκατηγορία της κύριας περιγραφής
- Location Description: Περιγραφή της τοποθεσίας όπου συνέβη το περιστατικό
- Arrest: Υποδεικνύει εάν έγινε σύλληψη ή όχι
- Domestic: Υποδεικνύει εάν το περιστατικό εγκλήματος σχετίζεται με domestic violence (ενδοοικογενειακή βία σε ελεύθερη μετάφραση) ή όχι
- Beat: Υποδεικνύει το beat όπου συνέβη το περιστατικό. Ένα beat είναι η μικρότερη γεωγραφική περιοχή της αστυνομίας
- District: Υποδεικνύει την αστυνομική περιφέρεια όπου συνέβη το περιστατικό
- Ward: Η πτέρυγα (περιφέρεια του Δημοτικού Συμβουλίου) όπου συνέβη το περιστατικό
- Community Area: Υποδεικνύει την κοινοτική περιοχή όπου συνέβη το περιστατικό. Το Σικάγο έχει 77 κοινοτικές περιοχές
- FBI Code: Υποδεικνύει την ταξινόμηση του εγκλήματος όπως περιγράφεται στο Εθνικό Σύστημα Αναφοράς Βάσει Συμβάντων (NIBRS) του FBI
- X Coordinate: Η συντεταγμένη x της τοποθεσίας όπου συνέβη το περιστατικό στην προβολή State Plane Illinois East NAD 1983. Αυτή η θέση μετατοπίζεται από την πραγματική θέση για μερική επεξεργασία, αλλά εμπίπτει στο ίδιο μπλοκ
- Y Coordinate: Η συντεταγμένη y της τοποθεσίας όπου συνέβη το περιστατικό στην προβολή State Plane Illinois East NAD 1983. Αυτή η θέση μετατοπίζεται από την πραγματική θέση για μερική επεξεργασία, αλλά εμπίπτει στο ίδιο μπλοκ.
- Year: Χρονιά που συνέβη το περιστατικό
- Updated On: Ημερομηνία και ώρα τελευταίας ενημέρωσης της εγγραφής
- Latitude: Το γεωγραφικό πλάτος της τοποθεσίας όπου συνέβη το περιστατικό. Αυτή η θέση μετατοπίζεται από την πραγματική θέση για μερική επεξεργασία, αλλά εμπίπτει στο ίδιο μπλοκ
- Longitude: Το γεωγραφικό μήκος της τοποθεσίας όπου συνέβη το περιστατικό. Αυτή η θέση μετατοπίζεται από την πραγματική θέση για μερική επεξεργασία, αλλά εμπίπτει στο ίδιο μπλοκ
- Location: Ο συνδυασμός γεωγραφικού μήκους και πλάτους

### 1.1.3 Πηγή

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

## 1.2 Καθαρισμός και Επεξεργασία Δεδομένων

### 1.2.1 Διαγραφή Στηλών

Για τον καθαρισμό των δεδομένων αρχικά αφαιρέσαμε από το dataset τις στήλες Year, καθώς έχουμε τη στήλη Date και Updated On, εφόσον δεν μας ενδιαφέρει η πληροφορία αυτής της στήλης.

```
In [4]: df = df.drop(columns=['Year', 'Updated On'])
```

Επομένως, οι στήλες που έχουν μείνει είναι οι εξής:

- ID
- Case Number
- Date
- Time
- Block
- IUCR
- Primary Type
- Description
- Location Description
- Arrest
- Domestic
- Beat
- District
- Ward
- Community Area
- FBI Code
- X Coordinate
- Y Coordinate
- Longitude
- Latitude
- Location

### 1.2.2 Longitude - Latitude

Έπειτα, μετά από έρευνα βρήκαμε το εύρος τιμών που μπορεί να πάρει το γεωγραφικό μήκος και πλάτος οποιουδήποτε σημείου που ανήκει στο Σικάγο. Το διάστημα αυτό είναι το [-87.907127, -87.524050] για το γεωγραφικό μήκος και [41.644335, 42.023131] για το γεωγραφικό πλάτος. Επομένως, κρατήσαμε μόνο τις εγγραφές εκείνες για τις οποίες οι στήλες Longitude και Latitude βρίσκονται στα αντίστοιχα διαστήματα.

```
In [6]: df = df.loc[(df['Longitude'] >= -87.907127) & (df['Longitude'] <= -87.524050)]  
df = df.loc[(df['Latitude'] <= 42.023131) & (df['Latitude'] >= 41.644335)]
```

### 1.2.3 Παρόμοιες τιμές

Παρατηρήσαμε επίσης, ότι κάποιες τιμές στην στήλη Primary Type αναφέρονται στο ίδιο έγκλημα, αλλά έχουν καταχωρηθεί σαν διαφορετικό string, όπως για παράδειγμα το NON - CRIMINAL και το NON-CRIMINAL, που η μόνη διαφορά τους είναι το κενό πριν και μετά την παύλα.

```
In [8]: df['Primary Type'] = df['Primary Type'].replace('NON - CRIMINAL', 'NON-CRIMINAL')  
df['Primary Type'] = df['Primary Type'].replace('CRIM SEXUAL ASSAULT', 'CRIMINAL SEXUAL ASSAULT')
```

#### 1.2.4 Διαγραφή NON-CRIMINAL

Παρατηρώντας τις τιμές της στήλης 'Primary Type', αποφασίσαμε να διαγράψουμε όλες τις τιμές που αναφέρονται ως NON - CRIMINAL, καθώς δεν αποτελούν εγκλήματα.

```
In [ ]: df = df[(df['Primary Type'] != 'NON-CRIMINAL')]
```

#### 1.2.5 Date - Time

Ακόμη, θεωρήσαμε καλή πρακτική να χωρίσουμε τη στήλη DateTime, η οποία περιέχει την ημερομηνία και την ώρα της καταγραφής του συμβάντος, σε δύο νέες, την Date και την Time, έτσι ώστε η ημερομηνία και η ώρα να βρίσκονται σε διαφορετικές στήλες. Παράλληλα, η ώρα περνάει σε μορφή 24 ωρών και διώχνουμε τα PM και AM. Για παράδειγμα η τιμή "2020-05-23 01:30:00 PM" της στήλης DateTime θα μεταφερθεί στις στήλες Date και Time με τις τιμές "2020-05-23" και "13:30:00" αντίστοιχα και η στήλη DateTime θα διαγραφεί.

```
In [12]: df.rename(columns={'Date': 'DateTime'}, inplace=True)
df['DateTime'] = pd.to_datetime(df['DateTime'])
df['DateTime'] = pd.to_datetime(df['DateTime'], format='%Y-%m-%d %I:%M:%S %p')
df['DateTime'] = df['DateTime'].dt.strftime('%Y-%m-%d %H:%M:%S')
```

```
In [13]: df['DateTime'] = pd.to_datetime(df['DateTime'])
df.insert(3, 'Date', df['DateTime'].dt.date)
df.insert(4, 'Time', df['DateTime'].dt.time)
```

```
In [18]: df = df.drop(columns=['DateTime'])
```

#### 1.2.6 Remove commas (,)

Επίσης, κάποιες τιμές του dataset και συγκεκριμένα string τιμές, είχαν κάποια κόμματα. Αυτό οδηγούσε στην λανθασμένη εισχώρηση των δεδομένων στη βάση, καθώς τα κόμματα άλλαζαν τον τρόπο που μπαίνουν τα δεδομένα στις στήλες των πινάκων, εφόσον η πηγή μας είναι ένα csv αρχείο. Επομένως, αφαιρέσαμε όλα τα κόμματα από το dataset.

```
In [17]: df = df.replace(',', '', regex=True)
```

#### 1.2.7 Boolean σε Bit

Τέλος, μετατρέψαμε τις τιμές των στηλών Arrest και Domestic σε 0 και 1, από True και False αντίστοιχα.

```
In [19]: df['Arrest'] = df['Arrest'].apply(lambda x: 0 if not x else 1)
df['Domestic'] = df['Domestic'].apply(lambda x: 0 if not x else 1)
df
```

Τα δεδομένα μας, μετά τον καθαρισμό, έχουν συνολικό μέγεθος 1.25 GB και 6.578.291 συνολικές εγγραφές.

## 2. Data Warehouse – SQL Server

## 2.1 Δημιουργία SQL Server

Για να περάσουμε τα δεδομένα μας στη βάση, χρειάστηκε να δημιουργήσουμε μία βάση δεδομένων την οποία ονομάσαμε `chicagoCrimes`. Πιο συγκεκριμένα, αρχικά, για την εισαγωγή των δεδομένων στη βάση, δημιουργήσαμε τον πίνακα `crimes`, με τους ακόλουθους τύπους δεδομένων.

```
USE [chicagoCrimes]
GO

CREATE TABLE [crimes] (
    [ID] bigint,
    [Case Number] varchar(10),
    [Date] date,
    [Time] time,
    [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)
)
```

## 2.2 Καταχώρηση Δεδομένων στον SQL Server

Για την καταχώρηση των δεδομένων στη βάση, χρησιμοποιήσαμε το `BULK INSERT` το οποίο φαίνεται στην επόμενη σελίδα. Πιο συγκεκριμένα:

- Το `FIRSTROW` χρησιμοποιείται για τον καθορισμό του αριθμού της πρώτης σειράς στο αρχείο που περιέχει δεδομένα. Σε αυτήν την περίπτωση, η πρώτη σειρά δεδομένων είναι η σειρά 2, όπως καθορίζεται από το `FIRSTROW = 2`
- Το `FORMAT` χρησιμοποιείται για τον καθορισμό της μορφής των δεδομένων στο αρχείο. Σε αυτήν την περίπτωση, τα δεδομένα είναι σε μορφή CSV, όπως ορίζεται από το `FORMAT = CSV`
- Το `MAXERRORS` χρησιμοποιείται για τον καθορισμό του μέγιστου αριθμού σφαλμάτων που επιτρέπονται πριν από τον τερματισμό της λειτουργίας μαζικής εισαγωγής. Σε αυτήν την περίπτωση, δεν επιτρέπονται σφάλματα και η μαζική εισαγωγή θα αποτύχει στο πρώτο σφάλμα, όπως καθορίζεται από `MAXERRORS = 0`



- Το FIELDQUOTE χρησιμοποιείται για τον καθορισμό ενός χαρακτήρα που χρησιμοποιείται ως χαρακτηριστικό για τον τερματισμό της γραμμής. Ο καθορισμένος χαρακτήρας είναι τα διπλά εισαγωγικά (")
- Το FIELDTERMINATOR χρησιμοποιείται για τον καθορισμό του χαρακτήρα που διαχωρίζει τα πεδία στο αρχείο εισόδου. Ο καθορισμένος χαρακτήρας είναι το κόμμα (,)
- Το ROWTERMINATOR χρησιμοποιείται για τον καθορισμό των χαρακτήρων που σηματοδοτούν το τέλος μιας σειράς στο αρχείο εισόδου. Ο καθορισμένος χαρακτήρας είναι \n

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

Επομένως, έχουμε δημιουργήσει τον πίνακα crimes όπως φαίνεται στην παρακάτω εικόνα:

ID	Case Number	Date	Time	Block	ICR#	Primary Type	Description	Location Description	Arrest	Domestic	Bait	Dist	Ward	Community Area	FBI Code	X Coordinate	Y Coordinate	Latitude	Longitude	Location	
1	10224738	HY411648	2015-09-05	13:30:00	043XX S WOOD ST	0485	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	0	1	924	9	12	61	088	118074	1075917	41.815117282	-87.609999562	41.815117282 -87.609999562
2	10224739	HY411615	2015-09-04	11:30:00	008XX N CENTRAL AVE	0870	THEFT	POCKET-PICKING	CTA BUS	0	0	1511	15	29	25	06	110875	1904859	41.895000471	-87.765400451	41.895000471 -87.765400451
3	10224740	HY411699	2015-09-04	12:45:00	036XX W BARRY AVE	2023	NARCOTICS	POSSE HEPICOMBRA/TAG	SIDEWALK	1	0	1412	14	28	21	18	1192037	1920384	41.817405763	-87.716649687	41.817405763 -87.716649687
4	10224741	HY411620	2015-09-05	13:00:00	009XX N LARAMIE AVE	0560	ASSAULT	SIMPLE	APARTMENT	0	1	1522	15	28	25	08A	1341786	1900286	41.801803443	-87.755217152	41.801803443 -87.755217152
5	10224742	HY411435	2015-09-05	10:55:00	082XX S LOOMS BLVD	0610	BURGLARY	FORCIBLE ENTRY	RESIDENCE	0	0	614	6	21	71	05	108430	1890165	41.744378879	-87.658430635	41.744378879 -87.658430635
6	10224743	HY411629	2015-09-04	18:00:00	021XX W CHURCHILL ST	0620	BURGLARY	UNLAWFUL ENTRY	RESIDENCE GARAGE	0	0	1434	14	32	34	05	1181628	1912157	41.914625603	-87.681630909	41.914625603 -87.681630909
7	10224744	HY411655	2015-09-04	02:00:00	028XX W CERRADO RD	0860	THEFT	RETAIL THEFT	GROCERY FOOD STORE	1	0	1034	10	25	31	06	1185754	1883113	41.851988805	-87.680219118	41.851988805 -87.680219118
8	10224745	HY411654	2015-09-05	11:30:00	031XX W WASHINGTON BLVD	0330	ROBBERY	STRONGARM- NO WEAPON	STREET	0	1	1222	12	27	27	03	1195836	1900915	41.80281374	-87.704325717	41.80281374 -87.704325717
9	10224746	HY411662	2015-09-05	14:00:00	071XX S PULASKI RD	0820	THEFT	\$500 AND UNDER	PARKING LOT/GARAGE/NO	0	0	833	8	13	65	06	1192838	1857556	41.763647552	-87.722344893	41.763647552 -87.722344893
10	10224749	HY411626	2015-09-05	11:00:00	052XX N MILWAUKEE AVE	0460	BATTERY	SIMPLE	SMALL RETAIL STORE	0	0	1623	16	45	11	08B	1137969	1934340	41.979684415	-87.768014257	41.979684415 -87.768014257
11	10224750	HY411632	2015-09-05	03:00:00	009XX W 103RD ST	0520	OTHER OFFE.	TELEPHONE THREAT	APARTMENT	0	1	912	9	34	49	25	1171071	1336676	41.707154919	-87.624244893	41.707154919 -87.624244893
12	10224751	HY411661	2015-09-05	12:50:00	018XX E 47TH ST	0486	BATTERY	DOMESTIC BATTERY SIMPLE	STREET	0	1	222	2	4	39	08B	1185907	1874105	41.809678314	-87.593638934	41.809678314 -87.593638934
13	10224752	HY411601	2015-09-03	13:00:00	028XX W SCHILLER ST	0810	THEFT	OVER \$500	STREET	0	0	1424	14	1	34	06	1162574	1909428	41.907127255	-87.678232016	41.907127255 -87.678232016
14	10224753	HY411489	2015-09-05	11:45:00	088XX S JUSTICE ST	0487	BATTERY	AGGRAVATED DOMESTIC B.	APARTMENT	0	0	612	6	21	71	04B	1167400	1891512	41.740097343	-87.662166183	41.740097343 -87.662166183
15	10224754	HY411604	2015-09-05	13:30:00	007XX N LEAVINGTON AVE	1320	CRIMINAL DA.	TO VEHICLE	STREET	0	0	1531	15	28	25	14	1141889	1904448	41.893869916	-87.754341096	41.893869916 -87.754341096
16	10224756	HY410204	2015-07-08	00:00:00	102XX S TORRENCE AVE	0620	BURGLARY	UNLAWFUL ENTRY	OTHER	0	0	434	4	10	51	05	1195858	1836950	41.707490122	-87.559650325	41.707490122 -87.559650325
17	10224757	HY411388	2015-09-04	09:55:00	088XX S PALLAS ST	0610	BURGLARY	FORCIBLE ENTRY	RESIDENCE	1	0	2221	22	21	71	05	1160554	1846067	41.733173536	-87.685421087	41.733173536 -87.685421087
18	10224758	HY411688	2015-09-05	12:35:00	058XX W GRACE ST	0486	BATTERY	DOMESTIC BATTERY SIMPLE	STREET	0	1	1633	16	38	15	08B	1130014	1934956	41.848424769	-87.775439328	41.848424769 -87.775439328
19	10224759	HY411581	2015-09-04	22:30:00	108XX S AVENUE L	1320	CRIMINAL DA.	TO VEHICLE	STREET	0	0	432	4	10	52	14	1201814	1839627	41.713031786	-87.536487809	41.713031786 -87.536487809
20	10224761	HY411620	2015-09-02	00:01:00	048XX W WYNNEMAC AVE	1310	CRIMINAL DA.	TO PROPERTY	RESIDENCE	0	0	1623	16	45	12	14	1142583	1933126	41.972552295	-87.751076326	41.972552295 -87.751076326
21	10224762	HY411593	2015-09-04	15:00:00	108XX S AVENUE L	1320	CRIMINAL DA.	TO VEHICLE	STREET	0	0	432	4	10	52	14	1201814	1839627	41.713031786	-87.536487809	41.713031786 -87.536487809
22	10224763	HY411606	2015-09-05	13:45:00	020XX E 71ST ST	0330	ROBBERY	STRONGARM- NO WEAPON	SIDEWALK	0	0	331	3	5	43	03	1186886	1905321	41.764247597	-87.575997062	41.764247597 -87.575997062
23	10224764	HY411550	2015-09-03	21:00:00	100XX S AVENUE L	1320	CRIMINAL DA.	TO VEHICLE	STREET	0	0	432	4	10	52	14	1201814	1839626	41.713127829	-87.536488623	41.713127829 -87.536488623
24	10224765	HY411129	2015-09-05	08:45:00	077XX S SOUTH SHORE DR	0560	ASSAULT	SIMPLE	APARTMENT	0	1	421	4	7	43	08A	1191240	1894764	41.756330319	-87.552716204	41.756330319 -87.552716204
25	10224766	HY411188	2015-09-05	06:20:00	078XX S PHILLIPS AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	0	1	421	4	7	43	08B	1193460	1891495	41.756420222	-87.560152335	41.756420222 -87.560152335
26	10224767	HY410417	2015-09-04	16:15:00	078XX S SAGINAW AVE	031A	ROBBERY	ARMED- HANDGUN	SIDEWALK	0	0	422	4	7	46	03	1195224	1892594	41.75135862	-87.560154545	41.75135862 -87.560154545
27	10224768	HY410631	2015-09-03	20:30:00	078XX S ESCANABA AVE	0810	THEFT	OVER \$500	VEHICLE NONCOMMERCIAL	0	0	421	4	7	43	06	1196887	1893400	41.752596188	-87.554055126	41.752596188 -87.554055126
28	10224769	HY411672	2015-09-05	14:30:00	009XX S STATE ST	0890	THEFT	FROM BUILDING	RESTAURANT	0	0	123	1	2	32	06	1171646	1897564	41.876307424	-87.627209951	41.876307424 -87.627209951
29	10224770	HY411586	2015-09-05	12:40:00	005XX W 61ST PL	0430	BATTERY	AGGRAVATED- OTHER DAN.	RESIDENCE PORCH/HALL	1	0	711	7	20	68	04B	1171953	1864134	41.782597483	-87.638979769	41.782597483 -87.638979769
30	10224772	HY411581	2015-09-04	23:00:00	011XX N ASHLAND AVE	0810	THEFT	OVER \$500	STREET	0	0	1213	12	1	24	06	1165522	1907528	41.901851233	-87.667456951	41.901851233 -87.667456951
31	10224773	HY411685	2015-09-05	07:00:00	028XX W FIFTH AVE	0620	BURGLARY	UNLAWFUL ENTRY	APARTMENT	0	0	1124	11	2	27	05	1193626	1899592	41.880252848	-87.699244681	41.880252848 -87.699244681
32	10224775	HY411686	2015-09-05	09:00:00	068XX S HEATING AVE	2025	OTHER OFFE.	HARASSMENT BY TELEPHO.	RESIDENCE	0	1	833	8	13	65	25	1146215	1899113	41.763888657	-87.746235842	41.763888657 -87.746235842
33	10224776	HY411687	2015-09-05	14:45:00	071XX S PULASKI RD	0810	THEFT	OVER \$500	PARKING LOT/GARAGE/NO	0	0	833	8	13	65	06	1192838	1857556	41.763647552	-87.722344893	41.763647552 -87.722344893
34	10224778	HY411675	2015-09-05	14:44:00	047XX N KEELER AVE	0560	ASSAULT	SIMPLE	SIDEWALK	1	0	1722	17	39	14	08A	1147526	1931300	41.967448812	-87.732951137	41.967448812 -87.732951137

## 2.3 Δημιουργία Dimensions

Για τη δημιουργία των Dimensions, χωρίσαμε τις στήλες σε κατηγορίες που έβγαζαν νόημα και δημιουργήσαμε τα dimensions. Τα demensions στα οποία καταλήξαμε ήταν τα εξής:

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

### 2.3.1 Case Dimension

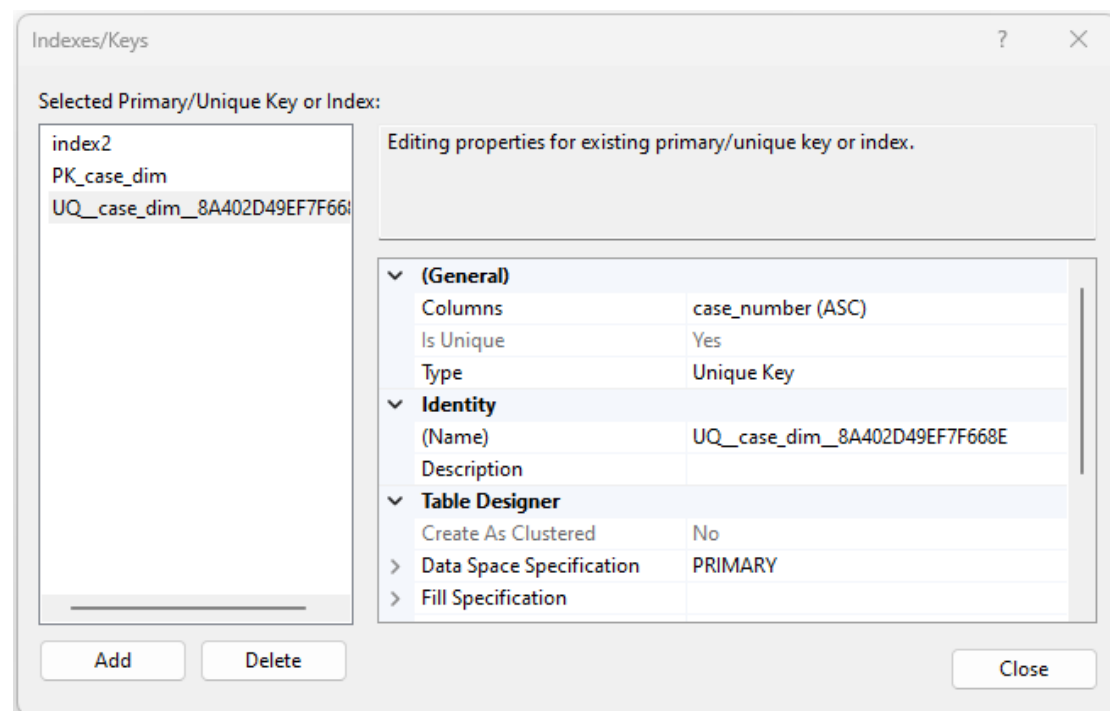
Το Case Dimension αναφέρεται σε όλα τα δεδομένα που χρειάζονται για να μπορέσει κάποιος να καταλάβει το είδος του εγκλήματος. Πιο συγκεκριμένα, για την δημιουργία του

συγκεκριμένου dimension συγκεντρώσαμε τις εξής στήλες από τον πίνακα crimes: Case Number, IUCR, Primary Type, Description. Επομένως, χρειάστηκε να δημιουργήσουμε τον πίνακα case\_dim στην βάση, έτσι ώστε να περάσουμε εκεί τα κατάλληλα δεδομένα.

Αυτός είναι ο πίνακας που δημιουργήθηκε στη βάση:

	Column Name	Data Type	Allow Nulls
🔑	case_id	int	<input type="checkbox"/>
	case_number	varchar(10)	<input type="checkbox"/>
	iucr	varchar(4)	<input type="checkbox"/>
	primary_type	varchar(50)	<input type="checkbox"/>
	description	varchar(100)	<input type="checkbox"/>
			<input type="checkbox"/>

Στον πίνακα αυτό βλέπουμε ότι το case\_id έχει καταχωρηθεί primary key σε αυτόν τον πίνακα και ως ξένο κλειδί στο fact table, ενώ το case\_number έχει καταχωρηθεί ως unique.



Έπειτα καταχωρούμε τα δεδομένα στην συγκεκριμένη διάσταση από τον αρχικό πίνακα crimes. Ενδεικτικό κομμάτι κώδικα που χρησιμοποιήσαμε για τον σκοπό αυτό είναι το εξής:

```
INSERT INTO case_dim (case_number, iucr, primary_type, [description])
SELECT [Case Number], [IUCR], [Primary Type], [Description]
FROM crimes
```

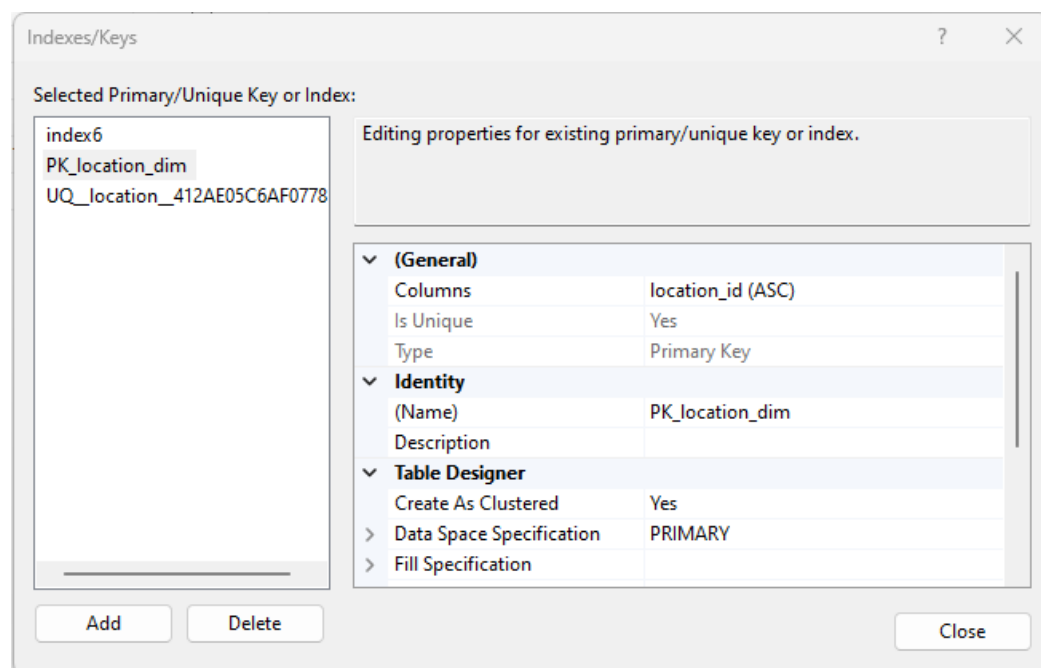
### 2.3.2 Location Dimension

Το Location Dimension αναφέρεται στην συγκέντρωση όλων των δεδομένων που δημιουργούν την πληροφορία της τοποθεσίας. Πιο συγκεκριμένα, για την δημιουργία του συγκεκριμένου dimension συγκεντρώσαμε τις εξής στήλες από τον πίνακα crimes: Block, Location Description, Beat, District, Ward, Community Area, X Coordinate, Y Coordinate, Longitude, Latitude, Location. Επομένως, χρειάστηκε να δημιουργήσουμε τον πίνακα location\_dim στην βάση, έτσι ώστε να περάσουμε εκεί τα κατάλληλα δεδομένα.

Αυτός είναι ο πίνακας που δημιουργήθηκε στη βάση:

	Column Name	Data Type	Allow Nulls
🔑	location_id	int	<input type="checkbox"/>
	block	varchar(50)	<input type="checkbox"/>
	beat	int	<input type="checkbox"/>
	district	int	<input type="checkbox"/>
	ward	int	<input type="checkbox"/>
	community_area	int	<input type="checkbox"/>
	x_coordinate	int	<input type="checkbox"/>
	y_coordinate	int	<input type="checkbox"/>
	longitude	float	<input type="checkbox"/>
	latitude	float	<input type="checkbox"/>
	location	varchar(100)	<input type="checkbox"/>
			<input type="checkbox"/>

Στον πίνακα αυτό βλέπουμε ότι το location\_id έχει καταχωρηθεί primary key σε αυτόν τον πίνακα και ως ξένο κλειδί στο fact table, ενώ το location έχει καταχωρηθεί ως unique.



Έπειτα καταχωρούμε τα δεδομένα στην συγκεκριμένη διάσταση από τον αρχικό πίνακα crimes. Ενδεικτικό κομμάτι κώδικα που χρησιμοποιήσαμε για τον σκοπό αυτό είναι το εξής:

```


INSERT INTO location_dim (block, beat, district, ward, community_area, x_coordinate, y_coordinate, longitude, latitude, location)
SELECT [Block], [Beat], [District], [Ward], [Community Area], [X Coordinate], [Y Coordinate], [Longitude], [Latitude], [Location]
FROM crimes

```

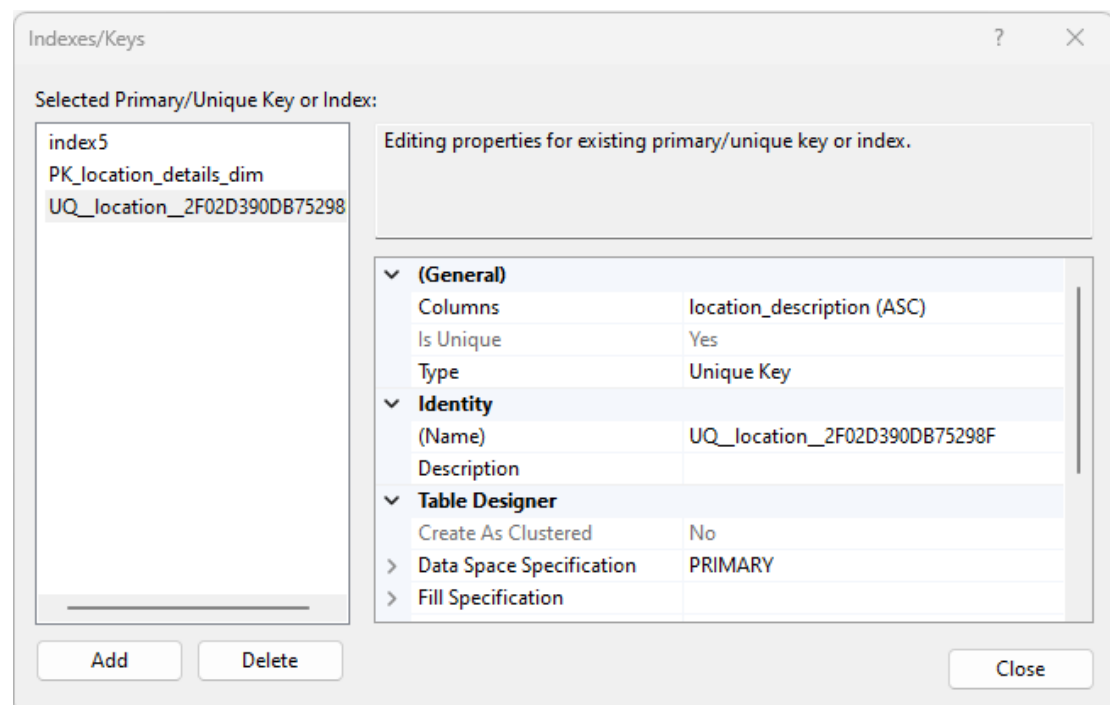
### 2.3.3 Location Details Dimension

Το συγκεκριμένο dimension αναφέρεται σε λεπτομέρειες, όσον αφορά την τοποθεσία. Πιο συγκεκριμένα, δίνει την πληροφορία για το συγκεκριμένο μέρος που έγινε το έγκλημα, όπως για παράδειγμα, ένα εγκαταλειμμένο κτίριο ή ένα αεροπλάνο. Για την δημιουργία του συγκεκριμένου dimension συγκεντρώσαμε τις εξής στήλες από από τον πίνακα crimes: Block, Location Description, Beat, District, Ward, Community Area, X Coordinate, Y Coordinate, Longitude, Latitude, Location. Επομένως, χρειάστηκε να δημιουργήσουμε τον πίνακα location\_details\_dim στην βάση, έτσι ώστε να περάσουμε εκεί τα κατάλληλα δεδομένα.

Αυτός είναι ο πίνακας που δημιουργήθηκε στη βάση:

	Column Name	Data Type	Allow Nulls
	location_details_id	int	<input type="checkbox"/>
	location_description	varchar(100)	<input type="checkbox"/>
			<input type="checkbox"/>

Στον πίνακα αυτό βλέπουμε ότι το location\_details\_id έχει καταχωρηθεί primary key σε αυτόν τον πίνακα και ως ξένο κλειδί στο fact table, ενώ το location\_description έχει καταχωρηθεί ως unique.



Έπειτα καταχωρούμε τα δεδομένα στην συγκεκριμένη διάσταση από τον αρχικό πίνακα crimes.

```

INSERT INTO location_details_dim (location_description)
SELECT [Location Description]
FROM crimes

```

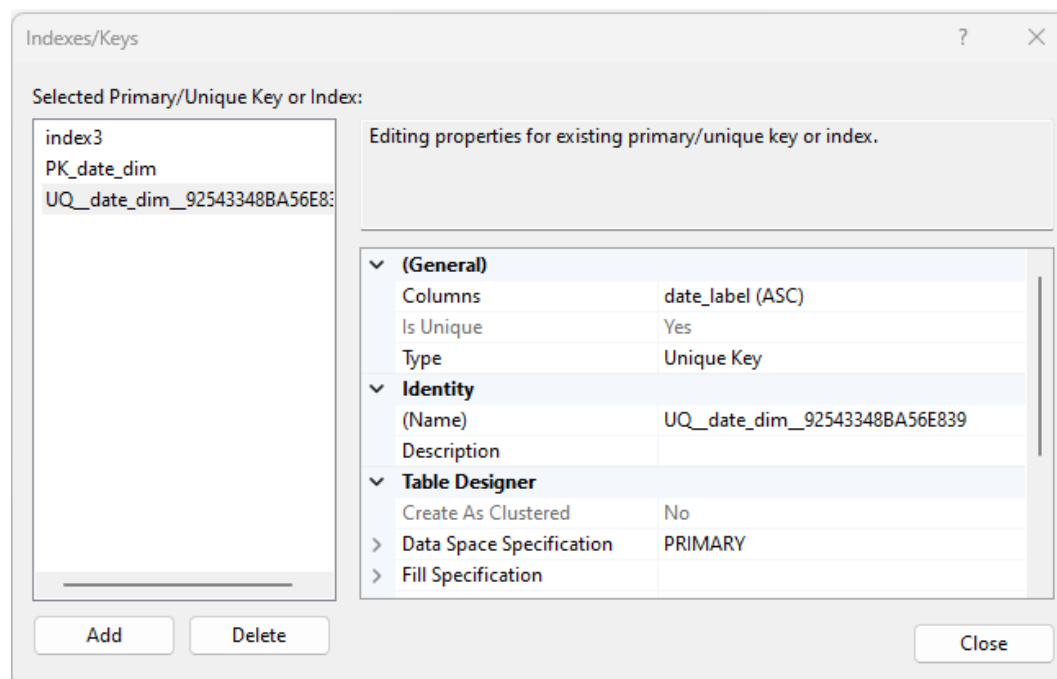
#### 2.3.4 Date Dimension

Φυσικά, από την δημιουργία κατηγοριών και dimensions δεν θα μπορούσε να λείπει η ημερομηνία. Πιο συγκεκριμένα, εκτός από την στήλη Date που χρησιμοποιήσαμε από τον αρχικό πίνακα crimes, δημιουργήσαμε και τις στήλες: Year, Month και Day. Επομένως, χρειάστηκε να δημιουργήσουμε τον πίνακα date\_dim στην βάση, έτσι ώστε να περάσουμε εκεί τα κατάλληλα δεδομένα.

Αυτός είναι ο πίνακας που δημιουργήθηκε στη βάση:

	Column Name	Data Type	Allow Nulls
🔑	date_id	int	<input type="checkbox"/>
	date_label	varchar(50)	<input type="checkbox"/>
	year	int	<input type="checkbox"/>
	month	int	<input type="checkbox"/>
	day	int	<input type="checkbox"/>
			<input type="checkbox"/>

Στον πίνακα αυτό βλέπουμε ότι το date\_id έχει καταχωρηθεί primary key σε αυτόν τον πίνακα και ως ξένο κλειδί στο fact table, ενώ το date\_label έχει καταχωρηθεί ως unique.



Έπειτα καταχωρούμε τα δεδομένα στην συγκεκριμένη διάσταση από τον αρχικό πίνακα crimes.

```

INSERT INTO date_dim (date_label, [year], [month], [day])
SELECT [Date], YEAR([Date]), MONTH([Date]), DAY([Date])
FROM crimes

```

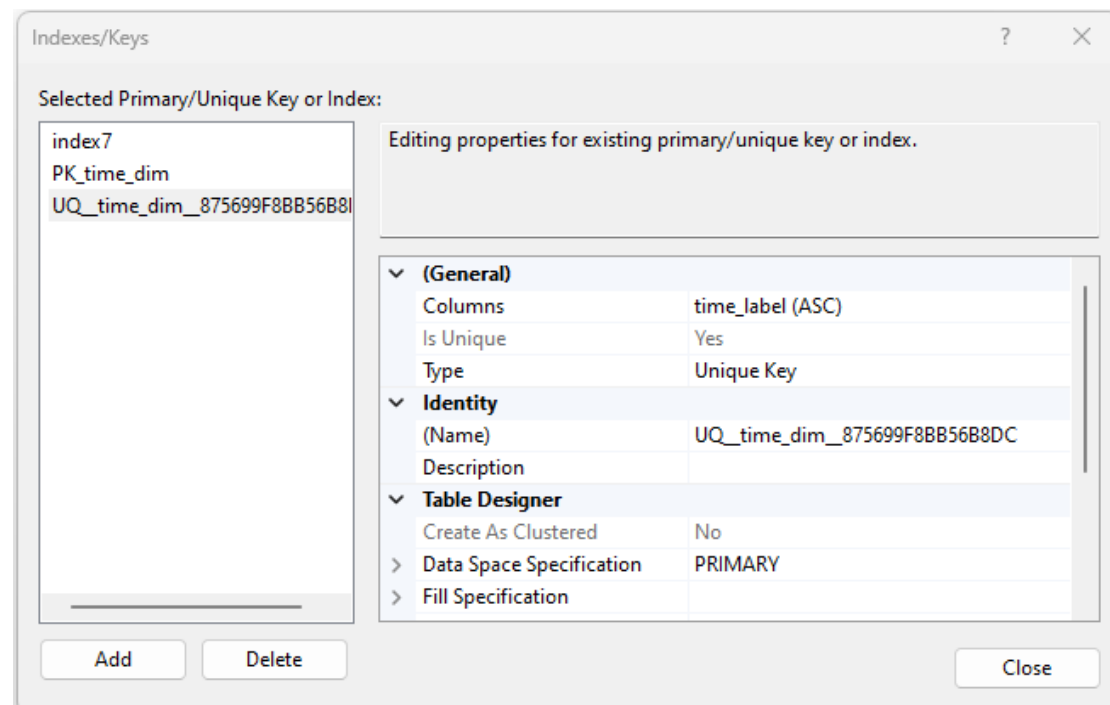
### 2.3.5 Time Dimension

Σε συνέχεια του Date Dimension, δημιουργήσαμε και το Time, το οποίο αναφέρεται στην ακριβή ώρα (μέχρι 30 λεπτά), που έγινε το περιστατικό. Πιο συγκεκριμένα, περιέχει τις στήλες: Time, Hour και Min. Επομένως, χρειάστηκε να δημιουργήσουμε τον πίνακα time\_dim στην βάση, έτσι ώστε να περάσουμε εκεί τα κατάλληλα δεδομένα.

Αυτός είναι ο πίνακας που δημιουργήθηκε στη βάση:

	Column Name	Data Type	Allow Nulls
🔑	time_id	int	<input type="checkbox"/>
	time_label	varchar(50)	<input type="checkbox"/>
	hour	int	<input type="checkbox"/>
	min	int	<input type="checkbox"/>
			<input type="checkbox"/>

Στον πίνακα αυτό βλέπουμε ότι το time\_id έχει καταχωρηθεί primary key σε αυτόν τον πίνακα και ως ξένο κλειδί στο fact table, ενώ το time\_label έχει καταχωρηθεί ως unique.



Έπειτα καταχωρούμε τα δεδομένα στην συγκεκριμένη διάσταση από τον αρχικό πίνακα crimes.


```
INSERT INTO time_dim (time_label, [hour], [min])  
SELECT [Time], DATEPART(hour, [Time]), DATEPART(minute, [Time])  
FROM crimes;
```

### 2.3.6 Arrest Dimension

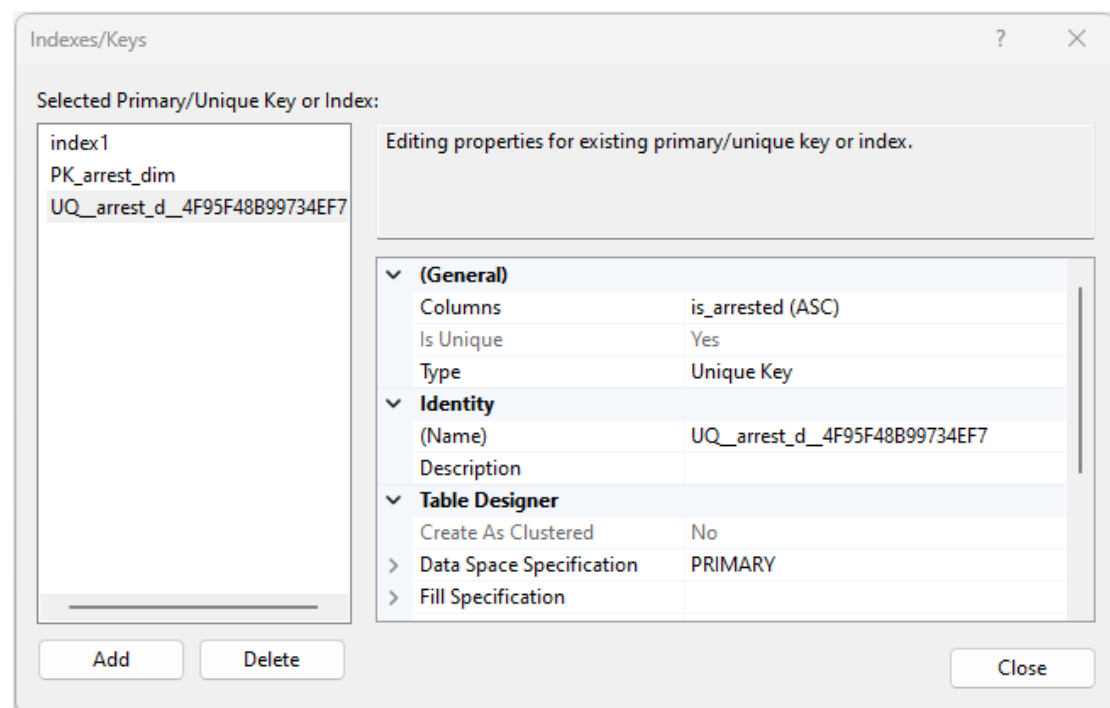
Το συγκεκριμένο dimension διαθέτει την πληροφορία για το αν έχει γίνει σύλληψη του ατόμου που διέπραξε το έγκλημα ή όχι. Πιο συγκεκριμένα, για την κατασκευή αυτού του dimension χρησιμοποιήσαμε από τον πίνακα crimes τη στήλη Arrest. Επομένως, χρειάστηκε

να δημιουργήσουμε τον πίνακα `arrest_dim` στην βάση, έτσι ώστε να περάσουμε εκεί τα κατάλληλα δεδομένα.

Αυτός είναι ο πίνακας που δημιουργήθηκε στη βάση:

	Column Name	Data Type	Allow Nulls
	arrest_id	int	<input type="checkbox"/>
	is_arrested	bit	<input type="checkbox"/>
			<input type="checkbox"/>

Στον πίνακα αυτό βλέπουμε ότι το `arrest_id` έχει καταχωρηθεί `primary key` σε αυτόν τον πίνακα και ως ξένο κλειδί στο `fact table`, ενώ το `is_arrested` έχει καταχωρηθεί ως `unique`.



Έπειτα καταχωρούμε τα δεδομένα στην συγκεκριμένη διάσταση από τον αρχικό πίνακα `crimes`.

```
INSERT INTO arrest_dim (is_arrested)
SELECT [Arrest]
FROM crimes
```

### 2.3.7 Domestic Dimension

Το συγκεκριμένο `dimension` διαθέτει την πληροφορία για το αν τα άτομα που εμπλάκηκαν στο περιστατικό, τόσο το θύμα όσο και ο δράστης, έχουν κάποια συγγένεια. Πιο συγκεκριμένα από τον πίνακα `crimes` έχει χρησιμοποιηθεί η στήλη `Domestic`. Επομένως, χρειάστηκε να δημιουργήσουμε τον πίνακα `domestic_dim` στην βάση, έτσι ώστε να περάσουμε εκεί τα κατάλληλα δεδομένα.

Αυτός είναι ο πίνακας που δημιουργήθηκε στη βάση:

	Column Name	Data Type	Allow Nulls
🔑	domestic_id	int	<input type="checkbox"/>
	is_domestic	bit	<input type="checkbox"/>
			<input type="checkbox"/>

Στον πίνακα αυτό βλέπουμε ότι το domestic\_id έχει καταχωρηθεί primary key σε αυτόν τον πίνακα και ως ξένο κλειδί στο fact table, ενώ το is\_domestic έχει καταχωρηθεί ως unique.

Έπειτα καταχωρούμε τα δεδομένα στην συγκεκριμένη διάσταση από τον αρχικό πίνακα crimes.

```


INSERT INTO domestic_dim (is_domestic)
SELECT [Domestic]
FROM crimes

```

## 2.4 Δημιουργία Fact Table

Η μόνη στήλη που δεν χρησιμοποιήσαμε στα dimensions από τον πίνακα crimes, είναι η FBI Code. Τη στήλη αυτή την βάζουμε μέσα στο fact table μαζί με τα foreign keys από τα dimensions και το primary key του fact table, το id. Επομένως, δημιουργείται το fact table όπως φαίνεται παρακάτω:



	Column Name	Data Type	Allow Nulls
	date	int	<input type="checkbox"/>
	time	int	<input type="checkbox"/>
	location	int	<input type="checkbox"/>
	location_details	int	<input type="checkbox"/>
	[case]	int	<input type="checkbox"/>
	arrest	int	<input type="checkbox"/>
	domestic	int	<input type="checkbox"/>
	id	int	<input type="checkbox"/>
	fbi_code	varchar(50)	<input checked="" type="checkbox"/>
			<input type="checkbox"/>

Παρατηρούμε ότι η μοναδική τιμή στο fact table που δέχεται null τιμές είναι το fbi\_code. Παρακάτω φαίνεται ο τρόπος με τον οποίο ορίσαμε τα ξένα κλειδιά.

Foreign Key Relationships

?

×

Selected Relationship:

FK\_criminal\_case\_fact\_arrest\_dim

FK\_criminal\_case\_fact\_case\_dim

FK\_criminal\_case\_fact\_date\_dim

FK\_criminal\_case\_fact\_domestic\_c

FK\_criminal\_case\_fact\_location\_d

FK\_criminal\_case\_fact\_location\_di

FK\_criminal\_case\_fact\_time\_dim

Add

Delete

Editing properties for existing relationship.

▼ (General)

Check Existing Data On Creat Yes

> Tables And Columns Specific

▼ Identity

(Name) FK\_criminal\_case\_fact\_arrest\_dim

Description

▼ Table Designer

Enforce For Replication Yes

Enforce Foreign Key Constrai Yes

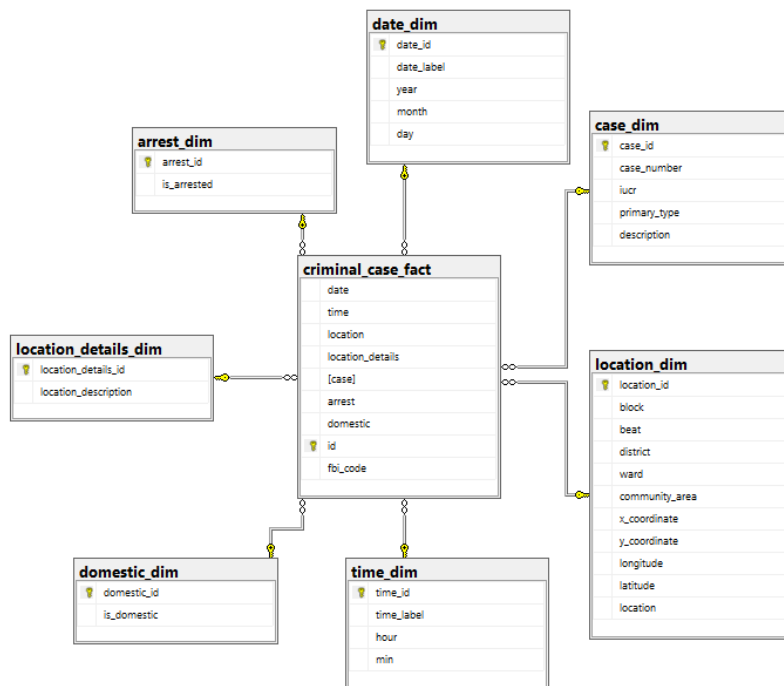
> INSERT And UPDATE Specifica

Close

These are some rows from the fact table:

	date	time	location	location_details	case	arrest	domestic	id	fbi_code
1	1014428	3282492	5064086	1524819	3915697	8	5	634	01A
2	1014428	1298126	5972281	6563907	3922051	8	5	635	01A
3	1035840	1576830	5071549	6563907	3946484	8	5	639	01A
4	1035840	1954951	4799525	6532716	3953052	6578290	5	640	01A
5	1035840	2619782	5097572	6525420	3853563	6578290	5	641	01A
6	3872087	1954951	3303320	6525420	3872087	6578290	5	644	01A
7	3884541	3644220	3884541	6525420	3884541	6578290	5	646	01A
8	3884541	2907029	5696503	6525420	3891120	6578290	5	647	01A
9	1381041	6528098	4445277	6564301	3896764	6578290	6578149	648	01A
10	1381041	1078154	3985131	6564301	3985131	6578290	6578149	652	01A

## 2.5 Star Schema



## 3. Visual Studio

### 3.1 Δημιουργία Project

Αρχικά δημιουργήσαμε ένα Multidimensional Analysis Services Project, αφού πρώτα εγκαταστήσαμε τα plugins SSIS και SSAS.

Στην συνέχεια, μέσω του wizard, περάσαμε ως data source την βάση `chicagoCrimes` που δημιουργήσαμε στον local server μας στο SQL Server Management Studio.

**Data Source Wizard**

**Select how to define the connection**  
You can select from a number of ways in which your data source will define its connection string.

☐ Create a data source based on another object  
☒ Create a data source based on an existing or new connection

Data connections:

LocalHost.chicagoCrimes
-------------------------

Data connection properties:

Property	Value
Data Source	.
Initial Catalog	chicagoCrimes
Integrated Se...	SSPI
Provider	SQLNCLI11.1

New... Delete

Έπειτα, περάσαμε τα data source views , μέσω του wizard, επιλέγοντας όλα τα dimensions και το fact table (criminal\_case\_fact).

**Data Source View Wizard**

**Select Tables and Views**  
Select objects from the relational database to be included in the data source view.

Available objects:

Name	Type
crimes (dbo)	Table

Filter:

☐ Show system objects

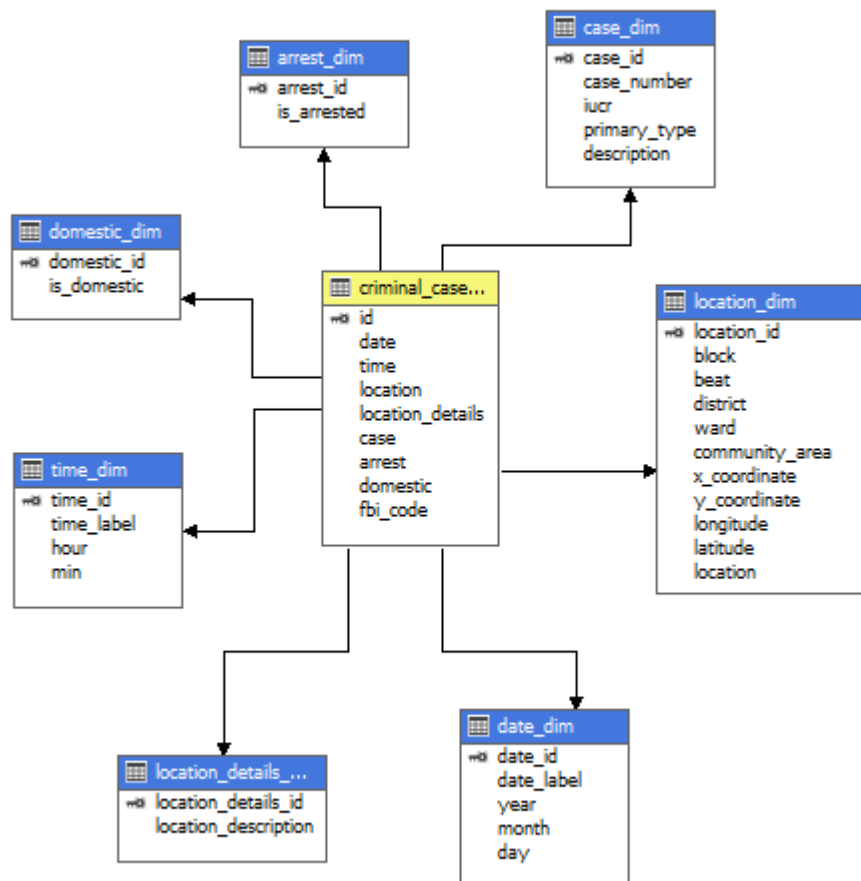
Included objects:

Name	Type
criminal_case_fact (d...	Table
domestic_dim (dbo)	Table
location_dim (dbo)	Table
arrest_dim (dbo)	Table
location_details_dim (...)	Table
time_dim (dbo)	Table
case_dim (dbo)	Table
date_dim (dbo)	Table

Add Related Tables

### 3.2 Δημιουργία Datacube

Τέλος, επιλέγοντας όλες τις διαστάσεις για την δημιουργία του κύβου, και επιλέγοντας ποιες από αυτές θα γίνουν measures και θα χρειαστεί το count, παράξαμε τον τελικό μας κύβο που παρουσιάζεται παρακάτω:



### 3.3 Υπολογισμός Μετρικών

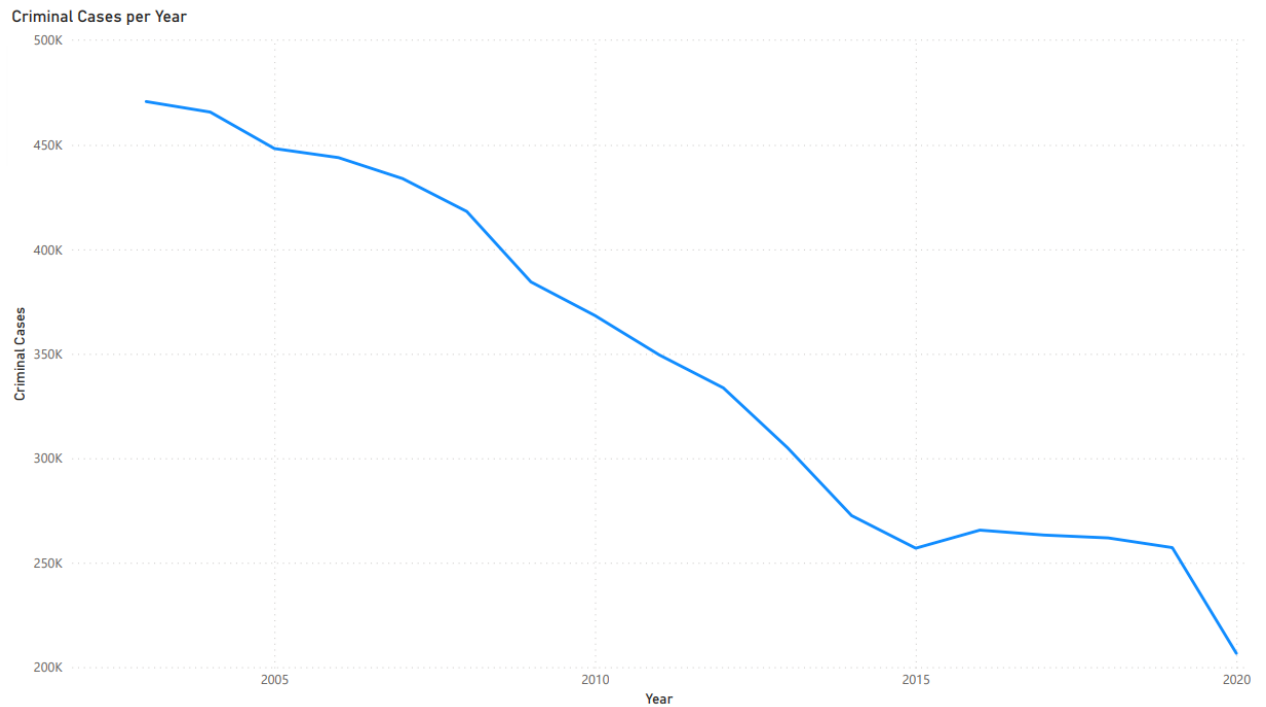
Έπειτα, προχωρήσαμε στην καταγραφή κάποιων μετρικών βασισμένων στο dataset μας, ώστε και να καταλάβουμε καλύτερα τα δεδομένα μας, αλλά και να δημιουργήσουμε μετρικές που μπορεί να μας φανούν χρήσιμες σε περαιτέρω ανάλυση.

Πιο συγκεκριμένα, οι μετρικές που βγάλαμε ήταν οι εξής:

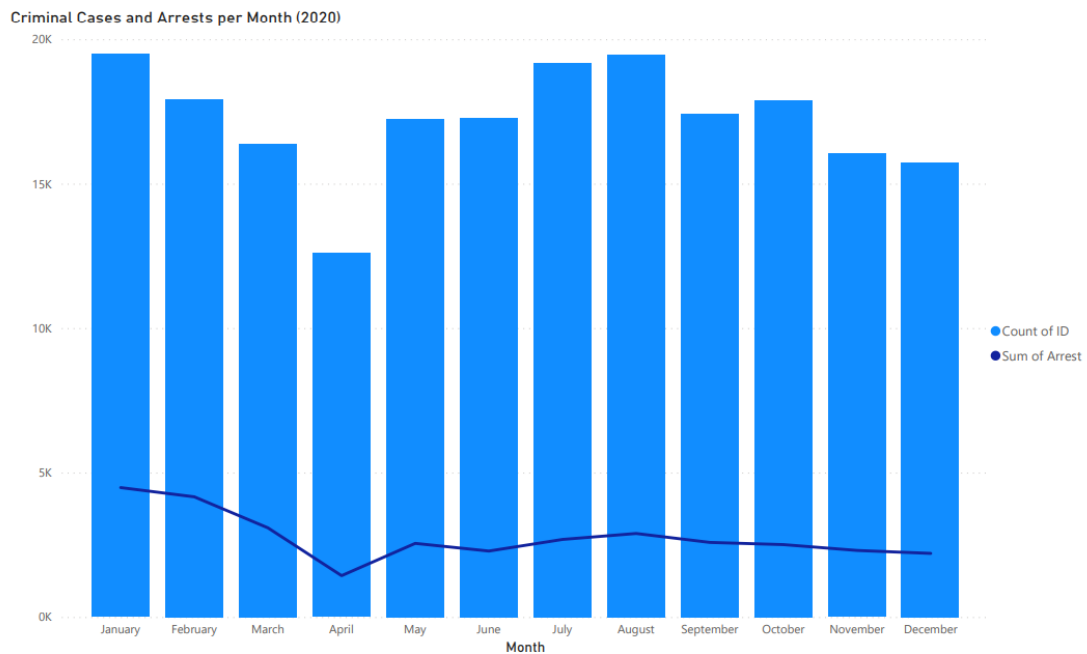
- 6.578.291 συνολικά εγκλήματα
- το 27% των παραβατών έχουν συλληφθεί
- το 13% των συνολικών εγκλημάτων αφορούν συγγενικά πρόσωπα
- τα περισσότερα εγκλήματα γίνονται τον Ιούλιο (9.4%)
- τα λιγότερα εγκλήματα γίνονται τον Φεβρουάριο (6.6%)
- το 5.7% των εγκλημάτων γίνεται στις 12:00 το μεσημέρι (ώρα με τα περισσότερα εγκλήματα)

- το 1.3% των εγκλημάτων γίνεται στις 05:00 το πρωί (ώρα με τα λιγότερα εγκλήματα)
- το 21% των εγκλημάτων αφορούν κλοπές
- the average crimes per year are 313.252

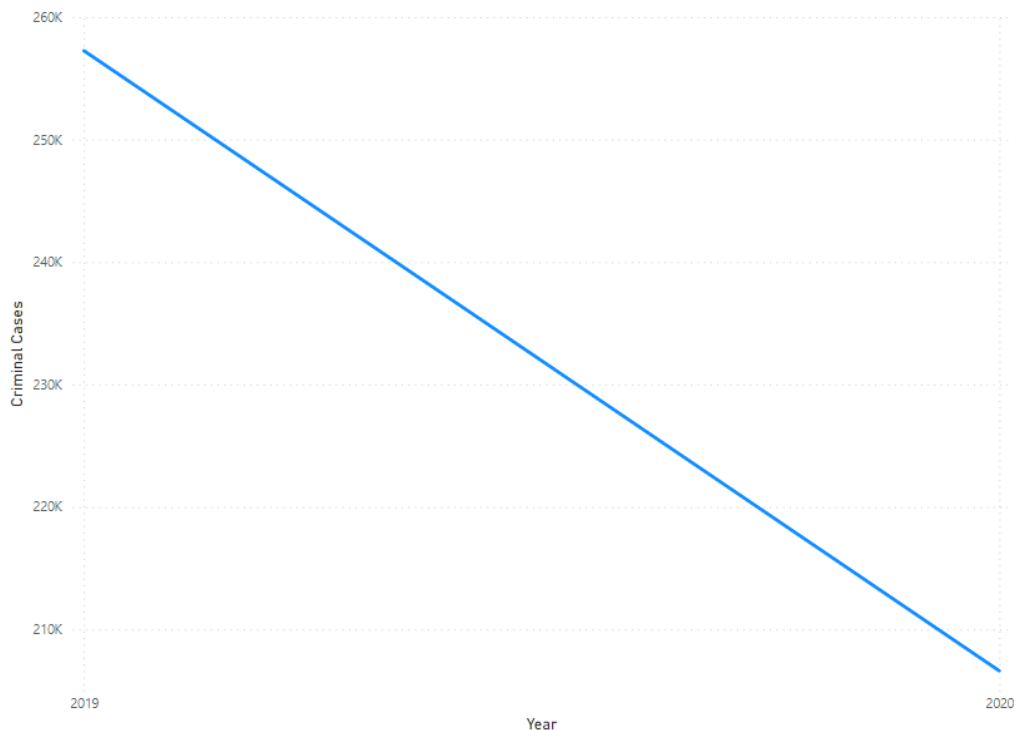
#### 4. Οπτικοποίηση



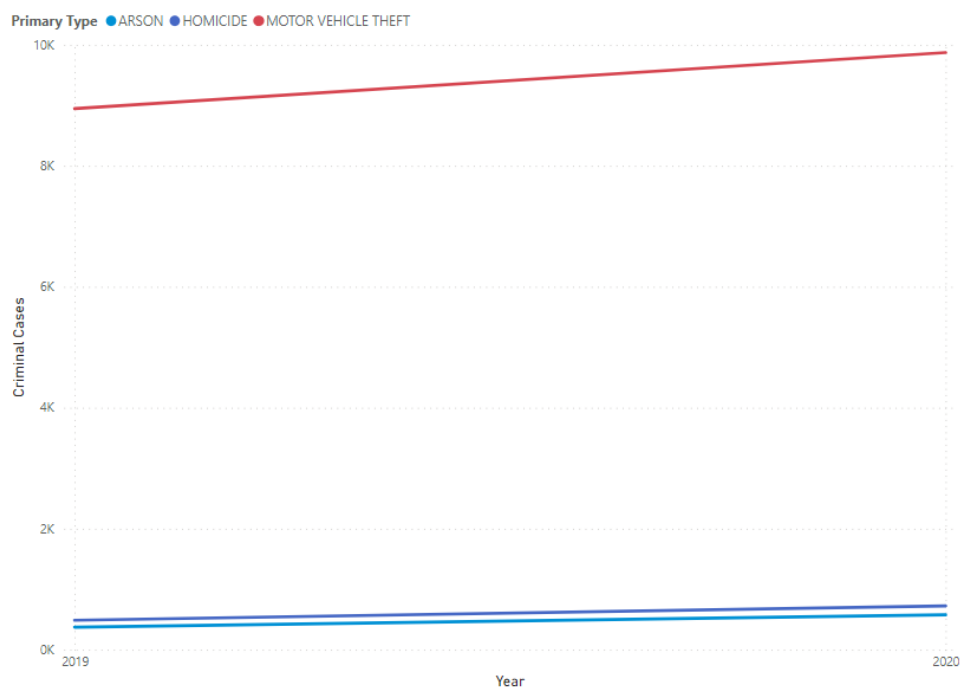
Εδώ βλέπουμε τις συνολικές ποινικές υποθέσεις κατά τις χρονιές 2003-2020, οι οποίες είναι μειούμενες με εξαίρεση το 2016, όπου παρατηρήθηκε αύξηση από 256K σε 265K. Επιλέξαμε να μην λάβουμε υπόψη το 2021 για το οποίο είχαμε δεδομένα μόνο για τους 2 πρώτους μήνες.



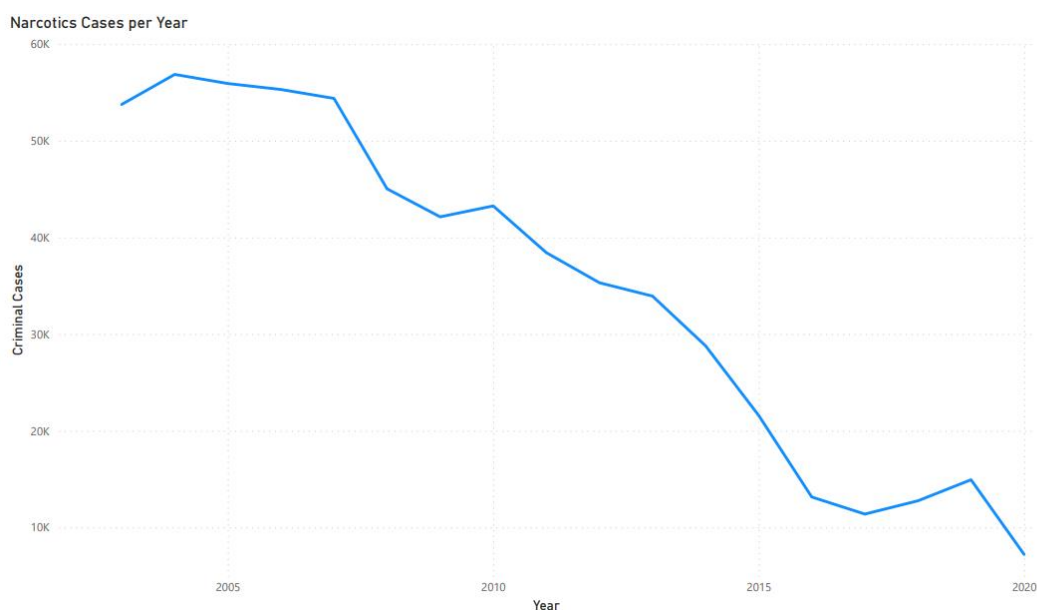
Εδώ βλέπουμε για τη χρονιά 2020, τις συνολικές ποινικές υποθέσεις ανά μήνα σε σχέση με τις συλλήψεις. Και τα δύο παρουσιάζουν καμπή τον Απρίλιο, γεγονός που ενδέχεται να σχετίζεται με την έξαρση του COVID-19. Οι συλλήψεις μετά τον Μάιο φαίνονται να εξομαλύνονται.



Εδώ βλέπουμε ότι υπάρχει σημαντική μείωση των ποινικών υποθέσεων από 257K σε 209K τις χρονιές 2019-2020. Ωστόσο, όπως βλέπουμε παρακάτω το 2020, το Σικάγο σημείωσε σημαντική αύξηση της εγκληματικότητας στις ανθρωποκτονίες, καθώς σε άλλους τύπους εγκλημάτων, όπως οι ληστείες αυτοκινήτων και οι εμπρησμοί.

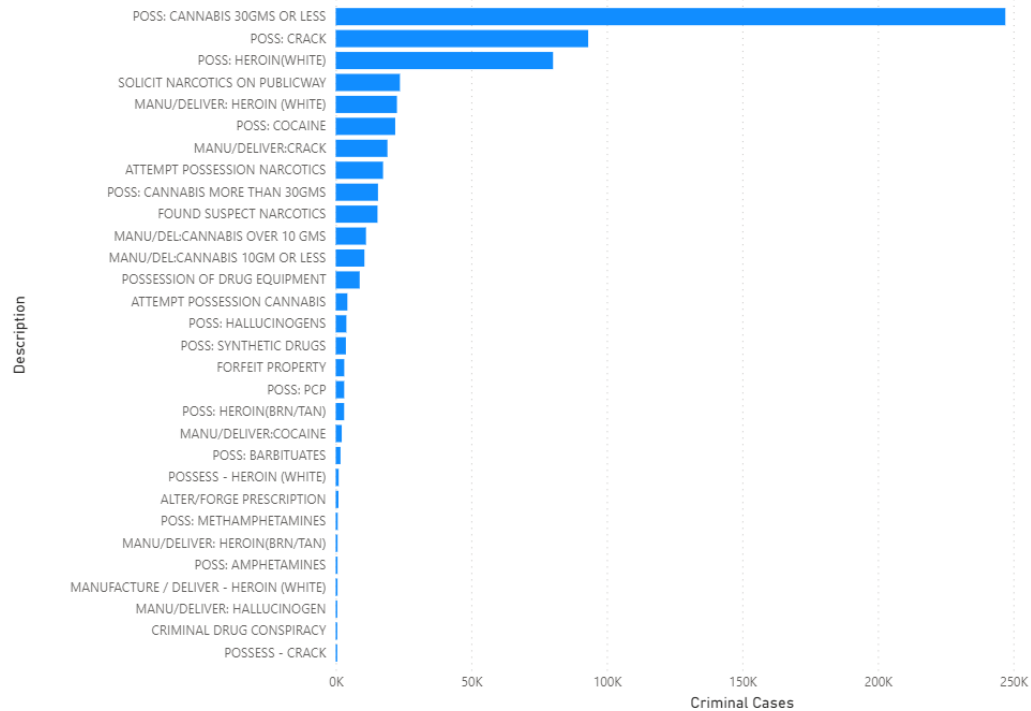


Είναι σημαντικό να σημειωθεί ότι τα ποσοστά εγκληματικότητας μπορεί να επηρεαστούν από πολλούς παράγοντες, όπως η φτώχεια, τα δημογραφικά στοιχεία, η τοποθεσία, οι στρατηγικές αστυνόμευσης και πολλοί άλλοι. Επιπλέον, η πανδημία COVID-19 είχε επίσης αντίκτυπο στα ποσοστά εγκληματικότητας σε πολλές πόλεις σε όλο τον κόσμο, συμπεριλαμβανομένου του Σικάγου.



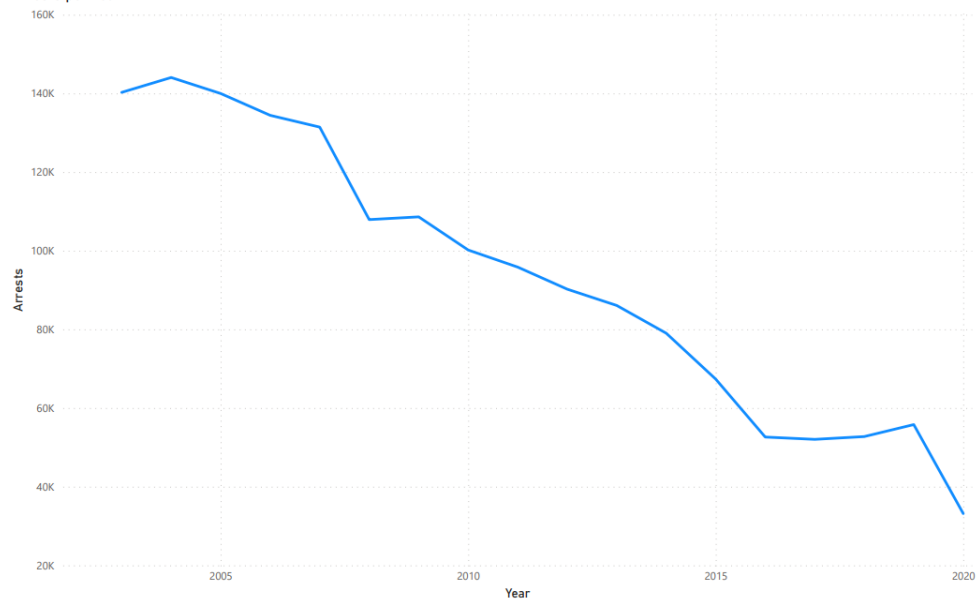
Η συνολική τάση για περιπτώσεις ναρκωτικών στο Σικάγο τα τελευταία χρόνια ήταν ένας συνδυασμός αυξήσεων και μειώσεων, με ορισμένα χρόνια να παρατηρούνται αυξήσεις στον αριθμό των αδικημάτων ναρκωτικών και άλλα χρόνια να σημειώνονται μειώσεις.

Narcotics Cases by Description



Στο διάγραμμα φαίνονται οι περιγραφές των υποθέσεων που αφορούν αδικήματα σχετικά με ναρκωτικά. Ενώ στο Σικάγο το ναρκωτικό με τη μεγαλύτερη κατανάλωση είναι η ηρωίνη, οι περισσότερες υποθέσεις αφορούν την κάνναβη, ενώ για υποθέσεις με ηρωίνη η αστυνομία του Σικάγο ενδιαφέρεται περισσότερο για τους χρήστες (possession), παρά για τους διακινητές (manu/deliver).

Arrests per Year

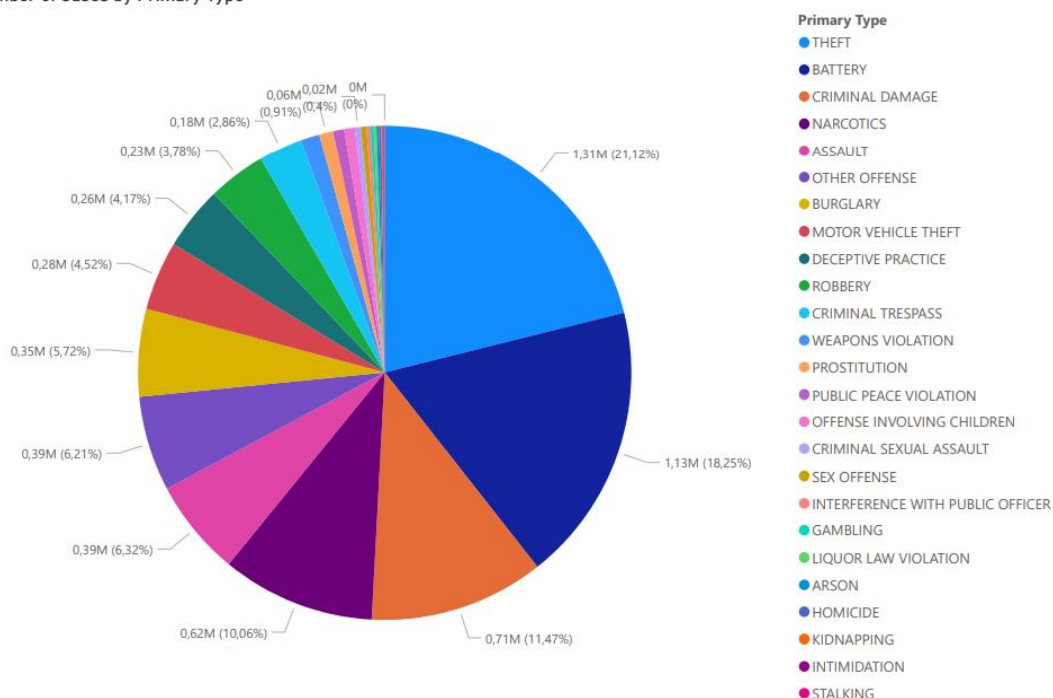


Οι συλλήψεις του Αστυνομικού Τμήματος του Σικάγο ακολουθούν μια πτωτική πορεία από το 2003, όμως τα χρόνια 2016-2019 υπάρχει μια μικρή αύξηση. Ωστόσο η μείωση των συλλήψεων από το 2019 στο 2020 είναι σημαντική.



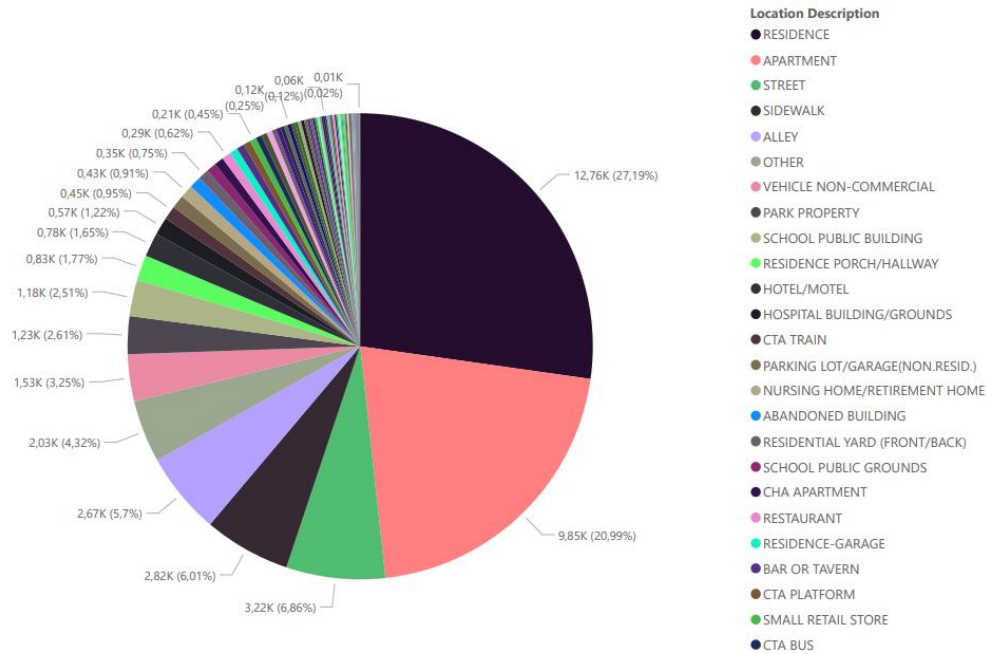
Είναι σημαντικό να σημειωθεί ότι ο αριθμός των συλλήψεων που έγιναν από την αστυνομία είναι μόνο ένα μέτρο που χρησιμοποιείται για την αξιολόγηση της επιβολής του νόμου από τις αρχές. Ο αριθμός των συλλήψεων από μόνος του δεν υποδεικνύει απαραίτητα το επίπεδο εγκληματικότητας ή τη δημόσια ασφάλεια σε μια κοινότητα, καθώς θα μπορούσε να επηρεαστεί από άλλους παράγοντες όπως αλλαγές στις στρατηγικές αστυνόμευσης, αλλαγές στην αναφορά εγκλημάτων, το μέγεθος του πληθυσμού και τα δημογραφικά στοιχεία και το συνολική εγκληματική τάση.

Number of Cases by Primary Type



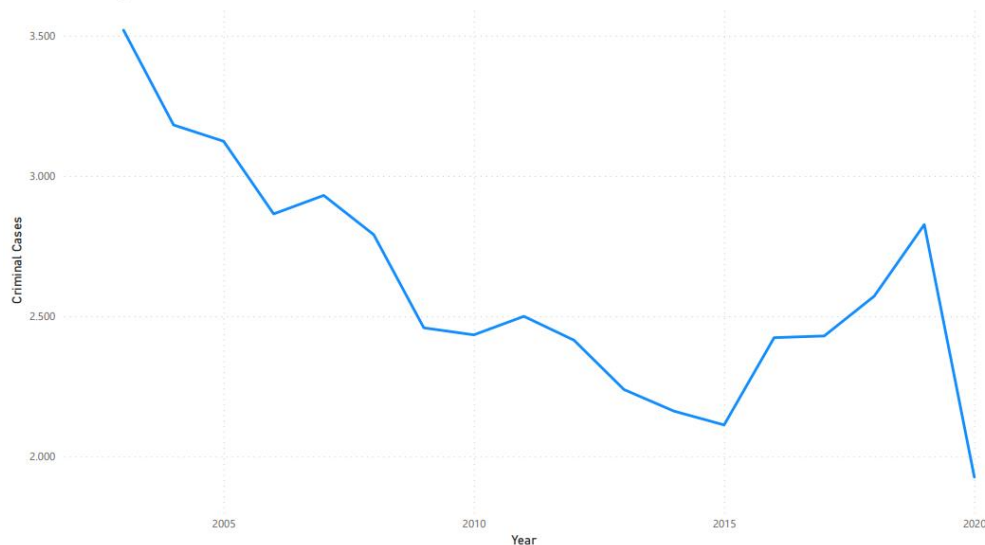
Εδώ βλέπουμε πως περισσότερο από το 50% των ποινικών υποθέσεων πρόκειται για Theft (κλοπή), Battery (βιαιοπραγία) και Criminal Damage (φθορά ξένης περιουσίας). Ακολουθούν υποθέσεις ναρκωτικών, άλλες υποθέσεις, διαρρήξεις, κλοπή ιχ, εξαπάτηση, ληστεία κλπ.

Sexcrime Cases by Location Description



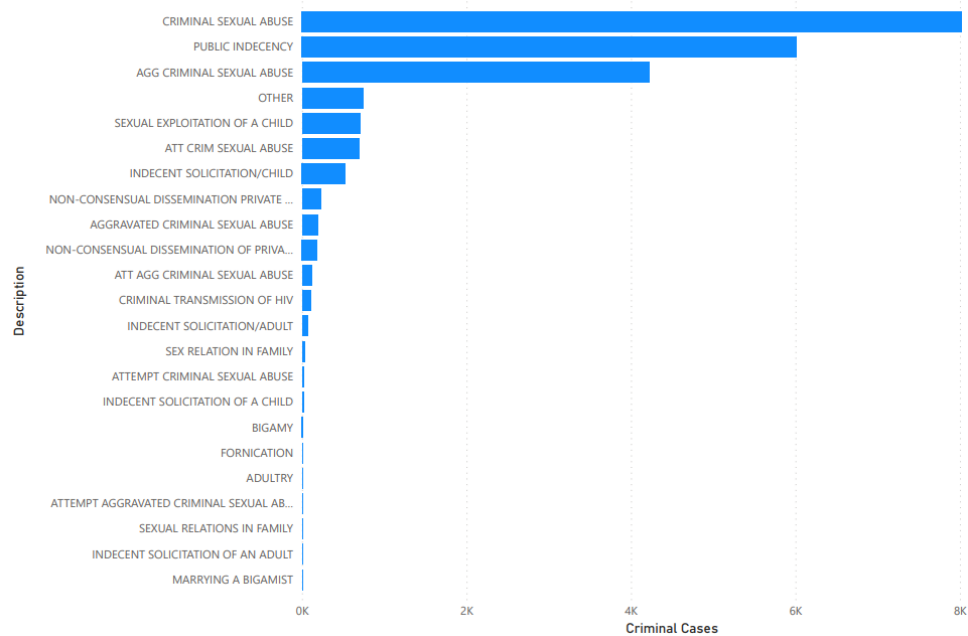
Το διάγραμμα δείχνει που λαμβάνουν χώρα Sexcrime Cases, δηλαδή αυτά που πρόκειται για Criminal Sexual Assault (CSA) και Sexual Offense. Το CSA είναι ένας συγκεκριμένος όρος που αναφέρεται συνήθως στον πιο σοβαρό τύπο σεξουαλικής επίθεσης, που συνήθως περιλαμβάνει μη συναινετική σεξουαλική επαφή. Το sexual offense περιλαμβάνει μια πιο ευρεία κατηγορία εγκλημάτων, μερικές από τις οποίες είναι λιγότερο ποινικά κολάσιμες από το CSA. Βλέπουμε πως σχεδόν οι μισές υποθέσεις λαμβάνουν χώρα μέσα σε κατοικία, και μετά στον δρόμο, γεγονός που εντείνει την ανασφάλεια.

Sexcrime Cases per Year



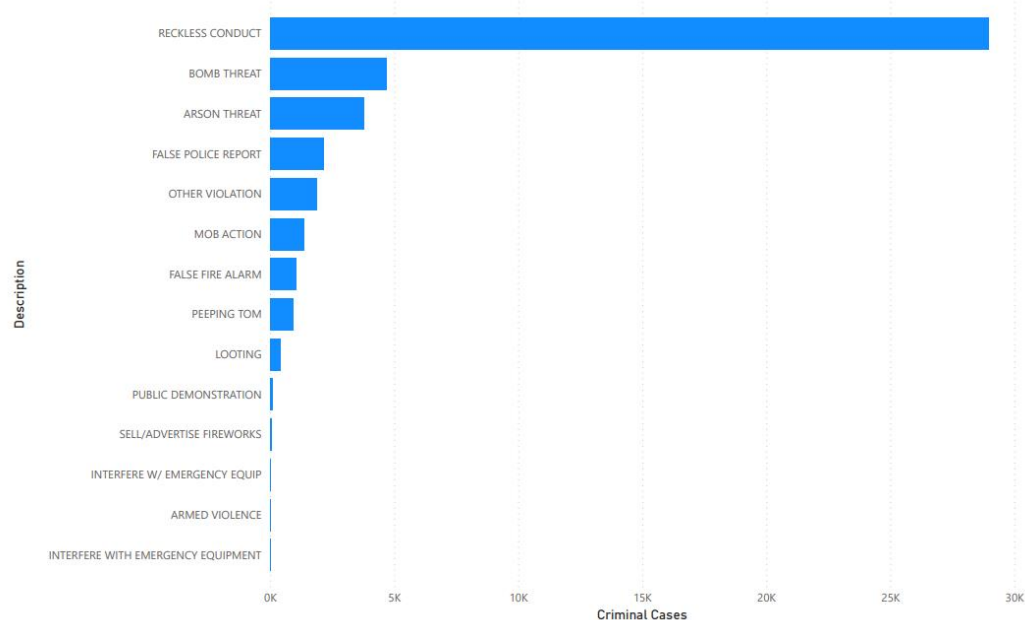
Το διάγραμμα μαρτυρά πως η ανθρωπότητα δεν τα πηγαίνει καθόλου καλά με το Sexcrime rate, αφού από το 2015, έως το οποίο παρουσίαζε γενική πτώση, μέχρι και το 2020 ακολουθεί ανοδική πορεία, που βέβαια ανατράπηκε ενδεχομένως λόγω της έξαρσης του COVID-19.

Sex Offense Cases by Description



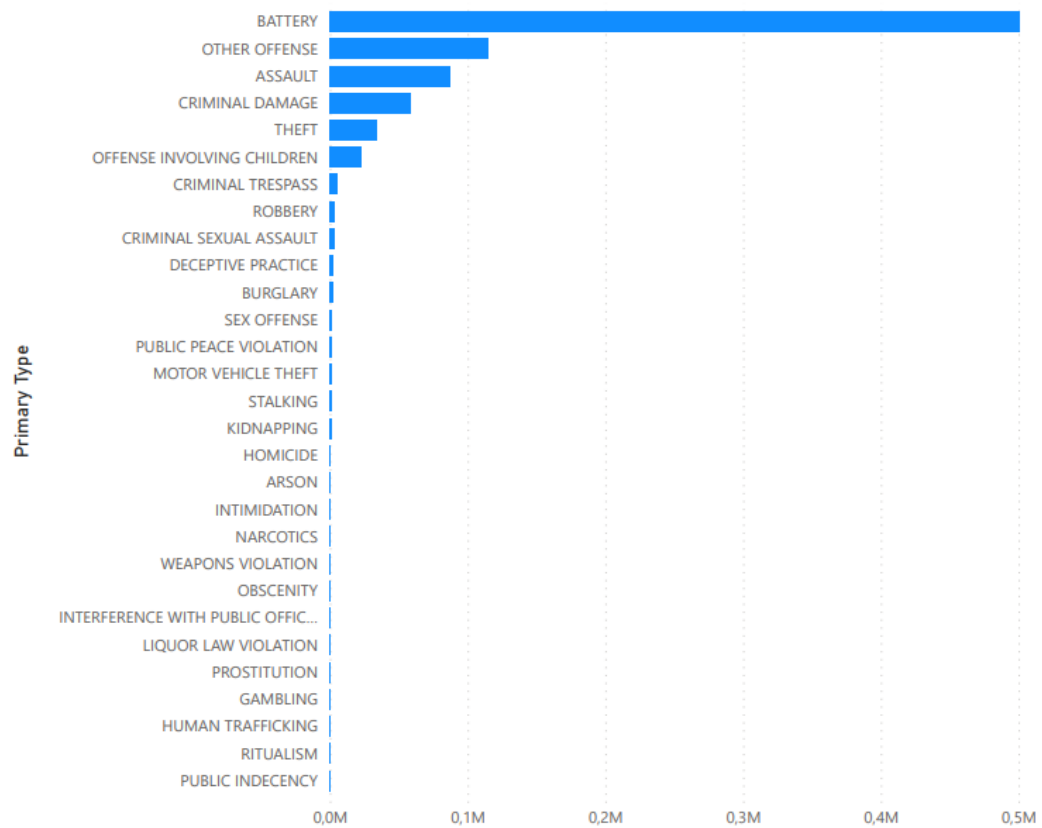
Στο διάγραμμα βλέπουμε τις περιγραφές των υποθέσεων που αφορούν αδικήματα σεξουαλικής προσβολής. Δυστυχώς δεν είναι λίγες οι περιγραφές που σχετίζονται και με παιδιά.

Public Peace Violation Cases by Description



Στο διάγραμμα βλέπουμε τις περιγραφές των υποθέσεων που αφορούν αδικήματα κατά της δημόσιας ειρήνης. Στη δεύτερη θέση βρίσκεται το σοβαρότατο αδίκημα βομβιστικής απειλής με 4.68K περιστατικά μέσα σε 18 χρόνια.

Domestic related Cases by Primary Type



Στο διάγραμμα εμφανίζονται οι τύποι εγκλήματος που είναι περισσότερο συχνοί στα domestic related cases (μεταξύ γνωστών/συγγενικών ατόμων στην ουσία). Αυτοί είναι η βιαιοπραγία, η προσβολή, η φθορά ξένης περιουσίας, η κλοπή, αλλά και εγκλήματα με παιδιά.

## 5. Εξόρυξη Δεδομένων

### 5.1 Clustering

Για το clustering χρησιμοποιήσαμε τον αλγόριθμο K-Means. Ομαδοποιήσαμε τα δεδομένα σύμφωνα με το District, το Ward και το IUCR. Πιστεύουμε ότι αυτό θα μας βοηθήσει να προσδιορίσουμε ποια τμήματα της πόλης αντιμετωπίζουν επιθέσεις και ποιου τύπου.

Αρχικά κρατάμε από το dataset μόνο τις στήλες: Ward, IUCR και District. Πιο συγκεκριμένα, η στήλη IUCR περιέχει και κάποια γράμματα. Αυτό που κάνουμε είναι να αφαιρέσουμε αυτά τα γράμματα από όλες πεδία περιέχουν γράμμα και να μετατρέψουμε τη στήλη IUCR σε int.

```
In [4]: sub_data = crimes[['Ward', 'IUCR', 'District']]
sub_data = sub_data.apply(lambda x:x.fillna(x.value_counts().index[0]))
sub_data['IUCR'] = sub_data.IUCR.str.extract('(\d+)', expand=True).astype(int)
sub_data.head()
```

Επομένως, δημιουργούμε τον παρακάτω πίνακα (βλέπουμε μόνο τα πρώτα 10 στοιχεία του).

Out[6]:

	Ward	IUCR	District
0	12	486	9
1	29	870	15
2	35	2023	14
3	28	560	15
4	21	610	6
5	32	620	14
6	25	860	10
7	27	320	12
8	13	820	8
9	45	460	16

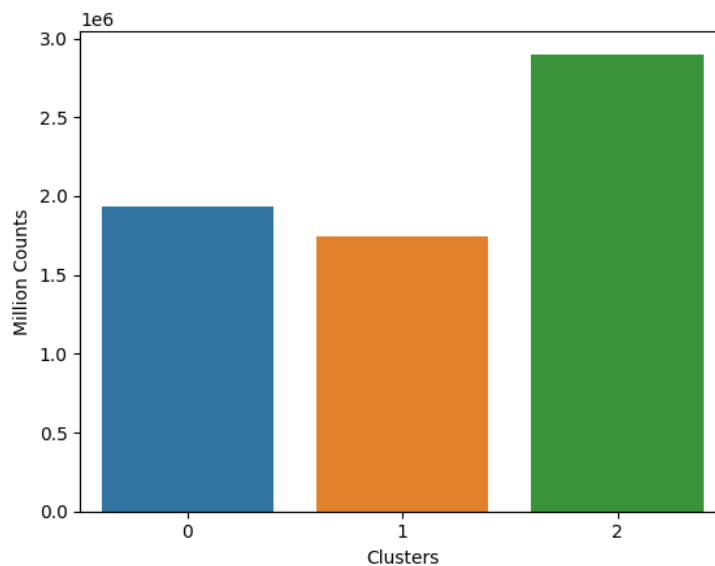
Αν προσπαθήσουμε να τρέξουμε τον K-Means σε αυτά τα δεδομένα, θα ομαδοποιηθούν τα δεδομένα σύμφωνα με τις ευκλείδειες αποστάσεις των κωδικών IUCR. Αυτό που κάνουμε λοιπόν είναι να κανονικοποιήσουμε τα δεδομένα.

Η κανονικοποίηση είναι μια τεχνική που χρησιμοποιείται για την κλίμακα των δεδομένων έτσι ώστε να εμπίπτουν σε ένα συγκεκριμένο εύρος, συνήθως μεταξύ 0 και 1. Αυτό γίνεται αφαιρώντας την ελάχιστη τιμή από κάθε στοιχείο της στήλης και στη συνέχεια διαιρώντας το αποτέλεσμα με το εύρος (η διαφορά μεταξύ της μέγιστης και της ελάχιστης τιμής).

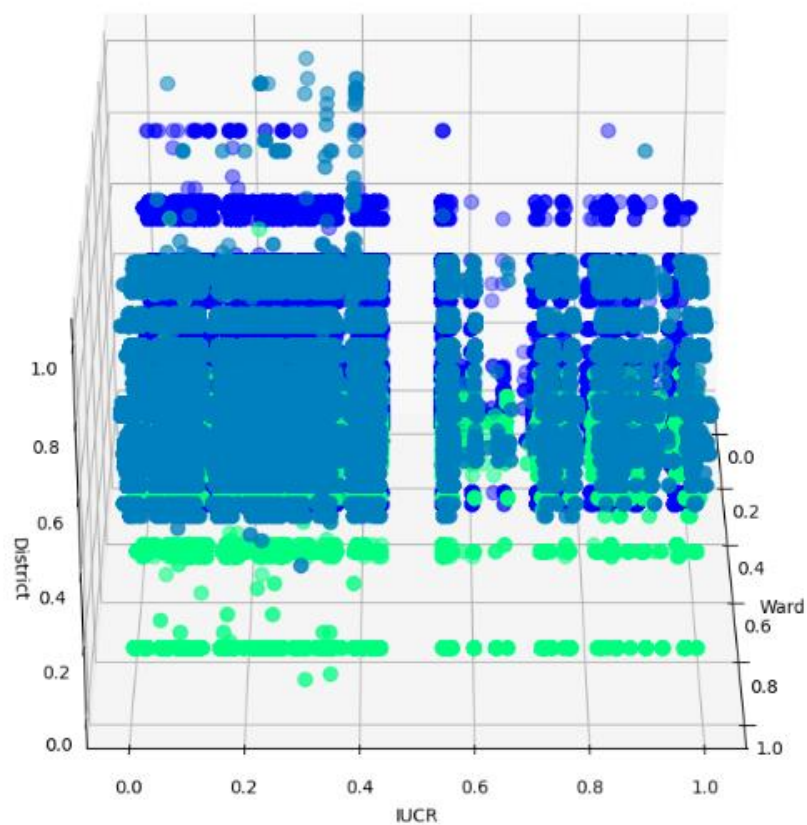
Ακολουθεί ενδεικτικό κομμάτι κώδικα:

```
In [7]: sub_data['IUCR'] = (sub_data['IUCR'] - sub_data['IUCR'].min())/(sub_data['IUCR'].max()-sub_data['IUCR'].min())
sub_data['Ward'] = (sub_data['Ward'] - sub_data['Ward'].min())/(sub_data['Ward'].max()-sub_data['Ward'].min())
sub_data['District'] = (sub_data['District'] - sub_data['District'].min())/(sub_data['District'].max()-sub_data['District'].min())
```

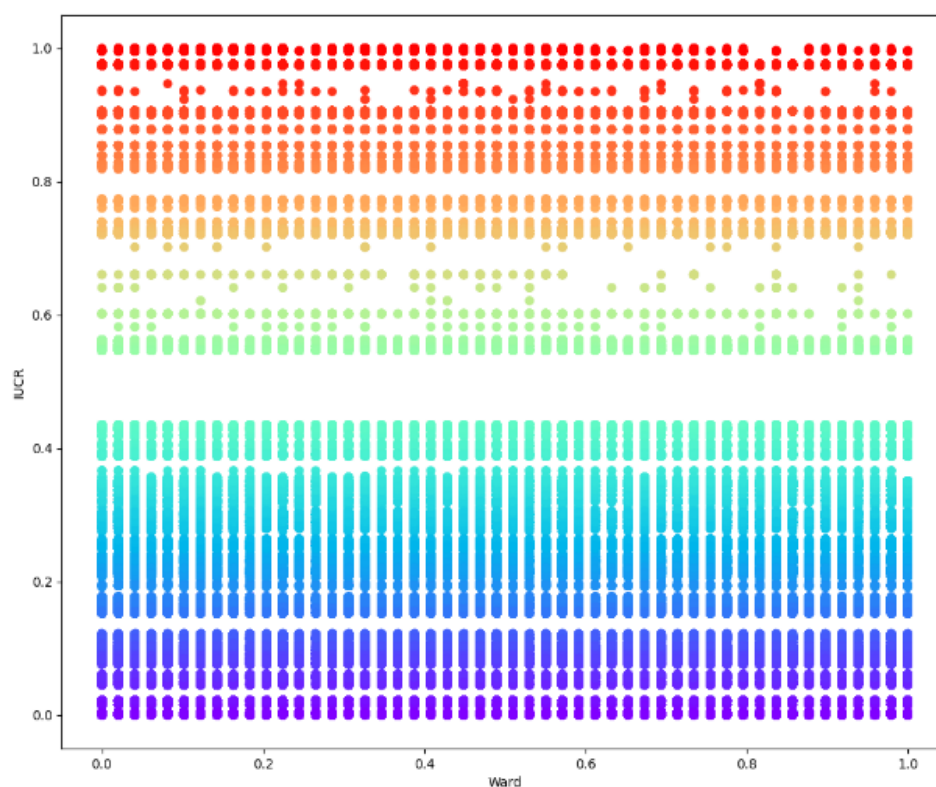
Τώρα που έχουμε τα δεδομένα στη μορφή που θέλουμε τρέχουμε τον αλγόριθμο K-Means. Έπειτα από αρκετές δοκιμές καταλήξαμε στο ότι ο καλύτερος αριθμός clusters είναι 3. Παρακάτω φαίνεται πόσες μετρήσεις έχει το κάθε cluster σε εκατομμύρια μονάδες.



Το ακόλουθο scatterplot οπτικοποιεί τις στήλες Ward, District και IUCR και χρωματίζει κάθε σημείο σύμφωνα με την ετικέτα συμπλέγματός του που λαμβάνεται από τον αλγόριθμο K-Means. Αυτή η γραφική παράσταση μπορεί να βοηθήσει στον εντοπισμό εάν υπάρχουν μοτίβα ή τάσεις στα δεδομένα, για παράδειγμα εάν τα σημεία σε ένα συγκεκριμένο cluster τείνουν να έχουν παρόμοιες τιμές για το Ward, το District και το IUCR. Βλέπουμε για παράδειγμα ότι δημιουργούνται οριζόντιες και κάθετες γραμμές. Οι οριζόντιες γραμμές δείχνουν ότι για μία συγκεκριμένη τιμή ή περιοχή τιμών του district ποια εγκλήματα (IUCR) εμφανίζονται. Οι κάθετες γραμμές δείχνουν ότι για ένα συγκεκριμένο τύπο εγκλήματος, σε ποιες περιοχές εμφανίζεται. Το ίδιο ισχύει και για τη διάσταση Ward.



Το ακόλουθο διάγραμμα έχει ακριβώς τον ίδιο σκοπό με το προηγούμενο. Η οπτικοποίηση γίνεται μεταξύ IUCR και Ward.



## 5.2 Συσχετίσεις εγκλημάτων

Σκεφτήκαμε ότι θα ήταν ωραίο να γνωρίζουμε ποια εγκλήματα εμφανίζονται μαζί στο Σικάγο. Δηλαδή, μία πιθανή συσχέτιση θα μπορούσε να είναι πως σε περιοχές με υψηλά ποσοστά κλοπών υπάρχουν και υψηλά ποσοστά διαρρήξεων.

Για να μπορέσουμε να βρούμε τέτοιου είδους συσχέτισεις εφαρμόσαμε τον apriori algorithm. Έτσι δημιουργήσαμε έναν πίνακα με index το community area και στήλες όλα τα είδη των εγκλημάτων.

	ARSON	ASSAULT	BATTERY	BURGLARY	CARRY LICENSE VIOLATION	CRIMINAL DAMAGE	CRIMINAL SEXUAL ASSAULT	CRIMINAL TRESPASS	DECEPTIVE PRACTICE	GAMBLING	...	OTHER OFFENSE	PROSTITUTION
Community Area													
0	0	0	0	0	0	0	0	0	0	0	...	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
73	0	0	0	0	0	0	0	0	0	0	...	0	0
74	0	0	0	0	0	0	0	0	0	0	...	0	0
75	0	0	0	0	0	0	0	0	0	0	...	0	0
76	0	0	0	0	0	0	0	0	0	0	...	0	0
77	0	0	0	0	0	0	0	0	0	0	...	0	0

78 rows x 31 columns

Για κάθε Community Area (κάθε γραμμή του πίνακα) υπάρχει σε κάθε στήλη ο αριθμός 0 αν το συγκεκριμένο έγκλημα δεν έχει συμβεί σε αυτή την περιοχή και 1 αν έχει συμβεί.

Με βάση αυτόν τον πίνακα εκτελούμε τον apriori algorithm.

```
In [36]: # Apply the Apriori algorithm to find frequent itemsets
frequent_itemsets = apriori(dummies, min_support=0.01, use_colnames=True)
frequent_itemsets
```

Βλέπουμε στον πίνακα που ακολουθεί τα εγκλήματα με τον μεγαλύτερο αριθμό εμφάνισης. Η στήλη support δίπλα από τη στήλη itemsets μας δείχνει πόσο συχνό είναι το κάθε έγκλημα.

	support	itemsets
0	0.192308	(ASSAULT)
1	0.538462	(BATTERY)
2	0.141026	(BURGLARY)
3	0.410256	(CRIMINAL DAMAGE)
4	0.025641	(CRIMINAL TRESPASS)
...	...	...
1274	0.012821	(THEFT, CRIMINAL TRESPASS, MOTOR VEHICLE THEFT...
1275	0.012821	(THEFT, MOTOR VEHICLE THEFT, CRIMINAL DAMAGE, ...
1276	0.012821	(THEFT, MOTOR VEHICLE THEFT, CRIMINAL DAMAGE, ...
1277	0.012821	(THEFT, MOTOR VEHICLE THEFT, CRIMINAL DAMAGE, ...
1278	0.012821	(THEFT, MOTOR VEHICLE THEFT, CRIMINAL DAMAGE, ...

1279 rows x 2 columns



Έπειτα, προχωράμε στην εύρεση συσχετίσεων μεταξύ των εγκλημάτων. Όπως φαίνεται στον πίνακα παρακάτω, συχνά εμφανίζονται μαζί για παράδειγμα το ASSAULT και το BATTERY, το NARCOTICS με το ASSAULT κλπ.

```
In [30]: # Find the association rules
association_rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.5)
association_rules.head(50)
```

	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

Φαίνεται επίσης ότι οι συσχετίσεις στον παραπάνω πίνακα είναι αρκετά ισχυρές καθώς έχουν  $lift > 1$ . Όσο μεγαλύτερη είναι αυτή η τιμή τόσο πιο ισχυρή είναι και η συσχέτιση των δύο εγκλημάτων.

### 5.3 Δέντρο Απόφασης

```
from sklearn.preprocessing import LabelEncoder

# Transforming non-numeric values

le = LabelEncoder()
df['new'] = le.fit_transform(df['IUCR'])

# Split the dataset into features and target
X = df[['Community Area', 'new']]
y = df['Arrest']

# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# Train the decision tree model
clf = tree.DecisionTreeClassifier(random_state=42, max_depth=4)
clf = clf.fit(X_train, y_train)

# Make predictions on the test set
y_pred = clf.predict(X_test)

# Evaluate the model's accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy: ", accuracy)
```

Accuracy: 0.864379751896198

Ο σκοπός αυτού του μοντέλου δέντρου αποφάσεων είναι να προβλέψει εάν ένα περιστατικό οδηγεί σε σύλληψη ή όχι. Χρησιμοποιεί τα features του συνόλου δεδομένων «Community Area» και «Primary Type» για να κάνει προβλέψεις.

Ο παραπάνω κώδικας δημιουργεί αρχικά δύο ξεχωριστές μεταβλητές, X και y. Το X είναι ένα σύνολο χαρακτηριστικών που περιέχει τις στήλες 'Community Area' και 'Primary Type' και το y είναι η μεταβλητή στόχος 'Arrest'. Στη συνέχεια, το σύνολο δεδομένων χωρίζεται σε σύνολα εκπαίδευσης και δοκιμής χρησιμοποιώντας τη συνάρτηση train\_test\_split. Αυτό επιτρέπει στο μοντέλο να εκπαιδευτεί στο σετ εκπαίδευσης και στη συνέχεια να αξιολογήσει την απόδοσή του στο σετ δοκιμής.

Στη συνέχεια, ο DecisionTreeClassifier δημιουργείται με τις ακόλουθες παραμέτρους:

random\_state = 42: εξασφάλιση αναπαραγωγικότητας των αποτελεσμάτων

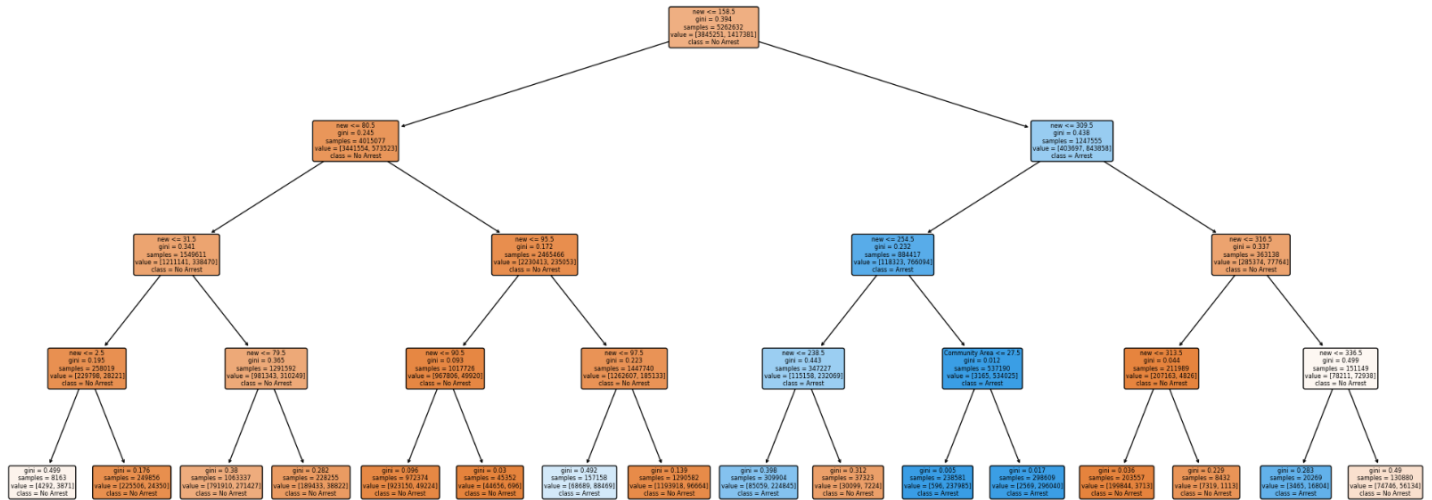
max\_depth=4: αυτό σημαίνει ότι το δέντρο θα κατασκευαστεί σε μέγιστο βάθος 4 επιπέδων

Το μοντέλο εκπαιδεύεται στο σετ προπόνησης (X\_train και y\_train) χρησιμοποιώντας τη μέθοδο fit. Έπειτα χρησιμοποιείται για να κάνει προβλέψεις στο σύνολο δοκιμών (X\_test) χρησιμοποιώντας τη μέθοδο πρόβλεψης.

Τέλος, η ακρίβεια του μοντέλου αξιολογείται συγκρίνοντας τις προβλέψεις (y\_pred) με τις πραγματικές τιμές της μεταβλητής στόχου στο σύνολο δοκιμής (y\_test) χρησιμοποιώντας τη συνάρτηση accuracy\_score. Η βαθμολογία ακρίβειας, που εκτυπώνεται στο τέλος, δείχνει

πόσο καλά έχει αποδώσει το μοντέλο, με άλλα λόγια πόσα από τα παραδείγματα δοκιμής είχαν προβλεφθεί σωστά από το μοντέλο. Συγκεκριμένα το μοντέλο δίνει ακρίβεια περίπου ίση με 0.86. Αυτό σημαίνει ότι το μοντέλο είναι σε θέση να προβλέψει σωστά εάν ένα περιστατικό οδηγεί σε σύλληψη ή όχι με ακρίβεια 86,47%. Με άλλα λόγια, εάν πάρουμε 100 τυχαία παραδείγματα από το σύνολο δοκιμών, κατά μέσο όρο το μοντέλο θα προβλέψει σωστά 86 από αυτά.

Παρακάτω βρίσκεται η οπτικοποίηση του μοντέλου μας



Το Gini impurity μετρά την πιθανότητα ότι εάν διαλέξουμε ένα αντικείμενο τυχαία αυτό θα ταξινομηθεί λανθασμένα. Κατά τη δημιουργία του δέντρου απόφασης, ο στόχος είναι να επιλέξουμε το split που μας δίνει το χαμηλότερο Gini impurity.