

**ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ**



ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΕΡΓΑΣΙΑ 7ου ΕΞΑΜΗΝΟΥ

**Επιχειρηματική Ευφυΐα και Ανάλυση
Μεγάλων Δεδομένων**

Υποβληθείσα στον Καθηγητή:
Χατζηαντωνίου Δαμιανό

Οι σπουδαστές:
Κόκοτα Ναταλία - ΑΜ: 8210060
Αιμιλία Δήμητρα Κτενά - ΑΜ: 8210073

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ.....	4
1. ΕΥΡΕΣΗ ΚΑΙ ΔΗΜΙΟΥΡΓΙΑ ΤΟΥ DATASET.....	4
1.1 Περιγραφή.....	4
1.2 Καθαρισμός και Επεξεργασία Δεδομένων.....	6
1.3 Εμπλουτισμός Δεδομένων με Κοινωνικοοικονομικούς Δείκτες.....	9
DATA WAREHOUSE – SQL SERVER.....	12
2.1 ΔΗΜΙΟΥΡΓΙΑ SQL SERVER	12
2.2 ΚΑΤΑΧΩΡΗΣΗ ΔΕΔΟΜΕΝΩΝ ΣΤΟΝ SQL SERVER	12
2.3 ΔΗΜΙΟΥΡΓΙΑ DIMENSIONS.....	13
2.3.1 Case Dimension.....	14
2.3.2 Location Dimension.....	15
2.3.3 Location Details Dimension	17
2.3.4 Date Dimension.....	18
2.3.5 Time Dimension	19
2.3.6 Arrest Dimension.....	21
2.3.7 Domestic Dimension	22
2.3.8 Socioeconomic Dimension	22
2.4 ΔΗΜΙΟΥΡΓΙΑ FACT TABLE	23
2.5 STAR SCHEMA	25
VISUAL STUDIO - ΔΗΜΙΟΥΡΓΙΑ KYBOY	26
3.1 ΔΗΜΙΟΥΡΓΙΑ PROJECT	26
3.2 ΔΗΜΙΟΥΡΓΙΑ DATA SOURCE ΚΑΙ DATA SOURCE VIEWS	26
3.3 ΔΗΜΙΟΥΡΓΙΑ ΤΟΥ CUBE.....	28
3.4 ΑΝΑΠΤΥΞΗ ΚΑΙ ΕΓΚΑΤΑΣΤΑΣΗ ΤΟΥ CUBE	28
ΟΠΤΙΚΟΠΟΙΗΣΗ ΜΕ POWER BI.....	29
4.1 ΥΠΟΛΟΓΙΣΜΟΣ ΒΑΣΙΚΩΝ ΜΕΤΡΙΚΩΝ	29
4.2 VISUALIZATIONS	30
4.2.1 Crimes Count by Hour	30
4.2.2 Monthly Crime Counts for 2011 and 2012	30
4.2.3 Monthly Homicide Crime Counts.....	31
4.2.4 Distribution of Crime Counts by Primary Crime Type.....	32
4.2.5 Domestic vs. Non-Domestic Crimes Distribution.....	33
4.2.6 Domestic vs. Non-Domestic Crimes by Type	34
4.2.7 Narcotics Crimes Count by Description	35
4.2.8 Crime Hotspots in Chicago	36
4.2.9 Socioeconomic Indicators vs. Crime Count	37
DATA MINING	38
5.1 CLUSTERING	38
5.1.1 Anomaly detection (Outlier analysis)	40
5.2 ASSOCIATION RULES	42

5.3 DECISION TREE	45
-------------------------	----

Εισαγωγή

Στην παρούσα εργασία, εξετάζουμε τη διαχείριση ενός μεγάλου dataset που περιλαμβάνει στοιχεία εγκλημάτων στην πόλη του Σικάγο, ΗΠΑ. Η διαδικασία περιλαμβάνει τον καθαρισμό των δεδομένων, την εισαγωγή τους σε μια αποθήκη δεδομένων και τη δημιουργία ενός κύβου δεδομένων για την εξαγωγή σχετικών μετρικών. Στη συνέχεια, αξιοποιούμε ένα εργαλείο οπτικοποίησης δεδομένων (Power BI) για την κατασκευή γραφημάτων. Τέλος μέσω βιβλιοθηκών της Python, τα δεδομένα της αποθήκης θα χρησιμοποιηθούν σε λειτουργίες εξόρυξης δεδομένων, με εφαρμογή τεχνικών όπως clustering, δέντρα αποφάσεων και συσχετίσεις.

Το dataset που επιλέξαμε αφορά την περίοδο από το 2001 έως το 2024 και περιλαμβάνει περισσότερες από 7 εκατομμύρια εγγραφές εγκλημάτων. Το συγκεκριμένο dataset επιλέχθηκε λόγω της πολυπλοκότητάς του και των πολλαπλών δυνατοτήτων ανάλυσης που προσφέρει.

Η εργασία αυτή στοχεύει όχι μόνο στη διαχείριση δεδομένων μεγάλης κλίμακας αλλά και στη χρήση σύγχρονων εργαλείων και τεχνικών ανάλυσης που ενισχύουν τη λήψη αποφάσεων και προσφέρουν χρήσιμα συμπεράσματα για την κοινωνία.

1. Εύρεση και Δημιουργία του Dataset

1.1 Περιγραφή

1.1.1 Πλαίσιο

Το dataset που χρησιμοποιούμε αφορά ποινικές υποθέσεις που έλαβαν χώρα στην πόλη του Σικάγο από την 1η Ιανουαρίου 2001 έως και τις 15 Ιουλίου 2023. Τα δεδομένα εξάγονται από το σύστημα CLEAR (Citizen Law Enforcement Analysis and Reporting) του Αστυνομικού Τμήματος του Σικάγο. Για την προστασία του απορρήτου των θυμάτων εγκλημάτων, οι διευθύνσεις εμφανίζονται μόνο σε επίπεδο μπλοκ και δεν προσδιορίζονται συγκεκριμένες τοποθεσίες. Επιπλέον, περιλαμβάνονται προκαταρκτικές ταξινομήσεις εγκλημάτων, οι οποίες μπορεί να τροποποιηθούν μεταγενέστερα βάσει πρόσθετης έρευνας. Είναι επίσης πιθανό να υπάρχουν μηχανικά ή ανθρώπινα λάθη. Για αυτόν τον λόγο, το Αστυνομικό Τμήμα του Σικάγο δεν εγγυάται την ακρίβεια, την πληρότητα ή την επικαιρότητα των δεδομένων.

Για την παρούσα εργασία, αποφασίσαμε να επικεντρωθούμε στα δεδομένα της περιόδου **2011-2012**. Η απόφαση αυτή βασίστηκε στην ανάγκη να περιορίσουμε το μέγεθος του dataset, κάνοντάς το πιο διαχειρίσιμο και αποτελεσματικό για την ανάλυση. Η συγκεκριμένη χρονική περίοδος επιλέχθηκε, καθώς εμπίπτει στο διάστημα 2008-2012, το οποίο καλύπτει τους κοινωνικοοικονομικούς δείκτες του Σικάγο του δεύτερου dataset που θα χρησιμοποιήσουμε, του οποίου τα δεδομένα αφορούν την χρονική περίοδο αυτή. Με αυτόν τον τρόπο, διασφαλίζουμε την αντιστοιχία των δεδομένων, ενώ ταυτόχρονα διατηρούμε ένα πιο περιορισμένο και εύκολα επεξεργάσιμο σύνολο δεδομένων, το οποίο περιλαμβάνει περίπου 680.000 εγγραφές.

1.1.2 Περιεχόμενο

Το αρχικό dataset περιλάμβανε περίπου 7.2 εκατομμύρια εγγραφές και 22 στήλες. Οι στήλες αυτές περιέγραφαν τα εξής για κάθε μοναδικό έγκλημα:

ID: Μοναδικό αναγνωριστικό για την εγγραφή.

Case Number: Ο αριθμός RD της Αστυνομίας του Σικάγο (Records Division Number), μοναδικός για κάθε περιστατικό.

Date: Ημερομηνία που συνέβη το περιστατικό.

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 Uniform, που 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 (αρχικά ως True/False).

Domestic: Υποδεικνύει εάν το περιστατικό σχετίζεται με ενδοοικογενειακή βία ή όχι (αρχικά ως True/False).

Beat: Υποδεικνύει το beat όπου συνέβη το περιστατικό, τη 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 κοινωνική περιοχή όπου συνέβη το περιστατικό. Το Σικάγο έχει 77 κοινωνικές 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 στην προβολή State Plane Illinois East NAD 1983.

Y Coordinate: Η συντεταγμένη y της τοποθεσίας στην προβολή State Plane Illinois East NAD 1983.

Year: Η χρονιά που συνέβη το περιστατικό.

Updated On: Η ημερομηνία και ώρα τελευταίας ενημέρωσης της εγγραφής.

Latitude: Το γεωγραφικό πλάτος της τοποθεσίας.

Longitude: Το γεωγραφικό μήκος της τοποθεσίας.

Location: Ο συνδυασμός γεωγραφικού μήκους και πλάτους.

1.1.3 Πηγή

Το dataset προέρχεται από την πλατφόρμα Kaggle και είναι διαθέσιμο στον παρακάτω σύνδεσμο:

[Chicago Crime Dataset 2001-Present](#)

1.2 Καθαρισμός και Επεξεργασία Δεδομένων

1.2.1 Διαγραφή Στηλών

Αρχικά, αφαιρέσαμε τις στήλες **Year** και **Updated On**, καθώς περιέχουν πληροφορίες που είναι ήδη διαθέσιμες στη στήλη **Date** και δεν προσφέρουν επιπλέον χρήσιμα δεδομένα για την ανάλυση μας.

```
data.drop(columns=['Year'], inplace=True)
data.drop(columns=['Updated On'], inplace=True)
```

Python

1.2.2 Γεωγραφικό Μήκος και Πλάτος

Για να διασφαλίσουμε ότι όλα τα δεδομένα αφορούν περιοχές εντός της πόλης του Σικάγου, φιλτράραμε τις εγγραφές χρησιμοποιώντας τα ακόλουθα εύρη τιμών:

- **Γεωγραφικό μήκος (Longitude):** [-87.9401, -87.5237]
- **Γεωγραφικό πλάτος (Latitude):** [41.6445, 42.0230]

Κρατήσαμε μόνο τις εγγραφές όπου οι τιμές στις στήλες **Longitude** και **Latitude** εμπίπτουν σε αυτά τα διαστήματα.

```
valid_longitude = (-87.9401, -87.5237)
valid_latitude = (41.6445, 42.0230)

data = data[(data['Longitude'] >= valid_longitude[0]) &
            (data['Longitude'] <= valid_longitude[1]) &
            (data['Latitude'] >= valid_latitude[0]) &
            (data['Latitude'] <= valid_latitude[1])]
```

Python

1.2.3 Παρόμοιες Τιμές

Παρατηρήσαμε ότι στη στήλη **Primary Type**, ορισμένες τιμές, όπως **NON - CRIMINAL** και **NON-CRIMINAL**, είναι ουσιαστικά ίδιες αλλά εμφανίζονται διαφορετικά λόγω διαστημάτων ή

γραμματοσειράς. Χρησιμοποιώντας κατάλληλη κανονικοποίηση, αφαιρέσαμε κενά πριν και μετά την παύλα και ενοποιήσαμε τις τιμές.

Επιπλέον, τροποποιήσαμε μερικές τιμές για λόγους ομοιομορφίας, όπως:

CRIM SEXUAL ASSAULT → CRIMINAL SEXUAL ASSAULT

```
[12] data['Primary Type'] = data['Primary Type'].replace({
    'NON - CRIMINAL': 'NON-CRIMINAL',
    'CRIM SEXUAL ASSAULT': 'CRIMINAL SEXUAL ASSAULT'
})
```

Python

1.2.4 Διαγραφή NON-CRIMINAL

Οι εγγραφές που κατατάσσονται ως NON-CRIMINAL διαγράφηκαν, καθώς δεν αντιπροσωπεύουν εγκληματικές πράξεις και δεν είναι σχετικές με την ανάλυσή μας.

```
[14] data = data[data['Primary Type'] != 'NON-CRIMINAL']
```

Python

1.2.5 Διαχωρισμός Ημερομηνίας και Ώρας

Η στήλη **Date**, η οποία περιέχει ημερομηνία και ώρα, χωρίστηκε σε δύο νέες στήλες:

- **Date:** Περιέχει μόνο την ημερομηνία.
- **Time:** Περιέχει μόνο την ώρα, σε μορφή 24 ωρών (π.χ. 13:30:00).

Μετά τον διαχωρισμό, η αρχική στήλη **Date** προσαρμόστηκε ανάλογα.

```
[16] data['Date'] = pd.to_datetime(data['Date'], format='%m/%d/%Y %I:%M:%S %p')
data['Time'] = data['Date'].dt.time
data['Date'] = data['Date'].dt.date
```

Python

1.2.6 Περιορισμός Εγγραφών στα Έτη 2011-2012

Όπως προαναφέρθηκε, για να μειώσουμε το μέγεθος του dataset, κρατήσαμε μόνο τις εγγραφές που αφορούν την περίοδο από 01-01-2011 έως 31-12-2012. Αυτό μας επιτρέπει να διαχειριστούμε τα δεδομένα πιο αποτελεσματικά.

```
[17] from datetime import date  
  
start_date = date(2011, 1, 1)  
end_date = date(2012, 12, 31)  
data = data[(data['Date'] >= start_date) & (data['Date'] <= end_date)]
```

Python

1.2.7 Αφαιρεση Κομμάτων

Αφαιρέσαμε όλα τα κόμματα από τις τιμές τύπου string, καθώς αυτά προκαλούν προβλήματα κατά την εισαγωγή δεδομένων στη βάση δεδομένων από αρχεία CSV.

```
[19] data = data.replace(' ', '', regex=True)
```

Python

1.2.8 Μετατροπή Boolean σε Bit

Μετατρέψαμε τις τιμές στις στήλες **Arrest** και **Domestic** από **True/False** σε **0/1**:

- Στήλη **Arrest**:
 - **1**: Έγινε σύλληψη
 - **0**: Δεν έγινε σύλληψη
- Στήλη **Domestic**:
 - **1**: Το περιστατικό σχετίζεται με ενδοοικογενειακή βία
 - **0**: Το περιστατικό δεν σχετίζεται με ενδοοικογενειακή βία.

```
[21] data['Arrest'] = data['Arrest'].astype(int)  
data['Domestic'] = data['Domestic'].astype(int)
```

Python

1.2.9 Αντιμετώπιση Ελλιπών Τιμών

Για να διασφαλίσουμε τη συνοχή των δεδομένων:

- a. Συμπληρώσαμε την κενή τιμή στη στήλη **Location Description** με **UNKNOWN**.

```
[26] data['Location Description'].fillna('UNKNOWN', inplace=True)
```

Python

- b. Διαγράψαμε τις εγγραφές που περιείχαν κενές τιμές στις στήλες **District** και **Community Area**.

```
[27] data.dropna(subset=['District'], inplace=True)  
data.dropna(subset=['Community Area'], inplace=True)
```

Python

- c. Αντικαταστήσαμε την κενή τιμή στη στήλη **Ward** με την τιμή -1, καθώς ήταν μόνο μία.

```
[28] data['Ward'].fillna(-1, inplace=True)
```

Python

1.2.10 Εξαγωγή Καθαρισμένων Δεδομένων

Μετά τον καθαρισμό, αποθηκεύσαμε το καθαρισμένο dataset σε νέο αρχείο CSV (**cleaned_chicago_crime.csv**). Το νέο dataset περιλαμβάνει **684.175 εγγραφές**.

1.3 Εμπλουτισμός Δεδομένων με Κοινωνικοοικονομικούς Δείκτες

1.3.1 Λογική πίσω από τον συνδυασμό των δεδομένων

Ο εμπλουτισμός του κυρίου dataset των εγκλημάτων με κοινωνικοοικονομικούς δείκτες του Σικάγο επιλέχθηκε για να προσφέρει μια πιο ολοκληρωμένη ανάλυση των υπαρχόντων δεδομένων.

Συγκεκριμένα:

- Επιτρέπει την εξερεύνηση της σχέσης μεταξύ εγκληματικότητας και κοινωνικοοικονομικών παραμέτρων, όπως είναι το εισόδημα, το ποσοστό ανεργίας και η φτώχεια.
- Παρέχει τη δυνατότητα δημιουργίας clusters εγκληματικότητας που βασίζονται σε οικονομικούς δείκτες.

Τα δεδομένα κοινωνικοοικονομικών δεικτών προέρχονται από την πλατφόρμα δεδομένων του Σικάγο και καλύπτουν τις 77 κοινότητες της πόλης.

Πηγή:

Το dataset προέρχεται από την πλατφόρμα Kaggle και είναι διαθέσιμο στον παρακάτω σύνδεσμο: [Socioeconomic Indicators in Chicago](#)

1.3.2 Βήματα Καθαρισμού και Επεξεργασίας Συμπληρωματικού dataset

Προετοιμασία Δεδομένων

A. Μετονομασία Στηλών

Για τη διασφάλιση της συνοχής και της ευχρηστίας, οι στήλες του dataset κοινωνικοοικονομικών δεικτών μετονομάστηκαν ως εξής:

- **Community Area Number → Community Area**
- **PER CAPITA INCOME → Per Capita Income**
- **PERCENT HOUSEHOLDS BELOW POVERTY → Poverty Rate**
- **PERCENT AGED 16+ UNEMPLOYED → Unemployment Rate**
- **HARDSHIP INDEX → Hardship Index**

```
[5]    socioeconomic_data.rename(columns={  
        'Community Area Number': 'Community Area',  
        'PER CAPITA INCOME ': 'Per Capita Income',  
        'PERCENT HOUSEHOLDS BELOW POVERTY': 'Poverty Rate',  
        'PERCENT AGED 16+ UNEMPLOYED': 'Unemployment Rate',  
        'HARDSHIP INDEX': 'Hardship Index'  
    }, inplace=True)
```

Python

B. Αντιμετώπιση Μη Έγκυρης Τιμής (NaN):

Η στήλη **Community Area** περιείχε μια φαινομενικά μη έγκυρη τιμή (**NaN**), καθώς αφορούσε στο σύνολο της πολιτείας του Σικάγο (επιπλέον εγγραφή), η οποία αντικαταστάθηκε με την τιμή -1 για να είναι δυνατή η μετατροπή της στήλης σε ακέραιο αριθμό.

```
[6]    socioeconomic_data['Community Area'] = socioeconomic_data['Community Area'].fillna(-1).astype(int)  
    socioeconomic_data['Community Area'] = socioeconomic_data['Community Area'].astype(int)
```

Python

C. Επιλογή Σημαντικών Στηλών:

Από το dataset επιλέχθηκαν οι στήλες που προσφέρουν χρήσιμες πληροφορίες για την ανάλυση, συγκεκριμένα:

- **Community Area**
- **Per Capita Income**: Το κατά κεφαλήν εισόδημα ανά κοινότητα.
- **Poverty Rate**: Το ποσοστό νοικοκυριών κάτω από το όριο της φτώχειας.
- **Unemployment Rate**: Το ποσοστό ανεργίας.
- **Hardship Index**: Ένας σύνθετος δείκτης δυσκολιών που λαμβάνει υπόψη πολλαπλούς κοινωνικούς παράγοντες.

```
[7] socioeconomic_data = socioeconomic_data[['Community Area',
                                         'Per Capita Income',
                                         'Poverty Rate',
                                         'Unemployment Rate',
                                         'Hardship Index']]
```

Python

Ενοποίηση με το Dataset Εγκλημάτων

Το καθαρισμένο dataset κοινωνικοοικονομικών δεικτών συγχωνεύτηκε με το ήδη καθαρισμένο dataset των εγκλημάτων, χρησιμοποιώντας τη στήλη **Community Area** ως κοινό κλειδί. Η ενοποίηση πραγματοποιήθηκε με τη μέθοδο **left join**, ώστε να παραμείνουν όλες οι εγγραφές του dataset εγκλημάτων, ακόμα και αν δεν υπάρχουν διαθέσιμοι κοινωνικοοικονομικοί δείκτες για ορισμένες περιοχές.

1.3.3 Τελικό Dataset

Το τελικό dataset περιλαμβάνει όλες τις αρχικές πληροφορίες εγκλημάτων, εμπλουτισμένες με τους επιλεγμένους κοινωνικοοικονομικούς δείκτες και αποθηκεύτηκε ως **full_chicago_crime_dataset.csv**. Αποτελείται από **684.175 εγγραφές** και **25 στήλες**.

Data Warehouse – SQL Server

2.1 Δημιουργία SQL Server

Για την εισαγωγή των δεδομένων στη βάση, πρώτα εγκαταστήσαμε το SQL Server και δημιουργήσαμε τη βάση δεδομένων, την οποία ονομάσαμε chicagoCrimes. Στη συνέχεια, συνδεθήκαμε στον SQL Server και προχωρήσαμε στη δημιουργία του πίνακα crimes στην εν λόγω βάση. Ο πίνακας αυτός σχεδιάστηκε με τους κατάλληλους τύπους δεδομένων για κάθε στήλη, ώστε να μπορέσουμε να αποθηκεύσουμε σωστά τα δεδομένα του dataset μας. Με την ολοκλήρωση αυτής της διαδικασίας, ήμασταν έτοιμοι να εισάγουμε τα δεδομένα στον πίνακα χρησιμοποιώντας κατάλληλες εντολές για να εξασφαλίσουμε την ακεραιότητα των δεδομένων και την αποδοτική διαχείριση τους.

```
CREATE TABLE crimes (
    ID INT PRIMARY KEY,
    CaseNumber NVARCHAR(50),
    Date DATE,
    Time TIME,
    Block NVARCHAR(255),
    IUCR NVARCHAR(10),
    PrimaryType NVARCHAR(100),
    Description NVARCHAR(255),
    LocationDescription NVARCHAR(255),
    Arrest BIT,
    Domestic BIT,
    Beat INT,
    District INT,
    Ward INT,
    CommunityArea INT,
    FBI_Code NVARCHAR(10),
    X_Coordinate FLOAT,
    Y_Coordinate FLOAT,
    Latitude FLOAT,
    Longitude FLOAT,
    Location NVARCHAR(255),
    PerCapitaIncome INT,
    PovertyRate FLOAT,
    UnemploymentRate FLOAT,
    HardshipIndex FLOAT
);
```

2.2 Καταχώρηση Δεδομένων στον SQL Server

Για την εισαγωγή των δεδομένων στον πίνακα crimes, χρησιμοποιήσαμε την εντολή **BULK INSERT**. Αυτή η εντολή είναι ιδιαίτερα χρήσιμη όταν έχουμε μεγάλο όγκο δεδομένων, όπως το dataset μας με 684.175 γραμμές.

Πιο συγκεκριμένα, χρησιμοποιήσαμε τις εξής παραμέτρους στην εντολή:

- **FIRSTROW:** Ορίζει την πρώτη γραμμή του αρχείου CSV (στην περίπτωση μας η πρώτη γραμμή δεδομένων είναι η σειρά 2, καθώς η πρώτη γραμμή του αρχείου περιέχει επικεφαλίδες των στηλών).
- **FIELDTERMINATOR:** Ορίζει τον χαρακτήρα που διαχωρίζει τα πεδία του αρχείου (**κόμμα**).
- **ROWTERMINATOR:** Ορίζει τον χαρακτήρα που δηλώνει το τέλος κάθε σειράς δεδομένων (**0x0a**, που αντιστοιχεί στον χαρακτήρα αλλαγής γραμμής).
- **FORMAT:** Ορίζει τη μορφή των δεδομένων στο αρχείο.
- **TABLOCK:** Βελτιστοποιεί την εισαγωγή δεδομένων για μεγάλες ποσότητες.
- **MAXERRORS:** Ορίζει τον μέγιστο αριθμό σφαλμάτων που επιτρέπονται πριν από την αποτυχία της διαδικασίας (σε αυτή την περίπτωση δεν επιτρέπονται σφάλματα).

Η εντολή για την εισαγωγή δεδομένων είναι η εξής:

```
BULK INSERT crimes
FROM 'C:\Program Files\BD\final_full_chicago_crime_dataset.csv'
WITH (
    FIRSTROW = 2,
    FIELDTERMINATOR = ',',
    ROWTERMINATOR = '0x0a',
    FIELDQUOTE = """",
    FORMAT = 'CSV',
    TABLOCK,
    MAXERRORS = 0
);
```

Έχουμε δημιουργήσει, λοιπόν τον πίνακα **crimes** όπως φαίνεται στην παρακάτω εικόνα:

ID	CaseNumber	Date	Block	IUCR	PrimaryType	Description	LocationDescription	Arest	Domestic	Beat	District	Ward	CommunityArea	FBICode	X_Coordinate	Y_Coordinate	Latitude	Longitude	Location	
1	19729	HS584556	2011-01-01	067XX S HALSTED ST	0110	HOMICIDE	FIRST DEGREE MURDER	APARTMENT	1	0	2022	20	48	77	01A	1172129	1860226	41.771907116	-87.644581938	(41.771907116,-87.644581938)
2	19730	HT100325	2011-01-01	011XX W HOLLYWOOD AVE	0110	HOMICIDE	FIRST DEGREE MURDER	APARTMENT	1	1	2022	20	48	77	01A	1167659	1930019	41.95473908	-87.6596893	(41.95473908,-87.6596893)
3	19736	HT101536	2011-01-02	001XX W 75TH ST	0110	HOMICIDE	FIRST DEGREE MURDER	CLUB	1	0	623	6	6	69	01A	1176502	1855256	41.758168911	-87.628701252	(41.758168911,-87.628701252)
4	19738	HT104380	2011-01-04	013XX W HASTINGS ST	0110	HOMICIDE	FIRST DEGREE MURDER	STREET	1	0	1231	12	2	28	01A	1167623	1893850	41.85472772	-87.660133858	(41.85472772,-87.660133858)
5	19739	HT104792	2011-01-04	045XX N MAGNOLIA AVE	0110	HOMICIDE	FIRST DEGREE MURDER	PARKING LOT	1	0	2311	19	46	3	01A	1167100	1930371	41.95449939	-87.661002797	(41.95449939,-87.661002797)
6	19741	HT108310	2011-01-05	080XX S CONSTANCE AVE	0110	HOMICIDE	FIRST DEGREE MURDER	STREET	0	0	412	4	8	45	01A	1190024	1848005	41.737959537	-87.57937793	(41.737959537,-87.57937793)
7	19742	HT111240	2011-01-08	080XX S PRINCETON AVE	0110	HOMICIDE	FIRST DEGREE MURDER	AUTO	1	0	623	6	17	44	01A	1175662	1851654	41.748305725	-87.631614065	(41.748305725,-87.631614065)
8	19743	HT111412	2011-01-08	040XX S MICHIGAN AVE	0110	HOMICIDE	FIRST DEGREE MURDER	HALLWAY	1	1	221	2	3	38	01A	1177882	1876231	41.815938011	-87.62000861	(41.815938011,-87.62000861)
9	19744	HT111676	2011-01-09	025XX W 51ST ST	0110	HOMICIDE	FIRST DEGREE MURDER	ALLEY	0	0	911	9	14	63	01A	1160431	1870775	41.801104078	-87.687172659	(41.801104078,-87.687172659)
10	19746	HT114232	2011-01-10	041XX W WILCOX ST	0110	HOMICIDE	FIRST DEGREE MURDER	APARTMENT	1	0	1115	11	28	26	01A	1148800	1899014	41.878827643	-87.72999546	(41.878827643,-87.72999546)
11	19747	HT112275	2011-01-11	059XX S ASHLAND AVE	0110	HOMICIDE	FIRST DEGREE MURDER	AUTO	0	0	2221	22	21	73	01A	1167292	1843448	41.725707061	-87.66279215	(41.725707061,-87.66279215)
12	19747	HT116872	2011-01-12	066XX E 3RD ST	0110	HOMICIDE	FIRST DEGREE MURDER	STREET	0	0	313	3	20	42	01A	1181413	1863372	41.780330981	-87.61045304	(41.780330981,-87.61045304)
13	19748	HT116874	2011-01-12	026XX W 15TH ST	0110	HOMICIDE	FIRST DEGREE MURDER	AUTO	0	0	1023	10	28	29	01A	1159146	1892630	41.861103162	-87.691286207	(41.861103162,-87.691286207)
14	19749	HT119996	2011-01-15	057XX W CHICAGO AVE	0110	HOMICIDE	FIRST DEGREE MURDER	STREET	0	0	1511	15	29	25	01A	1137681	1904755	41.894789245	-87.76978852	(41.894789245,-87.76978852)
15	19750	HT116513	2011-01-15	057XX S MORGAN ST	0110	HOMICIDE	FIRST DEGREE MURDER	HOUSE	0	0	733	7	17	68	01A	1170861	1858122	41.766161253	-87.6495291358	(41.766161253,-87.6495291358)
16	19751	HT122016	2011-01-16	026XX S TRIPP AVE	0110	HOMICIDE	FIRST DEGREE MURDER	STREET	0	0	1031	10	22	30	01A	1148478	1886193	41.843651461	-87.730612367	(41.843651461,-87.730612367)
17	19752	HT123587	2011-01-17	070XX W 63RD ST	0110	HOMICIDE	FIRST DEGREE MURDER	STREET	1	0	724	7	16	68	01A	1170297	1863035	41.779655416	-87.651251789	(41.779655416,-87.651251789)
18	19753	HT123735	2011-01-17	008XX W 119TH ST	0110	HOMICIDE	FIRST DEGREE MURDER	STREET	0	0	524	5	34	53	01A	1172848	1825938	41.677800201	-87.64295433	(41.677800201,-87.64295433)
19	19760	HT133119	2011-01-24	113XX S LAFLIN ST	0110	HOMICIDE	FIRST DEGREE MURDER	HOUSE	0	0	2234	22	34	75	01A	1168149	1829279	41.687070509	-87.660058693	(41.687070509,-87.660058693)

2.3 Δημιουργία Dimensions

Η δημιουργία των Dimensions είναι μια κρίσιμη φάση στην επεξεργασία δεδομένων για την ανάλυση. Στην παρούσα εργασία, για να αναπαραστήσουμε αποτελεσματικά τα δεδομένα, χωρίσαμε τον πίνακα **crimes** σε διαφορετικές κατηγορίες που έχουν νόημα και είναι χρήσιμες για τη μελλοντική ανάλυση. Κάθε dimension αναπαριστά μια οντότητα που έχει σημασία για τα δεδομένα μας και διευκολύνει τις σχέσεις στο **Fact Table**.

Ακολουθήσαμε μια διαδικασία δημιουργίας διαστάσεων, με σκοπό να κατανοήσουμε καλύτερα τις σχέσεις των δεδομένων, και να προσφέρουμε μια οργανωμένη βάση για τη δημιουργία του κύβου και την ανάλυση των δεδομένων μετέπειτα.

Τα dimensions που δημιουργήσαμε είναι τα εξής:

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

Ας αναλύσουμε κάθε dimension ξεχωριστά:

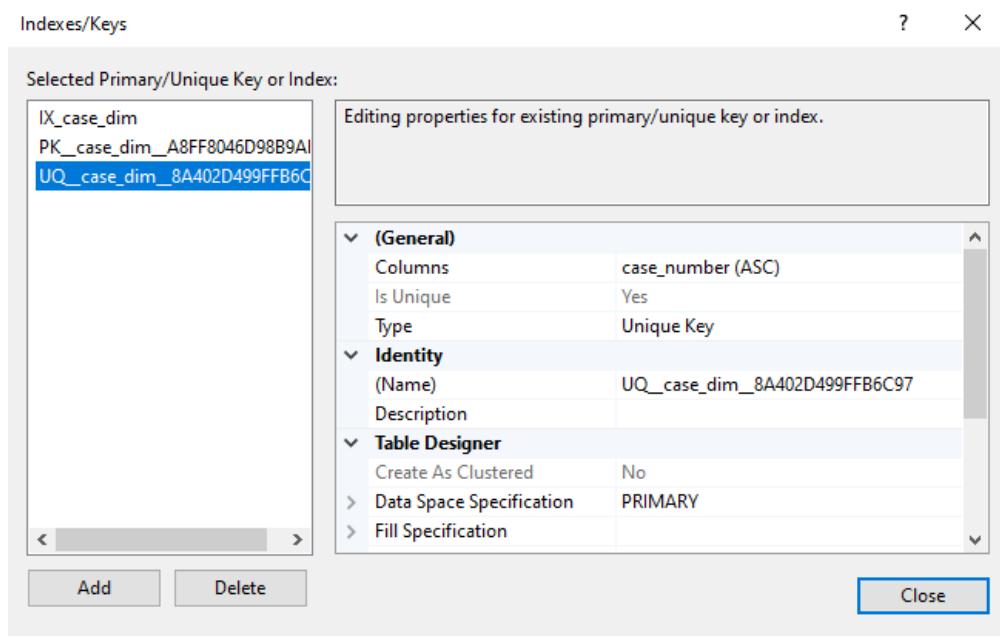
2.3.1 Case Dimension

Στο **Case Dimension**, συγκεντρώσαμε τα δεδομένα που σχετίζονται με το ίδιο το έγκλημα, τα οποία χρησιμοποιούνται για την κατηγοριοποίηση και την κατανόηση των διαφόρων τύπων εγκλημάτων που καταγράφονται. Για την δημιουργία αυτού του dimension, συγκεντρώθηκαν οι εξής στήλες από τον πίνακα crimes:

- **Case Number:** Ο αριθμός της υπόθεσης, ο οποίος είναι μοναδικός για κάθε περιστατικό.
- **IUCR:** Ο κωδικός του εγκλήματος σύμφωνα με το Illinois Uniform Crime Reporting Code.
- **Primary Type:** Ο κύριος τύπος του εγκλήματος, π.χ., κλοπή, επιθέσεις κ.ά.
- **Description:** Η δευτερεύουσα περιγραφή του εγκλήματος, η οποία παρέχει περισσότερες λεπτομέρειες για την υπόθεση.
- **FBI Code:** Ο κωδικός αναφοράς του FBI για την κατηγοριοποίηση του εγκλήματος σύμφωνα με το Εθνικό Σύστημα Αναφοράς του FBI (NIBRS).

Αυτός είναι ο πίνακας που δημιουργήθηκε στη βάση:

Column Name	Data Type	Allow Nulls
date_id	int	<input type="checkbox"/>
date_label	date	<input type="checkbox"/>
year	int	<input type="checkbox"/>
month	int	<input type="checkbox"/>
day	int	<input type="checkbox"/>
		<input type="checkbox"/>



Αφού δημιουργήθηκε ο πίνακας, καταχωρήσαμε τα δεδομένα από τον πίνακα **crimes** με τη χρήση του **INSERT INTO**.

```
INSERT INTO case_dim (case_number, iucr, primary_type, description, fbicode)
SELECT DISTINCT CaseNumber, IUCR, PrimaryType, Description, FBICode
FROM crimes;
```

2.3.2 Location Dimension

Το **Location Dimension** αναφέρεται στην αποθήκευση και οργάνωση των δεδομένων που περιγράφουν την τοποθεσία κάθε εγκλήματος. Αυτά τα δεδομένα είναι κρίσιμα για την ανάλυση της γεωγραφικής κατανομής των εγκλημάτων και την κατηγοριοποίηση τους σε διάφορες γεωγραφικές περιοχές του Σικάγο, όπως η συνοικία, το μπλοκ, ο τομέας αστυνομίας, και οι συντεταγμένες.

Για τη δημιουργία αυτού του dimension, συγκεντρώθηκαν οι εξής στήλες από τον πίνακα crimes:

- **Block**
- **Beat**
- **District**
- **Ward**
- **Community Area**
- **X Coordinate, Y Coordinate**
- **Longitude, Latitude**
- **Location**

Αυτός είναι ο πίνακας που δημιουργήθηκε στη βάση:

Column Name	Data Type	Allow Nulls
location_id	int	<input type="checkbox"/>
block	nvarchar(255)	<input type="checkbox"/>
location_description	nvarchar(255)	<input type="checkbox"/>
beat	float	<input type="checkbox"/>
district	float	<input type="checkbox"/>
ward	float	<input type="checkbox"/>
community_area	float	<input type="checkbox"/>
x_coordinate	float	<input type="checkbox"/>
y_coordinate	float	<input type="checkbox"/>
latitude	float	<input type="checkbox"/>
longitude	float	<input type="checkbox"/>
location	nvarchar(255)	<input type="checkbox"/>
		<input type="checkbox"/>

Αυτή η εντολή δημιουργεί το **location_dim** με τον **location_id** να είναι το πρωτεύον κλειδί, ενώ οι άλλες στήλες περιλαμβάνουν πληροφορίες σχετικά με την τοποθεσία του εγκλήματος, συμπεριλαμβανομένων των συντεταγμένων και των γεωγραφικών περιοχών του Σικάγο. Η στήλη **location** καταχωρείται ως μοναδική, για να αποφεύγονται οι διπλότυπες καταχωρήσεις των τοποθεσιών.

Indexes/Keys

Selected Primary/Unique Key or Index: **UQ_location_412AE05CB7C885C8**

Editing properties for existing primary/unique key or index.

(General)

Columns	location (ASC)
Is Unique	Yes
Type	Unique Key

Identity

(Name)	UQ_location_412AE05CB7C885C8
Description	

Table Designer

Create As Clustered	No
Data Space Specification	PRIMARY
Fill Specification	

Add Delete Close

Μετά τη δημιουργία του πίνακα, καταχωρήθηκαν τα δεδομένα από τον πίνακα crimes χρησιμοποιώντας την εντολή **INSERT INTO**, με τον εξής τρόπο:

```
INSERT INTO location_dim (block, location_description, beat, district, ward, community_area, x_coordinate, y_coordinate, latitude, longitude, location)
SELECT DISTINCT [Block], [LocationDescription], [Beat], [District], [Ward], [CommunityArea], [X_Coordinate], [Y_Coordinate], [Latitude], [Longitude], [Location]
FROM crimes;
```

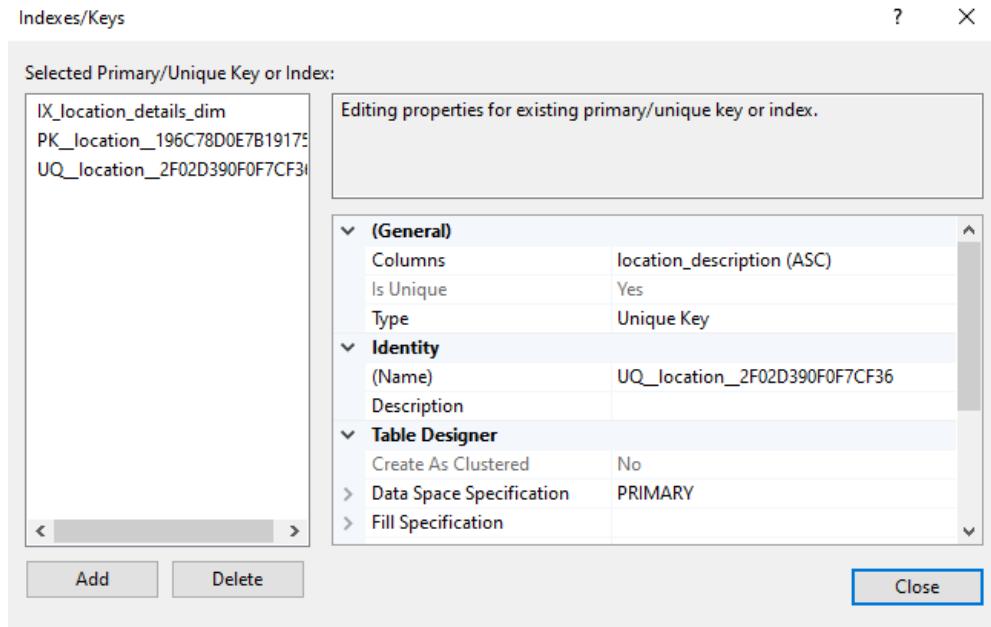
2.3.3 Location Details Dimension

To **Location Details Dimension** αναφέρεται σε πληροφορίες που προσδιορίζουν πιο εξειδικευμένα την τοποθεσία του εγκλήματος. Αυτή η διάσταση επικεντρώνεται στην περιγραφή της τοποθεσίας και προσφέρει μια πιο τεκμηριωμένη ανάλυση, που μπορεί να αφορά συγκεκριμένους τύπους τοποθεσιών, όπως εγκαταλειμμένα κτίρια, δρόμους και άλλα. Για τη δημιουργία αυτού του dimension, συγκεντρώθηκαν οι εξής στήλες από τον πίνακα crimes:

- **Block**
- **Location Description**
- **Beat**
- **District**
- **Ward**
- **Community Area**
- **X Coordinate, Y Coordinate**
- **Latitude, Longitude**

Αυτός είναι ο πίνακας που δημιουργήθηκε στη βάση:

Column Name	Data Type	Allow Nulls
location_details_id	int	<input type="checkbox"/>
location_description	nvarchar(255)	<input type="checkbox"/>
block	nvarchar(255)	<input type="checkbox"/>
beat	float	<input type="checkbox"/>
district	float	<input type="checkbox"/>
ward	float	<input type="checkbox"/>
community_area	float	<input type="checkbox"/>
latitude	float	<input type="checkbox"/>
longitude	float	<input type="checkbox"/>
		<input type="checkbox"/>



Μετά τη δημιουργία του πίνακα, καταχωρήθηκαν τα δεδομένα από τον πίνακα crimes ως εξής:

```
INSERT INTO location_details_dim (location_description, block, beat, district, ward, community_area, latitude, longitude)
SELECT DISTINCT
    [LocationDescription],
    [Block],
    [Beat],
    [District],
    [Ward],
    [CommunityArea],
    [Latitude],
    [Longitude]
FROM crimes;
```

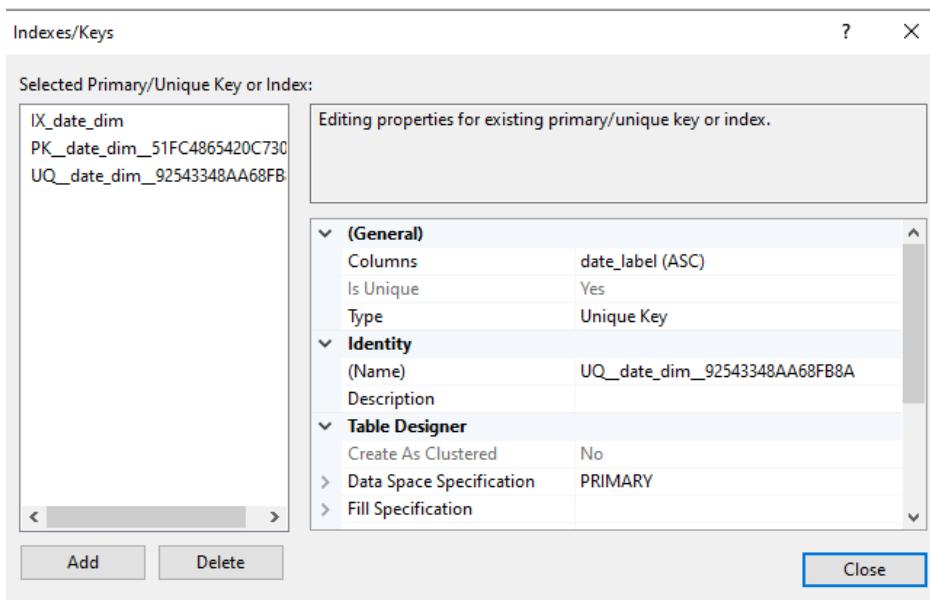
2.3.4 Date Dimension

Το **Date Dimension** είναι απαραίτητη για την αποθήκευση των δεδομένων που αφορούν την ημερομηνία κάθε περιστατικού και για την ανάλυση της χρονικής διάστασης των εγκλημάτων. Ο βασικός στόχος της συγκεκριμένης διάστασης είναι να δημιουργήσει δεδομένα που αφορούν την ημερομηνία του περιστατικού, και να παρέχει επιπλέον πληροφορίες για την ανάλυση, όπως το έτος, τον μήνα και την ημέρα. Για τη δημιουργία αυτού του dimension, χρησιμοποιήθηκαν οι εξής στήλες από τον πίνακα crimes:

- **Date**
- **Year**
- **Month**
- **Day**

Αυτός είναι ο πίνακας που δημιουργήθηκε στη βάση:

Column Name	Data Type	Allow Nulls
date_id	int	<input type="checkbox"/>
date_label	date	<input type="checkbox"/>
year	int	<input type="checkbox"/>
month	int	<input type="checkbox"/>
day	int	<input type="checkbox"/>



Αυτή η εντολή δημιουργεί τον πίνακα **date_dim**, όπου το **date_id** είναι το πρωτεύον κλειδί, και οι υπόλοιπες στήλες προσδιορίζουν την ημερομηνία και την κατηγοριοποίηση αυτής.

Τα δεδομένα από τον πίνακα **crimes** καταχωρούνται στον πίνακα με την εξής εντολή:

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

2.3.5 Time Dimension

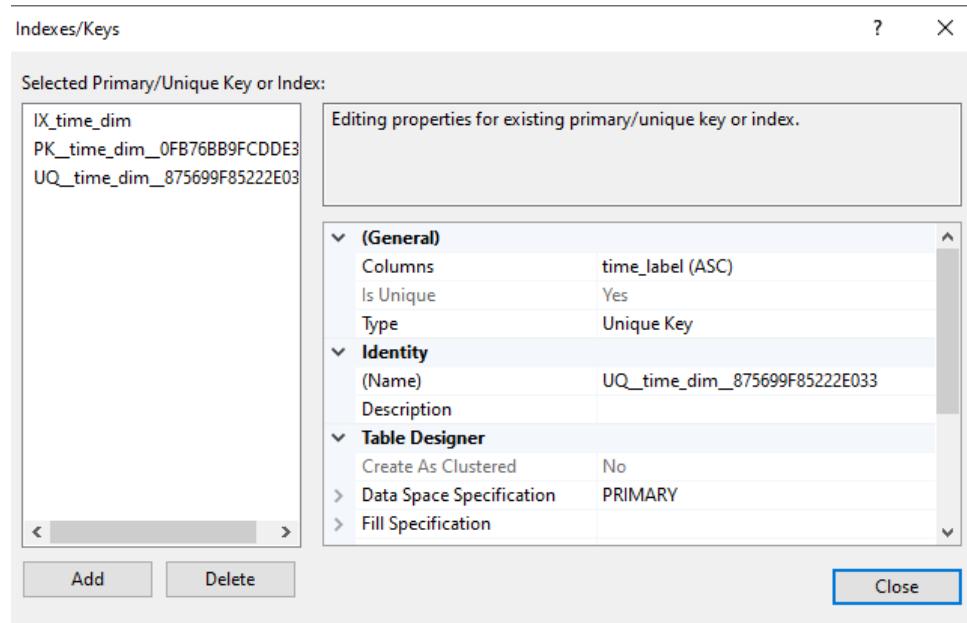
Το **Time Dimension** δημιουργήθηκε για την αποθήκευση των δεδομένων που αφορούν την ακριβή ώρα του περιστατικού. Αυτή η διάσταση είναι απαραίτητη για την ανάλυση της χρονικής κατανομής

των εγκλημάτων και για την εύρεση τάσεων σε σχέση με την ώρα της ημέρας. Η διάσταση αυτή περιλαμβάνει τα εξής δεδομένα:

- **Time**
- **Hour**
- **Minute**

Αυτός είναι ο πίνακας που δημιουργήθηκε στη βάση:

Column Name	Data Type	Allow Nulls
time_id	int	<input type="checkbox"/>
time_label	time(7)	<input type="checkbox"/>
hour	int	<input type="checkbox"/>
minute	int	<input type="checkbox"/>
		<input type="checkbox"/>



Για την εισαγωγή των δεδομένων στον πίνακα **time_dim** από τον πίνακα **crimes** χρησιμοποιήσαμε την εντολή:

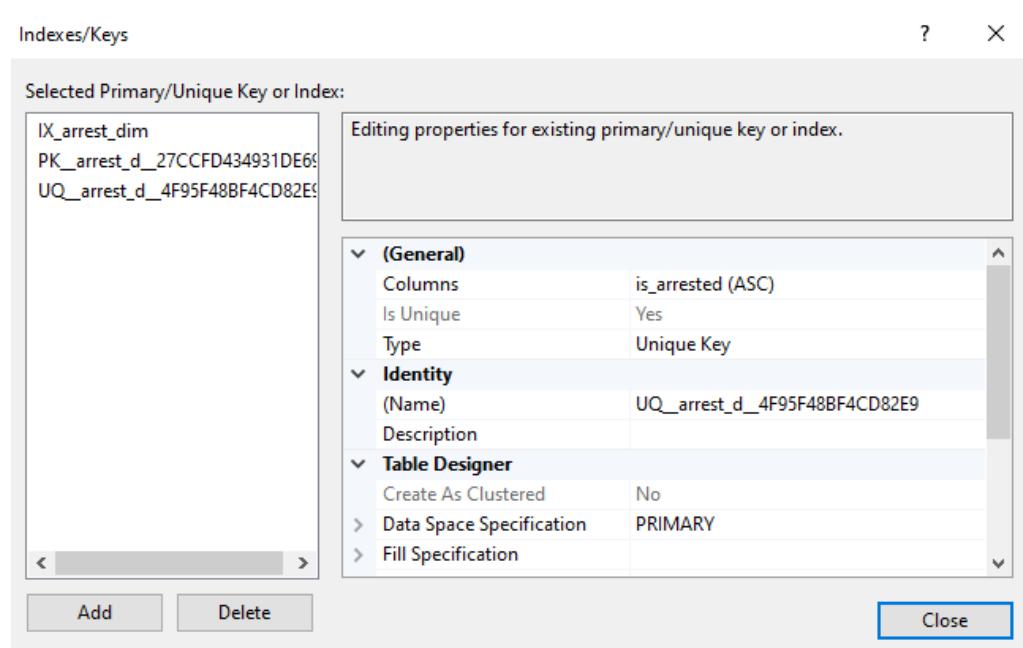
```
INSERT INTO time_dim (time_label, hour, minute)
SELECT DISTINCT
    [Time],
    DATEPART(hour, [Time]) AS hour,
    DATEPART(minute, [Time]) AS minute
FROM crimes;
```

2.3.6 Arrest Dimension

Το **Arrest Dimension** περιέχει την πληροφορία για το αν υπήρξε σύλληψη του δράστη του εγκλήματος. Η στήλη **Arrest** από τον πίνακα **crimes** μας παρέχει αυτήν την πληροφορία, η οποία καταχωρείται ως 1 (σύλληψη) ή 0 (καμία σύλληψη). Η δημιουργία του πίνακα **arrest_dim** έγινε ως εξής: ?????

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"/>



Για την εισαγωγή των δεδομένων στον πίνακα από τον πίνακα **crimes** χρησιμοποιήσαμε την εντολή:

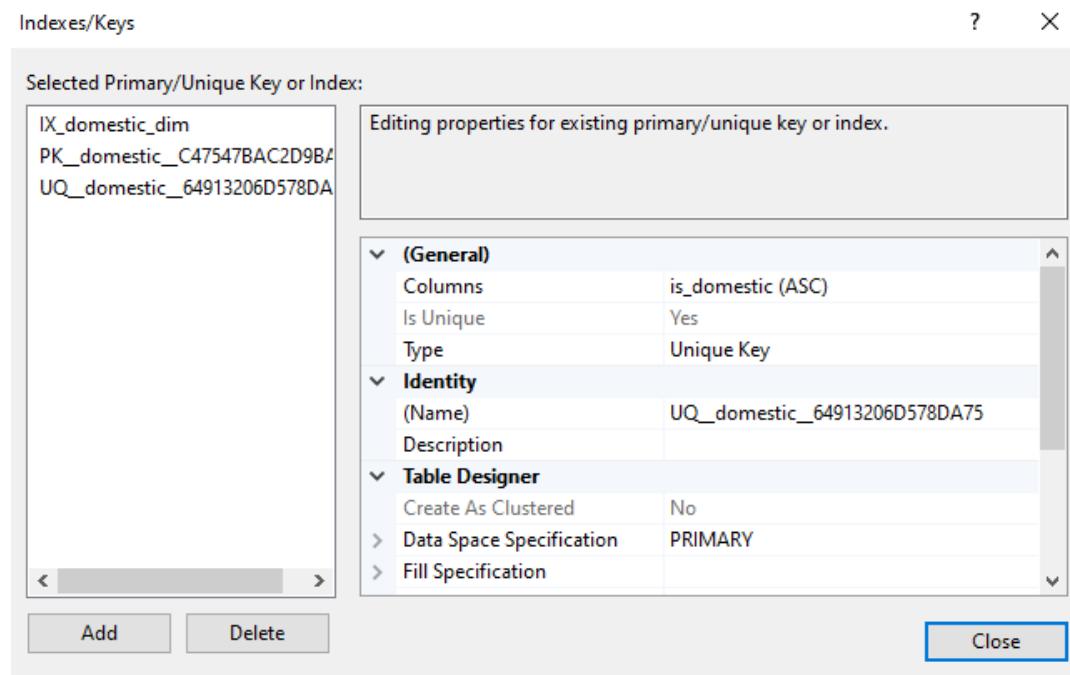
```
INSERT INTO arrest_dim (is_arrested)
SELECT DISTINCT [Arrest]
FROM crimes;
```

2.3.7 Domestic Dimension

Η **Domestic Dimension** παρέχει πληροφορίες για το αν το έγκλημα σχετίζεται με ενδοοικογενειακή βία. Η στήλη **Domestic** από τον πίνακα crimes μας προσφέρει τις τιμές 1 (σχετίζεται με ενδοοικογενειακή βία) ή 0 (δεν σχετίζεται με ενδοοικογενειακή βία).

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"/>



Για την εισαγωγή των δεδομένων στον πίνακα **domestic_dim** από τον πίνακα crimes χρησιμοποιήσαμε την εντολή:

```
|INSERT INTO domestic_dim (is_domestic)
SELECT DISTINCT [Domestic]
FROM crimes;
```

2.3.8 Socioeconomic Dimension

Η **Socioeconomic Dimension** δημιουργήθηκε με σκοπό να ενσωματώσει τις κοινωνικοοικονομικές συνθήκες της ανάλυσής μας.

Η δημιουργία του πίνακα **socioeconomic_dim** περιλαμβάνει τις εξής στήλες:

- **Community Area**
- **Per Capita Income**
- **Poverty Rate**
- **Unemployment Rate**
- **Hardship Index**

Αυτός είναι ο πίνακας που δημιουργήθηκε στη βάση:

Column Name	Data Type	Allow Nulls
socioecon_id	int	<input type="checkbox"/>
community_area	int	<input type="checkbox"/>
per_capita_income	int	<input type="checkbox"/>
poverty_rate	float	<input type="checkbox"/>
unemployment_rate	float	<input type="checkbox"/>
hardship_index	float	<input type="checkbox"/>
		<input type="checkbox"/>

Αντίστοιχα, τα δεδομένα καταχωρούνται με την εξής εντολή:

```
INSERT INTO socioeconomic_dim (community_area, per_capita_income, poverty_rate, unemployment_rate, hardship_index)
SELECT DISTINCT [CommunityArea], [PerCapitaIncome], [PovertyRate], [UnemploymentRate], [HardshipIndex]
FROM crimes;
```

2.4 Δημιουργία Fact Table

Για τη δημιουργία του **fact table**, χρησιμοποιήσαμε τα ξένα κλειδιά για να συνδέσουμε τα dimensions με τον πίνακα αυτόν. Τα ξένα κλειδιά περιλαμβάνουν: το **case_id** που συνδέεται με το **case_dim**, το **location_id** με το **location_dim**, το **location_details_id** με το **location_details_dim**, το **date_id** με το **date_dim**, το **time_id** με το **time_dim**, το **arrest_id** με το **arrest_dim**, το **domestic_id** με το **domestic_dim**, και το **socioecon_id** με το **socioeconomic_dim**.

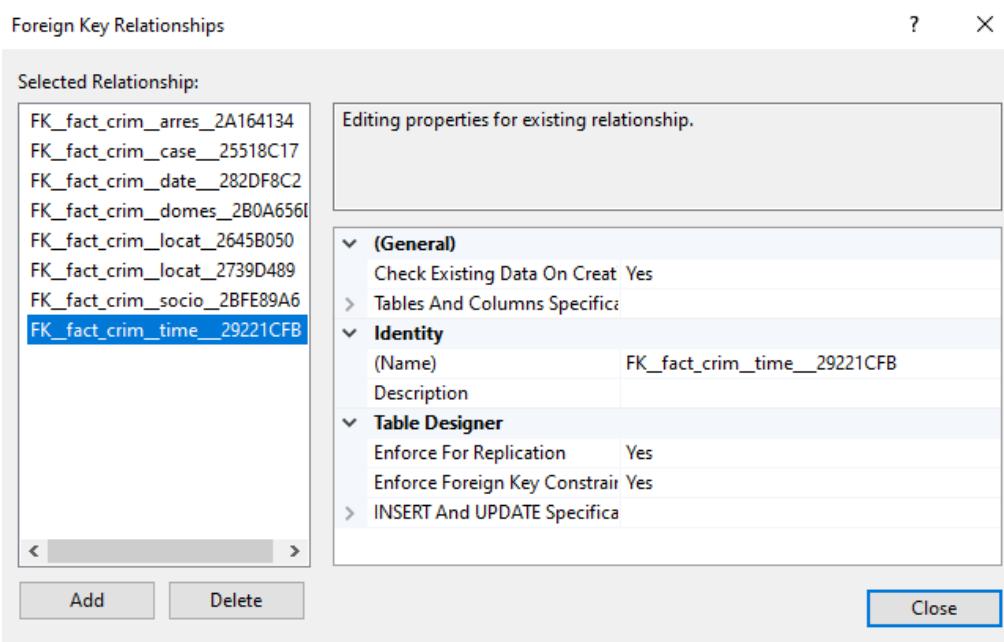
Πέρα από τα ξένα κλειδιά, προσθέσαμε επίσης τα μετρήσιμα values που σχετίζονται με τα κοινωνικοοικονομικά δεδομένα των περιοχών, όπως το **PerCapitalIncome**, το **PovertyRate**, το **UnemploymentRate** και το **HardshipIndex**. Αυτές οι στήλες επιτρέπουν την ανάλυση της εγκληματικότητας σε σχέση με τους κοινωνικοοικονομικούς δείκτες. Οι υπόλοιπες στήλες, καθώς περιλαμβάνουν περιγραφικά δεδομένα, βρίσκονται ήδη στα dimensions και επομένως δεν θα προστεθούν στον fact table.

Column Name	Data Type	Allow Nulls
id	int	<input type="checkbox"/>
case_id	int	<input type="checkbox"/>
location_id	int	<input type="checkbox"/>
location_details_id	int	<input type="checkbox"/>
date_id	int	<input type="checkbox"/>
time_id	int	<input type="checkbox"/>
arrest_id	int	<input type="checkbox"/>
domestic_id	int	<input type="checkbox"/>
socioecon_id	int	<input type="checkbox"/>
per_capita_income	int	<input checked="" type="checkbox"/>
poverty_rate	float	<input checked="" type="checkbox"/>
unemployment_rate	float	<input checked="" type="checkbox"/>
hardship_index	float	<input checked="" type="checkbox"/>
		<input type="checkbox"/>

```

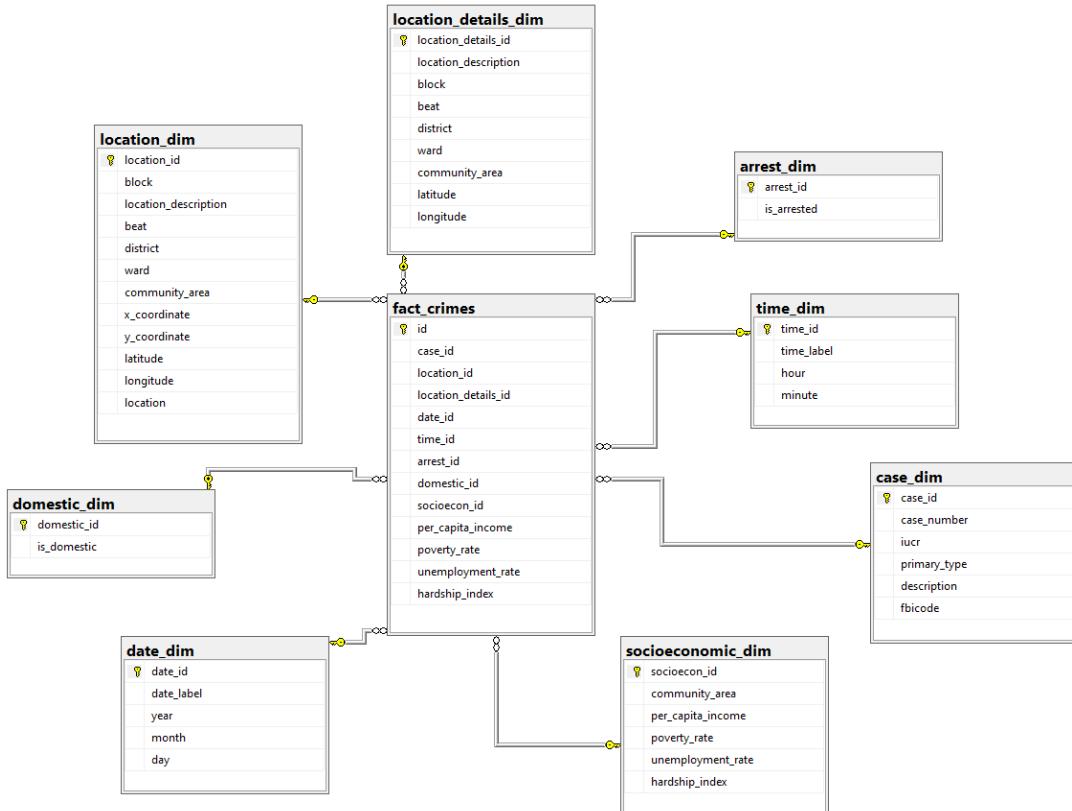
FOREIGN KEY (case_id) REFERENCES case_dim(case_id),
FOREIGN KEY (location_id) REFERENCES location_dim(location_id),
FOREIGN KEY (location_details_id) REFERENCES location_details_dim(location_details_id),
FOREIGN KEY (date_id) REFERENCES date_dim(date_id),
FOREIGN KEY (time_id) REFERENCES time_dim(time_id),
FOREIGN KEY (arrest_id) REFERENCES arrest_dim(arrest_id),
FOREIGN KEY (domestic_id) REFERENCES domestic_dim(domestic_id),
FOREIGN KEY (socioecon_id) REFERENCES socioeconomic_dim(socioecon_id)

```



2.5 Star Schema

Παρακάτω φαίνεται το **Star Schema** που δημιουργήσαμε για τη διαχείριση των δεδομένων. Στο κέντρο βρίσκεται το **fact table**, το οποίο συνδέεται με τα dimensions μέσω ξένων κλειδιών, επιτρέποντας την ανάλυση των δεδομένων σε διάφορες διαστάσεις, όπως το **Case**, το **Location**, το **Date**, το **Time**, το **Arrest**, το **Domestic**, και το **Socioeconomic**.



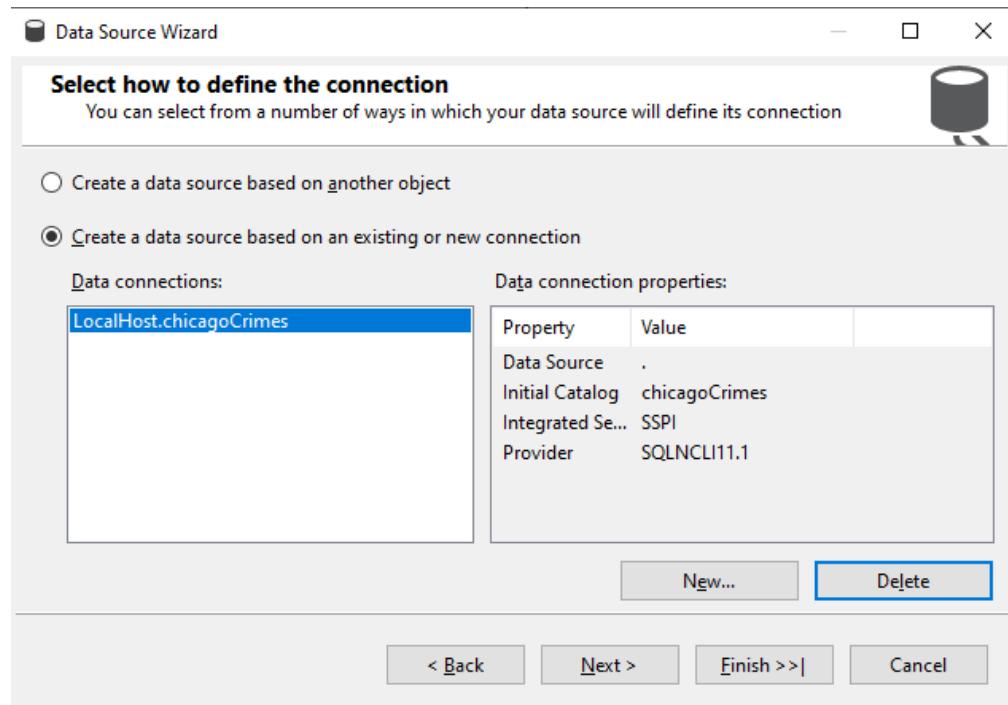
Visual Studio - Δημιουργία κύβου

3.1 Δημιουργία Project

Αρχικά, εγκαταστήσαμε τα απαιτούμενα plugins SSIS και SSAS στο Visual Studio. Στη συνέχεια, δημιουργήσαμε ένα νέο **Multidimensional Analysis Services Project** με το όνομα **ChicagoCrimesMulti**, το οποίο θα χρησιμοποιήσουμε για την ανάλυση των δεδομένων που συλλέξαμε από το SQL Server Management Studio.

3.2 Δημιουργία Data Source και Data Source Views

Με τον **wizard** του Visual Studio, καθορίσαμε την data source συνδέοντας το project με τη βάση δεδομένων **chicagoCrimes**. Στη συνέχεια, δημιουργήσαμε τα Data Source Views, επιλέγοντας τα απαραίτητα **dimensions** και το **fact table** που χρησιμοποιούνται στην ανάλυση.



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
sysdiagrams (dbo)	Table

Included objects:

Name	Type
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
socioeconomic_dim (...)	Table

Filter: ▾

Add Related Tables

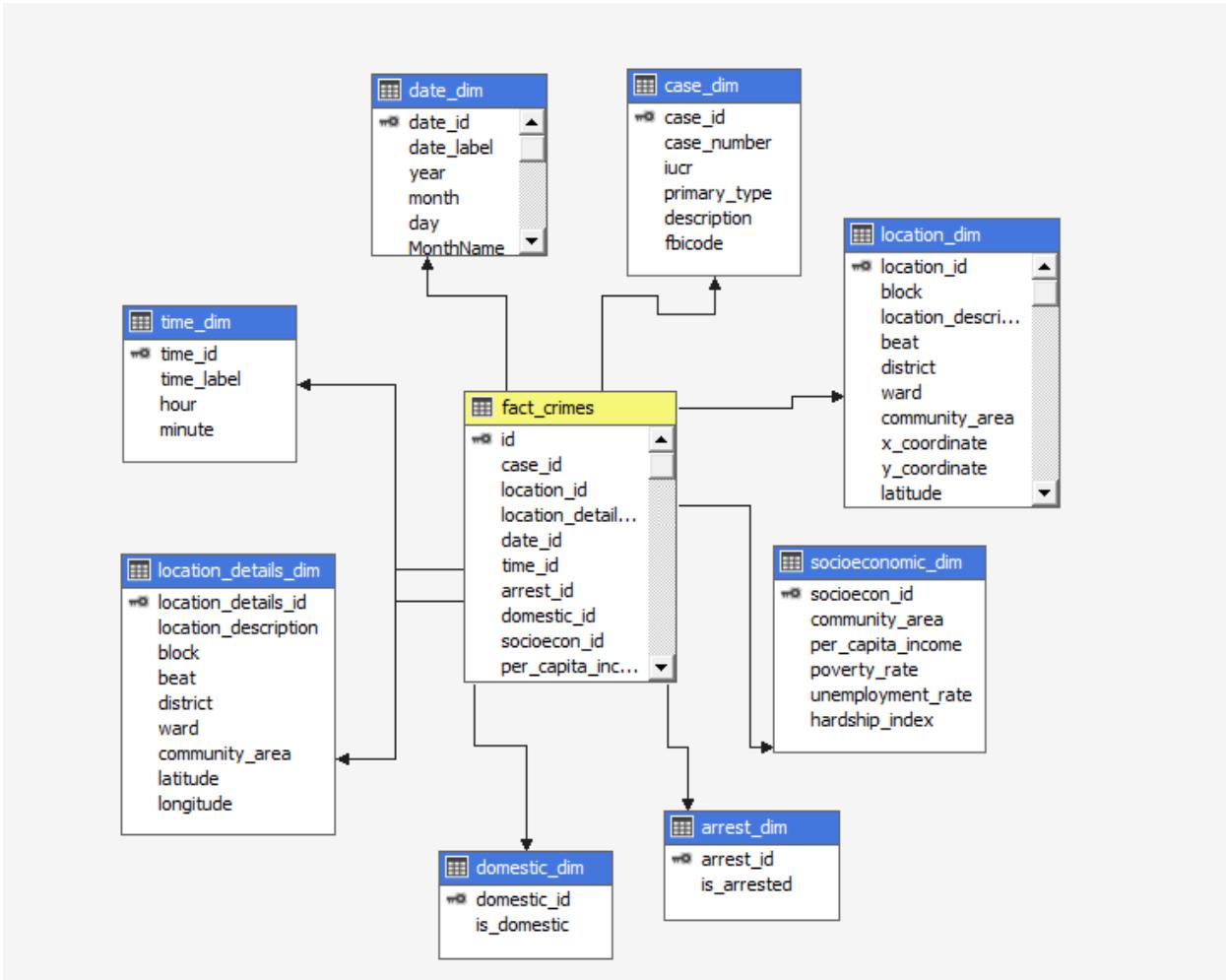
Show system objects

< Back Next > Finish >> | Cancel

3.3 Δημιουργία του Cube

Έπειτα, για τη δημιουργία του κύβου, ορίσαμε τα **measures** (μετρήσεις), όπως είναι το Crime Count, το Poverty Rate, το Per Capita Income, και το Unemployment Rate, και επιλέξαμε τα dimensions που σχετίζονται με τα μέτρα αυτά.

Ο κύβος παρουσιάζεται παρακάτω:



3.4 Ανάπτυξη και Εγκατάσταση του Cube

Το τελικό βήμα μετά τη δημιουργία του κύβου είναι το **deploy** του στον server Analysis Services (SSAS). Το deploy καθιστά τον κύβο προσβάσιμο στους τελικούς χρήστες και στα εργαλεία οπτικοποίησης, όπως το Power BI, επιτρέποντάς του να φορτώνει τα δεδομένα από την πηγή τους για γρήγορη εκτέλεση ερωτημάτων και ανάλυση.

Οπτικοποίηση με Power BI

4.1 Υπολογισμός βασικών μετρικών

Πριν προχωρήσουμε στην παρουσίαση των γραφημάτων και των οπτικοποιήσεων, θελήσαμε να κάνουμε μία προκαταρκτική ανάλυση των δεδομένων μας προκειμένου να κατανοήσουμε καλύτερα τις μετρικές και τις τάσεις που ενδέχεται να εμφανιστούν. Ορισμένες από τις κύριες μετρικές που προκύπτουν είναι οι εξής:

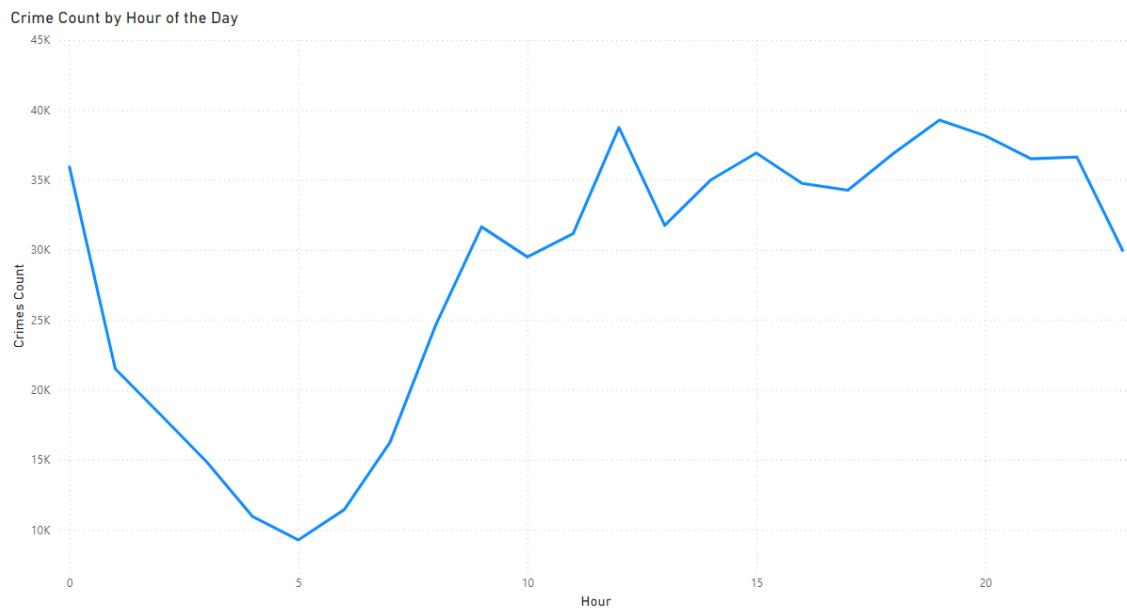
- Συνολικός αριθμός εγκλημάτων: **684.175**
- Συνολικός αριθμός συλλήψεων: **186.329**, δηλαδή περίπου **27%** των εγκλημάτων έχουν οδηγήσει σε σύλληψη
- Τα εγκλήματα που σχετίζονται με ενδοοικογενειακή βία αντιπροσωπεύουν περίπου το **14%** του συνόλου
- Οι κλοπές αποτελούν περίπου το **22%** των συνολικών εγκλημάτων
- Η ώρα με τα περισσότερα εγκλήματα είναι η **19:00** (39.281 εγκλήματα) ενώ η ώρα με τα λιγότερα εγκλήματα η **05:00** (9269 εγκλήματα)
- Ο **Ιούλιος** είναι ο μήνας με τα περισσότερα εγκλήματα (64.847), ενώ ο **Φεβρουάριος** καταγράφει τα λιγότερα (45.867)

Επιπλέον, παρατηρήσαμε αρκετά σημαντικές συσχετίσεις μεταξύ των κοινωνικοοικονομικών δεικτών και της εγκληματικότητας, με την ανεργία και το ποσοστό φτώχειας να έχουν έντονη σχέση με την αύξηση των εγκλημάτων.

4.2 Visualizations

4.2.1 Crimes Count by Hour

Το διάγραμμα δείχνει τη συχνότητα των εγκλημάτων καθ' όλη τη διάρκεια της ημέρας, με σημαντικές κορυφώσεις τις βραδινές ώρες από τις 17:00 έως τις 00:00, με την κορύφωση να είναι στις 19:00, υποδεικνύοντας ότι η εγκληματικότητα τείνει να αυξάνεται με την αύξηση της κοινωνικής δραστηριότητας. Παρατηρείται μια εμφανής μείωση στις πρώιμες πρωινές ώρες (3 π.μ. έως 6 π.μ.), που υποδεικνύει χαμηλότερα ποσοστά εγκλημάτων κατά τη διάρκεια πιο ήσυχων περιόδων.



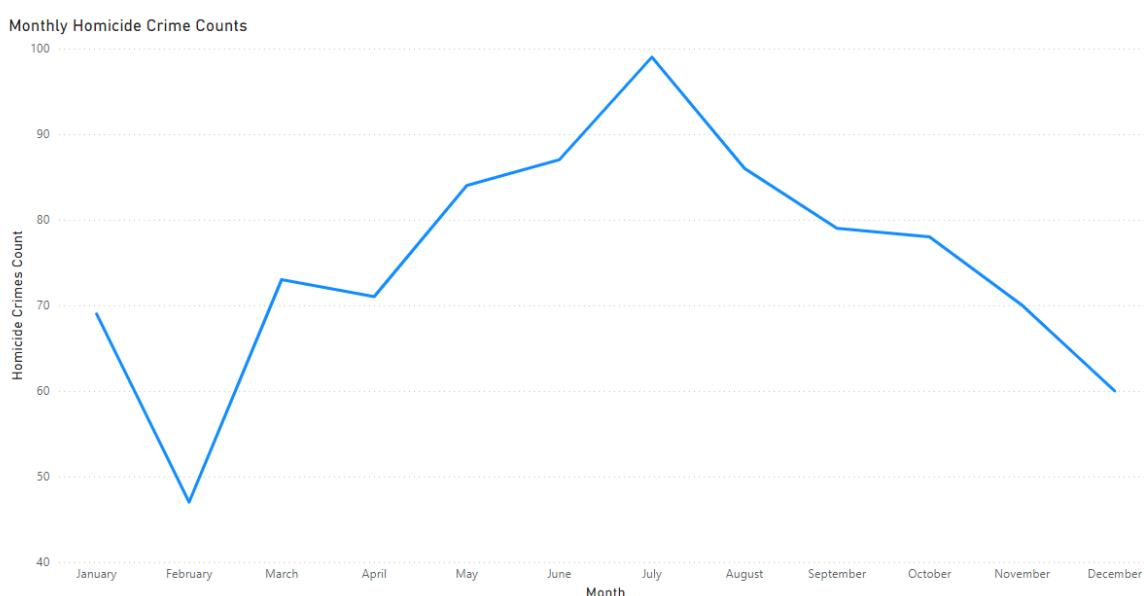
4.2.2 Monthly Crime Counts for 2011 and 2012

Στο διάγραμμα αυτό βλέπουμε τον μηνιαίο αριθμό εγκλημάτων για τα έτη 2011 και 2012, όπου φαίνεται να υπάρχει μια εποχιακή τάση με αύξηση των εγκλημάτων κατά τους θερινούς μήνες και μείωση τους κατά τον χειμώνα. Και τα δύο έτη ακολουθούν παρόμοιο μοτίβο, αλλά το 2011 παρουσιάζει ελαφρώς υψηλότερες κορυφές, υποδεικνύοντας αυξημένη εγκληματικότητα. Αυτό μπορεί να σχετίζεται με τις υποκείμενες οικονομικές πιέσεις από την οικονομική κρίση του 2008, οι οποίες συχνά οδηγούν σε υψηλότερους δείκτες κλοπών και βίας. Σε έρευνα μας διαπιστώσαμε ότι το 2011, παρατηρήθηκε σημαντική αύξηση στις επιθέσεις και τις κλοπές στο Σικάγο, υποδεικνύοντας μια ευρύτερη διάσταση εγκληματικής δραστηριότητας εκείνη τη χρονιά. Οι μήνες Δεκέμβριος, Ιανουάριος και Φεβρουάριος είχαν λιγότερα εγκλήματα, πιθανόν λόγω ταξιδιών για τις γιορτές και χαμηλότερης θερμοκρασίας που μείωσαν τη δραστηριότητα σε εξωτερικούς χώρους.



4.2.3 Monthly Homicide Crime Counts

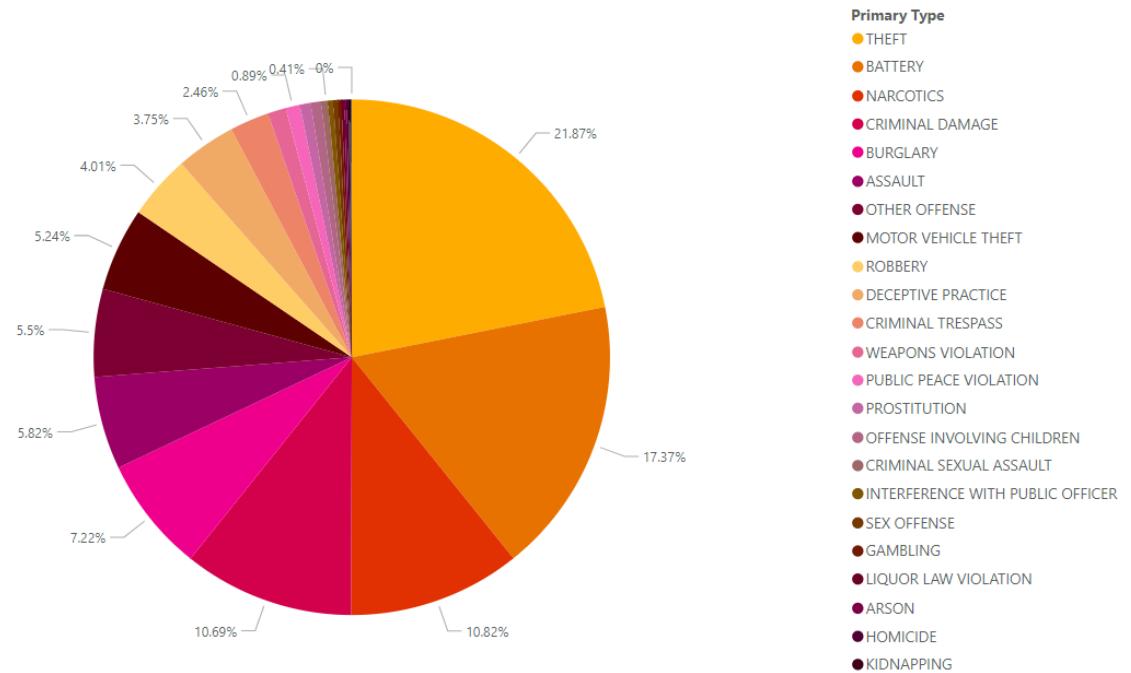
Αυτό το διάγραμμα δείχνει τις μηνιαίες καταγραφές εγκλημάτων ανθρωποκτονίας για τα έτη 2011 και 2012. Αποφασίσαμε να εστιάσουμε σε αυτό το είδος εγκλήματος ξεχωριστά, καθώς αυτά τα χρόνια συγκεκριμένα είχαν παρουσιάσει σημαντική αύξηση σε σχέση με τα προηγούμενα χρόνια, κυρίως λόγω της έντονης δραστηριότητας συμμοριών. Όπως φαίνεται, ιδιαίτερα τους μήνες Ιούνιο και Ιούλιο, υπάρχει μια απότομη αύξηση, υποδηλώνοντας ότι οι ανθρωποκτονίες αυξάνονται το καλοκαίρι λόγω παραγόντων όπως είναι η αυξημένη κοινωνική δραστηριότητα και η υψηλή θερμοκρασία ενώ η τάση μειώνεται προς το χειμώνα.



4.2.4 Distribution of Crime Counts by Primary Crime Type

Αυτό το γράφημα πίτας δείχνει την κατανομή των εγκλημάτων κατά τύπο, με τις κλοπές (21.87%), τις χειροδικίες (17.37%) και τα ναρκωτικά (10.69%) να καλύπτουν σχεδόν το 50% των εγκλημάτων. Οι στοχευμένες παρεμβάσεις λοιπόν σε αυτές τις κατηγορίες εγκλημάτων από τις κατάληξες αρχές μπορούν να μειώσουν σημαντικά τα συνολικά ποσοστά εγκληματικότητας.

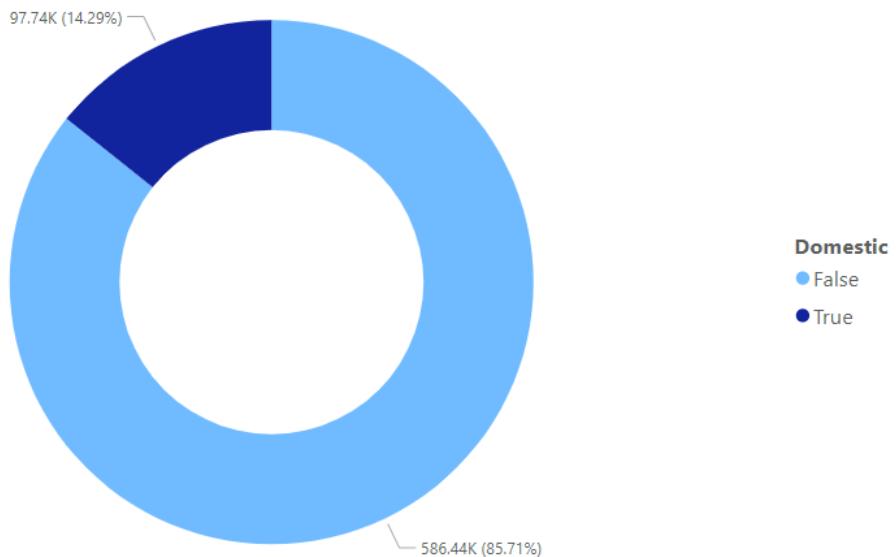
Distribution of Crime Counts by Primary Crime Type



4.2.5 Domestic vs. Non-Domestic Crimes Distribution

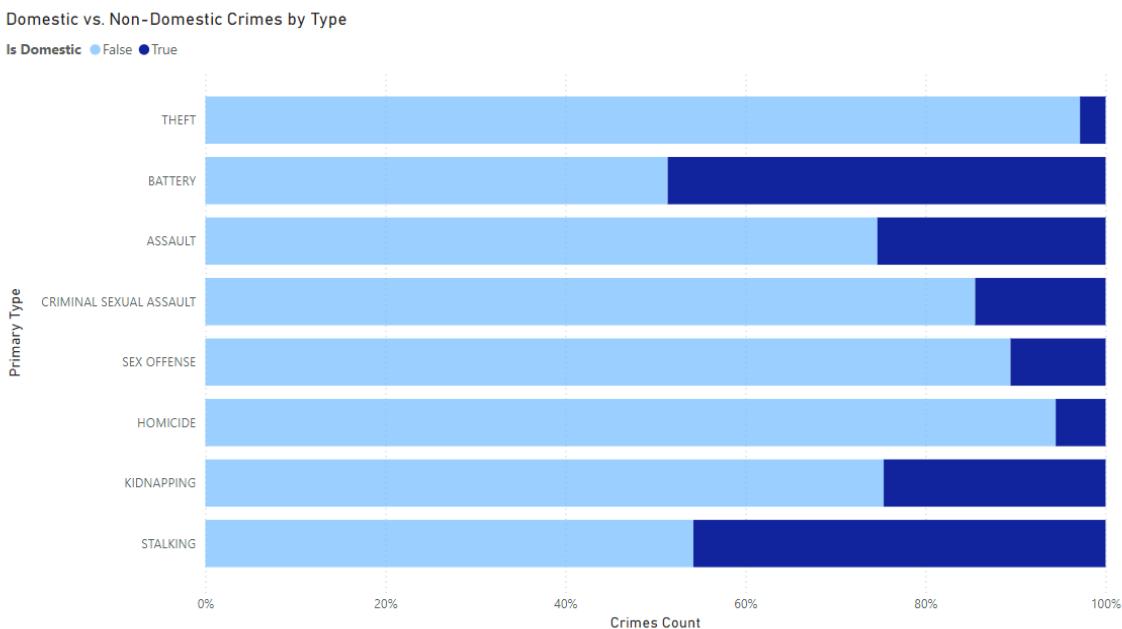
Αυτό το διάγραμμα δείχνει την κατανομή των ενδοοικογενειακών και μη-ενδοοικογενειακών εγκλημάτων στο Σικάγο. Το 85,71% των εγκλημάτων είναι μη-ενδοοικογενειακά, υποδεικνύοντας ότι τα περισσότερα εγκλήματα συμβαίνουν εκτός του οικογενειακού πλαισίου.

Domestic vs. Non-Domestic Crimes Distribution



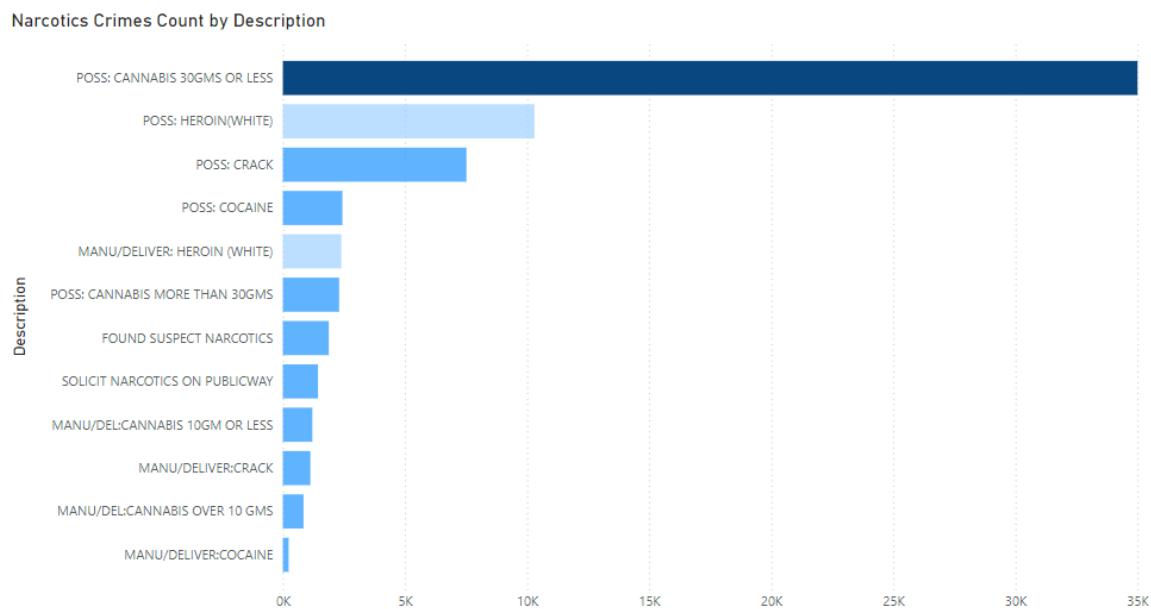
4.2.6 Domestic vs. Non-Domestic Crimes by Type

Ωστόσο, αυτό το διάγραμμα δείχνει την ανησυχητική έκταση της οικογενειακής εμπλοκής σε σοβαρά εγκλήματα όπως είναι η χειροδικία, η απαγωγή και η παρακολούθηση. Ένα σημαντικό ποσοστό αυτών των εγκλημάτων φαίνεται πως συμβαίνει μέσα σε οικογενειακά περιβάλλοντα, υποδεικνύοντας την έκταση της οικογενειακής βίας στην πολιτεία. Τα υψηλά αυτά ποσοστά στα ενδοοικογενειακά εγκλήματα δείχνουν την ανάγκη για άμεση επέμβαση από τις κατάλληλες αρχές για την αντιμετώπιση της οικογενειακής κακοποίησης και την υποστήριξη των θυμάτων.



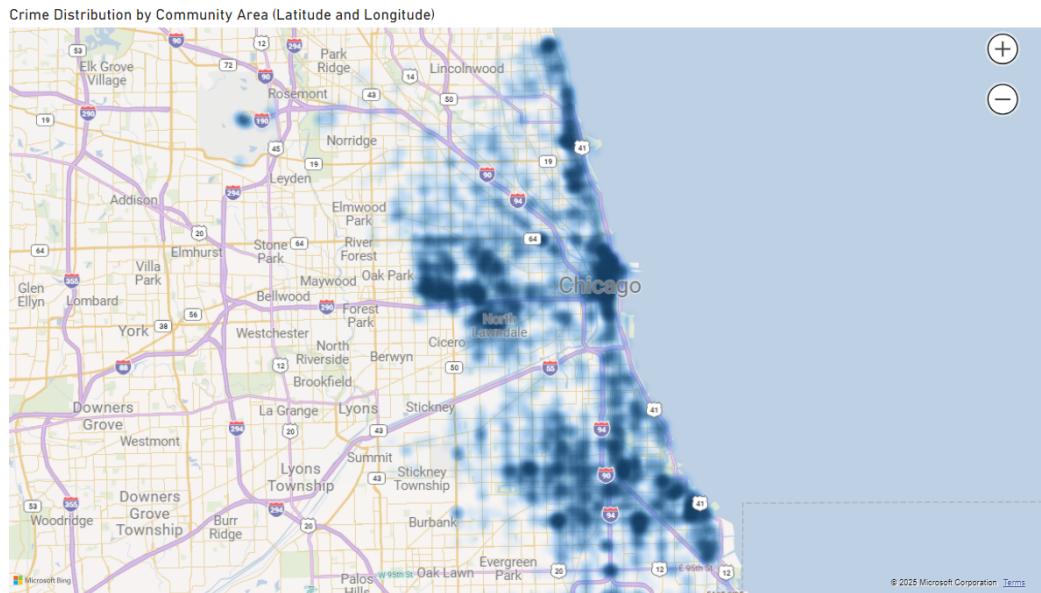
4.2.7 Narcotics Crimes Count by Description

Αυτό το διάγραμμα δείχνει την κατανομή των εγκλημάτων που σχετίζονται με ναρκωτικά, επισημαίνοντας ότι η πιο συχνή κατηγορία είναι η κατοχή κάνναβης (30γρ ή λιγότερο), η οποία αντιπροσωπεύει μεγάλο ποσοστό του συνόλου. Η κατοχή ηρωίνης επίσης καταγράφει σημαντικό αριθμό περιστατικών, ενώ η διακίνησή της είναι λιγότερο συχνή. Τα δεδομένα δείχνουν ότι η αστυνομία επικεντρώνεται περισσότερο στους χρήστες (κατοχή) παρά στους διακινητές (κατασκευή ή διανομή) όταν πρόκειται για εγκλήματα που αφορούν ηρωίνη.

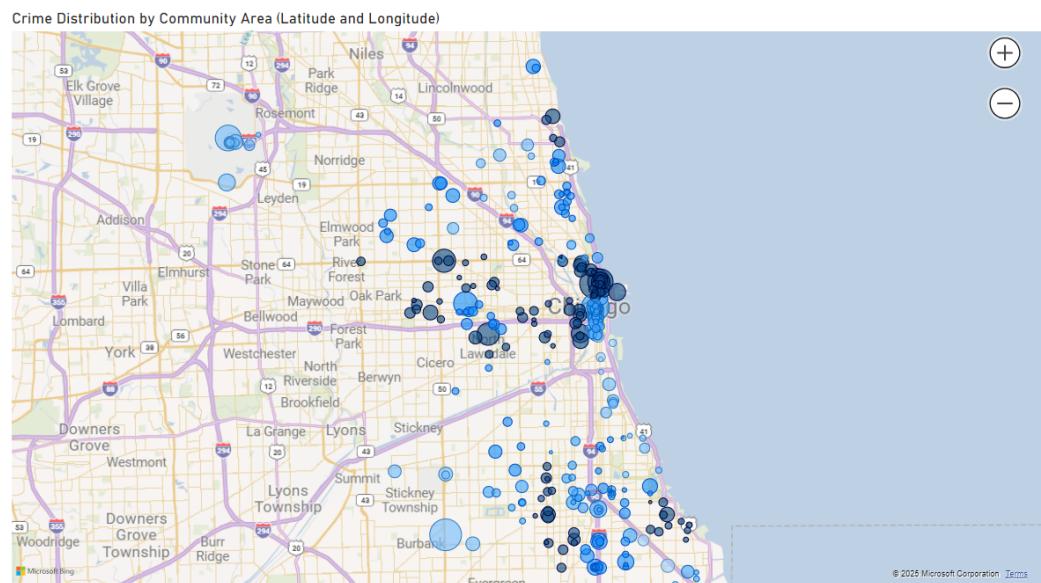


4.2.8 Crime Hotspots in Chicago

Ο πρώτος χάρτης απεικονίζει τη γενική κατανομή των εγκλημάτων στο Σικάγο με τη χρήση ενός **heatmap**. Οι διάφορες αποχρώσεις του μπλε αναπαριστούν τη intensity των εγκλημάτων, με τις πιο σκούρες αποχρώσεις να υποδεικνύουν περιοχές με υψηλότερο αριθμό εγκλημάτων. Οι κεντρικές και βόρειες περιοχές της πόλης παρουσιάζουν συγκέντρωση εγκλημάτων, κάτι που μπορεί να οφείλεται σε παράγοντες όπως η πυκνότητα του πληθυσμού και η αστικοποίηση.



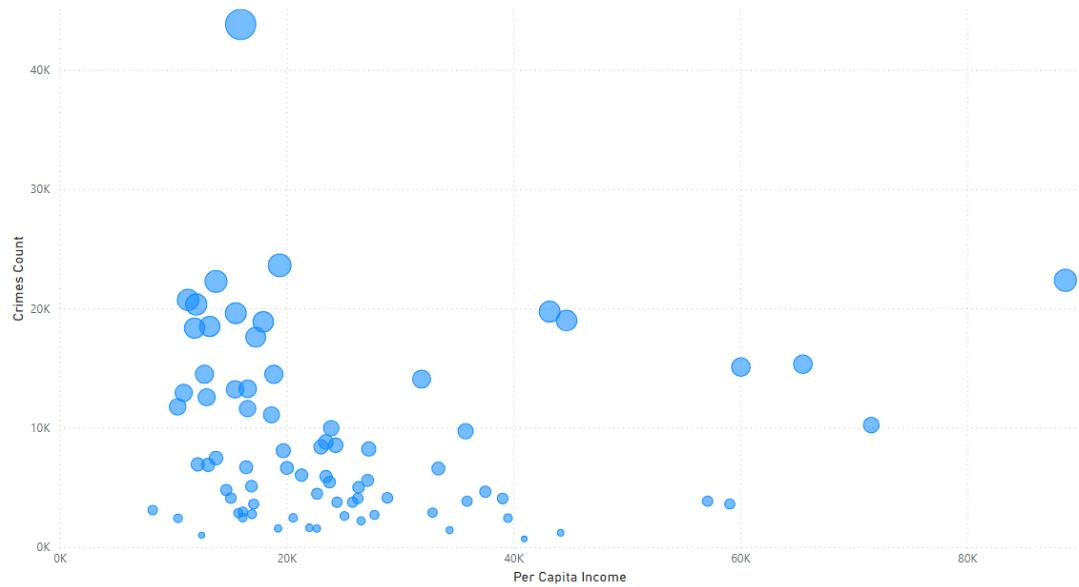
Ο δεύτερος χάρτης επικεντρώνεται στην κατανομή των εγκλημάτων ανά κοινωνική περιοχή, όπου το μέγεθος των φυσαλίδων αναπαριστά τη συχνότητα των εγκλημάτων σε κάθε περιοχή. Οι μεγαλύτερες φυσαλίδες υποδεικνύουν περιοχές με υψηλότερο ποσοστό εγκλημάτων, κυρίως στην αστική περιοχή, ενώ μικρότερες φυσαλίδες βρίσκονται σε περισσότερο σε προαστιακές περιοχές.



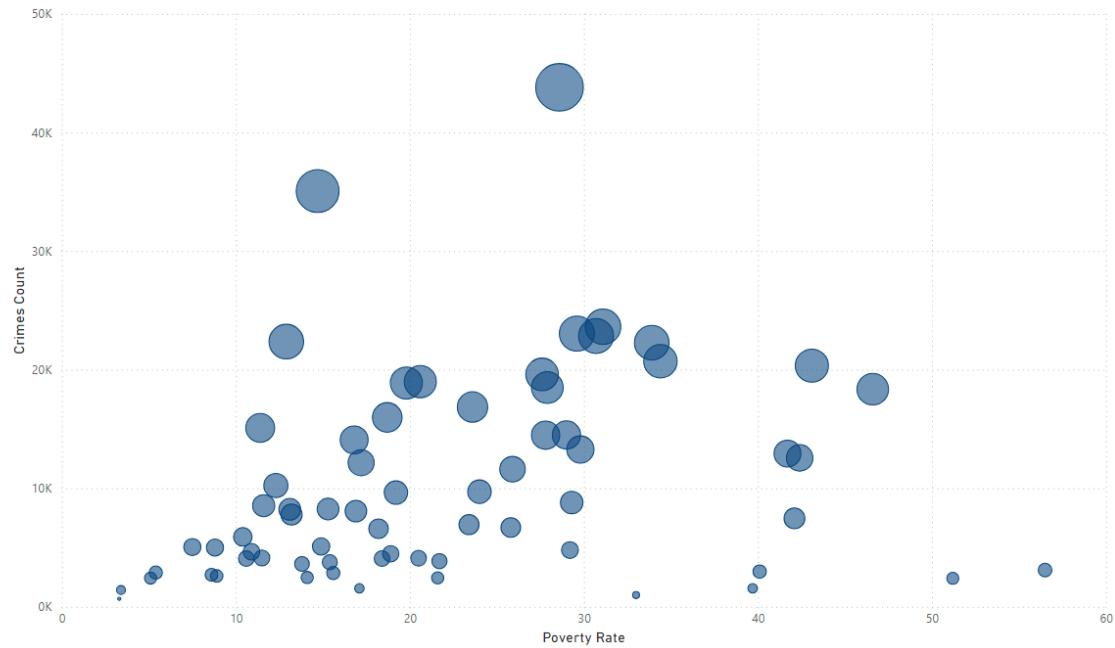
4.2.9 Socioeconomic Indicators vs. Crime Count

Και τα δύο scatter plots δείχνουν ότι οι περιοχές με υψηλότερους δείκτες εγκληματικότητας συνήθως βρίσκονται σε περιοχές με χαμηλότερο κατά κεφαλήν εισόδημα και υψηλότερο ποσοστό φτώχειας. Αν και υπάρχουν κάποιες εξαιρέσεις, όπως το community area 25 που θα αναλυθεί περαιτέρω στη συνέχεια, η γενική τάση υποδεικνύει ότι περιοχές με υψηλότερη φτώχεια ή/και χαμηλότερο εισόδημα έχουν πιο συχνά εγκλήματα.

Crime Count vs. Per Capita Income



Crime Count vs. Poverty Rate



Data mining

5.1 Clustering

Η πρώτη μέθοδος που θα υλοποιήσουμε είναι η Συσταδοποίηση για να ερευνήσουμε τη σχέση μεταξύ των ποσοστών φτώχειας και των επιπέδων εγκληματικότητας στα 77 community areas του Chicago στην περίοδο των 2 ετών.

Για το καινούριο dataframe χρησιμοποιούμε τις υπάρχουσες στήλες **Community Area** και **Poverty Rate** και για να βρούμε τον συνολικό αριθμό εγκλημάτων ανά community area παίρνουμε το άθροισμα των γράμμων (εγκλημάτων) στις οποίες εμφανίζεται κάθε community area.

Επομένως, το dataframe που προκύπτει έχει ως index τα 77 community areas και περιέχει για το καθένα το ποσοστό φτώχειας και τον αριθμό των συνολικών εγκλημάτων.

Community Area	Poverty_Rate	Total_Crimes
0	1	23.6
1	2	17.2
2	3	24.0
3	4	10.9
4	5	7.5
...
72	73	16.9
73	74	3.4
74	75	13.2
75	76	15.4
76	77	18.2

Πριν την υλοποίηση της συσταδοποίησης κανονικοποιούμε με τη μέθοδο **Min-Max** τα δεδομένα προς μελέτη, ώστε οι διαφορετικές μονάδες μέτρησης, μεγέθη και κατανομές τους, να μην επηρεάζουν τους υπολογισμούς απόστασης και ανάθεσης συστάδων.

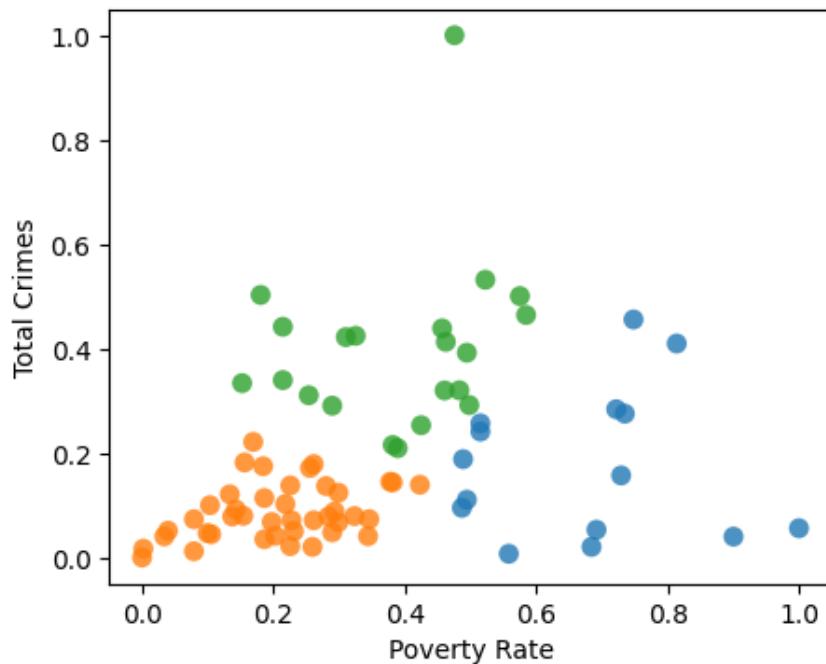
Το **Min-Max** αφαιρεί την ελάχιστη τιμή από κάθε στοιχείο της στήλης και στη συνέχεια διαιρεί το αποτέλεσμα με το εύρος (η διαφορά μεταξύ της μέγιστης και της ελάχιστης τιμής) ώστε τα δεδομένα να προσαρμόζονται σε κλίμακα [0, 1].

```
In [69]: scaler = MinMaxScaler()
df_cluster = scaler.fit_transform(df_cluster[['Poverty_Rate', 'Total_Crimes']])
```

Στη συνέχεια, τρέχουμε τον αλγόριθμο **K-Means** ο οποίος ομαδοποιεί τα datapoints σε συστάδες ελαχιστοποιώντας την ευκλείδεια απόστασή τους. Έπειτα από δοκιμές, κρίναμε ότι ο κατάλληλος αριθμός συστάδων είναι 3.

```
In [70]: from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans.fit(df_cluster)
```

Οπτικοποιούμε τα αποτελέσματα με Scatterplot. Το κάθε datapoint αντιπροσωπεύει ένα community area και το χρώμα του υποδεικνύει σε ποιο cluster ανήκει. Είναι σημαντικό να σημειωθεί ότι η μεταβλητή **Poverty Rate** δείχνει το ποσοστό των νοικοκυριών κάτω από το όριο της φτώχειας, συνεπώς όσο μεγαλύτερη τιμή έχει τόσο περισσότερο “φτωχό” θεωρείται το community area.



Συστάδα 1

Χαρακτηρίζεται από κοινοτικές περιοχές με χαμηλά ποσοστά φτώχειας και χαμηλά έως μεσαία επίπεδα εγκληματικότητας.

Συστάδα 2

Αντιπροσωπεύει κοινοτικές περιοχές με χαμηλά προς μεσαία ποσοστά φτώχειας και μεσαία έως υψηλά επίπεδα εγκληματικότητας.

Συστάδα 3

Αποτελείται από κοινοτικές περιοχές με σχετικά υψηλότερα ποσοστά φτώχειας καλύπτοντας όλα τα επίπεδα εγκληματικότητας.

Σημείωση: Μεσαία επίπεδα εγκληματικότητας θεωρούνται αυτά στο εύρος 0.2 - 0.4 και όχι στο μέσον του plot (0.4 – 0.6).

5.1.1 Anomaly detection (Outlier analysis)

Στο παραπάνω scatterplot παρατηρούμε ότι ένα datapoint που ανήκει στην [2η συστάδα](#) απέχει κατά πολύ από όλα τα υπόλοιπα.

Αποφασίσαμε, λοιπόν, να εκτελέσουμε τον αλγόριθμο **Local Outlier Factor (LOF)**, ώστε να εντοπίσουμε “outlier” community areas που αποκλίνουν από τις γενικές παρατηρήσεις για τις περιοχές του Chicago όσον αφορά το ποσοστό φτώχειας και τα συνολικά εγκλήματα.

```
In [18]: from sklearn.neighbors import LocalOutlierFactor
clf = LocalOutlierFactor(n_neighbors=3, contamination=0.1)

# Fit the model and predict outliers
y_pred = clf.fit_predict(df_cluster)

# Negative outlier factor (the lower, the more likely the sample is an outlier)
X_scores = clf.negative_outlier_factor_

# Add the results to the DataFrame
df_cluster['outlier'] = y_pred # -1 for outliers, 1 for inliers
df_cluster['outlier_score'] = X_scores

outliers = df_cluster[df_cluster['outlier'] == -1]

print(outliers)
```

Ορίζουμε την παράμετρο **n_neighbors = 3**, ώστε το local density κάθε datapoint να συγκρίνεται με αυτό των 3 πλησιέστερων γειτόνων του και το **contamination = 0.1** ώστε ο αλγόριθμος να υποθέτει ότι το 10% των datapoints είναι outliers.

Το **LOF** προβλέπει εάν κάθε datapoint είναι ακραίο αποδίδοντάς του την τιμή -1 ή εσωτερικό με τιμή 1.

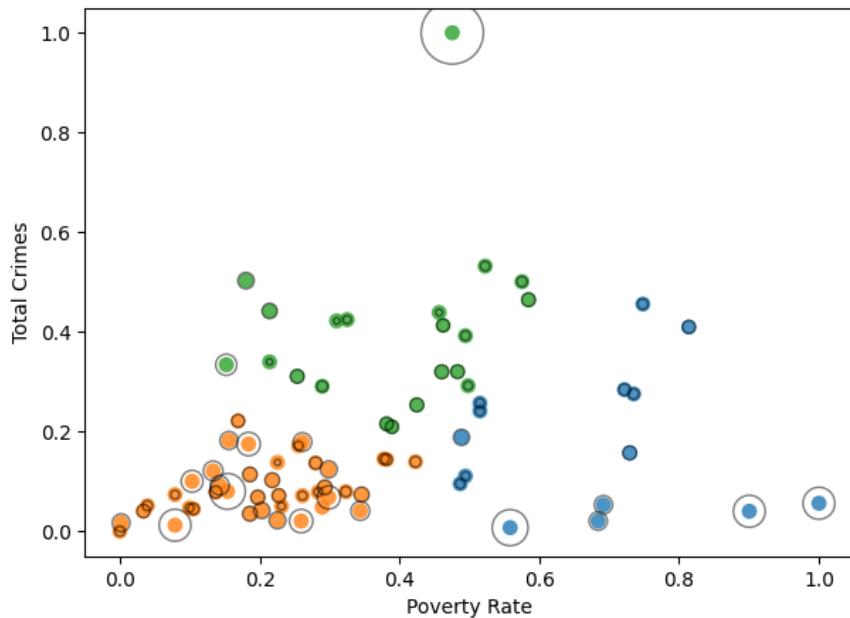
Τέλος, εκτυπώνουμε τα αποτελέσματα των outliers:

	Poverty_Rate	Total_Crimes	Cluster	outlier	outlier_score
11	0.078947	0.011987	1	-1	-1.754817
15	0.184211	0.174824	1	-1	-1.363182
24	0.475564	1.000000	2	-1	-4.189103
36	0.900376	0.039973	0	-1	-1.757419
46	0.558271	0.007118	0	-1	-1.991073
47	0.154135	0.079830	1	-1	-1.966865
53	1.000000	0.056041	0	-1	-1.755992
54	0.259398	0.020311	1	-1	-1.361803

Όσο χαμηλότερο είναι το **outlier_score** τόσο πιθανότερο είναι ένα datapoint να είναι outlier.

Παρατηρούμε ότι το **outlier_score** του datapoint 24 (που αντιπροσωπεύει το community area 25) είναι κατά πολύ μικρότερο των υπολοίπων.

Οπτικοποιούμε τα αποτελέσματα πάνω στο προηγούμενο scatterplot βάζοντας έναν κύκλο διαμέτρου ανάλογης του **outlier_score** πάνω στα datapoints που σύμφωνα με τον αλγόριθμο είναι outliers.



Φαίνεται ότι το datapoint 24 περιβάλλεται από κύκλο αισθητά μεγαλύτερο των υπολοίπων outliers.

Ερμηνεία αποτελέσματος

Ο αυξημένος αριθμός εγκλημάτων στο community area 25 (Austin) οφείλεται:

- Πρώτον, στο γεγονός ότι αποτελεί το τρίτο πολυπληθέστερο community area του Chicago και επομένως τα συνοικικά εγκλήματα είναι αυξημένα εφόσον λαμβάνονται υπόψιν ανεξάρτητα του πληθυσμού.
 - Δεύτερον, στο ότι αποτελεί κέντρο δραστηριότητας συμμοριών και επακολούθως υπάρχουν αυξημένα ποσοστά βίαιου εγκλήματος και ιδιαίτερα ανθρωποκτονιών.

5.2 Association rules

Στη συνέχεια, μέσω της μεθόδου Κανόνων Συσχέτισης θα ερευνήσουμε ποια εγκλήματα εμφανίζονται με μεγαλύτερη συχνότητα μαζί στην ίδια περιοχή.

Αρχικά, δημιουργούμε ένα dataframe με τα δεδομένα που μας ενδιαφέρουν, δηλαδή τα **Community Areas** και τα **Primary Types** των περιστατικών.

```
In [7]: sub_data = df_filtered[['Community Area', 'Primary Type']]
```

Φτιάχνουμε έναν πίνακα pivot με το **Community Area** ως index και τα **Primary Types** ως στήλες. Οι τιμές του πίνακα είναι True/False ανάλογα με το αν στη συγκεκριμένη περιοχή έχει διαπραχθεί ο συγκεκριμένος τύπος εγκλήματος.

```
In [9]: pivot_df = sub_data.pivot_table(index='Community Area', columns='Primary Type',
                                         aggfunc=lambda x: True, fill_value= False)
pivot_df
```

Χρησιμοποιώντας το pivot table που μόλις φτιάξαμε, εκτελούμε τον αλγόριθμο **Apriori**.

```
In [9]: from mlxtend.frequent_patterns import apriori  
frequent_itemsets = apriori(pivot_df, min_support=0.5, use_colnames=True)  
frequent_itemsets
```

Στον παρακάτω πίνακα παρουσιάζονται τα itemsets, δηλαδή εγκλήματα και οι συνδυασμοί εγκλημάτων, με συχνότητα εμφάνισης πάνω από το threshold που έχουμε ορίσει: **min_support = 0.5**.

Αυτό σημαίνει ότι τα συγκεκριμένα itemsets εμφανίζονται σε πάνω από τα μισά Community Areas του Chicago.

	support	itemsets
0	0.870130	(ARSON)
1	1.000000	(ASSAULT)
2	1.000000	(BATTERY)
3	1.000000	(BURGLARY)
4	1.000000	(CRIMINAL DAMAGE)
...
11796474	0.571429	(BATTERY, OFFENSE INVOLVING CHILDREN, BURGLARY...)
11796475	0.519481	(BATTERY, OFFENSE INVOLVING CHILDREN, BURGLARY...)
11796476	0.506494	(BATTERY, OFFENSE INVOLVING CHILDREN, BURGLARY...)
11796477	0.532468	(BATTERY, OFFENSE INVOLVING CHILDREN, BURGLARY...)
11796478	0.545455	(BATTERY, OFFENSE INVOLVING CHILDREN, BURGLARY...)

Έπειτα παράγουμε τα **Association Rules**.

```
In [9]: from mlxtend.frequent_patterns import association_rules  
num_itemsets=len(pivot_df)  
association_rules(frequent_itemsets, num_itemsets, metric="confidence", min_threshold=0.5)
```

	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

Ο πίνακας εμφανίζει τα rules με **confidence** μεγαλύτερο ίσο του **0.5**, τιμή που έχουμε ορίσει εμείς (**min_threshold = 0.5**).

Το **confidence** ενός rule, για παράδειγμα Assault -> Battery, είναι η πιθανότητα να δούμε το επακόλουθο (Battery) σε μια περιοχή δεδομένου ότι σε αυτήν την περιοχή έχει εμφανιστεί και το προηγούμενο (Assault). Η μέτρηση αυτή δεν είναι συμμετρική. Το **confidence** είναι 1 (μέγιστο) για ένα rule εάν τα δύο εγκλήματα εμφανίζονται πάντα μαζί (δηλαδή σε όλα τα community areas).

Μία άλλη μετρική που μας ενδιαφέρει είναι το **Lift** που χρησιμοποιείται για να μετρήσει πόσο πιο συχνά το προηγούμενο και το επακόλουθο ενός rule εμφανίζονται μαζί από ό,τι θα περιμέναμε αν ήταν στατιστικά ανεξάρτητα. Εάν στο προηγούμενο παράδειγμα το Assault και το Battery ήταν ανεξάρτητα, η βαθμολογία Lift θα ήταν ακριβώς 1.

Παρατηρούμε ότι στον πίνακα όλα τα **Lift** είναι μεγαλύτερα του 1, που σημαίνει ότι τα crime types των rules εμφανίζονται μαζί πιο συχνά από ό,τι θα αναμενόταν αν ήταν στατιστικά ανεξάρτητα.

5.3 Decision Tree

Τέλος, θα δημιουργήσουμε ένα μοντέλο Δέντρου Αποφάσεων με στόχο την πρόβλεψη αν ένα περιστατικό σύμφωνα με τον τύπο (**Primary Type**) και την περιοχή (**Community Area**) οδηγεί σε σύλληψη ή όχι (**Arrest**).

Για να χρησιμοποιήσουμε τα δεδομένα του IUCR (που αποτελεί την κωδικοποίηση του Primary Type), τα μετατρέπουμε σε αριθμητικές τιμές.

```
In [4]: from sklearn.preprocessing import LabelEncoder  
  
le = LabelEncoder()  
df['Primary Type'] = le.fit_transform(df['IUCR'])
```

```
In [5]: from sklearn.model_selection import train_test_split  
from sklearn import tree  
from sklearn.metrics import accuracy_score  
  
features = df[['Community Area', 'Primary Type']]  
target = df['Arrest']
```

Features: Είναι ο πίνακας χαρακτηριστικών, ο οποίος περιλαμβάνει τις στήλες **Community Area** και **Primary Type**.

Target: Είναι η μεταβλητή-στόχος, δηλαδή η στήλη **Arrest**, που δηλώνει αν πραγματοποιήθηκε σύλληψη με τιμές True/False.

```
In [6]: features_train, features_test, target_train, target_test = \  
train_test_split(features, target, test_size=0.2)
```

Η συνάρτηση **train_test_split** χωρίζει το σύνολο δεδομένων σε δύο μέρη:

Σετ εκπαίδευσης (**features_train, target_train**): Το 80% των δεδομένων, που χρησιμοποιείται για την εκπαίδευση του μοντέλου.

Σετ δοκιμής (**features_test, target_test**): Το 20% των δεδομένων, που χρησιμοποιείται για την αξιολόγηση της απόδοσης του μοντέλου.

```
In [7]: clf = tree.DecisionTreeClassifier(random_state=42, max_depth=4)  
clf = clf.fit(features_train, target_train)
```

DecisionTreeClassifier: Δημιουργεί ένα μοντέλο δέντρου αποφάσεων με δύο βασικές παραμέτρους:

random_state = 42: Εξασφαλίζει ότι το αποτέλεσμα είναι αναπαραγώγιμο κάθε φορά που εκτελείται ο κώδικας.

max_depth = 4: Περιορίζει το μέγιστο βάθος του δέντρου σε 4 επίπεδα, αποτρέποντας το μοντέλο από να κάνει overfit στα δεδομένα εκπαίδευσης.

fit: Εκπαιδεύει το μοντέλο χρησιμοποιώντας τα δεδομένα εκπαίδευσης (**features_train**, **target_train**).

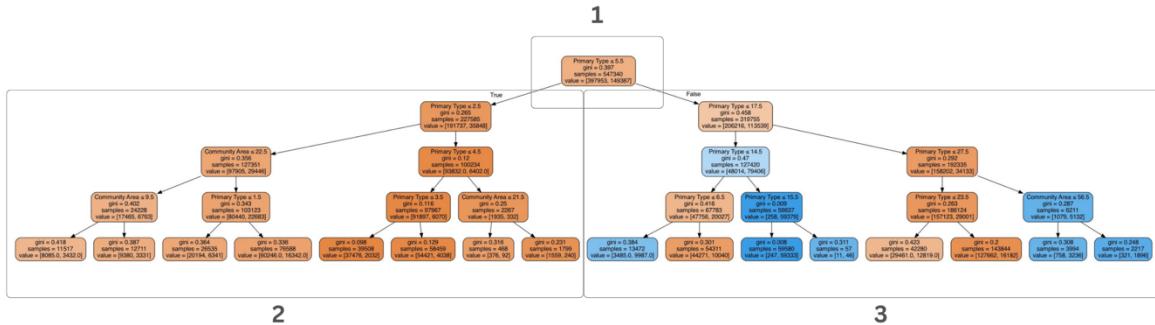
```
In [8]: target_pred = clf.predict(features_test)
```

Η μέθοδος **predict** χρησιμοποιεί το εκπαιδευμένο μοντέλο για να κάνει προβλέψεις σχετικά με το σετ δοκιμής (**features_test**). Το αποτέλεσμα αποθηκεύεται στη μεταβλητή **target_pred**.

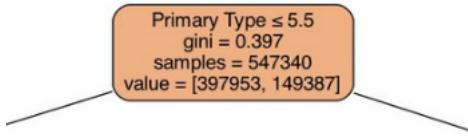
```
In [9]: accuracy = accuracy_score(target_test, target_pred)
        print("Accuracy: ", accuracy)
```

Accuracy: 0.8542039682829685

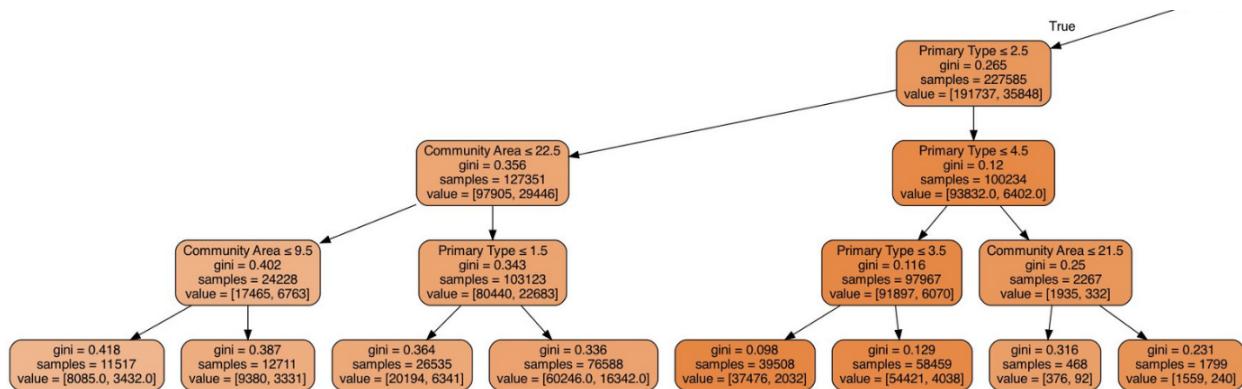
Η ακρίβεια του μοντέλου, που προκύπτει από τη σύγκριση των προβλέψεων με τις πραγματικές τιμές του test set, είναι 85%. Αυτό σημαίνει ότι το μοντέλο προβλέπει σωστά αν ένα περιστατικό οδηγεί σε σύλληψη ή όχι στο 85% των περιπτώσεων.



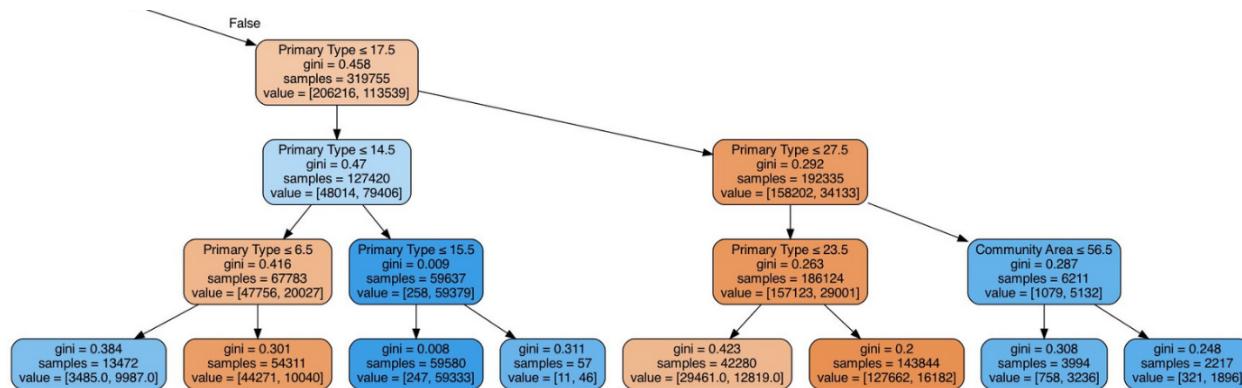
Απόσπασμα 1:



Απόσπασμα 2:



Απόσπασμα 3:



To **Gini Impurity** μετρά την πιθανότητα ένα τυχαία επιλεγμένο στοιχείο ενός συνόλου να προβλεφθεί λανθασμένα. Όσο μικρότερη είναι η τιμή του, τόσο καλύτερο είναι το split.