ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ
SCHOOL OF BUSINESS

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
DEPARTMENT OF MANAGEMENT SCIENCE & TECHNOLOGY

Μάθημα «Επιχειρηματική Ευφυΐα και Ανάλυση Μεγάλων Δεδομένων »
7ο Εξάμηνο

Θέμα εργασίας:

< CRIMES IN CHICAGO >

ΑΝΑΛΥΤΙΚΗ ΠΕΡΙΓΡΑΦΗ ΕΡΓΑΣΙΑΣ

Ομάδα Εργασίας **9**:
Λιάγγου Δέσποινα (8190094)
Φραγκιαδάκης Χριστόφορος (8190168)

<11/01/2023>

Περιεχόμενα

1. Περιγραφή του dataset

Το dataset αναφέρεται στα εγκλήματα που συνέβησαν στο Σικάγο μεταξύ των ετών 2019-2022. Παρακάτω φαίνεται ένα δείγμα των αρχικών δεδομένων:

	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...	Ward	Community Area	FBI Code	X Coordinate
0	12016034	JD193556	01/01/2020 12:00:00 AM	018XX N WINNEBAGO AVE	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	APARTMENT	False	False	...	32.0	22.0	11	1160263.0
1	12220321	JD430436	01/01/2020 12:00:00 AM	091XX S DREXEL AVE	1752	OFFENSE INVOLVING CHILDREN	AGGRAVATED CRIMINAL SEXUAL ABUSE BY FAMILY MEMBER	RESIDENCE	False	True	...	8.0	47.0	17	1184157.0
2	12013828	JD191019	01/01/2020 12:00:00 AM	044XX S LAVERGNE AVE	0281	CRIMINAL SEXUAL ASSAULT	NON-AGGRAVATED	APARTMENT	False	False	...	22.0	56.0	02	1143770.0
3	12019692	JD197444	01/01/2020 12:00:00	032XX N LINCOLN	1153	DECEPTIVE	FINANCIAL IDENTITY	APARTMENT	False	False	...	47.0	6.0	11	1164983.0

Y Coordinate	Year	Updated On	Latitude	Longitude	Location
1912391.0	2020	03/26/2020 03:45:12 PM	41.915306	-87.686639	(41.915306069, -87.686639247)
1844395.0	2020	12/19/2020 03:45:59 PM	41.728192	-87.600985	(41.728192429, -87.600985433)
1874726.0	2020	03/28/2020 03:47:02 PM	41.812274	-87.748177	(41.81227369, -87.748176594)
1921507.0	2020	04/01/2020 03:50:17	41.940222	-87.669039	(41.940221932,

INITIAL DATASET

Επεξήγηση στηλών:

→ ID: Μοναδικός κωδικός του κάθε εγκλήματος (της κάθε εγγραφής του συνόλου δεδομένων)
→ Case Number: Μοναδικός κωδικός του κάθε περιστατικού σύμφωνα με την αστυνομία
→ Date: Ημερομηνία και ώρα του κάθε εγκλήματος
→ Block: Διεύθυνση στην οποία έλαβε χώρα το έγκλημα
→ IUCR: Κωδικοποίηση του τύπου εγκλήματος
→ Primary type: Πρωτεύουσα περιγραφή του τύπου εγκλήματος (IUCR)
→ Description: Δευτερεύουσα περιγραφή του τύπου εγκλήματος (IUCR)
→ Location Description: Το μέρος όπου έγινε το περιστατικό
→ Arrest:Αν υπήρξε σύλληψη ή όχι
→ Domestic: Αν το έγκλημα σχετίζεται με ενδοοικογενειακή βία
→ Ward: Περιφέρεια όπου συνέβη το περιστατικό
→ Community Area: Ευρύτερη περιοχή που συνέβη το περιστατικό
→ FBI Code: Ταξινόμηση εγκλήματος από το FBI
→ X coordinate: Συντεταγμένη τοποθεσίας (X'X)
→ Y coordinate: Συντεταγμένη τοποθεσίας (Y'Y)
→ Year: Έτος εγκλήματος
→ Updated On: Ημερομηνία και ώρα τελευταίας ενημέρωσης εγγραφής
→ Latitude: Γεωγραφικό πλάτος τοποθεσίας

→ Longitude: Γεωγραφικό μήκος τοποθεσίας
→ Location: Συντεταγμένες τοποθεσίας

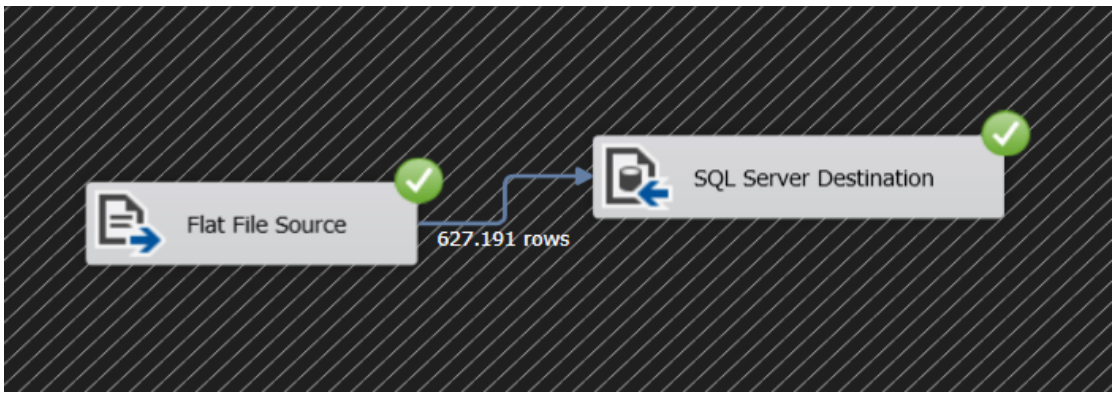
Συγκεκριμένα, το αρχικό σύνολο δεδομένων περιείχε 639.406 εγγραφές – εγκλήματα (ID), τα οποία συνέβησαν ανάμεσα σε 32.624 διευθύνσεις (Block) του Σικάγο από τις 1/1/20 μέχρι και 12/7/22. Η κάθε διεύθυνση, ανήκει σε μία από τις 78 ευρύτερες περιοχές του Σικάγο (Community Area). Επιπλέον, υπάρχουν 177 τόποι που αφορούν το πού έλαβε μέρος το έγκλημα (πχ δρόμο, σπίτι, σχολείο, αμάξι, τράπεζα κλπ.) Όσον αφορά τον τύπο του εγκλήματος, υπάρχουν 33 μοναδικές ονομασίες (Primary type) που το προσδιορίζουν.

2. Διαδικασία ETL – Κατασκευή Star Schema/Κύβου

Σε πρώτη φάση πραγματοποιήθηκε η διαδικασία του Data Cleaning, ώστε να καθαριστούν τα δεδομένα και να είναι έγκυρα και αξιόπιστα προκειμένου να επεξεργαστούν και να αναλυθούν. Για τη διαδικασία αυτή χρησιμοποιήθηκε το εργαλείο Visual Studio Code και η βιβλιοθήκη της γλώσσας Python pandas. Το αρχικό σύνολο δεδομένων αποτελούνταν από 3 αρχεία csv, ένα για κάθε έτος, τα οποία συγχωνεύθηκαν και στην συνέχεια επεξεργάστηκαν. Συγκεκριμένα, ελέγχθηκαν τα δεδομένα για null τιμές, για έγκυρες τιμές και για διπλές εγγραφές. Επίσης ελέγχθηκαν και δεδομένα που αποτελούσαν τυχόν "outliers" ως προς την τοποθεσία. Έτσι, τελικά διαγράφθηκε ένα ποσοστό περίπου 2% των δεδομένων. Ακολούθως, διαγράφθηκαν στήλες που ήταν περιττές για την ανάλυση που θα ακολουθήσει (Case Number, Description, Ward, FBI Code, X coordinate, Y coordinate, Updated On, Location). Επιπλέον, για λόγους διευκόλυνσης της μετέπειτα ανάλυσης, προστέθηκαν κάποιες στήλες που ήταν απαραίτητες, όπως τα ονόματα των Community Areas, καθώς και οι 9 ευρύτερες γεωγραφικές περιοχές του Σικάγο (Region), όπου ανήκει καθεμία από τις 77 Community Areas. Τα παραπάνω βρέθηκαν σε διαφορετικές πηγές δεδομένων και συγχωνεύθηκαν με τα ήδη υπάρχοντα. Τέλος, για την ημερομηνία χωρίστηκαν σε ξεχωριστές στήλες η ώρα, ο μήνας σε αριθμό και ο μήνας ονομαστικά.

Στη συνέχεια προκειμένου να δημιουργηθεί ένα πολυδιάστατο μοντέλο σε μία αποθήκη δεδομένων έγινε χρήση των εργαλείων SSIS (SQL Server Integration Services) και SSAS (SQL Server Analysis Services) χρησιμοποιώντας την εφαρμογή Visual Studio. Επίσης, προκειμένου τα παραπάνω να συνδεθούν σε μία βάση δεδομένων, χρησιμοποιήθηκε η εφαρμογή Microsoft SQL Server Management Studio.

Αναλυτικότερα, τα μεταμορφωμένα δεδομένα εισήχθησαν στο εργαλείο SSIS και σχεδιάστηκε μία ροή με σκοπό να κρατάει ενημερωμένη την αποθήκη δεδομένων όποτε τα δεδομένα ενημερώνονται και τροποποιούνται. Παρακάτω φαίνεται αρχικά το δημιουργημένο Data Flow όπου διαβάζει τα δεδομένα από το αρχείο (Flat File Source) και τα μεταφέρει στη βάση δεδομένων (SQL Server Destination) δημιουργώντας έναν προσωρινό πίνακα σε γλώσσα SQL που περιέχει όλα τα δεδομένα του αρχείου.

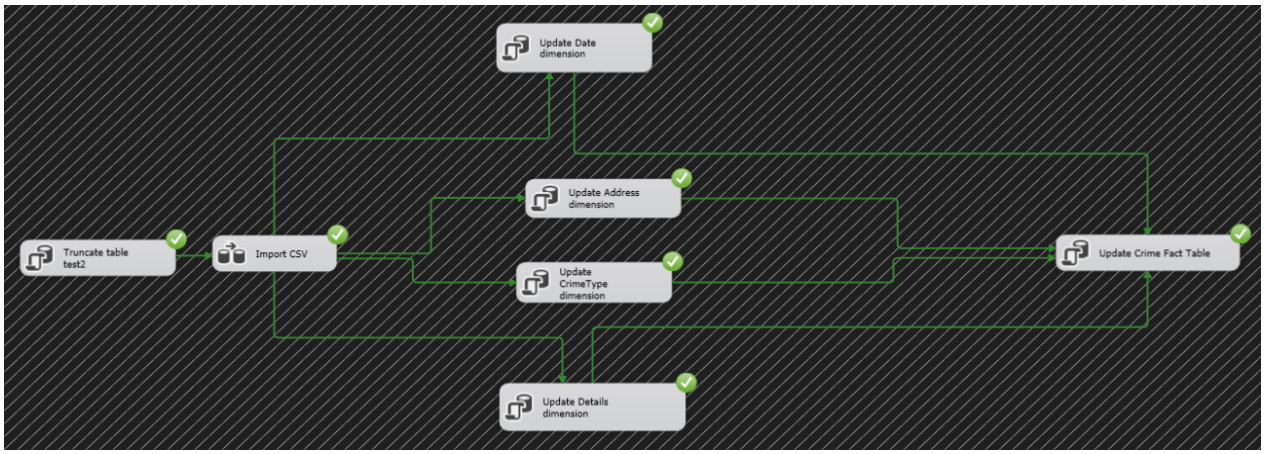


DATA FLOW – VISUAL STUDIO

Έπειτα έγινε μία σειρά από επιμέρους εργασίες ώστε να ελέγχεται η ροή και να ενημερώνονται τα δεδομένα σωστά. Οι εργασίες αυτές απεικονίζονται στο Control Flow, στον παρακάτω πίνακα.
Αρχικά, προστέθηκε το task "Truncate table test2" όπου κάθε φορά που ενημερώνεται και διαβάζεται το αρχείο, διαγράφει τα δεδομένα του προσωρινού πίνακα που δημιουργείται (test2) ώστε αυτά να μην πολλαπλασιάζονται κάθε φορά που τρέχει η ροή.

Στη συνέχεια, δημιουργήθηκαν στην αποθήκη δεδομένων οι πίνακες διαστάσεων (dimension tables) και ο κεντρικός πίνακας (fact table) καθώς και μία σειρά από SQL queries τα οποία εκτελούνται μέσω των tasks "Update dimension table or fact table" και η δουλειά τους είναι να

ενημερώνουν και να γεμίζουν αρχικά τους dimension πίνακες και στη συνέχεια τον κεντρικό fact πίνακα.



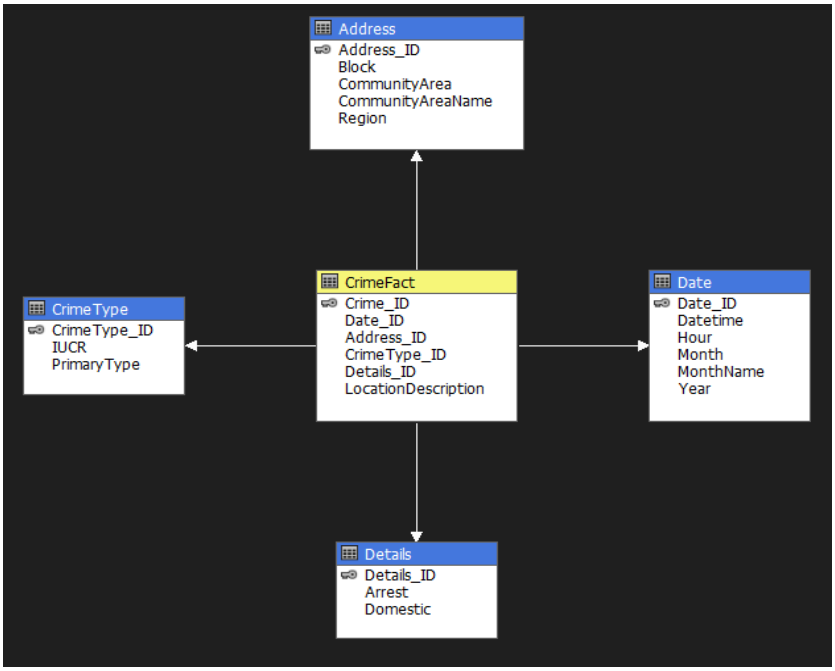
CONTROL FLOW – VISUAL STUDIO

Ακολούθως, κατασκευάστηκε ένα πολυδιάστατο μοντέλο στον OLAP Server (SSAS), το οποίο αποτελεί έναν κύβο και σχηματίζει το star schema.

Πιο συγκεκριμένα, αποτελείται από τα dimensions

a. Address: περιέχει τη διεύθυνση, τον αριθμό του Community Area στο οποίο ανήκει, το όνομα αυτού του Area, καθώς και σε ποιο από τα 9 Region ανήκει.
b. Date: περιέχει όλα τα στοιχεία της ημερομηνίας που καταγράφηκε το έγκλημα (το String Datetime που σπάει σε επιμέρους στήλες Hour, Month, Month ολογράφως, Year)
c. Details: περιέχει τα attributes Domestic και Arrest που μπορούν να πάρουν μόνο την τιμή True ή False
d. CrimeType: περιέχει το IUCR που δείχνει την κωδικοποίηση του εγκλήματος, και τον τύπο του εγκλήματος σε String

Και από τον Crime Fact ο οποίος είναι ο Fact table του κύβου, και περιέχει τα κλειδιά όλων των dimensions καθώς και το attribute “Location Description”



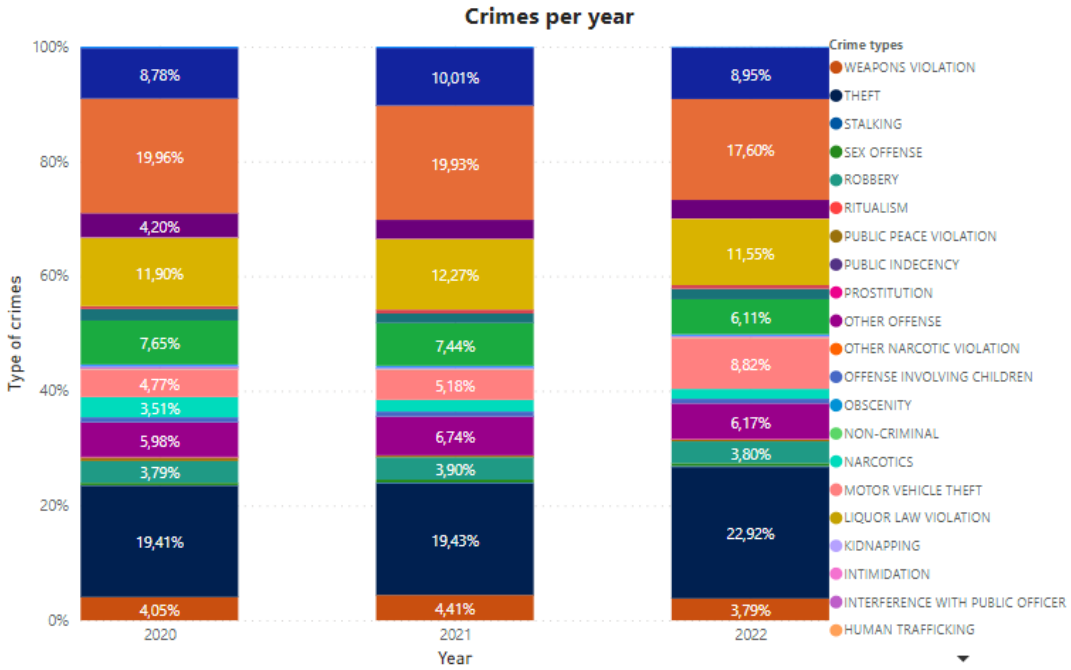
STAR SCHEMA – VISUAL STUDIO

3. Οπτική αναπαράσταση των δεδομένων

Προκειμένου να οπτικοποιηθούν τα δεδομένα, χρησιμοποιήθηκε το εργαλείο power BI αξιοποιώντας τον κύβο που ορίστηκε προηγουμένως. Σε ορισμένες περιπτώσεις κάποια διαγράμματα δημιουργήθηκαν μέσα από το Jupyter Notebook χρησιμοποιώντας τη βιβλιοθήκη της γλώσσας Python pandas.

Παρακάτω αναλύονται κάποια descriptives όπου αναδεικνύεται η συνολική εικόνα των δεδομένων.

- Ποια εγκλήματα συμβαίνουν κάθε χρόνο και σε τι βαθμό;



Top 5 crimes of 2020

2020	40157	THEFT
2020	15830	DECEPTIVE PRACTICE
2020	24629	CRIMINAL DAMAGE
2020	41310	BATTERY
2020	18171	ASSAULT
Total	**140097**	

Top 5 crimes of 2021

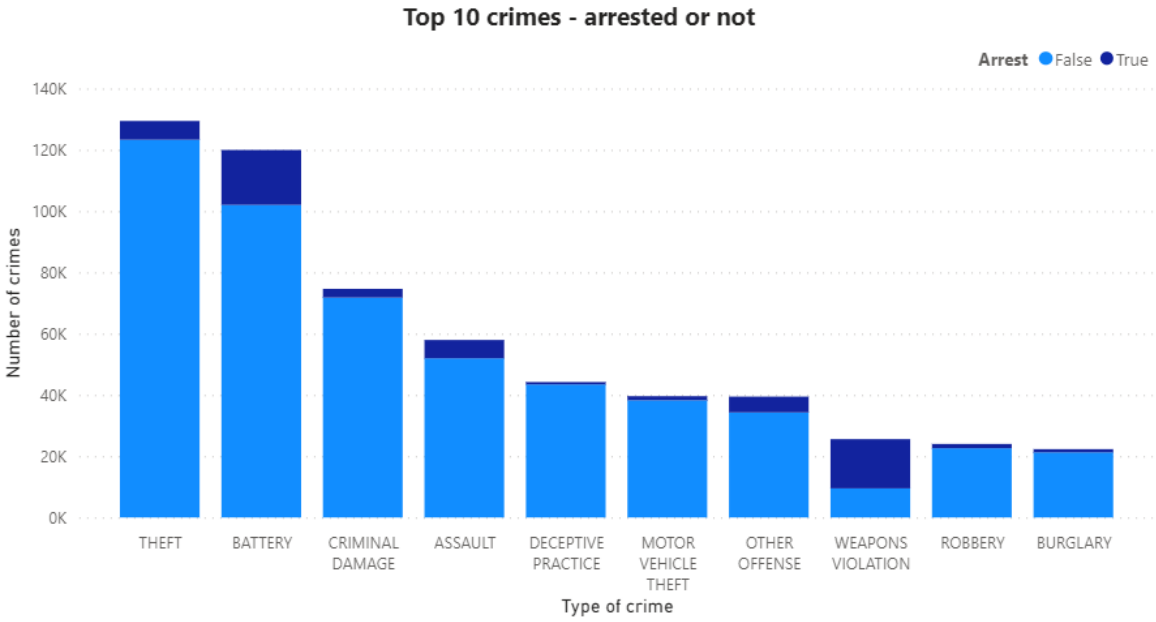
2021	39195	THEFT
2021	15014	DECEPTIVE PRACTICE
2021	24759	CRIMINAL DAMAGE
2021	40206	BATTERY
2021	20203	ASSAULT
Total	**139377**	

Top 5 crimes of 2022

2022	50075	THEFT
2022	19269	MOTOR VEHICLE THEFT
2022	25237	CRIMINAL DAMAGE
2022	38447	BATTERY
2022	19562	ASSAULT
Total	**152590**	

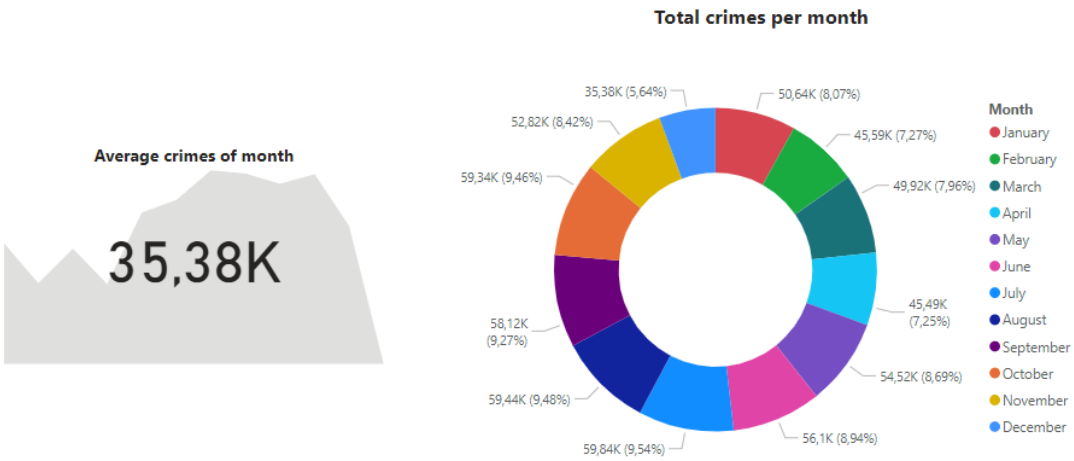
Παρατηρήσεις: Και στα 3 έτη τα περισσότερα εγκλήματα παρατηρούνται να είναι η κλοπή (Theft), η εγκληματική ενέργεια (Criminal Damage), η βιαιοπραγία (Battery) και η κακοποίηση (Assault). Στο μόνο τύπο εγκλήματος που διαφοροποιούνται τα έτη είναι η παραπλάνηση (Deceptive practice) που βρίσκεται 2^{η} στην κατάταξη του 2020 και 2021, ενώ το 2022 αυτό φαίνεται να εξασθενεί και να παίρνει τη θέση του η κλοπή μηχανών (Motor vehicle theft). Τα κορυφαία 5 εγκλήματα για καθένα από τα έτη αποτελούν το 68,8% του συνόλου των εγκλημάτων (22,3% για το 2020, 22,2% για το 2021 και 24,3% για το 2022). Σημειώνεται ότι και τα 3 έτη έχουν περίπου τον ίδιο συνολικό αριθμό εγκλημάτων. Συγκεκριμένα το 2020 καταλαμβάνει το 32,9% του συνόλου των δεδομένων, το 2021 καταλαμβάνει το 32,2% και το 2022 το 34,9%. Οπότε έχουν την ίδια βαρύτητα στα αποτελέσματα.

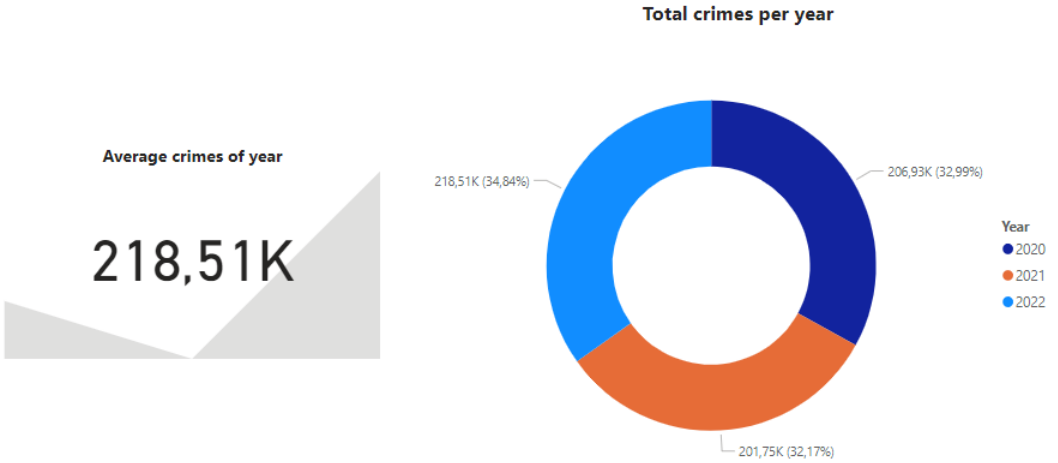
- Ποια είναι τα συχνότερα 10 εγκλήματα και σε τι ποσοστό αυτά οδηγούν σε συλλήψεις;



Παρατηρήσεις: Σε γενικές γραμμές φαίνεται πως δεν γίνονται πολλές συλλήψεις σε σχέση με τον αριθμό των εγκλημάτων. Οι περισσότερες συλλήψεις γίνονται κατά τις επιθέσεις με όπλο (Weapons violation) (62,8%) και ακολουθούν η βιαιοπραγία (Battery) (14,9%) και η βίαιη επίθεση – προσβολή (Assault) (10,3%). Οι λιγότερες συλλήψεις παρατηρούνται στην παραπλάνηση (Deceptive Practice) (1,7%) και στην κλοπή μηχανών (Motor vehicle theft) (3,3%).

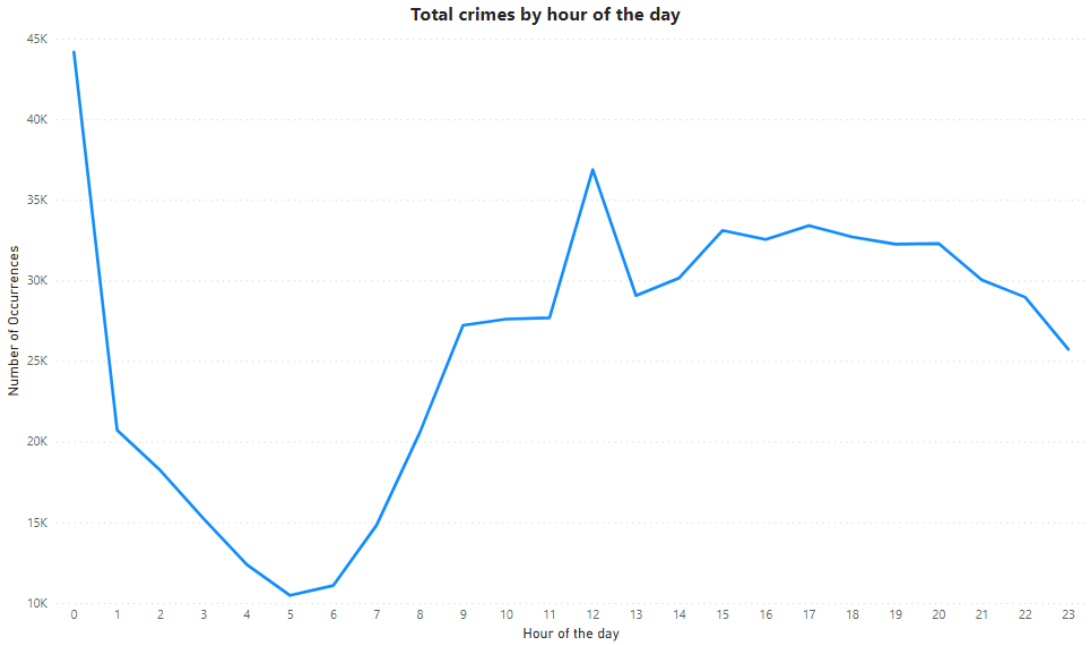
➢ Πόσα εγκλήματα γίνονται κατά μέσο όρο τον μήνα και το έτος;





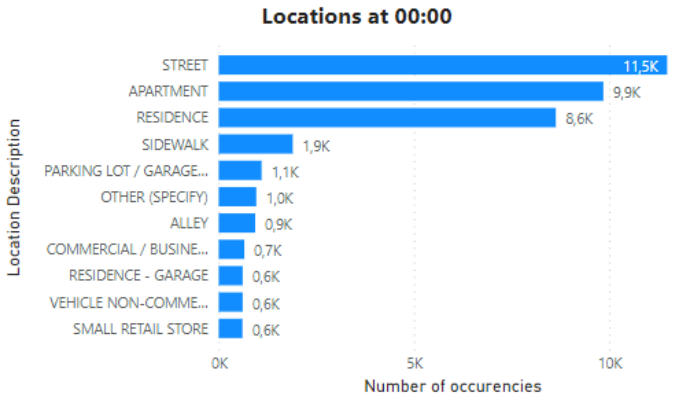
Παρατηρήσεις: Τον μήνα γίνονται κατά μέσο όρο περίπου 35 χιλιάδες εγκλήματα ενώ τον χρόνο περίπου 218 χιλιάδες. Επιπλέον, φαίνεται πως τα περισσότερα εγκλήματα γίνονται την περίοδο από Ιούλιο μέχρι Οκτώβριο. Τέλος, αύξηση παρατηρείται στα εγκλήματα το 2022.

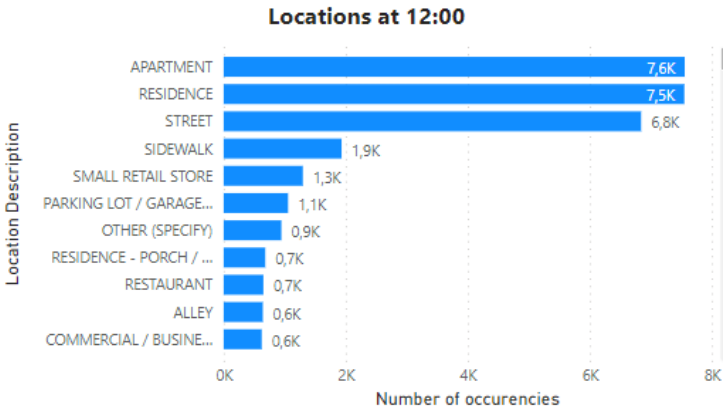
- Πόσα εγκλήματα συμβαίνουν κάθε ώρα της ημέρας;

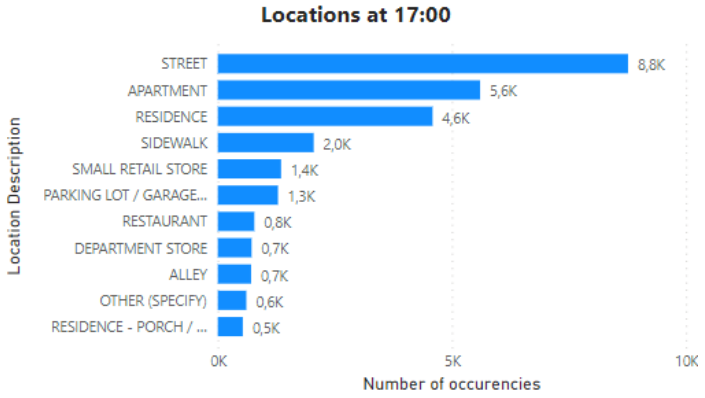


Παρατηρήσεις: Τα περισσότερα εγκλήματα φαίνεται να συμβαίνουν στις 00:00 τα ξημερώματα (7%) όπως επίσης φαίνεται να αυξάνονται σημαντικά στις 12:00 το μεσημέρι (5,8%) και στις 17:00 το απόγευμα (5,3%). Τα λιγότερα εγκλήματα γίνονται στις 5 τα ξημερώματα (1,6%).

- Σε ποιες τοποθεσίες γίνονται εγκλήματα αυτές τις «επικίνδυνες» ώρες της ημέρας;

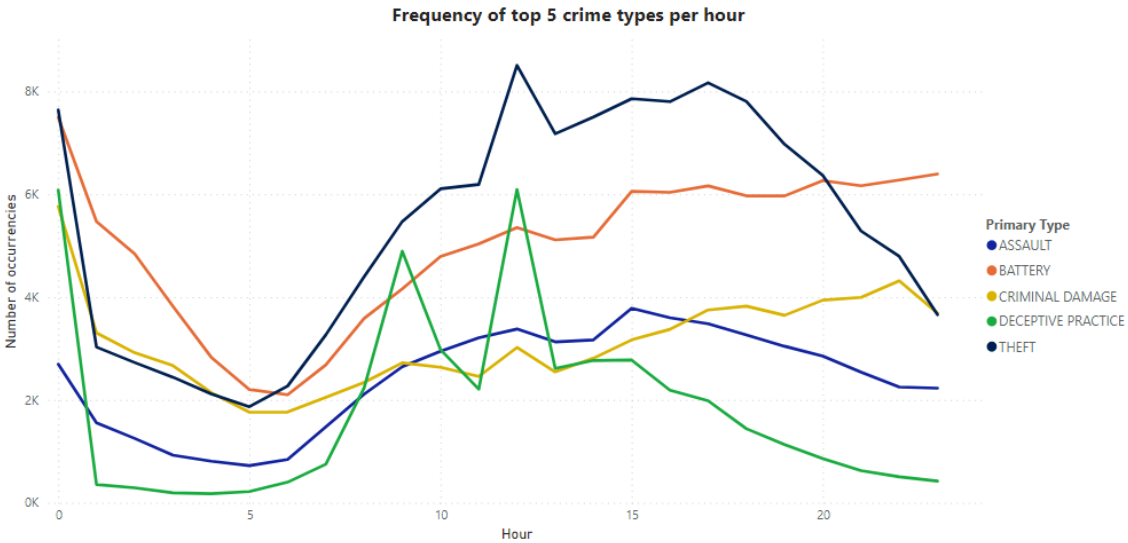






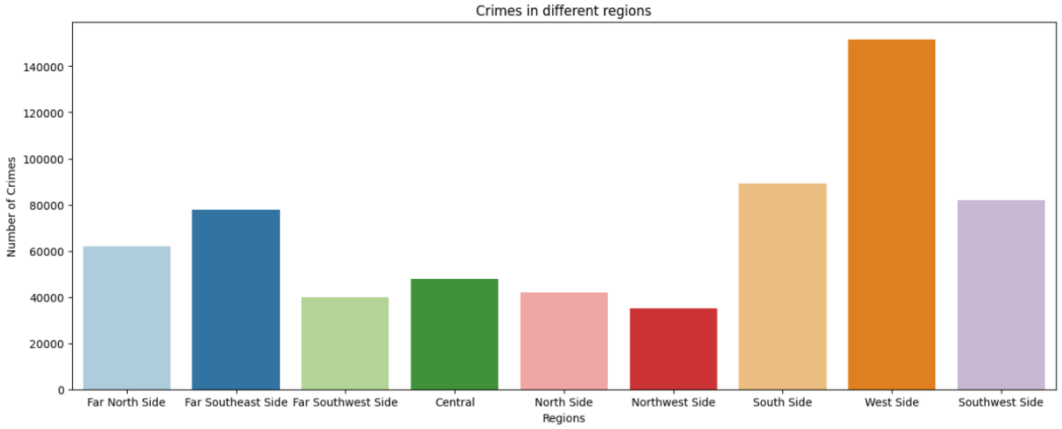
Παρατηρήσεις: Στις 00:00 τα ξημερώματα, στις 12:00 το μεσημέρι και στις 17:00 το απόγευμα τα περισσότερα εγκλήματα λαμβάνουν χώρα στον δρόμο, σε διαμέρισμα και σε κατοικία. Επίσης ένα μικρό ποσοστό αλλά όχι αμελητέο, διαδραματίζεται στο πεζοδρόμιο, σε γκαράζ και σε μικρά καταστήματα λιανικής.

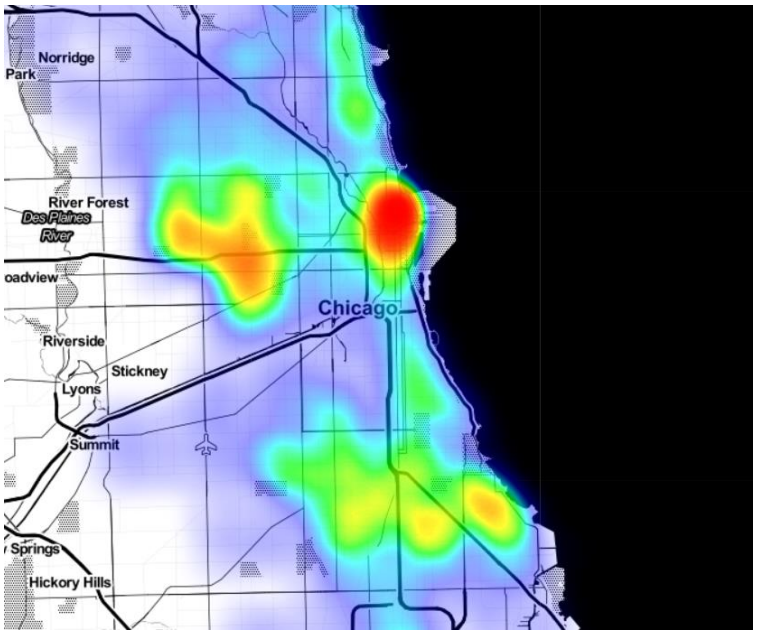
- Παρατηρώντας τα 5 κορυφαία εγκλήματα κάθε ώρα, ποια είναι η συμπεριφορά τους;



Παρατηρήσεις: Ενώ οι κλοπές (Theft) είναι τα πιο κοινά εγκλήματα κατά τη διάρκεια της ημέρας, συμβαίνουν πιο συχνά μεταξύ αργά το πρωί μέχρι και το απόγευμα. Τα ξημερώματα και το βράδυ μειώνονται αισθητά. Αυτό βγάζει νόημα, καθώς οι περισσότερες κλοπές αυτές τις ώρες γίνονται σε κατοικίες και διαμερίσματα (όπως είδαμε παραπάνω) εξαιτίας της απουσίας των ανθρώπων από τα σπίτια τους. Από την άλλη, από το απόγευμα μέχρι αργά το βράδυ φαίνεται να αυξάνονται τα φαινόμενα βιαιοπραγίας (Battery) και εγκληματικής ενέργειας (Criminal Damage). Τις πρωινές ώρες και νωρίς το μεσημέρι η παραπλάνηση (Deceptive practice) αυξάνεται σε μεγάλο βαθμό, ενώ τέλος η κακοποίηση (Assault) παρατηρείται κυρίως τις μεσημεριανές ώρες.

- Σε ποιες περιοχές του Σικάγο καταγράφονται τα περισσότερα εγκλήματα;





Παρατηρήσεις: Το παραπάνω heat map επιβεβαιώνει τη διασπορά του αριθμού των εγκλημάτων στο δυτικό και στο νότιο κομμάτι της πόλης, όμως δείχνει και μία επιπλέον πληροφορία. Εάν εστιάσουμε στην πιο έντονη περιοχή, θα βρούμε το North State Street, το οποίο είναι ο εμπορικότερος δρόμος του Σικάγο και βρίσκεται στο κέντρο (στον συγκεκριμένο δρόμο πάνω από 70% των εγκλημάτων που καταγράφονται είναι κλοπή(Theft)). Ο συγκεκριμένος δρόμος είναι αυτός στον οποίο έχουν γίνει οι περισσότερες καταγραφές των εγκλημάτων.

- Ποια γειτονιά του Σικάγο είναι η πιο επικίνδυνη;

Για να δοθεί μία ορθολογική απάντηση σε ένα τέτοιο ερώτημα, δεν πρέπει να αξιολογήσουμε μόνο τον αριθμό των εγκλημάτων που έχουν συμβεί. Ο παρακάτω πίνακας δείχνει ότι η γειτονιά με τα περισσότερα εγκλήματα είναι το Austin, και θεωρείται από τις πιο κακόφημες γειτονιές του Σικάγο.

Community Area Name	Number of Crime Records
Austin	11435
Near North Side	9628
Near West Side	8624

Εάν όμως συμπεριλάβουμε στην ανάλυση μας και τον πληθυσμό της κάθε γειτονιάς, μπορούμε να δημιουργήσουμε μία μετρική που υποδεικνύει το ποσοστό εγκληματικότητας για κάθε γειτονιά (Πληθυσμός/Πλήθος εγκλημάτων).

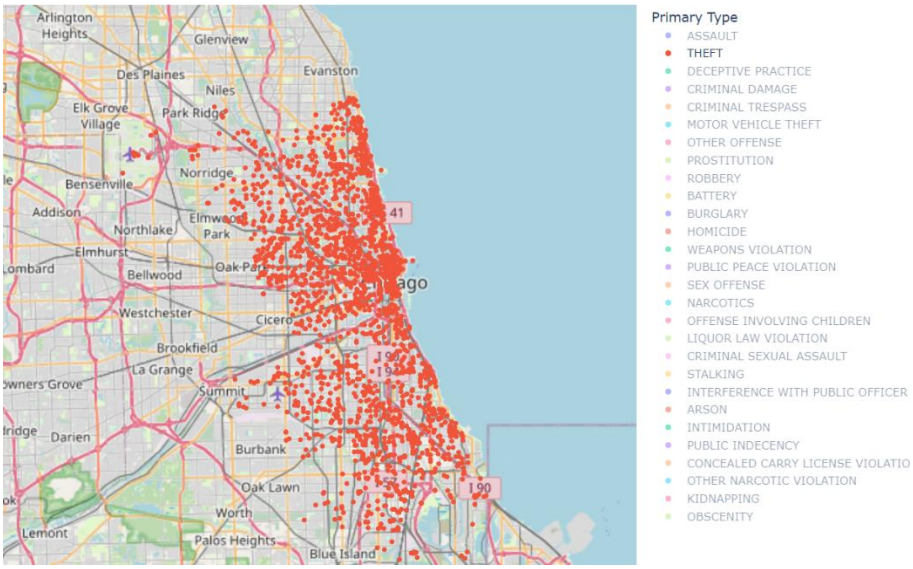
index	Name	Population	Crime Records	Crime Rate
37	Fuller Park	2567	587	0.228672
26	West Garfield Park	17433	3571	0.204841
68	Englewood	24369	4360	0.178916
44	Chatham	31710	5615	0.177073
29	North Lawndale	34794	6141	0.176496
69	Greater Grand Crossing	31471	5534	0.175844
27	East Garfield Park	19992	3455	0.172819
40	Washington Park	12707	2061	0.162194
54	Riverdale	7262	1155	0.159047
67	West Englewood	29647	4444	0.149897

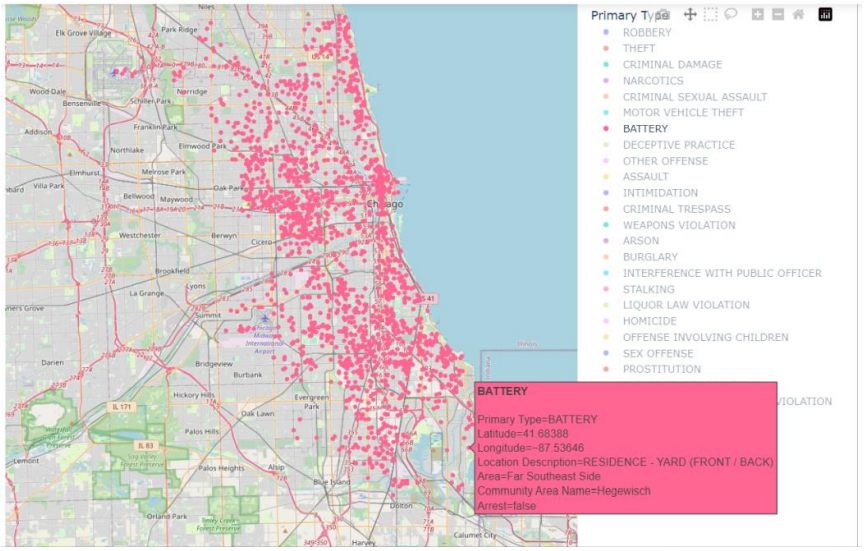
Εάν κρίνουμε με βάση αυτή τη μετρική, θα δούμε ότι το Fuller Park είναι το υψηλότερο στη σχετική λίστα, με εγκληματικότητα 22,8%, δηλαδή 22.867 θύματα ανά 100.000 ανθρώπους. Επιπλέον, αν θέλουμε να αναζητήσουμε το Austin με βάση την εγκληματικότητα, θα το βρούμε πολύ χαμηλότερα και συγκεκριμένα στην 22η θέση από τις 77. Ακόμη, αν κοιτάξουμε στις τελευταίες θέσεις επιβεβαιώνεται ότι το βόρειο κομμάτι του Σικάγο είναι το ασφαλέστερο, καθώς έχει γειτονιές με πολύ μικρό αριθμό εγκλημάτων και βαθμό εγκληματικότητας.

Name	Population	Crime Records	Crime Rate	Crime per 100.000 People	Area
Dunning	43147	1411	0.032702	3270.215774	Northwest Side
Norwood Park	38303	1142	0.029815	2981.489701	Far North Side
Forest Glen	19596	556	0.028373	2837.313737	Far North Side
Mount Greenwood	18628	493	0.026466	2646.553575	Far Southwest Side
Edison Park	11525	274	0.023774	2377.440347	Far North Side

- Τι συμβαίνει με τα εγκλήματα το 2022;

  Όπως αναφέρθηκε πιο πάνω η κλοπή και η βιαιοπραγία είναι με διαφορά σε σχέση με τις υπόλοιπες, τα πιο συχνά είδη εγκλημάτων. Σε ποιες περιοχές όμως εμφανίζονται έντονα αυτά τα εγκλήματα;





Παρατηρήσεις: Παρατηρώντας το διάγραμμα, βλέπουμε ότι η κλοπή (Theft), είναι πιο έντονη στο δυτικό κομμάτι της πόλης, κυρίως στο κέντρο του και στη συνέχεια βορειότερα. Αυτό μπορεί να οφείλεται στο γεγονός ότι οι πιο εύπορες περιοχές του Σικάγο, βρίσκονται στο κέντρο και στη συνέχεια στο βόρειο μέρος του (Near North Side, Forest Glen, The Loop, Edison Park). Αντίθετα, το φαινόμενο της βιαιοπραγίας (Battery) κατανέμεται πιο ομαλά σε ολόκληρη την πόλη, με εξαίρεση το βόρειο μέρος. Σημαντικός παράγοντας στο τελευταίο είναι το γεγονός ότι στο βόρειο μέρος ο καταγεγραμμένος αριθμός των εγκλημάτων είναι πολύ μικρότερος σε σχέση με ολόκληρη την πόλη.

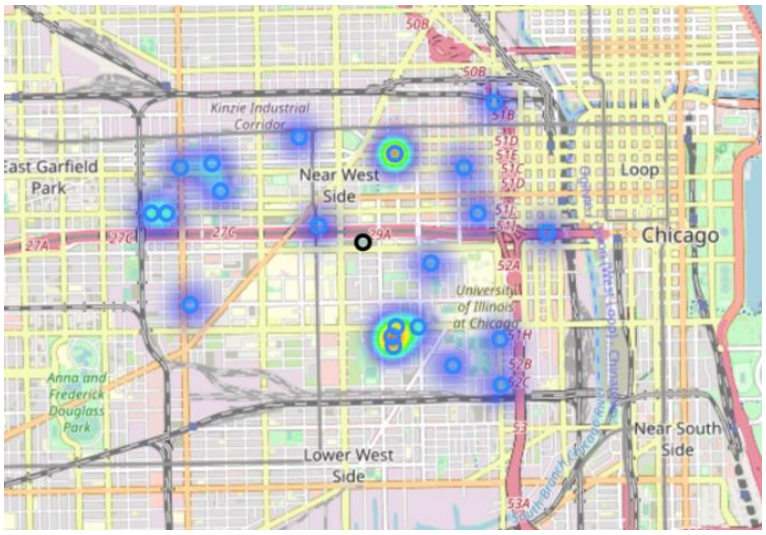
4. Σενάριο

Μετά την πρώτη ανάλυση λοιπόν των εγκλημάτων του Σικάγο, γεννάται η απορία: γιατί ενώ είναι ξεκάθαρη η έξαρση των φαινομένων σε συγκεκριμένες περιοχές, οι συλλήψεις είναι πολύ λίγες;

Η πιο εμφανής απάντηση είναι ότι πιθανότατα τα οχήματα της αστυνομίας βρισκόταν στο λάθος μέρος ανίκανα να φτάσουν έγκαιρα στον προορισμό τους.

Τι θα μπορούσε να γίνει ώστε να γνωρίζει η αστυνομία τον τρόπο που πρέπει να οργανώσει τους περιορισμένους πόρους της, ώστε να τους τοποθετήσει στο σωστό μέρος την κατάλληλη στιγμή;

Για τον σκοπό αυτό, χρειάζεται να προβούμε σε εκτενέστερη ανάλυση, και να εξετάσουμε εγκλήματα που έγιναν σε συγκεκριμένο μέρος σε συγκεκριμένη ημερομηνία. Στο παρακάτω παράδειγμα, βλέπουμε τα εγκλήματα σε κύκλο που συνέβησαν στη γειτονιά Near West Side (η οποία είναι 3η με τα περισσότερα εγκλήματα στο Σικάγο), την ημερομηνία 26 Ιουνίου 2022 (ημερομηνία στην οποία πραγματοποιήθηκε το Pride Parade στο Σικάγο). Το σημείο με τον μαύρο κύκλο δείχνει το κέντρο αυτών των εγκλημάτων. Εάν τότε, είχε τοποθετηθεί ένα περιπολικό σε εκείνο το σημείο, θα μπορούσε να ανταποκριθεί άμεσα στα επείγοντα περιστατικά.



Εάν λοιπόν σε αντίστοιχες μελλοντικές περιπτώσεις επανατοποθετηθούν με αυτόν τον τρόπο τα οχήματα της αστυνομίας (ορίζοντας δηλαδή ένα κεντρικό σημείο στο οποίο θα κάνουν την περιπολία τους, αξιολογώντας τα ιστορικά δεδομένα), θα βελτιστοποιηθεί η άμεση απόκριση τους και θα αποφευχθούν τυχόν άλλα εγκλήματα.

Παρακάτω, παρατίθεται ο κώδικας με τον οποίο βρίσκουμε τα συγκεκριμένα αυτά εγκλήματα, ενώ επίσης βρίσκουμε και προσθέτουμε στο heat map το κέντρο τους.

```
specificdate = result[result["Date"].str.contains("06/26/2022")]
specificdate = specificdate[specificdate["Community Area Name"] == 'Near West Side']

from folium import plugins
from folium.plugins import MarkerCluster

# Map points of events
m2 = folium.Map([41.8668,-87.6664], zoom_start=14)
for index, row in specificdate.iterrows():
    folium.CircleMarker([row['Latitude'], row['Longitude']],
                        radius=5,
                        popup=row['Primary Type'],
                        fill_color="#3db7e4",
                       ).add_to(m2)

dfmatrix = specificdate[['Latitude', 'Longitude']].values
# plot heatmap
m2.add_child(plugins.HeatMap(dfmatrix, radius=15))
m2
```

```
lat = []
long = []
for index, row in specificdate.iterrows():
    lat.append(row["Latitude"])
    long.append(row["Longitude"])
lat1=sum(lat)/len(lat)
lat2=sum(long)/len(long)
folium.CircleMarker([lat1,lat2],
                        radius=5,
                        popup="CENTER LOCATION",
                        color='black',
                        fill_color="#3db7e4",
                       ).add_to(m2)
m2
```

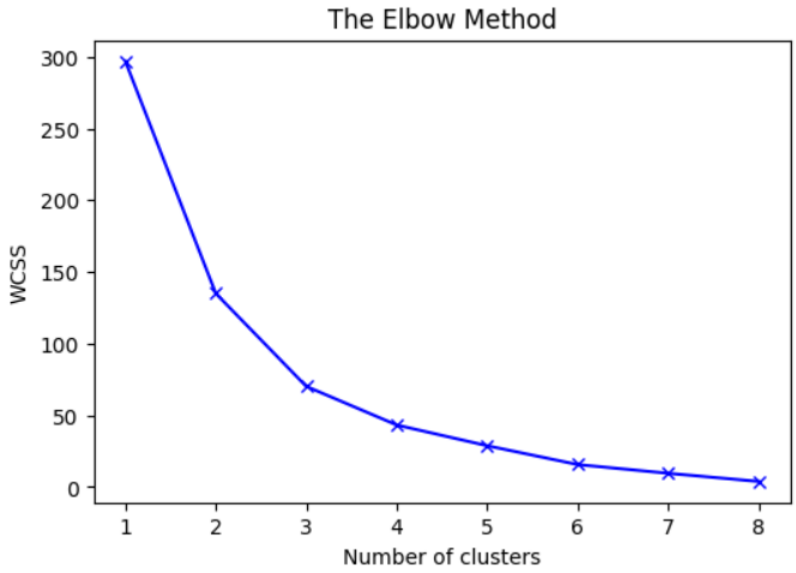
5. Υλοποίηση Data Mining

Προκειμένου να υλοποιηθούν τα παρακάτω Data Mining tasks, χρησιμοποιήθηκαν αρκετές βιβλιοθήκες της Python.

5.1 Clustering

Η ανάλυση Clustering είναι ιδανική για την εύρεση μοτίβων, τμηματοποίησης των πελατών, και στη συγκεκριμένη περίπτωση, εύρεσης ομοιοτήτων. Στο συγκεκριμένο task, θα πραγματοποιήσουμε clustering, χρησιμοποιώντας το Region (9 περιοχές) και το Primary Type (περιγραφή εγκλήματος, 34 είδη). Προκειμένου όμως να γίνει αυτή η ανάλυση, είναι απαραίτητο να βρούμε τον αριθμό K, ο οποίος θα καθορίσει τον αριθμό των συστάδων βέλτιστο.

Για τον λόγο αυτόν, θα χρησιμοποιήσουμε τον Elbow method, ο οποίος υπολογίζει τα τετράγωνα των αποστάσεων κάθε παρατήρησης με το κέντρο της συστάδας (WCSS). Όσο μικρότερη είναι η τιμή αυτού του αριθμού, τόσο πιο συμπαγές είναι το cluster. Ο σκοπός είναι να επιλέξουμε την τιμή για την οποία η επόμενη της, δεν θα προκαλέσει σημαντική διαφορά ως προς το πόσο συμπαγές είναι το cluster. Το σημείο όπου η τιμή WCSS κάνει το γράφημα να μοιάζει με ημι-λυγισμένο βραχίονα, δείχνει τον βέλτιστο αριθμό cluster.



Στη συγκεκριμένη περίπτωση, επιλέγουμε για αριθμό cluster την τιμή 3.

Εφαρμόζοντας τον αλγόριθμο K-Means με αριθμό cluster = 3 μπορούμε να δούμε ποια regions υπάρχουν σε κάθε cluster

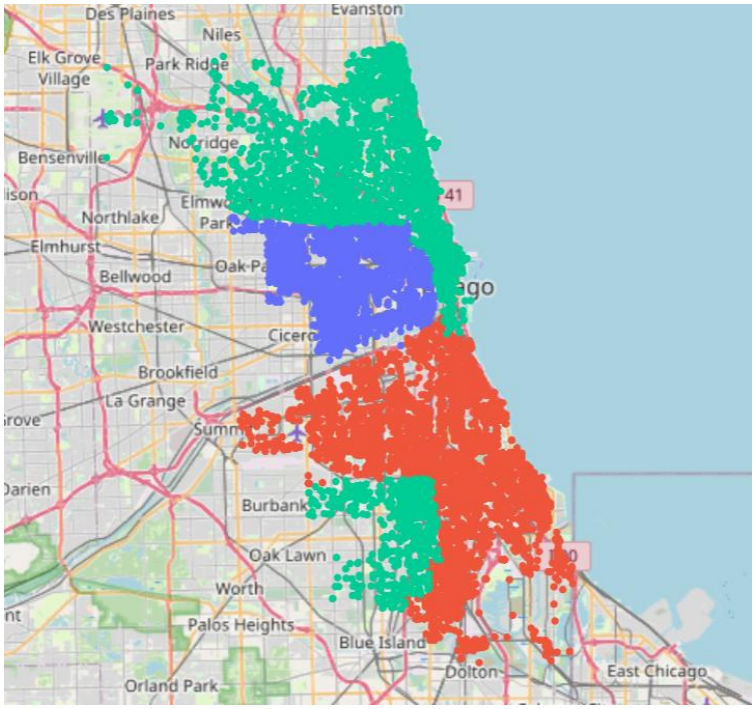
```
table2 = df.pivot_table(index='Area', columns='Primary Type', values='ID', aggfunc='count')
kmeans = KMeans(n_clusters = 3, init = 'k-means++',random_state = 42)
y_kmeans = kmeans.fit_predict(df_standardized)
#beginning of the cluster numbering with 1 instead of 0
y_kmeans1 = y_kmeans + 1

#list called cluster
cluster = list(y_kmeans1)
#Adding cluster to our data set
table2['cluster'] = cluster

#Mean of clusters 1 to 3
kmeans_mean_cluster = pd.DataFrame(round(table2.groupby('cluster').mean(),1))
kmeans_mean_cluster.plot(kind='bar',stacked = True)
plt.rcParams["figure.figsize"] = (2,1)
plt.title("Cluster analysis")
plt.xlabel("Clusters")
```

Primary Type Area	cluster
Far Southeast Side	1
South Side	1
Southwest Side	1
West Side	2
Central	3
Far North Side	3
Far Southwest Side	3
North Side	3
Northwest Side	3

Μπορούμε να παρατηρήσουμε ότι υπάρχει μία γεωγραφική σχέση στον τρόπο με τον οποίο ομαδοποιήθηκαν τα cluster. Παρακάτω ακολουθούν κάποια σχόλια που εξηγούν τον λόγο που έχει πραγματοποιηθεί αυτή η ομαδοποίηση.



Cluster 1

Regions: Far Southeast Side, South Side, Southwest Side

Στο συγκεκριμένο cluster συμμετέχουν οι περιοχές που βρίσκονται κυρίως στο νότιο μέρος του Σικάγο. Οι συγκεκριμένες περιοχές βρίσκονται στις υψηλότερες θέσεις ως προς τον αριθμό των εγκλημάτων, ενώ επίσης το κύριο έγκλημα που καταγράφεται σε κάθε μία από αυτές τις περιοχές είναι η βιαιοπραγία.

Cluster 2

Regions: West Side

Στο συγκεκριμένο cluster συμμετέχει μόνο το Δυτικό Σικάγο. Παρά το γεγονός αυτό, το συγκεκριμένο cluster έχει τον μεγαλύτερο αριθμό εγκλημάτων. Η διαφορά σε σχέση με τις υπόλοιπες περιοχές είναι τόσο μεγάλη στο βαθμό που η περιοχή κατατάσσεται σε ένα cluster μόνη της. Τα εγκλήματα που κυριαρχούν στο Δυτικό Σικάγο είναι η κλοπή και η βιαοπραγία.

Cluster 3
Regions: Central, Far North Side, Far Southwest Side, North Side, Northwest Side
Στο συγκεκριμένο cluster συμμετέχουν οι περιοχές που αποτελούν το βόρειο κομμάτι του Σικάγο, καθώς και το κέντρο του. Οι περιοχές αυτές ομαδοποιούνται σε ένα cluster για τους λόγους ότι ο αριθμός των καταγεγραμμένων εγκλημάτων είναι συγκριτικά αρκετά χαμηλότερος με την υπόλοιπη πόλη, ενώ επίσης το κυρίαρχο έγκλημα σε κάθε μία από αυτές τις περιοχές είναι η κλοπή.

Εάν κατατάξουμε τα εγκλήματα με τον μεγαλύτερο έως τον μικρότερο αριθμό εγκλημάτων, προκύπτει: Cluster 2 > Cluster 1 > Cluster 3

5.2 Random Forrest Classifier

Σε αυτό το data mining task, με το dataset που έχουμε, ο σκοπός είναι να χρησιμοποιήσουμε μηχανική μάθηση για να προβλέψουμε ένα είδος εγκλήματος στο Σικάγο με βάση το μέρος και τον χρόνο το οποίο έγινε. Για τον σκοπό αυτό θα χρησιμοποιηθεί ο αλγόριθμος Random Forest.

Ο αλγόριθμος αυτός αποτελεί ένα είδος μάθησης που λειτουργεί με την κατασκευή ενός πλήθους δέντρων απόφασης κατά το χρόνο εκπαίδευσης. Είναι ένας ισχυρός αλγόριθμος που μπορεί να χειριστεί προβλήματα παλινδρόμησης και ταξινόμησης, ενώ επίσης μπορεί να χειριστεί αρκετά καλά δεδομένα τα οποία δεν είναι ισορροπημένα (για τα δεδομένα των εγκλημάτων ισχύει).

Αρχικά, πρώτα κάνουμε import τις βιβλιοθήκες που θα μας βοηθήσουν στον παραπάνω σκοπό

```
import numpy as np # linear algebra
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix
import plotly.express as px
```

Λαμβάνοντας υπόψη την ανάλυση που έχει γίνει ήδη παραπάνω, πρέπει να επεξεργαστούμε τα δεδομένα μας προτού χρησιμοποιήσουμε τον αλγόριθμο Random forest. Ο συγκεκριμένος αλγόριθμος χρησιμοποιεί πολλαπλά δέντρα αποφάσεων, όπου μεμονωμένα κατασκευάζονται από τυχαίο υποσύνολο δεδομένων. Κάθε δέντρο στο δάσος κάνει μία πρόβλεψη, και αυτή που προβλέπεται από τα περισσότερα δέντρα είναι η τελική πρόβλεψη του αλγορίθμου. Ο Random forest μπορεί να διαχειριστεί δεδομένα που είναι μόνο ακέραιες ή λογικές τιμές, και όχι strings. Για αυτόν τον λόγο, είναι απαραίτητο να κάνουμε τις απαραίτητες αλλαγές στα δεδομένα μας.

Το καινούργιο dataframe που θα δημιουργηθεί μετά τις μετατροπές θα περιέχει :

Ημερομηνία και ώρα: Η στιγμή στην οποία πραγματοποιήθηκε το έγκλημα είναι αναγκαία. Για παράδειγμα, εάν είναι Κυριακή βράδυ όπου παρατηρείται υψηλή κίνηση στους δρόμους της πόλης, είναι πιο πιθανή η ύπαρξη συγκεκριμένων εγκλημάτων (για παράδειγμα διακίνιση ναρκωτικών ουσιών). Θα χρησιμοποιήσουμε διαφορετικές στήλες για τον μήνα, την ημέρα και την ώρα του εγκλήματος.

Τοποθεσία: Για τον σκοπό αυτό, θα χρησιμοποιήσουμε τη στήλη Area η οποία είναι τύπου string, επομένως θα προστεθεί με την χρήση μεταβλητών Dummy, αλλά επίσης και η στήλη Block, όπου η κατάληξη του προσδιορίζει αν βρισκόμαστε σε δρόμο, λεωφόρο κλπ.

Γεωγραφικό μήκος και πλάτος: Οι μεταβλητές αυτές θα προστεθούν, αφού αφαιρέσουμε οποιαδήποτε outliers. Με αυτόν τον τρόπο το μοντέλο μας θα έχει μία καλύτερη επίγνωση της τοποθεσίας.

Με βάση τα παραπάνω κατασκευάζεται ο αλγόριθμος Preprocess Features

```
def preprocessFeatures(dfx):

    df = pd.get_dummies(dfx[['Area']])
    df['Hour_Min'] = pd.to_datetime(dfx['Date']).dt.hour + pd.to_datetime(dfx['Date']).dt.minute / 60
    # Add a feature that contains the expontential time
    df['Hour_Min_Exp'] = np.exp(df['Hour_Min'])

    df['Day'] = pd.to_datetime(dfx['Date']).dt.day
    df['Month'] = pd.to_datetime(dfx['Date']).dt.month
    df['Year'] = pd.to_datetime(dfx['Date']).dt.year

    month_one_hot_encoded = pd.get_dummies(pd.to_datetime(dfx['Date']).dt.month, prefix='Month')
    df = pd.concat([df, month_one_hot_encoded], axis=1, join="inner")

    # Convert Carthesian Coordinates to Polar Coordinates
    df[['Latitude', 'Longitude']] = dfx[['Latitude', 'Longitude']]
    df['dist'], df['phi'] = cart2polar(dfx['Latitude'], dfx['Longitude'])

    # Extracting Street Types
    df['Is_ST'] = dfx['Block'].str.contains(" ST", case=True)
    df['Is_AV'] = dfx['Block'].str.contains(" AV", case=True)
    df['Is_TR'] = dfx['Block'].str.contains(" TR", case=True)
    df['Is_DR'] = dfx['Block'].str.contains(" DR", case=True)
    df['Is_PL'] = dfx['Block'].str.contains(" PL", case=True)

    return df
```

Επιπλέον, ο αλγόριθμος θα προβλέψει το είδος του εγκλήματος, τα οποία είναι 33. Προκειμένου να έχουμε μία αυξημένη ακρίβεια στον αλγόριθμό μας, κρατάμε τα 15 πιο συχνά εγκλήματα (τα οποία αποτελούν το 98.4% του συνόλου των δεδομένων), καθώς επίσης ομαδοποιούμε κάποια ήδη εγκλημάτων που έχουν μεγάλες ομοιότητες (για παράδειγμα η κλοπή, η ληστεία και η κλοπή μοτοσυκλέτας τοποθετούνται σε μία κατηγορία). Ως αποτέλεσμα, τα είδη των εγκλημάτων μετατρέπονται σε 6.

Χωρισμός δεδομένων

Σε αυτό το σημείο, πριν χρησιμοποιήσουμε το μοντέλο προβλέψεων, χωρίζουμε τα δεδομένα μας σε ξεχωριστά σύνολα, προκειμένου να επιτευχθεί εκπαίδευση και δοκιμή. Για αυτόν τον λόγο χρησιμοποιείται η συνάρτηση train_test_split με αναλογία διαχωρισμού 70%. Στη συνέχεια, ακολουθεί η εξαγωγή των δεδομένων για την εκπαίδευση του μοντέλου.

Εκπαίδευση μοντέλου

Σε αυτό το σημείο, αφού έχουμε διαχειριστεί τα δεδομένα μας, εκτελούμε τον αλγόριθμο Random Forest , βάζοντας σαν παράμετρο μέγιστου βάθους δέντρων 20 και αριθμό δέντρων 100.

```
clf = RandomForestClassifier(max_depth=20, random_state=0, n_estimators = 100)
clf.fit(x_train, y_train.ravel())
y_pred = clf.predict(x_test)

results_log = classification_report(y_test, y_pred)
print(results_log)
```

```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00      2669
           1       0.39      0.96      0.56     16990
           2       0.28      0.03      0.06     12489
           3       0.19      0.00      0.01      8440
           4       0.50      0.00      0.00      1644
           5       0.00      0.00      0.00       772

    accuracy                           0.39     43004
   macro avg       0.23      0.17      0.10     43004
weighted avg       0.29      0.39      0.24     43004
```

Για να δούμε πως αποδίδει το μοντέλο, δημιουργούμε τη μεταβλητή προβλέψεων y_pred στο σύνολο δεδομένων δοκιμής x_test. Στη συνέχεια, χρησιμοποιούμε τις έγκυρες τιμές y_test καθώς και την y_pred η οποία είναι το αποτέλεσμα που έχει προβλέψη το μοντέλο μας.

```
True label: 4 Predicted label: 1
True label: 0 Predicted label: 0
True label: 0 Predicted label: 0
True label: 0 Predicted label: 0
True label: 1 Predicted label: 0
True label: 1 Predicted label: 0
True label: 1 Predicted label: 0
True label: 0 Predicted label: 0
True label: 0 Predicted label: 0
True label: 0 Predicted label: 0
True label: 0 Predicted label: 0
True label: 0 Predicted label: 0
True label: 0 Predicted label: 0
True label: 0 Predicted label: 0
True label: 1 Predicted label: 0
True label: 0 Predicted label: 0
True label: 1 Predicted label: 1
True label: 0 Predicted label: 0
True label: 3 Predicted label: 0
```

Παρατηρούμε ακρίβεια 39%. Εκ πρώτης όψεως το αποτέλεσμα δεν φαίνεται ιδιαίτερα εντυπωσιακό, όμως αν συνυπολογίσουμε ότι χρησιμοποιούμε αποτελέσματα μόνο ενός έτους (2022, δηλαδή περίπου 210.000 εγγραφές), καθώς και το γεγονός ότι υπάρχουν 6 δυνατές κατηγορίες εγκλήματος, η απόδοση κρίνεται αποδεκτή

Το μοντέλο πολύ συχνά προβλέπει τον αριθμό 0, (δηλαδή το έγκλημα που έχει αρχειοθετηθεί με την τιμή 0), και στις περισσότερες φορές αμελεί τις υπόλοιπες κατηγορίες. Ο λόγος που συμβαίνει αυτό είναι η άνιση κατανομή των εγκλημάτων στα δεδομένα μας. Ως αποτέλεσμα, το μοντέλο μπορεί να προβλέπει με μεγάλη ακρίβεια συγκεκριμένα εγκλήματα, όμως με αρκετά χαμηλότερη ακρίβεια κάποια άλλα εγκλήματα.

6. Πηγές

https://www.kaggle.com/datasets/onlyrohit/crimes-in-chicago?resource=download&fbclid=IwAR2CbYaDRwKgWVlj5yJsYn1m2VMwNvXyhPBZDzCBRVG-WABM_ihdljQ-qzs

https://towardsdatascience.com/exploring-clustering-and-mapping-torontos-crimes-96336efe490f

https://towardsdatascience.com/crime-location-analysis-and-prediction-using-python-and-machine-learning-1d8db9c8b6e6

https://medium.com/analytics-vidhya/crime-data-pattern-analysis-and-visualization-using-k-means-clustering-ceeb963a2b47

https://medium.com/analytics-vidhya/everything-you-need-to-know-about-k-means-clustering-88ad4058cce0

https://medium.com/web-mining-is688-spring-2021/americas-crime-rates-2019-a-k-means-clustering-analysis-8c361f9db831

https://www.edureka.co/blog/implementing-kmeans-clustering-on-the-crime-dataset/

https://sigmamagic.com/blogs/crime-analysis-using-k-means-clustering/

https://github.com/tanvipenumudy/Winter-Internship-Internity/blob/main/Day%2011/Day-11%20Notebook-3%20%28Crime%20Data%20Analysis%29.ipynb

https://www.relataly.com/predicting-crimes-in-san-francisco-creatingsf-crime-map-using-xgboost/2960/

https://www.softwaretestinghelp.com/dimensional-data-model-in-data-warehouse/

https://www.softwaretestinghelp.com/data-mining-examples/#6_Crime_Prevention

https://en.wikipedia.org/wiki/Community_areas_in_Chicago

https://chicagostudies.uchicago.edu/grid

https://chicagostudies.uchicago.edu/neighborhoods

https://wmich.edu/writing/rules/addresses

https://www.allaroundmoving.com/5-richest-neighborhoods-in-chicago-to-move-in/

https://estatousa.com/most-dangerous-neighborhoods-in-chicago/

https://propertyclub.nyc/article/most-dangerous-neighborhoods-in-chicago