POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Systems and Methods for Big and Unstructured Data Project

Author(s): **Chiara Fossà 10724941**

Group Number: **63**

Academic Year: 2023-2024

Contents

1 | Introduction

In this introduction is described the problem I'd like to face using the dataset "Mexican Crime Statistics:Comprehensive (2015-2023)" and the Neo4j technology and why I picked the Neo4j DB technology based on the dataset features.

1.1. The dataset

The problem I intend to face using the data provided by the **"Mexican Crime Statistics: Comprehensive (2015-2023)"** dataset consists in the acquisition of in-depth information regarding criminal activities reported throughout Mexico over a multi-year period of time. The dataset comes directly from the official Mexican government website and offers a detailed collection of criminal incidents during a period of time that goes from 2015 to October 2023. Making use of the unique characteristics of the chosen dataset my goal is to conduct a comprehensive analysis in the fields of criminology, public policy and data science.

More in detail, analyzing the Mexican Crime Statistics dataset is interesting for several reasons which appear both from its characteristics and from the nature of the information contained:

- **The broad geographical and temporal coverage:** The dataset covers a period of nine years (from 2015 to 2023 (until October)) allowing a long-term perception of criminal dynamics. Furthermore, the data comes from all the different regions of Mexico and this allows us to carry out in-depth research not only in temporal but also spatial terms and allows us to detect geographical variations in criminal patterns.
- **The detailed crime categorizations:** The presence of fields, subfields, methods and typologies allows crimes to be categorized in a detailed way, allowing both the general and more specific aspects of criminal activities to be explored. This way it is possible to focus on specific trends by focusing on subgroups of crimes of particular interest. Information on the method and categorization of the crime is a

detail that enriches the analysis, allowing us to better understand the context and circumstances linked to each crime and offers a clear vision of the legal interests involved in each crime, allowing us to analyze the impacts both at personal and social level.

- **The official resources and data reliability:** The data is aggregated from official reports and Mexican government records and this ensures a high degree of both reliability and authenticity. Authenticity is key to ensuring that the data analysis that can be carried out on this data is based on accurate and verifiable information.

The information obtained can be useful in disciplines such as criminology, public policy analysis, data science, sociology, emergency management and public safety, criminal epidemiology, economics, forensic psychology, urban and regional planning and investigative journalism.

1.2. The tecnology

I chose Neo4j for its ability to effectively represent and manage data with complex relationships. With its graph structure, each element can be connected to many others through characteristic relationships, allowing the space in the database to be managed more efficiently, avoiding repetitions. In the case of this dataset this is particularly relevant, since there are articolated links between crime types, subtypes, modalities, geographical entities and time periods. Analyzing with Neo4j allows to dynamically approach the understanding of criminal models. I can explore relationships between different elements such as linking a specific crime type to a particular geographic entity or investigating crime subtypes associated with a specific modality.

I hope it will be interesting to analyze this dataset with Neo4j for the possibility of revealing complex and non-obvious relationships between the data. Furthermore, Neo4j is flexible in managing the complexity of connections between data and this makes it an ideal tool to fully exploit the information potential of the dataset I have chosen.

2 | Dataset

The Mexican Crime Statistics dataset represents a transformed, organized and translated set of data that comes directly from official crime statistics in Mexico. This dataset has been translated and structured in a regulated way to facilitate the analysis and interpretation of the information.

2.1. Data Card for Mexican Crime Statistics Dataset

The full title of this dataset is **Mexican Crime Statistics: Comprehensive Incident Dataset between 2015 and 2023** and which is precisely defined as **An Extensive Compilation of Criminal Incidents in Mexico, Sourced from Official Government Data**.
In fact its source, as already mentioned in the previous chapter, is the **Official Mexican Government Website**.

This dataset constitutes a collection of criminal incidents that have been reported throughout Mexico. In particular, it contains detailed records of various criminal activities offering interesting insights into criminal patterns and trends in the different states of Mexico.

Composition of the Dataset:

- Data Type: Structured, CSV format
- Record Number: Shape (332416, 9)
- Date Range: 2015-2023 (until October, when I downloaded it November and December were not yet available)

Possible uses suggested by the creators of the dataset:

- Intended Use: Research in criminology, analysis of public policies, analysis of criminal trends.

- Analysis Example: Trends in crime rates over time, regional crime analysis, frequency analysis of crime types.

Dataset Description

year: This is the year the crime was reported. A numeric field representing the calendar year (for example, 2015).

entity_code: A numeric code representing a specific entity (state or region) within Mexico. Each number corresponds to a unique entity. Here is the list of associations between code and entities:

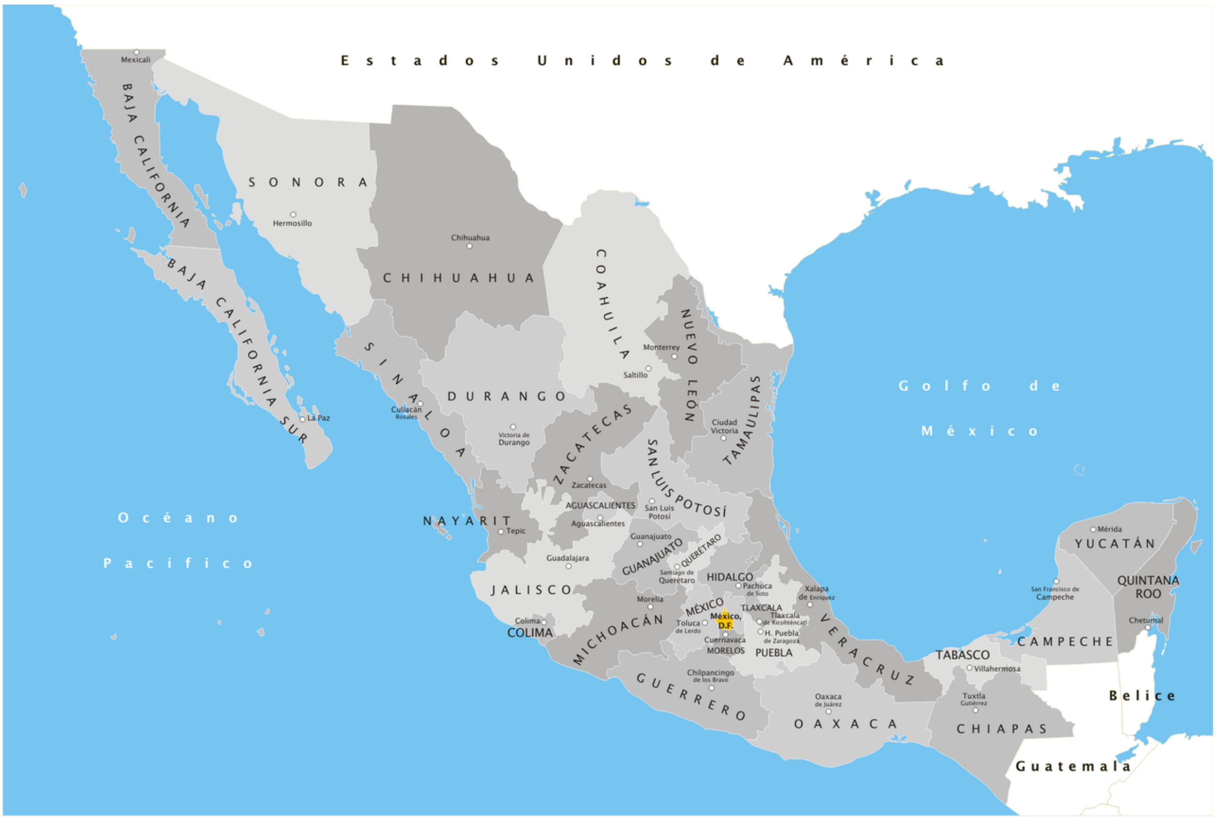


Figure 2.1: Federal states of Mexico

1. Aguascalientes
2. Baja California
3. Baja California Sur
4. Campeche
5. Coahuila de Zaragoza

6. Colima
7. Chiapas
8. Chihuahua
9. Ciudad de México
10. Durango
11. Guanajuato
12. Guerrero
13. Hidalgo
14. Jalisco
15. México
16. Michoacán de Ocampo
17. Morelos
18. Nayarit
19. Nuevo León
20. Oaxaca
21. Puebla
22. Querétaro
23. Quintana Roo
24. San Luis Potosí
25. Sinaloa
26. Sonora
27. Tabasco
28. Tamaulipas
29. Tlaxcala
30. Veracruz de Ignacio de la Llave
31. Yucatán

32. Zacatecas

entity: The name of the state or region of Mexico where the crime occurred. A text field (for example, Aguascalientes).

affected_legal_good: A categorical field describing the general category of legal good affected by the crime (such as personal or social interest). Examples include ’Personal Freedom’ and ’Sexual Freedom and Safety’.

type_of_crime: A categorical field indicating the general type of crime. More specific than ’affected_legal_good’ but less detailed than ’subtype_of_crime’. For example ’Kidnapping’, ’Sexual Abuse’ and ’Robbery’.

subtype_of_crime: A further categorization of the type of crime. It provides more specific details within the general type of crime. Examples include ’Sexual Harassment’, ’Simple Sexual Assault’ and ’Home Burglary’.

modality: describes the specific nature or method of the crime. This field provides details about how the crime was committed and any specific characteristics that differentiate it within its subtype. Examples: ’With violence’, ’Without violence’ and ’Sexual bullying’.

month: is the month in which the crime was reported. A text field representing the month (for example, January).

count: is the number of incidents reported for the specific crime type, subtype and mode in the specified entity and month.

2.2. The Neo4j database implementation

The database I created from the dataset is made up of **8 node types** and **5 relationship types**.

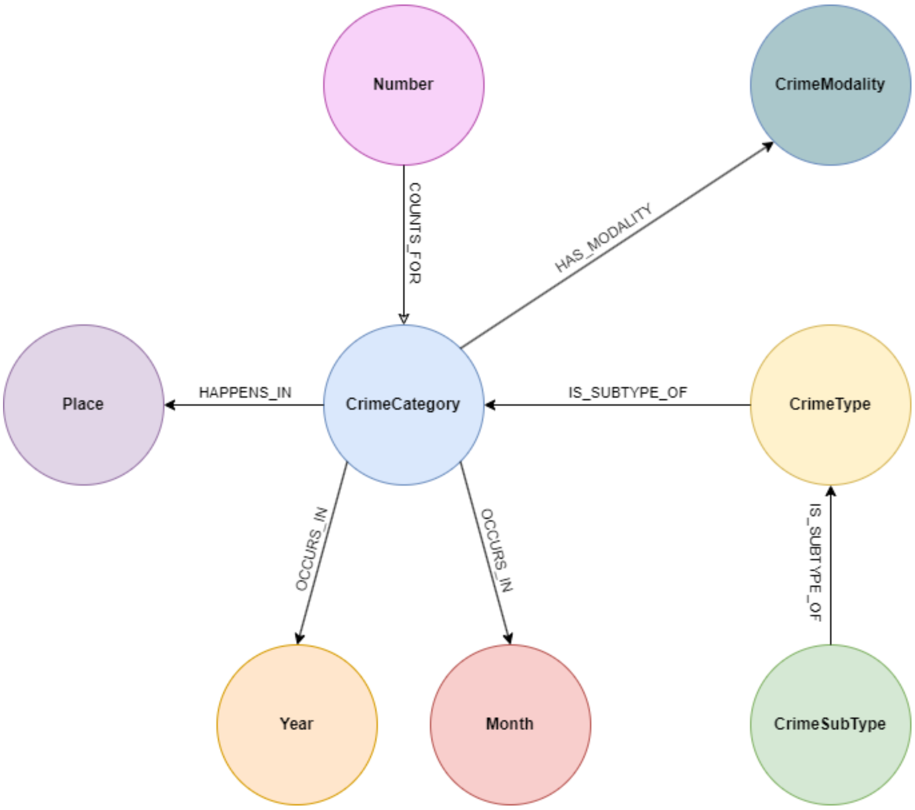


Figure 2.2: Neo4j database graph

The nodes are:

- **Place**: containing the attributes *entity_ code*, a number and *entity*, a textual name. These values represent respectevely the code associate to a state or a region and the name of that state or region
- **Year**: containing the attribute *year*, a number. This value represents an year between 2015 and 2023
- **Month**: containing the attribute *month*, a number. This value represents a month between January and December
- **CrimeCategory**: containing the attribute *crimeCategory*, a textual category. This attribute represents the general category of goods affected by a crime
- **CrimeType**: containing the attribute *crimeType*, a textual category. This attribute represents the general type of crime committed.
- **CrimeSubtype**: containing the attribute *crimeSubtype*, a textual category. This attribute represent the specific subtype of crime committed.

- **CrimeModality**: containing the attribute *crimeModality*, a textual category. This attribute represent the specific modality of crime committed.
- **Number**: containing the attribute *number*, a number. This attribute represents the number of occurrences of that specific crime in a specific place, with specific modalities.

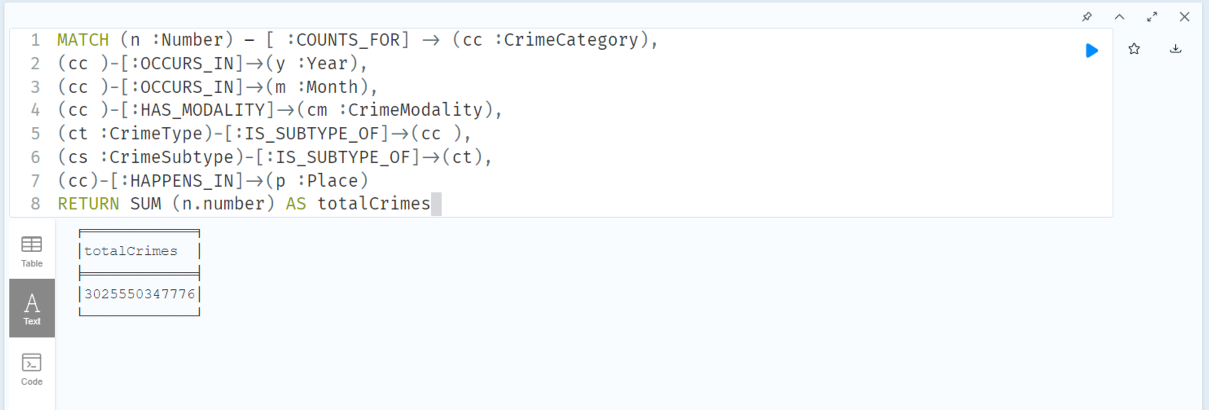
The relationships are:

- **HAPPENS_IN**: from CrimeCategory to Place
- **OCCURS_IN**: from CrimeCategory to Year and from CrimeCategory to Month
- **IS_SUBTYPE_OF**: from CrimeType to CrimeCategory and from CrimeSubtype to CrimeType
- **HAS_MODALITY**: from CrimeCategory to CrimeModality
- **COUNTS_FOR**: from Number to CrimeCategory

3 | Query

3.1. Query 1: How many crimes were committed between 2015 and 2023?

This first query tries to give an idea of the number of crimes we are dealing with in this dataset. It counts all the crimes present in the database of each month of each year examined adding the values present in the nodes Number.



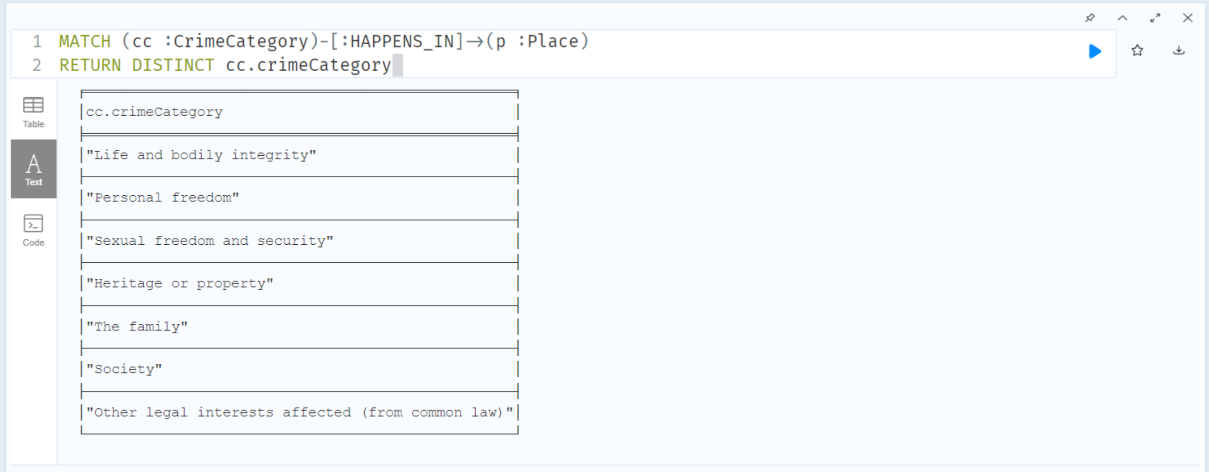
```
MATCH (n :Number) - [ :COUNTS_FOR] -> (cc :CrimeCategory),
(cc )-[:OCCURS_IN]->(y :Year),
(cc )-[:OCCURS_IN]->(m :Month),
(cc )-[:HAS_MODALITY]->(cm :CrimeModality),
(ct :CrimeType)-[:IS_SUBTYPE_OF]->(cc ),
(cs :CrimeSubtype)-[:IS_SUBTYPE_OF]->(ct),
(cc)-[:HAPPENS_IN]->(p :Place)
RETURN SUM (n.number) AS totalCrimes
```

```
╒═════════════╕
│totalCrimes  │
╞═════════════╡
│3025550347776│
└─────────────┘
```

Figure 3.1: Query 1

3.2. Query 2: Which categories do crimes committed in Mexico fall into?

This is an example of a query that can be run on each attribute of each node to get an idea of the types of information that can be found within each node. In fact, for each attribute, there is not a huge variety of data that may have been entered; the complexity of this database lies in how the nodes relate to each other. With queries of this type you can easily understand how crimes are classified in this database. I chose to operate this query on the CrimeCategory node data but the same thing can be done for nodes such as CrimeType, CrimeSubtype or CrimeModality.



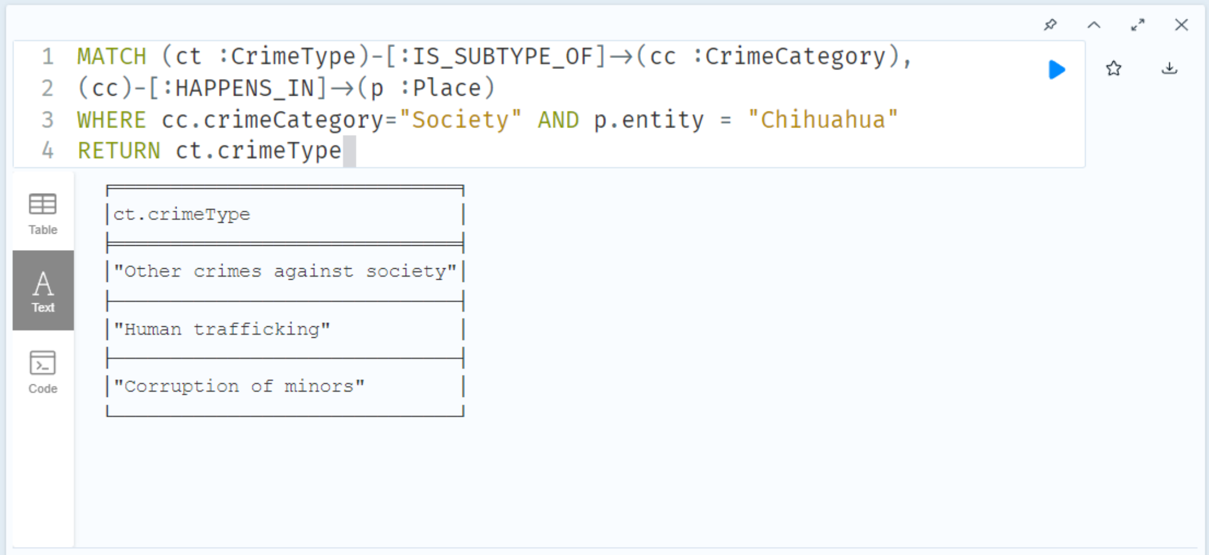
```
MATCH (cc :CrimeCategory)-[:HAPPENS_IN]→(p :Place)
RETURN DISTINCT cc.crimeCategory
```

```
╒═══════════════════════════════════════════════╕
│cc.crimeCategory                               │
╞═══════════════════════════════════════════════╡
│"Life and bodily integrity"                    │
├───────────────────────────────────────────────┤
│"Personal freedom"                             │
├───────────────────────────────────────────────┤
│"Sexual freedom and security"                  │
├───────────────────────────────────────────────┤
│"Heritage or property"                         │
├───────────────────────────────────────────────┤
│"The family"                                   │
├───────────────────────────────────────────────┤
│"Society"                                      │
├───────────────────────────────────────────────┤
│"Other legal interests affected (from common law)"│
└───────────────────────────────────────────────┘
```

Figure 3.2: Query 2

3.3. Query 3: Which crimes against Society happened in Chihuahua during the years?

This query returns the types of crimes against heritage or property committed in Chihuahua during the time period covered by the database



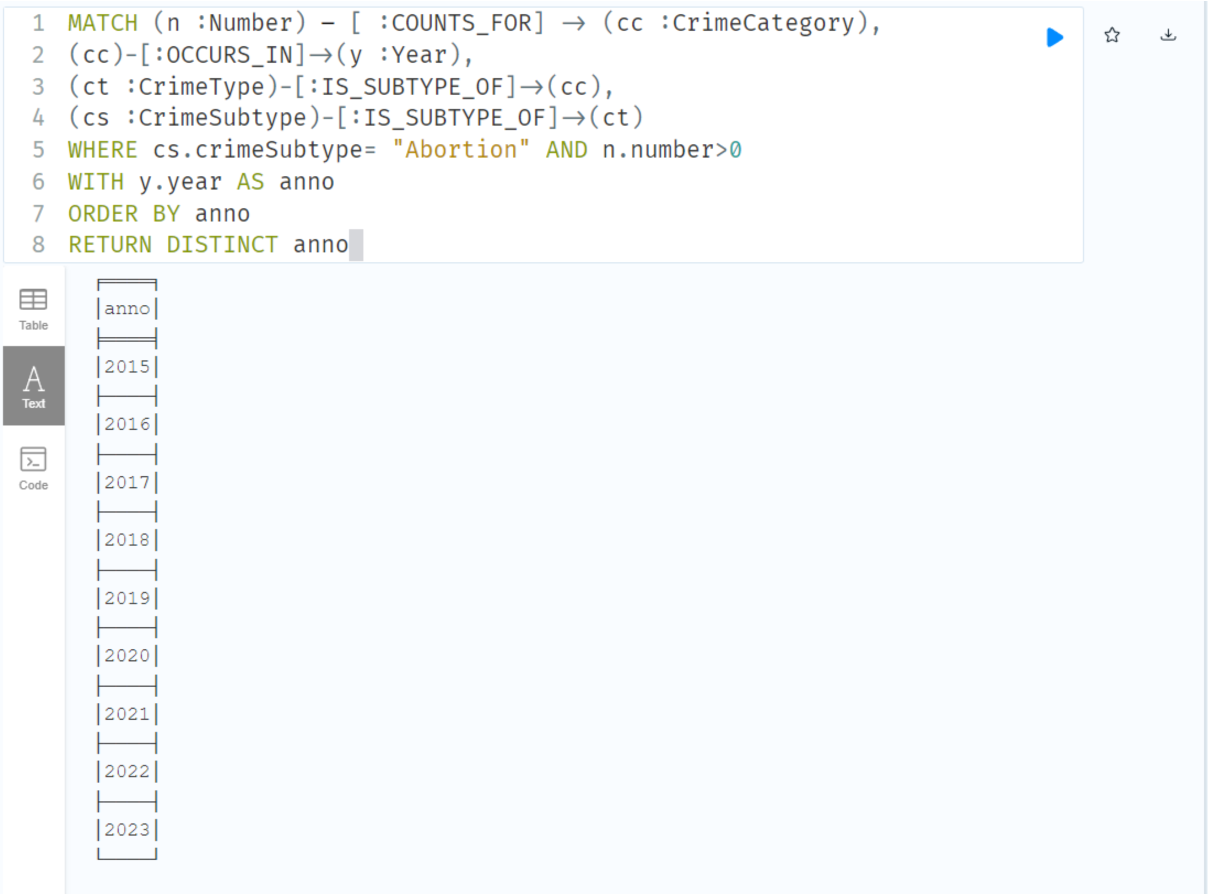
```
MATCH (ct :CrimeType)-[:IS_SUBTYPE_OF]→(cc :CrimeCategory),
(cc)-[:HAPPENS_IN]→(p :Place)
WHERE cc.crimeCategory="Society" AND p.entity = "Chihuahua"
RETURN ct.crimeType
```

```
╒═══════════════════════════════╕
│ct.crimeType                   │
╞═══════════════════════════════╡
│"Other crimes against society" │
├───────────────────────────────┤
│"Human trafficking"            │
├───────────────────────────────┤
│"Corruption of minors"         │
└───────────────────────────────┘
```

Figure 3.3: Query 3

3.4. Query 4: Abortion rights

On September 7, 2023, Mexico decriminalized abortion in all 32 states of the country. Previously there were states where it was considered a crime. This query searches in which years and in which states people were indicted for this reason.



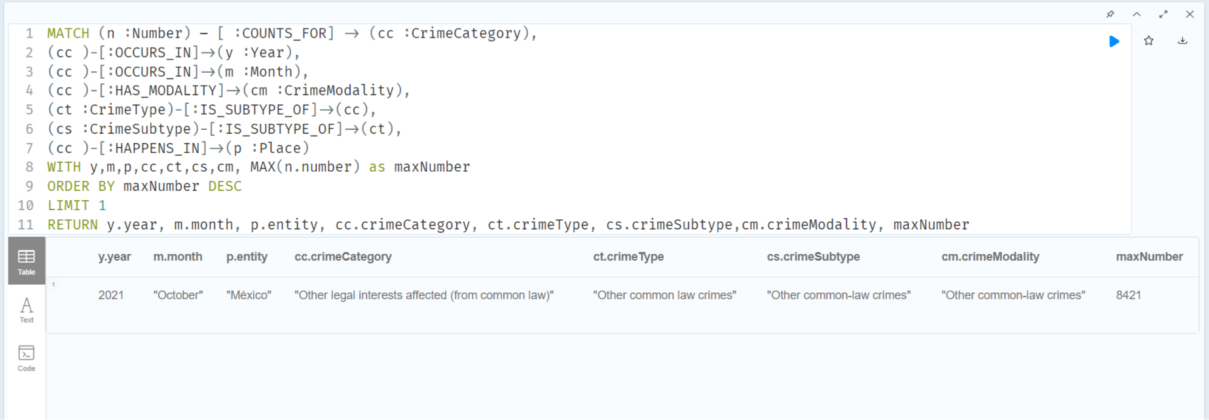
```
MATCH (n :Number) – [ :COUNTS_FOR] → (cc :CrimeCategory),
(cc)-[:OCCURS_IN]→(y :Year),
(ct :CrimeType)-[:IS_SUBTYPE_OF]→(cc),
(cs :CrimeSubtype)-[:IS_SUBTYPE_OF]→(ct)
WHERE cs.crimeSubtype= "Abortion" AND n.number>0
WITH y.year AS anno
ORDER BY anno
RETURN DISTINCT anno
```

```
╒════╕
│anno│
╞════╡
│2015│
├────┤
│2016│
├────┤
│2017│
├────┤
│2018│
├────┤
│2019│
├────┤
│2020│
├────┤
│2021│
├────┤
│2022│
├────┤
│2023│
└────┘
```

Figure 3.4: Query 4

3.5. Query 5: What is the crime that have the higher count in all the database and in which month and year?

This query tries to get all the details of the most committed specific crime in the entire dataset.



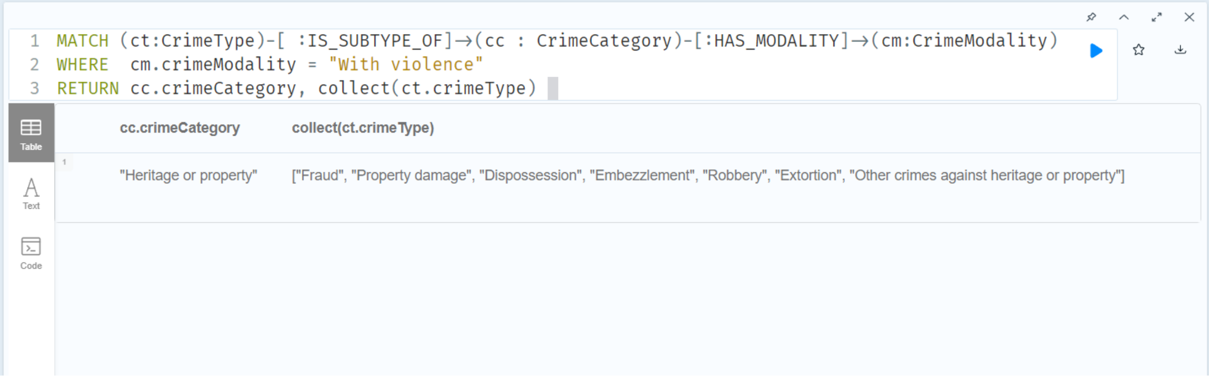
```
MATCH (n :Number) - [ :COUNTS_FOR] -> (cc :CrimeCategory),
(cc )-[:OCCURS_IN]->(y :Year),
(cc )-[:OCCURS_IN]->(m :Month),
(cc )-[:HAS_MODALITY]->(cm :CrimeModality),
(ct :CrimeType)-[:IS_SUBTYPE_OF]->(cc),
(cs :CrimeSubtype)-[:IS_SUBTYPE_OF]->(ct),
(cc )-[:HAPPENS_IN]->(p :Place)
WITH y,m,p,cc,ct,cs,cm, MAX(n.number) as maxNumber
ORDER BY maxNumber DESC
LIMIT 1
RETURN y.year, m.month, p.entity, cc.crimeCategory, ct.crimeType, cs.crimeSubtype,cm.crimeModality, maxNumber
```

	y.year	m.month	p.entity	cc.crimeCategory	ct.crimeType	cs.crimeSubtype	cm.crimeModality	maxNumber
1	2021	"October"	"México"	"Other legal interests affected (from common law)"	"Other common law crimes"	"Other common-law crimes"	"Other common-law crimes"	8421

Figure 3.5: Query 5

3.6. Query 6: What crimes can be committed with violence?

This query searches for all crimes committed with violence and for each category of crime that can be committed with violence it shows the collection of crime types that can be committed with violence. Apparently it is only the prerogative of "heritage or property".



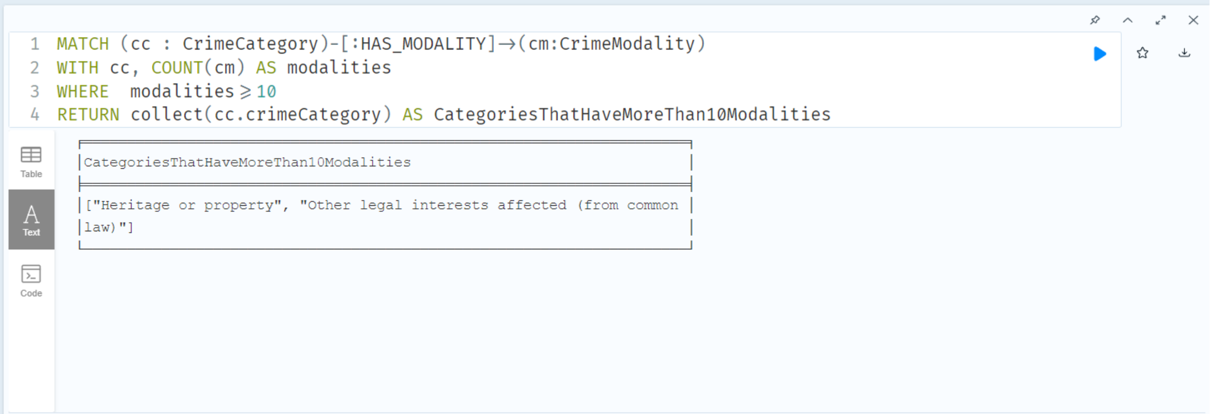
```
MATCH (ct:CrimeType)-[ :IS_SUBTYPE_OF]→(cc : CrimeCategory)-[:HAS_MODALITY]→(cm:CrimeModality)
WHERE  cm.crimeModality = "With violence"
RETURN cc.crimeCategory, collect(ct.crimeType)
```

	cc.crimeCategory	collect(ct.crimeType)
1	"Heritage or property"	["Fraud", "Property damage", "Dispossession", "Embezzlement", "Robbery", "Extortion", "Other crimes against heritage or property"]

Figure 3.6: Query 6

3.7. Query 7: There are so many ways to commit crimes

This query returns all categories of crimes that can be committed in at least 10 different ways.



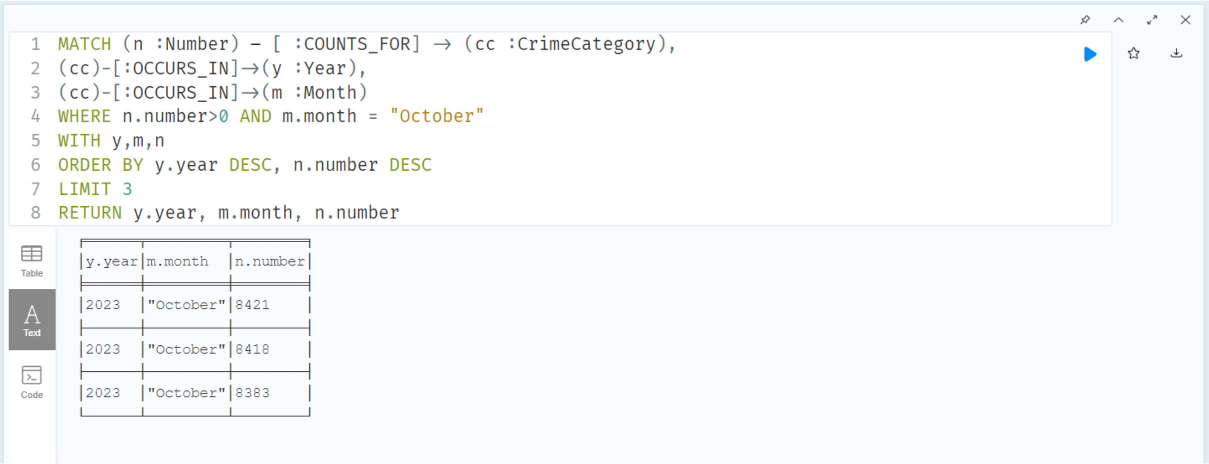
```
MATCH (cc : CrimeCategory)-[:HAS_MODALITY]→(cm:CrimeModality)
WITH cc, COUNT(cm) AS modalities
WHERE  modalities ⩾ 10
RETURN collect(cc.crimeCategory) AS CategoriesThatHaveMoreThan10Modalities
```

```
╒══════════════════════════════════════════════════════════════════════╕
│CategoriesThatHaveMoreThan10Modalities                                │
╞══════════════════════════════════════════════════════════════════════╡
│["Heritage or property", "Other legal interests affected (from common │
│law)"]                                                                │
└──────────────────────────────────────────────────────────────────────┘
```

Figure 3.7: Query 7

3.8. Query 8: The last records

How far does this database go? This query returns the number of 3 of the last crimes committed, recorded in this database.



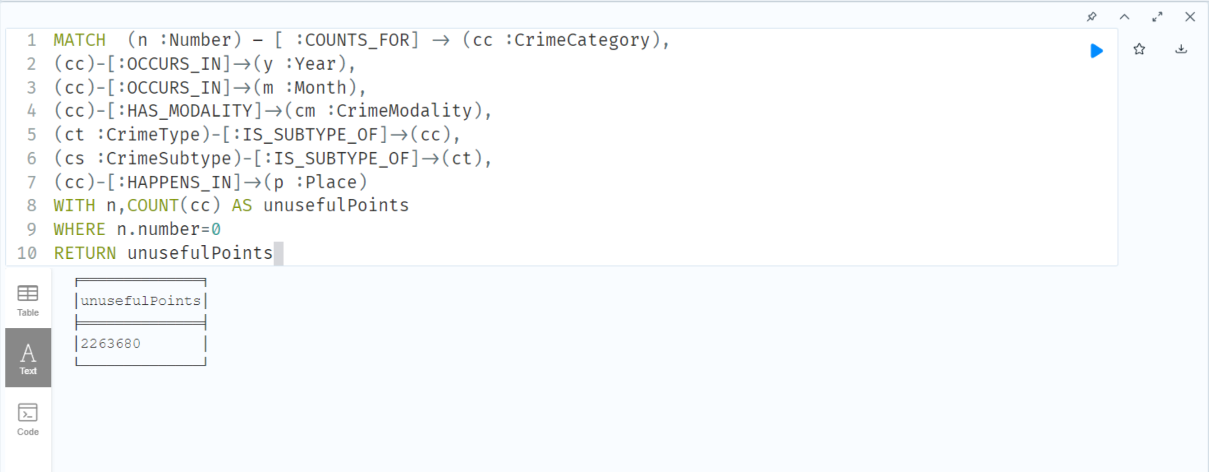
```
MATCH (n :Number) - [ :COUNTS_FOR] → (cc :CrimeCategory),
(cc)-[:OCCURS_IN]→(y :Year),
(cc)-[:OCCURS_IN]→(m :Month)
WHERE n.number>0 AND m.month = "October"
WITH y,m,n
ORDER BY y.year DESC, n.number DESC
LIMIT 3
RETURN y.year, m.month, n.number
```

```
┌──────┬─────────┬────────┐
│y.year│m.month  │n.number│
├──────┼─────────┼────────┤
│2023  │"October"│8421    │
├──────┼─────────┼────────┤
│2023  │"October"│8418    │
├──────┼─────────┼────────┤
│2023  │"October"│8383    │
└──────┴─────────┴────────┘
```

Figure 3.8: Query 8

3.9. Query 9: Speaking of useless data

I noticed that many of the data points in this dataset lead to a number of crimes equal to 0. Does it make sense to keep in memory crimes that did not occur? This query counts how many rows could have been removed to streamline the dataset.



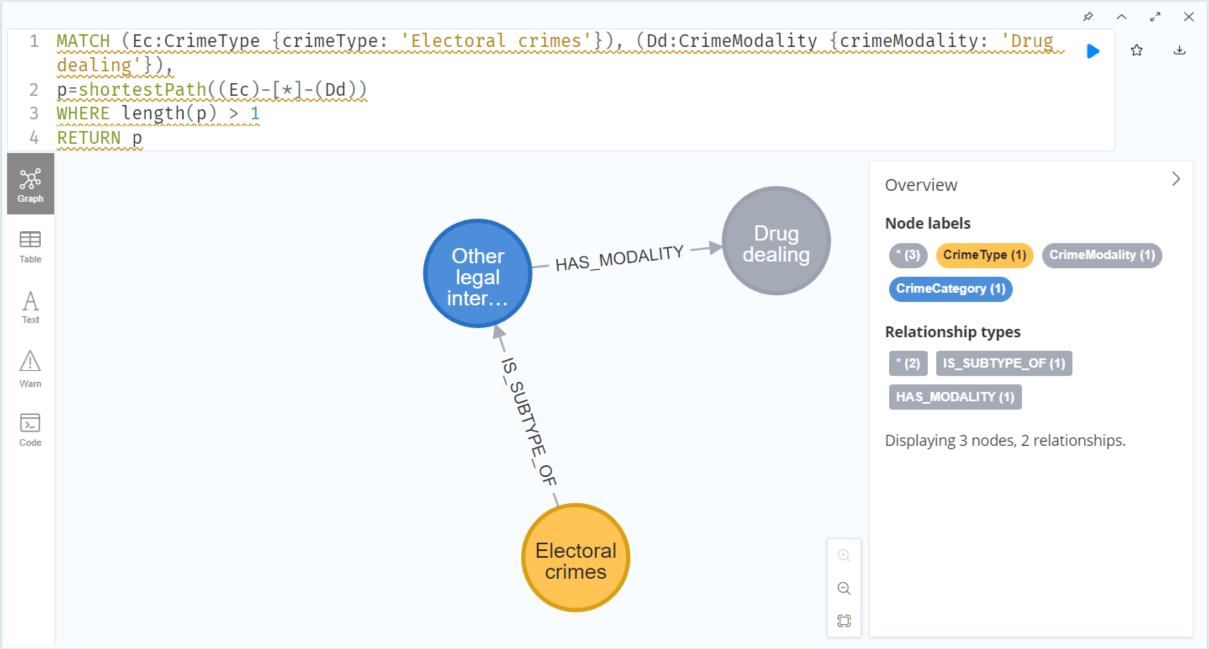
```
MATCH  (n :Number) - [ :COUNTS_FOR] -> (cc :CrimeCategory),
(cc)-[:OCCURS_IN]->(y :Year),
(cc)-[:OCCURS_IN]->(m :Month),
(cc)-[:HAS_MODALITY]->(cm :CrimeModality),
(ct :CrimeType)-[:IS_SUBTYPE_OF]->(cc),
(cs :CrimeSubtype)-[:IS_SUBTYPE_OF]->(ct),
(cc)-[:HAPPENS_IN]->(p :Place)
WITH n,COUNT(cc) AS unusefulPoints
WHERE n.number=0
RETURN unusefulPoints
```

```
╒══════════════╕
│unusefulPoints│
╞══════════════╡
│2263680       │
└──────────────┘
```

Figure 3.9: Query 9

3.10. Query 10: The road from electoral crimes to drug dealing is paved with good intentions

This query aims to show how conceptually close, according to the crime classification method of this dataset, a crime type (in this case I chose Electoral crimes) and a crime modality that does not belong to that crime type are (in this case I chose Drug dealing)



```
MATCH (Ec:CrimeType {crimeType: 'Electoral crimes'}), (Dd:CrimeModality {crimeModality: 'Drug dealing'}),
p=shortestPath((Ec)-[*]-(Dd))
WHERE length(p) > 1
RETURN p
```

Displaying 3 nodes, 2 relationships.

Figure 3.10: Query 10

A | Links

Open dataset link: Mexican Crime Statistics: Comprehensive Incident Dataset
Mexican govenrment link: Datos Abiertos de Incidencia Delictiva

List of Figures