DSC202 Data Management for Data Science

Final Project

Crime Analytics: Neo4J & PSQL Integration

By:

Kavya Sridhar PID: A69035242

Shreyash Reddy PID: A69034537

Susmit Singh PID: A69034300

Contents

1. Introduction

Crime significantly affects public safety and urban life, making data-driven analysis essential for prevention and informed decision-making. This project uses modern database technologies to explore patterns, trends, and relationships within crime data.

Using more than 6,000 incidents from the City of London (2022–2024), the data was cleaned and processed before being integrated into Neo4J and PostgreSQL. Neo4J, a graph database, helps uncover connections between crimes, locations, and categories, while PostgreSQL supports time series analysis and statistical querying.

Through this dual approach, we identify crime hotspots, seasonal trends, and influential locations. The system enables interactive querying and visualization, providing valuable insights for predictive policing and strategic planning.

2. Objective

The objective of this project is to develop a comprehensive crime analytics platform that integrates Neo4J and PostgreSQL to extract meaningful insights from crime data. With the strengths of both graph and relational databases, the project aims to detect hidden patterns, spatial-temporal trends, and relationships in huge crime datasets.

Some of the most important goals include identifying high-risk areas, analyzing crime types and their intersections, understanding seasonal and historical trends, and ranking important locations by connectivity. The use of algorithms like PageRank and Louvain in Neo4J allows for advanced network analysis, while PostgreSQL enables fast time series and statistical queries.

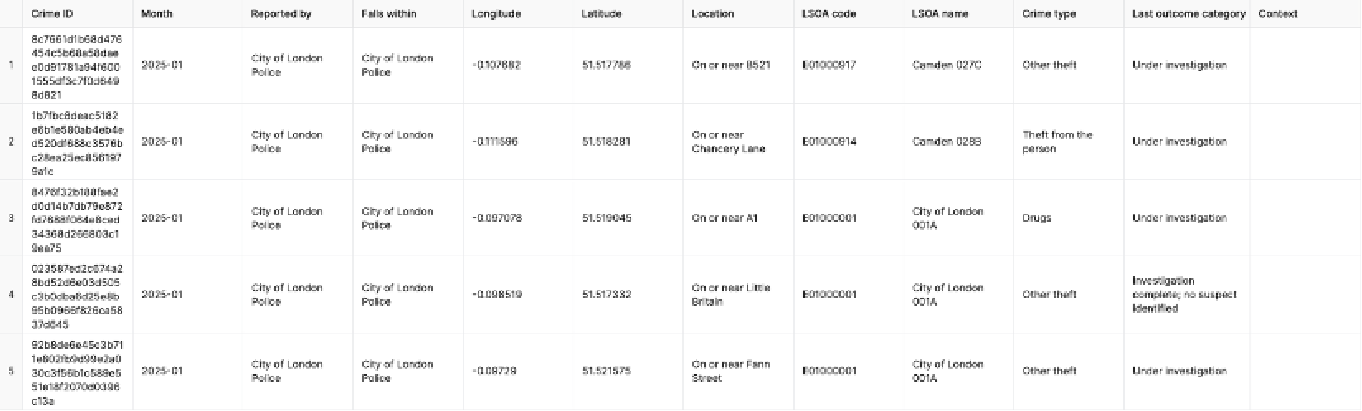
Through this approach, the project aims to enable data-driven decision-making by law enforcement agencies and policymakers, promote transparency, and aid in proactive crime prevention.

3. Dataset

The dataset used in this project is from the official ***City of London crime data repository***, available at `https://data.police.uk/data/`. It includes over 6,000 street-level crime incidents recorded between 2022 and 2024. The dataset is divided into two main files: crime records and outcome records.

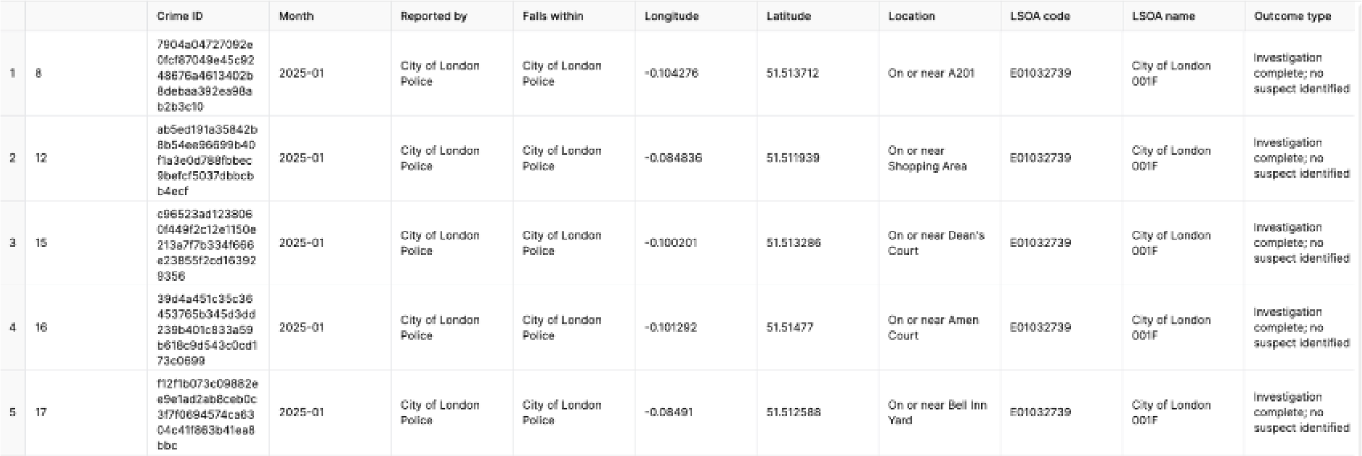
The crime records contain key attributes such as Crime ID, date(month and year), location details (including latitude and longitude), jurisdiction, and crime type(e.g., theft, assault, burglary). The outcome records complement this by providing the final status or resolution of each case, such as *”under investigation”* or *”no further action”*.

Before analysis, the dataset is cleaned to remove incomplete or duplicate records and ensure consistency across fields. Only records with valid and complete information such as Crime ID, location, category, outcome, and coordinates were retained. This clean, structured dataset forms the backbone of all queries and visualizations throughout the project.



	Crime ID	Month	Reported by	Falls within	Longitude	Latitude	Location	LSOA code	LSOA name	Crime type	Last outcome category	Context
1	8c7661d1b68d476454c5b68e58deec0d91781a94f6001555df3c7f0d6498d821	2025-01	City of London Police	City of London Police	-0.107882	51.517786	On or near B521	E01000917	Camden 027C	Other theft	Under investigation	
2	1b7fbc8deac5182e6b1e580ab4eb4ed520df688c3576bc28ea75ec8561979a1c	2025-01	City of London Police	City of London Police	-0.111586	51.518281	On or near Chancery Lane	E01000914	Camden 028B	Theft from the person	Under investigation	
3	8476f32b188fee2d0d14b7db79e872fd7688f064e8ced34368d266803c19ea75	2025-01	City of London Police	City of London Police	-0.097078	51.519045	On or near A1	E01000001	City of London 001A	Drugs	Under investigation	
4	023587ed2c874a28bd52d6e03d505c3b0dba6d25e8b95b0966f826ca5837d645	2025-01	City of London Police	City of London Police	-0.098519	51.517332	On or near Little Britain	E01000001	City of London 001A	Other theft	Investigation complete; no suspect identified	
5	92b8de6e45c3b711e802fb9d99e2a030c3f56b1c588e551e18f2070d0398c13a	2025-01	City of London Police	City of London Police	-0.08729	51.521575	On or near Fann Street	E01000001	City of London 001A	Other theft	Under investigation	

Figure 3.1: Crime Records



		Crime ID	Month	Reported by	Falls within	Longitude	Latitude	Location	LSOA code	LSOA name	Outcome type
1	8	7904a04727092e0fcf87049e45c9248676a4613402b8debaa392ea98ab2b3c10	2025-01	City of London Police	City of London Police	-0.104276	51.513712	On or near A201	E01032739	City of London 001F	Investigation complete; no suspect identified
2	12	ab5ed191a35842b8b54ee96699b40f1a3e0d788fbbec9befcf5037dbbcbb4ecf	2025-01	City of London Police	City of London Police	-0.084836	51.511939	On or near Shopping Area	E01032739	City of London 001F	Investigation complete; no suspect identified
3	15	c96523ad1238060f449f2c12e1150e213a7f7b334f666e23855f2cd163929356	2025-01	City of London Police	City of London Police	-0.100201	51.513286	On or near Dean's Court	E01032739	City of London 001F	Investigation complete; no suspect identified
4	16	39d4a451c35c36453765b345d3dd239b401c833a59b618c9d543c0cd173c0699	2025-01	City of London Police	City of London Police	-0.101282	51.51477	On or near Amen Court	E01032739	City of London 001F	Investigation complete; no suspect identified
5	17	f12f1b073c09882ee9e1ad2ab8ceb0c3f7f0694574ca6304c41f863b41ea8bbc	2025-01	City of London Police	City of London Police	-0.08491	51.512588	On or near Bell Inn Yard	E01032739	City of London 001F	Investigation complete; no suspect identified

Figure 3.2: Outcome Records

4. Methodology

The project was carried out in four key phases: Data preprocessing, Neo4J graph database setup, PostgreSQL relational database setup, and execution of use case queries.

4.1 Data Preprocessing

The raw dataset, sourced from the City of London's crime repository, was first cleaned to ensure accuracy and consistency. The following key steps were taken:

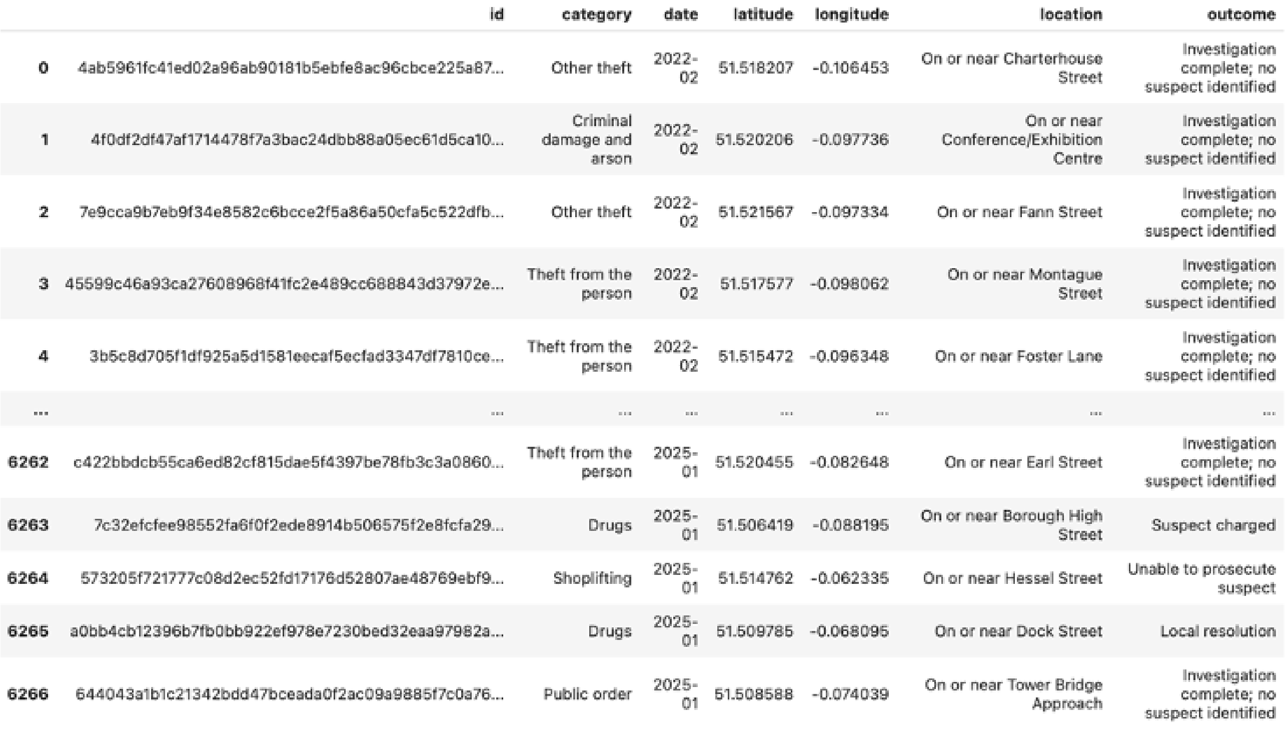
- **Handling Missing or Null Values:** To ensure data reliability, only valid (non-null) values were processed. Records with missing essential attributes such as Crime

ID, Location, Category, Date, Latitude, or Longitude were discarded. The Outcome Type field was also verified and assigned only if it contained a valid entry.

- **String Cleaning and Trimming:** All text fields, including Crime ID, Location, and Category, were stripped of leading and trailing spaces to maintain consistency. The Outcome Type field was also cleaned before mapping to ensure uniform formatting.

- **Mapping Crime Outcomes to Crime IDs:** A dictionary-based mapping approach was used to associate Crime IDs with their respective Outcome Types. This method prevented duplicate data processing and enabled efficient retrieval of outcome information for each crime record.

- **Ensuring Complete Records Before Appending:** A crime record was added to the database only if all key attributes were present, ensuring a clean and complete dataset. A valid record requires the presence of a Crime ID, Location, Crime Category, Outcome Type, Latitude, Longitude, and Date. Any records missing these fields were discarded to maintain the dataset's integrity and reliability.



	id	category	date	latitude	longitude	location	outcome
0	4ab5961fc41ed02a96ab90181b5ebfe8ac96cbce225a87...	Other theft	2022-02	51.518207	-0.106453	On or near Charterhouse Street	Investigation complete; no suspect identified
1	4f0df2df47af1714478f7a3bac24dbb88a05ec61d5ca10...	Criminal damage and arson	2022-02	51.520206	-0.097736	On or near Conference/Exhibition Centre	Investigation complete; no suspect identified
2	7e9cca9b7eb9f34e8582c6bcce2f5a86a50cfa5c522dfb...	Other theft	2022-02	51.521567	-0.097334	On or near Fann Street	Investigation complete; no suspect identified
3	45599c46a93ca27608968f41fc2e489cc688843d37972e...	Theft from the person	2022-02	51.517577	-0.098062	On or near Montague Street	Investigation complete; no suspect identified
4	3b5c8d705f1df925a5d1581eecaf5ecfad3347df7810ce...	Theft from the person	2022-02	51.515472	-0.096348	On or near Foster Lane	Investigation complete; no suspect identified
...	...	...	...	...	...	...	...
6262	c422bbdcb55ca6ed82cf815dae5f4397be78fb3c3a0860...	Theft from the person	2025-01	51.520455	-0.082648	On or near Earl Street	Investigation complete; no suspect identified
6263	7c32efcfee98552fa6f0f2ede8914b506575f2e8fcfa29...	Drugs	2025-01	51.506419	-0.088195	On or near Borough High Street	Suspect charged
6264	573205f721777c08d2ec52fd17176d52807ae48769ebf9...	Shoplifting	2025-01	51.514762	-0.062335	On or near Hessel Street	Unable to prosecute suspect
6265	a0bb4cb12396b7fb0bb922ef978e7230bed32eaa97982a...	Drugs	2025-01	51.509785	-0.068095	On or near Dock Street	Local resolution
6266	644043a1b1c21342bdd47bceada0f2ac09a9885f7c0a76...	Public order	2025-01	51.508588	-0.074039	On or near Tower Bridge Approach	Investigation complete; no suspect identified

Figure 4.3: Dataset post preprocessing

4.2 Neo4J graph database setup

To efficiently store and analyze crime data, Neo4J was set up as a graph database, ensuring data integrity, preventing duplication, and optimizing query performance. The setup process involved the following key steps:

- **Defining Constraints for data integrity:** To maintain a structured and efficient database, unique constraints were applied to prevent redundancy and ensure data accuracy. Unique constraints were enforced on places using latitude and longitude, ensuring that each recorded location is distinct. Similarly, crime categories were assigned unique identifiers to standardize classifications across the dataset. Time-based entries were structured to have unique year and month values, allowing for seamless chronological queries. Each crime outcome was uniquely mapped to cases, ensuring that resolution statuses were not duplicated. Finally, unique crime cases were maintained, preventing duplicate records and ensuring consistency in crime reporting.

- **Structuring crime data into graph format:** The dataset was transformed into a graph structure to enhance relationship-based analysis and querying. To ensure uniformity across different data sources, column names were standardized before insertion into the database. Additionally, date values were extracted into separate year and month fields, making it easier to analyze temporal trends. Crime records were then mapped to essential entities such as locations, categories, outcomes, and time. This structuring allowed for an efficient representation of crime data in Neo4J, enabling a more intuitive way to explore relationships between different elements.

- **Data insertion:** To handle large volumes of crime data efficiently, the dataset was processed in batches of 1000 records at a time. This batch-processing approach optimized performance by reducing the computational load during data insertion. Furthermore, error-handling mechanisms were implemented to ensure data consistency; problematic entries were automatically skipped instead of causing the entire batch process to fail. This ensured a smooth and robust data ingestion pipeline, allowing large datasets to be incorporated into the Neo4J database without significant disruptions.

- **Creating relationships between entities:** Neo4J's graph structure was leveraged to establish meaningful relationships between crime records. Each crime occurrence was linked to a specific place, making location-based queries more intuitive. Crimes were also associated with time by connecting them to their respective years and months, facilitating chronological analysis. Additionally, crimes were classified by type, enabling users to explore patterns within different crime categories such as theft, assault, and burglary. Lastly, crime cases were mapped to their corresponding outcomes, providing insight into law enforcement resolutions. These relationships enhanced the analytical capabilities of the database, making it easier to identify crime trends and patterns.

4.3 PostgreSQL relational database setup

To efficiently store and manage crime data, PostgreSQL was used as the relational database system. This setup involved the following key steps:

- **Connecting to PostgreSQL:** The connection to PostgreSQL was established using the `psycopg2` library in Python. This allowed interaction with the database by creating a cursor to execute queries. The connection ensured secure and efficient communication with the PostgreSQL server, enabling seamless data insertion and retrieval.

- **Creating the crime_data Table:** A table named crime_data was created to store crime records with a structured schema. The table includes multiple fields, each serving a specific purpose. The id column acts as the primary key and is auto-incremented for unique identification. The location field stores the textual description of where the crime occurred, while the date field records the time of occurrence in a date format. The category field classifies crimes based on type, and the outcome field captures the resolution or status of the crime incident.

- **Inserting Data into PostgreSQL:** Data was inserted into the crime_data table using the SQLAlchemy library, which provides an ORM (Object Relational Mapper) interface for seamless interaction with the database. This approach ensured that data was efficiently stored while maintaining the integrity and consistency of the database. SQLAlchemy's session-based handling allowed batch insertion and minimized performance overhead.

- **Ensuring Data Integrity and Performance:** To optimize data handling, the database was structured with indexing techniques and constraints that prevent duplication and maintain referential integrity. The id column was set as a primary key to enforce uniqueness, while required fields like location, date, category, and outcome were marked as NOT NULL to prevent incomplete data entries. This ensured efficient query execution and reliable data retrieval for analytical purposes.

5. Use Cases

5.1 Use cases leveraging Neo4J only

5.1.1 Top 10 crime hotspots based on frequency

This query is designed to identify crime hotspots by selecting locations where crimes have occurred in more than one distinct month and have more than five total reported crimes. By analyzing crime occurrences over time, the system can highlight areas that experience recurring criminal activities.

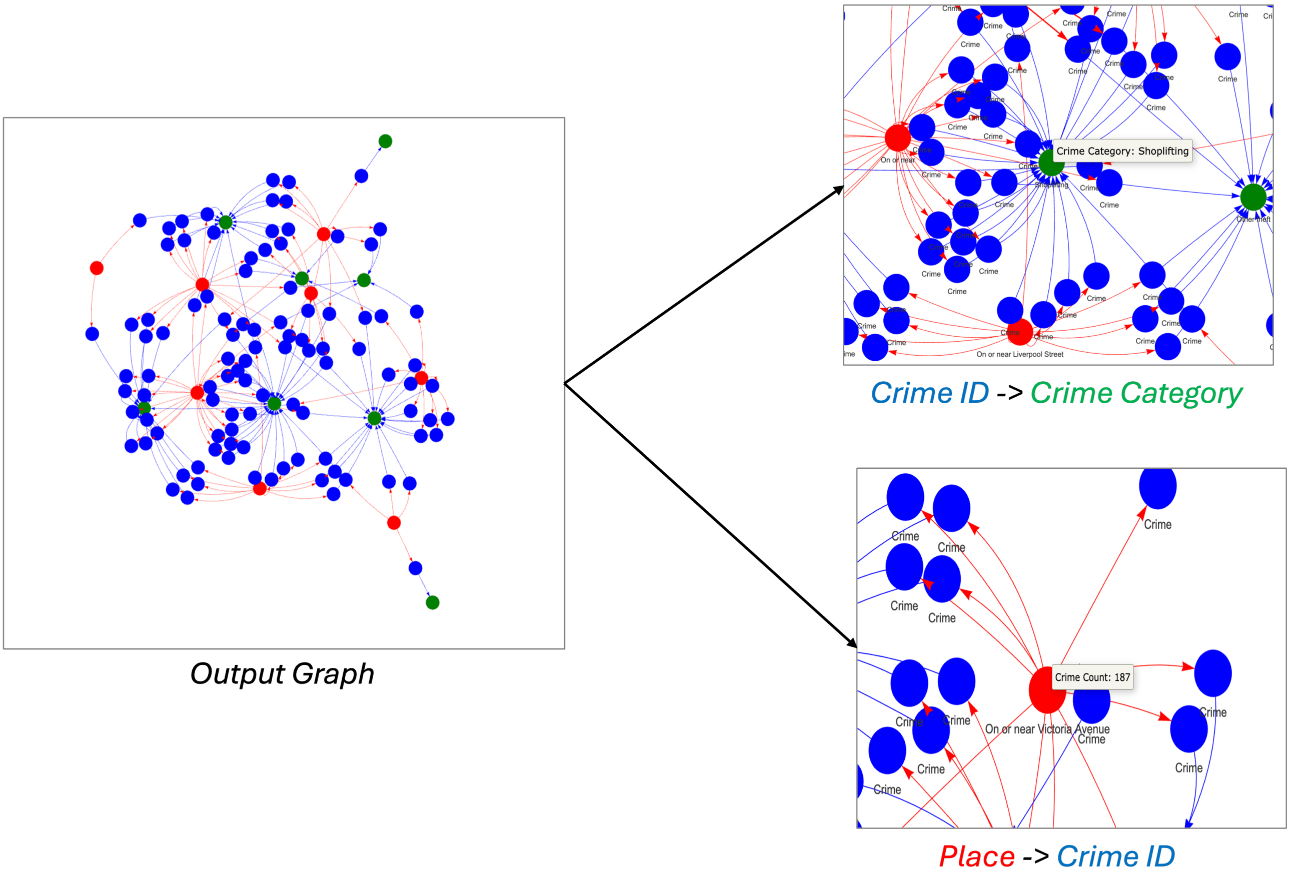
The query retrieves the top 10 such locations and fetches details of the latest 100 crimes associated with them. This includes information such as crime category, the month of occurrence, and the year. This approach enables law enforcement and analysts to focus on high-risk areas that require attention.

The MATCH clause in Cypher is used to model real-world relationships, such as "Crime occurred at Place" or "Crime belongs to a Category." Unlike PostgreSQL, which would require complex joins and indexing strategies to achieve similar results, Neo4J efficiently handles such queries by leveraging its graph-based architecture.

In the output visualization, both nodes and edges represent specific elements of the crime data. The **nodes** are colour-coded for clarity:

- Red Nodes represent places identified as crime hotspots.
- Blue Nodes indicate individual crimes.
- Green Nodes represent crime categories such as theft, assault, and burglary.

The **edges** establish relationships between these nodes. A red node (place) connects to a blue node (crime), labeled as ”Occurred At,” signifying that the crime took place at that specific location. A blue node (crime) further connects to a green node (crime category), labeled as ”Crime Type,” representing the classification of the crime. This structured representation enables efficient pattern detection, making it easier to analyze crime trends and associations.


Crime Category: Shoplifting
On or near Liverpool Street
Crime
Crime Count: 187
On or near Victoria Avenue
Crime

Output Graph

Crime ID -> Crime Category

Place -> Crime ID

Figure 5.4: Top 10 crime hotspots

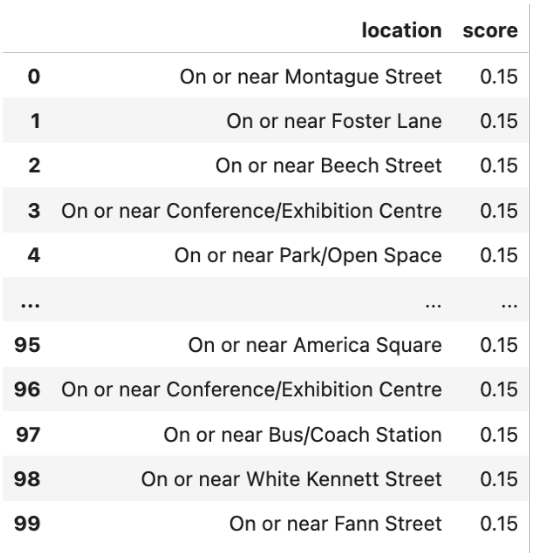
5.1.2 Using PageRank algorithm to rank places by their importance based on crime connections

This query projects a graph in Neo4j where locations are connected by crimes using the ”OCCURRED_AT” relationship. By analyzing how different places are linked through crime occurrences, the PageRank algorithm is applied to determine the most influential crime hotspots based on their connectivity.

The PageRank algorithm is specifically used to rank the top 100 locations with the highest influence in crime-related activities. The ranking is determined by analyzing how crimes interconnect between different locations, prioritizing places with a higher number of linked incidents. This method provides deeper insight into crime centrality and helps identify key locations that require focused law enforcement attention.

Neo4j proves to be more efficient for this type of analysis compared to PostgreSQL, which would require complex recursive Common Table Expressions (CTEs) and multiple join operations to achieve similar results. Neo4j's graph traversal capabilities allow it to process relationships directly, eliminating the need for expensive joins and recursive queries. This makes the approach significantly faster and more scalable when analyzing crime centrality and hotspot influence.

The output of this query is a ranked list of the **top 100 crime hotspots**, displaying each **location's name** alongside its **PageRank score**. The PageRank score represents the influence of each location based on its connectivity to crime incidents. Locations with higher scores indicate frequent and highly connected crimes, making them critical areas for crime analysis. This ranking allows authorities to prioritize investigations and allocate resources efficiently to the most impacted locations.



	location	score
0	On or near Montague Street	0.15
1	On or near Foster Lane	0.15
2	On or near Beech Street	0.15
3	On or near Conference/Exhibition Centre	0.15
4	On or near Park/Open Space	0.15
...	...	...
95	On or near America Square	0.15
96	On or near Conference/Exhibition Centre	0.15
97	On or near Bus/Coach Station	0.15
98	On or near White Kennett Street	0.15
99	On or near Fann Street	0.15

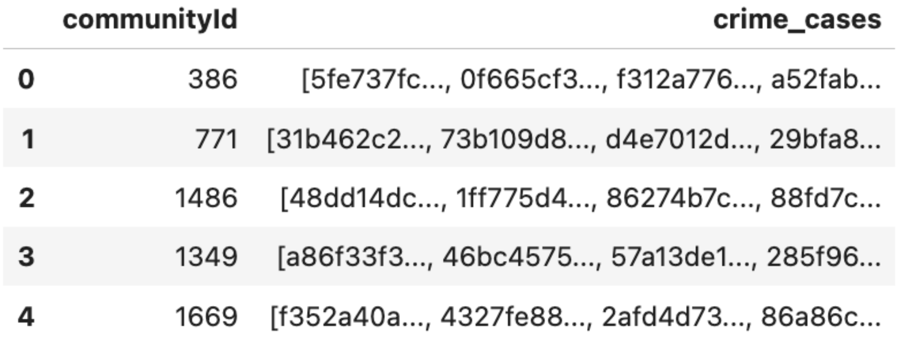
Figure 5.5: Locations ranked on the basis of their crime connections

5.1.3 Using Louvain algorithm to identify clusters of interconnected crimes, ranking the top 5 largest crime networks

The Louvain algorithm is applied in this query to detect crime communities by grouping closely connected crimes into clusters. By leveraging Neo4j's graph-based data model, this approach efficiently identifies patterns of interconnected criminal activities, revealing areas where crimes tend to form structured networks.

The query returns the **top 5 largest crime communities**, listing their community ID along with the associated crime case IDs. This helps in detecting crime clusters based on shared locations, involved individuals, or recurring patterns, enabling better crime analysis.

Identifying these clusters provides valuable insights into criminal behaviour and law enforcement strategies. By recognizing hotspots with frequent offences and understanding how crimes are interconnected, authorities can focus resources on disrupting recurring crime networks. This method enhances predictive policing and improves response strategies for crime prevention.



	communityId	crime_cases
0	386	[5fe737fc..., 0f665cf3..., f312a776..., a52fab...
1	771	[31b462c2..., 73b109d8..., d4e7012d..., 29bfa8...
2	1486	[48dd14dc..., 1ff775d4..., 86274b7c..., 88fd7c...
3	1349	[a86f33f3..., 46bc4575..., 57a13de1..., 285f96...
4	1669	[f352a40a..., 4327fe88..., 2afd4d73..., 86a86c...

Figure 5.6: Inconnected Crime Clusters

5.1.4 Top Co-Occurring Crime Categories: Patterns Associations

This query identifies crime categories that frequently co-occur at the same locations by analyzing pairs of crimes linked to the same place. It ranks the top 50 category pairs based on the number of locations where both crimes have been reported together. This approach helps in understanding crime patterns and relationships between different types of offences.

By analyzing these associations, law enforcement can gain insights into which crime types tend to occur together. This can aid in developing more effective crime prevention strategies, resource allocation, and policy-making to address specific crime clusters.

The output of the query is represented as a graph where **nodes** indicate crime categories. Each node represents a specific type of crime, such as theft, assault, or burglary. The **edges** between nodes signify a connection between two crime categories, meaning they have frequently co-occurred at the same place.

Additionally, **edge thickness** plays a crucial role in visualizing crime associations. The thicker the edge between two crime categories, the higher the number of locations where those crimes have occurred together. This highlights stronger crime associations and provides useful information for identifying high-risk crime patterns.

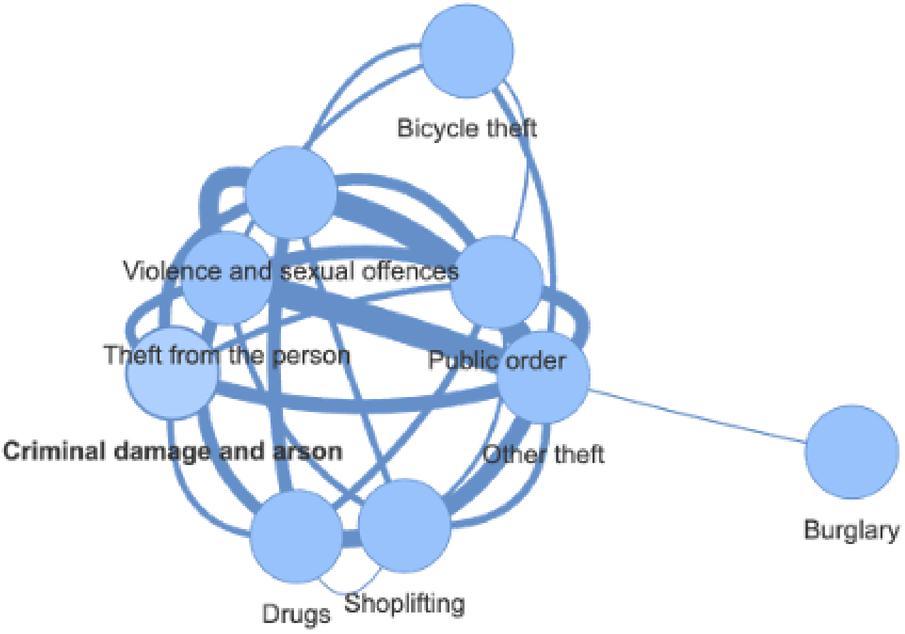


Figure 5.7: Top Co-Occurring Crime Categories

5.1.5 Crime case outcomes over time

This query retrieves crime case outcomes over time by counting the number of cases that resulted in each outcome type per month and year. By analyzing how cases are resolved over time, law enforcement agencies and policymakers can track changes in crime resolutions and identify trends in judicial and investigative outcomes.

The data is organized chronologically and visualized using a line graph, where each line represents a different outcome type. This graphical representation allows for an intuitive understanding of how various case resolutions, such as convictions, dismissals, or ongoing investigations, fluctuate over time. By examining these trends, stakeholders can assess the effectiveness of law enforcement interventions and judicial processes.

Understanding these changes over time is crucial for evaluating the efficiency of the justice system. Fluctuations in the number of resolved or unresolved cases can indicate shifts in investigative efficiency, policy changes, or variations in crime patterns. Additionally, the month-by-month breakdown of crime outcomes provides actionable insights that can aid in resource allocation and decision-making.

By leveraging this data-driven approach, authorities can enhance their ability to monitor judicial processes, ensure transparency, and optimize strategies for crime prevention and law enforcement. This methodology ultimately supports better crime resolution tracking and improved justice system efficiency.

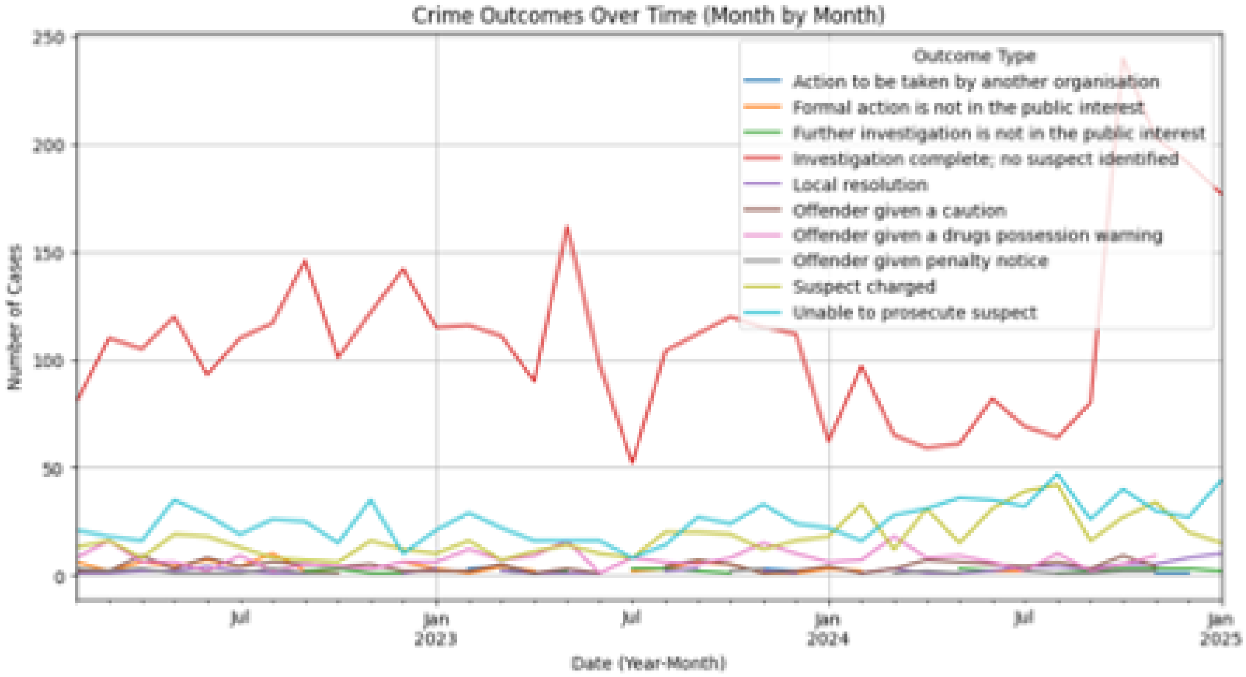


Figure 5.7: Crime case outcomes over time

5.1.6 Seasonal crime trends

This query analyzes seasonal crime trends by comparing the number of crimes that occurred during winter (December to February) and summer (June to August) across different locations. The goal is to uncover how crime patterns fluctuate depending on the time of year.

The query identifies the top 10 locations that exhibit the most significant differences in crime activity between the winter and summer months. It compares the crime counts for each season at these locations to highlight areas with noticeable seasonal variations.

The results are visualized using a stacked bar chart. This visualization technique emphasizes locations where crime levels change notably between seasons. By highlighting such seasonal fluctuations, the chart provides valuable insight into how criminal behaviour may be influenced by external factors such as weather, time of year, or social patterns.

Understanding these patterns is essential for law enforcement agencies, as it can help them allocate resources more effectively and implement timely crime prevention strategies. Seasonal insights like these are especially useful in proactive policing and urban safety planning.

In the visualization, blue bars represent crimes that occurred in winter, while orange bars represent those that occurred in summer. The difference in bar height reflects the scale of variation, helping to depict the seasonal impact on crime distribution across various locations.

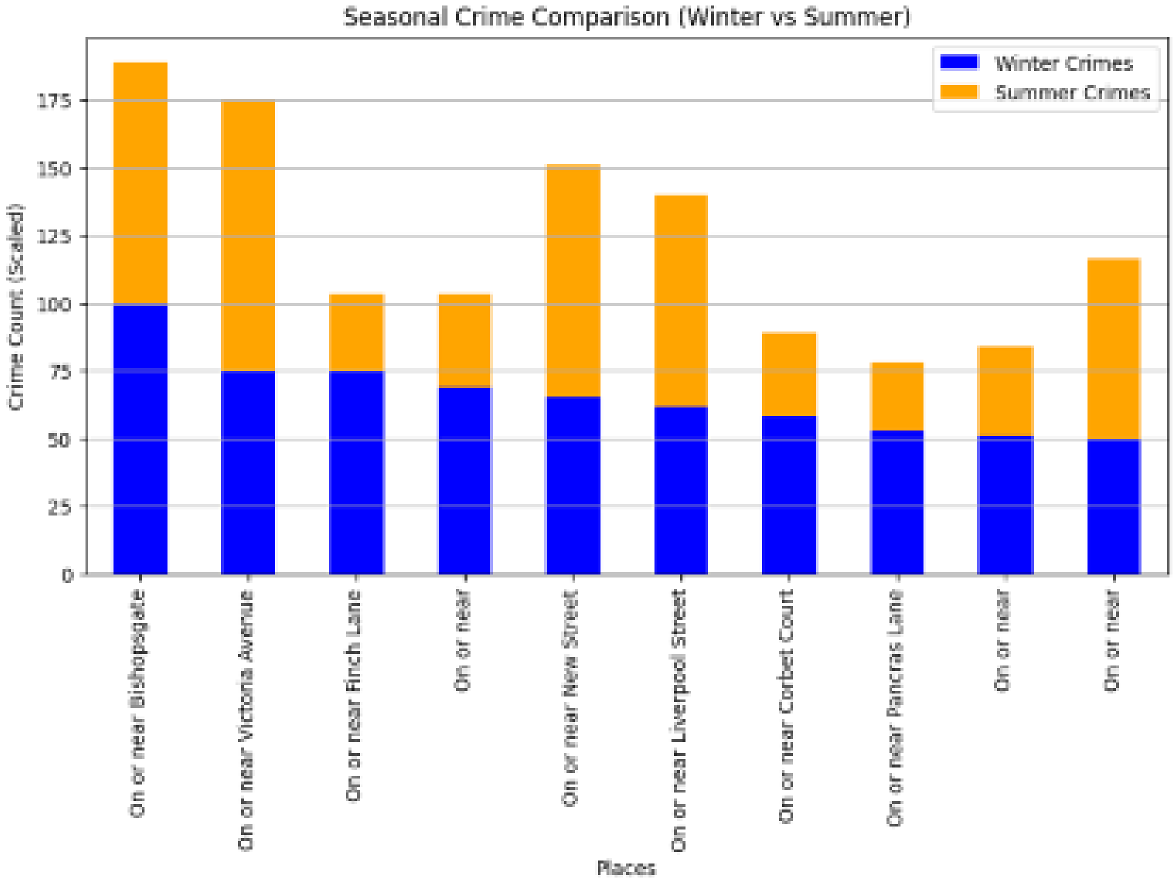


Figure 5.7: Seasonal crime trends

5.1.7 Identifying interconnected crime cases

This query focuses on identifying *multi-hop relationships* between different crime cases by linking them through shared locations and crime categories. It aims to uncover hidden connections between cases that may not be immediately obvious through traditional analysis.

The approach begins by identifying crime cases that occurred at the same place. These cases are then categorized by their crime type. Using this information, the system establishes a three-hop connection between different cases, relying on intermediate relationships to bridge the gaps. This method enables the detection of complex chains of criminal activity.

The results are visualized in a network graph where each node represents a distinct element: crime cases, shared locations, or crime categories. Through this graph, interconnected patterns of criminal activity emerge, offering valuable insights into how seemingly separate incidents may be part of broader crime networks.

In the visualization, different node colours represent different entities: **red nodes** represent crime cases, **orange nodes** represent intermediate cases, **blue nodes** indicate shared locations, and **green nodes** represent shared crime categories.

The edges between these nodes represent specific types of relationships. The **"Occurred At"** edge connects a crime case to its location. The **"Connected By Place"** edge links cases that occurred at the same location. The **"Linked By Category"** edge shows connections between cases sharing the same crime type. Finally, the **"Connected Case"** edge represents the final link in the multi-hop relationship, connecting initially distant cases through shared characteristics.

This query provides a powerful way to detect potential patterns of repeat or coordinated offenses, helping law enforcement agencies trace broader criminal connections and improve investigative efficiency.

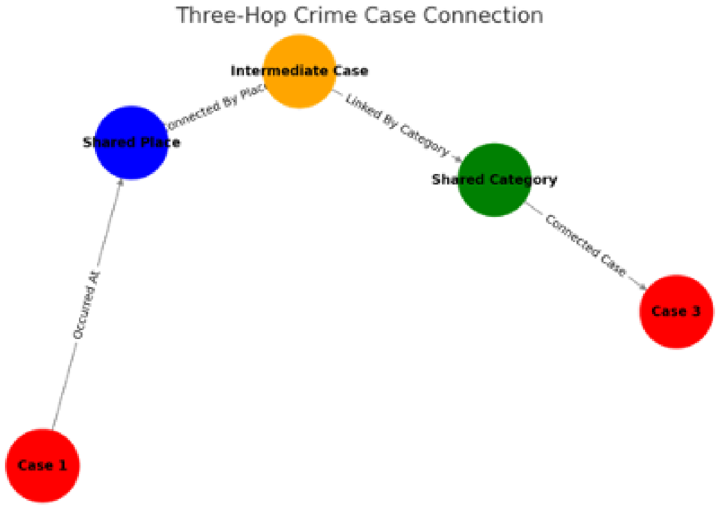


Figure 5.7: Three Hop Example

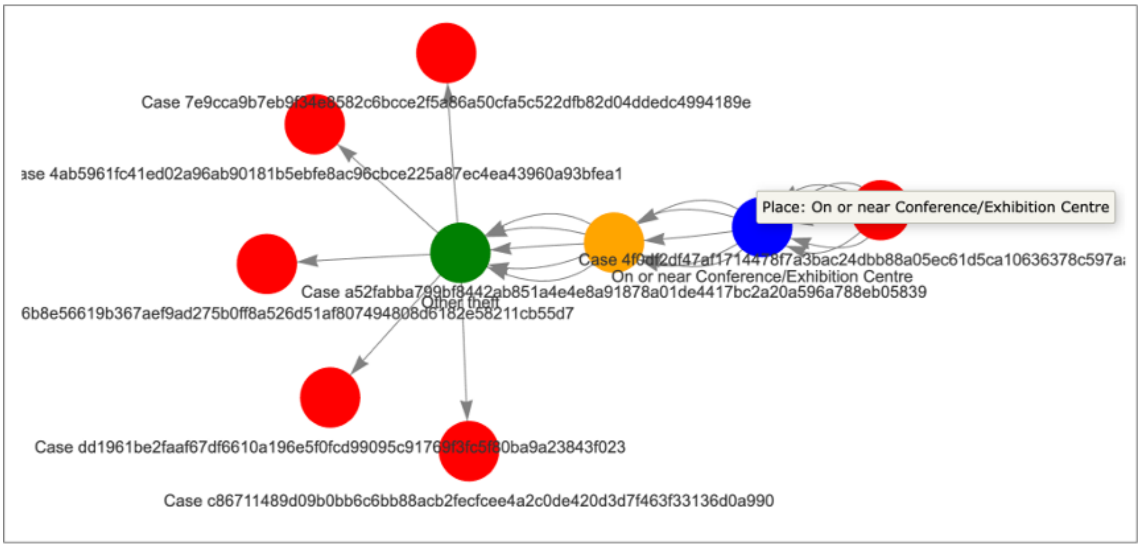

Case 7e9cca9b7eb9f34e8582c6bcce2f5a86a50cfa5c522dfb82d04ddedc4994189e
ase 4ab5961fc41ed02a96ab90181b5ebfe8ac96cbce225a87ec4ea43960a93bfea1
Place: On or near Conference/Exhibition Centre
Case 4f0df2df47af1714478f7a3bac24dbb88a05ec61d5ca10636378c597a
On or near Conference/Exhibition Centre
Case a52fabba799bf8442ab851a4e4e8a91878a01de4417bc2a20a596a788eb05839
Other theft
6b8e56619b367aef9ad275b0ff8a526d51af807494808d6182e58211cb55d7
Case dd1961be2faaf67df6610a196e5f0fcd99095c91769f3fc5f80ba9a23843f023
Case c86711489d09b0bb6c6bb88acb2fecfcee4a2c0de420d3d7f463f33136d0a990

Figure 5.7: Interconnected crime cases

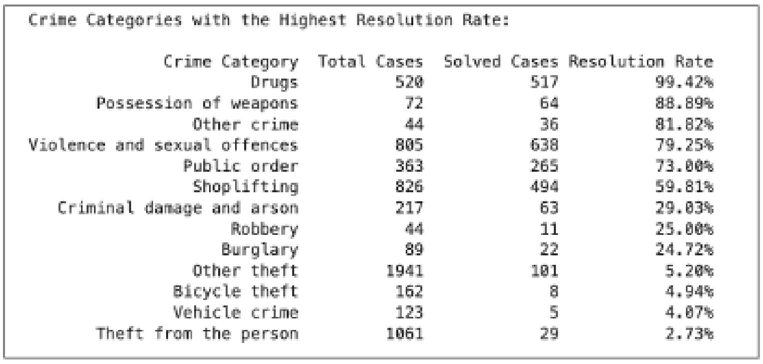
5.2 Use cases leveraging PostgreSQL only

5.2.1 Crime Categories with the Highest Resolution Rate

This query retrieves crime data and groups it by category to calculate the total number of cases, the number of solved cases, and the resolution rate for each category. The resolution rate is computed as the percentage of solved cases relative to the total number of cases in that category.

Unresolved cases are excluded from the solved count to ensure accuracy. Once calculated, the results are sorted in descending order by resolution rate, highlighting the crime categories with the highest success in resolution. This analysis provides valuable insights into which types of crimes are most effectively handled by law enforcement.

The output includes categories such as drugs and possession of weapons, which demonstrate high-resolution rates, and contrasts them with categories like vehicle crime or theft from the person, which have notably lower resolution percentages.



```
Crime Categories with the Highest Resolution Rate:

              Crime Category  Total Cases  Solved Cases Resolution Rate
                       Drugs          520           517          99.42%
       Possession of weapons           72            64          88.89%
                 Other crime           44            36          81.82%
Violence and sexual offences          805           638          79.25%
                Public order          363           265          73.00%
                 Shoplifting          826           494          59.81%
   Criminal damage and arson          217            63          29.03%
                     Robbery           44            11          25.00%
                    Burglary           89            22          24.72%
                 Other theft         1941           101           5.20%
               Bicycle theft          162             8           4.94%
               Vehicle crime          123             5           4.07%
       Theft from the person         1061            29           2.73%
```

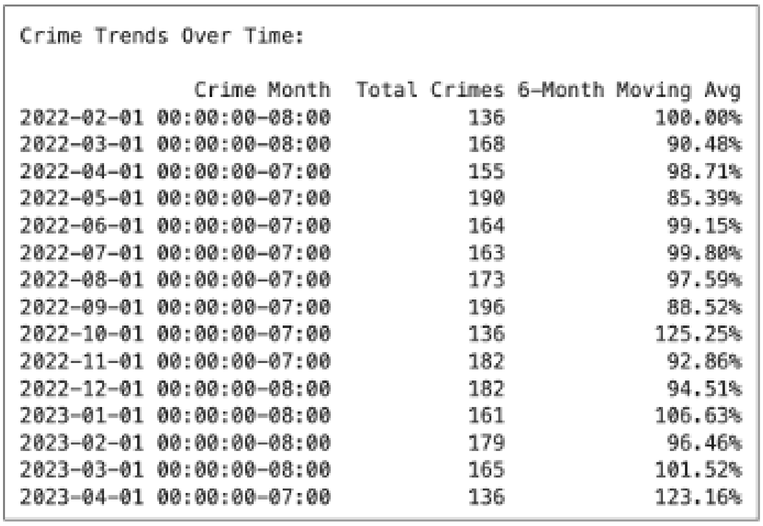
Figure 5.7: Crime Categories with the Highest Resolution Rate

5.2.2 Predicting Crime Trends Using Rolling Averages

This query analyzes crime trends over time by calculating the total number of crimes per month and applying a 6-month moving average to smooth out short-term fluctuations. This helps to identify long-term patterns in criminal activity.

The data is first aggregated by month using a common table expression (CTE), after which a moving average window function is applied. This statistical method helps in eliminating random spikes or drops in crime numbers, offering a clearer view of sustained trends.

The results show how crime rates evolve month-to-month, enabling analysts and policymakers to better forecast future patterns and allocate resources more effectively. This query provides a foundation for time-based predictive analysis in crime prevention strategies.



```
Crime Trends Over Time:

                 Crime Month  Total Crimes 6-Month Moving Avg
2022-02-01 00:00:00-08:00           136             100.00%
2022-03-01 00:00:00-08:00           168              90.48%
2022-04-01 00:00:00-07:00           155              98.71%
2022-05-01 00:00:00-07:00           190              85.39%
2022-06-01 00:00:00-07:00           164              99.15%
2022-07-01 00:00:00-07:00           163              99.80%
2022-08-01 00:00:00-07:00           173              97.59%
2022-09-01 00:00:00-07:00           196              88.52%
2022-10-01 00:00:00-07:00           136             125.25%
2022-11-01 00:00:00-07:00           182              92.86%
2022-12-01 00:00:00-08:00           182              94.51%
2023-01-01 00:00:00-08:00           161             106.63%
2023-02-01 00:00:00-08:00           179              96.46%
2023-03-01 00:00:00-08:00           165             101.52%
2023-04-01 00:00:00-07:00           136             123.16%
```

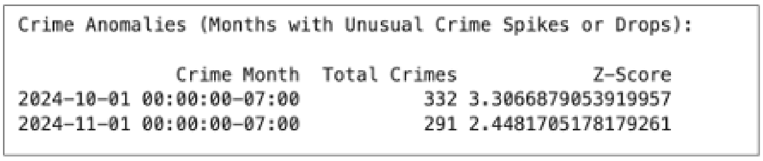
Figure 5.7: Crime Trends Using Rolling Averages

5.2.3 Identifying High Crime Time Series Anomalies

This query analyzes monthly crime data to detect anomalies using Z-scores. By applying statistical anomaly detection, it identifies months with unusually high or low crime counts compared to historical trends.

The process begins by calculating the total number of crimes recorded per month. The average and standard deviation of crime counts are then computed to establish a baseline for comparison. Each month's crime count is standardized using the Z-score formula, allowing the system to determine how far a particular month's crime level deviates from the mean.

Months with an absolute Z-score greater than 2 are flagged as significant anomalies, indicating substantial crime spikes or drops. The results are sorted in descending order of severity, highlighting months with extreme fluctuations in crime activity. Identifying these anomalies enables law enforcement and policymakers to investigate underlying causes, such as policy changes, major events, or seasonal variations, and respond accordingly.



```
Crime Anomalies (Months with Unusual Crime Spikes or Drops):

               Crime Month  Total Crimes           Z-Score
2024-10-01 00:00:00-07:00           332 3.3066879053919957
2024-11-01 00:00:00-07:00           291 2.4481705178179261
```

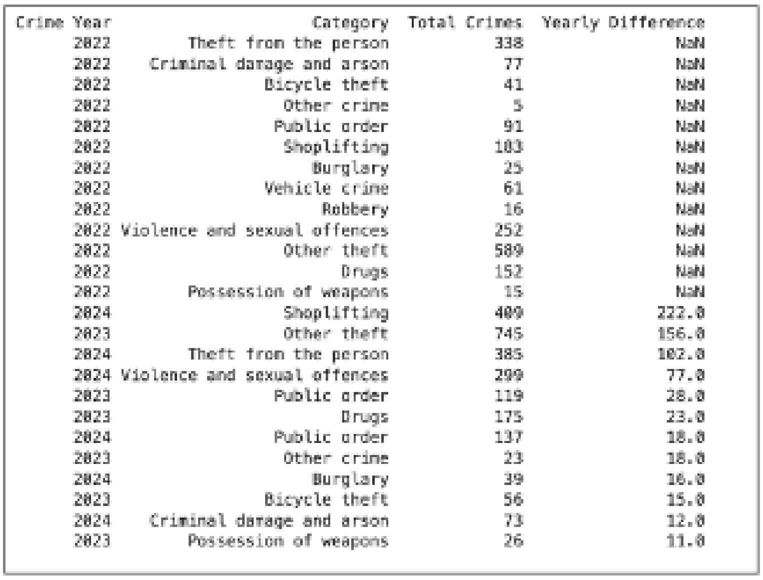
Figure 5.7: High Crime Anomalies

5.2.4 Crime Evolution Over the Years

This query analyzes yearly crime trends by category, calculating the total number of crimes recorded each year. The objective is to identify long-term trends and significant shifts in criminal activity.

The process involves computing the difference in crime counts from the previous year using a window function. By comparing yearly crime totals, this approach highlights categories experiencing the most notable increases or decreases in reported incidents.

The results are sorted in descending order based on the yearly difference, allowing analysts to focus on the top 25 crime categories with the most significant growth in incidents. This analysis provides valuable insights for law enforcement and policymakers, enabling them to develop strategic crime prevention measures and allocate resources more effectively based on evolving crime patterns.



```
Crime Year                     Category  Total Crimes  Yearly Difference
      2022        Theft from the person           338                NaN
      2022    Criminal damage and arson            77                NaN
      2022                Bicycle theft            41                NaN
      2022                  Other crime             5                NaN
      2022                 Public order            91                NaN
      2022                  Shoplifting           103                NaN
      2022                     Burglary            25                NaN
      2022                Vehicle crime            61                NaN
      2022                      Robbery            16                NaN
      2022 Violence and sexual offences           252                NaN
      2022                  Other theft           589                NaN
      2022                        Drugs           152                NaN
      2022        Possession of weapons            15                NaN
      2024                  Shoplifting           409              222.0
      2023                  Other theft           745              156.0
      2024        Theft from the person           385              102.0
      2024 Violence and sexual offences           299               77.0
      2023                 Public order           119               28.0
      2023                        Drugs           175               23.0
      2024                 Public order           137               18.0
      2023                  Other crime            23               18.0
      2024                     Burglary            39               16.0
      2023                Bicycle theft            56               15.0
      2024    Criminal damage and arson            73               12.0
      2023        Possession of weapons            26               11.0
```

Figure 5.7: Crime Evolution Over the Years

5.3 Use case leveraging PostgreSQL and Neo4J in combination

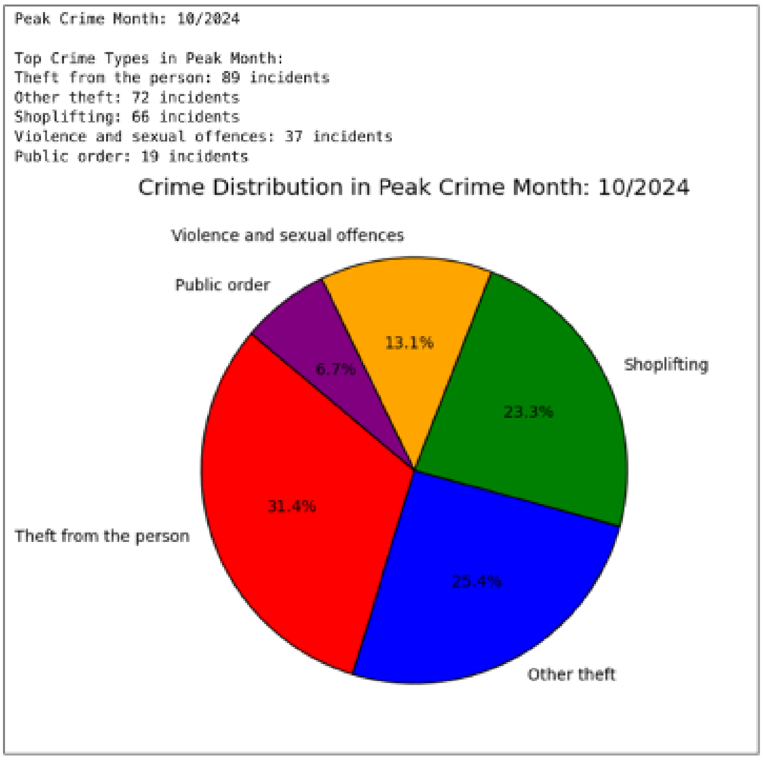
This query aims to identify the month with the highest crime rate and analyze the most common crime categories during that period. By combining the analytical power of PostgreSQL and Neo4J, this approach leverages both structured and graph-based data representations for deeper insights.

The process begins by querying a PostgreSQL database to determine the peak crime month. This is achieved by counting the total number of incidents recorded per month and identifying the period with the highest crime activity. PostgreSQL's robust aggregation functions make it efficient for extracting time-based crime trends.

Once the peak crime month is identified, the query retrieves the top five crime categories for that period using a Neo4J graph database. Crimes from the selected month are grouped by category, allowing for a structured understanding of crime distribution. Neo4J's ability to model relationships enhances the accuracy of identifying co-occurring crimes and patterns.

The results are then displayed through a pie chart visualization, illustrating the distribution of different crime types in the peak crime month. This graphical representation helps in understanding seasonal or periodic crime trends and assists law enforcement in allocating resources efficiently.

By integrating PostgreSQL for structured crime trend analysis and Neo4J for category-based crime relationships, this approach provides a comprehensive framework for crime analysis, allowing stakeholders to make informed decisions based on data-driven insights.



```
Peak Crime Month: 10/2024

Top Crime Types in Peak Month:
Theft from the person: 89 incidents
Other theft: 72 incidents
Shoplifting: 66 incidents
Violence and sexual offences: 37 incidents
Public order: 19 incidents
```

Figure 5.7: Crime distribution in the peak month

6. Future Work

In the future, this project can be improved in a number of very important ways. One giant leap is the addition of real-time crime data, allowing law enforcement to track crimes as they are happening and respond more rapidly. The system currently runs on historical data, but the addition of live feeds from crime records or open-data APIs would greatly enable it.

Another major enhancement is predictive modelling through machine learning. Based on historical crime trends, the system would be able to predict future hotspots, determine risk factors, and assist law enforcement in taking proactive measures before crimes are committed.

Enhanced visualization and reporting functions would also enhance usability. Interactive dashboards, heatmaps, and artificial intelligence (AI)-driven insights could make crime data more actionable for decision-makers. Integration with geographic information system (GIS) tools would facilitate richer spatial analysis of crime trends.

Finally, an expansion of the system to numerous cities from London would facilitate broader crime trend comparisons. More advanced graph analytics in Neo4J, including network-based crime detection and community analysis, would also facilitate improved understanding of organized crime patterns.