Final Project Report

INST 327 | 0303

May 13, 2025

Team 2: [redacted], Dunya Rahim, [redacted]

Introduction:

Crime influences how we view neighborhoods — from where we choose to live to how safe we feel walking home at night. Our project asks: do some areas in Montgomery County, Maryland, truly experience more crime, or are they perceived as dangerous due to stereotypes and reporting biases? By analyzing crime rates by zip code, we aim to uncover whether crime patterns reflect actual incidents or systemic inequalities in policing and reporting.

We're working with a large dataset from Montgomery County's Open Data portal, tracking over 300,000 reported crimes since 2016. This includes a wide range of offenses, from open cases to attempted crimes. By breaking the data down by zip code, we hope to identify trends — such as whether wealthier areas report crimes differently, or if certain public spaces are more vulnerable to specific offenses.

This report outlines our process of building a relational database to analyze these patterns. It covers our normalization decisions, challenges in schema design, and insights from complex SQL queries. We also reflect on ethical considerations, including how bias can shape public perception, and propose ways the database could be expanded to support deeper community analysis.

Database Description:

Normalization:

To effectively organize and analyze over 300,000 crime reports, we applied normalization principles to break the data into clean, non-redundant tables. Our goal was to eliminate duplication, reduce anomalies, and ensure each piece of information was stored only once. We followed normalization up to Third Normal Form (3NF), separating attributes based on functional dependency, removing partial dependencies, and avoiding transitive ones.

The original dataset was flat, with repeated values for agency, offense codes, zip codes, and crime types across thousands of rows. Through normalization, we identified and extracted recurring groups — like crime categories, locations, and police districts — into separate lookup tables such as crime_type, location, and agency. These were then linked to the main case_report

table using foreign keys. For instance, instead of repeating agency names, we stored them once in the agency table and referenced them via agency_id.

Table Name	Column Name	Data Type	Primary Key (PK)	Foreign Key (FK)	Nullable	Sample Data
Agency	agency_id	INT	Yes		No	1
Agency	agency_name	VARCHAR(100)			No	Montgomery PD
Crime_Type	crime_type_id	INT	Yes		No	1
Crime_Type	crime_category	VARCHAR(50)			No	Person
Offense	offense_code	INT	Yes		No	1234A
Offense	crime_name1	VARCHAR(45)			No	Person
Offense	crime_name2	VARCHAR(45)			Yes	Aggravated Assault
Offense	crime_name3	VARCHAR(45)			Yes	Assault with weapon
Offense	crime_type_id	INT		Crime_Type(crime_type_id)	No	1
Location	location_id	INT	Yes		No	1
Location	block_address	VARCHAR(100)			No	12603 Wisteria Drive
Location	city	VARCHAR(50)			No	Germantown
Location	zip_code	INT			No	20874
Location	state	VARCHAR(20)			No	MD
Location	district	VARCHAR(50)			No	Germantown District
Incident	incident_id	INT	Yes		No	123456
Incident	case_number	INT			Yes	CR20240405
Incident	start_date	DATETIME			No	2024-05-01 13:00
Incident	end_date	DATETIME			Yes	2024-05-01 14:00
Incident	dispatch_date	DATETIME			Yes	2024-05-01 12:55
Incident	victims	INT			No	1
Incident	agency_id	INT		Agency(agency_id)	No	1
Incident	location_id	INT		Location(location_id)	No	1
Incident	offense_code	VARCHAR(20)		Offense(offense_code)	No	1234A

Logical Design:

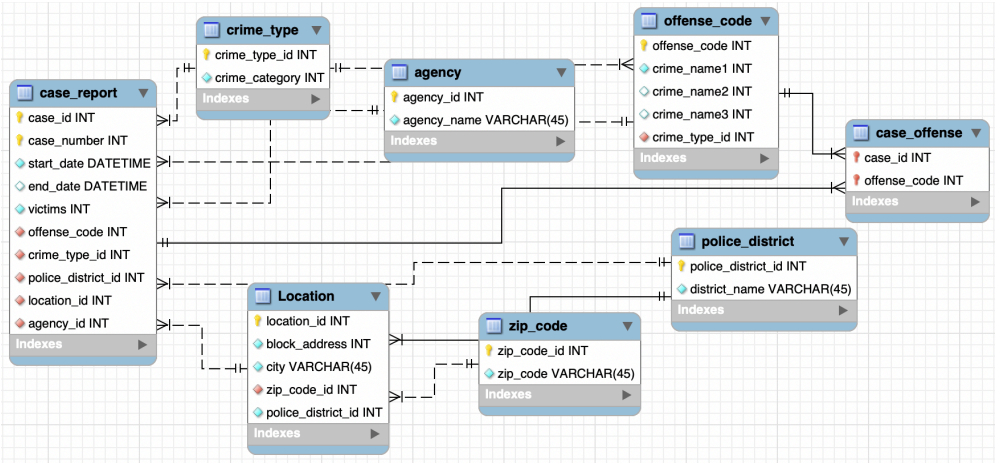
Our Entity-Relationship Diagram (ERD) represents the logical structure of our crime-tracking database, designed to model reported crime data in Montgomery County, Maryland. The ERD includes eight interconnected tables: case_report, agency, offense_code, crime_type, location, zip_code, police_district, and case_offense. Each table corresponds to a real-world entity relevant to understanding how, where, and by whom crimes are reported.

The case_report table serves as the primary fact table that captures individual crime incidents. It includes foreign keys linking to agency (which agency reported the case), crime_type (the broader classification of the crime), location (where the crime occurred), and police_district (jurisdictional coverage). To allow each case to be associated with multiple offenses, we introduced the case_offense table. This table creates a many-to-many relationship between case_report and offense_code by storing foreign keys to both, and uses a composite primary key to ensure each pair is unique. Since case_report links to location, and location connects to police_district, this structure allows us to trace individual offenses back to the geographic areas where they occurred, supporting more detailed and accurate analysis of crime patterns across districts.

These foreign key relationships allow the database to store crime details without redundancy and maintain consistency across related entities. Normalization principles guided our structure, particularly third normal form (3NF), to eliminate duplicate data, reduce anomalies,

and ensure referential integrity. Lookup tables like zip_code, police_district, and crime_type allow for future expansion without altering core table designs. All foreign key relationships are clearly established in the ERD using dotted lines, and each table includes a primary key to uniquely identify its records.

This ERD reflects a clean, efficient, and scalable design that supports complex querying while preserving the accuracy and clarity of the original dataset.



Physical Database

We built our physical database in MySQL Workbench based on our finalized ERD, using foreign keys to connect tables like Case_Report, Agency, and Case_Offense. This structure was shaped by the cleaned CSV files we had from the original CSV file for Montgomery County Crime (Montgomery County, 2025) and revised design choices made after testing relationships between entities.

Sample Data

We imported sample data directly from the original Montgomery County crime dataset, cleaning and organizing it by removing empty data in order to match our schema. Each table has at least 15 meaningful records, allowing us to run and test realistic queries for trends and insights.

CRUD Views/Queries Table:

View Name	Req. A	Req. B	Req. C	Req. D	Req. E
query_takomapark	X	X			X
query_rcpd	X	X			
query_dv	X	X			
query_cas	X	X			
query_citycount			X		
query_crimenamecount	X	X	X		
query_addresscount	X	X	X		
query_offense_linked	X			X	

Changes From Original Design:

One of the major changes from our original design was a complete restructuring of the ERD to align with normalization best practices. In our first attempt, we created tables without establishing any foreign key relationships, which resulted in isolated entities and redundant data across the schema. Additionally, some tables were connected in cyclical or illogical ways, and all relationships were represented using solid lines, even where they weren't linking tables. After receiving feedback, we re-evaluated the design and reorganized the schema based on the third normal form. This involved separating repeating or categorical data into their own tables, such as crime_type, zip_code, and police_district as well as linking them back to the case_report and location table the through foreign keys.

We also added surrogate keys (like crime_type_id and zip_code_id) in places where our original design used raw text or codes as primary identifiers. This made it easier to enforce referential integrity and reduced the chance of data inconsistencies. The revised ERD now clearly defines primary and foreign keys, uses dotted lines for non-linking relationships, and

avoids unnecessary duplication. These changes not only improved the clarity and correctness of our design but also ensured the database would be more scalable, maintainable, and aligned with relational database standards.

Database Ethics Considerations (Diversity Considerations, Target Audience, Privacy):

There are significant ethical obligations when working with crime data, particularly when it comes to confidential and at times excluding information associated with particular groups. Our team kept an eye out for any misuse or interpretation of the data throughout this project. The possibility of spreading negative stereotypes was one of our top concerns, especially how easy it is to categorize particular zip codes as "dangerous" based just on incident amounts. We acknowledged that these numbers might be deceptive in the absence of appropriate information, such as economic status or past law enforcement history.

In order to solve this issue, we purposefully kept clear of simpler classifications and concentrated on creating questions that encourage critical thinking. For instance, without making unjustified claims, we emphasized systemic themes, such as which agencies report the most or which regions have unusual patterns of crimes. There are major ethical considerations when working with crime statistics.

After finalizing our plan for the MoCo Crime Database, one of the biggest privacy concerns we had to think through was the use of people's addresses and zip codes. This kind of location data is obviously sensitive, but at the same time, it's necessary if we want to actually understand where crimes are happening and analyze patterns across Montgomery County.

As we built the database, we made sure we weren't just dumping in data blindly. We paid attention to whether addresses and zip codes were okay to include, and we kept in mind that even without names, this info could still feel personal. That's why we're clear that the database shouldn't be used to target anyone or leak private locations , that's not what this is for.

To protect the data, we're putting strong security measures in place and being mindful about who has access. We'll also make sure the database is only used for research purposes that are ethical and responsible. Including location info definitely adds value, but we wanted to make sure we did it the right way , carefully, thoughtfully, and with people's privacy in mind the whole time.

The target audience for this database includes anyone interested in gaining an accurate understanding of crime in Montgomery County. This could be individuals searching for a new

place to live who want to know whether a neighborhood is genuinely unsafe or simply misunderstood. It could also be civil rights advocates analyzing patterns of over-policing in specific areas. Even those who are simply curious can benefit, as crime data impacts everything from property values to public trust. Understanding local crime trends helps paint a clearer picture of the community and its challenges. Additionally, businesses, policymakers, and journalists may use this information to inform decisions, whether it's a shop owner evaluating risks for their customers, or a policymaker crafting better public safety strategies.

In our project, we're looking at crime rates in Montgomery County by zip code, which can help us explore DEI-related issues. But we also realize this approach has limitations and risks. Zip codes can cover a wide mix of neighborhoods , some wealthy, some lower income , with very different racial, economic, and historical backgrounds. That means lumping them together could hide important disparities in how crimes are reported or how police respond in certain areas.

For example, one zip code might include both high-income and under-resourced neighborhoods, and if we're not careful, the data could blur those lines and miss patterns of inequality. To really address DEI, the data should also include demographic info like income levels, race, and housing policy history. That would allow us to dig deeper into structural issues, like redlining or unequal policing. Without that context, there's a risk that the design ends up reinforcing bias by connecting crime to geography instead of looking at the bigger structural forces at play.

Lessons Learned:

Some important things that we learned from doing this project was that designing a database to analyze crime report data presents several complex challenges, starting with data consistency and standardization. Since crime data often comes from multiple sources – local police departments, federal agencies, and third-party reporting platforms – each with its own format, terminology, and categorization systems, it becomes very easy to develop inconsistencies within our database. We realized that this can make it difficult to aggregate and compare data accurately. We also realized that we had to create a normalization process that maps varying formats into a unified schema while preserving the original context and detail.

Another challenge that we faced was bias and representation issues. Since crime data is often shaped by systemic factors such as over-policing in certain communities or underreporting

in others, we realized that our database can also be used for analysis or decision-making like in predictive policing or resource allocation.

Potential Future Work:

Looking ahead, there are several ways this database could be expanded to provide deeper insights and greater utility. If we had more time, one possibility was to incorporate demographic data (such as income, race, or age by zip code) to analyze how socioeconomic factors correlate with crime patterns. This would allow for more intersectional analysis and could highlight whether certain communities are disproportionately affected or over-policed. We could also include time-based breakdowns (such as hour, day of the week, or season) to identify temporal crime trends. Adding tables for suspect or arrest information, where available, could open up additional research into clearance rates and law enforcement outcomes. However, we will have to be cautious not to upload all of the private information of the suspects, due to the fact that they may not always be guilty of the crime they're charged with. This would also improve our database' reputation because other companies and media outlets would be aware that our database doesn't violate the privacy rights of the people (criminal or not).

Another potential improvement would be to connect this database with 911 call data or court outcomes, offering a more complete picture of how reported incidents move through the criminal justice system. From a technical standpoint, we could build a front-end dashboard using a web framework or visualization tool (like Tableau or Power BI) to allow users, especially community organizations or policy makers, to interact with the data visually. These enhancements would make the system more robust, user-friendly, and valuable for real-world analysis and decision-making.

Another key improvement would be to implement auditing mechanisms. This is because our database wields significant power, often influencing background checks, bail decisions, and employment opportunities. Therefore, regular third-party reviews would definitely help ensure that records are accurate, timely, and free of systemic biases. Making summaries of these audits public would enhance transparency while holding agencies accountable for errors or misuse.

Citation:

Montgomery County, M. (2025, March 10). Crime: Open data portal. Crime | Open Data Portal. https://data.montgomerycountymd.gov/Public-Safety/Crime/icn6-v9z3/about_data