Data visualisation: Milestone 1
Dataset Selection and EDA

1. Introduction

New York, or *the city that never sleeps*, is one of the most crowded and lively cities in the world. Behind the bright lights, busy streets, and non-stop energy, there's also a less glamorous side: crime. Like any big city, New York has its share of problems—like theft, assault, and other types of crime.

Over time, crime in New York has changed. It's been affected by many things, like living conditions, money, neighborhoods, and local laws. Understanding how crime changes helps us understand the city itself—its challenges and how people try to make it safer.

In this project, we use data from the NYPD (New York Police Department), which includes many years of crime reports. This data tells us when and where crimes happened, and what kind of crimes they were. It gives us a chance to look at how safe the city really is.

By studying how crime changes over time, which areas are more affected, and who the victims are, we hope to better understand what's happening in the city. Our goal is to give useful information to people who want to visit or live in New York.

.

.2. Dataset:

Dataset link:
https://www.kaggle.com/datasets/aniket0712/nypd-complaint-data-historic?resource=download

The dataset selected for this project offers a rich and detailed overview of crime complaints reported in New York City over multiple years. It provides an extensive record of incidents, enabling both temporal and spatial analyses of criminal activity within the city. Below is a breakdown of the dataset's key characteristics:

- Source: The dataset is sourced from NYC Open Data, under the title *NYPD Complaint Data Historic*, and is publicly available for analysis and educational purposes.

- Scope: This dataset contains millions of individual crime complaints reported to the NYPD. Each row corresponds to a single criminal incident and includes information about when and where it happened, the type of crime, the status of the investigation (completed or attempted), and basic demographics of the people involved.

- Temporal Coverage: The data covers a broad time span, with some records going back as far as the early 2000s. It includes the exact date and time when the crime occurred (`CMPLNT_FR_DT`, `CMPLNT_FR_TM`), as well as the date it was reported (`RPT_DT`), making it possible to analyze both long-term trends and short-term patterns.

- Geographic Coverage: The dataset covers the entirety of New York City, including all five boroughs: Manhattan, Brooklyn, Queens, The Bronx, and Staten Island (`BORO_NM`). Each crime is associated with police precincts (`ADDR_PCT_CD`), patrol zones, and geographic coordinates (`Latitude`, `Longitude`), allowing for detailed spatial visualizations and hotspot detection.

- Key Variables:

    - Crime Type: `OFNS_DESC`, `LAW_CAT_CD` (e.g., felony, misdemeanor, violation)
    - Location Details: `PREM_TYP_DESC` (location type), `LOC_OF_OCCUR_DESC` (inside/outside), `X_COORD_CD`, `Y_COORD_CD`
    - Time Details: `CMPLNT_FR_DT`, `CMPLNT_FR_TM`, `RPT_DT`
    - Demographics: Victim and suspect age, sex, and race (`VIC_AGE_GROUP`, `VIC_SEX`, `VIC_RACE`, `SUSP_AGE_GROUP`, `SUSP_SEX`, `SUSP_RACE`)
- Granularity: The dataset is highly granular, offering data at the level of individual complaints. Each record is timestamped and geographically referenced, enabling in-depth analysis across both time and space.

- Data Quality: While the dataset is large and rich, it may contain missing or inconsistent values in some fields, especially those involving suspect information or precise locations. However, the volume and structure of the data allow for robust analysis once appropriate cleaning and preprocessing are applied.

In summary, this dataset provides a powerful foundation for studying crime in New York City. It enables us to explore spatial patterns, temporal trends, and demographic factors, and is well-suited for building visual tools to support public understanding of urban safety.

3. Problematic

Crime in large cities like New York is a complex and sensitive topic. Over the years, a significant amount of data has been collected by local authorities, but the sheer volume and variety of this information can make it difficult to understand the actual safety of the city. This is especially true for people who do not live in New York and want clear, accessible insights before visiting.

In this project, we aim to provide a visual and data-driven overview of crime in New York City, focusing on helping visitors and the general public understand where and when crimes happen, and who is affected by them. Our main goals are to show:

- which neighborhoods are safer or more at risk
- where crimes are most concentrated geographically
- how crime levels have changed over time
- who the victims are, based on age, sex, and race

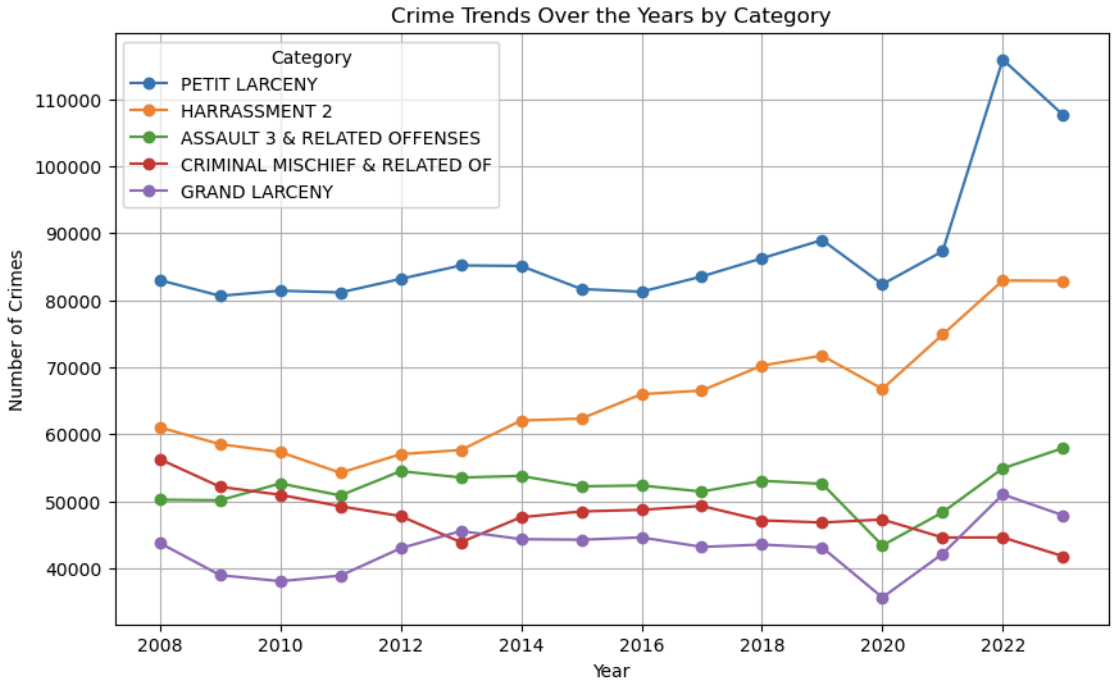
We do not aim to predict future crime or explain the causes of criminal behavior. Instead, our goal is to use the available historical data to build a clear and informative picture of New York's safety, so that people can make informed decisions when planning a trip or moving to the city.

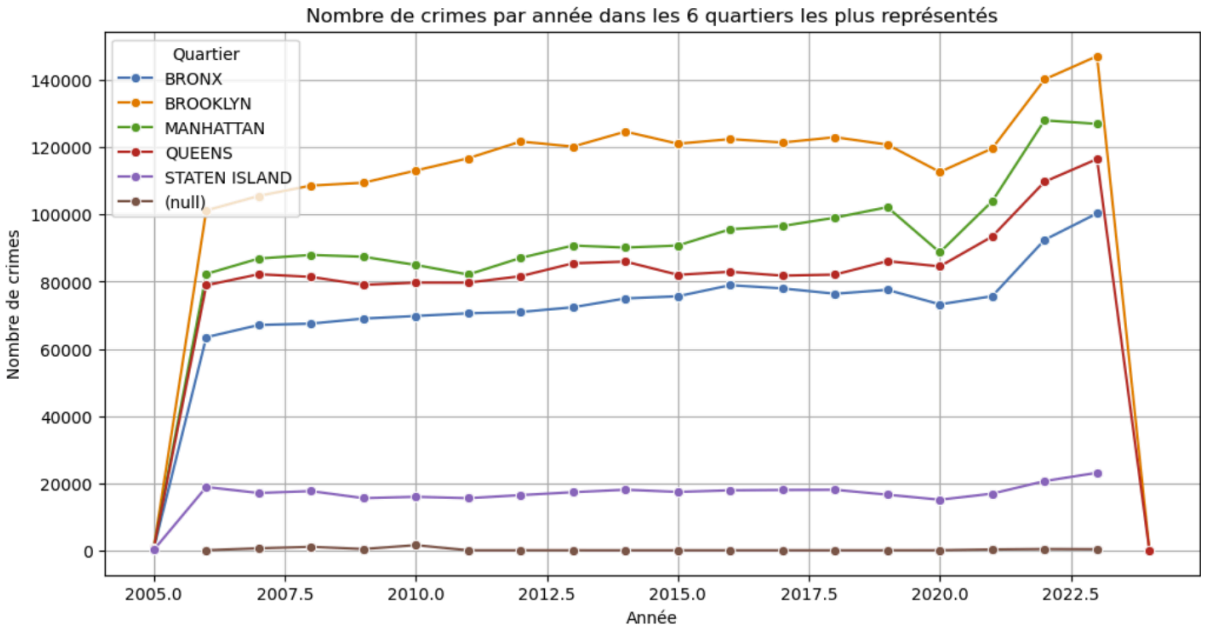
4. Exploratory Data Analysis

In this section, we analyze the crime trends in New York City based on our dataset. By examining crime variations over time and by hour of the day, we can identify patterns that help us understand when different types of crimes are most frequent.

4.1 Crime Trends Over Time

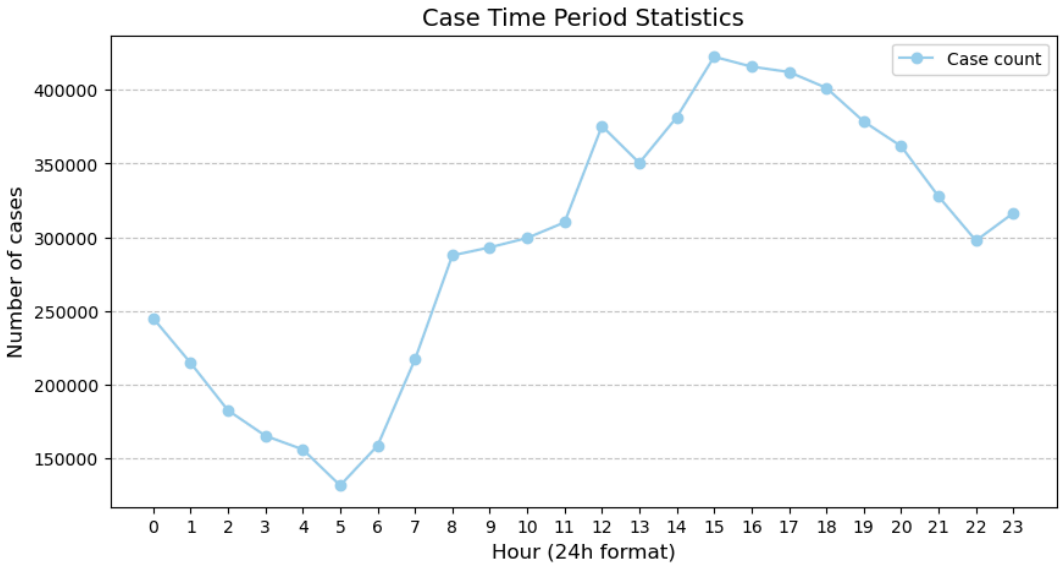
To assess how crime has evolved over the years, we analyzed the number of reported incidents for different offense categories. As shown in Figure 1, petit larceny (minor theft) has consistently been the most reported crime, with a sharp increase after 2020. This rise could be linked to post-pandemic economic difficulties or changes in law enforcement reporting. Other offenses, such as harassment, assault, and grand larceny, have remained relatively stable over time, with slight fluctuations.





4.2 Crime Occurrence by Hour of the Day

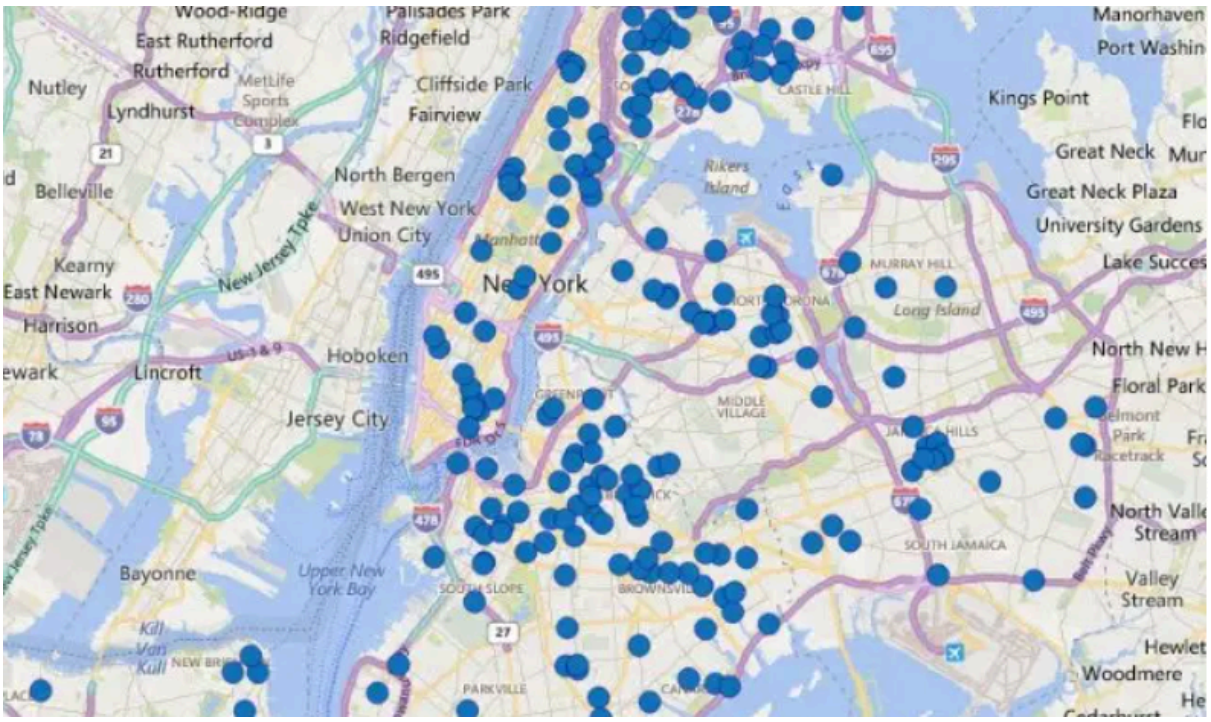
We also examined the distribution of crimes based on the time of day. As illustrated in Figure 2, crime rates are lowest between 3 AM and 6 AM, when the city is at its quietest. However, there is a sharp increase starting at 7 AM, peaking between 2 PM and 4 PM, which may correspond to increased public activity during work and school hours. Crime rates remain high throughout the evening and start declining after 9 PM.



4.3 Map of the crimes

Understanding where crimes occur is essential for evaluating urban safety. Using the latitude and longitude coordinates from the dataset, we plotted a crime heatmap to visualize crime of felony assaults density in the last week of february across New York City. As shown in Figure 3, crime is highly concentrated in Manhattan, Brooklyn, and the Bronx, with clusters of incidents also appearing in parts of Queens.

The high crime density in Manhattan can be attributed to its role as the city's commercial and entertainment hub, attracting large crowds daily. Similarly, Brooklyn and the Bronx, being densely populated boroughs, also experience significant crime activity. The outer regions, such as Staten Island, appear to have fewer reported crimes, indicating a lower crime density.



5. Related work

Urban crime and safety have been widely studied in both academic research and public policy. Over the years, many institutions and independent developers have created tools and visualizations to help make crime data more accessible and understandable. New York City, in particular, has been the focus of numerous crime mapping and analysis projects due to the availability of large-scale open data from the NYPD.

Several projects have been developed using crime data from New York City since 2006. One of the most well-known is CompStat, a system created by the NYPD to track and analyze crime trends, helping officers allocate resources more effectively. Another tool, Patternizr, uses machine learning to detect patterns in crimes, making it easier to connect related incidents. The Domain Awareness System (DAS), developed with Microsoft, integrates real-time data from cameras, 911 calls, and crime reports to improve law enforcement responses. Additionally, the NYPD launched a public crime statistics portal, allowing citizens to explore crime trends in their neighborhoods. These tools have helped make crime data more accessible and useful for both law enforcement and the public.