Project Proposal

Overview:

Title: High Crime Rate in New York City — Is Living in NYC Really Dangerous?

Objective: The goal of this project is to identify clear patterns in which boroughs and neighborhoods have higher crime rates and what types of crimes are most common. The project will analyze historical and statistical patterns of crimes and provide visual image.

Importance: I value safety over anything else. As a person who lives in New York City and may want to buy a home here later in my life, I want to use real statistical data to decide which neighborhoods are safer and which should be avoided.

Originality: While people often hear that "New York is dangerous," most don't know which areas are truly "dangerous" or what kinds of crimes dominate in which areas such as theft or assault. I want to show factual data and trends so people can make decisions based on real evidence instead of fear or rumors.

Background Research:

Key Term Definitions:

Crime Rate: The number of crimes happening within a given area per a specific population or per a specific time period.

Borough: New York City consists of five main divisions, including Bronx, Brooklyn, Manhattan, Queens, and Staten Island.

Choropleth Map: This term will be used to describe the higher or lower crime rate areas by showing the density of color on the map.

Existing Solutions:

New York government publishes crime rate in New York City, but they only provide raw data of the number of crimes instead of forecasting models with interactive maps, which shows the details of crime types. My project will provide the detailed information and predicting models with data visualizations.

Data:

Data Source:

The data will be from NYC Open Data.

1. NYPD Complaint Data from 2006 to 2019 (yearly updated)

   https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i/about_data

2. NYPD Complaint Data of current year, 2025 (quarterly updated)

https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/about_data

Data Volume:

Data 1 includes about 7 million rows and 36 columns; Data 2 includes around 300 thousand rows and 36 columns. They covers all valid felony, misdemeanor, and violation crimes. The data include almost 20 years of citywide records, which is large enough data for statistical modeling and time-series forecasting.

Data Richness:

My data contains time information(e.g., what time the accidents happened), location information(e.g., borough and specific details such as NYC park, playground, grocery store, residence, street etc.), crime description, level of offense, age, gender, race etc.

The Predictive Model:

Predictors (X):

- Month and year
- Borough(area)
- Crime type such as assault, robbery, larceny etc.
- Law category (felony, misdemeanor, violation)
- Season or weather(temperature may affect crime rates)

Python Dependencies: pandas, numpy, matplotlib, seaborn, scikit-learn, folium etc

Security and Privacy Considerations: NYPD data are de-identified, so there is no personal details included in the data.

The Visualization:

Summary Statistics Plots:

- Bar chart showing total crime counts by borough
- Line plot showing monthly crime trend from 2006 to 2025
- Histogram showing crime type distribution

Map Graphs:

- Choropleth map using folium showing crime intensity by borough with color density

Model Performance Plots:

I will create a line plot that compares the predicted crime rates from my model with the actual observed crime rates for each borough. This visualization will help demonstrate how closely the model's predictions follow real trends over time and show whether the forecasting method is accurate.