Boston Crime Project Report – DS2500 Professor Rushit Sanghrajka

Names: Ruyao (Anthony) Tian, Kaiyang Weng, Le Fan (Ethan) Fang

Problem Statement and Background:

Nobody wants their wallet stolen, or their body attacked, or their house set aflame; nobody likes crime, not even criminals if they thought it through. Though law enforcement does a great deal to minimize these occurrences, crime evidently still exists. One way to prevent a problem is to better anticipate it. Thus, knowing *what type of crimes tend to happen when and where*, would bolster society's ability to minimize it.

This project addresses the problem of crime in Boston, by analyzing publicly available crime data collected by the Boston Police Department. It covers three main areas of analysis: seasonality, geography, and daily temporality; along with a prediction model that encapsulates all three. The aim is to unveil insights about crime that can help law enforcement and lawmakers better prevent crime by getting ahead of the problem.

Introduction to the Data:

The dataset used for this project is the Crime Incident Reports (2023), sourced from the Boston Police Department through the "Analyze Boston" platform (data.boston.gov). Comprised of 146,150 rows, it documents unique crimes, their locations, and the time of police response. This mass of data has been collected by police officers, who note the initial details of every call.

Privacy and Ethical Concerns:

One privacy concern is that individuals may be identifiable through location data. Though there is only latitude and longitude, and no other personal info is present, simply having

location may be enough to deduce someone's identity; especially for crimes like "Investigate Property", where the location would be someone's home address.

Biaswise, the dataset discloses excluding crime not aligned with Massachusetts General Law (MGL Ch. 41, Section 98f). Furthermore, a crime that was never caught, is unfortunately— but obviously— not in the dataset. Both factors undermine the datasets' full representability.

Ethically, it is the case that certain crimes are higher in certain communities amongst certain groups. Data ought to be interpreted cautiously to avoid reinforcing stereotypes.

Data Science Approaches

1. Exploratory Data Analysis (EDA): Identified patterns, detected outliers, and visualized data distributions— in Seasonality, Geography, and Daily Temporality.

2. Data Filtering: Focused our analysis— in Seasonality, Geography, and Daily Temporality.

3. Geospatial Analysis with GeoPandas: Mapped crime hotspots to identify high-risk areas and visualize geographic trends— in Geography.

4. Predictive Modeling with K-Nearest Neighbors (KNN): Built a predictive model to estimate crime frequency based on month, hour, and location (Lat & Long).

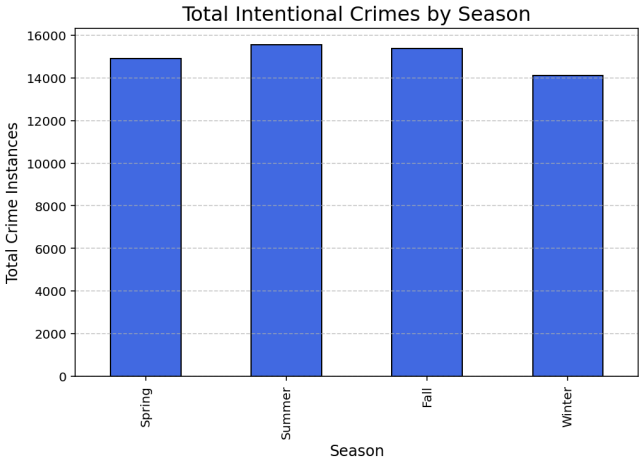
Results and Conclusions:

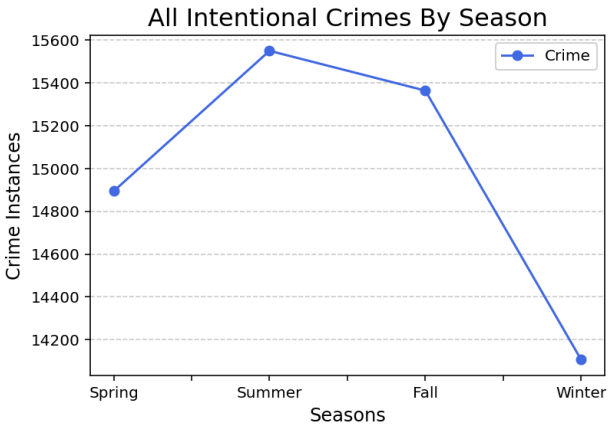
Seasonality:

To analyze seasonality, we first processed the data by filtering to the crime types and time-related columns within 2023. Then, our definition of *intentional crime*, crimes committed with malice, was used to further narrow our scope. This was done in interest of looking at the most targettable crimes; as crimes like "Larceny – Shoplifting"— clearly intentional— reoccur with more rhyme or reason, versus perhaps something like "Moving Vehicle – Accident"— unintentional. After this filter, four seasonal data frames were created, and the sum of each crime

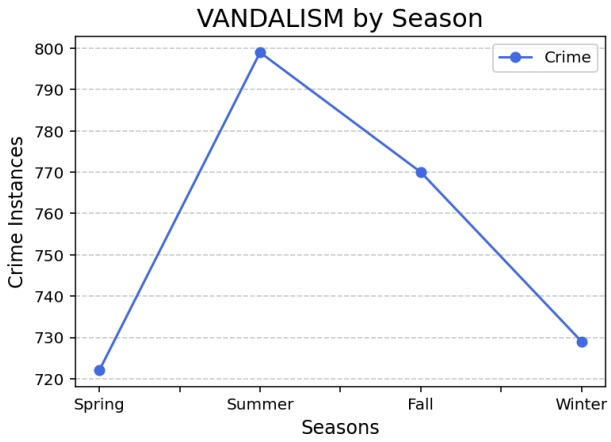
was calculated for each. The seasons were defined as: Spring, Summer, Fall, Winter; containing the months: [3, 4, 5], [6, 7, 8], [9, 10, 11], [12, 1, 2]— respectively.

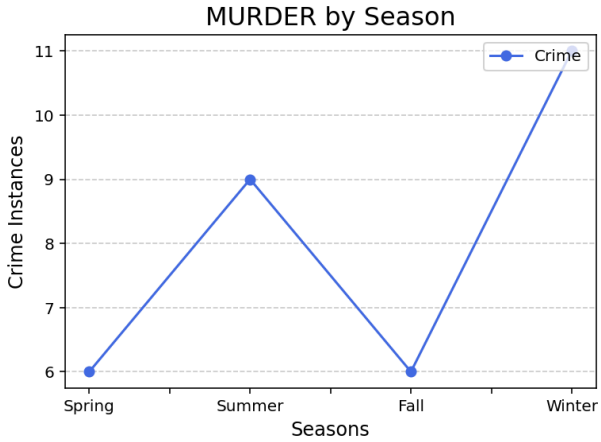
Visualizing intentional crimes by season in a bar graph showed that crime in the summer was noticeably higher than in the winter. This disparity was much easier to visualize after converting to line plot, with summer crimes, as a whole, reigned 1000 instances over winter.

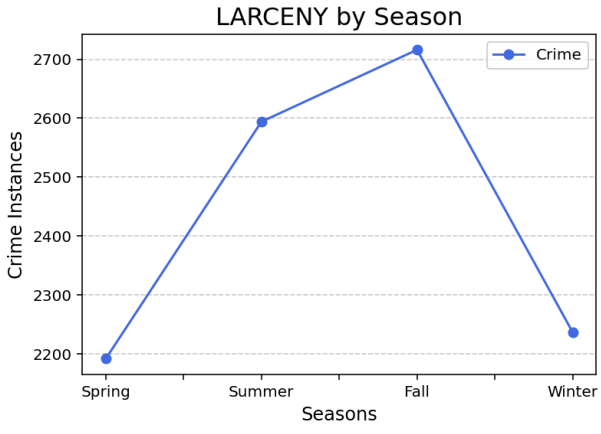


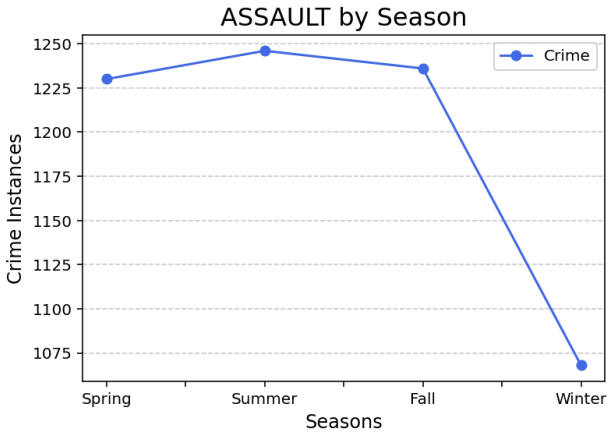


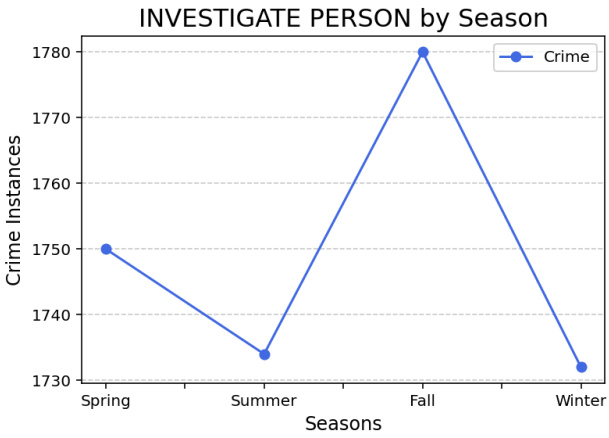
To find out how generalizable this pattern was, notable specific crimes were then visualized:

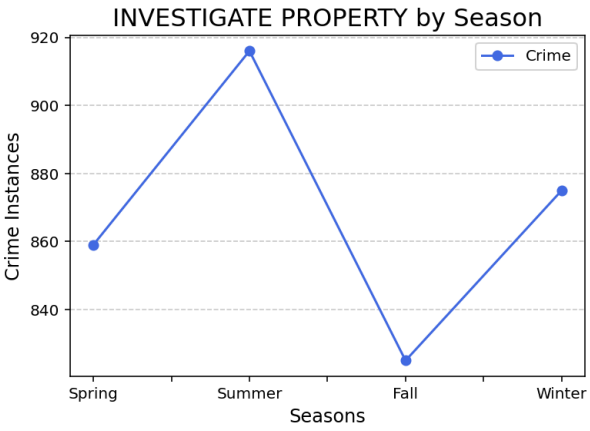






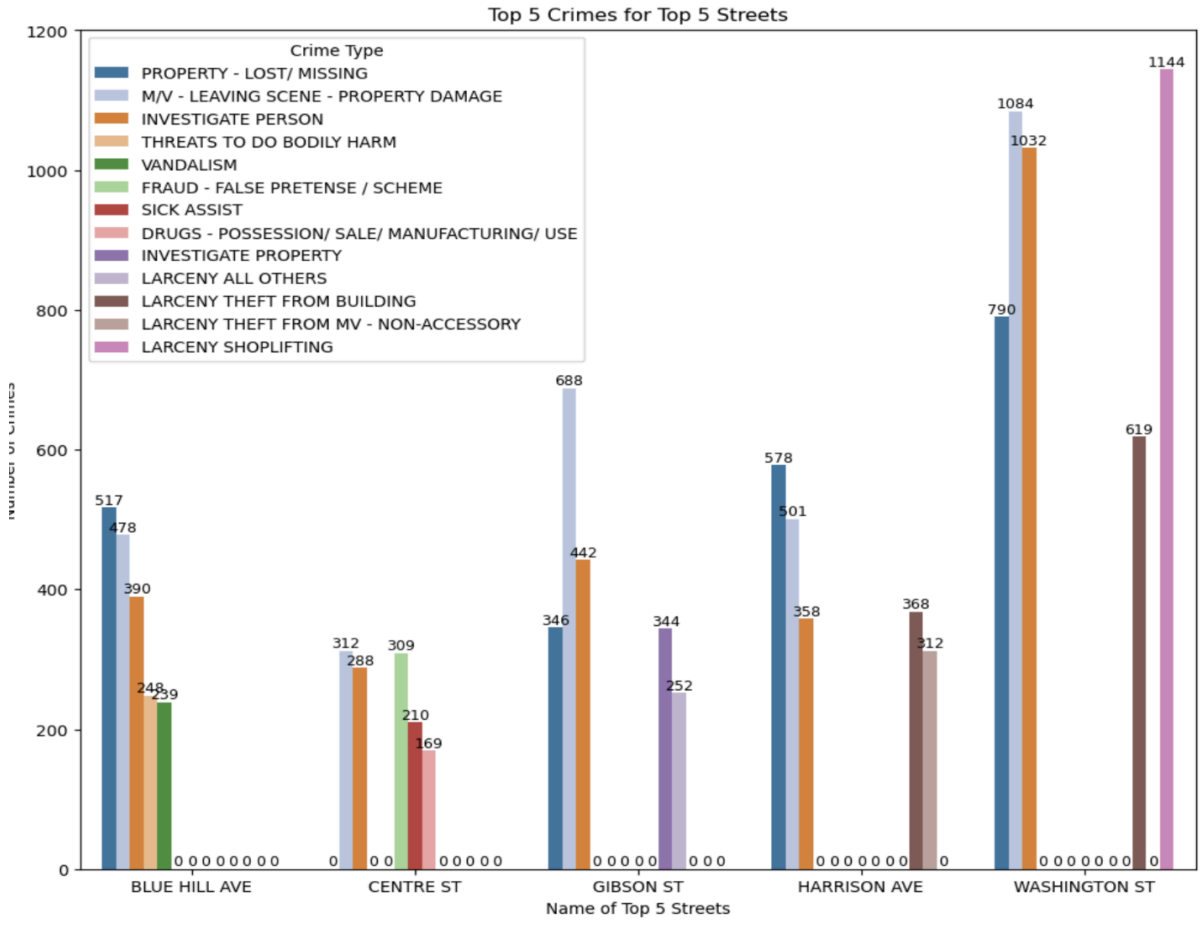






Even in just six crimes, picked for being most frequent or most severe and condensed for brevity, different annual distributions apply to different crimes. For example, while "Assault" is higher in the summer than winter, "Murder" sees the opposite pattern. Thus, when drawing actionable insights, it would be more useful to look at specific crimes.

Geography:

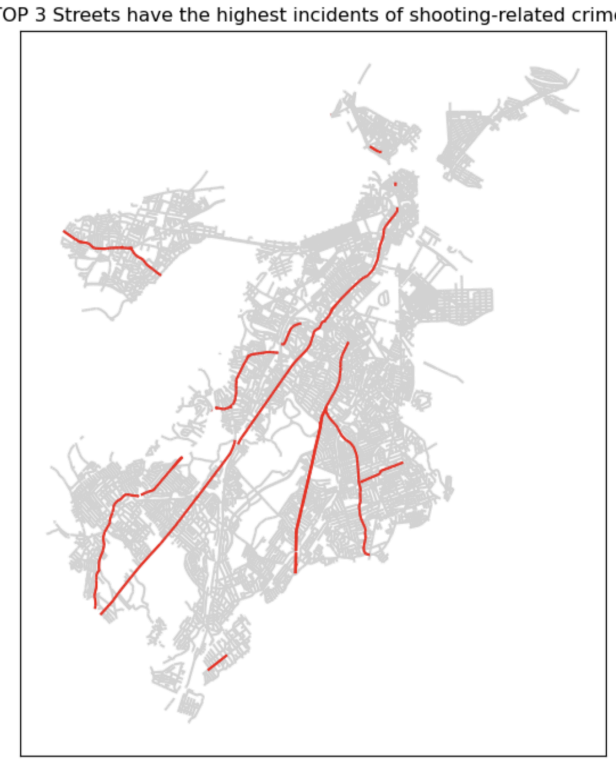


To analyze geography, two target features— streets and districts— were first grouped together, streets and districts. For each street and district, the total crime counts were calculated, then ranked in descending order of frequency.

The top five streets and districts with the highest crime are graphed above, along with the crimes that most often occur there. Notably, "Property – Lost/Missing", "Property Damage", and "Larceny Shoplifting" are common issues.

One commonality between these streets is their high concentrations of property-related crimes. Another is that they all happen to be near commercial areas, which may be related. Because commercial areas have businesses offering goods, theft crimes would occur there.

Afterwards, we wanted to look into shooting incidents specifically. Applying the same process above, the five streets, with the most shooting incidents, were retrieved. This is visualized in a bar graph and geospatial map below.

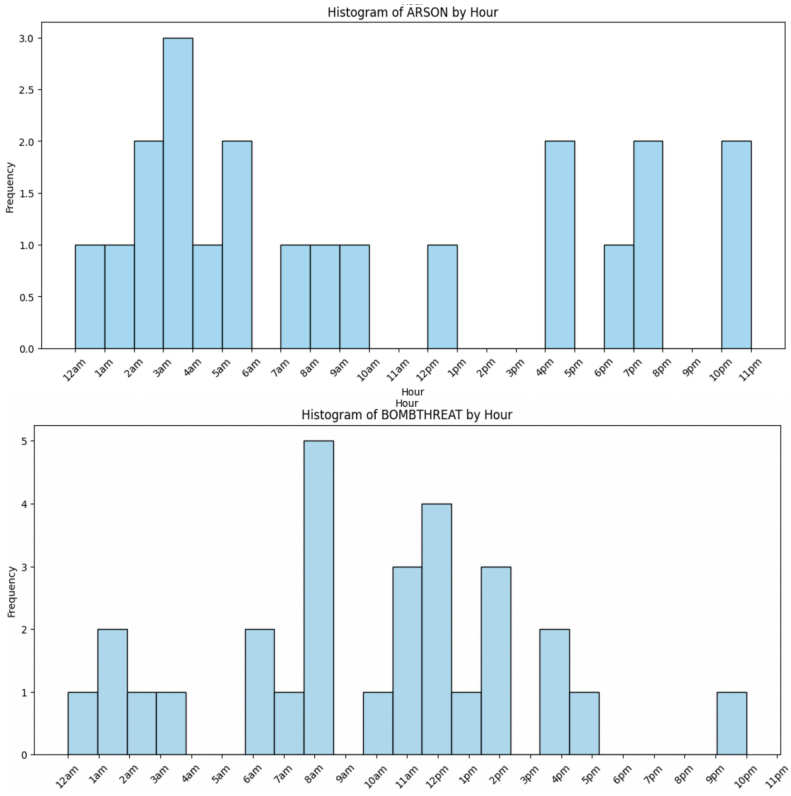


Knowing these streets could potentially help law enforcement minimize crime there. Street patrols and security cameras devoted in these areas can be increased. Policymakers can also use these streets, along with the specific crimes that are more common in them, when writing laws.

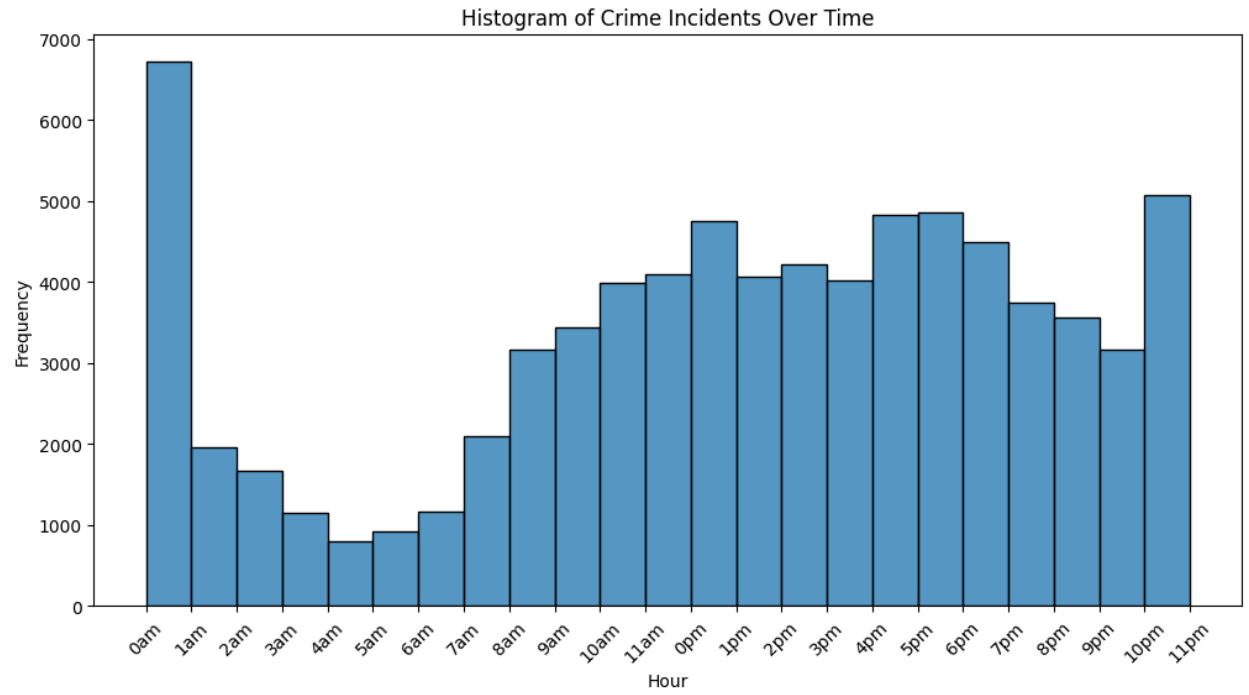
Daily Temporality:

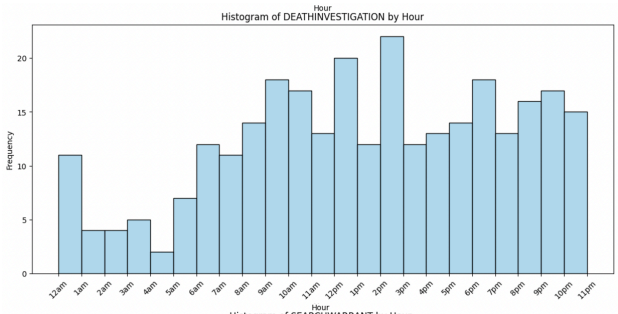
We plotted histograms of different types of crimes with the time they were committed. Same types of crime were grouped together; for example, first-degree murder and second-degree murder were simply considered as murder. This resulted in 81 crimes, subsequently visualized in 81 histograms. The below are not exhaustive, but a few representative examples.

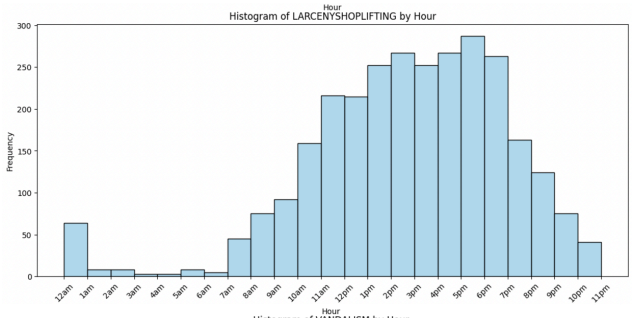
Many of the crimes had a low sample size; a fact that is relieving, as the less bomb threats there are, the better. However, low sample size also means it's hard to draw actionable insights from the data.

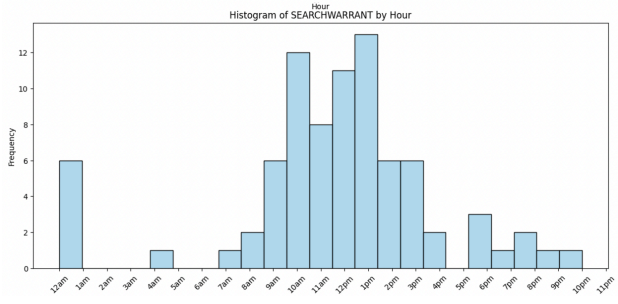


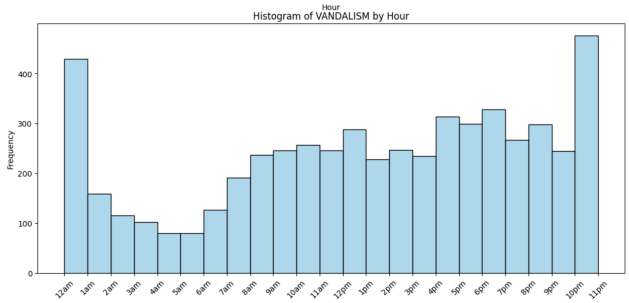
For crimes with a larger sample size though, one common trend was the frequency rising through the afternoon; except shoplifting, as shoplifting occurs only when stores are open— otherwise it'd be under burglary. Below is a general graph, along with some specific crimes that comprise it.









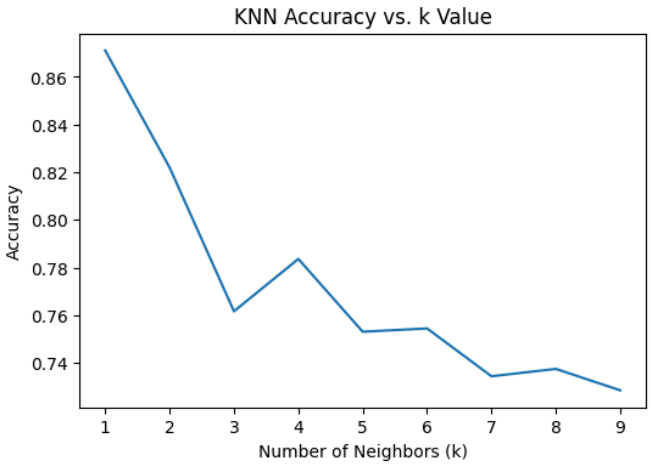


Predictive Model:

The KNN model demonstrated the feasibility of predicting crime likelihood based on location and time, providing a tool for proactive resource allocation.

We built a KNN model to estimate crime frequency based on month, hour, and location (Lat & Long). Features were normalized and values were scaled from 0-1. Geographic

coordinates were rounded to three decimals, corresponding to a 110-meter diameter area. Model Accuracy: Achieved approximately 78% accuracy using the KNN model, using 4 neighbors. We used 4 neighbors, as it avoids overfitting while also having decent accuracy.



Future Work

There are a couple next steps we'd like to take to expand our project, all centering around making our analysis' more accurate.

First, expanding our dataset to include previous disclosed years, from 2015-2022, would give us sevenfold the scope to analyze. The same functions would be reused to run the same processes over each old dataset. Histograms, maps, and our prediction model would then encapsulate crime throughout almost the past decade, rather than just one year; which may yield more reliable insights. This also opens opportunities to analyze crime throughout time periods. Immediately, comparing pre-pandemic, pandemic, and post-pandemic periods, is an alluring potential path.

Secondly, holidays and public events are special days that conceivably experience different crime rates. Take Black Friday for example, though not an official holiday, is still a unique day when shoplifting and theft may be higher. Analyzing special days, then comparing with our aggregated results, might be useful. Often, the most telling insights hide in specificity, and this is certainly a very specific realm containing tangible benefits.