Project Phase 1 - Data Intensive Computing

Students: Francesco Tinessa (50628509), Abhinna Manandhar (50626441), Abigail Udeobi (50446931)

Spring 2025

Crime Hotspots and Police Deployment in NYC: Examining Arrest Patterns

Background of NYPD Arrest Dataset (Link: NYPD Arrest Data): For our chosen dataset, we have decided to go with a New York Police Department Dataset. This dataset of over 260,000 lines documents the crimes committed throughout different precincts in New York City every month in the year 2024. The data includes the type of crime, arrest precinct, the perpetrator's race, and gender, along with the coordinates of where the crimes occurred. This dataset was released for the purpose of studying the nature of police enforcement activity.

Background of Problem and Problem Statement Important aspects of urban safety and policing throughout New York City include police activity and crime rates. As urban crime is often reoccurring throughout New York City it has become a significant problem since it heavily threatens public safety and police effectiveness. Crime rates have continued to increase over the years whether it be from physical violence in New York City subways or mass shootings occurring in big events on huge holidays. It has affected the overall sense of security of both local residents and visitors and has increased public safety concerns. As we have this dataset to perform a comprehensive analysis of arrest trends and the distribution of crime we aim to address the problems of the impact of demographic variables (such as gender, age, and location) influencing crime/arrest rates, if crime hotspots can possibly be ascertained from past arrest data, and the ways these findings can influence the adjustment of resource distribution by law enforcement.

Project Contribution to the Problem This project has the potential to increase public safety in New York City overall and allow police officers to allocate their resources more strategically. Police departments would be able to anticipate crime scenes by being more aware of arrest trends. Also, as New York City is known to be a famous tourist attraction with Times Square, stadiums, and parks being some of the most popular, this information from our project may be able to help the tourist industry by enabling travelers to make better-informed decisions about the safest places to visit and when.

Data Cleaning/Processing

1. Within the Offense Description column, we have filtered out the rows containing '9' for the reason of it not being part of a proper format.
2. We created a new column to combine "Latitude" and "Longitude for the reason of making the dataset more concise and less convoluted.

3. We consolidated the "Offense Detailed Description" column so we could map specific offenses to broader crime categories.

4. For the outlier offenses that don't fall under the broader crime categories, we assigned them "OTHER".

5. We removed rows with "UNKNOWN" in the "Perpetrator Race" column since they're invalid.

6. We consolidated the "Perpetrator Race" column from "Perpetrator Race" to specifically list the different races.

7. We removed the "Geo-referenced Location" column for the reason of inconsistent data entries.

8. We renamed column's abbreviated names for better clarity and to avoid confusion when examining.

9. We created a new column called "Arrest Day of Week" in order to analyze the crime on weekdays.

10. We removed null values from the "Offense Category Code" column for the reason of it being invalid/noisy data.

11. We removed the X and Y coordinates column for the sake of reducing noisy data and them missing some date as well.

EDA Operations

1. We plotted a vertical bar graph in order to visually show how many arrests occurred in different New York City boroughs. By seeing this graph, we are able to identify the NYC boroughs that have the highest and lowest amounts of arrests which can help law enforcement identify crime hot spots and allocate their resources in a much more efficient way.

2. We plotted a vertical bar graph in order to visually show how many arrests occurred in each offense category (M for misdemeanor, F for felony, V for violation, etc). By seeing this graph, we are able to identify the most frequent types of offenses committed, giving an overview of the occurrence of different crime categories in New York City.

3. We plotted a line graph in order to show the arrest trends for each month in the year. Based on this graph we visually see the fluctuations of the number of arrests that happen each month, with August and October being the months with the highest arrests and December being the month with the lowest.

4. The vertical bar graph is portrayed to display the variation in the number of arrests on different days of the week. The x and y axis represent the days of the week and the number of arrests respectively. This helps us to understand the correlation between the week day and the risk of any crime occurring.

5. Another vertical graph with the gender on the x-axis and the count of arrests on the y-axis visualizes the distribution of the genders of the arrested perpetrators. Based on this visualization we understand that there is a significant difference in the number of male and

female perpetrators. While there are over 200,000 male perpetrators arrested there are only about 50,000 females.

6. The bar-chart visualization with age groups in the x-axis and the number of arrests helps us to understand the distribution of the age groups of the perpetrators. From this we understand that most perpetrators that were arrested were between the age of 25 and 44.

7. The x-axis of the bar graph resembles different racial groups in the dataset and the y-axis represents the number of arrested perpetrators. This visualization provides us an insight of the number of arrested perpetrators based on their race. Additionally, this also allows us to understand more about any potential racial disparity in policing which can help aid different law enforcement practices.

8. The bar graph allows us to understand the volume of arrested perpetrators in various precincts. The x-axis labels all the available precincts in the database and the y-axis resembles the number of arrested perpetrators.

9. We plotted a scatter plot to visually show the geographical distribution of arrests all throughout New York City. This descriptive scatter plot can aid law enforcement in identifying locations of frequent criminal activity for crime predictions.

10. We plotted a line graph to visually show the monthly change in offense categories in New York City. Each of the five lines in the graph represents different offense categories such as unclassified. felony, infraction, misdemeanor, and violation. Each plot on the graph represents the number of offenses committed by each category throughout different months in the year and this visualization can aid us in understanding the fluctuations of each crime according to seasonal trends.

11. We plotted a line graph to visually show the monthly changes in different offenses in New York, similar to the line graph above showing the monthly change in offense categories. Each of the lines represent the offenses that were shown to repeat in the NYPD arrest data set: assault, drugs, theft, other. The purpose of this graph serves a similar one to the above graph, both being able to help aid the understanding of the fluctuations of each crime according to seasonal trends.

12. We plotted a vertical bar graph containing four variables this time, perpetrator sex, race, the number of offenses, and the offense type. This graph visually shows the categorization of each sex from each race when it comes to which offenses were committed. This would help law enforcement be better informed on the repeated criminal behaviors in each demographic.

13. We plotted a line graph visually showing the monthly change in the number of arrests of each gender throughout the different months of the year. This graph serves a similar purpose to the graph of arrest trends for each month except this shows the difference between the arrests of each gender, aiding law enforcement in their information database.

Resources Used

(a) https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

(b) https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html

(c) https://numpy.org/doc/2.2/user/index.html#user

(d) https://seaborn.pydata.org/

(e) cleaning and eda.ipynb (provided in class)