Mapping & Predicting Arrest Patterns within New York City

ISC 4941 – Data Science Capstone

Carolina Forero, Daniel Gutierrez, Kristen Kane, Natalia Rigol

Florida Atlantic University

Abstract:

The purpose of this project is to identify and predict arrest patterns within the precincts of New York City. More specifically, we are using NYPD arrest data to identify arrest patterns and examine how offenses may differ throughout each day of the week. These results may be utilized by law enforcement to allow for better use of resources in areas that need it the most. Communities may also have access to this information to enhance safety measures within neighborhoods that may have a higher influx of certain offenses. Our group has chosen to specifically avoid the use of any personal demographic factors within this analysis to prevent the possibility of factors, such as race or gender, being used as biased drivers when an arrest is being made. This study investigates whether spatial and temporal features can predict crime patterns in New York City using machine learning approaches. Analyzing 142,797 NYPD arrest records, we employed three methodologies: a Random Forest Classifier to predict offense types, a Random Forest Regressor to forecast arrest volumes, and a Conditional Probability Model to assess the probability of each top offense to occur within each precinct for every day of the week.

Introduction:

Areas with a very large population like New York often struggle with properly reducing crime while also trying to manage resource usage and budgeting. Crime tends to appear in certain clusters, making trends much more identifiable for proactive intervention methods. Currently, police rely on more reactive methods of intervention than predictive methods. The reason for this is that current predictive models favor more biased identification methods based on personal demographics, deeming them unreliable. Our project focuses on less personal demographics, such as location and time of arrest, to provide a more transparent approach to identifying trends in arrests and offense types. Properly addressing this issue will not only benefit law enforcement, but it also benefits emergency services, community organizations, and policymakers that are aiming to reduce unnecessary resource usage and increase public safety. Our group is attempting to address the gap in current predictive models and provide stakeholders with a model that is easily accessible and unbiased by utilizing geospatial data and temporal arrest patterns.

Methods:

Exploratory Data Analysis:

To get a better understanding of the spatial and temporal crime patterns within New York City, we created a series of geospatial maps to provide visualizations of arrests. First, we created a hotspot map that grouped the arrests by precinct and identified the most prominent offenses within each precinct. Then, we further broke down the arrests by day to produce a visualization with seven different hotspot maps for each day of the week, also showing the most prominent offenses within each precinct. Finally, we created a heatmap to visualize the total arrest counts for each day of the week, as well as each month within the dataset (January – June). These visualizations allowed us to get a better grasp on the arrest patterns that occur within the data.

Random Forest Classifier & Regressor:

To begin our methodology, we created month & weekday variables to extract from the date of the arrest for better analyzing and testing. For this model, we only used important features such as *ARREST_BORO*, *ARREST_PRECINCT*, *DAY_OF_WEEK*, and *MONTH* to avoid biased results. We then filtered the data by only using top 10 offenses for a better accuracy score. We also applied categorical encoding to convert the variables into numbers so the model can test faster and efficiently. We split the data 80% for training and 20% for testing.
For the random forest regressor, the target variable is the arrest count aggregating monthly totals, and the features include the month and offenses encoded. We evaluated the metrics by using sklearn.metrics to show accuracy, precision, recall, and feature importance. For the classification model, it received a 24% accuracy rate which is good when considering the fact that the arrests are very spatial, and each neighborhood has its own crime pattern. The precision got a good proportion of correct predictions per offense type. The feature importance showed that based on the arrest precinct, it can show which offense is likely to occur. For the regression model, the MAE shows the average prediction error in arrest counts, which is 67 errors per 1000 arrests, which is not bad at all. The R2 score shows a 97% performance of variance explained by the model.

Linear Models: Regression and Classification:

Linear Regression and Classification were used in this analysis because they offer clear and easy to understand patterns within the NYC Dataset. Linear Regression helps show how total arrests changed over time, allowing the model to find trends and make predictions for the future years. These predictions will show whether or not arrests are increasing or decreasing over time. As for us, we used 2021-2024 arrest data to predict total arrests for 2025 and 2026. Linear Classification is equally as valuable because it makes predictions on categories by finding patterns in the dataset. Our model uses data like location, month, and day to make predictions on which type of offense is most likely to occur. When using both these models, they can get a simple, yet powerful understanding of how the data works and the patterns within. These models also allow for easy-to-read graphs and charts which can help communicate our results.

Histogram Gradient Boosting Classification Model:

The Histogram Gradient Boosting model was created to predict offense type while utilizing only spatial and temporal features within the NYPD arrest dataset. The model utilizes characteristics such as precinct, borough, day, month, law category code, and a created weekend flag. The enhance the focus and predictability of the model, we chose to focus only on the top 5 most common offenses, considering they provide the most data entries. We implemented the HGB model through the use of scikit-learn because it is able to handle large datasets and group continuous data to create histograms before the decision trees are trained.

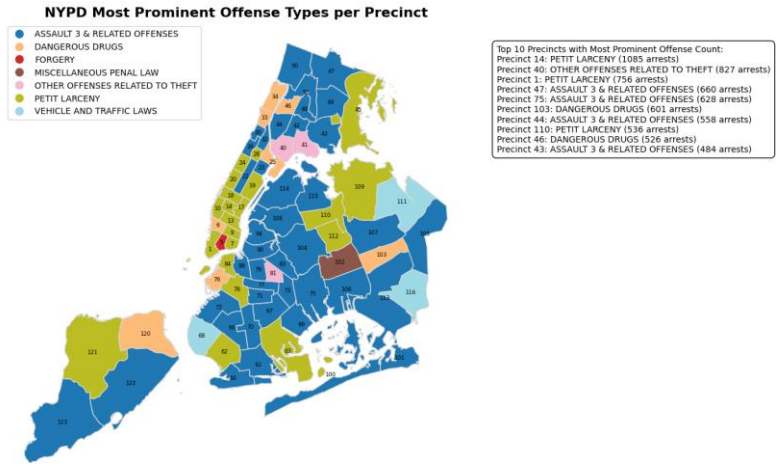
Conditional Probability Model:

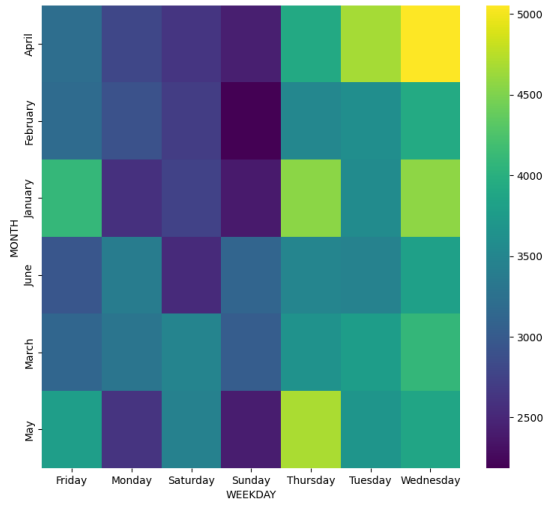
The conditional probability model helps to understand which offenses are most likely to occur in each precinct of New York City on any given day of the week. The model uses the

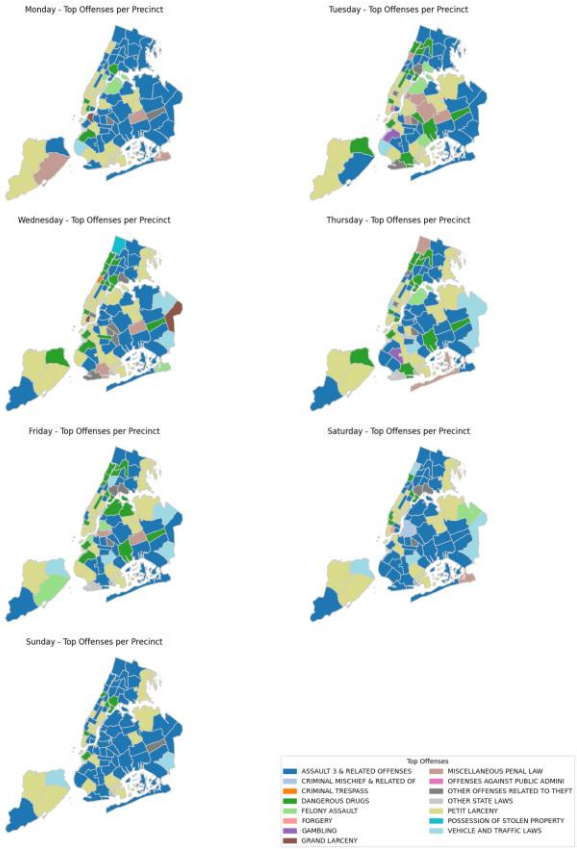
NYPD arrest dataset (Jan 2025 – June 2025) to group arrest by *ARREST_PRECINCT*, *DAY_OF_WEEK*, and *OFNS_DESC*, to calculate the probability of each offense based on the total amount of arrests made in each precinct on that day. This approach allows us to identify the crime patterns of the most prominent offense types without relying on predictive features. For example, on a Tuesday in precinct 1, there is a 40% chance that petit larceny will occur.

Results & Discussion:

EDA:







The results from both hotspot maps indicate that the most common offense types are Assault 3 & Related Offenses and Petit Larceny. When looking at the hotspot maps for each day of the week, you can see some variations with less prominent offense types. For example, Dangerous Drugs seems to be a more common offense type from Tuesday-Friday is multiple precincts. This within itself could allow law enforcement and communities to better allocate their resources each day. The heatmap indicates that Wednesdays in April have the highest arrest count. When comparing that information to the most prominent offenses hotspot map for Wednesday, you can see that the types of prominent offenses in each precinct vary greatly.

Classification Model:

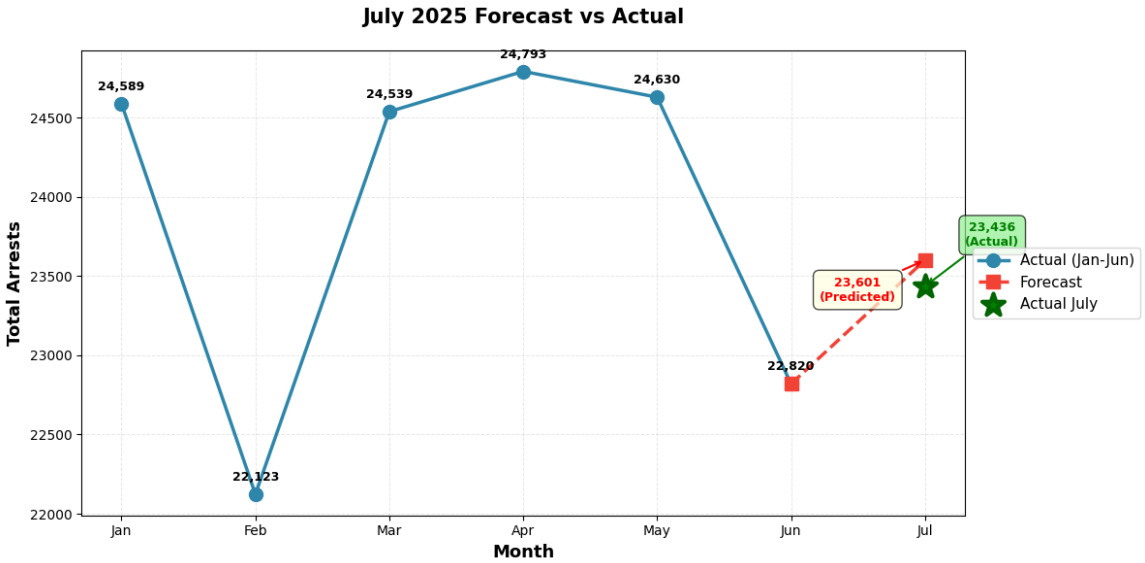
```
Original dataset: 142797 arrests
Filtered dataset: 101082 arrests
Offense types reduced from 56 to 10

Top 10 Offenses:
OFNS_DESC
ASSAULT 3 & RELATED OFFENSES       19590
PETIT LARCENY                      14461
DANGEROUS DRUGS                    11966
FELONY ASSAULT                     11479
MISCELLANEOUS PENAL LAW             9462
OTHER OFFENSES RELATED TO THEFT     8736
VEHICLE AND TRAFFIC LAWS            8325
CRIMINAL MISCHIEF & RELATED OF      6182
ROBBERY                             5519
DANGEROUS WEAPONS                   5362
Name: count, dtype: int64
Accuracy: 0.243
Number of offense types: 10

Feature Importance:
           Feature  Importance
1  ARREST_PRECINCT    0.494151
3            MONTH    0.250904
2          WEEKDAY    0.227639
0      ARREST_BORO    0.027306
```

The result of the Random Forest classification model shows better accuracy than the baseline model, which was 17%. We modified the model by utilizing the top 10 offenses to receive a better accuracy of 24%. The accuracy is still quite low but is understandable considering we only used 4 features: borough, precinct, day of week and month. This explains why the accuracy is so low, but we wanted to avoid any bias output. It also shows the number of arrests by offense from highest to lowest, showing Assault 3 & Petit larceny as the most common offenses to occur based on precinct.

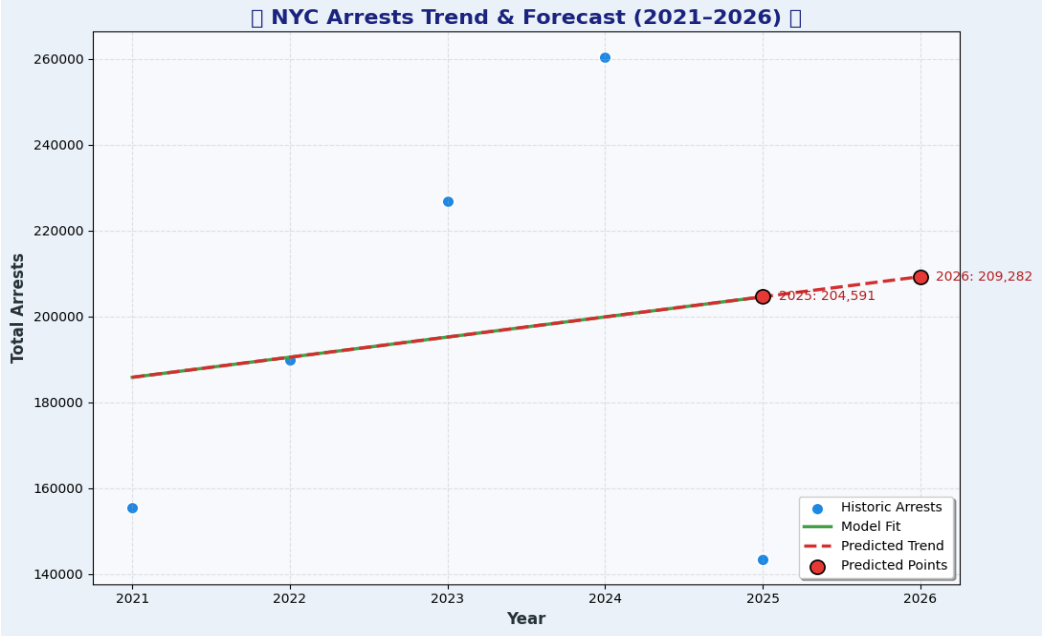
Regression Model:



The purpose of this Random Forest regression model is to predict the number of arrests that will be made in July 2025. Luckily, throughout the duration of this project, the NYPD dataset had been updated, adding the arrest data for the months of July, August, and September.

In the visual, we included the number of arrests that the model predicted, which is 23,601 and the actual number of arrests, which is 23,436. The difference can be explained by the $R^2 = 0.974$, performing with great accuracy.
This is the regression model predicting the number of arrests for the month of July. Compared to the regression model for historical crime data of NYC, this one only shows the number of arrests for a month compared to the following year of 2026, which is predicted to be higher than previous years.

Linear Regression:



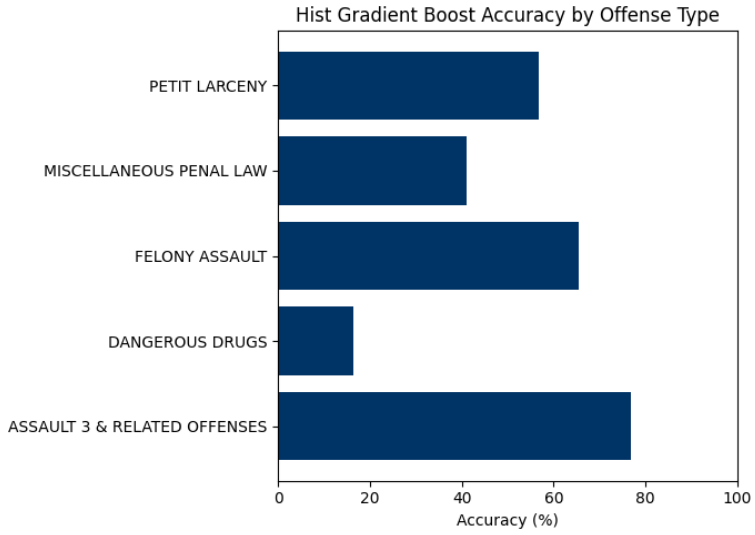
The linear regression model, which was trained on total arrest data in NYC from 2021-2024, shows an upward trend in arrests. Every year shows an increase in arrest data, with 2025 being the exception. This is because the 2025-year was not complete, yet the model is made to account for that information. To support our model, it achieved an accuracy score of roughly 80%, which is strong enough to rely on predictions. Although the R^2 value is 0.10, the prediction still aligns with the observed data. With these results, it gives us a reliable source that our model has reasonable predictions for the next two years. The Regression line (the red dotted line) smooths out the data and predicts that in 2025, there will be 204,591 arrests, which continues in an upward trend to 2026, increasing to 209,282. This shows that there might be a dip in arrests but will continue to point upwards until changes are made.

Linear Classification:

Year	Total Arrests	Top Month	Top Day	Top Borough	Top PD Description	Top Offense	Top Precinct
2021	155,404	October	Wednesday	K	ASSAULT 3	ASSAULT 3 & RELATED OFFENSES	14
2022	189,774	October	Wednesday	K	ASSAULT 3	ASSAULT 3 & RELATED OFFENSES	14
2023	226,872	May	Wednesday	K	ASSAULT 3	ASSAULT 3 & RELATED OFFENSES	14
2024	260,503	August	Wednesday	K	ASSAULT 3	ASSAULT 3 & RELATED OFFENSES	14
2025	143,494	April	Wednesday	K	ASSAULT 3	ASSAULT 3 & RELATED OFFENSES	14

The classification model shows a consistent pattern in crime behaviors across multiple years. The table shows data from 2021-2205, including total arrests, top month, top day, top borough, top ds description, top offense, and top precinct. The model predicted that 2025 would have 143,494 arrests, with April being the highest month, with the top day being Wednesday. It also predicted top borough being code “k”, for Brooklyn, while the most frequent type of offense was assault related. The model predicted precinct 14, which is for central Manhattan, which matches the areas dense foot traffic, having a mixture of tourists, residents, and workers, on top of the heavy influence in its late-night activities, especially in times square. With these predictions, it shows how classification can show crime patterns like when, where, and how. With this insight, police can be better prepared and can anticipate future incidents.

Histogram Gradient Boosting Classification Model:



This is the result of the HGB classification model, showing the model’s accuracy in predicting the top 5 offense types. Overall, the model received an accuracy of 55%. Assault-related offense types were most accurately predicted due to their consistent patterns within the data. Drug and property-related offenses received a lower accuracy, most likely due to their larger geographic distribution.

Conditional Probability Model:

Top Offense by Precinct and Day of Week

	Precinct	Day_of_Week	Top_Offense	count	total_arrests	Offense_Probability
0	1	Tuesday	PETIT LARCENY	163	405	40.20%
1	1	Wednesday	PETIT LARCENY	130	388	33.50%
2	1	Monday	PETIT LARCENY	89	273	32.60%
3	1	Thursday	PETIT LARCENY	117	376	31.10%
4	1	Saturday	PETIT LARCENY	98	315	31.10%
5	1	Friday	PETIT LARCENY	102	371	27.50%
6	1	Sunday	PETIT LARCENY	57	221	25.80%
7	5	Saturday	FORGERY	55	190	28.90%
8	5	Monday	DANGEROUS DRUGS	27	195	13.80%
9	5	Thursday	PETIT LARCENY	47	344	13.70%
10	5	Friday	PETIT LARCENY	28	211	13.30%
11	5	Sunday	ASSAULT 3 & RELATED OFFENSES	16	125	12.80%
12	5	Wednesday	GRAND LARCENY	35	291	12.00%
13	5	Tuesday	OFFENSES AGAINST PUBLIC ADMINI	44	380	11.60%
14	6	Tuesday	DANGEROUS DRUGS	68	209	32.50%
15	6	Friday	DANGEROUS DRUGS	59	216	27.30%

The conditional probability model shows that while some precincts have offense types that occur consistently throughout the week, other precincts do not always show the same consistency. For example, in Precinct 1, the most dominant offense type throughout the week is Petit Larceny, with its likelihood of occurring ranging from 25-40%. This suggests a reoccurring trend in the area, whereas Precinct 5 does not necessarily have the same pattern. Precinct 5 shows to have various offense types occurring throughout the week with a probability of 11-13%. This demonstrates that not all arrest patterns are the same throughout the city, indicating that different precincts should utilize different safety measures,

Conclusions & Next Steps:

The methods used to conduct exploratory data analysis reveal a clear pattern of both spatial and temporal arrests. Specific precincts show a consistently higher frequency of arrests with dominant offense types in those locations as well, which indicates that there is not an even distribution of crime throughout the city. The day-of-the-week hotspot visualization dives deeper into the patterns within the day to show the fluctuations of offense types throughout the week, with some specific offenses appearing more consistently on most days. When considering all the results from the EDA, it is clear that arrest patterns are not random, and law enforcements and communities may utilize this information to better understand that geography and time are essential when attempting to plan more predictive actions against crime.

One improvement we are looking forward to working on is the classification model to help us investigate which exact precinct the offense may occur in. So far, we know that the offenses are based on the most important feature, which is precinct, but on our model, it doesn't show exactly which precinct that will exactly happen in.

The results from the conditional probability model reinforced the findings that the hotspot maps and heat map provided by identifying which offense is most likely to occur on each day of

the week. By calculating total arrest per offense type for each day of the week, the model successfully identified how likely the dominant offense will occur. This method allows for a more defined understanding of the arrest patterns by providing the user with what offense is most probable under certain conditions, rather than just providing where and when arrest patterns tend to occur. This model can be used to form data-driven insights for both law enforcement and communities due to the enhanced details that it provides.

In terms of next steps, additional predictors could be utilized, if incorporated into the data, allowing for a more precise recognition of arrest patterns. This could include the time of day an arrest occurs, the socioeconomic factors of the area, or weather conditions. These results could be used for real-time decision making, such as resource allocation, scheduling patrols, or community planning. These findings may greatly improve safety planning in many areas throughout New York City.

Website with deliverables: https://kristennkane.github.io/NYPD-Capstone-FAU-2025/

References:

NYC Planning. Visit nyc.gov. (n.d.). https://www.nyc.gov/content/planning/pages/resources/datasets/police-precincts

(NYPD), P. D. (2025, April 15). *NYPD arrests data (historic): NYC Open Data*. NYPD Arrests Data (Historic) | NYC Open Data. https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u/about_data

(NYPD), P. D. (2025b, October 27). *NYPD arrest data (year to date): NYC Open Data*. NYPD Arrest Data (Year to Date) | NYC Open Data. https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc/about_data

Lau, T. (2020, April 1). *Predictive policing explained*. Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained