Report on NYC Arrest Data from 2019-2024

This short report is intended to discover trends in NYC arrests, determine the top 5 most frequent arrests in the city, compare arrest rates between specific precincts, and culminates with a proposal for a model that predicts crime to improve NYPD resource allocation.

Q1: Arrest Rate Trend

Utilizing New York City Arrest data, the first goal of the analysis was to visualize the overall number of arrests in all five boroughs and determine any notable patterns. Using the historic NYC arrest data, I obtained arrest counts from 2019 to 2023. I merged this subset with year-to-date arrest data to incorporate 2024 values. Seen in Figure 1 below, the number of arrests *is increasing* each year, except for 2020. The COVID-19 pandemic and subsequent shifts in social and behavioral patterns could serve as one potential explanation for this trend. Yet, the increase in arrests in the following years is interesting to note and warrants deeper investigation.

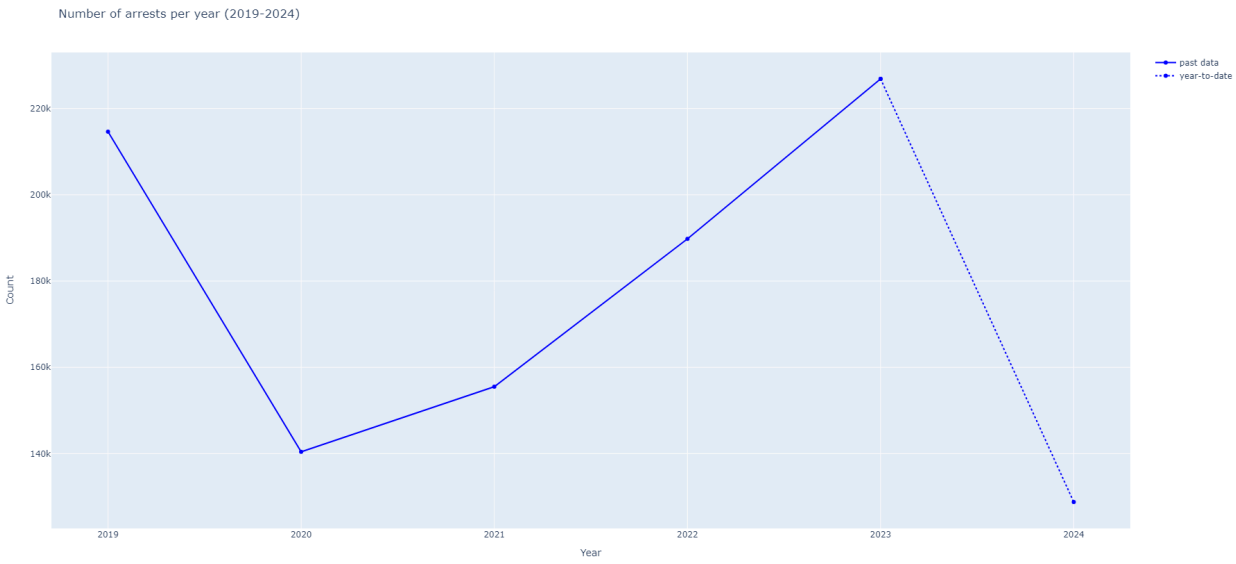

Figure 1 *depicts the total number of reported arrests per year between 2019 and 2024.*

Q2: Top 5 Arrests

To determine the top 5 most frequent arrest types, some pre-processing of the data was required. For arrest descriptions, I decided to impute the null values to reflect the information provided in the footnotes, that the information was either not available, unknown, or previously not reported. As such these values were simply replaced with the characters ‘U/NA/NR’. With the arrest counts, the top five most frequent arrest descriptions were:

1. Assault (3rd degree)
2. Petit Larceny
3. Assault (1st, 2nd, and unclassified)
4. Robbery (open areas and unclassified)
5. Traffic (unclassified misdemeanor)

These results are shown in Figure 2a. I also viewed these trends per year, to see the most frequent arrest types annually. As shown in Figure 2b, assault in the third degree has been the most frequent arrest type from 2019 to 2024, with petit larceny as the second most frequent each year as well. While most arrests saw a somewhat expected dip in numbers from 2019 to 2020, robbery and public administration arrests increased, earning the fourth and fifth spots in 2020, respectively. From 2020 to 2024, unclassified, 1st and 2nd-degree assaults have taken the third spot and are also slowly increasing each year. There was an almost 64% increase in reported larceny arrests between 2021 and 2022, in contrast to the 52% decrease from 2019 to 2020.

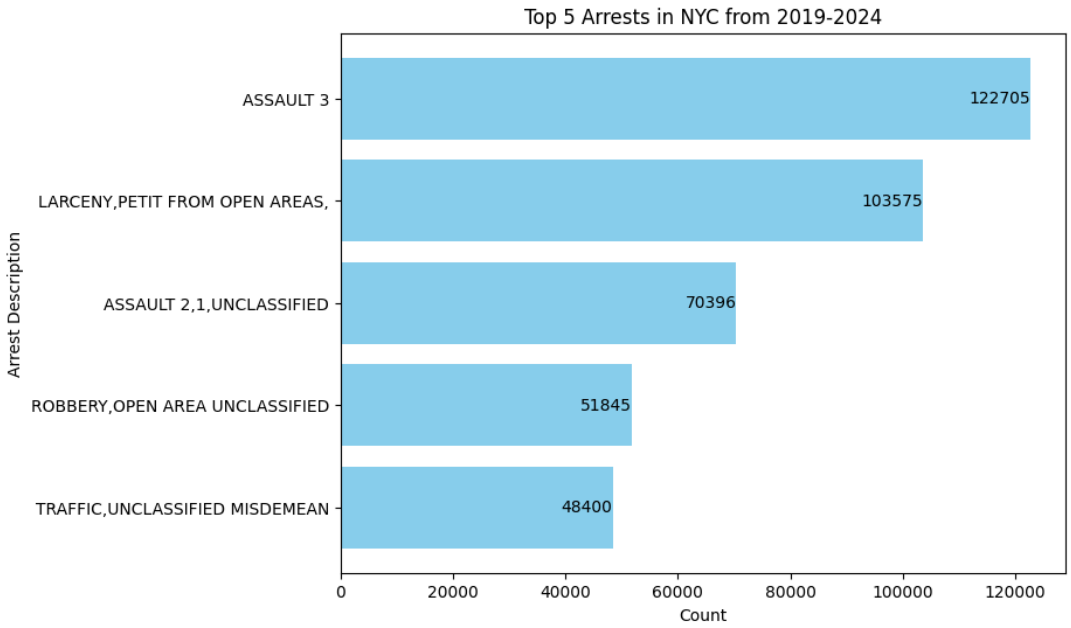

Figure 2a *shows the five most frequent arrest types in NYC from 2019-2024.*

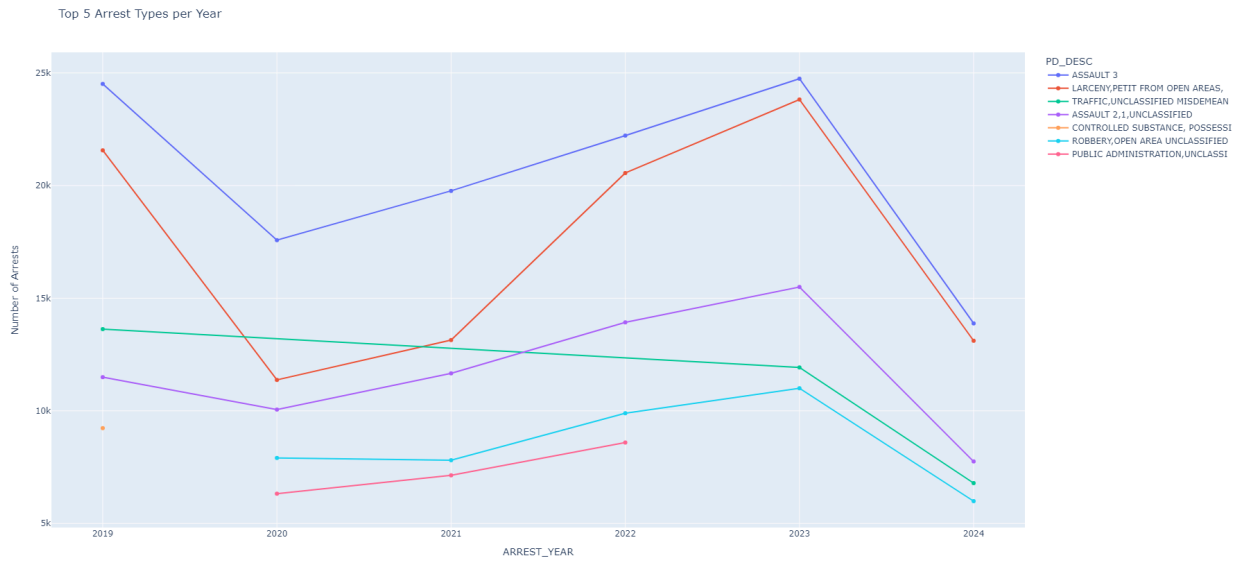

Figure 2b *shows the 5 most frequent types of arrests per year, from 2019-2024.*

Q3: Crime Comparison

The next step during analysis was to compare crime or arrest rates at two different precincts (Precinct 19 in the Upper East Side and Precinct 73 in Brownsville). To accomplish this, I utilized the two-sample or independent t-test, designed to compare averages between two independent samples or populations. This test assumes that the two samples are independent of one another, that the data is normal (when plotted, the data should generate a bell-like curve), and that the variances of the samples (how far each number is from the mean) are equal. Ensuring that these assumptions were met, I obtained monthly arrest data for each of these precincts and used the total number of precinct arrests per month between 2019-2024 as the outcome variable. With these numbers, I ran the t-test with a null hypothesis that there is no difference in crime (arrests) between Precincts 19 and 73, and an alternative hypothesis that there is a difference in crime between these precincts. In this test, the p-value is used to determine how likely the results of the test are due to chance, with a common standard of 0.05 as the threshold to meet for significance. Results from this test gave a p-value of below 0.05, signaling that the null hypothesis can be rejected and that the data provides evidence of a difference in the arrest rates between these two precincts that is statistically significant.

It is prudent to note this is a smaller range of data, so repeated sampling and even data encompassing a larger time frame may be appropriate measures to incorporate for gathering more evidence of where crime occurs more frequently between these two locations. However, based only on the descriptive statistics, average monthly arrest numbers are higher in Brownsville than in the Upper East Side (see Figure 3).

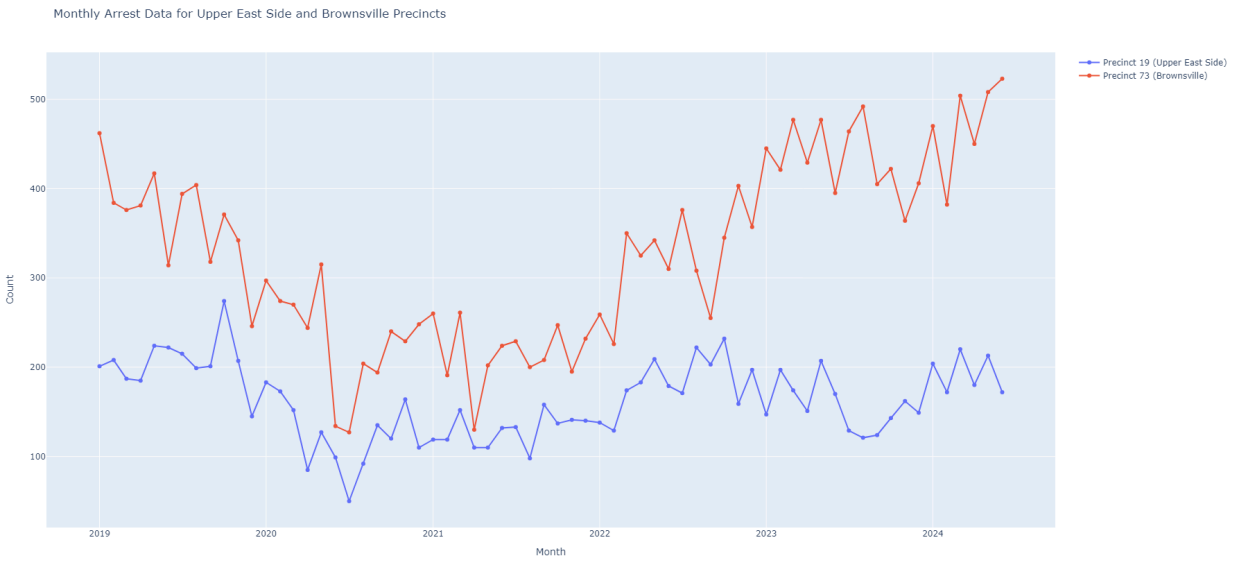

Figure 3 *shows the monthly arrest counts at the Upper East Side and Brownsville precincts*

Q4: Predictive Modeling

I propose that a predictive model be developed, with the purpose and intention of predicting arrests (and potentially more crime), so that the NYPD can better allocate its resources to preventing it. Depending on the outcome that would best fit the needs of the NYPD, I present options for consideration:

As one approach, I recommend predicting the level of offense based on the jurisdiction responsible for the arrest, location (arrest precinct, arrest borough, latitude and longitude), and arrest demographics (race, sex, age). This can be achieved using a model with classification capabilities such as Random Forest, XGBoost, or Support Vector Machines. These models are recommended because they, either alone or with ensemble learning methods, can combine multiple algorithms to make predictions. I propose utilizing the accuracy features of these models as the primary method of evaluating their performance, as well as using supplemental metrics as needed, such as precision and recall. In this case, the accuracy of the model is interpreted as the amount of correctly predicted offense levels, out of all predictions made by the model.

Another approach that could yield insights would be using a clustering algorithm to predict areas where certain crimes are committed, in the hopes of earlier intervention. One avenue for this would be utilizing data from the top 5 most frequent arrests, with information on jurisdiction responsible for arrest, arrest type, level of offense, and demographics, to generate clusters and find meaningful patterns, such as similarities and differences in arrests between precincts. The k-means clustering algorithm would be a good fit for accomplishing this task. This can be

evaluated using the Silhouette coefficient, to determine how well the model as grouped similarities and separated dissimilar observations.

Another model that can be beneficial for resource allocation is predicting when arrests (and potentially more crime) will occur. For this, we can assume that crime occurs more when certain opportunities arise recurringly throughout the year. Thus, I would recommend a model capable of time-series analysis with consideration of seasonality, like the Prophet model. With monthly arrest data over several years, this model can be used to predict when future crimes and arrests might occur with varying degrees of granularity (e.g. day, month, year). This approach can be applied at different levels: forecasting arrests by borough, predicting 'seasons' when the top 5 most frequent arrests occur, etc. Prophet also includes cross-validation functionality to account for forecast error.

Potential challenges that can be faced in developing these kind of models would be ensuring the accurate representation of the whole population(s), within the samples used for modeling. Correcting for data imbalances that exist (e.g. disproportionate number of arrests between boroughs, disaproportionate arrest types, etc) would be important for developing more accurate models. Additionally, with the systemic issues that can exist for arrests in urban areas, it's important to account for biases in the data that could also manifest in a model. Specifically in working with NYC arrest data, some ways to mitigate these issues would be to incorporate resampling, essentially taking more samples of underrepresented data, and taking fewer samples of overrepresented data. Additionally, utilizing varied algorithms and evaluation metrics to perform the same prediction task can be used to further verify and ensure more accurate results and representation. Finally, it’s important for those with investment in the data and its outcomes to have a full understanding of it, screen for bias at each stage of the data life cycle, and ensure that the models are used and deployed in an appropriate and ethical manner.