# NYPD Arrest Analysis

Marco Hui

2022-10-09

Q1: Has the arrest rate been decreasing from 2018-2022?

Arrests in New York City have shown a general decrease over previous years, from 20564 average monthly incidents in 2018 to 15539 average monthly incidents in 2022. This is evident from the green linear regression line in the month by month breakdown below.[1]

Diving deeper, number of arrests starts off at 23882 cases in January 2018, dips to just 7461 cases in July 2020, then rises back up to 16407 cases in Jun 2022. This nuance is captured by the orange loess curve. The significant drop in 2020 is most likely due to COVID-19 guidelines that discouraged people from being in public and limited human interactions across the city. When restrictions eased, the number of arrests rose back to a level more normal, which is still lower than where it began in 2018.


Monthly Arrest Count (Jan 2018 – Jun 2022)
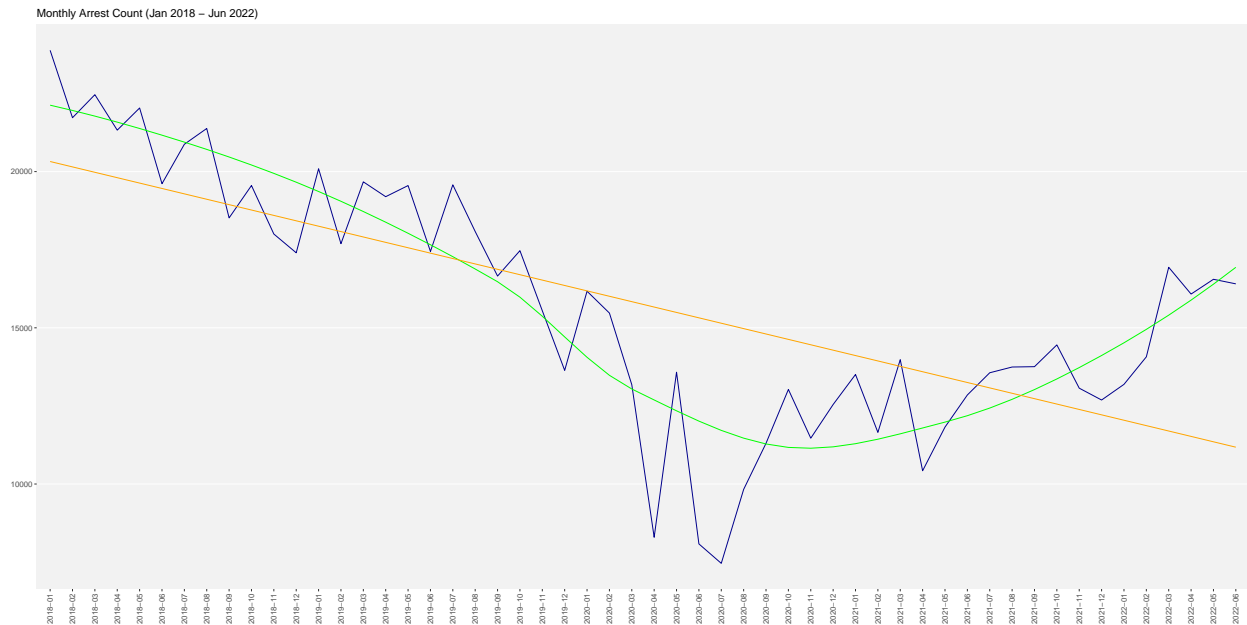
Q2: What are the top 5 most frequent arrests as described in the column 'pd_desc' in 2018-2022?

The top 5 most frequent arrests as described in the column 'pd_desc' between January 2018 and June 2022 are: assault (3), larceny (petit from open areas), assault (1, 2, unclassified), traffic (unclassified misdemean), and robbery (open area unclassified). The percentage of total for each type of arrest is labelled (the numbers are not very high because of how granular the data is).

---

[1]Without data on the population of New York City, it is difficult to calculate arrest rate data. Assuming that the population did not change drastically between 2019 and 2022, looking at arrest counts should be a reasonable proxy for arrest rate.
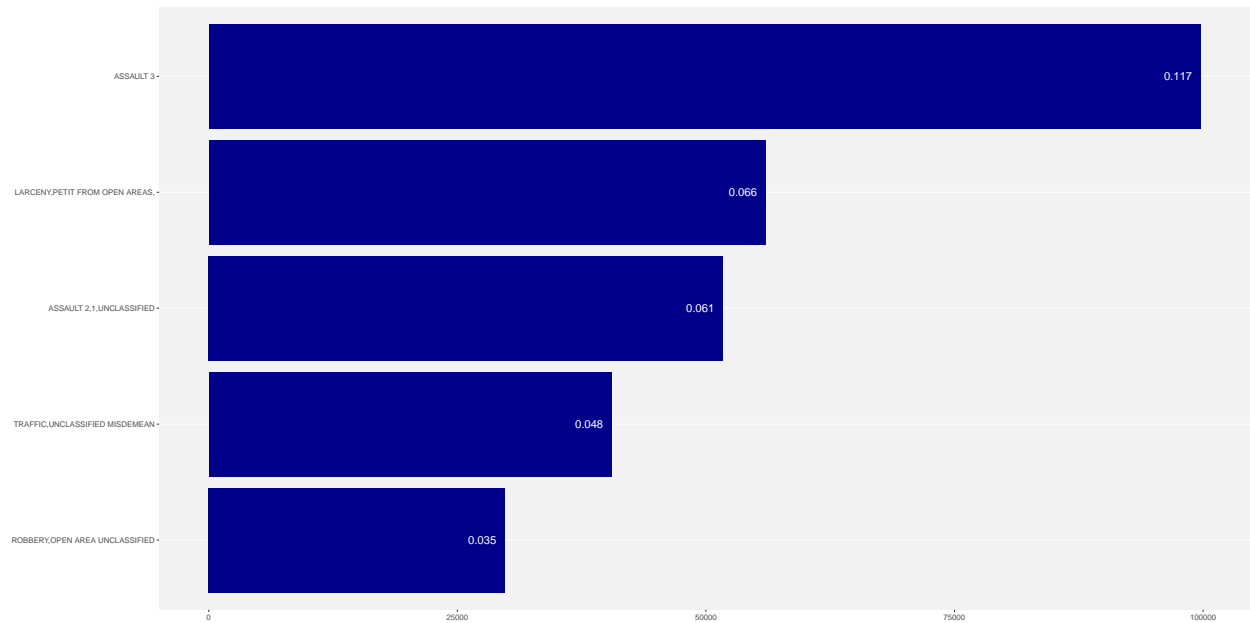
Following the overall decline of arrests in the past 30 months, these 5 offenses all trend downwards. Traffic arrests in particular has over time become the least frequent among the 5.
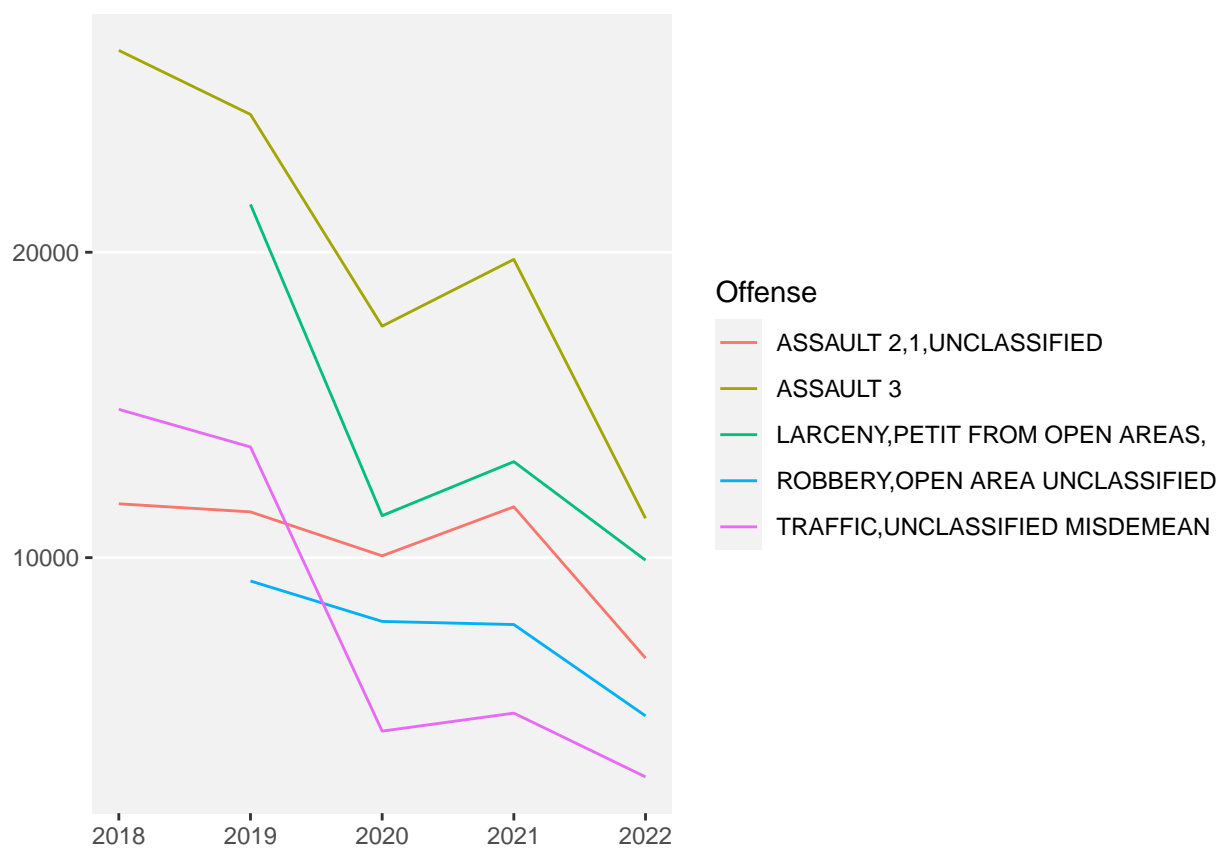
Regardless, the top 5 most frequent arrests remains fairly consistent from 2018 to 2022.
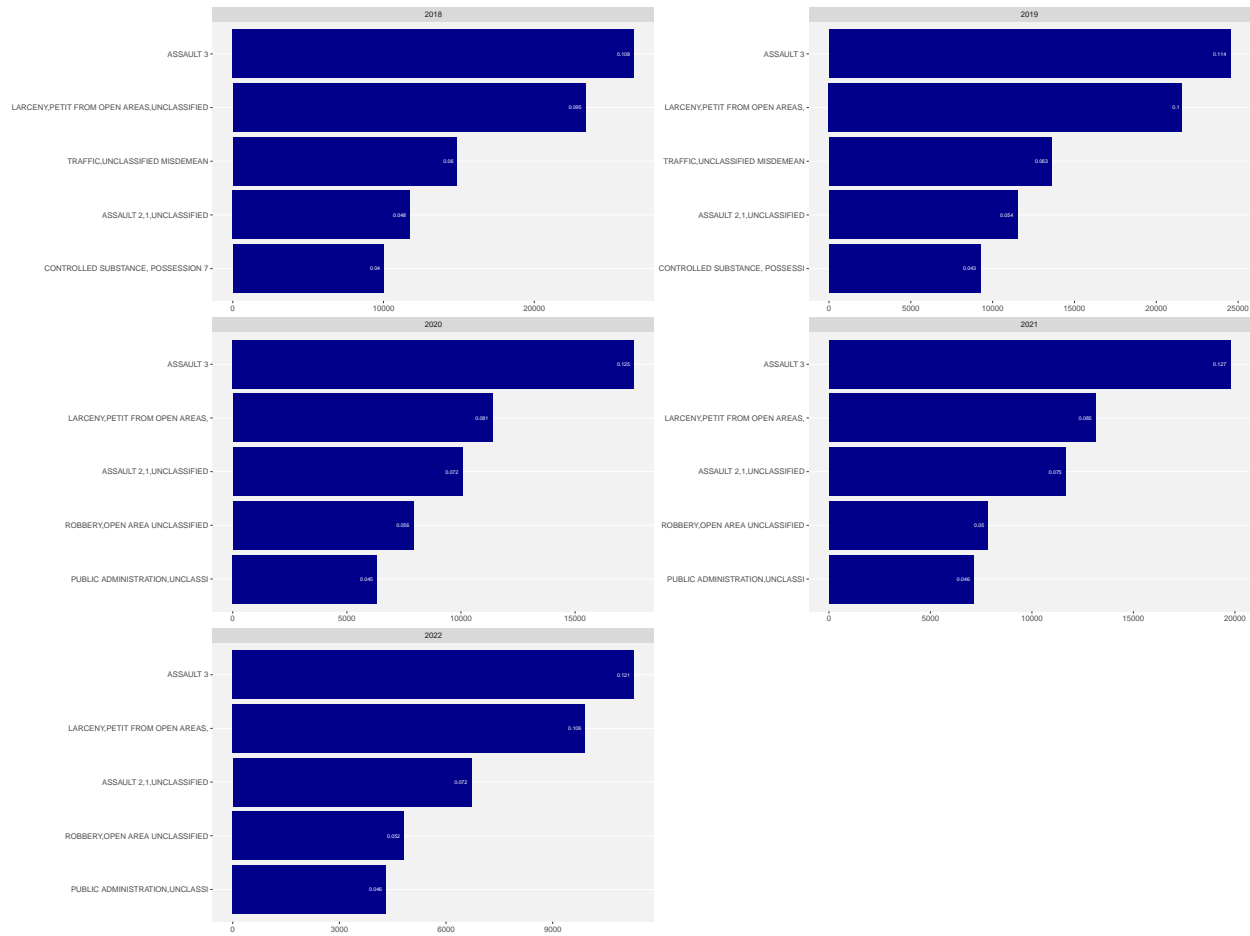
Q3: If we think of arrests as a sample of total crime, is there more crime in precinct 19 (Upper East Side) than precinct 73 (Brownsville)?

At first glance, there seems to be more crime in precinct 73 (Brownsville) than in precinct 19 (Upper East Side). Between January 2018 and June 2022, Brownsville has seen 17058 arrests cases while Upper East Side has seen just 8871 cases.

This difference in amplified considering Brownsville is noticeably smaller than Upper East Side in area. At 3.01 km^2, Brownsville has 5667 arrests per 1 km^2 over the past 30 months. At 4.60 km^2, Upper East Side has 1928 arrests per 1 km^2 over the past 30 months.[2]

The number of arrests in Brownsville is not only higher overall, but also much more variable in comparison.

---

[2]Resident population and general foot traffic would also be important to look at when judging crime between two areas. Something to consider for future analysis with more data.

Distribution of Monthly Arrest Count (Jan 2018 – Jun 2022)
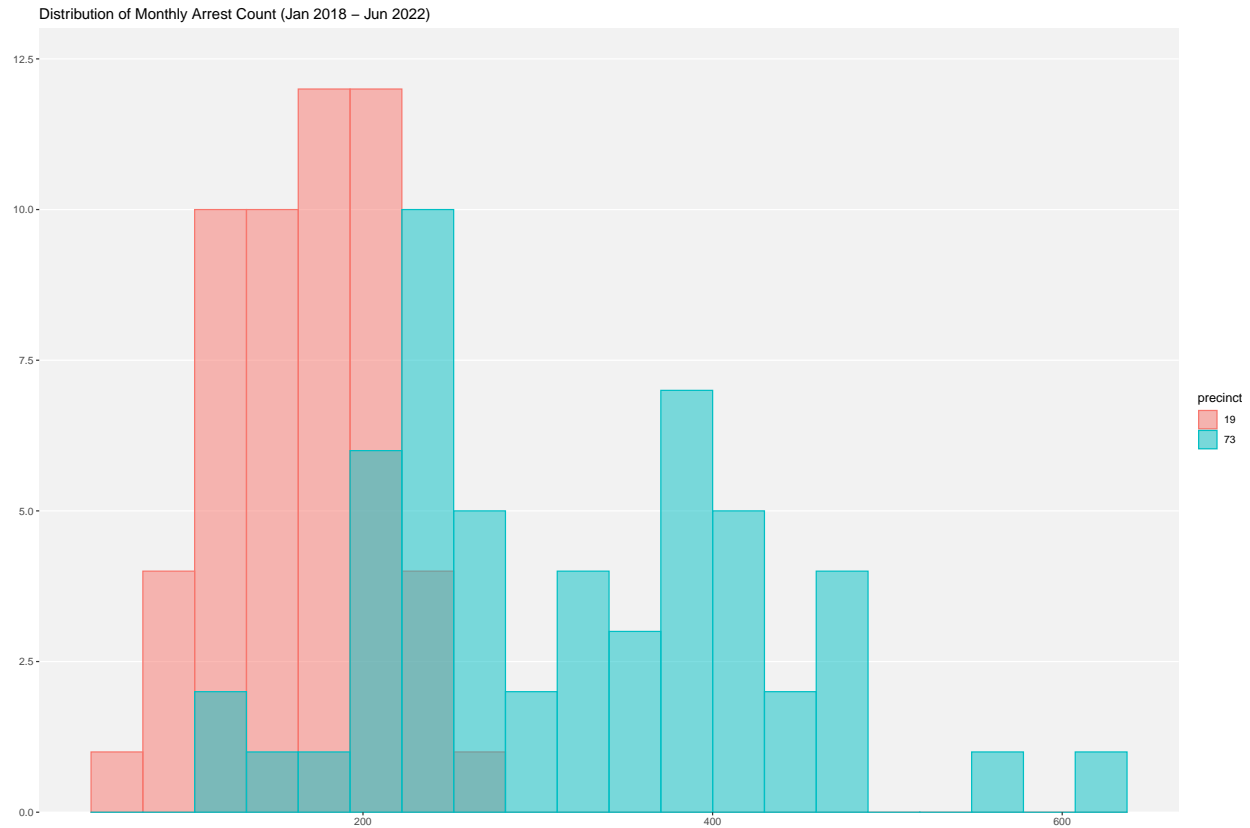
This makes sense when looking at the trends over time. As expected, crime in Upper East Side starts off and is generally lower. However, the occurrences of arrests in Brownsville have decreased more significantly over the past 30 months.

Monthly Arrest Count By Precinct (Jan 2018 – Jun 2022)
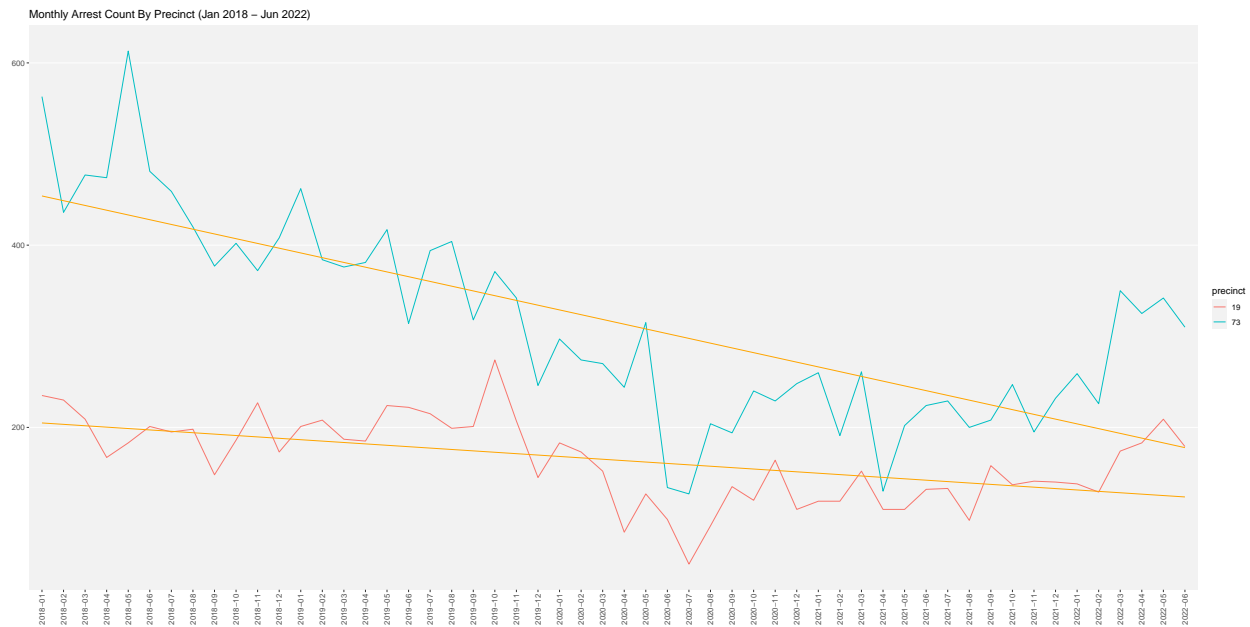
Q4: Given the available data, what model would you build to predict crime to better allocate NYPD resources?

Given the available data, I would build a random forest model.

The dependent variable would be the type of offense reported in OFNS_DESC. This was chosen because it has clear, distinct values and is meaningful; being able to anticipate the kind of crime allows the NYPD to better prepare themselves to handle the arrest. This also explains the choice of the multiclass classification random forest model. [3]

The independent variables would be various features of an incident such as AGE_GROUP, PERP_SEX, ARREST_BORO, etc. Some may have to be one-hot encoded into dummy variables.

The model would be generally be evaluated on its ability to correctly predict the type of offense in OFNS_DESC in a test set (a random subset from the total data) given just the inputted features. Some simple statistics I would look at to judge effectiveness would be the accuracy, precision, and recall scores.

A challenge that I forsee is the large number of possible output classifications. Although OFNS_DESC has fewer unique values than PD_DESC, there is still a decent amount. To make a more generalisable model, I would probably have to look through the possible types and consolidate some.

However, a good random forest model can give some insights on what attributes are more common for certain type of crimes, giving officers a better idea of what to look for.

Notes

Data used in this report is last updated 10/08/2022.

---

[3]Predicting whether or not a crime leads to an arrest would be interesting too and very relevant to studying prejudice in arrests. However, the lack of data on misdemeanors without leading arrest prevents us from constructing the necessary dependent variable.