Dr. Rachael Alexandroff

NYPD arrests: historic analysis

Summary
Overall the average number of arrests per day in NYC has been steadily decreasing by approximately 8 arrests a day each month from January 2015 to December 2018. There remains a disparity in trends across different arrest types and precincts with some kinds of arrests remaining flat or increasing in total number. While among both precincts examined the average number of arrests is decreasing, the average number remains statistically higher in certain precincts compared to others.

Initial data collection and assumptions
Data was downloaded from the NYC OpenData NYPD Arrests database via the Socrata Open Data API. This allowed me to filter for the four years of interest before downloading the data, and does not require the data to be stored on the hard drive of my computer. This was desirable given the large size of the dataset. To work with numbers and dates more easily, the data was 'cast' to the correct data type (numeric or datetime) before use. Finally, I created variables to store the year and month of the arrest record individually to make analysis and grouping of the data easier.

Overall arrest rate
The overall arrest rate has decreased from January 2015 to December 2018 across all precincts and all arrest-types (see Figure 1). This trend was identified by counting the total number of arrests per day in New York and using this value to find the average number of arrests per day in each month over the given timeframe. The standard deviation is calculated using the spread in the total number of arrests per day over the course of the month. Visually, it appears as if the average number of arrests per day is decreasing month over month.

To better understand and confirm this trend, I fit a line to the available data, using the measured standard deviation as the error. Fitting was done by minimizing the least squared error (predicted arrests – actual arrests)^2 to pick a model line that minimized the distance of each data point from the best fit line. The best fit model concludes that the average number of arrests per day has been decreasing by approximately eight arrests a month (so the average number of arrests a day in January 2016 was approximately 8 arrests less than the average number of arrests a day in December 2015). To better understand this trend requires a deeper-dive in to the data.

Arrests by type
To better understand the pattern of arrests, I broke them down by type of arrest. For this exercise I didn't combine different descriptions, even if they appeared similar to me, as I didn't feel I had enough familiarity with the dataset to make assessments on what data could be properly combined while still representing the breadth of the data correctly. One thing I would do in future would be to speak to someone with knowledge of how these codes are created to ascertain if any categories could or should be combined. That being said, I found that in 2018 the highest number of arrests were made for (in order of number of arrests): 1) ASSAULT 3, 2) LARCENY,PETIT FROM OPEN AREAS,UNCLASSIFIED, 3) TRAFFIC,UNCLASSIFIED MISDEMEAN, 4)

ASSAULT 2,1,UNCLASSIFIED, and 5) CONTROLLED SUBSTANCE, POSSESSION 7 with the incidence of the fifth-highest arrest type being only half as prevalent as the first (Figure 2).

Interestingly, there is a big gap between the first and second most common types of arrests and the third most common type. Also, there appears to be a significant decrease in the total number arrests of controlled substance possession from 2015 to 2018. This may relate to changing attitudes towards the criminalization of drug possession over the past decade, such as the passage of a state Good Samartitan Law in 2011 and Naloxone possession and administration law in 2006[1]. Any results here however remain separate from changing attitudes and legislation regarding the use or possession of marijuana which is not a 7th degree possession (assuming that is what the '7' in the description refers to)[2].

Total crime in precinct 19 vs. precinct 73

I would be hesitant to assume that the number of arrests is representative of the amount of total crime in a particular precinct. It is likely also correlated with the amount of police presence in a particular precinct as well as other factors such as historic policing records etc. I would want to take all of this in to account before making statements about the amount of total crime in particular precincts. The current results do show a higher record of arrests in precinct 73 than precinct 19 but appear to show that the gap in number of arrests is decreasing over time (Figure 3). Because we have less crime data for each precinct than the whole city, here I have calculated the total number of arrests per month in each precinct and then averaged them over the whole year to produce the given figure.

To test if the average number of arrests per month is in fact different for a given year I used a T-test to reject the null hypothesis (that is to reject the hypothesis that the average number of arrests per month in each precinct are in fact the same). We can use this test if we assume that the two samples are independent (higher crime in one precinct doesn't mean higher crime in another precinct) and that both samples are normally distributed. These assumptions should be fine for an initial assessment and I am able to reject the hypothesis that the average arrest rate in both precincts is the same with high confidence ($p << 0.01$).

Model to predict crime

One of the most straightforward ways to predict future crime would be to look at what features of a neighbourhood have correlated with crime in the past, and use these to predict future crime. It is difficult, however, to know about the actual crime rate as we only have data on arrests which tracks some function of actual crime historic policing leaving us with biased data. One way to deal with this might be to actively examine any correlations between neighbourhood factors (e.g. demographics like age, race, education level, earnings) and crime rate and remove (by fitting and subtracting off the correlation) any which are due to historical

[1] http://www.drugpolicy.org/new-york-state-drug-laws
[2] https://www.sunyrockland.edu/about/college-policies/general-administration-policies/drug-and-alcohol-free-campus/nys-penal-law-controlled-substances

bias. This would have to be done in conversation with the stakeholders of the model so as to ensure there was an agreement on which features to ignore. I would want to evaluate this model monthly with new data moving forward.

Figures

Figure 1: Average number of arrests a day each month from January 2015 to December 2018 in NYC. The actual data, with error bars equivalent to one standard deviation, are shown in green, while a fit to the data is shown in red.

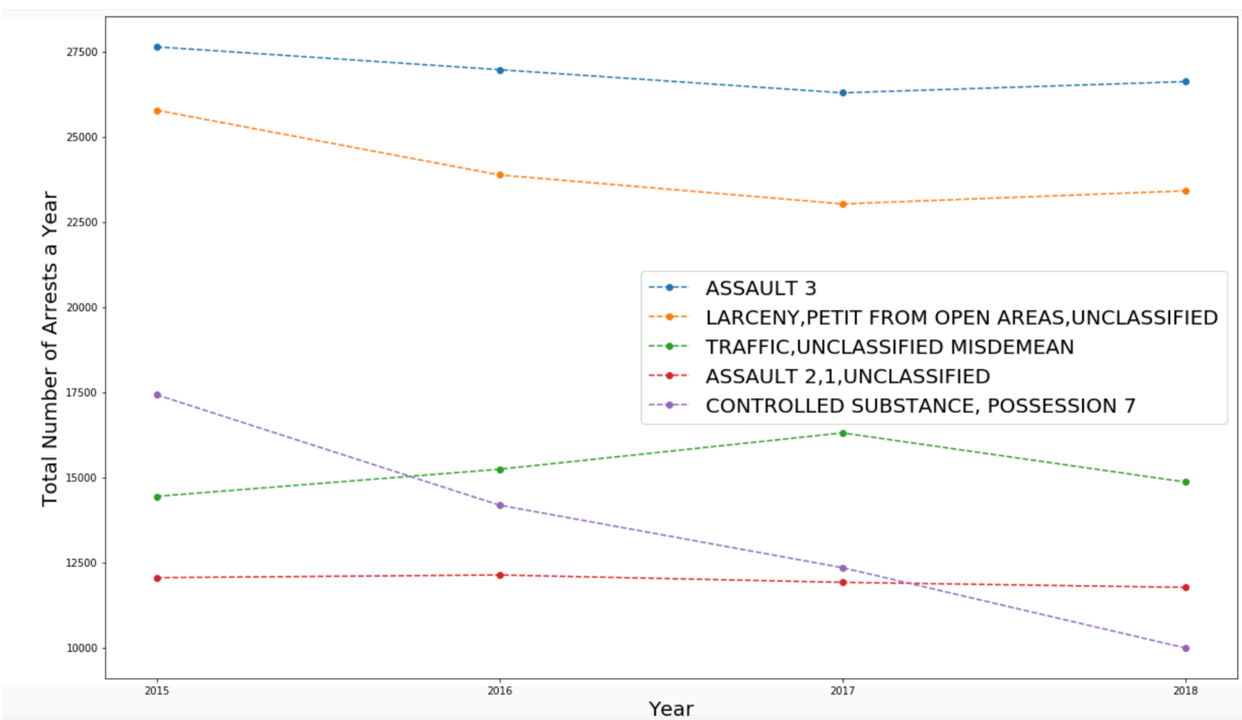

Figure 2: Total number of arrests per year for the five most common arrest classifications in 2018.

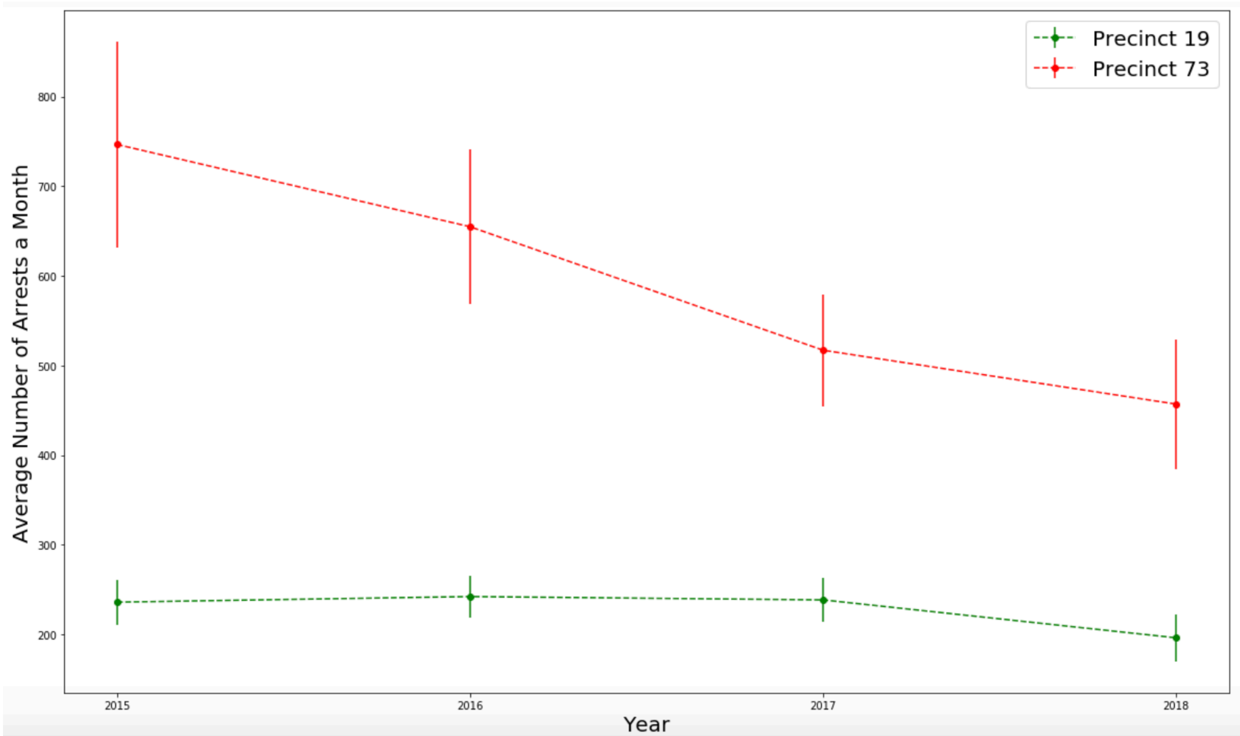

Figure 3: Average number of arrests per month, with error bars showing the standard deviation, for every year from 2015 to 2018 in two different NYC precincts.