NYCC 2015-2018 Arrest Analysis

James Wu

1 Introduction

The NYC Council's Public Safety Committee has asked the Data team for an analysis of arrests from 2015-2018. They want to know about whether arrest rates have decreased and overall trends in arrests during this period. Based on this and the districts of interest, the questions we will explore in this report are as follows:

- **Has the arrest rate been decreasing from 2015-2018**
- **What are the top 5 most frequent arrests types in 2018?**
- **How has the overall trends of arrest types changed from 2015-2018?**
- **Is there more crime in precinct 19 (Upper West Side) than in precinct 73 (Brownsville)? What are the trends?**

2 Data

We are using NYPD Arrests Dataset available at *https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u* for this study. Each record in this dataset represents an arrest effected in NYC by the NYPD and includes information about the type of crime, the location and time of enforcement, and the arrestee's demographic. As the dataset has records from 2006, we used only a subset of the data, including only 2015-2018 records. This data was reviewed for consistency and expected patterns, then prepared (i.e. formatting, creating new variables) for the study.

3 Methodology

All data preparation and analyses were conducted in R. In deciding appropriate models to answer the questions, we first explored how the data looked like with daily, monthly, and annual aggregated counts. We saw similar patterns in all three aggregations.

One particular property of the data is its time-element. When looking at the aggregated monthly arrest plots, the data seems to display signs of having a seasonal pattern. Every start of the year has a spike in arrests in January, followed by a sharp decline in February. There are usually higher arrests in March, May, and August following lower arrests in February, April, and July. The seasonality could be potentially explained by: (1) there may be less or more cops deployed for certain months due to holidays, or annual events, (2) people's behaviors may also be influenced by seasons, temperatures, holidays, etc. (3) the cyclical nature of new hires and people leaving, and other potential reasons.

Since we are concerned about overall trends, monthly and yearly aggregated data is primarily used for our linear models to reduce the effects of omitted variable bias and to remove seasonality. If seasonality is removed, the independence assumption for linear regressions is more believable.

4 Analysis and Results

4.1 Arrest rate trends

We found the general trend of arrests has been steadily decreasing from 2015-2018. There were 339,470 arrests in 2015, dropping to 246,773 arrests in 2018.

To quantify this trend, we used a linear regression model on the aggregated yearly data removing seasonality. The specified model is: $CRIME\_COUNT = \beta_0 + \beta_1 YEAR$. We found $YEAR$ variable to be statistically signicant at an alpha of .05, and negative which signifies a downwards trend. On average, the number of arrests have decreased by 30,673 per year.

4.2 Evolution of the top 5 arrest reasons

The top 5 arrest reasons in 2018 were 3rd degree assault, petit larceny, traffic misdemeanor, 2nd degree assault, and 7th degree possession of controlled substances in that order. In contrast, the top 5 arrest reasons in 2015 were theft of services, 3rd degree assault, petit larceny, 7th degree possession of controlled substances, and unclassified NYS laws in that order.

Theft arrests have decreased drastically over the years, as well as arrests under "unclassified NYS laws" (I suspect this may be because this classification may have been spread out to other offenses). 3rd degree assault, petit larceny, 2nd degree assault, and unclassified traffic misdemeanors have had relatively stable number of arrests throughout the years. Refer to Figure 2 in the Appendix for more details.

4.3 Precinct 19 vs. precinct 73

We found that precinct 73 (Brownsville) had a much higher arrest count than in precinct 19 (Upper East Side) with also a much larger variance month to month (refer to Figure 3). However, while arrests have been decreasing at a fast rate in Brownsville, arrest counts in the Upper East side have been relative stable from 2015-2018.

To quantify this trend, we used a linear regression model on the aggregated yearly data removing seasonality. The specified model is: $CRIME\_COUNT = \beta_0 + \beta_1 YEAR + \beta_2 PRECINCT73 + \beta_4 YEAR{*}PRECINCT73$, where $PRECINCT73$ is a indicator for whether it is Brownsville or not. We found that Brownsville has on average 2,143,694.1 - 1,060.90 * $YEAR$ more crime than Upper East Side. This also tells us that while both Brownsville and Upper East side has a decreasing trend (negative slopes), Brownsville's arrest count was decreasing on average 1,060.90 more than Upper East Side's decreasing trend of 147.9 per year.

5 Future Work - Predictive Modelling

To optimize NYC's police resources, we could potentially build a predictive model based on past arrests to predict the likelihood of a crime happening in any district. A logistic regression with added seasonality terms can be used for this.

The difficulty of this task could be the construction of the classifier. There are usually at least 1 arrest daily, so a binary term indicating whether an arrest was made in a day would not be particularly useful (ie. 100% every day, every where!) A solution for this could be creating subjective cutoff point(s) (e.g. high crime/medium crime/low crime) for the classifier.

There is also the problem of the ethics in equating arrests with crime. We run the risk of perpetuating prior biases due either historic bias or a change of law (e.g. marijuana decriminalization). We can reduce the probability of happening by excluding sensitive demographic data. All variables could be included in the model, excluding the repeat ones such as code and its description (choose one to avoid multicollinearity), *ARREST_KEY*, and *PERP_RACE*.

Our model may be evaluated by using a certain time-frame as the training data, such as 2015-2017, and another outside the timeframe as test data, like 2018. We can look the model's accuracy and mis-classification rate.