NYC Council Data Scientist Data Challenge Report

NYPD Arrest Data Analysis

Yushi Chen

New York City has a population of 8.5 million people, detecting crime and preventing crime is very important for public safety in New York City. Analyzing historical NYPD arrest data is very helpful to understand the trend that arrest and crime happen in the temporal and spatial dimension, it also delivers important information for the future allocation of NYPD resources. This project analyzes the NYPD arrest data from 2015 to 2018 and trying to find the overall trend in arrest and bring valuable insights and references for police enforcement in the City.

The data of the analysis used open-source NYPD arrest data. Each data record represents an arrest affected in NYC by the NYPD and includes information about the type of crime, the location and the time of enforcement. It also includes suspect demographic information. To understand the trend changes from 2015 - 2018, this project analyzed the data from the following perspectives:

1. Total number of arrest changes of 2015 – 2018

In the crime dataset from 2006 to 2018, each data record represents one arrest that happened in NYC. By aggregating the count for arrests in each year, the number of total crimes is represented in *Table 1* and *Figure 1*. According to *Table 1* and *Figure 1*, the number of arrests in 2015 was 339,470 and 246,773 in 2018. The number of arrests decreased in each year. There was a decreasing trend of the total number of arrest from 2015 – 2018 from the plot; however, to further test the linear relationship of the number of changes in these four years, I have calculated the Ordinary Least Square Coefficient by using Numpy in python, and found that the coefficient on the time variable is negative (-30673), so the observations are decreasing over time.

2. Top 5 most frequent arrests as described in the column 'pd_desc' in 2018

In the crime dataset, the “pd_desc” represents the description of internal classification corresponding with PD code, and the type of arrest in this column is more granular than the Offense Description. By aggregating the “pd_desc” column and sum up the total number of arrests in each type of arrest, the five largest number of arrest was represented in *Table 2*, and the top 5 arrest type is: “Assault

3", "Larceny, Petit from Open Area, Unclassified", "Traffic, Unclassified Misdemeanor", "Assault 2,1, Unclassified", "Controlled Substance, Possession7". According to *Figure 2,* the type arrest changes in different months in a year, the trend of "Assault 3", "Larceny, Petit from Open Area, Unclassified" were similar, which has a higher arrest number around May and August, and lowest arrest around December. Furthermore, the trend of arrest "Assault 2,1, Unclassified", "Controlled Substance, Possession7" were similar, which has a higher arrest number around August and October, and lower arrest number around June. Moreover, according to *Figure 3,* which represents the changes in the number of arrests from 2006 to 2018 in the top 5 types of arrests, the number of "Controlled Substance, Possession7" was decreasing over time from 2006 to 2018. The number of arrests of "Larceny, Petit from Open Area, Unclassified", "Traffic, Unclassified Misdemeanor" was generally in an increasing trend from 2006 to 2018; however, the speed of increasing got slower during 2015 to 2018 period. The number of arrests in "Assault 3" and "Assault 2,1, Unclassified" was relatively stable over the time from 2006 to 2018.

3. Is there more crime in precinct 19 (Upper East Side) than precinct 73 (Brownsville)? Describe the trend, variability and justify any statistical tests used to support this conclusion.

For this question, we assume the arrests as a sample of total crime and can represent the total crime. According to *Figure 4,* which represents the total changes of arrest in precinct 19 (Upper East Side) and Precinct 73 (Brownsville). The total number of arrests in Brownsville was higher than the Upper East Side. And in both two areas, the number of arrests was in the trend of decreasing. I have used the T-test and Analysis of Variance Test (ANOVA) to prove whether the means of two paired samples are significantly different. The null hypothesis of the test is there are no significant differences between the two sets of values. The P-value from the two tests was 0.002, which proved strong evidence against the null hypothesis and proves that the two sets of the sample are significantly different.

4. Other Exploratory Analysis Arrest number

4.1 weekday vs. weekend from 2015 - 2018

According to *Figure 5,* Wednesday has the highest number of arrests from 2015-2018, and Sunday has the lowest number of arrests. Furthermore, the number of arrests increased from Monday to Wednesday and decreased from Wednesday to Sunday.

4.2 Arrests changes over time in each borough

According to *Figure 6,* the pattern of changes in the number of arrests was similar in five boroughs. Brooklyn has the highest number of arrests, and Staten Island has the lowest number of arrests. Based on the trend shown on the figure, There was a higher amount of arrests in March and August, and a lower amount of arrests in January and December. According to *Figures 7 & 8,* misdemeanor was the highest amount of arrests, the second highest was Felony. Misdemeanor also has the highest amount of arrests in each borough.

4.3 The number of arrest for the level of offense

According to *Figure 9,* The age group of 25-44 has the largest amount of arrests, and over 65 has the lowest amount of arrests.

5. What model would you build to predict crime to better allocate NYPD resources?

After understanding the time-series information on arrest data and the characteristic of arrests changes over time. The next steps will be implementing various regression models and other machine learning models to predict the crime and assist NYPD for better-policing allocation. One of the challenges in this stage will be feature extraction. PCA will be a good method to decompose the dimension and analyze the correlation between features that we extracted, as well as analysis time series data will also useful for feature prediction. The features that will put in the machine learning model will be numerous, so I will need to enforce sparsity via regularization or some form of dimensionality reduction. For the machine learning model, I will try to implement a neural network, specifically a recurrent neural network. RNNs are good for capturing patterns in time series data, which is how this data is structured. In that case, we would implement the neural net in Torch.

Appendix:

	ARREST_DATE	ARREST_COUNT
0	2015	339470
1	2016	314864
2	2017	286225
3	2018	246773

Table 1: Number of arrest from 2015-2018

	PD_DESC	ARREST_COUNT
0	ASSAULT 3	26611
1	LARCENY,PETIT FROM OPEN AREAS,UNCLASSIFIED	23405
2	TRAFFIC,UNCLASSIFIED MISDEMEAN	14856
3	ASSAULT 2,1,UNCLASSIFIED	11763
4	CONTROLLED SUBSTANCE, POSSESSION 7	9982

Table 2: Top 5 arrest type in 2018

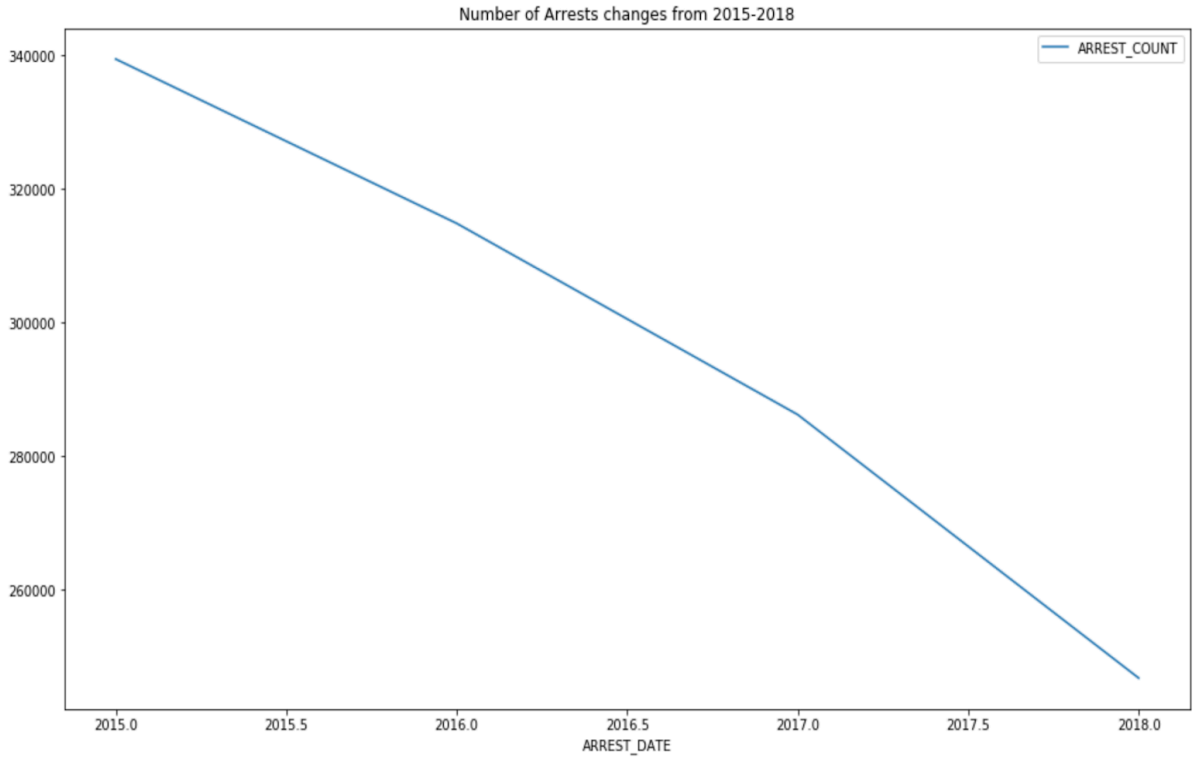

Figure 1: Number of arrest from 2015-2018

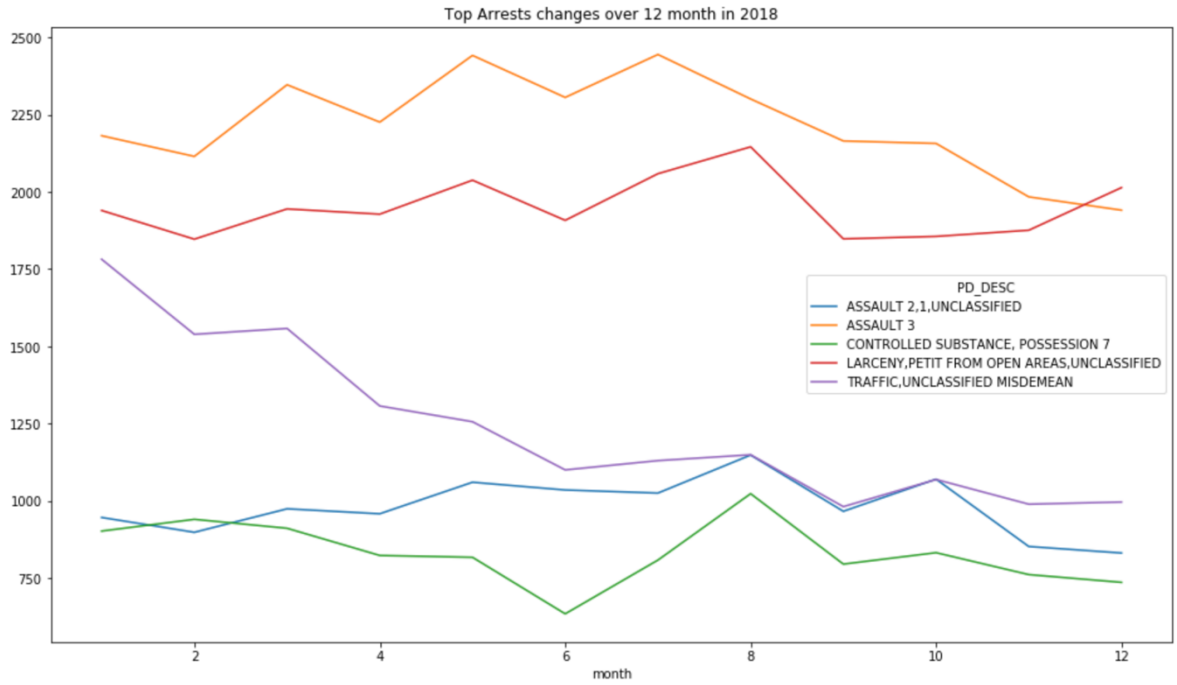

Figure 2: Top 5 arrest changes in each months in 2018

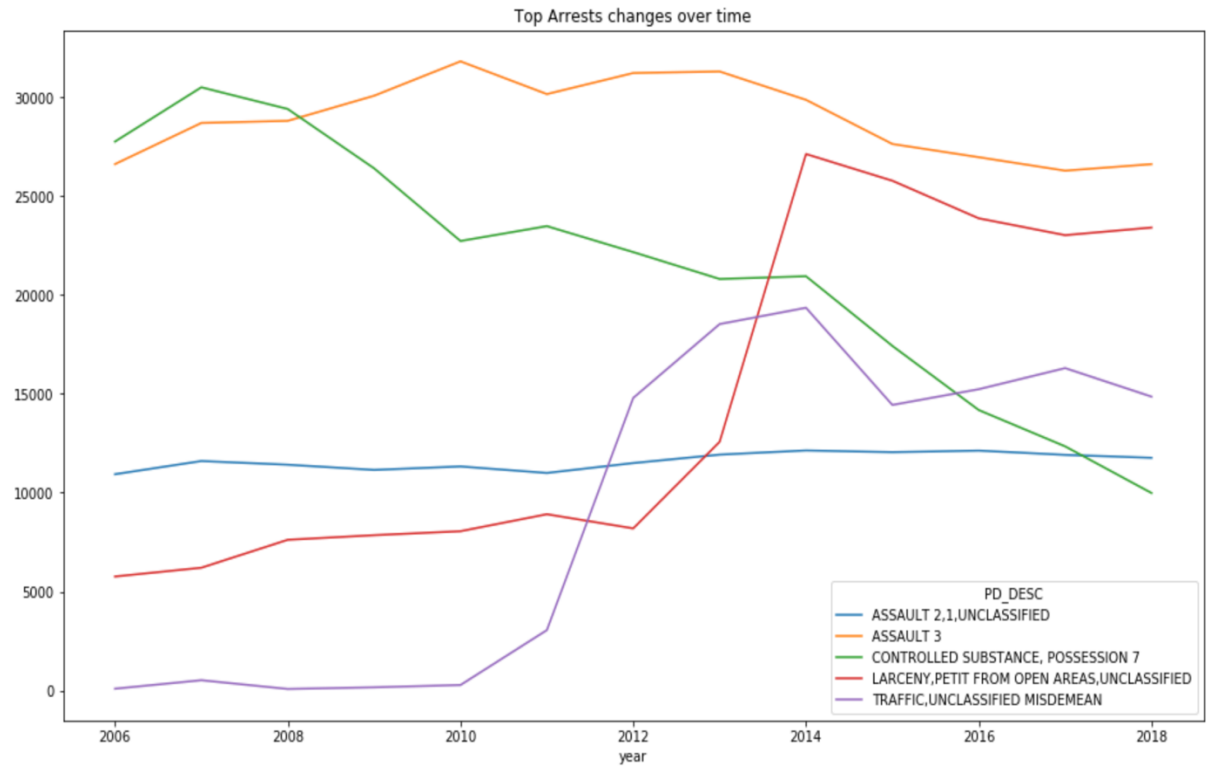

Figure 3: Top 5 arrest changes over time (2006 – 2018)

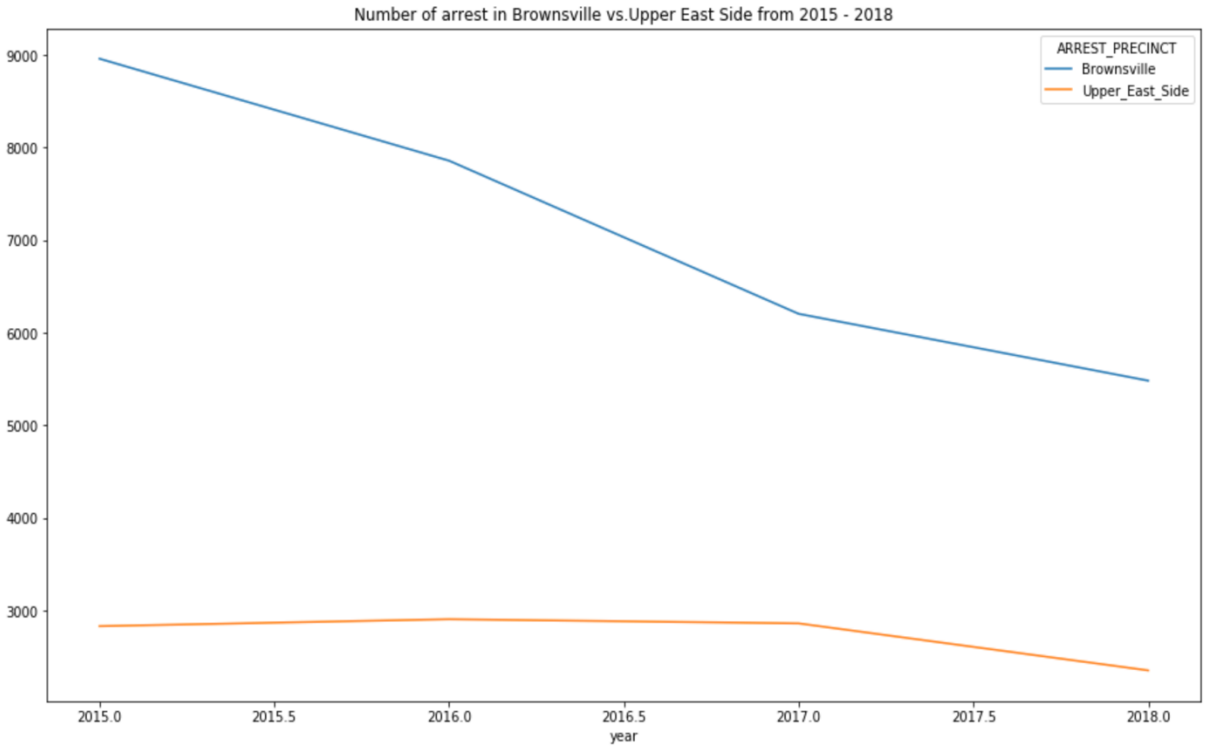

Figure 4: Number of arrest in Brownsville vs. Upper East Side from 2015 – 2018

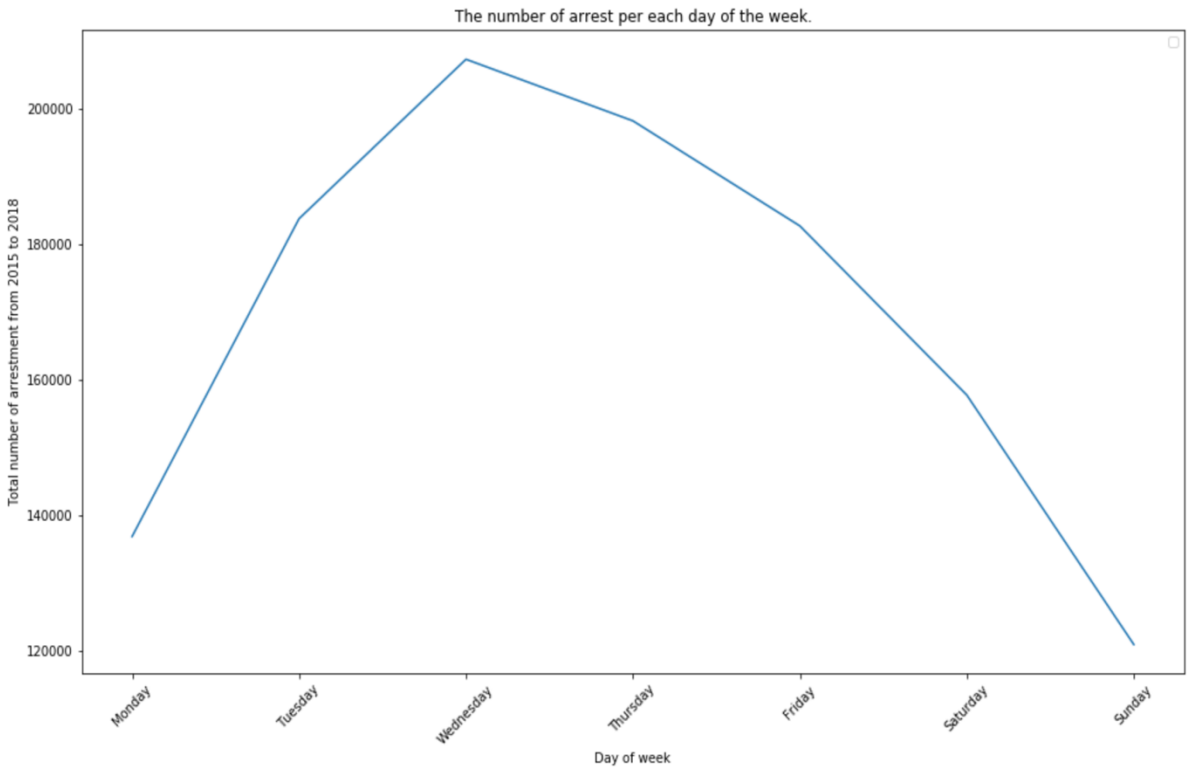

Figure 5: The number of arrest per each day of the week from 2015 – 2018

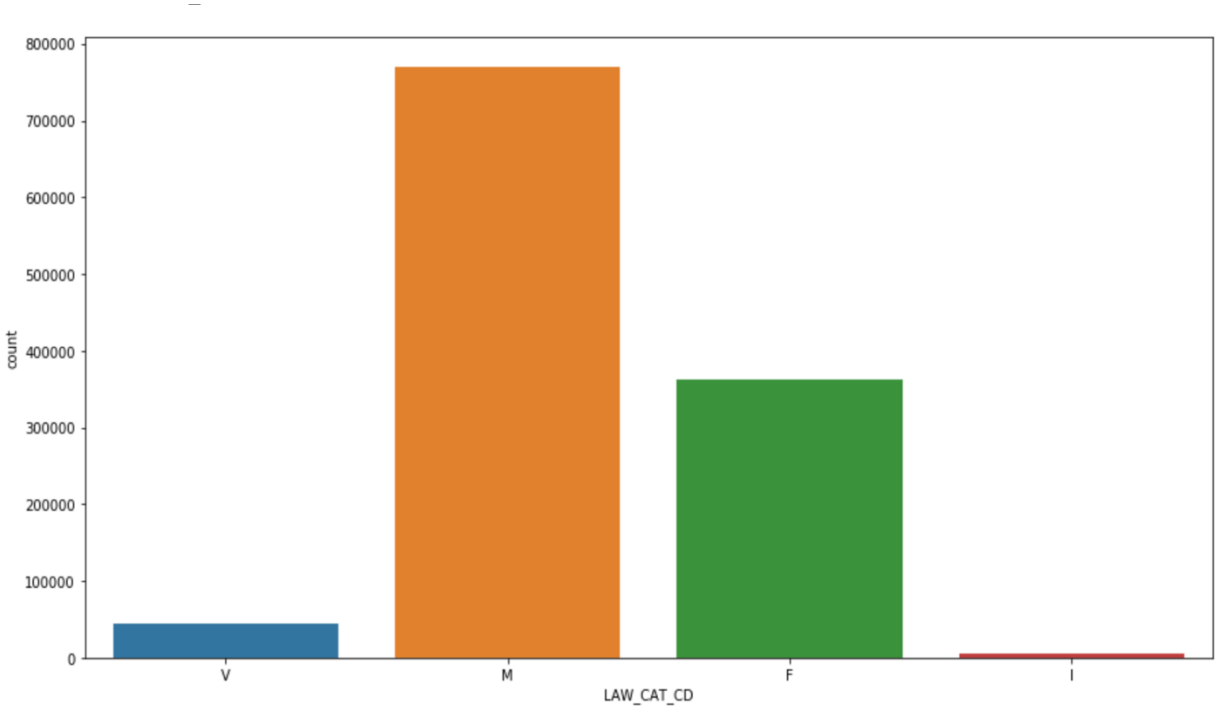

Figure 6: The number of arrest for the level of offense from 2015 – 2018

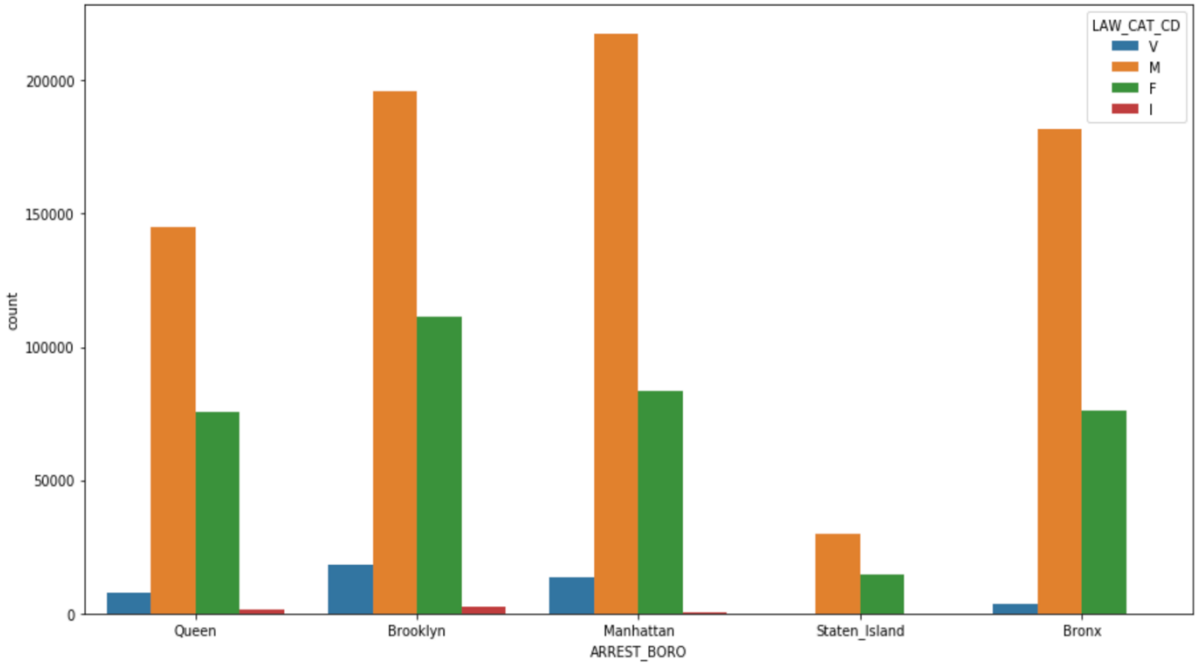

Figure 7: The number of arrest for the level of offense in each borough from 2015 – 2018

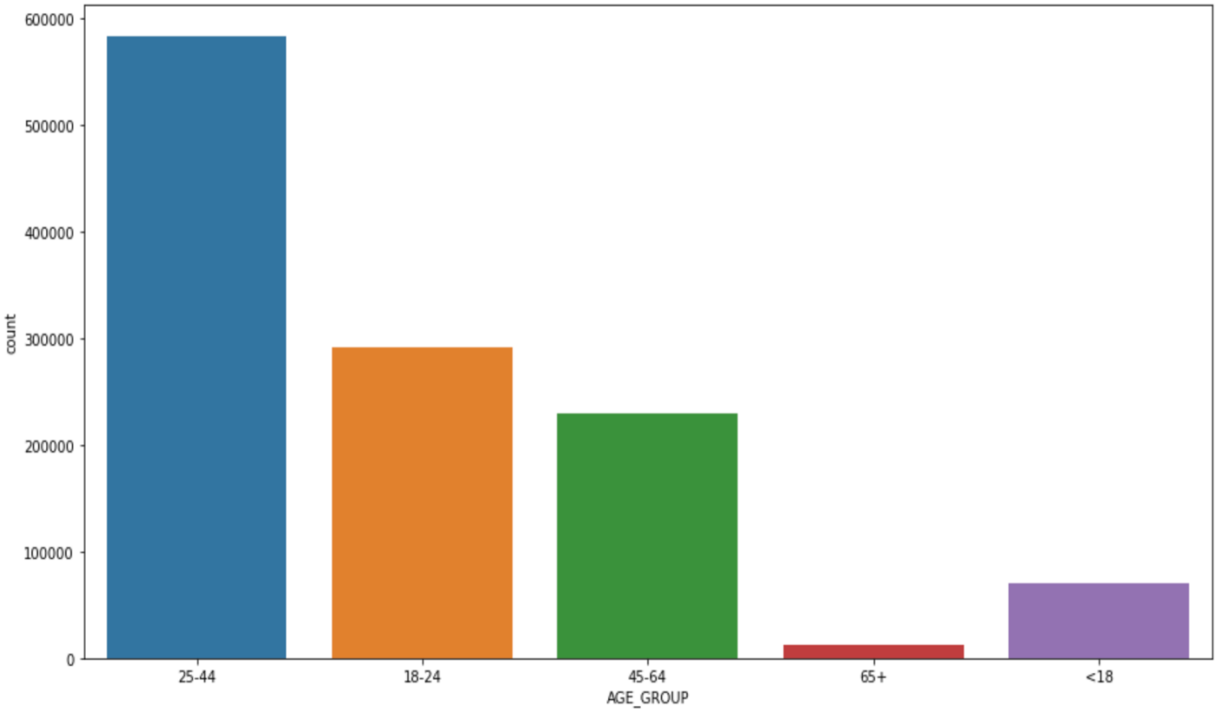

Figure 8: The number of arrest in each age groups