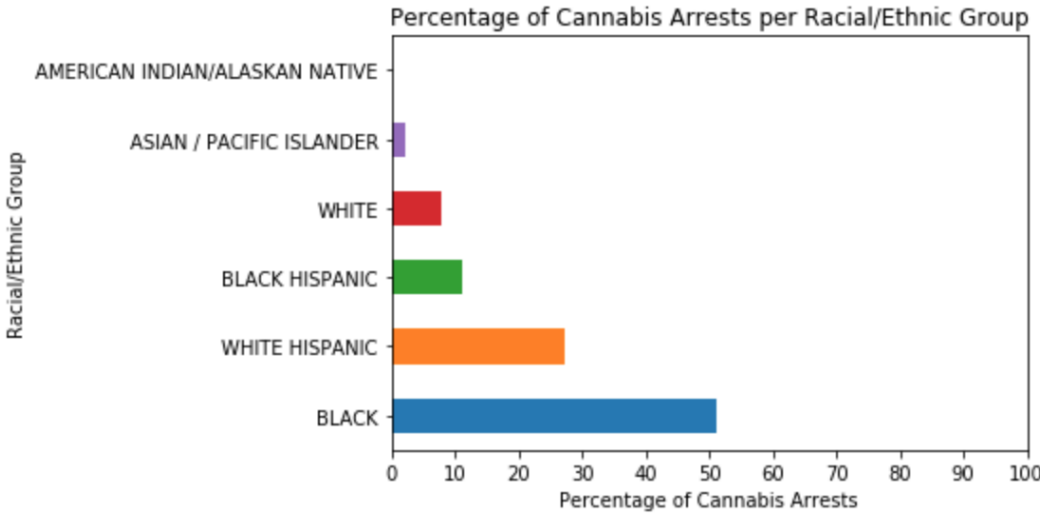
Project: Capstone Project 1: Data Story & EDA
Daniel Loew

A comprehensive exploratory data analysis (EDA) of the cannabis crime DataFrame with categorical features intact was conducted, which involved a series of visualizations designed to investigate the distribution of cannabis crime across demographic groups and geographic indicators. The combined cannabis and non-cannabis crime DataFrame with binarized categorical features for use in the machine learning classification models was also utilized to look for covariance and correlations between the feature set. These steps are detailed in the following Jupyter notebook:

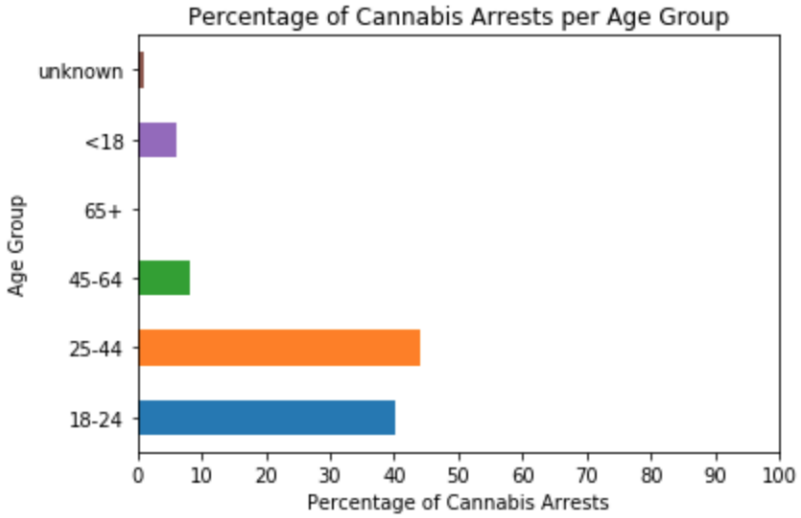
https://github.com/danloew/SpringboardFirstCapstone/blob/master/DataStory_EDA_Final.ipynb

In this exploratory data analysis (EDA) phase, the most important place to start is to look to see if this dataset from the NYPD corroborates the racial disparity in cannabis arrests reported elsewhere. However, only 34,837 cannabis cases (15.8%) have the crime suspect's race reported, which is unfortunate and begs the question as to how often the crime suspect's race is reported in non-cannabis crimes. As reported in the data cleaning notebook for this capstone project, 38.1% of non-cannabis crimes have the suspect's race reported. There is therefore a large difference between the percentage of cannabis crimes and non-cannabis crimes with the suspect's race reported, which will be the subject of a hypothesis test in the Statistical Methods section of this project to see if the difference is due to random chance.

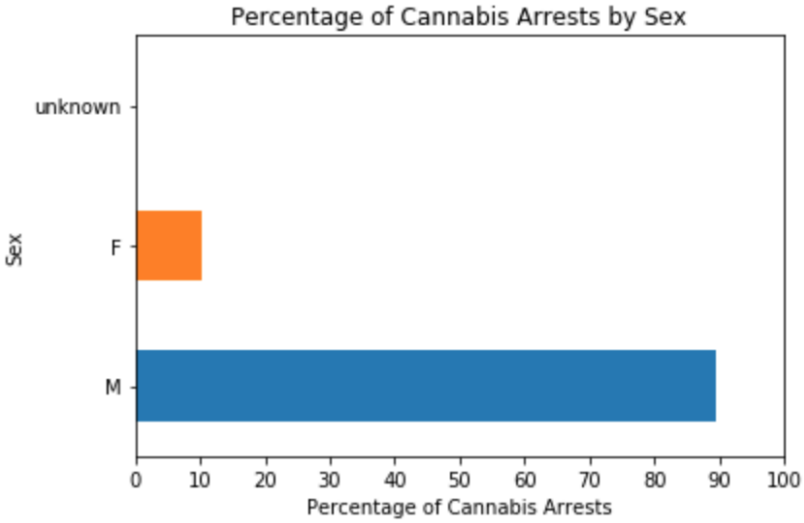
Although the cannabis crime suspects' race, sex and age data were very partial due to the NYPD's under-reporting, a DataFrame was loaded of just the cases where the suspects' race was reported in order to look at demographic distributions. Amongst the 34,837 cases in this DataFrame, 51% of cannabis arrests with the suspect's race reported were of African-Americans, 27% of white Hispanics, and 11% of black Hispanics, for a total of 89% of total cannabis crimes with the suspect's race reported being of African-American or Latino people. Only 8% of these arrests were of white people, and 2% were of Asian or Pacific Islander people. 0.2% were of American Indians or Alaskan Natives.



Looking just at age shows that 84% of cannabis arrests where the suspect's race was reported were made of people between the ages of 18-44.



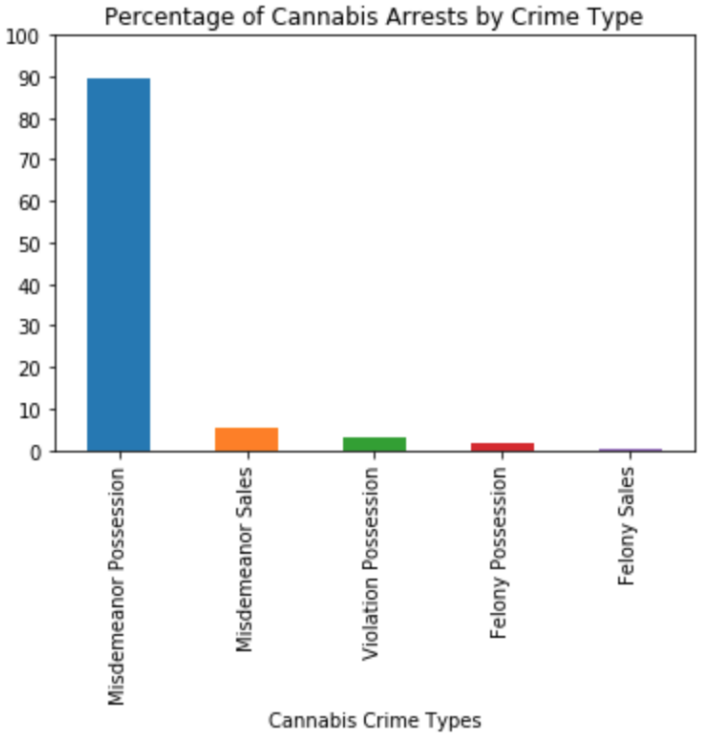
90% were of males.



As shown in the cross-tabulation in the notebook, 40% of cannabis arrests were of African-American males younger than 45, and 32% were of Hispanic/Latino males younger than 45, for a total of 72% of all cannabis arrests in New York City between 2006 and 2018 being of young African-American and Hispanic/Latino males. This corroborates the racial disparity data reported elsewhere. However, what was not reported elsewhere was the partial nature of the data on suspects' race for cannabis crimes. This will be partially addressed in the hypothesis tests run in the Statistical Data Analysis section below.

One of the striking things about cannabis arrests in New York City is that 92.6% of them are for simple misdemeanor and violation possession charges, which is the vast majority. 1.7% are for felony-level possession, 5.2% are for misdemeanor sales, and 0.5% are for felony sales, the latter being arguably the top priority if drug use prevention was the goal. It should be pointed out that

violations are less serious than misdemeanor charges, as they only involve fines and do not go on one's criminal record; violations have been the primary tool used in cannabis arrests after the recent decriminalization (New York State Penal Law).

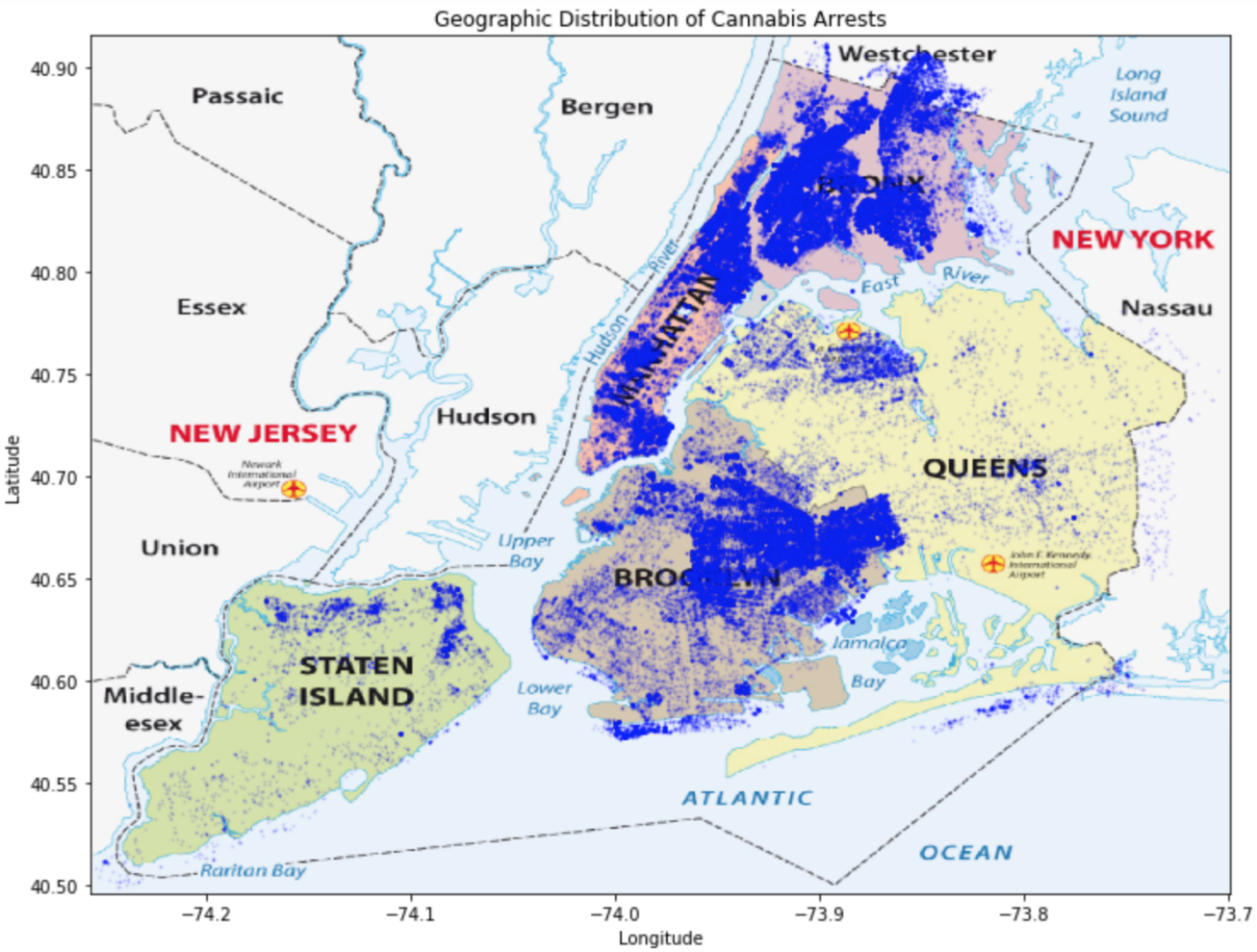


The same racial disparity described above holds true for all levels of cannabis crime except for the following differences. More violation possession arrests were made of white perpetrators than of black Hispanic perpetrators, but the difference was only 3%. Also, it should be noted that violation possession charges are the lowest level of cannabis arrests, and that the majority of violation possession charges were still of African-Americans and white Hispanics. More whites were arrested for felony possession charges than black Hispanics and the same amount of whites were arrested for felony sales charges as black Hispanics, but the difference was less than a percentage point and it bears mentioning that the sample sizes for these groups was very small.

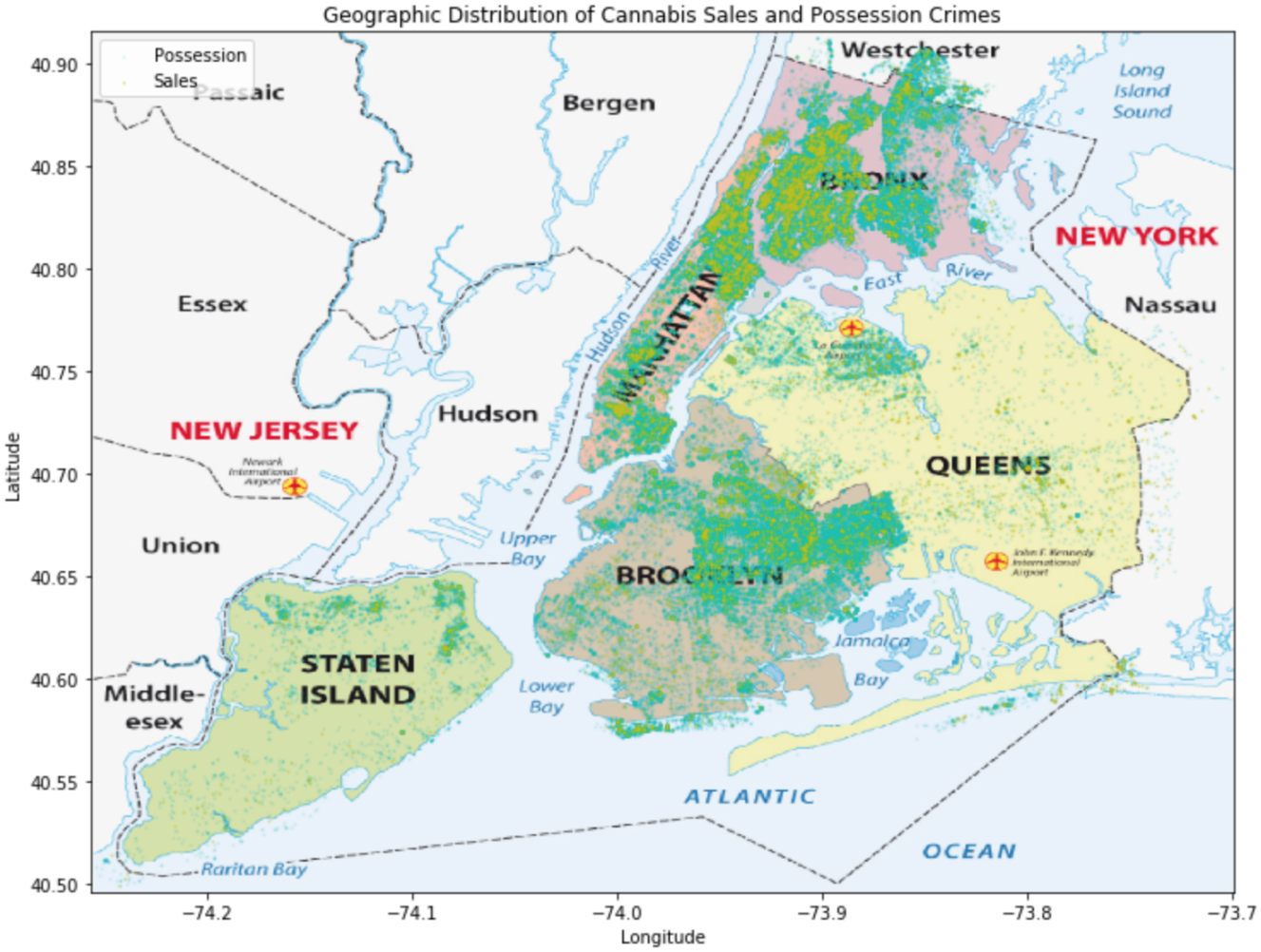
To look at other indicators of bias in cannabis arrests in New York City, five DataFrames were first made from the full DataFrame of all cannabis crimes (not just those with suspect race reported), one for each of the five cannabis crime levels. Scatter plots were created based on latitude and longitude of crime occurrence, which helps to illustrate the geographic distribution of the five types of cannabis arrests. Because there are references available with demographic information of the various parts of New York City, the visual concentration of arrests in certain parts of the city enable us to partially infer the race of cannabis crime suspects in the overall DataFrame, where only 16% of cases have suspect race reported.

The following visualizations show the geographic distribution of 1) cannabis arrests as a whole, 2) cannabis possession and sales arrests, 3) arrests for cannabis possession and its three types, and 4) arrests for cannabis sales and its two types. Please note that the overlay of the latitude/longitude

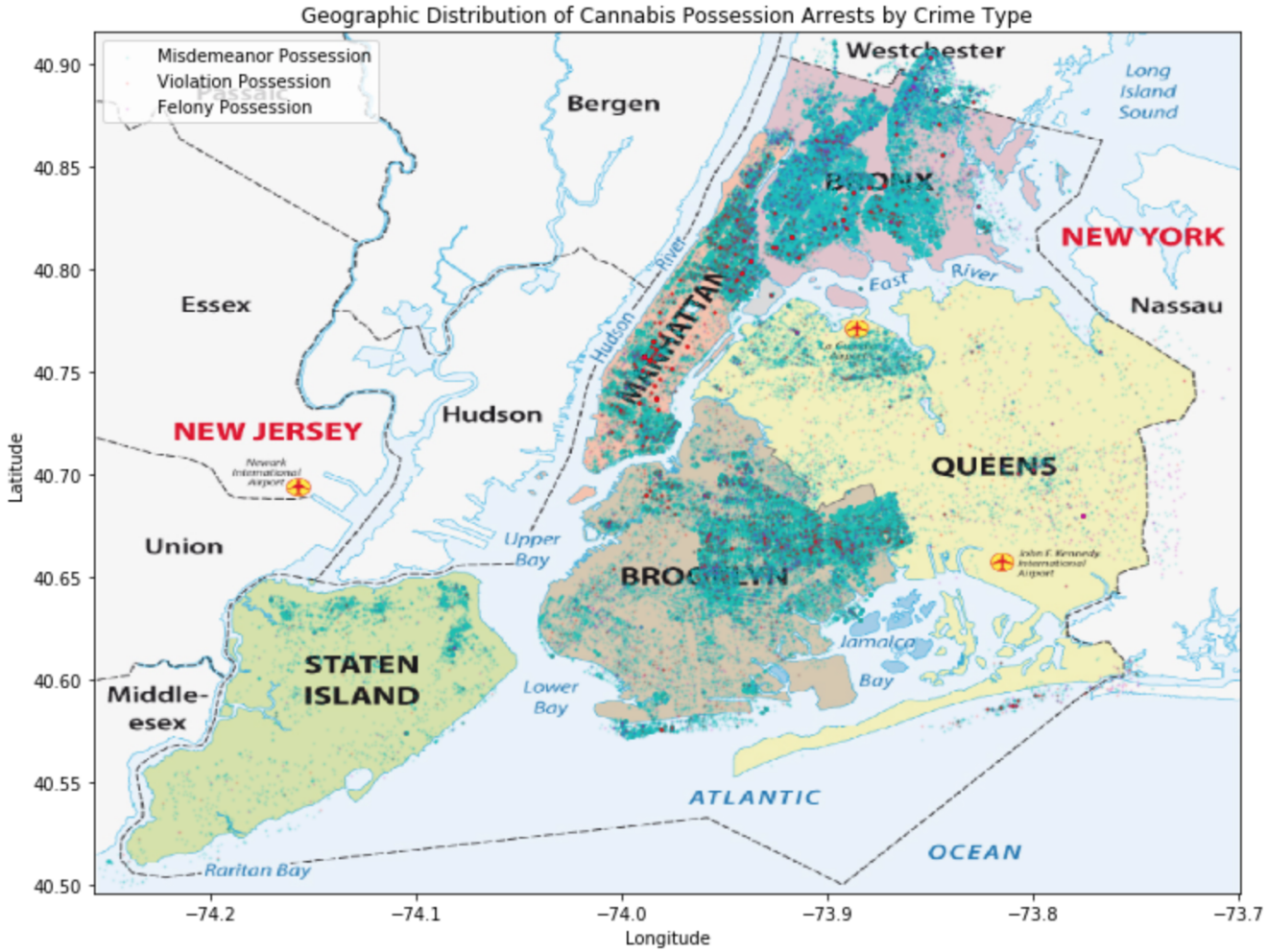
coordinates of each arrest are slightly warped in relation to the map image of NYC's five boroughs due to the curvature of the earth.



As can be seen, arrests were greatly concentrated in Manhattan, the Bronx, an area of Queens around LaGuardia Airport, northern Staten Island, northern Brooklyn, eastern Brooklyn, central Brooklyn, and to a lesser degree southern Brooklyn.

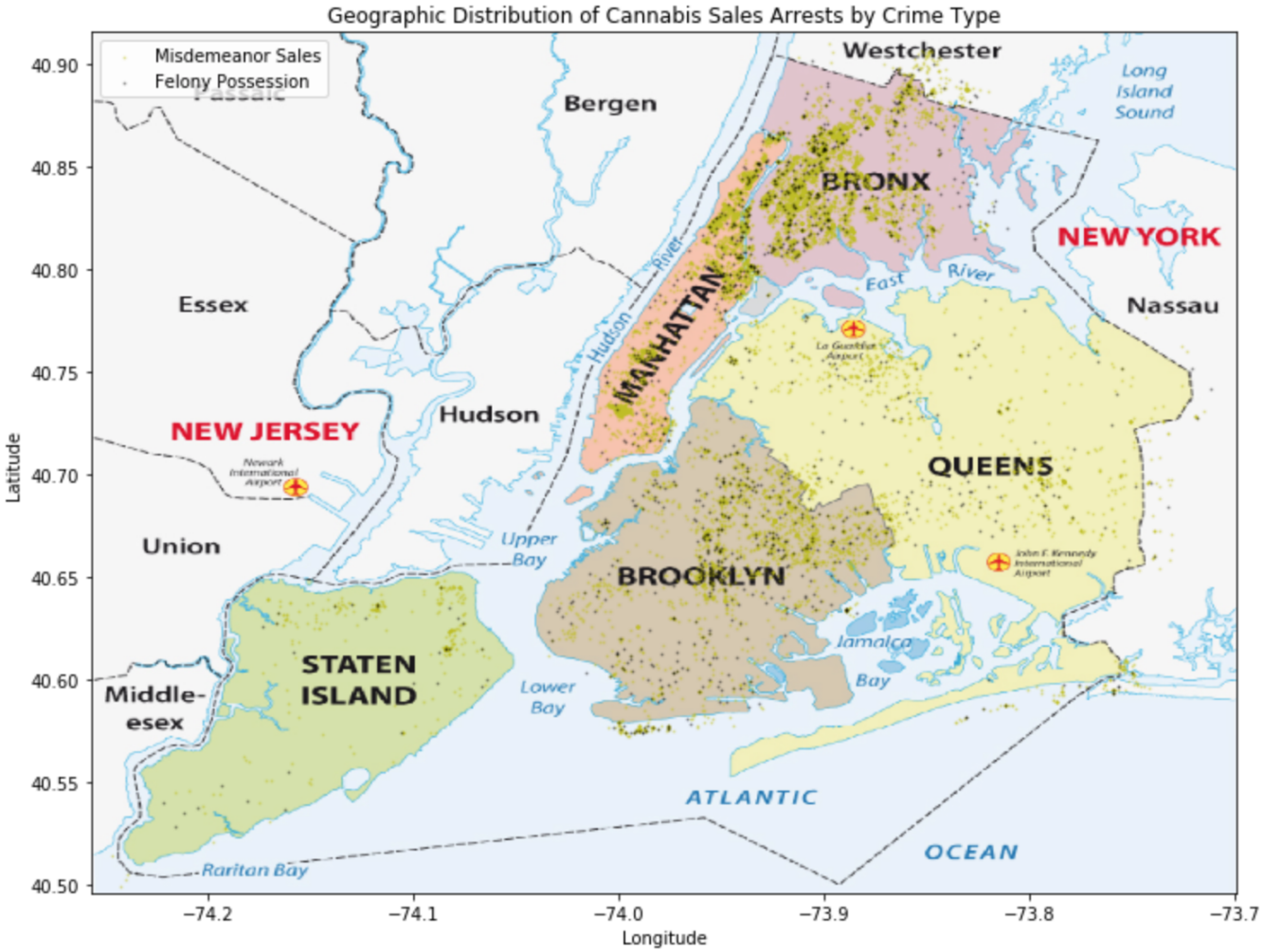


Possession crimes are plotted in cyan, and sales crimes are plotted in yellow. It can be seen that sales arrests are largely concentrated in uptown Manhattan and the Bronx, with pockets around the transit hub of Midtown, the West Village, and scattered points in Central and Eastern Brooklyn.



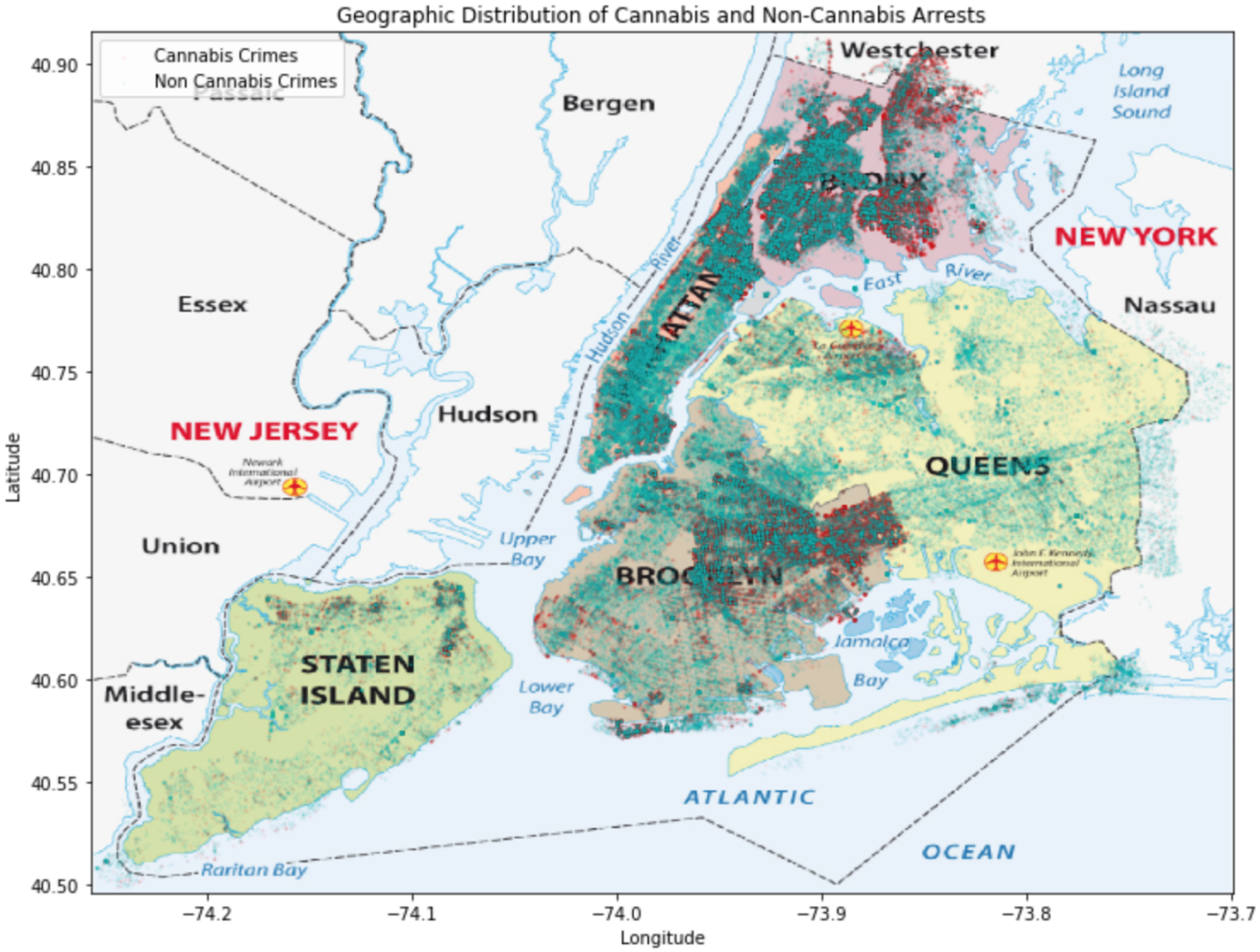
Misdemeanor possession arrests are displayed in cyan, violation possession arrests are displayed in red, and felony possession arrests are displayed in magenta.

As can be seen above, the vast majority of cannabis possession arrests are for misdemeanor possession, and are heavily concentrated in the Bronx, Inwood, Washington Heights, and Harlem, which have large populations of African-American and Latino residents. In Brooklyn, arrests are concentrated in neighborhoods like East New York, Cypress Hills, Brownsville, Crown Heights, Flatbush, Bedford-Stuyvesant, and Bushwick. Again, these neighborhoods have large populations of African-American and Latino residents. Violation and felony possession are peppered throughout, but they are concentrated in the neighborhoods already mentioned. Manhattan, Queens, south and west Brooklyn, and Staten Island have significantly fewer arrests. It bears mentioning that Staten Island is majority white, and the clusters of arrests there are centered around housing projects like Stapleton and Park Hill.



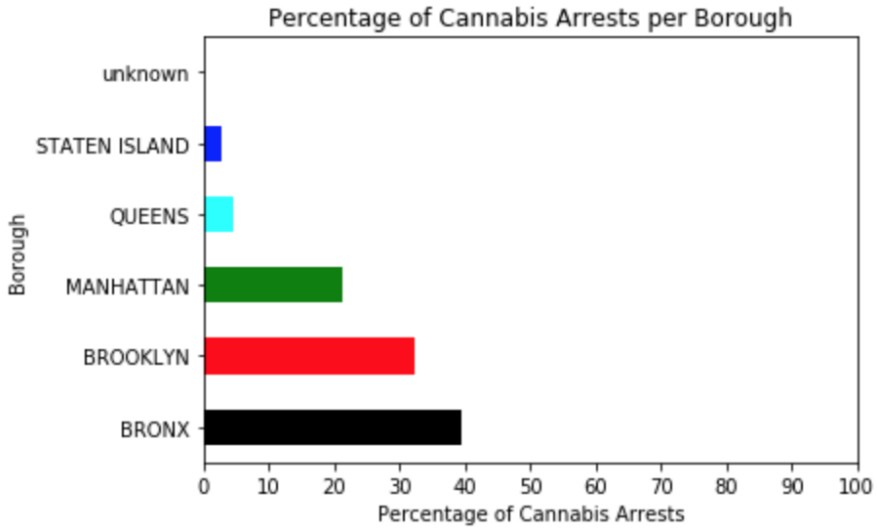
Misdemeanor sales arrests are displayed in yellow and felony sales arrests are displayed in black. One can see that these arrests tend to fall within the same neighborhoods as possession arrests, but are obviously much more sparse.

In order to visualize the locations of cannabis arrests and non-cannabis arrests on the same plot, the EDA versions of the cannabis and non-cannabis crime DataFrames were first concatenated and a 'cannabis_crime' flag feature was added to differentiate cannabis crimes from non-cannabis crimes. As a reminder, the EDA version of the non-cannabis DataFrame is a sample of all non-cannabis crimes with the same sample size as the universe of cannabis crimes (n=220,304).



All cannabis and non-cannabis crimes are displayed above on the same scatterplot. Cannabis crimes are plotted in red, and non-cannabis crimes are plotted in cyan. It can be seen clearly that against the background of all other crimes, cannabis crimes are largely concentrated in the Bronx (especially around Throggs Neck and the East Bronx) and Central Brooklyn (especially around East New York, Flatbush, and Brownsville).

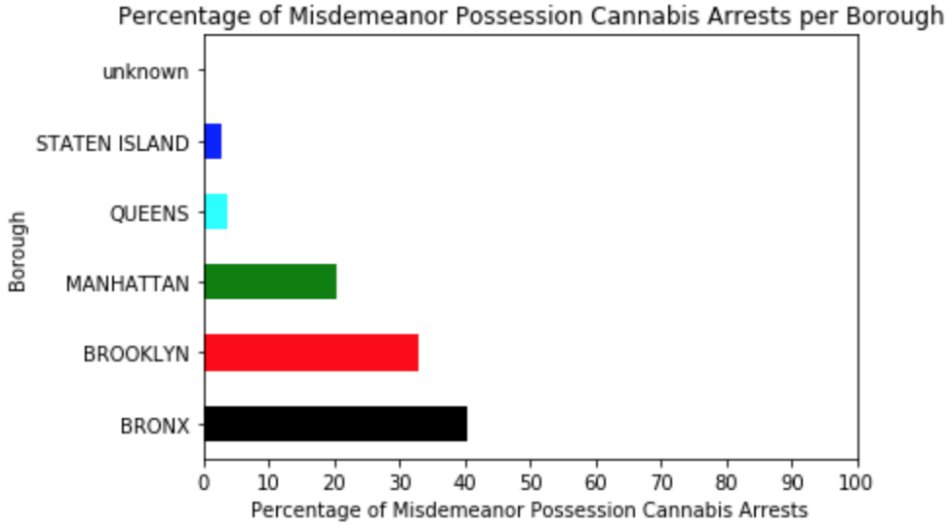
The most abstract geographic unit of New York City is the borough. The Bronx and Brooklyn are home to the majority of cannabis crimes overall, as shown in the following bar chart.

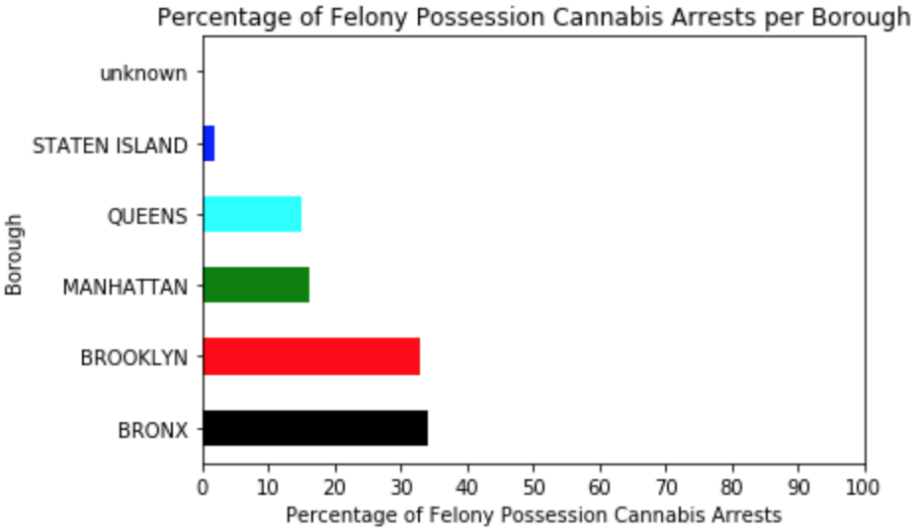


This clear disparity in cannabis arrests by borough is interesting because of the racial demographics of these two boroughs. Looking at the Census data estimated from 2018 and 2019 shows us the following racial and ethnic breakdown by borough. It is important to remember that Hispanic/Latino status is considered as an ethnicity, and people of any racial group (Black or African-American, White, Asian, Native Hawaiian/Pacific Islander, and American Indian/Alaskan Native) can also be of Hispanic/Latino ethnicity.

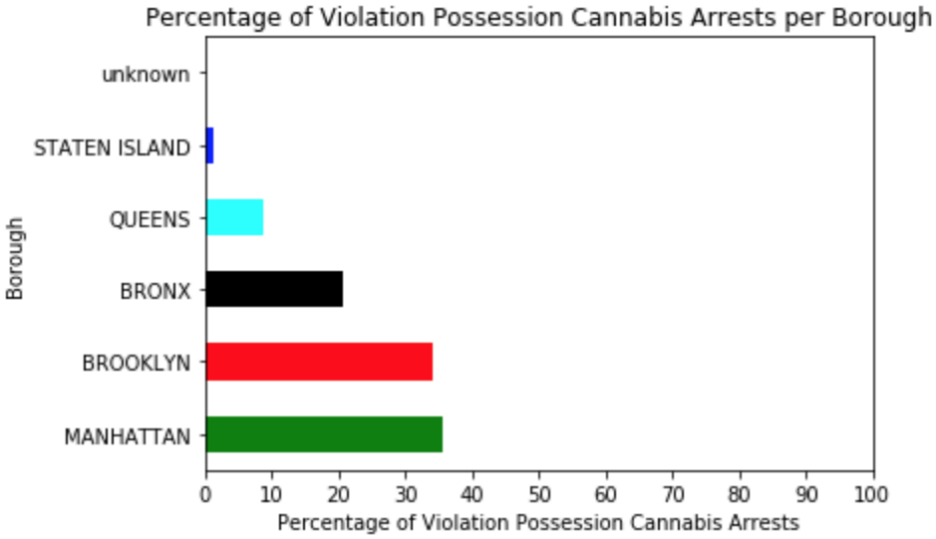
The Bronx's populace is 44% African-American, 56% Latino/Hispanic, 5% Asian and only 9% non-Latino White, while Brooklyn's populace is 34% African-American, 19% Latino/Hispanic, 13% Asian, and 36% non-Latino White. By contrast, Manhattan's populace is 18% African-American, 26% Latino/Hispanic, 13% Asian, and 47% non-Latino white. Queens is 21% African-American, 28% Latino/Hispanic, 25% non-Latino white, and 27% Asian. Staten Island is 12% African-American, 19% Latino/Hispanic, and 60% non-Latino white ("Census Bureau Quick Facts on New York City"). So even though suspect race is only reported 16% of the time for cannabis arrests, we can see that the majority of cannabis arrests are made in boroughs with high concentrations of African-Americans and Latinos/Hispanics. This racial/geographic data is only suggestive and will be further explored in the machine learning classification models.

Misdemeanor and felony possession charges are dominant in the Bronx and Brooklyn.

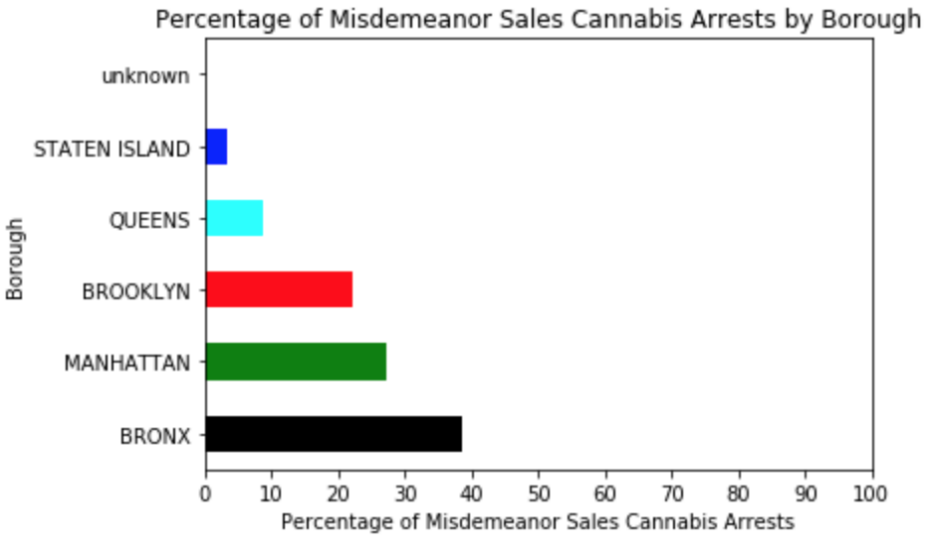


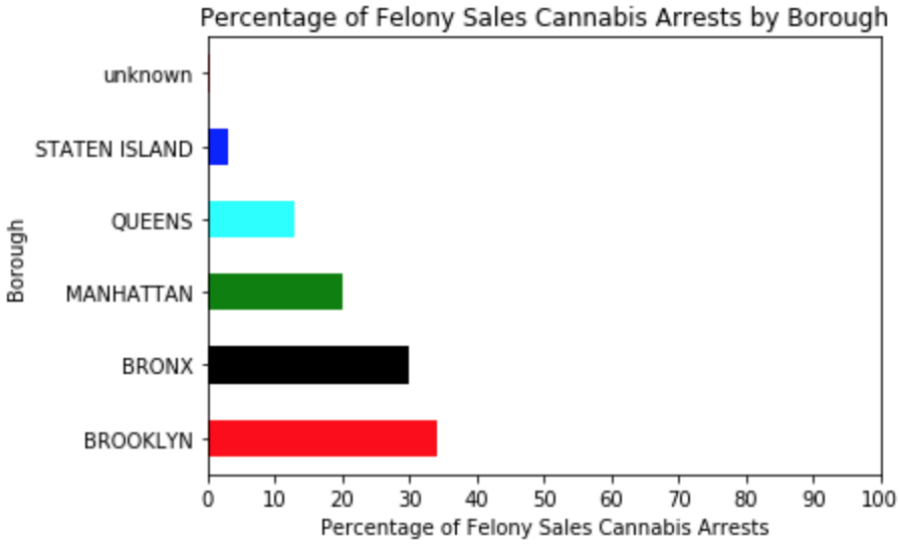


Violation possession charges (the lowest level of cannabis crime) are dominant in Manhattan. This supports the idea that cannabis crimes are punished very differently in New York City depending on which part of the city the crime takes place in.



Interestingly, Manhattan is second to the Bronx for misdemeanor sales arrests, and Brooklyn and the Bronx predominate for felony sales.





esponsible for these differences. Police precincts offer a route to explore these smaller geographic zones. The top 10 police precincts with the highest amounts of misdemeanor cannabis arrests and cannabis arrests overall are all in the Bronx and Brooklyn. The demographics in these neighborhoods reflects the racial disparity seen in cannabis arrests.

The precincts with the most violation possession charges differ however, being largely in Midtown Manhattan and to a lesser degree in Central Brooklyn.

Jamaica (in Queens), Washington Heights (in Manhattan), and Inwood (northernmost Manhattan) are also in the list of police precincts where the most felony possession charges are made. All of these neighborhoods have a predominantly African-American and Latino population.

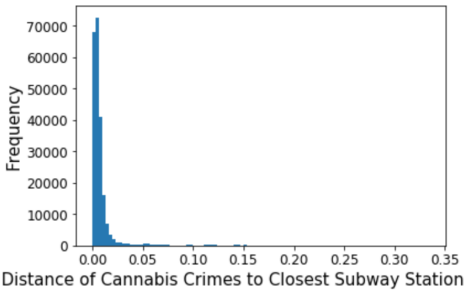
For misdemeanor sales, Greenwich Village and the West Village (in Manhattan), and Western Harlem are also common.

The Bedford-Stuyvesant neighborhood of Brooklyn and East Harlem also show up on the top 10 list of police precincts with the highest concentration of felony sales arrests. Again, both of these neighborhoods have a predominantly African-American and Latino population.

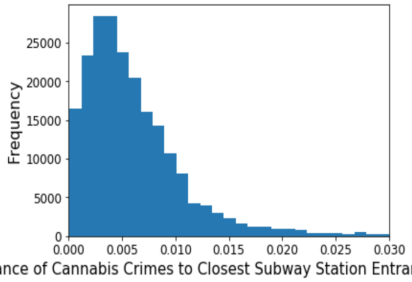
In the data cleaning notebooks, the latitude/longitude distance from each crime to the nearest NYC subway station entrance was calculated as a contributive predictor to the feature sets used in the classification models, as subway station entrances can often be where crime can occur. As a visual exploration, these distances are displayed in histograms below for cannabis crimes generally, as well as for misdemeanor possession, violation possession, felony possession, possession generally, misdemeanor sales, felony sales, and sales generally. Because New York City has so many subway station entrances, it can easily be seen that most cannabis crimes are very close to a subway station entrance.

For each crime category, the histograms were run once without any range limits for distance, and then again with a range limit for distance of 0.03 latitude/longitude units that helps illustrate differences in greater detail. The 'L1' distance is used instead of the 'L2' distance because

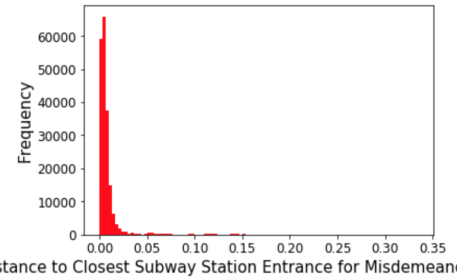
realistically, New Yorkers can't move through the city and engage in cannabis crime "as the crow flies".



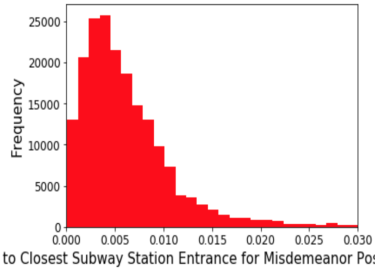
L1 Distance of Cannabis Crimes to Closest Subway Station Entrance



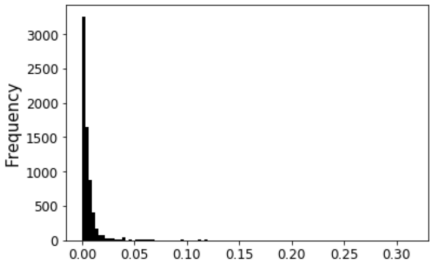
L1 Distance of Cannabis Crimes to Closest Subway Station Entrance magnified



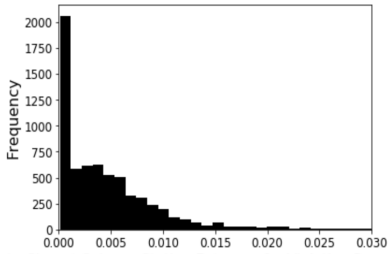
L1 Distance to Closest Subway Station Entrance for Misdemeanor Possession



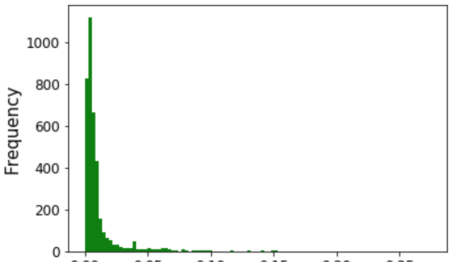
L1 Distance to Closest Subway Station Entrance for Misdemeanor Possession, magnified



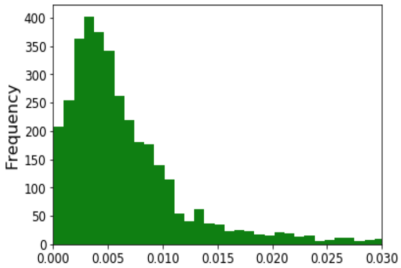
L1 Distance to Closest Subway Station Entrance for Violation Possession



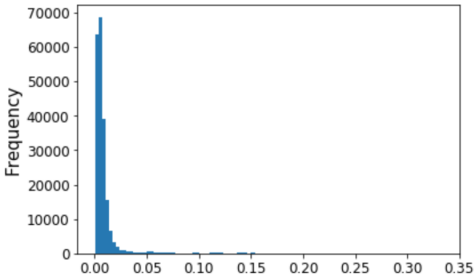
L1 Distance to Closest Subway Station Entrance for Violation Possession, magnified



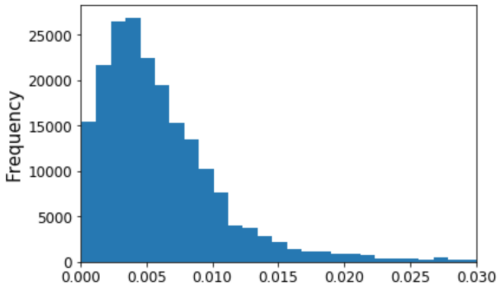
L1 Distance to Closest Subway Station Entrance for Felony Possession



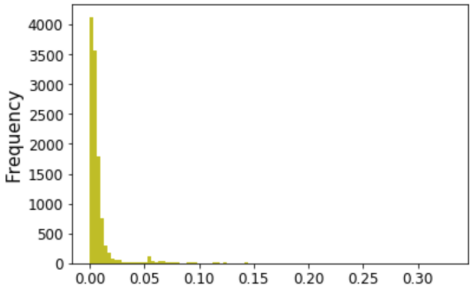
L1 Distance to Closest Subway Station Entrance for Felony Possession, magnified

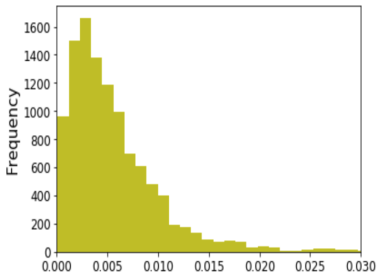


L1 Distance to Closest Subway Station Entrance for Possession

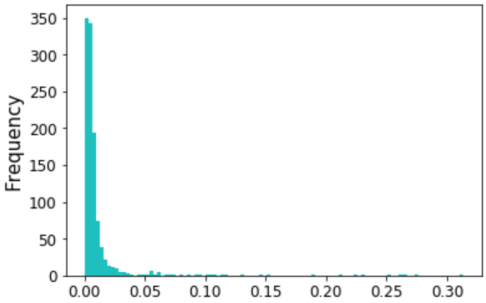


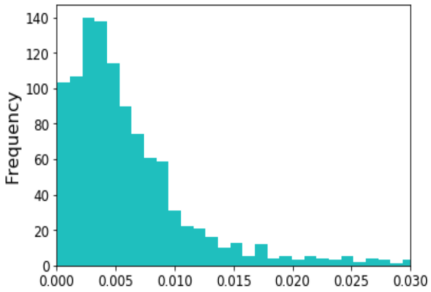
L1 Distance to Closest Subway Station Entrance for Possession, magnified



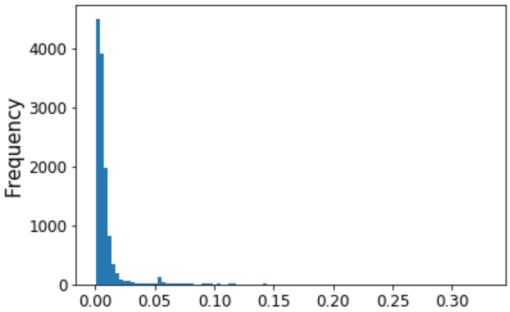


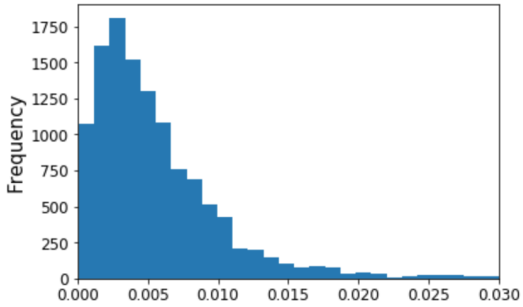
L1 Distance to Closest Subway Station Entrance for Misdemeanor Sales L1 Distance to Closest Subway Station Entrance for Misdemeanor Sales, magnified





L1 Distance to Closest Subway Station Entrance for Felony Sales L1 Distance to Closest Subway Station Entrance for Felony Sales, magnified





L1 Distance to Closest Subway Station Entrance for Sales L1 Distance to Closest Subway Station Entrance for Sales, magnified

As can be seen above, the shape of these histograms were generally very similar, with the exception of violation possession crimes. The majority of these crimes were within 0.001 latitude/longitude units, with a steep drop-off. This shows a higher proportion of low-level violation possession charges brought against people very close to subway entrances.

By looking at the feature describing the premises type that the arrest occurred in ('PREM_TYP_DESC'), it can be seen that the majority of cannabis arrests happened either on the street (58% of all cannabis arrests) or in the New York City public housing projects (19%). Less frequently, cannabis arrests were made in residential apartment houses (8%) and parks and playgrounds (6%). This pattern generally repeats itself through the five cannabis crime types, with the exception of violation possession arrests. 32% of these arrests occurred in the New York City subway system, which reflects the data reported above concerning the distance of crime to the closest subway entrance. Notably, a comparable percentage of misdemeanor and felony possession arrests occur on the street (58% and 61% respectively), while much more misdemeanor possession charges occurred in public housing projects than felony possession charges (20% and 7%, respectively).

As can be expected, the jurisdiction responsible for the majority of cannabis arrests between 2006 and 2018 was the NYPD, the New York City Housing Authority (NYCHA), and to a much lesser degree the N.Y. Transit Police, with 33% of violation possession arrests falling under this jurisdiction.

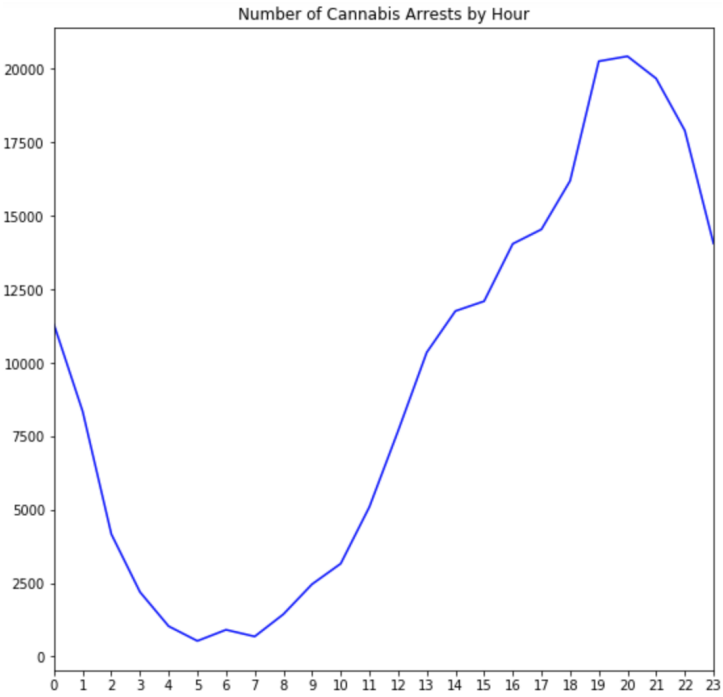
The fact that 19% of cannabis arrests fell under the jurisdiction of the NYCHA shows how heavily policed these public housing projects were.

Because of the fact that nearly 20% of all cannabis arrests occurred in NYC housing projects, it makes sense to look at the 'HADEVELOPT' feature, which tells which housing project cannabis arrests occurred in. Because there were so many unknown values in this feature (as roughly 80% of cannabis arrests occurred outside of N.Y. housing projects), it made sense for reporting purposes to first re-base the feature by removing the unknown values. After doing so, it was shown that the top 10 NYC housing developments with the highest proportion of cannabis arrests were all in the South Bronx or in economically disadvantaged areas of Brooklyn. When looking at violation possession charges specifically, Pink House, Hammel House, and Farragut House suddenly jumped into the top 10. Pink House is in the East New York neighborhood of Brooklyn, Hammel House is in the Rockaway Beach neighborhood of Brooklyn, and Farragut is in the Vinegar Hill neighborhood of Brooklyn. For felony possession and sales, Red Hook West in far western Brooklyn was the housing development project with the most cases.

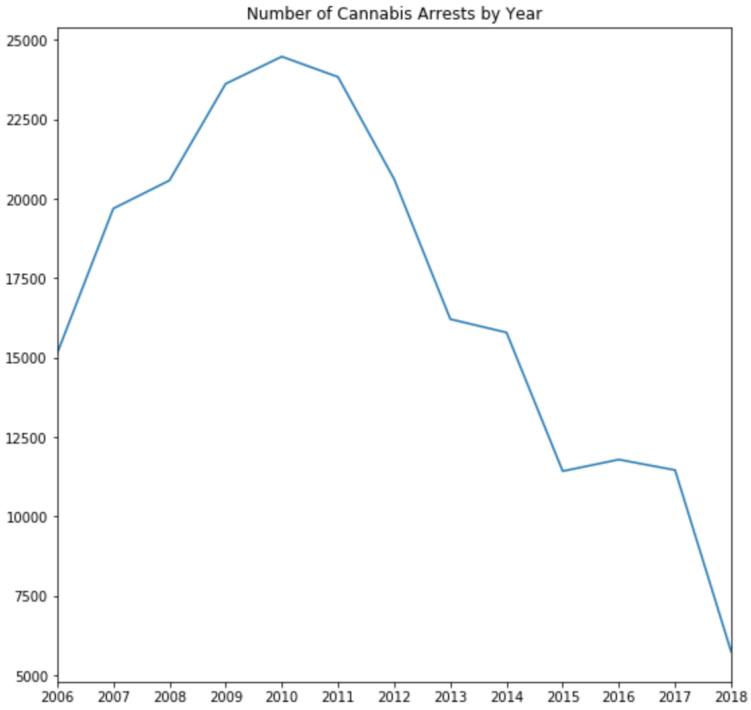
Cannabis arrests occurred more frequently during certain times of the day. 39% occur during the daytime (6 am - 6 pm), and 61% occur during the nighttime (6 pm - 6 am). The work day (9 am - 6 pm) composes most of the daytime arrests, and 37.5% of the total. Early morning (6 am - 7:30 am) and the morning rush hour (7:30 am - 9 am) have very little arrests (0.6% and 0.9 respectively), but this picks up during the lunch hour (12-1 pm), when 3.9% of the arrests are made. The long New York metropolitan area's evening rush hour (4:30 pm - 7 pm) straddles the daytime (6 am - 6 pm) and nighttime (6 pm - 6 am) windows, and one sees a fairly high concentration of arrests happening during this time window.

The nighttime saw the majority of cannabis arrests, at 61%. Overlapping with the evening rush hour, the dinner window of 6-8 pm had a high concentration of arrests for just a two hour window, and had nearly as many arrests that occurred in the 2.5 hour window of the evening rush hour. Evening (8-10 pm) had a similarly high concentration of arrests at 19% for a two hour window. Late night (10 pm - 6 am) had 26% of the arrests for an 8 hour window, showing that more than half of the nighttime arrests did not happen during the nightlife hours, but after work and before the working population would typically go to bed.

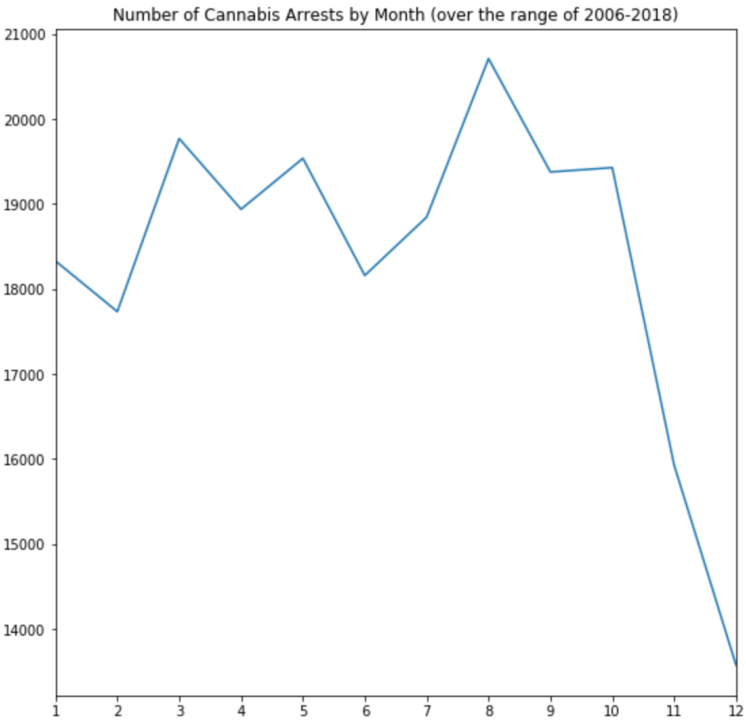
To easily visualize the number of arrests made per hour (across all years), the hour of the day is extracted and then displayed in the following line plot (using the 24 hour convention with ‘0’ indicating midnight).



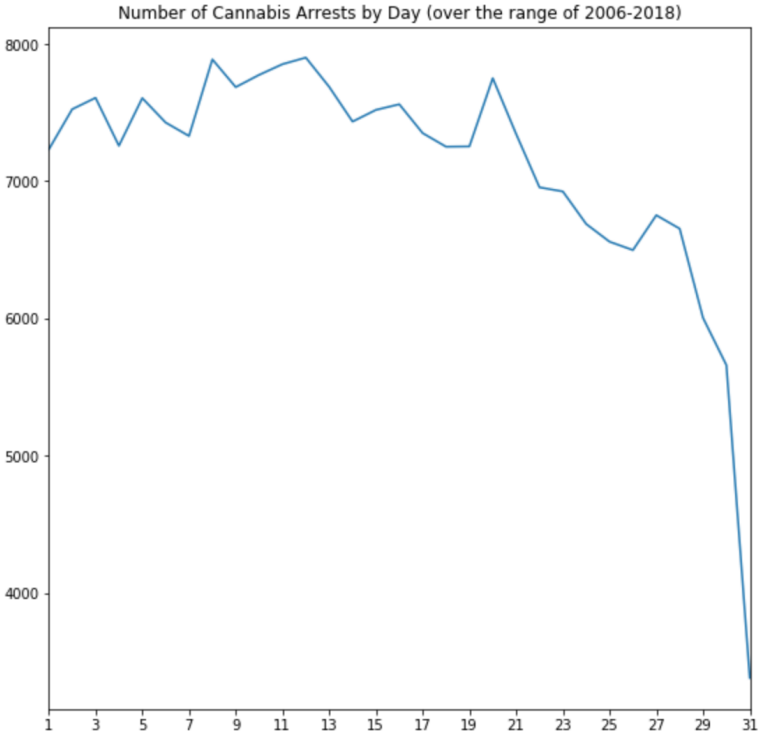
It has been well reported that during Mayor Bloomberg's time as mayor, cannabis arrests reached their peak. One can see that 2006 had 15,127 arrests, and that this increased to 24,468 arrests in 2010. This held fairly steady for 2011 (23,827), dropped a bit in 2012 (20,611) as criticism of Bloomberg's "stop and frisk" program mounted, and then dropped significantly in 2013 (16,206) when the "stop and frisk" program was judged as unconstitutional by Judge Scheindlin (Goldstein, NY Times, 2013). Mayor DeBlasio, who vowed to reverse the program, took office in 2014, but cannabis arrests remained fairly consistent in that year compared to 2013 (15,787). By 2015, the number was still fairly high but dropped significantly (11,424). This number stayed consistent through 2017, and then dropped by half in 2018 as discussions of cannabis legalization in New York intensified.



Each month of the year had about the same amount of cannabis arrests, but August had the highest number and the number dropped in November and December during the Holiday season.

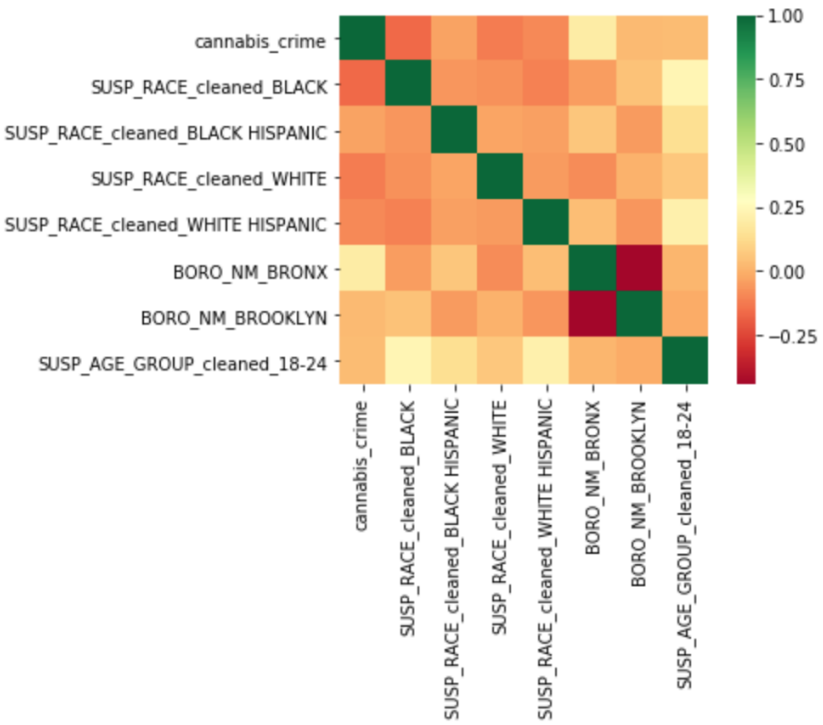


Each day of the month over the entire range of 2006 to 2018 had a fairly consistent number of cannabis arrests, ranging from 5,660 to 7,900 arrests a day . The number dropped somewhat in the last 10 days of the month. The 31st had roughly half the arrests as the rest of the month because not every month has 31 days.



Because of the importance of holidays to various cultural groups, and because of the differences in how certain groups of people are arrested for cannabis, it made sense to look at whether certain holidays had higher concentrations of cannabis arrests across the full year range of 2006 to 2018. Due to the cultural diversity of New York City, certain holidays were included that would not be typically celebrated in other parts of the United States. Intriguingly, the holidays with the highest number of cannabis arrests were Hindu, Jewish, and Muslim holidays. Diwali had 656 arrests, Yom Kippur had 707, Rosh Hashanah had 677, Eid al-Fitr had 644, and Eid al-Adha had 544. Inexplicably, Valentine's Day had 531 arrests. St. Patrick's Day also had a high number at 542, which may be due to co-occurring cannabis use that happens during the large amount of public drunkenness that occurs on New York City streets on that day. April 20th had the highest number of arrests, probably due to its cultural connection to cannabis.

Covariance matrices and correlation coefficients showing relationships between specific features of the feature set and the cannabis crime flag were created. First, a heatmap of several features of interest is displayed below. Intriguingly, there were not really any clear and strong correlations between racial/ethnic groups and boroughs that have the highest concentration of cannabis arrests. This hints at multi-componential interactions between the overall feature set and cannabis crime, which is more fully explored in the classification machine learning models.



The Pearson's R and covariance matrix was run repeatedly to try and identify strong correlations between the 'cannabis_crime' feature and the rest of the feature set. Intriguingly, there were not really any strong correlations. The only thing that stood out was the 0.59 covariance between the hour of the day and the 'cannabis_crime' feature. The lack of strong correlations and covariance points to the need of developing a series of strong machine learning models that can help provide a gestalt picture of the many features that differentiate cannabis crimes from all other crimes, cannabis possession from cannabis sales crimes, and the five legal levels of cannabis crime in New York City between 2006 and 2018.