Project: Capstone Project 1: Milestone Report
Daniel Loew

SPECIAL NOTE: This milestone report was a requirement of the Springboard curriculum, but it needs to be noted that it is not a final document and reflects the state of the project halfway through completion. It is kept on GitHub for historical purposes. The links to Jupyter notebooks lead to the *not final* versions that are stored on the Springboard Data Science Career Track repository and not the repository for the capstone project itself, titled 'Springboard-First-Capstone'.

Problem Statement:

It has been reported that there is a great racial disparity in cannabis arrests in New York City, which is a feature of the larger problem that minority groups in America have long borne the greatest negative impact of the Drug War. It has been reported that 9 out of 10 cannabis arrests made in New York City are of African-Americans and Latinos, even though the Substance Abuse and Mental Health Services Administration (SAMHSA, a branch of the U.S. Department of Health and Human Services) consistently reports in their surveys that people of different racial and ethnic groups use cannabis at roughly the same rates.

As cannabis legalization becomes an increasing reality across the country, civil rights groups like the ACLU and historians of civil rights could benefit from an analysis that takes an impartial look at several predictors of cannabis arrests to see if suspect race is truly the main predictive factor of who gets arrested for cannabis, or if there are other predictors that play a supportive or central role. This analysis could serve as evidence for civil rights cases as cannabis legalization unfolds across the U.S. and states like New Jersey try to set up legislation that tries to undo the damage that the Drug War has had on African-American and Latino communities.

Data Set (including wrangling):

The data that I will be using to carry out this project is the NYPD Complaint Data Historic dataset from the NYC Open Data project, available to the public at https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i . Using this dataset has its advantages, as it is arrest data directly reported from the NYPD officers making the cannabis arrests. The findings from this dataset could not be refuted by the NYPD, as it is their data. Also, it has data on all arrests, so comparative analysis with non-cannabis crimes can easily be done.

Data cleaning and wrangling was done to create several versions of the dataset for different purposes. To create DataFrames with just cannabis crimes, the following Jupyter notebook for cleaning and wrangling was used:

https://github.com/danloew/Springboard/blob/master/Data_Cleaning_cann_old.ipynb

To create DataFrames with just non-cannabis crimes, the following Jupyter notebook for cleaning and wrangling was used:

https://github.com/danloew/Springboard/blob/master/Data_Cleaning_ncann_old.ipynb

To create DataFrames of samples from the overall NYC crime dataset, the following Jupyter notebook for cleaning and wrangling was used:

https://github.com/danloew/Springboard/blob/master/Data_Cleaning_HT.ipynb

The following data cleaning and data wrangling steps were performed on the NYPD data set “NYPD Complaint Data Historic” that records data on all crimes committed in New York City between 2006 and the first part of 2018:

1. Dropped two columns that had multiple data types and were either of unlikely relevance to the predictive model that is central to this project, were unlabeled and therefore meaningless, or had too many values for predictive analysis. These two variables were ‘PARKS_NM’ and ‘HOUSING_PSA’, which showed which city park the crime reportedly occurred in, and a list of unlabeled numerical codes having to do with housing projects, respectively. ‘PARKS_NM’ had 1,130 values, far too many for the predictive model. There are also other geographic variables to use, and only 9.5% of cannabis crimes were recorded as occurring in city parks. The data dictionary had no information on the meaning of ‘HOUSING_PSA’.

2. A new dataset was subsetted from the larger dataset to only include cannabis crimes. This only includes crimes with penal codes (PD_CD) 566-570. These codes are 566) Marijuana Possession, 567) Marijuana Possession 4 & 5, 568) Marijuana Possession 1, 2, & 3, 569) Marijuana Sale 4 & 5, and 570) Marijuana Sale 1, 2, & 3.

3. Binary features (i.e., columns) were created for cannabis possession (PD_CDs 566-568), cannabis sales (PD_CDs 569 and 570), misdemeanors (LAW_CAT_CD = misdemeanor), violations (LAW_CAT_CD = violation), and felonies (LAW_CAT_CD = felony).

4. The features in step 3 were then used to create features for the following crime levels: misdemeanor cannabis possession, violation cannabis possession, felony cannabis possession, misdemeanor cannabis sales, violation cannabis sales, and felony cannabis sales. There were no violation sales cases, but the other five features will serve as the target variables for the machine learning classification of cannabis crime levels that will be carried out later in the capstone project pipeline.

5. The chained .isnull().sum() and .isna().sum() methods were then used to show how many missing values there were in each of the features in the feature set. A set of features had their missing values filled in with an ‘unknown’ value or some other similar

feature-specific value (like “not transit-related” for the TRANSIT_DISTRICT feature). This set included the following features (missing value n reported in parentheses): HADEVELOPT (193,639), CMPLNT_TO_DT (67,454), CMPLNT_TO_TM (67,381), TRANSIT_DISTRICT (217,273), STATION_NAME (217,273), BORO_NM (185), LOC_OF_OCCUR_DESC (92,077), PREM_TYP_DESC (1,731), SUSP_AGE_GROUP (186,008), SUSP_RACE (185,550), SUSP_SEX (185,587), PATROL_BORO (1), VIC_AGE_GROUP (188,373), VIC_RACE (54), and VIC_SEX (54). The date variable had missing values coded as ‘00/00/0000’, and the time variable had missing values coded as ‘00:00:00’.

6. Another set of features had missing values dropped, as imputation of these values would be specious and biasing to the results, and the number of missing values were rather small overall. This set included the following features (missing value n reported in parentheses): CMPLNT_FR_DT (2), CMPLNT_FR_TM (1), X_COORD_CD (472), Y_COORD_CD (472), Latitude (472), Longitude (472), and Lat_Lon (472).

7. Datetime crime start, crime end, and crime duration variables were created from the crime start date, crime start time, crime end date, and crime end time variables that came with the dataset.

8. The raw crime start time variable was used to create a set of time-window features that may be predictive of cannabis crimes. The start time variable 'CMPLNT_FR_TM' was used, as all crimes have a start time in the data set but not necessarily an end time. It also makes simple logical sense to timestamp a crime at the time that it starts. The derived time window features include daytime, night time, early morning, morning rush hour, the traditional work day, the lunch hour, evening rush hour, dinner hour, evening, and late night.

9. The distance of each cannabis crime from prominent NYC landmarks was encoded into continuous data features. All latitudes and longitudes were found from Google searches. Both driving/walking/biking distances ("_taxi" variables) and shortest distances ("_crow" variables) were computed. The landmarks included the World Trade Center, the New York Stock Exchange, Brooklyn Bridge, New York City Hall, Manhattan Bridge, Williamsburg Bridge, Washington Square Park, Union Square, Penn Station, Times Square, Rockefeller Center, Empire State Building, Lincoln Center, Central Park, Apollo Theatre, Yankee Stadium, Mets Stadium, the center of Queens Borough, the center of Prospect Park, the center of downtown Brooklyn, Staten Island Ferry Terminal, Port Authority Bus Station, New York Police Department headquarters, Manhattan Detention Center, Rikers Island, and the New York Supreme Court. With this step, each cannabis crimes’ distance from key geographic landmarks in New York City is known. This information will be used in the predictive analysis to follow.

10. Isolated year, month, and date features were extracted from the crime start datetime variable. Along with being useful data to have on their own, these extracted variables

were used to define holidays that fall on the same day every year. These included New Year’s Eve, New Year’s Day, Christmas Eve, Christmas Day, July 4th, Valentine’s Day, Halloween, and St. Patrick’s Day.

11. Boolean masks and variable assignment were used to create new binary features for major holidays that do not fall on the same day every year. The first step created separate boolean masks for the date of the holiday in question for each year, then a boolean mask was created for all years based on the boolean masks for each individual year, and then the holiday's feature was assigned. To keep the dataset tidy, the variables for each holiday’s year were then deleted, as they were no longer needed. These holidays include Martin Luther King Day, President’s Day, Easter, Diwali, the Puerto Rican Day Parade, Yom Kippur, Rosh Hashanah, Eid al-Fitr, Eid al-Adha, Hannukkah, Memorial Day, Labor Day, and Thanksgiving.

12. Cases outside of the stated year range of the dataset were dropped, that is, cases earlier than 2006.

13. Unclear values were recoded to ‘unknown’ for the suspect and victim age group, race, and sex variables. First, the value counts were called for these features for reference, a cleaned version of the feature was created by mapping unusual values to ‘unknown’, and then value counts were called on the cleaned version of the feature to ensure all unclear values had been mapped to ‘unknown’. Unclean versions of these features were dropped from the dataset. ‘JURISDICTION_CODE’ was also dropped, as it had numerical codes that were meaningless as they weren't labeled, and there is a descriptive variable for information on which jurisdiction the crime fell under.

14. A separate DataFrame was subsetted from the overall DataFrame of cannabis crimes, with just those cannabis crimes that had suspect race reported. This is for later machine learning classification of cannabis crime types for just the cases whose suspect race was reported.

15. A .csv file was then exported for both the overall DataFrame of cannabis crimes and the smaller DataFrame of cannabis crimes that had suspect race reported. These DataFrames keep their categorical features intact, to be used for visual and statistical EDA in the next report. This was done before the .get_dummies() method described in the next step that makes the DataFrames machine learning friendly.

16. The remaining categorical variables were transformed into individual binary features using .get_dummies(), so that each categorical value of each of these features ends up having its own feature. This was done for both the larger DataFrame of all cannabis crimes, and for the smaller subsetted DataFrame with cannabis crimes whose suspect race was reported. This is done for machine learning classification purposes later in the capstone project pipeline. The features that were transformed included: CRM_ATPT_CPTD_CD, SUSP_AGE_GROUP_cleaned, SUSP_RACE_cleaned,

SUSP_SEX_cleaned, VIC_AGE_GROUP_cleaned, VIC_RACE_cleaned, VIC_SEX_cleaned, BORO_NM, HADEVELOPT, JURIS_DESC, LOC_OF_OCCUR_DESC, PATROL_BORO, PREM_TYP_DESC, TRANSIT_DISTRICT, and STATION_NAME.

17. A list of variables was then dropped from both DataFrames as they were superfluous or were in a data type which would not be readable by machine learning methods (like 'string' and 'datetime'). The variables that were dropped were: possession, sales, misdemeanor, violation, felony, viol_sales, Lat_Lon, LAW_CAT_CD, PD_DESC, OFNS_DESC, CMPLNT_FR_DT, CMPLNT_TO_DT, RPT_DT, CMPLNT_FR_TM, CMPLNT_TO_TM, date_time_start, date_time_end, and duration.

18. The cleaned DataFrames were then exported to a .csv file for easy loading for the upcoming machine learning classification.

19. A separate DataFrame of cannabis crimes was created for a planned second round of machine learning classification of cannabis crimes vs non-cannabis crimes. The non-cannabis crime DataFrame was created in a separate Jupyter notebook that followed the exact same cleaning protocol as the cannabis crime DataFrame, but was randomly sampled. PD code and law category codes were compared between the overall population of non-cannabis crimes and the sample, to ensure that there was not an oversampling of any specific crime type, so as not to bias the results later.

20. Because this second round of classification won't need to differentiate between types of cannabis crime, and for easy merging with the cleaned non-cannabis crime DataFrame, the cannabis crime type target variables were dropped from the cannabis crime DataFrame. A new label variable was added to the cannabis crime and non-cannabis crime DataFrames for classification and supervised learning purposes.

21. Versions of cannabis crime and non-cannabis crime DataFrames were also created for just those cases where the suspect's race was reported, as that will be an important part of the analysis.

22. Csv exports were done for these last four DataFrames.

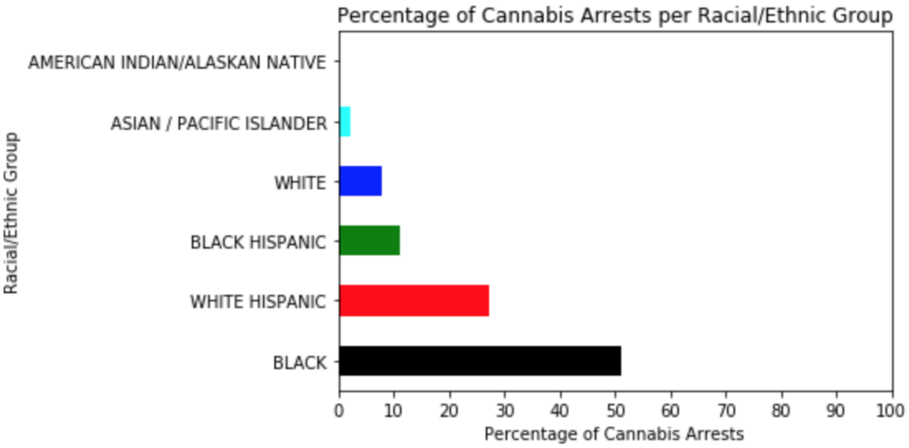
23. In order to explore exploratory hypothesis testing, a sample of the original NYPD "NYPD Complaint Data Historic" dataset was taken, which included cannabis and non-cannabis crimes. A cannabis crime flag was added, as were flags for the five cannabis crime types. A sample of the crimes with the suspect's race reported was also taken. All the prior data cleaning steps were taken on these sample DataFrames, and csv exports were made.

Exploratory Data Analysis:

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.

This difference was the subject of a hypothesis test with the null hypothesis that cannabis crimes are equally likely in having their suspect's race reported by the arresting NYPD officer as non-cannabis crimes. This null hypothesis was tested with a t-test, using SciPy Stats' "ttest_ind_from_stats" function. The null hypothesis was rejected with a p-value of zero, suggesting that the lower percentage of cannabis crimes having their suspect's race reported was not due to random chance, but to other factors. The cause for this difference is beyond the scope of this report, but warrants further research.

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.

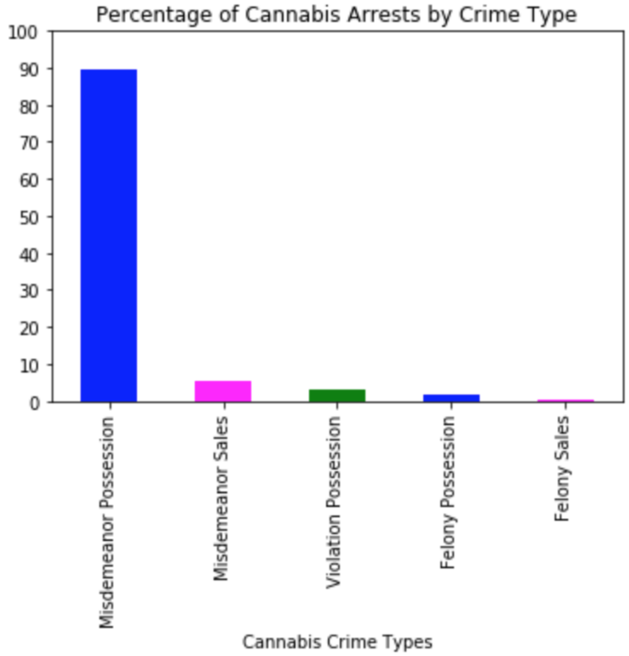


Looking just at age shows that 84% of cannabis arrests were made of people between the ages of 18-44. 90% were of males. 40% of cannabis arrests are of African-American males younger than 45, and 32% are of Hispanic/Latino males younger than 45, for a total of 72% of all cannabis arrests in New York City between 2006 and 2018 being of young African-American and Hispanic/Latino males.

A second null hypothesis was tested, whether African-Americans arrested for a crime are equally likely to be arrested for cannabis crimes as they are for non_cannabis crimes. This was also rejected with a p-value of zero, suggesting that the difference is not due to random chance and that there are unverified reasons why African-Americans are more often arrested for cannabis crimes than non-cannabis crimes. It bears reminding that SAMHSA surveys have consistently showed that people of different racial and ethnic groups use cannabis at the same rate. Interestingly, this result was replicated for whites, white Hispanics, black Hispanics, and Asians. Perhaps this simply shows that more people are being arrested for cannabis than for non-cannabis crimes.

Inter-group differences were then looked at, with the null hypothesis that African-Americans arrested for a crime are equally likely to be charged for cannabis crimes as white people arrested for a crime. This was rejected with a p-value very close to zero. This was replicated when comparing white Hispanics and whites, black Hispanics and whites, Asians and whites, African-Americans and white Hispanics, African-Americans and black Hispanics, and white Hispanics and black Hispanics. This suggests a hierarchy of likelihood in being arrested between different racial and ethnic groups. All of these findings corroborate the racial disparity data in cannabis arrests reported elsewhere.

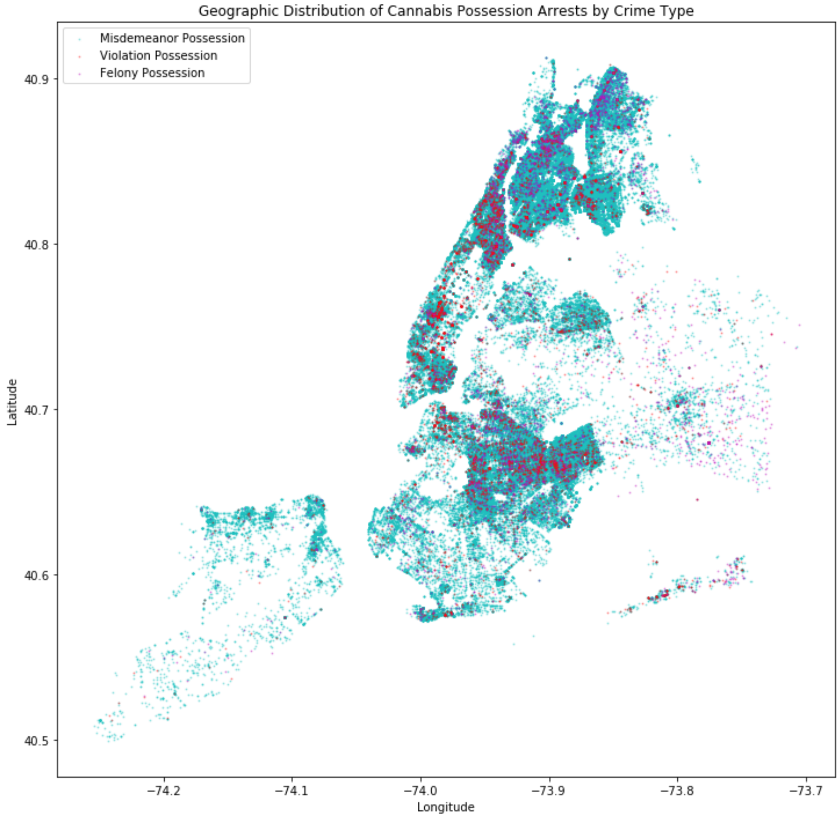
One of the striking things about cannabis arrests in New York City are that 92.6% of them are for simple misdemeanor and violation possession charges, which is the vast majority. 1.7% are for felony-level possession, 5.2% are for misdemeanor sales, and 0.5% are for felony sales, the latter being arguably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

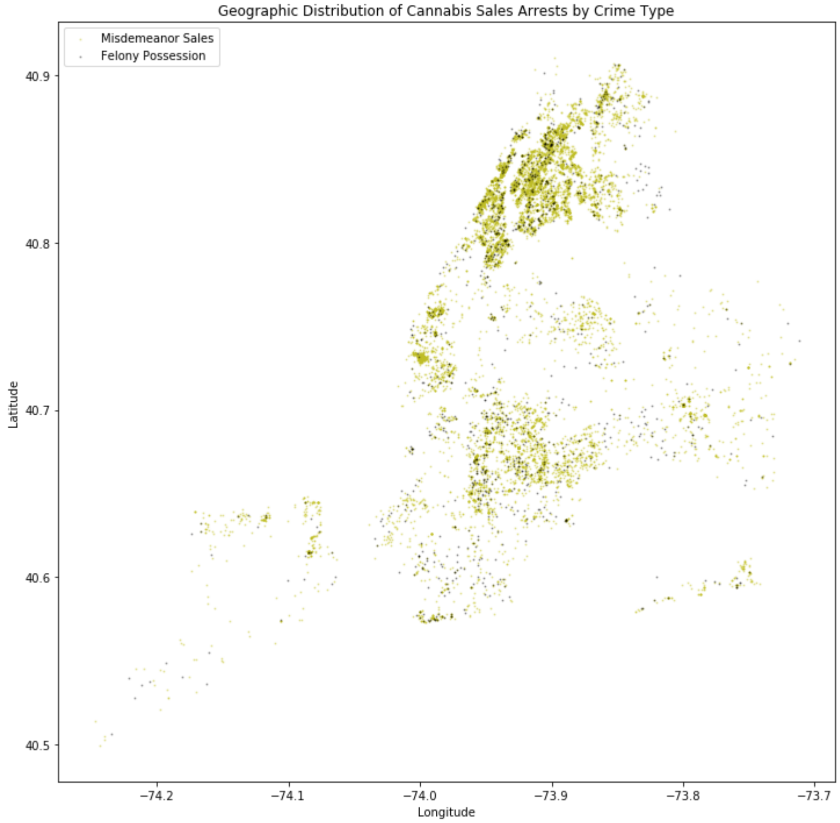


The same racial disparity described above holds true for all levels of cannabis crime. The only exceptions to the racial disparity are that more violation possession arrests are made of white perpetrators than of black Hispanic perpetrators, but the difference is only 3%. Also, it should be noted that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

The different levels of cannabis crime types were then looked at. The first null hypothesis to be tested was that African-Americans arrested for a cannabis crime are equally likely to be charged for misdemeanor cannabis possession as they are for violation cannabis possession. This was rejected at a p-value very close to zero. A null hypothesis was also tested that whites arrested for a cannabis crime are equally likely to be charged for misdemeanor cannabis possession as they are for violation cannabis possession. This was not rejected, with a p-value of 0.13. To see if more whites are arrested for violation possession (the least serious offense) than African-Americans, the null hypothesis was tested that African-Americans arrested for a cannabis crime are equally likely to be arrested for violation possession as are Whites arrested for a cannabis crime. This null hypothesis was not rejected, with a p-value of 0.43. A further null hypothesis was tested that African-Americans arrested for a cannabis crime are equally likely to be arrested for misdemeanor possession as they are for felony possession. This was rejected with a p-value very close to zero, suggesting that African-Americans may be more likely to be hit with a felony charge than a misdemeanor charge when it comes to possession. And finally, a null hypothesis was tested that African-Americans arrested for a cannabis crime are equally likely to be arrested for misdemeanor sales as they are for felony sales. This was also rejected at a p-value of 0.01, suggested that African-Americans may be more likely to be arrested on a felony level sales charge.

To look at other indicators of a 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 detailed above. 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.

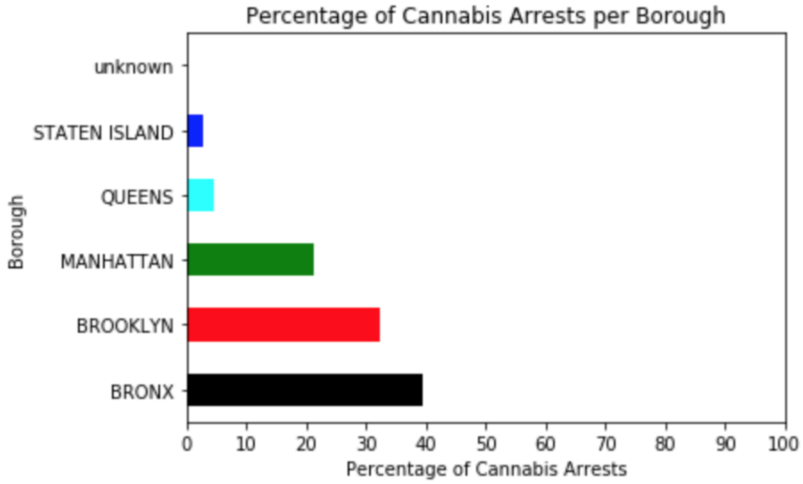




Because there are references with demographic information of the various parts of New York City, the visual concentration of arrests in certain parts of the city enable us to partially infer race of cannabis crime suspects in the overall DataFrame, where only 16% of cases have suspect race reported.

The vast majority of cannabis arrests are for misdemeanor possession, and in Manhattan they 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 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 are centered around housing projects like Stapleton and Park Hill. Because sales are so different than simple possession, and for ease of visualization, cannabis sales were visualized separately. Sales arrests tended to fall within the same neighborhoods as possession arrests.

The first geographic indicator of New York City is the borough. The Bronx and Brooklyn are home to the majority of cannabis crimes overall. This is interesting because of the racial demographics of these two boroughs. The Bronx's populace is 36% black, 48% Latino, and only 14.5% non-Latino white, and Brooklyn's populace is 36% black, 20% Latino, and 36% non-Latino white. By contrast, Manhattan's populace is 16% black, 25% Latino, and 48% non-Latino white. Queens is 19% black, 27% Latino, and 30% non-Latino white; and Staten Island is 11% black, 17% Latino, and 65% non-Latino white. Misdemeanor and felony possession charges are dominant in the Bronx and Brooklyn, while violation possession charges (lowest level) are dominant in Manhattan. This reflects the evidence that cannabis crimes are punished very differently in New York City dependent 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.



A series of hypothesis tests was also conducted looking at whether cannabis arrests are equally likely to be made in the five boroughs. All comparison tests were rejected at a p-value close to zero or zero. As was shown in earlier portions of this report, the boroughs are organized by percentage of overall cannabis arrests in the following descending order: the Bronx, Brooklyn, Manhattan, Queens, and Staten Island.

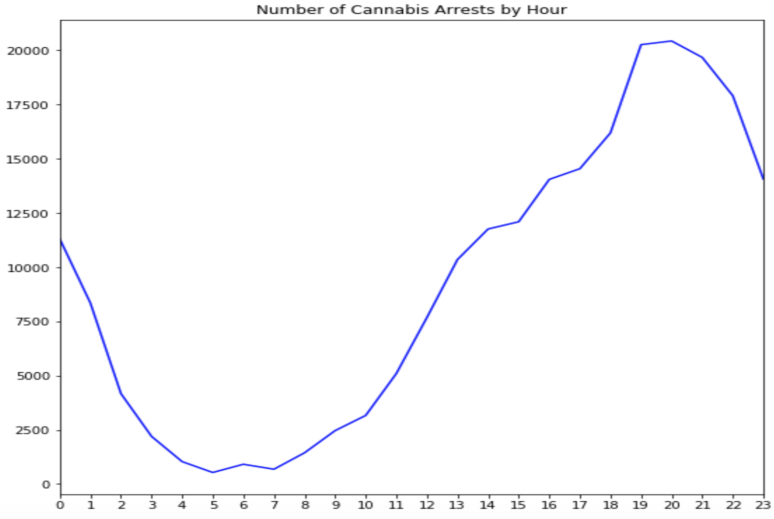
It would be interesting to see which neighborhoods of Manhattan are responsible for these differences. Police precincts offer a route to explore these smaller geographic zones. The top 10 police precincts with the highest amounts of misdemeanor cannabis arrests and cannabis arrests overall are all in 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.

An intriguing part of the NYPD dataset is a feature that describes the premises type that the arrest occurred in. As can be seen below, the majority of cannabis arrests happen either on the street or in the New York City housing projects. Violation possession charges also occur in the New York City subway system.
As can be expected, the jurisdiction responsible for the majority of cannabis arrests are the NYPD, the New York City Housing Authority (NYCHA), and to a much lesser degree the N.Y. Transit Police. The fact that 19% of cannabis arrests fall under the jurisdiction of the NYCHA shows how heavily policed these public housing projects are. The fact that the N.Y. Transit Police takes the NYCHA's place for violation possession charges show an interesting difference in enforcement of the different cannabis types, and reflects the fact that the premises type for violation possession is frequently in the N.Y. subway system.

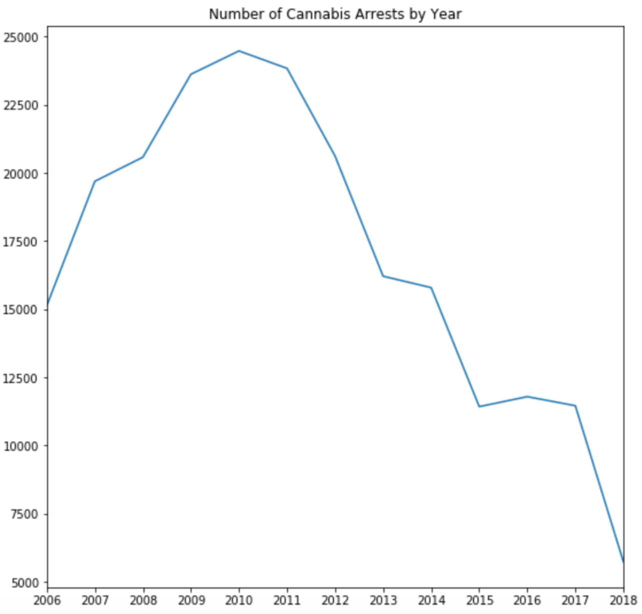
Because of the fact that nearly 20% of all cannabis arrests occur in New York City housing projects, the housing projects with the highest concentration of cannabis arrests were looked at. As is consistent with the rest of the data story, the top ten New York City housing developments with the highest proportion of cannabis arrests are all in the South Bronx or in economically disadvantaged areas of Brooklyn.

Cannabis arrests occur 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, but one sees a fairly high concentration of arrests (18.1%) happening during this time window.

As reported above, the nighttime sees the majority of cannabis arrests, at 61%. Overlapping with the evening rush hour, the dinner window of 6-8 pm has a high concentration of arrests for just a two hour window (17.4%), and has nearly as many arrests that occur in the 2.5 hour window of the evening rush hour. Evening (8-10 pm) has a similarly high concentration of arrests at 19% for a two hour window. Late night (10 pm - 6 am) has 26% of the arrests for an 8 hour window, showing that more than half of the nighttime arrests do not happen during the nightlife hours, but after work and before the working population would typically go to bed. The distribution of cannabis arrests over the typical day (averaged over all years from 2006-2018) is shown below:



It has been well reported that during Mayor Bloomberg's time as mayor, cannabis arrests reached their peak. One can see that 2006 has 15,127 arrests, and that this increases to 24,468 arrests in 2010. This holds fairly steady for 2011 (23,827), drops a bit in 2012 (20,611) as criticism of Bloomberg's "stop and frisk" program mounts, and then drops significantly in 2013 (16,206) when the "stop and frisk" program was judged as unconstitutional. 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 has about the same amount of cannabis arrests, but August has the highest number and the number drops in November and December during the Holiday season. Each day of the month has a fairly consistent number of cannabis arrests, ranging from 5,660 to 7,900 arrests a day. The number drops somewhat in the last 10 days of the month. The 31st has 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 makes sense to look at whether certain holidays have higher concentrations of cannabis arrests. Due to the cultural diversity of New York City, certain holidays are included that would not be typically celebrated in other parts of the United States. Intriguingly, the holidays with the highest number of cannabis arrests are Hindu, Jewish, and Muslim holidays. Diwali had 656 arrests, Yom Kippur has 707, Rosh Hashanah has 677, Eid al-Fitr has 644, and Eid al-Adha has 544. St. Patrick's Day also has 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.

The picture that emerges from exploring the descriptive statistics of cannabis arrests in New York City between 2006 and 2018 is one of racial bias against African-Americans and Hispanics for all five levels of cannabis crimes. This is further supported by looking at the geographic areas where these arrests are occurring, and seeing that from every angle the geographic areas being hit the most are boroughs, precincts, neighborhoods, and housing projects that are predominantly occupied by African-American and Hispanic residents. These arrests are largely happening during the evening and early nighttime hours of the day, and it was also seen that there are not huge spikes in holiday arrests except for those holidays intrinsically linked with religious minorities. Machine learning classification methods will be used later in the report cycle to look at which features are most predictive of cannabis arrests compared to non-cannabis arrests, and most predictive of the five different levels of cannabis crime.