Targets: Using Machine Learning Classification Models to Identify Salient Predictors of Cannabis Arrests in New York City, 2006-2018

Capstone Project 1: Overall Capstone Project

Author: Daniel Loew, https://www.linkedin.com/in/danielloew/

Introduction and Problem Statement

As has been reported (Levine, 2017; Harcourt & Ludwig, 2006; Mueller, Gebeloff, & Chinoy, 2018), data on low-level cannabis possession arrests in New York City have shown that they have been predominantly of young African-American and Latino men since at least 1987. Given the history of the drug war under Presidents Nixon, Reagan, Bush Sr., Clinton, and Bush Jr., and initially under Harry Anslinger during his years as the first Commissioner of the U.S. Treasury Department's Federal Bureau of Narcotics during the Jazz Age, this racial disparity of low-level cannabis arrests has likely remained constant since the Marijuana Tax Act of 1937 passed, which effectively made the plant illegal. However, data from the Substance Abuse and Mental Health Services Administration of the U.S. Department of Health and Human Services shows consistently that people of different racial groups use cannabis at effectively the same rate (SAMHSA, 2018). This is an issue of extreme concern for human and civil rights groups, as it allows discrimination in criminal justice under the guise of public health concerns.

This racial disparity in cannabis arrests continues to this day through the mayoral transition to Mayor DeBlasio from Mayor Bloomberg's policy era of stop-and-frisk arrests, even while cannabis arrests have dropped since their height around 2011 (Levine). At the same time, overall crime has dropped in New York City (Mueller, Gebeloff, & Chinoy, 2018). The New York Police Department (NYPD) has been pressed to explain this disparity, and has responded by saying that it is due to the fact that they receive more cannabis-related complaints from neighborhoods which are predominantly occupied by African-American and Latino residents.

The New York Times has done an analysis exploring this claim, and has shown that even between neighborhoods that have the same level of cannabis-related complaints, more cannabis arrests occur in neighborhoods with a majority of African-American and Latino residents (Mueller, Gebeloff, & Chinoy, 2018). One explanation for the racial disparity is that these neighborhoods are often more policed because of a higher rate of violent crimes there. Another explanation is that when people are arrested for cannabis, NYPD officers are able to check for open warrants and are therefore a way for police officers to cut down on other types of crime through these arrests. But these explanations do not fully illustrate the reasons that this racial disparity in low-level cannabis arrests persist during an era of criminal justice reform, given the data available.

While looking at both low-level and more serious cannabis arrests including felony sales, this report aims to provide a more complete picture of the predictors that influence cannabis arrests in New York City. In order to do so, machine learning classification methods will be applied to predict the following five target features: misdemeanor cannabis possession, violation cannabis possession, felony cannabis possession, misdemeanor cannabis sales, and felony cannabis sales. Additionally, these methods will be used to predict possession arrests vs. sales arrests. Violation sales were not used as a target feature as violation sales of cannabis are not a legal category.

Classification methods will also be used to try and differentiate predictors of cannabis crimes from non-cannabis crimes; in other words, all other crimes. This will be done using all cannabis crimes and a set of non-cannabis crimes of the same size randomly selected from the larger pool of non-cannabis crimes. By using machine learning methods to discover the features with the largest coefficients within classification models, the identification of biases in the most scientific sense of the term can be identified for future research and analysis by criminologists, human rights groups, legal scholars, public policy researchers, and more.

These methods will be used on all crimes between January 1st, 2006 and December 31st, 2018 in New York City as reported by the NYPD's Complaint Data historic dataset. A set of features from the original dataset and a set of features derived from this data will be used to create classification models that will attempt to identify several salient predictors of cannabis arrests in New York City during modern times. It is important to note that the data cleaning protocols and classification models developed in this project can also be modified to investigate the predictive factors of any type of crime that occurs in New York City.

Hopefully this project will present a fuller image of cannabis arrests that can be used to inform and improve drug policy in New York City and in the rest of the country, so that the suffering caused by Drug War policies can be reduced.

A summary of this project in the form of a Google slide deck is available at this link:

https://docs.google.com/presentation/d/1Tsno2V6mdiIv5j04ikVXX5ZFqq5gsULlFlDXnczozT4/edit?usp=sharing

Data Cleaning and Wrangling

The dataset used to build classification models of cannabis crime between the years of 2006 and 2018 was the "NYPD Complaint Data Historic" dataset downloaded from https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i. This dataset was chosen most importantly because it includes all valid misdemeanor, violation, and felony crimes reported to the NYPD, is drawn from data reported by the NYPD officer associated with each crime, and is available to the public via the NYC Open Data project. The classification models that were built from this dataset were used to identify the primary features that differentiate cannabis crime from all other crime, cannabis possession crime from cannabis sales crime, and the five more granular categories of cannabis crime: misdemeanor possession, violation possession, felony possession, misdemeanor sales, and felony sales. Doing so provides a comprehensive image of the coefficients of cannabis arrests within the bounds of the dataset provided by the NYPD.

The Jupyter notebooks with the data cleaning and wrangling steps detailed below can be found at the following two links:

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

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

The following data cleaning and wrangling steps were performed on the "NYPD Complaint Data Historic" dataset:

1. The dataset was loaded in full into a Pandas DataFrame.
2. The ‘PARKS_NM’ and ‘HOUSING_PSA’ variables were found to be of mixed data type. The ‘HOUSING_PSA’ variable was dropped from the DataFrame because it contained numeric codes for public housing developments whose character names were located elsewhere in the ‘HADEVELOPMENT’ variable, and there was no key in the data dictionary which provided correspondence between the numeric codes and character names of the housing developments. The ‘JURISDICTION_CODE’ feature was also dropped, as there was already a character data type feature called ‘JURIS_DESC’. The ‘PARKS_NM’ variable which contained the names of NYC parks where crimes were committed was kept, but was coerced to the ‘string’ data type.
3. To ensure that there were no duplicate rows, which can be a source of machine learning model leakage, the .drop_duplicates() method was used. No duplicate rows were found.
4. A new Pandas DataFrame was subsetted from the larger DataFrame to only include cannabis crimes. This only included crimes with penal codes (PD_CD) 566-570. These codes are listed in the NYPD’s data dictionary as:
    a. 566 - Marijuana, Possession (violation level)
    b. 567 - Marijuana, Possession, 4th & 5th degree (misdemeanor level)
    c. 568 - Marijuana, Possession, 1st, 2nd, and 3rd degree (felony level)
    d. 569 - Marijuana, Sales 4th & 5th degree (misdemeanor level)
    e. 570 - Marijuana, Sales, 1st, 2nd, and 3rd degree (felony level)
5. Binary features 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).
6. The binary features created in step 5 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 cannabis sales cases as that is not a valid legal category of crime, but the other five features are the target features for the machine learning classification of the five cannabis crime levels.
7. 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 another similar feature-specific value (like “not transit-related” for the TRANSIT_DISTRICT feature). The date variable had missing values coded as ‘00/00/0000’, and the time variable had missing values coded as ‘00:00:00’. This set included the following features (missing value n reported in parentheses):

    a. HADEVELOPT (193,639),
    b. CMPLNT_TO_DT (67,454),
    c. CMPLNT_TO_TM (67,381),
    d. TRANSIT_DISTRICT (217,273),
    e. STATION_NAME (217,273),
    f. BORO_NM (185),
    g. PARKS_NM (218,332),

    h. LOC_OF_OCCUR_DESC (92,077),
    i. PREM_TYP_DESC (1,731),
    j. SUSP_AGE_GROUP (186,008),
    k. SUSP_RACE (185,550),
    l. SUSP_SEX (185,587),
    m. PATROL_BORO (1),
    n. VIC_AGE_GROUP (188,373),
    o. VIC_RACE (54),
    p. and VIC_SEX (54)
8. 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. These dropped cases amounted to 475 out of 220,817 rows (or 0.2%), which was an acceptable loss for a uniform DataFrame with no missing values across all features. This set included the following features (missing value n reported in parentheses):
    a. CMPLNT_FR_DT (2),
    b. CMPLNT_FR_TM (1),
    c. X_COORD_CD (472),
    d. Y_COORD_CD (472),
    e. Latitude (472),
    f. Longitude (472), and
    g. Lat_Lon (472)
9. Datetime crime start, crime end, and crime duration features were created from the crime start date (‘CMPLNT_FR_DT’), crime start time (‘CMPLNT_FR_TM’), crime end date (‘CMPLNT_TO_DT’), and crime end time (‘CMPLNT_TO_TM’) features native to the NYPD’s dataset. These new features were named ‘date_time_start’, ‘date_time_end’, and ‘duration’. ‘Duration’ was created to measure the amount of time elapsed between the start of the crime complaints and their end.
10. The newly created ‘duration’ feature was checked for null values, which it had 67,212 of as this number of cases only had a ‘date_time_start’ value. Null values were edited to the specific value of zero days, hours, minutes, and seconds.
11. The new ‘duration’ feature was of TimeDelta data type, which is unable to be used by the Logistic Regression classifiers. Therefore, it was converted into the integer data type feature 'duration_days', in order to identify just how many days a case transpired over. The original TimeDelta ‘duration’ feature was then dropped from the DataFrame.
12. Discrete year, month, date, hour, minute and second features were extracted from the ‘date_time_start’ feature as integer data type features for use in the classification models. Unusual values were checked for in these features, and none were found except for several cases that did not occur within the prescribed year range of 2006 to 2018. Along with being useful features to have on their own, the extracted date and time features were used later in the data cleaning pipeline to create other features.
13. The cases with year values outside of the range of 2006 to 2018 were identified and dropped from the DataFrame.
14. Because the 'duration_days' feature uses 'date_time_end' in its valuation of the number of days a crime takes to occur, extracting datetime features from 'date_time_end' is not needed. In other words, the 'duration_days' feature stores the information of the date and time that the crime ends. Therefore, the 'date_time_end' feature was not used in the feature

set for classification models, and was dropped from the DataFrame. 'CMPLNT_TO_DT', 'CMPLNT_TO_TM', and 'end_year' were also no longer needed, and were dropped.

15. The 'RPT_DT', or date that the crime was reported by the NYPD, differed from the complaint date stored in the 'CMPLNT_FR_DT' feature 7.2% of the time. This difference was postulated as a possible predictive feature. Before creating a predictive feature based on the difference between the two dates, ‘RPT_DT’ was first converted into a feature of DateTime data type and then year, month, and date features were extracted from it so that any outliers outside of the prescribed date range of 2006 to 2018 could be identified and dropped. There were none such cases. The DateTime feature and extracted year, month, and date features were no longer needed and were dropped. A binary feature was then created to store the information of whether the complaint date matches the police-reported date. Because the original 'RPT_DT' feature contains the same information as 'CMPLNT_FR_DT' 92.8% of the time, the 'RPT_DT' feature itself will be dropped later in the data cleaning pipeline so as to avoid duplication of information that could theoretically adversely affect the classification models.
16. The native crime start time variable ‘CMPLNT_FR_TM’ was used to create a set of time-window features that may be predictive of cannabis crimes. The derived time window features included daytime, night time, early morning, morning rush hour, the traditional work day, the lunch hour, evening rush hour, dinner hour, evening, and late night.
17. Boolean masks and feature assignment were used to create new binary features for major holidays that always fall on the same day of the year. These holidays could be predictors of cannabis crimes because different holidays have special meaning to demographic groups that may be differentially targeted for cannabis arrests, and more generally holidays are drivers of behavior in the United States. These major holidays are New Year's Day, Valentine's Day, St. Patrick's Day, July 4th, Halloween, Christmas Eve, Christmas, and New Year's Eve. April 20th is included as it is an emerging day of cannabis celebrations and could be relevant. Intriguingly, this day has more cannabis arrests than any holiday. The creation of these binary fixed holiday features was done by first creating a DataFrame with the month and day of each fixed holiday. Then, the contents of the DataFrame were used by 'for' loops to create individual features storing whether a crime occurred on that particular fixed holiday or not. Fixed holiday features were then cast into integer type.
18. Pd.DataFrame and 'for' loops were used to create new binary features for crimes that occur on major holidays that do not fall on the same day every year, i.e., "floating holidays". This is done by first creating a DataFrame with the month, day, and year of each floating holiday, for each of the years between 2006 and 2018. Then, the contents of the DataFrame were used with a combination of 'for' loops, boolean assignment across the years, and integer data type casting to create individual features storing whether a crime occurred on that particular floating holiday (1) or not (0). These floating holidays included Martin Luther King Day, President’s Day, Easter, Diwali, the Puerto Rican Day Parade, Yom Kippur, Rosh Hashanah, Eid al-Fitr, Eid al-Adha, Hanukkah, Memorial Day, Labor Day, and Thanksgiving.
19. Outlier and erroneous values of several features native to the NYPD's data set were looked for and dropped if necessary. All values of these features were verified with online sources specified in the Jupyter notebooks. These features were the:
    a. police precinct (‘ADDR_PCT_CD’),
    b. NYC borough (‘BORO_NM’),
    c. location of crime occurrence in relation to the premises (‘LOC_OF_OCCUR_DESC’),

    d. premises type description of crime occurrence ('PREM_TYP_DESC'),
    e. jurisdiction of crimes ('JURIS_DESC'),
    f. NYC parks crimes occurred in ('PARKS_NM'),
    g. housing developments crimes occurred in ('HADEVELOPT')
    h. transit districts crimes occurred in ('TRANSIT_DISTRICT'),
    i. latitude and longitude crimes occurred in ('LATITUDE', 'LONGITUDE'),
    j. NYPD patrol borough crimes occurred in ('PATROL_BORO'), and
    k. MTA transit stations ('STATION_NAME')
20. The only features that had outliers were latitude and longitude. There was one case that occurred outside of New York City, which was discovered through a scatterplot of the two features. This case was dropped from the DataFrame.
21. Subway station entrances can be places where people sell, purchase, consume, and get cannabis delivered in New York City, and the Open NY project has latitudes and longitudes of all NYC subway stations and their entrances (available at https://data.ny.gov/widgets/i9wp-a4ja). Features were created that determined both the L1 (Euclidean or "taxi") and L2 (straight-line or "as the crow flies") distances in units of latitude/longitude to the closest subway station for each cannabis crime.
22. The distance of each cannabis crime from prominent NYC landmarks was encoded into continuous data features. All latitudes and longitudes for these landmarks were found via Google search. Both L1 (Euclidean or "taxi") and L2 (straight-line or "as the crow flies") distances in units of latitude/longitude 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, the Staten Island Ferry Terminal, the Port Authority Bus Station, the New York Police Department headquarters, Metropolitan Detention Center in Manhattan, Riker's Island, and the New York Supreme Court. With this step, each cannabis crimes' distance from key geographic landmarks in New York City is known.
23. Unclear values were recoded to 'unknown' for the suspect and victim age group, race, and sex features. First, the value counts were run for each feature. Then, a cleaned version of the features were created by mapping unusual values to the 'unknown' value. Approximately 84% of the cases had an 'unknown' value for these six demographic features.
24. For ease of use in doing visual and statistical EDA, two versions of the DataFrame were then exported before the binarization of categorical data features (which is implemented later in the data cleaning pipeline for use in the machine learning models). The first version of the DataFrame contained all cannabis crime cases, while the second version only contained cannabis crime cases where the suspect's race was recorded by the arresting officer (approximately 16% of the overall cases).
25. The hypothesis testing done in the Statistical Data Analysis notebook required a sample of all NYC crimes. After the data cleaning process described above, a 10% sample was taken and exported to a .csv file. The same process was done in the non-cannabis crime data cleaning notebook and the two sample DataFrames were concatenated in the Statistical Data Analysis notebook (see below).

26. The remaining categorical features were transformed into individual binary features via the Pandas .get_dummies() method in a separate machine learning version of the DataFrame, so that each value of categorical features has its own binary feature. This is done for later machine learning classification of different cannabis crime types within the universe of NYC cannabis crimes between 2006 and 2018. Several features were identified as being superfluous or as causing leakage in the machine learning pipeline, as they contain the same information as the target feature being predicted. Meanwhile, a set of valuable features were binarized and kept in the machine learning version of the DataFrame. A breakdown of which features were dropped and which were binarized follows:

    a. Because the five target features for the classification model of the five types of cannabis crime were already instantiated, some of the features that were used to create the target features were no longer needed. These included 'misdemeanor', 'violation', and 'felony'. 'Possession' was kept as the target feature for classification models differentiating cannabis possession crimes from cannabis sales crimes, which will be explored with a second round of classification. However, 'sales' was not needed, so it was dropped as well. So the 'misdemeanor', 'violation', 'felony', and 'sales' features were dropped, as NOT doing so would introduce leakage to the classification models.
    b. Although violation sales ('viol_sales') started as a target feature, no cases were designated as violation sales as there is no legal category of violation cannabis sales in New York City. So it was also dropped from the feature set.
    c. 'CMPLNT_NUM' was kept at this juncture as it was needed to label all cases as cannabis crimes for the classification model of cannabis vs. non-cannabis crimes (see below). Also, dropping it before concatenating the cannabis crimes with non-cannabis crimes for this model was shown to cause false positive duplicates in earlier versions of the notebook.
    d. The 'PD_DESC', 'PD_CD', and 'LAW_CAT_CD' features were superfluous as they contained the same information as the target features for each model, namely what kind of crime each case is. So to avoid leakage, 'PD_DESC', 'PD_CD', and 'LAW_CAT_CD' were dropped.
    e. 'OFNS_DESC' was also deemed to be superfluous, as it is another way of describing type of crime. Using 'OFNS_DESC' in the classification models will introduce leakage, as it contains the information that will be predicted by them. Also, all cases were either labeled as 'DANGEROUS DRUGS' or 'MISCELLANEOUS PENAL LAW'.
    f. 'KY_CD' was also dropped, as it duplicates the information stored in the target features and will introduce leakage to the classification models. 'KY_CD' is the numeric version of the string feature 'LAW_CAT_CD', as described in the NYPD's data dictionary.
    g. 'Lat_Lon' was dropped, as separate features for Latitude and Longitude were already in the DataFrame.
    h. 'X_COORD_CD' and 'Y_COORD_CD' were superfluous as they contained the same geo-coordinate information as latitude and longitude under a different data convention, so they were dropped as well.
    i. The 'CMPLNT_FR_DT', 'CMPLNT_FR_TM', and 'date_time_start' features were all dropped as they contained the same date and time information contained in the

'start_year', 'start_month', 'start_day', 'start_hour', 'start_minute', and 'start_seconds' features. In the case of 'date_time_start', it was a datetime formatted feature unable to be processed by the machine learning classifiers.

j. The 'rpt_cmplnt_dt_match' feature was kept, but its parent feature 'RPT_DT' was dropped so as not to duplicate 92.8% of the information found in the 'start_year', 'start_month', and 'start_day' features.
k. 'CRM_ATPT_CPTD_CD', the feature which stores information on whether a crime was completed or attempted, essentially contains the same information that exists in the target feature, i.e. a crime was completed. So it was dropped.
l. Victim info would not necessarily be available at the time of trying to predict whether a new crime is a cannabis crime or a non-cannabis crime, because not every cannabis crime has a victim and therefore victim age, race, and sex. Therefore, 'VIC_AGE_GROUP_cleaned', 'VIC_RACE_cleaned', and 'VIC_SEX_cleaned' was dropped.

27. A separate DataFrame (‘nyc_cann_ml_alt’) was created for machine learning classification of cannabis crimes and non-cannabis crimes through making a copy of the 'nyc_cann_ml' DataFrame. Because this round of classification won't need to differentiate between types of cannabis crime, and for easy merging with the cleaned non-cannabis crime DataFrame created in a separate notebook (see below), the cannabis crime type target features, along with 'possession', were dropped.
28. For the machine learning classification model differentiating cannabis crimes from non-cannabis crimes, a label feature was needed that pre-labels cannabis crimes and non-cannabis crimes. This feature called 'cannabis_crime' was created by assigning all rows in the ‘nyc_cann_ml_alt’ DataFrame with a ‘cannabis_crime’ value of 1. As the 'nyc_cann_ml_alt' DataFrame only contains cannabis crimes, and all cases have a positive integer complaint number ('CMPLNT_NUM' variable), assignment of '1' for the feature 'cannabis_crime' was implemented with a simple condition of the ‘CMPLTN_NUM’ being positive.
29. The cleaned machine learning DataFrame of just the universe of cannabis crimes (‘nyc_cann_ml’), and the cleaned machine learning DataFrame of the universe of cannabis crimes with the ‘cannabis_crime’ flag feature (‘nyc_cann_ml_alt’), were exported to separate .csv files. Both of these files had 220,304 cases. The ‘nyc_cann_ml’ DataFrame was used for the classification models differentiating cannabis possession crimes from cannabis sales crimes and the five models differentiating each of the five cannabis crime types from each other. The ‘nyc_cann_ml_alt’ DataFrame was used for the classification model differentiating cannabis crime from all other crimes, after concatenation with a DataFrame of non-cannabis crimes.

The non-cannabis crime DataFrame was created in a separate Jupyter notebook. This Jupyter notebook followed the exact same cleaning protocol as the cannabis crime DataFrame detailed above, except for the following differences:

1. A DataFrame was subsetted from the original NYPD dataset which only included non-cannabis crimes, i.e. crimes that did not have a police crime code (‘PD_CD’) of 566-570.
2. The ‘date_time_end’ feature derived from the ‘CMPLNT_TO_DT’ and ‘CMPLNT_TO_TM’ made clear that there were three cases that needed to be dropped because they had

erroneous values later than the circumscribed year range of 2006-2018. These three cases were dropped.

3. For non-cannabis crimes, there were 129 geographic outlier cases that occurred outside of the latitude/longitude range of the five boroughs of New York City. The frame of this study is crimes that occurred inside New York City between 2006-2018, so these outliers had to be dealt with. For situations where a significant proportion of cases are geographic outliers, their latitude/longitude values would be trimmed to the maximum or minimum point of the range, whichever is nearest. However, because of the vanishingly small proportion of outliers (129 out of 6,238,620 cases), removing these cases would not bias the results and they were dropped from the DataFrame.
4. A random sample of cases of equal size to the cleaned cannabis crime universe (n=220,304) was taken from the non-cannabis DataFrame. PD code and law category codes were compared between the cleaned universe of non-cannabis crimes and its random sample, to ensure that there was not an oversampling of any specific crime type that could bias the results later. There was no such oversampling.
5. The distance to the closest subway entrance feature was derived after the random sample was taken because of a shortage of computational resources. In other words, the analyst's computer was not up to the task of computing the values of this feature for approximately 6.2 million rows.
6. An EDA version of the sample DataFrame with the categorical features intact was exported to a csv file for use in the Data Story & Exploratory Data Analysis section below.
7. A value of '0' was assigned to the derived 'cannabis_crime' feature, so that all non-cannabis crimes in the sample were flagged as being non-cannabis crimes.
8. A .csv file named 'nyc_non_cann_for_ML' was exported and then concatenated with the 'nyc_cann_ml' DataFrame using the pandas.concat() method. The presence of duplicate rows was checked for with the pandas.drop_duplicates() method, and there were none.
9. The presence of null values was checked with the chained Pandas methods of .isnull().sum() and .isna().sum(). The features with null values were put into a list called 'fill_w_zero', and then any null values for these features were filled with zeroes. This was needed as there were many housing developments, transit stations, and city parks that exist in the cannabis crime DataFrame but not the non-cannabis DataFrame, or vice versa.
10. The 'CMPLNT_NUM' feature was dropped from the concatenated DataFrame of cannabis and non-cannabis crimes because it was not needed any longer, and would likely introduce leakage to the classification model. The leakage would occur as each record had a unique value for 'CMPLNT_NUM', which would create absolute overfitting if 'CMPLNT_NUM' was used as a predictive feature.
11. The concatenated DataFrame was finally exported to a .csv file for use in the machine learning classification model of differentiating cannabis crimes from non-cannabis crimes.

Exploratory Data Analysis (EDA) & Data Story: Descriptive Statistics, Data Visualizations, Covariance Matrices, and Correlations Between the Target Variables and the Feature Set

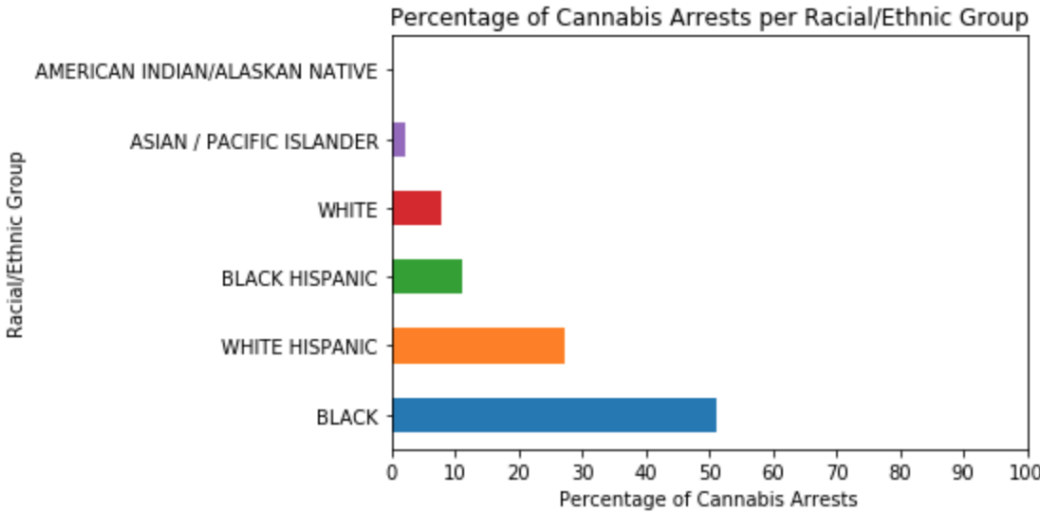
A comprehensive exploratory data analysis (EDA) of the cannabis crime DataFrame with categorical features intact was conducted, which involved a series of visualizations designed to investigate the distribution of cannabis crime across demographic groups and geographic indicators. The combined cannabis 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:

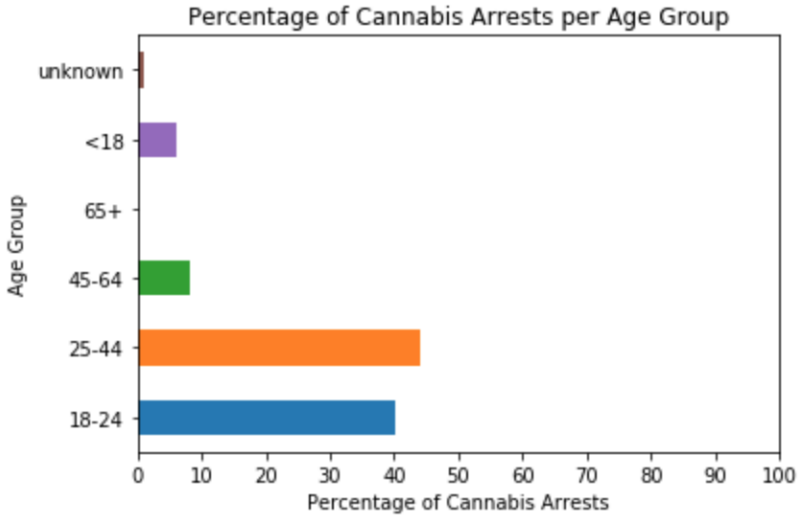
https://github.com/danloew/SpringboardFirstCapstone/blob/master/DataStory_EDA_Final.ipynb

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.

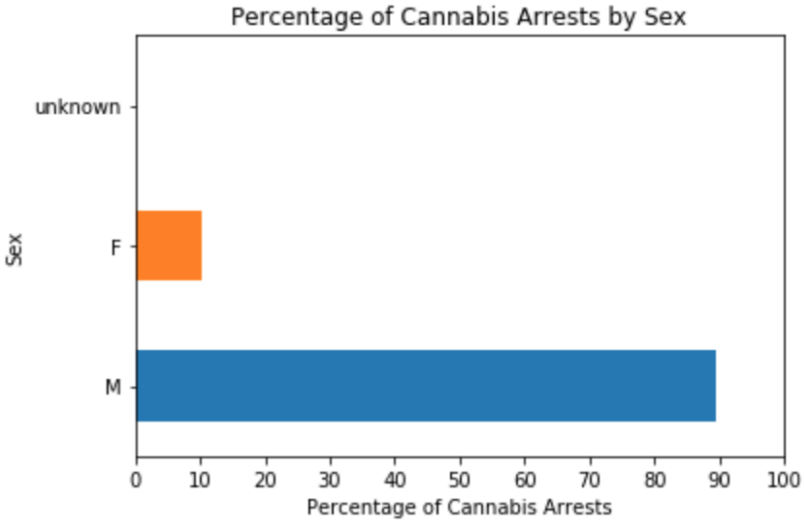
Although the cannabis crime suspects' race, sex and age data were very partial due to the NYPD's under-reporting, a DataFrame was loaded of just the cases where the suspects' race was reported in order to look at demographic distributions. Amongst the 34,837 cases in this DataFrame, 51% of cannabis arrests 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.

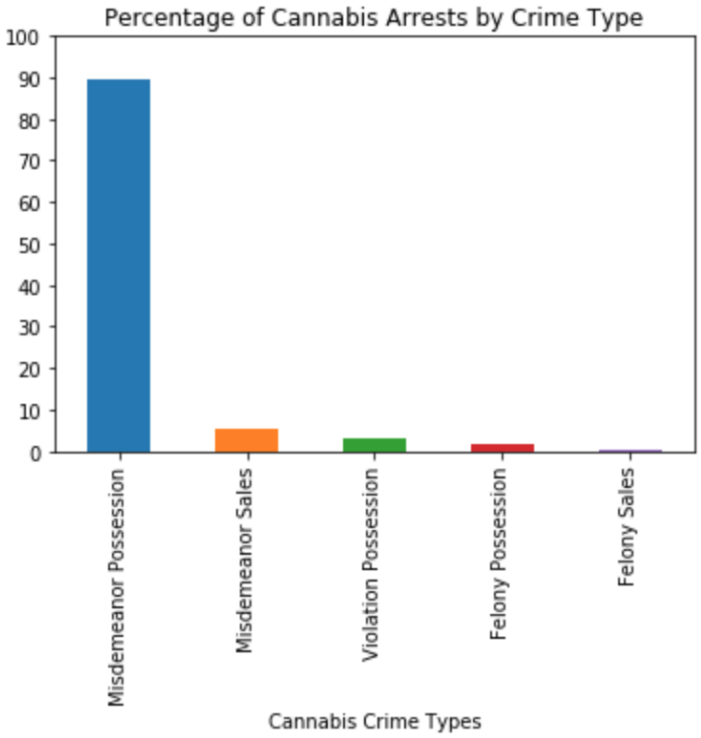


90% were of males.



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 cannabis arrests in New York City between 2006 and 2018 where the suspect's demographics were recorded 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.

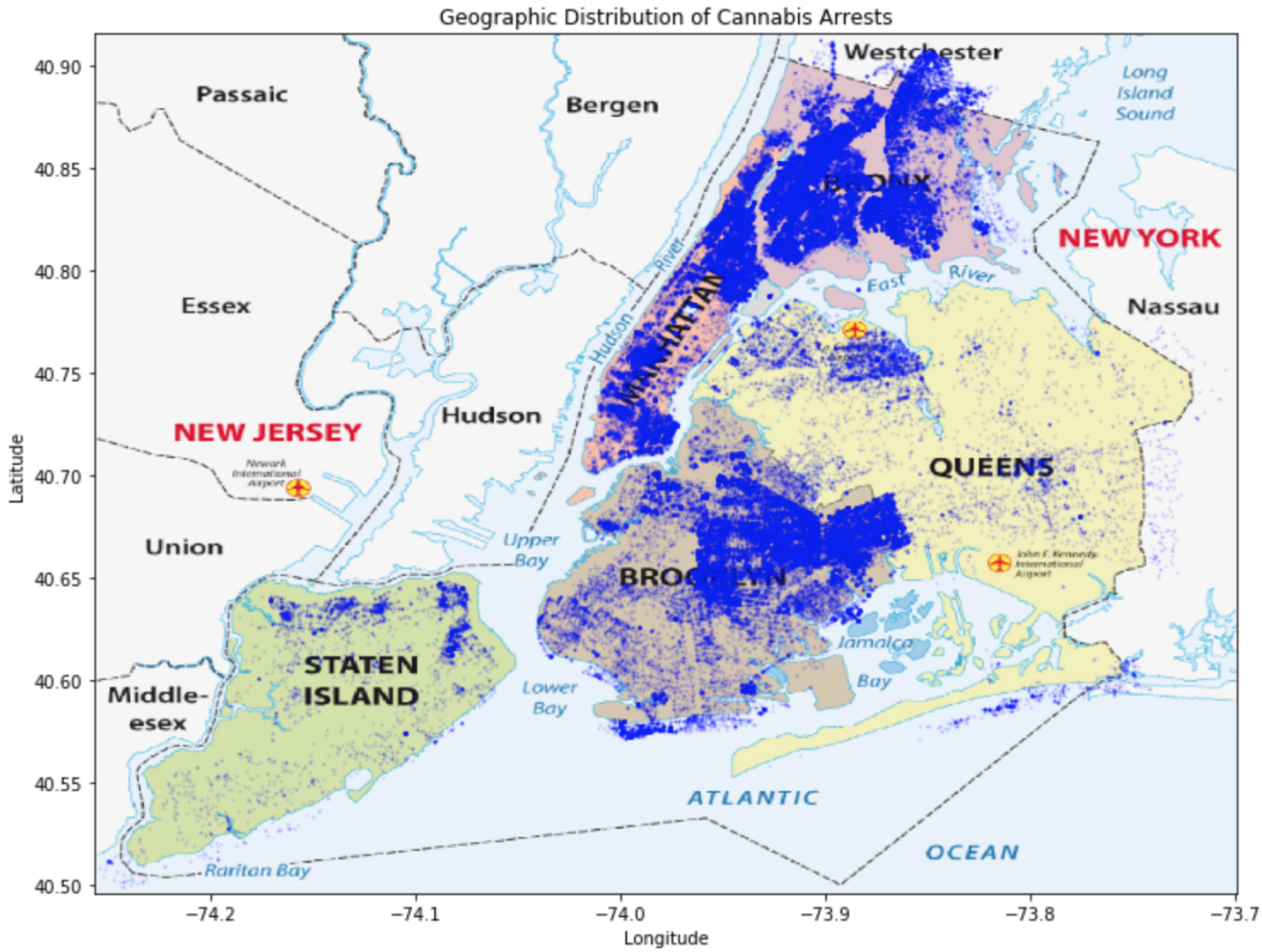
One of the striking things about cannabis arrests in New York City is that 92.6% of them are for simple misdemeanor and violation possession charges, which is the vast majority. 1.7% are for felony-level possession, 5.2% are for misdemeanor sales, and 0.5% are for felony sales, the latter being arguably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



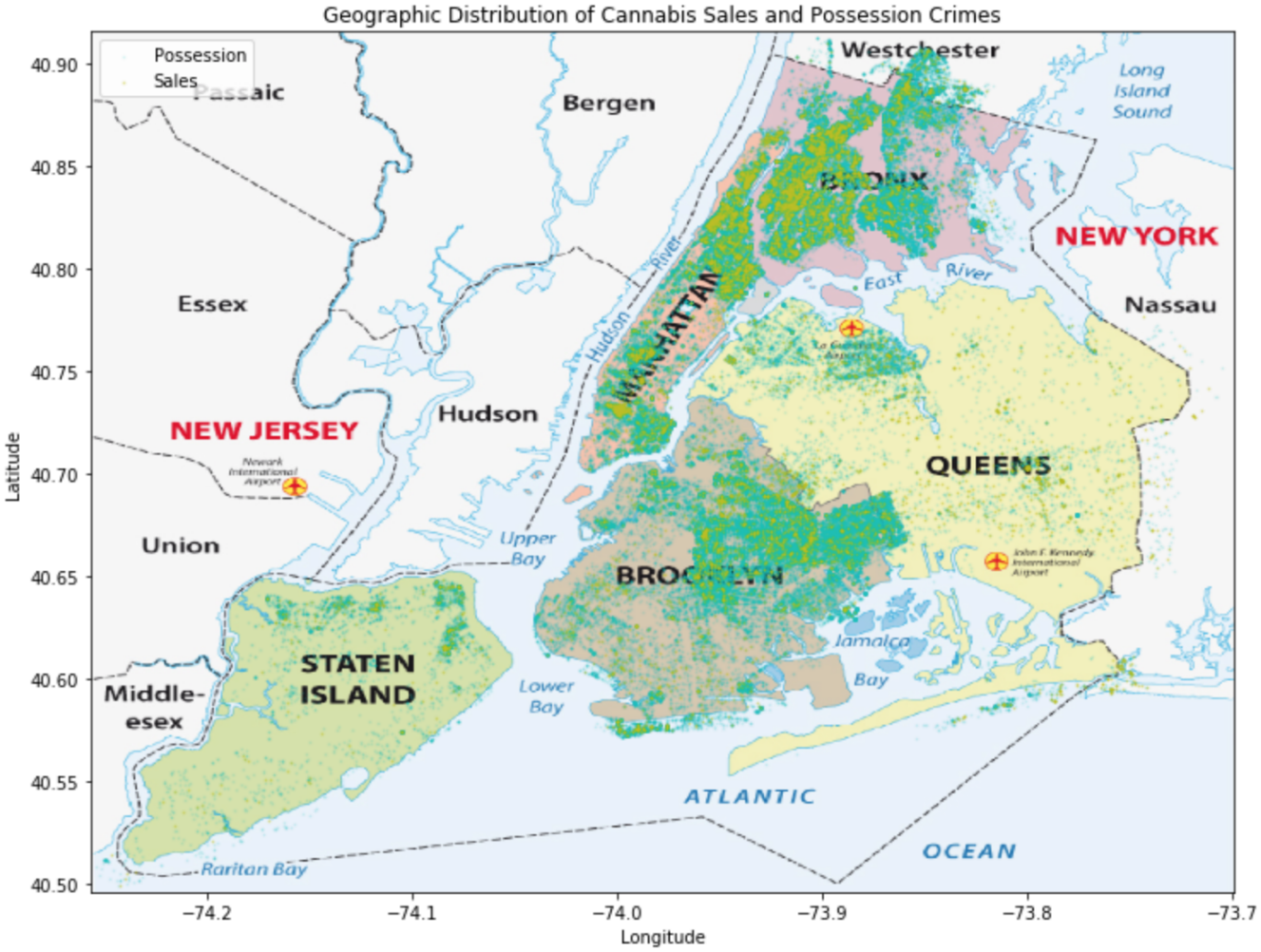
The same racial disparity described above holds true for all levels of cannabis crime except for the following differences. More violation possession arrests were made of white perpetrators than of black Hispanic perpetrators, but the difference was only 3%. Also, it should be noted that violation possession 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.

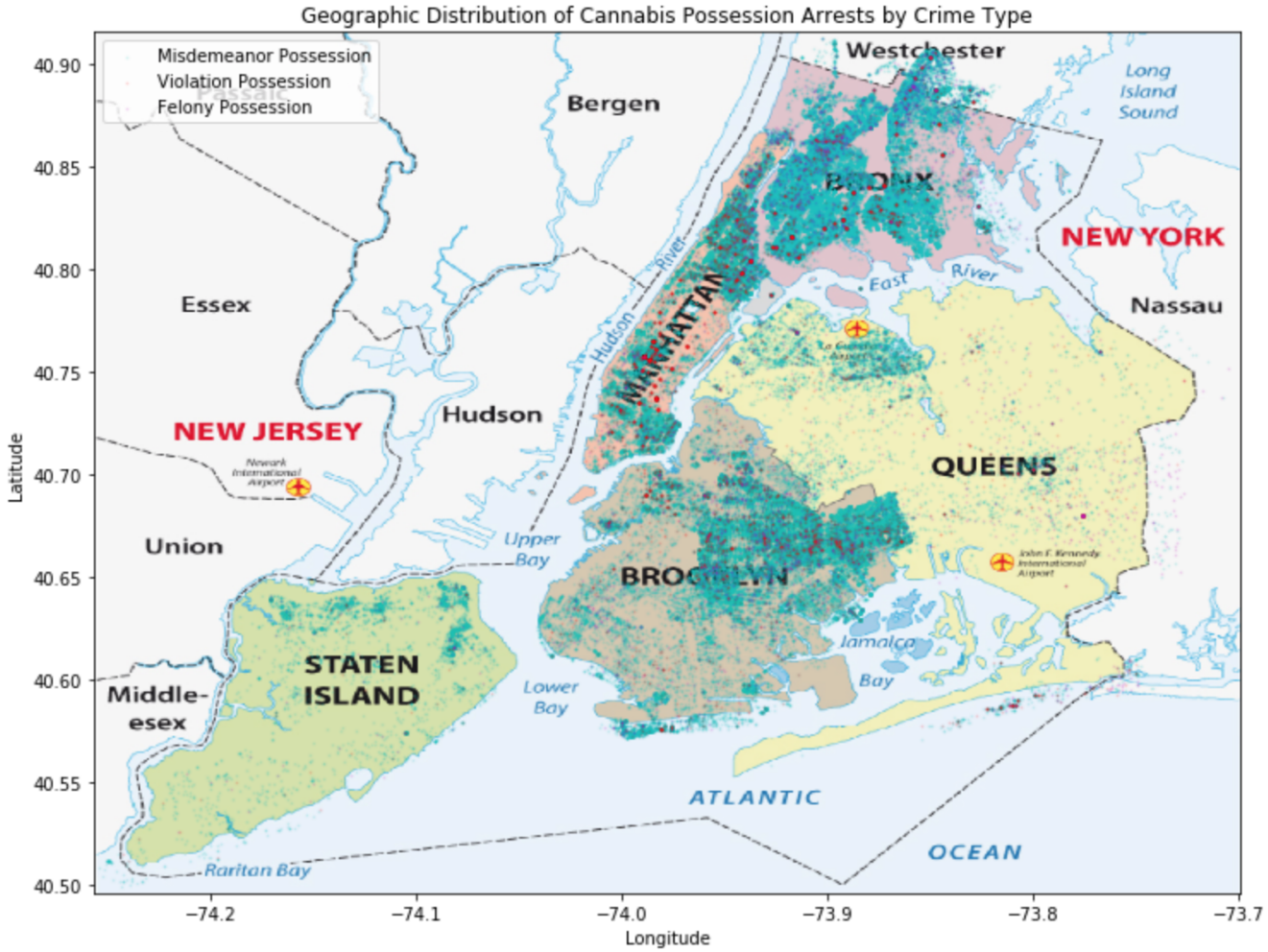
The following visualizations show the geographic distribution of 1) cannabis arrests as a whole, 2) cannabis possession and sales arrests, 3) arrests for cannabis possession and its three types, and 4) arrests for cannabis sales and its two types. Please note that the overlay of the latitude/longitude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

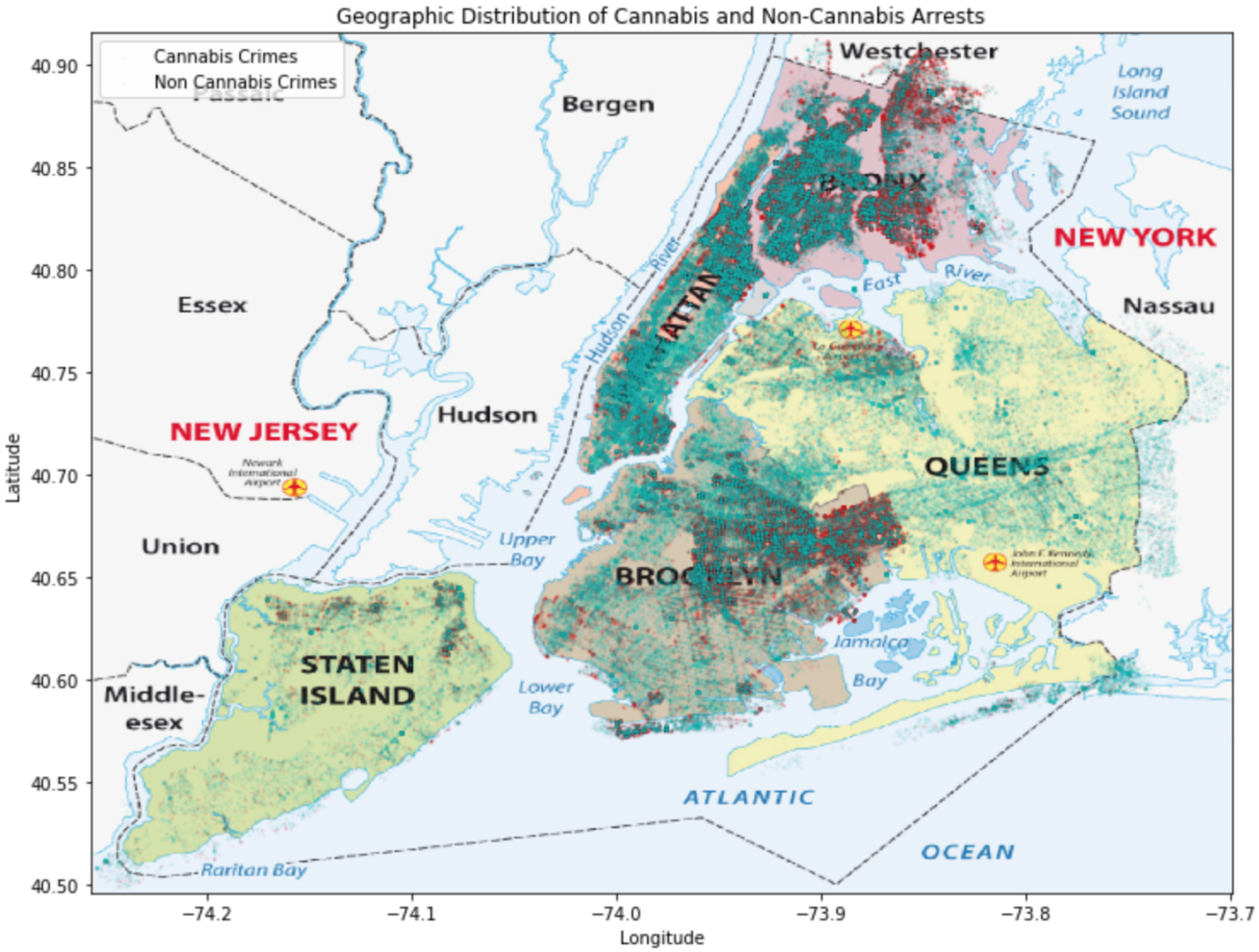


Misdemeanor possession arrests are displayed in cyan, violation possession arrests are displayed in red, and felony possession arrests are displayed in magenta.

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.

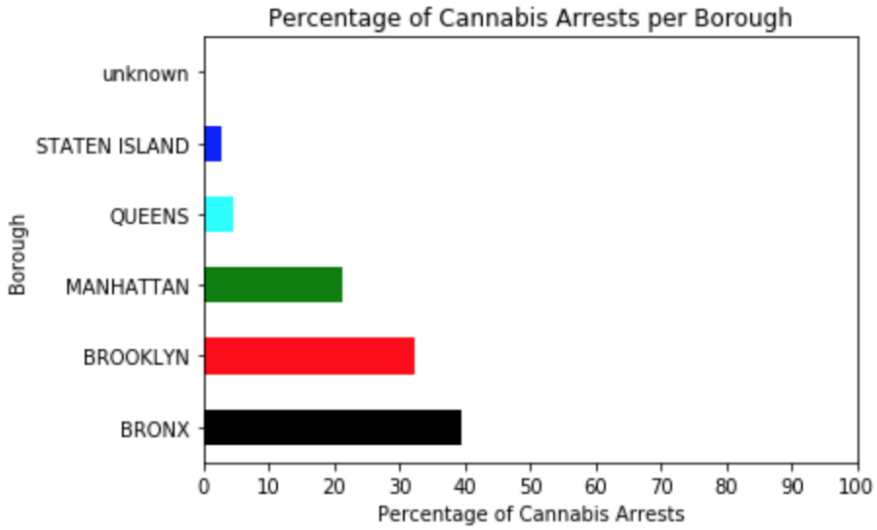
Misdemeanor sales arrests are displayed in yellow and felony sales arrests are displayed in black. One can see that these arrests tend to fall within the same neighborhoods as possession arrests, but are obviously much more sparse.

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).

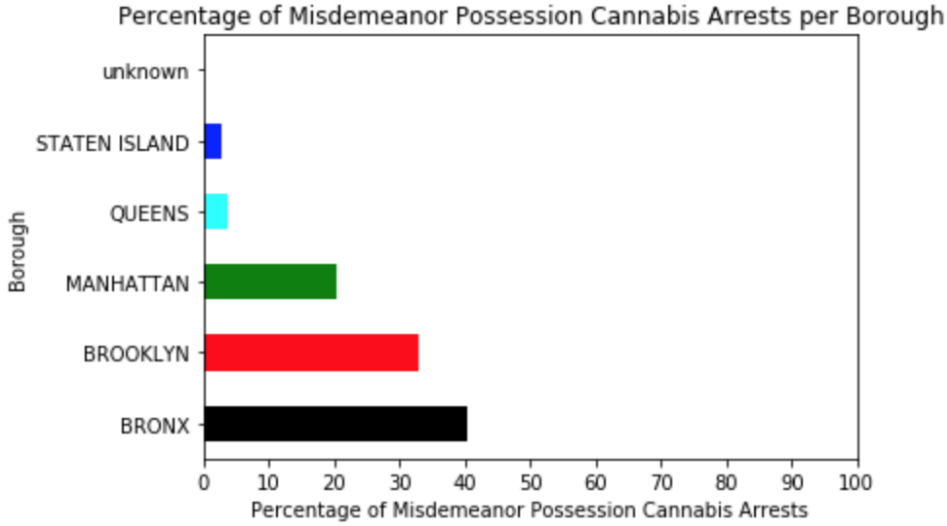
The most abstract geographic unit of New York City is the borough. The Bronx and Brooklyn are home to the majority of cannabis crimes overall, as shown in the following bar chart.

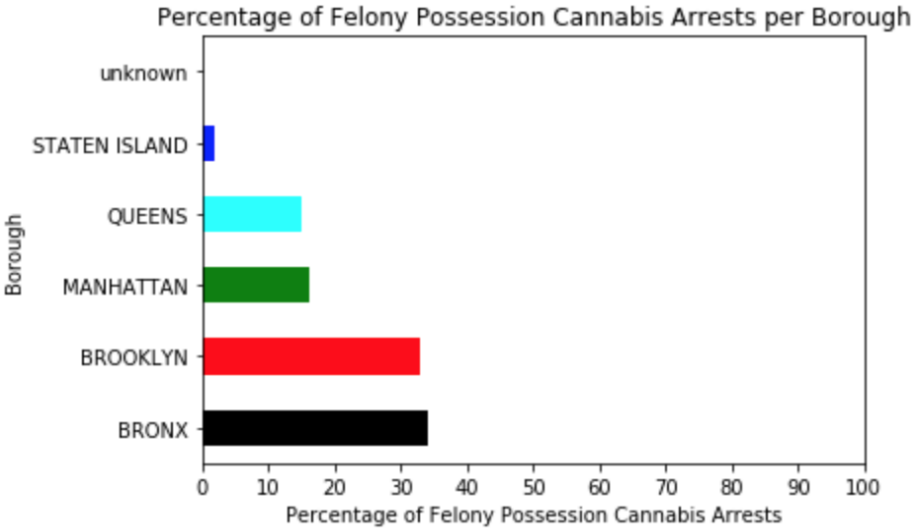


This clear disparity in cannabis arrests by borough is interesting because of the racial demographics of these two boroughs. Looking at the Census data estimated from 2018 and 2019 shows us the following racial and ethnic breakdown by borough. It is important to remember that Hispanic/Latino status is 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.

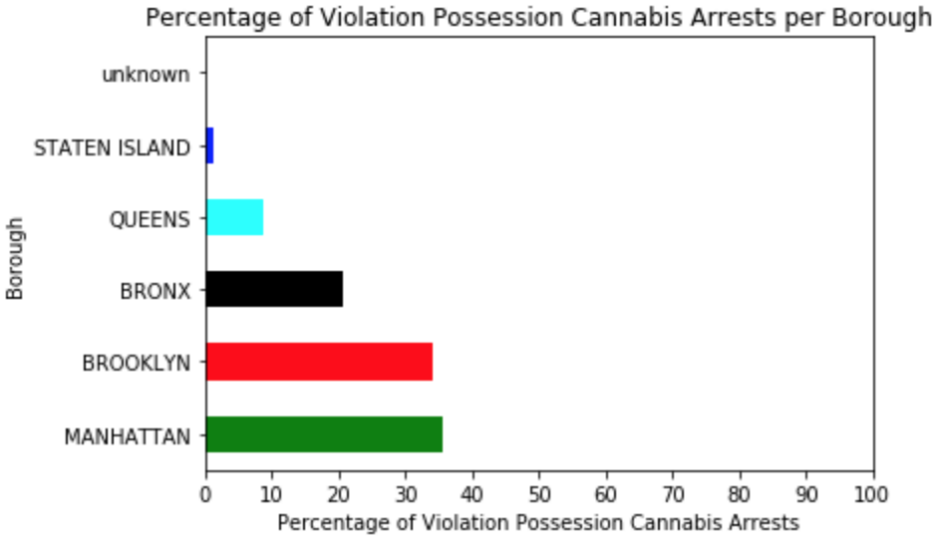
The Bronx's populace is 44% African-American, 56% Latino/Hispanic, 5% Asian and only 9% non-Latino White, while Brooklyn's populace is 34% African-American, 19% Latino/Hispanic, 13% Asian, and 36% non-Latino White. By contrast, Manhattan's populace is 18% African-American, 26% Latino/Hispanic, 13% Asian, 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.

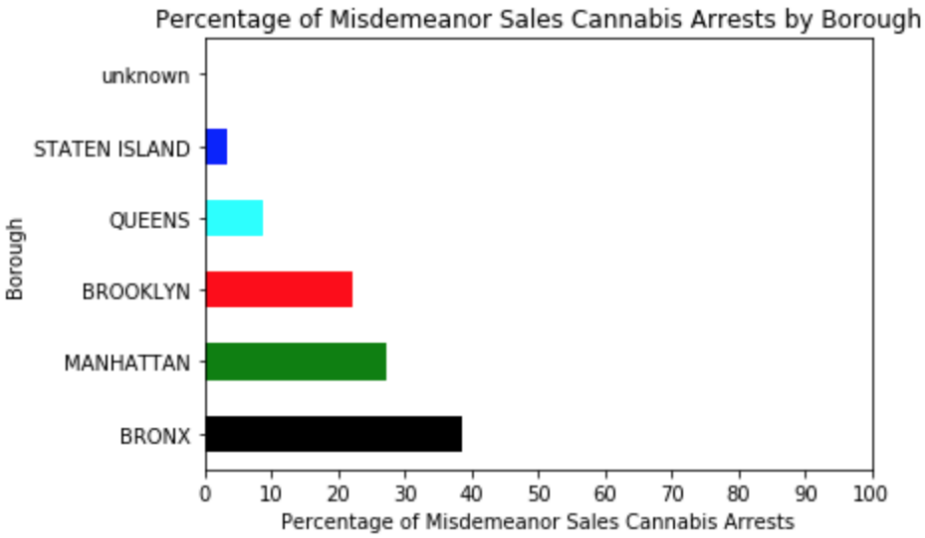


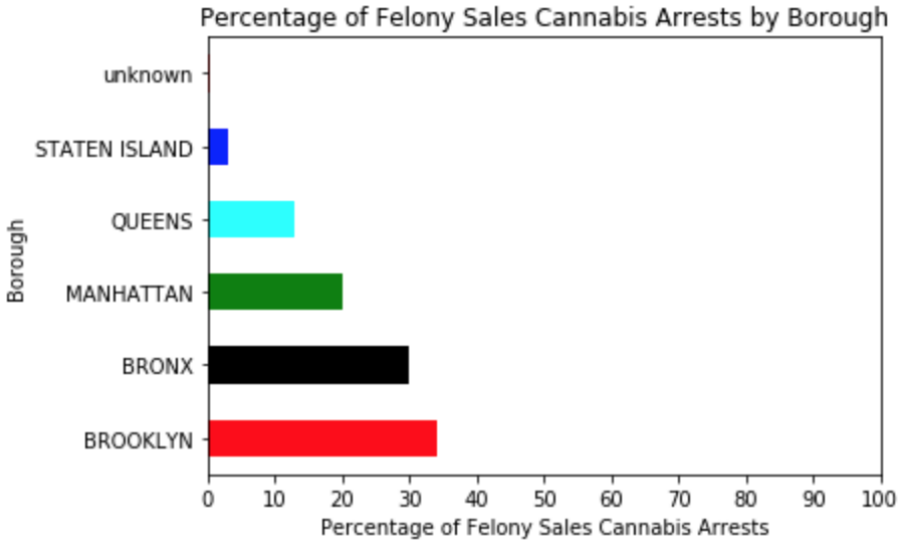


Violation possession charges (the lowest level of cannabis crime) are dominant in Manhattan. This supports the idea that cannabis crimes are punished very differently in New York City depending on which part of the city the crime takes place in.



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





It would be interesting to see which neighborhoods of New York City 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.

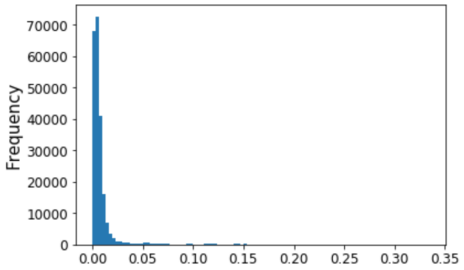
For misdemeanor sales, Greenwich Village and the West Village (in Manhattan), and Western Harlem are also common.

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.

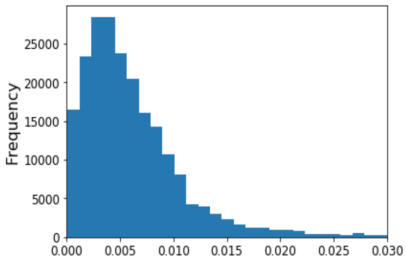
In the data cleaning notebooks, the latitude/longitude distance from each crime to the nearest NYC subway station entrance was calculated as a contributive predictor to the feature sets used in the classification models, as subway station entrances can often be where crime can occur. As a visual exploration, 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

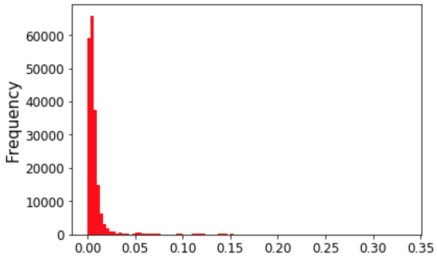
differences in greater detail. The 'L1' distance is used instead of the 'L2' distance because realistically, New Yorkers can't move through the city and engage in cannabis crime "as the crow flies".



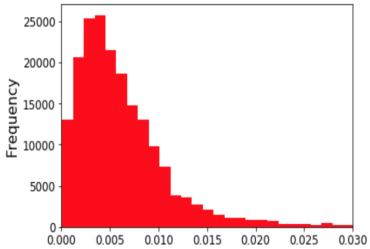
L1 Distance of Cannabis Crimes to Closest Subway Station Entrance



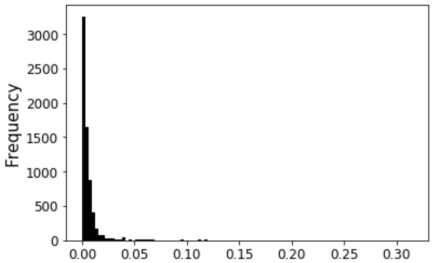
L1 Distance of Cannabis Crimes to Closest Subway Station Entrance magnified



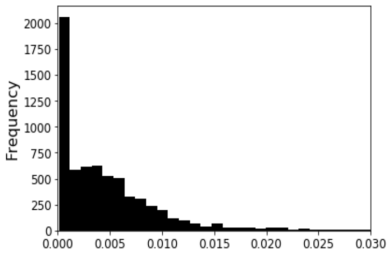
L1 Distance to Closest Subway Station Entrance for Misdemeanor Possession



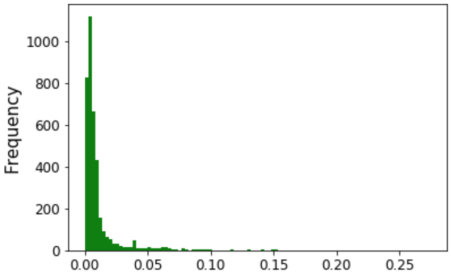
L1 Distance to Closest Subway Station Entrance for Misdemeanor Possession, magnified



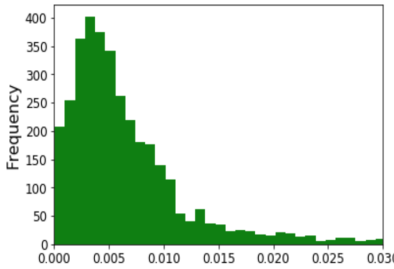
L1 Distance to Closest Subway Station Entrance for Violation Possession



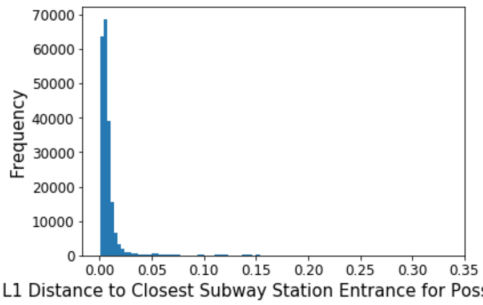
L1 Distance to Closest Subway Station Entrance for Violation Possession, magnified



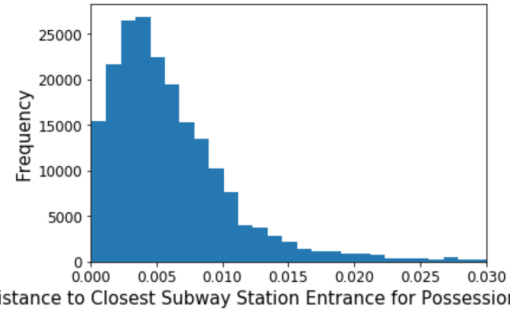
L1 Distance to Closest Subway Station Entrance for Felony Possession



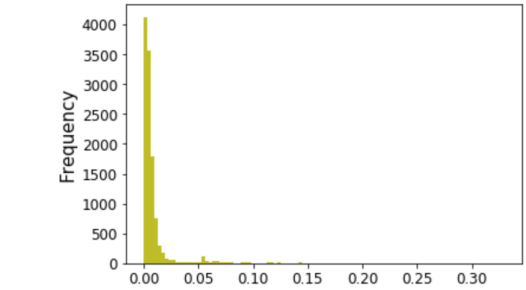
L1 Distance to Closest Subway Station Entrance for Felony Possession, magnified



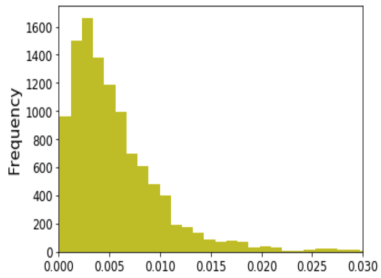
L1 Distance to Closest Subway Station Entrance for Possession



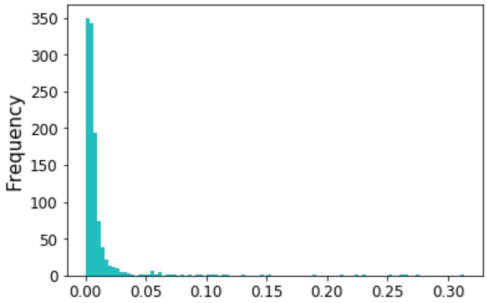
L1 Distance to Closest Subway Station Entrance for Possession, magnified



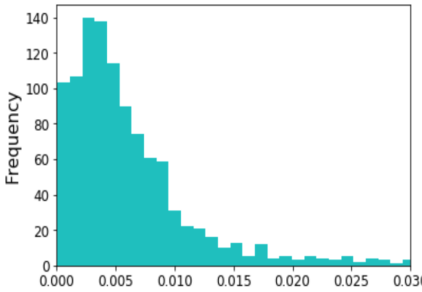
L1 Distance to Closest Subway Station Entrance for Misdemeanor Sales



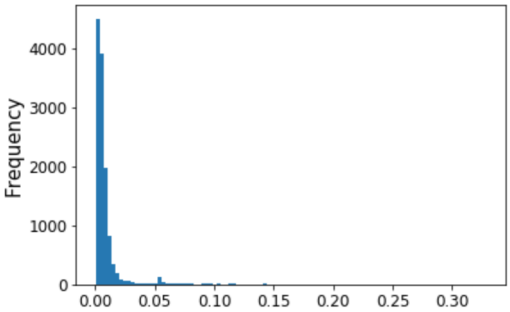
L1 Distance to Closest Subway Station Entrance for Misdemeanor Sales, magnified



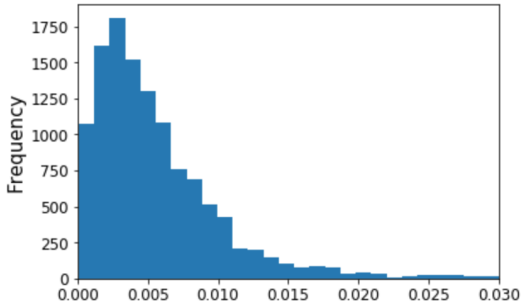
L1 Distance to Closest Subway Station Entrance for Felony Sales



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

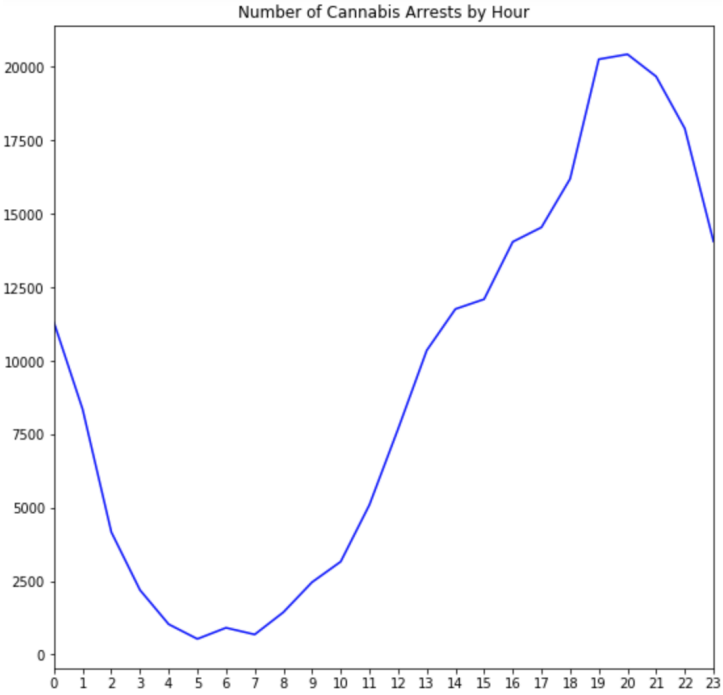
degree the N.Y. Transit Police, with 33% of violation possession arrests falling under this jurisdiction. The fact that 19% of cannabis arrests fell under the jurisdiction of the NYCHA shows how heavily policed these public housing projects were.

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.

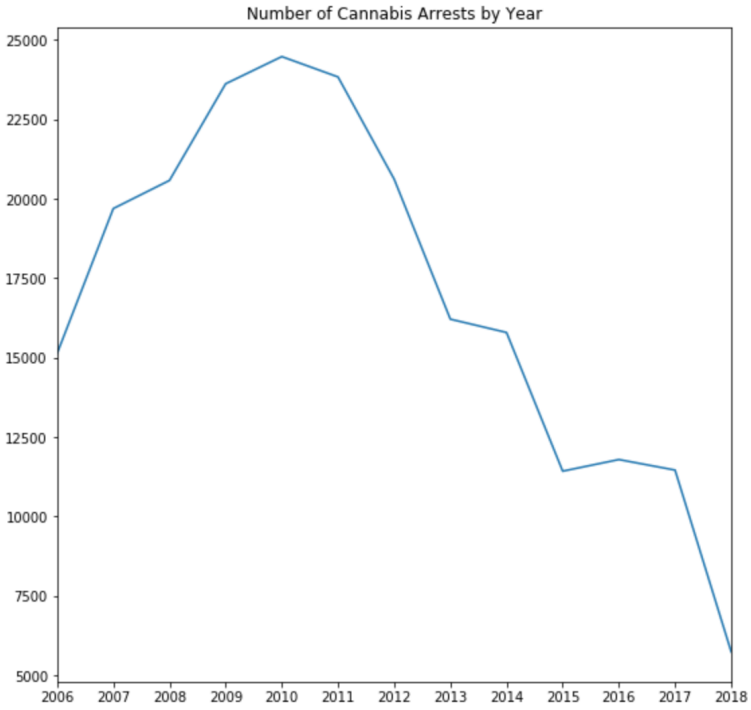
Cannabis arrests occurred more frequently during certain times of the day. 39% occur during the daytime (6 am - 6 pm), and 61% occur during the nighttime (6 pm - 6 am). The work day (9 am - 6 pm) composes most of the daytime arrests, and 37.5% of the total. Early morning (6 am - 7:30 am) and the morning 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.

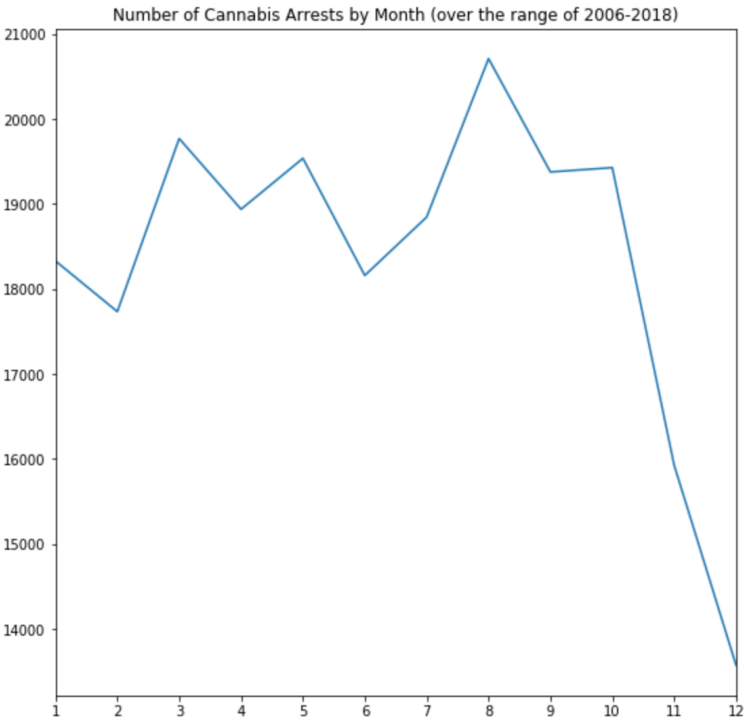
To easily visualize the number of arrests made per hour (across all years), the hour of the day is extracted and then displayed in the following line plot (using the 24 hour convention with ‘0’ indicating midnight).



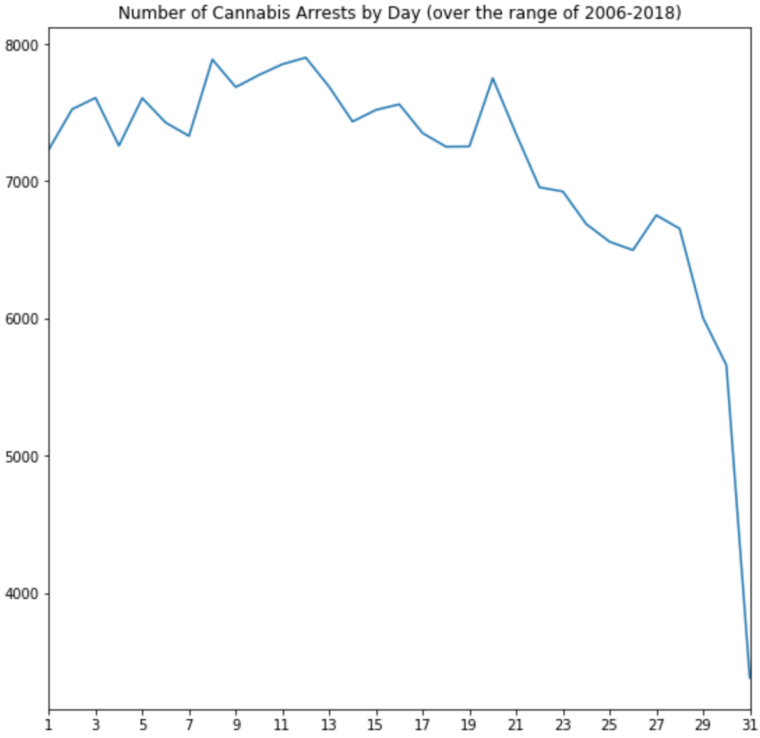
It has been well reported that during Mayor Bloomberg's time as mayor, cannabis arrests reached their peak. One can see that 2006 had 15,127 arrests, and that this increased to 24,468 arrests in 2010. This held fairly steady for 2011 (23,827), dropped a bit in 2012 (20,611) as criticism of Bloomberg's "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.

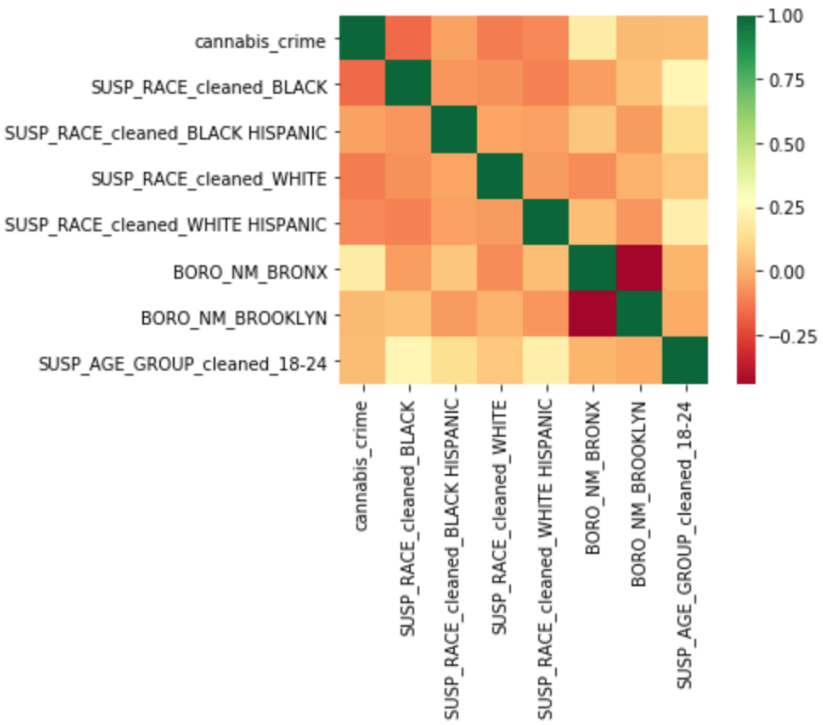


Each day of the month over the entire range of 2006 to 2018 had a fairly consistent number of cannabis arrests, ranging from 5,660 to 7,900 arrests a day . The number dropped somewhat in the last 10 days of the month. The 31st had roughly half the arrests as the rest of the month because not every month 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 664, 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.

Exploratory Data Analysis (EDA): Inferential Statistics Using Hypothesis Testing

In this second phase of EDA, inferential statistics were employed to formally test a series of hypotheses concerning the nature of cannabis crime in New York City between 2006 and 2018. These hypotheses concerned whether cannabis crimes were equally likely to have their suspects' demographic information recorded by the arresting officer, whether members of different demographic groups were equally likely to be arrested for cannabis, whether members of different demographic groups were equally likely to be arrested for different types of cannabis crime, and whether cannabis arrests were equally likely to occur in the different boroughs of New York City. The classic t-test was chosen in order to test these hypotheses by looking at whether there is a significant difference between the means of the two groups used in each of these tests, and therefore whether the differences seen were due to chance or not. Those null hypotheses that were rejected show that the difference seen between their two means are not due to chance, and that there was a confounding reason for this difference (that is outside of the scope of this project). The code for this series of hypothesis tests is available at:

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

The DataFrame used in this hypothesis testing stage was a 10% sample from the cleaned "NYPD Complaint Data Historic" dataset. As specified above under the “Data Cleaning and Wrangling” section, a 10% sample was randomly selected and exported to a .csv file inside both the cannabis crime and non-cannabis crime data cleaning notebooks. These samples still had their categorical features, as they were drawn before the binarization of categorical features done in preparation for machine learning. As the data cleaning for cannabis crimes was conducted separately from all other crimes, two .csv files were imported and concatenated in the “Statistical Data Analysis” notebook.

Null values were checked for, which were found only in the sample of non-cannabis crimes. These null values were in the features that specified the type of cannabis crime (‘misd_poss’, etc.), as the sample of non-cannabis crimes did not have these features created during its data cleaning phase. These null values were filled with zeroes, which was appropriate due to the fact that non-cannabis crimes by definition could not have values for features having to do with types of cannabis crime.

A cannabis crime flag feature named ‘cannabis_crime’ was created using the five requisite police codes (‘PD_CD’), wherein cannabis crimes were flagged with a ‘1’ and non-cannabis crimes were flagged with a ‘0’. For ease of use in running hypothesis tests, separate DataFrames for cannabis crimes ('cann') and non-cannabis crimes ('non_cann') were created.

Hypothesis testing using t-tests uncovered the following findings:

1. The difference seen between the percentage of cannabis crimes where the suspect's race was reported (15.8%) and the percentage of non-cannabis crimes where the suspect's race was reported (38.1%) was not due to chance, and that there was some mediating factor behind this difference. The t-score was approximately 67.6 and the p-value was 0.0. The mediating factor is beyond the scope of this analysis and is an area for future research.
2. African-Americans, Whites, Hispanic Whites, Hispanic African-Americans, and Asians and Pacific Islanders arrested for a crime were not equally likely to be arrested for cannabis crimes as they were for non-cannabis crimes, and that they were more likely to be arrested for non-cannabis crimes. The t-scores and p-values for each demographic group are as follows:
    a. African-Americans: T-score of approximately 41.3 and p-value of 0.0.
    b. Whites: T-score of approximately 27.9 and p-value of approximately 1.8e-171
    c. Hispanic Whites: T-score of approximately 23.7 and p-value of 2.8e-124
    d. Hispanic African-Americans: T-score of approximately 8.9 and p-value of 5.1e-19
    e. Asians: T-score of approximately 14.7 and p-value of 4.1e-49
3. Out of experimental curiosity, DataFrames of just those cases where the suspect’s race was reported were subsetted from both the cannabis and non-cannabis DataFrame. For only those crimes where the suspect’s race was reported, African-Americans arrested for a crime were equally likely to be arrested for cannabis crimes as they were for non-cannabis crimes. The t-score was approximately -1.54 and the p-valuewas0.12.
4. African-Americans arrested for a crime were not equally likely to be charged for cannabis crimes as White people arrested for a crime; they were significantly more likely with a t-score of approximately 34.0 and a p-value of 1.8e-249.
5. Hispanic Whites arrested for a crime were not equally likely to be charged for cannabis crimes as White people arrested for a crime; they were significantly more likely (but to a

lesser degree than with African-Americans). The t-score was approximately 18.4 and the p-value was 2.0e-75.

6. Hispanic African-Americans arrested for a crime were not equally likely to be charged for cannabis crimes as White people arrested for a crime; they were significantly more likely (but to a lesser degree than with African-Americans and Hispanic Whites). The t-score was approximately 2.7 and the p-value was 0.01.
7. Asians arrested for a crime were not equally likely to be charged for cannabis crimes as Whites arrested for a crime; they were significantly less likely. The t-score was approximately -11.0 and the p-value was 6.6e-28.
8. African-Americans arrested for a crime were not equally likely to be charged for cannabis crimes as Hispanic Whites arrested for a crime; they were significantly more likely. The t-score was approximately 17.1 and the p-value was 1.9e-65.
9. African-Americans arrested for a crime were not equally likely to be charged for cannabis crimes as Hispanic African-Americans arrested for a crime; they were significantly more likely. The t-score was approximately 31.9 and the p-value was 2.9e-220.
10. White Hispanics arrested for a crime were not equally likely to be charged for cannabis crimes as Hispanic African-Americans arrested for a crime; they were significantly more likely. The t-score was approximately 16 and the p-value was 2.3e-57.
11. African-Americans arrested for a cannabis crime were not equally likely to be charged for misdemeanor cannabis possession as they were for violation cannabis possession; they were significantly more likely to be arrested for misdemeanor possession (which carries more severe legal consequences than a violation possession charge). The t-score was approximately 10.6 and the p-value was 4.9e-26.
12. Whites arrested for a cannabis crime were equally likely to be charged for misdemeanor cannabis possession as they were for violation cannabis possession. The t-score was approximately 1.5 and the p-value was 0.1.
13. African-Americans arrested for a cannabis crime were equally likely to be arrested for violation possession as were Whites arrested for a cannabis crime. This suggests that violation possession charges were not charged differently among African-American and White suspects. The t-score was approximately 0.9 and the p-value was 0.4.
14. African-Americans arrested for a cannabis crime were not equally likely to be arrested for misdemeanor possession as they were for felony possession; they were more likely to be arrested for misdemeanor possession. The t-score was approximately 11.1 and the p-value was 3.7e-28.
15. African-Americans arrested for a cannabis crime were not equally likely to be arrested for misdemeanor sales as they were for felony sales; they were more likely to be arrested for misdemeanor sales. The t-score was approximately 2.5 and the p-value was 0.01.
16. Cannabis arrests were not equally as likely to happen in the five boroughs. The Bronx was the most likely, Brooklyn was the second most likely, Manhattan was the third, Queens was the fourth, and Staten Island was the fifth. The t-scores and p-values for each borough-related hypothesis test result were as follows:
    a. Cannabis arrests were not equally as likely to happen in the Bronx as they were in Manhattan; they were significantly more likely to happen in the Bronx. The t-score was approximately 42.2 and the p-value was 0.

b. Cannabis arrests were not equally as likely to happen in the Bronx as they were in Brooklyn; they were significantly more likely to happen in the Bronx. The t-score was approximately 14.9 and the p-value was 3.1e-50.
c. Cannabis arrests were not equally as likely to happen in the Bronx as they were in Queens; they were significantly more likely to happen in the Bronx. The t-score was approximately 98.5 and the p-value was 0.
d. Cannabis arrests were not equally as likely to happen in the Bronx as they were in Staten Island; they were significantly more likely to happen in the Bronx. The t-score was approximately 106.1 and the p-value was 0.
e. Cannabis arrests were not equally as likely to happen in Manhattan as they were in Brooklyn; they were significantly more likely to happen in Brooklyn. The t-score was approximately -27 and the p-value was 4.8e-159.
f. Cannabis arrests were not equally as likely to happen in Manhattan as they were in Queens; they were significantly more likely to happen in Manhattan. The t-score was approximately 55.2 and the p-value was 0.
g. Cannabis arrests were not equally as likely to happen in Manhattan as they were in Staten Island; they were significantly more likely to happen in Manhattan. The t-score was approximately 62.9 and the p-value was 0.
h. Cannabis arrests were not equally as likely to happen in Brooklyn as they were in Queens; they were significantly more likely to happen in Brooklyn. The t-score was approximately 82.2 and the p-value was 0.
i. Cannabis arrests were not equally as likely to happen in Brooklyn as they are in Staten Island; they were significantly more likely to happen in Brooklyn. The t-score was approximately 89.7 and the p-value was 0.
j. Cannabis arrests were not equally as likely to happen in Queens as they were in Staten Island. The t-score was approximately 9.5 and the p-value was 1.7e-21.

These hypothesis tests generally support that there is a racial hierarchy in terms of how frequently different racial groups are arrested for cannabis. From most likely to least likely, this hierarchy is African-Americans, White Hispanics, African-American Hispanics, Whites, and Asians and Pacific Islanders. These tests also support that there is a geographic hierarchy in terms of how frequently cannabis arrests occur in the five boroughs. From most likely to least likely, this hierarchy is the Bronx, Brooklyn, Manhattan, Queens, and Staten Island. Another very intriguing finding that has not been reported in the media is that there is a highly significant difference between the amount of cannabis crime with suspect race unreported and the amount of non-cannabis crime with suspect race unreported that cannot be ascribed to chance.

In-Depth Analysis Using Machine Learning

This project created machine learning classification models with Logistic Regression and Random Forest classifiers in order to predict cannabis arrests in New York City from the pool of all crime types, as well as to predict several sub-types of cannabis arrests from the pool of all cannabis arrests in New York City.

These models were created not only to predict cannabis crime and its subtypes in New York City between 2006 and 2018, but to also identify the most salient predictors of these crimes by looking at the coefficients of the best performing Logistic Regression classifier. Random Forest was also used as a way to nonlinearly predict the cannabis crime classes, but because the Random Forest method does not produce coefficients, it was not used to identify the most salient predictors of cannabis arrests. By identifying the most salient predictors, the biases in arrests can be scientifically identified in order to serve as a deep resource for future research in the areas of civil rights, criminology and sociology. Therefore, this project does not utilize machine learning classification to test one specific hypothesis, but instead uses it to create a descriptive landscape of cannabis crime in New York City in order to highlight arrest bias in all of the available data features provided by the NYPD in their dataset. It should be understood that the word “bias” is being used here in a scientific sense and not a political one, where the analyst is able to identify the features that show that cannabis law was not being enforced in an entirely equitable fashion. Of central importance is the fact that identifying the most salient predictors of cannabis arrests, and therefore biases in making these arrests, cannot elucidate the causes of these biases.

The cleaned DataFrames built off of the "NYPD Complaint Data Historic" dataset (https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i) were used to build a series of seven classification models. The target features of the models were designed as categorical binary features that contained information on the categorical crime type for each case in the cleaned DataFrame, so therefore classification was chosen instead of linear regression. To create the predictive feature set, the cleaned DataFrames utilized all features native to the NYPD's dataset in addition to a series of features derived from these native features (as detailed in the Data Cleaning notebooks). As mentioned above, the central classification method used was Logistic Regression, although Random Forest was also utilized to investigate whether a non-linear classification method could predict cannabis crime and its subtypes more accurately. Random Forest was partially chosen as it has been shown to be very effective with data sets that have a high number of both rows and features.

The seven different classification models classified

- cannabis crimes (class 1) from all other crimes (class 0),
- cannabis possession crimes (class 1) from cannabis sales crimes (class 0),
- misdemeanor cannabis possession crimes (class 1) from all other cannabis crimes (class 0),
- violation cannabis possession crimes (class 1) from all other cannabis crimes (class 0),
- felony cannabis possession crimes (class 1) from all other cannabis crimes (class 0),
- misdemeanor sales crimes (class 1) from all other cannabis crimes (class 0), and
- felony sales crimes (class 1) from all other cannabis crimes (class 0).

The first model identifying the strongest predictors of cannabis arrests in contrast to all other crimes used the cleaned universe of 220,304 NYC cannabis crimes committed between 2006 and 2018 combined with a random sample of 220,304 non-cannabis crimes committed during the same time period. The second model identifying the strongest predictors of cannabis possession in contrast to cannabis sales used the cleaned universe of NYC cannabis crimes during this time period. The third through seventh models identifying the strongest predictors that differentiate each of the five types of cannabis crimes listed above again used the cleaned universe of NYC cannabis crimes.

The best Logistic Regression classifier for each of the seven models was identified by evaluating the accuracy, precision, recall, and F1 scores of a set of Logistic Regression classifiers. Generally speaking, these classifiers were run on scaled and unscaled data (only during the first model classifying cannabis and non-cannabis crimes), and utilized different hyperparameter values for the solver, penalty, and regularization hyperparameter 'C' (during all of the models). The MinMax scaler was used for the scaled data. These classifiers and their accuracy, precision, recall, and F1 scores were stored in a hyperparameter tuning table. The best scoring Logistic Regression classifier was then identified for each of the seven models and further evaluated by examining their Receiver Operating Characteristic (ROC) curves and Precision Recall Curves (PRC).

The Jupyter notebooks for the seven classification models can be found at the following GitHub locations:

Cannabis and Non-Cannabis crimes:

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

Cannabis possession and sales crimes:

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

Misdemeanor cannabis possession from all other cannabis crimes:

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

Violation cannabis possession from all other cannabis crimes:

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

Felony cannabis possession from all other cannabis crimes:

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

Misdemeanor cannabis sales from all other cannabis crimes:

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

Felony cannabis sales from all other cannabis crimes:

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

The machine learning classification pipeline for creating these models followed a logical series of steps. Flow charts of all seven models' pipelines can be found in Appendix A, where idiosyncrasies of each model can be examined. Generally speaking, the pipeline consisted of the following steps:

1. The target feature y and predictive feature set X are defined. Feature selection was already accomplished during the data cleaning process, so all features in the cleaned DataFrame were included in the predictive feature set.
2. A stratified train-test split was carried out with a test size of 0.2, maximum iterations of 6,000, and a tolerance of 0.001. The maximum iterations and tolerance settings were decided upon through repeated experimentation of fitting and predicting, and were where all classifiers converged and accuracy didn't suffer.
3. Random Forest algorithms with 10 and 100 estimators were fitted to the target feature y and predictive feature set X, and accuracy, precision, recall, and F1 scores were returned to identify the Random Forest algorithm with the best predictive power. Generally, Random Forest algorithms outperformed the Logistic Regression algorithms on these classification metrics.
4. A series of Logistic Regression algorithms were fitted to the target feature y and predictive feature set X. Scaled and unscaled data were explored, and different hyperparameter values for the solver, penalty, and regularization hyperparameter 'C' were used in order to uncover the most powerfully predictive models. The Limited-memory Broyden-Fletcher-Goldfarb-Shanno ('lbfgs') and Stochastic Average Gradient descent A ('saga') solver types were used. Both the L2/Ridge Regression and L1/LASSO penalties were experimented with, and a range of 'C' values were used (e.g., 0.001, 0.01, 0.1, 1.0, 10.0, and 100.0). The 'lbfgs' solver was used as it is the default solver type for Sci-Kit Learn's Logistic Regression algorithm and the 'saga' solver was used in order to explore the L1/LASSO penalty, as 'lbfgs' cannot do L1/LASSO. Scaled and unscaled data were experimented with, as the 'saga' solver has been shown to perform better on scaled data.
5. Each algorithm was pickled for later reference.
6. The accuracy, precision, recall, and F1 scores were returned as classification metrics for all algorithms and stored in a hyperparameter tuning table, and the best performing algorithm was identified.

The first model classifying cannabis and non-cannabis crime implemented 36 algorithms to experiment with all possible combinations of unscaled data, scaled data, and hyperparameters. After it was discovered that 14 highest performing algorithms all used scaled data using the MinMax scaler, these algorithms were winnowed down in the model that classified cannabis possession and sales crime to 18 algorithms which only used scaled data but still tried all possible combinations of hyperparameters. The five highest performing algorithms from the second model were then tried for the third through seventh models classifying the five subtypes of cannabis crime, which also only used scaled data. Unlike the first model which had perfectly balanced classes, the second through seventh models had imbalanced classes, so the minority class was upsampled to the size of the majority class using a process involved Sci-Kit Learn's 'resample' utility after it was discovered that classification algorithms using both Logistic Regression and Random Forest returned unrealistically inflated classification metrics.

The highest performing algorithms for each of the seven models were then further evaluated using the Receiver Operating Characteristic (ROC) curve for the first balanced model, and both the ROC curve and the Precision-Recall Curve (PRC) for the second through seventh imbalanced models. These curves can be examined in Appendix B.

The classification and evaluation metrics for each best performing algorithm for each of the seven classification models is tabulated below:

	Accuracy	Precision	Recall	F1	ROC AUC	PRC AUC
Cannabis/Non-Cannabis Logistic Regression	0.843	0.84	0.84	0.84	0.913	N/A
Cannabis/Non-Cannabis Random Forest	0.856	0.86	0.86	0.86	0.924	N/A
Possession/Sales Logistic Regression	0.682	0.92	0.68	0.77	0.729	0.974
Possession/Sales Random Forest	0.931	0.92	0.93	0.92	0.712	N/A
Misdemeanor Possession Logistic Regression	0.716	0.86	0.72	0.77	0.719	0.948
Misdemeanor Possession Random Forest	0.88	0.86	0.88	0.87	0.718	N/A
Violation Possession Logistic Regression	0.813	0.96	0.81	0.87	0.813	0.338
Violation Possession Random Forest	0.971	0.97	0.97	0.97	.803	N/A
Felony Possession Logistic Regression	0.736	0.98	0.74	0.83	0.784	0.111
Felony Possession Random Forest	0.98	0.97	0.98	0.97	0.739	N/A
Misdemeanor Sales Logistic Regression	0.688	0.93	0.69	0.77	0.729	0.116
Misdemeanor Sales Random Forest	0.706	0.92	0.94	0.93	0.706	N/A
Felony Sales Logistic Regression	0.708	0.99	0.71	0.82	0.673	0.015
Felony Sales Random Forest	0.994	0.99	0.99	0.99	0.547	N/A

Using Classification Models' Coefficients to Identify Salient Predictors of Cannabis Arrests

A useful attribute of Sci-Kit Learn's LogisticRegression classifier is the coefficient. Although causality for cannabis arrests cannot be attributed to them, the best Logistic Regression model's coefficients can show us which features have a strong relationship with cannabis crimes. Although there are surely a host of interactions between individual features within the feature set, for the sake of exploring which features have the strongest relationship with cannabis crime, it is assumed that they are independent of each other. With the assumption that we treat the Logistic Regression model as having no interaction terms, what the coefficients for binary features can illuminate is that if a binary feature has a value of 1, there is a certain increased likelihood that a crime will be a cannabis crime, highlighting its relationship to cannabis crime itself.

For example, if a crime was committed on a park or playground (as denoted by the feature 'PREM_TYP_DESC_PARK/PLAYGROUND'), there is a 2.750705 unit increase in the log odds of it being a cannabis crime. When log odds are converted to odds, this 2.750705 unit increase is exponentiated to reveal a 15.65% increase in the odds of it being a cannabis crime (as discussed in the highly informative UCLA FAQ on interpreting odds ratios in Logistic Regression, available at https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/).This 15.65% odds increase shows that cannabis crimes often occur in parks and playgrounds, and have a strong relationship with these premises which differentiates them from all other crimes. For continuous features like the hour the crime complaint was made or its L1 latitude/longitude distance to the Williamsburg bridge, for each unit increase in the continuous feature there is an increase in the odds that a crime is a cannabis crime.

To look at which features have the strongest relationship to cannabis crime and its subtypes, the features that contribute at least a 2% increased likelihood of a crime being classified with a '1' value were identified. Namely, the '1' values for each of the seven models were cannabis crime generally, cannabis possession crime, cannabis misdemeanor possession crime, cannabis violation possession crime, cannabis felony possession crime, cannabis misdemeanor sales crime, and cannabis felony sales crime. For the second model, the '0' value of cannabis sales crime was also used to identify the features that have the strongest relationship to it. A summary of these features can be referenced in the "Coefficients" section of the machine learning notebooks, whose links are noted above in the "In-Depth Analysis Using Machine Learning" section.

Because there are well over a 1,000 features in the feature set, the interpretability of these features' coefficients and increased odds towards cannabis crime is better enabled by grouping the features into "feature families" (e.g. suspect race, borough, housing development). This grouping strategy involves taking the absolute value sum of all coefficients in the feature set, then summing the absolute values of the coefficients for each "feature family", and then dividing the absolute value coefficient sum of each "feature family" by the total absolute value coefficient sum to uncover the proportion of the total predictive value that each "feature family" has. This will help elucidate the "feature families" that have the strongest predictive power on whether a crime is a cannabis or non-cannabis crime, and therefore which feature families should be explored further as having the strongest relationship to cannabis crime. Any feature families that contain roughly 10% of the total coefficient sum are examined further, and the features within those families that show at least a 2%

increased likelihood of a crime being classified with a ‘1’ label (or ‘0’ label when it comes to the cannabis possession/sales model) are identified.

At a high level, the features whose coefficients are examined below for each of the seven models can be split into categories of premises type, age, race/ethnicity, distances to NYC landmarks (including subway entrances), police precincts, public housing projects, transit stations and districts, NY parks, holidays, and police jurisdictions. A specific profile is presented below for each model, whose details can be found in each of the machine learning notebooks found at the GitHub links already linked to above.

Cannabis Crime (1) as Differentiated from Non-Cannabis Crime (0):

The specific premises features with the highest coefficients show that cannabis arrests in New York City between 2006 and 2018 had the strongest relationship with parks and playgrounds, public housing buildings, the street, and open areas or lots. All of these features independently contributed at least an increased odds of 5% that a crime was a cannabis crime. Other common areas for cannabis arrests were marinas/piers, public parking lots or garages, public schools, apartment houses, tunnels (e.g., Holland Tunnel, Lincoln Tunnel), and the NYC subway, all of which independently contributed an increased odds between 2 and 5% of a crime being a cannabis crime.

The suspect age ranges of 18-24, less than 18, and 25-44 had positive coefficients, independently contributing an increased odds of 4.2, 2.5, and 2.0% to a crime being a cannabis crime, respectively. This reflects the finding in the Data Story & EDA notebook that a higher percentage of cannabis arrests were of younger people.

The L1 distance to the Williamsburg Bridge also had a notable relationship with cannabis crime, with an increased odds of 3.2%. Williamsburg and the Lower East Side are two neighborhoods on either side of the Williamsburg Bridge, both of which are known as countercultural or "hipster" neighborhoods, which may explain the relationship. There is also a Williamsburg housing development close to the bridge, which was shown below to have a strong relationship with cannabis crime.

Police precincts with the highest relationship to cannabis crime, in descending order, were the 71st, 75th, 67th, 77th, 73rd, 30th, 100th, 115th and 7th. Precinct descriptions come from https://www1.nyc.gov/site/nypd/bureaus/patrol/precincts-landing.page . These cover the following areas:

- 71st - Crown Heights, Wingate, Prospect Lefferts neighborhoods of Central Brooklyn (3.4% increase in odds)
- 75th - East New York and Cypress Hills neighborhoods of Easternmost Brooklyn (3% increase)
- 67th - East Flatbush and Remsen Village neighborhoods of Central Brooklyn (2.8% increase)
- 77th - Northern portion of Crown Heights and part of Prospect Heights neighborhoods in Central Brooklyn (2.7% increase)
- 73rd - Brownsville and Ocean Hill neighborhoods of northeastern Brooklyn (2.6% increase)
- 30th - Hamilton Heights, Sugar Hill, and West Harlem neighborhoods in the Western portion of Harlem (2.6% increase)

- 100th - Arverne, Belle Harbor, Breezy Point, Broad Channel, Neponsit, Rockaway Park, Rockaway Beach, and Roxbury neighborhoods of Queens (2.5% increase)
- 115th - Jackson Heights, East Elmhurst, and North Corona neighborhoods of northern Queens, including LaGuardia Airport (2.4% increase)
- 7th - Manhattan's Lower East Side (2.1% increase)

These precincts historically have predominantly African-American and Latino residents, possibly with the exclusion of the 115th and 7th.

Certain public housing developments had a strong relationship with cannabis crimes. They include:

- Williamsburg (3.3% increase): This housing development is noted above in relationship to the 'will_bridge_l1' feature, and is in the Williamsburg neighborhood of Brooklyn.
- Borinquen Plaza II (3.3% increase): This housing development is in Bushwick, in a classically Puerto Rican neighborhood of Brooklyn.
- Marcy Projects (2.9% increase): This housing development is in Bedford-Stuyvesant, in a classically African-American neighborhood of Brooklyn.
- Wyckoff Gardens (2.6% increase): This housing development is in Boerum Hill, a gentrified neighborhood of Brooklyn.
- Throggs Neck (2.5% increase): This housing development is in the Throggs Neck neighborhood of the Bronx, which was shown to have a large percentage of cannabis arrests in the Data Story and EDA notebook.
- Williams Plaza (2.3% increase): This housing development is in the South Williamsburg neighborhood of Brooklyn.
- Whitman (2.3% increase): This housing development is in the Fort Greene neighborhood of Brooklyn, which was historically an African-American neighborhood.
- Smith (2.2% increase): This housing development is in the Two Bridges neighborhood of Manhattan, near the Brooklyn Bridge.
- Armstrong I (2.2% increase): This housing development is in Bedford-Stuyvesant, in a classically African-American neighborhood of Brooklyn.
- Bushwick (2.1% increase): This housing development is in Bushwick, in a classically Puerto Rican neighborhood of Brooklyn.

The hour and minute of the crime had a positive relationship with cannabis crime, with an increased odds of a crime being a cannabis crime of 7.9% and 2.7% respectively. This relationship seems to reflect the fact that cannabis arrests happened more often during the latter part of the day, as is shown in the value counts for the 'start_hour' feature. So that as the hour increases towards the 23rd hour, there were more cannabis arrests (as was shown in the Data Story and EDA notebook). The relationship of the minute of the crime somewhat shows that there were more cannabis arrests made during the latter part of the hour than in the earlier part of the hour, although the value counts for the 'start_minute' feature didn't show an obvious pattern.

Transit District 23 is in Rockaway Park, Queens, and it had a 2.2% increased odds for cannabis crime. It's hard to say what relationship exists between this transit district and cannabis crime, but it does reflect the positive relationship between the 100th NYPD precinct noted above, which includes Rockaway Park. Because the majority of cannabis crimes were not transit related (approximately 98% had missing values for 'TRANSIT_DISTRICT' and 'STATION_NAME'), this particular Transit District in Rockaway Park, Queens must have had a concentrated effort or pattern of cannabis

arrests on behalf of the NYPD and transit police. This may suggest a cannabis distribution hub of some kind.

The relationship between cases with unknown suspect sex ('SUSP_SEX_cleaned_unknown') and cannabis crime was strong, with a coefficient of 2.5 and an increased odds of 12.2%. As was called below, the majority of cannabis arrests did not have the suspect's sex reported, suggesting that there was some confounding variable or variables for why the NYPD did not record the suspect's sex for cannabis crimes much more than for non-cannabis crimes. As was explored in the Data Story and EDA notebook, as well as in the Statistical Data Analysis notebook, about 16% of cannabis crimes had the suspect's race, sex, and age recorded by the arresting officer, while about 38% of non-cannabis crimes had these demographic data recorded. Therefore, there was some reason for this disparity, which warrants future research as the cause for this missing demographic data cannot be determined through the current analysis.

The feature storing whether the date of the crime complaint and the reported date of the crime match ('rpt_cmplnt_dt_match') also had a strong coefficient of 2.4 and an increased odds of 11.3%. The complaint date and the reported date of the crime matches for about 93% of cases (as shown below in the value counts for the 'rpt_cmplnt_dt_match' feature), showing that there was a confounding variable or variables for why there was a high rate of matching for cannabis crimes more so than for non-cannabis crimes. This may simply be because NYPD officers were more likely to report cannabis crimes on the same day of the arrest. The majority of cannabis arrests were misdemeanor arrests, where NYPD officers bring the arrestees to central booking and report the crime. Although again, the true reason for this relationship would have to be explored through future research.

The “feature families” that had the strongest relationship to cannabis crime generally, in descending order, were public housing developments (“projects”), premises types, transit stations, and police precincts. After looking at the specific features within the “feature families”, no new features with a significant relationship with cannabis crime were discovered.

However, the transit station “feature family” as a whole had a significant relationship with cannabis crime. The absolute sum of transit stations' coefficients was 45.1, and its percentage of the entire absolute sum of coefficients was 16%. This warrants looking closer at the strength of the relationship that transit stations have with cannabis crime. However, as can be seen in the Jupyter notebook, none of the individual transit stations had an increased odds of 2% that a crime was classified as a cannabis crime. Therefore, the 10 transit stations with the greatest relationship to cannabis crime are displayed in the Jupyter notebook, but were not explored any further. It was seen earlier that violation possession crimes had a fairly unique relationship with subway stations, so this relationship may be responsible for the rather large percentage of the entire absolute sum of coefficients that the transit station “feature family” has.

It bears mentioning that a crime suspect's age has a strong relationship with cannabis crime. The strongest relationship is for suspects between the ages of 18-24, the second strongest for those aged less than 18, and the third strongest for those aged between 25-44. The increased odds that a crime is a cannabis crime for people aged less than 18 years is 2.5%, for people aged 18-24 is 4.2%, and for people aged 25-44 is 2%. This data adds more credence to an age bias in cannabis arrests.

These relationships likely reflect that younger people are more likely to use cannabis, but it also shows that children are being arrested for cannabis use, and that the relationship between children and cannabis crime is stronger than that between those aged 25-44, even though the 25-44 year old group is much larger than the less than 18 year old group. This disparity in statistical relationships may be due to older people being more careful with their cannabis use and therefore may be less likely to be caught, or it could be that younger people are being targeted by the police for cannabis crimes. Obviously, the cause for these statistical relationships cannot be uncovered by this project, but this is definitely an area for future research.

Race has been at the center of the conversation around cannabis arrests in New York City, and this project has shown that the majority of cannabis arrests between 2006 and 2018 where the suspect's race/ethnicity was reported were of African-Americans and Hispanics, with only 8% of those arrests made of Whites. To explore this relationship deeper, the racial/ethnic groups' coefficients and increased odds of a crime being a cannabis crime were examined. African-American suspect race had a -0.31 coefficient, which shows that this feature is more closely related with non-cannabis crime than with cannabis crime. This is also true of White Hispanics, Whites, Black Hispanics, and Native Americans. Asians and Pacific Islanders have a very weak positive coefficient of 0.06. These coefficients simply show that there is a stronger (but still weak) relationship with non-cannabis crime than for cannabis crime, which makes sense as cannabis crime is only a subset of all crimes, and that African-Americans are generally more likely to be arrested for all types of crime in New York City.

What is interesting however, is that there is a strong 0.6 coefficient for unknown suspect race. This translates to a 1.8% increased likelihood that a crime is a cannabis crime. This shows that there is an underlying reason for NYPD officers not recording the suspect's race when it comes to cannabis crime. The precise reason, or reasons, are beyond the scope of this project. However, future study is highly warranted.

So the Logistic Regression model clearly shows that the suspect's race does not have a strong positive relationship with whether a crime was classified as a cannabis crime. This finding is confusing given the racial skew in cannabis arrests. It is important to again underline that the NYPD's data on suspect race for cannabis crimes was very partial with only 16% of crimes having their suspect's race reported. It is clear from the hypothesis testing done in the Statistical and Data Analysis notebook that the racial disparity in cannabis arrests is not due to chance and that there is a confounding reason for the disparity, but the web of confounding variables associated with this disparity needs to be explored further.

To further explore this seeming paradox, all of the suspect race features' coefficients were summed in absolute value terms, and then the absolute proportion of this summed coefficient was called for each racial group. This step was taken to show the proportion that each racial group has of the relationship between suspect race and whether a crime was a cannabis crime or non-cannabis crime.

The African-American proportion was the highest amongst crimes where the suspect's race was recorded at 0.23, the Hispanic White proportion was 0.09, the Hispanic Black proportion was 0.07, the White proportion was 0.09, the Native American proportion was 0.02, the Asian and Pacific Islander proportion was 0.04, and the unknown suspect race proportion was highest at 0.45

(because the majority of crimes in New York City do not have their suspect's race reported by the NYPD). This shows that African American status of arrestees did compose the majority of the relationship between suspect race and crime, both cannabis and non-cannabis.

DataFrames for crimes committed by each of the racial/ethnic groups were subsetted from the overall DataFrame, where the suspect's race was actually recorded. The value counts for the 'cannabis_crime' target feature were called, showing the counts and percentages of criminals of each racial group that were arrested for cannabis that had their race reported by the arresting officer.

It is important to note that the DataFrame that the Logistic Regression model was trained on was composed of all of the cannabis crimes between 2006 and 2018 and a representative sample of all other crimes of the same size as the cannabis crimes. Therefore, the percentages reported below are of this concatenated DataFrame of all cannabis crimes and a representative sample of all other cannabis crimes. Therefore, higher percentages can be interpreted as a bias in cannabis arrests for that racial/ethnic group.

For the DataFrame used, 30% of African-American criminals, 33% of White Hispanic criminals, 18% of White criminals, 40% of Black Hispanic criminals, 20% of Native American criminals and 19% of Asians or Pacific Islanders in New York City between 2006 and 2018 were arrested for cannabis. We see again that White criminals have the lowest arrest rate for cannabis, roughly equivalent with that of Native American and Asian/Pacific Islander criminals. African-American and Hispanic criminals are shown to be arrested at a higher percentage than the other racial/ethnic groups. This is another way to highlight the racial disparity in arrests, and helps to make better sense of the coefficients for the racial categories that were returned by the Logistic Regression model.

Cannabis Possession (1) as Differentiated from Cannabis Sales (0):

Possession:

The specific premises features with the highest coefficients and likelihoods show that cannabis possession arrests oddly had the strongest relationship with airport terminals, marinas/piers, and NYC buses. It's important to remember that not all arrests had a premises type recorded, but it's clear that there is a stronger relationship between these types of premises and cannabis possession than would be initially thought of. Among cannabis arrests, an arrest made at airport terminals was 51.1% more likely to be a possession arrest, which does make sense as selling cannabis at the heavily policed location of an airport terminal would be highly dangerous for the seller. Less outstanding, but still notable, is that cannabis arrests at a marina or pier were 4.4% more likely to be a possession arrest, and cannabis arrests on NYC buses are 2.3% more likely to be a possession arrest.

Unlike sales arrests, NYC parks had no notable relationship with possession arrests as a whole.

As a whole, the only transit station that had a notable relationship with possession arrests is the 42nd Street Port Authority Bus Terminal station, which had a 2.3% increased odds that it was a possession arrest. This may have to do with the fact that many homeless and transient people pass

through this bus terminal with small amounts of cannabis, and possession arrests are much more common than sales arrests.

On this topic, transit districts 11 and 12 had a relationship with possession arrests, with a 3% and 4% increased odds that cannabis arrests made there were possession arrests, respectively. Transit district 11 was centered around Yankee Stadium in the Bronx and transit district 12 was also located in the Bronx, around 180th Street (as can be found at www1.nyc.gov/site/nypd/bureaus/transit-housing/transit.page). This mirrors data found elsewhere in this project that possession arrests were highly concentrated in the Bronx.

Cannabis arrests with missing suspect race were 2.1% more likely to be possession arrests than sales arrests. This shows that there was some underlying reason why the suspect's race was not being recorded for possession arrests specifically. The causal reason can't be identified with this analysis, but future research is definitely needed in this area as the racial disparity has been a point of critical focus (especially in misdemeanor possession crimes).

Cannabis arrests with missing suspect sex were 2.1% more likely to be possession arrests than sales arrests. This shows that there was some underlying reason why the suspect's sex was not being recorded for possession arrests. The causal reason can't be identified with this analysis, but future research is definitely needed in this area.

The L1 distance to the Empire State Building also had a notable relationship with cannabis possession crime, with an increased odds of 6%. This likely had to do with the fact that the area around the Empire State Building is highly policed and is located in the area close to the major transit hub of New York City, which includes Penn Station and the Port Authority Bus Terminal. The extremely high traffic flow of commuter, tourist, transient, and homeless populations passing through this area increases the odds of any cannabis arrest, and of course possession arrests are more common than sales arrests.

The L2 distance to the closest subway station entrance also has a notable relationship with cannabis possession crime, with an increased odds of 5.4%. This shows that the closer to a subway station a cannabis arrest is in a straight line, the more likely it is a possession arrest and the less likely it is a sales arrest. This shows that drug deals are not typically centered around subway entrances any longer due to the underground bike messenger delivery services that proliferate throughout New York City.

If a cannabis crime was committed on Christmas Eve it was more likely to be a possession arrest, with a 2.1% increased odds. There are not many cannabis arrests on Christmas Eve (as shown in the Data Story & EDA notebook), so the relationship is likely partially rooted in the fact that there were more possession than sales arrests generally.

Police precincts with the highest relationship to cannabis possession crime, in descending order, were the 100th, 75th, 88th, 77th, 43rd and 26th. These cover the following areas:

- 100th - Arverne, Belle Harbor, Breezy Point, Broad Channel, Neponsit, Rockaway Park, Rockaway Beach, and Roxbury neighborhoods of Queens (3.5% increase)
- 75th - East New York and Cypress Hills neighborhoods of Easternmost Brooklyn (3.4% increase)

- 88th - Northern portion of Brooklyn containing Clinton Hill, Fort Greene Park, and Commodore Barry Park (3% increase)
- 77th - Northern portion of Crown Heights and part of Prospect Heights neighborhoods in Central Brooklyn (2.5% increase)
- 43rd - Southeast section of the Bronx, with the four primary commercial strips of Westchester Avenue, Castle Hill Avenue, White Plains Road, and Parkchester (2.4% increase)
- 26th - A portion of the Upper West Side of the Manhattan, encompassing Morningside Heights and Manhattanville, and is home to Columbia University, Riverside Park, Morningside Park, and General Grant National Memorial (2.2% increase)

These findings clearly show that the precincts with the strongest relationship to cannabis possession arrests were neighborhoods with traditionally Hispanic and African-American populations of lower socioeconomic status, even though some of these neighborhoods became increasingly gentrified throughout the time period between 2006 and 2018.

Interestingly, the jurisdictions with the strongest relationship to cannabis possession arrests were not the NYPD but the Tri-Borough Bridge and Tunnel Authority (5.5%) and the NY Transit Police (2.5%). This may be due to the fact that cannabis sales arrests just happen less frequently over the mass transit system and the Tri-Borough Bridge and Tunnel.

Certain public housing developments (or projects) had a strong relationship with cannabis possession crimes. They include:

- McKinley in the Morrisania neighborhood of the South Bronx (3.7% increase),
- 303 Vernon Avenue in the Bedford-Stuyvesant neighborhood of Central Brooklyn (3.6% increase),
- Morrisania Air Rights 42 in the Morrisania neighborhood of the South Bronx (3.6% increase),
- Edenwald in the Edenwald neighborhood of the North Bronx (3.4% increase),
- Williams Plaza in the South Williamsburg neighborhood of Brooklyn (3.2% increase),
- Weeksville Gardens in the Weeksville neighborhood of Central Brooklyn next to Crown Heights (2.9% increase),
- Tompkins in the Bedford-Stuyvesant neighborhood of Central Brooklyn (2.7% increase),
- LaFayette in the Clinton Hill neighborhood of Central Brooklyn (2.7% increase),
- Hughes Apartments in the Brownsville neighborhood of Central Brooklyn (2.4% increase),
- Claremont Rehab (Group 3) in the Morrisania neighborhood of the South Bronx (2.3% increase),
- Patterson in the South Bronx (2.3% increase),
- Wagner in East Harlem (2.2% increase),
- Throggs Neck Addition in the Throggs Neck neighborhood of the Bronx, which was shown to have a large percentage of cannabis arrests in the Data Story and EDA notebook (2.1% increase),
- Eastchester Gardens in the East Bronx (2% increase),
- Sumner in the Bedford-Stuyvesant neighborhood of Central Brooklyn (2% increase)

The “feature families” that had the strongest relationship to cannabis possession crime, in descending order, were public housing developments (“projects”), police precincts, premises types,

and transit stations. After looking at the specific features within the “feature families”, no new features with a significant relationship with cannabis possession crime were discovered.

When it comes to the suspect race feature’s relationship to cannabis possession crime, the three groups with a positive relationship with possession were Asians and Pacific Islanders with a 0.16 coefficient, Whites with a coefficient of 0.58, and an unknown suspect race coefficient of 0.75. The bidirectional (not absolute) mean coefficient of suspect race as a whole was roughly 0.16, which shows that suspect race was a predictor of whether a cannabis crime was classified as a possession or sales crime. Suspect race was associated with a roughly 1.2% increase in a crime being classified as a cannabis possession crime.

There is a weak relationship between cannabis possession and Asians and Pacific Islanders. For cannabis crimes, there is a 1.2% increased likelihood that the cannabis crime is a possession crime if the suspect belongs to these racial/ethnic groups.

There is a moderately strong relationship between cannabis possession and white suspects, however. The coefficient was 0.58 and the increased likelihood of a cannabis crime being marked as possession was 1.8%. Possible reasons for this relationship is that White people sell cannabis less often, that Whites are more likely to be arrested for possession for cannabis crimes where the amount confiscated is close to the possession/sales line of one ounce, or that Hispanic African-Americans, African-Americans, and White Hispanics are being targeted for cannabis sales crimes. These reasons are impossible to tease out with this study, as the data set used in this study doesn't contain information on how many White people are selling cannabis who do not get caught, but this finding does suggest a need for further research.

There is also a strong relationship between unknown suspect race and cannabis possession, with a 0.75 coefficient and an increased likelihood that a cannabis crime was a possession crime of 2.1%. Again, this shows that there was some confounding variable that was causing NYPD and other law enforcement officers to not record the suspect's race for possession crimes. Whether this was intentional on the part of the officers is impossible to say, but future research is definitely needed.

Sales:

The specific premises features with the highest coefficients and likelihoods show that cannabis sales arrests had the strongest relationship with liquor stores, groceries, bodegas, candy stores, fast food restaurants, unspecified stores, clothing stores and boutiques, telecomm stores, and gas stations. All of these premise types had between a 2 and 7% increased odds that a cannabis crime is a sales crime and not a possession crime. This suggests a focus on low-level sales arrests in small businesses where cannabis sales are a second "under the table" side business.

Unlike cannabis arrests generally and possession arrests specifically, cannabis sales arrests have several strong relationships with specific NYC parks. These parks represent "hot spots" of cannabis sales arrests, and are located near NYU in southern Manhattan, the Hell's Kitchen neighborhood of Manhattan, east Harlem, the Bronx, and a few Brooklyn neighborhoods. These include in descending order of increased likelihood that a cannabis crime is a sales crime:

- Washington Square Park next to the NYU campus (with a stunning increased odds of 30.3%),

- Parkside Playground Brooklyn in the Prospect Lefferts Gardens neighborhood of central Brooklyn (6.1% increase),
- St. James Park in the Fordham Manor neighborhood of central Bronx (5.8% increase),
- Park of the Americas in the Corona neighborhood of central Queens (3.9% increase),
- L/Cpl Thomas P. Noonan Jr. Playground in the Sunnyside neighborhood of central Queens (3.8% increase),
- Story Playground in the Soundview neighborhood of the east Bronx (3.5% increase),
- Maria Hernandez Park in the Bushwick neighborhood of north Brooklyn (2.9% increase),
- Matthews-Palmer Playground in the Hell's Kitchen neighborhood of Manhattan (2.6% increase),
- Harlem Art Park in East Harlem (2.6% increase),
- Stockton Playground close to the Williamsburg neighborhood of Brooklyn (2.4% increase),
- Harris Park in the Kingsbridge neighborhood of the central Bronx (2.4% increase), and
- Sixteen Sycamores Playground in the Fort Greene neighborhood of central Brooklyn (2.3% increase).

Also unlike cannabis arrests generally, cannabis sales arrests had several strong relationships with specific subway/transit stations. These include (in descending order of coefficients and likelihoods):

- Jackson Avenue in the South Bronx (with a large increased odds of 26.1%),
- Lexington Avenue either near Grand Central Station in Midtown Manhattan or in East Harlem as the data label is not specific enough (13.8% increase),
- Pennsylvania Avenue in the East New York neighborhood of eastern Brooklyn (7.9% increase),
- Kingston Avenue in the Crown Heights neighborhood of Central Brooklyn (4.8% increase),
- Junius Street in the Brownsville neighborhood of eastern Brooklyn (3.8% increase),
- 50th Street which could be the station near the Port Authority bus terminal or a station in the Bay Ridge neighborhood of southern Brooklyn (more likely the former) (3.4% increase),
- 42nd St. Times Square in midtown Manhattan (2.8% increase), and
- Stillwell Avenue in Coney Island (2.6% increase).

Unknown or unrecorded age had a 7.1% increased odds that the cannabis crime was a sales crime, showing that there was some underlying reason why the suspect's age was not being recorded for sales crimes. As was shown below in a value counts call, the majority of cannabis sales crimes (85%) did not have their suspect's age recorded.

Male suspects also had a strong relationship with cannabis sales crime, with an increased odds of 2.2% that a cannabis crime was a sales crime. This shows that males were more likely to be arrested for sales crimes than for misdemeanor crimes (although the vast majority of all cannabis arrests are of male suspects).

The L2 distance to the World Trade Center in downtown Manhattan also had a strong relationship with cannabis sales crime, with an increased odds of 6.1% that a cannabis arrest is a sales arrest. This is an unusual finding, as other geographical features don't show that there were a lot of cannabis sales arrests happening downtown. This finding would require future research, but is likely associated with a higher level of policing around the World Trade Center that may net more sales arrests involving the black market cannabis delivery networks.

The L1 distance to Yankee Stadium in the Bronx also had a moderately strong relationship with cannabis sales crime, with an increased odds of 5.2% that a cannabis arrest is a sales arrest. This makes sense, as there are many other geographical features that show that many cannabis sales arrests occur in the Bronx.

Police precincts with the highest relationship to cannabis sales crime, in descending order, were the 6th, 112th, 13th, 33rd, 34th, and 83rd. These cover the following areas:

- 6th - Southwestern Manhattan neighborhoods of Greenwich Village and the West Village (5.2% increase), a liberal enclave housing most of the NYU campus and traditionally a bastion of the counterculture
- 112th - Centrally located portion of Queens, housing Forest Hills and Rego Park, reflecting the strong relationship between the Pomonok Houses project in this portion of Queens, mentioned below (3.7% increase)
- 13th - Southern portion of Midtown, Manhattan. The precinct features the Peter Cooper Village/Stuyvesant Town residential complex, Gramercy Park, the lower portion of Rosehill, Madison Square Park, and Union Square Park (3.6% increase). Union Square Park may be responsible for this relationship.
- 33rd - Washington Heights neighborhood of Northern Manhattan (3.1% increase), a traditionally Hispanic neighborhood
- 34th - Communities of Washington Heights and Inwood, north of West 179th Street (2.7% increase), a traditionally Hispanic neighborhood
- 83rd - Northern region in Brooklyn comprising Bushwick (2.7% increase), a traditionally Hispanic neighborhood

Certain public housing developments (or projects) had a strong relationship with cannabis sales crimes. They include:

- Pomonok Houses in the Pomonok neighborhood of Queens (5.9% increase in odds towards sales),
- Wald in the Alphabet City neighborhood of Lower Manhattan (5.7% increase),
- Queensbridge North between the Astoria and Hunters Point neighborhoods of Queens (4.6% increase),
- Castle Hill in the Castle Hill neighborhood of the South Bronx (4.1% increase),
- Sedgwick in the Morris Heights neighborhood of the South Bronx (3.7% increase),
- Bay View in the East New York neighborhood of Brooklyn (3.5% increase),
- Astoria in the Astoria neighborhood of Queens (3.1% increase),
- Grant in the Manhattanville neighborhood of Harlem (3.1% increase),
- Highbridge Rehabs (Nelson Avenue) of the South Bronx (2.8% increase),
- Jacob Riis in the Alphabet City neighborhood of Lower Manhattan (2.5% increase),
- Soundview in the Soundview neighborhood of the South Bronx (2.4% increase),
- Ingersoll in downtown Brooklyn (2.4% increase),
- Marble Hill in the South Bronx across the Harlem River from the Inwood neighborhood of the South Bronx (2.3% increase),
- Sterling Place Rehab (Sterling-Buffalo) in the Weeksville neighborhood of Central Brooklyn close to Crown Heights (2.3% increase),

- Claremont Rehab (Group 3) in the Morrisania neighborhood of the South Bronx (2.2% increase),
- 33-35 Saratoga Avenue in the Bedford-Stuyvesant neighborhood of Brooklyn (2.2% increase),
- Woodson in the Brownsville neighborhood of Brooklyn (2% increase), and
- Queensbridge South between the Astoria and Hunters Point neighborhoods of Queens (2% increase)

The “feature families” that had the strongest relationship to cannabis sales crime, in descending order, were public housing developments (“projects”), police precincts, premises types, NYC parks, and transit stations. After looking at the specific features within the “feature families”, no new features with a significant relationship with cannabis sales crime were discovered.

There are weak relationships between cannabis sales arrests and Hispanic African-American (coefficient of 0.18), African-American (0.17), and White Hispanic suspects (0.03). This suggests that race as a whole does not have an overly strong relationship with cannabis sales as a whole, although racial/ethnic categories will be shown below to have a relationship with its subtypes.

Misdemeanor Cannabis Possession:

The specific premises features with the highest coefficients and likelihoods show that misdemeanor possession arrests had the strongest relationship with marinas and piers, parks and playgrounds, open areas and lots, public housing residences, public parking lots and garages, and social clubs, in descending order. These features had increased likelihoods of cannabis crimes being misdemeanor possession crimes between 6.7% and 2.3%, with marinas and piers having the highest likelihoods and social clubs having the least. Some of these features reflect how cannabis users are forced to utilize open public spaces to smoke and be caught by the police. Notably, the high rates of cannabis arrests in public housing buildings was again shown with the "PREM_TYP_DESC_RESIDENCE - PUBLIC HOUSING" feature's coefficient of 0.9 and increased likelihood of a misdemeanor possession charge of 2.9%. Also, the definition of the feature "PREM_TYP_DESC_SOCIAL CLUB/POLICY" was unclear, but it seems misdemeanor possession charges were occurring in social clubs in New York City. Whether these are private clubs being raided or normal public bars and music clubs is unclear.

The only NYC parks with a notable relationship with misdemeanor possession were Claremont Park in the South Bronx (2.3% increased odds of misdemeanor possession) and Riverside Park in the Riverside neighborhood of the Upper West Side of Manhattan (2.2% increased odds).

There were three transit subway stations with a notable relationship with misdemeanor possession:

- Nostrand Avenue in Crown Heights, which had a long-standing Afro-Caribbean population (2.4% increased odds),
- Queensboro Plaza near Long Island City, Queens (2.2% increased odds), and
- 168 St. Washington Heights in the traditionally Hispanic neighborhood of Washington Heights (2% increased odds)

There are three transit districts that had a relationship with misdemeanor possession:

- Transit District 20 in Jamaica, NY (2.8% increased odds),
- Transit District 34 in Coney Island, Brooklyn (2.7% increased odds), and
- Transit District 23 in Rockaway Park, Brooklyn (2.5% increased odds)

Cannabis arrests with missing suspect sex were 2.3% more likely to be misdemeanor possession arrests than all other cannabis arrests. This shows that there is some underlying reason why the suspect's sex is not being recorded for misdemeanor possession arrests. The causal reason can't be identified with this analysis, but future research is definitely needed in this area.

The L1 distance to Prospect Park also had a notable relationship with misdemeanor possession crime, with an increased odds of 3% that a cannabis crime is a misdemeanor possession crime. This may have to do with public smoking in the park, or its proximity to several heavily African-American neighborhoods like Crown Heights and Flatbush.

The L2 distance to the Williamsburg Bridge also had a notable relationship with misdemeanor possession crime, with an increased odds of 2.6% that a cannabis crime is a misdemeanor possession crime. Williamsburg and the Lower East Side are two neighborhoods on either side of the Williamsburg Bridge, both of which are known as countercultural or "hipster" neighborhoods, which may explain the relationship. There is also a Williamsburg housing development close to the bridge, which is shown below to have a strong relationship with cannabis crime.

The L2 distance to the closest subway station entrance also has a notable relationship with cannabis misdemeanor possession crime, with an increased odds of 2%. This shows that the closer to a subway station a cannabis arrest is in a straight line, the more likely it is a misdemeanor possession arrest and the less likely it is another type of cannabis crime.

Police precincts with the highest relationship to cannabis misdemeanor possession crime, in descending order, were the 75th, 77th, 73rd, 43rd, 100th and 115th. These cover the following areas:

- 75th - East New York and Cypress Hills neighborhoods of Easternmost Brooklyn (3.4% increase)
- 77th - Northern portion of Crown Heights and part of Prospect Heights neighborhoods in Central Brooklyn (2.7% increase)
- 73rd - Brownsville and Ocean Hill neighborhoods of northeastern Brooklyn (2.3% increase)
- 43rd - Southeast section of the Bronx, with the four primary commercial strips of Westchester Avenue, Castle Hill Avenue, White Plains Road, and Parkchester (2.3% increase)
- 100th - Arverne, Belle Harbor, Breezy Point, Broad Channel, Neponsit, Rockaway Park, Rockaway Beach, and Roxbury neighborhoods of Queens (2.1% increase)
- 115th - Jackson Heights, East Elmhurst, and North Corona neighborhoods of northern Queens, including LaGuardia Airport (2% increase)

Certain public housing developments (or projects) had a strong relationship with cannabis misdemeanor possession crimes. They include:

- 303 Vernon Avenue in the Bedford-Stuyvesant neighborhood of Central Brooklyn (3.2% increase),
- McKinley in the Morrisania neighborhood of the South Bronx (2.8% increase),

- Ocean Hill Houses in the Brownsville neighborhood of East Brooklyn (2.3% increase)
- Edenwald in the Edenwald neighborhood of the North Bronx (2.3% increase),
- Morrisania Air Rights 42 in the Morrisania neighborhood of the South Bronx (2.3% increase),
- Borinquen Plaza II in Bushwick, in a classically Puerto Rican neighborhood of Brooklyn (2.3% increase),
- Adams in the South Bronx (2.2% increase),
- Sheepshead Bay in the Sheepshead Bay neighborhood of South Brooklyn (2.1% increase),
- Williams Plaza in the South Williamsburg neighborhood of Brooklyn (2.1% increase),
- Dyckman in the Inwood neighborhood of North Manhattan (2% increase), and
- Mill Brook in the Mott Haven neighborhood of the South Bronx (2% increase)

The “feature families” that had the strongest relationship to misdemeanor cannabis possession crime, in descending order, were public housing developments (“projects”), NYC parks, transit stations, police precincts, and premises types. After looking at the specific features within the “feature families”, no new features with a significant relationship with misdemeanor cannabis possession crime were discovered.

Amongst cannabis crime arrestees, Asians and Pacific Islanders were the least associated with misdemeanor possession arrests (coefficient of -0.34), while African Americans (-0.18) , African American Hispanics (-0.14), and White Hispanics (-0.04) had a weaker negative relation with misdemeanor possession. Native Americans had effectively no relationship (0.03), Whites had a weak positive relationship with misdemeanor possession (0.21), and unknown suspect race had a moderately strong relationship with this level of cannabis crime (0.43 coefficient), again showing that there was a confounding variable or variables associated with the lack of recording the suspect's race.

Violation Cannabis Possession:

As detailed above, the F1 score for the violation possession class was only 0.18 when using the best performing Logistic Regression algorithm 'upsampled_vp_15'. However, the coefficients can still be valuable in identifying salient features of violation possession crimes to a certain degree if the F1 score is higher than zero. The only scenario where the coefficients would definitely not be valuable is if the algorithm only or almost only predicts one class. To check on whether the algorithm predicts both classes, and not just one, the predicted values 'upsampled_vp_pred_15' were converted to a Pandas Series and then its value counts were called. The violation possession class is predicted 8,654 times and the other cannabis crime types as a group were predicted 35,407 times. Therefore, the coefficients for the violation possession class were valuable and are analyzed below.

The specific premises features with the highest coefficients show that violation possession arrests that had a premise recorded had the strongest relationship with NYC busses, ferries and ferry terminals, other transit facilities, public schools, banks, factories and warehouses, check cashing businesses, department stores, airport terminals, bus stops, hospitals, and tunnels, in descending order. NYC busses really stand out with a 31.5% increased likelihood of a violation possession arrest. All of these other premise types had a 2% to 7% increased likelihood of being a violation possession arrest.

Unlike cannabis possession arrests generally, violation possession arrests had several strong relationships with specific NYC parks. These parks represent "hot spots" of violation possession arrests, and were located overwhelmingly in areas of the South Bronx and Brooklyn that have large populations of African-Americans, Hispanics and poor populations. These include in descending order of increased likelihood that a cannabis crime is a violation possession crime:

- Joyce Kilmer Park in the South Bronx (with a large increased odds of 22.3%),
- Paerdagat Park in the East Flatbush neighborhood of Brooklyn (18% increase),
- Neptune Playground in the Coney Island/West Brighton neighborhood of Brooklyn (17.8% increase),
- Linden Park in the East New York neighborhood of Eastern Brooklyn (12.8% increase),
- J. Hood Wright Park in the Washington Heights neighborhood of Northern Manhattan (10.5% increase),
- Weeksville Playground in the Weeksville neighborhood near Crown Heights in Central Brooklyn (9% increase),
- Coney Island Beach & Boardwalk in the Coney Island neighborhood of Southern Brooklyn (7.9% increase),
- St. Mary's Park Bronx in the South Bronx (7.8% increase),
- Remsen Playground in the Canarsie neighborhood of Eastern Brooklyn (7.7% increase),
- Fulton Park in the Stuyvesant Heights neighborhood of Central Brooklyn (5.6% increase),
- Mount Prospect Park in the Prospect Heights neighborhood of Central Brooklyn (5.6% increase),
- Livonia Park in the Brownsville neighborhood of Eastern Brooklyn (5.3% increase),
- De Hostos Playground in the South Williamsburg neighborhood of Northern Brooklyn (5% increase),
- Washington Square Park next to the NYU campus in Southern Manhattan (4.9% increase),
- Playground Ninety in the Jackson Heights neighborhood of Queens (4.6% increase),
- Msgr. McGolrick Park in the Greenpoint neighborhood of Northern Brooklyn (4.2% increase),
- Fort Greene Park in the Fort Greene neighborhood of Central Brooklyn (2.8% increase), and
- Central Park in Manhattan (2.2% increase)

The following transit districts had strong to weak relationships with violation possession crimes:

- Transit District 12 in the South Bronx (19.8% increase),
- Transit District 11 near the Yankee Stadium in the South Bronx (8.1% increase),
- Transit District 1 in Midtown Manhattan (4.6% increase),
- Transit District 4 near Union Square in Southern Manhattan (3.8% increase), and
- Transit District 3 in West Harlem (2.2% increase)

For each full latitude/longitude unit closer a cannabis crime is to the Brooklyn Bridge, there was a 439.5% increased odds that it is a violation possession charge and not another cannabis crime. This shows that violation possession charges were far more likely to be enforced in southern Manhattan and affluent areas of Brooklyn close to the Brooklyn Bridge.

Interestingly, for each full latitude/longitude unit closer a cannabis crime was to Riker's Island, there was a 15% increased odds that it was a violation possession charge and not another cannabis crime. This is harder to interpret, but it is clear that these arrests were not happening in the Riker's

Island prison, so there was a relationship with violation possession arrests occurring in the surrounding neighborhoods of the South Bronx and Queens.

Intriguingly, both Christmas and Hanukkah had weak relationships with violation possession crimes, at a 3.4% and 2.7% increased odds, respectively. This may suggest that police officers were somewhat more lenient on these holidays when enforcing cannabis law.

Police precincts with the highest relationship to violation possession crime, in descending order, were 84, 78, 100, 76, 94, 5, and 109. These cover the following areas:

- 84th - Northwestern section of Brooklyn, home to Brooklyn Heights, Boerum Hill, and Vinegar Hill (5.5% increase)
- 78th - Park Slope section of Brooklyn that contains Prospect Park (3.1% increase)
- 100th - Arverne, Belle Harbor, Breezy Point, Broad Channel, Neponsit, Rockaway Park, Rockaway Beach, and Roxbury neighborhoods of Queens (2.9% increase)
- 76th - South Brooklyn, including the neighborhoods of Carroll Gardens, Red Hook, Cobble Hill, parts of Gowanus, and the Columbia Street Waterfront District (2.6% increase)
- 94th - Northernmost portion of Brooklyn, consisting of, primarily, the neighborhood of Greenpoint (2.3% increase)
- 5th - Southeastern edge of Manhattan, home to Chinatown, Little Italy, and the Bowery (2.3% increase)
- 109th - Northeast portion of Queens, including Downtown Flushing, East Flushing, Queensboro Hill, College Point, Malba, Whitestone, Beechhurst, and Bay Terrace (2.3% increase)

It is interesting to see that all of these precincts except for the 109th are in affluent or gentrifying areas of NYC, reinforcing the theory that violation charges (the lowest level cannabis charge) were more likely in affluent neighborhoods.

Interestingly, the jurisdictions with the strongest relationship to violation possession arrests were not the NYPD but the N.Y. Transit Police (8% increase), NYC Parks (4.3% increase), and "other" (3.2% increase). This shows that the NYPD is not very involved with the lowest level of possession arrests.

There were many public housing developments (or projects) that had a strong relationship with violation possession crimes, again highlighting that public housing developments were home to many cannabis arrests. They included:

- Boynton Avenue Rehab in the Soundview neighborhood of the South Bronx (6.8% increase),
- Bronxdale (now named Sotomayor Houses) in the Soundview neighborhood of the South Bronx (6.6% increase),
- Soundview in the Soundview neighborhood of the South Bronx (6.5% increase),
- Bronx River in the Van Nest neighborhood of the South Bronx (6.1% increase),
- Monroe in the Soundview neighborhood of the South Bronx (5.5% increase),
- Throggs Neck Addition in the Throggs Neck neighborhood of the South Bronx (5.1% increase),
- Gompers in the Lower East Side of Manhattan (4.3% increase),
- Van Dyke I in the Brownsville neighborhood of Eastern Brooklyn (4.2% increase),
- Pink Houses in "The Hole" neighborhood near Brownsville in Eastern Brooklyn (4%

increase),
- Amsterdam on the Upper West Side of Manhattan (3.9% increase),
- Throggs Neck in the Throggs Neck neighborhood of the South Bronx (3.8% increase),
- Hammel in the Arverne neighborhood of Southern Brooklyn (3.7% increase),
- Rutgers in the Lower East Side of Manhattan (3.4% increase),
- Audubon in Upper Manhattan (3.3% increase),
- Highbridge Gardens in the Mount Eden neighborhood of the South Bronx (3.3% increase),
- Long Island Baptist Houses in the New Lots neighborhood near Brownsville in Eastern Brooklyn (3.1% increase),
- Hughes Apartments in the Brownsville neighborhood of Eastern Brooklyn (3.1% increase),
- Jacob Riis in the Alphabet City neighborhood of Lower Manhattan (3.1% increase),
- Eastchester Gardens in the East Bronx (2.9% increase),
- Cypress Hills in the Cypress Hills neighborhood of East Brooklyn (2.7% increase),
- Vladeck in the Lower East Side of Manhattan (2.7% increase),
- Albany in the Crown Heights neighborhood of Central Brooklyn (2.6% increase),
- Sumner in the Bedford-Stuyvesant neighborhood of Central Brooklyn (2.6% increase),
- Roosevelt II in the Bedford-Stuyvesant neighborhood of Central Brooklyn (2.6% increase),
- Wald in the Alphabet City neighborhood of Southern Manhattan (2.6% increase),
- Unity Plaza (Sites 4-27) in the Brownsville neighborhood of East Brooklyn (2.3% increase),
- Bushwick in the Bushwick neighborhood of Northern Brooklyn (2.4% increase),
- Howard in the Weeksville neighborhood close to Crown Heights in Central Brooklyn (2.2% increase),
- Jackson in the Concourse Village of the South Bronx (2.2% increase),
- Edenwald in the Northern part of the Bronx (2.2% increase),
- LaFayette in the Clinton Hall neighborhood of Central Brooklyn (2.2% increase), and
- Lenox Road-Rockaway Parkway in the Brownsville neighborhood of East Brooklyn (2.1% increase)

The “feature families” that had the strongest relationship to violation cannabis possession crime, in descending order, were public housing developments (“projects”), NYC parks, premises types, police precincts, and L1 distances from NYC landmarks. After looking at the specific features within the “feature families”, no new features with a significant relationship with violation cannabis possession crime were discovered.

Unlike misdemeanor possession crimes, African American suspects have a positive relationship with violation possession (0.39 coefficient), showing that if a cannabis crime's suspect is African-American, there is a 1.5% increased likelihood that it is a violation possession crime. This isn't an overly strong increased likelihood, and could be due to the connection between violation possession crimes and housing developments in the South Bronx. Again, unknown suspect race has the strongest relationship to violation possession crime (0.46 coefficient), showing that there was an unknown confounding variable for the arresting officer not recording the suspect's race.

Felony Cannabis Possession:

As detailed above, the F1 score for the felony possession class was only 0.08 when using the best performing Logistic Regression algorithm 'upsampled_fp_15'. However, the coefficients can still be valuable in identifying salient features of felony possession crimes to a certain degree if the F1 score

is higher than zero. The felony possession class was predicted 11,899 times and the other cannabis crime types as a group are predicted 32,162 times. Therefore, the coefficients for the felony possession class are valuable and are analyzed below.

The specific premises features with the highest coefficients show that felony possession arrests had the strongest relationship with hotels and motels, livery licensed taxis, drug stores, shoe stores, storage facilities, unclassified stores, yellow licensed taxis, commercial buildings, unlicensed livery taxis, variety stores, residential houses, beauty and nail salons, and telecomm stores, in descending order of increased likelihood. The increased likelihood of a cannabis crime being a felony possession crime for these features ranged from 2% to 20%, with hotels and motels having a 20% increased likelihood, livery licensed taxis having a 18% increased likelihood, and the other premise types ranging from 2% to 8% increased likelihood. This suggested that felony possession with intent to distribute was occurring inside of legitimate businesses as a side business, and that hotels/motels and livery licensed taxis were the most common premises where felony possession occurs. On a related note, "inside" of these premises was where felony possession arrests most often occur, denoted by the coefficient of 0.71 and increased likelihood of 2% for the "LOC_OF_OCCUR_DESC_INSIDE" feature.

There was interestingly only one NYC park that had a strong relationship with felony possession charges, and that was Canarsie Park in the Canarsie neighborhood of Eastern Brooklyn. This park had a coefficient of 7, and an increased likelihood of a cannabis arrest being a felony possession arrest of 1,094%! This park was obviously the site of many felony possession arrests and was likely connected in some way to a major distribution hub.

Felony possession arrests had several strong relationships with specific subway/transit stations. These included (in descending order of coefficients and likelihoods):

- 47-50 Sts./Rockefeller Center in Midtown Manhattan (557% increased likelihood),
- W. 4th Street in the East Village (188% increased likelihood),
- Gates Avenue in the Bushwick neighborhood of Northern Brooklyn (146% increased likelihood),
- 25th Street in the South Slope neighborhood of Southern Brooklyn (88% increased likelihood),
- 241st St.-Wakefield in the Northern Bronx (68.5% increased likelihood),
- 207 St.-Inwood in the Inwood neighborhood of far northern Manhatttan (34.9% increased likelihood),
- Newkirk Avenue the Flatbush neighborhood of southern Brooklyn (30.2% increased likelihood),
- 205th St.-Norwood in the North Bronx (14% increased likelihood), and
- Woodlawn in the North Bronx (12% increased likelihood)

The following transit districts had strong to weak relationships with felony possession crimes:

- Transit District 11 around the Yankee Stadium in the South Bronx (4.5% increased likelihood), and
- Transit District 32 in the Crown Heights neighborhood of Central Brooklyn (3.5% increased likelihood)

If a cannabis crime suspect was male, there was a 3% increased likelihood that the crime was a felony possession crime.

If a cannabis crime suspect was Asian or Pacific Islander, there was a 2% increased likelihood that the crime was a felony possession crime.

For each full latitude/longitude unit closer a cannabis crime was to downtown Brooklyn in L1 terms, there was a 48.8% increased odds that it was a felony possession charge and not another cannabis crime. This shows that felony possession charges had a strong relationship with downtown Brooklyn and nearby areas.

For each full latitude/longitude unit closer a cannabis crime is to the Staten Island Ferry in L2 terms, there was a 8.2% increased odds that it is a felony possession charge and not another cannabis crime. This shows that felony possession charges had a relationship with the area surrounding the Ferry Terminal on Staten Island.

For each full latitude/longitude unit closer a cannabis crime was to Central Park in L1 terms, there was a 7.7% increased odds that it was a felony possession charge and not another cannabis crime. This showed that felony possession charges had a relationship with the neighborhoods surrounding Central Park. The 'central_park_l1' feature cannot specify which neighborhoods.

For each full latitude/longitude unit closer a cannabis crime was to Union Square in L1 terms, there is a 6.4% increased odds that it is a felony possession charge and not another cannabis crime. This showed that felony possession charges have a relationship with the neighborhoods surrounding Union Square.

For each full latitude/longitude unit closer a cannabis crime was to Prospect Park in L2 terms, there was a 2.5% increased odds that it was a felony possession charge and not another cannabis crime. This shows that felony possession charges had a relationship with the neighborhoods surrounding Union Square.

The borough of Queens had a weak relationship with felony possession crimes, with a 2.5% increased likelihood of cannabis crimes committed there being felony possession crimes.

Cannabis arrests made in the early morning had a 4.5% increased likelihood of being felony possession arrests. This was likely due to the police practice of early morning raids on locales suspected of containing felony possession amounts with intent to distribute.

Police precincts with the highest relationship to violation possession crime, in descending order, were the 22nd, 90th, 94th, 79th, 83rd, 20th, 45th, 10th, 34th, 13th, 105th, and 107th. These cover the following areas:

- 22nd - This precinct covers all of Central Park (7% increase)
- 90th - The northwestern portion of Brooklyn that mainly consists of the neighborhood of Williamsburg (4.3% increase)
- 94th - The northernmost portion of Brooklyn, consisting of, primarily, the neighborhood of Greenpoint (3.9% increase)

- 79th - The northernmost portion of Brooklyn, consisting of, primarily, the neighborhood of Greenpoint (3.9% increase)
- 83rd - Northern region in Brooklyn comprising Bushwick (2.7% increase)
- 20th - A northern portion of Brooklyn that includes Bedford Stuyvesant and features Herbert Von King Park (3.4% increase)
- 45th - A portion of the northeastern section of the Bronx (3% increase)
- 10th - Chelsea, Clinton/Hell's Kitchen South and the Hudson Yards neighborhoods (2.8% increase)
- 34th - Washington Heights and Inwood neighborhoods of northern Manhattan, north of West 179th Street (2.8% increase)
- 13th - A southern portion of Midtown, Manhattan. The precinct features the Peter Cooper Village/Stuyvesant Town residential complex, Gramercy Park, the lower portion of Rosehill, Madison Square Park, and Union Square Park (2.5% increase)
- 105th - An easternmost portion of Queens involving Queens Village, Cambria Heights, Laurelton, Rosedale, Springfield Gardens, Bellerose, Glen Oaks, New Hyde Park, and Floral Park (2.3% increase)
- 107th - A portion of Northern Queens, containing Fresh Meadows, Cunningham Heights, and Hilltop Village (2% increase)

The jurisdictions with the strongest relationship to felony possession arrests were not the NYPD but the N.Y. State Police (11.3% increase), Port Authority (4.6% increase), and Tri-Borough Bridge and Tunnel Authority. This shows that felony possession arrests were often being made by police groups focused on crime issues that cross geographical boundaries and the trafficking of illegal goods. This suggests, but obviously does not prove, a lack of a line between felony possession and possession with intent to distribute. This makes sense as having more than an ounce of cannabis has long been considered as the amount that one would have if they are intending to distribute.

There were many public housing developments (or projects) that had a strong relationship with felony possession crimes. These included:

- Baruch Houses Addition in the Lower East Side neighborhood of southern Manhattan (26.9% increase),
- Southern Boulevard in the South Bronx (15.9% increase),
- Howard Avenue-Park Place in the Weeksville neighborhood close to Crown Heights in Central Brooklyn (14.7% increase),
- Unity Plaza (Sites 4-27) in the Brownsville neighborhood of East Brooklyn (13.8% increase),
- Eagle Avenue-East 163rd Street in the Forest Houses neighborhood of the South Bronx (11.5% increase),
- Stuyvesant Gardens I in the Bedford-Stuyvesant neighborhood of Central Brooklyn (10.7% increase),
- Lehman in East Harlem (8.8% increase),
- University Avenue Rehab in the Morris Heights neighborhood of the Central Bronx (8.2% increase),
- Farragut in the Vinegar Hill neighborhood just north of downtown Brooklyn (7.5% increase),
- Seward Park Extension in the Lower East Side neighborhood of southern Manhattan (7.5% increase),
- Bronxdale (now named Sotomayor Houses) in the Soundview neighborhood of the South

Bronx (7.4% increase),
- Todt Hill in the Manor Heights neighborhood of central Staten Island (7.3% increase),
- Wald in the Alphabet City neighborhood of Lower Manhattan (7.1% increase),
- Twin Parks West (Site 1 & 2) in the Fordham Heights neighborhood of the west Bronx (7% increase),
- Claremont Rehab (Group 4) in the Morrisania neighborhood of the South Bronx (6.6% increase),
- Morrisania in the Morrisania neighborhood of the South Bronx (5.9% increase),
- Baruch in the Lower East Side neighborhood of southern Manhattan (5.5% increase),
- Hernandez in the Lower East Side neighborhood of southern Manhattan (5% increase),
- Van Dyke I in the Brownsville neighborhood of Eastern Brooklyn (4.9% increase),
- Red Hook East in the Red Hook neighborhood of Western Brooklyn (4.8% increase),
- Longfellow Avenue Rehab in the Soundview neighborhood of the South Bronx (4.8% increase),
- Vladeck in the Lower East Side of Manhattan (4.7% increase),
- Soundview in the Soundview neighborhood of the South Bronx (4.7% increase),
- Wyckoff Gardens in the Boerum Hill neighborhood of Western Brooklyn (4.7% increase),
- Douglass in the Upper West Side of Manhattan (4.7% increase),
- Red Hook West in the Red Hook neighborhood of Western Brooklyn (4.7% increase),
- Brownsville in the Brownsville neighborhood of Eastern Brooklyn (4.6% increase),
- Murphy in the Crotona neighborhood of in the central Bronx (4.6% increase),
- Manhattanville in the Manhattanville neighborhood of West Harlem (4.5% increase),
- Pomonok Houses in the Pomonok neighborhood of Queens (4.2% increase),
- Woodside in the Woodside neighborhood of Central Queens (4% increase),
- Ingersoll in downtown Brooklyn (3.6% increase),
- Roosevelt II in the Bedford-Stuyvesant neighborhood of Central Brooklyn (3.6% increase),
- Morrisania Air Rights 42 in the Morrisania neighborhood of the South Bronx (3.4% increase),
- Howard in the Weeksville neighborhood close to Crown Heights in Central Brooklyn (3.1% increase),
- Taft in East Harlem (3% increase),
- Boulevard in the East New York neighborhood of eastern Brooklyn (2.8% increase),
- Rangel in the Washington Heights neighborhood of northern Manhattan (2.8% increase),
- Monroe in the Soundview neighborhood of the South Bronx (2.8% increase),
- Butler in the Claremont neighborhood of central Bronx (2.6% increase),
- Astoria in the Astoria neighborhood of Queens (2.6% increase),
- Albany in the Crown Heights neighborhood of Central Brooklyn (2.5% increase),
- Jacob Riis in the Alphabet City neighborhood of Lower Manhattan (2.4% increase),
- Linden in the East New York neighborhood of eastern Brooklyn (2.4% increase),
- Morris II in the Claremont neighborhood of central Bronx (2.4% increase),
- Melrose in the South Bronx (2.3% increase),
- O'Dwyer Gardens in the Coney Island neighborhood of southern Brooklyn (2.3% increase),
- Cooper Park in the East Williamsburg neighborhood of northern Brooklyn (2.1% increase),
- Saint Mary's Park in the Mott Haven neighborhood of the South Bronx (2.1% increase),
- Sterling Place Rehab (Sterling-Buffalo) in the Weeksville neighborhood of Central Brooklyn close to Crown Heights (2.1% increase), and
- Whitman in the Fort Greene neighborhood of Brooklyn (2% increase)

The “feature families” that had the strongest relationship to felony cannabis possession crime, in descending order, were public housing developments (“projects”), premises types, police precincts, transit stations, and L1 distances from NYC landmarks. After looking at the specific features within the “feature families”, no new features with a significant relationship with felony cannabis possession crime were discovered.

The suspect race categories that have a relationship with felony possession crime are unexpected, given the fact that the majority of cannabis crimes where the suspect's race was reported are of African-Americans and Latinos. The suspect race category that has the strongest relationship with felony possession cries are Asians and Pacific Islanders, with a coefficient of 0.7 and an increased likelihood of 2%. American Indians/Alaskan Natives have a weak positive relationship with felony possession (coefficient of 0.18), as do the felony possession crimes with no suspect race reported (0.13). Meanwhile, white Hispanics have a totally neutral relationship, whites have a very weak negative relationship (-0.07), African-American Hispanics have a weakly negative relationship (-0.14), and African-Americans have a moderate negative relationship with felony cannabis possession crimes (-0.42). This shows that in the context of cannabis possession, African-Americans have some kind of relationship with violation possession charges, but not really misdemeanor or felony possession charges.

Misdemeanor Cannabis Sales:

As detailed above, the F1 score for the misdemeanor sales class is only 0.18 when using the best performing Logistic Regression algorithm 'upsampled_ms_1'. However, the coefficients can still be valuable in identifying salient features of misdemeanor sales crimes to a certain degree if the F1 score is higher than zero. The misdemeanor sales class was predicted 14,477 times and the other cannabis crime types as a group were predicted 29,584 times. Therefore, the coefficients for the misdemeanor sales class were valuable and are called and analyzed below.

The specific premises features with the highest coefficients showed that misdemeanor sales arrests had the strongest relationship with groceries and bodegas, liquor stores, fast food restaurants, candy stores, unclassified stores, gas stations, bars and nightclubs, department stores, book and greeting card stores, and variety stores. The increased likelihood of a cannabis crime being a felony possession crime for these features ranged from 2% to 6%.

Unlike cannabis possession arrests generally and felony possession arrests, misdemeanor sales arrests had several strong relationships with specific NYC parks. These included in descending order of increased likelihood that a cannabis crime was a misdemeanor sales crimes:

- Washington Square Park next to the NYU campus (with an increased odds of 15.8%),
- Story Playground in the Soundview neighborhood of the east Bronx (6.5% increase),
- Parkside Playground Brooklyn in the Prospect Lefferts Gardens neighborhood of central Brooklyn (5.8% increase),
- Matthews-Palmer Playground in the Hell's Kitchen neighborhood of Manhattan (3.6% increase),
- Powell Playground in the Crown Heights neighborhood of central Brooklyn (3.5% increase),
- St. James Park in the Fordham Manor neighborhood of central Bronx (3.4% increase),

- Gorman Playground in the Jackson Heights neighborhood of central Queens (2.9% increase),
- Harlem Art Park in East Harlem (2.9% increase),
- Maria Hernandez Park in the Bushwick neighborhood of north Brooklyn (2.9% increase),
- Park of the Americas in the Corona neighborhood of central Queens (2.8% increase),
- Shore Park and Parkway in the Bay Ridge neighborhood of southern Brooklyn (2.8% increase),
- Frank D. O'Connor Playground in the Elmhurst neighborhood of central Queens (2.3% increase),
- Unnamed Park on Summit Avenue (unclear location) (2.2% increase),
- Cooper Park in the East Williamsburg neighborhood of northern Brooklyn (2.2% increase),
- Corona Golf Playground in the Corona neighborhood of central Queens (2.1% increase),
- L/Cpl Thomas P. Noonan Jr. Playground in the Sunnyside neighborhood of central Queens (2.1% increase), and
- Harris Park in the Kingsbridge neighborhood of the central Bronx (2.1% increase)

Misdemeanor sales arrests had several strong relationships with specific subway/transit stations. These included (in descending order of coefficients and likelihoods):

- Jackson Avenue in the South Bronx (with an increased odds of 11.1%),
- 8th Avenue in the Chelsea neighborhood of southern Manhattan (increased odds of 5.36%),
- Kingston Avenue in the Crown Heights neighborhood of Central Brooklyn (4.6% increase),
- Pennsylvania Avenue in the East New York neighborhood of eastern Brooklyn (4.2% increase),
- Lexington Ave either near Grand Central Station in Midtown Manhattan or in East Harlem as the data label is not specific enough (4.1% increase),
- Smith-9th Streets in the Carroll Gardens neighborhood of southern Brooklyn (3.2% increase),
- 7th Avenue in Midtown Manhattan (3.2% increase),
- Beach 60th Street in Rockaway Beach (3.1% increase),
- 50th Street which could be the station near the Port Authority bus terminal or a station in the Bay Ridge neighborhood of southern Brooklyn (more likely the former) (2.8% increase),
- 219th Street in the Baychester neighborhood of the north Bronx (2.7% increase),
- 155th Street in West Harlem (2.5% increase),
- Parsons/Archer-Jamaica Center in the Jamaica neighborhood of eastern Queens (2.5% increase),
- Halsey Street in the Bushwick neighborhood of northern Brooklyn (2.3% increase),
- 137th St.-City College in West Harlem (2.2% increase),
- 241st St.-Wakefield in the Northern Bronx (2.2% increase),
- 34th St.-Herald Sq. in midtown Manhattan (2.1% increase),
- 42nd St.-Times Square in midtown Manhattan (2% increase)

Unknown or unrecorded suspect age had a 5.6% increased odds that the cannabis crime was a misdemeanor sales crime, showing that there is some underlying reason why the suspect's age was not being recorded for misdemeanor sales crimes.

Cannabis arrests with males were 2.2% more likely to be misdemeanor sales arrests than all other cannabis arrests. This shows that men were more likely to engage in misdemeanor cannabis sales. The causal reason can't be identified with this analysis.

For each full latitude/longitude unit closer a cannabis crime was to Yankee Stadium in the Bronx in L1 terms, there was a 4.6% increased odds that a cannabis arrest was a misdemeanor sales arrest.

Police precincts with the highest relationship to misdemeanor sales crime, in descending order, were the 6th, 13th, 33rd, 83rd, 34th, and 112th. These cover the following areas:

- 6th - Southwestern Manhattan neighborhoods of Greenwich Village and the West Village (4.7% increase), a liberal enclave housing most of the NYU campus and traditionally a bastion of the counterculture
- 13th - Southern portion of Midtown, Manhattan. The precinct features the Peter Cooper Village/Stuyvesant Town residential complex, Gramercy Park, the lower portion of Rosehill, Madison Square Park, and Union Square Park (3.5% increase). Union Square Park may be responsible for this relationship.
- 33rd - Washington Heights neighborhood of Northern Manhattan (3.4% increase), a traditionally Hispanic neighborhood
- 83rd - Northern region in Brooklyn comprising Bushwick (3.4% increase), a traditionally Hispanic neighborhood
- 34th - Communities of Washington Heights and Inwood, north of West 179th Street (2.6% increase), a traditionally Hispanic neighborhood
- 112th - Centrally located portion of Queens, housing Forest Hills and Rego Park, reflecting the strong relationship between the Pomonok Houses project in this portion of Queens (2.1% increase)

The jurisdictions with the strongest relationship to misdemeanor sales arrests were not the NYPD but the Staten Island Rapid Transit Authority, with an increased likelihood of their cannabis arrests being misdemeanor sales arrests of 2.3%.

The patrol borough with the strongest relationship to misdemeanor sales arrests is the Manhattan North patrol borough, with an increased likelihood of 2.9%.

There were many public housing developments (or projects) that had a strong relationship with misdemeanor sales crimes. These include:

- Wald in the Alphabet City neighborhood of Lower Manhattan (5.9% increase),
- Claremont Rehab (Group 3) in the Morrisania neighborhood of the South Bronx (4.5% increase),
- Grant in the Manhattanville neighborhood of Harlem (4.4% increase),
- Randall Avenue-Balcom Avenue (4.1% increase)
- Castle Hill in the Castle Hill neighborhood of the South Bronx (3.9% increase),
- Pomonok Houses in the Pomonok neighborhood of Queens (3.8% increase),
- Bay View in the East New York neighborhood of Brooklyn (3.7% increase),
- Astoria in the Astoria neighborhood of Queens (3.6% increase),
- Bland in the Willets neighborhood of eastern Queens (3.3% increase),
- Jacob Riis in the Alphabet City neighborhood of Lower Manhattan (3.2% increase),

- Glenwood in the East Flatbush neighborhood of central Brooklyn (3.2% increase),
- Boynton Avenue Rehab in the Soundview neighborhood of the South Bronx (3.1% increase),
- Sedgwick in the Morris Heights neighborhood of the South Bronx (2.7% increase),
- Queensbridge South between the Astoria and Hunters Point neighborhoods of Queens (2.7% increase),
- Soundview in the Soundview neighborhood of the South Bronx (2.7% increase),
- Baruch Houses Addition in the Lower East Side neighborhood of southern Manhattan (2.7% increase),
- Throggs Neck in the Throggs Neck neighborhood of the Bronx, which was shown to have a large percentage of cannabis arrests in the Data Story and EDA notebook (2.7% increase),
- Boston Secor in the Eastchester neighborhood of the north Bronx (2.6% increase),
- Crown Heights in the Weeksville neighborhood near Crown Heights in central Brooklyn (2.4% increase),
- Low Houses in the Brownsville neighborhood of eastern Brooklyn (2.4% increase),
- Vladeck II in the Lower East Side of Manhattan (2.3% increase),
- Farragut in the Vinegar Hill neighborhood just north of downtown Brooklyn (2.3% increase),
- Sterling Place Rehab (Sterling-Buffalo) in the Weeksville neighborhood of Central Brooklyn close to Crown Heights (2.3% increase),
- Sutter Avenue-Union Street in the Brownsville neighborhood of eastern Brooklyn (2.3% increase),
- Union Avenue-East 166th Street in the Forest House neighborhood of the South Bronx (2.3% increase),
- Breukelen in the Canarsie neighborhood of eastern Brooklyn (2.2% increase),
- La Guardia in the Lower East side of southern Manhattan (2.2% increase),
- Amsterdam on the Upper West Side of Manhattan (2.2% increase),
- Manhattanville in the Manhattanville neighborhood of West Harlem (2.1% increase), and
- Ingersoll in downtown Brooklyn (2% increase)

The “feature families” that had the strongest relationship to misdemeanor cannabis sales crime, in descending order, were public housing developments (“projects”), transit stations, NYC parks, premises types, and police precincts. After looking at the specific features within the “feature families”, no new features with a significant relationship with misdemeanor cannabis sales crime were discovered.

African-Americans (Hispanic and non-Hispanic) arrested for cannabis were much more likely to be arrested for misdemeanor sales, with a coefficient of 0.45 and 0.43 respectively. White Hispanics arrested for cannabis were more likely to be arrested for misdemeanor sales, with a coefficient of 0.23. Asians and Pacific islanders arrested for cannabis were also more likely to be arrested for misdemeanor sales, with a coefficient of 0.14. Meanwhile, Whites arrested for cannabis were much less likely to be arrested for misdemeanor sales, with a coefficient of -0.38. American Indians and Alaskan Natives had a -0.53 coefficient. The cause for this difference cannot be determined by this analysis, but the difference is stark. It may be that more African-Americans and Hispanics were selling cannabis than Whites, or it may be that they were being targeted differently by the police. Further research is definitely needed to look at this stark difference.

Felony Cannabis Sales:

As detailed above, the F1 score for the felony sales class was only 0.02 when using the best performing Logistic Regression algorithm 'upsampled_fs_10'. However, the coefficients can still be valuable in identifying salient features of felony sales crimes to a certain degree if the F1 score is higher than zero. The felony sales class was predicted 12,888 times and the other cannabis crime types as a group was predicted 31,173 times. Therefore, the coefficients for the felony sales class were valuable and were called and analyzed below.

The specific premises features with the highest coefficients show that felony sales arrests had the strongest relationship with telecommunication stores with an increased likelihood of 174% and clothing stores and boutiques with an increased likelihood of 116%. Chain stores, beauty and nail salons, dry cleaners and laundromats, and livery licensed taxis had an increased likelihood of 53.8%, 49.4%, 38.2%, 21% respectively. This really suggested that there are many people with legitimate businesses that are distributing cannabis on the side, as it is unlikely that people not associated with these businesses are coming into the business and conducting felony level sales. Other premises where felony level cannabis sales were occurring were public schools, highways and parkways, open areas and lots, bars and nightclubs, tunnels, fast food restaurants, unclassified stores, bus terminals, groceries and bodegas, residential houses, commercial buildings, others, streets, residential apartment houses, public parking lots and garages, and public buildings. The increased likelihood of a cannabis crime being a felony sales crime for these other features ranged from 3% to 20%.

There was only one NYC park associated with felony sales. If a cannabis crime occurred on Coney Island Beach and Boardwalk, there was an increased likelihood of 45.5% that it was a felony sales crime.

Felony sales arrests had several strong relationships with specific subway/transit stations. These included (in descending order of coefficients and likelihoods):

- Junius Street in the Brownsville neighborhood of eastern Brooklyn (7,260% increase),
- 96th Street unusually on the Upper East Side of Manhattan (2,870% increase),
- 3rd Avenue-149th Street in the South Bronx (206% increase),
- Dekalb Avenue in the Fort Greene neighborhood of Brooklyn (136% increase),
- Stillwell Avenue-Coney Island in the Coney Island neighborhood of southern Brooklyn (112% increase),
- Pacific Street (unclear location) (59.2% increase), and
- Myrtle Avenue in the Bushwick neighborhood of central Brooklyn (15.7% increase)

The following transit districts had strong to weak relationships with felony sales crimes:

- Transit District 32 in the Crown Heights neighborhood of Central Brooklyn (22.2% increased likelihood),
- Transit District 33 in the Bedford-Stuyvesant neighborhood of Central Brooklyn (8.9% increased likelihood),
- Transit District 3 in West Harlem (6.5% increase), and
- Transit District 34 in Coney Island, Brooklyn (2.3% increased odds)

Intriguingly, the 65+ suspect age group had a 3.3% increased likelihood of felony sales crime. This was a rather modest increased likelihood in a small group of felony sales crimes, so it's likely just due to a few individual sellers and not a deeper trend of older people selling cannabis.

The only suspect race group with an increased likelihood of felony sales crimes was White Hispanics, with a 2.1% increased likelihood.

For each full latitude/longitude unit closer a cannabis crime was to Prospect Park in L2 terms, there was a 89,900% increased odds that it was a felony sales charge and not another cannabis crime. This shows that felony sales charges have an extremely strong relationship with Prospect Park and nearby areas.

For each full latitude/longitude unit closer a cannabis crime was to the Manhattan Bridge in L1 terms, there was a 388% increased odds that it was a felony sales charge and not another cannabis crime.

For each full latitude/longitude unit closer a cannabis crime was to the Lincoln Center in L1 terms, there was a 137% increased odds that it was a felony sales charge and not another cannabis crime. For each full latitude/longitude unit closer a cannabis crime was to the Lincoln Center in L2 terms, there was a 52.2% increased odds that it was a felony sales charge and not another cannabis crime.

For each full latitude/longitude unit closer a cannabis crime is to the Port Authority Bus Terminal in L1 terms, there was a 102% increased odds that it was a felony sales charge and not another cannabis crime. For each full latitude/longitude unit closer a cannabis crime was to the Port Authority Bus Terminal in L2 terms, there was a 63.6% increased odds that it was a felony sales charge and not another cannabis crime.

For each full latitude/longitude unit closer a cannabis crime was to the Brooklyn Bridge in L1 terms, there was a 50.1% increased odds that it was a felony sales charge and not another cannabis crime.

For each full latitude/longitude unit closer a cannabis crime was to Washington Square Park in L2 terms, there was a 49.3% increased odds that it was a felony sales charge and not another cannabis crime.

For each full latitude/longitude unit closer a cannabis crime was to Mets Stadium in L1 terms, there was a 21.3% increased odds that it was a felony sales charge and not another cannabis crime. For each full latitude/longitude unit closer a cannabis crime was to Mets Stadium in L2 terms, there was a 15.2% increased odds that it was a felony sales charge and not another cannabis crime.

For each full latitude/longitude unit closer a cannabis crime was to the Apollo Theatre in L2 terms, there was a 12.3% increased odds that it was a felony sales charge and not another cannabis crime.

For each full latitude/longitude unit closer a cannabis crime was to the Metropolitan Detention Center in L1 terms, there was a 10.5% increased odds that it was a felony sales charge and not another cannabis crime.

For each full latitude/longitude unit closer a cannabis crime was to Penn Station in L2 terms, there was a 9.2% increased odds that it was a felony sales charge and not another cannabis crime.

For each full latitude/longitude unit closer a cannabis crime was to the New York Stock Exchange in L2 terms, there was a 4.1% increased odds that it was a felony sales charge and not another cannabis crime.

For each full latitude/longitude unit closer a cannabis crime was to Times Square in L1 terms, there was a 4% increased odds that it was a felony sales charge and not another cannabis crime. For each full latitude/longitude unit closer a cannabis crime was to Times Square in L2 terms, there was a 7.8% increased odds that it was a felony sales charge and not another cannabis crime.

For each full latitude/longitude unit closer a cannabis crime is to Union Square in L2 terms, there was a 6.5% increased odds that it is a felony sales charge and not another cannabis crime.

For each full latitude/longitude unit closer a cannabis crime was to Riker's Island in L1 terms, there was a 3.4% increased odds that it was a felony sales charge and not another cannabis crime.

For each full latitude/longitude unit closer a cannabis crime was to the World Trade Center in L2 terms, there was a 2.4% increased odds that it was a felony sales charge and not another cannabis crime.

The borough of Queens had a relationship with felony sales crimes. If a cannabis crime was committed in Queens, there was a 25.2% increased likelihood that it was a felony sales crime, showing that even though Queens as a whole didn't have a lot of cannabis arrests, there is a significant amount of felony sales arrests.

The patrol boroughs with the strongest relationship to felony sales arrests were the Bronx and Brooklyn North patrol boroughs, with an increased likelihood of 16.8% and 2.4%, respectively.

The holidays with a notable relationship to felony sales arrests were Thanksgiving and Martin Luther King Day, Jr., with an increased likelihood of 7.4% and 3.3% respectively.

Cannabis arrests made in the early morning had a 9.1% increased likelihood of being felony sales arrests, and those made in the morning rush hour had a 3.2% increased likelihood. This is likely due to the police practice of early morning raids on locales suspected of containing felony possession amounts with intent to distribute.

Police precincts with the highest relationship to felony sales crime, in descending order, were the 45th, 10th, 123rd, 20th, 19th, 14th, 79th, 109th, 13th, 69th, 5th, 34th, 62nd, 71st, 72nd, 122nd, 106th, 25th, 67th, 90th, 83rd, 61st, 112th, and 47th. These cover the following areas:

- 45th - A portion of the northeastern section of the Bronx (10.1% increase),
- 10th - Chelsea, Clinton/Hell's Kitchen South and the Hudson Yards neighborhoods (6.6% increase)
- 123rd - A portion of the South Shore of Staten Island, including Tottenville, Huguenot, Rossville, Annadale, and Eltingville (5.2% increase)
- 20th - A northern portion of Brooklyn that includes Bedford Stuyvesant and features Herbert Von King Park (4.9% increase)
- 19th - The Upper East Side of Manhattan (4.8% increase)

- 14th - The southern portion of Midtown, Manhattan. The area contains commercial offices, hotels, Times Square, Grand Central Terminal, Penn Station, Madison Square Garden, Koreatown section, and the Manhattan Mall Plaza (4.6% increase)
- 79th - Northern portion of Brooklyn that includes Bedford Stuyvesant and features Herbert Von King Park (4.5% increase)
- 109th - Northeast portion of Queens, including Downtown Flushing, East Flushing, Queensboro Hill, College Point, Malba, Whitestone, Beechhurst, and Bay Terrace (4.5% increase)
- 13th - Southern portion of Midtown, Manhattan. The precinct features the Peter Cooper Village/Stuyvesant Town residential complex, Gramercy Park, the lower portion of Rosehill, Madison Square Park, and Union Square Park (4.1% increase)
- 69th - The Canarsie section of eastern Brooklyn (3.4% increase)
- 5th - Southeastern edge of Manhattan, home to Chinatown, Little Italy, and the Bowery (3% increase)
- 34th - Communities of Washington Heights and Inwood, north of West 179th Street (2.7% increase), a traditionally Hispanic neighborhood (2.8% increase)
- 62nd - Southwestern portion of Brooklyn, home to Bensonhurst, Mapleton, and Bath Beach (2.4% increase)
- 71st - Crown Heights, Wingate, Prospect Lefferts neighborhoods of Central Brooklyn (2.4% increase)
- 72nd - Northwestern portion of Brooklyn that includes Sunset Park and Windsor Terrace (2.3% increase)
- 122nd - A portion of the South Shore of Staten Island, which encompasses Eltingville, Great Kills, Bay Terrace, Oakwood Heights, Oakwood Beach, Lighthouse Hill, New Dorp, Grant City, Midland Beach, South Beach, Todt Hill, South Beach, Old Town, and Grasmere (2.3% increase)
- 106th - South Central Queens, and includes Ozone Park, South Ozone Park, Lindenwood, Howard Beach, and Old Howard Beach (2.2% increase)
- 25th - Northern portion of East Harlem, home to a large residential community as well as Marcus Garvey Park, Harlem Art Park, and the 125th Street Metro-North Station (2.2% increase)
- 67th - East Flatbush and Remsen Village neighborhoods of Central Brooklyn (2.2% increase)
- 90th - The northwestern portion of Brooklyn that mainly consists of the neighborhood of Williamsburg (2.1% increase)
- 83rd - Northern region in Brooklyn comprising Bushwick (2.1% increase)
- 61st - A southern portion of Brooklyn encompassing Kings Bay, Gravesend, Sheepshead Bay, and Manhattan Beach (2.1% increase)
- 112th - Centrally located portion of Queens, housing Forest Hills and Rego Park, reflecting the strong relationship between the Pomonok Houses project in this portion of Queens (2% increase)
- 47th - A northern portion of the Bronx, encompassing Woodlawn, Wakefield, Williamsbridge, Baychester, Edenwald, Olinville, Fishbay, and Woodlawn Cemetery (2% increase)

The jurisdictions with the strongest relationship to felony sales arrests, in descending order, were the following:

- N.Y. State Police (9.4% increase),

- Port Authority (5.1% increase),
- Other (3.5% increase),
- N.Y. Housing Police (2.5% increase),
- Police Dept NYC (2.4% increase),
- N.Y. Police Dept (2.4% increase)

There were many public housing developments (or projects) that had a strong relationship with felony sales crimes. These included:

- Douglass Addition on the Upper West Side of Manhattan (2,430% increase),
- Tapscott Street Rehab in the New Lots neighborhood of Central Brooklyn (512% increase),
- Sterling Place Rehab (Sterling-Buffalo) in the Weeksville neighborhood of Central Brooklyn close to Crown Heights (243% increase),
- Douglass on the Upper West Side of Manhattan (234% increase),
- Howard Avenue in the Brownsville neighborhood of Eastern Brooklyn (198% increase),
- Bushwick II (Groups B & D) in the Bushwick neighborhood of central Brooklyn (186% increase),
- Queensbridge North between the Astoria and Hunters Point neighborhoods of Queens (160% increase),
- Sack Wern in the Soundview neighborhood of the South Bronx (151% increase),
- Claremont Parkway-Franklin Avenue in the Claremont Village of the South Bronx (122% increase),
- Harlem River in the Harlem neighborhood of northern Manhattan (114% increase),
- Seward Park Extension in the Lower East Side neighborhood of southern Manhattan (107% increase),
- Lower East Side Rehab (Group 5) in the Lower East Side neighborhood of southern Manhattan (106% increase),
- Castle Hill in the Castle Hill neighborhood of the South Bronx (61% increase),
- Bayside-Ocean Bay Apts in the Arverne neighborhood of southern Brooklyn (59.7% increase),
- Soundview in the Soundview neighborhood of the South Bronx (56.3% increase),
- Albany II in the Crown Heights neighborhood of central Brooklyn (55.9% increase),
- Bushwick II (Groups A & C) in the Bushwick neighborhood of central Brooklyn (55.5% increase),
- Monroe in the Soundview neighborhood of the South Bronx (43.3% increase),
- Tilden in the Brownsville neighborhood of eastern Brooklyn (41.9% increase),
- Morris II in the Claremont neighborhood of central Bronx (39% increase),
- Van Dyke I in the Brownsville neighborhood of Eastern Brooklyn (38.8% increase),
- Bronx River in the Van Nest neighborhood of the South Bronx (38% increase),
- Fort Independence Street-Heath Avenue in the Van Cortlandt neighborhood of the Bronx (37% increase),
- Brownsville in the Brownsville neighborhood of Eastern Brooklyn (36.1% increase),
- Baruch in the Lower East Side neighborhood of southern Manhattan (34.9% increase),
- Red Hook West in the Red Hook neighborhood of Western Brooklyn (32.9%

increase),

- Brevoort in the Stuyvesant Heights neighborhood near Bedford-Stuyvesant in central Brooklyn (30.3% increase),
- Davidson in the Morrisania neighborhood of the South Bronx (29.9% increase),
- Roosevelt II in the Bedford-Stuyvesant neighborhood of Central Brooklyn (26.6% increase),
- Bay View in the East New York neighborhood of Brooklyn (25.9% increase),
- Wald in the Alphabet City neighborhood of Lower Manhattan (24.2% increase),
- Taft in East Harlem (23.7% increase),
- Glenwood in the East Flatbush neighborhood of central Brooklyn (22.3% increase),
- Pomonok Houses in the Pomonok neighborhood of Queens (18.9% increase),
- Moore Houses in the South Bronx (16.5% increase),
- Manhattanville in the Manhattanville neighborhood of West Harlem (15.6% increase),
- Richmond Terrace in northern Staten Island (15.3% increase),
- Jackson in the Concourse Village of the South Bronx (14.4% increase),
- Grant in the Manhattanville neighborhood of Harlem (12.5% increase),
- Saint Mary's Park in the Mott Haven neighborhood of the South Bronx (12.4% increase),
- Coney Island I (Site 1B) in the Coney Island neighborhood of southern Brooklyn (11.5% increase),
- Betances I in the South Bronx (11.2% increase),
- Ingersoll in downtown Brooklyn (9.4% increase),
- Gowanus in the Gowanus neighborhood of southern Brooklyn (9.2% increase),
- Bushwick in the Bushwick neighborhood of northern Brooklyn (9% increase),
- Borinquen Plaza I in the Bushwick neighborhood of northern Brooklyn (9% increase),
- Borinquen Plaza II in the Bushwick neighborhood of northern Brooklyn (8.8% increase),
- Butler in the Claremont neighborhood of central Bronx (6.4% increase),
- Johnson in East Harlem in northern Manhattan (6.3% increase),
- Williamsburg in the Williamsburg neighborhood of Brooklyn (5.9% increase),
- West Brighton I in northern Staten Island (5.7% increase),
- Armstrong I in Bedford-Stuyvesant, in a classically African-American neighborhood of Brooklyn (5.2% increase),
- Whitman in the Fort Greene neighborhood of Brooklyn (4.7% increase),
- Mitchel in the Mott Haven neighborhood of the South Bronx (3.5% increase),
- Sumner in the Bedford-Stuyvesant neighborhood of Central Brooklyn (3.5% increase),
- Lincoln in East Harlem in northern Manhattan (3.4% increase),
- Nostrand in the Sheepshead Bay neighborhood of southern Brooklyn (2.9% increase), and
- Throggs Neck in the Throggs Neck neighborhood of the South Bronx (2.6% increase)

The “feature families” that had the strongest relationship to felony cannabis sales crime, in descending order, were public housing developments (“projects”), premises types, transit stations, police precincts, and L1 and L2 distances from NYC landmarks. After looking at the specific features within the “feature families”, no new features with a significant relationship with misdemeanor cannabis sales crime were discovered.

For felony sales, there are several positive relationships with suspect racial/ethnic groups. The strongest relationship is with White Hispanics with a coefficient of 0.72 and an increased likelihood of a cannabis crime being committed by this group being a felony sales crime of 2.1%. Unknown/unrecorded suspect race also has a strong relationship with felony sales crime (coefficient of 0.49), showing that there is an unexplored reason why suspect race is not recorded for felony sales crimes. It is notable that White suspects and African-American suspects have a nearly equivalent relationship with felony sales crimes (0.41 and 0.40 respectively), while African-American Hispanics have no relationship. Asians and Pacific Islanders have a weak negative relationship with felony sales (-0.25), and Native Americans have a highly negative relationship with felony sales (-2.86).

Conclusion:

As can be seen in the coefficient analysis above, there were many commonalities and some interesting differences in the most salient predictive features for each cannabis crime type. At the most abstract level, the NYPD's dataset allows cannabis arrests to be analyzed on characteristics of the person arrested, the time they were arrested, and the place they were arrested.

Generally, young males belonging to African-American and Hispanic groups constituted a disproportionate percentage of those arrested for cannabis. When comparing cannabis crime to non-cannabis crime, all racial/ethnic groups had a stronger statistical relationship with non-cannabis crime than cannabis crime, as the latter only constituted approximately 3.5% of all crimes in New York City. However, African-Americans and Hispanics accounted for the majority of the relationship between race and crime generally.

Interesting differences emerged in the coefficient analysis when looking at race and ethnicity amongst those arrested for different cannabis crime types. When comparing possession crime to sales crimes, Whites had a stronger relationship to possession crimes (coefficient of 0.58), while Hispanic African-Americans, African-Americans, and White Hispanics had a fairly weak relationship to sales crimes (coefficients of 0.18, 0.17, and 0.03) respectively. Whites had a weak relationship with misdemeanor possession (coefficient of 0.21), African-Americans had a moderate relationship to violation possession (0.39), and Asians and Pacific Islanders had a strong relationship to felony possession (0.70).

African-Americans (Hispanic and non-Hispanic) arrested for cannabis were much more likely to be arrested for misdemeanor sales, with a coefficient of 0.45 and 0.43, respectively. White Hispanics and Asians/Pacific Islanders arrested for cannabis were also more likely to be arrested for misdemeanor sales, with a coefficient of 0.23 and 0.14, respectively. Meanwhile, Whites arrested for cannabis were much less likely to be arrested for misdemeanor sales, with a coefficient of -0.38. American Indians and Alaskan Natives had a -0.53 coefficient. The cause for this difference cannot be determined by this analysis, but the difference is stark. It may be that more African-Americans and Hispanics were selling cannabis than Whites, or it may be that they were being targeted differently by the police. Further research is definitely needed to look at this stark difference.

For felony sales, White Hispanics had the strongest relationship with a coefficient of 0.72 and an increased likelihood of a cannabis crime being committed by this group being a felony sales crime of 2.1%. It is notable that White suspects and African-American suspects had a nearly equivalent relationship with felony sales crimes (0.41 and 0.40 respectively), while African-American Hispanics had no relationship. Asians/Pacific Islanders and Native Americans had negative relationships with felony sales (-0.25 and -2.86).

Unknown/unrecorded suspect race also had several strong relationships with cannabis crime (besides misdemeanor sales), again showing that there are salient reasons why suspect race was not recorded for these crimes. The coefficients and increased likelihoods between the feature for unrecorded suspect race and the various cannabis crime types are as follows:

- Cannabis crime generally: 0.60 coefficient, 1.8% increased likelihood of a crime being a cannabis crime
- Possession: 0.75 coefficient, 2.1% increased likelihood of a cannabis crime being for possession
- Misdemeanor Possession: 0.43 coefficient, 1.5% increased likelihood of a cannabis crime being for misdemeanor possession; these crimes were also more likely to have their suspect's sex underreported.
- Violation Possession: 0.46 coefficient, 1.6% increased likelihood of a cannabis crime being for violation possession
- Felony Possession: 0.13 coefficient, 1.1% increased likelihood of a cannabis crime being for felony possession; these crimes were more likely to be related to male suspects
- Misdemeanor Sales: -0.39 coefficient, 1.5% decreased likelihood of a cannabis crime being for misdemeanor sales; curiously, unreported age and sex had a positive relationship with this class of crime
- Felony Sales: 0.49 coefficient, 1.6% increased likelihood of a cannabis crime being for felony sales

It is interesting that misdemeanor sales crimes have their suspect race recorded far more than the other cannabis crime types, and warrants further research as to why.

A crime suspect's age also had a strong relationship with cannabis crime. The strongest relationship was for suspects between the ages of 18-24, the second strongest for those aged less than 18, and the third strongest for those aged between 25-44. These relationships likely reflect that younger people were more likely to use cannabis, but it also shows that children were being arrested for cannabis use and that the relationship between children and cannabis crime is stronger than that between those aged 25-44, even though the 25-44 year old group was much larger than the less than 18 year old group. Furthermore, unknown or unrecorded age contributed a moderate increased odds for sales arrests, suggesting an underlying reason why the suspect's age was not being recorded for sales crimes specifically. These findings do point towards a need for future research.

Overall, cannabis arrests occurred either early in the morning for crimes connected to sales, or later in the evening and night for crimes connected to possession, when people were off of work and socializing. They peaked in 2010, but remained at high levels throughout the year range before dropping off in 2018. They remained fairly consistent throughout the months of the year, peaking in August and dropping off significantly in the holiday season. National and cultural holidays common

to the general populace did not have an overly strong relationship with cannabis crime, but holidays specific to minority populations did, with the festive inclusion of both St. Patrick's Day and April 20th.

Although cannabis arrests occurred in nearly every geographical part of New York City between 2006 and 2018, one can look at the police precinct data to see where they were most highly concentrated. These areas of highest concentration are in neighborhoods of uptown Manhattan like Harlem, Washington Heights, and Inwood; lower Manhattan neighborhoods like Greenwich Village, the Lower East Side, and areas around Union Square; various neighborhoods of the Southeast and North Bronx, neighborhoods in northern Brooklyn like Bushwick, Williamsburg and Greenpoint; central Brooklyn neighborhoods like Crown Heights, Clinton Hill, Fort Greene, Bedford-Stuyvesant, Flatbush and Brownsville; neighborhoods in easternmost Brooklyn like East New York and Cypress Hills; small areas of southern Queens and areas of northern Queens like Corona, Elmhurst and Jackson Heights; and a portion of the southern shore of Staten Island.

Sales arrests were more related to lower Manhattan, while possession arrests are more related to upper Manhattan. Misdemeanor possession charges were more connected to African-American and Hispanic neighborhoods of Brooklyn, the Bronx, and Queens, while violation possession charges were more connected to gentrified parts of Brooklyn. Felony possession was curiously connected to the precinct that was composed of Central Park, as well as the northern gentrified neighborhoods of Greenpoint, Williamsburg and Bushwick. Misdemeanor sales were strongly connected to lower Manhattan neighborhoods like Greenwich Village and Union Square, as well as areas in uptown Manhattan like Washington Heights and Inwood, Bushwick Brooklyn, and central areas of Queens like Rego Park and Forest Hills. Felony sales charges were related to many areas of the five boroughs, but were most strongly related to a northeastern section of the Bronx. A variety of transit stations and transit districts in these neighborhoods also showed a relationship to cannabis crime and its subtypes (as can be seen in each crime type's coefficient section).

More affluent areas in lower Manhattan and Brooklyn that have a close proximity to the Brooklyn Bridge have a stronger relationship to the low level violation possession crime class, as do areas of the Bronx and Queens that have a close proximity to Rikers Island. The features that compute the distance to NYC landmarks also reinforce the clustering of cannabis arrests around areas of Brooklyn, Upper Manhattan and the South Bronx, as well as a significant cluster around the transit hubs of midtown Manhattan where homelessness is prevalent.

Public housing developments had a strong relationship with cannabis crime and were home to 19% of total arrests, and most of these developments were located in the neighborhoods where street arrests were also made. The specific housing developments and their coefficients and increased likelihoods can be found in the Jupyter notebooks, but for cannabis crimes generally they are located in the Williamsburg and Bushwick neighborhoods of northern Brooklyn, the classically African-American neighborhoods of Bedford-Stuyvesant and Fort Greene in Central Brooklyn, in the gentrified neighborhood of Boerum Hill in southwestern Brooklyn, in the Throggs Neck neighborhood of the South Bronx, and in the Two Bridges neighborhood of Lower Manhattan near the Brooklyn Bridge.

Certain public housing developments (or projects) had a strong relationship with cannabis possession crimes. They included housing developments in the Morrisania neighborhood of the

South Bronx, the classically African-American neighborhoods of Bedford-Stuyvesant, Clinton Hill, and Brownsville in Central Brooklyn, the Edenwald neighborhood of the North Bronx, the Williamsburg neighborhood of northern Brooklyn, the small Weeksville neighborhood of Central Brooklyn next to Crown Heights, East Harlem, and the Throggs Neck and Eastchester neighborhoods of the South Bronx.

Cannabis sales crimes in housing developments were most closely related to projects in the Pomonok neighborhood of Central Queens, the Alphabet City neighborhood on the Lower East Side of Manhattan, Queensbridge in far western Queens, the South Bronx, the East New York neighborhood of eastern Brooklyn, the Astoria neighborhood of western Queens, west Harlem, downtown Brooklyn, the small Weeksville neighborhood of Central Brooklyn close to Crown Heights, the classically African-American neighborhoods of Bedford-Stuyvesant and Brownsville in Central Brooklyn.

Housing developments with the strongest relationship to misdemeanor possession were located in the classically African-American neighborhoods of Bedford-Stuyvesant and Brownsville in Central Brooklyn, the South Bronx, the North Bronx, the classically Puerto Rican neighborhood of Bushwick in northern Brooklyn, the Sheepshead Bay neighborhood of southern Brooklyn, and the classically Hispanic neighborhood of Inwood in northern Manhattan. Of note is that the premises type feature for public housing had a coefficient of 0.9 for misdemeanor possession crimes; none of the other cannabis crime subtypes had this relationship.

The housing developments with the highest relationship to violation possession are highly concentrated in the South and North Bronx, the Lower East Side of Manhattan (including the classically Hispanic neighborhood of Alphabet City), the Bedford-Stuyvesant, Clinton Hill, Brownsville and Cypress Hills neighborhoods of central Brooklyn, and one project in the Upper West Side of Manhattan.

The housing developments with the highest relationship to felony possession are located in the Lower East Side of southern Manhattan (including Alphabet City), the South, West and Central Bronx, the central Brooklyn neighborhood of Weeksville near Crown Heights, the central Brooklyn neighborhoods of Brownsville, Fort Greene, and Bedford-Stuyvesant, East and West Harlem, the tiny Vinegar Hill neighborhood of downtown Brooklyn, central Staten Island, the Red Hook neighborhood of far western Brooklyn, Central Queens, the Weeksville neighborhood next to Crown Heights in Central Brooklyn, the East New York neighborhood of eastern Brooklyn, the Coney Island neighborhood of far southern Brooklyn, and the Williamsburg neighborhood of northern Brooklyn.

The housing developments with the highest relationship to misdemeanor sales are located in the Lower East Side of Manhattan (including Alphabet City), the South Bronx, Harlem, the Pomonok neighborhood of Queens, the East New York neighborhood of eastern Brooklyn, the Astoria neighborhood of western Queens, the Willets neighborhood of eastern Queens, the heavily Afro-Caribbean East Flatbush neighborhood of central Brooklyn, the Weeksville neighborhood and Crown Heights in central Brooklyn, the Brownsville neighborhood of central Brooklyn, Vinegar Hill near downtown Brooklyn, the Canarsie neighborhood of eastern Brooklyn, west Harlem, and downtown Brooklyn.

The housing developments with the highest relationship to felony sales are located in the Upper West Side of Manhattan, the New Lots and Weeksville neighborhoods of Central Brooklyn, Brownsville, Fort Greene, East Flatbush, Bedford-Stuyvesant and Crown Heights in Central Brooklyn, the Bushwick neighborhood in northern Brooklyn, Queensbridge in western Queens, the South Bronx, west and east Harlem, the Lower East Side neighborhood of southern Manhattan, the Arverne neighborhood of southern Brooklyn, the Red Hook neighborhood of far western Brooklyn, the East New York neighborhood of eastern Brooklyn, Alphabet City in southern Manhattan, the Pomonok neighborhood of central Queens, northern Staten Island, the Coney Island and Sheepshead Bay neighborhoods of far southern Brooklyn, downtown Brooklyn, the Gowanus neighborhood of southern Brooklyn, and the Bushwick and Williamsburg neighborhoods of northern Brooklyn.

Certain New York City parks in these neighborhoods also have a relationship with cannabis crime, but mostly with sales, violation possession, and misdemeanor sales. When looking at the most abstract classification of cannabis crimes as differentiated from all other crimes, no parks show a relationship with cannabis crime. The same goes for possession crimes as a whole. For misdemeanor possession, only Claremont Park in the South Bronx and Riverside Park on the Upper West Side of Manhattan show moderate relationships. For felony possession, Canarsie Park in Canarsie Brooklyn shows an extremely strong relationship. For felony sales, only Coney Island Beach and Boardwalk show a strong relationship. A variety of parks in the Bronx, Queens, Brooklyn, and Manhattan have relationships with sales, violation possession, and misdemeanor sales crimes (see each crime type's coefficient section for details).

The jurisdictions with statistical relationships to cannabis crime differed by subtype. The jurisdictions with the strongest relationship to cannabis possession arrests were not the NYPD but the Tri-Borough Bridge and Tunnel Authority and the NY Transit Police. For violation possession, the jurisdictions with the strongest relationship were the N.Y. Transit Police, NYC Parks, and "other". The jurisdictions with the strongest relationship to felony possession arrests were not the NYPD but the N.Y. State Police, the Port Authority (4.6% increase), and the Tri-Borough Bridge and Tunnel Authority. For misdemeanor sales, the jurisdiction with the strongest relationship was the Staten Island Rapid Transit Authority. The jurisdictions with the strongest relationship to felony sales arrests, in descending order, were the N.Y. State Police, the Port Authority, other, the N.Y. Housing Police, and the New York Police Department.

For those crimes where a premise type was recorded by the arresting officer, possession arrests were more likely to occur in public spaces while sales arrests were more likely to occur in private businesses and residential locations. Several types of small businesses were uniquely related to cannabis sales, suggesting that small business owners in New York City may have been supporting the underground cannabis market to a certain extent.

The aim of this project was to use modern machine learning methods to provide a more comprehensive and complicated image of cannabis arrests in New York City between the years of 2006 and 2018 than was previously available using public data. Although much of the data supports what was reported in the media about the racial disparity in who was arrested for cannabis during this time, the findings reported here present a more detailed and nuanced resource for scholars and policy makers to study the issue and provide guidance to the law enforcement community as to how

to enforce cannabis law and general drug law in an equitable fashion. It is clear that African Americans and Hispanics in specific low-income neighborhoods bore the brunt of the negative effects of being arrested for cannabis, although the data is partial to a certain degree and invites further research. Of particular concern is the unreported fact that the vast majority of cannabis crimes did not have their demographic information collected by the arresting officer, which makes it difficult to fully understand the racial disparity and the motivations behind New York City's cannabis policy during this time.

The Jupyter notebooks created in this project can be modified to investigate any type of crime reported in the NYPD's data set. An offering that this project makes to the research community is a protocol for using the NYPD's data set in a clean way for predicting a wide variety of crimes, and for better understanding the statistical relationships between crime and its personal, temporal, and geographic characteristics. By understanding these relationships, new avenues of criminological and sociological research can be opened and explored. By using machine learning to open up the full landscape of statistical relationships, new insights and directions for research can be created that would not normally be generated by the human researcher's thoughts. By marrying the machine and the mind of the researcher, a more complete analysis can emerge of the challenges that crime brings to our society and that drug law brings to those that are differentially impacted by it.

Citations:

"New York State Penal Law". Article 221, No. 221 of 2016. Retrieved November 13, 2016.

Census Bureau Quick Facts on New York City:
https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork,bronxcountybronxboroughnewyork,kingscountybrooklynboroughnewyork,newyorkcountymanhattanboroughnewyork,queenscountyqueensboroughnewyork,richmondcountystatenislandboroughnewyork/PST045218

"Police Dept's Focus on Race Is at Core of Ruling Against Stop-and-Frisk Tactic", Joseph Goldstein, New York Times, August 14, 2013,

https://www.nytimes.com/2013/08/15/nyregion/racial-focus-by-police-is-at-core-of-judges-stop-and-frisk-ruling.html

Harcourt, B.E. & Ludwig, J., "Reefer Madness: Broken Windows Policing and Misdemeanor Marijuana Arrests in New York", University of Chicago Law School: Chicago Unbound, Working Papers, 2006,
https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1250&context=public_law_and_legal_theory

Levine, H., Sociology Department, Queens College, "Unjust and Unconstitutional", Marijuana Arrest Research Project and the Drug Policy Alliance, July 2017,
https://www.drugpolicy.org/sites/default/files/Marijuana-Arrests-NYC--Unjust-Unconstitutional--July2017_2.pdf

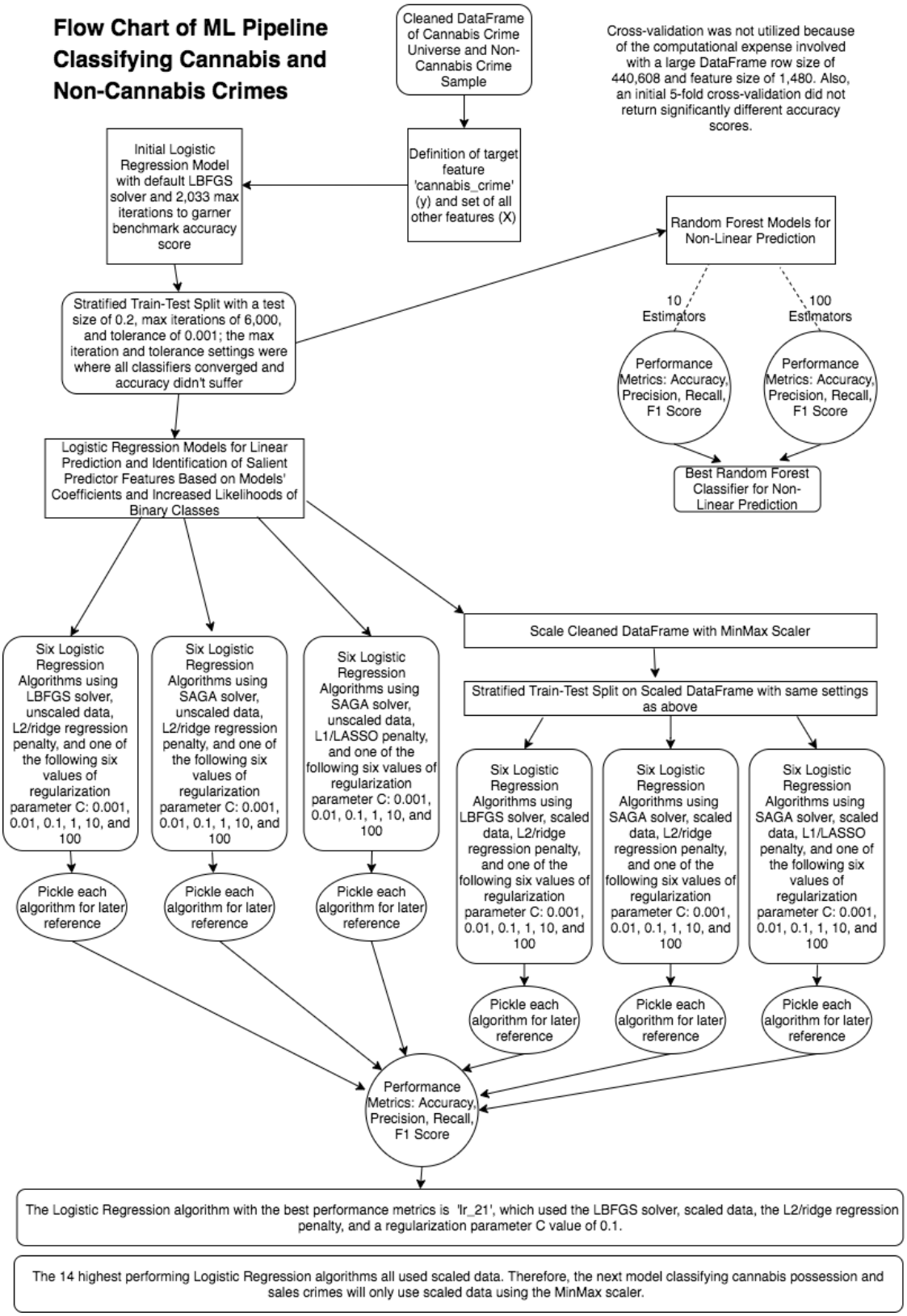
Mueller, B., Gebeloff, R., Chinoy, S., "Surest Way to Face Marijuana Charges in New York: Be Black or Hispanic", New York Times, May 13, 2018,
https://www.nytimes.com/2018/05/13/nyregion/marijuana-arrests-nyc-race.html

Substance Abuse and Mental Health Service Administration (SAMHSA), Results from the 2018 National Survey on Drug Use and Health: Detailed Tables (Washington, D.C.: SAMHSA, August 2019),

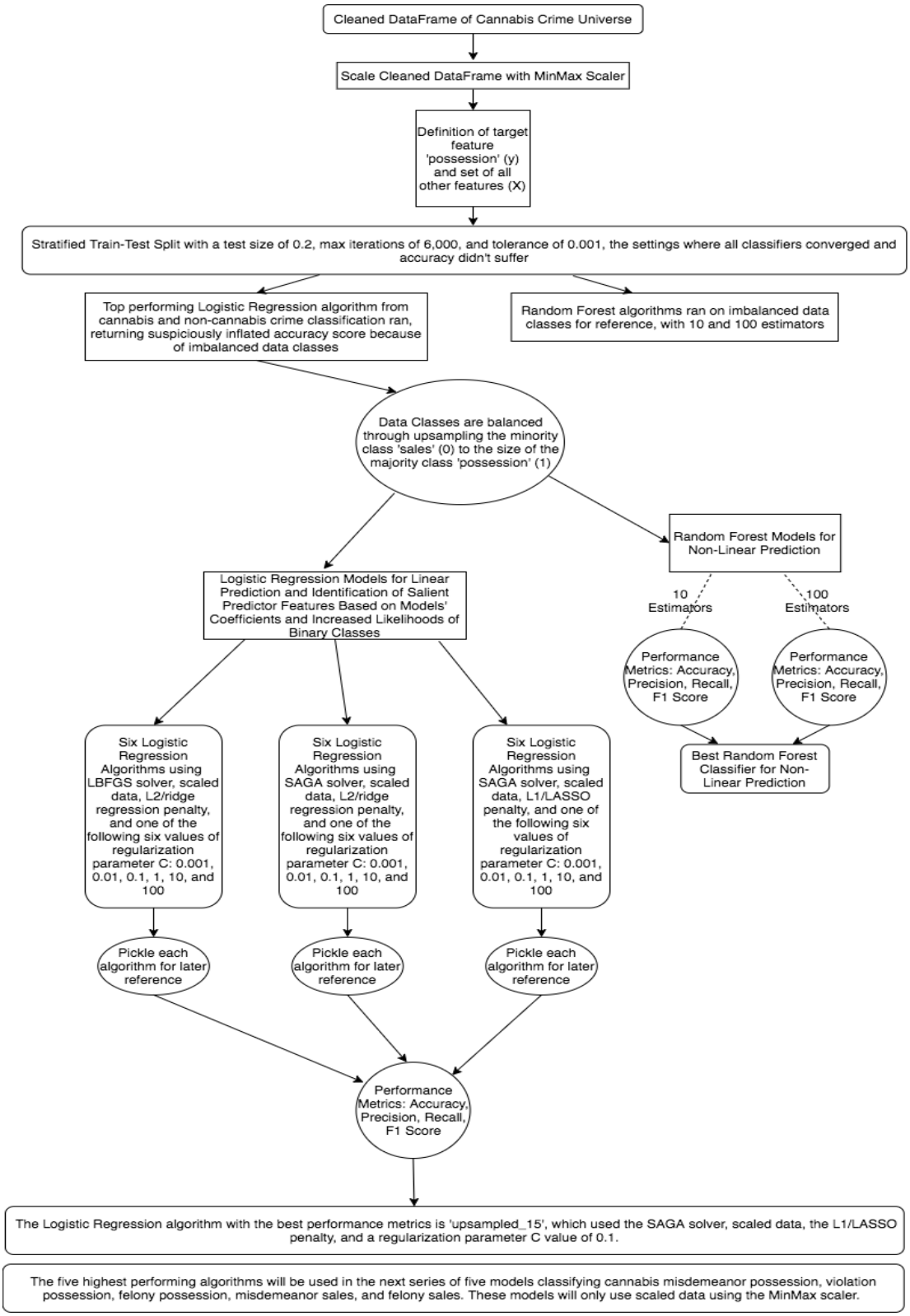
https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHDetailedTabs2018R2/NSDUHDetailedTabs2018.pdf

Table 1.26B: Marijuana Use in Past Year among Persons Aged 12 or Older, by Age Group and Demographic Characteristics: Percentages, 2017 and 2018.

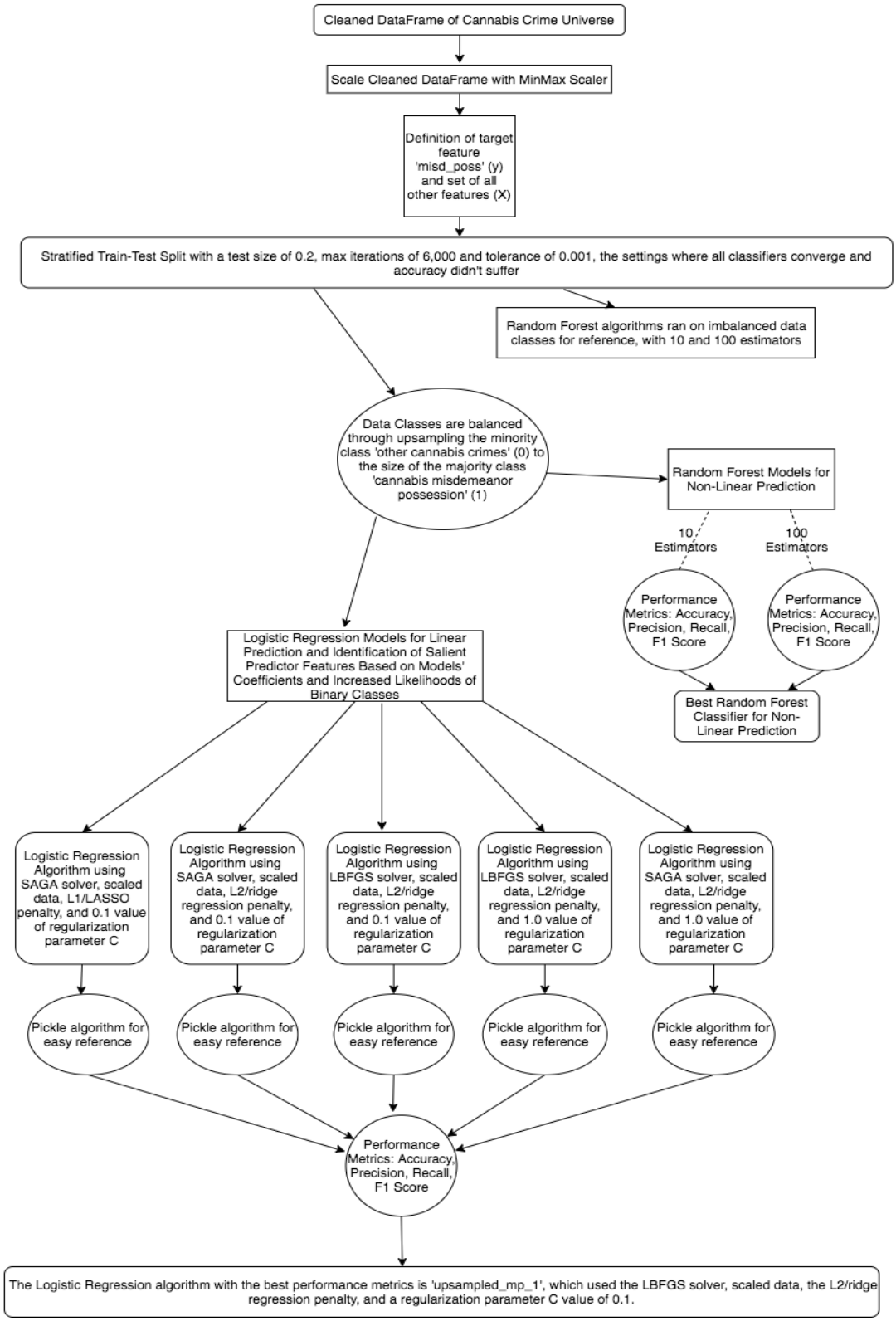
Appendix A - Flow Charts of Classification Pipeline



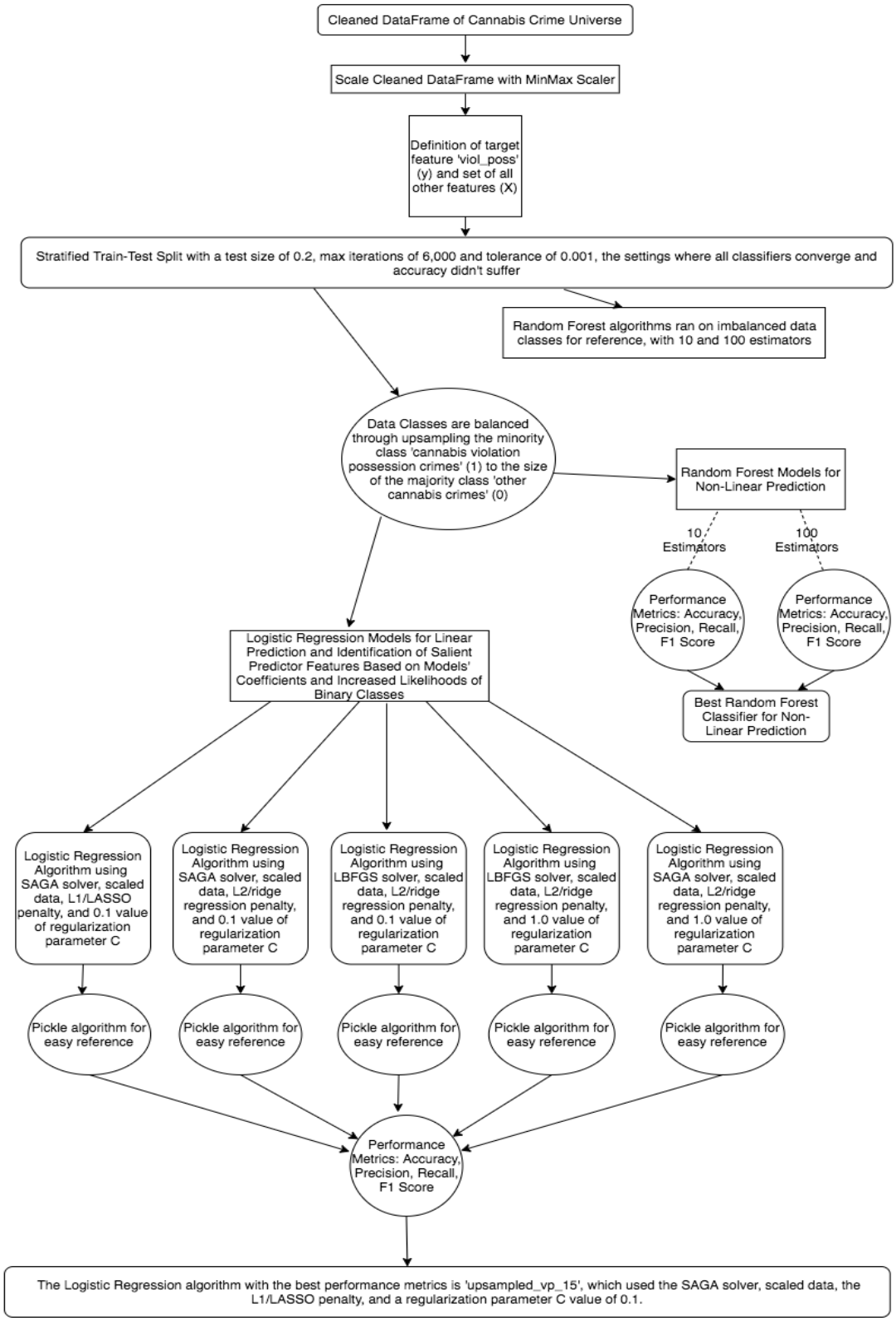
Flow Chart of ML Pipeline
Classifying Cannabis
Possession & Sales Crimes



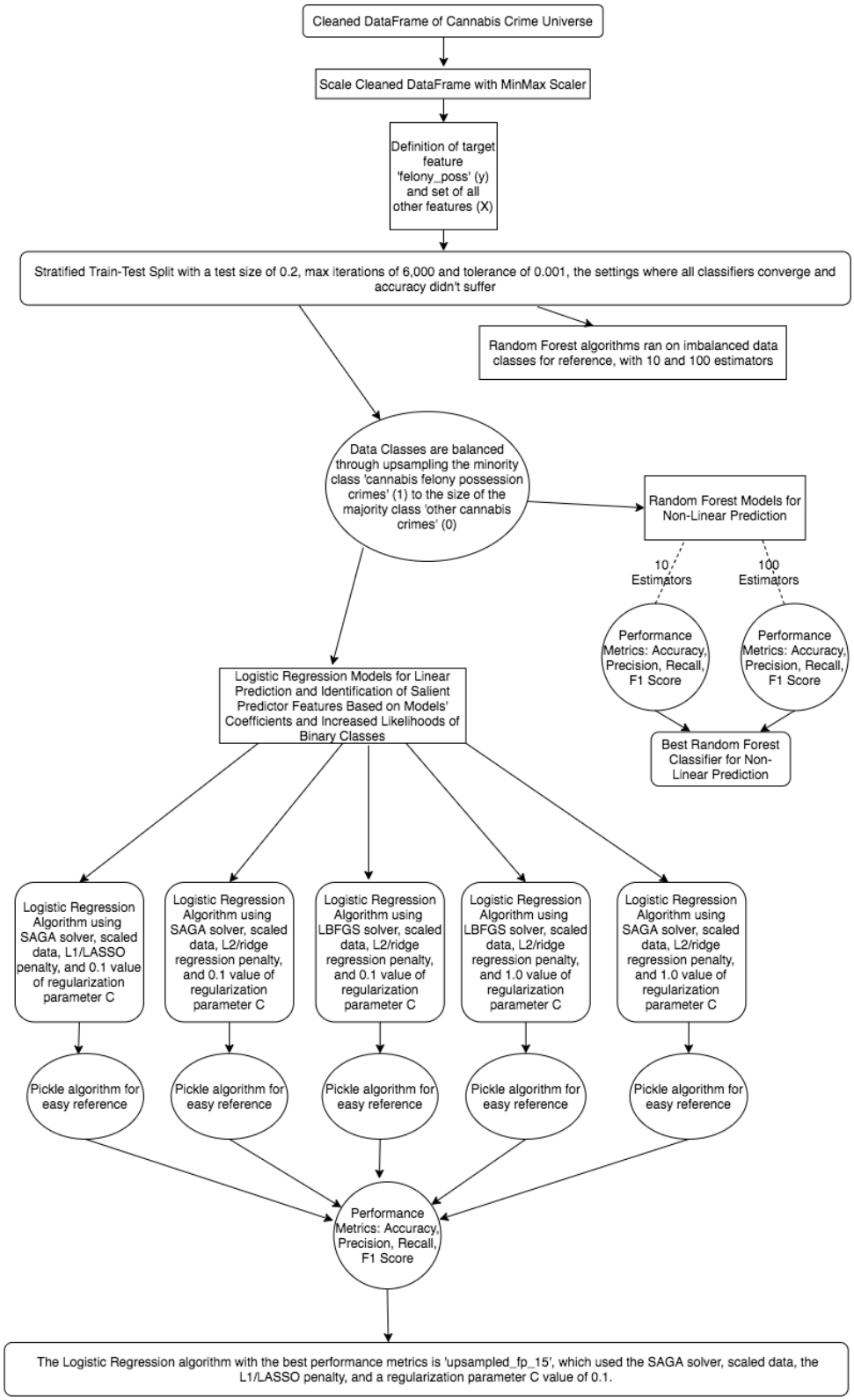
Flow Chart of ML Pipeline Classifying Misdemeanor Possession Crimes



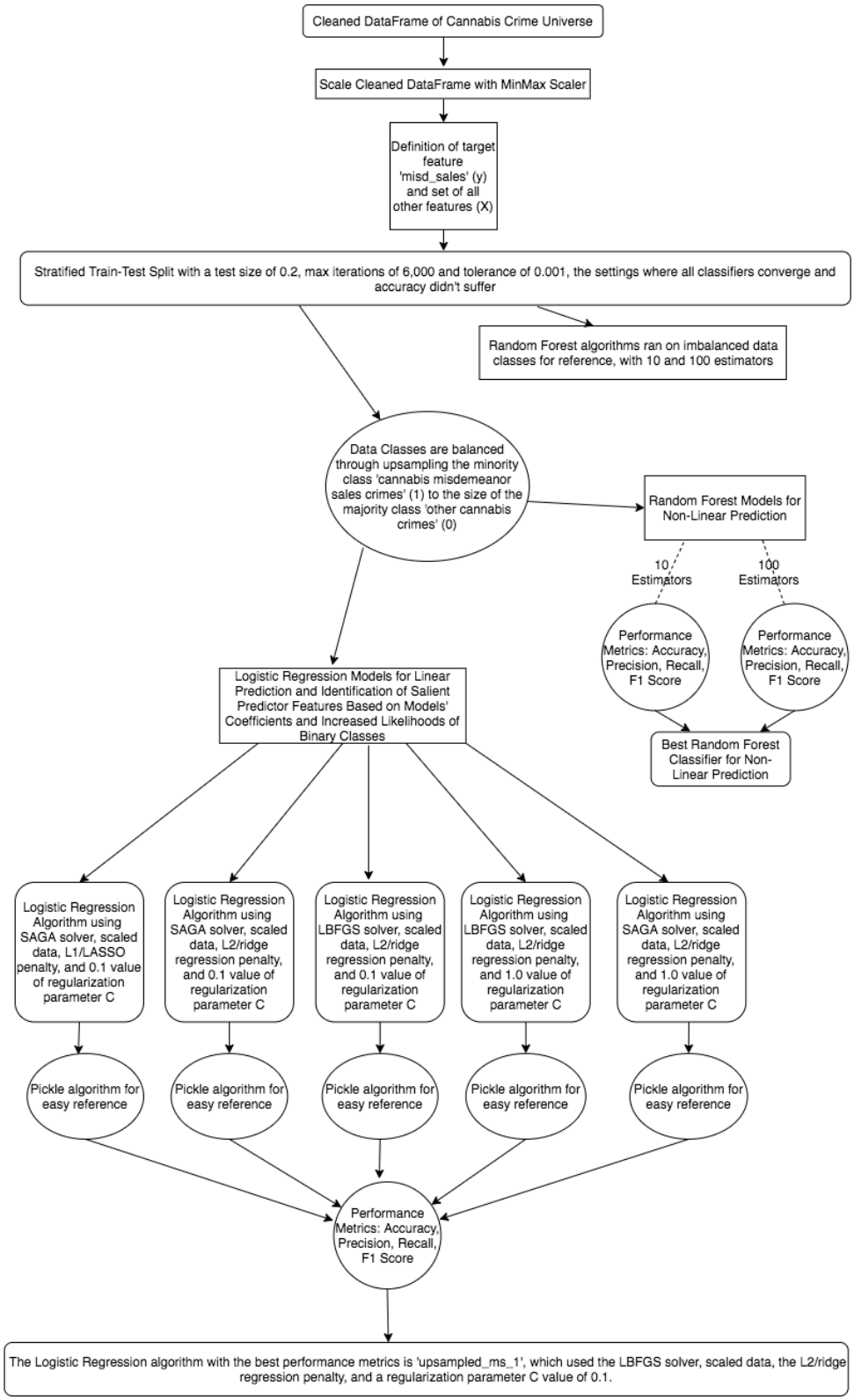
Flow Chart of ML Pipeline Classifying Violation Possession Crimes



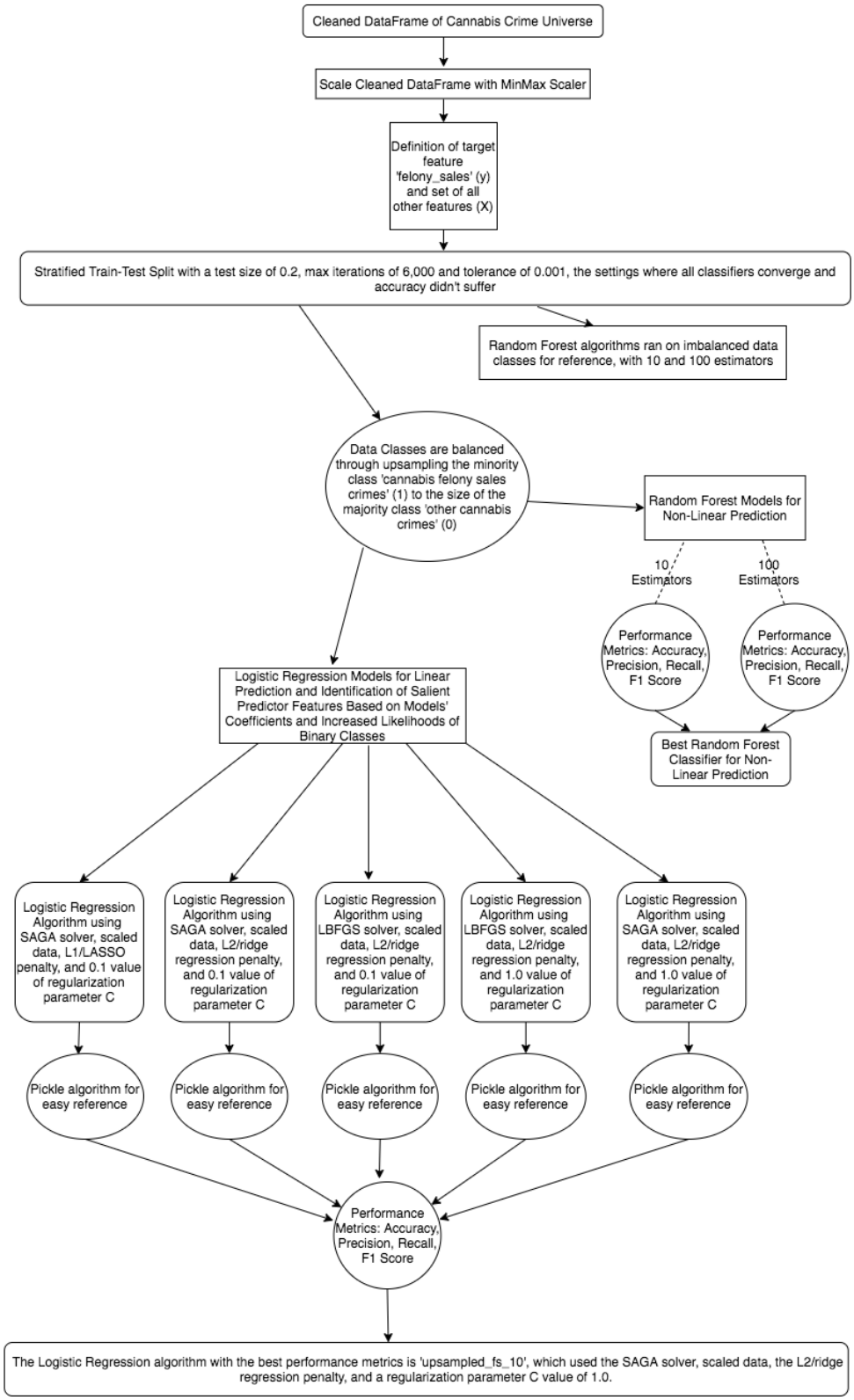
Flow Chart of ML Pipeline Classifying Felony Possession Crimes



Flow Chart of ML Pipeline Classifying Misdemeanor Sales Crimes

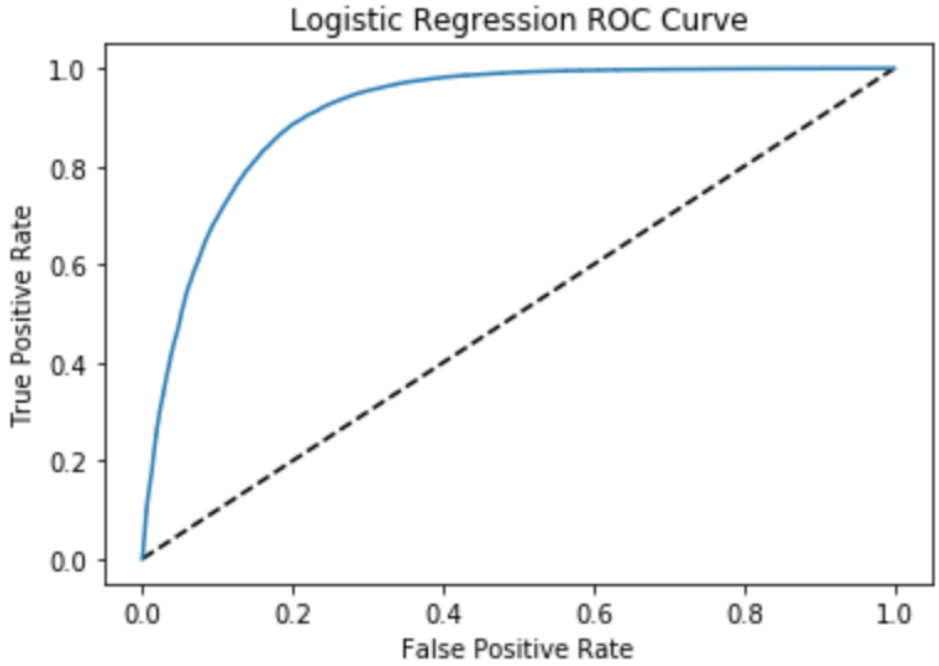


Flow Chart of ML Pipeline Classifying Felony Sales Crimes

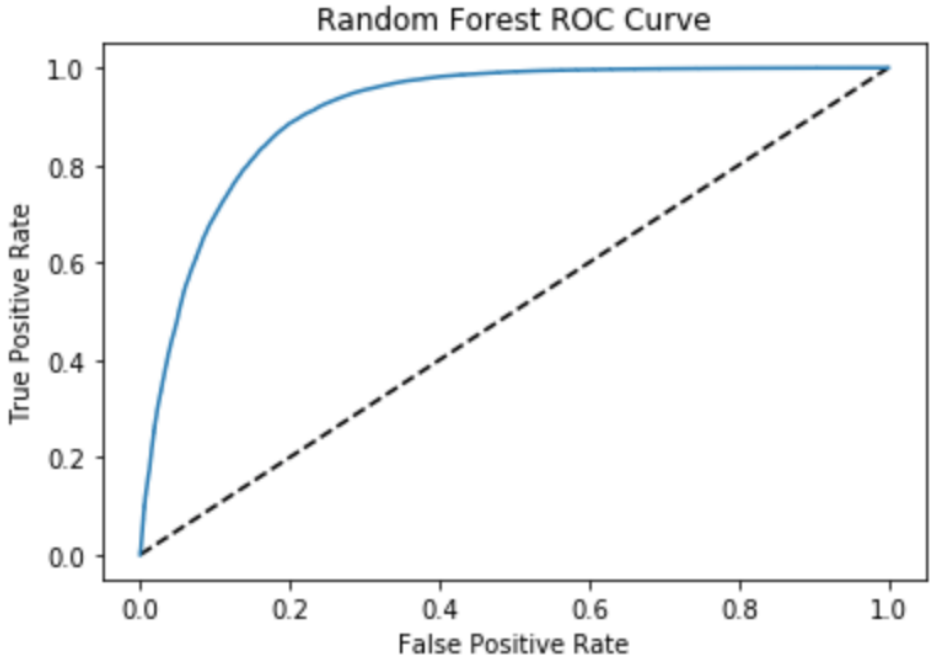


Appendix B - ROC and Precision-Recall Curves

Cannabis and Non-Cannabis Crimes:

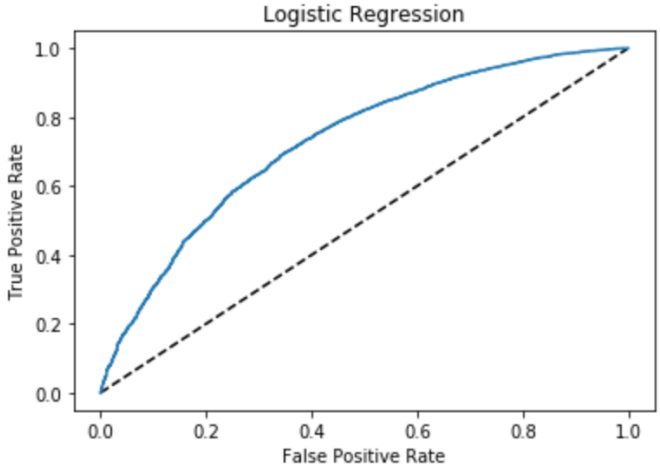


Area Under the Curve (AUC) score: 0.913

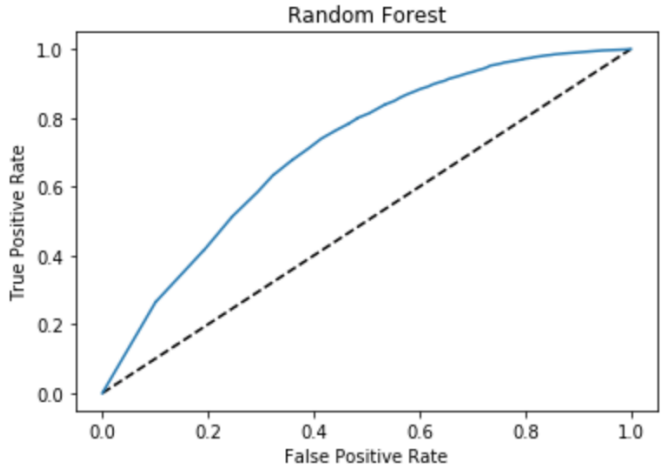


AUC score: 0.924

Cannabis Possession and Sales Crimes:
ROC curves:

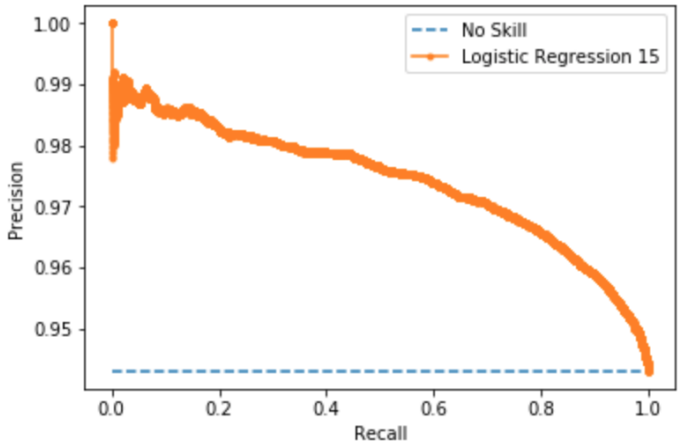


AUC score: 0.729



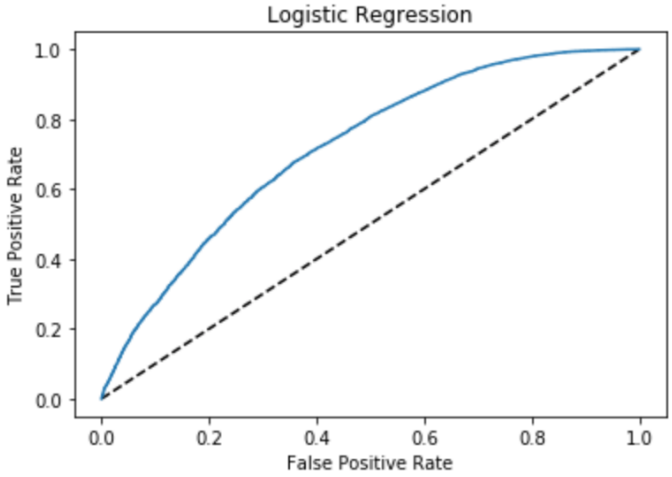
AUC score: 0.712

Precision-Recall Curve:

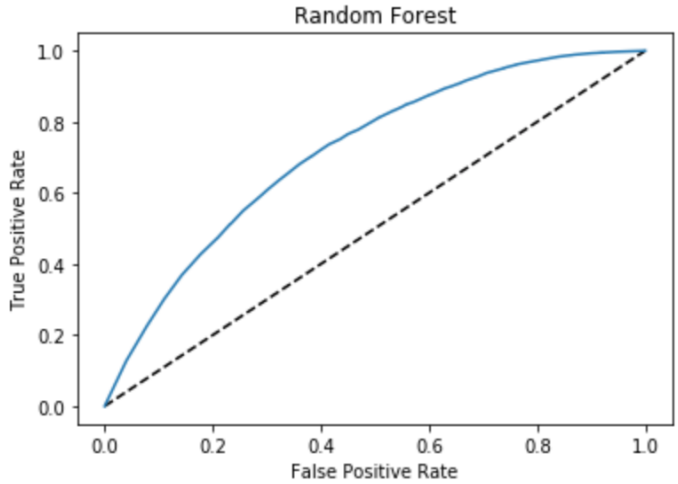


AUC score: 0.974

Cannabis Misdemeanor Possession:
ROC curves:

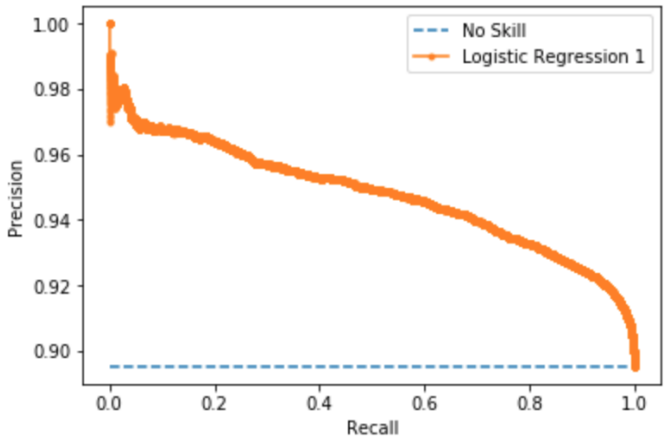


AUC score: 0.719



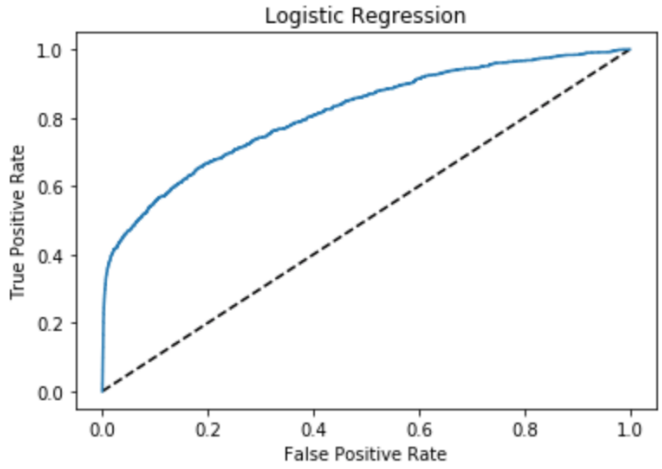
AUC score: 0.718

Precision-Recall Curve:

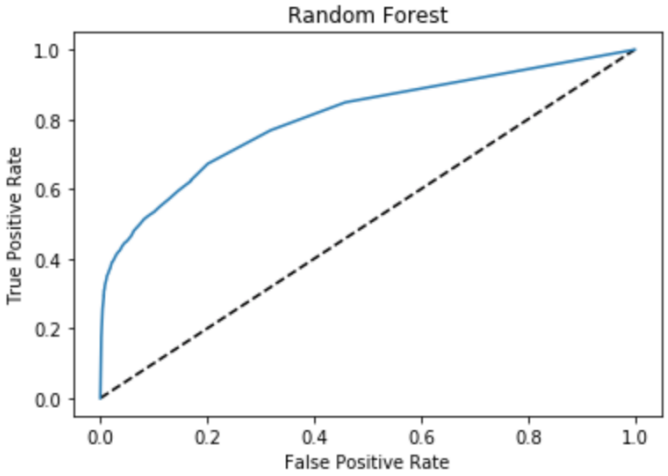


AUC score: 0.948

Cannabis Violation Possession:
ROC curves:

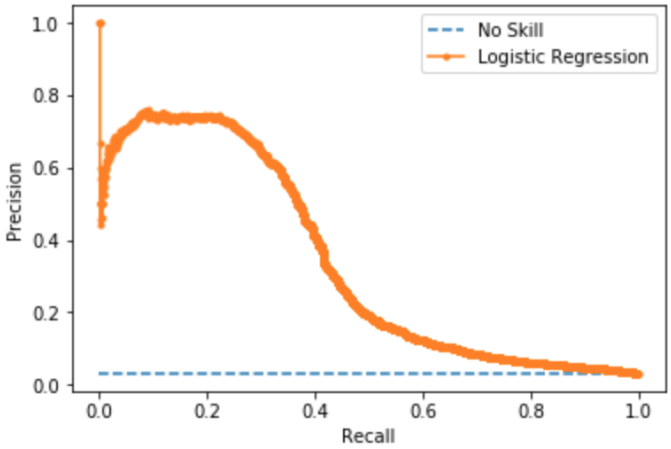


AUC score: 0.813



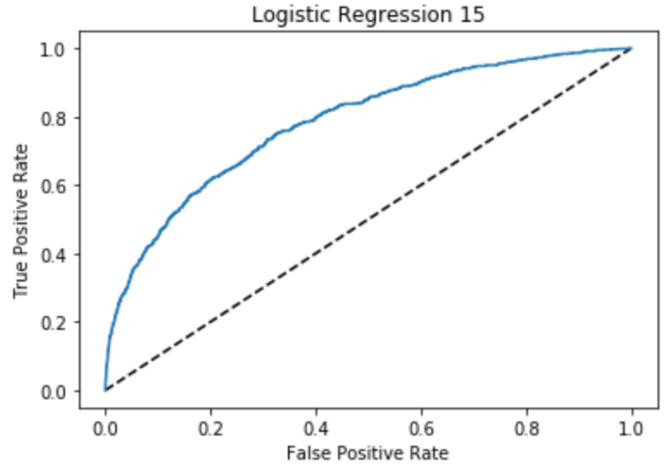
AUC score: 0.803

Precision-Recall Curve:

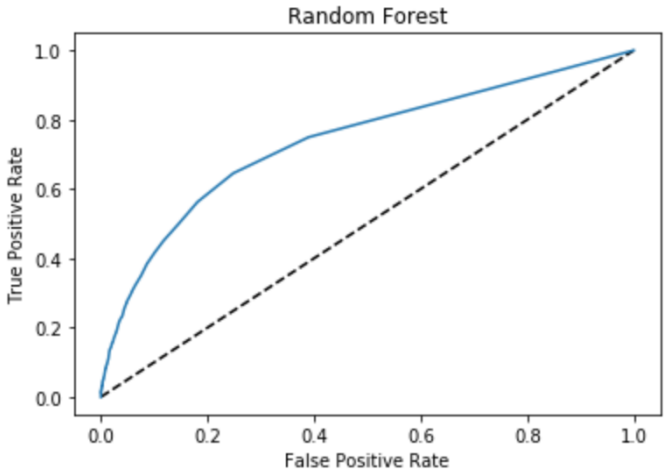


AUC score: 0.338

Cannabis Felony Possession:
ROC curves:

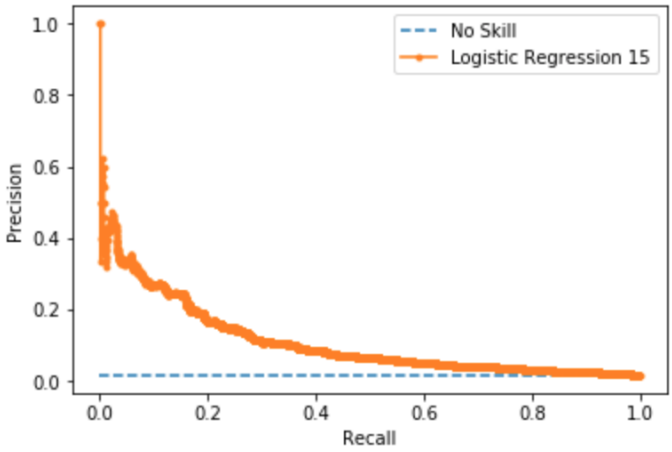


AUC score: 0.784



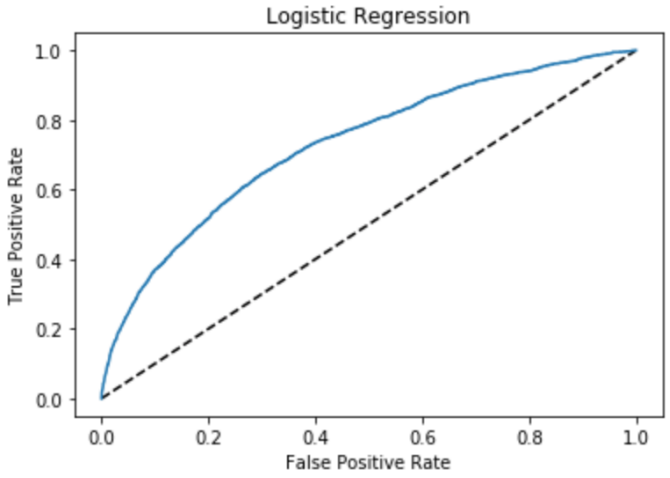
AUC score: 0.739

Precision-Recall Curve:

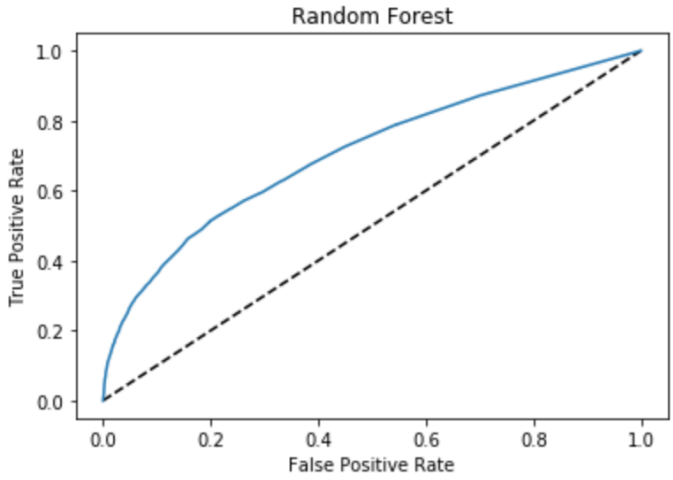


AUC score: 0.11

Cannabis Misdemeanor Sales:
ROC curves:

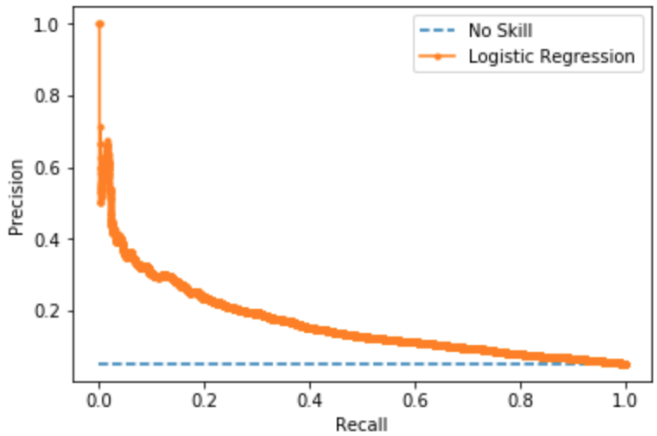


AUC score: 0.729



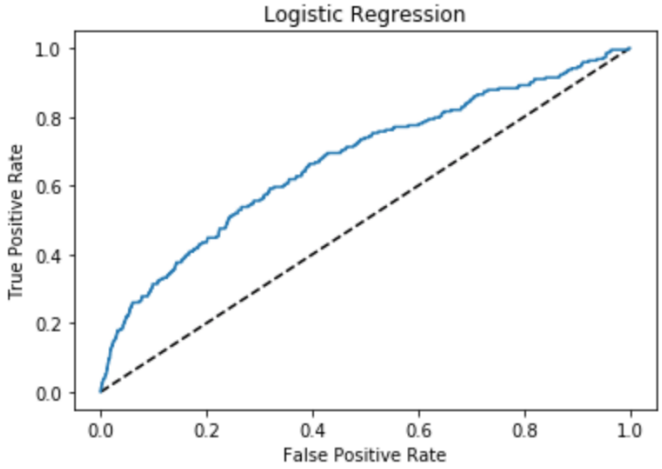
AUC score: 0.706

Precision-Recall Curve:

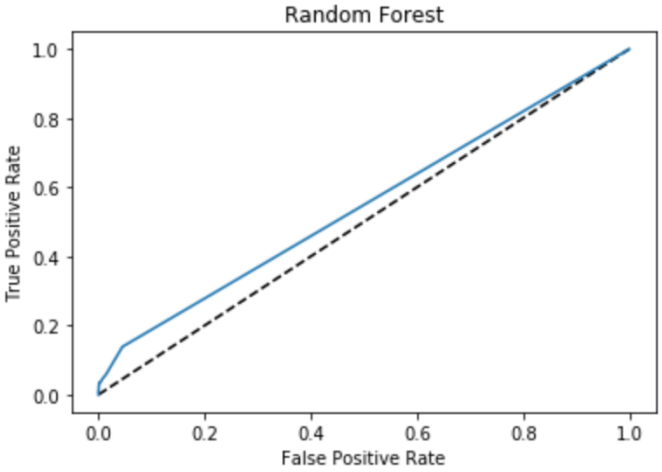


AUC score: 0.116

Cannabis Felony Sales:
ROC curves:

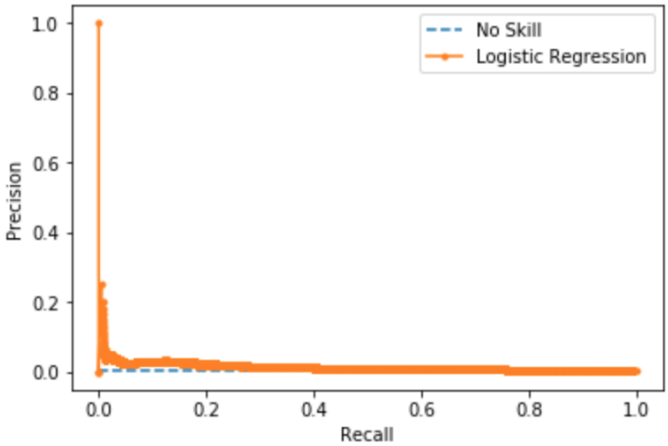


AUC score: 0.673



AUC score: 0.547

Precision-Recall Curve:



AUC score: 0.015