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

By Daniel Loew

The Problem: Racial Disparity in NYC Cannabis Arrests

It has been widely reported that arrests for cannabis in New York City over at least the last decade have been heavily biased towards African-American and Hispanic men, which is part of a wider problem of how police departments interact with communities of color.

Given the available data, how can a deeper understanding be achieved of how cannabis law was enforced?

Beyond the racial disparity, what other statistical relationships and biases can be discovered between features of New York City and cannabis arrests?

The Solution: Machine Learning Classification

To better understand the problem, modern techniques of machine learning classification can acquire knowledge of the features that differentiate cannabis crime from all other crime, and predict if a crime is a cannabis crime or not. These features are a set of demographic, geographic, and temporal variables that include the suspect’s race.

The features’ coefficients arising from Logistic Regression models can identify which features have the strongest positive relationships with cannabis crime which do not exist with other types of crime. This approach can also be repeated to uncover the relationships that differentiate the types of cannabis crime from each other.

Classification as a Path to Understanding

By uncovering these statistical relationships, further research into cannabis crime by criminologists, human rights groups, legal scholars, and public policy researchers can be supported.

The data cleaning protocols and classification models developed in this project can also be modified to investigate the predictive features of any type of crime that occurred in New York City between 2006 and 2018.

The Research Process (Part I)

- The NYPD’s self-reported crime data, available through NYC Open Data, were cleaned to create datasets ready for analysis. This cleaning used the original 34 features to create a predictive feature set of over 1,400 features.
- The data was explored through visualizations and descriptive statistics.
- A series of hypotheses on the likelihoods of cannabis arrests among racial/ethnic groups and geographic areas was tested.
- Several skillful machine learning models were created, including Logistic Regression classification algorithms which identified the coefficients for all features in the feature set.

The Research Process (Part II)

These models identified features of the data set that were most predictive of, and therefore had the strongest statistical relationship, to:

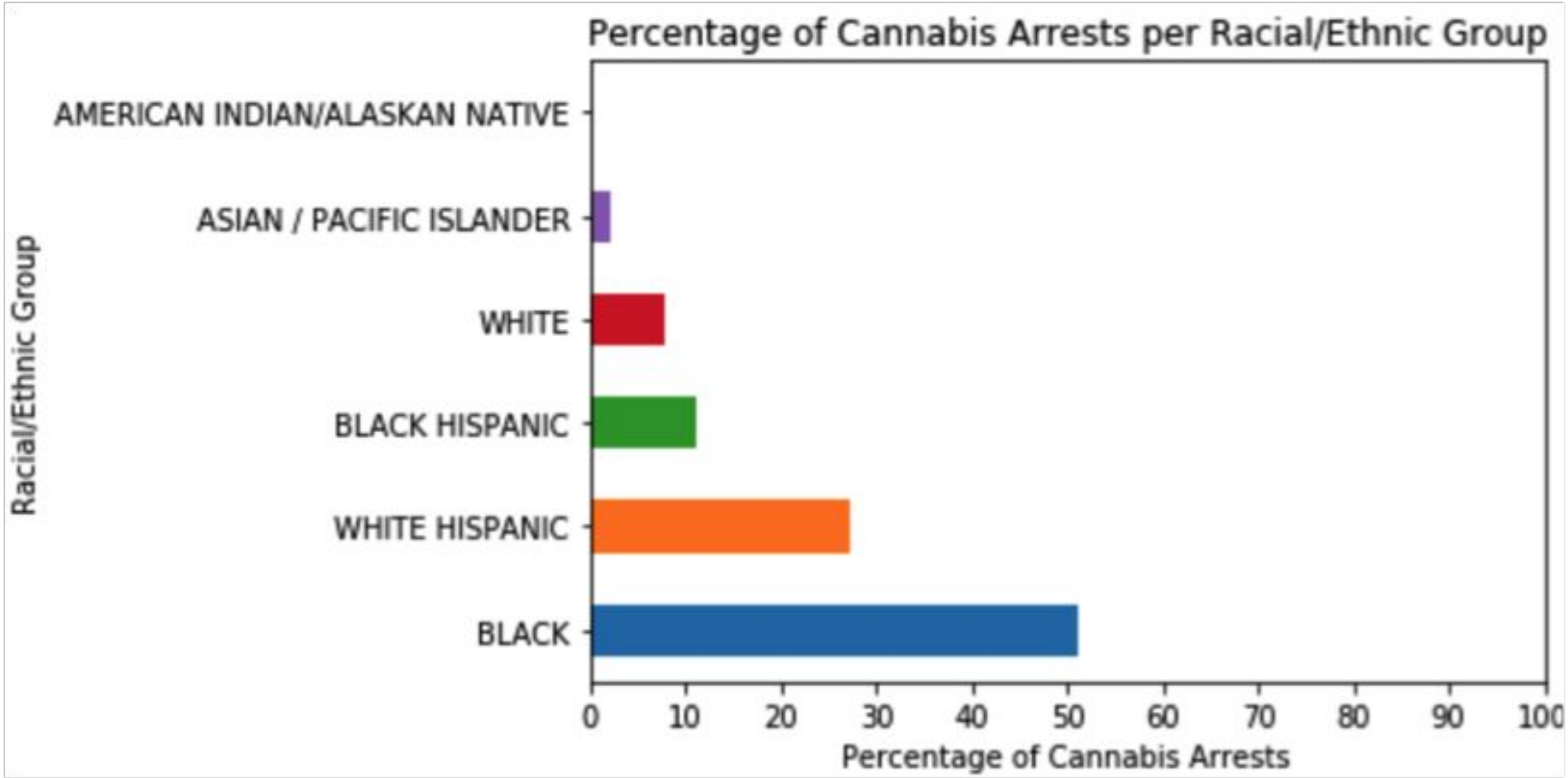
- Cannabis Crime Overall,
- Possession,
- Sales,
- Misdemeanor Possession,
- Violation Possession,
- Felony Possession,
- Misdemeanor Sales, and
- Felony Sales

Does the NYPD’s dataset corroborate the widely reported racial disparity in cannabis arrests?

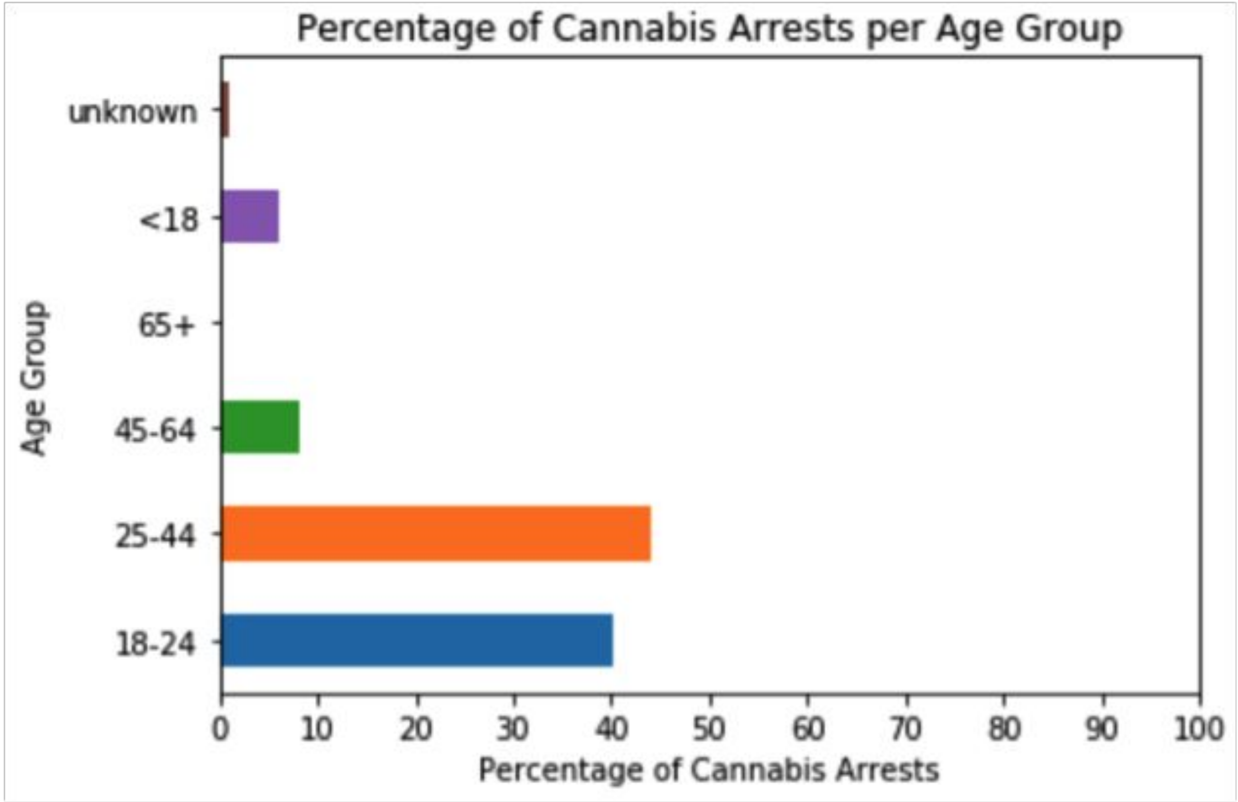
The answer to this question is a qualified “yes”, as the NYPD only reported the suspect’s race/ethnicity, age, and sex on **16% of the 220,304 cannabis arrests made in New York City between 2006 and 2018, while reporting these statistics for 38% of all other types of NYC crime during this time period.**

It was verified by the NYC Open Data Project that these demographic data are not required to be collected by the NYPD, and were subject to the arresting officer’s interpretation (email verifying this available upon request).

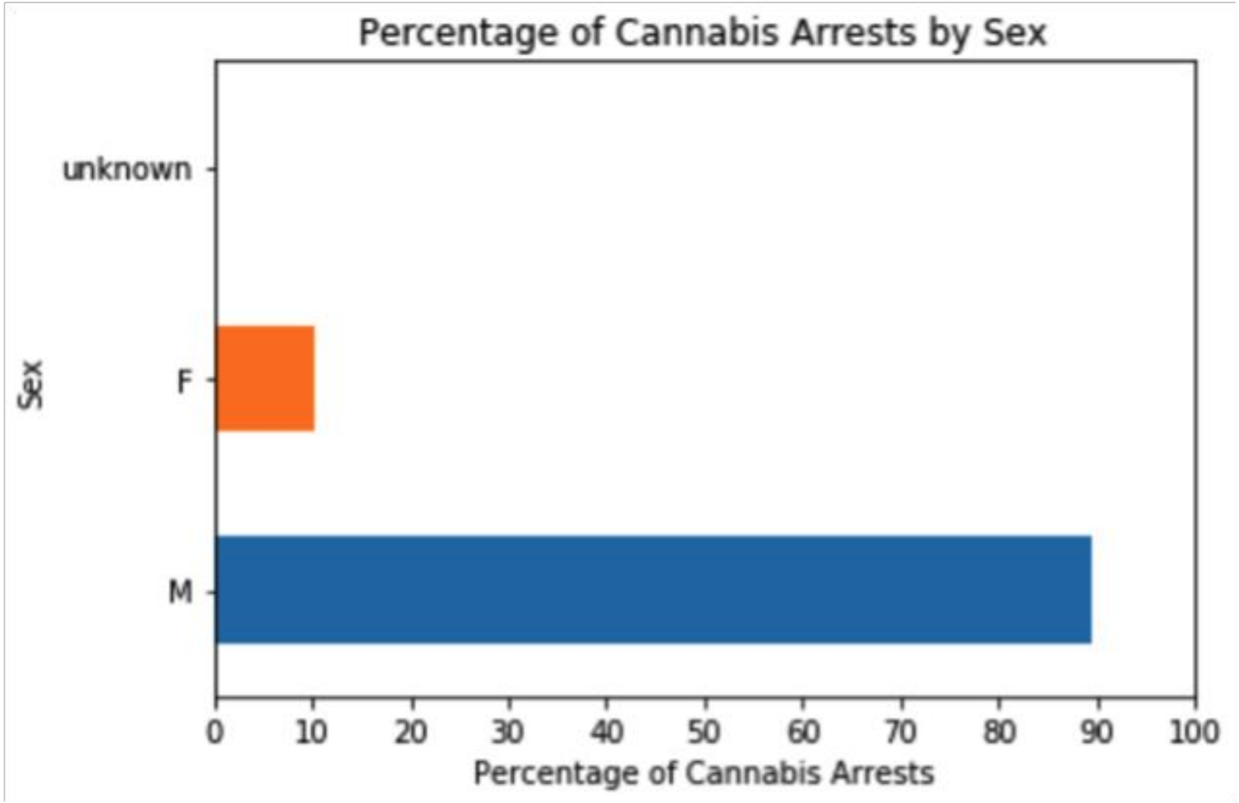
African-American and Hispanic suspects represent the majority of this 16% of cannabis arrests.



Suspects under the age of 45 represent the majority of this 16% of cannabis arrests



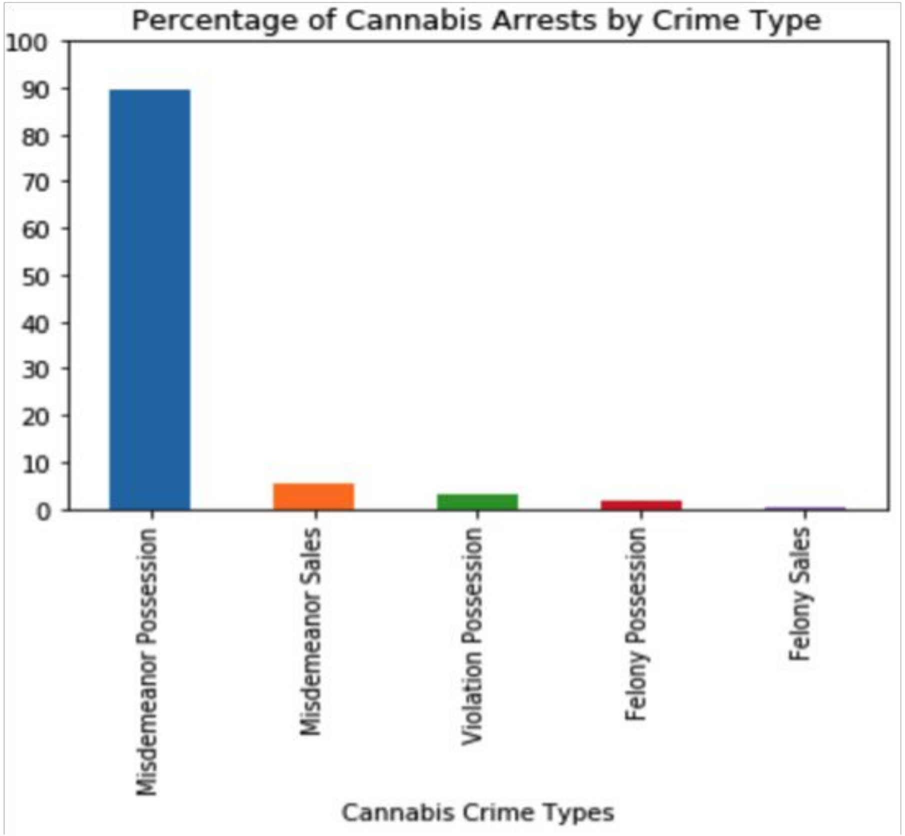
The majority of this 16% of cannabis arrests were male.



Out of these 16% of cannabis arrests...

- 40% of cannabis arrests were of African-American males younger than 45.
- 32% were of Hispanic/Latino males younger than 45.
- Therefore, 72% were of young African-American and Hispanic/Latino males.

The majority of all arrests were for misdemeanor possession.



Felony sales arrests were a vanishingly small percentage of overall cannabis arrests.

Misdemeanor sales, felony possession, and violation possession charges (the lowest level crime), were also a very small percentage of arrests.

This shows that users were targeted more than sellers, and were not given the lowest level charge.

Geography as an indicator of race/ethnicity

It is unfortunate that the NYPD recorded the suspect’s demographics so rarely, but the latitude and longitude of each arrest was recorded for the majority of cannabis crimes.

Although the demographics of different areas of New York City do vary, especially in the age of gentrification, and the population of New York City is very mobile, one can partially infer race/ethnicity of cannabis crime suspects based on the areas the arrests occurred in.

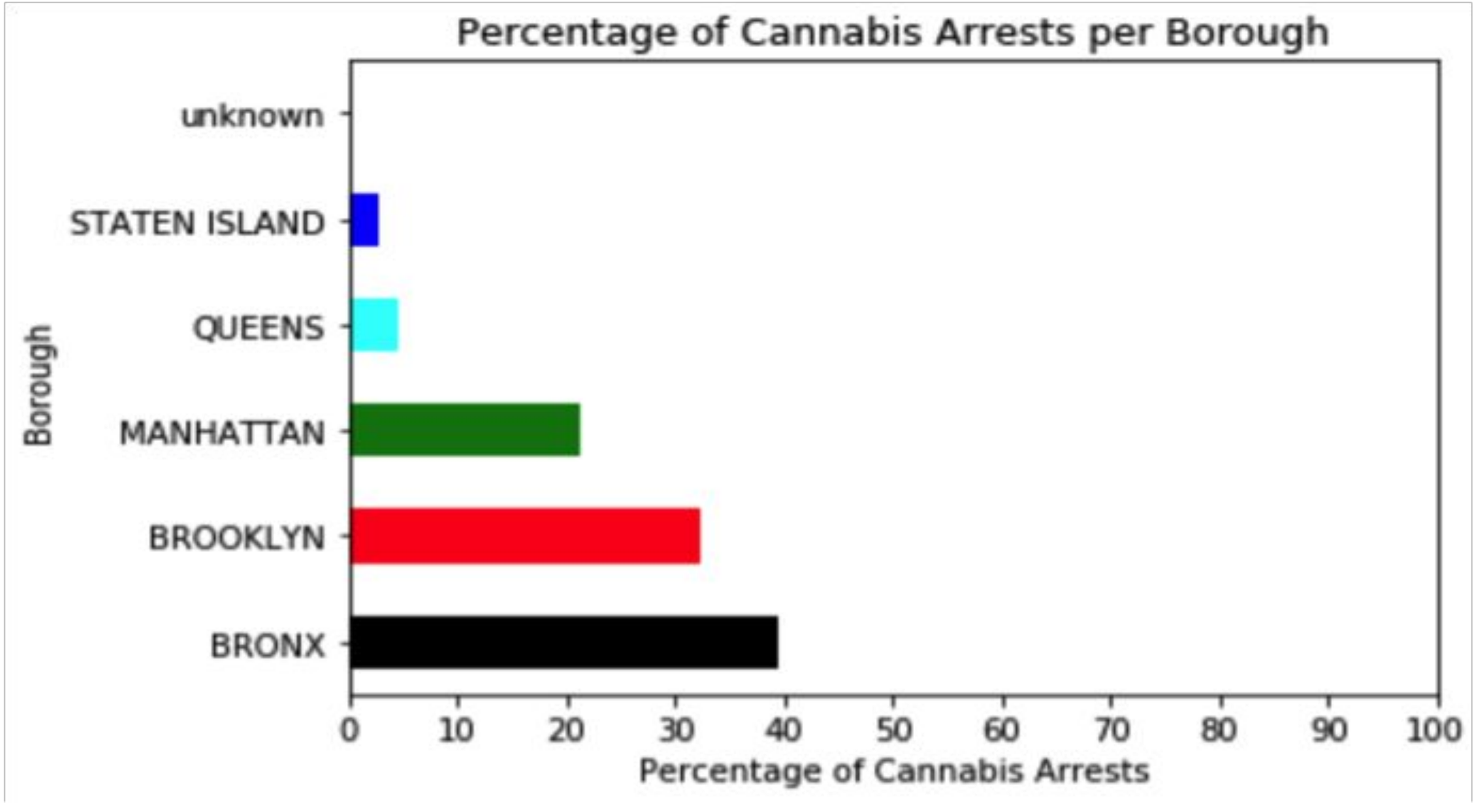
Estimated Percentages of Borough Population by Race, 2019*

	African-American	Latino/Hispanic	Asian	Non-Hispanic White
Bronx	44%	56%	5%	9%
Brooklyn	34%	19%	13%	36%
Manhattan	18%	26%	13%	47%
Queens	21%	28%	27%	25%
Staten Island	12%	19%	10%	60%

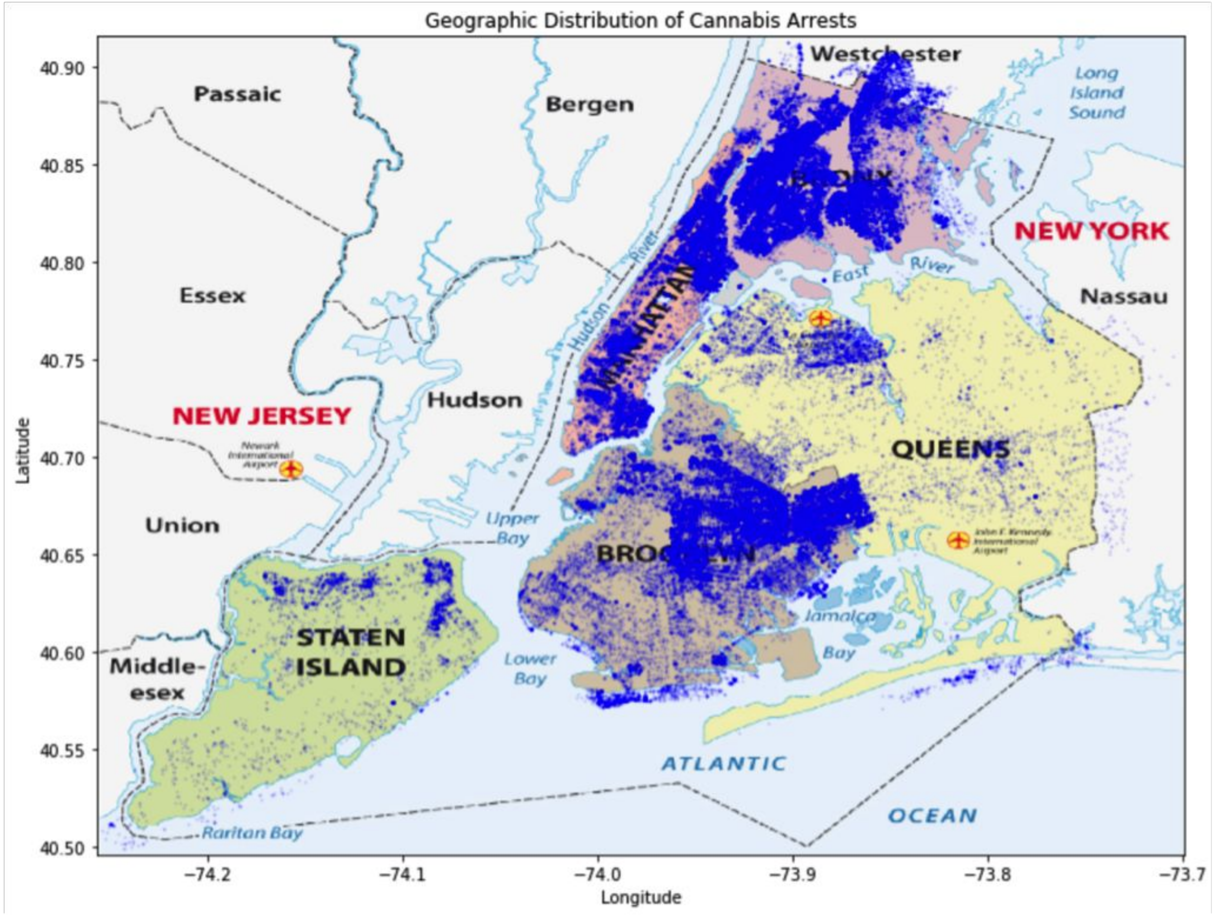
* Census Estimates

NOTE: Latino/Hispanic is an ethnicity and can apply to people of multiple races; this is why the row total sums to more than 100%.

Percentages of All Cannabis Arrests by Borough



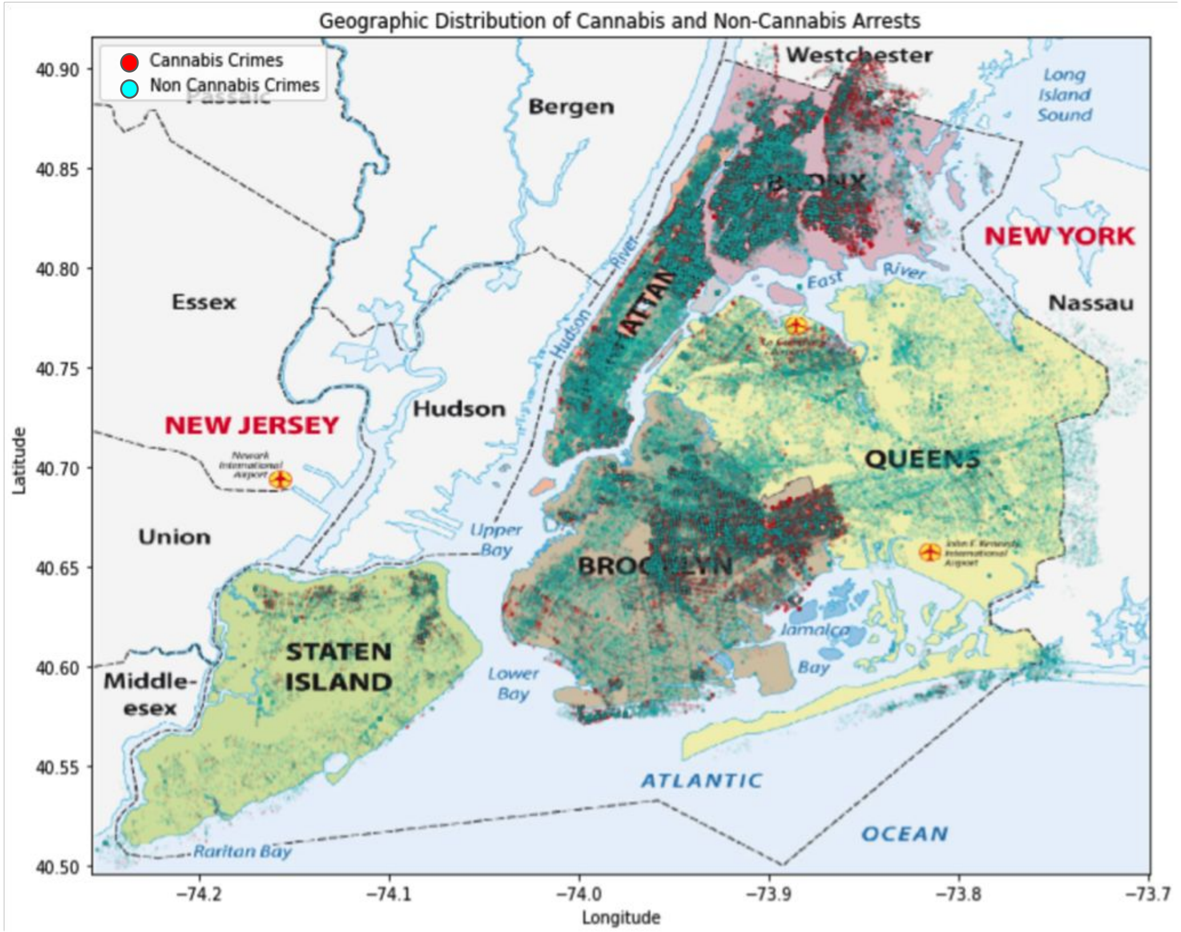
Location of All Cannabis Crimes (2006-2018)



NOTE: Each pixel represents a cannabis arrest.

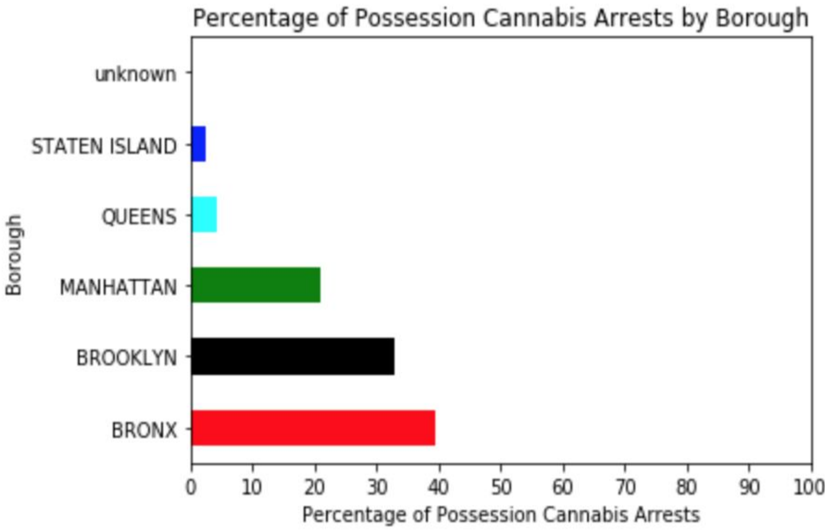
The overlay of the latitude/longitude coordinates of each arrest are slightly warped in relation to the map image of NYC's five boroughs due to the curvature of the earth.

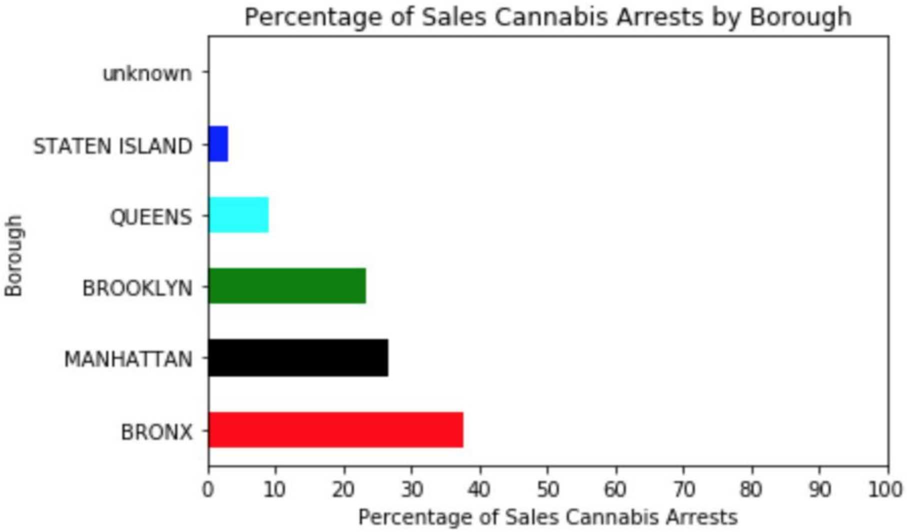
Location of Cannabis and Non-Cannabis Crimes (2006-2018)



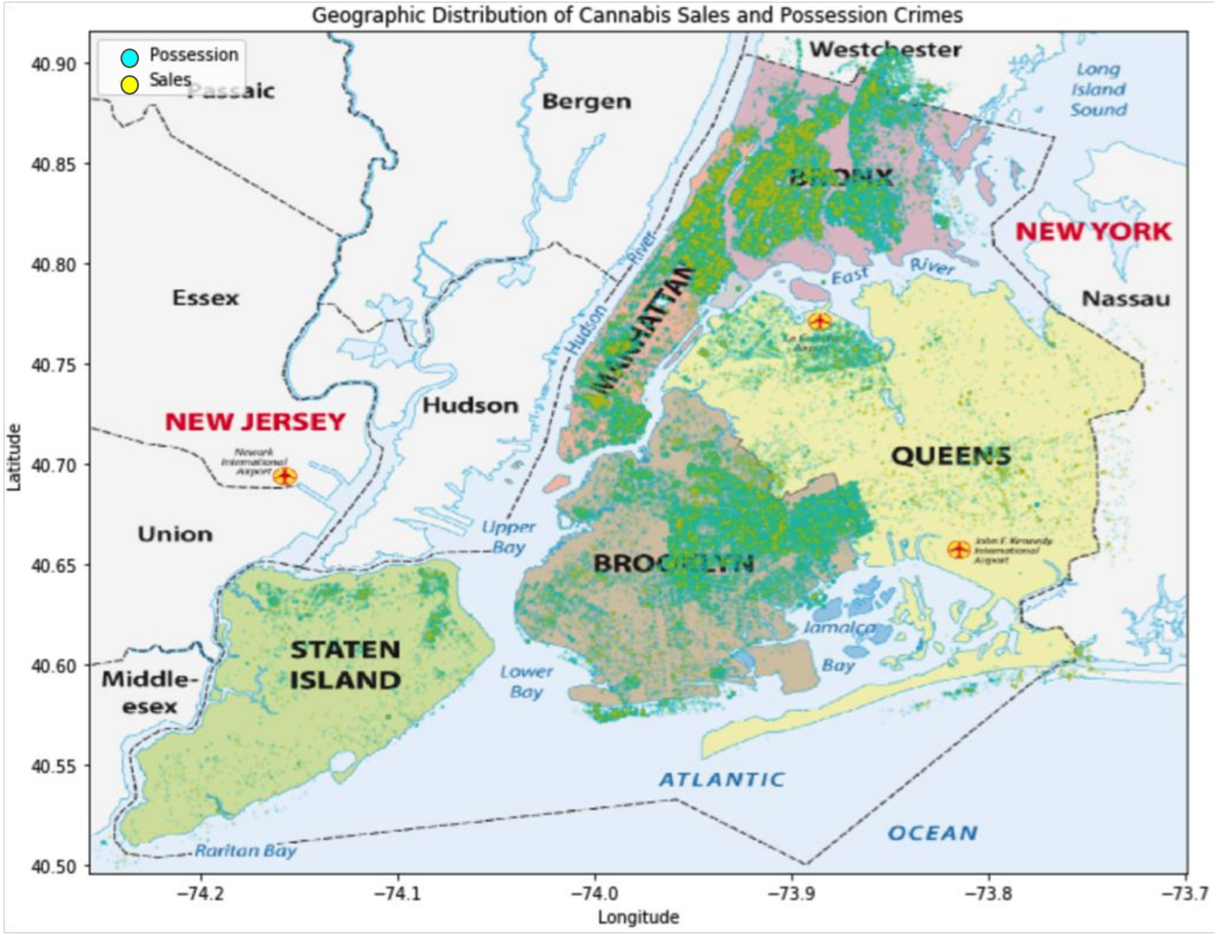
NOTE: The location of every NYC crime between 2006 and 2018 is shown, each pixel representing either a cannabis crime (in red) or a different type of crime (in cyan)..

Percentages of All Possession and Sales Arrests by Borough

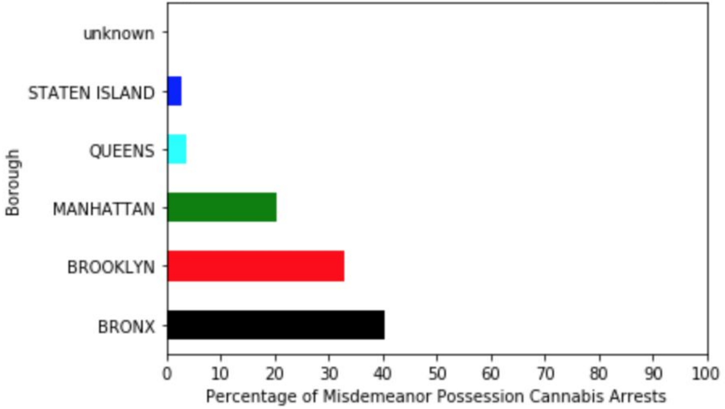


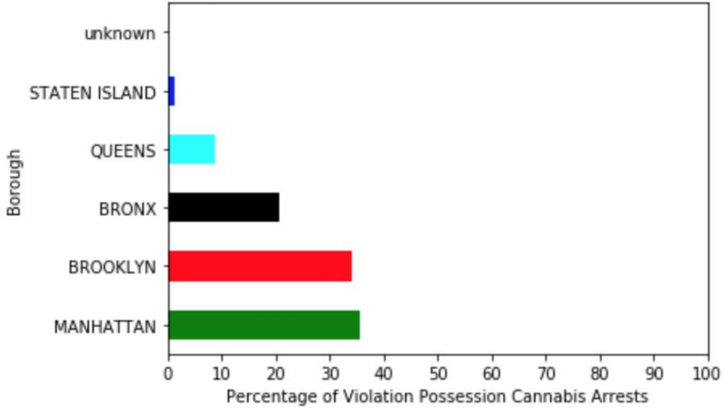


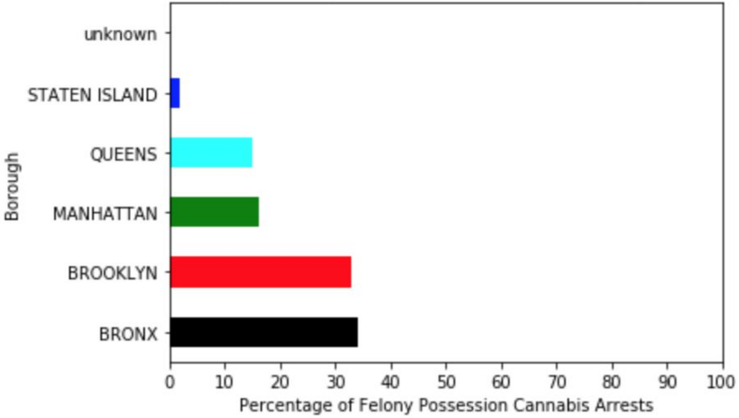
Location of All Sales and Possession Crimes (2006-2018)



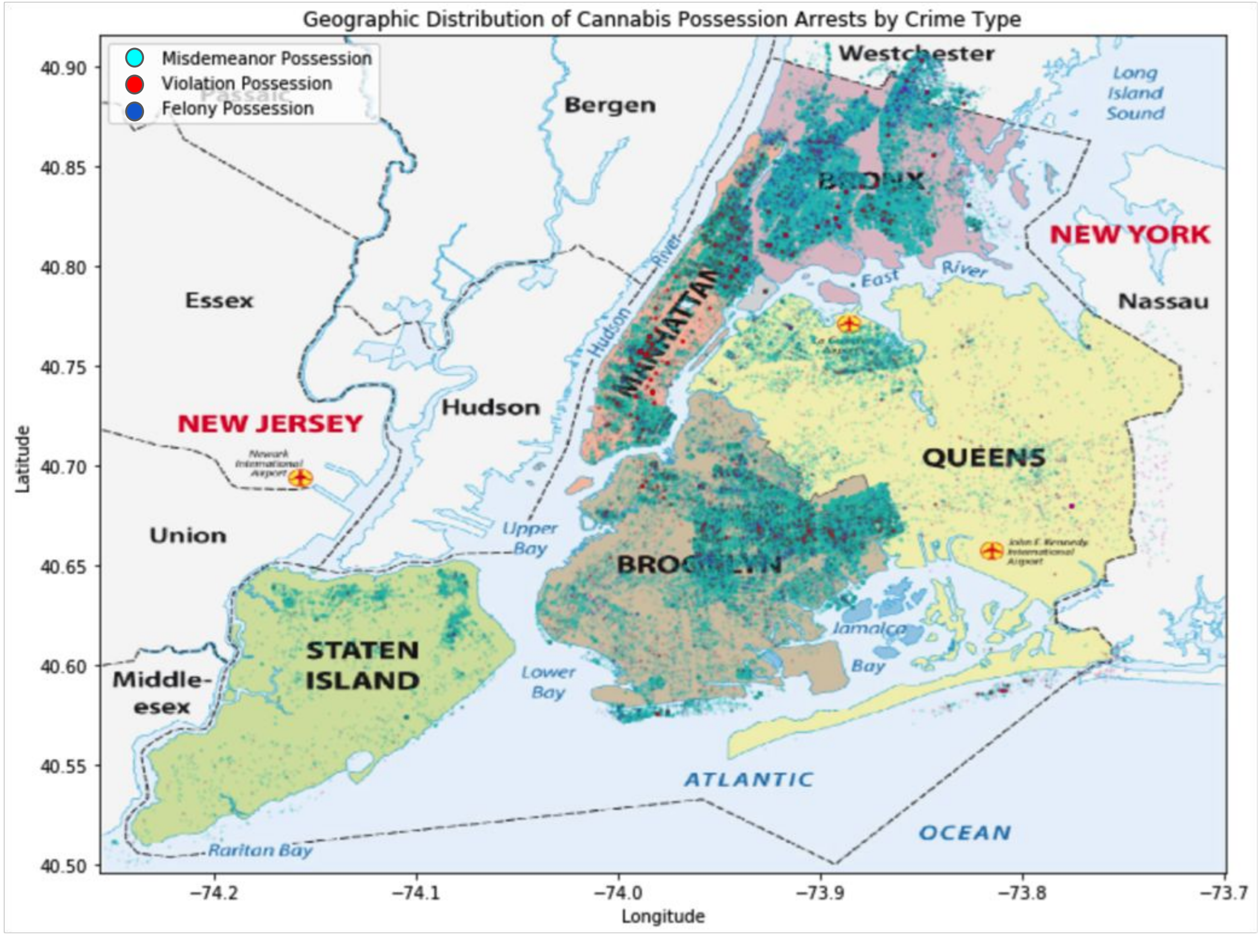
Percentages of Possession Arrests by Borough


Percentage of Misdemeanor Possession Cannabis Arrests per Borough

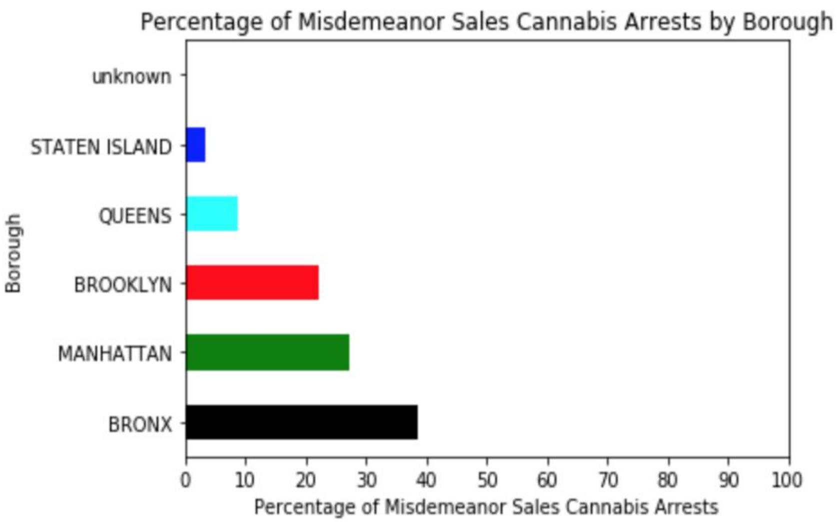

Percentage of Violation Possession Cannabis Arrests per Borough

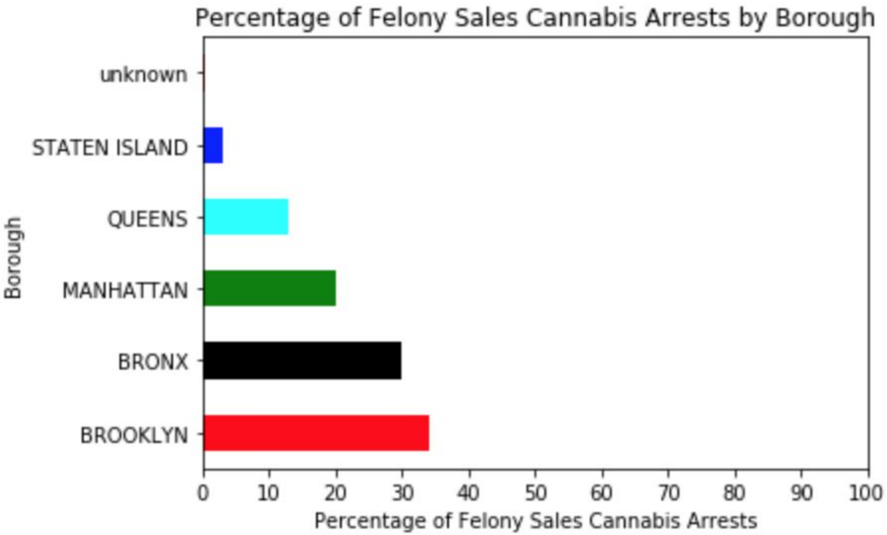

Percentage of Felony Possession Cannabis Arrests per Borough

Location of All Possession Crimes (2006-2018)

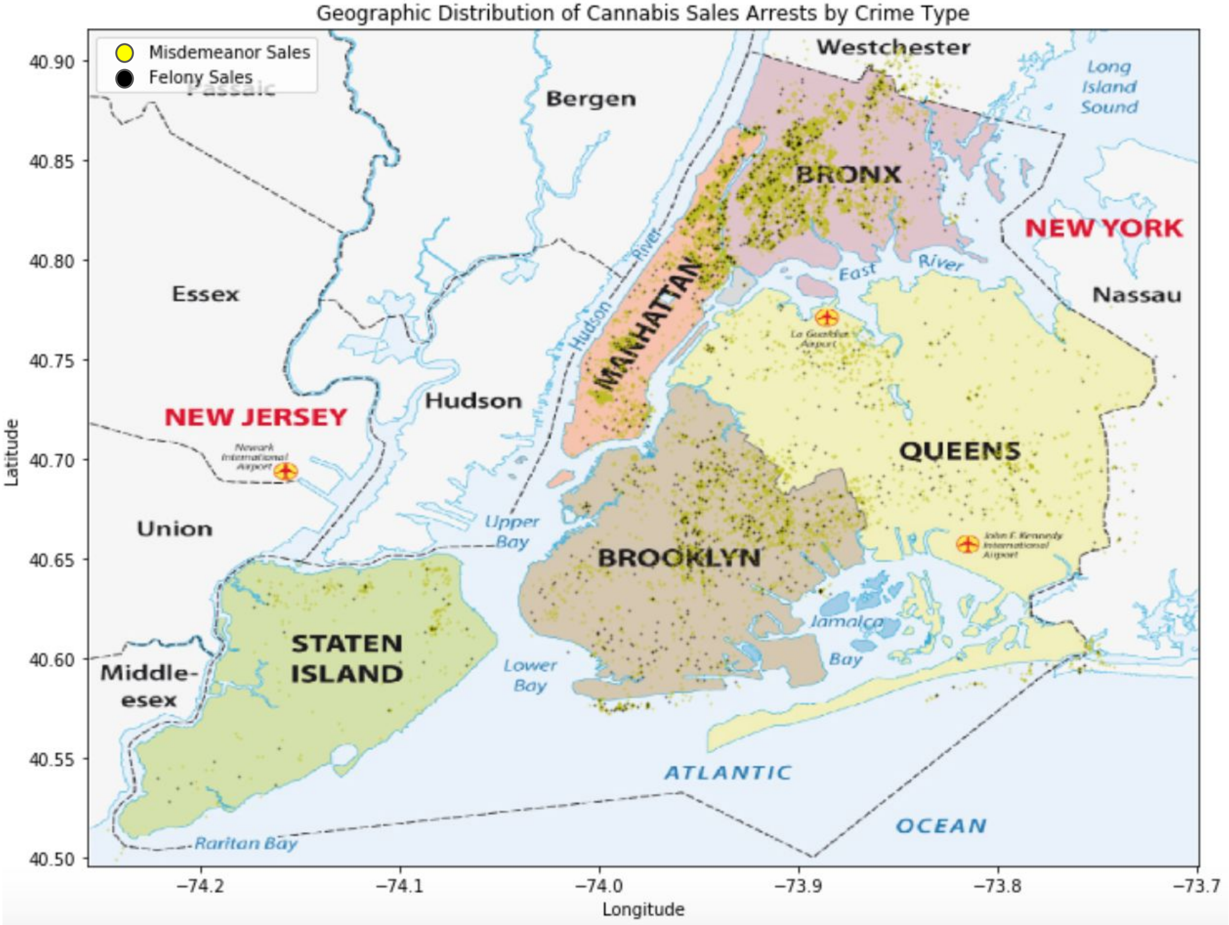


Percentages of Sales Arrests by Borough





Location of All Sales Crimes (2006-2018)



Premises of Cannabis Arrests

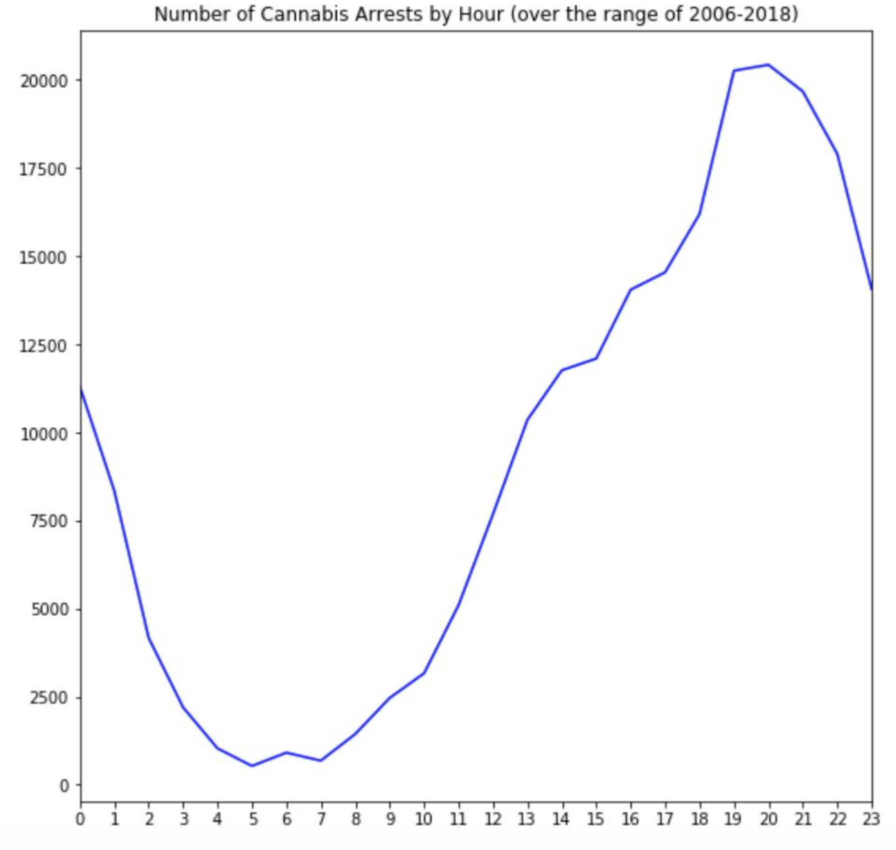
Cannabis arrests generally occurred primarily in the following locations:

- the street (58%),
- public housing projects (19%),
- residential apartment houses (8%),
- parks and playgrounds (6%)

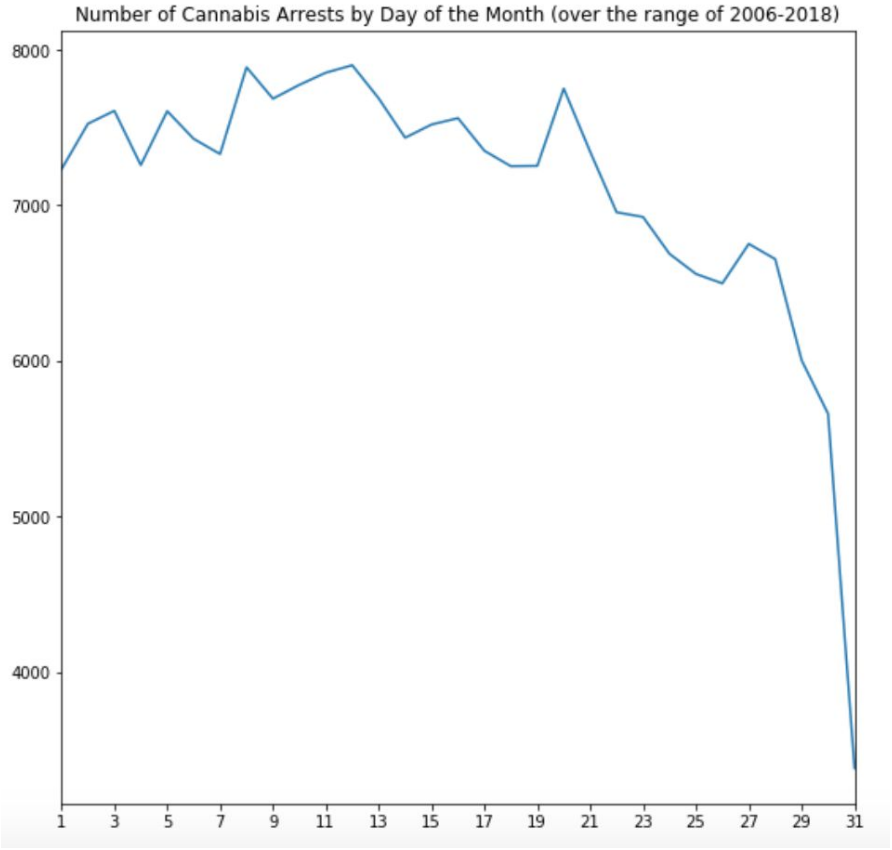
This pattern holds true in all types of cannabis arrests except for violation possession arrests, of which 32% occurred in the subway system, and for misdemeanor possession arrests, which occurred in public housing projects much more than felony possession arrests (20% and 7%, respectively).

The top 10 NYCHA housing developments with the highest proportion of cannabis arrests were all in economically disadvantaged areas of the Bronx and Brooklyn.

Cannabis Arrests by Hour and Day of the Month (2006-2018)

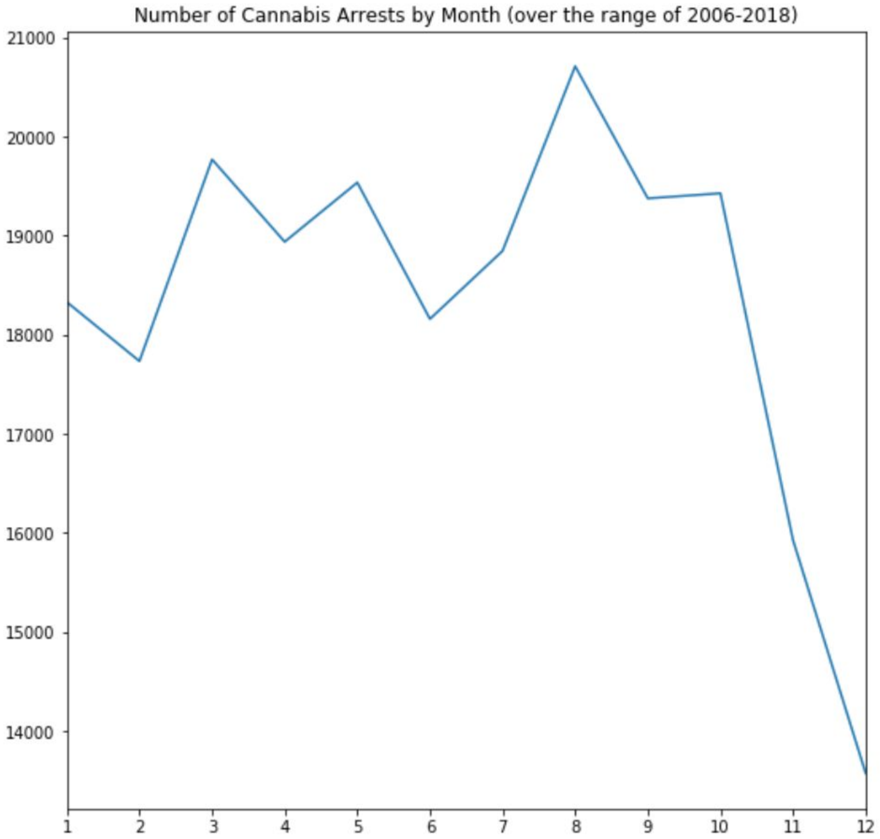


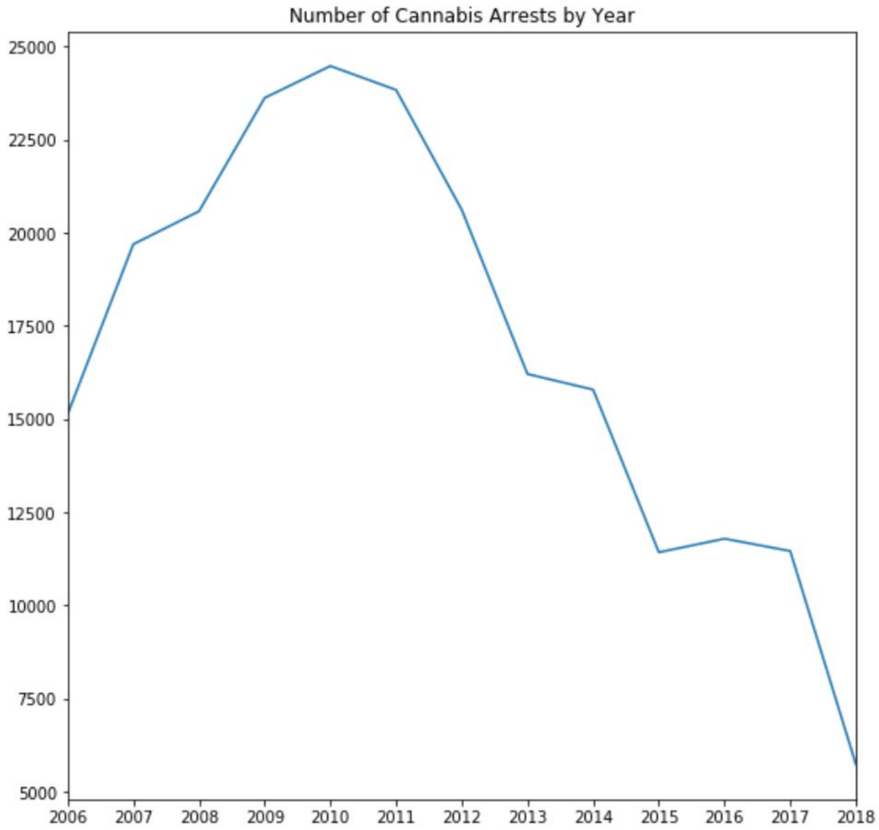
Number of Cannabis Arrests by Hour (over the range of 2006-2018)



Number of Cannabis Arrests by Day of the Month (over the range of 2006-2018)

Cannabis Arrests by Month and Year (2006-2018)





Holidays Cannabis Arrests Occurred On

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. Over the full span of 2006 to 2018, the holidays with the greatest number of cannabis arrests were:

- April 20th - 737 (non-official holiday)
- Yom Kippur - 707
- Rosh Hashanah - 677
- Eid al-Fitr - 664
- Diwali - 656
- Eid al-Adha - 544
- St. Patrick’s Day - 542
- Valentine’s Day - 531

Hypothesis Testing: Unreported Demographic Info

A t-test was run to determine that 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 therefore some mediating factor behind this difference. The t-score was approximately 67.6 and the p-value was 0.0.

Hypothesis Testing: Likelihood of Cannabis Arrest Based on Race/Ethnicity and Geography*

A series of t-tests generally support that there is a hierarchy in terms of how likely different racial/ethnic groups were to be arrested for cannabis in NYC between 2006 and 2018, and that these differences in likelihood were not due to chance. 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 likely cannabis arrests were made in the five boroughs, and these differences were not due to chance. From most likely to least likely, this hierarchy is the Bronx, Brooklyn, Manhattan, Queens, and Staten Island.

*Specific t-scores and p-values are available in the Overall Report and the ‘DataStory_HT_Final’ Jupyter notebook.

Logistic Regression for Predicting Class of Crime

A series of Logistic Regression algorithms with varying hyperparameters (non-normalized versus normalized data, C values, and regularization type) were trained and tested on the cleaned datasets in order to find the most predictive algorithm for the following models:

- Predicting whether a crime was a cannabis crime or not
- Predicting whether a cannabis crime was a:
    - Possession or Sales Crime
    - Misdemeanor Possession Crime or another type of cannabis crime
    - Violation Possession Crime or another type of cannabis crime
    - Felony Possession Crime or another type of cannabis crime
    - Misdemeanor Sales Crime or another type of cannabis crime
    - Felony Sales Crime or another type of cannabis crime

Evaluating the Logistic Regression Algorithms

The best performing Logistic Regression algorithms for each of the seven models was selected by evaluating them on the following metrics:

- Accuracy - the proportion of correctly predicted data points out of all the data points
- Precision - a measure of result relevancy, the number of true positives divided by the sum of true positives and false positives
- Recall - a second measure of result relevancy, the number of true positives divided by the sum of true positives and false negatives
- F1 Score - the harmonic mean of precision and recall, often seen as the most salient performance metric for machine learning models
- Area Under the Curve (AUC) for ROC and Precision-Recall curves - measures of a model’s diagnostic ability for all possible discrimination thresholds between binary classes

Performance Metrics for Best Performing LR Algorithms

	Accuracy	Precision	Recall	F1	ROC AUC	PRC AUC
Cannabis/Non-Cannabis	0.843	0.84	0.84	0.84	0.913	N/A
Possession/Sales	0.682	0.92	0.68	0.77	0.729	0.974
Misdemeanor Possession	0.716	0.86	0.72	0.77	0.719	0.948
Violation Possession	0.813	0.96	0.81	0.87	0.813	0.338
Felony Possession	0.736	0.98	0.74	0.83	0.784	0.111
Misdemeanor Sales	0.688	0.93	0.69	0.77	0.729	0.116
Felony Sales	0.708	0.99	0.71	0.82	0.673	0.015

LR Coefficients Reveal Salient Predictors of Cannabis Crime

Although causality for cannabis arrests cannot be attributed to them, the best Logistic Regression models' coefficients can show us which features have a strong statistical relationship with cannabis crime. 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, which highlights that feature's relationship to cannabis crime itself. For continuous features, for each unit increase in the continuous feature there is an increase in the odds that a crime is a cannabis crime.

Individual Features Differentiating Cannabis Crime, Pt. 1

If a crime in New York City between 2006 and 2018 has these features, it is X% more likely to be a cannabis crime.

Feature	**Coefficient**	**% More Likely Cannabis Crime**
Occurred in a park or playground	2.75	15.7%
Unrecorded/unknown suspect sex	2.50	12.2%
Complaint and recorded date matches	2.42	11.3%
Occurred in a public building (unspecified)	2.18	8.8%
Hour of complaint (continuous data type)	2.06	7.9%
Occurred in a public housing project	1.92	6.8%
Occurred on the street	1.64	5.2%

Individual Features Differentiating Cannabis Crime, Pt. 2

Feature	**Coefficient**	**% More Likely Cannabis Crime**
Occurred in open areas or lots	1.64	5.2%
Unreported premises	1.58	4.9%
Other premises	1.44	4.2%
Suspect 18-24 years old	1.43	4.2%
Occurred in a marina or pier	1.35	3.8%
Occurred in a parking lot or public garage	1.33	3.8%
Police Precinct 71 - Crown Heights, Wingate, Prospect Lefferts neighborhoods of Central Brooklyn	1.21	3.4%
Williamsburg Housing Project	1.21	3.3%

Individual Features Differentiating Cannabis Crime, Pt. 3

Feature	Coefficient	% More Likely Cannabis Crime
Borinquen Plaza II Housing Project	1.19	3.3%
L1 distance to Williamsburg Bridge	1.17	3.2%
Occurred in a public school	1.12	3.1%
Police Precinct 75 - East New York and Cypress Hills in Easternmost Brooklyn	1.09	3%
Marcy Housing Project	1.08	2.9%
Occurred in a residential apartment house	1.06	2.9%
Police Precinct 67 - East Flatbush and Remsen Village in Central Brooklyn	1.03	2.8%

Full list of coefficients and likelihoods available in Jupyter notebooks

LR Coefficients of Feature “Families”

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, etc.). 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.

These feature families are 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.

Most Prominent Feature Families for Cannabis Crime

The feature families with the largest percentage of the total predictive value of the feature set for cannabis crimes generally are, in descending order:

- Public housing developments, or “projects”: 23%
- Premises Types: 17%
- Transit/Subway Stations: 16%
- Police Precincts: 13%
- NYC Parks: 10%
- Holidays: 3%
- L1 distances from NYC landmarks: 3%
- Transit Districts: 2%
- Suspect Age: 2%

Suspect Race’s Role in Cannabis Arrests

The race of someone arrested for a crime in NYC between 2006 and 2018 is actually largely negatively correlated with that crime being for cannabis. However, it has to be reinforced that only 16% of cannabis arrests had their suspect’s race recorded.

Suspect Race/Ethnicity	**Coefficient**
African-American	-0.31
White Hispanic	-0.13
White Non-Hispanic	-0.12
African-American Hispanic	-0.10
Native American	-0.03
Asian or Pacific Islander	0.06
Unreported Suspect Race	**0.60**

How could this be?

How could the majority of cannabis crime suspects in New York City between 2006 and 2018 be African-American and/or Hispanic while their race/ethnicity has a negative coefficient when it comes to cannabis crime?

These weak negative coefficients actually show that for those who have been arrested for a crime, race/ethnicity generally has a stronger relationship with other types of crime than cannabis crime as these coefficients come from the binary classification of cannabis and non-cannabis crime, and the world of crime is much larger than just cannabis.

However, the fact that unreported suspect race has a coefficient of 0.6 shows that there is a strong relationship with a cannabis crime suspect’s race being unreported which does not exist with other types of crime. The reasons for this lack of reporting should be investigated by future researchers and policy makers.

Proportion of Race’s Relationship to Crime

To further explore this seeming paradox, all of the suspect race features' coefficients were summed in absolute value terms, and then the 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 classified as a cannabis crime or non-cannabis crime in New York City between 2006 and 2018.

The African-American proportion was the highest 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. This does show a racial bias in how people are arrested for crime generally.

Features Differentiating Cannabis Possession, Pt. 1

If a cannabis crime in New York City between 2006 and 2018 has these features, it is X% more likely to be a possession crime.

Feature	Coefficient	% More Likely Possession
Occurred in an airport terminal	3.94	51.2%
L1 distance to Empire State Building	1.79	6%
Tri-Boro Bridge & Tunnel Jurisdiction	1.71	5.5%
L2 distance to closest subway entrance	1.69	5.5%
Occurred at a marina or pier	1.47	4.4%
Transit District 12 - Bronx around 180th St.	1.40	4.1%
McKinley Housing Project	1.32	3.7%

Individual Features Differentiating Cannabis Possession, Pt. 2

Feature	Coefficient	% More Likely Possession
303 Vernon Avenue Housing Project	1.27	3.6%
Morrisania Air Rights 42 Housing Project	1.27	3.6%
Police Precinct 100 - Rockaway Park and surrounding neighborhoods, south Queens	1.26	3.5%
Police Precinct 75 - East New York and Cypress Hills in easternmost Brooklyn	1.23	3.4%
Edenwald Housing Project	1.21	3.4%
Williams Plaza Housing Project	1.17	3.2%
Transit District 11 - Bronx Yankee Stadium	1.10	3%
Police Precinct 88 - Clinton Hill, Fort Greene Park, and Commodore Barry Park in Northern Brooklyn	1.09	3%

Individual Features Differentiating Cannabis Possession, Pt. 3

Feature	Coefficient	% More Likely Possession
Weeksville Gardens Housing Project	1.08	2.9%
Tompkins Housing Project	0.99	2.7%
LaFayette Housing Project	0.98	2.7%
Police Precinct 77 - Crown Heights and Prospect Heights in Central Brooklyn	0.93	2.5%
N.Y. Transit Police Jurisdiction	0.92	2.5%
Police Precinct 43 - Southeast Bronx	0.89	2.4%
Hughes Apartments Housing Project	0.88	2.4%
Claremont Rehab (Group 5) Housing Project	0.85	2.3%

Features Differentiating Cannabis Sales, Pt. 1

If a cannabis crime in New York City between 2006 and 2018 has these features, it is X% more likely to be a sales crime.

Feature	Coefficient*	% More Likely Sales
Occurred in Washington Square Park	-3.41	30.3%
Jackson Ave. Subway Station, Bronx	-3.26	26.1%
Lexington Ave. Subway Station, Manhattan	-2.62	13.8%
Pennsylvania Ave. Subway Station, East New York, Brooklyn	-2.07	7.9%
Unreported Age	-1.97	7.1%
Occurred in or near a liquor store	-1.95	7.1%
L2 distance to World Trade Center	-1.81	6.1%

*Coefficients are negative because sales was the 0 class in the possession (1) vs. sales classification.

Individual Features Differentiating Cannabis Sales, Pt. 2

Feature	Coefficient	% More Likely Sales
Parkside Playground, Brooklyn	-1.81	6.1%
Occurred in a grocery store or bodega	-1.80	6.1%
Pomonok Housing Project	-1.78	5.9%
St. James Park, Bronx	-1.75	5.8%
Wald Housing Project	-1.74	5.7%
Police Precinct 6 - West Village, Manhattan	-1.64	5.2%
L1 distance to Yankee Stadium, Bronx	-1.64	5.2%
Kingston Ave. Subway Station, Brooklyn	-1.58	4.8%
Occurred in a candy store	-1.57	4.8%

Individual Features Differentiating Cannabis Sales, Pt. 3

Feature	**Coefficient**	**% More Likely Sales**
Queensbridge North Housing Project	-1.53	4.6%
Occurred in a fast food restaurant	-1.46	4.3%
Castle Hill housing project	-1.41	4.1%
Park of the Americas, Corona, Queens	-1.35	3.9%
Noonan Playground, Woodside, Queens	-1.33	3.8%
Junius St. Subway Station, Brownsville, BK	-1.32	3.8%
Police Precinct 112 - Forest Hills, Queens	-1.31	3.7%
Sedgwick housing project	-1.3	3.7%
Occurred in an unclassified store	-1.3	3.7%

Most Prominent Feature Families for Possession and Sales

The feature families with the largest percentage of the total predictive value of the feature set for classifying possession and sales crimes are, in descending order:

- Public housing developments, or “projects”: 32%
- Police Precincts: 16%
- Premises Types: 13%
- NYC Parks: 11%
- Transit/Subway Stations: 10%
- Transit Districts: 3%
- Jurisdiction: 2%
- L1 distances from NYC landmarks: 2%

Suspect Race’s Role in Sales and Possession Arrests

There is a weak relationship with sales arrests for African-Americans and Hispanics, and a strong relationship with possession arrests for non-Hispanic Whites and cases whose race was not reported.

Suspect Race/Ethnicity	**Coefficient**
African-American Hispanic	-0.18
African-American	-0.17
White Hispanic	-0.03
Asian or Pacific Islander	0.16
White Non-Hispanic	0.58
Unreported Suspect Race	0.75

Features Differentiating Misdemeanor Possession*

If a cannabis crime in New York City between 2006 and 2018 has these features, it is X% more likely to be a misdemeanor possession crime.

Feature	Coefficient	% More Likely Misdemeanor Possession
Occurred at a marina or pier	1.91	6.8%
Occurred at a park or playground	1.31	3.7%
Police Precinct 75 - East New York and Cypress Hills in Easternmost Brooklyn	1.23	3.4%
303 Vernon Ave. Housing Project	1.17	3.2%
Occurred in an open area or lot	1.13	3.1%
L1 distance to Prospect Park	1.1	3%
Occurred in a public housing residence	1.1	2.9%

*A much more extensive list is available in the Overall Report and Jupyter notebooks

Most Prominent Feature Families for Misdemeanor Possession

The feature families with the largest percentage of the total predictive value of the feature set for classifying misdemeanor possession crimes are, in descending order:

- Public housing developments, or “projects”: 30%
- NYC parks: 18%
- Transit/Subway Stations: 15%
- Police precincts: 10%
- Premises types: 10%
- L1 distances from NYC landmarks: 3%
- Transit districts: 3%
- L2 distances from NYC landmarks: 3%

Suspect Race’s Role in Misdemeanor Possession

Among those arrested for cannabis, white arrestees have a weakly positive relationship to misdemeanor possession and those with unreported suspect race have a moderately strong relationship, while cannabis arrestees of other racial groups have no relationship or weakly negative relationships to misdemeanor possession.

Suspect Race/Ethnicity	**Coefficient**
Asian or Pacific Islander	-0.34
African-American	-0.18
African-American Hispanic	-0.14
White Hispanic	-0.04
Native American	0.03
White Non-Hispanic	0.21
Unreported Suspect Race	0.43

Features Differentiating Violation Possession*

If a cannabis crime in New York City between 2006 and 2018 has these features, it is X% more likely to be a violation possession crime.

Feature	Coefficient	% More Likely Violation Possession
L1 distance to Brooklyn Bridge	6.09	439.5%
Occurred on an NYC city bus	3.45	31.5%
Joyce Kilmer Park in the Bronx	3.11	22.3%
Transit District 12, 180th St., the Bronx	2.99	19.8%
Paerdegat Park in East Flatbush, Brooklyn	2.93	18.8%
Neptune Playground, Coney Island, Bklyn	2.88	17.8%
L1 distance to Riker's Island	2.71	15%

*A much more extensive list is available in the Overall Report and Jupyter notebooks

Most Prominent Feature Families for Violation Possession

The feature families with the largest percentage of the total predictive value of the feature set for classifying violation possession crimes are, in descending order:

- Public housing developments, or “projects”: 35%
- NYC parks: 16%
- Premises types: 15%
- Police precincts: 12%
- L1 distances from NYC landmarks: 7%
- Transit districts: 4%
- Holidays: 3%
- Police Jurisdictions: 3%

Suspect Race’s Role in Violation Possession

Among those arrested for cannabis, African-Americans have a moderately strong relationship with violation possession, and again unreported race has a moderately strong relationship with violation possession.

Suspect Race/Ethnicity	Coefficient
Asian or Pacific Islander	-0.14
White Hispanic	-0.10
African-American Hispanic	-0.09
Native American	0
White Non-Hispanic	0.002
African-American	0.39
Unreported Race	0.46

Features Differentiating Felony Possession*

If a cannabis crime in New York City between 2006 and 2018 has these features, it is X% more likely to be a felony possession crime.

Feature	**Coefficient**	**% More Likely Felony Possession**
Canarsie Park in Canarsie, Brooklyn	7	1,094%
Rockefeller Ctr Subway Station, Manhattan	6.32	557.2%
W. 4th St. Subway Station, Manhattan	5.24	188.1%
Gates Ave. Subway Station, Brooklyn	4.99	146.3%
25th St. Subway Station, Manhattan	4.48	87.9%
241st St./Wakefield Subway Station, Bronx	4.23	68.5%
L1 distance to downtown Brooklyn	3.89	48.8%

*A much more extensive list is available in the Overall Report and Jupyter notebooks

Most Prominent Feature Families for Felony Possession

The feature families with the largest percentage of the total predictive value of the feature set for classifying felony possession crimes are, in descending order:

- Public housing developments, or “projects”: 36%
- Premises types: 16%
- Police precincts: 13%
- Transit/Subway stations: 12%
- L1 distances from NYC landmarks: 4%
- Police Jurisdictions: 4%
- NYC parks: 3%
- Holidays: 3%
- L2 distances from NYC landmarks: 2%

Suspect Race’s Role in Felony Possession

Among those arrested for cannabis, African-Americans arrested for cannabis have a moderately negative relationship with felony possession, and arrested Hispanics have a weakly negative relationship with it. Asians/Pacific Islanders arrested for cannabis have a strong relationship with felony possession.

Suspect Race/Ethnicity	**Coefficient**
African-American	-0.42
African-American Hispanic	-0.14
White non-Hispanic	-0.07
White Hispanic	0
Unreported Suspect Race	0.13
Native American	0.18
Asian or Pacific Islander	0.71

Features Differentiating Misdemeanor Sales*

If a cannabis crime in New York City between 2006 and 2018 has these features, it is X% more likely to be a misdemeanor sales crime.

Feature	Coefficient	% More Likely Misdemeanor Sales
Occurred in Washington Square Park	2.76	15.8%
Jackson Ave. Subway Station, Bronx	2.41	11.1%
Occurred in grocery store or bodega	1.89	6.6%
Story Playground, Bronx	1.87	6.5%
Wald Housing Project	1.77	5.9%
Parkside Playground, Brooklyn	1.77	5.8%
Unreported suspect age	1.72	5.6%

*A much more extensive list is available in the Overall Report and Jupyter notebooks

Most Prominent Feature Families for Misdemeanor Sales

The feature families with the largest percentage of the total predictive value of the feature set for classifying misdemeanor sales crimes are, in descending order:

- Public housing developments, or “projects”: 34%
- Transit/Subway stations: 16%
- NYC parks: 14%
- Premises type: 11%
- Police precincts: 10%
- L1 distances from NYC landmarks: 3%
- L1 distances from NYC landmarks: 2%
- Jurisdictions: 2%
- Transit Districts: 2%
- Holidays: 1%

Suspect Race’s Role in Misdemeanor Sales

Among those arrested for cannabis, African-Americans (Hispanic and not) arrested for cannabis have a moderately positive relationship with misdemeanor sales, and arrested White Hispanics and Asians/Pacific Islanders have a weakly positive relationship with it. Whites, those whose race is unreported, and Native Americans have a moderately negative relationship.

Suspect Race/Ethnicity	**Coefficient**
Native American	-0.50
Unreported suspect race	-0.39
White	-0.38
Asians and Pacific Islanders	0.14
White Hispanic	0.23
African-American	0.43
African-American Hispanic	0.45

Features Differentiating Felony Sales*

If a cannabis crime in New York City between 2006 and 2018 has these features, it is X% more likely to be a felony sales crime.

Feature	Coefficient	% More Likely Felony Sales
L2 distance to Prospect Park	11.41	89,851%
Junius Street Subway Station in Bronx	8.89	7,259%
96th St Subway Station, Manhattan	7.96	2,869%
Douglass Addition Housing Project	7.79	2,426.6%
Tapscott St Rehab Housing Project	6.24	511.5%
L1 distance to Manhattan Bridge	5.96	388%
Sterling Place Rehab Housing Project	5.49	243.1%

*A much more extensive list is available in the Overall Report and Jupyter notebooks

Most Prominent Feature Families for Felony Sales

The feature families with the largest percentage of the total predictive value of the feature set for classifying felony sales crimes are, in descending order:

- Public housing developments, or “projects”: 42%
- Premises type: 13%
- Transit/Subway Stations: 9%
- Police Precincts: 8%
- L1 distance from NYC landmarks: 7%
- L2 distances from NYC landmarks: 7%
- Holidays: 3%
- Transit Districts: 2%
- NYC parks: 2%
- Jurisdictions: 2%

Suspect Race’s Role in Felony Sales

Among those arrested for cannabis, White Hispanics have a strongly positive relationship with felony sales, and non-Hispanic Whites and African-Americans have a moderately positive relationship with it. Those whose race is unreported also have a moderately positive relationship with felony sales. African-American Hispanics have no relationship with felony sales.

Suspect Race/Ethnicity	**Coefficient**
Native American	-2.86
Asian or Pacific Islander	-0.25
African-American Hispanic	0.02
African-American	0.40
White, non-Hispanic	0.41
Unreported race	0.49
White Hispanic	0.72

Conclusion

Although suspect race/ethnicity and other demographics went largely unreported in NYC cannabis crime, by looking at the relationship that cannabis arrests have with specific neighborhoods and the strong relationship that they have with public housing developments in these neighborhoods, it is clear that there was a bias in enforcing cannabis law between 2006 and 2018 towards those of African-American and Hispanic descent and towards residents of public housing, even while cannabis use remains prevalent throughout New York City and is used at the same rates between different racial and ethnic groups (as reported by SAMHSA surveys).

Conclusion

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. 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 the years of 2006-2018.

Conclusion

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.