Project: Capstone Project 1: Data Story & EDA

Daniel Loew

This report summarizes what was done in the Data Story & EDA Jupyter notebook, that can be found at this link:

<https://github.com/danloew/Springboard/blob/master/7.2%20First%20Capstone%20Data%20Story%20%26%20EDA%20Final.ipynb>

As has been reported, 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, even though 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 (see notebook for citations). While looking at both low-level and more serious cannabis arrests including felony sales, this report aims to provide a more full picture of the factors that trigger cannabis arrests in New York City. In order to do so, machine learning classification methods will be applied to predict different levels of cannabis crime severity based on a set of predictive features, as well as classifying cannabis crimes or non-cannabis crimes based on these features.

These ML methods will be used on all cannabis crimes between January 1st, 2006 and December 31st, 2018 in New York City as reported by the NYPD's Complaint Data historic dataset (<https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i>). As more fully reported in the Data Wrangling report of this capstone project, three useless features of this dataset were dropped, and then the dataset was filtered to only include cannabis arrests. The five crime level target features were then derived from penal code and law category features native to the original dataset, resulting in differentiation of misdemeanor, violation, and felony possession cases, as well as misdemeanor and felony sales cases.

Missing data was either filled in with “unknown” values for select features, and a small number of rows were dropped from the dataset for cases whose missing location data would have made later analysis difficult. A few of these rows also had missing datetime data. Cases from years prior to 2006 were also dropped. Datetime-formatted crime start, crime end, and duration features were derived from separate string date and time features, and isolated year, month, and day features were extracted. Daily time-window features such as lunch hour and evening rush hour were derived from the datetime features, and several holiday features were derived as well. Proximity to city landmarks was calculated for all records through the derivation of a set of landmark-defined features. Clean versions of victim and suspect race, age, and sex data were created after coding noise values to an “unknown” code. These cleaning steps were repeated for a sample of non-cannabis crimes of equal size to the cannabis crime DataFrame. Versions of these DataFrames were sub-setted to only include records where suspect race was reported. EDA-friendly versions of the DataFrames were saved that maintained their categorical variable structure. ML-friendly versions of the DataFrame were also created and saved, where categorical features were split into binary features using pandas’ “.get\_dummies()” method.

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.

51% of cannabis arrests with the suspect's race reported were of African-Americans, 27% of white Hispanics, and 11% of black Hispanics, for a total of 89% of total cannabis crimes with the suspect's race reported being of African-American or Latino people. Only 8% of these arrests were of white people, and 2% were of Asian or Pacific Islander people. Looking just at age shows that 84% of cannabis arrests were made of people between the ages of 18-44. 90% were of males. 40% of cannabis arrests are of African-American males younger than 45, and 32% are of Hispanic/Latino males younger than 45, for a total of 72% of all cannabis arrests in New York City between 2006 and 2018 being of young African-American and Hispanic/Latino males. This corroborates the racial disparity data reported elsewhere.

One of the striking things about cannabis arrests in New York City are that 92.6% of them are for simple misdemeanor and violation possession charges, which is the vast majority. 1.7% are for felony-level possession, 5.2% are for misdemeanor sales, and 0.5% are for felony sales, the latter being arguably the top priority if drug use prevention was the goal. The same racial disparity described above holds true for all levels of cannabis crime. More violation possession arrests are made of white perpetrators than of black Hispanic perpetrators, but the difference is only 3%. Also, it should be noted that violation possession charges are the lowest level of cannabis arrests, and that the majority of violation possession charges are still of African-Americans and white Hispanics. More whites are arrested for felony possession charges than black Hispanics and the same amount of whites are arrested for felony sales charges as black Hispanics, but the difference is less than a percentage point and it bears mentioning that the sample sizes for these groups are very small.

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

The vast majority of cannabis arrests are for misdemeanor possession, and in Manhattan they are heavily concentrated in the Bronx, Inwood, Washington Heights, and Harlem, which have large populations of African-American and Latino residents. In Brooklyn, arrests are concentrated in neighborhoods like East New York, Cypress Hills, Brownsville, Crown Heights, Flatbush, Bedford-Stuyvesant, and Bushwick. Again, these neighborhoods have large populations of African-American and Latino residents. Violation and felony possession are peppered throughout, but they are concentrated in the neighborhoods mentioned. Manhattan, Queens, south and west Brooklyn, and Staten Island have significantly fewer arrests. It bears mentioning that Staten Island is majority white, and the clusters of arrests are centered around housing projects like Stapleton and Park Hill. Because sales are so different than simple possession, and for ease of visualization, cannabis sales were visualized separately. Sales arrests tended to fall within the same neighborhoods as possession arrests.

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

It would be interesting to see which neighborhoods of Manhattan are responsible for these differences. Police precincts offer a route to explore these smaller geographic zones. The top 10 police precincts with the highest amounts of misdemeanor cannabis arrests and cannabis arrests overall are all in the Bronx and Brooklyn. The demographics in these neighborhoods reflects the racial disparity seen in cannabis arrests. The precincts with the most violation possession charges differ however, being largely in Midtown Manhattan and to a lesser degree in Central Brooklyn. Jamaica (in Queens), Washington Heights (in Manhattan), and Inwood (northernmost Manhattan) are also in the list of police precincts where the most felony possession charges are made. All of these neighborhoods have a predominantly African-American and Latino population. For misdemeanor sales, Greenwich Village and the West Village (in Manhattan), and Western Harlem are also common. The Bedford-Stuyvesant neighborhood of Brooklyn and East Harlem also show up on the top 10 list of police precincts with the highest concentration of felony sales arrests. Again, both of these neighborhoods have a predominantly African-American and Latino population.

An intriguing part of the NYPD dataset is a feature that describes the premises type that the arrest occurred in. As can be seen below, the majority of cannabis arrests happen either on the street or in the New York City housing projects. Violation possession charges also occur in the New York City subway system.

The next portion of this report will be focused on using hypothesis testing to see how likely it is that the racial and geographic disparities seen in cannabis arrests are due to simply chance. Machine learning will be used later in the report cycle to look at which features are most predictive of cannabis arrests. As can be expected, the jurisdiction responsible for the majority of cannabis arrests are the NYPD, the New York City Housing Authority (NYCHA), and to a much lesser degree the N.Y. Transit Police. The fact that 19% of cannabis arrests fall under the jurisdiction of the NYCHA shows how heavily policed these public housing projects are. The fact that the N.Y. Transit Police takes the NYCHA's place for violation possession charges show an interesting difference in enforcement of the different cannabis types, and reflects the fact that the premises type for violation possession is frequently in the N.Y. subway system.

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

Cannabis arrests occur more frequently during certain times of the day. 39% occur during the daytime (6 am - 6 pm), and 61% occur during the nighttime (6 pm - 6 am). The work day (9 am - 6 pm) composes most of the daytime arrests, and 37.5% of the total. Early morning (6 am - 7:30 am) and the morning rush hour (7:30 am - 9 am) have very little arrests (0.6% and 0.9 respectively), but this picks up during the lunch hour (12-1 pm), when 3.9% of the arrests are made. The long New York metropolitan area's evening rush hour (4:30 pm - 7 pm) straddles the daytime (6 am - 6 pm) and nighttime (6 pm - 6 am) windows, but one sees a fairly high concentration of arrests (18.1%) happening during this time window.

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

It has been well reported that during Mayor Bloomberg's time as mayor, cannabis arrests reached their peak. One can see that 2006 has 15,127 arrests, and that this increases to 24,468 arrests in 2010. This holds fairly steady for 2011 (23,827), drops a bit in 2012 (20,611) as criticism of Bloomberg's "stop and frisk" program mounts, and then drops significantly in 2013 (16,206) when the "stop and frisk" program was judged as unconstitutional. Mayor DeBlasio, who vowed to reverse the program, took office in 2014, but cannabis arrests remained fairly consistent in that year compared to 2013 (15,787). By 2015, the number was still fairly high but dropped significantly (11,424). This number stayed consistent through 2017, and then dropped by half in 2018 as discussions of cannabis legalization in New York intensified. Each month of the year has about the same amount of cannabis arrests, but August has the highest number and the number drops in November and December during the Holiday season. Each day of the month has a fairly consistent number of cannabis arrests, ranging from 5,660 to 7,900 arrests a day. The number drops somewhat in the last 10 days of the month. The 31st has roughly half the arrests as the rest of the month, because not every month has 31 days.

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

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