Project: Capstone Project 1: Project Proposal
Daniel Loew

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

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

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

As with all data science projects, initial EDA will look at what the descriptive statistics on demography, jurisdiction, and geography tell us about cannabis arrests. Visualizations will be carried out to help with this EDA, as will hypothesis testing that looks at whether people of different demographic groups are equally likely to be arrested for cannabis, and how the level of cannabis arrest differs between these groups. Geographic factors like borough will also be subjected to hypothesis testing. Later in the project, machine learning classification methods will be used to unpack which features most help predict a cannabis arrest versus a non-cannabis arrest. Then, classification methods will also be used to see which features help predict which level of cannabis arrest will be made. These levels are misdemeanor possession, violation possession, felony possession, misdemeanor sales, and felony sales.

Deliverables for this project will include Jupyter notebooks with the code for data cleaning, exploratory data analysis, hypothesis testing, and machine learning classification methods. A summary report will also be created, as will slides including visualizations of several important features of the analysis. It is the aim of this project to more clearly elucidate the factors that predict cannabis arrests generally, and specific types of arrests, in order to provide a deeper understanding of the equitability of cannabis arrests in New York City and whether further research efforts are justified to look at what may amount to civil rights violations by way of U.S. drug policy.