Project: Capstone Project 1: Statistical Data Analysis

Daniel Loew

For this portion of the first capstone project, hypothesis testing about the demographic and geographic aspects of cannabis arrests in New York City was the aim. In order to do this, a sample of the original NYPD "NYPD Complaint Data Historic" dataset was taken. A cannabis crime flag was added, as were flags for the five cannabis crime types. A sample of the crimes with the suspect’s race reported was also taken. All the prior data cleaning steps were taken on these sample DataFrames, and can be reviewed in this notebook:

<https://github.com/danloew/Springboard/blob/master/1stCapstone_Data_Cleaning_full_set_for_HT_sample.ipynb>

All hypothesis testing was done in the following notebook:

<https://github.com/danloew/Springboard/blob/master/8.5_1st_Capstone_Statistical_Data_Analysis.ipynb>

All hypothesis testing was done using a t-test of independent samples computed from statistics, through SciPy Stats’ ttest\_ind\_from\_stats function.

As was reported in earlier notebooks, there was a discrepancy between the percentage of cannabis crimes with the suspect’s race reported (15.8%) and the percentage of non-cannabis crimes with the suspect’s race reported (38.1%). A null hypothesis was tested that cannabis crimes are equally likely in having their suspect's race reported by the arresting NYPD officer as non-cannabis crimes. This hypothesis was rejected with a p-value of zero. This test shows that the difference in reporting is not due to random chance; the reason of course is unclear but suggests the need for further research.

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

Inter-group differences were then looked at, with the null hypothesis that African-Americans arrested for a crime are equally likely to be charged for cannabis crimes as white people arrested for a crime.

This was rejected with a p-value very close to zero. This was replicated when comparing white Hispanics and whites, black Hispanics and whites, Asians and whites, African-Americans and white Hispanics, African-Americans and black Hispanics, and white Hispanics and black Hispanics. This suggests a hierarchy of likelihood in being arrested between different racial and ethnic groups.

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

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

This initial statistical EDA shows that there are some interesting demographic and geographic differences in how people are arrested for cannabis crimes in NYC. This will be further explored with machine learning classification methods.