Project: Capstone Project 1: Statistical Data Analysis
Daniel Loew

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-value i 0.12.
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.