Names: Vedika Srivastava, Samantha Shih, Oliver Samuels

Date: 11/22/22

### Project Deliverable 3 (v1 Draft Complete Report) (Due: 11/16)

Client Based Projects: All data should have been collected. All project questions should have been reviewed, answered, and submitted in a written document outlining findings. You will also be asked to submit the associated data and a README explaining what each label/feature in your dataset represents. Your team should meet with the client before this deliverable.

Checklist

* ~~All data is collected~~
* Refine the preliminary analysis of the data performed in PD1&2
* Answer another key question
* Attempt to answer overarching project question
* Create a draft of your final report
* Refine project scope and list of limitations with data and potential risks of achieving project goal
* Submit to Gradescope with the above report and modifications to original proposal

Here are some guidelines - just note that every project is different and some items may not be in scope for the project:

1. Background + Motivation (what is this project about, who will use it, for what purpose, in what context etc.)

2. Previous work (have there been other / different attempts toward these goals in the past?)

3. Data collection

4. Data visualization and exploration

5. Results obtained / questions answered

6. Interpretation + limitations of results

7. Challenges faced

8. Suggestions for the future of the project

1. **Background & Motivation**
2. <https://docs.google.com/spreadsheets/d/1EAC-N-UvhkwQxPfNJplmNXfky3UtzT4c/edit?usp=sharing&ouid=106251797320733438460&rtpof=true&sd=true>

The key question we are trying to answer is whether there is a discrepancy in how prosecutors sentence people based on race and if there is a way we can analyze said discrepancy. The general questions we want go deeper into are the following:

1. Are if Black and hispanic defendants more likely to face harsher versions of similar drug and weapon charges than White defendants?
2. Are Black and Latine defendants more likely to have harsher dispositions in cases with a drug or weapons charge than White defendants?
3. Are Black and Latine defendants more likely to face a harsher type of sentence in cases with a drug or weapons charge than White defendants?
4. Are Black and Latine defendants more likely to have their cases tried in a higher court (more severe offenses, often harsher sentences) than White defendants?
5. Are Black and Latine defendants more likely to face the “trifecta” of Assault and Battery Against a Police Officer, Disorderly Conduct, and Resisting Arrest than White defendants?

These are the primary questions we are attempting to answer throughout the data collection in order to better see how ethnicity may affect people's experiences in the judicial system.

1. **Data collection**

The data for each county was provided by the CPCS’s existing DAMION Database. The District Attorney Management Information Office Network (DAMION) case management system is currently used by the Middlesex County District Attorney's Office (MDAO) to perform a variety of tasks, including maintaining case, victim, and witness information and tracking court events. The Massachusetts District Attorney Association (MDAA) implemented DAMION for all 11 district attorney offices. Each office can tailor the system to its specific requirements. Our group split the Districts given to us as the following, Norfolk, Worcester, Berkshire, Northwestern, Middlesex, Essex and Plymouth. Each dataset gives a variety of information including the case, sentencing type or translation, Crime type, judge name, Disposition type and descriptions, gender, race as well as other logistic information. By dividing said information into groupings and creating visualizations for each we were able to come up with multiple conclusions.

1. **Data visualization and exploration**
   1. **Norfolk**

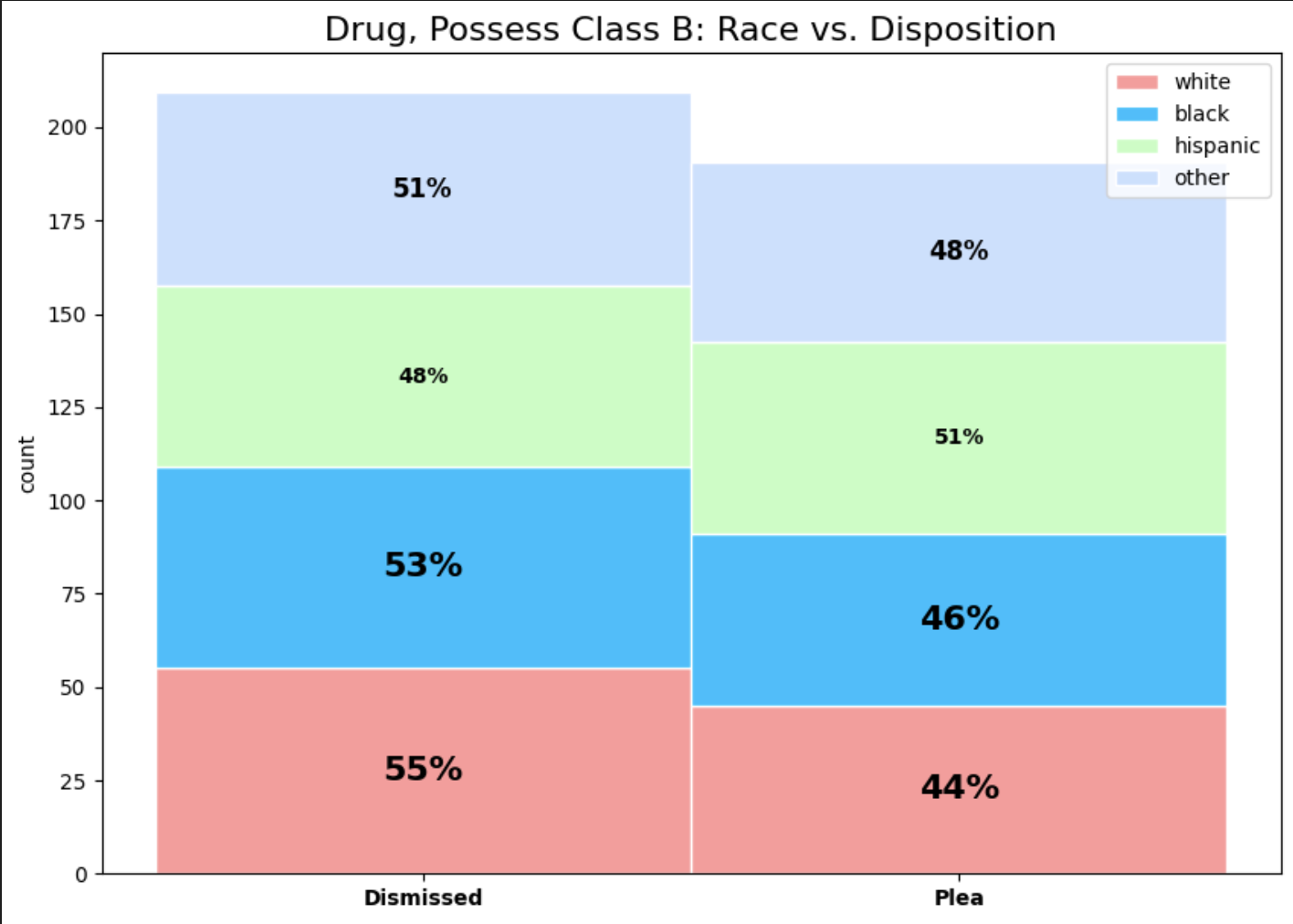


Figure 1. Drug, Possess Class B C94, § 34 Dispositions per Race Group

Drug possession charges are one of the crimes that most disproportionately affect different groups of people, primarily based on race. One of the most common drug-related charges in Norfolk, MA is “Drug, Possess Class B C94, § 34,” the unlawful possession of particular controlled substances, including heroin and marijuana. Visualized here (Figure 1) is the percentage of individuals charged with this offense that had their case dismissed versus plea. The individuals are grouped by race: white, Black, Hispanic, or other. The race group with the highest rate of dismissed “Drug, Possess Class B C94, § 34” cases is white people with about 55% of cases being dismissed. The race group with the lowest dismissal rate is Hispanic at about 48%, which is a significant difference. The race group with the highest rate of plea cases is Hispanic with about 51% of cases resulting in a plea. The race group with the lowest plea rate is white people with about 44%.

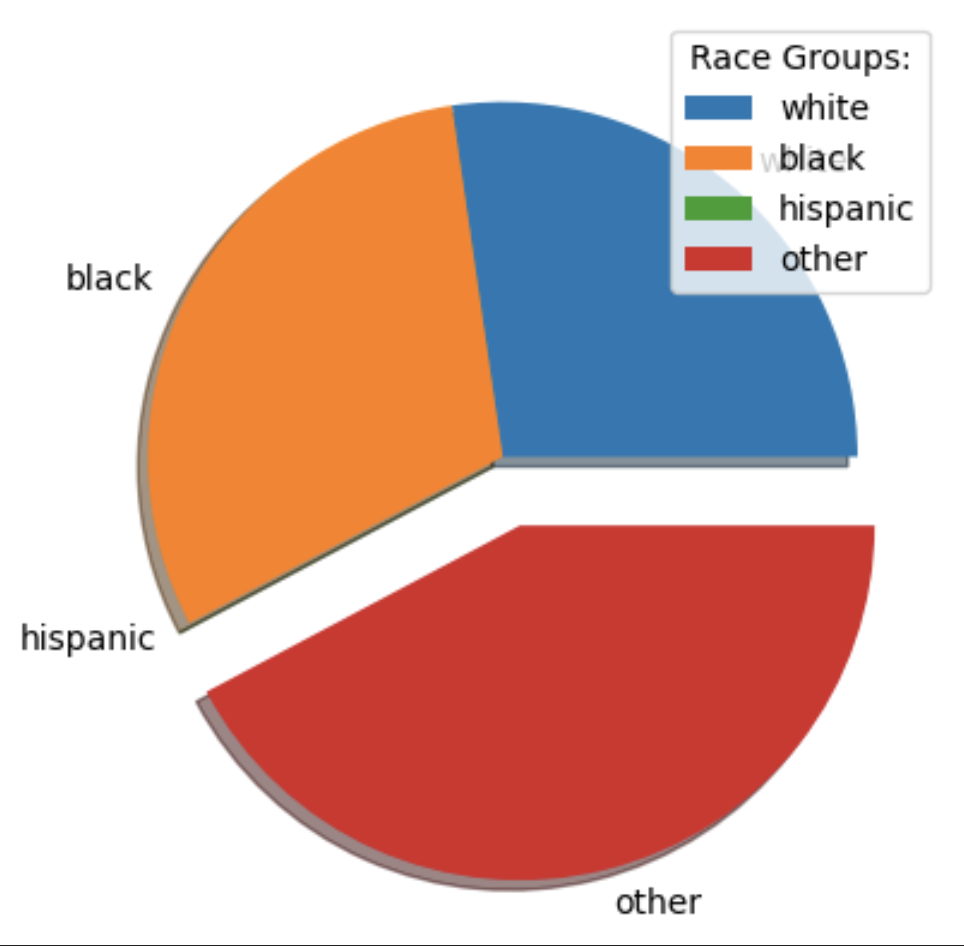


Figure 2. Assault and Battery Guilty Percentage per Race

One of the top charges in Norfolk county was Assault and Battery, 265/13A/B (A&B c265 § 13A). Figure 2 depicts the percentage of individuals that were charged with this crime per race that was found guilty as a result of their disposition. In Norfolk county, the percentages were 0.96% for white individuals, 1.07% for Black individuals, 0.00% for Hispanic individuals, and 1.48% for individuals of other races.

* 1. **Worcester**

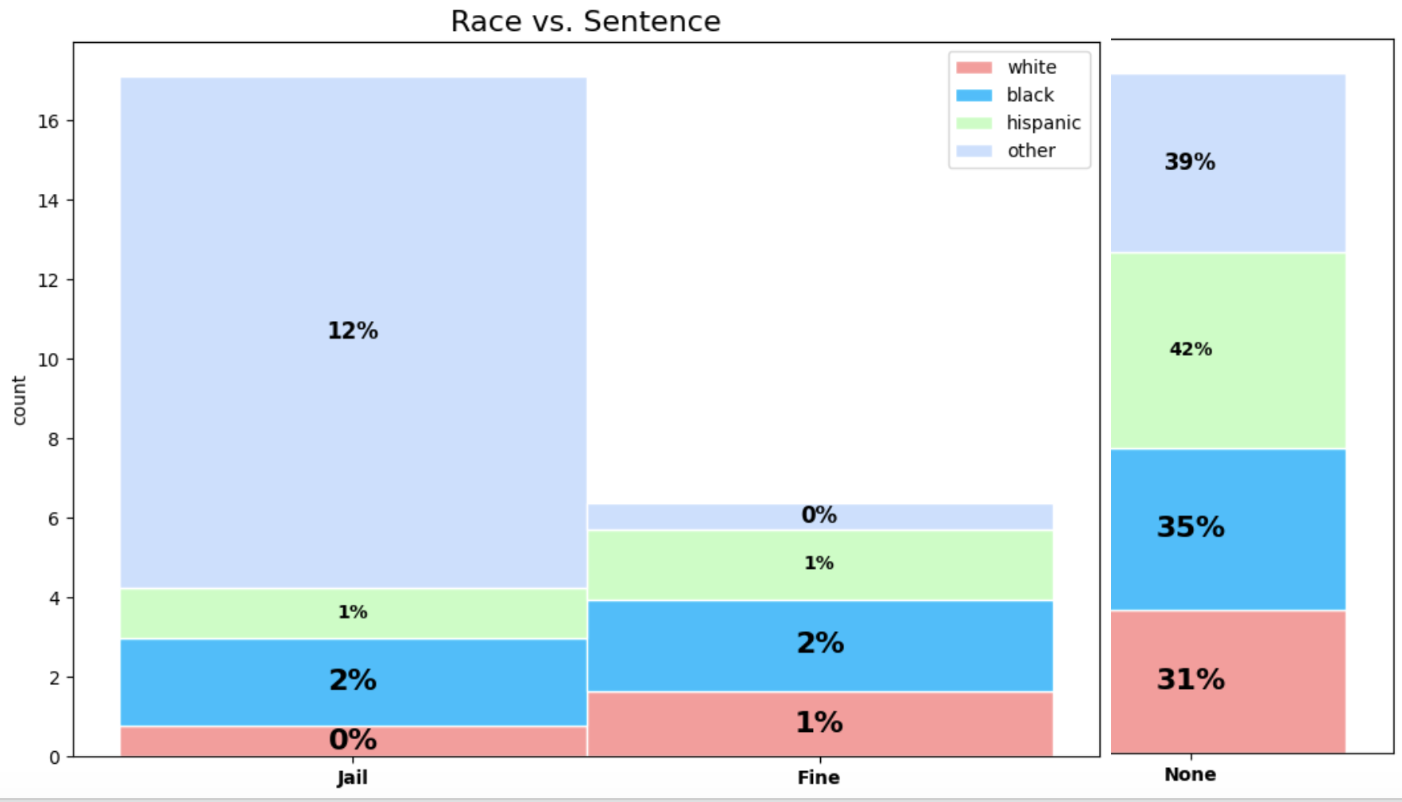


Figure 3. Drug, Possess Class B C94, § 34 Dispositions per Race Group

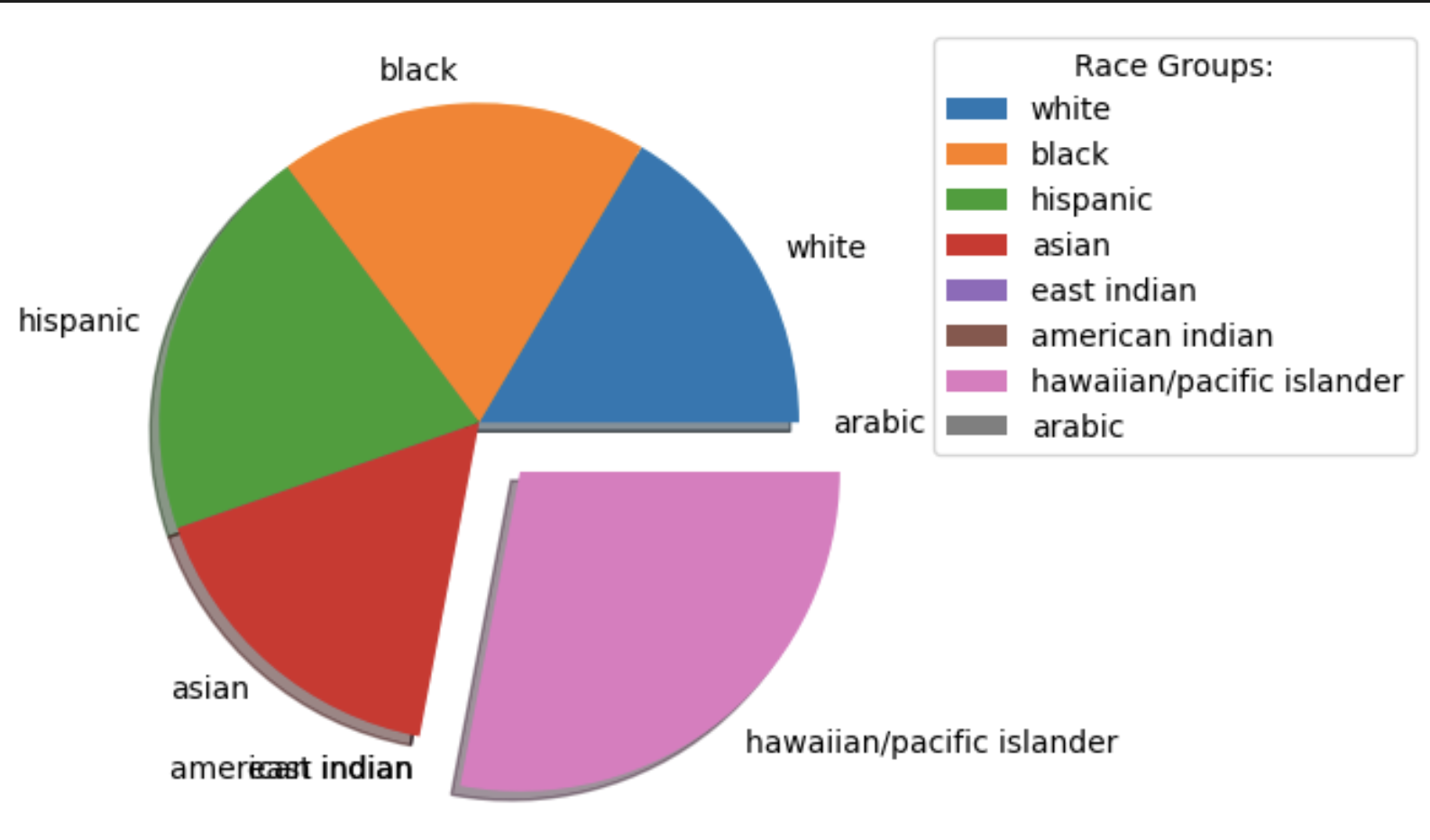


Figure 4. Assault and Battery Guilty Percentage per Race

Similarly, one of the top charges in Worcester county was Assault and Battery, 265/13A/B (A&B c265 § 13A). The Worcester data had more detailed records on each defendant’s race, so this figure has more races defined. Figure 4 depicts the percentage of individuals that were charged with this crime per race that was found guilty as a result of their disposition. In Worcester county, the percentages were 5.37% for white individuals, 6.09% for Black individuals, 6.53% for Hispanic individuals, 0.00% for East Indian individuals, 5.41% for Asian individuals, 0.00% for American Indian (Native American) individuals, 9.09% for Hawaiian/Native Pacific Islanders, and 0.00% for Arabic individuals. Hawaiian/Native Pacific Islanders appear to experience a much higher rate of guilty dispositions for assault and battery charges compared to other races, despite having the lowest number of people charged for that crime of this race (11 Hawaiian/Native Pacific Islanders individuals).

* 1. **Berkshire**

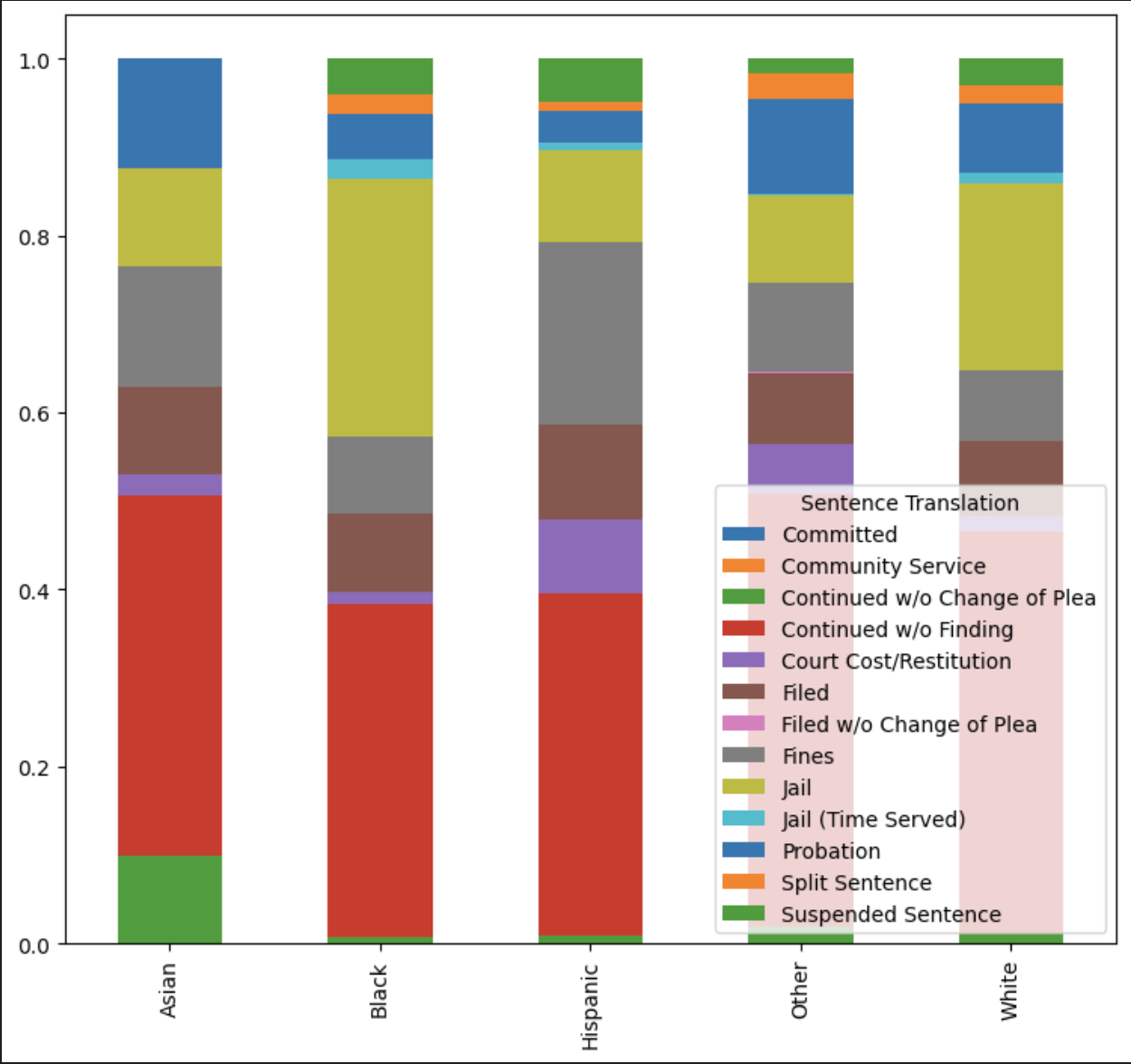


Figure 1. Sentence Translation Percentage Based on Race

The preliminary analysis of Berkshire's data is shown in figure 1. This figure shows the percentage ratio of all the Sentence translations in Berkshire for each race/ethnicities. The ethnicities shown are Asian, Black, Hispanic, Other and White. Other is the grouping of all of the races that were too infrequent to put into their own major group for the sake of the data.As can be seen in the graph that the most common Sentence Translations are Continued w/o finding, Jail, Fined and Committed, but the ratios per race are what need to be analyzed. Seeing the clearly disproportionate Jail sentencing for Black Ethnicity in comparison to the other ethnicities the next step was to analyze specifically the Jail sentencing ratios of each ethnicity

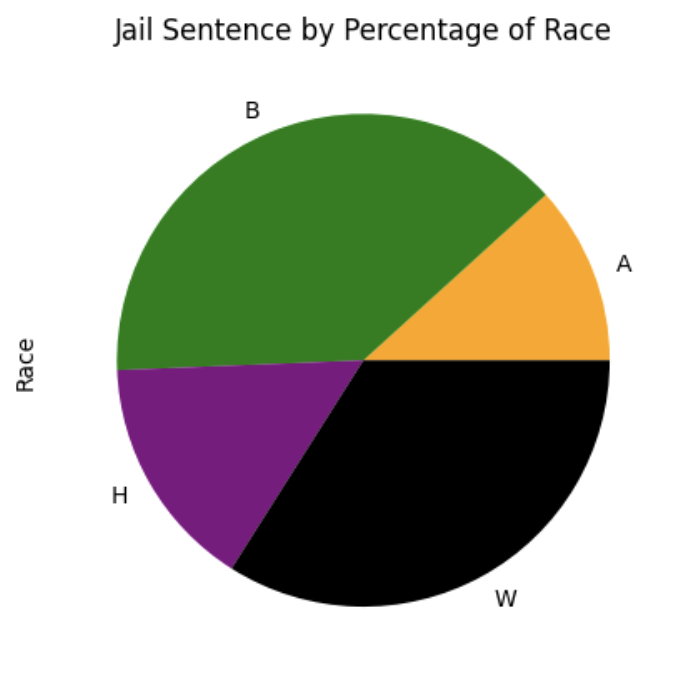
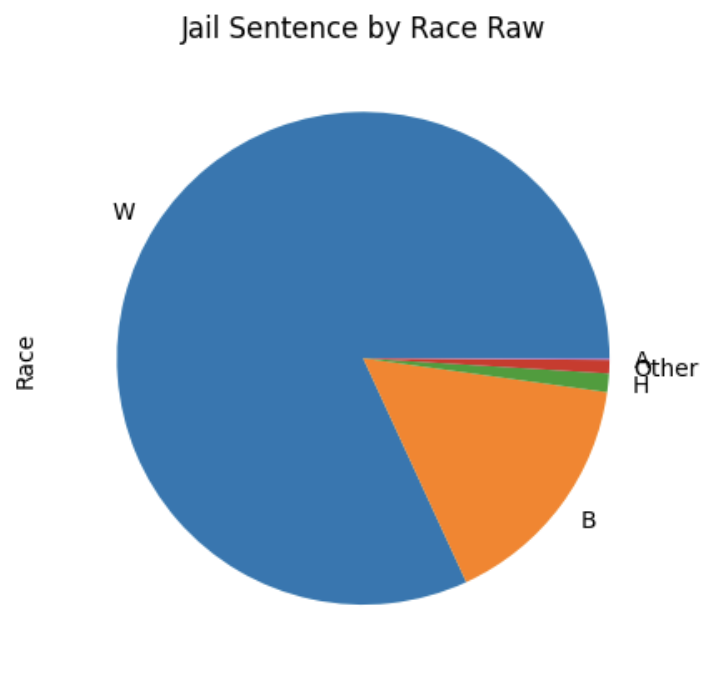


Figure 3 and Figure 4, Raw vs Percentage Ratio of Ethinc Jail Sentences

A Jail sentence is the second most common Sentence translation in Berkshire so the Jail sentence data was split into two visual representations with one being the (raw) total number of people sentenced to Jail and the other being the percentage of total people from that ethnic group having been arrested. As it can be seen even with extraordinarily low numbers in comparison to their white counterparts, Black, Hispanic and Asian ethnicities take up a much higher percentage than what proportionally should be expected given their low total numbers.

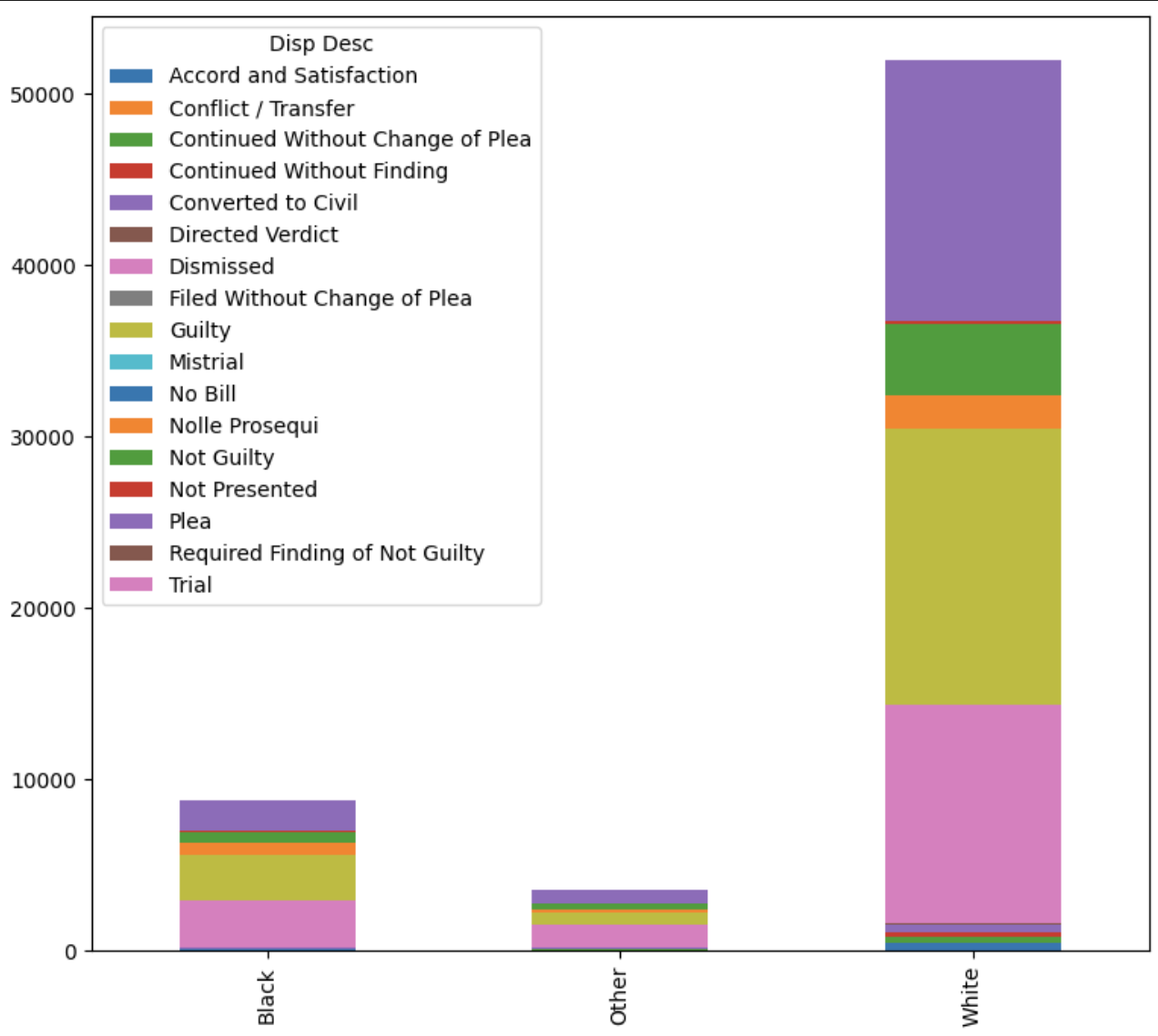


Figure 5. Disposition description By Race

Similarly to the Sentence Translation Figure 5 is the preliminary analysis of the Disposition description classified by Race/ethnicity. The most common Dispositions are clearly dismissed Guilty, Converted to Civil, and not guilty. Since there were so many other smaller classifications of ethnicities or missing disposition data once again a separate Other grouping was made.(Other graphs showing percentages of most common groupings to be shown in final/ graphing documentation)

* 1. **Northwestern**

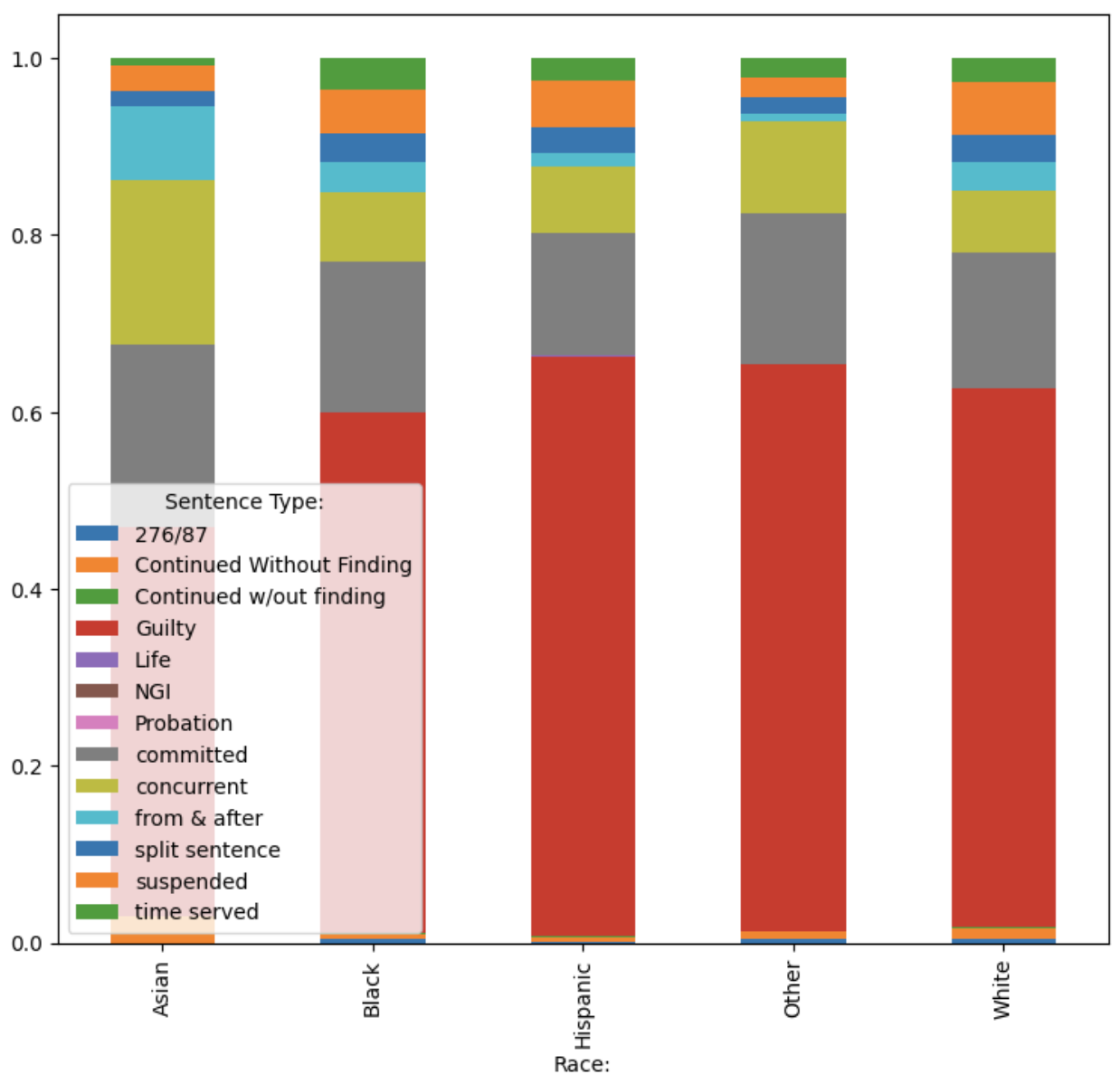
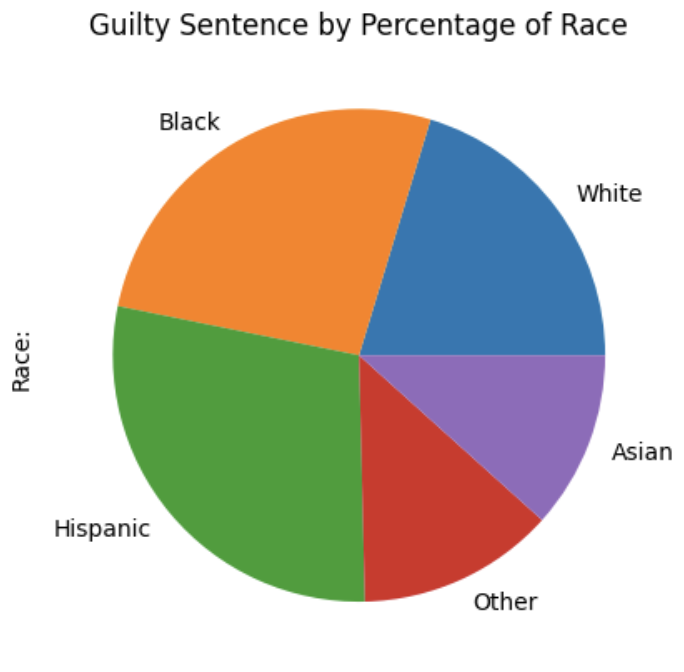
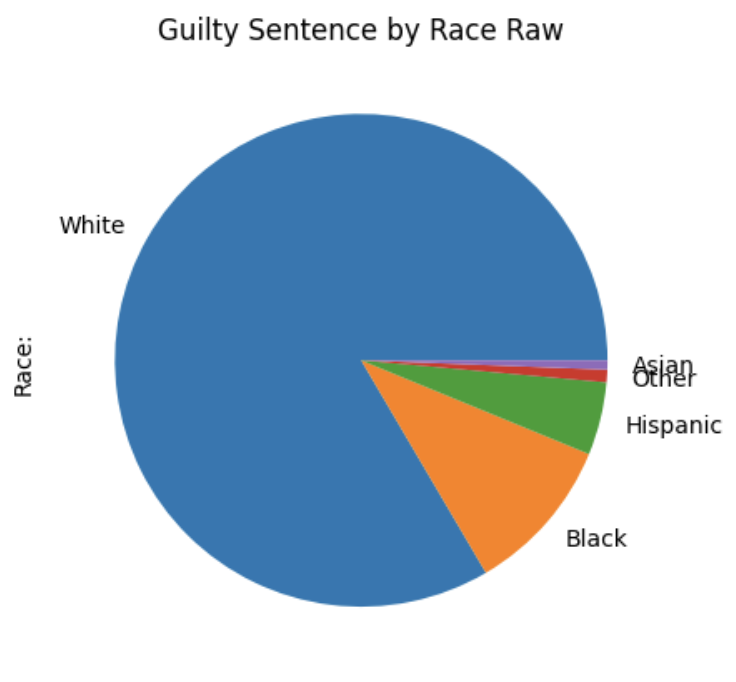
****

Figure 1. Sentence Type by Race

Figure 1, similarly to the first figure in Berkshire, is meant to be a preliminary analysis of the Sentencing Type which is the equivalent of Sentence translation of Berkshire to see trends in race. The bar graphs are split up by Race with Asian, Black, Hispanic, Other and White. The clear most common ones are guilty sentencing, committed, concurrent and suspended sentences. Since Guilty is clearly the most dominant of the sentences I it is further expanded on in the next figures.

****

**Figures 2 and 3: Raw and percentage ratio of Guilty Sentence by Race**

In FIgures 2 and 3 we can see that though hispanic asian and Black Ethnicities are the lowest overall occuring the ratios of those found guilty are significantly higher in comparison to their white counterparts. This is similar to the findings in figures 3 and 4 of Berkshire both go to show the clear disparity between the ethnicities in sentencing and trials

* 1. **Essex**

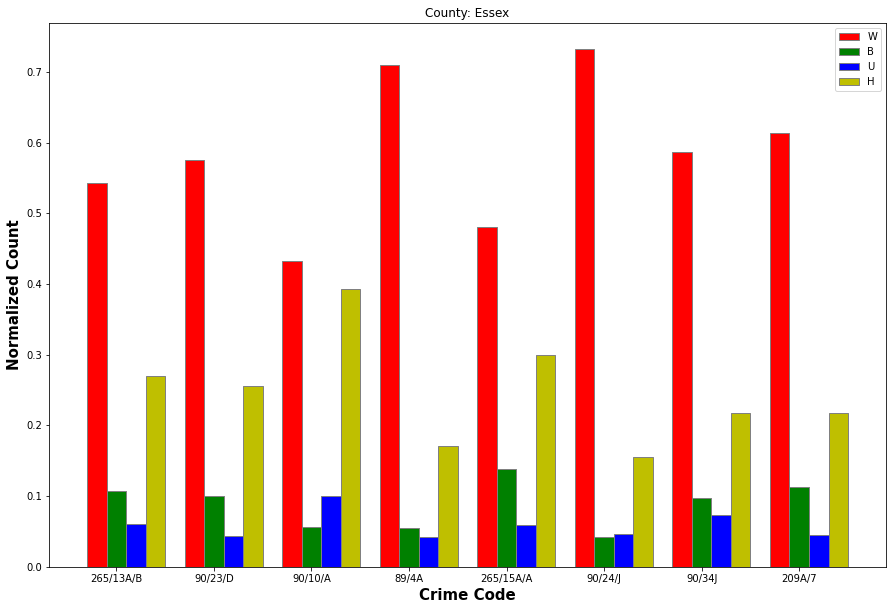


Figure 5. The top 10 crimes committed in Essex and what percentage

of these crimes were committed by what ethnicity

In Essex county, from preliminary analysis I found out that there were 29247 Races present in the dataset; however , we don't have enough case representation in the dataset for all these races; hence I decided to do visualization of only the 4 races that had enough representation in the dataset. Coincidentally the 4 prominent races came out to be White, Black, Hispanic and Native Americans/Indian Americans/Alaskans, which are also the Races under review. In Figure 5. the race represented by W - Whites have committed the crimes 90/24/J, the race represented by H - Hispanics have committed crimes 90/10/A, the race represented by B - Blacks have committed crimes 265/15A/A while the race U - Native Americans/Indian Americans/Alaskans committed crimes 90/10/A. In Essex we can also see that Whites and Hispanics are the Races that committed the most crimes, it is explainable for Whites because they have more population as compared to Hispanics. We can also say that the Hispanics commit the most crimes in Essex.

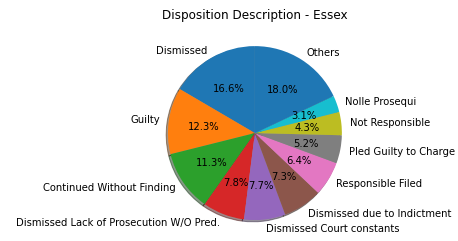


Figure 6. Representation of the outcome of various cases in Essex

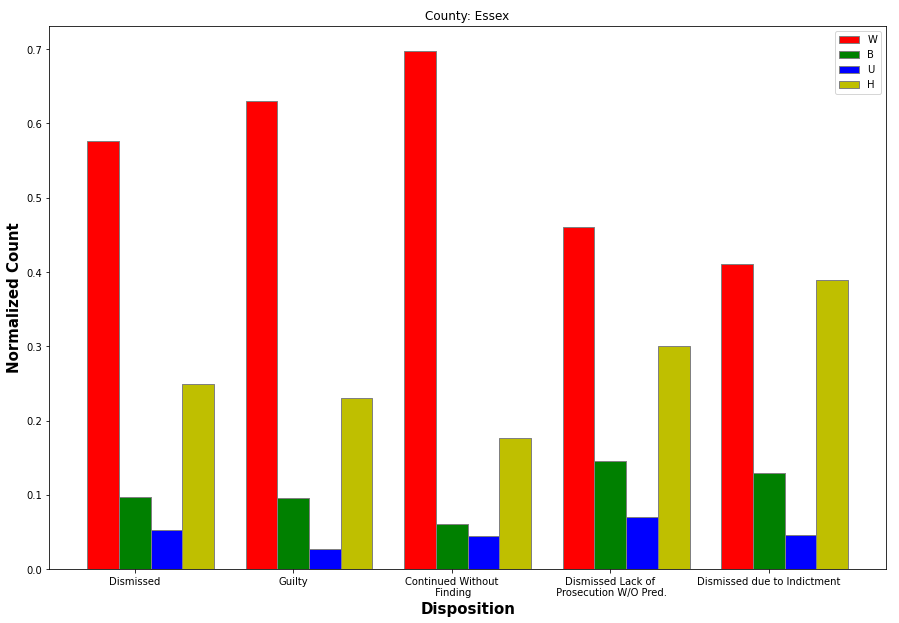


Figure 7. What percentage of each prominent Race was given what kind of Disposition in Essex

Figure 6 is a generic representation of what kind of disposition was charged in Essex, this also gives a rough idea of how strict punishments are carried out in Essex. There are around 2414 unique dispositions that were carried out but most of them just had 1 case that had that disposition carried out hence it doesn’t make sense to use all these for deriving results. The prominent dispositions have been represented in Figure 6 while the other less prominent ones have been grouped together in the “Others” category. We can see that there is a 60-40 ratio of Dismissed cases and Guilty charges. It is also important to note the harshness of disposition given to each of the Race, with reference to Figure 7 we can see that while Hispanics committed the most crimes (reference Figure 5), most of their cases were dismissed (due to indictment).

* 1. **Plymouth**

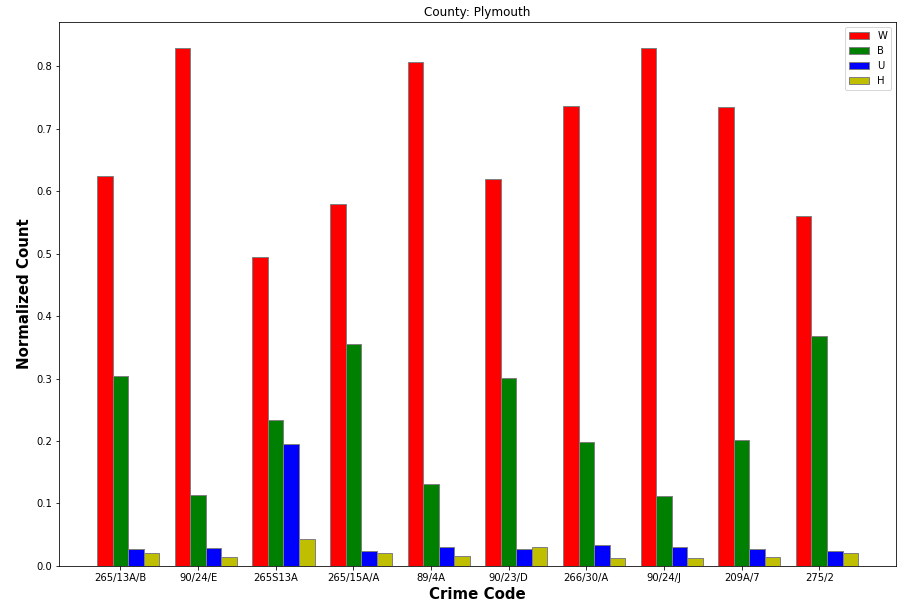


Figure 8. The top 10 crimes committed in Plymouth and what percent of these crimes were committed by what ethnicity

In Plymouth, similar to Essex, again all the races present in the dataset didn't have substantial representation so have taken only the prominent races for visualization purposes; the prominent races came out to be the same as that of Essex which makes the comparison between these two counties ever easier. With reference to Figure 8 we can see that the majority of people committing any type of crime in Plymouth are Whites and also it is essential to note that according to Figure 9 most of the cases in Plymouth are dismissed. This further solidifies the belief that Race has something to do with the king of judgment passed. Figure 10 further explores this hypothesis. Again most of the cases of the race Whites were continued without findings. For Plymouth visualization alone doesn’t tell us a lot about the relation between Race and Disposition.

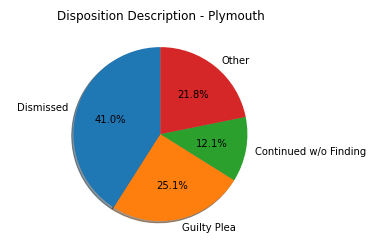


Figure 9. Representation of the outcome of various cases in Plymouth

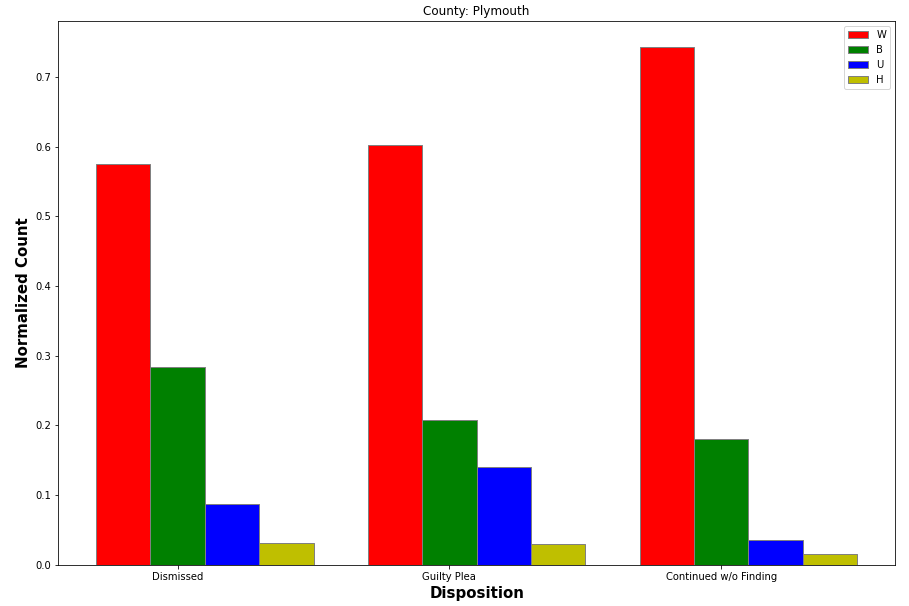


Figure 10. What percentage of each prominent Race was given what kind of Disposition in Plymouth

1. **Results & Questions Answered + Interpretation** 
   1. **Norfolk**

The overarching key question was whether there was bias, specifically racial bias, present during trial proceedings. By analyzing the outcomes of different cases under the same charge, we can see if the outcome is influenced by what identity group the defendant belongs to. In the figures for Norfolk, white individuals are consistently the racial group with the lowest rate of plea trials or guilty verdicts. Specifically, one of the charges we examined was illegal drug-possession charges. We wanted to determine if Black and Hispanic people were more likely to face a harsher outcome than white defendants. Figure 1 depicts Black and Hispanic individuals had the highest rate of receiving a plea deal rather than having their case dismissed. This aligns with the initial question: Black and Hispanic individuals are least likely to have their case dismissed while white individuals have the highest dismissal rate for drug-related charges.

* 1. **Worcester**

Similarly to Norfolk, individuals that identified as a racial minority were found to be at a disadvantage compared to white individuals of the same charge. One of the most common charges in Worcester was Assault and Battery. Within that charge, the racial groups with the highest rate of a guilty verdict were Hawaiian/Native Pacific Islanders, Hispanic, and Black. It is clear that defendants of different races face more lenient consequences. Racial minorities are consistently facing harsher trials in comparison to their white counterparts and this pattern depicts the racial bias present in the Massachusetts justice system.

* 1. **Berkshire**

One of the questions asked was if there is a discrepancy in Sentence translation based on race specifically for Black and hispanic ethnicities, it's clearly shown that there is an overall higher ratio of jail time and fines.. For Figure 1 we can see clearly that Hispanic Ethnicity people are fined more on average than their white counterparts, that black ethnicity people are Jailed at a much higher average than all of their counterparts, that Asian ethnicity people are given Suspended Sentences significantly more than their counterparts.

* 1. **Northwestern**

A major question asked is if there is a bias in sentencing toward black or hispanic ethnicities. Both Figure 2 and 3 of Northwestern explicitly show how hispanic and Black ethnicities clearly have a proportionally larger number of guilty verdicts in comparison to their white counterparts, even more so for hispanic defendants than black though both are severe.

* 1. **Essex**

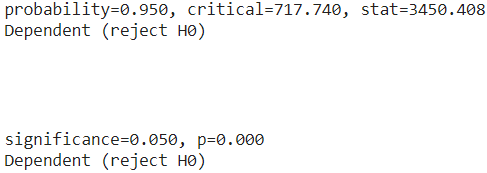
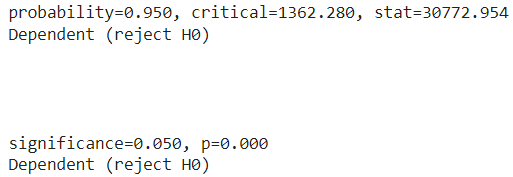
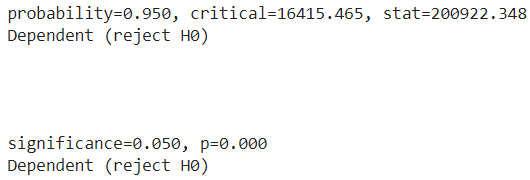


Figure 17 (a) Figure 17 (b) Figure 17 (c)

Figure 17(a), 17(b) and 17(c) are the analysis of the Chi-Square Test of Independence results corresponding to Figure 11, Figure 12 and Figure 13. We can observe that in all the cases the null hypothesis is not true hence confirming that there exists a relationship between the Race and the Charge that was filled. The Essex county data was missing an essential parameter of Sentence (Code/Type/Translation/Description) hence making it difficult to understand if Race also had a role to play in the kind of Sentence that was carried.

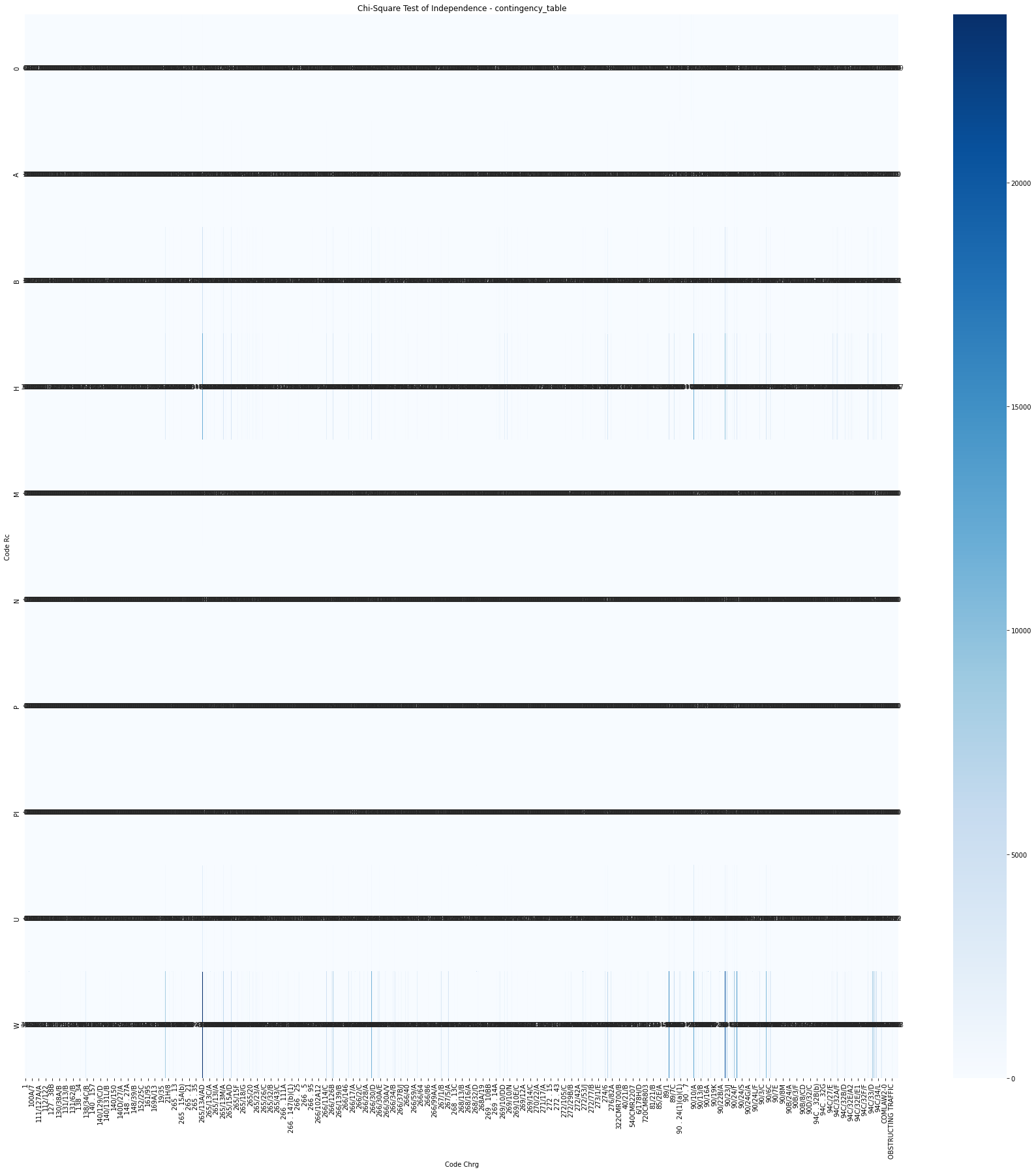


Figure 11. Contingency Table of Chi-Square Test of Independence on Race vs Charge when all the Charges were considered in Essex

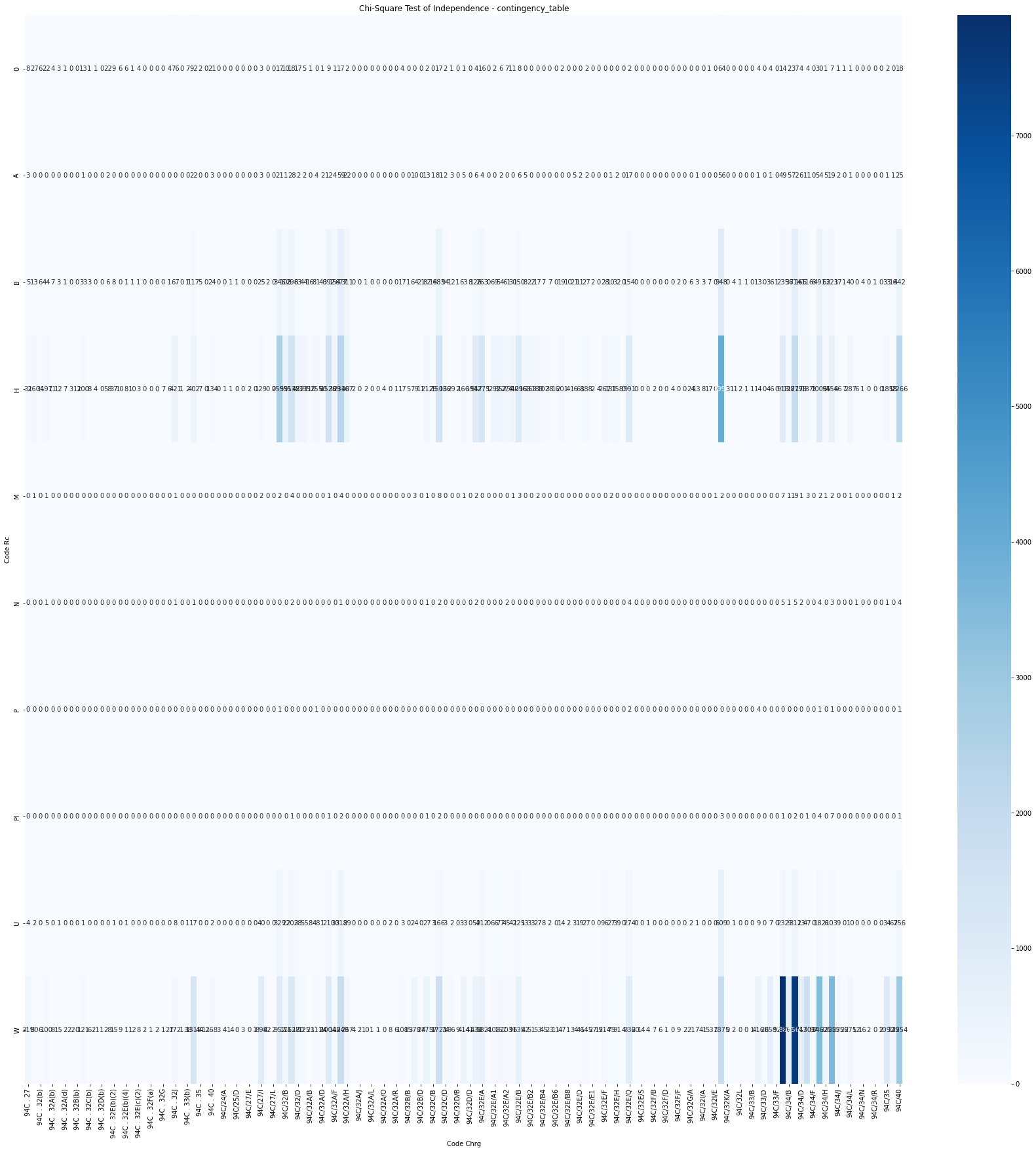


Figure 12. Contingency Table of Chi-Square Test of Independence on Race vs Charge when only Drug Charges (94C) were considered in Essex

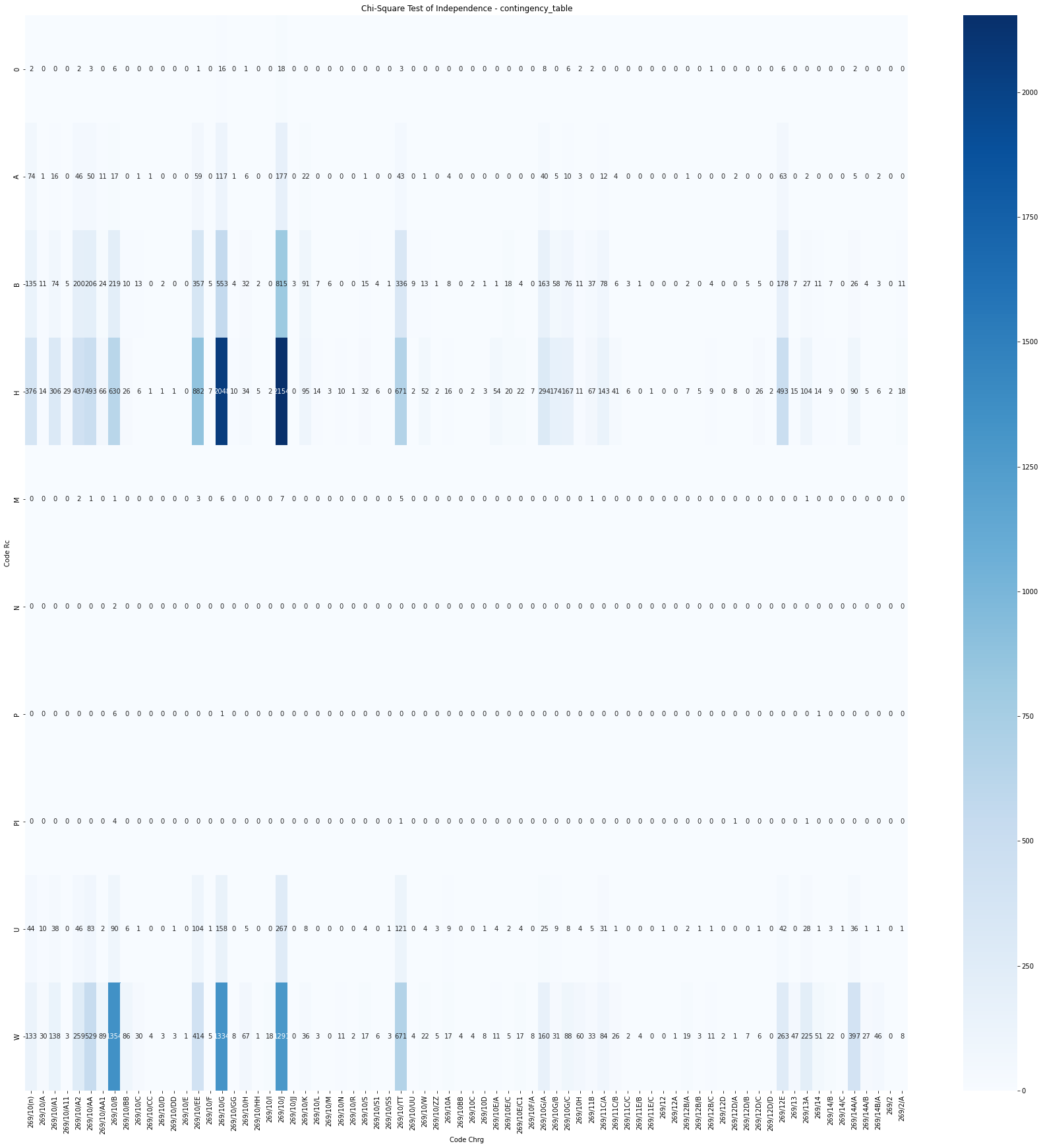


Figure 13. Contingency Table of Chi-Square Test of Independence on Race vs Charge when only Firearms Charges (269) were considered in Essex

* 1. **Plymouth**

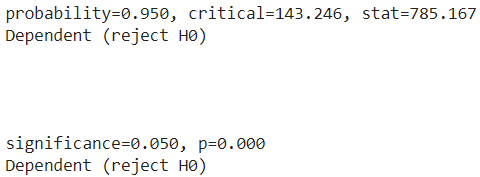
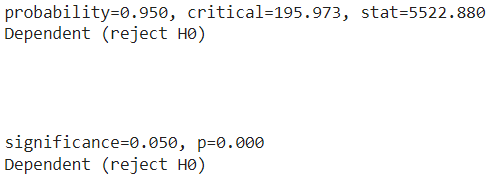
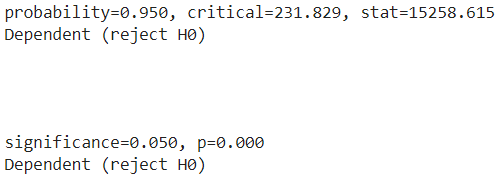


Figure 18 (a) Figure 18 (b) Figure 18 (c)

Figure 18(a), 18(b) and 18(c) are the analysis of the Chi-Square Test of Independence results corresponding to Figure 14, Figure 15 and Figure 16. We can observe that in all the cases the null hypothesis is not true hence confirming that there exists a relationship between the Race and the Sentence Type that was given.

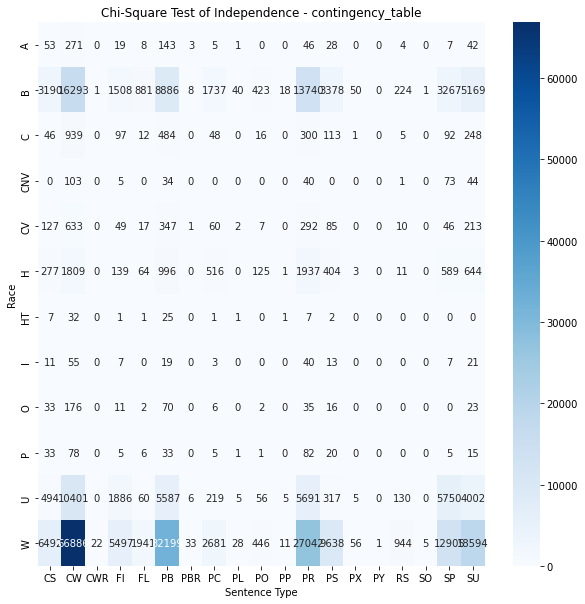


Figure 14. Chi-Square Test of Independence on Race vs Sentence Type when all the Charges were considered in Plymouth

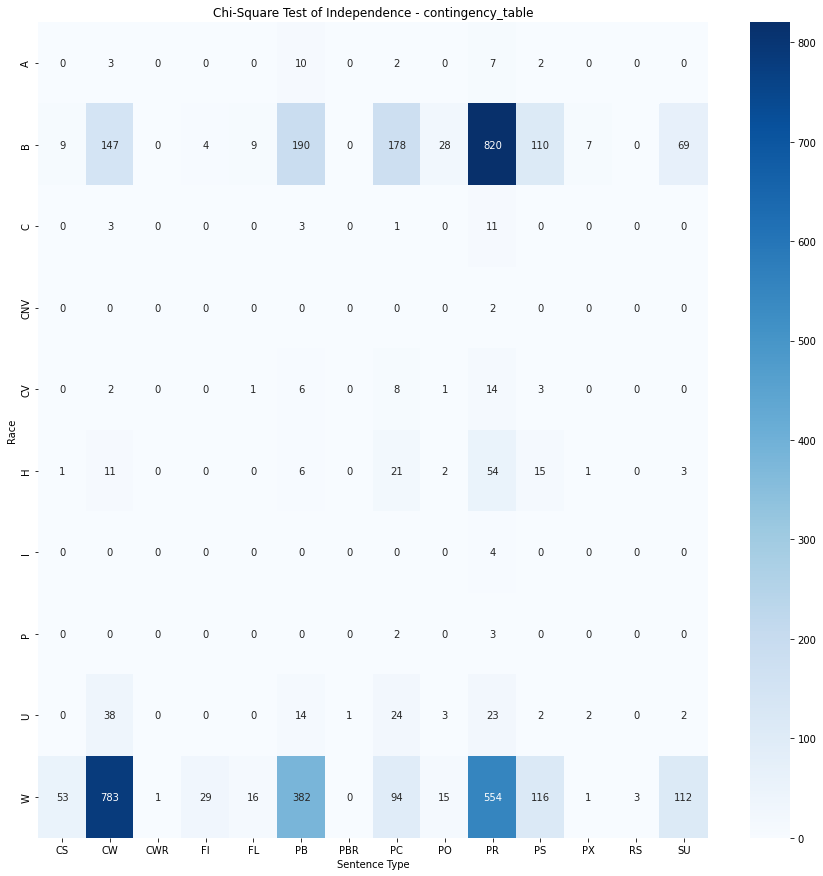
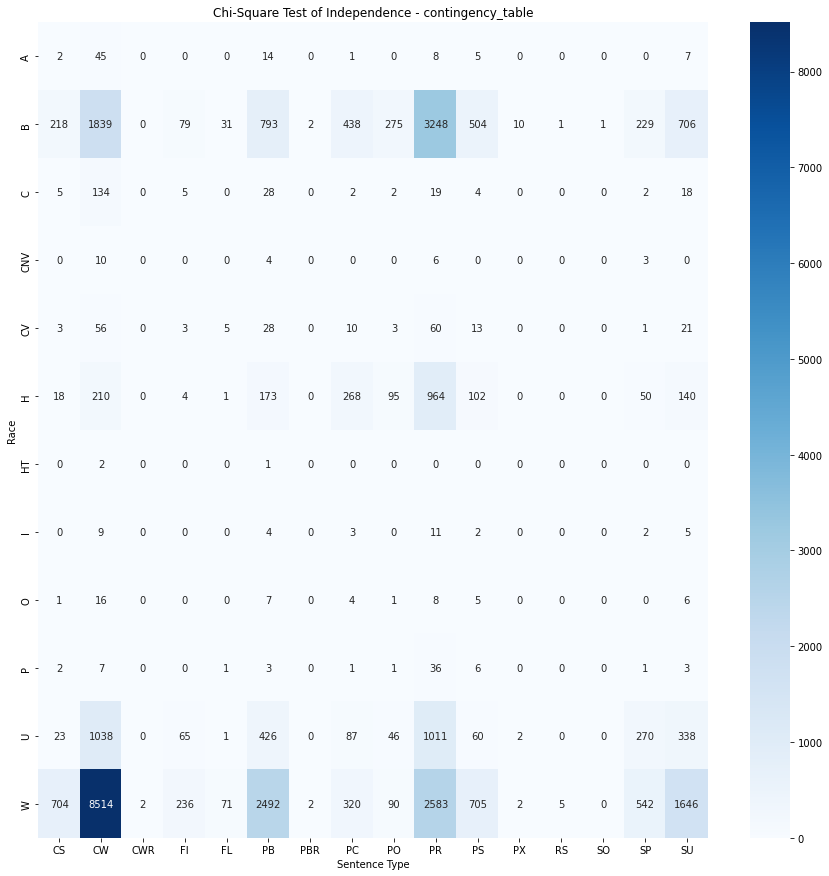


Figure 15. Chi-Square Test of Independence on Figure 16. Chi-Square Test of Independence on Race

Race vs Sentence Type when only Drug Charges vs Sentence Type when only Firearm Charges (269) (94C) were considered in Plymouth were considered in Plymouth

1. **Challenges Faced**

A common challenge we faced across all the data sets was inconsistency among the different datasets, making it difficult to compare the different locations against each other. The data provided for each county doesn’t have the same columns making it difficult to understand and process data, some counties are missing the essential column of Sentence (Type/Translation/Description). The interpretation of meaning of data values was also another challenge we faced. Though we were given the data set key that allowed us to further interpret the data such as specific races, disposition description meanings and types of translations the sheer amount of data analysis was made much more difficult as a result unfortunately. The varied formatting of documents as mentioned above mixed with the initial lack of understanding made for a confusing start to say the least.

We also experienced an unexpected change in the group late in the project. One of our group members decided to drop the class as a whole and unfortunately, he did not make progress on two of the counties, Middlesex and Sufflok. Since it was late in the process and each of us had already been data exploring and analyzing our own two assigned counties, we have decided to exclude Middlesex and Sufflok in our final report due to the lack of group resources. We also had various time constraints as a team. Unfortunately, the unavailability of each of our team members this semester prevented us from meeting formally to work on the project as much as we would have liked to to fully develop some of our ideas and visualizations as a team leading to a more fragmented approach which though did work could have been slightly different or varied in overall results.

1. **Suggestions**

A few suggestions for the future of this project would be as follows. Having a definitive data labeling system. One of the biggest problems we have had with our data collection is the massive amounts of variation in categories and inputs as well with the data. If there was a definitive key or system that each district could go by, the data analysis would be much quicker and a lack of a value is easier to interpret over duplicate values that mean the same thing but are worded differently.