Introduction

CPCS provides legal representation in Massachusetts for those unable to afford an attorney in all matters in which the law requires the appointment of counsel. This includes representation in criminal, delinquency, youthful offender, child welfare, mental health, sexually dangerous person and sex offender registry cases, as well as related appeals and post-conviction matters. The present project focuses on young people (up to age 18) accused of committing various crimes.

This cohort is represented by CPCS' Youth Advocacy Division (YAD) who work within the Massachusetts juvenile justice system. YAD represents court-involved youth and partners with the Education Law Project (EdLaw) practice that focuses on youth entangled in the school-based discipline.

Our goal is to help the YAD to analyze their clients' attributes, such as the distribution of their race, gender, home address. Also, we combine the census tract data to analyze the client's living environment. Finally, we help YAD to analyze whether their social services would help their clients to get a better outcome.

Strategic questions:

1. Are there disproportionate charge-types that vary based on various geographic areas (county, police department, school district, court area, etc.)? Are there racial or other biases which correlate with various charge types or disproportionate levels of specific charges? Is there a correlation of court involvement with demographic or sociological indicators such as income, housing density, schooling, environmental justice, demographics? Possible regression analysis.
2. What are prominent topics surrounding engagement of social service advocates provided?
3. What are success rates for social services provided in the course of a case? What are traits of cases with successful social service interventions? Identify key features here.

Data elements:

Our dataset includes:

1. Clients' demographic data.

2. Clients' cases data, including charge level and dispositions.
3. Cases runsheet including very detailed notes, which is the log between lawyers and clients.

Additional information:

1. Census tract data in Massachusetts.
2. School information in Massachusetts.
3. Court information in Massachusetts.
4. YAD office locations

Terminologies

Charge level

Each case depends on what the client did, lawyers would predict one or more charge levels for the case, and we use the maximum charge level of each case. The charge separates into 6 levels, from 1 to 6, as higher the level is, the case would be more serious.

Disposition level

Disposition is how the case was settled, and there are many different special types of disposition. According to the disposition serious level, we categorize them into 5 levels, from 1 to 5, as higher the level is, the case would be more serious.

5 levels are: 1 - No CORI; 2 - No admission/finding of delinquency or YO; 3 - Admission/finding of delinquency or YO stayed in community; 4 - Admission/finding of delinquency or YO removed from community; 5 - Adult imprisonment imposed.

Social services

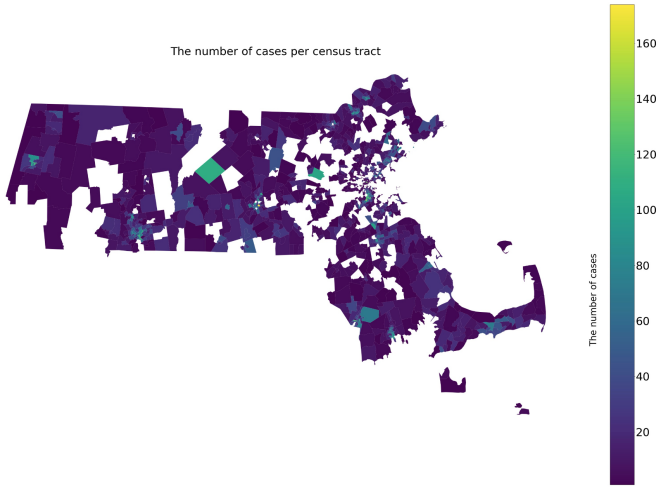
CPCS help clients find social services based on their interests, such as sports teams, music clubs, etc, in order to help them participate in their community and take advantage of their own life. In addition, social services could show more positive sides of clients, which could lead to better possible legal dispositions.

Detention

Detention means relatively severe disposition. We also split cases by yes/no detention to do analysis.

Distribution of clients in Massachusetts

The picture on the right shows how many cases there are in each census tract area in Massachusetts. The darker the color in the picture indicates the fewer cases in the area, and the lighter color indicates the more cases in the area. White indicates that there is no case in this area in the database. According to statistics, the area with the most cases is Census Tract 7327, Worcester County, Massachusetts, with 174 cases. On average, there are about 17 cases in each tract region of Massachusetts.



It can be seen from the figure that there are many cases in the central-eastern coastal areas (need to zoom in to see clearly). They are small in area, but the average number of cases in each area is higher than that in most parts of Massachusetts.

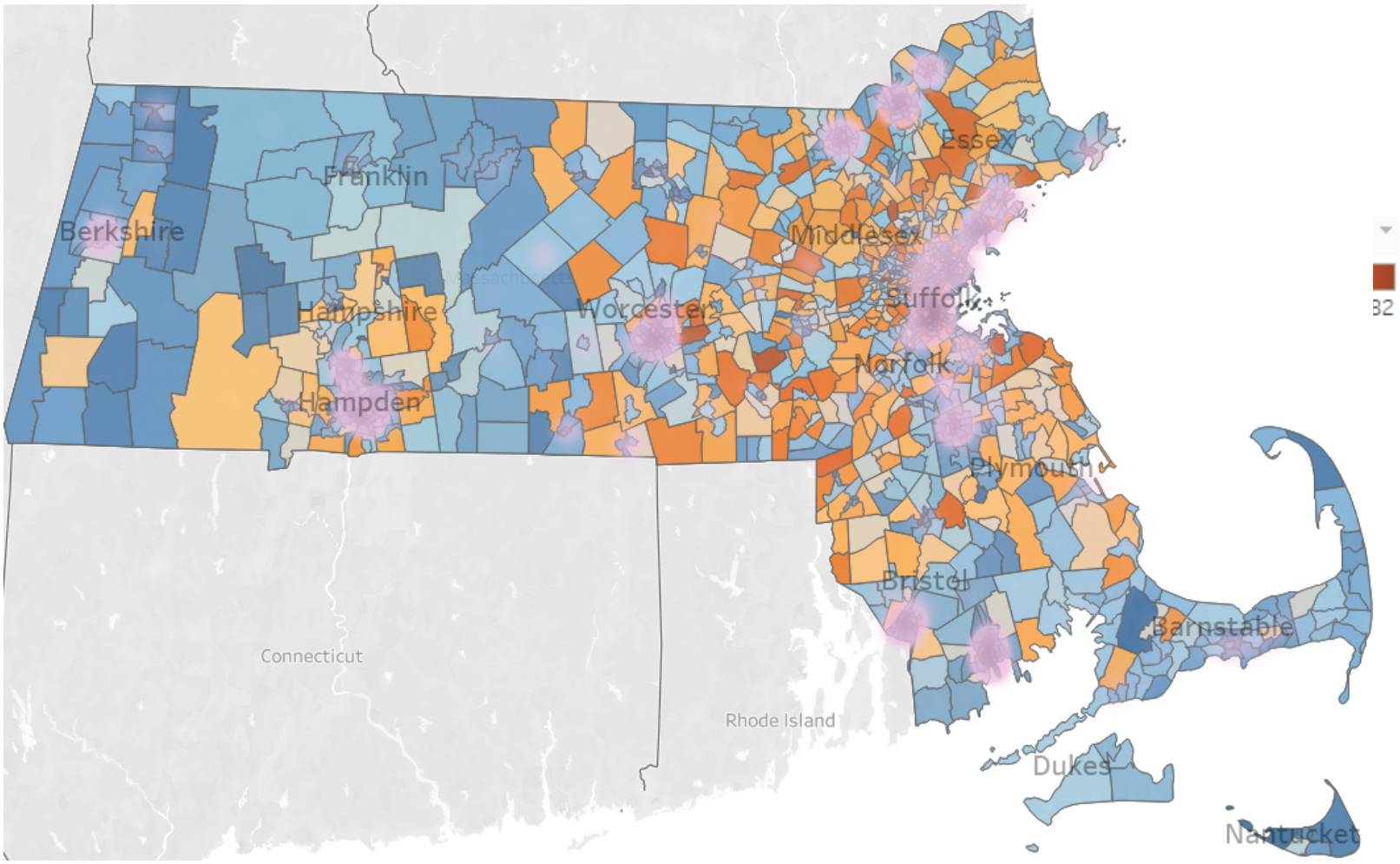
Distribution of Cases with Education Level in Massachusetts

Education level is defined as the sum of the number of people who are:

1. 18 to 24 years with Bachelor's degree or higher
2. 18 to 24 years with Some college or associate's degree
3. 25 years and over with Associate's degree
4. 25 years and over with Bachelor's degree
5. 25 years and over with Bachelor's degree or higher
6. 25 years and over with Graduate or professional degree
7. 25 years and over with High school graduate or higher
8. 25 years and over with Some college, no degree

The map at the bottom shows the distribution of education level, where the darker blue denotes that the education level of this tract is lower, while the darker red denotes that the education level of this tract is higher. The purple bubbles on the map denote that there are some case records.

If we mute the purple bubbles in Tableau, we can see that most of the cases happened in certain tracts that have lower education levels. Maybe CPCS can pay more attention to these tracts. This analysis is related to the first strategic question.

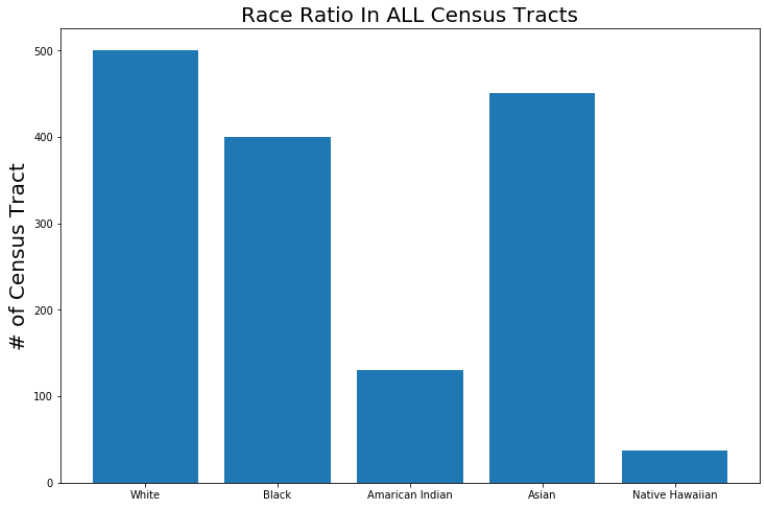

Essex
Franklin
Berkshire
Middlesex
Suffolk
Hampshire
Worcester
Norfolk
Hampden
Plymouth
Bristol
Barnstable
Connecticut
Rhode Island
Dukes
Nantucket
32

Concentrated Census Tract

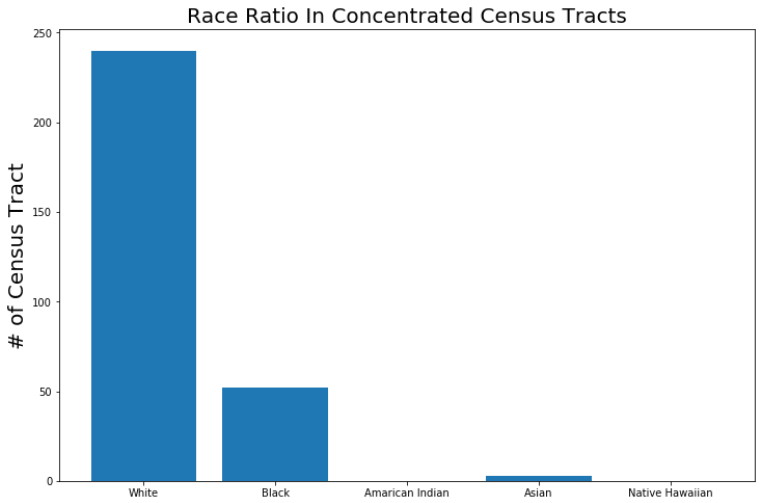
We decided to select the top 20% of the most concentrated census tracts as our analysis dataset, which covers about 60% of all cases. We choose 20% as the representative because it covers a fair number of cases in a fairly small number of census tracts.

1. **Race**

The following graphs are race ratio in all census tracts and race ratio in concentrated census tracts. We consider 5 races in a census tract. We label a census tract by the dominant race living in that census tract. The following bar graphs represent the number of census tracts of five different races.
NOTE: this race information of census tract is different from the clients' race information. The census tract information only provides the below five different races, whereas CPCS also provides races like Hispanic white and Hispanic black, which is not provided in census tract.



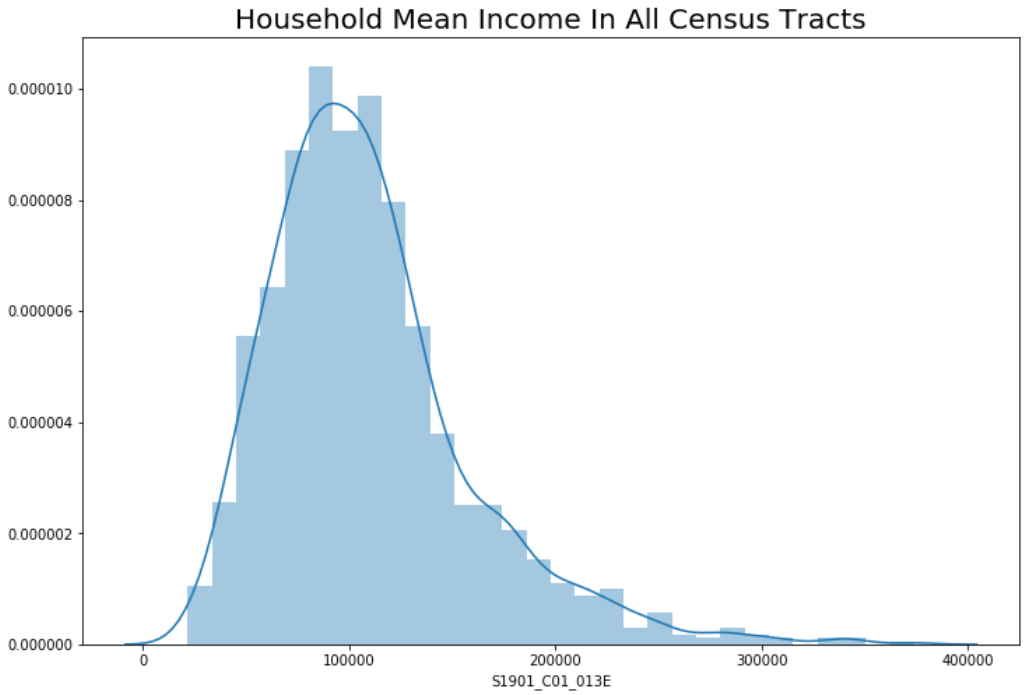
On the above graph, we can see that the white, black, and Asian census tracts are the most ones.



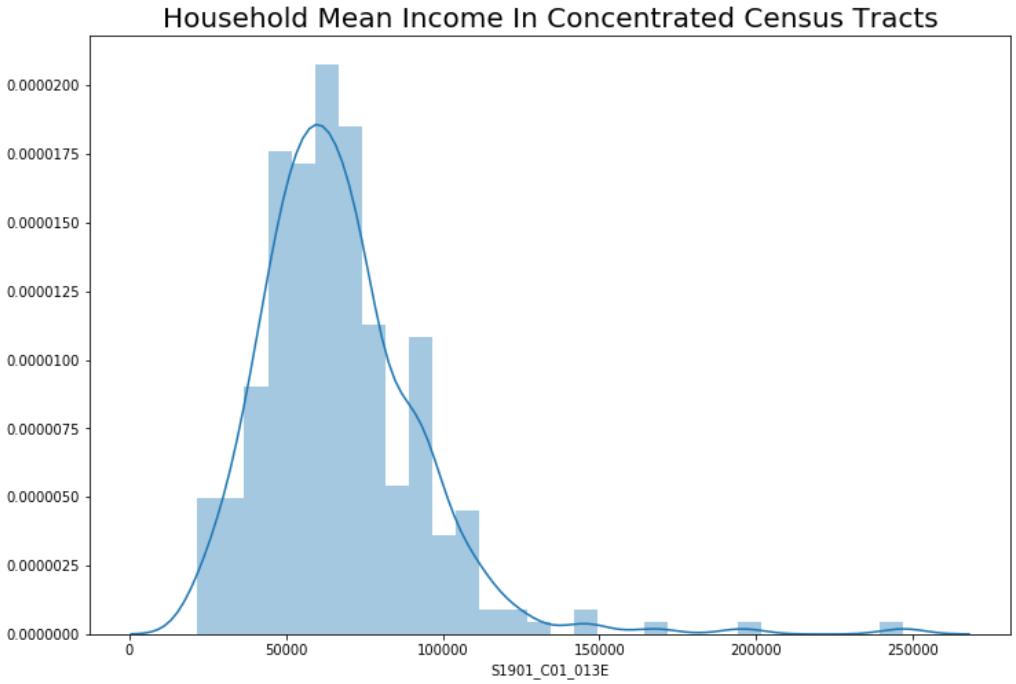
On the above graph, which is only showing the concentrated census tracts. We can see that the census tracts with the most cases are white and black census tracts, but white census tract has a higher ratio than the previous graph.

2. Income

The following graph is the household mean income in all census tracts. We can see the most samples clusters around 100000.



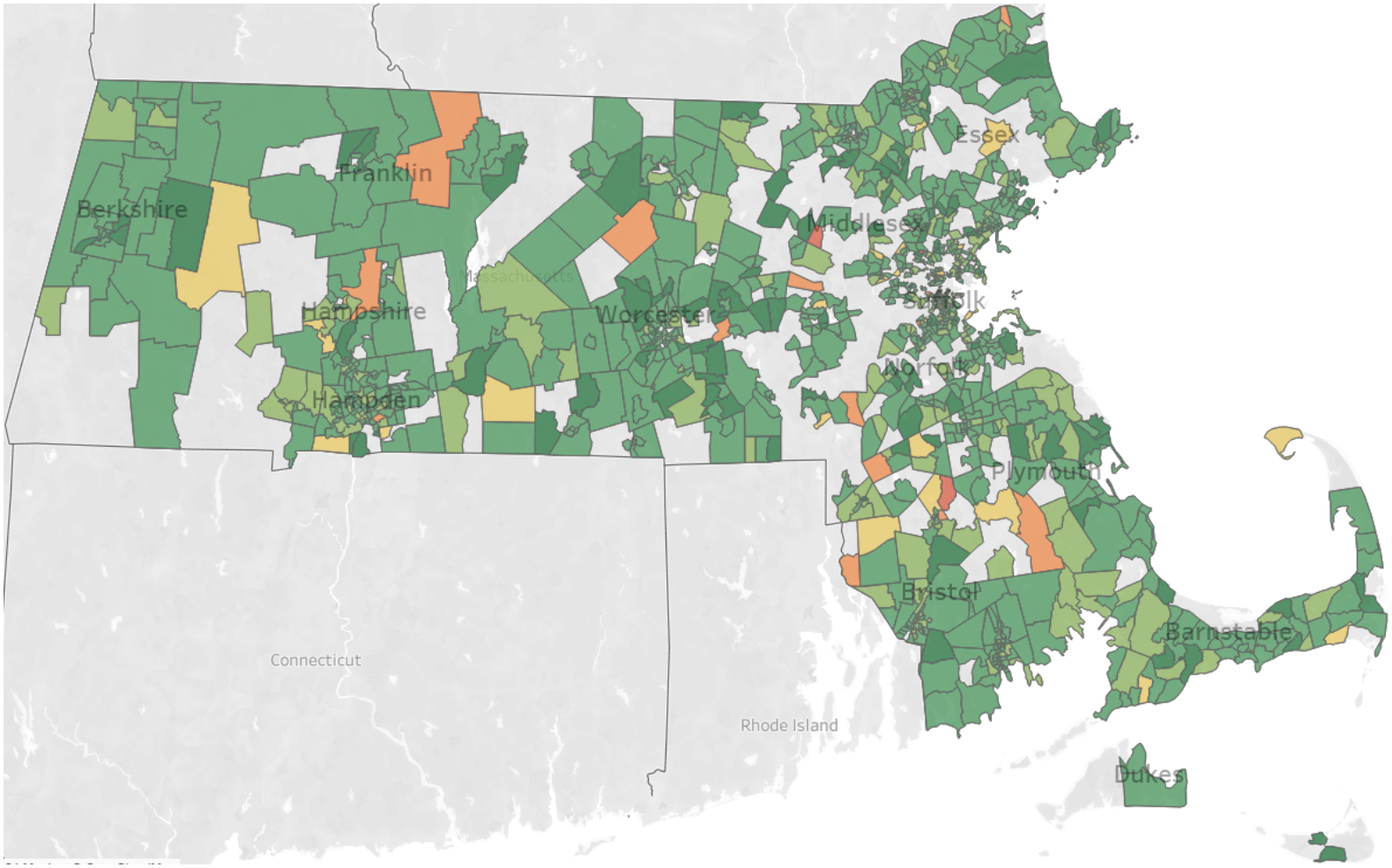
The following graph is the household mean income in concentrated census tracts.



We can see that both graphs share the same distribution, but there is a shift to the left on the household income in concentrated census tracts, which means that the concentrated census tracts tend to have a lower mean household income than the all census tracts. This analysis is related to the first strategic question.

Charge grid level distribution by census tract area

The average charge grid level for each census tract area is the mean of maximum charge grid leves of all cases that happened there. The darker green denotes that the average charge grid level in this tract is lower/less severe, while the darker red denotes that the average charge grid level in this tract is higher/more severe.

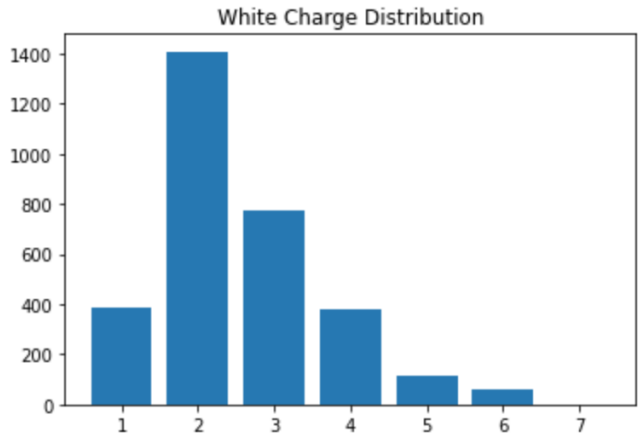


Charge level distribution

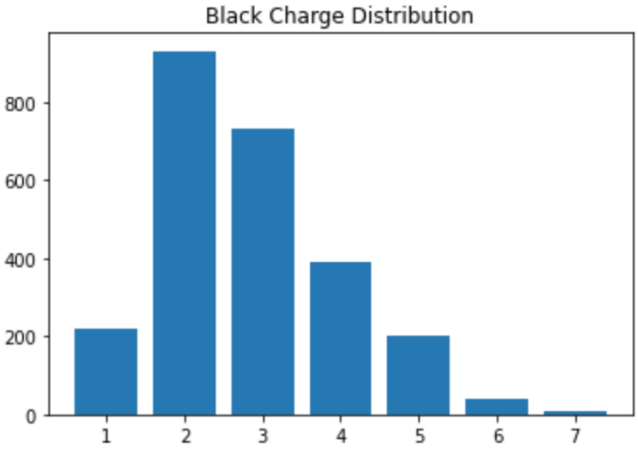
Draw the distribution per race

The vertical axis in each figure is the number of clients at a certain charge level, and the horizontal axis represents the charge grid level. If one client or one case has multiple charge levels, we use the maximum one.

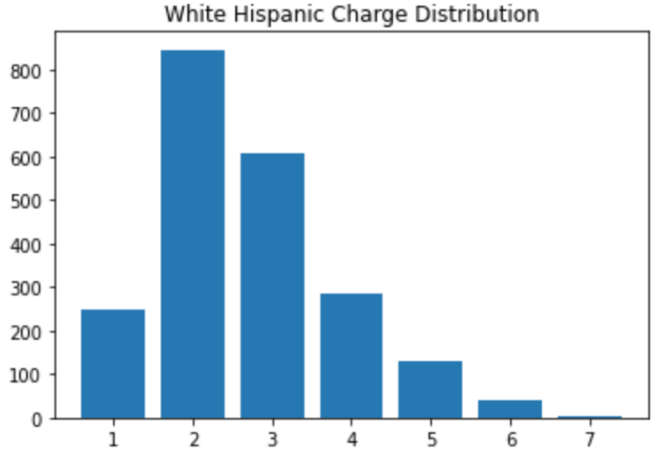
White: Total: 3129 Average: 2.557 Distribution: [388, 1408, 774, 382, 116, 59, 2]



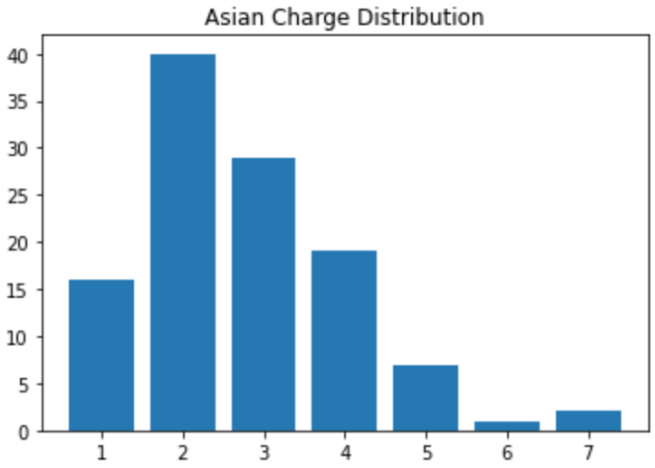
Black: Total:2522 Average: 2.834 Distribution: [220, 931, 731, 389, 200, 41, 10]



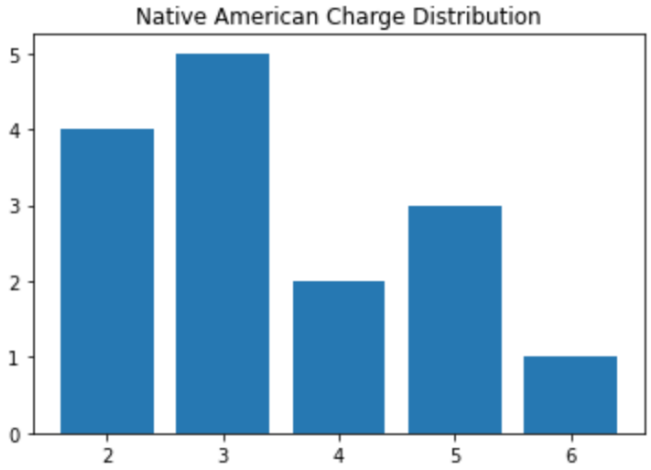
White Hispanic: Total:2162 Average: 2.696 Distribution: [247, 846, 609, 285, 130, 42, 3]



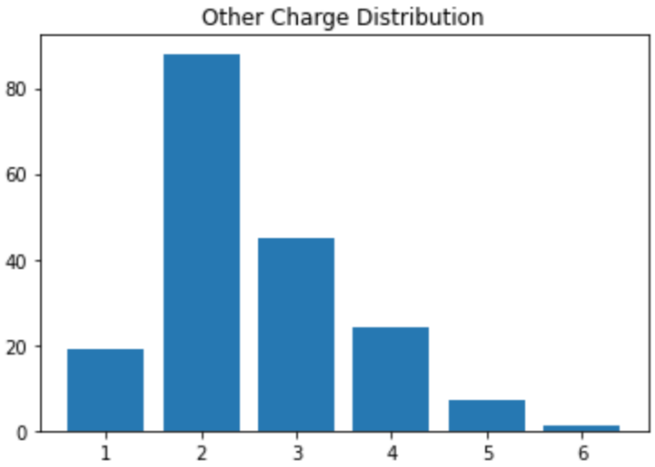
Asian: Total:114 Average: 2.754 Distribution: [16, 40, 29, 19, 7, 1, 2]



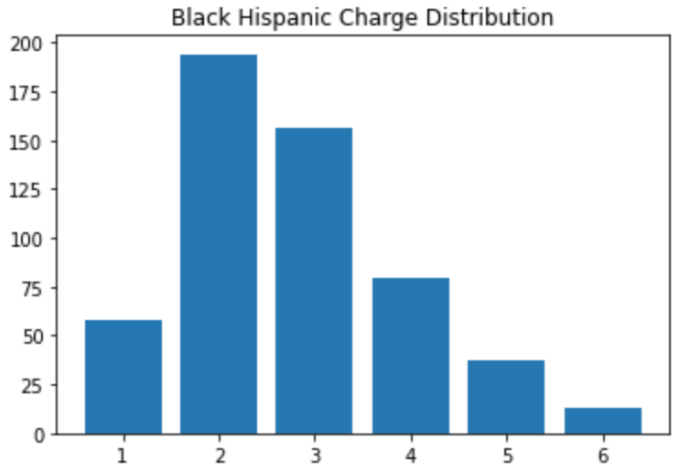
Native American:Total:15 Average: 3.467 Distribution: [0, 4, 5, 2, 3, 1, 0]



Other: Total:184 Average: 2.538 Distribution: [19, 88, 45, 24, 7, 1, 0]



Black Hispanic: Total:537 Average: 2.780 Distribution: [58, 194, 156, 79, 37, 13, 0]



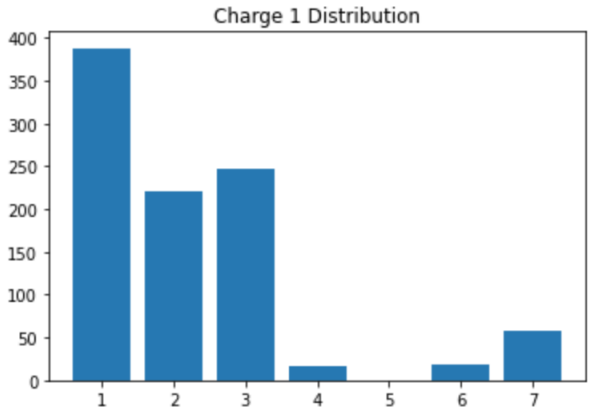
Summary

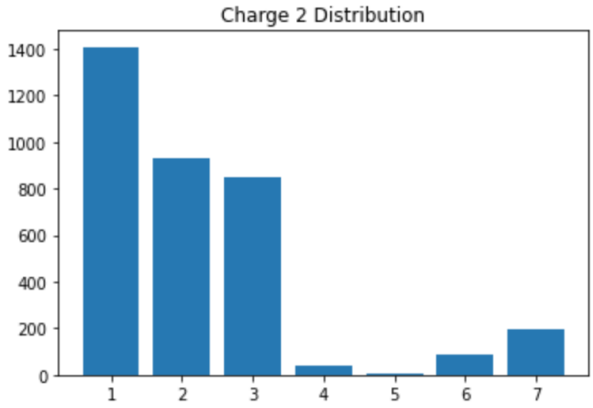
By drawing the charge grid level distribution of each race, through observation, we can conclude that the charge grid level distribution between different races is not obviously related to race.

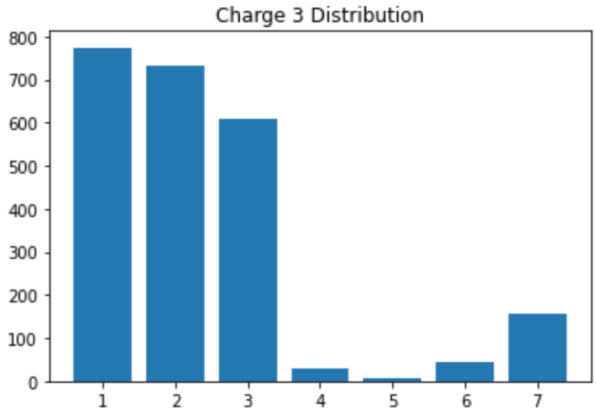
Among the 5 race clients, the lowest average charge grid level is the White client, with an average of 2.557; the highest average charge grid level is the Native American client, with an average of 3.467, but there are only 15 Native American clients in the database, so we think this average value for the Native American may be not persuasive; secondly, the Black clients have the second highest charge grid level, with an average of 2.834.

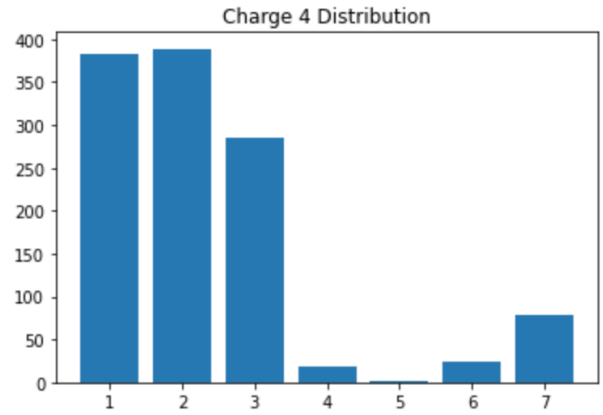
Draw the distribution per charge level

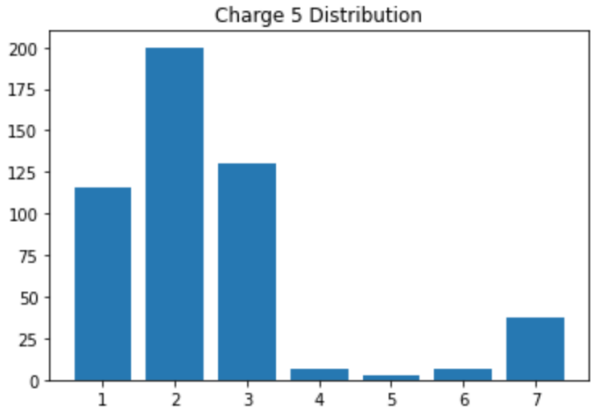
The following 6 histograms show the client distribution of each race in the certain charge level. The horizontal axis in each histogram represents the 7 race types, from left to right are 1. White, 2. Black, 3. White Hispanic, 4. Asian, 5. Native American, 6. Other, 7. Black Hispanic; the vertical axis represents the number of clients of a certain race in the certain charge level.

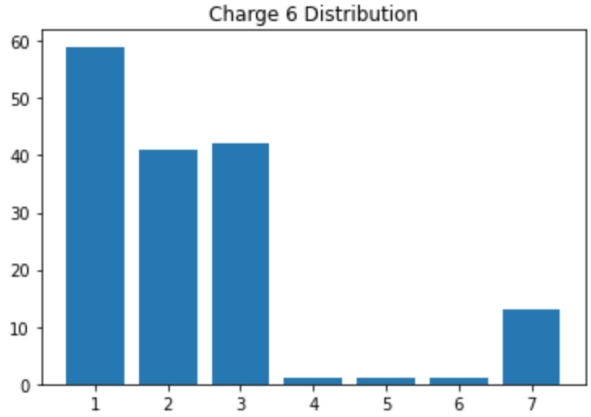




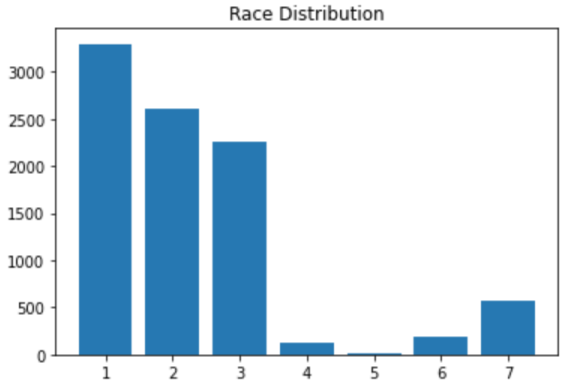








Summary



The histogram on the right shows the race distribution of clients.The horizontal axis represents the 7 race types, from left to right are 1. White, 2. Black, 3. White Hispanic, 4. Asian, 5. Native American, 6. Other, 7. Black Hispanic; the vertical axis represents the number of clients of a certain race.

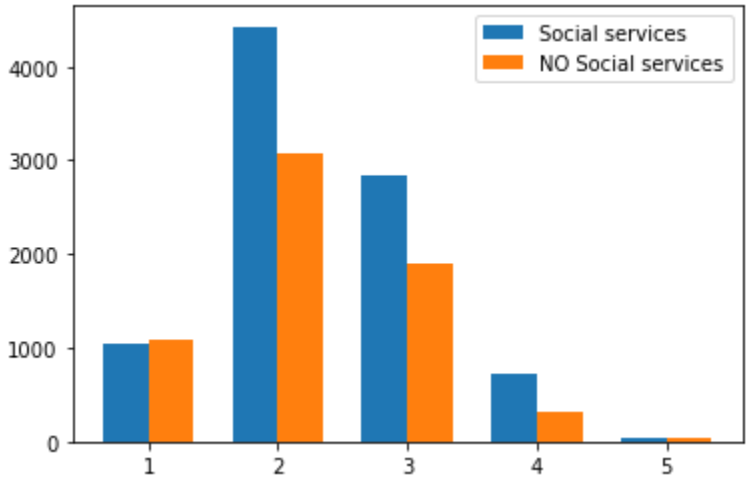
We can observe that the race distribution when the charge level is 3 and 4 is similar to the total client race distribution. By comparing with the total client race distribution, the histogram when charge level=5 shows that the proportion of Black clients is larger and the proportion of White clients is smaller; when charge level=1 and charge level=6, the proportion of Black clients is smaller. This analysis is related to the first strategic question.

Social Services and Dispositions

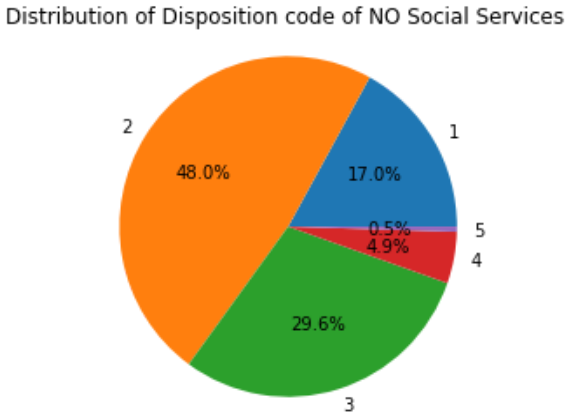
Here we tried to analyze whether the social services provided to the clients can produce better legal results on clients' cases. This analysis is helpful to answer the third strategic question.

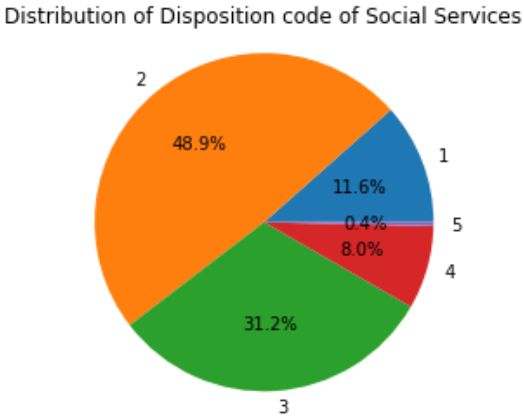
First, the following is the distribution of dispositions over all clients. The x-axis represents disposition code 1-5, and 0 means null or n/a. The y-axis represents the number of clients.

Clients receiving social services. **Distribution**: [3569, 1845, 9252, 5474, 1327, 58] **Total**: 21525
Clients NOT receiving social services. **Distribution**: [6053, 1771, 6100, 3487, 556, 44] **Total**: 18011



The following pie graphs show the proportion of each disposition code.



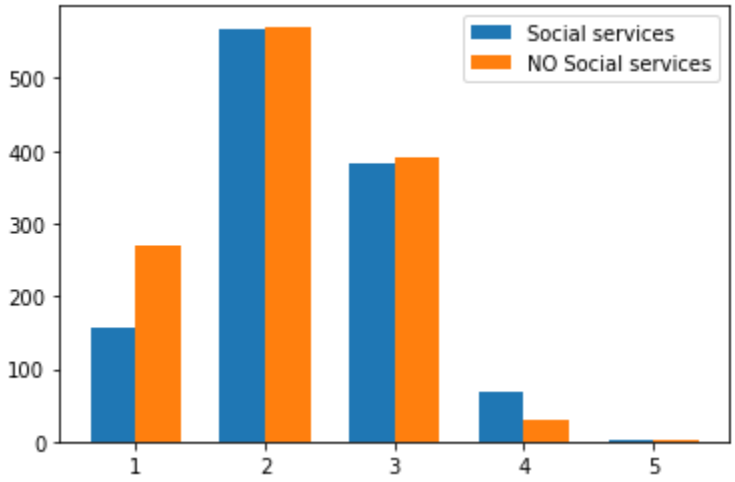


Second, we try to figure out whether there is a difference in distributions between different charge levels that clients received. The following tables are separated by seven charge levels 1-7, and for

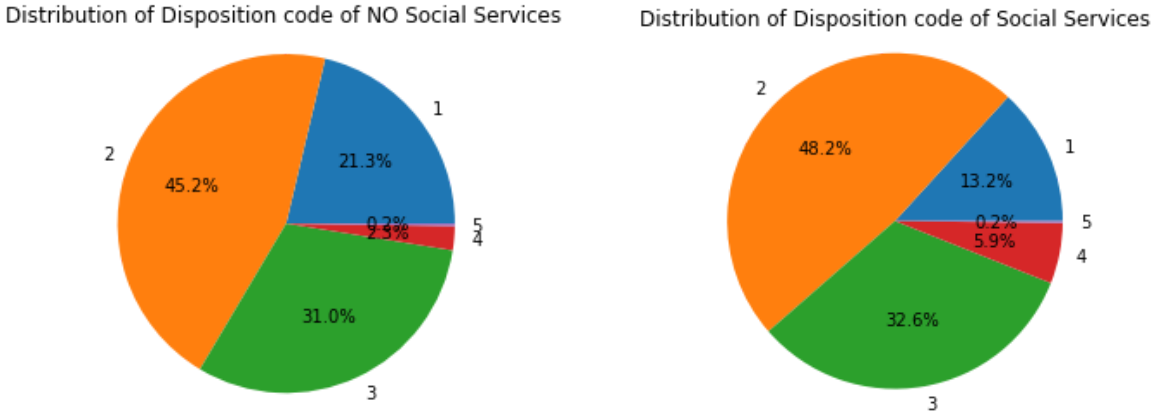
each charge level, the distributions of receiving social services and not receiving social services are shown.

For charge level one:

Clients receiving social services. **Distribution**: [523, 368, 1597, 1055, 234, 6] **Total**: 3783
Clients NOT receiving social services. **Distribution**: [969, 496, 1273, 894, 73, 8] **Total**: 3713

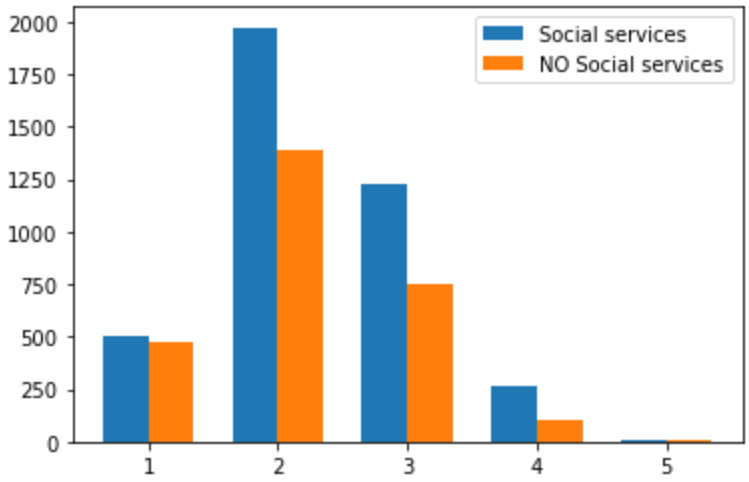


The following pie graphs show the proportion of each disposition code.

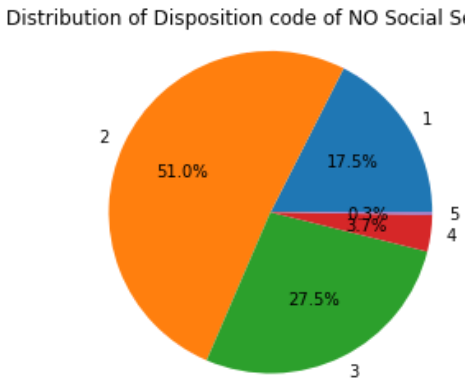


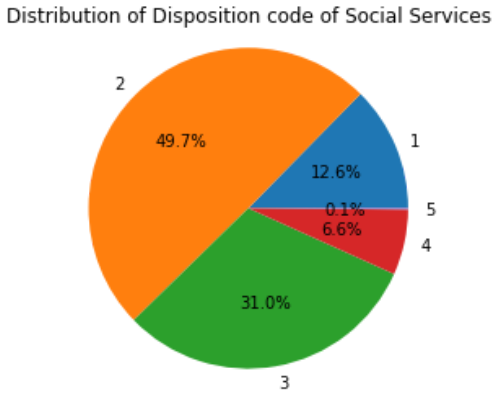
For charge level two:

Clients receiving social services. **Distribution**: [1156, 808, 3707, 2315, 491, 13] **Total**: 8490
Clients NOT receiving social services. **Distribution**: [2229, 717, 2487, 1414, 203, 13] **Total**: 7063



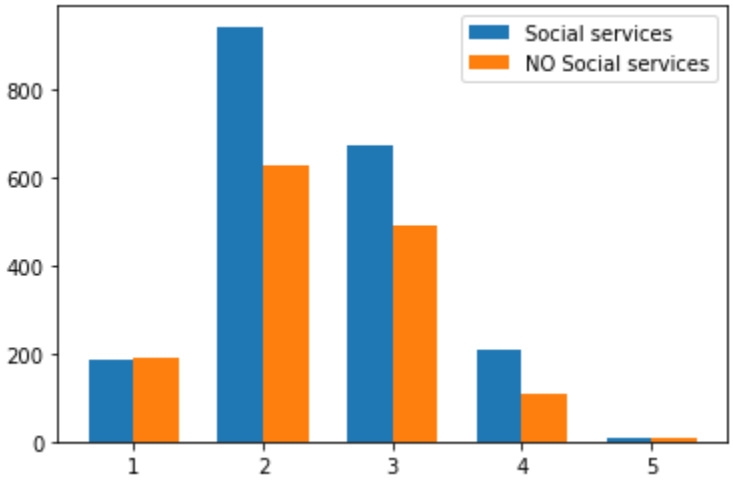
The following pie graphs show the proportion of each disposition code.



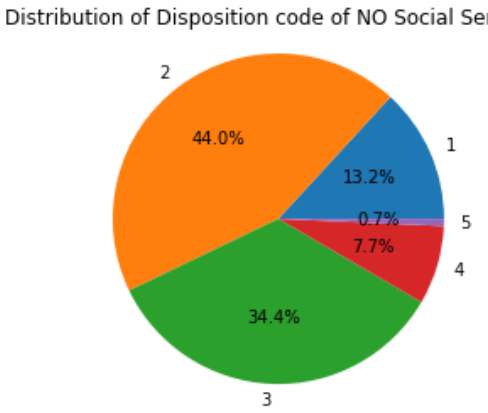


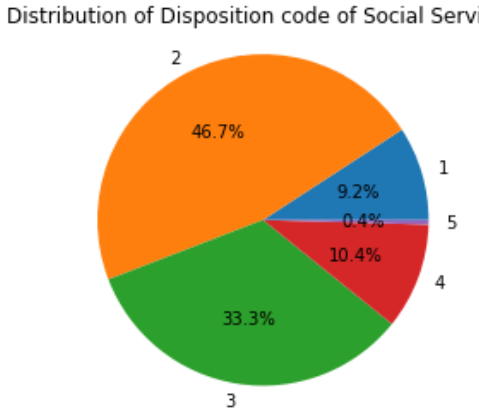
For charge level three:

Clients receiving social services. **Distribution**: [527, 250, 1591, 954, 272, 15] **Total**: 360
Clients NOT receiving social services. **Distribution**: [1081, 233, 943, 631, 133, 11] **Total**: 3032



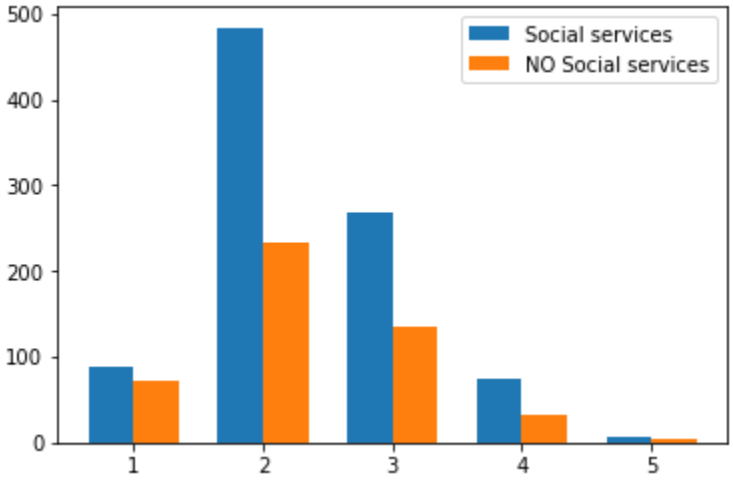
The following pie graphs show the proportion of each disposition code.



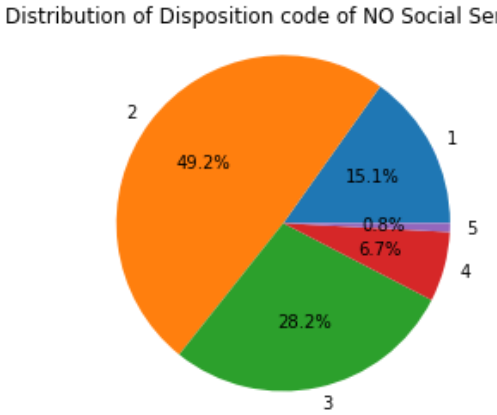


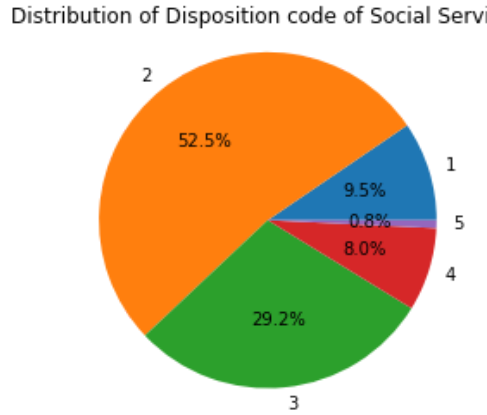
For charge level four:

Clients receiving social services. **Distribution**: [197, 106, 744, 346, 105, 7] **Total**: 1505
Clients NOT receiving social services. **Distribution**: [344, 78, 318, 159, 42, 5] **Total**: 946



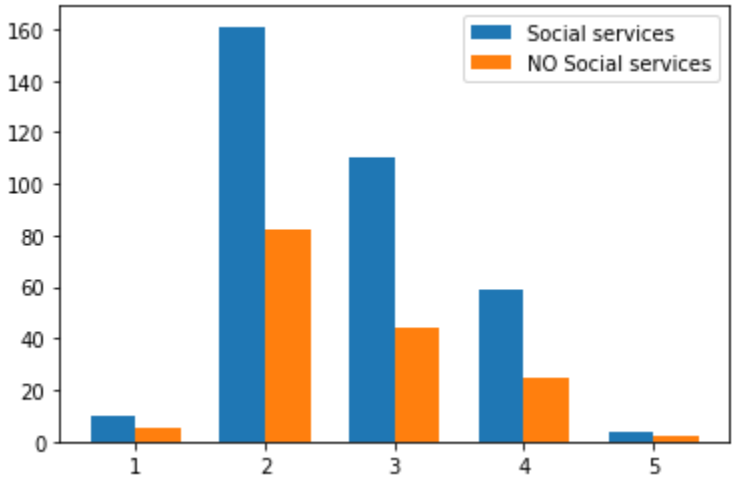
The following pie graphs show the proportion of each disposition code.



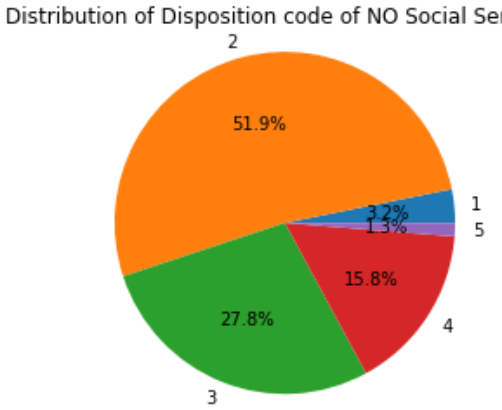


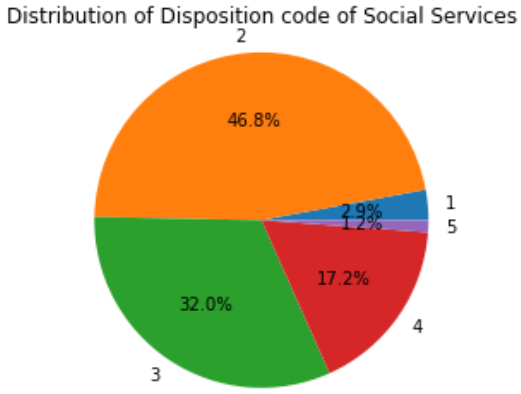
For charge level five:

Clients receiving social services. **Distribution**: [108, 11, 272, 141, 60, 4] **Total**: 596
Clients NOT receiving social services. **Distribution**: [215, 7, 114, 44, 24, 2] **Total**: 406



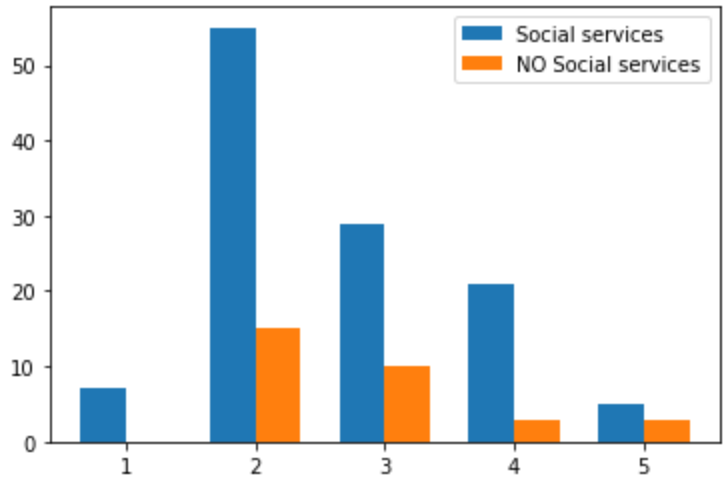
The following pie graphs show the proportion of each disposition code.



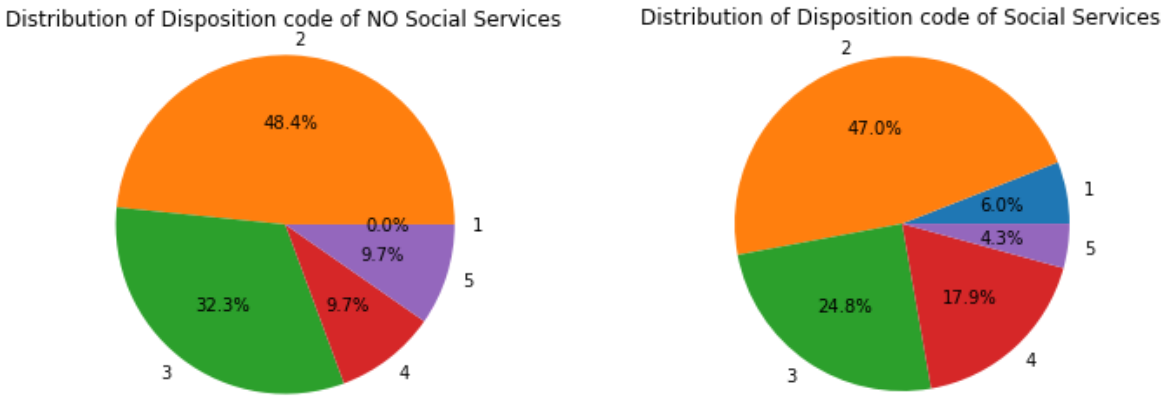


For charge level six:

Clients receiving social services. **Distribution**: [33, 9, 111, 50, 21, 6] **Total**: 230
Clients NOT receiving social services. **Distribution**: [49, 1, 27, 17, 1, 2] **Total**: 97



The following pie graphs show the proportion of each disposition code.



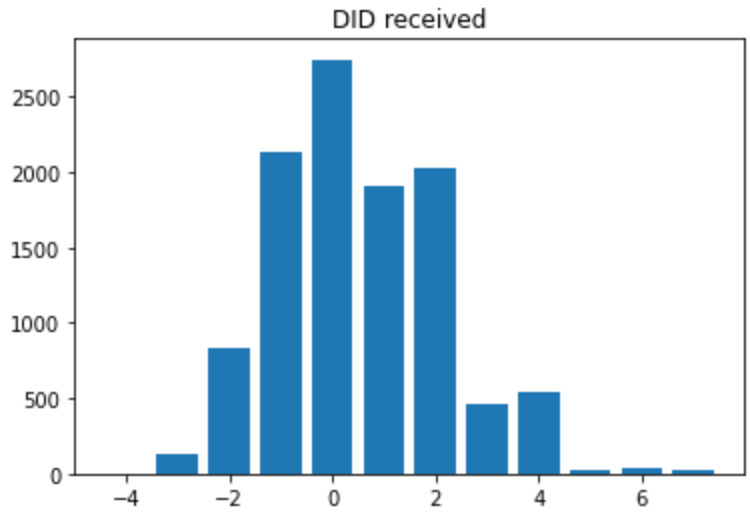
Based on the pie graphs above, we can see that the relation between whether clients participate in social services or not and clients' legal disposition is poor. Therefore, dispositions might not be a good way to measure social services.

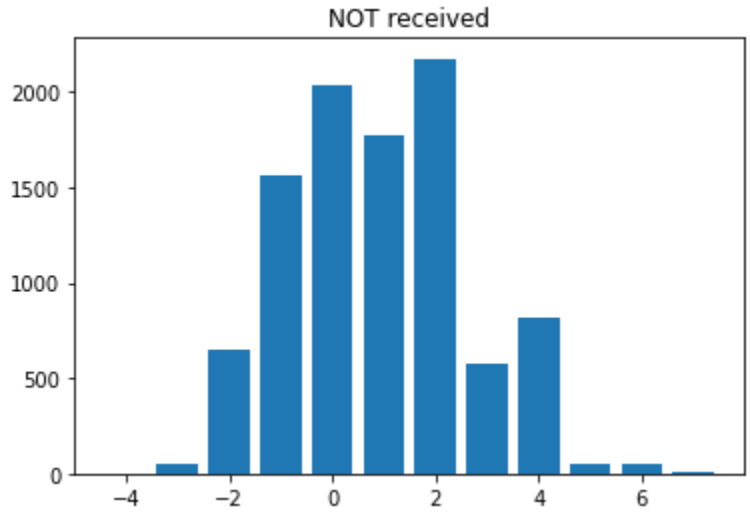
Analysis of the difference between charge level and disposition code

For a certain, we believe, the higher the charge level, the severer the case, and the lower the disposition outcome, the better the lawyer service. Thus, beside the disposition code itself, we then define another metric which equals charge level minus disposition code. This new metric can well measure how a case is improved by the lawyer service.

Given the finding before, the proportions of different charge levels are almost the same in cases with social service and those without. The same thing also happens when it comes to the proportions of different disposition codes. However, when we do a minus between these two, things change.

Here is the difference distribution of two types:





While the graphs seem similar, the difference is cases not received social services have an apparently higher proportion in difference 2.
Moreover, as we compute the averages of the differences, we have the following table:

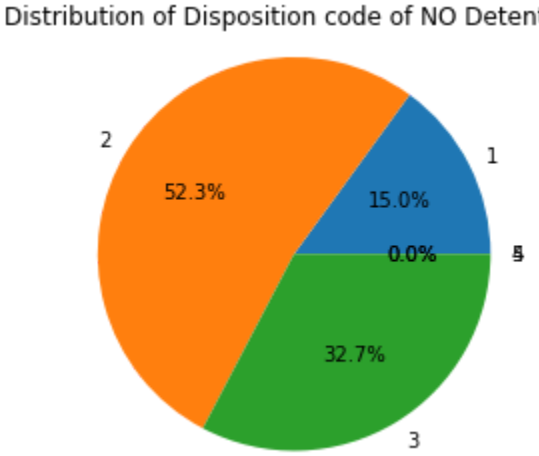
	Received	Not Received
Average of Difference between charge level and disposition code	0.5349	0.8904

The result indicates literally the cases without social services have a better improvement from their original charge, however, this does not mean social services are

useless and not worthy because there can be a natural difference between the people who choose and those who do not choose social services.
Perhaps those who choose social services think that they cannot overturn their prosecution, which means that the cases of those people are inherently more difficult than those of others, so this can explain part of this phenomenon. This analysis is related to the third strategic question.

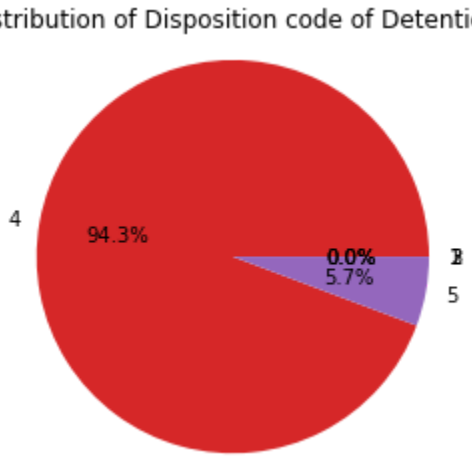
Disposition and detention cohorts

The following pie graphs show the proportion of each disposition code WITHOUT detention.



Here only disposition 1, 2, and 3, which are lighter disposition outcomes, are associated with NO detention.

The following pie graphs show the proportion of each disposition code WITH detention.

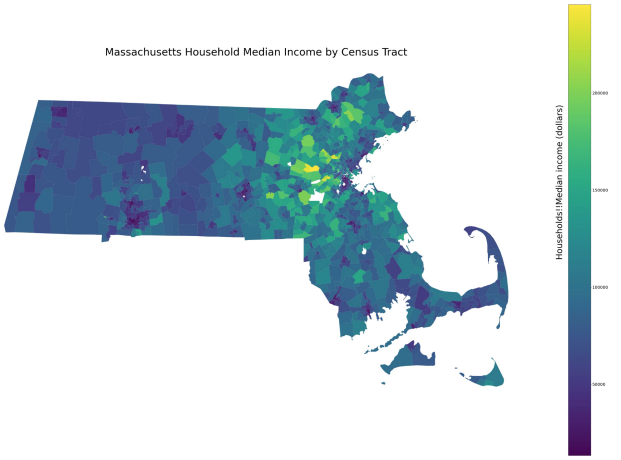


Here only disposition 4 and 5, which are heavier disposition outcomes, are associated with detention.

Analysis Overall Census Tract:

1. Income

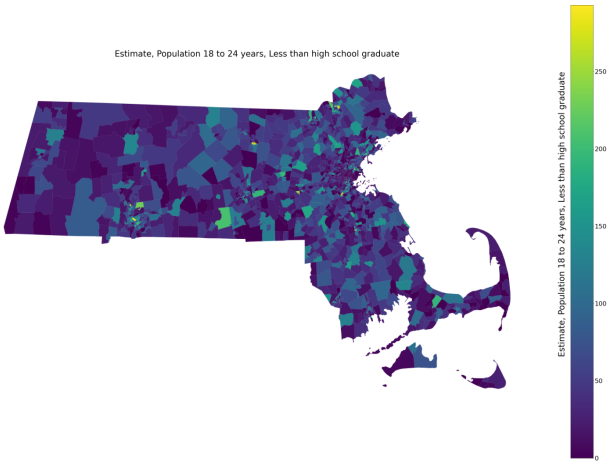
The picture on the right is the heatmap of Massachusetts's household's median income in 2019. The darker the color, the lower the median income of the area; on the contrary, the lighter the color, the higher the median income of the area.



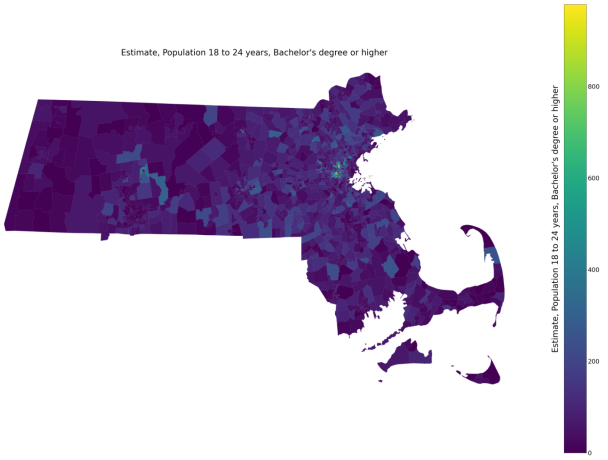
It can be seen from the figure that the household median income of eastern Massachusetts is relatively higher than that of the central and western regions.

2. Education Attainment

Population 18 to 24 years old that have attained less than high school graduates in Massachusetts.

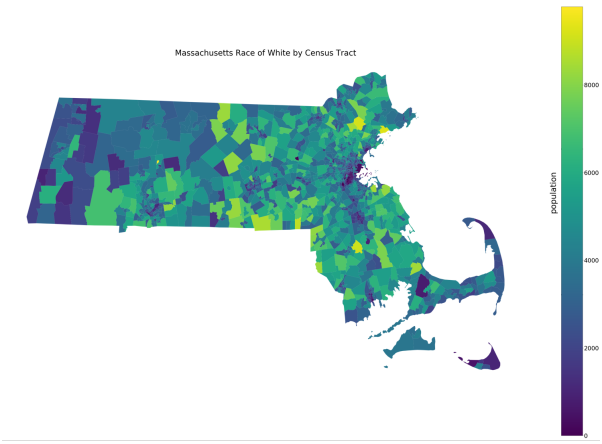


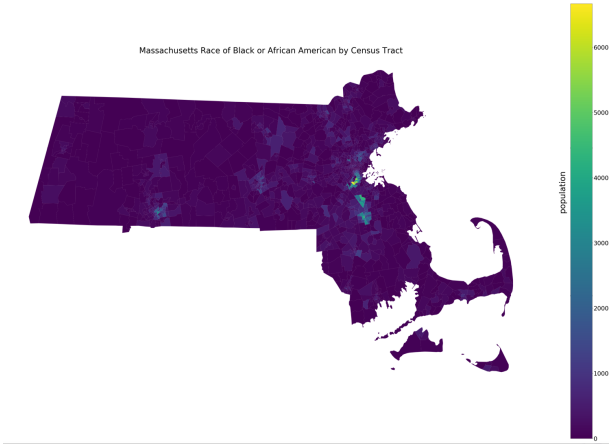
Population 18 to 24 years old that have attained Bachelor's degree or higher in Massachusetts.

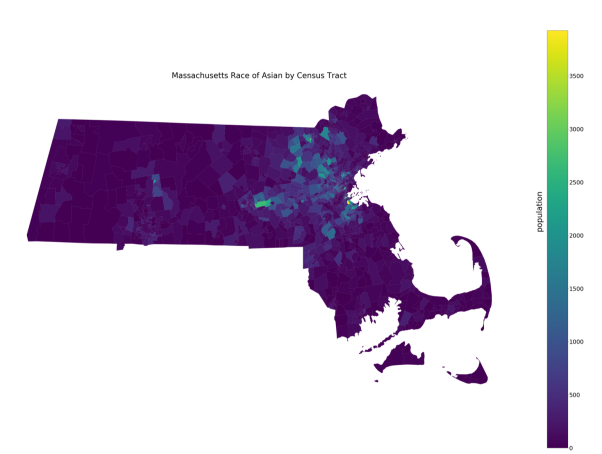


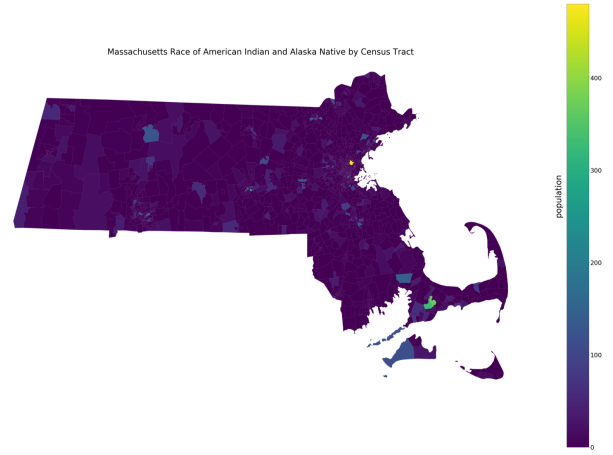
3. Race

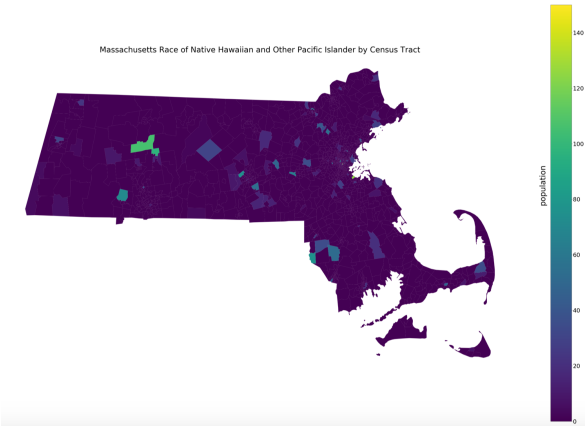
The following graphs are distributions of 5 different races in all census tracts. The lighter the color of a census tract, the more the population of that race in that census tract. The deeper the color of a census tract, the less the population of that race in that census tract.











Runsheet Note Analysis:

1. Keywords analysis using TF-IDF

We split all the notes from the runsheets into two parts, which are clients that received social services and those that did not. Then TF-IDF is deployed to extract keywords on these two datasets, respectively.
The results are shown below:

Received unique	going, home, know
Not received unique	client, judge, mother
Words in common	ada, also, asked, back, call, called, case, court, date, get, mom, said, school, she, spoke, told, would

Through the table, we find the notes of both types of clients are really similar. While there is a difference in the keywords extracted, the difference is too opaque for us to infer more information.
Nevertheless, an interesting finding is that these notes always mention females, such as she, mom, and mother, without any male-related nouns or pronouns such as he, dad, or father.
The above information indicates that clients are likely to be mostly female or clients are always in contact with females.However, after statistics, we found that only 5,745 of the 25,708 clients were women, which overturned our previous conjecture. In this case, as

both datasets have “mom” as the keyword, we believe mother of the client plays an important role in the case.

2. Sentiment analysis

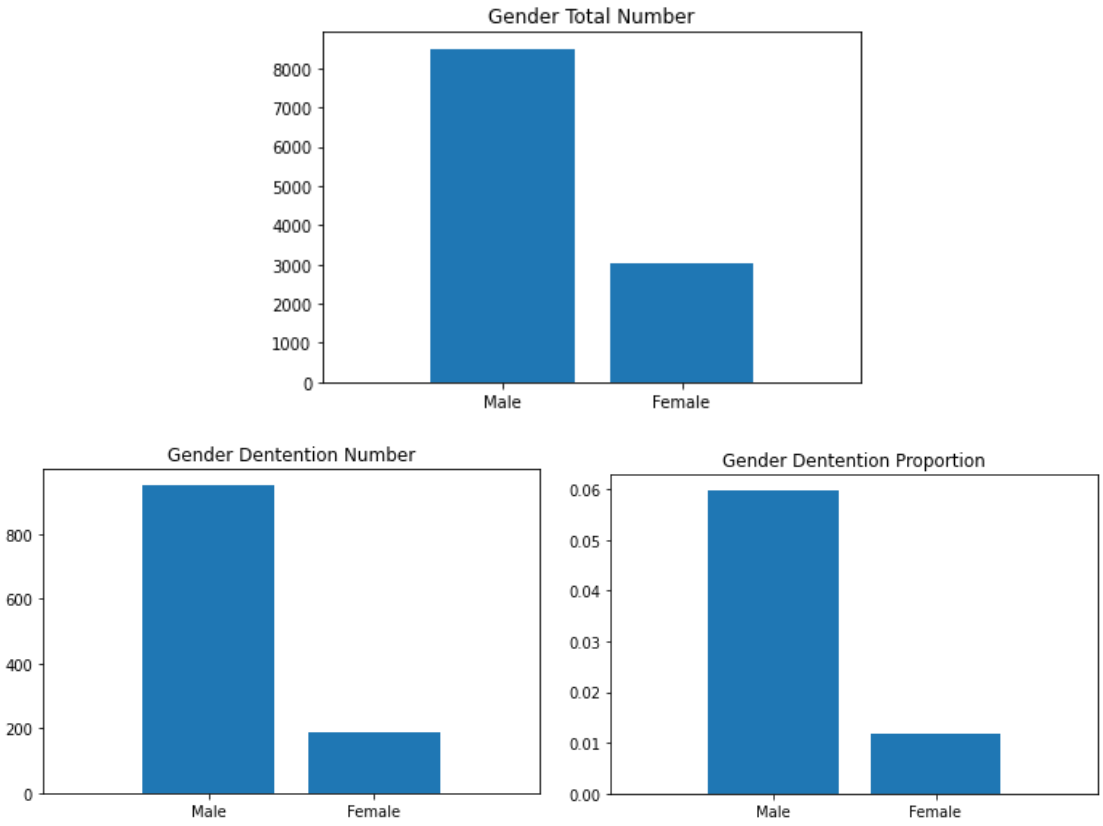
We also analyzed the sentiment of the datasets above, however, the datasets show same sentiment results as follow:

	Negative	Neural	Positive	Compound
Received	0.058	0.83	0.112	0.318
Not received	0.058	0.83	0.112	0.318

We speculate that this phenomenon is because these notes are written by government officials, so the content is neutral and objective.

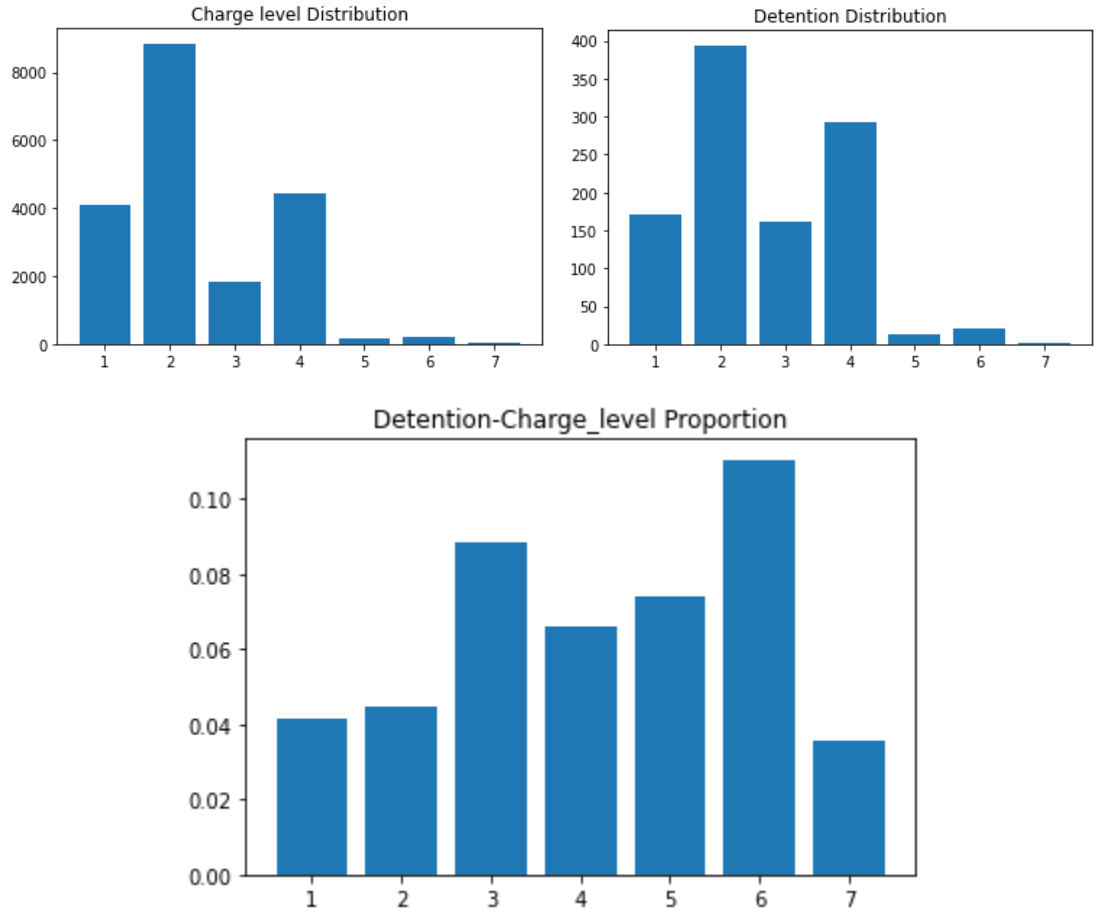
Correlation with detention

Correlation between gender and detention



These figures show that both the number of male cases and the proportion of male detention greatly exceeds female's data.

Correlation between charge level and detention



Generally, the higher the charge level, the higher the detention rate.
However, because we have too few charge level 7 cases, the detention rate of charge level 7 is accidently lower than that of either other charge level.
Another cause may explain the irregular graph that is not strictly a "the higher the level, the more the proportion" case, is when we deal with cases which have multiple charge levels, we always choose the maximum, exaggerating the severity to some extent. This analysis is related to the first strategic question.

Limitations

1. Because of the high confidentiality of the CPCS data, there are many legal barriers slowing down our progress in the first month.
2. The data we got at first was incomplete and poorly structured. We spent a fair amount of time cleaning the data, but most of it was abandoned after we received the second batch of the data.
3. CPCS data only contains the client's address, race and gender, which leads to limited perspectives when analyzing the client. Analysis involving the clients' family income and education level of clients and their family members could have been done if related clients' data was provided.
4. The success rate of social services is unavailable and unpredictable because of the lack of information and the difficulty of defining success. There are many perspectives of measuring a social service: the engagement level of clients, the duration of a social service, the effect of a social service to a community, the final legal disposition of clients, etc. The only information we have is the legal disposition of clients, and based on our analysis, the relation between whether clients participate in social service or not and clients' legal dispositions is very poor, which means that the dispositions might not be a good measuring bar for social services.
5. We lack good quantifying methods to quantify the qualitative data, and this could be a good perspective to think about for the continuation of this project in the future.

Data & Code

Since our data is confidential, we cannot upload our data to the Course Github repo, and our codes would contain some confidential data. All of our data and codes are saved in this google drive folder that can be accessed by our clients and some Spark! staff. If you need access to this folder, please contact us!