Cradle to Prison Pipeline
Team 1 -
Olivia Bene
Parth Ghayal
Gia Bach Nguyen
Karan Vyas
CS506

Cradle to Prison Pipeline Final Report

Background + Motivation

The Cradle to Prison Pipeline (C2P) project is the work of a combination of the Center for Public Interest Advocacy and Collaboration (CPIAC) at Northeastern's School of Law, Northeastern's College of Art, Media + Design (CAMD), Department of Sociology and Anthropology at the College of Social Science and Humanities at Northeastern (CSSH) and Boston Area Research Initiative (BARI). The project is working to solve the entrenched problem of mass incarceration. To do this, the collective will identify the key factors and underlying contributors that led to incarceration. Identifying these causes along the pipeline means being able to identify them in real-time in today's youth and intervene.

Being based at Northeastern University, this is currently a Massachusetts-focused project and uses data collected from various prisons in the state. However, this is being developed as a framework for other states to follow. C2P is meant to shine a light on the areas of the state that have been left behind academically and/or geographically. By highlighting the areas that are struggling most, C2P can get resources allocated to those who need them most and reduce mass incarceration.

Previous Work

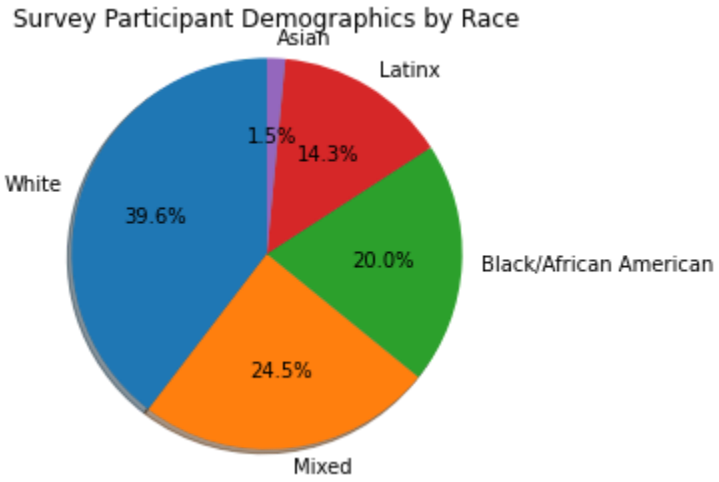
We are the second group working with the survey data in this project. Over the summer, a group of students translated two hundred handwritten surveys from country jails into usable formats. After that, about 300 surveys were scanned into PDFs and the data was extracted into CSV format with AWS Textract and OpenCV.

The data needs to be cleaned and organized to make better estimates. This can be done in many different ways depending on what you want the data to say. The last group focused on making the information from the address questions usable. This was a key area, as C2P is also factoring geographical segregation. They extracted zip codes for empty columns from the address given if the zipcode was left blank, filling in many empty rows.

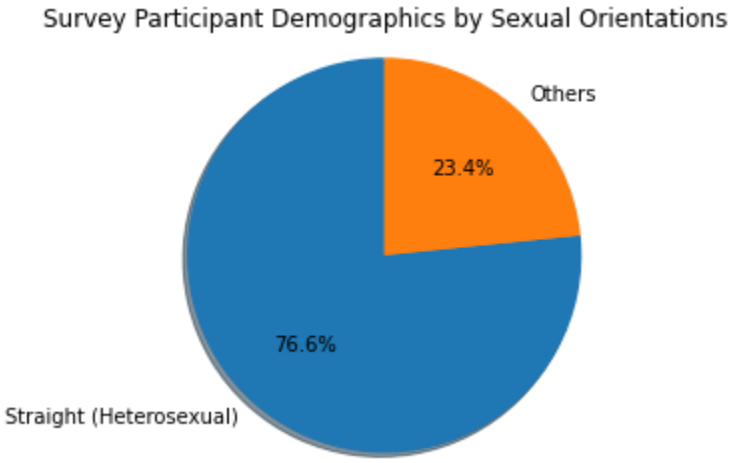
Data Collection

The data has been collected from a survey given to those currently incarcerated at Massachusetts correctional facilities. It is eleven pages long and the fifty-seven questions span topics of childhood, mental health, family, and personality through text and checkboxes.

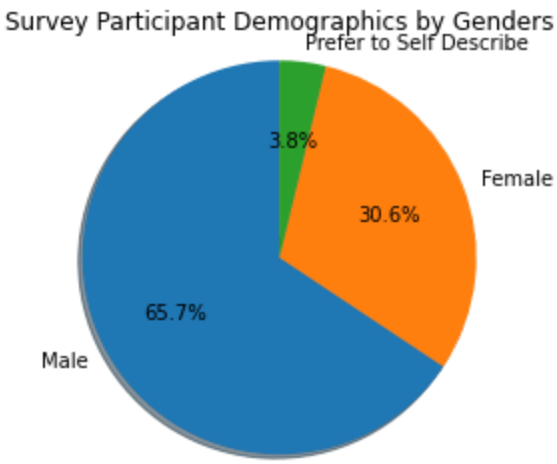
Data Visualization and Exploration



White participants are the highest percentages of respondents, so we should expect White to be significantly represented in the following analysis.



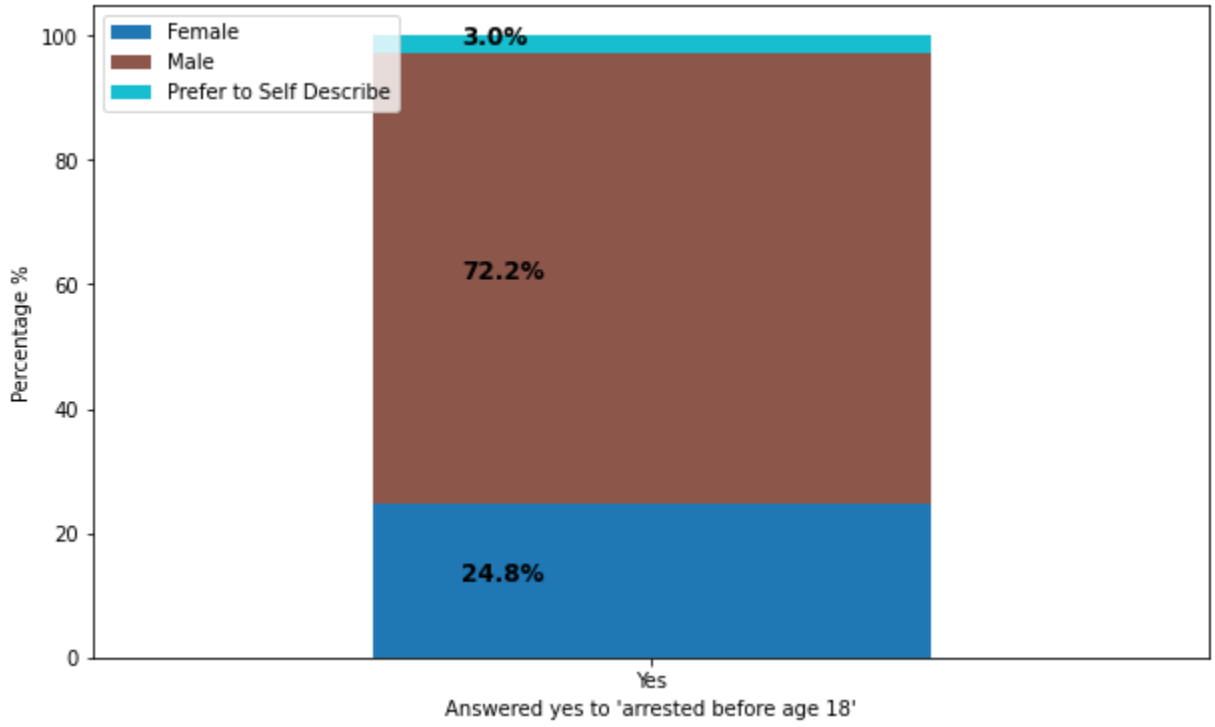
76.6% of the respondents are heterosexual, while only 23.4% are others (lgbtqia+). We also could expect heterosexual to be a majority in any sexual orientational analysis.



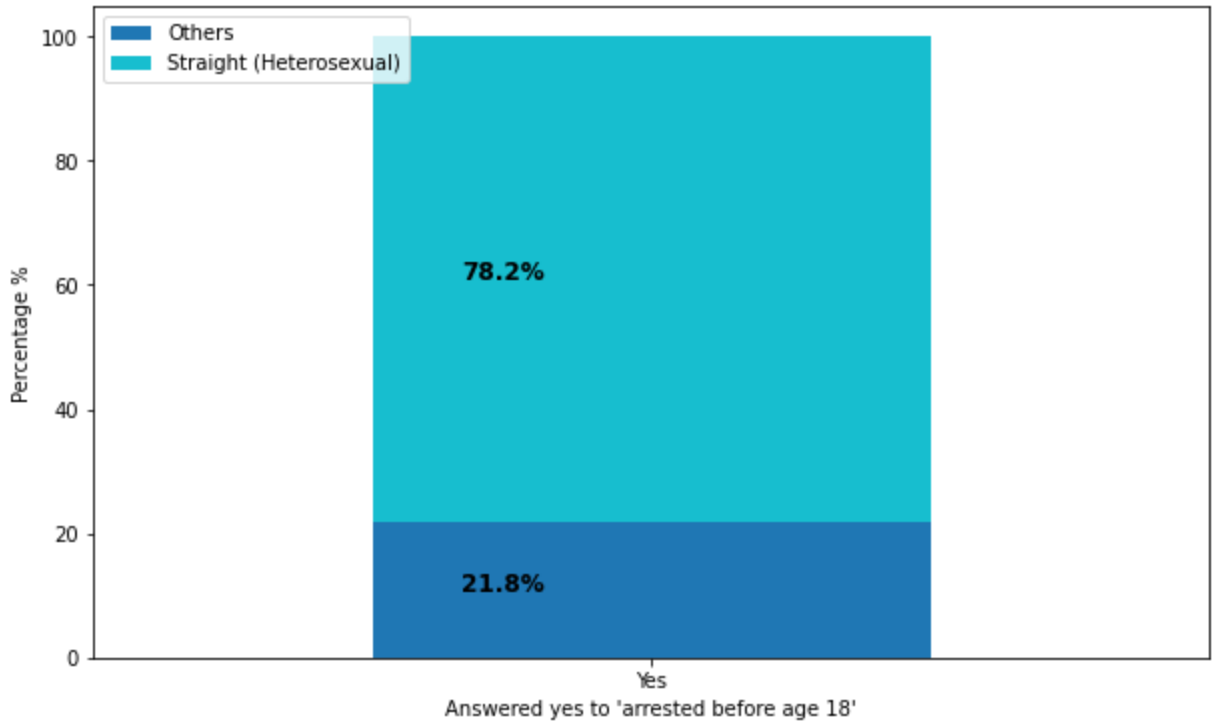
65.7% of the respondents are male which more than doubles the female representation, of 30.6% respectively. Only 3.8% of those who responded use self-described genders.

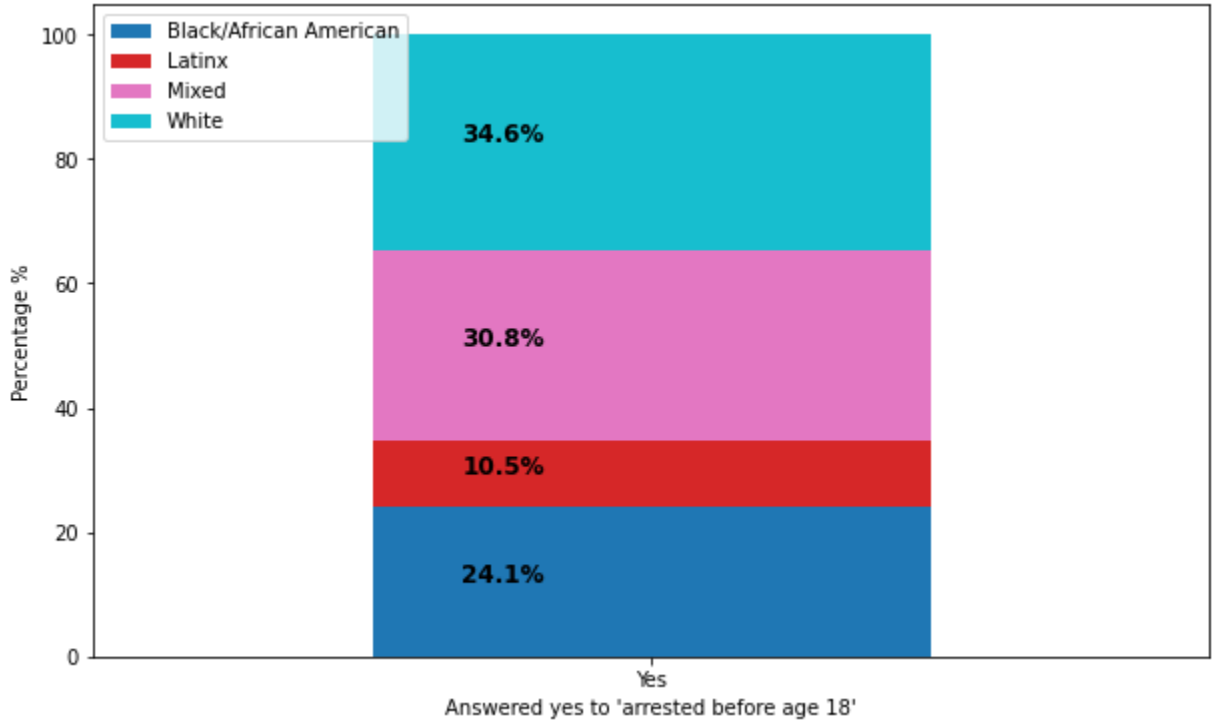
Results Obtained + Questions Answered

- What percentage of survey respondents answered yes to "arrested before age 18" separated by race, gender, and sexual orientation?



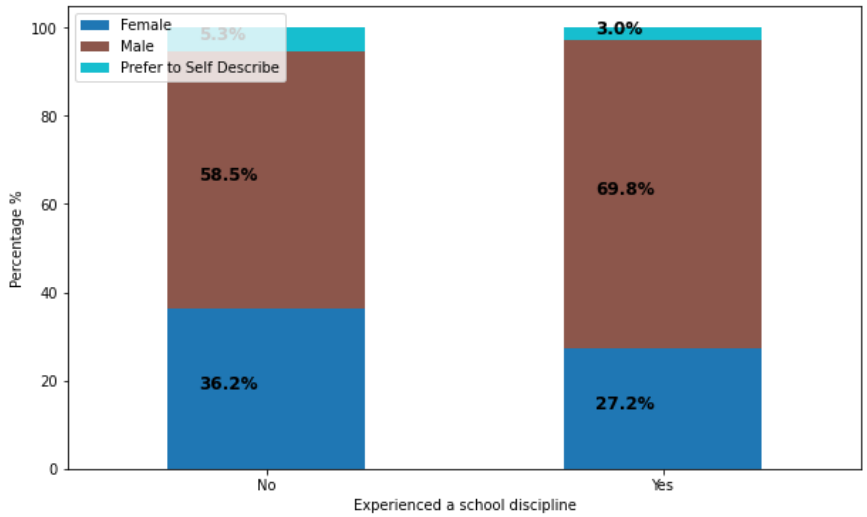
72.2% of those who answered yes are male while 24.8% are female.



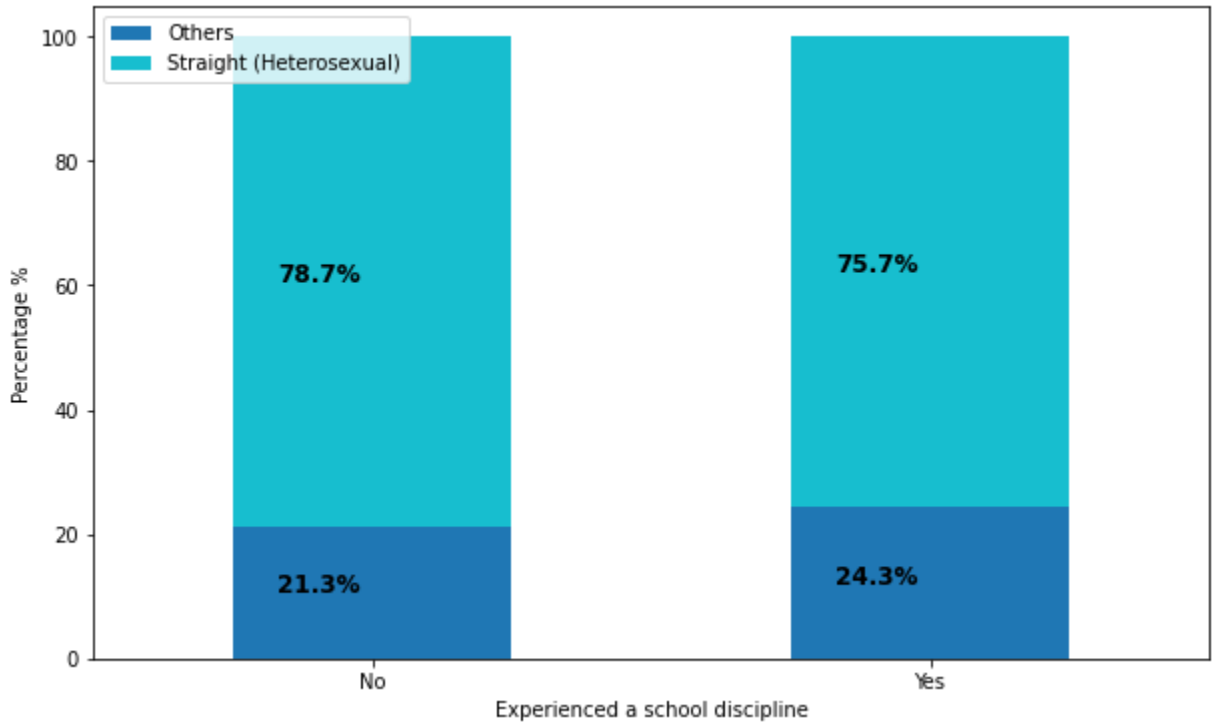


These values may not be representative because Asian respondents were missing.

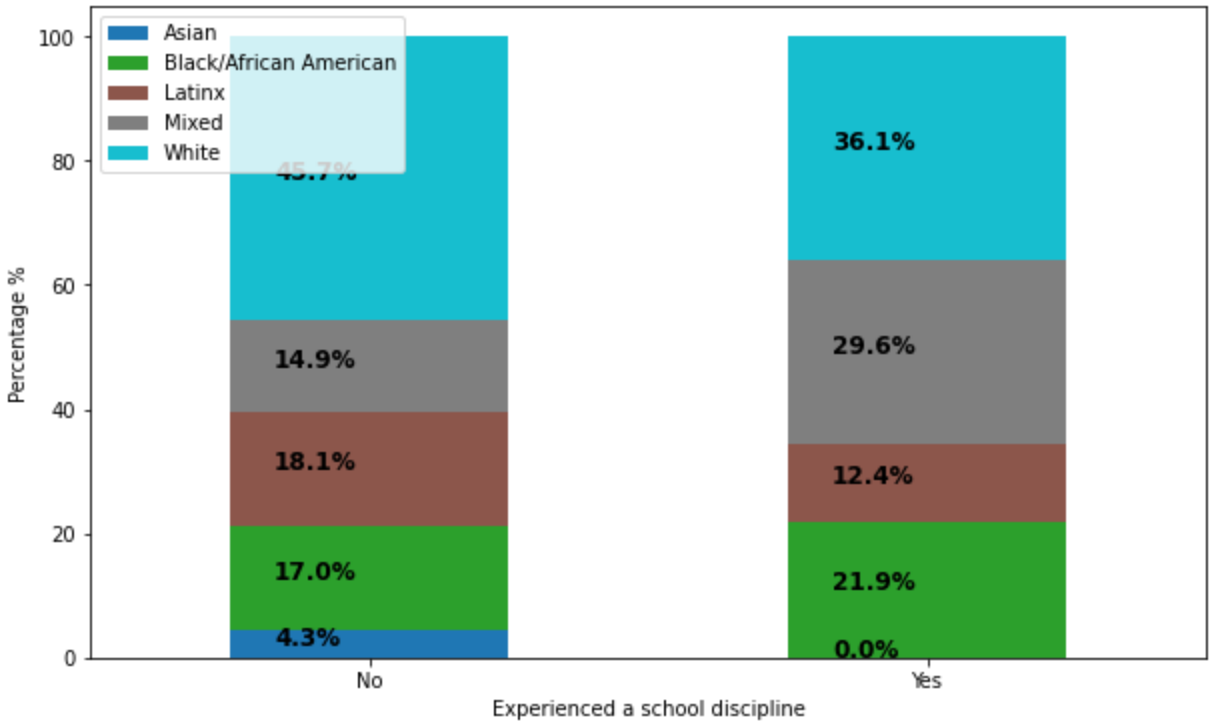
- What percentage of survey respondents experienced school discipline (suspension/expulsion) while in school separated by race, gender, and sexual orientation?



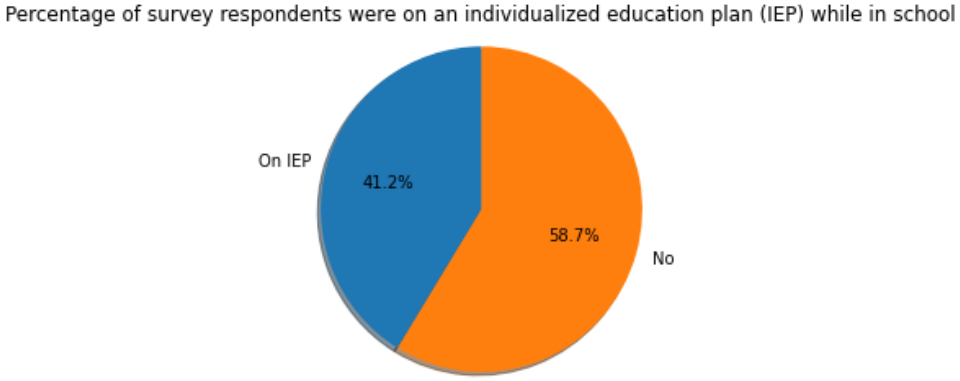
Of those who have experienced school discipline, 69.8% are male, while only 27.2% are female.



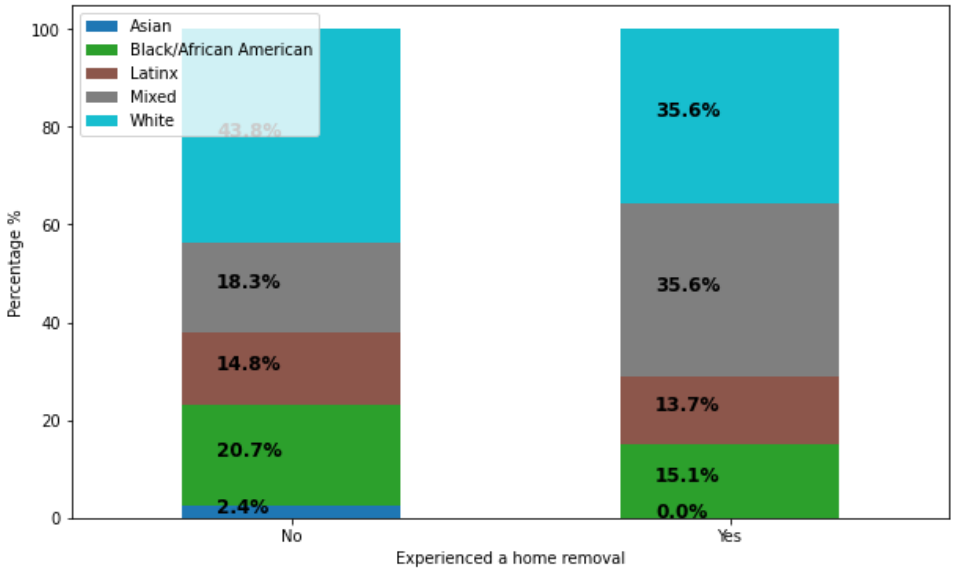
Both sexual orientational groups displayed a similar proportion in both answers. Maybe sexual orientations do not affect school discipline.



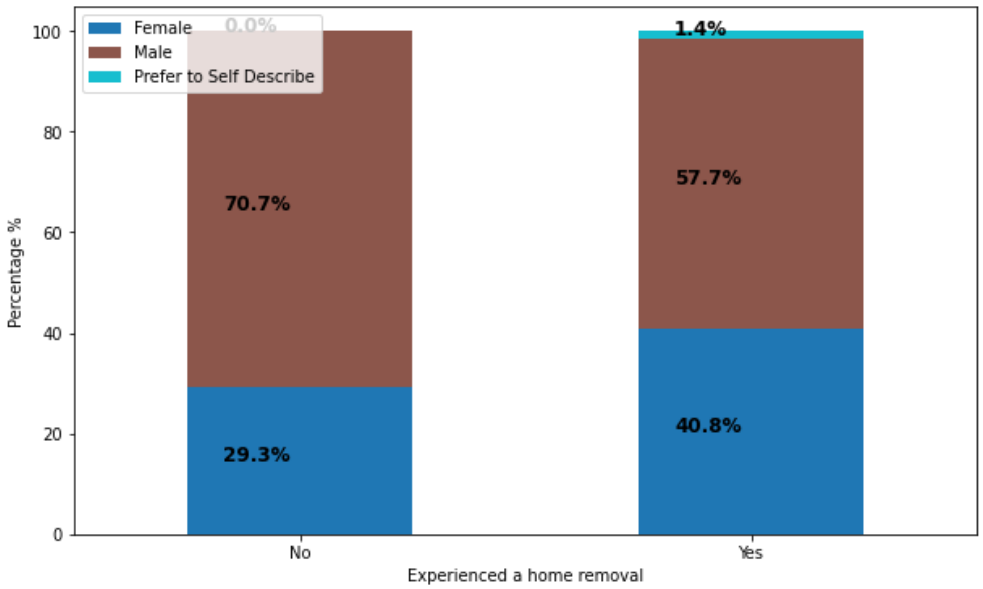
- What percentage of survey respondents were on an individualized education plan (IEP) while in school?



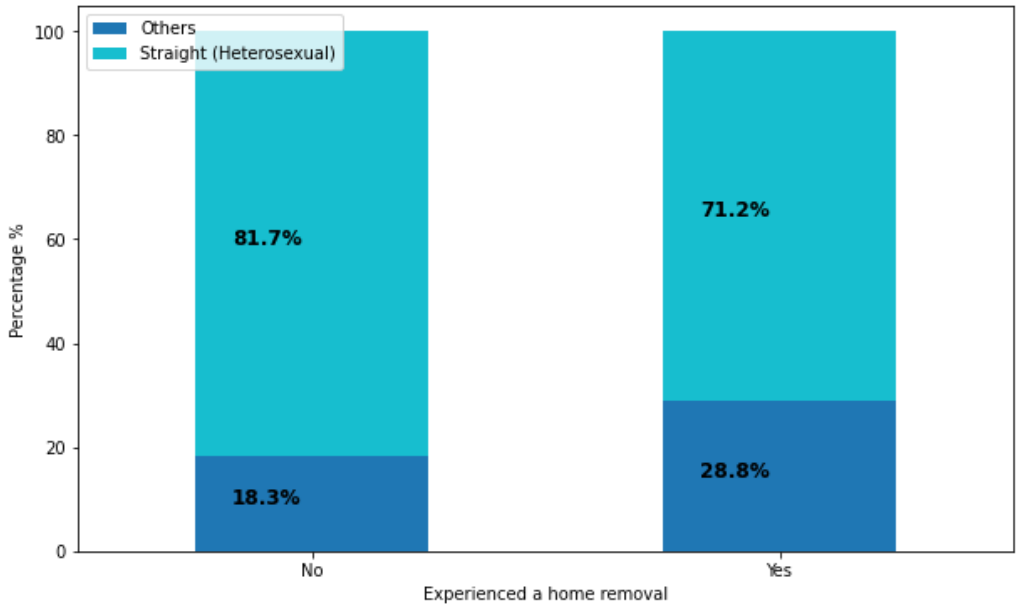
- What percentage of survey respondents experienced a home removal, separated by race, gender, and sexual orientation?



In races, the majority of those who have experienced a home removal were White, and Mixed races (both had 35.6%), while Asian represented 0%.

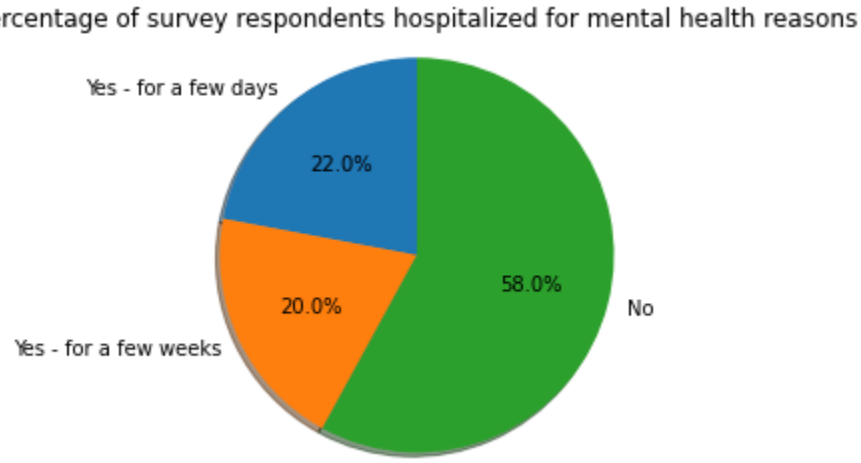


57.7% of the respondents who have experienced home removal are male, while 40.8% are female. However, all respondents who identified as “Prefer to Self Describe” experienced a home removal.



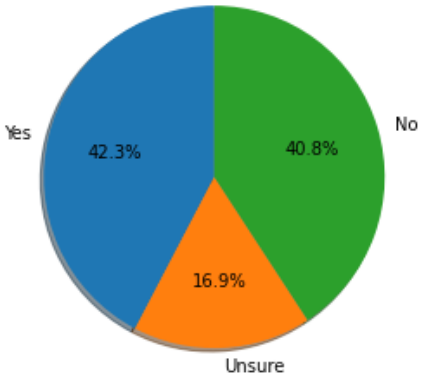
By dividing Heterosexual and others (lgbtqia+), 71.2% who have experienced a home removal are straight, while only 28.8% of those are lgbtqia+.

- Which factors along the pipeline are most predictive of a respondent experiencing incarceration (i.e. arrests before age 18, removal from home, suspension/expulsion, etc.)?

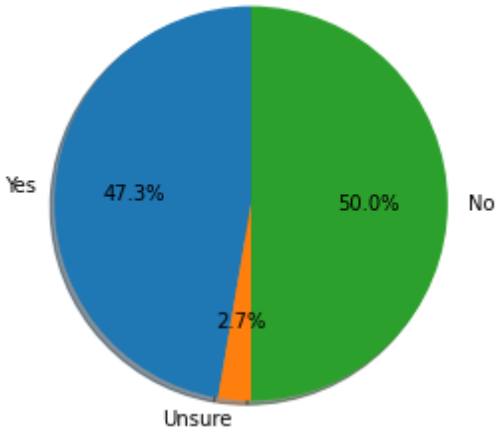


:

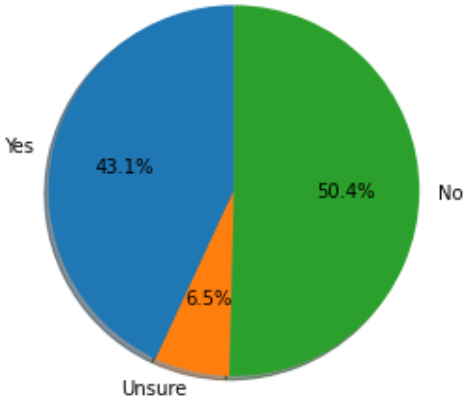
Percentage of survey respondents living with someone (while under 18) who was depressed, mentally ill, or suicidal



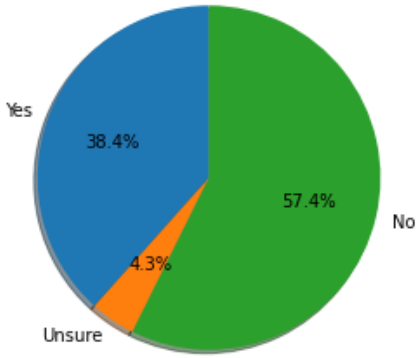
Percentage of survey respondents living with someone (while under 18) who was drinking too much



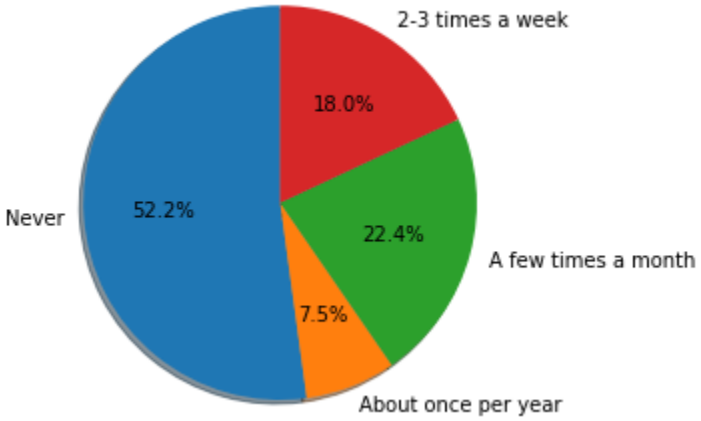
Percentage of survey respondents living with someone (while under 18) who did illegal drugs or medication



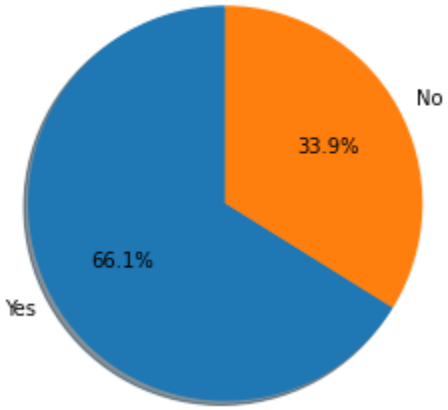
Percentage of survey respondents living with someone (while under 18) who served time or was sentenced to time



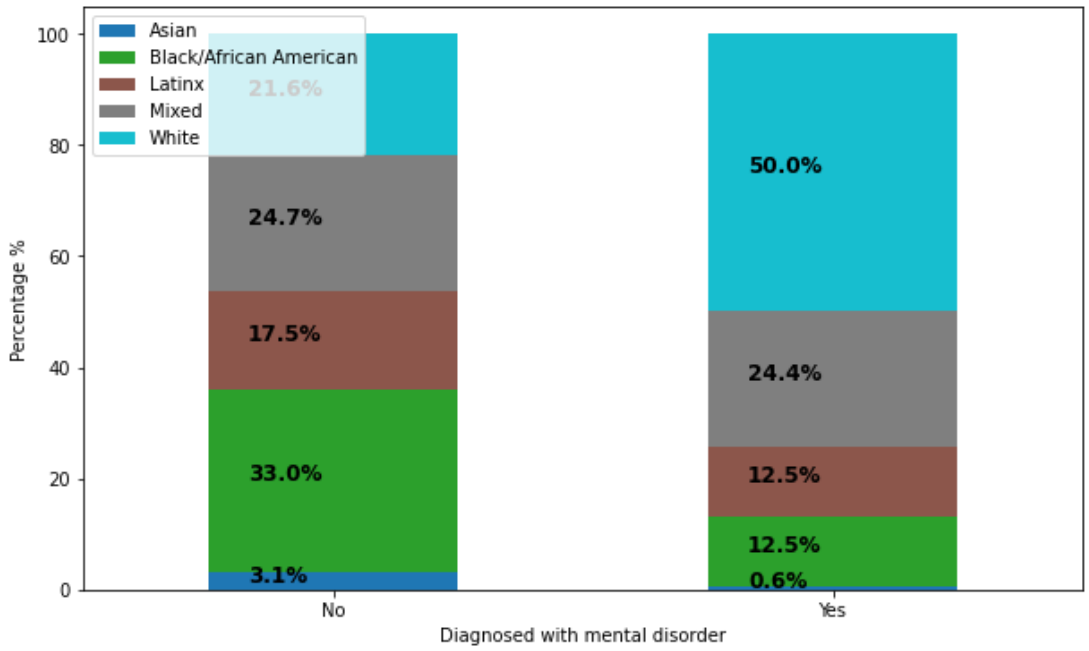
Percentage of survey respondents living with someone (while under 18) who physically hurt them



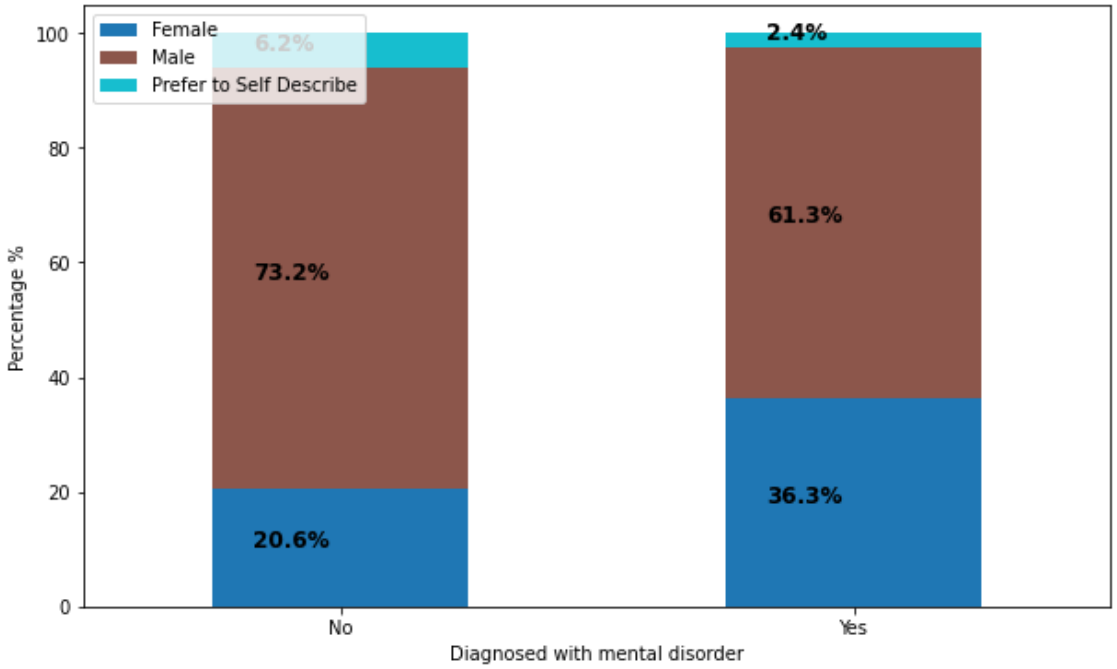
Percentage of survey respondents diagnosised with mental disorder

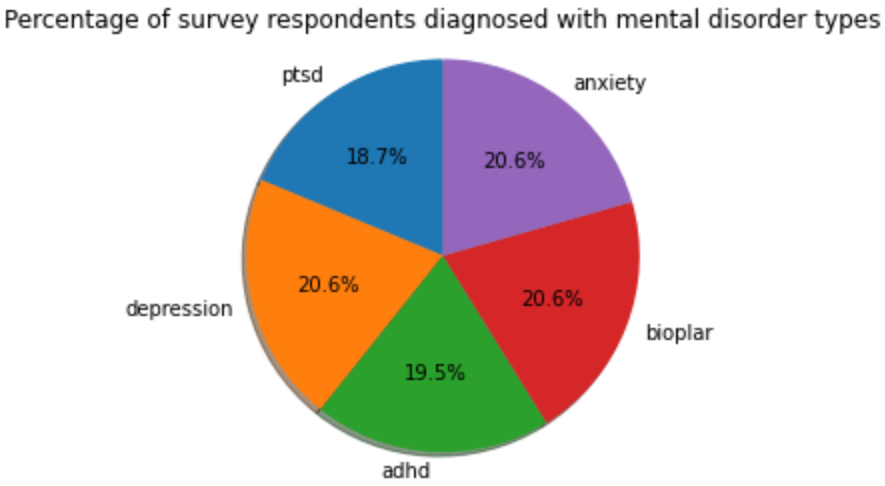


Percentage of survey respondents diagnosed with mental disorder by races

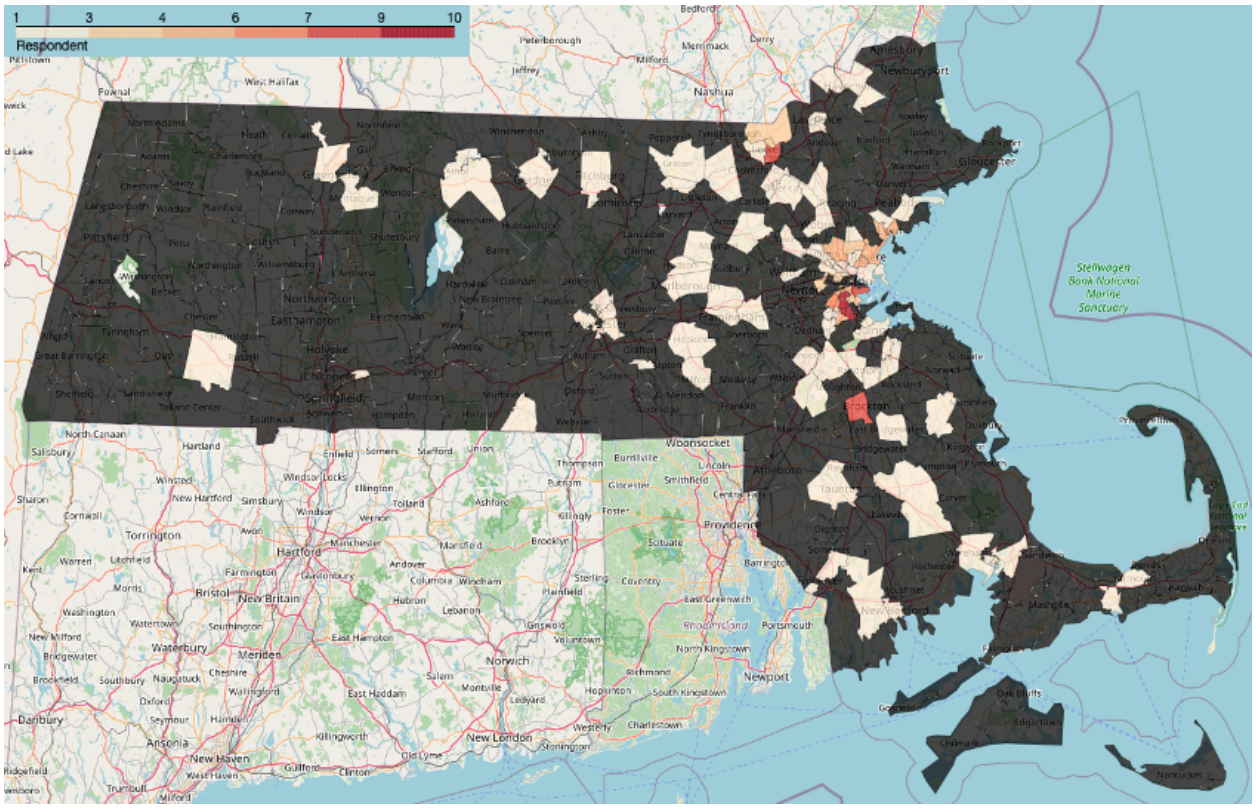


Percentage of survey respondents diagnosed with mental disorder by genders

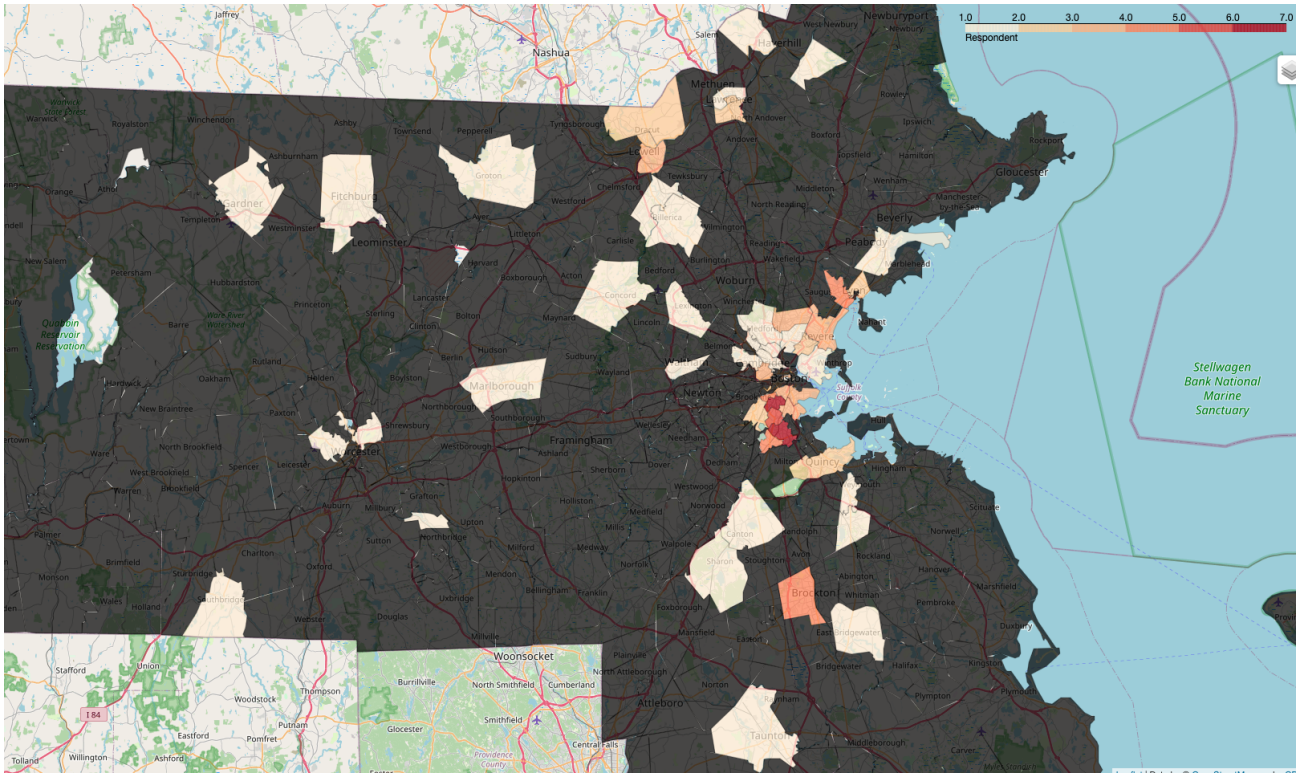




- Where are most respondents from? (zip code analysis)
  The cities and towns are shown on this map that have the most survey respondents also happen to be areas where the per capita income is on the lower end in the state and poorly performing schools. This is not a coincidence. These areas are and have been, at a disadvantage for many years, as they also have a much more diverse population than the areas that are not as affected.



- Where are the most respondents who experienced an arrest before 18 from? (zip code analysis)



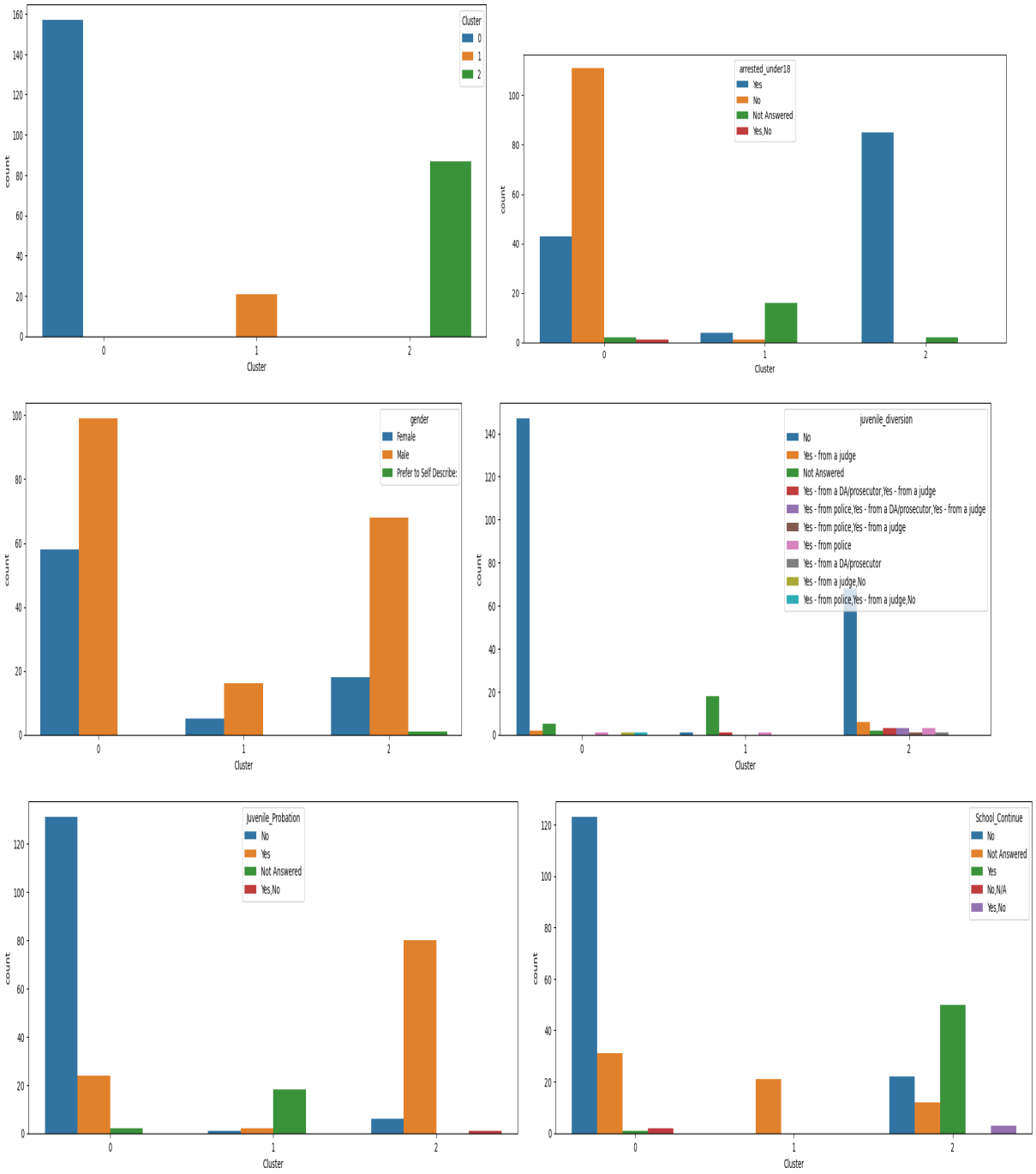
This map lines up with our analysis from the previous question.

Models and Interpretations:

1. Clustering Analysis
- Clustering of various factors related to Juvenile Justice Involvement.

	arrested_under18	juvenile_diversion	Juvenile_Probation	School_Continue	race_ethnicity	gender
Cluster 1	No	No	No	No	White	Male
Cluster 2	Not Answered	Not Answered	Not Answered	Not Answered	White	Male
Cluster 3	Yes	No	Yes	Yes	Black/African America	Male

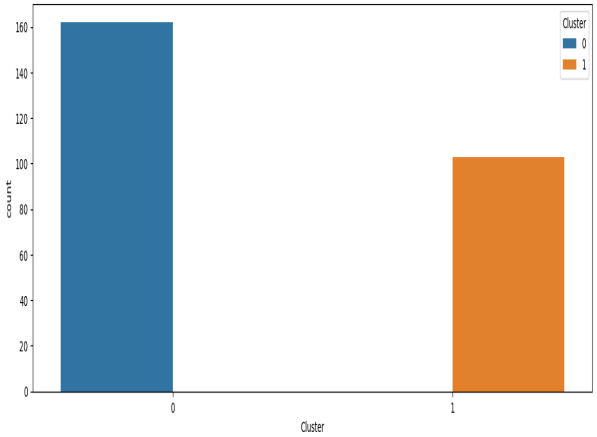
Centroids of the clusters

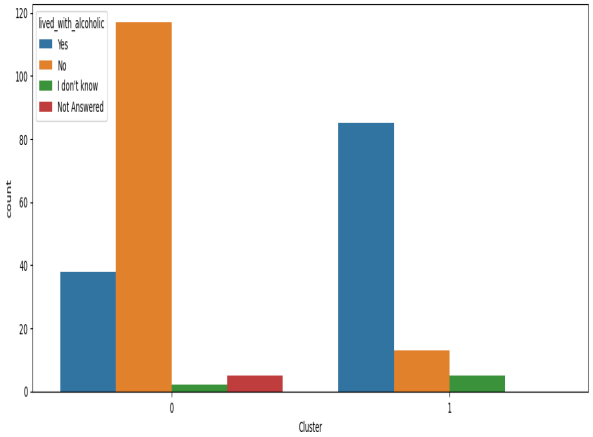


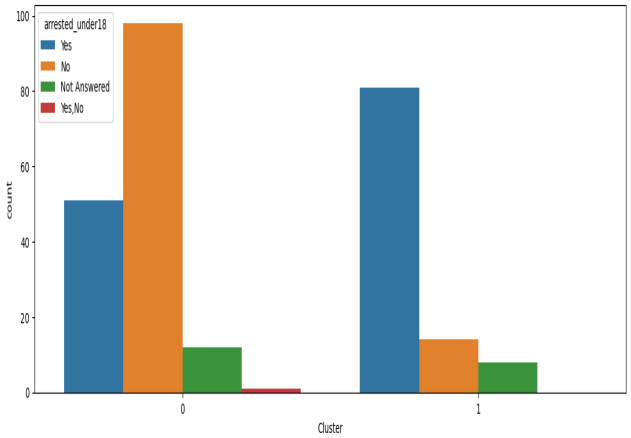
For this clustering, we used Kmode clustering because all the columns contain categorical values. Most of the data has been assigned to the first cluster. As we can see in the second

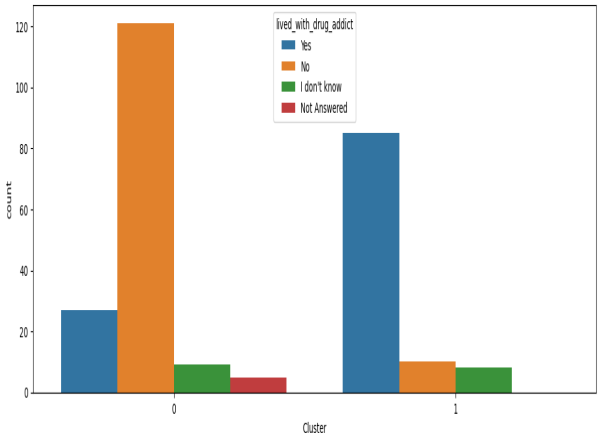
diagram, most of the people who were arrested before age 18 belong to cluster 3. Also we can see that juvenile probation was given to the people only in cluster 0. Also from the diagram, we can interpret that those who were given juvenile probation were not able to continue school.

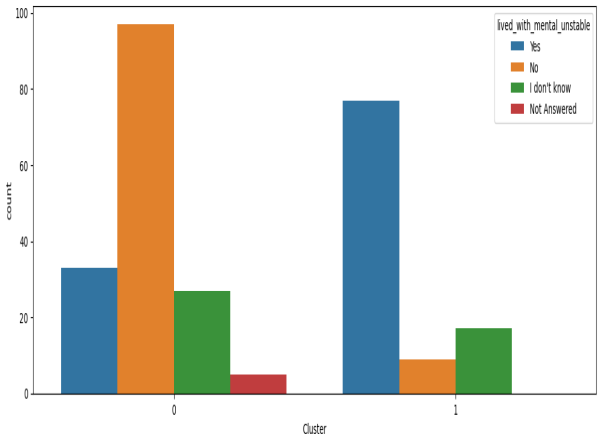
- Clustering of various factors related to childhood environment and juvenile arrest

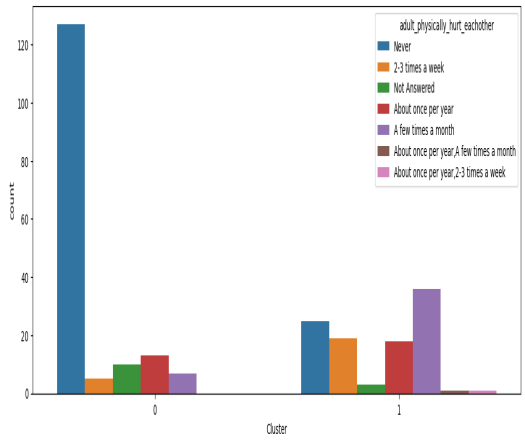


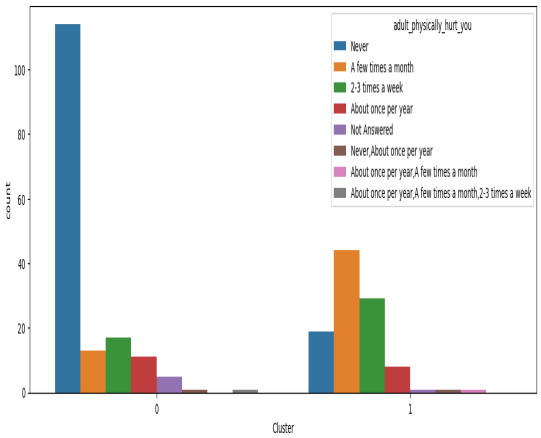


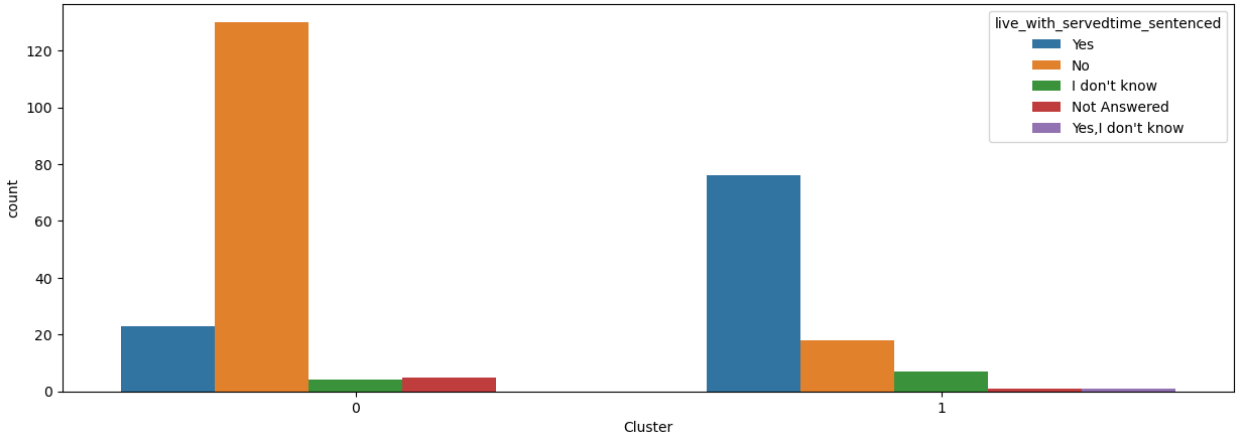










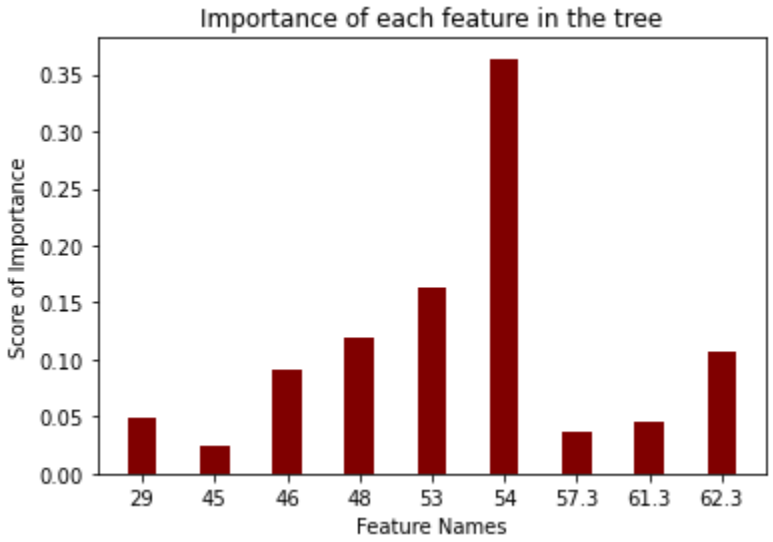


As we can see from the diagram, most of the people who were arrested before age 18 belong to cluster 1. Also from the diagram we can interpret that most of the people who were arrest under age 18 were living with alcoholic, drug addict and mentally unstable person.

2. Dicision Tree Analysis:

Survey questions that we chose:
#29 home removal (yes/no)
#45 positive experience with the police during adolescence (yes/no)
#46 negative experience with the police during adolescence (yes/no)
#48 spend time in youth correctional facility (yes/no)
#53 relationship with lawyers in the juvenile justice system (yes/no/mixed)
#54 juvenile convictions (yes/no)
#55 feel treated fairly in the juvenile justice system (yes/no/mixed)
#57.3 did they go to high schools (yes/no)
#61.3 did they get suspended at least twice (yes/no)

#62.3 did they get expelled from high schools (yes/no)



Using a decision tree to examine how the juvenile protection program performs (i.e: keep the youth from the pipeline before age 18), based on home safety, juvenile justice involvement, and schooling, I found that juvenile convictions and relationship with the layer in the juvenile justice systems are the most important predictors of whether the respondents go to jail before 18. While juvenile convictions seem obviously reasonable, the relationship with the juvenile justice lawyer is interesting.

3. Multiple Regression Analysis

We going to perform a multiple regression to analyze the significance of features:
By fitting into a multiple regression model, we got the following p-values for each features:

	coef	std err	t	P>|t|	[0.025	0.975]
29	0.1045	0.065	1.598	0.112	−0.024	0.233
45	−0.0304	0.081	−0.376	0.707	−0.190	0.129
46	0.2437	0.059	4.117	0.000	0.127	0.360
48	0.3258	0.083	3.949	0.000	0.163	0.489
53	−0.0169	0.036	−0.477	0.634	−0.087	0.053
54	0.2827	0.079	3.580	0.000	0.127	0.438
57.3	0.1960	0.081	2.407	0.017	0.035	0.357
61.3	0.0915	0.064	1.435	0.153	−0.034	0.217
62.3	−0.0065	0.071	−0.091	0.928	−0.147	0.134

From here, we can see that question 46, 48, 54, and 57.3 are statistically significant with 95% confidence level, which corresponds to negative experience with the police during adolescence, youth correctional facility participation, juvenile convictions, and high school attendance.

Limitations:
Since most of the survey answers were either in category type or in text, there are few models like clustering and decision tree can meaningfully visualize the data. Quantitative modelings like regression analysis are limited to analysis the data.

Challenges

The biggest challenge was the data itself. If it was a survey of only check boxes, it would be very easy to sort and see trends. Adding text made it more difficult. This added more human error. The text responses had many spelling errors and/or unclear script. This can cause a loss of data and is difficult to correct. Many questions of the survey were also left unanswered, which made it difficult to get an accurate answer. Along the same line of spelling errors, any question with a “prefer to self-describe:” option made things more complex. Some respondents selected this option, then wrote an option that was already listed or misspelled their choice.

We decided to follow a few basic rules for cleaning to try and get around these challenges. It was difficult to decide, as there are many ways that could be considered best practices. To correct for null values we filled appropriate numeral columns with the mean or median value of the column. Making all zero would skew the results as would dropping all the empty rows.

Suggestions for the Future of the Project

In the next phase of this project, it would be beneficial to translate the zipcode data into an interactive format. Something such as a heatmap with many layers that show various pipeline factors across the state based on the questions we answered to look for geographical trends. This would be useful for presenting the project to others, as it shows tangible data across the state in an understandable and digestible form.

So far, we only focus on races, genders, and sexual orientation. Zip code analysis shows a potential perspective to dwell in. It may bring more insight to look at geographical trends both in of states and out of states. This would require more surveys with more question about geographical information like ratings of local school, heath care system, laws, diversity.

Another suggestion that would yield better results is finding a solution to the spelling errors. The data could be combed over by hand, but the point of cleaning is for it to be done efficiently. Many responses cannot be factored into the data analysis because the code will not understand it is an error. One solution could be to use a word vectorization of common misspellings to map them to the correct spelling.