Cradle-2-Prison Pipeline

1. **Background / Motivation:**
   - In the light of understanding the reason behind the mass incarceration of current prisoners in Massachusetts, Northeastern University has started a project in collaboration with College of Art, Media + Design (CAMD), Center for Public Interest Advocacy and Collaboration (CPIAC) at the School of Law, Department of Sociology and Anthropology at the College of Social Science and Humanities (CSSH) and Boston Area Research Initiative (BARI) to tackle this matter.
   - Survey questionnaires have been sent and filled out by prisoners from State Prison and County Jail. Specifically, the survey addresses critical data gaps in data relating to childhood experiences and systems involvement along the pipeline. The survey was designed to test the hypothesis that children of color are disproportionately impacted by an array of systems—child welfare, public education, mental health, school discipline, and juvenile justice—which operate to increase the likelihood of adult incarceration. The collection of data through this survey is intended to support advocacy efforts and inform policy decisions to help dismantle the cradle-to-prison pipeline and address mass incarceration in meaningful ways.
   - With this idea in mind, Team 2 has been working to determine which factors in the questionnaire heavily influence the current prisoners being incarcerated.

2. **Team members: Team 2**
   - Quan Pham (Team Leader)
   - Melody Chan
   - Nicole Chiulli
   - Carlos Gianello

3. **Previous Work:**
   a. Data Collection
      - Questionnaires have been sent and filled out by the prisoners from County Jail and State Prisons.
      - Teams from the previous semester have been working with AWS Textract and OpenCV to generate CSV dataset files based on the questionnaire textbox and written answer data.
   b. Data Cleaning
      - Most of the work from previous teams did not have a full attempt to clean the whole dataset, just a few columns that would be useful for data analysis.
   c. Data Analysis
      - Several preliminary analysis has been carried out, and 1 team has applied a Decision Tree model included in the analysis.

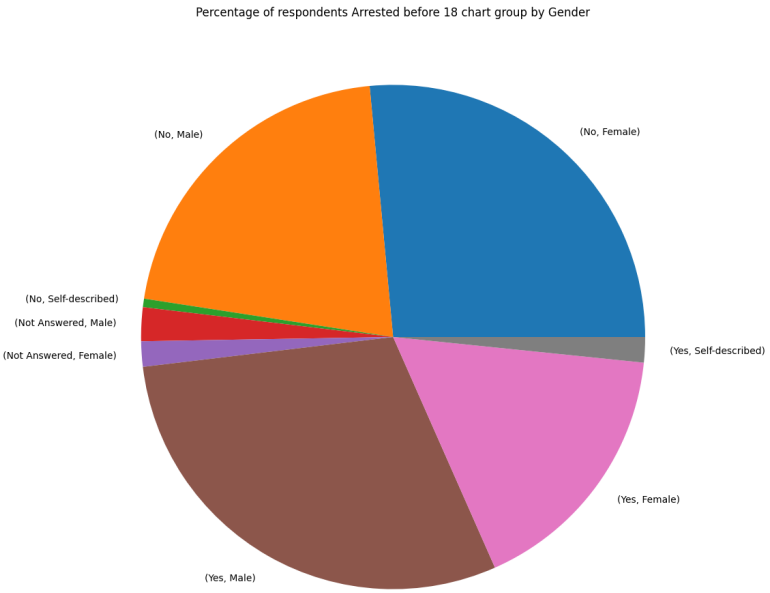
4. **Data Analysis**
   a. Workflow:
      i. Since the dataset is divided into 2 phases, we will be handing the dataset from phase 2. First, we will be cleaning the dataset phases 2 by cleaning the columns of the dataset that will be beneficial for the analysis.
      ii. Once we finalized the cleaning version of the dataset, we can further improve the current cleaning code for cleaning the combined dataset.
      iii. We then will carry out the preliminary analysis on phase 2 dataset, and comparison with the combined dataset. Finally, based on the several columns as predictor

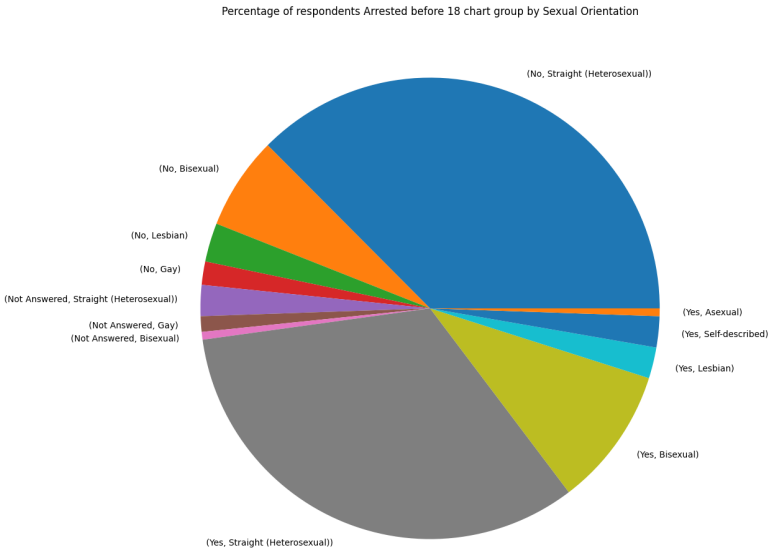
features, we will apply the use of Machine Learning to determine which features in the questionnaire will primarily weigh the most on the prisoners being incarcerated.

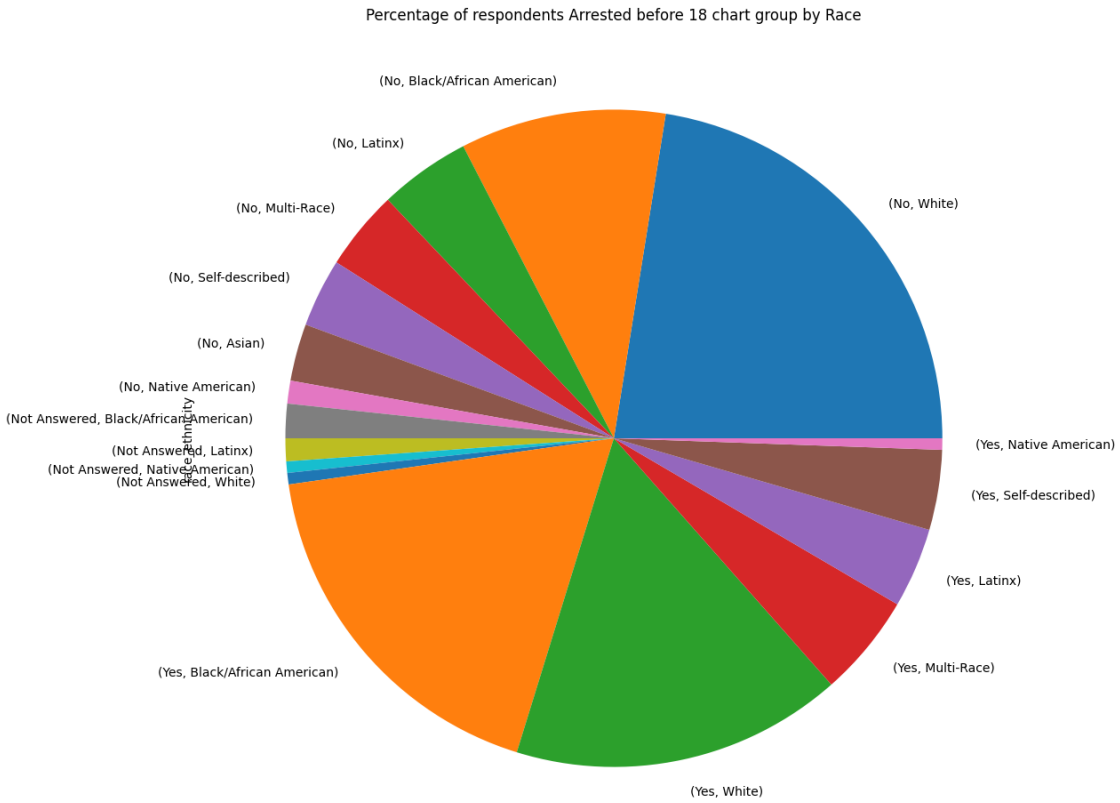
b. Numerical Analysis:

i. Question 1: What percentage of survey respondents answered yes to "arrested before age 18" separated by race, gender, and sexual orientation? [1]

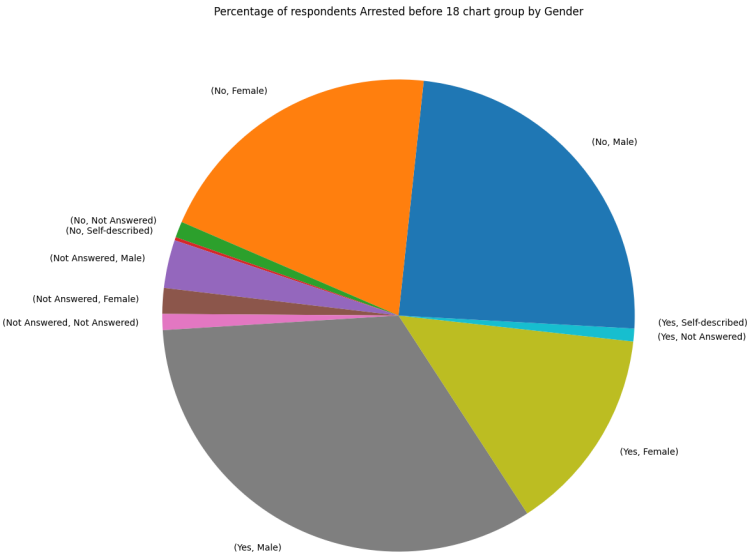
- Phase 2 Data only

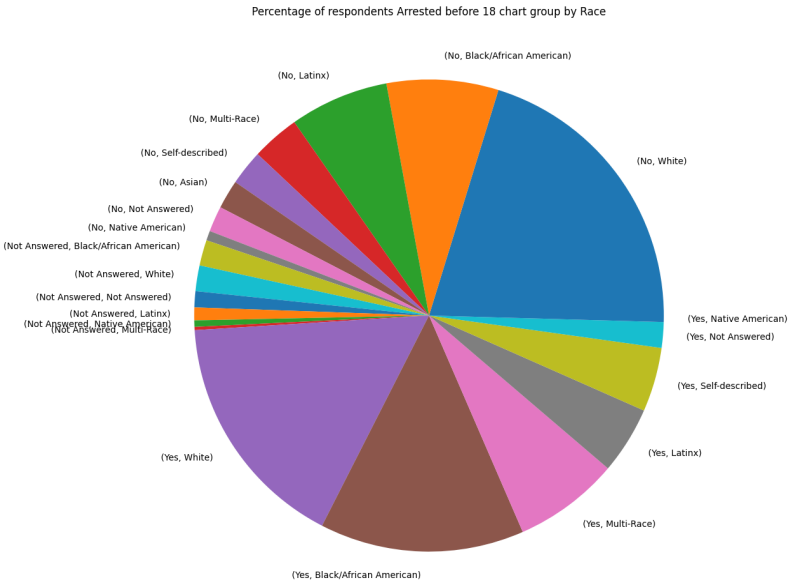


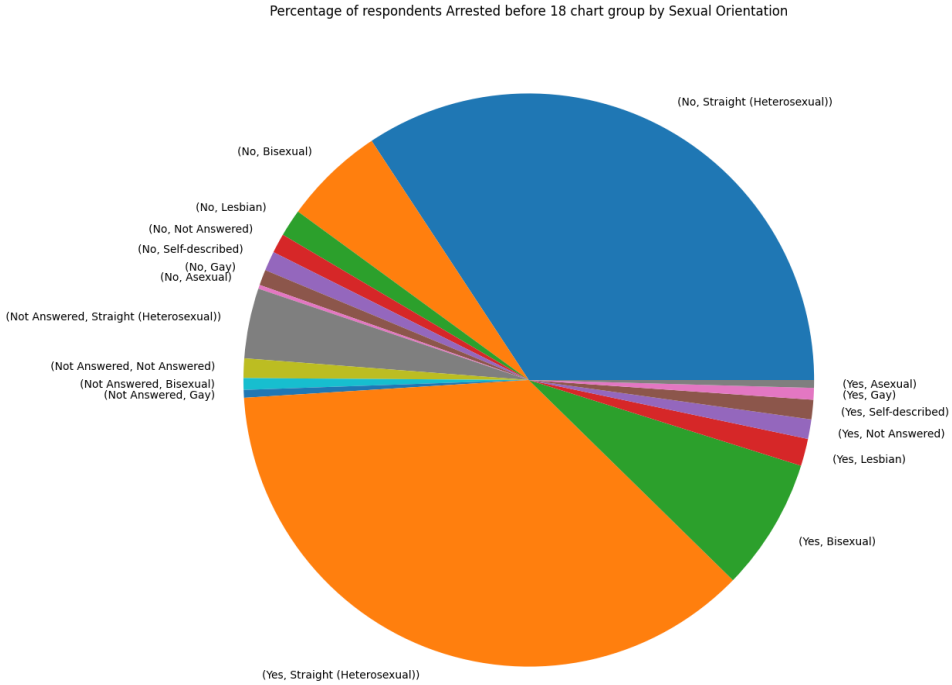




- Full dataset (phase 1 + 2 combined)

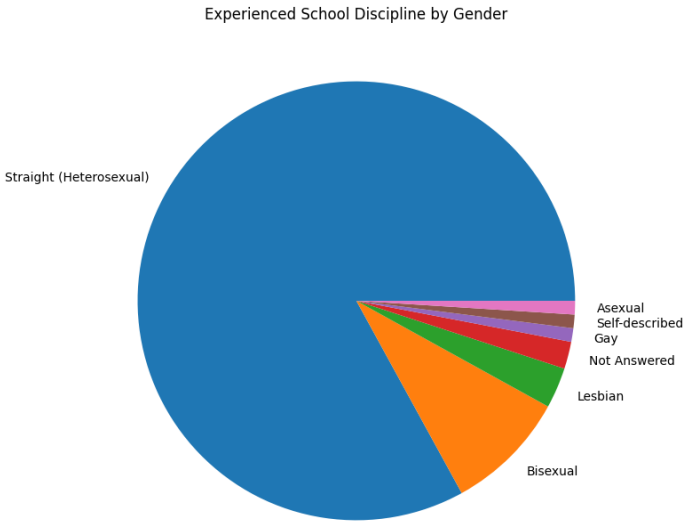


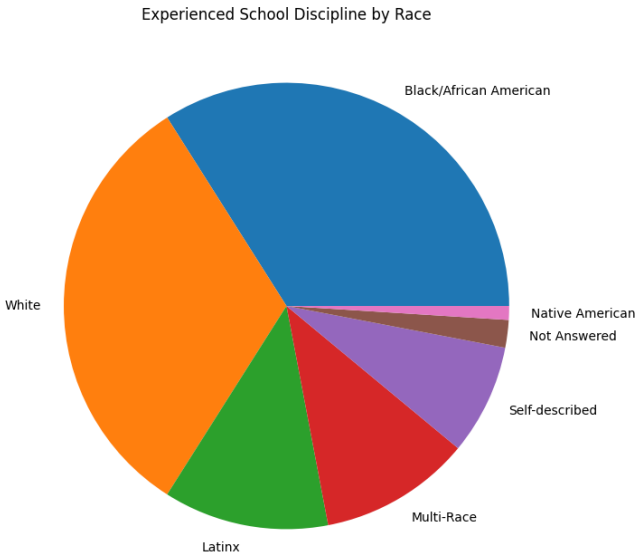


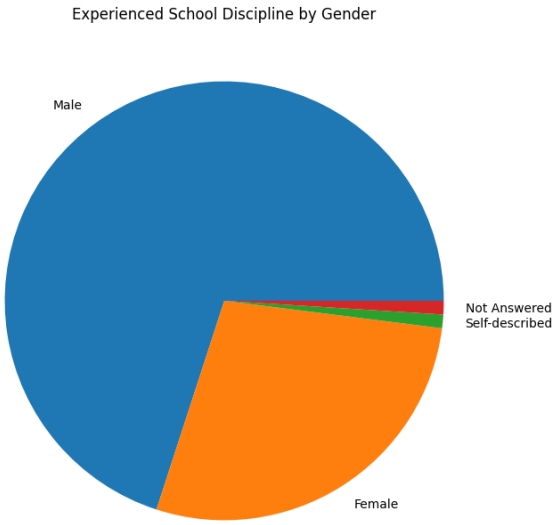


ii. Question 2: What percentage of survey respondents experienced school discipline (suspension/expulsion) while in school separated by race, gender, and sexual orientation?[2]

- Phase 2 Data only
- Full dataset (phase 1 + 2 combined)

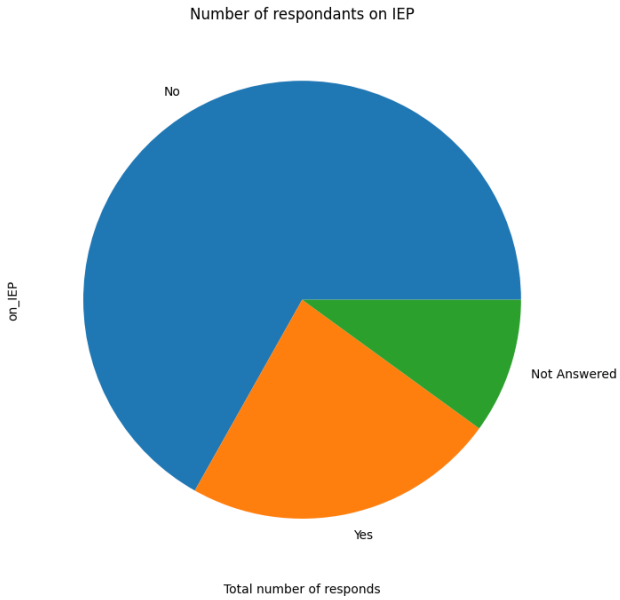




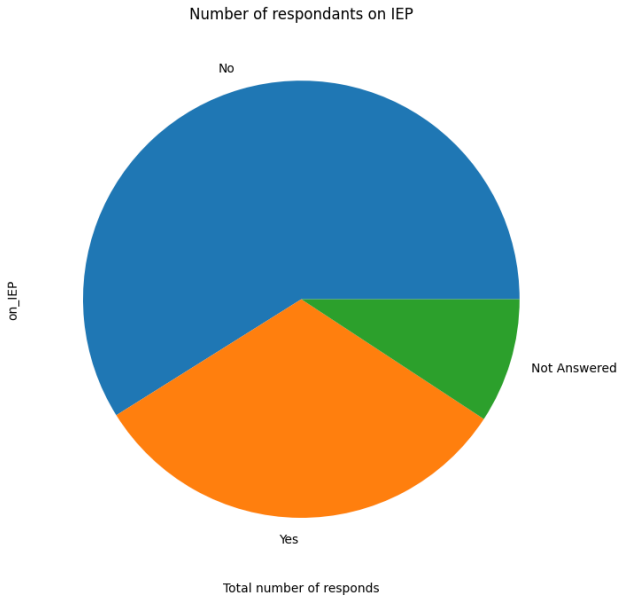


iii. Question 3: What percentage of survey respondents were on an individualized education plan (IEP) while in school?[3]

- Phase 2 Data only

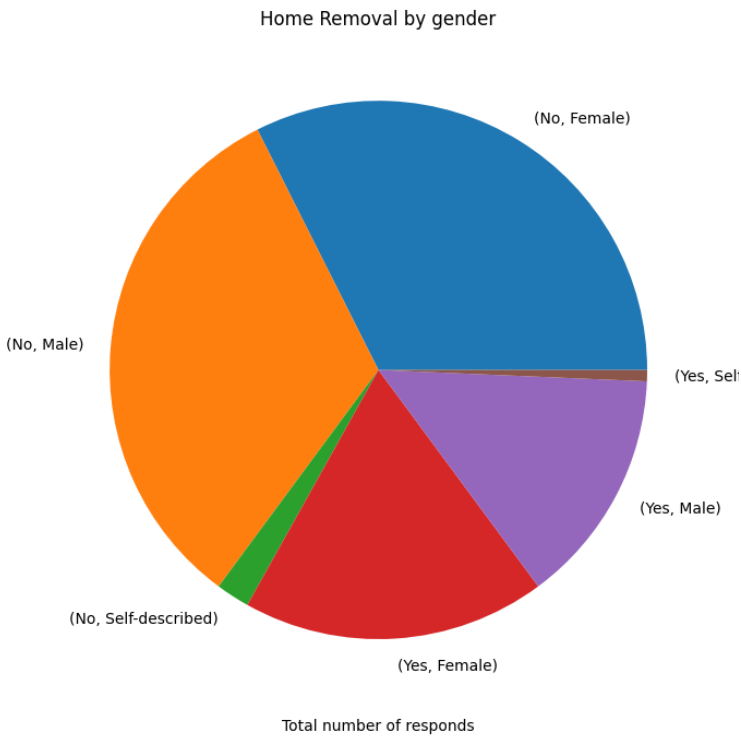


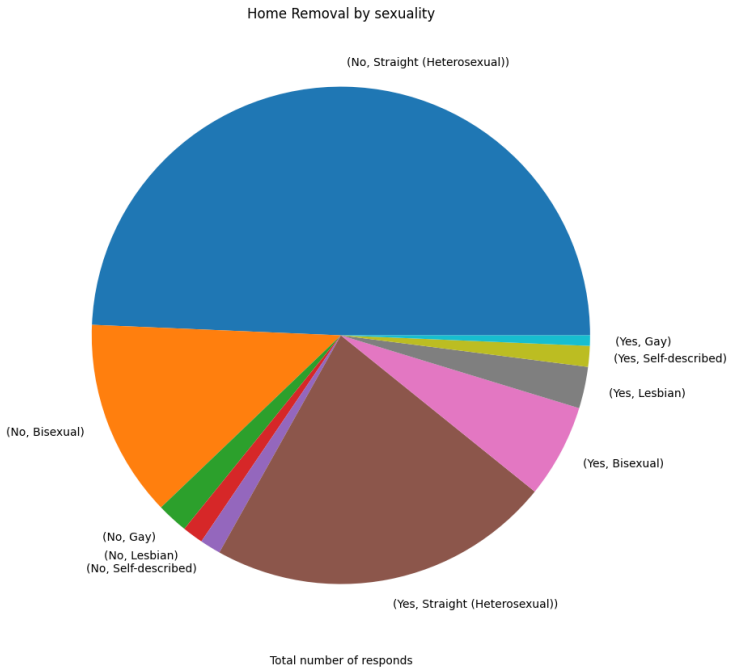
- Full dataset (phase 1 + 2 combined)

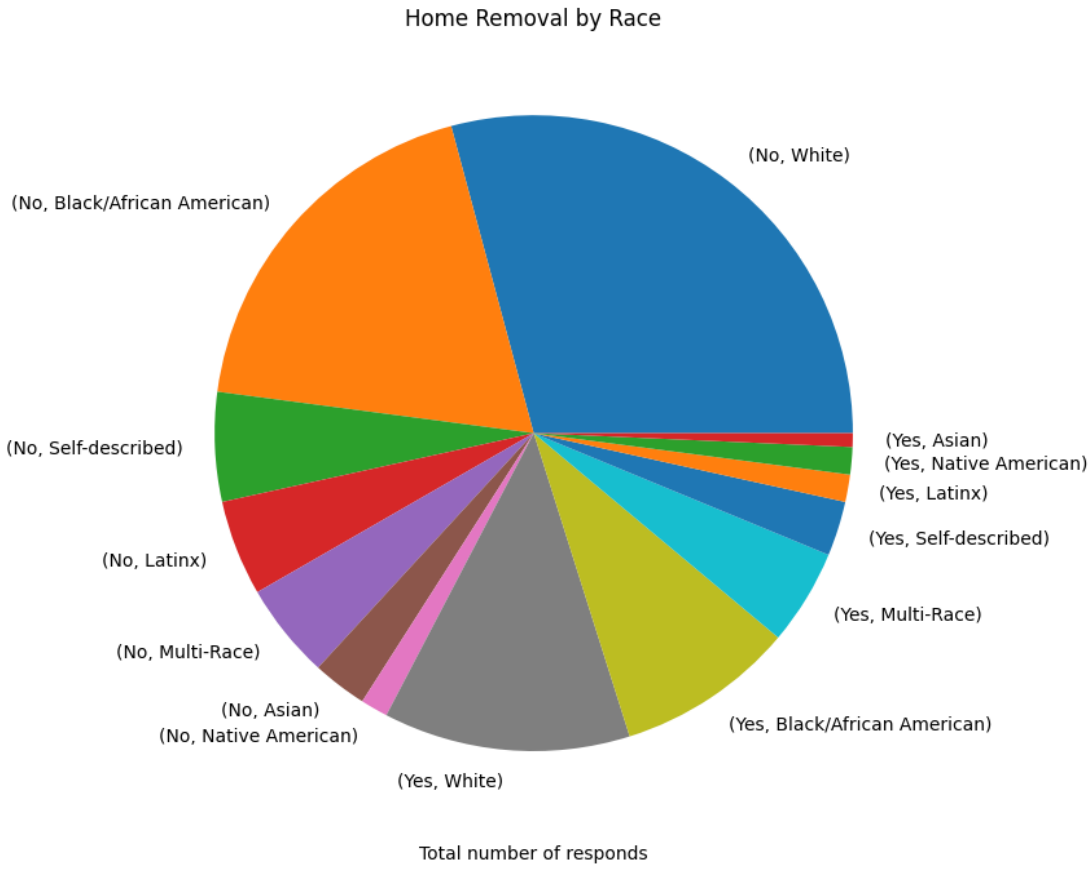


iv. Question 4: What percentage of survey respondents experienced a home removal, separated by race, gender, and sexual orientation? [4]

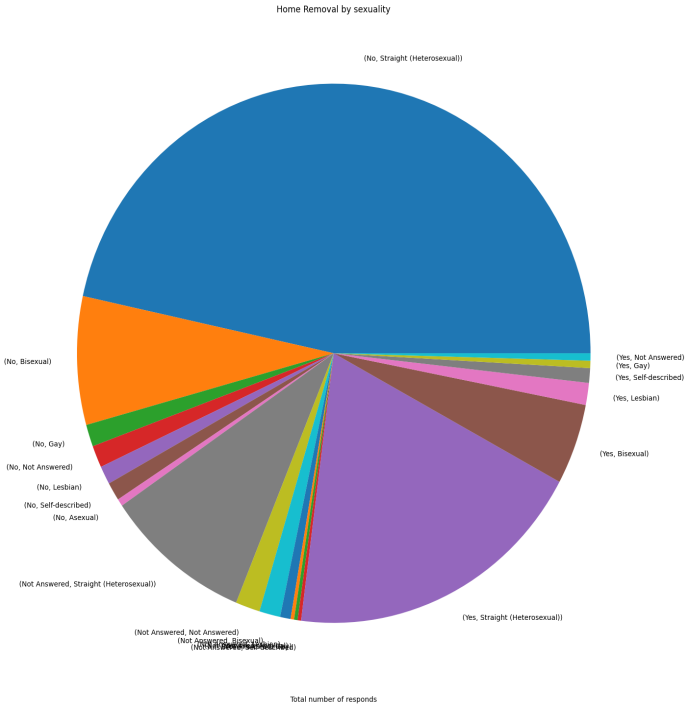
- Phase 2 Data only

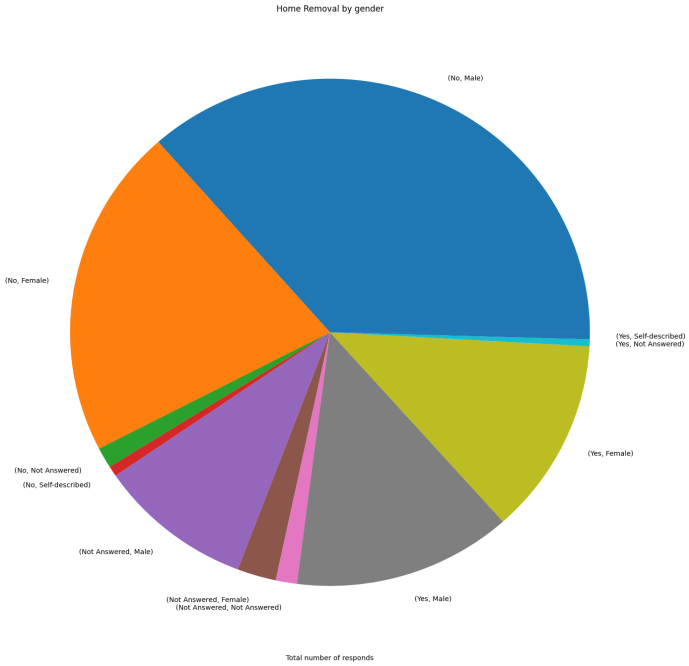


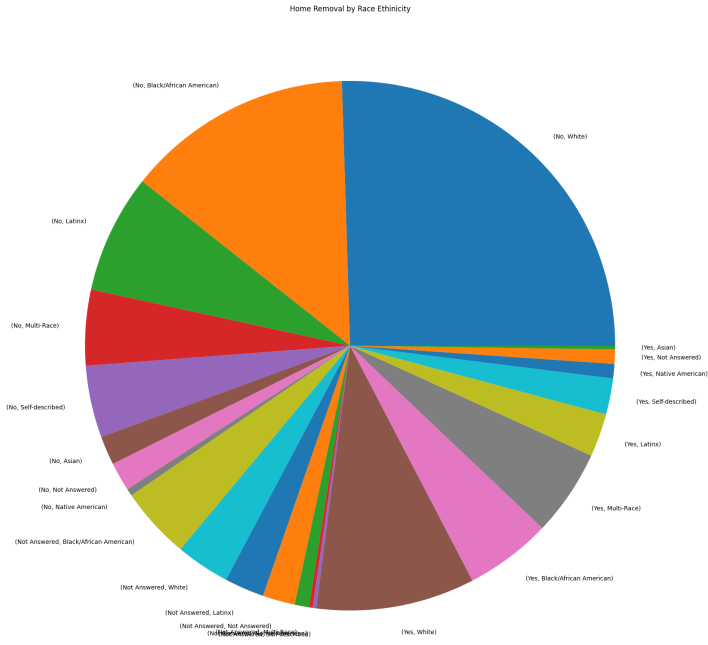




- Full dataset (phase 1 + 2 combined)

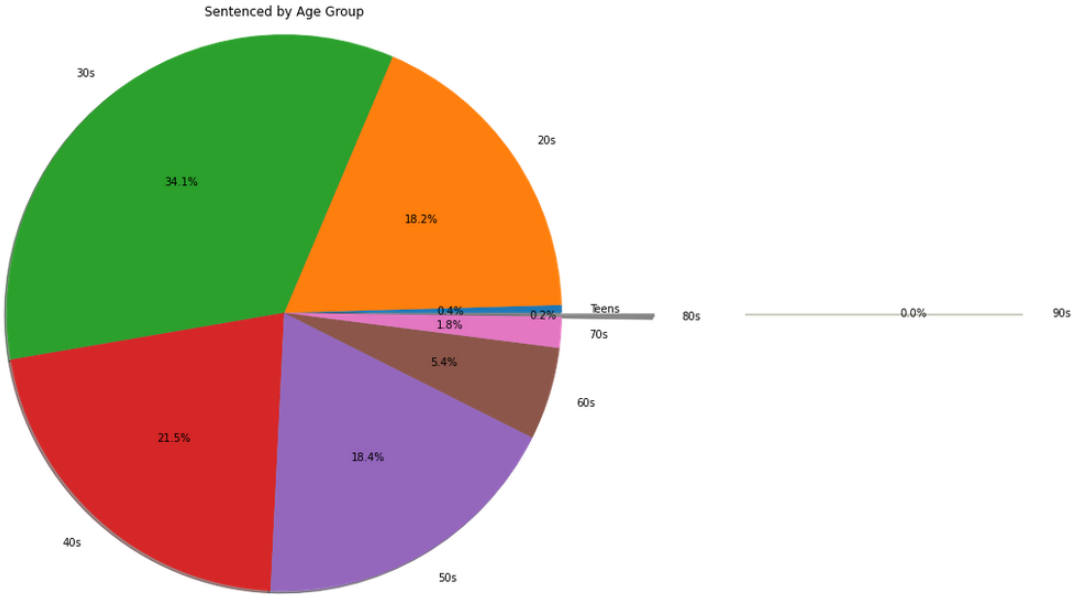


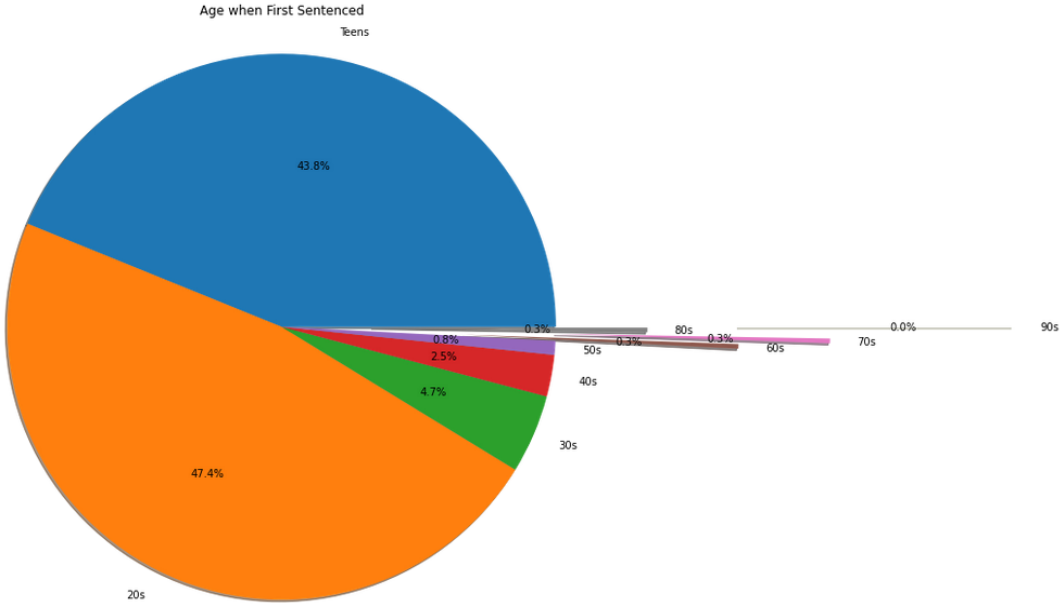


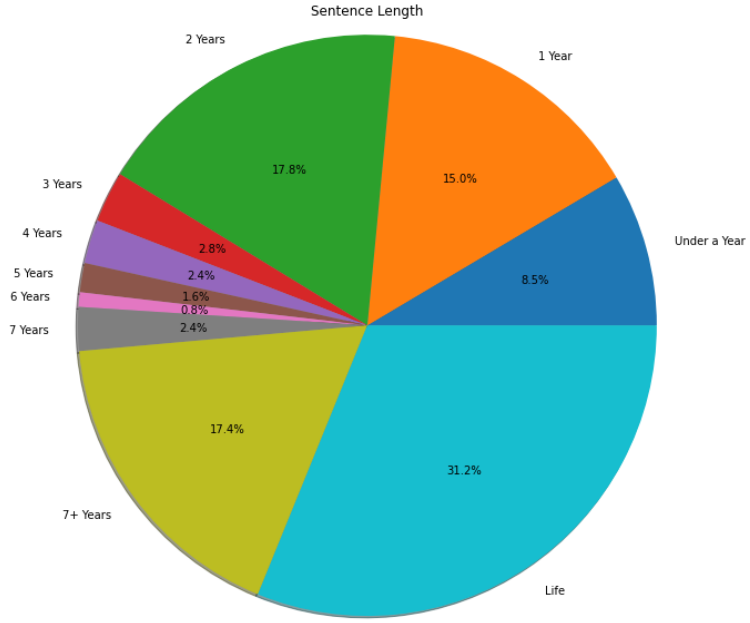


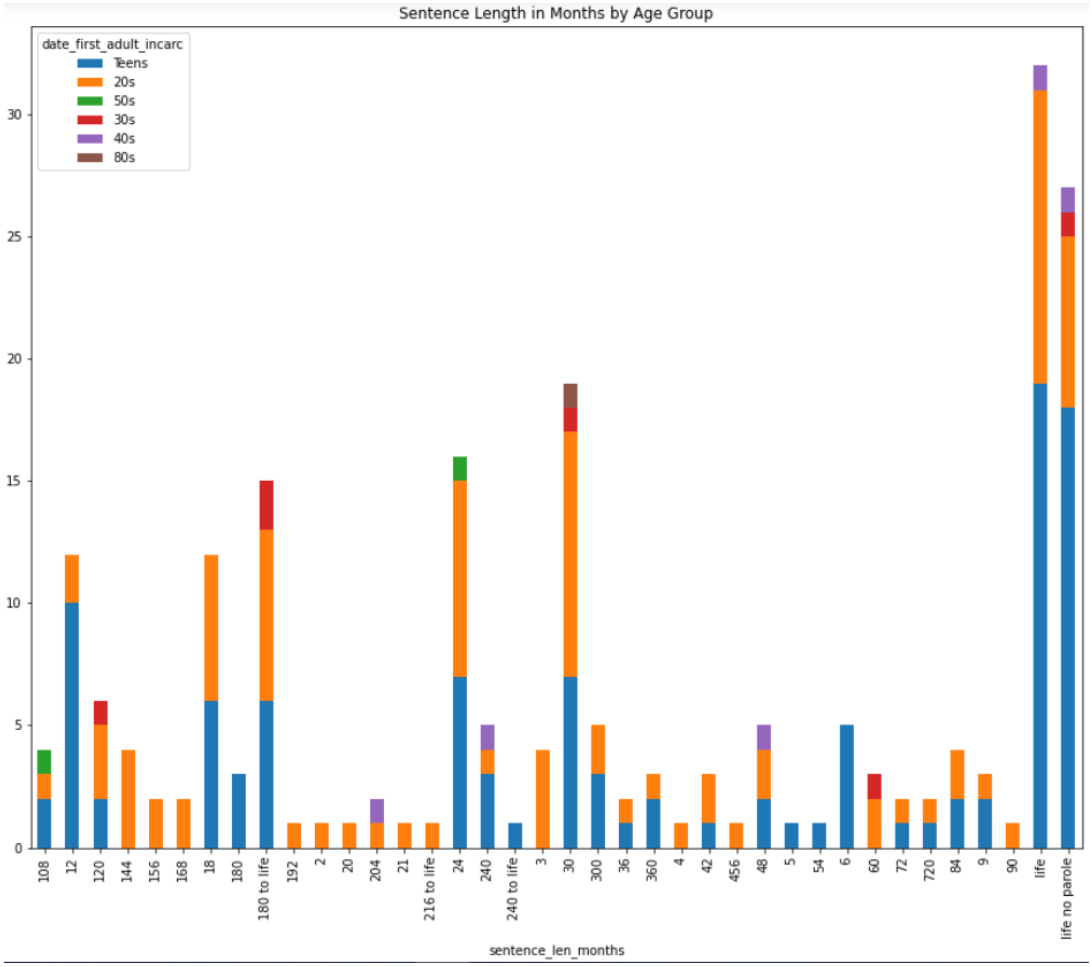
v. Question 5: What is the age distribution of survey respondents and when were they first sentenced / incarcerated? How does this impact their sentence length?[5]

- Full dataset (phase 1 + 2 combined)


Sentenced by Age Group
30s
20s
34.1%
18.2%
0.4%
Teens
0.2%
80s
0.0%
90s
1.8%
70s
5.4%
60s
21.5%
18.4%
40s
50s

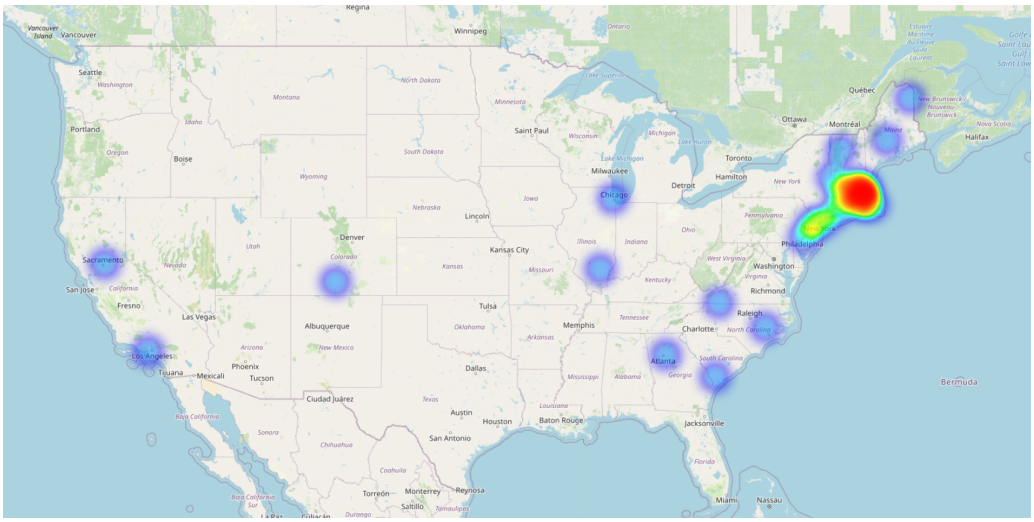

Age when First Sentenced
Teens
43.8%
0.3%
80s
0.0%
90s
0.8%
0.3%
0.3%
50s
60s
70s
2.5%
40s
4.7%
30s
47.4%
20s


Sentence Length
2 Years
1 Year
17.8%
15.0%
3 Years
4 Years
2.8%
Under a Year
5 Years
2.4%
8.5%
6 Years
1.6%
0.8%
7 Years
2.4%
17.4%
31.2%
7+ Years
Life

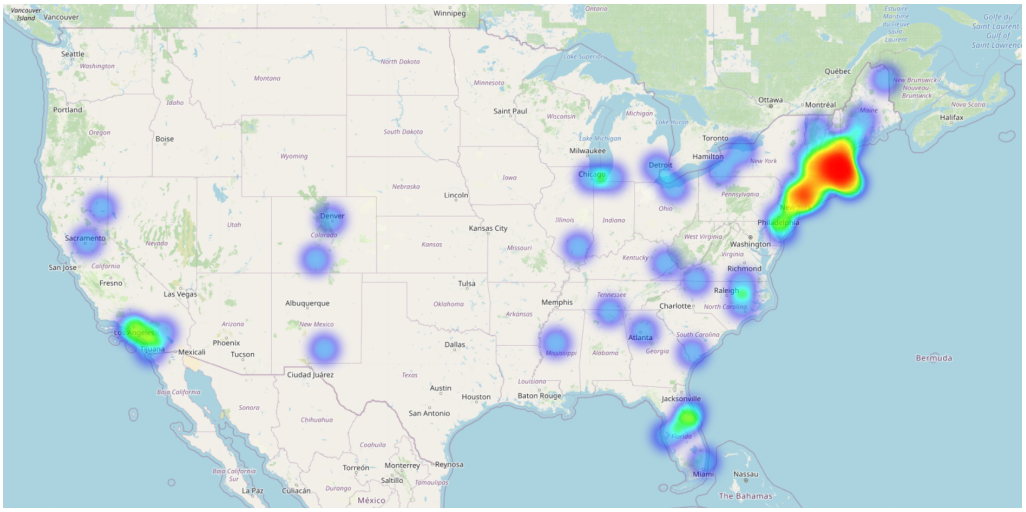


vi. Question 6: Where are most respondents from? [6] (zip code analysis)

- For this question, we will create a HeatMap of the prisoners based on the zip code provided in the questionnaire.
- Phase 2 Data only



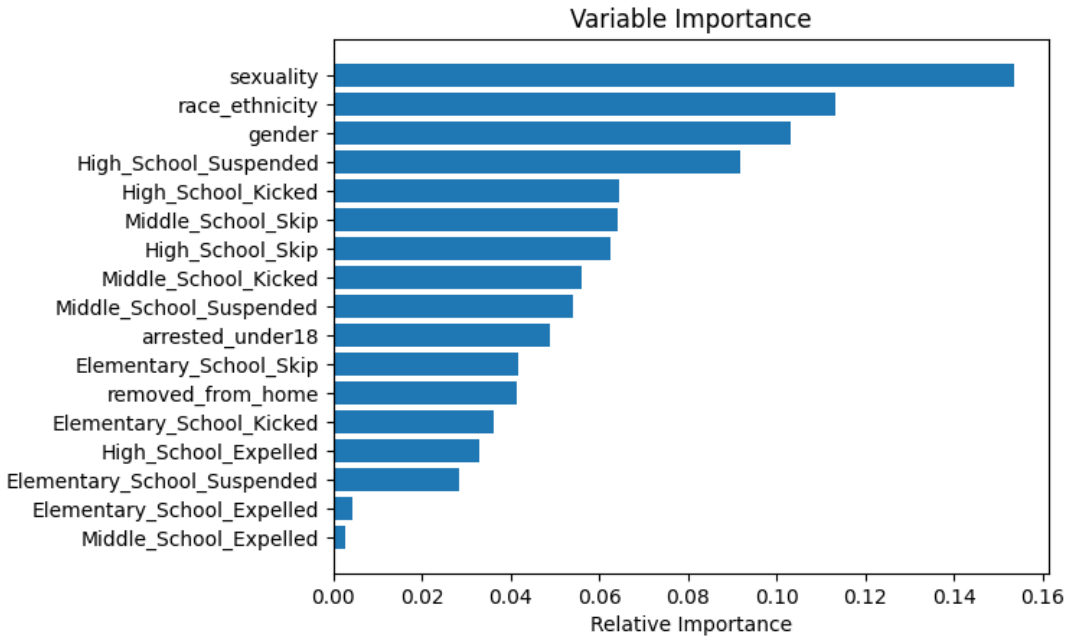
- Full dataset (phase 1 + 2 combined)



c. Machine Learning Analysis

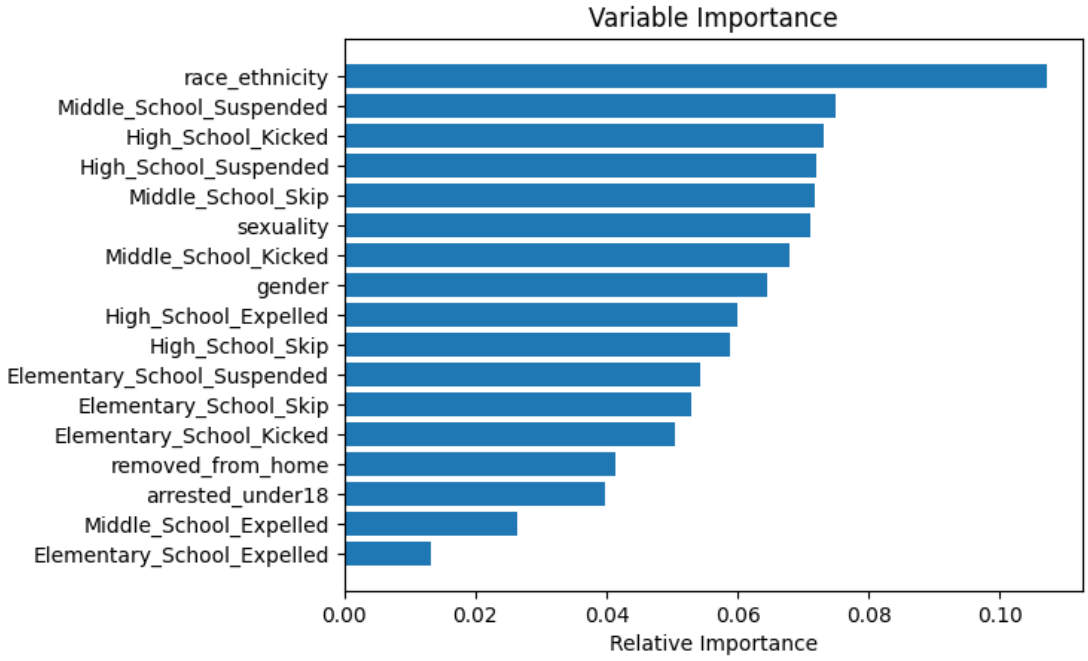
i. Which factors along the pipeline are most predictive of a respondent experiencing incarceration (i.e. arrests before age 18, removal from home, suspension/expulsion, etc.)

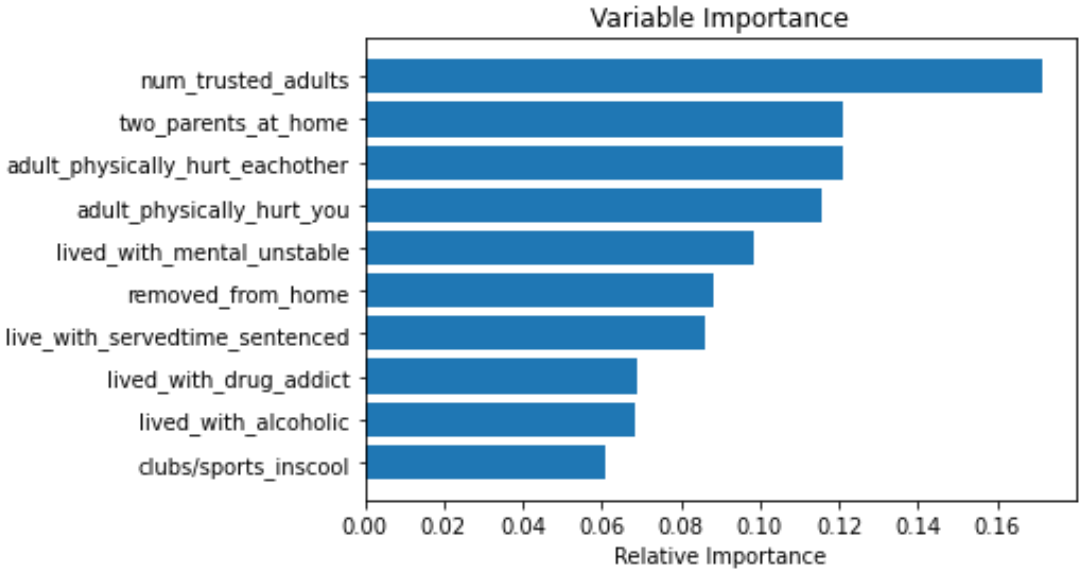
- Upon trying to understand the hypothesis, our team will rephrase the question as: Based on several features (i.e arrests before 18, home removal, etc), can we make a prediction on whether with the future detainers, with a new set of answers, will be detained only or sentenced to prison. And find the most impact features in the current dataset to make those predictions.
- We are basing this hypothesis on our team assumptions of the dataset:
    - Understanding that detained and sentenced is different.
        - Detained means you don't necessarily end up in prison (there is a chance you will).
        - Sentence means you are currently in prison.
    - In these surveys, detainers have a high chance of not being in prison.
- For this hypothesis, we will use Random Forest Classifier and its Feature Importance to determine which features act as the most important in determining the decision of prisoners being incarcerated. Support Vector Machine is also used, however, its feature importances are quite difficult to understand.
- Some different sets of features we are using to feed as training data currently are:
    - Quan's feature set: Gender, Race/Ethnicity, Sexual Orientation, and Elementary, Middle and High school experience.
    - Melody's feature set: Family/Childhood environment
    - Nicole's feature set: Address factors (grew up in certain locations, moved multiple times, etc)
    - Carlos' feature set: Current age / custody, age when first incarcerated, current crime, sentence length
- Further features will be added to tackle the problem stated in the project.

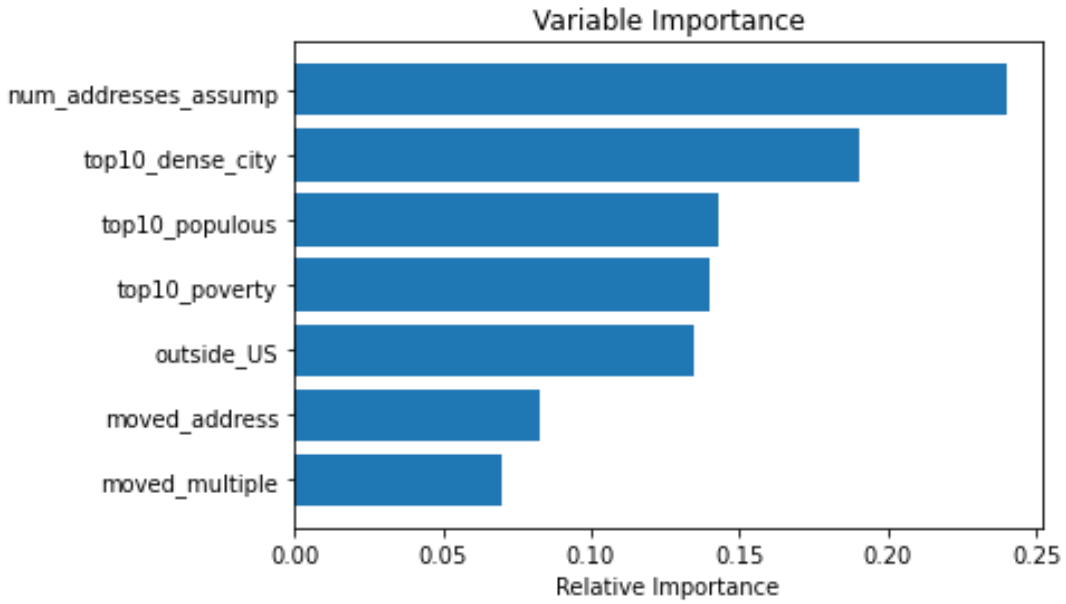


ii. Phase 2 dataset:6.

iii. Full dataset (phase 1 + 2 combined):



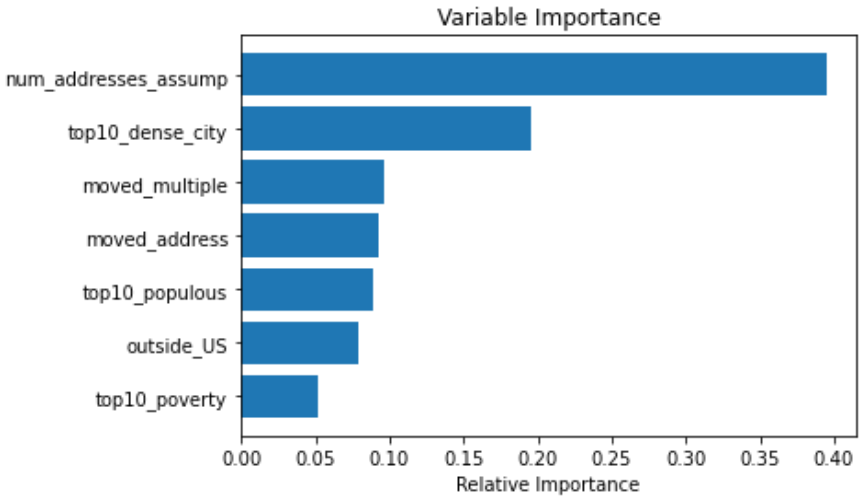


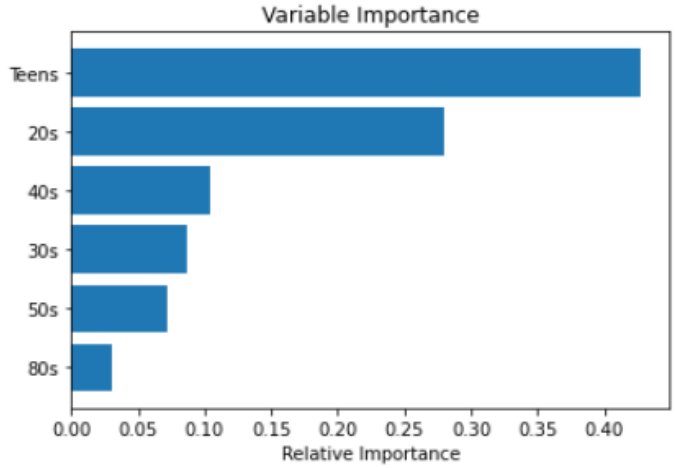


** [4] Notes on Address features:

- Num_addresses_assump is the number of addresses written, with entries such as ‘too many to name/ too many to remember’ being assumed to be 5 for quantification reasons
- Top10_dense_city checks whether zipcode is within one of the top ten most densely populated cities in the US
- Top10_populous checks whether zipcode is within one of the top ten most populated cities in the us
- Top10_poverty checks whether zipcode is in one of the ten zipcodes with the highest poverty rate within the state
- Outside_US checks whether an address outside of the US was reported
- Moved_address checks whether the individual moved addresses at least once
- Moved_multiple checks whether the individual moved addresses more than once

** [4] Additional Address analysis findings to note: Address features can be predictive of sentence length. Particularly, the features below predicted whether one's sentence would be longer than 1 year with 92% accuracy. As seen below, the number of addresses (ie, number of moves) is the most predictive, with the next most important feature being whether the zipcode given is in a top10 densest city. Notably, whether one's zipcode was in the top-10 highest poverty rates for that state does not seem to have an impact on the sentence length.





5. Interpretation

a. Numerical Analysis:

- [1] As we can see from phase 2 dataset only, Sexuality, plays the most important role in the prediction of whether the prisoners are being incarcerated or not.
- [5] As we can see from the combined dataset, in terms of age, only 0.4 percent of survey respondents were teens at the time of the survey. However when looking at when they were first incarcerated we can see that teens jump and make up 43.8 percent of the total. This makes sense when looking at the sentence length chart since 32.8 percent of the total is 1 and 2 years combined, slightly more than some form of life sentence which makes up for 31.2 percent of the total. The life sentence also coincides with the former story of people sentenced in their teens / 20s when comparing it to the chart of when the survey was taken which reflects respondents being in their 30s, 40s and 50s making up almost 3/4 of the total.

b. ML Analysis:
   - [2] As we can see from the full dataset chart above, regarding the importance of School Experience, we can notice there are a few important factors, such as prisoners being kicked out of / suspension rate from middle / high school to determine our model performance.
   - [3] Regarding the Family Childhood environment, the number of trusted adults and having at least 2 parents at home plays an important part in determining whether the prisoners will be sentenced to prison.
   - [4] The number of addresses given (ie, the number of times a person has moved homes in their life) plays a role in determining whether someone will be sentenced. Interestingly, whether someone lives in a high-density city is more significant than whether someone lives in a high-population city. Whether someone moves once versus more than once in their lifetime seems to be less important in determining whether someone will be sentenced. Limitations include the fact that some participants may have chosen to only record one address even if they did move, and also a large amount of participants did not record address at all (or it was not decipherable), so the dataset is rather small.
   - [5] According to the Machine Learning analysis, it would appear that age does in fact impact sentence length, it shows that teenagers and people in their 20s seem to receive longer sentencing and the trend is only broken by those in their 40s who appear to receive harsher sentence lengths compared to persons in their 30s. Aside from that exception, it is safe to assume that being sentenced at a younger age increases the probability that person will receive a longer sentence. Additional investigation would be necessary to determine the reason for this, perhaps teenagers and young adults commit more serious crimes, or repeat offenses, etc.

6. Difficulties

a. Data
   i. Several answers from the questionnaire are quite challenging to analyze, since it does not make sense, data incorrectly computed, or missing data.
   ii. The amount of data is still limited to make confidence predictions on the ML hypothesis.

b. Analysis
   i. Difficulties in expanding the analysis further since the understanding of the prison pipeline from students is limited.

7. Suggestions for the future project:

a. Data
   i. A finalized version of the data cleaned hopefully will be finalized to act as a final dataset for later team use.

b. HeatMap
   i. There are multiple responses from the respondents with various locations, we may suggest to visualize all locations from that respondents as markers and highlight the location marker for each respondent.

c. ML Analysis:
   i. More stronger models hopefully will be utilized to further determine the correct features to better determine the project's hypothesis.
   ii. More features should be added to further determine the most contributing factors to the incarceration rate in prisoners.