# Commercial Sexual Exploitation of Children Service Enhancement Data Analysis & Visualization Project

# Content

# Background

***MACA***

Massachusetts Children's Alliance (MACA) is a statewide non-profit organization, devoted to ending commercial sexual exploitation by promoting interventions, delivering effective educational programming, galvanizing informed and committed legislative support, and mobilizing. This organization has gathered 4 years of de-identified data on a case-by-case basis and aggregated county-level and state-level data as part of the program. The problem to be solved is that the organization has a lot of data but no capability for data analysis and visualization. The initial efforts have been undertaken, but they need a better data visualization on-trend and changes over time to attract their fund and the local community.

***What are PMT and MDT?***

MACA is the coalition of the 12 Children's Advocacy Center (CAC)) across the state, devoted to helping child victims of commercial sexual exploitation. There are two types of services they provide to child victims. One service is PMT, in which CAC provides direct service to a child. Another one is the Multi-Disciplinary Team (MDT), in which professionals of different backgrounds and expertise come together to devise a coordinated approach in the best interest of the child.

The detailed service of PMT and MDT:

| Direct services (PMT) | MDT |
|---|---|
| Information & Referral | Assistant District Attorney |
| Personal Advocacy/Accompaniment | Forensic Interviewer |
| Emotional Support & Safety Services | Police |
| Shelter/Housing Services | Medical Professional |
| Criminal/Civil Justice Assistance | Victim Advocate |
| | Mental Health Professional |
| | Department of Children & Families (DCF) |

# Project Goal

We are a group of MSBA students from Northeastern University to help the abused children in our own power. Cooperating with MACA, we will analyze and explore the difference in race, gender, and age over the past year and the difference between PMT and DMT through data visualization. Meanwhile, we find that limited human resource also is a big problem for this organization. Our team will

allocate their resource and improve the utilization of organization resources by building the predictive models to forecast the new child abuse case in MA.

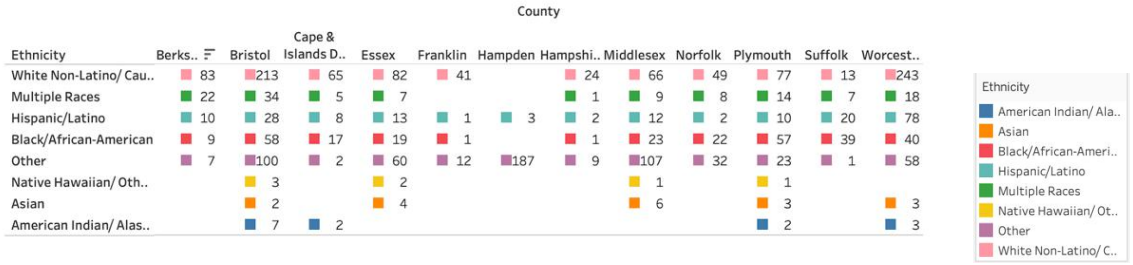
## County Overview

To better understand the situation of child abuse in Massachusetts, our team searched outside resources about the county overview of the child abuse on the internet. From 2012 to 2020, the average number of child abuse in the U.S inclined from 656372 to 618399. The peak of victims is in 2015, which hits 683221 (n.d., Statista). White is the biggest population of the victim of child abuse. The age ranging from 0 to 12 is the largest group that suffer more child abuse case. Overall, the trend of child abuse in the U.S. is decreasing.

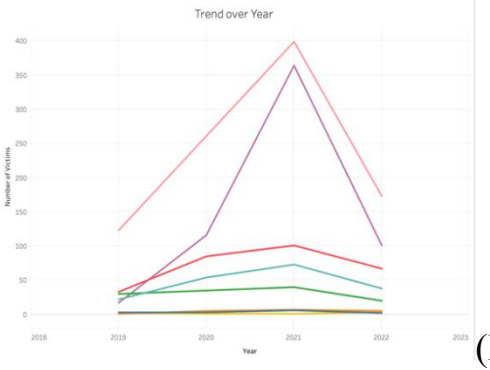
## Data Exploration and Visualization

### *Population Demographic*

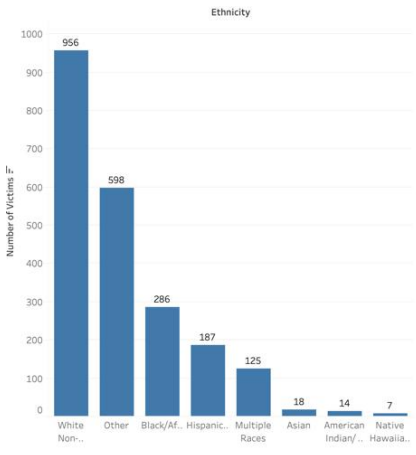
From Figure 1, it's very clear to learn the ethnic composition of victims by county. Bristol County and Plymouth County are the only two counties where victims of all races are covered. Hampden County is the only county without Whites Non-Latino/ Caucasian and Black/African American victims. Asian and American Indian/ Alaska Native victims are the smallest group distributed in Bristol, Essex, Middlesex, Plymouth, and Worcester.




( Figure 1 )

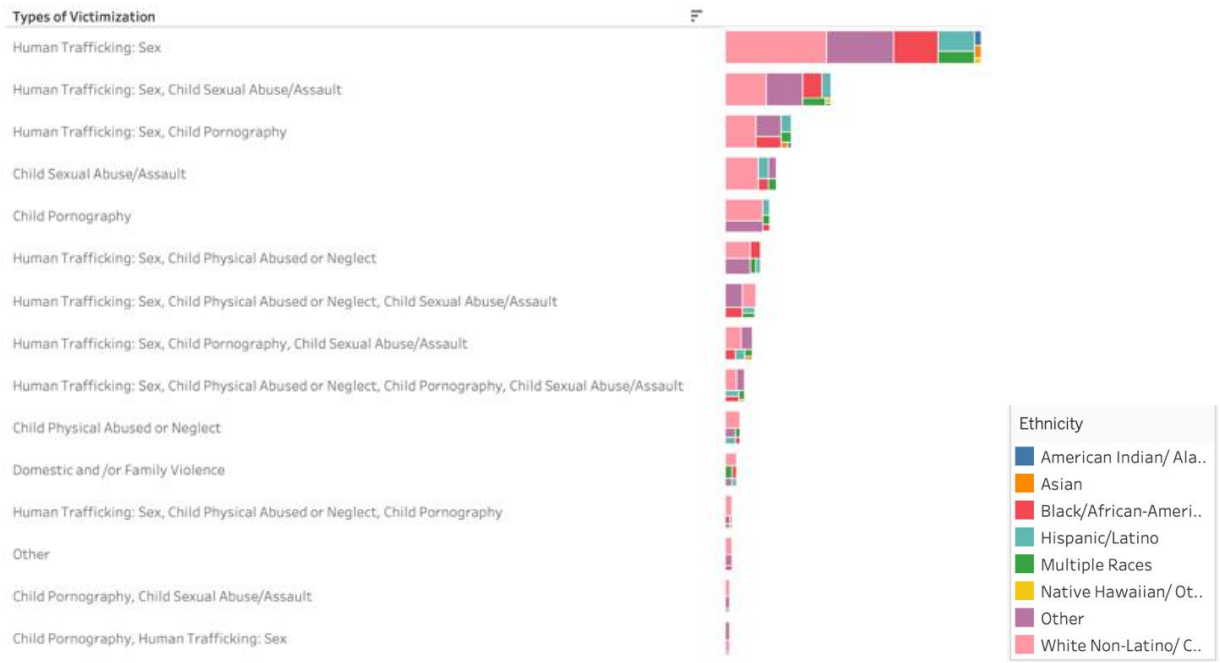



(Figure 2)




(Figure 3)

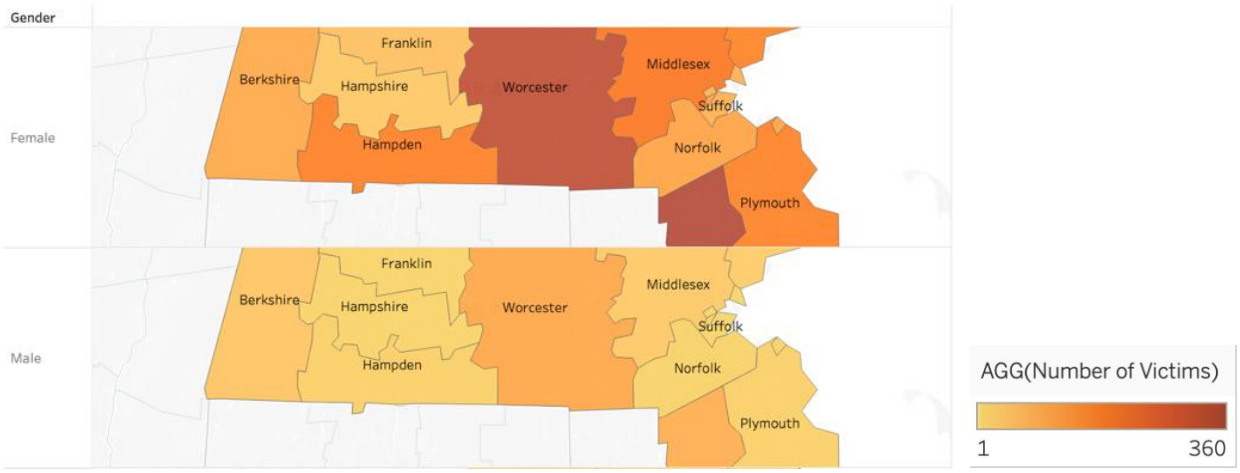
It is plain to see from Figures 2 & 3 that from 2018 to 2022, Whites Non-Latino/ Caucasians are the main victim group. During two years from 2019 to 2020, the number of white Non-Latino/ Caucasian victims soared from 120 to 400, with an average annual increase of more than 80%, the largest increase; the number of blacks and Latino victim groups also increased, with an average annual increase of about 55%, which is relatively small. However, according to the Official racial structure and total population data from the *American Community Survey* and the *United States Census Bureau*, there are 5,460,000 (78%) of White Non-Latino/ Caucasian people in the whole population, although the number of Whites Non-Latino/ Caucasian victims is 956 (44%), which is high. The proportion of White Non-Latino/ Caucasian victims is relatively low compared to the proportion of the whole population. However, compared to Black/African American and multiracial population groups in Massachusetts, which accounted for 700,000 (10%), the proportion of victims, which accounted for 411 (18.7%), is higher than their share of the population, so that we can conclude that black children are suffering higher rates of abuse.




***Victimization types***

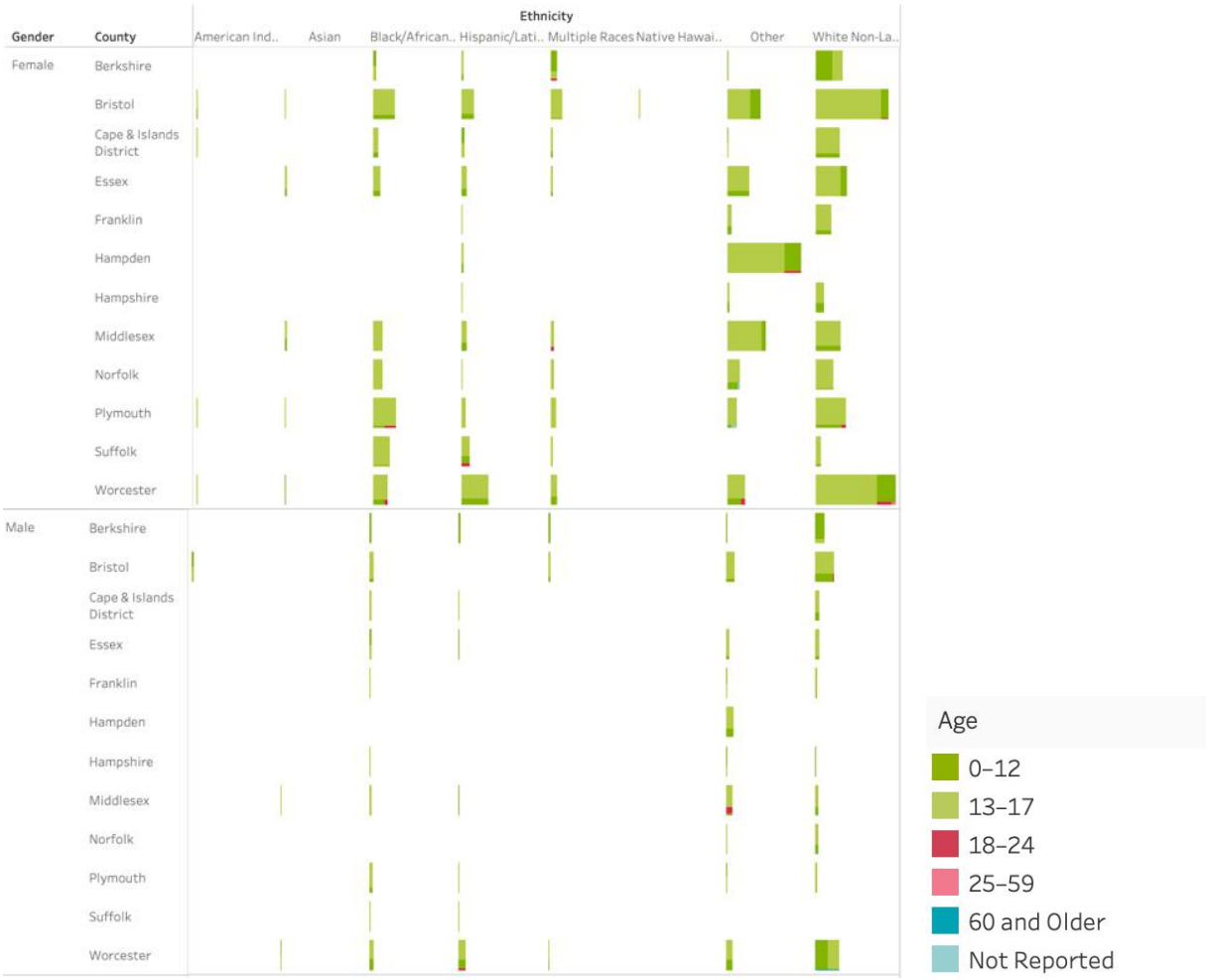
The graph above shows 26 different types of victimization and the racial makeup of each type of victimization. Some victims recorded that they suffered from monotype of abuse, and some suffered from multi-types. Most cases reported are the type of Human Trafficking: sex, which induces a person to be compelled to participate in commercial sex acts, or the person, who has not attained 18 years of

age, is induced to perform such act(s). The specific examples would be commercialized sex, seduction, transporting persons for prostitution, etc. Most victims of sexual abuse are female, according to *Retrospective surveys*, which indicate that 5% to 25% of North American women have experienced some form of childhood sexual abuse, the prevalence of childhood sexual abuse is estimated to be greater for Blacks (8%–19%) than for whites (6%–9%) or Hispanics (5%–8%).




***Victim distribution by gender***

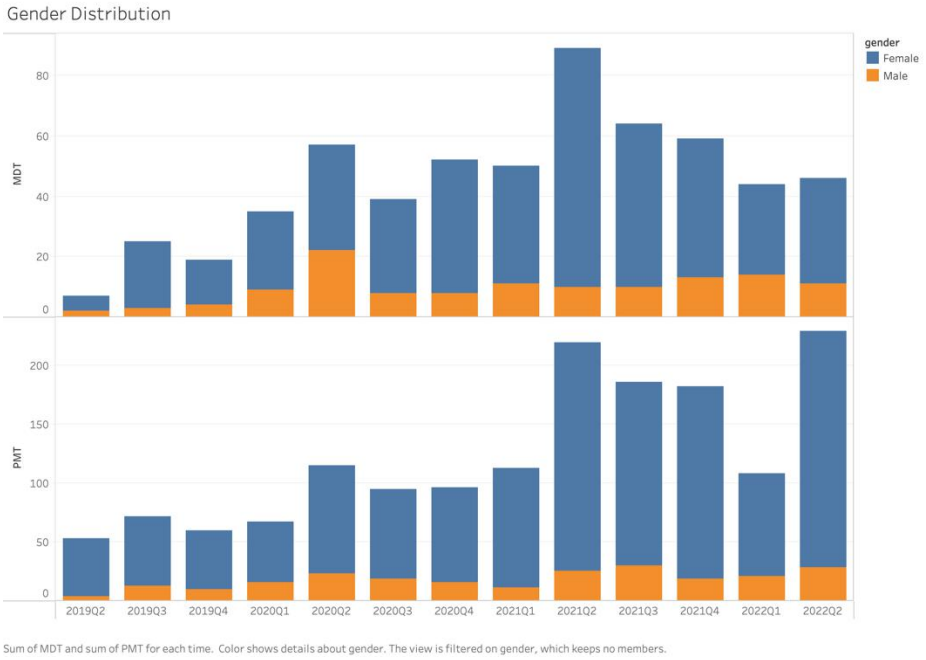
The image above shows the gender distribution of victims across 12 different Massachusetts counties as we mentioned before. The higher is several of victims, the darker are shaded areas. Both male and female victims are mainly concentrated in Worcester and Bristol counties, which have low Per Capita Income within 12 counties in Massachusetts. Also, Worcester ranked No.2, and Bristol ranked No.4 in the list of the least educated counties in Massachusetts. In Worcester County, there are 36.4% of the population 25 years and over with a bachelor's degree or higher, and only 28.7% in Bristol.




***Customer Segment***

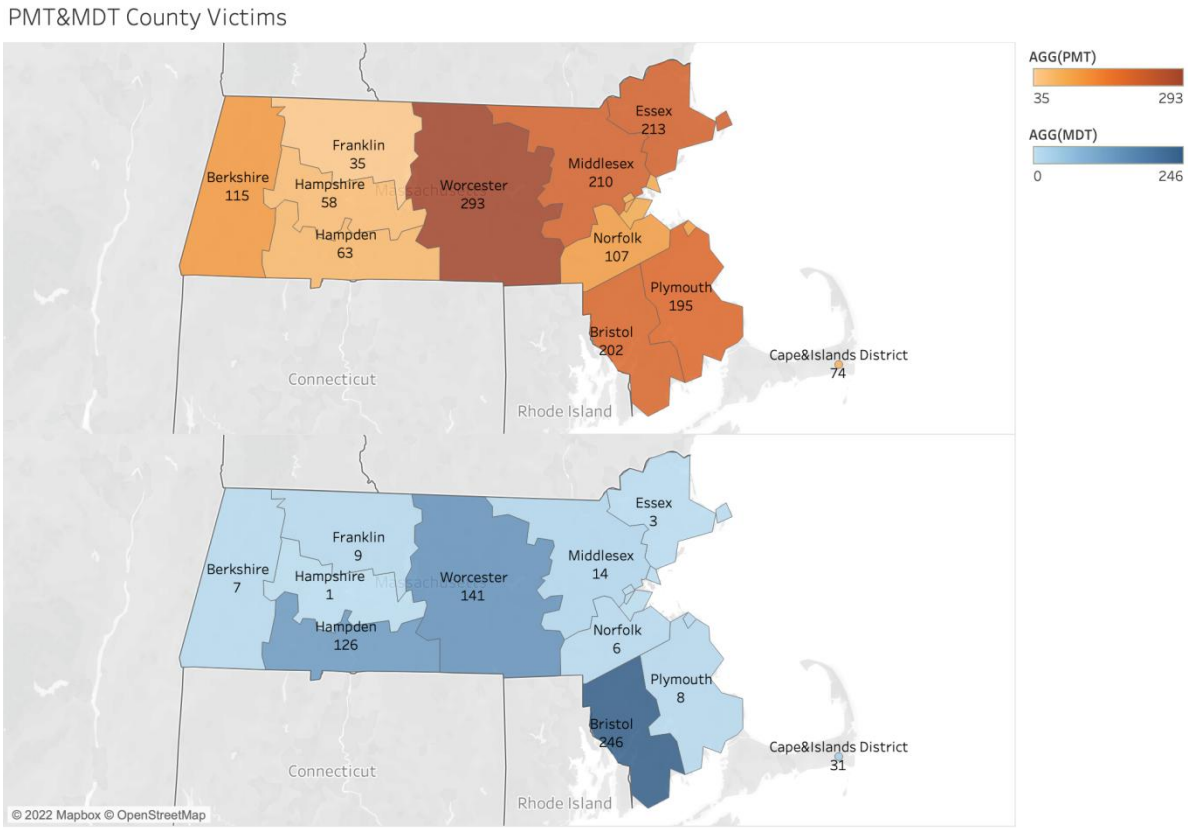
The classified segments, which are for a further recommendation, based on gender, ethnicity, age, and county are shown above. It can be easily seen from the image that among the different ethnic groups of female victims, the main age group of victims is concentrated in 13-17 years old, mainly White Non-Latino/ Caucasian, and Black/African American victims, and concentrated in Bristol and Worcester counties. Among the male victim groups of different races, the age groups of the victims are evenly distributed between 0-12 years old and 13-17 years old, mainly White Non-Latino/ Caucasian victims, and concentrated in Worcester County. Based on the simple analysis above, the two major segments are simply summarized as the first group of people who are 13-17 years old, Whites Non-Latino/ Caucasian and Black/African American female, living in Worcester or Bristol County, and the second group of people who are 0-17 years old, Whites Non-Latino/ Caucasian and Black/African American male, living in Worcester County

***Population variation between PMT and DMT***




From the graph, we can find female victims are way too much than male victims. The most important reason could be that female has less power. Also, women are more tolerant than men, they think more when asking for help, such as family integrity, other people's feelings,
reputation, etc. For gender distribution, the trend is almost the same. Only in 2020Q2, did the male proportion has a sudden increase, which could be the reason for the start period of Covid-19, and all children stay at home for a long period. Otherwise, for both MDT and PMT, the trend is flat, and the perpetrators of verbal violence and behavioral violence (the perpetrators are not divided into men and women, but mainly men) are often targeted by those who the perpetrator thinks lack the ability to fight back or is inconvenient to fight back.

The real strong people will face people who are stronger than themselves, dare to fight against the strong, and will not bully those who are physically weaker than themselves. People transfer their stress to those closest to them in the form of verbal and behavioral violence, whatever the reason, violence in everyday life is not a legitimate reason to exist in this society. Whether you are a man or a woman, opposing violent words and deeds in daily life, opposing domestic violence, protecting children, maintaining healthy and harmonious family relationships, maintaining good social order, and building a harmonious society are the common goals of all men and women.




Map based on Longitude (generated) and Latitude (generated) and Latitude (generated). The marks are labeled by county. For pane Latitude (generated): Color shows sum of PMT. The marks are labeled by county and sum of PMT. For pane Latitude (generated) (2): Color shows sum of MDT. The marks are labeled by county and sum of MDT.

Through this map, we can more intuitively find the huge gap between PMT and MDT in the central and southeast regions. Specially In Worcester and Plymouth, the population of PMT is several times higher than the population of MDT. It is necessary to allocate more MDT resources in those counties.

# Methodology

### *Data Wrangling*

The raw data that comes from the MACA organization consists of 13 children abuse data sets from Oct. 2018 to DEC. 2021, which is a four fiscal year. Different from the data we usually see in data analysis, those data are built on two-dimensionality like vertical/horizontal axis, increasing the difficulty of the data cleaning process.

On the vertical side, it includes three major components, population demographic, direct services, and subgrantee annually reported outcomes. Each component has more detailed attributions. Population demographic has the number of cases reported, ethnicity, gender, age, and type of victimization. The second component has the type of service they provide, which has information & referral, personal advocacy/accompaniment, emotional support & safety services,

shelter/housing service, and criminal/civil justice assistance. The last component is the client feedback process. It includes meeting the client's request, receiving a survey, and completing a survey. Before exploring the horizontal side, we need to provide the background of MACA (Massachusetts Children Alliance). It has 12 programs to solve child abuse cases across Massachusetts and cooperate with others. To classify the difference, they identify MDT cases as direct service by them and MDT cases as coordination only.

To build a model analyzing those data, our team aggregates 13 data sets into one good formal data set. We do not use all variables from the original data set and select three main aspects that are related to predicting child abuse, which are race, gender, and age. In addition, we add one more variable to accurate our research, which is the unemployed rate. According to the CDC's definition and potential factor of child abuse, the unemployment rate influences child abuse (n.d. CDC). Thus, we collect the unemployed rate of each county in Massachusetts from the government website. However, there are many abnormal values in the unemployed rate during the pandemic of covid-19. To increase the accuracy of this data, we transform this data into a logical format. So, we compare the unemployed rate of each country with the average employment rate of Massachusetts. If the value is above the average state, the column would 1. Otherwise, the value would be 0.

A detailed description of our data lists below:

| Variables | Data Type | Description |
|---|---|---|
| Case | Numerical | The quarterly child abuse case is reported by each CACs. (Target variable) |
| CAC | Categorical | CAC stands for Children's Advocacy Center which provides direct service and coordinated service to a child who suffers abuse. |
| Race | Numerical | It includes American Indian/ Alaska Native, Asian, Black/African-American, Hispanic/Latino, Native Hawaiian/ Other Pacific Islander, White Non-Latino/ Caucasian, Some Other Race, Multiple Races, race.not.report, and race.not.track. Each of them is an independent column and stands for the race of the abused child in the quarter. |

| | | |
|---|---|---|
| gender | Numerical | It includes male, female, other, gender.not.report, and gender.not.track. Each of them is an independent column and stands for the gender of the abused child in the quarter |
| Age | Numerical | It includes 0-12, 13-17, 18-24, 25-59, 60 and older, age.not.report, and age.not.track. Each of them is an independent column and stands for the age range of the abused child in the quarter |
| Unemployed rate | Logical | Comparing the county unemployed rate with the average unemployment rate of MA, the logical data is applied. |
| Quarter | Logical | The seasonality trend of the child abuse case is based on quarter. |

Since we put so much effort into creating the data set, our new data set is in a good format with 126 rows and 40 columns. The case (the new child abuse case happened each quarter) is our target variable. The rest of them are independent variables. The only step of our data cleaning is to create the dummy variables for CAC and t quarter.

The detail of our new data set:

| case | CAC | Quarter | n Indian/ Alask | Asian | </African-Amer | -ispanic/Latinc | iian/ Other Pac | Jon-Latino/ Cai | ome Other Rac | Multiple Races | ace.not.r |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 28 | BRI | Q2 | 0 | 0 | 2 | 0 | 1 | 14 | 0 | 11 | 0 |
| 3 | HMPSH | Q2 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| 4 | SUF | Q2 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| 25 | WOR | Q2 | 0 | 0 | 1 | 4 | 0 | 18 | 0 | 0 | 2 |
| 21 | BER | Q3 | 0 | 0 | 0 | 2 | 0 | 12 | 0 | 7 | 0 |
| 26 | BRI | Q3 | 2 | 0 | 6 | 0 | 0 | 15 | 0 | 2 | 1 |
| 5 | CAP | Q3 | 1 | 0 | 0 | 0 | 0 | 3 | 0 | 1 | 0 |
| 7 | HMPSH | Q3 | 0 | 0 | 0 | 1 | 0 | 6 | 0 | 0 | 0 |
| 18 | PLY | Q3 | 0 | 0 | 4 | 1 | 0 | 7 | 0 | 5 | 1 |
| 4 | SUF | Q3 | 0 | 0 | 3 | 1 | 0 | 0 | 0 | 0 | 0 |
| 16 | WOR | Q3 | 0 | 0 | 1 | 4 | 0 | 10 | 0 | 0 | 1 |
| 7 | BER | Q4 | 0 | 0 | 2 | 1 | 0 | 4 | 0 | 0 | 0 |
| 25 | BRI | Q4 | 0 | 0 | 5 | 3 | 1 | 13 | 0 | 3 | 0 |
| 6 | CAP | Q4 | 0 | 0 | 2 | 0 | 0 | 4 | 0 | 0 | 0 |
| 6 | HMPSH | Q4 | 0 | 0 | 0 | 3 | 0 | 3 | 0 | 0 | 0 |
| 11 | PLY | Q4 | 0 | 1 | 3 | 2 | 0 | 5 | 1 | 1 | 0 |
| 7 | SUF | Q4 | 0 | 0 | 5 | 2 | 0 | 0 | 0 | 0 | 0 |
| 17 | WOR | Q4 | 0 | 0 | 1 | 1 | 0 | 7 | 4 | 0 | 4 |
| 5 | BER | Q1 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 1 | 0 |

***Modeling***

To validate the model, our team splits the whole dataset into a training and testing dataset. The test site is about forty percent of the dataset. The cross-validation method cannot be applied since our dataset only has 125 rows. If we applied this method, the data in each k-fold is not enough to use. Three models can be applied to this regression model: linear regression, KNN, and random forest.

The first model we use is linear regression. To begin with, we apply all variables to linear regression. The R-squared of the model is 0.917. It means 91.7% of the new case can be explained by the independent variables in this regression model. As mentioned above, there are dozens of factors related to child abuse. Based on the data set that we have, our team build four dimensionality to create the model. Therefore, the overfitting problem will not exist in our model. Another observation from our model is the coefficient between a dependent variable and independent variables. Based on the results from the model, we find that Americans (Coef:5.7371) have the biggest impact on new child abuse cases in the race and Native Hawaiian is the least one (Coef: -5.9988). And female has a bigger influence on the new child abuse case than male. Essex, Hampshire, Middles and Worcester are the four areas having the largest coefficient with the new child abused case.

The detail OLS Regression Results:

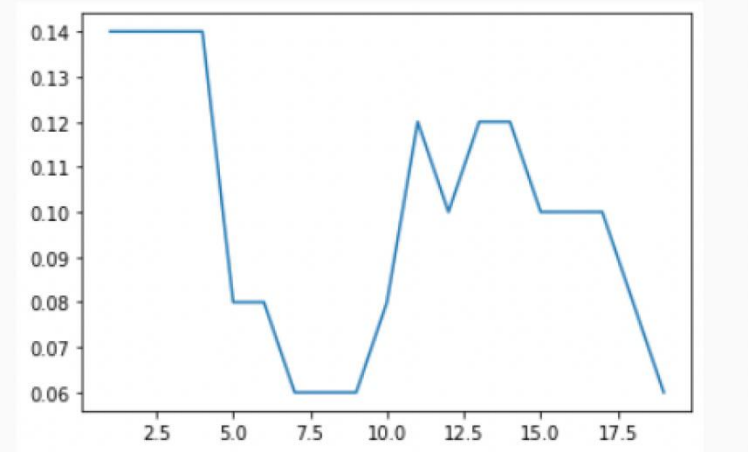
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | case | R-squared: | 0.917 |
| Model: | OLS | Adj. R-squared: | 0.833 |
| Method: | Least Squares | F-statistic: | 10.99 |
| Date: | Fri, 29 Apr 2022 | Prob (F-statistic): | 2.33e-11 |
| Time: | 19:16:18 | Log-Likelihood: | -218.92 |
| No. Observations: | 75 | AIC: | 513.8 |
| Df Residuals: | 37 | BIC: | 601.9 |
| Df Model: | 37 | | |
| Covariance Type: | nonrobust | | |

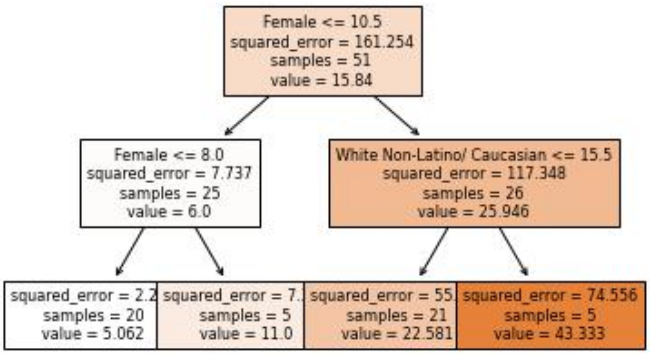
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| American Indian/ Alaska Native | 5.7371 | 7.561 | 0.759 | 0.453 | -9.582 | 21.056 |
| Asian | 2.9265 | 4.821 | 0.607 | 0.548 | -6.842 | 12.695 |
| Black/African-American | 1.9230 | 4.720 | 0.407 | 0.686 | -7.642 | 11.488 |
| Hispanic/Latino | -0.3529 | 4.597 | -0.077 | 0.939 | -9.666 | 8.961 |
| Native Hawaiian/ Other Pacific Islander | -5.9988 | 6.194 | -0.969 | 0.339 | -18.548 | 6.550 |
| White Non-Latino/ Caucasian | 0.8763 | 4.635 | 0.189 | 0.851 | -8.516 | 10.268 |
| Some Other Race | 1.2130 | 4.394 | 0.276 | 0.784 | -7.689 | 10.115 |
| Multiple Races | 1.2139 | 4.695 | 0.259 | 0.797 | -8.299 | 10.727 |
| race.not.report | 0.4082 | 4.548 | 0.090 | 0.929 | -8.807 | 9.623 |
| race.not.track | 2.3804 | 4.642 | 0.513 | 0.611 | -7.025 | 11.785 |
| Male | 0.2302 | 5.071 | 0.045 | 0.964 | -10.045 | 10.505 |
| Female | 1.4841 | 5.146 | 0.288 | 0.775 | -8.942 | 11.910 |
| Other | 1.8312 | 3.067 | 0.597 | 0.554 | -4.383 | 8.045 |
| gender.not.report | 2.3644 | 10.648 | 0.222 | 0.826 | -19.211 | 23.940 |
| gender.not.track | 0.5327 | 0.555 | 0.959 | 0.344 | -0.593 | 1.658 |
| 0-12 | -0.5442 | 6.759 | -0.081 | 0.936 | -14.239 | 13.150 |
| 13-17 | -1.7848 | 6.850 | -0.261 | 0.796 | -15.664 | 12.094 |
| 18-24 | -1.6127 | 7.036 | -0.229 | 0.820 | -15.870 | 12.644 |
| 25-59 | -27.4940 | 15.702 | -1.751 | 0.088 | -59.310 | 4.322 |
| 60 and older | -13.0656 | 12.241 | -1.067 | 0.293 | -37.868 | 11.737 |
| age.not.report | 2.6059 | 10.961 | 0.238 | 0.813 | -19.603 | 24.814 |
| age.not.track | -10.5781 | 16.493 | -0.641 | 0.525 | -43.996 | 22.839 |
| above average | -0.0965 | 2.976 | -0.032 | 0.974 | -6.126 | 5.933 |

The second model is KNN. This model is to calculate similarity by distance. To set the best K-value, our team use the stimulation from 1 to 20. According to the score of model performance, the model is the same when k value ranges from 1 to 4. Comparing with MSE, our team find when k is equal to 4, the model gets better performance.

Score for K-value (x-axis: K-value, y-axis: score):



The third model is Random Forest. This model is to construct a multitude of the decision tree, which can offset some outliers and noise. To get the best parameter for the model, our team use the GridSearchCV method. The result shows that the max depth is 2, the min sample leaf is 5, and the estimator is 10.




***Evaluation***

| model | MSE | MAE | MAPE |
|---|---|---|---|
| Linear Regression | 70.7593 | 5.46771 | 0.469232 |
| Linear Regression with feature seletion | 174.238 | 9.71397 | 0.702601 |
| KNN | 38.24 | 4.2 | 0.28096 |
| randomforest | 18.3096 | 2.96772 | 0.418487 |

In our project, we use three evaluation measures: MSE. MAE, MAPE. The lower score the model has, the better its performance the model is. The random forest defeats the rest of them in MSE and MAE. Also, the MAPE of random forests is not much high. So, a random forest should be the best model for our project.

## Limitation and Suggestion

***Limitations***

When we are doing the model, we find that there are a few potential limitations that could influence the accuracy of the model. The first one is the size of the sample. The new data set we create only has 126 rows. For some CAC, they have a few rows. For instance, the CAC in Essex only has 5 rows, but the CAC in Berkshire has more than 10 rows. This imbalance size of the data would bring bias to the model. The second is the insufficient variables. As talking before, we use race, gender, age, the unemployed rate, and a quarter (the seasonality trend) to build the model. However, the factor related to child abuse is more than twenty

factors. To increase the accuracy of the prediction of the model, more variables should be added.

***Suggestion1***

Communicating with MACA, we find that the way of information transmission between MACA and CAC is excellent. By sending a fixed format excel, MACA collects each CAC's data and arranges them into one data set. This process exists in abundant columns and rows. It is unavoidable to make mistakes. Therefore, MACA should update its way of information transmission. A database system is an ideal choice. Each CAC has an account to update its data in Cloud. And Employees at MACA could automatically manipulate and analyze data. Also, the database cloud is cheap online, even free. AWS, a data-based cloud provider, offers a free database. MACA can outsource this project to professionals. It will not cost much, but it can bring long-term benefits.

***Suggestion 2***

Combined with the above analysis of the age, gender, and race of victims, the portraits of victim teenagers are concentrated in Whites Non-Latino/ Caucasian, and Black/African American, aged 13-17, living in Worcester and Bristol counties. Therefore, according to the conclusion drawn from this article before, we highly recommend that the MACA organization should work closely with the local high schools, colleges, and universities in these two counties to provide more knowledge and information about sex for local teenagers of this portrait group, to improve these teenager's self-protection awareness. Colleges and universities can carry out corresponding education courses around the following suggestions:

1. Children and teenagers need to understand the concept of sexual consent. The premise of sexual activity is based on the premise of mutual consent. Strictly speaking, having sex with a person without their consent can be considered rape or sexual assault.
2. Adolescents and children need to understand that sex is not a means of making money. Filming pornography and engaging in prostitution is strictly illegal.
3. Always take safety precautions to protect yourself when having sex. If you have unprotected sex, you must seek medical attention immediately and ask your doctor to take appropriate measures.
4. If you have sex without consent or are sexually abused, please report to the police in time to protect your rights.

***Suggestion 3***

Given the high proportion of victims in Worcester County and Bristol County, we suggest that MACA can tilt resources towards these two counties more. Specific recommendations are as follows:

1. A big reason for the high number of victims is not knowing how to get help. Based on this, MACA can set up a mobile information service station in the community, where professionals (therapists, police, etc.) can stay and provide counseling services so that more people can know that there is such a channel where they can get help.
2. Open an official account on the current mainstream social media (Instagram, Twitter, etc.), arrange for therapists, police, and other professionals to stream regularly every week to help Adolescents and children and provide online counseling services.

# Reference

1. Centers for Disease Control and Prevention. (2022, April 6). *Risk and protective factors child abuse and neglect violence prevention injury Center CDC*. Centers for Disease Control and Prevention. Retrieved May 1, 2022, from https://www.cdc.gov/violenceprevention/childabuseandneglect/riskprotectivefactors.html

2. *Labor Market Information*. Mass.gov. (n.d.). Retrieved May 1, 2022, from https://lmi.dua.eol.mass.gov/LMI/LaborForceAndUnemployment#

3. Published by Statista Research Department, & 27, J. (2022, January 27). *Child abuse in the U.S. - total number of victims 2020*. Statista. Retrieved May 1, 2022, from https://www.statista.com/statistics/639375/number-of-child-abuse-cases-in-the-us/