Report

Group Members: Hanfei Qi, Anmol Singh, Wuraola Olawole, Ting Lian

## Abstract

In the most recent years, America has seen an increase in the occurrences of hate crimes. These crimes are often based on biases such as religious, race, and sexual orientations and these have become a source of concern to the FBI. In the initial analysis published by FiveThirtyEight, it was shown that income inequality was the main predictor of hate crimes, We worked with a dataset containing details on hate crimes occurring in several states to explore and identify other variables that are associated with hate crimes and predict this outcome.

Our method and approach to exploring these associations started with data exploration. Here we generated a table to observe the datasets and then we also used visual and graphical representation to further explore the distribution of our data. In our modeling approach, we tested the association published by FiveThirtyEight from their findings. We also tested for multicolinearity to assess for highly correlated variables; we performed a stepwise regression procedure to remove variables that did not contribute significantly to our model and checked for interactions. Finally, we performed a model diagnostics and checked for model assumptions and goodness of fit.

In our analysis, we found that there were missing data, we found that the distribution of the outcome was skewed and we also found some outliers. We removed the outliers and did a log transformation for the outcome. Some of the variables in our data were highly correlated e.g unemployment and median household income. We conducted some research using external sources to better understand the relationship between these variables. In our stepwise regression procedure, we did a backward elimination which left us with only three predictors: gini index, percent popopulation with highschool degree, and unemployment. In checking for interactions we found 3 significant 2-way interactions but upon further analysis(stratification) and tests they were not statistically significant. Finally in our model diagnostics we found that

conclusion

## Introduction

Hate crimes in the United States has become a severe problem and their occurrence has been rising in recent years [[1]](https://www.bbc.com/news/world-us-canada-54968498). According to the FBI, a hate crime is a “criminal offense against a person or property motivated in whole or in part by an offender’s bias against a race, religion, disability, sexual orientation, ethnicity, gender, or gender identity.”[[2]](https://www.fbi.gov/investigate/civil-rights/hate-crimes) These types of crimes can have lasting impact and cause devastating effects to people due to the horrific nature of the crimes, which is why they are the highest priority of the FBI’s civil rights program [[2]](https://www.fbi.gov/investigate/civil-rights/hate-crimes).

A previous study used data from FBI and a self-reported survey to analysis the association between hate crime rates (outcome) and different variables (potential predictors) [[3]](https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/). The author concluded that the income inequality was the most significant predictor of hate crime [[3]](https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/). In this project, our goal is to use the author’s data to build our own model, check if the author’s conclusion is correct, as well as assess if any other factors may play a role in hate crime occurrence. Potential predictors include level of state unemployment (low/high), level of state urbanization (high/low), median household income per state, percentage of adults (>25 yrs.) with a high school degree, percentage of population that are not US citizens, percentage of population that are non-white, Gini index that measuring income inequality (range 0-100). Finding out what factors play a role in the occurrence of hate crimes may be able to help us curb the incidence of these horrible crimes in the United States.

## Methods

### Data Exploration

There are 8 variables in the dataset. Numerical variables are: hate\_crimes\_per\_100k\_splc, median\_household\_income, perc\_population\_with\_high\_school\_degree, perc\_non\_citizen, gini\_index, and perc\_non\_white, while categorical variables include: unemployment, and urbanization. Both categorical variables contain two levels: low, and high. All coding process was done by using RStudio.

Firstly, we generated a descriptive statistics table (Table 1). This included the mean, standard deviation (SD), median, 25% quantile (Q1), 75% quantile (Q3), minimum value, maximum value, and count of missing values for each numerical variable. For categorical variables, we obtained counts of each level, and a count of missing values.

Secondly, we generated a density plot of outcome to show its distribution by using the ggplot function (Figure 1). We, furthermore, used the boxcox function to find the optimal transformation of the outcome and then double-checked the distribution of transformed outcome (Figure 2,3).

Finally, we generated a scatter plot of hate\_crimes\_per\_100k\_splc versus state, from low hate crime rate to high crime rate to observe any potential outliers (Figure 4).

### Modeling

#### Testing Original Association

We first wanted to check if the original association presented in the prior study on hate crimes [3] holds true. We thus performed a linear regression analysis with hate\_crimes\_per\_100k\_splc as the response variable and gini\_index as the predictor. For the original data we found that the association was significant at a threshold of 0.05. However, for the log transformed data we found that the significance decreased as the relationship was not significant at a threshold of 0.05. We thought this could be a problem with model diagnostics so we checked them using graphical displays. Doing this we found definite outliers that are affecting the association from the leverage plots. We confirmed these outliers by using the standard formula of a lower outlier being less than Q1 - 1.5(IQR) and an upper outlier being greater than Q3 - 1.5(IQR) where IQR is the interquartile range. Doing this we confirmed that the District of Columbia and Oregon were outliers. We then removed the outliers and tested the original association again. The model assumptions were met when the outliers were deleted and thus we proceeded to check for multicolliniarity in the model.

#### Checking Multicolliniarty

To check for multicolliniarity in the model we started by creating a correlation matrix of all the variables that could be used in the model and then isolating the pairs of variables that had a correlation above 0.6. This returned three pairs of variables that were highly correlated: . Furthermore, we conducted some research to figure out why these variables are so highly correlated and found sources that support the relationships we saw. Now that we know that there are highly correlated variable pairs in the model we need to perform a stepwise regression procedure to eliminate variables that do not contribute to our model.

#### Stepwise Regression Procedure:

For this procedure we started out with a model with all possible predictor variables and used the R function step in the backward direction to eliminate non-essential predictors one by one based on their AIC value. The final model we ended up with used unemployment, perc\_population\_with\_high\_school\_degree, and gini\_index as predictors.

#### Checking Interactions:

After finding our final model we checked for interactions between the predictors in our model to see how that affects the associations between the predictors and the outcome.

#### Model Diagnostics

For model diagnosis, we plotted four diagnosis plots of our final model to check the assumptions. We also checked the existence of outliers using the studentized residuals. Since there is no candidate for outlier, we did not further explore the existence of the influential point. Finally, the existence of multicollinearity was checked by the variance inflation factor (VIF).

## Results

### Data Exploration

Table 1 showed that there were 4 NA’s in the variable hate\_crimes\_per\_100k\_splc and 3 NA’s in the variable perc\_non\_citizen. The low level of unemployment and urbanization were similar, about 50% across 51 states. The distribution of the outcome variable was highly skewed to right (Figure 1). The box-cox transformation indicated that a natural logarithm transformation should be applied to the outcome variable (Figure 2). The distribution of transformed outcome variable was approximately normal (Figure 3). The scatter plot indicated that data from District of Columbia and Oregon could be potential outliers (Figure 4).

### Modeling

#### Checking Original Association

When checking the association between income inequality and hate crime rate using the original data, we found that there is an significant linear relationship exists with p-value equals 0.024 (Table3). However, in the data exploration part, we found that the original data is heavily left skewed. So we also check the association using the log transformed data. The p-value for the association using transformed data was 0.311 (Table 4) indicating that the linear association between income inequality and hate crime rate is not significant.

#### Checking Multicollinearity

We found that studies have shown that high income is correlated with increased education and thus it would make sense that the median income and percentage of high school diploma holders are highly correlated [[4]](https://budgetmodel.wharton.upenn.edu/issues/2016/2/22/education-and-income-growth). Furthermore, a study conducted by the pew research center found that only 17.7% of immigrants are white non-hispanic which makes sense why the percentage of non citizens and the percentage of white people are very highly correlated as well [[5]](https://www.pewresearch.org/hispanic/2020/08/20/facts-on-u-s-immigrants-current-data/).

#### Stepwise Regression Procedure

<<<<<<< HEAD

#### Checking Interactions

We explored all 2-way interactions among all the variables. Here, we observed that there were three different interactions: median household income and urbanization (Figure 5), percentage of population with highschool degree and urbanization (Figure 6), median household income and gini index (Figure 7). Upon stratified analysis we found that 2 of the interactions were not significant, only percentage of population with highschool degree and urbanization was significant. We performed further test (ANOVA) on this interaction and found no statistical evidence of an interaction (Table 2).

#### Model Diagnostics

We plotted the four diagnosis graphs after we got the final model (Figure 8). Points in the Residuals vs. Fitted plot show a random patter and are evenly spread above and below the line of 0. The red line is approximately horizontal and is bouncing around the line of 0. This graph shows that this model fit the assumption of homoscedasiticity. Similar pattern also shows in the Scale-Location plot indicating that the variance of the model is equal. No point in the Residual vs. Leverage plot is beyond the boundary of Cook’s distance so we could assume that there is no significant influence point. In the Normal Q-Q Plot, all points align in an approximately straight line with no significant departure which indicates that the residuals are normal. Overall, these four graphs show that this fitted model is good for represent the data. There is no observation has absolute studentized residual value greater than 2.5. So we conlcude that there is no outlier in Y in the data and assume that there is no influential point. The VIF values for the three predictors are 1.335 for unemploymentlow, 1.806 for gini\_index and, 1.971 for perc\_population\_with\_high\_school\_degree (Table 5). None of the predictor has variance inflation factor (VIF) value greater than 5 indicating that there is no multicollinearity in the final model.

## Conclusion/Discussion

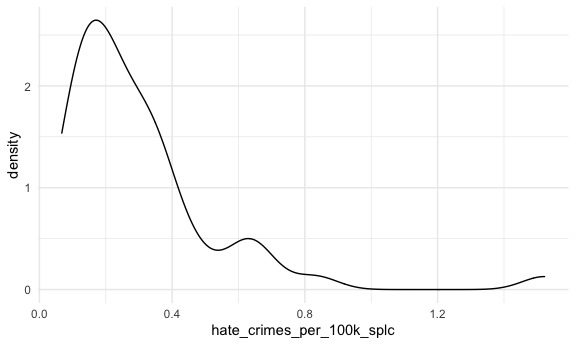
The final model included unemployment, perc\_population\_with\_high\_school\_degree. and gini\_index as predictors, without any interactions. We concluded that gini\_index was not the only main predictor.

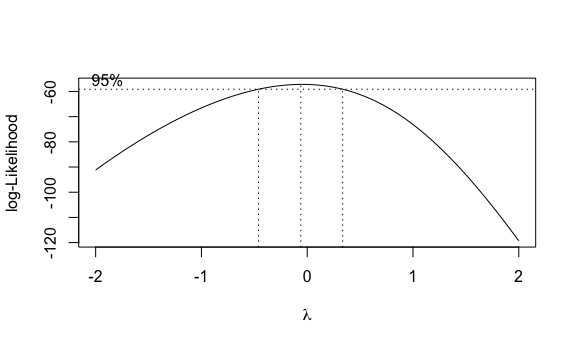
## References

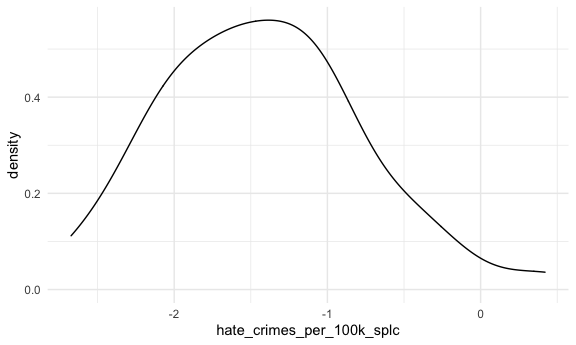
## Appendix

Table 1: Descriptive Statistics

|  |  |
| --- | --- |
|  | Overall (N=51) |
| hate\_crimes\_per\_100k\_splc |  |
| - Mean (SD) | 0.304 (0.253) |
| - Median (Q1, Q3) | 0.226 (0.143, 0.357) |
| - Min - Max | 0.067 - 1.522 |
| - Missing | 4 |
| unemployment |  |
| - high | 24 (47.1%) |
| - low | 27 (52.9%) |
| - Missing | 0 |
| urbanization |  |
| - high | 24 (47.1%) |
| - low | 27 (52.9%) |
| - Missing | 0 |
| median\_household\_income |  |
| - Mean (SD) | 55223.608 (9208.478) |
| - Median (Q1, Q3) | 54916.000 (48657.000, 60719.000) |
| - Min - Max | 35521.000 - 76165.000 |
| - Missing | 0 |
| perc\_population\_with\_high\_school\_degree |  |
| - Mean (SD) | 0.869 (0.034) |
| - Median (Q1, Q3) | 0.874 (0.841, 0.898) |
| - Min - Max | 0.799 - 0.918 |
| - Missing | 0 |
| perc\_non\_citizen |  |
| - Mean (SD) | 0.055 (0.031) |
| - Median (Q1, Q3) | 0.045 (0.030, 0.080) |
| - Min - Max | 0.010 - 0.130 |
| - Missing | 3 |
| gini\_index |  |
| - Mean (SD) | 0.454 (0.021) |
| - Median (Q1, Q3) | 0.454 (0.440, 0.467) |
| - Min - Max | 0.419 - 0.532 |
| - Missing | 0 |
| perc\_non\_white |  |
| - Mean (SD) | 0.316 (0.165) |
| - Median (Q1, Q3) | 0.280 (0.195, 0.420) |
| - Min - Max | 0.060 - 0.810 |
| - Missing | 0 |

 Figure 1: Distribution of hate\_crimes\_per\_100k\_splc.

 Figure 2: Box-Cox transformation.

 Figure 3: Distribution of hate\_crimes\_per\_100k\_splc after applied logarithm transformation.

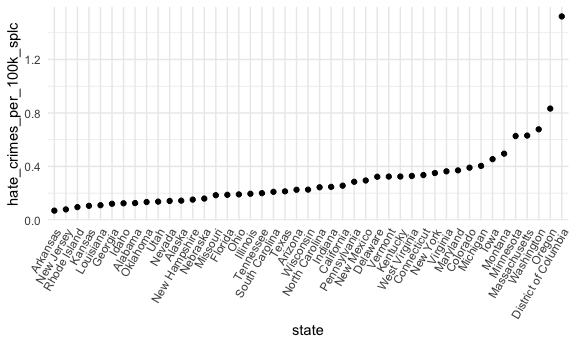
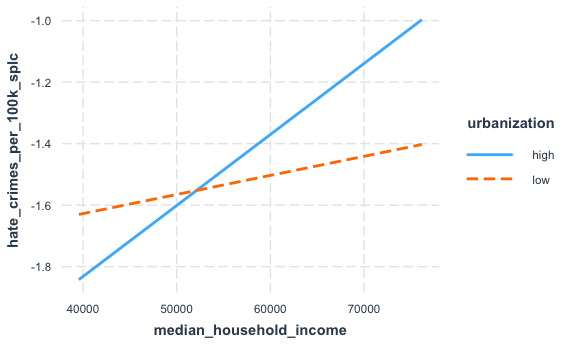
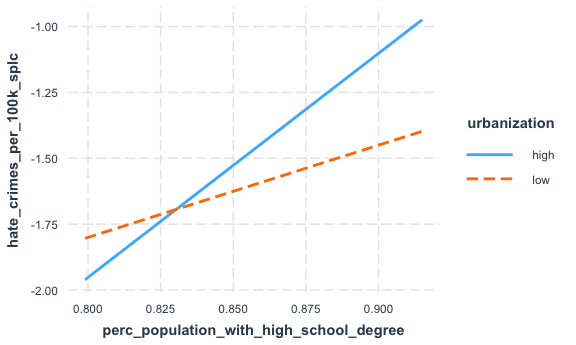
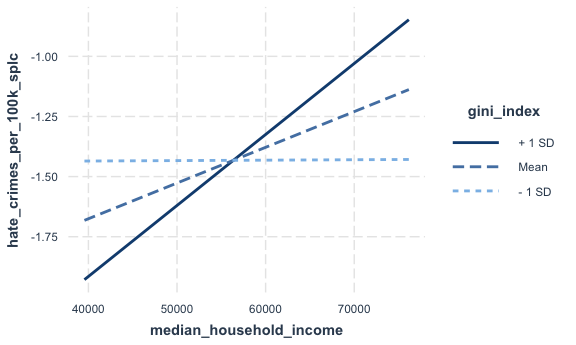
 Figure 4: Scatter plot of hate\_crimes\_per\_100k\_splc of each state, from lowest hate crime rate to highest crime rate.

Table 2: ANOVA Test, model with interaction

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| term | df | sumsq | meansq | statistic | p.value |
| perc\_population\_with\_high\_school\_degree | 1 | 1.2012907 | 1.2012907 | 4.0336007 | 0.0515635 |
| urbanization | 1 | 0.3131419 | 0.3131419 | 1.0514437 | 0.3114931 |
| perc\_population\_with\_high\_school\_degree:urbanization | 1 | 0.2719223 | 0.2719223 | 0.9130397 | 0.3451949 |
| Residuals | 39 | 11.6150159 | 0.2978209 | NA | NA |

 Figure 5: An interaction plot, showing the interaction between median household income and urbanization

 Figure 6: An interaction plot, showing the interaction between percentage of population with highschool degree and urbanization

 Figure 7: An interaction plot, showing the interaction between median household income and gini index

# use this code to do work count!  
# wordcountaddin::text\_stats("Report.Rmd")

Table 3: Testing Association Between Income Inequality & Hate Crime Rate Using Original Data

income\_hate\_model\_full\_data = crime\_df%>%  
 lm(hate\_crimes\_per\_100k\_splc~gini\_index,data=.)  
  
income\_hate\_model\_full\_data%>%  
 broom::tidy()%>%  
 knitr::kable(caption = "Testing Association between Income Inequality and Hate Crime using all the data", format = "html")

Testing Association between Income Inequality and Hate Crime using all the data

term

estimate

std.error

statistic

p.value

(Intercept)

-1.527463

0.7833043

-1.950025

0.0574197

gini\_index

4.020510

1.7177215

2.340606

0.0237445

Table 4: Testing Association Between Income Inequality & Hate Crime Rate Using Transformed Data

income\_hate\_model\_trans = crime\_df%>%  
 mutate(  
 hate\_crimes\_per\_100k\_splc = log(hate\_crimes\_per\_100k\_splc)  
 )%>%  
 lm(hate\_crimes\_per\_100k\_splc~gini\_index,data=.)  
  
income\_hate\_model\_trans%>%  
 broom::tidy()%>%  
 knitr::kable(caption = "Testing Association between Income Inequality and Hate Crime using log transformed data", format = "html")

Testing Association between Income Inequality and Hate Crime using log transformed data

term

estimate

std.error

statistic

p.value

(Intercept)

-3.675547

2.195289

-1.674288

0.1010115

gini\_index

4.931538

4.814087

1.024398

0.3111231

final\_rec = lm(formula = hate\_crimes\_per\_100k\_splc ~ unemployment + perc\_population\_with\_high\_school\_degree + gini\_index, data = crime\_trans)  
plot(final\_rec)

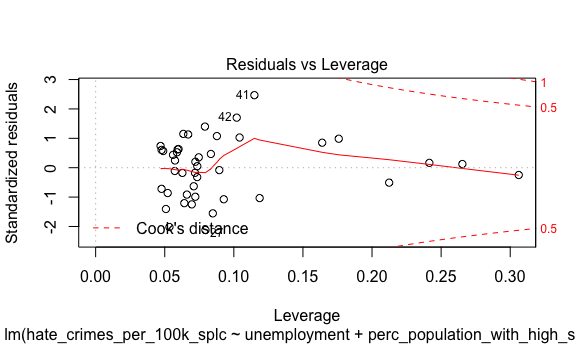
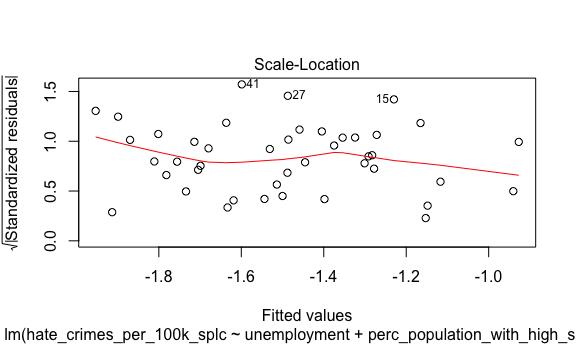
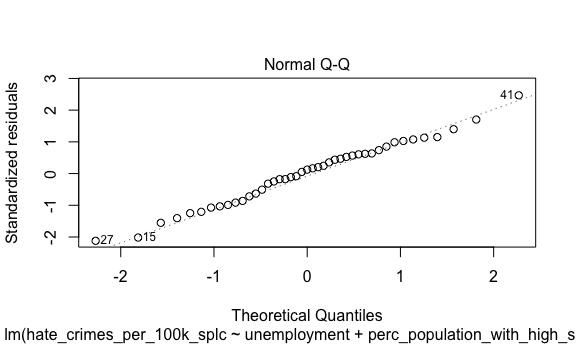
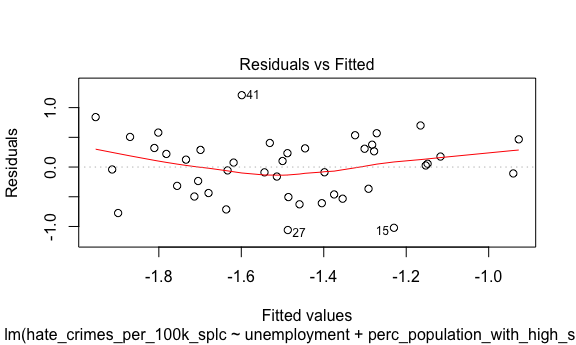
 Figure 8: Four Diagnosis Graphs For The Final Model

Table 5: The VIF Values For The Final Model

vif(final\_rec)

## unemploymentlow perc\_population\_with\_high\_school\_degree   
## 1.335118 1.970627   
## gini\_index   
## 1.806199