Report

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

## Abstract

## Introduction

Hate crime is a type of offense based on discrimination. Offender of a hate crime is motivated in whole or in part by his/her bias against a race, disability, ethnicity, gender, religion, sexual orientation, or gender identity (<https://www.fbi.gov/investigate/civil-rights/hate-crimes>).

Since hate crimes have a significant effect to families and communities compare to most type of crimes, they become the highest priority of the FBI’s civil rights program (<https://www.fbi.gov/investigate/civil-rights/hate-crimes>).

An article used data from FBI and a self-reported survey to analysis the association between hate crime rates (outcome) and different variables (potential predictors). The author concluded that the income inequality was the most significant predictor of hate crime. In this project, our goal is to use the author’s data to build our own model and check if the author’s conclusion is correct. 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).

## 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 level: low, and high. All coding process was done by using RStudio.

First, generated a descriptive statistics table to observe the data. Obtained 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, obtained count of each level, and count of missing values.

Second, generated density plot of outcome to show its distribution by using ggplot function. Used boxcox function to find the optimal transformation of the outcome, then double-checked the distribution of transformed outcome.

Finally, generated a sacatter plot of hate\_crimes\_per\_100k\_splc versus state, from low hate crime rate to high crime rate, so we could observe any potential outliers.

### Modeling

### Model Diagnostics

## Results

### Data Exploration

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

### Modeling

### Model Diagnostics

## Conclusion/Discussion

## References

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