P8130 Final Report

Amanda Tsai, Na Yun Cho, Rosemary Hahm, Ruwen Zhou, Yubei Liang

Abstract

In today's world where diversity abounds in every aspect of people's lives, hate crimes still remain a big issue that leave devastating effects on not only individuals but also communities as well. To better address and prevent future hate crimes, this study aims to identify variables that are most closely associated with hate crime rates. In doing so, we examined the data in every U.S. state that were recorded during the first weeks of November in 2016 by the Southern Poverty Law Center and examined a variety of potential factors that could be associated with hate crimes. With previous knowledge that income inequality is one of the main predictors of hate crime rates, we looked more into this factor and assessed its relationship with other variables. Through various model selections and statistical analyses, we concluded that on average, hate crime rate in the U.S. is linearly associated with an increase in the percentage of adults with a high school degree and a higher index of income inequality. The association between hate crime rate and income inequality in addition to the percentage of adults with a high school degree was stronger than that between hate crime rate and income inequality alone. Based on these results, future studies can look into identifying more factors that are closely related with hate crime rates globally and in the U.S. over the years.

Introduction

The current highest priority of the FBI's civil rights program is hate crimes. A hate crime, as defined by the FBI, 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."(FBI, n.d.). The number of hate crimes committed yearly in the United States has been growing and, as of 2020, has risen to the highest level in more than a decade, with 7,134 reported cases from 2019 (Balsamo, 2020). This number could be severely lower than the actual count, as hate crime data is voluntarily reported by law enforcement and only 2,172 out of the 15,000 participating agencies reported to the FBI last year (Balsamo, 2020). However, with the increasing incidence of hate crimes, there is a growing urgency to find trends within the hate crime data that can assist law enforcement agencies in addressing potentially problematic issues or provide lawmakers with justification for certain legislation and aid the detection and prevention of future incidents.

10 days after the 2016 election, more hate crimes were reported to the Southern Poverty Law Center on average per day than in the time between 2010 and 2015 (Majumder, 2017). Using the data reported in this time frame, which includes details on hate crimes that occurred in the United States by state, we seek to address the strength of association between a variety of potential variables and the incidence of hate crimes. The variables include the levels of unemployment, level of state urbanization, the median household income per state, percentage of adults over the age of 25 with high school degrees, the percentage of the population that are non-us citizens, the percentage of the population that are non-white, and the Gini index number that measures income inequality for each state (Majumder, 2017)

```
# Load libraries
rm(list = ls())
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(arsenal)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(boot)
```

```
# Load and tidy data, missing values of outcome variable are removed
hatecrimes_df <- read.csv("./data/HateCrimes.csv") %>%
 filter(hate_crimes_per_100k_splc != 'N/A') %>%
  mutate(
    unemployment = as.factor(unemployment),
    urbanization = as.factor(urbanization),
    hate_crime_rate = as.numeric(hate_crimes_per_100k_splc)) %>%
  select(-hate_crimes_per_100k_splc)
```

```
# Summary statistics
my_labels<-list(hate_crime_rate = "Hate Crimes(per 100k)", unemployment = "Unemplotment Level", urbaniz

#make controls
my_controls <- tableby.control(
               total = F,
               test=F,  # No test p-values yet
               numeric.stats = c("N", "meansd", "medianq1q3", "min", "max", "Nmiss2"),
               cat.stats = c("N", "countpct"),
               stats.labels = list(
                     meansd = "Mean (SD)",
                     medianq1q3 = "Median (Q1, Q3)",
                     min = "Min",
                     max = "Max",
                     Nmiss2 = "Missing",
                     countpct = "N (%)"))

#table 1
tab1<-tableby(~ hate_crime_rate + unemployment + urbanization + median_household_income + perc_populati
summary(tab1, title = "Descriptive Statistics: Hate Crimes per 100K Population and Possible Influential
```

```
##
##
## Table: Descriptive Statistics: Hate Crimes per 100K Population and Possible Influential Variables
```

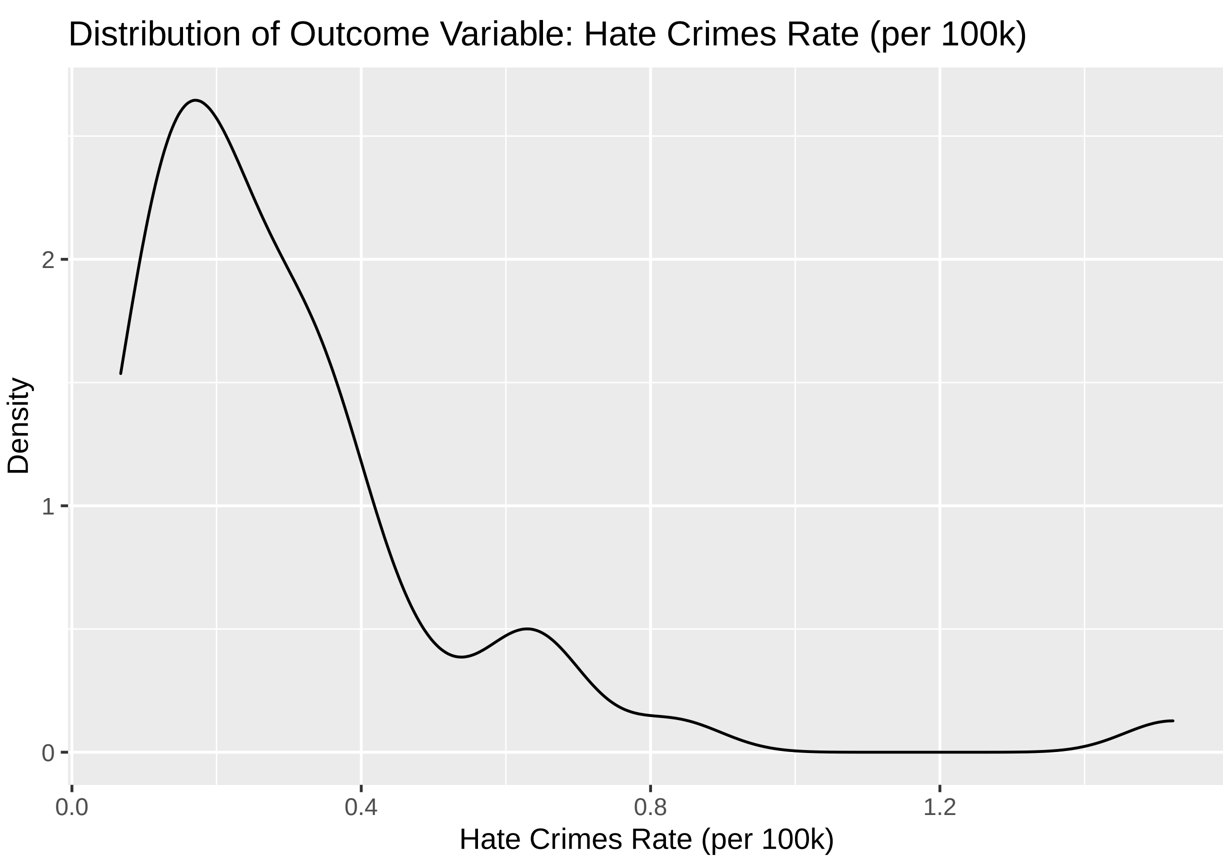
```
##
## |                                 |          Overall (N=47)          |
## |:--------------------------------|:--------------------------------:|
## |Hate Crimes(per 100k)            |                                  |
## |-  N                             |                47                |
## |-  Mean (SD)                     |          0.304 (0.253)           |
## |-  Median (Q1, Q3)               |       0.226 (0.143, 0.357)       |
## |-  Min                           |              0.067               |
## |-  Max                           |              1.522               |
## |-  Missing                       |                0                 |
## |Unemplotment Level               |                                  |
## |-  N                             |                47                |
## |-  high                          |            24 (51.1%)            |
## |-  low                           |            23 (48.9%)            |
## |Urbanization Level               |                                  |
## |-  N                             |                47                |
## |-  high                          |            24 (51.1%)            |
## |-  low                           |            23 (48.9%)            |
## |Median Household Income(dollar)  |                                  |
## |-  N                             |                47                |
## |-  Mean (SD)                     |       54802.298 (9255.117)       |
## |-  Median (Q1, Q3)               | 54310.000 (47629.500, 60597.500) |
## |-  Min                           |            35521.000             |
## |-  Max                           |            76165.000             |
## |-  Missing                       |                0                 |
## |High School Degree Rate(%)       |                                  |
## |-  N                             |                47                |
## |-  Mean (SD)                     |          0.866 (0.034)           |
## |-  Median (Q1, Q3)               |       0.871 (0.839, 0.895)       |
## |-  Min                           |              0.799               |
## |-  Max                           |              0.915               |
## |-  Missing                       |                0                 |
## |Non-Citizen Rate(%)              |                                  |
## |-  N                             |                45                |
## |-  Mean (SD)                     |          0.055 (0.031)           |
## |-  Median (Q1, Q3)               |       0.050 (0.030, 0.080)       |
## |-  Min                           |              0.010               |
## |-  Max                           |              0.130               |
## |-  Missing                       |                2                 |
## |Non-White Rate(%)                |                                  |
## |-  N                             |                47                |
## |-  Mean (SD)                     |          0.315 (0.150)           |
## |-  Median (Q1, Q3)               |       0.300 (0.205, 0.420)       |
## |-  Min                           |              0.060               |
## |-  Max                           |              0.630               |
## |-  Missing                       |                0                 |
## |Gini Index                       |                                  |
## |-  N                             |                47                |
## |-  Mean (SD)                     |          0.456 (0.021)           |
## |-  Median (Q1, Q3)               |       0.455 (0.441, 0.468)       |
## |-  Min                           |              0.419               |
## |-  Max                           |              0.532               |
## |-  Missing                       |                0                 |
```

Data Description

The original dataset of hate crime rate per 100k population was recorded by the Southern Poverty Law Center during the first weeks of November, 2016. Variable names include state, unemployment, urbanization, median_household_income, perc_population_with_high_school_degree, perc_non_citizen, gini_index, perc_non_white, hate_crime_rate. During the data cleaning process, 4 'N/A' observations of the outcome variable 'hate_crime_rate' were removed. Predictor variables 'unemployment' and 'urbanization' were converted to factors with levels of 'high' and 'low'. The rest of the predictor variables were numeric except 'state'. On average, 0.304 hate crime was committed per 100K population, which was as high as 1.522 in District of Columbia, the federal district. Both employment level and urbanization level were low in around half of the states and high in the rest. The median household income was 54802 dollars across the country with a standard deviation 9255. The variability of income could be addressed through the Gini Index, which had a mean value of 0.456. In other words, there is a big income gap in this country, on average. High school degree rate had a mean value of 86.6% with standard deviation of 3.4%. Moreover, Non-citizen rate had mean value of 5.5% with standard deviation of 3.1%. Lastly, non-white rate had mean value of 31.5% with standard deviation of 15%.

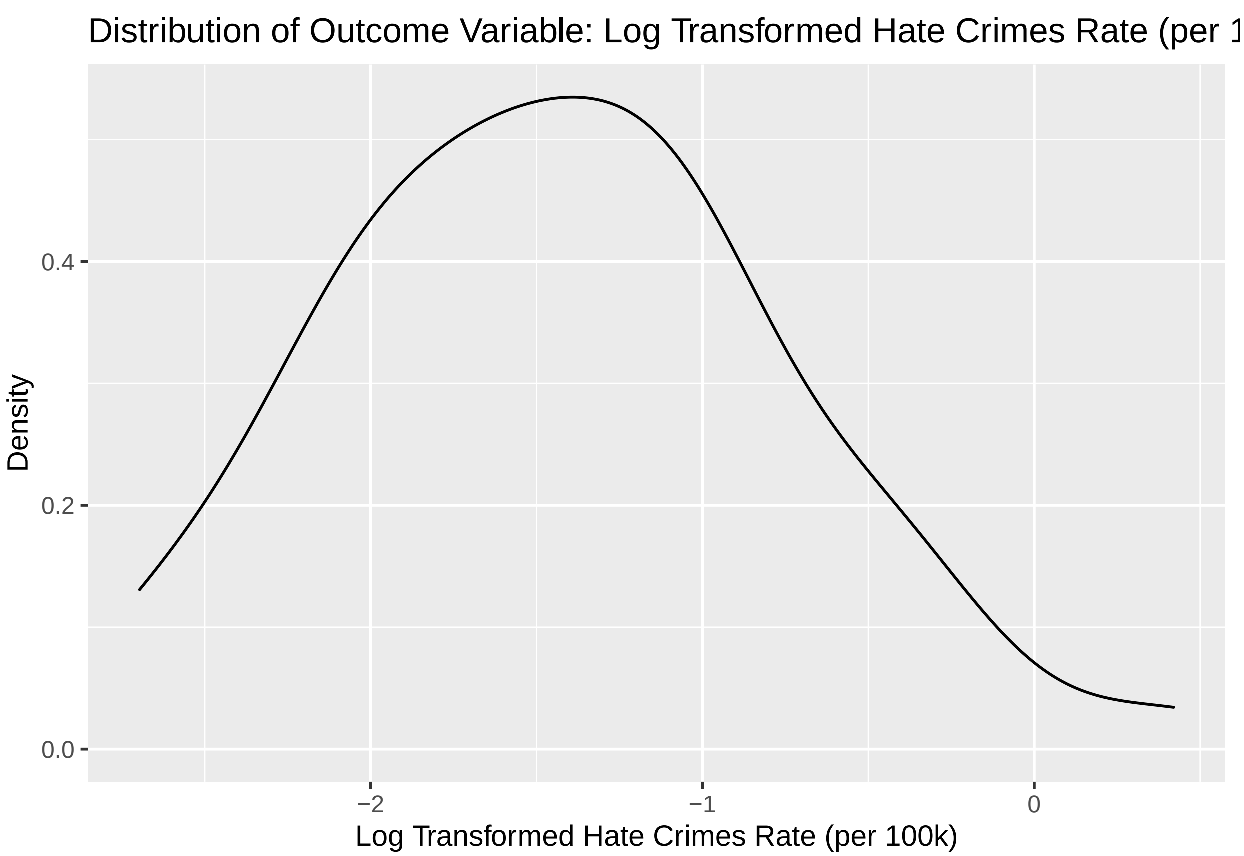
Exploratory Data Analysis

```
#Distribution of Outcome Variable
hatecrimes_df %>%
  ggplot(aes(x = hate_crime_rate)) +
  geom_density() +
  labs(x = 'Hate Crimes Rate (per 100k)', y = 'Density', title = 'Distribution of Outcome Variable: Hat
```



First, we checked the distribution of the outcome variable to detect if any transformation for normality would be needed (Figure 1). Because the distribution was highly skewed to the right, we considered doing a log transformation to the hate_crime_rate variable. The distribution of the outcome variable looked much more normal after doing a log transformation (Figure 2). We also formed a correlation matrix (Table 1), which showed a strong positive association(0.651) between perc_population_with_high_school_degree and median_household_income, and a stronger positive association(0.753) between perc_non_white and perc_non_citizen.

```
hatecrimes_df %>%
  ggplot(aes(x = log(hate_crime_rate))) +
  geom_density() +
  labs(x = 'Log Transformed Hate Crimes Rate (per 100k)', y = 'Density', title = 'Distribution of Outco
```




```
# Correlation matrix for numeric variables
hatecrimes_df <- hatecrimes_df %>%
  drop_na()
hatecrimes_df %>%
  select(hate_crime_rate, median_household_income, perc_population_with_high_school_degree, perc_non_ci
  cor()
```

```
##                                         hate_crime_rate median_household_income
## hate_crime_rate                               1.0000000              0.34378921
## median_household_income                       0.3437892              1.00000000
## perc_population_with_high_school_degree       0.2628198              0.65113832
## perc_non_citizen                              0.2435066              0.30173941
## perc_non_white                                0.1111650              0.03905399
```

```
## gini_index                                           0.3805028               -0.12952158
##                                            perc_population_with_high_school_degree
## hate_crime_rate                                                           0.2628198
## median_household_income                                                   0.6511383
## perc_population_with_high_school_degree                                   1.0000000
## perc_non_citizen                                                         -0.2621288
## perc_non_white                                                           -0.4958932
## gini_index                                                               -0.5371591
##                                            perc_non_citizen perc_non_white
## hate_crime_rate                                   0.2435066     0.11116503
## median_household_income                           0.3017394     0.03905399
## perc_population_with_high_school_degree          -0.2621288    -0.49589321
## perc_non_citizen                                  1.0000000     0.75261020
## perc_non_white                                    0.7526102     1.00000000
## gini_index                                        0.4798976     0.54840351
##                                            gini_index
## hate_crime_rate                             0.3805028
## median_household_income                    -0.1295216
## perc_population_with_high_school_degree    -0.5371591
## perc_non_citizen                            0.4798976
## perc_non_white                              0.5484035
## gini_index                                  1.0000000
```

```
#Potential outliers/Influential points of outcome variable

# (1) Identify unusual states by using the interval formed by 2.5 and 97.5 percentiles
lower_bound <- quantile(hatecrimes_df$hate_crime_rate, 0.025)
upper_bound <- quantile(hatecrimes_df$hate_crime_rate, 0.975)

outlier <- which(hatecrimes_df$hate_crime_rate < lower_bound | hatecrimes_df$hate_crime_rate > upper_bou
hatecrimes_df[outlier, ] %>% select(state, hate_crime_rate)
```

```
##                   state hate_crime_rate
## 4              Arkansas      0.06906077
## 9  District of Columbia      1.52230172
## 28           New Jersey      0.07830591
## 34               Oregon      0.83284961
```

```
# (2) Identify unusual states by using the rule that depicts outliers (value less than Q1 - 1.5(IQR), v

hatecrimes_df %>%
  filter(hate_crime_rate > 0.678) %>%
  select(state, hate_crime_rate)
```

```
##                  state hate_crime_rate
## 1 District of Columbia       1.5223017
## 2               Oregon       0.8328496
```

```
# (3) Identify unusual states by using studentized residual, 9th row returned: District of Columbia.
fit_full <- lm(hate_crime_rate ~ gini_index + median_household_income + perc_population_with_high_school
stu_res<-rstandard(fit_full)
outliers_y<-stu_res[abs(stu_res)>2.5]
outliers_y
```

```
##        9
## 3.726868
```

Moreover, one of the ways we tried to identify states with unusual rates was to identify states that have the hate crime rates below the 2.5 percentile and above the 97.5 percentile. From doing so, Arkansas, District of Columbia, Mississippi, and Oregon were selected. We also identified outlier states by using the rule that depicts outliers, which are defined as values less than Q1-1.5(IQR) and values greater than Q3+1.5(IQR). From doing so, District of Columbia and Oregon were selected. Lastly, we also used the studentized residual model to identify potential outliers. This returned District of Columbia as the only outlier.

Further Data Analysis

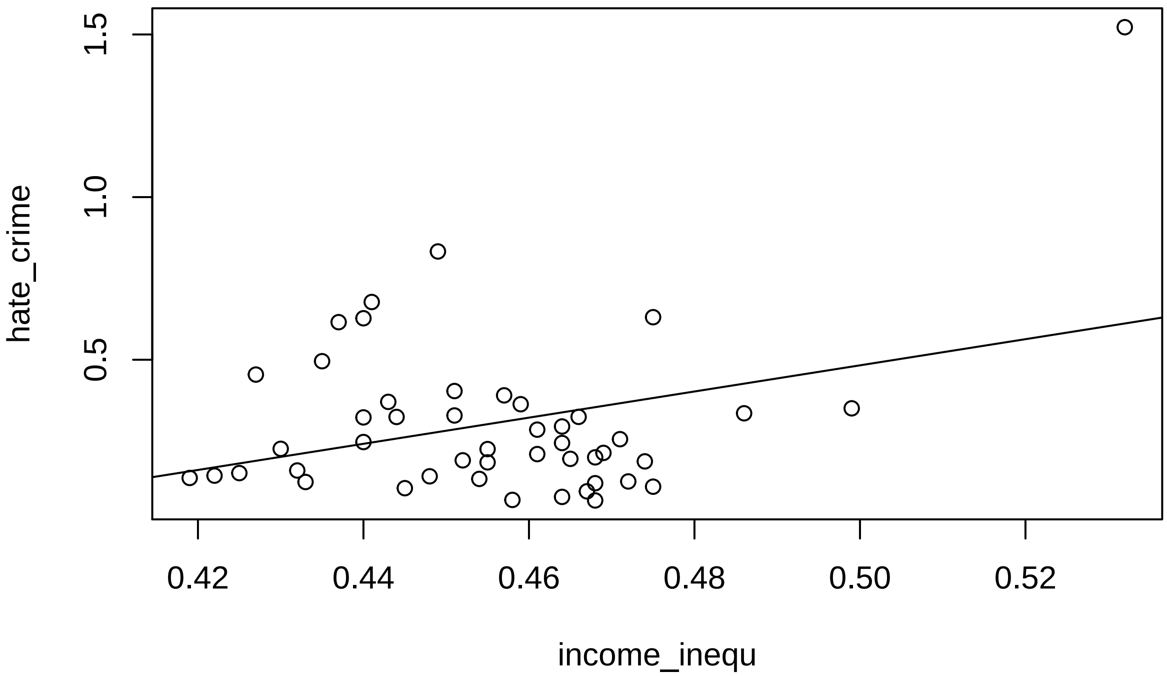
```
hatecrime = read.csv("./data/HateCrimes.csv")
hatecrime = hatecrime[!(hatecrime$hate_crimes_per_100k_splc == "N/A" ),]
hatecrime = hatecrime %>%
  mutate(med_income = median_household_income,
         perc_HSdegree = perc_population_with_high_school_degree,
         income_inequ = gini_index,
         hate_crime = as.numeric(hate_crimes_per_100k_splc)
         ) %>%
  select(-median_household_income, -perc_population_with_high_school_degree, -hate_crimes_per_100k_splc

hatecrime_lm = lm(hate_crime ~ income_inequ, data = hatecrime)
summary(hatecrime_lm)
```

Income inequality v.s. Hate crime rate

```
##
## Call:
## lm(formula = hate_crime ~ income_inequ, data = hatecrime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28669 -0.14565 -0.04991  0.07356  0.91085
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.5275     0.7833  -1.950   0.0574 .
## income_inequ   4.0205     1.7177   2.341   0.0237 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2412 on 45 degrees of freedom
## Multiple R-squared:  0.1085, Adjusted R-squared:  0.08872
## F-statistic: 5.478 on 1 and 45 DF,  p-value: 0.02374
```

```
plot(hate_crime ~ income_inequ, data = hatecrime) + abline(hatecrime_lm)
```

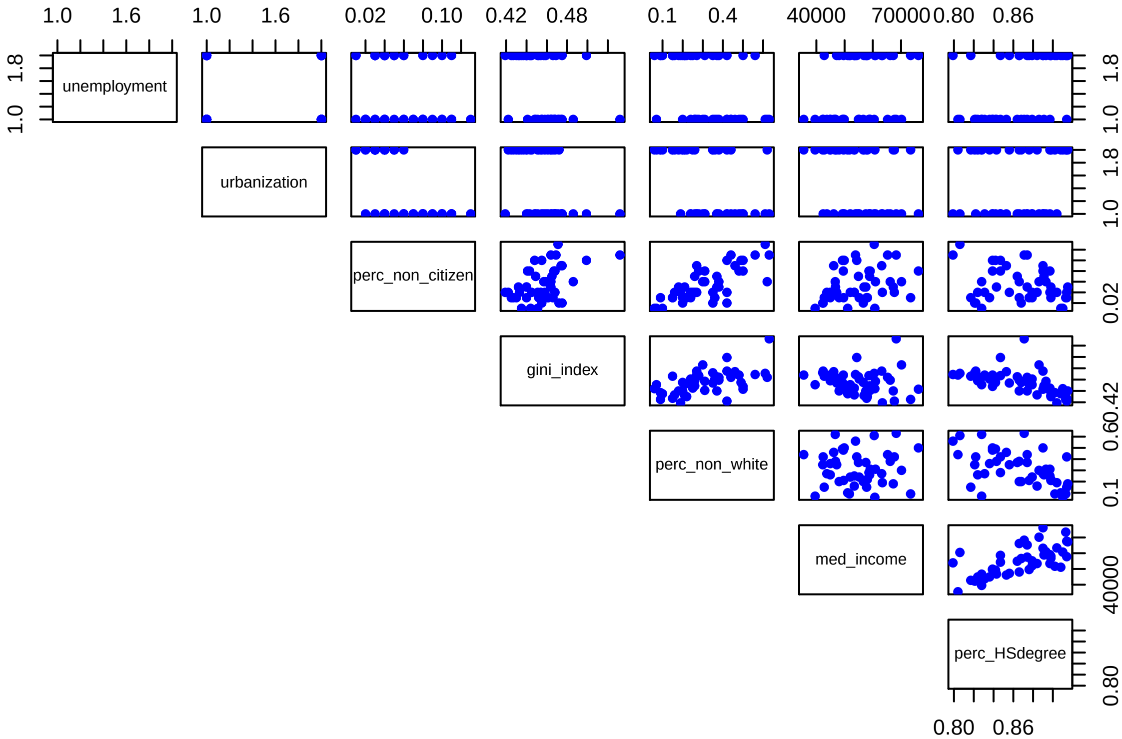



```
## integer(0)
```

From making a simple linear regression model that only includes income inequality as the variable, we were able to find that the p-value (0.02374) and the adjusted r-squared value (0.08872) were low. While the low p-value was a good indicator of statistical significance, the adjusted r-squared revealed that only 8.872% of the data will adequately fit the regression model. Also, looking at the fitted model, we could see that not all the points are along the regression line and that there is a major outlier point as well. Therefore, we sought to find a more precise model by constructing a multiple regression model.

First MLR model selection First, we made a matrix plot and examined the findings (Figure 3). We could see that there is a relationship between `med_income`, `perc_HSdegree` and `income_inequ`. So we used these three variables to predict the regression model.

```
plot(hatecrime[2:8], pch=16, col="blue", lower.panel = NULL)
```

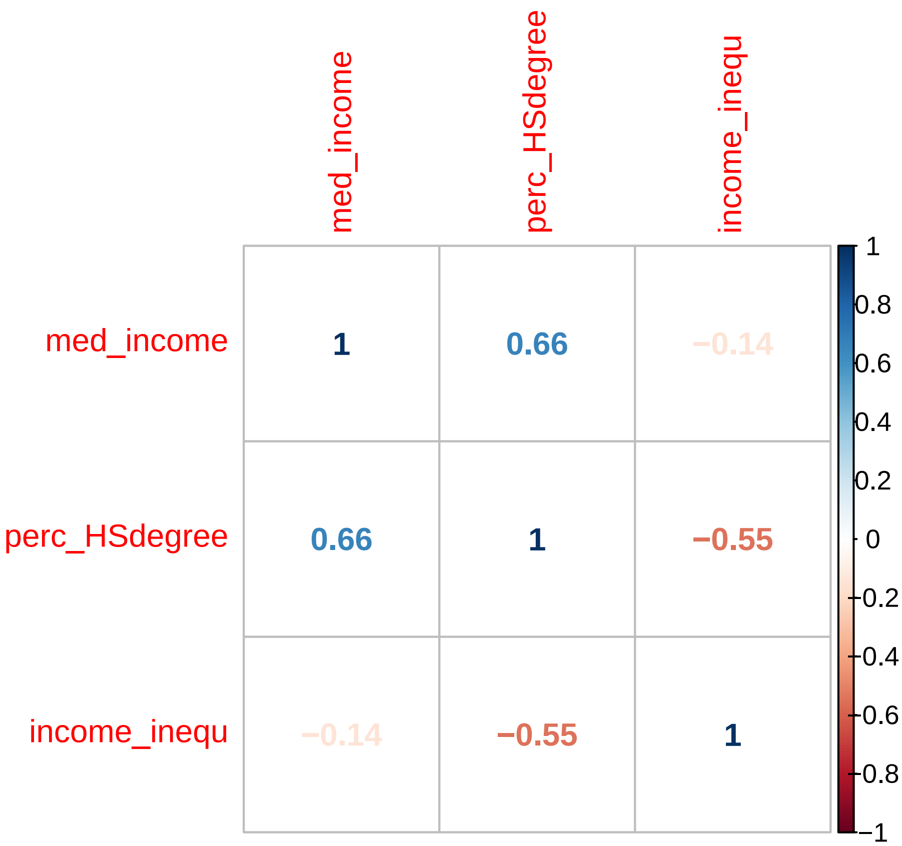



```r
hatecrime_mlr1 = lm(hate_crime ~ med_income + perc_HSdegree + income_inequ , data = hatecrime)
summary(hatecrime_mlr1)
```

```
##
## Call:
## lm(formula = hate_crime ~ med_income + perc_HSdegree + income_inequ,
##     data = hatecrime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33615 -0.12472 -0.02392  0.11543  0.52780
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.355e+00  1.705e+00  -4.900 1.40e-05 ***
## med_income    -6.851e-07  4.406e-06  -0.155 0.877157
## perc_HSdegree  5.414e+00  1.434e+00   3.774 0.000487 ***
## income_inequ   8.799e+00  1.762e+00   4.993 1.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1936 on 43 degrees of freedom
## Multiple R-squared:  0.4514, Adjusted R-squared:  0.4132
## F-statistic:  11.8 on 3 and 43 DF,  p-value: 9.083e-06
```

From examining the summary output after running the multiple linear regression model (Table 2), we could see that the adjusted R squared increased from 0.08872 to 0.4132. In addition, we could see that by improving one index measuring income inequality, the average hate crime rate per 100,000 population increased by 8.799. Similarly, seeing an improvement in the percentage of population with high school degree by one percent leads to an extra 5.414 rate in average hate crime rate. These coefficients that correspond to 'income inequality' and 'percentage of population with high school degree' have p-values under 0.05, which indicate that they are statistically significant. However, the coefficient corresponding to 'median income' has a large p-value of 0.877157, which shows it is not statistically significant.

```
hatecrimecor = hatecrime %>%
  select(med_income,perc_HSdegree,income_inequ) %>%
  cor()
corrplot(hatecrimecor, method = "number")
```



To be more specific, we could test using the global (overall) F-test to see if all our coefficients are equal to zero or if there is at least one coefficient that is not equal to zero. With the F-statistic of 11.8 with df 3 and 43, and critical value of 2.82, we could reject the null(because F-stat > F-crit) and conclude that there is at least one coefficient that is not equal to zero. Therefore, we had some evidence that at least one of the predictors is associated with hate crime rate.

Next, we created another correlation matrix of these three variables in the model (Figure 4). Here, we could see that `med_income` is associated with the percentage of population with HS degree, but there was no evidence that it is related with the index measurement of income inequality.

We then considered removing `med_income` from the model and saw how the model fit changed. Looking at the regression analysis summary, we could see that there was a significant change on both the F-statistics and the adjusted R-squared value. the F-statistic changed from 11.8 to 18.08 and the adjusted R-squared increased from 0.4132 to 0.4262. These changes indicated that the model without 'med_income' is stronger than the model that includes 'med_income'.

Thus, we could conclude with the multiple linear regression model of:

$$HateCrimeRate = -8.212 + 5.256 * percHSdegree + 8.702 * incomeInequality$$

Second MLR model selection using Backward Elimination method We also used the Backward Elimination method and yielded a model with the same predictors (Table 3). This validated the first model that we produced above.

$$HateCrimeRate = -8.103 + 5.059 * percHSdegree + 8.825 * incomeInequality$$

The coefficients in this model slightly differed from the coefficients in the original model becauase this Backward Elimination method required users to drop all N/A values.

Model diagnostics To check the model assumptions and goodness of fit, we created a histogram (Figure 5). The overall shape of the histogram looked close to a normal distribution, so we could see that normality has been met.

We also created diagnostic plots, and noticed that the residuals do not have any pattern around the 0 line, which indicates that there is a linear relationship (Figure 6). But some values stood out because they were outliers. In addition, the distances to the zero line were not equal, so we could see that there is a high interval of error terms of variability. The Q-Q plot also showed that the overall distribution of outcome is normal, except point 9, which suggested this point as an influential point in Residuals vs Leverage.

After outlier point 9 was removed, the coefficients of both `perc_HSdegree` and `income_inequ` changed greatly. The coefficient of `perc_HSdegree` decreased from 5.256 to 3.284 and the coefficient of `income_inequ` decreased from 8.702 to 3.136.

Bootstraps Since the distribution of our outcome variable(hate_crime_cate) was not normal, we used the bootstrap method to check the variability of model estimates. We found that the coefficient of income inequality had the greatest standard error(1.742). This implied that income inequality has a more uncertain extent of influence on hate crime rate compared to the percentage of adults with a high-school degree.