Hate crimes near you? Here are some possible factors...

STA141A Spring 2021 Final Project

By: Olivia Jungwon Yoon, Christine Phan, Su-Ting Tan, Sihua Cai, Henman Tan

Introduction & Background

Hate crimes are defined as criminal acts motivated by prejudice on the basis of perceived characteristics such as race, religion, gender, and sexual orientation. In today's hostile climate, hate crimes have frequently made national headlines, with anti-Asian and antisemitic hate crimes having particularly been on the rise. In our project, we will be analyzing data behind FiveThirtyEight's 2017 article, "Higher Rates Of Hate Crimes Are Tied To Income Inequality" to explore how the frequency of hate crime occurrences for each state in the United States may correlate with regional socioeconomic factors such as income, diversity, and education attainment, may correlate with frequency of hate crime occurrences.

Although the purpose of this data is mainly to explore how economic disparity is related to hate crimes, it is also to question increased hate incidents after the presidential election in 2016 as there was controversy about the impact of Trump's administration on the wealth gap. We are not to discuss nor criticize the administration. Yet, we would like to investigate how socioeconomic status affects crimes related to hate. We will possibly consider this factor in our analysis by looking at correlation between the increase in hate crimes and the share of voters for Donald Trump in a region and other variables.

We will be using a hate crimes data set posted on Kaggle in 2018 by FiveThirtyEight. The data set is a compilation of socioeconomic data from the Census Bureau, the FBI, the Kaiser Family Foundation, the Southern Poverty Law Center, and the United States Elections Project.
The data includes 12 variables, one categorical variable 'state' and 11 numeric variables that indicate education, geographic heterogeneity, economic health, income inequality, proportion of voters for Donald Trump, and frequency of hate crimes before and after the 2016 election.

Statistical Questions of Interest

The primary goal of analyzing our dataset is to see which of the above factors most greatly affect the amount of hate crimes that occur in each state.

1. Are there correlations between the different variables that are included in the dataset?
2. What affects the frequency of hate crimes in a state? Which of the 11 socioeconomic variables are the most influential?
3. Do the hate crime rates vary across different regions? What causes this variation?

Study Design and Methodology

Our first step will involve creating correlation tables in determining if a relationship exists between the variables. Additionally, we will conduct analysis of multiple linear regression models using the *lm()* function in R. Creating these linear regression models will help us explain the relationship between hate crimes and the regressors that are most influential to hate crimes. We will also conduct data visualization and graphical analysis using *ggplot2* in aim to display any variation of hate crimes across different states.

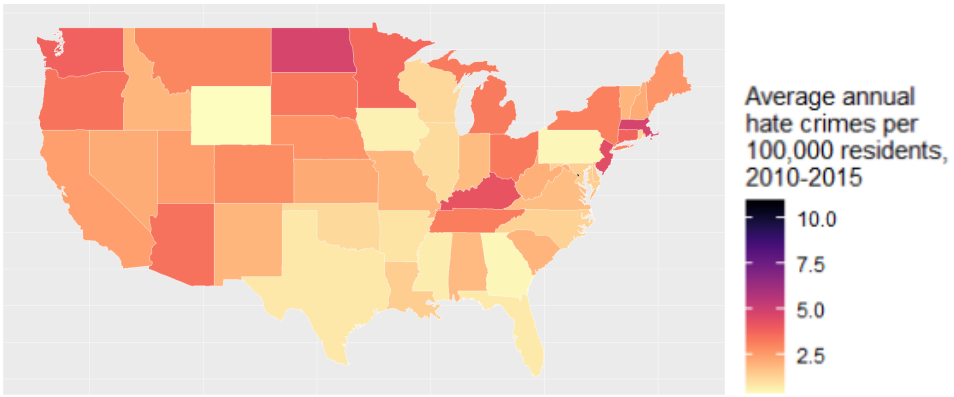
Data Preparation

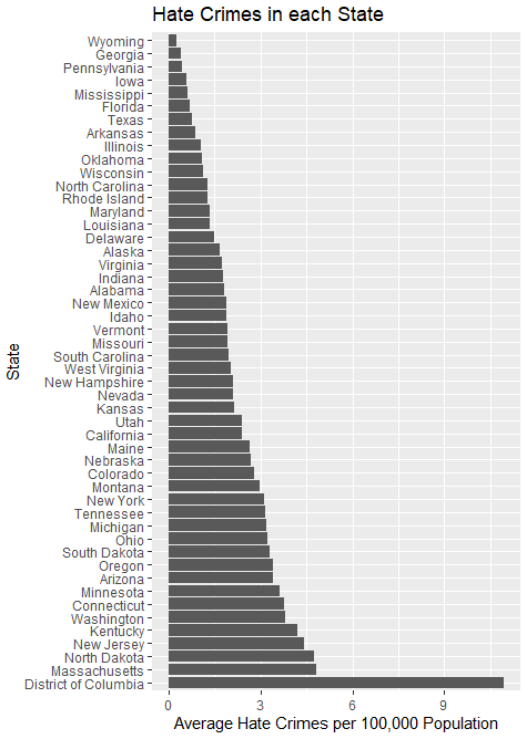
We will be using a hate crimes data set posted on Kaggle in 2018 by FiveThirtyEight. We will also use supplemental data sets to update the original dataset. We will use this dataset from Kaiser Family Foundation to fill in NA values in the `non_citizen` column of the original dataset. In the original dataset, the hs_degree data was gathered in 2009, whereas the other data was collected in 2015 or 2016, so we will use this dataset from the United States Census Bureau to update the `hs_degree` column with data

collected in 2015. We also added an additional variable to indicate the region of every state, as classified by the United States Census Bureau.

Data Visualization

Average Annual Hate Crimes per 100,000 residents, 2010-2015

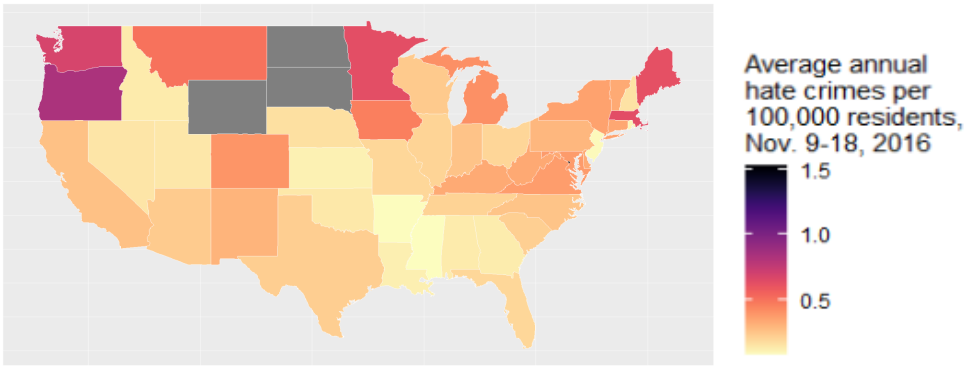


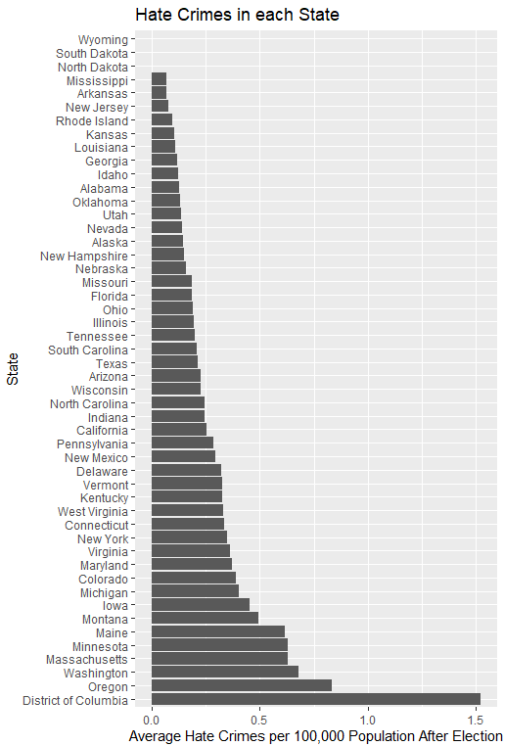


Shown left is a bar graph of the response variable, the average hate crimes per 100,000 population for each state before the 2016 election. We excluded Hawaii from this graph because the dataset did not provide hate crime frequency information for that state. The District of Columbia has the highest average of all the states, even doubling that of Massachusetts, which has the second highest average.

Shown above is the same information represented on a map of the contiguous United States.

Average Annual Hate Crimes per 100,000 residents, 2016





Shown left is a bar graph of the average hate crimes per 100,000 population for each state in the 10 days after the 2018 election.

The states that are excluded from this graph are Wyoming, South Dakota, North Dakota, and Hawaii, because no crime rate data was provided for these 4 states for the 10 days after the 2016 election.

Again, we see that the District of Columbia has the highest average of about 1.5, compared to the second highest state, which has an average of about 0.75.

Shown above is the same information represented on a map of the United States.

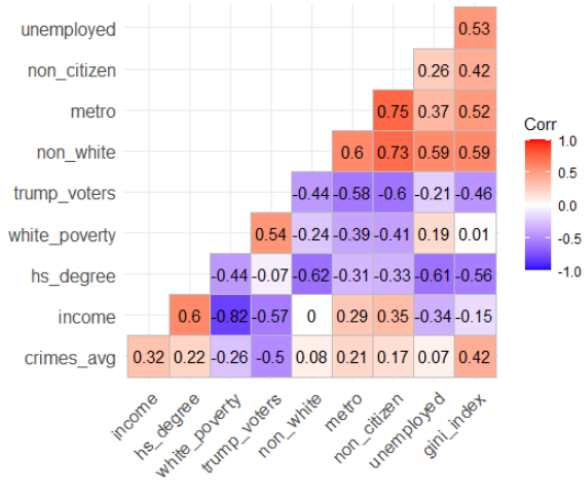
Results & Analysis

1. Are there correlations between the different variables that are included in the dataset?

We used the "ggcorrplot" function in R to see if there was any correlation between hate crimes and the other variables that are included in the dataset. After running it, we see that the most highly correlated variables are between white poverty and median income, with a correlation value of -0.82. This result makes sense because in states where there are a lot of people in poverty, the median household income is not as high, so there is a negative relationship between those two variables.

Also, another interesting part is that we had expected there to be a negative relationship with lower levels of education being associated with higher hate crime rates, but there is a slight positive relationship between the crime rate and rate of high school degree obtained. It appears that there is not a

strong linearity between the predictor variables and outcome variable (hate crimes). We will closely investigate this more in our model assumption.



2. What affects the frequency of hate crimes in a state? Which of the 11 socioeconomic variables are the most influential?

To answer this question, we would initially investigate our model assumptions and select a model using the Akaike Information Criteria method of selection.

Checking Model Assumptions

We chose to use the multiple full term regression model as a point of reference to check the model assumptions using the *lm()* function. There were 5 assumptions we investigated. First, we made sure that error terms are independent by using the Durbin-Watson test. To check linearity, normal distribution, and homoscedasticity, we ran the model plot. Washington DC appeared to be a potential influential data point, and our linearity, homoscedasticity, and normal distribution assumptions appeared to be violated most likely because of this. We ran a BoxCox function to see what transformation to take. Since the highest point of our lambda was around 0.5, we took a square root transformation.

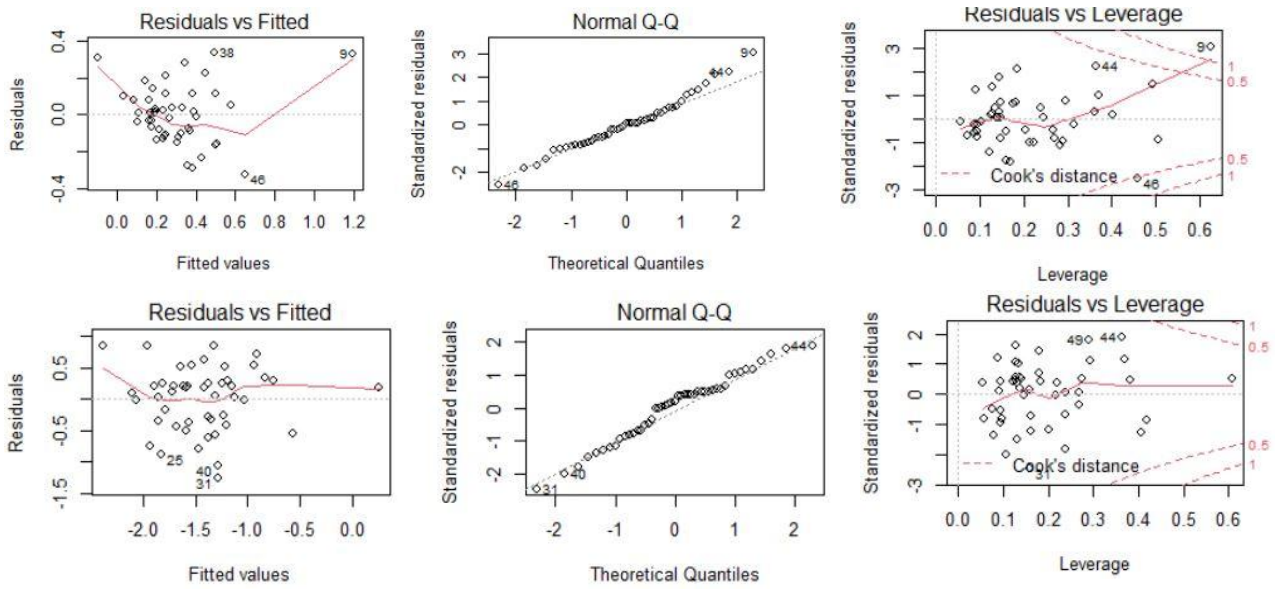
After the transformation, DC had a smaller effect on the model, and errors also seemed to be normally distributed, as verified in the Shapiro-Wilks test. Yet, it still appeared that it might be a leverage point as seen in our Residuals vs Leverage plot and a slightly weird trend in distribution errors. Before considering removing the influential point, we will try adding region characteristic predictors (S, NW..) as covariate to increase the accuracy of the fit by reducing the bias while increasing the variance. However, after adding it, there wasn't much change of the tendency of heteroskedasticity, so we will not consider adding the region variable into our model.

To check whether the DC observation is an outlier, we calculated the studentized residual, which was less than 3, so we decided to keep it for our further model selection process as it wasn't recorded as an error after checking the original data. It seems that other assumptions are met pretty well after transformation, but we should investigate more on heteroskedasticity, and it might cause a worry that linear regression might not be suitable for this data.

Lastly, to test multicollinearity, we used the *vif()* function to quantify the severity of the assumption in an ordinary least square regression. Three variables, income, hs_degree, non_white seem to be highly correlated with VIF values over 5. We decided to drop the variable of rate of non-white residents from the model since it has the least correlation with the outcome variable. This makes sense since 60% of the US citizen population is white, and the non-white variable could potentially be explained with other variables such as non_citizen.

We wanted to analyze the hate crime rate after the 2016 election, so we used the same procedure to check the assumption and all the assumptions seem to be held more confidently this time. We went with the log transformed model and removed the variable non_white as it appeared to have a high VIF value but less related to the outcome. Something to note is that three states were not included in this after election model, compared to the before election model, meaning that since we did not include three states, we might have missed some important implications.

Figures below: lm(crimes rate after election ~.) plot :before transformation



Figures above: lm(log(crimes rate after election) ~.) plot: after transformation

Model Selection

After checking the model assumptions, we then selected the variables to include in our model to best explain our data. Our base model is the transformed linear model with only the intercept and no variables, and our full model includes all terms after removing the highly correlated variable. We chose to use the Akaike Information Criterion (AIC) that is commonly used to compare models in both directions (forward and backward) to achieve our model selection.

Our final model by AIC selected the proportion of population with a high school degree and the Gini index as statistically significant socioeconomic variables to include in the model due to their strong associations with 2010-2015 hate crime rates. This resulted in the following regression equation:

$$\sqrt{Y} = -15 + 10X_{hs} + 16X_{gini}$$.

Since we took a square root transformation, the negative intercept should not be much of a concern. The positive coefficient of 10 for the high school degree variable suggests that there is a positive relationship between high school education and hate crime rates.

The Gini index coefficient of 16 indicates that there is also a positive relationship between income inequality and hate crime rates. We would expect states with higher levels of income inequality to have higher hate crime rates, and this is what the analysis of the FiveThirtyEight article focused on. The Southern Poverty Law Center collected the post-election hate crime statistics included in this data set, and Mark Potok, their journal editor-in-chief, also noted the significance of income inequality as a factor in hate crime rates: "It's typically not your objective situation that makes you angry and resentful, but rather your situation relative to others you see around you. So, where income inequality is very high, so is anger and resentment against those 'other' people who you fear are doing better than you."

To see if this holds with post-election data, we again performed model selection using AIC, now with the post-election hate crime rates as the response variable. This time, our final model by AIC selected the state rates of Trump voters, high school degrees, and white poverty as statistically significant socioeconomic variables associated with the post-election hate crime rates. It is important to note that the Southern Poverty Law Center collected these hate crime rates in the 10 days after the 2016 presidential election, when societal discontent was particularly high. That may explain why with our post-election data, Trump voter support and white poverty were found to be more significant factors than income inequality. With these chosen variables, we had the following regression equation:

$$log_{10}Y = -8 - 4.1X_{trump}, +8.8X_{hs} + 8.8X_{w.poverty}$$

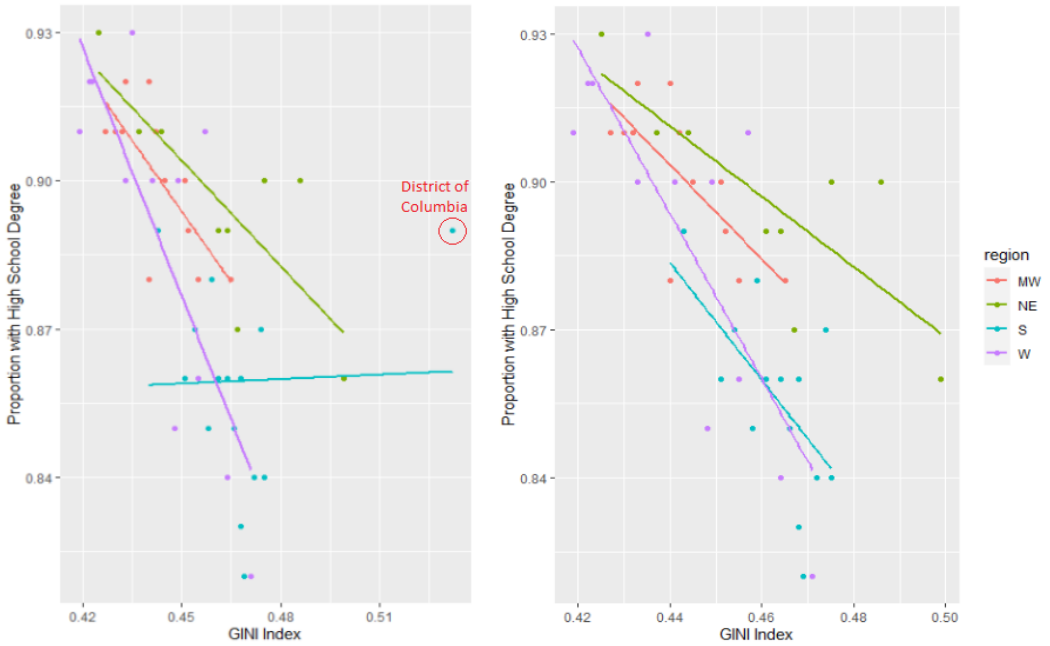
Since we took a log transformation to better fulfill the model assumptions, the interpretation of the negative intercept should not matter. The high school degree variable again unexpectedly has a positive coefficient. The negative coefficient (-4.1) for the Trump voter variable suggests that there is a negative relationship between the levels of Trump support in a state and their post-election hate crime rate. We found this to be interesting, as we had expected there to be a positive relationship. However, we can posit that states with greater levels of Trump support may have smaller populations of marginalized groups who are more often targeted in hate crimes. This would result in lower reported rates of hate crimes in the 10 days following the election. More long-term data in the months or years following the election may also be needed to conclusively determine trends. The positive coefficient (8.8) on the white poverty variable suggests that high rates of white poverty are correlated with higher hate crime rates. This is what we expect, as white poverty and income inequality have likely have similar relationships with hate crime rates. As Mark Potok emphasized, economic discontent seems to be a driving factor behind resentment of perceived "others" and hate crimes.

3. Does the frequency of hate crimes vary across different regions? What causes this variation?

As of now, we have only explored hate crimes across America as a whole, but as Professor Furfaro suggested in the proposal comments, we are also interested to see if there are any regional patterns. We used the United States Census Bureau's regional groupings to classify each state into one of the following four regions: Midwest, Northeast, South, and West.

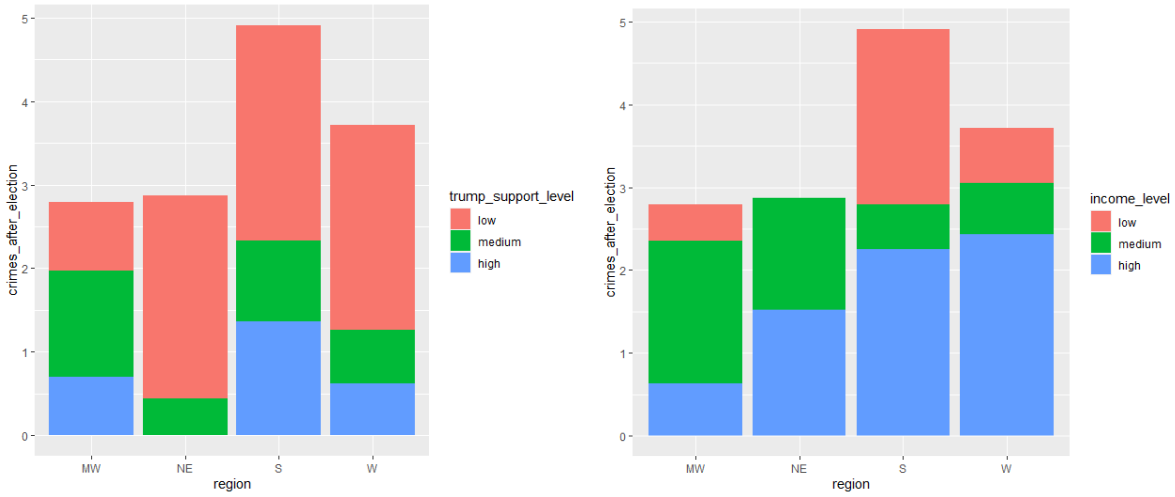
Using the same predictor variables (proportion of population with high school degree and Gini index) that we chose to include in our model in the previous question, we fitted four separate lines for each region. The left plot is for pre-election crime frequencies for all states except Hawaii. It is immediately obvious that the line for states in the South is extremely different from those of the other regions. This is in large part due to the District of Columbia observation. The District of Columbia has an extremely high Gini index and percentage of high school graduates compared to the other states in the south, which affects the slope and intercept of the model for states in the south. Thus, the DC observation is an influential point.

In the plot on the right, we removed the DC observation and fitted models with Gini index and proportion with high school degree again, and we can see now that the line for the South has a slope and intercept that does not deviate far from those of the other regions.



We also wanted to see how hate crime rates in the different regions compared to each other, and how the regions varied in terms of levels of Trump support (which is acquired from the share of 2016 U.S. presidential voters who voted for Trump) and income, since these were regional characteristics that we found to be significant based on our previous analysis. We divided the Trump support and income variables into levels of low medium, and high, and visualized the regional distributions in the following stacked bar graphs:

We can see from the first graph that Southern states had the highest hate crime rates after the 2016 election, and that they also had the largest proportion of high Trump support compared to the other regions. From the second graph, we can see that the South notably has the highest proportion of people with low income compared to the other regions as well.



Conclusion

After the transformation and AIC model selection, our conclusion was overall similar to the results in the FiveThirtyEight article, as it found that variables that indicate income inequality, as measured by Gini index scores and white poverty rates, were significantly associated with hate crime rates. Economic discontent seems to be a driving factor behind resentment of perceived "others" that leads to hate crimes.

While we were fitting our models, we saw that Washington DC was an outlier and would make a big difference if we included it in our model. Since Washington DC has an extremely high Gini index and percentage of high school graduates compared to other states, this explains why it makes a big difference to our models and plots, and hence why we had to apply transformations to our dataset.

One important factor we should note is that our analysis has limitations. Even though we controlled potential confounders, correlation between income inequality and hate crimes doesn't imply causation. As quoted in the article, hate crimes in communities may differ from what we analyzed in states. Other factors such as law enforcement laws and the amount of people who report these hate crimes could affect the results we have obtained. Conclusions about long-term trends from our post-election data also have limitations, as the post-election hate crime rates only reflect statistics from the 10 days immediately following the 2016 election.

Member Contributions

Olivia Jungwon Yoon (ojyoon@ucdavis.edu)	Data cleaning, model assumptions, model selection, analysis of plots.
Christine Phan (chphan@ucdavis.edu)	Data visualization, model selection interpretation, analysis of plots.
Su-Ting Tan (sttan@ucdavis.edu)	Correlation matrices, analysis of correlation and plots, conclusion of final results.
Sihua Cai (Sarah) (shscai@ucdavis.edu)	Data preparation, investigation of cross-regional differences in hate crime frequencies through linear models, analysis of plots.

Henman Tan (hetan@ucdavis.edu)	Analysis of variation on hate crime frequency across different regions, analysis of plots.

Links Referenced

- FiveThirtyEight 2017 article, “Higher Rates Of Hate Crimes Are Tied To Income Inequality”
- FiveThirtyEight 2018 Kaggle Dataset, “FiveThirtyEight Hate Crimes Dataset”
- Kaiser Family Foundation 2015 Dataset, “Population Distribution by Citizenship Status”
- United States Census Bureau 2015 Dataset, “Educational Attainment”

Code Appendix

```
require(knitr)
library(ggplot2)
library(dplyr)
library(corrplot)
library(ggcorrplot)
library(maps)
library(viridis)
library(MASS)
library(ISLR)
library(GGally)
library(olsrr)
library(car)

# loading the data
crimes <- read.csv("hate_crimes.csv")
citizen_bystate <- read.csv("citizen_bystate.csv")
edu_attainment_2015 <- read.csv("edu_attainment_2015.csv")

# simplifying variable names
crimes <- crimes %>% rename(income = median_household_income,
                            unemployed = share_unemployed_seasonal,
                            metro = share_population_in_metro_areas,
                            hs_degree = share_population_with_high_school_degree,
                            non_citizen = share_non_citizen,
                            white_poverty = share_white_poverty,
                            non_white = share_non_white,
                            trump_voters = share_voters_voted_trump,
                            crimes_avg = avg_hatecrimes_per_100k_fbi,
                            crimes_after_election = hate_crimes_per_100k_splc)

# replacing high school degree column if the state order matches
crimes$hs_degree <- ifelse(crimes[,1] == edu_attainment_2015[,1],
                           edu_attainment_2015$high_school_degree, "state does not match")

# replacing/filling in non-citizen column
crimes$non_citizen <- ifelse(crimes[,1] == citizen_bystate[,1],
                             citizen_bystate$Non.Citizen, "state does not match")

# adding region variable
crimes$region <- as.factor(citizen_bystate[,2])

# selecting variables to include

# before election
before_crimes <- subset(crimes, select = c('income', 'unemployed', 'metro', 'hs_degree',
                                           'non_citizen', 'white_poverty', 'gini_index',
                                           'non_white', 'trump_voters', 'crimes_avg'))

#after election
after_crimes <- subset(crimes, select = c('income', 'unemployed', 'metro', 'hs_degree',
                                          'non_citizen', 'white_poverty', 'gini_index',
```

```
                                          'non_white', 'trump_voters', 'crimes_after_election'))

# which states are missing before election crimes?
before_crimes[!complete.cases(before_crimes$crimes_avg),]
## hawaii

# removing hawaii from before election crimes
before_crimes <- before_crimes[-12,]

#mean 2.4

# which states are missing after election crimes?
after_crimes[!complete.cases(after_crimes$crimes_after_election),]
## hawaii, north dakota, south dakota, wyoming


# checking regions of states missing crimes data
# north dakota
crimes$region[35] ## MW
# south dakota
crimes$region[42] ## MW
# wyoming
crimes$region[51] ## W

# removing hawaii, nd, sd, wyoming from after election crimes
after_crimes <- after_crimes[-c(12,35,42,51),]

ggplot(crimes, aes(x = reorder(state, -crimes_avg), y = crimes_avg)) +
  geom_bar(stat = "identity") +
  labs(title = "Hate Crimes in each State") +
  ylab("Average Hate Crimes per 100,000 Population Before Election") +
  xlab("State") +
  coord_flip()
ggplot(crimes, aes(x = reorder(state, -crimes_after_election), y = crimes_after_election)) +
  geom_bar(stat = "identity") +
  labs(title = "Hate Crimes in each State") +
  ylab("Average Hate Crimes per 100,000 Population After Election") +
  xlab("State") +
  coord_flip()
par(mfrow = c(1,2))
crimes$crimes_avg <- as.numeric(crimes$crimes_avg)
crime.df <- data.frame(region = crimes$state, avg_crime = crimes$crimes_avg)
crime.df$region <- tolower(crimes$state)

states_map <- map_data("state")
crimes_map <- left_join(states_map, crime.df, by = "region")

ggplot(crimes_map, aes(long, lat, group = group))+
  geom_polygon(aes(fill = avg_crime), color = "white")+
  scale_fill_viridis(option = "magma", direction = -1) +
  labs(title = "Hate Crime Rates in the United States Before Election",
    fill = "Average annual \nhate crimes per \n100,000 residents, \n2010-2015") +
  theme(axis.text.x = element_blank(),
```

```
        axis.text.y = element_blank(),
        axis.title.y=element_blank(),
        axis.title.x=element_blank(),
        axis.ticks = element_blank())

crimes$crimes_after_election <- as.numeric(crimes$crimes_after_election)
crime.df <- data.frame(region = crimes$state, avg_crime = crimes$crimes_after_election)
crime.df$region <- tolower(crimes$state)

states_map <- map_data("state")
crimes_map <- left_join(states_map, crime.df, by = "region")

ggplot(crimes_map, aes(long, lat, group = group))+
  geom_polygon(aes(fill = avg_crime), color = "white")+
  scale_fill_viridis(option = "magma", direction = -1) +
  labs(title = "Hate Crime Rates in the United States After Election",
    fill = "Average annual \nhate crimes per \n100,000 residents, \nNov. 9-18, 2016") +
  theme(axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.title.y=element_blank(),
        axis.title.x=element_blank(),
        axis.ticks = element_blank())

plot(before_crimes)
ggscatmat(before_crimes)

crimes2 <- crimes[,c(-1, -11, -13)] # removing qualitative variables
crimes_matrix <- as.matrix(crimes2)
corr_matrix <- cor(before_crimes)

signif_4 <- function(x){
  return(signif(x, 3))
}

corr_matrix <- apply(corr_matrix, c(1,2), signif_4)
corr_matrix2 <- corr_matrix

kable(corr_matrix2)

# correlogram of all variables
corrplot(corr_matrix)



label_names <- c("Median Income", "Unemployment", "Metro Population", "High School Education",
                 "Non-Citizen", "White Poverty", "Gini Index", "Non-White",
                 "Voters who Voted Trump", "Hate Crimes")

ggcorrplot(corr_matrix) +
    scale_x_discrete(labels = label_names) +
    scale_y_discrete(labels = label_names)

ggcorrplot(corr_matrix, hc.order = TRUE, type = "lower", lab = TRUE)
```

```
bfit<- lm(crimes_avg ~ ., data = before_crimes)

qplot(1:length(bfit$res),bfit$res,geom="line") +
    geom_hline(yintercept = 0, colour = "red")

par(mfrow=c(2,2))
plot(bfit)
summary(bfit)

boxcox(bfit)

bfit_tr <- lm(sqrt(crimes_avg) ~., data=before_crimes)
plot(bfit_tr)
summary(bfit_tr)
hist(bfit_tr$residuals)

before_crimes_cov <- before_crimes
before_crimes_cov$region <- as.factor(crimes[-12,]$region)
bfit_cov <- lm(sqrt(crimes_avg) ~., data=before_crimes_cov)
plot(bfit_cov)
summary(bfit_cov)

bfit_cov_no <- lm(sqrt(crimes_avg) ~., data=before_crimes_cov[-9,])

bfit_no_out <- lm(sqrt(crimes_avg) ~ ., data = before_crimes[-9,])
plot(bfit_no_out)

shapiro.test(bfit_no_out$residuals)

summary(bfit_tr)
summary(bfit_no_out)

par(mfrow=c(2,2))


afit <- lm(crimes_after_election ~ ., data = after_crimes)
plot(afit)
summary(afit)

boxcox(afit)

afit_tr <- lm(log(crimes_after_election) ~., data=after_crimes)
plot(afit_tr)
summary(afit_tr)
shapiro.test(afit_tr$residuals)

cor(before_crimes)
vif(bfit_tr)

bfit_re <- lm(sqrt(crimes_avg) ~ .-non_white, data = before_crimes)
cor(after_crimes)
vif(afit_tr)
afit_re <- lm(log(crimes_after_election) ~.-non_white, data=after_crimes)
```

```r
cor(after_crimes[,-8])
vif(afit_re)
# model with intercept base with washington dc
model0b <- lm(sqrt(crimes_avg) ~1, data=before_crimes)

# full multiple regression model
modelFb <- bfit_re


# AIC
M1b <- step(model0b, scope=list(lower=model0b, upper=modelFb), direction = "both", k=2)
summary(M1b)$coefficients

lm(sqrt(crimes_avg) ~ hs_degree + gini_index + hs_degree:gini_index, data=before_crimes)

ggplot(data=crimes, mapping=aes(gini_index, hs_degree, color = region)) +
  geom_point() +
  geom_smooth(method="lm", se = FALSE) +
  xlab("GINI Index") +
  ylab("Proportion with High School Degree")+
  ggtitle("GINI VS HS Degree: Influential Included")

ggplot(data=crimes[-9,], mapping=aes(gini_index, hs_degree, color = region)) +
  geom_point() +
  geom_smooth(method="lm", se = FALSE) +
  xlab("GINI Index") +
  ylab("Proportion with High School Degree")+
  ggtitle("GINI VS HS Degree: Influential Point Removed")

# model with intercept base with washington dc
model0b2 <- lm(sqrt(crimes_avg) ~1, data=before_crimes[-9,])

# full multiple regression model
modelFb2 <- lm(sqrt(crimes_avg) ~.-non_white, data=before_crimes[-9,])


# AIC
M1b2 <- step(model0b2, scope=list(lower=model0b2, upper=modelFb2), direction = "both", k=2)
summary(M1b2)

lm(sqrt(crimes_avg) ~ hs_degree + gini_index + hs_degree:gini_index, data=before_crimes[-9,])

# model with intercept
model0a <- lm(log(crimes_after_election) ~1, data=after_crimes)

# full multiple regression model
modelFa <- afit_re

# AIC
M1a <- step(model0a, scope=list(lower=model0a, upper=modelFa), direction = "both", k=2)
summary(M1a)$coefficients

# model with intercept
```

```r
model0a2 <- lm(log(crimes_after_election) ~1, data=after_crimes[-9,])

# full multiple regression model
modelFa2 <- lm(log(crimes_after_election) ~.-non_white, data=after_crimes[-9,])

# AIC
M1a2 <- step(model0a2, scope=list(lower=model0a2, upper=modelFa2), direction = "both", k=2)
summary(M1a2)$coefficients

dwt(bfit)
dwt(afit)

crimes <- crimes %>%
  mutate(
    trump_support_level = cut_number(trump_voters, 3, labels=c("low", "medium", "high"))
  )

crimes <- crimes %>%
  mutate(
    income_level = cut_number(income, 3, labels=c("low", "medium", "high"))
  )

ggplot(data = crimes, aes(x = region, y = crimes_after_election, fill = trump_support_level)) +
  geom_bar(position="stack", stat="identity")

ggplot(data = crimes, aes(x = region, y = crimes_after_election, fill = income_level)) +
  geom_bar(position="stack", stat="identity")
```