Methods

Data Cleaning

```
hate_df =
  read_csv("./data/HateCrimes.csv") %>%
  mutate(
    state = as.factor(state),
    unemployment = as.factor(unemployment),
    urbanization = as.factor(urbanization),
    hate_crimes_per_100k_splc = as.numeric(hate_crimes_per_100k_splc)
  )
```

Descriptive Statistics

```
# Table labels
my_labels =
  list(
    unemployment = "Antibody IgM",
    urbanization = "Urbanization",
    median_household_income = "Median Household Income",
    perc_population_with_high_school_degree = "Percent with HS Degree",
    perc_non_citizen = "Percent Non-Citizen",
    gini_index = "Gini Index",
    perc_non_white = "Percent Non-White",
    hate_crimes_per_100k_splc = "Hate Crimes per 100k"
)

# Table controls
my_controls = tableby.control(
  total = F,
  test = F,
  numeric.stats = c("N", "meansd", "medianq1q3", "range", "Nmiss2"),
  cat.stats = c("N", "countpct"),
  stats.labels = list(
    meansd = "Mean (SD)",
    medianq1q3 = "Median (Q1, Q3)",
    range = "Min - Max",
    Nmiss2 = "Missing",
    countpct = "N (%)",
    N = "N"
    )
  )
```

```r
# Generate table
descriptive_tab =
  tableby( ~ unemployment +
           urbanization +
           median_household_income +
           perc_population_with_high_school_degree +
           perc_non_citizen +
           gini_index +
           perc_non_white +
           hate_crimes_per_100k_splc,
           data = hate_df,
           control = my_controls)

summary(
  descriptive_tab,
  title = "Descriptive Statistics: Hate Crimes Data",
  labelTranslations = my_labels,
  text = T)
```

```
##
## Table: Descriptive Statistics: Hate Crimes Data
##
## |                        |              Overall (N=51)         |
## |:-----------------------|:-----------------------------------:|
## |Antibody IgM            |                                     |
## |-  N                    |                 51                  |
## |-  high                 |             24 (47.1%)              |
## |-  low                  |             27 (52.9%)              |
## |Urbanization            |                                     |
## |-  N                    |                 51                  |
## |-  high                 |             24 (47.1%)              |
## |-  low                  |             27 (52.9%)              |
## |Median Household Income |                                     |
## |-  N                    |                 51                  |
## |-  Mean (SD)            |          55223.608 (9208.478)       |
## |-  Median (Q1, Q3)      | 54916.000 (48657.000, 60719.000)    |
## |-  Min - Max            |          35521.000 - 76165.000      |
## |-  Missing              |                 0                   |
## |Percent with HS Degree  |                                     |
## |-  N                    |                 51                  |
## |-  Mean (SD)            |            0.869 (0.034)            |
## |-  Median (Q1, Q3)      |        0.874 (0.841, 0.898)         |
## |-  Min - Max            |            0.799 - 0.918            |
## |-  Missing              |                 0                   |
## |Percent Non-Citizen     |                                     |
## |-  N                    |                 48                  |
## |-  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              |                                     |
## |-  N                    |                 51                  |
## |-  Mean (SD)            |            0.454 (0.021)            |
```

```
## |-  Median (Q1, Q3)     |       0.454 (0.440, 0.467)    |
## |-  Min - Max           |          0.419 - 0.532        |
## |-  Missing             |               0               |
## |Percent Non-White      |                               |
## |-  N                   |              51               |
## |-  Mean (SD)           |          0.316 (0.165)        |
## |-  Median (Q1, Q3)     |       0.280 (0.195, 0.420)    |
## |-  Min - Max           |          0.060 - 0.810        |
## |-  Missing             |               0               |
## |Hate Crimes per 100k   |                               |
## |-  N                   |              47               |
## |-  Mean (SD)           |          0.304 (0.253)        |
## |-  Median (Q1, Q3)     |       0.226 (0.143, 0.357)    |
## |-  Min - Max           |          0.067 - 1.522        |
## |-  Missing             |               4               |
```

As a note, I didn't include the "states" variable as the output was huge and not that helpful. Suggest we include a note somewhere that data from 50 states + Washington, DC.

Distribution of Outcome Data

Histogram of raw outcome data (hate crimes per 100k).

```
hate_df %>%
  ggplot(aes(x = hate_crimes_per_100k_splc, y = ..density..)) +
  geom_histogram(bins = 11) +
  geom_density(alpha = 0.2, color = "blue") +
  labs(
    x = "Hate Crimes per 100k",
    y = "Density"
  )
```
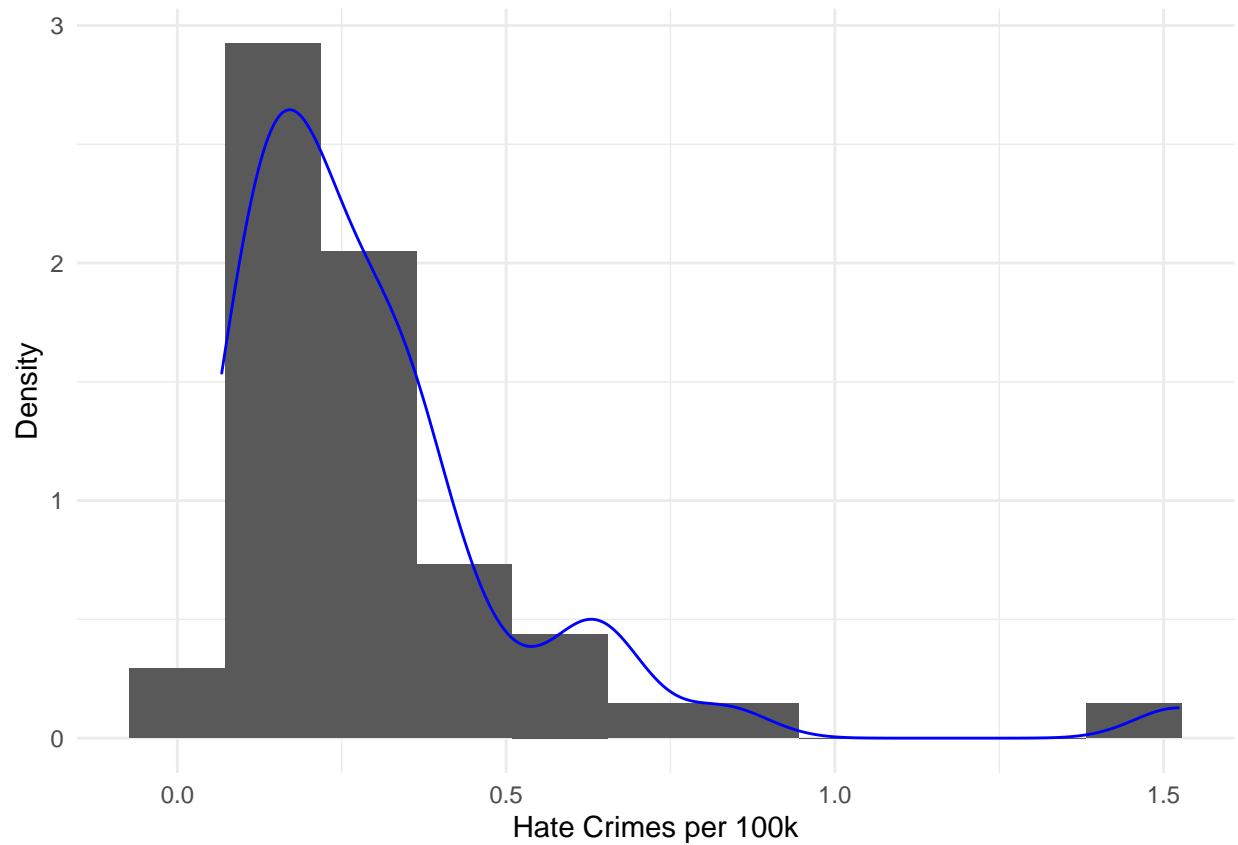
These data look skewed :(

Histogram of log-transformed outcome data (hate crimes per 100k).

```
hate_df %>%
  ggplot(aes(x = log(hate_crimes_per_100k_splc), y = ..density..)) +
  geom_histogram(bins = 11) +
  geom_density(alpha = 0.2, color = "blue") +
  labs(
    x = "Hate Crimes per 100k (Log Scale)",
    y = "Density"
  )
```
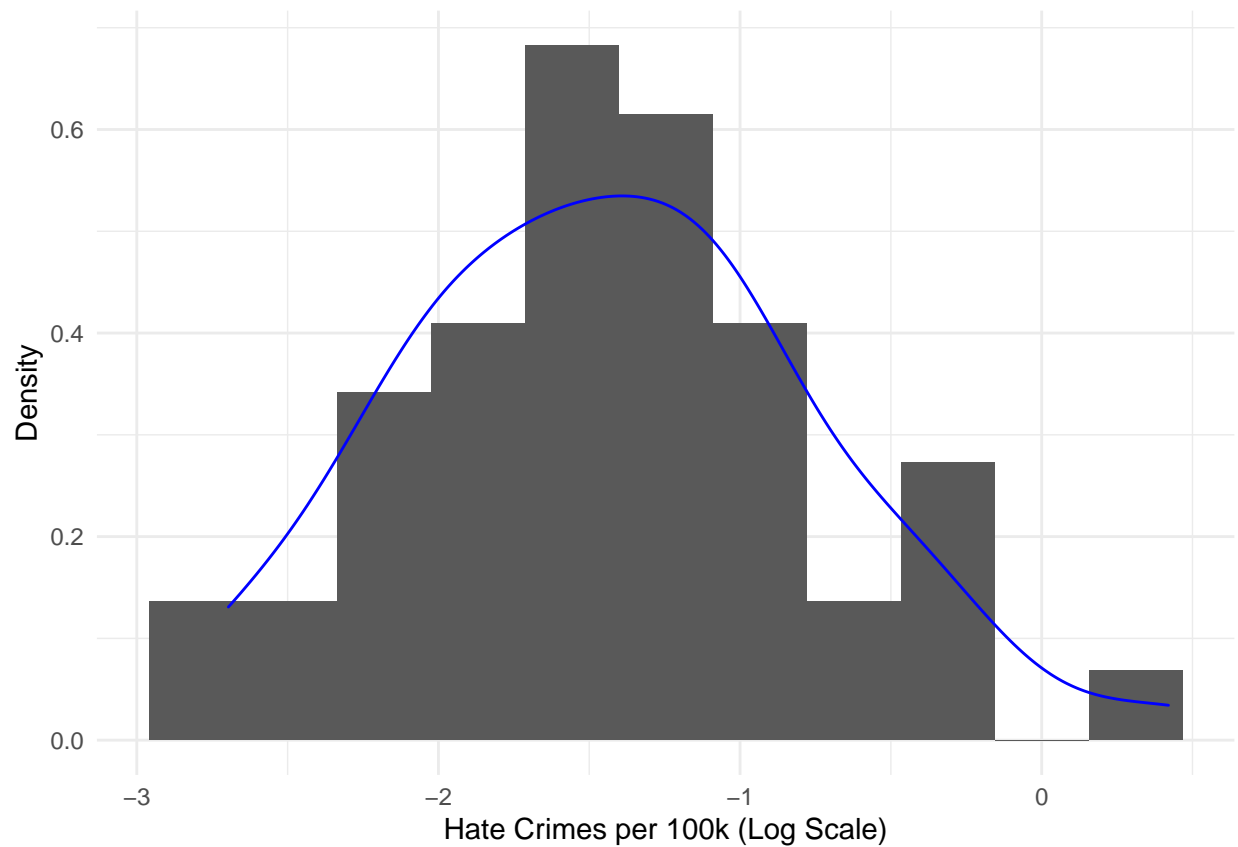
Looks better!

Box plot of the (raw) outcome data.

```
hate_df %>%
  ggplot(aes(y = hate_crimes_per_100k_splc)) +
  geom_boxplot() +
  labs(
    y = "Hate Crimes per 100k"
  ) +
  theme(
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank()
    )
```
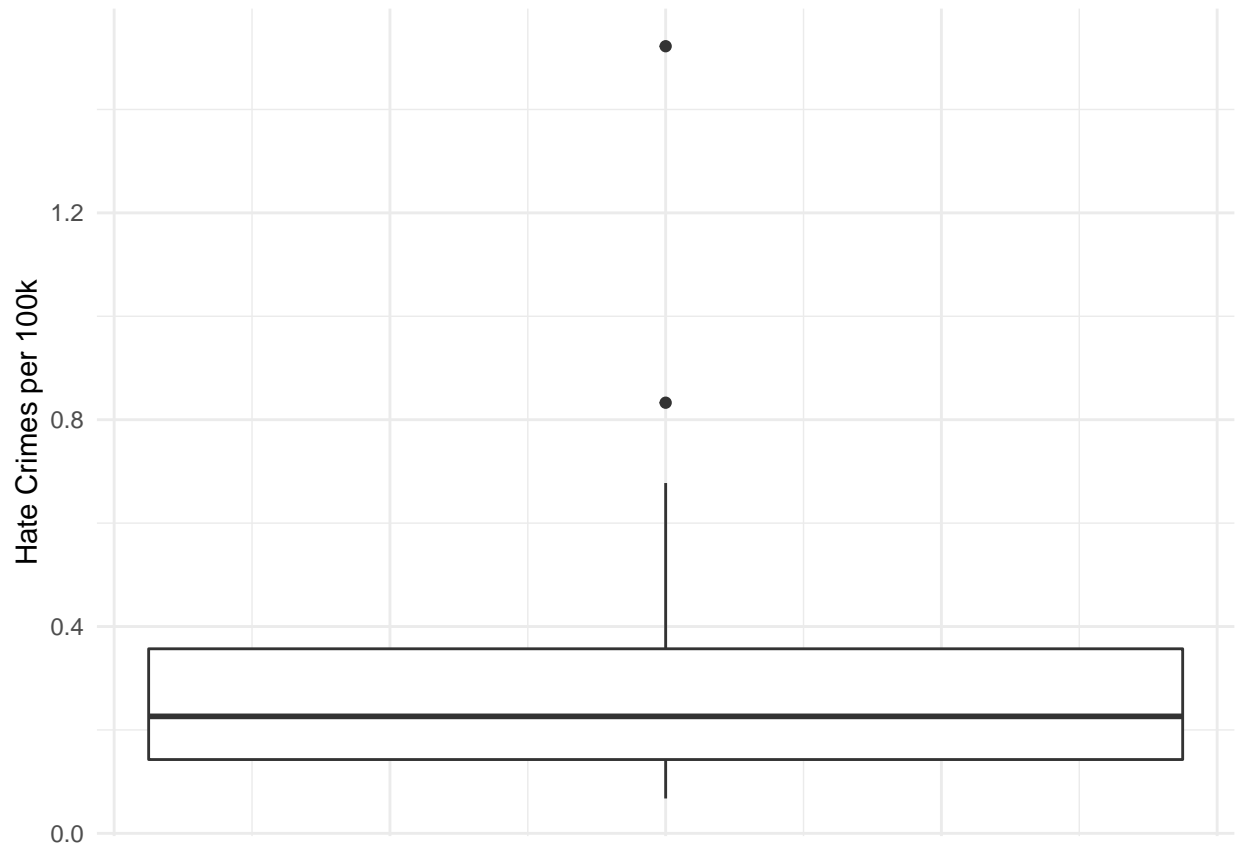
Just based on the boxplot, it ks like there are two states with potential usually high rates (Washington, DC and Oregon).

Examining Potential Multicollinearity

```
hate_df %>%
  select(
    hate_crimes_per_100k_splc,
    median_household_income,
    perc_population_with_high_school_degree,
    perc_non_citizen,
    gini_index,
    perc_non_white
    ) %>%
  cor(use = "complete.obs") %>% # Ignoring NA values
  round(., 2)
```

```
##                                         hate_crimes_per_100k_splc
## hate_crimes_per_100k_splc                                    1.00
## median_household_income                                      0.34
## perc_population_with_high_school_degree                      0.26
## perc_non_citizen                                             0.24
## gini_index                                                   0.38
```

6

```
## perc_non_white                                                     0.11
##                                        median_household_income
## hate_crimes_per_100k_splc                                      0.34
## median_household_income                                        1.00
## perc_population_with_high_school_degree                        0.65
## perc_non_citizen                                               0.30
## gini_index                                                    -0.13
## perc_non_white                                                 0.04
##                                        perc_population_with_high_school_degree
## hate_crimes_per_100k_splc                                              0.26
## median_household_income                                                0.65
## perc_population_with_high_school_degree                                1.00
## perc_non_citizen                                                      -0.26
## gini_index                                                            -0.54
## perc_non_white                                                        -0.50
##                                        perc_non_citizen gini_index
## hate_crimes_per_100k_splc                          0.24       0.38
## median_household_income                            0.30      -0.13
## perc_population_with_high_school_degree           -0.26      -0.54
## perc_non_citizen                                   1.00       0.48
## gini_index                                         0.48       1.00
## perc_non_white                                     0.75       0.55
##                                        perc_non_white
## hate_crimes_per_100k_splc                        0.11
## median_household_income                          0.04
## perc_population_with_high_school_degree         -0.50
## perc_non_citizen                                 0.75
## gini_index                                       0.55
## perc_non_white                                   1.00
```

Based on this output, the following pairs of variables have a correlation of 60% or higher:

- Percentage non-citizens & percentage non-white (0.75)

- Median household income & percentage of population with a high school degree (0.65)

```r
hate_df %>%
  select(
    hate_crimes_per_100k_splc,
    median_household_income,
    perc_population_with_high_school_degree,
    perc_non_citizen,
    gini_index,
    perc_non_white
    ) %>%
  pairs()
```
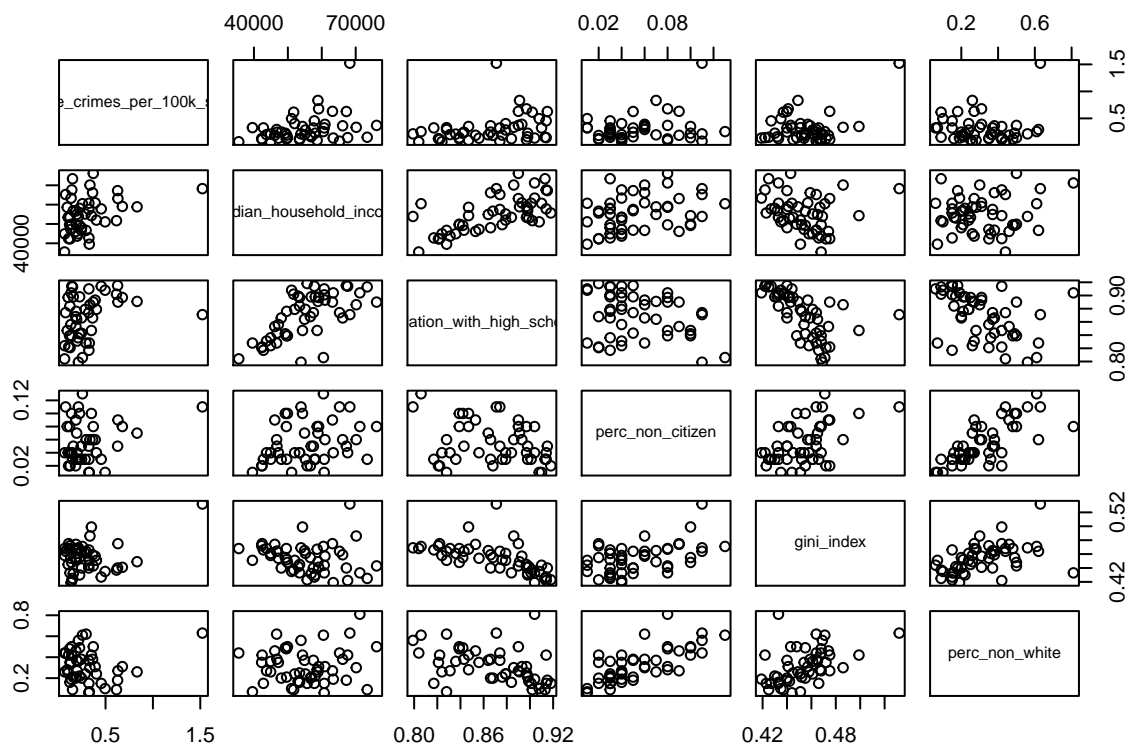
Jacy's Ideas

```r
##Project ideas
hate = read.csv("/Users/jacysparks/Downloads/HateCrimes.csv")
head(hate)
dim(hate)
hate$hate_crimes_per_100k_splc = as.character(hate$hate_crimes_per_100k_splc)
hate$hate_crimes_per_100k_splc = as.numeric(hate$hate_crimes_per_100k_splc)
summary(hate)
##Four NA's for outcome
##NA for Wyoming, South Dakota, North Dakota, and Idaho
hate[,c(1,9)]
##Could remove
hate = na.omit(hate)

##3 NA's for non citizen


##Create indicators
names(hate)[names(hate)=="unemployment"] = "High.Unemployment"
names(hate)[names(hate)=="urbanization"] = "High.Urban"
names(hate)[names(hate)=="median_household_income"] = "Med.Income"
names(hate)[names(hate)=="perc_population_with_high_school_degree"] = "HS.Degree"
```

```r
names(hate)[names(hate)=="perc_non_citizen"] = "Non.Citizen"
names(hate)[names(hate)=="perc_non_white"] = "Non.White"
names(hate)[names(hate)=="hate_crimes_per_100k_splc"] = "Hate.Crime"
hate$High.Unemployment = ifelse(hate$High.Unemployment=="high",1,0)
hate$High.Urban = ifelse(hate$High.Urban=="high",1,0)

##Outcome var is skewed
hate$Hate.Crime = log(hate$Hate.Crime)
hist(hate$Hate.Crime)
##Much better

reg = lm(Hate.Crime~.-state,data=hate)
summary(reg)

pairs(hate[,4:9],lower.panel=NULL)
cor(hate[,4:9])
#Percent white and percent non-white highly correlated


##Check linearity
for(i in 4:8){
  plot(hate[,i],hate$Hate.Crime,main=colnames(hate)[i])
}
plot(reg)
```