

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"  
  )  
)
```

```
# 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)|
```

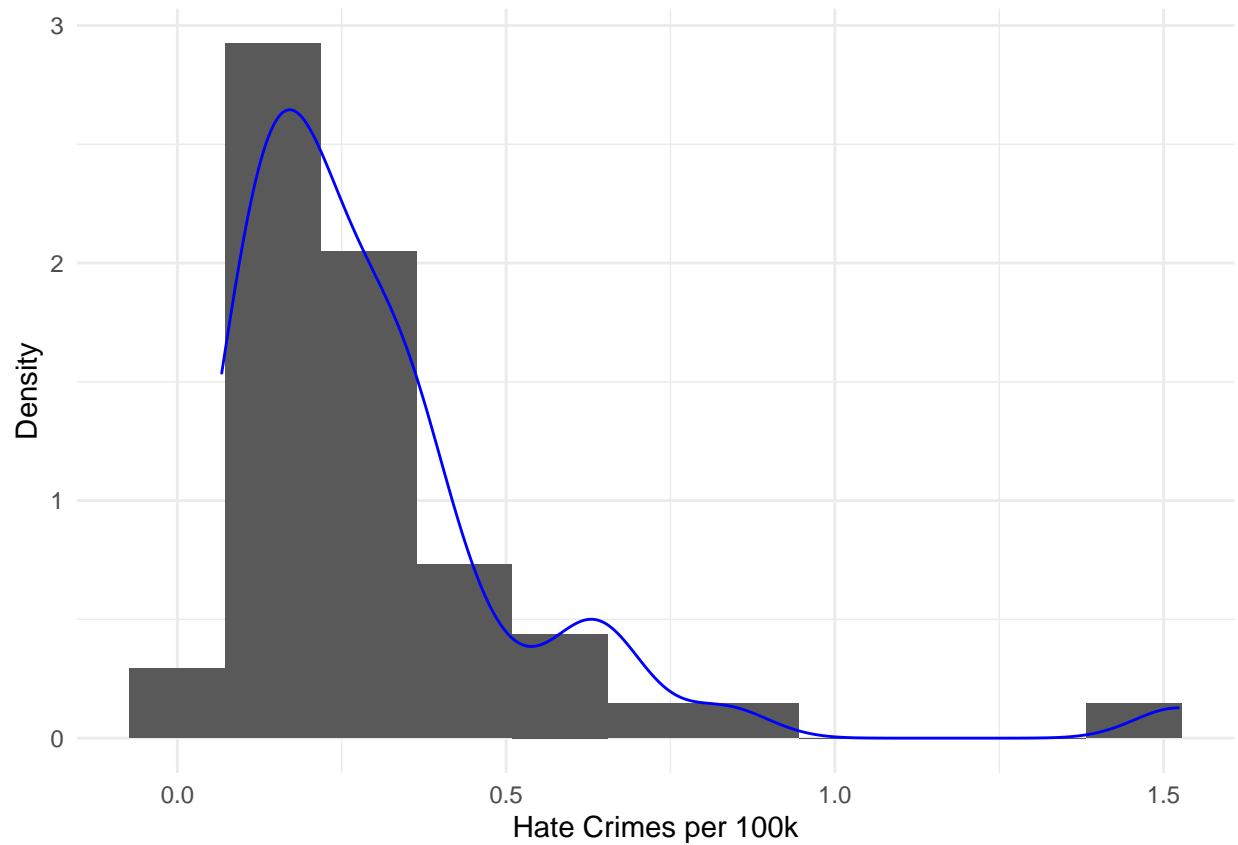
##	- Median (Q1, Q3)		0.454 (0.440, 0.467)	
##	- Min - Max		0.419 - 0.532	
##	- Missing		0	
##	Percent Non-White			
##	- N		51	
##	- Mean (SD)		0.316 (0.165)	
##	- Median (Q1, Q3)		0.280 (0.195, 0.420)	
##	- Min - Max		0.060 - 0.810	
##	- Missing		0	
##	Hate Crimes per 100k			
##	- N		47	
##	- Mean (SD)		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).

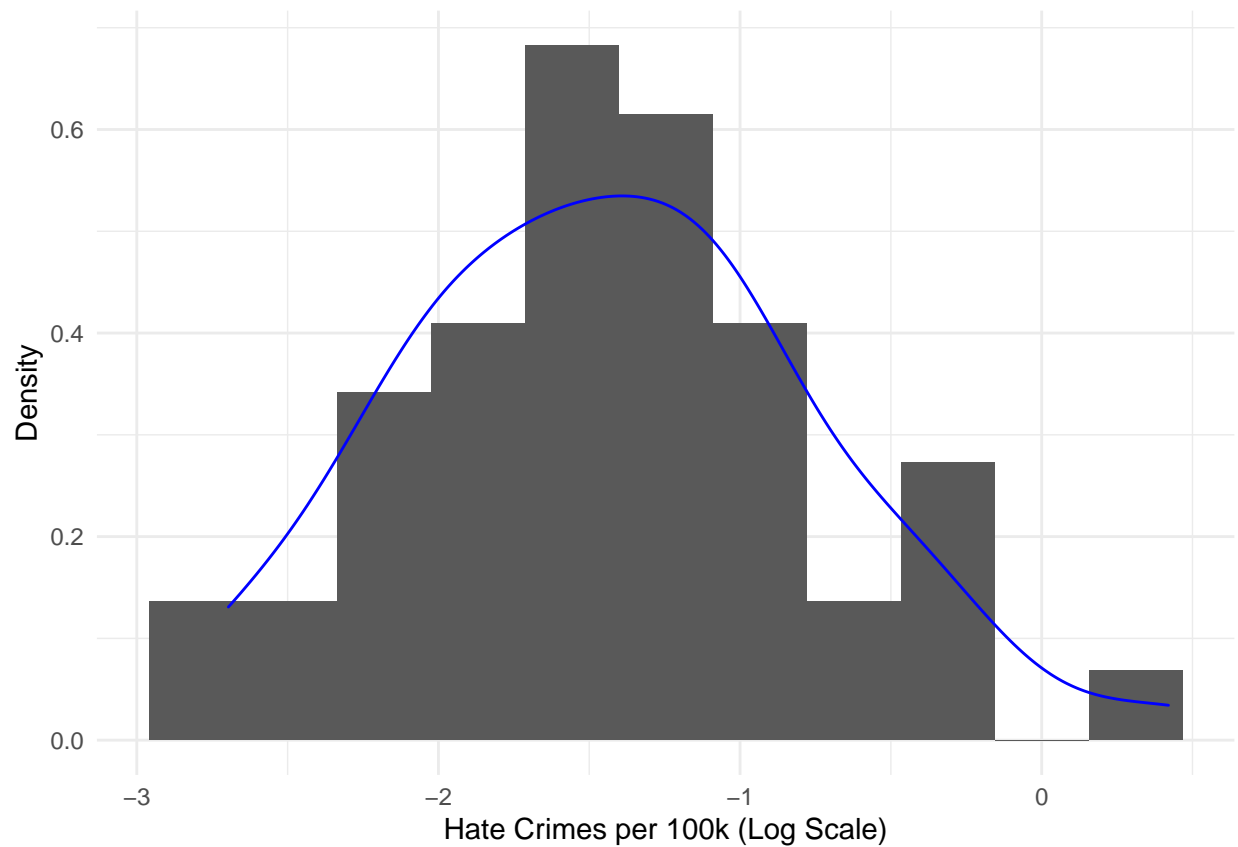
```
hate_df %>%
  ggplot(aes(x = hate_crimes_per_100k_splc, y = ..density..)) +
  geom_histogram(bins = 11) +
  geom_density(alpha = 0.2, color = "blue") +
  labs(
    x = "Hate Crimes per 100k",
    y = "Density"
  )
```



These data look skewed :(

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

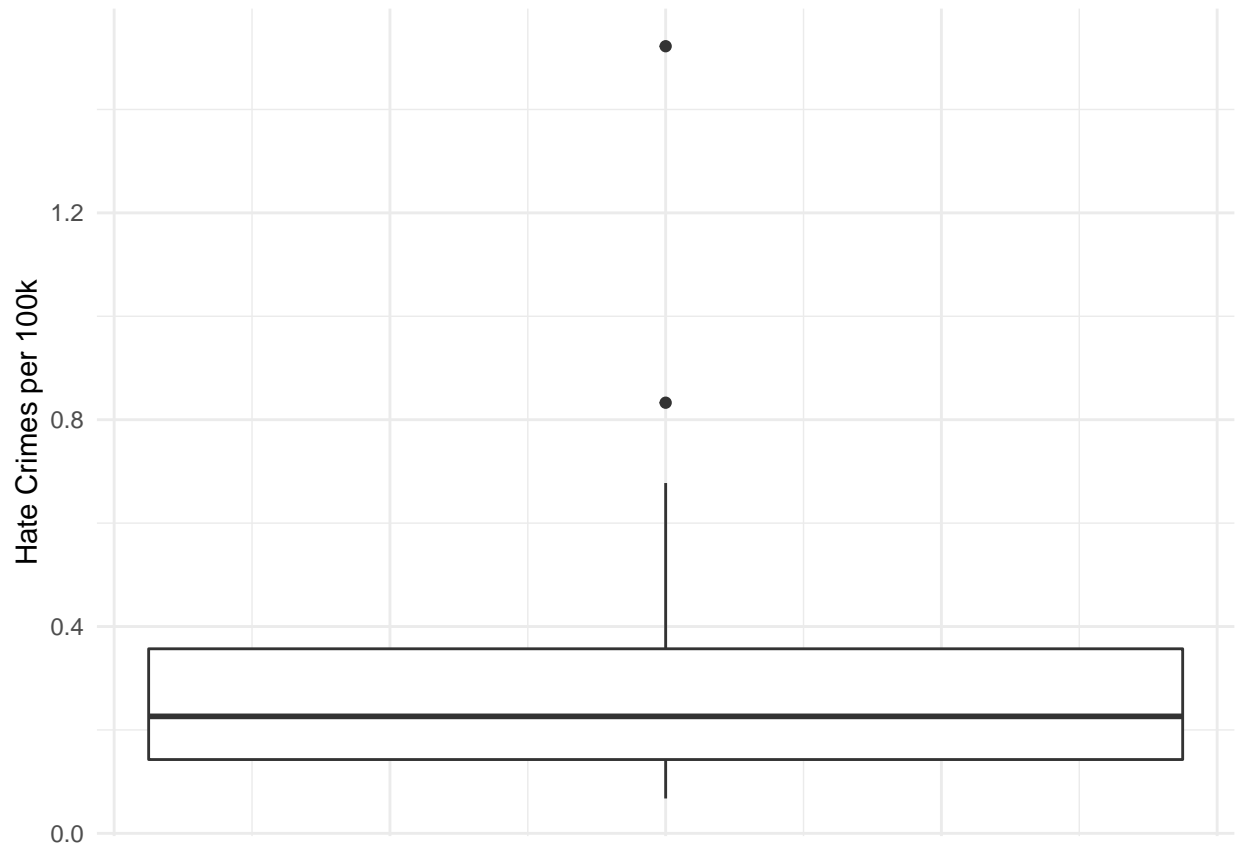
```
hate_df %>%  
  ggplot(aes(x = log(hate_crimes_per_100k_splc), y = ..density..)) +  
  geom_histogram(bins = 11) +  
  geom_density(alpha = 0.2, color = "blue") +  
  labs(  
    x = "Hate Crimes per 100k (Log Scale)",  
    y = "Density"  
  )
```



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
```

```

## 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

```

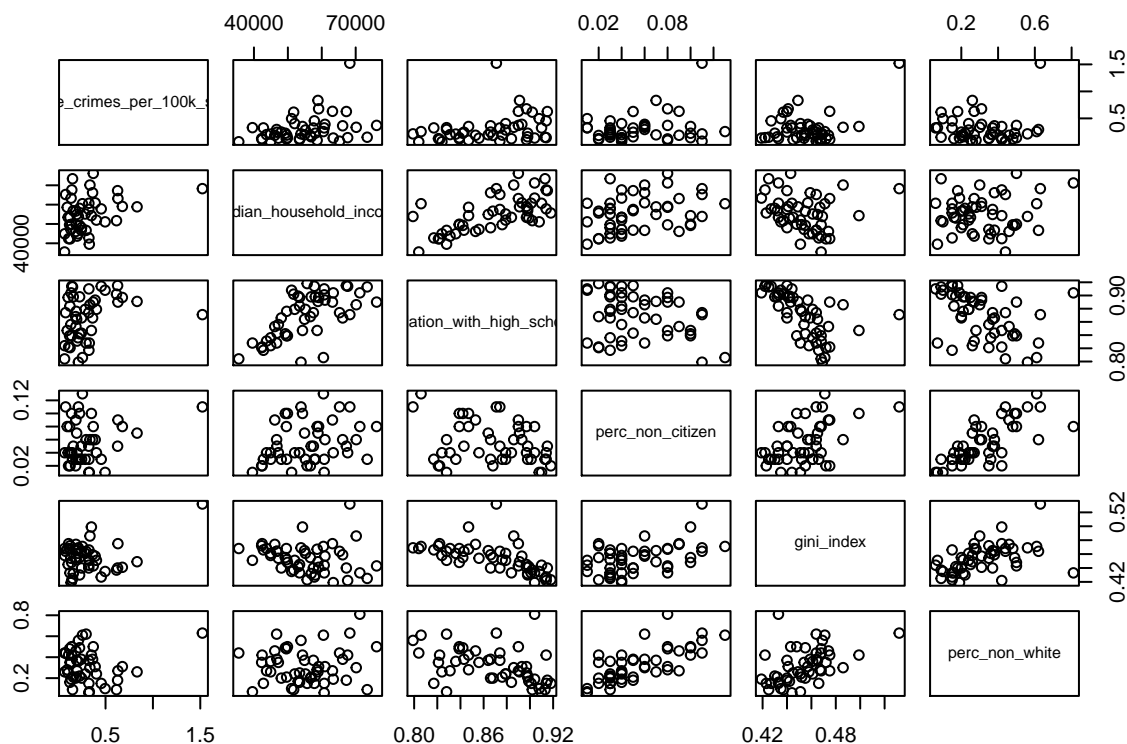
Based on this output, the following pairs of variables have a correlation of 60% or higher:

- Percentage non-citizens & percentage non-white (0.75)
- Median household income & percentage of population with a high school degree (0.65)

```

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()

```



Simple Linear Regression Using Income Inequality (Per FiveThirtyEight)

Fitting SLR using income inequality (measured by Gini index) per FiveThirtyEight findings.

```
slr_income_lm = lm(hate_crimes_per_100k_splc ~ gini_index, data = hate_df)
slr_income_log_lm = lm(log(hate_crimes_per_100k_splc) ~ gini_index, data = hate_df)

summary(slr_income_lm)
```

```
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ gini_index, data = hate_df)
##
## 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 .
## gini_index     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
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.1085, Adjusted R-squared:  0.08872
## F-statistic: 5.478 on 1 and 45 DF,  p-value: 0.02374
```

```
summary(slr_income_log_lm)
```

```
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ gini_index, data = hate_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32883 -0.36358 -0.02325  0.38705  1.47219
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.676      2.195  -1.674   0.101
## gini_index      4.932      4.814   1.024   0.311
##
## Residual standard error: 0.6761 on 45 degrees of freedom
## (4 observations deleted due to missingness)
## Multiple R-squared:  0.02279,    Adjusted R-squared:  0.001073
## F-statistic: 1.049 on 1 and 45 DF,  p-value: 0.3111
```

Gini index appears to be a significant predictor only when using the raw outcome data (not the log-transformed outcome data).

Trying Stepwise Approach

As a first step, trying stepwise approach.

```
hate_nona_df = # Removing rows with missing values from the dataset
hate_df %>%
  drop_na()

full_lm = lm(
  hate_crimes_per_100k_splc
  ~ unemployment +
    urbanization +
    median_household_income +
    perc_population_with_high_school_degree +
    perc_non_citizen +
    gini_index +
    perc_non_white,
  data = hate_nona_df
)

full_log_lm = lm(
```

```

log(hate_crimes_per_100k_splc)
~ unemployment +
  urbanization +
  median_household_income +
  perc_population_with_high_school_degree +
  perc_non_citizen +
  gini_index +
  perc_non_white,
data = hate_nona_df
)

step(full_lm, direction = "both") # Trying "both" directions

## Start:  AIC=-137.03
## hate_crimes_per_100k_splc ~ unemployment + urbanization + median_household_income +
##   perc_population_with_high_school_degree + perc_non_citizen +
##   gini_index + perc_non_white
##
##
##           Df Sum of Sq   RSS   AIC
## - perc_non_white      1    0.00001 1.5008 -139.03
## - unemployment       1    0.00135 1.5021 -138.99
## - median_household_income 1    0.00258 1.5034 -138.95
## - urbanization        1    0.00618 1.5070 -138.85
## - perc_non_citizen     1    0.01750 1.5183 -138.51
## <none>                  1.5008 -137.03
## - perc_population_with_high_school_degree 1    0.34889 1.8497 -129.62
## - gini_index           1    0.77465 2.2754 -120.30
##
## Step:  AIC=-139.03
## hate_crimes_per_100k_splc ~ unemployment + urbanization + median_household_income +
##   perc_population_with_high_school_degree + perc_non_citizen +
##   gini_index
##
##
##           Df Sum of Sq   RSS   AIC
## - unemployment       1    0.00148 1.5023 -140.99
## - median_household_income 1    0.00269 1.5035 -140.95
## - urbanization        1    0.00617 1.5070 -140.85
## - perc_non_citizen     1    0.02422 1.5250 -140.31
## <none>                  1.5008 -139.03
## + perc_non_white      1    0.00001 1.5008 -137.03
## - perc_population_with_high_school_degree 1    0.38759 1.8884 -130.69
## - gini_index           1    0.77888 2.2797 -122.22
##
## Step:  AIC=-140.99
## hate_crimes_per_100k_splc ~ urbanization + median_household_income +
##   perc_population_with_high_school_degree + perc_non_citizen +
##   gini_index
##
##
##           Df Sum of Sq   RSS   AIC
## - median_household_income 1    0.00243 1.5047 -142.91
## - urbanization            1    0.00693 1.5092 -142.78
## - perc_non_citizen        1    0.02401 1.5263 -142.27
## <none>                    1.5023 -140.99

```

```

## + unemployment          1  0.00148 1.5008 -139.03
## + perc_non_white         1  0.00015 1.5021 -138.99
## - perc_population_with_high_school_degree 1  0.40517 1.9074 -132.24
## - gini_index             1  0.78876 2.2910 -124.00
##
## Step: AIC=-142.91
## hate_crimes_per_100k_splc ~ urbanization + perc_population_with_high_school_degree +
##   perc_non_citizen + gini_index
##
##              Df Sum of Sq   RSS   AIC
## - urbanization      1  0.00762 1.5123 -144.69
## - perc_non_citizen   1  0.02232 1.5270 -144.25
## <none>                1.5047 -142.91
## + median_household_income      1  0.00243 1.5023 -140.99
## + unemployment              1  0.00122 1.5035 -140.95
## + perc_non_white              1  0.00034 1.5044 -140.92
## - gini_index                  1  0.78737 2.2921 -125.97
## - perc_population_with_high_school_degree 1  0.86254 2.3672 -124.52
##
## Step: AIC=-144.69
## hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
##   perc_non_citizen + gini_index
##
##              Df Sum of Sq   RSS   AIC
## - perc_non_citizen      1  0.01471 1.5270 -146.25
## <none>                1.5123 -144.69
## + urbanization          1  0.00762 1.5047 -142.91
## + median_household_income      1  0.00311 1.5092 -142.78
## + unemployment          1  0.00192 1.5104 -142.74
## + perc_non_white         1  0.00028 1.5120 -142.69
## - gini_index              1  0.78804 2.3004 -127.81
## - perc_population_with_high_school_degree 1  0.85561 2.3679 -126.51
##
## Step: AIC=-146.25
## hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
##   gini_index
##
##              Df Sum of Sq   RSS   AIC
## <none>                1.5270 -146.25
## + perc_non_citizen      1  0.01471 1.5123 -144.69
## + perc_non_white         1  0.00522 1.5218 -144.40
## + unemployment          1  0.00136 1.5257 -144.29
## + median_household_income      1  0.00068 1.5263 -144.27
## + urbanization          1  0.00001 1.5270 -144.25
## - perc_population_with_high_school_degree 1  0.85432 2.3813 -128.25
## - gini_index             1  1.06513 2.5922 -124.44
##
## Call:
## lm(formula = hate_crimes_per_100k_splc ~ perc_population_with_high_school_degree +
##   gini_index, data = hate_nona_df)
##
## Coefficients:
##
##              (Intercept)

```

```
## -8.103
## perc_population_with_high_school_degree
## 5.059
## gini_index
## 8.825
```

```
step(full_log_lm, direction = "both") # Trying "both" directions
```

```
## Start: AIC=-40.88
## log(hate_crimes_per_100k_splc) ~ unemployment + urbanization +
## median_household_income + perc_population_with_high_school_degree +
## perc_non_citizen + gini_index + perc_non_white
##
## Df Sum of Sq RSS AIC
## - perc_non_white 1 0.00456 12.719 -42.859
## - perc_non_citizen 1 0.01570 12.730 -42.820
## - median_household_income 1 0.02556 12.740 -42.785
## - urbanization 1 0.05519 12.770 -42.680
## - unemployment 1 0.37413 13.089 -41.570
## <none> 12.715 -40.875
## - perc_population_with_high_school_degree 1 1.51318 14.228 -37.815
## - gini_index 1 2.90660 15.621 -33.611
##
## Step: AIC=-42.86
## log(hate_crimes_per_100k_splc) ~ unemployment + urbanization +
## median_household_income + perc_population_with_high_school_degree +
## perc_non_citizen + gini_index
##
## Df Sum of Sq RSS AIC
## - perc_non_citizen 1 0.01114 12.730 -44.820
## - median_household_income 1 0.02946 12.749 -44.755
## - urbanization 1 0.05718 12.777 -44.657
## - unemployment 1 0.41699 13.136 -43.408
## <none> 12.719 -42.859
## + perc_non_white 1 0.00456 12.715 -40.875
## - perc_population_with_high_school_degree 1 1.73309 14.452 -39.111
## - gini_index 1 2.90620 15.626 -35.599
##
## Step: AIC=-44.82
## log(hate_crimes_per_100k_splc) ~ unemployment + urbanization +
## median_household_income + perc_population_with_high_school_degree +
## gini_index
##
## Df Sum of Sq RSS AIC
## - median_household_income 1 0.01910 12.750 -46.752
## - urbanization 1 0.11092 12.841 -46.429
## - unemployment 1 0.41466 13.145 -45.377
## <none> 12.730 -44.820
## + perc_non_citizen 1 0.01114 12.719 -42.859
## + perc_non_white 1 0.00000 12.730 -42.820
## - perc_population_with_high_school_degree 1 1.92883 14.659 -40.471
## - gini_index 1 3.00737 15.738 -37.277
##
## Step: AIC=-46.75
```

```

## log(hate_crimes_per_100k_splc) ~ unemployment + urbanization +
##   perc_population_with_high_school_degree + gini_index
##
##
##           Df Sum of Sq   RSS   AIC
## - urbanization      1    0.09183 12.841 -48.429
## - unemployment      1    0.40424 13.154 -47.348
## <none>                      12.750 -46.752
## + median_household_income      1    0.01910 12.730 -44.820
## + perc_non_white      1    0.00293 12.747 -44.763
## + perc_non_citizen      1    0.00077 12.749 -44.755
## - gini_index      1    3.02492 15.774 -39.172
## - perc_population_with_high_school_degree 1    3.15688 15.906 -38.797
##
## Step:  AIC=-48.43
## log(hate_crimes_per_100k_splc) ~ unemployment + perc_population_with_high_school_degree +
##   gini_index
##
##
##           Df Sum of Sq   RSS   AIC
## - unemployment      1    0.3655 13.207 -49.166
## <none>                      12.841 -48.429
## + urbanization      1    0.0918 12.750 -46.752
## + perc_non_citizen      1    0.0417 12.800 -46.576
## + perc_non_white      1    0.0049 12.837 -46.446
## + median_household_income      1    0.0000 12.841 -46.429
## - perc_population_with_high_school_degree 1    3.3459 16.187 -40.010
## - gini_index      1    4.0555 16.897 -38.079
##
## Step:  AIC=-49.17
## log(hate_crimes_per_100k_splc) ~ perc_population_with_high_school_degree +
##   gini_index
##
##
##           Df Sum of Sq   RSS   AIC
## <none>                      13.207 -49.166
## + unemployment      1    0.3655 12.841 -48.429
## + urbanization      1    0.0531 13.154 -47.348
## + perc_non_citizen      1    0.0288 13.178 -47.265
## + perc_non_white      1    0.0026 13.204 -47.175
## + median_household_income      1    0.0001 13.207 -47.167
## - gini_index      1    3.7171 16.924 -40.007
## - perc_population_with_high_school_degree 1    4.4569 17.664 -38.081
##
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ perc_population_with_high_school_degree +
##   gini_index, data = hate_nona_df)
##
## Coefficients:
##              (Intercept)
##                   -18.95
## perc_population_with_high_school_degree
##                   11.55
##              gini_index
##                   16.49

```

This procedure is retaining the following two predictors:

- Percent population with high school degree
- Gini index

Jacy's Ideas

```
##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"
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)
```