Methods

Data Cleaning

Read and clean hate crime data.

```
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

Create table of descriptive statistics.

```
# Table labels
my_labels =
  list(
    unemployment = "Unemployment",
    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 = "Table 1: Descriptive Statistics: Hate Crimes Data",
  labelTranslations = my_labels,
  text = T)
```

```
##
## Table: Table 1: Descriptive Statistics: Hate Crimes Data
##
## |                        |          Overall (N=51)          |
## |:-----------------------|:--------------------------------:|
## |Unemployment            |                                  |
## |-  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                  |
```

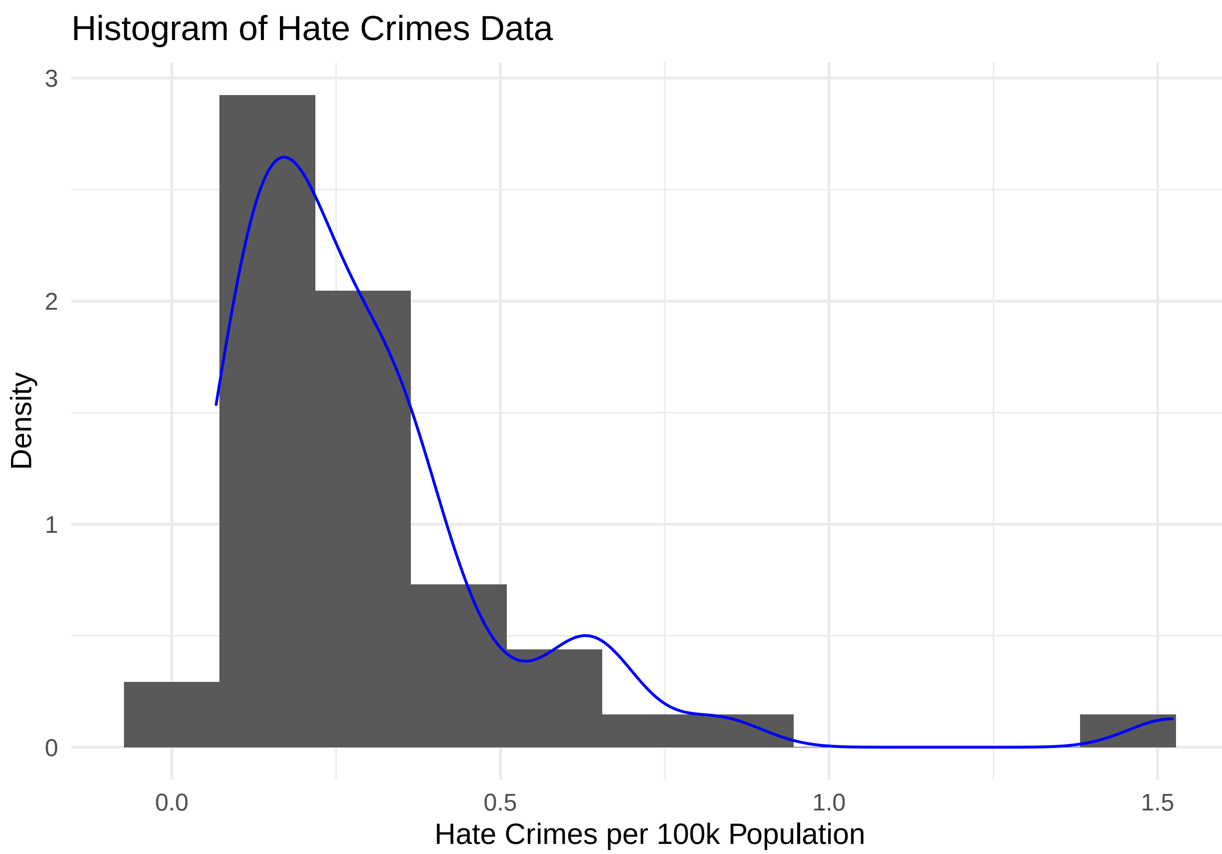
Remove Missing Data

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

Distribution of Outcome Data

Plot a histogram of raw outcome data (hate crimes per 100k) to assess distribution shape.

```r
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 Population",
    y = "Density",
    title = "Histogram of Hate Crimes Data"
  )
```

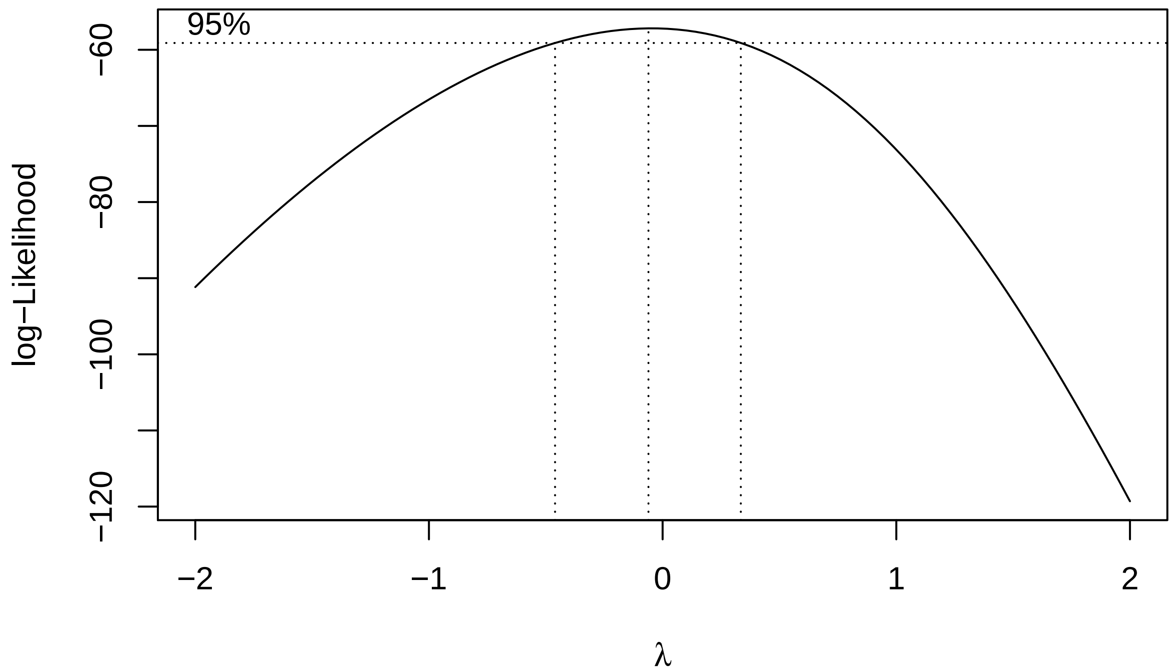


These data look skewed, so we use the Box-Cox method to determine the best transformation for the data. First, fit a linear model with all main effects.

```
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
)
```

Run boxcox on this model.

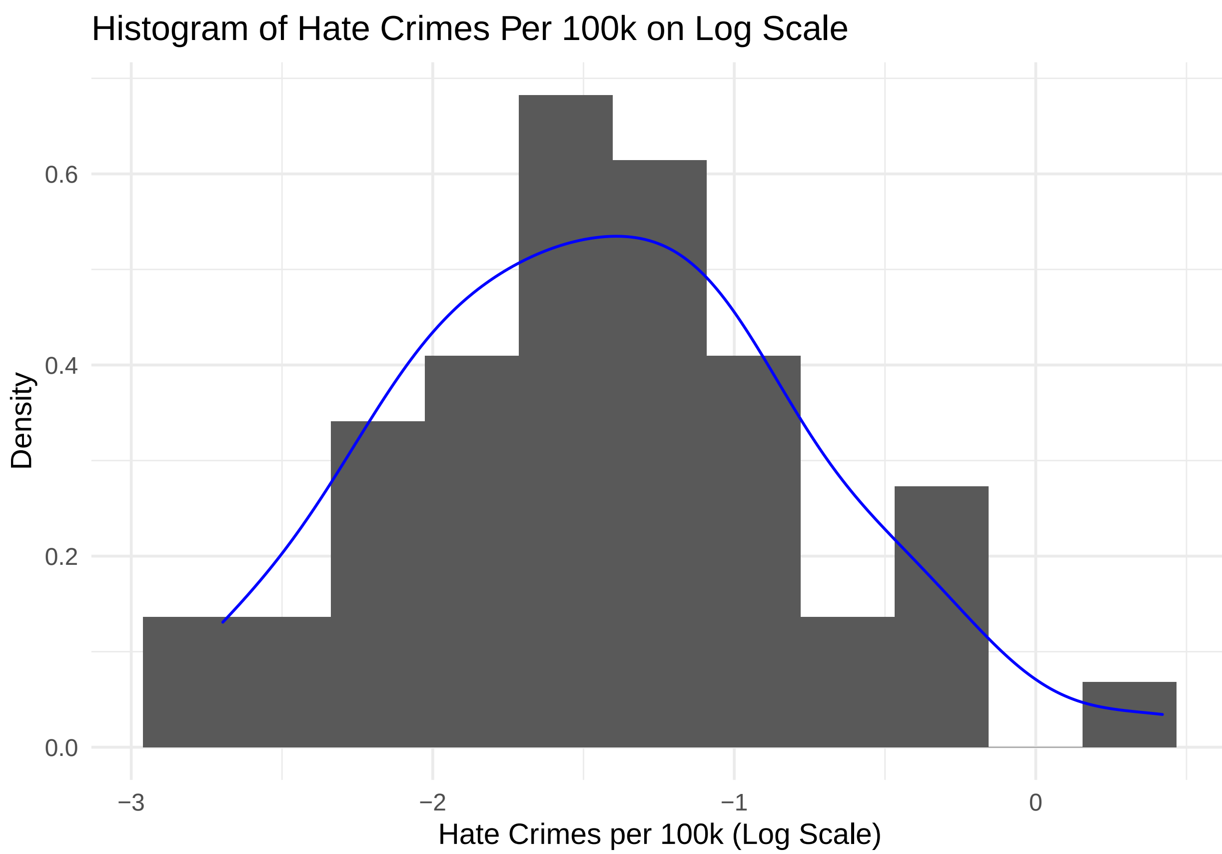
```
MASS::boxcox(full_lm)
```



The optimal value of Y^a is near 0, indicating that a natural log transformation of the outcome for all practical intents and purposes is best. We proceed with the log transformation.

Create a 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",
    title = "Histogram of Hate Crimes Per 100k on Log Scale"
  )
```



This plot looks less skewed.

Examining Potential Multicollinearity

Examine correlations between predictors.

```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
    ) %>%
  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
```

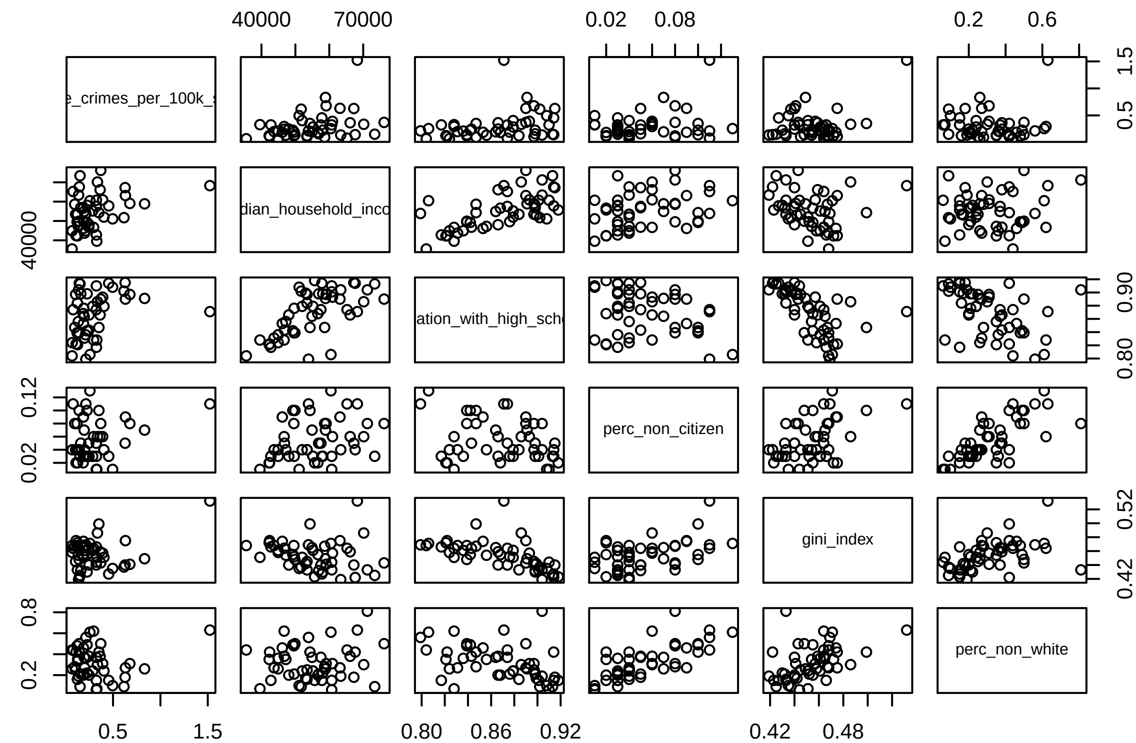
```
## perc_non_white                                                    0.11
##                                           median_household_income
## hate_crimes_per_100k_splc                                    0.34
## median_household_income                                      1.00
## perc_population_with_high_school_degree                      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)

Use a pairs plot to visually assess potential multicollinearity.

```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()
```



Selecting a Model Using the Stepwise Approach

First, consider all main effects in the model, using a log transformation of the outcome.

```
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
)

summary(full_log_lm)
```

```
##
## Call:
## lm(formula = 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)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28845 -0.41144  0.01898  0.31334  1.13022
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             -1.857e+01  5.553e+00  -3.344  0.00190
## unemploymentlow                          2.179e-01  2.088e-01   1.043  0.30353
## urbanizationlow                         -9.885e-02  2.467e-01  -0.401  0.69092
## median_household_income                 -4.732e-06  1.735e-05  -0.273  0.78658
## perc_population_with_high_school_degree  1.121e+01  5.341e+00   2.098  0.04275
## perc_non_citizen                         1.168e+00  5.464e+00   0.214  0.83189
## gini_index                               1.670e+01  5.744e+00   2.908  0.00611
## perc_non_white                          -1.232e-01  1.069e+00  -0.115  0.90887
##
## (Intercept)                             **
## unemploymentlow
## urbanizationlow
## median_household_income
## perc_population_with_high_school_degree *
## perc_non_citizen
## gini_index                              **
## perc_non_white
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5862 on 37 degrees of freedom
## Multiple R-squared:  0.3146, Adjusted R-squared:  0.1849
## F-statistic: 2.426 on 7 and 37 DF,  p-value: 0.03768
```

Use the Stepwise approach.

```
step(full_log_lm, direction = "both")
```

This procedure retains the following two predictors:

- Precent population with high school degree
- Gini index

Fit a linear regression model based on the results of the stepwise procedure.

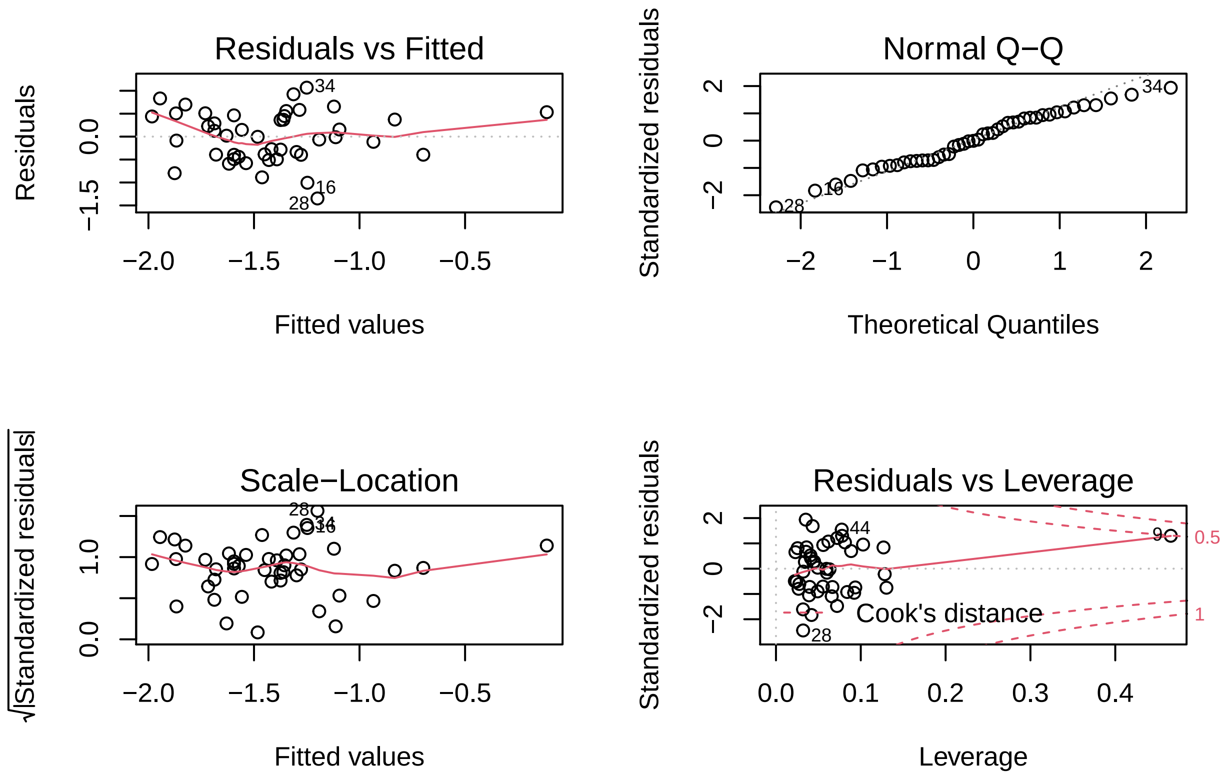
```
stepwise_log_lm = lm(
  log(hate_crimes_per_100k_splc)
  ~ perc_population_with_high_school_degree +
    gini_index,
  data = hate_nona_df)

summary(stepwise_log_lm)
```

```
##
## Call:
```

```
## lm(formula = log(hate_crimes_per_100k_splc) ~ perc_population_with_high_school_degree +
##     gini_index, data = hate_nona_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34787 -0.39659 -0.00387  0.44892  1.06723
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              -18.947      4.254  -4.454 6.14e-05
## perc_population_with_high_school_degree   11.554      3.069   3.765 0.000512
## gini_index                                16.486      4.795   3.438 0.001334
##
## (Intercept)                             ***
## perc_population_with_high_school_degree ***
## gini_index                              **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5608 on 42 degrees of freedom
## Multiple R-squared:  0.288,  Adjusted R-squared:  0.2541
## F-statistic: 8.496 on 2 and 42 DF,  p-value: 0.0007974
```

Check model assumptions using this model.

```r
par(mfrow = c(2, 2))
plot(stepwise_log_lm)
```



Influential Points

Determine if DC is an influential point quantitatively using the Cook's value. Note that on the previous assumptions plot, DC is point 9 and looks like it has a high Cook's Distance.

Check for influential points and create a dataframe without the Washington, DC point to see what impact that has on the model.

```
influence.measures(stepwise_log_lm)
```

```
## Influence measures of
##   lm(formula = log(hate_crimes_per_100k_splc) ~ perc_population_with_high_school_degree +      gini_
##
##        dfb.1_ dfb.p___  dfb.gn_n    dffit cov.r   cook.d    hat inf
## 1  -0.081618  0.13038 -4.05e-03 -0.19261 1.109 1.25e-02 0.0666
## 2  -0.013893 -0.10481  1.49e-01 -0.30406 1.109 3.09e-02 0.0922
## 3   0.054057 -0.05763 -3.27e-02  0.08493 1.108 2.45e-03 0.0412
## 4  -0.304896  0.34577  1.64e-01 -0.41699 0.987 5.63e-02 0.0720
## 5   0.191496 -0.26261 -4.75e-02  0.32205 1.122 3.47e-02 0.1027
## 6  -0.039118  0.04304  2.54e-02  0.06064 1.119 1.25e-03 0.0450
## 7   0.251790 -0.19747 -2.54e-01 -0.29040 1.186 2.84e-02 0.1302
## 8   0.068771 -0.02700 -9.62e-02  0.16236 1.060 8.85e-03 0.0360
## 9  -1.027893  0.67191  1.19e+00  1.22378 1.778 4.91e-01 0.4655   *
## 10  0.046163 -0.00709 -8.54e-02 -0.14609 1.078 7.20e-03 0.0398
```

```
## 11 -0.053388  0.10169 -2.60e-02 -0.21260 1.032 1.50e-02 0.0390
## 12 -0.075257  0.00825  1.31e-01 -0.20352 1.067 1.39e-02 0.0491
## 13  0.031597 -0.01738 -4.38e-02 -0.10108 1.077 3.46e-03 0.0275
## 14  0.063828 -0.04013 -7.24e-02  0.10745 1.099 3.92e-03 0.0407
## 15 -0.024348  0.13848 -1.15e-01  0.30940 1.082 3.18e-02 0.0813
## 16  0.133702 -0.22611  3.91e-03 -0.39474 0.874 4.90e-02 0.0419
## 17  0.235394 -0.30556 -7.86e-02  0.38074 1.031 4.75e-02 0.0780
## 18 -0.093402  0.17570 -3.85e-02 -0.28997 1.056 2.79e-02 0.0658
## 19 -0.013703  0.04710 -2.68e-02  0.12712 1.079 5.46e-03 0.0355
## 20 -0.182960  0.15692  1.69e-01  0.21549 1.139 1.57e-02 0.0886
## 21 -0.019768  0.03790 -2.98e-03  0.13067 1.052 5.74e-03 0.0255
## 22 -0.154266  0.24423  1.03e-02  0.34035 1.041 3.82e-02 0.0722
## 23 -0.000508 -0.00222  8.76e-04 -0.07363 1.081 1.84e-03 0.0223
## 24 -0.056507  0.14587 -6.12e-02  0.27564 1.054 2.52e-02 0.0618
## 25 -0.014601 -0.04569  7.99e-02 -0.17059 1.098 9.82e-03 0.0554
## 26 -0.031721  0.03033  2.39e-02 -0.03988 1.142 5.43e-04 0.0600
## 27  0.002872 -0.10873  1.21e-01 -0.27889 1.104 2.60e-02 0.0840
## 28  0.248803 -0.20839 -2.46e-01 -0.47387 0.701 6.58e-02 0.0320   *
## 29  0.127658 -0.16623 -4.07e-02  0.22682 1.069 1.72e-02 0.0558
## 30  0.049492 -0.02110 -7.20e-02 -0.08242 1.229 2.32e-03 0.1282   *
## 31  0.015806 -0.02440  4.52e-04  0.04951 1.107 8.36e-04 0.0341
## 32  0.008525 -0.01768  2.00e-03 -0.07960 1.081 2.15e-03 0.0241
## 33 -0.054224  0.05286  3.67e-02 -0.13131 1.055 5.80e-03 0.0267
## 34 -0.140534  0.20997  2.97e-02  0.38273 0.841 4.55e-02 0.0351
## 35  0.011274 -0.01075 -9.53e-03 -0.02090 1.110 1.49e-04 0.0322
## 36 -0.029075  0.09228 -6.52e-02 -0.29814 0.917 2.85e-02 0.0319
## 37  0.026918 -0.03342 -1.04e-02  0.04908 1.120 8.21e-04 0.0442
## 38  0.003366 -0.00523  2.32e-05  0.00839 1.131 2.40e-05 0.0492
## 39  0.214374 -0.27630 -7.60e-02  0.31860 1.170 3.41e-02 0.1268
## 40 -0.065943 -0.02807  1.58e-01 -0.23735 1.140 1.90e-02 0.0936
## 41  0.003208 -0.00467 -7.03e-04 -0.00639 1.147 1.39e-05 0.0631
## 42 -0.011898  0.00781  1.74e-02  0.09922 1.067 3.33e-03 0.0229
## 43 -0.070951  0.17415 -6.38e-02  0.36609 0.911 4.27e-02 0.0432
## 44  0.372189 -0.38111 -2.51e-01  0.45593 0.978 6.69e-02 0.0776
## 45 -0.000251 -0.00039  9.22e-04 -0.00177 1.143 1.07e-06 0.0595
```

```r
stu_res <- rstandard(stepwise_log_lm)
outliers_y <- stu_res[abs(stu_res) > 2.5]
outliers_y
```

```
## named numeric(0)
```

```r
summary(stepwise_log_lm)
```

```
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ perc_population_with_high_school_degree +
##     gini_index, data = hate_nona_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34787 -0.39659 -0.00387  0.44892  1.06723
##
```

```
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            -18.947      4.254  -4.454 6.14e-05
## perc_population_with_high_school_degree  11.554      3.069   3.765 0.000512
## gini_index                              16.486      4.795   3.438 0.001334
##
## (Intercept)                            ***
## perc_population_with_high_school_degree ***
## gini_index                             **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5608 on 42 degrees of freedom
## Multiple R-squared:  0.288,  Adjusted R-squared:  0.2541
## F-statistic: 8.496 on 2 and 42 DF,  p-value: 0.0007974
```

```
full_nodc = lm(log(hate_crimes_per_100k_splc) ~ perc_population_with_high_school_degree + gini_index, da
summary(full_nodc)
```

```
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ perc_population_with_high_school_degree +
##     gini_index, data = hate_nona_df[-9, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26341 -0.43818  0.03534  0.42669  1.10132
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            -14.611      5.359  -2.726  0.00938 **
## perc_population_with_high_school_degree   9.509      3.419   2.781  0.00814 **
## gini_index                              10.811      6.429   1.682  0.10025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.556 on 41 degrees of freedom
## Multiple R-squared:  0.1595, Adjusted R-squared:  0.1185
## F-statistic: 3.889 on 2 and 41 DF,  p-value: 0.02841
```

```
# Re-fit model without Gini index variable
hsdeg_lm_nodc = lm(log(hate_crimes_per_100k_splc) ~ perc_population_with_high_school_degree, data = hate
summary(hsdeg_lm_nodc)
```

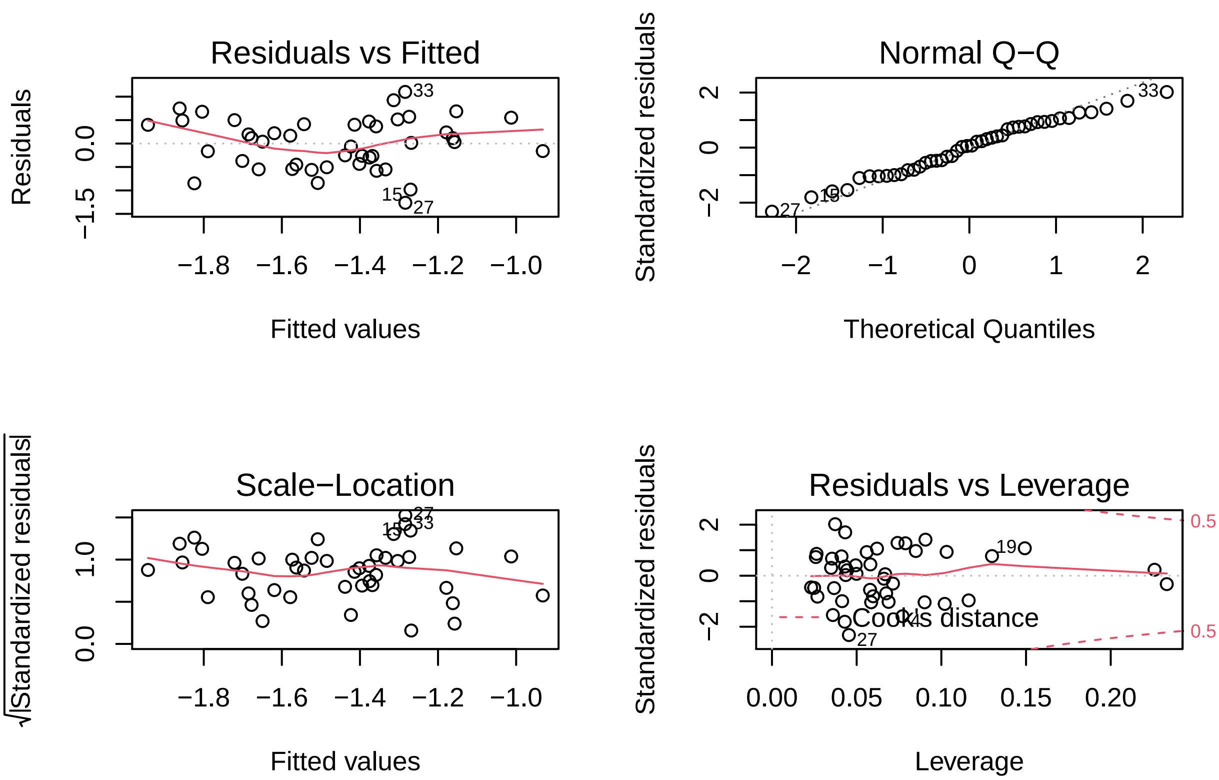
```
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ perc_population_with_high_school_degree,
##     data = hate_nona_df[-9, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12633 -0.48978  0.08266  0.44786  1.14080
##
```

```
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -6.413      2.274  -2.820   0.0073 **
## perc_population_with_high_school_degree    5.712      2.623   2.178   0.0351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.568 on 42 degrees of freedom
## Multiple R-squared:  0.1015, Adjusted R-squared:  0.08009
## F-statistic: 4.744 on 1 and 42 DF,  p-value: 0.03507
```

Using Cook's Distance and studentized residuals, DC could be an influential point. Cook's Distance is 0.491, close to the threshold of 0.5. The studentized residual is not greater than 2.5 for any variable. However, DFFIT is greater than 1 for DC which could be cause for concern. Comparing the regression analysis shows that this point is influential, so we need to consider deleting it.

Check model assumptions after removing DC as a point.

```r
par(mfrow = c(2, 2))
plot(full_nodc)
```



See if Gini index is significant by itself, with a log transformed outcome and DC point removed.

```r
gini_nodc_lm = lm(log(hate_crimes_per_100k_splc) ~ gini_index, data = hate_nona_df[-9,])
summary(gini_nodc_lm)
```

```
##
```

```
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ gini_index, data = hate_nona_df[-9,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.20509 -0.49941 -0.03609  0.41639  1.27579
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.0103     2.3622  -0.428    0.671
## gini_index   -0.9987     5.1997  -0.192    0.849
##
## Residual standard error: 0.5989 on 42 degrees of freedom
## Multiple R-squared:  0.0008775,  Adjusted R-squared:  -0.02291
## F-statistic: 0.03689 on 1 and 42 DF,  p-value: 0.8486
```

Gini index is not significant in this model.

Interactions

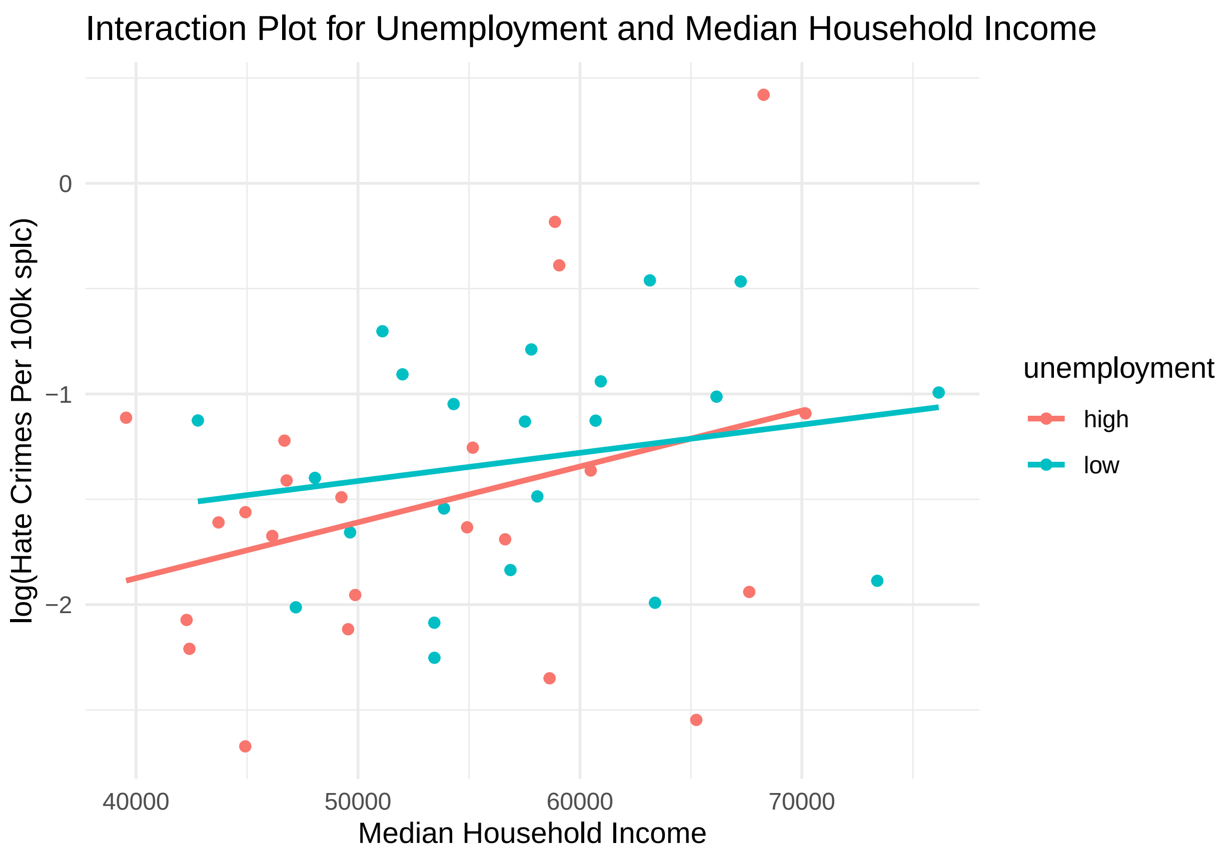
Check for interactions.

Fit a linear model with all two-way interactions and investigate which interactions could be significant.

```
full_int_lm = lm(log(hate_crimes_per_100k_splc)~(.-state)^2, data = hate_nona_df)
summary(full_int_lm)
```

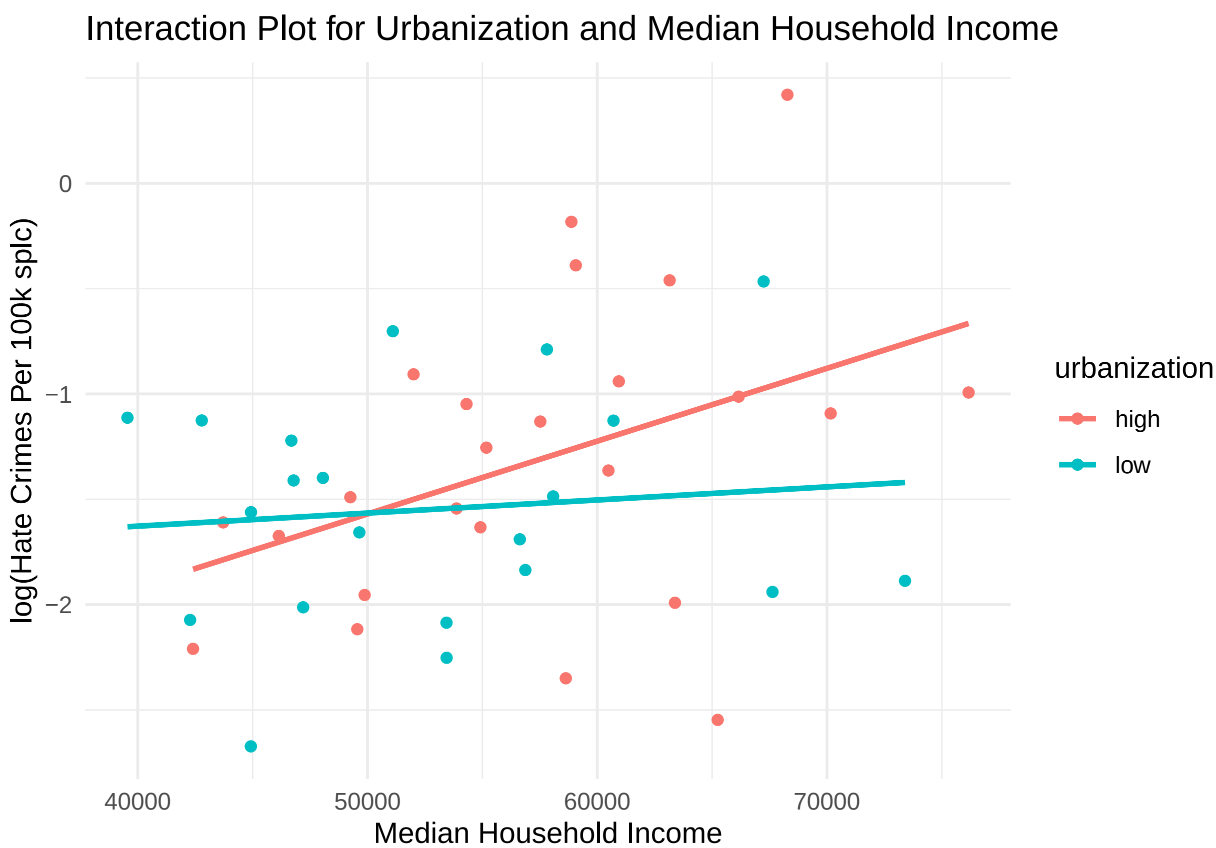
```
hate_nona_df %>%
  ggplot(aes(x = median_household_income, y = log(hate_crimes_per_100k_splc), color = unemployment)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Median Household Income",
       y = "log(Hate Crimes Per 100k splc)",
       title = "Interaction Plot for Unemployment and Median Household Income")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
hate_nona_df %>%
  ggplot(aes(x = median_household_income, y = log(hate_crimes_per_100k_splc), color = urbanization)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Median Household Income",
       y = "log(Hate Crimes Per 100k splc)",
       title = "Interaction Plot for Urbanization and Median Household Income")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



There were three significant interactions, two of which occur between a categorical and continuous variable, between median household income and unemployment, and median household income and urbanization. Looking at the interaction plots between these variables we do see that the lines for the two levels of the categorical variables cross, indicating interaction.

Perform a stratified analysis on these

```
# Stratified analysis for unemployment

unemployment_low = hate_nona_df %>%
  filter(unemployment == "low") %>%
  select(-unemployment)

unemployment_high = hate_nona_df %>%
  filter(unemployment == "high") %>%
  select(-unemployment)

unemployment_low_lm = lm(log(hate_crimes_per_100k_splc)~.-state, data = unemployment_low)
summary(unemployment_low_lm)
```

```
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ . - state, data = unemployment_low)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.81949 -0.33627  0.04041  0.26165  0.78586
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -1.499e+01  8.468e+00  -1.770    0.097
## urbanizationlow                      -3.689e-01  3.455e-01  -1.068    0.303
## median_household_income               1.328e-05  2.279e-05   0.582    0.569
## perc_population_with_high_school_degree  5.270e+00  7.077e+00   0.745    0.468
## perc_non_citizen                     -5.389e+00  1.040e+01  -0.518    0.612
## gini_index                            2.002e+01  1.026e+01   1.951    0.070
## perc_non_white                       -8.127e-01  1.862e+00  -0.436    0.669
##
## (Intercept)                          .
## urbanizationlow
## median_household_income
## perc_population_with_high_school_degree
## perc_non_citizen
## gini_index                           .
## perc_non_white
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5112 on 15 degrees of freedom
## Multiple R-squared:  0.3469, Adjusted R-squared:  0.08563
## F-statistic: 1.328 on 6 and 15 DF,  p-value: 0.3047
```

```r
unemployment_high_lm = lm(log(hate_crimes_per_100k_splc)~.-state, data = unemployment_high)
summary(unemployment_high_lm)
```

```
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ . - state, data = unemployment_high)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3070 -0.3050 -0.0516  0.3564  1.0279
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -2.722e+01  1.006e+01  -2.707   0.0156
## urbanizationlow                       2.858e-01  4.470e-01   0.639   0.5317
## median_household_income              -3.394e-05  3.327e-05  -1.020   0.3229
## perc_population_with_high_school_degree  2.065e+01  9.995e+00   2.066   0.0554
## perc_non_citizen                      8.021e+00  9.071e+00   0.884   0.3896
## gini_index                            2.044e+01  9.414e+00   2.171   0.0453
## perc_non_white                       -4.979e-01  1.638e+00  -0.304   0.7651
##
## (Intercept)                          *
## urbanizationlow
## median_household_income
## perc_population_with_high_school_degree .
## perc_non_citizen
## gini_index                           *
## perc_non_white
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6922 on 16 degrees of freedom
## Multiple R-squared:  0.3624, Adjusted R-squared:  0.1233
## F-statistic: 1.516 on 6 and 16 DF,  p-value: 0.2357
```

Median household income coefficients have different signs when unemployment is low vs high

```r
urbanization_low = hate_nona_df %>%
  filter(urbanization == "low") %>%
  select(-urbanization)

urbanization_high = hate_nona_df %>%
  filter(urbanization == "high") %>%
  select(-urbanization)

# Stratified analysis for urbanization

urbanization_low_lm = lm(log(hate_crimes_per_100k_splc)~.-state, data = urbanization_low)
summary(urbanization_low_lm)
```

```
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ . - state, data = urbanization_low)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94281 -0.32039 -0.02037  0.44346  0.90076
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            -1.748e+01  1.649e+01  -1.060    0.307
## unemploymentlow                         3.375e-03  4.615e-01   0.007    0.994
## median_household_income                -1.582e-05  3.439e-05  -0.460    0.652
## perc_population_with_high_school_degree  1.150e+01  1.214e+01   0.947    0.360
## perc_non_citizen                       -1.703e+00  1.608e+01  -0.106    0.917
## gini_index                              1.560e+01  1.961e+01   0.796    0.439
## perc_non_white                         -7.016e-01  1.908e+00  -0.368    0.719
##
## Residual standard error: 0.6062 on 14 degrees of freedom
## Multiple R-squared:  0.1546, Adjusted R-squared:  -0.2077
## F-statistic: 0.4266 on 6 and 14 DF,  p-value: 0.8492
```

```r
urbanization_high_lm = lm(log(hate_crimes_per_100k_splc)~.-state, data = urbanization_high)
summary(urbanization_high_lm)
```

```
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ . - state, data = urbanization_high)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.39221 -0.29276  0.04506  0.28314  1.11647
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             -2.371e+01  7.403e+00  -3.203  0.00522
## unemploymentlow                          3.247e-01  2.889e-01   1.124  0.27665
## median_household_income                 -9.553e-06  2.410e-05  -0.396  0.69672
## perc_population_with_high_school_degree  1.801e+01  8.066e+00   2.233  0.03929
## perc_non_citizen                         1.995e+00  6.777e+00   0.294  0.77203
## gini_index                               1.441e+01  6.744e+00   2.137  0.04740
## perc_non_white                           1.197e+00  1.897e+00   0.631  0.53637
##
## (Intercept)                             **
## unemploymentlow
## median_household_income
## perc_population_with_high_school_degree *
## perc_non_citizen
## gini_index                              *
## perc_non_white
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6168 on 17 degrees of freedom
## Multiple R-squared:  0.4537, Adjusted R-squared:  0.2609
## F-statistic: 2.353 on 6 and 17 DF,  p-value: 0.07723
```

```
## Median household income coefficients have different magnitudes when urbanization is low vs high
```

Stratifying on unemployment we see that the coefficient for median household income is positive when unemployment is low and negative when unemployment is high. When we stratify on urbanization, we see that the coefficient for median household income has a higher magnitude when urbanization is low (although both coefficients are negative). These stratified analyses indicate that interactions do exist.

Include these interaction terms in the full model to check for significance.

```r
interaction_log_lm = lm(
  log(hate_crimes_per_100k_splc)
  ~ perc_population_with_high_school_degree
  + gini_index
  + median_household_income
  + unemployment + urbanization
  + median_household_income*unemployment
  + median_household_income*urbanization,
  data = hate_nona_df)

summary(interaction_log_lm)
```

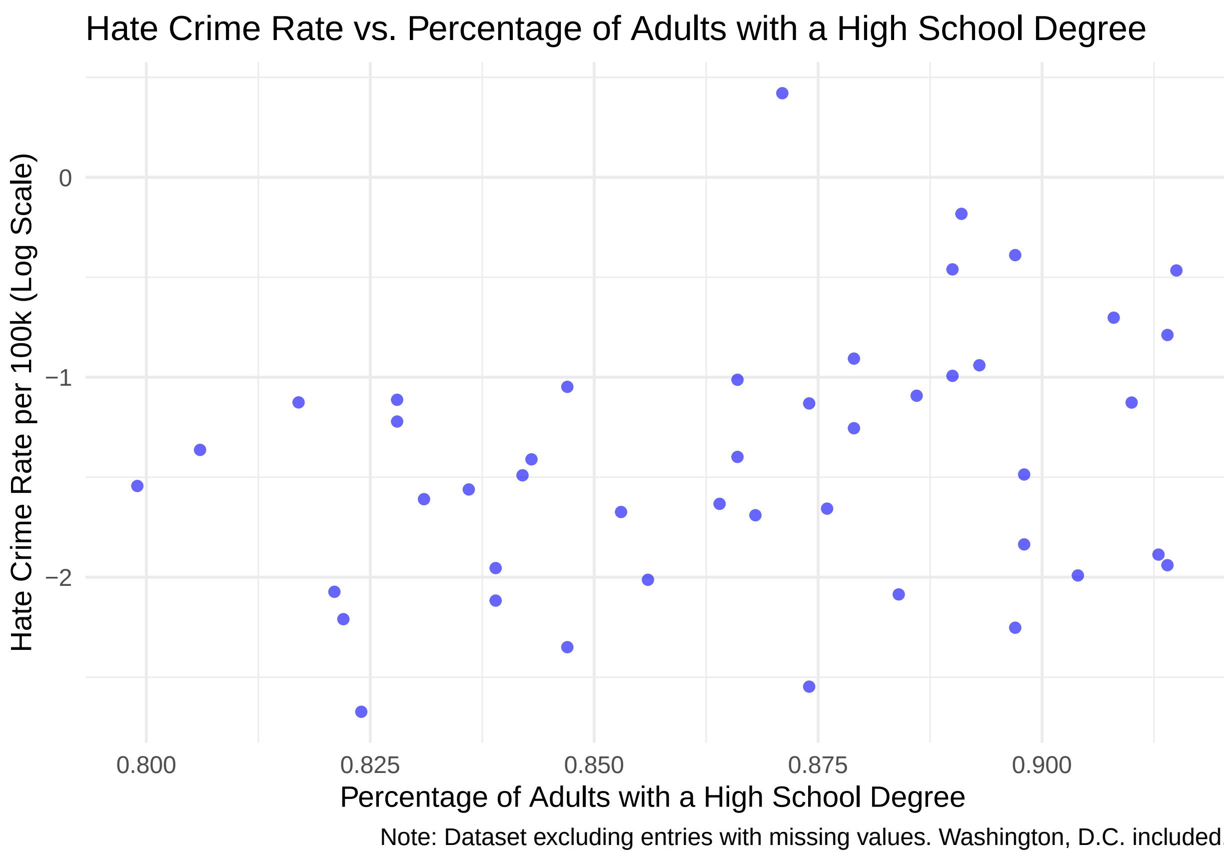
```
##
## Call:
## lm(formula = log(hate_crimes_per_100k_splc) ~ perc_population_with_high_school_degree +
##     gini_index + median_household_income + unemployment + urbanization +
##     median_household_income * unemployment + median_household_income *
```

```
##     urbanization, data = hate_nona_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31163 -0.36613  0.00301  0.33587  1.09203
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             -1.964e+01  5.275e+00  -3.722 0.000655
## perc_population_with_high_school_degree  1.246e+01  4.738e+00   2.631 0.012346
## gini_index                               1.589e+01  5.906e+00   2.691 0.010641
## median_household_income                  2.032e-06  1.962e-05   0.104 0.918057
## unemploymentlow                         -3.161e-01  1.189e+00  -0.266 0.791826
## urbanizationlow                          1.333e+00  1.171e+00   1.138 0.262301
## median_household_income:unemploymentlow  9.229e-06  2.112e-05   0.437 0.664710
## median_household_income:urbanizationlow -2.716e-05  2.156e-05  -1.260 0.215666
##
## (Intercept)                             ***
## perc_population_with_high_school_degree *
## gini_index                              *
## median_household_income
## unemploymentlow
## urbanizationlow
## median_household_income:unemploymentlow
## median_household_income:urbanizationlow
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5737 on 37 degrees of freedom
## Multiple R-squared:  0.3435, Adjusted R-squared:  0.2193
## F-statistic: 2.766 on 7 and 37 DF,  p-value: 0.02042
```

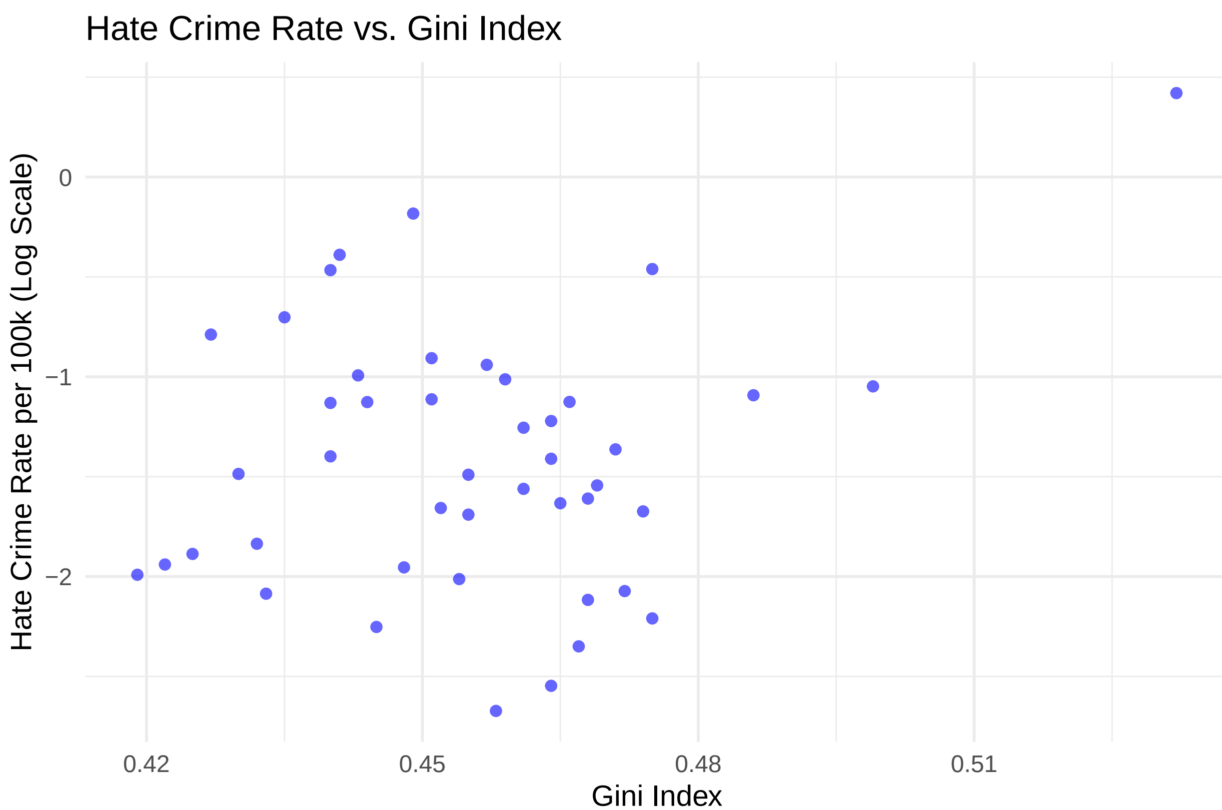
Holding Gini index and percent of population with high school degree constant, both interactions are not significant predictors of hate crime incidents. They are not included in the final model.

Scatter Plots for Predictors in Final Model

```
hate_nona_df %>%
  ggplot(aes(x = perc_population_with_high_school_degree, y = log(hate_crimes_per_100k_splc))) +
  geom_point(color = "blue", alpha = 0.6) +
  labs(
    title = "Hate Crime Rate vs. Percentage of Adults with a High School Degree",
    x = "Percentage of Adults with a High School Degree ",
    y = "Hate Crime Rate per 100k (Log Scale)",
    caption = "Note: Dataset excluding entries with missing values. Washington, D.C. included."
    )
```




```r
hate_nona_df %>%
  ggplot(aes(x = gini_index, y = log(hate_crimes_per_100k_splc))) +
  geom_point(color = "blue", alpha = 0.6) +
  labs(
    title = "Hate Crime Rate vs. Gini Index",
    x = "Gini Index",
    y = "Hate Crime Rate per 100k (Log Scale)",
    caption = "Note: Dataset excluding entries with missing values. Washington, D.C. included."
    )
```



Note: Dataset excluding entries with missing values. Washington, D.C. included.