Shepherd_M8_DSCI609

Emily Shepherd

2022-10-16

Preliminary Data Wrangling and Details

The data for income and crime rate was predominantly state-level data. I removed the entries that were for the entire United States. I also ensured that the years were listed in the same way for both data sets before combining into a single data frame. The unemployment data involved more cleaning and wrangling. The unemployment data was county level. Since you cannot just add percentages, I needed the total unemployed and total workforce. There were many missing entries for total unemployed workers. I used the unemployment rate and total workforce to determine the number unemployed. I omitted missing data and added by state. I then calculated the unemployment rate for each state. I noticed that there were several pieces of missing data in my unemployment data. There was no unemployment data for Delaware, New Jersey, or Connecticut, because those states did not provide the total workforce numbers. I also had to delete years. I converted state abbreviations to full names and reordered to match the combined income and crime data frame.

I noticed that all of the data for income and unemployment were numerical data. To complete ANOVA, I would need categorical data. I created levels for both the income and unemployment data. For the income data, I used 'low' for income less than $47338, 'medium' for $47338 to $62388, and 'high' for $62388 and above. For unemployment, I used two factors. From a bit of research, I found that under 5% is considered full employment, so under 5% was "ideal" and 5% or more was categorized as 'high.' This enabled me to complete ANOVA and neccessary post-hoc tests.

Part 1: One Way ANOVA: Crime vs Income

```
##              Df   Sum Sq Mean Sq F value   Pr(>F)
## level         2   667030  333515   15.91 1.93e-07 ***
## Residuals   538 11276193   20959
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = crime ~ level, data = all.three)
##
## $level
##                   diff        lwr       upr     p adj
## low-high     73.34670   29.36887 117.32452 0.0002945
## medium-high  -4.85542  -44.98928  35.27844 0.9564180
## medium-low  -78.20212 -111.86073 -44.54350 0.0000002
```

Conclusion: Because of the low p-value, we would reject the null hypothesis that there is no difference in crime rate based on income and therefore, conclude there is an association between average income and crime rate.

From the post-hoc test, we can conclude that there is not a significant difference in crime rate for middle and high income areas. However, the differences between crime rates in both low income and middle income areas and also, low income and high income areas are statistically significant.

Part 2: One Way ANOVA: Crime vs Unemployment

```
##              Df   Sum Sq Mean Sq F value Pr(>F)
## employ        1   108319  108319   4.933 0.0268 *
## Residuals   539 11834904   21957
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = crime ~ employ, data = all.three)
##
## $employ
##                  diff       lwr       upr     p adj
## ideal-high -29.71554 -55.99672 -3.434372 0.0267606
```

Conclusion: Because the p-value is less than 0.05, we conclude that there is a statistically significant association between crime and unemployment. The TukeyHSD confirmed this association.

Part 3: Two Way ANOVA of Crime Rate with Income and Unemployment

Two Way ANOVA with no interaction

```
##              Df   Sum Sq Mean Sq F value   Pr(>F)
## level         2   667030  333515  15.927 1.91e-07 ***
## employ        1    31305   31305   1.495    0.222
## Residuals   537 11244888   20940
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = crime ~ level + employ, data = all.three)
##
## $level
##                   diff        lwr       upr     p adj
## low-high     73.34670   29.38886 117.30453 0.0002925
## medium-high  -4.85542  -44.97103  35.26019 0.9563788
## medium-low  -78.20212 -111.84544 -44.55880 0.0000002
##
## $employ
##                  diff       lwr      upr     p adj
## ideal-high -15.55012 -41.21568 10.11544 0.2345033
```

Two Way ANOVA with interactions

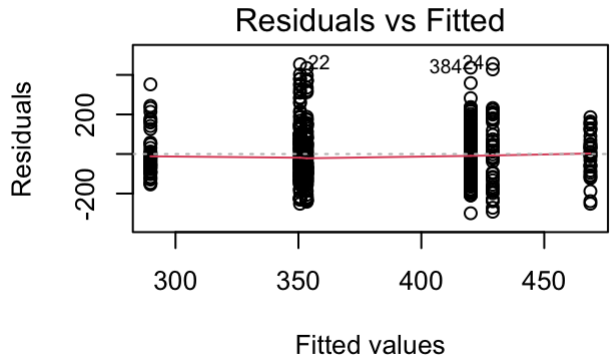
```
##               Df   Sum Sq Mean Sq F value   Pr(>F)
## level          2   667030  333515  16.597 1.02e-07 ***
## employ         1    31305   31305   1.558    0.213
## level:employ   2   494274  247137  12.299 6.00e-06 ***
## Residuals    535 10750613   20095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

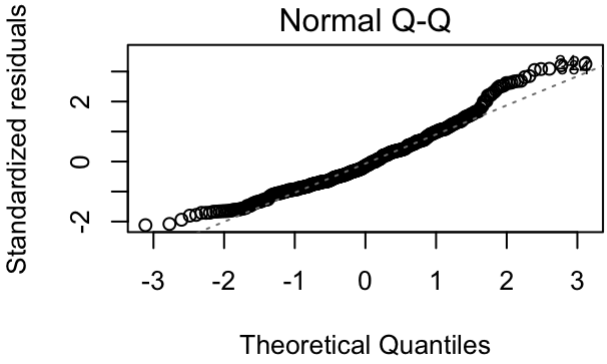
Determine the Best Model

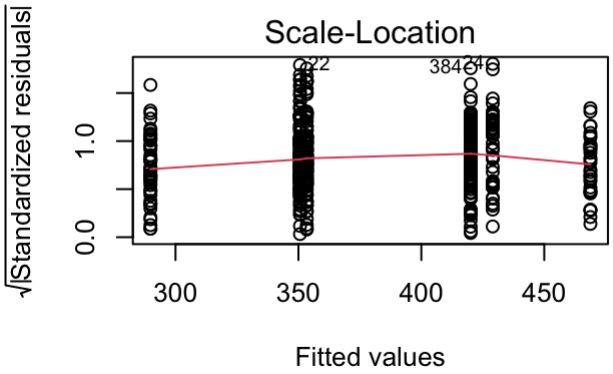
```
##
## Model selection based on AICc:
##
##              K    AICc Delta_AICc AICcWt Cum.Wt       LL
## interactions 7 6903.81       0.00      1      1 -3444.80
## one.way.inc  4 6923.49      19.69      0      1 -3457.71
## two.way      5 6924.03      20.22      0      1 -3456.96
## one.way.unem 3 6947.63      43.82      0      1 -3470.79
```

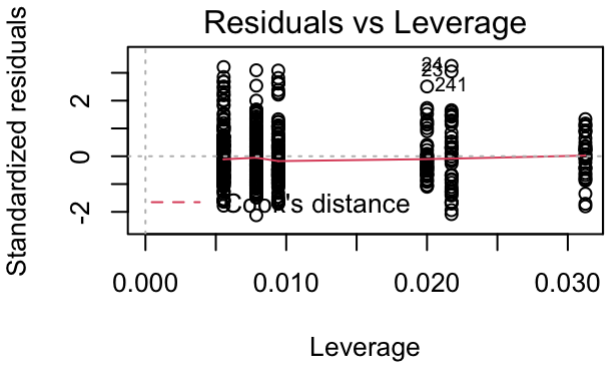
Conclusion: Based on the AICc, the interactions model would be the best fit, because it has the lowest value.

Plot the Best Model

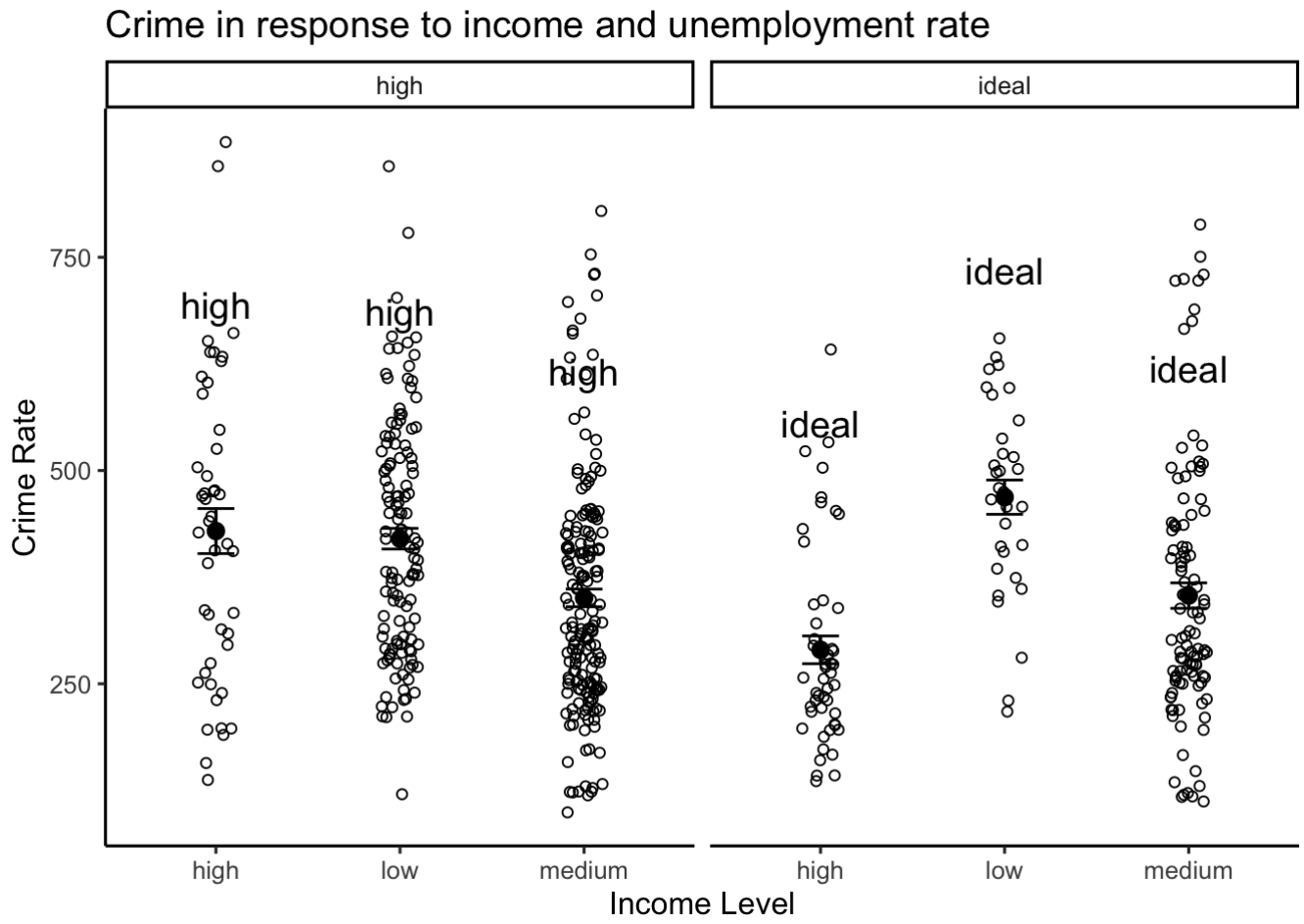








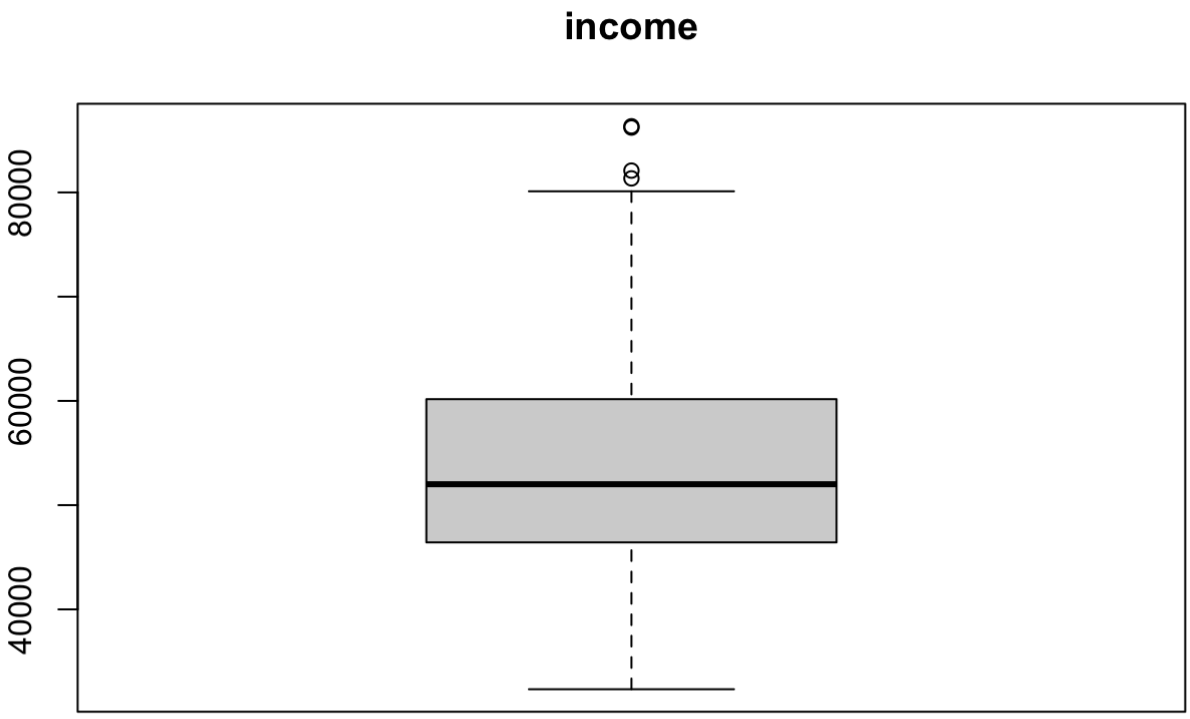
Plot the Raw Data



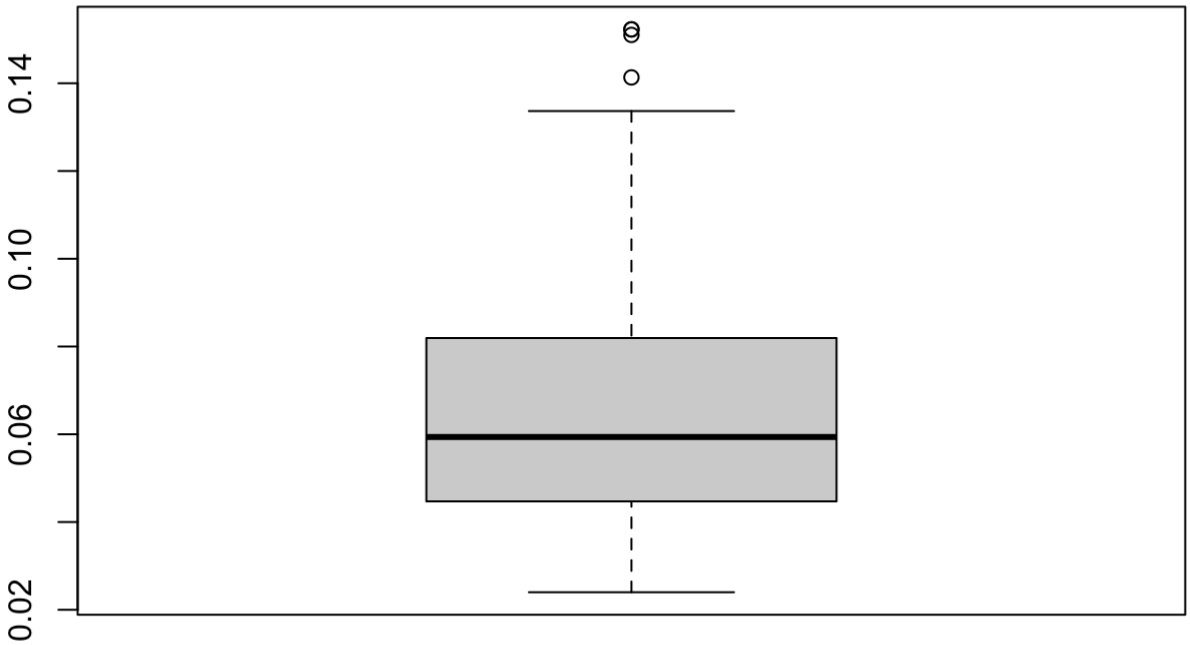
Conduct linear regression

For linear regression, I used the original quantitative variables for income and unemployment rate.

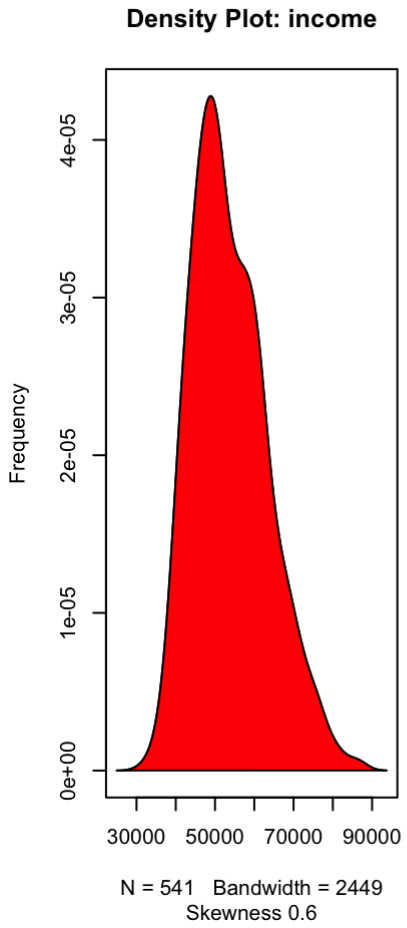
Check for outliers

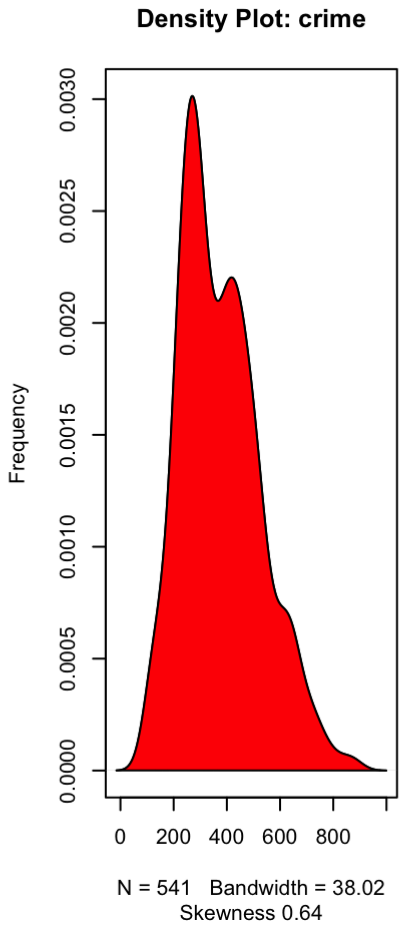

income
80000
60000
40000

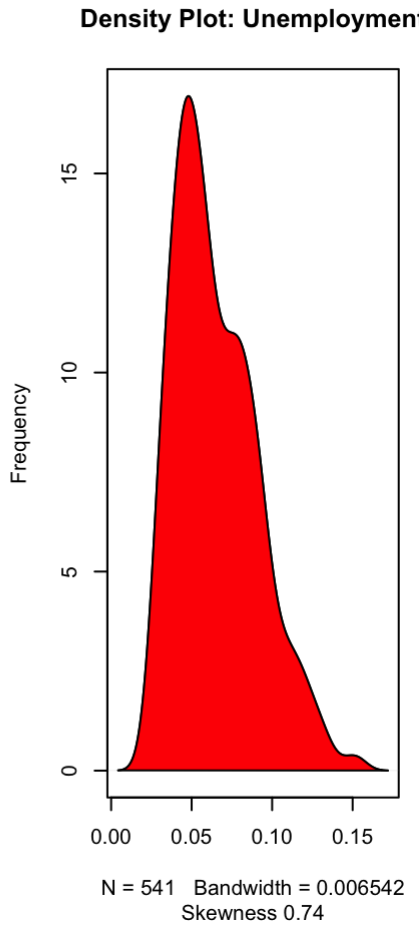

crime
800
600
400
200



Remove outliers and check for normality

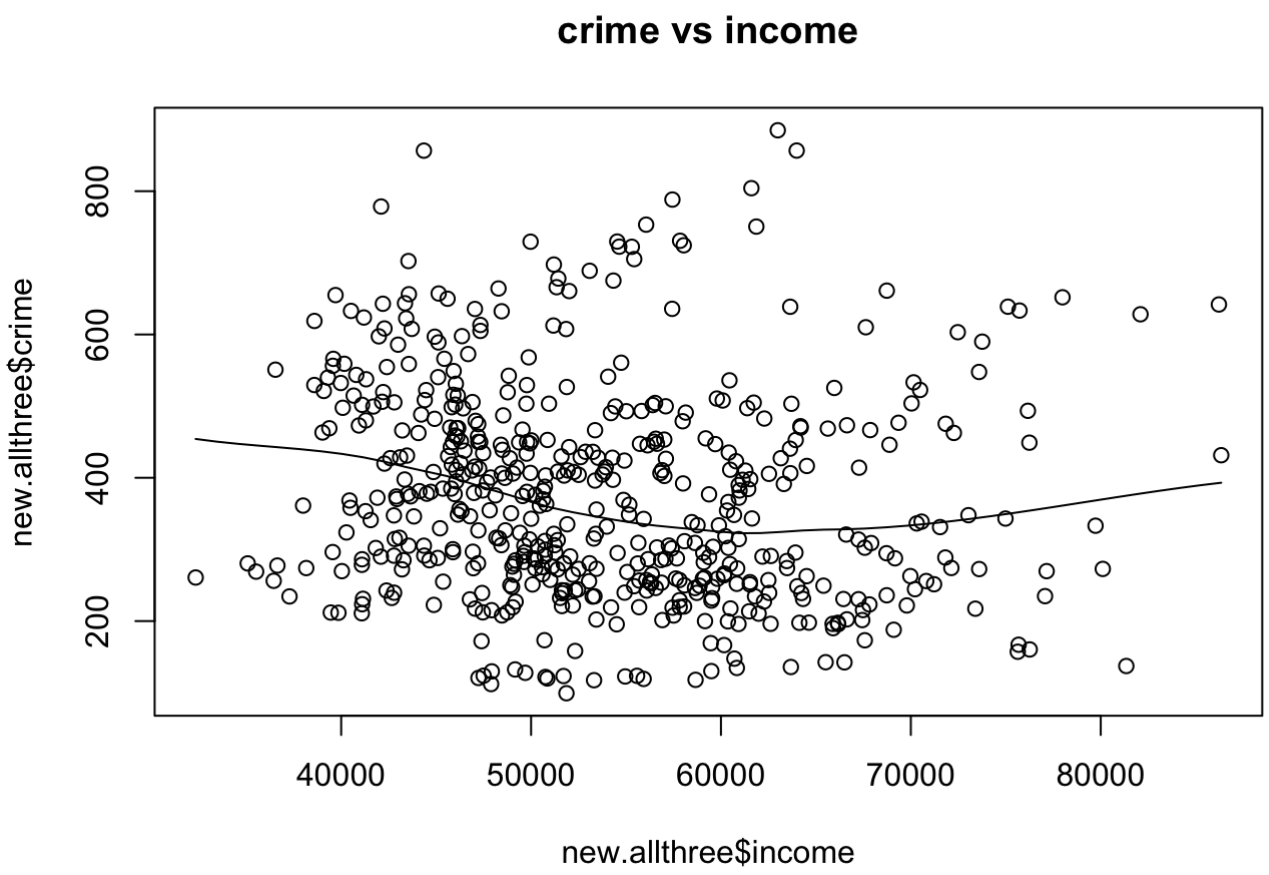






Correlation and Scatterplot for Crime vs Income

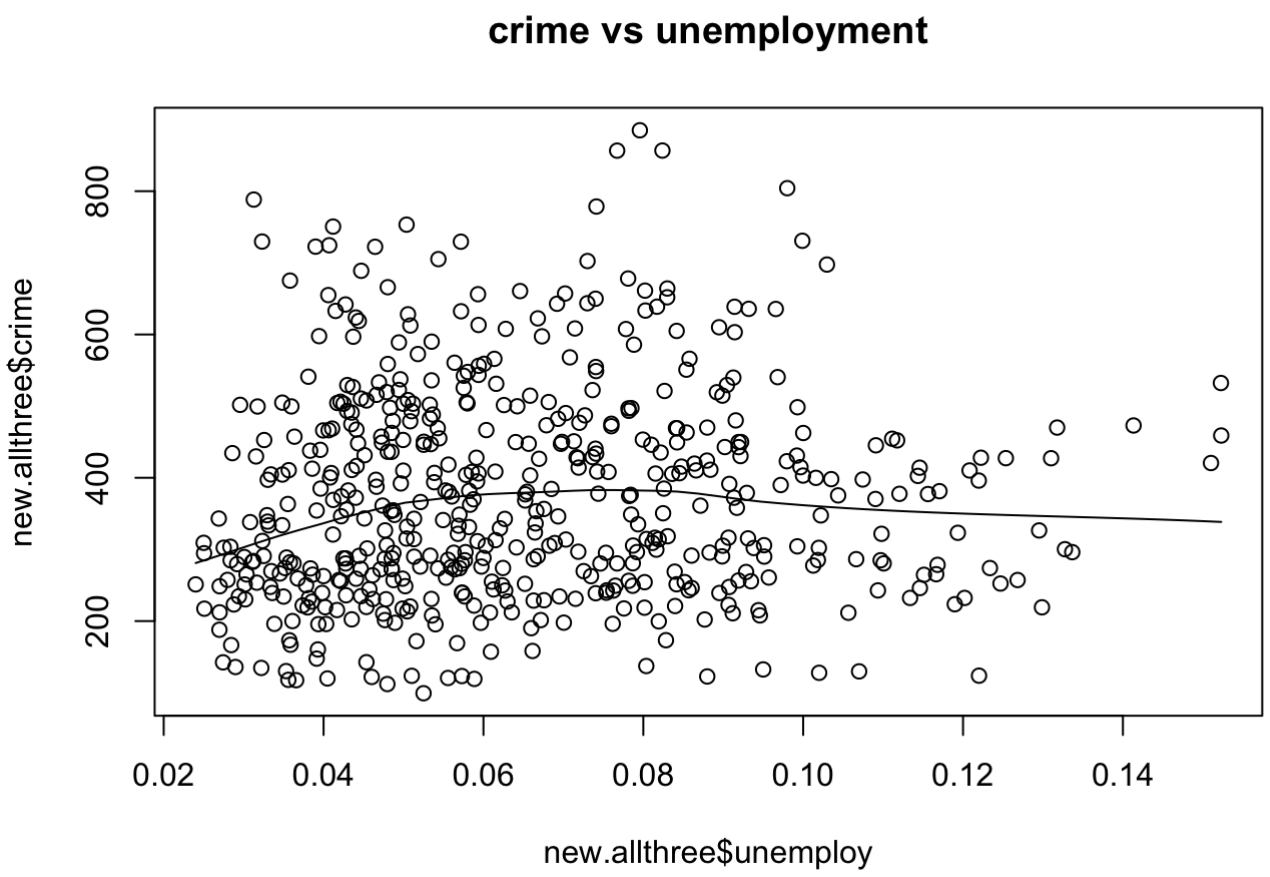
```
## [1] -0.1390983
```



Analysis: The scatterplot does not appear to have significant direction. With a correlation equal to -0.1391, there is a weak association between income and crime rate.

Correlation and Scatterplot for Crime vs Unemployment

```
## [1] 0.06747085
```



Analysis: The scatterplot does not appear to have significant direction. With a correlation equal to 0.0647, there is a weak association between unemployment rate and crime rate.

Linear Regression for Crime vs Income

```
##
## Call:
## lm(formula = crime ~ income, data = new.allthree)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -280.33 -110.31  -23.25   90.03  529.41
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.916e+02  3.616e+01  13.595  < 2e-16 ***
## income      -2.159e-03  6.622e-04  -3.261  0.00118 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 147.4 on 539 degrees of freedom
## Multiple R-squared:  0.01935,    Adjusted R-squared:  0.01753
## F-statistic: 10.63 on 1 and 539 DF,  p-value: 0.00118
```

```
## [1] 6941.942
```

```
## [1] 6954.822
```

AIC = 6941.942

BIC = 6954.822

p-value = 0.00118

Conclusion: At the 0.05 significance level, there is a statistically significant association between income and crime rate.

Linear Regression for Crime vs Unemployment

```
##
## Call:
## lm(formula = crime ~ unemploy, data = new.allthree)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -274.04 -111.92  -20.00   93.14  503.79
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   350.01      17.45   20.05   <2e-16 ***
## unemploy      392.06     249.72    1.57    0.117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 148.5 on 539 degrees of freedom
## Multiple R-squared:  0.004552,   Adjusted R-squared:  0.002705
## F-statistic: 2.465 on 1 and 539 DF,  p-value: 0.117
```

```
## [1] 6950.043
```

```
## [1] 6962.923
```

AIC = 6950.042

BIC = 6962.923

p-value = 0.117

Linear Models Conclusion: There is not a statistically significant association between unemployment rate and crime rate.

Multiple Linear Regression

```
##
## Call:
## lm(formula = crime ~ income + unemploy, data = new.allthree)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -277.82 -109.38  -23.09   93.38  525.56
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.727e+02  4.527e+01  10.442  < 2e-16 ***
## income      -2.025e-03  6.902e-04  -2.934  0.00349 **
## unemploy     1.796e+02  2.583e+02   0.695  0.48733
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 147.5 on 538 degrees of freedom
## Multiple R-squared:  0.02023,    Adjusted R-squared:  0.01659
## F-statistic: 5.554 on 2 and 538 DF,  p-value: 0.004098
```

Multiple Linear Regression Conclusion: Based on the p-value of 0.0041, at the 0.05 significance level, there is a statistically significant association among crime rate vs income and unemployment.

CONCLUSION: As stated in the above analyses, when using categorical variables to represent unemployment (ideal and high) and income (low, medium, high), the best model was the two-way ANOVA with interactions. When using the original quantitative data for income and unemployment rate, with outliers removed, the best model to predict crime rate would be the multiple linear regression model, because R-squared for multiple linear regression was slightly higher than the model of crime rate from income alone. There was not a statistically significant association between crime rate and unemployment rate alone.