descriptive_stats

Qetsiyah Wang

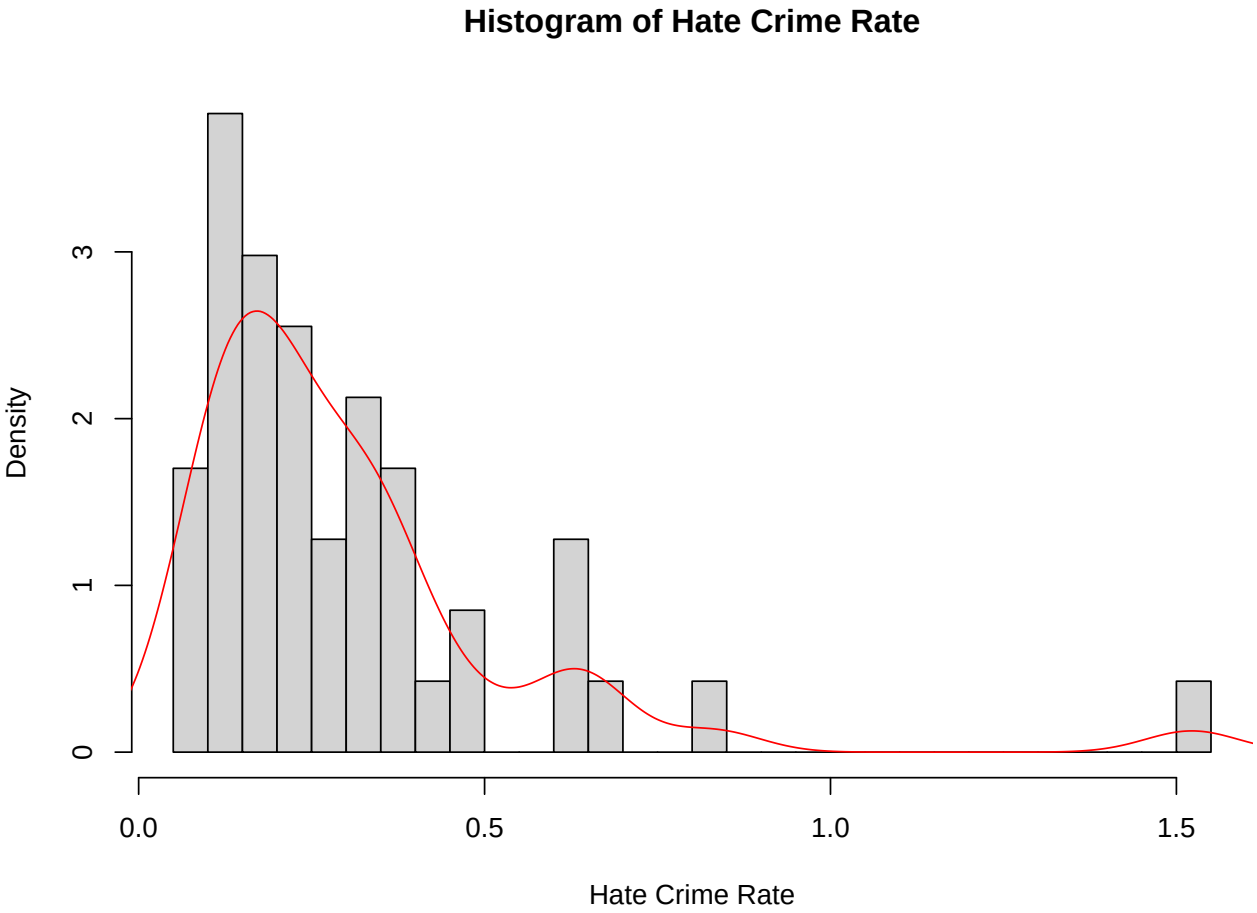
12/13/2020

Descriptive Statistics For Hate Crime Rate

Distribution of Hate Crime Rate

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.06745 0.14271 0.22620 0.30409 0.35694 1.52230       4
```

Descriptive Statistics about the Hate Crime Rate was shown in the table. The mean of the crime rate is 0.3041 and the median is 0.2262. Comparing two measures of location, mean is larger than the median, meaning that there would be a positive skewness within the distribution curve of hate crime rate. The third quartile is 0.3569, which shows the difference of 1.1654, meaning that outliers of crime rate exist. Range of the crime rate is 1.4549. Based on the general review on descriptive statistics of hate crime rate, we generate the histogram for more data visualization.

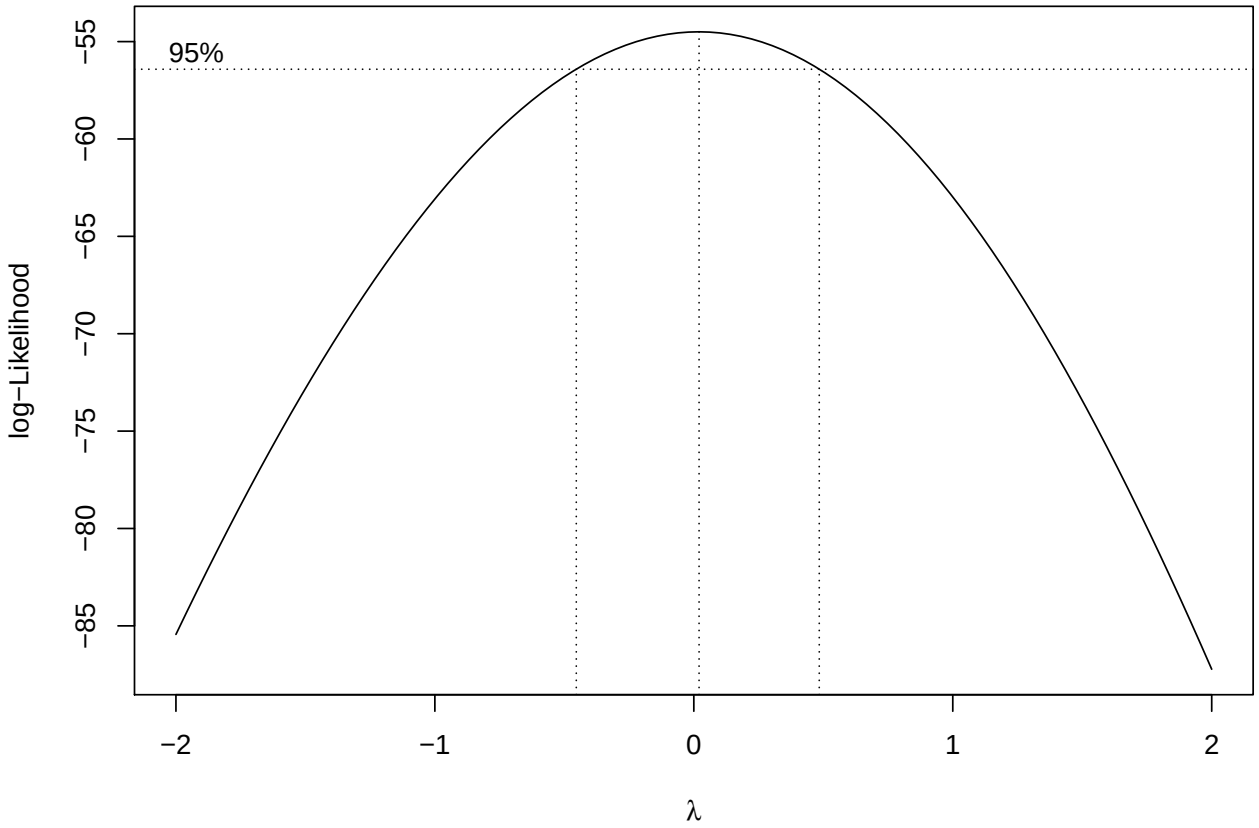


Same observations could be obtained from visualization. According to the Histogram of Hate Crime Rate,

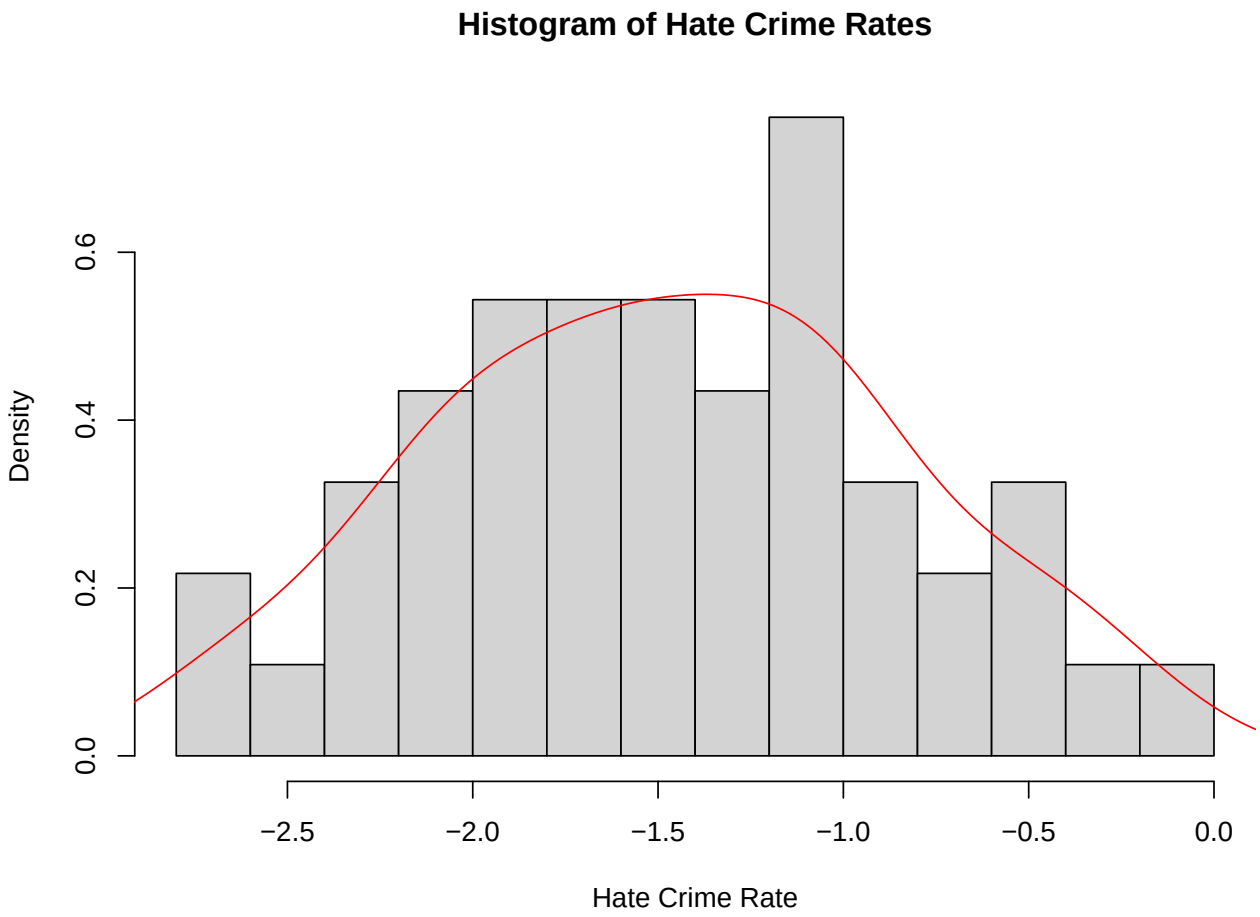
the distribution of hate crime rate is right skewed with a long right tail. The distribution curve is generally following the bell shape. Large proportion of data fell between 0 and 0.5 with 0.8039. However, the second slight peak could be observed from the crime rate range of 0.5-1.0. Furthermore, the maximum value of crime rate 1.5223 indeed acted as a outlier that strongly affected the distribution of hate crime rate. Generally, the distribution curve of Hate Crime Rate is the bell shape, however, the normality assumption could be regarded as being violated because there is long positive skewness observed from the distribution histogram. So, potential transformation of the outcome of interest (i.e,. Hate Crime Rate) needs to be performed for further establishment on data analysis, such as regression or prediction model.

```
## # A tibble: 5 x 2
##   state                hate_crimes_per_100k_splc
##   <chr>                                    <dbl>
## 1 District of Columbia                      1.52
## 2 Hawaii                                   NA
## 3 North Dakota                             NA
## 4 South Dakota                             NA
## 5 Wyoming                                  NA
```

For enhancing the efficacy of potential transformation on adjusting hate crime rate, outliers, presented by District of Columbia, could be considered to be excluded due to its negative impact on the normality assumption. Also, from descriptive statistic table, 4 states,Hawaii, North Dakota, South Dakota, Wyoming , showed missing values on hate crime rate, which did not need to be included inside our exploration.

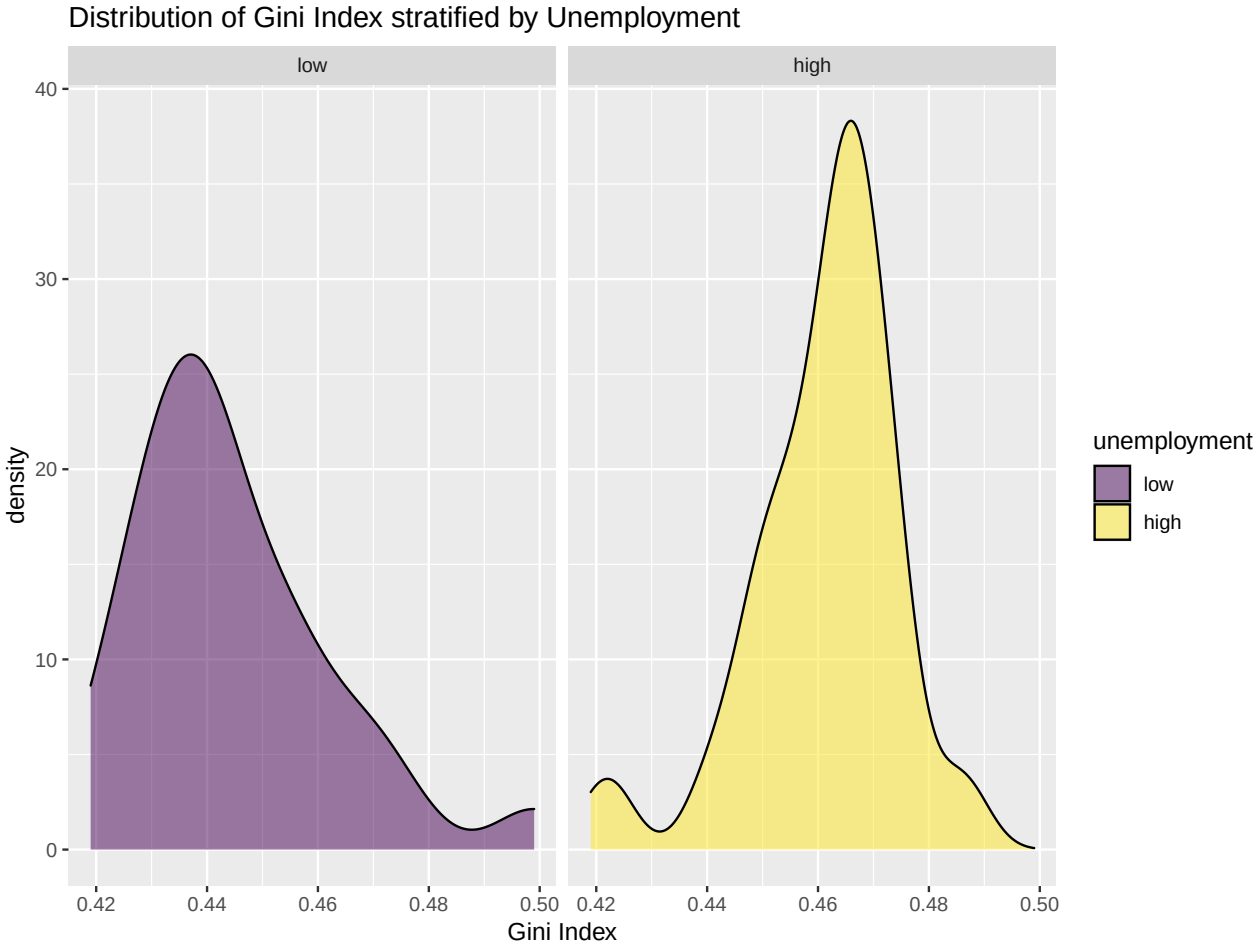


For finding out the effective transformation what we needed to perform for the outcome (i.e,. Hate Crime Rates), Box-Cox Transformation was used and shown in the figure above. Based on Box-Cox, the recommended transformation for the outcome was to take the natural logarithm as the lambda is equal to 0.

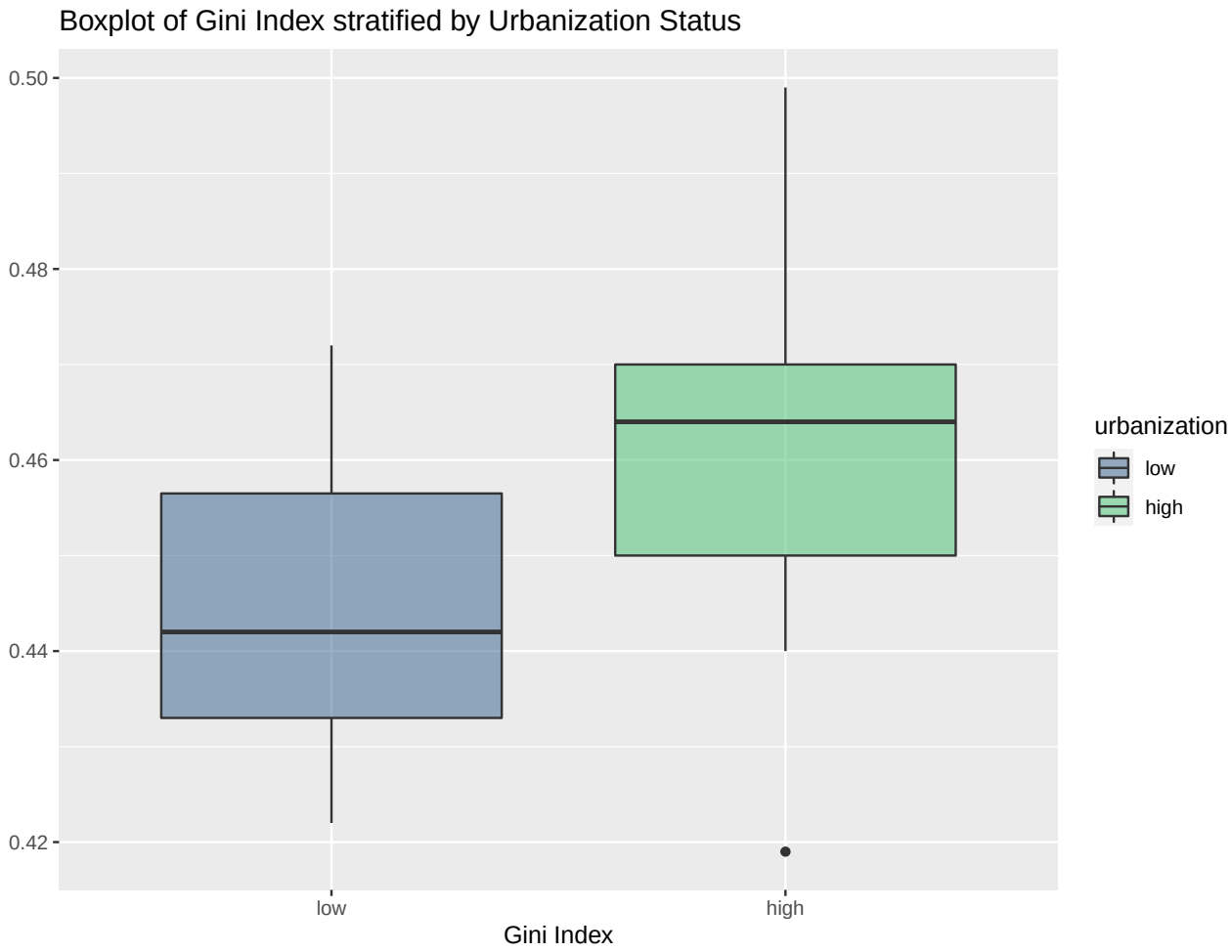


As taking the natural logarithm of hate crime rates, without showing any positive skewness, the distribution curve of hate crime rate generally presented as the bell shape and followed the normal distribution with a smooth peak. But, natural log histogram did not show any other peak besides the middle one, meaning that the potential transformation guaranteed the normality assumption for further model establishments.

Exploration with Gini Index - Income Inequality



Based on the article, income inequality is measured from the parameter of Gini Index, the distribution of the income across the population. The coefficient ranges from 0% to 100%, showing from perfect equality to perfect inequality, respectively. According to two distribution curves with the stratification of unemployment status for each state, high unemployment showed significantly higher Gini index between 46% to 48% than low unemployment status. States with low unemployment status presented both low Gini Index and low proportion at certain Gini Index than high unemployment status, meaning that the income inequality is more serious or significant within states with high unemployment status. In other words, states, showing high level of unemployment would have higher proportion of total income belonged to population with the higher income level.


Boxplot of Gini Index stratified by Urbanization Status
0.50
0.48
0.46
0.44
0.42
low
high
Gini Index
urbanization
low
high