Assignment 1

##Overview/Introduction: I chose the hate-crimes dataset. The corresponding article is "Higher Rates Of Hate Crimes Are Tied To Income Inequality" and the weblink is - https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/ ##The article discusses income inequality as the predictor of hate crimes and hate incidents in states, with state with more inequality having higher rates of hate incidents per capita. While the article doesn't conclusively refute the argument that hate crimes increased after the 2016 election (an implicit assumption that Trump voters felt emboldened and engaged in a reinvigorated hate crime spree post-election), it does conclude that the same factors were linked to hate crimes both during the pre-and-post election periods.

##Conclusions: I carried out some visualization to validate the following question "Did we see a higher percentage of hate crimes in states with a higher percentage of Trump voters or were there hgher hate crimes in states with lower median incomes?" However, the initial visualization did not yield clear results and an exploratory data analysis, complete with hypothesis and multiple regression analysis must be carried out to answer/validate the question.

Data Preparation - I loaded in the package fivethirtyeight and the corresponding dataset (hate_crimes) which is available on github. From the original dataset, I removed the share_non_citizen column as it is unrelated to my hypothesis and created a new subset called df. My predictor variable is median_house_inc and included in the df dataset. As part of data clean-up, I also changed the name of the column share_unemp_seas to share_unemp_seasadj as it was unclear whether the original column name referred to unemployment among seasonal workers or referred to the seasonally adjusted unemployment rate. In a similar vein, I changed the column name of avg_hatecrimes_per_100k_fbi to avg_annual_hatecrimes_per_100k_fbi as it refers to the average annual rate.

##Further data preparation steps included removing states with NA's which excluded 4 of the 51 records, I used the na.omit function on the entire subset (df) only because the hate_crimes_per_100k_splc is the only column with NA's. I also included a classification column (class1) that identifies and assigns a value of 1 for states where share of votes for Trump equals or exceeds 50% and a value of 0 for states where the share is less than 50%. I calculated the median income nationally to be $57,617 and similarly created a new column (class2) to identify states with at or above- average median incomes and below-average median incomes.

```r
#Include relevant packages
library(ggplot2)
library(fivethirtyeight)
```

```
## Some larger datasets need to be installed separately, like senators and
## house_district_forecast. To install these, we recommend you install the
## fivethirtyeightdata package by running:
## install.packages('fivethirtyeightdata', repos =
## 'https://fivethirtyeightdata.github.io/drat/', type = 'source')
```

```r
library(dplyr)
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```r
#Read in data set
data(hate_crimes)
#Exclude the share_non_citizen column
df <- subset(hate_crimes, select = -c(share_non_citizen))
#Rename columns
df <- rename(df, share_unemp_seasadj = share_unemp_seas, avg_annual_hatecrimes_per_100k_fbi = avg_hatec
#Check the number of rows and columns
row <- nrow(df); row
```

```
## [1] 51
```

```r
col <- ncol(df); col
```

```
## [1] 12
```

```r
#Omit rows with NA's in the hate_crimes_per_100k_splc column which excludes 4 states
naomit1 <- na.omit(df); naomit1
```

```
## # A tibble: 47 x 12
##    state state_abbrev median_house_inc share_unemp_sea~ share_pop_metro
##    <chr> <chr>                   <int>            <dbl>           <dbl>
##  1 Alab~ AL                      42278            0.06             0.64
##  2 Alas~ AK                      67629            0.064            0.63
##  3 Ariz~ AZ                      49254            0.063            0.9
##  4 Arka~ AR                      44922            0.052            0.69
##  5 Cali~ CA                      60487            0.059            0.97
##  6 Colo~ CO                      60940            0.04             0.8
##  7 Conn~ CT                      70161            0.052            0.94
##  8 Dela~ DE                      57522            0.049            0.9
##  9 Dist~ DC                      68277            0.067            1
## 10 Flor~ FL                      46140            0.052            0.96
## # ... with 37 more rows, and 7 more variables: share_pop_hs <dbl>,
## #   share_white_poverty <dbl>, gini_index <dbl>, share_non_white <dbl>,
## #   share_vote_trump <dbl>, hate_crimes_per_100k_splc <dbl>,
## #   avg_annual_hatecrimes_per_100k_fbi <dbl>
```

```r
naomit <- na.omit(df); naomit
```

```
## # A tibble: 47 x 12
##    state state_abbrev median_house_inc share_unemp_sea~ share_pop_metro
##    <chr> <chr>                   <int>            <dbl>           <dbl>
##  1 Alab~ AL                      42278            0.06             0.64
##  2 Alas~ AK                      67629            0.064            0.63
##  3 Ariz~ AZ                      49254            0.063            0.9
##  4 Arka~ AR                      44922            0.052            0.69
##  5 Cali~ CA                      60487            0.059            0.97
##  6 Colo~ CO                      60940            0.04             0.8
##  7 Conn~ CT                      70161            0.052            0.94
##  8 Dela~ DE                      57522            0.049            0.9
##  9 Dist~ DC                      68277            0.067            1
## 10 Flor~ FL                      46140            0.052            0.96
## # ... with 37 more rows, and 7 more variables: share_pop_hs <dbl>,
## #   share_white_poverty <dbl>, gini_index <dbl>, share_non_white <dbl>,
## #   share_vote_trump <dbl>, hate_crimes_per_100k_splc <dbl>,
## #   avg_annual_hatecrimes_per_100k_fbi <dbl>
```

```r
#Initialize variables for boxplots
vote <- naomit1$share_vote_trump
crime <- naomit1$hate_crimes_per_100k_splc
income <- naomit1$median_house_inc
state <- naomit1$state
inequality <- naomit1$gini_index
#Initialize the two new classification columns and assign values based on formula
naomit$class1 <- 0
naomit$class2 <- 0
naomit$class1[vote >= 0.5] <- 1; naomit
```

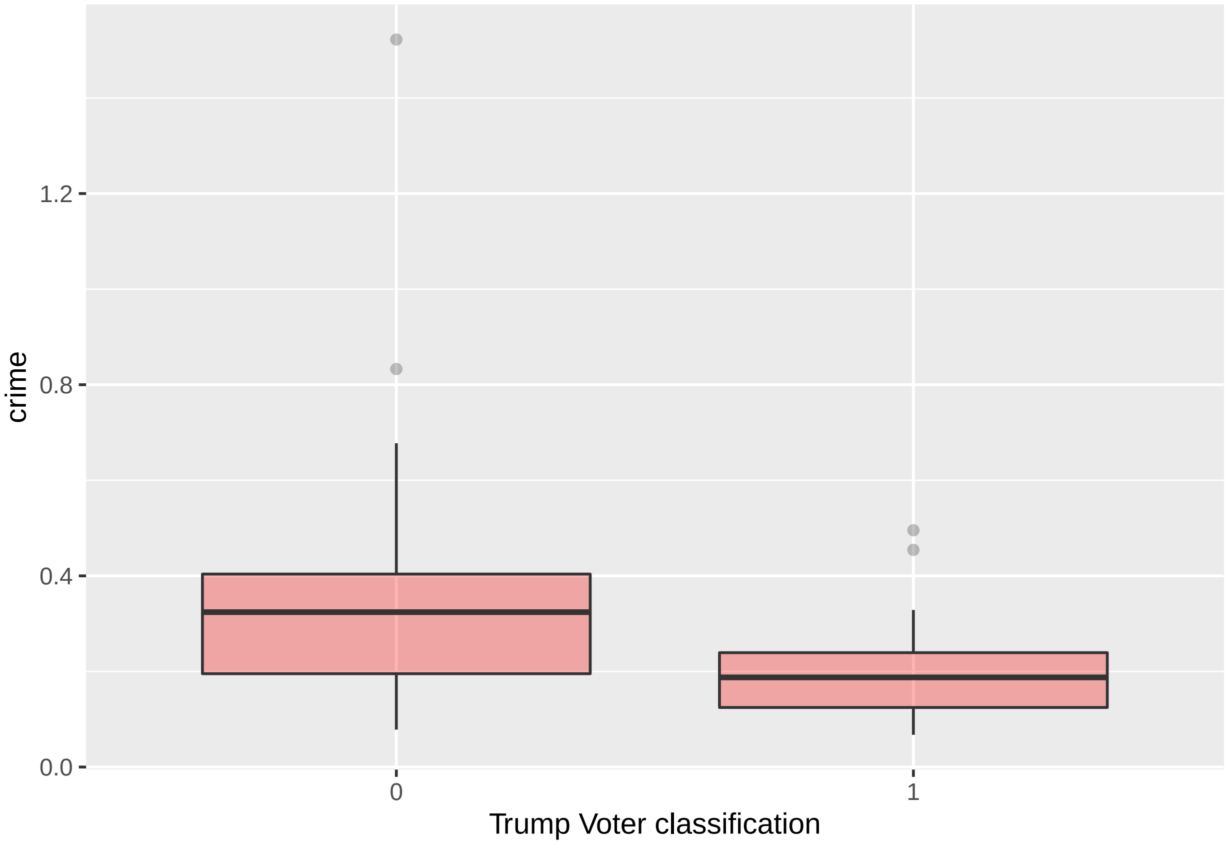
```
## # A tibble: 47 x 14
##    state state_abbrev median_house_inc share_unemp_sea~ share_pop_metro
##    <chr> <chr>                   <int>            <dbl>           <dbl>
##  1 Alab~ AL                      42278            0.06             0.64
##  2 Alas~ AK                      67629            0.064            0.63
##  3 Ariz~ AZ                      49254            0.063            0.9
##  4 Arka~ AR                      44922            0.052            0.69
##  5 Cali~ CA                      60487            0.059            0.97
##  6 Colo~ CO                      60940            0.04             0.8
##  7 Conn~ CT                      70161            0.052            0.94
##  8 Dela~ DE                      57522            0.049            0.9
##  9 Dist~ DC                      68277            0.067            1
## 10 Flor~ FL                      46140            0.052            0.96
## # ... with 37 more rows, and 9 more variables: share_pop_hs <dbl>,
## #   share_white_poverty <dbl>, gini_index <dbl>, share_non_white <dbl>,
## #   share_vote_trump <dbl>, hate_crimes_per_100k_splc <dbl>,
## #   avg_annual_hatecrimes_per_100k_fbi <dbl>, class1 <dbl>, class2 <dbl>
```

```r
naomit$class2[income >= 57617] <- 1; naomit
```

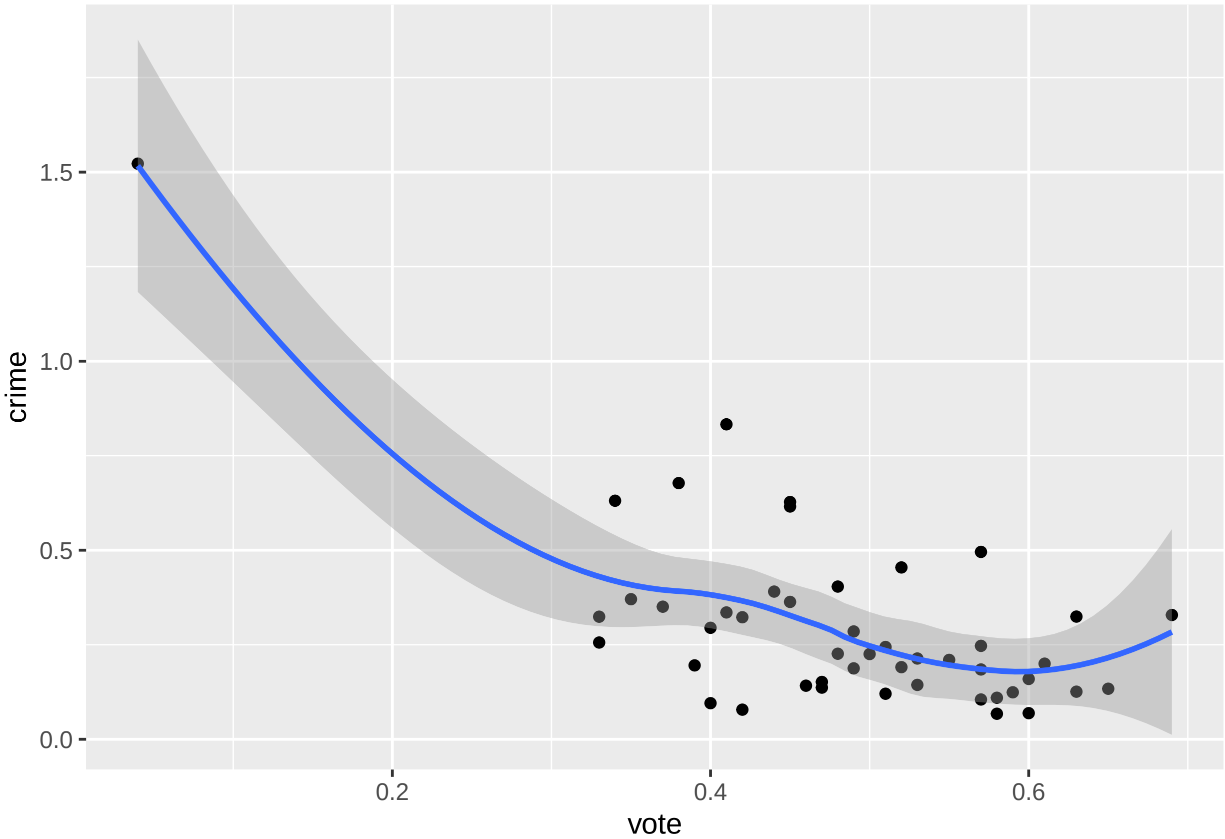
```
## # A tibble: 47 x 14
##    state state_abbrev median_house_inc share_unemp_sea~ share_pop_metro
##    <chr> <chr>                   <int>            <dbl>           <dbl>
##  1 Alab~ AL                      42278            0.06             0.64
##  2 Alas~ AK                      67629            0.064            0.63
##  3 Ariz~ AZ                      49254            0.063            0.9
##  4 Arka~ AR                      44922            0.052            0.69
##  5 Cali~ CA                      60487            0.059            0.97
##  6 Colo~ CO                      60940            0.04             0.8
##  7 Conn~ CT                      70161            0.052            0.94
##  8 Dela~ DE                      57522            0.049            0.9
##  9 Dist~ DC                      68277            0.067            1
## 10 Flor~ FL                      46140            0.052            0.96
## # ... with 37 more rows, and 9 more variables: share_pop_hs <dbl>,
## #   share_white_poverty <dbl>, gini_index <dbl>, share_non_white <dbl>,
## #   share_vote_trump <dbl>, hate_crimes_per_100k_splc <dbl>,
## #   avg_annual_hatecrimes_per_100k_fbi <dbl>, class1 <dbl>, class2 <dbl>
```

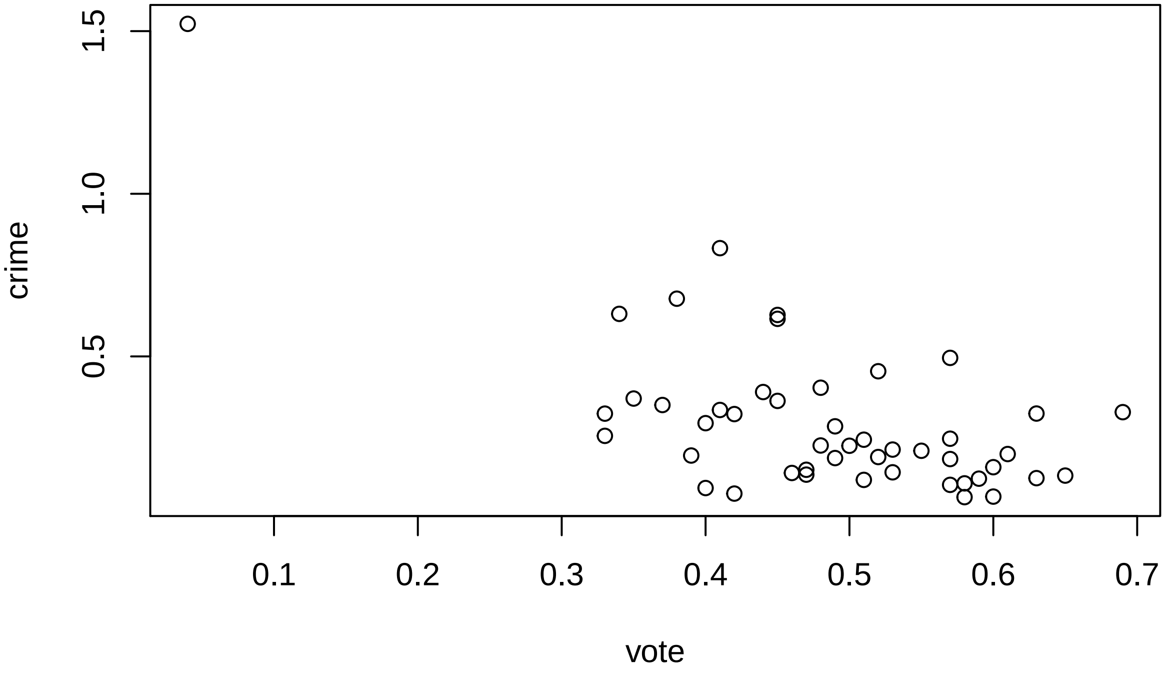
Visualization

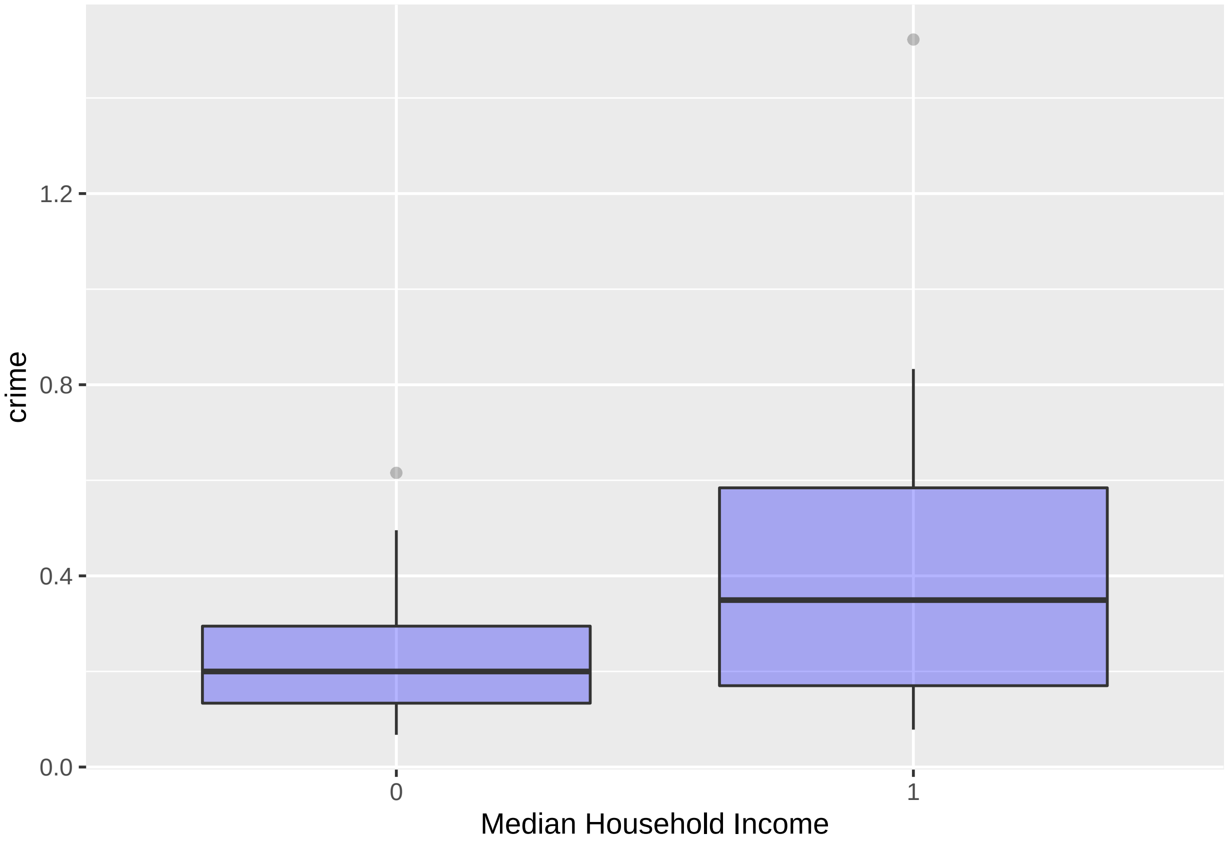
#After the data prep, I then plotted a boxplot of crime versus Trump voter classification and the results were interesting, the crime rate distribution was higher in states that did not vote for Trump and lower in states that saw a higher percentage of the vote share for Trump. Since, the Trump vote variable is numerical and not categorical, perhaps a correlation is a better metric to assess the relationship between crime and vote rather than simply relying onn the visual box plot. With this intent, I calculated correlation rates, the calculated correlation between crime and vote is -0.65 indicating a negative relationship. Hence, states with a higher percentage of Trump voters saw low crime which is visually depicted in the box plot. I carried out a similar analysis using median incomes and crime rates, I used the median US income in 2016 from the American Community Survey of $57,617 for the box plot classification of crime against income. Again, this variable didn't seem to intuitively explain the share of hate crimes. Washington DC seemed to be an outlier and removing this data point didn't increase correlation or intuition of the model.



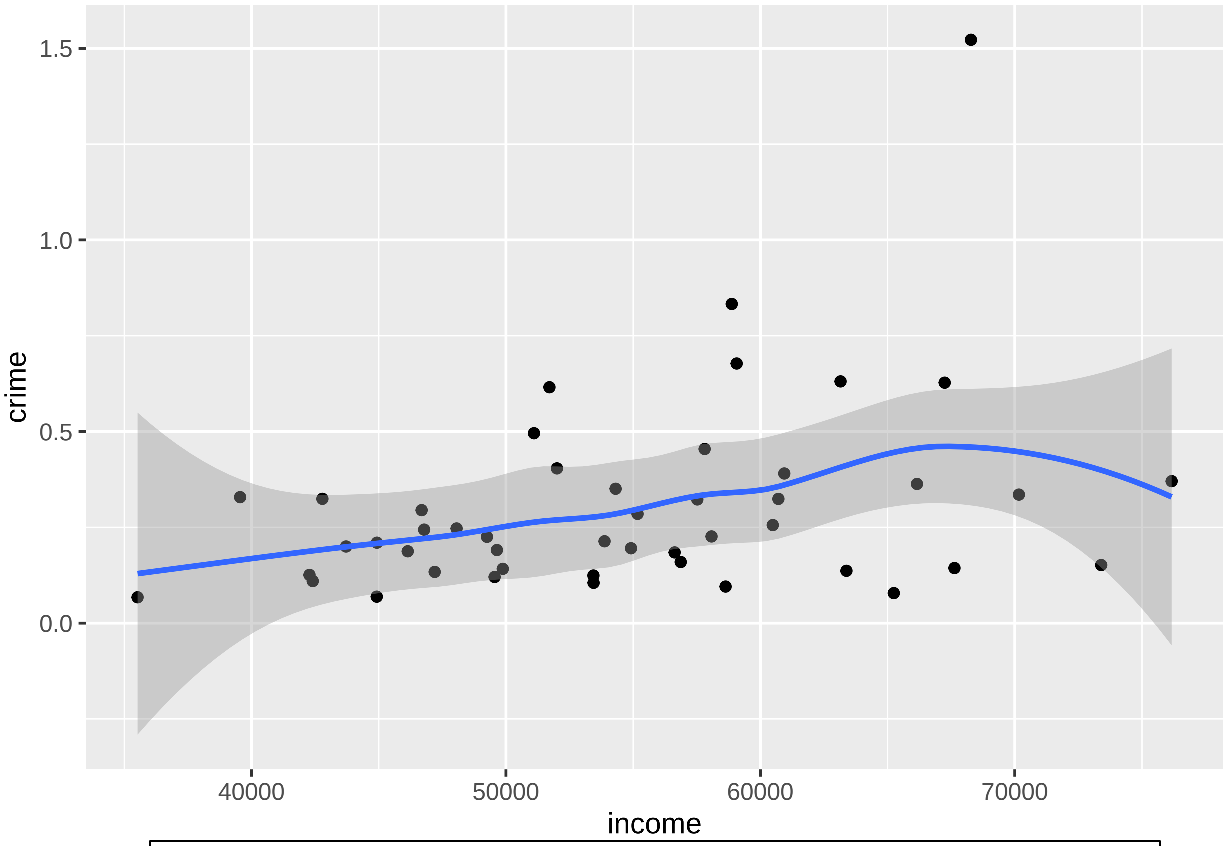
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

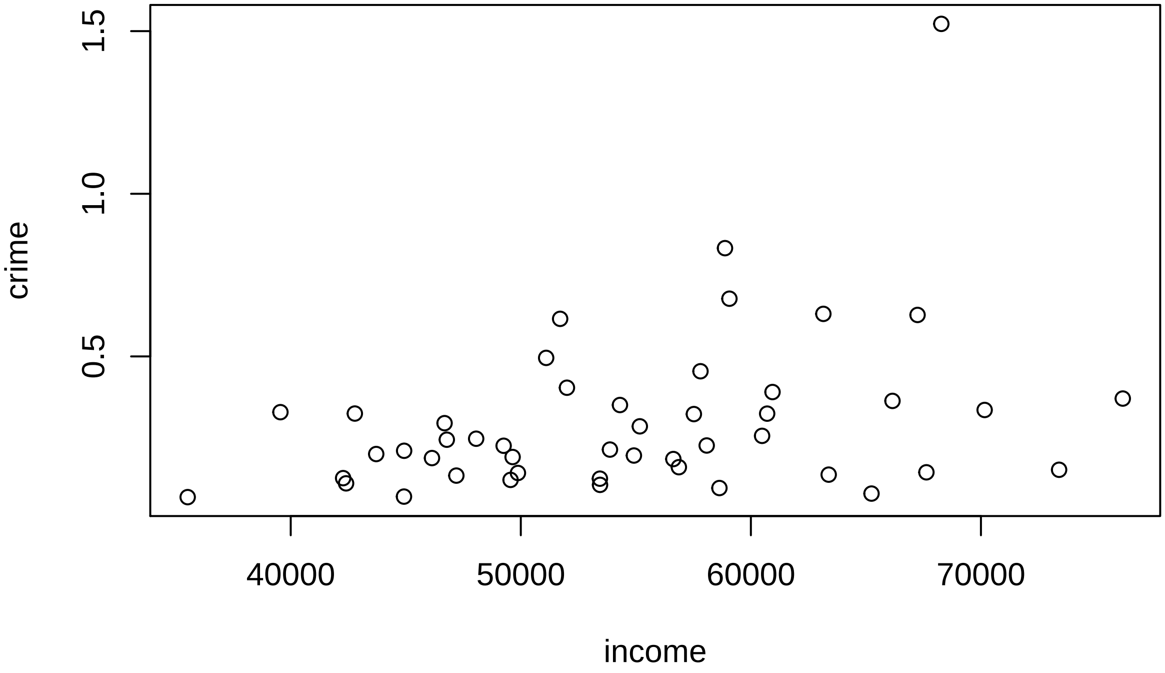

crime
1.5
1.0
0.5
0.0
0.2
0.4
0.6
vote


crime
1.5
1.0
0.5
0.1
0.2
0.3
0.4
0.5
0.6
0.7
vote



```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```





```
## [1] -0.6570672
```

```
## [1] 0.3507143
```