HW3

Evan Cranmer and Xiaoqing Liu

2024-09-20

1a Summary of "Crim" in Boston dataset and a Histogram of Crime Rates

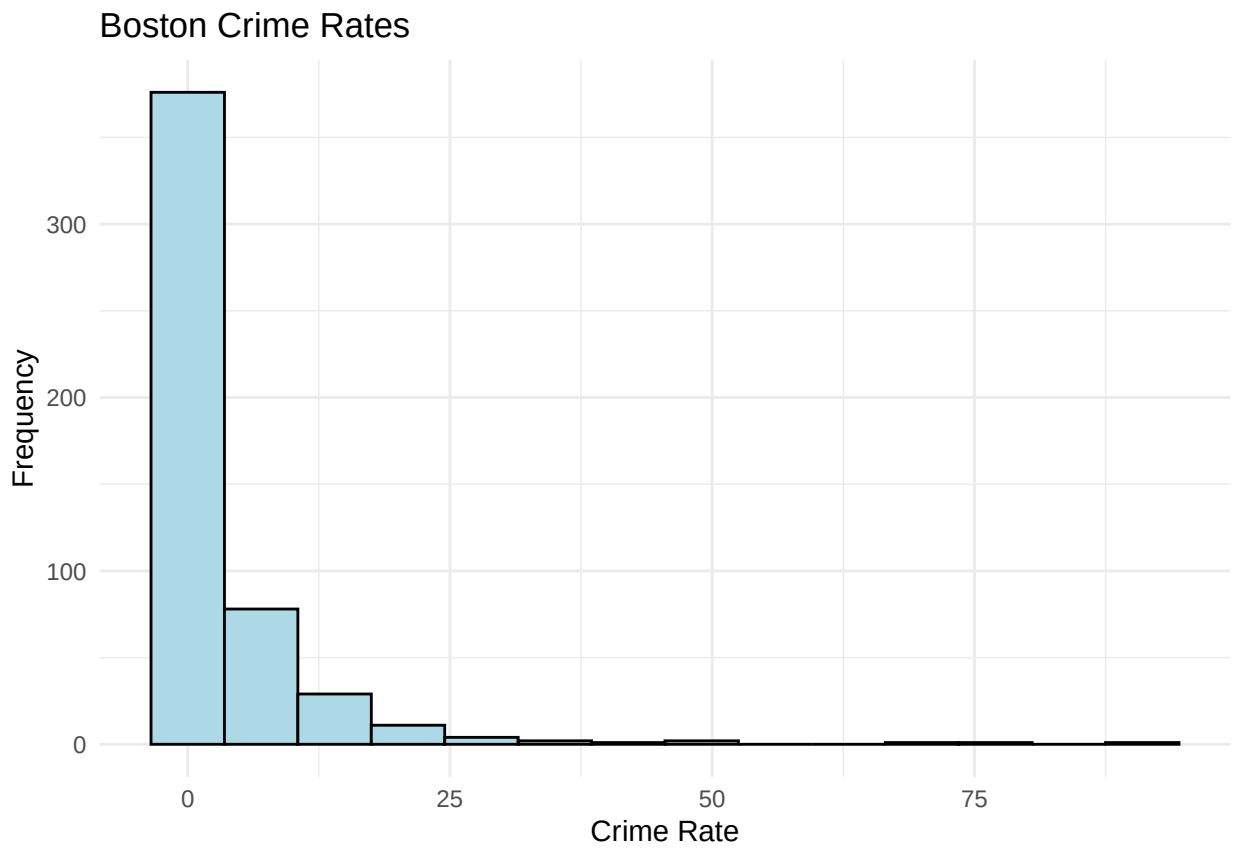
```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
data("Boston")

#getting familiar with the Boston dataset
?Boston

summarycrim <- summary(Boston["crim"])
summarycrim
```

```
##       crim
##  Min.   : 0.00632
##  1st Qu.: 0.08205
##  Median : 0.25651
##  Mean   : 3.61352
##  3rd Qu.: 3.67708
##  Max.   :88.97620
```

```
#subsetting crim data as a dataframe
crim_df <- data.frame(crim = Boston$crim)

# Plotting crim as a histogram
ggplot(crim_df, aes(x = crim)) +
  geom_histogram(binwidth = 7, fill = "lightblue", color = "black") +
  labs(title = "Boston Crime Rates", x = "Crime Rate", y = "Frequency") +
  theme_minimal()
```



After calculating summary statistics, we see that the median value of crime rat for Boston suburbs is 0.25, and the average is 3.61. After displaying the crime rates graphically, we can confirm most of our suburb crime rates are in the 0-12.5 range.

1b: Suburbs and Crime Rate

```
library(dplyr)
#subsetting suburbs where crim > 25
suburbs25 <- Boston %>%
  filter(crim > 25) %>%
  dplyr::select(crim, ptratio, tax, medv)

#counting the amount of cases we have. There are 11 suburbs
count(suburbs25)
```

```
##    n
## 1 11
```

```
#creating summary table for crime rate above 25
suburbs25summary <- suburbs25 %>%
  summarise(
    avg_ptratio = mean(ptratio),
    avg_tax = mean(tax),
    avg_medv = mean(medv)
```

```
  )

suburbs25summary
```

```
##   avg_ptratio avg_tax avg_medv
## 1        20.2     666 9.354545
```

```
#comparing the rest of the suburbs

suburbsrest <- Boston %>%
  filter(crim < 25) %>%
  dplyr::select(crim, ptratio, tax, medv)


suburbsrestsummary <- suburbsrest %>%
  summarise(
    avg_ptratio = mean(ptratio),
    avg_tax = mean(tax),
    avg_medv = mean(medv)
  )

suburbsrestsummary
```

```
##   avg_ptratio  avg_tax avg_medv
## 1    18.41677 402.5091 22.82566
```

Discussion of 1b

After subsetting the data to show suburbs with a crime rate higher than 25, we found that there are a total of 11 suburbs in the dataset. After collecting the values of pupil-teacher ratios, property tax rates, and median home values for these 11 suburbs, we found that the parent teacher ratio is the same for all 11 suburbs, at a value of 20.2. They also have a full-property tax rate of 666 per 10,000 dollars. For this reason, 20.2 and 666 are the average values for parent teacher ratios and property tax rates, respectively. The average median home value per is 9.35 in $1000s.

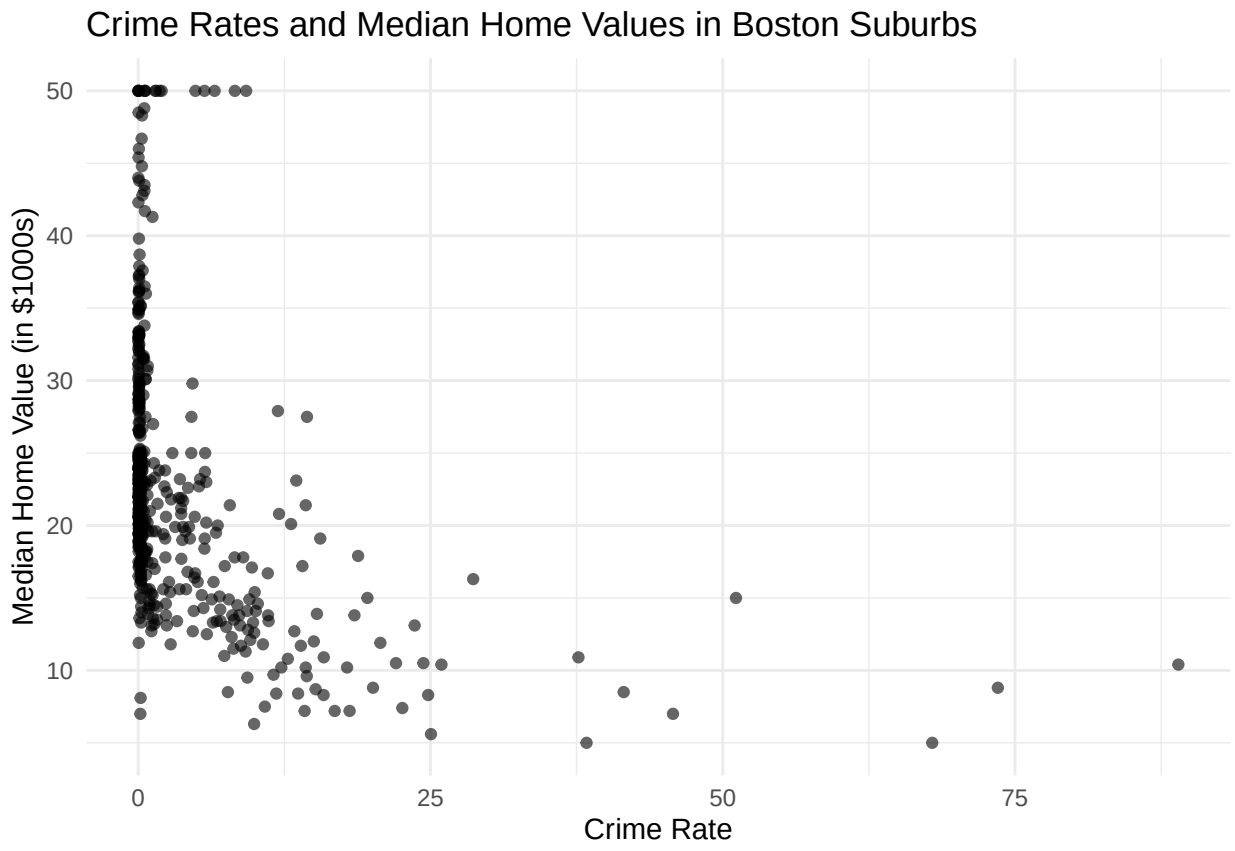
Comparing with all other suburbs

After subsetting the data for all other suburbs with a crime rate less than 25, and collecting the same summary statistics, we find different averages for each column. The average pupil-teacher ratio is 18.41, which is lower than that of the first dataset (20.2). The average full-property tax rate of 402.51 is less than that of high crime areas (666). However. the median value of homes is higher, with a value of 22.83, which is more than double that of high-crime suburbs (9.35).

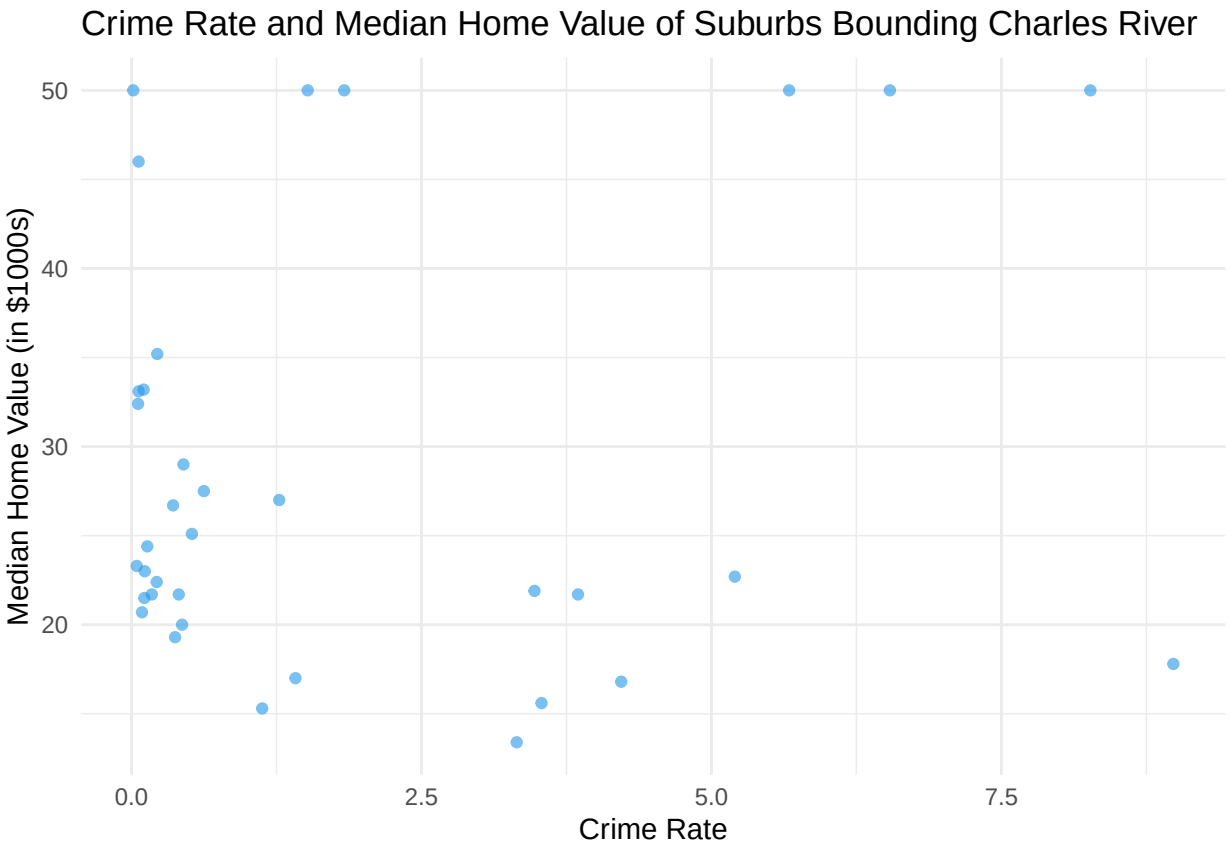
1c Creating a scatter plot of Crime Rates and Median Home Values

```
# 2. Create the scatter of crime rate and median home valuesfor all suburbs
crimeplot <- ggplot(Boston, aes(x = crim, y = medv)) +
  geom_point(alpha = 0.6) +
  labs(
```

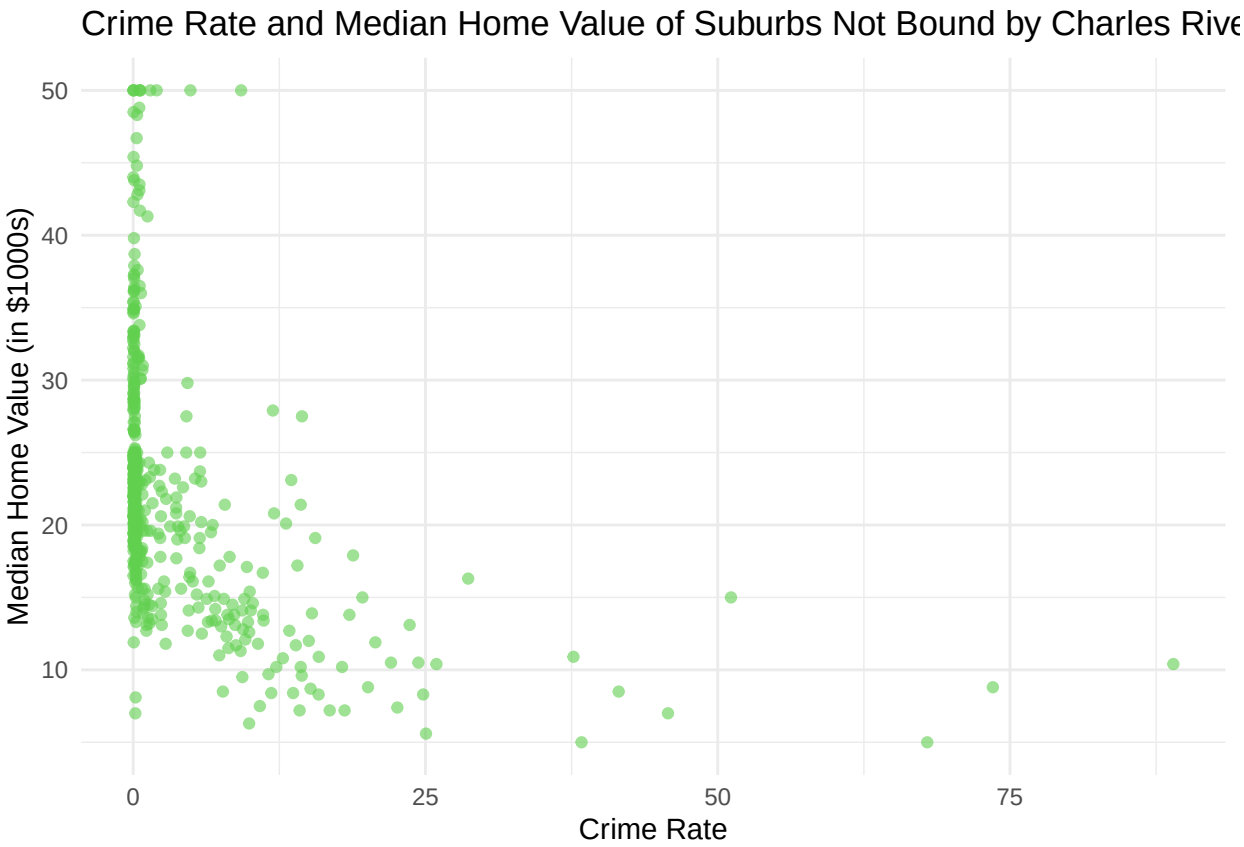
```
    title = "Crime Rates and Median Home Values in Boston Suburbs",
    x = "Crime Rate",
    y = "Median Home Value (in $1000s)",
  ) +
  theme_minimal()

crimeplot
```




```
#repeating for suburbs bounding Charles River

charlescrime <- Boston %>%
  filter(chas == 1) %>%
  ggplot(., aes(x = crim, y = medv)) +
  geom_point(alpha = 0.6, color = 4) +
  labs(
    title = "Crime Rate and Median Home Value of Suburbs Bounding Charles River",
    x = "Crime Rate",
    y = "Median Home Value (in $1000s)",
  ) +
  theme_minimal()

charlescrime
```




```
notcharlescrime <- Boston %>%
  filter(chas == 0) %>%
  ggplot(., aes(x = crim, y = medv)) +
  geom_point(alpha = 0.6, color = 3) +  # Plot points with some transparency
  labs(
    title = "Crime Rate and Median Home Value of Suburbs Not Bound by Charles River",
    x = "Crime Rate",
    y = "Median Home Value (in $1000s)",
  ) +
  theme_minimal()

notcharlescrime
```



1d: Analyzing the scatterplots

While the scatterplot showing suburbs bound by the Charles River looks more spread out, if we look at the x-axis we see that crime rate is actually at a different scale. No suburb bound by the Charles river has a crime rate higher than 10. If we contrast that to suburbs not bound by the Charles River, we can see that the majority of crime rates lie in that lower 0-5 range, but there are still many cases where crime rate is higher than 12.5. By taking a look at the median home values, this one is a bit harder to differentiate. There is not a major difference in median value of home prices because both have most cases in the 10-30 range, with fewer suburbs reaching 40-50.

1d Regression of crime rates as a function of median home values

```
crimemodel <- lm(crim ~ medv, data = Boston)

#summarizing our model
summary(crimemodel)
```

```
##
## Call:
## lm(formula = crim ~ medv, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -9.071 -4.022 -2.343  1.298 80.957
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.79654    0.93419   12.63   <2e-16 ***
## medv        -0.36316    0.03839   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.934 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

Interpreting Our Results from 1d

Based on the results of our model, we can interpret the intercept as the predicted crime rate of 11.80 when the median home value is 0 (which does not make intuitive sense). For each one unit increase in medv, which correspons to a $1000 increase in median value of homes, the crime rate decreases by about 0.36 per capita.

1E: Calculating our model by hand

After calculating the slope and intercept of our model we got the same results. Results are below.

```
#calculating slope
crimeslope <- sum((Boston$medv - mean(Boston$medv)) * (Boston$crim - mean(Boston$crim))) / sum((Boston$m

#verifying it matches -0.36... it does!
crimeslope
```

```
## [1] -0.3631599
```

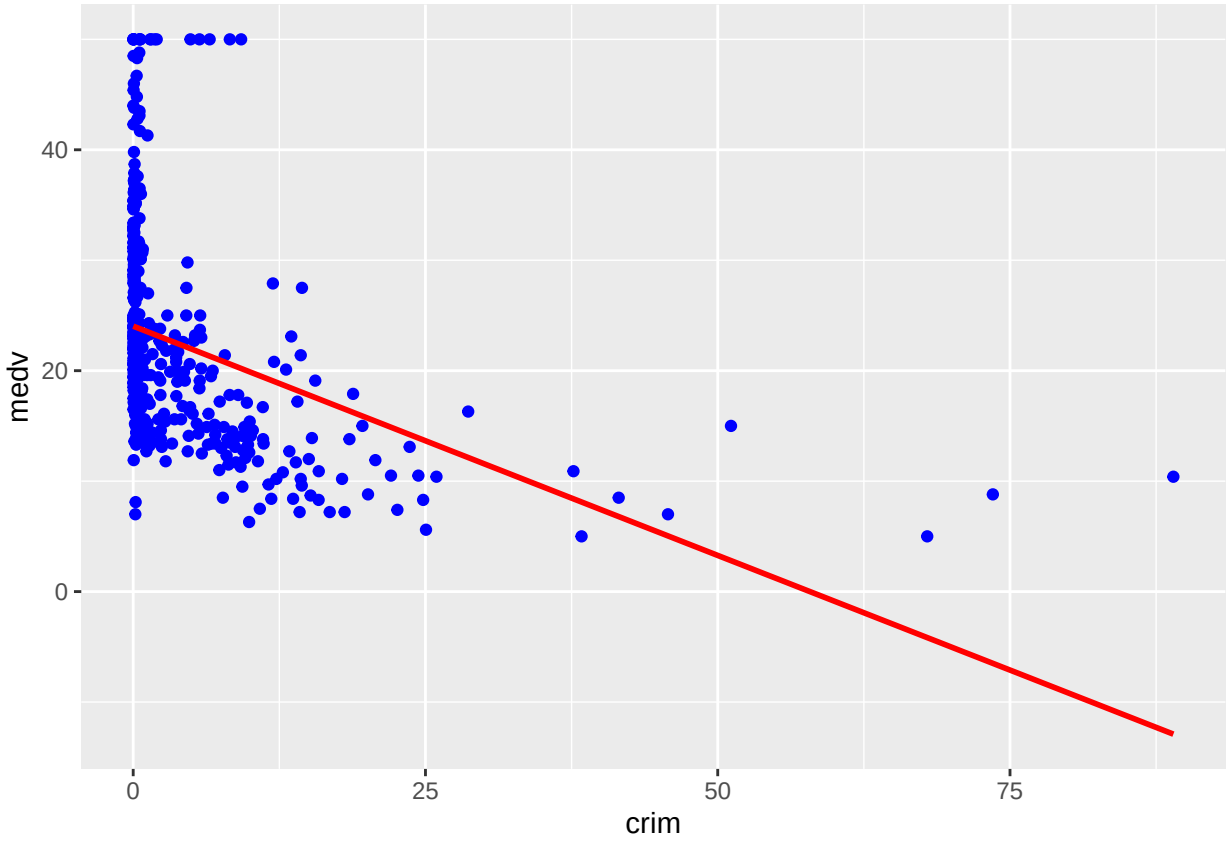
```
crimeintercept <- mean(Boston$crim) - crimeslope * mean(Boston$medv)

#verifying it matches 11.79654... it does!
crimeintercept
```

```
## [1] 11.79654
```

1F: Scatterplot with regression line and discussion of results

```
#creating scatter plot of crime rates and median home values with a regression line
ggplot(Boston, aes(y=medv, x=crim))+
  geom_point(color="blue")+
  geom_smooth(method='lm', color="red", se=FALSE)
```

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

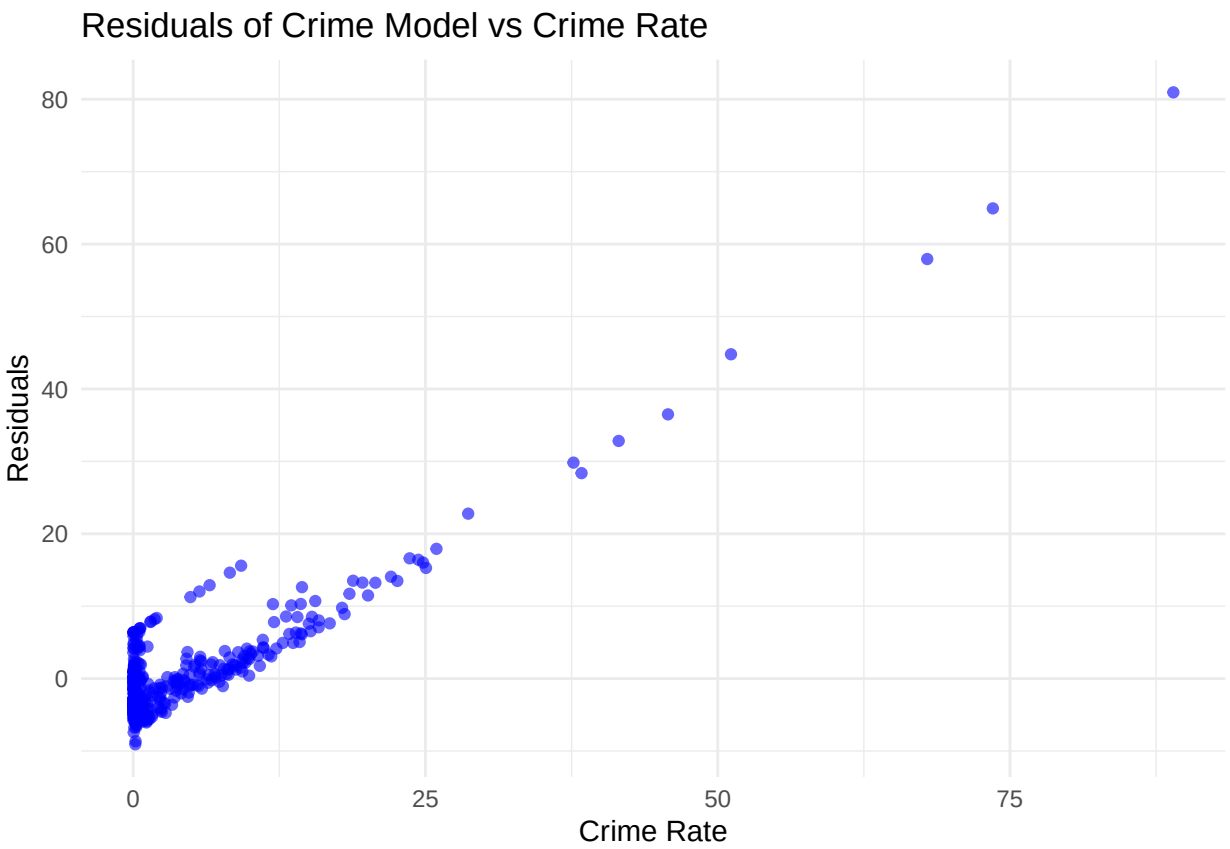



```
#subsetting residuals to examine

BostonResiduals<- Boston %>%
  mutate(residuals = residuals(crimemodel))
```

This is not a good fit for the model because the linear model poorly fits the data. Firstly, the regression is poorly accounting for the majority of the suburbs with a crime rate around 0, with fewer cases as we stray further from 0. Next, the line has a poor fit because it is heteroscedastic. The residuals increase as the crime rate increases.

1G: Calculating our model by hand

```
#taking our residuals and plotting them

ggplot(BostonResiduals, aes(x = crim, y = residuals)) +
  geom_point(alpha = 0.6, color = "blue") +  # Residuals plot
  labs(
    title = "Residuals of Crime Model vs Crime Rate",
    x = "Crime Rate",
    y = "Residuals"
  ) +
  theme_minimal()
```



After plotting all of the residuals, we can graphically see how the heteroscedasticity of the residuals in the previous graph exists. In this graph, we see more bias when crime rates increase, as the observed values of our model stray further away from the predicted regression line.