

# Project 2: Group 1 - Boston Housing Racial Bias

Will Johnson

12/3/2021

## Exploring our Dataset

```
# Download our dataset
library(mlbench)
data(BostonHousing2)
data <- BostonHousing2
head(data)
```

```
##      town tract      lon      lat medv cmedv      crim zn indus chas  nox
## 1  Nahant  2011 -70.9550 42.2550 24.0  24.0 0.00632 18  2.31   0 0.538
## 2 Swampscott 2021 -70.9500 42.2875 21.6  21.6 0.02731  0  7.07   0 0.469
## 3 Swampscott 2022 -70.9360 42.2830 34.7  34.7 0.02729  0  7.07   0 0.469
## 4 Marblehead 2031 -70.9280 42.2930 33.4  33.4 0.03237  0  2.18   0 0.458
## 5 Marblehead 2032 -70.9220 42.2980 36.2  36.2 0.06905  0  2.18   0 0.458
## 6 Marblehead 2033 -70.9165 42.3040 28.7  28.7 0.02985  0  2.18   0 0.458
##      rm age  dis rad tax ptratio      b lstat
## 1 6.575 65.2 4.0900  1 296    15.3 396.90  4.98
## 2 6.421 78.9 4.9671  2 242    17.8 396.90  9.14
## 3 7.185 61.1 4.9671  2 242    17.8 392.83  4.03
## 4 6.998 45.8 6.0622  3 222    18.7 394.63  2.94
## 5 7.147 54.2 6.0622  3 222    18.7 396.90  5.33
## 6 6.430 58.7 6.0622  3 222    18.7 394.12  5.21
```

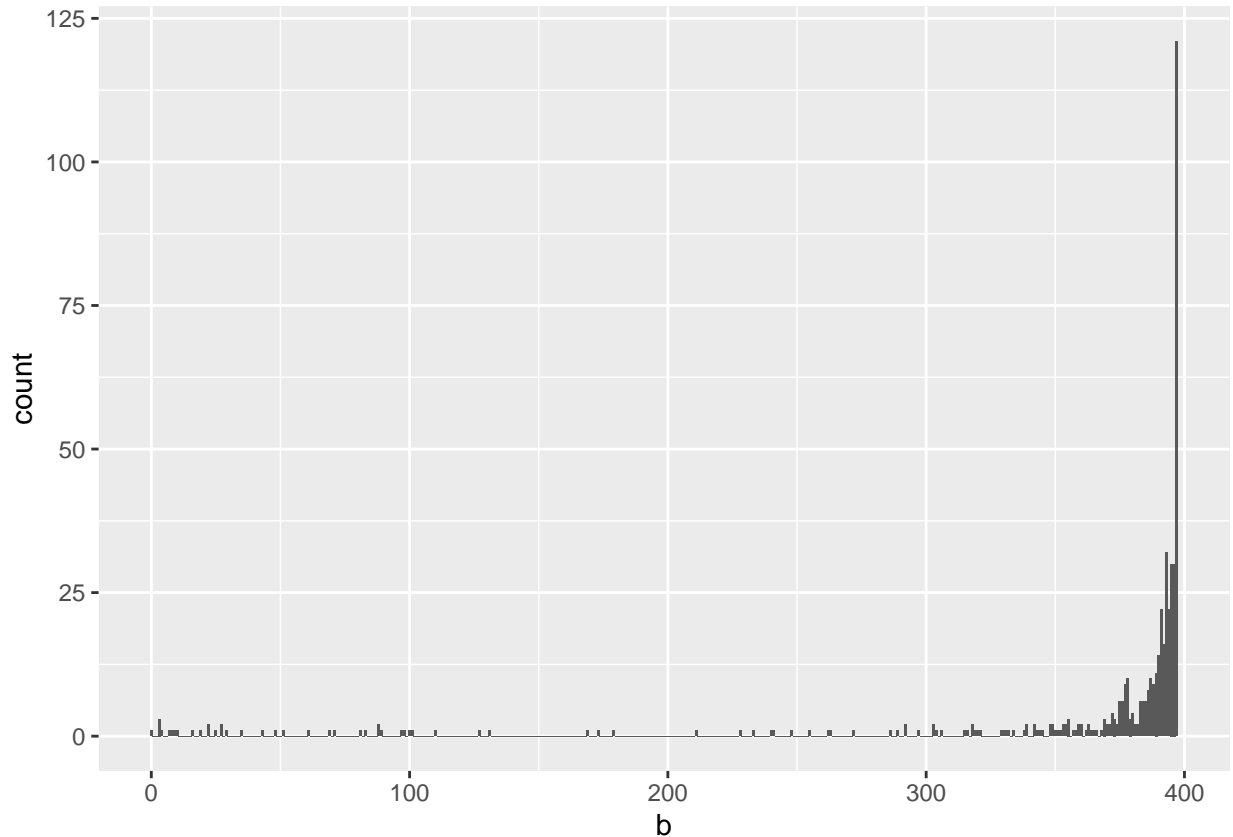
## “Majority Black” response variable

The variable “b” is described in the census documents as...

$1000(B - 0.63)^2$  where B is the proportion of blacks by town

Originally, we had planned to set this as a “majority black” binary variable. Based on the description, we thought the values would be between 0 and 1, or possibly 0 and 100. But the variable itself was a strange one instead. Take a look at the distribution of this variable b.

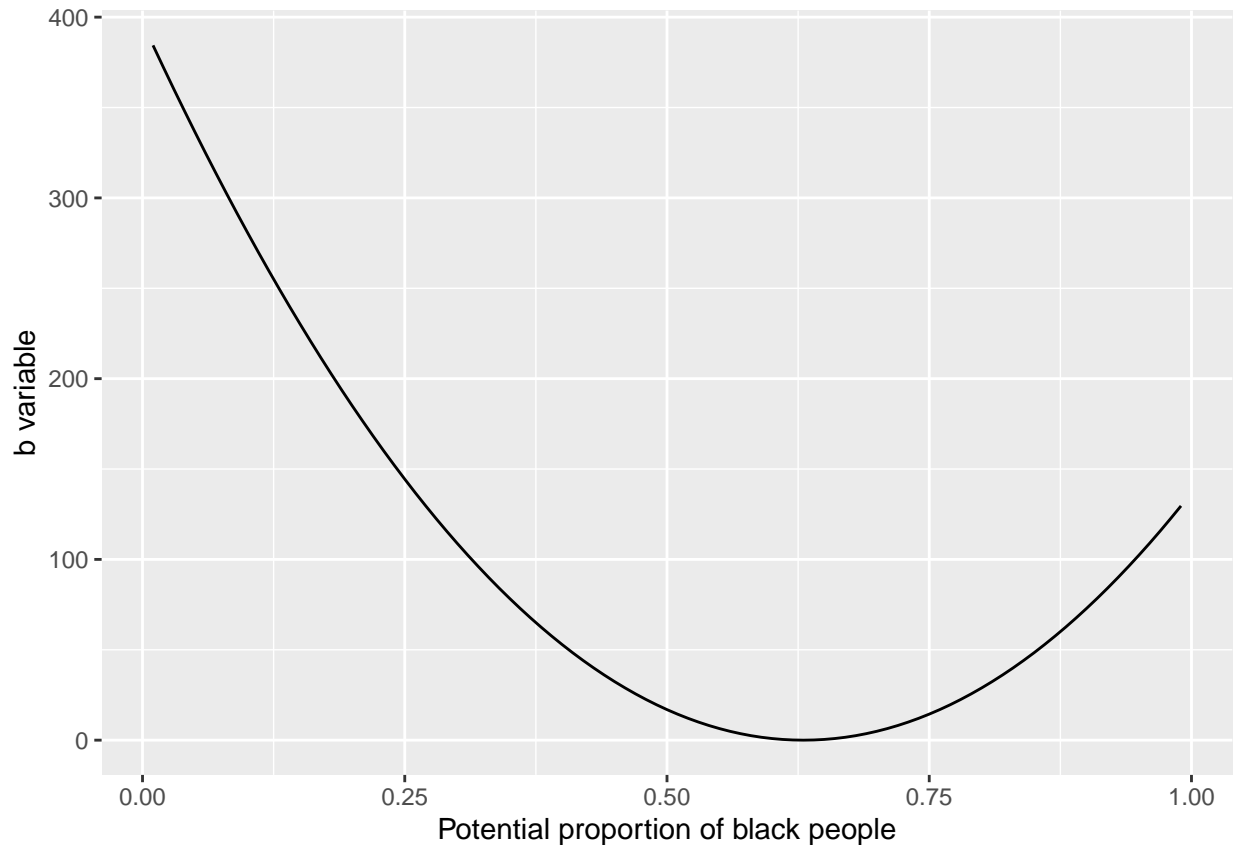
```
# plot the distribution on the "b" variable.
ggplot(data, aes(x = b))+
  geom_histogram(binwidth = 1)
```



Not only is the range very odd (0 to 400), the majority of these towns are in the high-end, suggesting that the data has been transformed so that the higher the value is, the whiter the neighborhood.

The most interesting part about our findings is that if we try to reverse engineer the original B value (based on their seemingly arbitrary formula  $b=1000(B-.63)^2$ ), it's not possible to get the answer. The result of the previously-mentioned formula is a quadratic curve, plotted below:

```
# Using the dataset's documentation, plot potential values for proportions of black people
# against each value of `data$b`.
props <- c()
i <- .01
for(x in 1:99){
  props <- c(props, 1000*(i-.63)**2 )
  i = i + .01
}
parab <- data.frame(props, seq(.01, .99, by=.01))
colnames(parab) = c('x', 'y')
ggplot(data = parab, aes(x=y, y=x ))+
  geom_line()+
  labs(x = "Potential proportion of black people", y='b variable')
```



Looking at this, it appears that the dataset is impossible to perfectly break back down into its original B variable because for some values, there are potentially 2 values of B for one value of the transformed b variable. Specifically, anything more than 26% black in a neighborhood becomes vague.

Some brief external research suggests that this transformation on the variable was set to create a pseudo-parabolic relationship to account for an initial drop in value when the proportion was too mixed-race. After 75% black, the effect was expected to rise again because of preference for a neighborhood of one's own race or "self-segregation" [Harrison,Rubinfeld, 96].

To quote an assistant professor of Mathematics at CUNY who researched this:

Harrison and Rubinfeld appear to have decided on a threshold of 63% at which to switch the regime of price decline to price increase (i.e. a so-called "ghetto threshold") [Carlisle,1]

Given that this was their decision at the time, we'll follow the precedent they set in the 1970's. We'll assess if there is a relationship between where a home falls on that threshold to see its effect on the price of a home, as well as try to predict where on that threshold a house will fall for our logistic model.

Plugging our 26% threshold into that formula, we get:

$$1000 * (0.26 - 0.63)^2 = 136.9$$

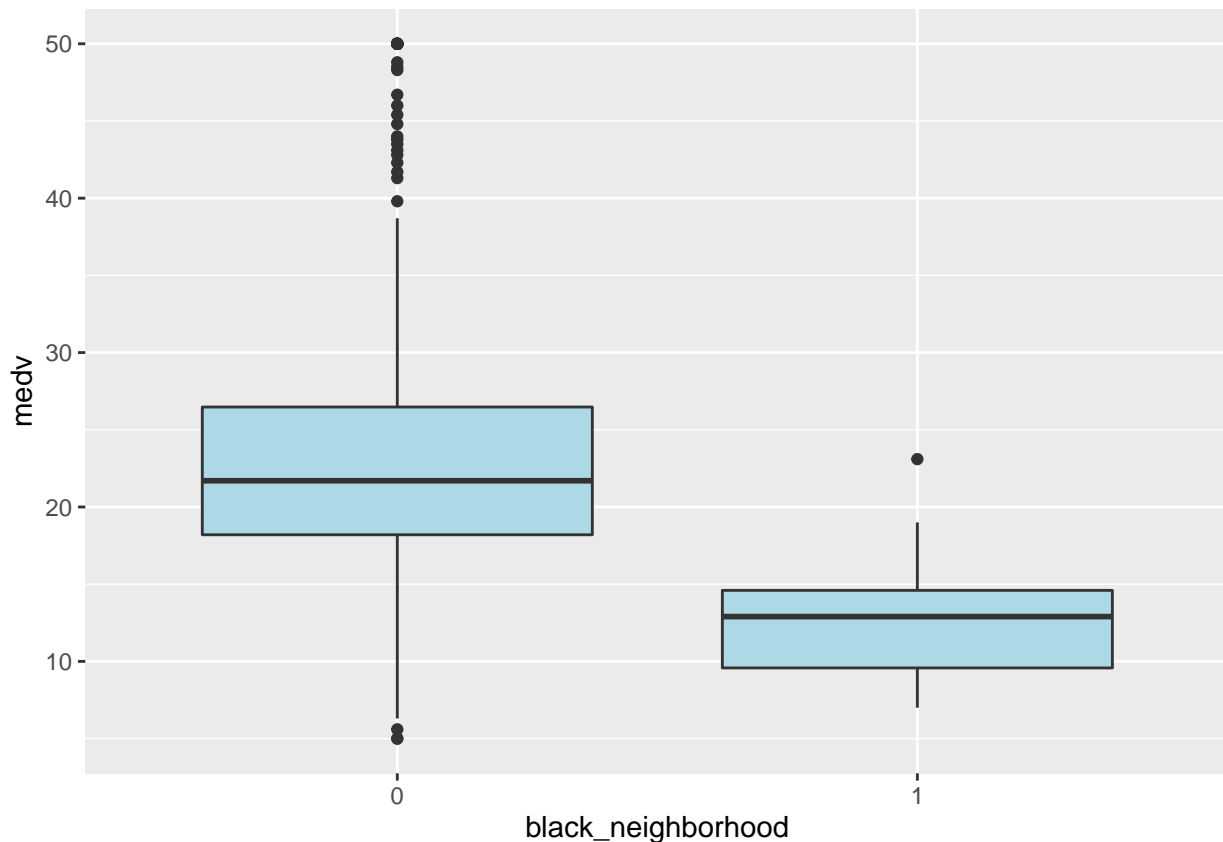
so `black_neighborhood = 1` if `b <= 136.9`

```
# code to create binary black_neighborhood variable
data <- data %>%
  mutate(black_neighborhood = (b <= 136.9) * 1)
# what % of our neighborhoods are black neighborhoods by this definition?
sum(data$black_neighborhood)/nrow(data)
```

```
## [1] 0.07114625
```

Finally, let's chart this variable to see how some of the other's fall across it. Namely our price variable

```
data$black_neighborhood <- as.factor(data$black_neighborhood)
# plot our new variable and the price distributions across it.
ggplot(data, aes(x= black_neighborhood, y=medv))+
  geom_boxplot(fill="light blue")
```



Definitely seems like there's a relationship between these two and price. Looking at it, we should assume there should be predictive value including this variable in predicting price.

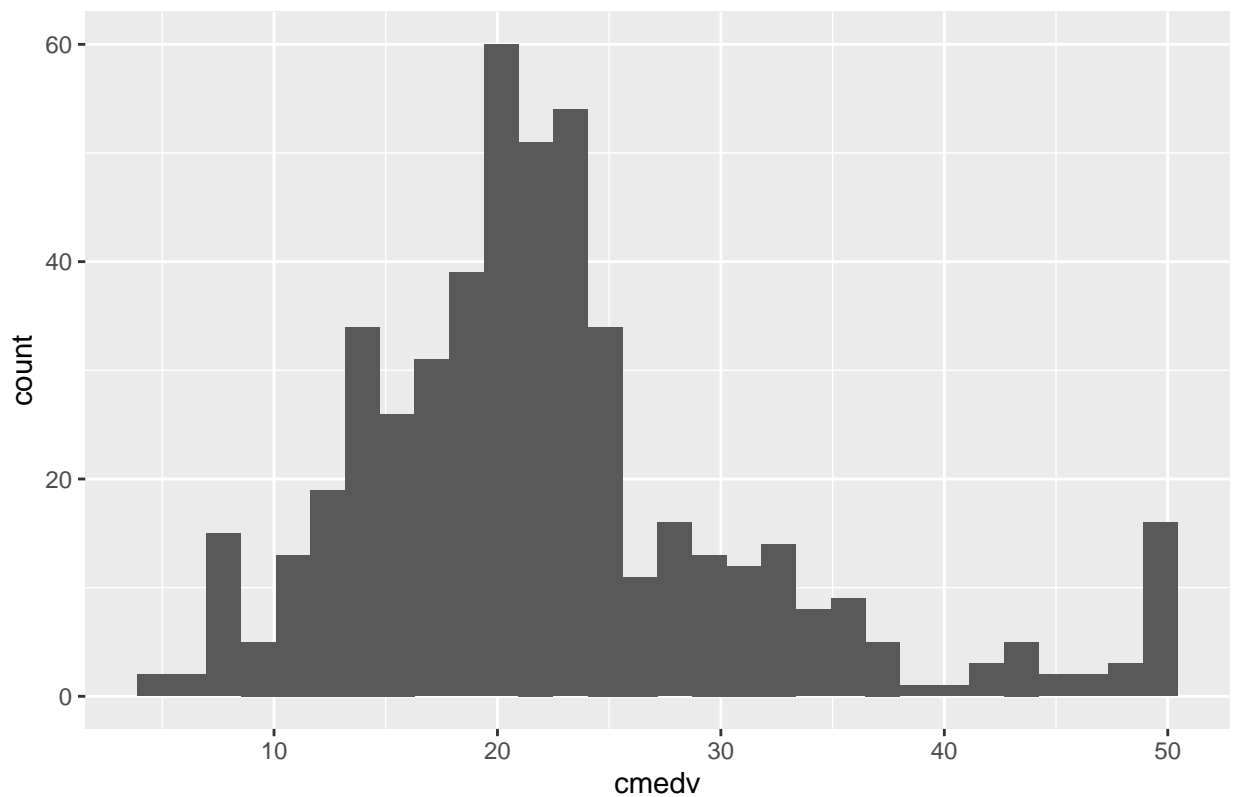
## Part 1 - EDA : How much does the relative blackness of a neighborhood affect price?

First we'll start by looking at a handful of other visuals to see which variables seem to affect price beyond our "black\_neighborhood" variable. We'll start with looking out our prices overall.

```
ggplot(data, aes(x=cmedv))+
  geom_histogram()+
  labs(title = "Distribution of Median Housing Values per Area")
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

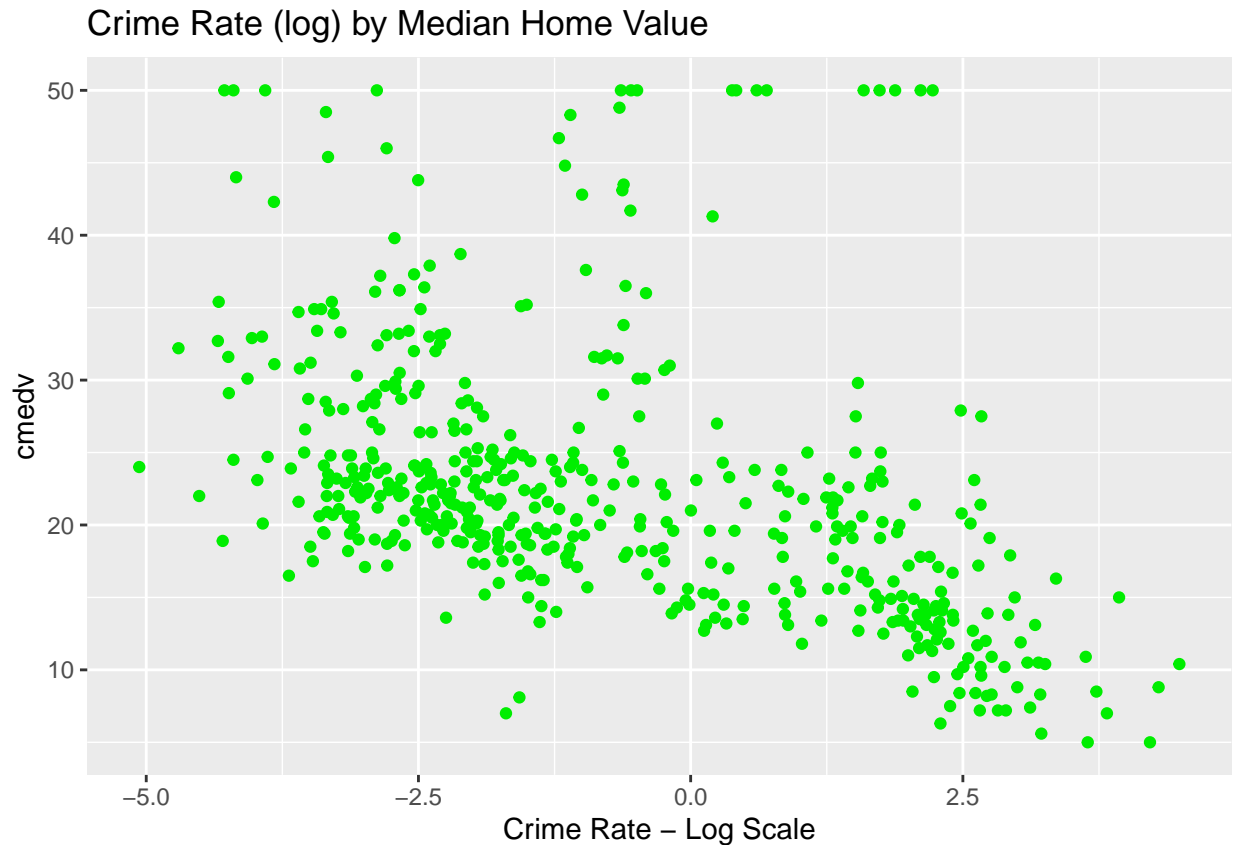
Distribution of Median Housing Values per Area



Looks mostly normally distributed with a handful of towns with a particularly high median value. We should look out for those to be potential outliers later.

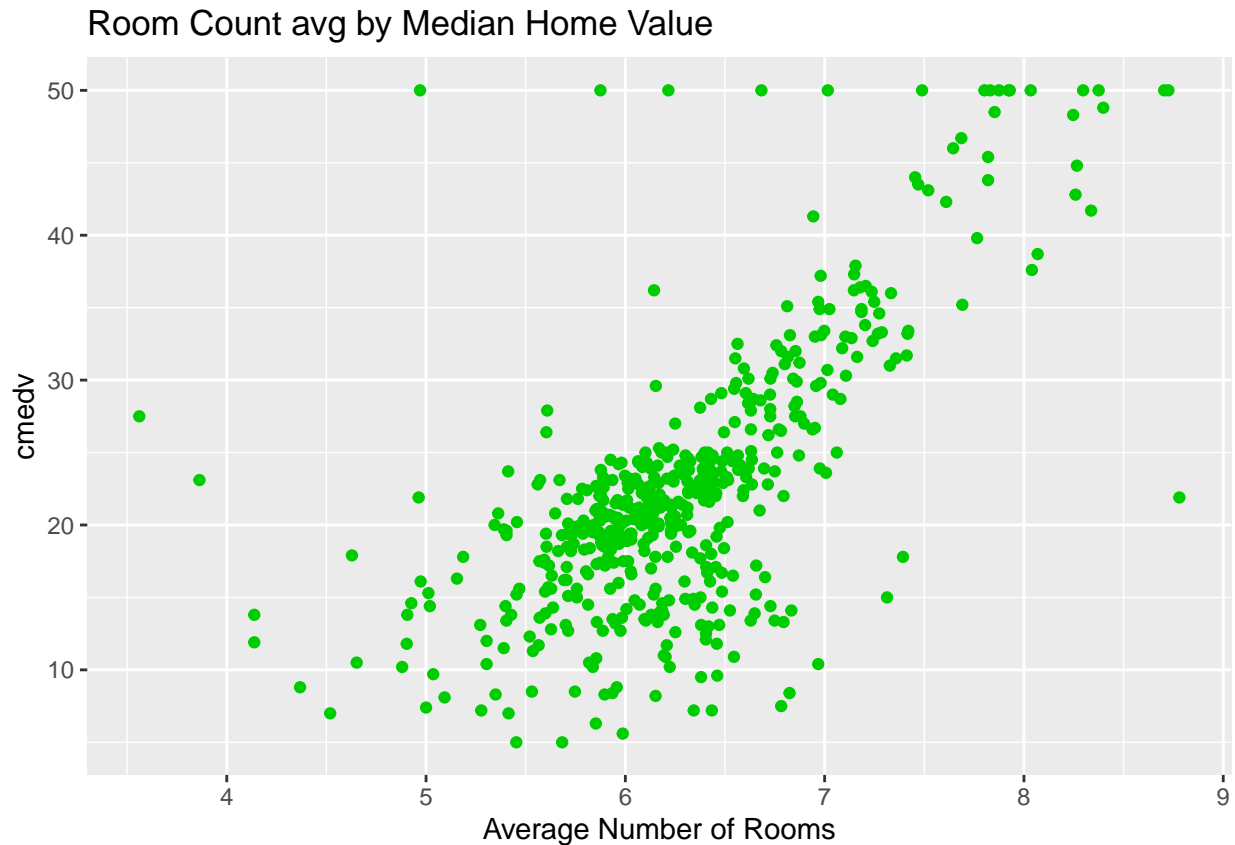
Next, some scatterplots between quantitative variables.

```
ggplot(data, aes(x = log(crim), y = cmedv))+  
  geom_point(color = 'green2')+  
  labs(title="Crime Rate (log) by Median Home Value",  
        x = "Crime Rate - Log Scale")
```



Vaguely, it looks as though the more the crime rate increases, the values of houses start to decrease. We used a log scale for crime here because otherwise the majority of towns had too low a crime rate to see this trend. Oddly, there are some expensive towns that also have relatively higher crime rates per capita as well. Next, let's see if the size of the houses make a difference based on the average number of rooms per house.

```
ggplot(data, aes(x = rm, y = cmedv))+  
  geom_point(color = 'green3')+  
  labs(title="Room Count avg by Median Home Value",  
        x = "Average Number of Rooms")
```

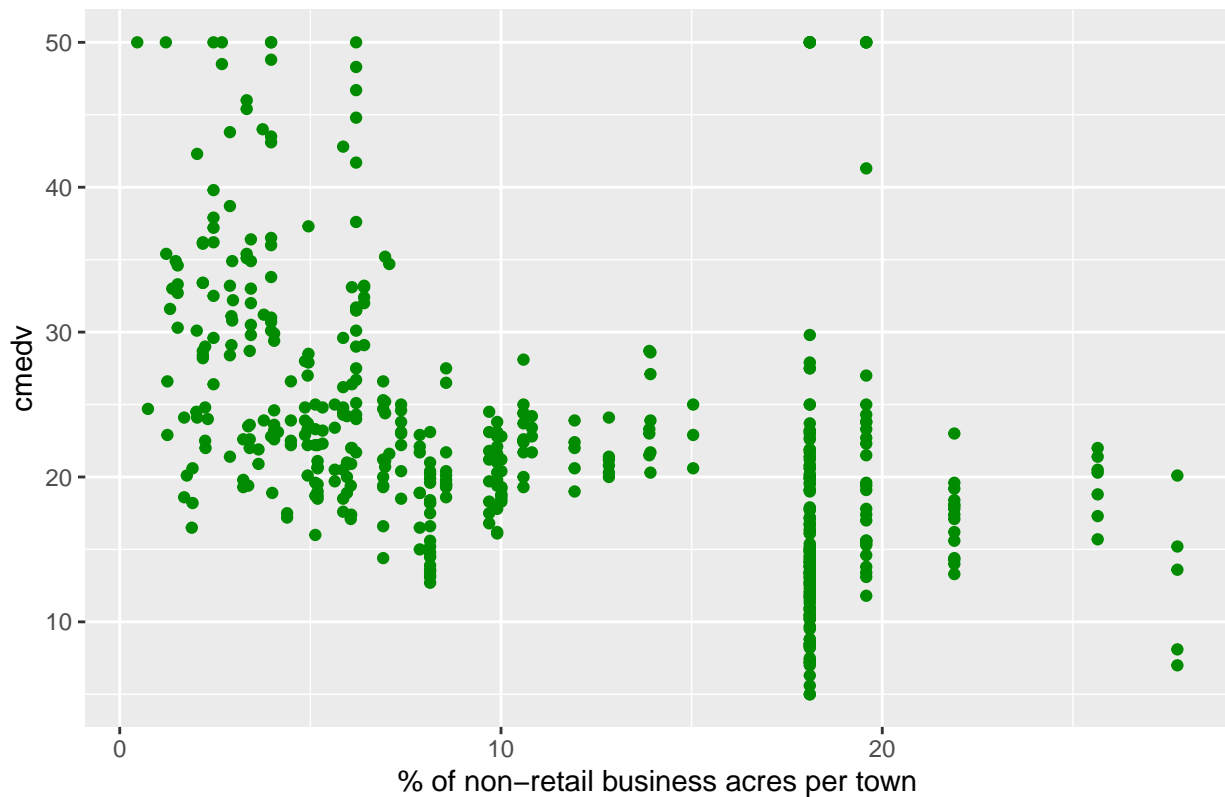


Pretty clear relationship here, though again we have some odd high-leverage points here (by observation) where there are some towns where the median value is around 50,000 but the average number of rooms is far lower. Let's see if there's another variable that can account for those handful of towns that don't fit our patterns.

Let's try the industry proportion. Maybe things are expensive in a business neighborhood where high expenses and crime rates can happen simultaneously.

```
ggplot(data, aes(x = indus, y = cmedv))+
  geom_point(color = 'green4')+
  labs(title="Proportion of non-retail business by Median Home Value",
       x = "% of non-retail business acres per town")
```

Proportion of non-retail business by Median Home Value



```
p1 <- ggplot(data, aes(x = nox, y = cmedv))+
  geom_point(color = 'purple1')+
  labs(title="NO2 concentration by Med. Value",
        x = "Nitric Oxides Concentration (parts per 10m)")

p2 <- ggplot(data, aes(x = dis, y = cmedv))+
  geom_point(color = 'purple2')+
  labs(title="Distance to Work by Med. Value",
        x = "Weighted Distances to Employment Centres")

p3 <- ggplot(data, aes(x = rad, y = cmedv))+
  geom_point(color = 'purple3')+
  labs(title="Highway Access by Med. Value",
        x = "Index of Accessibility to Highways")

p4 <- ggplot(data, aes(x = tax, y = cmedv, color=rad))+
  geom_point(alpha=.5)+
  labs(title="Property Tax Rate by Med. Value",
        x = "Property Tax rate per $10,000")

# Proportions
p5 <- ggplot(data, aes(x = ptratio, y = cmedv))+
  geom_point(color = 'orange2')+
  labs(title="Student-Teacher Ratio by Med. Value",
        x = "Studnets per Teacher")
```



```

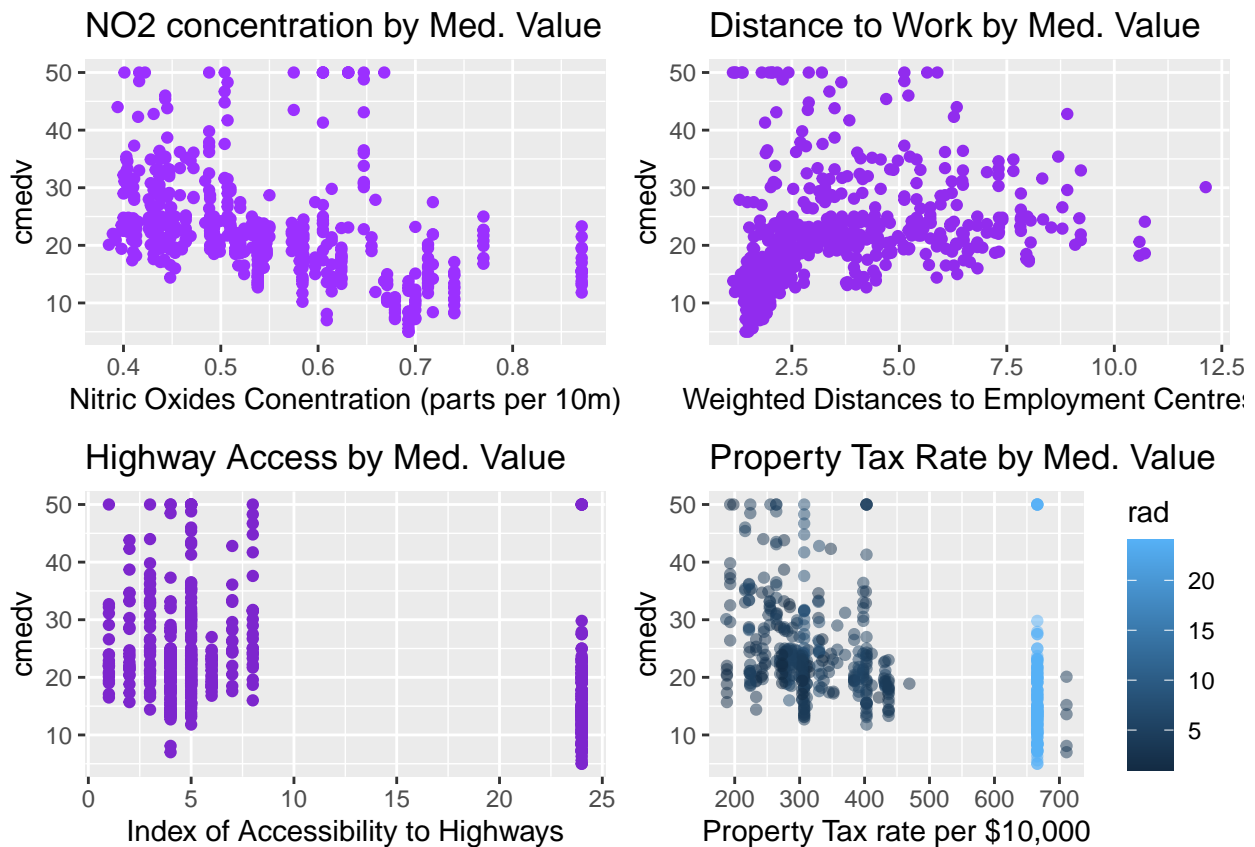
p6 <- ggplot(data, aes(x = age, y = cmedv))+
  geom_point(color = 'orange3')+
  labs(title="Age by Med. Value",
        x = "% of homes build prior to 1940")

p7 <- ggplot(data, aes(x = lstat, y = cmedv))+
  geom_point(color = 'orange4')+
  labs(title="% of Lower Class by Med. Value",
        x = "% of 'Lower Status' Citizens")

# Boolean Variables
p8 <- ggplot(data, aes(x = chas, y = cmedv))+
  geom_boxplot(fill = "lightblue2")+
  labs(title="House By the River (y/n) by Med. Value",
        x = "Homes Line Charles River")

grid.arrange(p1,p2,p3,p4,
  ncol = 2,
  clip = TRUE
)

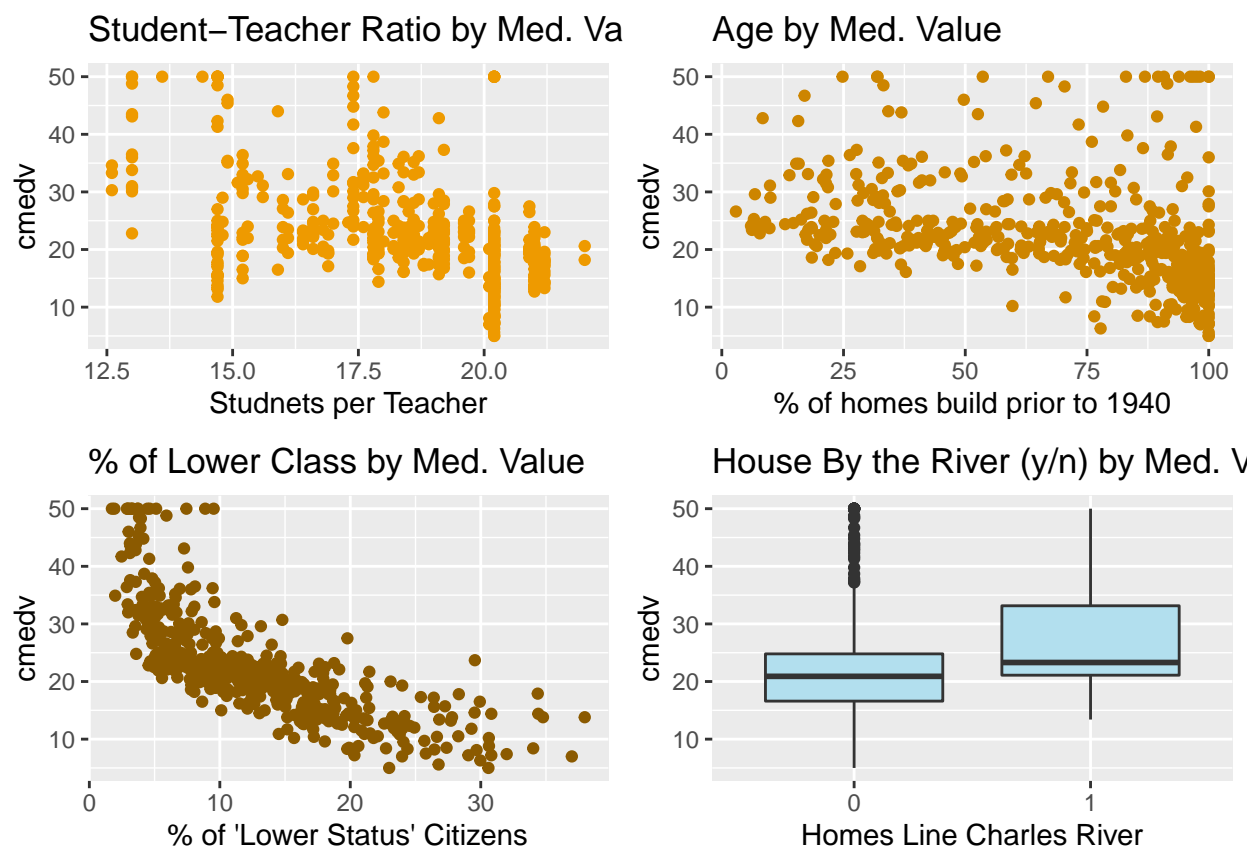
```



```

grid.arrange(p5,p6,p7,p8,
  ncol = 2,
  clip = FALSE
)

```



Something important to note about the above relationships. The Tax rate chart seems to cleave our dataset into two. The only other predictor that does this is the Index of Radial Highway access. Layering both together into the same chart via color, you can see that the group in the highest tax bracket is entirely made up of those areas with close access to the highways. Accessibility to highways seems to create a rift between groups in our dataset.

Other interesting relationships is that being next to the river seems to have a positive impact on price, whereas the typical age of the homes seems to have a negative impact (but only after 75 years). Finally, there's this indexed metric called `lstat` which measures the "Percentage of Lower Status of the Population". Looking further into the 1970 Census paper, it looks like this metric measures the following.

```
""" > Proportion of population that is lower status = 1/2 (proportion of adults without, some high school
education and proportion of male workers classified as laborers). The logarithmic specification implies that
socioeconomic status distinctions mean more in the upper brackets of society than in the lower classes.
Source: 1970 U. S. Census [Harrison & Rubinfeld, 82] """
```

So it looks as though this variable is a combination of several class-related factors that have been aggregated to a given township. It's a shame that we don't have those variables split out so we can use them in our model as individual predictors, but there is clearly a negative relationship between the % of "lower class" people in the town and the median value. This predictor is very much a product of it's time.

## Judgement Call

There are several reasons as to why there are clusters of neighborhoods at \$50k. It's possible that the researchers had blank values and filled those in with the maximum value they had data for. It's also possible that the original researchers decided to filter out any neighborhoods that were above 50k, which could make sense for their research.

For the purposes of our research, we will be *ignoring any of these neighborhoods valued at cmedv = \$50k*. Further reserach might yield insight into why those neighborhoods look this way and a published paper should try to do so, but for this assignment we'll keep our dataset within those bounds. We lose 16 neighborhoods doing so, or around 3% of our data.

```
data <- data %>%
  filter(cmedv<50)
```

## Part 2: Regression

To start with, we'll run a very basic linear model with all the other predictors in it to see our benchmark of performance with no changes or alterations.

### Benchmark Model Summary

```
benchmark <- lm(cmedv ~ crim + zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio + lstat +
summary(benchmark)
```

```
##
## Call:
## lm(formula = cmedv ~ crim + zn + indus + chas + nox + rm + age +
##      dis + rad + tax + ptratio + lstat + black_neighborhood, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5708  -2.3286  -0.4998   1.7388  15.6359
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    35.063889   3.894947   9.002 < 2e-16 ***
## crim          -0.103303   0.025655  -4.027 6.58e-05 ***
## zn             0.036976   0.011069   3.340 0.000902 ***
## indus         -0.047988   0.048897  -0.981 0.326895
## chas1          0.499567   0.728094   0.686 0.492965
## nox          -12.660202   2.997474  -4.224 2.88e-05 ***
## rm             3.752773   0.351088  10.689 < 2e-16 ***
## age          -0.024629   0.010472  -2.352 0.019081 *
## dis          -1.252766   0.157775  -7.940 1.46e-14 ***
## rad            0.254382   0.052210   4.872 1.50e-06 ***
## tax          -0.014307   0.002945  -4.859 1.61e-06 ***
## ptratio       -0.791543   0.103424  -7.653 1.10e-13 ***
## lstat        -0.356155   0.041690  -8.543 < 2e-16 ***
## black_neighborhood1 -2.970706   0.719234  -4.130 4.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.692 on 476 degrees of freedom
## Multiple R-squared:  0.7845, Adjusted R-squared:  0.7786
## F-statistic: 133.3 on 13 and 476 DF,  p-value: < 2.2e-16
```

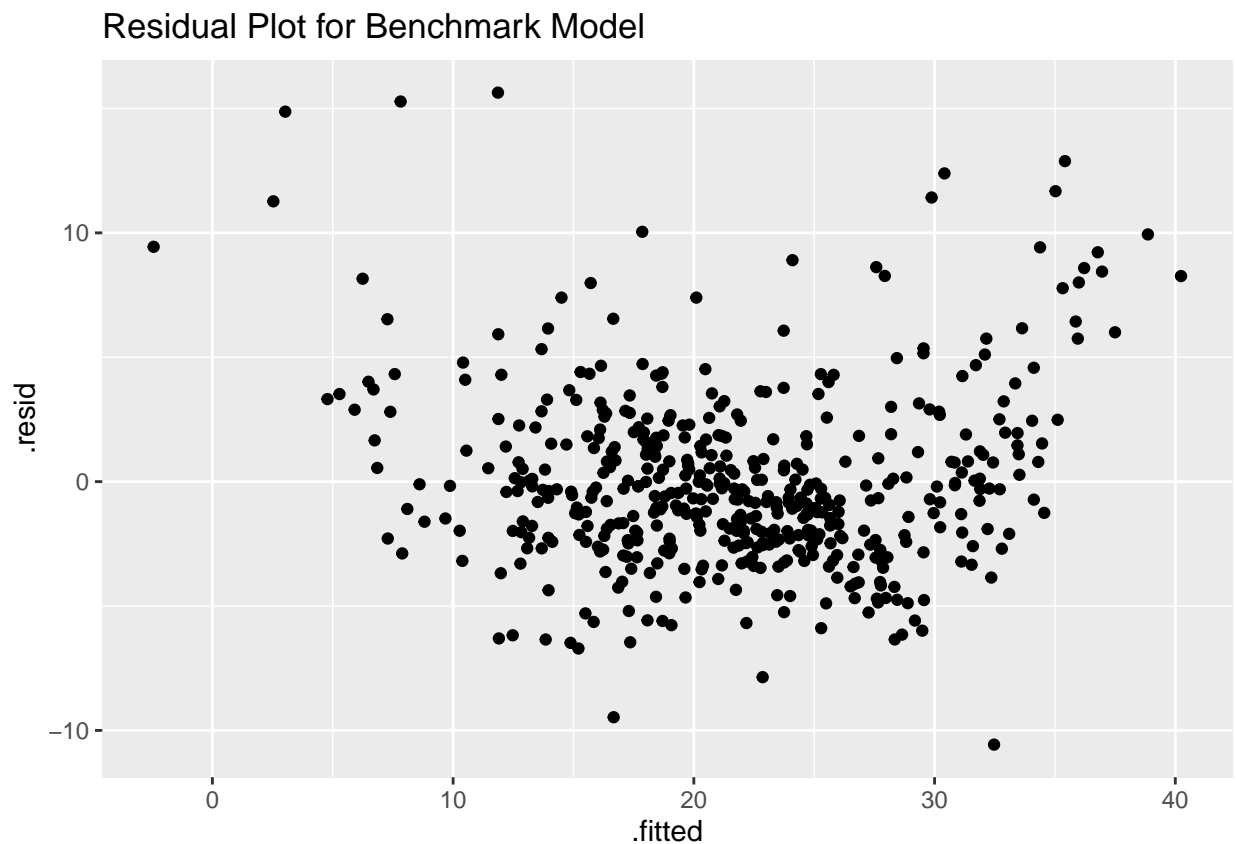
With the above benchmark set, it appears we are off to a good start. Our adjusted R-squared suggests that nearly 74% of the variance in our data is explained by our model. Our F-statistic is large and the relative p-value suggests the following result to a hypothesis test:

$H_0$ : \$ There is no difference between our model and the intercept alone.  $H_A$ : \$ There is difference between our model and the intercept alone.

Since our p-value is below .05, we can reject the null and conclude that some combination of the above predictors improves the model beyond the intercept alone.

To help us validate that this model is working, let's take a look at our residual values.

```
# plot our residual plot
ggplot(benchmark, aes(x = .fitted, y = .resid)) +
  geom_point() +
  labs(title='Residual Plot for Benchmark Model')
```



There are several assumptions that we should be checking for here. 1. Is the variance of our residuals consistent? 2. Is the mean of our Residuals 0? 3. Are the observations independent of each other?

There are other assumptions we can verify as well, but we'll start with these since we can identify issues with these using just the residual plot. From observing the residual plot, we can see that none of the above assumptions are met completely. The variance in our residuals seems to fluctuate as our fitted values increase. Likewise, the variance is not centered around 0 and a slight curve to a line is appearing, violating assumptions 2 and 3.

## NEXT STEPS \*\*\*\*\*

- log-transform variables to help meet assumptions
- Identify high-leverage points or outliers and see what we should do with them.
- Backward Selection (excluding `black_neighborhood`) to try to improve model performance.
- Final conclusions: How much does being a “Black Neighborhood” impact the median home price, other predictors held constant?

## Part 4: Logistic Regression

## Part 5: Check the prompt for more stuff.

## Bibliography

Michael Carlisle . “racist data destruction?” Medium, 13 June. 2019, <https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8>. Accessed 15 November. 2021.

David Harrison, Daniel L Rubinfeld, “Hedonic housing prices and the demand for clean air”, Journal of Environmental Economics and Management, Volume 5, Issue 1, 1978, Pages 81-102, ISSN 0095-0696.