Project 2: Group 1 - Boston Housing Racial Bias

Lauren Bassett, Will Johnson, Anoop Nath, Aishwarya Pradhan

12/3/2021

Exploring our Dataset

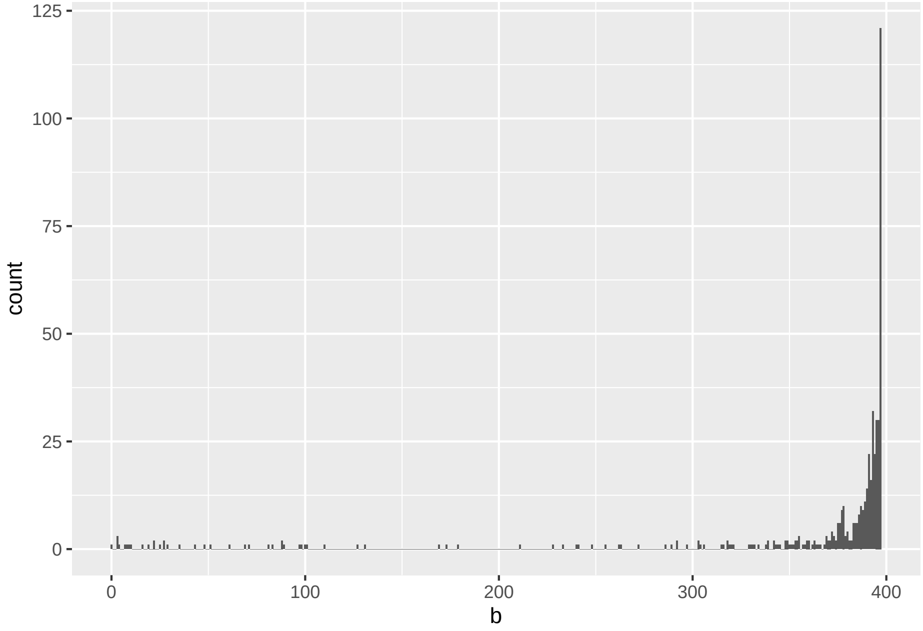
```
##         town tract      lon     lat medv cmedv    crim zn indus chas   nox
## 1     Nahant  2011 -70.9550 42.2550 24.0  24.0 0.00632 18  2.31    0 0.538
## 2 Swampscott  2021 -70.9500 42.2875 21.6  21.6 0.02731  0  7.07    0 0.469
## 3 Swampscott  2022 -70.9360 42.2830 34.7  34.7 0.02729  0  7.07    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)ˆ2 where B is the proportion of blacks by town Initially, our goal was to categorize the data using a "majority black" binary variable. Based on the description, we thought the values would be between 0 and 1, or possibly 0 and 100. Instead, the variable is distributed on a quadratic curve.

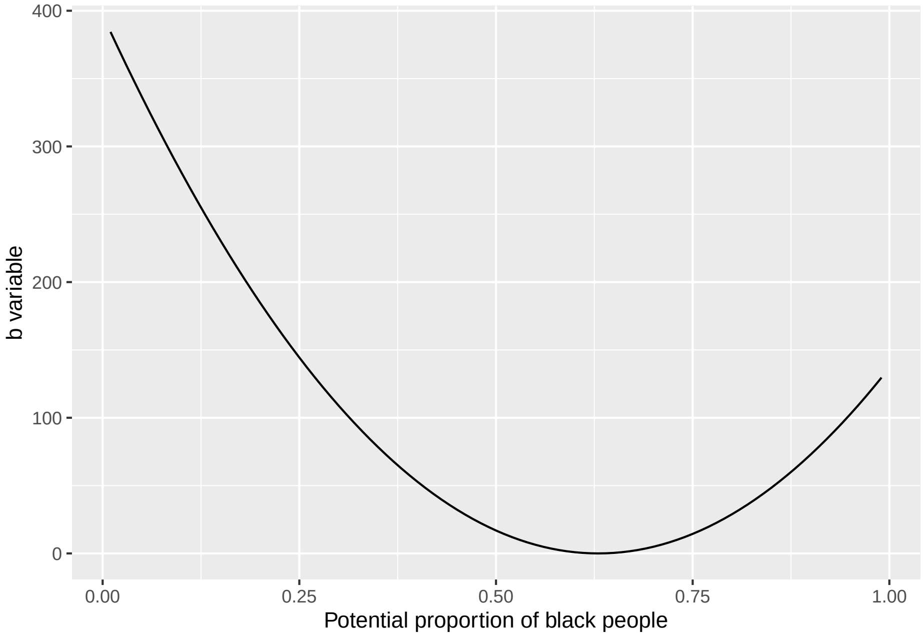
Histogram of b-variable



The proportion variable ranges from 0 to 400, and most of the data has values on the upper end of the range. The data has been transformed so that higher values represent neighborhoods with higher white populations.

If we try to reverse engineer the original B value (based on the 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:

Potential proportion of Black Population vs. b-variable



It is impossible to retrieve the original B variable because there are two possible values of B for each transformed variable. Neighborhoods with a black population higher than 26% become harder to determine.

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].

Researcher Michael Carlisle, an assistant professor of Mathematics at CUNY, said:

> 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]

We will continue to 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.

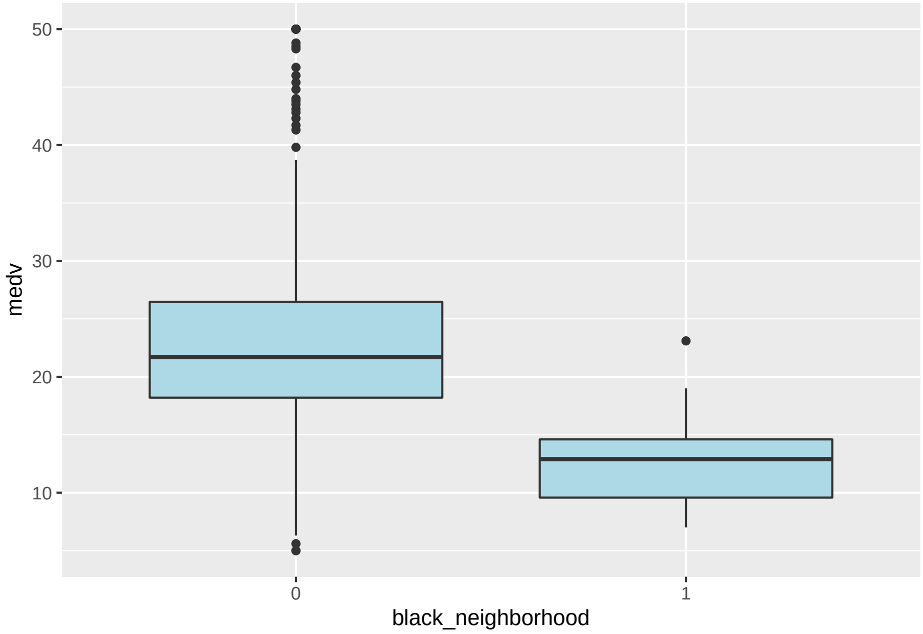
Plugging our 26% threshold into that formula, we get:

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

so black_neighborhood = 1 if b <= 136.9

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

After generating the binary variable, we generated visualizations to see how the other variables interact. Specifically, we are focused on the price variable.

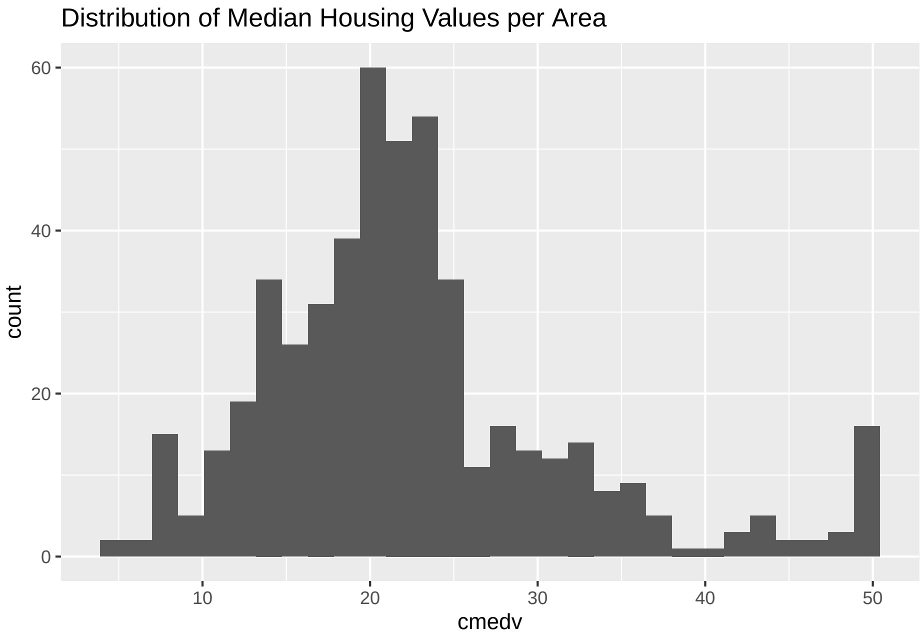


The visualizations above show a relationship between neighborhoods that are predominately black and the price of housing. Thus, we can assume that the black neighborhood indicator can be used as a predictor for housing prices.

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

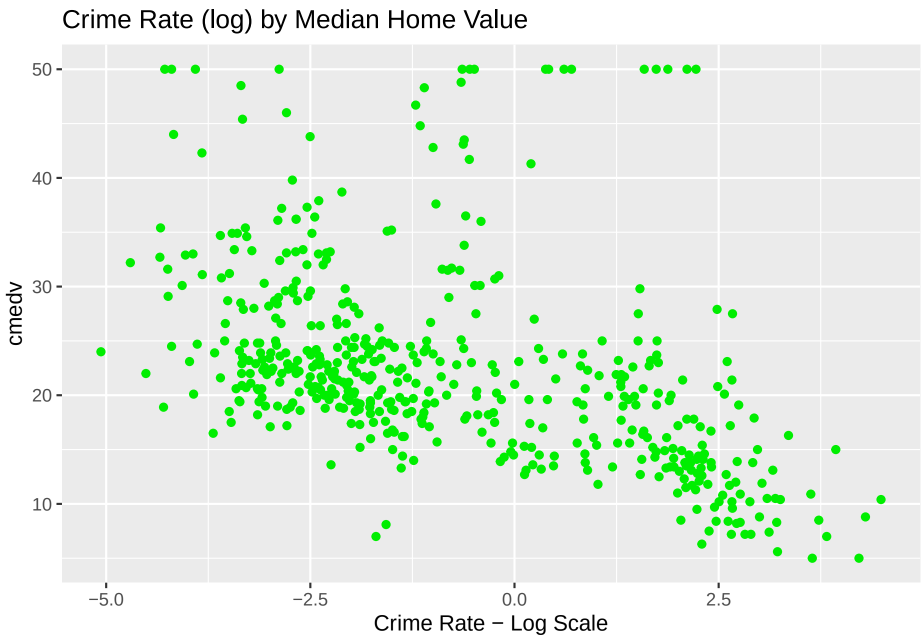
First, we will generate views to determine what other predictors affect the price of housing. We begin by looking at the distribtion of house prices across the entire dataset.

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



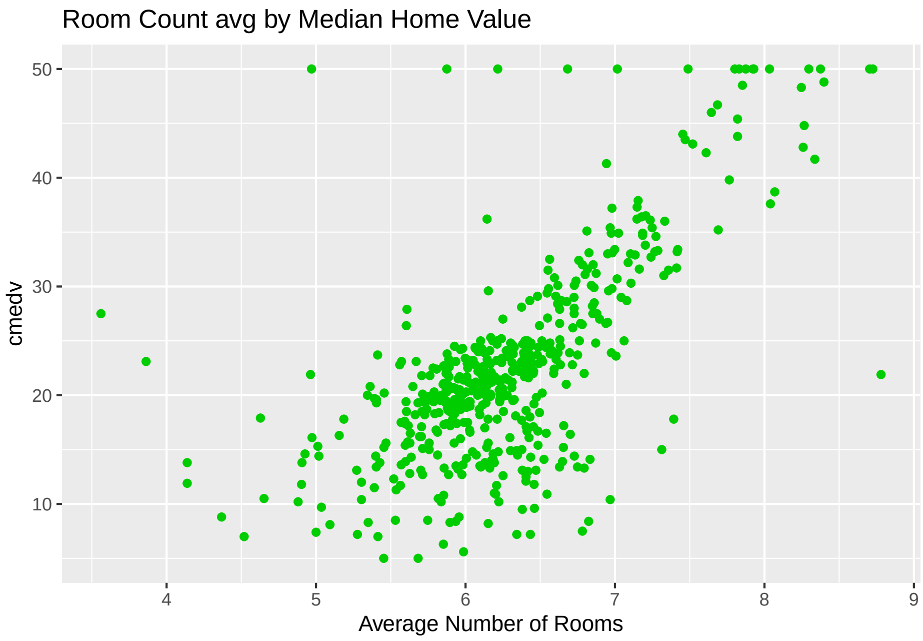
The price of housing appears normally distributed. There are a few towns with noticably higher prices. These may be potential outliers in the dataset.

Next, we generate scatterplots to describe the quantiative variables.



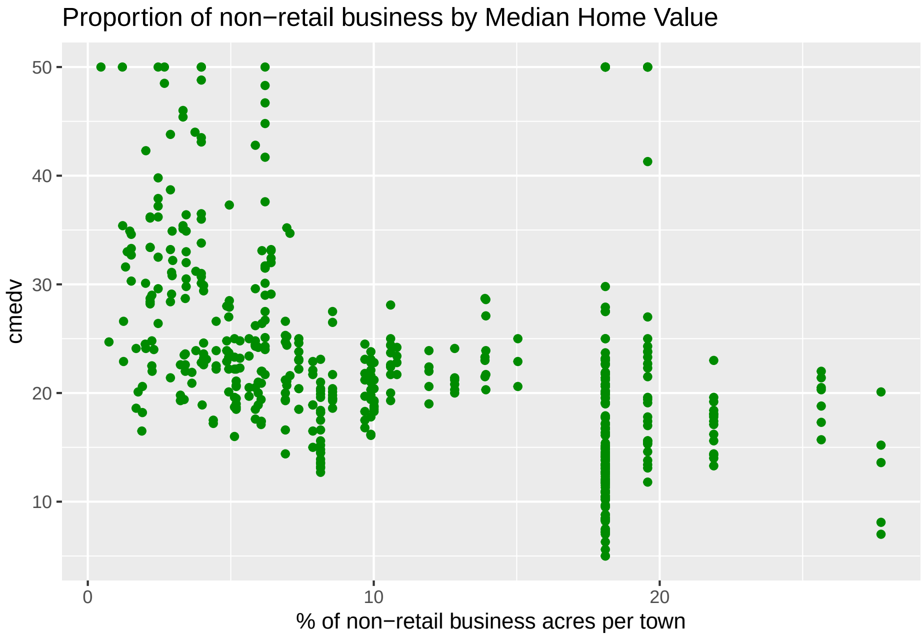
It appears that as crime rate increases, the median value of the home decreases. However, as most areas in the data set had extremely low crime rates, we used a log scale for crime to better highlight this trend. It is important to note that the relationship between crime and housing prices is not constant, as there are some expensive areas in the data set that have relatively higher crime rates per capita.

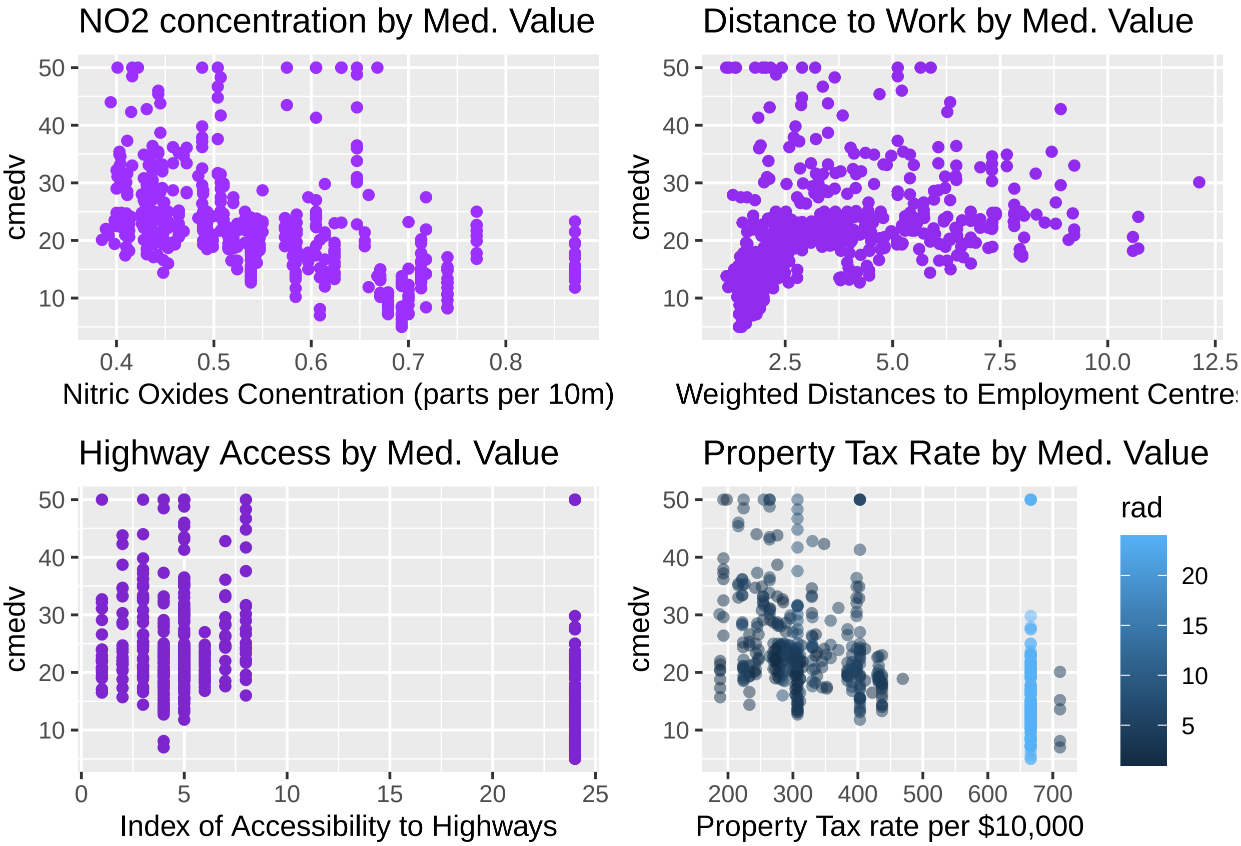
Next, we assess how the size of the home, approximated by the number of rooms, affects the median price.

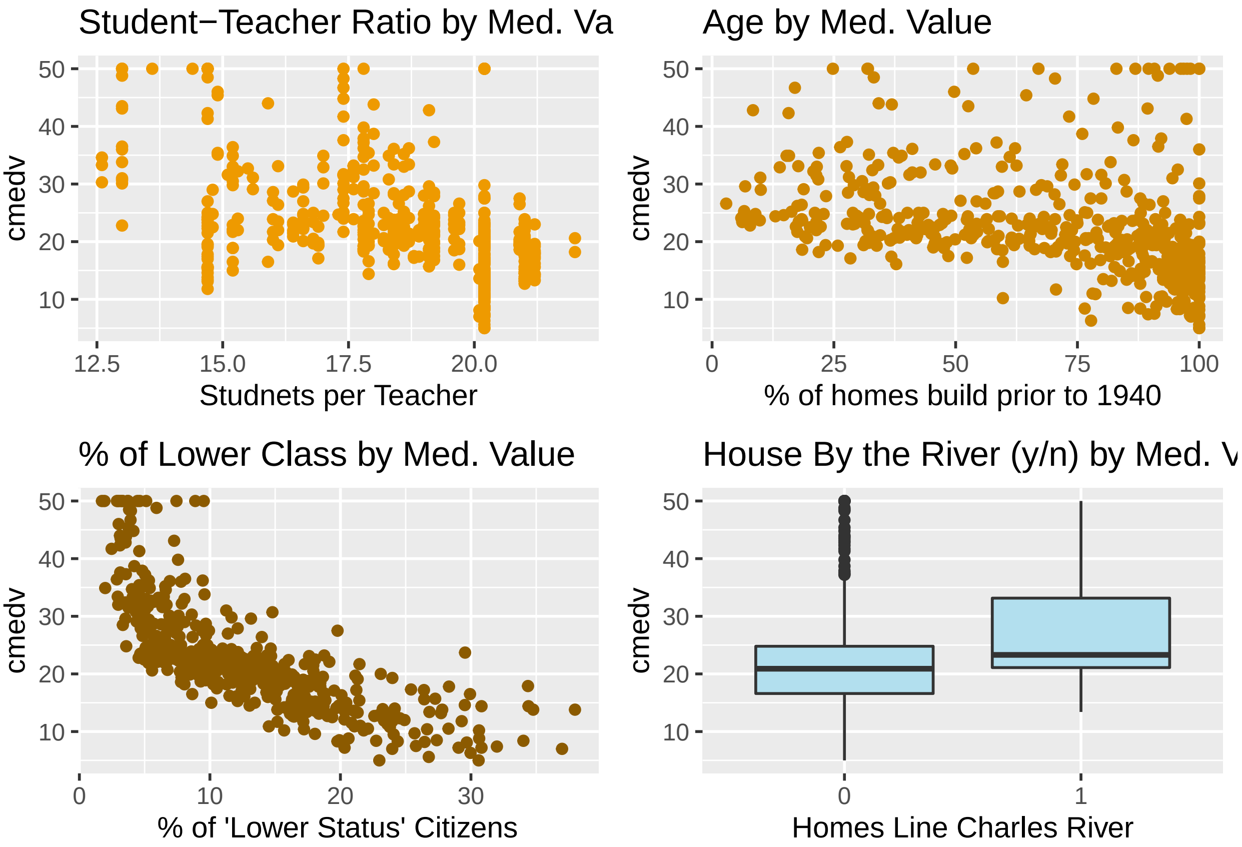


The relationship between number of rooms and median price is strong. However, there are some high-leverage points (by observation) where the median home value is around $50,000 but the average number of rooms is far lower. There may be other predictors that are influencing these high-leverage points.

One possible confounding variable could be the industry proportion. We believe business neighborhoods with high expenses may also have relatively high crime rates.







The views reveal key insights about our data set. The Tax rate chart divides the 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 divide between groups in our dataset.

Other relationships of note are the positive relationship between proximity to the river on price, and the negative relationship between age of the home on price. 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]

It appears that the variable is a combination of several class-related factors that have been aggregated to a given township. Unfortunatley, these variables are not able to be used as individual predictors, but there is a strong negative relationship between the % of "lower class" people in the town and the median value. ### 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.

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

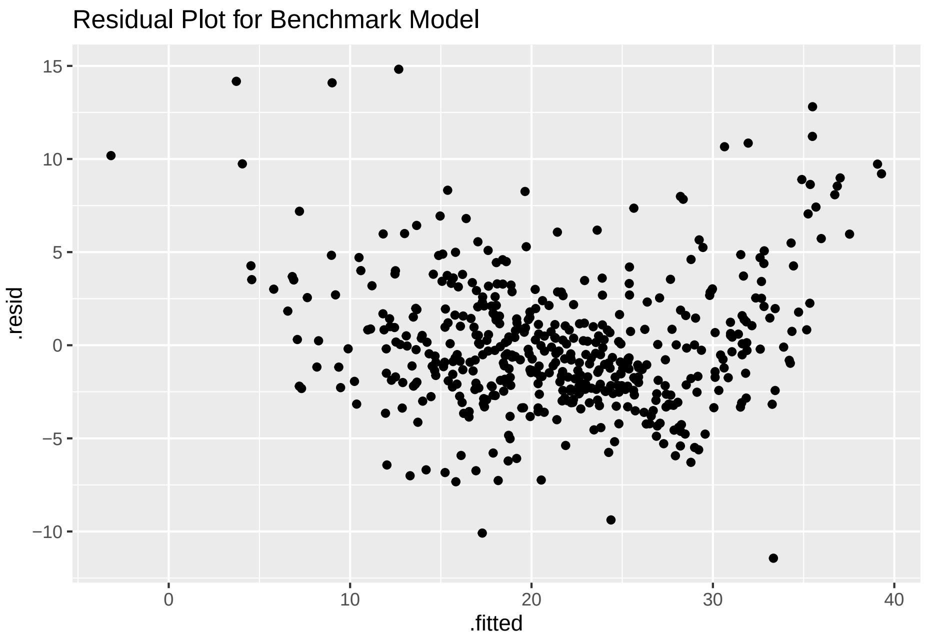
```
##
## Call:
## lm(formula = cmedv ~ crim + zn + indus + chas + nox + rm + age +
##     log(dis) + rad + tax + ptratio + lstat + black_neighborhood,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.436  -2.226  -0.503   1.640  14.819
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          40.396674   4.062053   9.945  < 2e-16 ***
## crim                 -0.131224   0.025892  -5.068 5.76e-07 ***
## zn                    0.022322   0.010256   2.176  0.03001 *
## indus                -0.069418   0.048739  -1.424  0.15502
## chas1                 0.741284   0.720308   1.029  0.30395
## nox                 -16.533140   3.092671  -5.346 1.40e-07 ***
## rm                    3.821131   0.346070  11.041  < 2e-16 ***
## age                  -0.029841   0.010469  -2.850  0.00456 **
## log(dis)             -6.059214   0.698115  -8.679  < 2e-16 ***
## rad                   0.276878   0.051704   5.355 1.33e-07 ***
## tax                  -0.015357   0.002913  -5.271 2.06e-07 ***
## ptratio              -0.783251   0.102289  -7.657 1.07e-13 ***
## lstat                -0.375410   0.041209  -9.110  < 2e-16 ***
## black_neighborhood1  -2.794221   0.711324  -3.928 9.83e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.651 on 476 degrees of freedom
## Multiple R-squared:  0.7893, Adjusted R-squared:  0.7835
## F-statistic: 137.1 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, we look at the residual values.



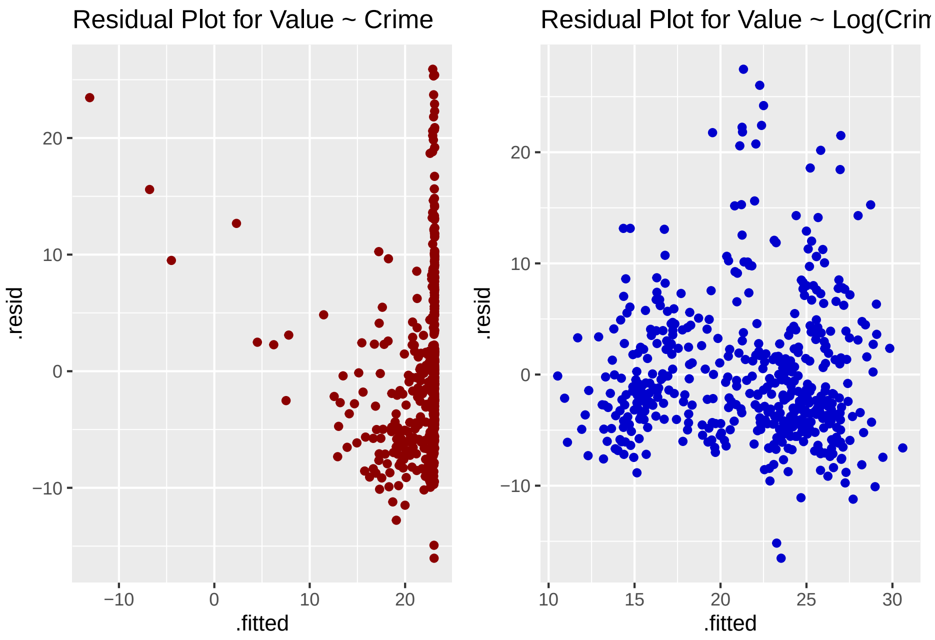
Next, we need to validate that our model complies with the following assumptions: 1. Is the variance of our residuals consistent? 2. Is the mean of our Residuals 0?

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.

Transformations

Ideally, we won't need to transform our response variable at all because we're hoping to interpret coefficients directly. Normally we would consider transforming the response variable first, since transforming the predictor does not affect the variance of the error terms and any transformations we do to the predictors might be skewed if we need to eventually transform the response. But the above suggests a curve rather than wild variance, so we'll start with predictors.

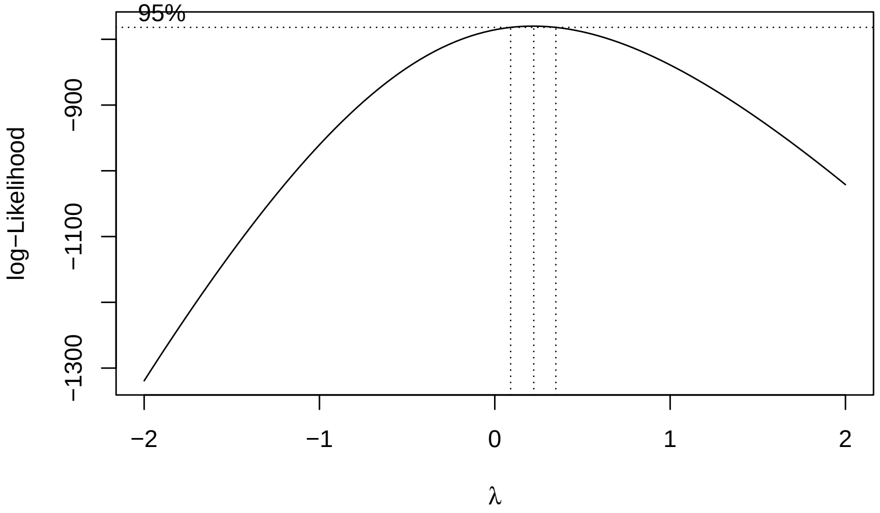
Since we did a log transformation on the crime rates when comparing them to cmedv, we'll start with seeing how well that improves the residuals for a simple model using just crime as a predictor.



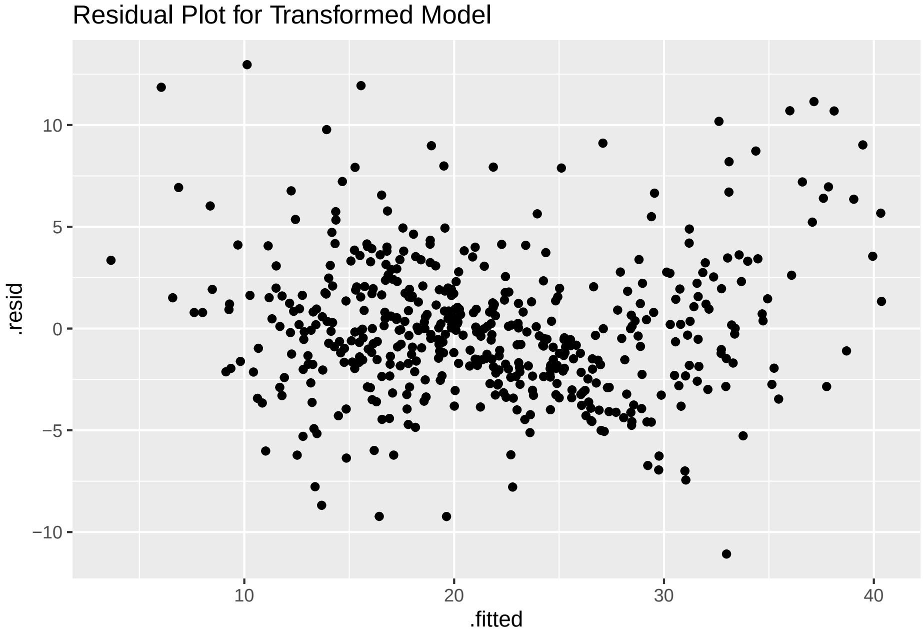
This is a large improvement. Let's also look at the `Distance to Work`, `House Age`, and `% Lower Class` features next, as they also seem to follow none-linear patterns. It's not inherently clear which one will work with which, so we'll use BoxCox to help us figure out which of these variables should be transformed in which ways.

Below, we'll show an example of one of several BOXCOX plots we used to help make our decisions. The other plots were very similar to this one below, that shows the boxcox plot for `cmedv` predicted by the `lstat` predictor, which measure the % of the population in "lower class".

BOXCOX for cmedv ~ Class



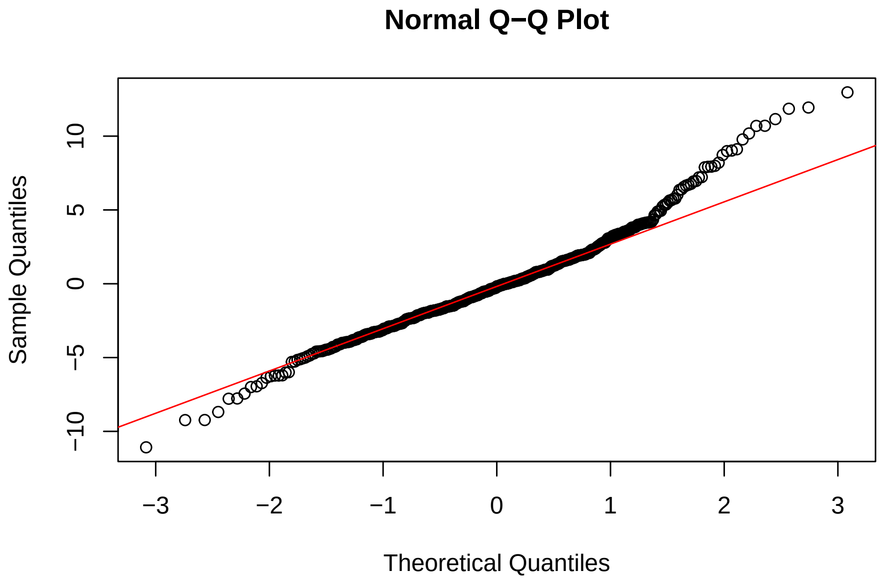
Since the lambda variable is optimized *right* next to 0, we'll try a log transformation for this variable. The others that were closer to .5, we tried a square-root transformation instead. The BOXCOX plot is really a guide to help us figure out where to look, but the end decision we made by looking at how our residuals looked once we've made these transformations.



The above fitted residuals look much better. There's very little curvature in the residuals nor does there appear to be any stark increase or decrease in the variance of the residuals. One thing to note is that if our goals were to maximize the accuracy of predictions, we might care less about the interpretability of our predictor and response variables and therefore we might transform our response variable as well to see how it affects performance. Since we do care about interpretability, this is where we'll stop our transformations and move on.

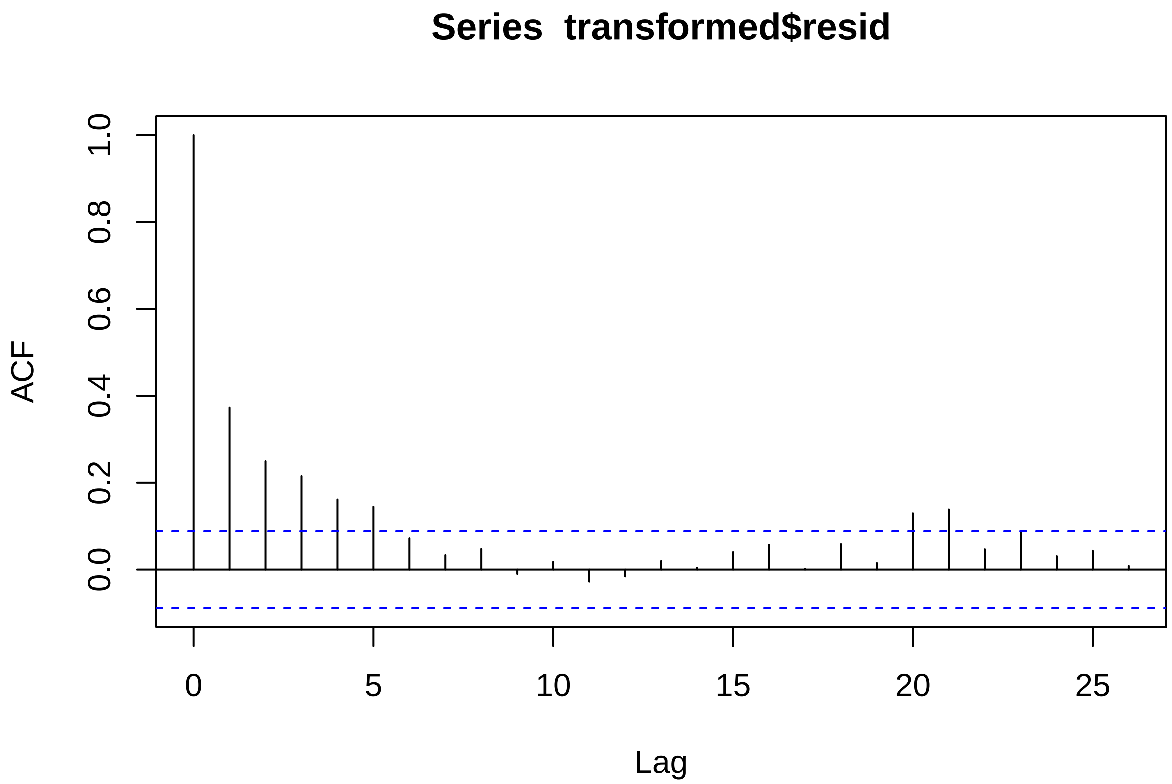
Two more assumptions to check for

First, let's check the QQ norm plot to see how well our residuals follow the assumed normal distribution.



There is room for improvement, but since this particular assumption is not the most important for us to bend to and the samples fit close enough to our expectations, we'll move on to checking whether our residuals are independently related to each other by checking the ACF plot below.

ACF Plot of our Model Post Transformations



Interestingly, there does appear to be a slight relationship in one residual from the next. A pattern appears to emerge where one observation seems oddly close to the next one in line. What this might tell us is that there is a relationship between the observations in one township and the neighboring areas. This makes sense logically, as one expensive town is more likely to be neighbored by another expensive town rather than completely random placements of high and low value towns sporadically throughout Boston.

There is little we can do about this relationship and the data we have thus far, so while our assumption is not met to 100% our satisfaction, we feel comfortable moving forward with the data we have so far. We will be sure to note this discrepancy in our final conclusions.

We ran both forwards and backwards predictor selections on our model. Both selection processes selected the same variables. The variables removed are chas, zn, indus, and sqrt(age).

```
##      368
## 3.989351
```

By using Bonferroni method for outlier detection, we find that observation 368 in our dataset is an outlier.

```
##        157        366        368        492        493
## 0.07941124 0.11826110 0.09374476 0.07941509 0.07759307
```

We identify the high leverage observations, which are the data points that are the most influencial. We ran leverages analysis, Cook's distance, and DFFITs to find the most influcencial data point. The outlier observation 368 was repeatedly included as an influcencial observation, however, Cook's distance resulted in no values. Though observation 368 is of concern, the analysis indicates that there is only one true outlier.

```
## named numeric(0)
```

What does Our Outlier Look Like? Upon closer inspection of this outlier point, we don't observe any notable reason to drop this neighborhood from our dataset. Our final coefficient of interest (`black_neighborhood`) is largely unaffected by the addition of this point, and since the model's performance is still performing better than our benchmark, we'll include it in our final model as well.

```
##                   town tract      lon    lat medv cmedv    crim zn indus chas
## 384 Boston East Boston   502 -71.0215 42.227 12.3  12.3 7.99248  0  18.1    0
##     nox   rm age    dis rad tax ptratio     b lstat black_neighborhood
## 384 0.7 5.52 100 1.5331  24 666    20.2 396.9 24.56                  0
```

Final MLR Model Coefficients and Performance

```
## 
## Call:
## lm(formula = cmedv ~ log(crim) + nox + rm + sqrt(dis) + rad + 
##     tax + ptratio + log(lstat) + black_neighborhood, data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -10.4582  -2.1354  -0.2416   1.6838  13.0511 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          53.64213    4.33311  12.380  < 2e-16 ***
## log(crim)            -0.21378    0.19892  -1.075    0.283    
## nox                 -12.95793    2.84647  -4.552 6.73e-06 ***
## rm                    2.89755    0.33290   8.704  < 2e-16 ***
## sqrt(dis)            -4.12189    0.55042  -7.489 3.36e-13 ***
## rad                   0.22788    0.05427   4.199 3.20e-05 ***
## tax                  -0.01361    0.00246  -5.533 5.18e-08 ***
## ptratio              -0.80182    0.09098  -8.813  < 2e-16 ***
## log(lstat)           -7.00753    0.48013 -14.595  < 2e-16 ***
## black_neighborhood1  -3.04513    0.67253  -4.528 7.53e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.49 on 480 degrees of freedom
## Multiple R-squared:  0.8059, Adjusted R-squared:  0.8022 
## F-statistic: 221.4 on 9 and 480 DF,  p-value: < 2.2e-16
```

Above we can see that the coefficient for `black_neighborhood` is - 3.4, which equates to a drop in expected median housing value by about $3,400 if the neighborhood is considered black, based on our assumptions and in the presence of the other predictors we've included into our final model.

Confindence Interval for `black_neighborhood`

```
##     2.5 %    97.5 % 
## -4.366596 -1.723656
```

We can be 95% confident that the true impact of a neighborhood being a black neighborhood is between -$1,724 and -$4,366. This is a good step towards providing evidence of racial disparities in housing values, but one additional step we can take is to see if we can predict the type of neighborhood a tract is (black or non) based on the median value or if other predictors account for more value in the next model.

Part 4: Logistic Regression

We wanted to see if a logistic model could be useful in predicting whether a neighborhood was a black neighborhood.

Similar to the multiple regression model, we will run a basics logistic model with all the predictors included. This will give us a sense of our benchmark, as well as which predictors could be useful. We will also split the data set into two: 75% for training and 25% for testing. We will use the training data frame to create a model and the testing data frame to measure it’s performance.

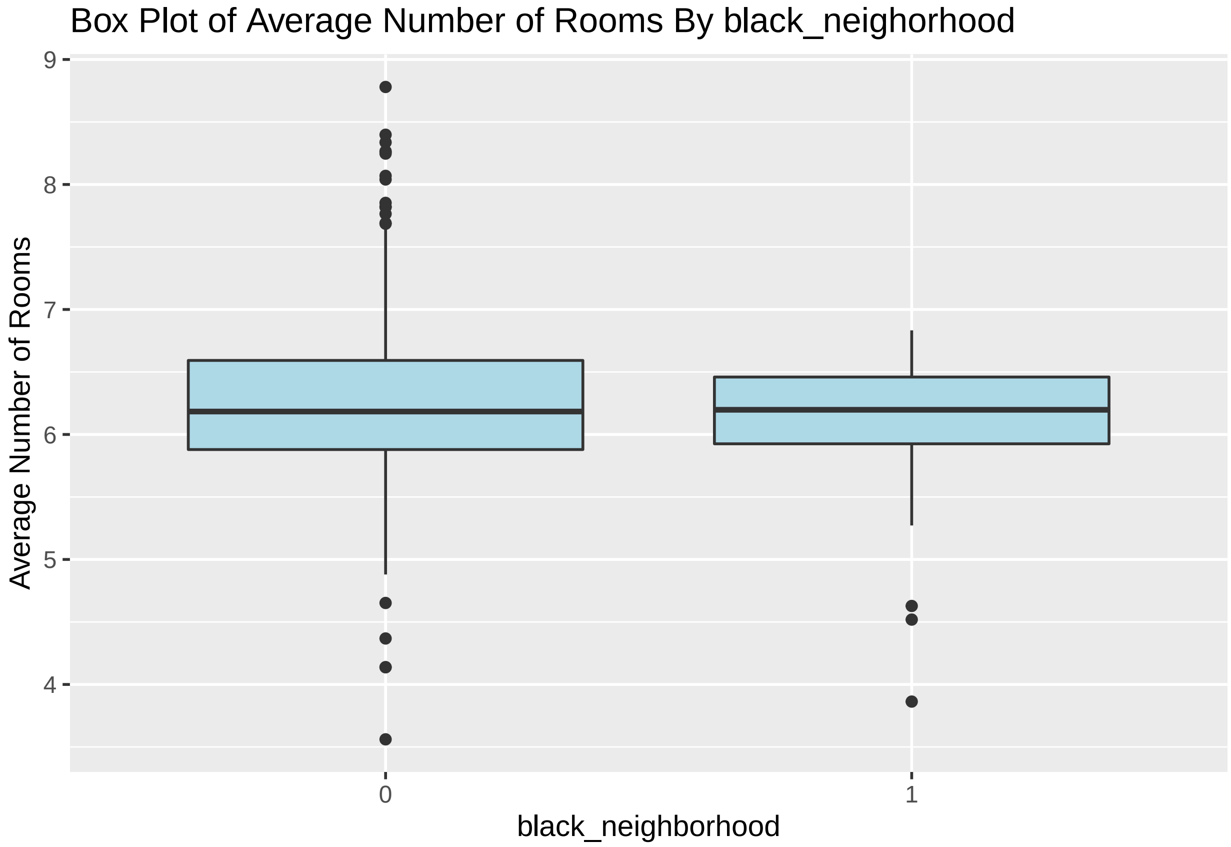
Benchmark Model Summary

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Call:
## glm(formula = black_neighborhood ~ cmedv + zn + indus + chas +
##     nox + rm + age + dis + rad + tax + ptratio + lstat, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4149  -0.2329  -0.1133   0.0000   3.1416
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.132933   8.596957   0.015  0.98766
## cmedv       -0.221945   0.081909  -2.710  0.00674 **
## zn          -1.036152  87.944079  -0.012  0.99060
## indus        0.046073   0.148600   0.310  0.75653
## chas1        0.001174   1.437339   0.001  0.99935
## nox         -1.037090   4.755350  -0.218  0.82736
## rm           0.993528   0.503762   1.972  0.04858 *
## age         -0.024885   0.023927  -1.040  0.29832
## dis         -0.063927   0.545764  -0.117  0.90675
## rad          0.130773   0.111379   1.174  0.24035
## tax          0.001001   0.008228   0.122  0.90320
## ptratio     -0.276180   0.366694  -0.753  0.45135
## lstat       -0.011128   0.067258  -0.165  0.86859
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 192.88  on 366  degrees of freedom
## Residual deviance: 119.56  on 354  degrees of freedom
## AIC: 145.56
```

```
##
## Number of Fisher Scoring iterations: 20
```

It appears that all of the variables except cmedv (median home value) and rm (average number of rooms) are not significant. Since we've already confirmed a relationship between average home value and black neighborhood in Part 1, it's not surprising to see a similar relationship in our logistic model. Let's chart the variables for average number of rooms and black neighborhood to see if there's a relationship.



It seems like the average number of rooms is similar between black neighborhoods and non-black neighborhoods at around 6.25. However, the variance for the average seems to be much higher for non-black neighborhoods.There also seems to be a large cluster of outliers above 7.5 rooms for non-black neighborhoods.

Now, we run a logistic regression on the test data using only the variables for average home value and average number of rooms.

```
##
## Call:
## glm(formula = black_neighborhood ~ cmedv + rm, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4926  -0.2859  -0.1914  -0.1087   3.1807
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.45286    2.56601  -2.515 0.011912 *
## cmedv       -0.30283    0.04947  -6.122 9.24e-10 ***
## rm           1.47395    0.44733   3.295 0.000984 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 192.88  on 366  degrees of freedom
## Residual deviance: 135.18  on 364  degrees of freedom
## AIC: 141.18
##
## Number of Fisher Scoring iterations: 7
```

The reduced model looks much better since the p-values for both coefficients are close to zero. It’s interesting to see that the coefficient for rm is positive. This indicates that holding the average home value constant, increasing the average number of rooms increases the probability that the neighborhood is black.

The coefficient for the average home value is negative. This indicates that holding the average number of rooms constant, increasing the average home value decreases the probability that the neighborhood is black.

We will now perform the following hypothesis test to see whether we should go with the full model or reduced model:

H_o : *Betas for all coefficients except for cmedv and rm are equal to zero*

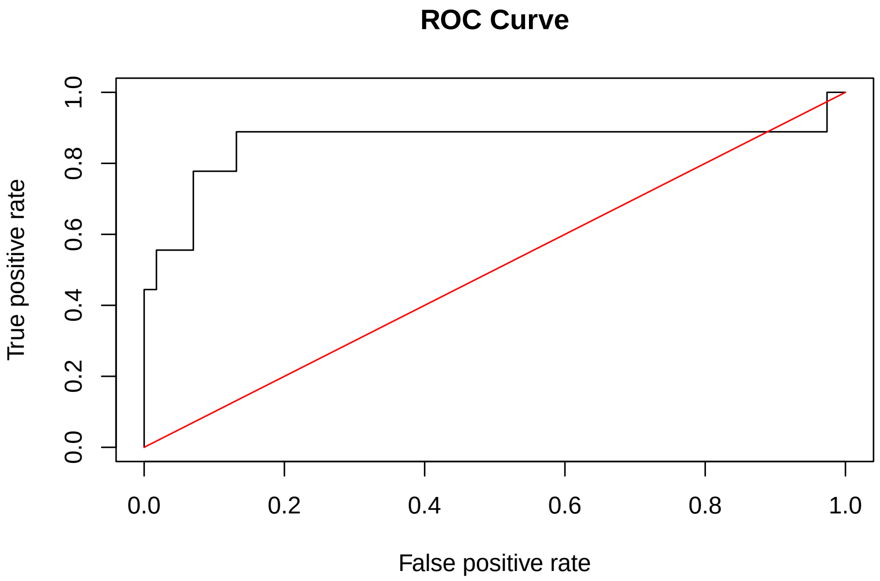
H_a : *At least one of the coefficients in H0 is not zero*

```
## [1] 15.62019
```

```
## [1] 0.1110332
```

The p-value of 0.111 is higher than our alpha of 0.05. Therefore, we fail to reject the null and decide to go with the reduced model.

Let’s now see how our reduced logistic model with two variables perform against the testing data frame. We’ll also perform the following hypothesis test to see whether we should go with the full model or reduced model.



```
## [[1]]
## [1] 0.8596491
```

We'll plot the ROC curve and then calculate the AUC.

The ROC curve is well above the diagonal line except for at the very end. It looks like the logistic regression performs much better than random guessing.

This is also confirmed by a very strong AUC of 0.8596. Since the AUC is well above 0.5, we can conclude that the model does better than random guessing and does have predictive value.

Part 5: Conclusions

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.