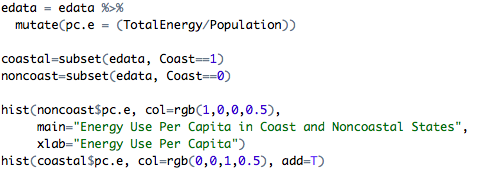
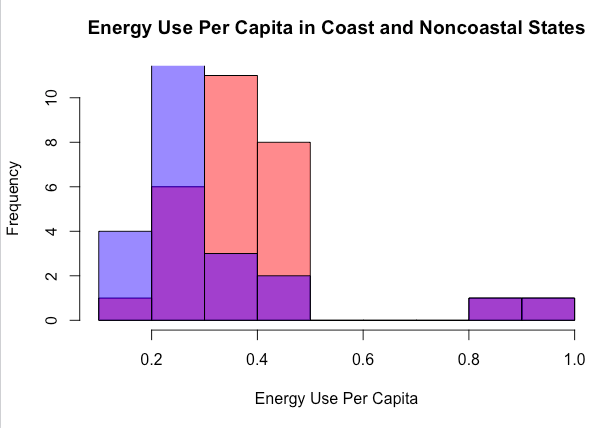
1. Does ***per capita*** energy consumption differ depending on whether a state is found on the coast or not?
2. Please write the null and alternate hypothesis (1 point).

Null: There is no difference in per capita energy consumption between states on the coast and states not on the coast. Alternative: There is a difference in per capita energy consumption between states on the coast and states not on the coast.

b. Please create a visual plot to answer this question (1 point).

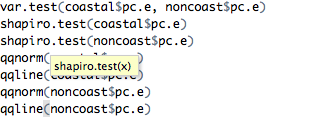
In this histogram the blue represents the coastal states, the red represents non-coastal states and purple represents areas of overlap





c. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

A two sample t-test. It should be a Welch’s t-test since the sample sizes are not equal. I can only assume that the observations are independent since each state is different. This data is not exactly just a random sample since it contains all of the states. I tested for equal variance and the p-value was 0.5, so variances are equal. However, Shapiro Wilk tests revealed that the data was not normally distributed (this appears to be due to states with low populations like Wyoming, Alaska, and North Dakota, which use more energy per capita than other states). I may still be able to run a t test because I am not working with a sample population- all of the states are represented; the assumption of normality relies on the idea that I am taking a sample of a population, and using that sample to try to estimate the true mean.



d. Please run the statistical test and interpret the result (1 point).

The p-value for the t test is 0.22 so I fail to reject my null hypothesis, suggesting that there is not a difference between the energy consumption per capita between coastal and non-coastal states.

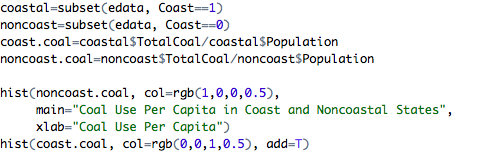
../Screen%20Shot%202017-03-28%20at%2011.08.29%20AM.png

2. Does ***per capita*** coal consumption differ depending on whether a state is found on the coast or not?

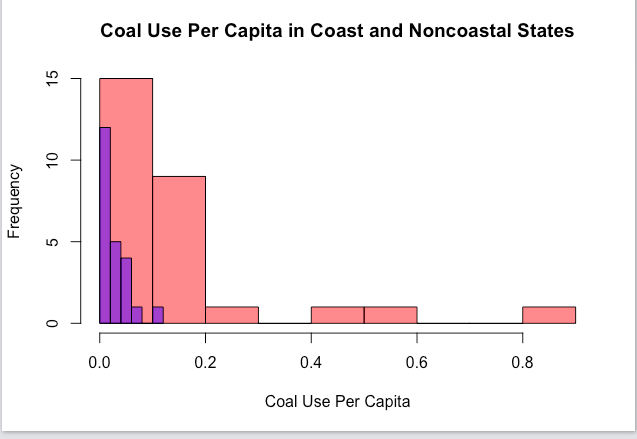
a. Please write the null and alternate hypothesis (1 point).

Null: Per capita coal consumption does not vary between coastal and non-coastal states. Alternative: Per capita coal consumption is greater in non-coastal states than non-coastal states.

b. Please create a visual plot to answer this question (1 point).

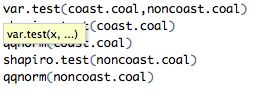


Non-coastal coal consumption per capita is in red/pink and the overlap between coastal and non-coastal coal consumption per capita is in purple.



c. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

I would like to run a two sample Welch’s t-test, but based on my histogram I am suspicious about whether or not this data meets the appropriate assumptions. A variance test yields a p value dramatically below 0.05, suggesting that it is highly unlikely that the two samples have equal variance (but the Welch’s t test does not assume equal variance anyway so that’s okay). Shapiro Wilk tests reveal p value of less than 0.05 so I cannot assume the data is normally distributed. However since the number of states I’m working with is greater than 40, I can ignore that assumption.



1. Please run the statistical test and interpret the result (1 point).

A t test yields a p value of 0.0019, so I reject my null hypothesis.

../Screen%20Shot%202017-03-28%20at%2011.13.14%20AM.png

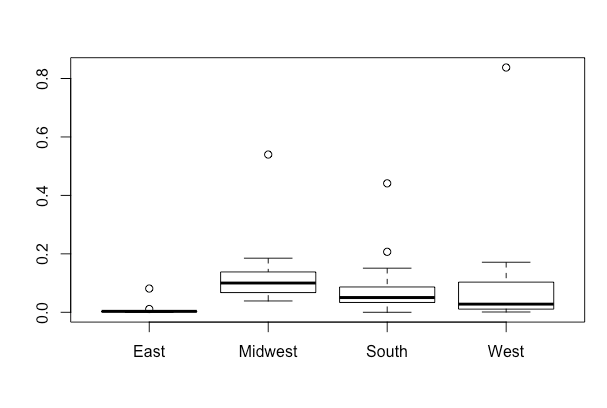
3. Does ***per capita*** coal consumption differ depending on the region in which a state is found?

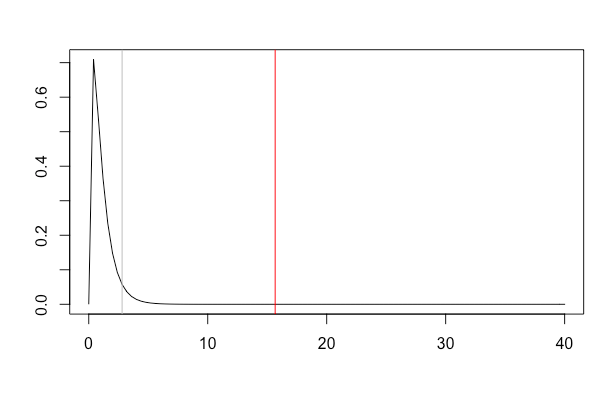
* 1. a. Please write the null and alternate hypothesis (1 point).

Null: Per capita coal consumption does not vary depending on what region a state is located in. Alternative: Per capita coal consumption is different depending on what region a state is located in.

b. Please create a visual plot to answer this question (1 point).

../Screen%20Shot%202017-03-28%20at%2011.14.38%20AM.png





c. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

A one-way anova. Observations should be reasonably independent (although you could argue that geographical considerations impact coal use per capita). Next I need to test for equal variance. A levene test yields a p-value of 0.52, so I will proceed assuming equal variance. Shapiro tests reveal that the data is not normally distributed.

d. Please run the statistical test and interpret the result (1 point).

../Screen%20Shot%202017-03-28%20at%206.57.16%20PM.png

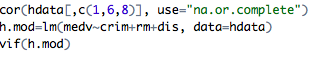
The p-value is 0.262 suggesting no significant difference between the groups, and I should accept my null hypothesis.

4. What is the correlation between ***per capita*** coal use and ***per capita*** GDP? Does this seem like a strong correlation to you? Why or why not? (2 points)

../Screen%20Shot%202017-03-28%20at%206.59.47%20PM.png

The correlation is 0.035 which does not seem particularly strong given that 0 is no correlation and 1.0 is perfect correlation, and 0.03 is much closer to zero.

5. Please select three covariates that you will include in your model as independent variables. Please check if these variables are highly correlated with one another to make sure you do not run into problems of multi-collinearity. Check if this model has issues with multi-collinearity using the variance inflation factor. **Report correlation values and VIF values in your answer** (3 points).



Variables: Crim, RM, DIS (Per capita crime rate, number of rooms, and weighted distance to Boston’s five employment center).

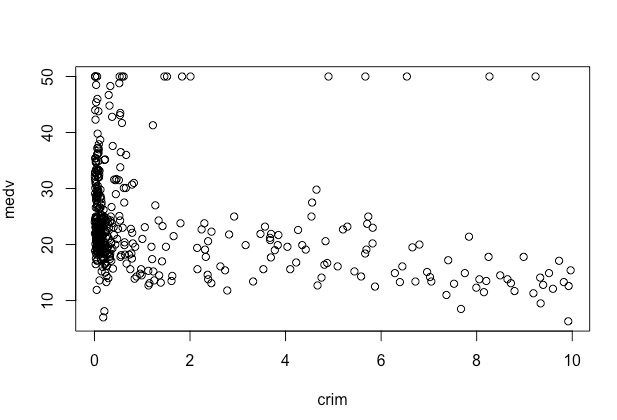
The correlation value for crime and rooms is -0.14, the correlation value for crime and distance is -0.46, and the correlation value for distance and rooms is 0.14

The VIF for crime is 1.28, for rooms it is 1.03, and for distance it is 1.28. All of these values are well below 10 or other thresholds that would indicate multicollinearity.

6. Plot the relationship between each of your three independent variables and the dependent variable (medv). **Include each plot in this answer and state whether and how you think each variable is related to median housing prices** (medv; 3 points).

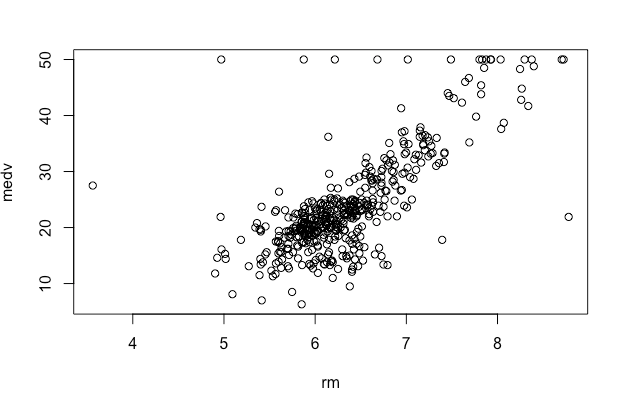
../Screen%20Shot%202017-03-28%20at%207.03.43%20PM.png

Crime & Housing Value



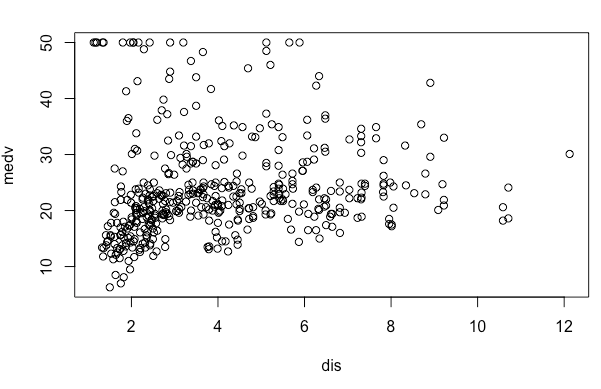
It looks like low crime rates are associated with most higher housing prices (although there are also low housing prices in low crime areas), but as crime increases, housing prices mostly level off with a few high priced houses as various higher crime rates. Overall the correlation appears negative, and it might not be linear since it drops off so sharply then continues in an almost level way.

Number of Rooms & Housing Value



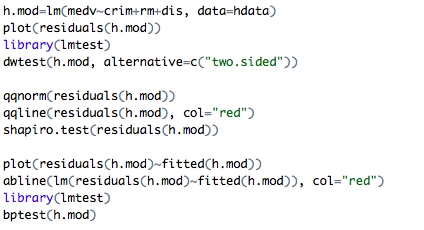
This relationship appears to be a positive correlation, though not all values reflect that. For the most part as the number of rooms increases value increases.

Weighted Distance from Boston’s Five Employment Centers and Housing Value



The relationship between these two variables appears to be almost no correlation. Most of the lowest valued homes do appear to have the lowest weighted distance to the employment centers, but there are also medium and high valued homes near the employment centers, and low valued homes further from the employment centers.

7. Run your multiple linear regression model. Check whether any assumptions are violated. Please state **which assumptions** you checked, **whether they were violated**, and **how you know** whether or not they were violated. If any assumptions are violated (e.g. normality), we will give you bonus points if you are able to identify a way to overcome this problem (3 points, plus additional 1 point bonus).



I first plotted the residuals and there did appear to be a pattern- they were mostly clustered in a line with a few outliers particularly near the left side. Next, I ran a Durbin-Watson test which yielded an extremely small p-value (<2.2e-16) suggesting that the residuals are not independent and that there is autocorrelation.

Next I tested to see if the residuals are normally distributed. A qqnorm plot revealed that they did not appear normal (largely due to values to the left of the plot which went above the line), and a Shapiro Wilk test confirmed that the errors were not normally distributed. This suggests that a linear model may not be the best fit but perhaps transforming the data would help.

Next I tested for homoscedasticity. Again, I initially created a plot. There the residuals were highly clustered indicating that there was not a constant variance of the errors. A bp test led to a p value of 4.568e-07 indicating that there is a pattern to the errors like I saw.

The fact that the residuals are not normally distributed and not independent suggests that the relationship may not be linear.

8. Interpret the results of the linear regression model. State **what the coefficient and its significance means** for the intercept and each of your three independent variables. Please explain what each regression coefficient means and do not just state that the coefficient is significant or not significant. For 1 bonus point, add in an interaction term, rerun the model, and interpret the result (3 points plus additional 1 point bonus).

The coefficient for the intercept is -34.39 and it is significant, so the intercept is different from zero. It is what the value of a house would be if all of the variables were set to zero, though this doesn’t make real logical sense, since a house with zero rooms and zero weighted to distance to Boston’s five employment centers is not a house. It is simply the value where the line meets the intercept.

The coefficient for crime is -0.74 and it is significant, so as crime increases house values decrease by -0.74.

The coefficient for rooms is 9.48 and it is also significant. As the number of rooms increases the value of a house also increases by 9.48.

The coefficient for distance is not significant so distance has no effect on house value.

I added an interaction term for the interaction between distance and number of rooms. In this model, again the intercept is significantly different from zero although it is -16.8 now while it was -34 in the previous model.

Crime still has significant a negative impact on housing values (coefficient -0.77).

Rooms still have a positive impact on house value. In this model the coefficient is 6.73 so as rooms increase house value increases by a factor of 6.73.

In this model distance has a significant coefficient of -5.24 so as weighted distance to Boston’s five employment centers increases housing value decreases by a factor of -5.24.

The interaction between rooms and distance has a significant positive coefficient of 0.78 so as there are more rooms at greater distance from employment centers, value increases.

9. Discuss the fit of your model and whether you think it is a good or bad fit. Why (2 points)?

The multiple r squared is 0.58 which indicates that 58 percent of the variation in house values can be explained by this model; however, since the data failed to meet assumptions, this model needs improvement. Since the data is heteroscedastic the p values pay be inaccurate. So while they appeared to be significant, I actually cannot say that for sure. Since errors are neither normally distributed nor independent there is a good chance that a linear model is not actually appropriate. I think there are problems due to the fact that the independent variables are on dramatically different scales; perhaps standardizing the data in some way (such as around the mean) would have been helpful.