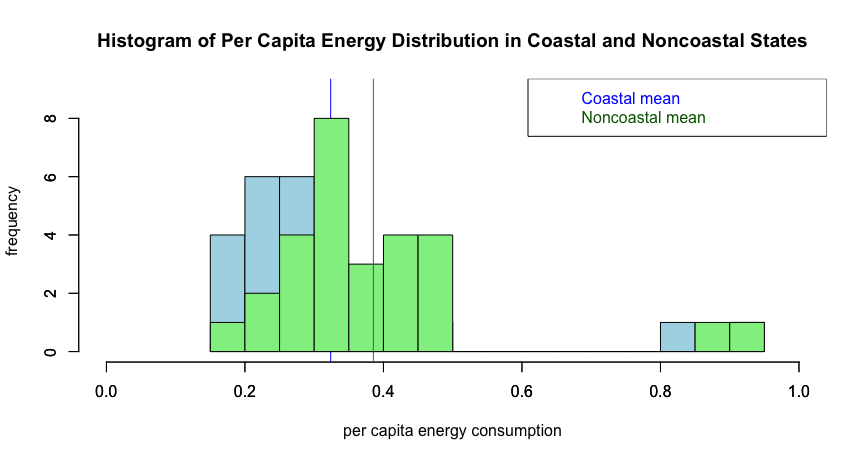
If you run any ANOVAs, you can use the Levene test for equality of variances. If your data violate an assumption about normality and a normal distribution is required for your analyses, you can get bonus points for transforming your data. Otherwise please run the statistical test anyway as if your data were normally distributed but make it clear that you violated this assumption in your answer.

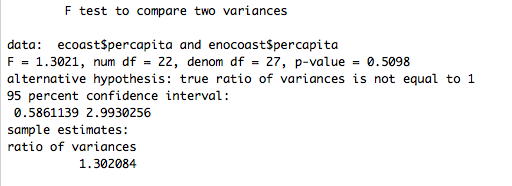
Please use the R script provided to load data and build your script from there.

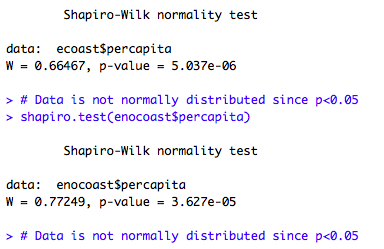
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized by whether they are in the South, West, Midwest, or East of the country (Region) or on the coast (Coast, 0 = no; 1 = yes). Depending on the questions below, you may need to construct your own variable that is a combination of the variables included in the dataset (e.g. when per capita is used). 14 points total.

1. Does ***per capita*** energy consumption differ depending on whether a state is found on the coast or not?
   1. Please write the null and alternate hypothesis (1 point).
      1. Ho: Mean per capita energy consumption in coastal states is not significantly different than mean per capita energy consumption in noncoastal states.
      2. Ha: Mean per capita energy consumption in coastal states is not equal to mean per capita energy consumption in noncoastal states.
   2. Please create a visual plot to answer this question (1 point).

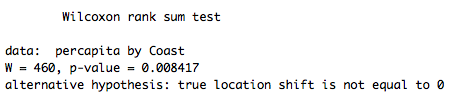


* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
     1. I would run a two tail, two sample t test. My first assumption is that the samples have equal variances. I ran a var.test to confirm this and since p>0.05, the true ratio of variances is equal to 1, meaning the variances are equal. A second assumption is that each observation is sampled independently, which we assume they are. A third assumption is that the samples are normally distributed. I ran a shapiro test for both coastal and noncoastal states and since the p values for both were <0.05, we conclude that the samples are not normally distributed. Further, we don’t have a large enough sample size for either sample (n<30 in both), so we can’t assume that they follow the Central Limit Theorem.

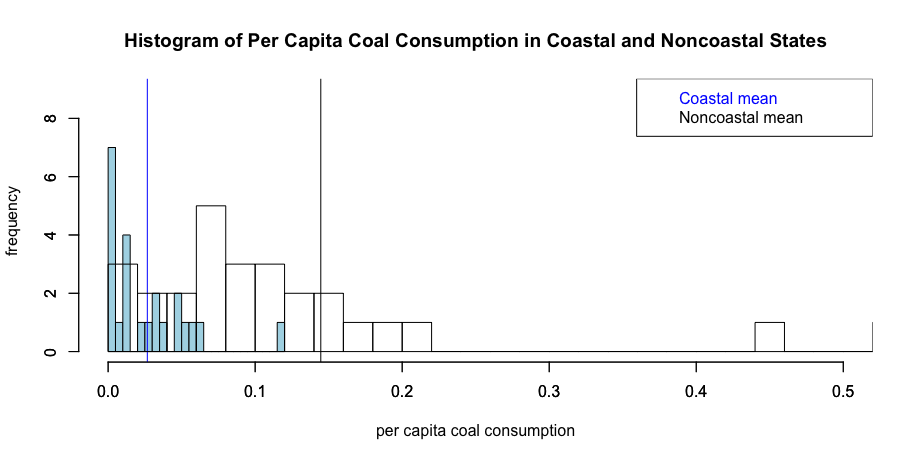




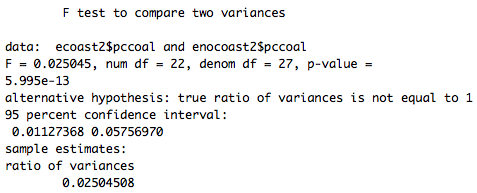
* 1. Please run the statistical test and interpret the result (1 point).
     1. Since our data is not normally distributed, I ran a Mann-Whitney-Wilcoxon Test to compare per capita energy consumption in coastal and noncoastal states. Since p<0.05, we reject the null hypothesis and conclude that the mean per capita energy consumption in coastal v. noncoastal states is significantly different.

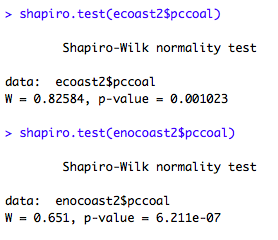


1. Does ***per capita*** coal consumption differ depending on whether a state is found on the coast or not?
   1. Please write the null and alternate hypothesis (1 point).
      1. Ho: Mean per capita coal consumption in coastal states is not significantly different than mean per capita coal consumption in noncoastal states.
      2. Ha: Mean per capita coal consumption in coastal states is not equal to mean per capita coal consumption in noncoastal states.
   2. Please create a visual plot to answer this question (1 point).

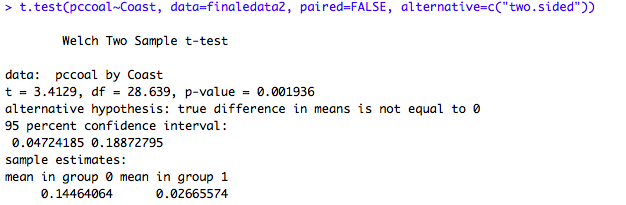


* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
     1. I would run a two tail, two sample t test. My assumptions are the same as in Question 1: samples have equal variances, samples are normally distributed, and observations are independent of each other. The var.test p value <0.05, so our data violates the assumption of equal variances. The shapiro test p values for both datasets are <0.05 and we have less than 30 data points in each, which means our data also violates the assumption of normality.

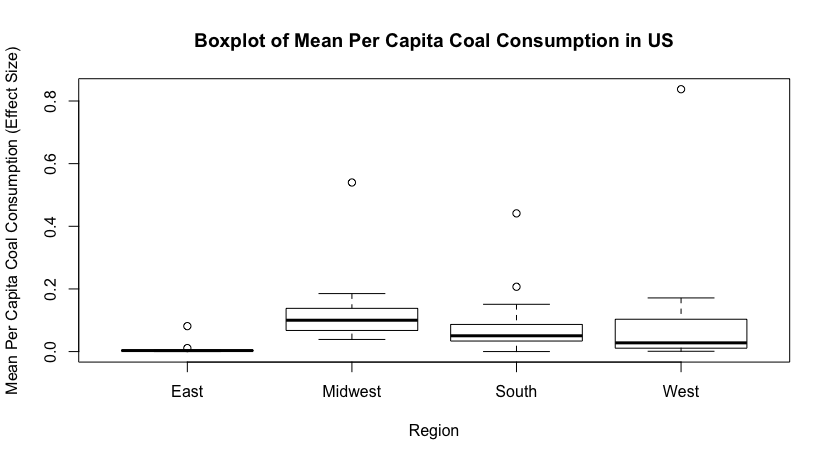




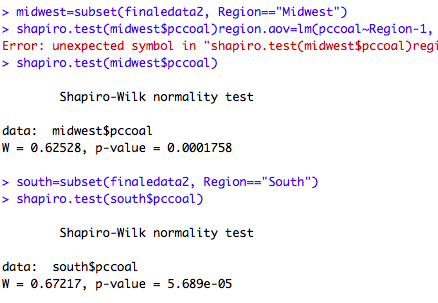
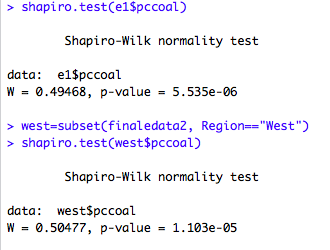
* 1. Please run the statistical test and interpret the result (1 point).
     1. Although our assumptions of normality and equal variance were violated, I proceeded with running the t test. Since p<0.05, we reject the null hypothesis and conclude that per capita coal consumption in coastal states is significantly different than per capita coal consumption in noncoastal states.

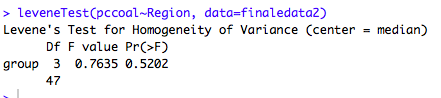


1. Does ***per capita*** coal consumption differ depending on the region in which a state is found? Pccoal = DV, continuous; Region = IV, categorical (4 levels…E, W, MW, S)
   1. Please write the null and alternate hypothesis (1 point).
      1. Ho: There is no regional difference in per capita coal consumption amongst states in the four regions of the US.
      2. Ha: There is a regional difference in per capita coal consumption amongst states in the four regions of the US.
   2. Please create a visual plot to answer this question (1 point).



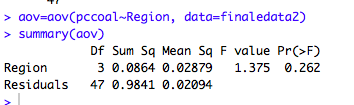
* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
     1. I would run an ANOVA. We make three assumptions about an ANOVA. First, we assume that the populations of interest are normally distributed. In running shapiro tests on each set of regional data, however, none of the p values were greater than 0.05, meaning that we reject the null hypothesis that the populations are normally distributed. Second, we assume that the samples are independent of one another. Third, we assume that each population has the same variance. To test this, I ran a Levene test and since p>0.05, we fail to reject the null hypothesis that the variances of the populations are equal. Therefore, we only violate one assumption for the ANOVA.



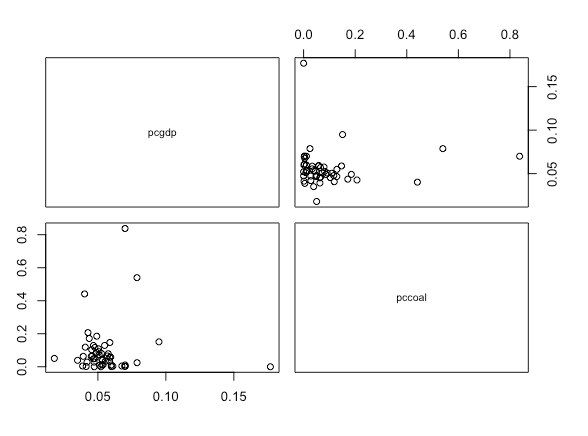


* 1. Please run the statistical test and interpret the result (1 point).

Since p>0.05 (p=0.262), we fail to reject our null hypothesis that there is no regional difference in per capita coal consumption amongst states in the four regions of the US.



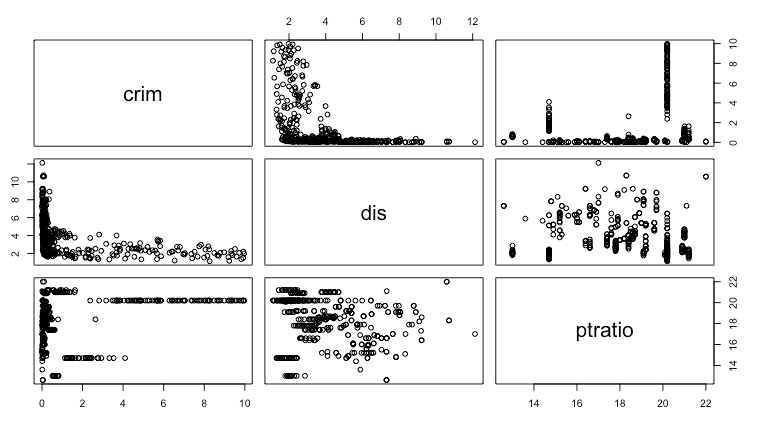
1. What is the correlation between ***per capita*** coal use and ***per capita*** GDP? Does this seem like a strong correlation to you? Why or why not? (2 points)
   1. The correlation between per capita coal consumption and per capita GDP is 0.035, which means that the two variables are not correlated. In order for there to be a correlation, the correlation coefficient should be greater than 0.5 (+/-), as -1 and 1 indicate high levels of correlation while 0 indicates no correlation. A plot of the relationship between the two variables graphically demonstrate the lack of correlation.

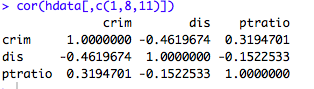




For questions 5-9, please use the ‘housedata.csv’ dataset that shows housing information for the Boston area. Information on what each of the variables are can be found here: <http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names>. In this exercise, the goal is to create a multiple linear regression model to predict housing value prices (medv). Please do not use an interaction term (unless stated in the question) since they can be challenging to interpret! 14 points + 2 bonus points.

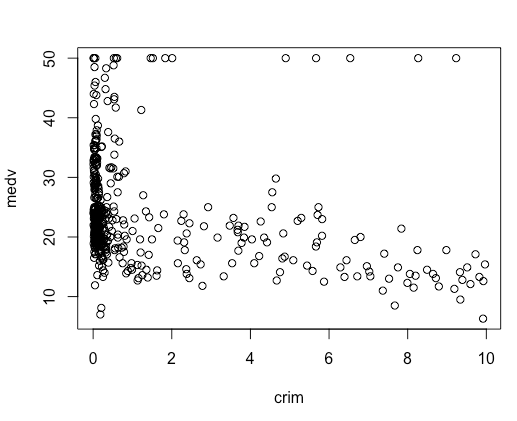
1. Please select three covariates that you will include in your model as independent variables. Please check if these variables are highly correlated with one another to make sure you do not run into problems of multi-collinearity. Check if this model has issues with multi-collinearity using the variance inflation factor. **Report correlation values and VIF values in your answer** (3 points).
   1. My independent/explanatory variables are: weighted distances to five Boston employment centers (DIS), per capita crime rate (CRIM), and pupil-teacher ratio (PTRATIO). While none of the covariates are highly correlated, per capita crime rate and index of accessibility to radial highways are at least moderately correlated (r = -0.46). However, the VIF for is 1.305, which indicates that there are no issues with collinearity.



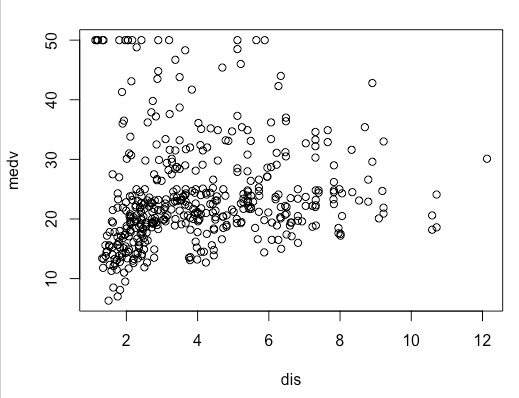




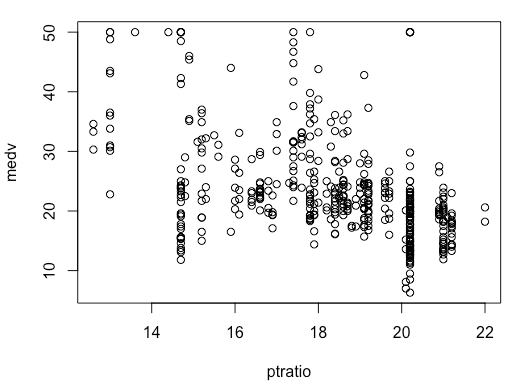
1. Plot the relationship between each of your three independent variables and the dependent variable (medv). **Include each plot in this answer and state whether and how you think each variable is related to median housing prices** (medv; 3 points).
2. Based on the plots below, none of my independent variables appear to have a linear relationship with MEDV. As such, it is difficult to say anything about the effect of each variable on median housing prices in Boston.
   1. Relationship between housing value prices and per capita crime rate

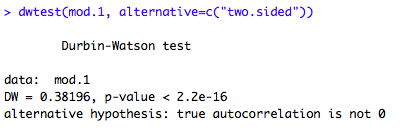


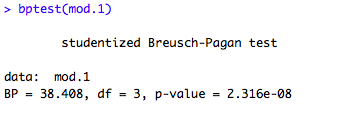
* 1. Relationship between housing value prices and distance to five Boston employment centers



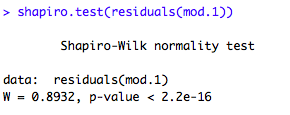
* 1. Relationship between housing value prices and pupil:teacher ratio



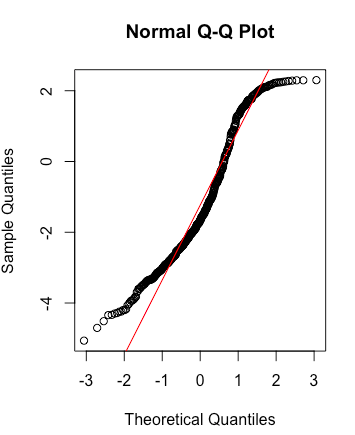
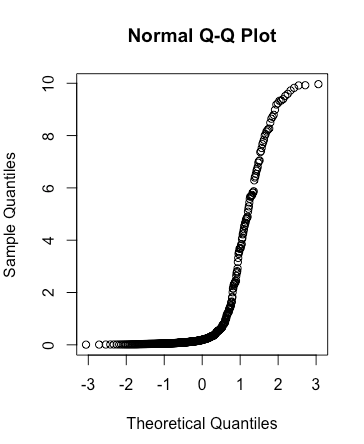
1. Run your multiple linear regression model. Check whether any assumptions are violated. Please state **which assumptions** you checked, **whether they were violated**, and **how you know** whether or not they were violated. If any assumptions are violated (e.g. normality), we will give you bonus points if you are able to identify a way to overcome this problem (3 points, plus additional 1 point bonus).
   1. My first assumption is that the residuals are independent of each other. I ran a Durbin-Watson test to check this and since my p value <0.05, I reject the null hypothesis that true autocorrelation is 0 and conclude that my residuals are not independent. 
   2. My second assumption is that there is a constant variance of the residuals (homoscedasticity). I used bptest to check this and since my p value < 0.05, I can conclude that there is heteroscedasticity of my residuals.



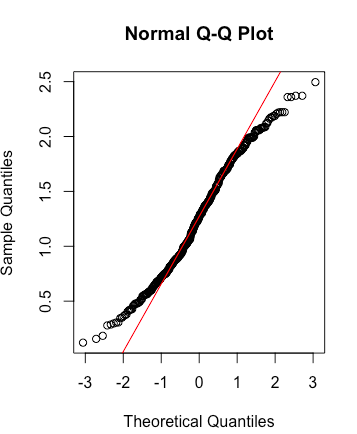
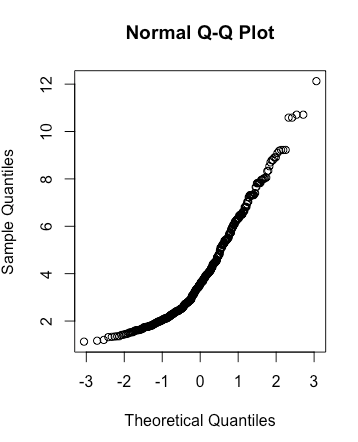
* 1. My third assumption is that the residuals are normally distributed. I used a shapiro test to check this and since p<0.05, I conclude that my residuals are not normally distributed.



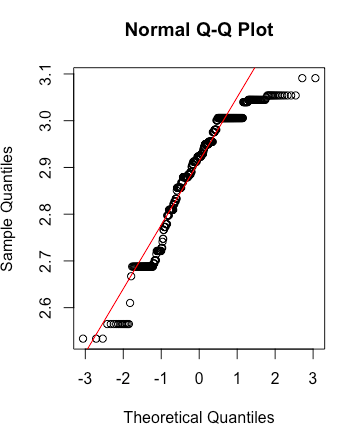
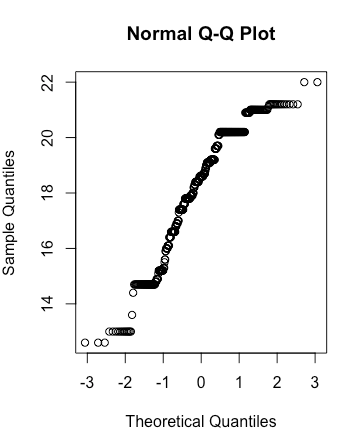
* 1. My final assumption is that there is a linear relationship between the variables. As my plots in Question 5 indicate, there does not appear to be a linear relationship between any of my independent variables and the dependent variable. I did qqnorm plots to visualize each predictor variable to see how I might transform the data to make it better fit our assumptions. Both the CRIM and DIS variables appear as though they could be exponential distributions, so I log transformed them to see if they became more normal.
     1. CRIM: Pre log transformation (Left); Post log transformation (Right)



* + 1. DIS: Pre log transformation (Left); Post log transformation (Right)



* + 1. I also log transformed the PTRATIO variable, although the qqplot didn’t appear to be an exponential distribution. The log transformation did not make the data more normal.



1. 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).

Even though none of my assumptions were met, I ran a multiple linear regression. Intercept estimate (54.49) is the estimated housing value price when all of the independent variables equal 0. Since p<0.05, we know this is significantly different than 0.

Crim estimate (-0.53635) means that for every unit increase in per capita crime rate, housing value price decreases by 0.536 units (when controlling for pupil:teacher ratio and distance to employment centers). Since p<0.05, we can conclude that per capita crime has a significant effect on housing value price.

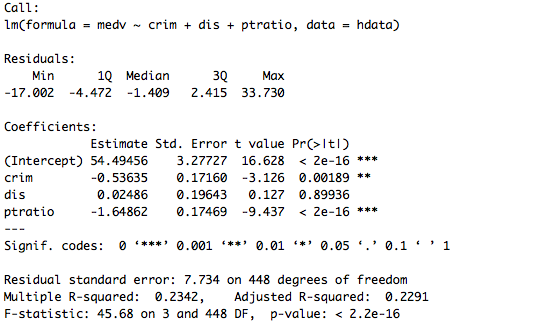
Dis estimate (0.02486) means that for every unit increase in distance to five Boston employment centers, housing value price increases by 0.024 units (when controlling for per capita crime rate and pupil:teacher ratio). However, since p>0.05, accessibility to radial highway does not have a significant effect on housing value prices.

Ptratio estimate (-1.64862) means that for every unit increase in pupil:teacher

ratio, housing value price decreases by 1.64 units (when controlling for per capita

crime rate and distance to employment centers). Since p<0.05, ptratio has a

significant effect on housing value price.



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

Although my f statistic p value <0.05 (and therefore including my IVs in the model explains variability in home value significantly better than the null model), my model is not a great fit. The adjusted R2 value is 0.2291, meaning that my model only explains 22.91% of the variability in housing value prices. This means that there are other factors that should be included in my model or other models that could do a better job of predicting housing value price. Additionally, my data did not meet the assumptions required to run a linear regression.