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

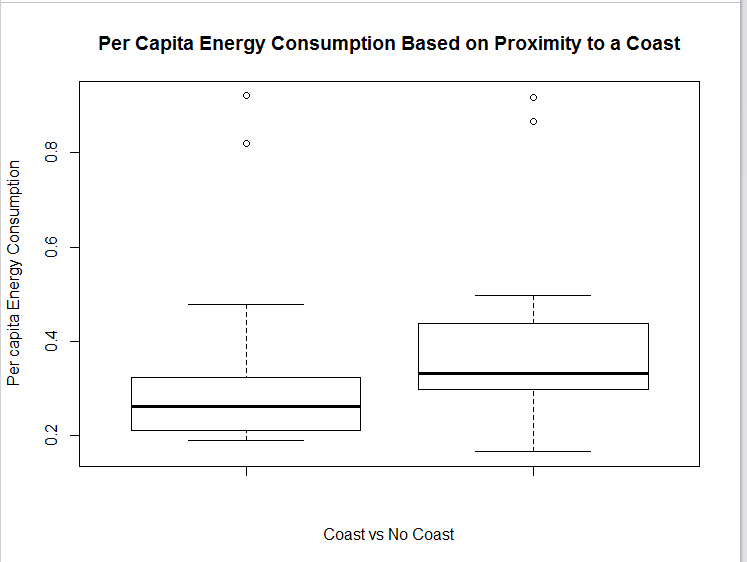
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized 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).

Null: Per capita energy consumption does not differ depending on whether a state is found on the coast or not.

Alternate: Per capita energy consumption does differ depending on whether a state is found on the coast or not.

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



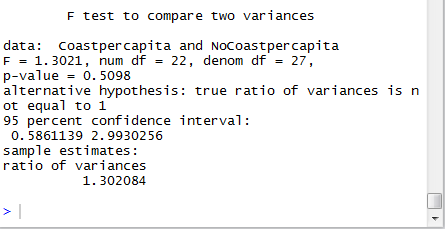
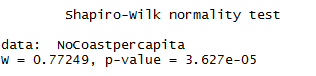
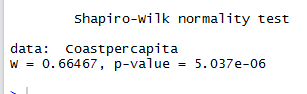
* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

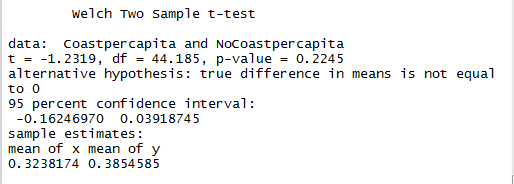
Independent: Coast (categorical)

Dependent: per capita energy consumption (continuous)

**Test to use: Unpaired 2-sample t-test**

Assumptions:

* Samples have equal variance 🡪 var.test()
  + 
* Samples are normally distributed 🡪 Shapiro Test
  + ’
  + 
* Each observation is sampled independently (for this, I am assuming it is met since each household does not likely influence the others)
  + **All assumptions are met**
  1. Please run the statistical test and interpret the result (1 point).



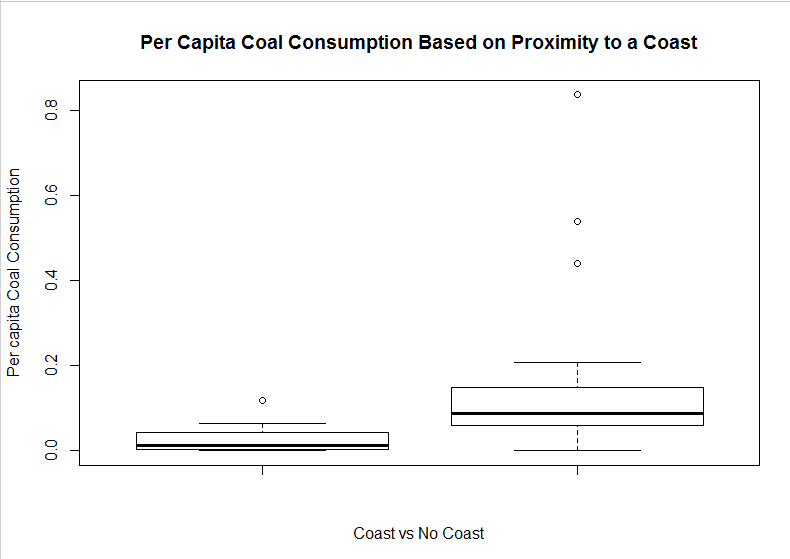
Fail to reject the null hypothesis because the p-value is larger than 0.05. Per capita energy consumption does not differ depending on whether a state is found on the coast or not.

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).

Null: Per capita coal consumption does not differ depending on whether a state is found on the coast or not.

Alternate: Per capita coal consumption does differ depending on whether a state is found on the coast or not.

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



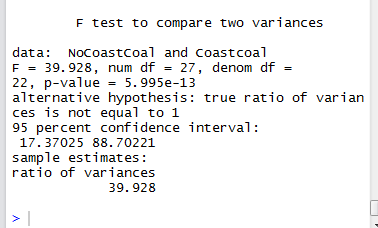
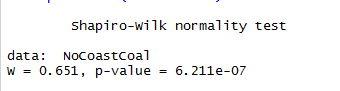
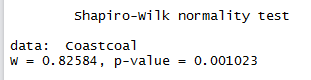
* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

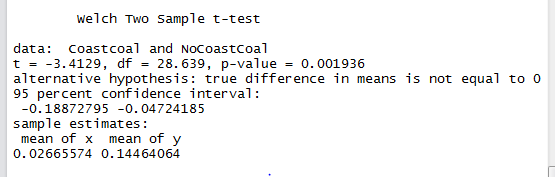
Independent: Coast (categorical)

Dependent: per capita coal consumption (continuous)

**Test to use: Unpaired 2-sample t-test**

Assumptions:

* Samples have equal variance 🡪 var.test()
  + 
* Samples are normally distributed 🡪 shapiro.test()
  + 
  + 
* Each observation is sampled independently (for this, I am assuming it is met since each household does not likely influence the others)
  + **All Assumptions are Met**
  1. Please run the statistical test and interpret the result (1 point).



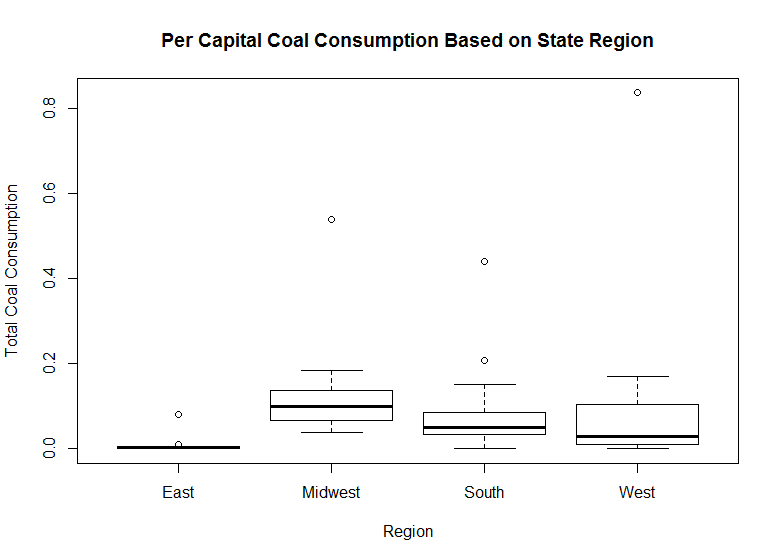
* P-value is significant, indicating that we can reject the null hypothesis. Per capita coal consumption does differ depending on whether a state is found on the coast or not.

1. Does ***per capita*** coal consumption differ depending on the region in which a state is found?
   1. Please write the null and alternate hypothesis (1 point).

Null: Per capita coal consumption does not differ depending on the region in which a state is found.

Alternate: Per capita coal consumption does differ depending on the region in which a state is found.

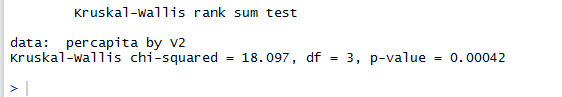
* 1. Please create a visual plot to answer this question (1 point).



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

Would use a one-way ANOVA if it passed the normality test, but since it does not use the **Kruskal Wallis test.**

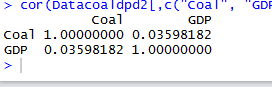
Assumptions:

* Samples have equal variance 🡪 leveneTest()
* Samples are normally distributed 🡪 shapiro.test() **\*\*\*Fails this test\*\*\***
* Each observation is sampled independently (for this, assuming it is met since each household does not likely influence the others)
  1. Please run the statistical test and interpret the result (1 point).

P-value is significant, so reject the null hypothesis. There is a significant difference between coal consumption and region in which the state is found.

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)

*Using Pearson Correlation Coefficient, the R output is:*

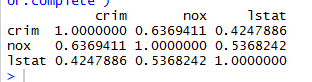


This shows that the correlation between per capita coal use and per capita GDP is 0.03598182, which is fairly weak (1 or -1 represents perfectly correlated data, and numbers that grow further away from 1 are less correlated).

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).

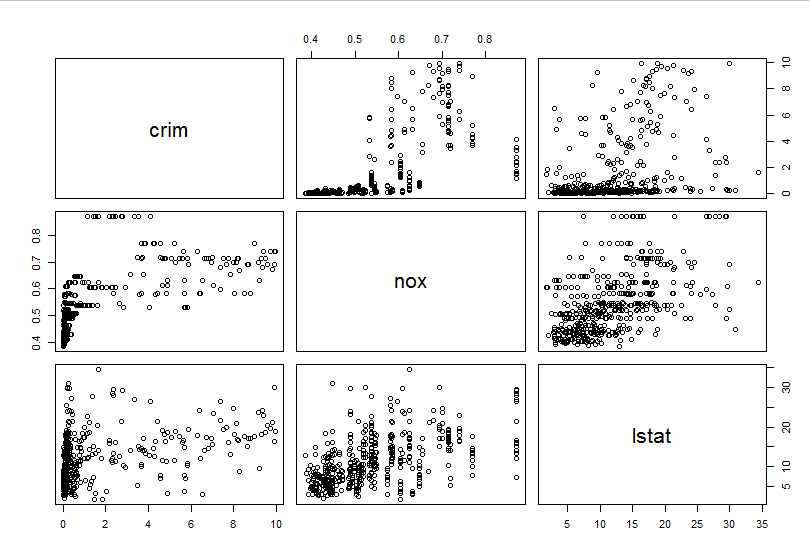
Correlation Values:

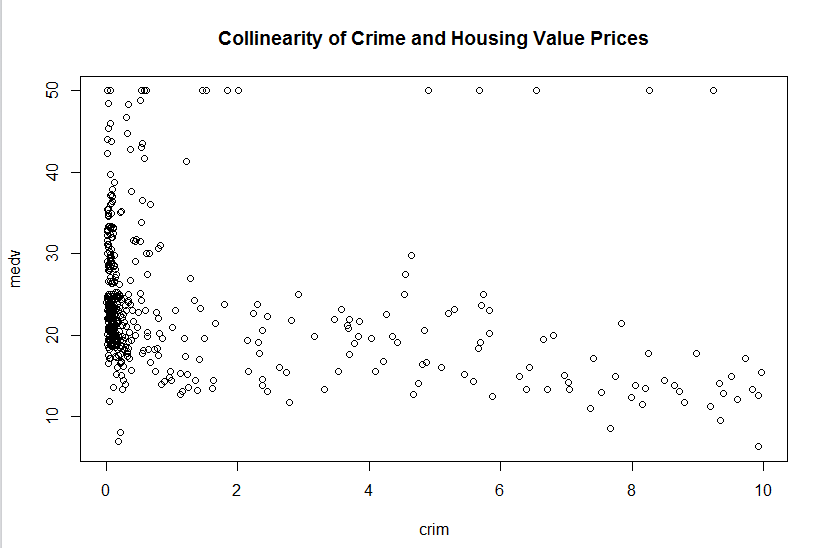


**VIF = 2.008428**

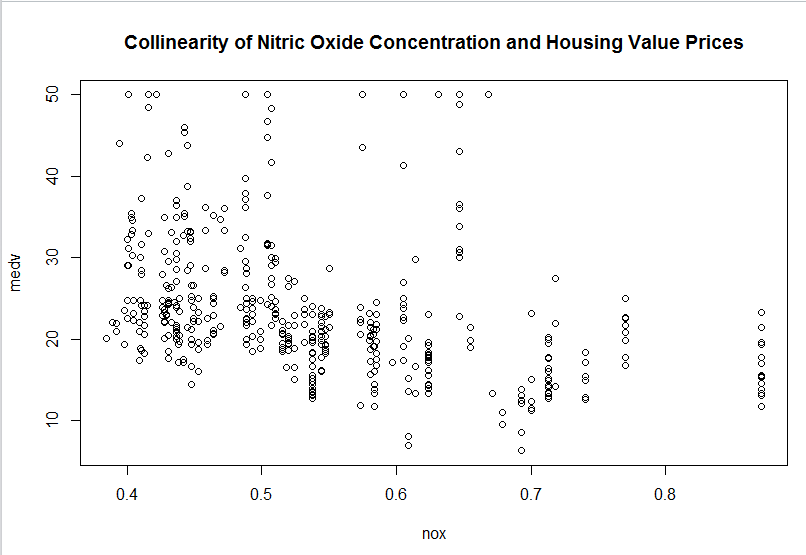
Crim (per capita crime rate by town) and nox (nitric oxides concentration) have the highest correlation values. The second highest is nox and lstat (% lower status of the population) and the least correlated is crim and lstat. Although Crim and Nox are highly correlated, the variance inflation factor is smaller than 5, meaning that multicollinearity is not an issue that would cause us to discard variables from our model. That being said, given the correlation value of crim and nox, there is a possibility of a bias that could result from their collinearity (which is important to keep in mind).

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).

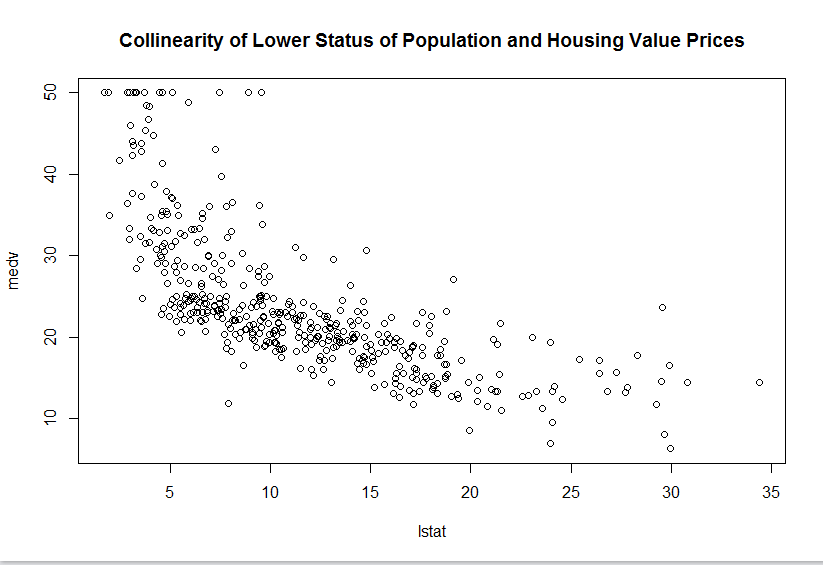




Crime and housing value prices appear to be generally negatively correlated.

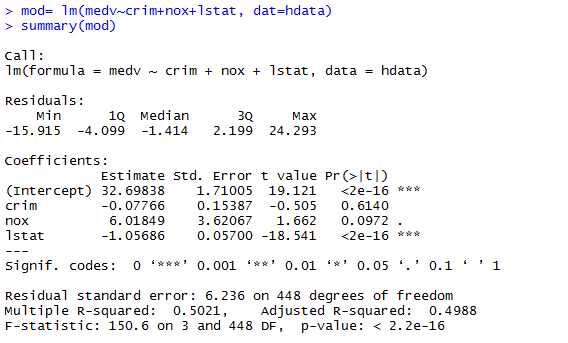


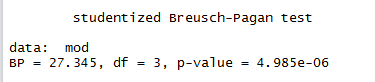
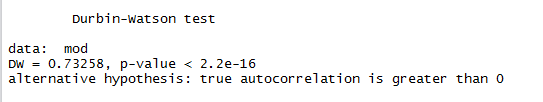
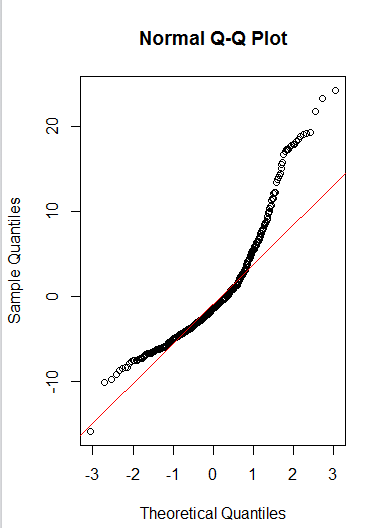
Nitric Oxide Concentration and Housing Value Prices seem to be generally negatively correlated.



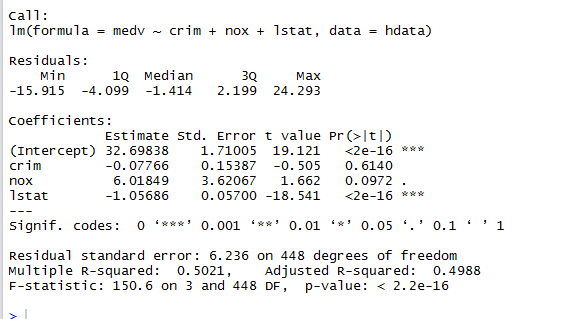
Lower status and housing value prices appear to be generally negatively correlated.

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).



* Assumptions:
  + **There is a linear relationship between variables** (follows this assumption, I know from calculating correlation in the above question)
* Homoscedasticity (failed this assumption, I know since I tested using the bptest)
  + 
    - Null hypothesis of homoscedasticity is rejected – this it is not homoscedastic.
* Statistical independence of the errors (lack of correlation) (follows this assumption, I know since I tested using the dwtest)
  + 
    - Passed because p value is significant, thus fail to reject null hypothesis that there is autocorrelation. (there should be no autocorrelation)
* Normality of the error distribution (failed, I know because I tested using the qqnorm). However, I can still use this test since the other assumptions are met)
  + 

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).



* Intercept: Significant. It represents the median value of owner-occupied homes when crime rate is zero.
* Crim estimate: Not significant. It represents the effect crime rate has on the median value of owner-occupied homes.
* Nox estimate: Not significant. It represents the effect nitric oxide concentration has on the median value of owner-occupied homes.
* Lstat estimate: Significant. It represents the % lower status of the population on the medium value of owner-occupied homes.

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

* Goodness of fit is determined by how much variance is explained by the independent variables in the model (crim, nox, and lstat). The adjusted R^2 for this case is 0.9293, which is a high R^2 value. This indicates that my model is a good fit.