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

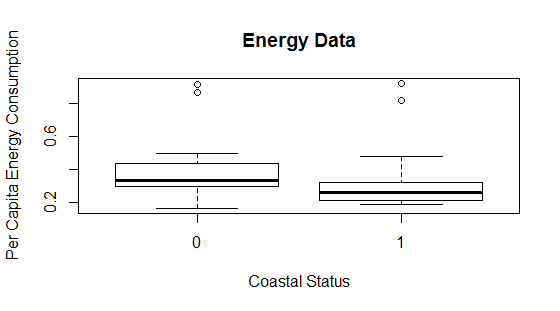
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized 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 hypothesis: per capita energy consumption is not related to whether a state is found on the coast. The mean energy consumption for states on the coast is not different from the mean energy consumption of states that are non-coastal.*

*Alternate hypothesis:* *per capita energy consumption is related to whether a state is found on the coast. The mean energy consumption for coastal states is different from non-coastal states.*

* 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).

*We should use a t-test because the independent variable is categorical (with only two values) and the dependent variable is continuous.*

*We will check for normality using shapiro.test (see code). We will also check for equal variances with var.test (see code Shapiro.test suggests that the data are not normal, so I transformed the data; var.test suggests that the variances are equal. We must also make sure it meets the assumptions that the observations are collected independently, but it is difficult to test that.*

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

Welch Two Sample t-test

data: coast\_log$LogEnergyPC and notcoast\_log$LogEnergyPC

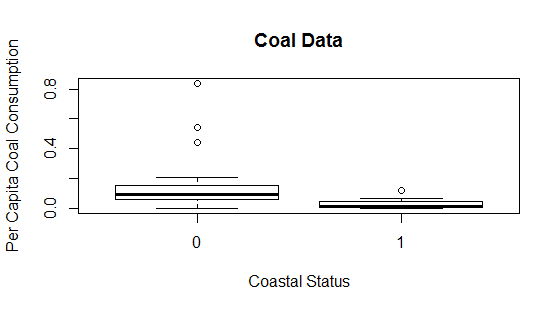
t = -1.8941, df = 43.234, p-value = 0.06492

*We got a p-value of greater than .05. If we set our alpha value at .05, this means that we do not have enough evidence to reject the null hypothesis that the mean energy consumption of coastal and non-coastal states are different from each other.*

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 hypothesis: per capita coal consumption is not related to whether a state is found on the coast. The mean coal consumption for states on the coast is not different from the mean coal consumption of states that are non-coastal.*

*Alternate hypothesis:* *per capita coal consumption is related to whether a state is found on the coast. The mean coal consumption for coastal states is different from non-coastal states.*

* 1. Please create a visual plot to answer this question (1 point).
  2. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

*We should use a t-test because the independent variable is categorical (with only two values) and the dependent variable is continuous.*

*We will check for normality using shapiro.test (see code). We will also check for equal variances with var.test (see code). Shapiro.test suggests that the data are not normal, so I transformed the data; var.test suggests that the variances are equal. We must also make sure it meets the assumptions that the observations are collected independently, but it is difficult to test that.*

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

Welch Two Sample t-test

data: coast\_cubert$CubertCoalPC and notcoast\_cubert$CubertCoalPC

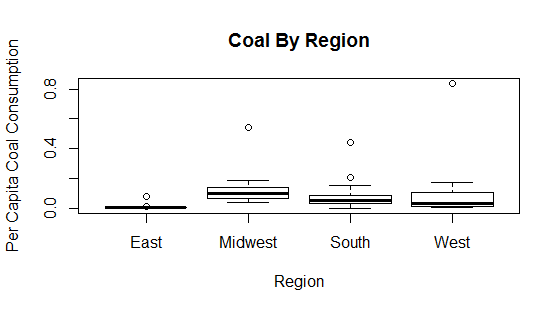
t = -4.4689, df = 45.271, p-value = 5.217e-05

*The p-value is < .05, meaning that we have enough evidence to reject the null hypothesis, and accept the alternate hypothesis that the mean coal consumption is different in coastal states than non-coastal states.*

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 hypothesis: per capita coal consumption is not related to a state’s region. The mean coal consumption for states does not differ by region.*

*Alternate hypothesis:* *per capita coal consumption is related to a state’s region. The mean coal consumption for coastal states differs by region.*

* 1. Please create a visual plot to answer this question (1 point).
  2. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

*We should use an ANOVA because the independent variable is categorical (with more than two values) and the dependent variable is continuous. We will check for normality using shapiro.test (see code). We will also check for equal variances with leveneTest (see code).*

*Shapiro.test suggests that the data are not normal, so I transformed the data, even though that is not quite as critical for an ANOVA. leveneTest suggests that the variances are equal.*

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

Df Sum Sq Mean Sq F value Pr(>F)

Coast 1 0.5012 0.5012 18.19 9.1e-05 \*\*\*

Residuals 49 1.3505 0.0276

*The p value is less than .05, so we can safely reject the null hypothesis and conclude that there is a significant difference between mean coal consumption across different state's regions.*

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)

*The correlation is 0.0356. Since -1 is a perfectly negative linear correlation, and +1 is a perfectly positive linear correlation, this result does not seem to be particularly strong correlation because its absolute value is very small.*

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).

*We will use AGE, NOX, and CRIM.*

*None of the correlation values have an absolute value greater than .33, so I feel confident that we will not have to worry about multi-collinearity. See below:*

age nox crim medv

age 1.0000000 0.7071478 0.4476638 -0.2998932

nox 0.7071478 1.0000000 0.6369411 -0.3327782

crim 0.4476638 0.6369411 1.0000000 -0.2862450

medv -0.2998932 -0.3327782 -0.2862450 1.0000000

*The VIF values also appear to be satisfactory:*

> vif.age

[1] 1.098824

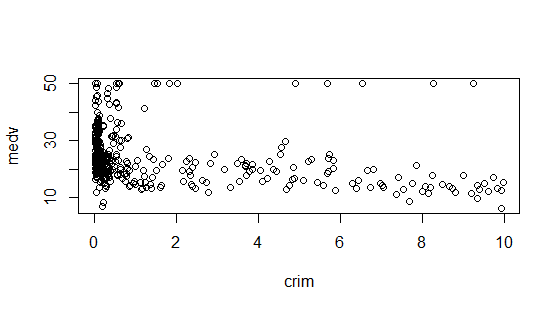
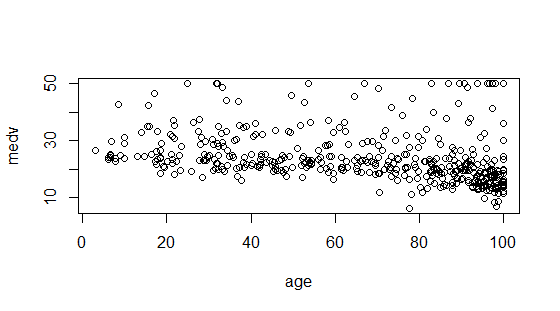
> vif.nox

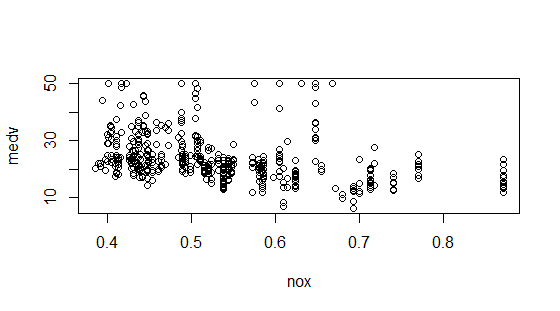
[1] 1.124532

> vif.crim

[1] 1.089249

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).

*Age is more closely related to median housing price as the house’s age increases (maybe because antique houses are considered more valuable).*

*Low crime rates are strongly related to median housing price, but high crime rates are not as highly related. Maybe because houses that have extra security features are more expensive, and tend to be in higher-priced neighborhoods.*

*Levels of NOX are not very strongly correlated with median housing price, except at slightly lower levels of NOX.*

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).

*I checked for multi-collinearity. Based on my answer to question 5, I don’t think we need to worry about that since the VIF values were all small.*

*I also checked for linearity by creating regression lines on my scatterplots. All of them were linear.*

*However, shapiro.test showed that the data were not normal. I attempted to overcome this by log-transforming them and was able to make them slightly more normal. I tried to use a box-cox transformation but couldn’t get it to work, so I decided to proceed as normal, since we have a large sample size.*

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).

Call:

lm(formula = medv ~ age + nox + crim, data = hdata)

Residuals:

Min 1Q Median 3Q Max

-15.118 -4.856 -2.447 2.433 32.236

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 33.78324 2.32797 14.512 <2e-16 \*\*\*

age -0.04066 0.01954 -2.081 0.0380 \*

nox -12.45743 5.59998 -2.225 0.0266 \*

crim -0.44325 0.20192 -2.195 0.0287 \*

*With no other input from other variables, the median value of housing prices would be $33.78 thousand dollars.*

*Every one-unit increase in proportion of old homes causes the housing price to decrease by $-.041 thousand dollars.*

*Every one-unit increase in nitric oxide concentration causes the housing price to decrease by $12.46 thousand dollars.*

*Every one-unit increase in per capita crime rate causes housing prices to decrease by $-.44 thousand dollars.*

*All of these effects are significant.*

*I also added an interaction term by incorporating the Charles River variable.*

Call:

lm(formula = medv ~ nox \* chas, data = hdata)

Residuals:

Min 1Q Median 3Q Max

-14.351 -4.937 -2.128 2.498 29.698

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 38.174 1.989 19.194 < 2e-16 \*\*\*

nox -27.623 3.632 -7.604 1.7e-13 \*\*\*

chas 7.780 6.216 1.252 0.211

nox:chas -1.890 10.310 -0.183 0.855

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.143 on 448 degrees of freedom

Multiple R-squared: 0.1511, Adjusted R-squared: 0.1455

F-statistic: 26.59 on 3 and 448 DF, p-value: 7.632e-16

*Houses located next to the Charles River with 0 nitric oxide concentrations have a median value of $38.17 thousand dollars.*

*Each additional unit of nitric oxide concentration decreases median property values by $27.62 thousand dollars.*

*Houses located next to the Charles River are worth $7.78 thousand dollars more than houses that aren't located next to the river.*

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

*Since the R-squared value is quite small for both models, I don’t think they are a very good fit. The first model only explains about 13% of the data, and the second only explains about 15%. It’s possible that this is due to the non-normality of the data; alternatively, maybe the variables that I chose are just not very good predictors of housing prices, at least not when used alongside each other.*