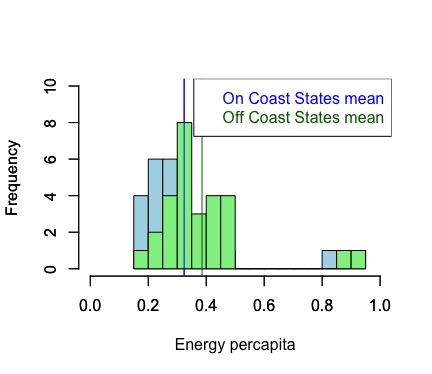
\*\*\*My answers in blue. See R script for more details.\*\*\*

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.

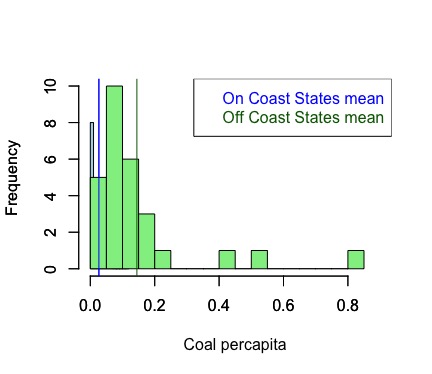
1. Does ***per capita*** energy consumption differ depending on whether a state is found on the coast or not?
   1. Please write the null and alternate hypothesis (1 point).
      1. Per capita energy consumption does not differ by states on or off the coast.
      2. There is a difference in per capita energy consumption between states on the coast and off the coast.
   2. Please create a visual plot to answer this question (1 point).



The Off Coast States mean appears to be higher than the On Coast mean from this histogram.

* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
     1. Statistical test = two sample unpaired t test
     2. Continuous data and random sample = yes, we can see this in the data set.
     3. Normality: yes, while both the on and off coast data produce a p-value of ~0 when using the Shapiro test, we can assume normality through the large enough sample size and central limit theorem.
     4. Equal variances: yes, meets this assumption, since the p-value for an F test is .51.
     5. Independent observations: yes, we assume this to be true given the data set.
  2. Please run the statistical test and interpret the result (1 point).
     1. Results: p-value = .22
     2. Therefore, mean per capita energy consumption between on and off shore states is not significantly different, as the probability that these means are different due to chance is greater than 0. Note: it’s interesting that if you run the Mann-Whitney test here if assuming non-normality based on the Shapiro test only, you get a significant p-value!

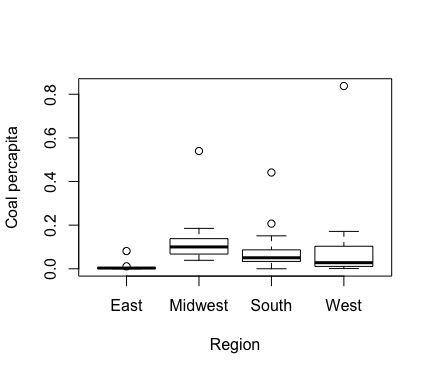
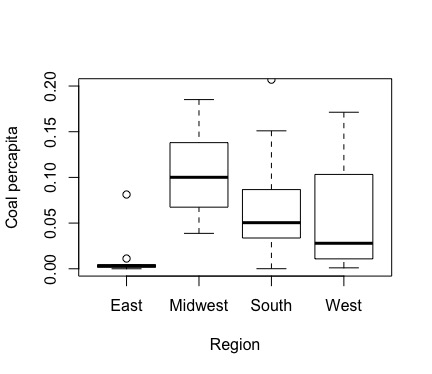
1. Does ***per capita*** coal consumption differ depending on whether a state is found on the coast or not?
   1. Please write the null and alternate hypothesis (1 point).
      1. Per capita coal consumption does not differ by states on or off the coast.
      2. There is a difference in per capita coal consumption between states on the coast and off the coast.
   2. Please create a visual plot to answer this question (1 point).



The Off Coast States mean appears to be higher than the On Coast mean from this histogram.

* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
     1. Statistical test = two sample unpaired t test
     2. Continuous data and random sample: yes, we assume this to be true based on the data
     3. Normality: yes, while both the on and off coast data produce a p-value of ~0 when using the Shapiro test, we can assume normality through the large enough sample size and central limit theorem.
     4. Equal variances: no, violates this assumption, where the F test shows a very small F value and a p-value of ~0
     5. Independent observations: yes, we assume this to be true based on the dataset.
     6. By not meeting one of these assumptions, I recognize that I increase my risk of Type 1 error.
  2. Please run the statistical test and interpret the result (1 point).
     1. Results: p~0
     2. Therefore, mean per capita coal consumptions between on and off shore states are significantly different and there is a very low probability that the difference in these means would be due to chance.

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).
      1. Per capita coal consumption does not differ by state region.
      2. There is a difference in per capita coal consumption between different state regions.
   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. Statistical test: ANOVA due to continuous dependent variable and categorical independent variable with levels
     2. Continuous data and random sample: yes, assumed based on data
     3. Normality: yes, while the per capita coal consumption variable produces a p-value of ~0 when using the Shapiro test, we can assume normality through the large enough sample size and central limit theorem.
     4. Homogeneity of variances: yes, the Levene test gives a p-value of .52, which says we cannot reject the null hypothesis that variances are equal.
     5. Observations independent: yes, this is assumed based on the dataset given
  2. Please run the statistical test and interpret the result (1 point).
     1. Results: p-value=.262. There is not a significant difference in coal consumption per capita among regions. Using the Tukey HSD test, we can confirm the lack of differences among regions if we wanted. Note: it’s interesting if I run the non-parametric test that these differences are significant!

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. Correlation = .04, very weak or basically 0
   2. We would want a value greater than .5 or less than -.5 for a strong correlation
   3. I was initially surprised by this value because usually energy consumption and GDP go together, yet due to coal’s decline as an energy source and the data framed in terms of per capita usage could contribute to the near 0 correlation of these data.

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. Covariates: rm (number of rooms per dwelling), tax (property tax), dis (distance to employment centers)
   2. Correlation values: all are below .5.

rm tax dis

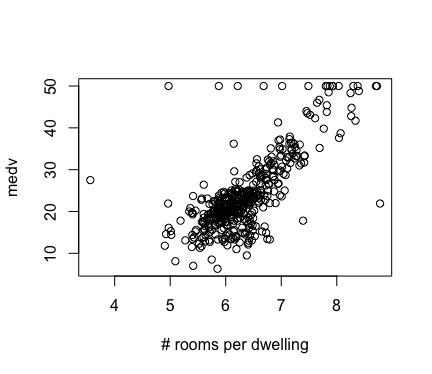
rm 1.0000000 -0.2149478 0.1387741

tax -0.2149478 1.0000000 -0.4441307

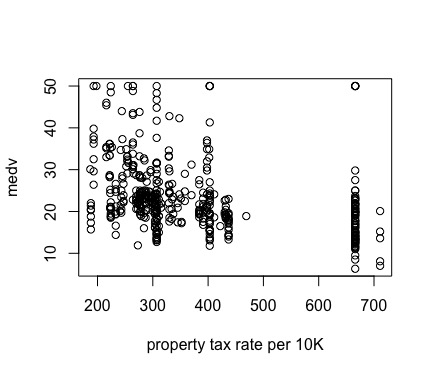
dis 0.1387741 -0.4441307 1.0000000

* 1. VIF value = 2.4, which means that none of the variables are highly correlated.

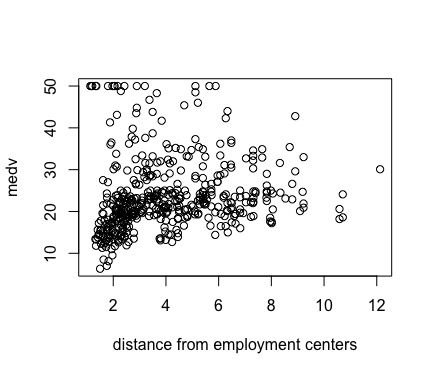
1. Plot the relationship between each of your three independent variables and the dependent variable (medv). **Include each plot in this answer and state whether and how you think each variable is related to median housing prices** (medv; 3 points).



Yes, the number of rooms has a pretty strong positive correlation on median housing prices, R2=.55



No, property tax rate seems to have a weak relationship with median housing prices, R2=.12



No, distance from central employment spots seems to have essentially no relationship with median housing prices, R2=.02

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. Test choice: regular linear model due to all continuous variables
   2. Normality: No, violated through qq-plot and Shapiro test’s result of p-value~0
   3. Errors/residuals normality: No, violated through the Shapiro test result of p-value of ~0
   4. Non-autocorrelated errors: No, violated through dw test where p-value is ~0
   5. Homoscedasticity: No, violated through the bptest where p-value~0
   6. Through violating these assumptions, I understand that it would be best to transform the data but am choosing not to do so for the purposes of this quiz. I also recognize that this increases my chances of committing Type 1 error.
2. 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).
   1. The intercept is essentially the average value of the dependent variable, or median house value in this case, when we consider all values in the dataset and the other variables are held constant at 0. The negative value appears here because it is impossible to have a house with 0 rooms, tax rate, or distance to employment centers, but if it were possible, the value of the house would be negative. When the intercept and variables are significant, we can also assume that there is a significant relationship between our response and independent variables.
   2. The number of rooms has a significant effect on median house prices in that it can explain a significant amount of variance greater than 0. For every increase in one room, housing prices go up 9.25 units ($).
   3. The property tax rate has a small significant effect on median house prices and can also explain a significant amount of variance greater than 0. For every increase in one percentage tax rate increase, median house prices go down .01 units.
   4. For distance to employment centers, it is only significant at the .1 alpha level. For every mile away from employment centers, median house prices go down .24 units.
   5. The F statistic is large and significant where the p-value~0, so there is a very low probability that this model would fit well due to chance.

Call:

lm(formula = medv ~ rm + tax + dis, data = hdata)

Residuals:

Min 1Q Median 3Q Max

-21.208 -2.779 -0.407 2.113 42.020

Coefficients:

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

(Intercept) -29.163771 2.947478 -9.894 < 2e-16 \*\*\*

rm 9.255521 0.411201 22.509 < 2e-16 \*\*\*

tax -0.012823 0.002003 -6.403 3.86e-10 \*\*\*

dis -0.237096 0.142970 -1.658 0.0979 .

---

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

Residual standard error: 5.68 on 448 degrees of freedom

Multiple R-squared: 0.587, Adjusted R-squared: 0.5842

F-statistic: 212.2 on 3 and 448 DF, p-value: < 2.2e-16

1. Discuss the fit of your model and whether you think it is a good or bad fit. Why (2 points)?
   1. This fit is pretty good due to the adjusted R-squared value of .584 and F statistic of 212.2, but most of the explanation in the data’s variability comes from the rm variable for number of rooms per dwelling. The tax and dis variables contribute very minimally to the fit of this model, and tax is likely significant at the .05 level because of correlations with rm. Still, since all remain significant in the full model, they all have some effect on predicting median house prices.