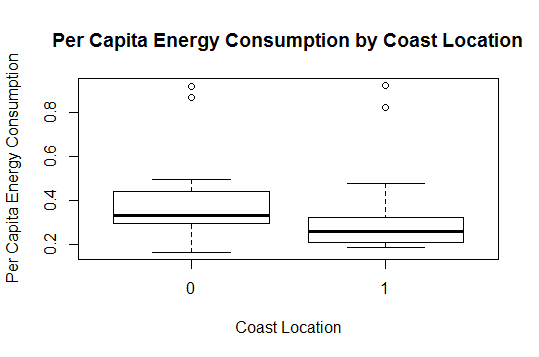
If you run any ANOVAs, you can use the Levene test for equality of variances (leveneTest). If your data violate an assumption about normality, please decide if this is really a problem. In many cases you can still run your parametric test with non-normal data assuming other conditions are met (see lecture notes). If you choose to run a parametric test any way despite the data not being normally distributed, state why you are able to do this. HINT: there is only one analysis in the entire exam (which is clearly marked) where you should run into real problems with normality. For this one analysis, you can get bonus points for transforming your data. If you are unable to transform your data, run the statistical test any way as if your data were normally distributed but make it clear that you violated this assumption in your answer (you won’t lose any points for violating this assumption). I’ve also updated Lecture15.R due to one mistake in the code.

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

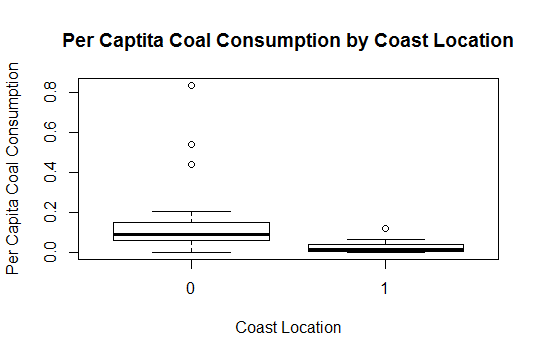
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized 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. Null: Per capita energy consumption does not differ between states found on the coast and states not found on the coast
      2. Alternative: Per capita energy consumption is higher for states found on the coast than states not found on the coast.
   2. Please create a visual plot to answer this question (1 point).



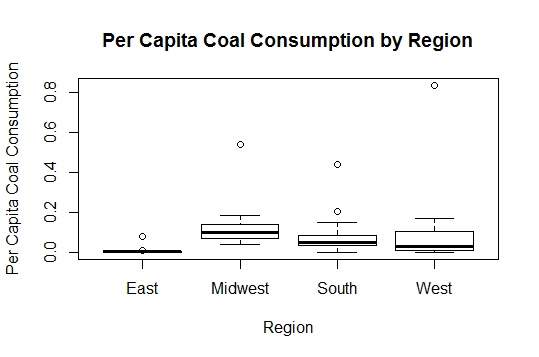
* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
     1. To Test the relationship between Per Capita energy consumption and coast location, we would run a two-sampled t-test. To do this, we must first check to make sure our assumptions are met. Although this data set is not ours, we can assume that the results contain a random sample. This means that our data was randomly selected from the population. We also assume that are observations are independent. We also assume that the sample size is small enough, meaning it is less than 10% of the entire population size. These are main assumptions we can make about our data. Next, we run test a test to see if values are nearly normal. To do this, we run the shapiro-wilks test and qqplot test. After running these, we see that the data is not normal. This can be shown both in the qqplot, as well as from the results of the Shapiro test. When running the Shaprio-wilks test, for both per capita energy consumption on the coast and off the coast, the p-values are significant (5.037x10^-6 and 3.62710^-5, meaning that are data is not normal. This data shows one of our assumptions is not met. However, we have a large sample size, an so we can infer that normality may be met. We then must test for variance. After running a variance test, we see that the p-value is 0.5, which is larger than our alpha (0.05). This means we cannot reject our null hypothesis, and so the variances are not equal, which is what we want to see. Therefore, the assumptions can be met.
  2. Please run the statistical test and interpret the result (1 point).
     1. After running the two sample t-test, we see that the p-value is not significant. This means that per capital energy consumption is not correlated with location to the coast. We see from this that there is not a significant difference between per capita energy consumption both on and off the coast in terms of location. Therefore, we fail to reject the null hypothesis. It is possible that outliers are skewing with our data. As can be seen when plotting the results, we see that there are several outliers that skew the normality. It is possible that we could take these out of the data set, and then it is possible we may see normality, and thus significant results.

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. Null: Per capita Coal consumption does not differ between states found on the coast and states not found on the coast.
      2. Alternative: Per capita coal consumption is higher in states that are on the coast than states that are not on the coast.
   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. To test whether per capita coal consumption is significantly different between coast locations, we would use a two-sample t-test. First, we assume that our data was randomly sampled from the population, we assume that observations are independent, and we assume that the sample size is small enough (less than 10% of the entire population). We can assume these things, as the data is not necessarily ours. Next, we test for normality, using both the shapiro-wilks test and a qqplot. After running these tests, we see that the p-values are significant for per capita coal consumption both on the coast and off the coast. This means we fail to reject our null hypothesis and that the data is not normally distributed. However, we have a large sample size, and thus, we can assume normality is achieved. Lastly, we must test our assumption of equal variance. When testing for this, we receive a significant p-value, and thus we fail to reject our null hypothesis, meaning that the variances are equal. We had wished to see a high p-value. Therefore, not all of our assumptions are met.
  2. Please run the statistical test and interpret the result (1 point).
     1. After running the two-sample t-test, we receive a significant p-value of 0.001. This tells us that per capita coal consumption is significantly different based on whether your location is on the coast or not. However, we did not have equal variances in this, and so our data cannot be completely accurate. This could be due to the fact that some of our data points are outliers and are potentially skewing the data. We could transform our data or potentially use a different statistical test to better fit our data and achieve more accurate results.

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. Null: Per capital coal consumption does not differ depending on the region in which the state is found
      2. Alternative: Per capita coal consumption is significantly different depending on which region the state is found.
   2. Please create a visual plot to answer this question (1 point).



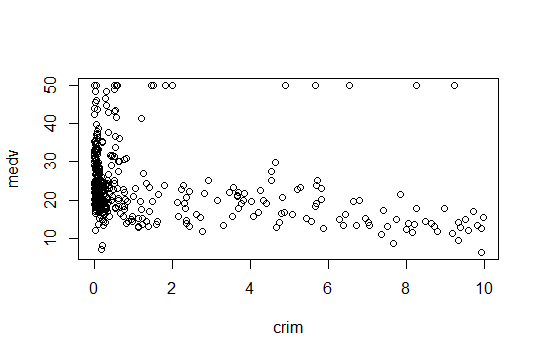
* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
     1. To test whether per capital coal consumption is statically significant between different regions in which a state is located, we need to run an ANOVA. For this ANOVA test, we first need to check the assumptions. We assume that the data was randomly sampled from the population, that the observations are independent, and that the sample size is small enough. Next, we test to see the values are nearly normal. The p-value was significant when running the shapiro-wilks test, and so our values are not normal. However, we have a large sample size, and so we assume the test is okay to run. Next, we test for variance using the Levenes test. From this, we see that the p-value is not significant, so we can assume equal variances.
  2. Please run the statistical test and interpret the result (1 point).
     1. After running the ANOVA, and with most of our assumptions met, we get a p-value that is not significant (0.267). From this, we can assume that per capita coal consumption is not significantly different based on what region a state is found in.

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. There is not a correlation between per capital coal use and per capital GDP. To test this in R, I ran a correlation test. I also visually plotted the correlation between these two variables. When looking at the correlation between these two variables, we get a value of 0.035. This means that these two variables are not highly correlated with one another. To achieve correlation, we want a high correlation coefficient, and we did not get that here.

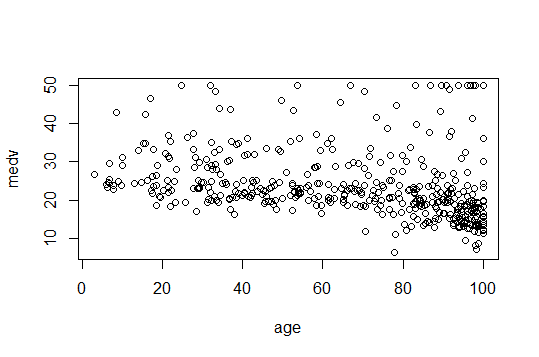
For questions 5-9, please use the ‘housedata.csv’ dataset that shows housing information for the Boston area. Information on what each of the variables are can be found here: <http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names>. In this exercise, the goal is to create a multiple 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. For my linear model, I selected the variables crime, age, and peer to teacher ratio. When testing for correlation, none of these variables were highly correlated with one another. Age and Crime’s value was 0.44, age and peer to teacher ratio was 0.19 and peer to teacher ratio and crime value was 0.31. These values are low and show these variables are not highly correlated with one another. When running VIF, we get a value of:
   2. crim age ptratio
   3. 1.345597 1.254989 1.117544
2. 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).

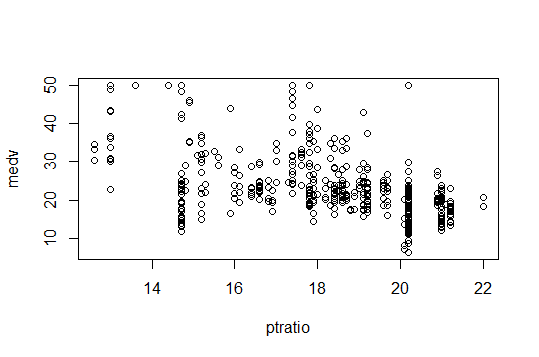
Median housing prices and crime were not highly correlated with one another. This can be shown by the plot below. If they were correlated with one another, we would see that median housing cost decreases as crime increases.



Median housing prices and age could potentially be slightly correlated with one another, especially as age increases. From the plot below, we see that as age increases, median housing cost somewhat decreases. This can be shown by the plot below:



Median housing prices and peer to teacher ratio were somewhat correlated with one another. It is challenging to say based on these plots without running and statistical tests, but we see some trends and some correlation from the plot below. We may say that as peer to teacher ratio increases, median housing cost decreases.



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. Now, it comes time to run the multiple linear regression. This model looks at the median housing prices, and the variables age, crime, and peer to teacher ratio, and the correlation between these values. Our goal is to see if we can predict one quantitative dependent variable from multiple dependent variables. For this linear regression to be run, however, we must make sure that our assumptions are met. We must check for residual independency, residual homoscedasticity, and residual normality. To do this, we run a series of different tests. First, I ran the Durbin-Watson (dwtest) test to check for independency. The dwtest looks at autocorrelation between the variables. The results from this test give me a significant p-value, which means that there is some autocorrelation. This means one of our assumptions is not being met.

Next, we run a bp-test, to look at the homoscedasticity of the data. Upon running this test, we receive a p-value of 1.5x10^-7. This is not what we want to see, as this means our residuals are not homoscedastic. This means that our residuals are not homoscedastic, our assumption of homoscedasticity is not met.

Lastly, we need to check for normality by running a qqplot or by using a shapiro-wilks test. After running the shapiro test, we see that our residuals are not normal. We know this because we have a significant p-value, which means we reject the null hypothesis of our data having equal variance. This means that our data does not have equal variance, and so our assumption of equal variance is violated.

From running these tests, we see that none of our assumptions are met. This means this model is potentially not the best fit for this data. While we don’t want overfitting, we do want the model to accurately predict our independent variable, in this case median housing cost. One way we could solve this problem is by log transforming the data. This will help to better fit the data to the model. If this does not work, another model or statistical test would better explain our data and the variance within our data set.

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).
   1. After running the model, I can see that some of my dependent variables can explain some of the variation in the model, however, not all of them do. To begin, we must identify the intercept or the y-intercept. For this model, the intercept is 57.32. This means that when there are no added variables, such as age, or crime, median housing cost will be this much at the intercept. Next, we move on to our three coefficients. For crime, every one increase in crime will result in a decrease in median housing cost by 0.25 (coefficent-0.25). This value however, is not significant, as stated by the p-value (0.121), and so crime does not explain a lot of variation in median housing cost, or at least not significantly. Next, we have age. As age increases by a factor of one, median housing cost decreases by 0.05 (coefficient of -0.05). This coefficient is significant, as can be shown by the p-value of 3.57x10^-5. This means that age explains some of the variation in median housing cost within our model, and could potentially be an accurate predictor of our independent variable. Lastly, we have peer to teacher ratio. For this, we have a coefficient of -1.6. This means that for every one increase in peer to teaching ratio, median housing costs decrease by this number. This is also statistically significant, as can be shown by the coefficient having a p-value of 2x10^-16. This is potentially good for our model, as peer to teacher ration could explain some of the variation in the model and accurately predict our independent variable.

Now, interpreting the model. From this model, we received an R-squared value of 0.26. This means that these variables do not explain a lot of the variation in the model, at least not accurately. This would make this model not very good in predicting the independent variable, which is what we would like our model to do. However, we did achieve a significant p-value of 2.2x10^-16, so it is possible that there is some significance within our model.

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

Based on the results, the fit of my model is only okay. While age and peer to teacher ratio can explain significant variation in median housing cost, crime does not, and so we can’t assume that this model would accurately predict our independent variable. This can be shown by the R-squared value, and how it is relatively low, meaning our dependent variables do not explain significant variation in our model. It is possible that if we log transformed some of the data, we may get different results. It is also possible that other variables better explain the variation in our model and can more accurately predict median housing cost. These are things to test for in the future.