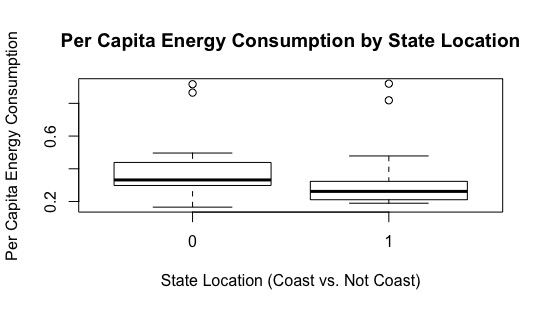
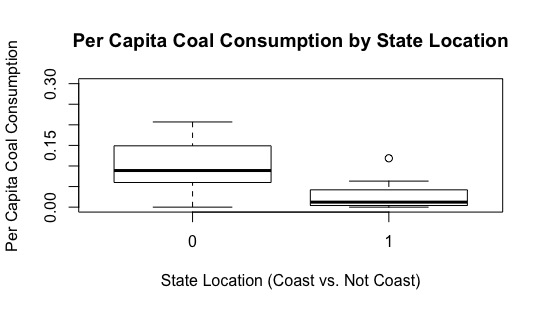
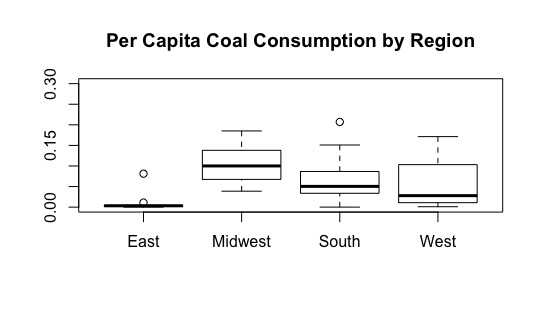
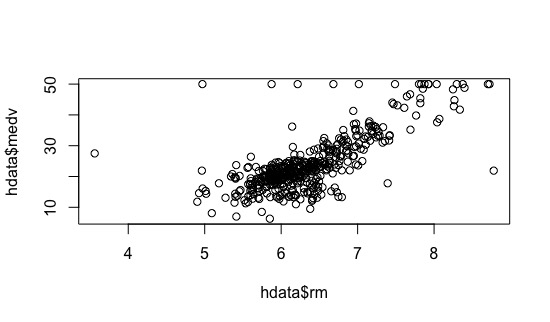
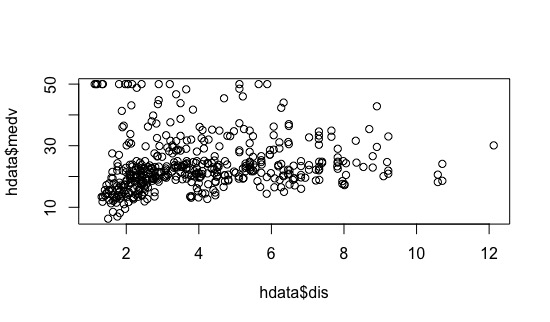
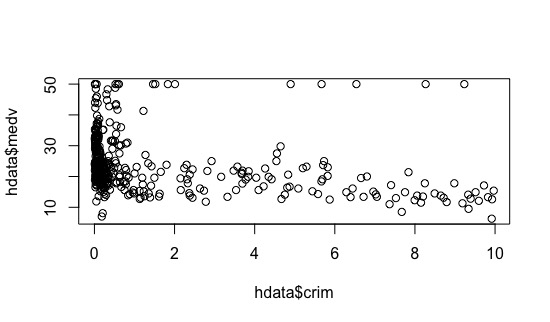
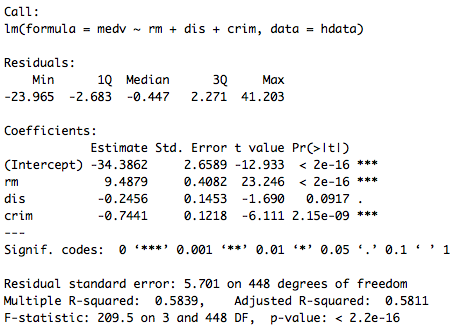
Please use the R script provided to load data and build your script from there.

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: There is no difference in per capita energy consumption depending on whether or not the state is on the coast.  
      Alternate: There is a difference in per capita energy consumption depending on whether or not the state is on the coast.
   2. Please create a visual plot to answer this question (1 point).  
        
      
   3. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).  
        
      I decided to use a two sample t-test to determine if there was a difference in per capita energy consumption between coastal and non-coastal states. The assumptions this statistical test needs to meet is equal variance of samples, samples are normally distributed, and observations are independent. We can assume the observations are independent, and when tested the variances are equal, but the data is not normally distributed. However, the sample size is sufficiently large, so we can assume normality via Central Limit Theorem and proceed with the test.
   4. Please run the statistical test and interpret the result (1 point).  
        
      When the two-sample t-test is run, the p-value of 0.2245 is not below the critical value of 0.05, so the difference in per capita energy consumption in coastal and non-coastal states is not significant.
2. 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: There is no difference in per capita coal consumption depending on whether or not the state is on the coast.  
      Alternate: There is a difference in per capita coal consumption depending on whether or not the state is on the coast.
   2. Please create a visual plot to answer this question (1 point).  
        
      
   3. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).  
        
      I decided to use a two sample t-test to determine if there was a difference in per capita coal consumption between coastal and non-coastal states. The assumptions this statistical test needs to meet is equal variance of samples, samples are normally distributed, and observations are independent. We can assume the observations are independent, but when tested, neither the variances are equal nor is the data normally distributed. However, the sample size is sufficiently large, so we can assume normality via the CLT and the t-test that R automatically runs (Welch’s t-test) accounts for unequal variance in the data, so we can proceed with the test.
   4. Please run the statistical test and interpret the result (1 point).  
        
      When the t-test is run, the p-value of 0.0019 is statistically significant, meaning that the per capita coal consumption does differ between coastal and non-coastal states.
3. 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: There is no difference in per capita coal consumption depending on the region a state is located in.  
      Alternate: There is a statistically significant difference in per capita coal consumption depending on the region a state is located in.
   2. Please create a visual plot to answer this question (1 point).  
        
      
   3. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).  
        
      I would initially think to run a One-Way ANOVA because my independent variable is categorical, with more than two “levels”, and my dependent variable is continuous. The assumptions of this test are normal distributions of populations, equal variances, and independence of the samples. We can assume independence of the samples. I used the Levene test for equal variances, and with a p-value of 0.5202, we can conclude variances between samples are equal. The samples are not normally distributed.
   4. Please run the statistical test and interpret the result (1 point).  
        
      The p-value of 0.262 is not statistically significant for the One-Way ANOVA, and thus we know that there is not a significant difference in per capita coal consumption between regions.
4. 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 between per capita coal consumption and per capita GDP is 0.0359. This does not seem like a very large correlation at all, because normally anything below 0.5 is considered okay, and this correlation is not even close to that. If I were to perform a multi-variate regression, I would not be concerned with multi-collinearity between these two variables.

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).  
     
   For my model, I selected # of rooms per dwelling (RM), per capita crime rate per town (CRIM), and weighted distances from major Boston city centers (DIS). None of the variables are too highly correlated to use in the model, although one pair is close: CRIM – RM (-0.14), CRIM – DIS (-0.46), and RM – DIS (0.14). None of the VIF values are close to or above 10, so we can safely assume there is no multi-collinearity between the independent variables (RM – 1.027, DIS – 1.28, CRIM – 1.28).
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).  
     
     
     
   Average number of rooms per dwelling seems to be pretty highly correlated with median house price. As number of rooms increases along the x axis, median house price increases in a fairly strong, linear way.  
     
     
     
   Distance from major city centers seems to be moderately correlated with median house value. The points are not as closely clumped as with # of rooms, but median house price seems to mostly increase as distance from city centers increases.   
     
     
     
   This variable, crime rate, seems to be least correlated with median house price, or at least not in a linear fashion. House prices are highest when crime rate is very low, and drop off steeply as crime rate increases. An exponential model might be a better fit for this variable.
3. 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).  
     
   The three assumptions of linear regression that must be met are independence of the errors, homoschedasticity, and a normal distribution of the errors. According to my calculations in R, all of these assumptions were violated for my model. I know this because all of the p-values for the tests I ran (dwtest, bptest, shapiro.test) were highly significant. I also plotted the residuals on the qq plot and they deviated from the 1:1 line. One possible way to overcome the violated assumptions, especially the non-normality of the residuals, is to log transform the data.
4. 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).  
     
     
     
   According to this model, both # of rooms and crime rate have a significant effect on median house price. The estimate for the intercept of -34.38, also known as the y-intercept, is what median house price would be when the independent variables are held constant. The estimate for RM of 9.48 is the slope coefficient for that variable, meaning that for every increase of 1 in RM, median house price increases by 9.48. The estimate of -0.24 is the slope coefficient for DIS, meaning that for every increase in 1 of DIS, median house price decreases by -0.24. The estimate of -0.74 is the slope coefficient for CRIM, meaning that for every increase of 1 in CRIM, median house price decreases by -0.74.
5. Discuss the fit of your model and whether you think it is a good or bad fit. Why (2 points)?  
     
   When I first began, I thought this model could be a good fit to explain median house price because none of the variables seemed too correlated with one another and when plotted individually, at least some of the variables seemed to be fairly strongly correlated. Also, when the model is run, the R2 is fairly high, and the p-values for two of the variables are significant. But after checking the assumptions for linear models and finding that all of them were violated, this raises questions as to the validity of the model in explaining the variability in median house price. Also, since of one the variables, crime rate, seems to follow a more exponential pattern, this makes me question whether a strictly linear model is the best fit for this data.