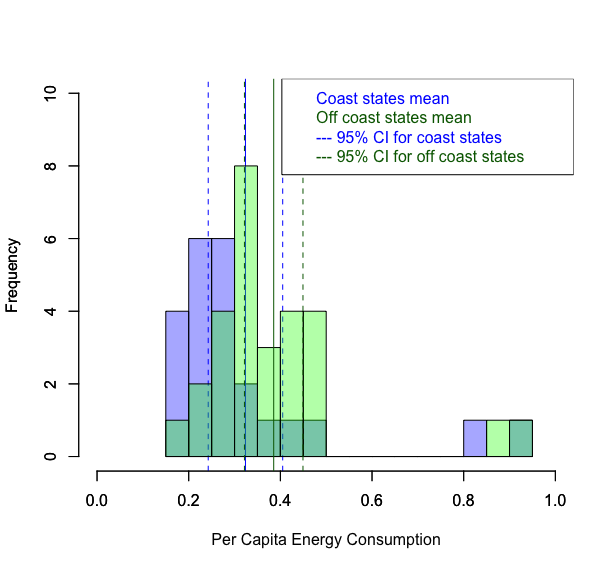
Please use the R script provided to load data and build your script from there.

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).
      * H0: The per capita energy consumptions do not differ between states on the coast and states off the coast;
      * Ha: The per capita energy consumptions are different between states on the coast and states off 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).
      * Statistics to use: At first, I hope to run a T-test, but I changed to use Wilcoxon-Mann-Whiteney U test because of the data are not normally distributed.
      * Assumptions for T-test:
2. Random sampled: **Unknown**
3. Independent observation: **Unknown**
4. Continuous data: **Yes**
5. Normal Distribution OR sample size > 30: **No.**   
   The data are not normally distributed and either of the sample sizes is smaller than 30 ---> Wilcoxon-Mann-Whiteney U test

* + - * + For states on the coast: (shapiro test)

p-value = 5.037e-06 < 0.05 --> reject the H0 that the distribution is not different from normal distribution -->(the distribution is different from normal distribution)

The sample size is 23 < 30

* + - * + For states off the coast: (shapiro test)

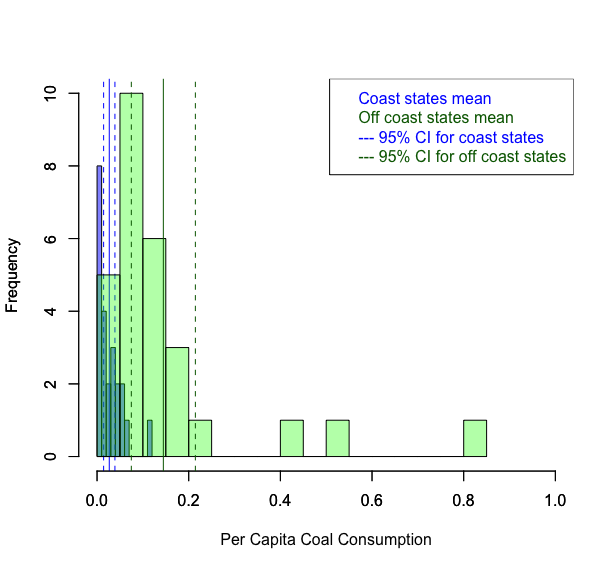
p-value = 3.627e-05 < 0.05 --> reject the H0 that the distribution is not different from normal distribution -->(the distribution is different from normal distribution)

The sample size is 28 < 30

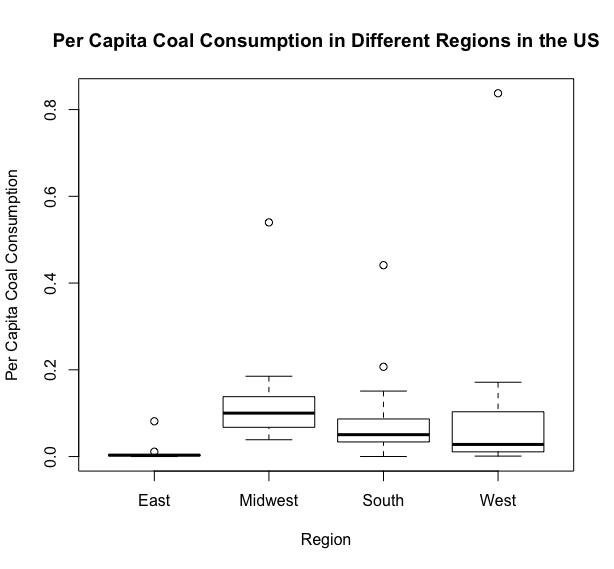
1. Equal variance: **Yes.**

* p-value = 0.5098 > 0.05, cannot reject the H0 (same variance) --> (the variances of the two data are not different)
  1. Please run the statistical test and interpret the result (1 point).
     + The result of Wilcoxon-Mann-Whiteney U test:   
       p-value = 0.008417 < 0.05
     + I reject the H0 that the total energy consumptions are not different between states on the coast and states off the coast
     + The result suggests that the total energy consumption are different between states on the coast and states off the coast)

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).
      * H0: The per capita coal consumptions do not differ between states on the coast and states off the coast.
      * Ha: The per capita coal consumptions are different between states on the coast and states off the coast.

* 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).
     + Statistics to use: Resampling
     + Assumptions for resampling: no assumption.
  3. Please run the statistical test and interpret the result (1 point).
     + The result of resampling: p-value = 0
     + P-value << 0.05 means that we can reject H0 and the per capita energy consumption are different between the states on the coast and the states off the coast.

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).
      * H0: The per capita coal consumptions do not differ among the regions
      * Ha: At least one region has per capita coal consumptions that is different among the regions.

* 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).
     + Statistics to use: At first I hope to run an one-way ANOVA, but I changed to run a Kruskal-Wallis test because the data are not normally distributed.
     + Assumptions for one-way ANOVA

1. Random sampled: **Unknown**
2. Independent observation: **Unknown**
3. Normal distribution: **No.** --> Kruskal-Wallis test
   * + - For region South (shapiro test)

p-value = 5.689e-05 < 0.05 --> reject the H0 that the distribution is not different from normal distribution   
🡪(the distribution is different from normal distribution)

* + - * For region West (shapiro test)

p-value = 1.103e-05 < 0.05 --> reject the H0 that the distribution is not different from normal distribution   
--> (the distribution is different from normal distribution)

* + - * For region East (shapiro test)

p-value = 5.535e-06 < 0.05 --> reject the H0 that the distribution is not different from normal distribution   
--> (the distribution is different from normal distribution)

* + - * For region Midwest (shapiro test)

p-value = 0.0001758 < 0.05 --> reject the H0 that the distribution is not different from normal distribution   
--> (the distribution is different from normal distribution)

1. Equal variance: **Yes.**
   * + The result of Levene Test:   
       P-value = 0.5202 > 0.05 --> cannot reject H0 (same variance among different regions)

--> (the variances of different groups are the same)

* 1. Please run the statistical test and interpret the result (1 point).
     + The result of Kruskal-Wallis test:

P-value = 0.00042 < 0.05 --> reject H0

Based on the result, the per capita coal consumptions are different among regions.

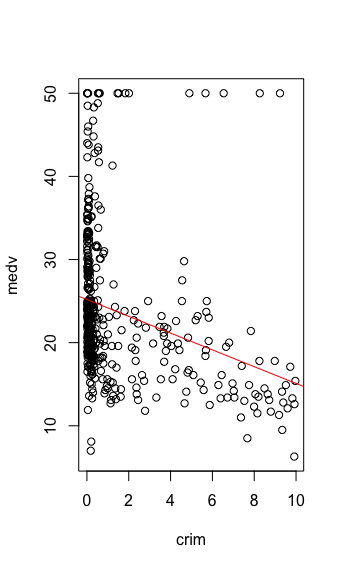
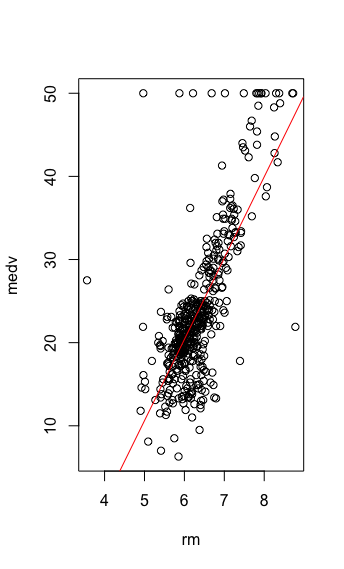
1. What is the correlation between ***per capita*** coal use and ***per capita*** GDP? Does this seem like a strong correlation to you? Why or why not? (2 points)
   * + The correlation coefficient of per capita coal consumption and per capita GDP is 0.03598182 which means there is a weak, positive relationship between the two variables.
     + I think the correlation is not strong because the correlation coefficient could range from -1 (strong and negative) to +1 (strong and positive) and 0.03598182 indicates a very weak positive relationship.

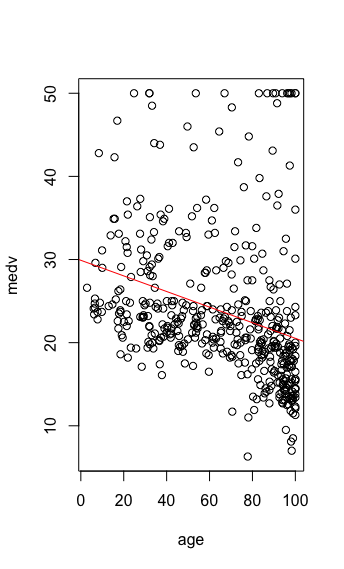
For questions 5-9, please use the ‘housedata.csv’ dataset that shows housing information for the Boston area. Information on what each of the variables are can be found here: <http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names>. In this exercise, the goal is to create a multiple 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).
   * + Three covariates: crim, rm, and age
     + Correlation coefficient:
       - The correlation coefficient of crim and rm: -0.1424577;   
         |-0.1424577|< 0.5
       - The correlation coefficient of crim and age: 0.4476638; 0.4476638 < 0.5
       - The correlation coefficient of rm and age: -0.1878709;

|-0.1878709| < 0.5

* + - The VIF is 2.439356 < 10.

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).
   1. medv~crim
      * Based on the plot, I think crim (per capita crime rate) might have a negative impact on medv (the median value of owner-occupied homes). Because there is a trend in the data that the higher the crim is, the lower the medv is.
      * 
   2. medv~rm
      * Based on the plot, I think rm (average number of rooms per dwelling) would have a positive impact on medv (the median value of owner-occupied homes). Because there is a trend in the data that the higher the rm is, the higher the medv is.
      * 

* 1. medv~age
     + Based on the plot, I think age (proportion of owner-occupied units built prior to 1940) might have a slightly negative impact on medv (the median value of owner-occupied homes). Because there is a trend in the data that the higher the age is, the lower the medv is, but some of the data do not fit the line pretty well.
     + 

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 the dependend variable and the independent variables:   
Based on the plots in question 6, there seems to be a linear relationship.

Homoscedasticity of the errors/residuals: **No.**   
The result of bp test:   
p-value = 8.811e-07 < 0.05 --> reject H0 (the errors have constant variance) -->(Heteroscedasticity)

Independence of the errors/residuals: **No.**The result of dw test:   
p-value < 2.2e-16 < 0.05 --> reject H0 (There are no autocorrelation presented in the errors) --> (There are autocorrelation presented in the errors)

Normality of the error/residuals distribution: **No.**The result of shapiro test:   
p-value < 2.2e-16 < 0.05 --> reject H0 (The errors are normally distributed) --> (The errors are not normally distributed)

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).
   * + Interpretation: (Note that some of the assumptions are violated which could result in inaccurate p-values for the regression coefficients and indicates that there is room for improvement in the model)
     + Intercept
       - the estimate for the Intercept is -32.07178, which is the median value of owner-occupied homes when the per capita crime rate, the average number of rooms per dwelling, and the proportion of owner-occupied units built prior to 1940 all = 0, and the intercept for the regression line.
       - The p-value <2e-16 < 0.05, which means the intercept is significantly different from 0

* + - Crim
      * the estimate for the crim is -0.49045, which is the effect of the per capita crime rate on the median value of owner-occupied homes when controlling for other variables that every unit of increase in per capita crime rate would decrease the median value of owner-occupied homes by 0.49045 unit.
      * The p-value = 4.93e-05 < 0.05, which means this effect is significantly different from 0.
    - Rm
      * The estimate for the rm is 9.25306, which is the effect of the average number of rooms per dwelling on the median value of owner-occupied homes when controlling for other variables that every unit of increase in the average number of rooms per dwelling would increase the median value of owner-occupied homes by 9.25306 unit.
      * The p-value < 2e-16 < 0.05, which means this effect is significantly different from 0.
    - Age
      * the estimate for the age is -0.03322, which is the effect of the proportion of owner-occupied units built prior to 1940 on the median value of owner-occupied homes when controlling for other variables that every unit of increase in the proportion of owner-occupied units built prior to 1940 would decrease the median value of owner-occupied homes by 0.03322 unit.
      * The p-value = 0.00202 < 0.05, which means this effect is significantly different from 0.

1. Discuss the fit of your model and whether you think it is a good or bad fit. Why (2 points)?
   * + The p-value for the F-statistic is < 2.2e-16 < 0.05, which means that the model is significantly different from "no model"/randomness. Or, the probability that "no model" (the intercept) will have the result of the model is < 2.2e-16.
     + However, the Adjusted R-squared is 0.5873, which means the model explains 58.73% of the variance in the data. This indicates that the model fit the data fairly but not very well because 41.27% of the variance of the data cannot be explained by the model.
     + In addition, the assumption of homoscedasticity of errors, independence of errors, and normality of the error distribution are all violated which indicates that there is room for improvement in the model.
     + Therefore, in conclusion, I would say that the model does not fit the data well.