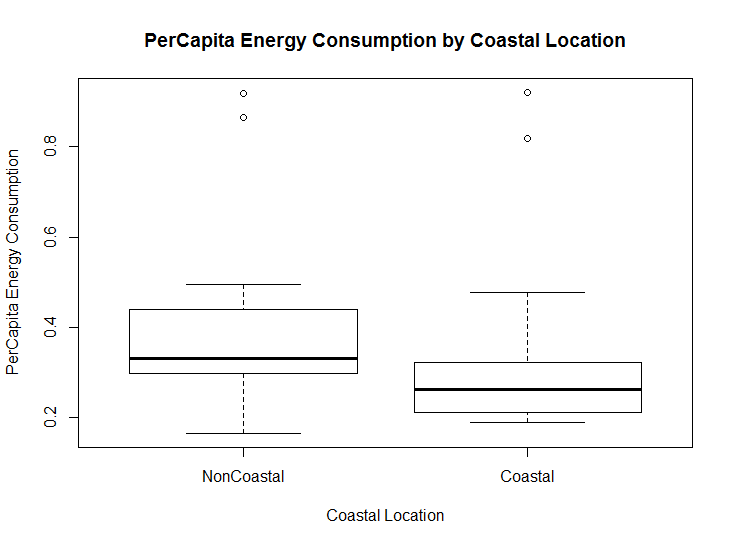
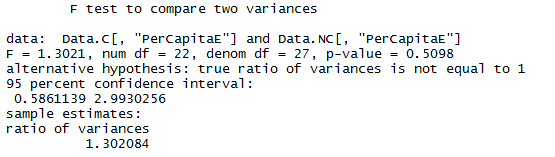
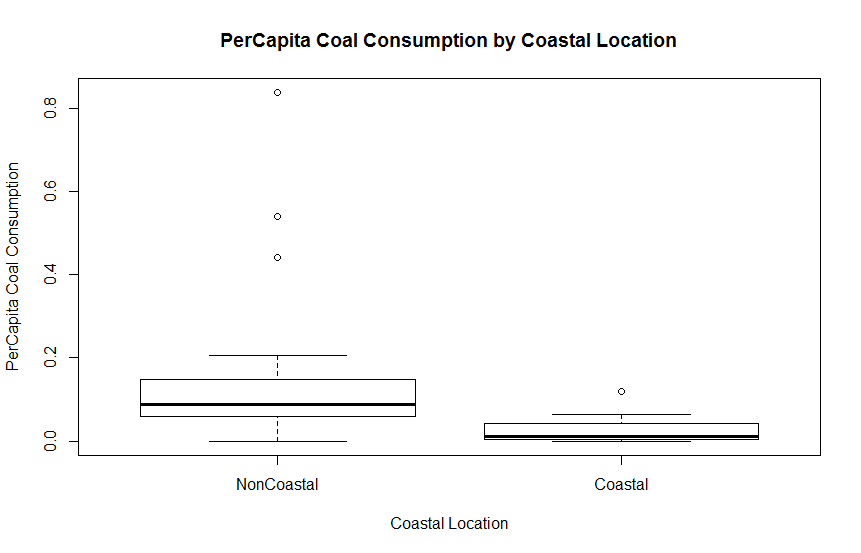
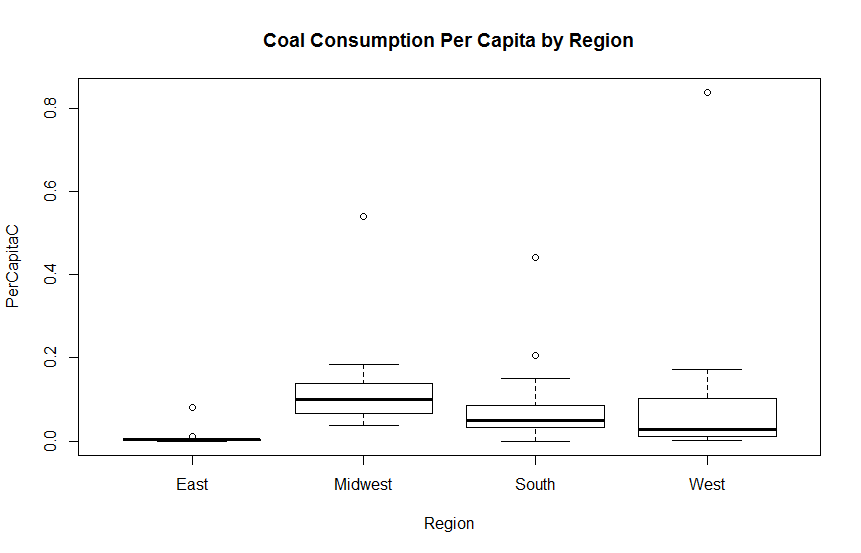
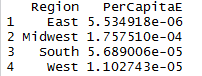
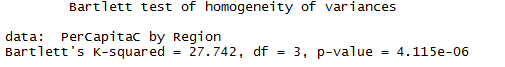
Please use the R script provided to load data and build your script from there.

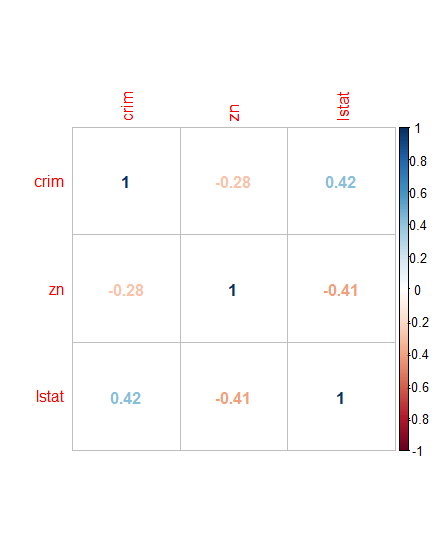
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized 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.

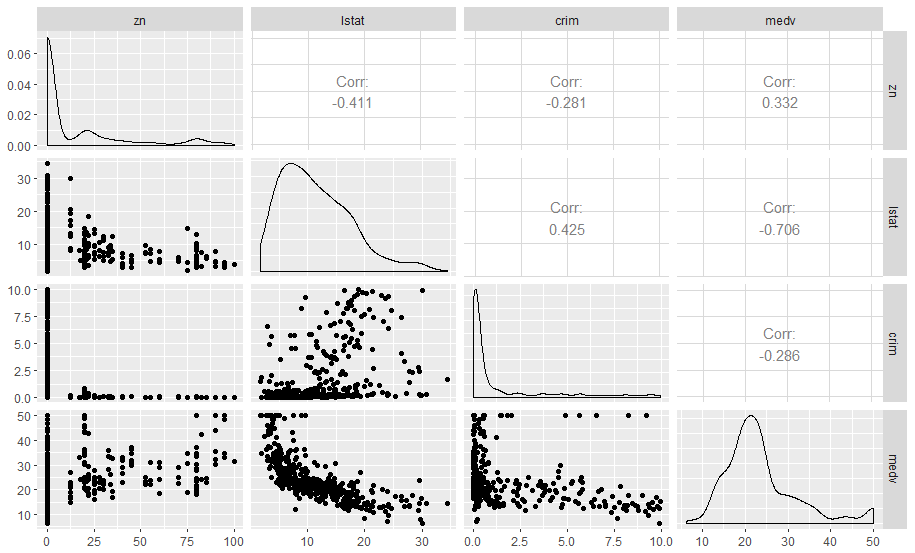
Variables needed– per capita energy consumption - total energy/population

* Per capita coal consumption – total coal/pop

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. H0 : Per Capita energy consumption does not depend on whether a state is found on the coast or not
      2. Ha : Per Capita energy consumption differs depending on whether a state is found on the coast or not
   2. Please create a visual plot to answer this question (1 point).
   3. 
   4. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
      1. Based on visual representation, a T.Test was chosen to compare the differences of energy consumption between the two groups. However due to small size and non-normality a non-parametric t.test the Mann Whitney will be run.
      2. Assumptions for
2. Continuous Data
   1. Data is continuous.
3. Testing for normality
   * + 1. A Shapiro Wilks test was run to test normality, Coastal States (p-value=5.0367e-06) and NonCoastal States (p-value=3.627e-05) had significant p-values, the alternative hypothesis could be accepted indicating that we can reject the null hypothesis for normality.
       2. 
4. Testing for Equal Variance
   * + 1. An F-test was run to determine equal variance, the p value of the F test is p=0.5098 meaning that we cannot reject the null, meaning that there is no significant difference between the two variances
       2. 
   1. Please run the statistical test and interpret the result (1 point).
      * 1. Due to the small sample size and normality variation a Mann-Whitney U was performed. The test resulted in a p-value of 0.008417 which indicated that we could reject the null and that the distribution between the two groups was not equal.
5. Does ***per capita*** coal consumption differ depending on whether a state is found on the coast or not?
6. Please write the null and alternate hypothesis (1 point).
   * 1. H0 : Per Capita coal consumption does not depend on whether a state is found on the coast or not
     2. Ha : Per Capita coal consumption differs depending on whether a state is found on the coast or not
7. Please create a visual plot to answer this question Please create a visual plot to answer this question (1 point).
8. 
9. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
   * 1. Based on violations of normality and variance for a parametric t.test, a non-parametric Kruskal Wallis will be used.
     2. Three assumptions are needed to determine that a non-parametric test is needed
        1. Data is continuous- yes
        2. Data follows a normal distribution
           1. A Shapiro-Wilks test was run to check for normality in the groups. Noncoastal (p=6.210e-07) and Coastal (p=1.02e-03) both had significant p-values, indicating that we can reject the null hypothesis for normality.
           2. 
           3. Data is not normally distributed
        3. Equal Variance
           1. F Test was ran to compare the variances. A P value of 1.966e-05 was obtained. Indicating that we can reject the null hypothesis for equal variance.
           2. Data is varied
10. Please run the statistical test and interpret the result (1 point).
    1. Kruskal-Wallis test had a p-value of 1.966e-05, indicating that I can reject the null hypothesis. At 0.05 significance level I can conclude that the consumption of coal per capita between the coastal and non-coastal states are non- identical
11. Does ***per capita*** coal consumption differ depending on the region in which a state is found?
12. Please write the null and alternate hypothesis (1 point).
    * 1. H0: Per Capita coal consumption does not differ depending on the region in which state it is found
      2. Ha Per Capita coal consumption does differ depending on the region in which a state is found.
13. Please create a visual plot to answer this question (1 point).
    1. 
14. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
    1. Due to the multiple groups in this data, small sample size, non-normality, and unequal variance a non-parametric test for ANOVA will be used – Kruskal-Wallis will be ran.
    2. The assumptions for this test are; each sample is independent, the distribution follows a normal distribution, the population variations are equal
       1. A Shapiro test was ran to test normality. All regions showed significant p values (E=5.534e06, MW=1.757, S=5.69e-05, and W=1.102e-05). The null hypothesis can be rejected; the data is not normally distributed.
       2. 
       3. Bartlett’s test was run to determine variation among the groups. We can reject the null hypothesis that the variances are equal (p=4.115e-06). Variances are not equal among the groups.
       4. 
15. Please run the statistical test and interpret the result (1 point).
    1. The results from the Kruskal Wallis gave a p-value of 0.00042, signifying that I can reject the null hypothesis. At 0.05 significance level I can conclude that the consumption of coal per capita between the regions are non- identical
    2. 
16. 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. Running a Kenall’s rank correlation analysis on the two groups, indicates that I cannot reject the null hypothesis (p=0.09752) that the variables are uncorrelated at the 0.05 level. The correlation coefficient for the groups are small. This indicates that there is not a linear correlation among the two groups. In addition, a pair plot of the two show a small clustering but no visible linear similarities.

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. The three variables selected from the data set were lstat, zn, and crim. After doing a correlation test, the linear relationships among the variables demonstrated weak-moderate relationships
          1. Lstat and zn had a correlation coefficient of -0.411 indicating a weak negative linear relationship
          2. Zm-cim had a correlation coefficient of -0.281 indicating a weak negative linear relationship
          3. Crim-stat have a 0.425 indicating a moderate positive relationship
   1. 
      * 1. When checking for multi-collinearity with the VIF values were zn=1.22, lstat=1.375, and crim=1.240873 indicating moderate correlation.
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).



1. Zn (proportion of residential land for lots over 25,000) appears to have little correlation with that of medv (median value). Lstat (% of lower status in the population) appears to have a high negative correlation to medv. Crim (per capita crime) and medv appear to have a slight negative correlation (not very strong) but there appears to be some variance in the chart.
2. 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. A use of the gvlma test indicated that many assumptions are not met for the linear regression including, skewness, kurtosis, link function, and heteroscedasticity. This was further enforced through multi plotting of residuals, QQ, and scale location plots.
         2. One way to overcome some of these assumption would be the removal of some of the higher outliers. A second part that would help possibly resolving some of the assumptions would be the transform the data around the mean. The scale location plots show high variance among the groups. Together those should improve the fit of the model to the line.
3. 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. For any given level of zoned land and per capita crime, improving one point of the median value of occupied homes will see a decrease in the percentage of lower status of the population by -0.099 (p value of 2e-16). All other groups being zero will see an increase in median value of occupied homes by a value of 34.76 (p=2e=16). There was no significant change for zoning of land (est=0.01, p-value=0.155) or per capita crime ( est=0.08, p-value=0.525).
4. Discuss the fit of your model and whether you think it is a good or bad fit. Why (2 points)?
   * + - 1. Due to the lack of assumptions met for the model I do not think it is a good fit. The R squared for the test was 0.49 which also indicated that the test does explains some variability but not much.