If you run any ANOVAs, you can use the Levene test for equality of variances. If your data violate an assumption about normality and a normal distribution is required for your analyses, you can get bonus points for transforming your data. Otherwise please run the statistical test anyway as if your data were normally distributed but make it clear that you violated this assumption in your answer.

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

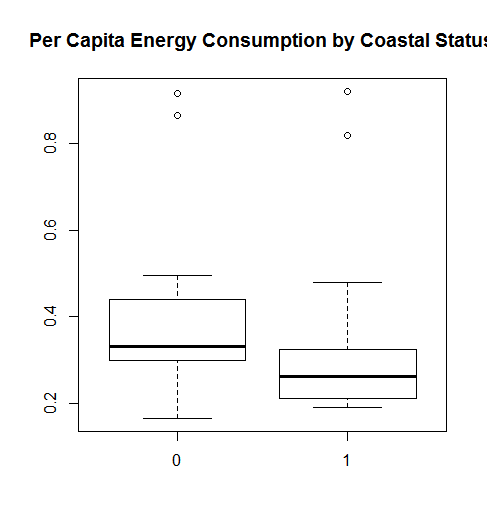
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized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

**Alternative hypothesis**: States found on coasts will have greater or less per capita energy consumption than states not on a coast.

**Null hypothesis**: States found on coasts will not have a significantly different per capita energy consumption than states not on a coast.

* 1. Please create a visual plot to answer this question (1 point).



0 = Inland

1 = Coast

* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

First I added a new column to the dataframe that divided total energy usage by population to give me a per capita energy usage. The variances are equal after an F-test. Then I checked for normality by running shapiro-wilkes tests on both samples and found that neither are normally distributed. The samples are independent of each other.

As such I decided to use a Wilcoxon-Mann-Whitney U test.

* 1. Please run the statistical test and interpret the result (1 point).

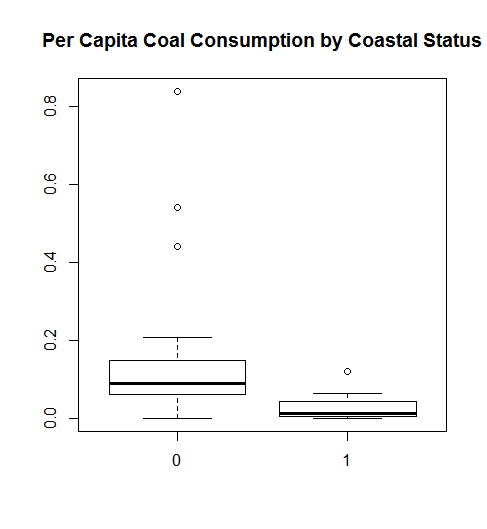
Running the Wilcoxon-Mann-Whitney U test gave me a significant p-value of 0.008417 and as such I can reject the null hypothesis. There is a significant difference in per capita energy usage if the state is coastal.

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).

**Alternative hypothesis**: States found on coasts will have greater or less per capita coal consumption than states not on a coast.

**Null hypothesis**: States found on coasts will not have a significantly different per capita coal consumption than states not on a coast.

* 1. Please create a visual plot to answer this question (1 point).



0 = Inland

1 = Coast

* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

First I added a new column to the dataframe that divided total coal usage by population to give me a per capita coal usage. The variances are not equal after an F-test. Then I checked for normality by running shapiro-wilkes tests on both samples and found that neither are normally distributed. The samples are independent of each other.

As such I decided to use a Wilcoxon-Mann-Whitney U test.

* 1. Please run the statistical test and interpret the result (1 point).

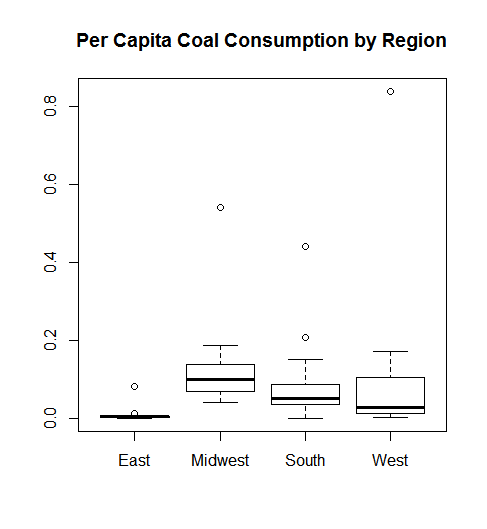
After running the Wilcoxon-Mann-Whitney U test I got a significant p-value of 2.051e-5 and as such I can reject the null hypothesis. There is a significant difference in per capita energy usage if the state is coastal.

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).

**Alternative hypothesis**: Per capita coal consumption is significantly different in at least one of the regions.

**Null hypothesis**: Per capita coal consumption is equal among all regions.

* 1. 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).

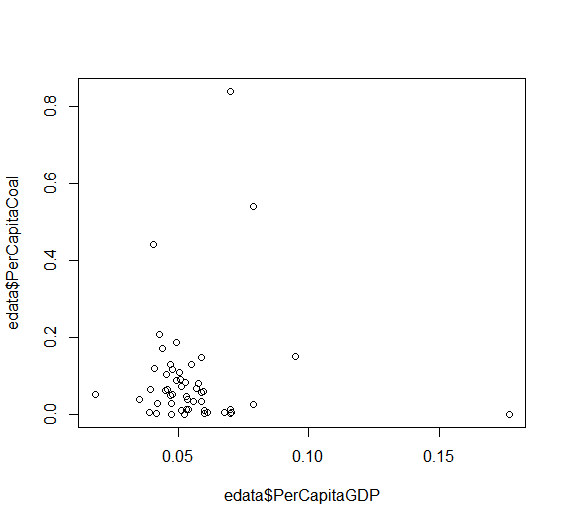
As I am comparing four different groups I decided to us an one-way ANOVA. However, after running shapiro-wilkes tests it is clear that none of the groups are normally distributed so I decided to run a Kruskal-Wallis rank sum test. I ran a Levene Test to determine if they have equal variances and got a p-value of 0.5202 so I can assume equal variances. The samples are independent of each other.

* 1. Please run the statistical test and interpret the result (1 point).

Upon running the Kruskal-Wallis I found a p-value of 0.00042 indicating that region does play a significant role in the per capita usage of coal.

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)

To test correlation I first made a plot to visualize the data and then I ran a Pearson correlation to determine the correlation coefficient. First I saw on the plot a distinct lack in visual correlation. This was confirmed when I saw the correlation coefficient is 0.03598182. This is a very weak correlation as strong correlations are closer to 1.0.

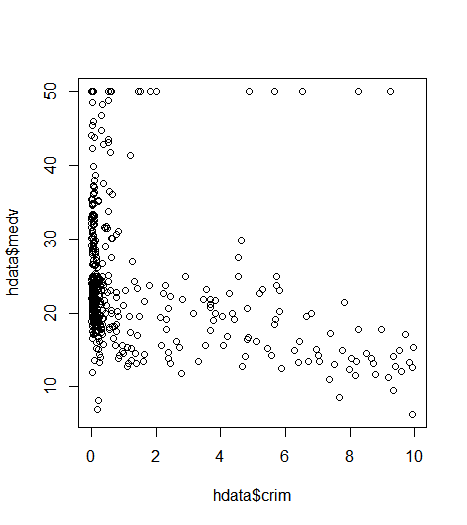
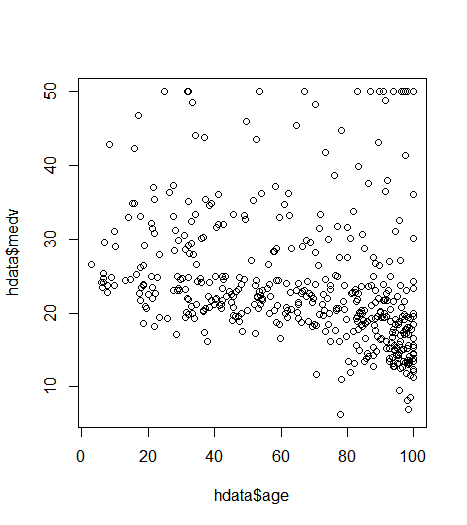


For questions 5-9, please use the ‘housedata.csv’ dataset that shows housing information for the Boston area. Information on what each of the variables are can be found here: <http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names>. In this exercise, the goal is to create a multiple 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.

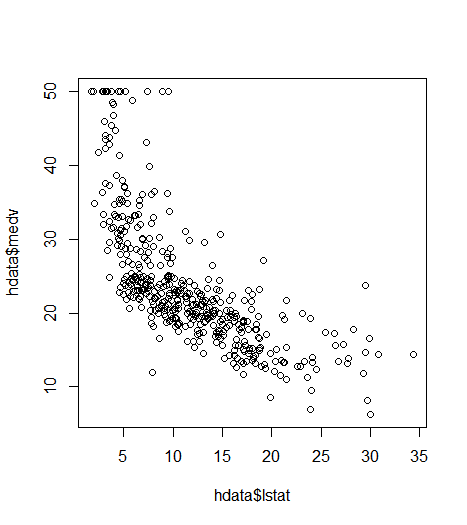
1. Please select three covariates that you will include in your model as independent variables. Please check if these variables are highly correlated with one another to make sure you do not run into problems of multi-collinearity. Check if this model has issues with multi-collinearity using the variance inflation factor. **Report correlation values and VIF values in your answer** (3 points).

The three variables I selected where “crim” (per capita crime rate by town), “age” (proportion of owner-occupied units built prior to 1940), and “lstat” (% lower status of population). After running a Pearson correlation test, the highest correlation rate I found between these variables was 0.5732663. As such I will not have issues with multi-collinearity. I also ran a VIF and found the following values: crim->1.320170, age->1.611567, lstat->1.572323. With these values it is clear that they are only very moderately correlated. Not enough to cause issues of multi-collinearity.

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).



Age Crim



Lstat

Based on the three plots it seems that age has a negative relationship with median housing price. As the proportion of owner-occupied units built prior to 1940 increases, the median housing price decreases but only slowly. As per capita crime rate by town increases median housing prices decreases, sharply at first and then slower. As the % of lower status of the population increases the median housing prices decrease, sharply at first and then slower.

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).

I first checked for normality by making qq-plots of all three independent variables. Crim and age are not normally distributed but lstat is just about normally distributed. I didn’t know how to transform it so I didn’t. I checked for independent errors and I checked homoscedasticity by sight on my plots. Lstat and age seem relatively homoscedastic but crim seems more heteroscedastic or just has a number of outliers.

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).

Coefficients:

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

(Intercept) 33.44194 0.78362 42.676 < 2e-16 \*\*\*

crim -0.08460 0.13333 -0.634 0.526080

age 0.05125 0.01307 3.921 0.000102 \*\*\*

lstat -1.13017 0.05899 -19.158 < 2e-16 \*\*\*

The intercept coefficient means that when crim, age, and lstat are all 0, the median housing price value is 33.44194. The crim coefficient is not significant but indicates a small decrease in medv with every increase in crim. Age is significant and shows a 0.05125 increase in medv for every 1 increase in age. Lstat is also significant and shows a **decrease** of 1.13017 medv for every 1 increase in lstat.

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

The Multiple R-squared for my model is 0.5156. As such this model explains about 51.6% of the variability of the data. While not terrible, this is not a good model. The closer to 100%, the better the model.