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

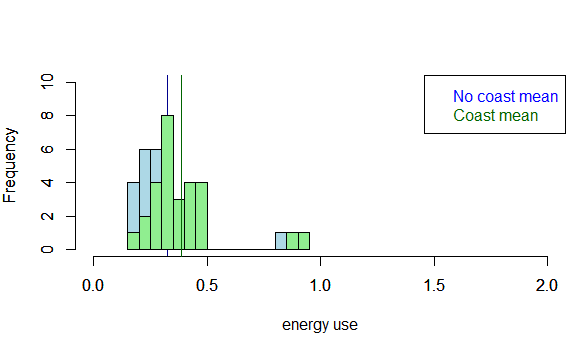
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized 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: Energy consumption does not differ depending on whether a state is found on the coast or not.**

**HA: Energy consumption does differ (greater or less) based on whether a state is found on the coast or not.**

* 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).

**1) Data are continuous**

**2) Sample is randomly selected from the population**

**3) Independent observations**

**4) Values are nearly normal or the sample size is sufficiently large**

**5) Equal variance between 2 populations**

**Test for normality:**

**shapiro.test(nocoast\_vec)**

**shapiro.test(coast\_vec)🡪 Does not pass test for normality; however the sample size is sufficiently large to still use parametric analysis (n > 30). So there is no real need to run the Wilcoxon:**

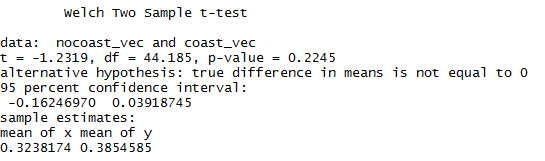
**wilcox.test(nocoast\_vec,coast\_vec, data=edata)**

**test for equal variance:**

**var.test(nocoast\_vec, coast\_vec)🡪Passes test for equal variance**

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

**t.test(nocoast\_vec,coast\_vec, paired=FALSE)**



**Based on the p-value for the t-test, *we cannot reject the null hypothesis* that energy consumption does not differ depending on whether a state is found on the coast or not.**

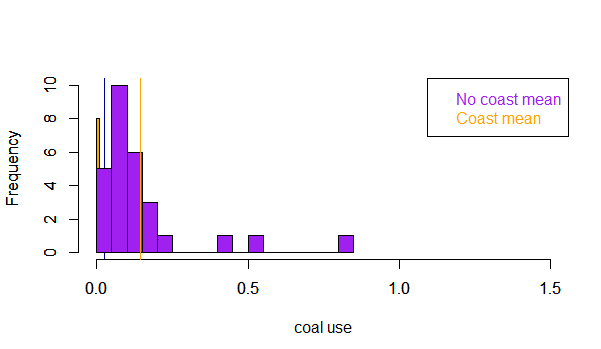
1. Does ***per capita*** coal consumption differ depending on whether a state is found on the coast or not?
   1. Please write the null and alternate hypothesis (1 point).

**H0: Per capita coal consumption does not significantly differ depending on whether a state is found on the coast or not.**

**HA: Per capita coal consumption significantly differs based on**

**whether a state is found on the coast or not.**

* 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).

**Test for normality:**

**shapiro.test(nocoast\_vec1)**

**shapiro.test(coast\_vec1)**

**#Does not pass test for normality; however, sample size is sufficiently large to run a parametric test (n>30), so there is no real need to run a non-parametric test (e.g. Wilcoxon).**

**wilcox.test(nocoast\_vec1, data=edata)**

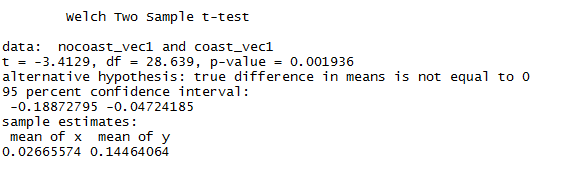
**wilcox.test(coast\_vec1, data=edata)**

**Test for equal variance**

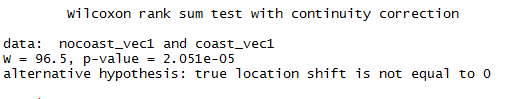
**var.test(nocoast\_vec1, coast\_vec1)**

**#Does not pass test for equal variance**

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



**From the t-test, we may *reject the null hypothesis* that there is no significant difference in coal use between coastal and non-coastal states.**



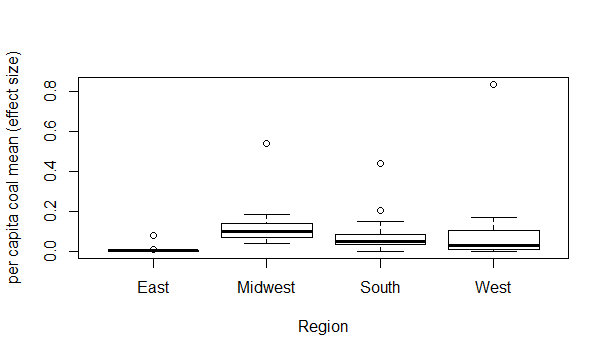
**The Wilcoxon Mann Whitney test confirms that there is a significant difference between coastal and noncoastal state energy consumption.**

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: Per capita coal consumption does not differ depending on the region in which a state is found.**

**HA: For at least one region, per capita coal consumption does differ depending on the region in which a state is found.**

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



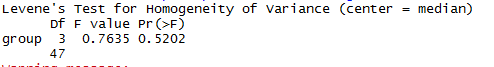
* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

**ANOVA:**

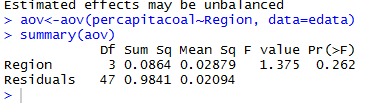
* + 1. **Data normally distributed**

**Even though the Shapiro-Wilk normality test gives p-values less than 0.05, which means that we could reject the null hypothesis that data are normal, our sample size is greater than 30; therefore we can assume normality based on the Central Limit Theorem, instead of using the Kruskal-Wallis test (nonparametric).**

* + 1. **Independent cases (yes)**
    2. **Outliers (from the box plot it appears as though there are a few outliers, but they are not clustered.)**
    3. **Equal variance (yes)**



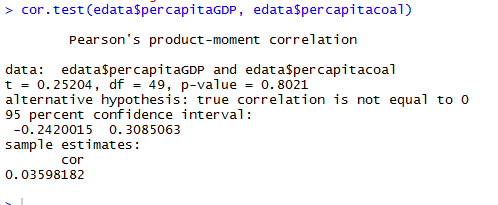
* 1. Please run the statistical test and interpret the result (1 point).



**ANOVA output *fails to reject* the null hypothesis that coal consumption does not significantly differ by region.**

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)

**edata["percapitaGDP"]<-edata$TotalGDP/edata$Population**



**The p-value is greater than 0.05 and the cor =0.036; therefore there is not a strong correlation between percapita coal use and per capita GDP.**

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).

**cor.test(hdata$crim, hdata$ptratio) = 0.32**

**cor.test(hdata$ptratio, hdata$rm) = -0.33**

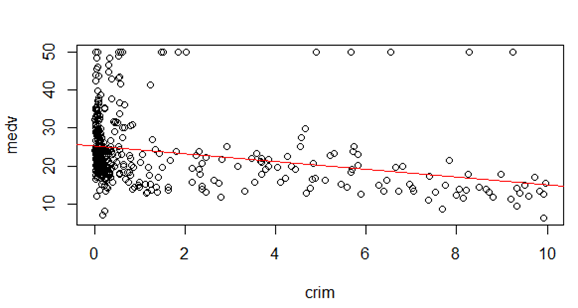
**cor.test(hdata$rm, hdata$crim) = -0.142**

**All correlation values are less than 0.5, which is an appropriate threshold to assume low correlation between my three chosen covariates.**

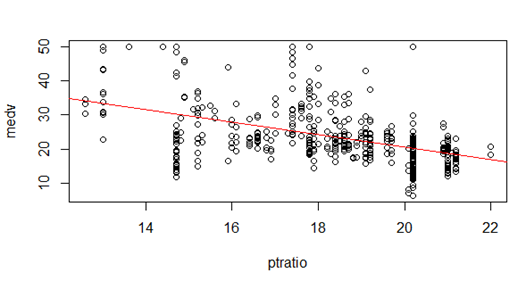


**Because the VIF is smaller than 5, we can say that these covariates are only slightly correlated, and therefore it is not necessary to remove any of them from our multilinear model.**

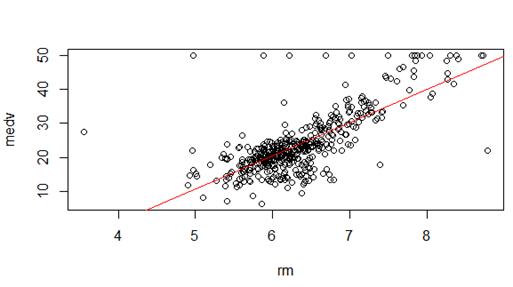
1. Plot the relationship between each of your three independent variables and the dependent variable (medv). **Include each plot in this answer and state whether and how you think each variable is related to median housing prices** (medv; 3 points).



**Although there is generally a negative correlation, the x-values are skewed, such that with lower per capita crime rates, there is greater variability in housing prices; while higher crime rates appear to be clustered with mid-low housing prices (which is not surprising).**

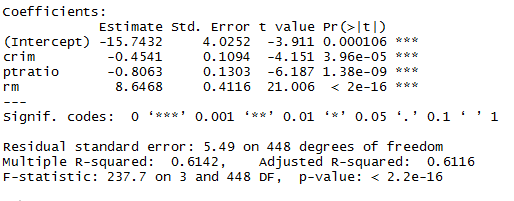


**From this plot it appears as though there is a slight/weak negative correlation between pupil-teacher ratio by town and housing value prices (when pupil-teacher ratios increase, housing prices decrease🡪 this is not surprising).**

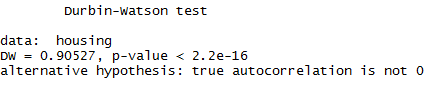


**From this plot, it appears as though average number of rooms per dwelling is positively correlated with housing value prices, especially at moderate value pricing and medium (about 6) number of rooms. In other words, mid-range valued houses do not generally have few (4) or many (8) rooms; this is not surprising.**

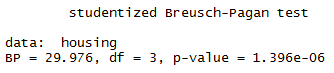
1. Run your multiple linear regression model. Check whether any assumptions are violated. Please state **which assumptions** you checked, **whether they were violated**, and **how you know** whether or not they were violated. If any assumptions are violated (e.g. normality), we will give you bonus points if you are able to identify a way to overcome this problem (3 points, plus additional 1 point bonus).



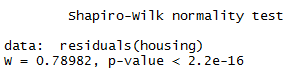
**1. Residual independence (dwtest) 🡪 Does not pass. Autocorrelation exists.**



**2**. **Residual homoscedasticity (bptest) 🡪 Does not pass. Residuals are heteroscedastic.**

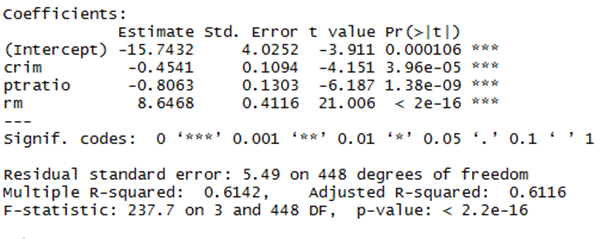


**3. Residual normality (shapiro.test) 🡪 Does not pass. Residuals are not normal.**



**4. There is a linear relationship between variables**

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).



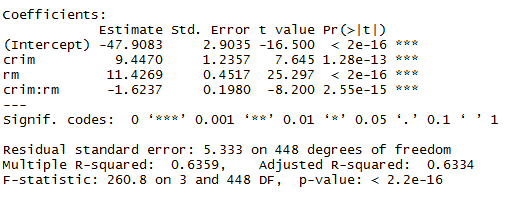
**Intercept: With all three covariables held constant (crime, pupil-teacher ratio, and room coefficient), median home value decreases by 15.743(1000) = $15743.**

**Crime (crim) coefficient (significant): For every increase in per-capita crime rate, the median value of owner occupied homes *decreases* by 0.45($1000) = $450.**

**Pupil-Teacher (ptratio) ratio coefficient (significant): For every increase in the pupil to teacher ratio, the median value of owner occupied homes *decreases* by 0.81($1000) = $810.**

**Rooms (rm) coefficient (significant): For every increase in the average number of rooms per dwelling, there is a 8.65($1000) = $8650 *increase* in median value of owner occupied homes.**

BONUS



**The coefficient for the interaction term suggests that per capita crime and average number of rooms per dwelling have a significant effect on each other. That is, the effect of crime on housing prices in a town varies significantly based on the average number of rooms per dwelling. (This interaction term would not make logical sense in the real world.)**

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

**IF my data were normal *(which they are not🡪 residuals show pattern, indicating that linear model is NOT the best fit),* based on my adjusted R-squared value (0.61), the low VIF, and the p-values for my coefficients, I think that my model is a good fit for predicting housing value prices; however, it could possibly be improved with the addition of more variables or a different combination of selected variables (in reality, number of rooms, crime rate, and pupil-teacher ratio may not be the MOST influential factors in determining housing value). A log transformation or square root transformation of the data may improve model fit:**



