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

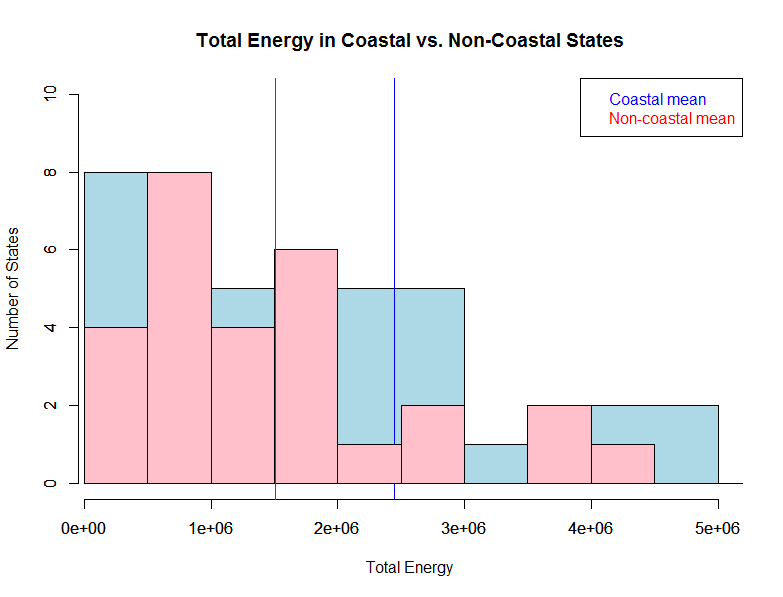
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized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

Alternate: Energy consumption in coastal states differ from non-coastal states.

* 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).

The data is neither normally distributed nor do they have equal variances (please see R-script for results). Assume normality since sample sizes are not too small (23,28). Use Welch’s T-test, considered robust against unequal variance.

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

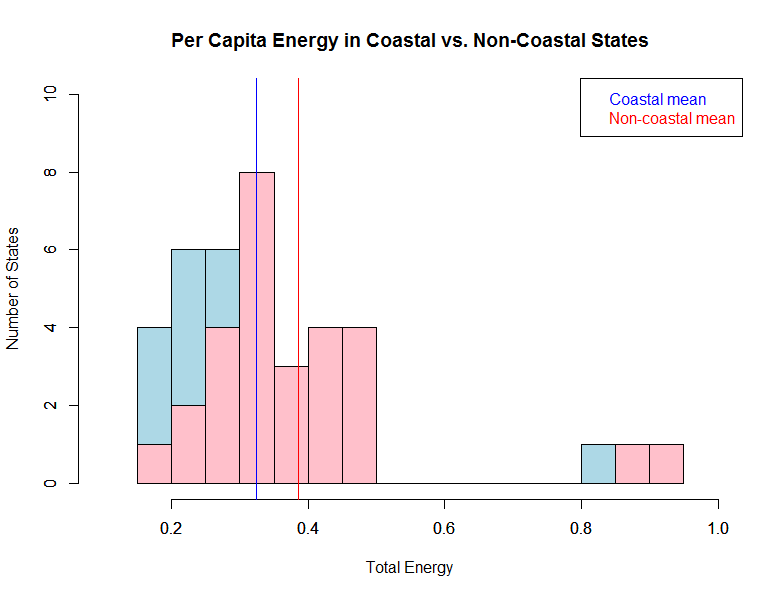
P-value=0.1494, not able to reject null-hypothesis. There is no difference in total energy consumption in coastal vs non-coastal states.

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).

Null: Per capita energy consumption in coastal states does not differ from non-coastal states.

Alternate: Per capita energy consumption in coastal states differ from non-coastal states.

* 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).

The data is not normally distributed, variance can be assumed equal (please see R-script for results). Samples sizes not too small (23, 28). Use a Welch’s t-test.

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

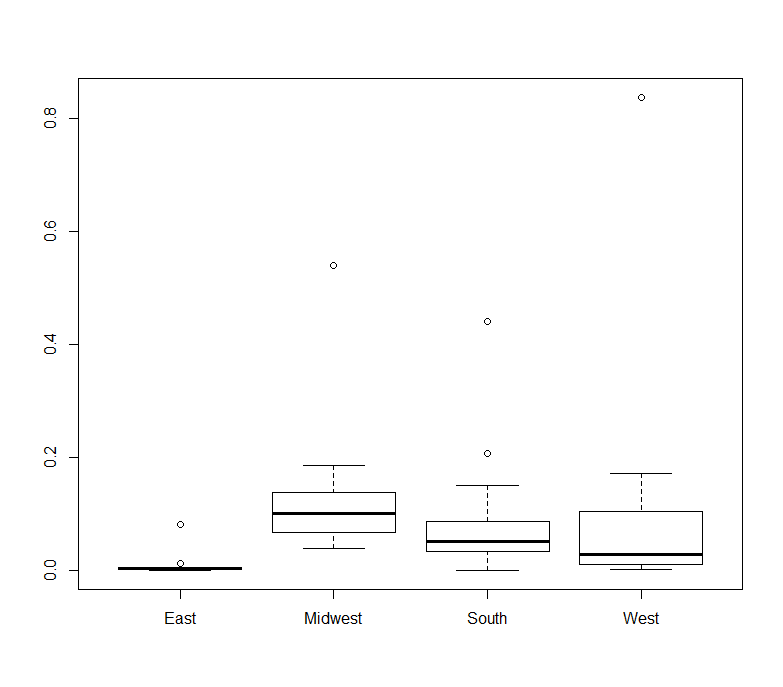
P-value=0.2245, cannot reject null-hypothesis. Per capita energy consumption is NOT different in coastal vs non-coastal states.

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).

Null: Per capita coal consumption is NOT different in different regions.

Alternate: Per capita coal consumption is different in different 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).

One-way ANOVA. Check variance and normality. (See R script) Variance equal, but data is not normal. One-way ANOVA tolerates violations of normality assumptions fairly well.

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

P-value: 0.262. Regions are NOT significantly different in their coal consumption.

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)

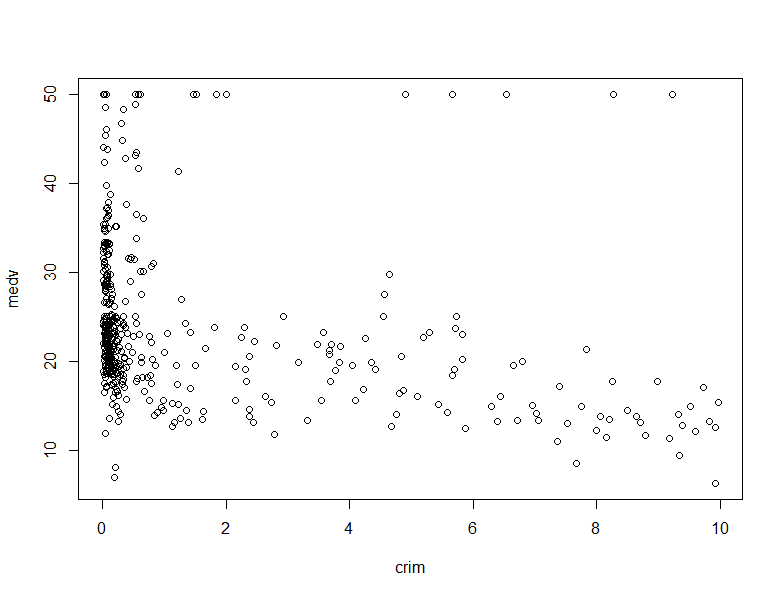
Correlation coefficient: 0.03598, indicating a very small positive correlation. This is not a strong correlation. With a correlation coefficient below the threshold of 0.5, they are considered weak.

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.

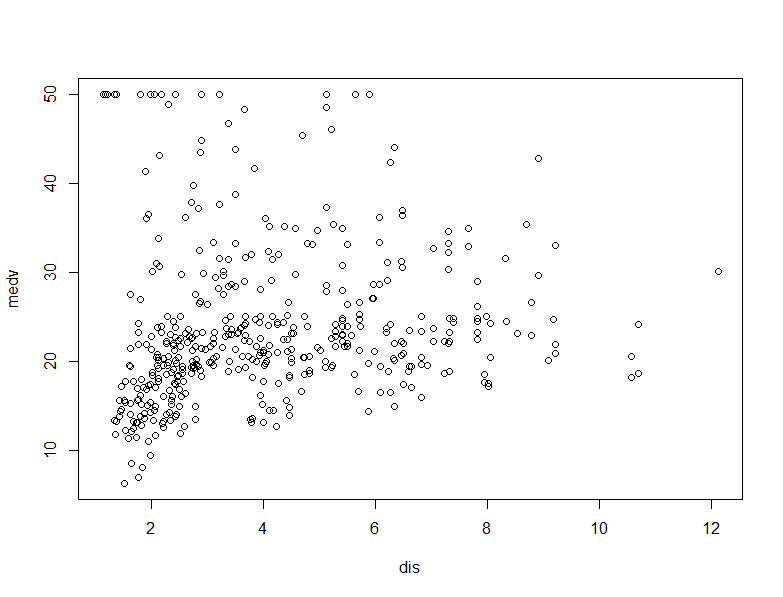
1. Please select three covariates that you will include in your model as independent variables. Please check if these variables are highly correlated with one another to make sure you do not run into problems of multi-collinearity. Check if this model has issues with multi-collinearity using the variance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

Selected variables: crime rates (CRIM), distance to employment centers (DIS) and teacher-pupil ratio [ptratio]. Correlation coefficients: -0.46, 0.32, -0.15 All below 0.5, suggesting they are not highly correlated. VIF: 1.31, <5, low and does not cause concern.

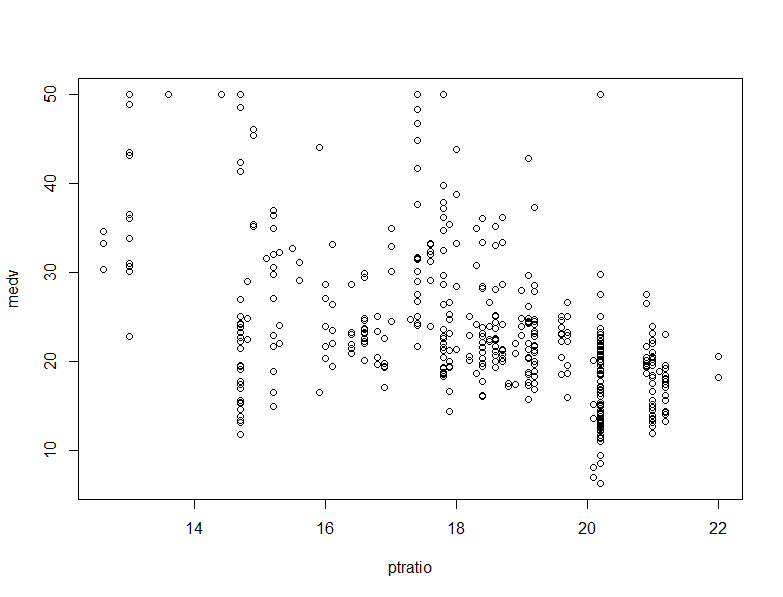
1. Plot the relationship between each of your three independent variables and the dependent variable (medv). **Include each plot in this answer and state whether and how you think each variable is related to median housing prices** (medv; 3 points).



Medv~crim: There is some relationship between crime rates and median property value, though it may not be a linear one. As crime increases, there is a trend towards lower median property values. However, there also appears to be a threshold effect – beyond the value of 1 for crime, it appears that there are very few homes above the value of $25,000.



Medv~dis: There does not appear to be a very strong/ clear relationship between distance to employment centers and median housing value. There is a vague trend towards slightly higher median housing values as distance increases, but there is also a much greater range of housing values in areas close to employment centers.



Medv~ptratio: there is a clear trend towards lower pupil-teacher ratios where median housing values are higher.

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).

Checked for independence of residuals (DW test), normality of residuals (qqplot), homoscedasticity of residuals (BP test). Results: DW test, p-value= < 2.2e-16, data is autocorrelated; qqplot shows residuals are clearly not normal, and BP test p-value = 2.316e-08, data is heteroscedastic. I log-transformed the crime and distance data which had an inverse exponential pattern to normalize them more.

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).
2. The intercept value: 54.49 indicates the value where the linear model intersects the Y-axis. Its p-value is very small, meaning that it is significantly different from zero. However, the value does not really have any meaning for the analysis that is being done here, since it doesn’t make very much sense for PT-ratio to be zero.
3. The coefficient of ‘crim’ is -0.536, indicating that there is a negative slope in the fitted line after controlling for the other factors. The model suggests that for every unit increase in crime rate, median housing prices fall by 0.536 units ($563). It is a significant relationship, as p-value is 0.00189.
4. The coefficient of ‘dis’ is 0.0248, but since its p-value is very high, there is not a significant relationship between distance from employment centers and median housing value.
5. The coefficient of ‘ptratio’ is -1.649, indicating a positive slope after controlling for other factors. P-value is extremely small, indicating a significant relationship. For every unit increase in PT-ratio, we can expect median housing value to decrease by 1.649 units ($1649).
6. Running the model with an interaction term between the two significant terms, crime and PT ratio:
   1. New intercept is 70.669
   2. With every unit increase of PT-ratio and where crime is 0, we expect a decrease of 2.402 ($2402) in median housing value. (significant)
   3. With every unit increase of crime and where PT-ratio is 0, we expect a decrease of 17.602 ($17, 602) in median housing value. This doesn’t make sense since PT-ratio cannot be 0. (significant)
   4. The effect of PT-ratio on median housing value increases by 0.8504 for every unit increase in crime rate. (significant)
7. Discuss the fit of your model and whether you think it is a good or bad fit. Why (2 points)?

Looking at the summary table of the model without interaction terms, the adjusted R-sq value is only 0.2291, meaning that the fitted line is only able to explain approximately 23% of the variation in the data. This does not seem to be a good fit to me. While median housing value is a complex subject with many unknowns that may not be easily accounted for, I would still expect to see an R-sq value upwards of 0.5 to consider it a good model.