Your script needs to be a self-contained and executable on my laptop. That being said please DO NOT download the data to your laptop and read in data locally. Please use the script I provided and build your script from there. Your quiz will **not be graded** if it is not executable on my machine.

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).
   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).
   4. Please run the statistical test and interpret the result (1 point).
2. 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).
   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).
   4. Please run the statistical test and interpret the result (1 point).
3. 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).
   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).
   4. Please run the statistical test and interpret the result (1 point).
4. 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)

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). 15 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 as your answer (3 points).
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).
3. 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 2 point bonus).
4. 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 mean not just state that coefficient is significant or not (3 points).
5. Discuss the fit of your model and whether you think it is a good or bad fit. **Why** (2 points)?