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).

H0 = Per capita energy consumption does not differ based on whether a state is found on the coast or not (difference in means is not statistically different from 0)

HA = Per capita energy consumption varies based on whether a state is found on the coast or not (difference in means is statistically different from 0)

* 1. Please create a visual plot to answer this question (1 point).
  2. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

I would use a two tailed T-test because the Independent Variable (per capita energy consumption) is continuous and the Dependent Variable (coast or not) is categorical. The sample is randomly selected and independent. The sample size is sufficiently large, greater than 30. There are equal variances.

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

The p-value is not statistically significant at .2245. We fail to reject the null hypothesis meaning that per capita energy consumption is not significantly different based 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 differ based on whether a state is found on the coast or not (difference in means is not statistically different from 0)

HA = Per capita coal consumption varies based on whether a state is found on the coast or not (difference in means is statistically different from 0)

* 1. Please create a visual plot to answer this question (1 point).
  2. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

I would use a two tailed T-test because the Independent Variable (per capita coal consumption) is continuous and the Dependent Variable (coast or not) is categorical. The sample is randomly selected and independent. The sample size is sufficiently large, greater than 30. There are equal variances.

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

The p-value is .001936. We can reject the null hypothesis, meaning that per capita coal consumption does vary based on whether a state is found on the coast or not.

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 = There is no difference in per capita coal consumption among the different regions.

HA = There is a significant difference in per capita coal consumption in at least one of the different regions.

* 1. Please create a visual plot to answer this question (1 point).
  2. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

I would use an ANOVA because there are more than two groups of the Dependent Variable (region). The population is normally distributed, the samples are independent of each other, the populations have the same variance, the sample size of the groups does not need to be equal.

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

We can reject the null hypothesis meaning that per capita coal consumption differs in at least one of the regions.

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)

There does not seem to be a strong relationship between per capita coal consumption and per capita GDP. There may be a trend towards high coal use indicating higher GDP, but this relationship is not statistically significant. By interpreting the plot, we can see that there is not a strong linear relationship given several outliers. By testing the relationship with a linear model, the p-value is not significant.

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).

Per capita crime rate (crim)

Nitric oxides concentration (nox)

Index of accessibility to radial highways (rad)

Correlation values

crim nox rad

crim 1.0000000 0.6369411 0.8979883

nox 0.6369411 1.0000000 0.5424992

rad 0.8979883 0.5424992 1.0000000

VIF values

crim nox rad

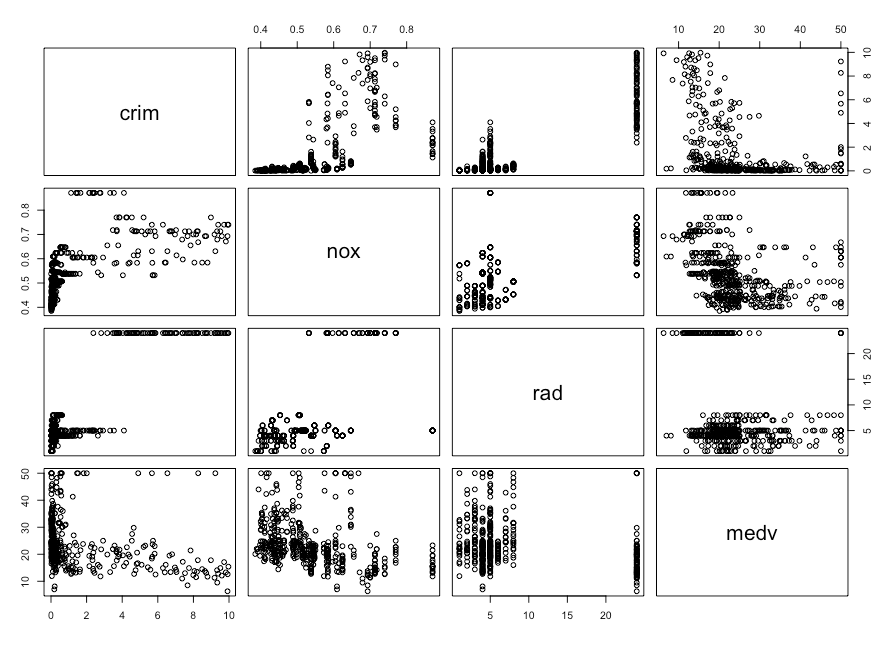
6.179491 1.695428 5.204105

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).

Crime and median housing prices appear to have a weak negative relationship.

NOx concentrations and median housing prices appear to have a negative relationship.

Accessibility to radial highways and median housing prices do not appear to be related.



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).

The assumption of a linear relationship between the variables was checked using the DW test which shows there is a linear correlation in the model.

The assumption of homoscedasticity was checked using the BP test. The p value of the BP test is greater than .05 so there is no pattern in the residual, the residual is not homoscedastic and this assumption was violated.

The assumption of normality of the error distribution was checked using the Shapiro test and qqplot. The results of these test show us that the residuals do not look normal. This assumption was violated.

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).

The intercept of medv is 33.8938. For every additional unit of crim, medv decreases by 0.9821. This relationship is significant at the .05 level, these variables have a weak negative relationship. For every additional unit of nox, medv decreases by 18.9606. This relationship is significant at the 0 level, these variables have a strong negative relationship. For every additional unit of rad, medv increases by 0.1925. This relationship is not statistically significant.

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

The model has an R2 of 0.1252504 and an adjusted R2 of 0.1194. These values indicate that this may not be the best fitting model for the data.