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

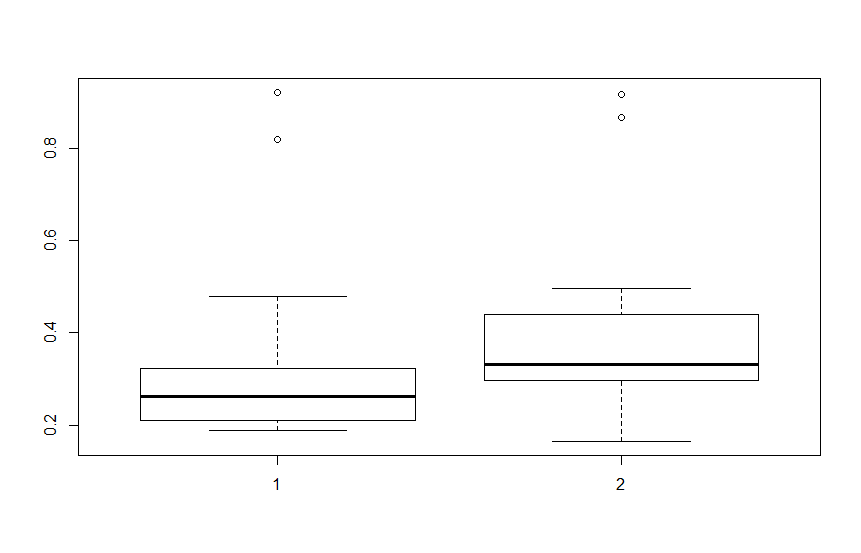
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized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

Ha: Per capita energy consumption in coastal states differs from that in non-coastal states.

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

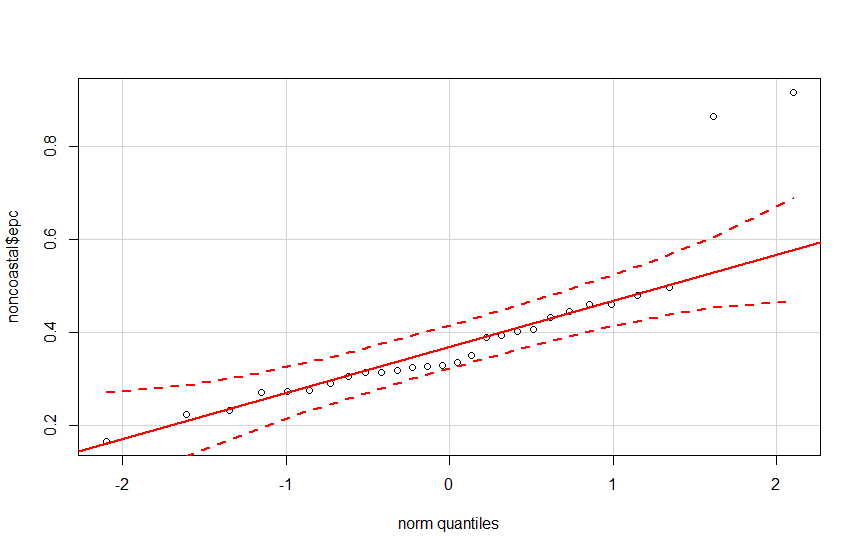
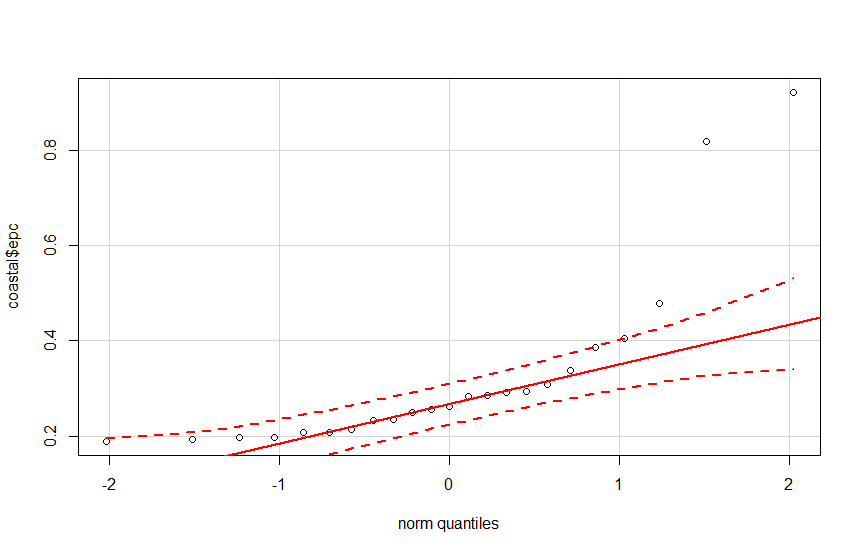


* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

Test: Two sample, two tailed, unpaired T-test

Assumptions: equal variance, normal distribution, and independent samples

I tested for equal variances using the var.test function. The p value is greater than 0.05, so we cannot reject the H0 of this test that the variances are equal (i.e. the variances are equal). I tested for normal distribution using the Shapiro.test function. The p-values were 5.037e-06 for coastal energy consumption per capita, and 3.627e-05 for noncoastal energy consumption per capita. Therefore, energy per capita is not normally distributed according to the Shapiro test. However, the qqPlots look close to normal, so I am going to assume normality. Also, the sample size is large enough to meet the assumption of normality or large enough sample size. I am also assuming independent random samples, as this is part of experimental design, and we do not have details regarding this.



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

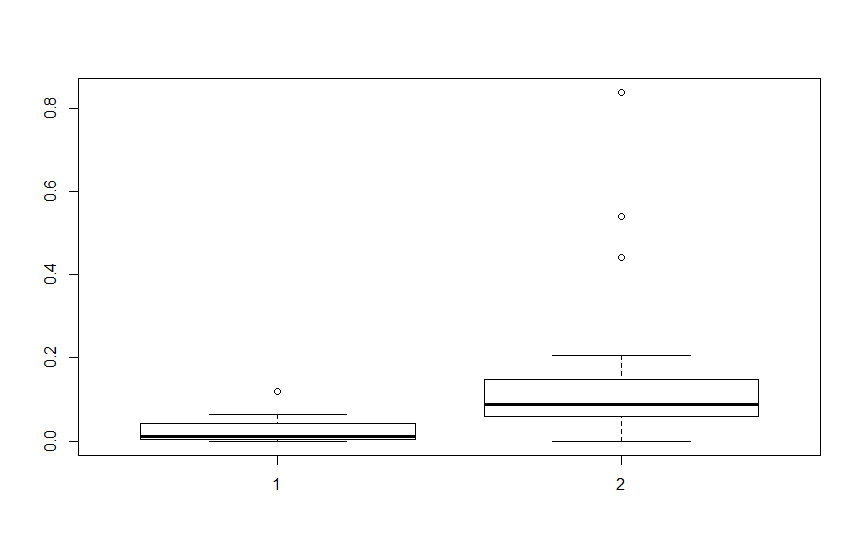
The p-value from the t-test is 0.2179, which is greater than 0.05, so we cannot reject the H0 that per capita energy consumption in coastal states does not differ from that in non-coastal states.

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).

H0: Per capita coal consumption in coastal states does not differ from that in non-coastal states.

Per capita coal consumption in coastal states differs from that in non-coastal states.

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

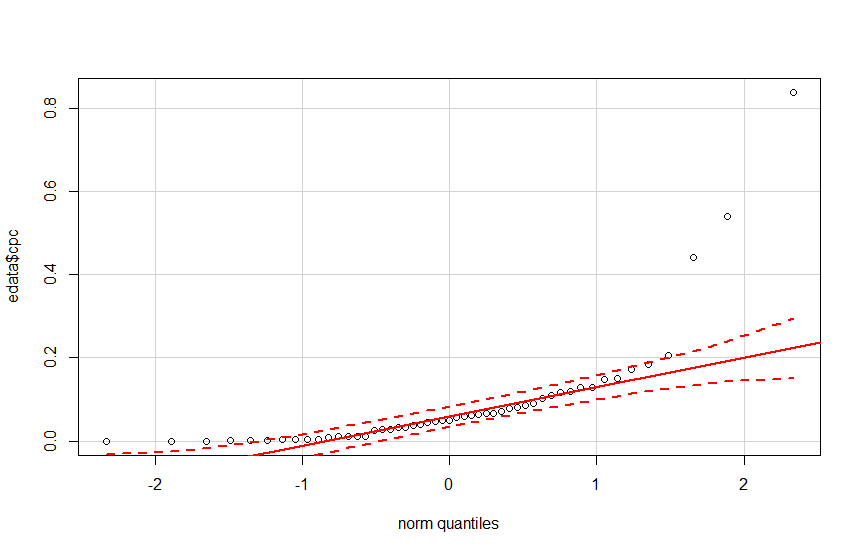


* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

Test: Two sample, two tailed, unpaired Welch’s T-test

Assumptions: equal variance, normal distribution, and independent samples.

I tested for equal variances using the var.test function. The p value is 5.995e-13, which is less than 0.05, so we reject the H0 of this test that the variances are equal (i.e. the variances are not equal). However, this is ok, because we can use Welch's T-test. I tested for normal distribution using the Shapiro.test function. The p-values were 0.001023 for coastal coal consumption per capita, and 6.211e-07 for noncoastal coal consumption per capita. Therefore, coal consumption per capita is not normally distributed according to the Shapiro test. From the qqPlots, they also did not look normal. However, I am going to violate this assumption and assume normality. Also, the sample size is large enough to meet the assumption of normality or large enough sample size. I am Also assuming independent random samples, as this is part of experimental design, and we do not have details regarding this.



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

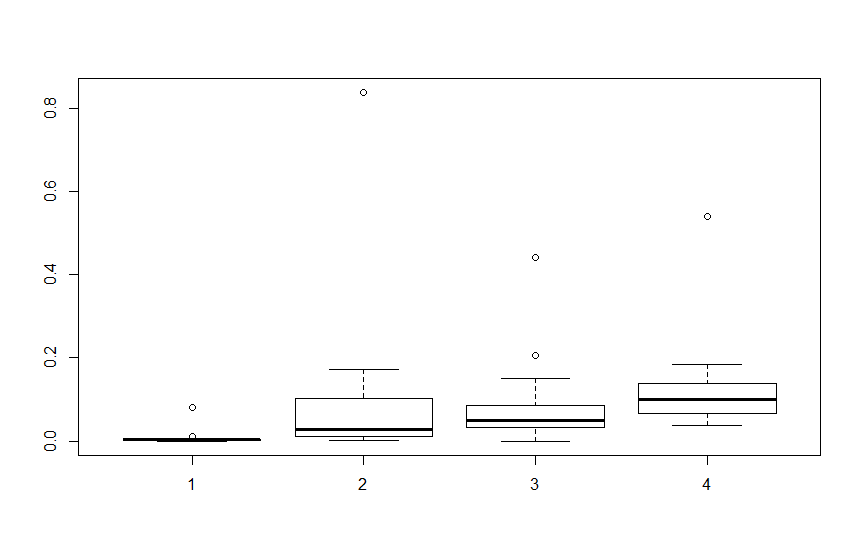
The p-value from the t-test is 0.001936, which is less than 0.05, so we can reject the H0 that per capita coal consumption in coastal states does not differ from that in non-coastal states. Therefore, per capita coal consumption is different in coastal than in noncoastal states.

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).

H0: Per capita coal consumption between states is not different based on region.

Ha: Per capita coal consumption between states differs by region.

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

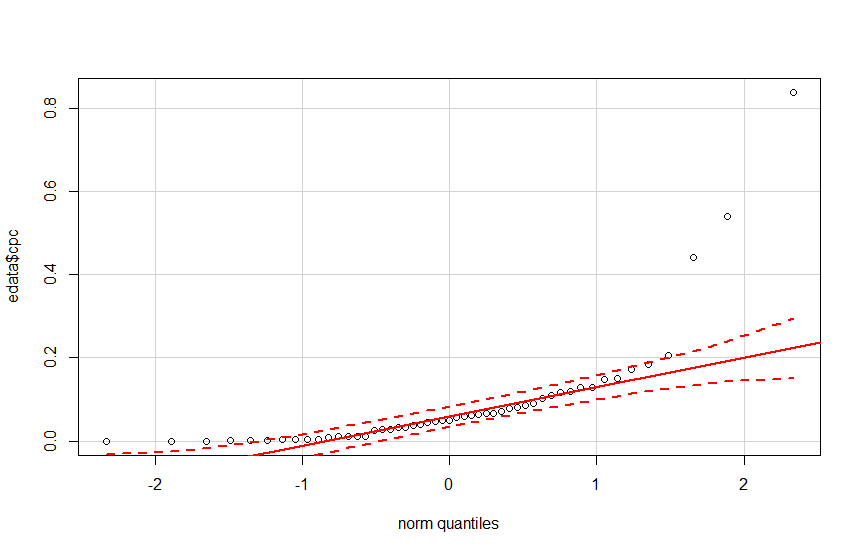


* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

Test: ANOVA

Assumptions: equal variance, normal distribution, and independent samples.

I tested for equal variance using the Levene Test. The p-value from the Levene test is greater than 0.05, so we fail to reject the null hypothesis of this test that the variances are equal (i.e. the variances are equal). I tested for normality using the Shapiro test and qqPlot function. The p-value from the shapiro test for the coal consumption per capita variable was 6.024e-11, which is less than 0.05, so the data are not normal. From the qqPlot it looks like the something is happening at the tails of the distribution. I violated this assumption and assumed normality to run the ANOVA.

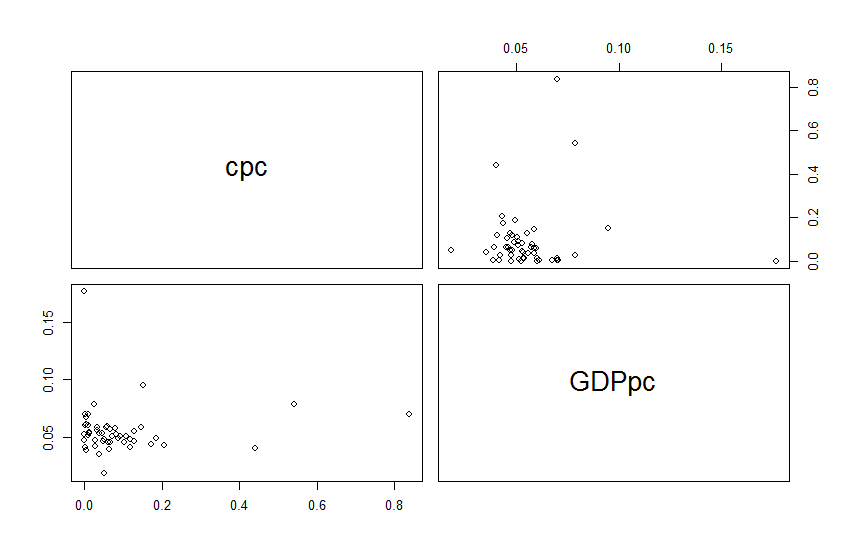


* 1. Please run the statistical test and interpret the result (1 point).

The p-value from the ANOVA summary is 0.262, which is greater than 0.05, so we fail to reject the null hypothesis that per capita coal consumption between states is not different based on region (i.e. per capita coal consumption does not differ based on 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)

The correlation coefficient is 0.03598182, which is not a strong correlation, because it is significantly lower than 0.50. Visually, the pairs output also shows that there does not appear to be a pattern between the variables.



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).

Chosen variables: nox, rm, ptratio

cor(nox, rm) = -0.2645944

cor(nox, ptratio) = 0.1034642

cor(rm, ptratio) = -0.3341642

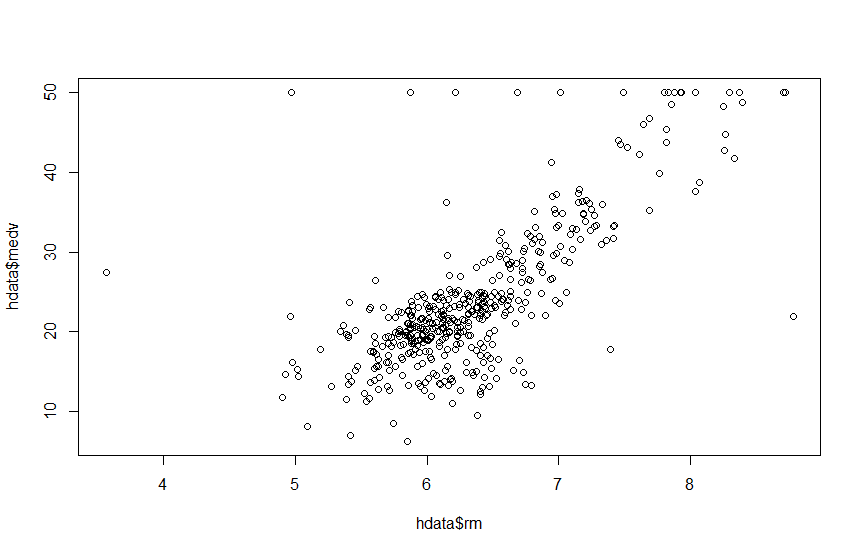
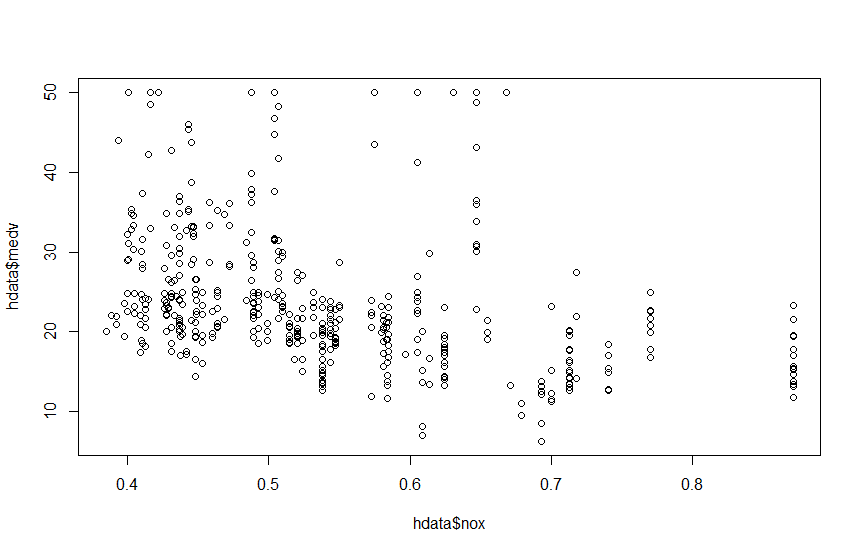
VIF.nox = 1.075575

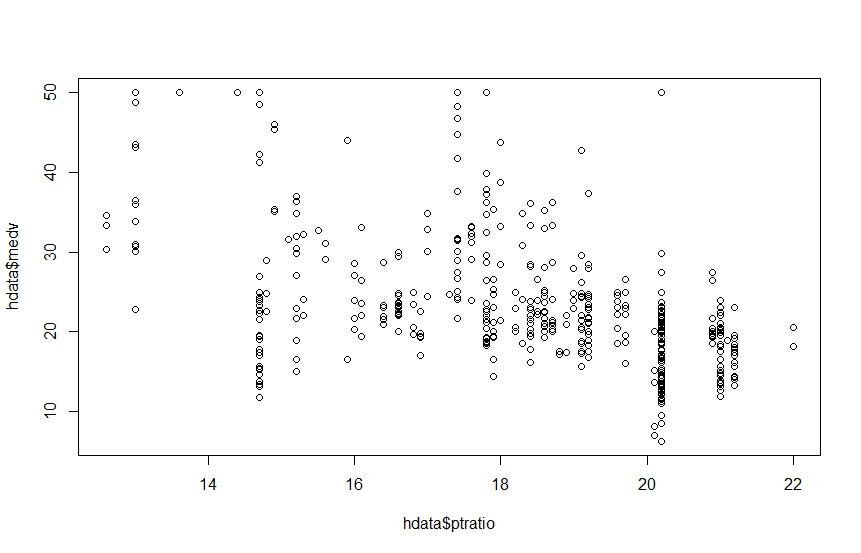
VIF.rm = 1.197816

VIF.ptratio = 1.126011

Based on the correlation values and VIF, there are not multi-collinearity issues.

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).





nox appears to have a slight negative correlation with median housing prices.

rm appears to have a strong positive correlation with median housing prices.

ptratio appears to have a slight negative correlation with median housing prices.

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 tested for residual independency using the dwtest function. The dwtest returns a p-value below 0.05, so there is autocorrelation in the residuals. Therefore, the assumption of residual independency is violated. I tested for heteroscedasticity using the bp test function. The bp test returns a p-value below 0.05, so there is heteroscedasticity. Therefore, the assumption of residual homoscedasticity is violated. I tested for normality of the variables using the Shapiro test. The p-values for all the variables were below 0.05, and the qqPlots confirmed that none of the variables were normally distributed. Therefore, the assumption of normality was also 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).

Coefficients:

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

(Intercept) -5.0099 4.4039 -1.138 0.256

nox -11.0969 2.3423 -4.738 2.91e-06 \*\*\*

rm 8.2246 0.4219 19.494 < 2e-16 \*\*\*

ptratio -0.9542 0.1240 -7.696 9.02e-14 \*\*\*

Intercept: the p-value indicates that the estimate of the intercept is not significantly different from zero, so when all variables in the model are set to zero, median housing price is 0.

nox: the p-value indicates that the estimate of the nox coefficient is significantly different from 0. For every unit increase in nox, median housing prices decrease by -11.0969 units, when controlling for the other variables.

rm: the p-value indicates that the estimate of the rm coefficient is significantly different from 0. For every unit increase in rm, median housing prices increase by 8.2246 units, when controlling for the other variables.

ptratio: the p-value indicates that the estimate of the ptratio coefficient is significantly different from 0. For every unit increase in ptratio, median housing prices decrease by 0.9542 units, when controlling for the other variables.

Interaction Term: medv~nox + rm\*ptratio

Coefficients:

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

(Intercept) -113.8979 17.4562 -6.525 1.85e-10 \*\*\*

nox -10.4033 2.2461 -4.632 4.76e-06 \*\*\*

rm 25.0855 2.6539 9.452 < 2e-16 \*\*\*

ptratio 5.2866 0.9781 5.405 1.06e-07 \*\*\*

rm:ptratio -0.9745 0.1516 -6.428 3.32e-10 \*\*\*



Intercept: the p-value indicates that the estimate of the intercept is significantly different from zero, so when all variables in the model are set to zero, median housing price is -113.8979.

nox: the p-value indicates that the estimate of the nox coefficient is significantly different from 0. For every unit increase in nox, median housing prices decrease by -10.4033 units, when controlling for the other variables.

rm: the p-value indicates that the estimate of the rm coefficient is significantly different from 0. For every unit increase in rm, median housing prices increase by 25.0855 units, when controlling for nox and setting ptratio to zero.

ptratio: the p-value indicates that the estimate of the ptratio coefficient is significantly different from 0. For every unit increase in ptratio, median housing prices decrease by 0.9745 units, when controlling for nox and setting rm to zero.

rm:ptratio: the p-value indicates that the interaction between rm and ptratio is significantly different from 0. As either variable increases one unit, the effect of the other on medv decreases by -0.9745.

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

The adjusted R-squared value is 0.6159, so the model explains 61.59% of the variation in median housing prices. This is a fairly good adjusted R-squared value. However, because the assumptions of residual independency, heteroscedasticity, and normal distributions were violated, a linear model is not a good model for these data.