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.

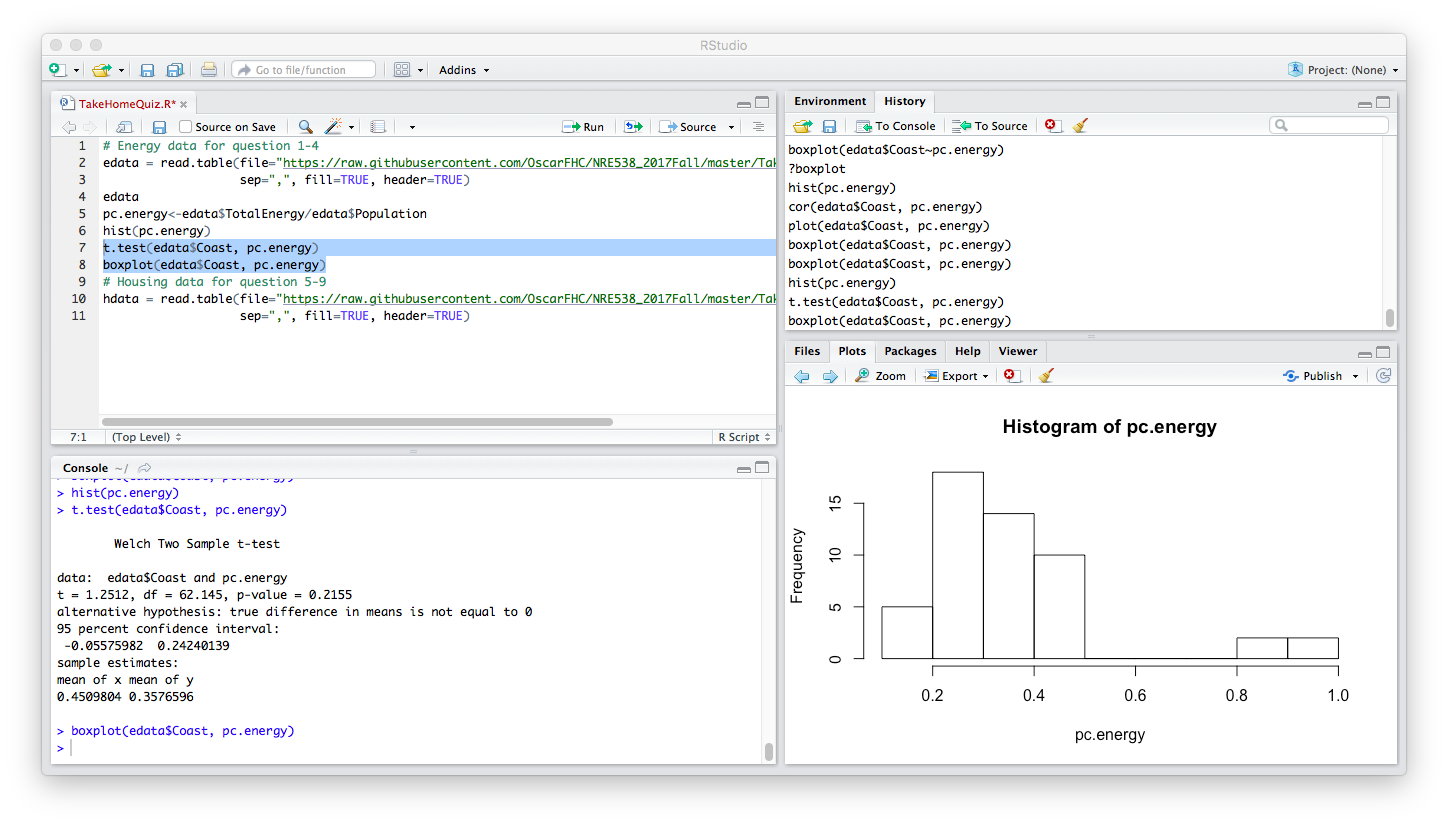
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized 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.

Ha: Per capita energy consumption differs based on whether the 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: two sample t-test

The assumption that the data is normally distributed is not met.

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

Welch Two Sample t-test

data: edata$Coast and pc.energy

t = 1.2512, df = 62.145, p-value = 0.2155

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.05575982 0.24240139

sample estimates:

mean of x mean of y

0.4509804 0.3576596

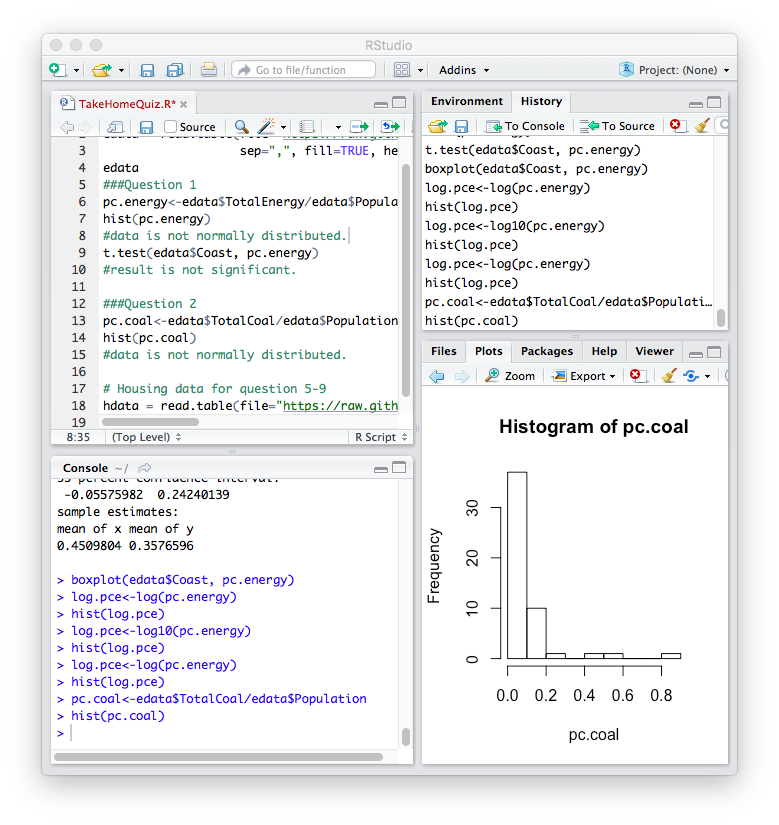
*As the p value is greater than 0.05, the result is not significant at the 95% CI. Thus, the difference in per capita energy consumed is not different depending on if the state is coastal or not.*

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.

Ha: Per capita coal consumption differs based on whether the 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: Two Sample t-test

The assumption that the data is normally distributed is not met.

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

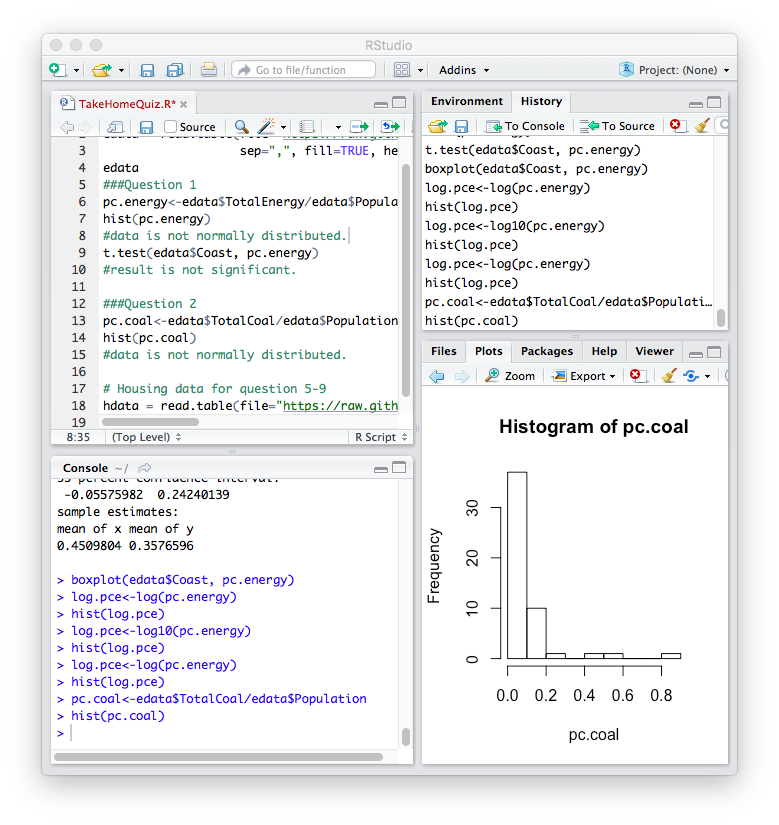
The p-value at 7.801\*10-6 is much lesser than 0.05 and is thus significant. Thus, the coal consumption differs based on whether the state lies on the coast or not.

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: Per capita coal consumption does not differ based on which region a state is found.

Ha: Per capita coal consumption differs based on which region 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).

Test: ANOVA

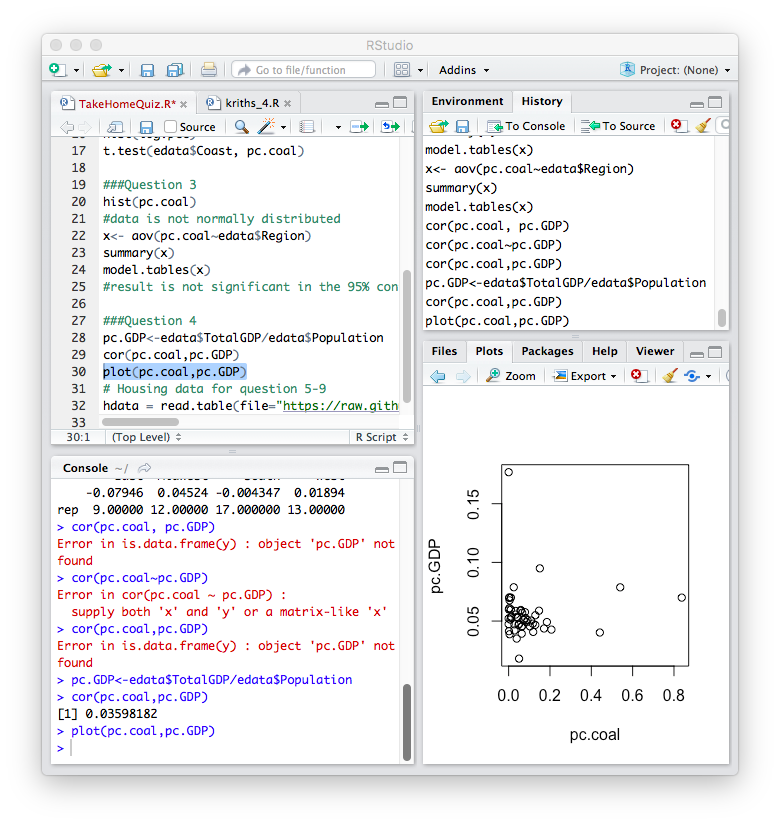
The assumption that the data is normally distributed is not met.

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

The p value is not significant at the 95% level. Thus, the per capita coal consumption does not differ based on which region the state is located in.

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 between per capita coal consumption and per capita GDP is 0.0359. this shows a positive correlation that is very weak (as the value is closer to 0). This is also seen by the scatter plot generated between per capita coal use and per capita GDP.

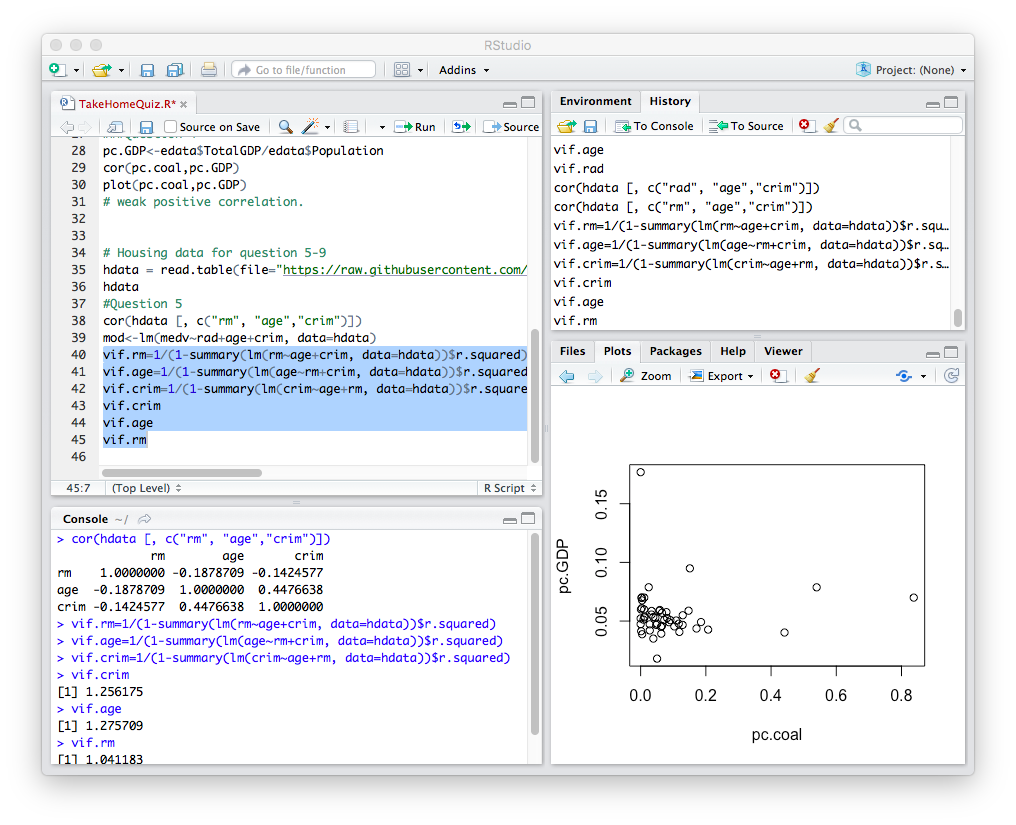


For questions 5-9, please use the ‘housedata.csv’ dataset that shows housing information for the Boston area. Information on what each of the variables are can be found here: <http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names>. In this exercise, the goal is to create a multiple 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.

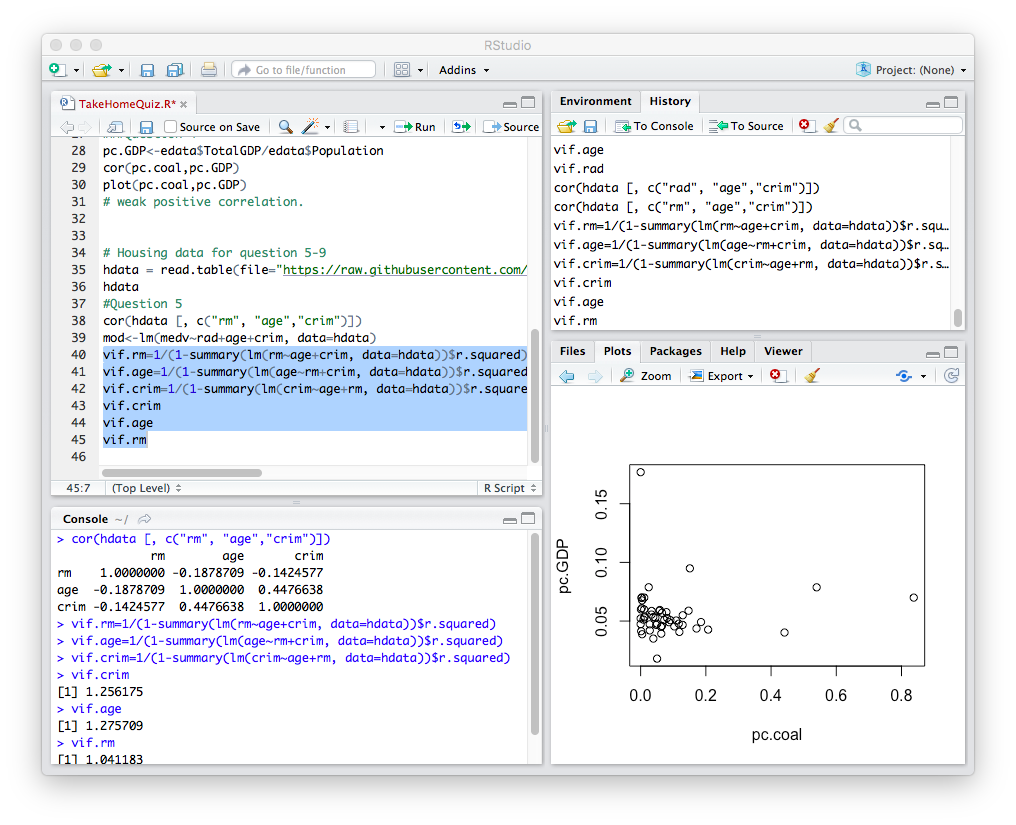
1. Please select three covariates that you will include in your model as independent variables. Please check if these variables are highly correlated with one another to make sure you do not run into problems of multi-collinearity. Check if this model has issues with multi-collinearity using the variance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

Variables: rm, age, crim

Correlation values:

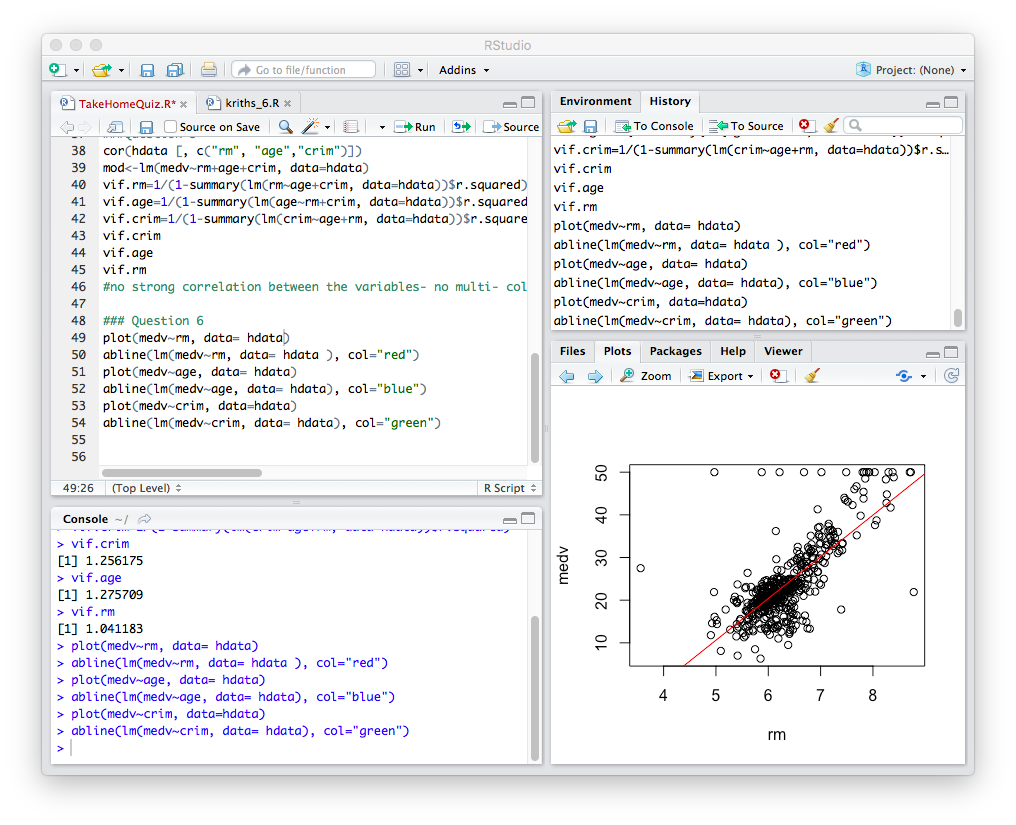
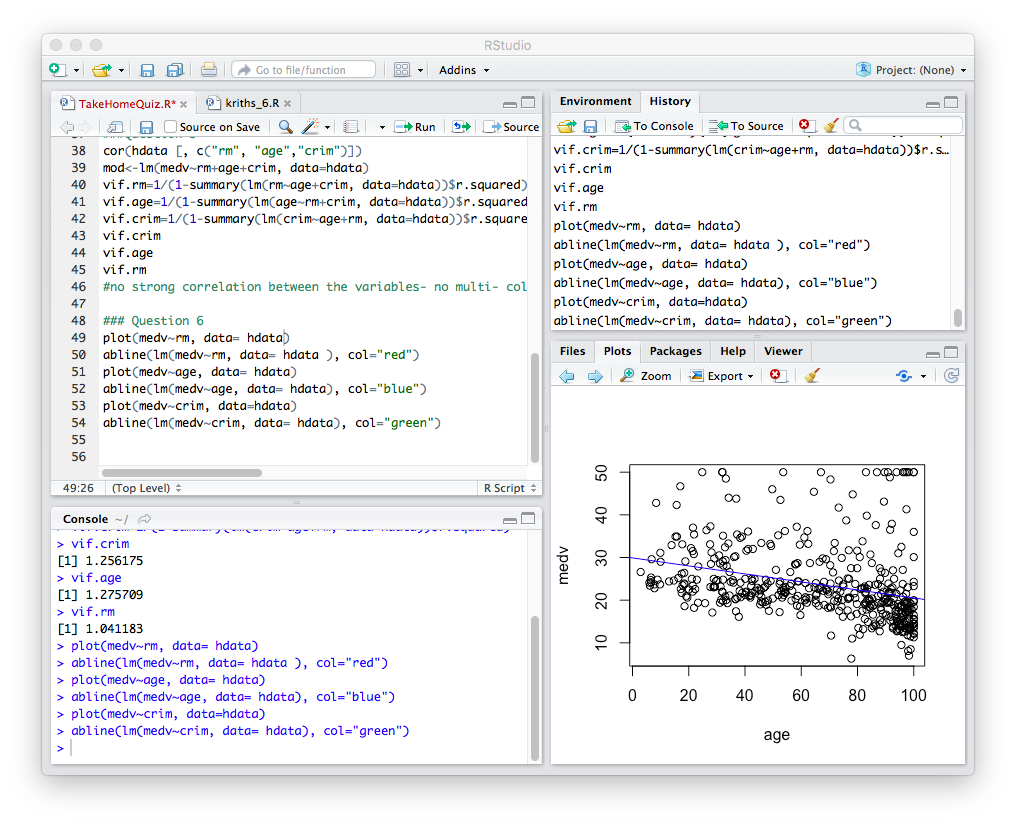


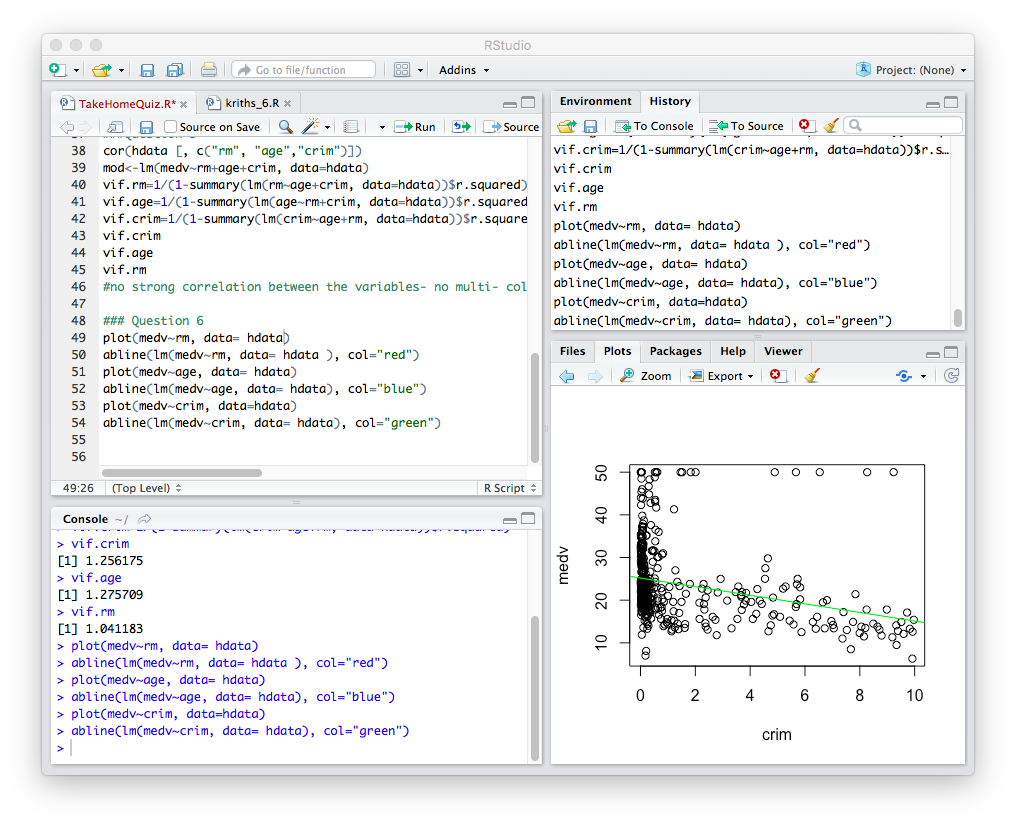
VIF values:



*Based on the correlation and VIF values, the variables chosen are not highly correlated, and hence, there is no issue of multi-collinearity.*

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).

A)B)



C)

Plot A- medv~rm: the median housing price increases with increase in average #rooms per dwelling.

Plot B- medv~age: the median housing price decreases with increase in proportion of owner-occupied units built prior to 1940

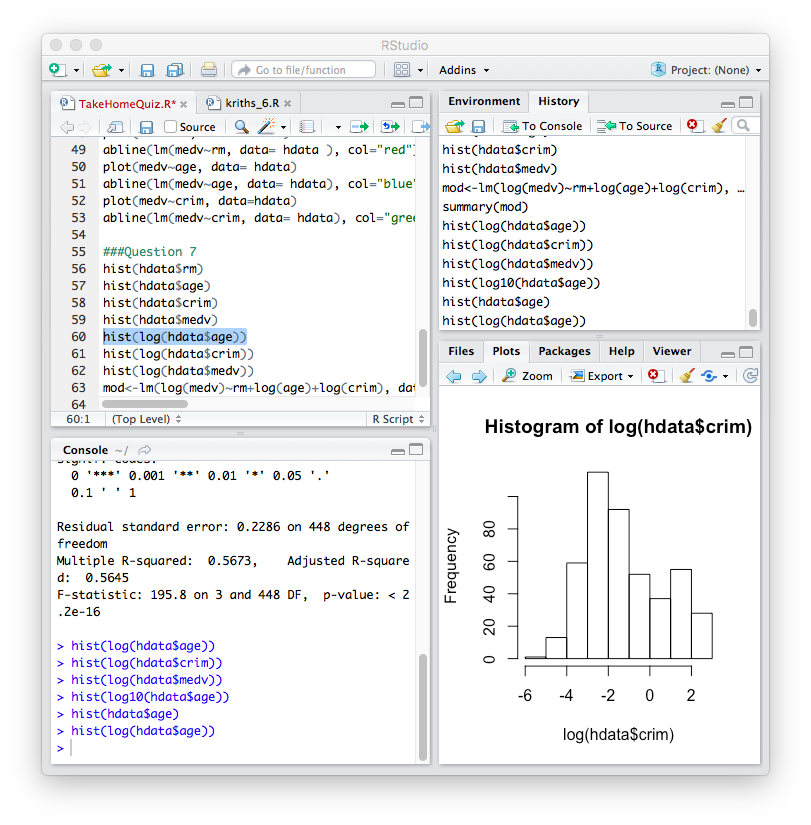
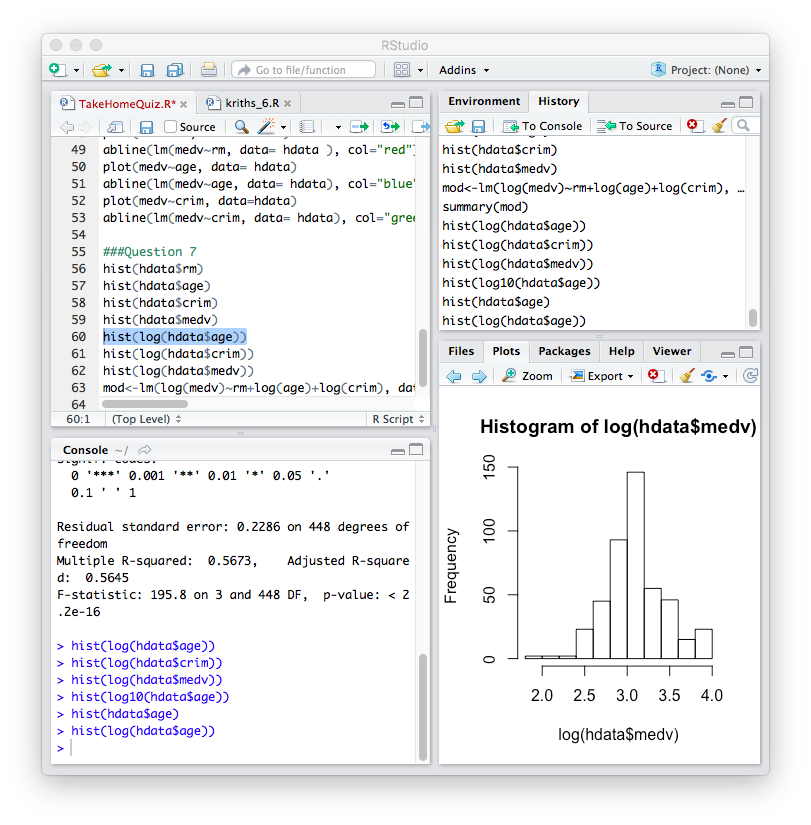
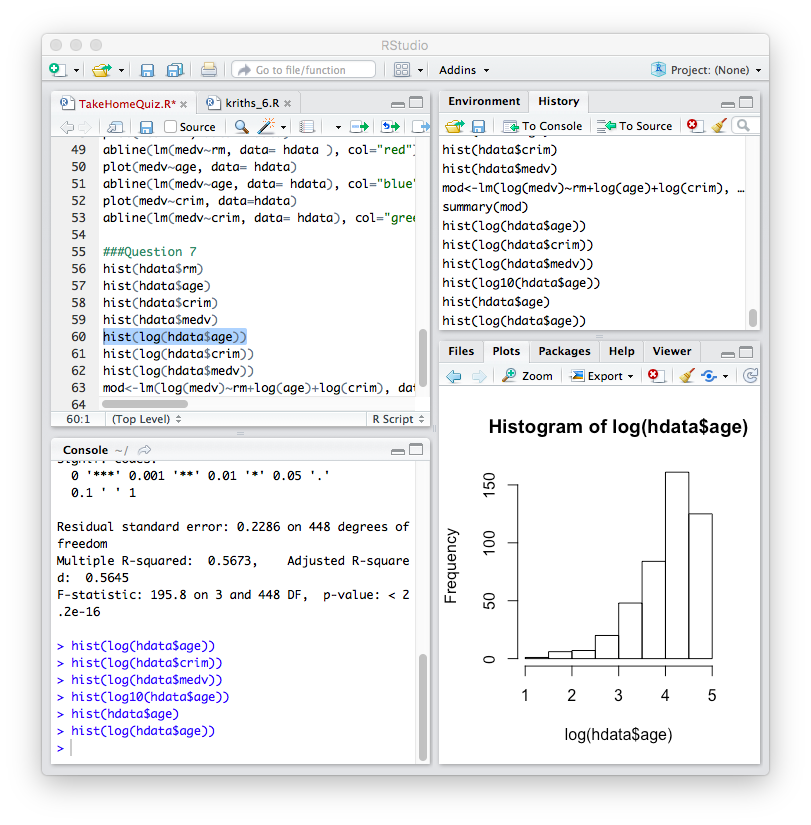
Plot C- medv~crim: the median housing price decreases with increase in per capita crime rate by town.

7. 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).

Assumptions:

1. Normality of data:

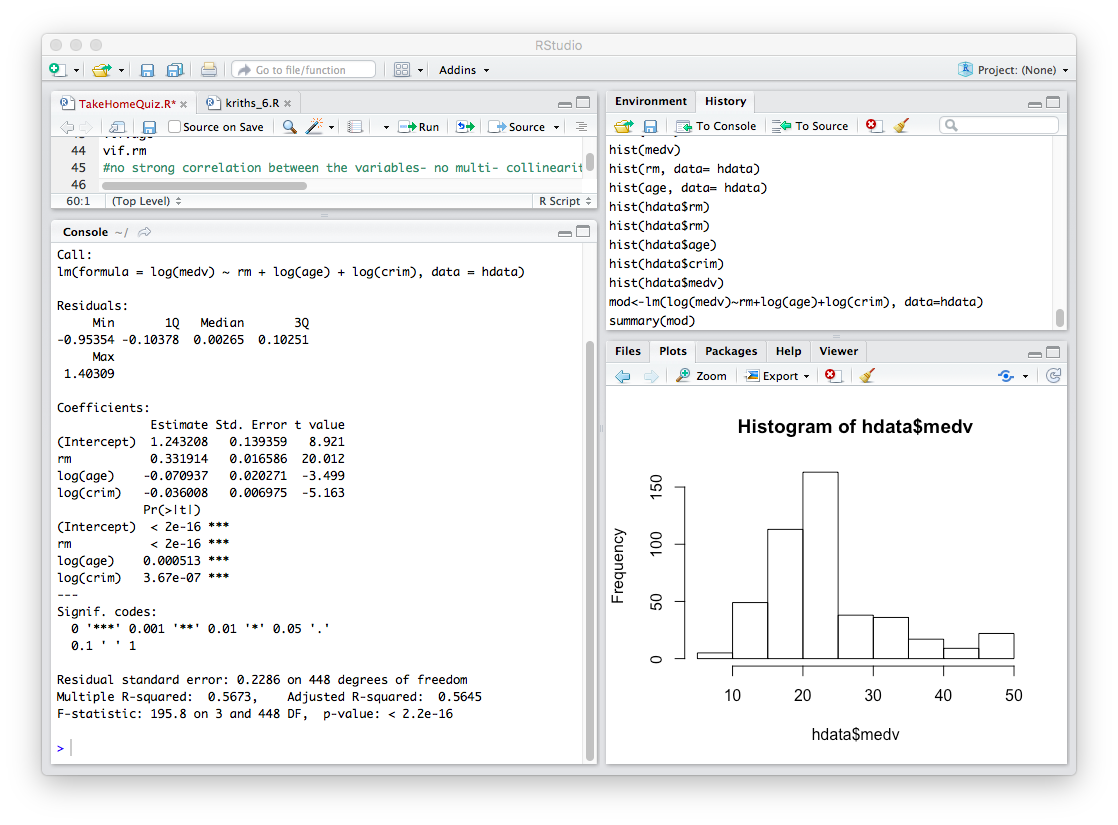
On plotting the histograms for rm, age, crim and medv, I observed that only rm was normally distributed. Thus, to meet the assumption for normal distribution of data, I log- transformed the following variables: age, crim and medv. This rectified the normality of crim and medv. However, there was only a slight change in the histogram for the age variable:



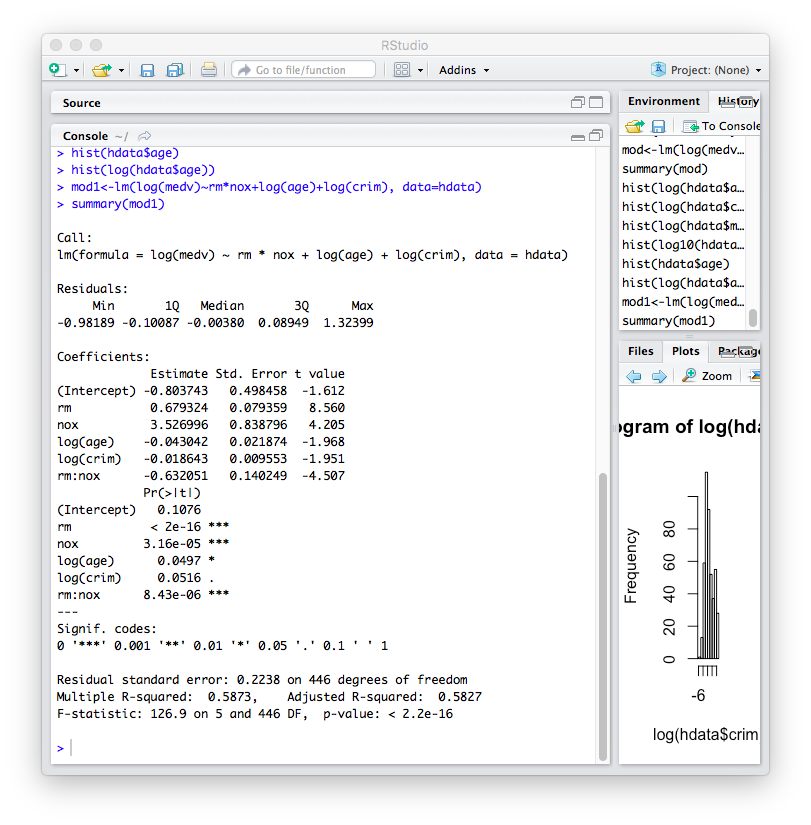
2. independence of variables:

I checked this by estimating the correlation values and the VIF values for each variable. As the values are very small, we can assume that the variables are independent of each other.

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).

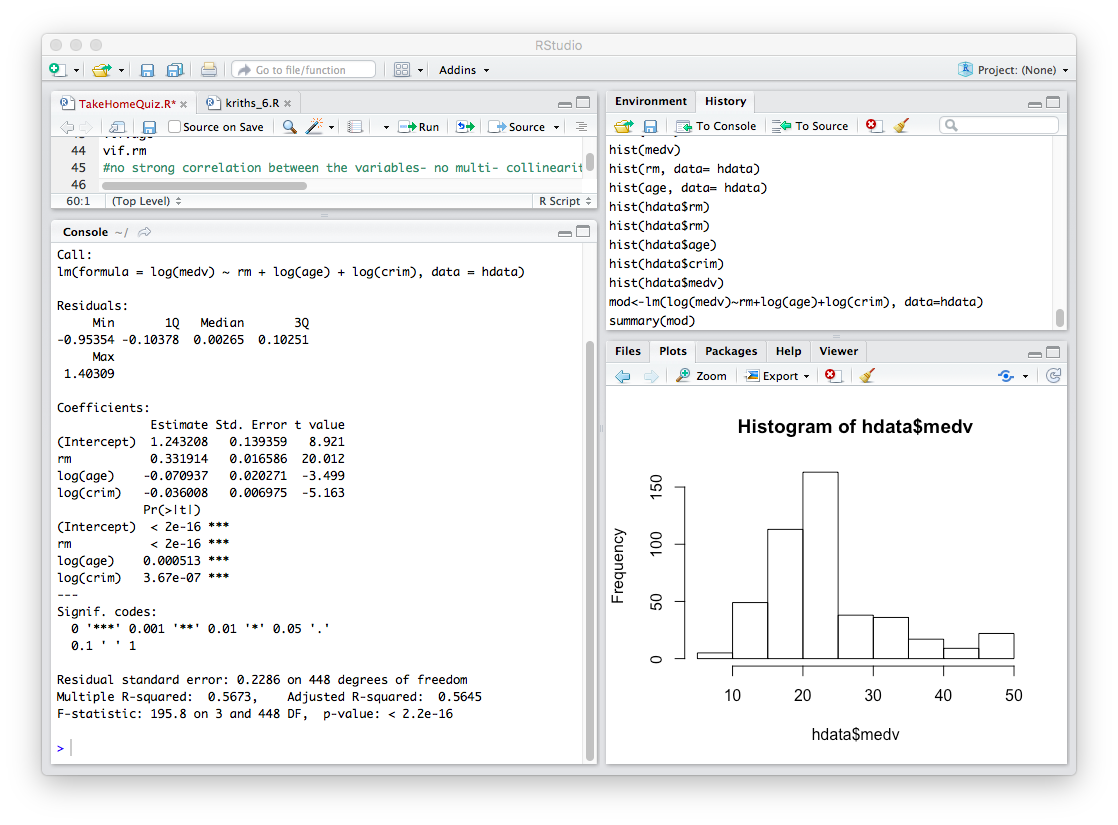


|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Coefficient | P value | Comment |
| Intercept | 1.24 | <2\*10-16 | The median value of owner occupied homes (when all other variables are zero) is 1240. This value is statistically significant. |
| Rm | 0.331914 | <2\*10-16 | As average number of rooms increase by 0.33, the median value of homes increase by 1%. This value is statistically significant. |
| Log(age) | -0.070937 | 0.000513 | As the proportion of owner occupied units built prior to 1940 decreases by 7%, the median value increases by 1%. This value is statistically significant. |
| Log(crim) | -0.036008 | 3.67\*10-7 | As the per capita crime rate decreases by 3%, the median house value increases by 1%. This value is statistically significant. |



As the number of rooms and the nitric oxide concentration decreased by 0.63, the median value of the house increased by 1%. This value is statistically significant.

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



From the above model (without the interaction term, we see that the adjusted R- squared value is 0.5645. Thus, we can consider this model to be a good fit for the data.

The inclusion of an interaction term increased the adjusted R-square value to 0.58. I would thus choose the second model over the first to fit the data better.