Introduction to Data Science with R

Homework Assignment #1

HA #1 is analyzing the data file **HPRICE1.xlsx**. In order to solve the assignment, you should:

- Create an R script that performs the required operations and submit the R script file only.
- Whenever there is a numerical question, ("What is the median of…"), running your code should calculate the required answer, assign it into a variable, and answer the question with a full sentence (i.e. the code should print "The median of something is 17").
- When asked to remove variables or observations, you should continue the assignment with the reduced data frame.
- Make sure that your file has meaningful variable names and is properly documented with comments, like in the R scripts presented in class.
- Please use only the R libraries presented in class and try to use pipelines as much as possible.

---

1. The first spreadsheet in the file contains the definitions of the variables, and the second spreadsheet contains the data. Check the documentation of read_excel in order to find out how to load the data from a spreadsheet which is not the first one in the file. How many observations are there in the data?
2. Remove all the variables that include log of other variables.
3. What is the percentage of houses in the dataset with more than 5 bedrooms? Use the function "round" in order to print the result in percentage terms with 1 significant digit (i.e. "17.8"). Remove these houses from the dataset.
4. What is the average price of a colonial house with more than 2000 sqrft?
5. Arrange the dataset according to lot size in a decreasing order. What is the assessed value of the house with the largest lot size?
6. How many houses are missing the lot size value? Remove them
7. How many colonial houses have a price per sqrft above 160?
8. Add a new variable to the dataframe with the price per bedroom and order the data frame from the smallest to the largest price per bedroom
9. Create a scatter plot, where the horizontal axis shows the size of the house in square feet and the vertical axis shows the assessed price.
10. Create the same plot, with different colors for colonial and non-colonial houses. Note that problem as ggplot treats colonial as a numerical and not a categorical variable. In order to solve the problem, create a new variable, is.colonial, which gets TRUE is colonial is 1 and FALSE otherwise, and use this new variable, within a single pipe and ggplot call. Also add a single regression line for all the observations in the plot, without the prediction interval
11. Create a new variable called size.type, which gets the value "Many bedrooms" if the number of bedrooms is above the median and the value "Few bedrooms" if below, using the "ifelse" function. Create a chart with two sub plots based on this

new variable, and where the size is based on the lot size and the color based on whether the house is colonial or not, using a single pipe and ggplot call.

12. Look up how to add a title to a ggplot chart and add a proper title to the chart.