**DS 710**

**Homework 8**

**R assignment**

1. In this problem, you will create and apply a function that rates cities based on how appealing they are for you to live in.

a.  What factors are important to you in deciding where to live?  The data set Best Cities.csv contains data on 10 U.S. cities, obtained from <http://www.census.gov/quickfacts/>, [www.walkscore.com](http://www.walkscore.com), and <http://www.wunderground.com/>.  Develop your own formula to rate how pleasant a city would be to live in, in your opinion, based on the variables in this data set.

* Your formula should yield a single number, which is higher for cities which are more pleasant.
* Your formula should use at least 3 different variables from the data set.
* Your formula should not rely on a comparison between cities.
* If there are other variables that you consider to be important, you may add them to the data set.  If so, upload the modified version of the data set with your homework, and include a link to the data source.
* Be creative!

b. Create an R function that computes the pleasantness score of a city, based on a vector of data about it.  You may assume that the vector contains data in the same order as it is listed in Best Cities.csv.

c.Use apply() to apply your function to each city in the Best Cities.csv data set.  Based on your criteria, which city is the best for you?  Does this assessment seem accurate?  If not, what would you want to change about your formula?

2.  Can we use statistical analysis of word lengths to identify the author of an anonymous essay?  In Homework 7, you wrote a Python function that counted the lengths of words in the 1770 essay by “A Mourner”.  Analysis of other articles published in *The Boston Gazette and Country Journal* in early 1770 finds that John Hancock wrote a 121-word article with a mean word length of 4.69 and standard deviation of 2.60.

a. Use R to assess whether it is plausible that John Hancock was “A Mourner”.

Note:  Because we only have summary statistics (mean and standard deviation) for John Hancock’s article, the t.test function won’t work to do a two-sample t-test.  You may either assume that the average word length of all of John Hancock’s writing (not just this sample) was 4.69 (and do a one-sample t-test), or install the BSDA package and use tsum.test() instead, as discussed here:  <http://stats.stackexchange.com/questions/30394/how-to-perform-two-sample-t-tests-in-r-by-inputting-sample-statistics-rather-tha>

Note 2: The null hypothesis for a 2-sample t-test of this question is

i.e., that A Mourner and Hancock have the *same* mean word length. In other words, the null hypothesis is that it *is* plausible that Hancock was “A Mourner.”

b. What assumptions must we make to draw a conclusion in part a?

3. In this problem, you will investigate whether sentences with more words tend to contain longer words.

1. In Homework 7, you used Python to count the number of words and mean length of words in each sentence of *Pride and Prejudice*.  Load this data set into R and make a scatterplot of mean length of words versus number of words per sentence.  Does linear regression appear appropriate here?
2. If your answer to part a was “no”, apply a transformation to make linear regression more appropriate.
3. Test whether sentences with more words tend to contain longer words.  State your conclusion in context.
4. Add a line to your scatterplot representing the regression model.  Write a sentence interpreting its slope.
5. Examine the residual diagnostic plots, and explain what they tell us in this case.

Submit a .doc, .docx, .rmd, or .pdf document containing your R code, plots, and interpretations.