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, 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.
    - You may pre-compute quantities like the standard deviation of the data, if you consider them important. What we're trying to avoid here is conditions like, "+3 if it's one of the top 3 warmest cities in the data set", which can't be determined by looking at each city separately.
- 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.

- Your function should not assume the existence of any variables which are not used as arguments of the function.

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. We want to use R to assess whether it is plausible that John Hancock was “A Mourner”, based on his mean word length. Explain why a 2-sided, 2-sample t-test is appropriate for this.

b. Explain why the t.test() function is *not* appropriate for the data we have available.

c. Write your own function for performing a 2-sided, 2-sample t-test for equality of means when the raw data are not available. Use the following information. (If the formulas do not show up correctly, please view the pdf version of the assignment.)

- The *test statistic* (a preliminary quantity needed to compute the p-value) is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE}$$

where $\bar{x}_1, \bar{x}_2$ are the sample means of the two samples, and SE is the standard error (similar to standard deviation, but for [in this case] the difference of sample means instead of the raw data).

- The standard error is

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where n_1, n_2 are the sample sizes and s_1, s_2 are the sample standard deviations. (This approach uses Welch’s t-test, which does not assume that the two populations have equal variances.)

- The p-value is

$$p = 2 * pt(-|t|, df)$$

where $-|t|$ is the negative absolute value of *t*, the test statistic, *df* is the number of degrees of freedom, and *pt*() is the R function *pt*().

- The degrees of freedom are

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

- Use additional functions as needed to organize your work.
- Your function should not assume the existence of any variables which are not used as arguments of the function.

d. Test your function by comparing it to t.test() on a pair of samples. You may wish to use rnorm() to generate random data from a normal distribution. If the p-value from your function doesn’t match the p-value from t.test(), then revise your code from part c.

e. Apply your function to assess whether it is plausible that Hancock was A Mourner. Write your conclusion as a sentence.

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

$$H_0\text{: } \mu_{Mourner} = \mu_{Hancock}$$

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.”

Submit a single .docx or .pdf document containing your R code, output/plots, and interpretations for both problems. Keep the code, output/plots, and interpretations for a given part of a problem together (i.e., don’t put all the code at the end of the document). Your R code should be the clean, final version—if you eventually got it working, there’s no need to show us your false starts.