STAT 331 Lab 3

Instructions

Please submit an HTML document created using R Notebook, but you are more than welcome to test your code out in an R script first. **Even if a question does not say "write code," you should write code for your answer!**

The Excel data set "hiphop" on the class PolyLearn site contains results from a study conducted by a linguist at the University of Minnesota. The researcher was interested in predicting musical taste based on familiarity with African American English (AAE). 168 subjects participated in the study, and each was asked to define 64 different AAE terms. The definitions given were used to create a "familiarity" score for each subject for each term. This score quantifies how well the subject knew the term on a scale of 1-5 (1 = not at all, 5 = very well). Before tackling the problems, **study the information on the following website**, which includes a description of each variable:

http://conservancy.umn.edu/bitstream/handle/11299/116327/5/explanationAAEHiphopChesley.txt

BE SURE TO SAVE YOUR WORK REGULARLY!!!

Exercises

1. The data set is on the course website. Read it into R, and display the first 6 observations.
2. Display the variable names of the data frame, and use the str() function to examine variable types.
3. Note that subj and word have been read in as factors (categorical variables), but we want R to treat them as character variables. To change that, you need to use the "as.is" option to read.csv. Give "as.is" the column numbers of columns that you don't want to be automatically converted to factors (in this case, columns 1 and 2). After reading in the data set with this option, check that it worked by using the str() command again.
4. What are the dimensions of the data set? Do the dimensions make sense considering the information given above about the study? Explain. *Hint: Examine the subj and word variables.*
5. Display the 64 AAE words that were tested, with no duplicates. There are various ways to remove duplicates; one is with the unique() function.
6. Use the summary() command to get an overview of the hiphop data set. Which variables contain missing values?
7. How many missing values are in the data set?
8. Calculate the mean and standard deviation of numPreferredArtists. Because this variable has missing-values, you will need to set the "na.rm" argument equal to TRUE.
9. Write code to create a new data frame called subject19 which only contains information for subject 19. What are the dimensions of this new data frame?
10. Display the familarity variable of the subject 19 data frame in three different ways.
11. Which word is examined on the 30th row of subject 19?
12. Write code to order this new data frame by familiarity from largest to smallest, retaining this sorting in the subject19 data frame (ie, you should not print out the data frame).

13. Display *only* the words, trial number, and familiarity scores of subject19 such that the familiarity score is greater than or equal to 4 and the trial is less than or equal to 20.