9.59 Lab in Psycholinguistics: Problem Set #1

The goal of this problem set is to get you comfortable manipulating and analyzing data in R. Here is what you should turn in:

1. A document (preferably a PDF but Word or txt is okay) that contains JUST the answers to the questions.
2. Your R code with comments.

OR

1. An R Markdown document (pdf or html) where you write your responses and show your code.

You will explore a data set provided in the languageR package from the Baayen book. It contains a set of English words with lexical decision time data and other measures like written and spoken frequency. Lexical decision time is one of the most important behavioral measures used in psycholinguistics. In a lexical decision task, a subject is presented with a word (or non-word like gnuppet) and asked to judge as quickly as possible whether or not it is a word. How fast they can make a decision reflects something about the psychological response of the subject to the word in question. In this problem set, you will find out what sorts of things are predictive of lexical decision time.

- Download rts.csv and put it in the same folder as your script. Read in the data as a data frame.
- RTlexdec is the lexical decision time for each data point in the data frame. Other relevant columns will be defined below.

1. What is the overall mean and the overall standard deviation for RTlexdec? Use the functions `mean()` and `sd()`.
2. How is lexical decision time affected by how long the word is? Calculate the means and standard deviations for RTlexdec by LengthInLetters. Look at the mean and standard deviation for each of the possible lengths. Using `ggplot`, make a scatterplot with LengthInLetters on the x axis and RTlexdec on the y-axis. What can you conclude?
3. Make a new data frame containing only words for which VerbFrequency and NounFrequency are greater than 0. How many data points were there originally in the dataset and how many does this eliminate from the original data frame? How is the mean RT affected and why?
4. The WrittenSpokenFrequencyRatio compares the frequency of a word in writing to its spoken frequency. So big numbers mean it's more likely to be written. What's the relationship between word length and WrittenSpokenFrequency? (You can answer in a number of different ways.)
5. How does the mean RTlexdec for words that start with "p" compare to the overall mean RTlexdec? What about words that start with "q"? Useful functions might be `mean()`, `str_sub()`, `mutate()`, `filter()`. Using `facet_wrap()`, show a histogram of RTs for words starting with "p" juxtaposed with a histogram of RTs for words starting with "q". What does this tell you about words that start with "p" and words that start with "q"?
6. What has a lower mean RTlexdec: nouns or verbs? What are the means? Find the log ratio $log(A/B)$ of the NounFrequency and VerbFrequency and add it as a column called NounVerbFreqRatio. Sometimes this gives you an answer that's not a number. In those cells, replace the non-number with NA. Using the function `is.na()`, how many NA's are produced?
7. Now we want to know how far the lexical decision time for some given words are from the means for the group. To do this, you can use z-transformed values. Add a column to the original data frame containing z-transformed values for RTlexdec. The formula for the z-score is $(x - \mu)/s$, where x is the particular value, μ is the mean value, and s is the standard deviation. Estimate z-scores separately for each of the two possible AgeSubject categories. That is, for each word, calculate one z-score using the mean and standard deviation for YOUNG people and one using the mean and standard deviation for

OLD people. Give z-scores for the words GULP and DOE. There should be two z-scores for each word (one for YOUNG and one for OLD). How do response times to “doe” and “gulp” compare to the mean in each group?

8. Consider and plot the correlations between each pair of 2 among the following: Familiarity, WrittenFrequency, FamilySize, and RTlexdec. Which factors are positively correlated with reaction time? Are you surprised by the results? Why or why not?

9. Using the z scores calculated before for young and old separately, find the words that have the biggest difference between young z score and old z score. List the 3 for which young is biggest relative to old and the 3 for which old is biggest relative to young. Are you surprised by the results? (For added fun, type these words into the Google Ngram viewer!)