# Activity: HipHop Lyrics

## Instructions

Please submit an HTML document created using R Markdown, 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 data set "hiphop" 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. Copy the following code into an R chunk, to load the data and gain access to the tidyverse package.

```
hiphop <- read.csv("https://raw.githubusercontent.com/kbodwin/STAT-331/master/
                    In-Class%20Activities/Data/hiphop.csv?
                    token=AVHCwTQaeq5UylWJxCcNN8qYww6UIaLqks5cP75ewA%3D%3D")
library(tidyverse)
```

2. What are the variable names and types in this dataset? Give a general overview, not a full list.
3. 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.*
4. Display the 64 AAE words that were tested, with no duplicates.
5. Get an overview of the hiphop data set. Which variables contain missing values?
6. How many missing values are in the whole data set?
7. Calculate the mean and standard deviation of numPreferredArtists. *Hint:Because this variable has missingvalues, you will need to set the "na.rm" argument equal to TRUE.*
8. 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?
9. Display the familarity variable of the subject 19 data frame in two different ways.
10. 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).