# Practice Midterm Exam

**Instructions**: Find a quiet place to work where you will not be interrupted for at least 60 minutes. Just as your lab and homework assignments, your exam will be graded based on the R script you submit, so make sure the R script you create for this practice exam is as detailed as possible, with clear comments accompanying each new block of code. It is important that you know what each line of code you use is doing and what its syntax is.

Once you are done with all questions, save your R script. I will go over this practice exam during next week’s office hours (5:30-7:30 pm on Monday, April 12th): <https://syracuseuniversity.zoom.us/j/96726336772>

If you cannot join the Zoom meeting but have questions about the practice test, please email me.

**Dataset Description**: The dataset used in this practice exam is called ***testData.csv*** and can be found here: <https://ist387.s3.us-east-2.amazonaws.com/data/testData.csv>

Read the file into a dataframe called “***data***” using the **read\_csv()** function from the ***tidyverse*** package.

Inspect your dataframe using the appropriate R function(s) – it contains the following variables: **Twitter user ID**, **number of followers**, **number of users the focal user follows**, **total number of tweets** a user has posted, and the user’s **state of residence**. Each observation, aka row, in the dataframe therefore represents the record of a unique Twitter user.

Don’t forget to “library” the appropriate R packages you think you will need to complete the tasks below.

# Research Questions (aka, your coding tasks):

1. Describe the **number of followers** variable using descriptive statistics provided by R. Do the same for the **number of users followed** variable.

*Hint*: What function(s) have we used so far to summarize variables in a dataset?

1. Describe the shape of the distribution for **number of followers**. Do the same for **number of users followed**.

*Hint*: How do we represent a distribution in R? Perhaps a histogram might help? Are the variables normally distributed? Or are their distributions right- or left-skewed?

1. On average, do the focal Twitter users in this dataset follow more accounts, or are they followed by more accounts?

*Hint*: Think of a statistical measure for each of the two variables that can help you make this determination. Perhaps looking at a measure of central tendency would help?

1. Create a new variable that represents the difference in number of followers and number of users followed for each focal Twitter user, aka observation, aka row. Describe the shape of the distribution of this new variable.
2. Create a scatterplot of the number of followers and number of users followed. Clearly label your axes so they are more descriptive. Does the scatterplot show a pattern or relationship?
3. Generate a linear model to predict the number of followers based on the number of tweets and the number of users followed. Generate another linear model to predict the number of users followed based on the number of tweets and the number of followers a user has.
4. Interpret the coefficients of the statistically significant predictors in the two models. Comment on the quality of each model. Which model is better – explain.
5. If you come across a Twitter user with 541 followers and 1128 tweets, what would be your model’s prediction about the number of people this user follows?
6. What would be your best guess about the number of people a user with 0 followers and 0 tweets follows?
7. Generate a map of the average number of tweets in each state where each state is shaded depending on its average number of tweets.

*Hint#1*: You may need to use the **aggregate()** function from your ggmap HW, since the current level of analysis in your data is individual users and you now want the level of analysis to be states.

*Hint#2*: The FUN (i.e. “function”) part of the **aggregate()** code in your HW may need to be modified – in your HW, we wanted to add up the income of individual ZIP code areas when aggregating, that’s why we used “sum,” but now we want the average – what might you want to replace “sum” with in that case?

1. Your map likely looks weird - what argument can you add to your map code to make sure states with no data still appear on the map, with an outline color and a fill color?
2. Based on your analysis so far, do you think the number of followers and the number of users followed are related?