STAT GR5206 Homework 9b [100 pts]
Due 8:00pm Monday, December 5 on Canvas

Please submit Homework 9a and Homework 9b in TWO SEPARATE documents. We are doing this so we can get the homework graded as soon as possible with the end of the semester approaching.

Your homework should be submitted on Canvas as an `R` Markdown file. Please submit the knitted .pdf file along with the .Rmd file. We will not accept any other formats. Please clearly label the questions in your responses and support your answers by textual explanations and the code you use to produce the result. Note that you cannot answer the questions by observing the data in the "Environment" section of RStudio or in Excel – you must use coded commands. We will print out your homeworks. Please do not waste paper by printing the dataset or any vector over, say, length 20.

In this homework you'll explore transforming data and get more practice with selective access and applying functions.

We continue examining the diffusion of tetracycline among doctors in Illinois in the early 1950s, building on our work from Homework 9a. You will need the data sets `ckm_nodes.csv` and `ckm_network.txt`.

Before you begin repeat the cleaning steps from Homework 9a: Load the `ckm_nodes.csv` data into a data frame called `nodes`. The variable `adoption_date` records the month in which the doctor began prescribing tetracycline, counting from November 1953. If the doctor did not begin prescribing it by month 17, i.e. February 1955, when the study ended, this is recorded as `Inf`. If it's not known when or if the doctor adopted tetracycline, their value is `NA`. Create a vector which records the *index numbers* of the doctors for whom `adoption_date` is not `NA`. Reassign `nodes` so it only contains those rows. (Do not drop rows if they have a value for `adoption_date` but are `NA` in some other column.) Use this cleaned version of `nodes` for the rest of the homework.

The file `ckm_network.txt` contains a binary matrix; the entry in row i, column j is 1 if doctor number i said that doctor j is a friend or close professional contact, and 0 otherwise. Load the file into `R` and call it `network`. Drop the rows and columns corresponding to doctors with missing `adoption_date` values. Use this reduced matrix, and its rows and column numbers for the rest of the homework.

1. Write a function `adopters` which takes two arguments, `month`, with no default value and `not.yet` defaulting to `FALSE`. If `not.yet` is `FALSE`, `adopters` should return a vector indicating the doctors who began prescribing tetracycline in that month. If `not.yet` is `TRUE`, then `adopters` should return the vector indicating the doctors

who began prescribing tetracycline *after* that month (or never did). Check that `adopters(2)` indicates 9 doctors began prescribing in month 2, and that `adopters(month = 14, not.yet = TRUE)` indicates that 23 doctors began prescribing after month 14, or never did. Your work from Homework 9a Part I may help here.

2. Create a vector which stores the number of contacts each doctor has. Do not use a loop. Check that doctor number 41 has 3 contacts.

3. Write a function `count_peer_pressure` which takes in the index number of a doctor and a month and returns the number of doctors whom that doctor names as contacts, *and* had begun prescribing tetracycline by that month or earlier. If it is working properly, doctor number 37 and month 5 should return 3. Your work from Homework 9a Part II may help here.

4. Write a function `prop_peer_pressure` which takes in the index number of a doctor and a month and returns the proportion of the doctor's contacts who are already prescribing tetracycline by that month. If a doctor has no contacts, your function should return `NaN`. Check that doctor 37, month 5 returns a proportion of 0.6, but doctor 102 in month 14 returns `NaN`. Your function should call, not repeat, your function from (3) and use your vector from (2).

5. Write a function which takes in a month and returns a vector of length 2. The first element of the returned value should be the average proportion of prescribers among contacts of doctors who *began* prescribing in that month. The other should be the average proportion of prescribers among contacts of doctors who began prescribing *later*, or never. Call your code from (1) and (4); avoid using a loop by using an appropriate function from the apply family.

6. Plot the average proportions from (5) over time on the same graph. Do not use a loop and add an appropriate legend. Do the doctors who adopt in a given month consistently have more contacts who are already prescribing than non-adopters?