STAT GR5206 Homework 9a [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.

Ideas which we take today as common such as 'viral marketing' and 'early adopters' grew from sociological studies on the diffusion of information. One of the most famous such studies tracked how a then-new antibiotic, tetracycline, spread among doctors in four small cities in Illinois in the 1950s. In this lab we will study this data and study the idea that the innovation (in this case tetracycline) 'spread' from one person to the next.

Download the two data files `ckm_nodes.csv` and `ckm_network.txt` which store information on each individual doctor in the four towns and on which doctors knew each other, respectively.

Part I

1. Load the `ckm_nodes.csv` data into a data frame called `nodes`. It should have 246 rows and 13 columns. 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`. Answer the following. (a) How many doctors *began* prescribing tetracycline in each month of the study? (b) How many never prescribed it? (c) How many are NAs?

2. Create a vector which records the *index numbers* of the doctors for whom `adoption_date` is not `NA`. Check that this vector has length 125. 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.

3. Create a plot of the number of doctors who began prescribing tetracycline each month versus time. (The number on the x-axis can be integers instead of formatted dates.) Produce another plot of the *total* number of doctors prescribing tetracycline in each month. The curve for total adoptions should first rise rapidly and then level out around month 6.

4. Create a logical vector which indicates for each doctor whether they had begun prescribing tetracycline by month 2. Convert it to a vector of index numbers. There should be twenty such doctors. Create a logical vector which indicates for each doctor whether they began prescribing tetracycline after month 14. Convert it to a vector of index numbers. There should be twenty-three such doctors.

Part II

5. 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`. Verify that gives you a square matrix which contains only 1s and 0s with 246 rows and columns. Drop the rows and columns corresponding to doctors with missing `adoption_date` values. Check that the result has 125 rows and columns. Use this reduced matrix, and its rows and column numbers for the rest of the homework.

6. 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.

7. Create a logical vector which indicates, for each doctor, whether they were contacts of doctor 37, *and* had begun prescribing tetracycline by month 5. Count the number of such doctors without converting the logical vector to a vector of indices. There should be three such doctors. What proportion of doctor 37's friends do those three doctors represent (use (6.) here)?