Stats 598z: Midterm exam 1

Important:
Write you name and PUID on all sheets, and include the number of sheets
There are 8 questions, each for 5 points (but not all equally easy)
Attempt all questions, and when appropriate include a brief justification of your answer
Don't spend time polishing your answers as the main idea is more important
For maximum points use vectorization, but you won't lose too many points by looping

1. `age_vec` is a vector of integers, containing the ages of a group of people. Write `R` code to:
   (a) give how many people have age 25, and their positions in `age_vec`.
   (b) give how many people are aged between 20 and 30 (inclusive)
   (c) give how many people are 2 years younger than the next person in the vector (ignore the last person).
2. Define a vector `my_vec <- 1:10`
   (a) What is the result of `my_vec[-1] - my_vec[-length(my_vec)]`?
   (b) What is the result of `my_vec[c(TRUE,FALSE)]`?
   (c) What is the result of `my_vec[11 - my_vec]`?
   (d) If you run `my_vec[1] <- 1`, is `my_vec` now any different from before? Explain.
3. (a) `mtrx1 <- matrix( c(1,2,3,4), nrow = 3, ncol = 4)` . Write out `mtrx1`.
   (b) Write a few lines of `R` to create an `m` × `n` matrix, whose (i, j)th element is ij
   (c) Briefly explain the difference between a matrix, list and a dataframe (and when you might use each). Is a dataframe a list?
4. Recall that `rnorm(n,a)` returns `n` Gaussian variables with mean `a`.
   (a) Write a function that takes `a` as an input parameter, and keeps sampling from a Gaussian with mean `a` until a negative value is encountered. The function returns the sum of all positive numbers that preceded it. You can use loops if you want.
   (b) Write a function that takes `m` and `n` as input, and returns a `m`×`n` matrix. The first column consists of Gaussians with mean 1, the second, Gaussians with mean 2 etc.
   (c) Briefly explain what global variables are and why using them is a bad idea.
5. (a) Briefly explain why vectorization is faster than looping.
   (b) You given a vector of length 1000 consisting of sequences of integers repeating a number of times (e.g. `c(1,1,1,4,4,1,1,1,2,2,...)`). Write `R` code to give a matrix, the first row being the element of each repetition, and the second row being the number of times it is repeated. For the example, the first row is `c(1,4,1,2,...)`, and the second row is `c(3,2,3,2,...)`.

6. (a) Briefly explain what lazy evaluation is. What is the difference between `&` and `&&`?

   (b) Suppose you run the lines:

   ```
   a <- 5
   add_5 <- function() {
      a <- a + 5
      return(a)
   }
   b <- add_5()
   ```

   What are the resulting values of a and b? Explain.

   (c) Suppose you further run the lines:

   ```
   wrapper <- function() {
      a <- 10
      return(add_5())
   }
   b <- wrapper()
   ```

   What are the resulting values of a and b? Explain.

7. (a) Create a dataframe whose first column (call it x) is the numbers 1 to 100. It's second column (call it y) is the reciprocal of x plus random Gaussian noise.

   (b) Use `ggplot` to plot x vs y as a green line.

   (c) Suppose you also wanted all positive y's to be dots of one color, and all negative y's to be dots of another. Give a few lines of `R` to do this.

8. (a) Briefly explain k-nearest neighbors. Explain why it is a nonparametric method. Given an example of parametric regression.

   (b) You are given `my_vec`, a vector of real numbers. Given a new number `query`, you want to find the 5 elements of `my_vec` that are closest to `query`. Give a few lines of `R` to do this.