Stats 598z: Midterm exam 1

Important:
Write you name and PUID on all sheets, and include the number of sheets
There are 7 questions, question 1 has 8 points, all the rest have 7.
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.
Also, not all questions are equally easy, even if they are for the same points.
For maximum points use vectorization, but you won't lose too many points by looping

---

1. `univ_ranking` is a dataframe, each row giving information about a different university. Its columns are `name, type, endowment, tuition, accept_rate, start_salary` and `score`. Write `R` code to give:

   (a) the total number of universities, and the number of universities of `type` equal to `"public"` and `"private"`.

   (b) the average endowment of all schools with `score` greater than `5`.

   (c) the name of the school with the highest `tuition`. Also, the `"public"` school with the highest `tuition`.

   (d) how many schools have `accept_rate` between 10 and 20 percent (inclusive).

   (e) What does the command `table(univ_ranking$type)` do?

2. (a) Coercion in `R` can happen in two directions, from a more general type to a less general one (e.g. double to boolean) and vice versa. Given an example of coercion for both these cases.

   (b) Explain the difference between `&` and `&&`. What is the output of `c(0,1,5,0,1) & c(0,0,0,NA,NA)`?

   (c) Explain why comparing two variables of type `double` using `==` is a bad idea. How would you do this instead? Provide one or two lines of `R` code.

3. (a) `my_mat` is a matrix with even number of rows. Write `R` code to print every alternate row (rows 1, 3, 5 etc.)

   (b) Write an `R` function that takes a square matrix as input, and sets elements above the diagonal (i.e. elements (i, j) with $i > j$) to 0.

   (c) What is the output of `matrix(c(1,2,3) > c(2,3), nrow=2,ncol=2)` (including errors/warnings)?

4. `my_vec_short` is a vector of numbers interpreted as `(start1, len1, start2, len2,...)`. This compactly represents a sequence `(start1, start1+1,..., start1+len1-1)` (of length `len1` starting at `start1`), followed by `(start2, start2+1,..., start2+len2-1)` (of length `len2` starting at `start2`) and so on. Thus, `(4,2,3,1,2,3,1,1)` expands to `(4,5,3,2,3,4,1)`.

   (a) Write a few lines of `R` code that expands `my_vec_short` to its longer form (call it `my_vec_long`).

   (b) Write a few lines of `R` code that compresses a vector like `my_vec_long` to its shorter form.

   (c) In `R`, lists are more flexible than atomic vectors. Explain the advantage of working with atomic vectors when possible, instead of using lists all the time.

5. (a) For the `movies` dataset of question 1, give `ggplot` commands to plot histograms of `tuition` at schools of different `types` in different color on the same plot. Give a sketch of what your plot might look like (I don't care about the values themselves, only how the plot looks).

   (b) Give `ggplot` commands to plot `start_salary` vs `endowment`, with schools with accept rate greater than 50% and less than 50% in different colors.

   (c) Explain the different between setting `color = 'blue'` inside and outside `aes()`.

6. (a) Give `R` code to fit a linear model giving `score` as a function of all other variables except `name`

   (b) Fit a linear model of `score` against `start_salary` and `start_salary` squared, ensuring that `start_salary` equal to 0 results in a `score` of 0.

   (c) Explain overfitting and underfitting, and how cross-validation helps deal with these problems. Does k-nearest neighbors tend to overfit or underfit as k increases? Explain briefly.

7. (a) What does the command `rnorm(5,,10)` do.

   (b) Briefly explain what lexical scoping is in `R`.

   (c) Explain what the code below does:
   ```
   x <- rnorm(1)
   x < 0 && x <- 0.
   ```
   Rewrite it using an `if` statement.

   (d) `my_df` is a dataframe with `m` rows and `n` columns. What does `length(my_df)` return? What does `length(my_df[1])` return? And `length(my_df[[1]])`?