Stats 598z: Midterm exam 2

Important:
Write you name and PUID on all sheets, and include the number of sheets
There are 6 questions, questions 1 and 6 have 9 points, the rest 8 points
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

---

1. (a) A dataframe `my_class` has columns `puid`, `degree`, `major`, and `grade`. In the following, use commands from `tidyverse` for full points, but you will get most points if you just use base R.
   - Write code to sort the rows of the dataframe in decreasing order of grade.
   - Write code to calculate the number of students of every `(degree, major)` pair as well as the mean score.
   - Write code to calculate the number of students of every `major` with a grade greater than 90.
   - Write code to plot histograms of `grade`, with each `(degree,major)` pair having its own color.

   (b) Provide short examples displaying some data in both tall and wide format.

2. You have a vector of filenames `my_files`. Write down the `R` regular expression command to find:

   (a) files whose names contain "`data`" in them

   (b) files whose names end with "`.txt`"

   (c) files whose names contain two or more numbers in a row.

   (d) Suppose all filenames contain a date, taking the form `MM/DD/YYYY` (and no other numbers). Write `R` code to replace the date in the names to `DD/MM/YYYY`.

3. (a) Briefly describe the simplex (or Nelder-Mead) algorithm at a high-level.

   (b) In the context of object-oriented programming, describe 1) inheritence 2) polymorphism 3) generic functions.

   (c) What are the outputs for `2 %>% '/'(4)` and `2 %>% '/'(4,.)`?

4. (a) Explain briefly what reactive programming is.

   (b) For the dataframe `my_class`, use `ggvis` to plot a histogram of grades, setting the number of bins with a slider.

   (c) Explain what the command `log1p(x)` does, and how it is different from `log(1+x)`.

   (d) Explain briefly why `.3*4 == 1.2` might behave unexpectedly, and how you might do this correctly.

5. (a) What is the motivation for LASSO? Write down the LASSO loss-function.

   (b) What is coordinate descent?

   (c) Recall that LASSO tries to find the `w` that minimizes the LASSO loss-function. Suppose you wanted to find the average value of the LASSO loss function for a fixed `X,y, lambda`, with `w` distributed as a Gaussian random variable. How would you do this? Provide a few lines of `R` code.

6. (a) Explain what the `set.seed()` function does, and why it is useful.

   (b) Consider rolling a fair die until the number 6 shows up. The number of rolls required is a random number. Write R code to get a Monte Carlo estimate of the average the value of this quantity.

   (c) Describe importance sampling at a high-level. Explain (e.g. with an example) how it is useful for estimating quantities involving rare events.