Stats 598z: Midterm exam 2

Important:
Write you name and PUID on all sheets, and include the number of sheets
There are 7 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

1. `my_vec` is a vector of length 1000 consisting of sequences of values repeating a number of times (e.g. `c(1,1,1,2,2,2,2,2,1,1,3,3,3,...)`) Write down `R` code to calculate
   (a) the number of change points (three in the snippet above)
   (b) the average length of each sequence
2. You have a vector `my_words`. Write down the regular expression to find components containing
   (a) `x` followed by one or more `z`'s followed by a `y`
   (b) `x` followed by one or more vowels followed by a `y`
   (c) `x` followed by one or more `+`'s followed by a `y`
   (d) You want to find `x` followed by one or more `z`'s followed by a `y`, and replace it with `y` followed by the same number of `z`'s followed by an `x` (e.g. `xzzzy` becomes `yzzzx`). Write down the `R` command for this.
3. (a) Write down the LASSO loss function. Explain the role of λ. What is the solution for $\lambda = \infty$? What happens when $\lambda = 0$?
   (b) In the homework, we solved LASSO for $\lambda = 1$ using the `optim` function. In this case, we had no restrictions on w. Explain what you would do if you wanted to solve LASSO, but wanted the components of w to be nonnegative. Provide `R` code for the function you would optimize `my_loss` as well as how you would call `optim`.
4. (a) Explain what a generic function is and when it is useful.
   (b) You have a function `do_calc` that does some calculations and returns a dataframe. You also have a function `plot_result` to plot it. Explain the steps involved so you can plot it just by calling the generic function `plot`.
   (c) Explain briefly when `NextMethod()` might be useful.
5. A data frame `my_df` has columns `state, city, year, temperature` and `rainfall`. Each row gives the average temperature and rainfall recorded in a city in a state for a particular year. Below, to get full points, use `melt()`, `dcast()` or `*pply`, but if not sure, use any other approach:
   (a) How would you convert it to a new data frame consisting of `state, city, year, measurement` and `value`, where `measurement` is either `temperature` or `rainfall`?
   (b) How would you convert it into a new dataframe giving the average temperature in a given state (averaged across all cities, and ignoring rainfall).
6. Let x be distributed as a Gaussian with mean 0 and variance 1, and y with mean m and variance 1. You want to calculate $p(x < y)$, the probability that x is less than y.
   (a) Describe a simple Monte Carlo approach to calculating this. Write a few lines of `R` that does this.
   (b) Describe for what values of m this might not be efficient, and explain very briefly how to address this
7. (a) Describe Metropolis-Hastings at a high level, and give the acceptance probability.
   (b) Describe rejection sampling at a high-level
   (c) You want to sample from the distribution $p(x) \propto |\sin(x)|$ over the interval $[-\pi, \pi]$ using rejection sampling. Suggest a suitable proposal distribution and write down its acceptance probability.