Statistical Methods for Data Science (Fall 2017)

Mini Project 2

Instructions:

- Due date: Sep 28, 2017.
- Total points = 20.
- Submit a typed report.
- You can work on the project either individually or in a group of no more than two students. In case of the latter, submit only one report for the group, and include a description of the contribution of each member.
- It is OK to discuss the project with other students in the class (even those who are not in your group), but each group must write its own code and answers. If the submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Do a good job.
- You must use the following template for your report:

  Mini Project #
  Name
  Names of group members (if applicable)
  Contribution of each group member
  Section 1. Answers to the specific questions asked
  Section 2: R code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.

---

1. (a) Simulate $n = 30$ draws from a $N(0, 1)$ distribution and draw a normal Q-Q plot.
   (b) Simulate $n = 30$ draws from a t distribution with 2 degrees of freedom and draw a normal Q-Q plot and an appropriate t Q-Q plot.
   (c) Simulate $n = 30$ draws from a gamma distribution with mean 2 and variance 3, and draw a normal Q-Q plot and an appropriate gamma Q-Q plot.
   (d) Repeat each of (a), (b), and (c) 3 more times, and comment on what you observe.
2. Consider the dataset stored in the file `singer.txt`. It contains heights in inches of the singers in the New York Choral Society in 1979. The data are grouped according to the voice part. There are four voice parts, namely, Bass, Tenor, Alto, and Soprano. The vocal range for each voice part increases in pitch from Bass to Soprano. Perform an exploratory analysis of the data by examining the distributions of the heights of the singers in the four groups. Comment on what you observe. Do the four distributions seem similar? If not, in what respects do they seem different? Which probability distribution, if any, would you use to model these data? Justify your answers. (Try to use `subset` and `by` functions in `R` for this part.)