EDA Homework 2

Due: Wednesday, February 10 at 5pm.

Submit exactly two files: (i) a PDF/HTML file with your write-up and graphs and (ii) a .r/.txt/.Rmd file with code to reproduce your graphs.

Let's look at the CyTOF dataset we discussed in class.

It is available at http://jfukuyama.github.io/teaching/stat670/notes/cytof_one_experiment.csv.

The csv has 50000 rows; if the computation is too slow because of the number of observations, feel free to subset down to a smaller number.

1. Convert the dataset from wide to long form as we did in class. Make a faceted plot showing the distributions of each of the markers and describe these distributions. What features do you notice? Are there groups of markers that seem more similar to each other?
2. Choose two of the markers and make a Q-Q plot comparing their distributions. What is the relationship between the distributions? Can you describe it as a simple shift, or is it something more complicated? Is it different for different parts of the distribution?
3. Compute some summary statistics of the markers, describing their center, spread, or any other relevant features you can think of. Plot these statistics and interpret what your results. How do you like these plots compared with the plots of the full distribution?