Stat 470/670 Homework 2

Due: Friday, January 27 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 results.

For the first set of questions, we will look again at the CyTOF data at
http://jfukuyama.github.io/teaching/stat670/notes/cytof_one_experiment.csv.

Each row in the dataset represents a cell, and each column in the dataset represents a protein, and the value is element i, j of the dataset represents the amount of protein j in cell i.

1. Use `pivot_longer` to reshape the dataset into one that has two columns, the first giving the protein identity and the second giving the amount of the protein in one of the cells. The dataset you get should have 1750000 rows (50000 cells in the original dataset times 35 proteins).
2. Use `group_by` and `summarise` to find the median protein level and the median absolute deviation of the protein level for each marker. (Use the R functions `median` and `mad`).
3. Make a plot with `mad` on the x-axis and `median` on the y-axis. This is known as a spread-location (s-l) plot. What does it tell you about the relationship betwen the median and the mad?

Next, for more practice pivoting, we will look at a dataset from `dcldata`. Install the package `dcldata` using

```
install.packages("remotes")
remotes::install_github("dcl-docs/dcldata")
```

(you don't need the first line if the `remotes` package is already installed). Load the package using `library(dcldata)`.

Load the dataset `example_gymnastics_2` using the command `data(example_gymnastics_2)`. Notice that the column names are of the form `event_year`.

4 Using either `pivot_longer` on its own or `pivot_longer` in combination with `separate`, reshape the dataset so that it has columns for country, event, year, and score.