**Apply data wrangling on my capstone**

Jihye Ko

The dataset that I am using is from kaggle.com, and it was from a 2014 survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. Luckily, all variables of this dataset were very well defined and easy to understand. This data has 26 variables, and among them, I chose ‘no\_employees (How many employees does your company or organization have?)’ as a dependent variable.

For data wrangling, all I needed to do with data was just finding missing values, and figuring out how to handle those missing values. In this case, only ‘no\_employees’ had mission values among variables which I am interested in. There are two ways to handle those missing values. First, just omit those values. Second, substitute with either mean or median. Since there are few missing values, I and my mentor thought that omitting them were not a good choice. Therefore, we decided to replace them with either mean or median. To choose which one would be a better choice, I made a histogram to see whether ‘no\_employees’ is normally distributed or skewed. If it is normally distributed, then it is better to use mean. If not, it is better to use median. Since ‘no\_employees’ was not normally distributed, I replaced missing values to median.