**Apply statistics to my capstone project**

Jihye Ko

Previously, I choose variables that I am interested in (self\_employed, family\_history, treatment, work\_interfere, no\_employees, remote\_work, tech\_company, benefits, care\_options, wellness\_program, seek\_help, anonymity, leave, mental\_health\_consequence, phys\_health\_consequence, coworkers, supervisor, mental\_health\_interview, phys\_health\_interview, mental\_vs\_physical, obs\_consequence), and among them I picked ‘no\_employees’ as a dependent variable. Since I wanted to figure out what other variables affect on the dependent variable, I needed to choose the most proper variables among them listed above.

After a long discussion with my mentor, we decided that figuring out which variables are strongly correlated is the first step. In my data, most of them are characters not numbers. In other words, since this data was from a survey, few variables have two answers (Yes/No), and other variables have three answers (Yes/No/Don’t know). Therefore, I categorized variables into two (variables with two answers and variables with three answers). For each category, I changed character value to numeric value to use corrPlot. By using corrPlot, I plotted the correlation on each category and found correlation coefficient. Then, on each category, I figured out which variables were strongly correlated to each other based on correlation coefficient values. The highest correlation coefficient value was 0.62 between ‘seek\_help’ (Does your employer provide resources to learn more about mental health issues and how to seek help?) and ‘wellness\_program’ (Has your employer ever discussed mental as part of an employee wellness program?). Therefore, I chose them as two independent variables.

Since I chose both dependent and independent variables, now it’s time to see their correlation. For further analysis, I needed to go into machine learning. But before that, I wanted to know brief relation between ‘no\_employees’ vs. ‘seek\_help’ & ‘wellness\_program’. So, I plotted them by using ggplot. I set ‘seek\_help’ as x-axis, ‘wellness\_program’ as y-axis, and ‘no\_employees’ as color. By doing so, I could see which answers of each variable are strongly correlated with which size of the company briefly.