**STAT 4355 Project Proposal**

Group Members: Ann Biju, Cindy Cho, Amulya Rayasam, Ally de Vera

Datasets: <https://www.kaggle.com/datasets/shariful07/student-mental-health> and <https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey>

Team Name: Mental Health Maniacs

**INTRODUCTION TO THE DATASETS:**

Our first dataset is concerning student mental health and how it affects their cumulative GPAs. This dataset looks at students across different fields of study and gauges the presence of various mental health issues among students. The second dataset looks at the presence of mental health issues of individuals working in tech. It also looks at how various aspects of their work affect the presence of these mental health conditions or vice versa.

**ANALYSIS GOAL:**

One goal of our analysis and project is to determine which factors (or variables) have a strong correlation with the mental health of college students, as well as, tech workers, in the United States. Another goal of this project is to unearth any possible links or strong patterns between college students majoring in STEM fields and their industry counterparts. We realized that as STEM majors ourselves, the knowledge gathered from such data would be precious to our understanding of a major facet of our futures.

**DATA VARIABLES:**

We have two datasets: the Student Mental Health dataset with 11 variables, and the Mental Health in Tech Survey dataset with 26 variables. We will select certain variables from these datasets to examine changes in mental health from the student period to working in the tech industry.

The Student Mental Health dataset includes the following variables:

* Timestamp
* Gender
* Age
* Course
* Current year of Study
* GPA
* Marital status
* Depression
* Anxiety
* Panic attack
* Seek any specialist for a treatment?

The Mental Health in Tech Survey dataset includes the following variables:

* Timestamp
* Age
* Gender
* Country
* States
* Self\_employed
* Family\_history: family history of mental illness
* treatment: Receive treatment for a mental health issue
* work\_interfere: Does your mental health condition affect your work?
* no\_employees: Number of employees in company/organization
* remote\_work: Work remotely at least 50% of the time
* tech\_company
* benefits: Does your employer provide mental health benefits?
* care\_options: Do you know the options for mental health care your employer provides?
* wellness\_program: Did your employer include mental health in an employee wellness program?
* seek\_help: Does your employer offer resources to educate about mental health concerns and how to get help?
* anonymity: Is your identity kept confidential if you use mental health or substance abuse treatment resources?
* leave: Is it easy for you to take medical leave for a mental health issue?
* mental\_health\_consequence: Do you believe that disclosing a mental health issue to your employer would result in negative outcomes?
* phys\_health\_consequence: Do you believe that discussing a physical health issue with your employer would have negative repercussions?
* coworkers: Are you comfortable discussing a mental health issue with your coworkers?
* supervisor: Are you comfortable discussing a mental health issue with your supervisor?
* mental\_health\_interview: Would you mention a mental health issue to a potential employer in interview?
* phys\_health\_interview: Would you mention a physical health issue to a potential employer in interview?
* mental\_vs\_physical: Do you think your employer regards mental health as important as physical health?
* obs\_consequence: Have you witnessed or been aware of any adverse effects for coworkers with mental health conditions in your workplace?

**ANALYSIS PLAN:**

**Preliminary Data Evaluation:**

* Firstly, we would assure that we are working with clean data, removing NA’s by either deleting or replacing them depending on what makes the most sense for each individual variable, making sure all variables are properly assigned to data types, and making sure all factor levels match when necessary (i.e. ‘year 2’ vs ‘Year 2’ vs ‘2’ vs ‘yr 2’... )
* Determine how our sub groups are going to be split and what our primary goals are for each group.

**Analysis:**

* From our data, a large portion of what we have to look at are binary attributes, so the most logical approach would be to find how each of our data with a wider range of values (state, age, degree, etc.) affect these values, likely using bar plots/ boxplots.

**Results and Conclusions:**

* Based on what each of our pairings find, we will reconvene and determine which variables have a notable correlation as well as which charts show the most clear evidence overall and we will work together to draw conclusions from each of them.

**RESPONSIBILITIES:**

Since our group is analyzing 2 datasets, we will have two partnerships of two group members each working on one assigned dataset. Once each dataset has been thoroughly analyzed, the whole group will come together to discuss their findings and compile the observations into a presentation format. This is where we will choose which plots and variables show the strongest evidence out of all of the graphs. At this stage in the project, it is vital for all four group members to be familiar with each others’ findings to be able to answer any queries from the audience, when time comes for presentations.

* Mental health in tech
  + Using treatment as criteria variable for deciding whether a person has mental illness or not, since no other variable can comfortably predict that label: for example, someone could have a mental illness present but it not affect their work, or someone could have a mental illness present but not have any family history of it
  + Based on categorical variables in data, we determined to create our largest model using the glm() function as opposed to the lm() function.
  + Using deviance residual specific to glm function for categorical variables https://towardsdatascience.com/diagnose-the-generalized-linear-models-66ad01128261#:~:text=Usually%2C%20the%20deviance%20residual%20is,using%20the%20'residuals'%20function.