Laura Abood - axu8rx
Kelly Gu - hnj4jk
Edison Huang - hmc4zu
Alisha Qian - sta2fu
Carlie Stewart - ayu6cp
Hana Wang - muf5cb

The purpose of the pre-analysis plan is to get you to commit to a framework for doing your analysis that makes sense, given the tools you've seen so far and the data you have. Pre-analysis plans are common in experimental fields, in order to bring some discipline to the data analysis.

- Goal: predict mental health status based on workplace conditions and other features
- Variable: Predict if employee is comfy sharing mental health concerns with supervisor

Github:
https://github.com/kkellygu/ds3001Project

You should address the following questions explicitly:

- **What is an observation in your study?**
    - An observation in our study is represented by each of the survey responses.
    - A possible observation could be employees who report experiencing high levels of stress, anxiety, poor management, lack of support, etc leading to burnout and/or possible depressive symptoms compared to employees who experience a more positive, healthy working environment and greater access to mental health resources.
- **Are you doing supervised or unsupervised learning? Classification or regression?**
    - Supervised learning allows the model to learn to predict an outcome based on input features. In our case, we have survey responses from employees regarding their workplace conditions and mental health status.
    - We will use classification because our target variable is categorical (yes, no, maybe). This allows us to categorize employees based on if they are comfortable sharing mental health concerns with their supervisor. Supervised classification is appropriate for this objective because it allows us to directly assess and model the relationship between workplace conditions and whether or not employees are comfortable sharing mental health concerns with their supervisor.
- **What models or algorithms do you plan to use in your analysis? How?**
    - Logical regression for binary classification
        - since most all of our data consists of yes/no responses, we can use logistic regression to begin to examine relationships between variables, like the availability of mental health support, to help estimate the likelihood of a certain outcome based on our independent variables
    - Decision trees

        - can help visualize how different factors will lead to a specific outcome – the data would get split into subsets based around the more prevalent variables and it’d work like a flowchart in going from top to bottom to then help see how certain patterns/pathways of yes/no contribute to mental health outcomes in the work environment
- **How will you know if your approach "works"? What does success mean?**
    - We know our model works if we can categorize test examples correctly with a f1 score (which measures recall and precision) of >= 0.85.
- **What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?**
    - Weaknesses that we anticipate happening are responses that may dominate the dataset. For example, if a large company has responses from the majority of their employees, this may impact the results of the cluster analysis. We can deal with this if it comes up by using oversampling or undersampling techniques.
    - The data itself is a weakness because we can only rely on user-reported yes/no responses, which significantly limits what we’re able to work with. In replying only yes/no, these kinds of responses may oversimplify underlying issues; there’s no room for explanation, different extents, etc. A simple yes or no doesn’t reveal anything deeper than surface level content, which could then lead to superficial conclusions made from the data which then may not fully be accurate of the population

You should address the following topics in the text, as appropriate:

- **Feature Engineering: How will you prepare the data specifically for your analysis? For example, are there many variables that should be one-hot encoded? Do you have many correlated numeric variables, for which PCA might be a useful tool?**
    - First, any missing values in our dataset should be cleaned. For our dataset, there are several categorical survey questions that should be one-hot encoded, such as “Does your employer provide mental health benefits as part of healthcare coverage?”, “Do you know the options for mental health care available under your employer-provided coverage?”, and more. These questions are primarily related to yes/no responses and can be transformed into binary features to be used for our model. There are also some numeric variables where PCA may be beneficial, such as the number of employees the company has. However, the dataset is predominantly made up of categorical data, which may limit the use of PCA.
- **Results: How will you communicate or present your results? This might be a table of regression coefficients, a confusion matrix, or comparisons of metrics like R^2 and RMSE or accuracy and sensitivity/specificity. This is how you illustrate why your plan succeeded or explain why it failed.**
    - Our preliminary approach to communicating our results is to use a confusion matrix, since this is the best way to present our model that is working on primarily categorical variables. With a confusion matrix, we can also report metrics such as

the F1 score to showcase the accuracy of our model. However, this may be subject to change according to the nature of our model and there could be better ways to present our results.