Homework # 4 (200 points)

For this assignment, we will be working with a very interesting mental health dataset from a real-life research project. All identifying information, of course, has been removed. The attached spreadsheet has the data (the tab name "Data"). The data dictionary is given in the second tab. You can get as creative as you want. The assignment is designed to really get you to think about how you could use different methods.

1. Conduct a thorough Exploratory Data Analysis (EDA) to understand the dataset. (**20 points**)

2. Use a clustering method to find clusters of patients here. Whether you choose to use k-means clustering or hierarchical clustering is up to you as long as you reason through your work. You are free to be creative in terms of which variables or some combination of those you want to use. Can you come up with creative names for the profiles you found? (**40 points**)

3. Let's explore using Principal Component Analysis on this dataset. You will note that there are different types of questions in the dataset: column: E-W: ADHD self-report; column X – AM: mood disorders questionnaire, column AN-AS: Individual Substance Misuse; etc. You could just use ONE of the sets of questionnaire, for example, you can conduct PCA on the ADHD score, or mood disorder score, etc. Please reason through your work as you decide on which sets of variables you want to use to conduct Principal Component Analysis. What did you learn from the PCA? Can you comment on which question may have a heavy bearing on the score? (**40 points**)

4. Assume you are modeling whether a patient attempted suicide (column AX). This is a binary target variable. Please use Gradient Boosting to predict whether a patient attempts suicides. Please use whatever boosting approach you deem appropriate. But please be sure to walk us through your steps. (**50 points**)

5. Using the same target variable (suicide attempt), please use support vector machine to model this. You might want to consider reducing the number of variables or somehow use extracted information from the variables. This can be a really fun modeling task! (**50 points**)