# BST-260 Project

* <https://docs.google.com/forms/d/e/1FAIpQLSfvmqoLqWyTVeUWqdkekeU92bGZWA66nx8Gl3YLMxCfkrQIgQ/viewform>

Question/topic (1-2 sentences)

Using regression to examine association between covariates and depression onset and aim to use machine learning methods in predicting individual level depression.

Dataset (description, source/URL, how many observations, summary of columns)

* <https://www.kaggle.com/datasets/francispython/b-depression>

<https://www.kaggle.com/datasets/diegobabativa/depression>

<https://www.kaggle.com/code/diegobabativa/kernel-b-depressed>

* 286+1143 = 1429 observations. 75 columns: including individual identifiers, where they’re from, survey date, gender, age, marital status, number of children, household size, years of education completed, Number of children <=18 or younger in Household, Non-durable expenditure (USD), Value of livestock (USD), Value of durable goods (USD), Value of cell phone (USD), Value of savings (USD), Land owned (acres) and other socioeconomic status indicators. (Column Description: <https://www.kaggle.com/datasets/francispython/b-depression/discussion/87722>)

Analyses/skill you will use (3-5 sentences)

* Regression: first I will be using the correlation content in exploratory data analysis in examining the association between each factor and outcome depression. I would also attempt regression models to fit with a number of the covariates that I found with higher correlation and see how that turns out in the summary output.
* Machine Learning: Aim to find the most optimal predictive model. Attempt linear regression, KNN, and random forest with cross-validation. I will use metrics such as confusion matrix to evaluate performance