**Assignment 2**

Topic: **Depression in Adolescents**

**Predictive:** “Can we predict the risk of depression among 12–17-year-old adolescents in New York City based on information about their demographic status, eating habits, neighborhood characteristics, and socioeconomic factors?”

To answer the above predictive question, I would rely on data-driven methods. For my study sample, I would collect data on demographic status, eating habits, SES, and neighborhood characteristics. After cleaning the data and observing the distribution of the available data, I would perform a correlation analysis to check whether any of my hypothesized predictors correlate with known depression endpoints. After operationalizing variables, I would train a machine learning model through different methods, such as logistic regression and random forest models, to predict the risk of depression. I would evaluate the performance of these models through precision-recall curves and F1 scores. After choosing the most parsimonious and efficient model, I would use cross-validation to validate the model and generalize findings. Although the process heavily depends on data-driven methods, the hypothesized predictors are a product of theory-driven methodology.

**Explanatory:** “Is low socioeconomic status of parents a cause of higher depression rates among 12-17-year-old adolescents in New York City?”

I would use a balanced combination of data-driven and theory-driven methods to answer the question. Based on literature reviews and preliminary research data, I would select what data is needed to assess a causal hypothesis – such as demographic data, data about adverse events in childhood, medical history, eating habits, etc. After exploring data distribution through numerical summaries, histograms, and correlation analyses, I would select variables to go into my final logistic regression model. I would use theory-driven methods to assign roles to variables other than SES to account for confounders, mediators, and effect modifiers. Finally, I would evaluate and validate my logistic regression model through likelihood ratio tests, p-values, and global F statistics.

It is paramount to note that I used theory to select and assign roles to the different variables in my explanatory model. However, in the predictive model, I assigned no roles to my variables and used machine learning (data-driven method) to rank the importance of predictors, effectively assigning them a role in the predictive analysis.