Intrduction

We begin by cleaning the data. Since our raw data file contains hundreds of features, we create a new data matrix with only the desired predictors (see the Data section for further details). We also create several new features by combining columns, which imroves the regression model's interpretability by creating more consise predictors. All this is done through the `code/scripts/data-processing-script.R` script.

Next we consider a cross-validated linear regression to predict the mean earnings of female students working and not enrolled 6 years after entry (MN_EARN_WNE_MALE0_P6) using the new data matrix, which we will use to create a scoring system for all schools. We use the mean earnings as a metric to represent the overall quality of a college's graduates, the intuition being that smarter more qualified candidates will be paid more. After preforming the regression, we check that all features are significant and use the weights of all significant terms.

We then use the regression equation, with only the significant terms included, to create a score for each school. After giving each collge a score, we will rank them by said score and present the top institutions to our client.