1. **What is an observation in your study?**

Each observation in the study is an institution in the US. Each observation includes key characteristics about a college or university like enrollment, student aid, costs, and student outcomes.

2. **What is our main question/prediction?**

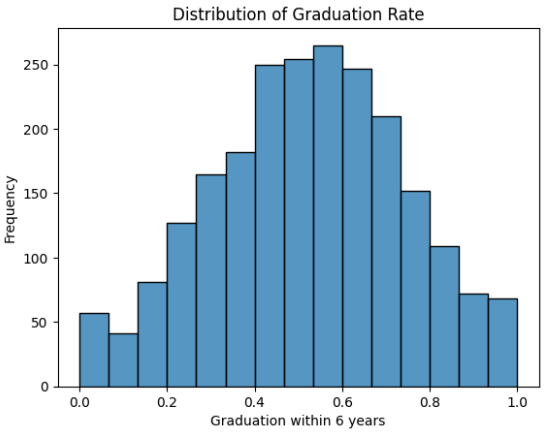
Our main question is: “how do socioeconomic factors, particularly average family income, predict graduation/completion rates at US colleges?” We plan to focus on C150_4 (completion rate for first-time, full-time students at four-year institutions within 1.5x expected completion) as the single target variable to narrow our analysis and allow us to more easily discern which variables are most strongly associated with higher C150_4 values. We hypothesize that lower average family income, higher average cost of attendance, higher DCS federal loan rate pooled, and higher DCS Pell grant rate pooled will be correlated with lower completion rates. This question is significant considering the tension between American ideals of “pulling yourself up by your bootstraps” and hidden structural barriers that influence educational outcomes & social mobility; exploring how differing financial backgrounds might lead to educational disparities can prompt deeper reflection on the notion that success is solely based on individual effort and illustrate the need for equitable policy that enables student success across socioeconomic divisions.

3. **Are you doing supervised or unsupervised learning? Classification or regression?**

We plan on using a combination of supervised and unsupervised learning. We’re doing supervised learning with OLS regression to predict student outcomes. We want to learn what factors most accurately predict successful student outcomes like earnings after 4 years, so all of the data points are already “labeled.” In other words, if we want to predict how much your income will increase depending on other factors like cost of attendance, all of the observations already include the income information which will act as the label. We also plan on doing regression since the vast majority of our variables are numeric. The only significant variable that is categorical is school type like public vs private, but the variables we want to predict are numeric. We will also create a second model using Principal Component Analysis to create proxies for the groups of variables in buckets (those representing socioeconomic status as well as school prestige) and incorporate those into a linear regression model.

4. **What models or algorithms do you plan to use in your analysis? How?**

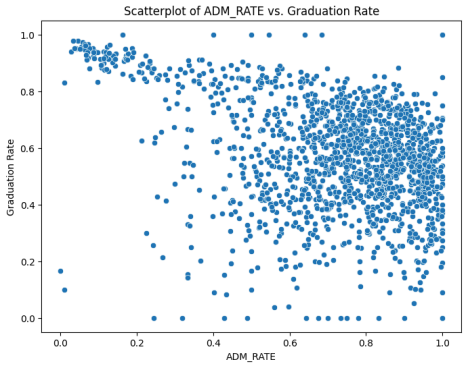
Based on the histogram to the right we can see that our target variable, graduation rate, is continuous, numeric, and the distribution seems relatively unimodal without skew, which indicates it could work for linear regression.



We plan on using multiple linear regression using the ordinary least squares (OLS) method to quantify the predictor values on C150_4 (completion rate for first-time, full-time students at four-year

institutions). This will allow us to determine which coefficients to use for each variable (i.e. Average Family Income Percentage Student receiving Pell Grant etc).

We chose this tool because it is easily interpreted and it will allow us to explore interactions between the variables as well as incorporating nonlinear, higher power terms into our model. For instance the relationship between admission rate and graduation rate (shown in the scatter plot below) appears to be non-linear, so we will attempt to incorporate this into our model.



To build this model we will do an 80/20 train test split, and build the model using the training data, putting aside the test data to test the model later to help prevent/be able to identify overfitting on the data.

In addition to the supervised learning method above where we include all variables that seem relevant, we realized that various explanatory variables we plan on using to build our model fall into two main categories: measures of average socioeconomic status at the school (Average Family Income, Percentage of Students receiving Pell Grants etc.) and measures of prestige of the school (Midpoint ACT score, Admission Rate etc.). So, because these variables are largely attempting to measure similar things, using all of them in the model would likely lead to multicollinearity and overfitting. We will make another model using the unsupervised learning algorithm of Principal Component Analysis to create proxies for the groups of variables in buckets (those representing socioeconomic status as well as school prestige). We will then compare this to the model trying to use them all, comparing both against the test set. One problem with doing this is that it reduces explainability of the model because it is difficult to make sense of what the transformed variables really mean. It is also susceptible to outliers and overfitting.

5. **How will you know if your approach "works"? What does success mean?**

Since our data is split into training and testing data, we will first train both of our models with the training data and then be able to assess the approach using the testing data set. Specifically, we will get the R squared and RMSE of the test set with the trained model. We will determine which model is best based on which performs better with the test set.

We will run linear regression of graduation rate on multiple variables such as acceptance rate and income. Then also check the R squared and RMSE to see how accurate these variables are in predicting the graduation rate. When we built a linear regression of graduation rate on acceptance rate, act score, and cost of attendance, the R squared score was 0.739 and RMSE: 0.0957. So this shows that these variables have a high rate of prediction and it has a low error in prediction. Just with this the model seems successful, because it is more accurate than not, however we would need to run more tests to be definite. We are trying to determine which model is most successful and will use multiple models to determine this.

We will also perform residual analysis; we will calculate the difference between observed and predicted values (residuals) and plot them against the predicted values. This will allow us to check for patterns related to linearity or potential outliers; if the residuals are all randomly scattered around zero, our model is capturing a linear relationship.

In addition, we want to know how applicable our model is to a real world scenario of a student choosing a school. To work towards this we will create a post-hoc situation where we use the best model (based on r-squared and RMSE from the test set) to rank schools based on our target variable given the explanatory variables. To do this we would create a simulated app where a prospective student could enter 5 schools they were interested in and we would use our model to rank those schools based on predicted graduation rate. To test this we will compare our rankings to the actual graduation rates in some sort of systematic way that is unfortunately still TBD (we will consult you :)).

6. **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?**

What we anticipate will be an issue could be wading through the exact measures that actually correlate with our question of which variables will answer the question of whether or not college is a “scam” and is college ultimately worth it. We are examining variables such as demographics, geographic influences, socioeconomic variations, and institutional differences between schools in terms of tuition. Most notably however, we did not include variables such as majors or what students studied at each of these institutions which could be relevant. We are specifically looking only at the year 2022 and 2023 which could potentially have Covid 19 impacts that might make our model less applicable to other years. Furthermore, the question of whether or not something is worth it is somewhat incomplete without not just data about completion but also data such as the employment rates after graduation and mean/median salaries after graduation at these institutions. Breakdowns of salary and income and employment/unemployment would give more comprehensive insight into whether or not college is a “scam” aka not worth the money put in.

Our weakness could come from the fact that there are simply so many variables that could potentially come into play and isolating the individual or power of one variable is also difficult due to the fact that these variables are also highly correlated. For this reason, we would engage in using PCA to deal with multicollinearity in such a high dimensional data set.
If our approach fails, we might come to surmise that there are additional variables not counted for or that we had focused on the wrong variables.

Also, we have many numeric variables like act score, cost of attendance, acceptance rate, and more. But the one categorical variable that is usable and could be a good predictor, public or private school, could skew the RMSE and R squared. WE want to be both categorical and numerical, but it is difficult to use both so finding a way to combine those results we get from both, could be difficult.

7. **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?**

The first thing that would have to be done is pre-processing the data by handling the missing data, identifying and managing the outliers, and transforming these categorical variables. Additionally, we will use the variable on the most common degree awarded (PREDDEG) to narrow our data frame to schools that primarily offer undergraduate degrees to increase similarity and comparability across our observations. One hot coding is used for binary categorical pieces of data. So the variable on whether the school is public, private nonprofit, or private for-profit (CONTROL_PEPS) would be good candidates for one-hot encoding since they are binary, categorical pieces of data. Secondly, for the many correlated numerical variables, we will perform PCA on them to reduce multicollinearity. An example of such variables would be PCTFLOAN_DCS_POOLED_SUPP and PCPPELL_DCS_POOLED_SUPP in order to provide insight into financial assistance and socioeconomic demographics between students and how this would relate to variables such as college completion rate.

8. **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.**

We will also compare test-set r-squared and RMSE from each of the models to communicate accuracy of the models. We will also present a table of regression coefficients to improve explainability and understandability of the model, such that those viewing the results can understand how each explanatory variable plays into the model.

We will compare the correlation between socioeconomic variables like average family income, average cost of attendance, DCS federal loan rate pooled, DCS Pell grant rate pooled, and completion/graduation rate. We will present this in a table. Higher correlation values would indicate stronger associations between the predictor variable and completion rate, indicating higher predictive power for the variable. We might also create visualizations based on the correlation matrix, such as a heatmap, for more effective visual summarizing/communication and to make pattern observation easier.

We will also communicate our results using the app discussed above where a student can enter schools and we will rank them by predicted graduation rate.