To develop the most effective regression model for predicting graduation rates using the IPEDS data, we began by refining the variables in the original dataset. Our first step was to eliminate variables that we believed lacked meaning or were redundant. Specifically, we removed all in-district measurements, as we believed they did not apply to most schools. Additionally, we excluded cbsatype, considering it too similar to the locale variable that we kept.

With this refined dataset, we first standardized the numerical predictors to ensure that all predictors were on a comparable scale. We constructed our first model with the rest of the data and chose to exclude interaction terms, using Stepwise Regression to identify the most significant predictors. The model with the lowest AIC included the following variables:

Categorical Variable:

* c21enprf

Numerical Variables:

* Cohort Size
* Grant Rate
* Pell Rate
* Loan Rate
* In-State Tuition
* In-State Fees
* Average Salary

Recognizing the potential for variable interactions to improve model performance, we constructed a second model incorporating two-way interaction terms. After evaluating different steps in the selection process, the step with the lowest AIC combined main effects and several interaction terms, as highlighted in the accompanying Excel document.

For our third and final model, we included two-way interactions between all included variables, ensuring that all potential relationships were accounted for. However, despite the added complexity, this model did not reduce AIC further and introduced an excessive number of predictors. Since it did not provide improvements, we decided to not look at this model any further.

After evaluating our models, we determined that model 2 was the best-performing, with the lowest AIC of -5729.4824 and 21 predictors. This model balanced predictive power and simplicity well, making it our final choice for predicting graduation rates.