Developing an Open-Source College and University Prediction Model for Tuition Costs of United States Institutions

Report Prepared by

Dylan Zelko; Mechanical Engineering B.S. Candidate

Jack Mandura; Mechanical Engineering B.S. Candidate

Hunter Babel; Electrical Engineering B.S. Candidate

Emmitt Stores; Chemical Engineering B.S. Candidate

Report Intended for EAS 345: Introduction to Data Science

December 2024

Table of Contents

1. Executive Summary

This report outlines the development of an open-source college and university tuition prediction model for United States institutions. Rising tuition costs pose a significant challenge for students and families, with current tools failing to provide personalized and predictive insights useful for assessing future situations. To address this gap, our team designed a model to estimate tuition costs based on institutional characteristics and historical data, offering a transparent, accessible, and user-friendly solution where users can type in compatible schools to gather data quickly.

The project utilized data from the U.S. Department of Education's College Scorecard, including metrics such as SAT scores, completion rates, and average faculty salaries. Through data cleaning and exploratory data analysis (EDA), over 7,000 initial entries were refined to a manageable dataset of approximately 900 institutions. Statistical techniques, including multiple linear regression, were employed to build and iterate upon the predictive model, with variables selected based on correlation strength and statistical significance.

The final model incorporates factors such as level of urbanization, school size, SAT scores, and completion rates. While our final model demonstrated acceptable performance compared to other iterations, limitations in R^2 values and error rates indicate room for further improvement. Advanced modeling techniques like principal component regression and ridge regression were identified as potential future approaches to enhance accuracy and reduce variance.

The results demonstrate the feasibility of creating an open-source tuition prediction model and provide a foundation for continued development and refinement. With future iterations, this tool has the potential to significantly impact how students and families navigate the complexities of college affordability.

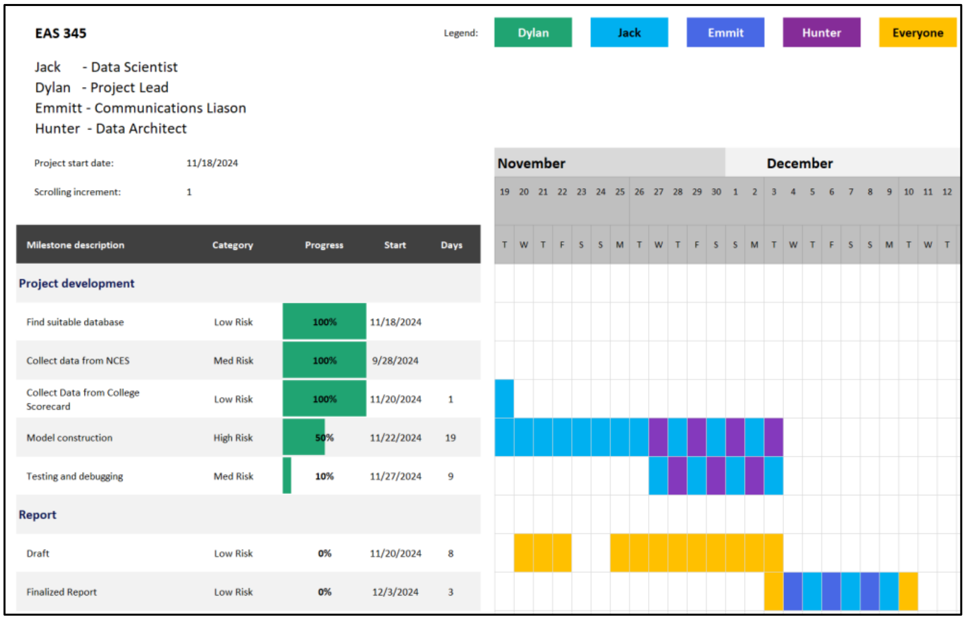
2. Project Management

This section of the report will begin with an in-depth look at the team member breakdowns. Defined roles, responsibilities, and workloads will be detailed for all four members. Then an analysis into the problems faced by the group and the resulting solutions. This section will end with a critical look at the timeline for the project, how it changed over time, and how the results were able to produce within a timely manner. This will include a Gantt chart which the team used to stay on track.

One of the first steps within this process was to assign team members with roles and responsibilities. This played a crucial part in the early success of the team when having to outline which member had which responsibility. With a team composed of four intelligent engineering undergraduates, the team was confident in any one member's ability to execute any given task. With this freedom, it allowed the team to play to individual strengths. Dylan Zelko headed the project, being given the title of Project Lead. This bestowed responsibility of ensuring deadlines were met and done so in a high-quality manner. As project lead, Dylan was also to give every deliverable the final approval. The workload put forth by Dylan was helping fellow members in data collection and cleaning, as well as assisting in the creation of various group documents submitted to the sponsor. Hunter Babel was assigned the role of Data Architect, giving responsibilities to handle data collection and data cleaning. Hunter completed his responsibilities with great precision, handling a large data set and making it manageable for the next team member. His workload not only included this data cleaning, but also in assisting in the production of the models used in this report and the final presentation. Jack Mandura boasted the role of Data Scientist, due to some experience in coding. Jack lead the project in terms of data sourcing, data cleaning, and the exploratory data analysis (EDA). He also created the models for tuition cost and future variable predictions. This would carry into filling out the respective reports, such as Milestone 3, Final Presentation, and this report. Emmitt Stores rounded out the team's roles with Communication Liaison and Scribe. As the team's liaison, it was Emmitt's responsibility to communicate with the sponsor, set up the various team meetings, and head the writing of this report. As scribe, Emmitt's duties were to record team minutes in meetings and ensure that team members who could not attend a given meeting were given a summary of said meeting. The workload for Emmitt dealt mainly with quality assurance of written documents, producing deliverables, and writing the bulk of most reports (action, final, etc.)

Throughout this project, very few challenges arose within the team that were outside of actual data. Most common was the issue of finding a time which fit well for all members of the team. Each member barred a busy schedule, with few times of overlapping availability between all four members. To resolve this, a meeting would only take place if at least three of the four members could be in attendance. This way, a majority of members were in attendance and confusion or falling behind on update was minimized. Assuring filling in the missing member was done through online communication and worked well. No other glaring issue had presented itself within the group. All members performed exceedingly well together and upheld a high degree of cohesion with one another.

Observing the evolution of this project's timeline includes minor changes to deadlines and delivery dates. Some of these changes were due to a change within the course structure, others were a result of a change in the group's ability to complete a task within a certain time. As aforementioned, all members of this group supported rather intense schedules, which led to some changes in the ability to complete something ahead of time as initially proposed. Our earliest timeline was created before official deadlines were created, resulting in it being nearly completely different from the final timeline (Fig. 1). The first projection was to have the project completed by November 17th, an entire month before the eventual deadline. This of course changed, allowing us to spend more time on the model, and this report.



EAS 345

Jack - Data Scientist
Dylan - Project Lead
Emmitt - Communications Liason
Hunter - Data Architect

Project start date: 11/18/2024

Scrolling increment: 1

Milestone description	Category	Progress	Start	Days
Project development				
Find suitable database	Low Risk	100%	11/18/2024	
Collect data from NCES	Med Risk	100%	9/28/2024	
Collect Data from College Scorecard	Low Risk	100%	11/20/2024	1
Model construction	High Risk	50%	11/22/2024	19
Testing and debugging	Med Risk	10%	11/27/2024	9
Report				
Draft	Low Risk	0%	11/20/2024	8
Finalized Report	Low Risk	0%	12/3/2024	3

Figure 1: Gantt Chart

The Gantt Chart produces a timeline for the process of producing the team's statistical model along with this report. The timeline shows equal contribution from team members. Emphasizing the ability of the team's ability to play to members strengths. Members Jack and Hunter headed the Project Development portion with members Emmitt and Dylan heading the report's draft, with ongoing help from the entire team. The Gantt Chart highlights the cohesiveness of the group, showing the ability to produce the project's model, and this report, in a timely manner.

3. Problem Definition

This section of the report will discuss the importance and impact the produced model can have on its target audience. It will begin with a thorough explanation of the problem, defining it clearly with supporting evidence. Included in this section is an analysis of the current solutions for the problem this group is looking to address and solve. The current solutions will be evaluated and explained both to their benefits and their shortcomings. This will allow for an easier understanding of why the model proposed in this document is an improvement over the current scenes.

Importance and Impact of the Model

The rising cost of higher education in the United States is a major concern for students and their families. The complexity of understanding and predicting tuition fees often adds to the burden, particularly when reliable tools are not freely available or tailored to individual needs [1]. This report aims to address the lack of a free predictive model that offers future insights into tuition costs, incorporating factors like price range, preferred school size, and other institutional characteristics. The team's goal is to develop an open-source tuition prediction model that provides an accessible solution for millions of students annually.

Definition of the Problem

Currently, there is no widely available, free tool that allows users to dynamically predict college tuition costs based on their selected institution's characteristics. While various online resources provide tuition data, these tools often present static [2]. This lack of predictive capability is a significant barrier for students and families seeking affordable options.

This project addresses this gap by developing a model that predicts tuition costs based on historical and institutional data. By leveraging open-source technology and publicly available data, the creation of a transparent, reliable, and customizable solution that democratizes access to tuition information can be done.

Current Solutions and Their Analysis

Several tools and platforms currently assist users in exploring college tuition data, including resources like the College Board's BigFuture tool and U.S. News and World Report's college rankings. These paid platforms aggregate tuition data reported by institutions and present it through user-friendly interfaces, often with filters for school size, location, or public/private status. Some platforms also feature financial aid estimators or calculators that approximate net costs based on family income. While these features are helpful, like previously stated, they lack predictive capability, relying on static data rather than dynamic predictions.

One major drawback is the absence of predictive insights. These tools present current tuition figures but do not account for possible trends, inflation, or changes in funding structures [3]. Additionally, financial aid data on some platforms are proprietary, requiring subscriptions or limiting functionality to paid users, creating accessibility barriers for underrepresented or low-income groups.

Novelty of the Proposed Approach

For most sources dealing with college selection, information that is crucial to the user, such as tuition cost, is locked behind paywalls. Our prediction model relies on open-sourced data, and therefore is a free tool for users. Also, unlike these existing tools that present static tuition data, our model predicts future tuition costs based on future estimated aspects of a given institution. This ensures users gain insights into potential future costs, helping them make more informed long-term decisions.

Potential Shortcomings

Despite its significant improvements to current tools, the proposed model has potential limitations. Reliance on publicly available data ties the quality and completeness of predictions to the robustness of the datasets. Incomplete or outdated data from institutions could affect the model's accuracy. The open-source nature of the model also requires continuous maintenance and updates, posing logistical challenges.

Another consideration is how effectively critical variables will be predicted. If these aspects can't be properly estimated, then the tuition model itself loses proper footing. This is an issue that is suppressed during model evaluation.

In addressing these challenges, this project's approach aims to balance innovation and practicality, offering a tool that is both effective and inclusive. By overcoming the shortcomings of existing solutions and prioritizing user needs, this model has the potential to significantly enhance how prospective students and families plan for college tuition.

4. Data Acquisition, Cleaning, and EDA

Data Acquisition and Definition

The data for this project was sourced from the College Scorecard of the Department of Education. This provided a large range of colleges within the United States; approximately seven thousand. This data set consists of any information that could be of concern about a school. This in turn provides a large pool of dependent and independent variables to sort through, and begin finding trends in relationships. Located on the website of the Department of Education, the data can be downloaded freely. The R code provided consists of the data set, the steps taken towards cleaning the data, the various visuals created for the analyzation of the data, and the initial steps taken towards creating a model.

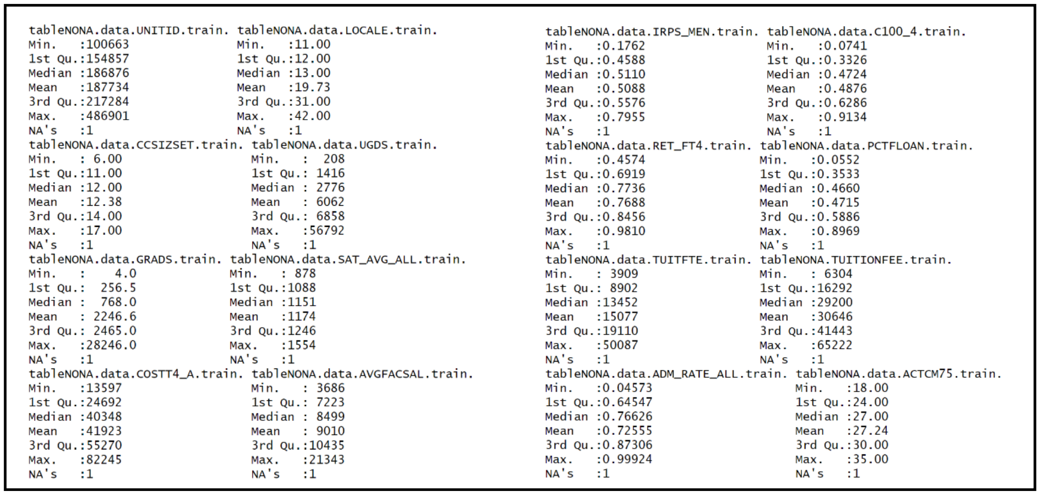
Within the repository, there is an RMD File and R File containing work concerning data cleaning, exploratory analysis, and modeling. Also included, is the data report containing the sources of data: The original data set marked as '*MERGED2022_23_PP.csv',* and each respective file for the years dating back to 2010. As well as the data dictionary that explains every variable listed within the data set, marked as '*CollegeScorecardDataDictionary.csv'*. Previous action reports and others can be located here as well.

Data Cleaning

The data set was downloaded and brought into R for cleaning: This began with creating a data frame consisting of only the desired variables found through the data dictionary. These data frames were initially large, with around *7000 x 4000* rows and columns, these columns were reduced to

17 entries. This table was then assessed for any columns consisting of too many null or *N/A* values, or even exclusively null values. These columns were deemed as unusable, and the respective variables were no longer considered for analysis. Other columns consisted of strings, an example being the school names. This required the names of schools to be substituted with their university ID number during calculations. However, this string data will be used later to identify each school during model use. Data that required exclusion involved the median debt and median household income of students: These columns were entered into the College Scorecard database as logical data types. However, the variables by definition are numbers and integers respectively. This led to the columns being completely *N/A*, and therefore unusable. Other years were checked that may have the data entered differently, but to no result. The data that was in fact usable consisted mostly of continuous data, such as tuition costs, and all ratios and rates. There were still some nominal data types, such as level of urbanization, but this proved less useful in testing. The most important variables were those that were discrete, such as number of students and SAT scores. The least useful data were those of ordinal types, such as school sizes, which lacked correlation to important variables.

For the remaining columns, most of the rows contained null values or factors that relate to an unknown value. When summarizing the tables, it was apparent as most minimums for variables were either zero or below zero; this skewed the means of most variables. This issue was circumvented by converting all null values to *N/A*, then removing the *N/A* values themselves. This removed most of the rows, and reduced the total count from seven thousand schools, to just over nine hundred. What remains is still a sufficient number of schools to divide into training and testing data, which is what followed. The cleaned data was divided into two halves: One half dedicated to training, while the other half is reserved for testing. The training data is the even entries in the data set, while the testing data is the odd entries. This data was summarized to view the new layout of means and outliers for each variable (Fig. 2).



```
tableNONA.data.UNITID.train. tableNONA.data.LOCALE.train.
Min.   :100663               Min.   :11.00
1st Qu.:154857               1st Qu.:12.00
Median :186876               Median :13.00
Mean   :187734               Mean   :19.73
3rd Qu.:217284               3rd Qu.:31.00
Max.   :486901               Max.   :42.00
NA's   :1                    NA's   :1
tableNONA.data.CCSIZSET.train. tableNONA.data.UGDS.train.
Min.   : 6.00                  Min.   :  208
1st Qu.:11.00                  1st Qu.: 1416
Median :12.00                  Median : 2776
Mean   :12.38                  Mean   : 6062
3rd Qu.:14.00                  3rd Qu.: 6858
Max.   :17.00                  Max.   :56792
NA's   :1                      NA's   :1
tableNONA.data.GRADS.train. tableNONA.data.SAT_AVG_ALL.train.
Min.   :    4.0             Min.   : 878
1st Qu.:  256.5             1st Qu.:1088
Median :  768.0             Median :1151
Mean   : 2246.6             Mean   :1174
3rd Qu.: 2465.0             3rd Qu.:1246
Max.   :28246.0             Max.   :1554
NA's   :1                   NA's   :1
tableNONA.data.COSTT4_A.train. tableNONA.data.AVGFACSAL.train.
Min.   :13597                  Min.   : 3686
1st Qu.:24692                  1st Qu.: 7223
Median :40348                  Median : 8499
Mean   :41923                  Mean   : 9010
3rd Qu.:55270                  3rd Qu.:10435
Max.   :82245                  Max.   :21343
NA's   :1                      NA's   :1
```

```
tableNONA.data.IRPS_MEN.train. tableNONA.data.C100_4.train.
Min.   :0.1762                 Min.   :0.0741
1st Qu.:0.4588                 1st Qu.:0.3326
Median :0.5110                 Median :0.4724
Mean   :0.5088                 Mean   :0.4876
3rd Qu.:0.5576                 3rd Qu.:0.6286
Max.   :0.7955                 Max.   :0.9134
NA's   :1                      NA's   :1
tableNONA.data.RET_FT4.train. tableNONA.data.PCTFLOAN.train.
Min.   :0.4574                Min.   :0.0552
1st Qu.:0.6919                1st Qu.:0.3533
Median :0.7736                Median :0.4660
Mean   :0.7688                Mean   :0.4715
3rd Qu.:0.8456                3rd Qu.:0.5886
Max.   :0.9810                Max.   :0.8969
NA's   :1                     NA's   :1
tableNONA.data.TUITFTE.train. tableNONA.TUITIONFEE.train.
Min.   : 3909                 Min.   : 6304
1st Qu.: 8902                 1st Qu.:16292
Median :13452                 Median :29200
Mean   :15077                 Mean   :30646
3rd Qu.:19110                 3rd Qu.:41443
Max.   :50087                 Max.   :65222
NA's   :1                     NA's   :1
tableNONA.data.ADM_RATE_ALL.train. tableNONA.data.ACTCM75.train.
Min.   :0.04573                    Min.   :18.00
1st Qu.:0.64547                    1st Qu.:24.00
Median :0.76626                    Median :27.00
Mean   :0.72555                    Mean   :27.24
3rd Qu.:0.87306                    3rd Qu.:30.00
Max.   :0.99924                    Max.   :35.00
NA's   :1                          NA's   :1
```

Figure 2: Summary of Training Data Set

This cleaning process was repeated for each data set of the years prior. Another issue encountered during cleaning was that most of these prior sets don't record data for all desired variables. Some of these variables would need to be discarded in the case of predicting future dependent values. For example, the ratio of male to female students was only recorded for the last two years. No other problems were found during cleaning.

Exploratory Data Analysis

After omitting the null valued and *N/A* entries, summaries were taken of the data. It was seen that no longer were the minimum and maximum bounds of any variable either zero or one. Other large outliers such as maximum tuition costs were no longer present as well: It appears that outrageous data values were mostly present with the schools with *N/A* entries for other variables, and were therefore removed. This indicated that cleaning was in fact successful, and a complete analysis can be made. During data cleaning, several columns were acquired that have a relation to tuition cost; Cost of Attendance (*COSTT4_A*), Tuition Revenue (*TUITFTE*), and Tuition Cost (*TUITIONFEE*) averaged between in-state and out-of-state. These three columns must be compared to determine which one best represents the cost of tuition. This was done through creating box plots of each column (Fig. 3). From the box plots, Tuition Cost has the most desirable distribution in the context of outliers and quartile ranges.

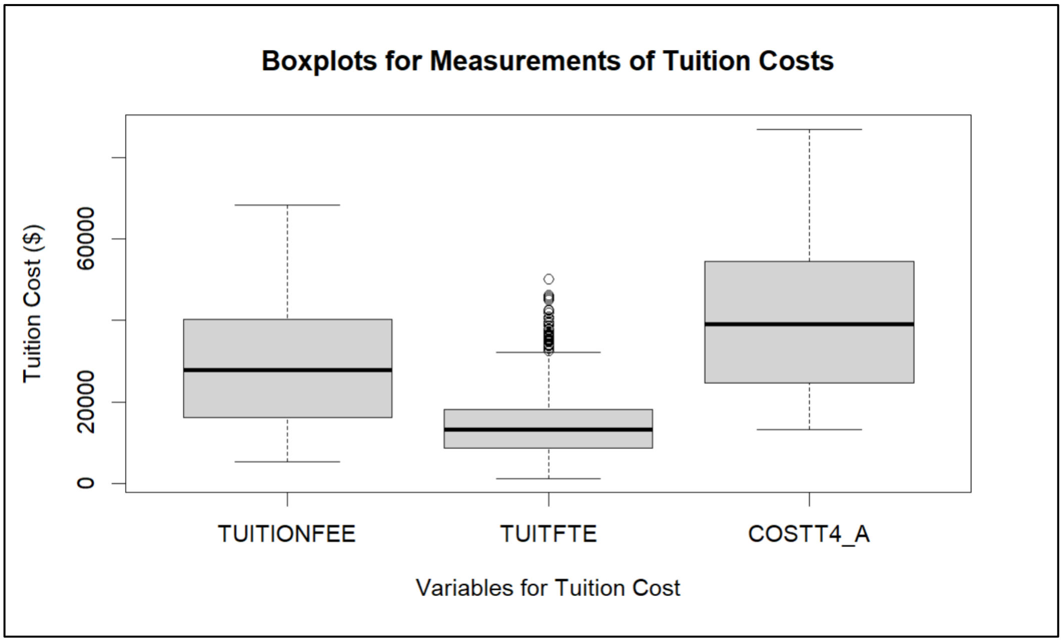


Figure 3: Comparison of Measurements for Tuition Cost

Box plots were generated for other variables as well, such as completion rate (Fig. 4). Here, distributions are somewhat vague, and show an even spread around fifty percent.

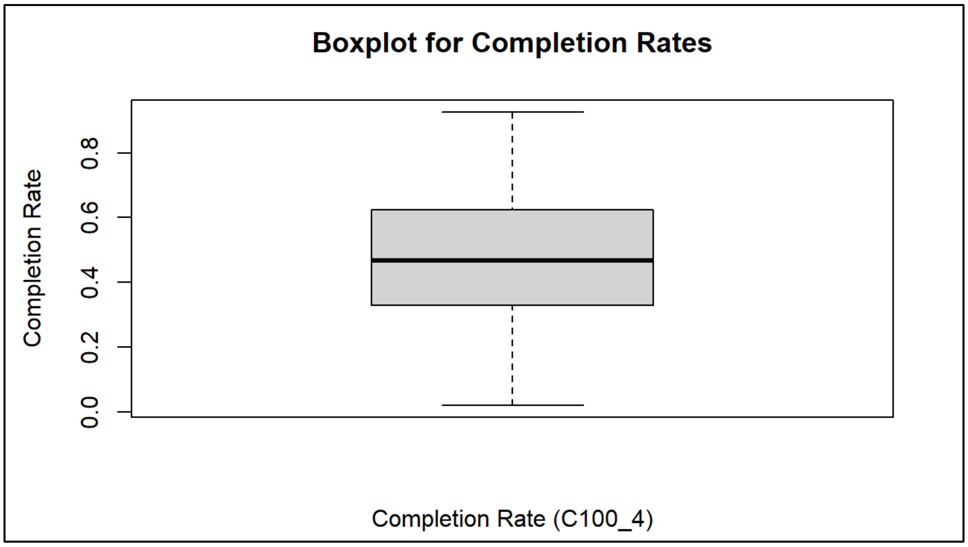


Figure 4: Visualization of Completion Rate Distribution

While using *summary()* was helpful to view the means and outer bounds of a variable, seeing the distribution via box plot is crucial in obtaining an understanding of how these variables behave. On top of box plots, bar charts and histograms were utilized as well to illustrate the same distributions. A bar chart was created (Fig. 5) to see how the schools compare to each other in terms of level of urbanization (*LOCALE*). It's seen that there's a higher count of schools in small rural towns than places more populated.

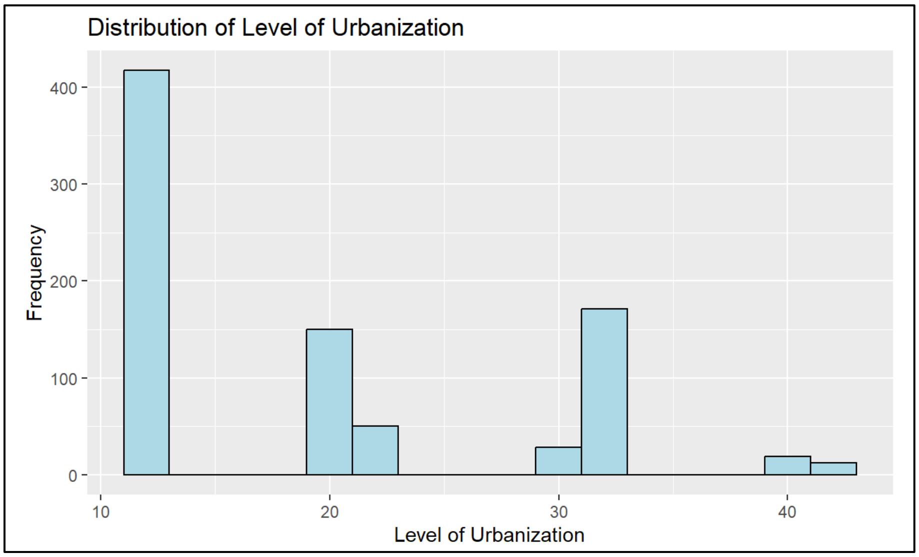


Figure 5: Bar Chart for Level of Urbanization of Schools

An interesting use of histograms was looking at the distribution of school sizes (Fig. 6), which shows a somewhat linear trend between school size (*CCSIZSET*), and how many of those schools exist. All of the variables mentioned and analyzed are tabulated (Table 1) with their respective definitions:

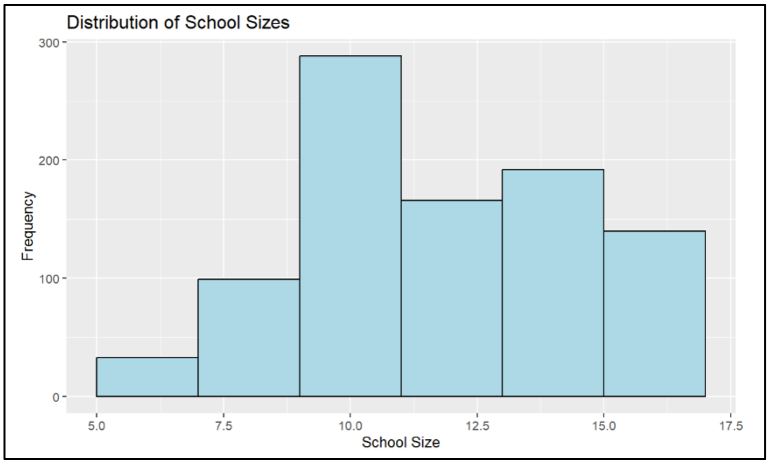
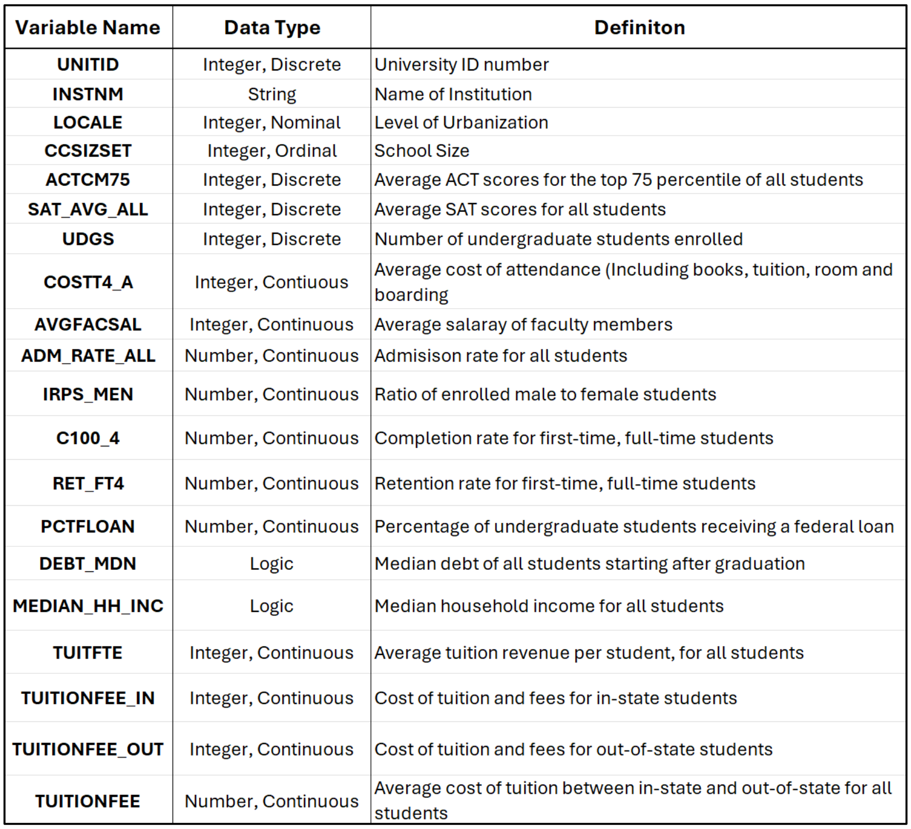


Figure 6: Histogram of School Sizes

Table 1: Variable Names, Data Types, and Definitions



Variable Name	Data Type	Definiton
UNITID	Integer, Discrete	University ID number
INSTNM	String	Name of Institution
LOCALE	Integer, Nominal	Level of Urbanization
CCSIZSET	Integer, Ordinal	School Size
ACTCM75	Integer, Discrete	Average ACT scores for the top 75 percentile of all students
SAT_AVG_ALL	Integer, Discrete	Average SAT scores for all students
UDGS	Integer, Discrete	Number of undergraduate students enrolled
COSTT4_A	Integer, Contiuous	Average cost of attendance (Including books, tuition, room and boarding
AVGFACSAL	Integer, Continuous	Average salaray of faculty members
ADM_RATE_ALL	Number, Continuous	Admisison rate for all students
IRPS_MEN	Number, Continuous	Ratio of enrolled male to female students
C100_4	Number, Continuous	Completion rate for first-time, full-time students
RET_FT4	Number, Continuous	Retention rate for first-time, full-time students
PCTFLOAN	Number, Continuous	Percentage of undergraduate students receiving a federal loan
DEBT_MDN	Logic	Median debt of all students starting after graduation
MEDIAN_HH_INC	Logic	Median household income for all students
TUITFTE	Integer, Continuous	Average tuition revenue per student, for all students
TUITIONFEE_IN	Integer, Continuous	Cost of tuition and fees for in-state students
TUITIONFEE_OUT	Integer, Continuous	Cost of tuition and fees for out-of-state students
TUITIONFEE	Number, Continuous	Average cost of tuition between in-state and out-of-state for all students

A correlation matrix was also used to see in a summary what variable has the highest correlation to what. From this matrix, we already start to see that SAT scores, completion rate, and retention rate have a larger correlation to tuition cost than other variables. The variables with the largest correlations are noted (Table 2). These are the variables that will first be experimented with when modeling.

Table 2: Correlation Matrix Results

Variable	Correlation
C100_4	73%
SAT_AVG_ALL	66%
ACTCM75	62%
RET_FT4	49%
ADM_RATE_ALL	-42%
AVGFACSAL	43%
UGDS	-21%

Overall, with our analysis, our end goal was to seek where most trends exist between the various dependent and independent variables. Therefore, several scatter plots were created to compare any two variables. With these plots containing best fit lines, some linear relationships began to become noticeable. One relationship in particular that showed a strong connection was between tuition costs and SAT scores (Fig. 7):

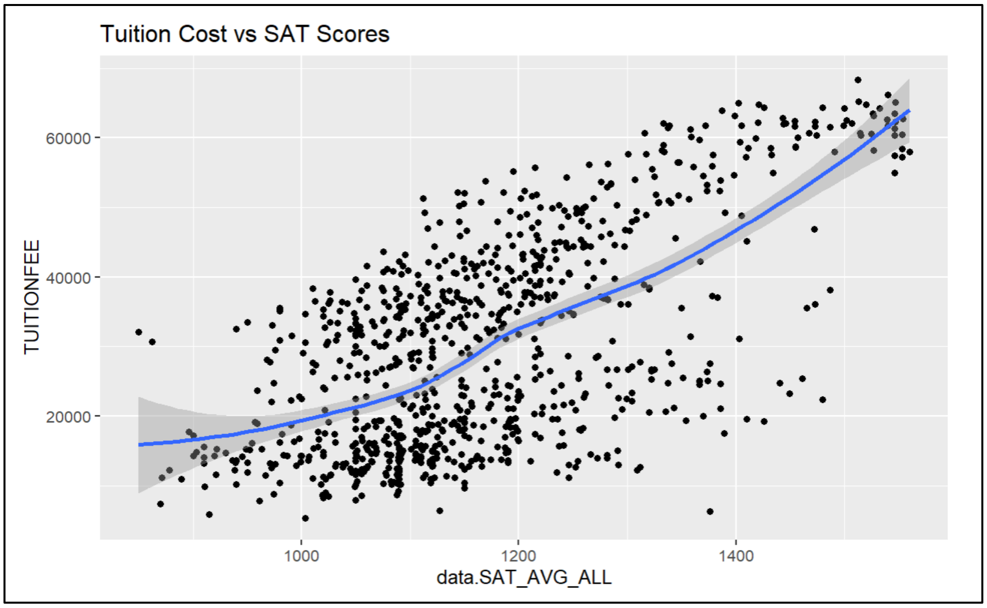


Figure 7: Scatter Plot of Tuition Cost Against SAT Scores

Similar trends were found between other dependent variables and tuition cost. Some trends that were previously thought to likely exist were shown to be random and nonexistent. An example of this being the relationship between tuition cost and school size, or even level of urbanization. Both were found to have little to no correlation when used with a scatter plot. On top of plots of independent variables, visualizations for tuition cost itself were created with the training data (Fig. 8). These help understand the dependent variable better, and see if there's any trends in the data. We see that the number of schools with lower tuition is the highest, and trend downward with increasing tuition cost.

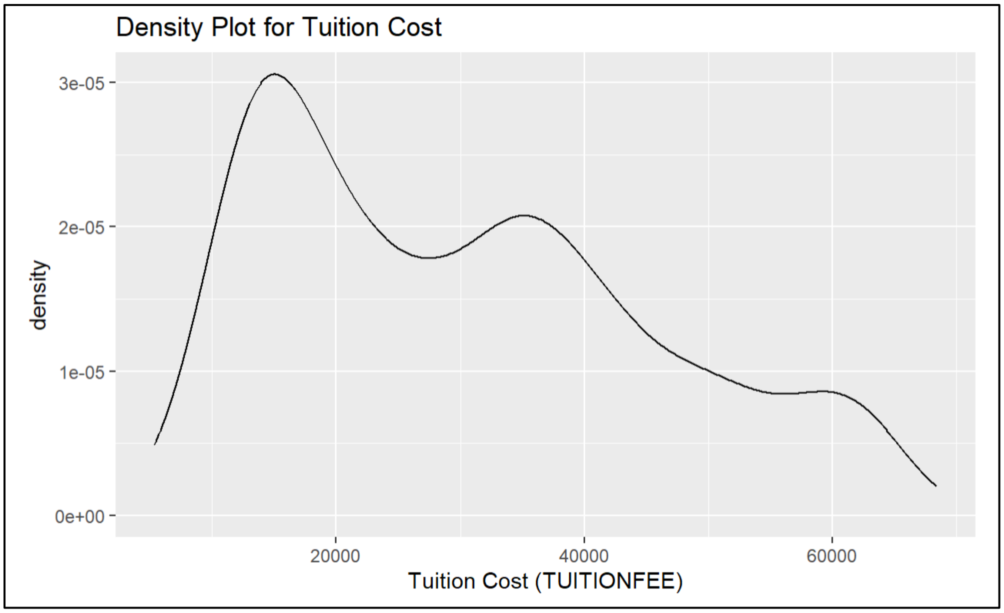


Figure 8: Visualization of Dependent Variable, Tuition Cost

After gaining an understanding of each variable for the most recent set, the remaining trends to find are those that each independent variable has with time. This will require the use of the data sets of previous years. Data sets were used for the last thirteen years, as data became random and skewed beyond 2010. Each independent variable was plotted against time, with the trend being visualized for one school at a time. Over the course of thirteen years, the trends of these variables are not linear. This makes sense, as trends are expected to change throughout a decade, and is seen with number of students (Fig. 9).

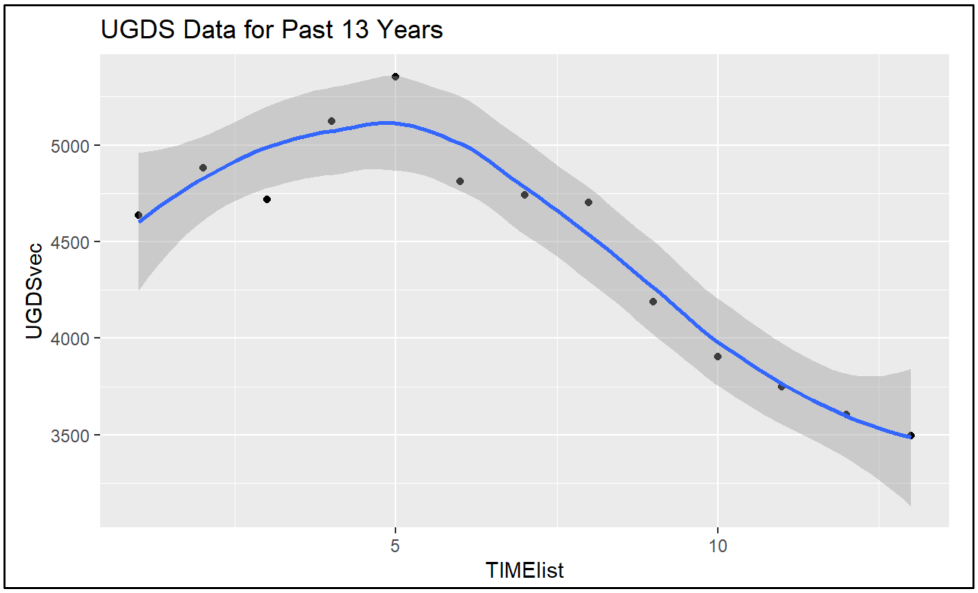


Figure 9: Plot of Number of Students Over the Last 13 Years

This issue was solved by accounting for the most recent years only. This smaller window allowed for linear trends to be seen, and greatly reduce error. For number of students, the trend is visible for a certain school when only viewing the most recent years (Fig. 10). From here, a clear and accurate prediction can be made from a single linear regression model.

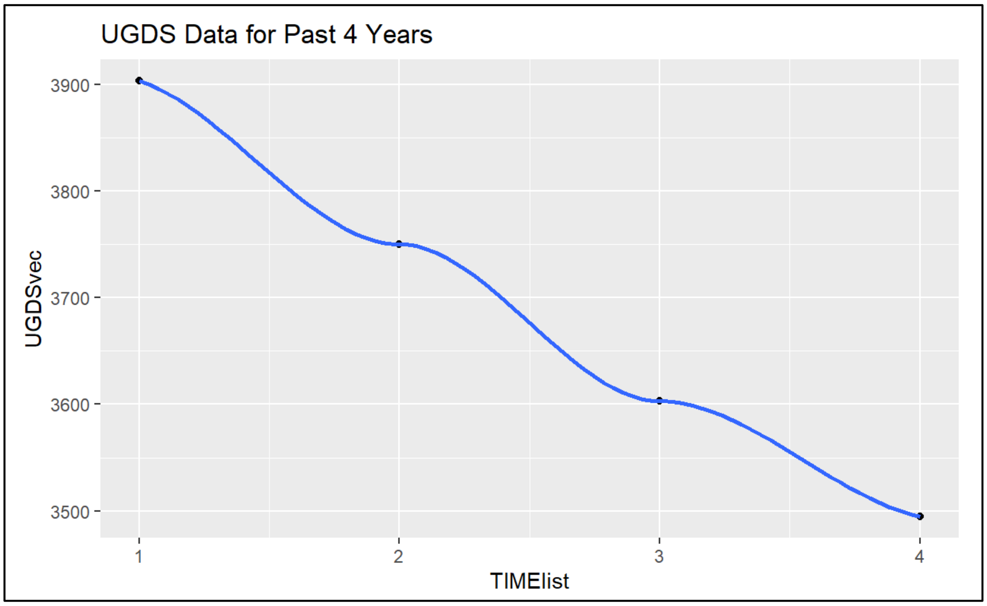


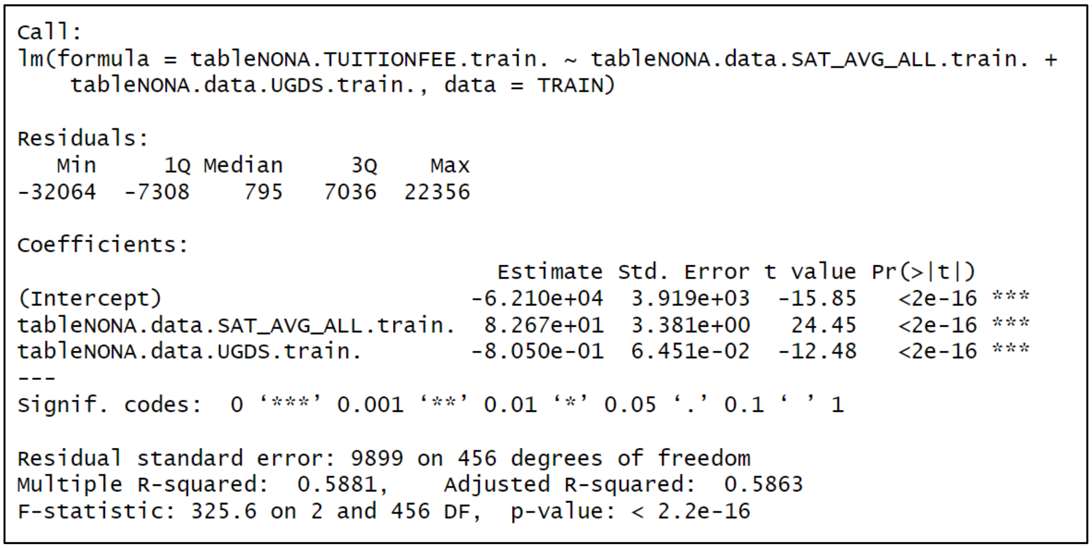
Figure 10: Plot of Number of Students Over the Last 4 Years

5. Model Creation and Validation

Modeling

While previously touched upon, our problem at-hand is to create a model that can predict the tuition cost of a college based on basic independent characteristics of the school. This requires the construction of a multiple linear regression model that will use several dependent variables such as SAT scores and number of students to estimate the dependent tuition cost. Regression functions of R will be utilized.

The function required for this model is *lm().* The inputs it takes are the desired dependent variable and the known independent variables. The outputs it spits out are the y-intercept and coefficients of the independent variables, which are located in the Estimate column. The Standard Error column shows the average amount that the estimates vary from the actual value. In our summary of *tuition_model2* (Fig. 11), we see that the standard error is high for the y-intercept and lower for the two coefficients. The y-intercept deals with tuition cost, which holds a much larger integer value than that of SAT scores. The t-value column simply provides the t-statistic of each coefficient and intercept: We want this to be relatively larger than zero, in order to reject the null hypothesis. This t-value goes hand in hand with the next column over which is the Pr column. This Pr column states the P-value for each coefficient. Obviously, we want this value to be as small as possible, as the P-value reflects the probability that the null hypothesis was rejected even though it’s true. This is clearly something that needs to be minimized.

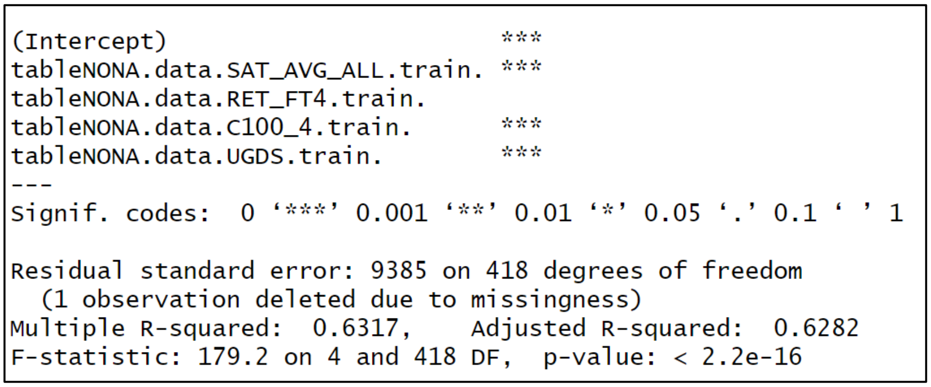


```
Call:
lm(formula = tableNONA.TUITIONFEE.train. ~ tableNONA.data.SAT_AVG_ALL.train. +
    tableNONA.data.UGDS.train., data = TRAIN)

Residuals:
   Min     1Q Median     3Q    Max
-32064  -7308    795   7036  22356

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -6.210e+04  3.919e+03  -15.85   <2e-16 ***
tableNONA.data.SAT_AVG_ALL.train.  8.267e+01  3.381e+00   24.45   <2e-16 ***
tableNONA.data.UGDS.train.        -8.050e-01  6.451e-02  -12.48   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9899 on 456 degrees of freedom
Multiple R-squared:  0.5881,    Adjusted R-squared:  0.5863
F-statistic: 325.6 on 2 and 456 DF,  p-value: < 2.2e-16
```

Figure 11: Initial Modeling Results for Linear Regression

The residual standard error explains the average amount that the estimate will deviate from the regression line, with the degrees of freedom representing how many rows of data we’re working with. In our case, the residual standard error is just under ten thousand, which is large considering the large range of tuition cost values. The degrees of freedom read out 456, which corresponds to our 456 entries in the training data set. The multiple R^2 value is an indication of how well the model fits the data; it represents how much of the variance in the dependent estimate can be explained by the independent estimates. Ideally, we want this R^2 value to be as close to one as possible, which corresponds to 100% variance explained. Currently, we see an R^2 value of 0.588 for our second model: This is pretty undesirable, as almost half of the variance in the tuition estimate cannot be explained by the two variables used. This can be increased as we increase the amount of variables in the regression model. Finally, the F-statistic is another indicator of the relationship between the independent and dependent variable. We want this to be a value a bit larger than one, with a small P-value. A small P-value associated with the F-statistic leans toward the idea that at least one independent variable was related to the dependent variable. In our second model, we have an F-statistic of 326, with an extremely small P-value. This means that at least one of the variables, SAT scores or number of students, is related to the tuition cost.

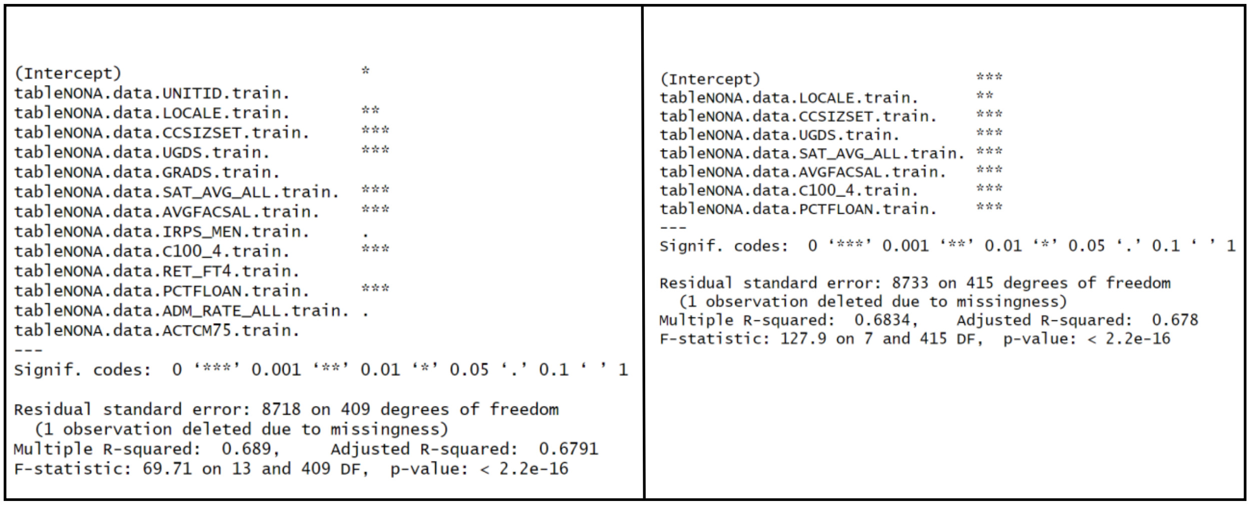
We now start with the variables picked from the correlation matrix with the highest correlation magnitudes. For this selection, the model was created and evaluated. The P-values of the independent variables were checked, and those with the highest values were removed. These variables were removed one at a time for each iteration, and was done until a satisfactory final version was reached, iteration 12 (Fig. 12):



```
(Intercept)                         ***
tableNONA.data.SAT_AVG_ALL.train. ***
tableNONA.data.RET_FT4.train.
tableNONA.data.C100_4.train.        ***
tableNONA.data.UGDS.train.          ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9385 on 418 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6317,    Adjusted R-squared:  0.6282
F-statistic: 179.2 on 4 and 418 DF,  p-value: < 2.2e-16
```

Figure 12: Final Iteration from Corr. Matrix: *tuition_model12*

From the summary of iteration 12, we see that there's an undesirable R^2 value of 0.63. We want a value that's at least 0.85. There's also a relatively high residual standard error of 9385. In the case of predicting tuition, we want an error much lower than nine thousand. After twelve iterations of the model using the correlation matrix, a desirable result wasn't reached. We'll now use the variables that were previously seen to have correlations to tuition from the EDA plots, instead of only considering the correlation matrix. Now utilizing variables such as level of urbanization and school size, we begin iterating just as before, removing one variable at a time based on P-values. The minimum P-value permitted is five percent. This was done for seven iterations until our final version, iteration 19 (Fig. 13). The initial version is shown as well to show how the model changed.



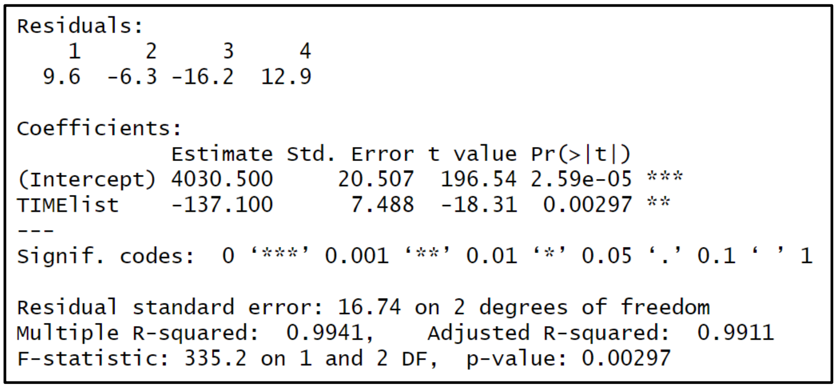
```
(Intercept)                       *
tableNONA.data.UNITID.train.
tableNONA.data.LOCALE.train.      **
tableNONA.data.CCSIZSET.train.    ***
tableNONA.data.UGDS.train.        ***
tableNONA.data.GRADS.train.
tableNONA.data.SAT_AVG_ALL.train. ***
tableNONA.data.AVGFACSAL.train.   ***
tableNONA.data.IRPS_MEN.train.    .
tableNONA.data.C100_4.train.      ***
tableNONA.data.RET_FT4.train.
tableNONA.data.PCTFLOAN.train.    ***
tableNONA.data.ADM_RATE_ALL.train. .
tableNONA.data.ACTCM75.train.
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8718 on 409 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.689,	Adjusted R-squared:  0.6791
F-statistic: 69.71 on 13 and 409 DF,  p-value: < 2.2e-16
```

```
(Intercept)                       ***
tableNONA.data.LOCALE.train.      **
tableNONA.data.CCSIZSET.train.    ***
tableNONA.data.UGDS.train.        ***
tableNONA.data.SAT_AVG_ALL.train. ***
tableNONA.data.AVGFACSAL.train.   ***
tableNONA.data.C100_4.train.      ***
tableNONA.data.PCTFLOAN.train.    ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8733 on 415 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6834,	Adjusted R-squared:  0.678
F-statistic: 127.9 on 7 and 415 DF,  p-value: < 2.2e-16
```

Figure 13: Iteration 12 and 19 of Tuition Model

From model 12 to 19, we see the slightest decrease in RSE and R^2 values, but a large increase in F-statistic values. The P-value is still extremely low, and there are no longer any independent variables with unsatisfactory P-values. This is our current final iteration of the model that calculates tuition costs.

We will pause on our tuition model, as we have to develop our regression models for each independent variable. Since each variable is checked on its own, we will use single linear regression models. It was previously explained that we initially looked at trends over the last thirteen years, but trends were not linear. This was adjusted to only the past four years, which provided a smaller range of data between values of previous years, and also showed linear trends. This allowed for extremely small windows for error, as was seen in Figure 8. The summary of the model that predicts future number of students (Fig. 14) indicates a very low error.



```
Residuals:
    1     2     3     4 
  9.6  -6.3 -16.2  12.9 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4030.500     20.507  196.54 2.59e-05 ***
TIMElist    -137.100      7.488  -18.31  0.00297 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.74 on 2 degrees of freedom
Multiple R-squared:  0.9941,	Adjusted R-squared:  0.9911 
F-statistic: 335.2 on 1 and 2 DF,  p-value: 0.00297
```

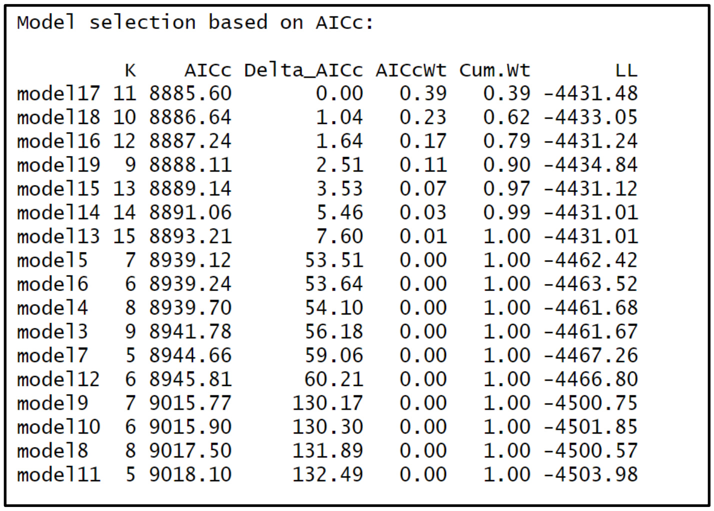
Figure 14: Prediction Model for Number of Undergraduates

We see an R^2 value of 0.99, which shows a very high level of model accuracy. This could already have been guessed from seeing the visual trend in Figure 8. The residual error is only around 16, which is very small considering there's many schools with thousands of students. This trend follows for the other prediction models of independent variables. With such high accuracy of these single linear regression models, it could be thought to simply use these models to predict tuition costs. However, the goal of this model is to calculate tuition costs with limited information given, as it will be a free open-source tool. Information such as tuition is limited behind paywall websites, and is assumed to be lacking if a tool such as this is needed. In the case where tuition cost can be accessed for the previous year, that can simply be used to estimate the following years cost. For the case where this tool is required, information such as tuition cost would not be accessible, and therefore cannot be used as an independent variable to estimate tuition.

Model Evaluation

For our tuition model, iteration 19 was thought to be the best-performing version. However, we want to gauge this through other means besides simply looking at error values. Therefore, we create an AIC table (Table 3) to compare AIC values of each model. These values depict how well each model fits to the data.

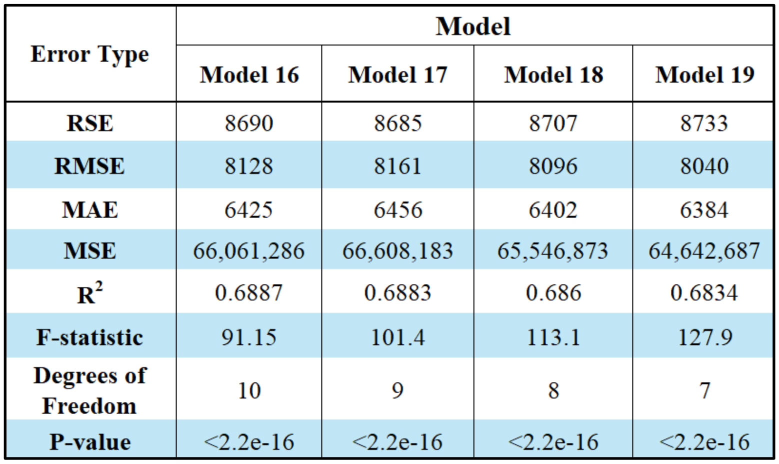
Table 3: AIC Table



```
Model selection based on AICc:

         K    AICc Delta_AICc AICcWt Cum.Wt       LL
model17 11 8885.60       0.00   0.39   0.39 -4431.48
model18 10 8886.64       1.04   0.23   0.62 -4433.05
model16 12 8887.24       1.64   0.17   0.79 -4431.24
model19  9 8888.11       2.51   0.11   0.90 -4434.84
model15 13 8889.14       3.53   0.07   0.97 -4431.12
model14 14 8891.06       5.46   0.03   0.99 -4431.01
model13 15 8893.21       7.60   0.01   1.00 -4431.01
model5   7 8939.12      53.51   0.00   1.00 -4462.42
model6   6 8939.24      53.64   0.00   1.00 -4463.52
model4   8 8939.70      54.10   0.00   1.00 -4461.68
model3   9 8941.78      56.18   0.00   1.00 -4461.67
model7   5 8944.66      59.06   0.00   1.00 -4467.26
model12  6 8945.81      60.21   0.00   1.00 -4466.80
model9   7 9015.77     130.17   0.00   1.00 -4500.75
model10  6 9015.90     130.30   0.00   1.00 -4501.85
model8   8 9017.50     131.89   0.00   1.00 -4500.57
model11  5 9018.10     132.49   0.00   1.00 -4503.98
```

From the AIC table, an interesting find was made. While model 19 was the final iteration, we see that iterations, 16, 17, and 18 rank higher in terms of AIC value. While these values are only marginally higher, they are higher nonetheless. These three models will therefore be evaluated and compared to model 19 and the testing data (Table 4).

Table 4: Tuition Model Comparison

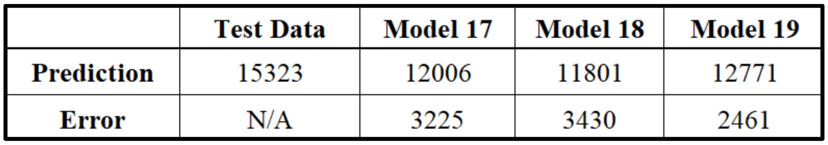


Error Type	Model			
	Model 16	Model 17	Model 18	Model 19
RSE	8690	8685	8707	8733
RMSE	8128	8161	8096	8040
MAE	6425	6456	6402	6384
MSE	66,061,286	66,608,183	65,546,873	64,642,687
R^2	0.6887	0.6883	0.686	0.6834
F-statistic	91.15	101.4	113.1	127.9
Degrees of Freedom	10	9	8	7
P-value	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16

From the tabulated error results, we see that model 16 has the largest standard, mean square, and root mean square error; as well as the lowest F-statistic. On the other hand, model 19 has the smallest mean square, root mean square, and mean absolute errors; as well as the highest F-statistic. Besides errors, the R^2 value for each model is relatively the same. When comparing the four models, it can be said that model 16 is the worse performing model. Models 17, 18, and 19 will be further analyzed and compared. This will be done with trials with schools from the testing set.

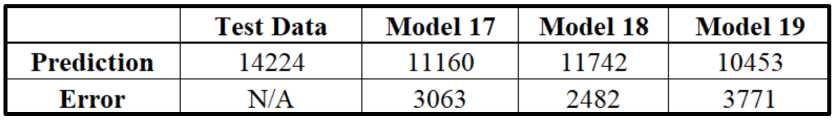
For some of our trials, we looked at schools such as *Alabama State University* and *Auburn University at Montgomery* (Table 5 and 6). For each trial, all independent variables of each model were predicted for the *2024* year. These variables were then used as inputs for each tuition model to predict tuition cost. These predictions were then compared to the values of the testing data; the errors were used to compare the models.

Table 5: Trial Results for *Alabama State University*



	Test Data	Model 17	Model 18	Model 19
Prediction	15323	12006	11801	12771
Error	N/A	3225	3430	2461

Table 6: Trial Results for *Auburn University at Montgomery*



	Test Data	Model 17	Model 18	Model 19
Prediction	14224	11160	11742	10453
Error	N/A	3063	2482	3771

With the first trial, we think that model 19 may be a clear outperformer. But with the second trial, we see that errors values vary between schools, and one model may outperform another for any given school. In terms of error, it’s difficult to choose a best model out of the three. This same overlap of errors is reflected in the similar AIC values and R^2 values. Based on all evaluation methods besides AIC tables, model 19 (Table 7) is the best performing model; and is therefore our final model.

Table 7: Tuition Model 19

Error Type	Model 19
RSE	8733
RMSE	8040
MAE	6384
MSE	64,642,687
AIC Rank	4th
R^2	0.6834
F-statistic	127.9
Degrees of Freedom	7
P-value	<2.2e-16

Model 19:

$$f_{19}(x) = -135x_1 - 1032x_2 - 0.56x_3 + 36.92x_4 + 1.5x_5 + 33130x_6 + 18060x_7 - 32130$$

Where, x_1 = LOCALE (Level of Urbanization),
x_2 = CCSIZSET (School Size),
x_3 = UGDS (Number of Undergraduates),
x_4 = SAT_AVG_ALL (Average SAT Scores),
x_5 = AVGFACSAL (Average Faculty Salary),
x_6 = C100_4 (Completion Rate),
x_7 = PCTFLOAN (Students Receiving Loans)

While the model is our "best" selection, there are some drawbacks to it: There's a low R^2 value, as we'd want a value over 0.85. The AIC rank was lower when compared to the other models, and the error values form the trials are relatively high. However, the F-statistic is high, with an extremely low associated P-value. However, when compared to the other models, the error values of version 19 are the most desirable. And considering the values of the tuition costs from the trials, the predictions of the models are somewhat accurate.

Future Work

While there wasn’t time to put these into practice, other modeling methods were researched to find any possibility they could be used with the tuition data. An intriguing method is principle component regression: When multiple linear regression has independent variables with high correlations and overall high variances with estimates, *pcr()* can be used to solve the issue. This model will find different linear combinations of the independent variables, and uses least squares to fit a linear regression model [4]. If implemented properly, this could solve our issue of high error during trials of model 19. Another regression method researched was ridge regression, which tackled a similar issue as pc regression, in that it helps where there’s high correlation between independent variables that causes high residual variance [5]. Using *glmnet(),* ridge regression minimizes mean squared error, which can help with model 19 in terms of the high errors in trials.

6. References

[1] Baum, S., & Ma, J. (2023). Trends in College Pricing and Student Aid 2023. The College Board.

[2] Gross, J., & Zerquera, D. (2022). "Gaps in Information Access and College Decision-Making Among Low-Income Students." Journal of Student Financial Aid, 51(1), 23-36.

[3] Heller, D. E. (2020). "Trends in the Affordability of Public Colleges and Universities." The Review of Higher Education, 43(2), 401-425.

[4] Bobbitt, Z. (2020, November 16). *Principal components Regression in R (Step-by-Step)*. Statology. https://www.statology.org/principal-components-regression-in-r/

[5] Bobbitt, Z. (2020, November 13). *Ridge regression in R (Step-by-Step)*. Statology. https://www.statology.org/ridge-regression-in-r/