EAS 345 Introduction to Data Science
Fall 2024

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

Group 3:

Jack Mandura – Data Scientist

Hunter Babel – Data Architect

Dylan Zelko – Project Lead

Emmitt Stores – Communications Liaison; Scribe

November 26, 2024

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, are the two excel files dealing with the data being used: The original data set marked as '*MERGED2022_23_PP.csv',* and the data dictionary that explains every variable listed within the data set, marked as '*CollegeScorecardDataDictionary.csv*'. Previous action and data source reports can be located here as well.

Data Cleaning

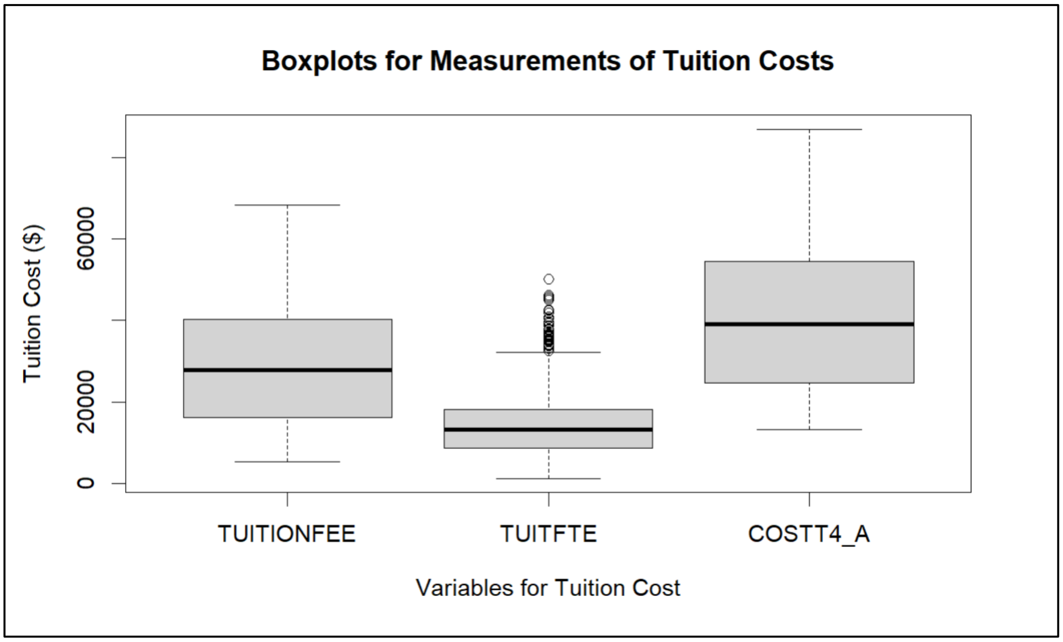
The data set was downloaded and brought into R for cleaning: This began with creating a data frame consisting of only the desired variables found through the data dictionary. This table was assessed for any columns consisting of too many null or N/A values, or even exclusively null values. These columns were deemed as unusable, and the respective variable was 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. 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.

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.

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. 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, Tuition Revenue, and Tuition Cost 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. 1). From the box plots, Tuition Cost has the most desirable distribution in the context of outliers and quartile ranges.

Figure 1: Comparison of Measurements for Tuition Cost



Box plots were generated for other variables as well. 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. An interesting use of histograms was looking at the distribution of school sizes (Fig. 2), which shows a somewhat linear trend between school size, and how many of those schools exist. All of the variables mentioned and analyzed are tabulated (Table 1) with their respective definitions:

Figure 2: Distribution of School Sizes

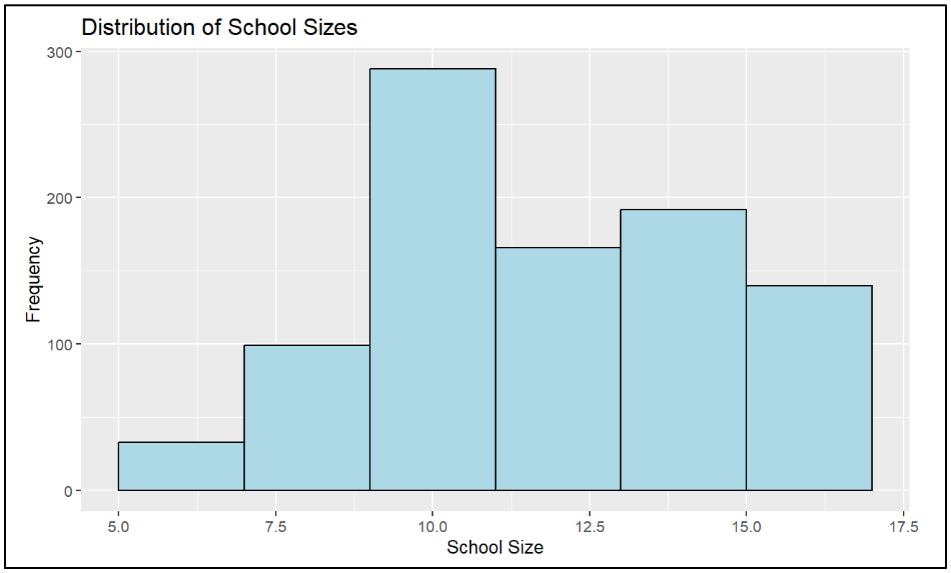


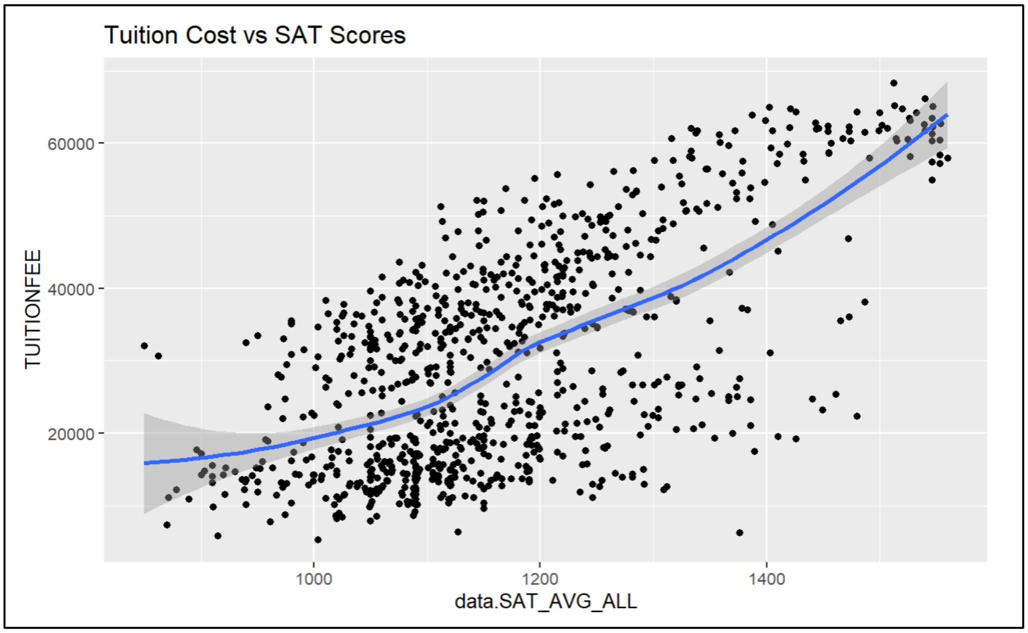
Table 1: Variable Names, Data Types, and Definitions

Variable Name	Data Type	Definiton
UNITID	Integer	University ID number
LOCALE	Integer	Level of Urbanization
CCSIZSET	Integer	School Size
SAT_AVG_ALL	Integer	Average SAT scores for all students
GRADS	Integer	Number of graduate students enrolled
UDGS	Integer	Number of undergraduate students enrolled
COSTT4_A	Integer	Average cost of attendance (Including books, tuition, room and boarding
AVGFACSAL	Integer	Average salaray of faculty members
ADM_RATE_ALL	Number	Admisison rate for all students
C100_4	Number	Completion rate for first-time, full-time students
RET_FT4	Number	Retention rate for first-time, full-time students
PCTFLOAN	Number	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	Average tuition revenue per student, for all students
TUITIONFEE_IN	Integer	Cost of tuition and fees for in-state students
TUITIONFEE_OUT	Integer	Cost of tuition and fees for out-of-state students
TUITIONFEE	Number	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 that other variables.

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. 3):

Figure 3: Scatter Plot of Tuition Cost Against SAT Scores



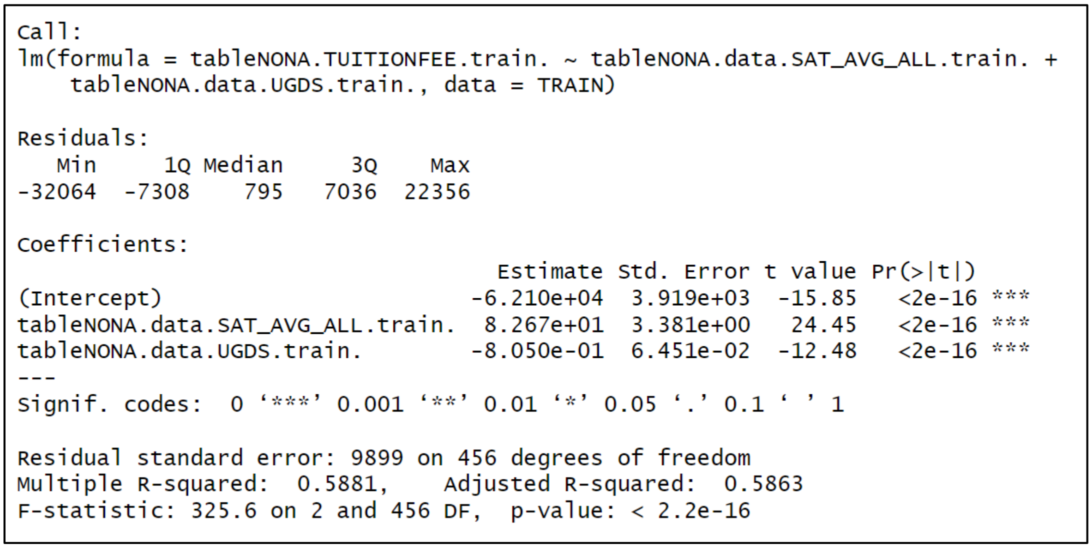
Similar trends were found between other dependent variables and tuition cost. Some trends that were previously thought to likely exist were shown to be random and nonexistent. An example of this being the relationship between tuition cost and school size, or even level of urbanization. Both were found to have little to no correlation when used with a scatter plot.

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 variables such as SAT scores and number of students to estimate 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. 4), 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.

Figure 4: Initial Modeling Results for Linear Regression



```
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
```

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.

Project Prognosis

After considering our current position in the project timeline, our project is likely to yield results. This determination was made based on our ability to stay on track, and the pace which our group has functioned. The coding aspect of our project has run smoothly, and with data acquired and cleaned, we expect to produce a functioning, accurate model. Looking towards the end of the semester, our Gantt Chart required revision. However, our project is still likely to be completed by the end of the semester. Minor adjustments had to be made to account for school breaks, and group member prior commitments. No issue has arisen which will affect the ability to complete this project on time. What remains to be accomplished is further developing and testing the linear model until factors such as the R2 and F-statistic values are within satisfactory ranges. Following this will be the completion of the final report and presentation.

GROUP NUMBER: 3

GROUP MEMBERS:

Name	Group Role(s)
Jack Mandura	Data Scientist
Hunter Babel	Data Architect
Dylan Zelko	Project Lead
Emmitt Stores	Communications Liaison; Scribe

College Scorecard – Department of Education
https://collegescorecard.ed.gov/data/ https://collegescorecard.ed.gov/data/data-documentation/
The College Scorecard is provided by the Department of Education, and provides complete coverage of all information about a college that’s a concern: Providing info about school sizes, student populations, admission rates, costs, etc. Information is provided for nearly seven thousand schools, including Title IV, public, and private institutions. The first link will take you to the page containing the download link: Within the folder downloaded, seek the file: *MERGED2022_23PP.csv* to download the data set. The second link will take you to the page containing the data dictionary, which defines all variables in the data set. This will be the first download link highlighted in green.
Data set which can be narrowed down by calling specific columns into R as desired.

DATE: 11/23/2024
GROUP NUMBER: 3
GROUP MEMBERS:

Name	Group Role(s)
Jack Mandura	Data Scientist
Hunter Babel	Data Architect
Dylan Zelko	Project Lead
Emmitt Stores	Communications Liaison; Scribe

Action Report

☐ Action Required
☒ No Action Required

Cause for Action

There is currently no cause for action within our group. Progress is moving along at good pace. Project is anticipated to produce results on-time.

Plan-of-Action

No Plan-of-Action needs to be made. Progress needs to keep being mad eat its current pace.

Timelines and Deadlines

No timeline or deadline needs to be set.

DATE: 11/2/2024
GROUP NUMBER: 3
GROUP MEMBERS:

Name	Group Role(s)
Jack Mandura	Data Scientist
Hunter Babel	Data Architect
Dylan Zelko	Project Lead
Emmitt Stores	Communications Liaison; Scribe

Action Report

☒ Action Required
□ No Action Required

Cause for Action

The changes which have been made at the time of the filing of this report were making corrections to our group proposal. The goal of this corrective action was to revise the report to develop a professional report with all of the desired information. The edited group proposal is attached to this corrective action report.

No other issues have presented themselves within our group.

Plan-of-Action

Description of proposed solution. What will you do about it? The solution to this corrective action was to reflect on the graded rubric we received and to make the appropriate changes. Specific changes made to the report include:

- **Changed report language to be more professional**
- **Added more details on the data cleaning process**
- **Included what models the team will be using**
- **Revised summary to be more specific to New York state schools**
- **Inserting a new paragraph in the introduction detailing our proposed tasks.**

Timelines and Deadlines

This action has been complete. It was completed on Saturday, November 2nd, 2024