Introduction

The idea that going to college and getting a college degree is the path to success gets harder and harder to swallow every year with rising tuition costs. We wanted to explore a dataset about college, so we attempted to find a more recent dataset that would be more relevant to our classmates, but to no avail. We eventually settled upon a dataset from the 1995 U.S. News report on American colleges and universities. The dataset originally had 1301 observations and 35 variables. As a university student, we are always interested about the statistics of our own school or having them compared with other universities. In this project, we try to address three main questions about the variables in the dataset.

1. Does a higher student/faculty ratio affect instructional costs?
2. Can we predict the number of applicants received?
3. Can we explain in-state tuition for public/private universities as a model of other variables?

Reading and Cleaning Data

When we first loaded the data set into R, the data set did not contain any of the column names. We had to correct this by referring to the documentation about the dataset. We then proceeded to clean the data by first dropping the columns of variables we did not want to investigate and then dropping the rows with missing values indicated by a '*'. Our final data had 480 observations and 23 variables which are listed in (see Appendix Figure 4).

Using `class()`, we discover that the numbers in some of the columns, e.g. X5 (number of applications received) were factors instead of numeric. Therefore, we used `transform()` to correct the data types to numeric, so we can do further analysis on the data. We also changed the school type from 1,2 to "Public" and "Private" respectively. As a last step, we used `sapply()` to make sure all the columns were in the proper data types.

Summary Statistics

Since we would like to know more about the difference between private school and public school in general, we have conducted some general summary statistics on a few variables. First, we made a boxplot for graduation rate over the school type. It was interesting to find the mean graduation rate for private school is 69.95%, while it is just 54.42% for public school (see Appendix Figure 1). We also made a box plot of the mean difference of instate tuition between private and public school (see Appendix Figure 2). Not a surprise, the mean in-state tuition for public school is $2392, while it is $12110 for private school—this is five times greater. We then

carried out an independent 2-group t-test to see if there is a true difference between the means. With a p-value of less than 0.05, we reject the null hypothesis that there is no true difference between the means of public and private in-state tuition fees.

<u>Does a higher student/faculty ratio affect instructional costs?</u>

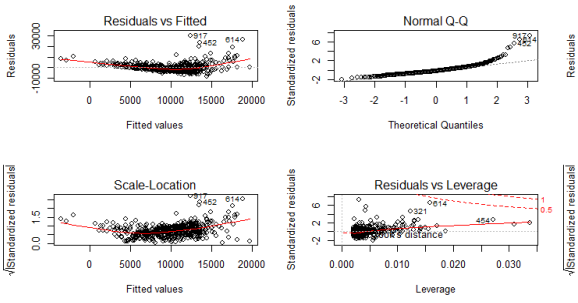
We used simple linear regression to answer this question. Our response variable is y="Instructional cost per student" and the explanatory variable is x="Student/faculty ratio". Using `lm()` in R, we got a linear equation $E(Y|X) = 22188 - 872 * x$. From the output below, we can see that the predictor is statistically significant with an extremely low p-value. When we looked at the diagnostic plots (1), we found an unusual trend in the residual and scale-location plot. This indicates that there is a non-constant variance and suggest that this linear model may not be valid. Using the transform function, we do a transformation by logging the response variable and got diagnostic plot (2). We can clearly see that there is no longer a trend in the residual plot the normality has also been improved from the QQ-plot. Multiple R-squared is the percentage of variance in the dependent variable that can be explained by the predictors. In output summary (1) and (2), we can see that the multiple R-squared has increased from 0.41 to 0.55, meaning that more of our data is explained by the transformed model.

```
Call:
lm(formula = Instructional_expenditure_per_student ~ Student_faculty_ratio,
    data = final.data)

Residuals:
    Min      1Q  Median      3Q     Max
-8138.7 -2501.2  -770.8  1569.7 30594.9

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          22188.72     700.87   31.66   <2e-16 ***
Student_faculty_ratio  -872.36      47.97  -18.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4159 on 478 degrees of freedom
Multiple R-squared:  0.409, Adjusted R-squared:  0.4077
F-statistic: 330.8 on 1 and 478 DF,  p-value: < 2.2e-16
```

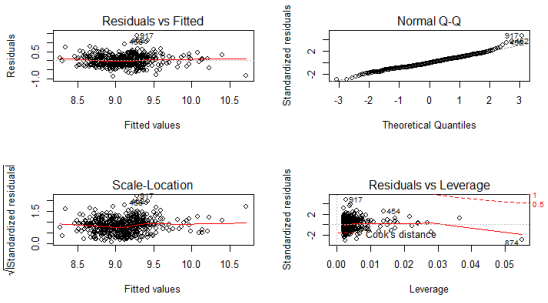
Summary Output (1)

```
Call:
lm(formula = log(Instructional_expenditure_per_student) ~ log(Student_faculty_ratio),
    data = final.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.84994 -0.20637 -0.01963  0.19500  1.38407

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             11.83513    0.11323  104.53   <2e-16 ***
log(Student_faculty_ratio) -1.05244    0.04324  -24.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2887 on 478 degrees of freedom
Multiple R-squared:  0.5534, Adjusted R-squared:  0.5525
F-statistic: 592.4 on 1 and 478 DF,  p-value: < 2.2e-16
```

Summary Output (2)



Diagnostic Plot (1)

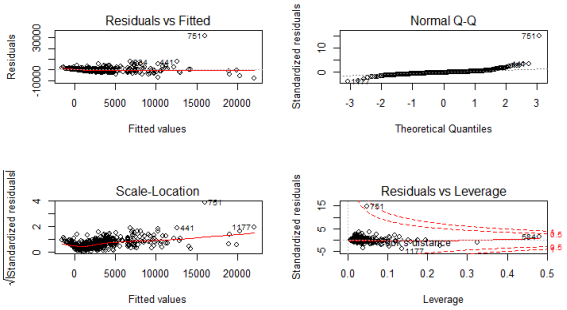


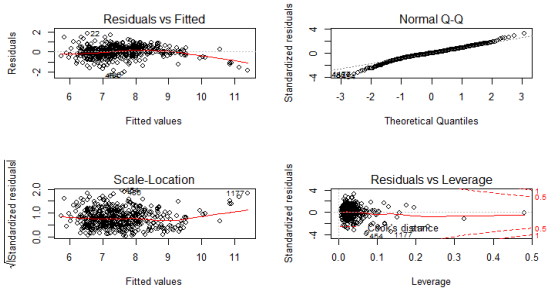
Diagnostic Plot (2)

From this simple linear regression, we can conclude that there is a negative relationship between instructional expenditure per student and student faculty ratio. In other words, the higher the student/faculty ratio, the lower the instructional expenditure per student. We believe our results make sense because when the student/faculty ratio is higher, more students will be sharing the instructional cost for the professor; therefore, the instructional expenditure per student will be lower.

Can we predict the number of applicants received?

As a second step of our project, we did a multiple linear regression to find out what variables can help predict the number of applicants received. To check the validity of the model, we repeated the steps that have been done for simple linear regression. Again, we discovered a fan shape in the residual plot, which means that a non-constant variance problem exists. We then attempted to log the response variable to see if it gets better. As shown in the below diagnostic plots, we can see an improvement.





Diagnostic plots before transformation

Diagnostic plots after transformation

Then, we got an output summary as shown below.

```
Call:
lm(formula = log(Applications_received) ~ . - Applicants_accepted -
    New_students_enrolled - Room_costs - Board_costs - Total_tuition,
    data = no.id.data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.0979 -0.2919  0.0076  0.3639  1.8681

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             3.931e+00  2.675e-01  14.694  < 2e-16 ***
School_typePublic                       5.186e-01  1.542e-01   3.363 0.000836 ***
Fulltime_undergrads                     1.178e-04  9.051e-06  13.019  < 2e-16 ***
Parttime_undergrads                     2.387e-06  2.119e-05   0.113 0.910319
Instate_tuition                        -1.337e-05  3.028e-05  -0.442 0.659046
Outstate_tuition                        5.365e-05  3.031e-05   1.770 0.077404 .
Room_board_costs                        1.245e-04  3.467e-05   3.590 0.000366 ***
Additional_fees                         1.756e-04  8.140e-05   2.157 0.031524 *
Estimated_book_costs                    2.769e-04  1.712e-04   1.617 0.106468
Estimated_personal_spending            -2.149e-06  4.375e-05  -0.049 0.960842
Percentage_faculty_Phd                  8.782e-03  3.371e-03   2.605 0.009473 **
Percentage_faculty_terminal_degree     -3.931e-03  3.662e-03  -1.073 0.283610
Student_faculty_ratio                   4.197e-02  9.853e-03   4.260 2.48e-05 ***
Percentage_alumni_donate               -4.782e-03  2.790e-03  -1.714 0.087175 .
Instructional_expenditure_per_student   3.037e-05  8.452e-06   3.593 0.000362 ***
Grad_rate                               1.043e-02  2.031e-03   5.135 4.16e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5728 on 464 degrees of freedom
Multiple R-squared:  0.7083, Adjusted R-squared:  0.6989
F-statistic: 75.11 on 15 and 464 DF,  p-value: < 2.2e-16
```

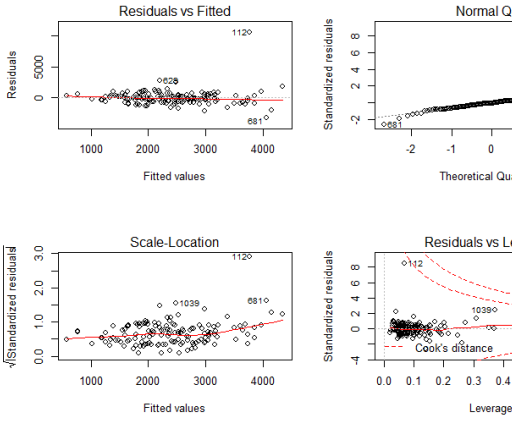
We found out that there are quite a number of insignificant variables. Since this indicates an incomplete model, using `leap()` with backwards stepwise strategy and BIC criteria, we removed variables that were useless to the model (see Appendix Figure 3.) After removing all irrelevant variables, we are left with all significant variables and got the multiple linear equation:

$$E(Y|X) = 5.48 + 0.0004 * X_8 - 0.00025 * X_9 + 0.0003 * X_{10} + 0.003 * X_{23}$$

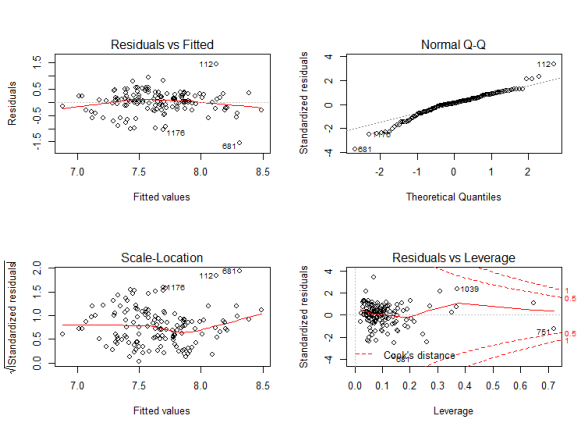
In conclusion, the most significant variable is the number of full time graduates. We may suggest that a large number of applications received are associated with the number of full time graduates in a university. A possible explanation for this is that students may believe that they have a better chance of getting into a university with a higher student population because this suggests that they accept more applicants. Furthermore, the number of applicants received also has a relationship with the amount of in-state tuition, the graduation rate and the number of part time undergraduates.

<u>Can we explain instate tuition for public/private universities as a model of other variables?</u>

The final question we sought to answer was if instate tuition for both public and private universities could be explainable and whether or not the same variables could be used to model instate tuition. As seen in Figure 2, we can observe that instate tuition for private schools is much higher than public schools. Following the same steps as before, we found that logging the response variable (instate tuition) results in a better residual plot and suggests that our model for public instate tuition is valid (see Diagnostic Plots (1), (2)).



Diagnostic Plot (1)



Diagnostic Plot (2)

The output shows that there are numerous insignificant variables, which means that is an incomplete model (see Output Summary (1)) . Using `stepAIC()` with backwards selection, we are able to reduce the model down to 9 variables and get the resulting equation (see Output Summary (2)):

$$\log(X_{10}) = 6.908 - 6.425 * 10^{-5} * X_5 + 1.774 * 10^{-4} * X_6 - 2.024 * 10^{-4} * X_7 + 2.389 * 10^{-4} * X_{12} - 3.004 * 10^{-4} * X_{15} - 5.805 * 10^{-3} * X_{19} + 1.225 * 10^{-2} * X_{21} + 4.372 * 10^{-5} * X_{22}$$

```
Call:
lm(formula = log(Instate_tuition) ~ Applications_received + Applicants_accepted +
    New_students_enrolled + Fulltime_undergrads + Parttime_undergrads +
    Room_board_costs + Additional_fees + Estimated_book_costs +
    Estimated_personal_spending + Percentage_faculty_Phd + Percentage_faculty_terminal_degree +
    Student_faculty_ratio + Percentage_alumni_donate + Instructional_expenditure_per_student,
    data = public.data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54260 -0.20583  0.04528  0.23547  1.44820

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            6.856e+00  4.957e-01  13.831  < 2e-16 ***
Applications_received                 -6.031e-05  2.645e-05  -2.280 0.024298 *
Applicants_accepted                    1.655e-04  5.364e-05   3.085 0.002508 **
New_students_enrolled                 -1.406e-04  1.232e-04  -1.141 0.255951
Fulltime_undergrads                   -8.623e-06  2.103e-05  -0.410 0.682530
Parttime_undergrads                   -1.218e-05  1.872e-05  -0.651 0.516492
Room_board_costs                       2.477e-04  5.615e-05   4.411  2.2e-05 ***
Additional_fees                       -3.239e-04  9.218e-05  -3.514 0.000617 ***
Estimated_book_costs                  -1.900e-04  3.141e-04  -0.605 0.546346
Estimated_personal_spending           -2.739e-05  6.587e-05  -0.416 0.678218
Percentage_faculty_Phd                 5.734e-03  5.643e-03   1.016 0.311498
Percentage_faculty_terminal_degree    -9.234e-03  5.594e-03  -1.651 0.101362
Student_faculty_ratio                  2.479e-03  1.575e-02   0.157 0.875159
Percentage_alumni_donate               1.218e-02  5.323e-03   2.287 0.023865 *
Instructional_expenditure_per_student  4.396e-05  2.676e-05   1.643 0.102944
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4421 on 124 degrees of freedom
Multiple R-squared:  0.3423, Adjusted R-squared:  0.268
F-statistic:  4.61 on 14 and 124 DF,  p-value: 1.115e-06
```

Output Summary (1)

```
Call:
lm(formula = log(Instate_tuition) ~ Applications_received + Applicants_accepted +
    New_students_enrolled + Room_board_costs + Additional_fees +
    Percentage_faculty_terminal_degree + Percentage_alumni_donate +
    Instructional_expenditure_per_student, data = public.data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.52171 -0.21454  0.03738  0.26077  1.51289

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            6.908e+00  2.979e-01  23.185  < 2e-16 ***
Applications_received                 -6.425e-05  2.552e-05  -2.518 0.013024 *
Applicants_accepted                    1.774e-04  5.217e-05   3.400 0.000894 ***
New_students_enrolled                 -2.024e-04  6.811e-05  -2.972 0.003524 **
Room_board_costs                       2.389e-04  5.332e-05   4.480 1.62e-05 ***
Additional_fees                       -3.004e-04  8.768e-05  -3.426 0.000821 ***
Percentage_faculty_terminal_degree    -5.805e-03  3.768e-03  -1.541 0.125809
Percentage_alumni_donate               1.225e-02  5.122e-03   2.391 0.018243 *
Instructional_expenditure_per_student  4.372e-05  1.811e-05   2.414 0.017168 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4367 on 130 degrees of freedom
Multiple R-squared:  0.327, Adjusted R-squared:  0.2856
F-statistic: 7.896 on 8 and 130 DF,  p-value: 1.274e-08
```

Output Summary (2)

For the private instate tuition model, we did not need to log-transform any of the variables (see Diagnostic Plot (1)). The resulting model after using `stepAIC()` is (see Output Summary (1)):

$$X_{10} = 1925 + 0.3149 * X_5 + 1.0912 * X_6 - 4.7603 * X_7 - 0.4696 * X_9 + 1.0592 * X_{12} - 0.8954 * X_{15} - 0.6075 * X_{17} + 31.6949 * X_{18} - 26.1317 * X_{19} - 76.6919 * X_{20} + 45.9813 * X_{21} + 0.1782 * X_{22}$$

Both models share a fair amount of predictors, but they also have the same significant predictor—Room and board costs. This shows that the majority of an instate student's tuition fees goes towards their lodging and food. There are also more explanatory variables for the instate tuition model for private universities, which shows that there is more variability. One explanation of this could be that public universities are usually on a fixed budget compared to private universities which would have more freedom to where they want to allocate their funds.

Conclusion

Most of our expectations about our investigative questions were confirmed. However, we did not expect the most significant predictor for the number of applications received to be the number of full-time undergraduates a university had. Personally, we felt that there should be other more important characteristics about a university that should factor into an applicant's decision to apply.

From this project, we gained a better grasp of R coding, more specifically manipulating data frames and using ggplot2. We also learned how to apply statistical techniques covered in other courses.

Appendix

Figure 1. *Box plot of Graduation Rate vs School Type*

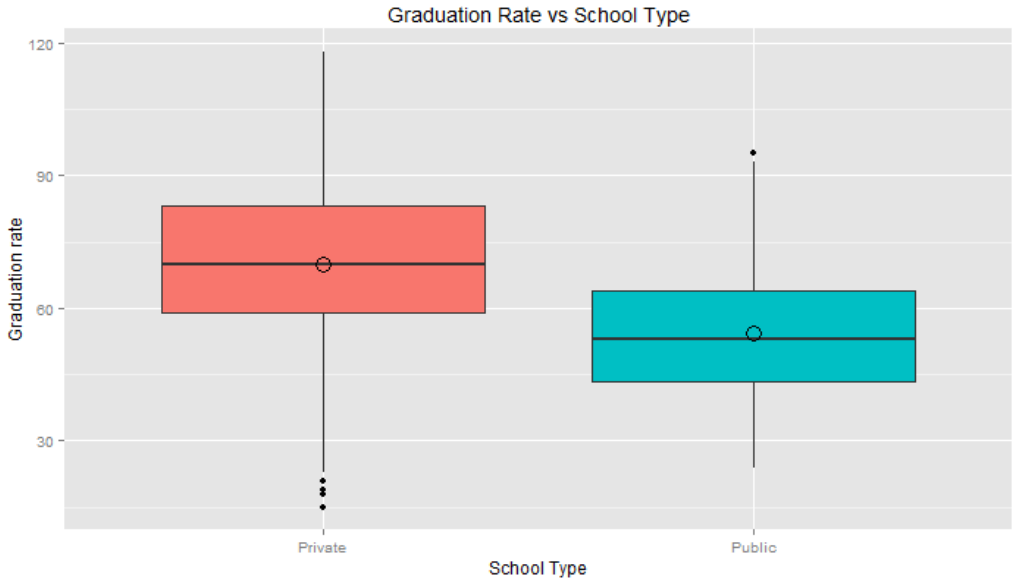


Figure 2. *In-state Tuition vs School Type*

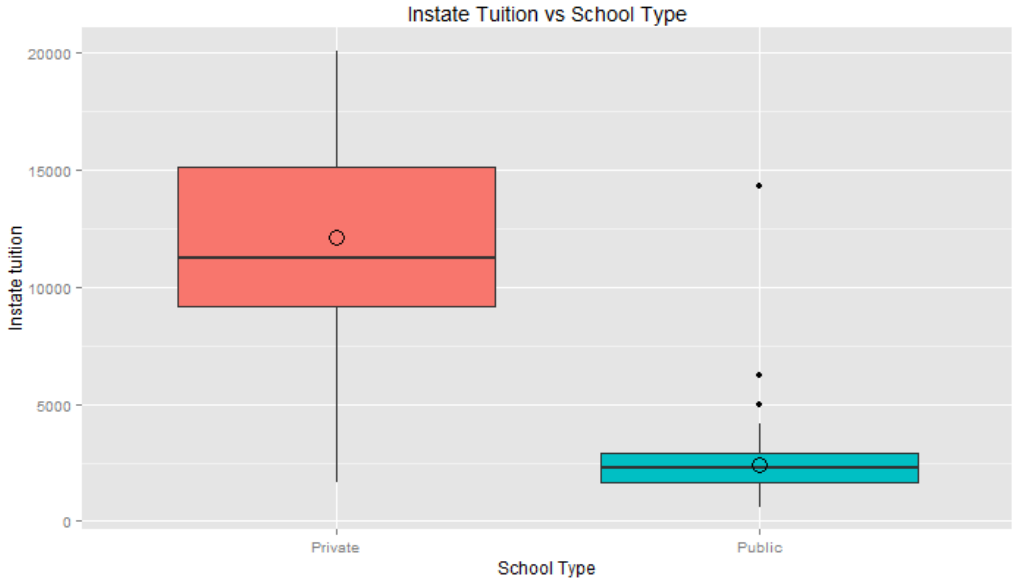
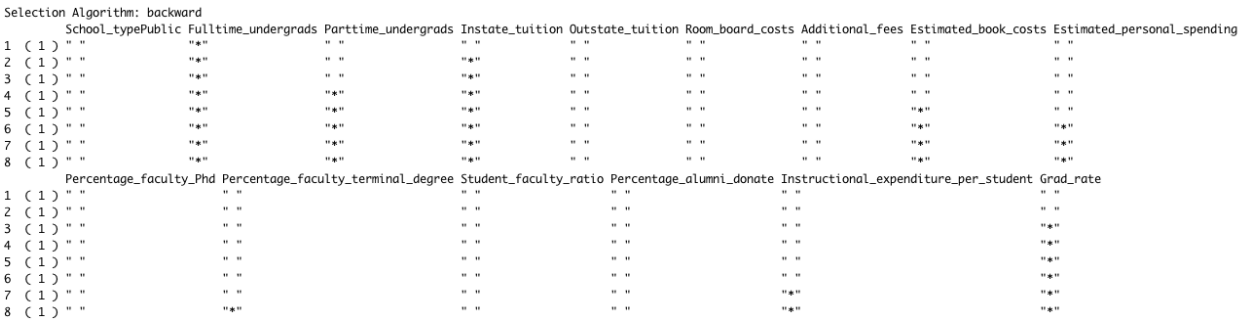


Figure 3. *Leap function*



```
Selection Algorithm: backward
         School_typePublic Fulltime_undergrads Parttime_undergrads Instate_tuition Outstate_tuition Room_board_costs Additional_fees Estimated_book_costs Estimated_personal_spending
1  ( 1 ) " "               "*"                " "                 " "             " "              " "             " "             " "                  " "
2  ( 1 ) " "               "*"                " "                 "*"             " "              " "             " "             " "                  " "
3  ( 1 ) " "               "*"                " "                 "*"             " "              " "             " "             " "                  " "
4  ( 1 ) " "               "*"                "*"                 "*"             " "              " "             " "             " "                  " "
5  ( 1 ) " "               "*"                "*"                 "*"             " "              " "             " "             "*"                  " "
6  ( 1 ) " "               "*"                "*"                 "*"             " "              " "             " "             "*"                  "*"
7  ( 1 ) " "               "*"                "*"                 "*"             " "              " "             " "             "*"                  "*"
8  ( 1 ) " "               "*"                "*"                 "*"             " "              " "             " "             "*"                  "*"
         Percentage_faculty_Phd Percentage_faculty_terminal_degree Student_faculty_ratio Percentage_alumni_donate Instructional_expenditure_per_student Grad_rate
1  ( 1 ) " "                    " "                                " "                  " "                     " "                                  " "
2  ( 1 ) " "                    " "                                " "                  " "                     " "                                  " "
3  ( 1 ) " "                    " "                                " "                  " "                     " "                                  "*"
4  ( 1 ) " "                    " "                                " "                  " "                     " "                                  "*"
5  ( 1 ) " "                    " "                                " "                  " "                     " "                                  "*"
6  ( 1 ) " "                    " "                                " "                  " "                     " "                                  "*"
7  ( 1 ) " "                    " "                                " "                  " "                     "*"                                  "*"
8  ( 1 ) " "                    "*"                                " "                  " "                     "*"                                  "*"
```

Figure 4. *Variable names and descriptions*

NOTE: Simplified variable names to X1…Xn instead of the variable name in R

X1	Federal ID number
X2	College Name
X3	State
X4	School Type (Public/Private)
X5	Number of applications received
X6	Number of applications accepted
X7	Number of new students enrolled
X8	Number of full-time undergraduates
X9	Number of part-time undergraduates
X10	Amount of in-state tuition per year
X11	Amount of out-state tuition per year
X12	Room and board costs per year
X13	Room costs per year
X14	Board costs per year
X15	Additional fees per year
X16	Estimated book costs
X17	Estimated personal spending
X18	Percentage of faculty with PhD’s
X19	Percentage of faculty with terminal degree
X20	Student/Faculty ratio
X21	Percentage of alumni who donate
X22	Instructional expenditure per student
X23	Graduation rate