College Ranking Predictions

Report

Introduction

Research Background

Every year, millions of students apply to colleges across the United States, and many of them use college rankings lists from sources such as *US News and World Report*, Forbes.com, and Niche.com to to help them decide where to apply and where to go. In recent years, these lists have been heavily criticized for focusing on "exclusivity and resources, rather than accessibility and economic mobility" (1).The system can also be easily manipulated by a university if that university prioritize certain metrics to raise their rankings, as seen in Northeastern University meteoric rise from #163 to #49 on the US News and World Report list in only 17 years. The president of Northeastern University even explicitly stated that it was a top priority of the university to raise its ranking (2).

It is important to know where these rankings come from and what they actually measure. In this project, we will explore how influential different metrics are in determining a college's ranking and in determining a higher ranking.

For the sake of clarity, when we say a "low rank," we are referring to schools with a lower numerical rank, such as #1 and #2. When we say a "high rank," we a referring to schools with a high numerical rank, such as #499 and #500.

Data

For our analysis, we will be joining the following two data set sets:

Data Set #1: Niche

- The first data set comes from Niche's "2023 Best Colleges in America" list. Niche aggregates data from a variety of sources, including the US Department of Education and reviews from students and alumni, to build their list of college rankings. The rankings list is updated monthly. However, Niche only receives data from the US Department of Education on an annual basis. The Niche data was scraped by Maia on October 17-19 2022. There are 500 observations, representing the top 500 schools in the United States. Each observation has two variables: `college` (institution name) and `rank`.

Data Set #2: US Department of Education

- The second data set comes from the US Department of Education's College Scorecard, which is an exhaustive summary of characteristics and statistics for all colleges and universities in the United States. The College Scorecard is updated by the Education Department as it collects new data, but most of the data comes from the 2020-2021 school year. Data used in the scorecard comes comes from data reported by the institutions, data on federal financial aid, data from taxes, and data from other federal agencies. There were 2,989 variables in the original data set, but we narrowed to 31 variables we thought could have an influence on rank. There are 6681 observations in the data set, representing all of the colleges and universities in the United States.

Variable Summary

Here is a summary of the variables we will be using in our analysis. We selected every one we thought could impact rank, but left some out, such as test score breakdowns, to avoid redundancy.

- `college`: Institution name
- `rank`: Rank of school on Niche list
- `REGION:` US geographic region
- `ACCREDAGENCY`: Accreditor for Institution
- `CONTROL`: Public, Private nonprofit, or Private for-profit
- `CCBASIC`: Carnegie Classification (basic)
- `ADM_RATE`: Admission rate
- `UGDS:` enrollment of undergraduate certificate/degree-seeking students
- `UGDS_WHITE`, `UGDS_BLACK`, `UGDS_HISP`, `UGDS_ASIAN`, `UGDS_AIAN`, `UGDS_NHPI`, `UGDS_2MOR`, and `UGDS_UNKN`: enrollment of undergraduate students of each racial/ethnic group.
- `NPT4_PUB`: Average net price for Title IV institutions (public institutions)
- `NPT4_PRIV`: Average net price for Title IV institutions (private institutions)
- `COSTT4_A`: Average cost of attendance (academic year institutions)
- `COSTT4_P`: Average cost of attendance (program-year institutions)
- `AVGFACSAL`: Average faculty salary
- `PCTPELL`: Percentage of undergraduates who receive a Pell Grant

- `C150_4`: Completion rate for first-time, full-time students at four-year institutions
- `AGE_ENTRY`: Average age of entry
- `FEMALE`: Share of female students
- `MARRIED`: Share of married students
- `FIRST_GEN`: Share of first-generation students
- `FAMINC`: Average family income
- `MD_FAMINC`: Median family income
- `ENDOWBEGIN`: Value of school's endowment at the beginning of the fiscal year
- `SAT_AVG`: Average SAT equivalent score of students admitted
- `ACTCMMID`: Midpoint of the ACT cumulative score

Data Preparation

1. To get the data, we scraped from Niche.com and downloaded data from the US Department of Education. The steps were done in an R script titled `niche-scrape.R`
2. We had to adjust for institution name discrepancies between the two data sets before joining the two. `University of South Florida - Sarasota-Manatee` and `University of South Florida - St. Petersburg` were removed because they did not exist in both data sets.

3. Some of the categorical variables in the US Department of Education dataset (`REGION`, `CONTROL`, and `CCBASIC`) used numbers to represent the different levels, so we looked at the data dictionary and replaced each number with the words that it represents.

4. All of the numerical variables are on different scales, so we created a scaled version of the data, with mean 0 and standard deviation 1. Here is the first row of the data set below.

college	rank	REGION	CONTROL	CCBASIC	[illegible]
Massachusetts Institute of Technology	1	New England	Private, Non-profit	Doctoral Universities: Very High Research Activity	New England Commission on Higher Education - - - [illegible] NA [illegible] NA [illegible] NA [illegible]

Exploratory Data Analysis

Means of Selected Numerical Variables by Rank Group

Interval	Mean Admission Rate	Mean SAT Average	Mean ACT Median	Mean % White Students	Mean % Asian Students	Mean Cost of Attendance
1 to 100	0.2767340	1422.703	32.05495	0.4865770	0.1500930	60691.49
101 to 200	0.6181010	1266.932	27.45455	0.6271786	0.0698429	46371.74
201 to 300	0.6818970	1215.864	26.02469	0.6142470	0.0729400	43882.71
301 to 400	0.7157737	1160.091	24.14474	0.5629051	0.0548111	40799.30
401 to 500	0.7430135	1145.662	23.86076	0.6184102	0.0436867	36583.50

As the rank group gets higher, the mean admission rate increases and the mean SAT Average, ACT Median, and cost of attendance decreases. 1-100 ranked schools have considerably fewer White students and considerably more Asian students than schools ranked above 100

We have included our exploratory analysis of categorical variables in our appendix.

Research Question and Hypothesis

Question: Which characteristics of a university are most associated with rankings on the Niche College Ranking list? Of these characteristics, what is the relationship between high and low rank?

Hypothesis: We hypothesize that SAT/ACT scores, acceptance rate, and family income will have the strongest association with rank because since Niche's audience is in large part students applying to college, we believe that they prioritize variables important in the college admissions process. Of these variables, we predict that SAT/ACT score will have strong negative relationship, acceptance rate will have a strong positive relationship, and family income will have a strong negative relationship with rank.

Methodology

We have split the first part of our analysis into two approaches. The first approach consists of looking at he linear relationship between the numerical explanatory variables and college rank using R-squared variables. The second approach consists of building a stepwise regression model between many explanatory variables and college rank. As the variables that appear in the final model will be most important for determining rank, we will use the model results to corroborate our results from the first approach. As we cannot find an R-squared value or other numerical metric to measure a relationship involving a categorical variable, we decided to simply use the stepwise regression model to determine if there is a strong association between those variables and rank.

In the second part of our analysis, we will combine the results of the two approaches and characterize the relationship between rank and the variables with the strongest association with it.

Approach #1: Individual Numerical Variable Analysis

First, we will create a linear regression models between each individual explanatory variable and college rank. Then, we will calculate the R-squared value for each respective model, rank the values from highest to lowest, and select the variables with the highest R-squared values.

Approach #2: Stepwise Regression Modeling

A stepwise regression model can manage large amounts of potential predictor variables and fine-tune the model to choose the best predictor variables from the available options. In our case, we have more than 25 variables to be examined and thus it is crucial to have a automated workflow for model selections.

In our research, we will use both forward and backward selections in the stepwise regression model by utilizing MASS package. We will evaluate the performance of each iteration of the model based on Akaike information criterion (AIC). AIC is used to compare different possible models and determine which one is the best fit for the data in statistic practice.

There are two main steps in this approach.

1. Create a correlation matrix to check correlation coefficients between variables so as to not use similar variables in our model. If two variables had an absolute value of r greater than 0.8, meaning they were too similar in how they factored into rankings, we only picked one of them to put into the model.
2. Compute the stepwise regression model using MASS package and mainly `stepAIC()` functions for model selections based on AIC. For the initial setting of the linear regression model, we will import all the valid variables into the model to predict the rank variable.

In the end, this will give us the best final model with much fewer variables. Those variables are the most influential factors to the rank of the college.

Final Variable Analysis

We will examine the final variables selected by both approaches and analyze their relationships with college rank by:

1. Interpreting the R-squared values and graphs to characterize the linear association for each variable and rank.
2. Calculate the linear regression slopes between each of the explanatory variables (scaled and non-scaled) and college rank. Then we will use the scaled slopes to determine which explanatory variable has the greatest influence on college on a school having a higher rank. We will interpret the relationships using the non-scaled slopes.

Results

Approach #1: Individual Numerical Variable Analysis

Table of R-squared Values

The table below gives the R-squared values from the linear regression models between each individual explanatory variable in our data set and colleges rank, arranged in descending order.

variable	r_squared
SAT_AVG	0.6291468
ACTCMMID	0.6062324
C150_4	0.5137627
AVGFACSAL	0.4692148
ADM_RATE	0.4048157
PCTPELL	0.2476545

`COSTT4_P` (Average cost of attendance for program-year institutions) has been removed because there are only two observations.

Average SAT (`SAT_AVG`), median ACT (`ACTCMMID`), graduation rate (`C150_4`), average faculty salary (`AVGFACSAL`), and admission rate (`ADM_RATE`), are the five variables with the strongest correlation to rank, based on their R-Squared values; therefore, they are the variables we will be examining later in our analysis. We chose five as a cutoff because there is a substantial difference between the R-squared value of these five and the next variable (`PCTPELL)`.

Approach #2: Stepwise Regression Model

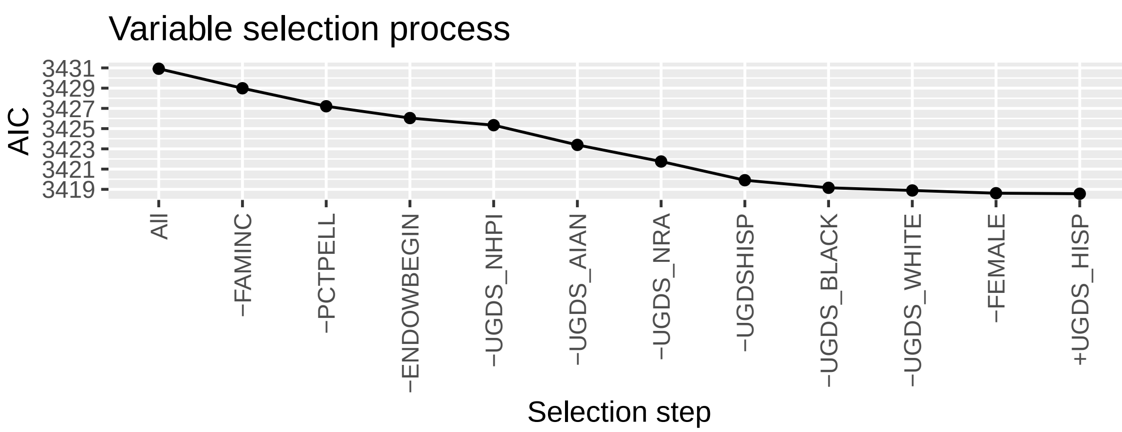
Remove Highly Correlated Variables

Below is the correlation table that displays the variable pairs with correlation values greater than or equal to 0.8.

Variable Pairs with r > 0.8	Correlation Coefficients
C150_4, SAT_AVG	0.8389
C150_4, ACTCMMID	0.8494
AGE_ENTRY, MARRIED	0.9059
FAMINC, MD_FAMINC	0.9538
SAT_AVG, ACTCMMID	0.9756

We will drop the variables `C150_4`, `MD_FAMINC`, `ACTCMMID`, `MARRIED` and preserve `SAT_AVG` and `FAMINC` to represent all other variables.

Compute Stepwise Regression



The plot above visually displays the decreasing AIC value as variables get taken out and potentially re-added. Each step represents a new iteration of the model after a certain variable is taken out. For example, the first step represents the first iteration with all the variables. Then, the variable `FAMINC` was removed, leading to a decrease in AIC. Then, one-by-one, it removes `PCTPELL`, `ENDOWBEGIN`, each level of undergraduate ethnicities, and female student population to continue making the best model for our data. Finally, the re-addition of `UGDS_HISP` led to the final iteration our of model with just 16 variables. We can tell this is the best model because its AIC value is the lowest.

Model Results

```
Final Model:
rank ~ REGION + CONTROL + CCBASIC + ACCREDAGENCY + ADM_RATE +
    UGDS + UGDS_ASIAN + UGDS_2MOR + UGDS_UNKN + COSTT4_A + AVGFACSAL +
    AGE_ENTRY + SAT_AVG + FIRST_GEN + UGDS_HISP
```

We can further tell that this is our best model because it has an r-squared coefficient of 0.7740974. The variables here can then be considered to be the variables that affect a college's ranking the most. Region (`REGION`) and type of school (`CONTROL`, `CCBASIC`) influence its ranking. There are certain qualities of its undergraduate population that are more statistically significant than others and were thus included in this model: overall undergraduate population, number of Asians, Hispanics, mixed students, and first-generation students, and average age when the undergraduated enrolled. Average SAT score was also significant enough to be include into the model, and `ACCREDAGENCY` and cost of attendance were also the final two variables considered significant enough for our best model.

Final Variable Analysis

Categorical Variable Analysis

All four categorical variables appeared in the final model, and therefore we can assume that they have a significant association with rank.

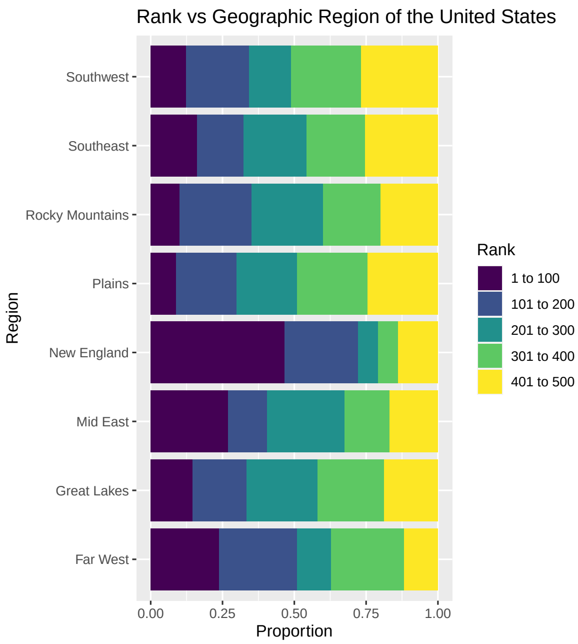
Note that for these graphs we dropped `NAs` and removed levels that only had one observation.

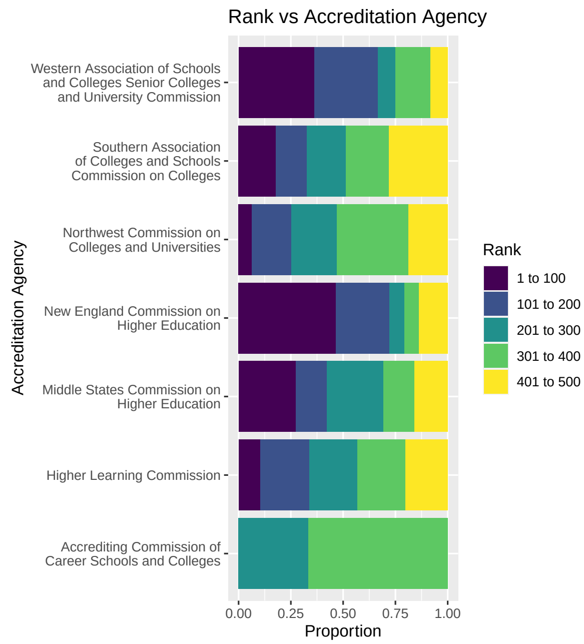
The Geographic Region graph shows that New England has the highest proportion of top-100 schools, while the Plains has the lowest. Apart from the `New England` and `NA` bars, the differences in proportions of rank groups do not vary dramatically between bars. It is possible that the strength of the correlation between rank and region is driven in large part by the association New England has with schools with the lowest 100 ranks.

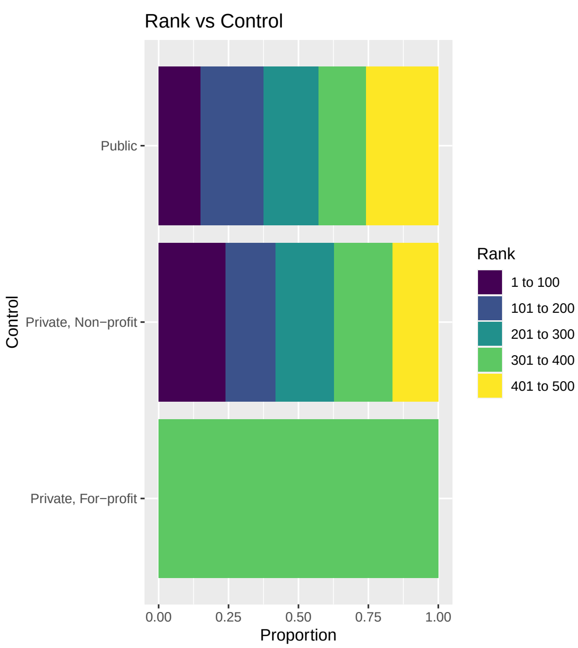
There does not appear to be a obvious pattern in the accreditation agency graph, which could be because some of the agencies corresponded to very few schools in the top 500. Additionally, as accreditation agency is often based on location, it reflects results similar to the region graph.

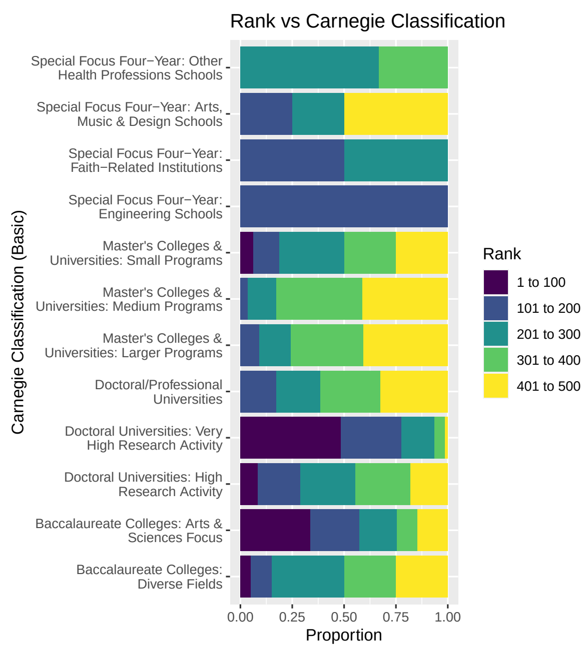
There are only 4 `Private, For-profit` schools in the top 500, and all of them are ranked between 301 and 400. The proportions of ranks between `Private, Non-profit` and `Public` are similar, although the first appears to have a larger proportion of `1-100` schools, and the latter a higher proportion of `401-500` schools.

There appear to be the greatest differences between bars of proportions of rank groups in the Carnegie Classification group, suggesting that this has the strongest association with rank. It appears that the lower the rank, the higher proportion of schools in `Doctoral Universities: Very High Research Activity` and `Baccalaureate Colleges: Arts & Sciences Focus`. However, the opposite appeared to be true for all other classifications with 3 or more rank


Rank vs Geographic Region of the United States
Region
Southwest
Southeast
Rocky Mountains
Plains
New England
Mid East
Great Lakes
Far West
0.00
0.25
0.50
0.75
1.00
Proportion
Rank
1 to 100
101 to 200
201 to 300
301 to 400
401 to 500


Rank vs Accreditation Agency
Accreditation Agency
Western Association of Schools and Colleges Senior Colleges and University Commission
Southern Association of Colleges and Schools Commission on Colleges
Northwest Commission on Colleges and Universities
New England Commission on Higher Education
Middle States Commission on Higher Education
Higher Learning Commission
Accrediting Commission of Career Schools and Colleges
0.00
0.25
0.50
0.75
1.00
Proportion
Rank
1 to 100
101 to 200
201 to 300
301 to 400
401 to 500


Rank vs Control
Control
Public
Private, Non−profit
Private, For−profit
0.00
0.25
0.50
0.75
1.00
Proportion
Rank
1 to 100
101 to 200
201 to 300
301 to 400
401 to 500


Rank vs Carnegie Classification
Carnegie Classification (Basic)
Special Focus Four−Year: Other Health Professions Schools
Special Focus Four−Year: Arts, Music & Design Schools
Special Focus Four−Year: Faith−Related Institutions
Special Focus Four−Year: Engineering Schools
Master's Colleges & Universities: Small Programs
Master's Colleges & Universities: Medium Programs
Master's Colleges & Universities: Larger Programs
Doctoral/Professional Universities
Doctoral Universities: Very High Research Activity
Doctoral Universities: High Research Activity
Baccalaureate Colleges: Arts & Sciences Focus
Baccalaureate Colleges: Diverse Fields
0.00
0.25
0.50
0.75
1.00
Proportion
Rank
1 to 100
101 to 200
201 to 300
301 to 400
401 to 500

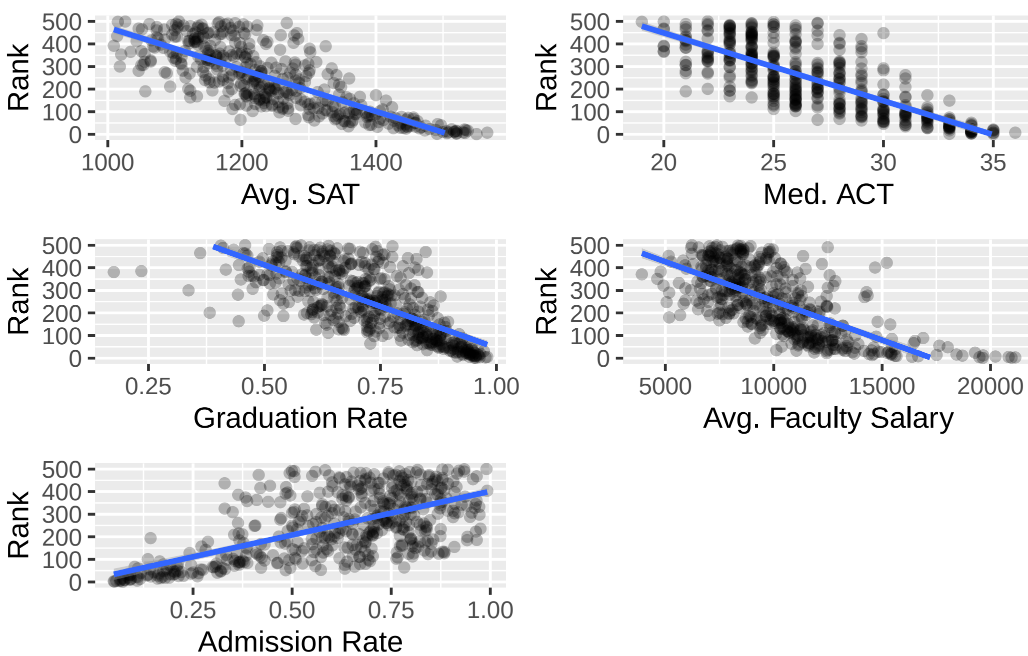
categories. Like accreditation agencies, some classifications corresponded to very few schools in the top 500.

Qualitatively, it appears that region and Carnegie Classification have the clearest relationship with rank; however, the final model indicates that they all have an association with rank.

Numerical Variable Analysis

All of the variables with the top 5 R-squared values appeared in the final stepwise regression model except for `ACTCMMID` (median ACT score) and `C150_4` (graduation rate). These were not used in the model because both of them had a high correlation with `SAT_AVG` (average SAT). Because `SAT_AVG` ended up in the final model, we can reasonably assume that they also have a strong association with rank based on the model's selection process. Therefore, we conclude that `SAT_AVG`, `ACTCMMID`, `C150_4`, `AVGFACSAL`, and `ADM_RATE` have the strongest association with rank and we will characterize the relationship below.

How do these 5 metrics influence Niche college rank?



R-squared, scaled slope, and un-scaled slope for linear regression models between 6 metrics and Niche College Rank

variable	R-Squared	Scaled Slope	Non-scaled Slope
Avg. SAT	0.6291468	-115.00	-0.9277
Med. ACT	0.6062324	-112.90	-30.0400
Graduation Rate	0.5137627	-103.50	-735.2000

variable	R-Squared	Scaled Slope	Non-scaled Slope
Avg. Faculty Salary	0.4692148	-98.89	-0.0347
Admission Rate	0.4048157	91.74	387.0200
% Students with Pell Grants	0.2476545	71.75	644.7600

Analysis

The clearest way to interpret these R-Squared values is the following: 63% of the variation in college rank can be explained by average SAT score. This same interpretation can be used for all of the variables.

Looking further at the relationships, there is a negative relationship between SAT/ACT/Graduation Rate/Faculty Salary and rank. This indicates that as the variables increase, the rank of the school decreases. There is a positive relationship between admission rate and rank, indicating that as this variable increases, the rank of a school increases.

Looking at the non-scaled slope allows us to interpret how much changes in the explanatory variable change rank. For example for ACT, we can say that on average, we can expect a 1-point increase in ACT score to drop the rank of a school by 30 places. For admission rate, since it is scaled from 0-1, we need to divide the slope by 100 to get an interpretable number. It indicates that a 1-point drop in admission rate will, on average, result in an estimated drop in rank of the school by 3.87 places.

Looking at the scaled slope allows us to tell which variables have the greatest "influence" on rank. In other words, which change in a numerical variable away from the mean has the greatest impact on decreasing a school's rank? An extremely interesting trend is that among the five most associated variables, schools that have a stronger R-squared also have a higher absolute value of scaled-slope, indicating that variables that have the strongest association to rank also have the greatest influence on decreasing rank. This is logical because Niche would likely tie their rankings to variables where there is the greatest differentiation between schools with higher and lower ranks. This is also concerning because if schools know which variables are most associated to rank and which ones have the greatest impact on decreasing it, it is fairly easy for them to know which variables to change if they wanted to manipulate the rankings.

Discussion

Based on our analysis and approaches, `SAT_AVG`, `ACTCMMID`, `C150_4`, `AVGFACSAL`, and `ADM_RATE` are the numerical explanatory variables most associated with college ranking, which partially confirm our initial hypothesis. SAT/ACT/Admission Rate were among the most correlated variables, but family income was not in the top five, possibly because financial aid allows students from various financial backgrounds to attend universities. These relationships indicate certain priorities in college rankings. The existence of the SAT, ACT, and admission rate variables in the top five highlight how rankings prioritize selectivity in college admissions. This makes us wonder if colleges focus on improving admissions selectivity over their quality of education and student outcomes. The inclusion of graduation rate and faculty salary do tell slightly different stories. While graduation rate indicates a focus on the ability of a university to meet the needs of its students, faculty salary may indicate the quality of the faculty both in teaching and research.

The stepwise regression model indicated that the categorical variables `REGION`, `ACCREDAGENCY`, `CONTROL`, and `CCBASIC` were also important to calculating rank. Unlike the numerical variables, it is interesting to note that these variables cannot change easily from year-to-year, so colleges cannot use them to manipulate their rankings. Additionally, there are some variables that appear in the final model that have a lower individual R-squared value than some that were taken out of the model. We believe that this is because AIC examines the collective predictive power of the variables rather than the individual predictive ability.

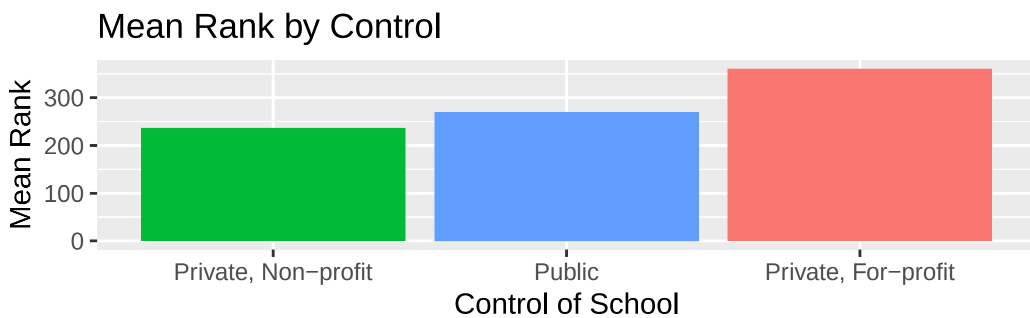
The limitations of our analysis are as follows: We left the categorical variables out of the first approach because we do not know a way to numerically analyze them, and in doing so, we could not subject these variables to the same two-step confirmation process that we did for the numerical values. Furthermore, we assumed that all variables had a linear relationship with ranks so we could use linear regression modeling to analyze them. Additionally, our linear models assume that rank is continuous and goes on forever. We recognize that this is not the case, but since the rank values have meaning and we have not learned how to properly work with ranked data in this class, we decided that a linear regression model was our best approach. All of these issues could be resolved by learning and implementing more appropriate statistical methods.

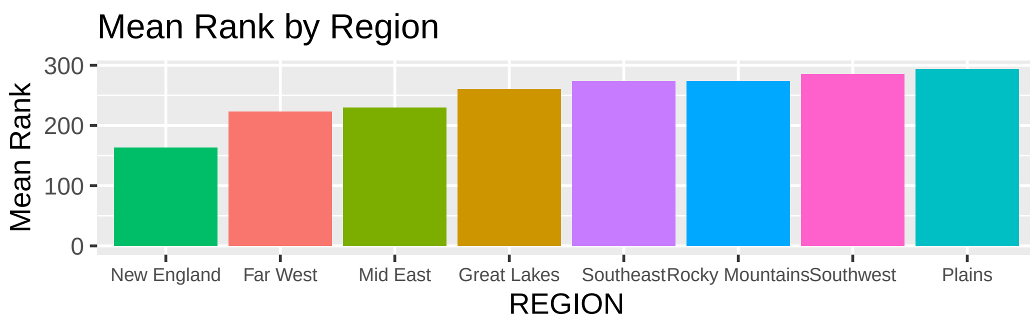
Finally, we believe that future avenues for this project could include analyzing and comparing more ranking systems, such as those created by US News and Forbes. It would be useful if students could understand what each system values and use the one most in line with their priorities. Additionally, we would like to look at over 1000 colleges to see if our results stay consistent between colleges throughout the country, and perhaps even throughout the world.

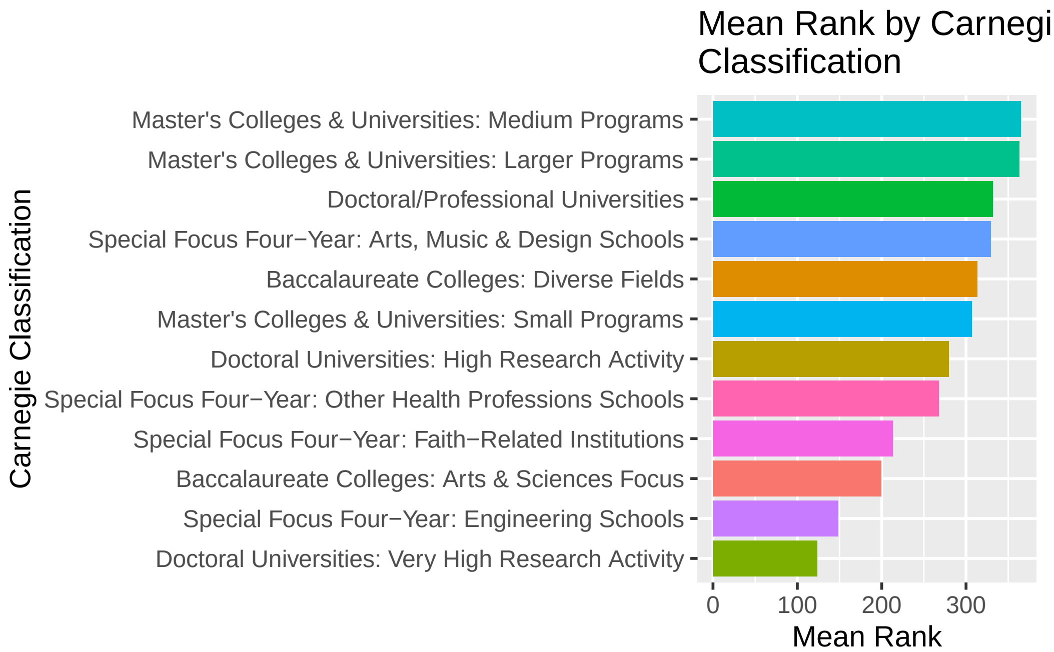
Appendix

Exploratory Analysis: Means of Rank by Categorical Variables

Below, we group the schools by the different categorical variables in our analysis and then take the mean rank for each of those groups. For Carnegie classification, any classification with only one school was removed from the analysis.







We observe that private non-profit colleges have a higher mean rank than public colleges or private for-profit colleges.

As far as region, schools from New England have the highest mean rank, while schools from the Plains have the lowest mean rank.

Looking at the Carnegie Classification, Doctoral Universities: Very High Research Activity have the highest mean rank, followed by Special Focus Four-Year: Engineering Schools. Master's Colleges and Universities: Medium Programs have the lowest mean rank.

Full R-square Value List

variable	r_squared
SAT_AVG	0.6291468
ACTCMMID	0.6062324
C150_4	0.5137627
AVGFACSAL	0.4692148
ADM_RATE	0.4048157
PCTPELL	0.2476545
UGDS_NRA	0.2040361
FIRST_GEN	0.1963961
UGDS_ASIAN	0.1946322
ENDOWBEGIN	0.1643675

variable	r_squared
COSTT4_A	0.1577993
FAMINC	0.1525416
AGE_ENTRY	0.1462016
MARRIED	0.1072270
FEMALE	0.0752810
MD_FAMINC	0.0661957
NPT4_PUB	0.0568142
UGDS	0.0548673
UGDS_2MOR	0.0476589
UGDS_BLACK	0.0463458
NPT4_PRIV	0.0419051
UGDS_WHITE	0.0304200
UGDS_NHPI	0.0222752
UGDS_UNKN	0.0126394
UGDS_AIAN	0.0089268
UGDS_HISP	0.0040103

References

- Learned how to do for loops from TA Eli Gnesin
- We used the `scale()` function found at https://www.statology.org/standardize-data-in-r/
- https://www.niche.com/colleges/search/best-colleges/
- https://collegescorecard.ed.gov/data/
- https://www.youtube.com/watch?v=ejR8LnQziPY
- https://stackoverflow.com/questions/57248708/stepwise-model-selection-in-an-r-tidyverse-workflow
- https://stackoverflow.com/questions/53135404/filter-correlation-matrix-r
- https://stackoverflow.com/questions/68093071/how-to-highlight-high-correlations-in-ggpairs-correlation-matrix
- http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram
- https://www.tutorialspoint.com/how-to-deal-with-missing-values-to-calculate-correlation-matrix-in-r
- https://www.displayr.com/how-to-create-a-correlation-matrix-in-r/
- https://stats.stackexchange.com/questions/550537/how-to-get-r-squared-after-doing-stepwise-model-selection-in-regression-in-r
- http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/
- https://www.researchgate.net/figure/R-2-and-RMSE-of-forward-stepwise-regression-models-vs-WHO-algorithm_tbl1_354396022
- https://www.r-bloggers.com/2016/05/visualizing-bootrapped-stepwise-regression-in-r-using-plotly/