Institutional Value in Higher Education

A Data-Driven Analysis of Cost, Quality, and Outcomes

*Note: Due to limited data in my files on homework 4, I added 3 more sources, which includes the FBI Crime Data with 8,365 entries, GeoNames Zip Code data with 9,918, and a Zippopotam.us API to account for missing zip codes in the Zip Code Data.

Abstract

My research aims to study cost, outcomes, and satisfaction ratings of undergraduate institutions throughout the United States. Does "more" really mean better? Through correlational research of over 6,000 American institutions, including both private and public universities and community colleges, I studied what factors make America's top schools rank so highly. My research involves six data sources, which includes two static web pages, a JSON API endpoint from a dynamic React-based webpage, two APIs, and a downloaded CSV file. Through various steps, including data extraction, cleaning, and aggregation, this project aims to explore which factors play a role in admissions rates, cost, and rankings.

Through statistical and exploratory data analysis, I examined the statistically significant factors that contribute to better student experiences and long-term outcomes. I discovered three overall themes in my research. The first theme was that admissions selectivity is most strongly associated with higher SAT scores, greater faculty pay, and better 4-year graduation outcomes, which suggests that prestige is closely tied to academic rigor and institutional investment. Secondly, crime and location, contrary to initial expectations, did not directly influence institutional rank or tuition, but did correlate with student satisfaction. This is likely due to top-tier institutions being located in urban centers that, despite higher crime rates, offer abundant opportunities and resources. Finally, success outcomes, as measured by post-graduation earnings and completion rates, are significantly associated with higher faculty compensation, lower student-faculty ratios, and greater cost of attendance, suggesting that expensive, well-resourced institutions may offer better long-term returns. I concluded that "more" really does mean better (sometimes), but we must also pay attention to key factors such as school administration, faculty student ratios, and support measures from the university to truly get a worthy return on investment for educational costs.

Motivation and Hypothesis

This research investigates the value proposition of higher education - identifying which schools offer the best return on investment in terms of student satisfaction, safety, support, and post-graduate success. I employed exploratory data analysis (EDA) to examine which factors play the biggest role in both student satisfaction and success outcomes, such as 4-year completion rate and median earnings after 8 years of entry. Many studies have highlighted the impact and importance of higher education, as it contributes to increased social mobility, critical thinking and problem solving skills, and civic engagement to name a few. At the start of my research, I hypothesized that key variables, such as SAT scores of accepted students, local crime rates, resources and facilities, school prestige, and cost of attendance will be highly correlated with admission rate. The findings of this work aim to pinpoint variables of successful

institutions to better invest in schools with limited funding, such as community colleges or state institutions.

I was particularly interested in this topic because when deciding on a Masters program, I was faced with choosing between a public and private institution. I ultimately opted for the private institution, which although was more costly, also provided more promise for network building and personalized support through career services mentorship opportunities. I wanted to explore whether this is the case for all schools with higher tuition rates or whether other factors contribute to educational satisfaction and success factors.

Data Overview

My datasets consisted of six data sources, which included two static web pages, a JSON API endpoint from a dynamic React-based webpage, two APIs, and a downloaded CSV file. After extracting data through API queries and web scraping into CSV files, I prepared the data for aggregation by normalizing institution and city names, removing unnecessary columns to prevent noisy outcomes, and added a zip code column to all datasets.

Three of my sources were webpages with university and geographical data. My first source is the Forbes Top 500 Ranked Universities webpage, which contains 500 entries of ranked universities in the United States with information on acceptance rate, base salary, and institution type. Because this dynamic page was rendered with React, I scraped this page from a JSON API endpoint found in the browser's Network tab in Developer Tools. My second data source is from a website called myPlan.com, which contains 611 entries of self reported rankings by university of overall satisfaction, prestige, personal safety, school resources & facilities, and more. This source was fairly simple to extract, as entries were found in an HTML table and URL endpoints differed for each variable, allowing for automated mass extraction based on the URL. My final webpage source was from a page called GeoNames.org that contained 9,918 entries of city postal code and county information by state for . Similarly to the myPlan data, extraction from this page was fairly simple as data could be parsed through an HTML table from unique URLs based on state. Below are snippets of the output files from Forbes, myPlan, and GeoNames respectively.

	rank	name	state	grade	medianBaseSalary	studentPopulation	campusSetting	schoolSize	description	uri	institutionType	carnegieClassification	cultyRatio	totalGrantAid	percentOfStudentsFinAid	percentOfStudentsGrant
1	1	Princeton University	NJ	A-	189400.0	8631	Urban	medium	r at Princeton University.	princeton-university	Private not-for-profit	ıry High Research Activity	5	46996888.0	62.0	61.0
2	2	Stanford University	CA	A+	177500.0	20490	Suburban	large	tanford commencements.	stanford-university	Private not-for-profit	ıry High Research Activity	5	67892960.0	68.0	53.0
3	3	tts Institute of Technology	MA	A	189400.0	12923	Urban	large	th's transition from Duke.	itts-institute-of-technology	Private not-for-profit	ıry High Research Activity	3	37931652.0	79.0	71.0
4	4	Yale University	CT	A+	168300.0	15828	Urban	large	school's bulldog mascot.	yale-university	Private not-for-profit	ıry High Research Activity	6	58441279.0	59.0	52.0
5	5	ity of California, Berkeley	CA		167000.0	47911	Urban	large	mous UC Berkeley logo.	rsity-of-california-berkeley	Public	ıry High Research Activity	19	7740284.0	60.0	52.0

	School	Prestige	Satisfaction	Resources & Facilities	Safety	Teacher Support	Administration	Campus Setting	Average Score
1	Yale University	96.8	86.3	88.3	66.1	74.7	68.8	86.6	81.09
2	University of California, Berkeley	96.1	83.8	87.8	55.5	52.7	49.8	79.4	72.16
3	Stanford University	95.8	86.3	92.5	89.7	71.0	63.7	92.7	84.53
4	Massachusetts Institute of Technology (MIT)	95.7	83.2	87.1	71.6	66.4	62.1	55.6	74.53
5	Columbia University in the City of New York	95.7	83.7	77.2	68.5	59.3	50.6	74.1	72.73

▾ 0
City "moody"
Postal Code "35004"
State "al"
County "st. clair"
▾ 1
City "adamsville"
Postal Code "35005"
State "al"
County "jefferson"

The following two data sources were accessed from free application programming interfaces (APIs). Given the limited scope of institutions included in the Forbes and myPlan datasets—primarily focusing on highly ranked or well-known schools—I needed a more comprehensive dataset that captured a broader overview of all colleges and universities across the United States. To get this information, I used a federal API provided by the Department of Education, which contains 6,482 universities and their associated median debt of completers, cost of attendance, average earnings after graduation, completion rates, and more. This API was easily accessible with a default rate limit of 1,000 requests per IP address per hour, but requires a unique API key obtained from their documentation page. The second API was a zip code retriever from Zippopotam.us. After using GeoNames data as a mapping dictionary to input zip code data for my data sources based on state and city, not all cities were included in this dataset. This is why I opted to use a robust zip code API with more city data. This API is free and does not require an API key. However, since the API enforces rate limits, I chose to use it in conjunction with my scraped data to minimize the number of queries and avoid exceeding the usage limit. Below are snippets of the output files from the Department of Education and Zippopotam respectively.

	School Name	City	State	Locale	ulty Salary	SAT Score	nission Rate	ar Completion Rate	Price (Public)	Price (Private)	dance (Academic Year)	Median Debt of Completers	Earnings (8 Years After Entry)	Earnings (8 Years After Entry)
1	Alabama A & M University	Normal	AL	12	8651	920	0.684	0.222366710013	14982	N/A	23167	31000	31992	30600
2	f Alabama at Birmingham	Birmingham	AL	12	11837	1291	0.8668	0.454072145779	16755	N/A	26257	22300	51431	44100
3	Amridge University	Montgomery	AL	12	4134	N/A	N/A	0.1328125	N/A	N/A	N/A	32189	34613	40800
4	y of Alabama in Huntsville	Huntsville	AL	12	10267	1259	0.781	0.369397217929	18240	N/A	25777	20705	61771	45900
5	Alabama State University	Montgomery	AL	12	8071	963	0.966	0.24620303757	13527	N/A	21900	31000	32520	27500

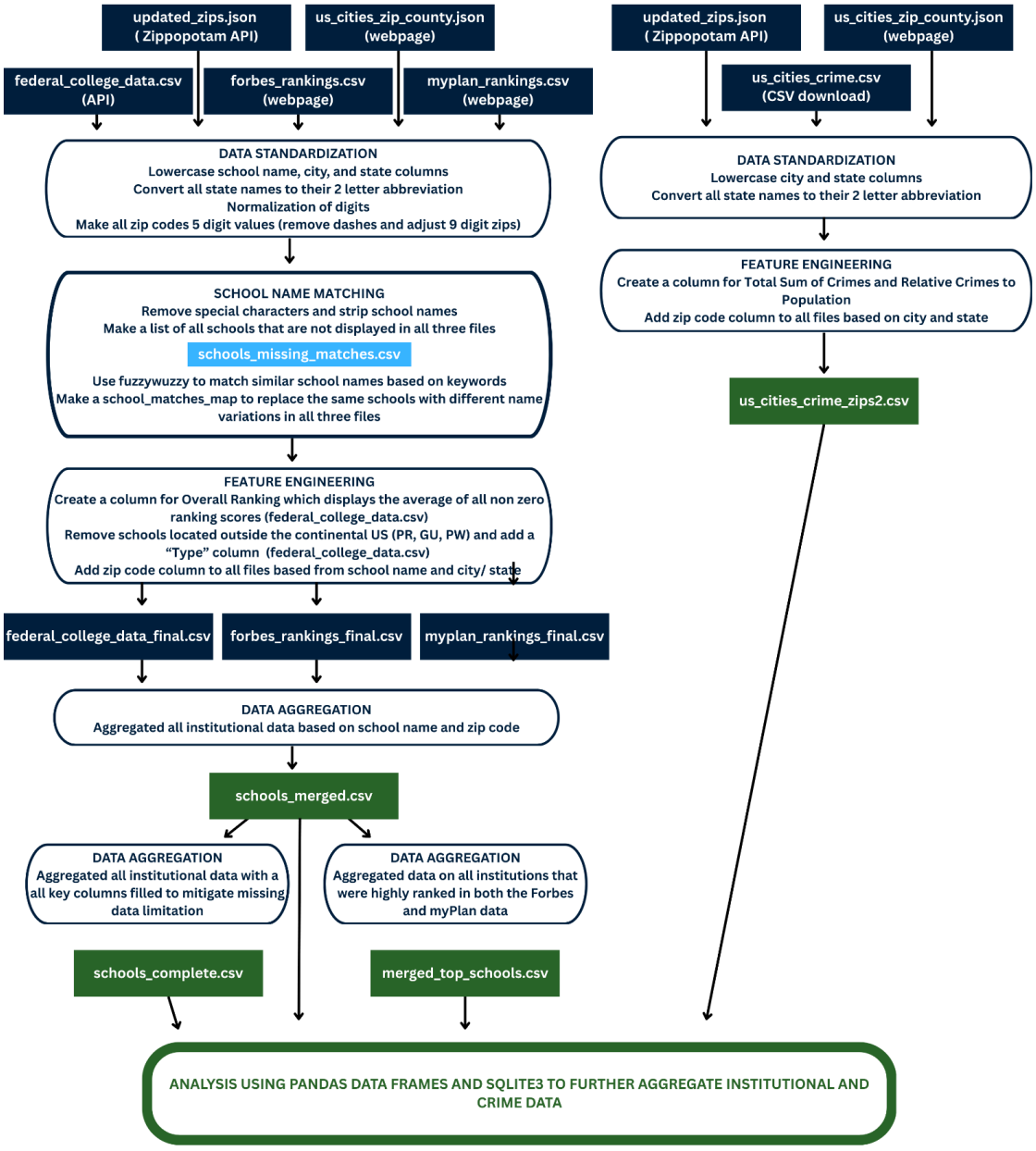
▾ 0
City "abbeville"
Postal Code "29620"
State "sc"
▾ 1
City "addison"
Postal Code "35540"
State "al"

My final data source was a CSV download from the Federal Bureau of Investigation (FBI) Crime Data Explorer. This dataset contained 8,365 entries and contained crime statistics for different cities in the United States, including violent crime, robbery, and burglary rates. Additionally, this dataset included population size, enabling me to analyze not just absolute crime rates, but also crime rates relative to city population—providing a clearer picture of which cities truly had higher crime levels. Below is a snippet of the downloaded file from the FBI.

	State	City	Population	Violent crime	onnegligent manslaughter	Rape	Robbery
1	al	abbeville	2371	6.0	0	0.0	0.0
2	al	adamsville	4158	17.0	0	1.0	5.0
3	al	addison	674	3.0	0	1.0	0.0
4	al	alabaster	34120	37.0	1	1.0	5.0
5	al	albertville	22887	68.0	0	11.0	3.0

Technical Specifications of Aggregation

After thorough data cleaning that included the normalization of digits, and standardization of city and institution names, fuzzy word matching to match schools such as “Columbia University” to “Columbia University of New York,” and feature engineering to add or remove relevant columbs, I aggregated datasets based on common identifiers that included institution name, zip code, or city. Below is a workflow that describes the cleaning and aggregation process. Additionally, I included a table with file descriptions to clarify the data that can be found in each file.



Several different aggregated files were created to capture different trends and to account for the limitation of missing data. Below is a table of the file names and descriptions of the data that they contain.

File Name	Description
schools_merged.csv	This file contains a list of all schools and their associated values, aggregated from the Forbes, myPlan, and federal API data. This document was unified based on institution name, which were normalized by lowercasing and removing non alphabetic characters. After normalization, further cleaning was done by processing unmatched schools using the fuzzywuzzy library to match schools with similar names based on keywords. For example, schools like "Columbia University" were matched to "Columbia University of New York" to ensure that schools were properly matched to its associated entry from all three datasets.
schools_complete.csv	One issue that I faced was missing values. This file is based off of the schools_merged.csv file and contains a list of schools that have all rows filled. The purpose of this document is to analyze well documented schools, which can be used for further training and modeling to develop imputation techniques that can label missing values from other schools.
us_cities_crime_zips2.csv	This file contains the original crime rate data by state with an added column for zip codes. This will allow us to compare crime rates of different institutions based on zip code.
merged_top_schools.csv	This file contains a list of schools that appear on top 300 ranked schools of both the Forbes and myPlan data. These schools represent elite schools as demonstrated by their Forbes rankings and their study body ratings from myPlan. The data was then aggregated with the federal API data to allow for exploratory data analysis on what truly makes these schools so highly ranked.

Maintainability and Extensibility

Maintainability: The code is structured with maintainability in mind, using modular functions that separate data collection, processing, and output (e.g., get_forbes_data, get_myplan_data, get_city_info, merge_and_save_data). Each function includes inline comments and descriptive variable names, making it easier for other developers to understand and modify. Key elements like URLs, fields, and headers are defined at the top of the code blocks, while filenames, global variables, and imports are defined at the top of the file. This allows for quick updates and verification. The use of Python's .get() method for accessing dictionary values ensures that the code does not break if some data fields are missing or rearranged in the API response. Additionally, error handling—such as null checking, safe filename validation, and handling unexpected JSON structures—helps prevent complete code failure and simplifies debugging. The scraping functions are also designed to adapt to moderate HTML changes, as I chose to export table classes (e.g., different tags or class names). Additionally, data extraction, cleaning, aggregation, and

analysis were all done in separate files to remove noise and improve clarity for developers reading the code.

Extensibility: The code is highly extensible and built to accommodate future enhancements with minimal disruption. For instance, adding new data fields from an API or JSON response only requires appending a few key-value pairs to an existing dictionary. Similarly, the ability to append new school ranking categories to a variable_sources list enables easy expansion of analysis scope without changing core functionality. The design supports integration with databases or alternative output formats beyond CSV, and these additions can be implemented with minor code changes. With APIs that serve large datasets—such as those covering over 6,000 institutions—the inclusion of parameters like start_page and end_page ensures flexible and efficient batch processing, preventing server overload or exceeding request limits. By clearly defining URLs, fields, and logic entry points, the code allows contributors to add new features and validate data sources independently, encouraging collaborative development and reducing debugging time.

Technical Challenges and Limitations

Two technical challenges I faced were bot detection from web scraping and missing data. Originally, I had planned to use the US News 2025 "Best National University Rankings" webpage, which contained 436 ranked universities in the United States with information on "tuition and fees" and "undergraduate enrollment." However, after a couple hours and various different attempts to scrape the data, I was unsuccessful in pulling the institution rankings from the webpage. At first, I tried scraping using BeautifulSoup to parse the static HTML content. However, this was unsuccessful because the webpage is a dynamic React page which renders content via asynchronous API calls to JSON endpoints. I then tried using Selenium to mimic user interaction, including scrolling and clicking the "Load More" button to reveal all 436 universities. However, using this method, I was not able to load all pages that included the 436 schools. Finally, I tried extracting data directly from a JSON API endpoint identified through the browser's Network tab in Developer Tools. However, this took a very long time to load and ultimately failed to return any data. At this point, I decided to find a similar source to replace the US News data and came across the Forbes Ranking data, which actually contained more listed institutions and worked well for my research.

Another challenge I faced was missing data from institutions and the inability to tag certain cities with a zip code. Popular cities and universities, such as Los Angeles and the University of Southern California, are very well documented. However, smaller areas with populations of 2,000 in rural America have much less documentation available. Many schools also do not track all their outcomes due to limited resources, funding, and different reporting requirements per state. To address this challenge, I aggregated datasets that reflect schools with almost all variables reported, such as schools_complete.csv and us_cities_crime_zips2.csv. In my analysis of certain variables, I also chose to drop schools with empty values because including incomplete records would compromise the integrity and comparability of the results, potentially introducing bias or skewing insights drawn from the data.

Analysis and Insights

Prior to starting my research, I hypothesized that factors such as SAT scores of accepted students, local crime rates, institutional resources, school prestige, and cost of attendance would be highly correlated with admission rates and overall institutional ranking. While some of these variables did have statistical significance to admission rates and overall institutional ranking, others surprisingly did not play a role. Through the implementation of statistical and exploratory data analysis, I was able to conclude what makes prestigious, highly ranked schools so extraordinary and whether attending a school with a higher cost of attendance is actually worth it.

Admissions Factors

My first finding was that selective institutions with lower admission rates correlate with higher SAT scores, institutional quality (as reflected by faculty pay, prestige, and resources), and better student outcomes (such as 4-year graduation rates and student satisfaction). Figure 1 presents a correlation heatmap, which helps visualize the strength of relationships between these variables. Notably, average SAT score and admission rate are strongly negatively correlated, indicating that more selective schools tend to admit students with higher SAT scores. Similarly, admission rate has a strong negative correlation with 4-year completion rate, faculty salary, and prestige, suggesting that students at more selective institutions are more likely to graduate in four years and have access to higher-paid faculty and prestigious environments.

Interestingly, while resources and facilities, student satisfaction, and campus setting are still negatively correlated with admission rate, the relationships are weaker. This implies that although these factors are present at selective schools, they are not as tightly linked to admissions criteria or outcomes as SAT scores or institutional prestige. Figure 2 further supports these findings by showing a clear negative linear trend between SAT scores and admission rates: institutions that accept fewer students generally admit those with higher test scores.

These insights confirm that while prestige and selectivity often go hand-in-hand with academic quality and student success, not all contributing factors—like student satisfaction or facilities—are equally weighted in that equation.

Figure 1: Correlational Analysis of Admission Rates and its contributing variables

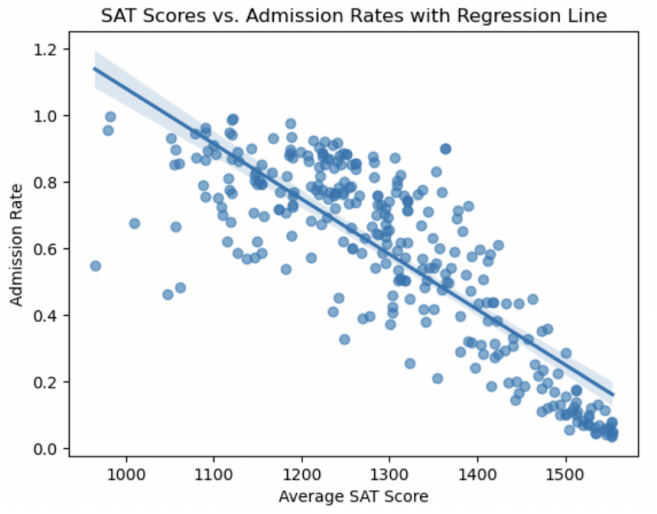


Figure 2: Cluster Map with Regression Line of Admission Rate v. Average SAT Score

Crime and Location

Contrary to my hypothesis, I found that local crime rates did not significantly impact institutional ranking or cost of attendance, but it did play a role in safety ratings and was a notable factor impacting student satisfaction. Looking at the relative crime rate of an institution based on its local population size, I found that the higher the crime rate, the higher the ratings for school administration, satisfaction, prestige, resources and facilities, and overall school ranking score. At first, the positive correlation between crime and prestige was quite interesting. However, after further thought, I realized that institutions located in bigger cities (with often higher crime rates due to the nature of a big city) have more job and internship opportunities, allowing the institution to attain more prestige. Figure 3 shows a correlation matrix for a variety of variables, with a key emphasis on the lower left quadrant, which describes crime rates and institutional factors. Due to the complexity of this correlation, I completed my analysis by calculating the p-value to identify statistically significant relationships and created a document titled *'sig_correlations_with_pvals.csv'* to identify important correlational relationships. This document can be found within my other attached files.

I also ran geographic analysis based on zip code to determine whether schools in certain zip codes resulted in better metrics by mapping low outlier ratings on a map of the United States as shown in Figure 4. To make the map, I downloaded a ZCTA (ZIP Code Tabulation Area) file, which was created by the US Census Bureau and shows geographic areas by zip codes. I then matched the zip codes in the school files to plot outliers. However, as we can see there are minimal patterns to see here as schools ranked poorly in factors such as prestige, satisfaction, resources and facilities, safety, teacher support, school administration, and campus setting produce minimal insightful patterns.

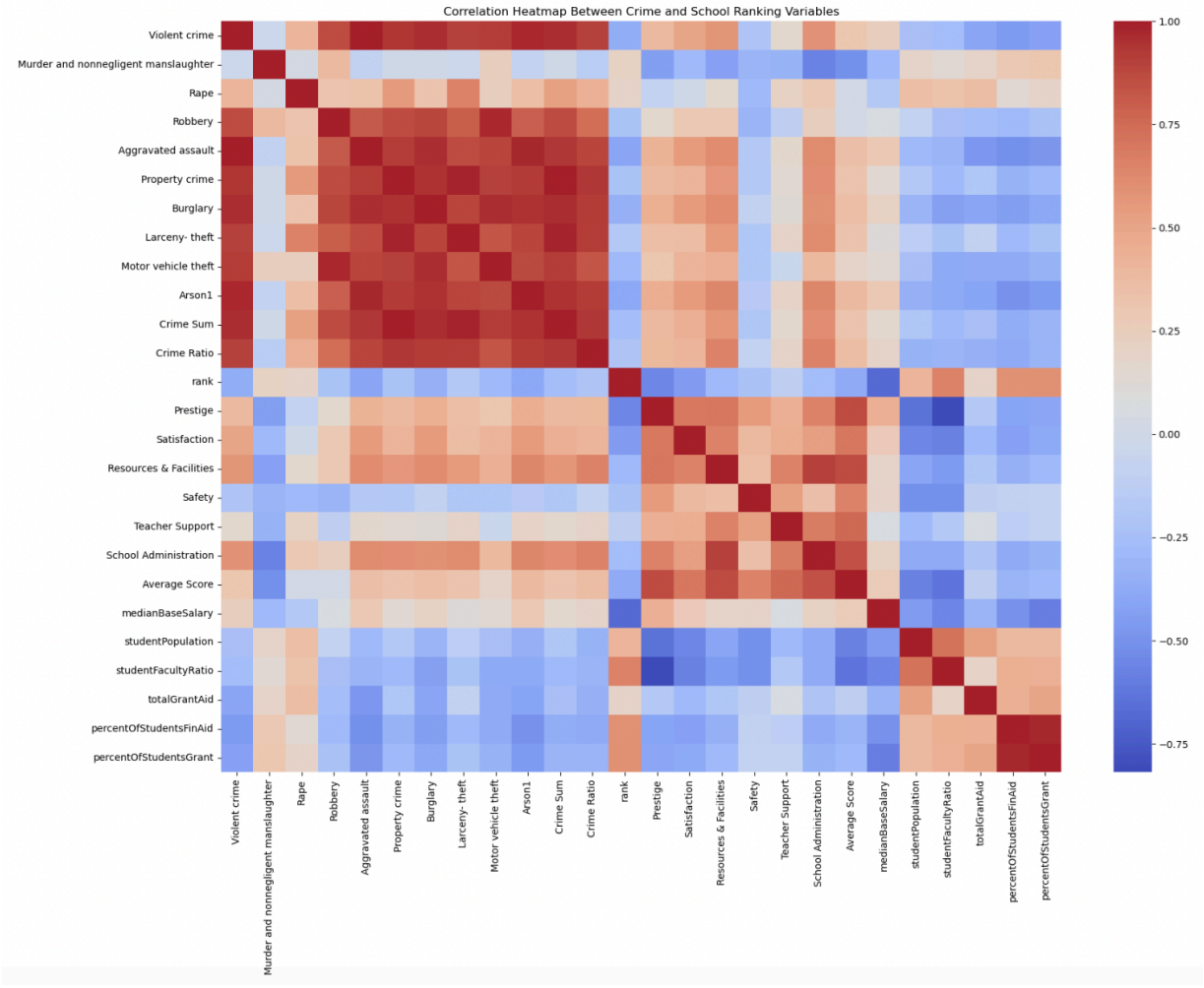


Figure 3: Correlational Analysis of Crime and its contributing variables

Figure 4: Map of the United States of America, Highlighting Low Ranked Schools

Success Factors

I measured success factors by 4-year completion rates, median earnings 8 years post-entry, and median baseline. Through my research, I found that these success outcomes are most strongly associated with cost of attendance, and faculty compensation. Using Random Forest analysis, I found which factors play the biggest role in predicting 4-year completion rate (Figure 5) and median baseline salary (Figure 6). I then used correlation analysis to explore the relationship between these variables (Figure 7). I found that the top factors that contribute to completion rate include cost of attendance, average faculty salary, and student faculty ratio. Faculty pay and cost of attendance had a positive correlation, meaning that they went up as the rate of completion went up. Student faculty ratio and 4 year completion rates were negatively correlated, highlighting the importance of small class sizes to improve overall student success. Median base salary was most impacted by median earnings (8 years after entry), school rank, and cost of attendance. Median earnings and cost of attendance is positively correlated to median base salary. Rank is negatively correlated to median base salary, meaning the lower (better) the school is ranked, the higher likelihood of getting a high paying job after graduation. This is a critical finding, as it supports my hypothesis that better ranked schools with a higher price point may actually offer the best return on investment. However, to confirm this suspicion, I did further analysis on the school cost and type (private vs. public) to really understand the impact on prestige and cost of attendance.

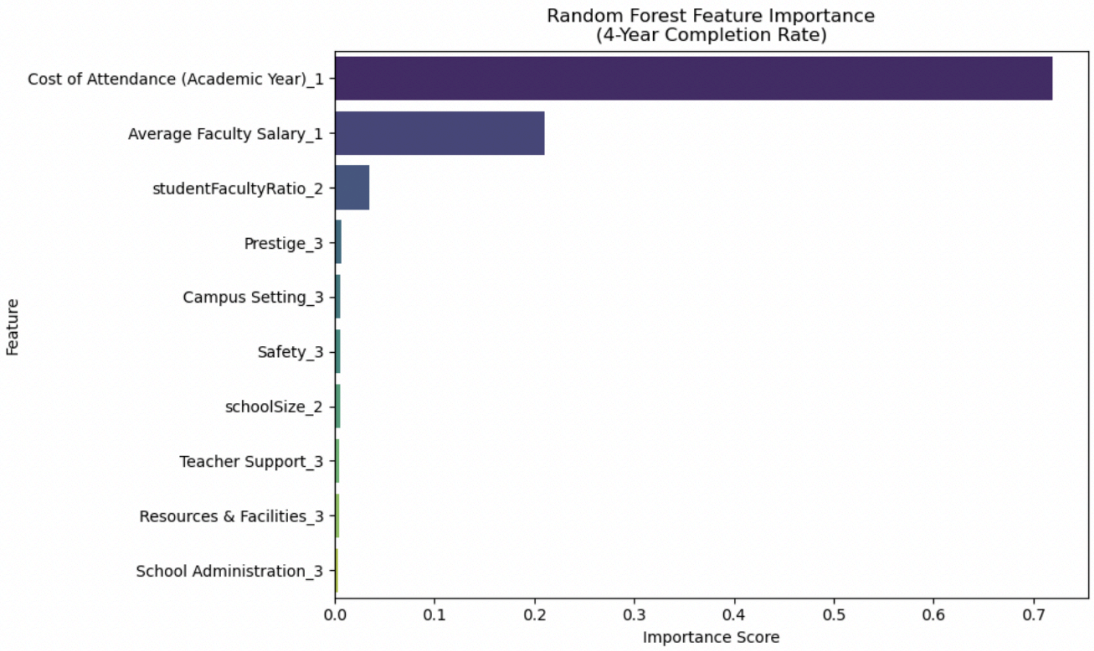


Figure 5: Random Forest Analysis To Examine Which Factors Contribute the Most To 4-Year Completion Rate

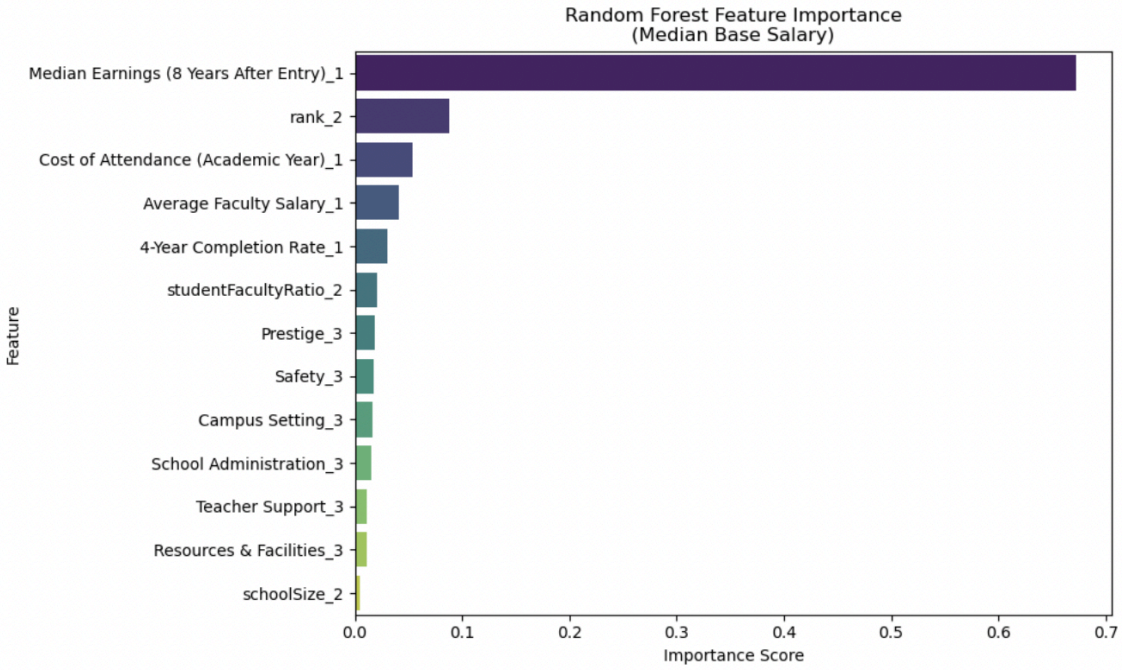


Figure 6: Random Forest Analysis To Examine Which Factors Contribute the Most To Median Base Salary

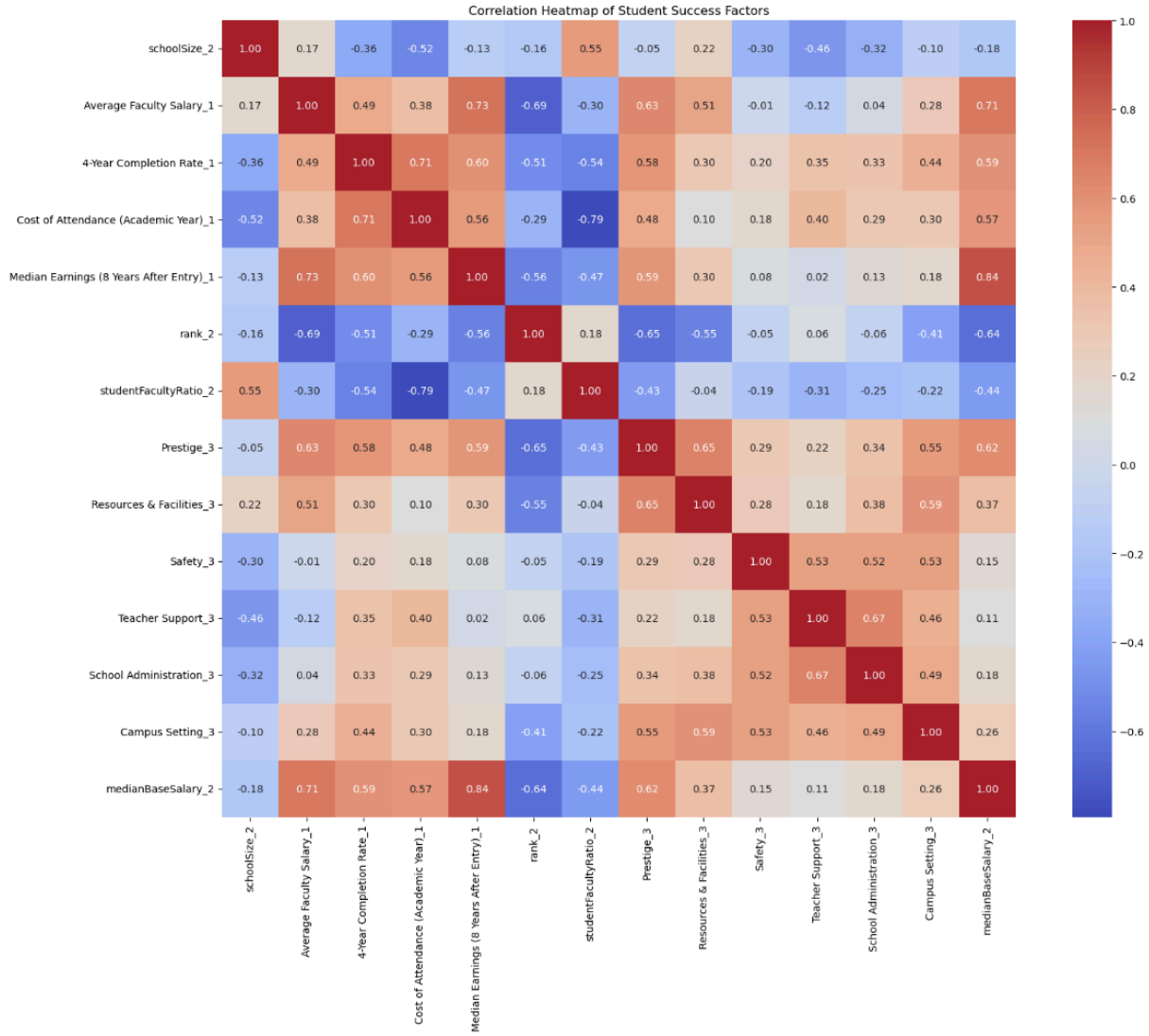


Figure 7: Correlational Analysis On Success Factors (4- Year Completion Rate, Median Base Salary, Earnings 8 Years After Entry)

Cost

I started my analysis by comparing the cost of attendance and median debt after completion by different states and different types of institutions. Figure 8 shows these variables by the private and public institutions. It was no surprise that the cost of attendance and median debt of completers was higher in private institutions due to their reliance on tuition for funding and the lack of public funding. However, it was interesting that the median debt of completers was not too different from public institutions. This can be because private institutions either provide more generous aid or because individuals attending these schools are not relying on loans to cover their educational expenses (due to savings, wealth, etc.). However, the former was disproved by analyzing the grant and aid metrics by institution type as seen in Figure 9.

To study whether higher cost of attendance and prestige equates to better student satisfaction, I did further analysis. Using the schools that were exclusively highly ranked by both the Forbes and myPlan datasets, I used principal component analysis (PCA) to identify two components. PC1 represents overall institution quality (as it is positively correlated with SAT scores, completion rate, and prestige), while PC2 represents student support (as it is associated with teacher support and safety). As seen in Figure 10, schools on the right represent PC1 and consist mostly of private schools with high prestige, SAT Scores, and 4-Year Completion Rate. Schools on the left represent PC2 and consist mostly of public schools. Further analysis in Figure 11 shows how splitting this data into quadrants gives us further insight into what makes costly schools worth the price point by identifying schools based on eliteness (schools with lower acceptance rates, higher SATs, prestige rankings, and cost of attendance) and school support (teacher support, safety, school administration, campus setting, and overall ratings). Some interesting and well known school names were shown in this data. For example, Johns Hopkins, Harvard, and the University of Southern California were all labeled as "elite, but less supportive." Schools like Texas State University and University of Oregon were labeled as "less elite, but supportive." Finally, schools like Princeton College, Dartmouth University, and Williams College were labeled as "elite and supportive." A full list of these school names can be found in the last few pages of this report under the appendix.

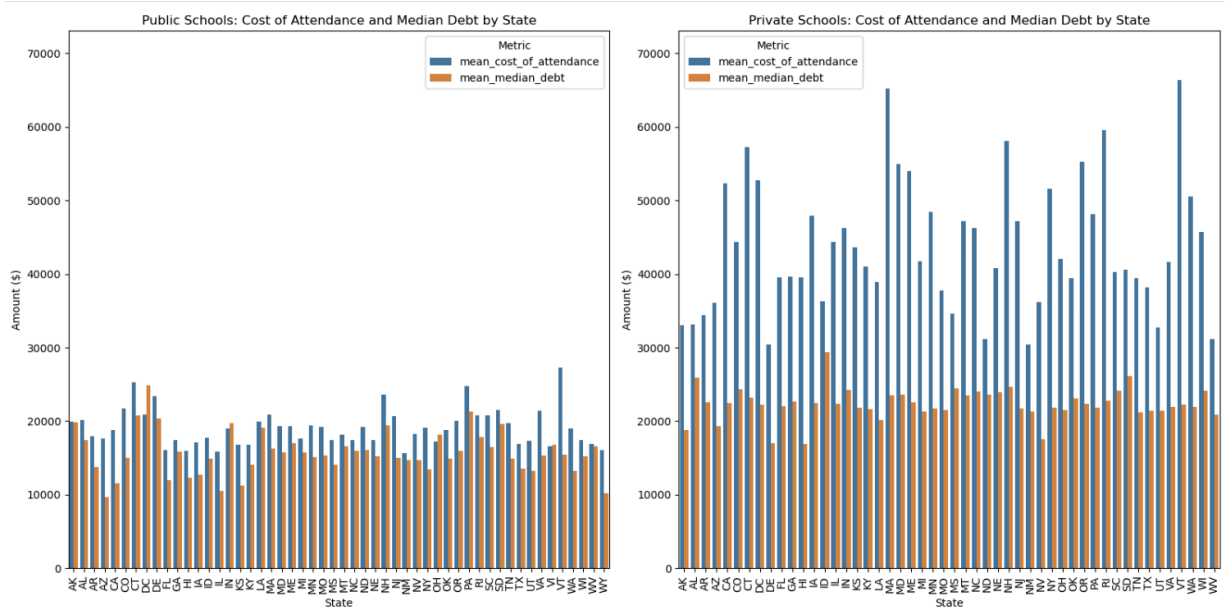


Figure 8: Analysis On Cost of Attendance and Debt of Completers by State for Private and Public Institutions

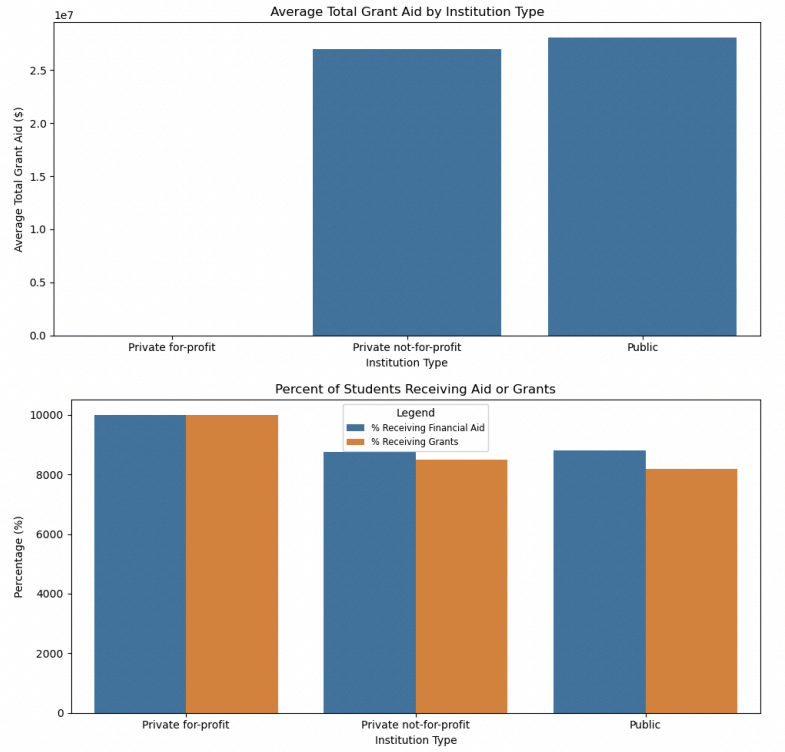


Figure 9: Grant Aid Availability, % of Students Receiving Aid, % of Students Receiving Grants by Institution Type

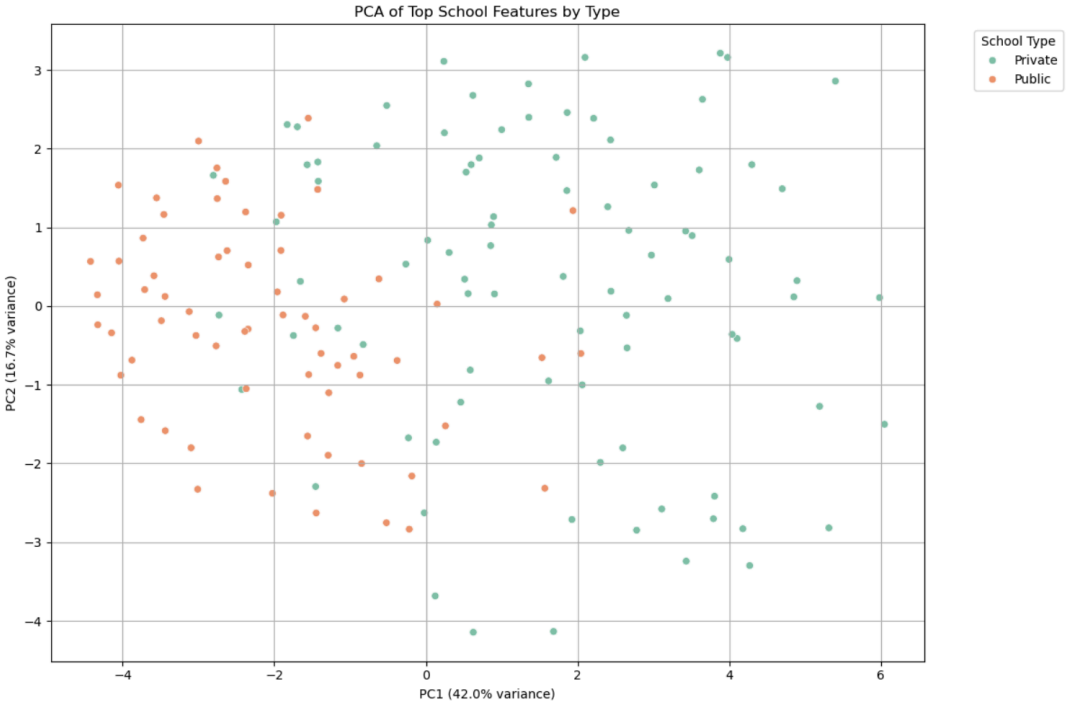


Figure 9: Analysis On Cost of Attendance and Debt of Completers by State for Private and Public Institutions

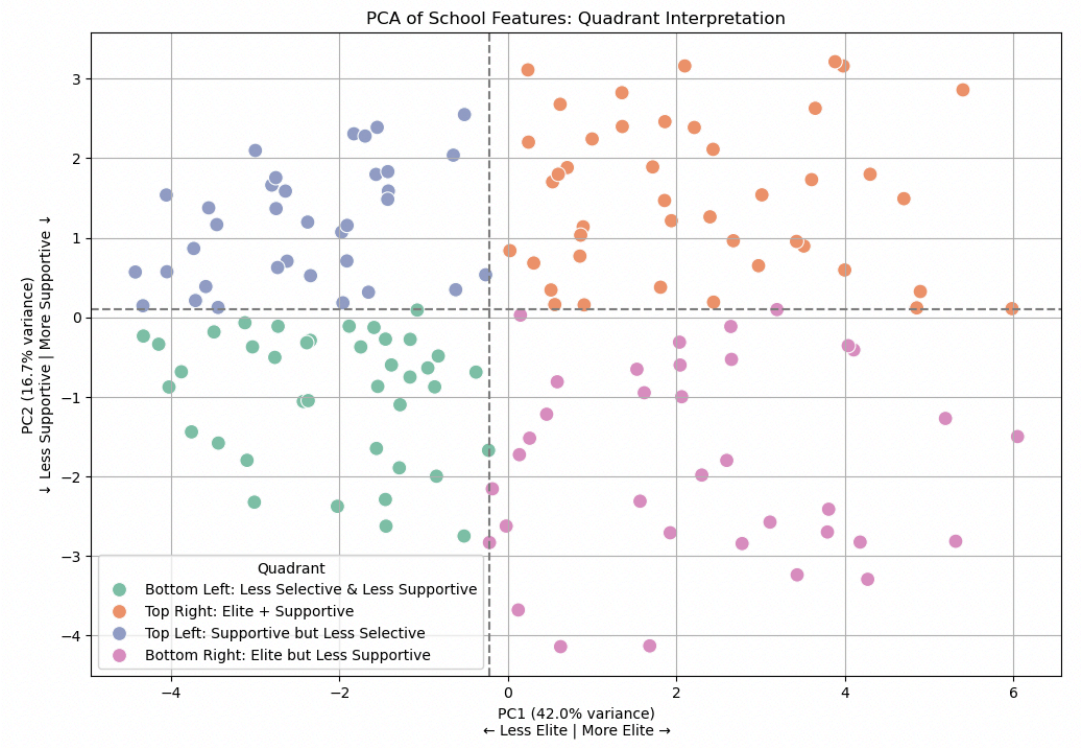


Figure 11: Principal Component Analysis (PCA) on School Eliteness and Support

Conclusions

This research sought to uncover the factors that influence college admission rates, institutional ranking, and student success, with a particular interest in how institutional characteristics like cost, prestige, and location shape these outcomes. Initially, I hypothesized that variables such as SAT scores, local crime rates, institutional resources, prestige, and cost of attendance would significantly correlate with admissions and ranking. While several of these variables proved influential—particularly SAT scores, prestige, and cost of attendance—others, like local crime rate and campus setting, played a more nuanced or negligible role.

Overall, I found that institutional prestige, student success, and satisfaction are not solely driven by cost or selectivity, but rather by a combination of academic investment and supportive learning environments. These findings help decode what differentiates elite institutions from others and highlight the tradeoffs students face when selecting a college.

Impact, Limitations, and Future Work

Although the data supports the notion that higher tuition costs often align with greater institutional quality, academic outcomes, and future earning potential (especially among private, highly ranked schools), it also affirms that prestige alone does not guarantee student support or success. With a passion

for education and a commitment to supporting underserved, marginalized communities, I hope to contribute to efforts that improve educational outcomes for students who may not have the opportunity to attend elite or costly institutions by identifying key areas for investment, such as faculty support, smaller student-faculty ratios, campus safety, and effective school administration. This data is the first step in building more supportive, equitable, and outcomes-driven educational environments.

This research has several limitations. First, data availability was inconsistent across schools, particularly for rural or smaller schools that do not have the resources to document student outcomes. Second, the analysis primarily focuses on four-year institutions in the U.S., limiting its applicability to international or two-year colleges. Additionally, student-level variables (e.g., demographics, financial background, academic preparedness) were not factored in and could further refine findings on student outcomes and institutional accessibility.

In the future, I hope to expand this work by incorporating longitudinal data to track how institutions change over time, as well as student demographic data to better evaluate issues of equity and access. I also plan to conduct qualitative and quantitative surveys with students across a diverse range of schools to better understand their lived experiences and identify gaps in support. I am especially interested in studying rural communities, which often face unique challenges such as limited access to resources that prevent them from offering the same level of academic support, extracurricular opportunities, and institutional investment as more urban or well-funded schools. By analyzing a wide variety of institutional settings, this research can help drive inclusive, student-centered improvements that inform both institutional practices and broader systemic reforms in the education sector.

Final Thoughts

While many college ranking systems focus exclusively on prestige, graduation rates, or alumni salaries, this analysis integrates academic, financial, geographic, and student experience data to provide a multidimensional view of institutional quality. By incorporating variables like crime, teacher support, and student satisfaction, this research highlights the complexity of school value and moves beyond the one-size-fits-all rankings used by platforms like Forbes.

Ultimately, this analysis examines institutions through a holistic approach to reveal that while prestige and cost often go hand in hand, not all expensive schools offer the same level of student support or return on investment. Students should consider both institutional “eliteness” and supportiveness when evaluating their options. By combining data-driven models with student-centered criteria, this research provides a more nuanced and equitable approach to understanding higher education value.

Appendix

Technical Specifications for Extraction

Each data source will be downloaded locally as a CSV file. Before extraction begins, the provided file name will be validated using the function below to ensure it's in the correct format. This function verifies that the filename is not empty, ends with .csv, and contains only valid characters (e.g., letters, numbers, underscores, and dashes). This prevents unnecessary processing in case of invalid filenames.

Function: validate_filename(filename: str)		
Libraries	re	**Example Calls:** validate_filename("colleges.csv") Output: True validate_filename("colleges") Output: False
Input	String - filename	
Output	Boolean - True if the filename is valid, False if the filename is invalid	

Source 1: Forbes Best Colleges Rankings

Variables Collected: Institution Name, Rank, State, Average Grade, Median Base Salary, Student Population, Campus Setting, School Size, Description, Institution Type, Carnegie Classification, Student to Faculty Ratio, Total Grant Aid, Percent of Students Receive Financial Aid, Percent of Students Receive Grants

Overall Description: To collect the Forbes Top Colleges dataset—which includes 500 U.S. undergraduate institutions—I identified a JSON API endpoint by inspecting the browser's Network tab in Developer Tools. I then developed a function to retrieve data by iterating through the JSON objects to extract ranking data per school. The data was then compiled into a structured pandas DataFrame and exported into a CSV file.

Function Description

get_forbes_data(filename: str)		
Libraries	pandas, requests, re	**Example Call** get_forbes_data("forbes_rankings.csv") Output (if successful): '500 universities added to forbes_rankings.csv.
Input	1. File name for a new CSV which will contain extracted data **(str)**	
Outputs	1. String - Error Message if data extraction was not successful or if the filename was invalid	

	2. String - Message confirming successful extraction AND CSV containing Institution Data that includes	

Source 2: myPlan College Rankings

Variables Collected: Institution Name, Prestige, Satisfaction, Resources & Facilities, Personal Safety, Teacher Support and Involvement, School Administration, Campus Setting, Aggregate score of all variables (Average Score)

Overall Description: I collected college ranking data by scraping multiple pages from MyPlan.com using customized URLs and HTML parsing with BeautifulSoup. For each variable, I iterated through pagination offsets based on the number of school entries, dynamically adjusting the pages accessed. Schools with a valid score were extracted and stored in a list, which was then merged with other variables based on school name. All variable data was gathered into a single CSV file and an average score was calculated for each school based on truthy values to prevent a reduced average score for schools with missing values.

Function Descriptions

<table>
<tr><th colspan="3">get_myplan_data(url: str, entries: int)</th></tr>
<tr><td>Libraries</td><td>BeautifulSoup, requests</td><td rowspan="3">Example Call
Input:
variable_sources = [
("prestige",
"https://www.myplan.com/education/colleges/college_rankings_8.php?sort=1&offset=", 611),
("satisfaction",
"https://www.myplan.com/education/colleges/college_rankings_1.php?sort=1&offset=", 613),
("resources",
"https://www.myplan.com/education/colleges/college_rankings_4.php?sort=1&offset=", 611)
]

rankings = {}
for name, url, entries in variable_sources:
rankings[name] = get_myplan_data(url, entries)

Output (if successful):
611 total entries were found and documented.
613 total entries were found and documented.</td></tr>
<tr><td>Input</td><td>1. URL (str) containing a ranking list of schools from myPlan.com
2. Number of entries (int) of schools ranked in the list, which will be used to paginate and parse through all results</td></tr>
<tr><td>Outputs</td><td>1. Error Message (str) if data extraction was not successful
2. Message confirming successful extraction (str) AND unique variable lists with university names and ranking scores (list)</td></tr>
</table>

merge_and_save_data(variables_list:list, filename:str)		
Libraries	csv, re	**Example Call** **Input:** merge_and_save_data([rankings["prestige"], rankings["satisfaction"], rankings["resources"]) **Output (if successful):** Saved to myplan_rankings.csv
Input	1. List of variables containing a a list of all variables and their college ranking scores **(list)** 2. File name for a new CSV which will contain extracted data **(str)**	
Outputs	1. Error Message **(str)** if the filename was invalid or if data was not successfully merged 2. Message confirming successful extraction (str) AND CSV containing Institution Data that includes	

Source 3: US Department of Education College Scorecard

Variables Collected: School Name, City, State, Locale, Average Faculty Salary, Average SAT Score, Admission Rate, 4-Year Completion Rate, Average Net Price (Public), Average Net Price (Private), Cost of Attendance (Academic Year), Median Debt of Completers, Median Earnings (8 Years After Entry), Mean Earnings (8 Years After Entry)

Overall Description: This script fetches higher education institution data from the U.S. Department of Education's College Scorecard API. It retrieves user-specified pages of data, focusing on selected fields that are clearly defined at the top of the code block. The script handles pagination, manages missing values, and appends the output to a CSV file without overwriting previous data.

Function Descriptions

get_myplan_data(url: str, entries: int)		
Libraries	csv, os, re, requests	**Example Call** **Input:** intervals = [(0,10), (10,20), (20,30)] api_file = "federal_college_data.csv" for i in intervals: get_data_to_csv (i[0], i[1], api_file) **Output (if successful):** Fetching page 1…
Input	1. Start page **(int)** to define which page to start accessing and extracting data (100 institutions per page) 2. End page **(int)** to define which page to stop extracting data 3. File name for a new CSV which will contain extracted data **(str)**	

Outputs	1. Error Message **(str)** if data extraction was not successful or if filename is invalid 2. Message confirming successful extraction **(str)** AND a CSV containing all extracted data	Fetching page 2… (...) Data from pages 0 to 10 saved to 'federal_college_data.csv - 1000 records' (...)

Source 4: GeoNames Zip Code and County Extraction

Variables Collected: city name, state name, zip code, county

Overall Description: This script fetches city information by accessing a URLs HTML code and parsing it with BeautifulSoup. For each variable, I iterated unique state URLs that required the full state name and the abbreviation. To speed up the extraction process, I looped through all states and executed the function, appending them to a csv file after extraction. I then turned this CSV file to a JSON file to use as a mapping dictionary when labeling zip codes by city for other documents.

Function Descriptions

get_city_info(state, abbrev, filename)		
Libraries	requests BeautifulSoup	**Example Call** **Input:** abbrev = “ca” state = “california” filename = "us_cities_zip_county.csv" **Output (if successful):** Data from california has been added to us_cities_zip_county.csv!
Input	1. State name (str) 2. State Abbreviation **(str)** 3. Filename of the CSV file that you want to extract your data to **(str)**	
Outputs	1. Error Message **(str)** if data extraction was not successful 2. Message confirming successful extraction **(str)** AND a CSV containing all extracted data	

Source 5: Zippopotam.us Zip Code API

Variables Collected: city name, state name, zip code

Overall Description: This script fetches zip code data when given a city name and state. If a city has more than one zip code associated with it, it will return the first zip code found, still providing us with the overall geographic information needed.

Function Descriptions

get_zip_codes_from_api(state: str, city: str)		
Libraries	requests	**Example Call** **Input:** state = “ca” city = “los angeles” **Output (if successful):** 90089
Input	1. State Abbreviation **(str)** 2. City Name **(str)**	
Outputs	1. Error Message **(str)** if API query was not successful 2. Zip code of associated city, state **(str)**	

PCA Cluster Analysis Schools Pertaining To Figure 11

Top Left: Supportive but Less Selective (35 schools):
- appalachian state university
- clarkson university
- clemson university
- college of charleston
- college of new jersey
- florida gulf coast university
- furman university
- gonzaga university
- gustavus adolphus college
- hope college
- illinois state university
- james madison university
- kansas state university
- kenyon college
- lewis & clark college
- mississippi state university
- new mexico institute of mining and technology
- oklahoma state university
- pacific lutheran university
- southwestern university
- suny college geneseo
- texas christian university
- texas state university
- towson university
- truman state university
- university of dayton
- university of hawaii manoa
- university of mississippi
- university of north carolina greensboro
- university of oklahoma norman
- university of oregon
- university of wisconsin eau claire
- utah state university
- virginia military institute
- western washington university

Top Right: Elite + Supportive (43 schools):
- amherst college
- babson college
- bentley university
- bowdoin college
- brandeis university
- brigham young university
- bryn mawr college
- bucknell university
- carleton college
- colgate university
- college of holy cross
- colorado college
- dartmouth college
- elon university
- franklin and marshall college
- grinnell college
- harvey mudd college
- lafayette college
- macalester college
- middlebury college
- mount holyoke college
- oberlin college
- pepperdine university
- pomona college
- princeton university
- reed college
- rice university
- rose hulman institute of technology
- santa clara university
- smith college
- southern methodist university
- swarthmore college
- trinity university
- university of richmond
- vanderbilt university
- vassar college
- villanova university
- wake forest university
- wellesley college
- wesleyan university

	- whitman college - college of william & mary - williams college
Bottom Left: Less Selective & Less Supportive (43 schools): - adelphi university - american university - auburn university - baylor university - depaul university - florida international university - florida state university - indiana university bloomington - iowa state university - miami university - michigan state university - north carolina state university raleigh - ohio state university - pennsylvania state university main campus - purdue university - seattle university - syracuse university - texas a&m university college station - university of alabama - university of arizona - university of arkansas - university of central florida - university of colorado boulder - university of delaware - university of georgia - university of illinois urbana champaign - university of iowa - university of kansas - university of maryland college park - university of miami - university of minnesota twin cities - university of missouri columbia - university of nebraska lincoln - university of north texas - university of south carolina - university of texas arlington - university of utah - university of wisconsin madison - west virginia university	Bottom Right: Elite but Less Supportive (35 schools): - boston college - boston university - brown university - carnegie mellon university - claremont mckenna college - columbia university in city of new york - cornell university - duke university - emory university - fordham university - georgetown university - georgia institute of technology - harvard university - johns hopkins university - lehigh university - massachusetts institute of technology - new york university - northeastern university - northwestern university - rensselaer polytechnic institute - stanford university - tufts university - tulane university - university of chicago - university of florida - university of michigan ann arbor - university of north carolina chapel hill - university of notre dame - university of pennsylvania - university of southern california - university of texas austin - university of virginia - virginia polytechnic institute and state university - washington university in st louis - yale university