Identifying Institutional Peers through Machine Learning

Introduction

Within today's environment of data-driven institutional management, higher education researchers employ comparative analysis as a tool for assessing and enhancing university outcomes. Academic leaders utilize benchmarking indicators such as six-year graduation rates to measure their institution's standing against comparable schools, recognize top-performing national universities, and discover effective strategies for reaching institutional objectives. This project applies clustering techniques to determine peer institutions for the University of Houston (UH) using federal postsecondary education data, aiming to support university advancement through systematic comparison.

Data and Methods

The analysis draws upon information from the Integrated Postsecondary Education Data System (IPEDS), maintained by the National Center for Education Statistics. IPEDS compiles information from all colleges and universities as a condition of their participation in federal student financial assistance programs. Information is submitted on an annual basis and encompasses indicators related to student enrollment, degree completions, retention and graduation metrics, personnel composition, institutional finances, tuition rates, and financial assistance programs. This information is gathered uniformly across participating institutions to enable valid comparisons.

Table 1: Institutional features

Variables
Institution size category
Sector of institution
Carnegie Classification 2018 - Basic
Percent admitted - total
Admissions yield - total
SAT Reading/Writing 25th percentile score
SAT Reading/Writing 75th percentile score
SAT Math 25th percentile score
SAT Math 75th percentile score
ACT Composite 25th percentile score

Variables
ACT Composite 75th percentile score
Undergraduate enrollment, Fall 2018
Graduate enrollment, Fall 2018
Percent American Indian or Alaska Native
Percent African American
Percent Hispanic/Latino
Percent White
Percent two or more races
Percent Nonresident Alien
Percent Asian/Pacific Islander
Percent women
FTE for last academic year, 2017-18
Percent of full-time first-time undergraduates awarded Pell grants
Student-to-faculty ratio
All academic rank faculty
Six-year graduation rate, BA Degree
Total degrees awarded (Bachelor's, Masters, Doctoral)
Core revenues, total dollars (GASB)
Core expenses, total dollars (GASB)
Instruction expenses as a percent of total core expenses (GASB)
Research expenses as a percent of total core expenses (GASB)
Student service expenses as a percent of total core expenses (GASB)
Endowment assets (year end) per FTE enrollment (GASB)
Tuition and fees as a percent of core revenues (GASB)
In-state average tuition for full-time undergraduates
Out-of-state average tuition for full-time undergraduates
Total price of attendance in-state living on campus

To establish a pool of potential peer institutions, the analysis applies three widely recognized filtering criteria: 1) institutions must be public four-year or higher degree-granting institutions, 2) hold the Carnegie Classification designation of Doctoral Universities with Very High Research Activity, and 3) maintain enrollment of 20,000 students or above.

Initially, 796 institutions were classified as public four-year or higher institutions. Further restriction to those with Carnegie Doctoral Universities designation yielded 94. The final filter based on institutional size produced a sample of 85 institutions available for analysis.

The dataset contains one observation per institution across 34 distinct institutional indicators. Clustering procedures require datasets without missing information. Nine institutions out of 85 (11%) contained at least one missing data point. Statistical imputation was conducted for these cases across features including revenues, revenue composition, expenditures, instructional spending proportions, research spending proportions, student services spending proportions, endowment values per student, and standardized test scores. Among the total data points, only 50 values required imputation, representing merely 1.7% of available information. Additionally, features were standardized to ensure comparability across different measurement scales.

Cluster Analysis

Clustering is a collection of statistical and machine learning techniques designed to compute similarity between institutions. This analytical approach involves collecting institutional attributes, measuring differences among universities on these attributes, and forming groups where internal similarity is maximized while differences between groups are maximized. Essentially, this technique organizes institutions according to their similarities and distinguishes them from alternative groupings.

Multiple approaches exist for computing similarity measures. This analysis employs the commonly applied Euclidean distance calculation. This formula computes each institution's distance from every other institution based on the 34 indicators. The visualization displays these distance calculations, where warmer colors represent greater dissimilarity and cooler colors indicate higher similarity. The diagonal represents perfect similarity where an institution intersects with itself.

K-means Clustering

The K-means approach represents a machine learning algorithm designed to divide data into a predetermined number of groups. This technique categorizes observations based on their characteristics, ensuring observations within each group are maximally similar while being maximally different from members of other groups. Each group possesses a center point representing the average position of the data defining those groups. Optimal groupings should demonstrate minimal internal variation and small distances from their center points.
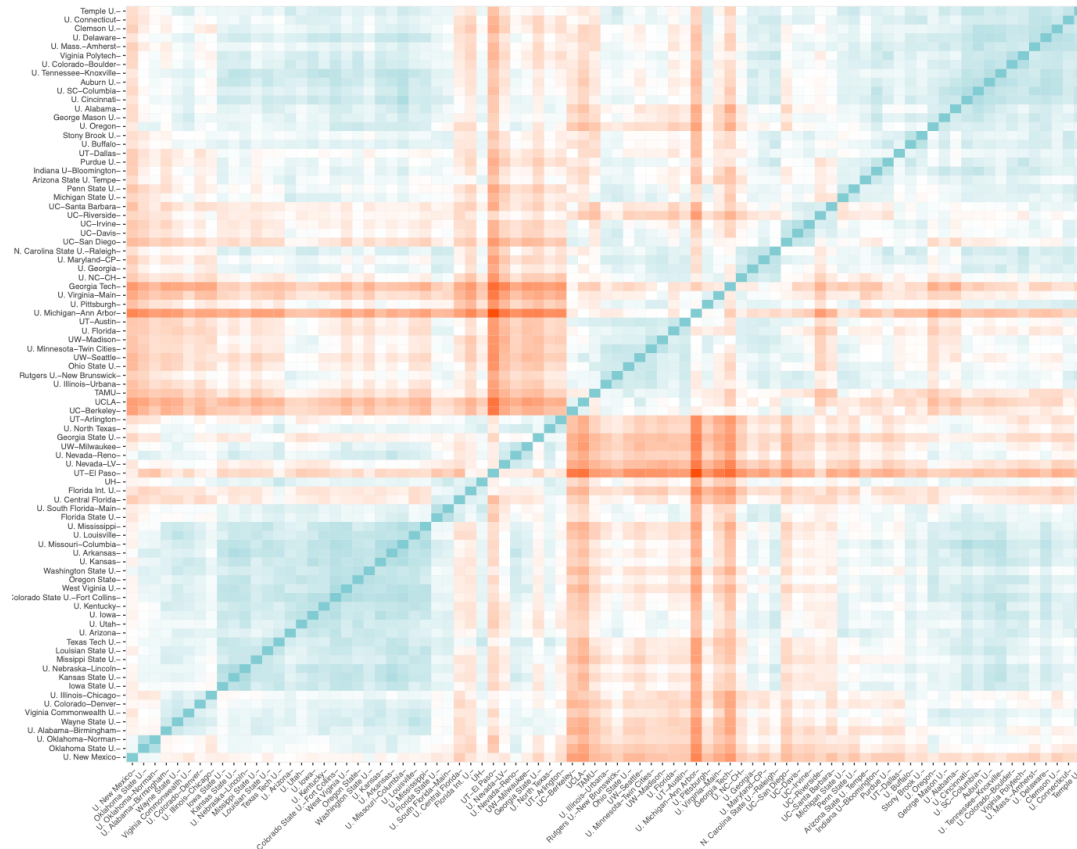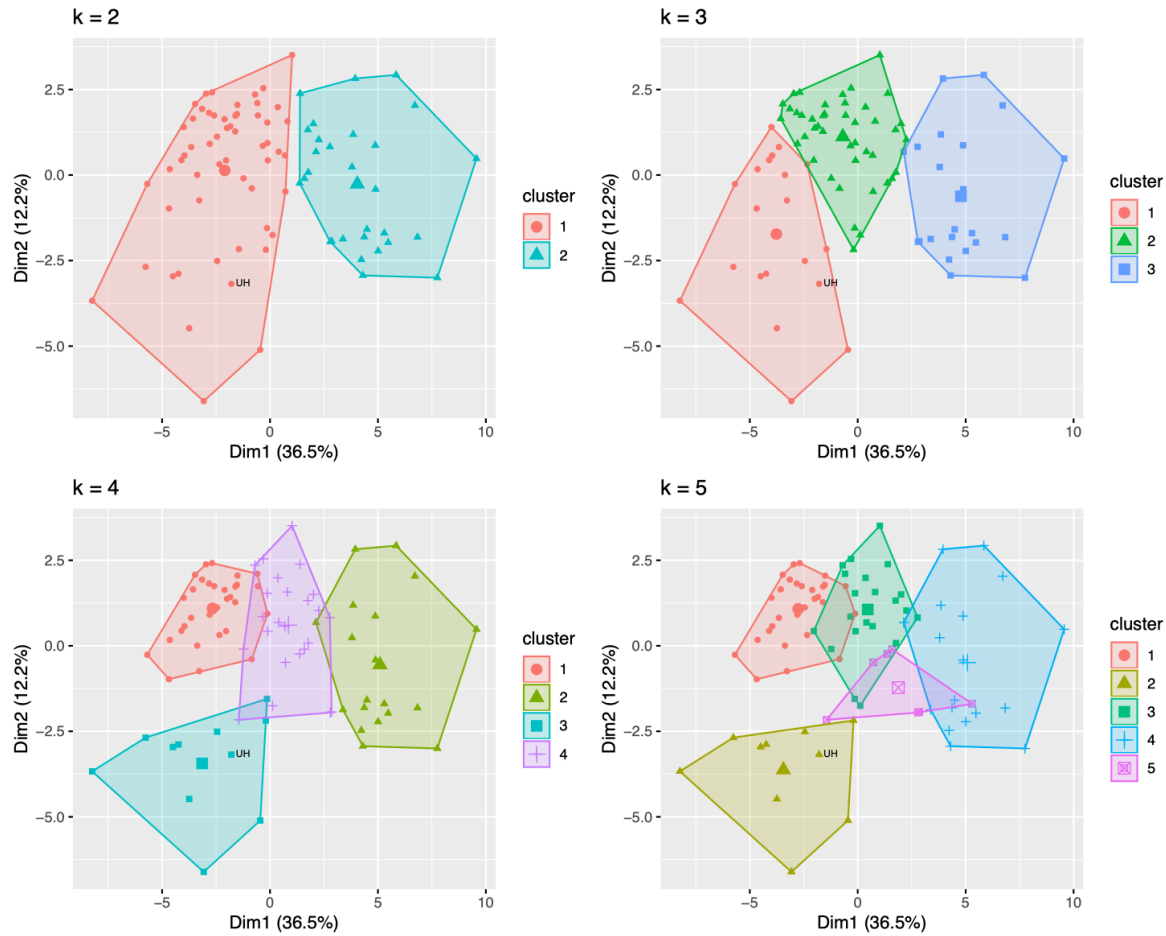
Figure 1: Euclidean Distance Matrix

Figure 2: displays four different specifications for group numbers from 2 to 5 groups. The x and y-axes are dimensions created by the clustering algorithm. They represent the amount of original information from the features used for clustering.

Each point represents one institution, with larger symbols indicating group centers. UH appears labeled in each grouping. The two-group specification demonstrates clear separation without overlap. The three, four, and five-group specifications all contain some overlap, indicating these groups are not sufficiently distinct from one another.

Despite some overlap in the five-group specification, this configuration proves optimal for determining peer institutions for two key reasons. First, the group containing UH and its peers shows no overlap with other groups, indicating this group is uniquely distinct from others in this specification. Second, and more significantly, this group in the five-group specification demonstrates the smallest internal variation among all specifications where UH appears as a member. This indicates the institution's group is not only uniquely different from others but contains the most similar members as measured by distances from the group center. Based on these considerations, the five-group specification was selected as optimal for peer group determination.

Examining the five-group specification in figure 3 and table 2 closely reveals nine institutional peers in the identified cluster.
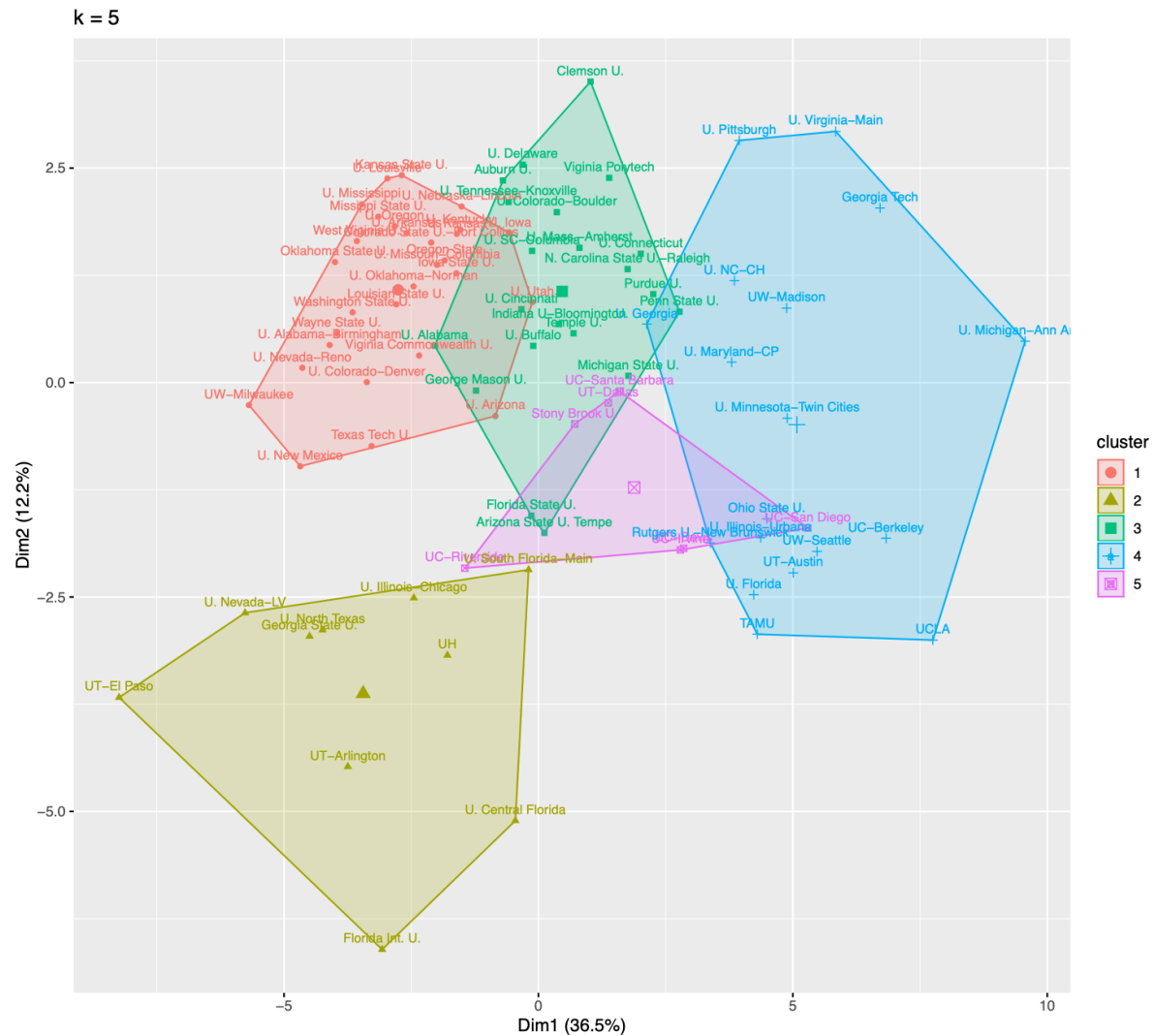

Figure 3: k=5 cluster by institution

Table 2: Institutional Peers of UH

Institution
Florida International University
Georgia State University
University of Central Florida
University of Illinois at Chicago
University of Nevada - Las Vegas
University of North Texas
University of South Florida - Main Campus
University of Texas at Arlington
University of Texas at El Paso

Table 3: Summary Statistics for Cluster 2 UH Peers.

Variable	Min	Max	Mean	UH	Std.Dev.
Percent Admitted	43%	100%	68.70%	62%	17.70%
Admissions Yield	26%	43%	36.2%	37%	5.6%
Undergraduate Enrollment	20,783	58,821	33,883	38,348	12,103
Graduate Enrollment	3,687	13,427	8,550	7,976	2,991.60
Percent Native American	0%	0%	0%	0%	0%
Percent African American	3%	38%	12.60%	10%	9.40%
Percent Hispanic	10%	80%	33.30%	32%	21.60%
Percent White	7%	48%	31%	25%	14.30%
Percent Two or More Races	1%	10%	4%	3%	2.50%
Percent Nonresident Alien	4%	12%	7.70%	8%	2.70%
Percent Asian/P.I.	1%	21%	10%	21%	6.60%
Percent Women	50%	63%	55.90%	50%	3.60%
Percent Undergrad Awarded Pell	32%	68%	47%	43%	11.20%
Students per Faculty	18	30	23	22	3.5
All Academic Ranked Faculty	780	2147	1333	1316	405.7
FTE 2017-18	20,214	56,819	36,695	38,366	10,341
Price of Attendance In-State On Campus	$22,548	$30,866	$24,872	$23,225	$2,464
Tuition in State	$4,478	$11,913	$7,017	$7,911	$2,247
Tuition Out-of-State	$15,473	$24,117	$20,262	$20,271	$2,872
Six-Year Graduation Rate, Fall 2018	39%	73%	56%	59%	11.20%
Total Degrees Awarded	4707	16076	10059.7	10134	3829.6
Core Revenues, Total Dollars	$452 M	$2,547 M	$1,072 M	$1,087 M	$571 M
Core Expenses, Total Dollars	$415 M	$1,946 M	$916 M	$940 M	$425 M
Tuition & Fees as % of Core Revenues	15%	42%	28.70%	36%	7.70%
Instructional Expenses as % Core Exp.	30%	50%	37.60%	32%	6.60%
Research Expenses as % Core Exp.	4%	29%	16.80%	20%	7.90%

Variable	Min	Max	Mean	UH	Std.Dev.
Student Services as % Core Exp.	4%	15%	7.70%	4%	4%
Endowment Assests per FTE	$2,844	$21,959	$8,804	$21,959	$5,759
SAT Read/Write 25th Percentile	470	590	542	570	37.4
SAT Read/Write 75th Percentile	570	670	632	650	28.2
SAT Math 25th Percentile	470	580	530	560	32.7
SAT Math 75th Percentile	560	670	628	660	32.9
ACT Composite Score 25th Percentile	17	25	21.4	22	2.5
ACT Composite Score 75th Percentile	22	30	26.5	28	2.3

Table 3 demonstrates how UH compares relative to peers across the 34 indicators. Notably, UH enrolls fewer White students and the maximum proportion of Asian/Pacific Islander students compared to the peer group. UH maintains equal gender distribution, which falls below the average. Although UH has larger undergraduate enrollment compared to peers, it maintains a more favorable student-to-faculty ratio.

The comprehensive cost of attendance for in-state students living on campus falls slightly below the mean. In-state tuition exceeds the average. However, out-of-state tuition aligns closely with the peer average.

Admitted students consistently achieve higher than average standardized test scores at both the 25th and 75th percentiles. Regarding student success, UH performs above average with a six-year graduation rate three percentage points higher than the peer group average. UH also awards more degrees than peers.

Financially, UH reports higher core revenues and core expenditures compared to peers. However, instructional and student services expenditures represent smaller proportions of core spending compared to the peer group average. Research represents a higher proportion of core expenses at UH compared to the peer group average.

Aspirational Peers

The nine institutions identified above represent comparable peers sharing similar institutional characteristics across multiple dimensions. University leadership can utilize these institutions to evaluate academic program quality and policies toward meeting institutional objectives. This section introduces an additional peer group type useful for identifying exceptional programs and policies producing substantial impacts on outcome measures.

Aspirational peers represent institutions with comparable institutional characteristics yet demonstrating significantly superior performance on key indicators. These are institutions with highly similar demographic and input characteristics such as racial/ethnic composition and standardized test scores that outperform UH in outcome measures including graduation rates or endowment values. To identify aspirational peers, the same clustering approach was applied, except this time institutions were grouped solely on demographic and non-performance indicators including admission rates, yield rates, standardized test percentile scores, undergraduate enrollment, racial/ethnic composition, gender distribution, financial aid recipients, full-time equivalent enrollment, faculty counts, and tuition rates. The visualization shows the optimal grouping specification for peer institutions based on these ten institutional characteristics.

The subsequent step involves identifying which institutions outperform the target institution in six-year graduation rates, graduate enrollment, and endowment values. The table presents these metrics for the seven identified aspirational peers.

Table 4: Aspirational Peers of UH

Institution	6-Yr Grad Rate	Grad Enrollment	Endowment	% Expenses as Research
University of Houston	59%	7,976	$21,959	20%
University of California-Davis	86%	7,449	$11,984	26%
University of California-Irvine	83%	6,296	$12,220	21%
University of California-Riverside	75%	3,341	$7,490	18%
University of California-San Diego	86%	7,602	$20,034	32%
University of California-Santa Barbara	82%	2,906	$8,389	22%
University of Texas at Dallas	72%	8,883	$21,332	17%
Stony Brook University	72%	8,734	$9,757	12%
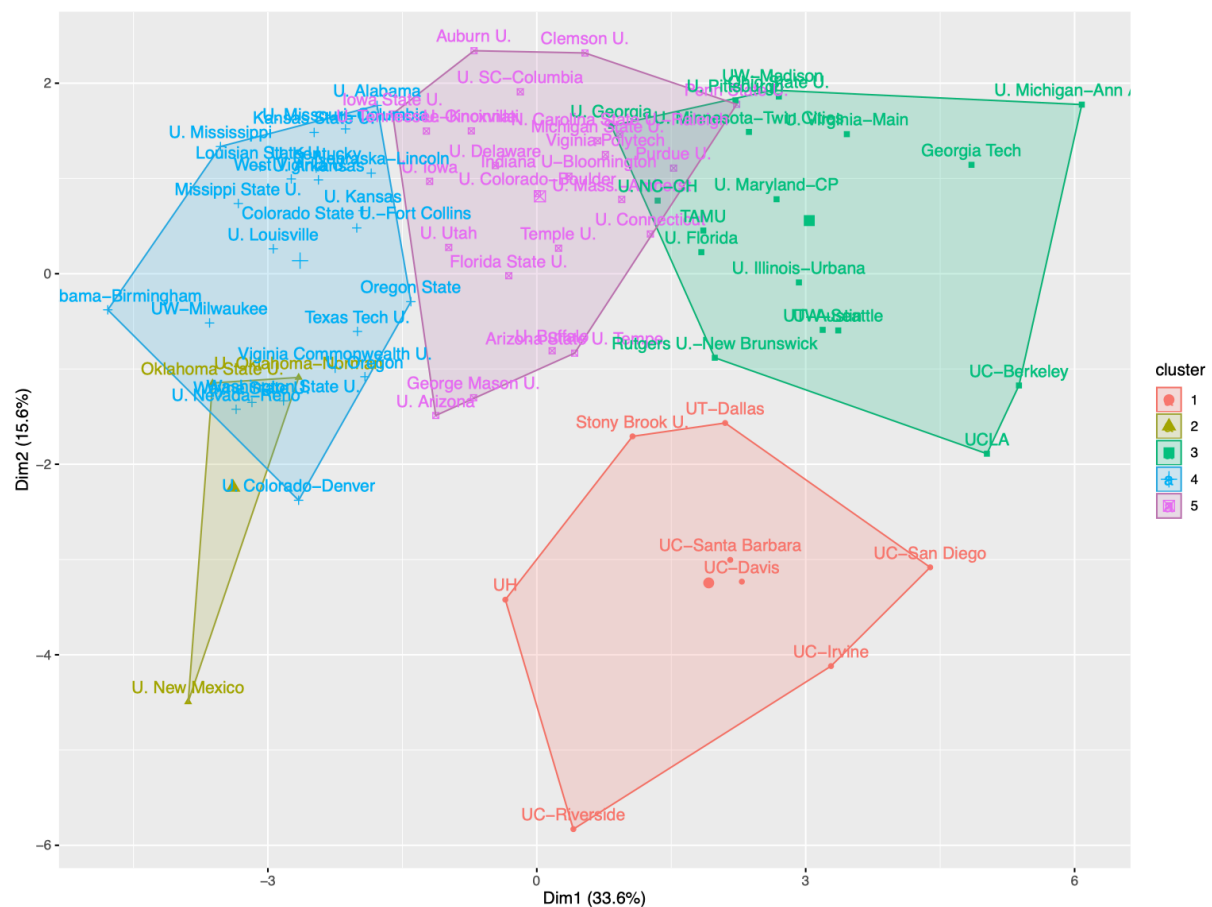
Figure 4: Aspirational Peer Cluster

In the Fall Address, the university president announced the institutional goal of achieving a 70% six-year graduation rate. Among the sample of 85 institutions, universities with graduation rates between sixty and seventy percent fall within one standard deviation above the institution's rate. All identified institutions in the cluster demonstrate graduation rates exceeding one standard deviation, ranging from 72% upward to

86%. Despite possessing highly similar institutional characteristics, these institutions qualify as aspirational peers through their superior performance in graduation rates.

Graduate enrollment represents another metric of institutional interest. Only two aspirational peers maintain larger graduate enrollment. Two aspirational peers fell outside the typical range compared to the institution.

UH maintains the largest endowment assets per student compared to institutions identified in the cluster, with all falling within the typical range. Regarding research expenditures as a proportion of core expenses, four of the seven cluster institutions reported greater research spending compared to the institution, with all within the typical range except one institution exceeding one standard deviation.

Across these performance indicators, each institution surpassed the target institution in six-year graduation rates. As a critical indicator of university performance, these seven institutions can be designated as aspirational peers despite mixed results for graduate enrollment, endowment assets, and research expenditure proportions. Another institutional objective involves membership in the Association of American Universities, an organization of leading comprehensive research universities distinguished by research breadth and quality and graduate education excellence. Membership is by invitation only. Although not included in the performance measure dataset, five of the seven cluster institutions hold membership, making them strong aspirational peers.

Conclusion

This analysis has identified nine peer institutions with characteristics comparable to the target institution using official federal data. Additionally, seven aspirational peers were identified with similar institutional characteristics but superior performance on key indicators. Cluster analysis was employed to statistically identify both peer groups for benchmarking purposes. These institutions can be examined to evaluate university policies and programs toward developing strategic plans for advancing institutional success. This analysis provides a systematic, evidence-based approach toward updating various peer groups previously identified historically. Historically, nineteen different peer groups have been identified by various departments and administrators encompassing one hundred nine different peer institutions. The cluster analysis results offer an unbiased, manageable list of peer institutions that can be programmatically updated annually as institutions change dynamically year-to-year. This enables continued momentum toward becoming a highly-ranked national university and joining the top quartile of universities in degree completion within six years or less.

Table 5: Mean values for k=5 clusters

Variable	1	2	3	4	5
Percent Admitted	79.50%	68.70%	64.30%	41.70%	42%
Admissions Yield	32%	36.20%	28.30%	37.20%	22.30%
Undergraduate Enrollment	22046	33883.4	28515.7	31635.9	24540.6
Graduate Enrollment	6165.8	8550	8180.1	13092.1	6458.7
Percent Native American	0.70%	0%	0%	0%	0%
Percent African American	6.60%	12.60%	6.40%	5.20%	3.40%
Percent Hispanic	10.60%	33.30%	7.40%	9.90%	21%
Percent White	64.20%	31%	62.90%	50.70%	25%
Percent Two or More Races	4.30%	4%	3.30%	3.90%	4.10%
Percent Nonresident Alien	6.50%	7.70%	10.60%	12.90%	17%
Percent Asian/P.I.	4.80%	10%	6.60%	14.30%	26.10%
Percent Women	52.70%	55.90%	50%	51.20%	51.30%
Percent Undergrad Awarded Pell	27.90%	47%	20.90%	19.50%	35.90%
Student-Faculty Ratio	17.3	23	16.4	17.1	20.7
All Academic Ranked Faculty	1670.7	1333	1811.5	3058.4	1700.1
FTE 2017-18	25736.9	36694.7	34512.5	43502.6	29953.1
Price of Attendance In-State On Campus	26298	24872	28996	29172	32793
Tuition in State	8649	7017	11144	10983	11016
Tuition Out-of-State	24448	20262	30477	34480	37655
Six-Year Graduation Rate, Fall 2018	62.70%	56%	77.50%	86.30%	79.70%
Total Degrees Awarded	6527.8	10059.7	9160.1	11967.6	8060.1
Core Revenues, Total Dollars	$1039 M	$1072 M	$1312 M	$2919 M	$1655 M
Core Expenses, Total Dollars	$995 M	$916 M	$1225 M	$2591 M	$1547 M
Tuition & Fees as % of Core Revenues	30.70%	28.70%	39.10%	23.50%	29.70%
Instructional Expenses as % Core Exp.	34.60%	37.60%	41.90%	32.50%	38.90%
Research Expenses as % Core Exp.	21.70%	16.80%	18.90%	27.60%	21.10%
Student Services as % Core Exp.	4.60%	7.70%	5%	4.10%	7.70%
Endowment Assests per FTE	30759	8804	28231	91521	13029
SAT Read/Write 25th Percentile	540.7	542	585	627.8	594.3
SAT Read/Write 75th Percentile	643.2	632	668.5	711.1	681.4
SAT Math 25th Percentile	529.4	530	579.5	639.4	605.7
SAT Math 75th Percentile	642	628	689	755.6	737.1
ACT Composite Score 25th Percentile	21.7	21.4	24.8	27.9	25.3
ACT Composite Score 75th Percentile	27.9	26.5	30.4	32.7	31.9

Table 6: Institution by Cluster, 1 and 2.

Cluster 1	Cluster 2
Colorado State University-Fort Collins	Florida International University
Iowa State University	Georgia State University
Kansas State University	The University of Texas at Arlington
Louisiana State University and Agricultural & Mechanical College	The University of Texas at El Paso
Mississippi State University	University of Central Florida
Oklahoma State University-Main Campus	University of Houston
Oregon State University	University of Illinois at Chicago
Texas Tech University	University of Nevada-Las Vegas
University of Alabama at Birmingham	University of North Texas
University of Arizona	University of South Florida-Main Campus
University of Arkansas	
University of Colorado Denver/Anschutz Medical Campus	
University of Iowa	
University of Kansas	
University of Kentucky	
University of Louisville	
University of Mississippi	
University of Missouri-Columbia	
University of Nebraska-Lincoln	
University of Nevada-Reno	
University of New Mexico-Main Campus	
University of Oklahoma-Norman Campus	
University of Oregon	
University of Utah	
University of Wisconsin-Milwaukee	
Virginia Commonwealth University	
Washington State University	
Wayne State University	
West Virginia University	

Table 7: Institution by Cluster 3 and 4

Cluster 3	Cluster 4
Arizona State University-Tempe	Georgia Institute of Technology-Main Campus
Auburn University	Ohio State University-Main Campus
Clemson University	Rutgers University-New Brunswick
Florida State University	Texas A & M University-College Station
George Mason University	The University of Texas at Austin
Indiana University-Bloomington	University of California-Berkeley
Michigan State University	University of California-Los Angeles
North Carolina State University at Raleigh	University of Florida
Pennsylvania State University-Main Campus	University of Georgia
Purdue University-Main Campus	University of Illinois at Urbana-Champaign
Temple University	University of Maryland-College Park
The University of Alabama	University of Michigan-Ann Arbor
The University of Tennessee-Knoxville	University of Minnesota-Twin Cities
University at Buffalo	University of North Carolina at Chapel Hill
University of Cincinnati-Main Campus	University of Pittsburgh-Pittsburgh Campus
University of Colorado Boulder	University of Virginia-Main Campus
University of Connecticut	University of Washington-Seattle Campus
University of Delaware	University of Wisconsin-Madison
University of Massachusetts-Amherst	
University of South Carolina-Columbia	
Virginia Polytechnic Institute and State University	

Table 8: Institution by Cluster 5

Cluster 5
Stony Brook University
The University of Texas at Dallas
University of California-Davis
University of California-Irvine
University of California-Riverside
University of California-San Diego
University of California-Santa Barbara

Table 9: Top 10 Peers by Distance Measures

Euclidean	Pearson Correlation	Manhattan
Texas Tech University	Florida Int. University	Texas Tech University
University of South Florida-Main	University of Central Florida	University of South Florida-Main
University of North Texas	UT-El Paso	Florida Int. University
Florida Int. University	Texas Tech University	University of North Texas
University of Arizona	UT-Arlington	Iowa State University
Florida State University	University North Texas	Louisian State University
Iowa State University	Georgia State University	UT-Arlington
Louisiana State University	University of South Florida-Main	University of Arizona
Arizona State University Tempe	University of Nevada-LV	Florida State University
George Mason University	Arizona State University Tempe	Oregon State

Table 10: Top 10 Peers by Distance Measures

Spearman Correlation	Kendall Correlation
Florida Int. University	Florida Int. University
Georgia State University	Georgia State University
UT-Arlington	University of North Texas
University of North Texas	UT-Arlington
Texas Tech University	Texas Tech University
University of Central Florida	University of Central Florida
UT-El Paso	UT-El Paso
University of South Florida-Main	University South Florida-Main
University of Nevada-LV	University of Nevada-LV
University of New Mexico	Arizona State University Tempe (tie)
	University of New Mexico (tie)

Table 11: Descriptive statistics for aspirational peer group

Variable	Min	Max	Mean	UH	Std.Dev.
Percent Admitted	29%	69%	44.50%	62%	15%
Admissions Yield	17%	37%	24.13%	37%	7.86%
SAT Read 25th Percentile	560	620	591.25	570	22.32
SAT Math 25th Percentile	550	640	600	560	32.95
Undergraduate Enrollment	17,522	38,348	26,266.50	38,348	7,109.98
Percent Native American	0%	0%	0%	0%	0%
Percent African American	2%	10%	4.25%	10%	2.96%
Percent Hispanic	11%	37%	22.38%	32%	8.73%
Percent White	14%	36%	25%	25%	7.63%
Percent Two or More Races	2%	7%	4%	3%	1.51%
Percent Nonresident Alien	8%	23%	15.88%	8%	5.49%
Percent Asian/P.I.	17%	33%	25.50%	21%	5.88%
Percent Women	43%	59%	51.12%	50%	4.58%
Percent Undergrad Awarded Pell	27%	50%	36.75%	43%	7.15%
FTE 2017-18	22,980	38,366	31,004.75	38,366	6,872.23
All Academic Ranked Faculty	873	2,564	1,652.13	1,316	686.56
Tuition In-State	$6,870	$13,034	$10,628.13	$7,911	$2,091.02
Tuition Out-of-State	$20,271	$40,434	$35,482.13	$20,271	$8,243.24