UNIVERSITY CLUSTERING BASED ON ACADEMIC PROFILE AND SURROUNDING VENUES

Shaleen Parikh

JUNE 22, 2020

Table of Contents

Introduction

In the United States, millions of high school students are faced with the task of applying to universities and may experience difficulty as they attempt to figure out which universities are a good match for them. With thousands of universities to choose from, it is hard to find a proper match. As there is also a fee for each application, it is also necessary for many students to limit the number of applications that they submit, meaning their choices of application should be correct for them.

In order to narrow down their choices, students may want to know about the available facilities near the universities and the academic profile of the universities. This information can help them find a good academic as well as environmental match.

Using the Foursquare API and other data sources to cluster universities, this project strives to help high school students and other future university students discover universities that have strong similarities based on their respective surrounding venues as well as their academic attributes. If students are able to find a proper match in both the academics and location of a university, they will likely have a better and happier college experience.

Together, the clusters will allow high schoolers and other university applicants to find personally suitable colleges based on both location and academic status.

Data

The IPEDS Data Center, https://nces.ed.gov/ipeds/use-the-data, provides more than 250 features for over 7,000 universities in a CSV format. In order to focus on the relevant information, I removed more than 200 of the features and dropped all universities with missing key information. I stored the resulting data in a Python pandas dataframe. The remaining features are university name, latitude, longitude, street address, city, average 25th and 75th percentile SAT math and reading scores, average 25th and 75th percentile ACT composite scores, average net price of attendance, total undergraduate applicants, total undergraduate admissions, and student-to-faculty ratio. There are a total of 1214 universities in the dataframe.

Another data source that will be used is the Foursquare API, which allows developers to enter search queries or location information to find surrounding places of interest. The latitude and longitude features of the universities from the dataframe will be passed to the API in order to receive JSON-formatted information about the surroundings of the universities, including the nearby venues. This will require the use of the explore keyword in order to get a prespecified number of venues within a prespecified radius of the universities.

Together, these data sources will allow for two types of clusters to be formed to characterize each university: one with the academic characteristics, which come from the IPEDS Data Center dataset, and one with the characteristics of surrounding venues, which come from the Foursquare API.

Methodology

In order to provide a more significant academic statistic than total undergraduate applicants and admissions, I created a new feature, acceptance rate, by dividing the number of admitted students for each university by the number of students that applied for each university. Acceptance rate is an indicator of selectivity of a university and a lower acceptance rate typically corresponds to a higher academic ranking for a university.

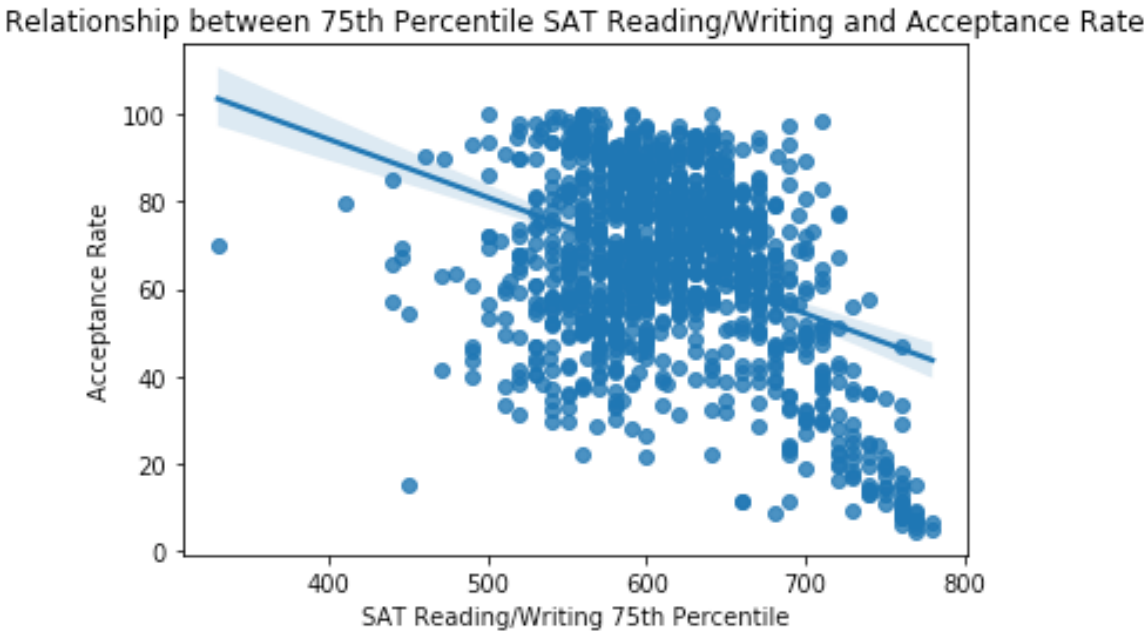

Figure 1

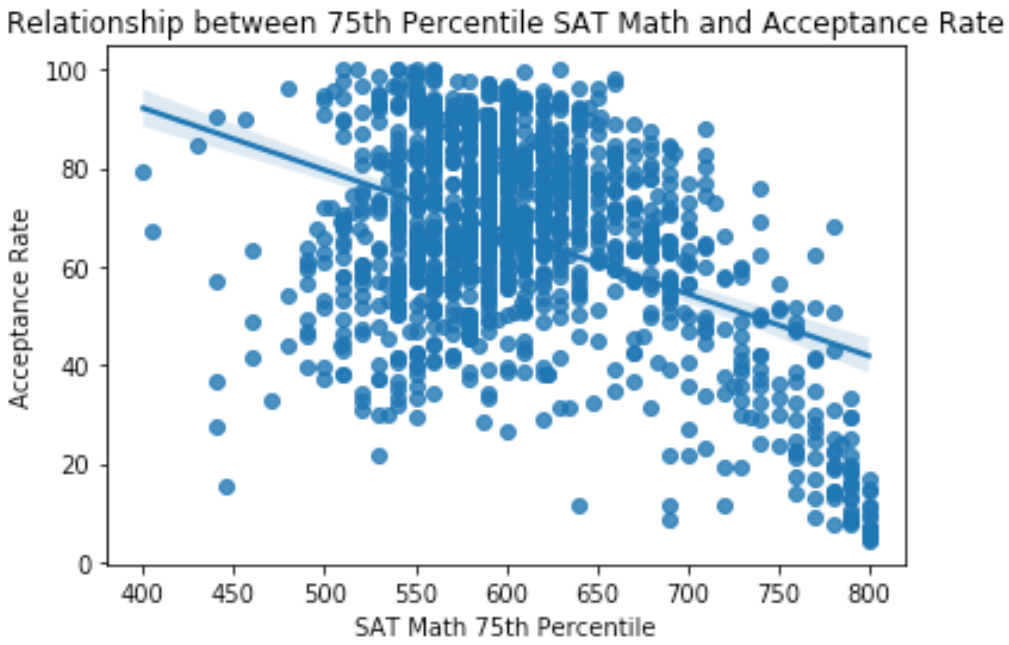

Figure 2

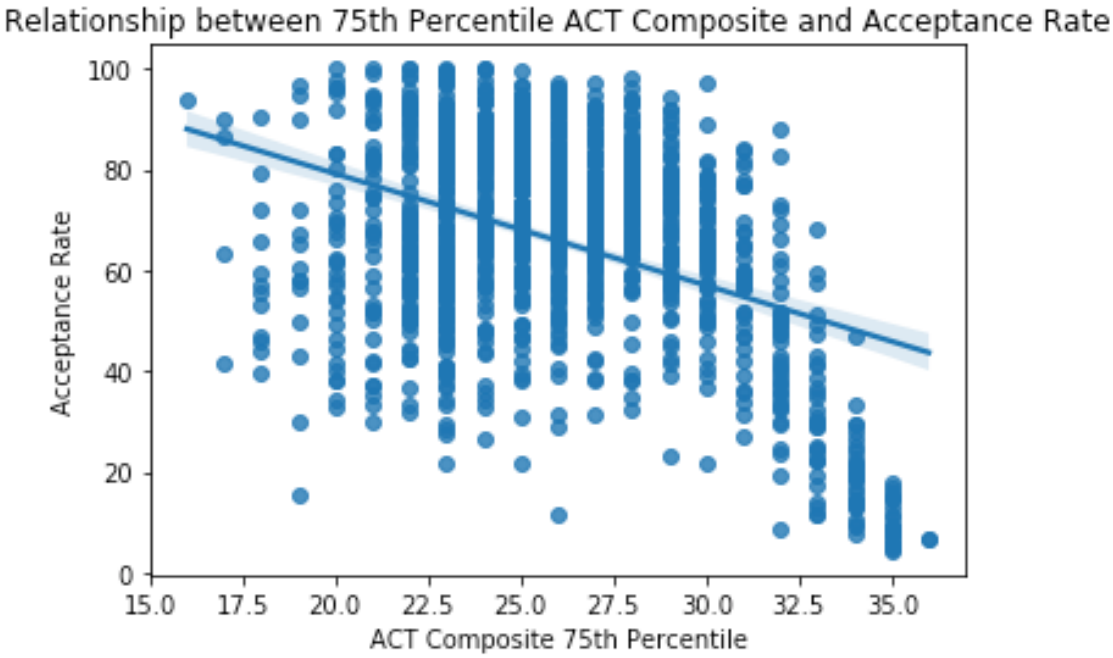

Figure 3

Other attributes correspond to university rankings as well. One of these is standardized test scores, including the commonly taken SAT and ACT, which test students' proficiency in mathematics, reading, writing, and science. As shown in Figures 1, 2, and 3, 75th percentile SAT and ACT scores clearly have a negative correlation to acceptance rate. Another is student-to-faculty ratio, which indicates the size of classes. A lower student-to-faculty ratio is generally desirable in top universities.

As acceptance rate, test scores of incoming students, and student-to-faculty ratio are commonly used to measure the selectivity and prestige of universities, I first used these attributes from the dataframe to cluster the universities on the basis of academics. In order to do so, I used the K-Means clustering algorithm. To find the optimal value of K, I implemented the elbow method by plotting values of K against their respective within-cluster sum of squares, or inertia.

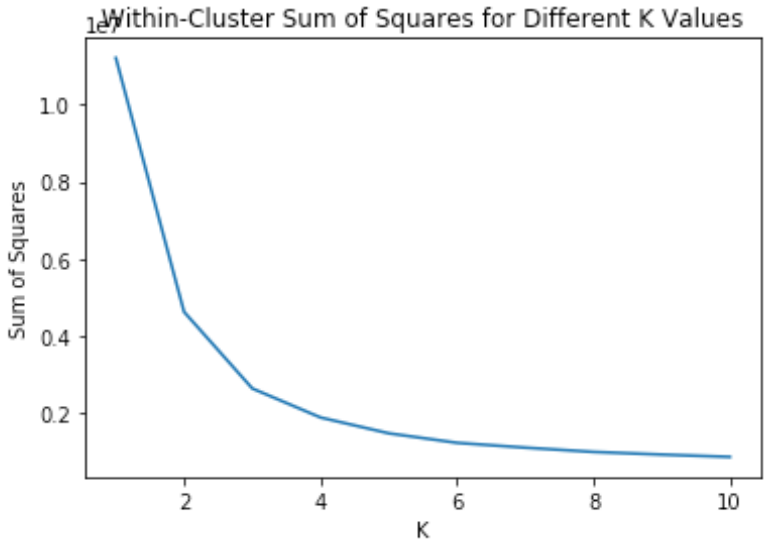

Figure 4

Since the rate of decline of the within-cluster sum of squares significantly decreased at K=4, I used 4 clusters to academically profile the universities.

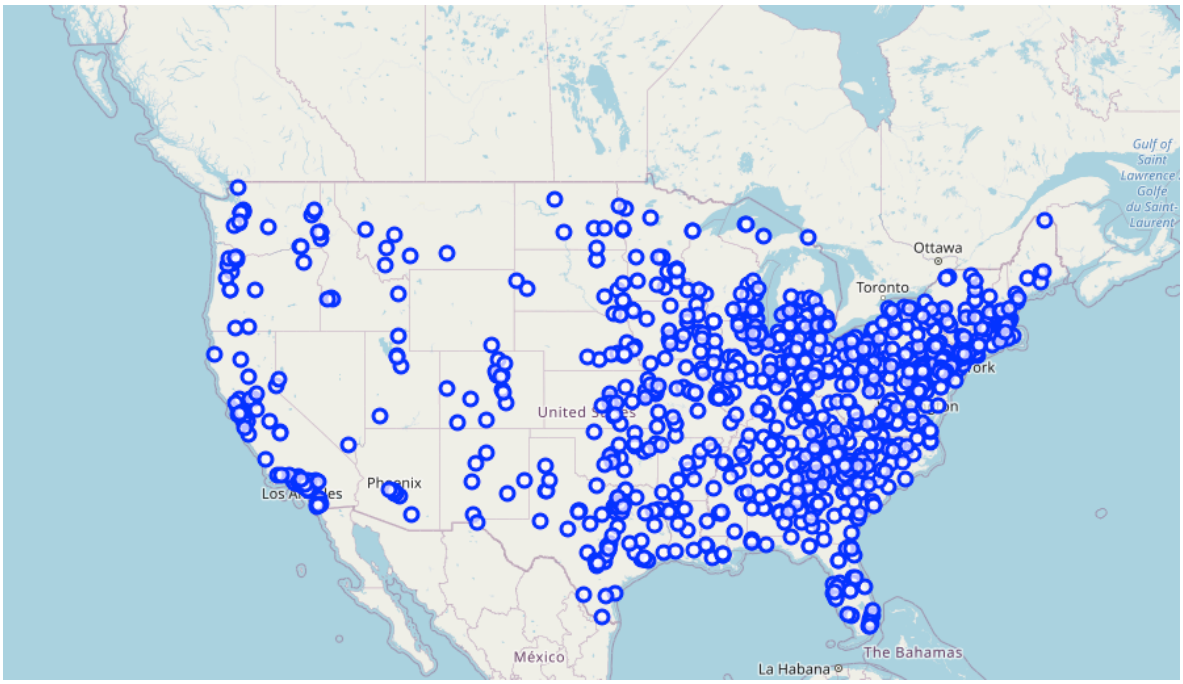

Figure 5

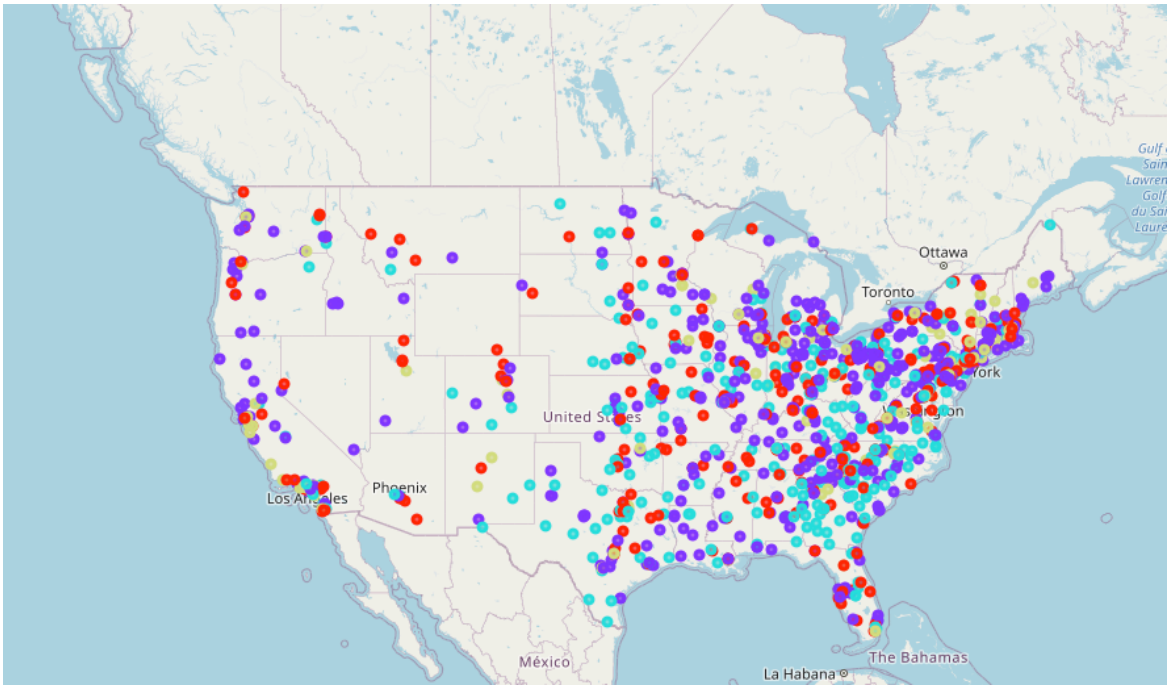

Figure 6

Academic Cluster	3	2	1	0
Color	Green	Blue	Purple	Red

Figure 7

Figure 5 shows the unclustered universities, and Figures 6 and 7 show the universities clustered by academics.

Next, I clustered the universities based on the categories of their nearby venues. To do so, I used the Foursquare API to receive the top 50 venues within 8000 meters (~5 miles) of each university. The venue categories were represented binarily and aggregated for each university. Then, the mean frequency for each venue category was calculated for each university. For example, Hartwick College has 3 times as many cafés than burger joints and BBQ joints within a 5-mile radius.

University	Art Museum	BBQ Joint	Beach	Burger Joint	Café	Chinese Restaurant
Hardin-Simmons University	0.00	0.00	0.0	0.04	0.000000	0.04
Harding University	0.00	0.04	0.0	0.00	0.020000	0.00
Hartwick College	0.00	0.02	0.0	0.02	0.060000	0.00
Harvard University	0.02	0.00	0.0	0.00	0.060000	0.00
Harvey Mudd College	0.00	0.00	0.0	0.00	0.020000	0.00
Haskell Indian Nations University	0.00	0.00	0.0	0.06	0.000000	0.00
Hastings College	0.00	0.00	0.0	0.02	0.020000	0.02
Haverford College	0.00	0.00	0.0	0.00	0.020000	0.00
Hawaii Pacific University	0.02	0.00	0.0	0.00	0.020000	0.04
Heidelberg University	0.00	0.00	0.0	0.00	0.020833	0.00

Figure 8

K-Means clustering was done to cluster the universities based on the frequency of each venue category, and the elbow method was again used to determine the optimal value of K. Ultimately, 5 venue clusters were formed.

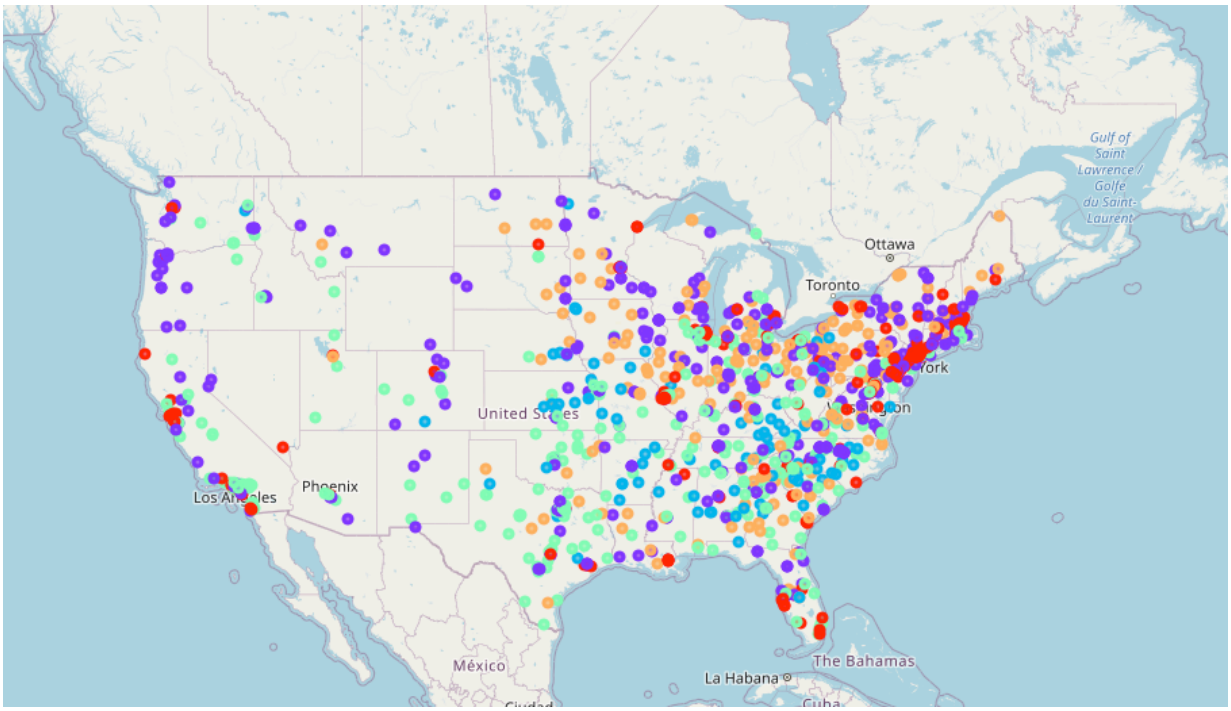

Figure 9

Venue Cluster	4	3	2	1	0
Color	Orange	Green	Blue	Purple	Red

Figure 10

Results

In order to assign key names for the academic clusters, I grouped the dataframe by academic cluster and examined the mean value of certain columns.

Academic Cluster	SAT Reading/Writing 25th Percentile	SAT Reading/Writing 75th Percentile	SAT Math 25th Percentile	SAT Math 75th Percentile	ACT Composite 25th Percentile	ACT Composite 75th Percentile	Student-to-Faculty Ratio	Acceptance Rate
0	556.886121	655.711744	543.807829	656.676157	22.761566	28.473310	14.003559	68.556511
1	501.500000	604.371849	494.199580	592.733193	19.598739	25.348739	14.214286	72.386219
2	452.962733	547.186335	445.906832	534.450311	17.024845	22.142857	14.090062	67.490585
3	646.022222	723.185185	658.748148	758.370370	29.074074	33.000000	10.940741	31.968272

Figure 11

It is clear in Figure 11 that universities in academic cluster 3 have a much lower acceptance rate and higher test scores than universities in academic clusters 0, 1, and 2, as well as a lower student-to-faculty ratio. This corresponds to the attributes of higher-ranked universities. The other three clusters are similar to each other in terms of their higher acceptance rates and student-to-faculty ratios, but they differ in standardized test scores. Universities in academic cluster 0 have noticeably higher 25th and 75th percentile SAT and ACT scores than those in academic clusters 1 and 2, but their mean acceptance rate is less than that of academic cluster 1 and greater than that of academic cluster 2. Universities in academic cluster 1 have the highest acceptance rate and the second-lowest test scores, and universities in academic cluster 2 have a lower acceptance rate and the lowest test scores. Figure 12 shows the academic profile names that were given to each academic cluster.

Academic Cluster	3	2	1	0
Academic Profile	Most Selective/ Highest Scores	Less Selective/ Lowest Scores	Least Selective/ Lower Scores	Less Selective/ Medium Scores

Figure 12

In order to assign venue profiles, the top 15 venue categories based on mean frequency were sorted for each university and examined for each cluster. For example:

University	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Wheeling Jesuit University	Café	American Restaurant	Italian Restaurant	Pizza Place	Brewery	Mexican Restaurant	Coffee Shop	Golf Course	Bar	Ice Cream Shop
William Peace University	Coffee Shop	Brewery	Bakery	Burger Joint	Cocktail Bar	Science Museum	Southern / Soul Food Restaurant	Sushi Restaurant	American Restaurant	Beer Store
Williams College	American Restaurant	Art Museum	Pizza Place	Sandwich Place	Sports Bar	New American Restaurant	Hotel	Coffee Shop	Grocery Store	College Bookstore
Winona State University	Coffee Shop	Convenience Store	Pizza Place	Sandwich Place	Bar	Ice Cream Shop	Café	Brewery	Grocery Store	Movie Theater
Winston-Salem State University	Brewery	Bar	Grocery Store	American Restaurant	Mediterranean Restaurant	Coffee Shop	Park	Bakery	BBQ Joint	Pizza Place
Wisconsin Lutheran College	Coffee Shop	Zoo Exhibit	Italian Restaurant	Bar	Café	Beer Garden	Ice Cream Shop	Park	Steakhouse	Gym
Xavier University	Zoo Exhibit	Ice Cream Shop	Pizza Place	American Restaurant	Coffee Shop	Brewery	Chinese Restaurant	Bakery	Burger Joint	Hot Dog Joint
Yale University	Pizza Place	Indian Restaurant	Sandwich Place	Sushi Restaurant	American Restaurant	Bar	Theater	Burger Joint	Restaurant	Park

Figure 13

To find the most common venue categories within a cluster, all the values in a given cluster's top eight categories for all combined universities were concatenated into a list, and the mode of venue category was calculated for each cluster. The mode for venue cluster 0 was park, the mode for venue cluster 1 was coffee shop, the mode for venue cluster 2 was fast food restaurant, the

mode for venue cluster 3 was Mexican restaurant, and the mode for venue cluster 4 was pizza place. Venue cluster 3 also had an abundance of other types of restaurants, such as Italian, Chinese, and Indian, and venue cluster 4 had a large number of bakeries and ice cream shops. Additionally, cluster 1 had a far greater percentage of coffee shops and cafés combined than cluster 0, and cluster 0 had a much larger percentage of trails, zoo exhibits, zoos, and trails combined than cluster 1. Hence, the venue profiles in Figure 14 were assigned to the venue clusters.

Venue Cluster	4	3	2	1	0
Venue Profile	Pizza and Dessert	Variety of Cuisines	Fast Food	Coffee and Casual	Outdoor Attractions

Figure 14

Discussion

The relationship between the academic cluster and venue cluster was examined. While there is no obvious relationship between the two, some interesting connections were found. 39.4% of the universities in academic cluster 3 have a venue profile of “Outdoor Attractions”, and 37.9% have a venue profile of “Coffee and Casual.” 27.3% of the universities in academic cluster 2 have a “Variety of Cuisines” venue profile, and 24.8% have a venue profile of “Pizza and Dessert.” 31.7% of the universities in academic cluster 1 have a venue profile of “Coffee and Casual” and 26.7% have a venue profile of “Variety of Cuisines.” Also, 41.9% of universities in academic cluster 0 have a venue profile of “Coffee and Casual.”

The inverse relationship was also observed. 39.3% of universities with a venue cluster of 4 have an academic profile of "Less Selective/Lowest Scores," and 38.8% have an academic profile of "Least Selective/Lower Scores." 43.8% of universities with a venue cluster of 3 have an academic profile of "Least Selective/Lower Scores," and 30.2% have an academic profile of "Less Selective/Lowest Scores." 54.9 % of universities with a venue cluster of 2 have an academic profile of "Less Selective/Lowest Scores," and 36.3% have an academic profile of "Least Selective/Lower Scores." 39.8% of universities with a venue cluster of 1 have an academic profile of "Least Selective/Lower Scores," and 30.8% have an academic profile of "Less Selective/Medium Scores." 34.5% of universities with a venue cluster of 0 have an academic profile of "Least Selective/Lower Scores," and 26.5% have an academic profile of "Less Selective/Medium Scores."

Some of the relevance of these percentages is lost when considering that there are far more universities that have an academic cluster of 0, 1, and 3 (238, 367, and 281, respectively) compared to academic clusters 2 and 4 (91 and 196, respectively).

Similarly, there are far more universities with venue clusters of 0, 1, and 2 (270, 460, and 311, respectively) compared to venue clusters 3 and 4 (132 and 270, respectively).

Logically, the venue profile and academic profile should not have a substantial relationship with one another, as they are largely independent attributes of a university. The venue profile has more to do with the surrounding city government and local businesses, and the academic profile has more to do with the performance of a university's students and the resources available to them.

Therefore, rather than interpreting the two as causations of each other, it is best to use the formed clusters to narrow down university choices as combined criteria. For example, one may wish to attend a highly university with many outdoor attractions. There is a good chance that a match can be found by finding universities under academic cluster 3 and venue cluster 0.

Conclusion

In this analysis, I grouped over 1,000 American universities based on both the academic characteristics of their students and resources and their nearby venues. By using standardized test scores, student-to-faculty ratio, and acceptance rate as indicators of a university's academic ranking, I used the K-Means algorithm to assign the universities to 4 academic clusters and assigned the universities in these clusters academic profiles based on their mean characteristics. On the other hand, I used the top 50 venues within a 5-mile radius to determine the most frequently occurring venue categories for each university and used this information in another implementation of the K-Means algorithm to group the universities into 5 venue clusters. The venue profiles were determined based on the mode of the combined venue categories for each cluster. Prospective United States university students can benefit from the resulting dataset by examining the universities that overlap in their preferred academic and venue clusters.