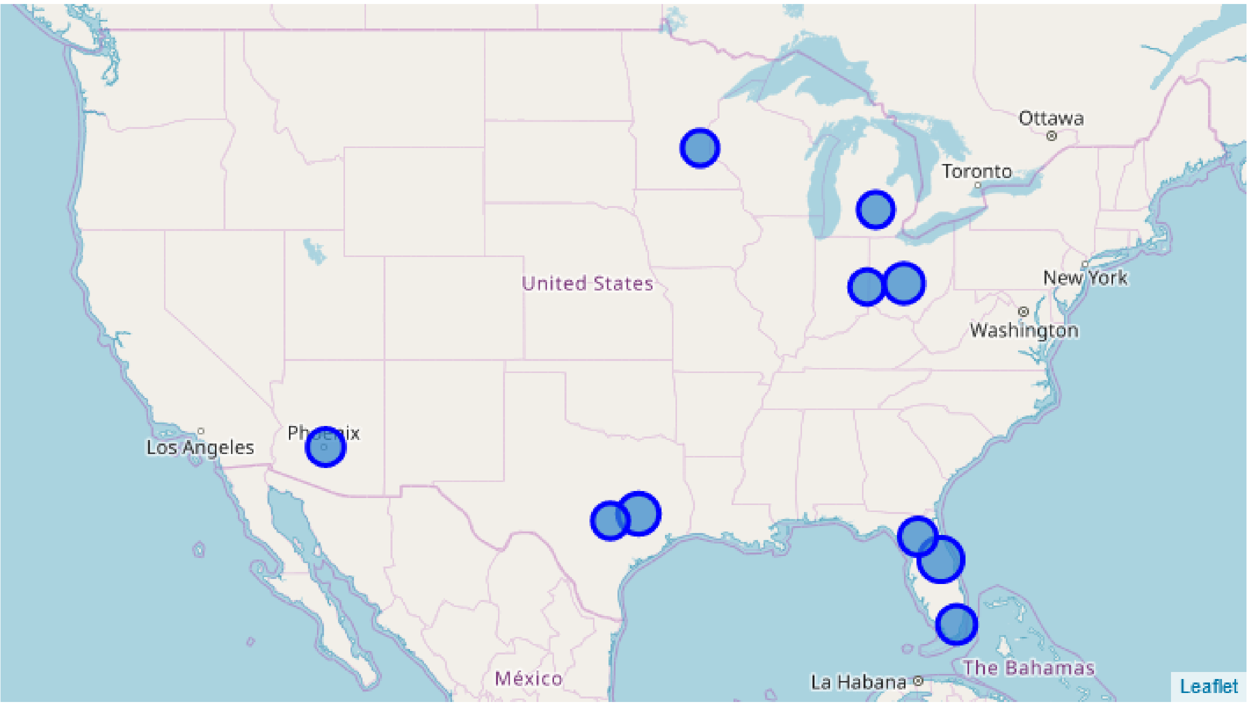
Applied Data Science Capstone

The Battle of Universities (Week 2)



Author: Gabriel Sánchez

Coursera: Applied Data Science Capstone

Date: 07-May-2020

URL: https://github.com/RieldaLearn/Coursera_Capstone

Notebook URL: https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/d259c49f-b8aa-4715-8c1a-3f0626c5f1c9/view?access_token=9528bd5ead92bcf6d98f8a546622bc91fa152d37947b18ef5369e5775f7210ea

Table of Contents

Introduction

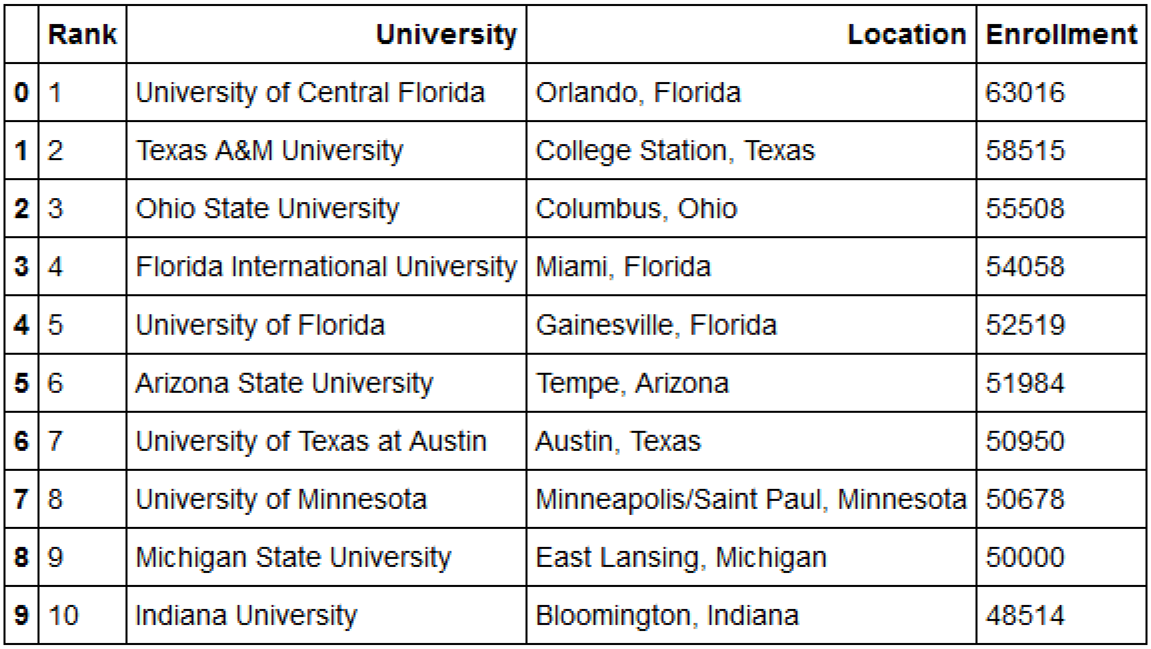
This project is submitted for consideration of the Coursera Applied Data Science Capstone course. For this project we will analyze the Foursquare venue data associated with the top ten universities in the United States based on enrollment.

Problem: What makes universities in the United States similar? Can venue data be used to determine similarities?

We will try to determine if venue similarities can help us discriminate venue categories for the top ten universities in the United States and group these into clusters using K-Means classification.

Data

I have searched for and identified a world atlas webpage that identifies the enrollment by academic year 2015-16 for the top ten universities in the United States.



	Rank	University	Location	Enrollment
0	1	University of Central Florida	Orlando, Florida	63016
1	2	Texas A&M University	College Station, Texas	58515
2	3	Ohio State University	Columbus, Ohio	55508
3	4	Florida International University	Miami, Florida	54058
4	5	University of Florida	Gainesville, Florida	52519
5	6	Arizona State University	Tempe, Arizona	51984
6	7	University of Texas at Austin	Austin, Texas	50950
7	8	University of Minnesota	Minneapolis/Saint Paul, Minnesota	50678
8	9	Michigan State University	East Lansing, Michigan	50000
9	10	Indiana University	Bloomington, Indiana	48514

Source: https://www.worldatlas.com/articles/largest-universities-in-the-united-states.html

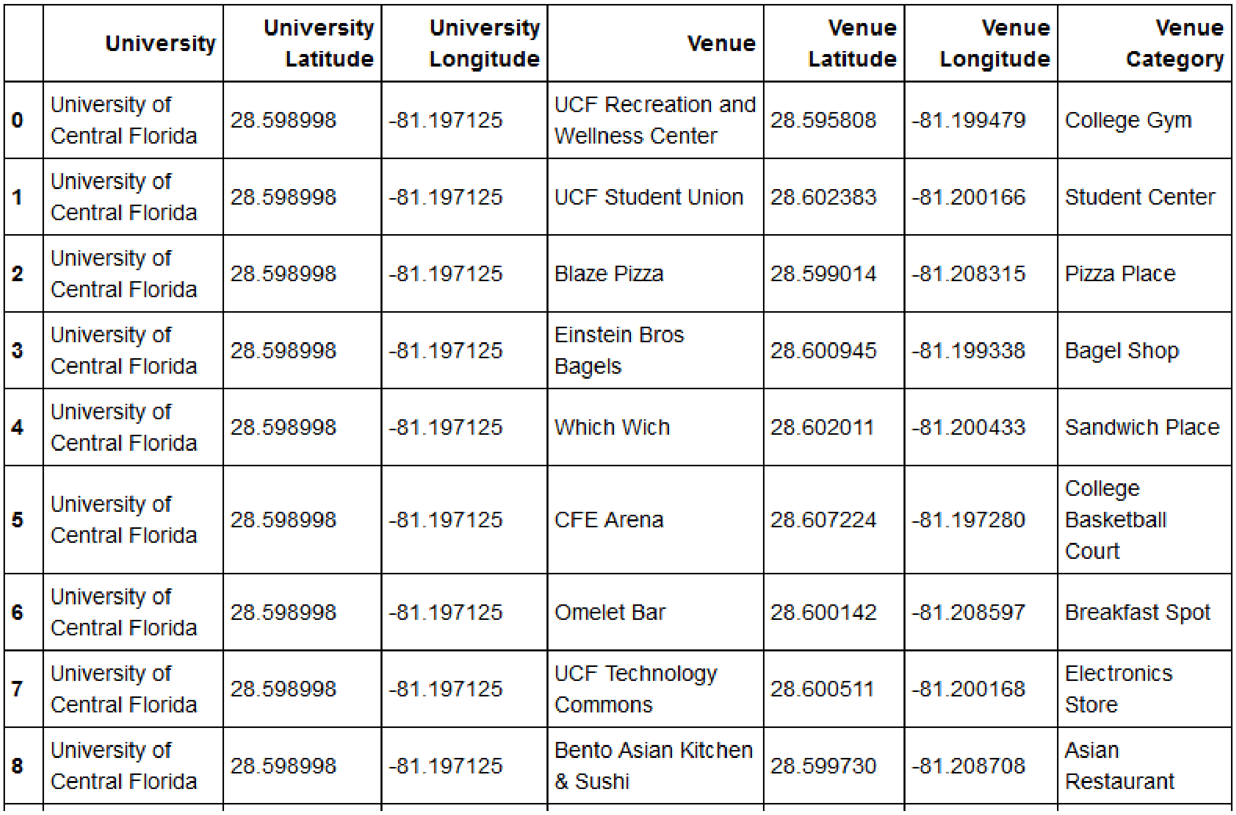
Using this source data on universities we will then capture and merge geolocation latitude and longitude data from the geopy geocoders Nominatim library.

Foursquare API will then be used to capture any venue and venue category data within one mile (1609 meters), defined as “walking distance”, from the center of each university location.

```
Example of Foursquare data elements: 'Venue', 'Venue Latitude', 'Venue
Longitude' and 'Venue Category'.

Example of dataframe population:
nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in
venue_list])
    nearby_venues.columns = ['University',
                  'University Latitude',
                  'University Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']
```

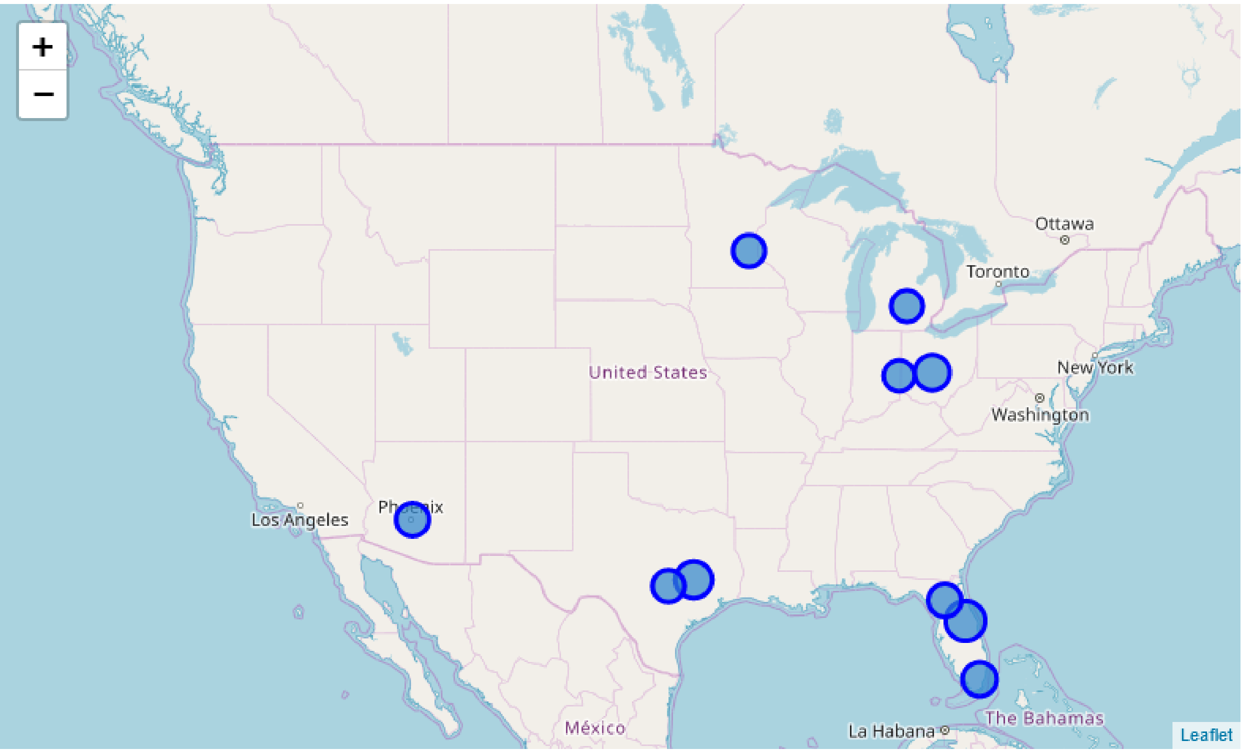
The data will be transformed into multiple dataframe objects required to analyze and map the results.



	University	University Latitude	University Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	University of Central Florida	28.598998	-81.197125	UCF Recreation and Wellness Center	28.595808	-81.199479	College Gym
1	University of Central Florida	28.598998	-81.197125	UCF Student Union	28.602383	-81.200166	Student Center
2	University of Central Florida	28.598998	-81.197125	Blaze Pizza	28.599014	-81.208315	Pizza Place
3	University of Central Florida	28.598998	-81.197125	Einstein Bros Bagels	28.600945	-81.199338	Bagel Shop
4	University of Central Florida	28.598998	-81.197125	Which Wich	28.602011	-81.200433	Sandwich Place
5	University of Central Florida	28.598998	-81.197125	CFE Arena	28.607224	-81.197280	College Basketball Court
6	University of Central Florida	28.598998	-81.197125	Omelet Bar	28.600142	-81.208597	Breakfast Spot
7	University of Central Florida	28.598998	-81.197125	UCF Technology Commons	28.600511	-81.200168	Electronics Store
8	University of Central Florida	28.598998	-81.197125	Bento Asian Kitchen & Sushi	28.599730	-81.208708	Asian Restaurant

The university data will be displayed on a Folium map with the University enrollment size proportioned as the radius of each specific university location.

Example: United States map with top ten university locations shown



Methodology

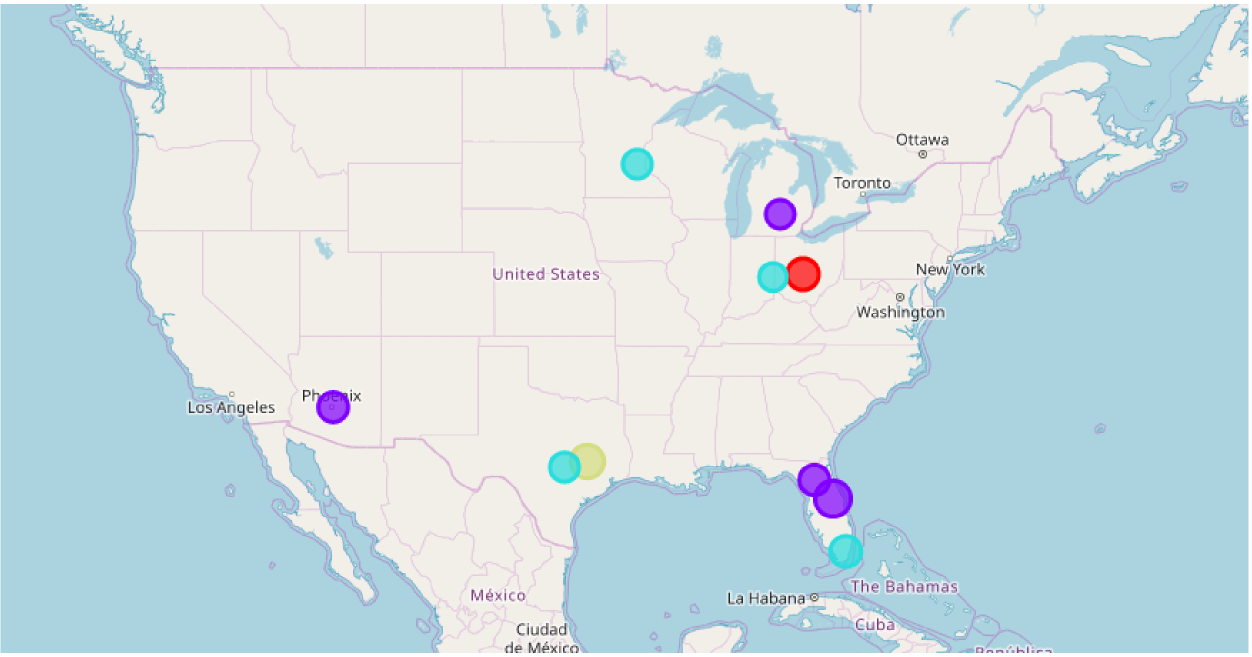
Geocoder is used to generate the latitude and longitude of each university location.

K-means analysis is used to cluster the venue into 4 distinct clusters, and plotted onto a map of the United States

Each cluster will be displayed and analyzed to determine a discriminating venue category, if one exists or can be determined.

Results

The K-means cluster results are displayed using a Folium map showing the four university clusters. So we can use the FourSquare API and venue data to determine similarities and common “grouping” features.

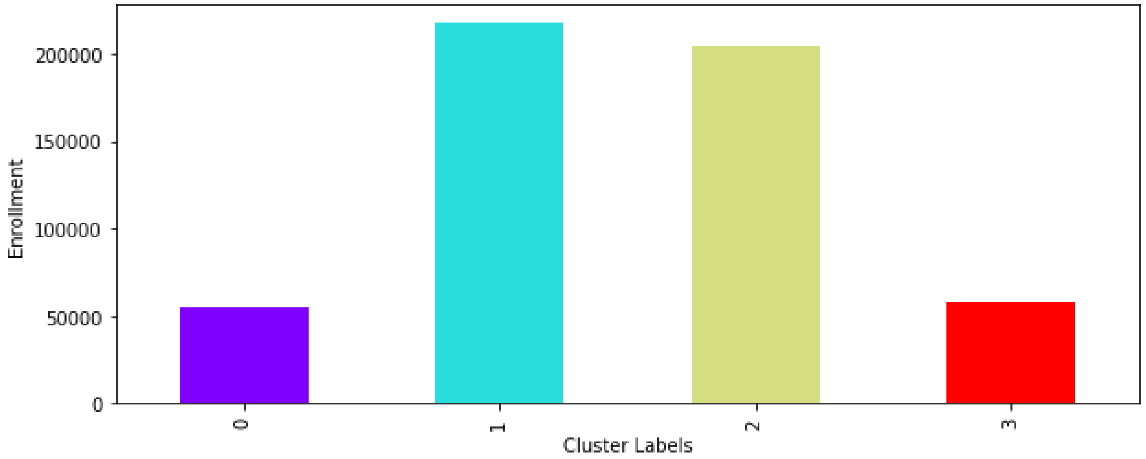


Although I expected these universities to all have common venue categories, there were 188 unique venue categories and the distribution and ranking of Top 10 venue categories did vary somewhat.

Additionally, three universities reached the “100 venue” limit possibly due to their proximity to a major city center. The three universities that reached this limit are Arizona State University, University of Minnesota and University of Texas at Austin.

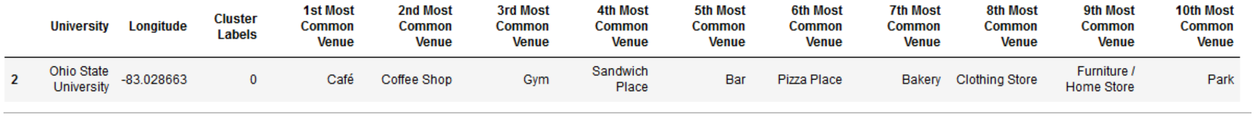
Discussion

The distribution of k-mean clusters by enrollment, also show the clusters are not evenly weighted. Ohio State make up the entirety of Cluster 0 and Texas A&M University makes up the entirety of Cluster 3, both having approximately 50,000 enrolled students.



Because Ohio State and Texas A&M University make up their own clusters we can look at the "Top 10" most common venue categories for each to ascertain a common feature or grouping construct.

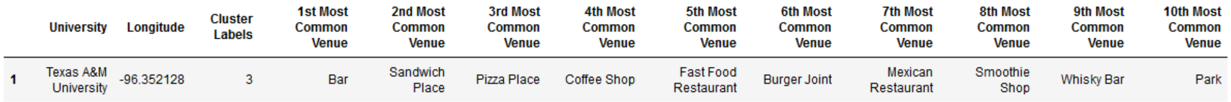
Ohio State "Top 10" are shown below:



	University	Longitude	Cluster Labels	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
2	Ohio State University	-83.028663	0	Café	Coffee Shop	Gym	Sandwich Place	Bar	Pizza Place	Bakery	Clothing Store	Furniture / Home Store	Park

Based on a preliminary review, café, coffee and gym appear in the top three venue categories so a "guess" at socializing and exercise are possible central tendencies for this cluster.

Texas A&M University "Top 10" are shown below:

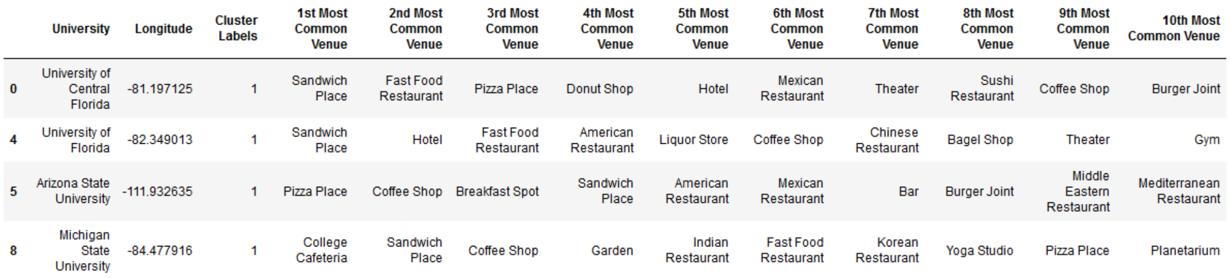


	University	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Texas A&M University	-96.352128	3	Bar	Sandwich Place	Pizza Place	Coffee Shop	Fast Food Restaurant	Burger Joint	Mexican Restaurant	Smoothie Shop	Whisky Bar	Park

Based on a preliminary review, Bar, Sandwich Place and Pizza Place appear in the top three venue categories so a "guess" at socializing and drinking alcohol are possible central tendencies for this cluster.

The remaining clusters are made up of four universities each, so identifying a common grouping feature is more difficult.

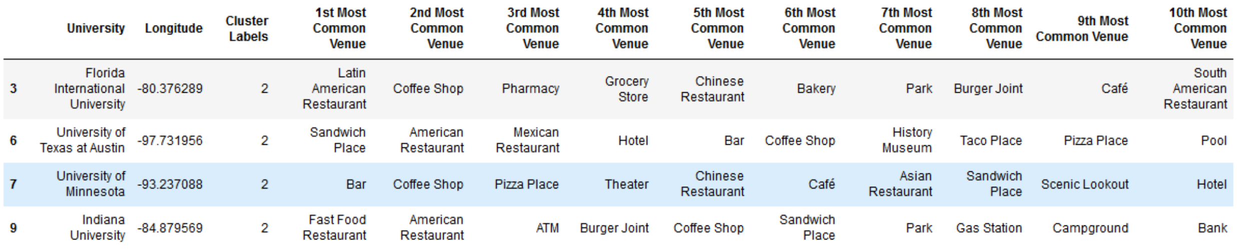
Cluster 1 contains four universities; Arizona State University, University of Central Florida, University of Florida and Michigan State University.



	University	Longitude	Cluster Labels	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
0	University of Central Florida	-81.197125	1	Sandwich Place	Fast Food Restaurant	Pizza Place	Donut Shop	Hotel	Mexican Restaurant	Theater	Sushi Restaurant	Coffee Shop	Burger Joint
4	University of Florida	-82.349013	1	Sandwich Place	Hotel	Fast Food Restaurant	American Restaurant	Liquor Store	Coffee Shop	Chinese Restaurant	Bagel Shop	Theater	Gym
5	Arizona State University	-111.932635	1	Pizza Place	Coffee Shop	Breakfast Spot	Sandwich Place	American Restaurant	Mexican Restaurant	Bar	Burger Joint	Middle Eastern Restaurant	Mediterranean Restaurant
8	Michigan State University	-84.477916	1	College Cafeteria	Sandwich Place	Coffee Shop	Garden	Indian Restaurant	Fast Food Restaurant	Korean Restaurant	Yoga Studio	Pizza Place	Planetarium

Sandwich Place, Fast food and Coffee Shop are prevalent venue categories for all four universities in this cluster group. So again, places to eat and socialize are common.

Cluster 2 also contains four universities; Florida International University, Indiana University, University of Minnesota, and University of Texas at Austin.



	University	Longitude	Cluster Labels	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
3	Florida International University	-80.376289	2	Latin American Restaurant	Coffee Shop	Pharmacy	Grocery Store	Chinese Restaurant	Bakery	Park	Burger Joint	Café	South American Restaurant
6	University of Texas at Austin	-97.731956	2	Sandwich Place	American Restaurant	Mexican Restaurant	Hotel	Bar	Coffee Shop	History Museum	Taco Place	Pizza Place	Pool
7	University of Minnesota	-93.237088	2	Bar	Coffee Shop	Pizza Place	Theater	Chinese Restaurant	Café	Asian Restaurant	Sandwich Place	Scenic Lookout	Hotel
9	Indiana University	-84.879569	2	Fast Food Restaurant	American Restaurant	ATM	Burger Joint	Coffee Shop	Sandwich Place	Park	Gas Station	Campground	Bank

The venue category groups are a bit more diverse for this cluster and include the first occurrence of ATM and Pharmacy, although places to eat, drink and socialize are still common.

Conclusion

There are appears to be similarities between the top ten universities in the United States and the venues and venue categories. So, we can use venue data and k-means clustering to determine similarities between the top 10 universities in the United States, the specific nature of the clusters remain elusive. Universities are places for students to socialize and learn and it is no surprise that the venues near universities are similar and encourage student to socialize and share their common experiences.

Although, as shown some universities remain in unique cluster groups. We can attempt a "best guess" approach as to why certain universities appear more similar based on venues and venue categories. For future research and better accuracy of this type of modeling, I recommend performing the analysis using a more comprehensive list of universities in the United States to determine if any core attribute or feature for venue clusters exist.