Finding a 4 Year Home

Joseph Kilroy

August 2019

Coursera Applied Data Science Capstone

1. Introduction

1.1 Background

Every year students go off to college where they will hopefully spend the next four years of their life. While this is the case for most students some spend one semester at an institution and return home because they were not happy. There are many reasons for this but I common reason is that they did not like the area. It could be anything from a location being too rural, too crowded, or not enough to do. I spent some time as a college counselor and would teach a course called college and career planning. During this course I would have students look at four categories while choosing schools. These were Admission Requirements, Academics, Cost, and Campus Life. I include Campus Life because this will be the next four years of their life. Academics should be the main focus of their life but if they are unhappy it will make it hard to stay motivated and succeed.

1.2 Problem

Students will spend four years getting a degree. Ideally, they will spend time researching the university and the area around it to make sure it is a good fit. Most colleges have plenty of on campus activities and venues but sometimes students need to get away from campus. Maybe it is going to a nice pizza joint with classmates after a big exam or taking the person you meet in Biology class to the movies. Some students like to work out at different gyms or do CrossFit. These could all be challenging if you pick a campus that does not have venues nearby that you are interested in. The goal of this project is to match students with a college that has similar venues as where they live. I will use Waukesha Wisconsin as the student's home base and use the four-year universities in Wisconsin as the schools of choice.

1.3 Interest

There are many people who might be interested in a project like this. Students, Teachers, and College Counselors would all be interested in using this project to help students find a university. People looking to move for a new job but like where they currently live could use this to find neighborhoods like where they currently live. Anyone who is looking at moving locations could use this to find a place to live that matches what they enjoy.

2. Data

2.1 Data Sources Prep

To begin I needed to collect some data about each school. I used the following Wikipedia site to get a list of all post-secondary schools in the state of Wisconsin. https://en.wikipedia.org/wiki/List_of_colleges_and_universities_in_Wisconsin. From this site I created a table with name and type of each school in the state of Wisconsin. Because I am only focusing on four-year universities I went through the list and deleted any college that was a tech school, specialty school, or a two-year college. I then went to each school website to get the address of the school. Once I had the address, I entered it into google to get the longitude and latitude of each school. Here is a copy of the CSV with the first 5 schools.

School	Address	Location	Type	Latitude	Longitude
Alverno College	3400 S 43rd St	Milwaukee	Master's university	42.983	-87.967
Beloit College	700 College St	Beloit	Baccalaureate college	42.503	-89.031
Cardinal Stritch University	6801 N Yates Rd	Milwaukee	Doctoral/research university	43.142	-87.906
Carroll University	100 N East Ave	Waukesha	Master's university	43.024	-88.221
Carthage College	2001 Alford Park Dr	Kenosha	Baccalaureate college	42.622	-89.822

The Latitude and Longitude is to get the venues from Foursquare. The other columns are to give students more information on the university they were matched with. Then I gathered the same data for Waukesha Wisconsin.

2.2 Data Program

In this stage of the project I used the latitude and longitude for each school to make a call to the foursquare and get a list of all venues within 3200 meters of the coordinates. I choose this distance because I figured it is a reasonable distance for someone without a car to travel. This data will be used to do a K-means cluster with Waukesha Wisconsin to determine the schools that are in the same cluster as Waukesha. Then I will create a data frame with limited venues to see if we get different results. I am doing the second test because I do not want to be matched based off venues that are not visited. For example, someone who does not own a pet would not want to be matched based off the number of pet stores.

3. Methodology

3.1 Set-up

After I created the data frame with the list of all venues with in a 3200-meter radius of all the colleges I decided to create a data frame with the counts of all the venues types. I figured there are a lot of local establishments that would not match from one town to another. For example, I really like Phil's Pizza but I know that it is a local establishment so it will not match with pizza places in other towns. By sorting the venue data frame by venue category, I will match the number of pizza places instead of trying to match the specific pizza place. Once I created a data frame for the count of each venue category for each school and a data frame for the count of the venue category for Waukesha, I merged the two data frames so I could do the k-means clustering.

3.2 Clustering

Using the data frame with Waukesha and the college venues I performed K-means clustering. After running some test runs with different numbers it appeared that 6 clusters were the best result for getting an even distribution of the schools across the clusters. As you will see in the tables in the report sections the schools clustered with Waukesha have a good match.

While this was a good first run, I did not want someone to be matched on venues that they do not utilize. For example, someone without any pets would not want to be clustered in a location with many pet stores. I picked the following categories and re-ran the k-means clustering to see if the result changed. The venues I choose were Sports Bar, Coffee Shop, Diner, Gym, Gym / Fitness Center, Mexican Restaurant, Pizza Place, Bank, Golf Course, and Basketball Court.

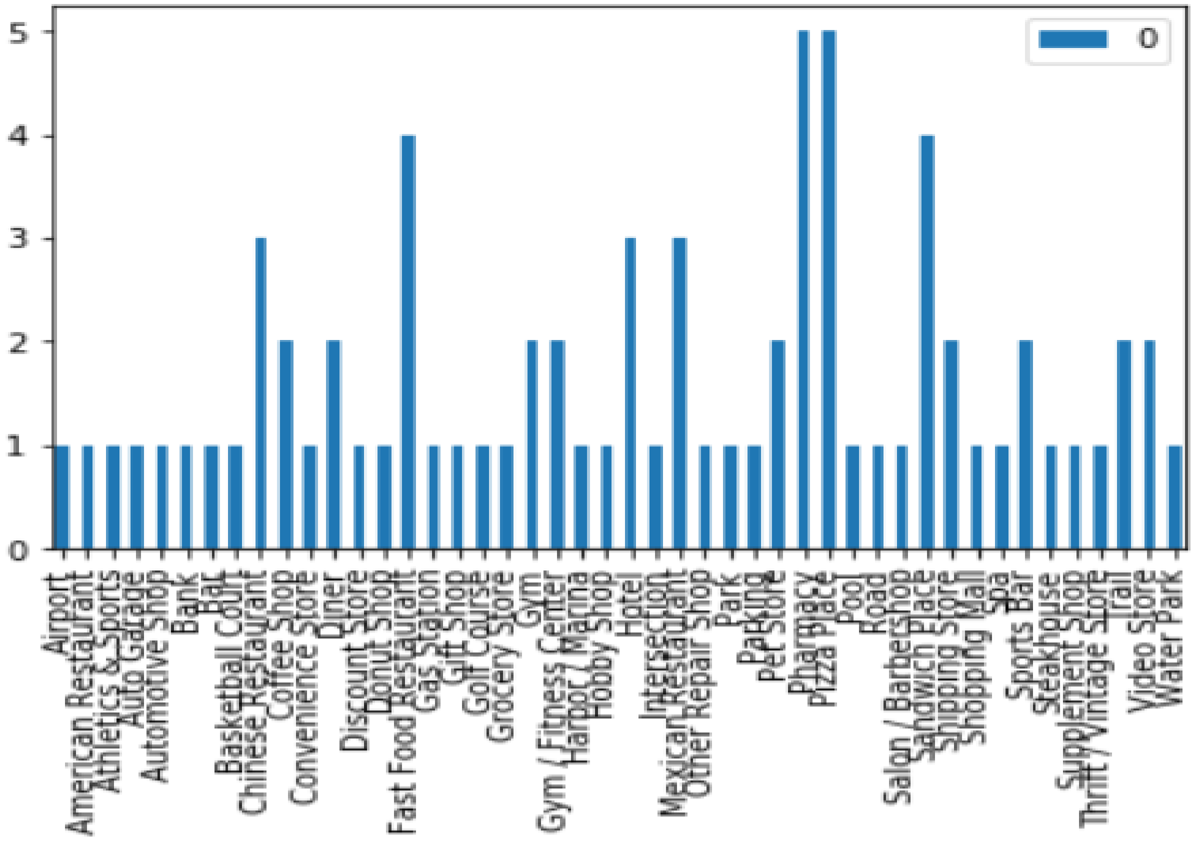
4 Results

4.1 Results for First Attempt

After running the k-means the first time with all venue categories the following schools were clustered with Waukesha, Cardinal Stritch University, Concordia University, Lakeland University, Northland College, Ripon College, Silver Lake College, University of Wisconsin-Green Bay, University of Wisconsin-Parkside, University of Wisconsin-River Falls, and University of Wisconsin- Stevens Point.

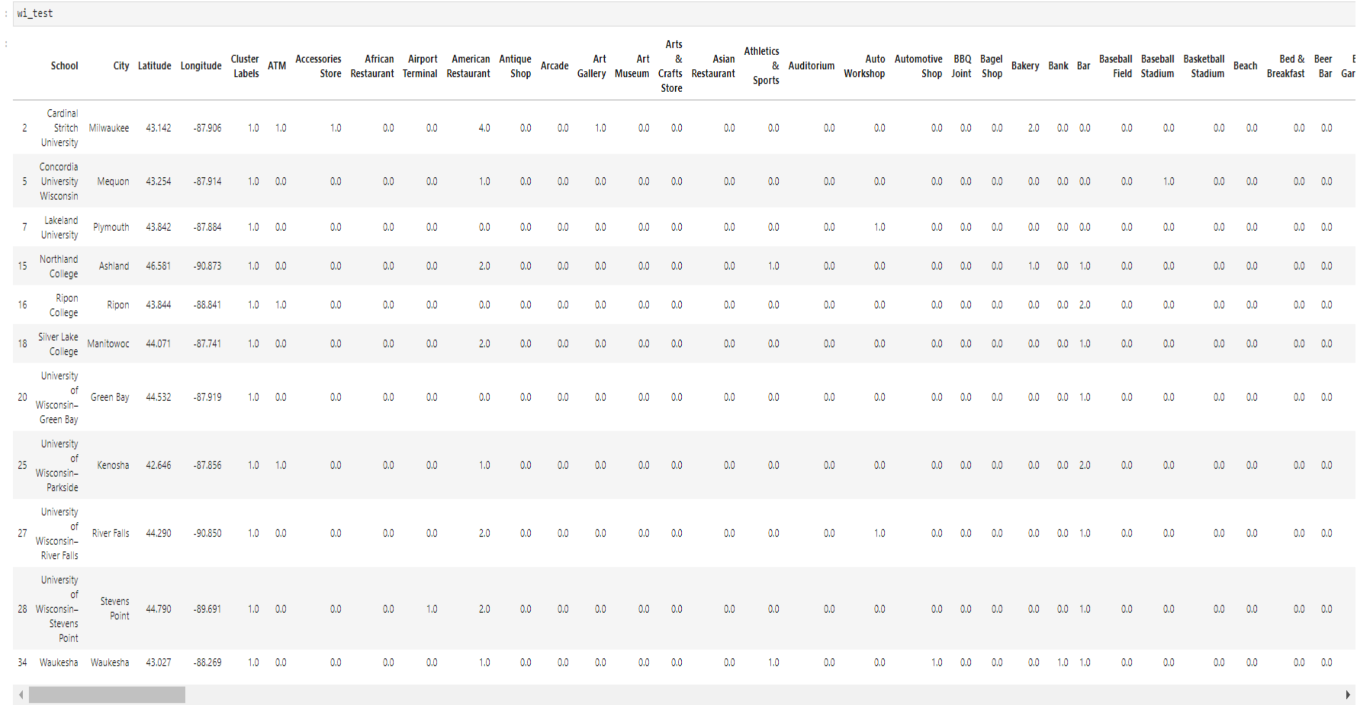
4.2 Comparing the Results from the First Attempt

Here is a table of the number of venues in Waukesha



As you can see there are only a few venues that have more then one venue.

Here is the table of venue counts for all the schools clustered with Waukesha. For the entire table please see the Jupyter notebook.



```
wi_test
```

	School	City	Latitude	Longitude	Cluster Labels	ATM	Accessories Store	African Restaurant	Airport Terminal	American Restaurant	Antique Shop	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auditorium	Auto Workshop	Automotive Shop	BBQ Joint	Bagel Shop	Bakery	Bank	Bar	Baseball Field	Baseball Stadium	Basketball Stadium	Beach	Bed & Breakfast	Beer Bar
2	Cardinal Stritch University	Milwaukee	43.142	-87.906	1.0	1.0	1.0	0.0	0.0	4.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	Concordia University Wisconsin	Mequon	43.254	-87.914	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
7	Lakeland University	Plymouth	43.842	-87.884	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	Northland College	Ashland	46.581	-90.873	1.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
16	Ripon College	Ripon	43.844	-88.841	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0
18	Silver Lake College	Manitowoc	44.071	-87.741	1.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
20	University of Wisconsin–Green Bay	Green Bay	44.532	-87.919	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
25	University of Wisconsin–Parkside	Kenosha	42.646	-87.856	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0
27	University of Wisconsin–River Falls	River Falls	44.290	-90.850	1.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
28	University of Wisconsin–Stevens Point	Stevens Point	44.790	-89.691	1.0	0.0	0.0	0.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
34	Waukesha	Waukesha	43.027	-88.269	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0

There are many venue categories that are not with-in the range of Waukesha. For example, Cardinal Stritch has both the venues ATM and accessories stores witch Waukesha does not have. If you look at the column for American Restaurant you will see that Waukesha has 1 American restaurant while Cardinal Stritch has 4. Which is fine if you like American restaurants but if we look at UW-Green Bay, Ripon College, and Lakeland they do not have any American restaurants.

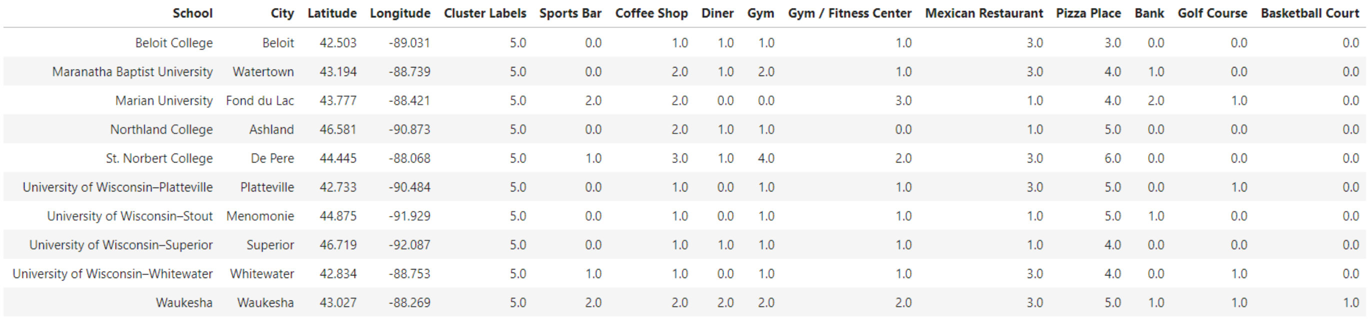
While this was a good start it seems that having to many venue categories is making it hard to find a good match. In the next run we will limit the venue categories and see if we get a closer match.

4.3 Run 2 Limited Venues

The venues I choose were Sports Bar, Coffee Shop, Diner, Gym, Gym / Fitness Center, Mexican Restaurant, Pizza Place, Bank, Golf Course, and Basketball Court. When cluster with k-means using 6 cluster Waukesha was clustered with the following schools. Beloit College, Maranatha Baptist University, Marian University, Northland College, St. Norbert College, University of Wisconsin-Platteville, University of Wisconsin-Stout, University of Wisconsin-Superior, and University of Wisconsin-Whitewater. As you can see this is a completely different list.

4.4 Comparing the Results for Run 2

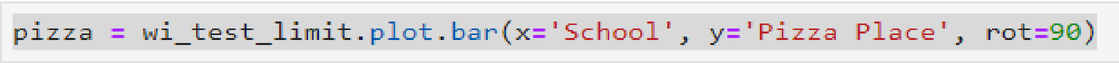
Here is the table of venue counts with just the limited venues.



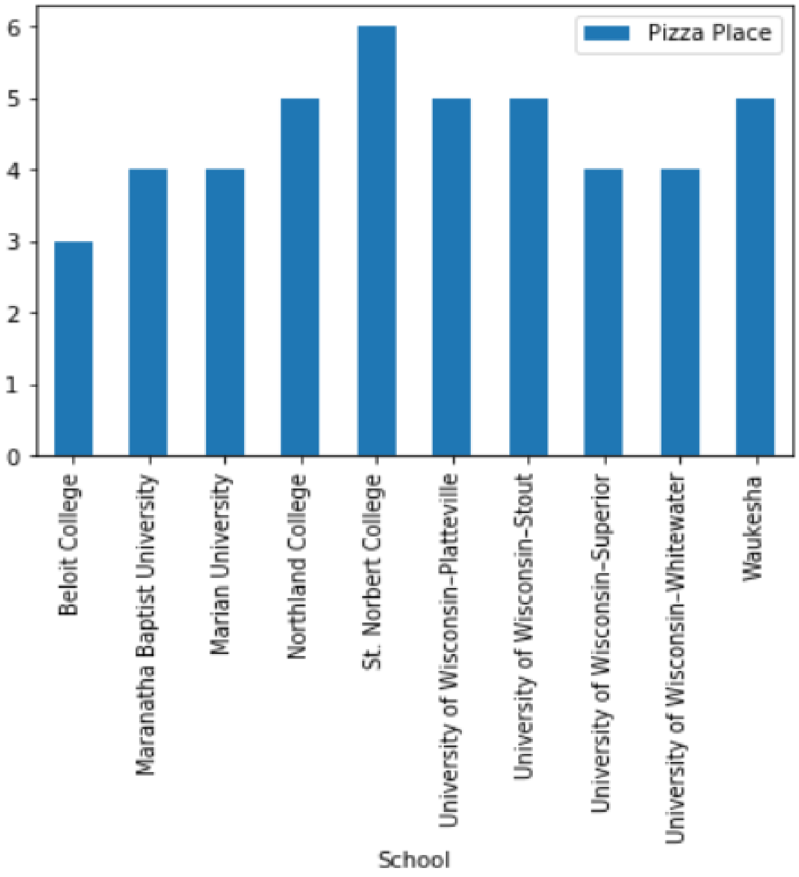
School	City	Latitude	Longitude	Cluster Labels	Sports Bar	Coffee Shop	Diner	Gym	Gym / Fitness Center	Mexican Restaurant	Pizza Place	Bank	Golf Course	Basketball Court
Beloit College	Beloit	42.503	-89.031	5.0	0.0	1.0	1.0	1.0	1.0	3.0	3.0	0.0	0.0	0.0
Maranatha Baptist University	Watertown	43.194	-88.739	5.0	0.0	2.0	1.0	2.0	1.0	3.0	4.0	1.0	0.0	0.0
Marian University	Fond du Lac	43.777	-88.421	5.0	2.0	2.0	0.0	0.0	3.0	1.0	4.0	2.0	1.0	0.0
Northland College	Ashland	46.581	-90.873	5.0	0.0	2.0	1.0	1.0	0.0	1.0	5.0	0.0	0.0	0.0
St. Norbert College	De Pere	44.445	-88.068	5.0	1.0	3.0	1.0	4.0	2.0	3.0	6.0	0.0	0.0	0.0
University of Wisconsin–Platteville	Platteville	42.733	-90.484	5.0	0.0	1.0	0.0	1.0	1.0	3.0	5.0	0.0	1.0	0.0
University of Wisconsin–Stout	Menomonie	44.875	-91.929	5.0	0.0	1.0	0.0	1.0	1.0	1.0	5.0	1.0	0.0	0.0
University of Wisconsin–Superior	Superior	46.719	-92.087	5.0	0.0	1.0	1.0	1.0	1.0	1.0	4.0	0.0	0.0	0.0
University of Wisconsin–Whitewater	Whitewater	42.834	-88.753	5.0	1.0	1.0	0.0	1.0	1.0	3.0	4.0	0.0	1.0	0.0
Waukesha	Waukesha	43.027	-88.269	5.0	2.0	2.0	2.0	2.0	2.0	3.0	5.0	1.0	1.0	1.0

For most of the categories there are better results. Basketball Court has all zeros except for Waukesha but when I went back and looked at the Venue Counts Waukesha was the only school with a Basketball court with-in 3200 meters.

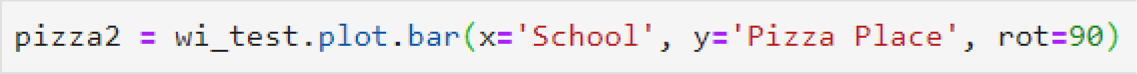
Here is a graph comparing the number of Pizza Places.



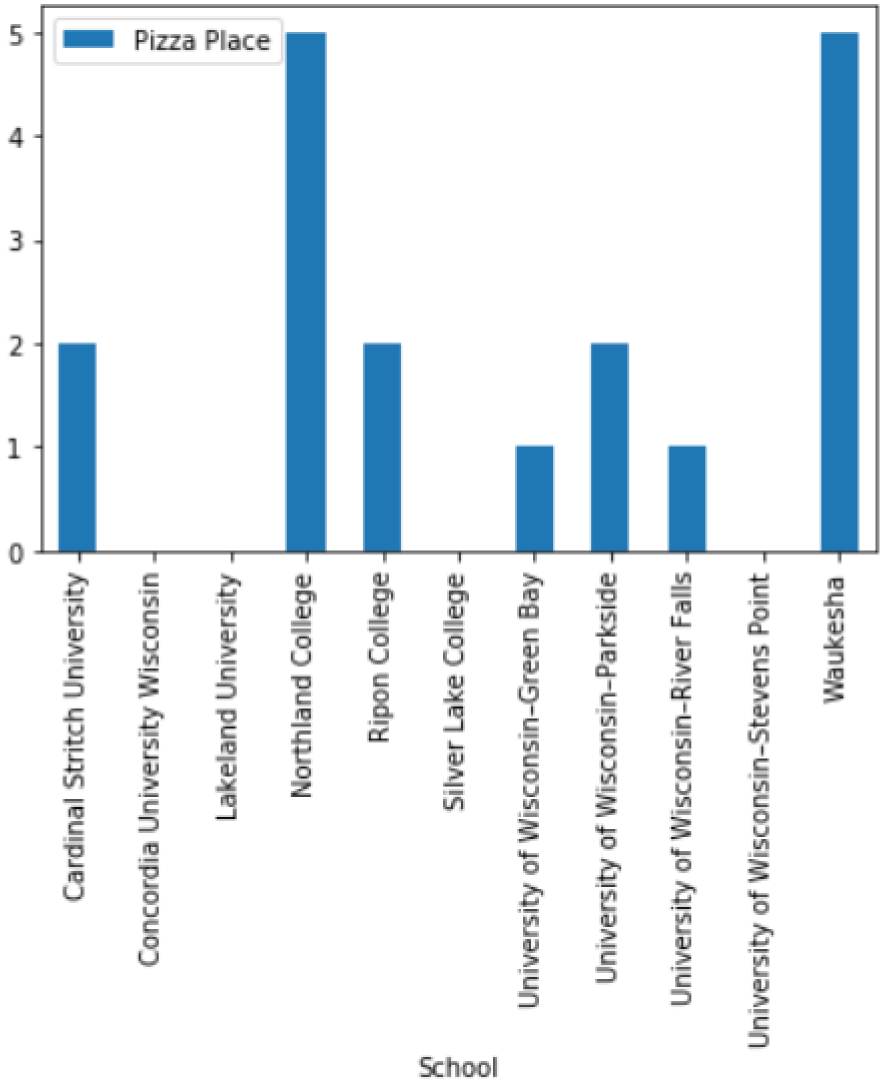
```
pizza = wi_test_limit.plot.bar(x='School', y='Pizza Place', rot=90)
```



Here is the same graph without using the limited venues for clustering.



```
pizza2 = wi_test.plot.bar(x='School', y='Pizza Place', rot=90)
```

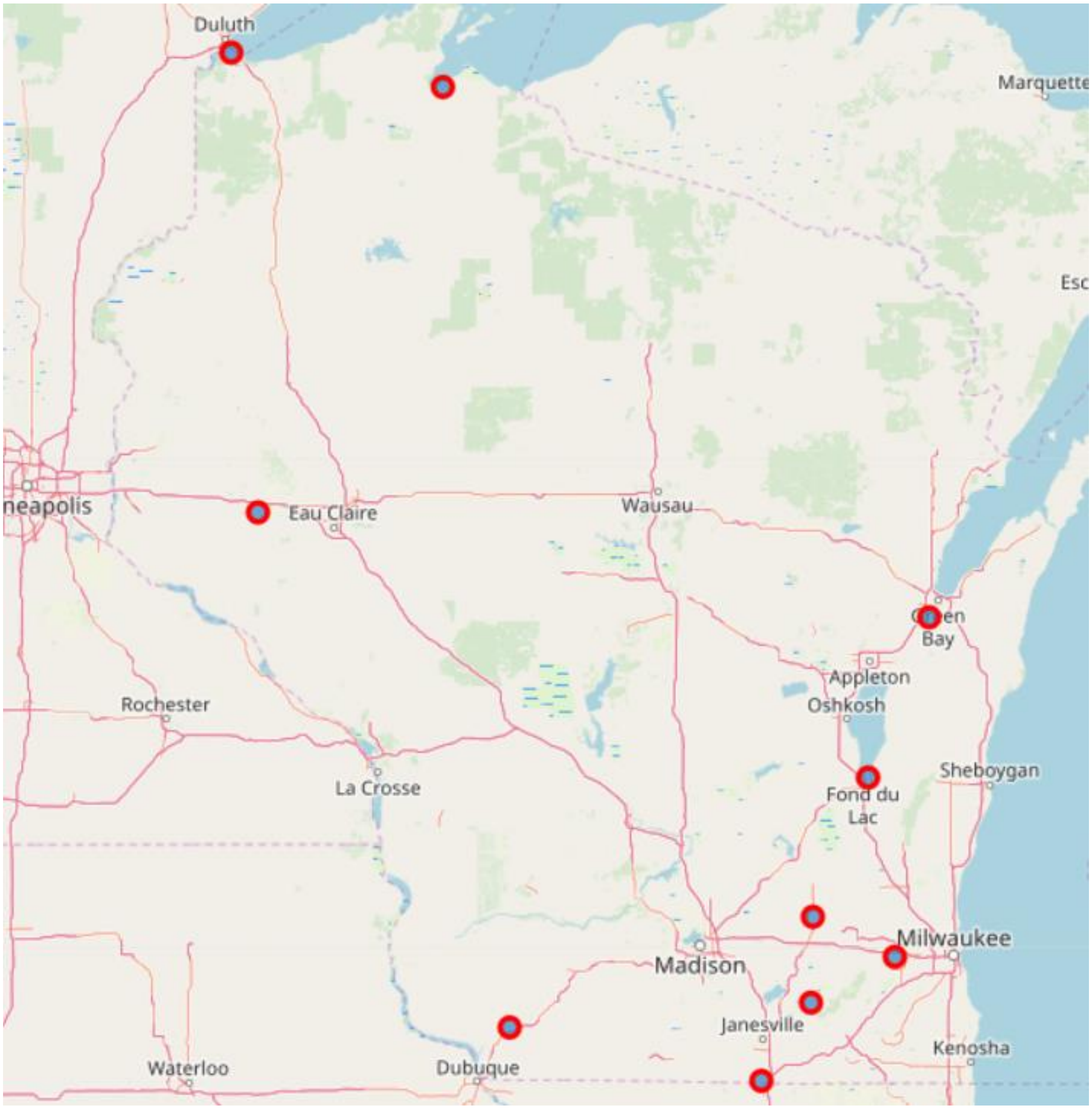


We can see that at least for the venue pizza places the limited clustering method has better results.

5.0 Information for the Client

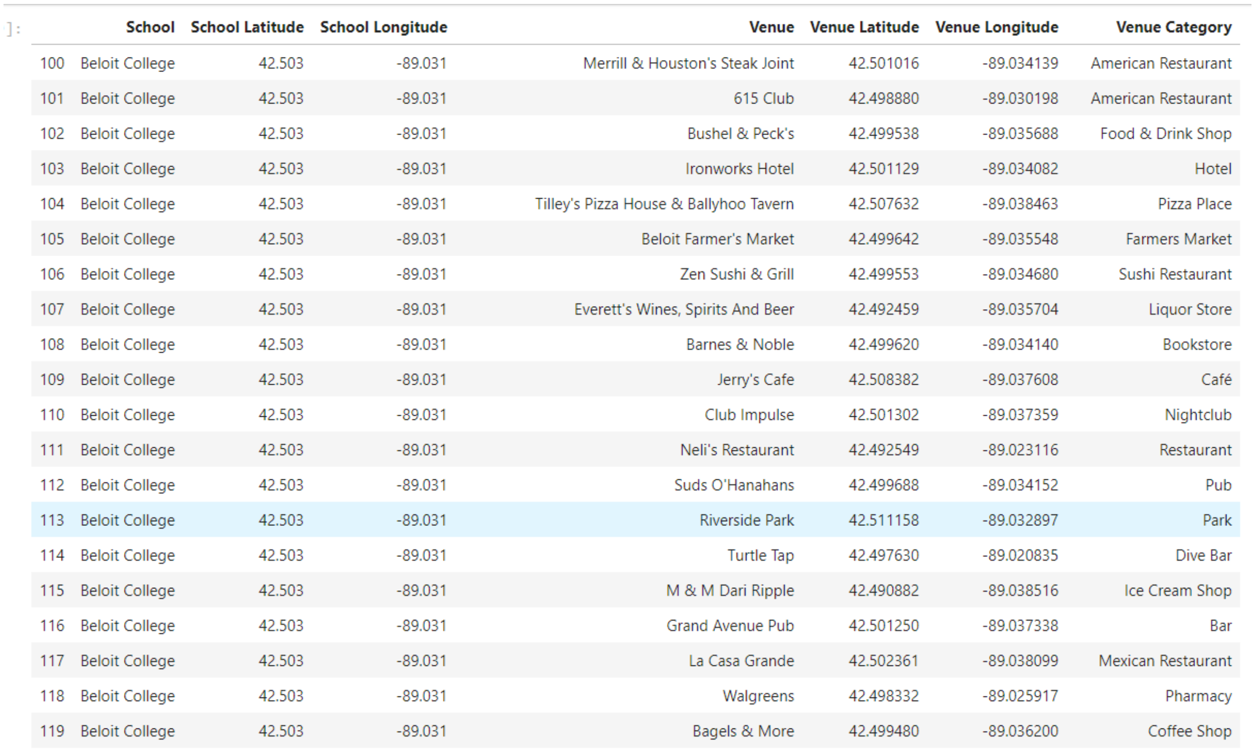
We can return some information to the client. After we have the clustered schools, we can return information on the area and the venues in the area.

Here is a map of all the schools cluster with Waukesha with run attempt 2.



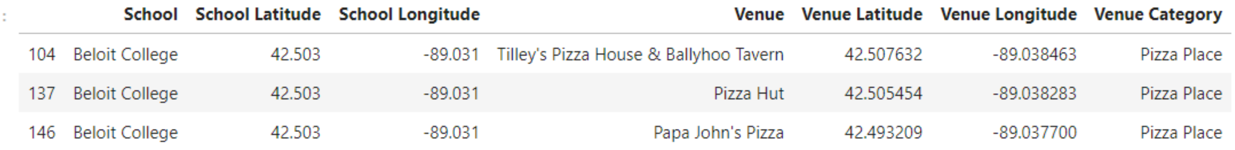
We can also return information about each school.

Here is the beginning of a list of all venues for Beloit College.



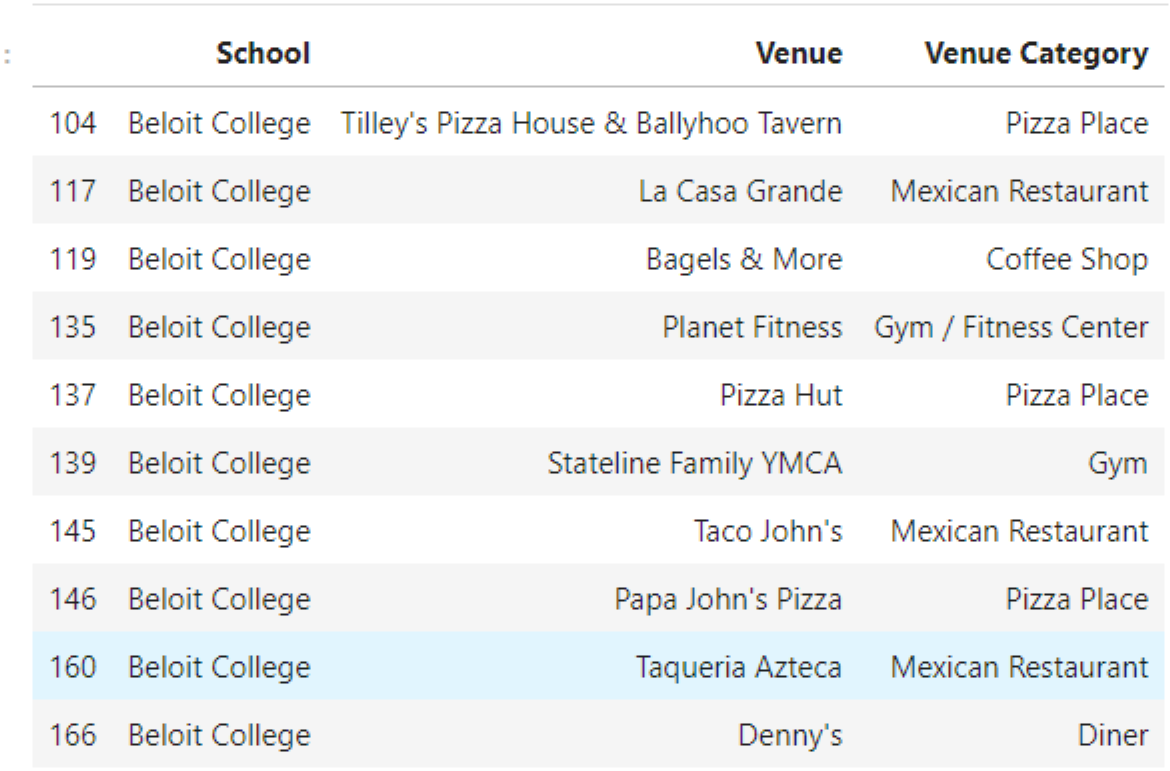
	School	School Latitude	School Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
100	Beloit College	42.503	-89.031	Merrill & Houston's Steak Joint	42.501016	-89.034139	American Restaurant
101	Beloit College	42.503	-89.031	615 Club	42.498880	-89.030198	American Restaurant
102	Beloit College	42.503	-89.031	Bushel & Peck's	42.499538	-89.035688	Food & Drink Shop
103	Beloit College	42.503	-89.031	Ironworks Hotel	42.501129	-89.034082	Hotel
104	Beloit College	42.503	-89.031	Tilley's Pizza House & Ballyhoo Tavern	42.507632	-89.038463	Pizza Place
105	Beloit College	42.503	-89.031	Beloit Farmer's Market	42.499642	-89.035548	Farmers Market
106	Beloit College	42.503	-89.031	Zen Sushi & Grill	42.499553	-89.034680	Sushi Restaurant
107	Beloit College	42.503	-89.031	Everett's Wines, Spirits And Beer	42.492459	-89.035704	Liquor Store
108	Beloit College	42.503	-89.031	Barnes & Noble	42.499620	-89.034140	Bookstore
109	Beloit College	42.503	-89.031	Jerry's Cafe	42.508382	-89.037608	Café
110	Beloit College	42.503	-89.031	Club Impulse	42.501302	-89.037359	Nightclub
111	Beloit College	42.503	-89.031	Neli's Restaurant	42.492549	-89.023116	Restaurant
112	Beloit College	42.503	-89.031	Suds O'Hanahans	42.499688	-89.034152	Pub
113	Beloit College	42.503	-89.031	Riverside Park	42.511158	-89.032897	Park
114	Beloit College	42.503	-89.031	Turtle Tap	42.497630	-89.020835	Dive Bar
115	Beloit College	42.503	-89.031	M & M Dari Ripple	42.490882	-89.038516	Ice Cream Shop
116	Beloit College	42.503	-89.031	Grand Avenue Pub	42.501250	-89.037338	Bar
117	Beloit College	42.503	-89.031	La Casa Grande	42.502361	-89.038099	Mexican Restaurant
118	Beloit College	42.503	-89.031	Walgreens	42.498332	-89.025917	Pharmacy
119	Beloit College	42.503	-89.031	Bagels & More	42.499480	-89.036200	Coffee Shop

We could also give them a list for each specific venue they are interested in. Here is a list of all the pizza places by Beloit College.

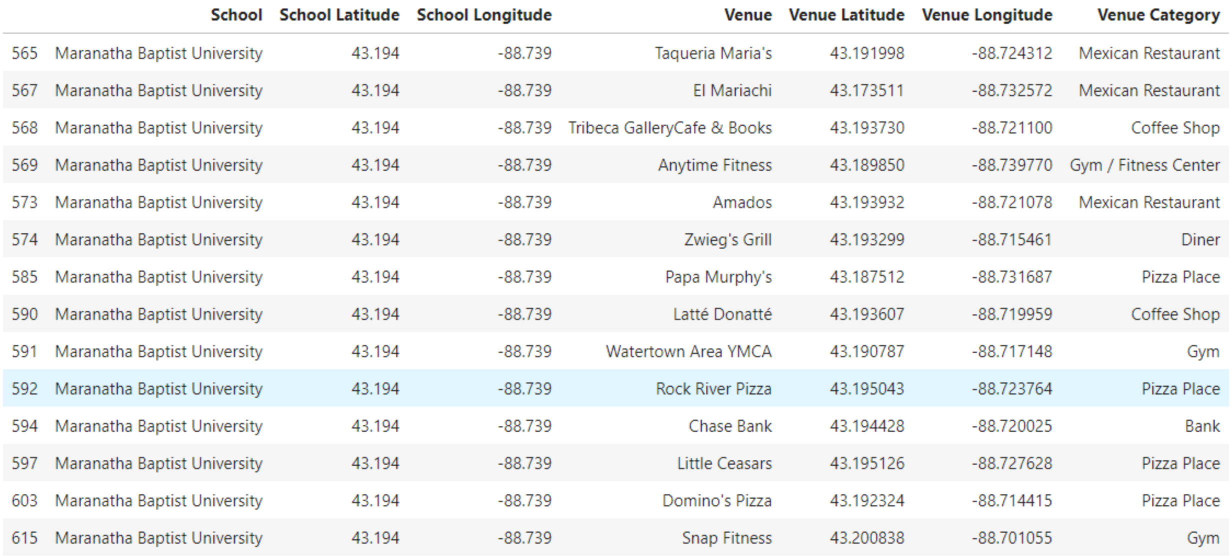


	School	School Latitude	School Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
104	Beloit College	42.503	-89.031	Tilley's Pizza House & Ballyhoo Tavern	42.507632	-89.038463	Pizza Place
137	Beloit College	42.503	-89.031	Pizza Hut	42.505454	-89.038283	Pizza Place
146	Beloit College	42.503	-89.031	Papa John's Pizza	42.493209	-89.037700	Pizza Place

We could provide a list of all the venues for the limited selection. Here are the list for Beloit College and Maranatha Baptist Bible College.



	School	Venue	Venue Category
104	Beloit College	Tilley's Pizza House & Ballyhoo Tavern	Pizza Place
117	Beloit College	La Casa Grande	Mexican Restaurant
119	Beloit College	Bagels & More	Coffee Shop
135	Beloit College	Planet Fitness	Gym / Fitness Center
137	Beloit College	Pizza Hut	Pizza Place
139	Beloit College	Stateline Family YMCA	Gym
145	Beloit College	Taco John's	Mexican Restaurant
146	Beloit College	Papa John's Pizza	Pizza Place
160	Beloit College	Taqueria Azteca	Mexican Restaurant
166	Beloit College	Denny's	Diner



	School	School Latitude	School Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
565	Maranatha Baptist University	43.194	-88.739	Taqueria Maria's	43.191998	-88.724312	Mexican Restaurant
567	Maranatha Baptist University	43.194	-88.739	El Mariachi	43.173511	-88.732572	Mexican Restaurant
568	Maranatha Baptist University	43.194	-88.739	Tribeca GalleryCafe & Books	43.193730	-88.721100	Coffee Shop
569	Maranatha Baptist University	43.194	-88.739	Anytime Fitness	43.189850	-88.739770	Gym / Fitness Center
573	Maranatha Baptist University	43.194	-88.739	Amados	43.193932	-88.721078	Mexican Restaurant
574	Maranatha Baptist University	43.194	-88.739	Zwieg's Grill	43.193299	-88.715461	Diner
585	Maranatha Baptist University	43.194	-88.739	Papa Murphy's	43.187512	-88.731687	Pizza Place
590	Maranatha Baptist University	43.194	-88.739	Latté Donatté	43.193607	-88.719959	Coffee Shop
591	Maranatha Baptist University	43.194	-88.739	Watertown Area YMCA	43.190787	-88.717148	Gym
592	Maranatha Baptist University	43.194	-88.739	Rock River Pizza	43.195043	-88.723764	Pizza Place
594	Maranatha Baptist University	43.194	-88.739	Chase Bank	43.194428	-88.720025	Bank
597	Maranatha Baptist University	43.194	-88.739	Little Ceasars	43.195126	-88.727628	Pizza Place
603	Maranatha Baptist University	43.194	-88.739	Domino's Pizza	43.192324	-88.714415	Pizza Place
615	Maranatha Baptist University	43.194	-88.739	Snap Fitness	43.200838	-88.701055	Gym

6.0 Moving Forward

The method which limits the venue categories seems to work better than the method that uses all categories. Most people would want to be matched on venues they visit so if they could choose the category it would give them better results.

Going forward, I should try other clustering methods, but k-means is the only one I have had time to practice. I also limited the number of venues and the distance from each venue, so I did not use up all my calls to foursquare. If I paid for an account, I would increase both these numbers. It would also be nice if people could rank things, they enjoy instead of matching them to a location but that programming skills is beyond my current skill level.

Appendix 1 Data frame list

I use a few different data frames though out my code. Mainly because I needed to use them at the end for the reports so instead of modifying a data frame, I generally just created a new one. Here is a list of them if you want to follow along with the code.

df	Schools and their longitude and latitude
school_venues	Venues of all the schools
waukesha_df	Waukesha Latitude and longitude
waukesha_venues	Waukesha venues
waukesha_grouped	Waukesha venues grouped by venue category
waukesha_area	Sorting Waukesha venues
school_grouped	School venues grouped by venue category
wi	Combined school_grouped and Waukesha_grouped
wi_merged	Merge df and Waukesha_df to make a data frame with all information about the schools and clusters
wi_cluster	Combine wi_merged with wi so that the schools location are on the wi dataframe
wi_test	Data frame that has all the schools that are only in the same cluster as Waukesha
wi_limit	Limit the Wisconsin Data Frame to a few venue catagories
wi_limit_cluster	Combine the wi_merged data frame with the wi_limit
wi_test_limit	All the schools that are in the same cluster as Waukesha