IBM Applied Data Science Capstone

The Taste of Cincinnati

An Exercise in K-Means Clustering

Sanjana Are
5-5-2020

1. Introduction

1.1 Background

Known as the “Queen City of the West,” Cincinnati, Ohio is home to a diverse array of global cuisines. From the renowned American Tom & Chee to the homely French Graeter’s, locals and tourists alike are hardpressed to find a more culturally rich metropolis in the Midwest. In fact, thanks to a recent revitalization of the restaurant scene, USA Today has consistently named Cincinnati as “one of six small cities with big food scenes" since 2012. Furthermore, Cincinnati’s food scene ranks among the nation’s most affordable, with the average price of a three-course meal for 2 at a mid-range restaurant at only $40, second to only San Antonio, Texas at $35. With new restaurants taking hold every year, the city’s food scene shows no signs of slowing down.

1.2 Problem

For the tourist, or even city native, Cincinnati’s food scene can often be quite overwhelming. With a myriad of restaurants from different cuisines to choose from, it is hard to objectively determine which restaurants are the most worth visiting. In this paper, I hope to distinguish the upper echelon of restaurants using a machine learning model.

1.3 Interest

This classification is of interest to many different stakeholders, namely tourists and city natives who are seeking to explore Cincinnati’s bustling food scene. Furthermore, such a classification is also of interest to food critics whose livelihoods depend upon providing scathing critiques of renowned restaurants. Finally, this classification is of interest to restaurants themselves, who gain significant exposure by being in the upper echelon.

2. Data

2.1 Data Description

The raw data used in this analysis includes 64 restaurants within 1000 meters of the heart of Cincinnati (geographical coordinates 39.1014537°, -84.5124602°). Each restaurant’s name, latitude, longitude, unique ID, and venue type is recorded upon the initial read of the Foursquare API. Upon further querying, the number of Foursquare users who have liked a restaurant is also recorded. The number of likes and the venue type are then utilized to categorize each restaurant into a different cluster based on a k-means clustering algorithm. The latitude and longitude of each restaurant is utilized to map each cluster using the Folium package. A snapshot of the data is provided in *Figure 1* below.

	name	id	categories	lat	lng
0	Sotto	5154c81ae4b0c54802cba3c7	Italian Restaurant	39.102797	-84.511263
1	21c Museum Hotels - Cincinnati	4f1825f8e4b0b4cc23ba433b	Hotel	39.103165	-84.512087
2	Boca	5185a0d0498e2061f617db14	Restaurant	39.102785	-84.511302
3	Aronoff Center for the Arts	4b4607f3f964a520871426e3	Performing Arts Venue	39.103560	-84.511932
4	Sleepy Bee Cafe	5a6dea8297cf5a7b38b5e293	Café	39.100055	-84.512272
5	Contemporary Arts Center	4b48bd02f964a520d65426e3	Art Museum	39.102685	-84.511811
6	Orchids at Palm Court	4b1478a3f964a5208ba323e3	New American Restaurant	39.100626	-84.514335
7	Fountain Square	4b438206f964a520fbe125e3	Plaza	39.101448	-84.512519
8	Aster On Fourth	5a4e9cc26bd36b1ecb142000	Cocktail Bar	39.100030	-84.512254
9	Graeter's Ice Cream	4b4f5355f964a520680127e3	Ice Cream Shop	39.101487	-84.511860
10	Nada	4b317901f964a520910725e3	Mexican Restaurant	39.102941	-84.511680

Figure 1

2.2 Data Source

Data will be collected using the Foresquare API, a free tool that allows developers to access location-based experiences with diverse information about venues, users, photos, and check-ins. In addition, the API supports real time access to places, Snap-to-Place that assigns users to specific locations, and Geo-tag. JSON is the preferred response format.

3. Methodology

3.1 Imports

For this analysis, a variety of specific libraries were required. Most notably, *geopy* was required to convert an address into a longitude and latitude value, *sklearn* was required to run the k-means machine learning algorithm, *folium* was required to visually render a cluster on a map, and *json* was required to handle venue data stored in a JSON file.

3.2 Foresquare API Setup

To successfully establish a connection with the Foresquare API, a client ID, client secret, and version ID were instantiated. Then, 100 venues within 1000 meters of the heart of Cincinnati were read, and the results were stored in a JSON file.

3.3 Initial Data Collection

The JSON file described in *3.2* was parsed, with each venue's name, unique ID, category, latitude, and longitude stored in a pandas dataframe. A unique list of venue categories was then derived from this sample to identify all categories that do not fall within the realm of "restaurant." A new sample of 64 venues that could be categorized as such was then created. A snapshot of the data is provided in *Figure 2* below.

	name	id	categories	lat	lng
0	Sotto	5154c81ae4b0c54802cba3c7	Italian Restaurant	39.102797	-84.511263
1	Boca	5185a0d0498e2061f617db14	Restaurant	39.102785	-84.511302
2	Sleepy Bee Cafe	5a6dea8297cf5a7b38b5e293	Café	39.100055	-84.512272
3	Orchids at Palm Court	4b1478a3f964a5208ba323e3	New American Restaurant	39.100626	-84.514335
4	Aster On Fourth	5a4e9cc26bd36b1ecb142000	Cocktail Bar	39.100030	-84.512254
5	Graeter's Ice Cream	4b4f5355f964a520680127e3	Ice Cream Shop	39.101487	-84.511860
6	Nada	4b317901f964a520910725e3	Mexican Restaurant	39.102941	-84.511680
7	Maplewood Kitchen and Bar	57680d90498e9e7e4183734a	Breakfast Spot	39.101513	-84.515113
8	Abby Girl Sweets	4b4b728ff964a520059c26e3	Cupcake Shop	39.101057	-84.514314
9	FUSIAN	4bed7a8091380f47c9f09f18	Sushi Restaurant	39.102720	-84.512924
10	Metropole	50980e2bd63eb33c0a84d7f4	Restaurant	39.103174	-84.511813

Figure 2

3.4 Advanced Data Collection

In order to provide adequate data for the k-means clustering algorithm, the number of Foursquare users who liked a venue was collected from the Foursquare API using the unique ID value of each venue. This data was then concatenated into the existing dataframe in the *likes* column. In order to better understand this new data, the matplotlib library was utilized to plot the distribution of Foursquare data in a histogram. This data revealed a notable right skew in the data, the results of which can be seen in *Figure 3*.

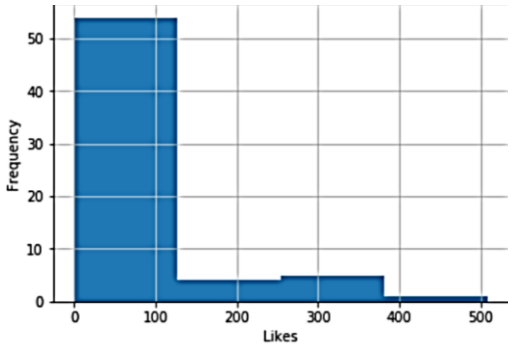

Figure 3

For the sake of the k-means clustering algorithm, the quantitative *likes* data was converted into more broad qualitative data, with ratings of “poor”, “below average”, “average”, and “great” assigned based on the 25th, 50th, and 75th quartiles of the data. This new categorization was then added into the dataframe, as seen in *Figure 4*.

	name	id	categories	lat	lng	likes	likes category
0	Sotto	5154c81ae4b0c54802cba3c7	Italian Restaurant	39.102797	-84.511263	205	great
1	Boca	5185a0d0498e2061f617db14	Restaurant	39.102785	-84.511302	95	great
2	Sleepy Bee Cafe	5a6dea8297cf5a7b38b5e293	Café	39.100055	-84.512272	23	below avg
3	Orchids at Palm Court	4b1478a3f964a5208ba323e3	New American Restaurant	39.100626	-84.514335	61	above avg
4	Aster On Fourth	5a4e9cc26bd36b1ecb142000	Cocktail Bar	39.100030	-84.512254	16	poor
5	Graeter's Ice Cream	4b4f5355f964a520680127e3	Ice Cream Shop	39.101487	-84.511860	119	great
6	Nada	4b317901f964a520910725e3	Mexican Restaurant	39.102941	-84.511680	358	great
7	Maplewood Kitchen and Bar	57680d90498e9e7e4183734a	Breakfast Spot	39.101513	-84.515113	98	great
8	Abby Girl Sweets	4b4b728ff964a520059c26e3	Cupcake Shop	39.101057	-84.514314	16	poor
9	FUSIAN	4bed7a8091380f47c9f09f18	Sushi Restaurant	39.102720	-84.512924	66	above avg
10	Metropole	50980e2bd63eb33c0a84d7f4	Restaurant	39.103174	-84.511813	85	above avg

Figure 4

A similar process was repeated to group the *categories* field into the more broad *food type* column (“european food,” “other food,” “hispanic food,” “asian food,” “american food,” and “bars”). Each original category was manually placed into a new category, and the new categories were merged into the existing dataframe, as shown in *Figure 5*.

	name	id	categories	lat	lng	likes	likes category	food type
0	Sotto	5154c81ae4b0c54802cba3c7	Italian Restaurant	39.102797	-84.511263	205	great	european
1	Boca	5185a0d0498e2061f617db14	Restaurant	39.102785	-84.511302	95	great	other
2	Sleepy Bee Cafe	5a6dea8297cf5a7b38b5e293	Café	39.100055	-84.512272	23	below avg	other
3	Orchids at Palm Court	4b1478a3f964a5208ba323e3	New American Restaurant	39.100626	-84.514335	61	above avg	american
4	Aster On Fourth	5a4e9cc26bd36b1ecb142000	Cocktail Bar	39.100030	-84.512254	16	poor	bar
5	Graeter's Ice Cream	4b4f5355f964a520680127e3	Ice Cream Shop	39.101487	-84.511860	119	great	other
6	Nada	4b317901f964a520910725e3	Mexican Restaurant	39.102941	-84.511680	358	great	hispanic
7	Maplewood Kitchen and Bar	57680d90498e9e7e4183734a	Breakfast Spot	39.101513	-84.515113	98	great	other
8	Abby Girl Sweets	4b4b728ff964a520059c26e3	Cupcake Shop	39.101057	-84.514314	16	poor	other
9	FUSIAN	4bed7a8091380f47c9f09f18	Sushi Restaurant	39.102720	-84.512924	66	above avg	asian
10	Metropole	50980e2bd63eb33c0a84d7f4	Restaurant	39.103174	-84.511813	85	above avg	other

Figure 5

3.5 K-Means Clustering

To prepare for k-means clustering, the data underwent one-hot encoding (the process by which a qualitative variable is removed, and a new binary variable is added for each unique value). Because there are 4 unique values in the *likes category* column and 6 unique values in the *food type* column, the resulting dataframe had 11 columns – 10 binary variables and the *name* field. The new dataframe can be seen in *Figure 6*.

	name	above avg	below avg	great	poor	american	asian	bar	european	hispanic	other
0	Sotto	0	0	1	0	0	0	0	1	0	0
1	Boca	0	0	1	0	0	0	0	0	0	1
2	Sleepy Bee Cafe	0	1	0	0	0	0	0	0	0	1
3	Orchids at Palm Court	1	0	0	0	1	0	0	0	0	0
4	Aster On Fourth	0	0	0	1	0	0	1	0	0	0

Figure 6

To determine the optimal k-value, or number of clusters, the mean squared error (MSE) of each k-value from 1 to 10 was plotted. The optimal k-value of 6 was determined based on the decreasing returns seen in any k-value past 6. This can be seen in *Figure 7* below.

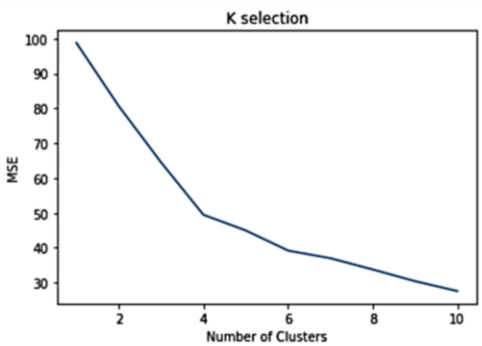

Figure 7

Finally, the k-means algorithm was run on the one-hot encoded data with a k-value of 6. Cluster labels were added to the original dataframe, as seen in *Figure 8*.

	name	id	categories	lat	lng	likes	likes category	food type	cluster
0	Sotto	5154c81ae4b0c54802cba3c7	Italian Restaurant	39.102797	-84.511263	205	great	european	1
1	Boca	5185a0d0498e2061f617db14	Restaurant	39.102785	-84.511302	95	great	other	5
2	Sleepy Bee Cafe	5a6dea8297cf5a7b38b5e293	Café	39.100055	-84.512272	23	below avg	other	0
3	Orchids at Palm Court	4b1478a3f964a5208ba323e3	New American Restaurant	39.100626	-84.514335	61	above avg	american	3
4	Aster On Fourth	5a4e9cc26bd36b1ecb142000	Cocktail Bar	39.100030	-84.512254	16	poor	bar	2
5	Graeter's Ice Cream	4b4f5355f964a520680127e3	Ice Cream Shop	39.101487	-84.511860	119	great	other	5
6	Nada	4b317901f964a520910725e3	Mexican Restaurant	39.102941	-84.511680	358	great	hispanic	1
7	Maplewood Kitchen and Bar	57680d90498e9e7e4183734a	Breakfast Spot	39.101513	-84.515113	98	great	other	5
8	Abby Girl Sweets	4b4b728ff964a520059c26e3	Cupcake Shop	39.101057	-84.514314	16	poor	other	4
9	FUSIAN	4bed7a8091380f47c9f09f18	Sushi Restaurant	39.102720	-84.512924	66	above avg	asian	3
10	Metropole	50980e2bd63eb33c0a84d7f4	Restaurant	39.103174	-84.511813	85	above avg	other	3

Figure 8

3.6 Data Visualization

Using the original dataframe with the added cluster labels from *3.5*, each individual cluster was visualized on a map using the Folium library, as seen in *Figure 9* below.

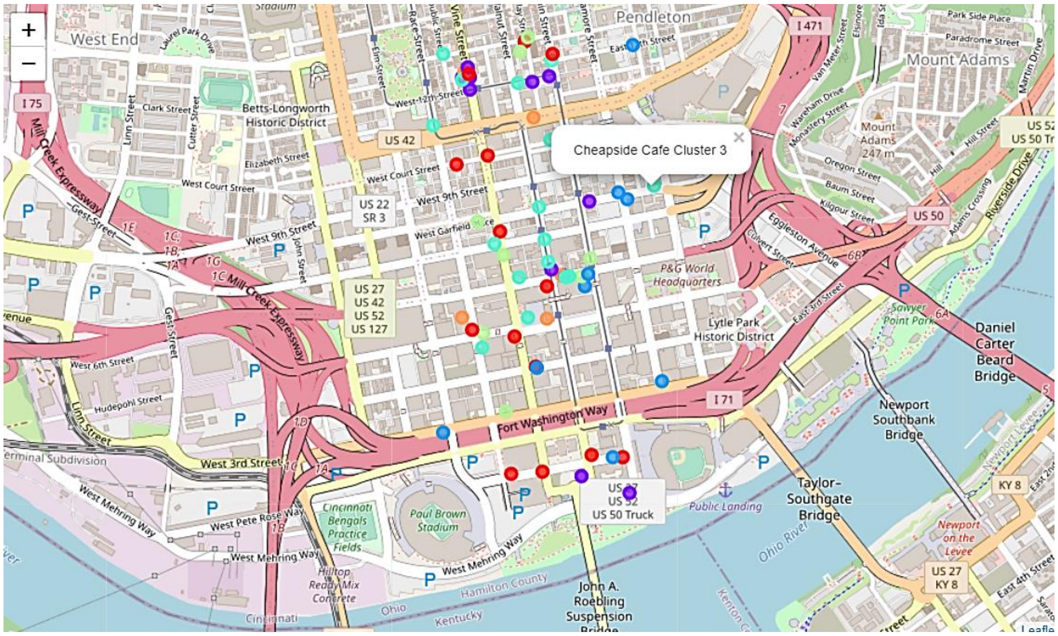

Figure 9

To better understand the distribution of venues in each cluster, a waffle map was also created using *mpatches* from *matplotlib*. This can be seen in *Figure 10* below.

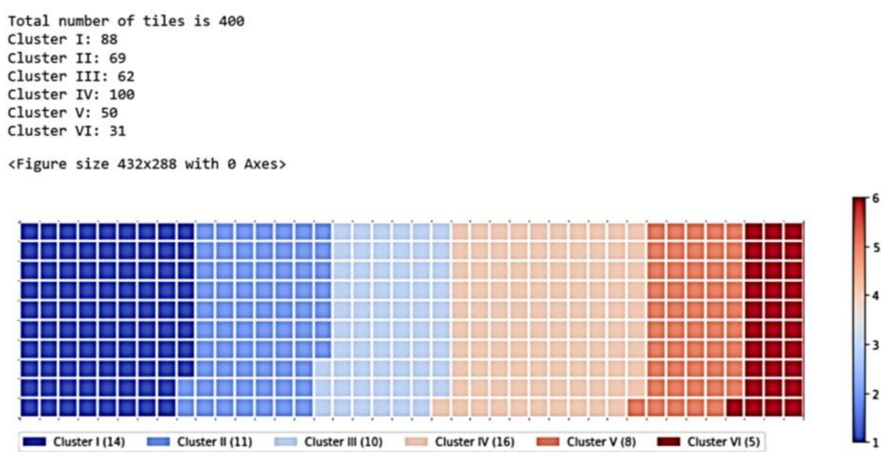

Figure 10

4. Results

The results of clustering algorithm reveal the following 6 distinct clusters (*Figure 11*):

	name	id	categories	lat	lng	likes	likes category	food type	cluster
2	Sleepy Bee Cafe	5a6dea8297cf5a7b38b5e293	Café	39.100055	-84.512272	23	below avg	other	0
11	Mita's	55cd2670498e0624984040e8	Latin American Restaurant	39.101161	-84.514726	39	below avg	hispanic	0
15	Bru Burger Bar	56605066498e467affa580ef	Burger Joint	39.102452	-84.511868	53	below avg	american	0
18	Morton's The Steakhouse	4b17fd3af964a520f1ca23e3	Steakhouse	39.100958	-84.513089	35	below avg	american	0
19	Jean-Robert's Table	4c620363e1621b8d79a72253	French Restaurant	39.104101	-84.513649	32	below avg	european	0
21	Taste of Belgium - The Banks	57f69a75498ee867debabeb9	Bistro	39.096929	-84.512024	51	below avg	bar	0
24	Pies & Pints	585a9629cf44517dc01928d3	Pizza Place	39.096872	-84.513252	29	below avg	european	0
26	Le's Pho and Sandwiches	4feb428aebcabf0f092bfe57	Vietnamese Restaurant	39.106369	-84.514132	25	below avg	asian	0
37	Ruth's Chris Steak House	507a2452e4b0ce42c6bdef75	Steakhouse	39.097418	-84.510127	37	below avg	american	0
39	Queen City Exchange	57f8fbf4498ecadc3af56807	Bar	39.106092	-84.515353	22	below avg	bar	0
48	Condado	5ae22dbb4c954c002cead32b	Mexican Restaurant	39.097370	-84.508927	21	below avg	hispanic	0
49	Macaron Bar	54065433498e90c0cac1960c	Bakery	39.109393	-84.511661	31	below avg	other	0
51	Longfellow	589f74662c55ec7cf24bb656	Bar	39.109734	-84.512704	26	below avg	bar	0
57	Abigail Street	4ea47d609911214fd1adc8cb	Tapas Restaurant	39.108788	-84.514857	41	below avg	hispanic	0

	name	id	categories	lat	lng	likes	likes category	food type	cluster
0	Sotto	5154c81ae4b0c54802cba3c7	Italian Restaurant	39.102797	-84.511263	205	great	european	1
6	Nada	4b317901f964a520910725e3	Mexican Restaurant	39.102941	-84.511680	358	great	hispanic	1
20	Arnold's Bar & Grill	4b3d1320f964a520588d25e3	Bar	39.104977	-84.510209	140	great	bar	1
28	Yard House	514b70e8e88971e7503c73b1	American Restaurant	39.096804	-84.510510	348	great	american	1
34	Moerlein Lager House	4bc5e8b3f360ef3b98c3da2d	Gastropub	39.096311	-84.508673	509	great	bar	1
36	Taste of Belgium OTR	4e33323afa76f00388beb648	American Restaurant	39.108309	-84.514806	315	great	american	1
44	Senate Restaurant	4b7c72d2f964a5200c942fe3	Gastropub	39.108696	-84.514819	177	great	bar	1
46	Bakersfield	4f2dde37e4b007725af12e10	Taco Place	39.108687	-84.515130	297	great	hispanic	1
50	A Tavola	4da8cef45da3ba8a47624c59	Italian Restaurant	39.108973	-84.514891	160	great	european	1
58	rhinehaus	507ad9a5e4b04c2b97f7a51b	Sports Bar	39.108544	-84.512414	101	great	bar	1
60	Japp's Since 1879	4ce30b2b7e9b721e823334f1	Cocktail Bar	39.108726	-84.511784	123	great	bar	1

	name	id	categories	lat	lng	likes	likes category	food type	cluster
4	Aster On Fourth	5a4e9cc26bd36b1ecb142000	Cocktail Bar	39.100030	-84.512254	16	poor	bar	2
16	Chick-Fil-A	5bc8c19ce0c0c9002c9c5aeb	Fast Food Restaurant	39.102455	-84.510390	0	poor	american	2
27	Wahlburgers	5a6a03209e3b651f15217a6b	Burger Joint	39.102808	-84.510259	14	poor	american	2
35	Silverglades on 8th	4b7a1290f964a520cb212fe3	Deli / Bodega	39.105252	-84.509121	18	poor	european	2
41	Cuban Pete Sandwiches	5527f715498ea7c14e73bdce	Cuban Restaurant	39.106475	-84.511285	13	poor	hispanic	2
45	Restaurant L	57e1a32e498e8e1645bd1d5e	French Restaurant	39.099618	-84.507433	9	poor	european	2
54	Crown Republic Gastropub	5b2b062c464d65002cc786dd	Gastropub	39.105043	-84.508747	6	poor	bar	2
55	Kitty's Sports Grill	560c708e498e541b90075039	Sports Bar	39.098082	-84.515842	10	poor	bar	2
59	Boomtown Biscuits & Whiskey	59e6b5ac7564f74cb2b6737b	American Restaurant	39.109662	-84.508557	18	poor	american	2
63	The Stretch	5830d14765e7c7635dfebb23	Bar	39.097372	-84.509334	7	poor	bar	2

	name	id	categories	lat	lng	likes	likes category	food type	cluster
3	Orchids at Palm Court	4b1478a3f964a5208ba323e3	New American Restaurant	39.100626	-84.514335	61	above avg	american	3
9	FUSIAN	4bed7a8091380f47c9f09f18	Sushi Restaurant	39.102720	-84.512924	66	above avg	asian	3
10	Metropole	50980e2bd63eb33c0a84d7f4	Restaurant	39.103174	-84.511813	85	above avg	other	3
13	Via Vite	4b44357ef964a520b5f225e3	Italian Restaurant	39.101519	-84.512572	83	above avg	european	3
14	Jeff Ruby's Steakhouse	4b92ff3af964a520842d34e3	Steakhouse	39.103869	-84.511959	80	above avg	american	3
22	Knockback Nat's	4b47f539f964a5201b4526e3	Bar	39.103713	-84.513886	90	above avg	bar	3
29	Mr. Sushi	4b4e4102f964a520f7e626e3	Japanese Restaurant	39.102716	-84.511143	55	above avg	asian	3
31	Cheapside Cafe	53760551498e68b83d4e746c	Café	39.105442	-84.507739	91	above avg	other	3
32	Tom & Chee	4cf7b1015c0e6a31b4b74099	Sandwich Place	39.106820	-84.511710	92	above avg	other	3
38	Taqueria Mercado	4bd5f1944e32d13ad0c6c180	Mexican Restaurant	39.104830	-84.512175	83	above avg	hispanic	3
40	Revolution Rotisserie & Bar	54f49879498ee35c902f8ed1	American Restaurant	39.107260	-84.516241	55	above avg	american	3
42	Gomez Salsa	52f418c211d209ba43106c1e	Mexican Restaurant	39.108532	-84.512971	65	above avg	hispanic	3
43	Goodfellas Pizzeria	54a236e8498ebc7d8ce5f804	Pizza Place	39.109171	-84.511684	69	above avg	european	3
52	Sundry and Vice	55156c74498ece57a4ab8508	Cocktail Bar	39.109378	-84.515852	68	above avg	bar	3
56	Krueger's Tavern	548a3ae3498e96c3ac6dbf16	Gastropub	39.108610	-84.515140	84	above avg	bar	3
61	Igby's	50481d09e4b0d1f2d8d750aa	Bar	39.102769	-84.511007	74	above avg	bar	3

	name	id	categories	lat	lng	likes	likes category	food type	cluster
8	Abby Girl Sweets	4b4b728ff964a520059c26e3	Cupcake Shop	39.101057	-84.514314	16	poor	other	4
12	Total Juice Plus	4b7d68a2f964a520cdbc2fe3	Juice Bar	39.103375	-84.513543	14	poor	other	4
17	Cafe De Paris	4ba8cd26f964a52007f039e3	Café	39.104384	-84.514501	9	poor	other	4
23	Silver Ladle	4f5fd1b9e4b0005742f4f53e	Sandwich Place	39.102741	-84.510546	18	poor	other	4
25	Lola's	55f31264498ebc576e3e6bee	Coffee Shop	39.098720	-84.513461	8	poor	other	4
33	Izzy's	4b9a70f9f964a5208cb535e3	Sandwich Place	39.103291	-84.510168	18	poor	other	4
53	Brown Bear Bakery	515d1365e4b0d2cd6d15e8e6	Bakery	39.109888	-84.512572	16	poor	other	4
62	SugarSnap!	502fcd7ce4b04de6f3c87eb2	Cupcake Shop	39.109456	-84.512902	12	poor	other	4

	name	id	categories	lat	lng	likes	likes category	food type	cluster
1	Boca	5185a0d0498e2061f617db14	Restaurant	39.102785	-84.511302	95	great	other	5
5	Graeter's Ice Cream	4b4f5355f964a520680127e3	Ice Cream Shop	39.101487	-84.511860	119	great	other	5
7	Maplewood Kitchen and Bar	57680d90498e9e7e4183734a	Breakfast Spot	39.101513	-84.515113	98	great	other	5
30	Coffee Emporium	4b460755f964a5207d1426e3	Coffee Shop	39.107498	-84.512390	281	great	other	5
47	1215 Wine Bar & Coffee Lab	4f349e4ee4b0debe1e46886a	Coffee Shop	39.108851	-84.515014	103	great	other	5

Figure 11

Cluster 2 and Cluster 4 consist of all the “poor” restaurants, Cluster 0 consists of all the “below average” restaurants, and Cluster 3 consists of all the “above average” restaurants. Of notable interest to tourists, locals, food critics, and restaurants are Cluster 1 and Cluster 5, both of which consist of “great” restaurants. Cluster 1 features food from various ethnicities (“european,” “hispanic,” and “american”) while Cluster 5 features alternative restaurants such as ice cream, coffee shops, and breakfast spots. Depending on which type of food a stakeholder is craving, they may choose between Cluster 1 and Cluster 5.

5. Discussion

Surprisingly, restaurants are evenly distributed over all 6 clusters despite the overwhelming right skew in the Foursquare likes data shown in *Figure 3*. While the upper echelon of “great” restaurants seems to be limited, there are roughly equal proportions of “poor,” “average,” and “above average” restaurants in Cincinnati. Specifically, there are very few “great” alternative restaurants in Cincinnati, a point of interest to potential entrants. Thus, our recommendations for the relevant stakeholders are as following:

1. Locals/Tourists/Restaurants: Explore restaurants in Cluster 1 and Cluster 5.
2. Potential Entrants: Seek to establish a restaurant such that it falls in Cluster 5.

Of course, the analysis presented here is far from complete. Only 2 factors are analyzed: the number of Foursquare users who liked a specific restaurant and the broad categorization of each restaurant. Several other factors are also of interest: the general price-point of each restaurant, each restaurant’s proximity to other tourist attractions, and each restaurant’s Michelin rating, to name just a few. While this data was not immediately available for this analysis, with the use of web scraping techniques, it can be obtained for future analyses.

6. Conclusion

As diverse as the food scene in Cincinnati is, the upper echelon of restaurants is limited. This is to the benefit to tourists, locals, and food critics who may find it difficult to determine which restaurants are the most worth exploring. Specifically, for all relevant parties, the restaurants in Cluster 1 and Cluster 5 should be of special interest. Furthermore, as the food scene in Cincinnati continues to expand, restaurants must seek to position themselves in the upper echelon of restaurants, the segment that is the least saturated in the Cincinnati area.