Neighborhood Venues Analysis for Café & Bar Crawls

IBM Data Science Capstone Project

Anthony Canterbury

September 29th, 2019

Table of contents:

1 Introduction

Within the city of Cincinnati, OH a nonprofit organization wants to put on a series of outreach events which take donating participants to multiple neighborhood cafés and bars to socialize. The goal of this "crawl" is not only to take donations but to make the charity or issue as visible as possible in these neighborhoods.

To make these "night & day crawls" successful the collections of venues in these locations must meet several criteria:

- First the location must have at least two cafés and two bars in proximity. Seven venues max per cluster. *Can seven be considered within walking distance (less than 0.5 miles apart)?*
- The venues must be open sometime on Saturdays between 3pm and 10pm.
- The venues must be popular (well-reviewed).
- The venues must not be expensive.

The problem to answer is which clusters of venues in these neighborhoods are most ideal to recommend for partnering with for these events based on available data.

The methodology in the analysis could be generalized to build recommendations for most types of outings within any city.

2 Data

The data required for neighborhood venue cluster analysis:

- List of Neighborhoods within Cincinnati, OH. A list of the 50 neighborhoods with approximate geographical boundaries comes from the "Cincinnati Area Geogrpahic Information System" https://data-cagisportal.opendata.arcgis.com/datasets/cincinnati-sna-boundary
- List of neighborhood venues with hours, reviews (ratings and likes) and approximate prices. This data will be supplied by the Foursquare API.

While bars are easy to filter for, a café is a little more subjective. For this analysis we will consider anything labeled a café, coffee shop, teahouse, or dessert shop (e.g. donut shop, ice cream shop, pastry shop) a café. Beyond basic cleaning and formatting for all datasets some of the data from the Foursquare API will most certainly need to be predicted due to missing data for prices and review ratings. Missing data is expected since new venues appear regularly in many of the neighborhoods.

2.1 Parameters

To gather and analyze the data appropriately given the questions outlined in the introduction we start with several parameters and their client values:

- **MAX_PRICE = 3**, the Foursquare ranking for prices range 1 to 4.

- **MAX_WALK = 0.8**, roughly 0.5 miles in kilometers.
- **VENUE_PRIME = ['bar', 'pub', 'brewery', 'lounge']**, bar type patterns to filter categories.
- **VENUE_SECONDARY = ['caf', 'coffee', 'tea', 'desert', 'ice cream', 'donut']**, cafe type patterns to filter categories (note 'caf' is used because of different spellings).
- **MIN_PRIME = 2**, minimum number of venues that match bars.
- **MIN_SECONDARY = 2**, minimum number of venues that match cafes.
- **MAX_VENUES = 7**, maximum number of venues per crawl event.
- **WEEK_DAY = 6**, Saturday.
- **START_TIME = 1500**, 3 PM.
- **END_TIME = 2200**, 10 PM.
- **PRIORITY_ORDER = {'Rating': 4, 'Count': 2, 'Likes': 1**}, gives weighting scale for the various deciding features

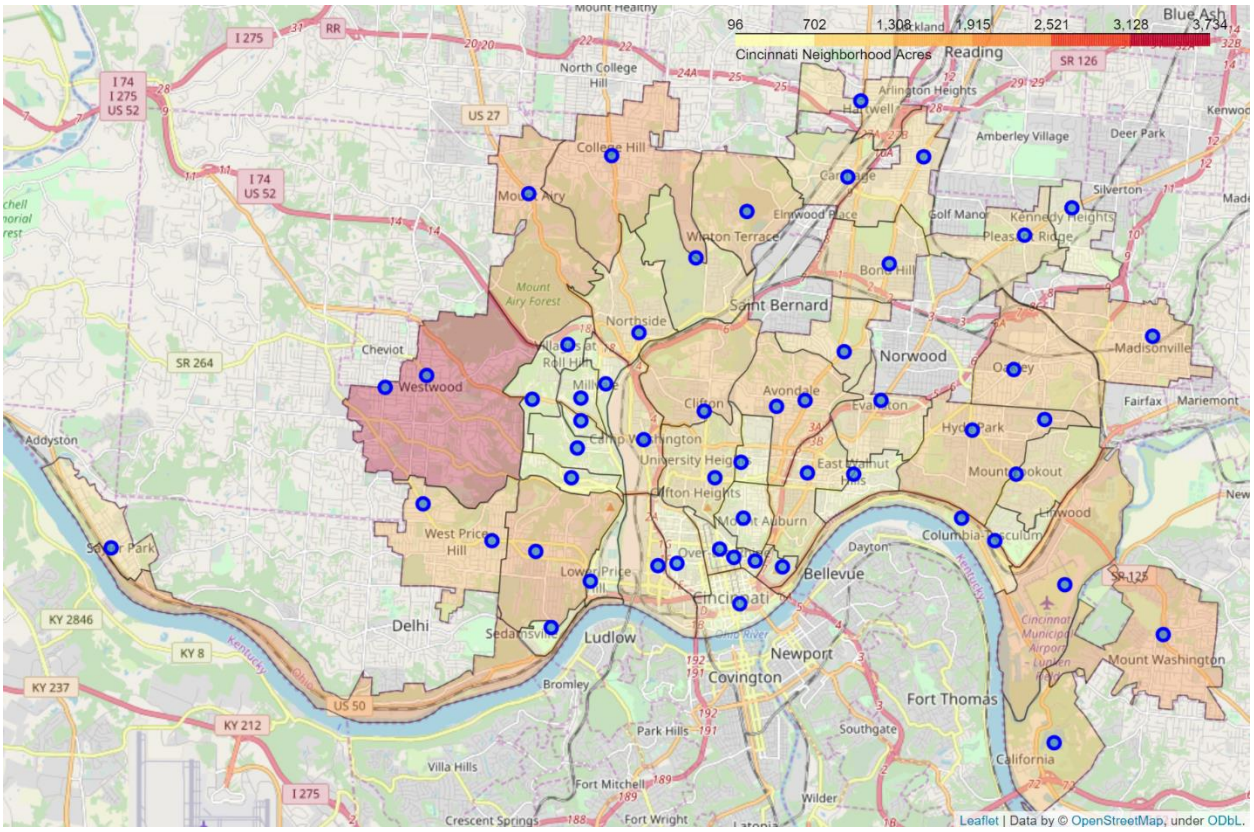
2.2 Acquiring and Cleaning the Data

2.2.1 Cincinnati Neighborhoods Data

We start gathering the data using the neighborhood data from the "Cincinnati Area Geogrpahic Information System". There are two geographic API datasets that are important the SNA Boundary data for the city neighborhoods, and the Business Districts. These datasets give land area in acres and boundary coordinates which we need to parse for the center point of the business district.

In the case of multiple business districts for a neighborhood we collect each one in a different district. If no district was found, I took the center point of the neighborhood itself. I only removed the neighborhood of Riverside due to a lack of business district and lack of businesses.

Choropleth map of the Neighborhoods with Business District Centers



The coloring by acres is just for aesthetics.

2.2.1 Venue Name, Location, and Category Data

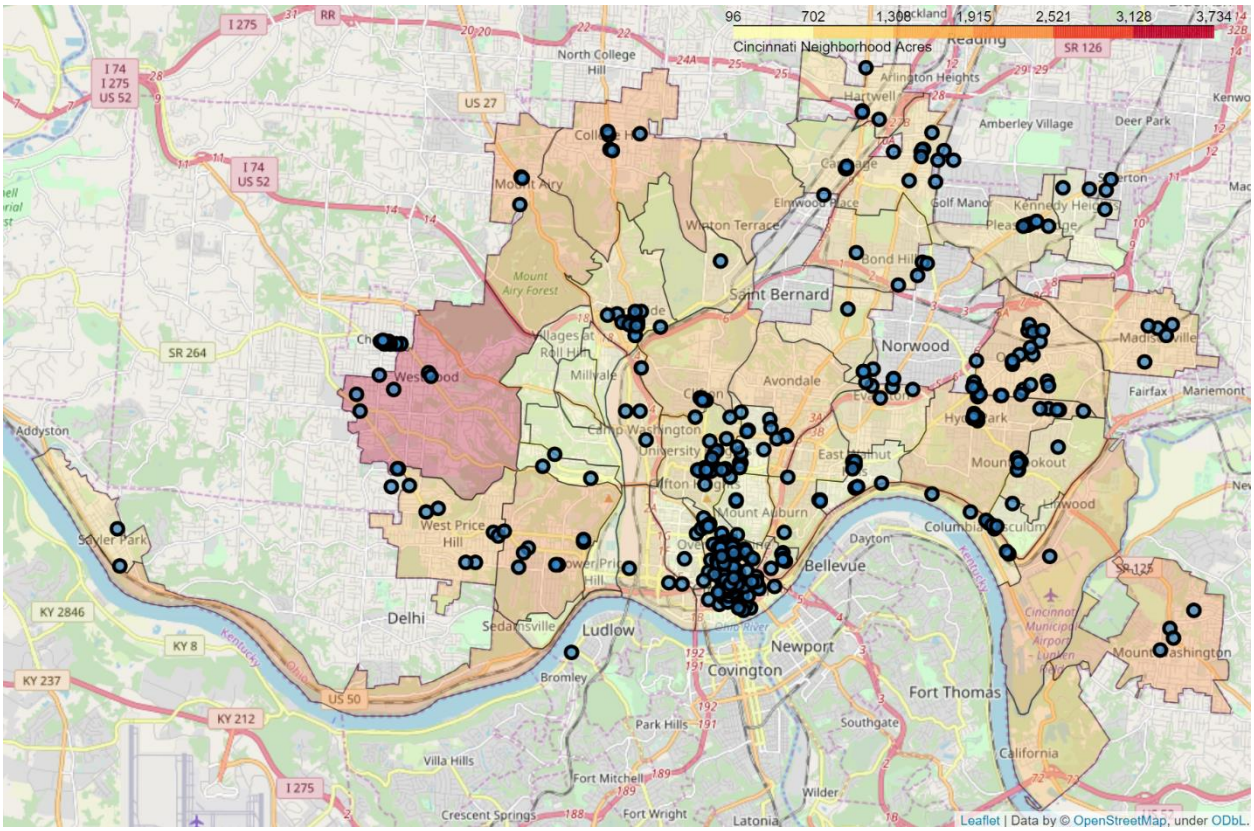
Using the center points, I locate all venues that match our criteria from the Foursquare API. I used the "venues-explore" endpoint with the parameters: latitude, longitude, radius = 1000 meters, limit = 100, section = drinks and coffee. Drinks and coffee matched all our venue criteria from researching the data. We are only concerned with the geo-location, name, and category from these requests. I examined the unique venue categories and removed several categories that did not fit our criteria (e.g. cafeterias).

With this new table we have 381 unique venues, sample below.

	Neighborhood	BusinessDistrict	NeighborhoodLatitude	NeighborhoodLongitude	VenueName	VenueId	VenueLatitude	VenueLongitude	VenueCategory
0	Linwood	0	39.104213	-84.415924	Dennert H Distribtg	4f324944198 36c91c7c8b7f7	39.108777	-84.421232	Wine Bar
1	East Walnut Hills	1	39.128889	-84.476823	The Woodburn Brewery & Taproom	55461bf6498eac118325e62e	39.129030	-84.476892	Beer Bar
2	East Walnut Hills	1	39.128889	-84.476823	Myrtle's Punch House	5473d783498ec0bbca9021d6	39.124276	-84.476130	Cocktail Bar
3	East Walnut Hills	1	39.128889	-84.476823	The Growler House	545d54ab498ea427d9af9d2d	39.129763	-84.477778	Bar
4	East Walnut Hills	1	39.128889	-84.476823	BrewRiver Gastropub	4fea02ede5e8dfeeb65b5000	39.121758	-84.475027	Gastropub
5	East Walnut Hills	1	39.128889	-84.476823	The Skunk Lounge	5182cdbd498e1c1b38b47f1c	39.124213	-84.476246	Lounge
6	East Walnut Hills	1	39.128889	-84.476823	Cliche	5d6459abca17630008abf539	39.123820	-84.477040	Bar

• • •

Map of Venues



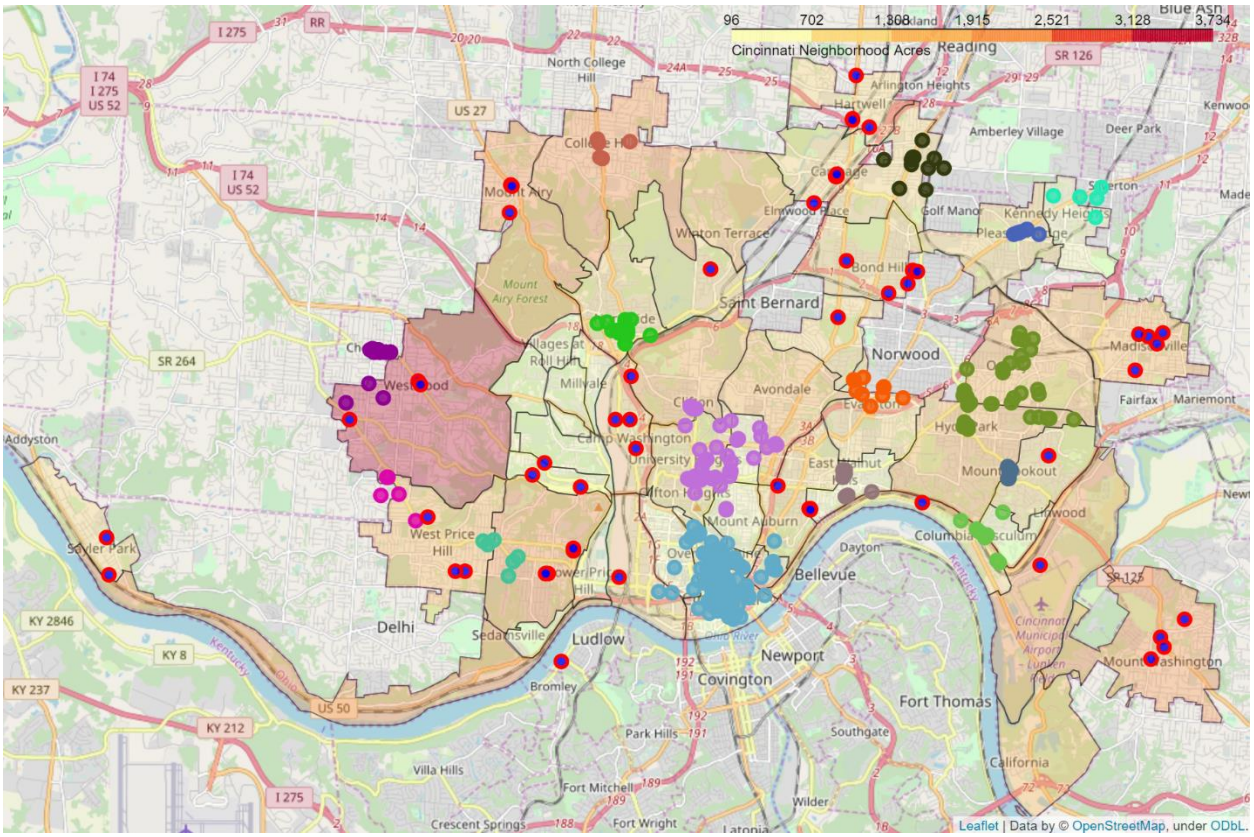
To restrict the amount of premium API calls to Foursquare the data was store in CSV files, and the data was cleaned for venues outside the criteria before each API request.

Removing unneeded venues before collecting individual venue details was performed using clustering amongst other technics. Specifically, I removed outliers from a Density-Based Spatial Clustering of Applications with Noise (DBSCAN). A DBSCAN is a density-based clustering non-parametric algorithm that groups closely packed points. The DBSCAN grouped the venue

geographical points based on our maximum walking distance parameter (MAX_WALK), and the minimum number of venues per event crawl.

In the map below, the clusters/groups are color coded with red-outlined blue points as outliers.

Highlighted Venue Distance Outliers Before Removal



After removing the outliers, we are left with 325 venues within 15 clusters.

2.2.2 Venue Open Hours Data

Are these venues open during the day and time of the event? To gather the hours of operations I used the “venue-hours” endpoint of the Foursquare API. Only roughly a third of the venues had hours of operations in the returned data. Filling in the gaps took some unsupervised machine learning.

Event times: On Saturday 3pm to 10pm.

K-nearest neighbors is an instance-based learning, or lazy learning, algorithm for regression and classification. KNN was employed to determine if the remaining empty venue hours would be open or closed. The features that make the most sense for determining if a venue is open during a particular time period are their category/type of venue and the cluster they belong to. So, we assume that venues in the same general location and similar type would keep similar hours.

Note: For a more detailed analysis the individual hours could have been determined. But the criteria did not call for it.

The accuracy of the model is then measured by the Jaccard similarity coefficient and F1 score which are rated from 0 (worst) to 1(perfect) score for similarity to true values and precision and recall to true value respectively. Modeling and fitting these values give us a pretty good accuracy:

- Jaccard Score ~ 0.915
- F1 Score ~ 0.875

With this data we can determine that nine of the venues are most likely closed on the scheduled event time. This leaves us with 316 venues. Re-clustering doesn't affect the number or position of the clusters.

2.2.3 Venue Ratings, Likes, and Prices Data

The last dataset to collect was pulled from the "venue-details" endpoint of the Foursquare API. This dataset provides each venue's rating, likes, and price range. Ratings are ranked from 0 to 10. Likes are a basic count. Price is a range from 1 to 4, which symbolize cheap, moderate, expensive, and very expensive. Again, this data contains missing values that had to be handled.

Predicting the ratings with the current dataset features is pointless so we'll be filling in the missing data with the median for each cluster when we need to use the ratings for recommendations.

The price of the establishments could be very important but thankfully we are only missing seven venue's prices.

Event Max Price: 3 or expensive.

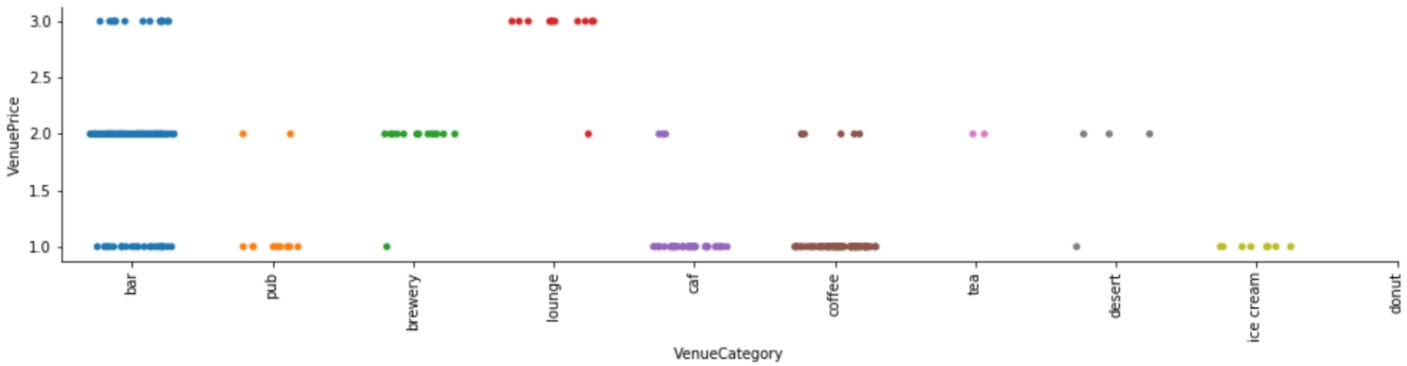
From a correlation matrix of the pertinent venue features we can see that category is the only marginal effect on Price.

Venue Feature Correlations

	DbCluster	VenueCategory	VenueRating	VenueLikes	VenuePrice
DbCluster	1.000000	-0.006894	-0.174062	-0.133704	-0.068473
VenueCategory	-0.006894	1.000000	0.050535	-0.010160	-0.532074
VenueRating	-0.174062	0.050535	1.000000	0.498003	0.063367
VenueLikes	-0.133704	-0.010160	0.498003	1.000000	0.007484
VenuePrice	-0.068473	-0.532074	0.063367	0.007484	1.000000

For a more detailed look at the relationship between categories and prices we create a categorical scatterplot.

Venue Category to Price Scatterplot



From the plot we can infer that there is a strong connection between many of the categories and prices. The bar is the only mixed case here. Again we find the KNN model to be the classifier for the job although it's not the only one that could work.

Using the KNN model fit we achieve these decent accuracies:

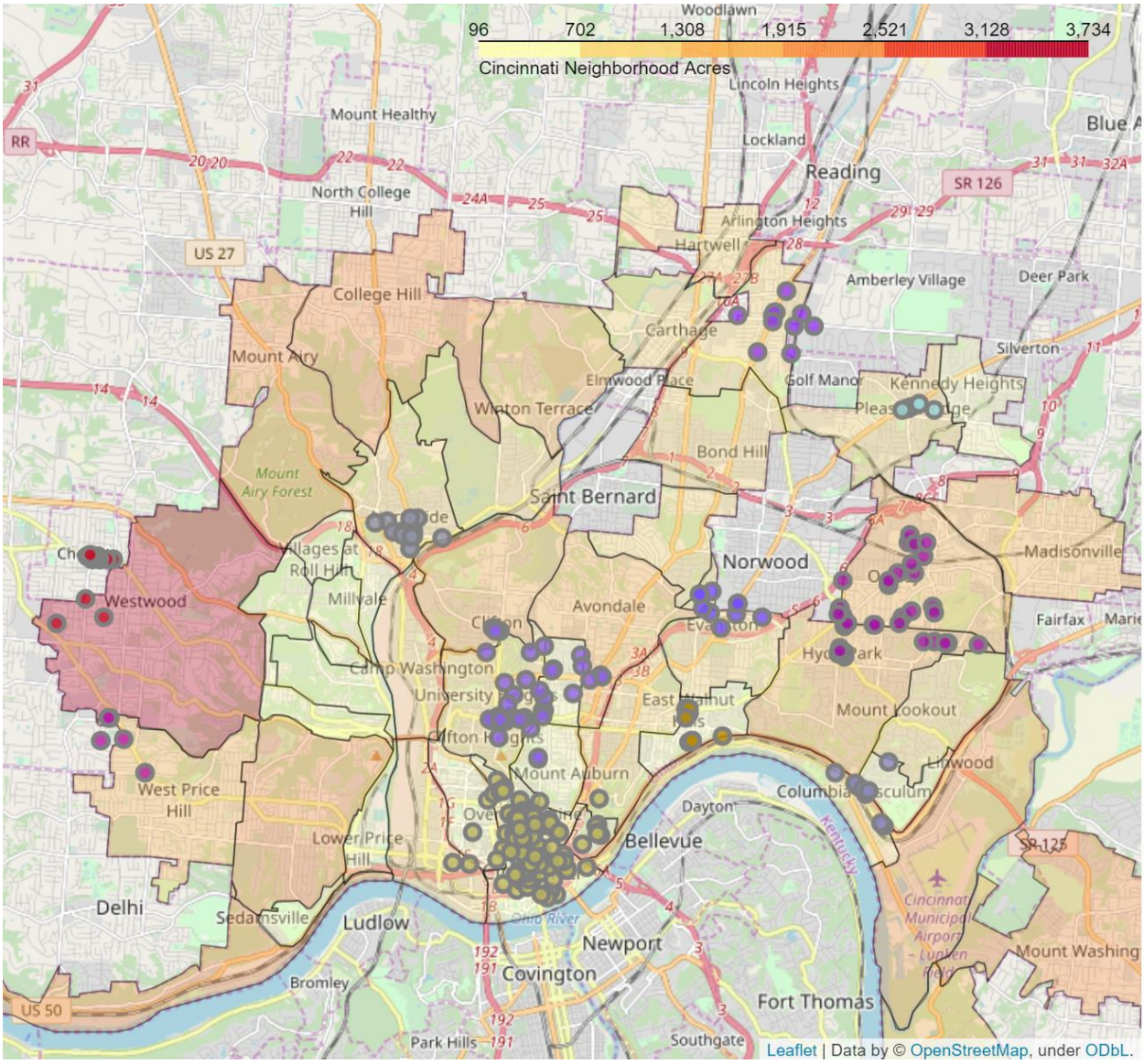
- Jaccard Score ~ 0.844
- F1 Score ~ 0.812

There are no scores outside our range of prices for the event so we are left with 316 venues in our 15 clusters.

2.3 Final Cleaning of the Data

We have one final criteria from our client parameters to address, minimum primary and secondary categories. For any cluster there must be at least two cafes and two bars. After removing clusters that don't have the minimum categories, we use a DBSCAN on the remaining eleven clusters.

Cincinnati Valid Venue Crawl Group Map



Although not employed for this study, we could improve the dataset values by gathering from multiple API sources (e.g. Google Places, Yelp).

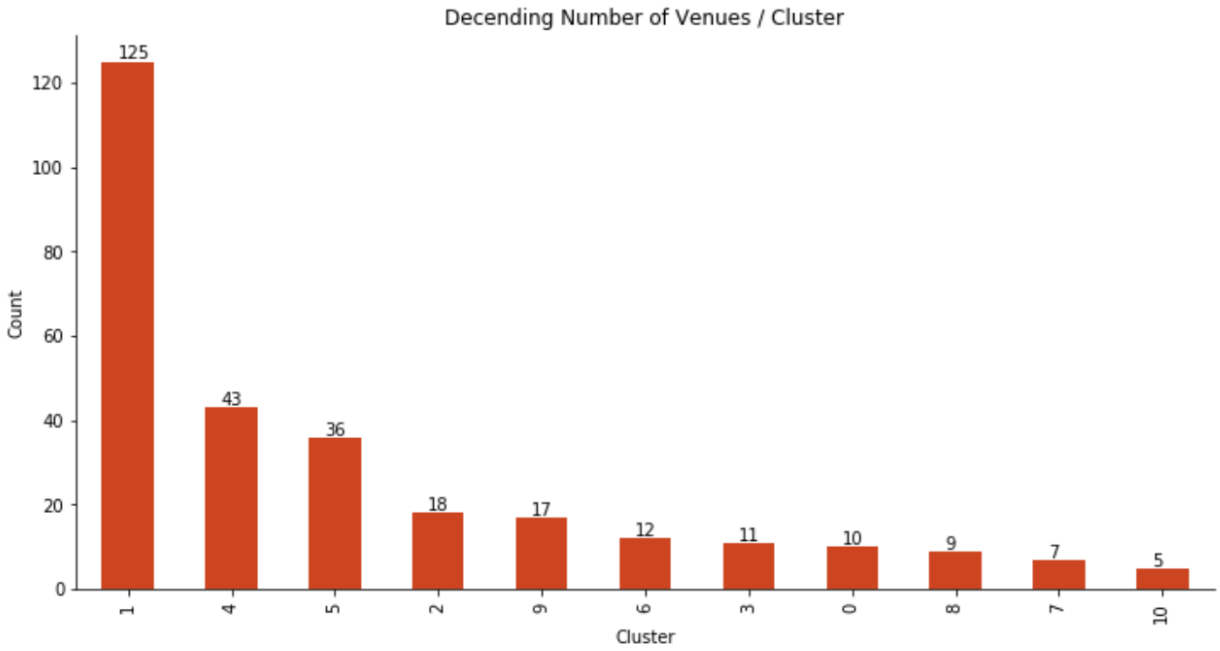
3 Data Analysis

As was expected and you can see from the last map the downtown and adjacent neighborhoods have the most qualifying venues for the crawl. But now it's time to test quality over quantity. Let's list the neighborhoods in each cluster.

Cluster Neighborhoods

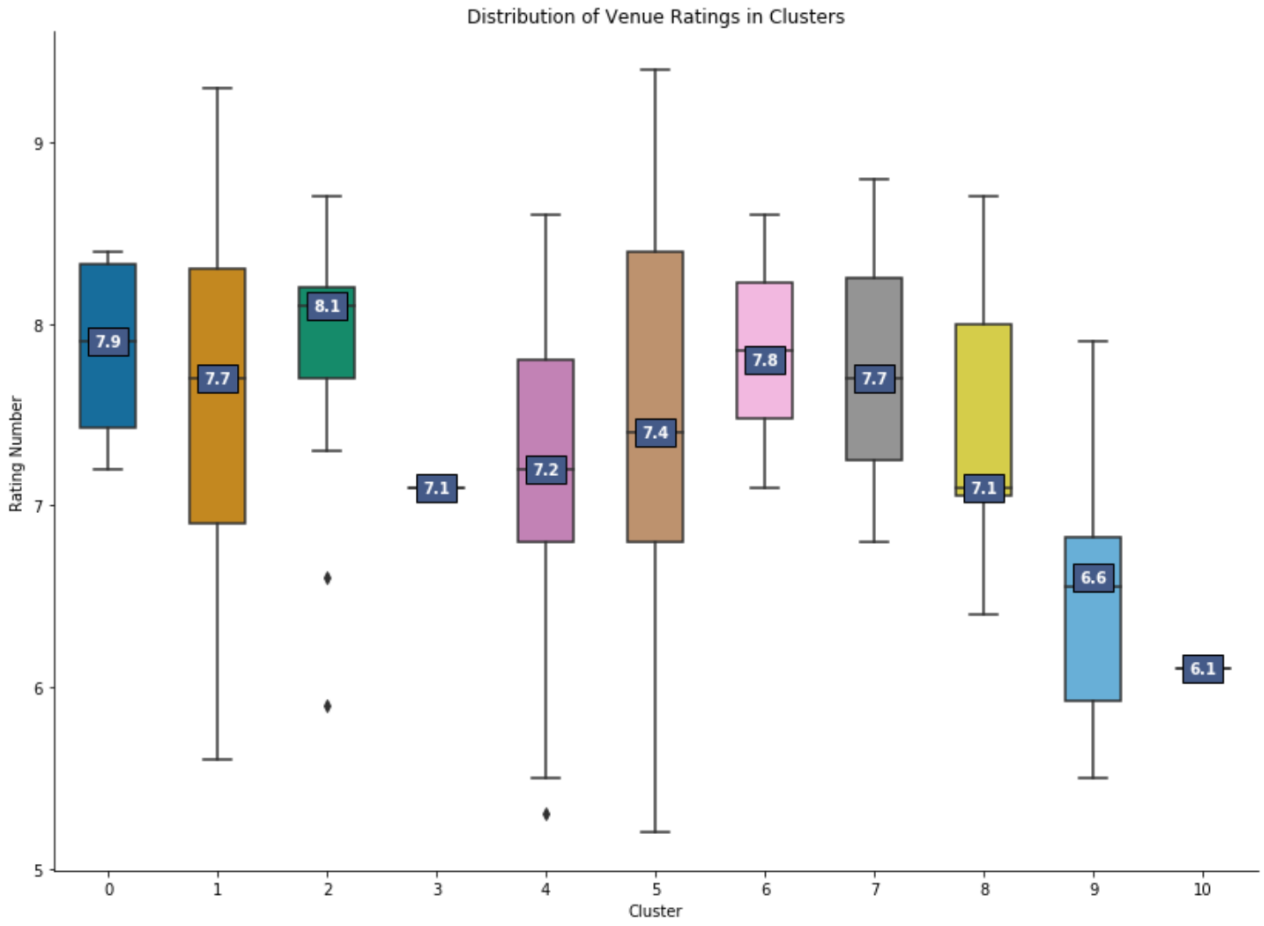
DbCluster	Neighborhoods
0	[East Walnut Hills]
1	[Queensgate, West End, Mt. Auburn, Downtown, Over-the-Rhine, Pendleton]
2	[Northside]
3	[Roselawn]
4	[Mt. Auburn, Avondale, Clifton, Clifton Heights, Corryville, Walnut Hills]
5	[Hyde Park, Oakley]
6	[Evanston, North Avondale]
7	[Pleasant Ridge, Kennedy Heights]
8	[East End, Columbia Tusculum, Mt. Lookout]
9	[Westwood]
10	[West Price Hill]

Surprisingly the only overlap of neighborhoods occurs for between cluster one and four with Mt. Auburn. Several clusters bleed into outside neighborhoods that are not incorporated in the city of Cincinnati. The main reason for the client to pick the city itself could be for various reasons but we'll assume it's ok to allow a little bleed outside the lines.



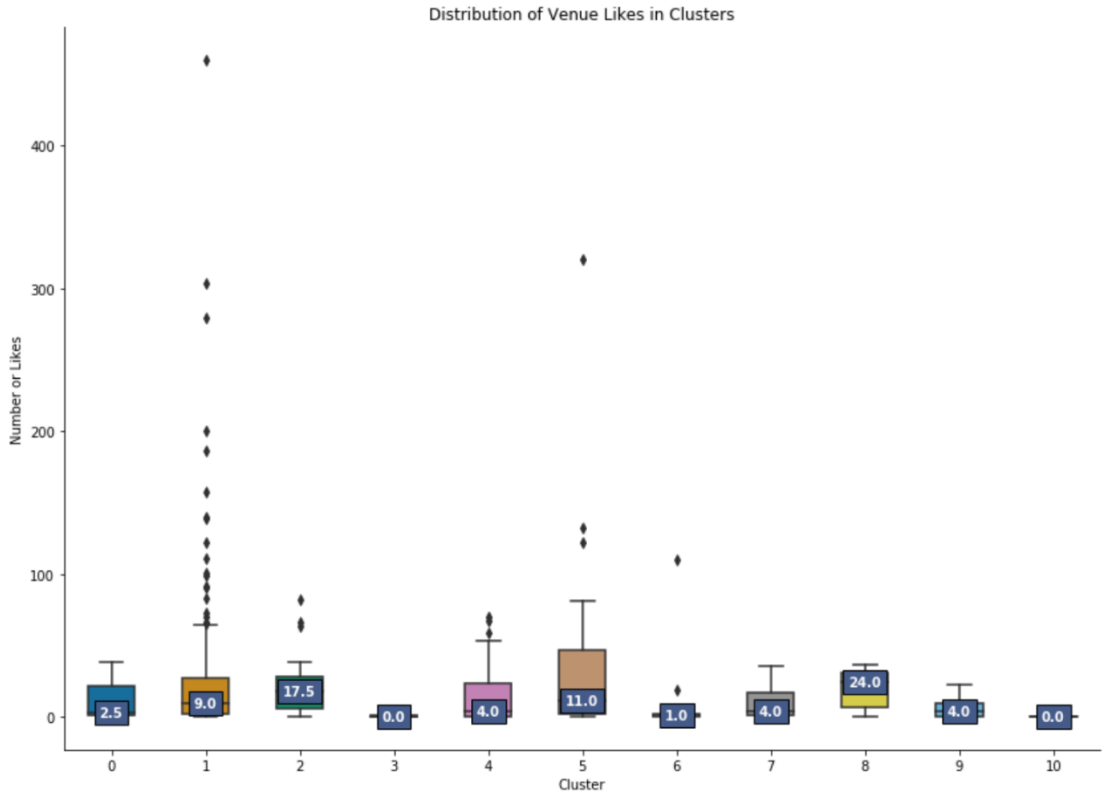
We can see here that the number of venues in the downtown cluster is close to more than all the other clusters combined. Cluster 4, Clifton, and Cluster 5, Hyde Park, are the only two others with a significant number of venues.

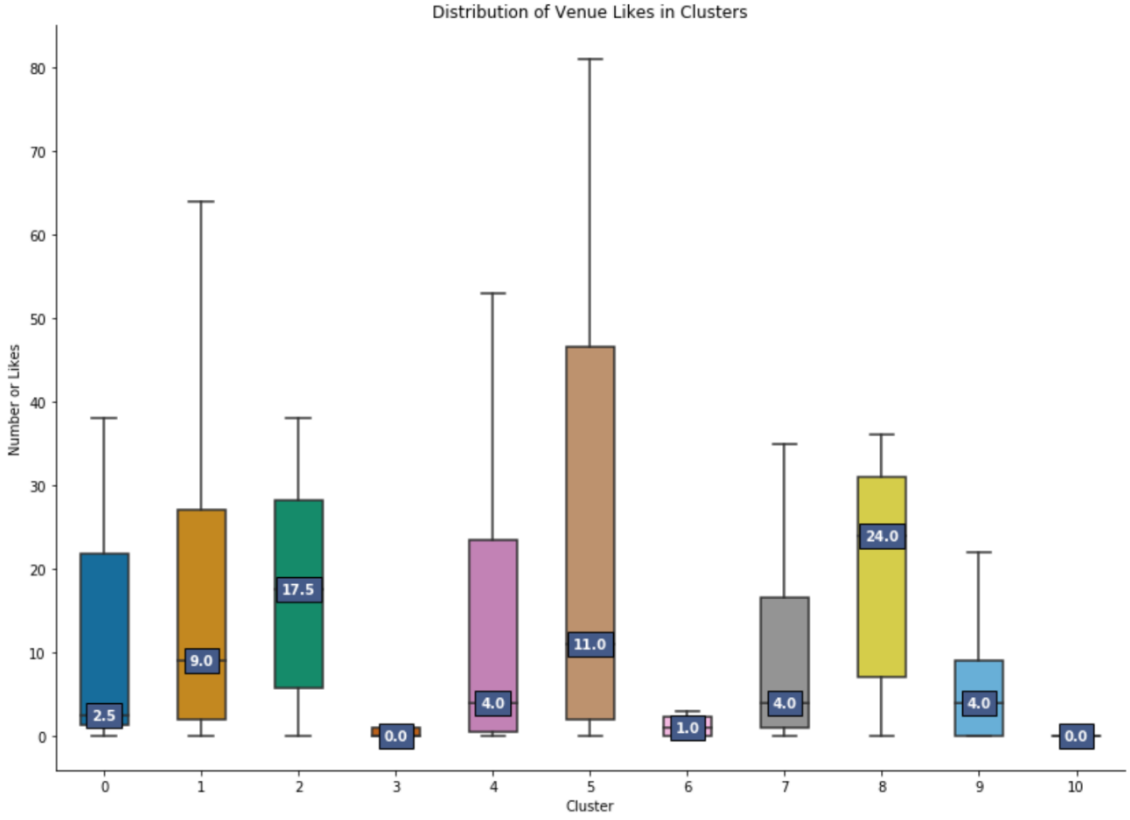
To analyze the distribution of ratings in the clusters we should use a boxplot since the quartiles, especially the median, are the most important values when it comes to this type of analysis.



The accuracy of the ratings should be considered pretty good with values like this. All the median ratings fall between six and eight which is what one might expect. The highs/maxes and lows/mins are far apart when there are a significant number of venues. Nothing is a perfect 10. There are only a couple outliers.

Venue counts of likes are a different story completely. They vary wildly probably due to businesses gaming the system (e.g. discounts for likes). In these next two boxplots the first includes outliers while the last does not.


Distribution of Venue Likes in Clusters
Number or Likes
Cluster
0
1
2
3
4
5
6
7
8
9
10
2.5
9.0
17.5
0.0
4.0
11.0
1.0
4.0
24.0
4.0
0.0


Distribution of Venue Likes in Clusters
Number or Likes
Cluster
0
1
2
3
4
5
6
7
8
9
10
2.5
9.0
17.5
0.0
4.0
11.0
1.0
4.0
24.0
4.0
0.0

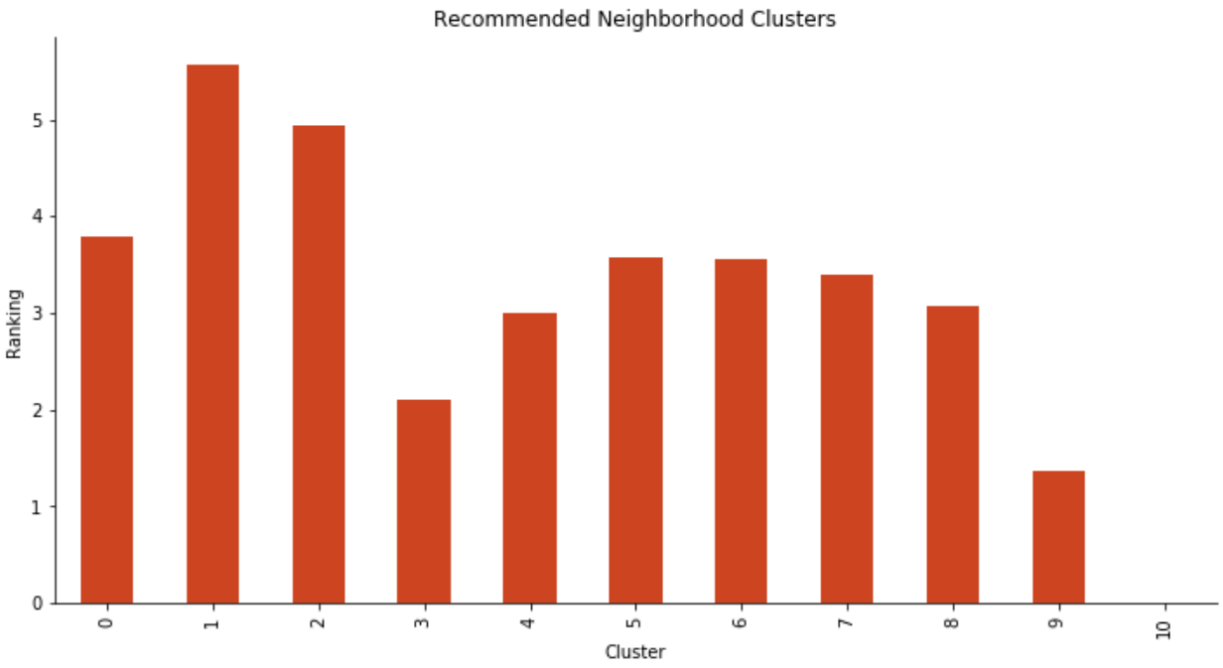
Not a lot can be gleaned from the venue likes but we can tell that this is not a great feature to base your recommendations off of unless maybe the recommendation are linked to a group of people with like interests.

To summarize the data analysis:

- The largest cluster of venues falls into the Downtown neighborhoods, which includes Queensgate, West End, Mt. Auburn, Downtown, Over-the-Rhine, and Pendleton.
- The ratings medians are similar across all cluster which is hoped for, most venues get ratings in the 7s (so a C grade). The overall distribution of the ratings does vary
- The venue likes count varies wildly and is not the best feature to base recommendations off.

4 Recommendation Results

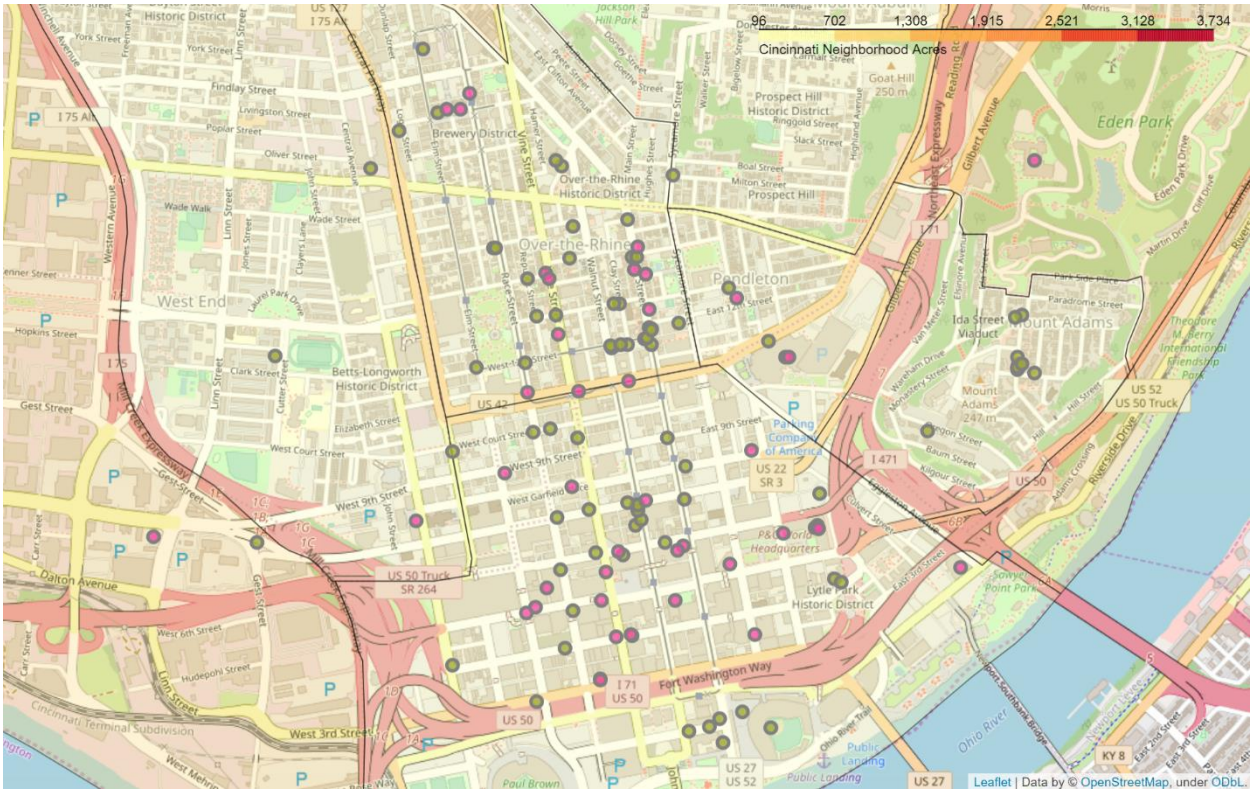
Based on the priority rankings (**PRIORITY_ORDER**) which weighs venue groups by "ratings" twice as important as "count" twice as important as "likes" we now view their rankings.



Higher the ranking the better the group/cluster of venues is. The Downtown just barely edges out Northside. The Downtown wins primarily based on its number of venues, where Northside has the most consistent high rated venues.

If we map the distribution of primary (bars) and secondary (cafes) venues for the Downtown group we can see that its fairly evenly distributed.

Downtown Bar and Café Distribution



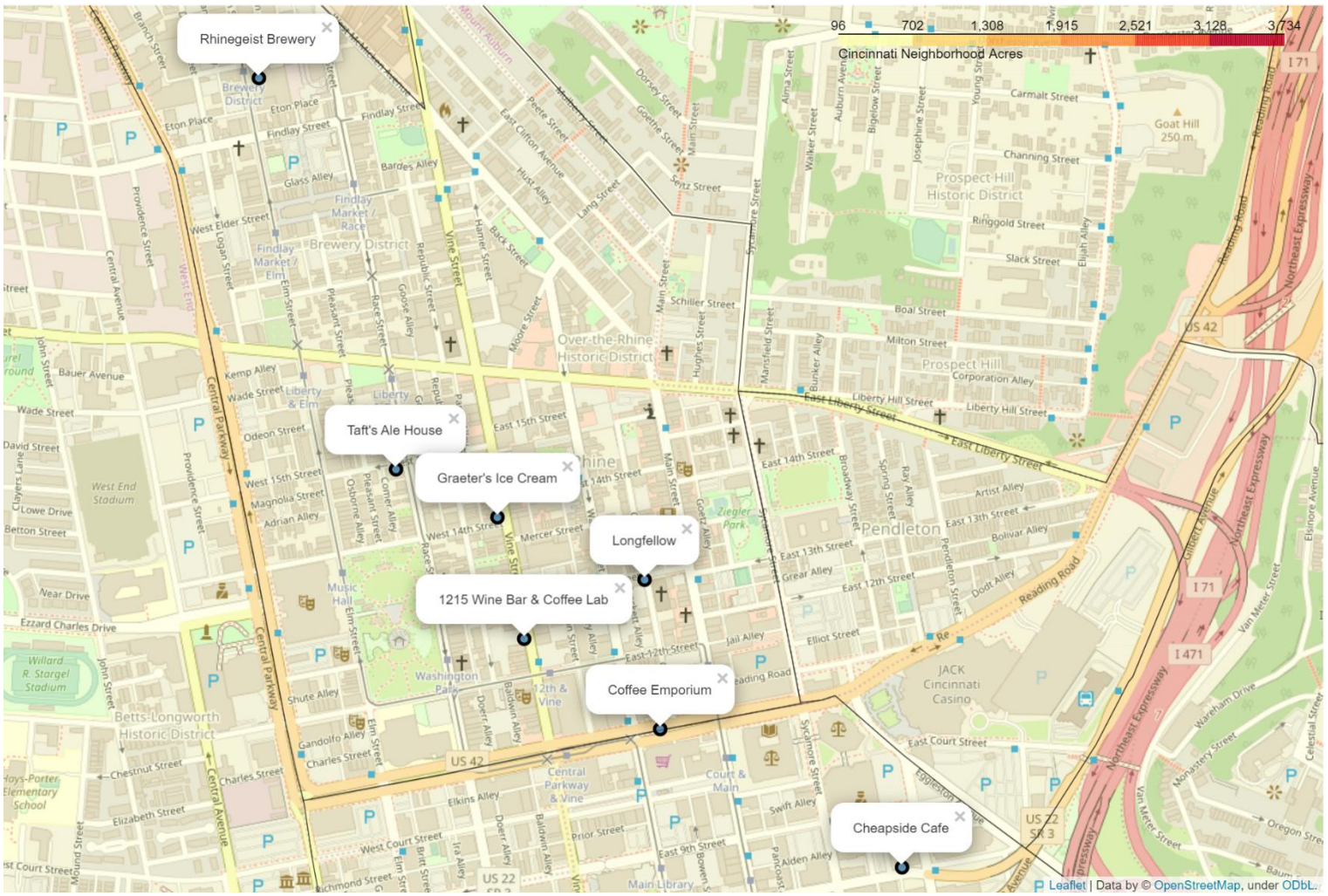
This map tells us that we are safe to pick any area in the downtown to hold our event. But what is the best group of venues based on ratings and much less likes.

To find the best group of venues we sort the venues by ratings and likes and then we iterate over the list of venues until we have a cluster that is within walking distance. I employed DBSCAN once again to verify the venues are all within walking distance.

Here are the top venues for our Night & Day Charity Crawl:

VenueName	VenueLatitude	VenueLongitude	VenueRating	VenueLikes	VenuePrice
Rhinegeist Brewery	39.117221	-84.520129	9.3	460.0	1.0
Taft's Ale House	39.111378	-84.517476	9.3	304.0	2.0
Coffee Emporium	39.107498	-84.512390	9.1	279.0	1.0
Graeter's Ice Cream	39.110662	-84.515525	9.0	51.0	2.0
Cheapside Cafe	39.105442	-84.507739	8.9	91.0	1.0
Longfellow	39.109734	-84.512704	8.9	25.0	2.0
1215 Wine Bar & Coffee Lab	39.108851	-84.515014	8.8	101.0	2.0

Top Venues for Event Crawl



5 Conclusions

This study analyzed the distribution of venues within the city limits of Cincinnati, OH, in order to produce recommended venues for a series of nonprofit outreach events, “night & day crawls”. I feel like the results were fairly successful, and the methodology could be applied to many types of “crawl” events. The parameters for the algorithm are simple enough to change to fit a different clients’ criteria.

The downtown neighborhoods won the top recommendation which was not that big of a surprise. And I was happy to see that most of the top venues are also some of my favorite places to visit in the area.

The biggest complication during the study was collecting and cleaning the data. The Foursquare API had to be called over various days due to the limit on the premium requests “venue-hours” and “venue-details”. The data was stored in CSV files to allow for multi-day runs.

6 Future Considerations

I would like to bring in other location APIs to compare the results. I think that some of the criteria could be tweaked and added to, e.g. an option to ignore certain venues would be helpful. Hours of operation should be collected in more detail, e.g. many of the cafes would not be open the entire event time. Although originally planned for I would like to take the time to create an optimal path through the recommended venues.

Resources

https://nbviewer.jupyter.org/github/ajcanterbury/Coursera_Capstone/blob/master/Data%20Science%20Capstone%20Notebook.ipynb

Disclaimer: This study is for entertainment only. Don't be a drunken fool!