Data science capstone project

The Final Battle of Neighbourhoods

The NGO path of events on Social Life Awareness in Cincinnati, Ohio

By Putrawan

Table of content

1. Introduction

A Non-Government Organisation (NGO) plans to create a new path of events to grab donors to increase awareness of social life. The NGO that stands in the city of Cincinnati, Ohio will partnering with venues where people are present to have social life such as bar and café. The goal is not only for charity but also to lift visibility and increase sponsorship interest. To make this path of events run succes, the NGO set criterias as below:

1. The selected venues proximity must consist at least 2 bar and 2 cafés. The cluster of venues as availabe planned budget and location distance is max 7 venues.
2. All the venues candidate must have open hours around 3pm –10pm on every saturdays as the day is considered the right time for having optimum participants.
3. The venues must also have a well-known reputation
4. The candidates for venues must not classified expensive.

A data scientist is needed to answers the requirement as part of the responsibility for the sponsors and donors involved. The methodology of analysis can be generalized to define recommendations for further NGO plan on the path of events in other cities.

2. Data Usage Description

A data scientist then digs credible data sources that will be used to meet the need by using the following data source:

a. List of Neighbourhoods in Cincinnati, Ohio. There are 50 neighbourhoods that source comes from the “Cincinnati Area Geogrpahic Information System” https://data-cagisportal.opendata.arcgis.com/datasets/cincinnati-sna-boundary This dataset provides land area in acres and boundary coordinates which we need to parse for the center point of the business district.

b. Use the Foursquare API to get list of neighborhood venues with hours, reviews and approximate prices. Then, set the “venues-explore” endpoint with the parameters: latitude, longitude, radius = 1000 meters, limit = 100, section = drinks and coffee. We will explore the geo-location, name, and category.

The data parameters set as below:

- **VENUE_PRIME = ['bar', 'pub', 'brewery', 'lounge']**, bar patterns categories.
  VENUE_SECONDARY = ['caf', 'coffee', 'tea', 'desert', 'ice cream', 'donut'], cafe patterns categories.
- **MAX_VENUES = 7**, maximum of venues per event.

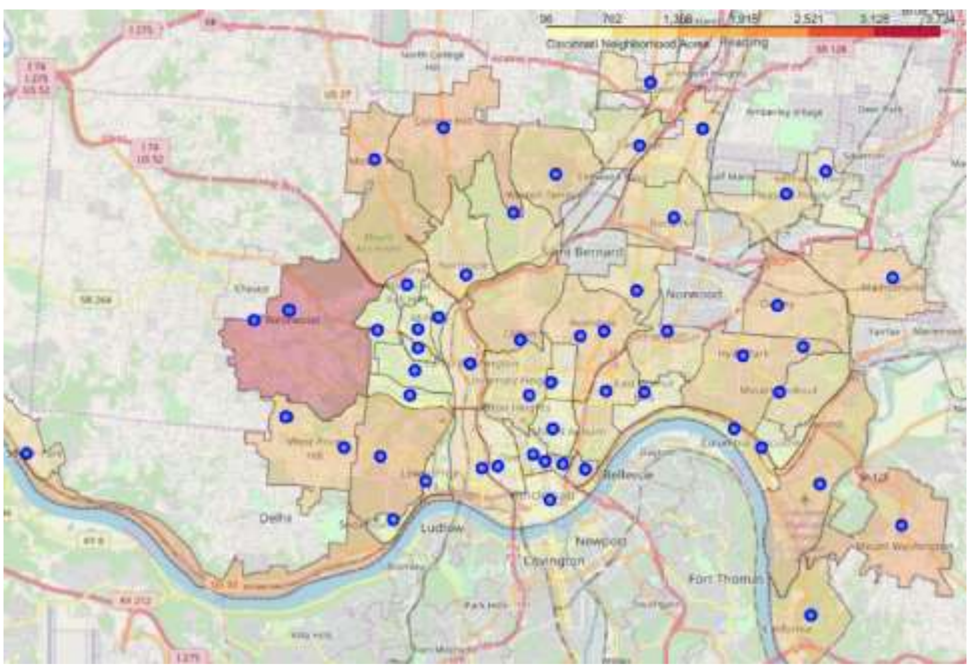
- **MAX_WALK = 0.8**, around 0.5 miles
- **MAX_PRICE = 3**, the Foursquare ranks for prices range 1 to 4.
- **MIN_PRIME = 2**, minimum number of venues that match bars. **MIN_SECONDARY = 2**, minimum number of cafes.
- **WEEK_DAY = 6**, Saturday.
- **START_TIME = 1500**, 3 PM. **END_TIME = 2200**, 10 PM.
- **PRIORITY_ORDER = {'Rating': 4, 'Count': 2, 'Likes': 1}**, provides weighting scale.

To some extend of limitation, anything labeled as coffee shop, teahouse, pastry shop and similar ones will be defined as café. As the data also comes from the Foursquare API, beyond cleaning and formatting datasets, prediction is also needed due to missing data on prices and ratings review in some of the neighbourhoods. That condition can also be found, as new venues appear on the avialable data.

3. Methodology

3.1 Exploring geodata from city APIs and display **Choropleth map**

First, exploring are two geographic API datasets that are important. The SNA Boundary data for the city neighborhoods, and the Business Districts. It provides us with coordinates which we need to parse for the center point of the business district.

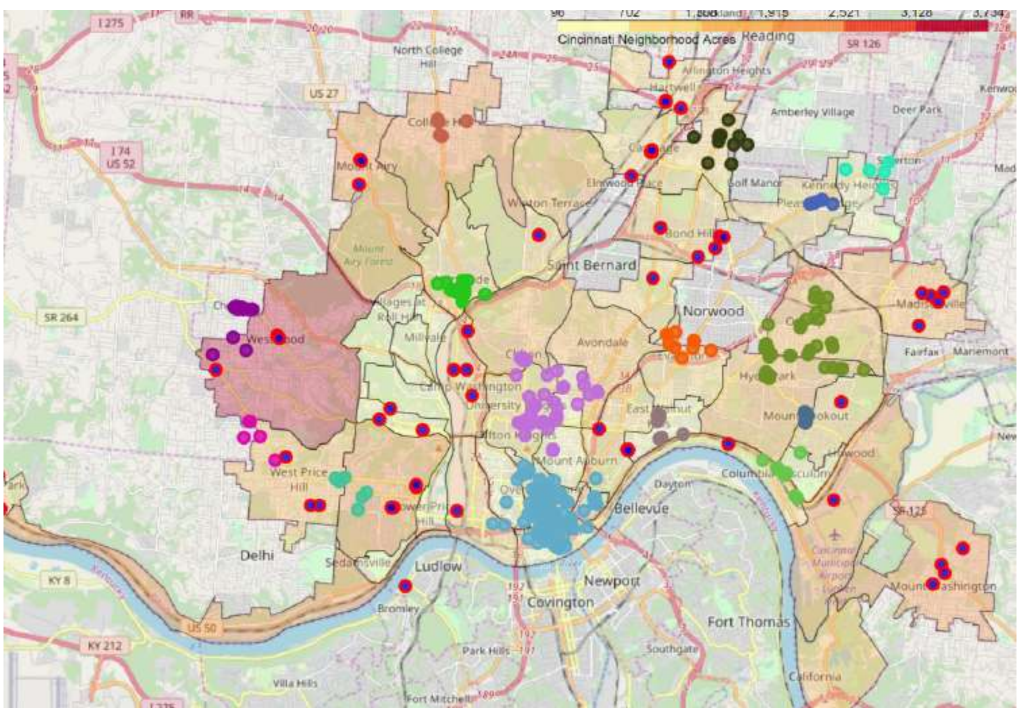


For multiple districts for a neighborhood we collect each one in a different district. If no district was found, focus on center point of the neighborhood. Then consider to remove the neighborhood of Riverside due to a lack of businesses.

3.2 <u>Define venues that match criteria and clean outliers from **DBSCAN**</u>

Next, define venues that match criteria by pulling data from Foursquare API and set center points. The ‘venue-explore’ endpoints is used combine with latitude, longitude, radius = 1000 meters, limit = 100, section = drinks and coffee as parameters. Coffee and Drink match the venue criteria. Focus on the geo-location, name, and category. Cleaning the data also examined, there are 381 venues that not match criteria. The clean data stored in .csv files. Outliers also removed from DBSCAN.

	Neighborhood	BusinessDistrict	NeighborhoodLatitude	NeighborhoodLongitude	VenueName	VenueId	VenueLatitude	VenueLongitude	W
0	Linwood	0	39.104213	-84.415924	Dennert H Distribtg	4f32494419836c91c7c8b7f7	39.108777	-84.421232	
1	East Walnut Hills	1	39.128889	-84.476823	The Woodburn Brewery & Taproom	55461bf6498eac118325a62e	39.129030	-84.476892	
2	East Walnut Hills	1	39.128889	-84.476823	Myrtle's Punch House	5473d783498ec0bbca9021d6	39.124276	-84.476130	
3	East Walnut Hills	1	39.128889	-84.476823	The Growler House	545d54ab498ea427d9af9d2d	39.129763	-84.477776	
4	East Walnut Hills	1	39.128889	-84.476823	BrewRiver Gastropub	4fea02ede5e8dfeeb65b5000	39.121758	-84.475027	
5	East Walnut Hills	1	39.128889	-84.476823	The Skunk Lounge	5182cdbd498e1c1b38b47f1c	39.124213	-84.476246	
6	East Walnut Hills	1	39.128889	-84.476823	Cliche	5c6459abca17830006abf539	39.123820	-84.477040	
7	Queensgate	0	39.106472	-84.533758	City West Brewing Company	580ceb4a38faa26bf32db135	39.108208	-84.525736	
8	Queensgate	0	39.106472	-84.533758	The Playhouse	4e8a404977c80797 4bd89725	39.106017	-84.541503	
9	Queensgate	0	39.106472	-84.533758	Royal Imports	4f3246f419836c91c7c7cd1e	39.102755	-84.526396	



The resut revealed that 325 venues within 15 clusters match criteria.

3.3 <u>Define venues hours and use **K-Nearest Neighbours machine learning**</u>

Define the venues that their hours of operations fit with the Event times: Saturday 3pm – 10 pm. Therefore “venue-hours” endpoint is exercised from Foursquare API. Unsupervised machine learning is also being used to patch the gap. The KNN (K-Nearest Neighbors) is used to determine weather the venues are open or closed during particular time period. The venues also have similar assumption that they have operation hours similar with others.
Jaccard similarity coefficient and F1 score is used to define the accuracy of the model. Rated from 0 (as worst) and 1 (as perfect) score for similiarity to true values and precision and respectively recall to true value. The Jaccard score ~ 0.915 and F1 score ~ 0.875, are the basis to determine the nine of venues that are most likely closed on the event time. The rest 316 venues with Re-clustering is found not affected the clusters position.

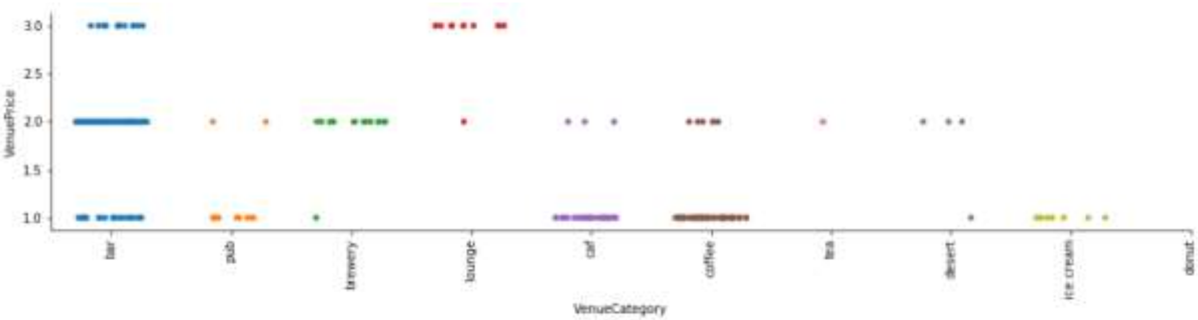
3.4 <u>Define venue features correlation, categorical relations with **Scatterplot**</u>

By using the Foursquare API, get into venue details endpoint. This 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 stand for cheap, moderate, expensive, and very expensive. However, this data contains missing values to be handled.

Venue Feature Correlations

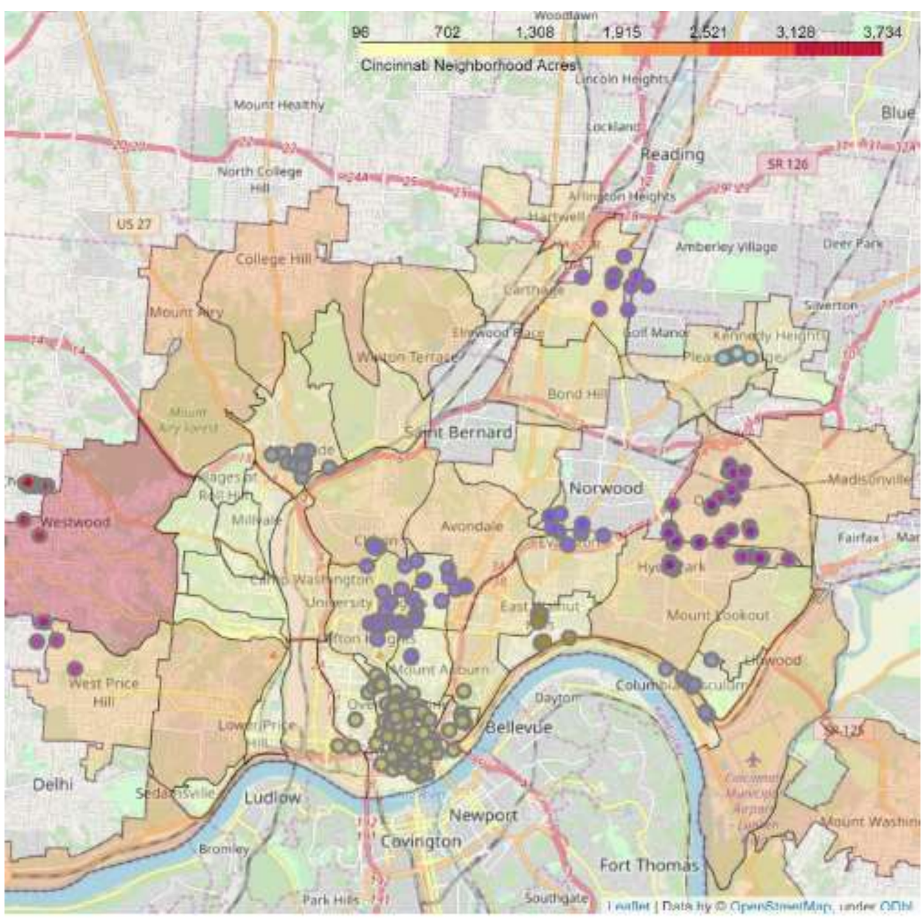
	DbCluster	VenueCategory	VenueRating	VenueLikes	VenuePrice
DbCluster	1.000000	-0.006894	-0.174062	-0.133704	-0.068473
VenueCategory	-0.006894	1.000000	0.050535	-0.010160	-0.532074
VenueRating	-0.174062	0.050535	1.000000	0.498003	0.063367
VenueLikes	-0.133704	-0.010160	0.498003	1.000000	0.007484
VenuePrice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

Fill the missing data with the median of each cluster to use the ratings for recommendations. The price establishments are very important and there are only 1 missing venue’s prices from the max seven venues. Event Max Price: 3 or expensive. From a correlation matrix of the pertinent venue features, and the marginal effect on prices is the category.



Categorical scatterplot is used to look detail of relationship between categories and price. There is a strong connection between many categories and price than infered from the plot, except the bar mixed case. The KNN model is used to define achievement of these accuracies. Jaccard score ~ 0.844, F1 Score ~ 0.812 make it the event venues left with 15 cluster and 316 venues.

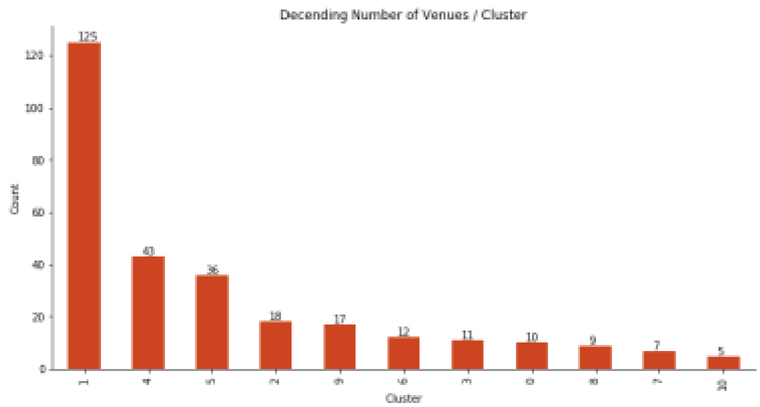
Based on parameter that being address, it arrived on final criteria. Minimum and secondary categories exercised that it came with no minimum categories. DBSCAN showed that it can used the11 remaining clusters.



4. Result analysis

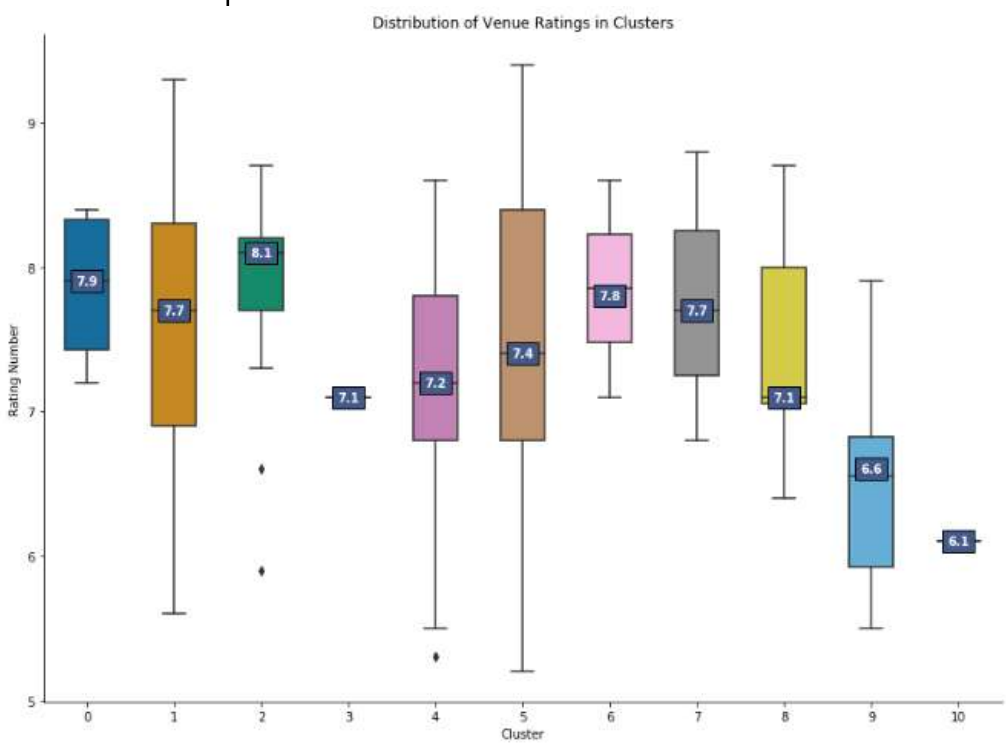
The most qualified venues for the events are the downtown and its adjacents. The only overlap of neighborhoods is found between cluster 1 and 4, with Mt. Auburn. Several clusters bleed into outside neighborhoods that are not incorporated in the city of Cincinnati.

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



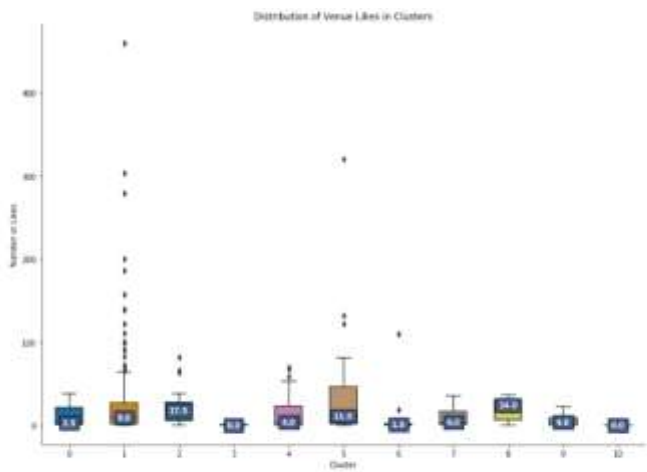
The venues in the downtown cluster is close to more than other combined clusters. Cluster 4 (clifton), and cluster 5 (Hyde Park) are significant number of venues.

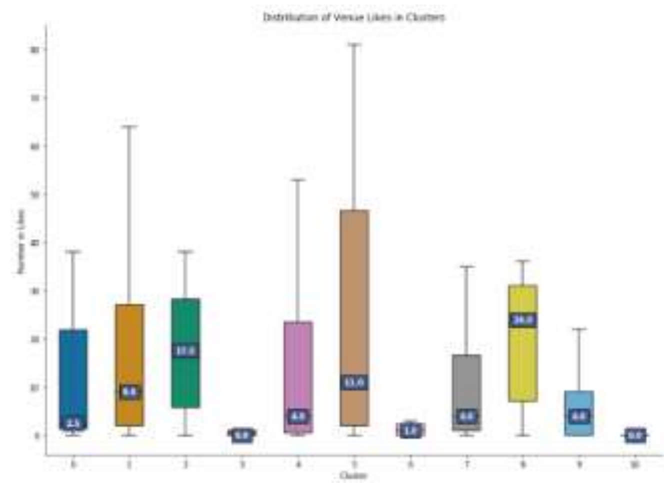
Boxplot revealed that distribution of ratings in the clusters since quartiles, and the median are the most important values.



The rating accuracy is good. The median ratings between 6 and 8 as expected. There are only a couple of outliers, and when there are significant numbers of venues, the highs and lows are far apart.

The gimmic plan in the business-like discount package and other marketing tools influence the venue counts of like. Two boxplots below show outliers and not outliers.



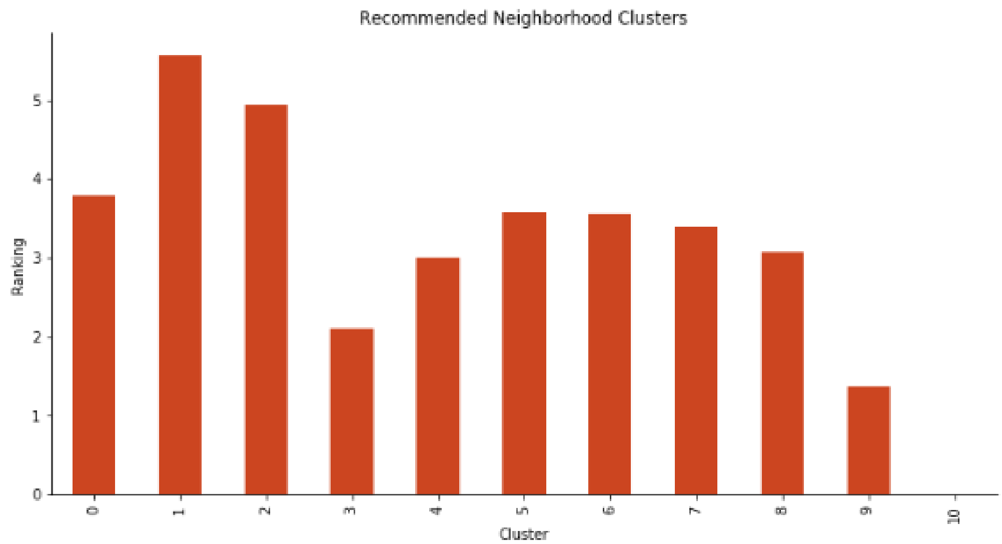


To summarize the result of analysis

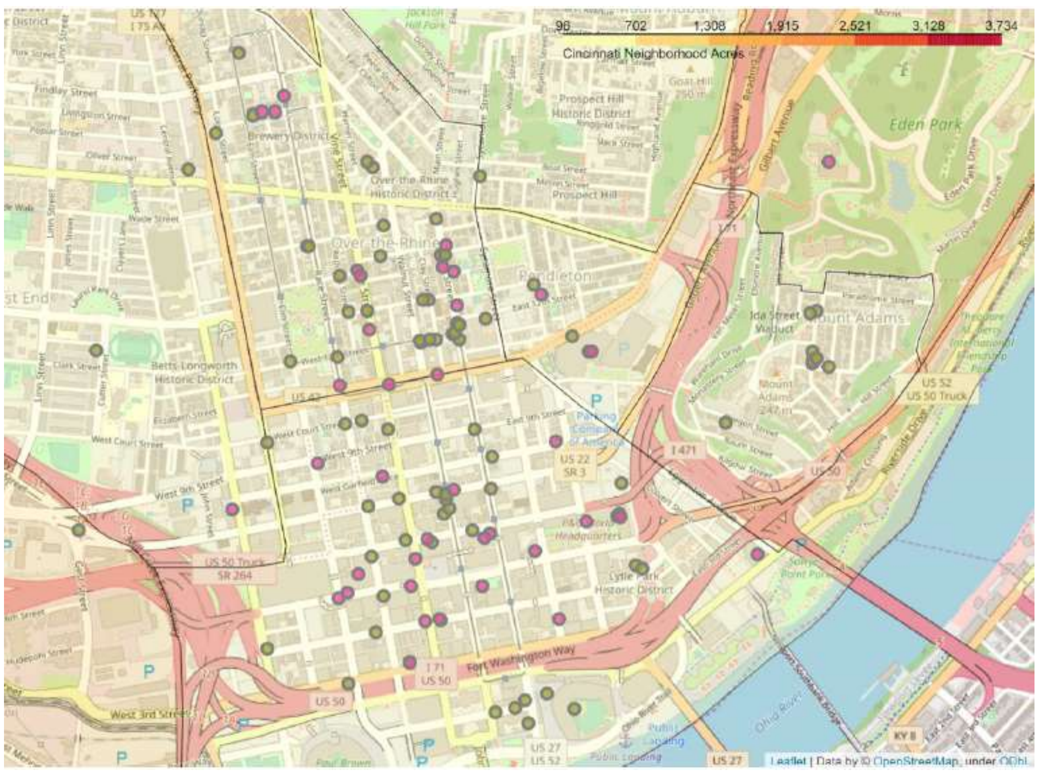
The Downtown neighbourhoods including Queensgate, West End, Mt. Auburn, Downtown, Over-the-Rhine, and Pendleton are the largest cluster of venues. The ratings medians are similar across all cluster and the overall distribution of ratings are varied. The venue likes cannot be used to reflect real recommendations.

5. Discussion and recommendations

The cluster of venues show that the higher the ranking the better the venues. Downtown wins primarily based on its number of venues, with Northside is the top-rated venues



The Downtown group is fairly evenly distributed. This map show that the best group of venues based on ratings and much less likes and it is safe to pick any area in the downtown to hold our event.



DBSCAN once again verifies the venues are all within walking distance, and the top venues is listed below:

	VenueName	Neighborhood	VenueLatitude	VenueLongitude	VenueRating	VenueLikes	VenuePrice	PrimaryCategory	DbCluster
0	Rhinegeist Brewery	Over-the-Rhine	39.117221	-84.520129	9.3	460.0	1.0	1	0
1	Taft's Ale House	West End	39.111378	-84.517476	9.3	304.0	2.0	1	0
2	Coffee Emporium	Downtown	39.107498	-84.512390	9.1	279.0	1.0	0	0
3	Graeter's Ice Cream	Over-the-Rhine	39.110662	-84.515525	9.0	51.0	2.0	0	0
4	Cheapside Cafe	Downtown	39.105442	-84.507739	8.9	91.0	1.0	0	0
5	Longfellow	Over-the-Rhine	39.109734	-84.512704	8.9	25.0	2.0	1	0
6	1215 Wine Bar & Coffee Lab	Over-the-Rhine	39.108851	-84.515014	8.8	101.0	2.0	0	0

Top Venues for Event Crawl

6. Conclusion

The downtown neighborhoods won the top recommendations as this study analyzed the distribution of venues. The challenge found when collecting and cleaning data. External sources such as Foursquare API has limitation and must be combined with .csv data file stored for days run.

This work can be optimized by other APIs location as comparation, and enrich the criteria such as hours of venue operational.