Cleveland Business Expansion

Introduction

As businesses expand, key strategic decisions need to be made about where new locations are opened. This project explores a hypothetical case of a coffee shop owner, based in Independence, Ohio looking to expand into the rest of Greater Cleveland area. There are a host of variables to consider in this case, including the current business climate and the local economic health [1]. However, another consideration could be to search for communities that are similar to where you have already found success, but do not have much competition in terms of other coffee businesses already opened. If you are not familiar with such a large area, how can you determine which communities may be similar to your current business location?

Data Sources

In order to solve this problem, open and publically available data related to the Cleveland area was searched. The zipcodes for various cities within Greater Cleveland were found on Trip Savvy's website [2], and then the latitude and longitude data for each one of these zipcodes was retrieved through Eric Hurst's GitHub account [3]. This was a convenient source, as he already had the data already nicely formatted from the US Census website [4]. Next, the data sources were joined using a vlookup in Excel. For some further refinement, python was used to group by zip code and combine cities that fell into a particular zip code. A subset of the resulting data frame is shown below:

	Zipcode	Latitude	Longitude	City
0	44004	41.855940	-80.791866	Ashtabula
1	44011	41.445100	-82.005109	Avon Lake
2	44012	41.498342	-82.017368	Avon
3	44017	41.370548	-81.861757	Berea
4	44021	41.443260	-81.144465	Burton
5	44022	41.446269	-81.402969	Chagrin Falls , Moreland Hills
6	44023	41.384751	-81.285759	Bainbridge
7	44024	41.577765	-81.192433	Chardon
8	44026	41.528147	-81.324706	Chesterland
9	44039	41.386000	-82.024711	North Ridgeville

Figure 1 - Subset of Main Data Frame

One way cities can be compared is with their already existing venues. What categories of venues are frequented most often? Data about nearby venues within each of these cities was obtained from Foursquare's API. Foursquare is considered the most " trusted, independent location data platform for understanding how people move through the real world" [5].

Methodology

The folium library was used to plot the cities in the Greater Cleveland area based on the main data frame created.

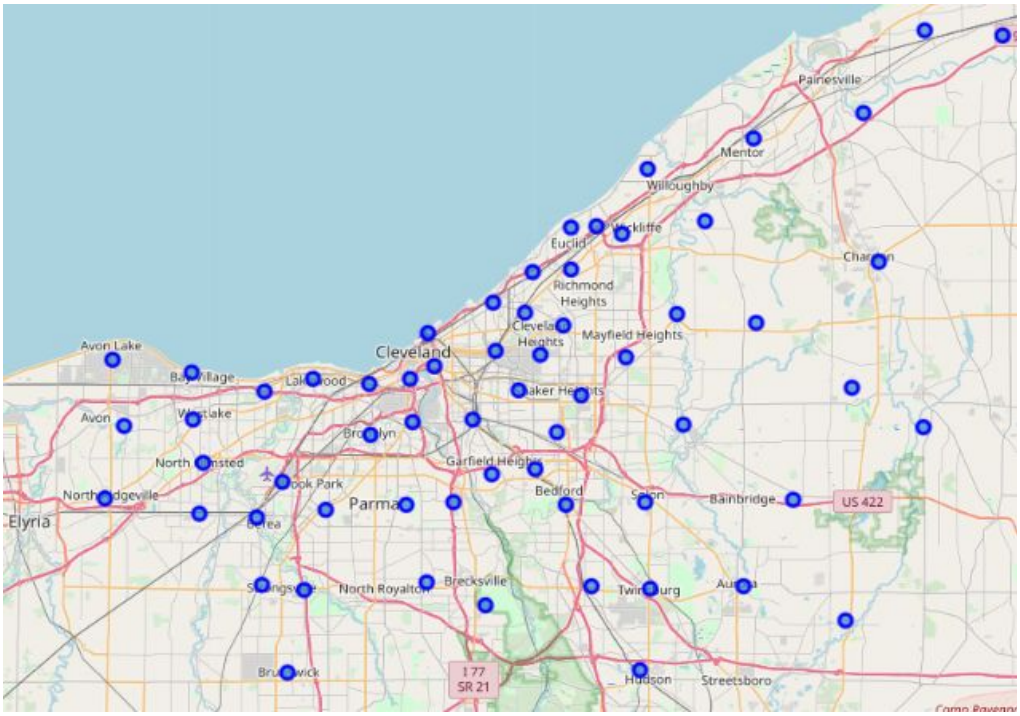

Figure 2 - Communities of Greater Cleveland

Looping through each city, a list of venues and their corresponding categories were obtained. See the below figure for a subset of what this data looked like.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ashtabula	41.855940	-80.791866	Domino's Pizza	41.854736	-80.786767	Pizza Place
1	Ashtabula	41.855940	-80.791866	Circle K	41.854803	-80.786620	Convenience Store
2	Ashtabula	41.855940	-80.791866	Fay's Coffee Kitchen	41.855420	-80.788670	Coffee Shop
3	Avon	41.498342	-82.017368	Las Cazuelas	41.494242	-82.017597	Mexican Restaurant
4	Avon	41.498342	-82.017368	The Landing	41.495411	-82.017943	Plaza
5	Avon	41.498342	-82.017368	Romeo's Pizza	41.495279	-82.021231	Pizza Place
6	Avon	41.498342	-82.017368	Neon beach	41.495279	-82.021231	Spa
7	Avon	41.498342	-82.017368	Erie Burger Company	41.494867	-82.020644	Burger Joint
8	Avon	41.498342	-82.017368	The Landings	41.493943	-82.017721	Shopping Mall

Figure 3 - Venue Data Subset

Upon further data exploration, it was found that there were 166 unique venue categories, ranging from “Fabric Shop” to “Organic Grocery”. One-hot encoding was then applied to the data, for ease of further analysis. Each venue category type became a column header with values either set to 0 or 1 depending on if that row contained that venue category or not. Please note that all of the values shown in the below example show 0 as there are hidden columns not displayed.

	Neighborhood	Accessories Store	American Restaurant	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	Auto Dealership	Auto Garage	Automotive Shop	...	Toy / Game Store	Trail	Video Game Store	Video Store
0	Ashtabula	0	0	0	0	0	0	0	0	0	...	0	0	0	0
1	Ashtabula	0	0	0	0	0	0	0	0	0	...	0	0	0	0
2	Ashtabula	0	0	0	0	0	0	0	0	0	...	0	0	0	0
3	Avon	0	0	0	0	0	0	0	0	0	...	0	0	0	0
4	Avon	0	0	0	0	0	0	0	0	0	...	0	0	0	0

Figure 4 - Venue Data Subset after One-Hot Encoding

Next, the data was grouped by the neighborhood, with the values from the one-hot encoding being averaged. This resulted in only one row for each community.

	Neighborhood	Accessories Store	American Restaurant	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	Auto Dealership	Auto Garage	Automotive Shop	...	Toy / Game Store	Trail	Video Game Store
0	Ashtabula	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000	0.000000	...	0.000000	0.0	0.000000
1	Aurora	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000	0.000000	...	0.000000	0.0	0.000000
2	Avon	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000	0.000000	...	0.000000	0.0	0.000000

Figure 5 - Venue Data Subset after Grouping by Neighborhood

A frequency chart was created that outputted venue category by neighborhood. For the Independence and Seven Hills neighborhood, coffee shops, baseball fields, and parks were recorded as equal frequency. Next, a function was written to display the top 5 venues for each neighborhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Ashtabula	Convenience Store	Coffee Shop	Pizza Place	Yoga Studio	Dessert Shop
1	Aurora	Gym / Fitness Center	Golf Course	BBQ Joint	New American Restaurant	Yoga Studio
2	Avon	Plaza	Pizza Place	Mexican Restaurant	Burger Joint	Shopping Mall
3	Bay Village	Park	Theater	Dessert Shop	Beach	Soccer Field
4	Beachwood , Shaker Heights	Department Store	Yoga Studio	Diner	Fabric Shop	Entertainment Service

Figure 6 - Top Venues by Neighborhood

Based on this data, an unsupervised k means algorithm was implemented to cluster neighborhoods in the Greater Cleveland area based upon the top five venues in each community. The results, displayed below, were mapped using the folium library. The red markers represent the cluster that contains Independence.

Figure 7 - Clustering of Neighborhoods in Greater Cleveland

Results

Upon clustering, the machine learning algorithm grouped Madison, Perry, Wickliffe/Willowick, Cleveland Heights/South Euclid, Parma, and Bay Village with the Independence/Seven Hills neighborhood. See the below figure for details regarding the most common venue types for each neighborhood.

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
14	Madison	Construction & Landscaping	Farmers Market	Fabric Shop	Entertainment Service	Electronics Store
19	Perry	Athletics & Sports	Business Service	Yoga Studio	Diner	Fabric Shop
21	Wickliffe , Willowick	Golf Course	Park	Yoga Studio	Dessert Shop	Entertainment Service
39	Cleveland Heights , South Euclid	Park	Rental Car Location	Yoga Studio	Dessert Shop	Entertainment Service
46	Independence , Seven Hills	Coffee Shop	Park	Baseball Field	Yoga Studio	Diner
48	Parma	Campground	Park	Business Service	Construction & Landscaping	Diner
53	Bay Village	Park	Theater	Dessert Shop	Beach	Soccer Field

Figure 8 - Inspection of Cluster Containing the Neighborhood of Independence/Seven Hills

Discussion

Upon inspection of this cluster, it appears that they all share similar venue categories with Independence, having ‘park’, ‘yoga studio’, and/or ‘diner’ in their top five list, with the exception of Madison. It seems Madison was added to this cluster based on its commonality of ‘entertainment service’ to some of the other neighborhoods in the cluster. Interestingly, although the algorithm does not know the relation between baseball field and soccer field, those also happened to relate Bay Village to Independence.

The Bay Village, Wickliffe/Willowick and Cleveland Heights/South Euclid neighborhoods could be potential candidates to explore as places to open a new coffee shop based on their similarity to Independence, and also because ‘dessert shop’ is listed in their top 5 venue categories types.

Conclusion

Although this clustering is useful for exploration, a business decision as important as deciding upon where to add a new location should not be based upon this analysis alone. Based upon this analysis, I would suggest Bay Village, Wickliffe/Willowick and Cleveland Heights/South Euclid as places to explore opening another coffee shop location. Further research should be done to determine if there are already coffee shops located in each of these communities and if there are, why they did not make the top 5 list for those communities.

References

[1] https://www.nerdwallet.com/blog/small-business/best-places-start-business-ohio/

[2] https://www.tripsavvy.com/cleveland-area-zip-codes-753059

[3] https://gist.github.com/erichurst/7882666

[4] https://www.census.gov/programs-surveys/geography.html

[5] https://foursquare.com/