Capstone Project: The Battle of Neighborhoods
Finding a location for a new bakery in Cincinnati

Maria Fox

September 4, 2020

Contents

1 Introduction/Business Problem

A new bakery is hoping to open in the Greater Cincinnati, Ohio, area, but the owners aren't sure where to open it. While Cincinnati is a fairly small city, there are distinct neighborhoods with different personalities and characteristics that might influence the success of a small new bakery in town. Additionally, the owners would be competing with several well-known local bakeries that have locations across town, such as Busken Bakery and Servatii.

The owners are looking for recommendations for where to locate their bakery. They would like to know **where other successful bakeries are located and the businesses around these successful bakeries** to determine if certain areas or other venues might influence the bakery's success. Based on those characteristics, I will suggest neighborhoods that would be good candidates for opening the new bakery.

2 Data

I will use data from the following sources to determine between one and three options for the new bakery's location:

- Cincinnati neighborhoods on Wikipedia to web scrape neighborhood names
- Google maps to also pull neighborhoods/city names from Northern Kentucky
- Foursquare API to extract venues in each neighborhood

- Foursquare API to search for all bakeries
- Foursquare API to get the number of likes for each bakery

3 Methodology

With Wikipedia's page of Cincinnati neighborhoods, I web scraped the content using the `BeautifulSoup` python package to extract the list of 46 neighborhoods (one neighborhood, The Villages of Roll Hill, was removed from the list because its geographic coordinates could not be located). The geographic location for each neighborhood was extracted using `geopy`. Additionally, 14 neighborhoods in Northern Kentucky were included to make up the Greater Cincinnati area. Northern Kentucky locations were chosen by inspecting Google Maps within a radius similar to Cincinnati's. Northern Kentucky neighborhood locations were then extracted using `geopy`, and the Cincinnati and Northern Kentucky data were concatenated in Python. Neighborhood locations were visualized in Python using 'folium' to denote neighborhood centroids.

I used the Foursquare API to "explore" popular venues in each neighborhood. Rather than defining a set radius for each neighborhood to explore, I allowed the Foursquare API to calculate a suggested radius based on the density of venues in each neighborhood. I extracted venue names, venue IDs, venue locations, distance of each venue to the neighborhood center, calculated neighborhood radius, and venue category.

Because many neighborhoods were relatively close together and suggested radii overlapped, some venues were duplicated in the resulting list (as noted by the venue ID). Duplicate venues were removed by inspecting the distances to the neighborhood center and keeping whichever was closest to a neighborhood center.

Since the "explore" feature might not find all of the bakeries in a given neighborhood, I searched for bakeries separately using the Foursquare API "search" function with the same suggested radius for each neighborhood as was used for exploring. Duplicate bakeries were removed in the same manner as other venues described above. Since some bakeries were not categorized as Bakery under the venue category, I created an additional column that denoted all venues in the bakery list as bakeries. Before merging the neighborhood venues and bakeries lists, I removed all venues containing the word "bakery" in the first (explore) list and denoted these venues as *not* bakeries.

I determined which bakeries in the Greater Cincinnati were most popular using the number of "likes" recorded in Foursquare (NOTE: this popularity is only reflective of Foursquare users' opinions and therefore may be biased and/or inaccurate). I used the Foursquare API "likes" endpoint to extract the number of likes for each bakery. I then extracted unique neighborhoods that contained bakeries with 10 or more "likes" to determine which neighborhoods contained the most popular bakeries in town.

I converted the original list of "explore" venue categories into dummy variables and calculated mean frequency of each venue category across neighborhoods. Venue category frequencies were visualized for the neighborhoods containing the most popular bakeries using `seaborn` heatmaps. These heatmaps visually emphasized venue categories that were more prominent in the neighborhoods with popular bakeries.

Finally, dummy-coded data frames for all venues and bakeries were concatenated, and venue categories converted to mean frequencies.

I then used k-Means clustering from `scikit-learn` to cluster neighborhoods based on similar venue category frequencies. The optimal number of clusters was determined by calculating the minimum within sum-of-squares (WSS) across different candidate numbers of clusters (2-15). I then determined which cluster(s) the popular bakery neighborhoods belonged to, and using category

frequencies and heatmaps, determined which shared features best described the cluster(s). Members of the same cluster(s) would indicate good candidate neighborhoods for the client's new bakery location.

Based on the mean frequency of venue categories in the neighborhoods with the most popular bakeries, I then calculated the Euclidean distances between those frequencies and the frequencies from each candidate neighborhood in the relevant cluster(s). I sorted the Euclidean distances, and retained the five neighborhoods with the smallest Euclidean distances as those that were most similar to the neighborhoods with popular bakeries. Those five neighborhoods then represent the best candidates for the new bakery.

4 Results

According to Foursquare, seven neighborhoods in the Greater Cincinnati area contained the most popular bakeries: Hyde Park, Kennedy Heights, College Hill, Westwood, Pendleton, Mount Washington, and Northside. The popular bakeries, along with their number of likes, are reported in Table 1. Several of these popular bakeries are actually restaurants that contain bakeries, such as the two Perkins locations.

Neighborhood	Venue	N Likes
Hyde Park	Busken Bakery	67
Kennedy Heights	Ferrari's Little Italy and Bakery	36
College Hill	Perkins Restaurant & Bakery	23
Westwood	Perkins Restaurant & Bakery	17
Pendleton	Brown Bear Bakery	16
College Hill	North College Hill Bakery	12
Mount Washington	Mt. Washington Creamy Whip	12
Northside	Bonomini Bakery	12
Pendleton	Shadeau Breads	10

Table 1: Neighborhoods and the bakeries with 10 or more likes in Foursquare.

These neighborhoods shared certain characteristics, including a relatively high frequency of American restaurants, bars, pizza places, and ice cream shops (Figure 1, Table 2).

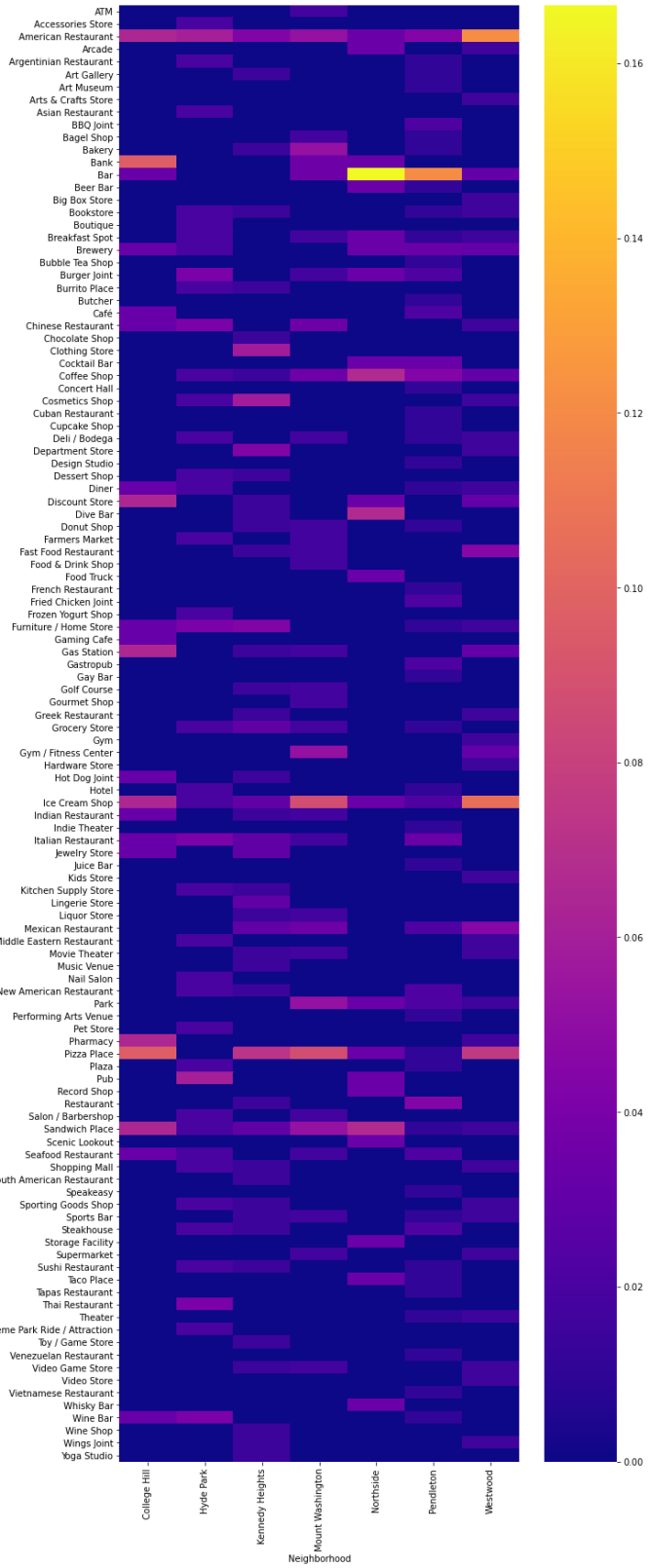

Figure 1: Venue category frequency in neighborhoods with popular bakeries

Venue Category	Mean Frequency
American restaurant	0.060
Bar	0.055
Pizza Place	0.054
Ice Cream Shop	0.052
Sandwich Place	0.037
Coffee Shop	0.030

Table 2: Venue categories with frequencies > 0.025 in the popular bakery neighborhoods

A k-Means cluster analysis defined eight distinct clusters of neighborhoods based on the venue category frequencies. Clusters 0 and 1 contained all seven of the neighborhoods with popular bakeries, so those two clusters were chosen as the candidate clusters for the new bakery.

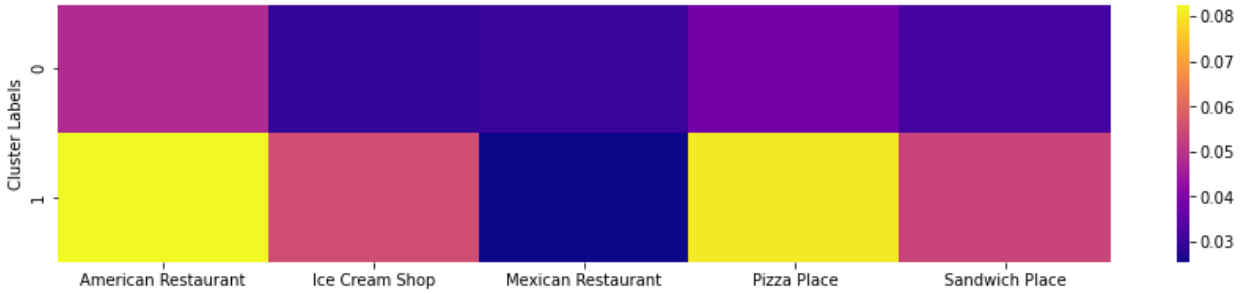

Figure 2: Venue category frequency in neighborhoods with popular bakeries

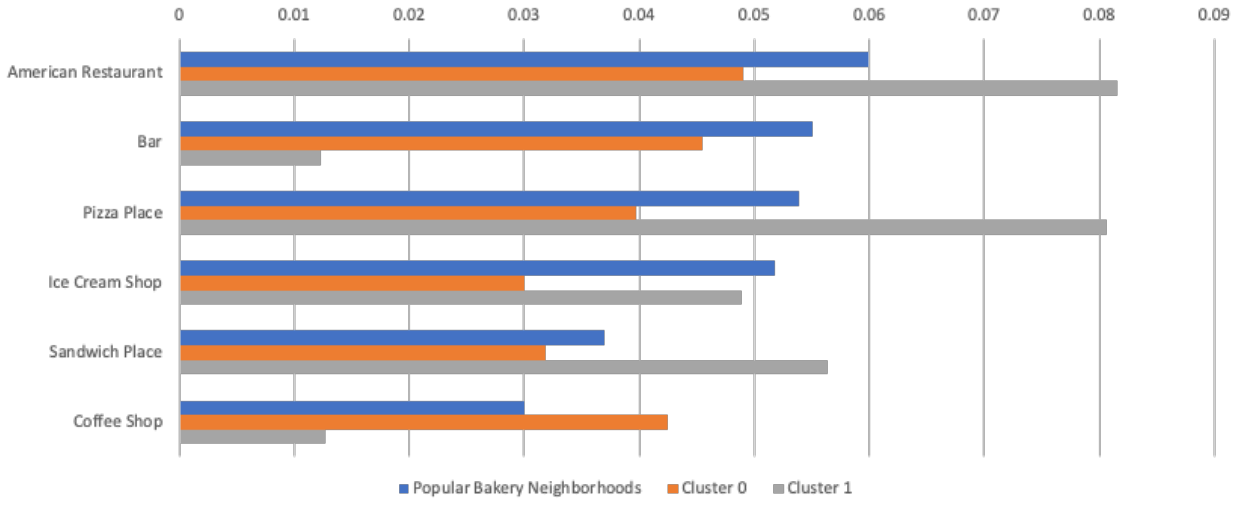

Figure 3: Venue category frequency in neighborhoods with popular bakeries

The vast majority of neighborhoods were included in clusters 0 and 1 (34 and 17, respectively; only 9 neighborhoods were not included in those first two clusters), and the feature most characteristic of the two clusters was American restaurants (Figure 2, Figure 3).

Euclidean distances between the mean frequency of the most frequent venue categories in the

popular bakery neighborhoods (Table 2) and each candidate neighborhood in the relevant clusters (clusters 0 and 1) revealed that the top five candidates, based on the lowest Euclidean distances, were Winton Hills, California, Southgate, Newport, and Clifton (Figure 4). All five of the candidate neighborhoods belonged to cluster 0.

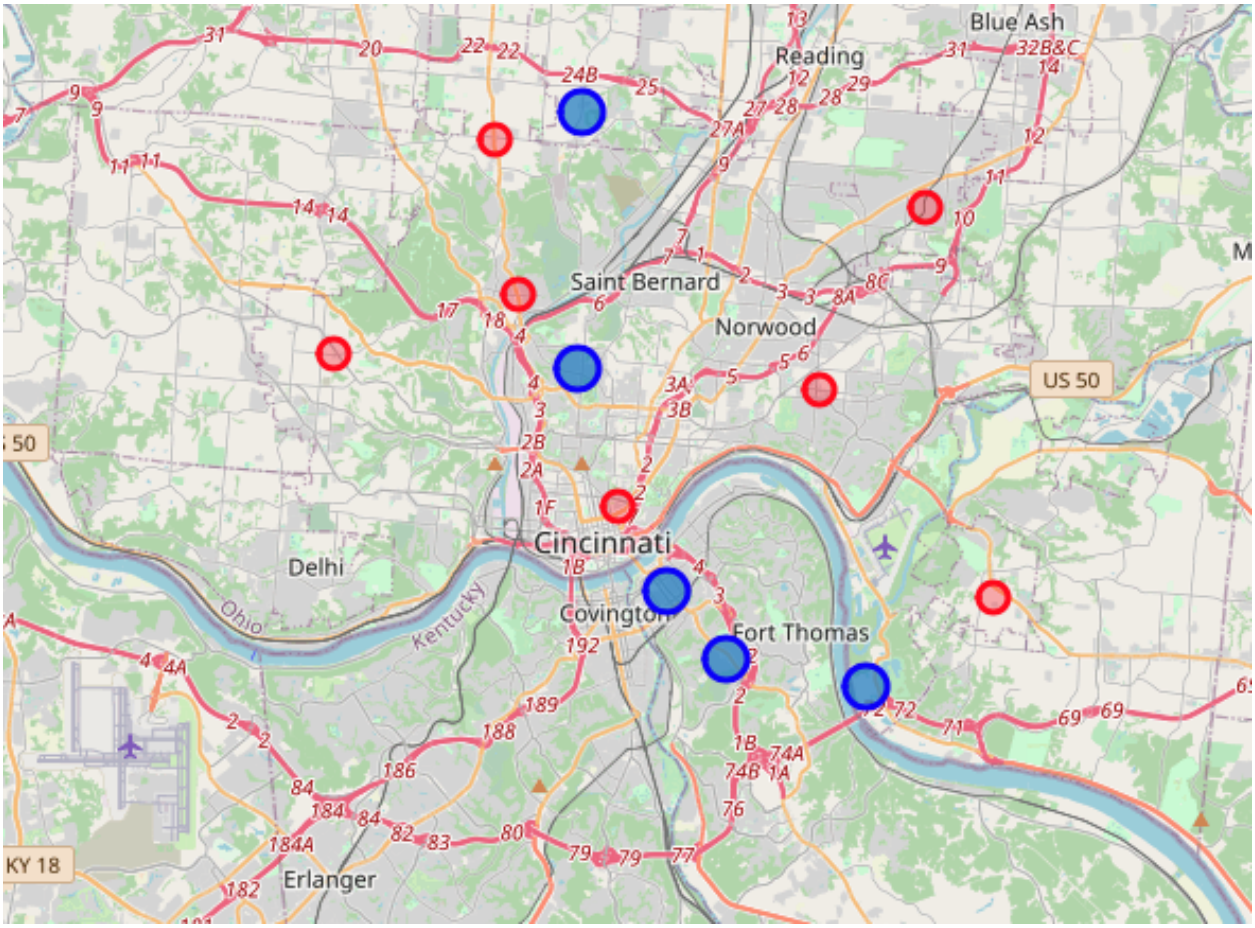

Figure 4: Candidate neighborhoods (blue circles) and neighborhoods with popular bakeries (red circles) plotted on a map of Greater Cincinnati

5 Discussion

In this project, I analyzed the Greater Cincinnati neighborhoods that contained the most popular bakeries as indicated by Foursquare likes. Using k-Means clustering and Euclidean distances, I found candidate neighborhoods that were most similar to the neighborhoods that contained the most popular bakeries. This resulted in a final list of five candidate neighborhoods throughout the Greater Cincinnati area, including Winton Hills, California, Southgate, Newport, and Clifton.

The most popular bakeries in town were located in Hyde Park, Kennedy Heights, College Hill, Westwood, Pendleton, Mount Washington, and Northside. Two of the most popular bakeries were Perkins restaurants, so it is unclear whether these venues were popular because of the restaurant or bakery component.

Out of the candidate neighborhoods that were most similar to the neighborhoods with popular

bakeries, only two of those neighborhoods (Newport and Southgate) already have bakeries. Newport contains two bakeries, Cookie Jar Bakery and Mrs. Fields Bakery Cafe, both of which only had one like on Foursquare. This suggests that if the clients wanted to open a bakery in Newport, there likely wouldn't be much competition from the existing bakeries. Southgate had one bakery, the Hostess Bakery Outlets, with one like on Foursquare. This would also not likely provide a lot of competition, since it is a bakery outlet for a national brand and would be dissimilar from the small local bakery that the clients want to open.

The five candidate neighborhoods are very similar in venue profiles to the popular bakery neighborhoods, and any one of those locations would make a good option. They could choose from locations in Kentucky or Ohio, with options closer or further away from the city center.

6 Conclusion

Using Foursquare API data, I located the most popular bakeries in the Greater Cincinnati area. I used K-means clustering to identify candidate neighborhoods with similar features to the popular bakery neighborhoods, and Euclidean distances to further refine the list of candidate neighborhoods.

Five candidate neighborhoods were identified for the new bakery location, based on similarity to the popular bakery neighborhoods. Three of these candidate neighborhoods (Winton Hills, California, and Clifton) were located in Cincinnati, and two (Southgate and Newport) located in Kentucky.