The Battle of the Neighborhoods

Introduction

Background

This is the final report for the IBM Data Science Professional Certificate course. In this project, I will analyze the neighborhoods in Milwaukee, WI to determine the best location to open a new restaurant and bar. To do this, I will be using python libraries and Foursquare data.

Scenario

A company notices the high growth rates of the number of young people in Milwaukee, WI, and wants to capitalize on it. They want to open a new restaurant/bar in the city that will attract young people by providing a fun place to eat and hang out with their friends.

Problem

The company is unfamiliar with the area and they need to choose the best location for the new location so that it gets a lot of foot traffic, is close to nightlife, and is in a convenient place for customers to get to. They are targeting a younger audience so it is also important to be aware of the locations demographic. They have contracted our firm to find the best place for this new hang out.

Data

Data Sources

For this project, I used:

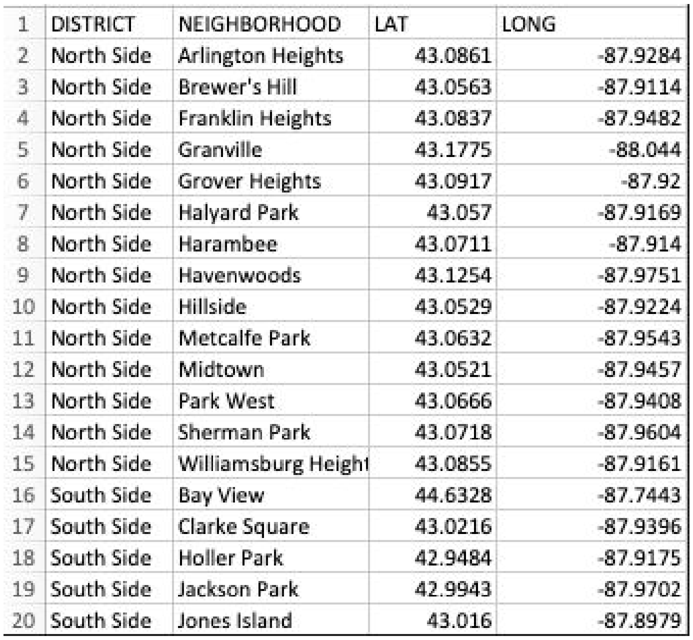
1. Wikipedia: Information about neighborhood names and districts in Milwaukee
2. Foursquare: Venue data for each neighborhood
3. Google/Google Maps: Latitude and longitude data for each neighborhood

Data Collection

The company is only interested in locations inside the city, so we will only be looking at neighborhoods inside the city of Milwaukee, not Milwaukee county as a whole. In addition, there are several neighborhoods with smaller sub-neighborhoods, but for this project, we will be ignoring the sub-neighborhoods.

First, I needed to create a CSV file with all the neighborhoods and their respective latitude and longitude. I got each neighborhood name from the Wikipedia page and the latitude and longitude from google searches. Some neighborhoods would not appear from google search so I went on google maps and manually found the center of the neighborhood and dropped a pin.

I will be using the Foursquare API to get venue data. I will use the neighborhood coordinates to query the API to get data about venues near each location.



1	DISTRICT	NEIGHBORHOOD	LAT	LONG
2	North Side	Arlington Heights	43.0861	-87.9284
3	North Side	Brewer's Hill	43.0563	-87.9114
4	North Side	Franklin Heights	43.0837	-87.9482
5	North Side	Granville	43.1775	-88.044
6	North Side	Grover Heights	43.0917	-87.92
7	North Side	Halyard Park	43.057	-87.9169
8	North Side	Harambee	43.0711	-87.914
9	North Side	Havenwoods	43.1254	-87.9751
10	North Side	Hillside	43.0529	-87.9224
11	North Side	Metcalfe Park	43.0632	-87.9543
12	North Side	Midtown	43.0521	-87.9457
13	North Side	Park West	43.0666	-87.9408
14	North Side	Sherman Park	43.0718	-87.9604
15	North Side	Williamsburg Height	43.0855	-87.9161
16	South Side	Bay View	44.6328	-87.7443
17	South Side	Clarke Square	43.0216	-87.9396
18	South Side	Holler Park	42.9484	-87.9175
19	South Side	Jackson Park	42.9943	-87.9702
20	South Side	Jones Island	43.016	-87.8979

MilwaukeeNeighborhoods.csv

Methodology

Feature Selection

I will be analyzing the neighborhoods based on the top 100 venues within a 950-meter radius. The ten most common categories for each neighborhood will then be used for features. In addition, we will be looking to see if there are college facilities and apartments nearby.

Python packages used

- Requests - HTTP requests
- Folium - Map rendering
- NumPy - Matrix operations
- Pandas - Data handling
- Matplotlib - Python plotting
- Sklearn - Machine learning

Algorithms

A K-Means was clustering algorithm was used for this analysis. It is an unsupervised algorithm that groups based on the similarity of features. Our dataset is unlabelled, which makes a k-means algorithm an appropriate choice. We will be using this to determine what types of different areas the city of Milwaukee has in terms of common venues. The algorithm will find patterns of similar venues and group them together. One challenge of k-means is that because of the randomly selected initial centroids, the result may not be a global optimum. To counter this, I will be running the algorithm several times and comparing the results. For the number of clusters, I will be using 5. This will help avoid over-fitting the relatively small number of neighborhoods.

Results

Clusters

Cluster 1 is mainly the upper west side of Milwaukee. It primarily consists of bars, with several parks, bed & breakfasts, and activities, like rock climbing or pool halls.

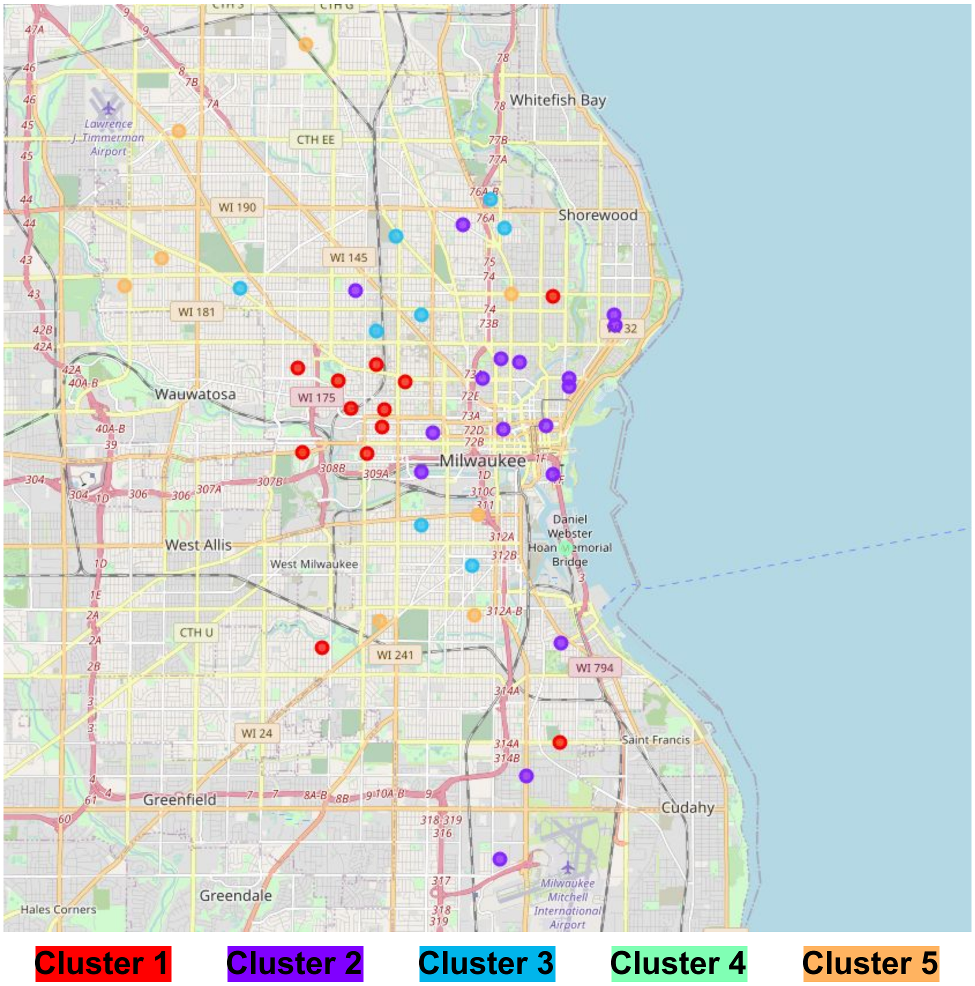
Cluster 2 is mostly the east coast, reaching a little west into downtown. It is primarily bars and places to eat or hang out, such as music venues and coffee shops. Additionally, this cluster has the University of Milwaukee in it.

Cluster 3 is on the north and south edges of downtown. Gyms, grocery stores, parks, and fast-food restaurants are the most common venues.

Cluster 4 is Jones Island, with the ferry port, Mariana, and a few parks that are right on the lake.

Cluster 5 is all far away from downtown and getting close to suburbs. Mexican restaurants are very popular, along with gas stations, grocery stores, and parks.

Milwaukee Neighborhoods with cluster labels



Discussion

Recommendations

Looking at the specific cluster data, we can quickly rule out cluster 4 as that is an island with little to no night-life, making it difficult to get to and out of the way for potential customers.

Clusters 3 and 5 have much more residential and daytime venues, such as parks, gyms, shops, and grocery stores, along with being far away from downtown and night-life.

Depending on the type of restaurant that is being opened, I would recommend either cluster 1 or 2. Cluster 1 is almost all bars, which would attract a lot of people to those areas at night, while cluster 2 is mostly restaurant and coffee shops, which would see more traffic during the day. In addition, cluster 2 has the University of Milwaukee campus, and college students love to go out and explore new places.

Conclusion

In this project, I used venue data retrieved from Foursquare to determine where the best location for a new restaurant/bar in Milwaukee, WI. This can information can now be used by the company to make an informed decision on where the most active areas for a certain category of venue are.