Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Coffee Shop in Cincinnati

By: Vineeth Kumar Kondamadugu

(July 2020)

Introduction:

once heard someone say *"A bad day with coffee is better than a good day without it"*. For many People, visiting Coffee Shop is a great way to relax and enjoy themselves during weekends and holidays. They can make it as their hobby, dine at, spend some time out with friends, find peace and even work at coffee shops. Coffee shops are like a one-stop destination for all coffee lovers. Property developers are also taking advantage of this trend to build more coffee shops to cater to the demand. As a result, there are many best coffee shops in the city of Cincinnati and many more are being built. Opening coffee shops allows lease owners as well as owners to earn consistent rental income. Of course, as with any business decision, opening a new coffee shop requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the coffee shop is one of the most important decisions that will determine whether the shop will be a success or a failure.

Business Problem

The objective of this capstone project is to analyze and select the best locations in the city of Cincinnati, Ohio to open a new Coffee Shop. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Cincinnati, Ohio, if a property developer or retailer or owner himself is looking to open a new coffee shop, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to property developers, investors, retailers and individuals looking to open or invest in new shopping malls in the city of Cincinnati, Ohio, USA. Current trend says that coffee shops are more successful with the growing busy workstyle of people around globe not only in Cincinnati. Simply put, Cincinnati has an impressive variety of coffee shops. Whether you're interested in finding a place to get some work done, looking for a place to meet a friend, seeking out a place to enjoy a caffeinated beverage, or searching for a spot to host a meetup, Cincinnati has coffee shops sprinkled throughout the city where you can do all of these things.

Data:

To solve the problem, we will need the following data:

- List of neighborhoods in Cincinnati. This defines the scope of this project which is confined to the city of Cincinnati, Ohio, United States of America.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to Coffee Shops. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Cincinnati) contains a list of neighborhoods in Cincinnati, Ohio, with a total of 52 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the "Coffee Shop" category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used

Methodology:

Firstly, we need to get the list of neighborhoods in the city of Cincinnati. Fortunately, the list is available in the Wiki page (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Cincinnati). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas Data Frame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Kuala Lumpur.

(52, 3)

	Neighborhood	Latitude	Longitude
1	Avondale, Cincinnati	39.14771	-84.49490
2	Bond Hill, Cincinnati	39.17460	-84.46715
3	California, Cincinnati	39.06536	-84.42365
4	Camp Washington, Cincinnati	39.13691	-84.53730
5	Carthage, Cincinnati	39.19733	-84.48062

Figure: data frame head of neighborhoods of Cincinnati

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of

occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Coffee Shop” data, we will filter the “coffee Shop” as venue category for the neighborhoods.

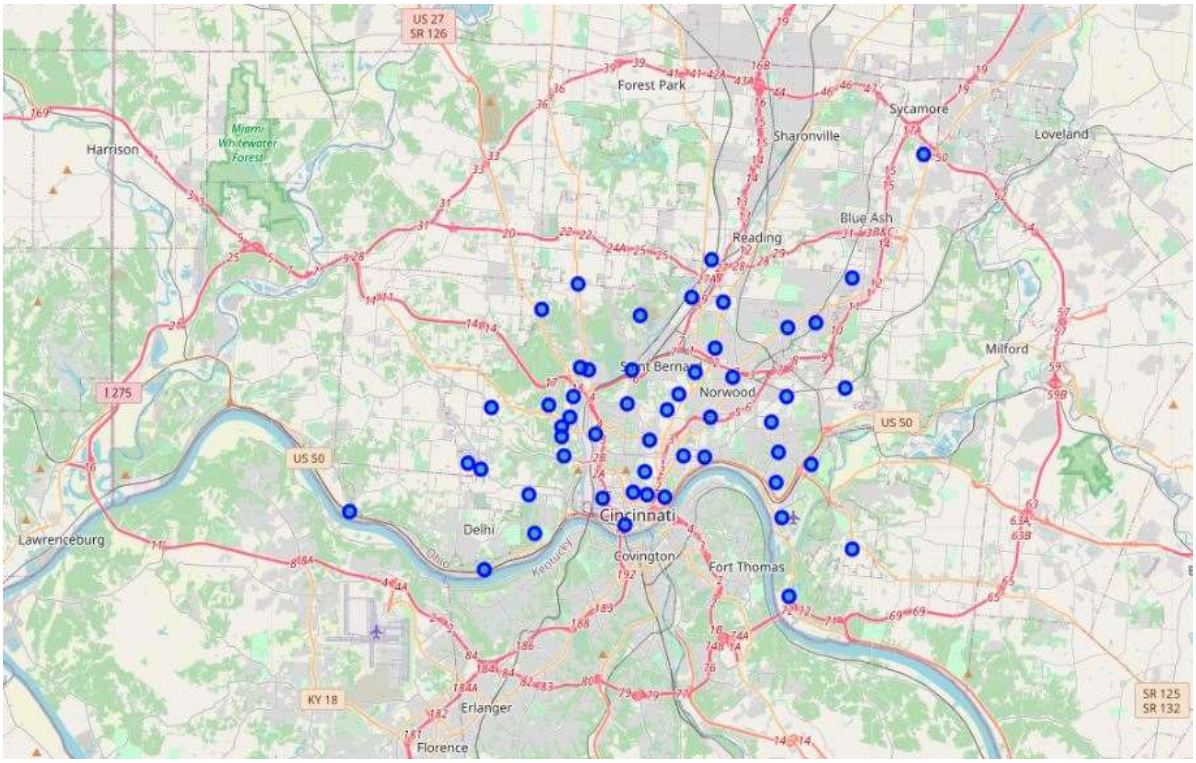

Figure: neighborhoods map of Cincinnati

(2978, 7)

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Avondale, Cincinnati	39.14771	-84.4949	Hippo Cove	39.145257	-84.506104	Zoo Exhibit
1	Avondale, Cincinnati	39.14771	-84.4949	Dobsa Giraffe Ridge	39.143495	-84.506975	Zoo Exhibit
2	Avondale, Cincinnati	39.14771	-84.4949	Jungle Trails	39.146071	-84.506643	Zoo Exhibit
3	Avondale, Cincinnati	39.14771	-84.4949	Cincinnati Zoo & Botanical Garden	39.142740	-84.509266	Zoo
4	Avondale, Cincinnati	39.14771	-84.4949	Marge Schott-Unnewehr Elephant Reserve	39.143109	-84.508114	Zoo Exhibit
5	Avondale, Cincinnati	39.14771	-84.4949	Kroger Lords of the Arctic	39.145949	-84.507424	Zoo Exhibit

Figure: Data frame of venues with their venue categories in Cincinnati

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, here we choose the k value as 3 and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Coffee Shop”. The results will allow us to identify which neighborhoods have higher concentration of coffee shops while which neighborhoods have fewer number of coffee shops. Based on the occurrence of coffee shops in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new coffee shops.

	Neighborhoods	ATM	Accessories Store	Advertising Agency	Airport	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automotive Shop	BBQ Joint
0	Avondale, Cincinnati	0.026316	0.00000	0.0	0.0	0.0	0.026316	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.013156	0.00000	0.0
1	Bond Hill, Cincinnati	0.020833	0.00000	0.0	0.0	0.0	0.041667	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0
2	CUF, Cincinnati	0.000000	0.01087	0.0	0.0	0.0	0.054348	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.01087	0.0
3	California, Cincinnati	0.000000	0.00000	0.0	0.0	0.0	0.062500	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.031250	0.00000	0.0
4	Camp Washington, Cincinnati	0.011905	0.00000	0.0	0.0	0.0	0.035714	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000	0.0

Figure: Data frame head of neighbourhoods after one hot encoding on venue categories and mean of frequency of occurances

Results:

The results from the k-means clustering show that we have categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Coffee Shop”:

- Cluster 0: Neighborhoods with high concentration number of coffee shops
- Cluster 1: Neighborhoods with low number to no existence of coffee shops
- Cluster 2: Neighborhoods with moderate concentration of coffee shops

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.

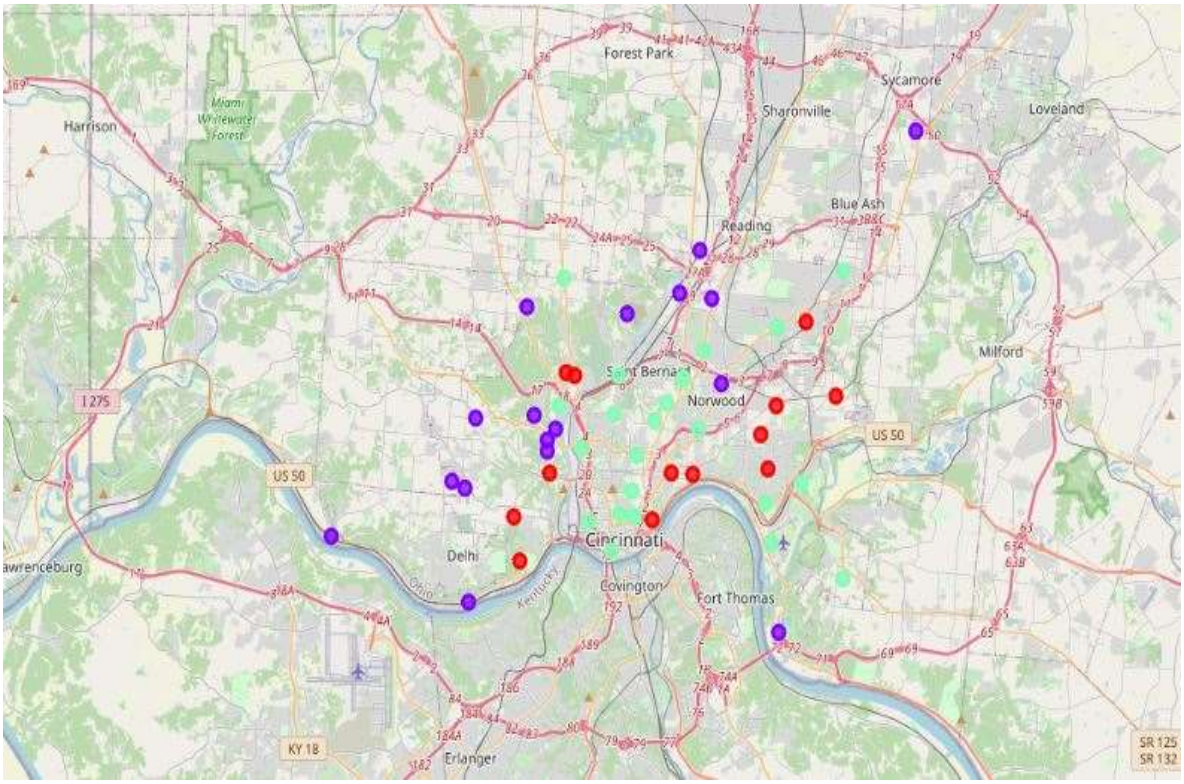

Figure: 3 Clusters of neighborhoods of Cincinnati

Discussion:

As observations noted from the map in the Results section, most of the shopping malls are concentrated in the central area of Cincinnati with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to no coffee shops in the neighborhoods. This represents a great opportunity and high potential areas to open new coffee shops as there is very little to no competition from existing shops. Meanwhile, coffee shops in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of coffee shops. From another perspective, the results also show that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still have very few coffee shops. Therefore, this project recommends property developers and individuals to capitalize on these findings to open new coffee shops in neighborhoods in cluster 1 with little to no competition. Property individuals with unique selling propositions to stand out from the competition can also open new coffee shop in neighborhoods in cluster 0 with moderate competition. Lastly, individuals and developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of coffee shops and suffering from intense competition.

	Neighborhood	Coffee Shop	Cluster Labels	Latitude	Longitude
35	Pill Hill, Cincinnati	0.046512	0	39.166175	-84.54628
21	Madisonville, Cincinnati	0.048387	0	39.157380	-84.39103
13	East Walnut Hills, Cincinnati	0.056338	0	39.126660	-84.47332
42	Sedamsville, Cincinnati	0.050000	0	39.093040	-84.57292
23	Mount Adams, Cincinnati	0.050000	0	39.109200	-84.49639
19	Kennedy Heights, Cincinnati	0.062500	0	39.185740	-84.40819
26	Mount Lookout, Cincinnati	0.050000	0	39.128730	-84.43022
44	South Fairmount, Cincinnati	0.055556	0	39.127560	-84.55538
30	Northside, Cincinnati	0.043478	0	39.165470	-84.54113
31	Oakley, Cincinnati	0.060000	0	39.153330	-84.42486
37	Price Hill, Cincinnati	0.066667	0	39.110340	-84.57603
47	Walnut Hills, Cincinnati	0.043478	0	39.127200	-84.48541
18	Hyde Park, Cincinnati	0.050000	0	39.142190	-84.43423

Figure: Cluster 0 of the neighborhoods where there is high frequency of coffee shops

	Neighborhood	Coffee Shop	Cluster Labels	Latitude	Longitude
50	Westwood, Cincinnati	0.015385	1	39.148570	-84.598490
39	Riverside, Cincinnati	0.000000	1	39.077460	-84.602230
24	Mount Airy, Cincinnati	0.000000	1	39.191720	-84.568930
29	North Fairmount, Cincinnati	0.000000	1	39.135890	-84.556870
22	Millvale, Cincinnati	0.000000	1	39.144430	-84.552070
40	Roselawn, Cincinnati	0.000000	1	39.194780	-84.462430
51	Winton Hills, Cincinnati	0.000000	1	39.189320	-84.511070
41	Sayler Park, Cincinnati	0.000000	1	39.102910	-84.681310
15	English Woods, Cincinnati	0.000000	1	39.140220	-84.557210
14	East Westwood, Cincinnati	0.000000	1	39.149910	-84.564190
10	Covedale, Cincinnati	0.000000	1	39.121430	-84.604080
46	Villages at Roll Hill, Cincinnati	0.000000	1	39.259697	-84.345048
5	Carthage, Cincinnati	0.000000	1	39.197330	-84.480620
3	California, Cincinnati	0.000000	1	39.065360	-84.423650
2	CUF, Cincinnati	0.010870	1	39.162000	-84.456890
49	Western Hills, Cincinnati	0.000000	1	39.123986	-84.611668
17	Hartwell, Cincinnati	0.000000	1	39.213390	-84.469370

Figure: Cluster 1 of the neighborhoods where there is no or very little frequency of coffee shops

	Neighborhood	Coffee Shop	Cluster Labels	Latitude	Longitude
43	South Cumminsville, Cincinnati	0.037736	2	39.153460	-84.550050
45	Spring Grove Village, Cincinnati	0.028571	2	39.165078	-84.515793
38	Queensgate, Cincinnati	0.020000	2	39.205725	-84.386436
48	West End, Cincinnati	0.040000	2	39.108820	-84.532720
36	Pleasant Ridge, Cincinnati	0.038462	2	39.183580	-84.424690
0	Avondale, Cincinnati	0.039474	2	39.147710	-84.494900
33	Paddock Hills, Cincinnati	0.035714	2	39.164370	-84.478390
32	Over-the-Rhine	0.040000	2	39.111450	-84.515220
28	North Avondale, Cincinnati	0.034483	2	39.154320	-84.488440
27	Mount Washington, Cincinnati	0.029412	2	39.086250	-84.386510
20	Linwood, Cincinnati	0.025641	2	39.123910	-84.410390
16	Evanston, Cincinnati	0.028169	2	39.144520	-84.469470
12	East End, Cincinnati	0.025000	2	39.099960	-84.427780
11	Downtown Cincinnati	0.040000	2	39.097230	-84.519640
9	Corryville, Cincinnati	0.040000	2	39.134320	-84.505740
8	Columbia-Tusculum, Cincinnati	0.028571	2	39.115690	-84.431540
7	College Hill, Cincinnati	0.023256	2	39.202870	-84.547680
6	Clifton, Cincinnati	0.028571	2	39.150270	-84.518250
4	Camp Washington, Cincinnati	0.023810	2	39.136910	-84.537300
1	Bond Hill, Cincinnati	0.020833	2	39.174600	-84.467150
34	Pendleton, Cincinnati	0.040000	2	39.110302	-84.506903
25	Mount Auburn Historic District	0.040000	2	39.120340	-84.508280

Figure: Cluster 2 of the neighborhoods where there is moderate frequency of coffee shops

Limitations and Suggestions for Future Research:

In this project, we only consider one factor i.e. frequency of occurrence of coffee shops, there are other factors such as population and income of residents and surroundings that could influence the location decision of a new coffee shops. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new coffee shops. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion:

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new coffee shops. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new coffee shop.

References

Category: Neighborhoods of Cincinnati, Ohio. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Cincinnati

Foursquare Developers Documentation. *Foursquare*. Retrieved from https://developer.foursquare.com/docs