Capstone Project – Battle of the Neighborhoods

Applied Data Science Capstone Project by IBM/Coursera

Using Data to Establish Location of a New Bar/Nightclub in Milton ON

By: Richard Moss

Date: June 28, 2021

Introduction: Business Problem

In this project I will try to find an optimal location for a bar and nightclub. Specifically, this report will be targeted to stakeholders interested in opening a bar in Milton, Ontario, Canada.

Since there are lots of restaurants, bars, and pubs in Milton I will try to detect locations that are not already crowded with these amenities. I am particularly interested in areas with no bars, lounges, or pubs. I would also prefer locations as close to city center as possible, close to bus routs, and most importantly: close to the location of Milton's two new University Campuses in association with Wilfrid Laurier University and Conestoga College, known as the Milton Education Village, assuming that first two conditions are met.

I will use data science methodologies to generate a few neighborhoods of interest based on these criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Data

Based on definition of our problem, factors that will influence my decision are:

- number of existing bars, pubs, nightclubs and lounges in the neighborhood
- number of bus stops or stations
- distance of neighborhood from new University Campuses

I decided to use radial locations of 2 kilometers around the geographical center of each neighborhood.

Following data sources will be needed to extract/generate the required information:

- Centers of each neighborhood were acquired from Google Earth and stored in a CSV file. This CSV file was loaded into a pandas data frame, where **Foursquare API** venues method was used to extract nearby venues for analysis of each neighborhood.
- Number of venues and their type and location in every neighborhood will be obtained using **Foursquare API.**
- Coordinate of new University Campus are 43.484091, -79.883314, and the distance from each neighborhood to this location will be analyzed.

Neighborhood Database Creation

The neighborhood database was created by using geocode to acquire the latitude and longitude of each location. The location of the Milton Education Village was found in the same manor, and then each neighborhood was plotted using Folium in blue, with the key Milton Education Village location shown in red on the map below.

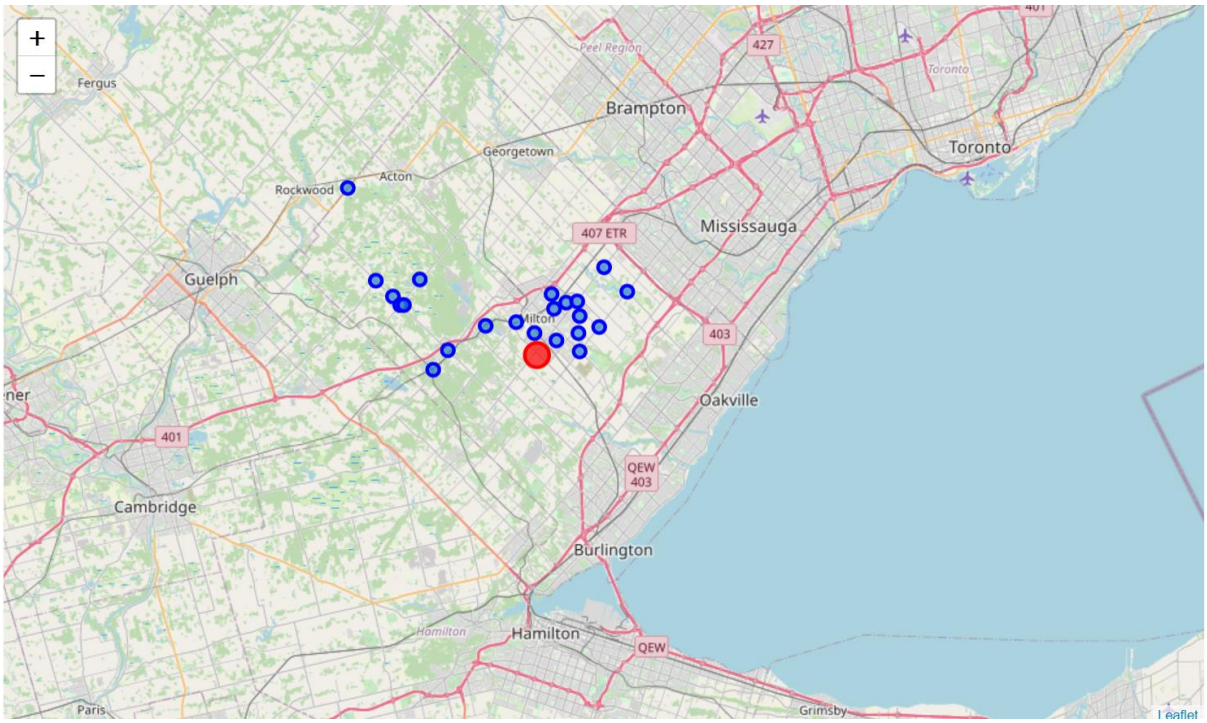

Figure 1. Milton Neighborhoods (blue) and Milton Education Village Development Site (Red)

The distance from each neighborhood to the development were calculated based on the Pythagorean theorem, plotting each neighborhood on an x-y coordinate system using pyproj, and converting the latitude and longitude coordinates. The distance was then appended to the data frame and sorted. The five closest to the neighborhoods are shown in the data frame section below, with the distance measured in meters.

	Neighborhood	Latitude	Longitude	x-coord	y-coord	Distance to Milton Education Village
19	Scott	43.501648	-79.886141	-5.327883e+06	1.056847e+07	2811.650124
21	Wilmott	43.495457	-79.860700	-5.329232e+06	1.056565e+07	3195.957976
17	Peru	43.510885	-79.905953	-5.326128e+06	1.057059e+07	5024.993714
4	Boyne	43.486744	-79.834765	-5.330988e+06	1.056281e+07	5685.718321
1	Ash	43.501779	-79.836021	-5.328577e+06	1.056266e+07	6198.572993

Figure 2. Database of Milton Neighborhoods sorted by Distance to Milton Education Village

Foursquare API Implementation

Now that I have my location candidates, I can use Foursquare API to get info on venues in each neighborhood.

I'm interested in venues in the 'food', 'travel and transport' and 'nightlife spot' categories, but within the food category I am only interested in fast food, and my travel and transport query will focus on buses. The capture below indicates the shape of the data frame, which means that 226 venues of interest based

on the categories I passed to the Foursquare API were found within 2km of each neighbourhood. Some of these are shown below.

```
(226, 7)
  Neighborhood  Neighborhood Latitude  Neighborhood Longitude  \
0        Scott              43.501648              -79.886141
1        Scott              43.501648              -79.886141
2        Scott              43.501648              -79.886141
3        Scott              43.501648              -79.886141
4        Scott              43.501648              -79.886141

                              Venue  Venue Latitude  Venue Longitude  \
0      The Works Gourmet Burger Bistro       43.512589       -79.883587
1                         Troy's Diner       43.515083       -79.881386
2             Halifax Donair and Pizza       43.515091       -79.881277
3                      Jay's Ice Cream       43.511843       -79.884502
4  Jay's Ice Cream & Sunshine's Gelato       43.511750       -79.884449

        Venue Category
0         Burger Joint
1  American Restaurant
2          Pizza Place
3       Ice Cream Shop
4       Ice Cream Shop
```

Figure 3. Header of Venues of Interest Data Frame

The code used to generate this data frame is shown below.

```
# Category IDs corresponding to venus of interest were taken from Foursquare web site (https://developer.foursquare.com/docs/res

categories_of_interest = ['4bf58dd8d48988d16c941735', '52e81612bcbc57f1066b7a00', '4bf58dd8d48988d1c9941735',
                          '4bf58dd8d48988d16e941735', '4d4ae6fc7a7b7dea34424761', '4bf58dd8d48988d1ca941735',
                          '4bf58dd8d48988d1c7941735', '4d4b7105d754a06376d81259', '4bf58dd8d48988d1fe931735',
                          '4bf58dd8d48988d12b951735', '52f2ab2ebcbc57f1066b8b4f']


def getNearbyVenues(names, latitudes, longitudes, radius=2000):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)
        for category in categories_of_interest :
            # create the API request URL
            url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&categoryId={}&radius
                CLIENT_ID,
                CLIENT_SECRET,
                VERSION,
                lat,
                lng,
                category,
                radius,
                LIMIT,
                )

            # make the GET request
            results = requests.get(url).json()["response"]['groups'][0]['items']

            # return only relevant information for each nearby venue
            venues_list.append([(
                name,
                lat,
                lng,
                v['venue']['name'],
                v['venue']['location']['lat'],
                v['venue']['location']['lng'],
                v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                'Neighborhood Latitude',
                'Neighborhood Longitude',
                'Venue',
                'Venue Latitude',
                'Venue Longitude',
                'Venue Category']

    return(nearby_venues)
```

Figure 4. Code Used to Acquire Venues of Interest

Methodology

In this project I will direct my efforts on detecting areas of Milton that have low nightlife establishment density, and high fast food and transportation density. I will evaluate potential areas of interest derived from these criteria based on distance from the Milton Education Village development.

In first step I collected the required **data: location and type (category) of every venue of interest within 2km from each neighborhood center**.

Second step in the analysis will be calculation and exploration of '**venue-type density**' across each neighbourhood.

In third and final step I will focus on most promising areas and within those create **clusters of locations that meet some basic requirements** established in discussion with stakeholders. I will present map of all such locations but also create clusters (using **k-means clustering**) of those locations to identify general zones / neighborhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

Of note during this process, Scott, Wilmott, Peru, Boyne and Ash are the closet neighborhoods to the development, and this will be a key evaluation factor.

Analysis

Let's perform some basic explanatory data analysis and derive some additional info from the raw data. First, I counted the **number of venues of interest in each neighborhood**.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Agerton	1	1	1	1	1	1
Ash	10	10	10	10	10	10
Beaty	13	13	13	13	13	13
Boyne	3	3	3	3	3	3
Campbellville	3	3	3	3	3	3
Clarke	24	24	24	24	24	24
Darbyville	1	1	1	1	1	1
Dempsey	28	28	28	28	28	28
Guelph Junction	1	1	1	1	1	1
Hawthorne Village	8	8	8	8	8	8
Kelso	4	4	4	4	4	4
Omagh	3	3	3	3	3	3
Peru	25	25	25	25	25	25
Scott	29	29	29	29	29	29
Timberlea	47	47	47	47	47	47
Wilmott	26	26	26	26	26	26

Figure 5. Count of Venues of Interest in Venues Data Frame Grouped by Neighborhood

These venues were then one hot encoded and grouped by the mean frequency of each category. This allowed me to group determine the most common venues in each neighborhood, two examples of which are shown below.

```
----Agerton----
                 venue  freq
0  Fast Food Restaurant   1.0
1   American Restaurant   0.0
2             Gastropub   0.0
3            Sports Bar   0.0
4             Speakeasy   0.0


----Ash----
                 venue  freq
0           Restaurant   0.3
1          Pizza Place   0.3
2  Fast Food Restaurant   0.2
3          Burger Joint   0.1
4             Gastropub   0.1
```

Figure 6. Venue Frequency in Agerton and Ash, Milton

I then took the 10 most common venue categories in each neighborhood and created a new pandas data frame by neighborhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agerton	Fast Food Restaurant	American Restaurant	Gastropub	Sports Bar	Speakeasy	Restaurant	Pub	Pizza Place	Nightlife Spot	Lounge
1	Ash	Restaurant	Pizza Place	Fast Food Restaurant	Burger Joint	Gastropub	American Restaurant	Ice Cream Shop	Sports Bar	Speakeasy	Pub
2	Beaty	Pizza Place	Restaurant	Fast Food Restaurant	Burger Joint	Bus Line	Bus Station	Gastropub	American Restaurant	Italian Restaurant	Sports Bar
3	Boyne	Pizza Place	Fast Food Restaurant	Ice Cream Shop	American Restaurant	Gastropub	Sports Bar	Speakeasy	Restaurant	Pub	Nightlife Spot
4	Campbellville	Steakhouse	Speakeasy	Bus Station	Gastropub	Sports Bar	Restaurant	Pub	Pizza Place	Nightlife Spot	Lounge

Figure 7. Data frame of most Common Venues by Neighborhood

Finally, I used k-means clustering to create a map which groups the neighborhoods by similarity of establishments in the neighborhood and proximity to neighborhoods of similar types, shown below.

Figure 8. K-Means Clustering of Milton Neighborhoods

This concludes our analysis. As it turns out, a lot of these neighborhoods are particularly similar. Perhaps this is to be expected when our focus is on a medium sized suburban area. However, this can be used to our benefit. We can use the 10 most common venue types for the 5 closest neighborhoods the Milton

Education Village to best choose our new bar/nightclub location. This also shows that cluster 1 contains the greatest number of locations of interest, focused on the Milton downtown core, and any location of interest should fall in cluster 1 (labelled cluster 0 on our map).

Results and Discussion

Let us look at each of the 5 neighborhoods on our shortlist in detail to determine the best spot for the new bar.

1) Scott

Scott is the closest neighborhood to the Milton Education Village, and has a high density of pizza places, fast food shops, ice cream shops which make it a good candidate for my new bar. However, there is not easy access to transportation, and there are other pubs and sports bars in the area which could act as competition.

2) Wilmott

Wilmott's 5 most frequent venues are all food establishments, however, the sixth and seventh spot on the venue density list tells us that there are other bars in Wilmott, as well as pubs in ninth, and no access to transportation. This makes Wilmott a less appealing option.

3) Peru

Peru is an incredibly promising location. Peru's top four venues are fast food establishments, fifth and ninth show Peru has a high density of bus lines and stations, there are no pubs or bars, and while there is some nightlife spot density, it is at tenth place on the list.

4) Boyne

Boyne has a number of nightlife spots, sports bars and speakeasys and no transportation access. With this in mind, and it being further from the development site from Peru, it is less appealing.

5) Ash

Ash has a large number of fast-food joints, but with no access to transportation, and it being further from the development site as well as spots 8-10 being direct competitors to the new bar I am hoping to establish.

Conclusion

Purpose of this project was to identify Milton neighborhoods close to the Milton Education village with a low number of nightlife establishments (particularly bars and nightclubs) and a high number of fast food restaurants and transportation access in order to aid stakeholders in narrowing down the search for optimal location for a new bar or nightclub. By calculating venue-of-interest density distribution from Foursquare data I have first identified neighborhoods that justify further analysis, and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby establishments. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations).

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc. however, it is strongly recommended that Peru be considered as an optimal neighborhood for the reasons listed above.