IBM Data Science Capstone Project

Finding the optimal location for a bubble tea store in City of Calgary

Anyi W.
January 26, 2021

Introduction

Bubble tea as a new type of beverage consumption, has won its own beverage market over the past ten years, especially among younger generation. It becomes one of the most populate beverage over the world, not only because of the innovative idea of adapting a variety of flavors, but also remaining healthy by using real natural tea as base. In 2019, global bubble tea market was worth USD 2.1 billion, and expecting to reach $4.3 billion and a compound annual growth rate (CAGR) of 7.8% in 2027. [i]

Business Problem

Increasing popularity has attracted many investors to open bubble tea stores in their own city, chained or non-chained. The main purpose of this project is to help investors to find a potential optimal neighbourhood to open a new bubble tea store in City of Calgary. The project is aiming to provide an analysis of population density of the city of Calgary, using Machine Learning methodologies to cluster neighbourhoods, and accessing to Foursquare API to obtain the venues in the neighbourhoods. The recommendation of the optimal location will be made based on analysis of population density, density of restaurant.

Data Sources

Following data sources will be used in order to perform the analysis of the project:

- City of Calgary - Community by Sector, which contains the communities names in City of Calgary, belonging sectors and corresponding coordinates. (*Last Updated: September 11,* 2020 https://data.calgary.ca/Base-Maps/Communities-by-Sector/e6xg-kaxf)
- A list of top venues in these neighbourhood is acquired using Foursquare API

Data Cleaning

The City of Calgary - Community by Sector data was obtained in csv format. And it contains more comprehensive aspects of the community distribution. For this project, only data related to residential communities will be used. Therefore, we will do some clean-up to make it more accessible and simple. Detailed procedures and codes are below:

	CLASS	CLASS_CODE	COMM_CODE	NAME	SECTOR	SRG	COMM_STRUCTURE	longitude	latitude	location
0	Residential	1	THS	TWINHILLS	EAST	DEVELOPING	BUILDING OUT	-113.877110	51.045111	(51.045111353378694, -113.87710975220665)
1	Residential	1	WIL	WILLOW PARK	SOUTH	BUILT-OUT	1960s/1970s	-114.056204	50.956623	(50.95662292848714, -114.05620363150967)
2	Residual Sub Area	4	05D	05D	NORTHEAST	NaN	UNDEVELOPED	-113.958662	51.179598	(51.17959764644064, -113.95866183876556)
3	Industrial	2	ST4	STONEY 4	NORTHEAST	NaN	EMPLOYMENT	-114.002762	51.176204	(51.17620448693238, -114.00276157771617)
4	Residential	1	PKH	PARKHILL	CENTRE	BUILT-OUT	1950s	-114.065552	51.018181	(51.01818071993347, -114.06555236114401)

Figure 1 – Sample data of Community by Sector data, before data cleaning

1. Keep residential communities:

```
# This practise only focus on Residential data, only keep rows with class = Residential
city_data = city_data[city_data["CLASS"] == "Residential"]
city_data.head()
```

	CLASS	CLASS_CODE	COMM_CODE	NAME	SECTOR	SRG	COMM_STRUCTURE	longitude	latitude	location
0	Residential	1	THS	TWINHILLS	EAST	DEVELOPING	BUILDING OUT	-113.877110	51.045111	(51.045111353378694, -113.87710975220665)
1	Residential	1	WIL	WILLOW PARK	SOUTH	BUILT-OUT	1960s/1970s	-114.056204	50.956623	(50.95662292848714, -114.05620363150967)
4	Residential	1	PKH	PARKHILL	CENTRE	BUILT-OUT	1950s	-114.065552	51.018181	(51.01818071993347, -114.06555236114401)
5	Residential	1	PAT	PATTERSON	WEST	BUILT-OUT	1980s/1990s	-114.177047	51.063838	(51.06383775082155, -114.17704650860274)
6	Residential	1	RCK	ROSSCARROCK	WEST	BUILT-OUT	1950s	-114.145495	51.043280	(51.04328023810093, -114.14549516107789)

Figure 2 – Sample data of Community by Sector data, selecting Residential communities

2. Remove not necessary columns

```
#Only keep community name, lat and long
df = city_data[["NAME","longitude","latitude"]].copy()
df.head()
```

	NAME	longitude	latitude
0	TWINHILLS	-113.877110	51.045111
1	WILLOW PARK	-114.056204	50.956623
4	PARKHILL	-114.065552	51.018181
5	PATTERSON	-114.177047	51.063838
6	ROSSCARROCK	-114.145495	51.043280

Figure 3 – Sample data of Community by Sector data, after removed necessary columns

3. Change column name and capitalize string content

```
#Change column name from NAME to Community
df.rename(columns= {"NAME":"Community"}, inplace = True)
df.columns
```

```
Index(['Community', 'longitude', 'latitude'], dtype='object')
```

```
#Capitalize community column
df['Community'] = df['Community'].str.capitalize()
df.head()
```

	Community	longitude	latitude
0	Twinhills	-113.877110	51.045111
1	Willow park	-114.056204	50.956623
4	Parkhill	-114.065552	51.018181
5	Patterson	-114.177047	51.063838
6	Rosscarrock	-114.145495	51.043280

Figure 4 – Sample data of Community by Sector data, after capitalizing community names

4. Reset index of the data frame

```
#Reset index
df.reset_index(drop = True, inplace = True)
df.head()
```

	Community	longitude	latitude
0	Twinhills	-113.877110	51.045111
1	Willow park	-114.056204	50.956623
2	Parkhill	-114.065552	51.018181
3	Patterson	-114.177047	51.063838
4	Rosscarrock	-114.145495	51.043280

Figure 5 – Sample data of Community by Sector data, after reset index of the Pandas data frame

Now the city data is ready to be used!

Methodologies

Data Organization

To explore distribution of venues in communities in Calgary, we need to get the list of communities in Calgary and corresponding coordinates (Latitude and Longitude), which is already included in the City of Calgary - Community by Sector City of Calgary. The dataset is obtained in csv format, we will need to convert to Python Pandas data frame, clean and normalize the data to be readable and understandable. Detailed procedures can be found in last section in this document – *Data Cleaning*. Below is the sample data post clean-up:

	Community	longitude	latitude
0	Twinhills	-113.877110	51.045111
1	Willow park	-114.056204	50.956623
2	Parkhill	-114.065552	51.018181
3	Patterson	-114.177047	51.063838
4	Rosscarrock	-114.145495	51.043280

Figure 6 – Sample Data of Calgary Community

Data Visualization

Visualize the Calgary communities in a Calgary map by using **Python Folium library**. Which represents most neighborhoods located in the center of the city, we can possibly conclude the population density in the center area is the highest.

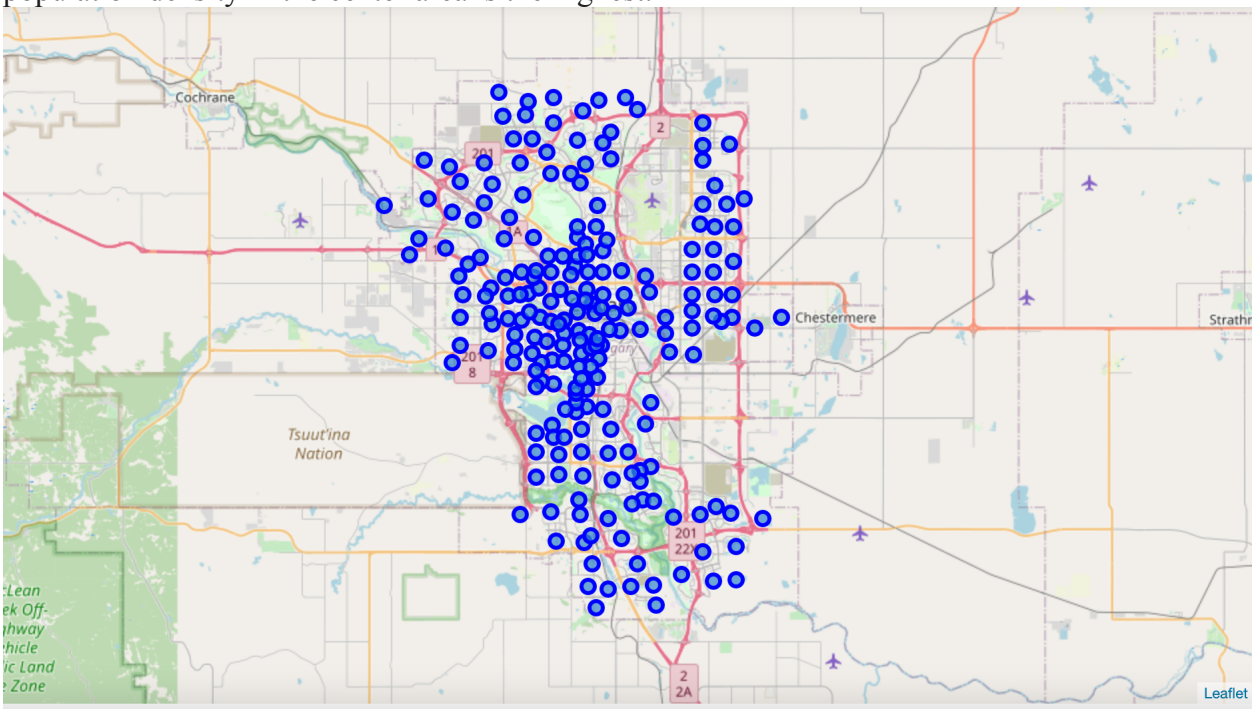

Figure 7 – Residential communities plotted in Calgary map

Foursquare API

The venue data for the City of Calgary was obtained by logging in to Foursquare developer API. From there, we gathered the venue data to analyze the distribution of venue in City of Calgary. A python function was created in order to process all the venue in the city of Calgary and also appended to a new data frame contains the neighborhoods, coordinates and venue categories of each venue related to, which will be used for upcoming analysis.

```
def getNearbyVenues(names, latitudes, longitudes, radius=500):
    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)
    # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limi
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            100000)

     # 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 8 – Python function created to possess venue data

```
# Run the function created above on each neighborhood and createa new dataframe Calgary_venues
Calgary_venues = getNearbyVenues(names=df['Community'],
                                 latitudes=df['latitude'],
                                 longitudes=df['longitude'])
```

```
Symons valley ranch
Bridlewood
Oakridge
Hidden valley
Cliff bungalow
Marlborough park
Rundle
Rideau park
Hillhurst
Martindale
```

Figure 9 – Run the function to make the list

```
# Review the new created dataframe
print(Calgary_venues.shape)
Calgary_venues.head()
```

```
(1185, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Willow park	50.956623	-114.056204	Michael Hill	50.952635	-114.059239	Jewelry Store
1	Parkhill	51.018181	-114.065552	Axe Music	51.017012	-114.063163	Music Store
2	Parkhill	51.018181	-114.065552	Annex Ale Project	51.015039	-114.062072	Brewery
3	Parkhill	51.018181	-114.065552	Stanley Park	51.017171	-114.071570	Park
4	Parkhill	51.018181	-114.065552	Salt & Pepper	51.014624	-114.065525	Mexican Restaurant

Figure 10 – List of Venues in Calgary sample data

We can also extract the restaurants from the list to see how the restaurants distributed in the city of Calgary. By plotting the restaurants from the newly created list of venues can also provide some useful information for us to make decision. From where we can see most restaurants located in the very center of the city, which is known as downtown Calgary. Bubble tea as a type of beverage, can be consumed after visiting restaurant. This can help us narrow down the optimal location of new bubble tea locations.

```
calgary_restro = Calgary_venues[Calgary_venues['Venue Category'].str.contains("Restaurant")]
calgary_restro.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
4	Parkhill	51.018181	-114.065552	Salt & Pepper	51.014624	-114.065525	Mexican Restaurant
5	Parkhill	51.018181	-114.065552	Alloy	51.016225	-114.060050	American Restaurant
6	Parkhill	51.018181	-114.065552	Sushi Ichiban	51.017674	-114.063167	Sushi Restaurant
7	Parkhill	51.018181	-114.065552	Seoul BBQ Restaurant	51.014878	-114.064625	Korean BBQ Restaurant
12	Parkhill	51.018181	-114.065552	McDonald's	51.019193	-114.061605	Fast Food Restaurant

Figure 11 – Restaurants Sample Data

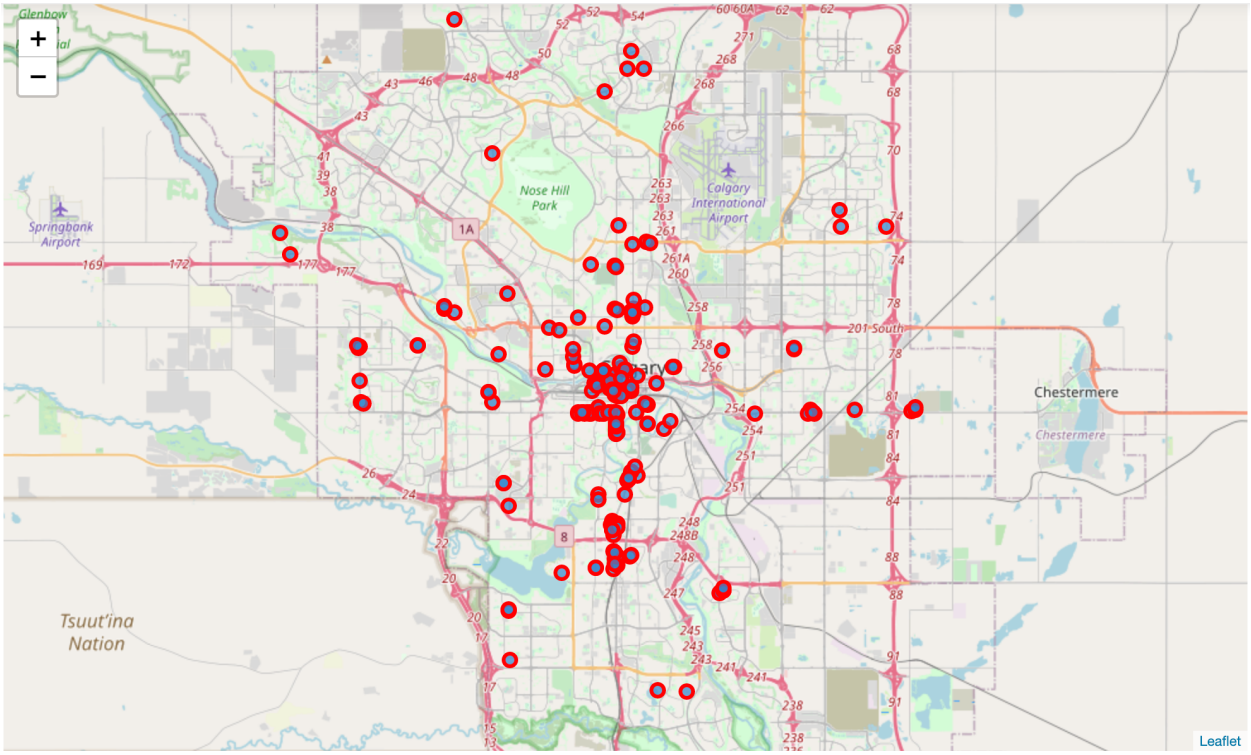

Figure 12 – Distribution of Restaurants in Calgary map

Analyze Each Neighborhoods

This project also analyzed each neighborhood by the top 5 most common venue categories, which will be valuable for the neighborhood clustering. To do this, one-hot encoding was used to enable the data to be used for machine learning algorithm.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Abbeydale	Wings Joint	Convenience Store	Health & Beauty Service	Sandwich Place	Diner
1	Acadia	Pool	Gym / Fitness Center	Women's Store	Fast Food Restaurant	Gas Station
2	Albert park/radisson heights	Light Rail Station	Train Station	Rock Club	Farmers Market	Gas Station
3	Altadore	Pub	Ice Cream Shop	Dog Run	Spa	Women's Store
4	Applewood park	Liquor Store	Construction & Landscaping	Trail	Home Service	Park

Figure 13 – One-hot Coding Sample Data

Clustering Neighborhoods

Since we are dealing with unsupervised data, the most common clustering method recommended in the project is K-Means Clustering. We used K-Means Clustering algorithm to segment neighborhoods in the city.

K was decided to be 10 for this case. However, there are some outliners which don't have various of venues opened in the neighborhoods. Therefore, we will eliminate those neighborhoods.

	Community	longitude	latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Twinhills	-113.877110	51.045111	NaN	NaN	NaN	NaN	NaN	NaN
1	Willow park	-114.056204	50.956623	1.0	Jewelry Store	Women's Store	Fast Food Restaurant	Gastropub	Gas Station
2	Parkhill	-114.065552	51.018181	1.0	Fast Food Restaurant	Sushi Restaurant	Bar	Snack Place	Brewery
3	Patterson	-114.177047	51.063838	1.0	Bar	Gas Station	Pizza Place	Vietnamese Restaurant	Convenience Store
4	Rosscarrock	-114.145495	51.043280	1.0	Sporting Goods Shop	Ice Cream Shop	Japanese Restaurant	Fast Food Restaurant	Juice Bar

```
calgary_merged = calgary_merged.dropna()
calgary_merged['Cluster Labels'] = calgary_merged['Cluster Labels'].astype('int64')
calgary_merged.head()
```

	Community	longitude	latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Willow park	-114.056204	50.956623	1	Jewelry Store	Women's Store	Fast Food Restaurant	Gastropub	Gas Station
2	Parkhill	-114.065552	51.018181	1	Fast Food Restaurant	Sushi Restaurant	Bar	Snack Place	Brewery
3	Patterson	-114.177047	51.063838	1	Bar	Gas Station	Pizza Place	Vietnamese Restaurant	Convenience Store
4	Rosscarrock	-114.145495	51.043280	1	Sporting Goods Shop	Ice Cream Shop	Japanese Restaurant	Fast Food Restaurant	Juice Bar
5	Acadia	-114.053702	50.972407	1	Pool	Gym / Fitness Center	Women's Store	Fast Food Restaurant	Gas Station

Figure 14 – Clustering data before & after clean up

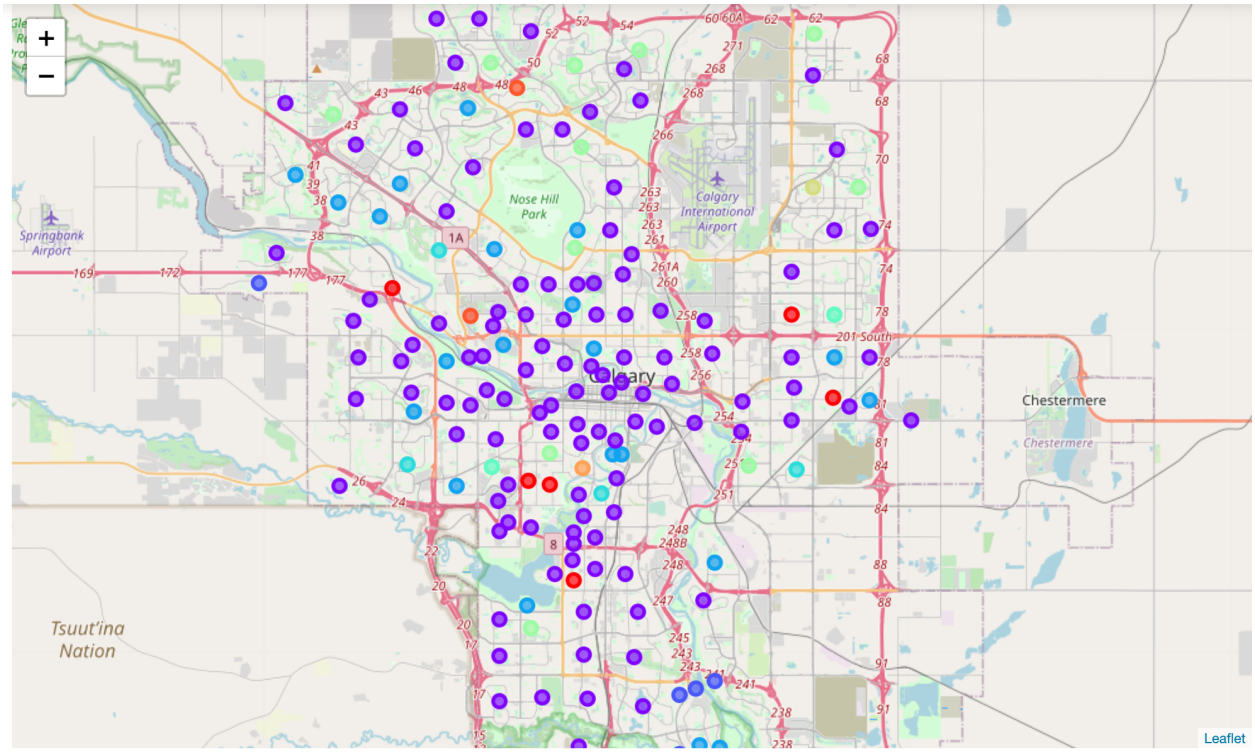

Figure 15 – Clusters distributed in Calgary Map

Results

The analysis above can be concluded as below:

1. Center of the city has the highest neighborhood density
2. Most restaurants located in the center of Calgary which has a smaller scale compared to the neighborhood density. The scope of the locations can be narrowed down.
3. Clustering: Based on the common venues category for each cluster, cluster 3 is the most ideal locations to open a new bubble tea store. Because this cluster contains the least café, but most parks, shopping centers, which potentially will bring a lot of visits.

Discussion

Based on the distribution of restaurant and the neighborhood density concluded by this project, center of the city is most recommended to open the new bubble tea store. However, there are some important factors also should be considered but not covered in this project. For example, existing bubble tea stores allocation. Locating in a competitive location could possibly affect the numbers of customer visit.

Conclusion

In this project, we accessed to Community location from the city of Calgary and venues locations in the city of Calgary from Foursquare API. And Python package – Folium has been used to plot the neighborhoods and venues to visualize the distributions in the Calgary Map. K-Means clustering algorithm was used to cluster the venues by categories to see the potential popularity of the store in the future. Based on the analysis made in this project, it is recommended to have the new store opened in the center of the city of Calgary, where has the highest population density and most restaurant distribution. This project is created to make recommendations on finding the optimal location for a new bubble tea stores. However, more factors should be considered to finally determine the location.

References:

[i] *Shankar Bhandalkar, Allied Market Research,* https://www.alliedmarketresearch.com/press-release/bubble-tea-market.html