Data Science Capstone Final Report

Date: February 2021

1. Introduction

Background: In this assignment, I will develop a data science tool to support entrepreneurs who want to start a new horeca business (Hotel, Restaurant Catering) with information about:

- Which neighbourhood in Toronto is the most suitable to put their business location i.e. shops/restaurants.
- Compare their business to other businesses that can potentially create a good niche for their business to grow.

The tool to support entrepreneurs is developed based on a few theories that have been established and verified in market analysis and business development. These include:

- Businesses, especially those in the Horeca sector, tend to thrive together. It is therefore relevant to invest in a new business in neighbourhoods that are already known, established with other similar businesses.
- In a thriving neighbourhood, businesses tend to thrive together due to their complementary natures. For instance, restaurants are often found in close proximity to bars, coffeeshops, and shopping streets. It is therefore relevant to have a good idea of what kind of other businesses are present at the targeting neighbourhood, which can help to design the business idea and fit in the overall business landscape.

2. Data preparation and data cleansing

For this assignment, I will use location data from Foursquare to develop insights to support entrepreneurs to choose a neighbourhood for their new horeca business. The data will focus on the city of Toronto, Canada. The following categories of data will be used:

- Neighbourhood name
- Venue name
- Location (Lat and Lon data)
- Business types

Standard data cleansing methods will be applied, including removing NAs, removing ‘not specified’ entities, merge similar entries.

3. Methods

K-means clustering will be used to group venues to clusters, and give insights on what kind of venue/business is most common in a neighbourhood.

4. Results

The neighbourhoods of Toronto are scrapped from Wikipedia and cleansed and save in a dataframe:

	Postal Code	Borough	Neighbourhood
0	M1B	Scarborough	Malvern, Rouge
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae
5	M1J	Scarborough	Scarborough Village
6	M1K	Scarborough	Kennedy Park, Ionview, East Birchmount Park
7	M1L	Scarborough	Golden Mile, Clairlea, Oakridge
8	M1M	Scarborough	Cliffside, Cliffcrest, Scarborough Village West
9	M1N	Scarborough	Birch Cliff, Cliffside West
10	M1P	Scarborough	Dorset Park, Wexford Heights, Scarborough Town...
11	M1R	Scarborough	Wexford, Maryvale
12	M1S	Scarborough	Agincourt
13	M1T	Scarborough	Clarks Corners, Tam O'Shanter, Sullivan
14	M1V	Scarborough	Milliken, Agincourt North, Steeles East, L'Amo...
15	M1W	Scarborough	Steeles West, L'Amoreaux West
16	M1X	Scarborough	Upper Rouge
17	M2H	North York	Hillcrest Village
18	M2J	North York	Fairview, Henry Farm, Oriole
19	M2K	North York	Bayview Village
20	M2L	North York	York Mills, Silver Hills

Location data (Lat and Lon) are retrieved online, and added to the dataframe:

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

K-means clustering are preformed, and the most common venues are listed for each neighbourhood. This information will help entrepreneur to decide which neighbourhood are suitable to start their business.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Agincourt	Lounge	Latin American Restaurant	Clothing Store	Breakfast Spot	Skating Rink
1	Alderwood, Long Branch	Pizza Place	Gym	Coffee Shop	Skating Rink	Sandwich Place
2	Bathurst Manor, Wilson Heights, Downsview North	Bank	Coffee Shop	Diner	Supermarket	Sushi Restaurant
3	Bayview Village	Café	Japanese Restaurant	Bank	Chinese Restaurant	Women's Store
4	Bedford Park, Lawrence Manor East	Italian Restaurant	Coffee Shop	Sandwich Place	Liquor Store	Thai Restaurant
5	Berczy Park	Coffee Shop	Cocktail Bar	Cheese Shop	Restaurant	Beer Bar
6	Birch Cliff, Cliffside West	College Stadium	General Entertainment	Café	Skating Rink	Dim Sum Restaurant
7	Brockton, Parkdale Village, Exhibition Place	Café	Nightclub	Breakfast Spot	Coffee Shop	Performing Arts Venue
8	Business reply mail Processing Centre, South C...	Light Rail Station	Yoga Studio	Spa	Burrito Place	Skate Park
9	CN Tower, King and Spadina, Railway Lands, Har...	Airport Lounge	Airport Service	Airport Terminal	Boat or Ferry	Boutique
10	Caledonia-Fairbanks	Park	Women's Store	Pool	Doner Restaurant	Dim Sum Restaurant
11	Canada Post Gateway Processing Centre	Coffee Shop	Hotel	Gym	Gas Station	Food Truck
12	Cedarbrae	Hakka Restaurant	Bakery	Lounge	Athletics & Sports	Bank
13	Central Bay Street	Coffee Shop	Café	Sandwich Place	Italian Restaurant	Bubble Tea Shop

5. Discussion and conclusion

The developed tool based on data scrapping and K-means clustering could help to support decision making of entrepreneurs who want to start a business in a new neighbourhood by pointing out what kind of business is already available in that neighbourhood, and show whether they can find synergies with existing businesses to create a niche.

Further work is suggested to develop a automatic recommending system, where entrepreneurs can provide in a table their business information using keywords. This table is then used to compare against the table developed for the neighbourhood, and result in a list of top 5 most suitable neighbourhoods.