The Battle of the Neighborhoods (Toronto)

1. Introduction

1.1 Problem

The business problem that will be attempted to solve through this project will be: “Where is the best possible location to set up a tourism-related business, F&B outlet and an office in Toronto”. Starting a new business is a complicated process starting from brainstorming ideas, gathering capital, registering with the local government to setting up of establishment, any hiccups at the various stages could result in the failure of the business. Location is an important consideration when setting up and establishment. Different businesses have different location requirements based on its targeted audience. For example, an Asian restaurant would get better business when opened in an Asian community whereas a Michelin-star restaurant would do better when opened in the city-center as the spending abilities of the community is higher and thus more likely to visit. As for offices, accessibility is important to ensure the attractiveness of the company amongst employees.

1.2 Background

According to the 2019 Ease of Doing Business ranking by The World Bank, Canada was ranked 3rd in ease of starting business and 23rd overall, which meant that Canada is an attractive location for new entrepreneur to kick start their business. Home to more than 6 million people, Toronto among all the cities in Canada, is one of the greatest cities in the world to do business as it consistently ranked among the top for global competitiveness, innovation and quality of life. Its people is also highly skilled and multilingual as 64% of Toronto residents between 25 and 64 have finished their post-secondary education. Being within a 90-minute flight away from the USA also makes Toronto an accessible and attractive choice for many Multi-National Companies.

2. Data

2.1 Data Sources

The 3 main data sources are:

- Wikipedia Toronto Postal Code List (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- Postal Code Corresponding Coordinates (http://cocl.us/Geospatial_data)
- Foursquare API

2.2 Data Extraction

Toronto Wikipedia page has a list of information of the neighborhoods and the corresponding boroughs and postal code. By making use of the BeautifulSoup library, the list can be extracted via the table HTML element and stored into a data frame. (Named as toronto_df) Due to the large number of unassigned boroughs and neighborhoods, cleaning will be required.

	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
...	...	...	...
175	M5Z	Not assigned	Not assigned
176	M6Z	Not assigned	Not assigned
177	M7Z	Not assigned	Not assigned
178	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...
179	M9Z	Not assigned	Not assigned

Due to the large number of neighborhoods and them being mostly grouped by the postal code, the corresponding coordinates for each postal code will be required for the searching of venues. Coursera Applied Data Science Capstone Week 3 provided a .csv file for its final assignment which contains a list of longitudes and latitudes for each Postal Code. The read_csv is a function in the Pandas library which allows the values in the .csv file to be directly loaded into a data frame. The data frame can then be merged with the Toronto_df using the postal code as the joining key.

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Foursquare API is a gateway provided by Foursquare, a location technology platform, for developers to extract venue-related information. The API requires a client ID, client secret, version to access most of its search features and is only obtainable via opening a Foursquare developer account. As business proportion can provide an insight into the profile of the area, Foursquare API explore feature will be used to obtain a list of venues within a pre-defined radius of a given coordinate. For this analysis, the radius has been set to 2000m and the limit of venues set to 300.

	PostCode	PostCode Latitude	PostCode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M1B	43.806686	-79.194353	African Rainforest Pavilion	43.817725	-79.183433	Zoo Exhibit
1	M1B	43.806686	-79.194353	Images Salon & Spa	43.802283	-79.198565	Spa
2	M1B	43.806686	-79.194353	Toronto Pan Am Sports Centre	43.790623	-79.193869	Athletics & Sports
3	M1B	43.806686	-79.194353	Toronto Zoo	43.820582	-79.181551	Zoo
4	M1B	43.806686	-79.194353	Gorilla Exhibit	43.819080	-79.184235	Zoo Exhibit

2.3 Data Cleaning

Cleaning of data is an important process as it prepares the data for analysis. For the data frame with the postal code, borough and neighborhoods, all rows with Borough and Neighborhood as 'Not assigned' were removed. To ensure that there are no duplicated postal codes, all rows are grouped by the Postal Code and Borough and neighborhood names were joined by ','.

	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
...	...	...	...
98	M9N	York	Weston
99	M9P	Etobicoke	Westmount
100	M9R	Etobicoke	Kingsview Village, St. Phillips, Martin Grove ...
101	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...
102	M9W	Etobicoke	Northwest, West Humber - Clairville

As clustering requires numerical data, the venues data collected via the Foursquare API would need to be cleaned. Venue Category is a categorical variable which describes the category of a venue nearby. By using Pandas get_dummies, venue categories can be converted to variables so clustering can be done.

	PostCode	Accessories Store	Afghan Restaurant	African Restaurant	Airport	American Restaurant	Amphitheater	Antique Shop	Aquarium	Argentinian Restaurant	...	Volleyball Court	Warehouse Store	Whisky Bar
0	M1B	0.0	0.0	0.019608	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
1	M1C	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
2	M1E	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
3	M1G	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
4	M1H	0.0	0.0	0.000000	0.0	0.020000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
98	M9N	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
99	M9P	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
100	M9R	0.0	0.0	0.000000	0.0	0.014706	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
101	M9V	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
102	M9W	0.0	0.0	0.000000	0.0	0.078947	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0

By making use of the venues collected from the Foursquare API, clustering of the various postal codes using the venue categories will allow a better understanding of profile of the various postal codes which can enable identification of possible locations most suitable to the 3 types businesses: tourism-related business, F&B outlet and office.