Exploring neighborhoods in Toronto to open an Indian Restaurant

Introduction

Toronto, Canada is one of the most diverse cities in the world [1] and consists of many immigrants from all over the world. According to the 2016 census report [1] people of Asian origins account for about 17.7% of Canadian population. From this overall Asian community, South Asians are considered a visible minority in Canada as per the classification of Statistics Canada [2] and are growing at a substantial pace due to the new immigration programs such as Express Entry.

Because of this diverse multicultural nature of the city, the food culture in also as diverse. As the South Asian population grow so does the demand for South Asian food. Indian cuisine is one of the largest and perhaps the most popular south Asian cuisine and is admired by taste buds all around the world. In this study, I try to analyze the available public data using data science techniques to determine the most suitable neighborhood in Toronto city to open an Indian restaurant to serve the rapidly growing Indian community.

Business Problem

If you are a businessman or an investor, exploring business opportunities in Canada, restaurant busines is undoubtably one of the top candidates. Choosing the correct location for a restaurant is one key element the investment decision. However, it is not obvious as to what factors should be considered prior selecting a location. Extracted from "How to start a restaurant" article published at *entrepreneur.com* [3], following are some of the components deemed as key factors in selecting restaurant location:

- Demographics
- Traffic density
- Crime rates
- Competitors
- Property affordability

However, one would not be able to sit down and research on all of these aspects manually and come to a decision as to which location is most suitable. Therefore, It would be very convenient and reliable if there was a data driven method to consider these factors and recommend the best location or neighborhood to open the restaurant. Therefore, in this study the business problem I try to address is, **"What neighborhood(s) is/are most suitable to open an Indian cuisine restaurant in the city of Toronto, in Canada?"**

Within the limited timeframe it is not practical to conduct a full-fledged study considering all of the above factors but only few where data is accessible. Therefore, I will choose only a few aspects from the above list and try to answer the problem using a data science approach.

Target Audience and stakeholders

This project is intended for two main groups of individuals or organizations, namely,

- **Investors:** who are exploring opportunities to step into the restaurant business
- **Restaurant owners**: who are planning to expand or diversify their business into different cuisines and neighborhoods.

Apart from them, anyone who is interested in understanding the food scene in Toronto could also benefit from this study.

Data

The data section discusses about the datasets and the data sources considered for this project. All data in this project are publicly available data sourced through world wide web and Canadian government data portals and reports.

Data sources

Toronto Neighborhoods Data

Toronto city officially has 140 neighborhoods covering multiple districts of old Toronto, East York, Etobicoke, North York, Scarborough, and York. There are multiple ways of dividing the neighborhoods for different purposes but for the scope of this project I will be using the City Designated Neighborhoods (CDN) as the reference for each neighborhood.

To map the neighborhoods in an interactive map, perhaps the most suitable and reliable way would be to find an official **geojson** file from the city's official web site. The Toronto City Open Data Portal has such a geojson format file with all the 140 neighborhoods, their names, CDN, boundaries, centroids and more. This file can be found in the Toronto City Open Data Portal at the below link.

https://open.toronto.ca/dataset/neighbourhoods/

A sample data frame created from this file using the "**geopandas"** package is shown below. The required data columns are area code, longitude, latitude and geometry.

```
Toronto_Geojson = gpd.read_file(Toronto_geojson_url)
Toronto_Geojson.head(5)
```

	_id	AREA_ID	AREA_ATTR_ID	PARENT_AREA_ID	AREA_SHORT_CODE	AREA_LONG_CODE	AREA_NAME	AREA_DESC	X	Y	LONGITUDE	LATITUDE	OBJECTID	Shape__Area	Shape__Length	geometry
0	6581	25886861	25926662	49885	94	94	Wychwood (94)	Wychwood (94)	None	None	-79.425515	43.676919	16491505	3.217960e+06	7515.779658	POLYGON ((-79.43592 43.68015, -79.43492 43.680...
1	6582	25886820	25926663	49885	100	100	Yonge-Eglinton (100)	Yonge-Eglinton (100)	None	None	-79.403590	43.704689	16491521	3.160334e+06	7872.021074	POLYGON ((-79.41096 43.70408, -79.40962 43.704...
2	6583	25886834	25926664	49885	97	97	Yonge-St.Clair (97)	Yonge-St.Clair (97)	None	None	-79.397871	43.687859	16491537	2.222464e+06	8130.411276	POLYGON ((-79.39119 43.68108, -79.39141 43.680...
3	6584	25886593	25926665	49885	27	27	York University Heights (27)	York University Heights (27)	None	None	-79.488883	43.765736	16491553	2.541821e+07	25632.335242	POLYGON ((-79.50529 43.75987, -79.50488 43.759...
4	6585	25886688	25926666	49885	31	31	Yorkdale-Glen Park (31)	Yorkdale-Glen Park (31)	None	None	-79.457108	43.714672	16491569	1.156669e+07	13953.408098	POLYGON ((-79.43969 43.70561, -79.44011 43.705...

In addition to this, I was also determined to manually explore each neighborhoods' geo coordinates (latitude, longitude) as there could be a difference in the geometric centroid (of the shape) of a neighborhood versus the actual neighborhood center in terms of activity.

Therefore, as a secondary dataset I used the following Wikipedia page to find the 140 neighborhoods division.(https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Toronto#Table)

I used web scraping methods embedded in pandas (using *Dataframe.read_html*) to scrape a table in the above web page and then used "Nominatim" geocoder against the neighborhood names to find the coordinates.

The longitude and latitude value returned by Nominatim for each neighborhood was them inserted to a Pandas data frame. The coordinates of each neighborhood would later be used to search for venues in the surrounding area.

Following is a sample geocoded neighborhood dataset created using Nominatim.

Out[14]:

	CDN	Neighbourhood_Name	Borough	Areas_Covered	Neighbourhood	Latitude	Longitude
0	129	Agincourt North	Scarborough	Agincourt and Brimwood	Agincourt North	43.808038	-79.266439
1	128	Agincourt South-Malvern West	Scarborough	Agincourt and Malvern	Agincourt South	43.785353	-79.278549
2	20	Alderwood	Etobicoke	Alderwood	Alderwood	43.601717	-79.545232
3	95	Annex	Old City of Toronto	The Annex and Seaton Village	Annex	43.670338	-79.407117
4	42	Banbury-Don Mills	North York	Don Mills	Banbury	43.742796	-79.369957
5	34	Bathurst Manor	North York	Bathurst Manor	Bathurst Manor	43.665519	-79.411937
6	76	Bay Street Corridor	Old City of Toronto	Bay Street, Financial District	Bay Street Corridor	43.667342	-79.388457
7	52	Bayview Village	North York	Bayview Village	Bayview Village	43.769197	-79.376662
8	49	Bayview Woods-Steeles	North York	Bayview Woods	Bayview Woods	43.798127	-79.382973
9	39	Bedford Park-Nortown	North York	Bedford Park, Ledbury Park, and Nortown	Bedford Park	43.737388	-79.410925

Neighborhood Demographic data

It is important to understand the demographics of each neighborhood in this study and hence I used the Statistics Canada 2016 census reports to obtain the population data end ethnic origin data for each

neighborhood. The Toronto City Open Data Portal has a huge number of useful datasets which is publicly available. “Neighborhood Profile 2016” is such dataset with rich insights into the neighborhood demographics.

I downloaded the “Neighborhood Profile 2016” CSV file examined it to extract the relevant data points.

Below are two inks to the original data set “Neighborhood profile 2016”

https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/

https://open.toronto.ca/dataset/neighbourhood-profiles/

Population and their ethnic origin

Indian food is admired by many nationals and people of different ethnic origin around the world. But it is not a secret that Indians themselves and in more general **south Asians** are the ethnic group that would miss or crave for this type of food the most. Therefore, I decided to focus my study considering the population with south Asian ethnic origin.

This population data can help to later map the restaurant locations with the target audience (Majority south Asian neighborhoods) determining if a neighborhood is underserved.

In the “Neighborhood profile 2016.csv”file, I was able to find the total population with a south Asian origin as well as individual ethnic origin country wise population numbers, by looking at the column “category” = “Ethnic origin” and the column “topic”= “Ethnic origin population”.

After further inspecting the above selected data, I found that the below values in the “characteristic” column would be most suitable for my study.

1) South Asian origins
2) South Asian origins; n.i.e. (not indicated elsewhere)

Therefore, I extracted these two data points along with its country wise subdivisions and saved it as “TorontoSouthAsianPopulation.csv”. A sample dataset from the extracted CSV file looks like this.

```
# import the data from neighborhood profiles

SouthAsianPoplution = pd.read_csv('TorontoSouthAsianPopulation.csv')
SouthAsianPoplution
```

	Characteristic	City of Toronto	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	Banbury-Don Mills	Bathurst Manor	Bay Street Corridor	Bayview Village	Bayview Woods-Steeles	Bedford Park-Nortown	Bee
0	South Asian origins	350040	5260	4170	575	1320	2940	565	2515	1410	805	535	
1	Bangladeshi	18155	10	80	0	35	50	0	85	35	0	0	
2	Bengali	9455	30	60	0	45	10	80	85	20	0	0	
3	Bhutanese	125	0	0	0	0	0	0	0	0	0	0	
4	East Indian	202675	2090	2180	495	965	2365	360	1865	1025	570	430	
5	Goan	1280	10	30	20	10	30	0	0	0	20	15	
6	Gujarati	1990	20	10	0	0	40	0	20	0	0	0	
7	Kashmiri	345	0	10	0	25	0	0	0	0	0	0	
8	Nepali	3080	0	0	0	10	0	0	20	0	0	0	
9	Pakistani	36135	230	410	50	125	325	30	340	140	65	25	
10	Punjabi	6205	55	20	20	15	10	10	60	0	0	40	
11	Sinhalese	1735	0	65	0	10	20	0	10	15	15	15	
12	Sri Lankan	58185	2230	1145	30	85	135	40	125	155	75	30	

Location based venues data from Foursquare API

My intention is to explore restaurants belonging to different cuisines in each of the neighborhood in order to understand the distribution of restaurants and its possible relationship to the ethnic distribution, as well as to determine underserved neighborhoods.

As the source of location-based venue data I chose Foursquare API. Foursquare is one of the largest data sources serving location-based venue data and is free to use (with certain limitations) with a developer account.

Reading the Foursquare API documentation revealed that there is a “category ID” for each venue data point and that the categories are also organized in a “Category Hierarchy”. This means that if we search for a venue, providing the parent venue category ID as a parameter, we will also receive the venue details of any child category ID under that parent ID.

By using the parent category ID – “Food - 4d4b7105d754a06374d81259” I was able to retrieve any type of venue that is categorized under this parent category. The results were limited using LIMIT and RADIUS parameters in the API, to a 1 km radius and 200 venues. A sample data set is given below.

```
Toronto_food_venues.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wychwood	43.676919	-79.425515	Annabelle Pasta Bar	43.675445	-79.423341	Italian Restaurant
1	Wychwood	43.676919	-79.425515	Loblaws	43.684188	-79.415485	Grocery Store
2	Wychwood	43.676919	-79.425515	Booyah	43.681744	-79.418011	Ice Cream Shop
3	Wychwood	43.676919	-79.425515	Loblaws	43.671657	-79.421364	Grocery Store
4	Wychwood	43.676919	-79.425515	Knockout Ice Cream	43.670717	-79.434551	Ice Cream Shop

References

[1] https://www.worldatlas.com/articles/the-most-diverse-cities-in-the-world.html
[2] https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=PR&Code1=01&Geo2=PR&Code2=01&Data=Count&SearchText=canada&SearchType=Begins&SearchPR=01&B1=Ethnic%20origin&TABID=1
[3] https://www.entrepreneur.com/article/73384

https://web.archive.org/web/20160718232555/http://www.statcan.gc.ca/eng/concepts/definitions/minority01a
Neighbourhoods Wikipedia page:
https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Toronto#Table
Postal codes:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Toronto Neighbourhoods GeoJason Data
https://open.toronto.ca/dataset/neighbourhoods/

Neighbourhold profile 2016
https://open.toronto.ca/dataset/neighbourhood-profiles/

How to find the optimal ‘k’ for k-means clustering

https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb