Open a new restaurant in Toronto

Wen Gao

May 16, 2021

1. Introduction

1.1 Background

There are many good restaurants in Toronto, and it is not easy to choose a great location, if a businessman wants to open a new restaurant in Toronto. This new location needs to be in a popular neighborhood, but there should not be many restaurants already, so the market of restaurants is not saturated yet. In this case, the businessman can have a chance to predict a great start and less competition with others. Therefore, the businessman has more chance to earn money from this new restaurant.

1.2 Problem

This project aims to solve the problem that if there is a businessman wants to open a new restaurant in Toronto, but the businessman doesn't know where he should choose as the new restaurant's location. This project will help the businessman to find a better location.

1.3 Interest

The businessman in Tonronto will have high interest in this project, as the businessman will be suggested a better location to open the new restaurant, and has a better chance to earn more money from it.

2. Data acquisition and cleaning

2.1 Data sources

There are 3 data sources in this project:

1. To get the whole overview of neighborhoods, postal code, borough info: We will use wikipage info in this url:

1) https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2. Then we need to match the neighborhoods with accurate latitude and longitude, so we can try to get the top venues within the neighborhoods from Foursqaure API. This part of data will be got from coursera IBM data science course resource. The url for getting the data is:

1) https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv

3. Finally we need to get the top 10 venues in each neighborhoods in Toronto, using Foursqaure API. This is the API that we will use:

1) https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}

2.2 Data cleaning

In total, we will combine different data into 1 dataframe, and then apply cluster analysis on it. Below are the details of how to do the data cleaning for each data source.

1. Firstly we will get the neighborhoods data of Toronto from wikipage, which will be in HTML format. We will transfer the HTML data into pandas dataframe by using python soup function, and we will also do some cleaning of the dataframe as below.

1) The dataframe will consist of three columns: PostalCode, Borough, and Neighborhood and will look like below:

a)

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park	Ontario Provincial Government

2) Since there are some values in Borough is “Not assigned”, we will ignore these rows as they are redundant and useless data.

3) There are more than one neighborhood can exist in one postal code area. For example, M5A is listed twice and has two neighborhoods: Harbourfront and

Regent Park. In the recommended approach from IBM data science course, we should combine these 2 rows into one row with the neighborhoods separated with a comma as “Harbourfront, Regent Park”.

4) There is also “Not assigned” value in neighborhood column, in this case, the value will be the same as borough.

2. For the CSV file of latitude and longitude of neighborhoods in Toronto, it is already in clean structure, and the only thing we need to do it to combine the data with the 1st dataframe, and get a new dataframe with 5 columns: PostalCode, Borough, Neighborhood, latitude and longitude as below.

1)

	PostalCode	Borough	Neighborhood	Postal Code	Latitude	Longitude
0	M3A	North York	Parkwoods	M3A	43.753259	-79.329656
1	M4A	North York	Victoria Village	M4A	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	M5A	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	M6A	43.718518	-79.464763
4	M7A	Queen's Park	Ontario Provincial Government	M7A	43.662301	-79.389494

3. For the response from Foursquare data, it is in json format. In this case, we need to transfer the json data into pandas dataframe. The final dataframe will have 7 columns: Neighborhood, Neighborhood, LatitudeNeighborhood, LongitudeVenueVenue, LatitudeVenue, LongitudeVenue, Category as below:

1)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant

3. Methodology

3.1 Exploratory Data Analysis

After neighborhoods data of Toronto has been cleaned, we got the dataframe including info of neighborhoods, postal code, borough, latitude, longitude. Then we can use this latitude and longitude info to request API in Foursquare, to get the top 10 venues in each neighborhood.

After we got the 10 venues in each neighborhood, we need to use cluster method to group the neighborhoods into 5 groups. Within the groups, we will calculate the counts of venues for each neighborhood, and the counts of restaurants for each neighborhood. Then we use the value of “counts of venues for each neighborhood” to divide “counts of restaurants for each neighborhood”, in this case, we get the percentage of the restaurants among the venues for each neighborhood.

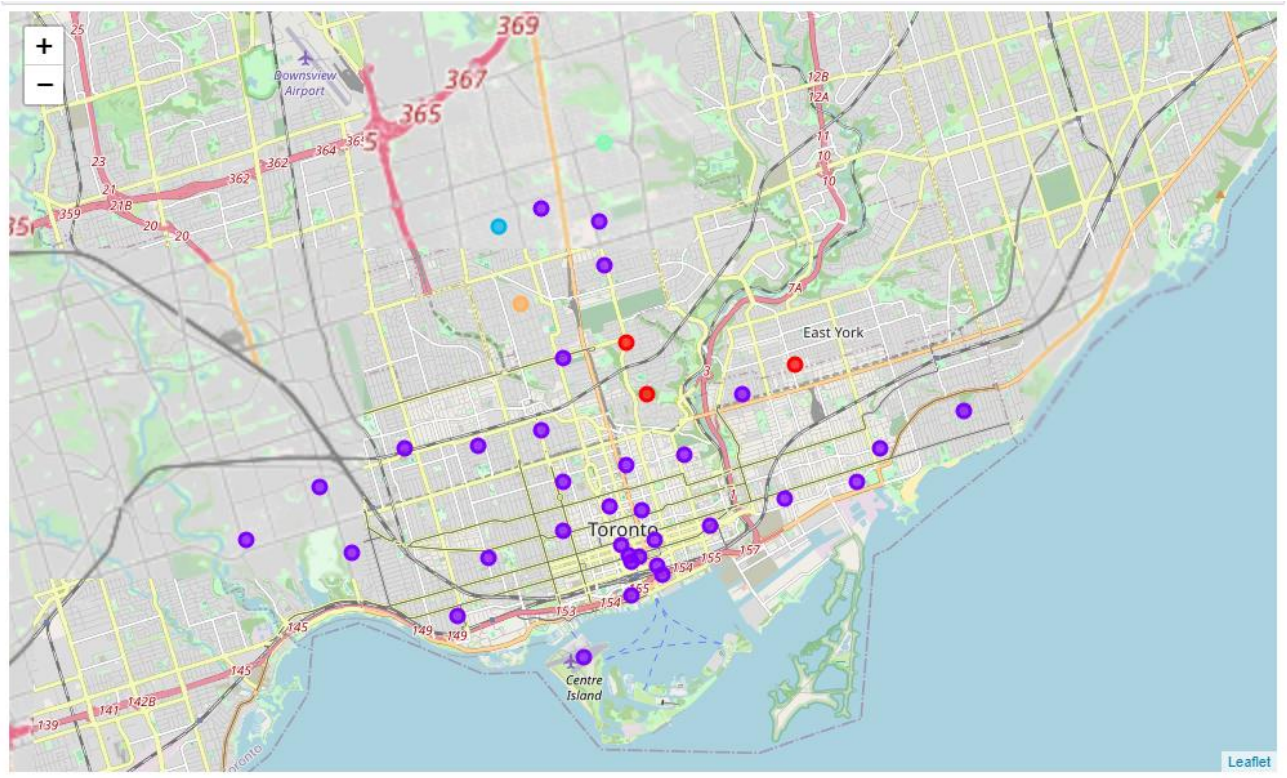
Since the neighborhoods in the same clustered group should have similar attributes, we can assume that within the same group, the less percentage a neighborhood has, the better that this location is great for opening a new restaurant.

During the analysis, I found out that cluster 2, 3, 4 only has 1 neighborhood, so there are not enough data to analyze them. Therefore, I have dropped them from the dataframe, and chose to only analyze cluster 0(has 2 neighborhoods) and 1(has 33 neighborhoods- should be more accratue prediction then cluster 0 as

there are more data to be analyzed).

3.2 Modeling

In this project, I used cluster modeling as the method, because we don't have clearly defined output for how to group the neighborhoods. In this case, using clustering modeling is a good choice, as we only need to define how many groups we want, and the machine learning algorithm will help us to automatically group our neighborhoods in Toronto in an intelligent way. Below is the clustering map overview:



4. Results

After the whole analysis, we have below results:

1. In cluster 0, both neighborhoods Rosedale and The Danforth East are good choices to open a new restaurant, as there is no restaurant in top 10 venues yet. The overview is as below:

	Neighborhood	total_venues_count_per_neighboorhood	is_restaurant	percentage
0	Rosedale	4	0	0.0
1	The Danforth East	4	0	0.0

1)

2. In cluster 1, the top 5 neighborhoods for the choices to open a new restaurant are:

1) "CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport",

2) "The Beaches",

3) "Davisville North"

4) "Dufferin, Dovercourt Village"

5) "Brockton, Parkdale Village, Exhibition Place"

And here is the overview:

	Neighborhood	total_venues_count_per_neighboorhood	is_restaurant	percentage
2	CN Tower, King and Spadina, Railway Lands, Har...	17	0	0.000000
29	The Beaches	4	0	0.000000
8	Davisville North	8	0	0.000000
9	Dufferin, Dovercourt Village	15	1	0.066667
1	Brockton, Parkdale Village, Exhibition Place	22	2	0.090909

5. Discussion

5.1 Data concerning

1. In this project, as we only get the neighborhoods data from wikipedia, and some of the values of neighborhoods and borough are “not assigned”, so the source data of neighborhoods in Toronto is not very accurate.
2. As we only choose to get top 10 venues in each neighborhood from Foursquare API, it is not big enough to provide very accurate prediction actually. In addition, after the clustering method applied, there are 3 clusters only have 1 neighborhood, so we need to abandon them. This will also have an influence on our prediction.

5.2 Methodology concerning

In the real life, there are multiple factors, that will influence where the best location is for a new restaurant opening. For example, we need to distinguish which type of restaurant that a businessman wants to open. If it is a Chinese restaurant, then we need to investigate how much Chinese people there are in each neighborhood. Furthermore, we also need to comprehensively consider other factors, e.g. what the income levels are in each neighborhood.

This project still has many dimensions that can be improved.

6. Conclusions

In all, from this project, we can have a primary prediction of which neighborhoods in Toronto are more suitable to open a new restaurant. The best neighborhoods to open a new restaurant in Toronto that I suggested are:

1) "CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport",

2) "The Beaches",

3) "Davisville North"

4) "Dufferin, Dovercourt Village"

5) "Brockton, Parkdale Village, Exhibition Place"

6) "Rosedale"

7) "The Danforth East"

However, this is only a reference for a businessman's decision. We still need bigger data to improve this analysis in many dimensions, to make a better choice.