Optimum location for an Indian restaurant in Toronto

Dinesh Goyal
Sep 19, 2019

1. Introduction

Background

After learning about data science and how it can be used to solve real life problems, I've started my own company, DataMiners Inc. DataMiners has subscribed to various data sources all over the world. Using this data along with other packages like FourSquare which provides location data, our company can help clients solve various business problems. One of the common use of our expertise is to help potential business owners who want to know the optimum location to start a new business in a given geographical area like a city or a neighborhood.

Problem Description

Our latest client is a father son team who would like to open an Indian restaurant in the Toronto area. They would like to serve mainly north indian cuisine, be open for lunch and dinner, and cater to mdeium to upper medium class families. They have come to us to help them with exploring Toronto Neighborhoods, understand the existing restaurant scene in the city, and some suggestions for best locations for their new restaurant.

Target Audience

The father and son team who wants to learn about Toronto neighborhoods and find a suitable location to open an Indian restaurant

2.0 Data acquisition and cleaning

Data Sources

To solve this problem, we'll need 3 sets of data as described below:

1. List of neighborhoods in Toronto, Canada. The list of neighborhoods is given in a Wikipedia page which we can make use of (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M_) The data that we are most interested in is in the "Neighborhood' column. If a 'Neighborhood' name is not given, then we'll use the coressponding 'Borough' name. Example data:

Postal Code	**Borough**	**Neighborhood**
C01	Downtown Toronto	Downtown
C01	Downtown Toronto	Harbourfront
C01	Downtown Toronto	Little Italy

2. Latitude and Longitude of these neighborhoods. We'll try to use publicly available Google APIs for this. This data would be needed to get the existing restaurants details in each neighborhood. This data can also be obtained using a CSV file as Google APIs don't work some times.

3. Various data for the venues in these neighborhoods. From this data, we'll try to figure out the number and location of existing restaurants, specially Indian restaurants. We'll use FourSquare APIs to get this information. The data most useful would be 'Venue Categroy' as it describes the kind of restaurant that already exists in a neibhborhood. We'll skip neighborhoods that have "Indian Restaurants' already, and focus on neighborhoods with restaurants from Asian countires since most asians tend to like spicy food. Example data after some formatting:

Neighborhood	**Venue**	**Longitude**	**Lattitude**	**Venue Category**
Downtown	Sakoon	16.78456	14.5678	Indian Restaurant
Downtown	Ho Choi	12.12453	24.4092	Chinese Restaurant
Little Italy	Taste of Italy	13.5556	34.4562	Italian Restaurant

Data Cleaning

As the data was scrapped from a Wiki page, some values had a '\n' character at the end. So we used the rstrip() method to get rid of any new line or eol characters.

The neighborhood data from Wikipedia has main issues: either the Neighborhood name or Borough name was missing. We decided to totally skip the rows where Borough was missing. For rows, where Borough was given but no value was present for the Neighborhood, we assigned the Borough's value to the Neighborhood name.

Raw Data:

Postal Code	**Borough**	**Neighborhood**
C01	Downtown Toronto	Downtown
C01	Not Assigned	Harbourfront
C01	Downtown Toronto	Not Assigned

After Cleaning:

Postal Code	**Borough**	**Neighborhood**
C01	Downtown Toronto	Downtown
C01	Downtown Toronto	Downtown Toronto

Feature Selection

After data cleaning, the next step was to decide which features of data to use for our analysis. For the Wiki data about neighborhoods, we didn't make use of the Borough name as Neighborhood names were sufficient for our purpose.

From the FourSquare API location data, we didn't make use of the Venue Latitude or Venue Longitude data since we were only interested in the Neighborhood geographic informaiton.

3.0 Methodology

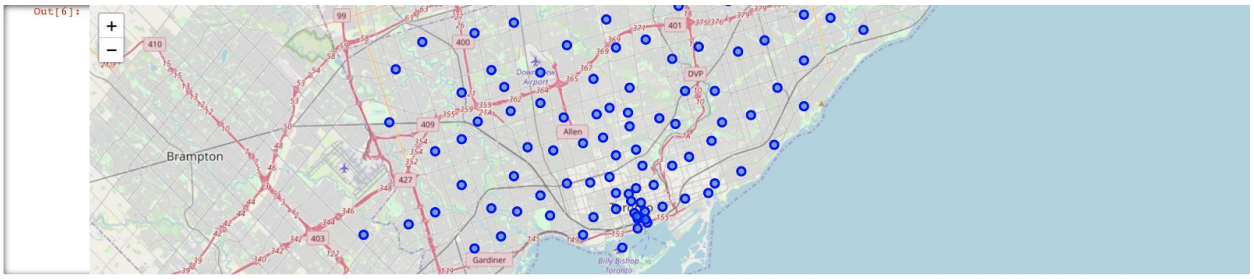
First, we need to get the list of all the neighborhoods in Toronto, Canada. Luckily this data was available on a Wikipedia page. Using standard Panda HTML page scrapping methods, I was able to scrap the web page and bring the entire tabular dtaa into a panda dataframe.

After the necessary cleaning (see the Data Cleaning section), I needed to get longitude and latitude coordinates of each of the neighborhoods so that we could make use of the FourSquare APIs for venue data. To get the coordinates, the Geocoder google APIs were not working consistently, so I decided to use the csv file provided in the course. Then the data from the Wiki page and CSV file was merged to create a dataframe containing a list of Toronto neighborhoods along with their longitude and latitude values.

Out[4]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront,Regent Park	43.654260	-79.360636
3	M6A	North York	Lawrence Heights,Lawrence Manor	43.718518	-79.464763
4	M7A	Queen's Park	Not assigned	43.662301	-79.389494

This is what the the map of Toronto looks like showing all it's neighborhoods:



Next, I used Foursquare API to get a list of venues in each neighborhood. Limited use of this API is availabel for free which is what we made use of for this project. Foursquare API allows us to get information such as venue name, category (bar, restaurant, gym, etc), and latitude and longitude values.

	name	categories	lat	lng
0	The Greater Good Bar	Bar	43.669409	-79.439267
1	Parallel	Middle Eastern Restaurant	43.669516	-79.438728
2	Happy Bakery & Pastries	Bakery	43.667050	-79.441791
3	Planet Fitness Toronto Galleria	Gym / Fitness Center	43.667588	-79.442574
4	Blood Brothers Brewing	Brewery	43.669944	-79.436533

Having venue data such as above, allowed me to get a sense of types of venues in each neighborhood. I could find out the number, and frequency of each category in each neighborhood.

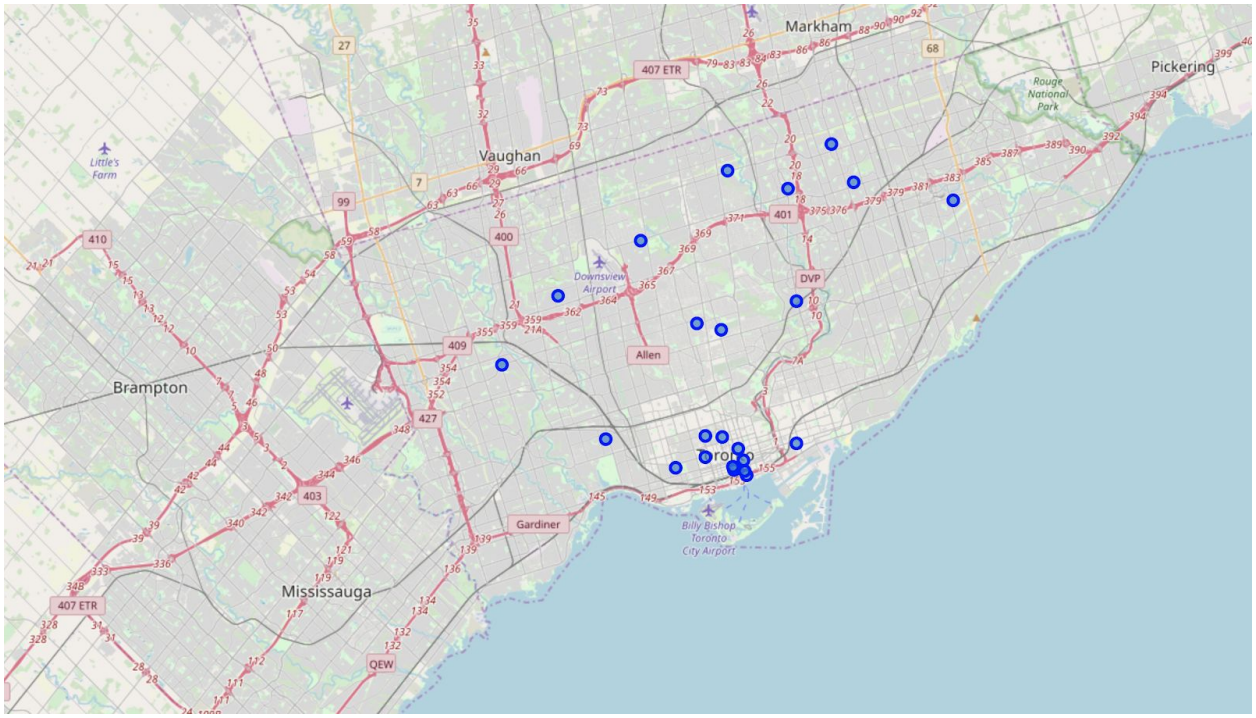
After I obtained data for all the venues in Toronto, I dropped all the neighbordhoods that already had an Indian restaurant. Thinking was that it would un necessarilty create competition and possible hurt the current and the new business owners.

Given our assumption that having many asian (non-indian) restaurnats is a good indication of type of food people like in these neighborhoods, we only selected neighborhoods that have at least one of the following types of restaurants:

- Thai
- Malai
- Afgan
- Chinese
- Asian
- Ethiopian

Note: Even though Ethiopia is not in Asia, it's food is somewhat similar to Indian food. So the thinking was that a neighborhood containing Ethiopian restaurnat is a reasonable good location for an Indian restaurant as well.

I plotted these neighborhoods on the Tornoto map, and got the following:

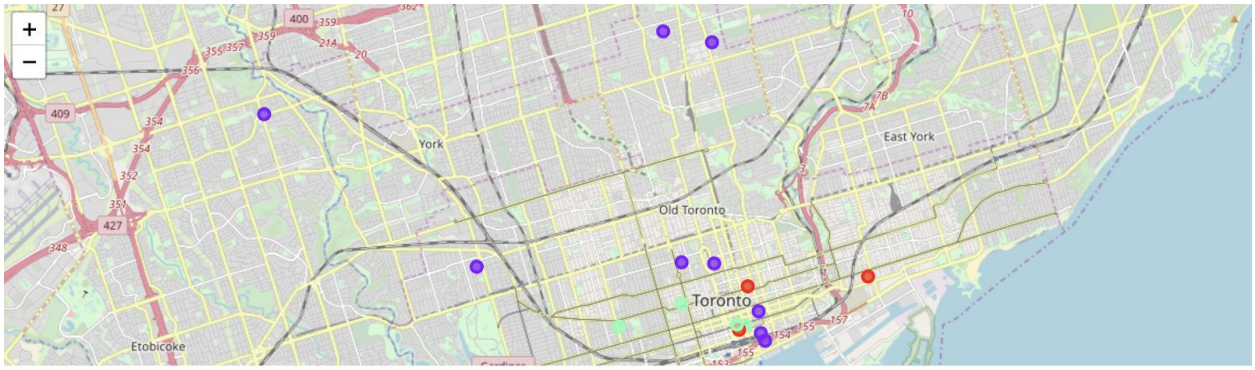


The last step was to show the density of asian restaurants in the city of Tornonto. To acomplish this, I decided to use the k-means clustering mehtod. K-means clustering algorithm identifies k number of centeriods, and then allocates every data point to the nearest cluster. It then tries to move the location of the centroids so that its distance to the points that are assigned to it is as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. For a more thorough analysis, additional data can be used for k-means clustetring like population, income level, etc. For this project, I have clustered the neighborhoods in Toronto into three clusters based on the total number of asian restaurnats in each neighborhood.

Finally, using the folium map APIs, we can show these clusters on a geograpic map of Toronto, and be able to recommend the neighborhoods most suitable for a new Indian restaurant.

4.0 Results

The results from our simple k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Asian restaurants are in each neighborhood as shown by this folio map:



- Cluster Red: Neighborhoods with just 1 Asian restaurant
- Cluster Blue: Neighborhoods with 2 or 3 Asian restaurants
- Cluster Green: Neighborhoods with 4 to 5 (highest number) of Asian restaurants.

5.0 Recommendations

Based on our analyis, the idle locations to open an Indian restaurant are in the Little Portugal, China Town, First Canadian Place and Commerce Court neighborhoods.

Neighborhoods like Westmount, North Toronto, Cedarbrae, and BayView village should be avoided as they don't have very few asian restaurants.

All other neighborhoods not shown on the map didn't have a single asian restaurnant. This could either mean it's an opportunity or there is no interest in asian type food in that area. It would require further analysis to figure out which of these two reasons is correct..

6.0 Suggestions for Future Research

For this study, I considered two main factors for makig the recommendations:

- Similarity of asian (thai, chinese, malay, etc.) food to Indian food.
- Assuming that areas which have many asian restaurants but no Indian restaurants, are a good bet that folks in these neighborhoods would also like Indian food.

With more time and resources, I would have liked to take into account the following as well:

- Populaiton of each neighborhood.

- Ethinic and income breakdown
- Residential vs. business mix - for a business, lunch is more porfitable, while for a residential area, more people might come for dinner.
- Proximity to famous landmarks in the city

7.0 Conclusion

This report illustrates how a typical real world problem can be solved using Data Science tools and techniques. In this study, I looked at various neeighborhoods in the Toronoto metropolitan area. For each neighorhood, I examined the venues to get a sense of the restaurant scene in the area. This was done to help a client on which neighborhood would be best to open an Indian restaurant. By makig use of publically available data, a set of geo location API framework, and machine learning techniques (K-means in this exampke) I was able to provide a recommendation to the client.

8.0 References

- List of neighborhoods in Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Foursquare Developer Documentation: https://developer.foursquare.com/docs
- Jupitor notebook containing all the source code: https://nbviewer.jupyter.org/github/digoyal/Coursera_Capstone/blob/master/Coursera_Capstone_Week5.ipynb