Toronto Neighborhood Classification for Business Expansion

By: David Marcus Thierry
January 7, 2021

Introduction

IBM Data Science Professional Certificate Capstone Project
This project aims to utilize all Data Science Concepts learned in the IBM Data Science Professional Course. We define a Business Problem, the data that will be utilized and using that data, we are able to analyze it using Machine Learning tools. In this project, we will go through all the processes in a step-by-step manner from problem designing, data preparation to final analysis and finally will provide a conclusion that can be leveraged by the business stakeholders to make their decisions.

Objective
Toronto is one of the most densely populated areas in Canada. Being the land of opportunity, it brings in a variety of people from different ethnic backgrounds to the core city of Canada, Toronto. Being the largest city in Canada with an estimated population of over 6 million, there is no doubt about the diversity of the population. Multiculturalism is seen through the various neighborhoods including; Chinatown, Corso Italia, Little India, Kensington Market, Little Italy, Koreatown and many more. Downtown Toronto being the hub of interactions between ethnicities brings many opportunities for entrepreneurs to start or grow their business. It is a place where people can try the best of each culture, either while they work or just passing through. Toronto is well known for its great food.
The objective of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the 'best' neighborhood in Toronto to open a restaurant. Pizza and Pasta are one of the most bought dishes in Toronto originating from Italy. Toronto is the fourth largest home to Italians with a population of over 500k, there are numerous opportunities to open a new Italian restaurant. Through this project, we will find the most suitable location for an entrepreneur to open a new Italian restaurant in Toronto, Canada.

Target Audience
Entrepreneurs who are passionate about opening an Italian restaurant in a metropolitan city such as Toronto would be extremely interested in this project. The project is also for business owners and stakeholders who want to expand their businesses and wonder how data science could be applied to the questions at hand.

Data

The data that will be required will be a combination of CSV files that have been prepared for the purposes of the analysis from multiple sources which will provide the list of neighborhoods in Toronto (via Wikipedia), the Geographical location of the neighborhoods (via Geocoder package) and Venue data pertaining to Italian restaurants (via Foursquare). The Venue data will help find which neighborhood is best suitable to open an Italian restaurant.

Data Acquisition

Source 1: Wikipedia Table

The Wikipedia site provided almost all the information about the neighborhoods. It included the postal code, borough and the name of the neighborhoods present in Toronto. Since the data is not in a format that is suitable for analysis, scraping of the data was done. Wikipedia was fairly easy to scrape as it has a nice underlying structure.

	Postal Code	Borough	Neighbourhood
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	Downtown Toronto	Queen's Park, Ontario Provincial Government
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village
6	M1B	Scarborough	Malvern, Rouge
7	M3B	North York	Don Mills
8	M4B	East York	Parkview Hill, Woodbine Gardens
9	M5B	Downtown Toronto	Garden District, Ryerson

Source 2: Geographical Location Data

The second source of data provided us with the Geographical coordinates of the neighborhoods with the respective Postal Codes. The file was in CSV format, so we had to attach it to a Pandas DataFrame.

	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

Source 3: Venue Data using Foursquare

Using examples from previous Coursera instructional notebooks, we can adapt code to work for us. First, we call the api using the displayed code and then load it into use using a Pandas DataFrame as well.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	The Danforth West, Riverdale	43.679557	-79.352188	Pantheon	43.677621	-79.351434	Greek Restaurant

Methodology

Data Cleansing

After all the data was collected and put into data frames, cleansing and merging of the data was required to start the process of analysis. When getting the data from Wikipedia, there were Boroughs that were not assigned to any neighborhood therefore, the following assumptions were made:

1. Only the cells that have an assigned a borough will be processed. Borough's that were not assigned get ignored.
2. More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma.
3. If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

After the implementation of the following assumptions, the rows were grouped based on the borough as shown below.

	Postal Code	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront

Using the Latitude and Longitude collected from http://cocl.us/Geospatial_data ...

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711

I merged the two tables together based on Postal Code which resulted in the below dataframe.

	Postal Code	Latitude	Longitude	Borough	Neighborhood
0	M1B	43.806686	-79.194353	Scarborough	Malvern, Rouge
1	M1C	43.784535	-79.160497	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	43.763573	-79.188711	Scarborough	Guildwood, Morningside, West Hill

After, the venue data pulled from the Foursquare API was merged with the table above providing us with the local venue within a 500-meter radius shown below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	The Danforth West, Riverdale	43.679557	-79.352188	Pantheon	43.677621	-79.351434	Greek Restaurant

Data Exploration

Now after cleansing the data, the next step was to analyze it. We then created a map using Folium and color-coded each Neighborhood depending on what Borough it was located in.

Machine Learning

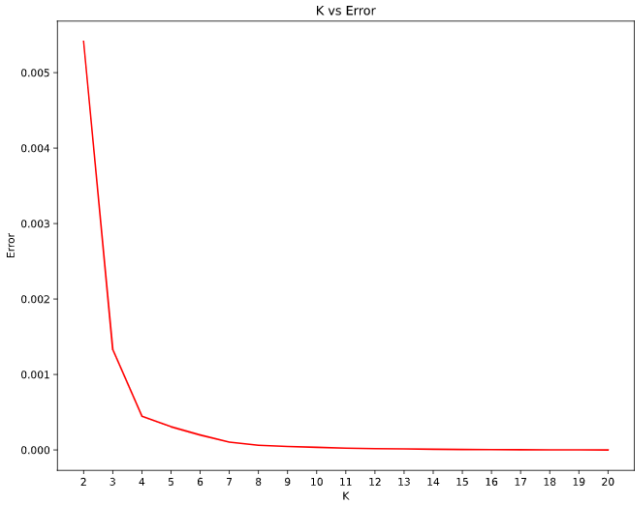
Then to analyze the data we performed a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called One hot encoding. For each of the neighborhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighborhood.

	Neighborhoods	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	...	Tibetan Restaurant	Toy / Game Store	Trail	Train Station
0	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	1	0
1	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	0	0
2	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	0	0
3	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	0	0
4	The Danforth West, Riverdale	0	0	0	0	0	0	0	0	0	...	0	0	0	0

Next, let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category.

	Neighborhoods	Italian Restaurant
0	Berczy Park	0.000000
1	Brockton, Parkdale Village, Exhibition Place	0.045455
2	Business reply mail Processing Centre, South C...	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000
4	Central Bay Street	0.047619

Now we will cluster the neighborhoods. We will use k-means clustering. But first we will find the best K using the Elbow Point method.



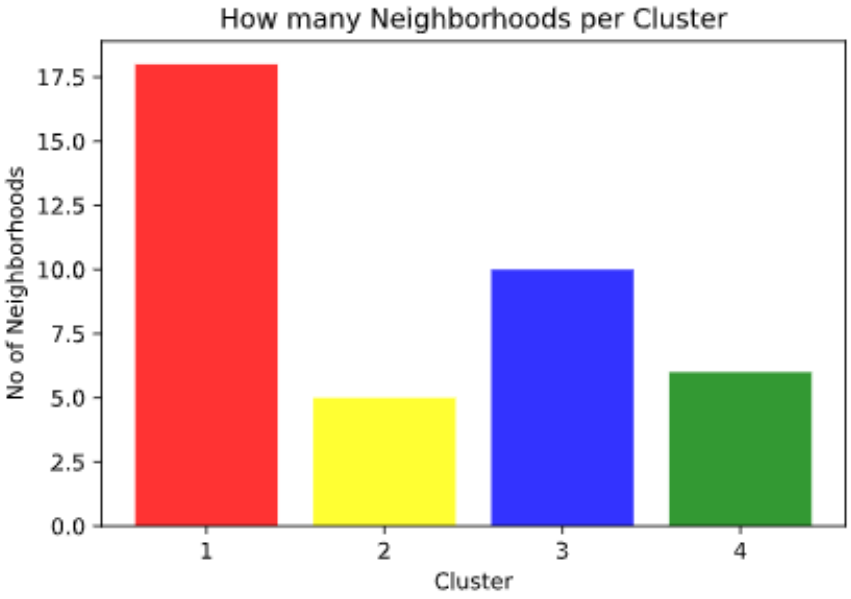
We see that the optimum K value is 4 so we will have a resulting of 4 clusters. After running our clustering algorithm, we can also see that there are a total of 40 locations with Italian Restaurants in Toronto. The final dataframe with has Neighborhood, proportion of Italian restaurants and Cluster Labels.

	Neighborhood	Italian Restaurant	Cluster Labels
0	Berczy Park	0.000000	0
1	Brockton, Parkdale Village, Exhibition Place	0.045455	3
2	Business reply mail Processing Centre, South C...	0.000000	0
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0
4	Central Bay Street	0.047619	3

The important thing to notice here is that we feature engineered the Italian restaurant proportion per neighborhood, and then used k-means clustering with 4 clusters to “attach” categories to these proportions. We can think of each cluster found by k-means as an interval between the entire range of these proportions. Pretty cool way to generate a feature and then utilize clustering to measure your intuition.

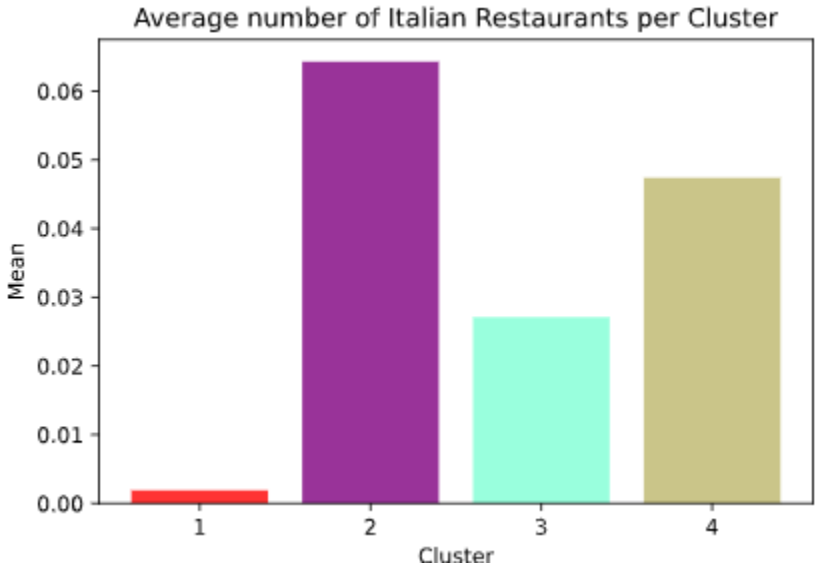
Data Analysis

We have a total of 4 clusters (0,1,2,3). Before we analyze them one by one let's check the total amount of neighborhoods in each cluster.



From the bar graph that was made using Matplotlib, we can compare the number of Neighborhoods per Cluster. We see that Cluster one has 18 neighborhoods while cluster two has 10 neighborhoods (second most). Cluster four has 8 neighborhoods and cluster three has only 6.

Then we compared the average Italian Restaurants per cluster. From the bar graph that was made using Matplotlib, we can compare the average Italian Restaurants per cluster. We see that Cluster one has the least Italian restaurants while cluster two has the most. Cluster three has around half the average number of Italian restaurants as cluster four has.



Cluster Analysis

This information is crucial as we can see that even though there is only 1 Italian restaurant in Cluster 1, it has the highest number of neighborhoods (0.1304) while Cluster 2 has the least neighborhoods but has the highest average of Italian Restaurants (0.0009). Also, from the map, we can see that neighborhoods in Cluster 2 are the most sparsely populated. Now let us analyze the clusters one by one.

Cluster 1

- Cluster 1 was mainly in the East Toronto Area. First Canadian Place, Underground city, Richmond, Adelaide, King were among some neighborhoods that were in that cluster. Cluster 1 had 173 unique Venue locations and out of those only 1 were Italian Restaurants. Cluster 1 had the lowest average of Italian Restaurants equating to 0.0.

Cluster 2

- There was a total of 140 neighborhoods, 63 different venues and 9 Italian Restaurants. Therefore, the average amount of Italian Restaurants that were near the venues in Cluster 2 is the highest being 0.07. In the map and from our numbers, we can see that nodes of Cluster 2 were dispersed mostly evenly throughout Toronto making it one of the most sparsely populated clusters.

Cluster 3

- Cluster 3 had the second to lowest average of Italian Restaurants. Cluster 3 was mainly located in the Downtown area but also had some neighborhoods in West Toronto, East Toronto and in North York. Neighborhoods such as Ryerson, Toronto Dominion Center, Garden District, Queen's Park and many more were included in this cluster. There was a total of 162 unique venues and out of those 20 were Italian Restaurants.

Cluster 4

- Cluster 4 venues were located in the Downtown, West, East Toronto areas. Neighborhoods such as Central Bay Street, St. James Town, Cabbagetown were some of the neighborhoods that made up this cluster. There was a total of 91 unique Venues in Cluster 4 with 10 Italian Restaurants. This made up the second-highest average of Italian Restaurants in that cluster which was approximately 0.047.

Discussion

Most of the Italian Restaurants are in cluster 2 represented by the yellow clusters. The Neighborhoods located in the East Toronto area that have the highest average of Italian Restaurants are The Danforth West and Riverdale. Even though there is a huge number of Neighborhoods in cluster 1, there is little to no Italian Restaurant. We see that in the Downtown Toronto area (cluster 3) has the second last average of Italian Restaurants. Looking at the nearby venues, the optimum place to put a new Italian Restaurant in cluster 1 as there are many Neighborhoods in the area but little to no Italian Restaurants, therefore, eliminating any competition. The second-best Neighborhoods that have a great opportunity would be in areas such as which is in Cluster 4. Having 90 neighborhoods in the area with 10 Italian Restaurants gives a good opportunity for opening a new restaurant, relative to cluster 3 where there are more neighborhoods and more Italian restaurants.

Some of the drawbacks of this analysis are — the clustering is completely based on data obtained from the Foursquare API. Also, the analysis does not take into consideration of the Italian population across neighborhoods as this can play a huge factor while choosing which place to open a new Italian restaurant. This concludes the optimal findings for this project and recommends the entrepreneur to open

an authentic Italian restaurant in these locations such as Cluster 1 because of the possibility that can expect little to no competition upon grand opening.

Conclusion

In conclusion, to end off this project, we had an opportunity on a business problem, and it was tackled in a way that it was similar to how a genuine data scientist would do. We utilized numerous Python libraries to fetch the information, control the content and break down and visualize those datasets. We have utilized Foursquare API to investigate the settings in neighborhoods of Toronto, get a great measure of data from Wikipedia which we scraped with the BeautifulSoup Web scraping Library. We also visualized utilizing different plots present in seaborn and Matplotlib libraries. Similarly, we applied AI strategy to anticipate the error given the information and utilized Folium to picture it on a map.

Places that have room for improvement or certain drawbacks give us that this project can be additionally improved with the assistance of more information and distinctive Machine Learning strategies. Additionally, we can utilize this venture to investigate any situation, for example, opening an alternate cuisine or opening of a Movie Theater and so forth. Ideally, this task acts as an initial direction to tackle more complex real-life problems using data science.