IBM Data Science Capstone Project

Analyzing the Neighborhoods of Toronto for Starting a New Restaurant.

Tope Orire

05/20/2021

1. Introduction

1.1. Background

Toronto is said to be a fast-growing city, that is unsurprising really with the number of new skyscrapers constantly popping up. The city is getting pretty much larger at the blink of an eye. The city has benefited from good planning from onset, making it a wonderful city to explore and be super-curios about. This is an incredible city for anyone to visit in Canada with tons of markets, delicious food, wonderful museums, and funky bars. Toronto is also the most multicultural city in the world, offering tons of diversity and culture (including lots of delicious Asian food). This report is a case study that analyzes the top neighborhoods in Toronto that would then be the best location to build a new restaurant for potential investors. I will be focusing majorly on most common venues in a particular neighborhood to determine how profitable a restaurant business will be there.

1.2. Business Problem

Potential Investors are interested in starting a new restaurant in Toronto. I have offered to help study the neighborhood and suggest a location that is best fit for this purpose. My main goal for this project is to properly analyze the right data about the several neighborhoods of Toronto using the data science skills I possess and then suggest to the potential investors, the best location to start a new restaurant.

2. Data

2.1. Neighborhood Data

The data used throughout this project analysis was scrapped from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M . The data was extracted using web scraping with the BeautifulSoup library in Python. This provides the up-to-date list of all neighborhoods present in Canada. The table (Figure 1) below shows an example of five neighborhoods after extracting data and cleanup.

	Postal Code	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

Figure 1: Example of five neighborhoods after extracting data and cleanup

2.2. **Geographical Coordinates**

The geographical coordinates of Toronto were extracted using Geospatial Data in Python. These coordinates are important for plotting the map of Toronto and later, the map of other neighborhoods in the city. These maps are useful while visualizing our data later in this report. I have provided a figure (figure 2) below to show the latitude and longitudes coordinates of some neighborhoods after merging with the data above.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Figure 2: showing the latitude and longitudes coordinates of some neighborhoods after merging.

2.3.Venue Data

The venue data has been extracted by querying the Foursquare API. Getting this data is necessary becomes it helps to see recommendations for all neighborhoods in the city. This data is then used to study popular venues in different neighborhoods. The query returned 1,583 neighborhoods which was more than enough for an in-depth analysis of the city. To be more specific, I have limited the scope of the study to focus on four boroughs around Toronto. So, I would mainly be analyzing Downtown, East, North and Central Toronto.

3. Methodology

3.1. Tools

The *matplotlib* function was mainly used to make explanatory and summary plot of data. The *json* package was useful when opening and reading geojson file. The *Folium* package for Python was used to plot neighborhoods on the map. *BeautifulSoup* was used to extract the data from the Wikipedia page. The *requests* library was then used to make query request to the Foursquare API. Finally, *K-means* module from *sklearn* package was used to cluster the data.

3.2. Feature Extraction

We had extract Features from the Foursquare API. These features are needed to build the unsupervised learning model to be used in this project. One Hot Encoding method was used to take all unique categories and then create a column for each category. In this method, if each feature is a category that belongs to a venue, it is converted into binary. This means that 1 refers to found category while 0 means not found. Below is a table showing the code used (Figure 3) and the result of the method after grouping by neighborhoods (Figure 4).

```
# one hot encoding
toronto_onehot = pd.get_dummies(toronto_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
toronto_onehot['Neighborhood'] = toronto_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [toronto_onehot.columns[-1]] + list(toronto_onehot.columns[:-1])
toronto_onehot = toronto_onehot[fixed_columns]

toronto_onehot.head()
```

Figure 3: one hot encoding

	Neighborhood	Yoga Studio	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	...	Theme Restaurant	Tibetan Restaurant	Toy / Game Store	T
0	Berczy Park	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0
1	Brockton, Parkdale Village, Exhibition Place	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0
2	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0.0	0.0	0.058824	0.058824	0.058824	0.117647	0.176471	0.117647	...	0.0	0.0	0.0	0
3	Central Bay Street	0.016393	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0
4	Christie	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0

Figure 4: the result of the method after grouping by neighborhoods.

3.3. Unsupervised Learning

The unsupervised learning approach used to cluster the neighborhoods is the K-means technique. The K-Means Method is a clustering algorithm that search clusters within the data, it then minimizes the data dispersion for each cluster. A new dataframe (Figure 5) that combined all the results into one table was created and used in the model. This table shows the most common venues in each neighborhood. For this project, 5 clusters will be used for the k-means clustering model.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Berczy Park	Coffee Shop	Cocktail Bar	Bakery	Restaurant	Seafood Restaurant	Cheese Shop	Beer Bar	Pharmacy	Farmers Market	Fish Market
1	Brockton, Parkdale Village, Exhibition Place	Café	Breakfast Spot	Coffee Shop	Intersection	Burrito Place	Stadium	Bar	Furniture / Home Store	Bakery	Nightclub
2	CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Lounge	Airport Terminal	Plane	Harbor / Marina	Sculpture Garden	Rental Car Location	Boat or Ferry	Bar	Boutique
3	Central Bay Street	Coffee Shop	Sandwich Place	Café	Italian Restaurant	Burger Joint	Salad Place	Bubble Tea Shop	Pizza Place	Japanese Restaurant	Discount Store
4	Christie	Grocery Store	Café	Park	Athletics & Sports	Coffee Shop	Italian Restaurant	Candy Store	Nightclub	Restaurant	Baby Store

Figure 5: most common venues in each neighborhood

3.4. Data Visualization (Mapping)

The Folium library was used to plot the maps of neighborhoods in Toronto (Figure 6) as well as the that of the neighborhoods in consideration for this project(Figure 7).

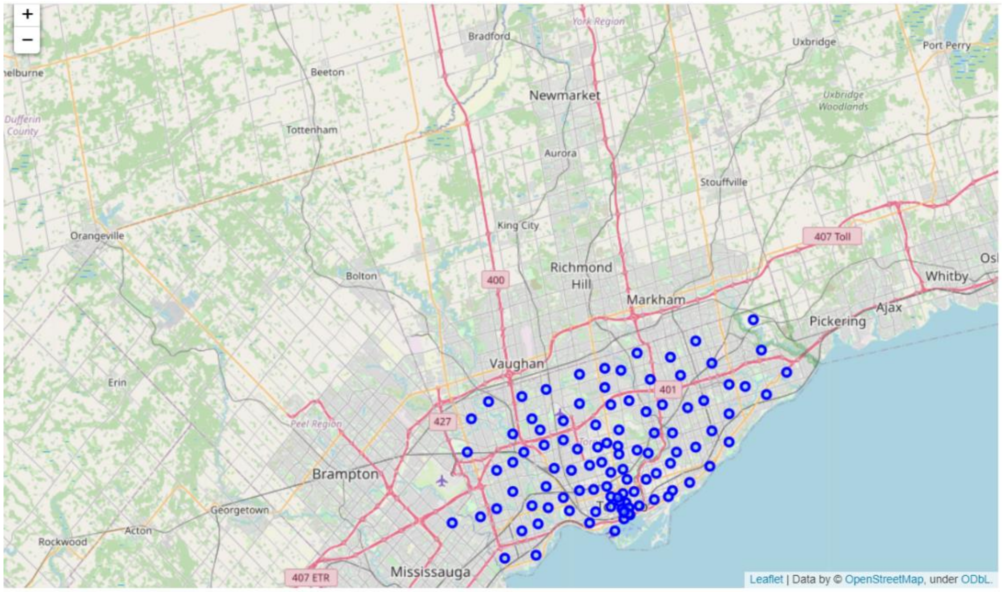

Figure 6: maps of neighborhoods in Toronto

Figure 7: neighborhoods in consideration for this project.

4. Results

I was able to cluster 5 neighborhoods in Toronto and this clusters provide a label from 0 to 4 for each cluster. It is interesting to see how this clusters came out because I am sure people in Toronto will probably agree that these clusters are reasonable and very much expected. Each cluster label was added to the dataframe in Figure 5 and they provide the borough Location, cluster Label and most common venues ranked from 1 to 10 as you would see in the Figures below.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	East Toronto	0	Health Food Store	Trail	Pub	Yoga Studio	Museum	Martial Arts School	Mediterranean Restaurant	Men's Store	Mexican Restaurant	Middle Eastern Restaurant
3	East Toronto	0	Fast Food Restaurant	Park	Sandwich Place	Pizza Place	Italian Restaurant	Fish & Chips Shop	Liquor Store	Food & Drink Shop	Light Rail Station	Steakhouse
24	Central Toronto	0	Trail	Jewelry Store	Bus Line	Sushi Restaurant	Yoga Studio	Museum	Mediterranean Restaurant	Men's Store	Mexican Restaurant	Middle Eastern Restaurant
26	Downtown Toronto	0	Café	Bookstore	Bakery	Japanese Restaurant	Bar	Restaurant	Bank	Beer Bar	Beer Store	Sandwich Place
28	Downtown Toronto	0	Airport Service	Airport Lounge	Airport Terminal	Plane	Harbor / Marina	Sculpture Garden	Rental Car Location	Boat or Ferry	Bar	Boutique
32	West Toronto	0	Bakery	Pharmacy	Wine Shop	Park	Recording Studio	Music Venue	Café	Middle Eastern Restaurant	Brewery	Supermarket
35	West Toronto	0	Mexican Restaurant	Café	Thai Restaurant	Arts & Crafts Store	Flea Market	Cajun / Creole Restaurant	Speakeasy	Fried Chicken Joint	Fast Food Restaurant	Furniture / Home Store

Figure 8: results for cluster label 0

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
9	Central Toronto	1	Park	Tennis Court	Yoga Studio	Museum	Market	Martial Arts School	Mediterranean Restaurant	Men's Store	Mexican Restaurant	Middle Eastern Restaurant

Figure 9: results for cluster label 1

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
23	Central Toronto	2	Garden	Home Service	Fast Food Restaurant	Music Store	Martial Arts School	Mediterranean Restaurant	Men's Store	Mexican Restaurant	Middle Eastern Restaurant	Miscellaneous Shop

Figure 10: results for cluster label 2

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	East Toronto	3	Greek Restaurant	Coffee Shop	Italian Restaurant	Ice Cream Shop	Furniture / Home Store	Yoga Studio	Frozen Yogurt Shop	Restaurant	Fruit & Vegetable Store	Liquor Store
4	East Toronto	3	Coffee Shop	Brewery	Gastropub	Café	Bakery	Music Store	American Restaurant	Cheese Shop	Clothing Store	Comfort Food Restaurant
6	Central Toronto	3	Gym / Fitness Center	Food & Drink Shop	Department Store	Hotel	Breakfast Spot	Dance Studio	Park	Sandwich Place	Middle Eastern Restaurant	Mexican Restaurant
7	Central Toronto	3	Clothing Store	Cosmetics Shop	Coffee Shop	Yoga Studio	Sporting Goods Shop	Park	Diner	Chinese Restaurant	Rental Car Location	Restaurant
8	Central Toronto	3	Sandwich Place	Pizza Place	Dessert Shop	Italian Restaurant	Café	Gym	Coffee Shop	Sushi Restaurant	Indian Restaurant	Brewery
10	Central Toronto	3	Coffee Shop	Restaurant	American Restaurant	Sushi Restaurant	Supermarket	Fried Chicken Joint	Sandwich Place	Pub	Bank	Bagel Shop
12	Downtown Toronto	3	Café	Pizza Place	Coffee Shop	Bakery	Chinese Restaurant	Restaurant	Pub	Italian Restaurant	Liquor Store	Bank
13	Downtown Toronto	3	Coffee Shop	Sushi Restaurant	Japanese Restaurant	Restaurant	Gay Bar	Men's Store	Hotel	Fast Food Restaurant	Pub	Mediterranean Restaurant
14	Downtown Toronto	3	Coffee Shop	Bakery	Park	Breakfast Spot	Theater	Café	Pub	Farmers Market	Chocolate Shop	Cosmetics Shop
15	Downtown Toronto	3	Coffee Shop	Clothing Store	Café	Japanese Restaurant	Bubble Tea Shop	Middle Eastern Restaurant	Hotel	Cosmetics Shop	Italian Restaurant	Lingerie Store

Figure 11: results for cluster label 3

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	East York/East Toronto	4	Park	Convenience Store	Yoga Studio	Museum	Martial Arts School	Mediterranean Restaurant	Men's Store	Mexican Restaurant	Middle Eastern Restaurant	Miscellaneous Shop
5	Central Toronto	4	Park	Swim School	Bus Line	Yoga Studio	Movie Theater	Martial Arts School	Mediterranean Restaurant	Men's Store	Mexican Restaurant	Middle Eastern Restaurant
11	Downtown Toronto	4	Park	Playground	Trail	Yoga Studio	Movie Theater	Market	Martial Arts School	Mediterranean Restaurant	Men's Store	Mexican Restaurant

Figure 12: results for cluster label 4

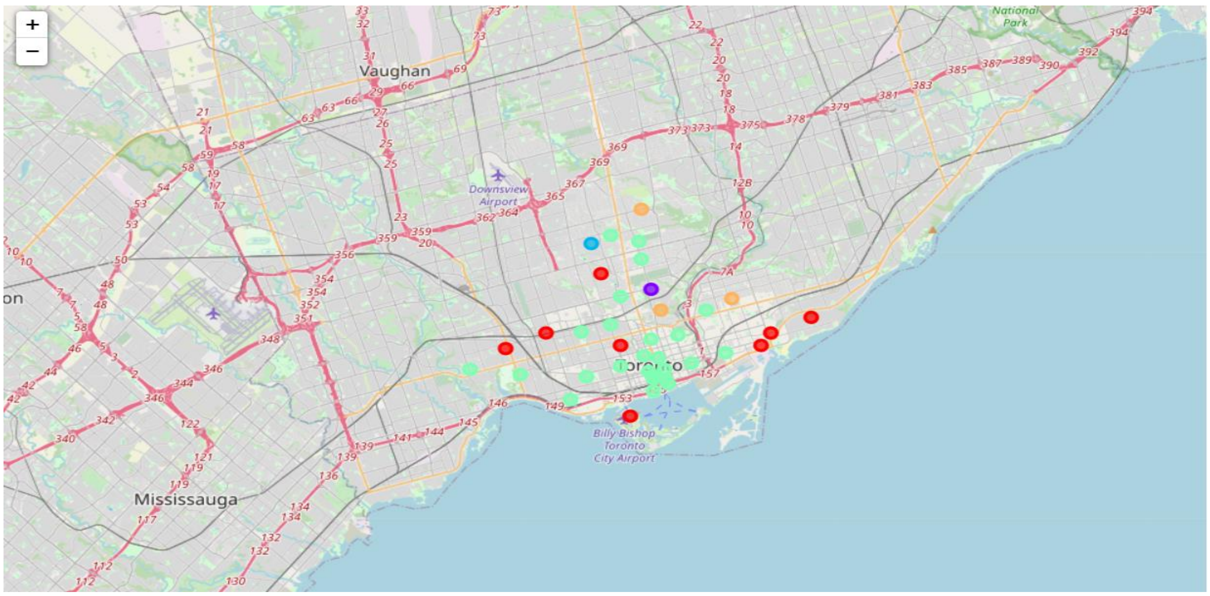

Figure 13: map visualization after clustering.

5. Discussion

We can see from the results shown above that Cluster 3 will provide the best neighborhoods to choose from when starting a new restaurant business. The K-Means model worked well in helping to cluster similar neighborhoods and show what their common venues are.

I will, therefore, recommend to the potential investors that neighborhoods in East, Central and Downtown Toronto will be best fit to starting their business. Although, it is also to be noted that there would be a high rate of competitions in these neighborhoods considering the number of restaurants already present in them. If the potential investors wish to work is a less competitive neighborhood, I will then suggest cluster 0, that shows us that there are a lot of people around this region visiting parks and recreational centers. This will also be a good spot to that a new restaurant.

6. Conclusion

This has been a great project for me to embark on. The data analysis and machine learning techniques used in this project were very helpful in determining solutions to the problem at hand. This report has been able to analyze and adequately recommend the best neighborhoods to start a new restaurant. The potential investors can then move forward with considering other factors like transportation, legal requirements and cost associated in starting a new restaurant. Those were not part of the scope of this project and therefore not considered.