Tea or Coffee?

Introduction

Problem Background:

Toronto is one of the most multicultural urban areas in the world. Each year tens of thousands of newcomers from around the globe choose our city as their new home. Their diverse cultures and communities have helped create Toronto's identity as a vibrant global city.

Toronto is Canada's largest city and a world leader in such areas as business, finance, technology, entertainment, and culture. Its large population of immigrants from all over the globe has also made Toronto one of the most multicultural cities in the world. Being Canada's largest City, the city is home to 2,956,024 persons who live within a 630 square kilometre area.

Selecting the right location for a business is one of the first and very important decision in running a business. Starting a new business in a metropolitan city such as Toronto can be challenging. Toronto has 39 neighbourhoods. It is important to evaluate different neighbourhoods based on the factors that are important for running a successful business such as the number of competitors within a geographical area.

Problem Description

An entrepreneur is interested in opening a Coffee Shop in Toronto but needs to identify the best location to do so. Therefore, the objective of this project is to determine what might be the 'best' neighbourhood(s) in Toronto to open a Coffee Shop.

Target Audience:

Entrepreneurs who are interested in opening a new Coffee Shop or expanding their existing chain in the Toronto area.

Data Overview

To identify the optimal location, we would need to determine the following:

- The number of neighbourhoods in Toronto
- The geographical location of these neighbourhoods
- The number of venues that are Coffee Shops

To determine the items listed above we would use the following datasets:

- Dataset 1: A Wikipedia page which will provide the postal code, borough and the name of the neighbourhoods present in Toronto (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- Dataset2: The CSV file of geographical coordinates from the Geocoder package which has the geographical coordinates of each postal code. (http://cocl.us/Geospatial_data)
- Dataset3: Foursquare API to get venue data pertaining to restaurants.

Methodology

Data Preparation

The first step in determining the optimal location for opening a new Coffee Shop would be to extract the data from the sources listed and store them in a format that is analysis friendly.

Each dataset was extracted from the data sources identified above, cleansed, and then merged.

The following actions were taken to clean up Dataset 1: Dataset 1: A Wikipedia page which will provide the postal code, borough and the name of the neighbourhoods present in Toronto (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

1. All rows with no Borough assigned were removed.
2. All rows with a Neighbourhood that was not assigned was replaced with the name of the Borough.
3. All rows were grouped based on the Postal Code as more than one neighbourhood can exist in one postal code area.

This data frame was then merged with Dataset 2: The CSV file of geographical coordinates from the Geocoder package to generate the table below.

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
37	M4E	East Toronto	The Beaches	43.676357	-79.293031
41	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
42	M4L	East Toronto	India Bazaar, The Beaches West	43.668999	-79.315572
43	M4M	East Toronto	Studio District	43.659526	-79.340923
44	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

Figure 1: Merged Tables from Dataset 1 and Dataset 2

The data frame above was then restricted to Toronto Neighbourhoods and merged with the data from Dataset 3: Foursquare API to generate a data frame including all the local venues within a 500-meter radius shown below.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood 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	MenEssentials	43.677820	-79.351265	Cosmetics Shop

Figure 2: Merged Tables from all datasets

Results

Exploratory Data Analysis

Now that the data has been prepped for analysis, we will begin exploring the data.

Using the list of venues generated using the Foursquare API, we were able to identify that there are 158 Coffee shops in the Toronto Area. The following table shows the top 10 neighbourhoods in terms of the number of Coffee Shops in each neighbourhood.

Neighbourhood	Venue Category
Toronto Dominion Centre, Design Exchange	14
Commerce Court, Victoria Hotel	14
First Canadian Place, Underground city	13
Harbourfront East, Union Station, Toronto Islands	12
Stn A PO Boxes	11
Central Bay Street	11
Queen's Park, Ontario Provincial Government	9
Church and Wellesley	8
Richmond, Adelaide, King	7
Regent Park, Harbourfront	7

Figure 3: Number of Coffee Shops by Neighbourhood.

Machine Learning Approach

In order to prep the data for using a machine learning algorithm, a method called One-hot Encoding was applied to the dataset for each venue to identify whether the a certain type of venue was located within a specific neighbourhood, a snapshot of this transformation is shown below.

	Neighbourhood	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant
0	The Beaches	0	0	0	0	0	0	0	0	0	0	0	0	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

Figure 4: Transformed Dataset using One-Hot Encoding

Using the transformed dataset, the rows were grouped by Neighbourhood and the mean value of Venue occurrences for each Venue was generated. This new dataset was then modified to show these values specifically for Coffee Shops.

	Neighbourhood	Coffee Shop
0	Berczy Park	0.087719
1	Brockton, Parkdale Village, Exhibition Place	0.090909
2	Business reply mail Processing Centre, South C...	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.066667
4	Central Bay Street	0.171875

Figure 5: Transformed Dataset displaying the mean occurrence of Coffee Shops.

The next stage of the analysis was to cluster the neighbourhoods based on the mean value of occurrences. To do this, we implemented the k means clustering algorithm to generate 4 clusters. Neighbourhoods that had a similar mean frequency of Coffee Shops were divided into 4 clusters. Each of these clusters was labelled from 0 to 3 because the indexing of labels begins with 0 instead of 1.

	Cluster Labels	Neighbourhood	Coffee Shop
0	2	Berczy Park	0.087719
1	2	Brockton, Parkdale Village, Exhibition Place	0.083333
2	0	Business reply mail Processing Centre, South C...	0.000000
3	2	CN Tower, King and Spadina, Railway Lands, Har...	0.062500
4	1	Central Bay Street	0.169231

Figure 5: Transformed Dataset displaying the Cluster labels.

Analysis of Clusters

Now, that we have generated our clusters, let us look at the number of neighbourhoods in each cluster as well as the mean occurrences of Coffee shops in each cluster.

Cluster Labels	Neighbourhood
0	12
1	10
2	16
3	1

Cluster Labels	Coffee Shop
0	0.000000
1	0.131837
2	0.073694
3	0.250000

Figure 6: Number of Neighbourhoods and Mean Coffee Shop occurrences by Cluster.

From the figures above, Cluster 4 (labelled as 3) has only 1 neighbourhood, while cluster 1(labelled as 0) has the most with 12. It is interesting to note that although there is only 1 neighbourhood in Cluster 4 (labelled as 3), it has the highest number of Coffee Shops (0.250), while Cluster 1(labelled as 0) has the second most neighbourhoods but has no occurrences of Coffee Shops.

Looking at the map below to get an understanding of how the clusters are spread out, it was observed that neighbourhoods in Cluster 1 (labelled as 0 - Blue) are the most sparsely populated, followed by in Cluster 3 (labelled as 2 - Purple).

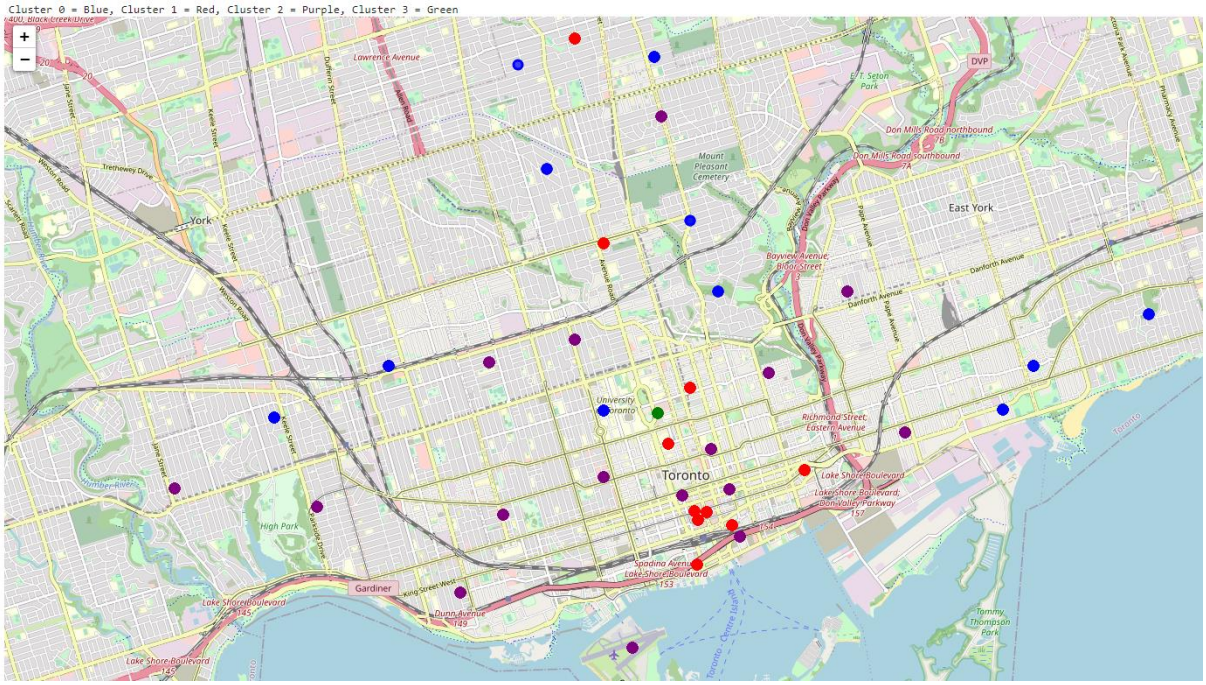


Figure 7: Geographical locations of Clusters.

We will now look at each Cluster Individually.

Cluster 1 (labelled as 0 – Blue)

Cluster 1 was in Downtown, Central, East and West Toronto with 12 neighbourhoods in that Cluster. Cluster 1 had 78 unique Venue locations and no Coffee Shops. Notably there exists 8 Cafés in this cluster.

There are 78 uniques categories.

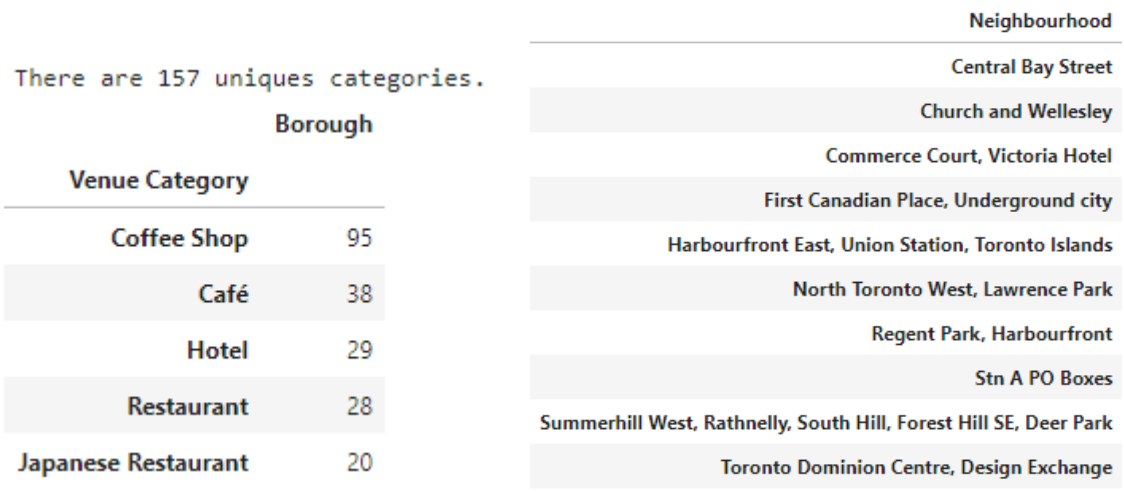
Venue Category	Borough
Park	10
Café	8
Bakery	5
Restaurant	4
Sandwich Place	4

Neighbourhood
Business reply mail Processing Centre, South Central Letter Processing Plant Toronto
Davisville North
Dufferin, Dovercourt Village
Forest Hill North & West, Forest Hill Road Park
High Park, The Junction South
India Bazaar, The Beaches West
Lawrence Park
Moore Park, Summerhill East
Rosedale
Roselawn
The Beaches
University of Toronto, Harbord

Figure 8: Number of Unique Venues, Coffee Shops and Neighbourhoods in Cluster 1.

Cluster 2 (labelled as 1 – Red)

Cluster 2 was in Downtown and Toronto with 10 neighbourhoods in that Cluster. Cluster 2 had 157 unique Venue locations and out of those 95 were Coffee Shops. Cluster 2 had the second highest average of Coffee Shops with a mean occurrence of 0.132 because the 95 Coffee Shops are spread out in 10 neighbourhoods.



There are 157 uniques categories.

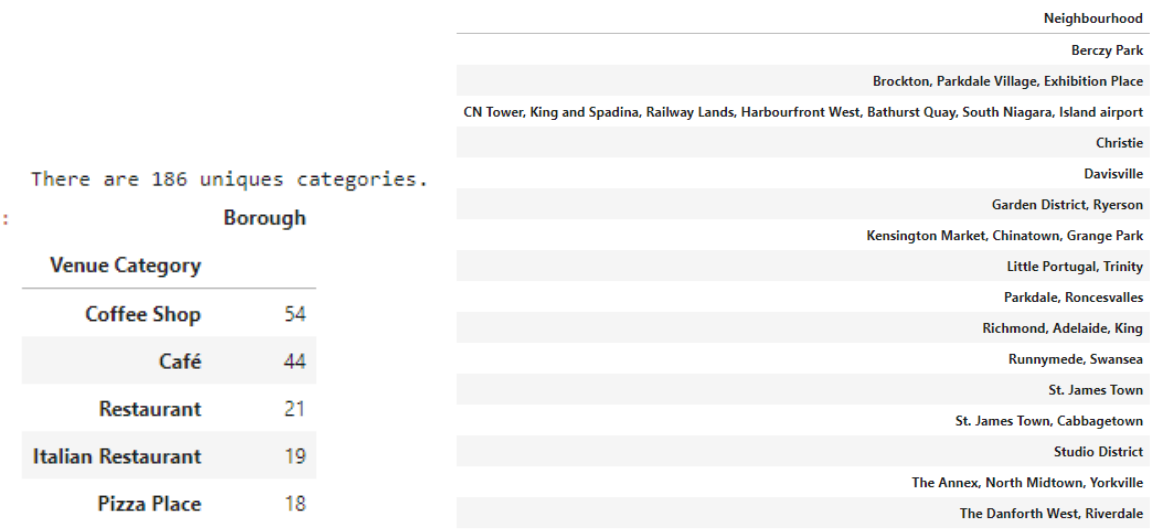
Venue Category	Borough
Coffee Shop	95
Café	38
Hotel	29
Restaurant	28
Japanese Restaurant	20

Neighbourhood
Central Bay Street
Church and Wellesley
Commerce Court, Victoria Hotel
First Canadian Place, Underground city
Harbourfront East, Union Station, Toronto Islands
North Toronto West, Lawrence Park
Regent Park, Harbourfront
Stn A PO Boxes
Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park
Toronto Dominion Centre, Design Exchange

Figure 9: Number of Unique Venues, Coffee Shops and Neighbourhoods in Cluster 2.

Cluster 3 (labelled as 2 – Purple)

Cluster 3 was in Downtown, Central, East and West Toronto with 16 neighbourhoods in that Cluster. Cluster 3 had 186 unique Venue locations and out of those 54 were Coffee Shops. Cluster 3 had the second lowest average of Coffee Shops with a mean occurrence of 0.0737 because the 54 Coffee Shops are spread out in 16 neighbourhoods.



There are 186 uniques categories.

Venue Category	Borough
Coffee Shop	54
Café	44
Restaurant	21
Italian Restaurant	19
Pizza Place	18

Neighbourhood
Berczy Park
Brockton, Parkdale Village, Exhibition Place
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport
Christie
Davisville
Garden District, Ryerson
Kensington Market, Chinatown, Grange Park
Little Portugal, Trinity
Parkdale, Roncesvalles
Richmond, Adelaide, King
Runnymede, Swansea
St. James Town
St. James Town, Cabbagetown
Studio District
The Annex, North Midtown, Yorkville
The Danforth West, Riverdale

Figure 10: Number of Unique Venues, Coffee Shops and Neighbourhoods in Cluster 3.

Cluster 4 (labelled as 3 – Green)

Cluster 4 was in Downtown Toronto in the Queen's Park neighbourhood. Cluster 3 had 186 unique Venue locations and out of those 9 were Coffee Shops. Cluster 4 had the highest average of Coffee Shops with a mean occurrence of 0.250 because the 9 Coffee Shops are in 1 neighbourhood.

Discussion

Most of the Coffee Shops are in cluster 2. There is a huge number of Neighbourhoods in cluster 1, but no coffee shops. We see that in the Downtown Toronto area (cluster 4) has the highest average of Coffee Shops. Looking at the nearby venues, the optimum places to put a new Coffee Shop is in Cluster 1 or Cluster 3 as there are many Neighbourhoods in the area but little to no Coffee Shops. therefore, reducing any competition. In Cluster 1, there are 12 neighbourhoods in the area with no Coffee Shop, so this is a good opportunity for opening a new shop.

The limitations associated with this analysis are:

- Clustering is completely based Foursquare API data.
- The analysis does not account for Coffee Shops across neighbourhoods which is very important in deciding where is the optimal location for opening a new shop.

In conclusion, it is suggested that the ideal location would be in Cluster 1, however more research would need to be done to determine the number of recreational sites such as Parks and Galleries etc. as well as high traffic areas so that the store can be located within proximity of these venues to fill the void that is missing. Also, there are also 8 Cafes within this area, so it would be important to determine their offerings as well as locations of each in relation to the high traffic areas to determine the Optimal location.

Conclusion

Entrepreneurs are looking to big cities to start a business. For this reason, they can make better business decisions by understanding the data available to them using the platforms where such information is provided.