IBM Applied Data Science Capstone

Opening a New Restaurant in Toronto, Canada

By Suyash Singh

June 2020

Introduction

With the rising economic growth in the Greater Toronto Area (GTA), people have higher spending power, and this can be seen in the growing number of hotels, restaurants and entertainment centers. The diversity in this region has allowed the rise of a variety of cuisines being offered. We have been approached by a property dealer to identify potential areas for a new restaurant business. We are tasked with the goal of identifying what cuisines would work in each area.

Business Problem

The investors are not picky about the cuisine and want to choose the cuisine that will give them the highest profits. They are looking to open a restaurant in the Toronto area. We are required to analyze the different areas of Toronto and identify the more popular cuisines in these areas. For this purpose, we will implement a k-means algorithm to cluster different areas of Toronto based on the popular restaurants in the region. At the end of the analysis, we will be able to identify regions with potential for a new restaurant and will recommend the cuisine to offer.

Data

Data Sources

The data used in this analysis come from:

1. Wikipedia
2. Foursquare API

Wikipedia(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) provides a table containing postal codes and the associated neighborhoods of Toronto. We use the BeautifulSoup package in Python to scrape the table from the website and modify it for our use.

We will then extract the latitude and longitude values for these neighborhoods using the Geocoder package in Python.

List of restaurants will be extracted using the Foursquare API. We will use the latitude and longitude values of the neighborhoods from previous step and find restaurants within 500 metres radius.

Once we have cleaned the data after extraction, we will apply k-means clustering algorithm to identify the identical clusters.

Data Extraction

Wikipedia

We use Wikipedia to scrape a table containing neighborhoods of Toronto and their postal codes. We use BeautifulSoup package to extract this table. After cleaning the data, the final table looks like the one below.

	Postal Code	Borough	Neighborhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
8	M9A	Etobicoke	Islington Avenue, Humber Valley Village
9	M1B	Scarborough	Malvern, Rouge
11	M3B	North York	Don Mills
12	M4B	East York	Parkview Hill, Woodbine Gardens
13	M5B	Downtown Toronto	Garden District, Ryerson

Figure 1: Postal Code and relevant Neighborhoods in Toronto

Geocoder (Python package)

We then use the postal code to extract their latitude and longitude values using the Geocoder package on python. The data after extracting these values looks like this:

	Postal Code	Borough	Latitude	Longitude
0	M3A	North York	43.753259	-79.329656
1	M4A	North York	43.725882	-79.315572
2	M5A	Downtown Toronto	43.654260	-79.360636
3	M6A	North York	43.718518	-79.464763
4	M7A	Downtown Toronto	43.662301	-79.389494
5	M9A	Etobicoke	43.667856	-79.532242
6	M1B	Scarborough	43.806686	-79.194353
7	M3B	North York	43.745906	-79.352188
8	M4B	East York	43.706397	-79.309937

Figure 2: Latitude and Longitude values of Postal Codes

FourSquare API

We wrote a function that retrieves nearby venues for each neighborhood. It does this using the FourSquare API. We also obtain the latitude and longitude values of these venues. We manipulate the dataset to obtain the count of top ten most common category of restaurants in each neighborhood and apply mean function on the columns to convert the values to mean values. This is done to implement the k-means algorithm. The final dataset that will be used for the clustering looks like this:

Neighborhood	Afghan Restaurant	American Restaurant	Asian Restaurant	Belgian Restaurant	Brazilian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant
Berczy Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Brockton, Parkdale Village, Exhibition Place	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Business reply mail Processing Centre, South C...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Central Bay Street	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Christie	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Church and Wellesley	0.013158	0.013158	0.000000	0.000000	0.000000	0.000000	0.013158

Figure 3: Final dataset

Methodology

In this analysis, we are going to develop a K-means clustering algorithm to cluster neighborhoods in Toronto based on the most common type of restaurants in the area. We generated a map of Toronto displaying the various boroughs using Folium.

Figure 4:Map generated using Folium

The final dataset on which we apply the k-means algorithm contains mean of all restaurants in each neighborhood. We apply the algorithm to see what kind of clusters we get.

Elbow Method (Cluster number selection)

We will first use the elbow method to identify the ideal number of clusters to be used. Here we calculate the sum of squared distances of the values in the dataset and plot its value with increasing value of K. The K at which we see an elbow is the ideal number of clusters to use.

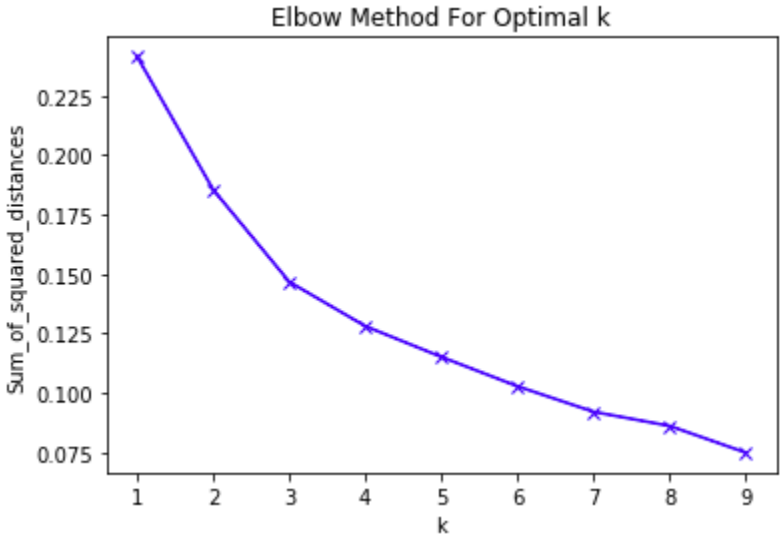

Figure 5: Number of Clusters VS Sum of Squared Distances

The optimal K value is at k = 5. Hence, we will use this value as the number of clusters for the k-means clustering algorithm. We plot the results of the clustering on a map of Toronto using Folium. The representation looks like this:

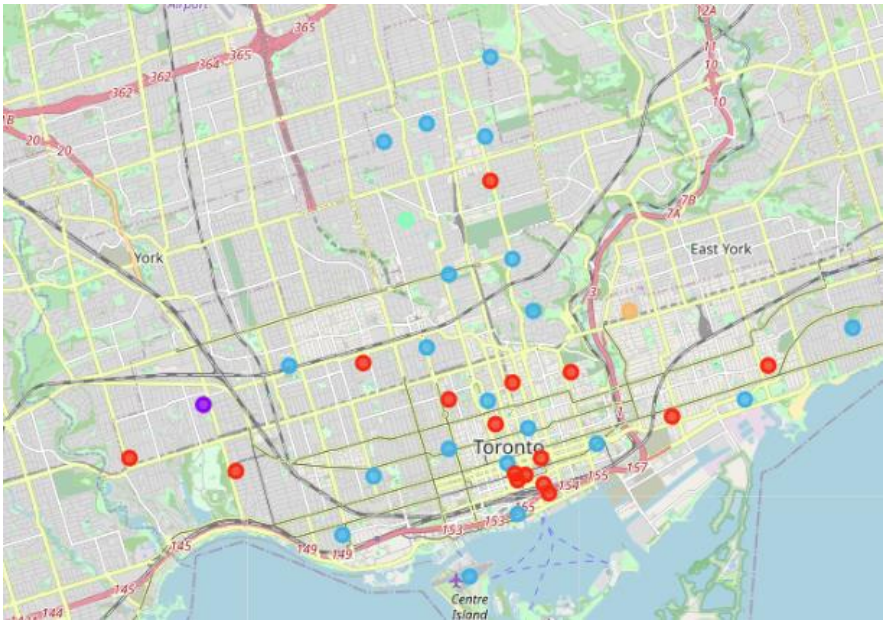

Figure 6: Map of Toronto after K- means Clustering

Results

The K-means clustering algorithm gave us 5 clusters based on the restaurants in Toronto area. The clusters are mentioned below.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
St. James Town	High Park, The Junction South	Regent Park, Harbourfront	Forest Hill North & West, Forest Hill Road Park	The Danforth West, Riverdale
Berczy Park		Queen's Park, Ontario Provincial Government		
Central Bay Street		Garden District, Ryerson		
Christie		The Beaches		
Toronto Dominion Centre, Design Exchange		Richmond, Adelaide, King		
India Bazaar, The Beaches West		Dufferin, Dovercourt Village		
Commerce Court, Victoria Hotel		Harbourfront East, Union Station, Toronto Islands		
Studio District		Little Portugal, Trinity		
Parkdale, Roncesvalles		Brockton, Parkdale Village, Exhibition Place		
Davisville		Lawrence Park		
University of Toronto, Harbord		Roselawn		
Runnymede, Swansea		Davisville North		
Stn A PO Boxes		North Toronto West, Lawrence Park		
St. James Town, Cabbagetown		The Annex, North Midtown, Yorkville		
First Canadian Place, Underground city		Moore Park, Summerhill East		
Church and Wellesley		Kensington Market, Chinatown, Grange Park		
		Summerhill West, R athnelly, South Hill, Forest Hill SE, Deer Park		

		Rosedale		
		Business reply mail Processing Centre, South Central Letter Processing Plant Toronto		
		CN Tower, King and Spadina, Railway La nds, Harbourfront West, Bathurst Qua y, South Niagara, Isl and airport		

The first cluster is strongly dominated by Italian and American Restaurants as well as a few other cuisines in smaller numbers. The second cluster has strong presence of Thai and Mexican Restaurants. The third cluster has many Vietnamese restaurants. We also notice in this group that the number of restaurants in this region aren't many. The fourth cluster has a strong presence of Sushi Restaurants and the final cluster is dominated by Greek and Italian Restaurants.

Conclusion

The clustering has allowed us to determine the kind of restaurant distribution in the Toronto region. Business owners looking to open a restaurant in areas under Cluster 1 should explore the Italian and American Restaurant. We see a strong presence of these restaurants in this cluster. Other cuisines have little or no presence in these regions.

Cluster 2 contains only contains the neighborhoods of High Park and The Junction South. These neighborhoods are good for a Thai and Mexican restaurant business. We see a strong presence of these restaurants here. Cajun and Fast Food restaurants also have higher presence than other clusters. This cluster generally looks good for the restaurant business.

Cluster 3 contains many Vietnamese Restaurants. We also see a presence of Vegetarian/ Vegan cuisines. There are lower number of restaurants in these clusters. This maybe because these neighborhoods may be housing areas or occupied by other structures that prevent the rise of a vibrant restaurant business. More data may be required to deduce why exactly these areas don't have many restaurants. We would not advise opening a restaurant business in these neighborhoods.

Cluster 4 is only dominated by Sushi restaurants. No other restaurant business is present in these neighborhoods. We believe only Sushi restaurants have scope for business in these regions.

Cluster 5 contains the Danforth West and Riverdale neighborhoods. These neighborhoods have many Greek and Italian Restaurants. We recommend exploring these cuisines in this region.

This analysis gives us information about what kind of cuisines are doing better business in different areas of Toronto region. We conclude that Clusters 1,2,4 and 5 have scope for a new restaurant business. Cluster 3 is the only cluster to avoid investing in.