Capstone Project: Battle of the Neighborhoods

Analyzing the city of Toronto for selecting the ideal location for a new Italian restaurant

-Pranav Natarajan

Introduction

For most of us, we enjoy the occasional fine-dine experience at a quant local restaurant. It is indeed a great way to relax and enjoy time with friends, family and loved ones during weekends and holidays. With the decline in commercial real estate lease costs, it would be a good opportunity for restaurateurs to open a restaurant in the city of Toronto taking advantage of the lower real estate costs. Toronto is a growing city with a rising youth population and a multi-cultural and cosmopolitan environment. While there are a multitude of cuisines that people savor, and Italian cuisine is one of the most sought-after cuisines when it comes to a fine dine experience and is personal favorite. Of course, as with any business decision, opening a new restaurant requires serious consideration and is a lot more complicated with various facets to it. Particularly, the location of the restaurant is one of the most important decisions that will determine whether it will be a success or a failure.

Business Problem

The objective of this capstone project is to analyze and select the best locations in the city of Toronto to open a new Italian restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Toronto , if an entrepreneur is looking to open a new Italian restaurant, where would you recommend that they open it?

Target audience

This project is particularly useful to property developers, investors and restaurateurs looking to open or invest in a new Italian restaurant in the financial capital and growing city of Toronto, Canada. This project is timely as the city is currently facing a rebalancing of commercial real estate prices due to onset of the unfortunate COVID pandemic. Based on various news sources commercial real estate has dropped around 10% and could fall further as social distancing norms and mandates cause pressures in the short run. This would give sufficient time to ramp up the restaurant and as the vaccine become more easily available social distancing mandates would be loosened. The pent-up demand would surely result in increased consumer spending once retail and recreational outlets resume services.

Data

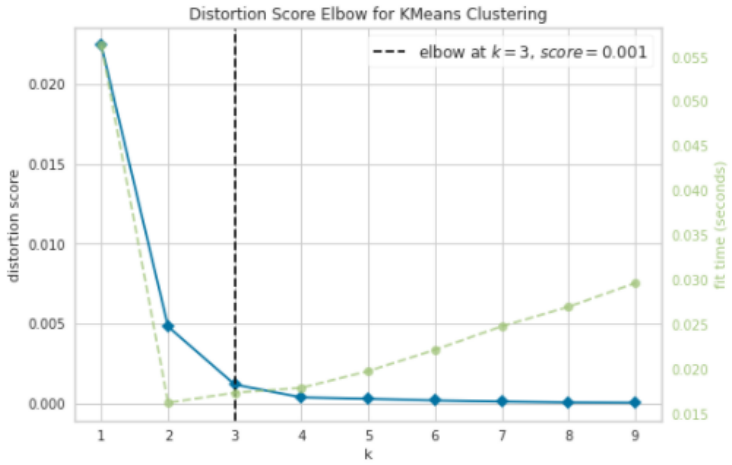
To solve the problem, we will need the following data:

- List of neighborhoods in Toronto.
- Latitude and longitude coordinates of those neighborhoods to plot the map and to get the relevant venue level data.
- Venue data related to Italian restaurants which we will use to perform clustering on the neighborhoods.

Methodology

Firstly, we need to get the list of neighborhoods in Toronto which is available in the Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. We perform web scraping using Python requests and Beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the Geocoder package that allows us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are looking into Italian restaurants, we will filter based on Italian restaurants as venue category for the neighborhoods. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will use the "KElbowVisualizer" from the K-means package to visualize and decide the optimal number of clusters for the data. Based on the analysis we cluster the neighborhoods into 3 clusters based on their frequency of occurrence for "Italian restaurants".



The results will allow us to identify which neighborhoods have higher concentration of Italian restaurants while which neighborhoods have fewer number. Based on the occurrence of in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable.

Sources of data and extraction methods

This Wikipedia page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M contains a list of neighborhoods in Toronto. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and BeautifulSoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods. After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare API will provide many categories of the venue data, we are particularly interested in the Italian restaurant in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

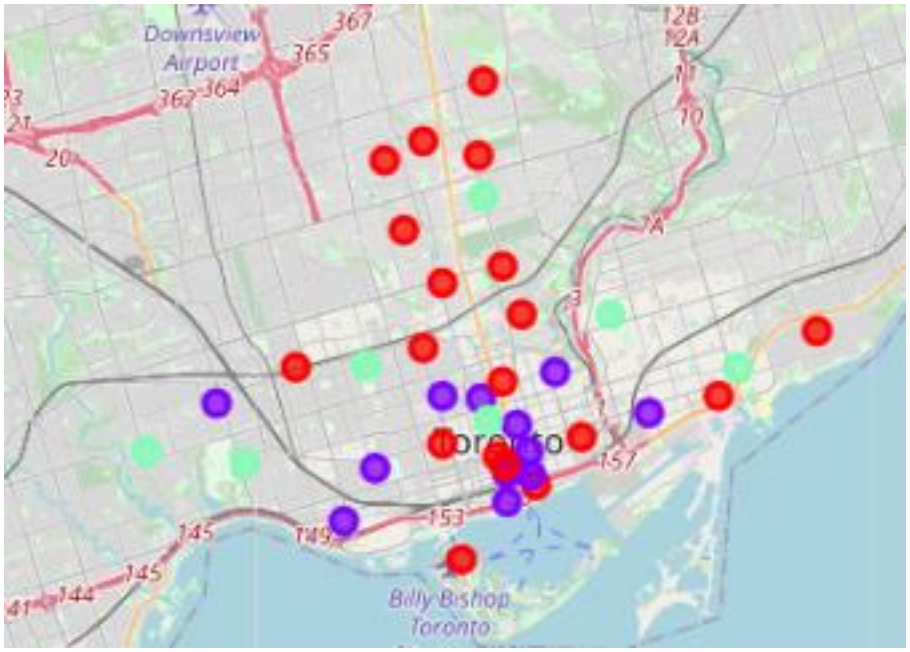
Results

We found the most common venues in each neighborhood to get a better sense of the data

After which we performed k-means clustering to find out the optimal number of 3 clusters for the data

- Cluster 1 (Red dots): Had the least amount of Italian restaurants
- Cluster 2 (Blue dots): Had moderate number
- Cluster 3 (Light green dots): Had the highest incidence

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
Berczy Park	Coffee Shop	Cocktail Bar	Seafood Restaurant	Restaurant	Cheese Shop	Bakery	Farmers Market
Brockton, Parkdale Village, Exhibition Place	Café	Breakfast Spot	Nightclub	Coffee Shop	Pet Store	Stadium	Bar
Business reply mail Processing Centre, South C...	Light Rail Station	Skate Park	Restaurant	Recording Studio	Fast Food Restaurant	Farmers Market	Auto Workshop
CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Lounge	Coffee Shop	Harbor / Marina	Rental Car Location	Sculpture Garden	Boutique
Central Bay Street	Coffee Shop	Sandwich Place	Italian Restaurant	Café	Burger Joint	Bubble Tea Shop	Salad Place



Discussion

As observations noted from the map in the Results section, most of the Italian restaurants are concentrated in the downtown (southside) of Toronto, with the highest number in cluster 3 and minimal to no restaurants in cluster 1. On the other hand, cluster 2 had a moderate number of Italian Restaurants in the neighborhoods. Cluster 1 represents a sizable opportunity with a high potential to open a successful Italian restaurant given the low incidence in the first place. Opening a restaurant in Cluster 1 would result in minimal competition. Meanwhile, restaurants in cluster 3 are likely suffering from intense competition due to oversupply and high concentration. From another perspective, the results also show that the oversupply of Italian restaurants mostly happened in the downtown area of the city, with the suburb area still have very Italian restaurants. Therefore, this project recommends property developers and restaurateurs to capitalize on these findings to open a new Italian in the neighborhoods in cluster 1 with little to no competition. Furthermore, the recent COVID pandemic has resulted in lower commercial property rates and reduced lease costs. This would be a great opportunity to capitalize on the persisting macro-economic environment for entrepreneurs with access to capital to set up a new restaurant. Closing in on a property and ramping it up over a period of 8 months would be in right time if there is a suitable roll out of vaccinations resulting in more relaxed social distancing

norms. Lastly, property developers are advised to avoid neighborhoods in cluster 3 which already have high concentration of Italian restaurants and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of Italian restaurants, there are other factors such as population and income of residents that could influence the location decision. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new Italian restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new Italian restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions.

References

Category: Suburbs in Toronto. Wikipedia. Retrieved from

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Foursquare Developers Documentation. Foursquare. Retrieved from

https://developer.foursquare.com/docs