Find Locations to open a Italian Restaurent Using Maching Learnning

1-Introduction

Background

For this Capstone project , I am creating a hypothetical scenario for a concept Italian restaurateur who wants to explore opening an authentic Italian restaurant in Toronto and it might present a great opportunity for this entrepreneur who is based in Canada , this entrepreneur is thinking of opening this restaurant in locations where Italian food is popular . One of the important decisions for this entrepreneur is finding the location to open such a restaurant , so in this project I am trying to help him find the most suitable location .

Business Problem

The objective of this capstone project is to find the most suitable location for the entrepreneur in order to open a new Italian restaurant in Toronto , Canada.

By using data science methods and Maching Learnning methods such as clustering , this project aims to provide solutions to answer the business question : In Toronto if a an entrepreneur wants to open an Italian restaurant ,

Where should they consider opening it ?

Target Audience

The entrepreneur who wants to find the location to open authentic Italian restaurant .

Data

To solve this problem , we will need below data :

- List of neiborhoods in Toronto , Canada :

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

- Latitude and Longgitude of these neihborhoods.

- Venue data related to Italian restaurents , this help us find the neiborhoods that are most suitable to open an Italian restaurant .

Extracting Data

- Scrapping of Toronto neighborhoods via Wikipedia
- Getting Latitude and data of these neighborhoods via Geocoder package
- Using Foursquare API to get venue data related to these neigbhorhoods

For this Capstone project , I am creating a hypothetical scenario for a concept Italian restaurateur who wants to explore opening

Methodology :

First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from wikipedia page (“https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M ”), I did the web scraping by utilizing pandas html table scraping method as it is easier and more convenient to pull tabular data directly from a web page into dataframe. However, it is only a list of neighborhood names and postal codes.

:

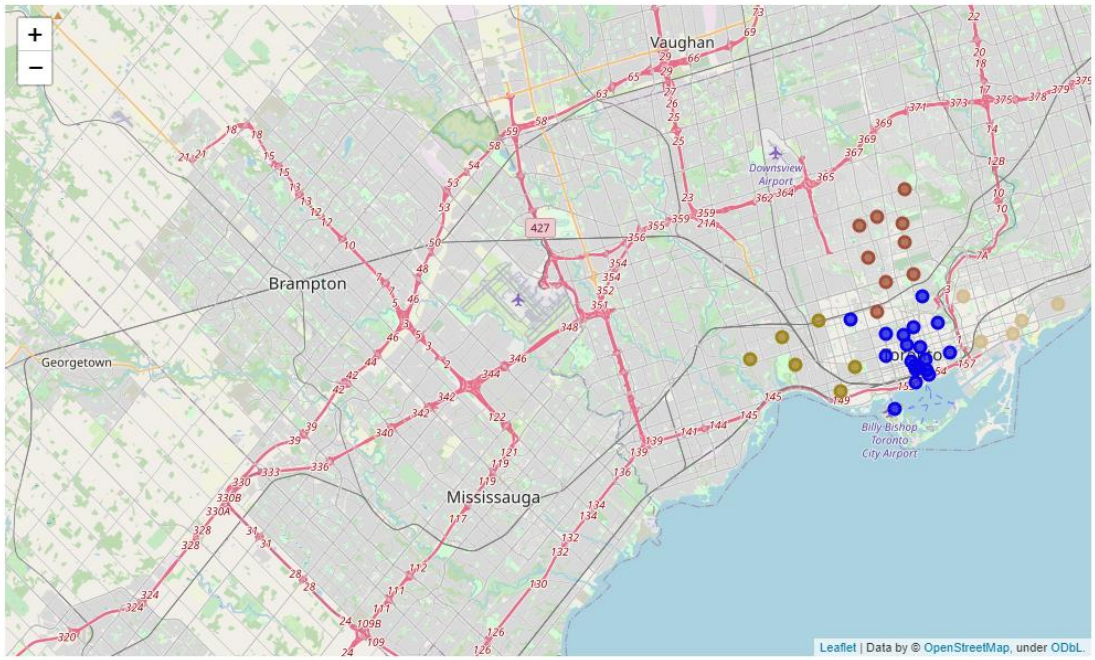
	Postal code	Borough	Neighborhood
0	M1B	Scarborough	Malvern / Rouge
1	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek
2	M1E	Scarborough	Guildwood / Morningside / West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

I will need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I tried using Geocoder package but it was not working so I used the csv file provided by IBM team to match the coordinates of Toronto neighborhoods. After gathering all these coordinates, I visualized the map of Toronto using Folium package to verify whether these are correct coordinates.

```
        fill_opacity=0.7).add_to(map_toronto)

map_toronto
```

Out[56]:



Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later. Here, I made a justification to specifically look for “Italian restaurants”

Out[73]:

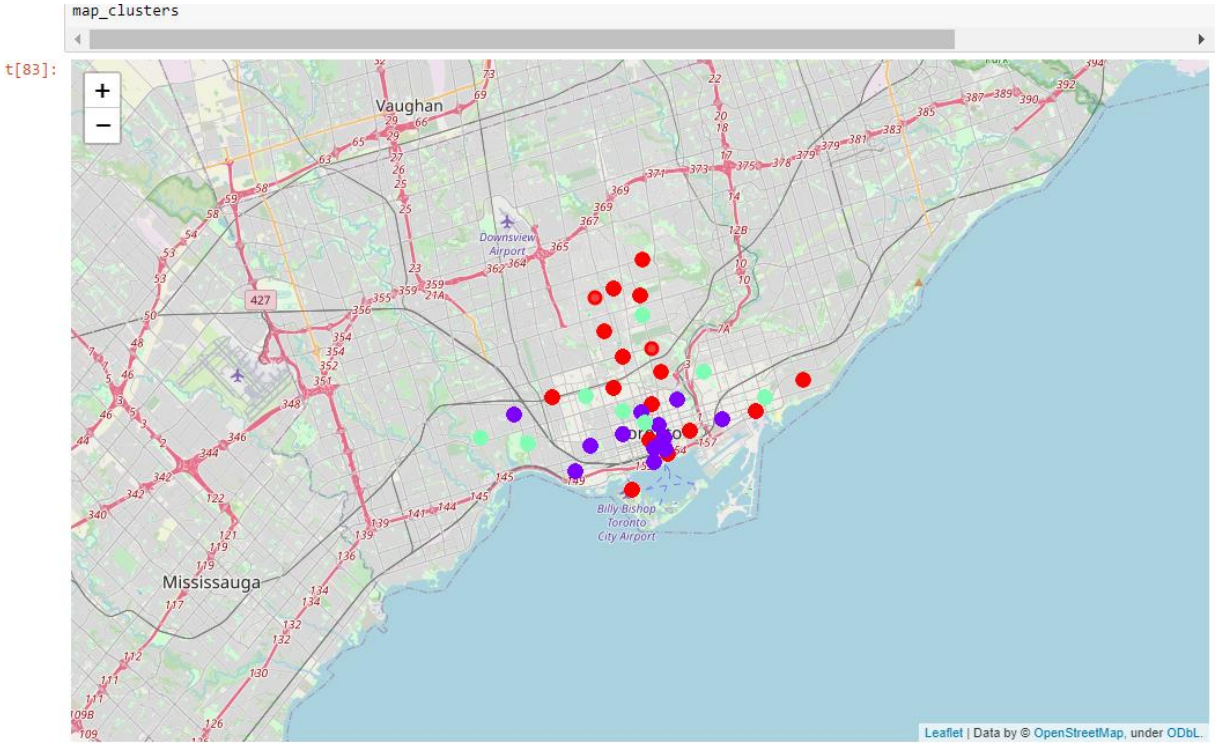
	Neighborhoods	Italian Restaurant
0	Berczy Park	0.000000
1	Brockton / Parkdale Village / Exhibition Place	0.043478
2	Business reply mail Processing CentrE	0.000000
3	CN Tower / King and Spadina / Railway Lands / ...	0.000000
4	Central Bay Street	0.079365

Cluster Neighborhoods

I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centeriods, 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 it is highly suited for this project as well. I have clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for “Italian food”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

Results

Clusters



The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Thai restaurants are in each neighborhood:

- Cluster 0: Neighborhoods with little or no Italian restaurants
- Cluster 1: Neighborhoods high number of Italian restaurants
- Cluster 2: Neighborhoods with noItalian restaurants

The results are visualized in the above map with Cluster 0 in red color, Cluster 1 in purple color and Cluster 2 in light green color

Recommendations :

Most of Italian restaurants are in Cluster 1 which is around Adelaide Toronto ,

it seems Cluster 2 might be a good location as there are not a lot of Italian restaurants in these areas. Therefore, this project recommends the entrepreneur to open an authentic Italian restaurant in these locations with little to no competition. Nonetheless, if the food is authentic, affordable and good taste, I am confident that it will have great following everywhere