Exploring neighborhoods in Toronto to start a restaurant

Web-Scraping, Foursquare API, Folium Map & Machine Learning

Author
VinodKumar Reddy Hebbatam

Preface:

As a part of the IBM Data Science professional program Capstone Project, we worked on the real datasets to get an experience of what a data scientist goes through in real life. Main objectives of this project were to define a business problem, look for data in the web and, use Foursquare location data to compare different neighborhoods of Toronto to figure out which neighborhood is suitable for starting a new restaurant business. In this project, we will go through all the process in a step by step manner from problem designing, data preparation to final analysis and finally will provide a conclusion that can be leveraged by the business stakeholders to make their decisions.

1. Introduction / Business Problem

My client company wants to start a restaurant business and is trying to decide which area in Toronto could be a right spot to do that. An area should have the biggest potentiality in developing restaurant business. Location should help company to reach as wide customers auditory as possible to make business profitable. As company does not possess a knowledge of Toronto districts and neighborhoods, it decided to hire a business consultant company which provides various consultant services for companies which want to establish new business in Toronto area. A consultant company utilizing Data Science methods is going to cluster Toronto city area and provide recommendations where is the best place to start a restaurant business.
As there are a lot of restaurants in Toronto already, we will take into consideration every district's population and its density too. We are very interested in districts located in Toronto Downtown as tourists can generate additional revenues too.

Target Audience

Who will be more interested in this project? What type of clients or a group of people would be benefitted?

1. Business personnel who wants to invest or open a restaurant in Toronto. This analysis will be a comprehensive guide to start or expand restaurants targeting the crowd.

2. Freelancer who loves to have their own restaurant as a side business. This analysis will give an idea, how beneficial it is to open a restaurant and what are the pros and cons of this business.

3. Crowd who wants to find neighborhoods with lots of option for restaurants.

4. Business Analyst or Data Scientists, who wish to analyze the neighborhoods of Toronto using Exploratory Data Analysis and other statistical & machine learning techniques to obtain all the necessary data, perform some operations on it and, finally be able to tell a story out of it.

2. Data Gathering and Preprocessing

2.1 Data Sources

a) I'm using "List of Postal code of Canada: M"

(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) wiki page to get all the information

about the neighborhoods present in Toronto. This page has the postal code, borough & the name of all the neighborhoods present in Toronto.

b) Then I'm using "https://cocl.us/Geospatial_data" csv file to get all the geographical coordinates of the neighborhoods.

c) To get information about the distribution of population by their ethnicity I'm using "Demographics of Toronto" (https://en.m.wikipedia.org/wiki/Demographics_of_Toronto#Ethnic_diversity) wiki page. Using this page I'm going to identify the neighborhoods which are densely populated as it might be helpful in identifying the suitable neighborhood to open a new restaurant.

d) To get location and other information about various venues in Toronto I'm using Foursquare's explore API. Using the Foursquare's explore API (which gives venues recommendations), I'm fetching details about the venues up present in Toronto and collected their names, categories and locations (latitude and longitude).

From Foursquare API (https://developer.foursquare.com/docs), I retrieved the following for each venue:

- Name: The name of the venue.
- Category: The category type as defined by the API.
- Latitude: The latitude value of the venue.
- Longitude: The longitude value of the venue.

2.2 Data Cleaning

a) Scraping Toronto Neighborhoods Table from Wikipedia

Scraped the following Wikipedia page, "List of Postal code of Canada: M" in order to obtain the data about the Toronto & the Neighborhoods in it.

Assumptions made to attain the below DataFrame:

- Dataframe will consist of three columns: PostalCode, Borough, and Neighborhood
- Only the cells that have an assigned borough will be processed. Borough that is not assigned are ignored.
- More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in row 11 in the above table.
- If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

Wikipedia — package is used to scrape the data from wiki.

Scrap data from Wiki page, download and Explore Dataset

```
source = requests.get("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M").text
soup = BeautifulSoup(source, 'lxml')

table = soup.find("table")
table_rows = table.tbody.find_all("tr")

table_contents=[]
table=soup.find('table')
for row in table.findAll('td'):
    cell = {}
    if row.span.text=='Not assigned':
        pass
    else:
        cell['PostalCode'] = row.p.text[:3]
        cell['Borough'] = (row.span.text).split('(')[0]
        cell['Neighborhood'] = (((((row.span.text).split('(')[1]).strip(')')).replace(' /',',')).replace(')',' ')).strip(
        table_contents.append(cell)

# print(table_contents)
df=pd.DataFrame(table_contents)
df['Borough']=df['Borough'].replace({'Downtown TorontoStn A PO Boxes25 The Esplanade':'Downtown Toronto Stn A',
                                     'East TorontoBusiness reply mail Processing Centre969 Eastern':'East Toronto
                                     'EtobicokeNorthwest':'Etobicoke Northwest','East YorkEast Toronto':'East Yor
                                     'MississaugaCanada Post Gateway Processing Centre':'Mississauga'})
df.head()
```

Clean the data and group all neighborhood

Group all neighborhoods with the same postal code

```
df = df.groupby(["PostalCode", "Borough"])["Neighborhood"].apply(", ".join).reset_index()
df.head()
```

1]:

	PostalCode	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

```
print("Shape: ", df.shape)
```

```
Shape:  (103, 3)
```

b) Adding geographical coordinates to the neighborhoods

Next important step is adding the geographical coordinates to these neighborhoods. To do so I'm extracting the data present in the Geospatial Data csv file and I'm combining it with the existing neighborhood dataframe by merging them both based on the postal code.

We are downloading geospacial data from the link https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv

```
df_geo_coor = pd.read_csv("C:\\Users\\Lenovo\\Downloads\\Geospatial_Coordinates.csv")
df_geo_coor.head()
```

]:

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

We need to couple 2 dataframes "df" and "df_geo_coor" into one dataframe.

DataFrame with latitude & longitude of Postal codes in Toronto

I'm renaming the columns to match the existing dataframe formed from 'List of Postal code of Canada: M' wiki page.After that I'm merging both the dataframe into one by merging on the postal code.

We need to couple 2 dataframes "df" and "df_geo_coor" into one dataframe.

```
df_toronto = pd.merge(df, df_geo_coor, how='left', left_on = 'PostalCode', right_on = 'Postal Code')
# remove the "Postal Code" column
df_toronto.drop("Postal Code", axis=1, inplace=True)
df_toronto.head()
```

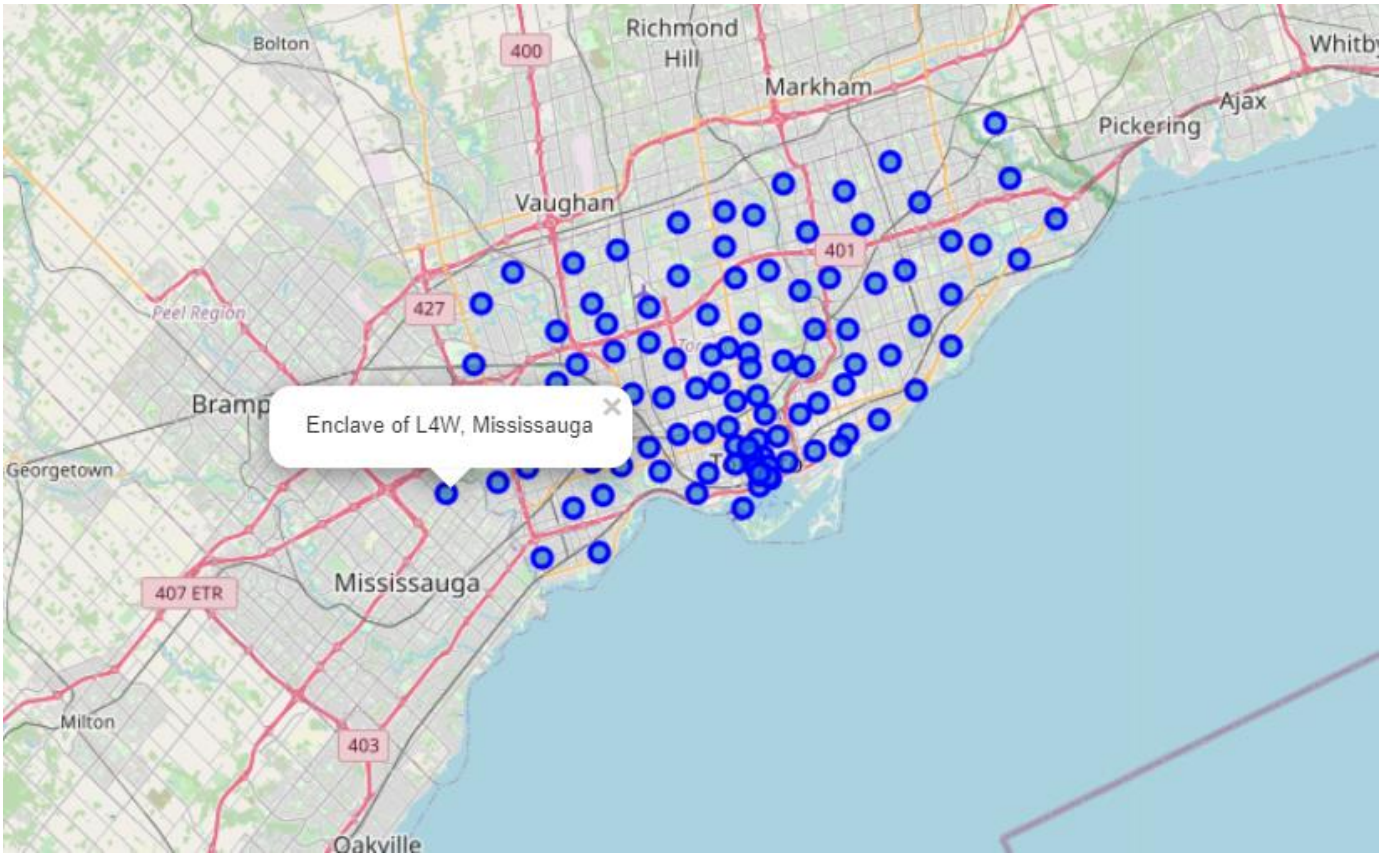
4]:

	PostalCode	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

Merged new dataframe with info about Neighborhoods, borough, postalcode, latitude & longitude in Toronto

3. Explore Neighborhoods in Toronto City

While exploring the neighborhoods we need to the latitude and longitude value of Toronto. Create a map of whole city with neighbors superimposed on the top, we will also add markers to the map



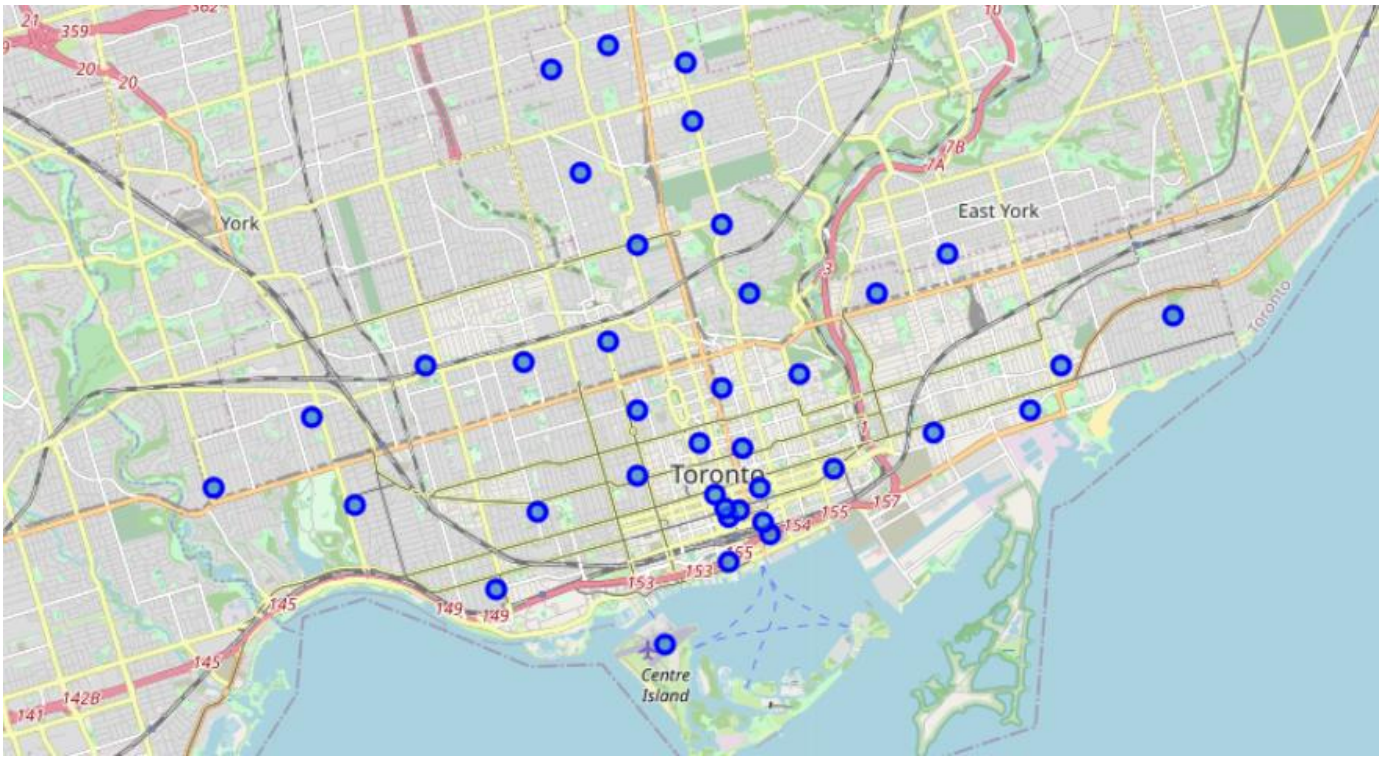
Further We are going to work with only the boroughs that contain the word "Toronto".

```
# "denc" = [D]owntown Toronto, [E]ast Toronto, [N]orth Toronto, [C]entral Toronto
df_toronto_denc = df_toronto[df_toronto['Borough'].str.contains("Toronto")].reset_index(drop=True)
df_toronto_denc.head()
```

:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4J	East York/East Toronto	The Danforth East	43.685347	-79.338106
2	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
3	M4L	East Toronto	India Bazaar, The Beaches West	43.668999	-79.315572
4	M4M	East Toronto	Studio District	43.659526	-79.340923

We will have a look at the map again to check the boroughs



Now we will make us of foursquare API's to group the data so that we can use it to create the clusters later. We will explore the first neighborhood and later we will explore all the neighborhoods.

```
toronto_denc_venues.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood 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 East	43.685347	-79.338106	Aldwych Park	43.684901	-79.341091	Park

```
toronto_denc_venues.groupby('Neighborhood').count()
```

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Berczy Park	57	57	57	57	57	57
Brockton, Parkdale Village, Exhibition Place	23	23	23	23	23	23
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	15	15	15	15	15	15
Central Bay Street	62	62	62	62	62	62
Christie	16	16	16	16	16	16

Lets find out how many unique categories can be curated from all the returned values

```
print('There are {} uniques categories.'.format(len(toronto_denc_venues['Venue Category'].unique())))
```

```
There are 236 uniques categories.
```

Since we have only analyzed one neighborhood, lets analyze all the other neighborhoods now.

3.7. Analyze Each Neighborhood

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

# add neighborhood column back to dataframe
toronto_denc_onehot['Neighborhood'] = toronto_denc_venues['Neighborhood']

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

toronto_denc_onehot.head()
```

3]:

	Yoga Studio	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	...	Tibetan Restaurant	Toy / Game Store	Trail	Train Station	Vegetarian / Vegan Restaurant	Video Game Store
0	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0

5 rows × 236 columns

Next, let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category, after this we will check the 10 most common venues in each neighborhood

	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	Seafood Restaurant	Restaurant	Cheese Shop	Pharmacy	Farmers Market	Bakery	Beer Bar	Belgian Restaurant
1	Brockton, Parkdale Village, Exhibition Place	Café	Breakfast Spot	Bakery	Coffee Shop	Furniture / Home Store	Burrito Place	Restaurant	Italian Restaurant	Stadium	Bar
2	CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Lounge	Boat or Ferry	Plane	Sculpture Garden	Harbor / Marina	Airport Terminal	Rental Car Location	Airport Gate	Airport Food Court
3	Central Bay Street	Coffee Shop	Sandwich Place	Italian Restaurant	Café	Burger Joint	Salad Place	Bubble Tea Shop	Poke Place	Portuguese Restaurant	Pizza Place
4	Christie	Grocery Store	Café	Park	Nightclub	Italian Restaurant	Baby Store	Restaurant	Candy Store	Athletics & Sports	Coffee Shop

4. Predictive Modelling:

4.1. Clustering Neighborhoods of Toronto:

First step in K-means clustering is to identify best K value meaning the number of clusters in a given dataset. To do so we are going to use the elbow method on the Toronto dataset with restaurant percentage.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	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
0	M4E	East Toronto	The Beaches	43.676357	-79.293031	1	Pub	Health Food Store	Trail	Women's Store	Discount Store	Distribution Center	Dog Run
1	M4J	East York/East Toronto	The Danforth East	43.685347	-79.338106	0	Park	Convenience Store	Metro Station	Women's Store	Diner	Ethiopian Restaurant	Escape Room
2	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188	1	Greek Restaurant	Coffee Shop	Italian Restaurant	Ice Cream Shop	Bookstore	Furniture / Home Store	Fruit & Vegetable Store
3	M4L	East Toronto	India Bazaar, The Beaches West	43.668999	-79.315572	1	Sandwich Place	Fast Food Restaurant	Ice Cream Shop	Movie Theater	Brewery	Restaurant	Italian Restaurant
4	M4M	East Toronto	Studio District	43.659526	-79.340923	1	Coffee Shop	American Restaurant	Bakery	Brewery	Café	Gastropub	Gym / Fitness Center

4.2 Examine the Clusters:

We have total of 5 clusters such as 1,2,3,4,5. Let us examine one after the other.

Cluster 1 displays the borough where resutaurant are located

Cluster 1 ¶

```
toronto_denc_merged.loc[toronto_denc_merged['Cluster Labels'] == 0, toronto_denc_merged.columns[[1] + list(range(5, toronto_d
```

)]:

	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	0	Park	Convenience Store	Metro Station	Women's Store	Diner	Ethiopian Restaurant	Escape Room	Electronics Store	Eastern European Restaurant	Dumpling Restaurant
11	Downtown Toronto	0	Park	Playground	Trail	Women's Store	Dessert Shop	Ethiopian Restaurant	Escape Room	Electronics Store	Eastern European Restaurant	Dumpling Restaurant

Cluster 2 got 33 borough's in the downtown and other locations and it looks ideal to open a restaurant

Cluster 2

```
toronto_denc_merged.loc[toronto_denc_merged['Cluster Labels'] == 1, toronto_denc_merged.columns[[1] + list(range(5, toronto_d
```

'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
0	East Toronto	1	Pub	Health Food Store	Trail	Women's Store	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Eastern European Restaurant
2	East Toronto	1	Greek Restaurant	Coffee Shop	Italian Restaurant	Ice Cream Shop	Bookstore	Furniture / Home Store	Fruit & Vegetable Store	Pizza Place	Brewery	Bubble Tea Shop
3	East Toronto	1	Sandwich Place	Fast Food Restaurant	Ice Cream Shop	Movie Theater	Brewery	Restaurant	Italian Restaurant	Fish & Chips Shop	Steakhouse	Sushi Restaurant
4	East Toronto	1	Coffee Shop	American Restaurant	Bakery	Brewery	Café	Gastropub	Gym / Fitness Center	Diner	Park	Middle Eastern Restaurant
6	Central Toronto	1	Gym / Fitness Center	Breakfast Spot	Hotel	Food & Drink Shop	Department Store	Park	Sandwich Place	Gym	Airport Terminal	Distribution Center
7	Central Toronto	1	Coffee Shop	Clothing Store	Yoga Studio	Rental Car Location	Spa	Shoe Store	Seafood Restaurant	Salon / Barbershop	Restaurant	Chinese Restaurant
8	Central Toronto	1	Sandwich Place	Dessert Shop	Pizza Place	Sushi Restaurant	Coffee Shop	Gym	Italian Restaurant	Thai Restaurant	Café	Indian Restaurant

Cluster 3, 4 and 5 got only one borough's and does looks good to open restaurant in these clusters

Cluster 3

```
toronto_denc_merged.loc[toronto_denc_merged['Cluster Labels'] == 2, toronto_denc_merged.columns[[1] + list(range(5, toronto_d
```

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
9	Central Toronto	2	Tennis Court	Women's Store	Diner	Falafel Restaurant	Event Space	Ethiopian Restaurant	Escape Room	Electronics Store	Eastern European Restaurant	Dumpling Restaurant

Cluster 4

```
toronto_denc_merged.loc[toronto_denc_merged['Cluster Labels'] == 3, toronto_denc_merged.columns[[1] + list(range(5, toronto_d
```

73]:

	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	3	Home Service	Garden	Women's Store	Dessert Shop	Event Space	Ethiopian Restaurant	Escape Room	Electronics Store	Eastern European Restaurant	Dumpling Restaurant

Cluster 5

```
toronto_denc_merged.loc[toronto_denc_merged['Cluster Labels'] == 4, toronto_denc_merged.columns[[1] + list(range(5, toronto_d
```

'4]:

	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
5	Central Toronto	4	Bus Line	Park	Swim School	Women's Store	Ethiopian Restaurant	Escape Room	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop

5. Results and Discussion:

5.1 Results

We have reached the end of the analysis, in this section we will document all the findings from above clustering & visualization of the dataset. In this project, we started off with the business problem of identifying a good neighborhood to open a new restaurant. To achieve that we looked into all the neighborhoods in Toronto, analyzed the population in each neighborhood & number of restaurants in those neighborhoods to come to conclusion about which neighborhood would be a better spot. We have used variety of data sources to set up a very realistic data-analysis scenario. We have found out that —

- In those 38 boroughs we identified that only Central Toronto, Downtown Toronto, East Toronto, East York, North York & Scarborough boroughs have high amount of restaurants with the help of Violin plots between Number of restaurants in Borough of Toronto.

- In all the ridings, Scarborough-Guildwood, Scarborough-Rouge Park, Scarborough Centre, Scarborough North, Humber River-Black Creek, Don Valley East, Scarborough Southwest, Don Valley North & Scarborough-Agincourt are the densely populated with crowd ridings.

- With the help of clusters examining & violin plots looks like Downtown Toronto, Central Toronto, East York are already densely populated with restaurants. So it is better idea to leave those boroughs out and consider only Scarborough, East Toronto & North York for the new restaurant's location.

- After careful consideration it is a good idea to open a new restaurant in Scarborough borough since it has high number of population which gives a higher number of customers possibility and lower competition since very less restaurants in the neighborhoods.

5.2 Discussion

According to this analysis, Scarborough borough will provide the least competition for the new upcoming restaurant as there is very less restaurants spread in the neighborhoods. Also looking at the population distribution looks like it is densely populated with crowd which helps the new restaurant by providing high customer visit possibility. So, definitely this region could potentially be a perfect place for starting a quality restaurants. Some of

the drawbacks of this analysis are — the clustering is completely based only on data obtained from Foursquare API and the data about the population distribution in each neighborhood is also based on the 2018 census which is not up-to date. Thus there is huge gap of 3 years in the population distribution data. Even Though there are lots of areas where it can be improved yet this analysis has certainly provided us with some good insights, preliminary information on possibilities & a head start into this business problem by setting the step stones properly.

6. Conclusion:

Finally to conclude this project, We have got a chance to on a business problem like how a real like data scientists would do. We have used many python libraries to fetch the data , to manipulate the contents & to analyze and visualize those datasets. We have made use of Foursquare API to explore the venues in neighborhoods of Toronto, then get good amount of data from Wikipedia which we scraped with help of Wikipedia python library and visualized using various plots present in seaborn & matplotlib. We also applied machine learning technique to predict the output given the data and used Folium to visualize it on a map.

Some of the drawbacks or areas of improvements shows us that this analysis can be further improved with the help of more data and different machine learning technique. Similarly we can use this project to analysis any scenario such as opening a different cuisine restaurant or opening of a new gym and etc. Hopefully, this project helps acts as initial guidance to take more complex real-life challenges using data-science.