IBM Data Science Capstone Project

Opening a Multi-cuisine Restaurant in Toronto

Submitted By: Zafreena Althaf

20th February 2020

Table of Contents

1.Introduction

Toronto is the provincial capital of Ontario and the most populous city in Canada. The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada. More than 50 percent of residents belong to a visible minority population group, and over 200 distinct ethnic origins are represented among its inhabitants. While the majority of Torontonians speak English as their primary language, over 160 languages are spoken in the city.

Such a diverse city would be the best place to open a Multi-cuisine restaurant.

1.1 Business Problem

The objective of this capstone project is to analyze and recommend the best neighborhoods in the city of Toronto to open a new Multi-cuisine restaurant serving different continental foods. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: Where would you recommend a new investor to open a Multi-cuisine restaurant in the city of Toronto?

In this project we would be mainly looking at the below:

- Areas for large scope of shopping- Because common most people shop at least once every week and that is usually the same day they decide to dine out!
- Areas having restaurant venues- So that we can perform clustering on them and accordingly check for the concentration of restaurants in these areas.

1.2 Target Audience

This project is useful for any investors who are willing to open a new Multi-cuisine restaurant in the city of Toronto. All potential investors in the food industry, particularly in Toronto can draw insights from this project. These insights will give a better understanding of the business environment and help reduce risk and reap generous returns.

2. Data acquisition and cleaning

Here we will see what data we need, what sources we got the data from and how data was extracted and cleaned and made ready for analysis**.**

2.1 Data Required

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

- List of neighborhoods in Toronto.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to areas with restaurants and shopping places.

2.2 Data sources

The Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) contains a list of neighborhoods in and around Toronto, along with its corresponding boroughs and postal codes. We will use the 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 https://cocl.us/Geospatial_data csv file 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 has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Restaurants and shops category in order to help us to solve the business problem put forward.

2.3 Data extraction and cleaning

Web scraping of wiki page is done using the BeautifulSoup package and dataframe of three columns: PostalCode, Borough, and Neighborhood is created.

We then only process the cells that have an assigned borough. Ignore cells with a borough that is 'Not assigned**'.**

Since more than one neighborhood can exist in one postal code area we do a groupby of Postal code and Borough and group neighborhoods in same borough.

If a neighborhood is 'Not assigned' neighborhood, then the neighborhood will be the same as the borough.

We will then get the geographical coordinates of the neighborhoods using https://cocl.us/Geospatial_data csv file which will give us the latitude and longitude coordinates of the neighborhoods. Which we will then merge with our original dataframe based on postalcode and now our data is ready!

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Heights,Lawrence Manor	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494

Additionally we can get our venue data using Foursquare API.

Venues were returned for each PostalCode

```
venues_df.groupby(["PostalCode", "Borough", "Neighborhood"]).count()
```

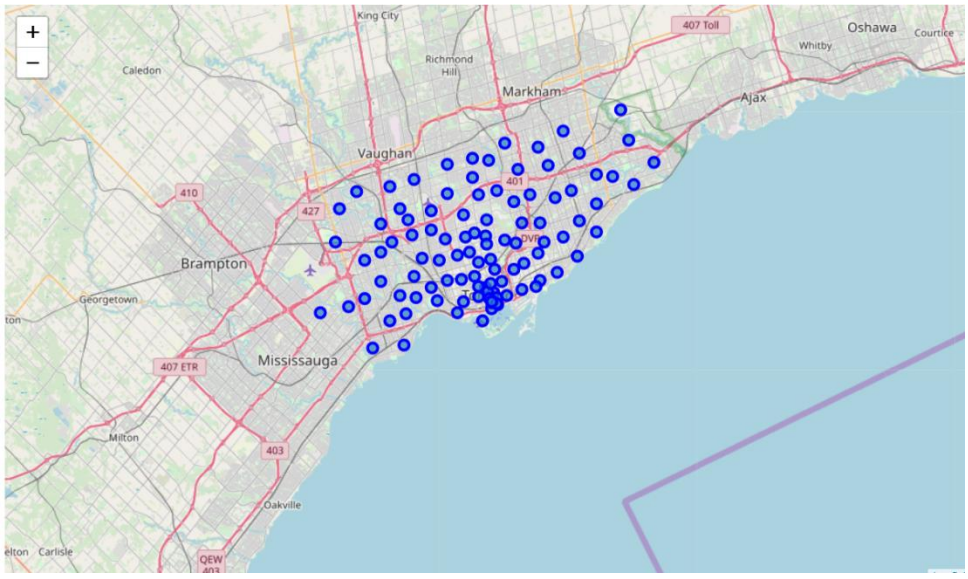
PostalCode	Borough	Neighborhood	BoroughLatitude	BoroughLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
M4E	East Toronto	The Beaches	7	7	7	7	7	7
M4K	East Toronto	The Danforth West,Riverdale	41	41	41	41	41	41
M4L	East Toronto	The Beaches West,India Bazaar	19	19	19	19	19	19
M4M	East Toronto	Studio District	42	42	42	42	42	42
M4N	Central Toronto	Lawrence Park	3	3	3	3	3	3
M4P	Central	Davisville North						

First the data related to shopping venues in each neighborhood is extracted using foursquare to check the neighborhoods with large scope for shopping.

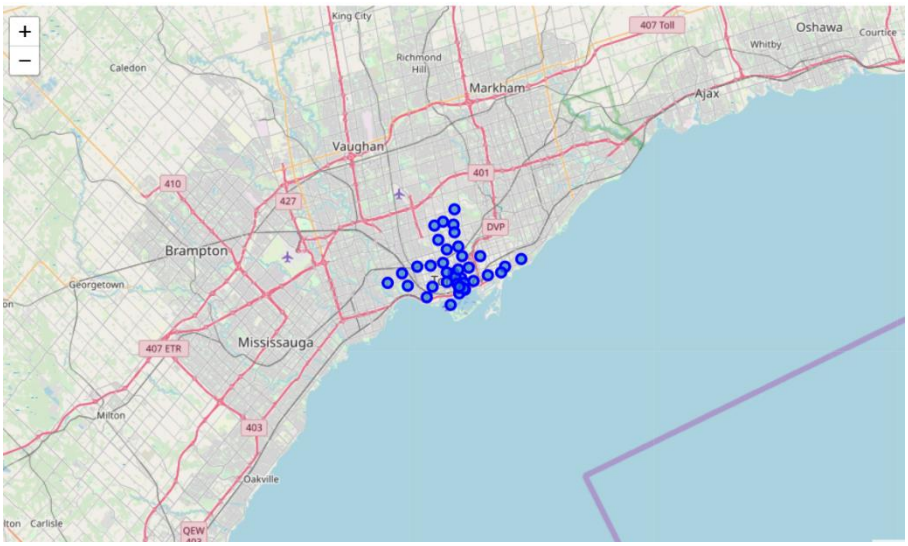
Next, we analyzed each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. Since we are analyzing the "Restaurants" data, we will filter the "Restaurants" as venue category for the neighbourhoods. On this we will run kmean cluster to get restaurant concentration in each neighborhood.

3. Methodology

Now that we have our data containing the neighborhoods and the coordinates ready we can plot the Map of Toronto with the neighborhoods superimposed on top.



In this project we are looking mainly at boroughs having 'Toronto' and we use 'folium' to visualize our maps.



Using Foursquare API, captured a list of top 100 venues that are within a radius of 500 meters. To do so, I first registered a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. Then added made API calls to Foursquare passing in the geographical coordinates of the neighbourhoods and the Foursquare credentials in a Python loop.

Foursquare returned the venue data in JSON format and extracted the venue name, venue category, venue latitude and longitude. With the data, checked how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues.

Number of unique categories that can be curated from all the returned venues

```
print('There are {} uniques categories.'.format(len(venues_df['VenueCategory'].unique())))
```

```
There are 234 uniques categories.
```

```
venues_df.head()
```

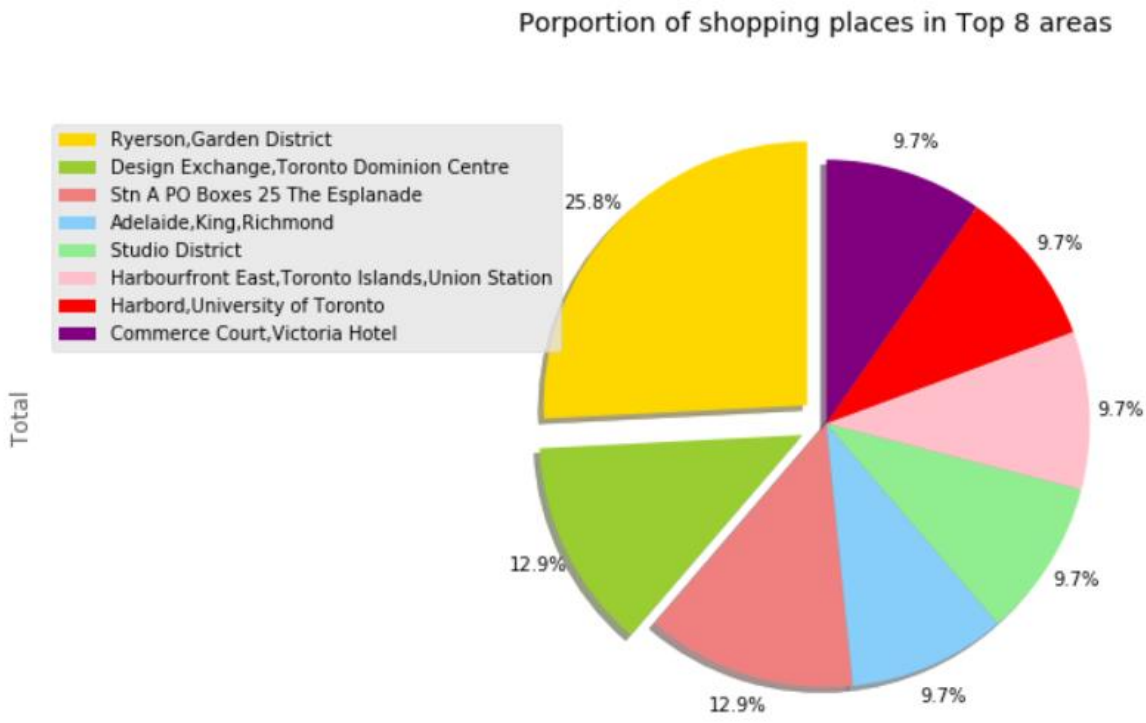
	PostalCode	Borough	Neighborhood	BoroughLatitude	BoroughLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	M5A	Downtown Toronto	Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	M5A	Downtown Toronto	Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	M5A	Downtown Toronto	Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653191	-79.357947	Gym / Fitness Center
3	M5A	Downtown Toronto	Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	M5A	Downtown Toronto	Harbourfront	43.65426	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot

First the data related to shopping venues in each neighborhood was extracted and we plotted a pie chart of the top 8 neighborhoods with highest number of shopping places.

Analyzing neighbourhoods with largest number of shopping areas

```
toronto_shopping = toronto_onehotEn[['Neighborhoods','Antique Shop','Arts & Crafts Store','Bookstore','Comic Shop', 'Boutique', 'Mar
toronto_shopping_group = toronto_shopping.groupby('Neighborhoods').sum(ascending=False)
toronto_shopping_group['Total'] =  toronto_shopping_group.sum (axis = 1)
toronto_shopping_group.sort_values("Total",ascending=False,axis=0,inplace=True)
toronto_shopping_group.head(5)
```

Neighborhoods	Antique Shop	Arts & Crafts Store	Bookstore	Comic Shop	Boutique	Market	Record Shop	Shoe Store	Shopping Mall	Sporting Goods Shop	Stationery Store	Supermarket	Thrift / Vintage Store	Toy Gam Stor
Ryerson,Garden District	0	0	2	1	0	0	0	1	1	1	0	0	0	
Design Exchange,Toronto Dominion Centre	0	0	1	0	0	0	0	0	1	2	0	0	0	
Stn A PO Boxes 25 The Esplanade	1	0	1	0	0	0	0	0	1	1	0	0	0	
Adelaide,King,Richmond	0	0	1	0	0	0	0	0	1	0	0	0	0	
Studio District	0	0	1	0	0	0	0	0	0	0	1	0	1	



Next, we analyzed each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. Since we are analyzing the "Restaurants" data, we will filter the "Restaurants" as venue category for the neighbourhoods.

Group rows by neighborhood and by taking the mean of the frequency of occurrence of each category

```
group_df = toronto_onehotEn.groupby(["PostalCode", "Borough", "Neighborhoods"]).mean().reset_index()

print(group_df.shape)
group_df
```

```
(39, 237)
```

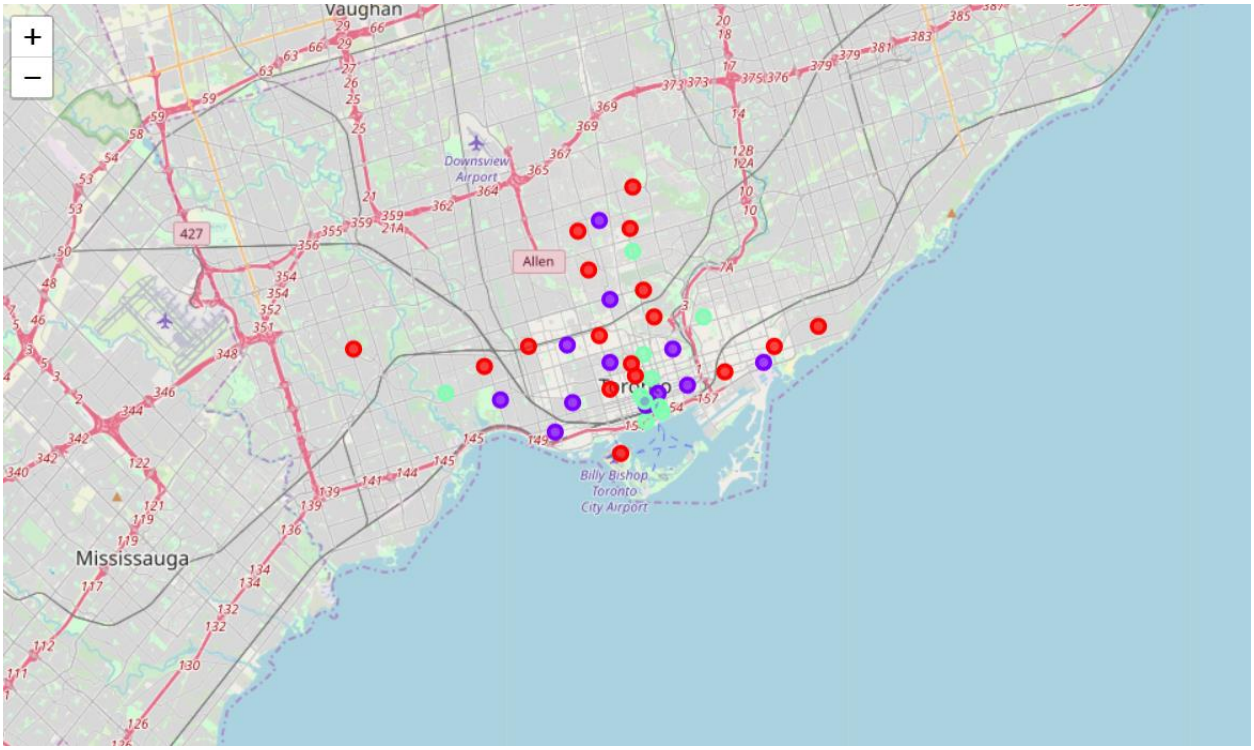
	PostalCode	Borough	Neighborhoods	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Toy / Game Store	Trail	Trai Statio
0	M4E	East Toronto	The Beaches	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.142857	0.0
1	M4K	East Toronto	The Danforth West,Riverdale	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.024390	...	0.000000	0.024390	0.0
2	M4L	East Toronto	The Beaches West,India Bazaar	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0
3	M4M	East Toronto	Studio District	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.047619	...	0.000000	0.000000	0.0
4	M4N	Central												

Lastly, we performed machine learning technique of 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 cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for "Restaurants". The results will allow us to identify the concentration of restaurants in different neighborhoods. The below image shows the neighborhoods divided into 3 clusters (0, 1 & 2)

	Neighborhood	Restaurant	Cluster Labels	PostalCode	Borough	Latitude	Longitude
0	The Beaches	0.000000	0	M4E	East Toronto	43.676357	-79.293031
1	The Danforth West,Riverdale	0.024390	2	M4K	East Toronto	43.679557	-79.352188
2	The Beaches West,India Bazaar	0.000000	0	M4L	East Toronto	43.668999	-79.315572
3	Studio District	0.000000	0	M4M	East Toronto	43.659526	-79.340923
4	Lawrence Park	0.000000	0	M4N	Central Toronto	43.728020	-79.388790
5	Davisville North	0.000000	0	M4P	Central Toronto	43.712751	-79.390197
6	North Toronto West	0.050000	1	M4R	Central Toronto	43.715383	-79.405678
7	Davisville	0.026316	2	M4S	Central Toronto	43.704324	-79.388790
8	Moore Park,Summerhill East	0.000000	0	M4T	Central Toronto	43.689574	-79.383160
9	Deer Park,Forest Hill SE,Rathnelly,South Hill	0.066667	1	M4V	Central Toronto	43.686412	-79.400049

The below map represents the 3 clusters in REDD, PURPLE & GREEN.



4. Results

The results from the k-means clustering shows that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence of Restaurants:

Cluster 0: Neighbourhoods with low concentration of Restaurants

Cluster 0

```
k_merged.loc[k_merged['Cluster Labels'] == 0]
```

	Neighborhood	Restaurant	Cluster Labels	PostalCode	Borough	Latitude	Longitude
0	The Beaches	0.0	0	M4E	East Toronto	43.676357	-79.293031
2	The Beaches West,India Bazaar	0.0	0	M4L	East Toronto	43.668999	-79.315572
3	Studio District	0.0	0	M4M	East Toronto	43.659526	-79.340923
4	Lawrence Park	0.0	0	M4N	Central Toronto	43.728020	-79.388790
5	Davisville North	0.0	0	M4P	Central Toronto	43.712751	-79.390197
8	Moore Park,Summerhill East	0.0	0	M4T	Central Toronto	43.689574	-79.383160
10	Rosedale	0.0	0	M4W	Downtown Toronto	43.679563	-79.377529
17	Central Bay Street	0.0	0	M5G	Downtown Toronto	43.657952	-79.387383
22	Roselawn	0.0	0	M5N	Central Toronto	43.711695	-79.416936
23	Forest Hill North,Forest Hill West	0.0	0	M5P	Central Toronto	43.696948	-79.411307

Cluster 1: Neighbourhoods with High concentration of Restaurants

Cluster 1

```
k_merged.loc[k_merged['Cluster Labels'] == 1]
```

	Neighborhood	Restaurant	Cluster Labels	PostalCode	Borough	Latitude	Longitude
6	North Toronto West	0.050000	1	M4R	Central Toronto	43.715383	-79.405678
9	Deer Park,Forest Hill SE,Rathnelly,South Hill,...	0.066667	1	M4V	Central Toronto	43.686412	-79.400049
11	Cabbagetown,St. James Town	0.066667	1	M4X	Downtown Toronto	43.667967	-79.367675
13	Harbourfront	0.042553	1	M5A	Downtown Toronto	43.654260	-79.360636
15	St. James Town	0.050000	1	M5C	Downtown Toronto	43.651494	-79.375418
20	Design Exchange,Toronto Dominion Centre	0.050000	1	M5K	Downtown Toronto	43.647177	-79.381576
21	Commerce Court,Victoria Hotel	0.050000	1	M5L	Downtown Toronto	43.648198	-79.379817
25	Harbord,University of Toronto	0.057143	1	M5S	Downtown Toronto	43.662696	-79.400049
30	Christie	0.055556	1	M6G	Downtown Toronto	43.669542	-79.422564
32	Little Portugal,Trinity	0.056604	1	M6J	West Toronto	43.647927	-79.419750
33	Brockton,Exhibition Place,Parkdale Village	0.043478	1	M6K	West Toronto	43.636847	-79.428191
35	Parkdale,Roncesvalles	0.071429	1	M6R	West Toronto	43.648960	-79.456325
38	Business Reply Mail Processing Centre 969 Eastern	0.055556	1	M7Y	East Toronto	43.662744	-79.321558

Cluster 2: Neighbourhoods with medium concentration of Restaurants

Cluster 2

```
k_merged.loc[k_merged['Cluster Labels'] == 2]
```

	Neighborhood	Restaurant	Cluster Labels	PostalCode	Borough	Latitude	Longitude
1	The Danforth West,Riverdale	0.024390	2	M4K	East Toronto	43.679557	-79.352188
7	Davisville	0.026316	2	M4S	Central Toronto	43.704324	-79.388790
12	Church and Wellesley	0.037037	2	M4Y	Downtown Toronto	43.665860	-79.383160
14	Ryerson,Garden District	0.020000	2	M5B	Downtown Toronto	43.657162	-79.378937
16	Berczy Park	0.017857	2	M5E	Downtown Toronto	43.644771	-79.373306
18	Adelaide,King,Richmond	0.030000	2	M5H	Downtown Toronto	43.650571	-79.384568
19	Harbourfront East,Toronto Islands,Union Station	0.030000	2	M5J	Downtown Toronto	43.640816	-79.381752
28	Stn A PO Boxes 25 The Esplanade	0.031915	2	M5W	Downtown Toronto	43.646435	-79.374846
29	First Canadian Place,Underground city	0.040000	2	M5X	Downtown Toronto	43.648429	-79.382280
36	Runnymede,Swansea	0.025000	2	M6S	West Toronto	43.651571	-79.484450

Also according to our pie chart the neighborhoods: (Ryerson,Garden District) , (Adelaide,King,Richmond) and (Harbourfront East,Toronto Islands,Union Station) all come under the top 8 largest shopping neighborhoods and these three also fall under cluster 2)
The neighborhoods: (Design Exchange,Toronto Dominion Centre),(Commerce Court,Victoria Hotel) and (Harbord,University of Toronto) are also under the top 8 largest shopping neighborhoods and belong to cluster 1

Neighborhoods	Antique Shop	Arts & Crafts Store	Bookstore	Comic Shop	Boutique	Market	Record Shop	Shoe Store	Shopping Mall	Sporting Goods Shop	Stationery Store	Supermarket	Thrift / Vintage Store	Toy Gam Stor
Ryerson,Garden District	0	0	2	1	0	0	0	1	1	1	0	0	0	
Design Exchange,Toronto Dominion Centre	0	0	1	0	0	0	0	0	1	2	0	0	0	
Stn A PO Boxes 25 The Esplanade	1	0	1	0	0	0	0	0	1	1	0	0	0	
Adelaide,King,Richmond	0	0	1	0	0	0	0	0	1	0	0	0	0	
Studio District	0	0	1	0	0	0	0	0	0	0	1	0	1	
Harbourfront East,Toronto Islands,Union Station	0	0	0	0	0	0	0	0	0	2	0	1	0	
Harbord,University of Toronto	0	0	2	0	0	0	0	0	0	0	0	0	0	
Commerce Court,Victoria Hotel	0	0	1	0	0	0	0	0	1	1	0	0	0	

5. Discussion

Cluster 0 has low concentration of restaurants meaning it might not be a good place for profitable restaurants.
Cluster 1 has high concentration of restaurants which means competition will be very high
Cluster 2 has a medium concentration so it would be ideal for restaurants.

Also families spent at least one weekday and especially weekends shopping and that is also usually the same day they dine out! So we should look for neighborhoods with lot of shopping possibilities.

As per our observation the top 3 ideal neighborhoods to start a multicuisine restaurant would be (Ryerson,Garden District) , (Adelaide,King,Richmond) and (Harbourfront East,Toronto Islands,Union Station) as they are one of the top largest shopping neighborhood and also fall under cluster 2.

6. Conclusion

To start a multicuisine restaurant we should consider the neighborhoods (Ryerson,Garden District) , (Adelaide,King,Richmond) and (Harbourfront East,Toronto Islands,Union Station). The final decision will depend on in depth analysis of these 3 neighborhoods as well as many other factors such as financing, operational agreements & business-related terms and conditions agreed upon by all the stakeholders involved.

7. References

- Beautiful Soup documentation - https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- Pandas documentation - https://pandas.pydata.org/pandas-docs/stable/
- Folium documentation - https://python-visualization.github.io/folium/
- Wiki: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M