Most common types of restaurants in Toronto Area

Coursera Capstone Project

Oscar Tarique

4/4/2019

Table of Contents

Table of Figures

1. Introduction

1.1. Background and problem

One of the major pitfalls before venturing on a project/starting a business is not performing sufficient market research such as gathering information about your competitors; getting a grasp of the customer base and possible growth opportunities. It is a crucial step to forge the success of any business venture. This can be applied to restaurants as well. Location and customer base are two key components to a restaurant's success. Many new restaurants fail by either being located in areas where customers will not be interested or having a restaurant serving a cuisine that is the exact same as several other restaurants in the area. Hence having information with regards to which areas have what kind of restaurants will give entrepreneurs an edge to make various informed business decisions such as on where to setup the restaurant; what type of cuisine to offer; financial implication; investments etc.

2. Methodology

2.1. Data Sources

Datasets required to tackle the problem are Toronto's geospatial data; neighborhood data in the Toronto area and restaurants within in each neighborhood. Data regarding Toronto's postal codes, boroughs and neighborhood was obtained using Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). Geospatial data containing the geographical coordinates of the neighborhoods was obtained from https://cocl.us/Geospatial_data. Finally, information regarding restaurants was collected using the Foursquare API.

2.2. Data Cleaning and Preparation

First the data or rather the table containing the required information on the Wikipedia page was scraped using the BeautifulSoup package in python. This was then processed into a dataframe with columns PostalCode, Borough and Neighborhood. Cells containing values of "Not assigned" in the boroughs column had their respective row removed. Postal codes containing the same value were grouped and their corresponding neighborhood values collated. Neighborhood values with "Not assigned" were given values the same as their borough.

The geospatial data was processed into another dataframe which contained columns Postalcode, Latitude and Logitude. Thus the two dataframes were merged using the Postalcode as the two dataframe's common key and processed into one final dataframe. This dataframe was used as the grounds to do all further analysis.

2.3. Cluster Analysis

K-means clustering algorithm was used to segment the Toronto city data into a fixed number of clusters. This is achieved by defining a target value of *k* that indicates the number of centroids in a dataset, which in turn is a center of a cluster. It is an iterative process that begins with random/predefined centroid positions; calculates the distance between each data point and a centroid and then allocates the data points to the closest cluster. This continues until the centroids have stabilised.

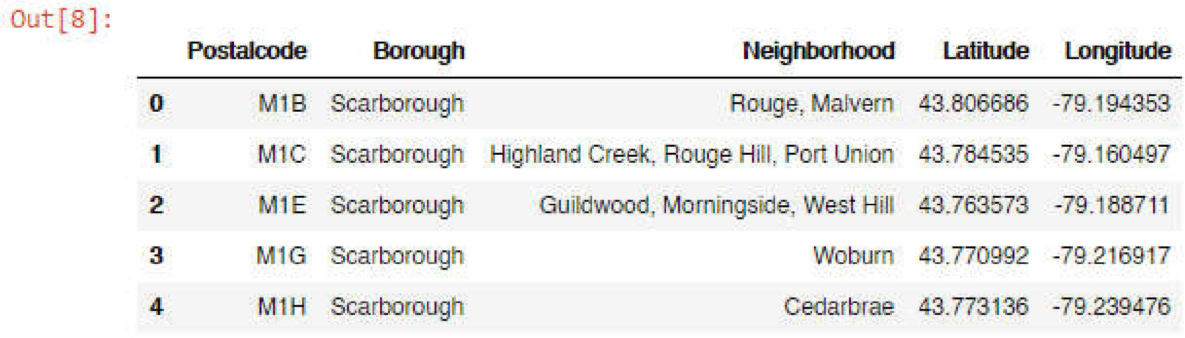
For this problem, the target value *k* was chosen to be five. This segmented the Toronto city data into five clusters using the latitude and longitude values. This was shown on a map using the Folium package on python.

2.4. Restaurant Analysis

The Foursquare API was used to get the most common restaurant cuisines in the neighborhood(s) within each cluster using their location data. This was achieved by first making calls to the Foursquare API to gather restaurant venues within a 500m radius of every neighborhood(s). This dataframe was then processed to rank the most common cuisines in the neighborhood(s). This was then merged with the clustered dataframe where exploratory data analysis was done.

3. Results and Discussion

The image in Figure 1 shows an excerpt (top 5 entries) of the cleaned dataframe containing the borough, neighborhood, postal code and location data for the city of Toronto.



Out[8]:

	Postalcode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	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

Figure 1 - Cleaned dataframe of Toronto City Dataset

The final shape of the dataframe is 103 rows and 5 columns.

Before beginning to cluster, the dataframe above can represented on a map using the Folium package in python as shown in Figure 2.

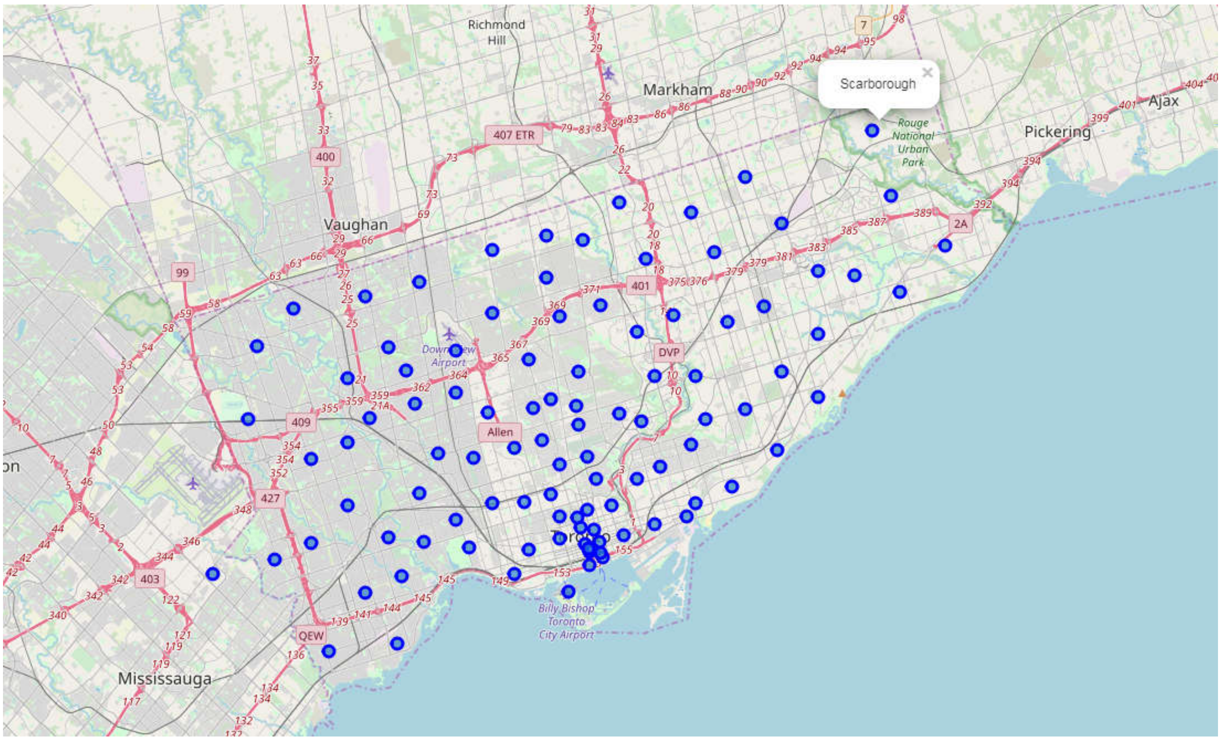
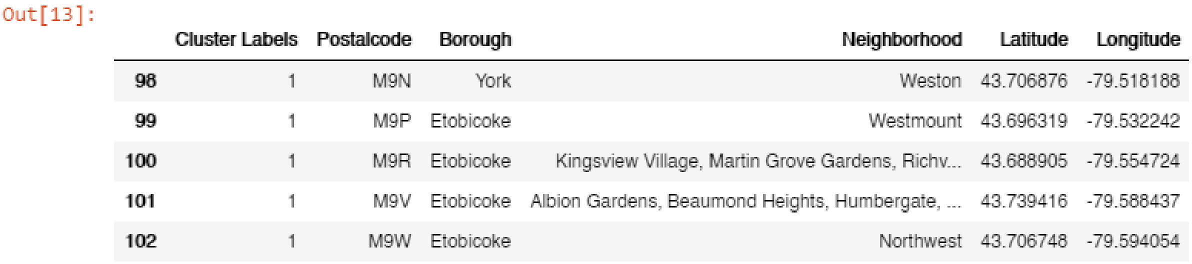


Figure 2 - Interactive map of Toronto City with location data

The K-means algorithm with a target value of five is used on the dataframe to segment the Toronto city data into 5 clusters. Figure 3 show an excerpt output (Bottom 5 entries) of the dataframe with the cluster labels included and Figure 4 visually represents the clusters on a map.



Out[13]:

	Cluster Labels	Postalcode	Borough	Neighborhood	Latitude	Longitude
98	1	M9N	York	Weston	43.706876	-79.518188
99	1	M9P	Etobicoke	Westmount	43.696319	-79.532242
100	1	M9R	Etobicoke	Kingsview Village, Martin Grove Gardens, Richv...	43.688905	-79.554724
101	1	M9V	Etobicoke	Albion Gardens, Beaumond Heights, Humbergate, ...	43.739416	-79.588437
102	1	M9W	Etobicoke	Northwest	43.706748	-79.594054

Figure 3 - Merged dataframe with cluster labels

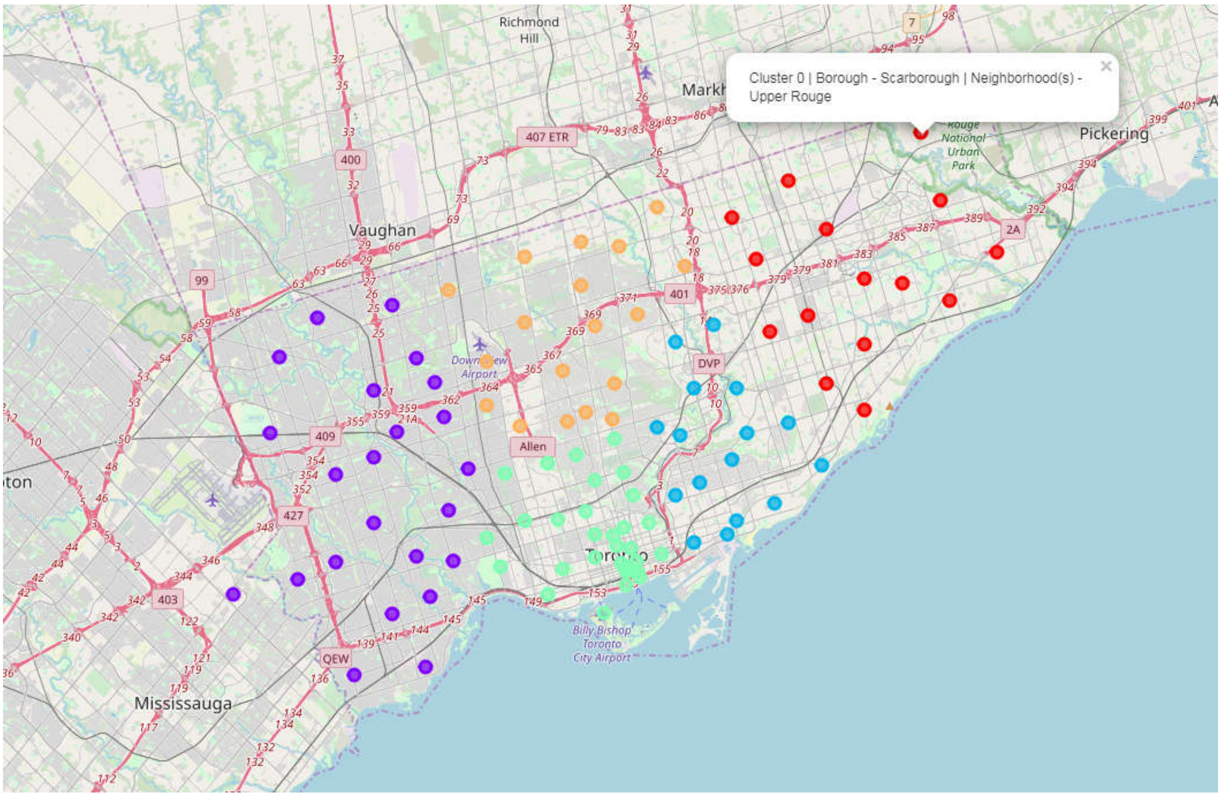
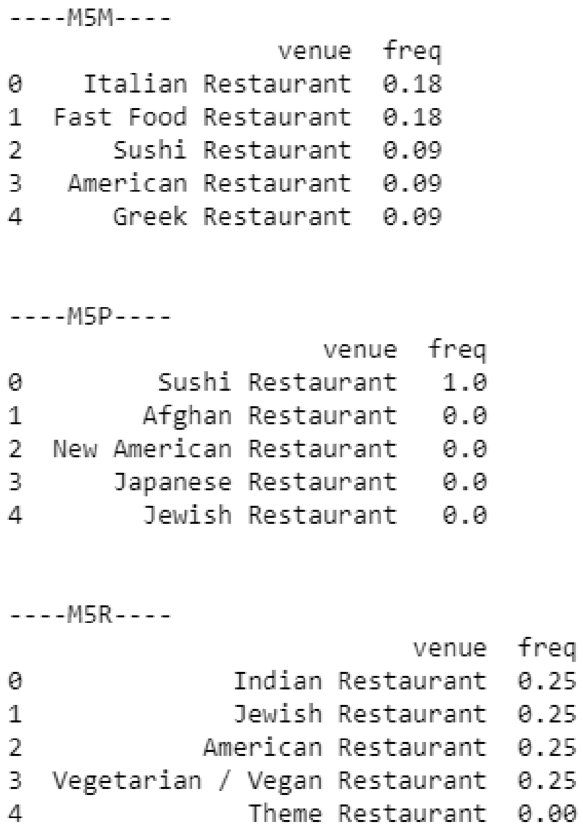

Cluster 0 | Borough - Scarborough | Neighborhood(s) - Upper Rouge
Richmond Hill
Markh
Vaughan
Pickering
Rouge National Urban Park
Downsview Airport
Allen
DVP
407 ETR
400
401
409
427
403
QEW
Toronto
Billy Bishop Toronto City Airport
Mississauga

Figure 4 - Map showing the clusters of Toronto City dataset

The Foursquare API was then used to gather information of the most common restaurants cuisines and rank them accordingly. This was done by first one hot encoding, grouping and calculating the frequency. An excerpt of the output of the results in shown in Figure 5, which shows the top 5 restaurants cuisines in the neighborhood(s). It is to be noted that as that the Postalcode column was used as the key instead of the neighborhood column. This is primarily due to the postalcode column having singular unique values as opposed to the neighborhood column, which contains numerous values.



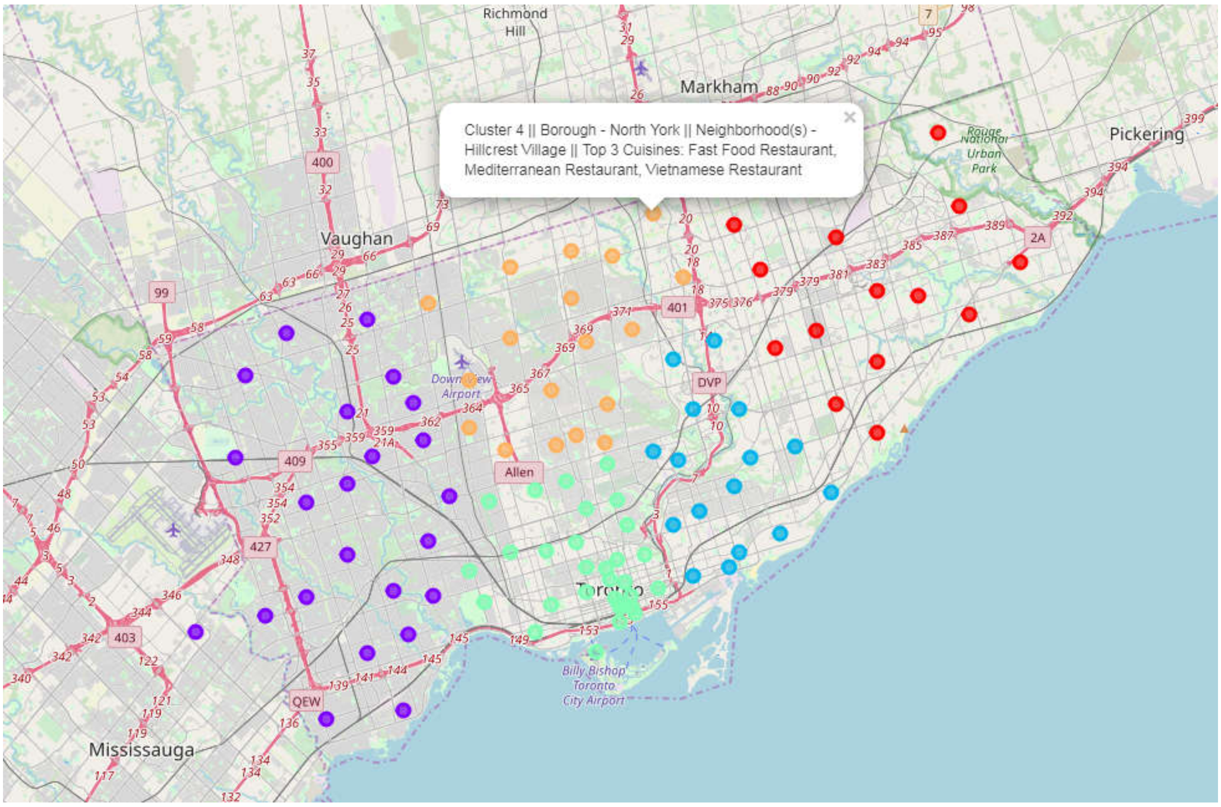
```
----M5M----
                   venue  freq
0     Italian Restaurant  0.18
1   Fast Food Restaurant  0.18
2       Sushi Restaurant  0.09
3    American Restaurant  0.09
4       Greek Restaurant  0.09


----M5P----
                       venue  freq
0           Sushi Restaurant   1.0
1          Afghan Restaurant   0.0
2  New American Restaurant     0.0
3        Japanese Restaurant   0.0
4          Jewish Restaurant   0.0


----M5R----
                             venue  freq
0                Indian Restaurant  0.25
1                Jewish Restaurant  0.25
2              American Restaurant  0.25
3  Vegetarian / Vegan Restaurant    0.25
4                 Theme Restaurant  0.00
```

Figure 5 - Top 5 cuisines in respective neighborhood(s)

Thereafter, this data was expanded to include the top 10 most common restaurant cuisines and merged with the first dataframe so that calls can be made should one be interested in gaining insight of the type restaurants available in a certain cluster or neighborhood. Figure 6 shows an updated Folium map that now contains the top 3 restaurants in the popup when clicking on a neighborhood.



Cluster 4 || Borough - North York || Neighborhood(s) - Hillcrest Village || Top 3 Cuisines: Fast Food Restaurant, Mediterranean Restaurant, Vietnamese Restaurant

Exploring the data, we can make the following observations:

In cluster 0 (red), which is predominantly the Scarborough borough on the Eastern side, the two most common restaurants in the Top 3 are Vietnamese and Doner Restaurants. There are 5 entries with NaN values which is due to there being no information being available on the Foursquare API. These areas could have niche restaurants and be further explored to introduce a new franchise if the market exists for it.

In cluster 1 (purple), which contains boroughs Etobicoke, York and North York on the western side, has far more NaN entries. This could indicate a possible source of development and worth further investigation. This is similar to cluster 4 (yellow), which is on the Northern side and consists mostly of the North York borough.

In cluster 2 (blue), which is on the Southern Eastern side, has Fast Food and Italian cuisines in 1st place. Similarly in cluster 3 (light green), which is in Downtown Toronto on the Southern side, is predominantly Italian cuisine

4. Conclusions

In conclusion, the study consisted of acquiring location data related to Toronto City and leveraging the Foursquare API to get restaurant cuisines around various neighborhoods in Toronto City. The location dataset was cleaned and K-mean cluster algorithm was employed on it to segment the city in 5 different clusters. Thereafter, the Foursquare API was used to get the Top 10 most common restaurant cuisines in every neighborhood(s) in each cluster. This was embedded into an interactive map using Folium that gives a popup with all relevant data. This information can be used by potential entrepreneurs to make informed business decisions on where to setup the restaurant and what type of cuisine to offer. This can potential avoid the pitfall of new restaurants failing by either being located in areas where customers will not be interested or having a restaurant serving a cuisine that is the exact same as several other restaurants in the area.