Find out customer's activity concentration by using Data Science

Different Places -Different Customers- Toronto

Data based classification of Toronto locations

UMESHPRATAP SINGH

Contents

1. Introduction

Many research companies and /or investors always keep trying to find out perfect location for a particular business interest. Typically , research companies rely on primary research methods like sample survey of customers in a particular area to gauge their preferences. However, such methods are costly and are likely to contain biasness in report based on sample selection.

With the help of Data Science, we can propose an alternative method of selecting location based on customer preference by using location-based API like Foursquare.

Based on already available data /customers feedback on Foursquare, we can use advance unsupervised machine learning tool like K-clustering method to identify locations which have a particular customers preference.

2. Background

One of our Market research client has been using traditional methods of market research like customer survey and other published market reports. However these methods are time confusing and every time it needs to be customize based on end user requirement. Hence, our client has asked us to help it by classifying Toronto locations based on customers' activities or preference using available data science tools , such that its all conclusive and versatile to be used at any location /preference.

3. Problem

Our marketing research firm wants to know more about booming real estate industry (in commercial /shop) of Canada's Financial Capital Toronto . It wants to highlight key business activities and preferred locations of shops based on customer preference to its clients.¶

Further, it has short listed Main Toronto as location to be drilled down but are not sure which part of Main Toronto, it should suggest as preferred location to its clients and which type of shop/commercial category it should advise to its client.

Solution logic :

- Typically , an area where most of shopping / commercial outlets are located , suggests that it has more commercial activities in and around that area.
- Further, based on simliar grouping, we can identify customer specific locations i.e what type of customer visit that place and/or what activity they perform.

Approach :

- We can demarcate Downtown Toronto area on the basis of pin codes and then using Foursquare API n identify which area has maximum stores/venues located within 500 meters of radius.

- Further, with using clustering algorithm, we can divide each location based on type of venues /shops present , indicating what type of customers /activity is preferred in each locations

4 .Raw Data

We will require mainly two data sets.

1. Data set require for downloading pin code wise demarcation of Toronto city.

- For this we will be using url form Wikipedia https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and will do web scrapping
- further , we will required latitude and longitude of each neighborhood, for that we will get data from http://cocl.us/Geospatial_data so we will further refine the data set to reduce the number of columns .

2. Foursquare API data :

- For category selection, feedback and popularity index , we will need Foursquare API to get required data and map them on Mumbai data set.

  Api which we will be using are as follows

- Venue search :
  - Request : GET https://api.foursquare.com/v2/venues/search
  - Returns a list of venues near the current location, optionally matching a search term.
- Explore a place :
  - Request : https://api.foursquare.com/v2/venues/explore?
  - Returns total places near by

5. Data wrangling /preparation

Data format from both the site's is not in our desired format.

From Wikipedia :

First we need to do data scraping using beautifulSoup to bring data from site into data frame

We got data from Wikipedia in the following format :

	PostalCode	Borough	Neighborhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

It has some not assigned value of Borough which are not useful for us , as we can not work on these rows further, so we selected df with Borough != Not assigned.

Further, it does not contain Latitude and longitude details , which is required for our analysis as Foursquare links locations based on Latitude and longitude.

For that we got data from http://cocl.us/Geospatial_data. Then using merge method of pandas , we created our desired data set as follows.

	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	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Further, since our area of interest is only main Toronto , we selected our main Toronto data as follows

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
4	M4E	East Toronto	The Beaches	43.676357	-79.293031

From foursquare Api :

We got our data in with lots of columns which we don't required. We used filter columns to get desired data format.

filtered_columns = ['venue.name', 'venue.categories', 'venue.location.lat', 'venue.location.lng']

further, venue.categories was in nested list format as shown below :

	venue.name	venue.categories	venue.location.lat	venue.location.lng
0	Downtown Toronto	[{'id': '4f2a25ac4b909258e854f55f', 'name': 'N...	43.653232	-79.385296
1	Nathan Phillips Square	[{'id': '4bf58dd8d48988d164941735', 'name': 'P...	43.652270	-79.383516
2	UNIQLO ユニクロ	[{'id': '4bf58dd8d48988d103951735', 'name': 'C...	43.655910	-79.380641
3	Elgin And Winter Garden Theatres	[{'id': '4bf58dd8d48988d137941735', 'name': 'T...	43.653394	-79.378507
4	Richmond Station	[{'id': '4bf58dd8d48988d14e941735', 'name': 'A...	43.651569	-79.379266

We used foursquare function 'get_category_type' to get only name from venue.categories column

	name	categories	lat	lng
0	Downtown Toronto	Neighborhood	43.653232	-79.385296
1	Nathan Phillips Square	Plaza	43.652270	-79.383516
2	UNIQLO ユニクロ	Clothing Store	43.655910	-79.380641
3	Elgin And Winter Garden Theatres	Theater	43.653394	-79.378507
4	Richmond Station	American Restaurant	43.651569	-79.379266

Now our data is in desired format to work upon.

6. Methods

We used Folium for plotting map , pandas for data frame statistics and sklearn for kmeans machine learning algorithm.

1. Plotting all shortlisted Toronto location on map using folium

2. We called foursquare explore venue function for all our above locations to down load the desired data set in ‘main_toronto_venues’ and then using groupby function ,gropued all venues at same location/neighborhood .

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Berczy Park	57	57	57	57	57
Brockton, Parkdale Village, Exhibition Place	25	25	25	25	25
Business reply mail Processing Centre, South Central Letter Processing Plant Toronto	19	19	19	19	19
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	16	16	16	16	16
Central Bay Street	67	67	67	67	67
Christie	16	16	16	16	16
Church and Wellesley	77	77	77	77	77
Commerce Court, Victoria Hotel	100	100	100	100	100
Davisville	36	36	36	36	36
Davisville North	7	7	7	7	7
Dufferin, Dovercourt Village	14	14	14	14	14
First Canadian Place, Underground city	100	100	100	100	100
Forest Hill North & West, Forest Hill Road Park	4	4	4	4	4
Garden District, Ryerson	100	100	100	100	100
Harbourfront East, Union Station, Toronto Islands	100	100	100	100	100

By using one hot coding we got detailed break up of venue category name for each neighborhood . then we used mean function of pandas to sort all venues by their occurrence in each neighborhood as follows.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Berczy Park	Coffee Shop	Restaurant	Bakery	Cocktail Bar
1	Brockton, Parkdale Village, Exhibition Place	Café	Yoga Studio	Bakery	Breakfast Spot
2	Business reply mail Processing Centre, South C...	Light Rail Station	Yoga Studio	Garden Center	Skate Park
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Lounge	Airport Service	Airport Terminal	Boutique
4	Central Bay Street	Coffee Shop	Sandwich Place	Café	Italian Restaurant
5	Christie	Grocery Store	Café	Park	Restaurant
6	Church and Wellesley	Coffee Shop	Japanese Restaurant	Sushi Restaurant	Restaurant
7	Commerce Court, Victoria Hotel	Coffee Shop	Restaurant	Café	Hotel
8	Davisville	Pizza Place	Dessert Shop	Sandwich Place	Sushi Restaurant
9	Davisville North	Gym	Hotel	Breakfast Spot	Food & Drink Shop
10	Dufferin, Dovercourt Village	Bakery	Pharmacy	Brewery	Supermarket
11	First Canadian Place, Underground city	Coffee Shop	Café	Hotel	Japanese Restaurant
12	Forest Hill North & West, Forest Hill Road Park	Park	Jewelry Store	Trail	Sushi Restaurant

Finally using Kmeans method to cluster all location under unsupervised learning using following code

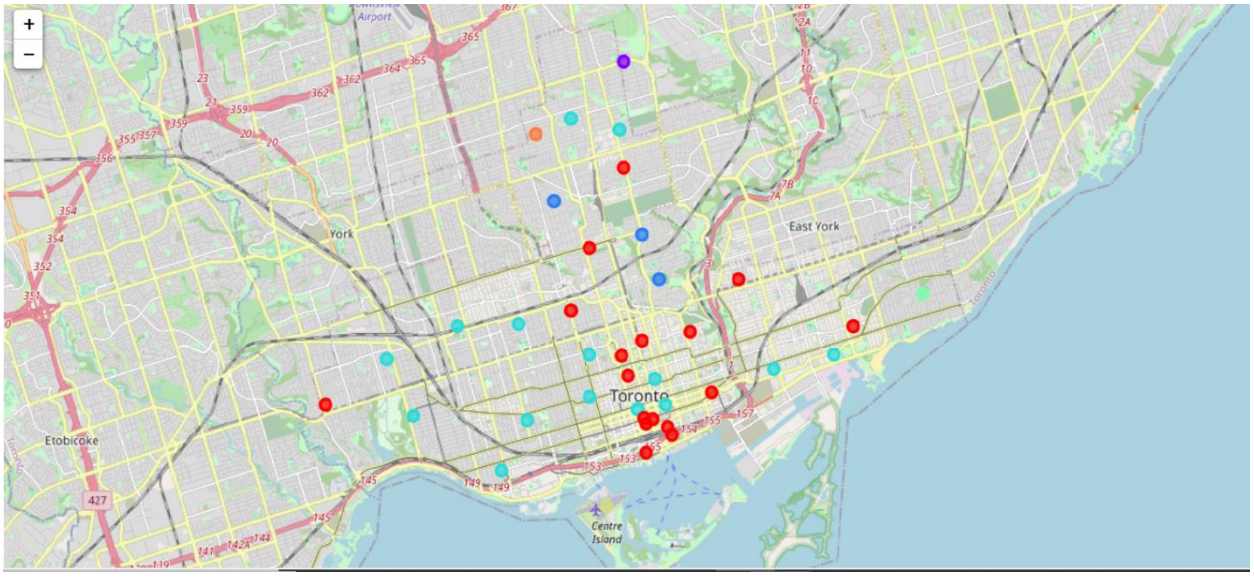
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(main_toronto_grouped_clustering)

```
# set number of clusters
kclusters = 7

main_toronto_grouped_clustering = main_toronto_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(main_toronto_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:38]
```

```
[175]: array([0, 3, 3, 5, 0, 3, 0, 0, 0, 3, 3, 0, 2, 3, 0, 3, 0, 3, 1, 3, 2, 3,
               3, 0, 0, 3, 2, 6, 0, 3, 0, 0, 3, 0, 0, 4, 0, 0], dtype=int32)
```

7. Findings

We found that kmeans algorithm defined clusters in following way .



Detail of each cluster is as follows

- **Cluster 1 : Office Center**

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Downtown Toronto	0	Coffee Shop	Pub	Bakery	Park
1	Downtown Toronto	0	Coffee Shop	Diner	College Auditorium	Beer Bar
5	Downtown Toronto	0	Coffee Shop	Restaurant	Bakery	Cocktail Bar
6	Downtown Toronto	0	Coffee Shop	Sandwich Place	Café	Italian Restaurant
10	Downtown Toronto	0	Coffee Shop	Aquarium	Hotel	Café
12	East Toronto	0	Greek Restaurant	Coffee Shop	Italian Restaurant	Furniture / Home Store
13	Downtown Toronto	0	Coffee Shop	Café	Hotel	Restaurant
15	East Toronto	0	Park	Brewery	Burrito Place	Ice Cream Shop
16	Downtown Toronto	0	Coffee Shop	Restaurant	Café	Hotel
24	Central Toronto	0	Sandwich Place	Café	Coffee Shop	Liquor Store
26	Central Toronto	0	Pizza Place	Dessert Shop	Sandwich Place	Sushi Restaurant
28	West Toronto	0	Coffee Shop	Pizza Place	Café	Italian Restaurant
31	Central Toronto	0	Pub	Coffee Shop	Bagel Shop	Pizza Place
34	Downtown Toronto	0	Coffee Shop	Café	Hotel	Pub
35	Downtown Toronto	0	Coffee Shop	Pizza Place	Italian Restaurant	Restaurant
36	Downtown Toronto	0	Coffee Shop	Café	Hotel	Japanese Restaurant
37	Downtown Toronto	0	Coffee Shop	Japanese Restaurant	Sushi Restaurant	Restaurant

first cluster has lots of meeting joints like coffee shops , cafe , restaurant and hotel. this indicate that this cluster has lots of commercial activity in a day and has lots of office. If some one wants to target office goers can open store in this cluster. We name it "Office center"

- **Cluster 3 : Open spaces**

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
21	Central Toronto	2	Park	Jewelry Store	Trail	Sushi Restaurant
29	Central Toronto	2	Restaurant	Park	Trail	Deli / Bodega
33	Downtown Toronto	2	Park	Playground	Trail	Yoga Studio

This cluster clearly suggests that it has lots of open spaces and is ideal for adventure sports shops complementing trail and park already present in each location . We can name this cluster as "Open Space"

- **Cluster 4 : Mixed use**

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
2	Downtown Toronto	3	Clothing Store	Coffee Shop	Café	Cosmetics Shop
3	Downtown Toronto	3	Café	Coffee Shop	Clothing Store	Restaurant
7	Downtown Toronto	3	Grocery Store	Café	Park	Restaurant
8	Downtown Toronto	3	Coffee Shop	Café	Restaurant	Bar
9	West Toronto	3	Bakery	Pharmacy	Brewery	Supermarket
11	West Toronto	3	Bar	Coffee Shop	Asian Restaurant	Restaurant
14	West Toronto	3	Café	Yoga Studio	Bakery	Breakfast Spot
17	East Toronto	3	Café	Coffee Shop	American Restaurant	Bakery
20	Central Toronto	3	Gym	Hotel	Breakfast Spot	Food & Drink Shop
22	West Toronto	3	Thai Restaurant	Mexican Restaurant	Café	Discount Store
23	Central Toronto	3	Clothing Store	Coffee Shop	Yoga Studio	Mexican Restaurant
25	West Toronto	3	Breakfast Spot	Gift Shop	Restaurant	Movie Theater
27	Downtown Toronto	3	Café	Restaurant	Bar	Bookstore
30	Downtown Toronto	3	Café	Vegetarian / Vegan Restaurant	Coffee Shop	Bar
38	East Toronto	3	Light Rail Station	Yoga Studio	Garden Center	Skate Park

This above cluster has cafe , hotel , bar, discount store , clothing store etc . indicating it has mixed customers are with part residential and part commerical space presence and hence we name it Mixed use

- Cluster 6 : Airport

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
32	Downtown Toronto	5	Airport Lounge	Airport Service	Airport Terminal	Boutique

This above cluster clearly indicates the presence of Airport and allied services and anybody interested in airport related business activity must prefer this

- Cluster 7 : Residential

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
19	Central Toronto	6	Home Service	Garden	Music Venue	Yoga Studio

The above cluster has home service , garden and yoga studio, clearly indicating that its a residential area and accordingly somebody should plan a shop/outlet in this area

Note : Cluster 2 and cluster 5 did not yielded in any meaningful insight and hence were excluded from the report.

8. Solution /conclusion

Based on clustering of main Toronto area we can draw conclusion as per following table

Sr. No	Name of clusters	Total number of locations	Suggestions
1	Office spaces	17	ideal for joints like coffee shops , restaurant ete
2	Open Spaces	3	Ideal for adventure sports shops
3	Mixed spaces	15	has mix of residential and commercial, good for shopping malls with joints
4	Air port	1	good for airport related joints/ services- duty free shops
5	Residential	1	good for home service, home decor etc

As per above table, main Toronto has most places occupied by office spaces , followed by mixed uses . the table clearly indicates which type of main activities /customer preferences are prevailing in each of the locations.

9. Future direction

We can further evolve this project so that it can be used for multiple cities and/or further drilling of shortlisted locations.

Also we can use additional data like per capita income for each location to make further in depth finding of customer choices and locations .