Research on Location Selection of Street Stores in Toronto

1. Introduction

A long time ago, I read a report by chance. Researchers at the University of Cambridge have analyzed the 620,000 check-in data related to Starbucks and McDonald's users posted on Twitter within 5 months, combined with traditional location characteristics, using machine learning algorithms to get a result: Foursquare check-in data helps Starbucks chose a better store.

First, the research team analyzed the traditional location factors that influence passenger traffic, such as the types of nearby merchants, whether the types of merchants in the area are rich, whether there are attractive landmarks, and whether they are close to the train station. Second, they used the check-in data to analyze which are the customer-intensive areas, the popularity of one store to another, and the characteristics of passenger flow between different locations.

The characteristics and numerical values of these location factors are used to describe the location characteristics of each store and to train several different machine learning algorithms: Support Vector, Regression, M5 decision trees and Linear Regression. Based on a random sample of two-thirds of store locations and their known popularity, each algorithm will be trained 1,000 times, and then the system will find the remaining one-third store's passenger density. Then we will get the ranking of the pros and cons of the site selection. This result can be compared with the real sign-in density of these site selections to get the correlation.

The research team found that Starbucks, McDonald's, and Dunkin' Donuts have different signature data patterns. The total number of Starbucks check-ins is 5 times the sum of McDonald's and Dunkin' Donuts, but the fact that this result ignores is that Starbucks stores in Manhattan are twice as large as McDonald's and Dunkin' Donuts. And the probability of Starbucks being built near the train station is also greater.

The location prediction models of the three chain stores have different characteristics. The degree of competition is the most important factor in the Starbucks prediction model, which means that it performs better when there are fewer surrounding

competitors; for McDonald 's, its customers are mostly from outside the current business district, and they often don't care much Itinerary. Dunkin' Donuts customers are mostly shopping crowds, and its customers often sign in at neighboring shops.

This study found that even with these differences, no matter which chain store is targeted, the combination of traditional geographic features and the flow of people based on Foursquare's check-in data analysis can get the best prediction results.

Therefore, based on the research results of the report talked above. My Capstone Project will sort out the most popular types of stores in Toronto's neighborhoods, and form a sorted list to provide useful reference for the location selection of the new stores. Furthermore, different types of stores have different location strategies, some radical and some conservative. But no matter what type, you can get reference from the results of this capstone project.

2. Data Description

The data used by Capstone Project mainly includes three parts. The first is Toronto 's postal codes data, which mainly includes the three fields of Postal code, Borough, and Neighborhood. The second is to integrate postal codes data with GPS data to form a table with longitude and latitude. The third is the data obtained from Foursquare, the latitude and longitude of each Venue and the type of store from each Neighborhood, so that the space can be visualized more conveniently.

2.1 Importing Dataset

First of all, we Import postal code and neighborhood data of Toronto from Wikipedia.

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

2.2 Cleaning the data

Then the data cleaning is needed, for the following reasons:

- More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, 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.

After data cleaning, we get the postal code and neighborhoods data table as follows.

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
5	M9A	Etobicoke	Islington Avenue

Secondly, we import the postal code, latitude and longitude data.

	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

Through the postal code and latitude and longitude data, we can easily know that as long as you merge through the postal code field, you can know the latitude and longitude of each neighborhood.

After preparing, acquiring and cleaning the above three parts of data, we can use the machine learning clustering analysis algorithm for a more in-depth analysis.

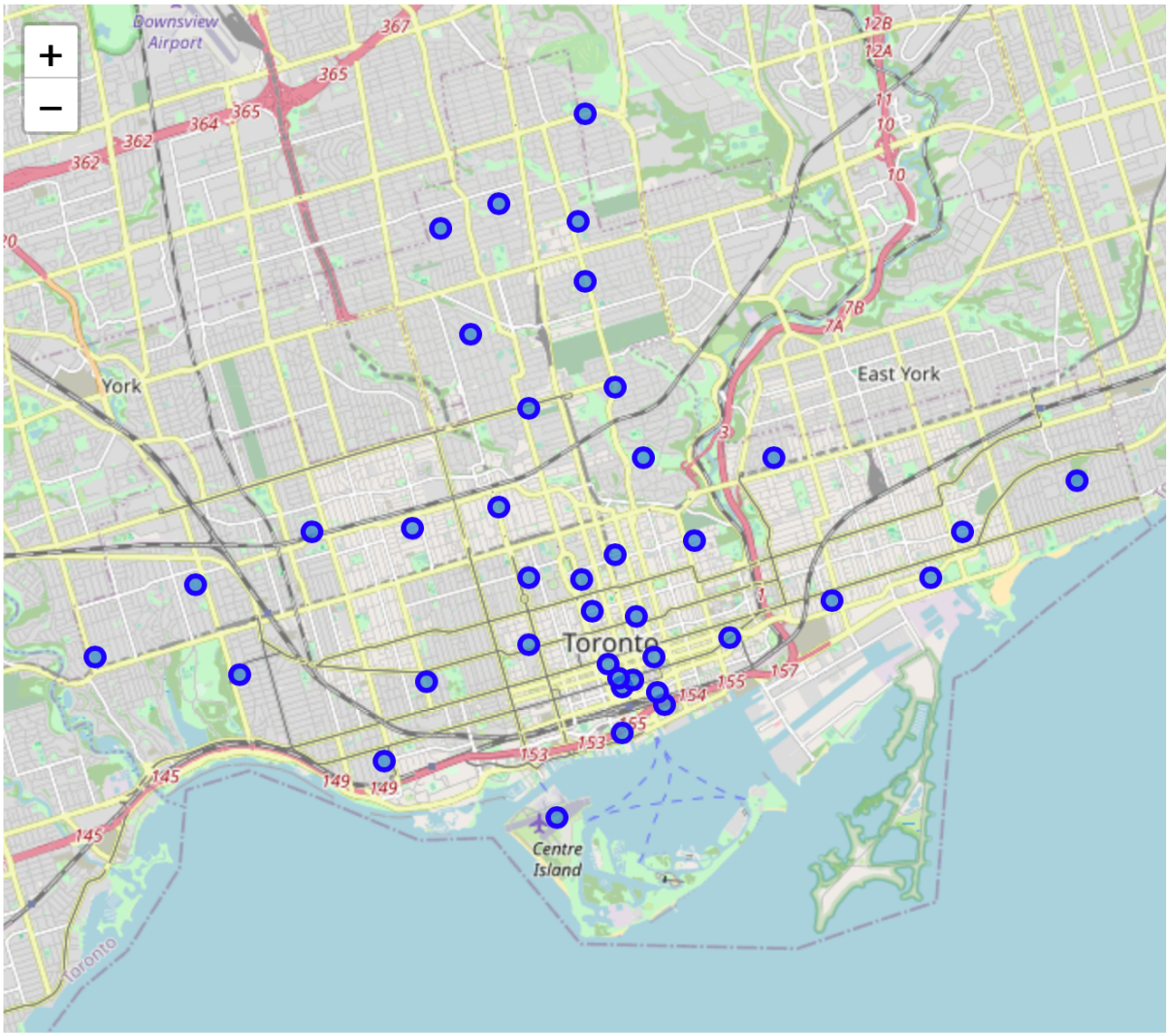
	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242

2.2 Selecting Toronto data and data visualization

Filter the data table obtained in the previous section, and only keep the data of Toronto, as shown in the right table. Both borough and neighborhood in the table are in Toronto. The dataframe has 4 boroughs and 39 neighborhoods. By this way, we only keep the data that we should pay attention to.

	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

Next, We use the folium function to draw the map, the geographical coordinates of Toronto are 43.6534817, -79.3839347. So we can get the data visualization of the neighborhoods of Toronto.



3. Methodology

In the previous section, we processed the basic data of Toronto, with the postal code, neighborhood, and corresponding latitude and longitude data, and drawn a visual map. The map visually shows the location distribution of neighborhoods.

In this section, we read the data of Foursquare Data and carry out specific machine learning and cluster analysis.

3.1 Getting Foursquare Data

We start utilizing the Foursquare API to explore the neighborhoods and segment them.

First, we define Foursquare credentials and version by adding the CLIENT_ID, CLIENT_SECRET and VERSION. Next, we get the top 100 venues that are in the first neighborhood within a radius of 500 meters. The result file returned through the API interface is in json format, and the data structure is shown in the figure.


```
{'meta': {'code': 200, 'requestId': '5ea03711c8cff26793583533'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Corktown',
  'headerFullLocation': 'Corktown, Toronto',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 48,
  'suggestedBounds': {'ne': {'lat': 43.6587599045, 'lng': -79.3544279001486},
   'sw': {'lat': 43.6497598955, 'lng': -79.36684389985142}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '54ea41ad498e9a11e9e13308',
       'name': 'Roselle Desserts',
       'location': {'address': '362 King St E',
```


As we know, all the information is in the items key. Before we proceed, let's borrow the get_category_type function from the Foursquare lab. Now we are ready to clean the json and structure it into a pandas dataframe. we know the number of venues were returned by Foursquare. 48 venues were returned by Foursquare.

```python
# function that extracts the category of the venue
def get_category_type(row):
    try:
        categories_list = row['categories']
    except:
        categories_list = row['venue.categories']

    if len(categories_list) == 0:
        return None
    else:
        return categories_list[0]['name']
```

3.2 Explore Neighborhoods in Toronto

Firstly, we create a function to repeat the same process to all the neighborhoods in Toronto.

```
def getNearbyVenues(names, latitudes, longitudes, radius=500):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&ll={},{
            CLIENT_ID,
            CLIENT_SECRET,
            lat,
            lng,
            VERSION,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

Then we run the above function on each neighborhood and create a new dataframe called toronto_venues. The size of the resulting dataframe is (1612, 7). There are 230 uniques categories.

3.3 Analyze Each Neighborhood

Analyze Each Neighborhood, let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category. And printing each neighborhood along with the top 5 most common venues.

```
----Berczy Park----
                venue  freq
0         Coffee Shop  0.07
1        Cocktail Bar  0.05
2  Italian Restaurant  0.03
3              Bakery  0.03
4                Café  0.03
```

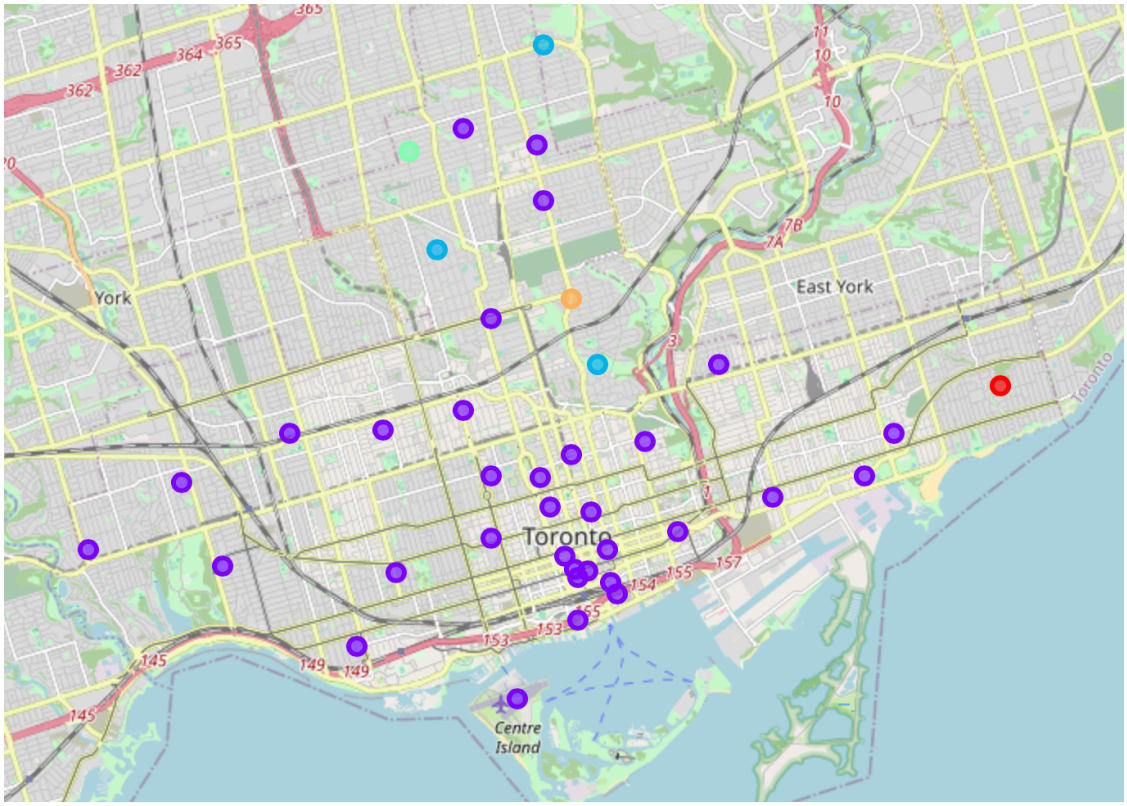
Next, we put that into a pandas dataframe. First, we write a function to sort the venues in descending order. Then we create the new dataframe and display the top 10 venues for each neighborhood. Result is shown in the following table.

	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	Cheese Shop	Farmers Market	Seafood Restaurant	Bakery	Restaurant	Italian Restaurant	Café	Beer Bar
1	Brockton, Parkdale Village, Exhibition Place	Café	Coffee Shop	Breakfast Spot	Gym	Bakery	Convenience Store	Performing Arts Venue	Pet Store	Climbing Gym	Restaurant
2	Business reply mail Processing CentrE	Yoga Studio	Spa	Garden Center	Garden	Fast Food Restaurant	Farmers Market	Light Rail Station	Comic Shop	Pizza Place	Butcher
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Lounge	Airport Terminal	Sculpture Garden	Plane	Airport	Airport Food Court	Rental Car Location	Boutique	Harbor / Marina
4	Central Bay Street	Coffee Shop	Italian Restaurant	Sandwich Place	Café	Salad Place	Fried Chicken Joint	Bubble Tea Shop	Burger Joint	Japanese Restaurant	Ice Cream Shop

Now we can finally use the cluster analysis algorithm of machine learning.

3.3 Cluster Neighborhoods

Run k-means to cluster the neighborhood into 5 clusters. Set number of clusters equals to 5, run k-means clustering, and check cluster labels generated for each row in the dataframe. We create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood. Adding clustering labels and merging toronto_grouped with toronto_data to add latitude/longitude for each neighborhood. Check the last columns, and in the final part, we visualize the resulting clusters.



4. Results

Now, we can examine each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, we can then assign a name to each cluster.

4.1 Cluster 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
4	East Toronto	0	Health Food Store	Trail	Pub	Women's Store	Department Store	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop

4.2 Cluster 2

Due to limited space, only the first five lines are shown here.

	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	Downtown Toronto	1	Coffee Shop	Pub	Bakery	Park	Breakfast Spot	Restaurant	Café	Theater	Mexican Restaurant	Shoe Store
1	Downtown Toronto	1	Coffee Shop	Diner	Sushi Restaurant	Gym	Restaurant	Park	Mexican Restaurant	Juice Bar	Italian Restaurant	Hobby Shop
2	Downtown Toronto	1	Clothing Store	Coffee Shop	Café	Restaurant	Japanese Restaurant	Bubble Tea Shop	Cosmetics Shop	Middle Eastern Restaurant	Tea Room	Bookstore
3	Downtown Toronto	1	Coffee Shop	Café	Cocktail Bar	Gastropub	Hotel	Italian Restaurant	American Restaurant	Seafood Restaurant	Farmers Market	Department Store
5	Downtown Toronto	1	Coffee Shop	Cocktail Bar	Cheese Shop	Farmers Market	Seafood Restaurant	Bakery	Restaurant	Italian Restaurant	Café	Beer Bar

4.3 Cluster 3

	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
18	Central Toronto	2	Park	Bus Line	Swim School	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop
21	Central Toronto	2	Park	Jewelry Store	Home Service	Trail	Sushi Restaurant	Eastern European Restaurant	Dumpling Restaurant	Electronics Store	Department Store	Doner Restaurant
33	Downtown Toronto	2	Park	Playground	Trail	Deli / Bodega	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Doner Restaurant

4.4 Cluster 4

Cluster 4 has only one neighborhood.

	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
19	Central Toronto	3	Garden	Women's Store	Department Store	Event Space	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Doner Restaurant

4.5 Cluster 5

	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
29	Central Toronto	4	Playground	Women's Store	Department Store	Event Space	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Doner Restaurant

Cluster 5 has only one neighborhood, too.

5. Discussion

Combining the above analysis, we can know that the most popular types of stores among clusters vary widely. But there are extremely high similarities in the same cluster. This has important strategic significance for the selection of new stores. Through analysis, the company can set its own operation and location strategy, choose to participate in fierce competition, or stay away from the same type of store, and go to a store that has not yet been developed. region. It all depends on the company's decision, but our use of data can provide strong support for the company's decision. This is also the most important significance of this report.

6. Conclusions

The conclusion is that through Foursquare data, combined with machine learning and cluster analysis, when the company decides to open a new chain store, it can combine its own location strategy to find the most suitable store area.