1.Introduction

Machine Learning (ML) is a form of AI that enables a system to learn from data rather than explicit programming. Machine learning has become one of the most important topics within organisations that are looking for new ways to leverage data to help the business and gain a new level of understanding. ML algorithms are advantageous since they offer solutions to problems related to the big quantities of different types of data in a speedy way.

Unsupervised learning is best suited when the problem requires a massive amount of data that is unlabelled. Unsupervised learning algorithms segment data into groups of examples (clusters) or groups of features. One of the popular unsupervised algorithms is K-Means. It is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining.

Foursquare is a location technology platform where people comment and rank food sites, coffee sites, malls and parks. I am consulting a customer who is going to choose one of the job options in New York and Toronto. We are trying to analyse the similarities and differences between neighbourhoods in downtown Toronto and Manhattan by using Foursquare location data along with a K-means clustering algorithm. We try to decide by looking most common venues in these areas and try to find a place for his pleasure.

2.Data

For this project, we have used the Foursquare API. Two CVS files about New York and Toronto are downloaded from the links attached below so that their longitude and latitude coordinates are obtained.

New York neighbourhoods: https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json

Toronto neighbourhoods: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

We have structured the data and focus on neighbourhoods in Downtown Toronto and Manhattan. Then we merged the data. A Foursquare API GET request is sent in order to acquire the surrounds venues that are within a radius of 500m. The data is formatted using one hot encoding with the categories of each venue. Then, the venues are grouped by neighbourhoods computing the mean of each feature. The similarities will be determined based on the frequency of the categories found in the neighbourhoods. These clusters help my customer to decide which areas are good for her taste.

3. Methodology

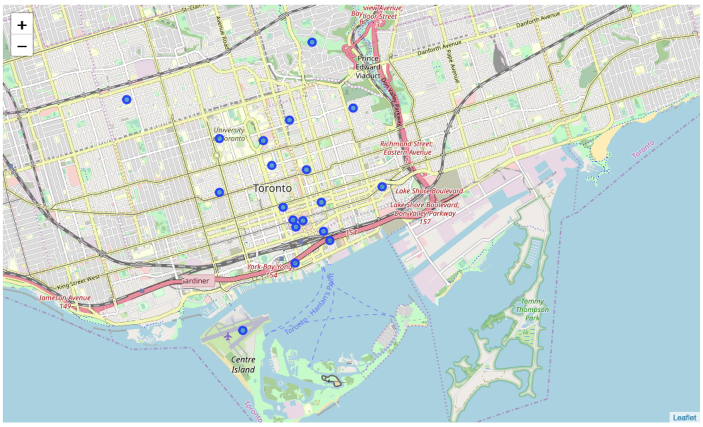
In order to find similarities between neighbourhoods, a clustering algorithm is implemented. In this case K-Means is used due to its simplicity and its efficiency. This algorithm search clusters within the data and the main objective function is to minimise the data dispersion for each cluster.

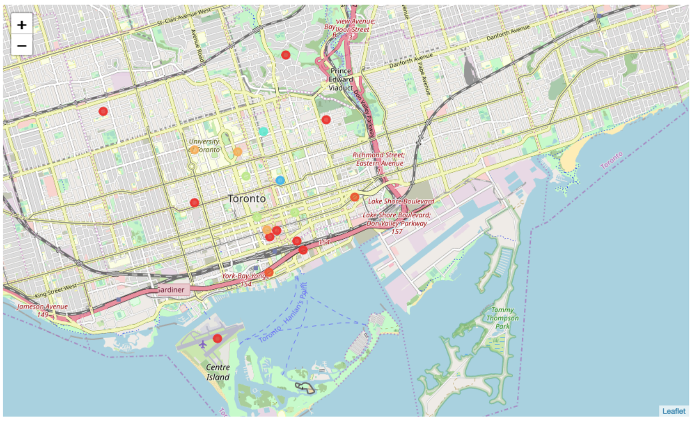
One Hot Encoding (Binary) is used in terms of categories. Therefore, each feature is a category that belongs to a venue. All the venues are grouped by the neighbourhoods, computing at the same time the mean. So we obtain a venue for each row and each column containing the frequency of occurrence of that particular category.

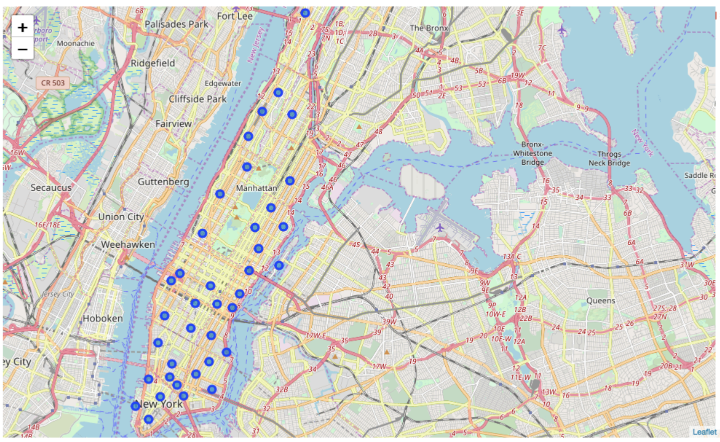
In K-means algorithm, one has to set the number of clusters, to find the best K, we use elbow method. A chart that compares error vs number of cluster is done and the elbow is selected.

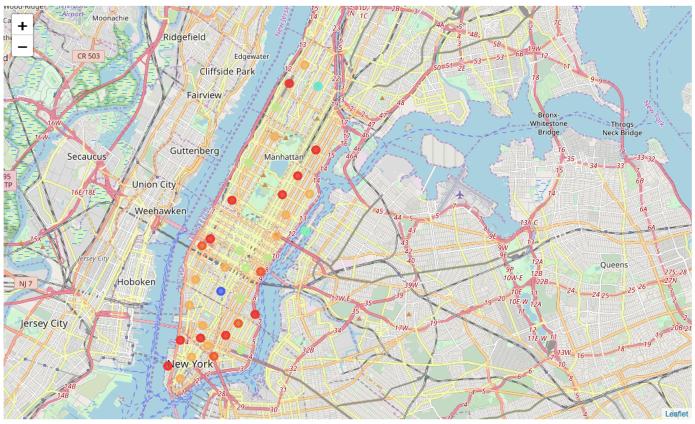
4. Results

By using Folium library, we show how k-means clustering applied in our data on the maps below.

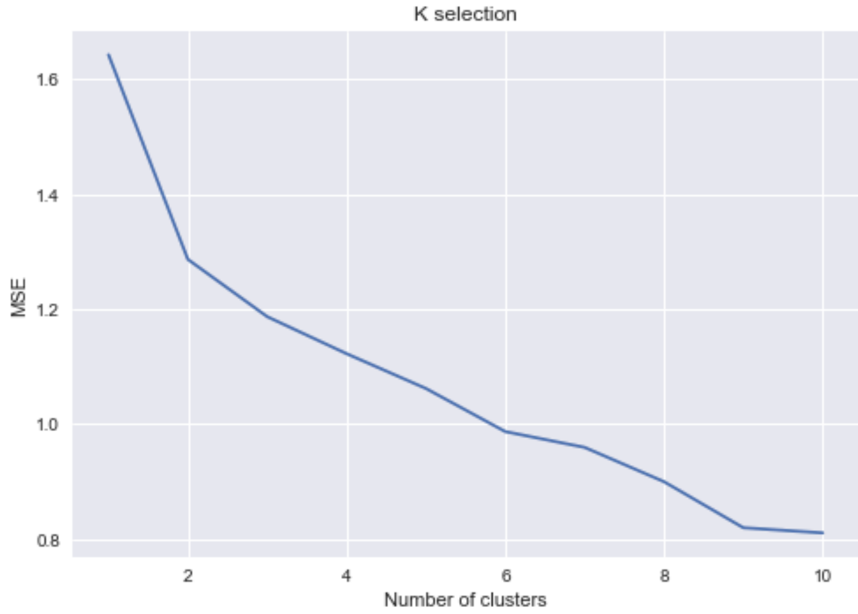




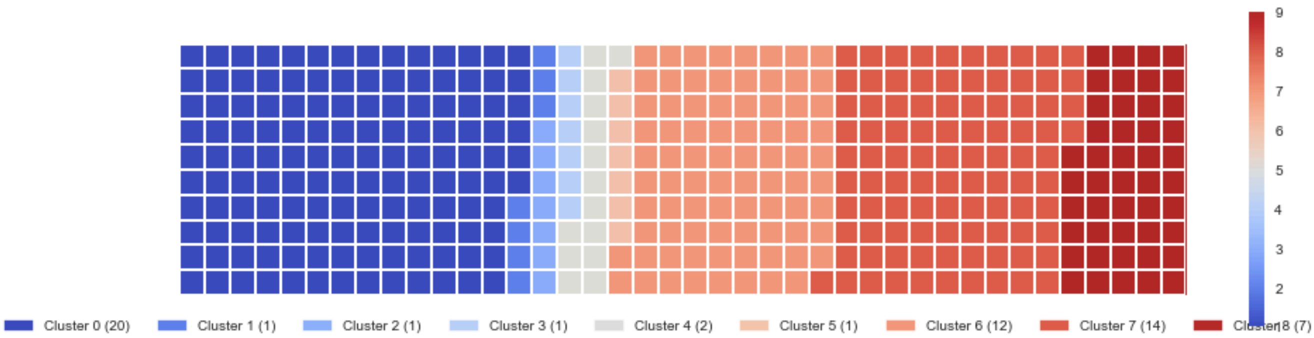




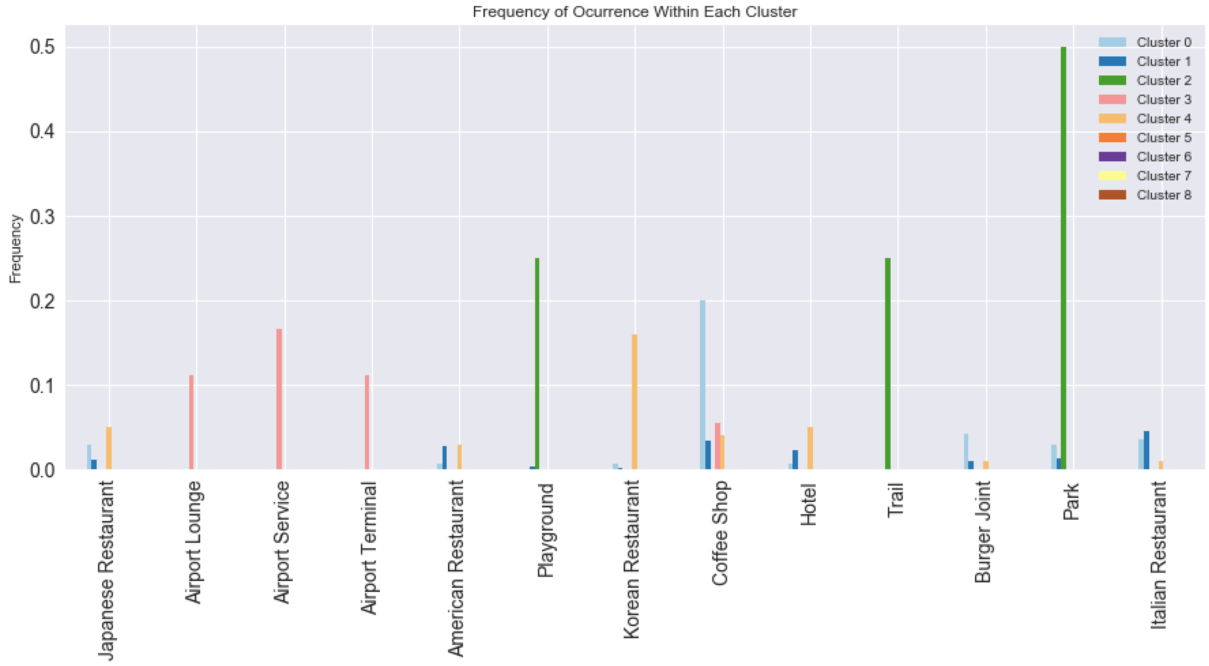
When we apply cluster algorithm, we have to choose the number of clusters, we determine it by looking different K's and their MSE's . As it is expected, the MSE decreases over the number of clusters. The elbow method here is implemented in order to select the appropriate number of groups. We chose K = 9. It is also possible to choose K = 5 or K= 6, because dealing with small number of clusters is easy but we prefer more accurate results.

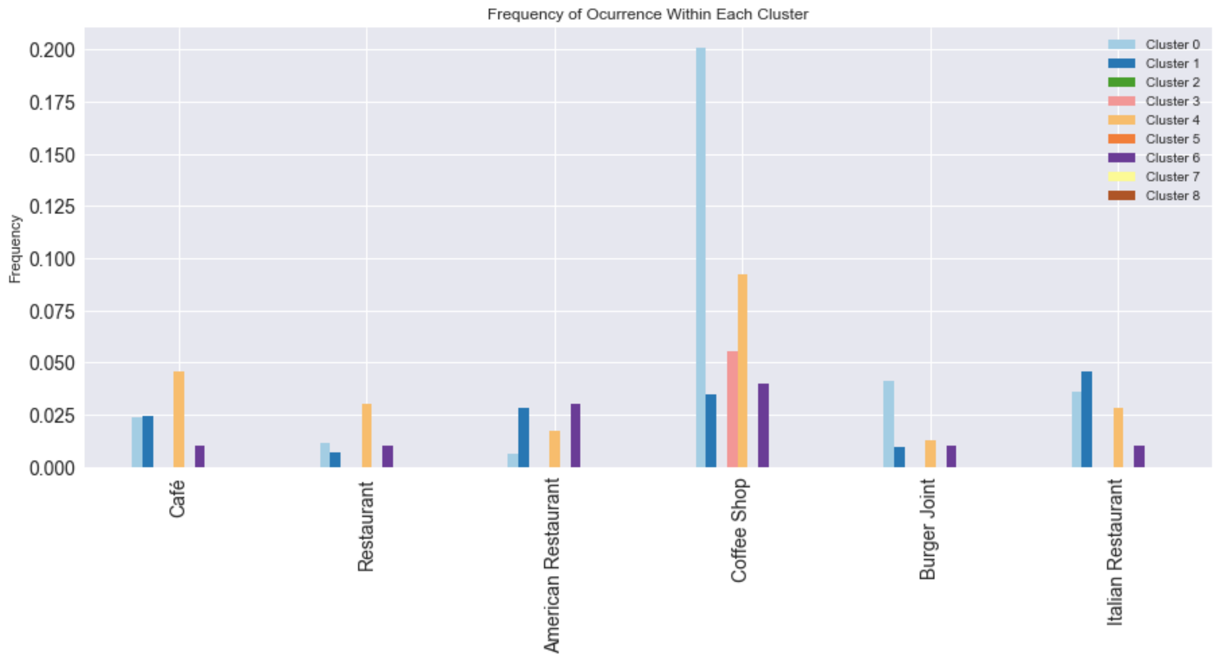


On the maps above, it is possible to see which neighbourhoods within Manhattan and New York are similar. Each colour represents the cluster for which that neighbourhood belongs. In the waffle chart below, the proportion of the neighbourhoods assigned to each cluster can be seen. For this reason a waffle chart is implemented. There are 4 major clusters and 5 minor clusters. I have attached the list of neighbourhoods and most common venues in the appendix.



We construct bar charts to see the frequency of the venues in clusters. It can be easily seen that Cluster 0, 6 and 4 have a vibrant city atmosphere since they have many cafes and restaurants. Cluster 2 has many green ares like parks and playgrounds. Cluster 3 has a neighbourhood which has an airport so there are many venues related with that. The bar charts are given below.





5. Discussion

This search can be used to see the differences and similarities between Manhattan and Downtown Toronto in terms of venue types and frequencies. We have limited amount of data, so we could not discuss the living conditions and living costs in these two cities. We need more data sets about these cities to make more meaningful decisions. This search on the other hand, still captures the similar neighbourhoods. For example, it segments neighbourhoods with parks and it classifies the neighbourhood with airports. Another evidence is that it puts the top similar neighbourhoods in one cluster like Tribeca and Soho. We can say that our algorithm and techniques are working well.

6. Conclusion

In this search, a segmentation of neighbourhoods of two different countries is done. This clustering involves the neighbourhoods in Manhattan, New York and the neighbourhoods in Downtown Toronto. We use the Foursquare API for venues. One Hot Encoding is used for converting the categories of the venues into a feature matrix. The features are the frequency of occurrence from each category in a neighbourhood.

We apply the K-Means clustering algorithm to find similarities between all the neighbourhoods. The elbow method is used for selecting the appropriate number of clusters. We choose K= 9.
Apparently we can say that Cluster 0, Cluster 4 and Cluster 6 have many coffee shops, different types of restaurants. It seems that they have neighbourhoods that are the hearts of these cities. Cluster 2 is dominant in terms of parks and playgrounds. Cluster 3 has neighbourhood with airport. Clusters 1, 2, 3 and 5 have only one neighbourhoods. Finally, my customer who wants me to compare Manhattan and Toronto can use this analysis to get a notion about the neighbourhoods.

APPENDIX

Cluster 0

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Manhattanville	Coffee Shop	Deli / Bodega	Mexican Restaurant	Seafood Restaurant	Italian Restaurant
East Harlem	Mexican Restaurant	Bakery	Latin American Restaurant	Thai Restaurant	Deli / Bodega
Upper East Side	Art Gallery	Italian Restaurant	Exhibit	Bakery	Gym / Fitness Center
Lincoln Square	Theater	Plaza	Café	Italian Restaurant	Concert Hall
Clinton	Theater	Gym / Fitness Center	Italian Restaurant	American Restaurant	Coffee Shop
Tribeca	American Restaurant	Italian Restaurant	Park	Café	Men's Store
Soho	Clothing Store	Boutique	Art Gallery	Women's Store	Men's Store
Battery Park City	Park	Coffee Shop	Hotel	Gym	Memorial Site
Carnegie Hill	Coffee Shop	Café	Pizza Place	Gym	Cosmetics Shop
Noho	Italian Restaurant	Hotel	Cocktail Bar	Pizza Place	French Restaurant
Civic Center	Coffee Shop	Gym / Fitness Center	Hotel	French Restaurant	Spa
Stuyvesant Town	Park	Boat or Ferry	Bar	Baseball Field	German Restaurant
Rosedale	Park	Playground	Trail	Yoga Studio	Dry Cleaner

Cabbagetown, St. James Town	Bakery	Coffee Shop	Italian Restaurant	Pizza Place	Café
Berczy Park	Coffee Shop	Seafood Restaurant	Cheese Shop	Beer Bar	Steakhouse
Design Exchange, Toronto Dominion Centre	Coffee Shop	Café	Hotel	Restaurant	Italian Restaurant
Commerce Court, Victoria Hotel	Coffee Shop	Café	Restaurant	Hotel	Gym
CN Tower, Bathurst Quay, Island airport, Harbo...	Airport Service	Airport Terminal	Airport Lounge	Airport Gate	Harbor / Marina
Stn A PO Boxes 25 The Esplanade	Coffee Shop	Café	Hotel	Restaurant	Seafood Restaurant
Christie	Grocery Store	Café	Park	Restaurant	Coffee Shop

Cluster 1

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Inwood	Mexican Restaurant	Lounge	Restaurant	Pizza Place	Café

Cluster 2

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Flatiron	Yoga Studio	American Restaurant	Japanese Restaurant	Café	Gym / Fitness Center

Cluster 3

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Ryerson, Garden District	Coffee Shop	Clothing Store	Café	Japanese Restaurant	Cosmetics Shop

Cluster 4

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Central Harlem	Cosmetics Shop	Chinese Restaurant	French Restaurant	Bar	American Restaurant
Church and Wellesley	Coffee Shop	Japanese Restaurant	Gay Bar	Sushi Restaurant	Restaurant

Cluster 5

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Roosevelt Island	Park	Sandwich Place	Deli / Bodega	Coffee Shop	Scenic Lookout

Cluster 6

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Marble Hill	Coffee Shop	Gym	Sandwich Place	Yoga Studio	Shopping Mall
Washington Heights	Café	Bakery	Grocery Store	Deli / Bodega	Mobile Phone Shop

Upper West Side	Italian Restaurant	Wine Bar	Bar	Coffee Shop	Mediterranean Restaurant
Midtown	Hotel	Sporting Goods Shop	Clothing Store	Coffee Shop	Bakery
Murray Hill	Sandwich Place	Coffee Shop	American Restaurant	Gym / Fitness Center	Japanese Restaurant
Lower East Side	Café	Pizza Place	Coffee Shop	Chinese Restaurant	Art Gallery
Gramercy	Italian Restaurant	Pizza Place	Mexican Restaurant	Bagel Shop	Thai Restaurant
Turtle Bay	Italian Restaurant	Coffee Shop	Sushi Restaurant	Steakhouse	Wine Bar
St. James Town	Coffee Shop	Café	Restaurant	Clothing Store	Diner
Central Bay Street	Coffee Shop	Italian Restaurant	Ice Cream Shop	Sandwich Place	Japanese Restaurant
Adelaide, King, Richmond	Coffee Shop	Café	Bar	Thai Restaurant	Cosmetics Shop
Chinatown, Grange Park, Kensington Market	Bar	Café	Vietnamese Restaurant	Coffee Shop	Chinese Restaurant

Cluster 7

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Hamilton Heights	Pizza Place	Café	Coffee Shop	Mexican Restaurant	Deli / Bodega
Yorkville	Italian Restaurant	Bar	Gym	Coffee Shop	Deli / Bodega
Lenox Hill	Coffee Shop	Italian Restaurant	Sushi Restaurant	Pizza Place	Cocktail Bar
Chelsea	Coffee Shop	Italian Restaurant	Bakery	Ice Cream Shop	Hotel

Greenwich Village	Italian Restaurant	Sushi Restaurant	Clothing Store	Café	Indian Restaurant
Little Italy	Bakery	Café	Salon / Barbershop	Cocktail Bar	Sandwich Place
West Village	Italian Restaurant	New American Restaurant	Cosmetics Shop	Park	Wine Bar
Manhattan Valley	Bar	Indian Restaurant	Pizza Place	Coffee Shop	Mexican Restaurant
Financial District	Coffee Shop	American Restaurant	Bar	Hotel	Pizza Place
Midtown South	Korean Restaurant	Hotel	Japanese Restaurant	Hotel Bar	Coffee Shop
Sutton Place	Gym / Fitness Center	Italian Restaurant	Furniture / Home Store	Coffee Shop	Gym
Harbord, University of Toronto	Café	Bookstore	Restaurant	Bar	Japanese Restaurant
First Canadian Place, Underground city	Coffee Shop	Café	Restaurant	Asian Restaurant	Gym
Queen's Park	Coffee Shop	Burger Joint	Park	Gym	Yoga Studio

Cluster 8

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Chinatown	Chinese Restaurant	Cocktail Bar	American Restaurant	Vietnamese Restaurant	Hotpot Restaurant
East Village	Bar	Ice Cream Shop	Cocktail Bar	Wine Bar	Chinese Restaurant
Morningside Heights	Park	Coffee Shop	American Restaurant	Bookstore	Burger Joint

Tudor City	Mexican Restaurant	Park	Café	Pizza Place	Diner
Hudson Yards	American Restaurant	Hotel	Gym / Fitness Center	Italian Restaurant	Café
Harbourfront	Coffee Shop	Pub	Bakery	Park	Theater
Harbourfront East, Toronto Islands, Union Station	Coffee Shop	Aquarium	Hotel	Italian Restaurant	Café