Battle of the Neighborhoods

Applied data science capstone project

Introduction / Business Problem

- Mr. Smith is looking for a house in Toronto, Canada as he will be starting to work there soon.
- There are some criteria of the neighborhood that need to be met before Mr. Smith decides where he wants to settle-down.
- As there are so many neighborhoods around Toronto City, he wishes to narrow down the scope. This means coming out with a list of neighborhoods that fit his preferences best and he can filter out from there on.

Introduction / Business Problem – continued

- He prefers to stay in a neighborhood that has the following:
  1. Ease of assess to public transport
  2. Close to schools for his children
  3. Close to supermarkets for regular grocery shopping
  4. A park or a garden

Introduction / Business Problem – continued

- He does not like to stay in a neighborhood with the following:
  1. A bar or a pub (He rarely drinks).
  2. Restaurants or coffee shops.
  3. Hotel
  4. Event halls

Data Requirements

The data analysis via K-means clustering algorithm will require the following:

1. The postal codes, boroughs and neighborhoods in Toronto, Canada
   - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
   - All postal codes start with M are located within the city of Toronto
   - For example:

Postal Code	Borough	Neighborhood
M3A	North York	Parkwoods

Data Requirements

The data analysis via K-means clustering algorithm will require the following:

2. Geospatial Data
    - http://cocl.us/Geospatial_data
    - The postal codes found from Wikipedia will be matched with the corresponding latitude and longitudes tabulated in this csv file
    - For example:

Postal Code	Latitude	Longitude
M3A	43.7532586	-79.3296565

Data Requirements – Continued

The data analysis via K-means clustering algorithm will require the following:

3. Venues available within 1 km from the center of neighborhood
    - Extracted from Foursquare
    - Venue category used to determine the type of venue located 1 km radius of the neighborhood.
    - For example: A bar, a supermarket and a school within the defined radius.
    - The frequency of each venue category can be calculated.

Methodology

1. Data cleaning is performed to remove postal codes that do not contain any boroughs or neighborhoods. No houses are there so these data have to be removed.
2. The longitudes and latitudes of each postal code is determined and link to each postal codes.
3. The longitudes and latitudes of each postal code is important for Foursquare API to determine what venues are located within a specified radius (1000 m) of the neighborhood.
4. Foursquare API is used to extract the 100 venues located around neighborhood. For example, Tim Hortons in Parkwoods.

Methodology – Continued

5. Venue categories of each venue is determined and the mean frequency of each venue category in the neighborhood is calculated.
6. The ten most common and ten least common venue category is determined.
7. The data is fed to the K-means algorithm to cluster the neighborhood based on the five most and five least common venue category.
8. Iteration is performed to determine the value of K that can differentiate the neighborhood based on the data best.
9. Based on the best K value determined, results of the K-means clustering is analyzed to see which cluster contains neighborhoods that fit to Mr. Smith's preferences.

Results – Cluster 1

- Looks like a food heaven here, full of various restaurants. Definitely not a place where Mr. Smith wants to stay.
- A total of 20 boroughs.

	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
3	North York	0.0	Fast Food Restaurant	Coffee Shop	Restaurant	Accessories Store	Dessert Shop	Vietnamese Restaurant	Fried Chicken Joint	Sushi Restaurant
6	Scarborough	0.0	Fast Food Restaurant	Trail	Coffee Shop	Hobby Shop	Restaurant	Bus Station	Supermarket	Caribbean Restaurant
7	North York	0.0	Restaurant	Coffee Shop	Japanese Restaurant	Asian Restaurant	Burger Joint	Gym	Supermarket	Bank
10	North York	0.0	Grocery Store	Fast Food Restaurant	Gym	Pizza Place	Gas Station	Park	Coffee Shop	Bank
13	North York	0.0	Restaurant	Coffee Shop	Japanese Restaurant	Asian Restaurant	Burger Joint	Gym	Supermarket	Bank
16	York	0.0	Convenience Store	Pizza Place	Coffee Shop	Sushi Restaurant	Grocery Store	Sandwich Place	Bakery	Field

Note: Only 6 boroughs are shown here for conciseness.

Results – Cluster 2

- Cluster full of coffee shops, not to Mr. Smith preferences too.
- A total of 42 boroughs.

	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
2	Downtown Toronto	1.0	Coffee Shop	Theater	Diner	Café	Park	Pub	Breakfast Spot	Italian Restaurant
4	Downtown Toronto	1.0	Coffee Shop	Park	Sushi Restaurant	Italian Restaurant	Japanese Restaurant	Gastropub	Ramen Restaurant	Restaurant
9	Downtown Toronto	1.0	Coffee Shop	Gastropub	Japanese Restaurant	Café	Theater	Italian Restaurant	Restaurant	Plaza
15	Downtown Toronto	1.0	Coffee Shop	Café	Restaurant	Gastropub	Seafood Restaurant	Hotel	Italian Restaurant	Theater
19	East Toronto	1.0	Pub	Coffee Shop	Pizza Place	Breakfast Spot	Bakery	Beach	Japanese Restaurant	Burger Joint
20	Downtown Toronto	1.0	Coffee Shop	Café	Hotel	Japanese Restaurant	Restaurant	Seafood Restaurant	Park	Gastropub

Note: Only 6 boroughs are shown here for conciseness.

Results – Cluster 3

- Cluster 3 has less coffee shops and restaurants than Cluster 1 and 2.
- It also has a park but it does not fit much of Mr. Smith preferences.
- A total of 2 boroughs.

	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
12	Scarborough	2.0	Breakfast Spot	Park	Playground	Burger Joint	Italian Restaurant	Yoga Studio	Event Space	Dumpling Restaurant
101	Etobicoke	2.0	Park	Italian Restaurant	Eastern European Restaurant	Ice Cream Shop	Gym / Fitness Center	Ethiopian Restaurant	Donut Shop	Dry Cleaner

Results – Cluster 4

- Cluster 4 looks like the best for now! Most has a convenience store, park, and even a bus stop!
- Mr. Smith can start working on his screening process from this cluster group.
- A total of 35 boroughs.

	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
0	North York	3.0	Park	Convenience Store	Pharmacy	Shopping Mall	Bus Stop	Food & Drink Shop	Laundry Service	Caribbean Restaurant
1	North York	3.0	Coffee Shop	Boxing Gym	Hockey Arena	Sporting Goods Shop	Park	Portuguese Restaurant	Pizza Place	Golf Course
5	Etobicoke	3.0	Pharmacy	Grocery Store	Skating Rink	Bank	Convenience Store	Playground	Bakery	Shopping Mall
8	East York	3.0	Pizza Place	Brewery	Fast Food Restaurant	Athletics & Sports	Soccer Stadium	Bus Line	Breakfast Spot	Café
11	Etobicoke	3.0	Park	Hotel	Pizza Place	Theater	Restaurant	Bank	Fish & Chips Shop	Clothing Store
14	East York	3.0	Coffee Shop	Park	Pizza Place	Sandwich Place	Thai Restaurant	Cosmetics Shop	Athletics & Sports	Bus Line

Note: Only 6 boroughs are shown here for conciseness.

Results – Cluster 5

- Cluster 5 only has the park that meets Mr. Smith's requirements.
- Not the best cluster! But still better than a neighborhood full of coffee shops and restaurants (i.e. Cluster 1 and 2).
- Total of 1 borough.

	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
45	North York	5.0	Park	Pool	Yoga Studio	Falafel Restaurant	Dry Cleaner	Dumpling Restaurant	Eastern European Restaurant	Electronics Store

Results – Cluster 6

- Nothing in here fits to Mr. Smith's preferences. He does not want a neighborhood common with hotels!
- Definitely not his preference.
- Total of 1 borough.

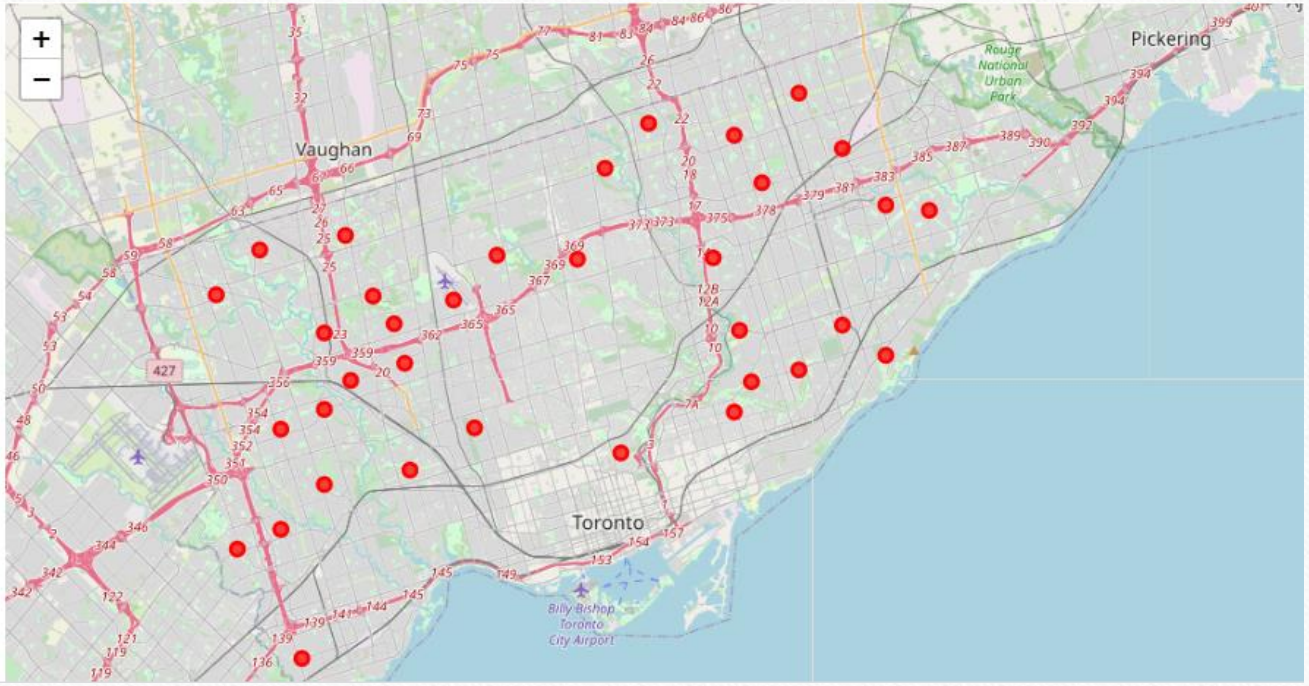
	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
94	Etobicoke	6.0	Hotel	Rental Car Location	Coffee Shop	Yoga Studio	Falafel Restaurant	Dry Cleaner	Dumpling Restaurant	Eastern European Restaurant

Discussion

- K-means clustering has been performed for the neighborhoods in Toronto based on the most and least common venue categories.
- Out of the 103 neighborhoods in Toronto, the neighborhoods have been clustered into 6 clusters. Each with different attributes.
- The results show that Cluster 4 neighborhoods fit best to Mr. Smith preferences. This include the presence of bus stop (public transport amenities) and convenience stores or shopping centres. Some has even a park!

Discussion – Continued

Mr. Smith can choose one of the 35 neighborhoods as his place to stay. These 35 neighborhoods have been pointed in the Toronto city map on the right.



Conclusion

- Mr. Smith can start further screening of the neighborhoods from Cluster 4. The clustering process has reduced the total of 103 neighborhoods in Toronto to only 35 for him to work on.
- The clustering process has definitely reduce his time and effort to look at the other neighborhoods that definitely do not fit into his preferences.
- Further analysis can be performed on Cluster 4, using other attributes such as average house prices and crime index to determine the best one for Mr. Smith.