Applied Data Science Capstone – IBM Coursera

By

Sanujit Senapati

The Battle of Neighborhoods

Table of contents

Introduction: Business Problem

A real estate development and investment company is trying to identify and shortlist retail opportunities in the Greater Toronto area based on trends and popularity. The company realizes the importance and relevance of social media in understanding the pulse of the market and seeks to use data as a key driver in decision making.

How can the company use social trends to select popular venues, understand and identify characteristics of the venues, and select new locations with similar characteristics which would have high growth potential?

In this study, as a Data Scientist, I provide a point of view of how data can be acquired, cleansed, curated and analyzed through machine learning technique to better drive the decision-making process.

Data

1. Toronto neighborhoods data from Wikipedia via https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, which includes the Postal Codes, Boroughs and Neighborhood in the Toronto area

2. Geo codes for each postal code above from Cognitive Class via http://cocl.us/Geospatial_data

3. Foursquare Places API for Venues – Regular endpoints include basic venue firmographic data, category, and ID. For the analysis, I have used the “explore” Regular API endpoint to get venue recommendations via https://developer.foursquare.com/docs/venues/explore.

Methodology

Neighborhood Candidate Selection

- Neighborhood and Geo Code data acquisition for Wikipedia and Cognitive Class
- Data merging, shaping and transformation
- Fetch coordinates using OpenStreetMap Nominatim API
- Visualization of neighborhoods using Folium
- Augment venue data from Foursquare Place API with the neighborhood

Exploratory Data Analysis

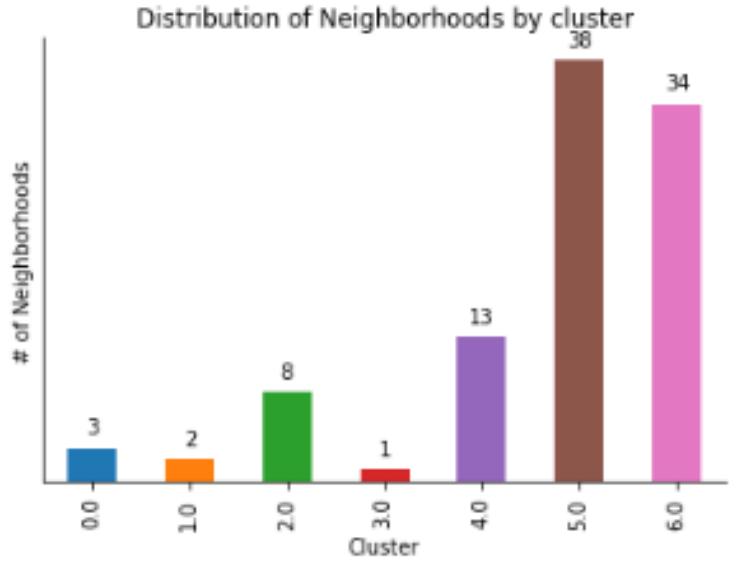
- Generate statistics such as:
  - Venue counts by neighborhood
  - Top 20 neighborhoods
- Visualize data by venue counts, top categories

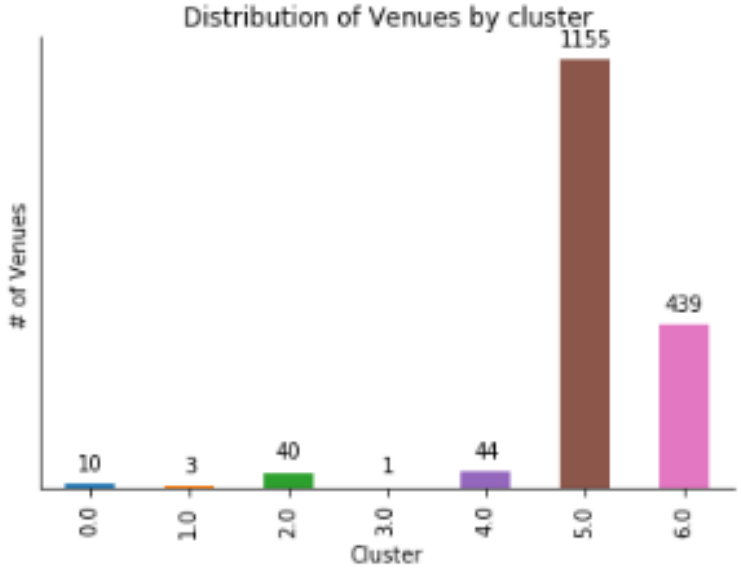
Analysis (Machine Learning)

- Use unsupervised learning – clustering to group neighborhoods and venue data
- Leverage k-means clustering to fit the data, define k based on the number of neighborhoods
- Combine cluster labels with source data
- Visualization of clusters overlaid on a map of Toronto
- Analyze cluster characteristics

Results

Distribution of Neighborhoods and Venues by Cluster





Results

Most Common Venues by Cluster

	Cluster Labels	Venue Category	Count
0	0.0	Bar	3
8	1.0	Baseball Field	2
26	2.0	Pizza Place	9
32	3.0	Garden	1
46	4.0	Park	15
99	5.0	Coffee Shop	105
309	6.0	Fast Food Restaurant	20

Results

Cluster 0 has 3 neighborhoods which are geographically apart (Scarborough on the east, North York in the northcentral, and Etobicoke) with a suburban flavor, bars, coffee shops, open areas like golf course and dog run, and women's stores.

	Borough	Neighborhood	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
1	Scarborough	Highland Creek, Rouge Hill, Port Union	0.0	Bar	Golf Course	History Museum	Women's Store	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Dog Run
29	North York	Northwood Park, York University	0.0	Coffee Shop	Massage Studio	Caribbean Restaurant	Bar	Women's Store	Dive Bar	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store
102	Etobicoke	Northwest	0.0	Drugstore	Bar	Rental Car Location	Women's Store	Dance Studio	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Dog Run	Dive Bar

Cluster 1 has 2 neighborhoods which are in the same boroughs as cluster 0, and with similar characteristics. While outdoor spaces like Parks and Golf Course are a common thread, the key difference is the Baseball Field which is a common thread in this cluster.

	Borough	Neighborhood	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
91	Etobicoke	Humber Bay, King's Mill Park, Kingsway Park So...	1.0	Baseball Field	Construction & Landscaping	Women's Store	Deli / Bodega	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run
97	North York	Emery, Humberlea	1.0	Baseball Field	Women's Store	Event Space	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run	Dive Bar

Results

Cluster 2 has 8 neighborhoods that are geographically apart like cluster 0. Pizza Place and Grocery Stores are the most common venues.

	Borough	Neighborhood	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 I Com V
13	Scarborough	Clarks Corners, Sullivan, Tam O'Shanter	2.0	Pizza Place	Pharmacy	Fast Food Restaurant	Italian Restaurant	Thai Restaurant	Fried Chicken Joint	Chinese Restaurant	Noodle House	Discount Store	Curlin
24	North York	Willowdale West	2.0	Grocery Store	Coffee Shop	Butcher	Pharmacy	Pizza Place	College Stadium	Comfort Food Restaurant	Electronics Store	Eastern European Restaurant	Dum Resta
31	North York	Downsview West	2.0	Grocery Store	Bank	Hotel	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Dive Bar	D S
81	York	The Junction North, Runnymede	2.0	Grocery Store	Convenience Store	Bus Line	Pizza Place	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run	D S
89	Etobicoke	Alderwood, Long Branch	2.0	Pizza Place	Pharmacy	Coffee Shop	Pub	Sandwich Place	Skating Rink	Gym	Gastropub	Garden Center	Disc
94	Etobicoke	Cloverdale, Islington, Martin Grove, Princess ...	2.0	Bank	Women's Store	Event Space	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run	Dive
96	North York	Humber Summit	2.0	Empanada Restaurant	Pizza Place	Curling Ice	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run	Dive Bar	Disc
99	Etobicoke	Westmount	2.0	Pizza Place	Playground	Middle Eastern Restaurant	Sandwich Place	Chinese Restaurant	Coffee Shop	Dog Run	Dive Bar	Discount Store	

Cluster 3 has 1 neighborhood which indicates characteristics that are unique to the venue categories. It also indicates that the neighborhood is small and does not have enough venues to fit other clusters.

	Borough	Neighborhood	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
63	Central Toronto	Roselawn	3.0	Garden	Women's Store	Dance Studio	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run	Dive Bar

Results

Cluster 4 has 13 neighborhoods which cover most of the boroughs. A common thread is the number of parks, playgrounds, and trails around the Don River Valley. Access to public transportation such as Bus Line is a key feature. Restaurants serving multi-cultural cuisine is common in this cluster.

	Borough	Neighborhood	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
14	Scarborough	Agincourt North, L'Amoreaux East, Milliken, St...	4.0	Playground	Park	Curling Ice	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run	Dive Bar
23	North York	York Mills West	4.0	Park	Bank	Women's Store	Dance Studio	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run
25	North York	Parkwoods	4.0	Bus Stop	Food & Drink Shop	Fast Food Restaurant	Park	Diner	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant
30	North York	CFB Toronto, Downsview East	4.0	Airport	Park	Bus Stop	Women's Store	Discount Store	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant
40	East York	East Toronto	4.0	Convenience Store	Coffee Shop	Park	Women's Store	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store
44	Central Toronto	Lawrence Park	4.0	Bus Line	Park	Swim School	Discount Store	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Diner
50	Downtown Toronto	Rosedale	4.0	Park	Playground	Trail	Diner	Dance Studio	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant
64	Central Toronto	Forest Hill North, Forest Hill West	4.0	Bus Line	Trail	Park	Sushi Restaurant	Jewelry Store	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant
74	York	Caledonia-Fairbanks	4.0	Park	Women's Store	Market	Pharmacy	Fast Food Restaurant	Diner	Deli / Bodega	Department Store	Dessert Shop
79	North York	Maple Leaf Park, North Park, Upwood Park	4.0	Park	Basketball Court	Construction & Landscaping	Bakery	Women's Store	Dog Run	Dim Sum Restaurant	Diner	Discount Store
90	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	4.0	River	Pool	Park	Women's Store	Dim Sum Restaurant	Dance Studio	Deli / Bodega	Department Store	Dessert Shop
98	York	Weston	4.0	Park	Convenience Store	Women's Store	Dance Studio	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run
100	Etobicoke	Kingsview Village, Martin Grove Gardens, Richv...	4.0	Pizza Place	Park	Bus Line	Dance Studio	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run

Results

Cluster 5 is the largest cluster with 38 neighborhoods and includes 1155 venues. It covers most of the boroughs. Although the most common venue is Coffee Shops and Cafes, this cluster has a wide coverage of restaurants serving multi-cultural cuisine.

	Borough	Neighborhood	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
3	Scarborough	Woburn	5.0	Coffee Shop	Korean Restaurant	Women's Store	Dance Studio	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run	Dive Bar
6	Scarborough	East Birchmount Park, Ionview, Kennedy Park	5.0	Discount Store	Hobby Shop	Coffee Shop	Department Store	Dive Bar	Deli / Bodega	Dessert Shop	Dim Sum Restaurant	Diner	Women's Store
8	Scarborough	Clifforest, Cliffside, Scarborough Village West	5.0	Movie Theater	American Restaurant	Motel	Women's Store	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run	Dive Bar
9	Scarborough	Birch Cliff, Cliffside West	5.0	College Stadium	Café	Skating Rink	General Entertainment	Women's Store	Discount Store	Department Store	Dessert Shop	Dim Sum Restaurant	Diner

Cluster 6 is the second largest cluster with 34 neighborhoods and includes 439 venues. Like cluster 5, it covers most of the boroughs. The most common venues is Fast Food Restaurant. Apart from restaurants, this cluster includes a wide range of retails outlets.

	Borough	Neighborhood	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	Scarborough	Rouge, Malvern	6.0	Fast Food Restaurant	Print Shop	Women's Store	Curling Ice	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run	Dive Bar	Discount Store
2	Scarborough	Guildwood, Morningside, West Hill	6.0	Electronics Store	Rental Car Location	Spa	Medical Center	Breakfast Spot	Mexican Restaurant	Pizza Place	Concert Hall	Comfort Food Restaurant	College Gym
4	Scarborough	Cedarbrae	6.0	Athletics & Sports	Lounge	Hakka Restaurant	Fried Chicken Joint	Thai Restaurant	Bakery	Bank	Caribbean Restaurant	Diner	Dim Sum Restaurant
5	Scarborough	Scarborough Village	6.0	Playground	Convenience Store	Dance Studio	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Dog Run	Dive Bar

Discussion

Observations	Predominance of 2 clusters across the neighborhoods and venue categories which indicates similarity or commonality of features. The remaining 4 clusters had more distinguishing or unique features.
	The k-means clustering approach relied of frequency of a category across the 255 unique categories. The feature set may be large compared to the number of samples, i.e. number of neighborhoods (99).
	I tried multiple values of k in the k-means clustering. For lower values of k, the larger clusters coalesced into a single cluster. For higher values of k, the number of smaller clusters increased but the larger clusters did not break up noticeably any further.
	The Foursquare data is primarily social and is crowdsourced. I noticed the API calls returned slightly different data sets when executed at various times of the day or day of the week.
Recommendations	I had planned initially to use the Premium Endpoint to fetch ratings but was unable to because of the daily limits of API calls. This extended data could have provided a social dimension, but the data would change frequently.
	Running the analysis and comparing results over a period as opposed to a snapshot would stabilize the findings.
	Consider other unsupervised learning methods for comparative analysis.
	Augment demographic data for neighborhoods to get additional insights.

Conclusion

In conclusion, this study was a positive step for the stakeholders to understand how data from various sources can be used via powerful tools and visualization techniques to derive insights.

From a personal perspective, it provided me with exposure to the data science methodology from a business problem, analysis, data acquisition, preparation, feature selection, model creation, train/fit and test/analyze results. The libraries for data acquisition, preparation, and visualization demonstrated the value of data science.