Toronto Venue Cluster Analysis

Author: Camilo Hoyos

Date: 2021-02-07

1. Introduction

1.1. Business Problem

Commercial companies are looking for a better set of data to identify where they should be looking to purchase space for expansion. With limited data, the ask is to use existing public data to identify which area is best to support their business needs.

1.2. Audience

Those interested in setting up a commercial location in Toronto would be interested in this project. It provides a density based cluster algorithm to identify where general commercial stores are located. Within each of these clusters, it provides the top five types of venues in that cluster. This can assist with identifying where to expand to disrupt competitors, or where to expand to capture an untapped market.

2. Data Review

2.1. Data

The data has two primary sources. First the identification of neighborhoods and boroughs uses Wikipedia's list of postal codes for Toronto. This data is not explicitly used for clustering, however, it does provide better identification when viewing the cluster map to identify the residing borough and neighborhood. It helps provide some of the initial data setup before clustering by providing longitude and latitude coordinates of postal codes. The primary source is using Foursquare's explore API. Foursquare mentions on their about page that "We help leading global companies tap into this intelligence to create better customer experiences and smarter business outcomes…" (*Our story*). By utilizing Foursquare's explore API, we can identify venues in Toronto by latitude and longitude. The API also provides us with the type of venue which allows for determining which venues are seen in which clusters. Lastly, there is a tertiary source used for Borough and Neighbourhood longitude and latitude values. This data is sourced from Cousera as available sources are limited.

3. Methodology

3.1. Web Scrape & Postal Code Coordinates

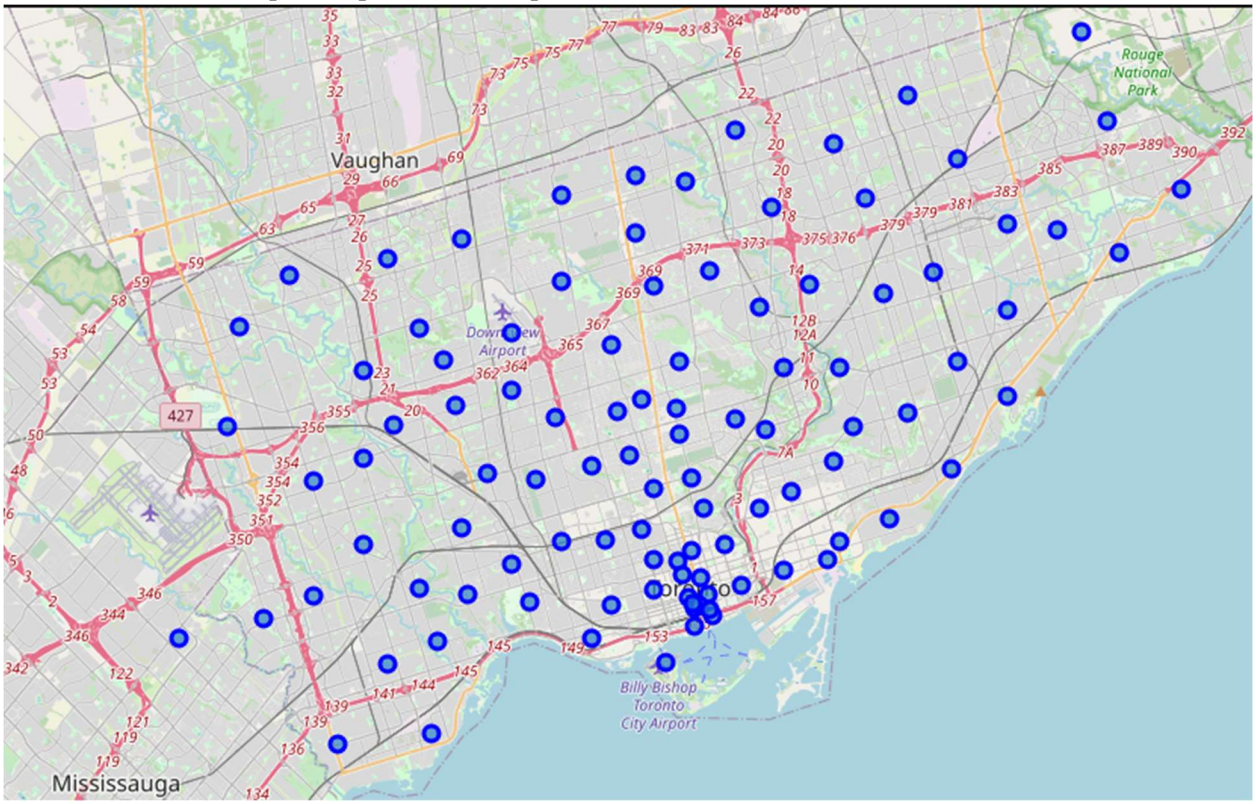
The first objective is to consume the Wikipedia page to gather a list of valid boroughs and neighborhoods. We use BeautifulSoup4 to gather the HTML page and target the specific table. Using a simple loop to read through the table grants a starting point. For data cleansing, all empty entries of borough or neighborhood are removed to only gather groups that have proper naming. Having cleaned up the table the longitude and latitude coordinates are joined and added to create a data frame with postal code, borough, neighborhood, latitude, and longitude. We end up with a total of 103 records.

[6]:

	Postal Code	Borough	Neighbourhood	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
...	...	...	...	...	...
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944
99	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
100	M7Y	East Toronto	Business reply mail Processing Centre, South C...	43.662744	-79.321558
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999

103 rows × 5 columns

With this list, we then plot all points on a map of Toronto to review.

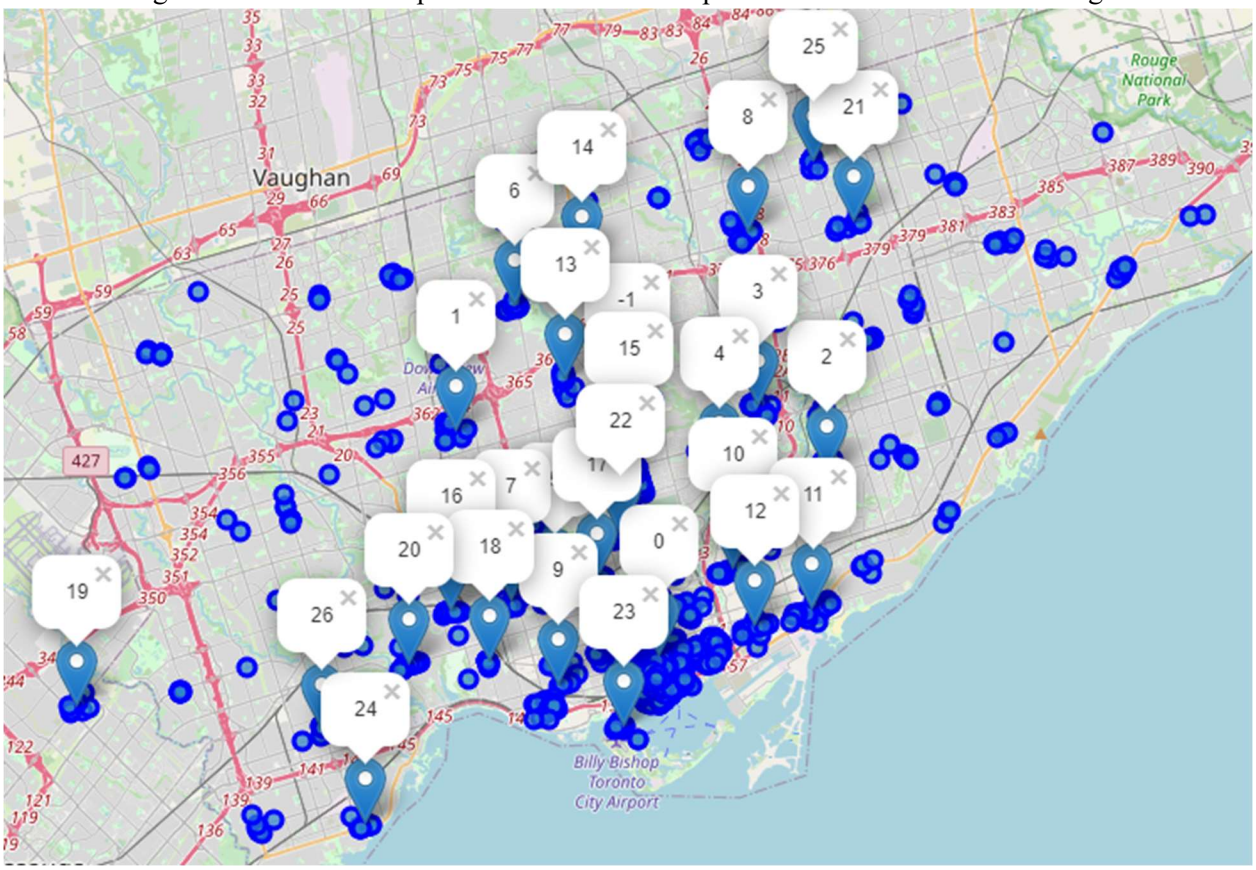


This provides a first glimpse into how boroughs and neighborhoods are clustered and spread through Toronto.

3.2. Foursquare API & DBSCAN

With the initial list of boroughs and neighborhoods, we utilize Foursquare's API to gather venues around each of these points. At the time of this report, we receive2098 records. We then use a destiny-based clustering algorithm known as 'DBSCAN'. DBSCAN is best for spatial data which is the primary need in identifying clusters within Toronto. We utilize the longitude and latitude coordinates from each of

the venues to identify clusters. DBSCAN will provide one outlier cluster known as '-1' and all other clusters will be identified as actual clusters. At the time of this report, 27 clusters were found and reported based on the density of venues in Toronto. The mean of the longitudes and latitudes provide a focal point on numbering the cluster. We then plot this on the same map of Toronto to review the findings.



Finally, with clusters identified and venues aggregated, the objective is to identify what the top five venue categories are in each cluster. We simply one-hot encode to count the types of venues in each cluster, sum the categories, and apply a descending sort. A brief review of the results:

[23]:

	Clus_Db	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	-1	Park	Pizza Place	Coffee Shop	Bakery	Grocery Store
1	0	Coffee Shop	Café	Hotel	Restaurant	Japanese Restaurant
2	1	Clothing Store	Furniture / Home Store	Accessories Store	Coffee Shop	Miscellaneous Shop
3	2	Pizza Place	Gym / Fitness Center	Flea Market	Bank	Intersection
4	3	Gym	Restaurant	Beer Store	Coffee Shop	Art Gallery

4. Results

4.1. Top Cluster Review

The largest cluster contains a coffee shop as the highest venue, followed directly by a cafe. We then see a hotel in the third position and then followed by two restaurants where Japanese is specifically called out in the fifth position. It is clear that the first cluster has a high count of coffee related locations, restaurants with a skew towards japanese restaurants, and where individuals stay when they travel.

	Clus_Db	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	0	Coffee Shop	Café	Hotel	Restaurant	Japanese Restaurant

4.2. Full Results

Some notable trends are that locations that provide coffee take the top common venue spot with a count of 6, and pizza is the runner up with a count of four.

Clus_Db	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Coffee Shop	Café	Hotel	Restaurant	Japanese Restaurant
1	Clothing Store	Furniture / Home Store	Accessories Store	Coffee Shop	Miscellaneous Shop
2	Pizza Place	Gym / Fitness Center	Flea Market	Bank	Intersection
3	Gym	Restaurant	Beer Store	Coffee Shop	Art Gallery
4	Coffee Shop	Sporting Goods Shop	Burger Joint	Bank	Restaurant
5	Grocery Store	Café	Park	Baby Store	Coffee Shop
6	Bank	Coffee Shop	Pet Store	Restaurant	Middle Eastern Restaurant
7	Bakery	Pharmacy	Coffee Shop	Furniture / Home Store	Bar
8	Clothing Store	Coffee Shop	Fast Food Restaurant	Restaurant	Japanese Restaurant
9	Bar	Café	Coffee Shop	Restaurant	Vegetarian / Vegan Restaurant
10	Greek Restaurant	Coffee Shop	Italian Restaurant	Restaurant	Furniture / Home Store
11	Fast Food Restaurant	Pizza Place	Park	Restaurant	Light Rail Station
12	Coffee Shop	Brewery	Bakery	Gastropub	American Restaurant
13	Italian Restaurant	Coffee Shop	Sandwich Place	Spa	Comfort Food Restaurant
14	Ramen Restaurant	Shopping Mall	Sandwich Place	Café	Coffee Shop
15	Pizza Place	Coffee Shop	Sandwich Place	Café	Dessert Shop
16	Mexican Restaurant	Café	Thai Restaurant	Restaurant	Flea Market
17	Sandwich Place	Café	Coffee Shop	History Museum	BBQ Joint
18	Breakfast Spot	Gift Shop	Coffee Shop	Eastern European Restaurant	Movie Theater
19	Coffee Shop	Hotel	Burrito Place	Gas Station	Fried Chicken Joint
20	Café	Coffee Shop	Sushi Restaurant	Pizza Place	Italian Restaurant
21	Pharmacy	Pizza Place	Fast Food Restaurant	Gas Station	Noodle House
22	Coffee Shop	Sushi Restaurant	Bank	Pizza Place	Bagel Shop
23	Airport Service	Airport Lounge	Boutique	Harbor / Marina	Bar
24	Pizza Place	Coffee Shop	Mexican Restaurant	Liquor Store	Fried Chicken Joint
25	Pizza Place	Coffee Shop	Fast Food Restaurant	Supermarket	Electronics Store
26	Hardware Store	Thrift / Vintage Store	Fast Food Restaurant	Burger Joint	Supplement Shop

4.3. Outlier Results

Lets not forget that DBSCAN finds items that are not in a dense area and labels them as outliers in the ‘-1’ cluster. We can see that Parks are not in densely located areas. Pizza places and coffee shops are so common that they even exist in the outlier. Finally we see bakery and grocery stores.

Clus_Db	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
-1	Park	Pizza Place	Coffee Shop	Bakery	Grocery Store

5. Discussion

5.1. Review

It is clear that coffee shops and pizza are adored in Toronto as they take the top spot in many places, and exist within the outliers cluster. It would appear as long as you open with a competitive edge such as a differentiator that one can open a coffee shop or pizza place and succeed. One should also take into account locations where hotels fall within the top five. Seeing these clusters would also gather the attention of tourists. One can see that hotels in the top five also are generally surrounded with other food venues in top visitation. Another observation is that cluster 26 does not have pizza or coffee or a general dine-in restaurant, this could be a prime location to capture an untapped market. Further analysis would be needed.

6. Conclusion

6.1. Conclusion

In conclusion, it is clear that restaurants, pizza places, and coffee shops thrive throughout Toronto. Bars, retail stores, parks, and bakeries follow up piecemeal throughout the Toronto clusters but depending on additional customer evaluation could be an untapped market in some of the more prominent and dense areas. There is also very little in terms of nightlife or art and entertainment identified in the top five venues. Toronto is easily a place to go and enjoy coffee in the morning, visit a different specialized restaurant for lunch, and close out with a pizza for dinner.

6.2. Future Analysis

This is a preliminary report, but much more can be done to draw insights. As it stands, this is using a non-Premium API call from Foursquare. Premium API calls from Foursquare also provide the number of visitations and ratings. Using more of these metrics can help identify which of these top venues thrive, as opposed to just its existence. It would be even more instrumental to identify which customers visit which types of venues more often, or if customers are often tourists visiting from out of the city. All being said, this report would only thrive as a starting point for a larger deep dive with a more robust and detailed data set.

References

Our story. (n.d.). Retrieved February 07, 2021, from https://foursquare.com/about/