The Ideal Location for a New Toronto Coffee Shop

Andrew Warrick

1. Introduction

The goal is to discover which Toronto neighborhood would be best-suited for a new coffee shop. We will use the data available from Foursquare to determine which area most has an ideal combination of a high amount of consumer activity and a low number of coffee shops (potential rivals). Whichever neighborhood can claim to possess both of these traits to the greatest degree we will determine to be the ideal new coffee shop location. This information should be highly useful to those seeking to establish a coffee shop in Toronto.

2. Data

The data we use consists of the Toronto city data (including its postal codes, boroughs, and neighborhoods), its corresponding latitude and longitude information, and the data provided from Foursquare.

Toronto postal codes, boroughs, and neighborhoods:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Toronto latitude and longitude: http://cocl.us/Geospatial_data

The Toronto Wikipedia table was converted into a csv file through an online converter which can be found here: https://wikitable2csv.ggor.de/

We then used python code to remove unassigned data, aggregate the location information by postal code, and merge this data with the latitude and longitude information.

3. Methodology

For our problem, we must determine two things: consumer activity and the existence of rival coffee shops. We determine the amount of consumer activity simply by examining the number of existing venues within a neighborhood's vicinity. When it comes to rival coffee shops, however, we decided that it is not sufficient to simply count the number of existing coffee shops in a neighborhood, but that the density of coffee shops would be a more useful metric for our problem. A neighborhood with 100 venues and 4 coffee shops would be less likely to be oversaturated by an additional coffee shop than a neighborhood with 2 coffee shops and only 4 venues.

After scraping and cleaning the data, we proceed by counting and obtaining the categories of each venue within a radius of 500 from the latitude and longitude coordinates of each neighborhood using Foursquare. We then count the number of those venues that are coffee shops for each neighborhood.

	Neighborhood	# Coffee Shops
0	Adelaide, King, Richmond	7.0
1	Agincourt	0.0
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.0
3	Albion Gardens, Beaumond Heights, Humbergate, ...	1.0
4	Alderwood, Long Branch	1.0

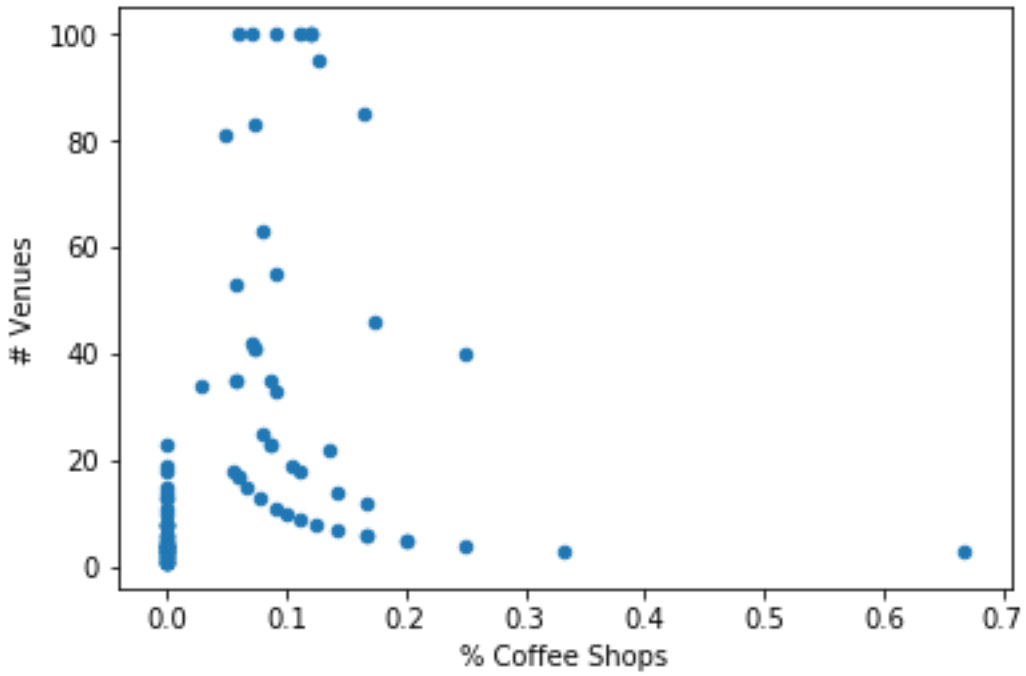
Table 1: Sample from the List of the Number of Coffee Shops per Neighborhood

Next, to determine the density of coffee shops per neighborhood, we count the total number of venues and divide the number of coffee shops by that number.

	Neighborhood	% Coffee Shops	# Venues
0	Adelaide, King, Richmond	0.070000	100.0
1	Agincourt	0.000000	5.0
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.000000	2.0
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.100000	10.0
4	Alderwood, Long Branch	0.111111	9.0

Table 2: Sample from the List of the Percentage of Venues that are Coffee Shops Compared to the Number of Venues

After this, we store the maximum values for the percentage of coffee shops and the number of venues for later analysis.



Plot 1: A Scatter Plot Comparing the Percentage of Coffee Shops vs. the Number of Venues

Finally, we developed a formula that would assign a score to each neighborhood between 0 and 1, where the neighborhood with a score closest to 1 would be the most ideal location for a new coffee shop.

$$Score = \frac{1}{2}\left(\frac{x}{\mathrm{m1}} + \frac{1}{m2}(m2 - y)\right)$$

Here, m1 represents the largest number of venues previously found, m2 represents the highest percentage of coffee shops, x is the number of venues in the neighborhood currently being examined, and y is likewise the percentage of coffee shops in that neighborhood. In our formula, the score approaches 1 as x approaches m1 and y approaches 0, and the score approaches 0 as x approaches 0 and y approaches m2.

4. Results

After calculating the score for each of our 99 neighborhoods, we determined that St. James Town would be the ideal location for a coffee shop, with a score of 0.955. Indeed, St. James Town had the maximum number of venues (100), and the percentage of those that were coffee shops was only 6%, having only 6 coffee shops.

5. Discussion

This approach successfully gave us a neighborhood that had a low saturation of coffee shops but a high amount of consumer activity. One flaw of our approach is that Foursquare appears to have a limit of 100 venues that it can obtain from any given neighborhood (at least for the free version we are using). Even though the limit of venues examined is an adjustable parameter when obtaining data from Foursquare, increasing the limit above 100 still results in a clearly artificial capping (as is apparent from the scatter plot) at 100 venues, so it appears that this is a limit set by Foursquare itself. While it is probable that the sample of venues obtained from neighborhoods with above 100 venues fairly accurately reflects the real ratio of coffee shops to total venues in those neighborhoods, not having the full number of venues opens up the possibility of skewed results if some of the unselected venues were coffee shops. It also prevents us from knowing which neighborhood truly has the highest consumer activity from the venue data.

6. Conclusion

This project demonstrates that business questions – such as where to open a new coffee shop – can be greatly informed via easily obtainable online data and simple algorithms. Going forward, it may be worthwhile to consider weighing the number of venues and the percentage of those that are coffee shops differently. In our model, the weight of those values is the same for determining our score, but outside business analysis may suggest that one is a more useful indicator for the viability of a new shop than the other, and therefore should account for a higher percentage of the final score. Additionally, the question of whether an upgraded Foursquare account could exceed the imposed limit of 100 venues should be explored.