The Optimal Location for a Coffee Shop in Toronto, CA

Using Clustering and Demographics

Calvin Todorovich

August 2020

1 Introduction

1.1 Background

Toronto is the biggest, and most populated city in all of Canada, and "a world leader in such areas as business, finance, technology, entertainment and culture," according to AboutToronto.ca. Furthermore, the average household income in Toronto is over $104,000, making it an excellent location for business. Opening a business, however, is a big risk, since owners have a difficult time finding profitable locations. There are many things to consider about an area, such as competition, finding your customer base, and through traffic, all based on the type of business you want to open, according to AllBusiness.com.

I decided to focus on just coffee shops, to keep my problem simple, but the way my project is set up, it is a one line fix if you wanted to scope out other types of businesses, like restaurants.

According to kukani.org, 70% of canadians drink caffeine daily. This number pales in comparison to the same statistic in Americans, which is over 90%. Therefore, using my model on the United States later could lead to immense interest, but for now, I will work with the Toronto data.

1.2 Problem

Profit is calculated with Total Revenue - Total Costs, which foursquare was unable to supply. I couldn't find a data set that matched these venues, so my goal for this project is to generate a metric to predict profit based on the demographics of the area in combination with foursquare to show the best locations to open a new coffee shop. This will require research on which demographic variables are good predictors of profit.

1.3 Interest

Being able to predict how well a business will do in a certain location would be especially useful to potential small business owners, who are taking a much bigger risk of opening a business compared to chains. Larger companies, on the other hand, would be able to provide me massive data sets to refine my model if they show interest.

2 Data Acquisition

2.1 Data Source

My first task was finding out which demographic data would be good predictors for this project. According to Rick Suttle [3], the best way to predict profit is to know your customer base. Depending on your business, you need to target the right age and income group, in the right location. Dependable demographics variables include income, age, education, and proximity to customers.

I used the postal codes data of Toronto, from Wikipedia. This data consisted of postal codes, boroughs, and neighborhoods [5]. The postal codes were used with the foursquare API to get data for venues in the area. I will refer to this data as Venues.

```
print(venues_df.shape)
venues_df.head()
```

```
(2146, 9)
```

	PostalCode	Borough	Neighborhood	BoroughLatitude	BoroughLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	M3A	North York	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	M3A	North York	Parkwoods	43.753259	-79.329656	Brookbanks Pool	43.751389	-79.332184	Pool
2	M3A	North York	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	M4A	North York	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	M4A	North York	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

I also gathered demographics data for those same neighborhoods from Wikipedia, which contained an exhaustive list of variables, such as population, area and income [4]. This data, which I will call demographics, was used in combination to calculate my Profit metric.

Name	FM	Census Tracts	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	% Renters	Second most common language (after English) by name	Second most common language (after English) by percentage	Map
Toronto CMA Average		All	**5,113,149**	**5903.63**	**866**	**9.0**	**40,704**	**10.6**	**11.4**			
Agincourt	S	0377.01, 0377.02, 0377.03, 0377.04, 0378.02, 0378.08, 0378.14, 0378.23, 0378.24	44,577	12.45	3580	4.6	25,750	11.1	5.9	Cantonese (19.3%)	19.3% Cantonese	
Alexandra Park	OCoT	0039.00	4,355	0.32	13,609	0.0	19,687	13.8	28.0	Cantonese (17.9%)	17.9% Cantonese	
Allenby	OCoT	0140.00	2,513	0.58	4333	-1.0	245,592	5.2	3.4	Russian (1.4%)	01.4% Russian	
Amesbury	NY	0280.00, 0281.01, 0281.02	17,318	3.51	4,934	1.1	27,546	16.4	19.7	Spanish (6.1%)	06.1% Spanish	
Armour Heights	NY	0298.00	4,384	2.29	1914	2.0	116,651	10.8	16.1	Russian (9.4%)	09.4% Russian	

2.2 Data Cleaning

	Neighborhood	Borough	Population	Density	AvgIncome	Commuting%
0	Crescent Town	EY\n	8,157\n	20,393\n	23,021\n	24.5\n
1	Governor's Bridge	EY\n	2,112\n	1129\n	129,904\n	7.1\n
2	Leaside	EY\n	13,876\n	4938\n	82,670\n	9.7\n
3	O'Connor-Parkview	EY\n	17,740\n	3591\n	33,517\n	15.8\n
4	Old East York	EY\n	52,220\n	6577\n	33,172\n	22.0\n

The first thing I had to do was ensure the borough names were consistent, so I changed the boroughs in the Venues data to match the encoding from Demographics. This was done with df.replace(), for example, replace 'North York' with 'NY', and so on.

The next step was to remove the characters left over from beautiful soup, such as the new line characters.

I did a mini one hot encoding on whether or not the venue was a coffee shop. This allowed me to create a cluster of only coffee shops, to highlight them on the map. Doing this also allowed me to average the True/False to get a percentage of how common coffee shops were in a Neighborhood/Borough.

2.3 Feature Selection

The dataset I was working with, unfortunately, didn't have age as a variable, so I had to adjust slightly.

Based on my research on the correlation between demographics and profit, I decided to use population, average income, density, and commuting percentage, as these variables would impact the sales of coffee.

- Population $\implies$ A raw amount of potential customers.
- Avg Income $\implies$ Higher income in the area means more disposable income for premium coffee.
- Density $\implies$ a closer proximity to customers, e.g. a cafe on the bottom level of an apartment complex.
- Commuting Percent $\implies$ Amount of customers you could get from outside of the neighborhood on a daily basis.

Going forward, I plan on scraping profit data, so I can train-test-split, and/or use a decision tree to optimize the model.

Here is my merged data set:

```
[53]  merged = pd.merge(toronto_cluster,demographics_df)
      merged.head()
      #Need to get the merge to keep all data in toronto cluster, so left join
```

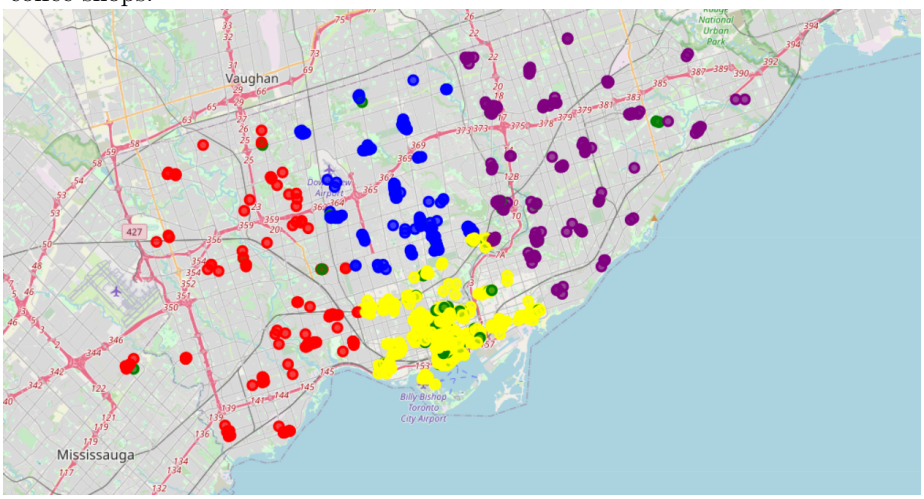
	VenueCategory	VenueLatitude	VenueLongitude	Labels	VenueName	marker_color	Coffee Distance	Neighborhood	Borough	Population	Density	AvgIncome	Commuting%
0	False	43.751976	-79.332140	4	Brookbanks Park	purple	2.283982	Parkwoods	NY	26533	5349	34811	14.0
1	False	43.751974	-79.333114	4	Variety Store	purple	2.256852	Parkwoods	NY	26533	5349	34811	14.0
2	False	43.723481	-79.315635	4	Victoria Village Arena	purple	0.036837	Victoria Village	NY	17047	3612	29657	15.6
3	True	43.725517	-79.313103	1	Tim Hortons	green	0.385463	Victoria Village	NY	17047	3612	29657	15.6
4	False	43.725819	-79.312785	4	Portugril	purple	0.338651	Victoria Village	NY	17047	3612	29657	15.6

I will explain what coffee distance is in the next section.

3 Exploratory Data Analysis

3.1 Target Variable (The profit score)

I used the foursquare data and k-clustering based on location and whether or not they were a coffee shop. This resulted in the map showing several districts, along with highlighting all coffee shops in Toronto. The green cluster represents the coffee shops.



The folium map was able to show which areas are dense with coffee shops. For example, there are dozens of coffee shops located in downtown Toronto, which means it would be more difficult to open a shop there due to competition.

This gave me the idea to look for areas of low competition, where coffee shops are not as popular. This could fill a void for many customers, who are traveling too far to get coffee. I used a for loop, and sklearn's nearest neighbors to iteratively calculate the distance to the nearest coffee shop for each venue. I will refer to this variable as 'coffee distance,' and it will be used in combination with the demographics.

```
[102]
for i in range(0,len(toronto_cluster)):
    X2 = dists2.to_numpy() #resets X and X2 every iteration so the point put in last iteration is gone
    X = dists.to_numpy()
    t_loc = X2[i,:]
    X = np.concatenate(([t_loc], X))

    nbrs = NearestNeighbors(n_neighbors = 2, algorithm = 'ball_tree').fit(X)
    distances, indices = nbrs.kneighbors(X)

    scaler = StandardScaler()
    scaler.fit(distances)

    #The first column is all zeroes, since it represents the distance between a point and itself
    #The second column represents the distance between a point and the nearest point

    Y = abs(scaler.transform(distances)[:,1]).tolist() #absolute distance
    toronto_cluster.at[i, 'Coffee Distance'] = Y[0] # = the distance from that point the the nearest coffee shop


toronto_cluster.head()
```

	VenueCategory	VenueLatitude	VenueLongitude	Labels	VenueName	marker_color	Coffee Distance
0	False	43.751976	-79.332140	3	Brookbanks Park	red	2.256931
1	False	43.751389	-79.332184	3	Brookbanks Pool	red	2.199774
2	False	43.751974	-79.333114	3	Variety Store	red	2.230391
3	False	43.723481	-79.315635	3	Victoria Village Arena	red	0.008802
4	False	43.725819	-79.312785	3	Portugril	red	0.302962

I grouped the data by neighborhoods, and calculated the profit score using the average of coffee distance, population, average income, commuting percent, and density, for each neighborhood. This takes every venue into account, shows the best locations by neighborhood.

4 Results

My working model is as follows:

$$Profit = \beta_1(\text{Population}) + \beta_2(\text{Income}) + \beta_3(\text{Density}) + \beta_4(\text{Commuting}) + \beta_5(\text{CoffeeDistance})$$

For now, my β_i's are all $\frac{1}{5}$, since I averaged my columns. ANOVA and model selection would be good methods to try after getting more data.

Here are the top 15 neighborhoods for a new coffee shop, based on demographics and competition.

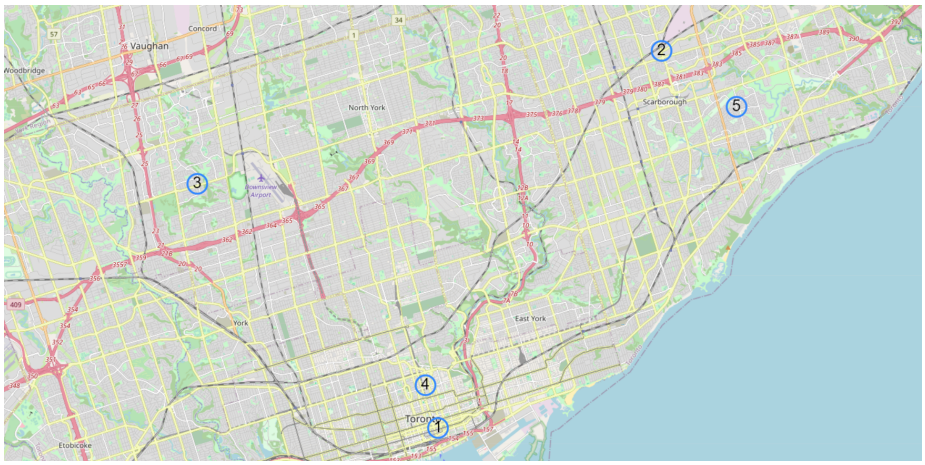
```
merge_group = merged.groupby('Neighborhood').mean() #The top 15 neighborhoods based on the average profit score metric
merge_group.sort_values('Profit Score', ascending = False).head(15)
```

Neighborhood	VenueCategory	VenueLatitude	VenueLongitude	Labels	Coffee Distance	Profit Score
St. James Town	0.058824	43.650098	-79.376167	1.941176	0.245395	0.449044
Agincourt	0.000000	43.792049	-79.259427	3.000000	3.842672	0.351064
Downsview	0.000000	43.742198	-79.501877	0.750000	2.855004	0.312063
Church and Wellesley	0.105263	43.666195	-79.382816	1.894737	0.204250	0.308067
Woburn	0.500000	43.771113	-79.220309	2.000000	0.299168	0.279654
Rosedale	0.000000	43.679754	-79.377335	2.000000	0.786679	0.277084
Parkwoods	0.000000	43.751780	-79.332480	3.000000	2.229032	0.259286
Bayview Village	0.000000	43.787903	-79.380860	4.000000	3.222711	0.232736
Humber Summit	0.000000	43.757837	-79.567048	0.000000	6.167868	0.231885
Weston	0.000000	43.705312	-79.515829	0.000000	1.921335	0.213516
Thorncliffe Park	0.050000	43.705567	-79.348375	2.150000	0.166435	0.172085
Victoria Village	0.250000	43.725467	-79.314735	2.500000	0.197702	0.155337
Don Mills	0.076923	43.730601	-79.342297	2.846154	0.522730	0.152831
Leaside	0.093750	43.708467	-79.361758	2.281250	0.246937	0.142927
Westmount	0.125000	43.694465	-79.532367	0.125000	0.177203	0.055747

1. St James Town is located in the heart of downtown, making it have high business opportunity, but lots of competition
2. Agincourt had the highest coffee distance of all neighborhoods, making it the most isolated
3. Downsview is situated right next to an airport, and also has high coffee distance
4. Church and Wellesley has close proximity to the University, making it a good business opportunity as well

5 Recommendations

Based on the results of my preliminary model, the best locations are St. James Town for the population and commuting, Church and Wellesley for the proximity to University of Toronto, or Agincourt, if you're more concerned about competition. Personally, I would choose Church and Wellesley, because it has a lack of competition on the West side of campus. Placing a cafe in a convenient location for students could lead to massive business, since college students are heavy coffee drinkers.



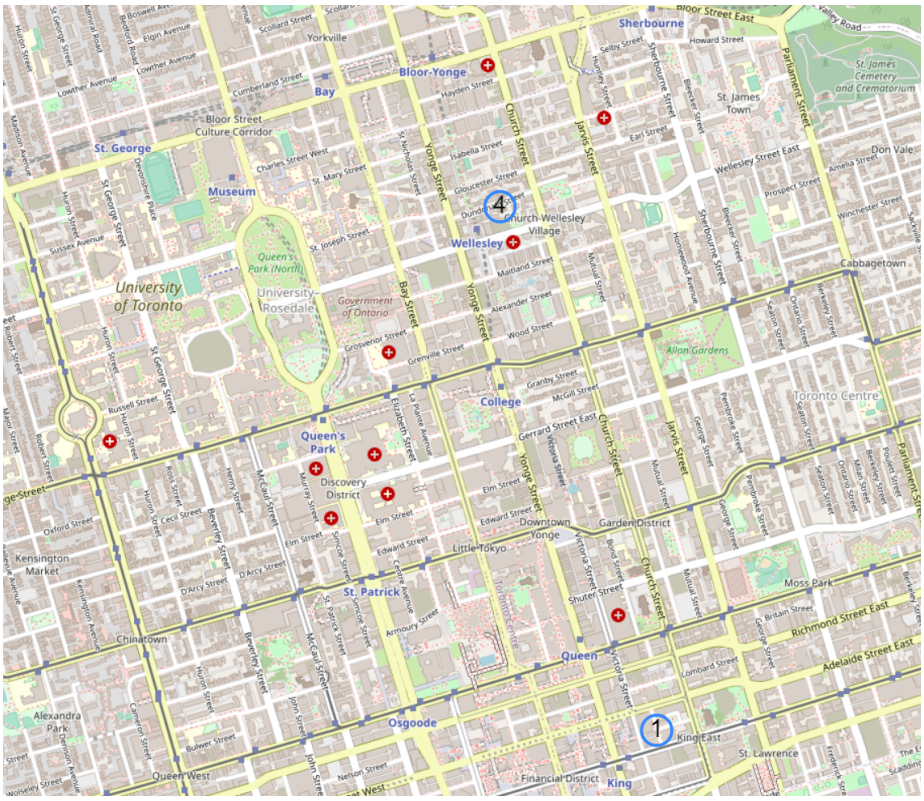
1. St. James Town
2. Agincourt
3. Downsview
4. Church and Wellesley
5. Woburn

6 Discussion

The areas I expected to show up in the results were there, such as near campus, and isolated areas on the outskirts of Toronto, but I didn't expect St. James Town to be number one, due to the amount of coffee shops in the area. My coffee distance variable seemed insightful, and showed areas that have no competition from other cafes.

7 Conclusion

In conclusion Church and Wellesley is the best location for a coffee shop, which is located just North of Downtown Toronto, with proximity to campus.



My model gave results that made sense, as the downtown area, where most coffee shops in Toronto are, had the most profitable location. Other locations that showed up in the results were areas with few coffee shops, where a potential business could corner the market.

I'm happy with my results, however, the model I came up with is one that will have to be proven once I get data for profit.

8 Citation

1. "Caffeine: America's Most Popular Drug." Kukani, www.kuakini.org/wps/portal/public/Health-Wellness/Health-Info-Tips/Miscellaneous/Caffeine--America-s-Most-Popular-Drug#:~:text=Every%20day%2C%20about%2090%20percent,it%20America's%20most%20popular%20drug.

2. City of Toronto. "Toronto at a Glance." City of Toronto, 29 July 2020, www.toronto.ca/city-government/data-research-maps/toronto-at-a-glance/.

3. Suttle, Rick. "The Demographic Variables That Affect a Business." Small Business - Chron.com, Chron.com, 20 Mar. 2019, smallbusiness.chron.com/demographic-variables-affect-business-24344.html.

4. "Demographics of Toronto Neighbourhoods." Wikipedia, Wikimedia Foundation, 25 July 2020, en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods.

5. "List of Postal Codes of Canada: M." Wikipedia, Wikimedia Foundation, 25 July 2020, en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.