IBM APPLIED DATA SCIENCE CAPSTONE

THE ADVENTURE OF SIMULATING WHERE TO OPEN A NEW COFFEE SHOP, CONSIDERING THE INFLUENCE OF REGIONS WITH TRANSPORT STATIONS

GUSTAVO SOUSA, BCS, CPRE-FL

CONTENTS

INTRODUCTION

The project seeks to return results of a research to a fictional coffee store company that have plans of expanding in New York and Toronto, considering opening near subway, train or bus stations.

BUSINESS PROBLEM

The moment of choosing and defining the place to begin a new business activity is critical. An infinity of variables – economic, cultural, social – must be considered and carefully assessed.

I decided to apply this study case considering two extremely popular kinds of venues: coffee shops and transportation stations. I exercised the imagination of a situation where, beyond the benefit of a huge flow of people nearby a station, our fictional entrepreneur is intending to use these venues on a two-month advertising campaign with flyers and promotional pamphlets.

The coffee shop franchising intends to open two new venues: one in New York and the other in Toronto. The best neighborhoods in each city to satisfy the business needs and objectives will be analyzed and assessed.

This research is also appliable in similar problems, considering crowded places and venues that can make use of this to gather clients.

DATA

The data sources used on this project are the datasets of neighborhoods from New Your and Toronto, and a sample list of venues distributed on the respective neighborhoods above. The venues' information is obtained via Foursquare, using the cities' datasets as input information.

NEW YORK AND TORONTO DATA

The data related to the New York neighborhoods was obtained on the New York University Spatial Data Repository webpage. A 2014 dataset is available for researchers and can be downloaded using the link https://geo.nyu.edu/catalog/nyu_2451_34572 .

The analyzed data was converted to a dataset like the following (just the top ten rows):

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446

The dataset has 306 neighborhoods distributed within 5 boroughs.

In turn, Toronto's neighborhoods data was obtained on the Wikipedia related webpage. The link is https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M .

In this case, it was necessary to extract the data using web scraping techniques and only after that convert it to the dataset. On this specific case, the neighborhoods are listed without coordinates' information, which lead us to a further necessary step to capture these coordinates.

For that, we can use Geopy, a Python client for several geolocation web service. The neighborhood postal code is used as input for obtaining the coordinate for that postal code.

The final analyzed and merged data was converted to a dataset like the following (just the top ten rows):

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern / Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood / Morningside / West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	Kennedy Park / Ionview / East Birchmount Park	43.727929	-79.262029
7	M1L	Scarborough	Golden Mile / Clairlea / Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffside / Cliffcrest / Scarborough Village W...	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff / Cliffside West	43.692657	-79.264848

The dataset has 103 neighborhoods distributed within 10 boroughs.

To complete this specific phase, it is necessary to capture the coordinates (latitude and longitude) of New York City and Toronto. For that, we again use Geopy, but now using the name of the city as input. The main coordinates of the city are returned.

These coordinates will be used to visualize the neighborhoods distributed around each city's main coordinates.

FOURSQUARE DATA

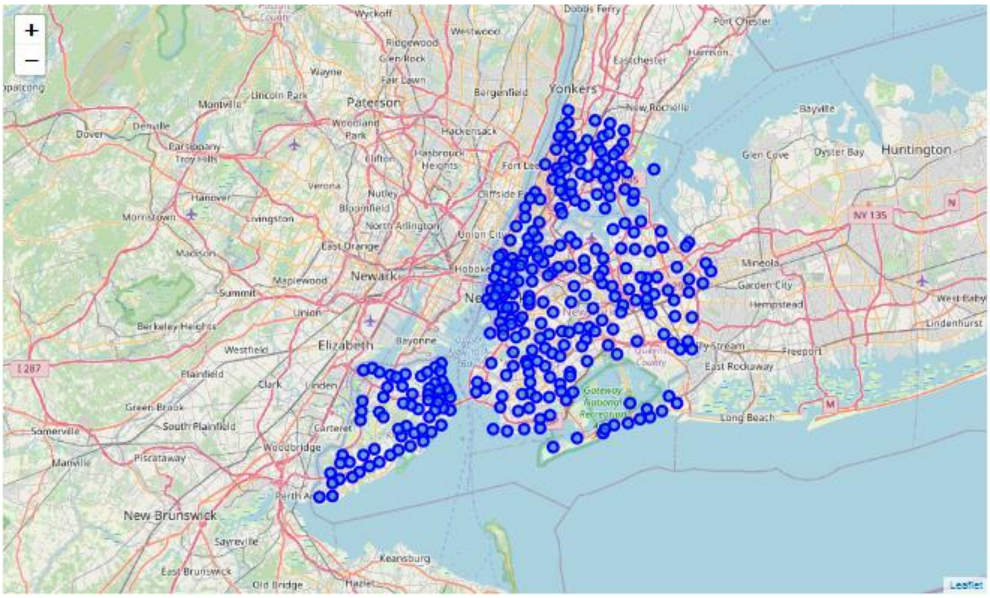
Foursquare service provides access to a huge amount of data related to points of interest, named by venues, in almost every city around the globe. A big part of the information provided (and captured) by the service is collaborative, what increases the database day by day.

On this case study we will capture the venues related to each neighborhood of our analyzed cities. After that it will be important to rank the most common venues nearby each neighborhood and proceed with the main purpose of our research.

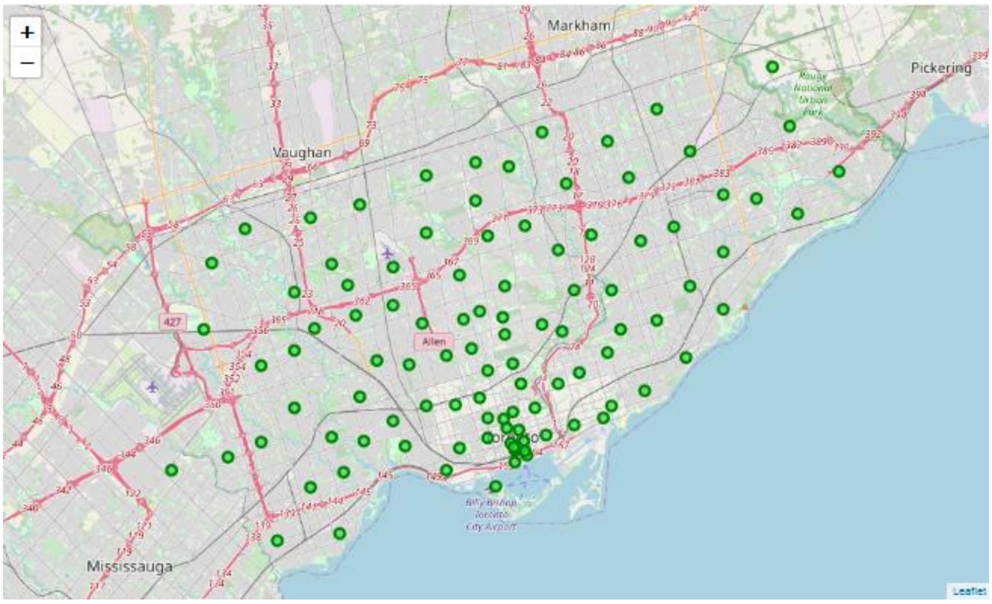
METHODOLOGY

The first step in our case study, after defining our initial data universe, and Foursquare as the main service to help us in the task of bringing more important information to our need, is to compose the datasets relating to the most common venues grouped by neighborhood. This will allow us to exercise our business rule of our problem situation.

We will use the Foursquare service, going through each neighborhood returning, within a radius of 500 meters, the main venues listed in the Foursquare tool, reaching a limit of 50 places. Before, follows below the distribution of the neighbors on the map in each city. This visualization helps in the understanding and interpretation of the data.



Representation of the New York neighborhoods



Representation of the Toronto neighborhoods

Through common credentials for access to Foursquare data, we compose our new base, with the places of interest grouped by neighborhood. We will use a Python function to retrieve this information.

As an initial result, below is the list of the first 10 neighborhoods (in alphabetical order) and the grouped, preliminary, number of places of interest associated with it.

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Allerton	32	32	32	32	32	32
Annadale	14	14	14	14	14	14
Arden Heights	4	4	4	4	4	4
Arlington	7	7	7	7	7	7
Arrochar	20	20	20	20	20	20
Arverne	17	17	17	17	17	17
Astoria	50	50	50	50	50	50
Astoria Heights	14	14	14	14	14	14
Auburndale	18	18	18	18	18	18
Bath Beach	49	49	49	49	49	49

New York sample

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Agincourt	5	5	5	5	5	5
Alderwood / Long Branch	10	10	10	10	10	10
Bathurst Manor / Wilson Heights / Downsview North	19	19	19	19	19	19
Bayview Village	4	4	4	4	4	4
Bedford Park / Lawrence Manor East	24	24	24	24	24	24
Berczy Park	50	50	50	50	50	50
Birch Cliff / Cliffside West	4	4	4	4	4	4
Brockton / Parkdale Village / Exhibition Place	23	23	23	23	23	23
Business reply mail Processing CentrE	18	18	18	18	18	18
CN Tower / King and Spadina / Railway Lands / Harbourfront West / Bathurst Quay / South Niagara / Island airport	17	17	17	17	17	17

Toronto sample

Next, we will move on to the most important step so that we can have a real and comparative view of the most representative venues within each neighborhood. We will do the process of standardizing or normalizing the data.

Normalization is a technique commonly applied in this phase of data preparation, whether in purely data analysis problems such as machine learning. The main idea is to change the values of numerical columns, which obey different scales in their minimum and maximum value ranges, to a common scale, which will not cause distortions to the value ranges. Also known as 0-1 transformation, it is thus known to assume minimum values of 0 and maximum values of 1 for each data column, indicating the percentage of incidence of each element.

Below is an example of a portion of the normalized database for the New York neighborhoods.

17	Bedford Stuyvesant	0.000000	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.0	0.0
18	Beechhurst	0.083333	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.0	0.0
19	Bellaire	0.000000	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.0	0.0
20	Belle Harbor	0.000000	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.0	0.0
21	Bellerose	0.000000	0.00	0.0	0.000000	0.000000	0.000000	0.043478	0.000000	0.000000	...	0.00	0.0	0.0
22	Belmont	0.000000	0.00	0.0	0.000000	0.000000	0.000000	0.020000	0.000000	0.000000	...	0.00	0.0	0.0
23	Bensonhurst	0.000000	0.00	0.0	0.000000	0.000000	0.000000	0.035714	0.000000	0.000000	...	0.00	0.0	0.0
24	Bergen Beach	0.000000	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.0	0.0
25	Bliss ville	0.000000	0.00	0.0	0.000000	0.000000	0.058824	0.000000	0.000000	0.000000	...	0.00	0.0	0.0
26	Bloomfield	0.000000	0.00	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00	0.0	0.0

The next step, once the two bases have been normalized, is to classify or rank the most frequent places of interest in each neighborhood. Thus, with the ten most frequent locations ranked in a descending order, we will be able to follow the final step of our analysis.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Allerton	Pizza Place	Deli / Bodega	Cosmetics Shop	Department Store	Supermarket
1	Annadale	Pub	Restaurant	Diner	Pizza Place	Train Station
2	Arden Heights	Pizza Place	Pharmacy	Pool	Playground	Coffee Shop
3	Arlington	Bus Stop	Intersection	American Restaurant	Deli / Bodega	Grocery Store
4	Arrochar	Bus Stop	Italian Restaurant	Deli / Bodega	Cosmetics Shop	Athletics & Sports
5	Arverne	Surf Spot	Sandwich Place	Metro Station	Bus Stop	Playground
6	Astoria	Middle Eastern Restaurant	Bar	Pizza Place	Mediterranean Restaurant	Greek Restaurant
7	Astoria Heights	Chinese Restaurant	Plaza	Laundromat	Bakery	Hostel
8	Auburndale	Furniture / Home Store	Supermarket	Pet Store	Pharmacy	Toy / Game Store
9	Bath Beach	Chinese Restaurant	Pharmacy	Gas Station	Sushi Restaurant	Italian Restaurant

First 10 neighborhoods of New York and their five most common venues

	Neighborhood	1st Most Common	2nd Most Common	3rd Most Common	4th Most Common	5th Most Common
0	Agincourt	Latin American Restaurant	Skating Rink	Lounge	Breakfast Spot	Women's Store
1	Alderwood / Long Branch	Pizza Place	Pharmacy	Skating Rink	Pool	Sandwich Place
2	Bathurst Manor / Wilson Heights / Downsview No...	Coffee Shop	Bank	Gift Shop	Pizza Place	Sandwich Place
3	Bayview Village	Japanese Restaurant	Chinese Restaurant	Bank	Café	Electronics Store
4	Bedford Park / Lawrence Manor East	Sandwich Place	Italian Restaurant	Restaurant	Coffee Shop	Women's Store
5	Berczy Park	Coffee Shop	Cocktail Bar	Bakery	Cheese Shop	Café
6	Birch Cliff / Cliffside West	College Stadium	Skating Rink	General Entertainment	Café	Empanada Restaurant
7	Brockton / Parkdale Village / Exhibition Place	Café	Breakfast Spot	Coffee Shop	Nightclub	Burrito Place
8	Business reply mail Processing CentrE	Gym / Fitness Center	Spa	Auto Workshop	Brewery	Burrito Place
9	CN Tower / King and Spadina / Railway Lands / ...	Airport Service	Airport Lounge	Airport Terminal	Sculpture Garden	Harbor / Marina

First 10 neighborhoods of Toronto and their five most common venues

We have come to the point where we have defined our data set on which we will apply our business rule. The first step now is to search only the neighborhoods that have, among their ten most frequent places of interest, concentrations of people in transportation stations. We defined these locations as bus, train, subway and busy bus stops. That is, this is our first filtering. All neighborhoods that do not meet this criterion are outside of our assessment. We use the search for keywords such as 'bus stop', 'bus station', 'metro station', 'subway' and 'train station' among the categories present.

After this filtering, we proceeded to the filtering that indicates a pre-disposition of the neighborhood for cafeteria-type venues. For this check we use the terms 'donut', 'bakery', 'cafe' and 'coffee'. The idea is to look for neighborhoods with a certain culture for companies of this type.

RESULTS

The results of our analysis are the following:

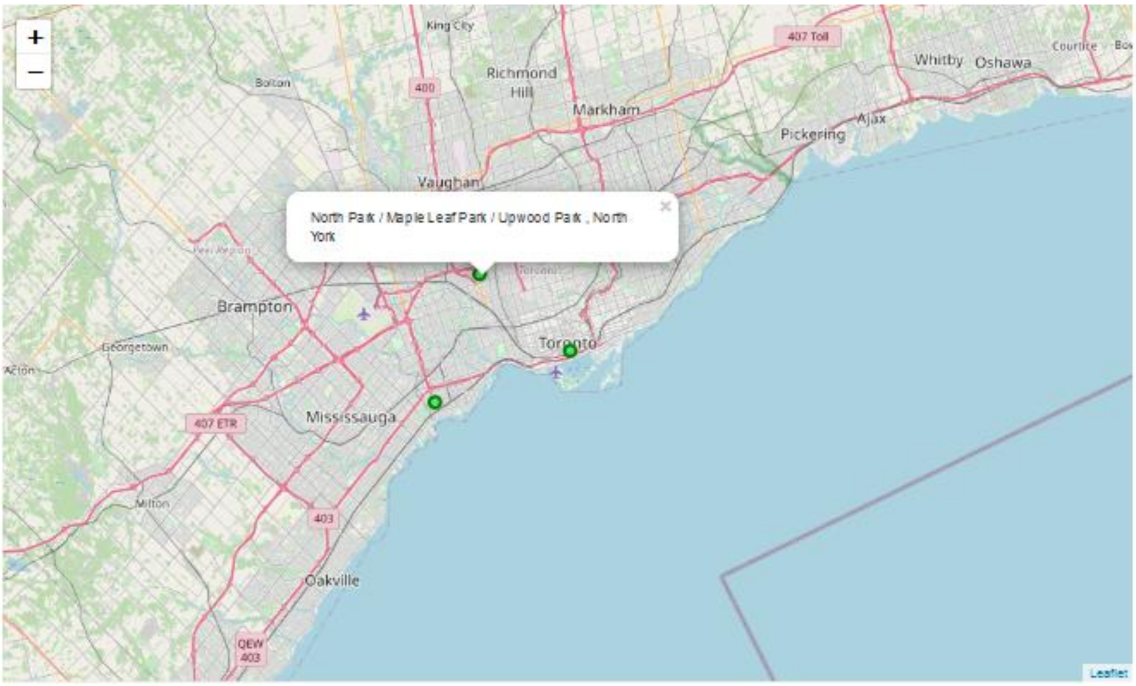
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Annadale	Pizza Place	Bakery	Park	Diner	Sushi Restaurant	Bagel Shop	Train Station	Restaurant	American Restaurant	Sports Bar
1	Claremont Village	Grocery Store	Chinese Restaurant	Pizza Place	Bus Station	Food	Liquor Store	Caribbean Restaurant	Gym	Bakery	Discount Store
2	Eastchester	Caribbean Restaurant	Deli / Bodega	Diner	Bowling Alley	Donut Shop	Metro Station	Bakery	Seafood Restaurant	Fast Food Restaurant	Pizza Place
3	Grasmere	Bus Stop	Bakery	Deli / Bodega	Restaurant	Park	Grocery Store	Pharmacy	Bank	Bagel Shop	Home Service
4	Gravesend	Italian Restaurant	Lounge	Pizza Place	Bus Station	Bakery	Furniture / Home Store	Donut Shop	Sporting Goods Shop	Record Shop	Men's Store
5	Manhattan Beach	Bus Stop	Café	Sandwich Place	Beach	Food	Ice Cream Shop	Playground	Women's Store	Exhibit	Eye Doctor
6	Mott Haven	Gym	Spanish Restaurant	Donut Shop	Pizza Place	Latin American Restaurant	Metro Station	Burger Joint	Bakery	Peruvian Restaurant	Electronics Store
7	South Jamaica	Bus Station	Vegetarian / Vegan Restaurant	Bakery	Supermarket	Caribbean Restaurant	Grocery Store	Sandwich Place	Field	Exhibit	Eye Doctor
8	Tompkinsville	Thrift / Vintage Store	Brewery	Rock Club	Bus Stop	Spanish Restaurant	Supermarket	Café	Caribbean Restaurant	Mexican Restaurant	Gastropub

In New York, eight neighborhoods met our analysis. We used a Kmeans clusterization to find some similarity between them, although the small sample was not very productive.

For the city of Toronto, the filtering was even more restrictive, presented the result below:

	Neighborhood	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	Alderwood / Long Branch	Pizza Place	Bakery	Park	Diner	Sushi Restaurant	Bagel Shop	Train Station	Restaurant	American Restaurant	Sports Bar
1	North Park / Maple Leaf Park / Upwood Park	Grocery Store	Chinese Restaurant	Pizza Place	Bus Station	Food	Liquor Store	Caribbean Restaurant	Gym	Bakery	Discount Store
2	Toronto Dominion Centre / Design Exchange	Caribbean Restaurant	Deli / Bodega	Diner	Bowling Alley	Donut Shop	Metro Station	Bakery	Seafood Restaurant	Fast Food Restaurant	Pizza Place

It would not make much sense to do a cluster view of only three neighborhoods. We just opted to put them on the map of Toronto. The visualization of the location will be one of the final inputs for the business decision.



DISCUSSIONS

We could observe that two apparently independent databases, when assertively related, can bring numerous different analyses about the existing information.

Evaluations only related to frequency representations can leave the influence of the bias to the situation as late as possible, where it will have practically no impact on the quality of the data.

It is also worth mentioning the importance of the collaborative data from the Foursquare database, which increasingly enables analyses of high cultural, social and economic impact.

CONCLUSION

For the purpose of reflection and observation, it is worth noting that this whole case study based on a fictitious business idea, culminates in a subjective opportunity assessment and obeys the criteria of market observation, which is complex. In our business situation, for example, it was taken as a principle that it would not be interesting to locate in a place where no similar category venue had been representative among the ten most frequent categories. This could indicate a neighborhood incompatible with the branch, or even a place where the risk of not achieving good results was high.

This is an inference, and shows that often the universe of data, even after all the cleaning, preparation and analysis, will still be subjected to a situation of subjectivity.