Capstone Project Report
The Battle of Neighborhoods

Author: Nuno Silva

December, 2020

Abstract

In this work we use data science techniques and venues data fetched from Foursquare to study a specific business problem. We compare neighborhoods from Toronto and Manhattan using clustering techniques, namely the k-means algorithm, to identify similar neighborhoods in distinct cities using the venues categories data. Our results show that Regent Hill in Toronto may be similar to Marble Hill or Financial District in Manhattan, a conclusion that may be helpful for the final decision of the business problem.

1. Introduction

In this work we use data science techniques and data fetched from Foursquare to study a specific business problem. The project was performed in the context of the Capstone Project - The Battle of Neighborhoods of IBM Data Science series.

1.1. Business problem and background

GONE is a successful grab-and-go restaurant that is established in X, Toronto. Recently, an opportunity to establish a franchise in Manhattan, New York appeared through their network of friends and USA collaborators. Knowing that neighborhood dynamics are one of the most important success factors of its venue in Canada, they are looking to understand the dynamics of the neighborhoods of New York to help making the decision.

2. Data Description

For our research we will use the neighborhoods location data provided in the course and fetch additional data regarding venues from Foursquare API. For each neighborhood we collected the data of 100 venues within a radius of 500 meters from the given location. In principle, this data and in particular each venue category will be able to characterize each neighborhood in terms of lifestyle and dynamics.

2.1. Data for Toronto

Regarding the data for Toronto we scrap the wikipedia for data of postal codes in the zone of Toronto and restrict to the Boroughs containing "Toronto", resulting in a dataset containing 39 neighborhoods. We then use the location data provided in the course materials to associate latitude and longitude coordinates. Through this location data we use Foursquare venues search API to collect the data for 100 venues within a radius of 500 meters of the given location of a neighborhood, resulting in a total of 1619 venues falling in 235 unique categories. Our final dataset is a frequency table of each unique venue category for each one of the 39 neighborhoods.

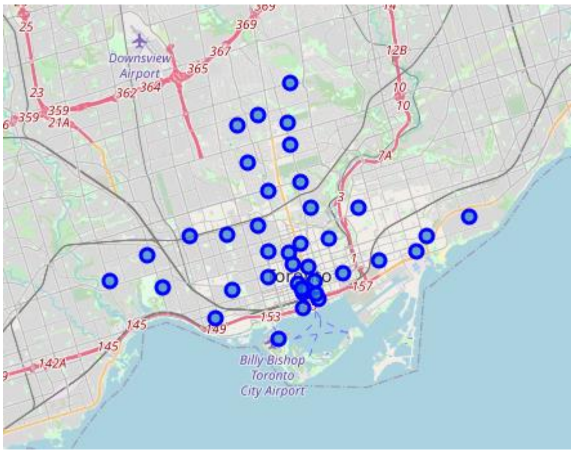
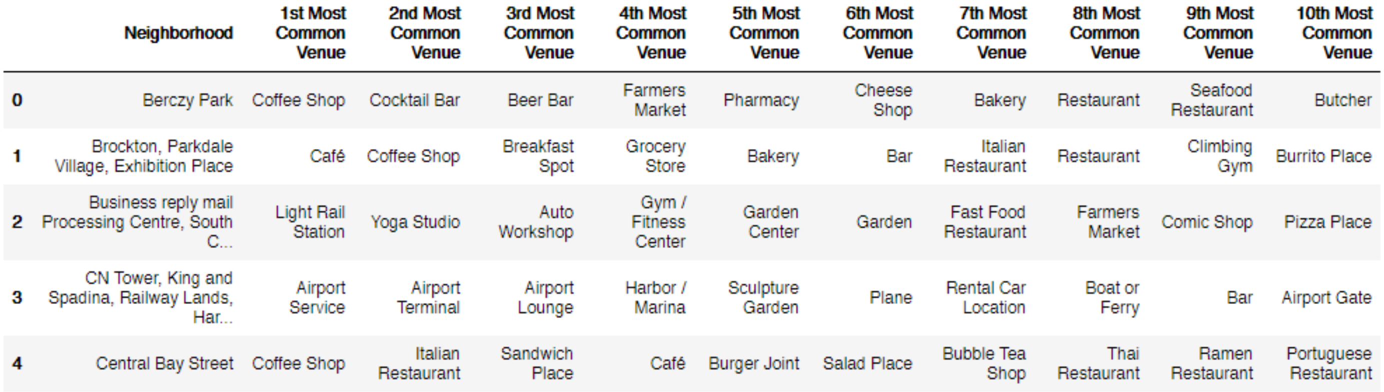


Figure 1 Map of the neighborhoods of Toronto.

Table 1 First five lines of a table containing the neighborhoods and the top10 most common venues in each of them.



	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	Berczy Park	Coffee Shop	Cocktail Bar	Beer Bar	Farmers Market	Pharmacy	Cheese Shop	Bakery	Restaurant	Seafood Restaurant	Butcher
1	Brockton, Parkdale Village, Exhibition Place	Café	Coffee Shop	Breakfast Spot	Grocery Store	Bakery	Bar	Italian Restaurant	Restaurant	Climbing Gym	Burrito Place
2	Business reply mail Processing Centre, South C...	Light Rail Station	Yoga Studio	Auto Workshop	Gym / Fitness Center	Garden Center	Garden	Fast Food Restaurant	Farmers Market	Comic Shop	Pizza Place
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Terminal	Airport Lounge	Harbor / Marina	Sculpture Garden	Plane	Rental Car Location	Boat or Ferry	Bar	Airport Gate
4	Central Bay Street	Coffee Shop	Italian Restaurant	Sandwich Place	Café	Burger Joint	Salad Place	Bubble Tea Shop	Thai Restaurant	Ramen Restaurant	Portuguese Restaurant

2.2. Data for Manhattan

Regarding the data for Manhattan we used the New York data from the course materials and restrict to the Boroughs containing "Manhattan", resulting in a dataset containing 40 neighborhoods. We then use the location data fetched from geopy to associate latitude and longitude coordinates. Through this location data we use Foursquare venues search API to collect the data for 100 venues within a radius of 500 meters of the given location of a neighborhood, resulting in a total of 3217 venues falling in 331 unique categories. Our final dataset is a frequency table of each unique venue category for each one of the 40 neighborhoods.

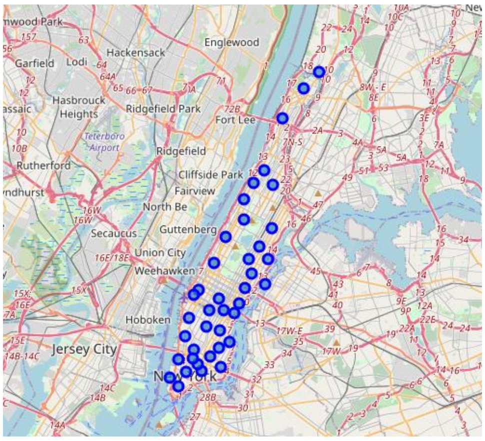
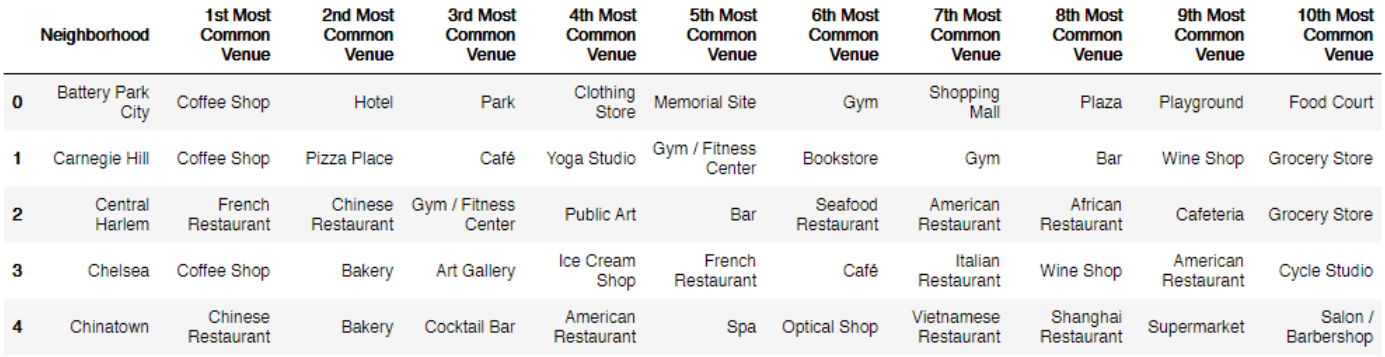


Figure 2 Map of the neighborhoods in Manhattan.

Table 2 First five lines of a table containing the neighborhoods and the top10 most common venues in each of them.



	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	Battery Park City	Coffee Shop	Hotel	Park	Clothing Store	Memorial Site	Gym	Shopping Mall	Plaza	Playground	Food Court
1	Carnegie Hill	Coffee Shop	Pizza Place	Café	Yoga Studio	Gym / Fitness Center	Bookstore	Gym	Bar	Wine Shop	Grocery Store
2	Central Harlem	French Restaurant	Chinese Restaurant	Gym / Fitness Center	Public Art	Bar	Seafood Restaurant	American Restaurant	African Restaurant	Cafeteria	Grocery Store
3	Chelsea	Coffee Shop	Bakery	Art Gallery	Ice Cream Shop	French Restaurant	Café	Italian Restaurant	Wine Shop	American Restaurant	Cycle Studio
4	Chinatown	Chinese Restaurant	Bakery	Cocktail Bar	American Restaurant	Spa	Optical Shop	Vietnamese Restaurant	Shanghai Restaurant	Supermarket	Salon / Barbershop

3. Methodology

We start by first exploring the cities individually through clustering techniques, taking as parameters the frequencies of each venue category for each neighborhood. For that we will use k-means algorithms of the *sklearn* library[1] and chose the optimal k value using the elbow method running the *KElbowVisualizer* routine from *yellowbrick* library[2]. We analyze the resulting data by printing each cluster data and visualize them in a map using folium[3] for a visual aid.

Finally, to understand if we can find a similar neighborhood as X in Manhattan we perform the same methodology applied to all the data, trying to understand if Manhattan and Downtown Toronto are similar.

Clusters in Toronto

For clustering the neighborhoods of Toronto in terms of its venues – which may give an interesting picture regarding lifestyles – we first need to determine the optical number of clusters to use in the k-means technique. For that we will ran the *KElbowVisualizer* routine on our data obtaining the results depicted on Figure 3. Clearly the elbow is not as pronounced as it should be yet the routine indicates that k=4 is the best for our case.

The results show that Regent Park belongs to a large cluster containing 35 neighborhoods of Toronto which suggest that in Toronto one would find many similar neighborhoods. Does the same happens in Manhattan?

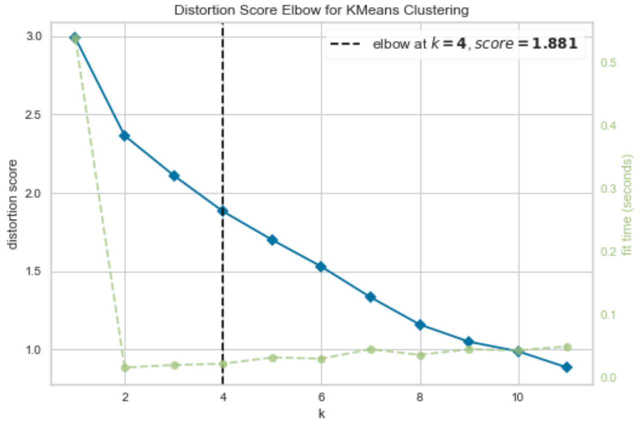


Figure 3 Results of the elbow methodology to determine the best

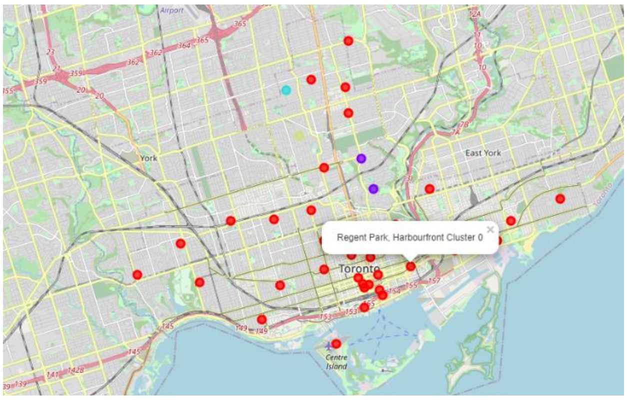


Figure 4 Results of the clustering process for Toronto neighborhoods.

Clusters in Manhattan

We then did the same procedure for the neighborhoods of Manhattan. The elbow technique resulted in a much larger optimal k=15 which suggests that Manhattan neighborhoods are more diverse comparing to Toronto. Yet the question is that if we can find a similar one to Regent Park in Manhattan and for that we must apply the procedure to the data of both cities simultaneously.

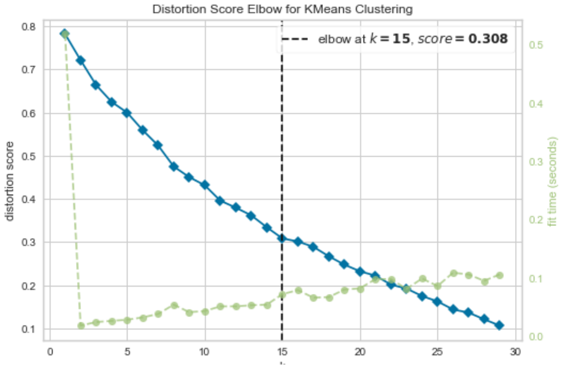


Figure 5 Results of the elbow methodology to determine the best k.

Clustering the neighborhoods of both cities

We proceeded and merged the data of both cities, applying the described procedure. The elbow technique suggested an optimal k=14 clusters, again suggesting that the neighborhoods are indeed more diverse comparing to Toronto and reflecting the influence of Manhattan's diversity. In particular we are interested in the Regent Park cluster and we found that the resulting cluster contains a total of 14 neighborhoods, with 12 belonging to the city of Toronto and only 2 located in Manhattan – Marble Hill and the Financial District.

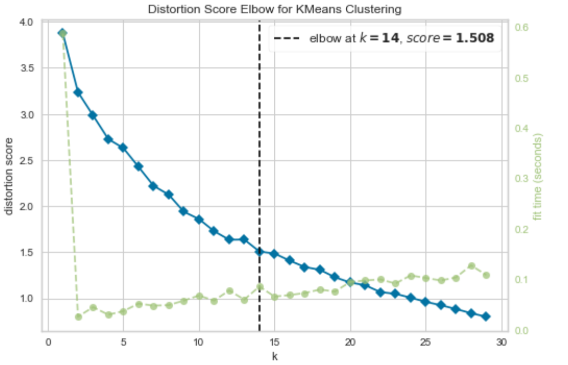
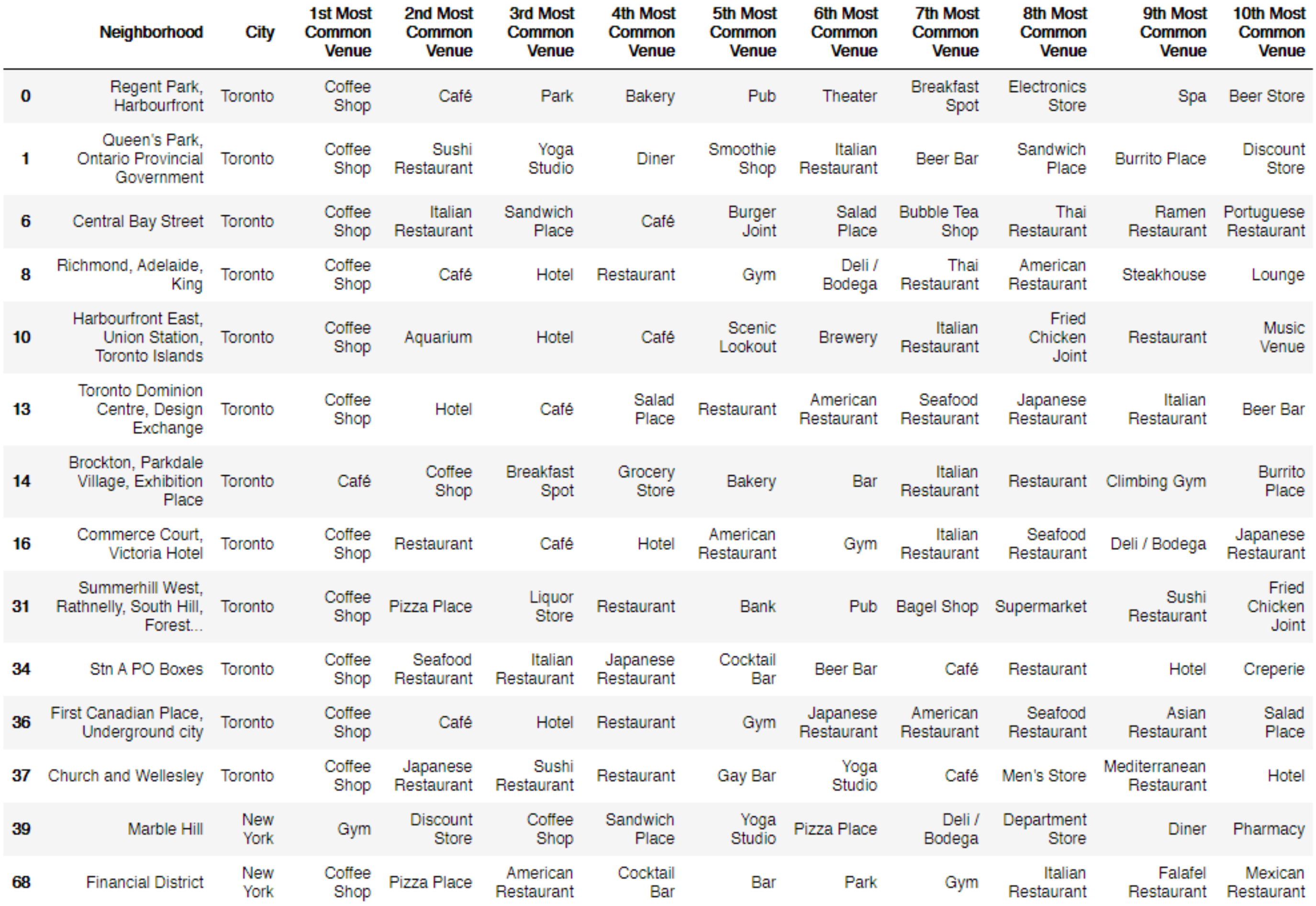


Figure 6 Results of the elbow methodology to determine the best k.

Table 3 Cluster containing Regent Park Neighborhood.



	Neighborhood	City	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	Regent Park, Harbourfront	Toronto	Coffee Shop	Café	Park	Bakery	Pub	Theater	Breakfast Spot	Electronics Store	Spa	Beer Store
1	Queen's Park, Ontario Provincial Government	Toronto	Coffee Shop	Sushi Restaurant	Yoga Studio	Diner	Smoothie Shop	Italian Restaurant	Beer Bar	Sandwich Place	Burrito Place	Discount Store
6	Central Bay Street	Toronto	Coffee Shop	Italian Restaurant	Sandwich Place	Café	Burger Joint	Salad Place	Bubble Tea Shop	Thai Restaurant	Ramen Restaurant	Portuguese Restaurant
8	Richmond, Adelaide, King	Toronto	Coffee Shop	Café	Hotel	Restaurant	Gym	Deli / Bodega	Thai Restaurant	American Restaurant	Steakhouse	Lounge
10	Harbourfront East, Union Station, Toronto Islands	Toronto	Coffee Shop	Aquarium	Hotel	Café	Scenic Lookout	Brewery	Italian Restaurant	Fried Chicken Joint	Restaurant	Music Venue
13	Toronto Dominion Centre, Design Exchange	Toronto	Coffee Shop	Hotel	Café	Salad Place	Restaurant	American Restaurant	Seafood Restaurant	Japanese Restaurant	Italian Restaurant	Beer Bar
14	Brockton, Parkdale Village, Exhibition Place	Toronto	Café	Coffee Shop	Breakfast Spot	Grocery Store	Bakery	Bar	Italian Restaurant	Restaurant	Climbing Gym	Burrito Place
16	Commerce Court, Victoria Hotel	Toronto	Coffee Shop	Restaurant	Café	Hotel	American Restaurant	Gym	Italian Restaurant	Seafood Restaurant	Deli / Bodega	Japanese Restaurant
31	Summerhill West, Rathnelly, South Hill, Forest...	Toronto	Coffee Shop	Pizza Place	Liquor Store	Restaurant	Bank	Pub	Bagel Shop	Supermarket	Sushi Restaurant	Fried Chicken Joint
34	Stn A PO Boxes	Toronto	Coffee Shop	Seafood Restaurant	Italian Restaurant	Japanese Restaurant	Cocktail Bar	Beer Bar	Café	Restaurant	Hotel	Creperie
36	First Canadian Place, Underground city	Toronto	Coffee Shop	Café	Hotel	Restaurant	Gym	Japanese Restaurant	American Restaurant	Seafood Restaurant	Asian Restaurant	Salad Place
37	Church and Wellesley	Toronto	Coffee Shop	Japanese Restaurant	Sushi Restaurant	Restaurant	Gay Bar	Yoga Studio	Café	Men's Store	Mediterranean Restaurant	Hotel
39	Marble Hill	New York	Gym	Discount Store	Coffee Shop	Sandwich Place	Yoga Studio	Pizza Place	Deli / Bodega	Department Store	Diner	Pharmacy
68	Financial District	New York	Coffee Shop	Pizza Place	American Restaurant	Cocktail Bar	Bar	Park	Gym	Italian Restaurant	Falafel Restaurant	Mexican Restaurant

Discussion

Our research used clustering techniques and data referent to the venues of each neighborhood to compare lifestyles and find similarities between distinct locations. Our results suggest that Marble Hill and the Financial District are the neighborhoods in Manhattan that are most similar to Regent Park and therefore, can be more suitable for a possible franchise in the target borough. To better explore and confirm these results additional search involving venues rate would be beneficial for more solid conclusions.

Conclusions

In this work we used data science and multiple source data to study a specific business problem. Our client is a grab-and-go restaurant owner from Regent Park, Toronto, that is looking to establish a franchise venue in Manhattan, New York. As a background, he believes from its experience that part of the success of its venue is own the lifestyle and neighborhood dynamics. To help in the business decision we fetched venue data for each neighborhood in Toronto and Manhattan from Foursquare to analyze each neighborhood in terms of lifestyle and possibly find one in Manhattan that is similar to Regent Park.

We compared the neighborhoods using clustering techniques, namely the k-means algorithm, to identify similar neighborhoods in distinct cities using the venues categories data. Our results show that Manhattan is far more diverse than Toronto, which means that while in Toronto there are many similar neighborhoods to Regent Park, the same may not verify in Manhattan. Indeed, applying the procedure to the whole dataset, we found that Regent Hill in Toronto seems to be similar only with Marble Hill or Financial District in Manhattan, a final conclusion that may be helpful for the decision of the business problem.

References

[1] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[2] Yellowbrick, https://www.scikit-yb.org/en/latest/

[3] https://python-visualization.github.io/folium/