Site Selection to open Sushi place

Dawit H. Hailu

dawit.hailu@mail.huji.ac.il

IBM Data Science Professional Certificate
December, 2018

Outline

Sushi place

The problem

- Your friend wants to open a Sushi place in Toronto
- You are asked to help in site selection for her business

You agreed to help your friend in selecting the best sites for her business by making use of your newly acquired data science tools

Data Collection

- Data about the neighborhoods in the city of Toronto is not readily available
- Therefore we have to scrape Wikipedia page[1] to obtain dataset about Borough in Toronto
- Then wrangle, clean it, and read it into a *pandas* dataframe before making use of it
- We make use of *BeautifulSoup* package to scrape the page
- We extract more information about the venues in the neighborhood from Foursquare API [2]
- location data to explore a geographical location

[1] https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
[2] https://foursquare.com/

Snippet of the dataset

	PostalCode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M5A	Downtown Toronto	Regent Park
6	M6A	North York	Lawrence Heights
7	M6A	North York	Lawrence Manor
8	M7A	Queen's Park	Not assigned
9	M8A	Not assigned	Not assigned
10	M9A	Etobicoke	Islington Avenue
11	M1B	Scarborough	Rouge

```
Toronto_df.shape
```

```
(289, 3)
```

Figure: Top 12 rows of the data before it is cleaned

Neighborhoods' coordinate is not included yet

	PostalCode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront,Regent Park
3	M6A	North York	Lawrence Heights,Lawrence Manor
4	M7A	Queen's Park	Queen's Park
5	M9A	Etobicoke	Islington Avenue
6	M1B	Scarborough	Rouge,Malvern
7	M3B	North York	Don Mills North
8	M4B	East York	Woodbine Gardens,Parkview Hill
9	M5B	Downtown Toronto	Ryerson,Garden District
10	M6B	North York	Glencairn
11	M9B	Etobicoke	Cloverdale,Islington,Martin Grove,Princess Gar...

Figure: Top 12 rows of the data after it is cleaned

Snippet of the dataset

	PostalCode	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	Harbourfront,Regent Park	43.654260	-79.360636
3	M6A	North York	Lawrence Heights,Lawrence Manor	43.718518	-79.464763
4	M7A	Queen's Park	Queen's Park	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
6	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
7	M3B	North York	Don Mills North	43.745906	-79.352188
8	M4B	East York	Woodbine Gardens,Parkview Hill	43.706397	-79.309937
9	M5B	Downtown Toronto	Ryerson,Garden District	43.657162	-79.378937
10	M6B	North York	Glencairn	43.709577	-79.445073
11	M9B	Etobicoke	Cloverdale,Islington,Martin Grove,Princess Gar...	43.650943	-79.554724

```
Toronto_data.shape
```

```
(103, 5)
```

- omitted Borough with "not assigned"
- combined Neighborhood who share same PostalCode area
- included coordinates of the neighborhood
- original size of our dataframe has 289 rows and 3 columns
- now we have 103 rows and 5 columns

Foursquare

- I will use the *Foursquare* API to explore neighborhoods in Toronto.
- I will use the **explore** function to get the most common venue categories in each neighborhood,
- and then use this feature to group the neighborhoods into clusters.
- I then will use the k-means clustering algorithm to complete this task.
- Finally, I will exploit the Folium library to visualize the neighborhoods in Toronto and their emerging clusters

Neighborhood dataset

- Neighborhood has a total of 11 boroughs and 103 neighborhoods.
- In order to segment the neighborhoods and explore them, we will essentially need a dataset
- the dataset should contain: the 11 boroughs and the neighborhoods that exist in each Borough
- we as well need the the latitude and longitude coordinates of each neighborhood.

Venues of each neighborhood

What we did

we first categorize the neighborhood in Toronto and then using *Foursquare* API we got top 5 popular venues in each neighborhood where we included coordinates of the neighborhood

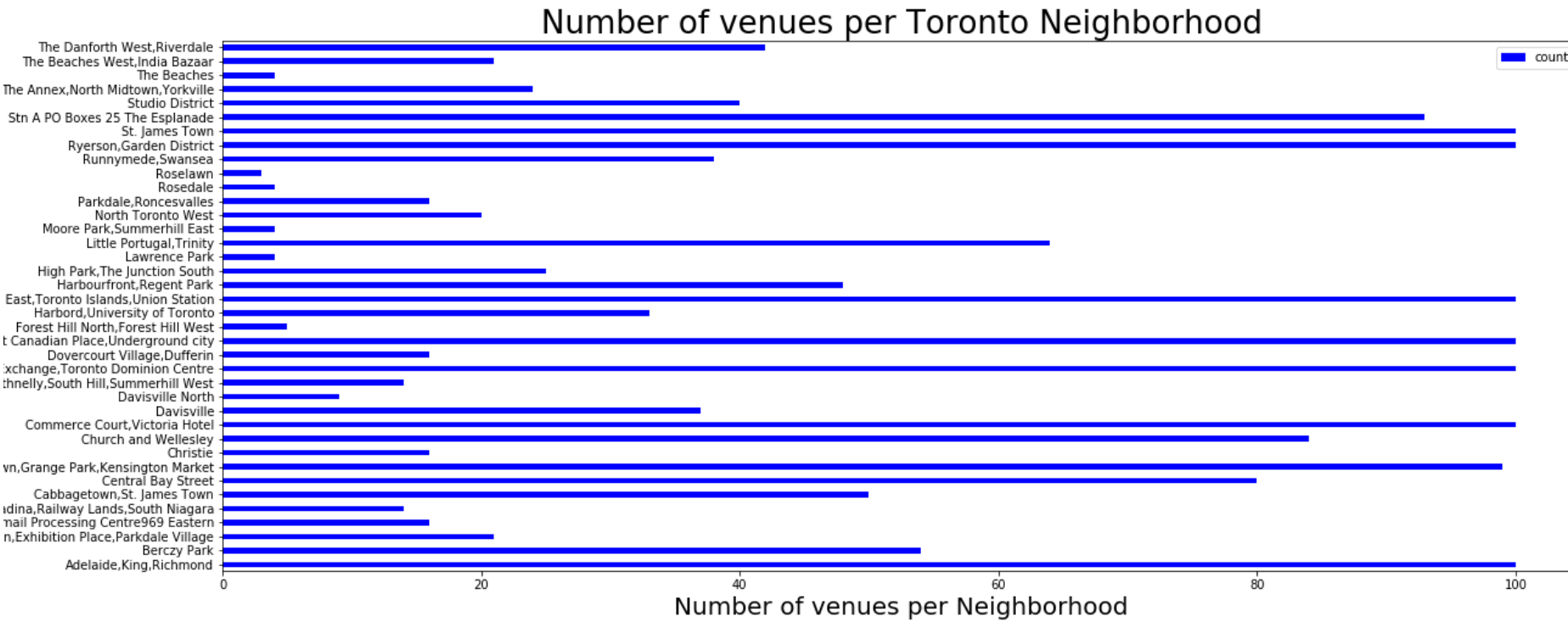
for Sushi place

We want to learn from experience as we hope to identify the popular venues wherever there is Sushi place

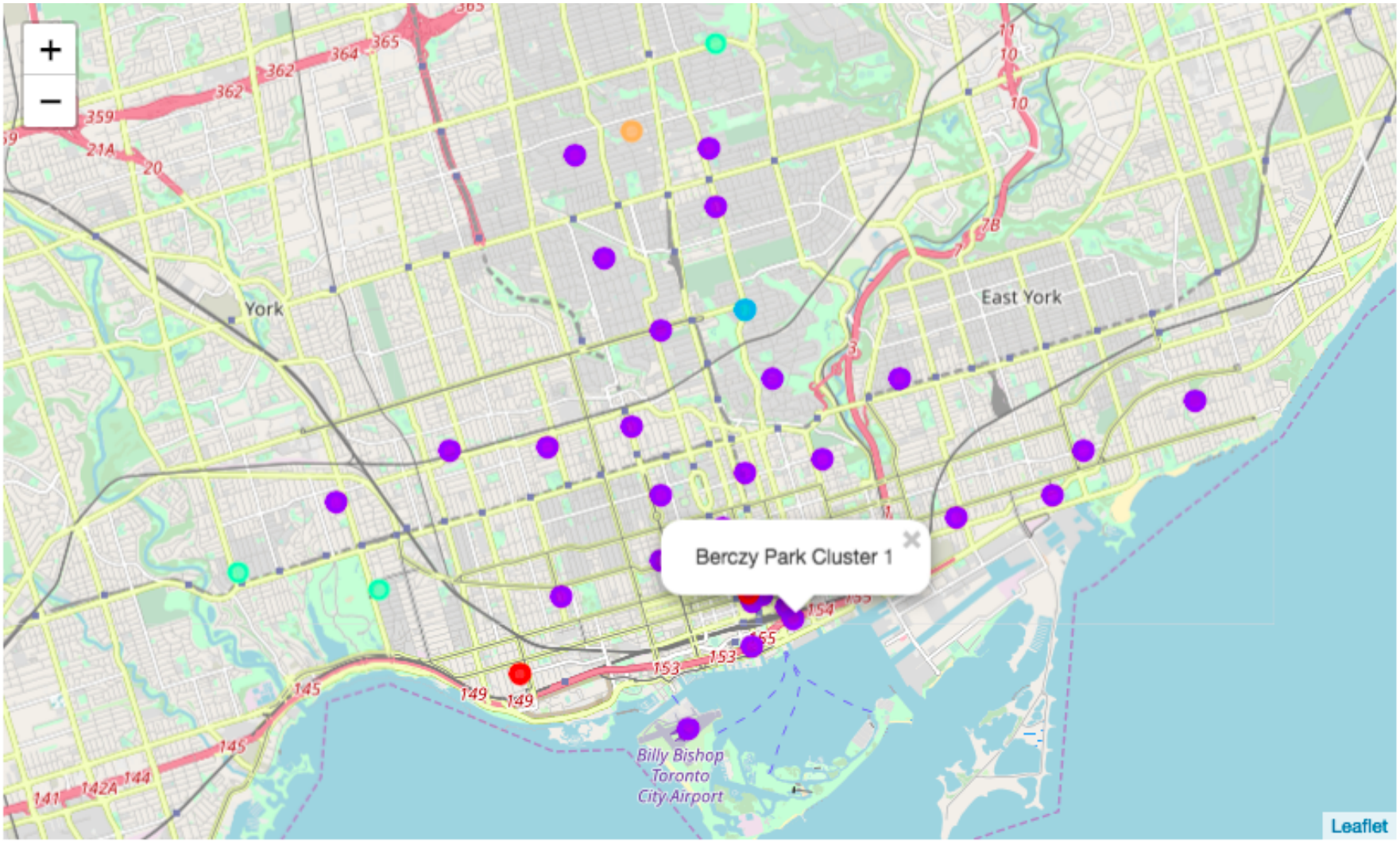
experience?

We search from the dataset for Japanese restaurant, Asian restaurant, sushi restaurant or Sushi place

Venues of each neighborhood



Neighborhood



Explore neighborhood

Average

We calculated the average number of Sushi near those neighborhoods to be 3.5.

Conditions

- the top 5 categories of venues in a neighborhood
- we demand less than 4 Sushi places in a neighborhood

Congested neighborhood

we further refine our choice by addressing the question that which of the sites will be able to cover the most number of neighborhoods

List of candidate neighborhood

Coffee Shop	98
Café	60
Restaurant	36
Hotel	33
Bar	26
Japanese Restaurant	24
Italian Restaurant	23
Pizza Place	23
Gym	18
American Restaurant	18

dtype: int64

Figure: Top venues near existing Sushi related places

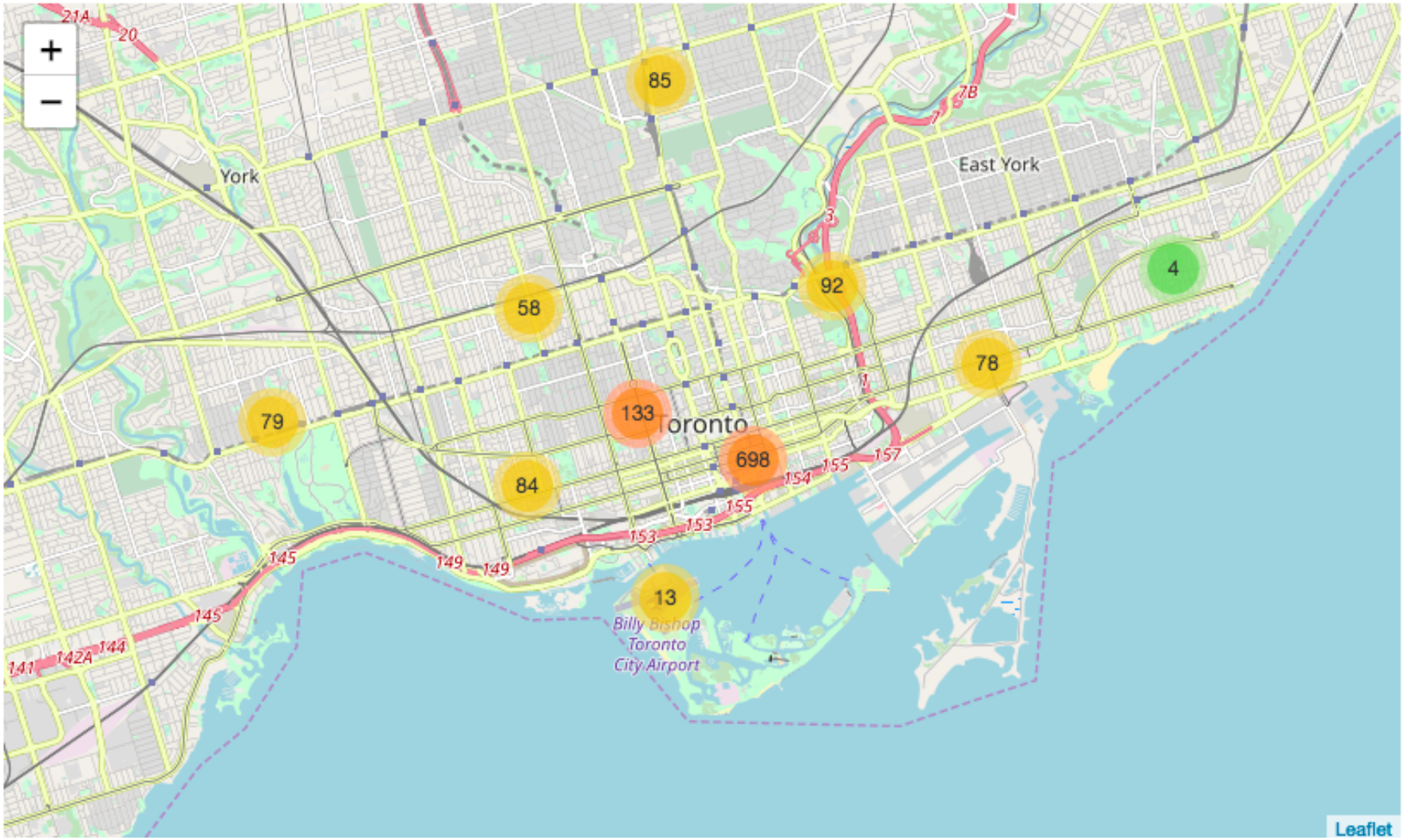
Top location candidates are

Neighborhood	
Berczy Park	0
Dovercourt Village,Dufferin	0
High Park,The Junction South	0
Studio District	0
The Danforth West,Riverdale	0
Cabbagetown,St. James Town	1
Chinatown,Grange Park,Kensington Market	1
Commerce Court,Victoria Hotel	1
Harbourfront,Regent Park	1
Harbourfront East,Toronto Islands,Union Station	2
Little Portugal,Trinity	2
St. James Town	2
Stn A PO Boxes 25 The Esplanade	2
Design Exchange,Toronto Dominion Centre	3
Harbord,University of Toronto	3

dtype: int64

Figure: Neighborhood along with number of existing Sushi places

Clustered neighborhood



Best candidate

Berczy Park

- In this project we managed to find the best neighborhood for a Sushi place, i.e. **Berczy Park**
- In order to identify the optimal location factors we took into consideration are: learning from experience by studying the popular venues around existing Sushi places and assessing the existence, or lack thereof, Japanese restaurant/ Sushi restaurant/ Asian restaurant.

Limitation of the study

Limitation?

it is worth mentioning that we haven't included the population density of the neighborhoods, as we did not have ready made dataset for population in each neighborhood. But as remedy we have further tried to cluster the neighborhoods' proximity to each other. The more neighborhood dense clusters our site belongs to the better location it is, as it will be easily accessible from those neighborhoods.

Thank you for your attention