Is life in Toronto linear or nonlinear?

Introduction

When people are hungry, they will go to a restaurant, when people are thirsty, they will find a coffee shop, and when people need new clothes, they will go to store. These venues are reflections of people's activities. The locations of venues are not random. For example, a food joint usually picks a location where most people visit, because the more people pass by, the more likely they pick up some food in this food joint.

And an interesting question is, from a geography point of view, are the venues located linearly or nonlinearly? Using locations of venues as a proxy, we can take a look at the life pattern of people.

In this project, we will explore the locations of venues of Toronto. And answer the question, do venues in Toronto located linearly or nonlinearly, which type of venue is located linearly or nonlinearly.

Data

The data used in this project are acquired through Wikipedia and Foursquare. First the postcodes, Borough and Neighborhood information is downloaded from Wikipedia page. After removing not assigned borough and combine neighborhood in the same borough, we get 103 neighborhoods.

	Postcode	Borough	Neighbourhood
0	M1B	Scarborough	Rouge,Malvern
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union
2	M1E	Scarborough	Guildwood,Morningside,West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

Figure 1 DataFrame of postcode, Borough and Neighborhood

Then for the locations of the neighborhood, we use latitude and longitude to represent the physical location of the neighborhood. The latitude and longitude could be achieved through geocoder, however the codes to get coordinates from geocoder always fail because max retries exceed with url. In this project we use the coordinate data directly from the webpage provided in week3 of Applied Data Science Capstone.

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	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

Figure 2 DataFrame of postcode, Borough and Neighborhood, with coordinates
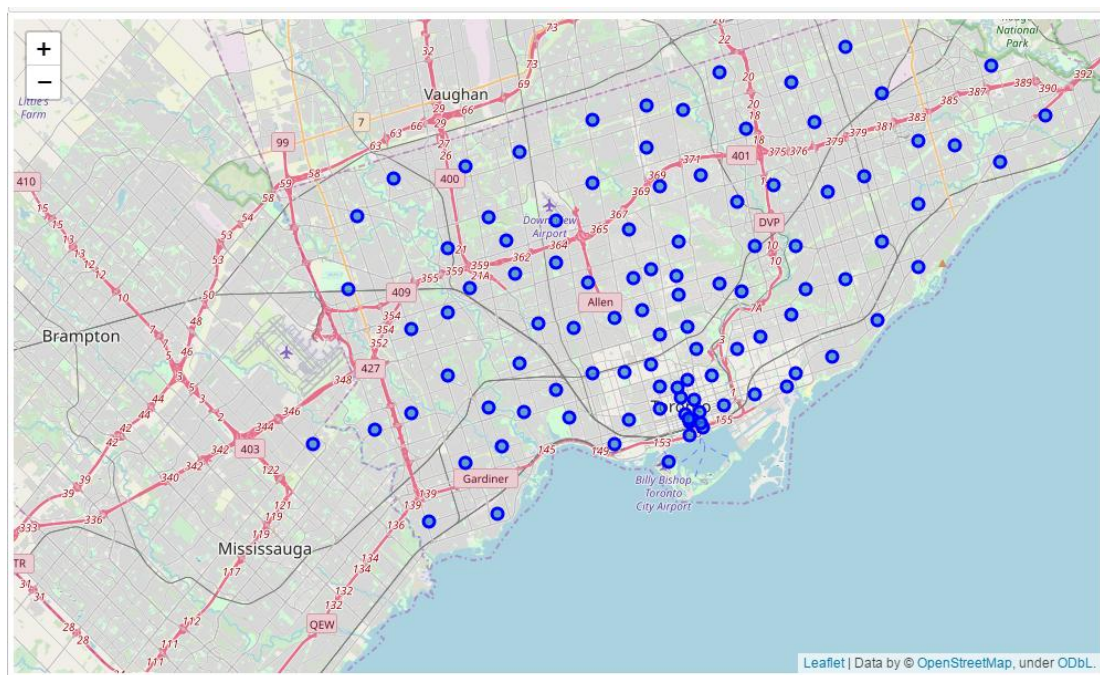

Figure 3 map of Toronto where neightborhoods are marked

Once we have the coordinates of the neighborhoods, venues in each neighborhood can be accessed using Foursquare API. The data we get contains the type of the venue, and the venue's coordinates. When accessing the venue data, we limit the maximum number of venues to be 100. For all the neighborhoods we get 2266 venue information. And this 2266 venues belong to 276 unique venue types.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge,Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge,Malvern	43.806686	-79.194353	Interprovincial Group	43.805630	-79.200378	Print Shop
2	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	RIGHT WAY TO GOLF	43.785177	-79.161108	Golf Course
3	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
4	Guildwood,Morningside,West Hill	43.763573	-79.188711	Swiss Chalet Rotisserie & Grill	43.767697	-79.189914	Pizza Place

Figure 4 sample of DataFrame containing venue information

The venues are grouped by their types and their neighborhoods. The dataframe looks like this:

	Postcode	Borough	Neighborhood	Latitude	Longitude	Café	Coffee Shop	Park	Pizza Place	Restaurant	Sandwich Place
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353	0.0	0.0	0.0	0.0	0.0	0.0
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	0.0	0.0	0.0	0.0	0.0	0.0
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711	0.0	0.0	0.0	1.0	0.0	0.0
3	M1G	Scarborough	Woburn	43.770992	-79.216917	0.0	2.0	0.0	0.0	0.0	0.0
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	0.0	0.0	0.0	0.0	0.0	0.0

Figure 5 sample of DataFrame where venue types are processed by one-hot method

Because regression analysis is used in this project, we need to filter out some venue types that do not have enough sample. In this project we choose the venue types that are not zero In at least 30 neighborhoods. Also before we feed the data into the model, for each venue type selected, the neighborhoods that do not have venues of this type are removed.

Reference

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
'https://cocl.us/Geospatial_data
Forsquare API