TABLE OF CONTENTS

Contents

Introduction

IDEA

When people want to start a new business in a city like a restaurant, it always a good idea to learn the details about distribution of the city which they want to start their business. The detail distribution which means according to different district or neighborhood such as what kind of restaurant located more or less, what is one particular neighborhood generally used for, what kind of people living here, and what kind of service they need most. By knowing these detail information, it would be easier for merchant to decide the location for their restaurant business.

FEATURES IMPORTANT FOR RESTAURANT LOCATION PICKUP

- Restaurant distribution: there's lots of different restaurant categories such as Chinese restaurant, fast food, Italian restaurant, and Mexican food. Avoiding duplicate food categories when there was already existing plentiful in a same neighborhood would reduce a fierce competition.
- Safety: choose a safety neighborhood is also very important for restaurant business. Since people willing to pick a safe place to hang out.
- Parking: Since some city like Los Angeles has limited parking lot which lead to very expensive parking fees. It will lead to an increased cost of both customers and business owners.
- Surrounding environment: surrounding environment can be considered directly related to customer amount. For example, a shopping mall around may lead to more customers.

These features have a very close connection to restaurant profits. There are still a lot more features need to be considered when start a new restaurant business. However, due to the limitation of data set, we only analyze several features that could be showed by our clustering method.

Data

The dataset for this capstone project consists of neighborhood information of Toronto obtained from Wikipedia website: *List of postal codes of Canada: M*. By using the BeautifulSoup package of Python, transforming the website data to pandas data frame which include postcode, borough, and neighborhood features. Then according to the transformed data frame obtained latitude and longitude by applying geopy package. Then using the latitude and longitude we can collect venues near each neighborhood for cluster analysis by using Foursquare API.

Figure 1 shows the first five rows of website data from Wikipedia website:

	Postcode	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

Figure 1

This data frame only contains index numbers and three features which are postcode, borough, and neighborhood. I used folium package to map the data point on map, which shows in figure 2:
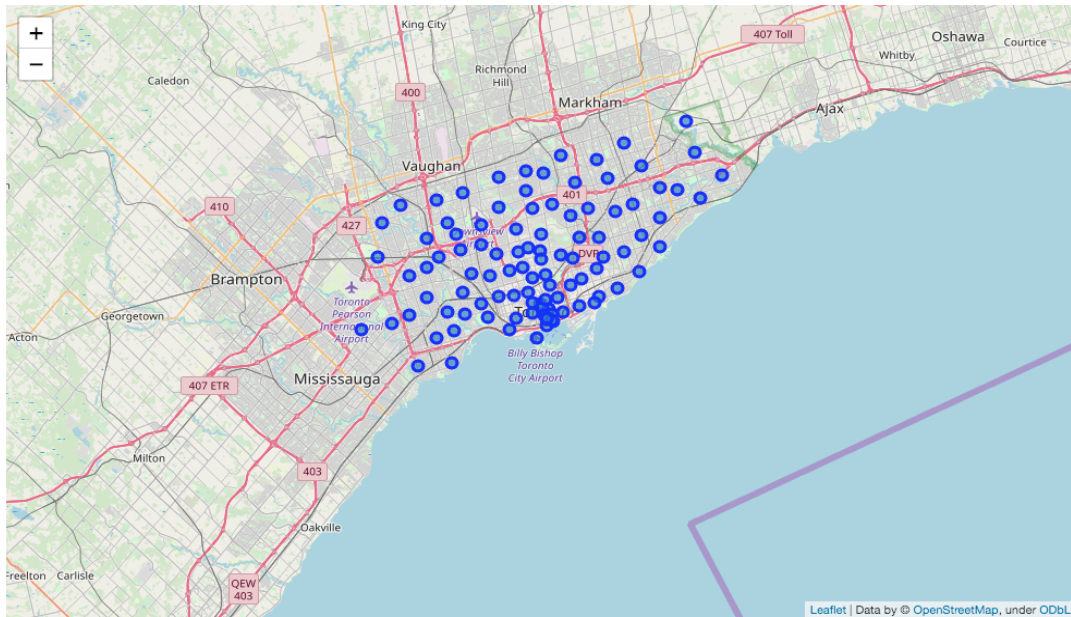
Figure 2

We can see there's boroughs not in Toronto, and the table does not contain latitude and longitude. The next step is to obtain latitude and longitude, then filtering the unnecessary values.

Figure 3 shows the table with latitude and longitude added and only included the borough which contains Toronto:

	Postcode	Borough	Neighbourhood	Latitude	Longitude
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
19	M4E	East Toronto	The Beaches	43.676357	-79.293031
20	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306

Figure 3

Then with obtained table, we can get the venues from Foursquare, and doing k-means clustering to obtain the most popular venue at each borough.