Applied Data Science Capstone

Capstone Project - The Battle of Neighborhoods

Created by Matej Kroc

Contents

1. Introduction

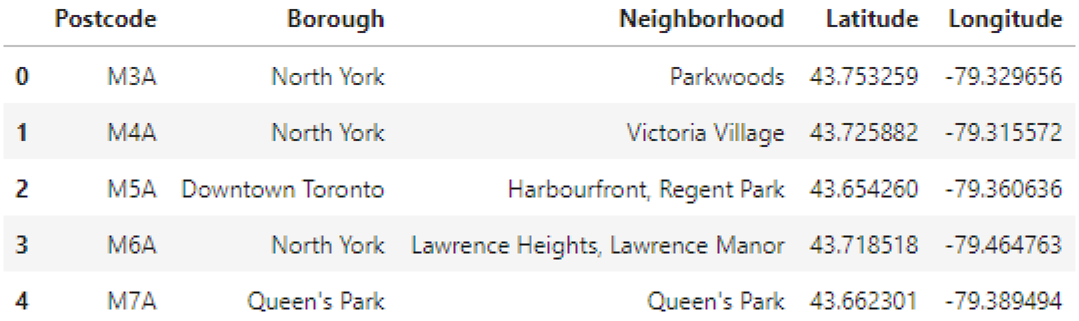
In this project, which is part of the Data Science Capstone course on Coursera, we attempt to give protentional stakeholders a better description of the Toronto, Canada neighborhoods. This is done for the purpose of recognizing which area/neighborhood is the best for opening a new restaurant. Finding appropriate place for a new restaurant is crucial in nowadays business. Finding gap on the market in certain area or building your restaurant in an area that is similar to other location where certain type of a restaurant is popular can be the breaking point for successful business.

We collected the data about neighborhood restaurants and analyzed them. Location information was obtained via the Foursquare API. By using data mining technique k-mean clustering we categorize our dataset into different clusters based on similarities in restaurants available in the areas. It the end, we analyzed different types of restaurants and what cluster would be the best option for the investor to build the new restaurant in.

2. Data description

Data for this project were obtained from different sources, cleaned and add to a single data frame. Python and Jupyter Notebook were used during the analysis.

We needed to obtain names and locations of neighborhoods in Toronto, Canada. Names of different boroughs and neighborhoods together with their postcodes were extracted from Wikipedia page containing information about Canadian postcodes. Only postal codes where the first letter is M were extracted since postal codes beginning with M are located within the city of Toronto. We omitted postal codes that have no borough and neighborhood assigned. Some neighborhoods are relatively small; therefore, some postal codes are the same for more than one neighborhood. We merged the data for these areas to obtain data with a unique postal code as an identifier for each neighborhood. Next, using the geocoder library available for Python we obtained location coordinates for each neighborhood based on the postal code. The final dataset contains 103 different postal codes. The first 5 entries in the dataset can be seen in Table 1. The map of Toronto with the marked neighborhood's location can be seen in Figure 1. An interactive map is also available at GitHub repository (GitHub, nbviewer(renders maps)) .



	Postcode	Borough	Neighborhood	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

Table 1: Dataset with neighborhood locations.

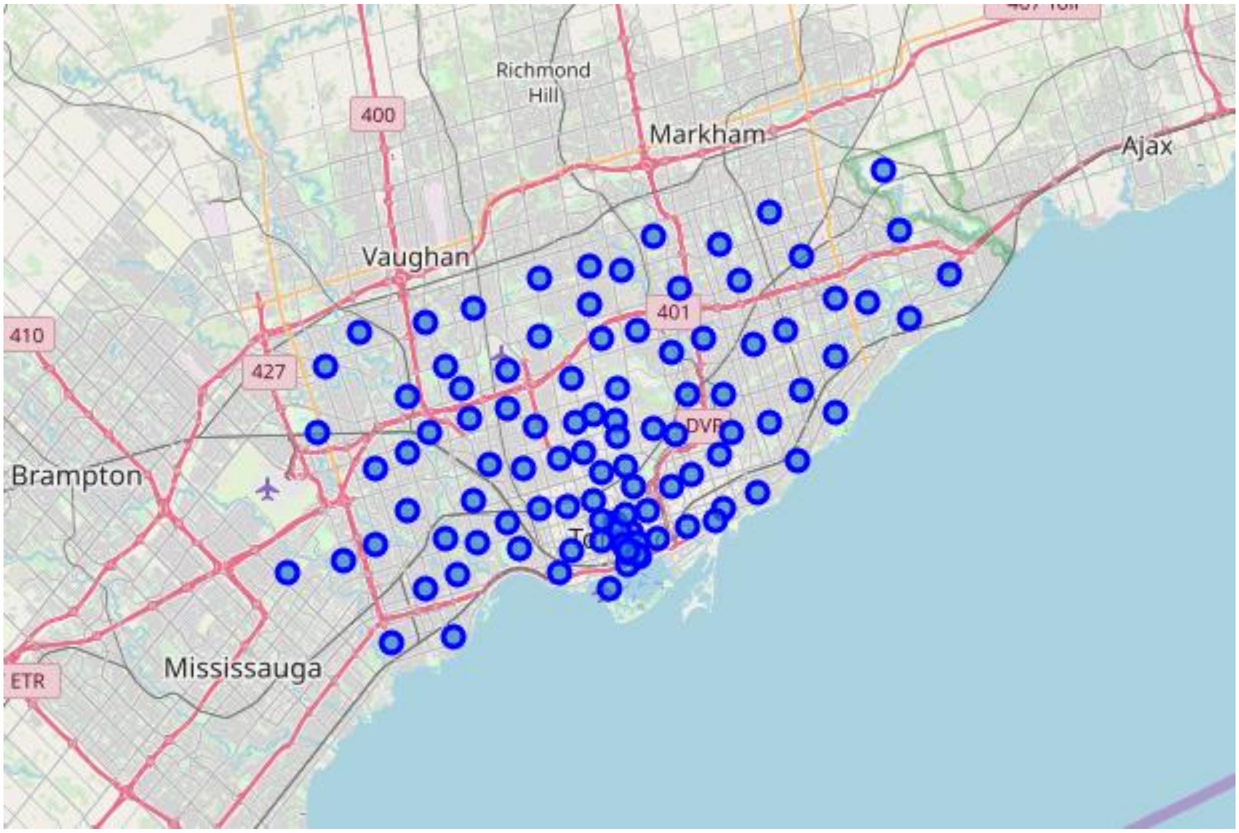
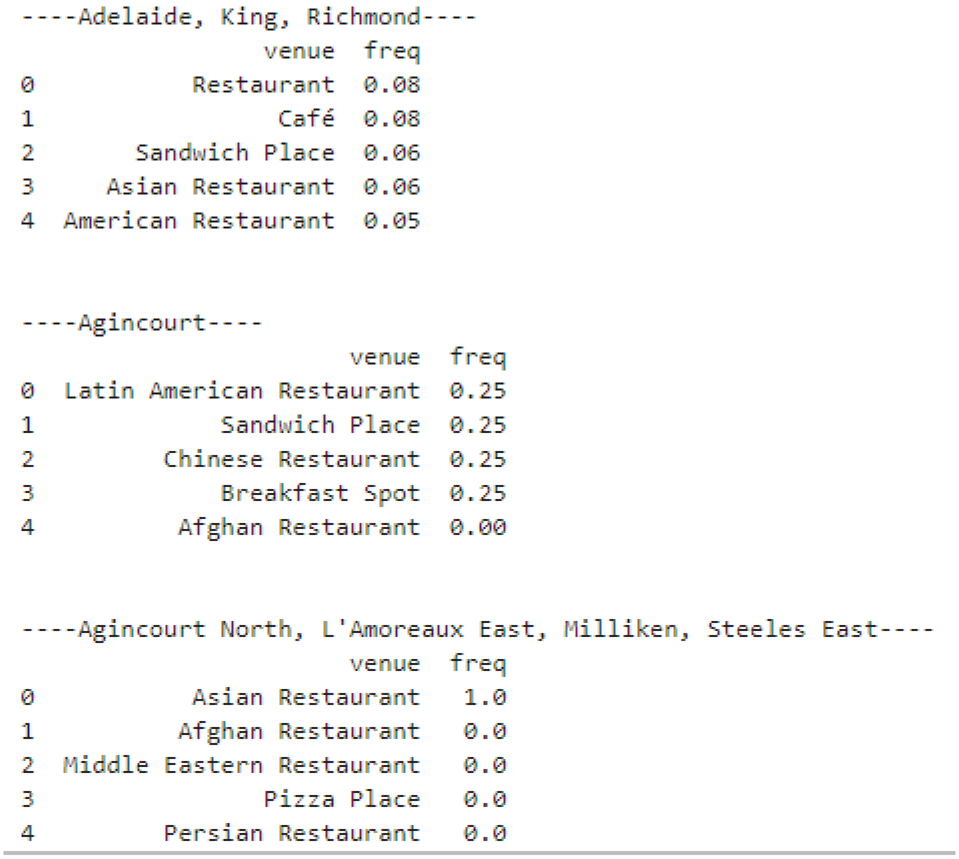


Figure 1: The map of Toronto with marked neighborhoods.

Then, using Foursquare API and endpoint explore with specified ID code category, we obtained venues that are categorized as food venues. We used a limit of 100 venues and a radius 500 meters. Venues were then grouped in a data frame and relative frequencies computed. Top 5 venues for 3 neighborhoods together with their relative frequencies are displayed in Figure 2.



```
----Adelaide, King, Richmond----
                 venue  freq
0           Restaurant  0.08
1                 Café  0.08
2       Sandwich Place  0.06
3     Asian Restaurant  0.06
4  American Restaurant  0.05


----Agincourt----
                       venue  freq
0  Latin American Restaurant  0.25
1             Sandwich Place  0.25
2         Chinese Restaurant  0.25
3             Breakfast Spot  0.25
4          Afghan Restaurant  0.00


----Agincourt North, L'Amoreaux East, Milliken, Steeles East----
                       venue  freq
0           Asian Restaurant   1.0
1          Afghan Restaurant   0.0
2  Middle Eastern Restaurant   0.0
3                Pizza Place   0.0
4         Persian Restaurant   0.0
```

Figure 2: Sample of venues and their relative frequencies.