Capstone Project - The Battle of the Neighborhoods

Data Science Specialization Capstone by IBM/Coursera

Table of contents

Introduction:

Toronto being financial capital of Canada, it attracts new immigrants as well as travellers. Indians are among top immigrants to Canada. As Indian population is rising in Toronto, want of Indian food & grocery is increasing. A restaurant chain name XYZ, which is popular in India is interested to open an Indian restaurant in Toronto. They want to analyse how the existing Indian restaurants in Toronto are distributed among the neighborhoods.

With the help of data science, we will classify Toronto neighborhoods with high, moderate & low density of Indian restaurants.

Data

Looking at the problem, we need:

- details of neighborhood in Toronto
- existing indian restaurants in the neighborhood

We will use following database/information

- details of neighborhood in Toronto by web scraping from Wikipedia using postal codes
- number of restaurants and their type and location in every neighborhood will be obtained using **Foursquare API**

Methodology

In this project, the following methodology will be adopted

- we will first obtain the Toronto neighborhood data, clean it & convert it to dataframe.
- Then we will get the co-ordinates of these neighborhoods.
- After merging location data with neighborhood we will get the nearby venues details using Foursquare API.
- After filtering venue details to only Indian restaurants, we will get the number of restaurants in each neighborhood.
- We will explore the restaurant density & cluster it with high, moderate & low density areas & plot it on map for better visualization.

Data Cleaning

Neighborhood

We first obtain neighborhood data using postal codes of Toronto from Wikipedia & build a dataframe from the same.

	Postcode	Borough	Neighborhood
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

Then we need to remove column Borough & Neighborhood values as not assigned.

	Postcode	Neighborhood
2	M3A	Parkwoods
3	M4A	Victoria Village
4	M5A	Harbourfront
5	M6A	Lawrence Heights
6	M6A	Lawrence Manor

Then we have obtained the geo co-ordinates of each neighbourhood & have merged the same in the above dataframe. So, final dataframe is look like this:

	Postcode	Neighborhood	Latitude	Longitude
0	M3A	Parkwoods	43.753259	-79.329656
1	M4A	Victoria Village	43.725882	-79.315572
2	M5A	Harbourfront	43.654260	-79.360636
3	M6A	Lawrence Heights	43.718518	-79.464763
4	M6A	Lawrence Manor	43.718518	-79.464763

First look of map with all neighborhoods as data points

Exploratory Data Analysis

We got the nearby venue details of all the neighborhood from F**oursquare API.** The venue details are obtained as follows:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

Then we performed grouping on the dataframe & added all venue categories as columns in the data. After filtering the data with Indian restaurant, we have taken mean of values & the final table is as follows:

	Neighborhood	Indian Restaurant
0	Adelaide	0.01
1	Agincourt	0.00
2	Agincourt North	0.00
3	Albion Gardens	0.00
4	Alderwood	0.00

Clustering the data

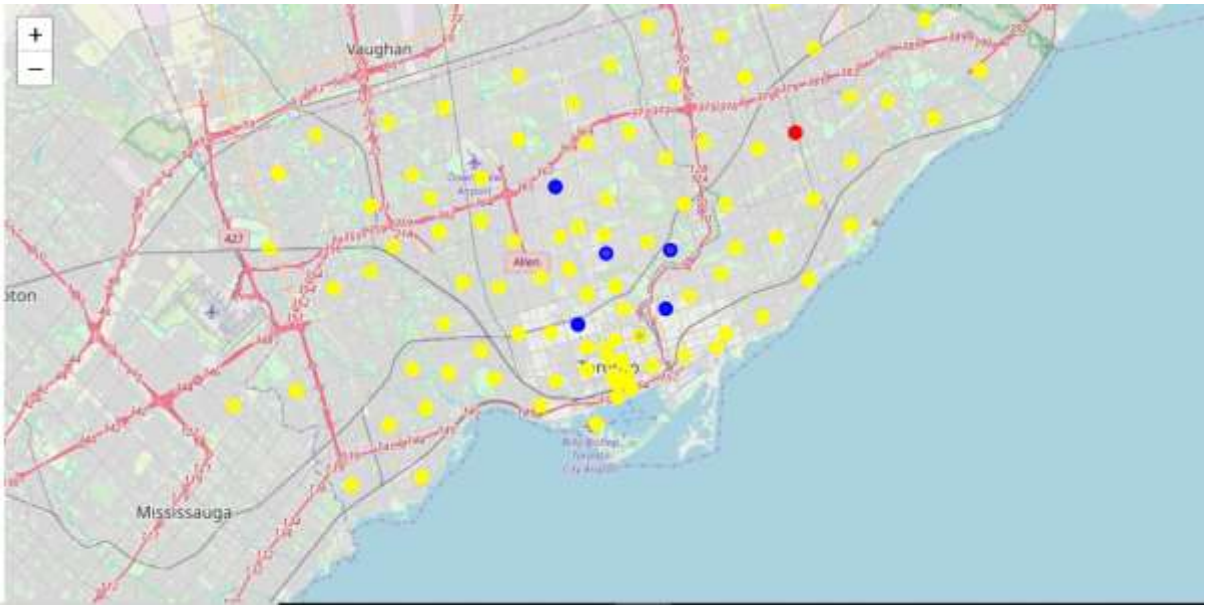
We use **k-means** method for clustering the data with 3 clusters. **k-means clustering** aims to partition n observations into **k clusters** in which each observation belongs to the **cluster** with the nearest **mean**, serving as a prototype of the **cluster**.

Our final data for output is look like this:

	Neighborhood	Indian Restaurant	Cluster Labels	Postcode	Latitude	Longitude
0	Adelaide	0.01	0	M5H	43.650571	-79.384568
1	Agincourt	0.00	0	M1S	43.794200	-79.262029
2	Agincourt North	0.00	0	M1V	43.815252	-79.284577
3	Albion Gardens	0.00	0	M9V	43.739416	-79.588437
4	Alderwood	0.00	0	M8W	43.602414	-79.543484

Result

After clustering the data into 3 clusters, we have plotted the points on map & each cluster is represented by separate colour.



Conclusion

There are few neighborhoods like Bedford Park, North Midtown, Yorkville etc with good number of Indian Restaurants. Whereas, other areas have very low numbers. To open a specific cousin restaurant mainly depends on the demand for that particular food. Opening a restaurant where existing Indian Restaurants are very low or none in numbers means low competition but it also means low demand. As the company wants to identify areas where there are high number of Indian restaurants compare to others, we have identified Cluster 2 for them.

Let's explore the cluster 2 neighborhood

	Neighborhood	Indian Restaurant	Cluster Labels	Postcode	Latitude	Longitude
9	Bedford Park	0.043478	2	M5M	43.733283	-79.419750
17	Cabbagetown	0.022222	2	M4X	43.667967	-79.367675

	Neighborhood	Indian Restaurant	Cluster Labels	Postcode	Latitude	Longitude
31	Davisville	0.030303	2	M4S	43.704324	-79.388790
94	Lawrence Manor East	0.043478	2	M5M	43.733283	-79.419750
113	North Midtown	0.043478	2	M5R	43.672710	-79.405678
133	Riverdale	0.023810	2	M4K	43.679557	-79.352188
164	The Annex	0.043478	2	M5R	43.672710	-79.405678
167	The Danforth West	0.023810	2	M4K	43.679557	-79.352188
174	Thorncliffe Park	0.111111	2	M4H	43.705369	-79.349372
198	Yorkville	0.043478	2	M5R	43.672710	-79.405678