Clustering the Toronto Neighborhoods

Husam Alhwadi

20th December 2020

Introduction:

1.1 background:

Toronto city is multicultural city where people from different countries live and enjoy visiting different restaurants which provide different cuisines from their original home countries.

1.2 Problem:

This diversity in cuisines makes knowing the number of restaurants and their variety of cuisines is so important for either investors who are looking to open new restaurant or for new immigrants who are looking to live in neighborhoods which have high number of restaurants and variety of cuisines.

1.3 **Interest**:

This report is an attempt to address the challenge which faces mainly two groups of people:

Group 1 The investors who plan to open new restaurant in one of Toronto city neighborhoods and don’t know exactly which neighborhoods they can choose it considering how big and diversity Toronto city is.

<u>Group 2</u> The immigrants who are looking to live in neighborhoods with adequate number of restaurants and variety of cuisines.

<u>Group 3</u> In addition to these two groups this report may be sound interesting for people who are interested in such type of research from different fields like Data science filed, city planning ..etc.

1.4 Proposed Solution:

Toronto neighborhoods will be clustered based on number of restaurants and variety of cuisines by using Data science unsupervised model to provide clear insight about Toronto neighborhoods from perspective of their restaurants and cuisines.

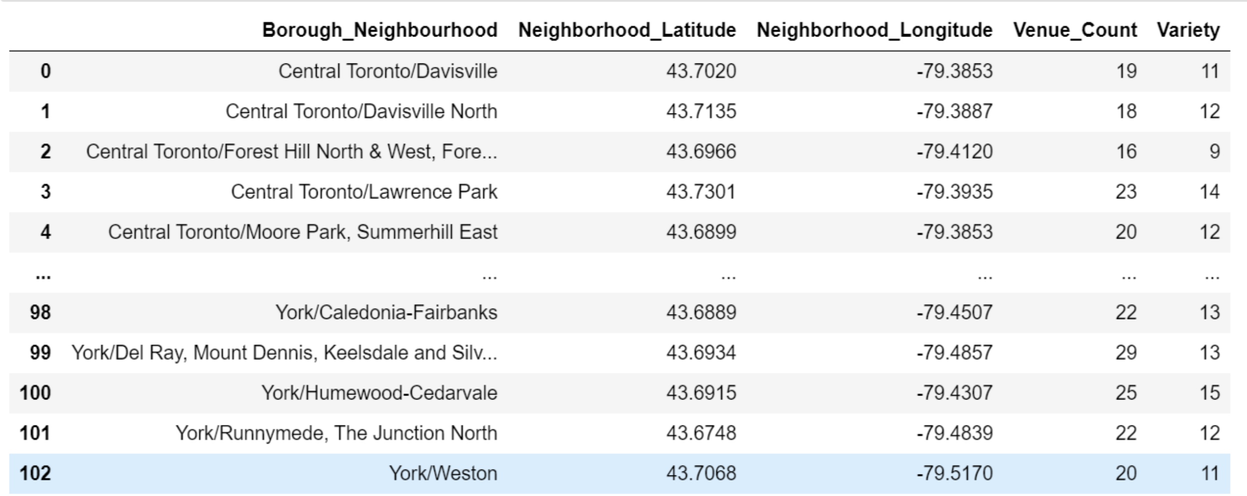
2. Data

2.1 Data Source

1-Data from Foursquare system (https://foursquare.com/) Foursquare free user account will be used to retrieve the available and relevant data from Foursquare system via using Foursquare API's.

2- List of Toronto boroughs and neighborhoods from Wikipediae (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

2.2 Data Sample:



	Borough_Neighbourhood	Neighborhood_Latitude	Neighborhood_Longitude	Venue_Count	Variety
0	Central Toronto/Davisville	43.7020	-79.3853	19	11
1	Central Toronto/Davisville North	43.7135	-79.3887	18	12
2	Central Toronto/Forest Hill North & West, Fore...	43.6966	-79.4120	16	9
3	Central Toronto/Lawrence Park	43.7301	-79.3935	23	14
4	Central Toronto/Moore Park, Summerhill East	43.6899	-79.3853	20	12
...	...	...	...	...	...
98	York/Caledonia-Fairbanks	43.6889	-79.4507	22	13
99	York/Del Ray, Mount Dennis, Keelsdale and Silv...	43.6934	-79.4857	29	13
100	York/Humewood-Cedarvale	43.6915	-79.4307	25	15
101	York/Runnymede, The Junction North	43.6748	-79.4839	22	12
102	York/Weston	43.7068	-79.5170	20	11

2.3 Data Processing

1-Getting the Latitude and longtiude for each neighborhood by using pgecode library (refer to link below for more details): https://pypi.org/project/pgeocode/

2- Explore and adding the venues for each neighborhood by using foursquare end point:

https://api.foursquare.com/v2/venues/explore

3- Adding the category for each venue by using foursquare end point:

https://api.foursquare.com/v2/venues/explore

3-Filtering the restaurants by using venue category filed.

4- Preparing DataFrame with columns for Toronto neighborhoods, neighborhood latitude, neighborhoods longitude, Venue Id, Venue Category.

5- Grouping the DataFrame records by using neighborhoods column to get the number of total restaurants for each neighborhood.

5- Process the column of venue category to split the categories for restaurants for each neighborhood.

6- Count the number of unique restaurant categories for each neighborhood.

2.4 Data Cleaning

1- Removing the duplicated categories within each neighborhood, duplicated categories occurs when dataframe record is grouped so to avoid multi-counting for same restaurant category removing duplicated categories has been proceed.

2- Continuous checking for dataframe to assure its freeness of null, unknown cells after manipulation its records.

2.5 Features Selections

Number of restaurants (Venue_Count) and the number of unique restaurants categories (Variety) will be selected as features to build the clustering model.

2.6 **Data Limitations**

dataset used in this report is so restricted and don’t represent the actual number of venues and their categories in Toronto neighborhoods Due to the limitation of free foursquare account which is used in this report as this user is eligible for 950 regular call type and 50 premium calls type per day and to retrieve up to 100 items when explore venue end point API’s is used.

3. **Methodology**

3.1 Exploratory Data Analysis

Explanatory analysis for Foursquare data has taken place at first step to understand which type of data can be retrieved adequately by using foursquare free account user type (please refer to 2.6 Data Limitation section to understand the limitation for this type of user) , accordingly decision has been taken to:

1- Use the available data at Venue/Explore end point because we can run up to 950 calls per day which will be sufficient to collect the relevant data.
2- Removing the rating data from data scope because of limitation of foursquare free account user which can get up to 50 rating per day which will not be sufficient.
3- Using number of restaurants and number of unique cuisines to segment the Toronto neighborhoods based on.

3.2 **Statistical Testing**

Populations for each neighborhood been collected in order to understand the correlation between neighborhood population and number of restaurants, however its not been involved in this study because lack of actual number of restaurants that can be retrieved by using foursquare free account user type, nevertheless; its high recommended to involve population in this study if actual number of restaurants and their cuisines been collected.

3.3 **Machine Learning**

Unsupervised machine learning model has been used to cluster the neighborhoods.

k-means cluster model has been selected in this study because it fulfills the requirements and Elbow method has been applied to select the optimal number of clusters (more details is available at notebook).

4. Results

Toronto neighborhoods have been clustered into 3 clusters which are:

A- Neighborhoods which have low number of restaurants and variety of cuisines like Downtown Toronto/Harbourfront East, Union Station, Toronto Islands Cluster neighborhood.

B- Neighborhoods which have medium number of restaurants and variety of cuisines like Scarborough/Upper Rouge neighborhood

A- Neighborhoods which have high number of restaurants and variety of cuisines like North York/Willowdale, Newtonbrook neighborhood.

5. Discussion

1- Its noticeable that as we move far from coast the number of restaurants and their variety increases, For example areas which are close to cost like downtown have low number of restaurants and variety whereas areas like North York/Willowdale have high number of restaurants and variety.

2-As stated early these results are based on limited volume of dataset which been retrieved from foursquare system by using free account type hence some neighbor maybe not clustered correctly due to this limitation in dataset

6. **Recommendations**

1-Its high recommended to reconduct this study by using full dataset associated with population for each neighborhood to cluster the neighborhoods properly.
2-Breaking down this study at level of restaurant cuisine for example Chinese or Indian cuisines to shedlight on specific category of subgroups likewise for other types of venues categories.
3-Involving other data like rating, tips…etc which can help model to outline the differences between different restaurants and venues.