Applying Data Science Methodology to Discover Business Opportunities in Toronto, Ontario, Canada

Qusay Sellat

Nov, 2019

1. Introduction

Toronto is the provincial capital of Ontario and the most populous city in Canada. It is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

However, it seems to be very hard for people outside the city to define the main characteristics of each neighborhood inside Toronto that distinguish it from the other neighborhoods.

1.1. Problem

This project analyses the data of Toronto’s various neighborhoods in order to discover the best businesses that can be done in each of the respective neighborhoods by doing a cluster analysis.

1.2. Interest

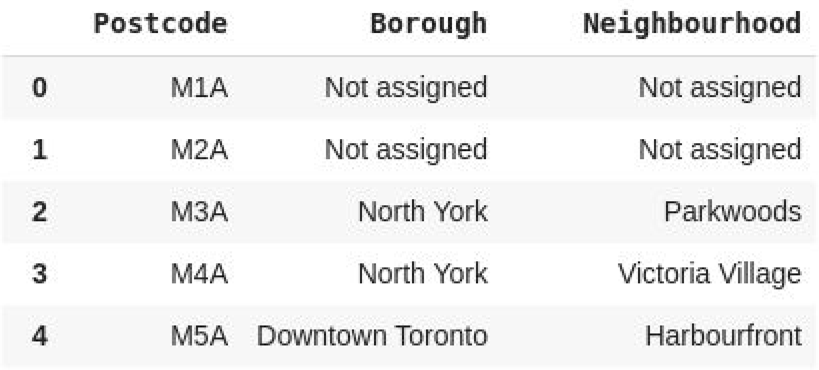
Some people outside Toronto have plans to invest in the city but they don’t know how and where. Therefore, this project can be helpful for those who have dreams of having a successful business into the city of Toronto.

2. Data Collecting and Pre-processing

2.1. Data Collecting

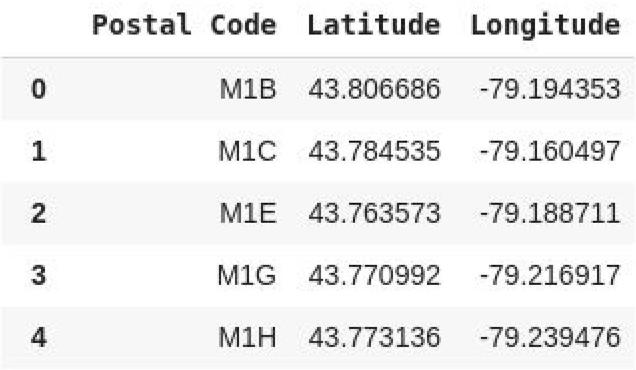
The data needed for the clustering of Toronto’s neighborhoods according to their characteristics can be collected from many sources. We store collected data into Pandas dataframes. The data used in this project comes from the following sources:

- Data about Toronto’s postcodes, boroughs, and neighborhoods. This data is read using Pandas library by scraping it from wikipedia url. In this report we will call this data **Neighborhoods** dataset.



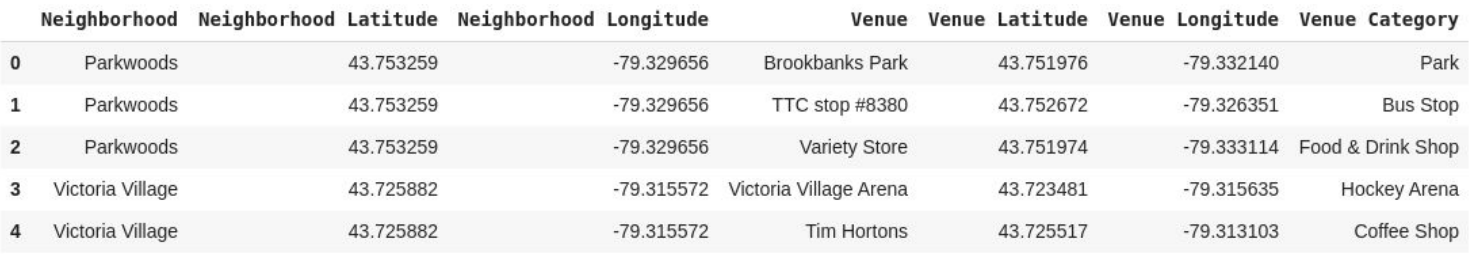
	Postcode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront

- Data about the coordinates (latitude and longitude) of various Toronto’s neighborhoods. This data can be collected using Geocoder Python package. However, This package is highly unreliable and I couldn’t use it to download the data. Fortunately, Coursera provided the data via a reliable link. In this report we will call this data **Coordinates** dataset.



	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

- Data about the various types of venues located in each of the neighborhoods represented in the data collected in the above steps. For the purpose of collecting this data, we use FourSquare API. In this report we will call this data **Venues** dataset.

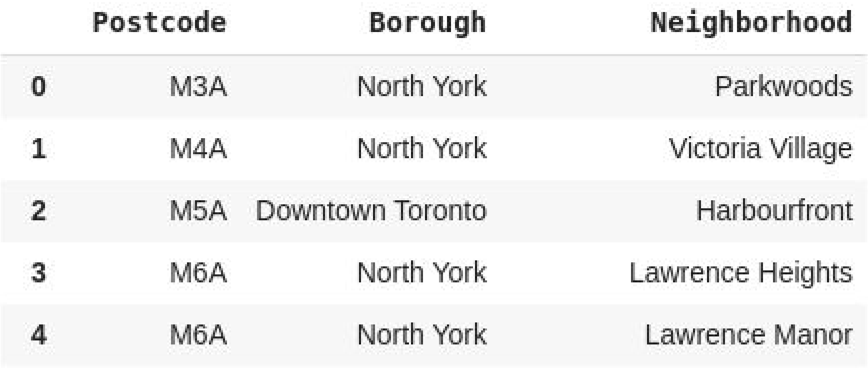


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	TTC stop #8380	43.752672	-79.326351	Bus Stop
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

2.2. Data Pre-processing

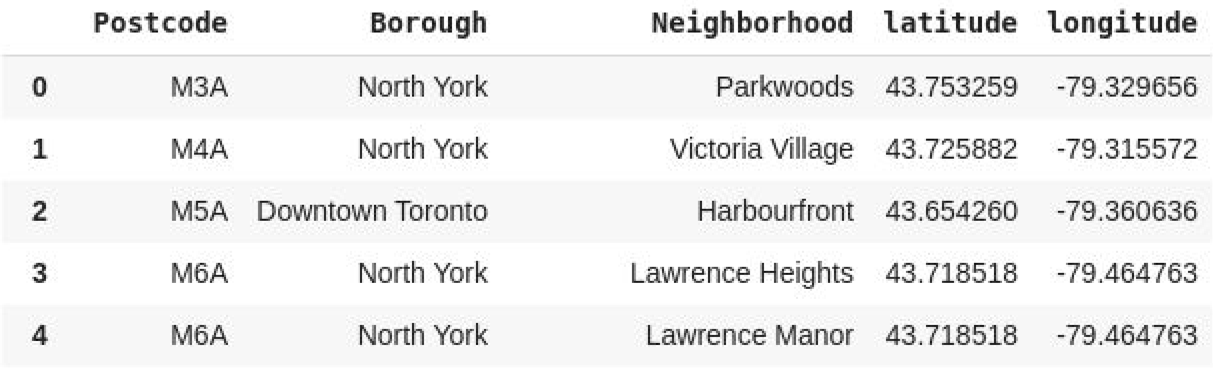
In order to benefit from the collected data, we must have some processing done.

- First of all, we notice that some cells under Neighborhoods dataset are Not assigned. For convenience, any row containing Not assigned Borough was dropped. Also, any Not assigned Neighborhood is replaced by the corresponding Borough.



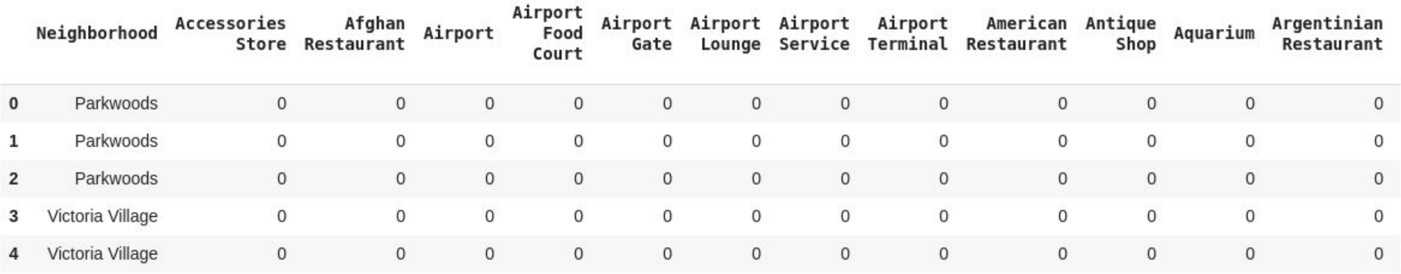
	Postcode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M6A	North York	Lawrence Heights
4	M6A	North York	Lawrence Manor

- Then we prepare Neighborhoods data set to contain latitude and longitude information by using Coordinates data set. The final **Neighborhoods Coordinates** dataset looks like this:



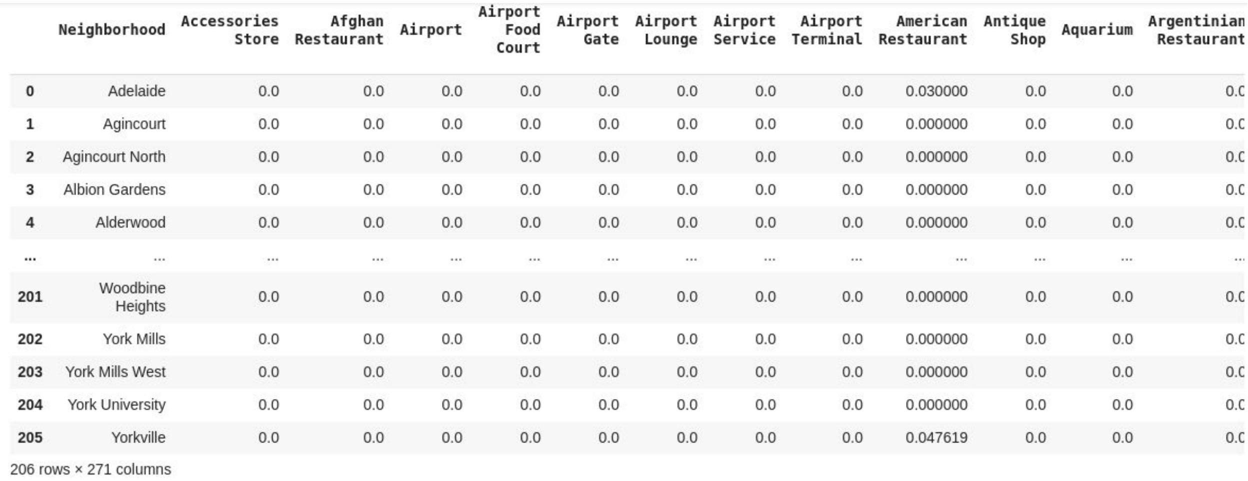
	Postcode	Borough	Neighborhood	latitude	longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Heights	43.718518	-79.464763
4	M6A	North York	Lawrence Manor	43.718518	-79.464763

- In order to do the clustering process on the neighborhoods of Toronto, we must first make a dataset that can be fit into a clustering algorithm like KMeans. One approach is to use **one-hot encoding t**o represent each of the returned avenues with the corresponding neighborhoods. So first we convert to one-hot encoding:



	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Argentinian Restaurant
0	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0
1	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0
2	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0
3	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0
4	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0

- We notice that the number of avenues is 4379 of 270 type. For each neighborhood, in order to fit the KMeans algorithm, it's important to know what avenues located in each neighborhood. For this reason we group the one-hot encoded dataset of the avenues by the neighborhood by applying the mean function to represent the importance of each avenue in describing the neighborhood. So the final dataset looks like the following, we will call it **Neighborhoods Avenues** dataset (because of sparsity nature of data, most entries are zero), and it will be the input of our clustering algorithm.



	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Argentinian Restaurant
0	Adelaide	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.030000	0.0	0.0	0.0
1	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
2	Agincourt North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
3	Albion Gardens	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
4	Alderwood	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
201	Woodbine Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
202	York Mills	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
203	York Mills West	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
204	York University	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
205	Yorkville	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.047619	0.0	0.0	0.0

206 rows × 271 columns