Capstone Project - The Battle of Neighbourhoods – Toronto

Applied Data Science Capstone
by IBM/Coursera

Author: Aman Agrawal

February 29, 2020

Table of contents

1. Introduction and Data description

Introduction/Business Problem

Coffee is a rich source of antioxidants that slow down the aging process of tissues and effectively protect the body against health loss. One cup of coffee has been shown to contain even more antioxidants than a glass of grapefruit, blueberry, raspberry or orange juice. Investor already has many cafes of his own brand in the world. Now he intends to conquer a new market. Considering the above, the investor intends to open a new cafe in Toronto. Unfortunately, he doesn't know the city well and doesn't know where to open a café. He wants to know if there are coffee shops in all the neighbourhoods. That's why he wants to open a business where there are already cafés, but they are not very popular. As a result, the business problem is:

- **Where to open a new a Successful Cafe in Toronto?**

I would like to do some research and recommend them the best place based on the number of cafes in different districts of Toronto. In order to solve this business problem, we intend to merge Toronto districts into a cluster in order to recommend locations.

Data

To consider the objective stated above, we can list the below data sources used for the analysis.

- Districts of Toronto Wikipedia page was scraped to pull out the necessary information;
- Coordinate data for each Districts of Toronto obtained through Nominatim search engine for OpenStreetMap data;

In order to investigate and target recommended locations in different locations depending on the presence of facilities and necessary objects, we will access the data through the FourSquare API and arrange it as a data frame for visualization. By combining data about districts in Toronto and data about amenities and essential facilities surrounding such properties from the FourSquare API, we will be able to recommend an appropriate location.

2. Methodology

The methodology in this project consists of two parts:

- **Data Understanding & Data Preparation**: Process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. Once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases.

- **Data Preparation & Data Exploration**: Process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. Once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases. It is necessary to visualise the districts of Toronto.
- **Modelling**: To help people find similar neighbourhoods in the safest borough we will be clustering similar neighbourhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 5 for this project that will cluster neighbourhoods into 5 clusters. The reason to conduct a K- means clustering is to cluster neighbourhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighbourhood.

Data Understanding & Data Preparation

Scrape the Wikipedia page and gathering data into a Pandas dataframe

To start with our analysis, we used the Beautiful Soup package to transform the data in the table on the Wikipedia page into the below pandas data frame. The data frame will consist of three columns: Postal Code, Borough, and Neighbourhood. Subsequently, we transform the data into a pandas data frame.

	PostalCode	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

Data cleaning

Only cells that have a district assigned to them will be processed. There may be more than one district in one of the postal codes. If a cell has a borough but a Not assigned neighbourhood, then the neighbourhood will be the same as the borough.

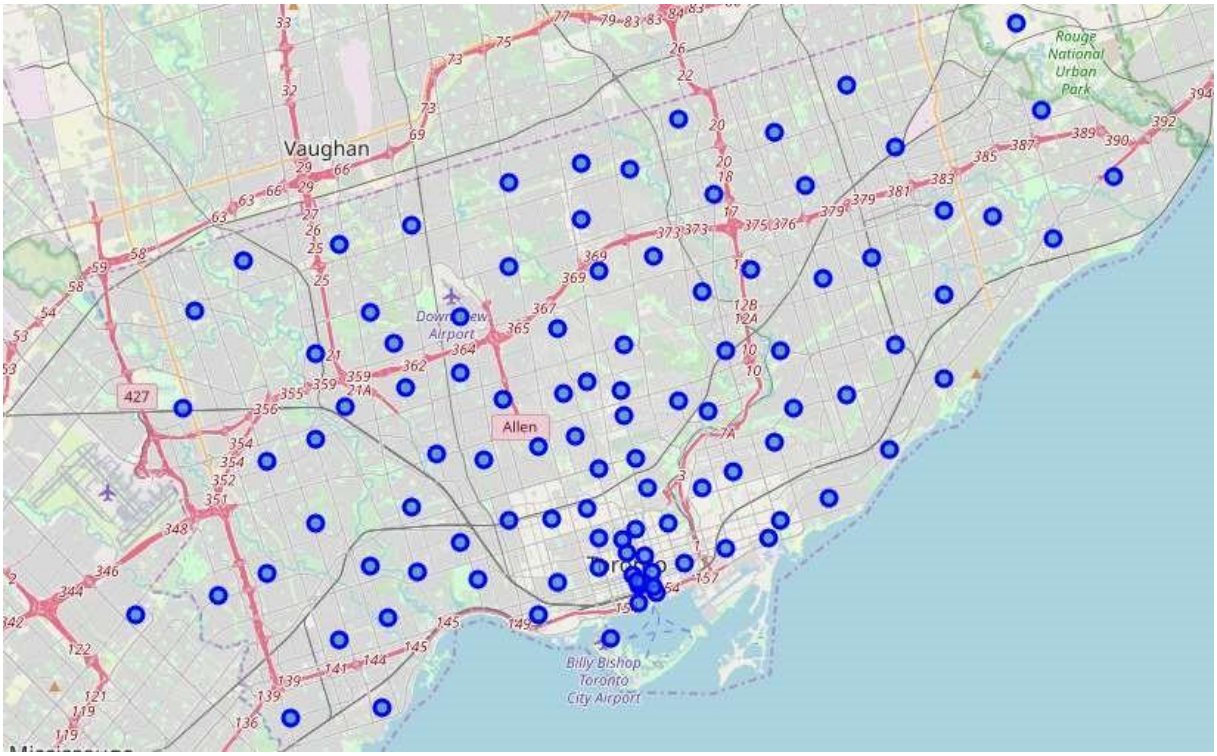
	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M6A	North York	Lawrence Heights, Lawrence Manor
4	M7A	Downtown Toronto	Queen's Park

Use geopy library to get the latitude and longitude values of Toronto

After we have built a dataframe of Toronto localities along with the district name and neighbourhood name, in order to utilize the Foursquare location data, we need to get the latitude and the longitude coordinates of each neighbourhood. It possible to export data to a csv file for easier loading later.

	PostalCode	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, Lawrence Manor	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494

Generating a map of Toronto and plotting the Neighbourhood data on it



Modelling

- Finding all the cafes within a 500-meter radius of each neighbourhood.
- Perform one hot encoding on the venues data.
- Grouping the venues by the neighbourhood and calculating their mean.
- Performing a K-means clustering (Defining K = 5)

Finding all the cafes within a 500-meter radius of each neighbourhood.

Foursquare is the most trusted, independent location data platform for understanding how people move through the real world. We have used, as a part of the assignment, the Foursquare API to retrieve information about the popular spots for each neighbourhoods of Toronto. The recommended location needs to have many eating and shopping venues nearby. Convenient public transport is also required.

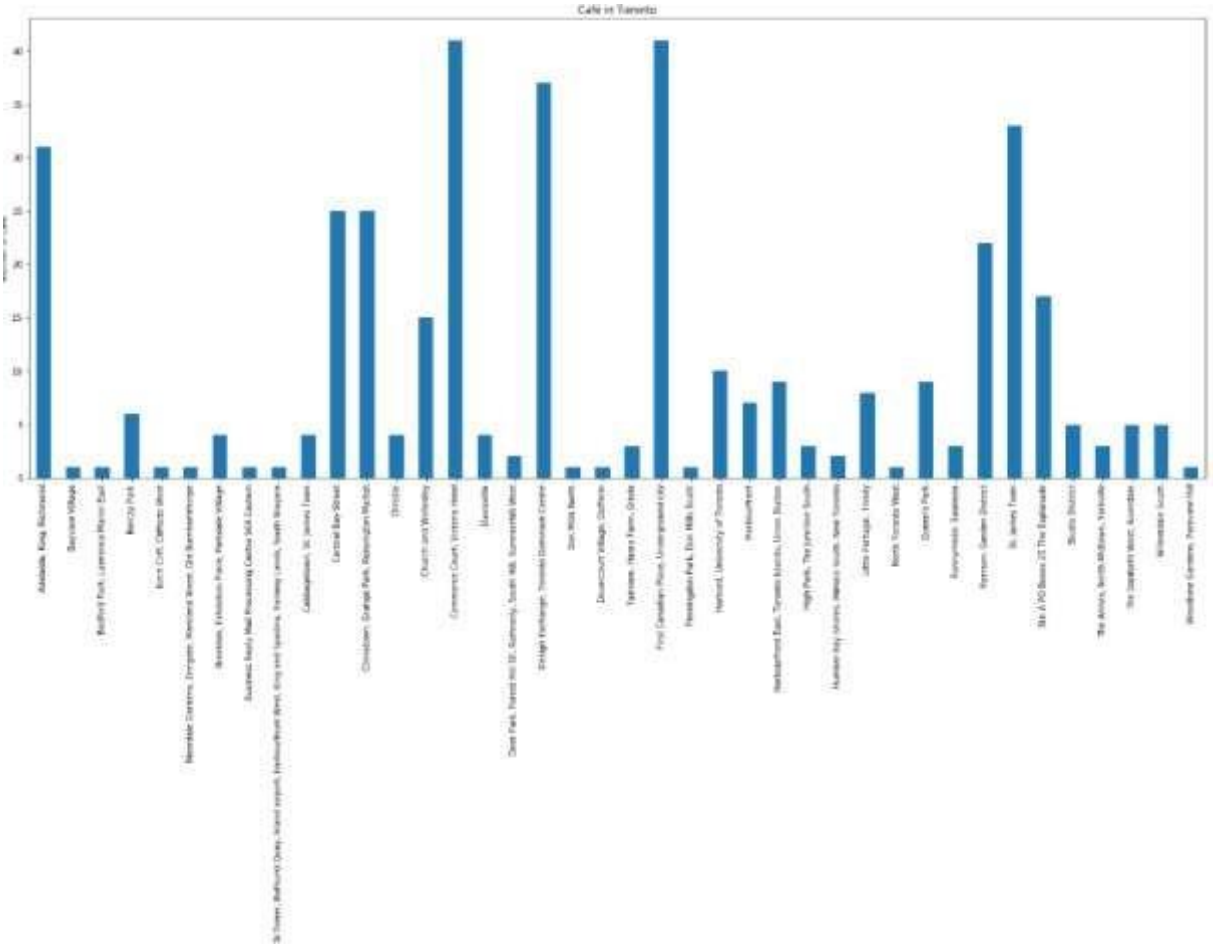
```
print('Total {} of venues are found in {} uniques categories.'
      .format(len(toronto_venues),len(toronto_venues['Venue Category'].unique())))
toronto_venues.head()
```

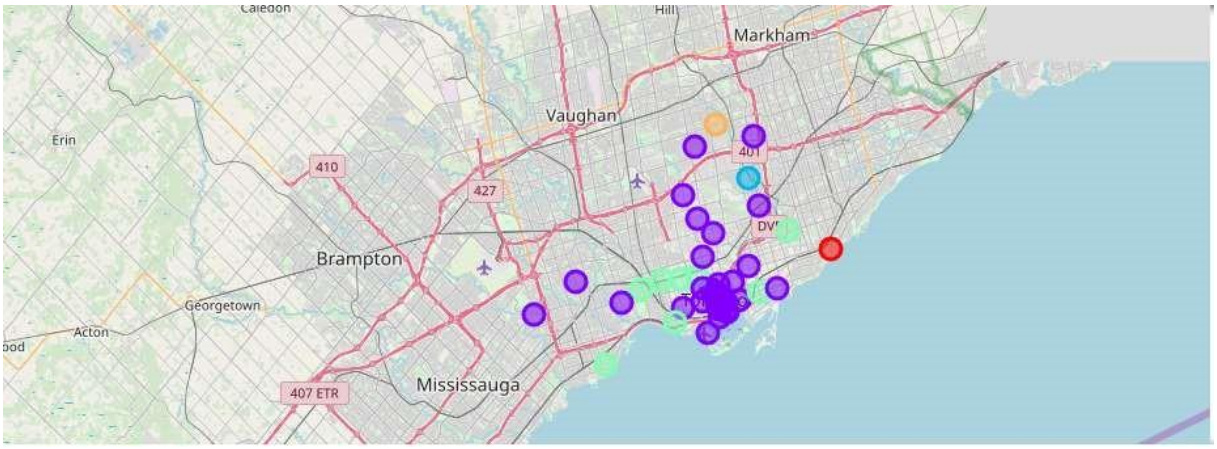
```
Total 465 of venues are found in 19 uniques categories.
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.358980	Café
1	Harbourfront	43.65426	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Café
2	Harbourfront	43.65426	-79.360636	ODIN Cafe + Bar	43.656739	-79.356503	Café
3	Harbourfront	43.65426	-79.360636	Dark Horse Espresso Bar	43.653081	-79.357078	Café
4	Harbourfront	43.65426	-79.360636	Caffe Furbo	43.649970	-79.358849	Café

All modelling process you find notebook [link]

Analysis





- Cluster 0 shape = (1, 12)
- Cluster 1 shape = (30, 12)
- Cluster 2 shape = (1, 12) • Cluster 3 shape = (8, 12)
- Cluster 4 shape = (1, 12)

Examine Cluster 0

```
k0 = toronto_merged.loc[toronto_merged['Cluster Labels'] == 0, toronto_merged.columns[[0] + list(range(5, toronto_merged.shape[1
k0
```

	PostalCode	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
58	M1N	0	Skating Rink	General Entertainment	College Stadium	Café	Cuban Restaurant	Donut Shop	Doner Restaurant	Dog Run	Discount Store	Diner

Examine Cluster 1

```
k1 = toronto_merged.loc[toronto_merged['Cluster Labels'] == 1, toronto_merged.columns[[1] + list(range(5, toronto_merged.shape[1
k1.head()
```

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Downtown Toronto	1	Coffee Shop	Bakery	Pub	Café	Park	Mexican Restaurant	Breakfast Spot	Restaurant	Chocolate Shop	Historic Site
4	Downtown Toronto	1	Coffee Shop	Gym	Park	Music Venue	Seafood Restaurant	Sandwich Place	Salad Place	Restaurant	Burger Joint	Burrito Place
5	Queen's Park	1	Coffee Shop	Gym	Park	Music Venue	Seafood Restaurant	Sandwich Place	Salad Place	Restaurant	Burger Joint	Burrito Place
9	Downtown Toronto	1	Coffee Shop	Clothing Store	Cosmetics Shop	Japanese Restaurant	Café	Bakery	Pizza Place	Bookstore	Restaurant	Diner
13	North York	1	Beer Store	Gym	Asian Restaurant	Coffee Shop	Bike Shop	Italian Restaurant	Sandwich Place	Supermarket	Café	Dim Sum Restaurant

Examine Cluster 2

```
k2 = toronto_merged.loc[toronto_merged['Cluster Labels'] == 2, toronto_merged.columns[[1] + list(range(5, toronto_merged.shape[1
k2
```

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
7	North York	2	Café	Japanese Restaurant	Basketball Court	Gym / Fitness Center	Caribbean Restaurant	Women's Store	Dim Sum Restaurant	Deli / Bodega	Department Store	Dessert Shop

Examine Cluster 3

```
k3 = toronto_merged.loc[toronto_merged['Cluster Labels'] == 3, toronto_merged.columns[[1] + list(range(5, toronto_merged.shape[1
k3.head()
```

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
8	East York	3	Pizza Place	Fast Food Restaurant	Pharmacy	Bus Line	Café	Athletics & Sports	Intersection	Gym / Fitness Center	Bank	Pet Store
25	Downtown Toronto	3	Grocery Store	Café	Park	Athletics & Sports	Nightclub	Candy Store	Restaurant	Diner	Italian Restaurant	Bank
31	West Toronto	3	Pharmacy	Bakery	Supermarket	Bar	Fast Food Restaurant	Café	Middle Eastern Restaurant	Recording Studio	Gym / Fitness Center	Grocery Store
43	West Toronto	3	Café	Coffee Shop	Breakfast Spot	Nightclub	Convenience Store	Italian Restaurant	Intersection	Restaurant	Burrito Place	Bar
54	East Toronto	3	Café	Coffee Shop	Brewery	American Restaurant	Italian Restaurant	Gastropub	Bakery	Stationery Store	Diner	Bank

Examine Cluster 4

```
k4 = toronto_merged.loc[toronto_merged['Cluster Labels'] == 4, toronto_merged.columns[[1] + list(range(5, toronto_merged.shape[1
k4
```

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
39	North York	4	Japanese Restaurant	Chinese Restaurant	Bank	Café	Women's Store	Cupcake Shop	Donut Shop	Doner Restaurant	Dog Run	Discount Store

Results and Discussion

The purpose of my experiment was to point out the right neighbourhood to open a café in Toronto. Based on the experiments, districts where cafés already exist were selected at the beginning. There are 40 of them, and then, to broaden the scope of clustering, all the popular places in the districts where the café already exists were found. It was indicated that "The biggest proportion of Café among other venues in a district in Toronto is 7 % in Woodbine Gardens, Parkview Hill." In some cases, cafés have not been identified as a popular place at all.

However, considering the grouping performed, I can recommend cluster 0 and 4 because there the cafes occupy the farthest place according to popularity. We must remember that in this experiment the distance from the centre was not considered. The investor can of course make a different choice, because the data is already prepared.

Conclusion

Different applications of this analysis are available based on a different methodology and possibly different data sources. The stakeholder problem has been resolved. This project helps the investor to better understand the area in relation to the most common places in the area. It is always helpful to use technology to be one step ahead, i.e. learn more about places before opening a new coffee shop in the district. The future of this project involves considering other factors, such as the cost of living in the areas concerned, in order to draw up a short list of neighbourhoods based on a predefined budget.