Opening an Italian Restaurant in Toronto

Sixuan Zhu

September 19th, 2020

# Introduction

## Background

Toronto, the capital of the province of Ontario, is a major Canadian city along Lake Ontario’s northwestern shore. It's a dynamic metropolis with a core of soaring skyscrapers, all dwarfed by the iconic, free-standing CN Tower, and a various selection of global cuisine restaurants. Sam Xiong is looking to open a new Italian restaurant in Toronto City.

## Problem

With all these selections, the competition for opening a new restaurant is tough. Sam would like to know which location would have the highest probability of success. In order to provide a data-driven and fact-based solution, we will need to take into accounts attributes such as the venue popularity of the neighborhood, the characteristics of each neighborhood, the similarity and dissimilarity of each neighborhood, and the top common venues within each neighborhood.

# Data acquisition and cleaning

## Data Source

Two main data sources for this project would be Toronto neighborhood data and Foursquare location data. For the Toronto neighborhood data, a [Wikipedia page](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) exists that has all the information we need to explore and cluster the neighborhoods in Toronto. Foursquare is the location data provider that share information about businesses and attractions around a certain a location point. We will need to retrieve Foursquare location data around Toronto City through API.

## Data cleaning

The first problem with the dataset is to get Toronto neighborhood data from the Wikipedia page. I scraped the Wikipedia page and wrangle the data, cleaned it, and then read it into a pandas dataframe so that it is in a structured format. Then I combined the neighborhood under the same postal code. Here the new data frame contains 103 neighborhoods.

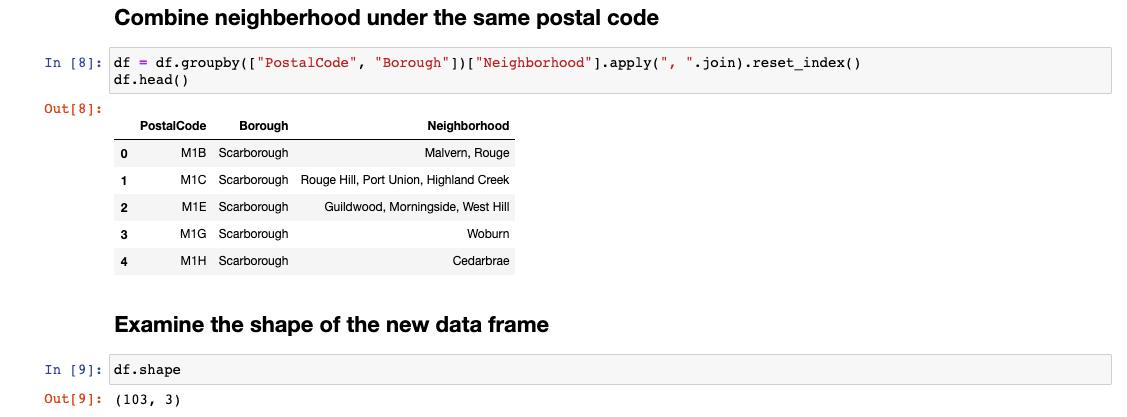
The second step is to get coordinates data and join them to the neighborhood dataset. I have obtained a dataset containing the coordinates data. The step is an intermediate step to connect the neighborhood dataset and Foursquare location data.

The third step is to get specific venue information for all neighborhood in Toronto city from Foursquare through the API program. Then I have written a function to create the API request and then return only the relevant information for each nearby venue, i.e. venue name, venue location, venue category, venue lat, venue lng, and venue name. I also limited the number of venues returned to 100 to avoid the list get too exhaustive.

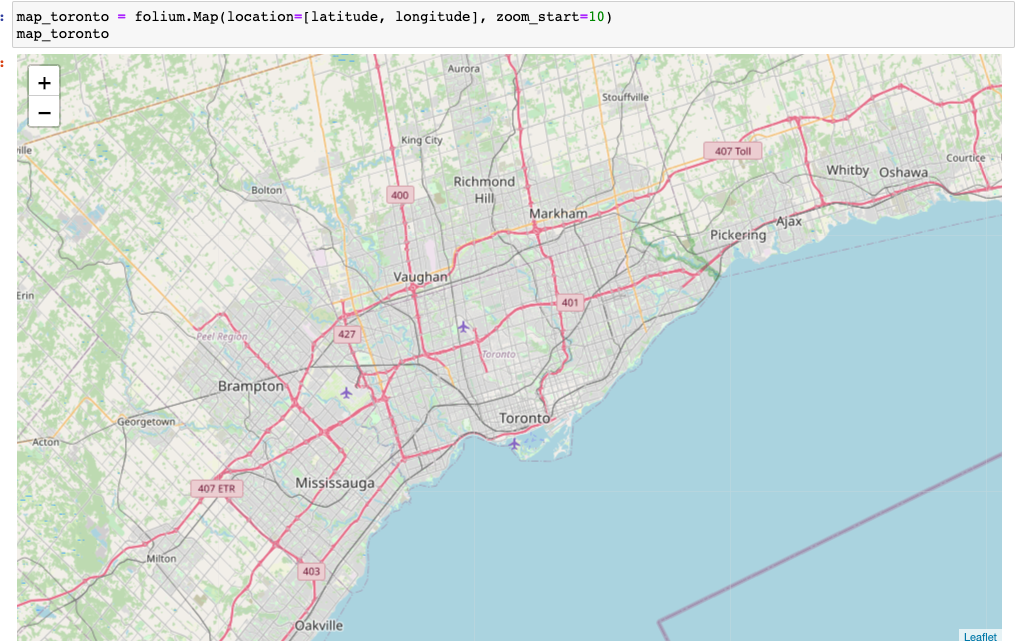
With all the data digested into pandas data frames, I can start to perform exploratory analysis and use clustering technique to analyze the characteristics of a neighborhood.

# Exploratory Analysis

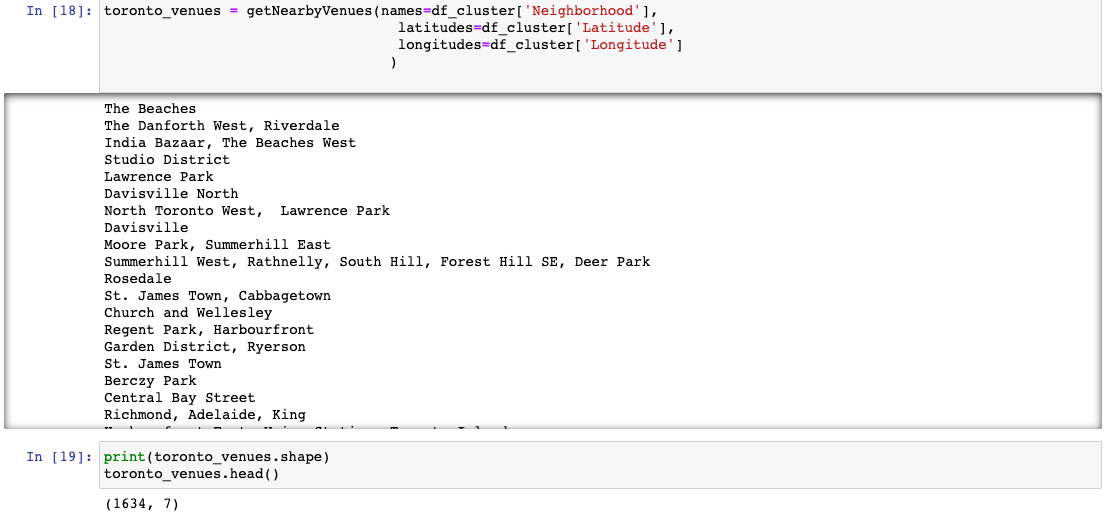
After the initial data cleaning steps, I would like to see how many neighborhoods I am looking at for the purpose of this project. I noted that, in total, 103 neighborhoods are filtered out and all of them are sit in Toronto city.



I would also like to see the overall map of Toronto city and have a generate idea of how the neighborhoods are scatter around in the city.



After this, I populated the venues of each neighborhood through Foursqaure API. And it turned with in total 1634 venues for all the 103 neighborhoods.



Among the 1634 venues, there are in total 238 unique categories, one of which category is Italian Restaurants. To see the characteristics of each neighborhood, I printed out each neighborhood along with the top 5 most common Venues.



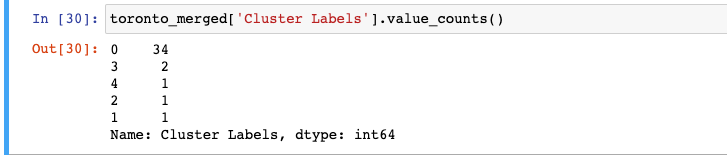
For example, in Christie neighborhood, the most popular venue is grocery stores with a frequency of 0.25. Italian restaurants are also popular here with a frequency of 0.06.

Now, with this understanding, I proceeded to apply K-means clustering technique to group the neighborhood that are similar to each other. I intend to find a cluster with many commercial venues such as park, café, restaurants, etc, which will be suitable for Sam to open a Italian restaurant.

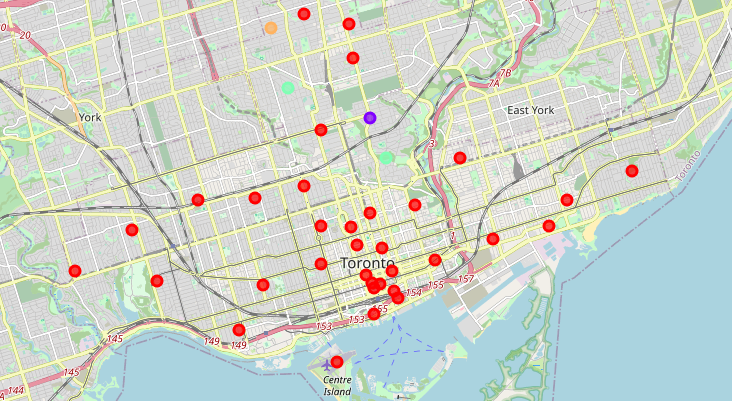
# Method

I have set the cluster number as 5 and applied the kmeans function from the scikit.learn package. 

Now that I have cluster labels, I would like to see the value counts of each cluster, i.e. the counts of neighborhoods under each cluster. I have noted the following:



So I have a main cluster with 34 neighborhoods in it and all the other clusters as niche clusters with 1 or 2 neighborhood in it. As a next step, I have visualized the clusters on top of the map of Toronto City:



As seen from the map, the outlier clusters are mostly located in central Toronto. The main cluster (cluster 0) are concentrated in Downtown Toronto and West Toronto.

# Results

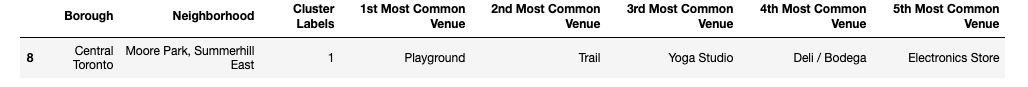
In order to provide advice on where to open a new Italian restaurant, I need to first analyze the characteristics of each neighborhood.

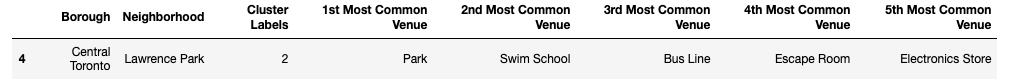
Cluster 0:



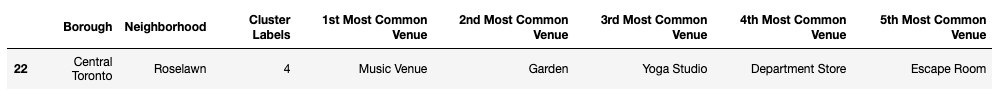
After a closer investigation, I have noted that neighborhoods in cluster 0 are mostly commercial areas with a lot of restaurants, bars, parks, coffee shops, cafes, etc. It would probably be a good candidate for opening a new restaurant.

Cluster 1, 2, 3, 4:









What I have observed from 1,2,3,4 clusters are that restaurants are in general not popular in these neighborhoods. Some popular venues in these clusters are department stores, swim schools, music venues, playgrounds, etc.

# Discussions

As discussed above, neighborhoods in cluster 1 would be good candidates for opening a new restaurant. But the second thing I would consider is: whether an Italian restaurant is popular in a neighborhood. As the data frame presented, I only would like to see neighborhood with Italian restaurants listed as one of the top most common venues. Thus, I have filtered the neighborhoods in cluster one and it returned with a list of 4 neighborhoods with Italian Restaurant listed as one of the Top 3 most common venues:



# Conclusion

With these insights, I think I will be able to provide Sam with advice on where to open a new Italian restaurant in Toronto City. The best candidate as shown in the above graph would be among these four neighborhoods: The Danforth West Riverdale, Davisville, Central Bay Street, Stn A PO boxes.

However, in order to make the decision for opening the restaurant, I believe there should be more factors that needs to be considered such as cost, convenience of transportation, etc, which are currently not considered in my analysis and posing a limitation on the conclusion that I have reached. But I strongly believe this would definitely be a great starting point for Sam to explore the choices of opening an Italian restaurant.