Predicting an optimal location of Coffee Shop

This is a final project of a professional certification course "IBM Data Science"

Introduction

Business problem

The objective of this project is to determine an optimal location for opening a new coffee shop in Toronto. The location significantly affects the profit of any company. Thus we could select the "right" location and get much more money. Using clustering approach allows us to solve this business problem.

Target audience

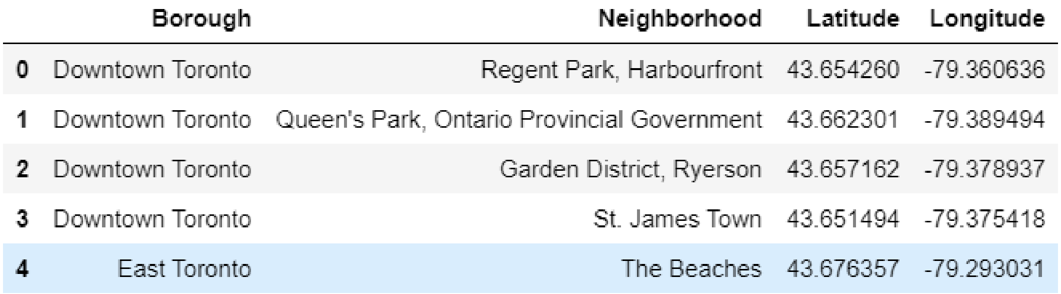
The target audience of this project is any person who cares about establishing new business connected with coffee shops.

Data

To solve this problem we will need the following data:

- The neighborhood data, which describe names and postal codes of neighborhoods.
- Geolocation data with longitude and latitude of neighborhoods.
- Venues data (especially data related to existing coffee shops).

Table 1. Example of data



	Borough	Neighborhood	Latitude	Longitude
0	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
1	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
2	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
3	Downtown Toronto	St. James Town	43.651494	-79.375418
4	East Toronto	The Beaches	43.676357	-79.293031

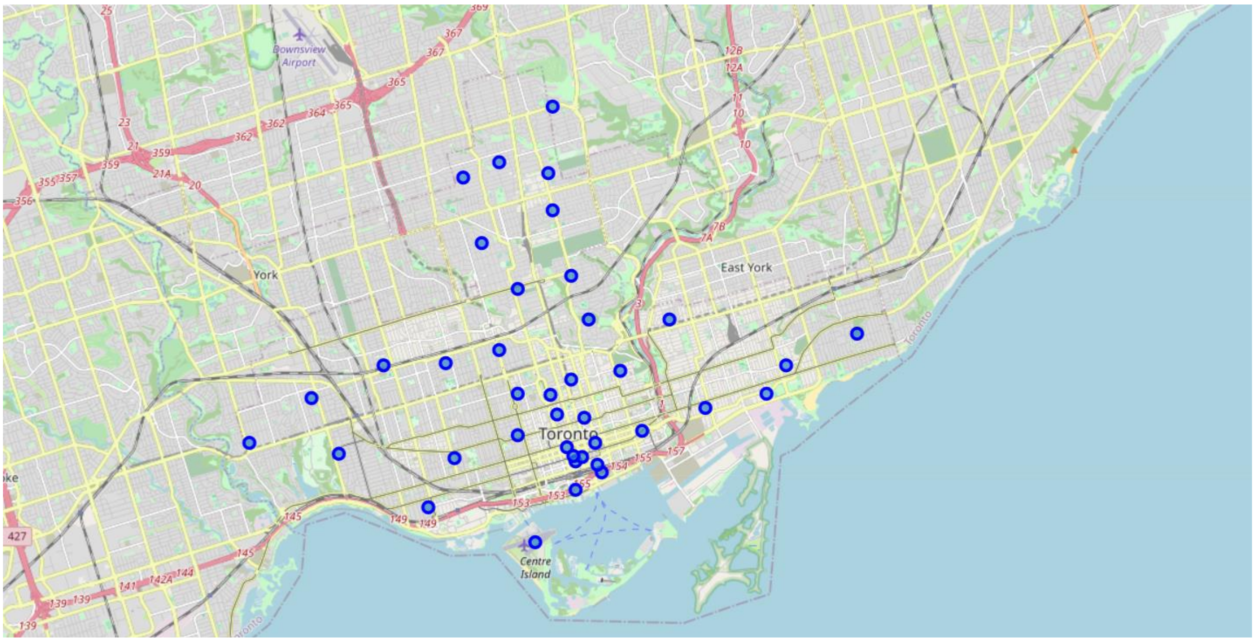
Data sources

Data can be obtained from these sources:

- The neighborhoods data were taken from [Wikipedia]
- Neighborhoods Geolocation data were taken from the provided [.json file]
- Venues data were obtained using [Foursquare API]

Methodology

First of all, we need to get the list of neighborhoods in Toronto. The data can be obtained from Wikipedia by web-scraping via beautiful soup library. Then we can add geospatial data with longitudes and latitudes. After that, we can create pandas dataframe and try to visualize data on the map using Folium.

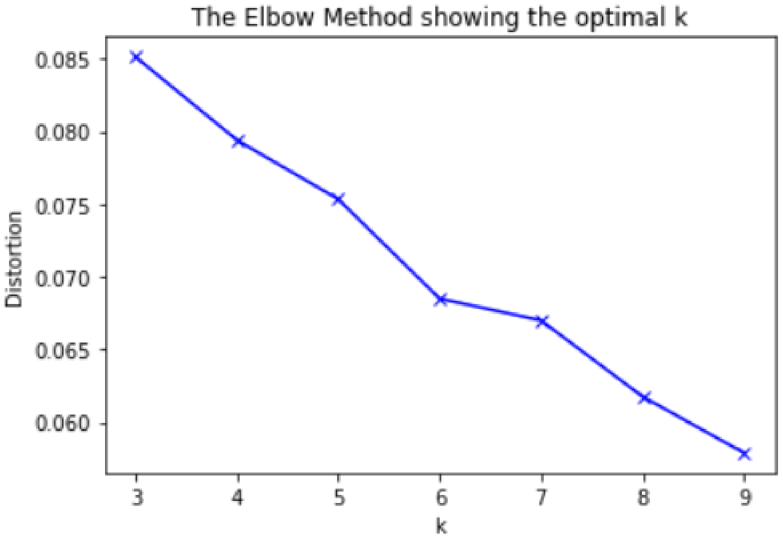


In the next step, we will use Foursquare API to get the top 100 venues that are within a radius of 1000 meters. For this step Foursquare Developer Account is needed. We can make API calls to Foursquare in order to get venue data. Foursquare API returns the venue data in JSON format and we will extract the venue name, venue category, venue latitude, and longitude. After that, we can check how many venues are in a neighborhood and what is the frequency of venue category in a particular neighborhood.

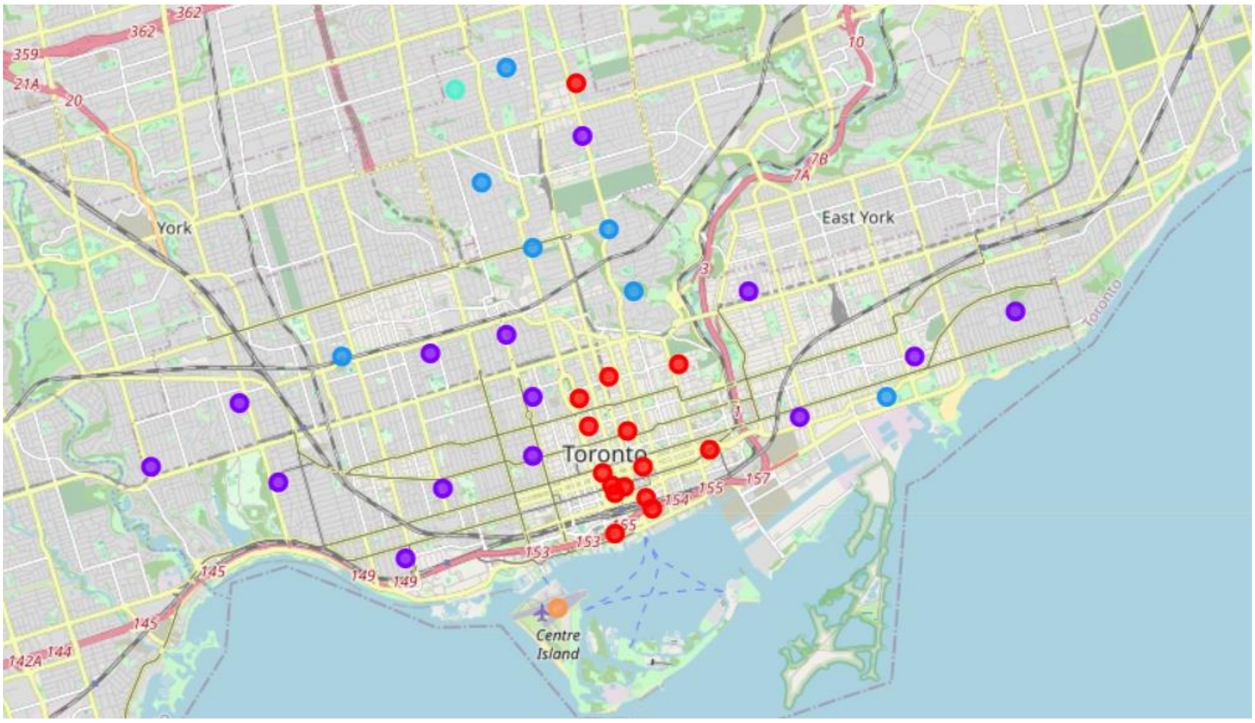
Now we can get 5 most popular venue categories in each neighborhood and define columns of our train dataset like unique categories from all 4 most popular.

Lastly, we will perform clustering by using k-means clustering. K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. It is one of the simplest and popular unsupervised machine learning algorithms.

We can determine the number of clusters k using the elbow method. In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. The elbow method gives to us k=6.



Now we can perform clustering, get distribution of neighborhoods by clusters, and visualize it using Folium.



Results

We can see that clusters 0, 4, 5 have a very high number of coffee shops in most neighborhoods, otherwise, clusters 1, 2, 3 has a low number of coffee shops. It provides a great opportunity for establishing a new business.

Based on the map, most coffee shops are concentrated in the center of the city, meanwhile, the suburban area still has very few coffee shops.

Discussion

This notebook provides a very simple way to visualize and analyze data, but the next steps can go deeper and analyze the relationship between cafes and other types of establishments. You can also use additional datasets, for example, with data on the cost of rent and the index of the popularity of the area

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 6 clusters based on their similarities, and lastly providing recommendations to the relevant to open a new coffee shop.