IBM

Data Science Project Week 5

Report

Introduction

This is the IBM Applied Data Science Capstone Project Certification Week 4 Assignment. I am creating a scenario for a concept that there may not be enough Indian Restaurants in Toronto Area. Therefore it might be a great opportunity for a business-man who is new shifted in Canada. As the Indian food is popular among Asian community, so this business-man might think of satarting his business in areas where asian community resides. With the purpose in mind, finding the location to open such a restaurant is one of the most important decisions for this person, so this project will help him in understanding the market properly.

Business Problem

The objective of this capstone project is to find the most suitable location for the business man to open a new Indian Restaurant in Toronto,Canada. By using data science methods and tools along with machine learning algorithms such as clustering, this project aims to provide solutions to answer the business question.

Target Audience

The business people and foodbloggers who are looking for indian restaurants or starting a new restaurant in Toronto

Data

To solve this problem, we will need below data:
● List of neighborhoods in Toronto, Canada
● Latitude and Longitude of these neighborhoods
● Venue data related to Indian restaurants. This will help us find theneighborhoods that are more suitable to open an Indian Restaurant.
Extracting the Data
● Scrapping of Toronto neighborhoods using Wikipedia portal
● Getting Latitude and Longitude data of these neighborhoods using Geocoder package
● Using Foursquare API to get venue data related to these neighborhoods

Methodology

First, we need the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

I did the web scraping by utilizing pandas HTML table scraping method as it is easier and more convenient to pull tabular data directly from a web page into the data frame.

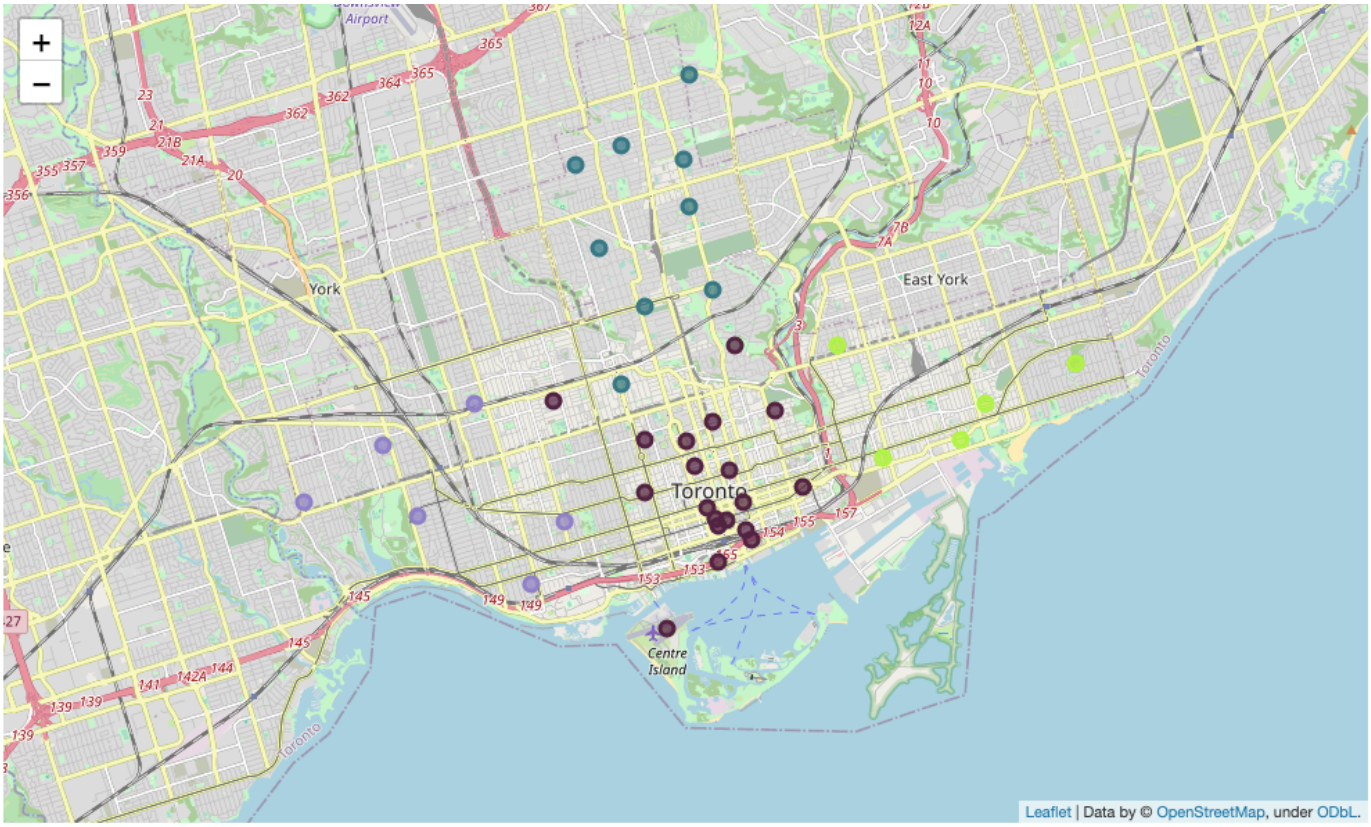
However, it only gives us the list of the neighborhood and postal codes and now we need to get their coordinates using Foursquare Api

I used the CSV file provided by IBM team to match the coordinates of Toronto neighborhoods. After gathering these coordinates, I visualize the map of Toronto using Folium package to verify whether these are correct coordinates. Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude, and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Here, I made a justification to specifically look for “Indian restaurants”. Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for “Indian food”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

RESULT

CLUSTER



The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Indian restaurants are in each neighborhood:

- Cluster 0: Neighborhoods with the less number of Indian restaurants.
- Cluster 1: Neighborhoods with no Indian restaurants.
- Cluster 2: Neighborhoods with a more number of Indian restaurants

The results are visualized in the above map with Cluster 0 in green, Cluster 1 in blue, Cluster 2 in red