Coursera IBM Data Science Capstone Project

# Introduction/Business Problem

## Introduction

This is the Coursera IBM Data Science Capstone project for the professional certificate. In this project I will be performing a data analysis on the best places to open an Italian Restaurant – PIZZA! in a populous city Toronto, Canada. Italy is not only known for pizza, but a lot of traditional and rich Mediterranean cuisine. However, during the colonization, traditional Italian food have been modernized, but there exist few natives in Canada who are yearning for good Italian Food. The beauty of Italian food, it just consists of two-four main ingredients that evolve into mouth-watering taste. So now it’s time to get Milan in Toronto.

## Business Problem:

The objective of this capstone project is to find the most suitable location to open a traditional Italian Cuisine in Toronto, Canada. By using various Data Science and Machine Learning methods and tools such as clustering and folium. This project will provide the best clusters of location to open the restaurant based on the neighborhood data of Toronto.

# Data Section

To Solve this problem, I got access to these data

1. The Toronto Neighborhood Postal Codes
2. Access to Latitude and Longitude of Toronto Neighborhoods
3. Data related to current Italian Restaurant.

# Extraction of Data

1. We will be scraping data using Beautiful Soup via the Toronto Codes website
2. Get the Latitude and Longitude of these data using a geocoder
3. Use the Foursquare application API to get the custom Italian venue data in the neighborhoods.

# Methodology

1. First, I need to get 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​​

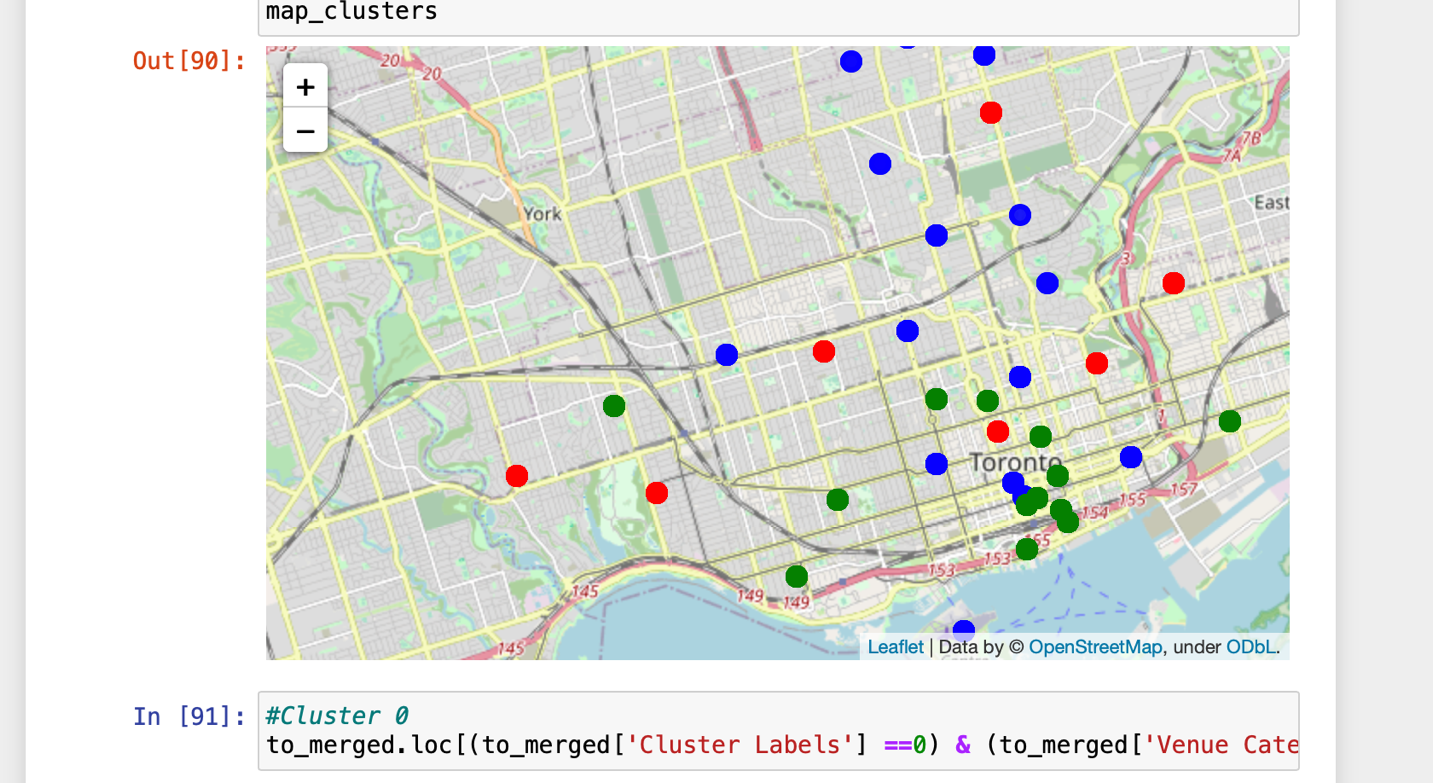
I did the web scraping by utilizing panda’s xml table scraping method as it is easier and more convenient to pull tabular data directly from a web page into the data frame.

However, it is only a list of neighborhood names and postal codes. I need to get their coordinates using Geocoder to use Foursquare to to get the list of venues near these neighborhoods. To get the coordinates, I used the CSV file provided by IBM team to match the coordinates of Toronto neighborhoods in previous exercise. 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 for week 5 to access the credential 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. Finally, I prepare the cluster map.

A close up of a map

Description automatically generated

This picture is the general venues in Toronto, Canada



This is the cluster map after for Italian Restaurants using the Foursquare API

In the folium map, I made the code to specifically look for Italian 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 to get accurate results.

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 “Italian food”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

RESULT

CLUSTERS

A close up of a map

Description automatically generated

Most of the Italian Restaurants are in the cluster 2, and then cluster 0

the Neighborhoods First Canadian Place, Underground city in Cluster 1 are like the best places to open a traditional Italian Restaurants.

Looking at nearby venues it seems cluster 0 and 1. I think it might be a good location as there are not a lot of Italian restaurants in these areas. Therefore, this project recommends these neighborhoods to open a hot Italian Restaurant.