IBM

Data Science Capstone Project Report

INTRODUCTION

This is a capstone project for IBM Data Science Professional Certificate. In this project, I am creating a hypothetical scenario for a concept that there may not be enough Italian Restaurants in Toronto Area. Therefore, it might be a great opportunity for an entrepreneur who is based in Canada.

BUSINESS PROBLEM

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Italian 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: In Toronto, if an entrepreneur wants to open an Italian Restaurant, where should they consider opening it?

TARGET AUDIENCE

The entrepreneur who wants to find the location to open an authentic Italian restaurant.

DATA

To solve this problem, we will need below data:

- List of neighbourhoods in Toronto, Canada
- Latitude and Longitude of these neighbourhoods
- Venue data related to Italian restaurants. This will help us find neighbourhoods that are more suitable to open an Indian Restaurant.

EXTRACTING THE DATA

- The scrapping of Toronto neighbourhoods via Wikipedia
- Getting Latitude and Longitude data of these neighbourhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighbourhoods

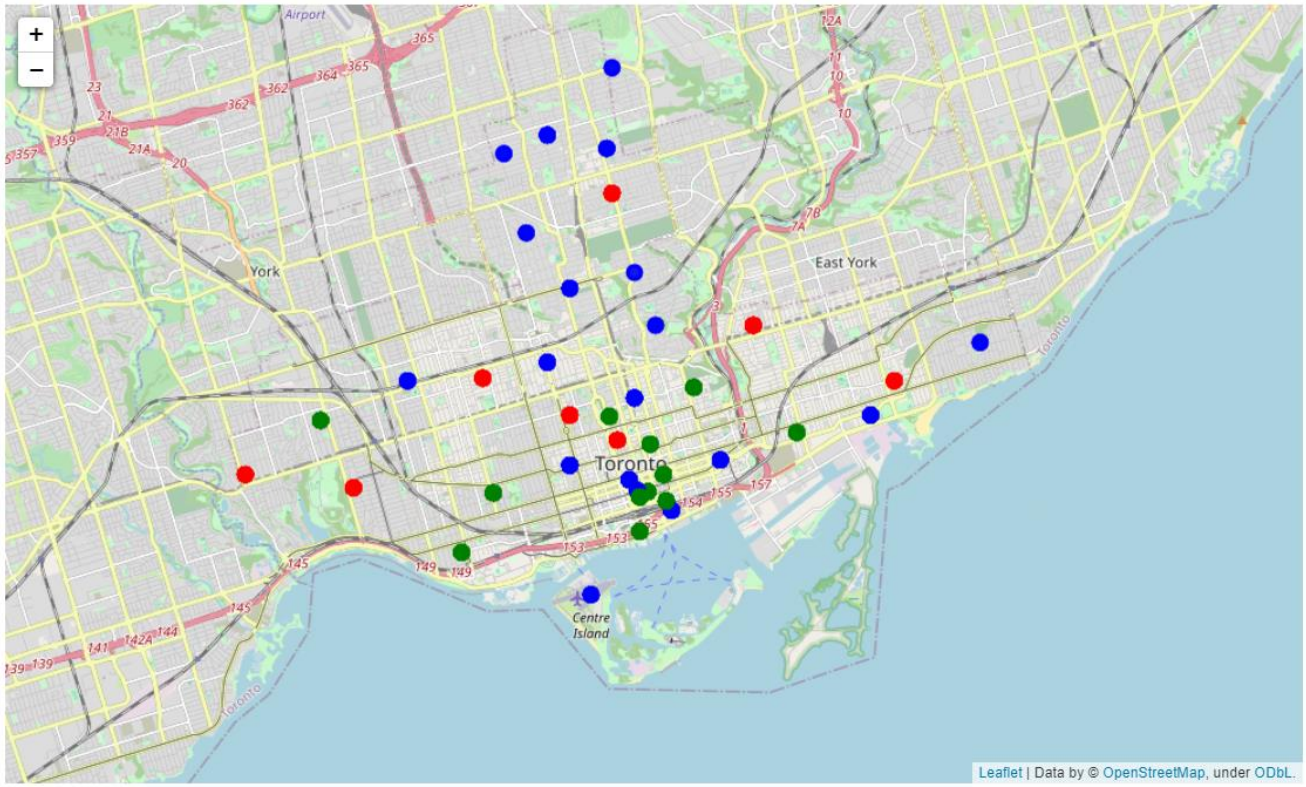
METHODOLOGY

First, I need to get the list of neighbourhoods in Toronto, Canada. This is possible by extracting the list of neighbourhoods 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 is only a list of neighbourhood names and postal codes. I need to get their coordinates to utilize Foursquare to pull the list of venues near these neighbourhoods. To get the coordinates, I tried using Geocoder Package but it was not working so I used the CSV file provided by IBM team to match the coordinates of Toronto neighbourhoods. 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 analyse each neighbourhood by grouping the rows by neighbourhood 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 “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. 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 neighbourhoods 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



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

RECOMMENDATIONS

- Cluster 0: Neighbourhoods with the smaller number of Italian restaurants.
- Cluster 1: Neighbourhoods with no Italian restaurants.
- Cluster 2: Neighbourhoods with a greater number of Italian restaurants.

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

Majority of Italian restaurants are in Cluster 2. Through this analysis we can recommend the neighbourhoods in cluster 0 and 1 to open Italian restaurants.