IBM
Data Science Capstone Week 4

Introduction

This is a capstone project for the IBM Data Science professional Certificate. In this project, I will be testing my hypothesis that there are not enough Japanese restaurants in the Toronto area. With this lack of Japanese restaurants, this may provide a good opportunity for entrepreneurs to open new restaurants. I will be finding the most suitable location for a new restaurant to open.

Business Problem

The main goal of this project is to find the best location to open a new Japanese restaurant in Toronto, Canada. I will be using data science methods and machine learning algorithms to provide an answer to this question.
Question: "Where should I open a new Japanese restaurant in Toronto, Canada?"

Target Audience

Any person who wants to find the best place to open a Japanese restaurant.

Data

We will need the below data to solve the problem:

1. List of neighborhoods in Toronto, Canada
2. Latitude and Longitude of the neighborhoods
3. Venue data related to Japanese Restaurants

Extracting the Data

1. Scraping data of Toronto neighborhoods located in Wikipedia data set. This will be done with the imported library of BeautifulSoup4 which is a web scraping library.
2. Getting the latitude and longitude information of the neighborhoods using the Geocoder package.
3. Using Foursquare API to get venue data related to these neighborhoods.

Methodology

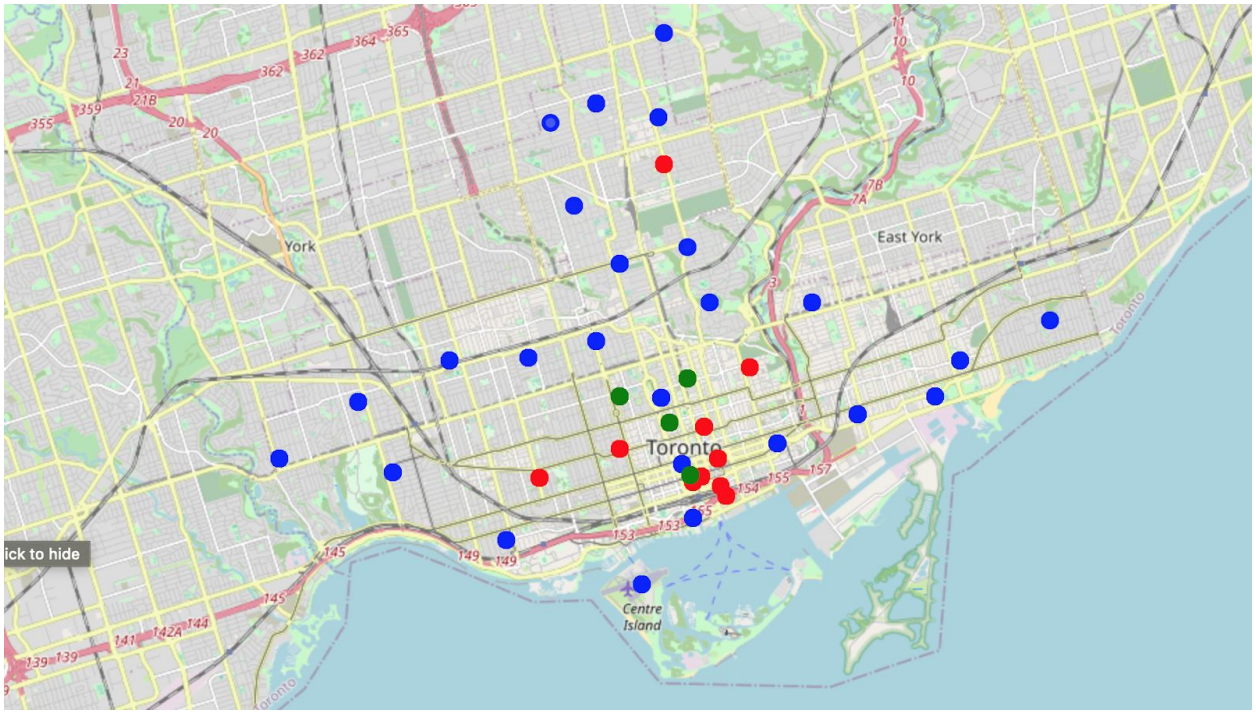
First I needed to gather a list of the neighborhoods within Canada. This can be done by extracting the information from this wikipedia link:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

I used the pandas library built in HTML table scraping method to input the information into a dataframe. This gave me a dataframe without coordinate information. Thus, I used the Geocoder package to obtain a list of latitude and longitude and then merged this list to the dataframe that was previously built. Next, I used the Folium library to create a map to visualize the neighborhoods in Toronto, Canada. Then I used the Foursquare API to access venue information around Toronto. Finally, I took the average of the unique venues to find how often each venue occurred in the neighborhood. This gave me insight into how many Japanese Restaurants there were.

I performed a K-means clustering algorithm on the data with specific venues equal to Japanese Restaurants. The K-Means Clustering algorithm identifies k number of centroids, and then allocates every point in the data to the nearest cluster while keeping the centroids as small as possible. This is an unsupervised machine learning algorithm. I used 3 clusters based on their frequency of occurrence for Japanese Food. This will allow me to recommend the ideal location for a new Japanese restaurant.

Results

Clusters:



This is the map generated by the K-Means clustering algorithm. The blue color indicates neighborhoods with the majority of Japanese restaurants.

Cluster 0: Blue, with a majority of Japanese Restaurants

Cluster 1: Green with no Japanese restaurants

Cluster 2: Red with less Japanese restaurants than Cluster 0.

Recommendations

I would recommend that a Japanese restaurant open up in either Cluster 1, or Cluster 2. This is due to the fact that these neighborhoods do not have Japanese restaurants, or very little Japanese restaurants. A good location would be Central Bay Street, Church and Wellesley, and The Danforth West, Riverdale.