Capstone Project
The Battle of Neighborhoods Week 5 Part 2

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 Indian Restaurants in Toronto Area. Therefore it might be a great opportunity for an entrepreneur who is based in Canada. As Indian food is popular among the Asian community, so this entrepreneur might think of opening its business in areas where the 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 entrepreneur and I am designing this project to help him find the most suitable location.

BUSINESS PROBLEM

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

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 neighborhoods that are more suitable to open an Indian Restaurant.

DATA SOURCES

- Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- Geocoder package
- Foursquare API

METHODOLOGY

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 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 neighborhood names and postal codes. I need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. 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 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.

THE MAP