Data Science Project for WK4

Introduction:

In this project, I will help an entrepreneur to discover the location for operating a Thai restaurant in Toronto. According to the weather and diversity in Toronto, it is a good opportunity to operate a Thai restaurant in Toronto. As Thai food is popular around the world, plus it is easily acceptable as well. According the data, there is not much Thai restaurants in Toronto. I will find out a place to open a Thai restaurant by cluster method.

Business Problem:

To find out the best place to open a Thai restaurant, I used data from Wikipedia to find out postcode in Toronto. Plus, data from Foursquare to find out all Thai restaurants in Toronto. These data could help me to identify where is the most Thai restaurants are located to avoid competitors and find out the suitable location.

List of data:

1. A postcode list of Toronto
2. List of neighborhoods in Toronto
3. Longitude and Latitude of neighborhoods.
4. Data related to Thai restaurant in Toronto, accessed by Foursquare API

Methodology:

Firstly, I have to get the list of postcodes and neighborhoods in Toronto. We extracted the list of postcodes from Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

Secondly, I use pandas' dataframe function to generate a table from website, which involves postcodes and neighborhoods. Table will be easier to pull data and compare them.

I also need to access neighborhoods' coordinates. To achieve this, I utilize Foursquare API to access venues, which are related to neighborhoods. I utilize CSV document which provided by IBM to get coordinates for neighborhoods in Toronto.

Thirdly, I visualize the map of Toronto by using Folium package to show the neighborhoods on map. I utilize Foursquare API to pull out a list, which includes top 100 venues within 500 meters radius. From Foursqaure API, I could pull out vary information involves venues' name, categories, latitude and longitude. Through data, I could identify data difference, group data by neighborhoods and take the mean on the frequency of occurrence of each venue category. This way is the first to prepare clustering for later.

I filtered specifically 'Thai restaurants. Then clustered neighborhoods into 3 clusters by K-means cluster method.

According to this method, I could identify the best location for operating a Thai restaurant.