Location determination for a new upmarket Sushi Restaurant in the City of Toronto Project

Dataset

The dataset used for analysis for this project was collected from publicly available data.

How the dataset used for analysis was created

The dataset was created by merging the geospatial data containing latitude and longitude data for postal codes (http://cocl.us/Geospatial_data) with scraped data (using the BeautifulSoup4 library) from the wikipedia table for postal codes and neighborhoods ('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:M'); the common key being the postal code. After cleaning, deleting 'non-assigned' postal codes and dropping the postal code column which is not needed for analysis, the final dataset was written out to the file csv.

This file was created for the assignment portion (Segmenting and Clustering Neighborhoods in Toronto) of this Capstone and was reused in the interest of saving a bit of time.

I. Original datasets

1- First source: http://cocl.us/Geospatial_data
The file obtained from the source URL provides a list of postal codes for the City of Toronto (first three letters of the six letter Canadian postal code) and the associated GPS coordinates (latitude and longitude).

2- Second source: 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:M'
The file obtained from the source URL provides a list of postal codes for the City of Toronto (first letter of the prefix – M) and the associated names for boroughs and neighborhoods.

II. Method

1. Download the csv file from the source URL mentioned in section I above containing the postal code prefixes and GPS coordinates.
2. Scrape the table containing the postal code prefixes and associated neighborhood information from the source URL mentioned in section I above.
3. Load the csv data into pandas dataframe.
4. Merge the two dataframes (inner join) with postal code as key.
5. Drop those rows that have ‘Not assigned’ in the Borough column and no data in the Neighborhood column.
6. Reset the index for the dataframe.
7. Ensure that the dataframe is clean.
8. The postal code column can be dropped as it will not be used for analysis.
9. Store the cleaned dataframe to csv which will be used to read back into a dataframe as required.

Author: Suresh Jacob
Date: 18 May, 2020