Peer-graded Assignment: Capstone Project - The Battle of Neighbourhoods (Week 1)

1. Introduction/ Business Problem

1.1. Background

Yoga is becoming an increasing popular activity with statistics showing that the number of Americans practicing Yoga grew by 50% between 2012 and 2016[1]. The statistics also show that 1 in 3 Americans have tried Yoga in the last 6 months and spend a total of $16 billion on classes and equipment each year. Yoga has a number of health benefits including improving flexibility, building muscle strength, perfecting posture and preventing cartilage and joint breakdown[2].

1.2. Problem

My client has come to me for advice on a suitable location to open a new Yoga studio in Toronto. They would like to discover how many Yoga studios there currently are in the city and in which neighbourhoods they are located. They would like to know how this compares to other similar size cities, where the Yoga studios are located in these cities and what other amenities are close by. The city I will use for segmentation and clustering comparison will be New York City.

1.3. Interest

Yoga is enjoyed by all ages. The chosen location could be in the CBD area to attract office workers to keep them active during their working day. Or alternatively the location could be in the suburbs where they may attract retirees or stay at home parents during the day. My client is flexible on location and would like to locate the studio in the location where he can expect to attract the most clients.

2. Data acquisition and cleaning

2.1. Data Sources

The following sources will be used for the investigation:

1. Neighbourhoods of New York data: https://geo.nyu.edu/catalog/nyu_2451_34572
2. Neighbourhoods of Toronto data: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
3. Geospatial Coordinates of Toronto: http://cocl.us/Geospatial_data
4. Foursquare - a technology company with a massive dataset of location data. A Foursquare API developer account is required and queries can be run on chosen locations using your Foursquare API Client ID and Client Secret.

2.2. Data Cleaning

2.2.1. New York dataset (data source 1)

A dataset that contains the New York boroughs and neighbourhoods as well as the latitude and longitude coordinates is already in existence and can be used for this project. A wget command is

[1] https://www.thegoodbody.com/yoga-statistics/
[2] https://www.yogajournal.com/lifestyle/count-yoga-38-ways-yoga-keeps-fit

used to download the dataset into a json file and then transfom it into a *pandas* dataframe for analysis– an empty dataframe is set up and then the data looped through.

2.2.2. Toronto dataset (data sources 2 and 3)

There is no existing dataset showing both neighbourhoods and coordinates so this is created using 2 datasets. The Neighbourhoods of Toronto data set is imported as a *pandas* dataframe (or by using the BeautifulSoup package) and the data is cleaned and the duplicates removed. Where data is missing for a Post Code, these boroughs are removed and in the case of missing data for a neighbourhood, then the borough is used.

The Geospatial csv data file is then also imported as a *pandas* dataframe and the 2 tables merged together using Post Code which is included in both datasets.

We then have a complete table including borough and neighbourhood and the longitude and latitude coordinates.

2.3. Methodology

2.3.1. Foursquare

Foursquare API will be used to locate existing Yoga studios in each location and to explore the area around the venue. The longitude and latitude coordinates will be used on both New York City and Toronto, the search query will be set to 'Yoga' and a search radius of 4km. A URL is generated using this data along with Client ID and Client Secret. A Get request is sent as a JSON file and then transformed into a *pandas* dataframe.

Folium is then used to visualise the Yoga studios on a map of the Cities.

We can then select a Yoga studio from the results to explore the area and see what else tends to be located around a Yoga studio and see if there are any patterns. If the Yoga studios have been rated we can also obtain to help determine the most popular ones.

2.3.2. K-means Clustering

We can use K-means clustering the neighbourhoods based on similarities and in each cluster identify the top 10 venues. We can also use this to determine which neighbourhoods, if any, have a Yoga studio within the top 10.