Applied Datascience Capstone Project

Question 2 - The Battle of Neighbourhoods

By How Chih Lee

Criteria: This capstone project is worth **70%** of your total grade. The project will be completed over the course of **2 weeks**. Week 1 submissions will be worth **30%** whereas week 2 submissions will be worth **40% of your total grade**.

For this week, you will be required to submit the following:

1. A description of the problem and a discussion of the background. (**15 marks**)
2. A description of the data and how it will be used to solve the problem. (**15 marks)**

---

2. Data Acquisition and Cleaning

Describe the data that you will be using to solve the problem or execute your idea. Remember that you will need to use the Foursquare location data to solve the problem or execute your idea. You can absolutely use other datasets in combination with the Foursquare location data. So make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using, even if it is only Foursquare location data.
*This submission will eventually become your **Data** section in your final report. So I recommend that you push the report (having your **Data** section) to your Github repository and submit a link to it.*

2.1 Data Sources

Data is sourced from multiple channels.

Foursquare provides location based data on different sorts of venues obtained through crowd-sourcing. Foursquare indexes each location through the longitude and latitude of that location, and provides the data based on the endpoints selected. Regular endpoints are provided free and include basic venue firmographic data, category of venue, and venue ID. Premium endpoints require payment and include rich content such as user ratings, URLs, photos, tips, menus, hours of operation, etc. Endpoints classifications and links can be found here https://developer.foursquare.com/docs/places-api/endpoints/

To solve our Location Problem, we will use “search?” and “explore?” functions to call up data on the category of venue, geographical coordinates, hours of operation, URL links to websites (if any), menus and pricing of competing F&B outlets in the neighbourhood. In Toronto, neighbourhoods are classified by Postal Code, but Foursquare does not map Toronto locations to their postal codes. We will separately obtain Postal Codes for each neighbourhood from the following Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, After that, we will use Google Maps Geocoder API to get the geographical coordinates of each Postal Code before adding the Postal Codes to the dataframe downloaded from Foursquare.

For the average spending power of GTA residents, this is difficult to obtain as income tax returns are private and confidential in Canada. We will use average housing prices in the different neighbourhoods as a proxy for average spending power as economic studies have shown that higher housing prices have a strong correlation to higher spending power in any given neighbourhood. [**Quote source**]. Historical average housing prices are indexed by Postal Codes and can be obtained by scraping data from https://housepricehub.com.

To solve our Promotion Problem, we will use the "tips?" and "trending?" functions to call up data from Foursquare endpoints that contain tips and reviews from customers and KOLs, User ID of KOLs, their names and contacts, and also real-time trending data on popular venues the KOLs recommend. We will use the "explore?" function to call up Other Venues under the categories of "fitness centres" and "yoga studios" that attract customers of a similar profile, who are health conscious or care about the environment. These Other Venues will be approached to cooperate in directed advertising for the target customers of Client V.

2.2 Data Cleaning

Postal Code data from Wikipedia shows 103 postal codes starting with M, which is reserved for the GTA, whereas data on historical average housing prices from housepricehub.com shows only 99 postal codes starting with M. Further inspection revealed that certain postal codes have been assigned by Canada Post to high-traffic areas like an Amazon warehouse in Mississauga and a Gateway sorting facility, etc. These postal codes have no residential homes and therefore no average housing prices.

Further inspection of housepricehub.com reveals that some housing prices have null entries for certain months due to lack of property transactions, so the historical average price needs to be inferred from months with valid entries and this average price is then used to backfill null entries.

Average housing prices then have to normalised across neighbourhoods, disregarding outliers at the very high end. In certain exclusive areas, multi-million dollar mansions dominate the neighbourhood and skew average housing prices to the high side. However, this is not a good indicator for Location as there are no commercial spaces for rent in such exclusive neighbourhoods for our RCO clients to locate their outlets due to zoning restrictions. We will disregard these outliers which can lead to misleading results.

Data scraped from different Data Sources are combined together and Postal Codes with no historical average housing prices are dropped. Certain venues have identical names, which could mean they belong to the same chain store, like McDonald's or Tim Horton's. These are kept in the database and distinguished using their geographical coordinates because different outlets of the same chain may have different tips and reviews from their customers, even though their menu and pricing may be the same.

Trending data often produces a null return. This could be because venues in Toronto were not popular enough with users of Foursquare to generate enough real-time reviews and tips to allow for the generation of trending data.

2.3 Feature Selection

After Data Cleaning, there are 99 neighbourhoods in the GTA with more than 1,600 venues and more than twenty features of which about ten were selected for their relevance to solving our Problem.

The features relevant to Part 1- Location are: Name, Neighbourhood, Postal Code, Geographical Coordinates, Average Housing Prices, Category of Venue, Type of F&B, Pricing of F&B, Opening Hours. With the specifications of Client V in mind, we will use unsupervised machine learning techniques like K-means Clustering to narrow down our search results to a handful of neighbourhoods where we recommend Client V to locate their outlet.

Using the 'Folium' visualisation library, we will create a map of Toronto, using the latitude and longitude data extracted using Google Maps Geocoder API and plot the different Venues on the map.

We will use one hot encoding to analyse the Venues by Venue Category and do a venue count for the most common types of venue per neighbourhood. We will use this information to cluster venues with competing businesses and distinguish these from clusters of restaurants / cafes with symbiotic businesses. We will also cluster neighbourhoods into three housing price categories: Above Average Housing Prices, Average Housing Prices and Below Average Housing Prices.

After narrowing down our search to a few specific neighbourhoods, we will add other features to solve Part 2 - Promotion. These features are: Tips and Reviews, Agree and Disagree Counts, User ID and User Name of Reviewers, Venue ID and Contacts of Other Venues like fitness centres and yoga studios. Again we will use the function "explore?" within a 500 meter radius together to identify Other Venues in the vicinity of our Location. We will use the function "tips?" to filter out KOLs with large followings who are active in that vicinity. We will approach these Other Venues and KOLs to discuss opportunities for cooperation and joint promotions.

Using data analysis and visualisation techniques, we aim to solve the Location and Promotion parts of the Problem for Client V, and hope to launch her business with great success.