Coursera Capstone

IBM Applied Data Science Capstone

RECOMMEND A LOCATION FOR SETTING UP A USED BOOKSTORE IN TORONTO

Business Problem

Pearl Lee recently graduated from a university in Toronto, ON, Canada. During her student time, she realized that a lot of students cannot afford to buy new textbooks and the queue to borrow such textbooks from university library was always very long. Also, there are only few used bookstores in Toronto and most of them are located in remote area.

As a student, she constantly thought of opening a used bookstore where students can buy textbooks at an affordable price. As soon as her graduation, she started to realise her idea: to open a used bookstore in Toronto. She teamed up with a data analyst team to find out where is the optimal place in Toronto to locate the bookstore.

The ideal location for the used bookstore would be:

- Reachable from Toronto's four universities (i.e. University of Toronto, York University, Ryerson University, and OCAD University);
- Located in a middle- or low-income neighbourhood (based on assumption that the high-income neighbourhood would be less interested in used books); and
- Located in a high density population neighbourhood.

Data Description

- List of Toronto boroughs and neighbourhoods which can be found at https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M to explore, segment, and cluster.
- Toronto's sociodemographic data which can be found at https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods.
- Information on venues in Toronto extracted from Foursquare.com

The Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M contains a list of neighborhoods in Toronto. We will use the read_html from pandas package to import the table. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods. The Geocoder package will also provide us the geographic information of the four universities in Toronto. After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we will explore the city of Toronto. We use the read_html from pandas packages to extract the list of neighbourhood data as well as the demographic information of each neighbourhood. We also use read_csv from pandas packages to extract the latitude and longitude of each neighbourhood. The data frames will then be merged to create a unified and complete table contains both geographic and demographic information of Toronto. We use matplotlib and folium package to present and visualize some basic information of the city.

Secondly, we will use Foursquare API to get the top 100 venues that are within a radius of 2500 meters. Using our Foursquare ID and secret key, we make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyses each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering.

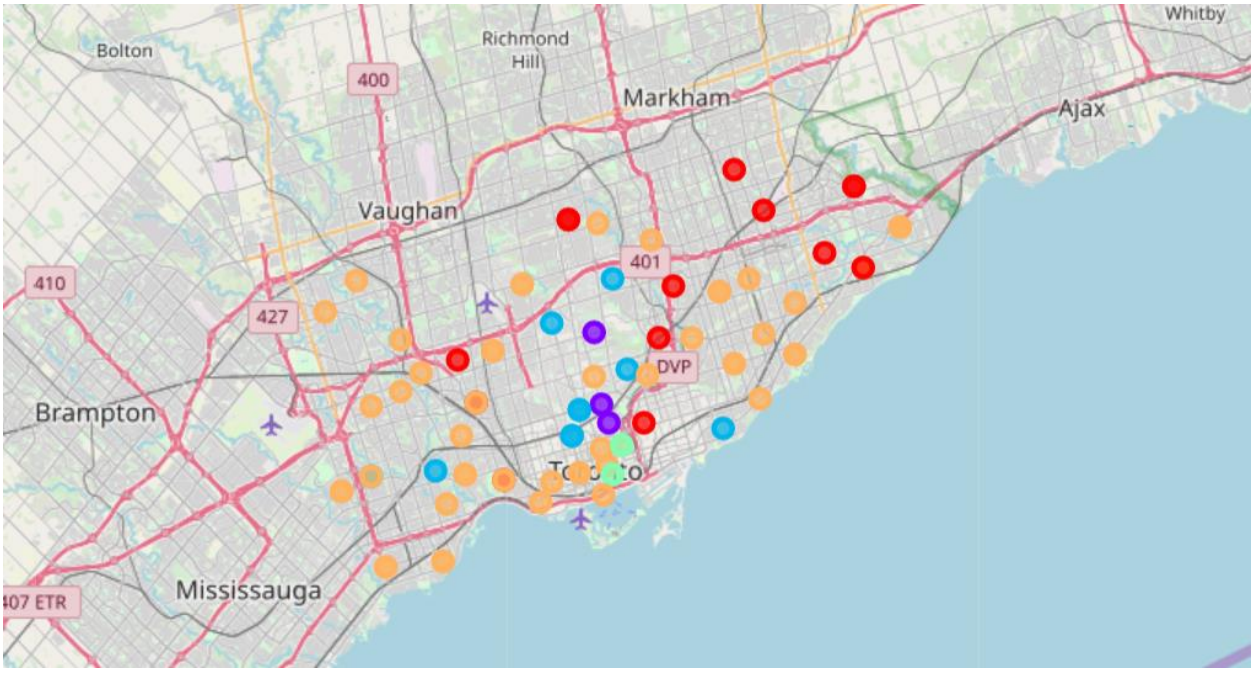
Lastly, we will use K-mean clustering to segment and cluster Toronto neighbourhoods to understand their similarity. With that understanding, we will be able to recommend a suitable place for Pearl Lee to set up her used bookstore. As noted in the Business Problem section, such location would be near universities, in a populous but 'not too rich' neighbourhoods.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 5 clusters based on:

- Population count;
- Population density;
- Average income; and
- Weighted average distance to universities.

The results of the k-means clustering are visualized in the map below:



Per our observation, cluster 5 appears to be the most suitable to locate our bookstore since it satisfies all our criteria. Other cluster would fail at least one criterion such as:

- Cluster 1 and 4 contain of neighborhoods relatively far from universities;
- Cluster 2 and 3 contain of high and extremely high-income neighborhoods.

The details of cluster 5 is as follow:

Neighbourhood	Population	Density	Income	2nd language	Postcode	Borough	Latitude	Longitude	Distance to University of Toronto	Distance to York University	Distance to Ryerson University	Distance to OCAD University	Weighted average distance to universities
Church and Wellesley	13397	24358	37653	Spanish (1.8%)	M4Y	Downtown Toronto	43.665860	-79.383160	1.082338	15.587613	0.980255	1.525760	6.166043
Kensington Market	3740	10389	23335	Cantonese (14.3%)	M5T	Downtown Toronto	43.653206	-79.400049	1.132020	15.972149	1.779208	0.714719	6.402278
Grange Park	9007	10793	35277	Unspecified Chinese (14.8%)	M5T	Downtown Toronto	43.653206	-79.400049	1.132020	15.972149	1.779208	0.714719	6.402278
Garden District	8240	15846	37614	Unspecified Chinese (3.0%)	M5B	Downtown Toronto	43.657162	-79.378937	1.517025	16.561667	0.050568	1.072329	6.540446
Cabbagetown	11120	7943	50398	Unspecified Chinese (1.6%)	M4X	Downtown Toronto	43.667967	-79.367675	2.348950	16.212696	1.461279	2.491697	7.075248

With our criterion in mind, it is apparent that Kensington Market is the ideal choice for Pearl Lee to set up the used bookstore because:

- Kensington Market is relative near the universities with the weighted average distance being 6.4km. Its distance to University of Toronto, Ryerson University, and OCAD University is less than 2km which is very convenient for students from such universities to reach the bookstore.
- Kensington Market's average income is less than the 1st quartile of the city, thus considered a 'low-income' neighborhood.

- Kensington Market's population density is higher than the 3rd quartile of the city, thus qualified our criterion.

It is also noted that the second popular language spoken in Kensington Market is Cantonese, and its adjacent neighborhood - Grange Park has very similar characteristics and is common with Chinese language, thus Pearl Lee may consider to offer books in that Cantonese and Chinese beside used student textbooks.

Discussion

In this project, we only consider three factors population, income, and distance from universities. However, there are other factors such as age distribution of each neighborhood or average rate of spending on education product that could influence the location decision of the bookstore. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a used bookstore. In addition, this project made use of the free Personal account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 5 clusters based on their similarities, and lastly providing recommendations to the businesswoman to choose an optimal location to set up a used bookstore. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 5, specifically, Kensington Market neighborhood is the most preferred location to open a used bookstore. The findings of this project also provide further recommendation to the businesswoman to operate additional offer (i.e. Cantonese and Chinese language books beside of used student textbooks), bring the added value to the businesswoman in her decisions to open a bookstore.