IBM Applied Data Science Capstone Project

1. Introduction

New York City is in reality a collection of many neighbourhoods scattered among the city's five boroughs—Manhattan, Brooklyn, the Bronx, Queens, and Staten Island—each exhibiting its own lifestyle. Moving from one city neighbourhood to the next may be like passing from one country to another. New York is the most populous and the most international city in the country.

New York City is a central hub for many industries. It attracts people from all over the globe with job opportunities but it keeps people with its rich diversity, restaurants, nightlife, and historic buildings. With new people always moving into the city and old buildings falling apart, it can be necessary for new developers to come in and construct new buildings for the ever growing workforce that inhabits the city. My client is looking to do just that. They are looking for the prime neighbourhood to build their new apartment for young millennials who work in Manhattan. What is the prime neighbourhood? Ideally, this neighbourhood would be affordable, close to common working neighbourhoods, and have all the necessary venues in close walking distance. To help my client find this neighbourhood, I will be using data to find a space that meets every qualification. Those neighbourhoods are East Harlem, Yorkville and Murray Hill.

[Central Park, Manhattan, New York City, flanked by the apartment buildings of the Upper East Side]

2. Data

In order to solve such a business problem, the following data will be required:

- List of neighbourhoods in Toronto
★ Description: Contains a list of all neighbourhoods in NewYork with their associated postal codes and boroughs
★ Collection method: Web scraping using BeautifulSoup package

- Values of latitude and longitude of the associated neighbourhoods
★ Description: Needed to specify the neighbourhoods' locations in order to further interact with Foursquare API
★ Source: GeoSpace data of NewYork
★ Collection method: Using GeoCoder package in python or CSV file containing the GeoSpace data

- Location data regarding venues that are present within those neighbourhoods
★ Description: Consists of all venues in the neighbourhoods of NewYork
★ Source: Foursquare database
★ Collection method: Using the latitude and longitude values mentioned above, we can communicate and request information from Foursquare using the Foursquare API
★ Folium is a library that will allow me to build maps to better visualize these neighbourhoods. Lastly, I used an article from StreetEasy that listed the top 6 most affordable neighbourhoods in Manhattan [2]. StreetEasy is an online real estate marketplace for New York City.

3. Methodology

Data collection

First and foremost in order to start the project, the data mentioned above have to be collected accordingly. By using the BeautifulSoup package, it allows us to web scrape the necessary table that contains the data of Toronto's neighbourhoods from Wikipedia in the form of HTML data. Combining it with Pandas function to read HTML data, we can then turn it into a dataframe to be further prepared in Python.

To obtain the values of latitude and longitude of all the neighbourhoods in Toronto (which is needed to interact with the Foursquare API), the GeoSpace data of NewYork is required.

Last but not least, location data regarding venues present within the Toronto neighbourhoods are to be collected using the Foursquare API. By defining our credentials (client ID and client secret), we can interact with the Foursquare API to gain access to their location data – in which we can select the location data of Toronto's neighbourhoods by including the previously mentioned latitudes and longitudes as parameters of the call request to the API.

Data wrangling/preparation

The dataframe of Toronto's neighbourhoods had missing values, whereby some postal codes had boroughs and neighbourhoods that were not assigned to them. As a result, rows with missing borough values are dropped, and neighbourhoods with missing values are replaced with the same values as the borough in their respective rows. The latitude and longitude values of all the neighbourhoods were also initially imported as a separate dataframe, so both dataframes were also merged with postal codes being the column to join based on. Afterwards, the location data obtained from using the

Foursquare API were also merged by grouping them via the neighbourhoods and by taking the mean of the frequency of occurrence of each venue category. To prepare for further data analysis, the data set was also further filtered to only contain the neighbourhoods and their associated mean frequencies of Chinese restaurants.

Method of analysis

It is important to first explain why clustering is an appropriate machine learning model to be used in formulating a solution for the aforementioned business problem. Clustering is an unsupervised machine learning model that groups a set of data points in such a way that data points within the same groups (also called as clusters) are more similar to each other compared to objects present in other clusters. This way, with the context of the business problem at hand, clustering can be used to group NeyYork neighbourhoods which have similar attributes.

4. Results

1. Starting with **Yorkville**, which is located in the Upper East Side of Manhattan, I found that the closest venues are a bagel shop, gym, wine shop, liquor shop, and diner. StreetEasy stated the median rent was $2,450 and the Q train is the only real form of public transit in the area. Its most common venues are Italian restaurants, coffee shops, gyms, and bodegas.

2. The next neighbourhood was **Murray Hill**, located on the east side of midtown. Murray Hill's closest venues are a tea room, a Japanese restaurant, a coffee shop, and a Hawaiian restaurant. The median rent for Murray Hill applies to all of Midtown East, which is $2,700 and it is within walking distance to Grand Central Station. The most common venues are coffee shops, sandwich places, Japanese restaurants, and hotels.

3. I then looked at **Hamilton Heights** which is located near the top of the island, near West 145th street. The closest venues are two cocktail bars, a cafe, an Italian restaurant, and a yoga studio. Median rent is listed as $2,278 and the A, D, and 1 trains run through the area. The most common venues are pizza places, cafes, bodegas, and coffee shops.

4. **East Harlem** is next on the list and is situated near the top right of Central Park. The closest venues are a beer bar, a Mexican restaurant, a Latin American restaurant, and a bakery. The median listing price is $2,100 and the area has access to the 4,5,6 and E train. The most common venues are Mexican restaurants, bakeries, Thai restaurants, and Latin American restaurants.

5. Next, I looked at **Inwood** which is located at the very top of the island near the Bronx. The closest venues are a farmers market, a wine shop, a bakery, and a bodega. The median rent is $1,850 and the trains available are the A and the 1. The most common venues are cafes, Mexican restaurants, restaurants, and lounges.

6. Finally, I looked at **Washington Heights** which is located near the Trans Manhattan Expressway. The closest venues are a restaurant, a cafe, a bodega, and another cafe. The median rent is $2,021 and the C, A, and 1 train are all accessible. The most common venues are cafes, bakeries, mobile phone shops, and bodegas.

5. Discussion

One of the many charms of New York City is how close in proximity everything is. When trying to differentiate these neighbourhoods, it is no surprise that they all have amazing venues near them that make them attractive in their own way.

One of the more obvious ways to divide them was their location within Manhattan. Especially because it is important to the client that these neighbourhoods are close to the Financial District and Midtown, some of these neighbourhoods are too out of the way for what would be an ideal fit.

To make matters easier, we are going to take Inwood, Washington Heights, and Hamilton Heights out of the running because of their location.

Now we can focus on Yorkville, Murray Hill, and East Harlem. East Harlem has the lowest median rent of the three options and very ideal public transit, buying building space would be most affordable in this neighbourhood, but it is still the farthest away from the Financial District and Midtown. Yorkville is closer than East Harlem but the most expensive of the three options. The Upper East Side is home to many old buildings and has a very residential tone. Younger millennials may want a closer proximity to nightlife but older millennials may appreciate the quiet.

Lastly, Murray Hill technically is a part of Midtown so it is the closest option to other working neighbourhoods and it has great accessibility due to its proximity to Grand Central Station. There are also many bars in the area so working professionals who like to unwind after work may find the location appealing.