Neighbourhood Recommendation for Indian Restaurant

Suraj Ravindra Gurav

January 12th, 2020

Section 1: Introduction / Business Problem

The final part of IBM Data Science Professional Certificate is Applied Data Science Capstone Project. The purpose is to work on the real dataset to implement the acquired data import, pre-processing and analysis skills. Objectives of this assignment are to define a business problem, look for data in the web, use Foursquare API to fetch venues in the neighbourhood, analyse the data and its visualization.

1. Description of the Problem and Discussion of the Background

Problem Statement: To recommend a neighbourhood for newly opening an Indian Restaurant in Toronto City

Changing lifestyles, better career opportunities and changing immigration laws are attracting many Indians towards Canada. About 40,000 Indian citizens obtained permanent residency in Canada in 2018, a 50% increase over the 26,600 Indian citizens who were awarded permanent residency in 2017 and this number is increasing continuously.

Toronto, the capital of the province of Ontario, is the most populous Canadian city. Its diversity is reflected in Toronto's ethnic neighbourhoods. As Toronto is shelter for a greater number of Indians than any other city in Canada, it is a good idea to start the restaurant here.

In this project a step-by-step process will be followed to come up with recommendation about the neighbourhood, which will be best to start a new Indian Restaurant. The success of restaurant depends on the people (consumers) and the quality of services and food offered. To make such recommendation, the Toronto neighbourhood data will be imported, processed, visualized and analysed.

2. Target Audience

These are the people who will be benefitted through this analysis. Also, the similar kind of analysis can be performed in the city of their choice, to make decision about the location of starting new restaurant.

2.1. Investors who want to invest in the restaurant business and looking for more profit, low risk location for the same.
2.2. Indian restaurant food chains, who wish to expand their business in Toronto.
2.3. Business analysts, data analyst and data scientists, who want to explore the neighbourhood data for Toronto or any other city of their choice, this project can be used as reference.
2.4. People who would like to have an Indian restaurant or wish to provide Indian food services, this project can be helpful for them to decide about their work area.

Section 2: Data acquisition and cleaning

To find solution to such business problem, a huge amount of data is required. Data once processed and analysed in proper way can explain different business scenarios.

The data requirement for this project are:

- Different boroughs and neighbourhood in these boroughs
- Geographical coordinates of these neighbourhoods
- List of venues in these neighbourhoods
- Details of those venues

This analysis can be performed on any city in the world and such recommendation system can be built for starting different types of business.

As this is demonstrative work, I wanted to choose some city from Germany. But, due to lack of availability of reliable data, I selected Toronto. The Toronto is the most populated city in Canada and its data is available on Wikipedia and other sources. Hence, the Toronto is chosen for this project.

1. Data Sources

The following data sources were used for this project.

1.1 Boroughs and Neighbourhoods

The Wikipedia page "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M" is used to get the information of the boroughs and neighbourhoods along with their postal codes in the city of Toronto

1.2 Geographical Coordinates

The geographical coordinates for the neighbourhoods can be accessed using Geocoder Python package documentation for which is available at "https://geocoder.readthedocs.io/". But, it is a bit unreliable, as one needs to keep it calling multiple times as it does not yield results always. Hence, I used the Geographical coordinates data for different neighbourhoods, which was available as csv file at "http://cocl.us/Geospatial_data".

1.3 Venues and their details in all neighbourhoods

List of all the Venues and their details were fetched by using FourSquare API developers account. The detailed documentation is available at "https://developer.foursquare.com/docs" Foursquare is a social location service that allows users to explore the world around them. Firstly, all the venues, its venue id and its category were fetched using FourSquare API. Later, the details about the venues such as likes, tips and ratings were obtained with the same API.

2. Data Collection

Different techniques such as web scraping, reading direct csv file and fetching data from FourSquare API are used to collect the data.

2.1 Boroughs and Neighbourhoods

Python library BeautifulSoup was used to scrape the webpage "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M" to extract the list of boroughs and the neighbourhoods.

```
In [2]: website_url = requests.get('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M').text
        soup = BeautifulSoup(website_url,'html.parser')

        my_table = soup.find('tbody')
```

The scraped data was collected in a table which was later converted to Pandas DataFrame as follows.

```
In [4]: df = pd.DataFrame(table_data)
        df.head()
```

Out[4]:

	Borough	Neighbourhood	Postalcode
0	Not assigned	Not assigned\n	M1A
1	Not assigned	Not assigned\n	M2A
2	North York	Parkwoods\n	M3A
3	North York	Victoria Village\n	M4A
4	Downtown Toronto	Harbourfront\n	M5A

2.2 Geographical Coordinates

As Geocoder package was not working properly the location data was obtained by reading csv file available at "http://cocl.us/Geospatial_data", which resulted in the following DataFrame.

```
In [15]: df_locations = pd.read_csv('https://cocl.us/Geospatial_data')
         df_locations.head()
```

Out[15]:

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

2.3 Venue and their details in neighbourhoods

Data regarding the 100 venues within 1000 meters radius from each neighbourhoods were obtained by using foursquare API and recorded in pandas DataFrame like this.

	Borough	Neighborhood	ID	Name
0	Downtown Toronto	Harbourfront	4af9a379f964a520c91222e3	Bombay Palace
1	Downtown Toronto	Harbourfront	52af6dc5498e33995b0bbf03	Sultan Of Samosas
2	Downtown Toronto	Queen's Park	4bedf8b5e24d20a17b567214	Kothur Indian Cuisine
3	Downtown Toronto	St. James Town	4af9a379f964a520c91222e3	Bombay Palace

3. Data Cleaning

Data cleaning deals with the processing of missing data, different data types such as categorical, numeric, addition of more columns and sometimes removal of columns, cells and some rows.

3.1 Boroughs and Neighborhoods

In DataFrame for Boroughs and Neghbourhoods, there are some rows where Boroughs were 'Not assigned' or some rows where Neighbourhoods were 'Not assigned'. In order to analyse such data frame, some assumptions should be made for this 'Not assigned' cell.

Here, as part of Data Cleaning, all the rows where Boroughs 'Not assigned' were removed. Also, the cells where Neighbourhood was 'Not assigned' were substituted with the corresponding Borough names. Also, there were more than one neighbourhood in some Boroughs, those were grouped by Postal Code to get such DataFrame.

```
df_toronto = df_filtered.groupby(['Postalcode','Borough'])['Neighbourhood'].apply(','.join).reset_index()
df_toronto
```

	Postalcode	Borough	Neighbourhood
0	M1B	Scarborough	Rouge,Malvern
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union
2	M1E	Scarborough	Guildwood,Morningside,West Hill
3	M1G	Scarborough	Woburn

3.2 Geographical Coordinates

The geographical coordinates were obtained by reading a csv file and stored in separate dataframe. This dataframe was then merged with Boroughs and Neighbourhoods dataframe to have the detailed information about borough, name, location, and postal code of every neighbourhood like this.

	Postalcode	Borough	Neighbourhood	Postal Code	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	M1B	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	M1C	43.784535	-79.160497

As the location coordinates of two neighbourhoods within the same postal code are quite close to each other, the coordinates mentioned here are considered for that postal code.

3.3 Venue Details

Only the venues which were categorized as Indian Restaurants, were fetched from all the neighbourhoods. As FourSqure API, couldn't find details for all the venues requested, it might result into some empty cells. The value of 0 is assigned to such cells such as 0 Likes, 0 Tips and 0 Ratings. The data frame for venue details look like this.

	Borough	Neighborhood	ID	Name	Likes	Rating	Tips
0	Downtown Toronto	Harbourfront	4af9a379f964a520c91222e3	Bombay Palace	14	7.6	13
1	Downtown Toronto	Harbourfront	52af6dc5498e33995b0bbf03	Sultan Of Samosas	9	6.6	4
2	Downtown Toronto	Queen's Park	4bedf8b5e24d20a17b567214	Kothur Indian Cuisine	16	7.9	18

3.4 How to use this data to solve the problem

This analysis start with Exploratory Data Analysis. It included the visualisation such number of neighbourhoods in each borough, grouping them as per the postal codes and how many such groups can be generated.

Using the latitude and longitude coordinates of the postal code, 100 venues in the 1000-meter radius from those coordinates were fetched using FourSqaure API. Also, the venue details were obtained and organized in the tabular dataframe.

In order to come up with some recommendation, the ranking of different neighbourhoods as per the average rating of Indian Restaurants in that neighbourhood is made. All the neighbourhoods, with Average rating for Indian Restaurants as 7.0 or more, were considered as good locations and accordingly their ranking is generated.

From all the analysis, the recommendations can be made as follows:

- The neighbourhood where average rating of Indian Restaurants is poor, can be good place to start the new restaurant to provide the best quality of food and service to be in Profit.
- The neighbourhood with very few Indian restaurants can be good location to start new restaurant of same category, as the competition will be comparatively low.
- The Neighbourhood with high ratings for Indian restaurants can also be good choice for risk taking investors as there can be high competition to provide good quality services at reasonable cost and to be in profit.