Applied Data Science Capstone

ANALYZING NEIGHBORHOODS OF TORONTO TO STARTING A NEW RESTAURANT

Capstone Project (Week 5)

INTRODUCTION: BUSINESS PROBLEM

Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America. Toronto's growing population coincides with increased development and investment in the city and surrounding region. This is why the Toronto area an excellent place to set up a new restaurant.

Opening a succesful restaurant is not an easy task. One of the things that need to be done is to determine the right location to open a new restaurant. In this project, we will study the neighborhoods and find the most suitable location to starting a new restaurant. Target audience of this project is someone who want to open new restaurant in Toronto and have no idea where to build it.

DATA

In this project, we need the following data :

- Neighborhoods of Toronto
  - Source
    https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
  - Description
    This data was scrapped from Wikipedia page using BeautifulSoup (python library) and will be store in a dataframe. The dataframe will consist of three columns: PostalCode, Borough, and Neighborhood.
  - Example

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village

- Geographical coordinates of the neighborhoods (latitude and longitude)
    - Source
      http://cocl.us/Geospatial_data
    - Description
      In order to utilize the Foursquare location data, we need to get the latitude and the longitude coordinates of each neighborhood. . The dataframe will consist of three columns: PostalCode, Latitude, and Longitude.
    - Example

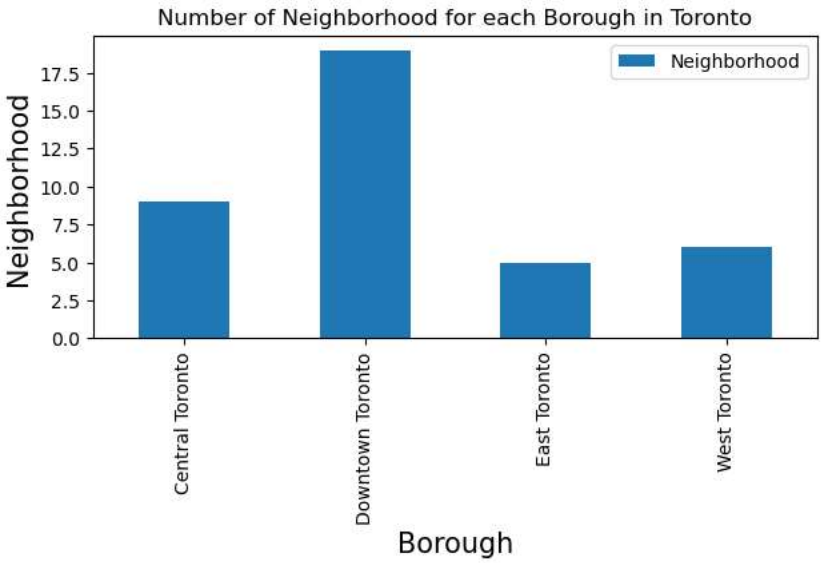
	PostalCode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

- Venue data related to restaurants in Toronto
    - Source : Foursquare
    - Description
      This data will consist of four columns: venue name, categories, latitude, and longitude.
    - Example

	name	categories	lat	lng
0	Roselle Desserts	Bakery	43.653447	-79.362017
1	Tandem Coffee	Coffee Shop	43.653559	-79.361809
2	Cooper Koo Family YMCA	Distribution Center	43.653249	-79.358008
3	Morning Glory Cafe	Breakfast Spot	43.653947	-79.361149
4	Body Blitz Spa East	Spa	43.654735	-79.359874

METHODOLOGY AND ANALYSIS

1. **Data preparation**
    a. First, we need to collect neighborhoods of Toronto data from Wikipedia page by using BeautifulSoup library. Store scrapped data in `df_postal` dataframe.
    b. Get geographical coordinates for each neighborhood, store that data ini `df_geo` dataframe. Merge `df_postal` and `df_geo`, then store it in `df_merge` dataframe. In this project we need to minimize the observation, so narrow down boroughs to Downtown Toronto, East Toronto, West Toronto, and Central Toronto.
    c. Analyze the number of neighborhoods for each borough. From below bar plot, we can see that Downtown Toronto has highest number of neighborhood.



d. Visualize map of Toronto with neighborhoods data

e. Collect venue data from foursquare in radius 500 meters and store it in `toronto_venues` dataframe. There are 1620 restaurants data with 44 uniques restaurant categories that we collect. List of restaurant category:

```
['Restaurant', 'French Restaurant', 'Italian Restaurant', 'Portuguese
Restaurant', 'Mexican Restaurant', 'Sushi Restaurant', 'Japanese Rest
aurant', 'Fast Food Restaurant', 'Ramen Restaurant', 'Thai Restaurant
', 'Modern European Restaurant', 'Seafood Restaurant', 'Middle Easter
n Restaurant', 'Ethiopian Restaurant', 'Chinese Restaurant', 'Vietnam
ese Restaurant', 'American Restaurant', 'New American Restaurant', 'V
egetarian / Vegan Restaurant', 'German Restaurant', 'Comfort Food Res
taurant', 'Asian Restaurant', 'Belgian Restaurant', 'Moroccan Restaur
ant', 'Greek Restaurant', 'Eastern European Restaurant', 'Indian Rest
aurant', 'Falafel Restaurant', 'Korean Restaurant', 'Colombian Restau
rant', 'Mediterranean Restaurant', 'Brazilian Restaurant', 'Gluten-fr
ee Restaurant', 'Latin American Restaurant', 'Cuban Restaurant', 'Tib
etan Restaurant', 'Caribbean Restaurant', 'Cajun / Creole Restaurant
', 'South American Restaurant', 'Filipino Restaurant', 'Doner Restaur
ant', 'Molecular Gastronomy Restaurant', 'Taiwanese Restaurant', 'The
me Restaurant']
```

2. Feature extraction

   In this project, we use One Hot Encoding to extract feature. The output of this method is the frequency of occurrence of restaurant category in each neighborhood. In the table below, we can see the result of One Hot Encoding.

	Neighborhood	Restaurant	French Restaurant	Italian Restaurant	Portuguese Restaurant	Mexican Restaurant	Sushi Restaurant
0	Berczy Park	0.034483	0.017241	0.000000	0.000000	0.0	0.017241
1	Brockton, Parkdale Village, Exhibition Place	0.045455	0.000000	0.045455	0.000000	0.0	0.000000
2	Business reply mail Processing Centre, South C...	0.066667	0.000000	0.000000	0.000000	0.0	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0.000000	0.000000	0.000000	0.0	0.000000
4	Central Bay Street	0.015152	0.015152	0.060606	0.015152	0.0	0.015152

Below table show us most common restaurant in each neighborhoods

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Berczy Park	Restaurant	Seafood Restaurant	French Restaurant	Thai Restaurant	Vegetarian / Vegan Restaurant	Greek Restaurant
1	Brockton, Parkdale Village, Exhibition Place	Restaurant	Italian Restaurant	Seafood Restaurant	German Restaurant	Vegetarian / Vegan Restaurant	New American Restaurant
2	Business reply mail Processing Centre, South C...	Restaurant	Fast Food Restaurant	Seafood Restaurant	German Restaurant	Vegetarian / Vegan Restaurant	New American Restaurant
3	CN Tower, King and Spadina, Railway Lands, Har...	Theme Restaurant	Taiwanese Restaurant	German Restaurant	Vegetarian / Vegan Restaurant	New American Restaurant	American Restaurant
4	Central Bay Street	Italian Restaurant	Thai Restaurant	Restaurant	Ramen Restaurant	Indian Restaurant	Falafel Restaurant

3. Clustering neighborhoods
   K-means algorithm used in this project due to its simplicity and its similarity approach to find patterns. Clustered the neighborhood in Toronto into 3 clusters based on their frequency of occurrence for “Restaurant”.

RESULTS

The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many restaurants are in each neighborhood. The meaning of markers in the map:

- Red circle (cluster 0) : neighborhoods with less number of restaurants.
- Purple circle (cluster 1) : neighborhoods with no restaurants.
- Green circle (cluster 2) : neighborhoods with more number of restaurants.

Most restaurants are in cluster 2 which is around Forest Hill North & West, Forest Hill Road Park. Cluster 0 show good location to starting a new restaurant in Toronto, because there are less restaurant in those area.

DISCUSSION

As mentioned earlier the most suitable neighbourhoods for starting new restaurant are present in the cluster number 4. The restaurant owner can go ahead and make a decision depending on other factors like availability and legal requirements that are out of scope of this project.

CONCLUSION

Data science can be very helpful in determining solutions for certain business problems. In this project we studied the neighbourhoods of Toronto and came up with a recommendation of neighbourhoods where our client can start their new restaurant.