Capstone Project: The battle of the neighborhoods - presentation

by Subhayan Sur

Date: 23-Sep-2019

Agenda

- Introduction
- Data Acquisition and Cleaning
- Methodology
- Modeling
- Results
- Discussion
- Conclusion

Introduction

- Background: Getting insights about localities is very important when someone intends to open a new restaurant. This helps make a strong business case with the backing of credible data.
- Problem: This project aims to select the best neighborhood in Toronto based on places where people prefer to frequent a lot and have lots of similar restaurants.
- Interest: Anyone who is interested to figure out what would be a good place to open a restaurant business in Toronto and explore its neighborhoods and common venues around each neighborhood.

Data Acquisition and Cleaning

- **Data Acquisition:**

1. The Borough and neighborhood information from Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

The columns were as follows:

i) Postcode: The postal codes for various neighborhoods of Toronto.

ii) Borough : The names of the boroughs, ‘not Assigned’ – if they’re not assigned to any borough.

iii) Neighborhood: The names of the neighborhoods, ‘not Assigned’ – if they do not have any assigned neighborhoods.

2. We used the geospatial data provided - http://cocl.us/Geospatial_data to perform the task of updating the coordinates and create clusters

3. The number of restaurants and their type and location in every neighborhood will be obtained using Foursquare API

4. We used Folium package in Python to create the map of Toronto using latitude and longitude values.

Data Acquisition and Cleaning

- **Data Cleaning:**

We performed the following data cleaning activities –

1. Updated the column names in the data frame with proper connotation.

2. Deleted the new lines from the data frame.

3. Deleted the records which doesn’t have a borough assigned.

4. Removed spaces at the beginning of the strings.

5. Combined the neighborhoods with same Postal Code and Boroughs using a comma separator.

Data Acquisition and Cleaning

- The dataset after performing cleaning and merging:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Methodology

- In this project we directed our efforts on detecting neighborhoods of Toronto that have high restaurant density, particularly those with which have restaurants as their most frequented venues.
- In Part 1, We gathered Toronto's neighborhood information from Wikipedia sources and created a clean data frame.
- Data set after Part 1, looks like below:

	PostalCode	Borough	Neighborhood
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
4	M1H	Scarborough	Cedarbrae
5	M1J	Scarborough	Scarborough Village
6	M1K	Scarborough	East Birchmount Park,Ionview,Kennedy Park
7	M1L	Scarborough	Clairlea,Golden Mile,Oakridge
8	M1M	Scarborough	Cliffcrest,Cliffside,Scarborough Village West

Methodology

- In Part 2, We prepared the data frames with Neighborhood coordinates, postal code, borough and neighborhood information that was used for clustering. We also used the geospatial data from the link.
- Data set after Part 2, looks like below:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

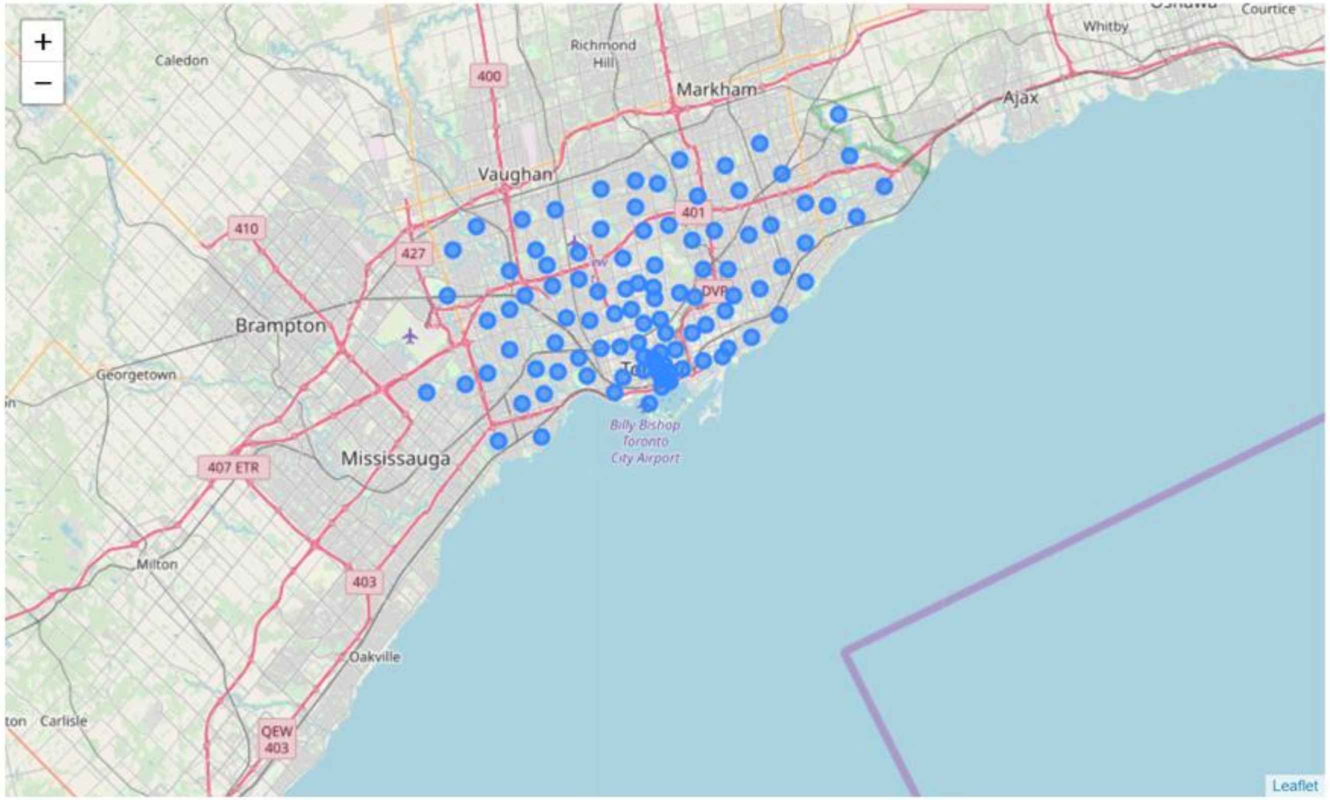
Methodology

- In the final Part 3, we focused on most promising areas by creating clusters of locations that meet our requirements established in discussion with stakeholders: we took into consideration locations with multiple restaurants in the vicinity, and we want locations that have various restaurants as their most common venue.
- Dataset after clustering in Part 3, looks like below:

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Scarborough	2.0	Fast Food Restaurant	Women's Store	Event Space	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Donut Shop	Dog Run
13	Scarborough	2.0	Pizza Place	Chinese Restaurant	Italian Restaurant	Noodle House	Bank	Breakfast Spot	Fried Chicken Joint	Fast Food Restaurant	Thai Restaurant	Dessert Shop
15	Scarborough	2.0	Chinese Restaurant	Coffee Shop	Fast Food Restaurant	Pizza Place	Sandwich Place	Bubble Tea Shop	Thrift / Vintage Store	Grocery Store	Pharmacy	Breakfast Spot
17	North York	2.0	Dog Run	Golf Course	Fast Food Restaurant	Mediterranean Restaurant	Pool	Discount Store	Department Store	Dessert Shop	Dim Sum Restaurant	Diner
24	North York	2.0	Coffee Shop	Butcher	Home Service	Discount Store	Pharmacy	Pizza Place	Grocery Store	Airport Terminal	Falafel Restaurant	Ethiopian Restaurant
34	North York	2.0	Coffee Shop	Portuguese Restaurant	Pizza Place	Hockey Arena	Intersection	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant	Diner
35	East York	2.0	Pizza Place	Fast Food Restaurant	Bank	Gym / Fitness Center	Pet Store	Pharmacy	Gastropub	Intersection	Athletics & Sports	Dessert Shop

Methodology

- Visualizing the clustered neighborhoods:



Results

- **Summary of the clusters**

  During the 5 rounds of clustering we came across data segments that describe the most frequented places by the people based on most common venues. In each clusters the segmentations show a different grouping. However, from Cluster 2 and Cluster 3, it became apparent that Scarborough has the greatest number of restaurants as the 'most common venue' for people and visitors.

Cluster 2 - Segmentation based on Miscellaneous stores, restaurants, cafes etc

```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 1, toronto_merged.columns[[1] +
```

]:

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
2	Scarborough	1.0	Electronics Store	Mexican Restaurant	Breakfast Spot	Intersection	Medical Center	Rental Car Location
3	Scarborough	1.0	Coffee Shop	Convenience Store	Korean Restaurant	Women's Store	Department Store	Dessert Shop
4	Scarborough	1.0	Fried Chicken Joint	Bakery	Hakka Restaurant	Bank	Athletics & Sports	Thai Restaurant
5	Scarborough	1.0	Women's Store	Playground	Curling Ice	Electronics Store	Eastern European Restaurant	Dumpling Restaurant

Results

- Summary of the clusters

Cluster 3 - Segmentation based on Restaurants, Pizza Places

```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 2, toronto_merged.columns[[1]
```

4]:

	Borough	Cluster Labels	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	Scarborough	2.0	Fast Food Restaurant	Women's Store	Event Space	Empanada Restaurant	Electronics Store	Eastern European Restaurant
13	Scarborough	2.0	Pizza Place	Chinese Restaurant	Italian Restaurant	Noodle House	Bank	Breakfast Spot
15	Scarborough	2.0	Chinese Restaurant	Coffee Shop	Fast Food Restaurant	Pizza Place	Sandwich Place	Bubble Tea Shop
17	North York	2.0	Dog Run	Golf Course	Fast Food Restaurant	Mediterranean Restaurant	Pool	Discount Store
24	North York	2.0	Coffee Shop	Butcher	Home Service	Discount Store	Pharmacy	Pizza Place
34	North York	2.0	Coffee Shop	Portuguese Restaurant	Pizza Place	Hockey Arena	Intersection	Deli / Bodega

Discussion

- The aim of the project is to help people who want to open a new restaurant in a neighborhood borough of Toronto based on the most places for restaurants.
- If we look at cluster 2 and cluster 3, we can see that Scarborough has shops, bus stations, bus lines, lounges, food joints, pizza places, cafes all in one place.
- Hence it becomes as obvious choice to open a new restaurant as people can choose to take public transport to visit that place, do shopping and finally go to restaurant for a meal before heading back.

Conclusion

- Purpose of this project was to identify Toronto neighborhoods in order to aid the stakeholder in narrowing down the search for optimal location for a new restaurant.
- We first created the boroughs and neighborhoods from various data sources and added their location coordinates using geocode. Then we created a map to visualize Toronto's neighborhoods. Next we utilized the Foursquare API to explore all the neighborhoods and their most frequented venues. Then we analyzed each Neighborhood by taking the mean of the frequency of venue-occurrence for each category. Then we created a data frame with the top 10 - most common venues in each neighborhood. Then we used k-means clustering of those locations in order to create major neighborhoods of interest hat were to be used as starting points for final exploration by stakeholders.
- Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in the recommended neighborhood (Scarborough), taking into consideration additional factors like attractiveness of restaurant location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every location etc.