Coursera Capstone Project – The Battle of the Neighborhoods

Pirtham Sangione

Introduction

Problem Description

As a data scientist with 'Entrepreneurs R US', you have been tasked with creating a Jupyter Notebook to help local entrepreneurs with their business needs. Specifically, providing insight into what type of venue is most frequently visited and where these venues are located within the vast Toronto region. The entrepreneur can then use this information to aid in their decision to invest in a particular industry.

Background

Small business owners are often considered the backbone of an economy. These businesses help grow the economy, provide employment, and reduce crime rates. Tasked with providing an online service to such a diverse, multicultural city like Toronto will provide the entrepreneurs with all the necessary information that will help them in picking the location for their new ventures based on their desire to support a specific ethnic group/community in Toronto.

Solution

Our solution will comprise of 3 crucial steps.

1. Scrape and preprocess public and commercial datasets for demographics such as census data, geographic locations, popular venues, etc.
2. Apply a K-Means machine learning algorithm to a particular borough of interest (can be changed upon the entrepreneur's request) to cluster popular venues in a certain neighborhood.
3. Visualize the results using an interactive map

We will then publish the results into a report and presentation that can be made available to the public.

Data

Dataset 1: List of postal codes of Canada: M

Description:

Wikipedia page containing a table that lists all the postal codes starting with the letter "M". Postal codes beginning with "M" are located within the city of Toronto. Additional information that is listed are the boroughs and corresponding neighborhoods.

Data Use:

Information is scraped using the "Beautiful Soup" library in Python.

Source:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Dataset 2: Census Profile, 2016 Census

Description:

Webpage on Statistics Canada website containg census information from the year 2016. Data was downloaded as a CSV file named "population.CSV".

Data Use:

Information is read into Jupyter Notebook using Pandas "read_csv" function and merged into existing data frame.

Source:

https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=CSD&Code1=3520005&Geo2=PR&Code2=35&Data=Count&SearchText=Toronto&SearchType=Begins&SearchPR=01&B1=All&GeoLevel=PR&GeoCode=3520005&TABID=1

Dataset 3: Geospatial Data

Description:

CSV file containg the geospatial coordinates of the Toronto postal codes.

Data Use:

Information is read into Jupyter Notebook using Pandas "read_csv" function and merged into existing data frame.

Source:

http://cocl.us/Geospatial_data

Dataset 4: Foursquare API

Description:

Location service that leveraged location data for venues in Toronto.

Data Use:

Connected to Foursquare API in Jupyter Notebook and created a function that got nearby venues in the area of interest.

Source:

https://foursquare.com/developers/apps

Methodology

Using our various data sources, we can then begin our analysis and visualization using Python. The following list describes the steps taken in our Jupyter Notebook.

1. Import relevant python libraries (Ex: pandas, NumPy, json, requests, etc.)

2. Scrape Wikipedia page for a table containing postal codes of Toronto using Beautiful Soup library.

3. Build Pandas data frame from Wikipedia table, then cleanse the data by reassigning variables and deleting NaN values.

4. Merge additional data sources containing populations and geospatial coordinates of Toronto to existing data frame.

5. Visualize all the boroughs and corresponding Neighborhoods of Toronto using Folium.

6. Choose a borough that you wish to conduct analysis on. We chose West Toronto.

7. Connect to Foursquare API and search for the most visited venues in West Toronto. Then create another data frame with location parameters such as venue name, latitude, longitude, neighborhoods, category.

8. Perform "one hot encoding" to the categorical data so that it can be manipulated into numerical data for the machine learning portion of the script.

9. Display the top 5 venues for each West Toronto neighborhood along with the frequency. (We chose to reduce the list of venues to top 3 for simplicity)

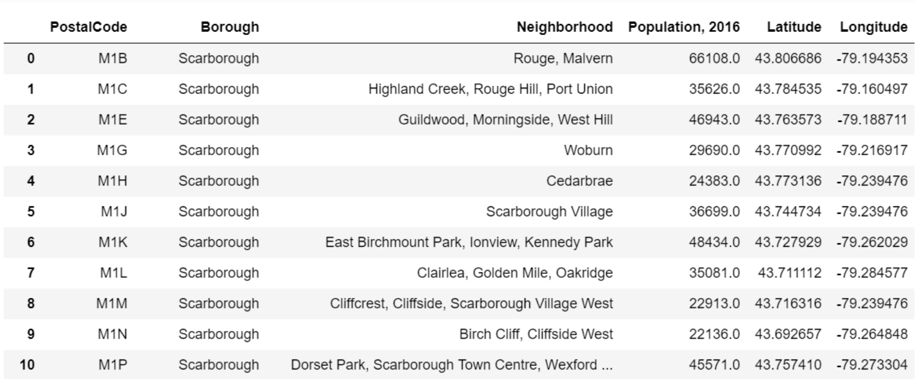
10. Perform K-Means Clustering on the data frame. Then merge new data frame with previous data frame that contains the venue information for neighborhoods in West Toronto.

11. Create an updated map with the various clusters showing which clusters are locating in which neighborhoods.

Results

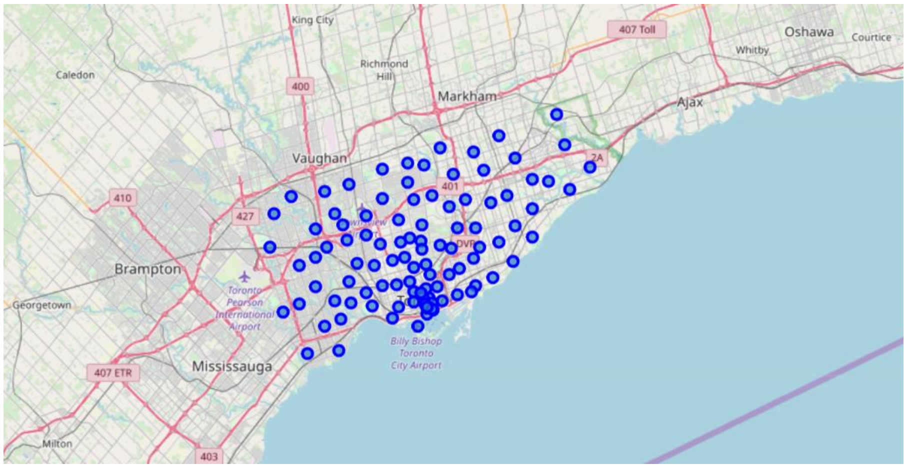
Following the above steps, here are some of the results that came from executing the Python script. The first portion of the results pertain to cleansing and wrangling of the data in order for manipulation in later segments of the Python script.

The data frame shown below was the result of steps 1-4. The column headers describe various features such as location, name and populations for all boroughs of Toronto.

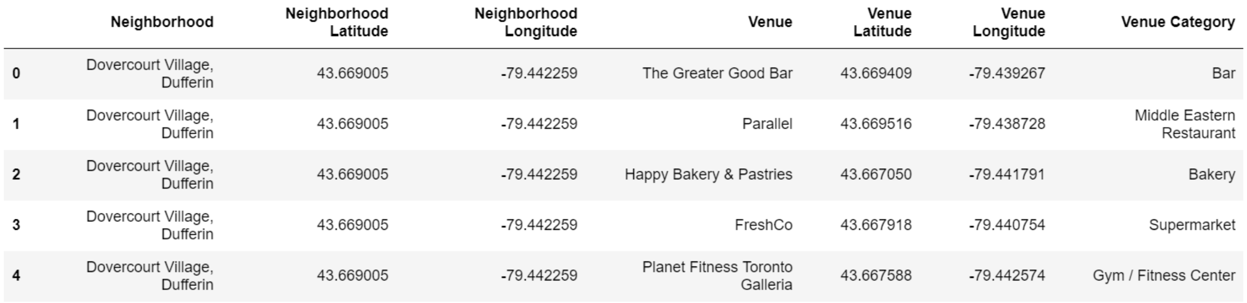


	PostalCode	Borough	Neighborhood	Population, 2016	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	66108.0	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	35626.0	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	46943.0	43.763573	-79.188711
3	M1G	Scarborough	Woburn	29690.0	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	24383.0	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	36699.0	43.744734	-79.239476
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park	48434.0	43.727929	-79.262029
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge	35081.0	43.711112	-79.284577
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West	22913.0	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	22136.0	43.692657	-79.264848
10	M1P	Scarborough	Dorset Park, Scarborough Town Centre, Wexford ...	45571.0	43.757410	-79.273304

In step 5 we used the folium library in Python to visualize Toronto. Each blue marker illustrates each borough and corresponding neighborhood.

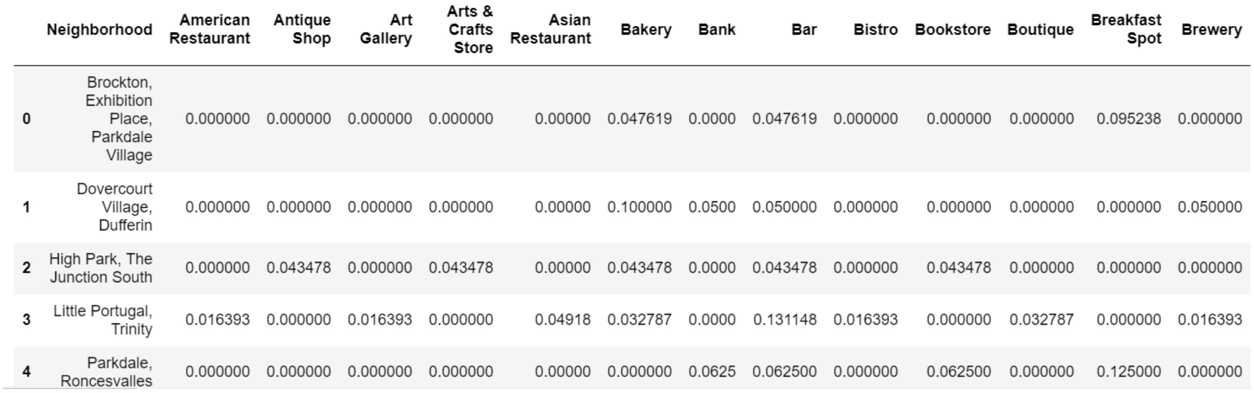


Steps 6 & 7 involved using the Foursquare API developer website and searching for the most popular venues for each neighborhood and its associated location data. The data frame below is an example of one of West Toronto's neighborhoods: The Dovercourt Village, Dufferin.



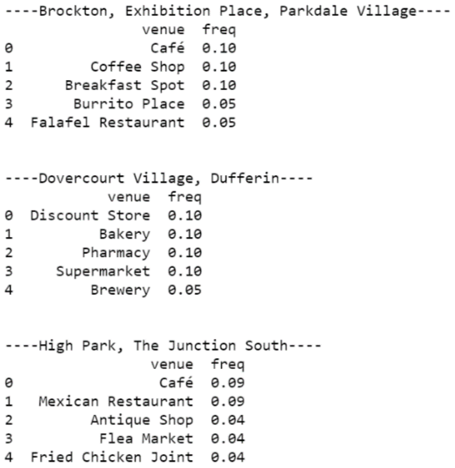
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Dovercourt Village, Dufferin	43.669005	-79.442259	The Greater Good Bar	43.669409	-79.439267	Bar
1	Dovercourt Village, Dufferin	43.669005	-79.442259	Parallel	43.669516	-79.438728	Middle Eastern Restaurant
2	Dovercourt Village, Dufferin	43.669005	-79.442259	Happy Bakery & Pastries	43.667050	-79.441791	Bakery
3	Dovercourt Village, Dufferin	43.669005	-79.442259	FreshCo	43.667918	-79.440754	Supermarket
4	Dovercourt Village, Dufferin	43.669005	-79.442259	Planet Fitness Toronto Galleria	43.667588	-79.442574	Gym / Fitness Center

Step 8 involved performing "one hot encoding" which turns categorical data into numerical data. This allows the machine learning algorithm to read the data.



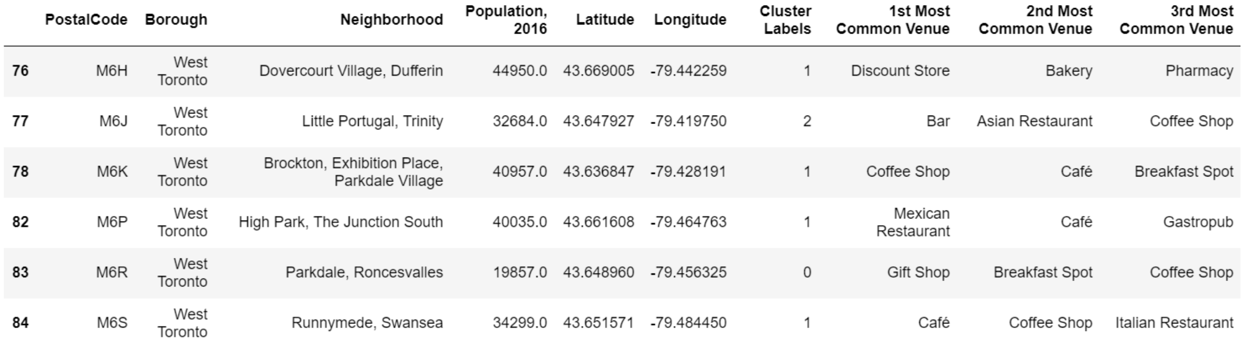
	Neighborhood	American Restaurant	Antique Shop	Art Gallery	Arts & Crafts Store	Asian Restaurant	Bakery	Bank	Bar	Bistro	Bookstore	Boutique	Breakfast Spot	Brewery
0	Brockton, Exhibition Place, Parkdale Village	0.000000	0.000000	0.000000	0.000000	0.00000	0.047619	0.0000	0.047619	0.000000	0.000000	0.000000	0.095238	0.000000
1	Dovercourt Village, Dufferin	0.000000	0.000000	0.000000	0.000000	0.00000	0.100000	0.0500	0.050000	0.000000	0.000000	0.000000	0.000000	0.050000
2	High Park, The Junction South	0.000000	0.043478	0.000000	0.043478	0.00000	0.043478	0.0000	0.043478	0.000000	0.043478	0.000000	0.000000	0.000000
3	Little Portugal, Trinity	0.016393	0.000000	0.016393	0.000000	0.04918	0.032787	0.0000	0.131148	0.016393	0.000000	0.032787	0.000000	0.016393
4	Parkdale, Roncesvalles	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.0625	0.062500	0.000000	0.062500	0.000000	0.125000	0.000000

In step 9, we employed some statistical measures such as frequency to get an idea of how often certain venues are searched in Foursquare. The output shown to the right illustrates the top 5 searched venues for every neighborhood in West Toronto. (Please note there are more neighborhoods shown in the Python script)



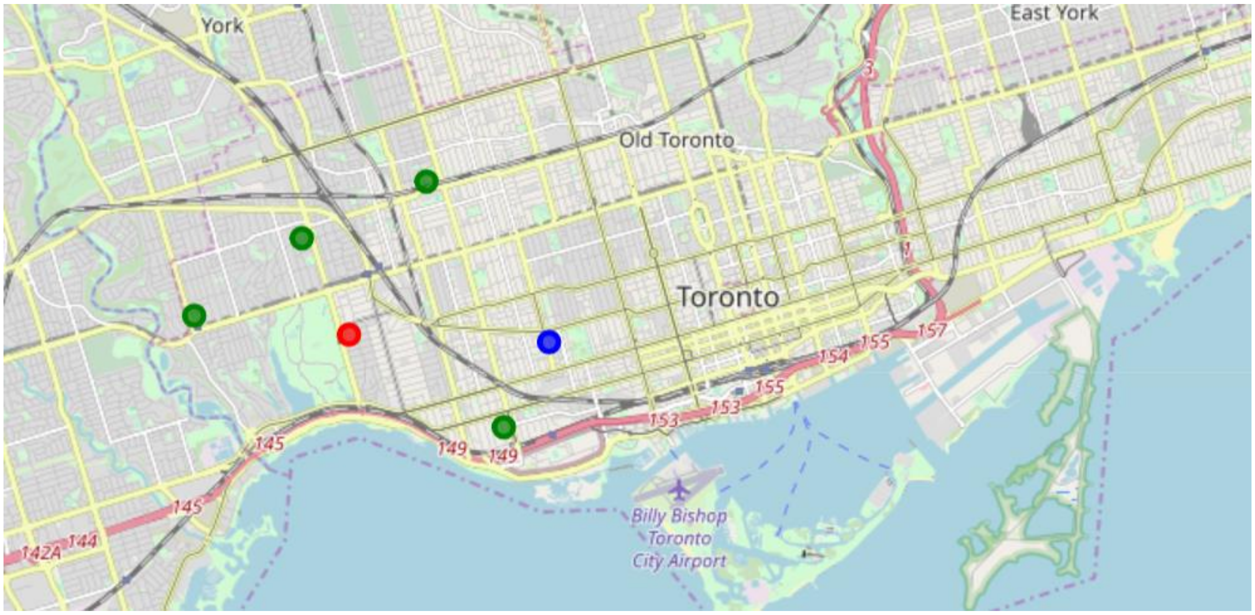
```
----Brockton, Exhibition Place, Parkdale Village----
                venue  freq
0                Café  0.10
1         Coffee Shop  0.10
2      Breakfast Spot  0.10
3       Burrito Place  0.05
4  Falafel Restaurant  0.05


----Dovercourt Village, Dufferin----
            venue  freq
0  Discount Store  0.10
1          Bakery  0.10
2        Pharmacy  0.10
3     Supermarket  0.10
4         Brewery  0.05


----High Park, The Junction South----
                venue  freq
0                Café  0.09
1  Mexican Restaurant  0.09
2        Antique Shop  0.04
3         Flea Market  0.04
4  Fried Chicken Joint  0.04
```

In the data analysis portion of the code, we performed a K-Means cluster algorithm to the data frame above. By doing so, we can take multiple clusters that group together certain venues based on a variety of factors such as location, type of venue, frequency, etc. In the discussion segment of the report, we will talk in length what these clusters represent to the entrepreneurs who utilize these results.



	PostalCode	Borough	Neighborhood	Population, 2016	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
76	M6H	West Toronto	Dovercourt Village, Dufferin	44950.0	43.669005	-79.442259	1	Discount Store	Bakery	Pharmacy
77	M6J	West Toronto	Little Portugal, Trinity	32684.0	43.647927	-79.419750	2	Bar	Asian Restaurant	Coffee Shop
78	M6K	West Toronto	Brockton, Exhibition Place, Parkdale Village	40957.0	43.636847	-79.428191	1	Coffee Shop	Café	Breakfast Spot
82	M6P	West Toronto	High Park, The Junction South	40035.0	43.661608	-79.464763	1	Mexican Restaurant	Café	Gastropub
83	M6R	West Toronto	Parkdale, Roncesvalles	19857.0	43.648960	-79.456325	0	Gift Shop	Breakfast Spot	Coffee Shop
84	M6S	West Toronto	Runnymede, Swansea	34299.0	43.651571	-79.484450	1	Café	Coffee Shop	Italian Restaurant

Finally, we output these clusters onto the original map. Each color represents a different cluster.



Discussion

Based upon the findings in the results section, the user can make a conscious decision about where the most popular types of venues are present in which neighborhoods.
The results from the cluster analysis show 3 main clusters. Cluster 1 has Dovercourt Village, Brockton, High Park and Runnymede, cluster 2 has Little Portugal and cluster 0 has Parkdale. The most common venue for cluster 1 is a café, bar for cluster 2 and gift shop for cluster 0. With this knowledge, an entrepreneur can either decide to open a popular venue such as a café in cluster 1 and compete for market share or choose an unpopular venue and take advantage of the opportunity to dominate that sector.

One observation that I can note is that for clusters 2 and 0, there is only 1 neighborhood as opposed to cluster 1 which showcases four neighborhoods. This will not be the case for every borough. The user/entrepreneur is recommended to experiment with various boroughs to get a broad idea of what type of venues and neighborhoods are clustered together in the city of Toronto.

Conclusion

To conclude, we hope you are satisfied with the information presented to you today. At Entrepreneurs R Us, we strive to provide accurate and insightful analysis tools for your business needs. Using both public and commercial datasets such as Foursquare, we can then manipulate the data to extract meaningful information. Our interactive K-Means Analysis script will allow you to achieve a greater understanding of how popular venues and neighborhoods can be classified together using machine learning.