Capstone Project: Final Report

Helping home buyers choose which neighborhood will best meet their needs.
Robert M. Taylor

1. Introduction

This project is intended for real estate agents who frequently work with out of town clients who may be unfamiliar with the area. Please imagine that you are a real estate agent in Toronto and often have a common problem when helping out of town clients. Many of your clients, who are unfamiliar with the Toronto area, will narrow down their choices to 2 or 3 houses that they liked. In making a final decision, your clients often tell you what types of venues and/or businesses they would like in their ideal neighborhood. The clients then often ask you to help them decide which of the 2 or 3 houses they are looking at are in a neighborhood with their ideal "wants." You have been learning data science, in the hopes of getting an upper hand in your industry and decide that you can use your newly acquired data science skills to help your clients.

You have a new client that has recently landed a lucrative data science job at a university in Toronto. You have shown the client and their partner several houses and they have narrowed their choice down to 3 lovely homes. Each home is in a different neighborhood in Toronto and, as usual, the clients have asked for your help in deciding which neighborhood would best fit their needs. Your clients have picked 3 homes in the neighborhoods of 1) Berczy Park, 2) Queen's Park, and 3) Rosedale. The clients have told you that they MUST have 3 things in the neighborhood they choose that are extremely important to them. They tell you that these 3 things are 1) a park, 2) coffee shops, and 3) a gym. Your project is to now utilize your new skills in python, data analytics, and Foursquare to determine which neighborhood is best for your clients.

2. Data

To solve this problem, I used 3 data sources. First, I used the following Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. I then scraped this page to get the postcode data from the table and then transformed it into a data frame. The resulting data frame consisted of 3 columns for 'Postcode', 'Borough', and 'Neighborhood.' I then cleaned and analyze this data and used Folium to visualize the home locations and neighborhoods of interest. Next, I used the geospatial data for Toronto. This data is accessed with: http://cocl.us/Geospatial_data. This data contains all of the latitude and longitude coordinates for postcodes in the Toronto area. Finally, the Foursquare location data source was employed to determine what venues (specifically those that our clients were interested in) were in the neighborhoods. I was then be able to utilize this data to make a final determination of which neighborhood was the best choice for our clients.

3. Methodology

I began by scraping the Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M to get the postcode data from the table. I then transformed the table into a pandas data frame. The resulting data frame ('wiki') contained 3 columns for 'Postcode', 'Borough', and 'Neighborhood.' I examined the resulting data frame using 'head' and 'shape.' I next dropped any "Not assigned" boroughs and also corrected the spelling of the column "Neighborhood."

I wanted to split up the data frame for ease of cleaning. So, next I made a new data frame called 'borough' that only contained the postcode and borough information and I removed duplicate postcode entries from the borough column. Similarly, I then made a new data frame called 'neighborhood' that only contained the "neighborhood" and the "postcode" features. This allowed me to combine multiple neighborhoods with the same postcode into a single row for the postcode. This was easier to do without the borough data being included. I also changed any "Not assigned" neighborhoods to the name of the borough.

Next I loaded in the Toronto geospatial data (http://cocl.us/Geospatial_data) as a pandas data frame and examined the resulting data frame that I call "geo." I also renamed the column "Postal Code" to "Postcode" to match the other two borough and neighborhood data frames, respectively. I was then able to add the latitude and longitude features from the geo data frame into the neighborhood data frame. Finally, I added back in the borough feature into the neighborhood data frame. This resulted in a "neighborhood" data frame that contained all of the information for postcode, borough, neighborhood, latitude, and longitude.

I next pulled out data for only the three neighborhoods of interest (i.e. Berczy Park, Queen's Park, and Rosedale). I was then able to utilize the geopy library to get the latitude and longitude of Toronto so that we could then create folium maps with the neighborhoods superimposed on top. Following this, I utilized the Foursquare location data to answer the question of which neighborhood best fit the client's needs.

To use Foursquare, I first defined credentials and located the 3 neighborhoods of interest in our data frame to locate the latitude and longitude for each neighborhood. I then began analysis with the first neighborhood (Berczy Park). The general method flow for analysis of each neighborhood with Foursquare was as follows:

1. Get latitude and longitude for neighborhood
2. Get URL request with a limit of 100 and radius of 500.
3. Send GET request
4. Write function to extract venues
5. Clean json and structure into panda data frame
6. Narrow search down to only client venues (park, coffee, gym)
7. Review results
8. Repeat process for other 2 neighborhoods

Next, to get all the Foursquare results from each neighborhood into a single data frame, I created a function to do all 3 of the previous Foursquare calls in a new data frame. I then ran this function to create the new data frame called "Toronto_venues." I eliminated all venues from this data frame EXCEPT the client venues (park, coffee, gym). This new data frame was called "WishList." Folum maps were generated to visualize all of the park, coffee, and gym venues within the 3

neighborhoods. One hot encoding was then used to generate new features from the different 3 venues and then grouped and the mean taken, to determine the frequency of each venue occurrence in each of the 3 neighborhoods. Finally, K-means was used with a k=3 and clustering was used with subsequent visualizations using folium. At the end, I ended up again reviewing the "wantlist_grouped" data frame which showed the frequency of the client venues in each neighborhood.

4. Results and Discussion

The original Wikipedia page containing postal codes in Canada was scraped and transformed into a data frame. The resulting data frame contained 3 columns for 'Postcode', 'Borough', and 'Neighborhood.' Additionally, I dropped (ignored) boroughs that were 'Not assigned'. I also corrected the spelling of Neighborhood in the data frame. The resulting data frames are shown in Figure 1. This initial data frame was subsequently split the into 2 new data frames called 'borough' and 'neighborhood' that contained only the postal code along with either the borough or neighborhood, respectively (Figure 2a). The borough data frame was then cleaned by eliminating duplicate postal codes. The neighborhood data frame was cleaned by combining multiple neighborhoods into their appropriate single postal codes and changing the "Not assigned" neighborhood to the name of the borough (Figure 2b).

Figure 1

A

[3]:

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

B

[5]:

	Postcode	Borough	Neighborhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M5A	Downtown Toronto	Regent Park
6	M6A	North York	Lawrence Heights

Figure 2

A

[6]:

	Postcode	Borough
2	M3A	North York
3	M4A	North York
4	M5A	Downtown Toronto
5	M5A	Downtown Toronto
6	M6A	North York

[11]:

Postcode	Neighborhood
M3A	Parkwoods
M4A	Victoria Village
M5A	Harbourfront,Regent Park
M6A	Lawrence Heights,Lawrence Manor
M7A	Not assigned

B

[15]:

Postcode	Neighborhood
M3A	Parkwoods
M4A	Victoria Village
M5A	Harbourfront,Regent Park
M6A	Lawrence Heights,Lawrence Manor
M7A	Queen's Park

The Geospatial Data for Toronto contained all of the latitude and longitude coordinates for postcodes in the Toronto area. This data was put into a data frame and then merged with the borough and neighborhood data frames (Figure 3). This merged data frame was then condensed into the tornonto_data data frame that only contained the 3 neighborhoods our clients were interested in (Figure 4). Finally, I used the Foursquare location data to determine what venues (specifically those that our clients were interested in) are in the neighborhoods. Figure 5 shows a map of Toronto with the 3 neighborhoods of interest superimposed over it. Foursquare was again used to pull all of the

Figure 3

[23]:

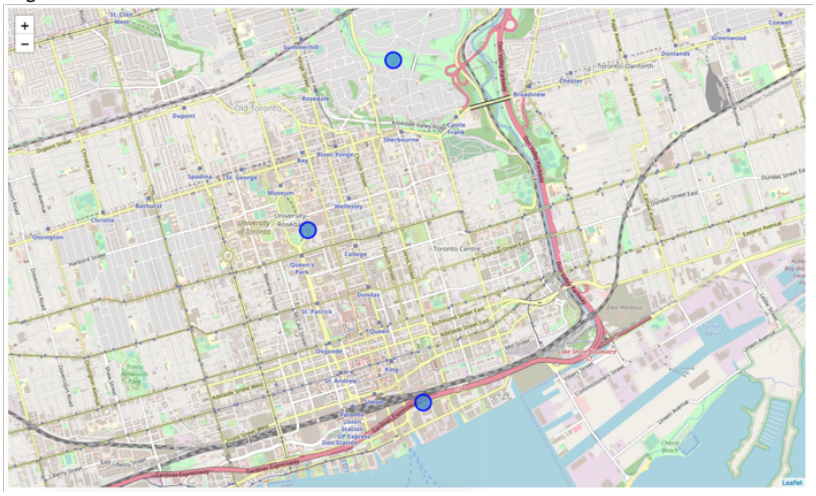
	Postcode	Neighborhood	Latitude	Longitude	Borough
93	M8W	Alderwood,Long Branch	43.602414	-79.543484	Etobicoke
88	M8V	Humber Bay Shores,Mimico South,New Toronto	43.605647	-79.501321	Etobicoke
102	M8Z	Kingsway Park South West,Mimico NW,The Queensw...	43.628841	-79.520999	Etobicoke
87	M5V	CN Tower,Bathurst Quay,Island airport,Harbourf...	43.628947	-79.394420	Downtown Toronto
101	M8Y	Humber Bay,King's Mill Park,Kingsway Park Sout...	43.636258	-79.498509	Etobicoke

Figure 4

[25]:

	Postcode	Neighborhood	Latitude	Longitude	Borough
20	M5E	Berczy Park	43.644771	-79.373306	Downtown Toronto
4	M7A	Queen's Park	43.662301	-79.389494	Queen's Park
91	M4W	Rosedale	43.679563	-79.377529	Downtown Toronto

Figure 5



venue information about the 3 neighborhoods. This was subsequently condensed down to only the 3 venues of interest to our clients (park, coffee shop, and gym). The condensed venue results for each neighborhood are shown in Figure 6. These venue results were then combined into a single data frame and the venues were superimposed over a map of Toronto (Figure 7).

Figure 6

Berczy

	name	categories	lat	lng
14	Berczy Park	Park	43.648048	-79.375172
27	Starbucks	Coffee Shop	43.644424	-79.369294
37	Starbucks	Coffee Shop	43.646957	-79.378265
39	Everyday Gourmet (Teas & Coffees)	Coffee Shop	43.648757	-79.371645
44	Starbucks	Coffee Shop	43.648738	-79.372519
55	Starbucks	Coffee Shop	43.644525	-79.368560

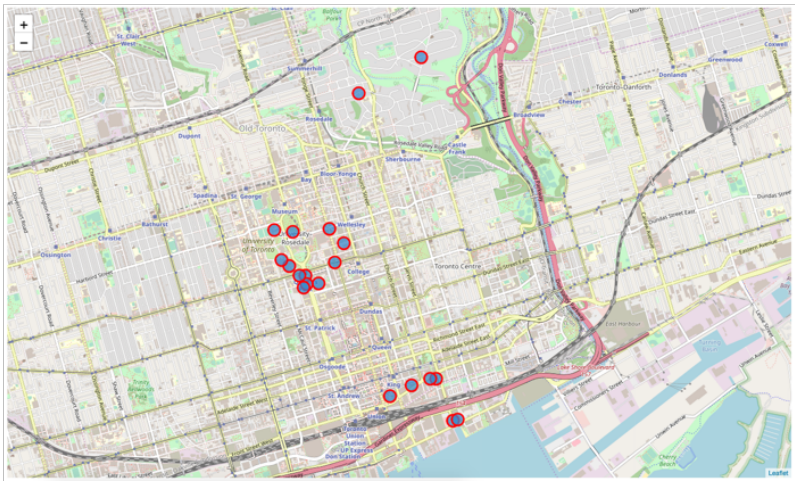
Queen's Park

	name	categories	lat	lng
0	Queen's Park	Park	43.663946	-79.392180
3	Coffee Island	Coffee Shop	43.664271	-79.386972
4	YMCA	Gym	43.662753	-79.384849
5	Coffee Public	Coffee Shop	43.660763	-79.386184
6	Starbucks (MaRS)	Coffee Shop	43.659456	-79.390411
18	Starbucks	Coffee Shop	43.658557	-79.390196
28	Hart House Gym	Gym	43.664172	-79.394888
30	Starbucks	Coffee Shop	43.660412	-79.392692
31	Tim Hortons	Coffee Shop	43.661038	-79.393797
33	Tim Hortons	Coffee Shop	43.658599	-79.388498
34	Tim Hortons	Coffee Shop	43.659415	-79.391221
36	Tim Hortons	Coffee Shop	43.658209	-79.390635

Rosedale

	name	categories	lat	lng
1	Whitney Park	Park	43.682036	-79.373788
2	Alex Murray Parkette	Park	43.678300	-79.382773

Figure 7



To determine the optimal neighborhood for our clients, I then chose to analyze the venue data using frequency and K-means. First, for venue frequency, I grouped rows by neighborhood and took the mean of the frequency of occurrence of each category. I also created a data frame to show the top 3 results from each neighborhood. These results are shown in Figure 8. Additionally, I used K-means (k=3) to cluster the venue results (Figure 9). Although K-means was not very pertinent to the final conclusion, I have included the results of clustering in Figure 11.

Figure 8

[56]:

	Neighborhood	Coffee Shop	Gym	Park
0	Berczy Park	0.833333	0.000000	0.166667
1	Queen's Park	0.750000	0.166667	0.083333
2	Rosedale	0.000000	0.000000	1.000000

[59]:

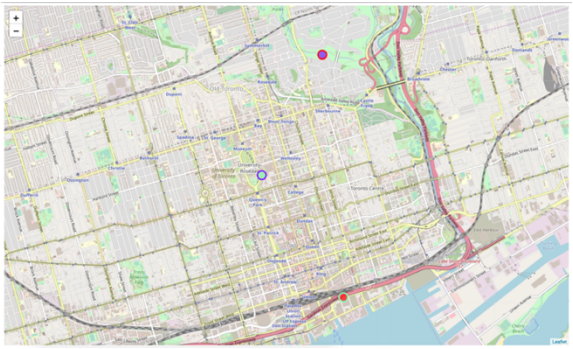
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Berczy Park	Coffee Shop	Park	Gym
1	Queen's Park	Coffee Shop	Gym	Park
2	Rosedale	Park	Gym	Coffee Shop

```
----Berczy Park----
          venue  freq
0   Coffee Shop  0.83
1          Park  0.17
2           Gym  0.00


----Queen's Park----
          venue  freq
0   Coffee Shop  0.75
1           Gym  0.17
2          Park  0.08


----Rosedale----
          venue  freq
0          Park   1.0
1   Coffee Shop   0.0
2           Gym   0.0
```

Figure 9



5. Conclusion

The venue frequency tables proved to be the most important in arriving at a final decision as to the best neighborhood for our clients. We can gather several important conclusions from the frequency tables (Figure 9). First, Rosedale only has parks. It does not have coffee shops or gyms. So, we eliminate Rosedale from our choices since it doesn't meet the client's requirements. Second, Berczy Park has quite a few coffee shops and 1 park but does not have a gym. So, again we will also eliminate Berczy Park since it doesn't meet the client's requirements. Finally, Queen's Park has coffee shops, some gyms, and a park. It definitely meets the client's requirements! So, my final conclusion was that I chose to suggest that the clients buy the house in Queen's Park since it has all the things they want in a neighborhood.

6. Further Information

This Capstone Project with all notebooks, final report, data, and other appropriate files can be found at:
https://github.com/CaptainNano77/Coursera_Capstone