Relocation Services

(Battle of the Neighborhoods)

Krishna Yellamilli

07/15/2020

1. Introduction

1.1. Background

We all moved at least once in our lifetime. Remember the experience of moving! Anyhow, indeed, moving or relocation is a challenging task for anybody. Their neighborhood gets changed totally but people like to have consistency in their lifestyles. Need to make the adaptation to the new surroundings as smooth as possible. Choose an educated neighborhood rather than a random neighborhood. One way to navigate this stressful event is to identify neighborhoods similar to or better than those in their existing neighborhood This is where geolocation exploration comes in handy!

1.2. Problem

The problem is the daunting task of locating one's next neighborhood. People need to relocate to a new place or a neighborhood whether it is due to

- a change in their job
- a change in their financial situation
- a personal reason
- an impending retirement

The relocation creates a problem in people's life-styles, impacting their enjoyment of life. Clients need options to relocate with minimal impact to their day-to-day lives

1.3. Audience

Any person or families that need to move perceive the relocation service as a big relief in locating a suitable neighborhood.

The relocation service will have a large client base comprising of

- Employers with relocating employees
- Individuals or families who are moving to new areas

2. Data

2.1. Sources

2.1.1. From the client: Need the following

- Current location
- Target location
- Another important non-venue care about attribute:
  - Say distance from their target which can be computed using the geographical location
  - It could be anything else that be related to a neighborhood

2.1.2. Web data

- Locale Details: Due to steep cost acquiring geo info of neighborhoods, the current project will be limited to neighborhoods in The City of Toronto, Ontario, Canada. Obtained local postal codes covering boroughs and neighborhoods from Wikipedia's **Toronto's Postal Codes Link**
- Used the provided geo location data from Google
- FourSquare venue data collected using its APIs

2.2. Data Pre-processing/Cleaning/Wrangling

After scraping the Wikipedia for the postal codes in the City of Toronto, Canada, obtained 103 postal codes with their neighborhoods as shown below:

	Postal Code	Borough	Neighborhood
0	M1J	Scarborough	Scarborough Village
1	M5M	North York	Bedford Park, Lawrence Manor East
2	M4C	East York	Woodbine Heights
3	M9A	Etobicoke	Islington Avenue, Humber Valley Village
4	M3M	North York	Downsview

```
# Let's check the size of the DF
T_DF.shape
```

```
(103, 3)
```

After collecting venues within 1000m (increased from 500m due to fewer venues in some postal codes as explained in the Exploratory Data Analysis (EDA)), got only 100 postal codes with venues as shown below:

Postal Code	Accessories Store	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports
M1B	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000
M1C	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000
M1E	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000
M1G	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000
M1H	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.125000
M1J	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000
M1K	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000
M1L	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000

```
T_venues_Grouped.shape
```

```
(100, 269)
```

Dropped the 3 postal codes with missing venues which may be dummy codes setup for Amazon box delivery, etc. and continued the analysis with the remaining 100 postal codes with venues for the client to choose from.

3. Methodology

3.1. Detailed steps of the problem solving method

1. Collect the categories of the venues in their current neighborhood using FourSquare data
2. Identify neighborhoods with venues that have similar characteristic to those of their current neighborhood
3. Since these neighborhoods can be located across a wide geographical area, incorporate their interested non-venue attribute into their decision making
    a. One obvious metric could be the distance of the neighborhood from their target
    b. Will compute the distance between the identified and target neighborhoods using their geolocations
4. Display the similar neighborhood data on an interactive geomap with popup capability showing the following to aid the client's selection process
    a. Neighborhood name
    b. Distance from the target
    c. Popular venues in that neighborhood
    d. Target will be red in color, while their current neighborhood will be in green
5. Output the client's choice/selection
    a. Their new chosen neighborhood
    b. Its comparison with their current neighborhood

3.2. Exploratory Data Analysis (EDA)

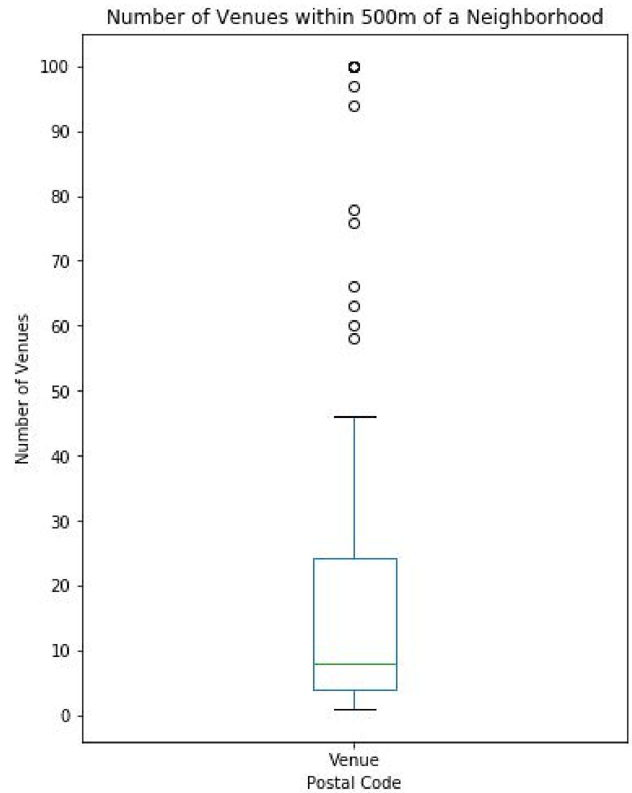
```
VenuesDF.describe()
```

```
count    100.000000
mean      21.180000
std       27.548077
min        1.000000
25%        4.000000
50%        8.000000
75%       24.250000
max      100.000000
Name: Venue Category, dtype: float64
```

The initial analysis of venues was set to 500m of a postal code with 100 venues limit.

- This resulted in some neighborhoods missing venues within 500m (As we can see from the description of the venue DataFrame on the right, only 100 neighborhoods have venues out of the 103).

- The number of venues is varying dramatically from neighborhood to neighborhood
    - Shown by large standard deviation of neary 28 in the description of the DataFrame
    - Depicted in the box plot of the data
- There are few neighborhoods with few venues as shown Q1 of 4.



To increase the number of possible venues in low venues populated neighborhoods, I increased the distance of search within the neighborhood to 1000m while keeping the limit of 100 venues per neighborhood. It did improve the number of venues slightly. It was probably because the venues are concentrated in certain neighborhoods. Here is the new description of the DataFrame with 1km search distance:

```
# With 1000m search distance
VenuesDF.describe()
```

```
count    100.00000
mean      21.36000
std       27.69769
min        1.00000
25%        4.00000
50%        8.00000
75%       24.25000
max      100.00000
Name: Venue, dtype: float64
```

Will continue the analysis of venues of the last set within 1000m of a postal code with 100 venues limit.

Due to wide variation in the number of venues in the neighborhoods, limited the client’s choice to either 5 or 8.

3.3. Clustering with Machine Learning using KMeans

Based on the exploratory data analysis there are nearly 300 different categories of venues in the City of Toronto. At this high dimensionality, it’s extremely difficult to cluster effectively, I resorted to Machine Learning and choose KMeans clustering technique due to the following reasons:

1) Simple to understand and implement:
    a) It’s based on geometric distance from the centroid.
    b) It minimizes the within-cluster sum-of-mean-squares.
    c) It needs just the additional number of clusters required.
    d) It auto initializes cluster centers in a way to speed up convergence of centroids.
    e) It chooses the best outcome out of the default setting of about a dozen (10 to be exact) different centroid seeds.
    f) It has a high number (300) of maximum iterations for achieving convergence of centroids.
2) Scales well with the increasing number of samples.

4. Results

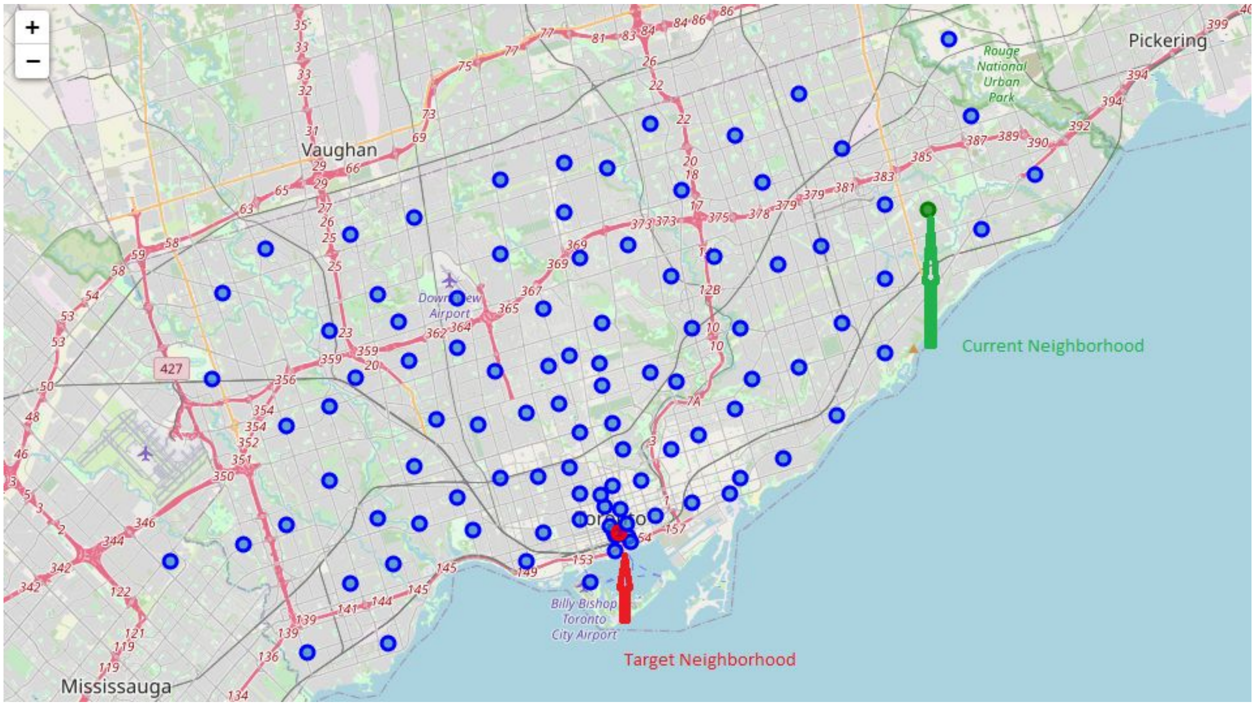
4.1. Interactive Guiding in Selection

Interactively, we can obtain the current neighborhood and the target neighborhood to which the client wants to get closer to within the city of Toronto presented to a prospective client. To make it easy for a client, his entered postal code will be converted to upper case for cross-checking the validity of the postal code entered. Here are sample interactive dialogues:

```
Your current postal code with 3 chars starting with M, a digit, and a char:
m9x

Your entered current postal code: M9X is not valid.
Let's try again.

Your current postal code with 3 chars starting with M, a digit, and a char:
m1g

Your entered current postal code: M1G

Your target postal code with 3 chars starting with M, a digit, and a char:
m5l

Your entered target postal code: M5L
```

4.2. Displaying All Neighborhoods in the City

After obtaining the client's information, I obtain the current neighborhood and the target neighborhood which will be displayed in green and red circles respectively, while other neighborhoods are displayed in blue circles.
Here is an interactive map of the neighborhoods displaying his current and target neighborhoods (Annotated for the clarity in the picture for the report):



4.3. Selection of Number of Clusters

Obtained the client's choice in clustering the neighborhoods on the number of clusters that the would like to choose:

```
Please choose either 5 or 8 for clustering of the neighborhoods: 5
Your chose 5 clusters
```

4.4. Current Neighborhood's Venues Details

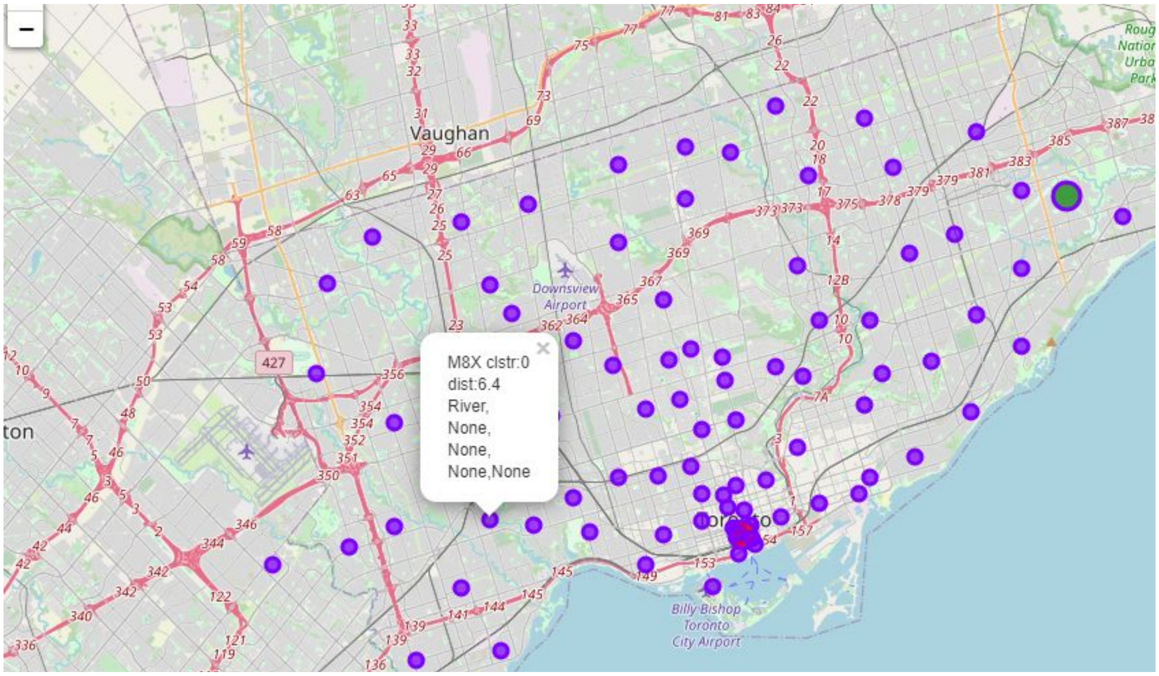
After clustering with KMeans, provided the top venues in the client's current neighborhood, its cluster ID, and its distance from the target neighborhood:

```
Current postal code: M1G
Current neighborhood cluster ID: 0
Current neighborhood 1st venue: Coffee Shop
Current neighborhood 2nd venue: Korean Restaurant
Current neighborhood 3rd venue: None
Current neighborhood 4th venue: None
Current neighborhood 5th venue: None
Current neighborhood distance from M5L postal code: 11.8 miles
```

4.5. Provided options and generated selective maps

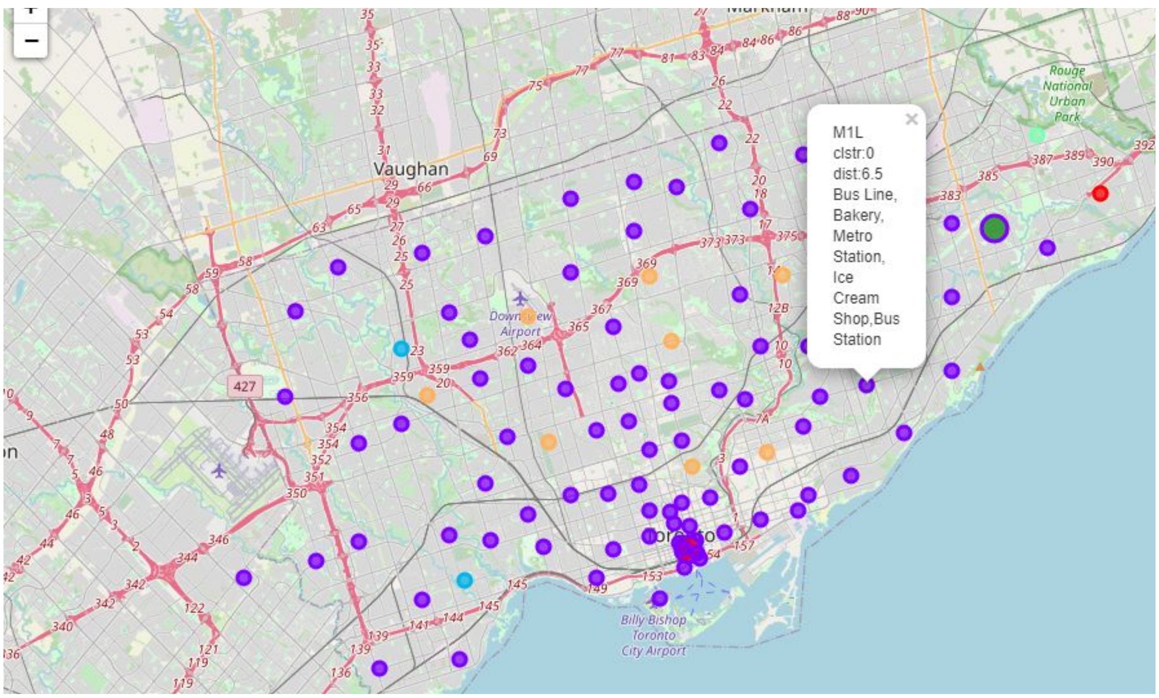
4.5.1. Displaying Neighborhoods similar to the current neighborhood

The neighborhoods in the cluster that belonged to the client's current neighborhood to explore neighborhoods that are similar to the client's neighborhoods:



4.5.2. Displaying all Neighborhoods

All the neighborhoods in the city in case the client wants to explore other neighborhoods:



4.6. Selection Aid

Aided client in his selection process of the new neighborhood showing the top five categories of venues in the neighborhoods as following:

```
Current postal code: M1G
Current neighborhood cluster ID: 0
Current neighborhood 1st venue: Coffee Shop
Current neighborhood 2nd venue: Korean Restaurant
Current neighborhood 3rd venue: None
Current neighborhood 4th venue: None
Current neighborhood 5th venue: None
Current neighborhood distance from M5L postal code: 11.8 miles

Hope you reviewed all the neighborhoods of your choice.
Let's gather the details of your your choice.
Your selection postal code with 3 chars starting with M, a digit, and a char: m4w
Your entered selection postal code: M4W
FYI: Selected cluster 3 is not similar to your current cluster 0
Would like to select a different postal code(y for yes)?y
Your selection postal code with 3 chars starting with M, a digit, and a char: m4x
Your entered selection postal code: M4X

Selected postal code: M4X
Selected neighborhood cluster ID: 0
Selected neighborhood 1st venue: Coffee Shop
Selected neighborhood 2nd venue: Bakery
Selected neighborhood 3rd venue: Pizza Place
Selected neighborhood 4th venue: Restaurant
Selected neighborhood 5th venue: Café
Selected neighborhood distance from M5L postal code: 1.5 miles
Are you OK with your selection (y for yes)?y
```

4.7. Confirmation of the selection

After confirmation of the neighborhood selection, displayed all the venues details of the current neighborhood as following:

	Venue Category	Venue	Venue Distance
0	Coffee Shop	Starbucks	0.222
1	Coffee Shop	Tim Hortons	0.308
2	Korean Restaurant	Korean Grill House	0.121

and all the venues in 1000m within the new neighborhood as following:

	Venue Category	Venue	Venue Distance
0	Diner	Cranberries	0.087
1	Japanese Restaurant	Kingyo Toronto	0.148
2	Restaurant	Murgatroid	0.091
3	Italian Restaurant	F'Amelia	0.056
4	Café	Merryberry Cafe + Bistro	0.108
5	Café	Cabbagetown Brew	0.108
6	Indian Restaurant	Butter Chicken Factory	0.098
7	Bakery	Absolute Bakery & Café	0.087
8	Jewelry Store	Fair Trade Jewellery Co.	0.184
9	General Entertainment	Toronto Dance Theatre	0.123
10	Butcher	St. Jamestown Delicatessen	0.157
11	Gastropub	House on Parliament	0.208
12	Pet Store	Pet Valu	0.263
13	Taiwanese Restaurant	Kenpai Snack Bar	0.252
14	Deli / Bodega	The Epicure Shop	0.276
15	Caribbean Restaurant	Mr. Jerk	0.290
16	Gift Shop	Labour Of Love	0.266
17	Pub	Stout Irish Pub	0.289
18	Park	Wellesley Park	0.300
19	Restaurant	The Pear Tree	0.213
20	Market	Sunny Green Vegetable and Fruit	0.240
21	Coffee Shop	Jetfuel Coffee	0.187
22	Bank	TD Canada Trust	0.229
23	Italian Restaurant	Hey Lucy	0.273
24	Thai Restaurant	Thai Room - Carlton	0.284
25	Liquor Store	LCBO	0.170
26	Playground	Winchester Park	0.232
27	Beer Store	The Beer Store	0.122
28	Bakery	Daniel et Daniel Event Creation & Catering	0.250
29	Coffee Shop	Tim Hortons	0.081
30	Sandwich Place	Subway	0.163
31	Pizza Place	Pizza Pizza	0.114
32	Pub	The Flying Beaver Pubaret	0.219
33	Coffee Shop	Tim Hortons	0.154
34	Pharmacy	Shoppers Drug Mart	0.274
35	Chinese Restaurant	China Gourmet	0.284
36	Pizza Place	Pizza Pizza	0.131
37	Outdoor Sculpture	Rock Towers	0.101
38	American Restaurant	Cabbage Town Kitchen	0.098
39	Grocery Store	Matt's No Frills	0.308
40	Coffee Shop	Tim Hortons / Esso	0.161
41	Gourmet Shop	Cabbage Town Organics	0.226
42	Snack Place	Park Snacks	0.238
43	Furniture / Home Store	Spruce	0.301

5. Discussion

5.1. Observations

5.1.1. Wide variation in the number of venues across neighborhoods

The number of venues is varying dramatically from neighborhood to neighborhood.

- Shown by large standard deviation of neary 28 in the description of the DataFrame
- Depicted in the box plot of the data
- There are few neighborhoods with few venues as shown Q1 of 4.

The clustering was skewed: some with few neighborhoods, while others with a lot of venues.

5.1.2. Met all the current neighborhood venues

Based on the chosen postal code, M4X, it’s very clear that the chosen postal code has all the venues of the current neighborhood as shown below

- All the venue categories that the current postal code has
- More venue categories
- A lot more venues in total 43 versus the current neighborhood's 3
- A lot closer to the target postal code 11 miles versus 2 miles

This is all good but would the client be able to afford the chosen target?

5.2. Recommendation

Based on the above observations (mentioned in 5.1), I would recommend an affordability or a cost of living factor to give another insightful metric for the client to consider a new neighborhood. May be something along the following lines:

- The average/median house price
- The average/median rental details

6. Conclusion

6.1. Summary

Aided the client interactively to choose for the relocation from the current postal code, M1G, which is about 11 miles away to the target area in postal code, M5L, to the new postal code, M4X, which is just within two miles from the target area and exceeding the number of venue categories and with a lot more variety of venues.

6.2. Future work

This relocation service project is not limited just to the City of Toronto, Ontario, Canada. This service can be extended to cover any neighborhood in the world with online data and resources available to scrape the web. As highlighted in section 5.2, I would also like to extend the relocation service to include some affordability or cost of living factors of neighborhoods such as house prices or rents to give clients a comprehensive picture of the neighborhoods.