Using Machine Learning to find the best housing option in Toronto

Carlos Eduardo Flores-Tinoco[a]

[a]IBM Data Science Professional Certificate 2020

This manuscript was compiled on November 5, 2020

A primary concern when moving into a new city is finding a place to live. Selecting a new place to live depends on many factors beyond the distance to work, such as the number and type of venues present in a particular region. However, the importance and need for each of these commodities is a matter of personal taste. Thus, the concept of an ideal house will vary between individuals. In his work, I used a non-supervised machine learning approach to identify common patterns between neighborhoods in Toronto based on the frequency and location of typical venues and places (e.g., restaurants, parks). Through this approach, I identified seven different clusters of neighborhoods. These clusters' properties can be matched to individual preferences to highlight the best community to move in. In sum, this approach displays the feasibility and advantages of machine learning to build recommendation systems for the housing market.

Data Science | Recommendation Systems

Migration is a critical factor that shapes the economics and diversity of every society. Globally, every week 3 million people move into cities(1). This trend has increased 181% over the past 20 years (2). Nowadays, urban population represents 56.2% of the worldwide community, which is expected to increase up to 60.4% by 2030 (2).

The continually increasing urbanization and migration present great opportunities for economic and cultural thriving. At the same time, it also presents challenges for the migrants and receiving communities(3). For instance, newcomers will have to learn and adapt to local rules and costumes. The rate at which newcomers adapt will directly impact their integration to the society.

Cities within themselves show a great range of diversity (2). We can exploit this diversity as a means to facilitate immigrants with potential communities with similar taste and traditions. Therefore, devising a personalized recommendation system that considers the needs and preferences of an incoming person will heavily aid in the selection of the best fitting neighborhood upon migration to a new city. Consequently, newcomers will face a smoother transition into their new home.

In this work, I used an unsupervised machine learning approach to generate neighborhood clusters based on the frequency and type of venues and commodities. Furthermore, I show how these clusters are the starting point to recommend neighborhoods that fit a user's preferences.

Data

Developing a neighborhood recommendation system based on Toronto venues. To offer the best fitting community, any given recommendation system for relocation purposes must consider the spatial distribution of the incoming place and match them to the user's needs. For instance, whether a newcomer requires nearby hospitals, restaurants, sports venues. Furthermore, it should also consider *negative* needs, such as dislike for heavily concurred places. Therefore, the classification of businesses has to be as extensive and inclusive as possible.

The City of Toronto, Ontario, Canada is one of the most diverse cities in the world (4) and one with a high immigration rate (5). Additionally, the distinct types of venues and places are well documented either by public (e.g., Toronto Open Data) (6–8) and private (e.g., FourSquare) (9) entities. Thus, the city of Toronto can be of great interest and use to develop a personalized recommendation system.

Mapping of venues and places of interest in each neighborhood of Toronto. To obtain a comprehensive map of the distinct places of interest in Toronto, I collected the geographical location of schools, hospitals, public transport and commercial places of interest (e.g., restaurants, shops, sports venues) from Toronto Open Data or Foursqare API (See Methods Section and Juypter Notebook for details). This process resulted in a list of 15180 spots of interest within 24 different categories (Table 1). Finally, to determine venues that are nearby a given neighborhood, I calculated the distance between venues to each of the 140 neighborhoods in Toronto, places of interest within less than 1.2km were considered within range.

Results

Clustering Toronto neighborhoods based on the frequency and diversity of venues. To group neighborhoods sharing similar venues, a non-supervised approach is the best since we do not have any labels to corroborate the efficiency of the training model. In particular, as the variables (Table 1) are discrete numeric data (i.e., not categorical data), K-means is a useful strategy to detect differences between communities. With this approach, I could assign the 140 neighborhoods into seven different clusters (Table 2 and Figure 1).

Neighborhood clusters display distinct geographical distribution. Venues within cities tend to be allocated in specific regions or quarters rather than having a random distribution. For instance, a shopping mall would have multiple and different venues within a small area. Moreover, the place where such mall has will depend on the surrounding shops. Therefore, it is reasonable to expect that clusters show unique patterns in their geographical location. Indeed, from their geographical dispersion (Figure 1), there are two types of clusters: clusters

This work is the final report from the Capstone Project: "Battle of the Neighborhoods". The Capstone Project is the 9th and final courser from the IBM Data Science Professional Certificate

This report and the full data analysis can be found at: https://github.com/caeduft/Coursera_Capstone

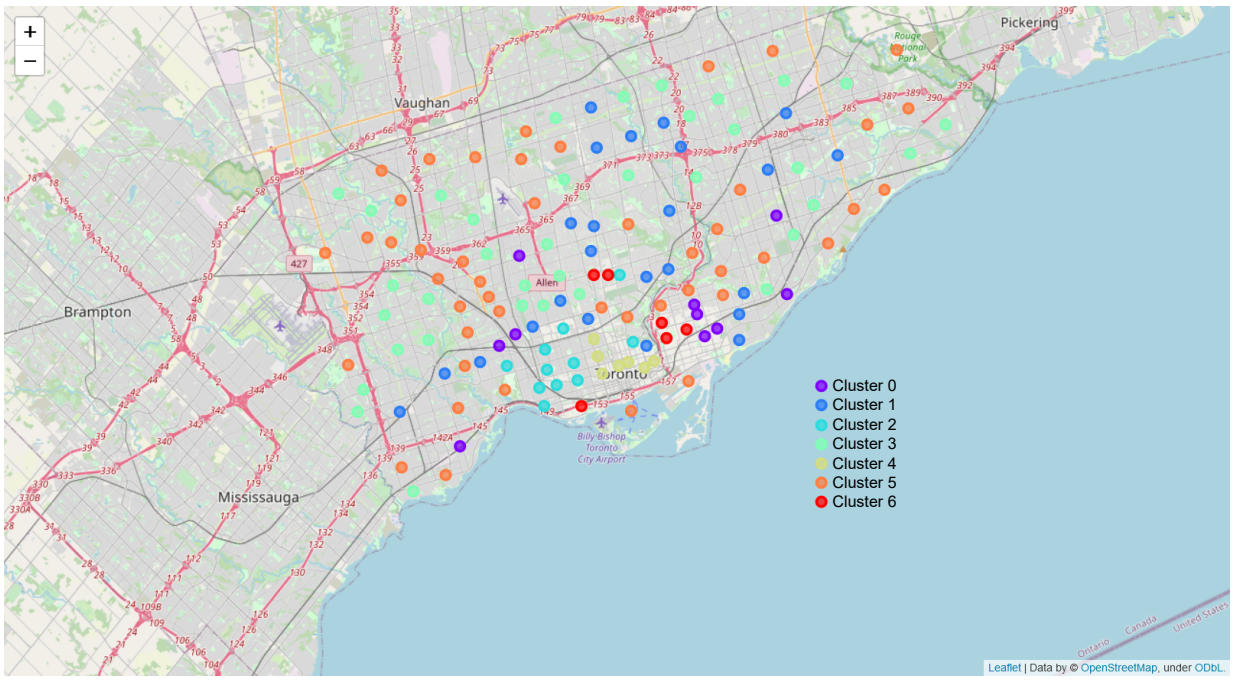

Fig. 1. Toronto neighborhoods can be grouped into seven different clusters. Clustering the 140 neighborhoods (points in the map) according to the type and frequency of venues revealed 7 different classes of communities (Clusters 0-6).

Table 1. Number of venues in the city of Toronto

Group[a]	Venue[b]	Description[c]
Banks	113	Financial services
Cafe	422	Coffe, tea, pastry
Cinema	75	Theaters, cinema
Food Afr ME	82	Africa and Middle East Food
Food Asian	392	Eastern asian food
Food Gral	486	General food venues
Food Indo	57	Indian, Pakistani, and Tibet Food
Food Lat Am	103	Latin American and Caribbean Food
Food Mediterrean	296	Iberic, French, Italian, Greek Food
Food Other Europe	13	German, Swiss, Eastern European Food
Health Centers	51	Hospitals and health centers
Hotel	41	Hotels and hostels
Indoor Attr	11	Bowling, Arcade
Large Oudoor Attr	25	Stadiums, arenas, airports
Markets	208	Markets, grocery stores
Medical	90	Pharmacies
Museum	203	Museums, galleries, community centers
Outdoor Attr	381	Parks, zoos, lakes
Pub	250	Bars, pubs, clubs, lounges
Public transport	9472	Public transport stops and terminals
Salon	35	Beauty salon, spa
Schools	1194	Schools
Shops	1012	Shops, boutiques and services
Sports	168	Gym, courts, sports venues

[a] Group name
[b] Number of venues for the group indicated
[c] Examples of venues allocated in each group

Table 2. Number of venues in the city of Toronto

Cluster	Neighborhood[a]
0	10
1	24
2	11
3	36
4	7
5	46
6	6

[b] Number of neighborhood for the cluster indicated

with high (Clusters 0,1,3, and 5) and low spread (Clusters 2,4,6). Furthermore, clusters that were showed less spread (Clusters 2,4,6) were within the downtown area (Figure 1), where I would expect to have a higher concentration of venues. In sum, the approach I used could cluster even on characteristics which were not used for training, thus, validating the method.

Geographical dispersion of clusters relates to the number and variety of venues within clusters. To confirm that the geographical distribution of clusters (Figure 1) corresponds to the differential presence of venues, I used hierarchical clustering to group the clusters using their multidimensional centroids (Figure 2). As expected, two types of major divisions can be observed. One of these divisions contains a high frequency of venues (Clusters 2, 4, and 6). These clusters are located near Toronto's downtown, where the majority of commercial venues are expected. On the other hand, Clusters 1, 3, and 5 are located further away from the city center and accordingly have fewer venues, possibly indicating more residential areas. Interestingly, Cluster 0 was not part of any of these two divisions

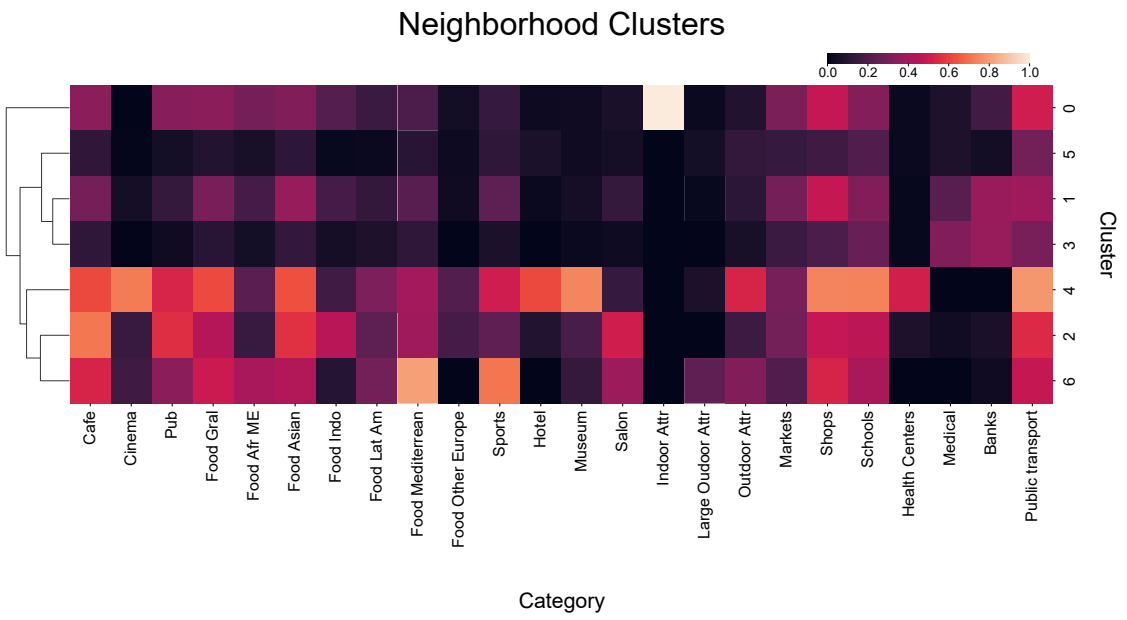

Fig. 2. Clusters display differential variety and number of venues Hierarchical clustering analysis of Toronto's Neighborhood Clusters using the centroids of each Cluster. Categories are the same as in Table 1.

(see Dendrogram in Figure 2) and showed an overall midpoint in venues' frequency between the other two divisions. Indeed, a closer inspection of the geographical profile of Cluster 0 reveals that Cluster 0 is surrounding Clusters 2, 4, and 6; thus bridging these two divisions.

Using the clusters to recommend a newcomer the best fitting neighborhoods. The ultimate goal of the neighborhood clustering is to provide personalized advice upon relocation to Toronto. To this end, an individual profile is correlated to the centroids of the clusters. This individual profile serves as the *ideal* neighborhood for a given user, for which the closest *real* neighborhood will be retrieved. Through this method, a user can use grading scales (absolute need, advantageous, neutral, dislike and heavily disliked) to define preferred places as well as undesired venues. As test cases, I used two contrasting profiles and looked for, according to the clusters, the recommended neighborhoods.

Case 1: A user that wants to be close to commercial venues. For this user, I set that food venues and public transport are absolutely required. At the same time, shops and salons (i.e. beauty salons) are nice to have, though not as crucial as food courts. For this particular profile, no negative preferences were set. The similarity of this profile to each cluster is shown in Table 3:

Table 3. Neighborhood fit to Profile 1

Cluster	2	6	1	0	5	4	3
Similarity[a]	0.53	0.40	0.32	0.16	0.16	0.07	0.02

[a] Correlation between the user profile and the cluster

As expected, the recommendation for this user is to look into cluster 2 and 6, which have a high ratio of food courts (Figure 2). Interestingly, cluster 4 was not recommended, which also has multiple food courts. However, cluster 4 also has numerous schools and museums. Therefore, the chances of a user of finding a food court are reduced by the probability of finding other venues. Hence, cluster 2 and 6 are the most fitting clusters.

Case 2: A user that prefers open spaces and wants to avoid large gatherings. For this user, parks and open places are absolutely required. Sports venues could be advantageous. This user will prefer to avoid large gatherings. Hence, places where people tend to concentrate like schools, pubs, and food courts, are a disadvantage. However, the user will need to get supplies; thus, markets will be advantageous. Furthermore, massive attractions, such as stadiums, are not an option. The similarity of this profile to each cluster is shown in Table 4:

Table 4. Neighborhood fit to Profile 2

Cluster	5	6	3	0	1	4	2
Similarity[a]	0.30	0.08	0.07	0.06	0.05	-0.05	-0.18

[a] Correlation between the user profile and the cluster

As expected, this user was recommended to search far away from the city center and look into cluster 5, which is characterized by having the lowest frequency of venues (Figure 2). However, within Cluster 5, the outdoor attractions and markets are well represented. Interestingly, by showing a negative correlation, the analysis clearly advised against cluster 2, bearing a high ratio of attractions; thus indicating that this cluster is the opposite to the user's desire.

Discussion

Defining a new place to relocate goes far beyond the distance to work/school, and it must be considered if the surroundings are within the expectations of those relocating. It is safe to assume that consumers will be willing to trade distance from their working places in exchange for better surroundings.

However, the definition of better surroundings is an individual preference rather than a global certainty. Thus, it would be of great use to have a tool that removes the hassle of searching over all the neighborhoods for defined commodities in a city and directly points newcomers to the best fitting areas in a personalized approach.

To this end, I defined neighborhood clusters based on the number and diversity of potential places of interest for a newcomer (Figure 1). Next, I used these clusters to determine the best fitting areas for a defined user. Although this process may seem longer and more redundant than a direct correlation between a user's profile and every neighborhood, there are two main advantages of using this approach. On the technical side, analysis on the centroids is faster and more portable than analysis on every neighborhood (i.e., consider running 7 correlations instead of 140 for every user). Most important, on the functional aspect, clusters of similar communities will also include venues and characteristics which were not included in the training (e.g., geographical location, Figure 1). This means that the model contains more information than the input, though on a hidden layer. Therefore, having a system containing these similar untrained characteristics will have a better overall efficiency compared to direct neighborhood comparison, thus, increasing consumer satisfaction.

During the test cases, the algorithm proved to yield the expected results; nonetheless, upcoming iterations of the algorithm will benefit from several considerations. For instance, some of the venue's categories used for the clustering analysis (Table 1) could have been too broad, such as Shops and Schools, or to specific, such as the food venues classification. Moreover, the clustering analysis revealed 7 clusters; however, the number of neighborhoods in each cluster varied significantly (Table 2). For example, Cluster 5 contains 46 communities. Although it is possible that these 46 neighborhoods are incredibly similar to each other, it is also possible that at this point, the information captured is not enough to separate them into more clusters. Furthermore, at this stage, beyond the test profiles, there is no feedback to the performance on the algorithm. Thus, a proper assessment will include the test on real user-profiles and their feedback. Addressing all these questions will improve the algorithm performance.

Conclusion

Altogether, the model I present in this work provides an initial standpoint for a functional personalized recommendation system of neighborhood relocation. Such a system can be easily used by individuals willing to find the best neighborhood to relocate as well as relocation agencies to provide better offers to their customers. Though this model was done for the city of Toronto, the methodology could be applied to any other city. Furthermore, this system is not limited to the housing market and could be used to forecast the success of business according to the surrounding venues.

Methods

Data Availability. All the input data sets used and the entire data analysis can be found at https://github.com/caeduft/Coursera_Capstone

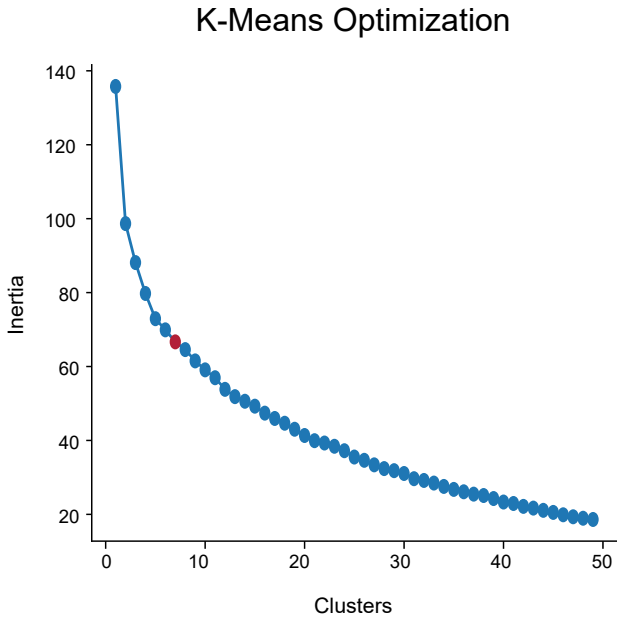

Fig. 3. Determination of the optimal number of K-Means clusters. Inertia obtained after K-Means clustering using different number of clusters. The optimal point (seven clusters) is highlighted in red.

Data Collection. Geographical location of places of interest in Toronto were acquired from the Toronto Open Data Portal (7) and Foursquare API (10). Specifically, information regarding: neighborhood name and location, neighborhood social metrics, hospitals and health centers, schools, public transports, museums and community centers were obtained from the Toronto Open Data Portal. Venues within 2500m from the center of each neighborhood were retrieved from the Foursquare API version '20180605' using the explore endpoint, latitude and longitude parameters. Due to the multiple data inputs, duplicated entry venues were removed based on the venue's name and coordinates. To simplify the categories and increase the consistency of the data set, venues were allocated into one of the 24 final categories (Table 1). Data was parsed and processed using JSON (v. 2.0.9), Requests (v. 2.24.0) (11), and Pandas (v. 1.1.1) packages (12) in Python v. 3.8.5 (13)

Distance calculation between venues and neighborhoods. The distance between a venue and the center of the neighborhood was calculated using the geodesic distance. The geodesic distance was obtained with the GeoPy package (v. 2.0.0) (14) using the `geopy.distance.distance` function and the coordinates as parameters.

K-Means clustering of Toronto neighborhoods. Since the categories have a different range of venues(e.g. public transport vs indoor attractions), prior clustering, the frequency of the venues was Min-Max normalized per group category. in this way, the differences between neighborhoods per group was retained but all catergories were within the range [0,1]. Thus, all groups will have the same weight during the clustering.

The normalized frequency of the venues was used to define clusters using K-Means. K-Means clusters were obtained with the Sklearn package (v. 0.23.2) (15) using the `KMeans(init = "k-means++", n_clusters = 7, n_init = 12,random_state=1)` function from the cluster module. The number of clusters (n_clusters parameter) was determined empirically by graphing the inertia against the number of clusters and determined that the optimum was 7 clusters (Figure 3).

Plotting neighborhood clusters in Toronto's map. Neighborhoods and their cluster association were plotted onto Toronto Maps using the Folium package (v. 0.11.0) (16). Maps were obtained from OpenStreet Maps (17).

Hierarchical clustering analysis. Hierarchical clustering of the K-Means clusters was done using the centroids provided by the K-means clustering with correlation as the metric distance. Metrics and heatmap were obtained with the `clustermap` function from the Seaborn package (v. 0.11.0) (18).

Correlation analysis between a user's profile and the neighborhood clusters. Pearson correlation between a defined user prefrences' profile and the Toronto neighborhood clusters was done using the `corr` function from the Pandas (v. 1.1.1) package (12)

To classify the user's preferences for each category the following scale was used:

`5: Absolute requirement, 4: Advantageous, 3: Neutral, 2: Disadvantage, 1: Completely undesired`

To reflect the degree of preference between categories, prior the correlation analysis, each value of the user's preferences was elevated to the power of 2. This transformed the scale from `[5,4,3,2,1]` to `[25,16,9,4,1]`.

The profiles used for the test cases were:

profile1 = 'Cafe': 5, 'Cinema' : 3, 'Pub' :3, 'Food Gral':5, 'Food Afr ME' : 5, 'Food Asian':5, 'Food Indo':5, 'Food Lat Am':5, 'Food Mediterrean':5, 'Food Other Europe':5, 'Sports':3, 'Hotel':3, 'Museum':3, 'Salon':4,'Indoor Attr':3, 'Large Oudoor Attr':3,'Outdoor Attr':3, 'Markets':3, 'Shops':4, 'Schools':3, 'Health Centers':3, 'Medical':3, 'Banks':3, 'Public transport':5

profile2 = 'Cafe': 2, 'Cinema' : 1, 'Pub' :1, 'Food Gral':2, 'Food Afr ME' : 2, 'Food Asian':2, 'Food Indo':2, 'Food Lat Am':2, 'Food Mediterrean':2, 'Food Other Europe':2, 'Sports':4, 'Hotel':2, 'Museum':2, 'Salon':2,'Indoor Attr':3, 'Large Oudoor Attr':1,'Outdoor Attr':5, 'Markets':4, 'Shops':2, 'Schools':1, 'Health Centers':2, 'Medical':3, 'Banks':2, 'Public transport':3

1. UN-Habitat (2009) *State of the World's Cities Report 2008/9: Harmonious Cities*. (United Nations Human Settlements Programme (UN-Habitat)).
2. UN-Habitat (2020) *World cities report 2020: The Value of Sustainable Urbanization*. (United Nations Human Settlements Programme (UN-Habitat)).
3. International Organization for Migration (IOM) (2015) *World migration report 2015*. (International Organization for Migration (IOM)).
4. City of Toronto: Social Development, Finance and Administration Division (2013) *Toronto Newcomer Strategy*. (City of Toronto).
5. City of Toronto (2017) *2016 Census: Housing, Immigration and Ethnocultural Diversity, Aboriginal peoples*. (City of Toronto).
6. City of Toronto: City Clerk's Office (2011) *Open Data Policy*. (City of Toronto).
7. City of Toronto (2020) Open Data. [Online; accessed 26-October-2020].
8. City of Toronto (2020) Open Data Licence. [Online; accessed 26-October-2020].
9. Foursquare (2020) Foursquare. [Online; accessed 26-October-2020].
10. Foursquare (2020) Foursquare API. [Online; accessed 26-October-2020].
11. Chandra RV, Varanasi BS (2015) *Python requests essentials*. (Packt Publishing Ltd).
12. McKinney W, , et al. (2010) Data structures for statistical computing in python in *Proceedings of the 9th Python in Science Conference*. (Austin, TX), Vol. 445, pp. 51–56.
13. Van Rossum G, Drake FL (2009) *Python 3 Reference Manual*. (CreateSpace, Scotts Valley, CA).
14. Geopy contributors (2020) GeoPy. [Online–Manual; accessed 03-November-2020].
15. Pedregosa F, et al. (2011) Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12(Oct):2825–2830.
16. Folium contributors (2020) Folium (https://pypi.org/project/folium/).
17. OpenStreetMap contributors (2017) Planet dump retrieved from https://planet.osm.org (https://www.openstreetmap.org).
18. Waskom M, et al. (2017) mwaskom/seaborn: v0.8.1 (September 2017).