Finding the Right Neighborhood: Toronto and the Twin Cities

Alexander Stetzer

June 12th, 2021

Introduction

The Twin Cities of Minnesota are a culturally diverse region of the Upper Midwest. Made up of the cities of Minneapolis and St. Paul, the Twin Cities are closely linked and seem almost like the same city, hence the name Twin Cities. Toronto, being a key city in Canada, also has this diverse culture. With many similar characteristics between these cities and the economic growth of these communities, movement between the two is possible. In order to find similar neighborhoods between the two, the goal is to use venue data to find similarities between neighborhoods. If a person needs to move from Toronto to the Twin cities or vice versa they might want to live in a place with similar venues. Using KMeans clustering and cosine similarity, neighborhoods with similar venues can be found.

The interested parties for the project are the people moving in between the cities and looking for new residences. A neighborhood with similar venues may allow for an easier time getting familiar with the surrounding area. Also the project could be used in reverse, instead of finding a similar neighborhood, one could use it to look for neighborhoods to avoid. Another group that could benefit from this project is Real Estate agencies. Using this data, the agencies could provide suggestions on places to look for their clients.

Data

The data used for the analysis primarily comes from Wikipedia. Wikipedia was chosen for the source of the data as it is easy to web scrape and gather cleaner data from. The main goal for gathering the data was to use postal codes, or zip codes in the case of the United States, to find the names for the neighborhoods. Then, using the postal codes, location data can be found. A problem started to arise because in the United States zip codes often overlap multiple cities or a single city has around five zip codes to itself. The scarcity of the zip code locations would mean that multiple neighborhoods would not be able to be separated and precision of the data would be too wide. To remedy this, neighborhood lists from wikipedia would be used and then the location data extracted.

The data for Minneapolis neighborhoods, gathered from here, was quite simple to webscrape as it is in an easy html code structure. Using BeautifulSoup, the communities were under h3 tags and the neighborhoods were under table tags. The communities were placed into one list and the neighborhoods into another. After the lists were constructed, the lists were combined with the city of Minneapolis placed in a city column.

For St. Paul, the data from the wikipedia page was quite messy (St. Paul Wiki). Districts, as they are called in St. Paul, were listed as the neighborhoods contained inside of them. Some of the districts were listed as the correct name, but in the paragraph explaining the district there was no mention of the neighborhoods contained inside. Due to the odd makeup of the data and the small size of neighborhoods, it was decided that a brute force method was best. Using the wiki page as a guideline, all seventeen districts were put into a dictionary with districts as keys and neighborhoods as values. In the case that a neighborhood did not appear in the district description, a google maps search was used to find the correct neighborhood names. Once all the necessary data was collected, the lists were combined, St. Paul was placed in a city column, and a dataframe constructed.

Since Canada uses a more accurate postal code system, Toronto data was an easier endeavor. Using the wikipedia page here, the postal codes were stripped from the website using BeautifulSoup. The data was easy to gather as it was all under tables and minimal cleaning was needed to get a complete dataframe with postal codes, communities, and neighborhoods. After the dataframe was constructed and a city column with Toronto placed in every row. After all the neighborhood dataframes were constructed they were concatenated into a single dataframe to complete the neighborhood set.

Once all of the neighborhoods were found, location data for each neighborhood was acquired. Geocoder was initially used to find the data, but was found extremely unreliable and took forever to get a proper response for any neighborhood. To alleviate the location problem Google's Geocoder API was used directly to find the location data for the Twin Cities neighborhoods. Looping through the neighborhood dataframe all of the latitude and longitude data was easily found. To find the location data for the neighborhoods of Toronto, a spreadsheet of postal code geospatial coordinates was used. This was done as it would best line up the way the neighborhood data was collected. The postal code was read into the system and was merged with the Toronto Neighborhood data only keeping the postal codes that were included in the neighborhood data.

With all location data collected, the next step was to find venue data. Using the Foursquare API, a master list of every venue from all of the neighborhoods was created. For ease of machine learning, one hot encoding is used to separate the venue categories into multiple columns. Since the main goal of the project is to find similarities in neighborhoods, the venues were grouped under each neighborhood that they belong to. Along with grouping the data, the mean of each neighborhood venue categories is used. This creates the dataset that is used for the final KMeans Clustering operation.

Methodology

Once the data is collected and wrangled the real fun begins. The main goal for the interested parties is to find a similar neighborhood to live in. With the Foursquare venue category data, a general idea of a neighborhood can be made. Neighborhoods that have similar venues, parks, elementary schools, and others, tend to have similar feelings.

To find these similar neighborhoods, KMeans clustering was used. KMeans clustering was the chosen clustering machine learning algorithm because of its ability to eek out similarities within the data. Five clusters were chosen for the final project as five clusters seemed to produce a great cluster set.

Even with a large amount of clusters, finding differences between the neighborhoods within the cluster is difficult. With KMeans, there tends to be a few clusters of 10-15 and then a major 100+ cluster. To help the interested parties, another further similarity solution is used. The solution that was used is cosine similarity. Equation 1 shows the formula used to find the similarity.

$$cos\theta \ = \ \frac{A \cdot B}{||A||*||B||} \tag{1}$$

In the equation, A and B are the mean one hot encoded neighborhood data as briefly shown in table 1.

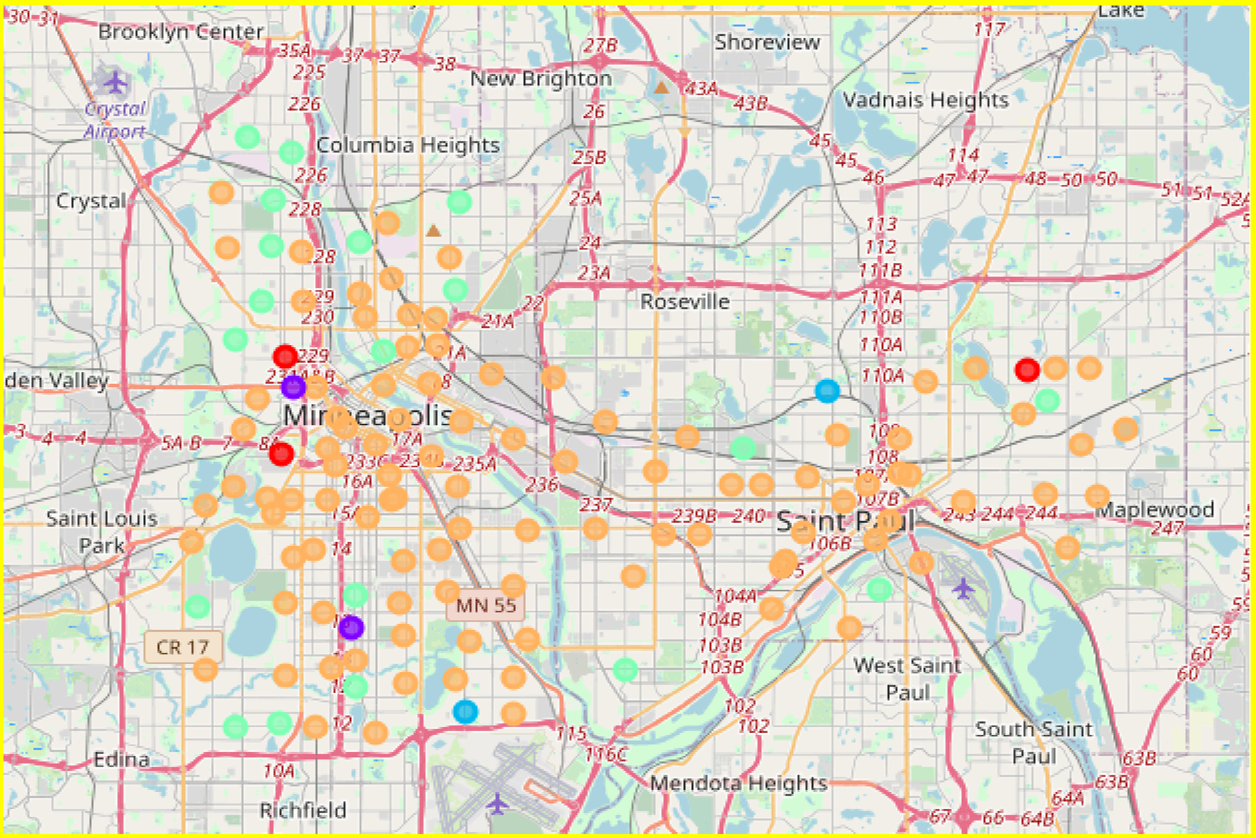
City	Neighborhood	ATM	Airport	Business Service	Cable Car
Minneapolis	Armatage	0.0	0.0	0.16667	0.0
Minneapolis	Jordan	0.16667	0.0	0.0	0.0

Table 1: Example of the mean one hot encoding data. Categories shortened to show examples as true dataFrame has 342 categories.

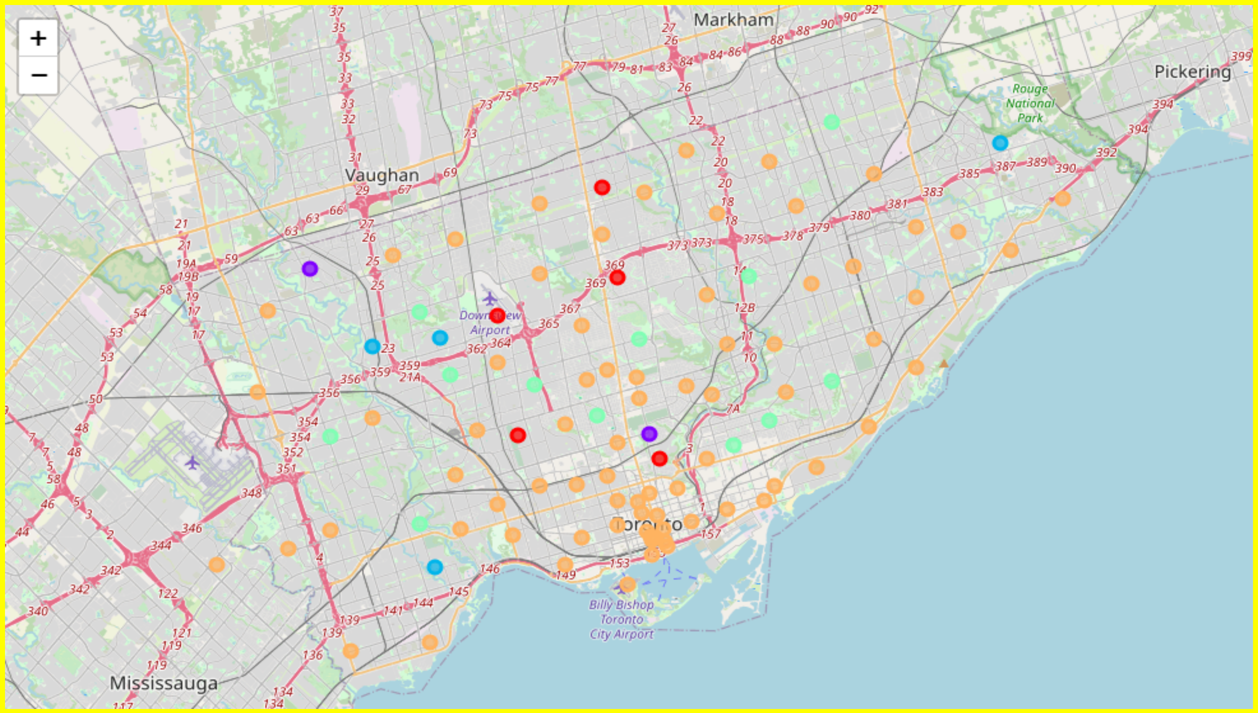
Using the neighborhood data from table 1, dot products between the two neighborhoods make up the numerator of equation 1. Then the length of the two neighborhood vectors is multiplied together, which makes up the denominator of the cosine similarity. In the end, the final cosine similarity will provide a score between one and zero. As the cosine similarity reaches closer to one, the more similar the neighborhoods are. With the KMeans clustering and the cosine similarity calculations, the final recommendations can be made. The final part is to allow an interested person to input their neighborhood and find recommendations based on the solutions. The neigh_finder function scans through each of the cosine similarities and finds all of the similarities with the chosen neighborhood. Then, neigh_finder function returns the top five similar neighborhoods with similarity scores and the top ten most common venues of each neighborhood.

Results

The main results for the project can be broken into three parts clustering, cosine similarity, and final printables. Maps 1 and 2 shows where the neighborhoods are and the clustering group that they belong to.



Map 1: Cluster Map of the Twin Cities Neighborhoods



Map 2: Cluster Map of Toronto Neighborhoods

As shown by both maps, there is a large cluster where all of the neighborhoods end up. Table 2 shows the amount of neighborhoods in each cluster.

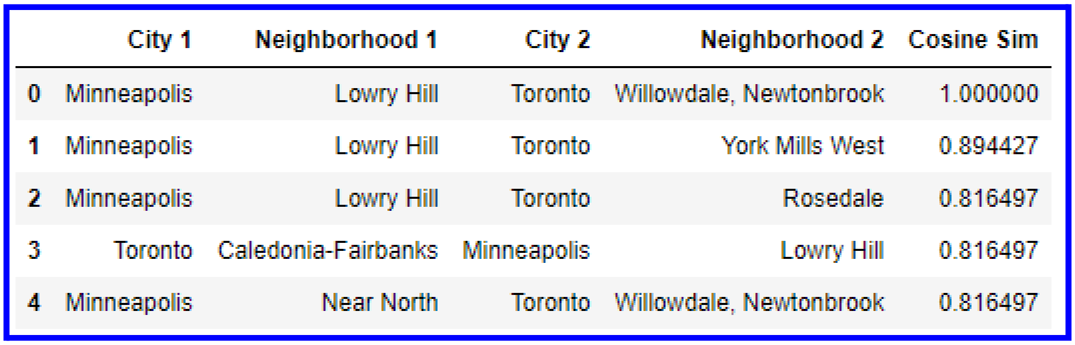
Cluster Number	0	1	2	3	4
Total Neighborhoods	8	4	6	31	177

Table 2: Total number of neighborhoods in each cluster. Cluster numbers tend to fluctuate, but this is a common example for the data set.

The need for the additional refinement is shown by the table. Four neighborhoods may be easy to sift through, but 177 is a bit too many and would take forever.

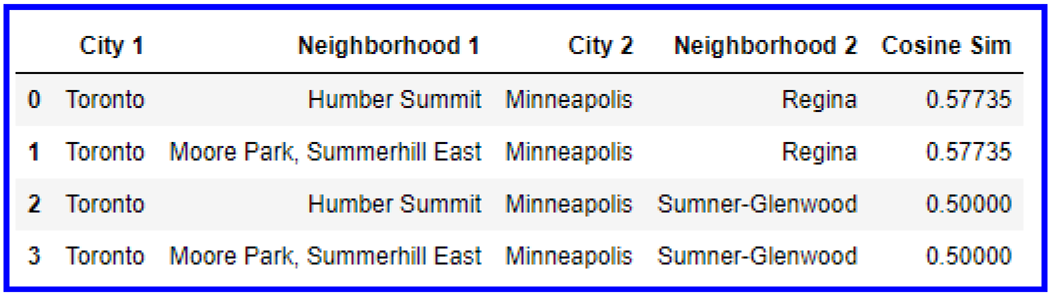
The cosine similarity results are split into each cluster as the refinement is done within the clusters. Table 3 shows the top five similarities from each cluster.

Cluster 0:



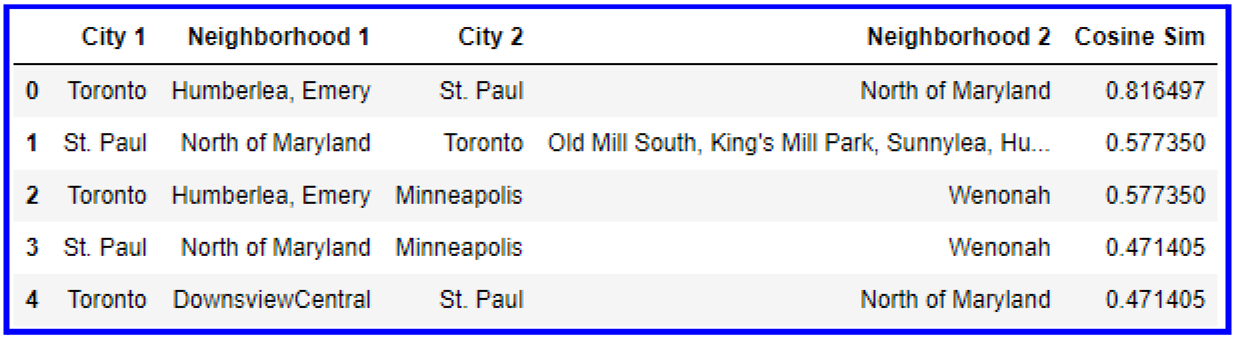
	City 1	Neighborhood 1	City 2	Neighborhood 2	Cosine Sim
0	Minneapolis	Lowry Hill	Toronto	Willowdale, Newtonbrook	1.000000
1	Minneapolis	Lowry Hill	Toronto	York Mills West	0.894427
2	Minneapolis	Lowry Hill	Toronto	Rosedale	0.816497
3	Toronto	Caledonia-Fairbanks	Minneapolis	Lowry Hill	0.816497
4	Minneapolis	Near North	Toronto	Willowdale, Newtonbrook	0.816497

Cluster 1:



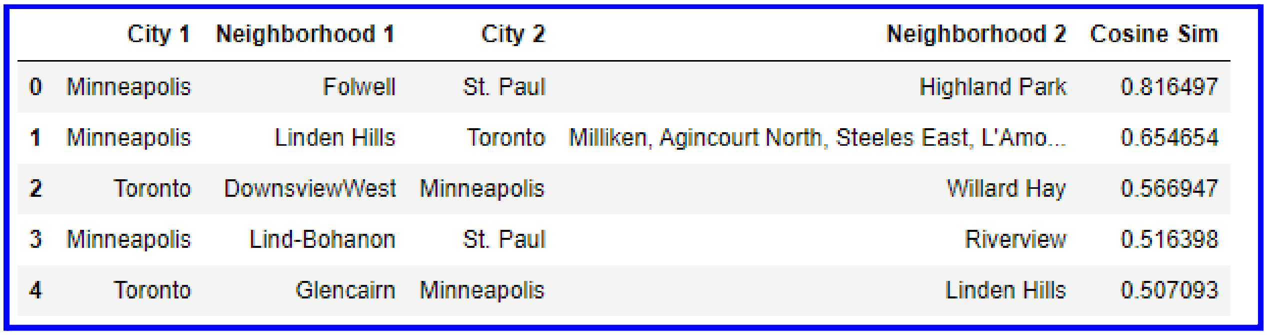
	City 1	Neighborhood 1	City 2	Neighborhood 2	Cosine Sim
0	Toronto	Humber Summit	Minneapolis	Regina	0.57735
1	Toronto	Moore Park, Summerhill East	Minneapolis	Regina	0.57735
2	Toronto	Humber Summit	Minneapolis	Sumner-Glenwood	0.50000
3	Toronto	Moore Park, Summerhill East	Minneapolis	Sumner-Glenwood	0.50000

Cluster 2:



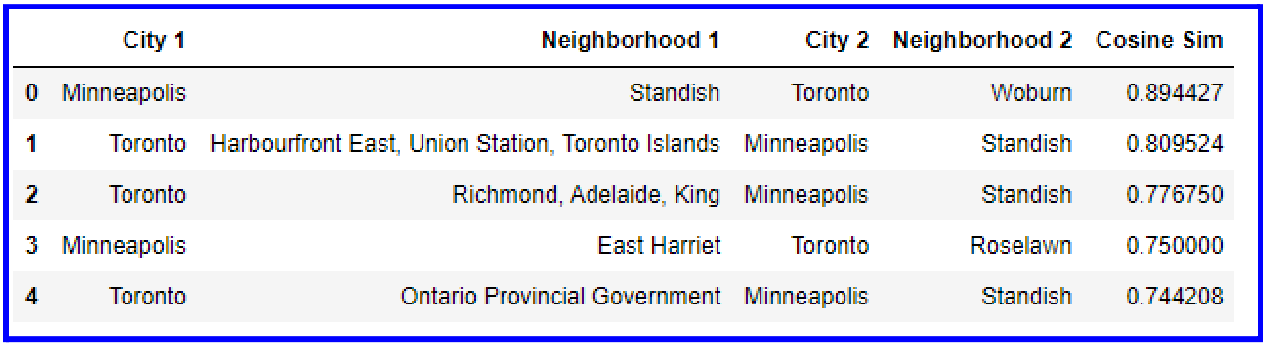
	City 1	Neighborhood 1	City 2	Neighborhood 2	Cosine Sim
0	Toronto	Humberlea, Emery	St. Paul	North of Maryland	0.816497
1	St. Paul	North of Maryland	Toronto	Old Mill South, King's Mill Park, Sunnylea, Hu...	0.577350
2	Toronto	Humberlea, Emery	Minneapolis	Wenonah	0.577350
3	St. Paul	North of Maryland	Minneapolis	Wenonah	0.471405
4	Toronto	DownsviewCentral	St. Paul	North of Maryland	0.471405

Cluster 3:



	City 1	Neighborhood 1	City 2	Neighborhood 2	Cosine Sim
0	Minneapolis	Folwell	St. Paul	Highland Park	0.816497
1	Minneapolis	Linden Hills	Toronto	Milliken, Agincourt North, Steeles East, L'Amo...	0.654654
2	Toronto	DownsviewWest	Minneapolis	Willard Hay	0.566947
3	Minneapolis	Lind-Bohanon	St. Paul	Riverview	0.516398
4	Toronto	Glencairn	Minneapolis	Linden Hills	0.507093

Cluster 4:



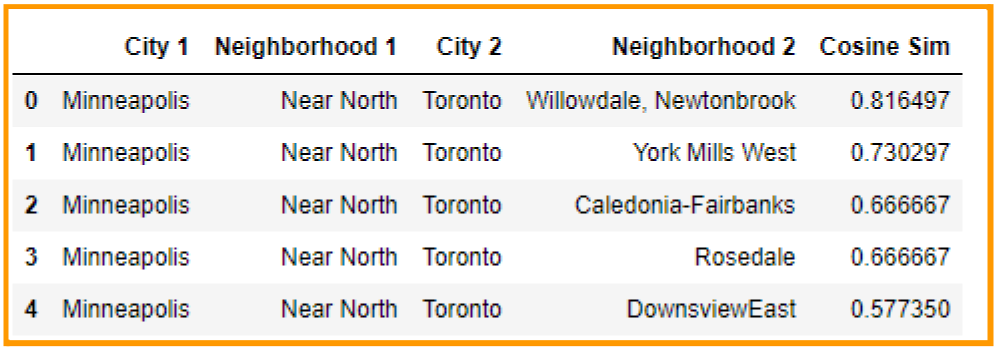
	City 1	Neighborhood 1	City 2	Neighborhood 2	Cosine Sim
0	Minneapolis	Standish	Toronto	Woburn	0.894427
1	Toronto	Harbourfront East, Union Station, Toronto Islands	Minneapolis	Standish	0.809524
2	Toronto	Richmond, Adelaide, King	Minneapolis	Standish	0.776750
3	Minneapolis	East Harriet	Toronto	Roselawn	0.750000
4	Toronto	Ontario Provincial Government	Minneapolis	Standish	0.744208

Table 3: Top 5 Cosine Similarities of Each Cluster

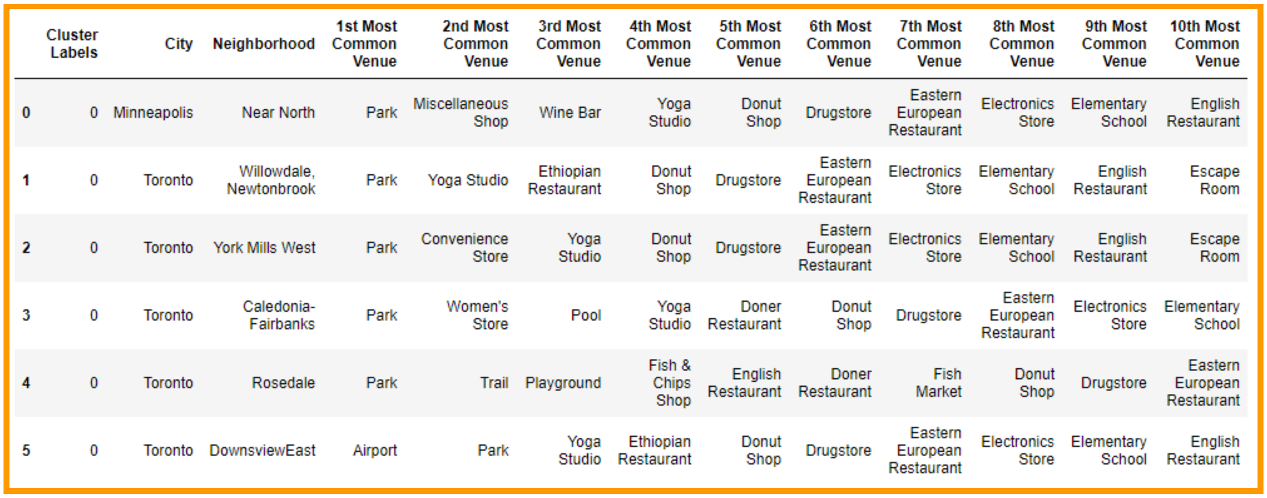
Looking through the clusters, the top five of each cluster seems to be very similar. Cluster 1 suffers from the fact that it only has four neighborhoods to choose from, but all are at least 50% similar.

The top five from each cluster are a good look into the cluster quality, but does not help the interested parties. That is where the neigh_finder function is involved. Table 4 shows the neigh_finder results for one neighborhood from each city and a couple different clusters.

Near North, Minneapolis:

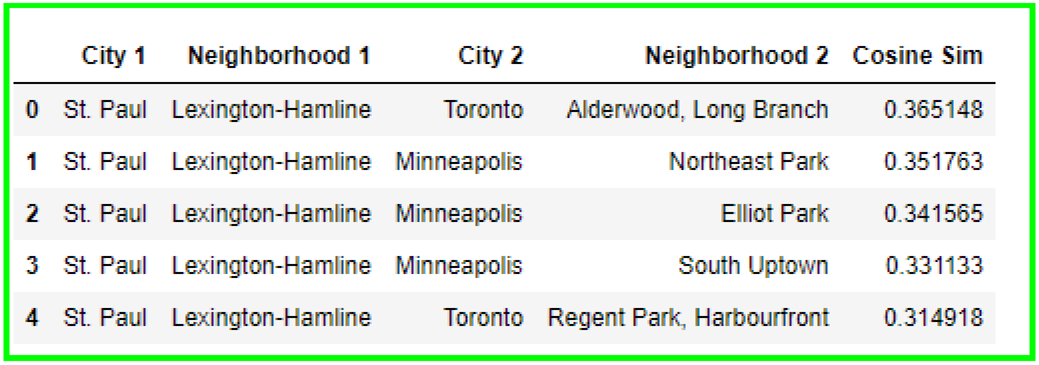


	City 1	Neighborhood 1	City 2	Neighborhood 2	Cosine Sim
0	Minneapolis	Near North	Toronto	Willowdale, Newtonbrook	0.816497
1	Minneapolis	Near North	Toronto	York Mills West	0.730297
2	Minneapolis	Near North	Toronto	Caledonia-Fairbanks	0.666667
3	Minneapolis	Near North	Toronto	Rosedale	0.666667
4	Minneapolis	Near North	Toronto	DownsviewEast	0.577350

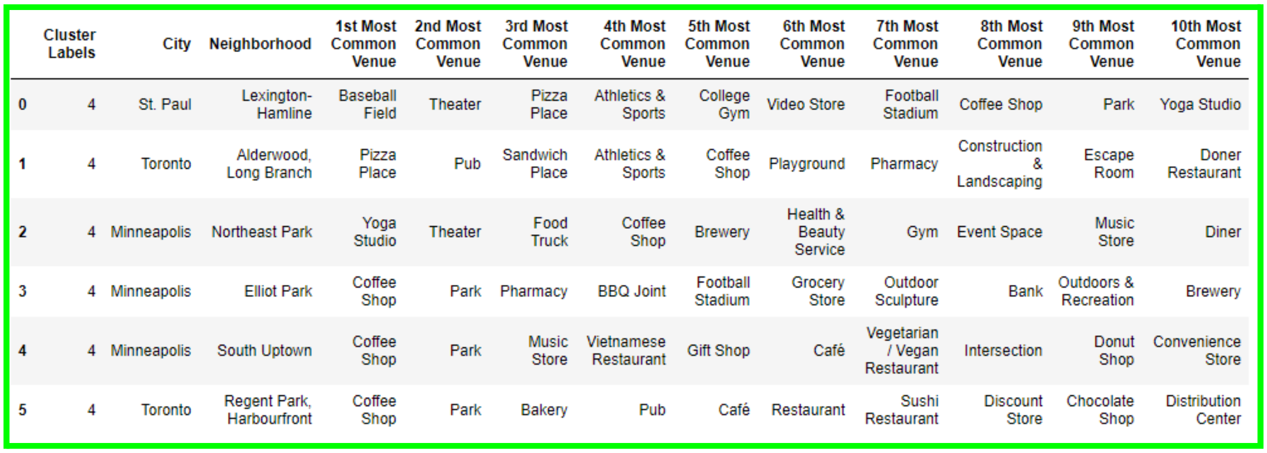


	Cluster Labels	City	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	0	Minneapolis	Near North	Park	Miscellaneous Shop	Wine Bar	Yoga Studio	Donut Shop	Drugstore	Eastern European Restaurant	Electronics Store	Elementary School	English Restaurant
1	0	Toronto	Willowdale, Newtonbrook	Park	Yoga Studio	Ethiopian Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	Electronics Store	Elementary School	English Restaurant	Escape Room
2	0	Toronto	York Mills West	Park	Convenience Store	Yoga Studio	Donut Shop	Drugstore	Eastern European Restaurant	Electronics Store	Elementary School	English Restaurant	Escape Room
3	0	Toronto	Caledonia-Fairbanks	Park	Women's Store	Pool	Yoga Studio	Doner Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	Electronics Store	Elementary School
4	0	Toronto	Rosedale	Park	Trail	Playground	Fish & Chips Shop	English Restaurant	Doner Restaurant	Fish Market	Donut Shop	Drugstore	Eastern European Restaurant
5	0	Toronto	DownsviewEast	Airport	Park	Yoga Studio	Ethiopian Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	Electronics Store	Elementary School	English Restaurant

Lexington-Hamline, St. Paul

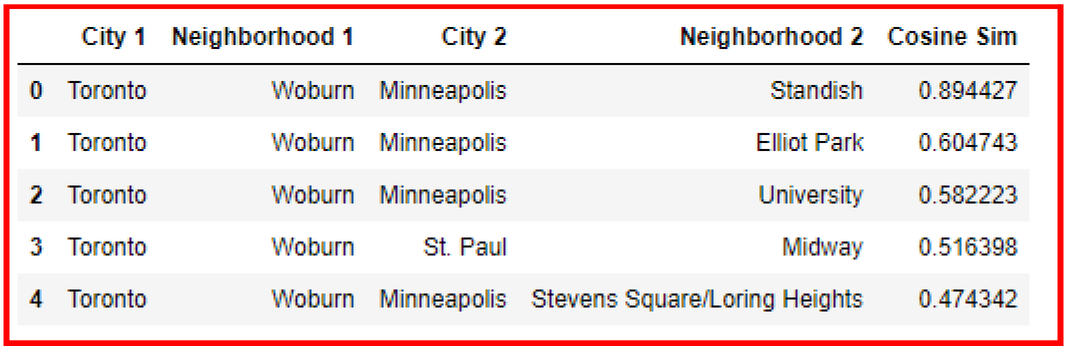


	City 1	Neighborhood 1	City 2	Neighborhood 2	Cosine Sim
0	St. Paul	Lexington-Hamline	Toronto	Alderwood, Long Branch	0.365148
1	St. Paul	Lexington-Hamline	Minneapolis	Northeast Park	0.351763
2	St. Paul	Lexington-Hamline	Minneapolis	Elliot Park	0.341565
3	St. Paul	Lexington-Hamline	Minneapolis	South Uptown	0.331133
4	St. Paul	Lexington-Hamline	Toronto	Regent Park, Harbourfront	0.314918

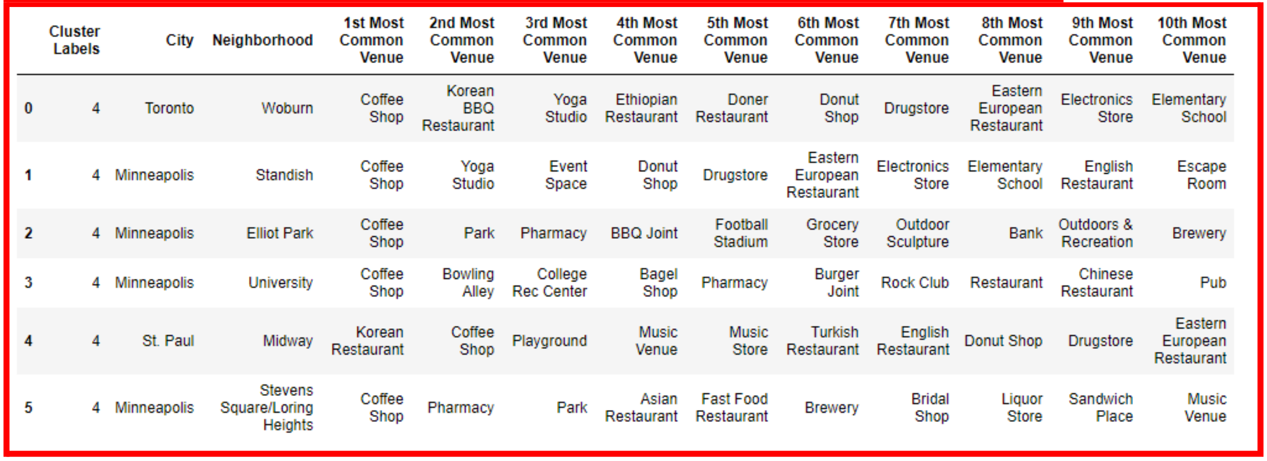


	Cluster Labels	City	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	4	St. Paul	Lexington-Hamline	Baseball Field	Theater	Pizza Place	Athletics & Sports	College Gym	Video Store	Football Stadium	Coffee Shop	Park	Yoga Studio
1	4	Toronto	Alderwood, Long Branch	Pizza Place	Pub	Sandwich Place	Athletics & Sports	Coffee Shop	Playground	Pharmacy	Construction & Landscaping	Escape Room	Doner Restaurant
2	4	Minneapolis	Northeast Park	Yoga Studio	Theater	Food Truck	Coffee Shop	Brewery	Health & Beauty Service	Gym	Event Space	Music Store	Diner
3	4	Minneapolis	Elliot Park	Coffee Shop	Park	Pharmacy	BBQ Joint	Football Stadium	Grocery Store	Outdoor Sculpture	Bank	Outdoors & Recreation	Brewery
4	4	Minneapolis	South Uptown	Coffee Shop	Park	Music Store	Vietnamese Restaurant	Gift Shop	Café	Vegetarian / Vegan Restaurant	Intersection	Donut Shop	Convenience Store
5	4	Toronto	Regent Park, Harbourfront	Coffee Shop	Park	Bakery	Pub	Café	Restaurant	Sushi Restaurant	Discount Store	Chocolate Shop	Distribution Center

Woburn, Toronto:



	City 1	Neighborhood 1	City 2	Neighborhood 2	Cosine Sim
0	Toronto	Woburn	Minneapolis	Standish	0.894427
1	Toronto	Woburn	Minneapolis	Elliot Park	0.604743
2	Toronto	Woburn	Minneapolis	University	0.582223
3	Toronto	Woburn	St. Paul	Midway	0.516398
4	Toronto	Woburn	Minneapolis	Stevens Square/Loring Heights	0.474342



	Cluster Labels	City	Neighborhood	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	4	Toronto	Woburn	Coffee Shop	Korean BBQ Restaurant	Yoga Studio	Ethiopian Restaurant	Doner Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	Electronics Store	Elementary School
1	4	Minneapolis	Standish	Coffee Shop	Yoga Studio	Event Space	Donut Shop	Drugstore	Eastern European Restaurant	Electronics Store	Elementary School	English Restaurant	Escape Room
2	4	Minneapolis	Elliot Park	Coffee Shop	Park	Pharmacy	BBQ Joint	Football Stadium	Grocery Store	Outdoor Sculpture	Bank	Outdoors & Recreation	Brewery
3	4	Minneapolis	University	Coffee Shop	Bowling Alley	College Rec Center	Bagel Shop	Pharmacy	Burger Joint	Rock Club	Restaurant	Chinese Restaurant	Pub
4	4	St. Paul	Midway	Korean Restaurant	Coffee Shop	Playground	Music Venue	Music Store	Turkish Restaurant	English Restaurant	Donut Shop	Drugstore	Eastern European Restaurant
5	4	Minneapolis	Stevens Square/Loring Heights	Coffee Shop	Pharmacy	Park	Asian Restaurant	Fast Food Restaurant	Brewery	Bridal Shop	Liquor Store	Sandwich Place	Music Venue

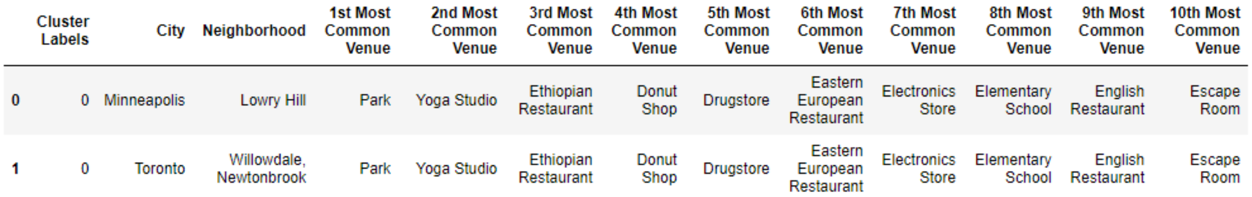
Table 4: Neigh_Finder Results for three neighborhoods of the dataset

Table 4 shows that each of the neighborhoods has very similar neighborhoods and can be proven by the top ten most common venues.

Discussion

The results show that certain neighborhoods between the cities have very similar venues. A cool observation shown through the maps is that central cities tend to share the same cluster. Sharing the same cluster makes sense because city centers tend to have a large amount of venues and can overlap due to their small size. The opposite is true as well for the outskirts of the cities. The outskirts tend to have less venues and of those venues, they tend to be very similar, i.e., parks, ball fields, and schools.

The cosine similarity portion also showed a cool observation between the neighborhoods of Lowry Hill, Minneapolis and Willowdale, Newtonbrook, Toronto. As shown in the Cluster 0 top five similarities these two neighborhoods have exactly the same venues. Table 5 also shows this connection. All of the similar venues make a decent amount of sense to be there, but one that stuck out was the Ethiopian Restaurants. With some knowledge of Minneapolis, there is a great amount of Eastern African influence due to Somoli refugees, but it is still interesting that they both feature these types of restaurants.



	Cluster Labels	City	Neighborhood	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	0	Minneapolis	Lowry Hill	Park	Yoga Studio	Ethiopian Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	Electronics Store	Elementary School	English Restaurant	Escape Room
1	0	Toronto	Willowdale, Newtonbrook	Park	Yoga Studio	Ethiopian Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	Electronics Store	Elementary School	English Restaurant	Escape Room

Table 5: Table of the top ten venues for two neighborhoods.

The last cool observation found was to see the different suggested neighborhoods from the find_neigh function. Often, looking through each of the neighborhoods, showed that many of the neighborhoods had similar most common venues, but with some slightly out of order.

Conclusion

In this project, Foursquare venue category data was analyzed and similar neighborhoods were found. The goal for the project was to find similar neighborhoods in different cities. This would allow for potential homebuyers or renters to find similar neighborhoods in their new city and help to ease the transition. With the final neigh_finder function, the similar neighborhoods and their common venue categories can be shown to the interested parties. The solution is definitely a good first step in the search for a potential new home. Improvements that could be made are comparisons between housing costs, crime rates, venue ratings, or many others. All of these comparisons are very important to any home buyer, and may potentially have a greater impact depending on the severity. In the end, the final deliverable for this project completed its goal and provides a good overview of Toronto and the Twin Cities for potential home buyers and renters.