Tannya Cai

Data Science Capstone Project

The Battle of the Neighborhoods

Summary

- The goal of this project was to examine potential areas of investment for a hotel in Toronto, taking into consideration the number of hotels in the neighborhood and the diversity of top venues in the region.
- The areas with the least number of hotels are: Bathurst Manor, Wilson Heights, Downsview North; Caledonia-Fairbanks; Davisville; High Park, The Junction South; Wexford, Maryvale, all with only 1 hotel category venue.
- The areas where there appears to be the most diversity in venues appear to be neighborhoods in Cluster 2 (solid blue), which consist of most neighborhoods.
- Top venues in Cluster 2 neighborhoods include coffee shops, restaurants, and many other attractions. Cluster 2 appears to be convenient spots for tourists (and thus hotels).
- Thus, investments in hotels in Toronto would be the overlapping areas with the least number of hotels, and those in Cluster 2.
- Final areas of potential hotel investment: Bathurst Manor, Wilson Heights, Downsview North; Davisville; High Park, The Junction South; and Wexford and Maryvale

Introduction and Business Problem

With nearly 3 million people, Greater Toronto Area (GTA) is a prospering area in Ontario, Canada where over 27 million tourists and students alike go to visit every year. As the area continues to expand its growing healthcare and finance presence, its academic and research institutions, and lively city life, there will be increasing interest from investors and developers in understanding the distribution of venues in the region, and how that can impact opportunities for new buildings and businesses, especially hotels for visiting travelers, in the area.

The goal here is to identify where the most optimal neighborhoods to buy or build a hotel, with the intention of being in an area with the least number of hotels as possible, and overlaying areas with the number of venues and attractions in the region.

Intended audience: This analysis is intended for investors and developers who are interested in gaining a deeper view of where optimal locations for might be starting up a hotel business, especially in the context of surrounding venues and attractions.

Data

Data sources

The data used will come from the following sources:

- Neighborhood and borough data: from Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Data on common venues: from Foursquare API
- Longitude and latitude data: from Geocoder Python package

The neighborhood and borough data will be used to determine the various neighborhoods in Toronto, with longitudinal and latitudinal data used to geographically locate the boroughs.

Subsequently, data on the various venues and attractions within each neighborhood can be pulled from the Foursquare API to build an understanding of the distribution of venues within each neighborhood. Additionally, hotel data can be pulled from the Foursquare API as well to understand distribution of hotels around the GTA.

Data cleaning

After scraping the Wikipedia list for postal codes and neighborhoods in the Toronto area, the rows in which there are no boroughs are dropped as they do not belong in any neighborhood, and thus are not useful in the analysis. Additionally, any “not assigned” neighborhoods were assigned the values of its corresponding borough to include them in the analysis. This gives a result of 103 rows and 3 columns with the ‘Postal Code’, ‘Borough’, and ‘Neighborhood’.

Subsequently, the geospatial data (longitude, latitude) was pulled from https://cocl.us/Geospatial_data and then joined to the previous dataframe with the neighborhood and borough data, using “Postal Code” as the ID.

This data is the basis of the neighborhood information for the subsequent analysis.

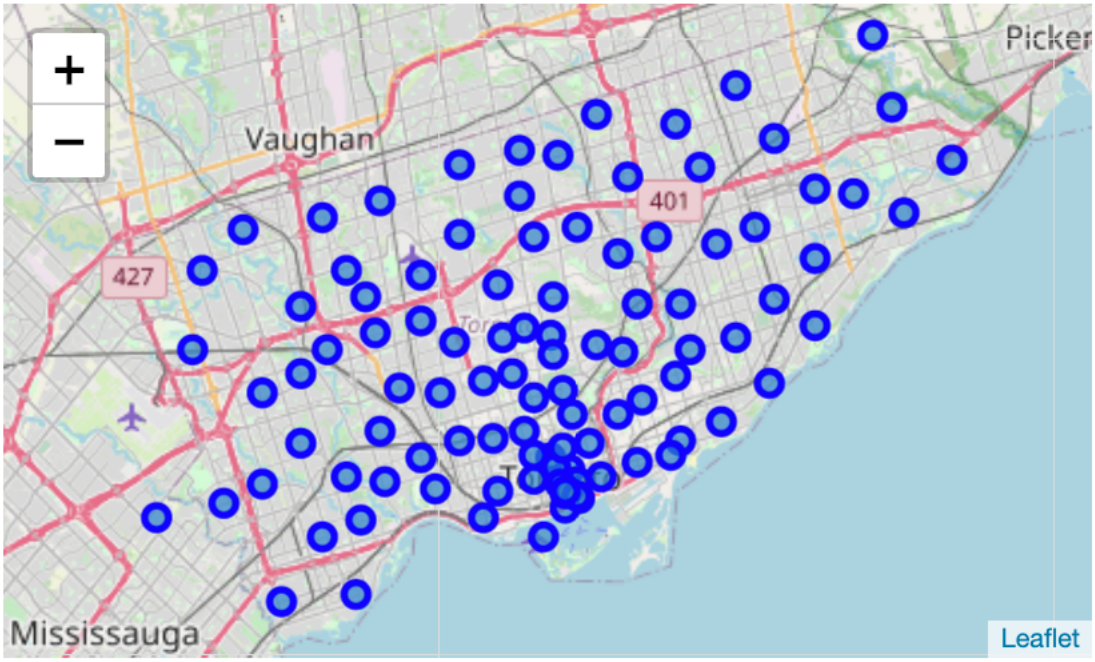
Methodology

Exploratory data analysis and using Foursquare

- Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any

inferential statistical testing that you performed, if any, and what machine learnings were used and why.

First, the boroughs and neighborhoods of Toronto were mapped using Folium to allow for better visualization.



Then, using Foursquare's API, a random neighborhood (in this case, the neighborhood fifth on the list, Queen's Park, Ontario Provincial Government) was chosen and examined in more detail. Its longitudinal and latitudinal data was pulled, and a search for the top 100 venues in a radius of 600m around that area was sent to Foursquare. This was done using the explore function. 57 venues were returned by Foursquare. This was repeated for all neighborhoods in Toronto, with the resulting dataframe having 2131 venues and 7 columns.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Parkwoods	43.753259	-79.329656	TTC stop - 44 Valley Woods	43.755402	-79.333741	Bus Stop
3	Parkwoods	43.753259	-79.329656	Bella Vita Catering & Private Chef Service	43.756651	-79.331524	BBQ Joint
4	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena

Examining this by neighborhood, we see 96 neighborhoods as below.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Agincourt	4	4	4	4	4	4
Alderwood, Long Branch	8	8	8	8	8	8
Bathurst Manor, Wilson Heights, Downsview North	21	21	21	21	21	21
Bayview Village	4	4	4	4	4	4
Bedford Park, Lawrence Manor East	25	25	25	25	25	25
...	...	...	...	...	...	...
Willowdale, Willowdale West	6	6	6	6	6	6
Woburn	4	4	4	4	4	4
Woodbine Heights	7	7	7	7	7	7
York Mills West	4	4	4	4	4	4
York Mills, Silver Hills	1	1	1	1	1	1

96 rows × 6 columns

Each neighborhood was then analyzed using one hot encoding to then determine the most common 10 venues in each area. The goal of this is to give some indication into which areas have the most diversity in type of venues, or which have venues that

would be most of interest to tourists (e.g., restaurants, attractions, etc.). This would make those areas good opportunities for hotel investment.

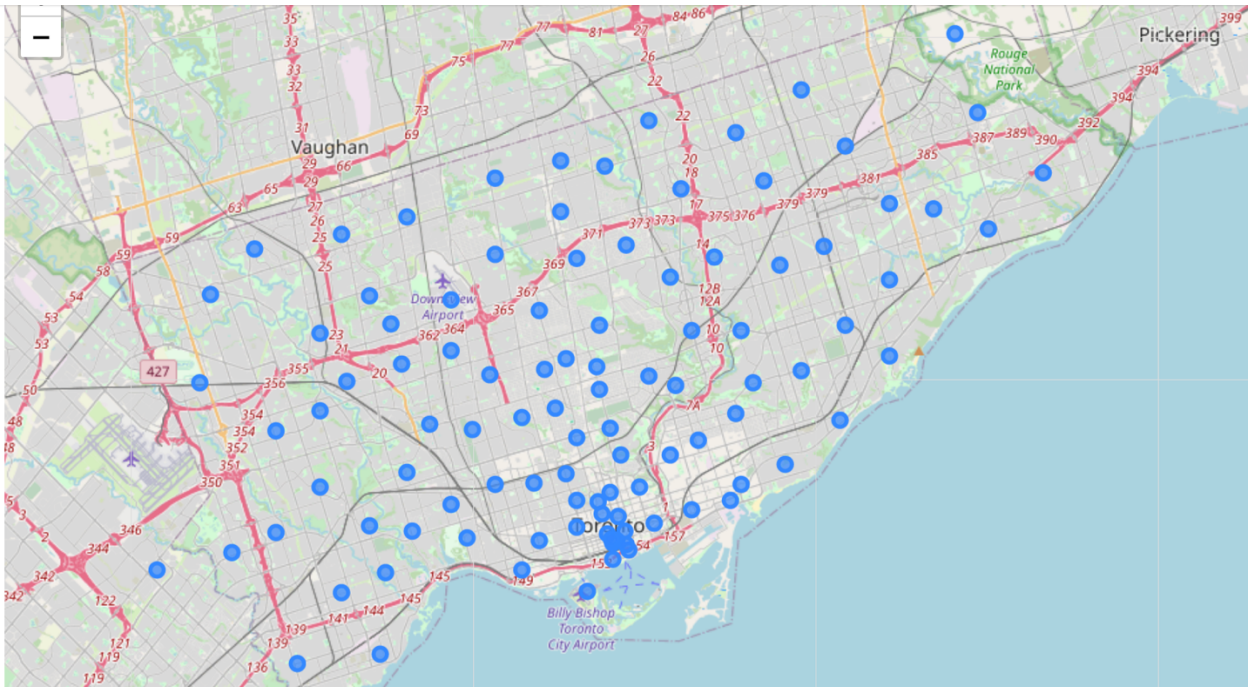
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agincourt	Lounge	Skating Rink	Breakfast Spot	Latin American Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Drugstore	Ethiopian Restaurant	Donut Shop
1	Alderwood, Long Branch	Pizza Place	Gym	Coffee Shop	Skating Rink	Pharmacy	Pub	Sandwich Place	Dim Sum Restaurant	Deli / Bodega	Department Store
2	Bathurst Manor, Wilson Heights, Downsview North	Coffee Shop	Bank	Middle Eastern Restaurant	Pizza Place	Supermarket	Sushi Restaurant	Deli / Bodega	Shopping Mall	Restaurant	Mobile Phone Shop
3	Bayview Village	Japanese Restaurant	Café	Bank	Chinese Restaurant	Dessert Shop	Diner	Discount Store	Distribution Center	Dog Run	Women's Store
4	Bedford Park, Lawrence Manor East	Restaurant	Coffee Shop	Italian Restaurant	Sandwich Place	Juice Bar	Butcher	Sushi Restaurant	Pizza Place	Pharmacy	Indian Restaurant

K-means clustering and machine learning

The next step involves clustering the neighborhoods using k-means to determine if there was some way of grouping the neighborhoods by type of venues to facilitate the search for an area suitable for hotel investment. K-means is a common unsupervised learning method of clustering data points. A k of 5 was chosen as it was the “kink” identified the elbow method, making it the optimal k value.

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd M Comr Ve
0	M3A	North York	Parkwoods	43.753259	-79.329656	2.0	Park	BBQ Joint	Foc Drink S
1	M4A	North York	Victoria Village	43.725882	-79.315572	2.0	Portuguese Restaurant	Financial or Legal Service	Intersec
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	2.0	Coffee Shop	Park	Bak
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	2.0	Clothing Store	Furniture / Home Store	Accesso S
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	2.0	Coffee Shop	Diner	Distribu Ce

Subsequently, the neighborhoods with their respective clusters are visualized using Folium maps. The different clusters are demonstrated with varying shades of blue and level of circle color filling.



Finally, each cluster was examined individually. It can be seen that the first cluster (label = 0), with 4 neighborhoods, largely consists of neighborhoods where pizza places are the 1st most common venues. Cluster label = 1 consists mostly of areas where the 1st most common venue is a park, and has 7 neighborhoods total. For Cluster label = 2, the largest cluster with 86 neighborhoods, the top most common venues are restaurants of various sorts, along with a diversity of other venues. Cluster label = 3 has only 1 neighborhood with the 1st most common venue being martial arts schools. Cluster label = 4 has 2 neighborhoods, with baseball field as the most common venue.

Hotel analysis

Now that general venues have been analysed, the number of hotels in each area is also examined. Using the “categoryId” term, only the hotels within radius = 500m of the set neighborhood longitude/latitude points were returned, with a limit of 100. This was first tested with one neighborhood, and then performed for every neighborhood. The results are shown below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.654260	-79.360636	Residence & Conference Centre	43.653040	-79.357040	Hotel
1	Regent Park, Harbourfront	43.654260	-79.360636	Locarno Hostel	43.658627	-79.365400	Hostel
2	Regent Park, Harbourfront	43.654260	-79.360636	Corktown Cottages	43.654332	-79.358005	Bed & Breakfast
3	Regent Park, Harbourfront	43.654260	-79.360636	Red Lion Inn	43.652245	-79.363126	Hotel
4	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	Toronto	43.660533	-79.387507	Vacation Rental

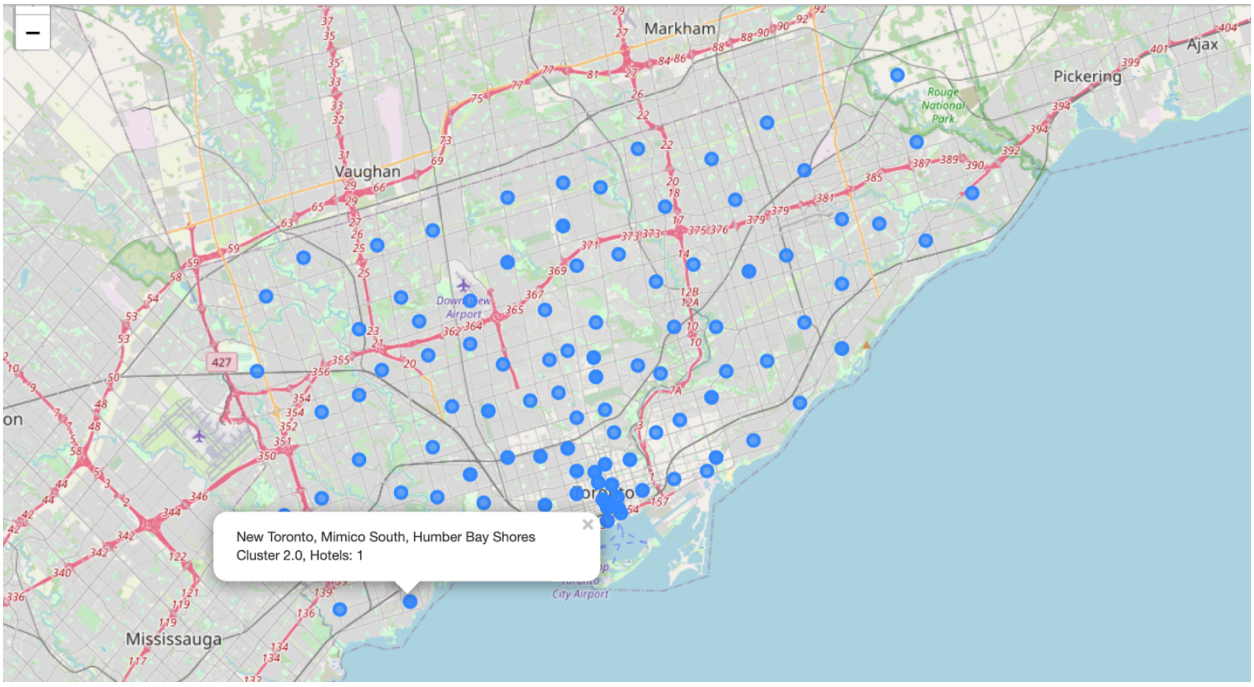
This was then grouped by neighborhood using groupby, and then cleaned to only have one column with “Neighborhood”, and one column with the “Number of venues with category of hotels”.

Neighborhood	Number of Category: Hotels
Bathurst Manor, Wilson Heights, Downsview North	1
Berczy Park	8
Brockton, Parkdale Village, Exhibition Place	2
Caledonia-Fairbanks	1
Canada Post Gateway Processing Centre	4
Central Bay Street	21
Christie	2
Church and Wellesley	21
Cliffside, Cliffcrest, Scarborough Village West	2
Commerce Court, Victoria Hotel	37
Davisville	1
Davisville North	2
Dufferin, Dovercourt Village	3
First Canadian Place, Underground city	41
Garden District, Ryerson	41

Now, we can merge this data with the information on venues clustering and neighborhood areas, and map it using Folium maps. The final dataframe is below.

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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	Hotels
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	2.0	Coffee Shop	Park	Bakery	Pub	Breakfast Spot	Café	Theater	Dessert Shop	Ice Cream Shop	Beer Store	4
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	2.0	Coffee Shop	Diner	Distribution Center	Sandwich Place	Portuguese Restaurant	Persian Restaurant	Park	Mexican Restaurant	Japanese Restaurant	Hobby Shop	10
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	2.0	Clothing Store	Coffee Shop	Café	Bubble Tea Shop	Japanese Restaurant	Italian Restaurant	Cosmetics Shop	Pizza Place	Theater	Lingerie Store	41
3	M4C	East York	Woodbine Heights	43.695344	-79.318389	2.0	Skating Rink	Park	Pharmacy	Beer Store	Bus Stop	Curling Ice	Colombian Restaurant	Department Store	College Gym	College Rec Center	1
4	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	2.0	Coffee Shop	Café	Clothing Store	Restaurant	Cocktail Bar	American Restaurant	Cosmetics Shop	Gym	Breakfast Spot	Park	27
5	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	2.0	Coffee Shop	Café	Bakery	Cocktail Bar	Beer Bar	Farmers Market	Restaurant	Cheese Shop	Seafood Restaurant	Gym	8
6	M6E	York	Caledonia-Fairbanks	43.689026	-79.453512	1.0	Park	Pool	Women's Store	Golf Course	Electronics Store	Dumpling Restaurant	Drugstore	Donut Shop	Doner Restaurant	Dog Run	1
7	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	2.0	Coffee Shop	Sandwich Place	Café	Italian Restaurant	Salad Place	Bubble Tea Shop	Burger Joint	Department Store	Japanese Restaurant	Miscellaneous Shop	21
8	M6G	Downtown Toronto	Christie	43.669542	-79.422564	2.0	Grocery Store	Café	Park	Candy Store	Italian Restaurant	Diner	Baby Store	Restaurant	Nightclub	Coffee Shop	2
9	M3H	North York	Bathurst Manor, Wilson Heights, Downsview North	43.754328	-79.442259	2.0	Coffee Shop	Bank	Middle Eastern Restaurant	Pizza Place	Supermarket	Sushi Restaurant	Deli / Bodega	Shopping Mall	Restaurant	Mobile Phone Shop	1
10	M5H	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568	2.0	Coffee Shop	Café	Gym	Hotel	Clothing Store	Restaurant	Thai Restaurant	Bar	Steakhouse	Pizza Place	38

Finally, we can add the hotel information to the popup of the map. The below map shows different clusters differentiated by their fill colors and shade of blue. The popup includes information on the name of the neighborhood, the cluster label (e.g., 2.0 here), and the number of hotels in the region, if any (here, hotels: 1, which means there is one main hotel here).



Results

From above, we can see the resulting dataframes and interactive maps generated. The final dataframe and map includes information on the neighborhoods, venues, their respective longitudes and latitudes, their cluster labels, the most common venues in the area, and the number of hotels in the region.

From the analyses above, it is clear that the areas with the greatest number of hotels are the following:

- Garden District, Ryerson, 41 hotels
- First Canadian Place, Underground City, 41 hotels
- Richmond, Adelaide, King, 38 hotels
- Commerce Court, Victoria Hotel, 37 hotels
- Toronto Dominion Centre, Design Exchange, 35 hotels

The areas with the least number of hotels are the following:

- Bathurst Manor, Wilson Heights, Downsview North, 1 hotel
- Caledonia-Fairbanks, 1 hotel
- Davisville, 1 hotel
- High Park, The Junction South, 1 hotel
- Wexford, Maryvale, 1 hotel

As determined above, the areas where there appears to be the most diversity in venues appear to be neighborhoods in Cluster 2 (solid blue), which consist of most of the neighborhoods in Toronto. Top venues in Cluster 2 neighborhoods include coffee shops, restaurants, and many other attractions, and thus areas falling under the cluster label 2 would be good spots for tourists to visit (and thus a good spot to build a hotel).

For areas where “least number of hotels” overlaps with “cluster label 2”, this would include the following:

- Bathurst Manor, Wilson Heights, Downsview North, 1 hotel (cluster label 2)
- Davisville, 1 hotel (cluster label 2)
- High Park, The Junction South, 1 hotel (cluster label 2)
- Wexford, Maryvale, 1 hotel (cluster label 2)

Discussion & Recommendations

As determined above, the areas where there appears to be the most diversity in top venues appear to be neighborhoods in Cluster 2 (solid blue), which consist of most of the neighborhoods in Toronto—this is good news for investors as it means there are many neighborhoods in which tourists would be eager to visit, and thus many

opportunities for investment. Investors would thus want to focus on Cluster 2 in idneitfying locations for investment.

When overlapping this with the areas with the least number of hotels, we see that Bathurst Manor, Wilson Heights, Downsview North; Davisville; High Park, The Junction South; and Wexford and Maryvale all in cluster label 2 and do not have many hotels, and thus little competition. This suggests that should a hotel be constructed in these regions, it may present an opportunity for strong returns as tourists would be happy to visits that neighborhood and would likely look for hotels in the area. Since there are not many hotels, the hotel the investor constructs would get that customer.

Future directions

Of course, there are still a lot of other factors that come into play when looking at a commercial real estate investment. Real estate in Toronto is notoriously expensive, and it would be wise to conduct additional analysis examining the amount of investment required to build a hotel, and whether that is financially sound. It would also be important to consider the rise of Airbnbs, and consider whether Airbnb listings in the region have essentially replaced hotels, or if there's still opportunity for more accommodations. Finally, it is possible that areas that have few hotels but high numbers of venues are either limited in the height of buildings that can be constructed (e.g., can't construct hotel with lots of floors), or are relatively unknown by travellers, both of which would affect investment outcomes. These are all areas of further analysis that would have to be conducted to ensure a sound investment.

Conclusion

In this report, the goal was to determine potential neighborhoods where a hotel investment would be wise. Specifically, the diversity and type of venues around an area that would appeal to tourists, as well as the number of hotels, were used as a proxy for a good investment. By leveraging machine learning and Foursquare's API, additional understanding was gained around which neighborhoods have diverse venues and thus might be better suited for visitors, and the distribution of hotels around Toronto so a hotel is not built in an already hotel-crowded region. From this analysis, it was determined that *Bathurst Manor, Wilson Heights, Downsview North; Davisville; High Park, The Junction South; and Wexford and Maryvale* were all strong candidates for investment, as both display strong diversity of venues, restaurants, and attractions in their top venues, while also having a relatively low number of hotels in the region. Future areas of analysis include examining financial returns on investment, considering Airbnb as a competitor, and regulatory challenges and traveller knowledge of those regions.