Predicting the success of a venue in Toronto.

Mario Andre Molina Caballero, data analyst.
mario.andre.molina@gmail.com

Applied data science capstone project.
IBM data science professional certificate.

August 19, 2020

1. Introduction

Over the course of the last six months working in this certification, we have reviewed different topics in data science fields, including methodologies, tools and data analysis, modelling, evaluation and visualization; lastly, we explored how to use the Foursquare API in order to query its comprehensive venues database. The purpose of this project is to apply all this acquired knowledge in order to 'attempt' to solve a real-world problem, I have decided to try to predict the success of a venue in Toronto, using the venue rating as the parameter for success.

Data science is an interdisciplinary field that combines knowledge from well-known and developed disciplines such as statistics, probability and mathematics with the capabilities that modern tools and software provide for us, including data mining, data bases and higher computing power and storage. All of this allows us to manipulate huge amounts of often unstructured data in order to identify trends, obtain previously hidden insights and even predicting future outcomes.
It is the job of the data scientist to obtain value from the data and communicate the findings to stakeholders in the project in order to allow them to make better decisions or decide a new strategy. In order to correctly convey the most important information obtained, a data scientist must build a narrative and employ data visualization techniques such as charts, maps and plots.

In the last course of the certification, we learned about Foursquare and how to query its database using their publicly available API. Foursquare is the leading independent location technology platform, its core product is the 'Places Database' which has information for more than 60 million points of interest including restaurants, shops and services across more than 190 countries. The information provided for each of these points of interest includes the venue name, category, location, rating and even user generated data such as photos and tips. All of this data is available through the 'Places API', after generating the credentials you can send API requests to the different Foursquare endpoints, each of them providing a different response.

Finally, I wanted to apply all of this knowledge to try to predict the rating a venue in Toronto would obtain given the location and its category. Toronto is the most populous city in Canada, with 2.7 million habitants as of 2016, a truly multicultural and cosmopolitan city. Given its importance as a cultural and economic hotspot in Canada and the world, it seemed interesting

to try and predict the success of a new business in this city, for it will surely be a challenging enterprise for any entrepreneur.

We will be using a Jupyter notebook running python hosted in the 'Skill Networks Lab' by IBM.

Some of the libraries that we will be using in our python script are:

- BeautifulSoup: for web scraping.
- Requests: for http request handling.
- Numpy: numerical computing tools.
- Pandas: data analysis and manipulation tools.
- IPython.display: for displaying images and web content.
- Geopy: python client that enables coordinate location.
- Matplotlib: for visualizations.
- Sklearn: machine learning tools.
- Folium: for map visualization.
- Seaborn: for visualization.

2. Data

The first thing was to obtain the list of neighborhoods in Toronto, with the 'BeautifulSoup' library, we could send a request directly to Wikipedia to read the following table into the Jupyter notebook:

Postal Code	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
M8A	Not assigned	Not assigned

Figure 2.1: Wikipedia list for Canadian postal codes.

After some manipulation, we obtained the following pandas dataframe:

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Figure 2.2: First dataframe obtained

The next step was to obtain the coordinates for the postal codes obtained, downloaded from the csv file: https://cocl.us/Geospatial_data, after merging the data, this was the dataframe obtained:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Figure 2.3: Dataframe with neighborhood and coordinates data.

The next step was to get venue information for Toronto, so using the Foursquare API, we queried their Places Database in order to obtain at most 100 venues for each neighborhood, after cleaning the response and merging with our previous dataframe, we obtained this new one, including venue name, ID, coordinates and category:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Id	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	4e8d9dcdd5fbbbb6b3003c7b	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	4cb11e2075ebb60cd1c4caad	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	4c633acb86b6be9a61268e34	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Portugril	4f3ecce6e4b0587016b6f30d	43.725819	-79.312785	Portuguese Restaurant
4	Victoria Village	43.725882	-79.315572	Tim Hortons	4bbe904a85fbb713420d7167	43.725517	-79.313103	Coffee Shop

Figure 2.4: Dataframe with venue data.

It's important to mention that this request made to the places database was pointing to the 'explore' endpoint, a 'regular' call, which means we get to do 99,500 of such calls each day.

In order to get the ratings for the venues retrieved, we had to query the 'details' endpoint, this is a 'premium' call and as such we only get 500 requests a day. From the previous steps we had obtained venue data for about 2,100 venues in Toronto, so in order to get the ratings for each of them, we had to split the full dataframe in 5 smaller ones and query for the ratings 500 venues a day, and at the end, concatenate the data back together.

Finally, we ended up with this data set of venues for Toronto:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Id	Venue Latitude	Venue Longitude	Venue Category	Rating
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	4e8d9dcdd5fbbbb6b3003c7b	43.751976	-79.332140	Park	6.9
1	Parkwoods	43.753259	-79.329656	Variety Store	4cb11e2075ebb60cd1c4caad	43.751974	-79.333114	Food & Drink Shop	No rating received for this venue
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	4c633acb86b6be9a61268e34	43.723481	-79.315635	Hockey Arena	7.3

Figure 2.5: Dataframe with full venue data, including rating.

We have 2,384 rows and 11 columns in our dataframe:

- 2384 venues.
- 96 neighborhoods.
- 264 categories.

Our target variable is venue rating, and we will be using the neighborhood and venue category as features in order to predict the rating of a new venue. As we can see we have 264 categories, I think the algorithm will have a hard time taking so many different values into account for the prediction, so we will reduce this range of options by binning similar categories together as following:

- Bin 1: coffee and breakfast spots.
  Including:

Coffee Shop	Breakfast Spot	Café	Smoothie Shop
Bubble Tea Shop	tea Room	Juice Bar	Bagel Shop
Donut Shop	Frozen Yogurt Shop	Cafeteria	

Table 2.1: Bin 1.

- Bin 2: sweets and dessert spots.
  Including:

Bakery	Chocolate Shop	Dessert Shop	Ice Cream Shop
Creperie	Cupcake Shop		

Table 2.2: Bin 2.

- Bin 3: Fast food restaurants.
  Including:

Pizza Place	Burrito Place	Fried Chickern Joint	Sandwich Place
Fast Food Restaurant	Burger Joint	Taco Place	Food Truck
Salad Place	Food Court	Snack Place	

Table 2.3: Bin 3.

- Bin 4: European cuisine restaurants.
  Including:

Portuguese Restaurant	French Restaurant	Mediterranean Restaurant	Italian Restaurant
Modern European Restaurant	Greek Restaurant	German Restaurant	Belgian Restaurant
Eastern European Restaurant			

Table 2.4: Bin 4.

- Bin 5: Asian cuisine restaurants.
  Including:

Vietnamese Restaurant	Sushi Restaurant	Japanese Restaurant	Asian Restaurant
Ramen Restaurant	Thai Restaurant	Chinese Restaurant	Dim Sum Restaurant
Poke Place	Indian Restaurant	Korean Restaurant	Hakka Restaurant
Noodle House	Indonesian Restaurant	Filipino Restaurant	Dumpling Restaurant
Taiwanese Restaurant	Afghan Restaurant		

Table 2.5: Bin 5.

- Bin 6: Latin cuisine restaurants.
  Including:

Mexican Restaurant	Caribbean Restaurant	Colombian Restaurant	Brazilian Restaurant
Latin American Restaurant	Cuban Restaurant		

Table 2.6: Bin 6.

- Bin 7: Middle eastern and African cuisine restaurants.
  Including:

Middle Eastern Restaurant	Ethiopian Restaurant	Moroccan Restaurant	Falafel Restaurant
Doner Restaurant			

Table 2.7: Bin 7.

- Bin 8: General food restaurants.
  Including:

Food & Drink Shop	Restaurant	Seafood Restaurant	Diner
Steakhouse	New American Restaurant	Poutine Place	BBQ Joint
American Restaurant	Vegetarian / Vegan Restaurant	Comfort Food Restaurant	Bistro
Gluten-free Restaurant	Cajun / Creole Restaurant	Molecular Gastronomy Restaurant	Soup Place
Theme Restaurant	Wings Joint		

Table 2.8: Bin 8.

- Bin 9: Grocery stores and markets.
  Including:

Farmers Market	Beer Store	Health Food Store	Wine Shop
Supermarket	Grocery Store	Market	Convenience Store
Liquor Store	Fish Market	Cheese Shop	Gourmet Shop
Fish & Chips Shop	Candy Store	Deli / Bodega	Fruit & Vegetable Store
Stationery Store	Butcher	Organic Grocery	

Table 2.9: Bin 9.

- Bin 10: Entertainment and activities.
  Including:

Jazz Club	Aquarium	Movie Theater	Indie Movie Theater
Gaming Cafe	Roof Deck	General Entertainment	Theater
Music Venue	Performing Arts Venue	Opera House	

Table 2.10: Bin 10.

- Bin 11: Nightlife, clubs and bars.
  Including:

Hookah Bar	Wine Bar	Lounge	Beer Bar
Irish Pub	Gay Bar	Nightclub	Speakeasy
Bar	Gastropub	Brewery	Pub
Cocktail Bar	Sports Bar	Hotel Bar	Sake Bar
Strip Club			

Table 2.11: Bin 11.

- Bin 12: Health and beauty services.
  Including:

Spa	Cosmetics Shop	Pharmacy	Tanning Salon
Medical Center	Drugstore	Health & Beauty Service	Supplement Shop

Table 2.12: Bin 12.

- Bin 13: Nature and outdoors.
  Including:

Park	Lake	Other Great Outdoors	River
Field	Trail	Playground	Beach
Garden	Scenic Lookout		

Table 2.13: Bin 13.

- Bin 14: Sport activities.
  Including:

Gym / Fitness Center	Yoga Studio	Gym	Baseball Field
Soccer Field	Golf Course	Dog Run	Skating Rink
Curling Ice	Pool	Climbing Gym	Swim School
Tennis Court	Dance Studio	Martial Arts Dojo	Skate Park

Table 2.14: Bin 14.

- Bin 15: Sport and event venues.
  Including:

Hockey Arena	Basketball Stadium	Baseball Stadium	Stadium
Convention Center	Building	Event Space	Concert Hall

Table 2.15: Bin 15.

- Bin 16: Landmarks, galleries, museums.
  Including:

Historic Site	Art Gallery	Art Museum	Fountain
Museum	Monument / Landmark	History Museum	Sculpture Garden
Garden Center			

Table 2.16: Bin 16.

- Bin 17: Department stores.
  Including:

Shoe Store	Electronics Store	Furniture / Home Store	Clothing Store
Hobby Shop	"Antique Shop	Boutique	Accessories Store
Women's Store	Athletics & Sports	Miscellaneous Shop	Arts & Crafts Store
Shopping Mall	Department Store	Bookstore	Pet Store
Comic Shop	Sporting Goods Shop	Toy / Game Store	Lingerie Store
Jewelry Store	Discount Store	Bike Shop	Camera Store
Baby Store	Bridal Shop	Mobile Phone Shop	Warehouse Store
Smoke Shop	Gift Shop	Video Game Store	Luggage Store
Record Shop	Men's Store	Flea Market	Thrift / Vintage Store
Optical Shop	Hardware Store		

Table 2.17: Bin 17.

- Bin 18: Transportation services.
  Including:

Intersection	Metro Station	Rental Car Location	Gas Station
General Travel	Bus Station	Train Station	Bus Line
Light Rail Station	Auto Garage	Harbor / Marina	Boat or Ferry

Table 2.18: Bin 18.

- Bin 19: General services and businesses.
  Including:

Distribution Center	Bank	Hotel	Plaza
Office	Salon / Barbershop	Tailor Shop	Business Service
Motel	Massage Studio	IT Services	Construction & Landscaping
Coworking Space	Lawyer	Hospital	Church
Bed & Breakfast	Auto Workshop	Neighborhood	Recording Studio

Table 2.19: Bin 19.

- Bin 20: College buildings.
  Including:

College Cafeteria	College Rec Center	College Auditorium	College Stadium
College Arts Building	College Gym		

Table 2.20: Bin 20.

- Bin 21: Airport and its services.
  Including:

Airport Service	Airport	Plane	Airport Terminal
Airport Food Court	Airport Lounge		

Table 2.21: Bin 21.

As we can see in Figure 2.5, some of the venues in our sample don't have a rating, we will deal with this by replacing the missing values for the mode of the corresponding category bin, by selecting the most common value instead of the average, we avoid averaging the worst and the best rating values for any unrated venue, as both cases would be undeserving.
Finally, we drop repeated venues to have our final dataset that we will be working with:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Id	Venue Latitude	Venue Longitude	Venue Category	Rating	Category Bin
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	4e8d9dcdd5fbbbb6b3003c7b	43.751976	-79.332140	Park	6.9	Bin 13
1	Parkwoods	43.753259	-79.329656	Variety Store	4cb11e2075ebb60cd1c4caad	43.751974	-79.333114	Food & Drink Shop	7.9	Bin 8
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	4c633acb86b6be9a61268e34	43.723481	-79.315635	Hockey Arena	7.3	Bin 15

Figure 2.6: Final dataframe with full venue data, including rating and category bins.

We have 1711 rows and 10 columns:

- There are 1711 venues.
- There are 264 unique categories.
- There are 96 neighborhoods.
- There are 21 category bins.

To get some quick insights about our data set, we obtained the following box plots, relating rating against different parameters:

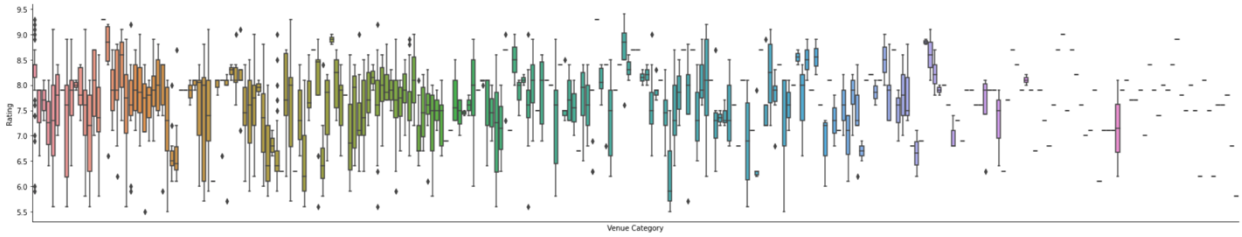

Figure 2.7: Plot for rating by venue category.

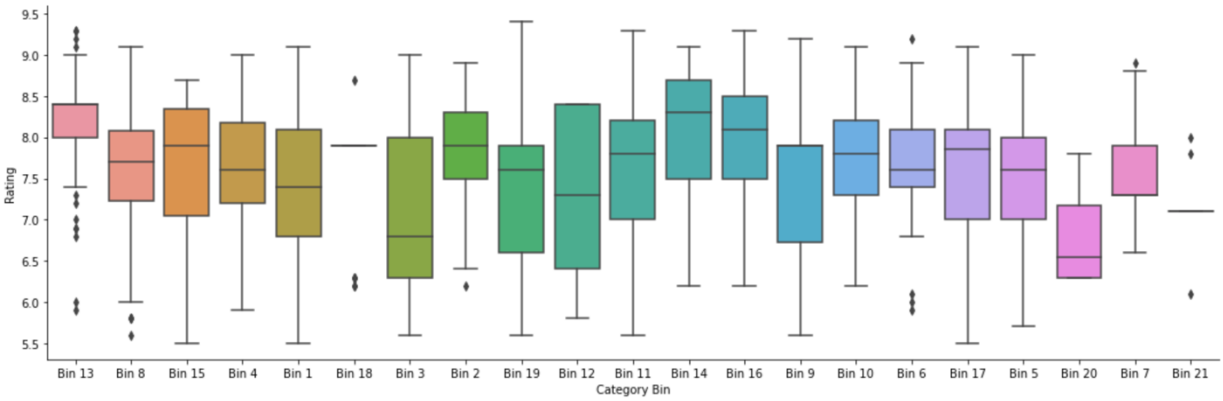

Figure 2.8: Plot for rating by category bin.

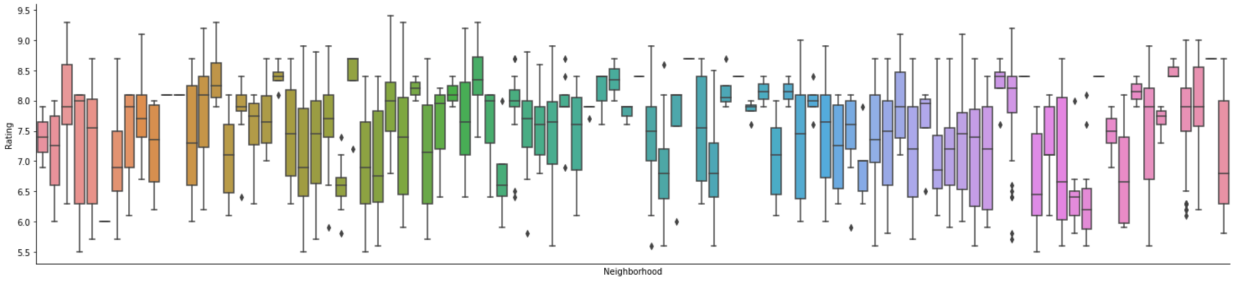

Figure 2.9: Plot for rating by category neighborhood.

3. Methodology

As we can see above, we have 96 different neighborhoods, as we did with the venue category, we will try to bin the neighborhoods together based on their similarity, to reduce the wide range of options we have. In order to do this we will use the K-means clustering technique, taking the top three most common category bins in each neighborhood as features.

First, we will obtain the top three most common category bins for each neighborhood as shown:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Agincourt	Bin 1	Bin 6	Bin 11
1	Alderwood, Long Branch	Bin 3	Bin 1	Bin 11
2	Bathurst Manor, Wilson Heights, Downsview North	Bin 9	Bin 3	Bin 17
3	Bayview Village	Bin 5	Bin 19	Bin 1
4	Bedford Park, Lawrence Manor East	Bin 1	Bin 8	Bin 3

Figure 3.1: Dataframe showing the three most common category bin for each neighborhood.

Using the most common bins in each neighborhood as features, we proceeded to cluster the neighborhoods using the k-means algorithm, with K = 6:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Id	Venue Latitude	Venue Longitude	Venue Category	Rating	Category Bin	Cluster Labels
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	4e8d9dcdd5fbbbb6b3003c7b	43.751976	-79.332140	Park	6.9	Bin 13	0
1	Parkwoods	43.753259	-79.329656	Variety Store	4cb11e2075ebb60cd1c4caad	43.751974	-79.333114	Food & Drink Shop	7.9	Bin 8	0
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	4c633acb86b6be9a61268e34	43.723481	-79.315635	Hockey Arena	7.3	Bin 15	1
3	Victoria Village	43.725882	-79.315572	Portugril	4f3ecce6e4b0587016b6f30d	43.725819	-79.312785	Portuguese Restaurant	6.4	Bin 4	1
4	Victoria Village	43.725882	-79.315572	Tim Hortons	4bbe904a85fbb713420d7167	43.725517	-79.313103	Coffee Shop	6.0	Bin 1	1

Figure 3.2: Dataframe showing cluster labels.

After we had obtained our clusters of neighborhoods, the idea was to try and define a prediction function for the venue rating in each of the clusters, the theory was that each type of neighborhood would value different type of venues more or less than the others, so a residential neighborhood would prefer grocery stores over nightclubs, and would rate the former higher than the later, for example.

4. Results

Now that the neighborhoods are clustered by similarity, we can explore each cluster and their relation to the venues in its neighborhoods.

Cluster	Number of venues
1	35
2	216
3	6
4	23
5	15
6	1416

Table 4.1: Number of venues by cluster.

We can see that most of the venues are located in neighborhoods that landed in Cluster 6, meaning these neighborhoods are very similar to each other.

Information on each cluster:

- Cluster 1

Neighborhoods	11
Category bins	12
Venues	35
1st most common bin	Outdoors
2nd most common bin	Grocery stores, markets
3rd most common bin	Sport activities

Table 4.2: Data in Cluster 1.

- Cluster 2

Neighborhoods	22
Category bins	19
Venues	167
1st most common bin	Fast food
2nd most common bin	Grocery stores, markets
3rd most common bin	Coffee and breakfast

Table 4.3: Data in Cluster 2.

- Cluster 3

Neighborhoods	3
Category bins	2
Venues	6
1st most common bin	Sport activities
2nd most common bin	European cuisine restaurants
3rd most common bin	Grocery stores, markets

Table 4.4: Data in Cluster 3.

- Cluster 4

Neighborhoods	3
Category bins	5
Venues	21
1st most common bin	Department stores and shops
2nd most common bin	Coffee and breakfast
3rd most common bin	Asian cuisine restaurants

Table 4.5: Data in Cluster 4.

- Cluster 5

Neighborhoods	6
Category bins	3
Venues	15
1st most common bin	Outdoors
2nd most common bin	Sport activities
3rd most common bin	Sport and event venues

Table 4.6: Data in Cluster 5.

- Cluster 6

Neighborhoods	51
Category bins	21
Venues	1190
1st most common bin	Coffee and breakfast
2nd most common bin	Department stores and shops
3rd most common bin	General food & drink, restaurants

Table 4.7: Data in Cluster 6.

5. Discussion

Using the folium library, we have located the neighborhoods clustered together in the map of Toronto, these maps will allow us to obtain a better comprehension of the results. We will go cluster by cluster analyzing the results as well as making some recommendations on what type of businesses are most likely to thrive there.

For further reference, the full list of neighborhoods included in each cluster will be added at the end of this document, as Annex 1.

- Cluster 1

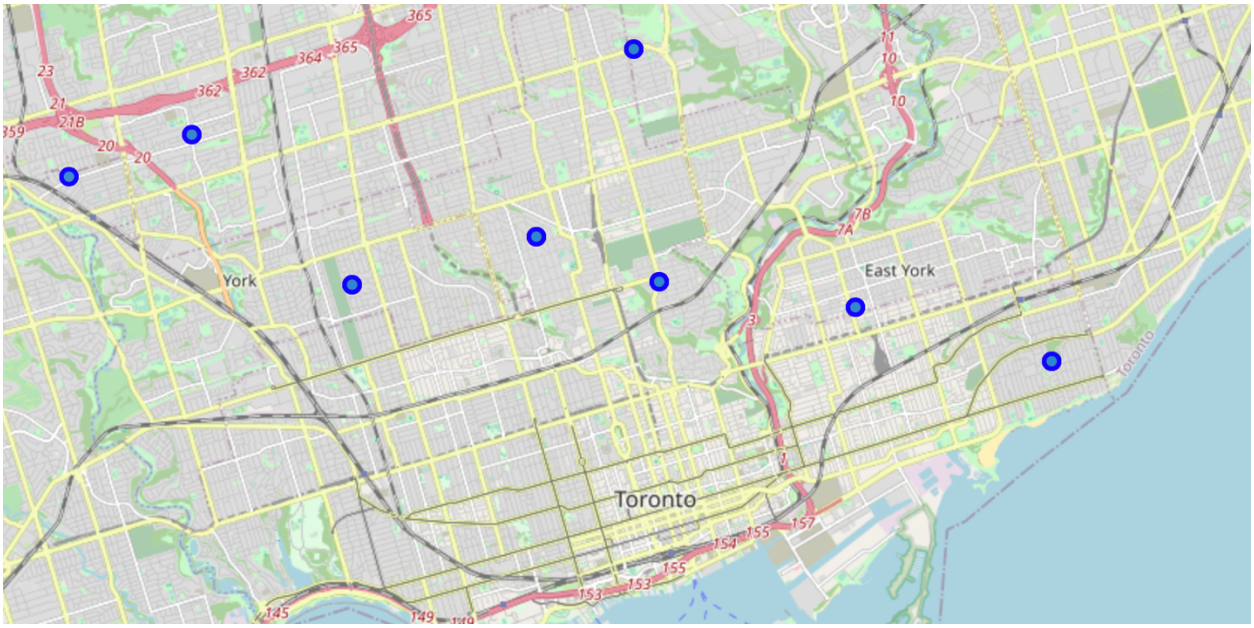

Figure 5.1: Map of neighborhoods in Cluster 1.

Based on the most common category bins for this cluster seen in table 4.2 and seeing the location of the neighborhoods clustered together in figure 5.1, we can describe the neighborhoods in cluster 1 as neighborhoods in the greater Toronto area, most likely residential neighborhoods such as suburbs with lots of natural spaces, places to practice sports and grocery stores.
We only have 35 venues registered for this cluster, and the most common type of places registered are actually nature and outdoor places such as parks, so in the case someone would try to place their business here, we would recommend for it to be either a grocery store or a gym. Due to the low number of venues for this clusters, we will refrain to try to predict the rating using any kind of machine learning algorithm, as the data sets are simply not big enough to produce any particularly significant results, this applies to all subsequent clusters except cluster 6, as we'll see.

- Cluster 2

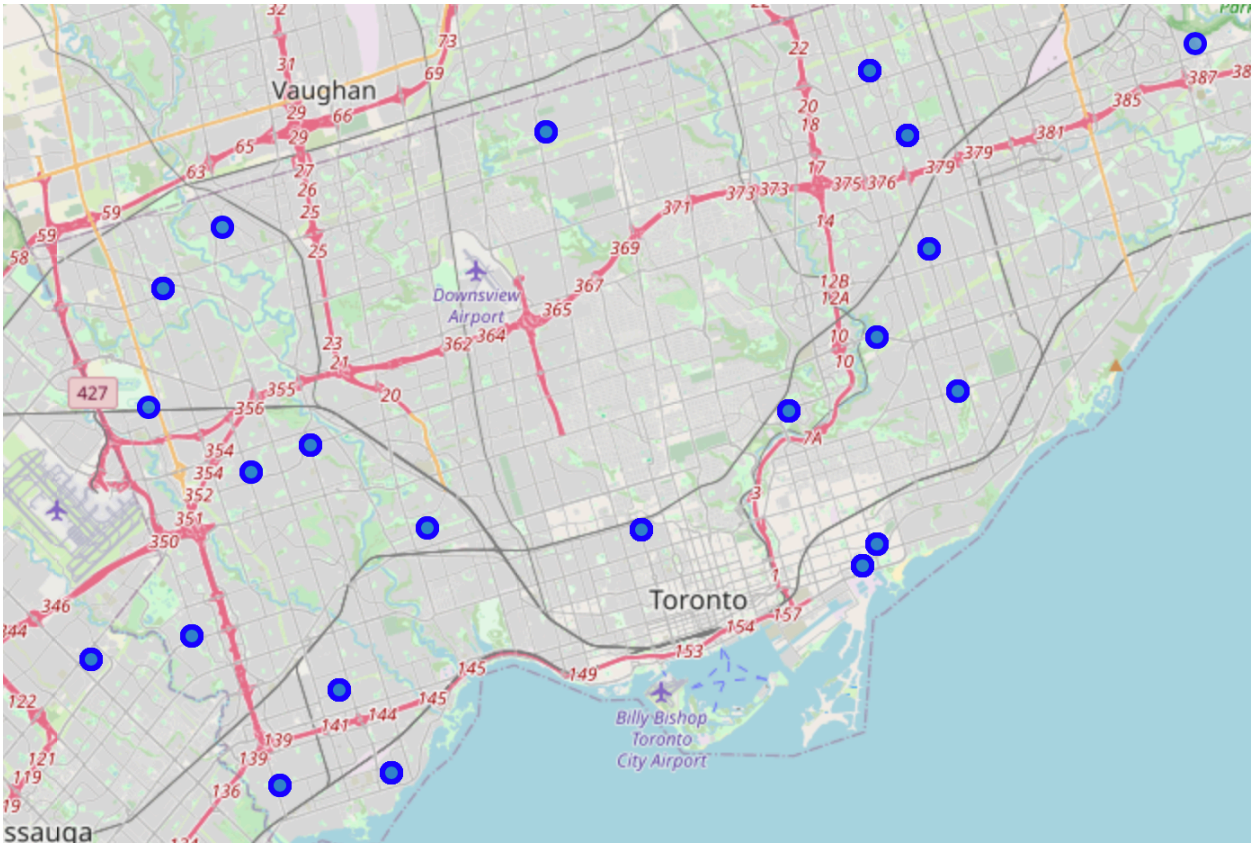

Figure 5.2: Map of neighborhoods in Cluster 2.

According to table 4.3, we can see that fast food and coffee spots along with grocery stores are the most common type of venues here, so maybe we could assume that these are neighborhoods with lots of office buildings, with large amounts of people going to work, thus needing quick meal options both for breakfast and lunch, these neighborhoods can be located in the figure 5.2 above.

We have 167 venues registered for this cluster, so we would recommend following the trend and set up a fast food restaurant or a coffee shop.

- Cluster 3

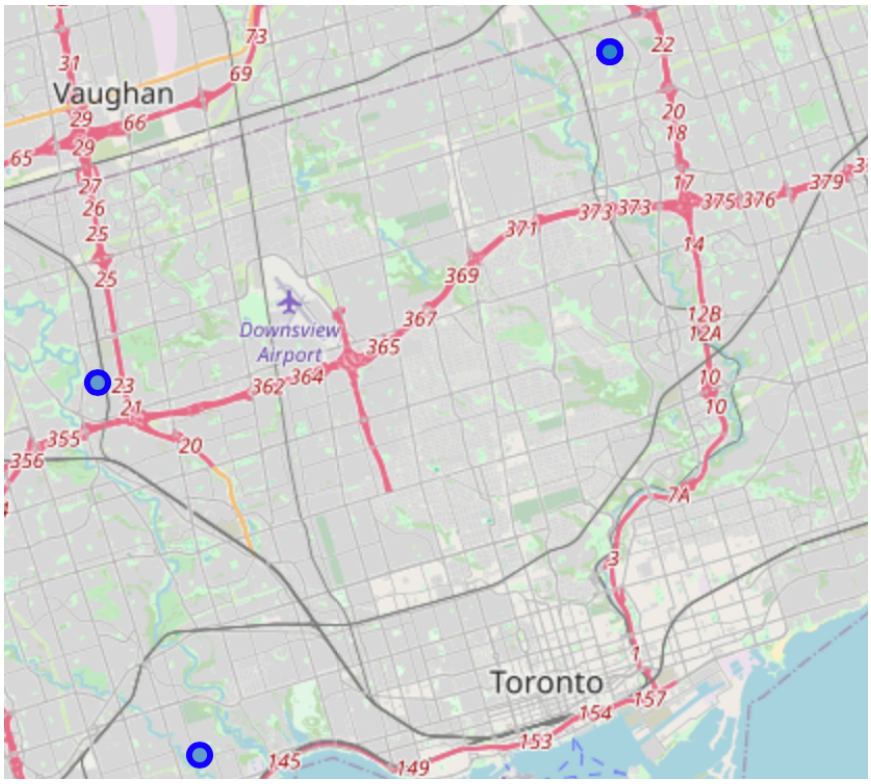

Figure 5.3: Map of neighborhoods in Cluster 3.

Based on the data retrieved from table 4.4, we can reach a conclusion akin to that of Cluster 1, as it seems to be similar to Cluster 3, these seem to be residential neighborhoods in the greater Toronto area.

We only have 6 venues registered for this cluster, so if someone wants to set up shop here, we would recommend for it to be a European cuisine restaurant or a grocery store or market.

- Cluster 4

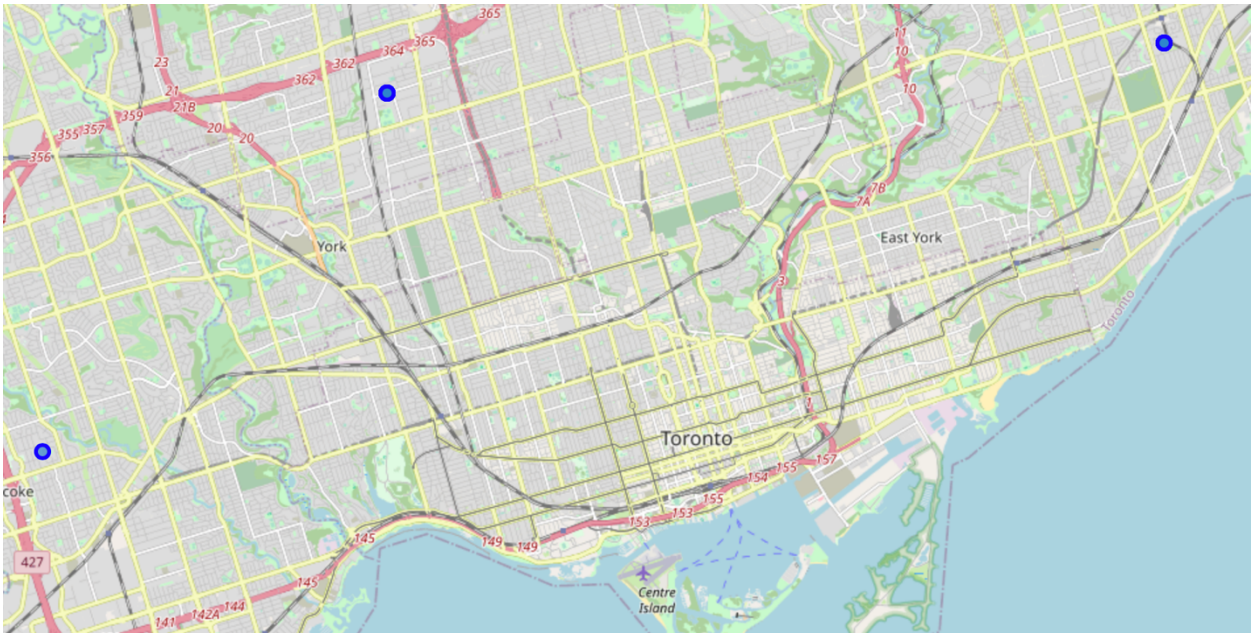

Figure 5.4: Map of neighborhoods in Cluster 4.

These neighborhoods could be identified as shopping areas where people go to have some leisure time, walk and have a look around, based on the most common type of venues as seen in table 4.5.

Based on the 21 venues registered there, we could recommend setting up a coffee shop or an Asian cuisine restaurant, as they seem to be the preferred businesses other than furniture and clothing stores in the area.

- Cluster 5

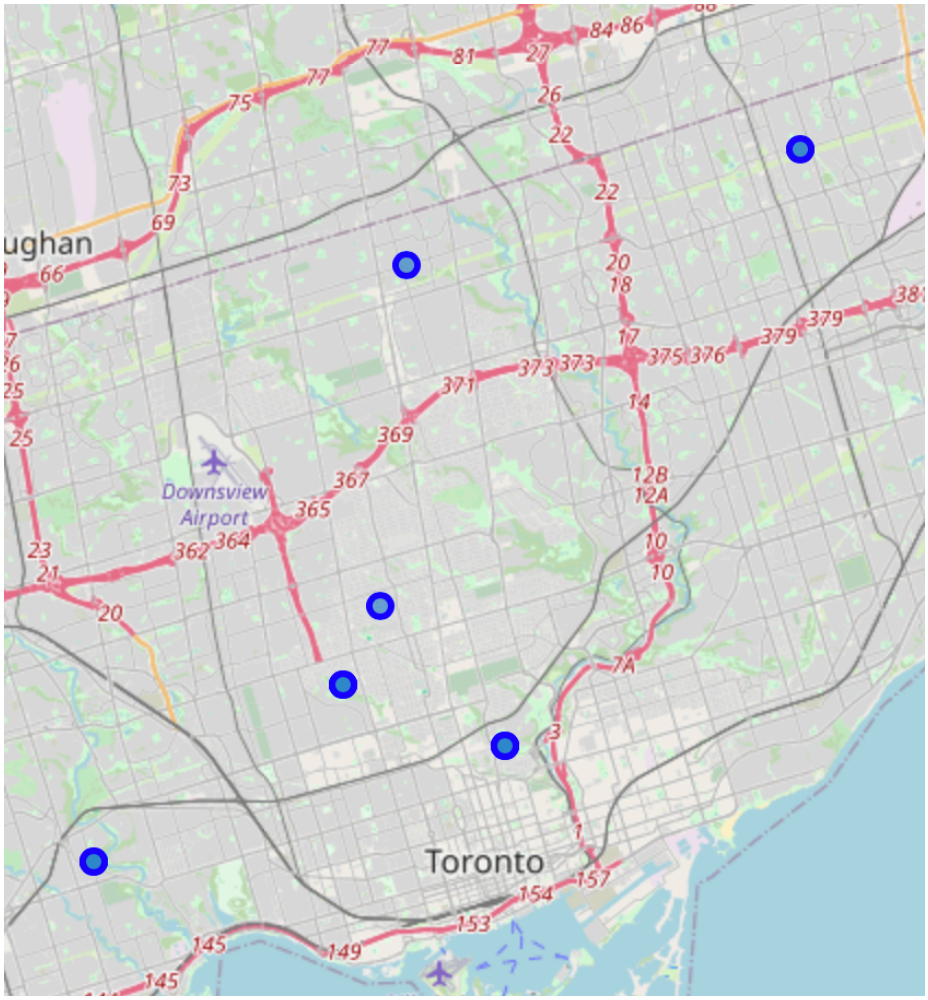

Figure 5.5: Map of neighborhoods in Cluster 5.

Based on the three most common venues for the neighborhoods in cluster 5, as seen in table 4.6, these neighborhoods could be identified as outskirts of the city, with natural spaces and big venues for sporting and other massive events.

With the information gathered on table 4.6, we can see that the other popular venues in these neighborhoods tend to be for sporting activities such as gyms, so a business in this category would be the recommendation.

- Cluster 6

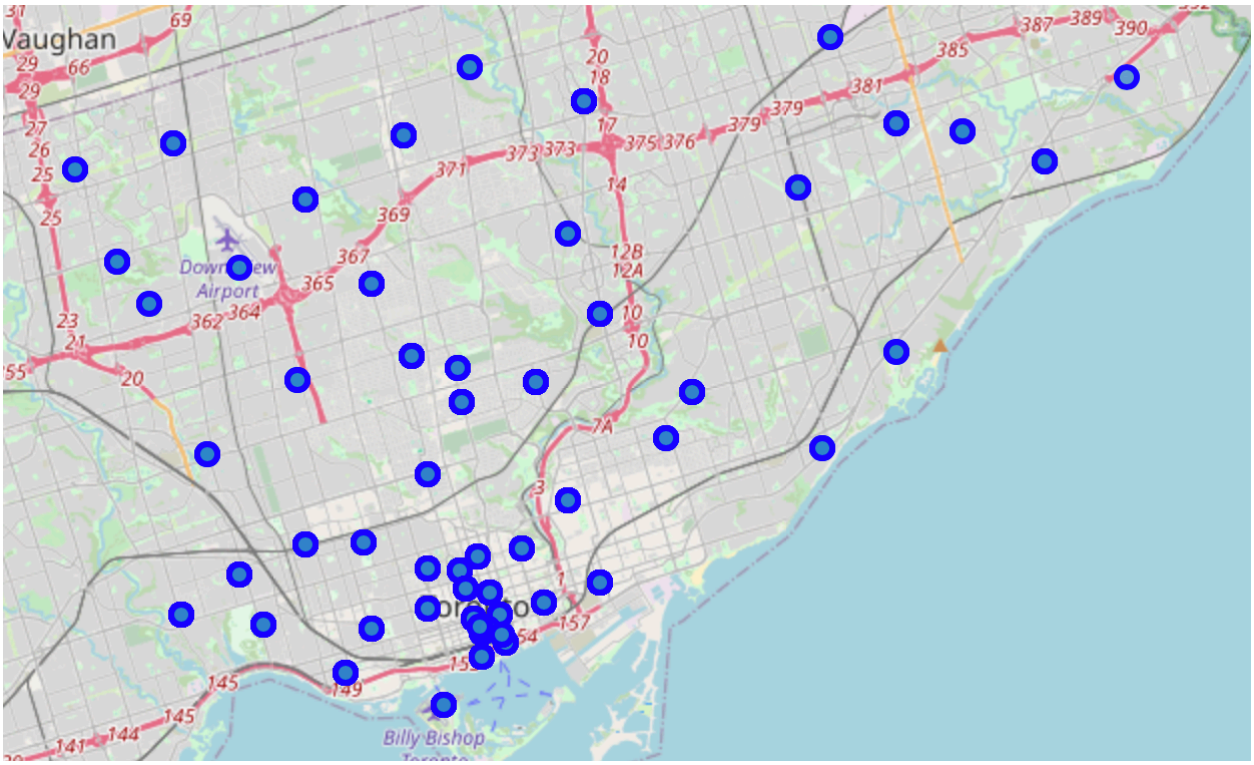

Figure 5.6: Map of neighborhoods in cluster 6.

Finally, we have cluster number 6, with 1190 venues belonging to all the 21 category bins, these areas are obviously the most populated, including all the central Toronto neighborhoods, the most common type of businesses here are coffee spots and department stores.

We can see the relationship between venue category bin and rating in this box plot below:

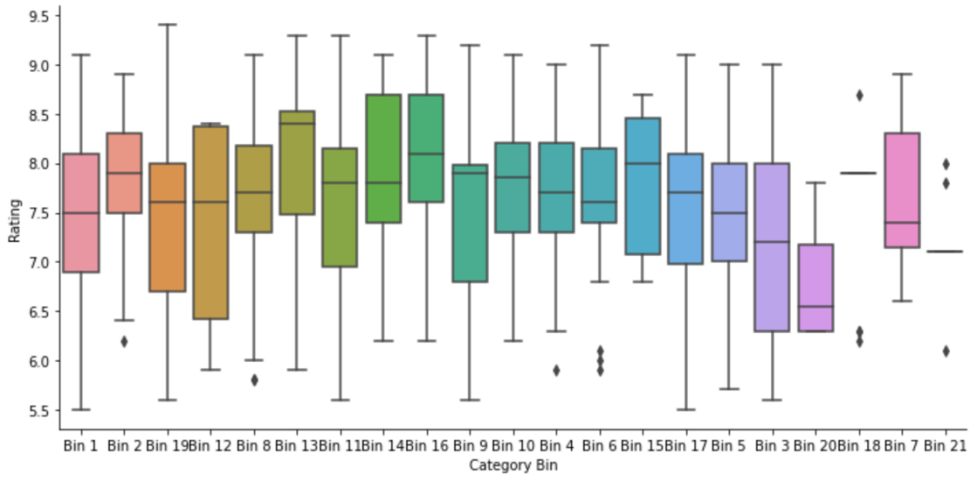

Figure 5.7: Box plot showing relationship between category bin and rating for cluster 6.

Sadly, we can see no real distinction between the category bins and their box plot for rating distribution, we can see the boxes have meaningful overlaps around the 7.5 and 8 ratings.
This means that the category bins alone will not suffice in order to try and predict the ratings, any function trying to accomplish this will likely return a value averaging all the ratings found around the 7.5 mark, just as the below box plot shows:

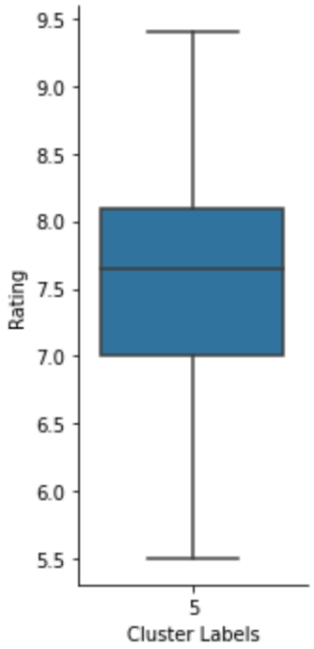

Figure 5.8: Box plot showing relationship between category bin and rating for cluster 6.

6. Conclusion

Coursing this certification really opened my eyes to the vast array of disciplines at work in the data science field and provided me with the knowledge on how to apply them to real world situations, as a data analyst, I can appreciate the usefulness of having this set of skills.
It was a comprehensive review of different topics, and I found the Machine Learning course particularly interesting, I was amazed by the seemingly simple algorithms that are capable of predicting future outcomes, finding hidden relationships or categorize items in a way that seemed impossible before. But the course that I found most useful was the Data Visualization with Python. I wasn't aware of all the different options we have to visualize different types of data in charts, plots and even maps. I certainly see the usefulness of the plotting directly from python in order to convey information about our data in a quick and digestible way, in some cases a simple plot will suffice to encapsulate data that we would otherwise had to explain with words, such a practice can easily get overwhelming if the relationships presented are not particularity simple ones.

Now specifically about this capstone project, the thing I liked the most was the real querying to the Foursquare database through its API, it was my first time running such operations and I found very enriching to retrieve data from them and then manipulating it for our own purposes.

Unfortunately, the initial goal of defining a function that would predict the rating of a particular venue based on its location and category was not met. While we were indeed capable of extracting insights from the dataset, it was its underlying nature that made us run into our dead end. It seems I tried to abstract the venues too much, which led us to have a dataset with high bias and low variance, meaning the information provided was too simple and the features for the data points weren't enough to differentiate them from one another, thus predictability wasn't possible. In short, the rating of the venues did not depend solely on their location and category, in order to accomplish it we would need to extract more data about each venue such as opening hours, venue capacity, availability of certain products, etc.

Nevertheless, as said before, we were still able to obtain knowledge from the data and extract key insights with the segmentation we ran. We visualized and explored the different neighborhoods in the city of Toronto and the type of businesses that usually thrive there, now we know Toronto really likes drinking coffee!

With the digital revolution we have witnessed during the last decade that led to the rapid rise of cloud services, applications and user generated content, it is now imperative for everyone and for every type of business to exploit the opportunities that this huge amount of data provides. Every business that wants to stay ahead of its competition needs to dominate this field to a certain degree of expertise, because, as we've reviewed, obtaining, manipulating and extracting knowledge from the data is no trivial task, but it is certainly rewarding.

7. References

- van der Aalst W. (2016) Data Science in Action. In: Process Mining. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-49851-4_1

- (2020). Places Database. Foursquare. https://developer.foursquare.com/docs/places-database/

- (2020). Places API. Foursquare. https://developer.foursquare.com/docs/places-api/

https://www.kdnuggets.com/2016/12/4-reasons-machine-learning-model-wrong.html

Annexes

Annex 1: Neighborhoods included in each cluster; they are grouped by postal code.

- Cluster 1:
    - Parkwoods
    - The Beaches
    - Caledonia-Fairbanks
    - Scarborough Village
    - East Toronto, Broadview North (Old East York)
    - North Park, Maple Leaf Park, Upwood Park
    - Lawrence Park
    - Weston
    - York Mills West
    - Forest Hill North & West, Forest Hill Road Park
    - Moore Park, Summerhill East

- Cluster 2:
    - Victoria Village
    - Malvern, Rouge
    - Eringate, Bloordale Gardens, Old Burnhamthorpe, Markland Wood
    - Thorncliffe Park
    - Golden Mile, Clairlea, Oakridge
    - India Bazaar, The Beaches West
    - Humber Summit
    - Runnymede, The Junction North
    - Westmount
    - Wexford, Maryvale
    - Willowdale, Willowdale West
    - The Annex, North Midtown, Yorkville
    - Canada Post Gateway Processing Centre
    - Kingsview Village, St. Phillips, Martin Grove Gardens, Richview Gardens
    - Clarks Corners, Tam O'Shanter, Sullivan
    - New Toronto, Mimico South, Humber Bay Shores
    - South Steeles, Silverstone, Humbergate, Jamestown, Mount Olive, Beaumond Heights, Thistletown, Albion Gardens
    - Steeles West, L'Amoreaux West
    - Alderwood, Long Branch
    - Northwest, West Humber - Clairville
    - Business reply mail Processing Centre, South Central Letter Processing Plant Toronto
    - Mimico NW, The Queensway West, South of Bloor, Kingsway Park South West, Royal York South West

- Cluster 3:
    - Hillcrest Village
    - Humberlea, Emery
    - Old Mill South, King's Mill Park, Sunnylea, Humber Bay, Mimico NE, The Queensway East, Royal York South East, Kingsway Park South East

- Cluster 4:
    - Lawrence Manor, Lawrence Heights
    - West Deane Park, Princess Gardens, Martin Grove, Islington, Cloverdale
    - Kennedy Park, Ionview, East Birchmount Park

- Cluster 5:
    - Humewood-Cedarvale
    - Willowdale, Newtonbrook
    - Roselawn
    - Milliken, Agincourt North, Steeles East, L'Amoreaux East
    - Rosedale
    - The Kingsway, Montgomery Road, Old Mill North

- Cluster 6:
    - Regent Park, Harbourfront
    - Queen's Park, Ontario Provincial Government
    - Don Mills
    - Parkview Hill, Woodbine Gardens
    - Garden District, Ryerson
    - Glencairn
    - Rouge Hill, Port Union, Highland Creek
    - Woodbine Heights
    - St. James Town
    - Guildwood, Morningside, West Hill
    - Berczy Park
    - Woburn
    - Leaside
    - Central Bay Street
    - Christie
    - Cedarbrae
    - Bathurst Manor, Wilson Heights, Downsview North
    - Richmond, Adelaide, King
    - Dufferin, Dovercourt Village

- Fairview, Henry Farm, Oriole
- Northwood Park, York University
- Harbourfront East, Union Station, Toronto Islands
- Little Portugal, Trinity
- Bayview Village
- Downsview
- The Danforth West, Riverdale
- Toronto Dominion Centre, Design Exchange
- Brockton, Parkdale Village, Exhibition Place
- Commerce Court, Victoria Hotel
- Cliffside, Cliffcrest, Scarborough Village West
- Studio District
- Bedford Park, Lawrence Manor East
- Del Ray, Mount Dennis, Keelsdale and Silverthorn
- Birch Cliff, Cliffside West
- Willowdale, Willowdale East
- Dorset Park, Wexford Heights, Scarborough Town Centre
- Davisville North
- High Park, The Junction South
- North Toronto West, Lawrence Park
- Parkdale, Roncesvalles
- Agincourt
- Davisville
- University of Toronto, Harbord
- Runnymede, Swansea
- Kensington Market, Chinatown, Grange Park
- Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park
- CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport
- Stn A PO Boxes
- St. James Town, Cabbagetown
- First Canadian Place, Underground city
- Church and Wellesley