Predicting the success of a Venue in Toronto.

Mario Andre Molina Caballero

mario.andre.molina@gmail.com

Applied data science capstone project.

IBM data science professional certificate.

Browse | Structure | SQL | Search | Insert | Export | Import | Operations

#	Column	Type	Collation	Attributes	Null	Default
1	ID	bigint(20)		UNSIGNED	No	None
2	post_author	bigint(20)		UNSIGNED	No	0
3	post_date	datetime			No	0000-00-00 00:00:00
4	post_date_gmt	datetime			No	0000-00-00 00:00:00
5	post_content	longtext	utf8_general_ci		No	None
6	post_title	text	utf8_general_ci		No	None
7	post_excerpt	text	utf8_general_ci		No	None
8	post_status	varchar(20)	utf8_general_ci		No	publish
9	comment_status	varchar(20)	utf8_general_ci		No	open
10	ping_status	varchar(20)	utf8_general_ci		No	open
11	post_password	varchar(20)	utf8_general_ci		No	
12	post_name	varchar(200)	utf8_general_ci		No	
13	to_ping	text	utf8_general_ci		No	None
14	pinged	text	utf8_general_ci		No	None

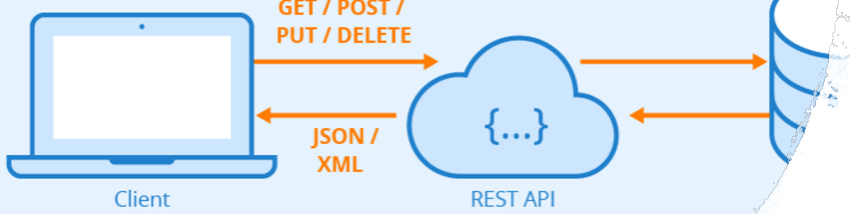
- Introduction

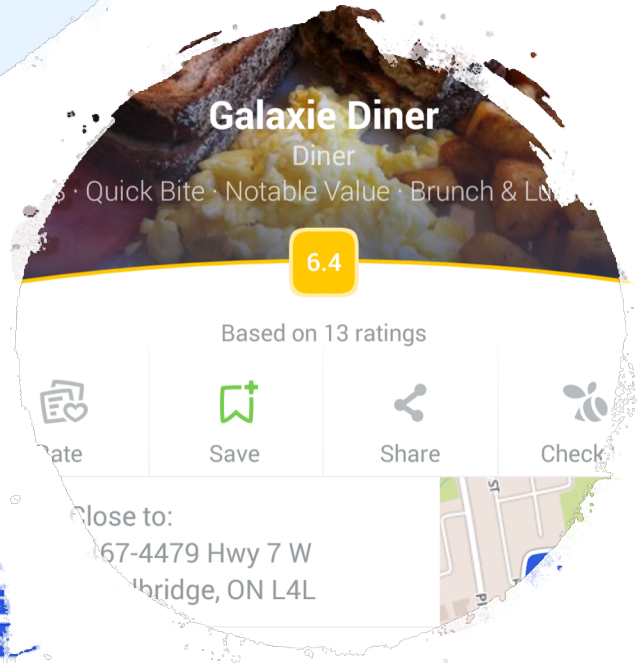
Data science is a multidisciplinary field that allows us to obtain insights and knowledge from huge amounts of unstructured data, with the purpose of generating better strategies or taking important decisions.

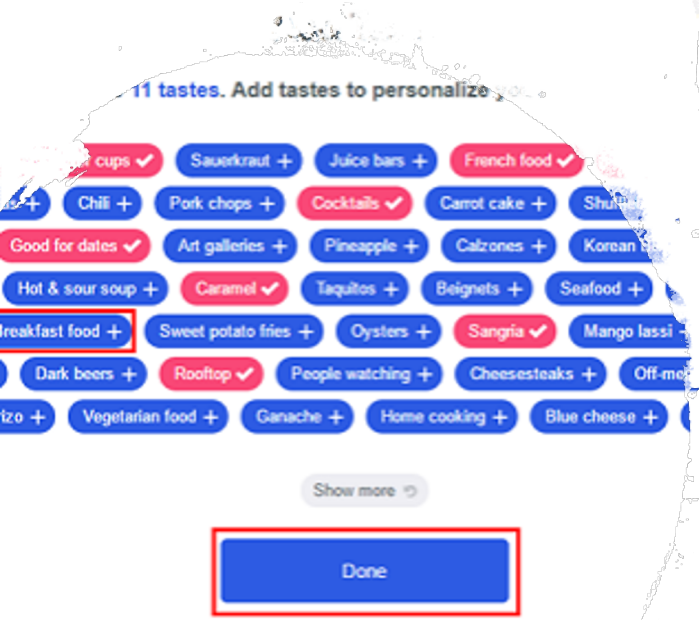
Toronto

The most populous city in Canada, what if I want to open a business there?

We will try to predict whether a business there would succeed or not, based on its category and its location in the city.







We will be using Foursquare, which allows us to query its ‘places database’ through API calls.

- Data

After multiple calls to the Foursquare database, and a lot of manipulation we ended up with a data set that looks like this:

	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

We have more than 260 different venue categories, so we proceeded to group similar categories together, in order to reduce the range of possibilities for a venue.

- Bin 1 for coffee and breakfast spots.
- Bin 2 for sweets and dessert spots.
- Bin 3 for fast food restaurants.
- Bin 4 for European cuisine restaurants.
- Bin 5 for Asian cuisine restaurants.
- Bin 6 for Latin cuisine restaurants.
- Bin 7 for middle eastern and African cuisine restaurants.
- Bin 8 for general food restaurants.
- Bin 9 for grocery stores and markets.
- Bin 10 for entertainment and activities.
- Bin 11 for nightlife, clubs and bars.

- Bin 12 for health and beauty services.
- Bin 13 for nature and outdoors.
- Bin 14 for sport activities.
- Bin 15 for sport and event venues.
- Bin 16 for landmarks, galleries, museums.
- Bin 17 for department stores.
- Bin 18 for transportation services.
- Bin 19 for general services and businesses.
- Bin 20 for college buildings.
- Bin 21 for airport and its services.

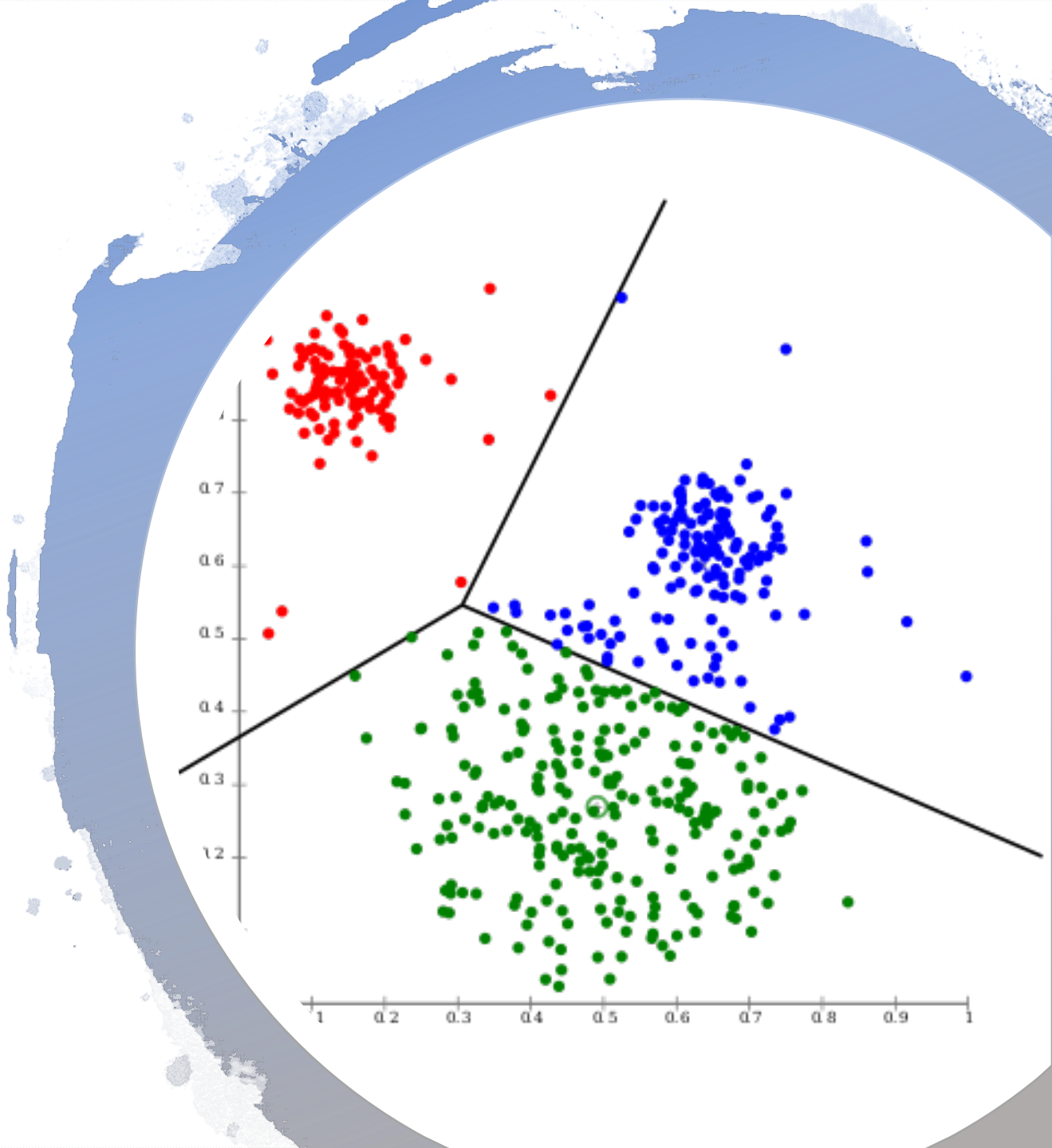
Finally, after dividing the venues by bins, and ensuring uniqueness across the rows, we ended up with a dataframe that looks like this:

	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

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

• Methodology

We will try to group the neighborhoods together in clusters, based on the similarity between them. We will use the three most common bin categories as features, and then we will use the K-means clustering algorithm, with k = 6.



Results

We will explore the results by cluster, for each of them we will include:

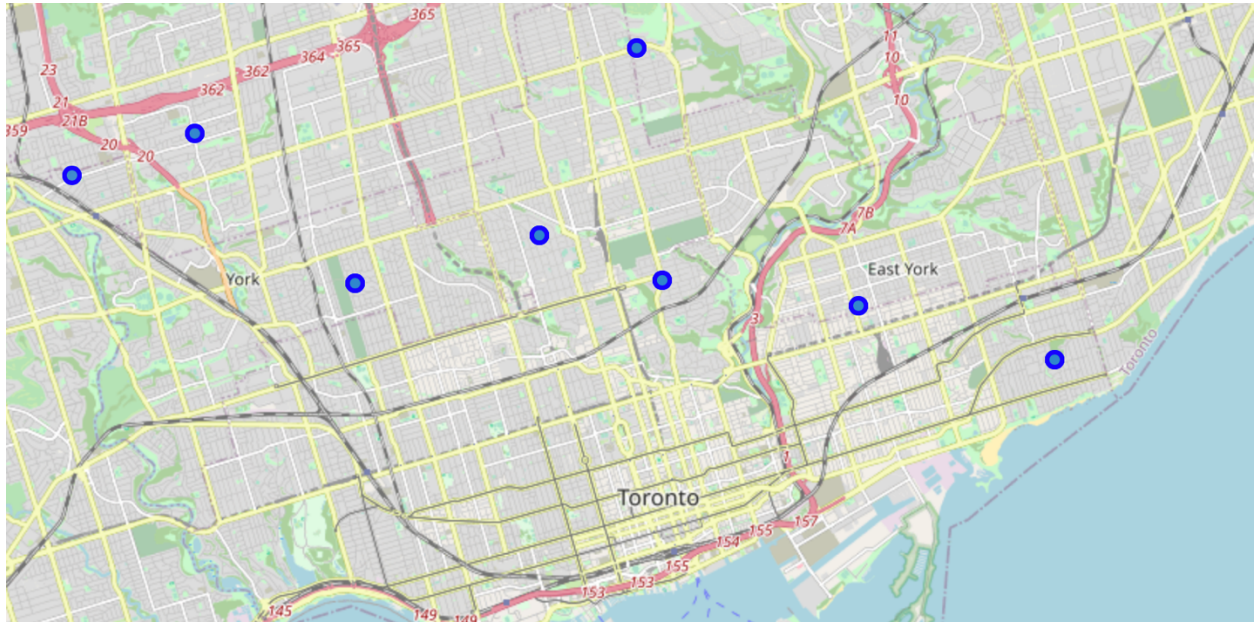
- Number of venues.
- Number of neighborhoods.
- Most common type of venue
- A map to visualize the neighborhoods in Toronto.
- A final recommendation on type of venue most likely to thrive there.

Cluster 1

- We may describe the neighborhoods here 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.

- If you were to open a business here, we would recommend for it to be either a grocery store or a gym.

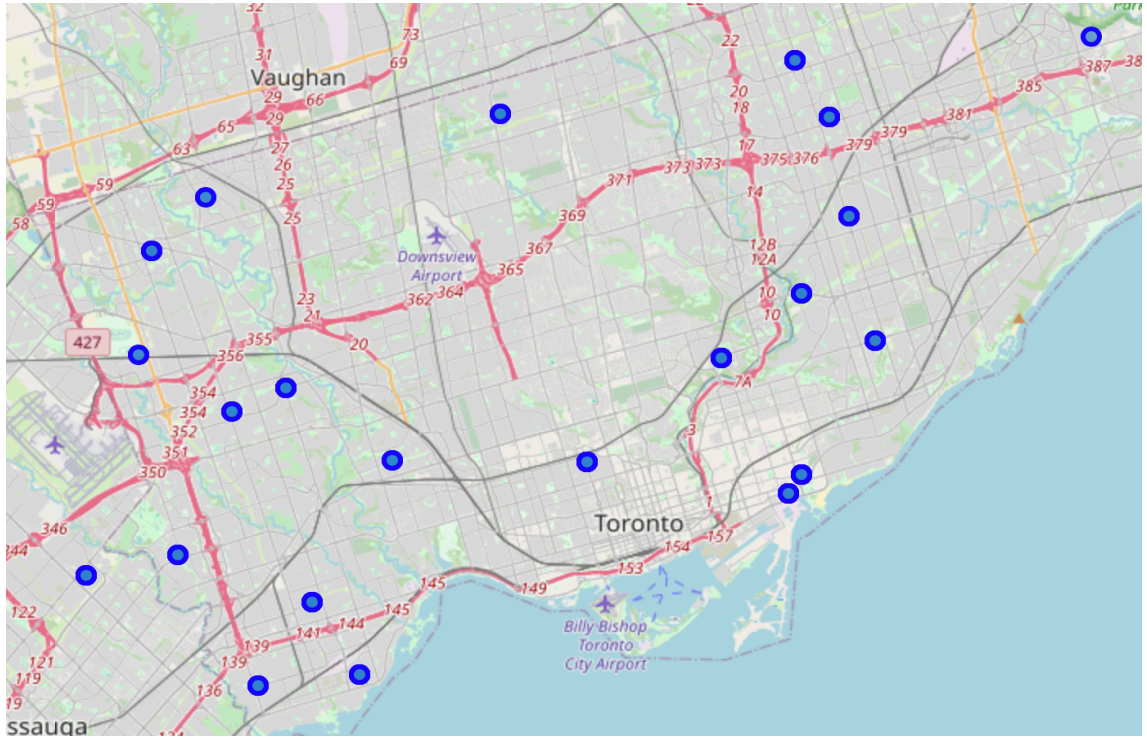
Neighborhoods	11
Category bins	12
Venues	35
1st most common bin	Outdoors
2nd most common bin	Grocery stores, markets
3rd most common bin	Sport activities



Cluster 2

- We may describe the neighborhoods here as neighborhoods with lots of office buildings, with large amounts of people going to work, thus needing quick meal options both for breakfast and lunch
- If you were to open a business here, we would recommend recommend following the trend and set up a fast food restaurant or a coffee shop.

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

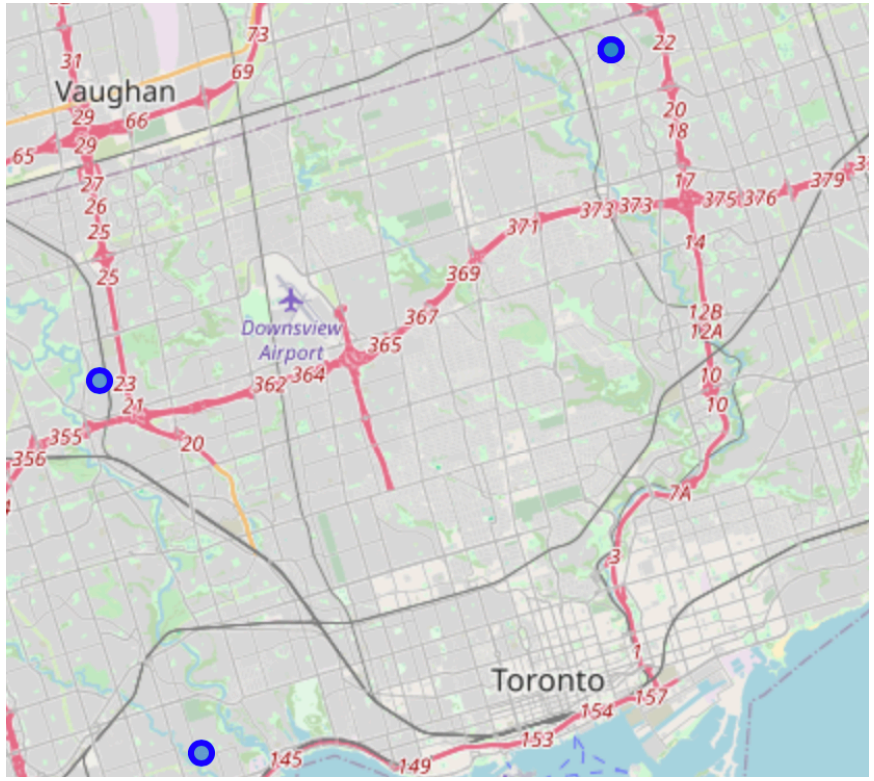


Cluster 3

- Similar to cluster 1, these seem to be residential neighborhoods in the greater Toronto area.

- If you were to open a business here, we would recommend recommend for it to be a European cuisine restaurant or a grocery store or market.

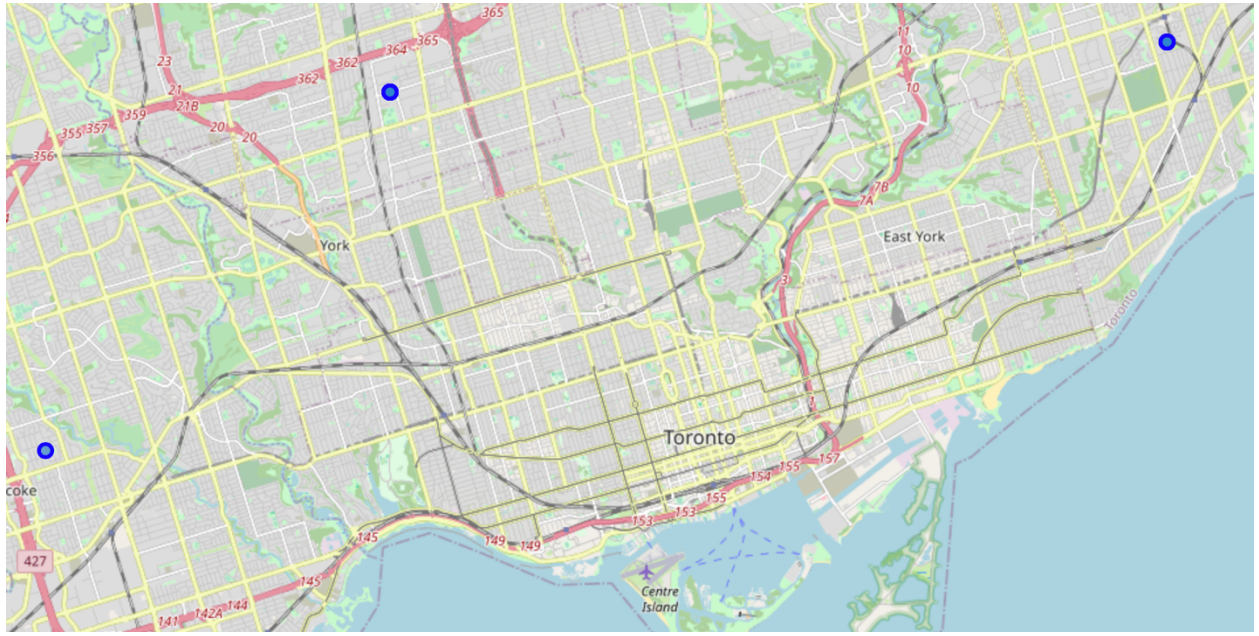
Neighborhoods	3
Category bins	2
Venues	6
1st most common bin	Sport activities
2nd most common bin	European cuisine restaurants
3rd most common bin	Grocery stores, markets



Cluster 4

- The neighborhoods here could be identified as shopping areas where people go to have some leisure time, walk and have a look around
- If you were to open a business here, 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.

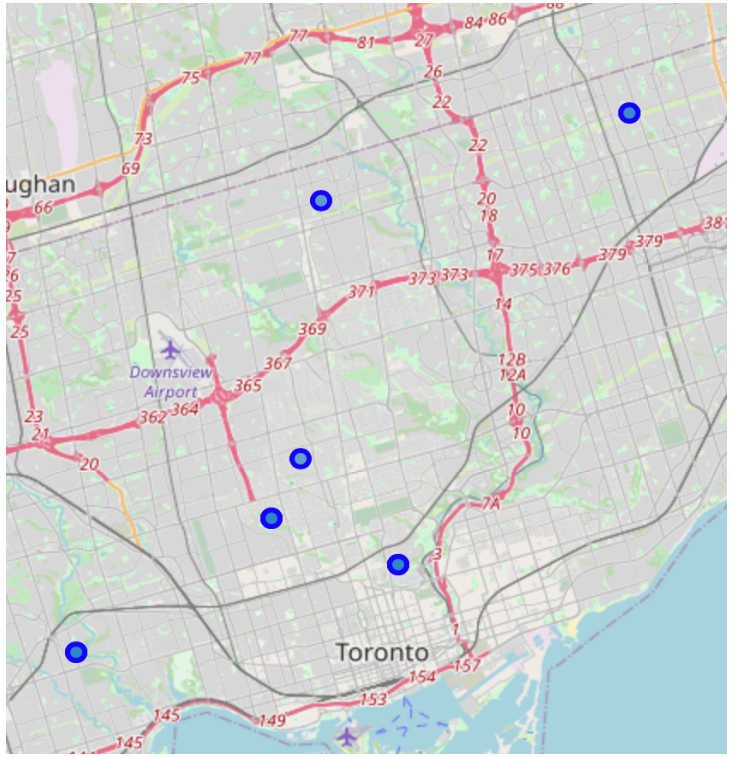
Neighborhoods	3
Category bins	5
Venues	21
1st most common bin	Department stores and shops
2nd most common bin	Coffee and breakfast
3rd most common bin	Asian cuisine restaurants



Cluster 5

- We may describe the neighborhoods here as outskirts of the city, with natural spaces and big venues for sporting and other massive events
- If you were to open a business here, we would recommend for it to be a gym.

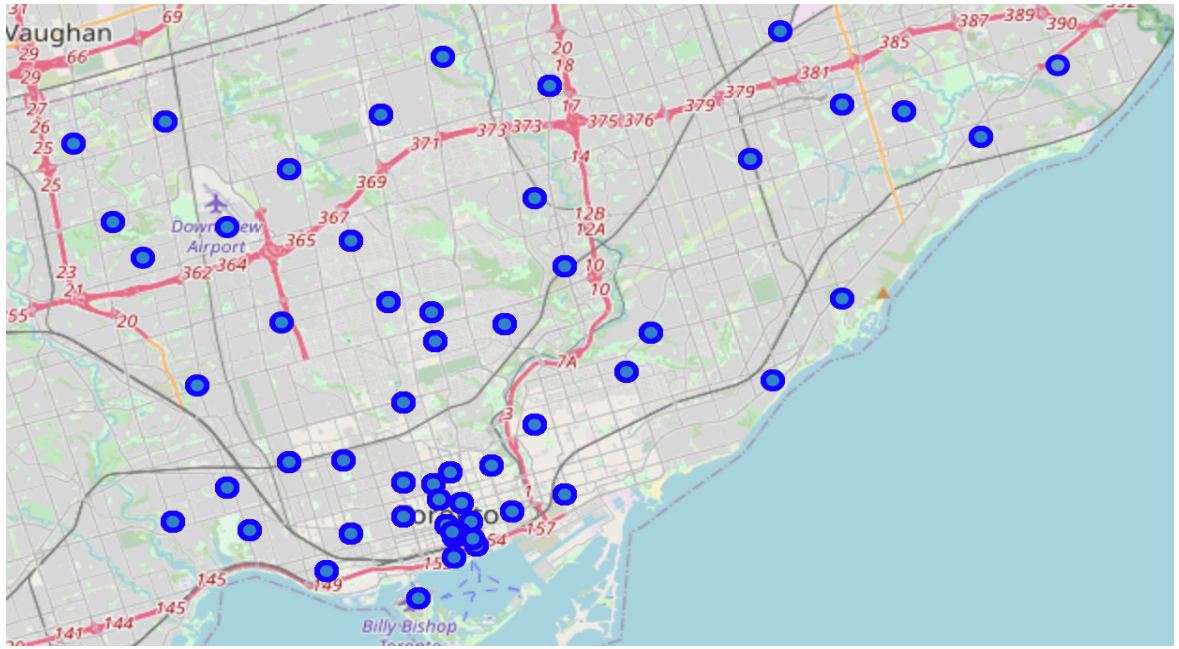
Neighborhoods	6
Category bins	3
Venues	15
1st most common bin	Outdoors
2nd most common bin	Sport activities
3rd most common bin	Sport and event venues



Cluster 6

- These areas are obviously the most populated, including all the central Toronto neighborhoods.
- If you were to open a business here, e would recommend tot follow the trend and set up a coffee spot or a department store.

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



• Conclusion

As we can see, most of the neighborhoods and venues ended up clustered in Cluster 6, as it seems all these neighborhoods were too similar to each other.

Because of the reduced features for each of the venues (we only had category and location) we ran into a common problem in the data science field, a data set with high bias and low variance. Thus the original idea of predicting the ratings was only half achieved with the description of the clustered neighborhoods.

Thanks to this certification I am now fully aware of all the needed knowledge to fully take advantage of the capabilities that data science provides. Coursing it also gave me some of this knowledge of tools and techniques.