IBM Applied Data Science Capstone

Recommending the ideal category

and location for a potentially successful eatery

Fowaad Barkat

1) Background and Problem:

Eateries, with a 60 percent fail rate in the first year, are a challenging business. There is a high upfront start up cost as well as significant competition from the nearby establishments. In order to increase the likelihood of a successful investment, a venture capital firm, interested in investing in a food place in Toronto, has tasked me with researching for the same. After much thought I decided to determine the locations and categories of the 3 most popular eateries in Toronto with the assumption they would have a strong correlation with potential success.

2) Data acquisition:

2.1) Neighborhood:

The link https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&direction=prev&oldid=926287641 was used to acquire information related to the various neighborhoods of Toronto.

2.2) Venue Data:

Foursquare will serve as the source of information related to restaurants in the various neighborhoods of Toronto.

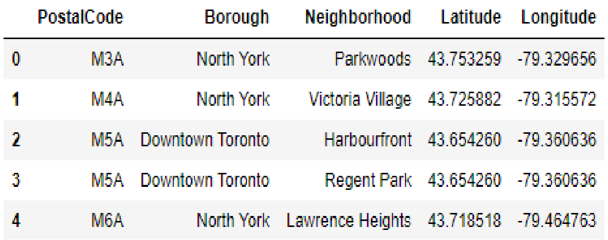
2.3) Geocoding:

Geopy will provide location specifics(longitude, latitude).

3)Methodology:

3.1) Webscraping:

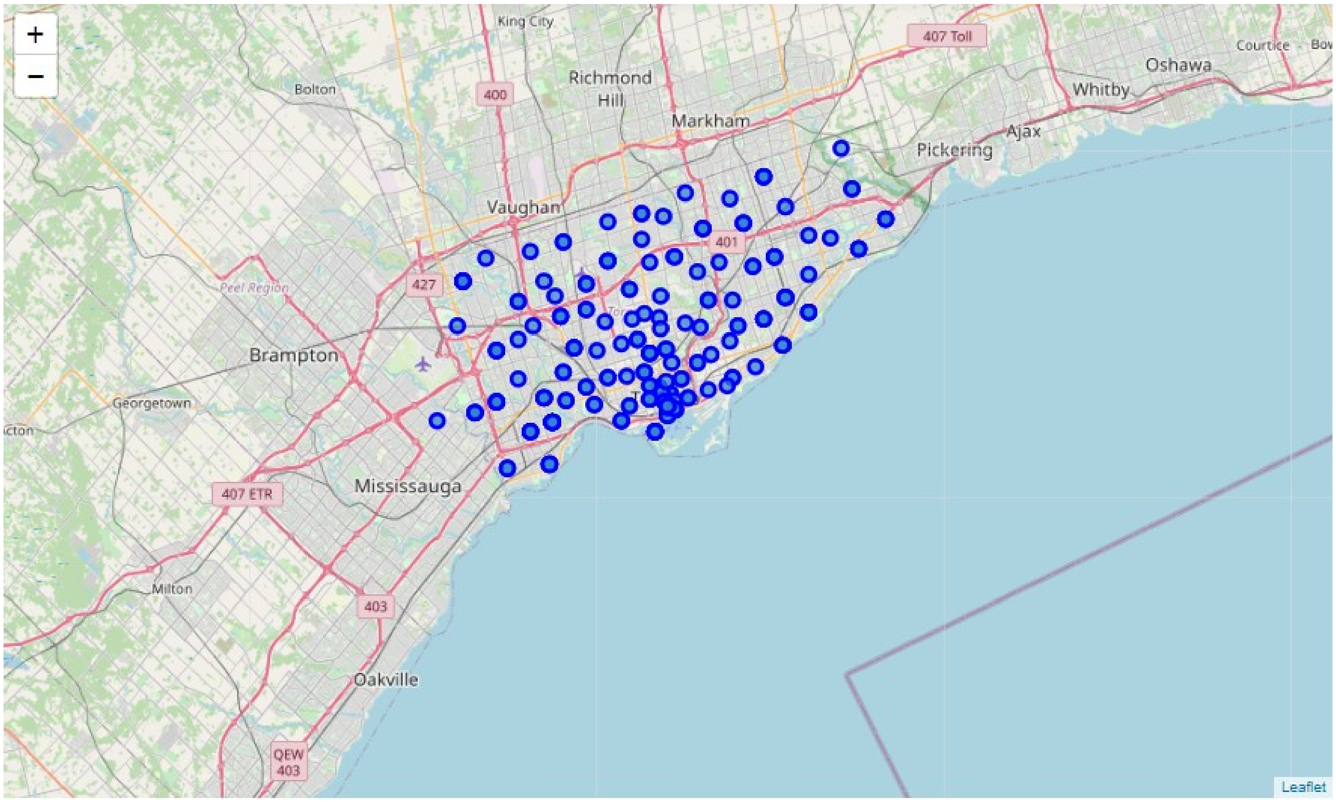
Employed Beautiful soup to scrape postal code, borough and neighborhood data from https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&direction=prev&oldid=926287641, subsequently used geopy to gather location data for each neighborhood. Arranging the combined collected data into the dataframe shown below.



	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	Harbourfront	43.654260	-79.360636
3	M5A	Downtown Toronto	Regent Park	43.654260	-79.360636
4	M6A	North York	Lawrence Heights	43.718518	-79.464763

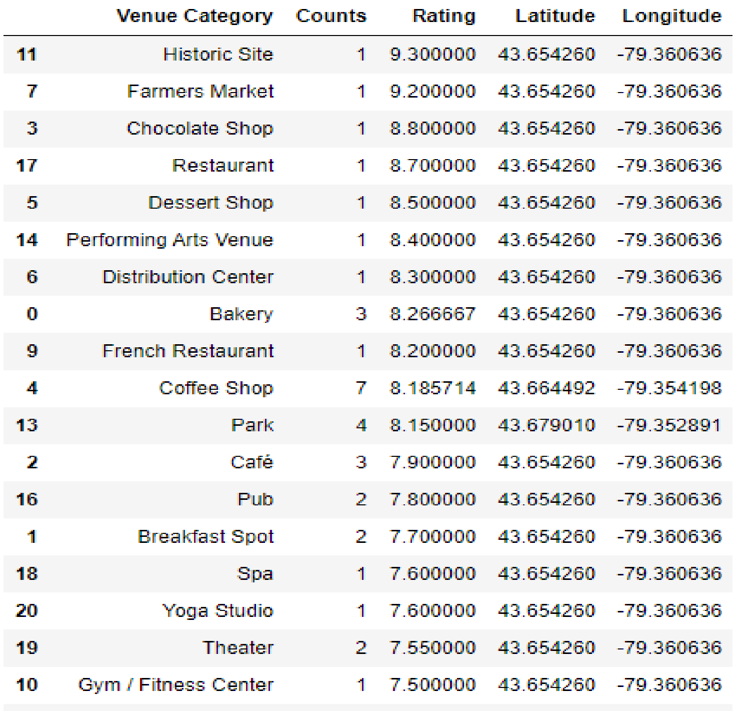
3.2) Folium:

Using the above data and incorporating it into Folium, constructed a map of the various neighborhood of Toronto.



3.3) Foursquare API and data cleaning:

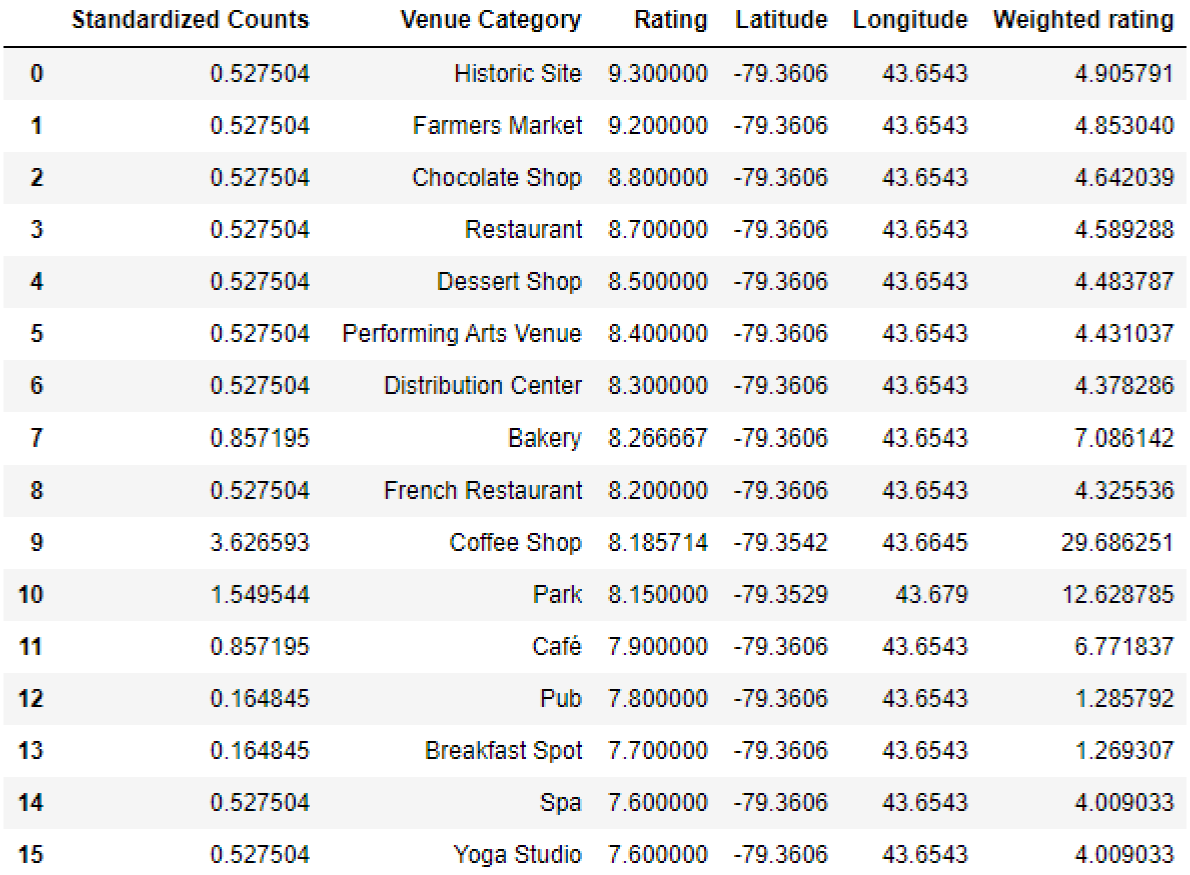
As we wish to acquire details of various eateries in Toronto, foursquare API is used to collect the necessary data, including venue, venue latitude, venue longitude and venue category. The information obtained is sequestered into a new dataframe grouped by venue category, sorted by ratings, with unrated entries being removed from the list. Also added a new counts column to the dataframe to showcase the number of ratings for each venue category.



	Venue Category	Counts	Rating	Latitude	Longitude
11	Historic Site	1	9.300000	43.654260	-79.360636
7	Farmers Market	1	9.200000	43.654260	-79.360636
3	Chocolate Shop	1	8.800000	43.654260	-79.360636
17	Restaurant	1	8.700000	43.654260	-79.360636
5	Dessert Shop	1	8.500000	43.654260	-79.360636
14	Performing Arts Venue	1	8.400000	43.654260	-79.360636
6	Distribution Center	1	8.300000	43.654260	-79.360636
0	Bakery	3	8.266667	43.654260	-79.360636
9	French Restaurant	1	8.200000	43.654260	-79.360636
4	Coffee Shop	7	8.185714	43.664492	-79.354198
13	Park	4	8.150000	43.679010	-79.352891
2	Café	3	7.900000	43.654260	-79.360636
16	Pub	2	7.800000	43.654260	-79.360636
1	Breakfast Spot	2	7.700000	43.654260	-79.360636
18	Spa	1	7.600000	43.654260	-79.360636
20	Yoga Studio	1	7.600000	43.654260	-79.360636
19	Theater	2	7.550000	43.654260	-79.360636
10	Gym / Fitness Center	1	7.500000	43.654260	-79.360636

3.4) Weighted rating:

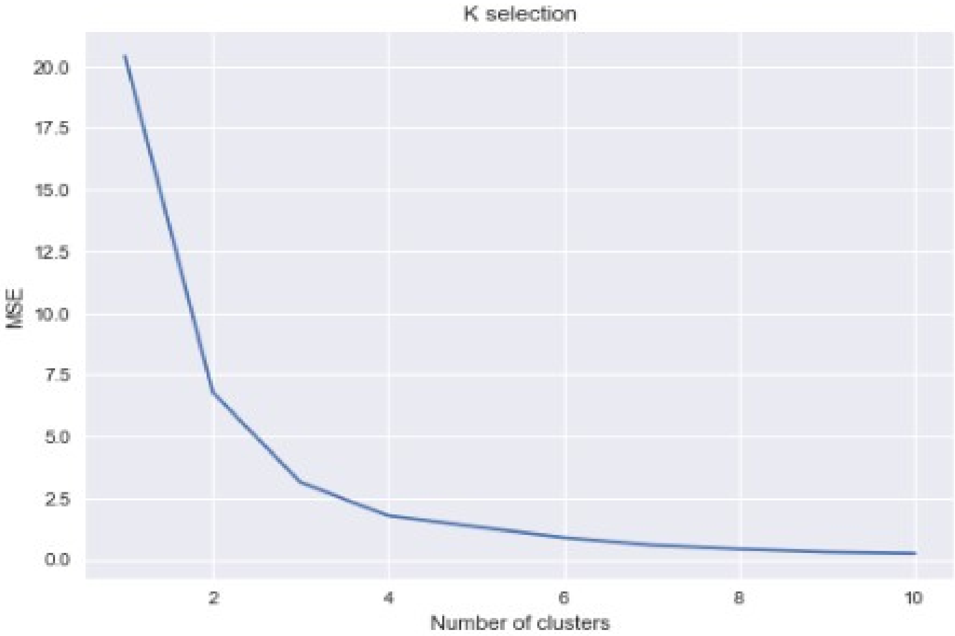
Standardizing the counts/number of ratings and multiplying it with the corresponding ratings to produce a weighted rating, creating a more objective measure for determining the popularity of each venue category.



	Standardized Counts	Venue Category	Rating	Latitude	Longitude	Weighted rating
0	0.527504	Historic Site	9.300000	-79.3606	43.6543	4.905791
1	0.527504	Farmers Market	9.200000	-79.3606	43.6543	4.853040
2	0.527504	Chocolate Shop	8.800000	-79.3606	43.6543	4.642039
3	0.527504	Restaurant	8.700000	-79.3606	43.6543	4.589288
4	0.527504	Dessert Shop	8.500000	-79.3606	43.6543	4.483787
5	0.527504	Performing Arts Venue	8.400000	-79.3606	43.6543	4.431037
6	0.527504	Distribution Center	8.300000	-79.3606	43.6543	4.378286
7	0.857195	Bakery	8.266667	-79.3606	43.6543	7.086142
8	0.527504	French Restaurant	8.200000	-79.3606	43.6543	4.325536
9	3.626593	Coffee Shop	8.185714	-79.3542	43.6645	29.686251
10	1.549544	Park	8.150000	-79.3529	43.679	12.628785
11	0.857195	Café	7.900000	-79.3606	43.6543	6.771837
12	0.164845	Pub	7.800000	-79.3606	43.6543	1.285792
13	0.164845	Breakfast Spot	7.700000	-79.3606	43.6543	1.269307
14	0.527504	Spa	7.600000	-79.3606	43.6543	4.009033
15	0.527504	Yoga Studio	7.600000	-79.3606	43.6543	4.009033

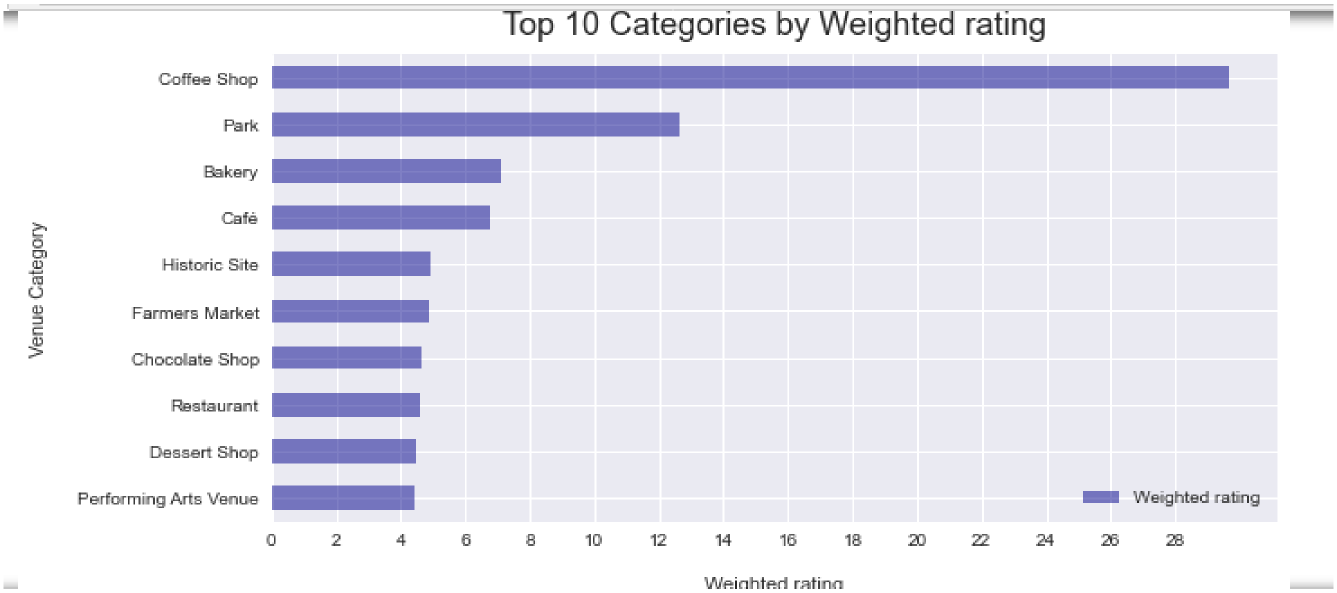
3.4) K-means clustering:

Employed machine learning technique K-means clustering to cluster the data through. Prior to it categorical variables were removed for K-means assesses nominal data. Location data for Neighborhood are removed in favor of the similar columns for venue categories. Elbow method is used to determine optimal clusters which was calculated to be 3.



4) Results and Discussion:

After using the visual tools to analyze data, coffee shops, bakeries and cafes are found to be three of the most popular categories of eateries.

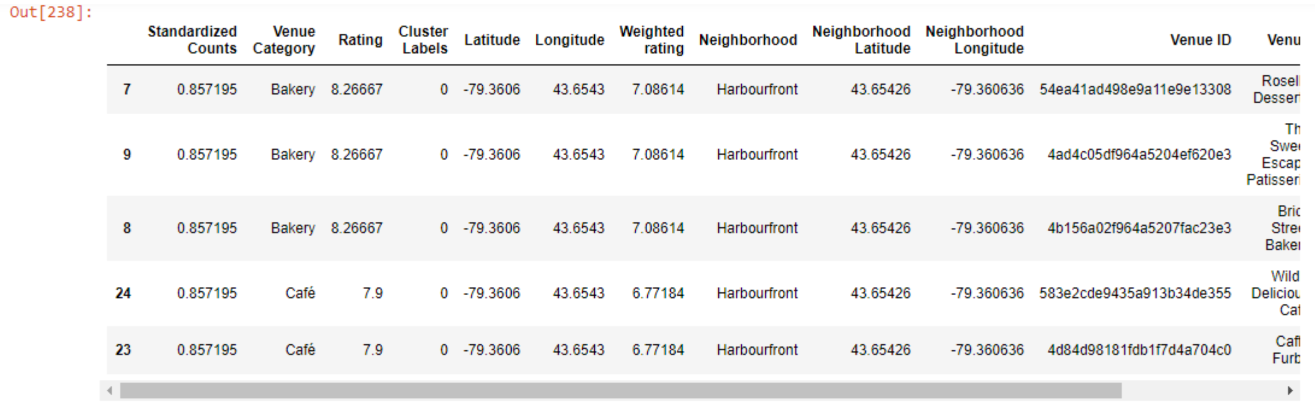


4.1) K-means Clustering:

Further analysis through K means revealed that the Harbourfront neighborhood contains the most popular restaurants.

4.1.a) Cluster 1:

Shows cluster of venue categories from overall data having the lowest weighted rating. Of this cohort, those belonging to the Harbourfront neighborhood have the highest weighted rating.

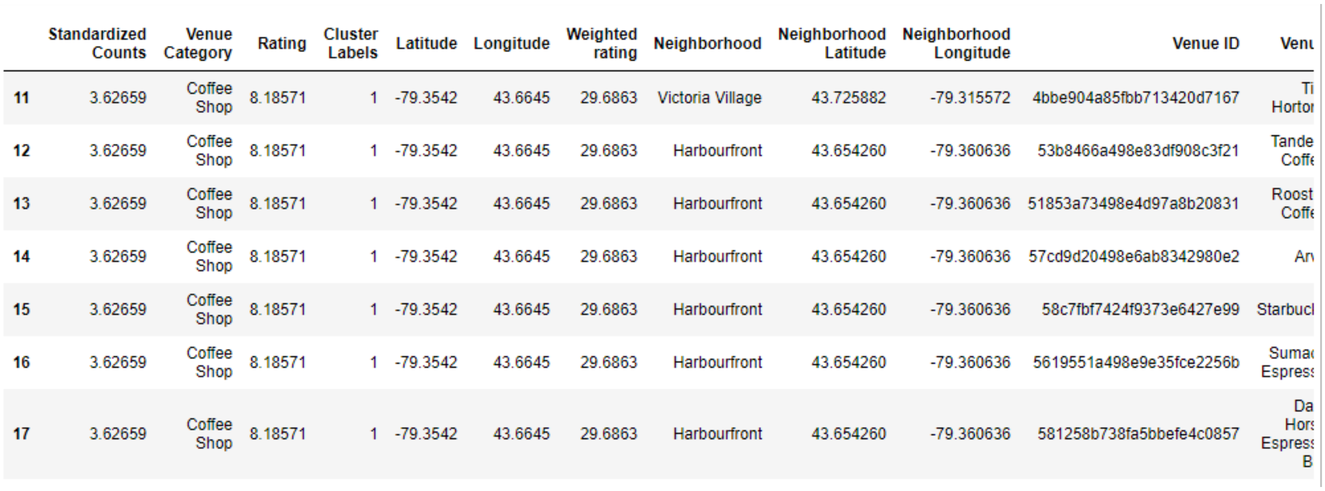


Out[238]:

	Standardized Counts	Venue Category	Rating	Cluster Labels	Latitude	Longitude	Weighted rating	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue ID	Venu
7	0.857195	Bakery	8.26667	0	-79.3606	43.6543	7.08614	Harbourfront	43.65426	-79.360636	54ea41ad498e9a11e9e13308	Rosel Desser
9	0.857195	Bakery	8.26667	0	-79.3606	43.6543	7.08614	Harbourfront	43.65426	-79.360636	4ad4c05df964a5204ef620e3	Th Swe Escap Patisser
8	0.857195	Bakery	8.26667	0	-79.3606	43.6543	7.08614	Harbourfront	43.65426	-79.360636	4b156a02f964a5207fac23e3	Bric Stre Bake
24	0.857195	Café	7.9	0	-79.3606	43.6543	6.77184	Harbourfront	43.65426	-79.360636	583e2cde9435a913b34de355	Wild Deliciou Ca
23	0.857195	Café	7.9	0	-79.3606	43.6543	6.77184	Harbourfront	43.65426	-79.360636	4d84d98181fdb1f7d4a704c0	Caf Furb

4.1.b) Cluster 2:

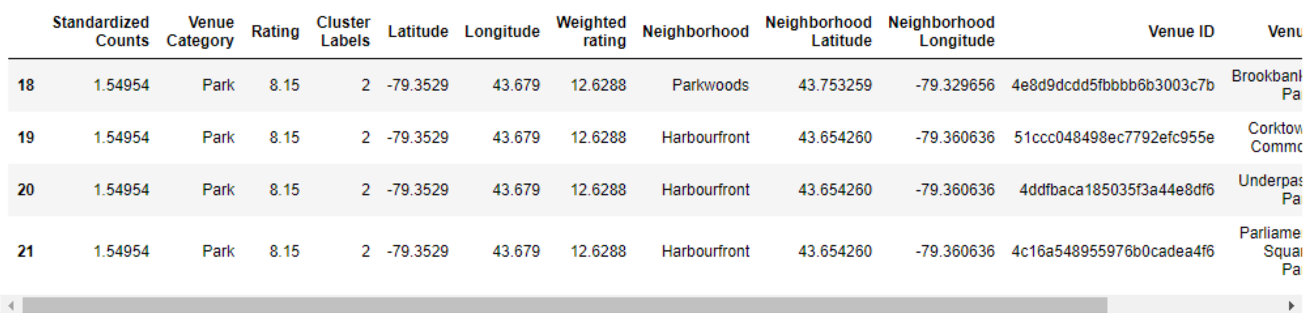
Contains the cohort of venue categories that have the highest weighted rating of eateries compared to all other clusters. Here too, except for one, all are located in Harbourfront.



	Standardized Counts	Venue Category	Rating	Cluster Labels	Latitude	Longitude	Weighted rating	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue ID	Venu
11	3.62659	Coffee Shop	8.18571	1	-79.3542	43.6645	29.6863	Victoria Village	43.725882	-79.315572	4bbe904a85fbb713420d7167	Ti Horto
12	3.62659	Coffee Shop	8.18571	1	-79.3542	43.6645	29.6863	Harbourfront	43.654260	-79.360636	53b8466a498e83df908c3f21	Tande Coffe
13	3.62659	Coffee Shop	8.18571	1	-79.3542	43.6645	29.6863	Harbourfront	43.654260	-79.360636	51853a73498e4d97a8b20831	Roost Coffe
14	3.62659	Coffee Shop	8.18571	1	-79.3542	43.6645	29.6863	Harbourfront	43.654260	-79.360636	57cd9d20498e6ab8342980e2	Ar
15	3.62659	Coffee Shop	8.18571	1	-79.3542	43.6645	29.6863	Harbourfront	43.654260	-79.360636	58c7fbf7424f9373e6427e99	Starbuc
16	3.62659	Coffee Shop	8.18571	1	-79.3542	43.6645	29.6863	Harbourfront	43.654260	-79.360636	5619551a498e9e35fce2256b	Suma Espress
17	3.62659	Coffee Shop	8.18571	1	-79.3542	43.6645	29.6863	Harbourfront	43.654260	-79.360636	581258b738fa5bbefe4c0857	Da Hors Espress B

4.1.c) Cluster 3:

This cluster only contains parks, which have the second highest weighted rating among-st all venue categories, and considering their popularity, can serve as a potentially good location for setting up an extension of the primary eatery location in the form of a food stand.



	Standardized Counts	Venue Category	Rating	Cluster Labels	Latitude	Longitude	Weighted rating	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue ID	Venu
18	1.54954	Park	8.15	2	-79.3529	43.679	12.6288	Parkwoods	43.753259	-79.329656	4e8d9dcdd5fbbbb6b3003c7b	Brookban Pa
19	1.54954	Park	8.15	2	-79.3529	43.679	12.6288	Harbourfront	43.654260	-79.360636	51ccc048498ec7792efc955e	Corktow Commo
20	1.54954	Park	8.15	2	-79.3529	43.679	12.6288	Harbourfront	43.654260	-79.360636	4ddfbaca185035f3a44e8df6	Underpas Pa
21	1.54954	Park	8.15	2	-79.3529	43.679	12.6288	Harbourfront	43.654260	-79.360636	4c16a548955976b0cadea4f6	Parliame Squa Pa

Due to free Sandbox type foursquare account provided in the course there were limitations on how many API calls and results returned, thus information from many different neighborhoods and greater number of similar venue categories, could not be obtained.

Conclusion:

In conclusion from clustering Toronto locations containing restaurants having varying ratings, the Harbourfront neighborhood of Toronto was found to contain the most popular restaurants. On analysis of data done prior to clustering it was also found that coffee shops, bakeries and cafes are three of the highest rated venue categories. It was also discovered that parks ranked second highest among-st venue categories and can serve as an appropriate location for an extension of the primary eatery establishment in the form of a food stand. Further investigation is needed to acquire details of the most popular restaurants, such as their best selling items and management strategies, the latter being pivotal for the success of any venture. Additional research should also be done to acquire more ratings of similar category venues as well as popular eateries of other neighborhoods in Toronto. In order to save time, plans to establish the eatery should be initiated by the venture capital firm. These can be reinforced with new information as it becomes available from additional required research.