Data Science Capstone

Final Project: An analysis of the Food Venues in Toronto

Amy King-Robertson

4/20/2021

Introduction/Business Problem

In the Data Science Capstone course, an exploratory data analysis was conducted to examine and cluster venues in areas of Toronto, Ontario with a Borough name that included the word 'Toronto'. Six boroughs were identified and neighborhoods within each of those boroughs were examined. The analysis revealed the most popular venues within the neighborhoods of those six boroughs.

This report will expand on that analysis to examine the most popular types of restaurants in all boroughs in Toronto, Ontario. The analysis will provide venue information from the Foursquare location data tool to identify the most popular restaurants and their food types in each neighborhood in Toronto. This information can be used by tourists, visitors, or individuals who are new to the city to identify the most popular restaurants in each neighborhood by type of food. This information could also be used by individuals who are considering opening a new restaurant of a specific type to identify where that type of food is most popular in Toronto or where fewer restaurants of that type or located, which may be a good opportunity for business development in less crowded areas.

Data Description

The data used in the analysis will contain Toronto area location identified by postal code from the wikipedia postal code website: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

The data will be downloaded in the form of a json file. This data will be used to identify neighborhoods within Toronto by postal code. Using the postal codes for each of the neighborhoods, the longitude and latitude can be identified and passed to the FourSquare application to obtain the venue listings. A sample of the Postal Code json data is below.

```
]: [{'PostalCode': 'M3A', 'Borough': 'North York', 'Neighborhood': 'Parkwoods'},
    {'PostalCode': 'M4A',
     'Borough': 'North York',
     'Neighborhood': 'Victoria Village'},
    {'PostalCode': 'M5A',
     'Borough': 'Downtown Toronto',
     'Neighborhood': 'Regent Park, Harbourfront'},
    {'PostalCode': 'M6A',
     'Borough': 'North York',
     'Neighborhood': 'Lawrence Manor, Lawrence Heights'},
    {'PostalCode': 'M7A',
     'Borough': "Queen's Park",
     'Neighborhood': 'Ontario Provincial Government'},
    {'PostalCode': 'M9A',
     'Borough': 'Etobicoke',
     'Neighborhood': 'Islington Avenue'},
    {'PostalCode': 'M1B',
     'Borough': 'Scarborough',
     'Neighborhood': 'Malvern, Rouge'},
```

Geocoder data to obtain each neighborhood's longitude and latitude was an option considered, but the application is unstable so instead the longitude and latitude will come from a pre-loaded file from the Data Science capstone course that was used in the Week 3 Assignment.https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv

This data file provides postal code, longitude, and latitude and will be merged with the postal code data to produce a dataset that includes the longitude and latitude for each neighborhood, which will be used with the FourSquare application.

A sample of the csv data that was read into a dataframe is provided below.

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Venue information for each neighborhood will be obtained using the FourSquare application. Information on FourSquare can be found here: https://foursquare.com/ or on this page for developer information: https://developer.foursquare.com/

FourSquare data will be obtained using the dataset produced with the two above data sources, identifying each Toronto neighborhood, postal code, longitude and latitude. The data from FourSquare will be received

in the form of a json file, with a sample of the data provided below.


```
{'meta': {'code': 200, 'requestId': '607c84dd1f7e8830e82b18d9'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Bay Street Corridor',
  'headerFullLocation': 'Bay Street Corridor, Toronto',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 74,
  'suggestedBounds': {'ne': {'lat': 43.6579817045, 'lng': -79.37772678059432},
   'sw': {'lat': 43.6489816955, 'lng': -79.39014261940568}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '5227bb01498e17bf485e6202',
       'name': 'Downtown Toronto',
       'location': {'lat': 43.65323167517444,
        'lng': -79.38529600606677,
```


Methodology

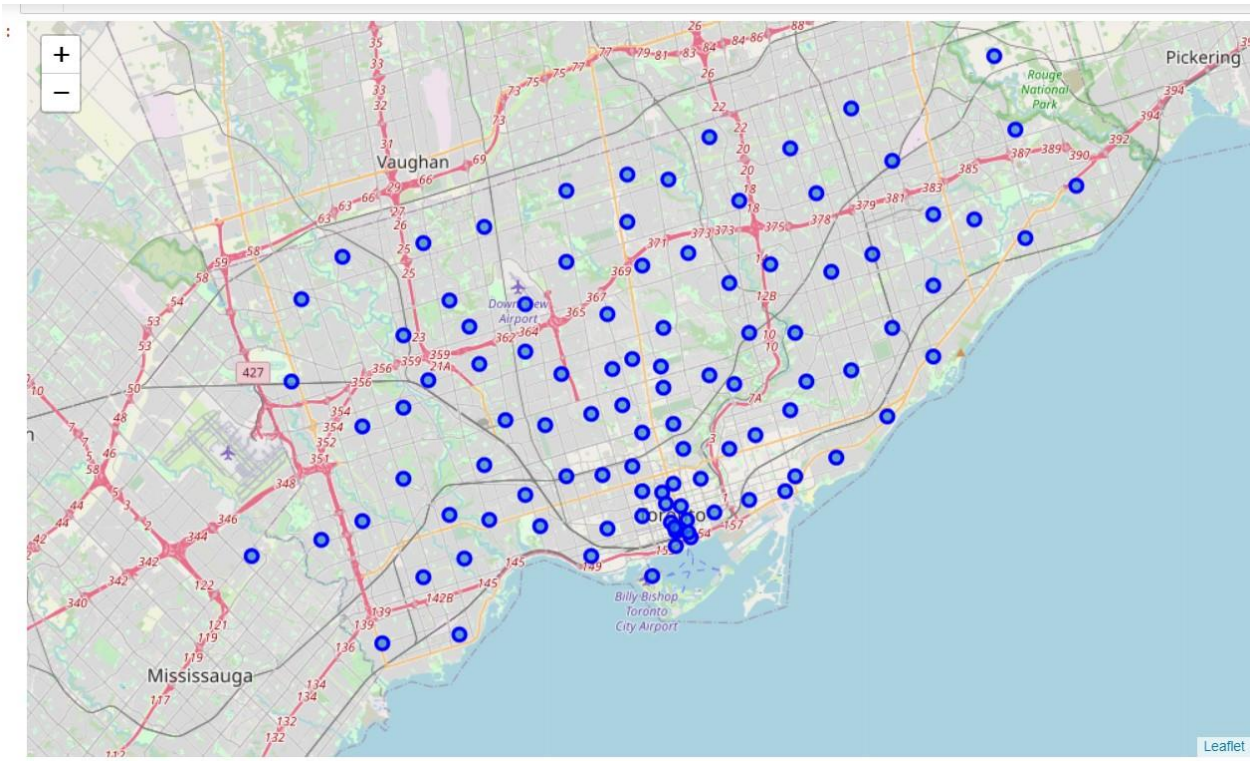
Data Setup and Cleaning

The data was obtained using the sources indicated above. A merged table was created to hold the latitude and longitude for each neighborhood, matching on postal code.

Out[7]:

	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	Queen's Park	Ontario Provincial Government	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills North	43.745906	-79.352188
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
10	M6B	North York	Glencairn	43.709577	-79.445073
11	M9B	Etobicoke	West Deane Park, Princess Gardens, Martin Grov...	43.650943	-79.554724

A Folium map was produced to visualize the neighborhoods identified.



The FourSquare API was then used to obtain the venues within the neighborhoods. Limits were placed on the results to include returning only 100 results within 500 meters of the data point. In addition, the query string also requested to return only results from a specific category, since the analysis was only concerned with identifying types of food eateries available in the area. According to the FourSquare developer site, the category id for 'Food' is 4d4b7105d754a06374d81259, so this category id was included in the query string.

The initial results included 2,106 venues with venue categories that did not match what was expected, returning venues such as parks, Distribution centers, and Hockey Arenas. Further analysis into the FourSquare data showed an additional column called, 'Icon' contained a field specifically identifying a venue as a 'food' venue. This column was added to the table for use in data cleaning.

Out[279]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Category_ID	Icon	
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park	4bf58dd8d48988d163941735	{'prefix': 'https://ss3.4sqi.net/img/categorie...	htt
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop	4bf58dd8d48988d1f9941735	{'prefix': 'https://ss3.4sqi.net/img/categorie...	h
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena	4bf58dd8d48988d185941735	{'prefix': 'https://ss3.4sqi.net/img/categorie...	ht
3	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant	4def73e84765ae376e57713a	{'prefix': 'https://ss3.4sqi.net/img/categorie...	ht
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop	4bf58dd8d48988d1e0931735	{'prefix': 'https://ss3.4sqi.net/img/categorie...	ht
5	Victoria Village	43.725882	-79.315572	Pizza Nova	43.725824	-79.312860	Pizza Place	4bf58dd8d48988d1ca941735	{'prefix': 'https://ss3.4sqi.net/img/categorie...	h
6	Regent Park, Harbourfront	43.654260	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery	4bf58dd8d48988d16a941735	{'prefix': 'https://ss3.4sqi.net/img/categorie...	ht
7	Regent Park, Harbourfront	43.654260	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop	4bf58dd8d48988d1e0931735	{'prefix': 'https://ss3.4sqi.net/img/categorie...	ht
8	Regent Park, Harbourfront	43.654260	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center	52e81612bcbc57f1066b7a37	{'prefix': 'https://ss3.4sqi.net/img/categorie...	h
9	Regent Park, Harbourfront	43.654260	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa	4bf58dd8d48988d1ed941735	{'prefix': 'https://ss3.4sqi.net/img/categorie...	http

The Icon column was expanded to extract the ‘prefix’ data and the table was reset to include only results that contained the word, ‘food’ in the prefix string. A review of the unique venue categories after completion resulted in the following.

```
In [283]: 1 toronto_rests['Venue Category'].unique()

Out[283]: array(['Food & Drink Shop', 'Portuguese Restaurant', 'Coffee Shop',
                 'Pizza Place', 'Bakery', 'Restaurant', 'Breakfast Spot',
                 'Farmers Market', 'Dessert Shop', 'French Restaurant', 'Café',
                 'Asian Restaurant', 'Vietnamese Restaurant', 'Italian Restaurant',
                 'Creperie', 'Sushi Restaurant', 'Burrito Place',
                 'Mexican Restaurant', 'Diner', 'Wings Joint',
                 'Fried Chicken Joint', 'Japanese Restaurant', 'Smoothie Shop',
                 'Sandwich Place', 'Fast Food Restaurant', 'Caribbean Restaurant',
                 'Gastropub', 'Ramen Restaurant', 'Burger Joint', 'Steakhouse',
                 'Thai Restaurant', 'Modern European Restaurant',
                 'New American Restaurant', 'Tea Room', 'Middle Eastern Restaurant',
                 'Chinese Restaurant', 'Ethiopian Restaurant', 'Seafood Restaurant',
                 'Bubble Tea Shop', 'Wine Bar', 'Ice Cream Shop', 'Poutine Place',
                 'Grocery Store', 'Supermarket', 'Dim Sum Restaurant', 'Food Truck',
                 'BBQ Joint', 'American Restaurant',
                 'Vegetarian / Vegan Restaurant', 'Fish Market',
                 'German Restaurant', 'Comfort Food Restaurant',
                 'Belgian Restaurant', 'Moroccan Restaurant', 'Bistro',
                 'Liquor Store', 'Donut Shop', 'Health Food Store', 'Cheese Shop',
                 'Greek Restaurant', 'Eastern European Restaurant', 'Gourmet Shop',
                 'Bagel Shop', 'Indian Restaurant', 'Juice Bar',
                 'Korean BBQ Restaurant', 'Fish & Chips Shop', 'Brewery',
                 'Poke Place', 'Falafel Restaurant', 'Salad Place',
                 'Korean Restaurant', 'Hakka Restaurant',
                 'Mediterranean Restaurant', 'Deli / Bodega', 'Frozen Yogurt Shop',
                 'Colombian Restaurant', 'Brazilian Restaurant',
                 'Gluten-free Restaurant', 'Noodle House',
                 'Latin American Restaurant', 'Cupcake Shop', 'Soup Place',
                 'Food Court', 'Cuban Restaurant', 'Fruit & Vegetable Store',
                 'Tibetan Restaurant', 'Taco Place',
                 'Molecular Gastronomy Restaurant', 'Butcher', 'Turkish Restaurant',
                 'Food Service', 'Cajun / Creole Restaurant', 'Organic Grocery',
                 'Dumpling Restaurant', 'Doner Restaurant', 'Filipino Restaurant',
                 'Taiwanese Restaurant', 'Snack Place', 'Theme Restaurant'],
                dtype=object)
```

Some venue categories were determined to be out of alignment with the project goal and were then removed from the dataset. The categories included:

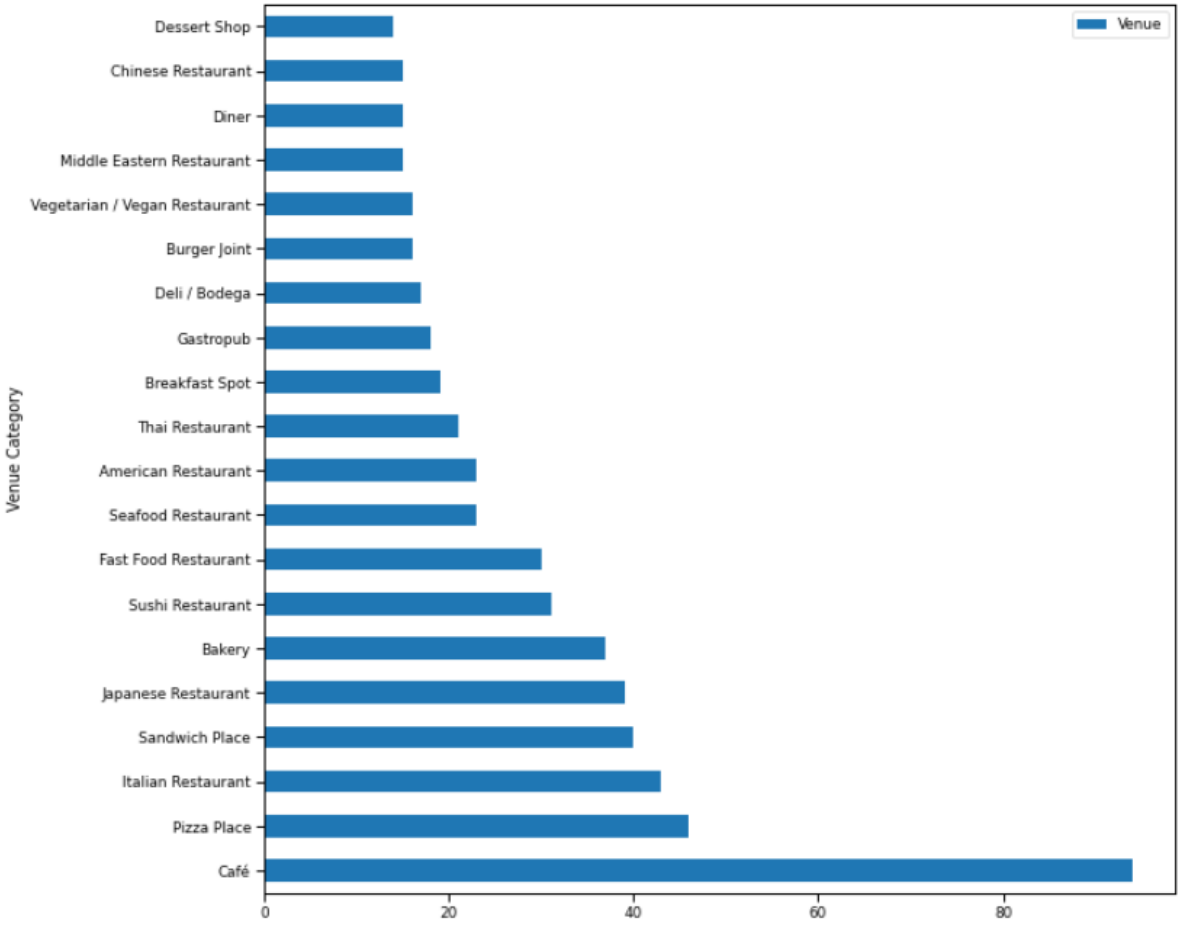
- Grocery Store
- Restaurant (as this category was too vague for this analysis)
- Food & Drink Shop
- Liquor Store
- Butcher
- Organic Grocery
- Supermarket

The venue category of ‘Coffee Shop’ was also removed after the analysis was completed, and then rerun because of the overwhelming number of Coffee Shops identified. With the goal of the project being to identify different places to eat by type of food, eliminating Coffee Shops was deemed to be in alignment with the project. This reduced the number of venues from 2,106 to 941.

```
In [491]: 1 toronto_food.shape #after categories removed
Out[491]: (941, 7)
```

Exploratory Data Analysis

An initial observation was conducted to identify the count of venues by venue category. graph below shows the top 20 venue categories. Cafes outnumbered any other venue category.



One hot encoding was used to determine the frequency of visits to a venue category by neighborhood to be used in determining the most popular venues for that neighborhood. An example of the data output is below.

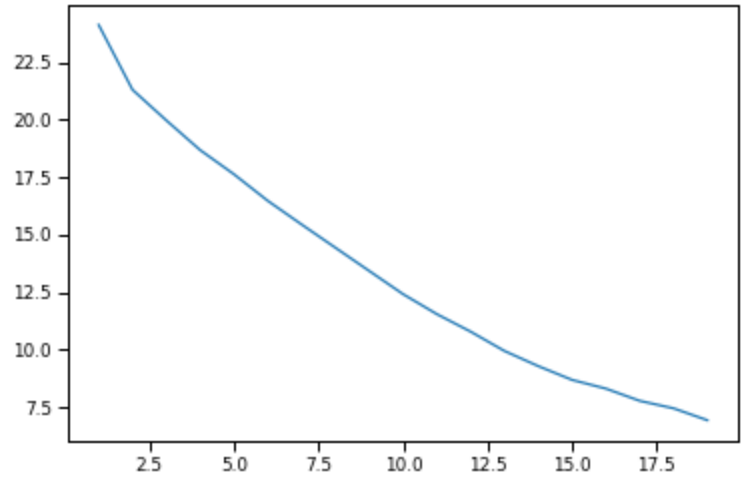
	Neighborhood	American Restaurant	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	Belgian Restaurant	Bistro	Brazilian Restaurant	Breakfast Spot	Brewery	Bubble Tea Shop	Burger Joint	Burrito Place	Café	C C Resta
0	Agincourt	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.000000	
1	Alderwood, Long Branch	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	
3	Bayview Village	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.333333	
4	Bedford Park, Lawrence Manor East	0.066667	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.066667	

The data was then sorted to identify the top ten most frequented venue categories per neighborhood.

	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	Latin American Restaurant	Breakfast Spot	Wings Joint	Farmers Market	Dim Sum Restaurant	Diner	Doner Restaurant	Donut Shop	Dumpling Restaurant	Eastern European Restaurant
1	Alderwood, Long Branch	Pizza Place	Sandwich Place	Ethiopian Restaurant	Deli / Bodega	Dessert Shop	Dim Sum Restaurant	Diner	Doner Restaurant	Donut Shop	Dumpling Restaurant
2	Bathurst Manor, Wilson Heights, Downsview North	Fried Chicken Joint	Ice Cream Shop	Pizza Place	Middle Eastern Restaurant	Diner	Sandwich Place	Sushi Restaurant	Deli / Bodega	Frozen Yogurt Shop	Filipino Restaurant
3	Bayview Village	Chinese Restaurant	Japanese Restaurant	Café	Wings Joint	Falafel Restaurant	Dim Sum Restaurant	Diner	Doner Restaurant	Donut Shop	Dumpling Restaurant
4	Bedford Park, Lawrence Manor East	Italian Restaurant	Sandwich Place	American Restaurant	Sushi Restaurant	Fast Food Restaurant	Greek Restaurant	Indian Restaurant	Juice Bar	Cupcake Shop	Pizza Place
5	Berczy Park	Farmers Market	Bakery	Cheese Shop	Seafood Restaurant	Comfort Food	Juice Bar	Japanese Restaurant	Creperie	Indian Restaurant	Eastern European

Machine Learning

1. K-Means clustering is a type of unsupervised machine learning that is used to identify groups or clusters in the data based on feature similarity. K-Means was used to build clusters based on the most frequented venue categories. It is important to identify the number of clusters (k) to segment the data. This requirement is not always a straight-forward task. In this analysis, determining the best option to use for k segmentation was attempted using the Elbow method. This method involves identifying the point of the elbow in which the inertia, which is the sum of squared distances of samples to their closest cluster center, starts to decrease in a linear manner. The result was continuous in this case, as displayed in the figure below, and was not helpful in identifying the best option for k.



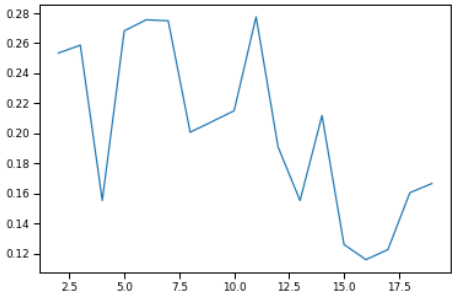
2. The Silhouette method was then used to attempt to identify the best number to use for k. There were multiple peaks, but after printing out the analysis numbers, the two largest peaks occurred when k was 6 and 11. The k chosen was 6 with a silhouette score of: 0.2754953671581569.

.

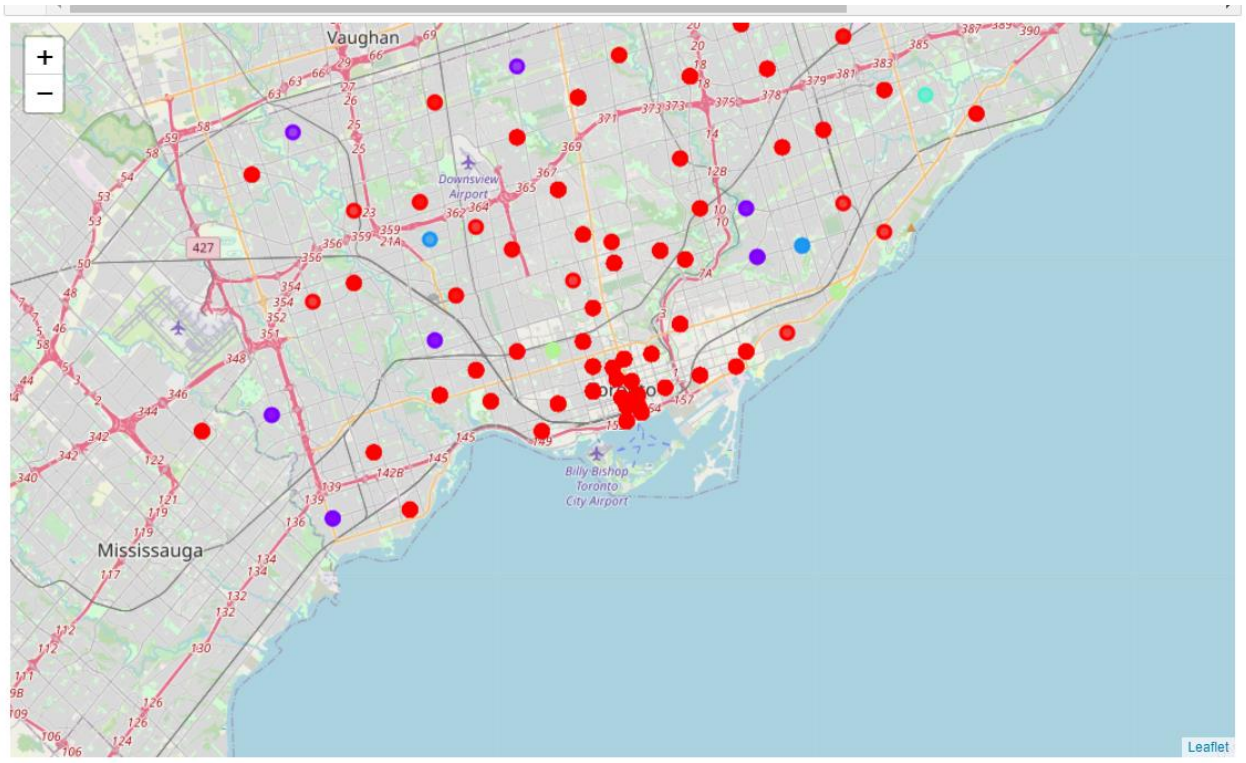
.]:
```
ss = []
k_ss = range(2,20)
for k in k_ss:
    kmeans = KMeans(n_clusters=k, random_state=0).fit(toronto_grouped_clustering)
    labels=kmeans.labels_
    ss.append(silhouette_score(toronto_grouped_clustering,labels,metric='euclidean'))

plt.plot(k_ss,ss)
#highest point is 3,6
print(k_ss,ss)
```

```
range(2, 20) [0.2534296855699702, 0.2586267078197195, 0.15543355956929764, 0.2681906188953071, 0.2754953671581569, 0.2748351568
6436366, 0.2007791832274612, 0.20784197773548915, 0.21503786203438893, 0.277492862125123, 0.1910118680870649, 0.155332734606342
73, 0.2118292025471107, 0.1261422930936441, 0.11607782678935158, 0.12273561344753234, 0.1605971090165302, 0.16664658013038022]
```



The clusters were then mapped using the Folium tool again with the color coding representing each cluster.



Results

Cluster 0

The analysis for Cluster 0 is below, displaying the top 5 venue categories that were identified as the first and second most common venue for the cluster. This cluster contains a high number of Cafes, Japanese Restaurants, and Italian Restaurants identified as first and second most common venues. This cluster could be relabeled to be identified as ‘Cafes, Japanese, and Italian’ cuisine.

```
*****Top 5 for categories in 1st and 2nd most common for Cluster: 0
     Cluster Labels 1st Most Common Venue  count
6                 0                 Café    398
28                0      Sushi Restaurant     71
26                0        Sandwich Place     50
27                0    Seafood Restaurant     44
4                 0       Bubble Tea Shop     42
     Cluster Labels  2nd Most Common Venue  count
18                0    Japanese Restaurant    202
17                0     Italian Restaurant    155
24                0  Vietnamese Restaurant     74
2                 0                Bakery     66
6                 0                 Café      63

**************************
```

Cluster 1

The analysis for Cluster 1 is below, displaying the top 5 venue categories that were identified as the first and second most common venue for the cluster. Pizza Places, Gastropubs, and Sandwich places were identified the strongest as first and second most common venues. This cluster could be relabeled to be identified as 'Pizza, pub, and sandwiches' cuisine.

```
*****Top 5 for categories in 1st and 2nd most common for Cluster: 1
     Cluster Labels 1st Most Common Venue  count
32                1           Pizza Place     14
     Cluster Labels 2nd Most Common Venue  count
29                1             Gastropub      3
31                1        Sandwich Place      3
26                1                 Café       2
27                1  Caribbean Restaurant      2
28                1    Falafel Restaurant      2

**************************
```

Cluster 2

There were few venue categories identified in Cluster 2. The first most common venue category identified was Bakery with 4 venues and Ice Cream Shop as the second most common venue category with 3. This cluster could be relabeled as 'Bakery and Ice Cream'.

```
*****Top 5 for categories in 1st and 2nd most common for Cluster: 2
     Cluster Labels 1st Most Common Venue  count
33               2                Bakery      4
     Cluster Labels 2nd Most Common Venue  count
32               2         Ice Cream Shop      3
33               2            Wings Joint      1

**************************
```

Cluster 3

Cluster 3 produced one result for the first most common venue category, ‘Korean BBQ Restaurant’ and one result for the second most common as ‘Wings Joint’. This cluster could be relabeled ‘Korean BBQ and Wings’.

```
*****Top 5 for categories in 1st and 2nd most common for Cluster: 3
     Cluster Labels  1st Most Common Venue  count
34               3  Korean BBQ Restaurant      1
     Cluster Labels 2nd Most Common Venue  count
34               3           Wings Joint      1

**************************
```

Cluster 4

Few venue categories were identified in Cluster 4 with the first most common venue as Cafe with 5 venues and Italian Restaurant as the second most common venue category with 4. This cluster could be relabeled as ‘Café and Italian’.

```
*****Top 5 for categories in 1st and 2nd most common for Cluster: 4
     Cluster Labels 1st Most Common Venue  count
35               4                 Café      5
     Cluster Labels 2nd Most Common Venue  count
35               4    Italian Restaurant      4
36               4           Wings Joint      1

**************************
```

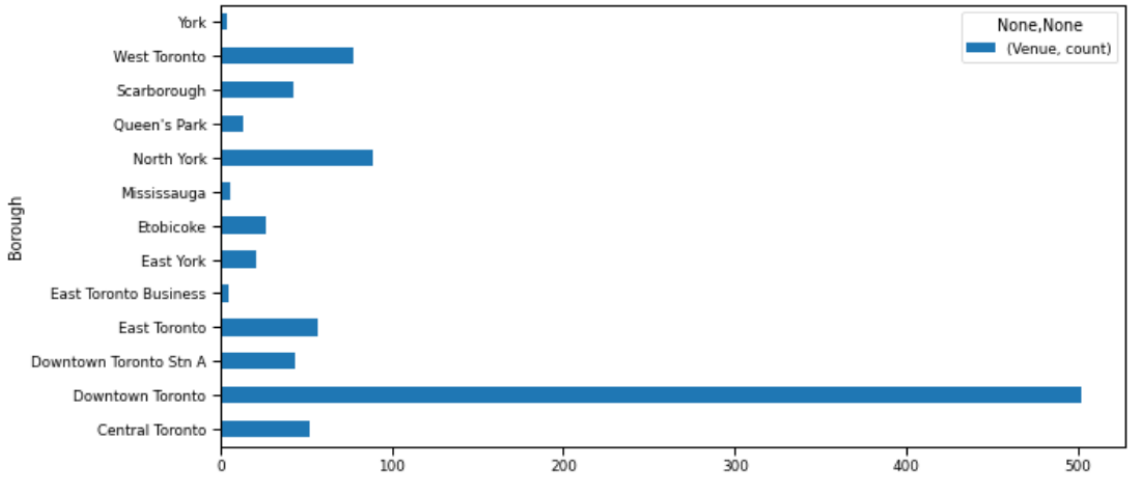
Cluster 5

Few venue categories were identified in Cluster 5 with the first most common venue as Mediterranean with 2 venues and Fast Food with 1. Fast Food was also identified as a top second most common venue category with 2. This cluster could be relabeled as 'Mediterranean and Fast Food.

```
*****Top 5 for categories in 1st and 2nd most common for Cluster: 5
     Cluster Labels     1st Most Common Venue  count
37                5  Mediterranean Restaurant      2
36                5      Fast Food Restaurant      1
     Cluster Labels 2nd Most Common Venue  count
37                5  Fast Food Restaurant      2
38                5           Wings Joint      1

**************************
```

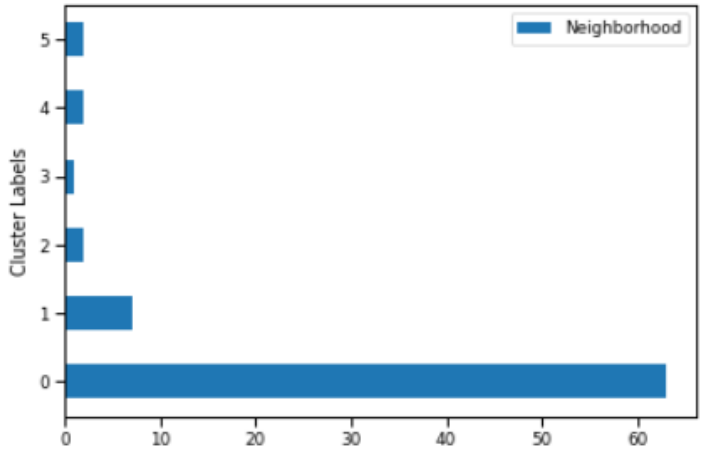
The number of venues showed that the majority were located in the Downtown Toronto borough.



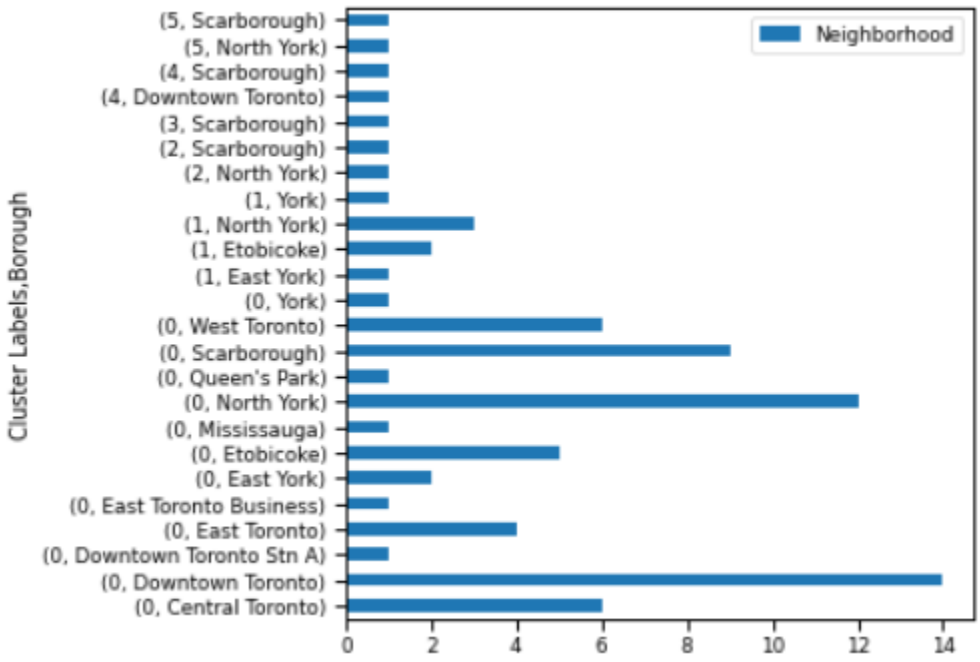
The data was summarized to identify the number of neighborhoods within each cluster. Cluster 0 contained the most number of neighborhoods, which was expected after seeing the results of the cluster analysis for Cluster 0, which contained the majority of venues over any other cluster.

```
In [528]:  1 clus_neigh = tor_food_borough[['Neighborhood','Cluster Labels']]
           2 clus_neigh.groupby(['Cluster Labels']).nunique().plot(kind='barh')
```

```
Out[528]: <AxesSubplot:ylabel='Cluster Labels'>
```



The graph below displays the number of neighborhoods within each cluster, borough. Downtown Toronto, North York, and Scarborough in Cluster 0 had the most number of neighborhoods assigned.



Discussion

Cafes were identified as one of the most frequented category venues, along with Pizza Places, Italian Restaurants, Sandwich Places, Japanese Restaurants, and Bakeries. This was interesting as there are very few Cafes or Japanese Restaurants in this writer's city, which is not located in Canada. Tourists or visitors looking for these types of venues will have many options that are highly frequented. For those considering opening a new restaurant of one of these venue categories, further analysis would be recommended. One could argue that these venue categories are the most successful in the Toronto area, and thus opening a new one of the same venue category could be profitable, as they are in high demand. On the other hand, one could argue that there would be an opportunity to open a different type of venue not identified to offer consumers a new or different food category option that may not be represented as highly.

Venue Category	Neighborhood
Café	94
Pizza Place	46
Italian Restaurant	43
Sandwich Place	40
Japanese Restaurant	39
Bakery	37
Sushi Restaurant	31
Fast Food Restaurant	30
Seafood Restaurant	23
American Restaurant	23
Thai Restaurant	21
Breakfast Spot	19
Gastropub	18
Deli / Bodega	17
Burger Joint	16
Vegetarian / Vegan Restaurant	16
Middle Eastern Restaurant	15
Diner	15
Chinese Restaurant	15
Dessert Shop	14

Conclusion

Using a clustering algorithm such as K-Means can provide an opportunity to explain, understand, and visualize a dataset and the similarity among the different data points. In this analysis, different clusters were identified based on the most frequented food venue category in the Toronto area, using the FourSquare API. With the majority of food venues occurring in the Downtown Toronto borough, it may have been helpful to either exclude this from the analysis, or to focus solely on that borough. That would be an opportunity for improvement of this study, or an opportunity for a future analysis.