Capstone Project: Italian Restaurants in Toronto

Mathias Weingärtner

14.09.2020

Introduction:

As part of the final assignsment of the Capstone course Applied Data Science I set up the following business case.

Business Case:

A manager of an italian restaurant chain (higher middle price category) wants to extend her business in Toronto. For the first flagstore she is looking for a neighborhood where it seems to be interesting to set up the location. Therefore she wants to know if there is a need for another new italian restaurant. If there is a need she wants to know the best neighborhood or borough to establish the restaurant. She is asking a market research agency to get the relevant information she can base her information on.

Key Question:

Before starting a detailed market research with customers of existing italian restaurants the market research agency wants to identify if there are areas which have a need for a new middle price category restaurants and, if so, which areas should be included in the detailed market survey.

Analytic Approach:

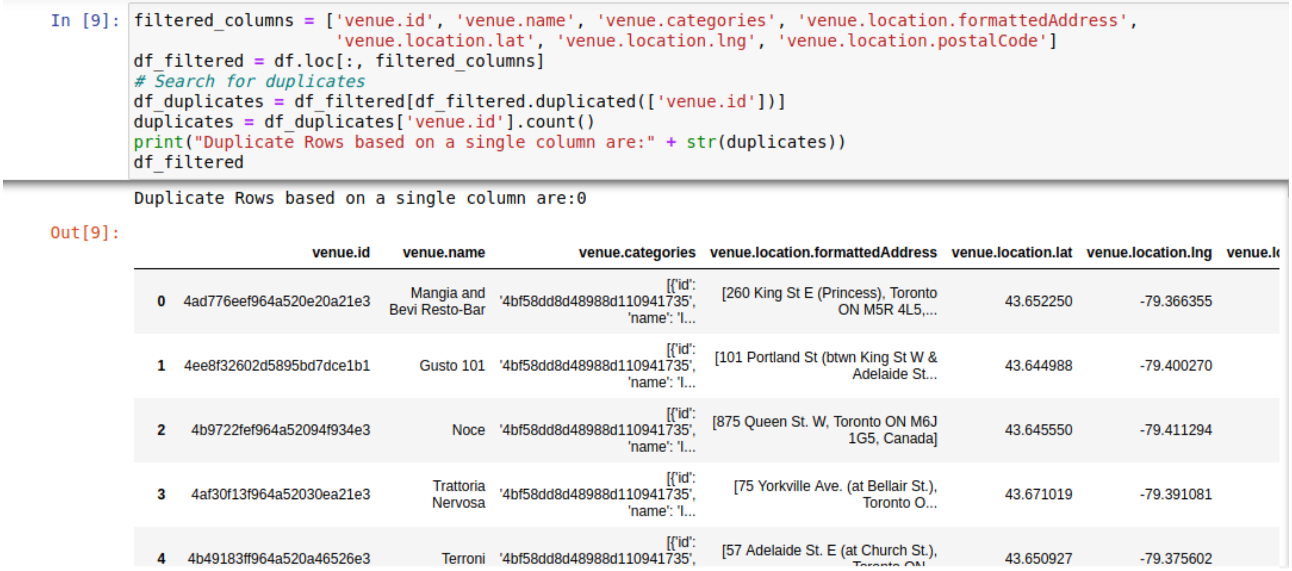
The market research agency assumes that first of all it has to identify areas of Toronto where italian restaurants are still existing. The idea behind this approach is that italian restaurants are established in Toronto for a long time. Therefore only in areas, where they are still existing, there is a need for italian restaurant food. In all other areas the italian food has no chance to be established since, if tried before, they did not succeed in surviving. In the next steps the relevant italian restaurants venues have to be clustered by rating and price category. The result has to be represented in a visualization and areas with clusters of a small amount of middle price catagory italian restaurants or a middle to high amount of middle price category restaurants with bad ratings has to be defined. In this areas the market research company will start their field research.

Data Requirements:

To cluster italian restaurants and to localize them in areas we need a representative list of italian restaurants with price categories and ratings and their coordinates. For mapping the coordinates with neighborhoods and boroughs we need a list with geospital data of Toronto. We assume for this exercise that in the Foursquare application we find a representative view of italian restaurants in Toronto (out of a statistical view we could verify this assumption for example by comparing the amount of italian restaurants in Toronto represented in Foursquare to the total of registered italian restaunrants in Toronto.) For the analytic part we use the venue and location information we get from the Forsquare api and combine it with the neighborhood and borough coordinates we receive from the prepared and downloaded csv file: 'Geospatial_Coordinates.csv'.

Methodology:

In a first step I use the Foursquare api to explore all restaurants in Toronto categorized as „Italaian restaurants“. Therefore I have to create a Foursquare endpoint „venues“ explore request and set up a pandas dataframe with the results.



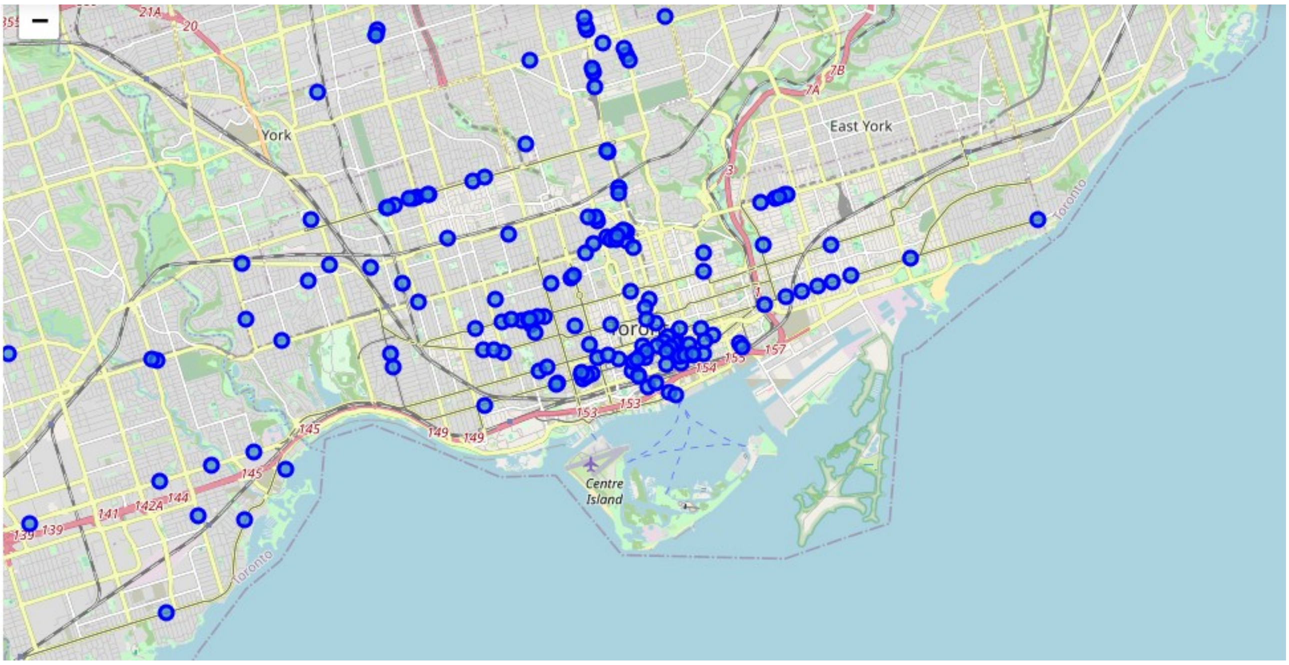
```
In [9]: filtered_columns = ['venue.id', 'venue.name', 'venue.categories', 'venue.location.formattedAddress',
                            'venue.location.lat', 'venue.location.lng', 'venue.location.postalCode']
        df_filtered = df.loc[:, filtered_columns]
        # Search for duplicates
        df_duplicates = df_filtered[df_filtered.duplicated(['venue.id'])]
        duplicates = df_duplicates['venue.id'].count()
        print("Duplicate Rows based on a single column are:" + str(duplicates))
        df_filtered
```

```
        Duplicate Rows based on a single column are:0
```

Out[9]:

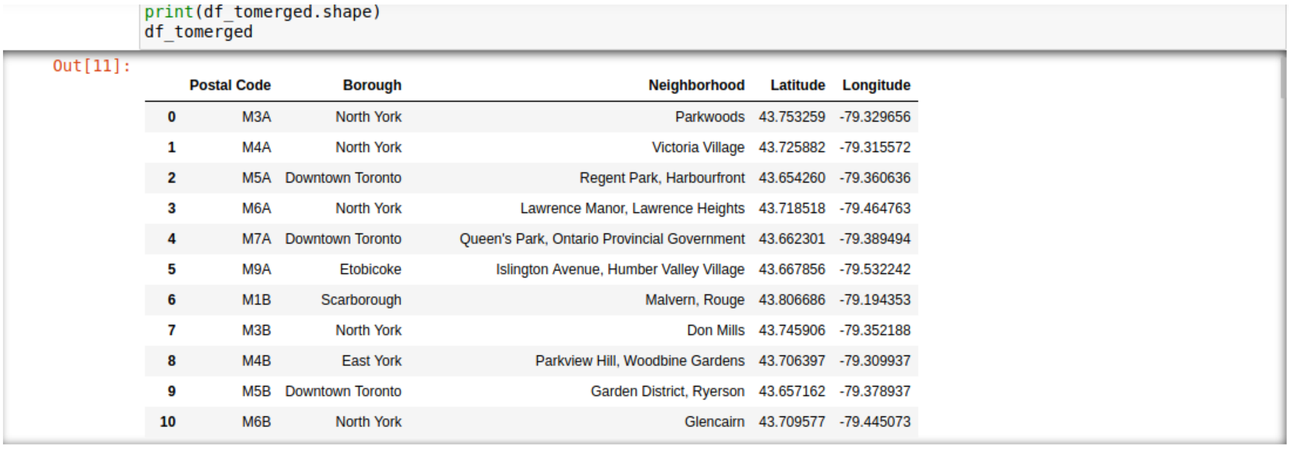
	venue.id	venue.name	venue.categories	venue.location.formattedAddress	venue.location.lat	venue.location.lng	venue.l
0	4ad776eef964a520e20a21e3	Mangia and Bevi Resto-Bar	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	[260 King St E (Princess), Toronto ON M5R 4L5,...	43.652250	-79.366355	
1	4ee8f32602d5895bd7dce1b1	Gusto 101	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	[101 Portland St (btwn King St W & Adelaide St...	43.644988	-79.400270	
2	4b9722fef964a52094f934e3	Noce	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	[875 Queen St. W, Toronto ON M6J 1G5, Canada]	43.645550	-79.411294	
3	4af30f13f964a52030ea21e3	Trattoria Nervosa	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	[75 Yorkville Ave. (at Bellair St.), Toronto O...	43.671019	-79.391081	
4	4b49183ff964a520a46526e3	Terroni	[{'id': '4bf58dd8d48988d110941735',	[57 Adelaide St. E (at Church St.), Toronto ON	43.650927	-79.375602	

I use the folium library and visualize the position of the restaurants in Toronto.



To combine neighborhood data with restaurant data I upload a geolocation data file of Toronto provided by the Coursera capstone course (see also data in my github account) and combine the data with the neighborhood information from wikipedia:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

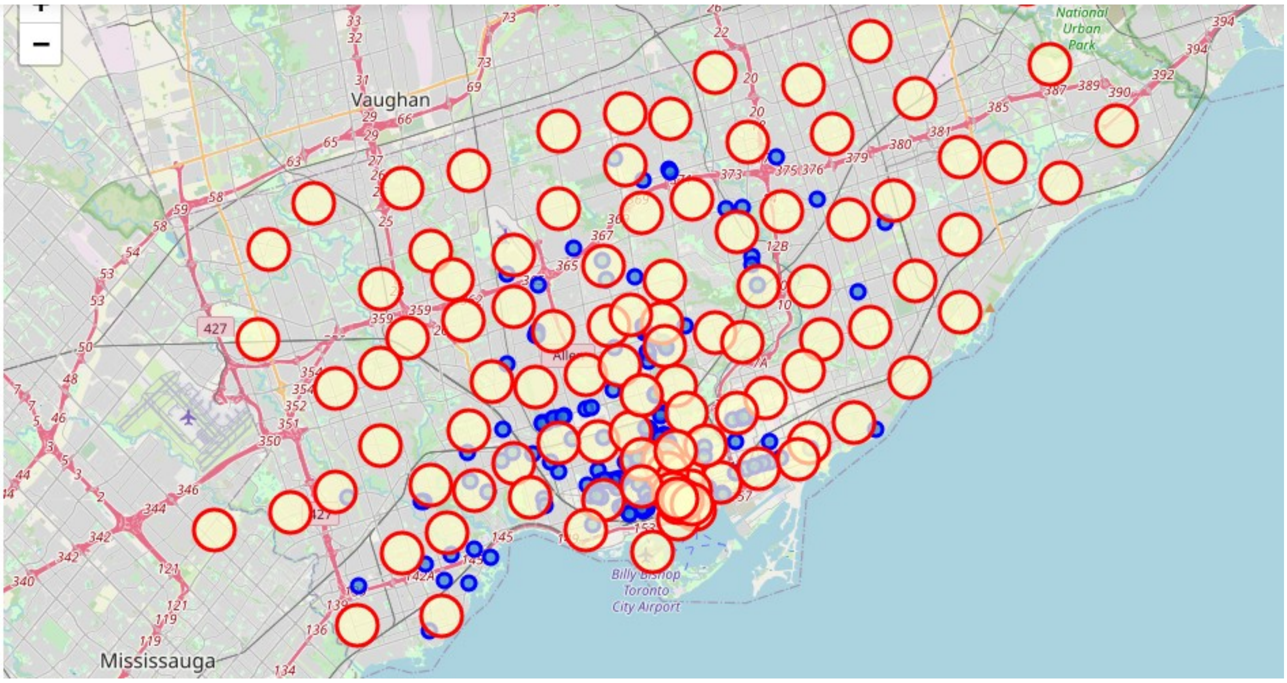


```
print(df_tomerged.shape)
df_tomerged
```

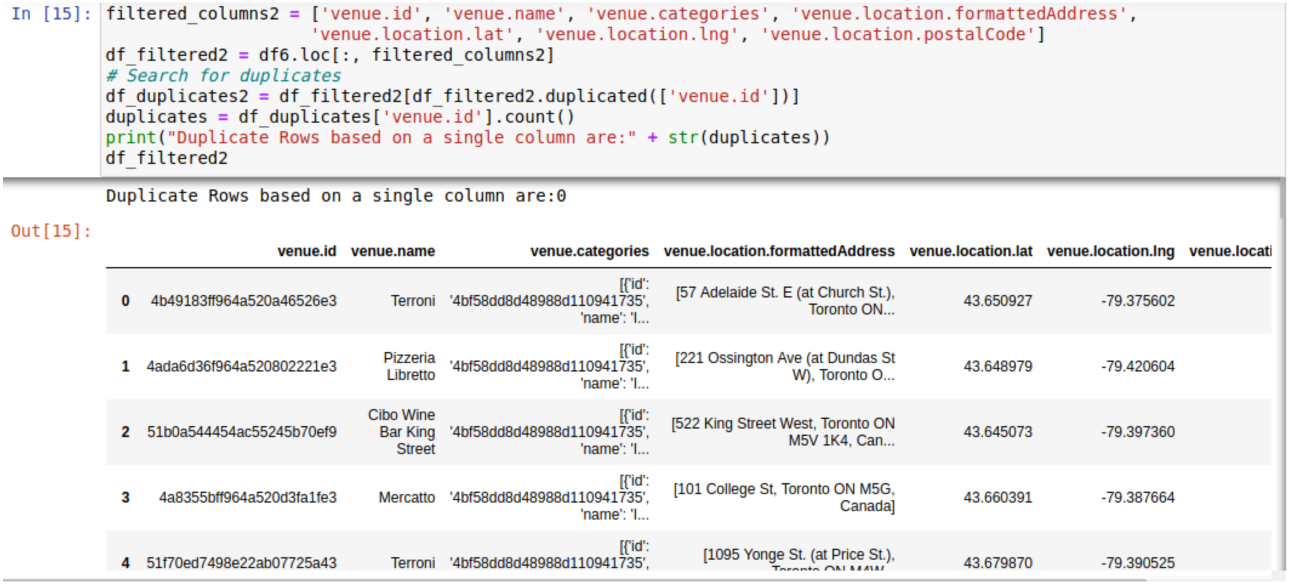
Out[11]:

	Postal Code	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
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills	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

Again I visualize neigborhoods and restaurants with folium to exclude neighborhoods on the first view.



In the next steps I use again the Foursquare api to explore venues with category „Italian restaurant“ and a higher middle price category. Again I set up a pandas dataframe.



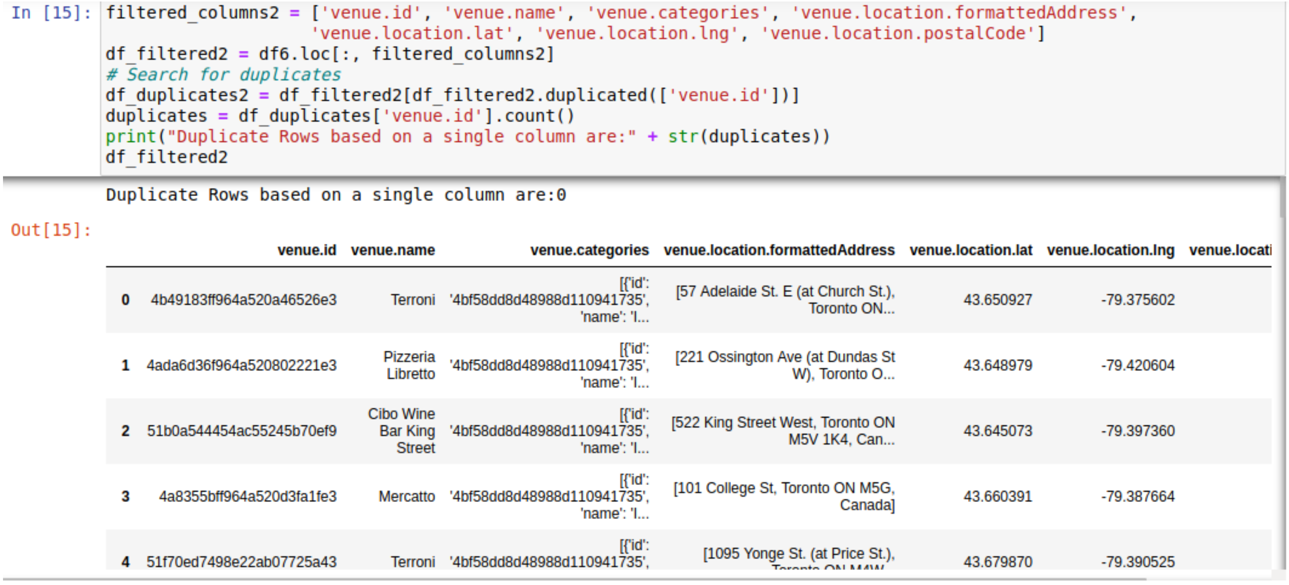
```
In [15]: filtered_columns2 = ['venue.id', 'venue.name', 'venue.categories', 'venue.location.formattedAddress',
                              'venue.location.lat', 'venue.location.lng', 'venue.location.postalCode']
         df_filtered2 = df6.loc[:, filtered_columns2]
         # Search for duplicates
         df_duplicates2 = df_filtered2[df_filtered2.duplicated(['venue.id'])]
         duplicates = df_duplicates['venue.id'].count()
         print("Duplicate Rows based on a single column are:" + str(duplicates))
         df_filtered2
```

```
         Duplicate Rows based on a single column are:0
Out[15]:
```

	venue.id	venue.name	venue.categories	venue.location.formattedAddress	venue.location.lat	venue.location.lng	venue.locati
0	4b49183ff964a520a46526e3	Terroni	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	[57 Adelaide St. E (at Church St.), Toronto ON...	43.650927	-79.375602	
1	4ada6d36f964a520802221e3	Pizzeria Libretto	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	[221 Ossington Ave (at Dundas St W), Toronto O...	43.648979	-79.420604	
2	51b0a544454ac55245b70ef9	Cibo Wine Bar King Street	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	[522 King Street West, Toronto ON M5V 1K4, Can...	43.645073	-79.397360	
3	4a8355bff964a520d3fa1fe3	Mercatto	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	[101 College St, Toronto ON M5G, Canada]	43.660391	-79.387664	
4	51f70ed7498e22ab07725a43	Terroni	[{'id': '4bf58dd8d48988d110941735',	[1095 Yonge St. (at Price St.),	43.679870	-79.390525	

Since I want to cluster the 48 hits of high middle price catgorized „Italian Restaurants“ by rating I have to get the details data for every single venue id.

With results I set up again a filtered pandas dataframe



```
In [15]: filtered_columns2 = ['venue.id', 'venue.name', 'venue.categories', 'venue.location.formattedAddress',
                              'venue.location.lat', 'venue.location.lng', 'venue.location.postalCode']
         df_filtered2 = df6.loc[:, filtered_columns2]
         # Search for duplicates
         df_duplicates2 = df_filtered2[df_filtered2.duplicated(['venue.id'])]
         duplicates = df_duplicates['venue.id'].count()
         print("Duplicate Rows based on a single column are:" + str(duplicates))
         df_filtered2
```

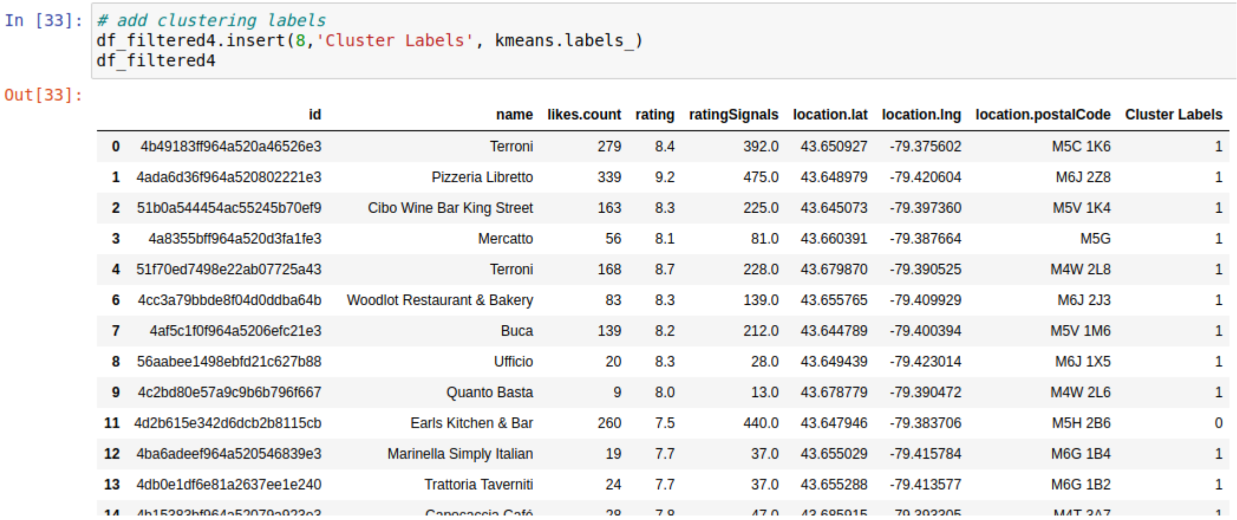
```
         Duplicate Rows based on a single column are:0
Out[15]:
```

	venue.id	venue.name	venue.categories	venue.location.formattedAddress	venue.location.lat	venue.location.lng	venue.locati
0	4b49183ff964a520a46526e3	Terroni	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	[57 Adelaide St. E (at Church St.), Toronto ON...	43.650927	-79.375602	
1	4ada6d36f964a520802221e3	Pizzeria Libretto	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	[221 Ossington Ave (at Dundas St W), Toronto O...	43.648979	-79.420604	
2	51b0a544454ac55245b70ef9	Cibo Wine Bar King Street	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	[522 King Street West, Toronto ON M5V 1K4, Can...	43.645073	-79.397360	
3	4a8355bff964a520d3fa1fe3	Mercatto	[{'id': '4bf58dd8d48988d110941735', 'name': 'I...	[101 College St, Toronto ON M5G, Canada]	43.660391	-79.387664	
4	51f70ed7498e22ab07725a43	Terroni	[{'id': '4bf58dd8d48988d110941735',	[1095 Yonge St. (at Price St.),	43.679870	-79.390525	

To cluster the venues in two groups representing high rated and low rated restaurants I decided to use the Kmeans clustering algorithm from klearn.cluster with cluster number of 2. Before running Kmeans I cleaned up the data by deleting rows with NaN values (this reduces the result by 7 datasets, but is necessary because Kmeans need values to be executed on) and reduce the dataframe to id and ratings only.

Procedure results in two cluster with two groups one closer to values >7.6 and the other lower or equal to 7.6.

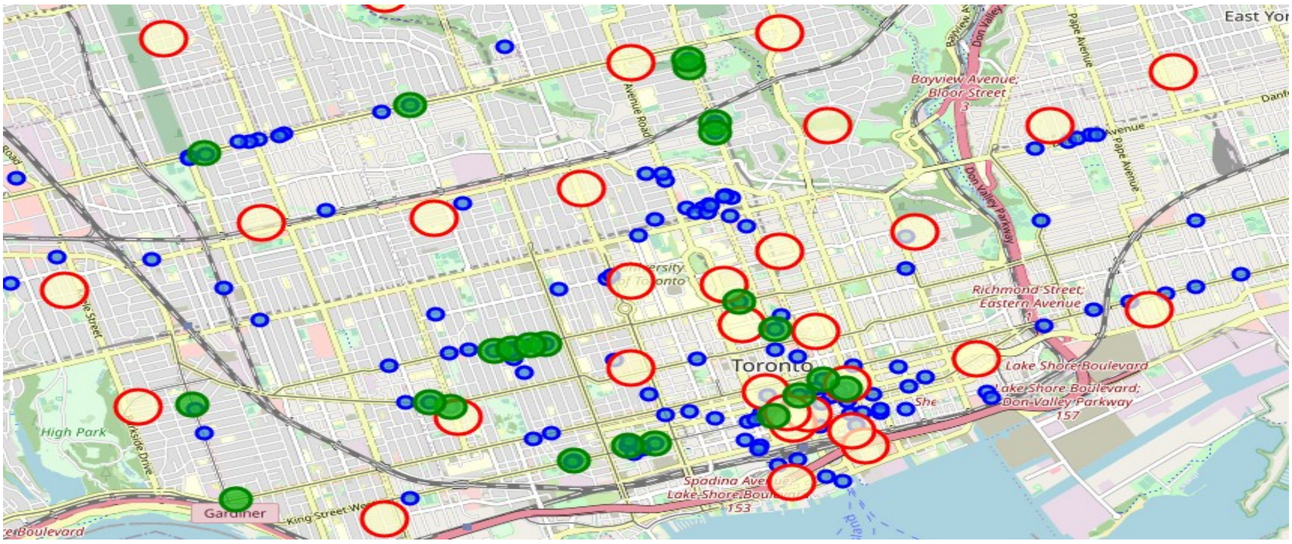
The ratings label are allocated to dataframe with data of high middle priced restaurants. The datframe is used to visualize the data with the geolocation data of Toronto boroughs.



In [33]:
```
# add clustering labels
df_filtered4.insert(8,'Cluster Labels', kmeans.labels_)
df_filtered4
```

Out[33]:

	id	name	likes.count	rating	ratingSignals	location.lat	location.lng	location.postalCode	Cluster Labels
0	4b49183ff964a520a46526e3	Terroni	279	8.4	392.0	43.650927	-79.375602	M5C 1K6	1
1	4ada6d36f964a520802221e3	Pizzeria Libretto	339	9.2	475.0	43.648979	-79.420604	M6J 2Z8	1
2	51b0a544454ac55245b70ef9	Cibo Wine Bar King Street	163	8.3	225.0	43.645073	-79.397360	M5V 1K4	1
3	4a8355bff964a520d3fa1fe3	Mercatto	56	8.1	81.0	43.660391	-79.387664	M5G	1
4	51f70ed7498e22ab07725a43	Terroni	168	8.7	228.0	43.679870	-79.390525	M4W 2L8	1
6	4cc3a79bbde8f04d0ddba64b	Woodlot Restaurant & Bakery	83	8.3	139.0	43.655765	-79.409929	M6J 2J3	1
7	4af5c1f0f964a5206efc21e3	Buca	139	8.2	212.0	43.644789	-79.400394	M5V 1M6	1
8	56aabee1498ebfd21c627b88	Ufficio	20	8.3	28.0	43.649439	-79.423014	M6J 1X5	1
9	4c2bd80e57a9c9b6b796f667	Quanto Basta	9	8.0	13.0	43.678779	-79.390472	M4W 2L6	1
11	4d2b615e342d6dcb2b8115cb	Earls Kitchen & Bar	260	7.5	440.0	43.647946	-79.383706	M5H 2B6	0
12	4ba6adeef964a520546839e3	Marinella Simply Italian	19	7.7	37.0	43.655029	-79.415784	M6G 1B4	1
13	4db0e1df6e81a2637ee1e240	Trattoria Taverniti	24	7.7	37.0	43.655288	-79.413577	M6G 1B2	1
14	4b15383bf964a52079a923e3	Capocaccia Café	28	7.8	47.0	43.685915	-79.393305	M4T 3A7	1



Results:

With the visualzation and the clustered restaurants we can answer the key question of the market research company. I identified the following neigborhoods / boroughs as suitable for the market research company starting a field research.

North York:low rating category 3 restaurants: esp. boroughs: **Bedford Park, Lawrence Manor East**

Central Toronto : low rating category 3 restaurants: esp. borough **Davisville**

Central Toronto: competition with one existing high rated restaurant: borough: **Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park**

Downtown Toronto and West Toronto are very well equipped with every type of Italian restaurant. So there seemed to be not a real opportunity for starting a new one.

Recommendation:

Start the field research in the following areas.

North York: low rating category 3 restaurants: esp. Boroughs: **Bedford Park, Lawrence Manor East**

Central Toronto low rating category 3 restaurants: esp. Borough **Davisville**

Central Toronto competition with one existing high rated restauran: Borough: **Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park**

In these areas there is a need for italian food, since italian restaurants are well established. In addition these areas have a lack of good rated middle priced (category 3) restaurants (**North York**: Boroughs: **Bedford Park, Lawrence Manor East** and **Central Toronto**: Borough **Davisville**) or there is only one competitor which indicates that the customers are willing to pay the price(**Central Toronto**: **Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park**).

Downtown Toronto and West Toronto are very well equipped with every type of Italian restaurant. So there seemed to be not a real opportunity for starting a new one.

Conclusion:

By using the Foursquare api, geolocation data and Kmeans clustering algorithm we can make a good guess where to start the field research in Toronto. This approach saves the market research company money because they could clearly segment their field of interest, do not need so much interview partner and can execute the field research faster. The final results can be delivered also faster to the end customer. The process of decision making is in total accelerated.

Sources:

Coursera Capstone Course Applied Datascience

Foursquare documnetation: https://developer.foursquare.com/docs/places-api/

Pandas documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html