Coursera Capstone

IBM Data Science Professional Certificate

Analysis of venues in the city of North York, Toronto

Mácio M. S. de Arruda
Dec, 2018

1. Introduction/Business Problem

The Brazilian tourism agency plans to organize their first trip to Toronto in Canada, and they plan to start with the city of North York.

To create the best itinerary for your clients, the travel agency is looking for which districts near North York visit, for that, he had commissioned an analysis of characteristics and the most relevant places in these neighborhoods.

Mainly, the tourism agency focuses on hotels, restaurants, parks, shops, places, squares, etc.

Question: So, what are the characteristics of the neighborhoods neighboring North York and what places should the Brazilian tourism agency visit to provide the best tour to its clients?

1.1. Context

Location data is now fetched from different open sources like: Wikipedia, Foursquare, Google. The exploration of this data through a data science methodology can help to extract from this data the information and insights for the benefit of those interested. To achieve good results, I will use knowledge learned throughout the course, such as data analysis, data visualization and Machine Learning.

This tourism agency has experience in several other countries and has always left its customers satisfied. But it is their first time organizing a trip to Canada. So, to achieve this goal and maintain your good reputation and acquire knowledge of locations in a new country, data science can be crucial for them to achieve their goals.

2. Description of the data

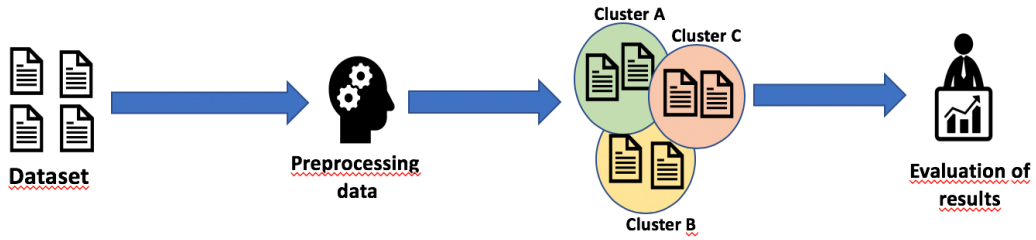
To solve this problem, I'll first get the data on the neighborhoods of North York (Toronto): Borough, Latitude, Longitude. This data is available on Wikipedia, so with a simple script, I was able to access the data in tabular format. Below is the example of this dataframe.

The second step, with the help of the API of Foursquare, got the information of the best places of the neighborhood and its neighbors. With a new dataframe with the complete information, you're all set to cluster the entire dataset.

17	M2H	North York	Hillcrest Village	43.803762	-79.363452
18	M2J	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556
19	M2K	North York	Bayview Village	43.786947	-79.385975
20	M2L	North York	Silver Hills, York Mills	43.757490	-79.374714
21	M2M	North York	Newtonbrook, Willowdale	43.789053	-79.408493
22	M2N	North York	Willowdale South	43.770120	-79.408493
23	M2P	North York	York Mills West	43.752758	-79.400049
24	M2R	North York	Willowdale West	43.782736	-79.442259
25	M3A	North York	Parkwoods	43.753259	-79.329656
26	M3B	North York	Don Mills North	43.745906	-79.352188
27	M3C	North York	Flemingdon Park, Don Mills South	43.725900	-79.340923
28	M3H	North York	Bathurst Manor, Downsview North, Wilson Heights	43.754328	-79.442259
29	M3J	North York	Northwood Park, York University	43.767980	-79.487262
30	M3K	North York	CFB Toronto, Downsview East	43.737473	-79.464763
31	M3L	North York	Downsview West	43.739015	-79.506944
32	M3M	North York	Downsview Central	43.728496	-79.495697
33	M3N	North York	Downsview Northwest	43.761631	-79.520999
34	M4A	North York	Victoria Village	43.725882	-79.315572
35	M4B	East York	Woodbine Gardens, Parkview Hill	43.706397	-79.309937
36	M4C	East York	Woodbine Heights	43.695344	-79.318389
37	M4E	East Toronto	The Beaches	43.676357	-79.293031
38	M4G	East York	Leaside	43.709060	-79.363452
39	M4H	East York	Thorncliffe Park	43.705369	-79.349372
40	M4J	East York	East Toronto	43.685347	-79.338106

3. Methodology

This section describes what steps have been taken during project development. Here you will find pictures about the workflow, maps, explanations of procedures performed and other relevant information.



(High level workflow. Mácio Arruda, 2018)

3.1. Toronto city code list

For the Toronto city code list, it was extracted from Wikipedia List of Postal Code through a Python script. After extracting the table from the Wikipedia link, a dataframe with the information extracted was assembled.

Get Wikipedia content with BeautifulSoup

```
url = "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
r = requests.get(url)
html = BeautifulSoup(r.content, "lxml")
```

Clean Not assigned and reset index

```
table = html.find_all('table')[0]
df = pd.read_html(str(table))[0].iloc[1:,].rename({0:"PostalCode",1:"Borough",2:"Neighborhood"},axis=1)
df = df[df.Borough!="Not assigned"]
df.reset_index(inplace = True, drop = True)
#df.drop('index',1,inplace=True)
df.head(12)
```

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M5A	Downtown Toronto	Regent Park
4	M6A	North York	Lawrence Heights
5	M6A	North York	Lawrence Manor
6	M7A	Queen's Park	Not assigned
7	M9A	Etobicoke	Islington Avenue
8	M1B	Scarborough	Rouge
9	M1B	Scarborough	Malvern
10	M3B	North York	Don Mills North
11	M4B	East York	Woodbine Gardens

Then, with postal codes in hand, I got the coordinate list of each item from the previous table.

```
url = "http://cocl.us/Geospatial_data/Geospatial_Coordinates.csv"
coor = pd.read_csv(url)
coor.columns = ['PostalCode', 'Latitude', 'Longitude']
coor.rename({'Postal Code':'PostalCode'},axis=1, inplace=True)
coor.head()
```

	PostalCode	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

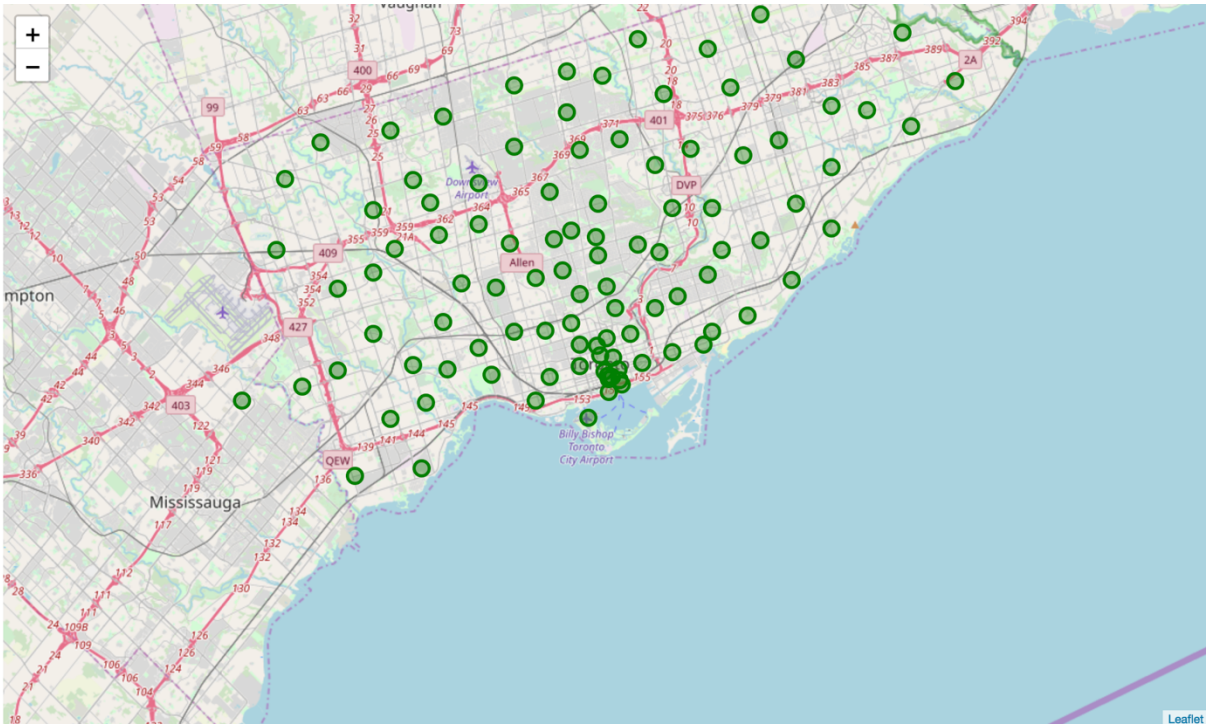
(Get coordinates)

Then the data was merged to compose the dataframe with neighborhood descriptions, neighbors and coordinates.

```
df_toronto = pd.read_csv('toronto_data.csv')
df_toronto.tail(50)
```

	Unnamed: 0	PostalCode	Borough	Neighborhood	Latitude	Longitude
53	53	M5A	Downtown Toronto	Harbourfront, Regent Park	43.654260	-79.360636
54	54	M5B	Downtown Toronto	Ryerson, Garden District	43.657162	-79.378937
55	55	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
56	56	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306
57	57	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383
58	58	M5H	Downtown Toronto	Adelaide, King, Richmond	43.650571	-79.384568
59	59	M5J	Downtown Toronto	Harbourfront East, Toronto Islands, Union Station	43.640816	-79.381752
60	60	M5K	Downtown Toronto	Design Exchange, Toronto Dominion Centre	43.647177	-79.381576
61	61	M5L	Downtown Toronto	Commerce Court, Victoria Hotel	43.648198	-79.379817
62	62	M5M	North York	Bedford Park, Lawrence Manor East	43.733283	-79.419750

Finally, a map was built with the previous dataframe and now it is possible to see Toronto's neighborhoods.



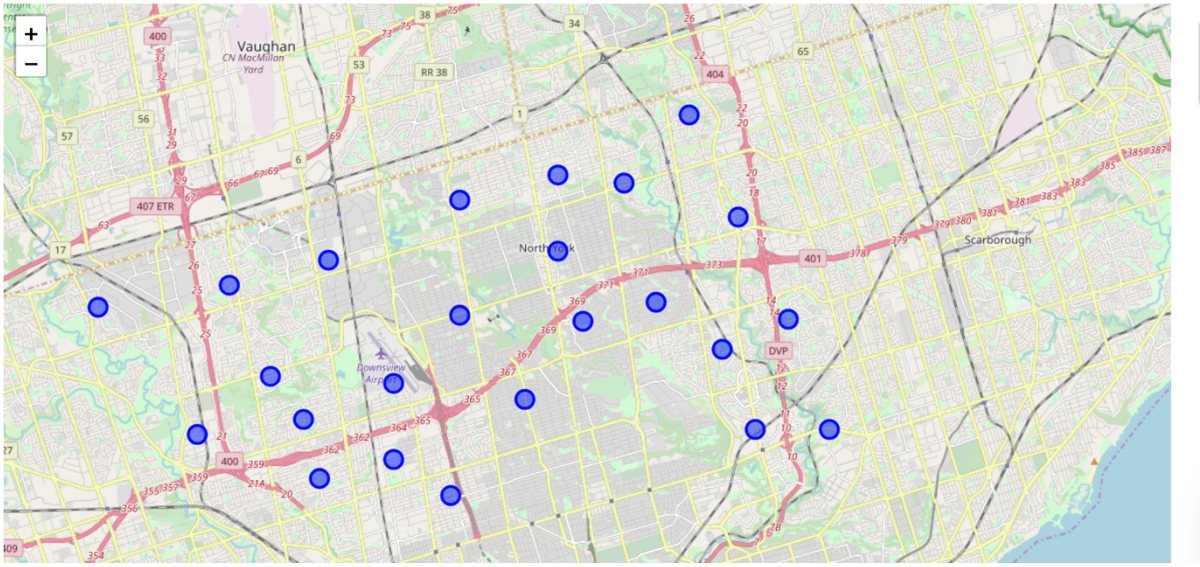
3.2. Create a map of North York and its neighborhoods

Below is a map of the city of North York and its neighborhoods, using the libraries: folium and pygeo:

```
address_nyork = 'North York,Toronto'
latitude_nyork = 43.773077
longitude_nyork = -79.257774
print('The geograpical coordinate of North York are {}, {}.'.format(latitude_nyork, longitude_nyork))
```

```
The geograpical coordinate of North York are 43.773077, -79.257774.
```

```
map_nyork = folium.Map(location=[latitude_nyork, longitude_nyork], zoom_start=13)

# add markers to map
for lat, lng, label in zip(nyork_data['Latitude'], nyork_data['Longitude'], nyork_data['Neighborhood']):
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=9,
        popup=label,
        color='blue',
        fill=True,
        fill_opacity=0.5).add_to(map_nyork)

map_nyork
```



3.3. Get the top 100 venues in the neighborhood 'Hillcrest Village', from North York

From the neighborhood Hillcrest Village, a neighborhood of North York, were selected the 100 best places of different types, within a radius of 1 km. With the help of the Foursquare API, I got the venue list.

```
neighborhood_latitude = nyork_data.loc[0, 'Latitude'] # neighbourhood latitude value
neighborhood_longitude = nyork_data.loc[0, 'Longitude'] # neighbourhood longitude value

neighborhood_name = nyork_data.loc[0, 'Neighborhood'] # neighbourhood name

print('Latitude and longitude values of "{}" are {}, {}.'.format(neighborhood_name,
                                                                 neighborhood_latitude,
                                                                 neighborhood_longitude))
```

```
Latitude and longitude values of "Hillcrest Village" are 43.8037622, -79.3634517.
```

```
LIMIT = 100
radius = 1000
url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&ll={},{}&v={}&radius={}&limit=
```

```
results = requests.get(url).json()
results
```


```
{'meta': {'code': 200, 'requestId': '5c28c99a4434b91ed84cd8c0'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Scarborough City Centre',
  'headerFullLocation': 'Scarborough City Centre, Toronto',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 62,
  'suggestedBounds': {'ne': {'lat': 43.78207700900001,
    'lng': -79.24533335909429},
   'sw': {'lat': 43.76407699099999, 'lng': -79.2702146409057}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
```


toronto_data.csv

From Foursquare's response, a dataframe was set up and then mergeed with the master dataframe previously assembled.

```
import json
from pandas.io.json import json_normalize

venues = results['response']['groups'][0]['items']
nearby_venues = json_normalize(venues)
filtered_columns = ['venue.name', 'venue.categories', 'venue.location.lat', 'venue.location.lng']
nearby_venues =nearby_venues.loc[:, filtered_columns]
nearby_venues['venue.categories'] = nearby_venues.apply(get_category_type, axis=1)
nearby_venues.columns = [col.split(".")[-1] for col in nearby_venues.columns]

nearby_venues.head(10)
```

	name	categories	lat	lng
0	Disney Store	Toy / Game Store	43.775537	-79.256833
1	Canyon Creek Chophouse	Steakhouse	43.776959	-79.261694
2	DAVIDsTEA	Tea Room	43.776613	-79.258516
3	Tommy Hilfiger Company Store	Clothing Store	43.776015	-79.257369
4	American Eagle Outfitters	Clothing Store	43.775908	-79.258352
5	St. Andrews Fish & Chips	Fish & Chips Shop	43.771865	-79.252645
6	Chipotle Mexican Grill	Mexican Restaurant	43.776410	-79.258069
7	Coliseum Scarborough Cinemas	Movie Theater	43.775995	-79.255649
8	SEPHORA	Cosmetics Shop	43.775592	-79.258242
9	Shoppers Drug Mart	Pharmacy	43.772747	-79.251123

3.4. Clustering the data with the KMeans algorithm

After all the work of aggregating data from several different sources, we generate the following dataframe with information of locations, neighborhoods, neighborhoods, latitude and longitude. We are now one step closer to being able to apply a clustering technique to the data.

```
nyork_venues.tail(10)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
225	Glencairn	43.709577	-79.445073	Glencairn Subway Station	43.708872	-79.440801	Metro Station
226	Glencairn	43.709577	-79.445073	Fraserwood Park	43.713550	-79.442482	Park
227	Maple Leaf Park, North Park, Upwood Park	43.713756	-79.490074	Rustic Bakery	43.715414	-79.490300	Bakery
228	Maple Leaf Park, North Park, Upwood Park	43.713756	-79.490074	Sporty	43.716503	-79.489838	Basketball Court
229	Maple Leaf Park, North Park, Upwood Park	43.713756	-79.490074	Maple leaf park	43.716188	-79.493531	Park
230	Maple Leaf Park, North Park, Upwood Park	43.713756	-79.490074	Mika's Trim	43.714068	-79.496113	Construction & Landscaping
231	Humber Summit	43.756303	-79.565963	Pizza Monza	43.755043	-79.567195	Pizza Place
232	Humber Summit	43.756303	-79.565963	The Famous Mama Mia	43.758820	-79.570637	Empanada Restaurant
233	Emery, Humberlea	43.724766	-79.532242	Strathburn Park	43.721765	-79.532854	Baseball Field
234	Emery, Humberlea	43.724766	-79.532242	Danbury Bankruptcy Sale	43.728140	-79.529923	Furniture / Home Store

Typically, machine learning techniques do not support non-numeric data, so we need to encode the data so that we can use some of the techniques. Below, we use the One Hot Encode technique to handle the data.

```
# one hot encoding
nyork_onehot = pd.get_dummies(nyork_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
nyork_onehot['Neighborhood'] = nyork_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [nyork_onehot.columns[-1]] + list(nyork_onehot.columns[:-1])
nyork_onehot = nyork_onehot[fixed_columns]

nyork_onehot.head(20)
```

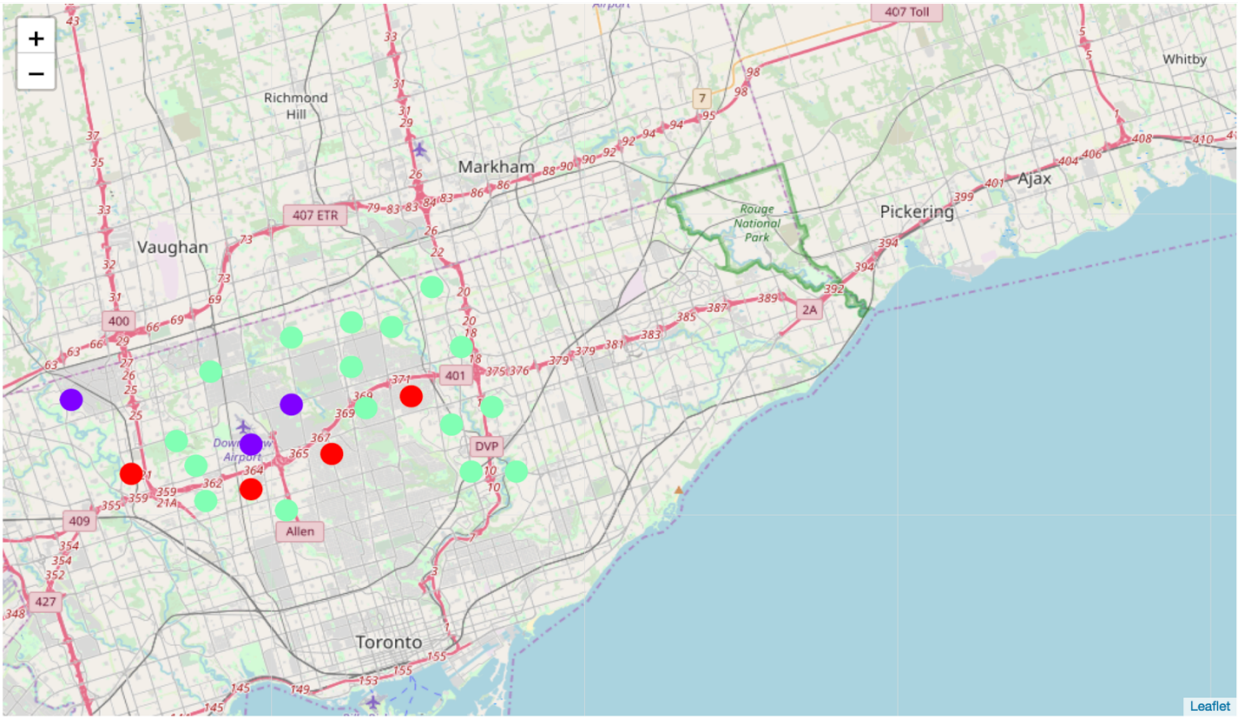
	Neighborhood	Accessories Store	Airport	American Restaurant	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Bakery	Bank	Bar	Baseball Field	Basketball Court	Beer Store	Bike Shop	Boutique	Bridal Shop
0	Hillcrest Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Hillcrest Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Hillcrest Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Hillcrest Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Hillcrest Village	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
5	Fairview, Henry Farm, Oriole	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Fairview, Henry Farm, Oriole	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Fairview, Henry Farm, Oriole	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
8	Fairview, Henry Farm, Oriole	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Fairview, Henry Farm, Oriole	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

We execute the kameans with k equal to 3 clusters.

```
# import k-means from clustering stage
from sklearn.cluster import KMeans

#nyork_data = nyork_data.drop(16)
# set number of clusters
kclusters = 3

nyork_grouped_clustering = nyork_grouped.drop('Neighborhood', 1)


# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(nyork_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

4. Results

Below, the map plotted with clusters generated by Kmeans. We can notice the similarity of several neighborhoods with regard to the category of venues.



Below, the dataframes of the 3 clusters obtained

Label 0:

```
nyork_merged.loc[nyork_merged['Cluster Labels'] == 0, nyork_merged.columns[[1] + list(range(5, nyork_merged.shape[1]))]]
```

	Borough	Cluster Labels	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
3	North York	0	Park	Martial Arts Dojo	Clothing Store	Coffee Shop	Comfort Food Restaurant	Construction & Landscaping	Cosmetics Shop	Deli / Bodega	Department Store	Dim Sum Restaurant
18	North York	0	Coffee Shop	Fast Food Restaurant	Italian Restaurant	Indian Restaurant	Comfort Food Restaurant	Café	Liquor Store	Pharmacy	Pizza Place	Juice Bar
19	North York	0	Clothing Store	Furniture / Home Store	Accessories Store	Event Space	Miscellaneous Shop	Boutique	Coffee Shop	Vietnamese Restaurant	Fried Chicken Joint	Food Truck
23	North York	0	Furniture / Home Store	Baseball Field	Women's Store	Empanada Restaurant	Comfort Food Restaurant	Construction & Landscaping	Cosmetics Shop	Deli / Bodega	Department Store	Dim Sum Restaurant

Label 1:

```
nyork_merged.loc[nyork_merged['Cluster Labels'] == 1, nyork_merged.columns[[1] + list(range(5, nyork_merged.shape[1]))]]
```

	Borough	Cluster Labels	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
11	North York	1	Coffee Shop	Middle Eastern Restaurant	Bank	Bridal Shop	Restaurant	Diner	Sandwich Place	Shopping Mall	Pizza Place	Pharmacy
13	North York	1	Park	Airport	Bus Stop	Electronics Store	Coffee Shop	Comfort Food Restaurant	Construction & Landscaping	Cosmetics Shop	Deli / Bodega	Department Store
22	North York	1	Pizza Place	Empanada Restaurant	Dog Run	Clothing Store	Coffee Shop	Comfort Food Restaurant	Construction & Landscaping	Cosmetics Shop	Deli / Bodega	Department Store

Label 2:

```
nyork_merged.loc[nyork_merged['Cluster Labels'] == 2, nyork_merged.columns[[1] + list(range(5, nyork_merged.shape[1]))]]
```

	Borough	Cluster Labels	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	North York	2	Dog Run	Golf Course	Pool	Athletics & Sports	Mediterranean Restaurant	Women's Store	Comfort Food Restaurant	Construction & Landscaping	Cosmetics Shop	Deli / Bodega
1	North York	2	Clothing Store	Fast Food Restaurant	Women's Store	Coffee Shop	Food Court	Restaurant	Tea Room	Kids Store	Bakery	Health Food Store
2	North York	2	Chinese Restaurant	Café	Bank	Japanese Restaurant	Electronics Store	Comfort Food Restaurant	Construction & Landscaping	Cosmetics Shop	Deli / Bodega	Department Store
4	North York	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	North York	2	Ramen Restaurant	Restaurant	Sandwich Place	Café	Fast Food Restaurant	Coffee Shop	Pizza Place	Japanese Restaurant	Middle Eastern Restaurant	Indonesian Restaurant
6	North York	2	Park	Bank	Electronics Store	Coffee Shop	Comfort Food Restaurant	Construction & Landscaping	Cosmetics Shop	Deli / Bodega	Department Store	Dim Sum Restaurant
7	North York	2	Pharmacy	Pizza Place	Coffee Shop	Butcher	Dog Run	Comfort Food Restaurant	Construction & Landscaping	Cosmetics Shop	Deli / Bodega	Department Store
8	North York	2	Park	Food & Drink Shop	Fast Food Restaurant	Women's Store	Dog Run	Coffee Shop	Comfort Food Restaurant	Construction & Landscaping	Cosmetics Shop	Deli / Bodega
9	North York	2	Caribbean Restaurant	Gym / Fitness Center	Café	Pool	Basketball Court	Japanese Restaurant	Dog Run	Comfort Food Restaurant	Construction & Landscaping	Cosmetics Shop
10	North York	2	Gym	Asian Restaurant	Coffee Shop	Beer Store	Bike Shop	Grocery Store	Fast Food Restaurant	Italian Restaurant	Japanese Restaurant	Dim Sum Restaurant
12	North York	2	Caribbean Restaurant	Furniture / Home Store	Miscellaneous Shop	Bar	Massage Studio	Coffee Shop	Women's Store	Dog Run	Construction & Landscaping	Cosmetics Shop
14	North York	2	Moving Target	Shopping Mall	Grocery Store	Bank	Coffee Shop	Comfort Food Restaurant	Construction & Landscaping	Cosmetics Shop	Deli / Bodega	Department Store
15	North York	2	Business Service	Food Truck	Baseball Field	Women's Store	Electronics Store	Comfort Food Restaurant	Construction & Landscaping	Cosmetics Shop	Deli / Bodega	Department Store
17	North York	2	Coffee Shop	Hockey Arena	Portuguese Restaurant	Intersection	Women's Store	Dog Run	Comfort Food Restaurant	Construction & Landscaping	Cosmetics Shop	Deli / Bodega
20	North York	2	Pizza Place	Metro Station	Pub	Japanese Restaurant	Park	Women's Store	Diner	Clothing Store	Coffee Shop	Comfort Food Restaurant
21	North York	2	Park	Construction & Landscaping	Bakery	Basketball Court	Electronics Store	Coffee Shop	Comfort Food Restaurant	Cosmetics Shop	Deli / Bodega	Department Store

5. Discussion

The main challenge of data science applications is certainly data manipulation. We need to get data from different sources, which are not always friendly and easy to manipulate.

When the data preparation step is unsuccessful, we need to choose the appropriate technique to be applied. The word is to experiment and analyze options.

Finally, the challenge of a data scientist is to organize the chronology of facts and tell the story. In the course, we learned that the data scientist, in addition to the various skills, need to be able to tell the story and persuade the person involved in the project.

6. Conclusion

This project analyzed data from different sources and was able to obtain results for the client. The applied process also proved that with a more comprehensive database, we could extrapolate the Toronto boundaries and get insightful insights from other places as well.

The challenge always ends up being: the data. It is not always possible to get them easily. But with deep research and the curiosity of a data scientist who wants to solve the customer's problem, the sky may be the limit.