IBM DATA SCIENCE CAPSTONE

Battle of the Neighborhoods

Jad Ghantous

Introduction/Business Problem:

People are interested in investing in businesses. Surely that is great since it promotes the economic activity in a country. Investing definitely requires money and that is why one must make sure the investment will have a return of profits. If we take a major city like Toronto do you think that all investments are created equally?

Using data science we shall find out what really works in Toronto so that anyone who is interested in investing in Toronto has an idea of what is popular and common and therefore visited. This will give any investor an idea of the category of the investment they should do.

Data:

The data was gathered were used to create and visualize the end result and that is the most common venues in Toronto using mainly 2 sources:

- Wikipedia
- Foursquare API

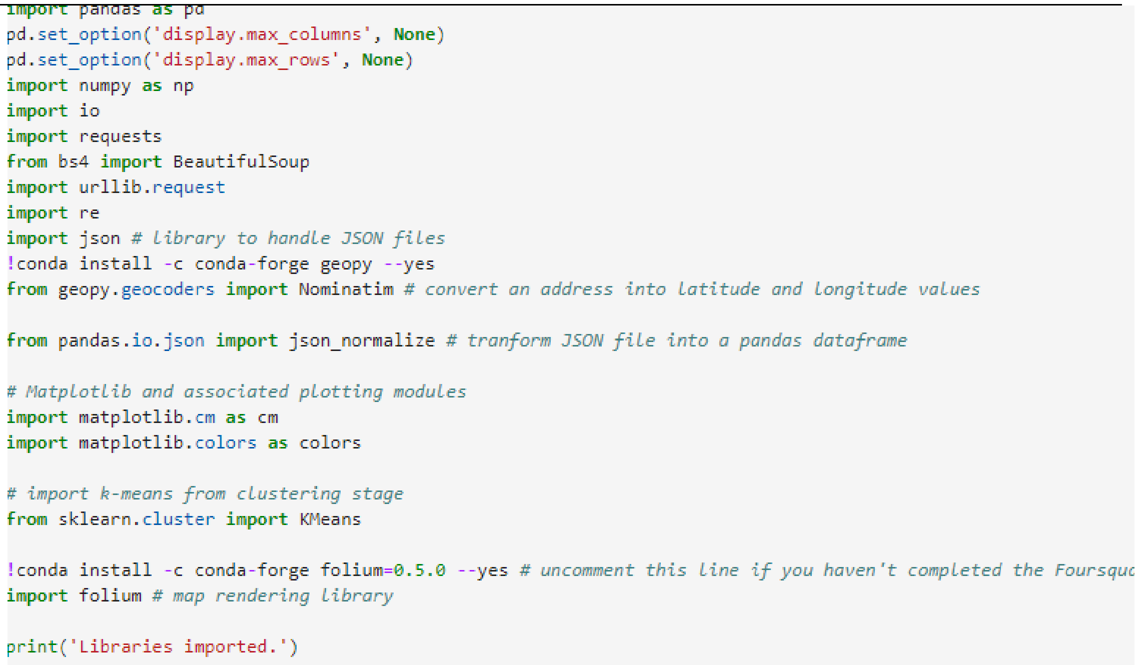
Methodology:

Several Steps were conducted in order to create the desired data science results:

1. Importing Python libraries and packages
2. Scraping Wikipedia
3. Sorting and shaping the data frame
4. Creating a table for Toronto
5. Creating a map of Toronto containing the different neighborhoods
6. Calling Foursquare API in order get venues for a certain neighborhood (as an experiment)
7. Calling Foursquare API and getting top venues for the different neighborhoods
8. Using machine learning to cluster the neighborhoods
9. Creating a map showing the clusters
10. Creating a table showing the most common venues in different neighborhoods in Toronto

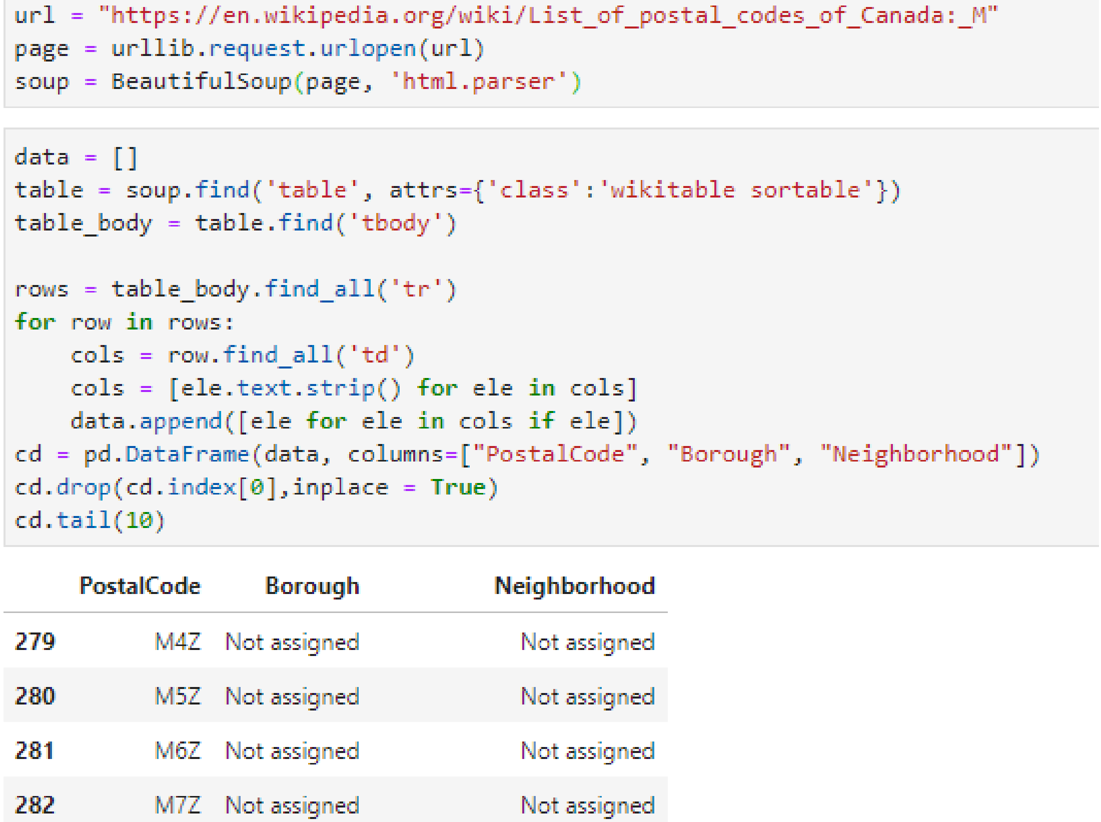
Results:

1



```python
import pandas as pd
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
import numpy as np
import io
import requests
from bs4 import BeautifulSoup
import urllib.request
import re
import json # library to handle JSON files
!conda install -c conda-forge geopy --yes
from geopy.geocoders import Nominatim # convert an address into latitude and longitude values

from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe

# Matplotlib and associated plotting modules
import matplotlib.cm as cm
import matplotlib.colors as colors

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

!conda install -c conda-forge folium=0.5.0 --yes # uncomment this line if you haven't completed the Foursqua
import folium # map rendering library

print('Libraries imported.')
```

2

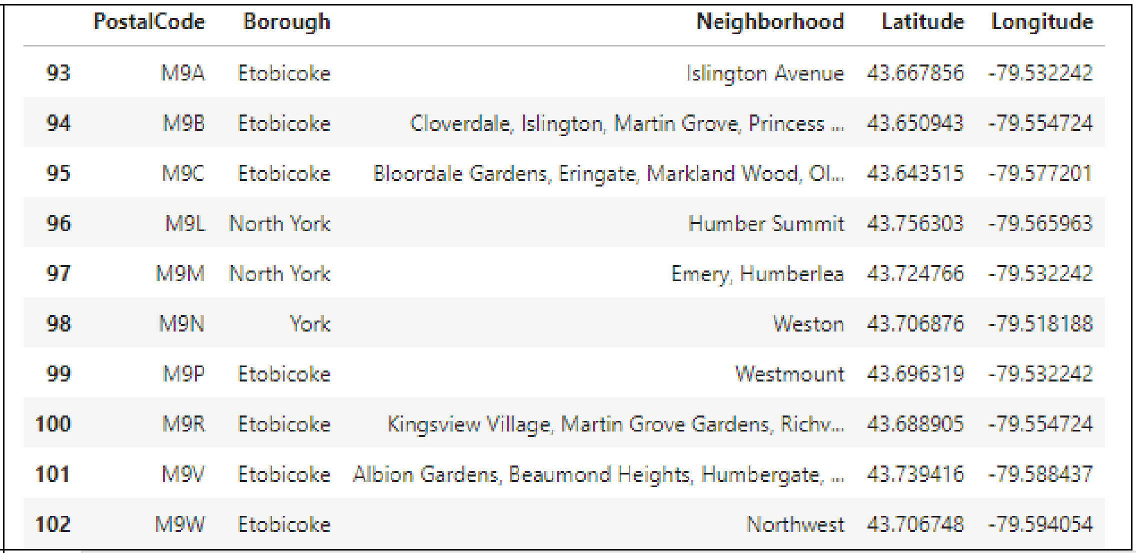


```python
url = "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
page = urllib.request.urlopen(url)
soup = BeautifulSoup(page, 'html.parser')
```

```python
data = []
table = soup.find('table', attrs={'class':'wikitable sortable'})
table_body = table.find('tbody')

rows = table_body.find_all('tr')
for row in rows:
    cols = row.find_all('td')
    cols = [ele.text.strip() for ele in cols]
    data.append([ele for ele in cols if ele])
cd = pd.DataFrame(data, columns=["PostalCode", "Borough", "Neighborhood"])
cd.drop(cd.index[0],inplace = True)
cd.tail(10)
```

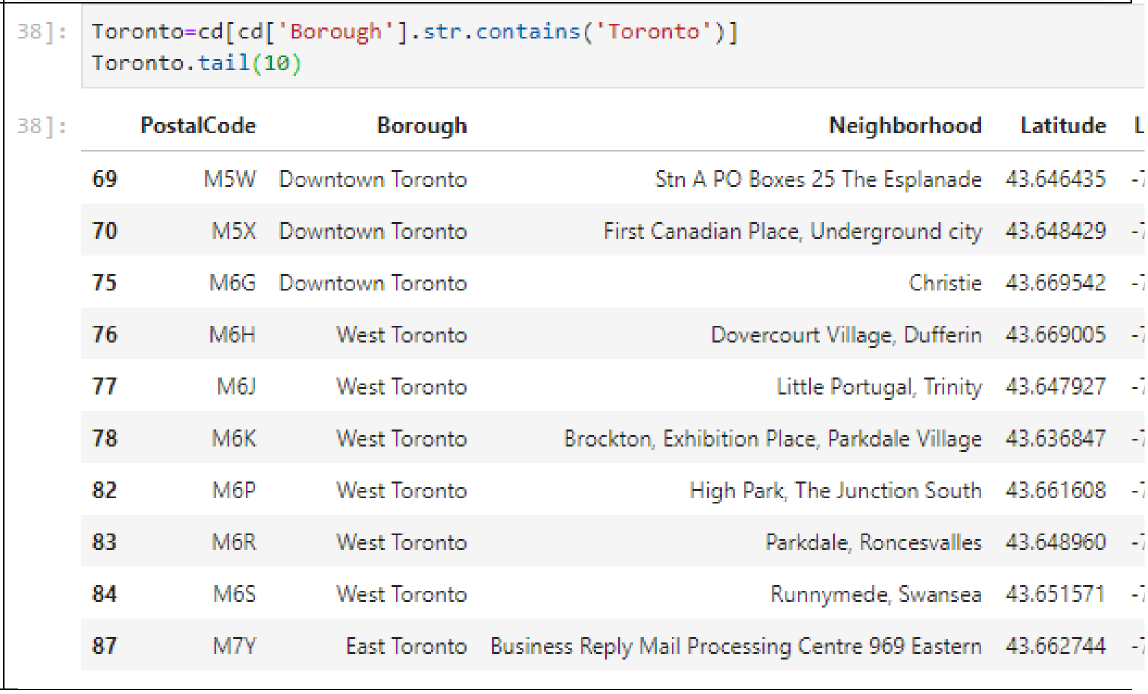
	PostalCode	Borough	Neighborhood
279	M4Z	Not assigned	Not assigned
280	M5Z	Not assigned	Not assigned
281	M6Z	Not assigned	Not assigned
282	M7Z	Not assigned	Not assigned

3



	PostalCode	Borough	Neighborhood	Latitude	Longitude
93	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
94	M9B	Etobicoke	Cloverdale, Islington, Martin Grove, Princess ...	43.650943	-79.554724
95	M9C	Etobicoke	Bloordale Gardens, Eringate, Markland Wood, Ol...	43.643515	-79.577201
96	M9L	North York	Humber Summit	43.756303	-79.565963
97	M9M	North York	Emery, Humberlea	43.724766	-79.532242
98	M9N	York	Weston	43.706876	-79.518188
99	M9P	Etobicoke	Westmount	43.696319	-79.532242
100	M9R	Etobicoke	Kingsview Village, Martin Grove Gardens, Richv...	43.688905	-79.554724
101	M9V	Etobicoke	Albion Gardens, Beaumond Heights, Humbergate, ...	43.739416	-79.588437
102	M9W	Etobicoke	Northwest	43.706748	-79.594054

4

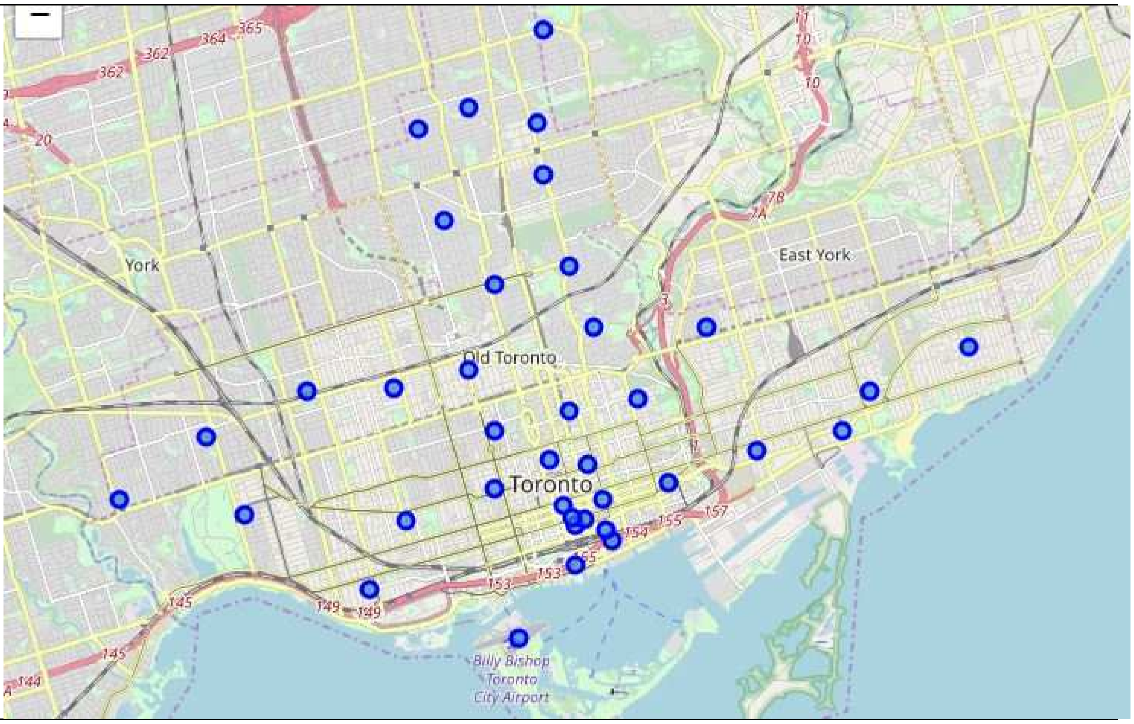


```
38]: Toronto=cd[cd['Borough'].str.contains('Toronto')]
     Toronto.tail(10)
```

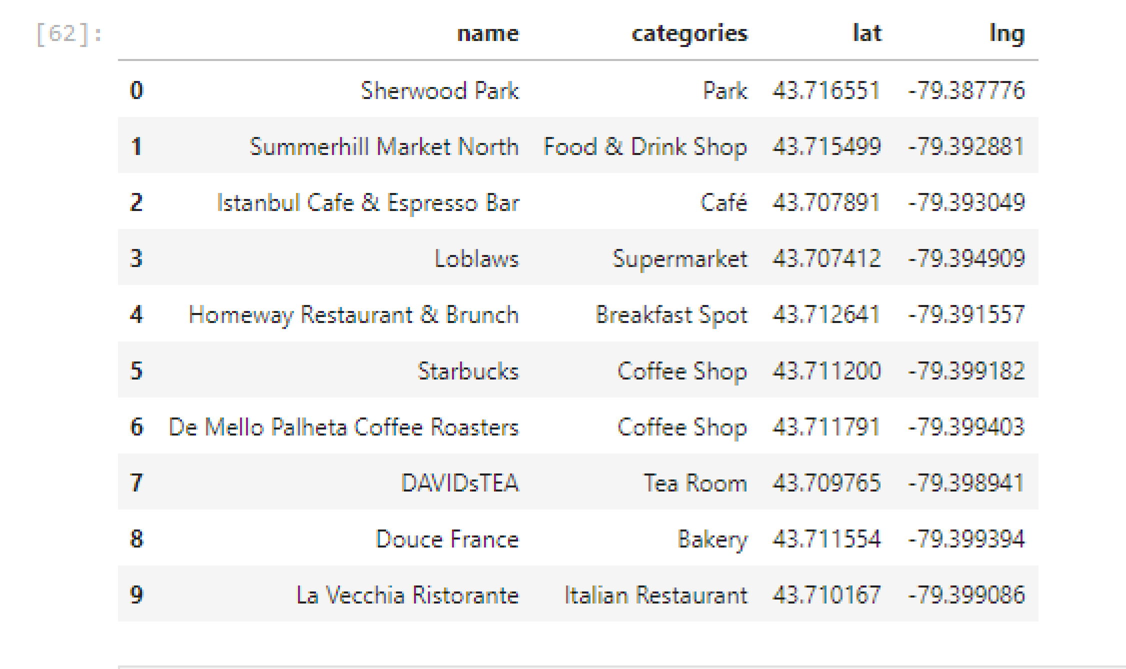
38]:

	PostalCode	Borough	Neighborhood	Latitude	L
69	M5W	Downtown Toronto	Stn A PO Boxes 25 The Esplanade	43.646435	-7
70	M5X	Downtown Toronto	First Canadian Place, Underground city	43.648429	-7
75	M6G	Downtown Toronto	Christie	43.669542	-7
76	M6H	West Toronto	Dovercourt Village, Dufferin	43.669005	-7
77	M6J	West Toronto	Little Portugal, Trinity	43.647927	-7
78	M6K	West Toronto	Brockton, Exhibition Place, Parkdale Village	43.636847	-7
82	M6P	West Toronto	High Park, The Junction South	43.661608	-7
83	M6R	West Toronto	Parkdale, Roncesvalles	43.648960	-7
84	M6S	West Toronto	Runnymede, Swansea	43.651571	-7
87	M7Y	East Toronto	Business Reply Mail Processing Centre 969 Eastern	43.662744	-7

5



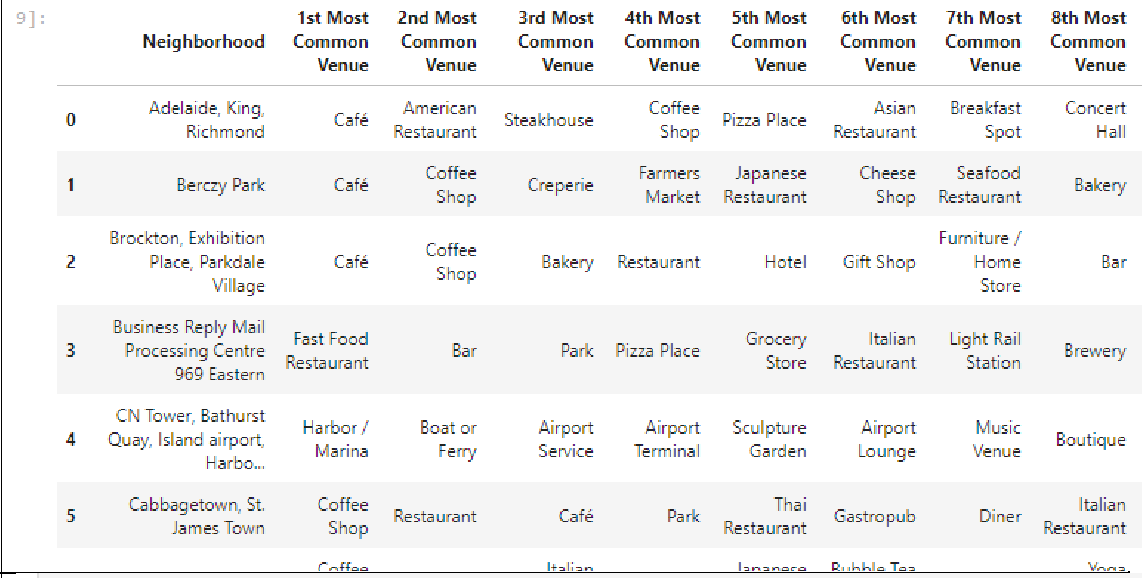
6



[62]:

	name	categories	lat	lng
0	Sherwood Park	Park	43.716551	-79.387776
1	Summerhill Market North	Food & Drink Shop	43.715499	-79.392881
2	Istanbul Cafe & Espresso Bar	Café	43.707891	-79.393049
3	Loblaws	Supermarket	43.707412	-79.394909
4	Homeway Restaurant & Brunch	Breakfast Spot	43.712641	-79.391557
5	Starbucks	Coffee Shop	43.711200	-79.399182
6	De Mello Palheta Coffee Roasters	Coffee Shop	43.711791	-79.399403
7	DAVIDsTEA	Tea Room	43.709765	-79.398941
8	Douce France	Bakery	43.711554	-79.399394
9	La Vecchia Ristorante	Italian Restaurant	43.710167	-79.399086

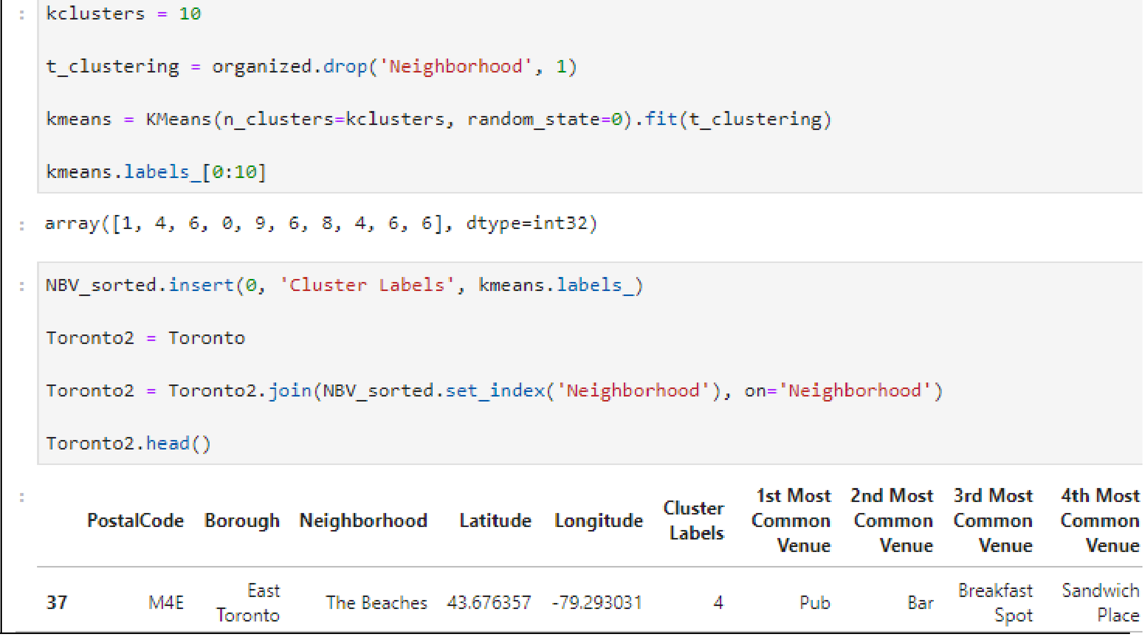
7



9]:

	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
0	Adelaide, King, Richmond	Café	American Restaurant	Steakhouse	Coffee Shop	Pizza Place	Asian Restaurant	Breakfast Spot	Concert Hall
1	Berczy Park	Café	Coffee Shop	Creperie	Farmers Market	Japanese Restaurant	Cheese Shop	Seafood Restaurant	Bakery
2	Brockton, Exhibition Place, Parkdale Village	Café	Coffee Shop	Bakery	Restaurant	Hotel	Gift Shop	Furniture / Home Store	Bar
3	Business Reply Mail Processing Centre 969 Eastern	Fast Food Restaurant	Bar	Park	Pizza Place	Grocery Store	Italian Restaurant	Light Rail Station	Brewery
4	CN Tower, Bathurst Quay, Island airport, Harbo...	Harbor / Marina	Boat or Ferry	Airport Service	Airport Terminal	Sculpture Garden	Airport Lounge	Music Venue	Boutique
5	Cabbagetown, St. James Town	Coffee Shop	Restaurant	Café	Park	Thai Restaurant	Gastropub	Diner	Italian Restaurant
		Coffee		Italian		Japanese	Bubble Tea		Yoga

8



```
: kclusters = 10

  t_clustering = organized.drop('Neighborhood', 1)

  kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(t_clustering)

  kmeans.labels_[0:10]
```

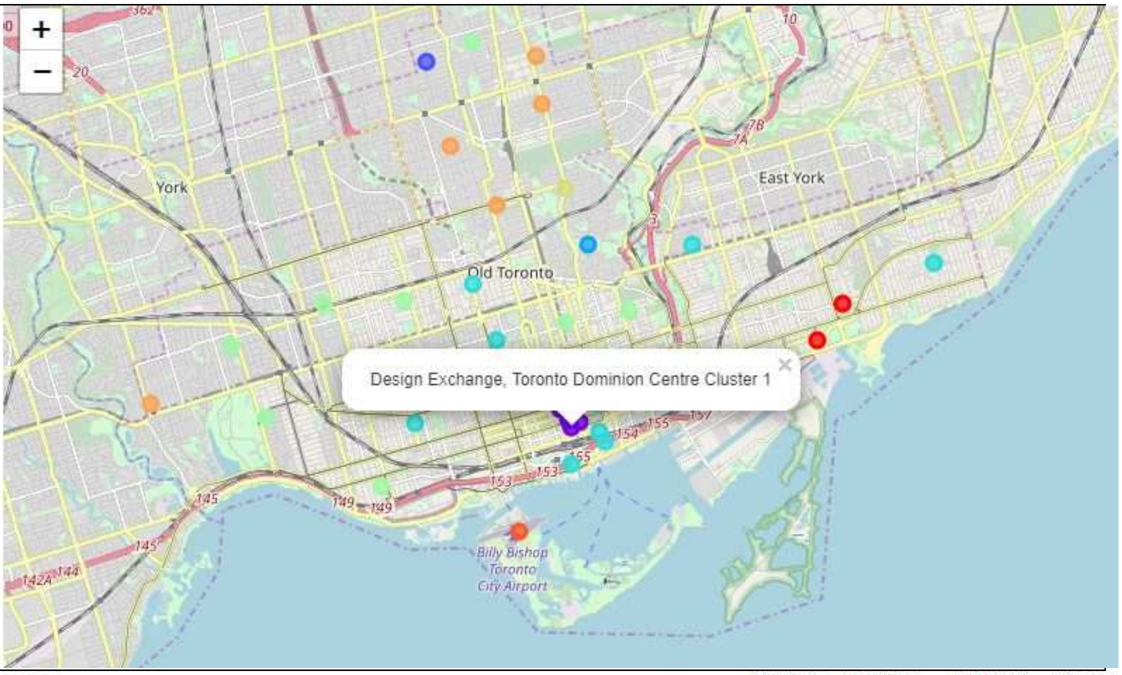
```
: array([1, 4, 6, 0, 9, 6, 8, 4, 6, 6], dtype=int32)
```

```
: NBV_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

  Toronto2 = Toronto

  Toronto2 = Toronto2.join(NBV_sorted.set_index('Neighborhood'), on='Neighborhood')

  Toronto2.head()
```

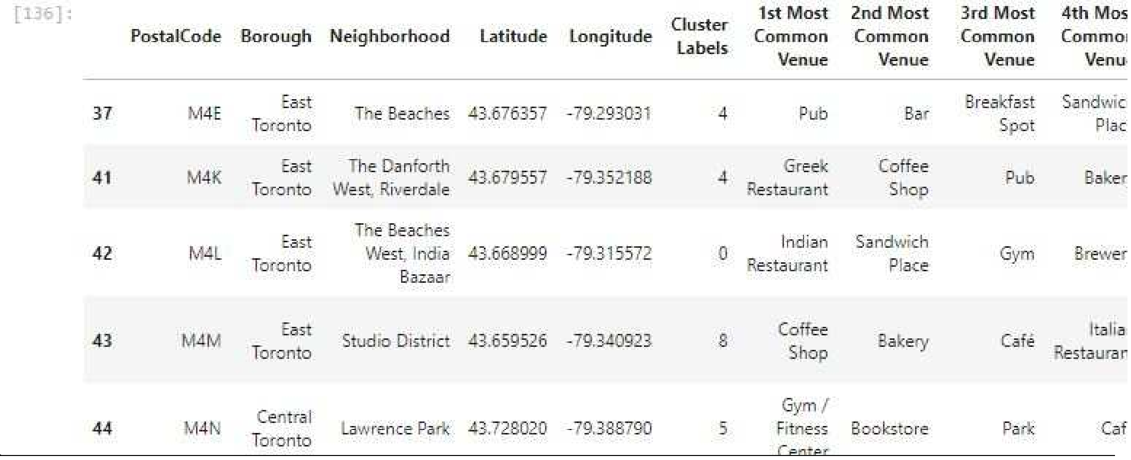
:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
37	M4E	East Toronto	The Beaches	43.676357	-79.293031	4	Pub	Bar	Breakfast Spot	Sandwich Place

9


+
−
York
East York
Old Toronto
Design Exchange, Toronto Dominion Centre Cluster 1
Billy Bishop Toronto City Airport

10



[136]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Mos Commo Venu
37	M4E	East Toronto	The Beaches	43.676357	-79.293031	4	Pub	Bar	Breakfast Spot	Sandwic Plac
41	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188	4	Greek Restaurant	Coffee Shop	Pub	Baker
42	M4L	East Toronto	The Beaches West, India Bazaar	43.668999	-79.315572	0	Indian Restaurant	Sandwich Place	Gym	Brewer
43	M4M	East Toronto	Studio District	43.659526	-79.340923	8	Coffee Shop	Bakery	Café	Italia Restauran
44	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790	5	Gym / Fitness Center	Bookstore	Park	Caf

Analysis:

As we can see from the dataset, the most popular venues include either food or beverages mainly and that should be a crucial idea who is considering what would work in a city like Toronto. After careful analysis, it would be wise to create an investment that is part of at least the first 5 most common venues as any other type of venue might not be able to create a return on the investment.

Conclusion:

As a conclusion, it was noticed that the nine courses we learned in the coursera IBM program all contributed to create a story. This is essentially what data science is about really.

References:

- Wikipedia
- Foursquare API
- Coursera/IBM lectures & labs