Capstone Project
The Battle of Neighborhoods (Week 2)

Table of Contents

Introduction

In this project our objective is to take some decisions based on information available with us and make certain predictions by using data science methodology. I have identified 2 real world problems and I shall try to predict the answer. The problems are:

- We have segmented and clustered the neighborhoods of New York City and the city of Toronto. Both cities are very diverse and are the financial capitals of their respective countries. One interesting idea would be to compare the neighborhoods of the two cities and determine how similar or dissimilar they are. Is New York City more like Toronto?
- In the city of Toronto, if someone is looking to open a restaurant, where would you recommend that they open it?

Data Source

To solve this problem, I need data from 2 cities: New York and Toronto.

New York Data

New York data is downloaded from the link https://cocl.us/new_york_dataset Then this data

In [12]:
```
!wget -q -O 'newyork_data.json' https://cocl.us/new_york_dataset
print('Data downloaded!')
```

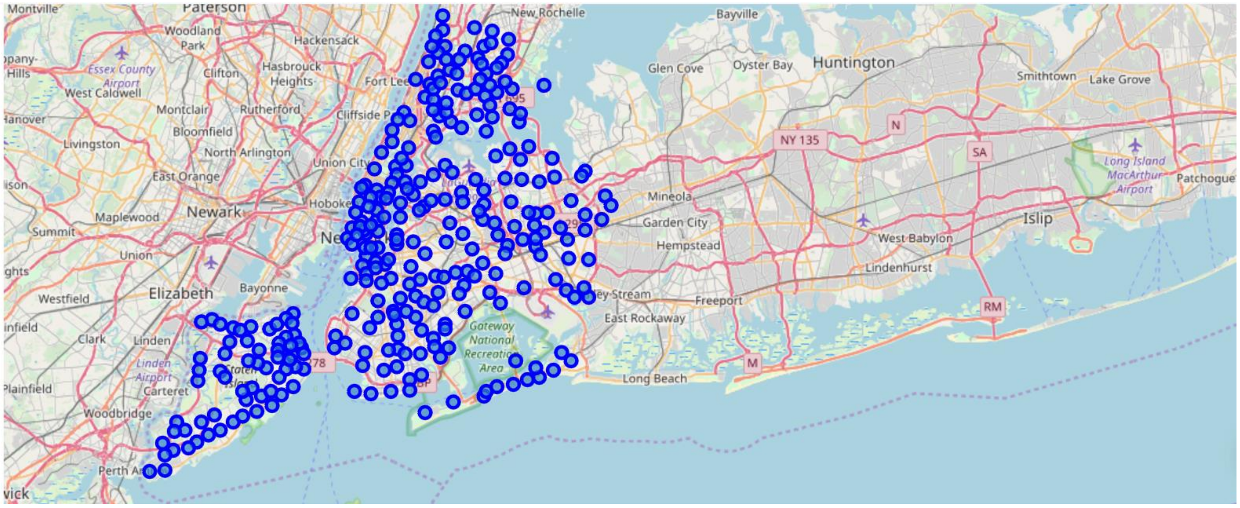
```
Data downloaded!
```

In [28]:
```
with open('newyork_data.json') as json_data:
    newyork_data = json.load(json_data)
newyork_data
```

Out[28]:

```
{'type': 'FeatureCollection',
 'totalFeatures': 306,
 'features': [{'type': 'Feature',
   'id': 'nyu_2451_34572.1',
   'geometry': {'type': 'Point',
    'coordinates': [-73.84720052054902, 40.89470517661]},
   'geometry_name': 'geom',
   'properties': {'name': 'Wakefield',
    'stacked': 1,
    'annoline1': 'Wakefield',
    'annoline2': None,
    'annoline3': None,
    'annoangle': 0.0,
    'borough': 'Bronx',
    'bbox': [-73.84720052054902,
     40.89470517661,
     -73.84720052054902,
     40.89470517661]}},
  {'type': 'Feature',
```

Data is plotted in a map:



Toronto Data

Toronto data is downloaded from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

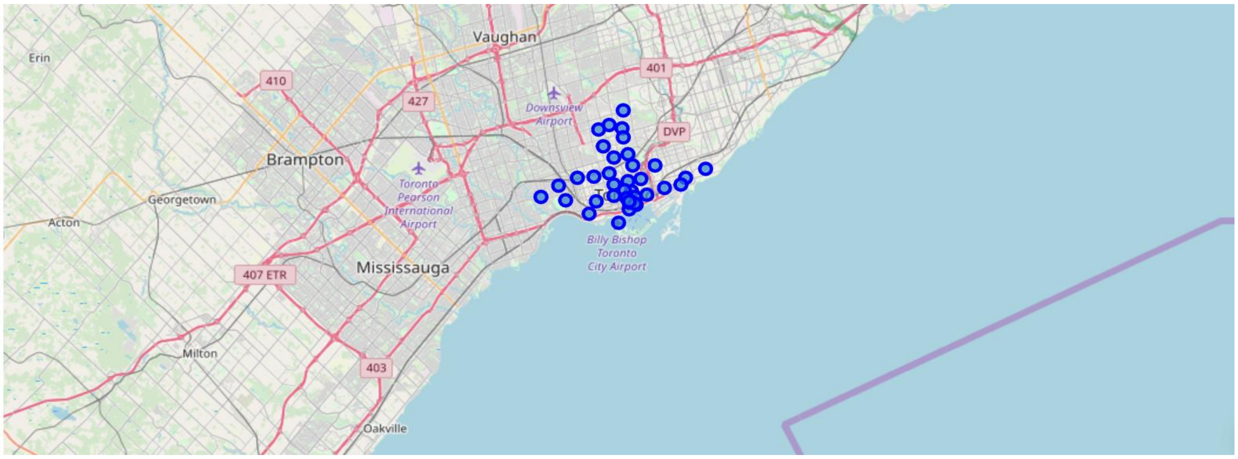
```
postal_code_df = pd.read_html('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M', header=0)[0]
df=postal_code_df[postal_code_df.Borough != 'Not assigned']
df=df.groupby(['Postcode','Borough'])['Neighbourhood'].apply(', '.join).reset_index()
df.loc[df.Neighbourhood == 'Not assigned', 'Neighbourhood'] = df.Borough
!wget -q -O 'geographical_data.csv' http://cocl.us/Geospatial_data
print('Data downloaded!')
geo_df = pd.read_csv('geographical_data.csv')
geo_df.rename(columns={'Postal Code':'Postcode'},inplace=True)
df1=pd.merge(df,geo_df[['Postcode','Latitude','Longitude']],on=['Postcode'])
df1.head()
```

```
Data downloaded!
```

[22]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Data is plotted in a map:



Methodology used

Both New York and Toronto data is downloaded and cleaned. This data now will be combined with FourSquare Location data and we can validate the results to identify the similarities and dis-similarities among these two cities.

We can use Toronto data and use K-Nearest-Neighbor algorithm to predict the location to open a restaurant.

New York Data grouped:

In [37]:
```
manhattan_grouped = manhattan_onehot.groupby('Neighborhood').mean().reset_index()
manhattan_grouped
```

Out[37]:

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auditorium	Aust Resta
0	Battery Park City	0.000000	0.00	0.00	0.000000	0.010000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.010000	0.01	
1	Carnegie Hill	0.000000	0.00	0.00	0.000000	0.010000	0.00	0.00	0.000000	0.010000	0.000000	0.01	0.000000	0.000000	0.000000	0.00	
2	Central Harlem	0.000000	0.00	0.00	0.069767	0.046512	0.00	0.00	0.000000	0.000000	0.023256	0.00	0.000000	0.000000	0.000000	0.00	
3	Chelsea	0.000000	0.00	0.00	0.000000	0.030000	0.01	0.00	0.000000	0.000000	0.030000	0.00	0.000000	0.010000	0.000000	0.00	
4	Chinatown	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.020000	0.000000	0.00	
5	Civic Center	0.000000	0.00	0.00	0.000000	0.030000	0.01	0.00	0.000000	0.000000	0.020000	0.00	0.000000	0.010000	0.000000	0.00	
6	Clinton	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	0.010000	0.00	0.000000	0.000000	0.000000	0.00	
7	East Harlem	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	
8	East Village	0.000000	0.00	0.00	0.000000	0.020000	0.01	0.00	0.020000	0.010000	0.010000	0.00	0.010000	0.000000	0.000000	0.00	
9	Financial District	0.010000	0.00	0.00	0.000000	0.030000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	
10	Flatiron	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	
11	Gramercy	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.01	0.000000	0.000000	0.010000	0.00	0.000000	0.000000	0.000000	0.00	
[illegible]	Greenwich	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	

Toronto data grouped:

```
In [43]: toronto_grouped = toronto_onehot.groupby('Neighborhood').mean().reset_index()
         toronto_grouped
```

Out[43]:

	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports
0	Adelaide, King, Richmond	0.000000	0.00	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.040000	0.000000	0.00	0.010000	0.010000	0.000000	0.030000	0.0000
1	Berczy Park	0.000000	0.00	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.000000	0.000000	0.00	0.017544	0.000000	0.000000	0.000000	0.0000
2	Brockton, Exhibition Place, Parkdale Village	0.074074	0.00	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.0000
3	Business Reply Mail Processing Centre 969 Eastern	0.052632	0.00	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.0000
4	CN Tower, Bathurst Quay, Island airport, Harbo...	0.000000	0.00	0.000000	0.0625	0.0625	0.0625	0.125	0.1875	0.125	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.0000
5	Cabbagetown, St. James Town	0.000000	0.00	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.0000
6	Central Bay Street	0.011236	0.00	0.000000	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.011236	0.000000	0.00	0.000000	0.011236	0.000000	0.000000	0.0000

Above data analysis clearly shows that both the cities are almost equally populated with parks, shops, restaurants, Hotels. We can come to a conclusion that both New York city and Toronto are very similar to each other in terms of amenities like parks, shops, restaurants, Hotels etc.

Second Problem (K-Nearest Neighbor)

In the city of Toronto, if someone is looking to open a restaurant, where would you recommend that they open it?

```
In [66]: k = 6
         #Train Model and Predict
         neigh = KNeighborsClassifier(n_neighbors = k).fit(X_train,y_train)
         neigh

Out[66]: KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                    metric_params=None, n_jobs=None, n_neighbors=6, p=2,
                    weights='uniform')
```

```
In [67]: yhat = neigh.predict(X_test)
         yhat[0:5]

Out[67]: array(['M4Y', 'M5E', 'M4Y', 'M5B', 'M5B'], dtype=object)
```

```
In [68]: from sklearn import metrics
         print("Train set Accuracy: ", metrics.accuracy_score(y_train, neigh.predict(X_train)))
         print("Test set Accuracy: ", metrics.accuracy_score(y_test, yhat))

    Train set Accuracy:  0.1
    Test set Accuracy:  0.0
```

Using K-Nearest Neighbor process we can come to the prediction that, the restaurant can be opened in one of the following zip codes: 'M4Y', 'M5E', 'M4Y', 'M5B', 'M5B'

Results

Above data analysis clearly shows that both the cities are almost equally populated with parks, shops, restaurants, Hotels. We can come to a conclusion that both New York city and Toronto are very similar to each other in terms of amenities like parks, shops, restaurants, Hotels etc.

Using K-Nearest Neighbor process we can come to the prediction that, the restaurant can be opened in one of the following zip codes: 'M4Y', 'M5E', 'M4Y', 'M5B', 'M5B'

Discussion

This analysis is done based on the available data collected from website.

Conclusion

After this analysis this can be concluded that both New York city and Toronto are very similar to each other in terms of amenities like parks, shops, restaurants, Hotels etc. Above data analysis clearly shows that both the cities are almost equally populated with parks, shops, restaurants, Hotels. For the second problem, Using K-Nearest Neighbor process we can come to the prediction that, the restaurant can be opened in one of the following zip codes: 'M4Y', 'M5E', 'M4Y', 'M5B', 'M5B'