IBM Coursera Data Science Specialization Capstone Project - Clustering and exploring neighbourhoods of New York and Toronto

By Aisha Kala

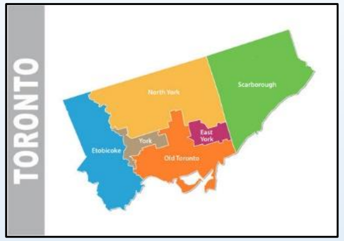
Introduction

- New York City as well as Toronto City are recognised cosmopolitan cities of the world, and as such, they both attract many people who come to visit or to live and work in these cities.
- For this project I have decided to cluster and explore the neighbourhoods of both New York, USA, and Toronto, Canada.
- The two cities were compared in terms of venues such as restaurants, art galleries, parks, coffee shops, café’s, and so forth as this information will be useful to people who are either looking to visit any of these cities or people that may intend to relocate to one of these cities.

Background – New York

- One of the most densely populated cities in the United States, with an estimated population of 8 336 817 distributed over 784 square kilometres.
- Has been described as the cultural, financial and media capital of the world and many popular television shows have made New York City a “must-see” destination for people from all over the world.
- Considered the financial hub of the United States of America.
- Comprises of 5 boroughs: Brooklyn, Queens, Manhattan, The Bronx and Staten Island and is situated on one of the world’s largest natural harbours.
- Many districts and landmarks are well known tourist destinations, attracting 65 million tourists annually which is why I have chosen to use New York City in my project.

Background - Toronto



- The capital city of Ontario in Canada, situated in Northern America.
- The most populous city in Canada, with an estimated population of 6 139 000 distributed over 630.2 square kilometres.
- Has been recognised as one of the most multicultural and cosmopolitan cities of the world.
- The diverse population is a reflection of its current and historical role as an important destination for immigrants to Canada.
- Like New York, Toronto is considered the financial and business capital of Canada.
- Comprises of 10 boroughs: Etobicoke, North York, Scarborough, York, East York, Central Toronto, East Toronto, West Toronto, Downtown Toronto and old Toronto.
- Attracts over 43 million tourists annually and is home to many museums, art galleries, historic sites, festivals and sporting activities.

Background

- It can be seen that there are many similarities between these cities.
- Both of these cities are large, multicultural, diverse and attractive and have many points of interest for residents as well as tourists, and both cities offer potential for entrepreneurs wanting to open a business in either city.
- The aim of this project is to compare the neighbourhoods of the 2 cities and find neighbourhoods that are similar based on the venues in these neighbourhood's that make them similar.

Target Audience

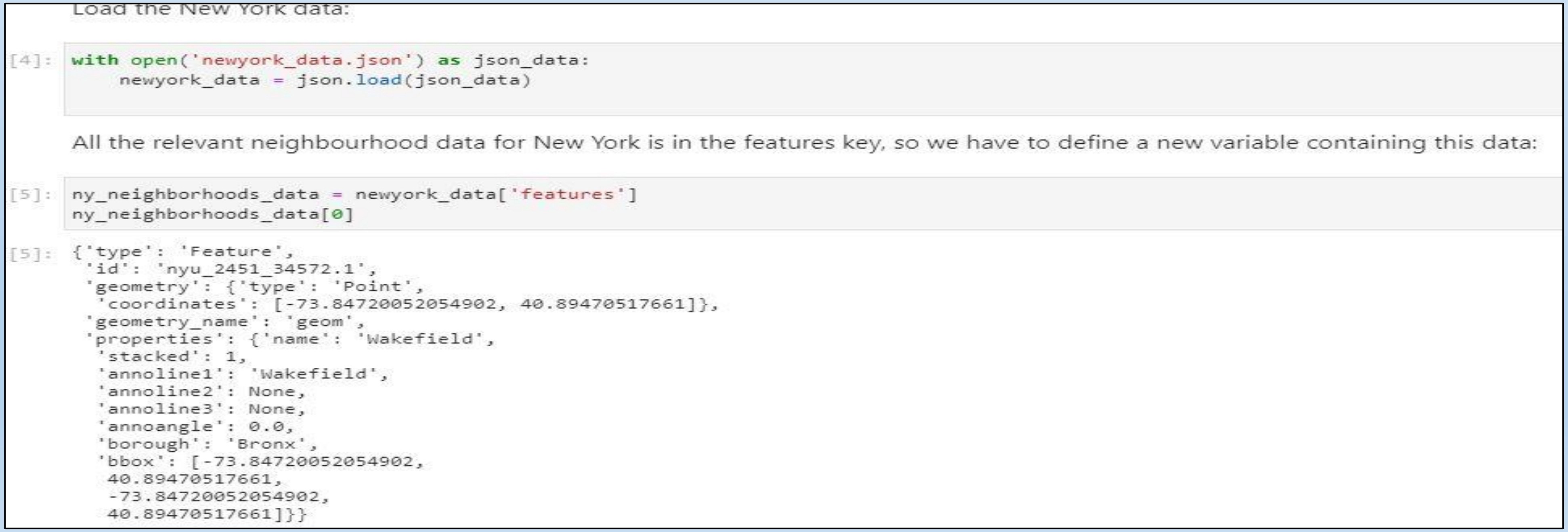
- Potential job seekers wanting to relocate to a large, multicultural city
- Tourists intending to visit either one or both of these cities and would like to know which areas would be of interest to them
- Entrepreneurs intending to start a business in either one or both of these cities

Data Acquisition and processing

Neighbourhood data

New York City

- Data for the boroughs and neighbourhoods of New York was obtained from the Coursera skills lab as New York was one of the cities explored during the lab (**NYU Spatial Data Repository**).
- The dataset is in a json file that contains the name of each borough, the neighbourhoods, and the geographical coordinates and so on. In order to utilise the data, it had to be stored in a pandas dataframe.



Load the New York data:

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

All the relevant neighbourhood data for New York is in the features key, so we have to define a new variable containing this data:

```
[5]: ny_neighborhoods_data = newyork_data['features']
     ny_neighborhoods_data[0]
```

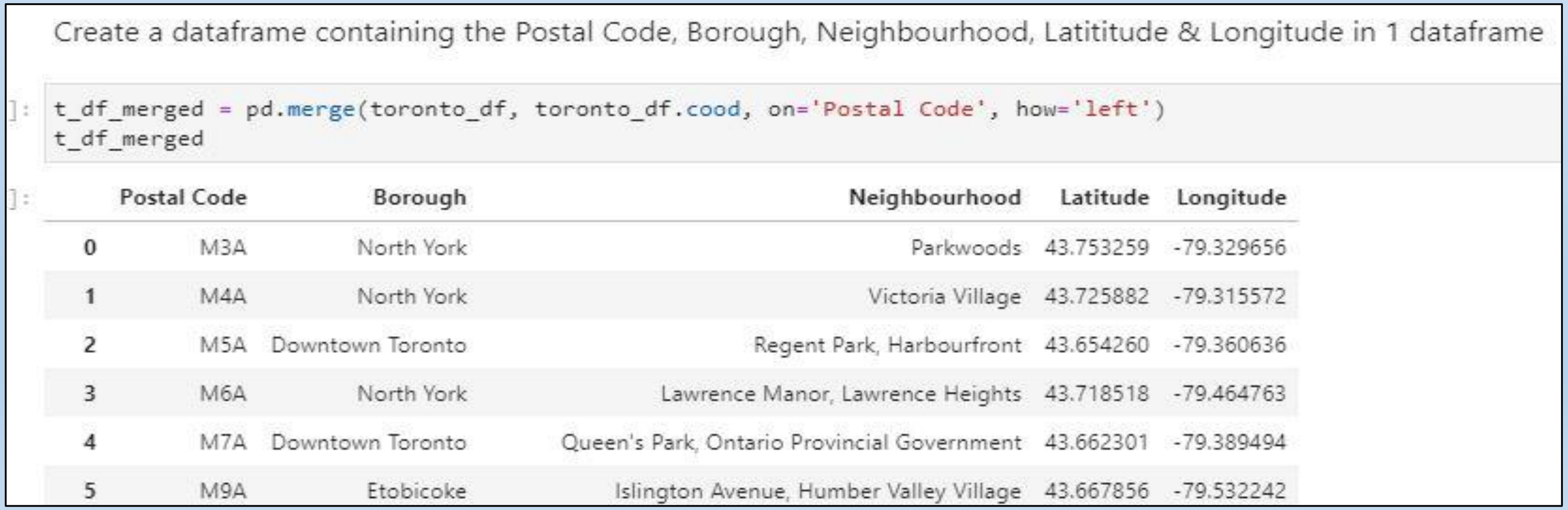
```
[5]: {'type': 'Feature',
      'id': 'nyu_2451_34572.1',
      'geometry': {'type': 'Point',
       'coordinates': [-73.84720052054902, 40.89470517661]},
      'geometry_name': 'geom',
      'properties': {'name': 'Wakefield',
       'stacked': 1,
       'annoline1': 'Wakefield',
       'annoline2':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

Data Acquisition and processing

Neighbourhood data

Toronto

- Unlike New York, there was no json file for Toronto, therefore the data for the boroughs of Toronto was obtained by scraping a Wikipedia page (**List of Postal Codes of Canada_M**.)
- This was merged with a csv file provided in the Coursera skills lab containing geographical coordinates and postal codes of the neighbourhoods.
- The resulting dataframe contained the postal codes, boroughs, neighbourhoods and geographical coordinates for the city.



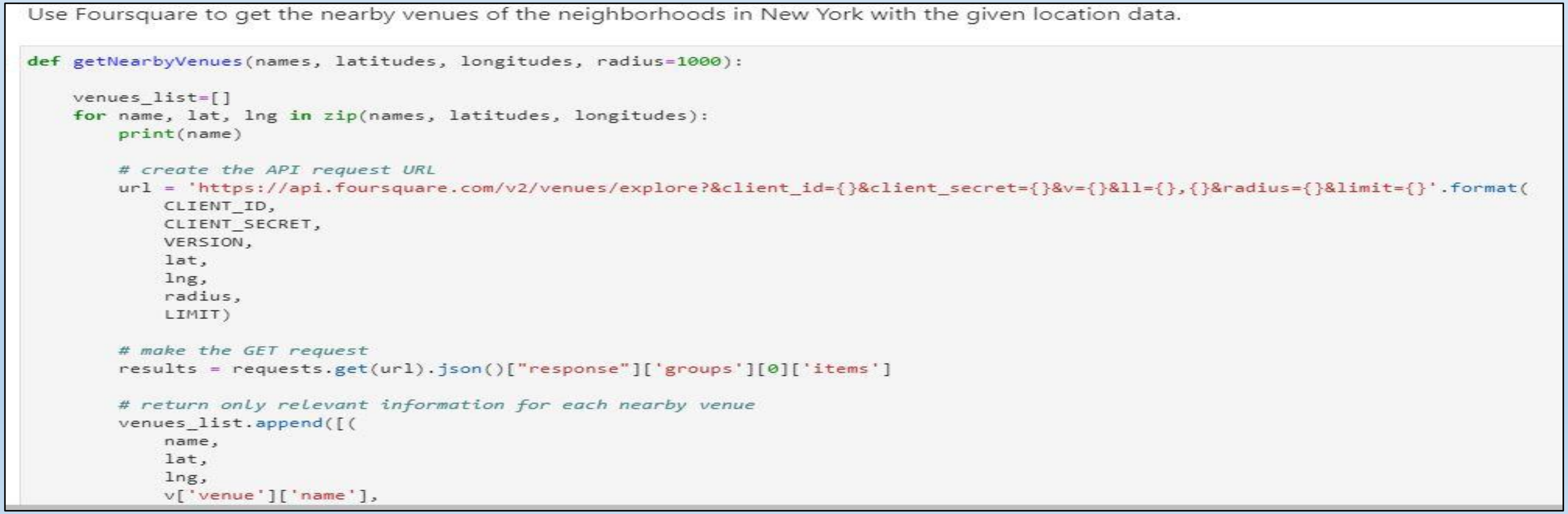
Create a dataframe containing the Postal Code, Borough, Neighbourhood, Latititude & Longitude in 1 dataframe

```
]: t_df_merged = pd.merge(toronto_df, toronto_df.cood, on='Postal Code', how='left')
   t_df_merged
```

	Postal Code	Borough	Neighbourhood	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

Venues data: Foursquare location data

- A social location service that allows users to explore places around them.
- The API provides location based experiences with diverse information about venues, users, photos, and check-ins.
- Was utilised to extract venue based information for all of the neighbourhoods in this project, by utilising the geographical coordinates from the dataframes.
- For this project, the top 5 most common venue data was extracted for both New York and Toronto.



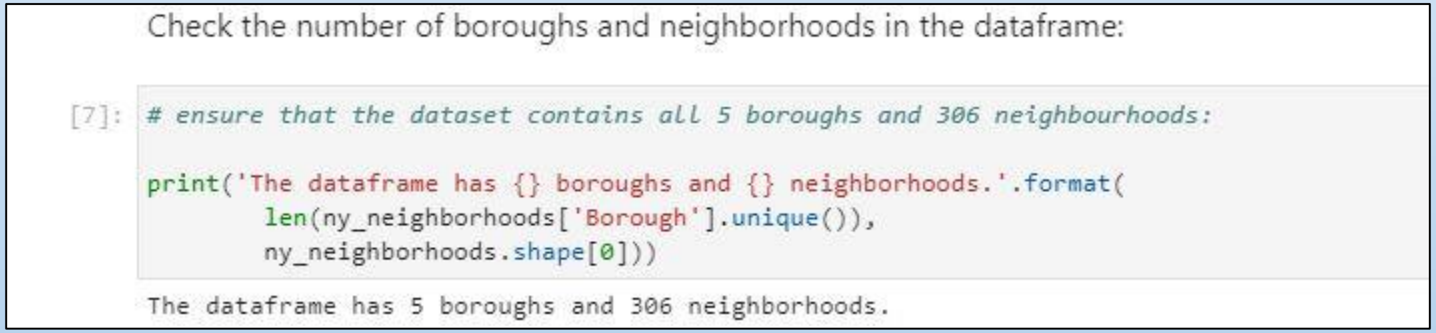
Use Foursquare to get the nearby venues of the neighborhoods in New York with the given location data.

```
def getNearbyVenues(names, latitudes, longitudes, radius=1000):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
```

Methodology

Exploratory Data Analysis

After obtaining the data for both cities, the shape of the data had to be checked to ensure that all the boroughs and neighbourhoods were present.



Check the number of boroughs and neighborhoods in the dataframe:

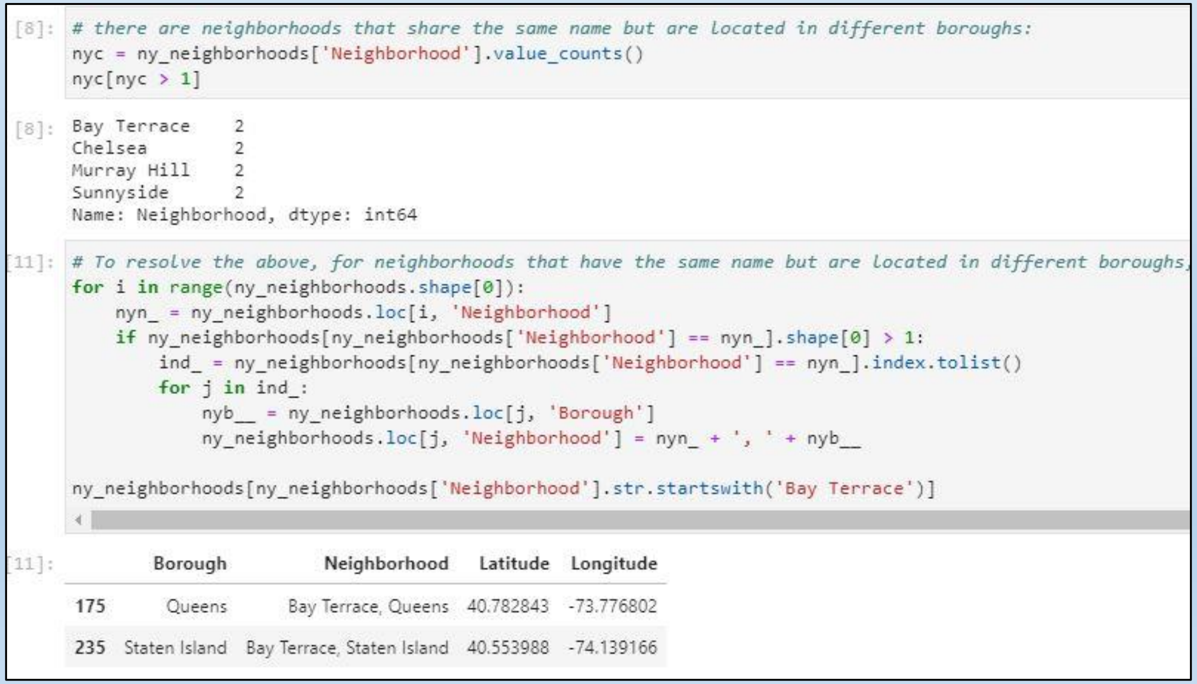
```
[7]: # ensure that the dataset contains all 5 boroughs and 306 neighbourhoods:

     print('The dataframe has {} boroughs and {} neighborhoods.'.format(
             len(ny_neighborhoods['Borough'].unique()),
             ny_neighborhoods.shape[0]))
```

```
The dataframe has 5 boroughs and 306 neighborhoods.
```

Methodology

Data pre-processing

- For the New York data, the data from the json file was placed into a pandas dataframe.
- While doing the exploratory data analysis, it was discovered that for New York, certain neighbourhoods had the same name but were in different boroughs.
- To correct this, the borough name had to be added to these neighbourhood names in order to differentiate them.
- For Toronto, certain boroughs were not assigned therefore these boroughs were removed. For the neighbourhoods in Toronto that we not assigned, the neighbourhood name was changed from “not assigned” to the respective borough name.



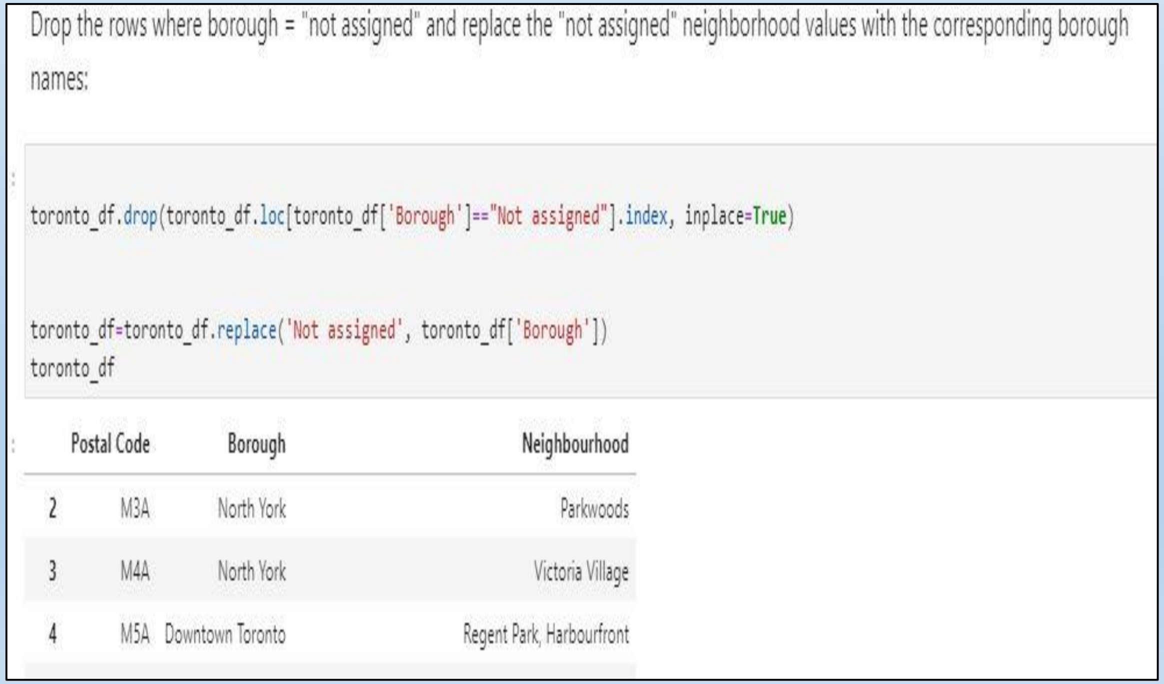
```
[8]: # there are neighborhoods that share the same name but are located in different boroughs:
     nyc = ny_neighborhoods['Neighborhood'].value_counts()
     nyc[nyc > 1]
```

```
[8]: Bay Terrace    2
     Chelsea        2
     Murray Hill    2
     Sunnyside      2
     Name: Neighborhood, dtype: int64
```

```
[11]: # To resolve the above, for neighborhoods that have the same name but are located in different boroughs,
      for i in range(ny_neighborhoods.shape[0]):
          nyn_ = ny_neighborhoods.loc[i, 'Neighborhood']
          if ny_neighborhoods[ny_neighborhoods['Neighborhood'] == nyn_].shape[0] > 1:
              ind_ = ny_neighborhoods[ny_neighborhoods['Neighborhood'] == nyn_].index.tolist()
              for j in ind_:
                  nyb__ = ny_neighborhoods.loc[j, 'Borough']
                  ny_neighborhoods.loc[j, 'Neighborhood'] = nyn_ + ', ' + nyb__

      ny_neighborhoods[ny_neighborhoods['Neighborhood'].str.startswith('Bay Terrace')]
```

[11]:

	Borough	Neighborhood	Latitude	Longitude
175	Queens	Bay Terrace, Queens	40.782843	-73.776802
235	Staten Island	Bay Terrace, Staten Island	40.553988	-74.139166



Drop the rows where borough = "not assigned" and replace the "not assigned" neighborhood values with the corresponding borough names:

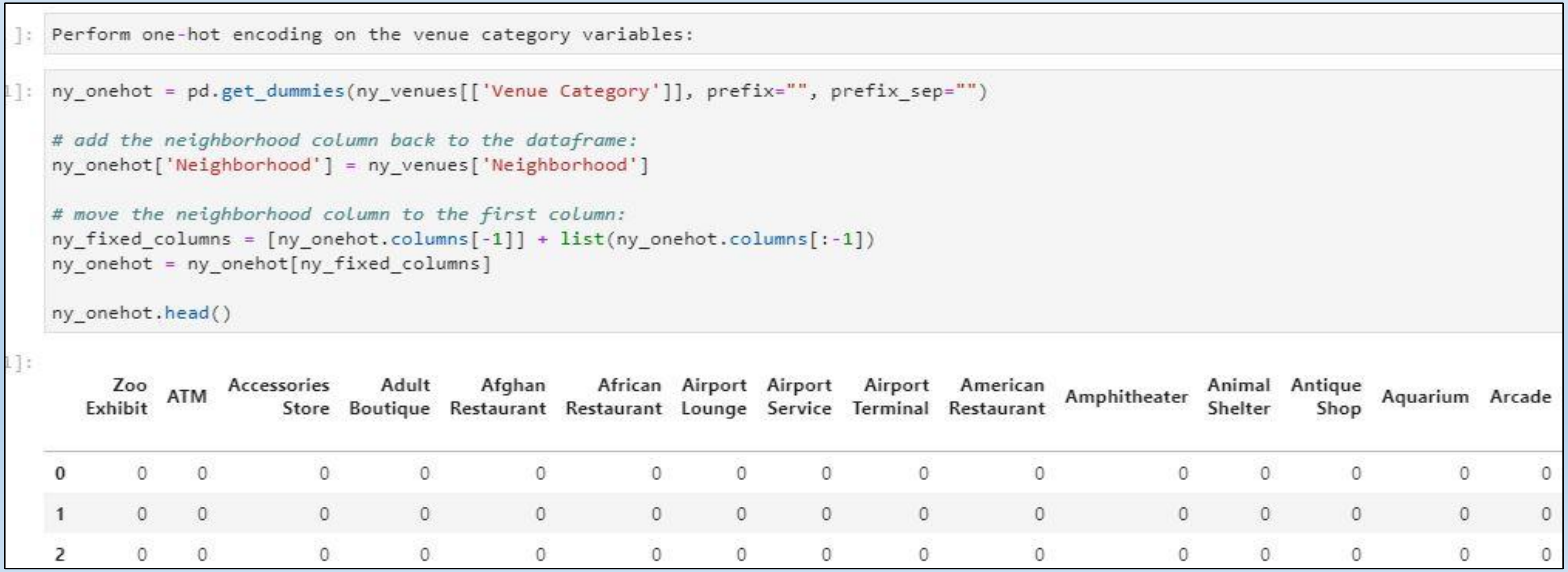
```
toronto_df.drop(toronto_df.loc[toronto_df['Borough']=="Not assigned"].index, inplace=True)


toronto_df=toronto_df.replace('Not assigned', toronto_df['Borough'])
toronto_df
```

	Postal Code	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

Methodology

One-hot encoding

- Machine learning algorithms cannot work with categorical data directly therefore categorical data must be converted to numerical data.
- This is required for both input and output variables that are categorical.
- This is done using one-hot encoding.
- One hot encoding is a representation of categorical variables as binary vectors.



Perform one-hot encoding on the venue category variables:

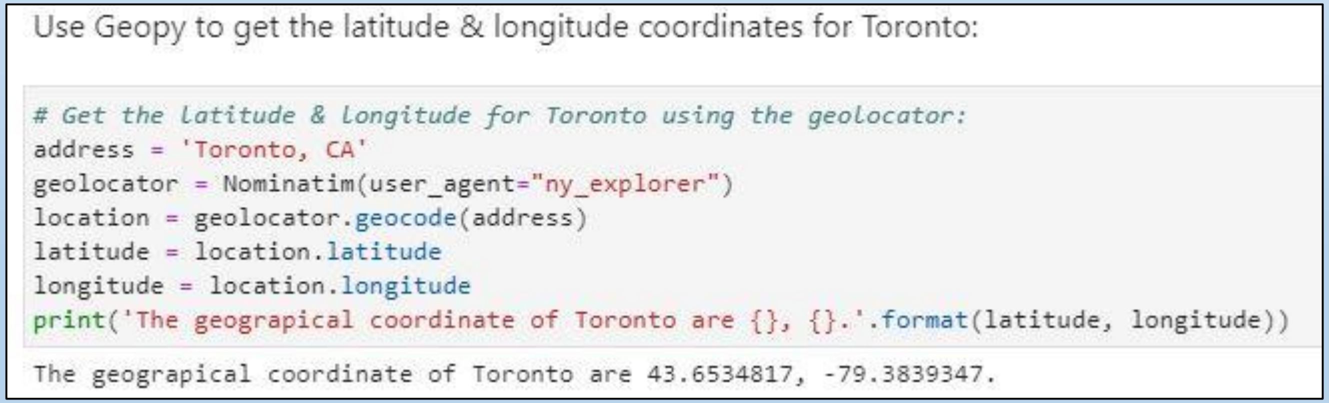
```
ny_onehot = pd.get_dummies(ny_venues[['Venue Category']], prefix="", prefix_sep="")

# add the neighborhood column back to the dataframe:
ny_onehot['Neighborhood'] = ny_venues['Neighborhood']

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

ny_onehot.head()
```

	Zoo Exhibit	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Aquarium	Arcade
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Methodology

Data Visualisation

- The neighbourhoods of New York and Toronto were plotted on a map using Folium.
- Folium is an interactive maps generator package in pandas.
- Maps were created for New York and Toronto neighbourhoods and also for the clustering of the neighbourhoods.
- For the mapping on New York and Toronto, geopy was used to obtain the latitude and longitude of both cities and this was subsequently utilised to generate the maps of New York and Toronto City.

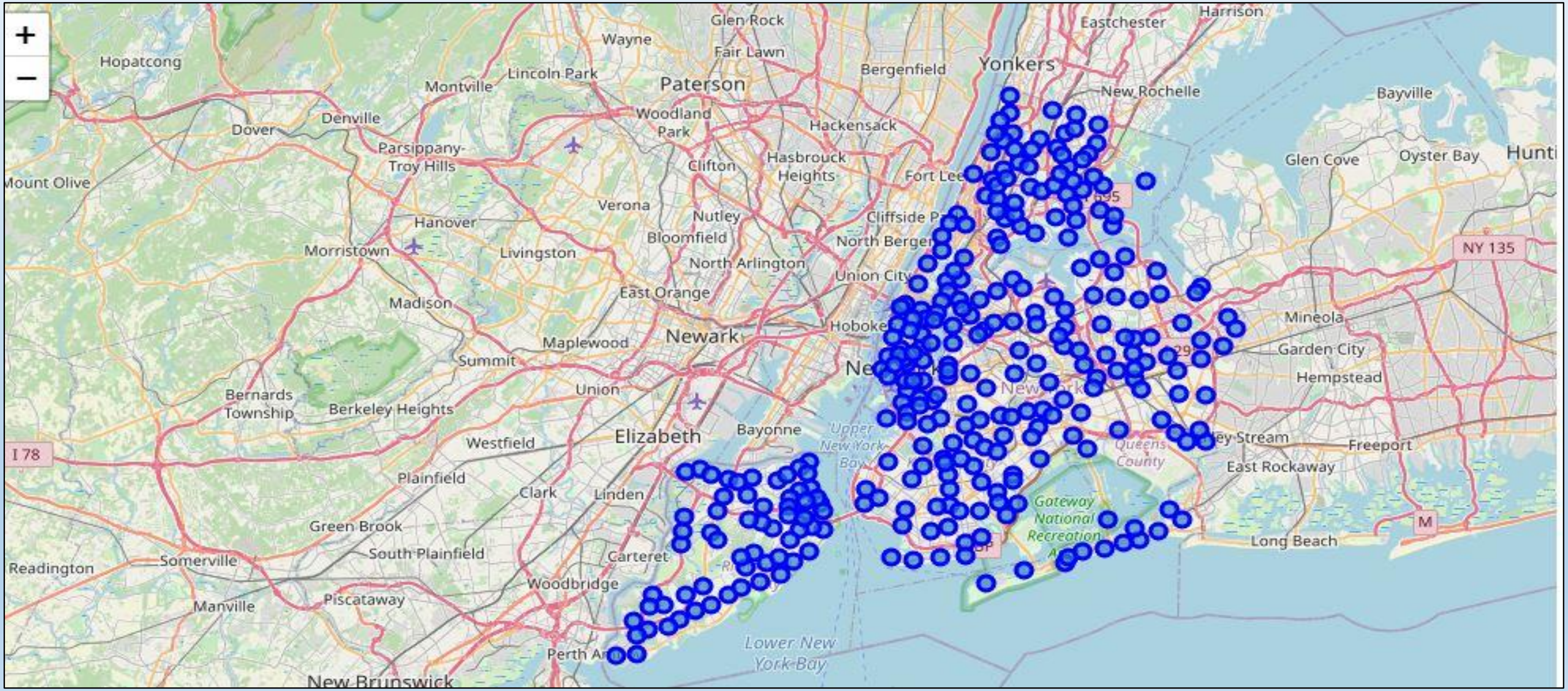


Use Geopy to get the latitude & longitude coordinates for Toronto:

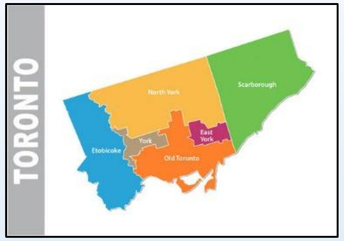
```
# Get the latitude & longitude for Toronto using the geolocator:
address = 'Toronto, CA'
geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Toronto are {}, {}.'.format(latitude, longitude))
```

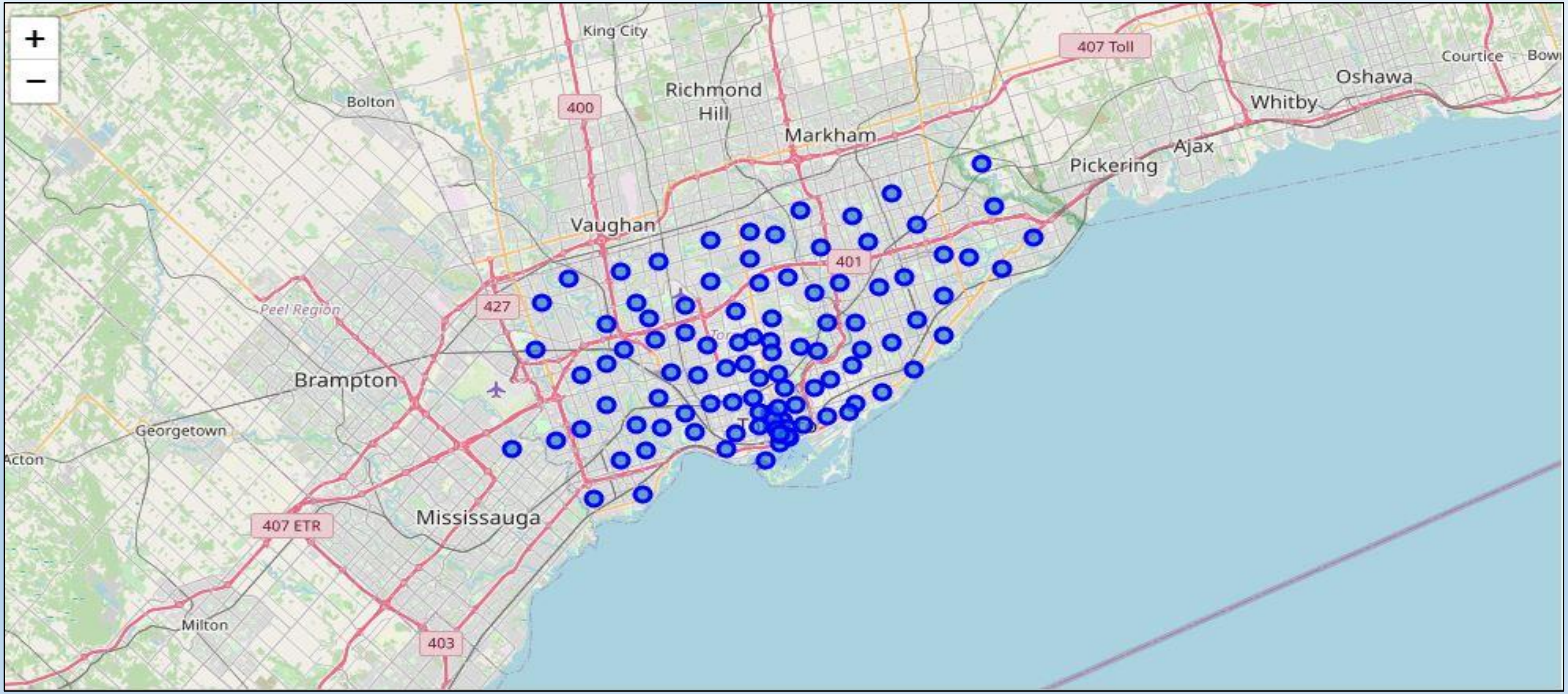
```
The geograpical coordinate of Toronto are 43.6534817, -79.3839347.
```

Data Visualisation - Map of New York



Data Visualisation - Map of Toronto

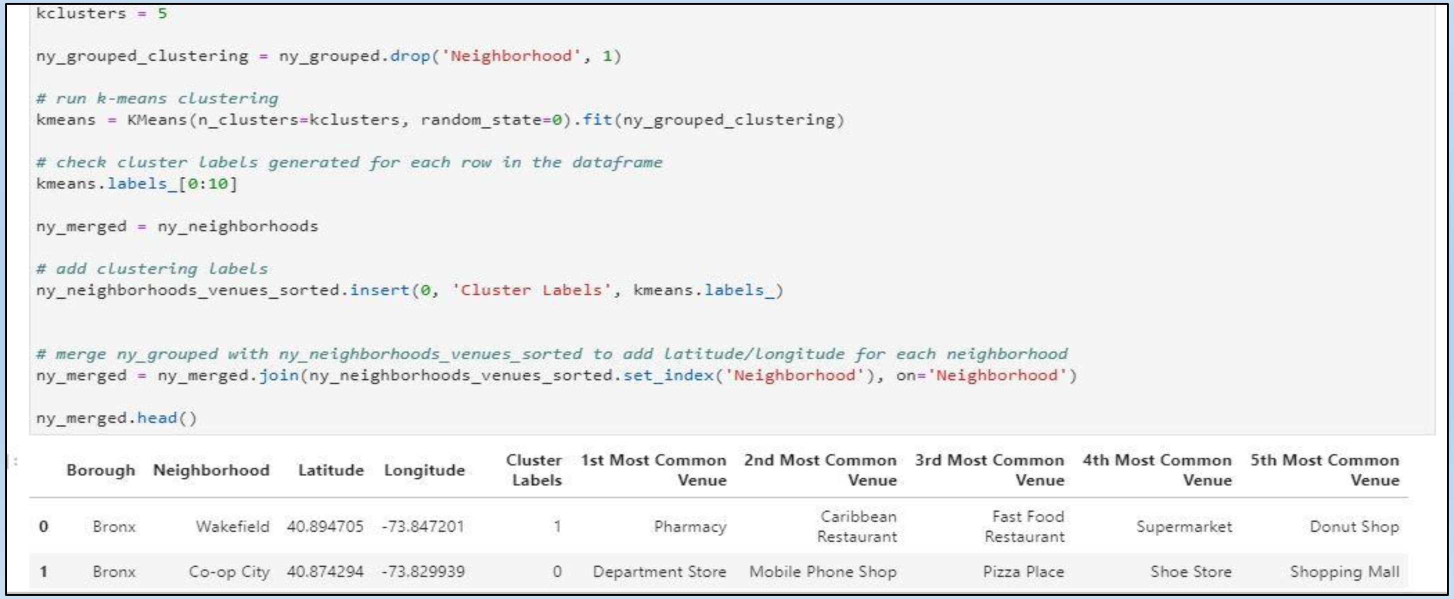




Methodology

Machine Learning – K-means clustering

- The neighbourhoods were compared using clustering and segmentation by an unsupervised machine learning technique called k-means clustering.
- The *k*-means clustering method is an unsupervised machine learning technique used to identify clusters of data objects in a dataset.
- Sci-Kit Learn is the package used for this algorithm and for this project I chose the number of cluster for both cities to be five.
- The output of the clustering is a label for each neighbourhood indicating to which cluster this neighbourhood belongs. These clusters were then visualized using folium.

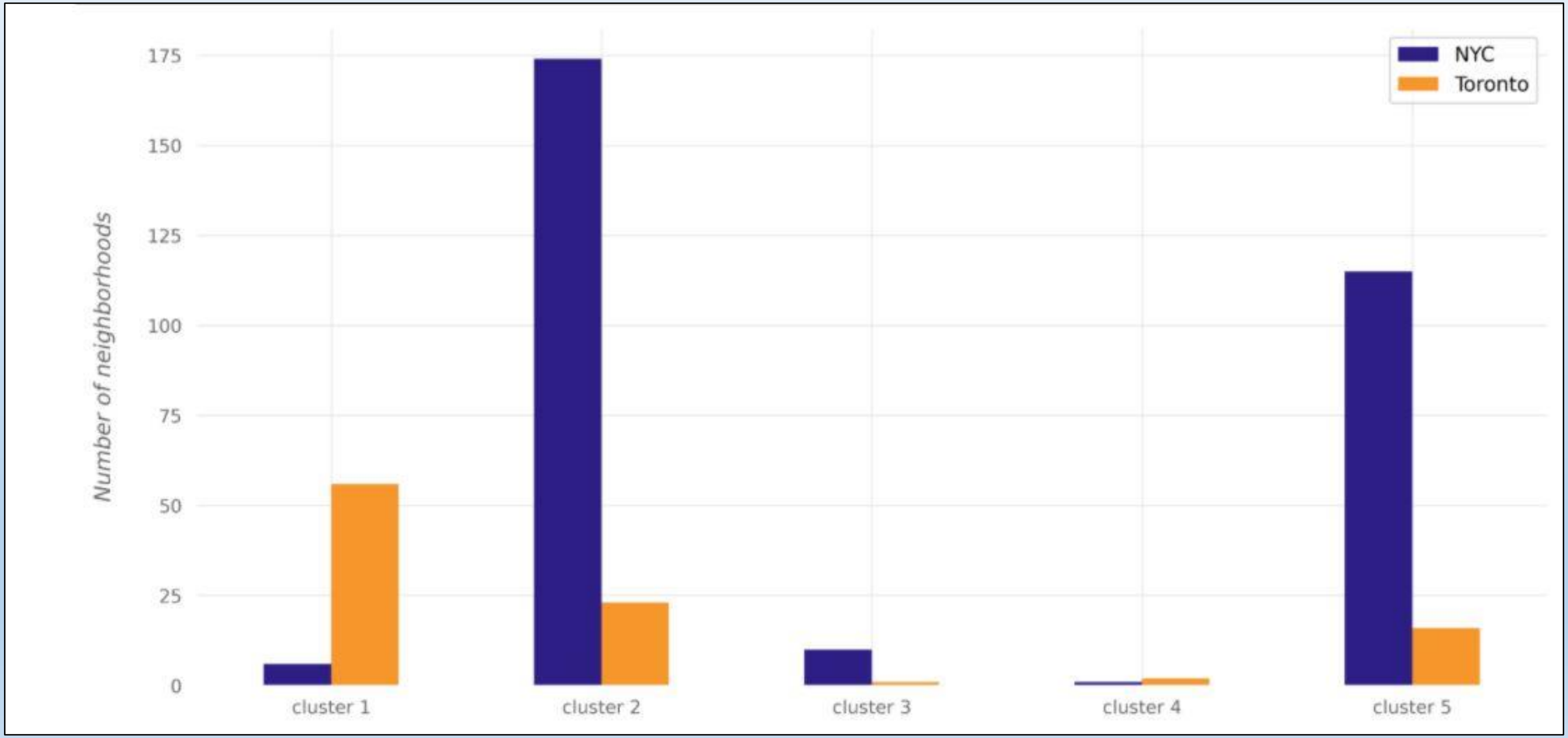


```
kclusters = 5

ny_grouped_clustering = ny_grouped.drop('Neighborhood', 1)

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

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

ny_merged = ny_neighborhoods

# add clustering labels
ny_neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)


# merge ny_grouped with ny_neighborhoods_venues_sorted to add latitude/longitude for each neighborhood
ny_merged = ny_merged.join(ny_neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')

ny_merged.head()
```

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bronx	Wakefield	40.894705	-73.847201	1	Pharmacy	Caribbean Restaurant	Fast Food Restaurant	Supermarket	Donut Shop
1	Bronx	Co-op City	40.874294	-73.829939	0	Department Store	Mobile Phone Shop	Pizza Place	Shoe Store	Shopping Mall

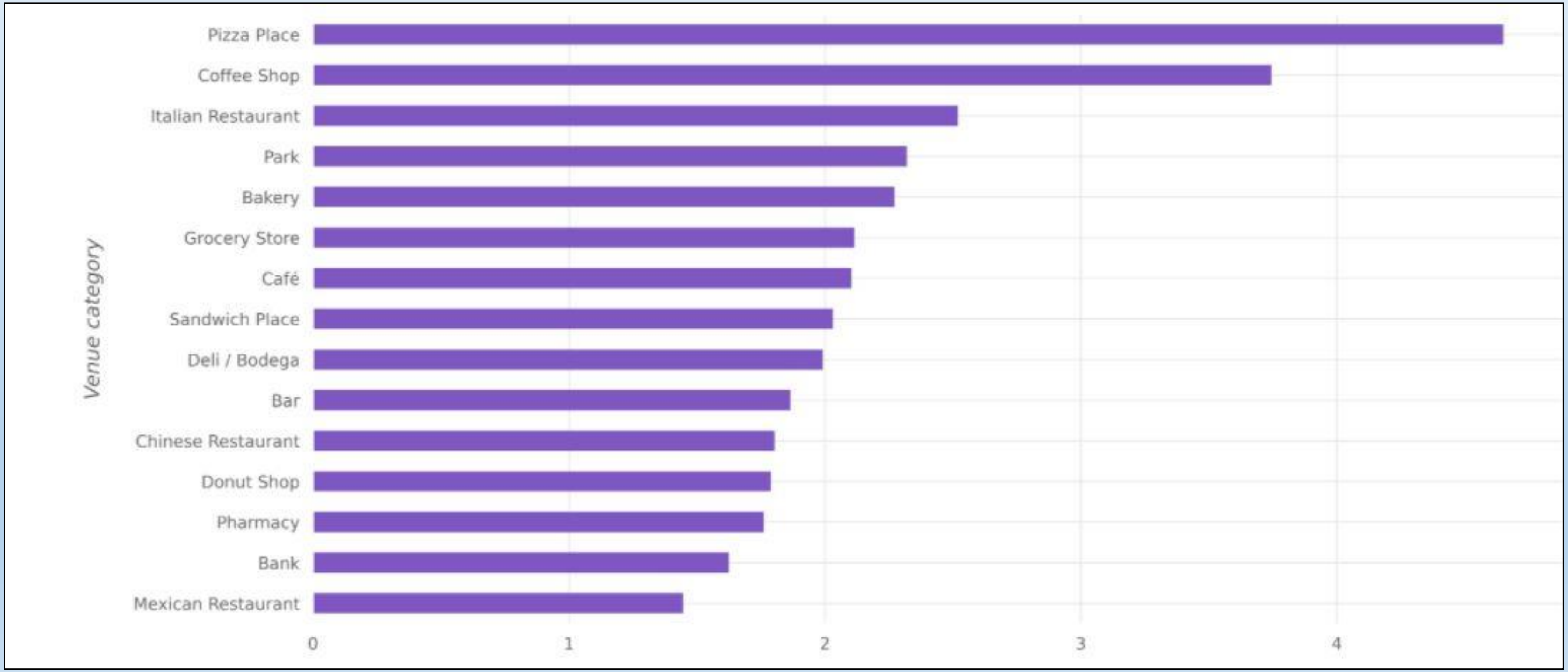
Results

Bar plot of the clusters for the combined New York-Toronto dataframe



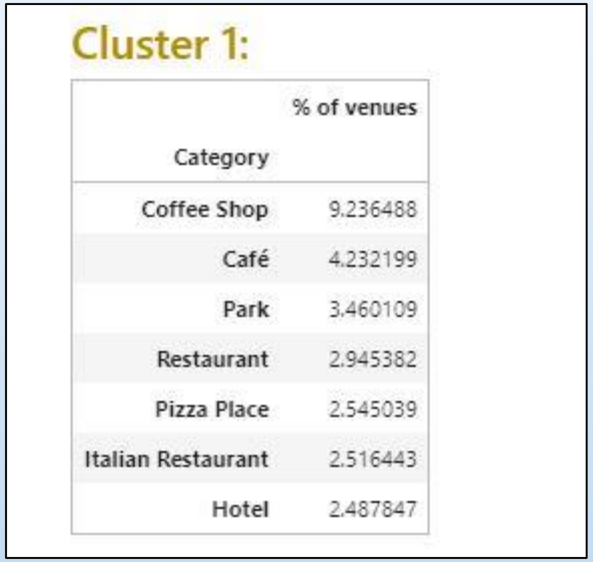
Results

Bar plot of the most common venues for the combined New York – Toronto dataframe



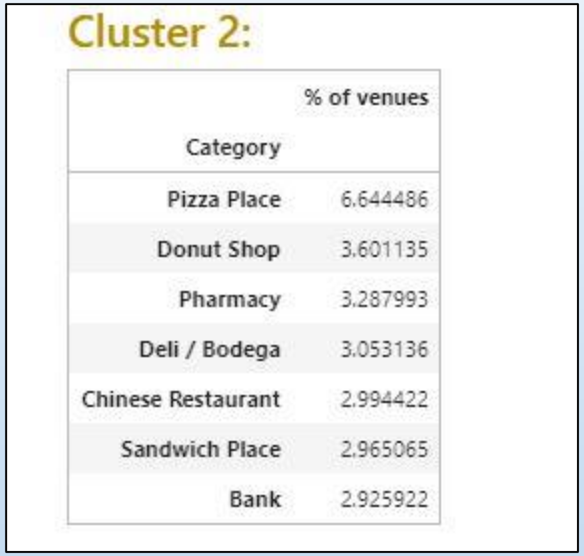
Results

Most common venues in each cluster



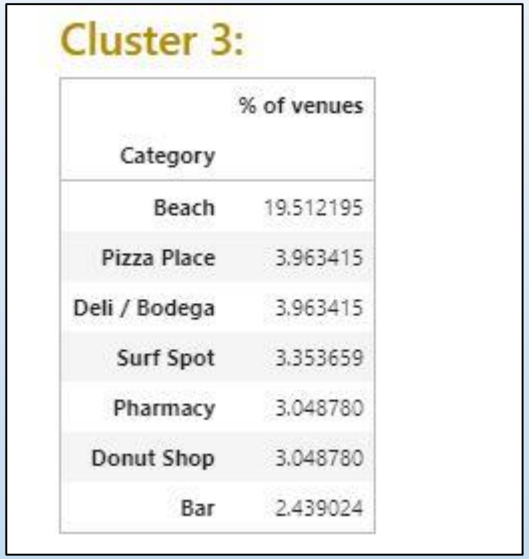
Cluster 1:

Category	% of venues
Coffee Shop	9.236488
Café	4.232199
Park	3.460109
Restaurant	2.945382
Pizza Place	2.545039
Italian Restaurant	2.516443
Hotel	2.487847



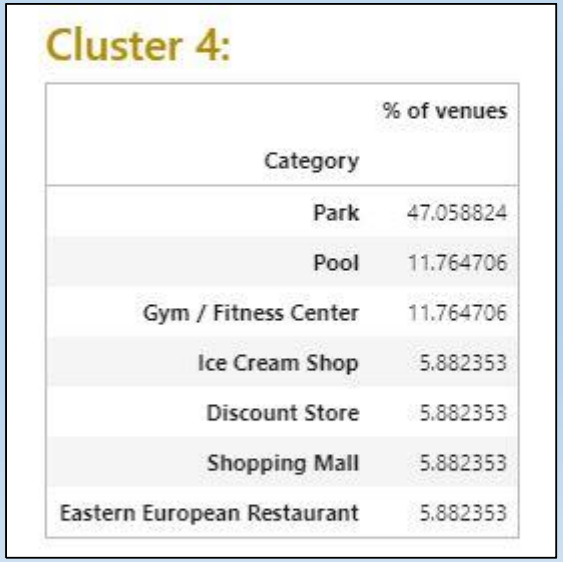
Cluster 2:

Category	% of venues
Pizza Place	6.644486
Donut Shop	3.601135
Pharmacy	3.287993
Deli / Bodega	3.053136
Chinese Restaurant	2.994422
Sandwich Place	2.965065
Bank	2.925922



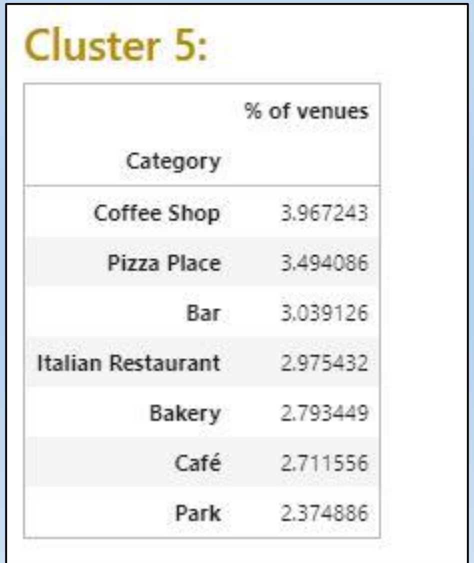
Cluster 3:

Category	% of venues
Beach	19.512195
Pizza Place	3.963415
Deli / Bodega	3.963415
Surf Spot	3.353659
Pharmacy	3.048780
Donut Shop	3.048780
Bar	2.439024



Cluster 4:

Category	% of venues
Park	47.058824
Pool	11.764706
Gym / Fitness Center	11.764706
Ice Cream Shop	5.882353
Discount Store	5.882353
Shopping Mall	5.882353
Eastern European Restaurant	5.882353



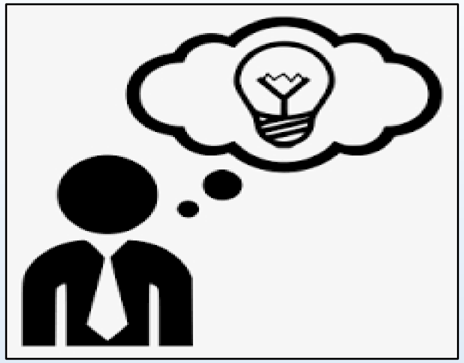
Cluster 5:

Category	% of venues
Coffee Shop	3.967243
Pizza Place	3.494086
Bar	3.039126
Italian Restaurant	2.975432
Bakery	2.793449
Café	2.711556
Park	2.374886

Discussion

- After clustering the neighbourhood’s one can easily explore the clusters and from the most common venues determine what sort of cluster it is.
- Both New York and Toronto were divided into 5 clusters.
- As can be seen from the clusters, both New York and Toronto clusters have a wide range of venues as would be expected of a large cosmopolitan city.
- Coffee shops were the most common venue for neighbourhoods in Toronto City.
- Pizza places were the most common venue for neighbourhoods in New York City.
- New York City clusters had more food venues than Toronto city clusters.
- The food venues for both cities covered a wide range of options.
- Many New York City neighbourhoods had baseball fields whereas in Toronto City neighbourhoods there were hockey stadiums and skate parks.
- Toronto City clusters had more parks and outdoor activities.
- The trendy neighbourhoods in New York had similar venues to the trendy neighbourhoods in Toronto, and these neighbourhoods fell into one cluster when the cities were combined for clustering.

Project Limitations



- This project did not look at user recommendations for common venues which can also be sourced from Foursquare. This may have broadened the project as we could have explored most recommended venues for both cities.
- Another limitation of this project is that the venue data was extracted from Foursquare during COVID-19, which may have affected the data as many places were restricting the number of customers and tourist numbers were decreased due to travel restrictions.
- The data in this study only focused on the most common venues per neighbourhood however this project could be broadened to include data such as top 10 places to visit as per data from a platform such a TripAdvisor in order to be more relevant for tourists.
- The data could also include average real estate prices per neighbourhood if one would like to compare rent prices of New York with that of Toronto, and this could also extend to include cost of living such as utilities, groceries, fuel and so forth which would be useful for anyone exploring relocation to any of these cities.

Conclusion

- Analysing, clustering and exploring cities and neighbourhoods of two large cities has revealed a basic idea of the types if venues and activities available in these cities, however this was a very generalised analysis and does not provide comprehensive information to objectively compare both cities.
- It was interesting to note the similarities and differences in the venues for both cities.
- This project provided an understanding of the application of data science principles in a real-life scenario.
- There are many ways to improve this analysis by exploring further areas such as average income, real estate pricing and cost of living.