Comparison Between NYC and Toronto

1 Introduction

In this capstone project, we will collect and analyze the neighborhoods data of two cities: One is the New York City (NYC), another is the Toronto. As the economic growth of world, the comparison of large cities such as NYC, Toronto, and Seoul has become important to manage it effectively. By using the data and Foursquare API, we will investigate which venues are common between NYC and Toronto.

By analyzing the data between two cities, we will be able to understand the differences and similarities of two cities, which will make in known to business people and people who want to live in these cities. To explain the details of the comparison, the data will be collected from 'IBM Data Science Professional Certificate' and scraped more data from the internet.

2 Data Acquisition and Processing

In this section, the acquisition and pre-processing of data will be discussed. Moreover, we will mention the analyzing methods.

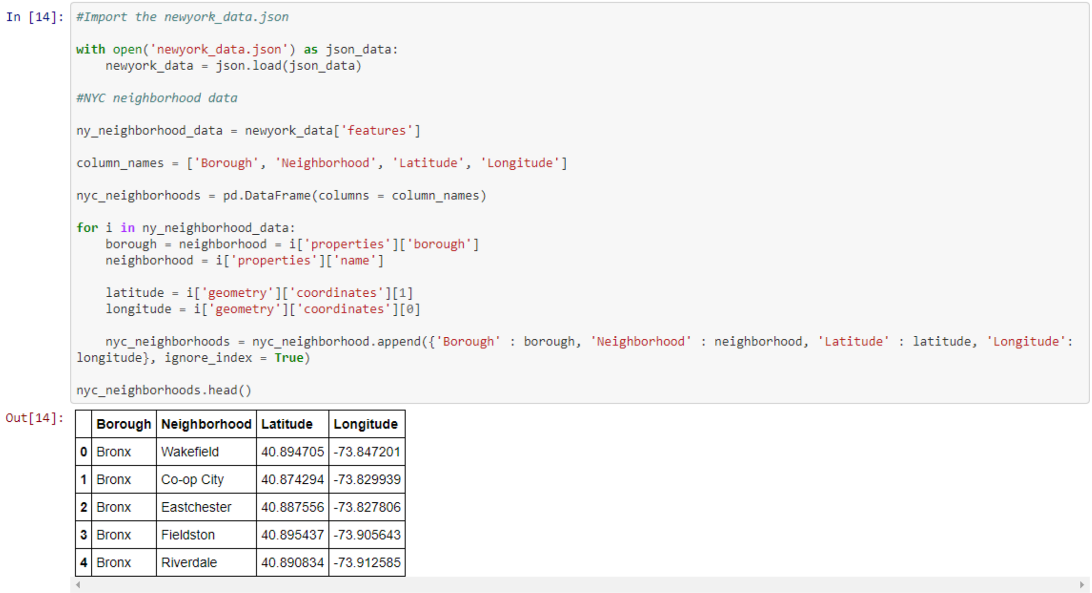
2.1 Data Acquisition

- **Neighborhood Data** : The data of two cities is imported from the "IBM Data Science Professional Certificate" and from "Wikipedia". First or all, 'Newyork data' was imported from the 'IBM developer skills network'. The neighborhood in Newyork data has a total of 5 boroughs and 306 neighborhoods. Second, Toronto data has been imported from the list of postal codes of Canada in 'Wikipedia'. Then, both data were grouped with respect to the neighborhood and borough.
- **Venues Data** : The data that describes the top 100 venues (restaurants, cafes, parks, etc.) in each neighborhood of the two cities. The data will be retrieved from Foursquare which is one of the world largest sources of location and venue data.

2.2 Data Pre-processing

Before implementing the data analysis, data should have to be processed to find and replace the missing value, and bring into the format which we want to. This process is called as the 'pre-processing' of the data, the process of converting data from raw form into another format in order to analyze it properly.

To implement the pre-processing of data, we firstly looked through the whole data of 'newyork.json'. We realized that the data consisted of longitude, latitude, borough, etc. From the whole data, as shown in the below Figure 1, we extracted the 'Borough', 'Neighborhood', 'Latitude', and 'Longitude'.



```
In [14]: #Import the newyork_data.json

         with open('newyork_data.json') as json_data:
             newyork_data = json.load(json_data)

         #NYC neighborhood data

         ny_neighborhood_data = newyork_data['features']

         column_names = ['Borough', 'Neighborhood', 'Latitude', 'Longitude']

         nyc_neighborhoods = pd.DataFrame(columns = column_names)

         for i in ny_neighborhood_data:
             borough = neighborhood = i['properties']['borough']
             neighborhood = i['properties']['name']

             latitude = i['geometry']['coordinates'][1]
             longitude = i['geometry']['coordinates'][0]

             nyc_neighborhoods = nyc_neighborhood.append({'Borough' : borough, 'Neighborhood' : neighborhood, 'Latitude' : latitude, 'Longitude':
         longitude}, ignore_index = True)

         nyc_neighborhoods.head()
```

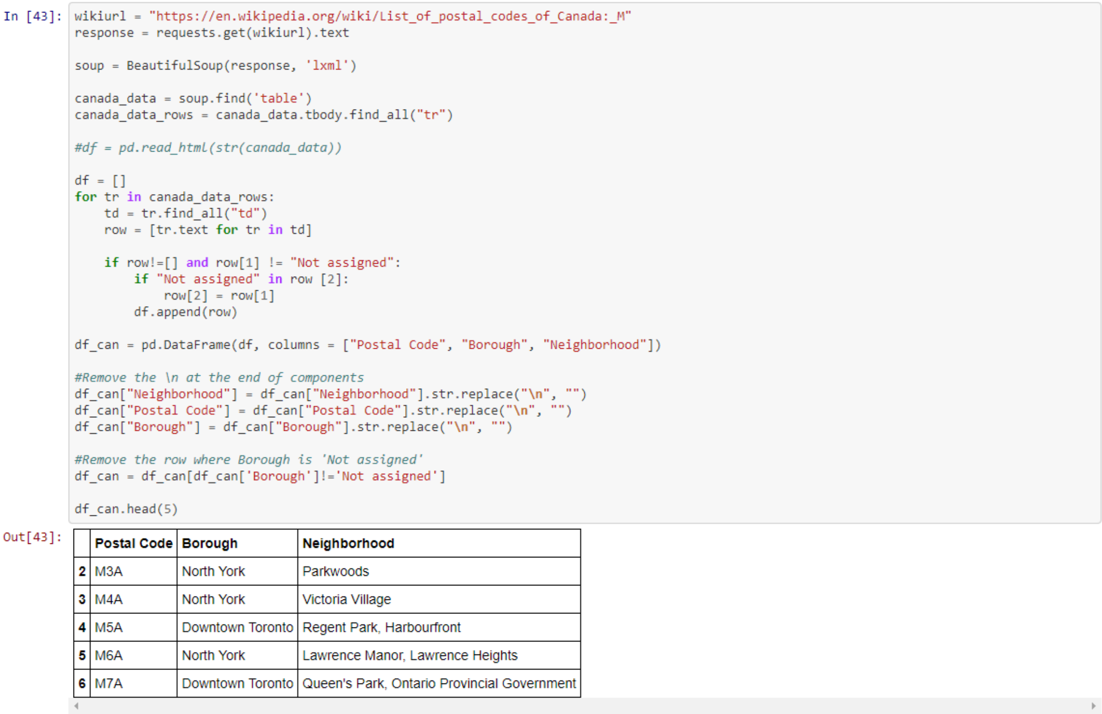
Out[14]:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Figure 1. Pre-processing of the New York data

For the Toronto data, we have imported from the Wikipedia and Foursquare API. Then, with the similar methods in Newyork data pre-processing, we implemented it and successfully organized as shown in Figure 2.

By utilizing the location information in processed data, the venue data that describes the top 100 venues (restaurants, cafes, parks, etc.) in each neighborhood of two cities. This can be easily implemented by using the Foursquare API. As described in the previous section, our purpose is to compare the most common venues between New York and Toronto. Thus, after organizing the data with the 4 attributes, we categorized the common venues from 1st to 10th most common venues as shown in Figure 3.



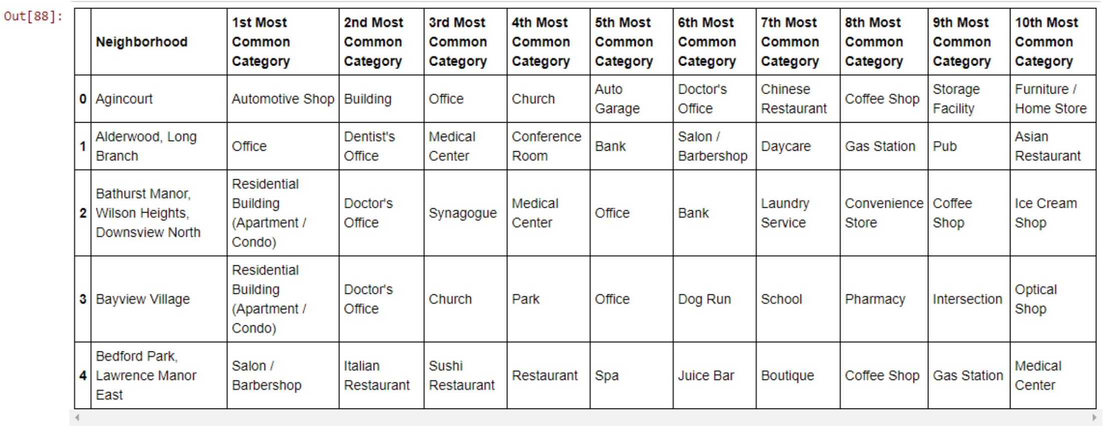
In [43]:

```
wikiurl = "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
response = requests.get(wikiurl).text

soup = BeautifulSoup(response, 'lxml')

canada_data = soup.find('table')
canada_data_rows = canada_data.tbody.find_all("tr")

#df = pd.read_html(str(canada_data))

df = []
for tr in canada_data_rows:
    td = tr.find_all("td")
    row = [tr.text for tr in td]

    if row!=[] and row[1] != "Not assigned":
        if "Not assigned" in row [2]:
            row[2] = row[1]
        df.append(row)

df_can = pd.DataFrame(df, columns = ["Postal Code", "Borough", "Neighborhood"])

#Remove the \n at the end of components
df_can["Neighborhood"] = df_can["Neighborhood"].str.replace("\n", "")
df_can["Postal Code"] = df_can["Postal Code"].str.replace("\n", "")
df_can["Borough"] = df_can["Borough"].str.replace("\n", "")

#Remove the row where Borough is 'Not assigned'
df_can = df_can[df_can['Borough']!='Not assigned']

df_can.head(5)
```

Out[43]:

	Postal Code	Borough	Neighborhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Figure 2. Pre-processing of the Toronto data



Out[88]:

	Neighborhood	1st Most Common Category	2nd Most Common Category	3rd Most Common Category	4th Most Common Category	5th Most Common Category	6th Most Common Category	7th Most Common Category	8th Most Common Category	9th Most Common Category	10th Most Common Category
0	Agincourt	Automotive Shop	Building	Office	Church	Auto Garage	Doctor's Office	Chinese Restaurant	Coffee Shop	Storage Facility	Furniture / Home Store
1	Alderwood, Long Branch	Office	Dentist's Office	Medical Center	Conference Room	Bank	Salon / Barbershop	Daycare	Gas Station	Pub	Asian Restaurant
2	Bathurst Manor, Wilson Heights, Downsview North	Residential Building (Apartment / Condo)	Doctor's Office	Synagogue	Medical Center	Office	Bank	Laundry Service	Convenience Store	Coffee Shop	Ice Cream Shop
3	Bayview Village	Residential Building (Apartment / Condo)	Doctor's Office	Church	Park	Office	Dog Run	School	Pharmacy	Intersection	Optical Shop
4	Bedford Park, Lawrence Manor East	Salon / Barbershop	Italian Restaurant	Sushi Restaurant	Restaurant	Spa	Juice Bar	Boutique	Coffee Shop	Gas Station	Medical Center

Figure 3. Organizing the data from 1st most common venues to 10th most common venues

3 Data Analysis

To analyze the data, we plotted the categorized data with respect to the number of venues in Toronto and New York. In the NYC, the world most populated cities, Residential Building is selected as the most common venues, and the Salon/Barbershop, Doctor's office followed (Figure 4). We can conjecture the reason of this. The reason is that the city needs to accomodate more than ten millions people. Thus, a lot of apartment and condo, and life related venues (restaurant, hospital, salon, etc.) have to be the priority in the city.

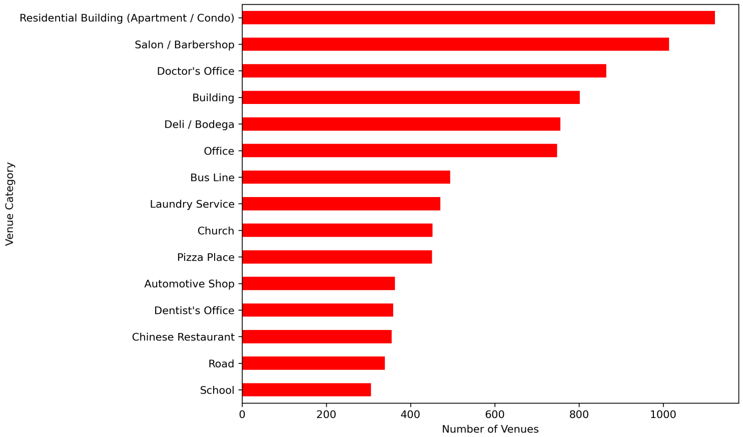


Figure 4. Most common venues in New York City

Similarly, as shown in the Figure 5, Toronto shows the similar distribution of the most common venues.

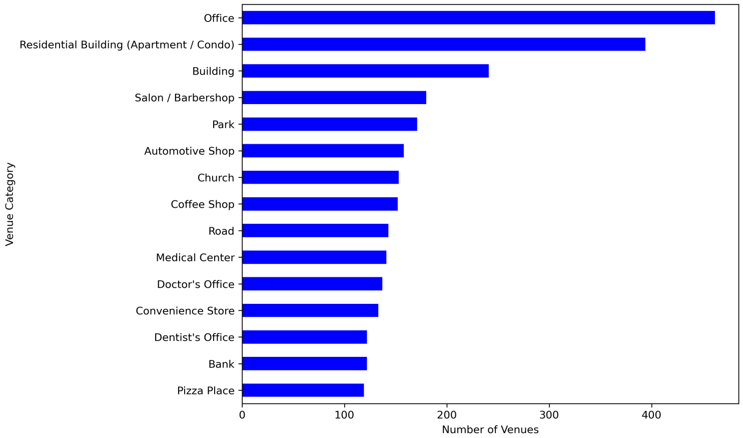


Figure 5. Most common venues in Toronto

4 Conclusion

In conclusion, it is a little bit mature to conclude but the data represents that the big cities show the similar patterns in these days. Due to the large number of population, they need to build more residential building like apartment and then they need to have convenient facilities to satisfy the people in the cities. From now on, this kind of fact is very interesting and important to build a city effectively due to the explosive increment of population in the dense area.