Coursera Capstone Project

Comparing the population of New York City and Toronto city

Project presented by: Vismay Patel

Table of Content:

Introduction / Business Problem:

In this project we will focus on the population comparision of the two cities, Canada and New York.

New York and Toronto, are both large and developed cities, and populated plays an essention role in determing following factors:

1. How much GDP the people of the region / city are contributing to.

2. Populus cities can create employment and job opportunities can come up, since the economic contribution of the city is expected to increase.

3. The government can generate more tax, and funding for the wellfare of the nation can be increased, which includes development and construction of new fecilities and good exchange of goods and services.

The current metro area population of Toronto in 2020 is 6,197,000, a 0.94% increase from 2019. The metro area population of Toronto in 2019 was 6,139,000, a 0.94% increase from 2018. The metro area population of Toronto in 2018 was 6,082,000, a 1.2% increase from 2017

New York City (NYC), often called simply New York, is the most populous city in the United States. With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles (784 km2), New York City is also the most densely populated major city in the United States.

Problem Statement:

1. Comapre the number of neighbourhoods each city has.

2. Derive insights from that data to predict which city could be more populated in coming years.

Data Collection:

The data for Toronto will be collected from wikipedia and the data for New York city will be used from IBM database repository. (The reason to opt for IBM data repository is, there is a library / system error which refrains the "groups" keyword to be executed while structuring venue from Foursquare API.)

Link for CANADA DATASET:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Link for NEW YORK dataset: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

Methodology:

1. Data extraction : Data wrangling, using beautiful soap and extract data from IBM data base

2. Data Cleaning / Data Preprocessing : Transforming the raw data into Data Frame and structuring the data using json.

3. Data Analysis/ Modeling : Performing visulaizations using folium library to represent data points of neighborhoods from both datasets. Generating bar graphs.

Data extraction and Pre-Processing:

Methods used for data extraction

Toronto dataset: data wrangling, Beautiful Soap

New York dataset: extracted from IBM database

Pre-process the data by converting it into data frame.

This was done using Beautiful soap and basic merging and drop operations.

The Toronto dataset was converted by beautiful soap.

New York data was converted by specifying selection parameters

Merging the dataset and modifying the data frame:

The data frame for both the data sets was modified, to merge the location attributes.

That is Latitude, Longitude were merged in Toronto dataset using a geographic location .csv file.

```
import types
import pandas as pd
from botocore.client import Config
import ibm_boto3

def __iter__(self): return 0

# @hidden_cell
# The following code accesses a file in your IBM Cloud Object Storage. It include
s your credentials.
# You might want to remove those credentials before you share the notebook.
client_5fe7869a6ea44b6484b25ae323282f99 = ibm_boto3.client(service_name='s3',
    ibm_api_key_id='h2S6E-y8Nmt_Sdq21QPiiOeEnN1QA4mX1egnTur29iJZ',
    ibm_auth_endpoint="https://iam.cloud.ibm.com/oidc/token",
    config=Config(signature_version='oauth'),
    endpoint_url='https://s3-api.us-geo.objectstorage.service.networklayer.com')

body = client_5fe7869a6ea44b6484b25ae323282f99.get_object(Bucket='capstoneproject
-donotdelete-pr-yspbcgfydcoh1v',Key='Geospatial_Coordinates.csv')['Body']
# add missing __iter__ method, so pandas accepts body as file-like object
if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType( __iter__, bod
y )

df_data_1 = pd.read_csv(body)
df_data_1.head()
```

	Postal Code	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

Figure 1.0

```
toronto_df = pd.merge(web_df1, df_data_1, how = 'left', left_on = 'PostalCode', r
ight_on = 'Postal Code')
toronto_df.drop("Postal Code", axis = 1, inplace = True)
toronto_df
```

	PostalCode	Borough	Neighborhood	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

Figure 1.1

New York Data Set extraction:

New York data was extracted by using IBM dataset using the json file.

Then the features of the neighborhoods were extracted so that they could be later structured in to a data frame. Further the data was visualized by using folium library.

```
!wget -q -O 'newyork_data.json' https://cf-courses-data.s3.us.cloud-object-storag
print('Data downloaded!')
```

```
Data downloaded!
```

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

newyork_data
```

```
{'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',
```

```
# declare the features in the dataset
neighborhoods_data = newyork_data['features']
```

```
neighborhoods_data[0]
```

Figure 1.2

Data Visualization and modeling data:

The folium map of neighborhoods was generated by using the geolocator and geocoder instance. This will give us insight to probelm 1 solution.

To find the number of neighborhoods and compare which city has more neighborhoods

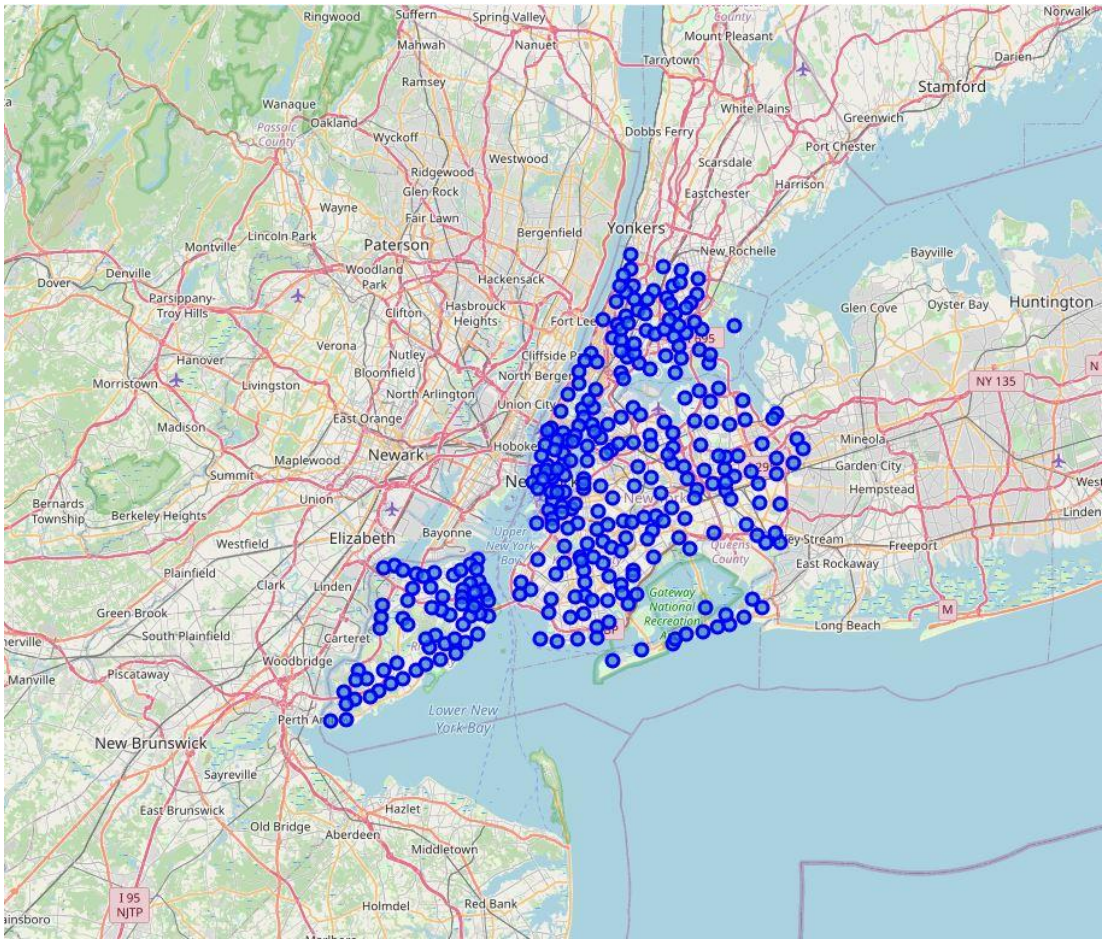

Figure 1.3

The above map shows all the neighborhoods of newyork city. This is generated from folium library.

The next visualization will be for the Toronto data set, consisting of all the neighborhoods. Comparing the neighborhoods will help us to derive at the conclusion and get better insights about the population of the two cities and their growth in the future.

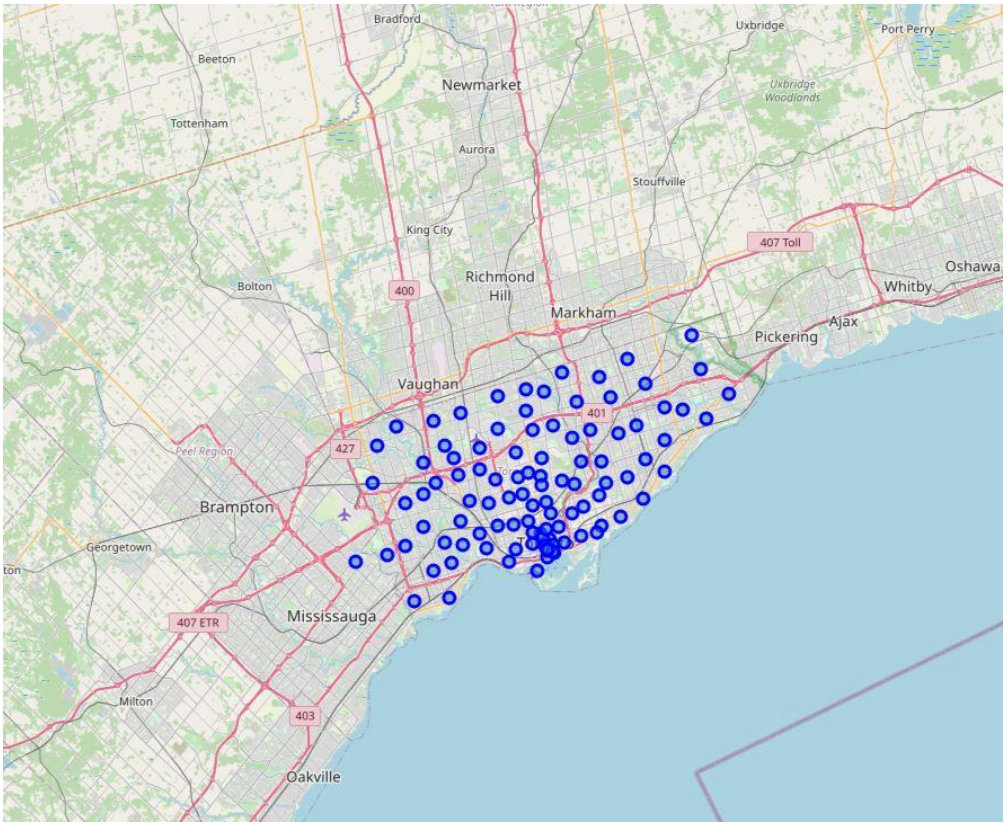

Figure 1.4

From the above graph we can clearly see that the number of neighborhoods in NYC are more than that of Totonto city.

To get better insights about which borough has higher neighborhoods, bar plots were generated using matplotlib.pyplot library.

Bar plot for New York data:

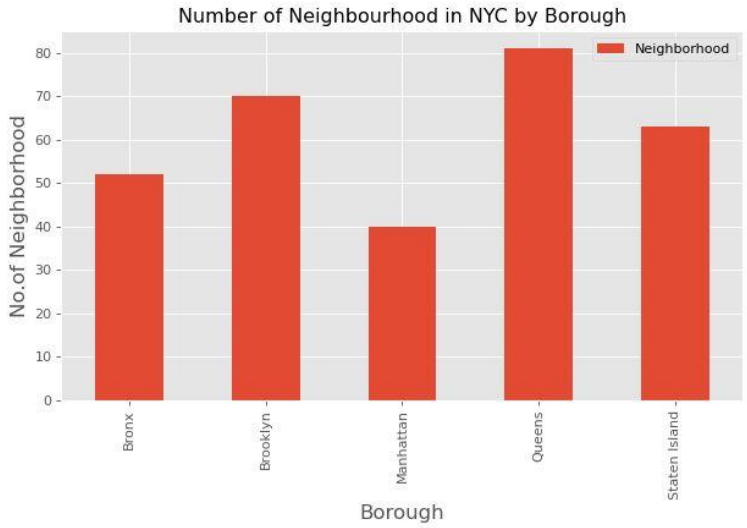

Figure 1.5

Bar Plot for Toronto dataset:

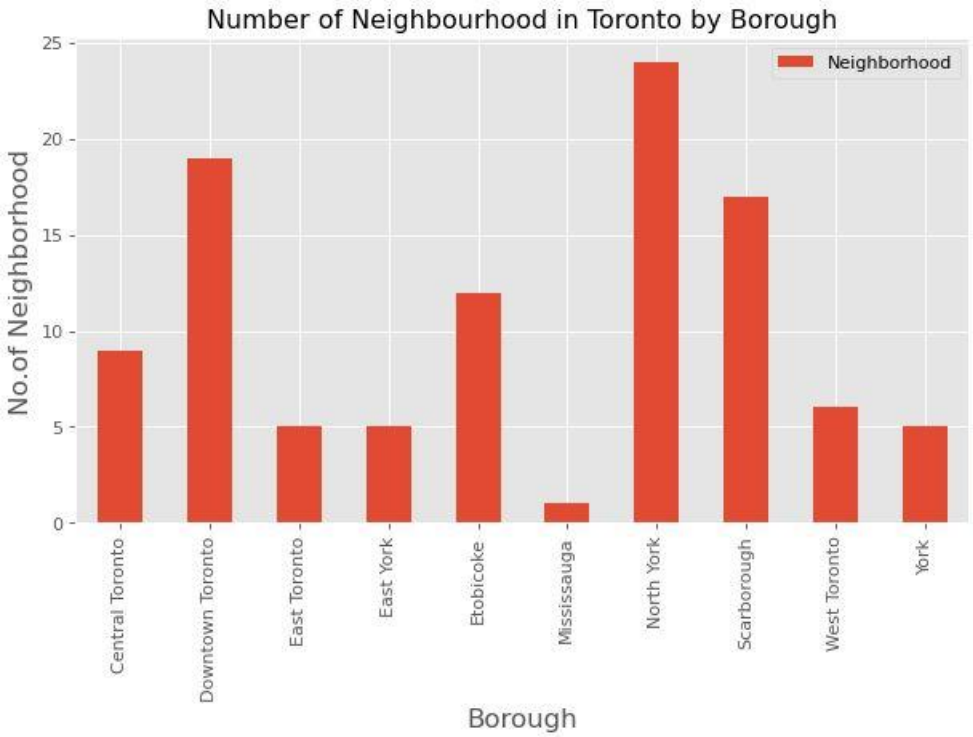

Figure 1.6

Next, the most populated Borough from both the cities were picked to get further insight to the data. From New York data the selected Borough was Queens. From Toronto data the selected borough was North York.

From Queens and North York we will find the total number of venues to understand how populated neighborhoods can help in understanding consumer goods relationship.

Results:

So, in the final stage we generate the neighborhoods of both the boroughs Queens and North York, with the help of folium library.

Below is the code with which the folium plot was generated. It also shows how we need to assign the address first and get the Latitude and Longitude values of the location. Once we have the locations we can generate the folium map, with those locations.

```
address = 'Queens, NY'

geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Queens are {}, {}.'.format(latitude, longitu
```

```
The geograpical coordinate of Queens are 40.7498243, -73.7976337.
```

```
# create map of Manhattan using latitude and longitude values
map_Queens = folium.Map(location=[latitude, longitude], zoom_start=11)

# add markers to map
for lat, lng, label in zip(Queens_data['Latitude'], Queens_data['Longitude'], Que
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_Queens)

map_Queens
```

Figure 1.7

The first visualization is of Queens Borough, and the second visualization will be for North York Borough.

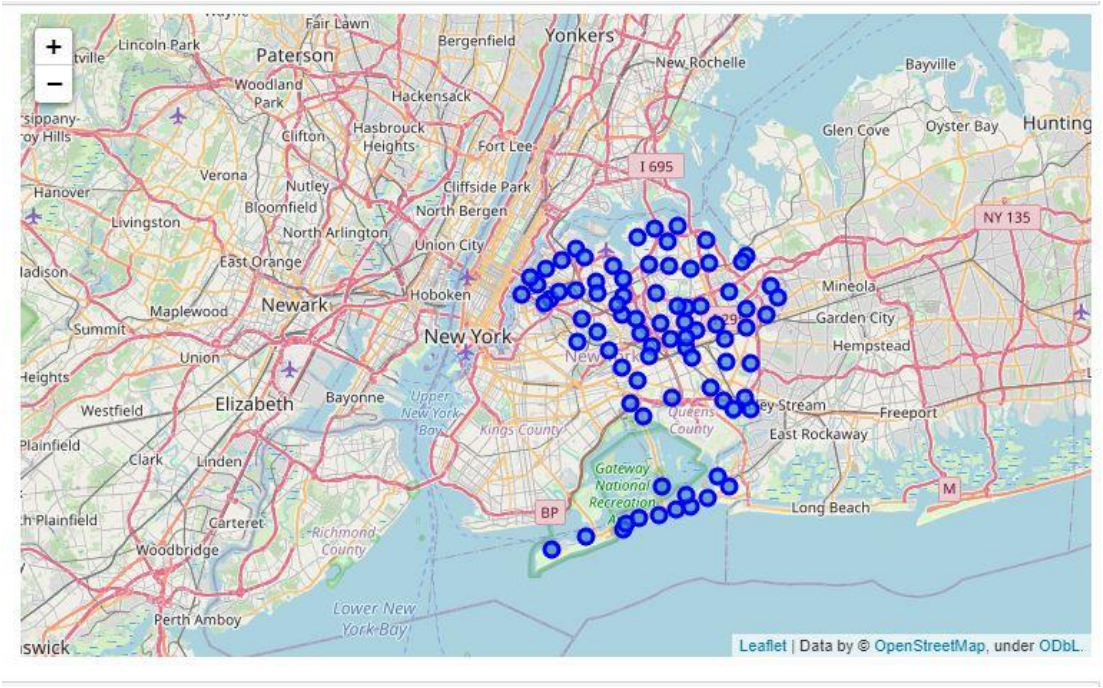

Figure 1.8

North York folium visualization:

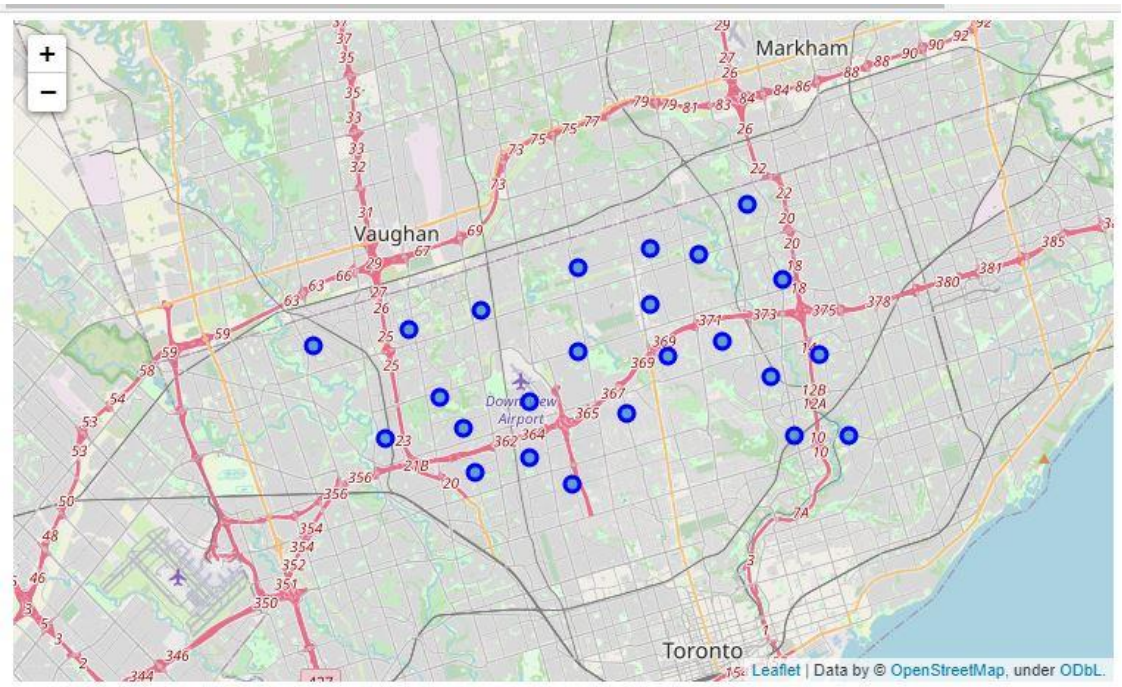

Figure 1.9

Improvement and discussion:

The comparision of both the cities can be done more effectively by generating the most trending venues, based on that we can understand the preferences of the people in both the cities.

Unfortunately the "groups" KeyWord always throws error after connecting to Foresquare API while generating venues using latitude and longitude. Despite numerous attempts to restart, stop the kernel or generate a new notebook.

Conclusion:

The main focus of the project was to compare the number of neighborhoods present in New York City and Toronto City.

This was done by fetching the data from the web and by performing the Data Science techinques. Folium maps are generated to present the number of neighborhoods of both the cities. Moreover, to get more clarity on the understanding of the data generated on the maps, bar plots were generated to find out the most populas borough with maximum neighborhoods.

Queens in New york city and North York in toronto city had the highest number of neighborhoods.

Thus, we can say that the Boroughs in NYC are more likely to contribute higher to the city's GDP, Employment, tax, and other factors. On the other hand toronto city boroughs do not have similar impact on the contribution they can have on GDP, Employment, tax etc.

We can clearly see that in the bar graphs since the total number of neighborhoods are less for toronto, than NYC.