IBM Data Science Professional Certificate

Capstone Project

Similarities and differences between neighbourhoods of New York and Toronto

Part 1 and Part 2

Author: Tomasz Tomaszewski

May, 2021

PART 1

Introduction

This work is a capstone project submission for "IBM Data Science Professional Certificate" online course on Coursera. Additionally it can showcase my knowledge and skills in data science/data analysis of real-world datasets and different scenarios.

In this report I will try to compare the cities of Toronto (Canada) and New York (USA). I will find some public datasets, explore them, anylize them and visualise them. All this work will be later presented in the form of final report presentation document.

Comparison of neighbourhoods of those cities will help to get insights on what kind of venues and points of interests are common for both cities and what kinds are very different between the cities.

New York, with an estimated 2019 population of over 8 millions is the most populous city in the United States and also the most densely populated major city in the United States. New York City serves as the cultural and financial capital of USA and possibly the world.

The second city of interest is Toronto, the capital of Canadian province of Ontario. The city it is the most populous city in Canada and the fourth most populous city in North America, with population of over 2.5 millions. Like New York, Toronto is also an international centre of business and culture, and is widely renowned as one of the most multicultural and cosmopolitan cities in the world.

Business problem

The report should give readers better understanding of similarities and differences between the two cities. This in turn will help to find suitable location either if one is interested in setting up a business in the city, or considers moving to the city or wants to visit city as a tourist.

Both final deliverable and the notebook can facilitate making decisions for many various stakeholders, including for example:
1) Students considering taking studies in any of the city in question.
2) Somebody who got an job offer in either NY or Toronto and would like to get to know the city before making decision to move.

3) Business management - exploration of the city could help in making decision in which district would be best to open a business (like bank, coffee shop, restaurant etc).
4) Citizens of various countries could find the area of the city that suits them best before they decide to move.
5) Turists will find it easier to make decision what to visit in the city or which city to choose for a visit.

PART 2

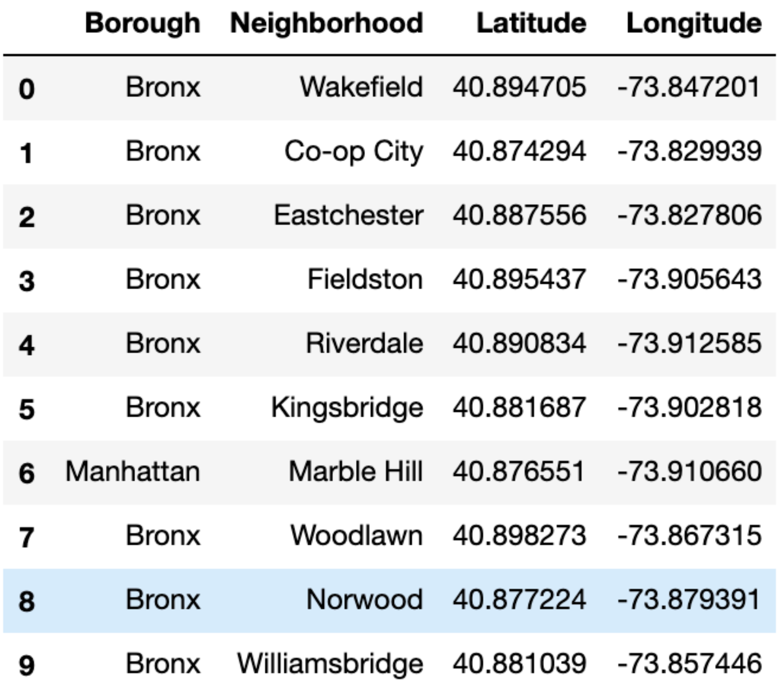
Description of data.

During earlier labs and courses modules of IBM Data Science Professional Certifficate there were many datasets presented and explored. Specifically for preparation of this report following sources of data will be used.

Datasets listing names of the neighborhoods of New Yorkand Toronto and their latitude and longitude coordinates.

For New York the source was provided by course teachers and the data set is extracted in the form of json file from following url:

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

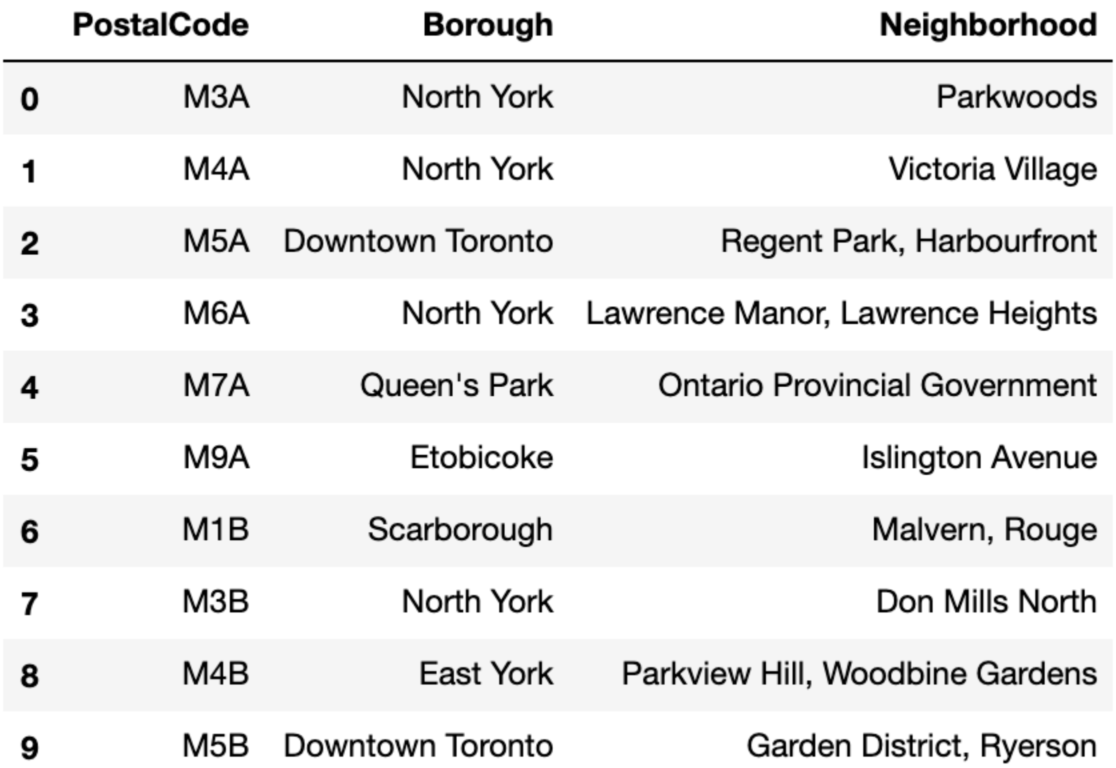


	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446

Data for Toronto were extracted via web scrapping technique with help of BeautifulSoup. Data source is the following link:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Thereafter Pandas dataframe has been created (see picture below):



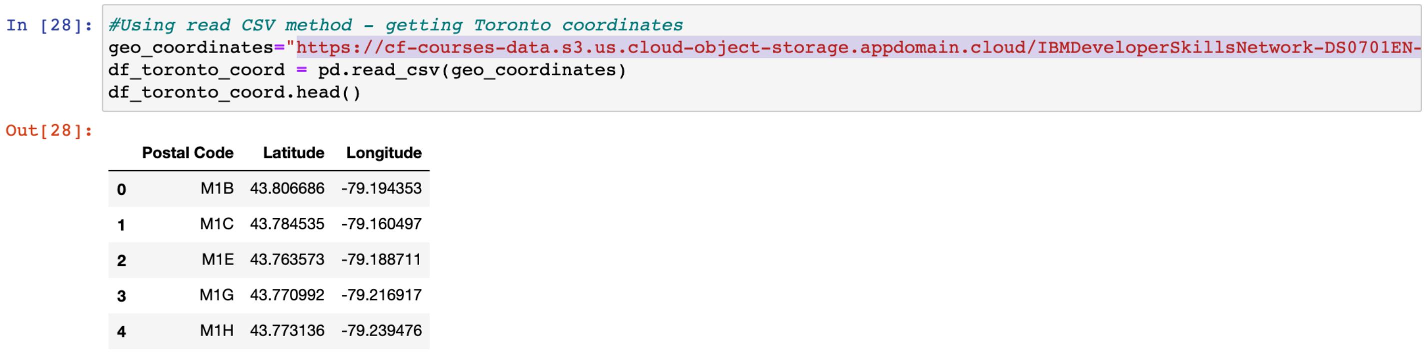
	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park	Ontario Provincial Government
5	M9A	Etobicoke	Islington Avenue
6	M1B	Scarborough	Malvern, Rouge
7	M3B	North York	Don Mills North
8	M4B	East York	Parkview Hill, Woodbine Gardens
9	M5B	Downtown Toronto	Garden District, Ryerson

The source of data, hence also dataframe contains only PostalCode, Borough and Neighbourhood of Toronto. This problem will be solved by finding latitude and longitude data.

In order to make the data source complete, geospatial coordinates of Toronto were extracted from a csv file (file and location provided in IBM Data Science course):

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv

See the snipets of the code for geospatial coordinates extraction:

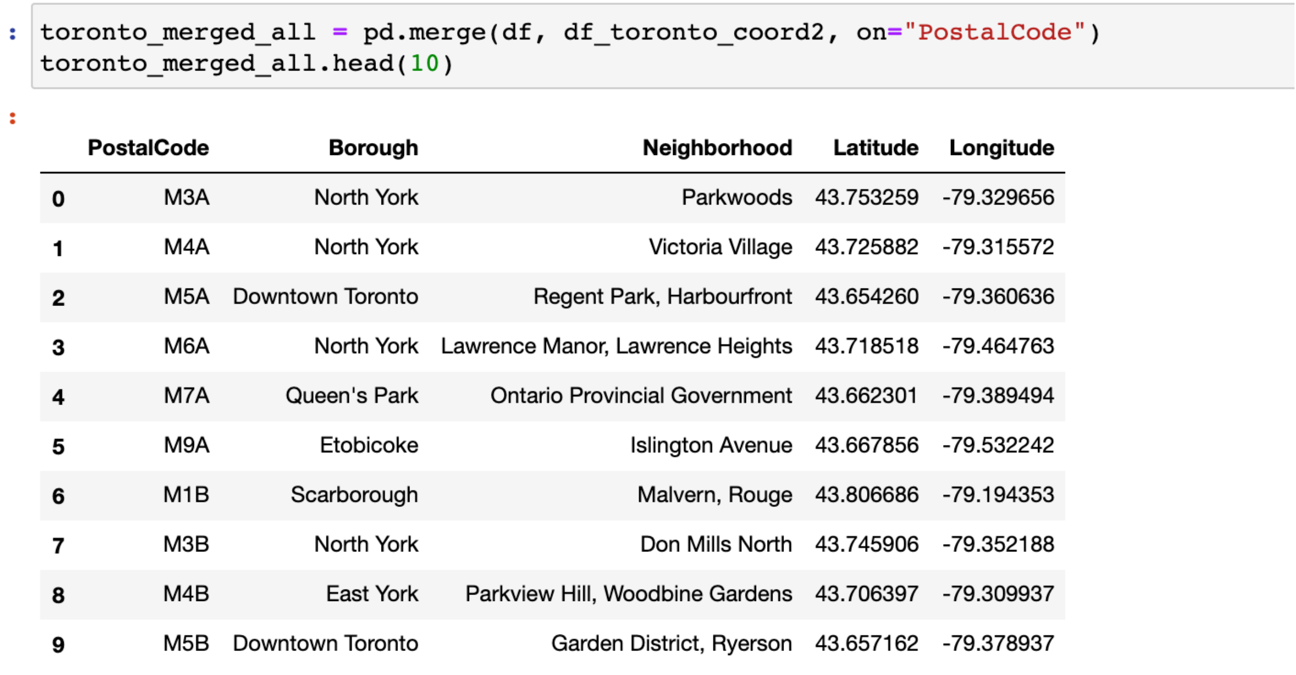


```
In [28]: #Using read CSV method - getting Toronto coordinates
         geo_coordinates="https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-
         df_toronto_coord = pd.read_csv(geo_coordinates)
         df_toronto_coord.head()
```

Out[28]:

	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

Finally, both Toronto data frames were merged together which resulted in final Toronto data frame similar to the data for New York (see below).



```
: toronto_merged_all = pd.merge(df, df_toronto_coord2, on="PostalCode")
  toronto_merged_all.head(10)
```

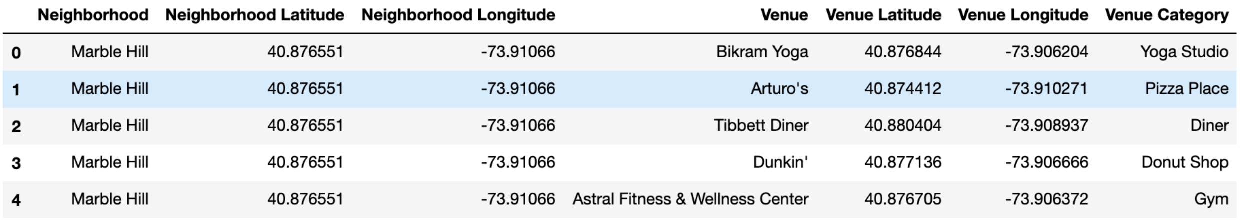
:

	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
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills North	43.745906	-79.352188
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937

Venues data were downloaded with help of Foursquare API, which is popular source of venue data and location data and utilisation of this tool was introduced during the course.

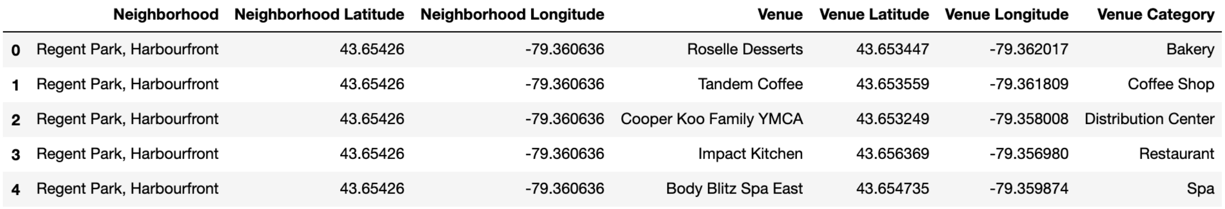
Different numbers of venues were found in different neighborhoods for respective city. Data were also saved in the form of pandas data frame.

Foursqure date retrieved for New York:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
1	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop
4	Marble Hill	40.876551	-73.91066	Astral Fitness & Wellness Center	40.876705	-73.906372	Gym

Foursqure date retrieved for Toronto:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.656369	-79.356980	Restaurant
4	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa

All retrieved data will be later wrangled, processed and analysed in later parts of this project and report in coming sections.