Comparing big multicultural city neighborhoods

Ivelina Yordanova

February 2, 2020

1. Introduction

a. Background

As of today, the most widely used measurement of economic growth and prosperity of a country is the GDP. The Gross Domestic Product (GDP) is “the total monetary or market value of all the finished goods and services produced within a country's borders in a specific time period”. Two of the top ten countries with highest GDP last year are in North America, the USA (at first place) and Canada (at tenth place). Overall the stats one can find online make it seem that both have a lot of similarities and whatever one country lack in a particular area, it makes up for in another. Often taxes are a key differentiator - Canada has an average tax rate of 28%, which is higher than the 18% in the United States. Canadians bring home $35,299 annually on average, whereas an average post-tax annual salary is $52,344. On the other hand, that’s compensated by the cost of living - rent for a one-bedroom apartment in Toronto consts $1,536.22 vs. $3,116.43 in New York City. It is interesting to compare how similar the life in those countries using the Foursquare API data for the venues in those cities and their location within the cities.

b. Problem

The problem this report will aim to solve is to find whether big multicultural cities tend to have a similar way of developing, a common cultural scene and common preferences of the people living there.
Analyzing the clustering of different venues by category one can find whether a specific type tend to be located in the center of the city or it's suburban areas.

c. Interest

Social scientists, anthropologists and economists would be interested in this information to make certain conclusions on how people live and how the urban lifestyle might develop. Another interesting use of this information would be for people who are considering places to start a new business.

2. Data acquisition and cleaning

a. Data sources

The data for New York neighborhoods is gathered from “https://cocl.us/new_york_dataset” which is the same used in the third module of this course. The dataset consist of 4 main columns - borough, neighborhood, latitude, longitude.

The data for the Toronto neighborhoods is extracted from the Wikipedia page of Toronto’s postal codes (“https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M”). The geospatial data comes from the “Geospatial_Coordinates.csv” file provided in module three of this course.

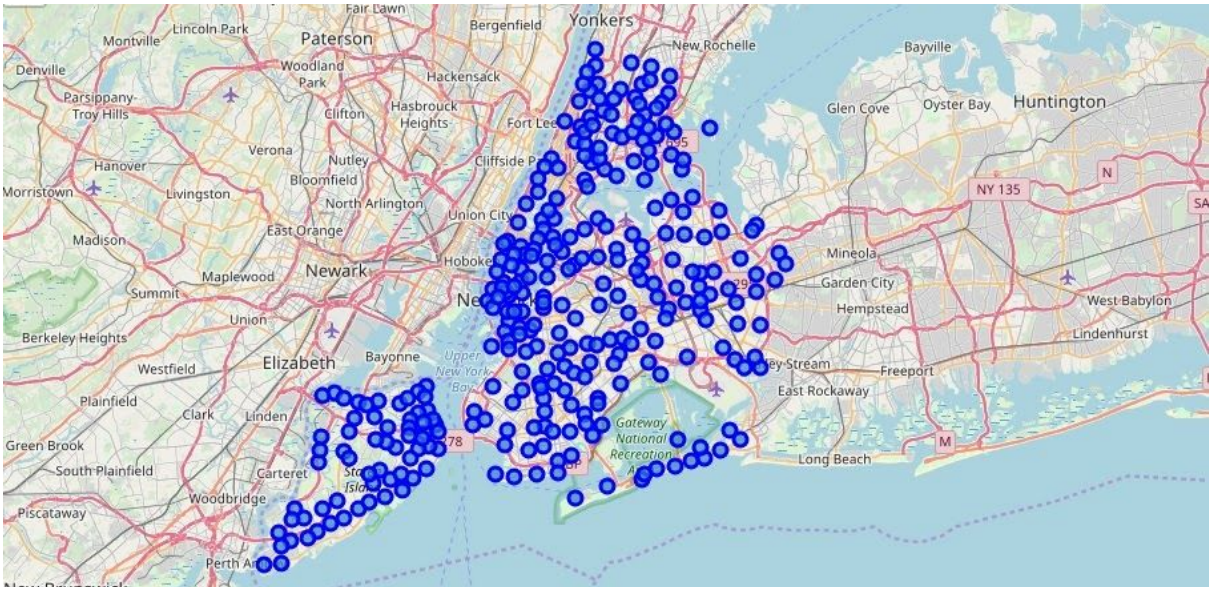
The venues information for both cities comes from using the Foursquare API with the basic account. One important limitation here is that the personal free account requires a credit card which I could not provide so this makes for a very little information available for analysis.

b. Data cleaning

The New York neighborhood dataset did not require any additional work, whereas the original table used as a source for the Toronto one consists of 3 columns - borough, neighborhood and postal code with a lot of incomplete details, which meant the whole dataset had to be cleaned up and normalized. It required adjustments like removing leftover html formatting/chars, dropping the rows where the borough is “Not assigned”, combining the rows where the neighborhoods belong to the same borough.

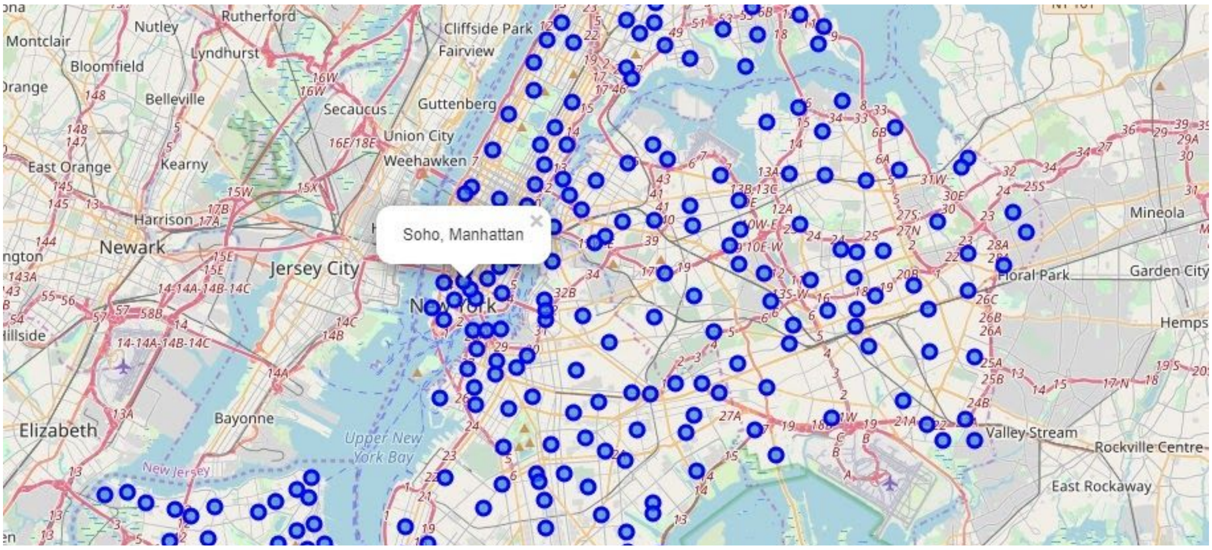
c. Feature selection

In order to select representative sample neighborhoods from both cities the datasets were represented on a map.

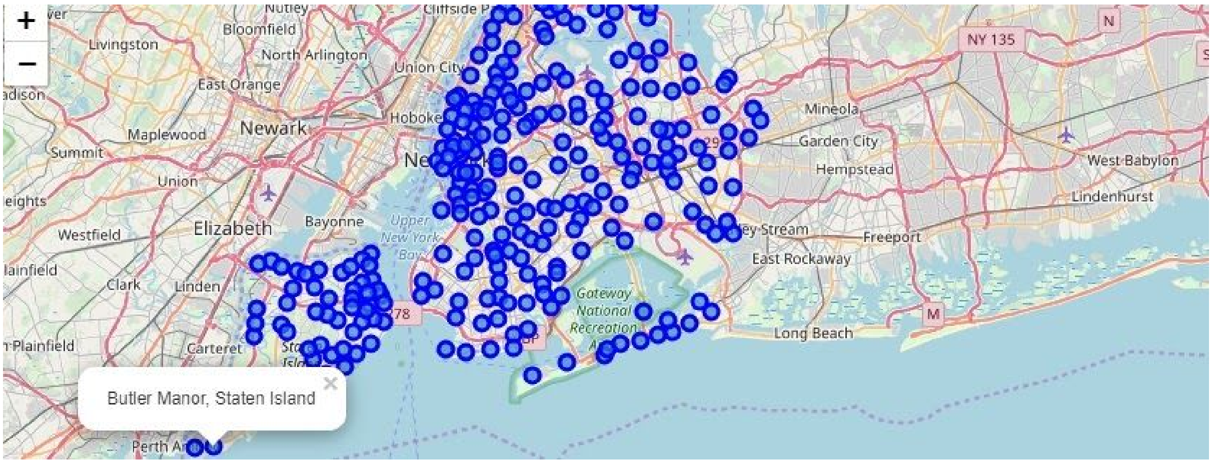


New York - all neighborhoods

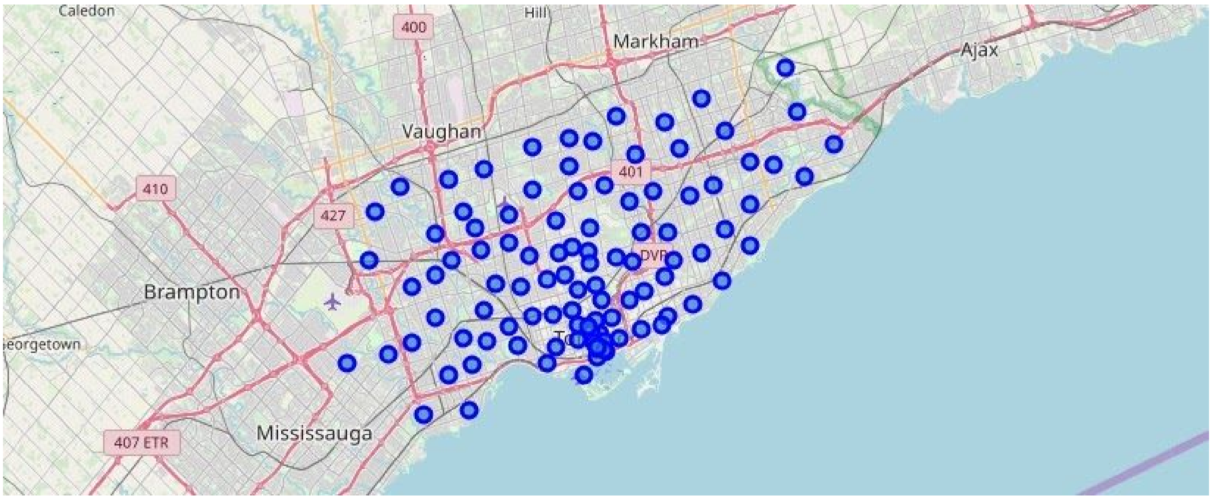
The most obvious choice for a neighborhood in the center of the city was Soho, Manhattan:



A representation of a neighborhood on the outskirts of the city is Butler Manor, Staten Island:



Similar process was used in order to choose the features for Toronto.



3. Exploratory Data Analysis

a. Calculation of target variable

In order to represent the categorical data in any sort of plot it has to be encoded.

The following is a list of the categories in the chosen New York neighborhoods:

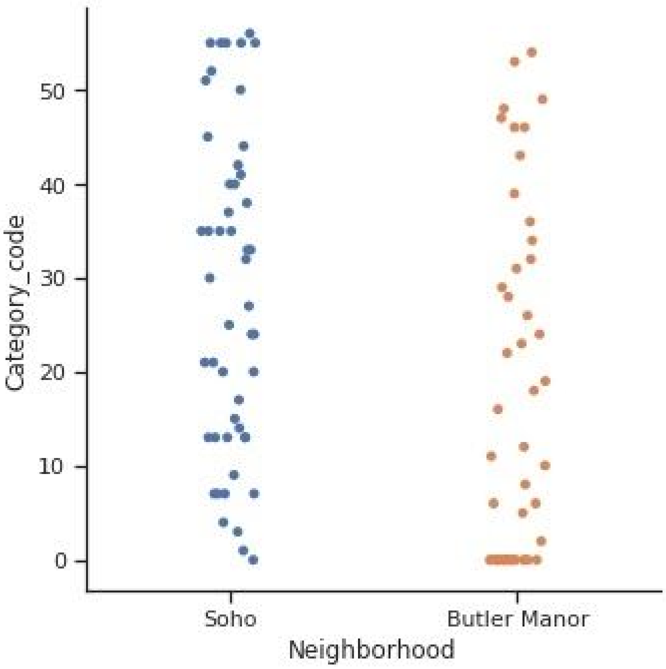
1	Advertising Agency
2	American Restaurant
3	Art Museum
4	Arts & Crafts Store
5	BBQ Joint
6	Baseball Field
7	Boutique
8	Bus Line
9	Café
10	Campground
11	Chinese Restaurant
12	Church
13	Clothing Store
14	Club House
15	Conference Room
16	Convenience Store
17	Cosmetics Shop
18	Dance Studio
19	Dentist's Office
20	Department Store
21	Design Studio
22	Elementary School
23	Event Space
24	Flea Market
25	Fried Chicken Joint
26	Furniture / Home Store
27	Gas Station
28	General Entertainment
29	Gift Shop
30	Grocery Store
31	Italian Restaurant
32	Jewelry Store
33	Lounge
34	Medical Center
35	Men's Store
36	Multiplex
37	Music Store
38	Music Venue
39	Nail Salon
40	Office
41	Optical Shop
42	Other Great Outdoors
43	Park
44	Pet Store
45	Playground
46	Pool
47	Road
48	Rock Club
49	School
50	Shoe Store
51	Supermarket
52	Trail
53	Women's Store
54	Yoga Studio

The following is the list of the encoded categories in the chosen Toronto neighborhoods:

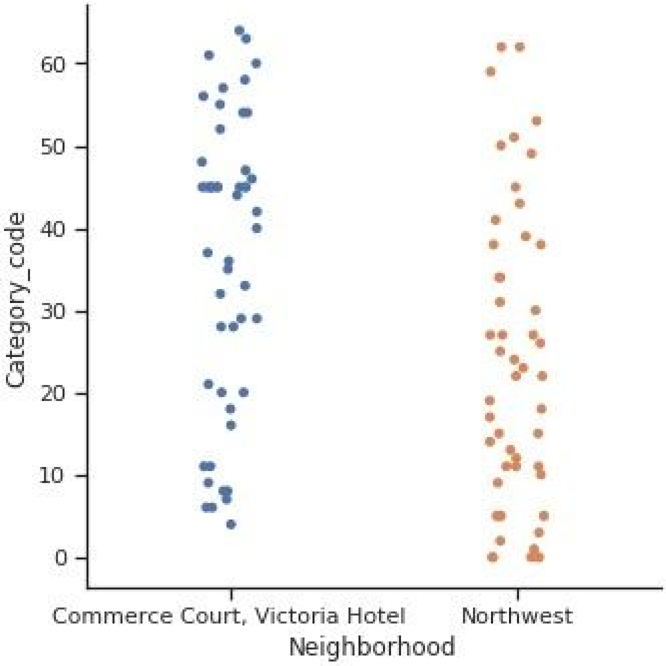
1	African Restaurant
2	Airport
3	Airport Lounge
4	Art Gallery
5	Automotive Shop
6	Bagel Shop
7	Bakery
8	Bank
9	Bar
10	Brewery
11	Building
12	Bus Line
13	Café
14	Car Wash
15	Caribbean Restaurant
16	Chinese Restaurant
17	Church
18	Coffee Shop

19	Corporate Amenity
20	Deli / Bodega
21	Dentist's Office
22	Doctor's Office
23	Dog Run
24	Drugstore
25	Electronics Store
26	Event Space
27	Factory
28	Financial or Legal Service
29	Food Court
30	Furniture / Home Store
31	Gas Station
32	General Travel
33	Gym
34	Hardware Store
35	Health Food Store
36	History Museum
37	Hot Dog Joint
38	Hotel
39	Kingdom Hall
40	Light Rail Station
41	Mediterranean Restaurant
42	Miscellaneous Shop
43	Moving Target
44	Newsstand
45	Office
46	Park
47	Pharmacy
48	Pizza Place
49	Playground
50	Racetrack
51	Rental Car Location
52	Restaurant
53	Road
54	Salon / Barbershop
55	Sandwich Place
56	Seafood Restaurant
57	Spa
58	Sushi Restaurant
59	Swiss Restaurant
60	Toy / Game Store
61	Train
62	Transportation Service
63	Vegetarian / Vegan Restaurant
64	Video Game Store

b. Relationship location - venue categories in the same city

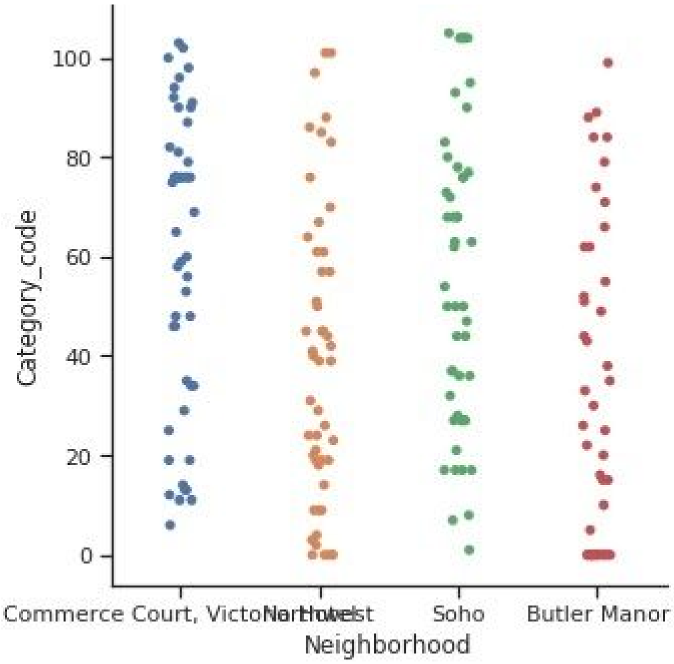


Comparing Soho and Butler Manor in New York, it becomes obvious that neighborhood in the center have a lot more venues in categories like Woman’s and Men’s Store, Yoga Studio, Cafes, Clothing Stores and Parks whereas the one furthest from the center has more venues in categories like Hardware Store, Grocery Store, Medical Center.



Comparing Commerce Court,Victoria Hotel and Northwest in Toronto, it is obvious that the neighborhood in the center has more venues of categories like Office, Park, Pizza Place, while in the neighborhood furthest from the center one would find more venues like Doctor’s, Dentist’s office, Drugstore, Airport and Airport lounges.

c. Relationship between similar neighborhoods in both cities

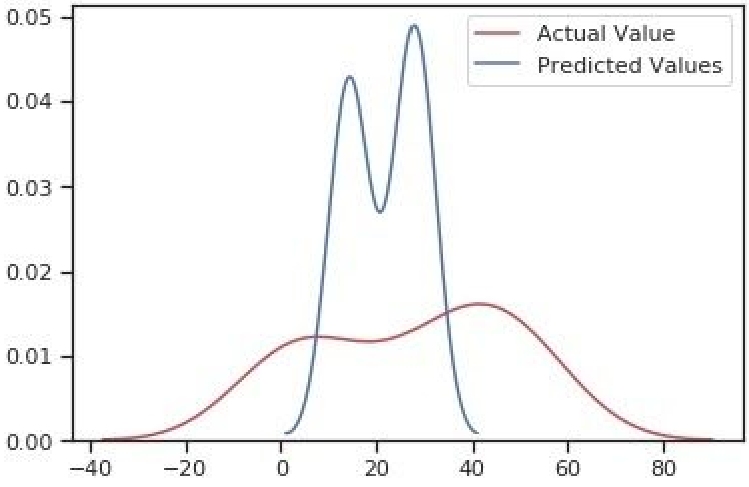


4. Predictive modelling

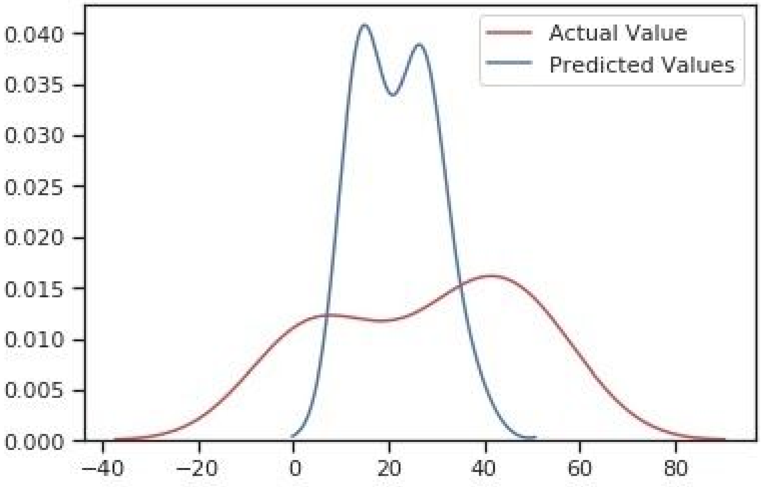
Initially, an assumption was made that one way in which the data can be used is to try and predict the types of venues one would find in a certain distance from the center of the city.

To put this theory to a test the venues dataset for New York was chosen. The "geopy.distance.VincentyDistance" distance calculation was used to determine the distance between each venue and the coordinates of the central neighborhood i.e. Soho.

Using a simple linear regression the predicted values are way off the actual ones:



Using the same test/train split with a polynomial feature for the model resulted in an even more non deterministic results:



5. Conclusion

In this study, I analyzed the relationship between a venue location in the city and what category it would fit in. I also explored the possibility of using distance between the venues and a center location to identify possible dependencies that can be used to predict venue categories and therefore choose the perfect place for a new business. The first turned out to be true, there certainly are common venue categories located in the neighborhoods in the city center as opposed to the ones on the outskirts. The latter unfortunately proven to be a dead end due to the high number of possible categories and the minimal amount of data I could use using the free account without a credit card.