City comparison

Introduction

This project aims at comparing different cities to see and identify similarities. Being an European moving to America, I saw big differences between European cities and American ones.

With this project we proved that actually there are similarities between American and Canadian cities and a big difference with European ones.

In this project we consider the city center, or Downtown, of 3 different cities:

- New York
- Toronto
- Rome

By analyzing the venues in each neighborhood for each Downton, we were able to cluster the neighborhoods and see how these 3 cities are similar or different. A farther analysis could include more cities for a more holistic view (for example we could include Montreal or Quebec city which are known to have a more European vibe and include London or Berlin which on the other end are known to be more international).

This project aimed at being a proof of concept and a starting point for further analysis. I honestly had bigger and more interesting thought, like analyzing the price and the characteristics of the venues, but unfortunately I found that that it would require an account update that I can not afford right now (venue characteristics are premium calls, 500 per day are really not enough to complete the project).

However, we still can appraise the difference between these three cities, as we will see later in the further sections.

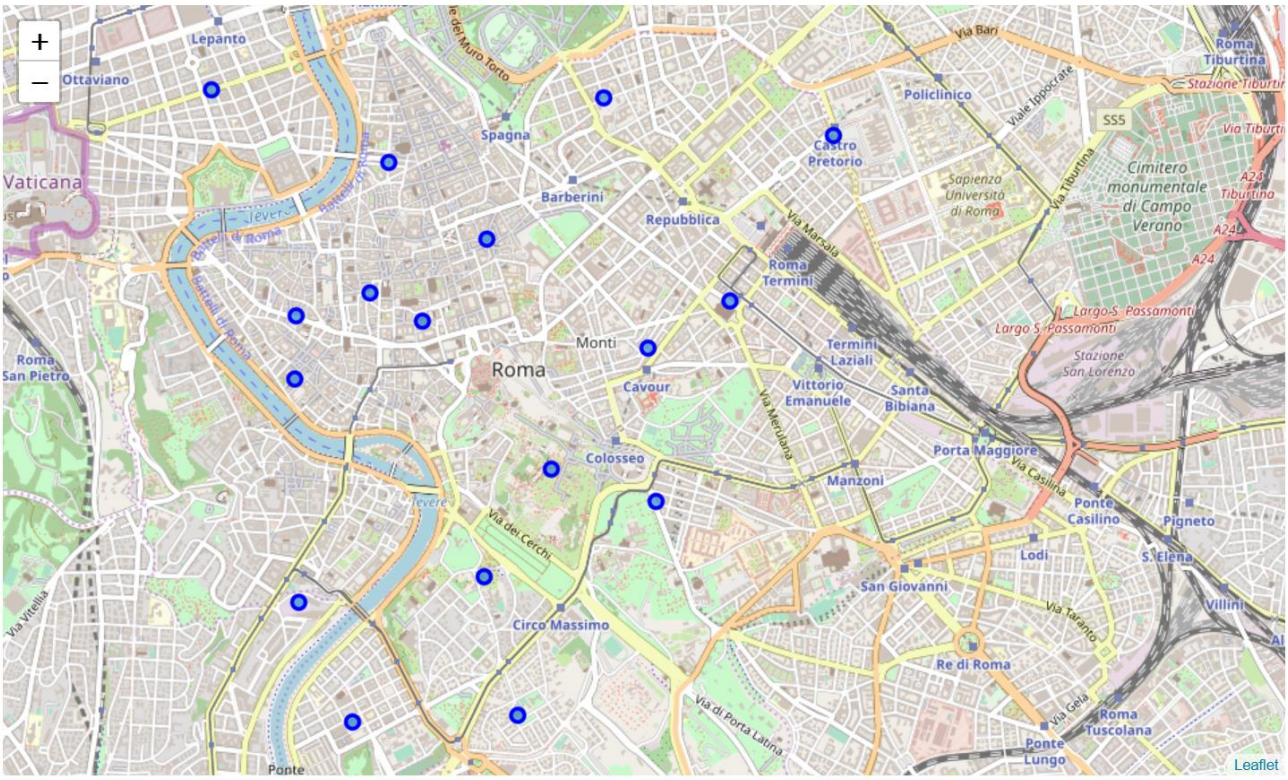
Data section

The data are retrieve mainly on internet. An easy research made me create 3 csv files containing the neighborhoods of each downtown of interest, which were enriched with the geo coordinates.

Here an example for Rome, the other 2 cities follows the same examples.

	Neighbourhood	latitude	longitude
0	MONTI	41.895813	12.493587
1	TREVI	41.900978	12.483285
2	COLONNA	41.833718	12.753184
3	CAMPO MARZIO	41.904647	12.477055
4	PIGNA	41.897116	12.479196
5	PONTE	42.040346	12.853573
6	PARIONE	41.897358	12.471103
7	REGOLA	41.894375	12.471030
8	S. EUSTACHIO	41.898437	12.475792
9	CAMPITELLI	41.890085	12.487416
10	S. ANGELO	42.134941	12.838770
11	R. XII – RIPA	41.884987	12.483107
12	TRASTEVERE	41.883765	12.471270
13	BORGO	41.840235	12.889186
14	CELIO	41.888552	12.494115
15	ESQUILINO	41.898044	12.498863
16	LUDOVISI	41.684213	12.776364
17	SALLUSTIANO	41.907724	12.490797
18	CASTRO PRETORIO	41.905911	12.505453
19	TESTACCIO	41.878065	12.474757
20	S. SABA	41.878352	12.485295
21	PRATI	41.908078	12.465706

Such Neighborhoods have been also mapped thanks to folium:



Using Foursquare API, each neighborhood have been explored and its most popular venues (including their categories) have been save in a dataframe

	Neighbourhood	Neighboyrhood Latitude	Neighboyrhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	MONTI	41.895813	12.493587	Fatamorgana	41.895610	12.493304	Ice Cream Shop
1	MONTI	41.895813	12.493587	Trieste	41.896305	12.494132	Pizza Place
2	MONTI	41.895813	12.493587	Grezzo	41.896681	12.494535	Pastry Shop
3	MONTI	41.895813	12.493587	Black Market	41.897096	12.494645	Cocktail Bar
4	MONTI	41.895813	12.493587	Analemma Cafe	41.894763	12.491536	Bar
5	MONTI	41.895813	12.493587	Montipalace Hotel	41.895384	12.493839	Hotel
6	MONTI	41.895813	12.493587	Libreria Caffè Bohemien	41.895444	12.492863	Cocktail Bar
7	MONTI	41.895813	12.493587	Aromaticus	41.896840	12.494611	Salad Place
8	MONTI	41.895813	12.493587	Relais Monti	41.896606	12.494637	Bed & Breakfast
9	MONTI	41.895813	12.493587	Gelateria dell'Angeletto	41.894815	12.491431	Ice Cream Shop
10	MONTI	41.895813	12.493587	The K Boutique Hotel	41.894849	12.493646	Hotel

Repeating this process for all 3 cities we obtain the dataframe that will be used for the cluster analysis

Methodology

This project involves using a dataframe with all the neighbourhoods names and coordinates of each town.

I tried to consider other variables such as price, outdoor seating etch for each venues, but unfortunately I couldn't go much further since it requires a premium call and I had only 50 of them available per day. So I had to stick with categories venue data as done in previous assignments.

The venues data analysis as been done for all 3 cities of interests, and the results were put together in a whole comprehensive dataframe.

Such dataframe as been analyze using the get_dummies function to find which venue was present in each neighborhood and with which frequency:

	City	Neighborhood	Abruzzo Restaurant	Accessories Store	Adult Boutique	Afghan Restaurant	Airport Lounge	Airport Service	American Restaurant	Animal Shelter	Antique Shop	Arepa Restaurant	Arge Rest
0	New York	Battery Park City	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.032258	0.00	0.00	0.00	0.(
1	New York	Carnegie Hill	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.00	0.(
2	New York	Central Harlem	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.047619	0.00	0.00	0.00	0.(
3	New York	Chelsea	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.00	0.(
4	New York	Chinatown	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.020408	0.00	0.00	0.00	0.(
5	New York	Civic Center	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.00	0.(

We can also find the most common venues for each neighborhood:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Battery Park City	Park	Department Store	Cupcake Shop	Food Court	Sandwich Place	Pharmacy	Coffee Shop	Smoke Shop	Gastropub	Salad Place
1	Carnegie Hill	Café	Gym / Fitness Center	Italian Restaurant	Gym	Spa	Dance Studio	Deli / Bodega	Korean Restaurant	Sports Bar	Pizza Place
2	Central Harlem	Cosmetics Shop	Breakfast Spot	Bagel Shop	Cycle Studio	Café	French Restaurant	Fried Chicken Joint	Music Venue	Lounge	Caribbean Restaurant
3	Chelsea	Mexican Restaurant	Nightclub	Café	Hotel	Speakeasy	Cupcake Shop	Asian Restaurant	Liquor Store	Coffee Shop	Event Space
4	Chinatown	Bubble Tea Shop	Chinese Restaurant	Japanese Restaurant	Hotpot Restaurant	Sandwich Place	Vietnamese Restaurant	Korean Restaurant	Spa	Noodle House	Record Shop

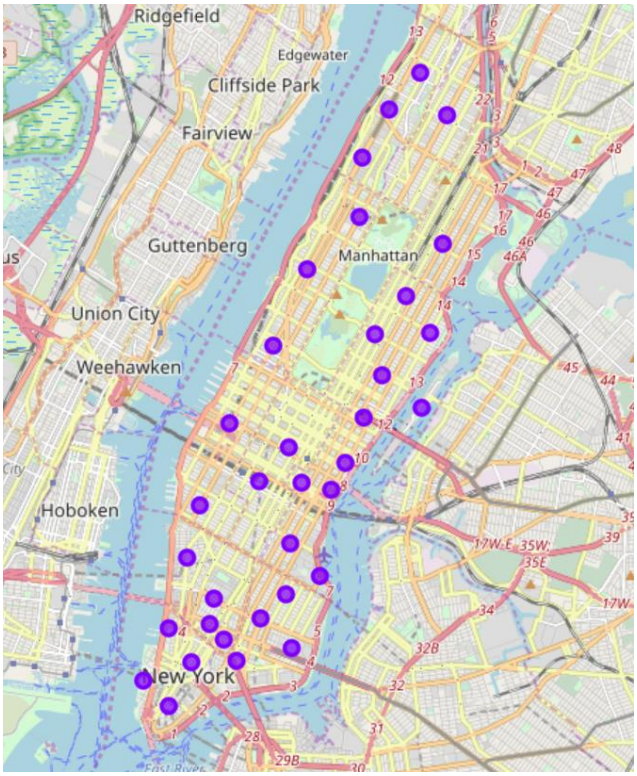
The dataframe containing the frequency for each venues (previous picture) was used to cluster the neighbourhoods using K mean.

K mean was chosen because we are looking at an unsupervised classification problem.

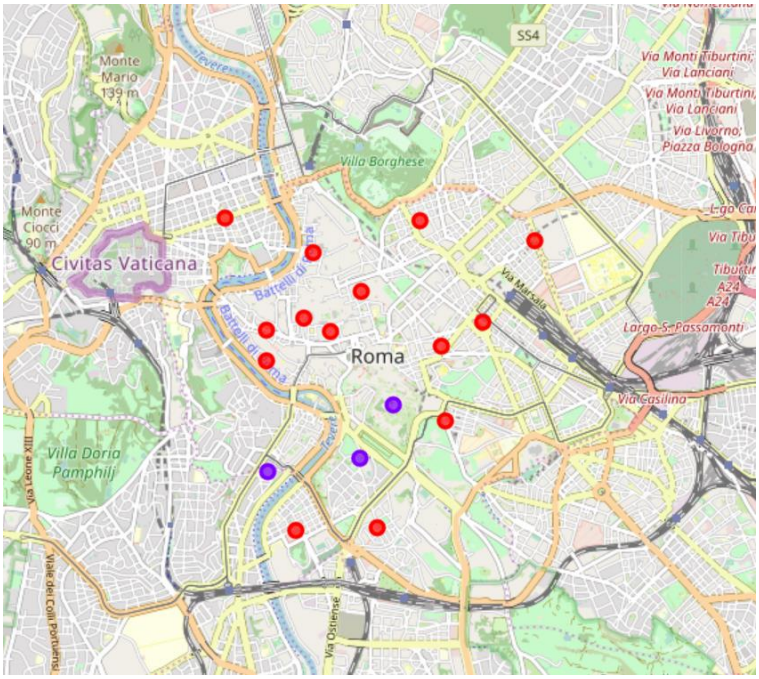
Results

Giving a different color for each cluster, we can see how the different neighborhoods are clustered:

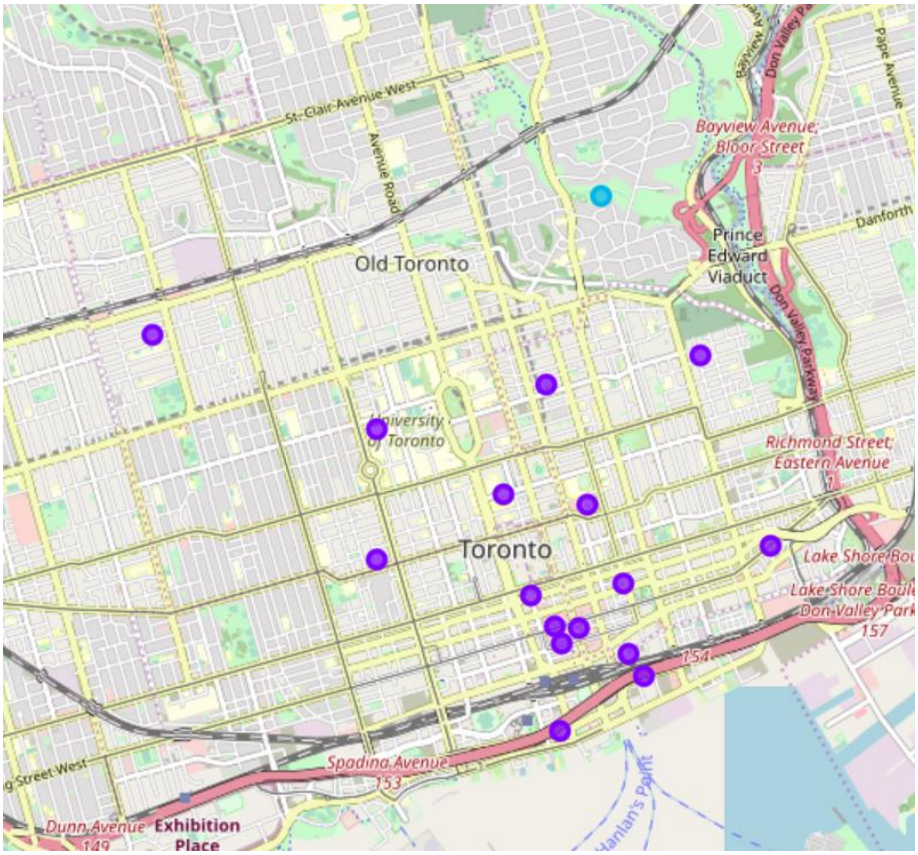
1) New York



2) Rome



3) Toronto



Discussion

We can easly see how Toronto and New York are very alike, while Rome appear to be completely different. Which is something I honestly did expect. With more time and resources, It will be interesting to add other cities, such as Montreal, Quebec city such as London Berlin and Paris, and why not some other part of the world too to discover similarities and differences. K mean helped us clustering the neighborhoods, and for my best knowledge the results are very good, since it is really true that Rome has a totally different vibe when compared to New York or Toronto.

Conclusion

This project provided a proof of concept of how the vibe of a city could be understood by analyzing the most common venues in its neighborhoods. K mean proved to be trusthworthy in clustering the neighbourhoods given the input data. Further and more interesting studies could be carried out including many more different cities, possibly giving more outputs on the nature of each neighbourhood (which have a more international vibe? Which has a more European vibe?). A further development could include a predicting Machine learning tool to obtain the same answers a priori (given a neighbourhood, what is its vibe likely to be?)