Analysis of Similarities among New York, Paris, Toronto

based upon Venues

Sohaib Shibli

10 June 2021

Introduction, Background, Importance

As per Google search "In 2018, there were 467 cities with between 1 and 5 million inhabitants and an additional 598 cities with between 500,000 and 1 million inhabit- ants. By 2030, the number of cities with 1 to 5 million inhabitants is projected to grow to 597". Researchers has categorized those cities in terms of population, immunities, sizes, social and economical activities etc.

The categorize of these cities changes with time facilitating study of their past and future. On going globalization and visibility of information through social media and internet resetting and redefining the people thought process about everything they do or they intend to do. Knowledge of similar cities or dissimilar cities is interesting in several ways and beneficial to people, organizations, tourists, businesses etc. around the world.

Such information can assist people to explore in understanding of political, social, and economic conditions and respective deviation if it exist. International organization such as IMF, world bank, news agencies etc. could formulate their policy based upon such kind of segmented cities. The Multinational Companies could also evaluate and decide their strategical target markets for their products. The tourists could plan their visits based upon such clustering and segmentation that matches their interests. The countries could also set benchmark cities and develop their own cities according to such benchmarks.

Description of Data and its usage in resolution

We will explore and evaluate similarities and dissimilarities among Toronto, New York and Paris. There are various approaches that could be followed, one of them is studying the venues around the cities within the radius of 1000 meters. Instead of studying individual venues, the count of venue category is found more useful for city to city comparison.

Since our analysis of similarities among cities are based upon the venues around the city thus we decided to use only the foursquare database instead of several other available sources. The major benefit of doing so is that same definition of venues, venues categories, location, etc. definition irrespective of the indiviual cities are used. The foursquare data is most comprehensive database for geographical locations with plenty of attributes that could be used for any kind of location based study.

The Foursqare offers various endpoints, endpoints groups that can be used in venues, users related searching, exploring, trending, and etc.

Data Sources:

https://api.foursquare.com

We may use various endpoints groups but we will be using foursquare endpoint for venue exploration only to study the venues categories to evaluate similarities in Toronto, New York and Paris. After combining venues we calculate the frequencies of the venues categories for each city within the radius around 1000 meter. The reference location is city center of each city and venues are captured accordingly. A frequency distribution of the count of the venues are used to access the similarities. The closer the frequency venue categories the similar the cities are to each other.

Methodology and exploratory data analysis

The venues data extracted is Jason format from Foursquare database using URLs of New York, Paris, and Toronto.

The Jason format data is then converted into pandas data frames for each city. Since it is decided to to analysis the categories of each city to find the pattern of similarities between them, so only the

categories and cities columns are selected in the data frames

These separate data frames of each country is combined into one data frame for exploratory data analysis purposes.

The categories are defined sometimes using different terms but similar to each other are collected and a new column main category is created to hold this data.

Category	Main Category
Strip Club	Adult Entertainment
Bar	Bar
Pub	Bar
Theater	Entertainment
Trail	Entertainment
Scenic Lookout	Entertainment
Gallery	Entertainment
Museum	Entertainment
Concert Hall	Entertainment
Cultural Center	Entertainment
Dance Studio	Entertainment

Category	Main Category
Fountain	Park
Garden	Park
Park	Park
Pedestrian Plaza	Plaza
Plaza	Plaza
Restaurant	Restaurant
Taco Place	Restaurant
Steakhouse	Restaurant
Breakfast Spot	Restaurant
Burger Joint	Restaurant
Café	Restaurant

Historic Site	Entertainment
Music Venue	Entertainment
Opera House	Entertainment
Auditorium	Entertainment
#VALUE!	Entertainment
Gym	Gym
Pilates Studio	Gym
Playground	Gym
Spa	Gym
Yoga Studio	Gym
Hotel	Hotel
Memorial Site	Memorial Site
Laundry Service	Others
Neighborhood	Others
Speakeasy	Others
University	Others

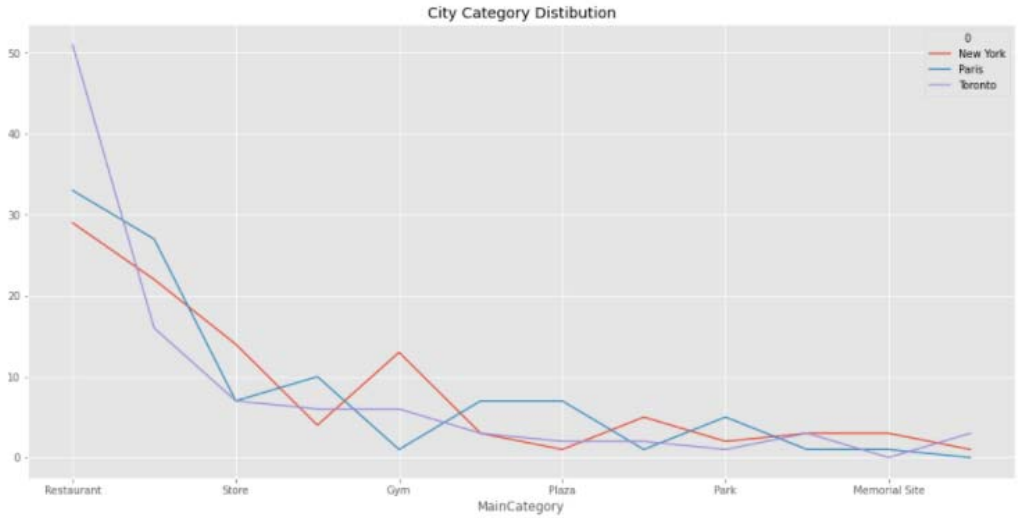
Creperie	Restaurant
Deli / Bodega	Restaurant
Creperie	Restaurant
Gastropub	Restaurant
Bistro	Restaurant
Poke Place	Restaurant
Pizza Place	Restaurant
Burrito Place	Restaurant
Sandwich Place	Restaurant
Shop	Shop
Bakery	Shop
Bookstore	Store
Farmers Market	Store
Store	Store

A summarized rankning table is built to see behaviour of categories. Since every time four square returns live data that may leads creation of new categories, these are categorized as and captured under “newly added”.

MainCategory	New York	Paris	Toronto	Total	cum_sum	cum_perc%
Restaurant	29	33	51	113	113	37.666667
Shop	22	27	16	65	178	59.333333
Store	14	7	7	28	206	68.666667
Entertainment	4	10	6	20	226	75.333333
Gym	13	1	6	20	246	82.000000
Bar	3	7	3	13	259	86.333333
Plaza	1	7	2	10	269	89.666667
Newly Added	5	1	2	8	277	92.333333
Park	2	5	1	8	285	95.000000
Hotel	3	1	3	7	292	97.333333
Memorial Site	3	1	0	4	296	98.666667
Others	1	0	3	4	300	100.000000

Results

After plotting these categories for each city shows that there are varying degree of similarities among these cities.



A correlation table is created among these cities as shown below:

	New York	Paris	Toronto	Total
New York	1.000000	0.847204	0.865275	0.939487
Paris	0.847204	1.000000	0.869501	0.948429
Toronto	0.865275	0.869501	1.000000	0.967143
Total	0.939487	0.948429	0.967143	1.000000

The correlation results shows that New York is much similar to Toronto, Paris is much similar to Toronto based upon the limited venues we collected from Four Square database.

Observations and recommendations

Since the Foursquare database returns only 100 venues for each call, which is not sufficient size for category analysis. Other challenge is defining the correct city center Latitude and Longitude, which is important for accurate collection of venues within the cities. The definition of category is also and its

assignment to main selected categories have risk if incorrect assignment. There must be some data that could simplify and standardize the venue definition.

Conclusion

Based upon the limited data extracted from Foursquare database, it is reflected from calculated correlation that New York is more 86.5% similar to Toronto based upon the venues categories data for both the cities. Interestingly, Paris shows higher correlation with Toronto.