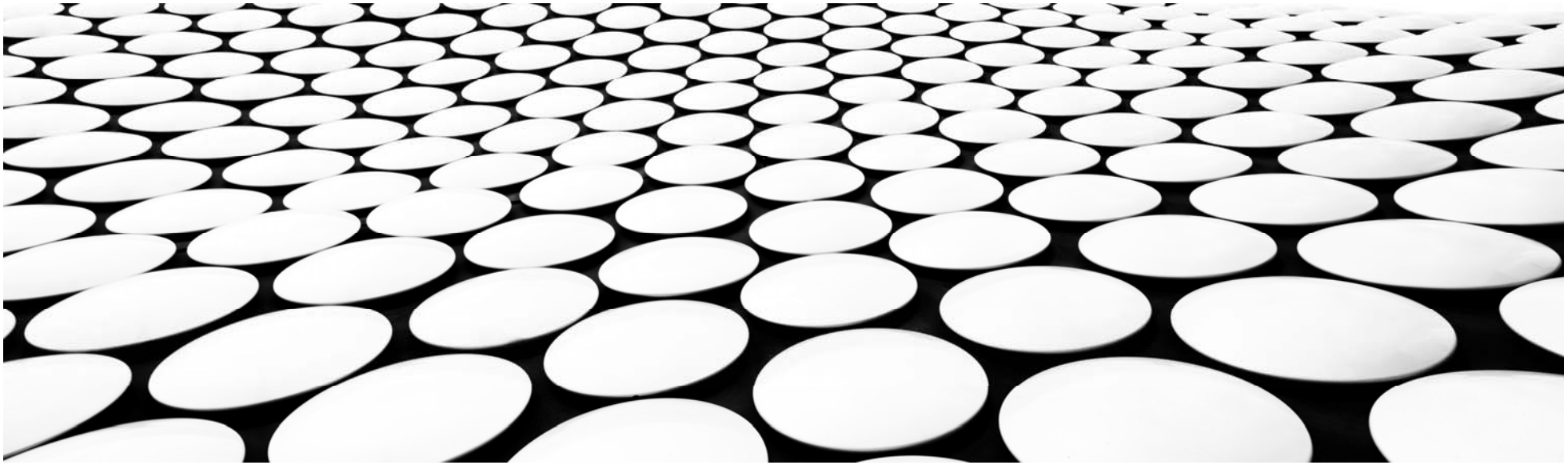
SIMILARITIES AMONG NEW YORK, PARIS, TORONTO

SOHAIB SHIBLI 11 JUN 2021



INTRODUCTION, BACKGROUND, IMPORTANCE

- 467 cities with between 1 and 5 million inhabitants
- 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
- 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.

DESCRIPTION OF DATA AND ITS USAGE IN RESOLUTION

- studying the venues around the cities
- Foursquare database to capture venues. The major benefit of doing so is that same definition of venues, venues categories, location, etc. definition irrespective of the individual 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

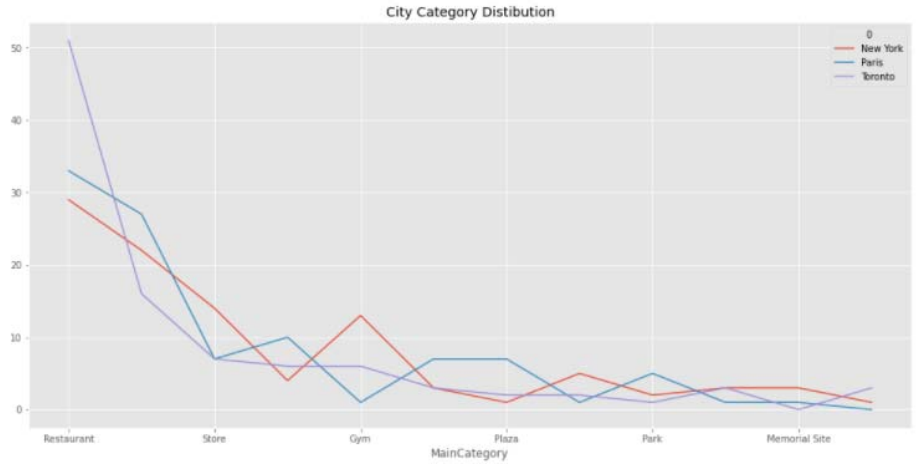
METHODOLOGY AND EXPLORATORY DATA ANALYSIS

- 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.

METHODOLOGY AND EXPLORATORY DATA ANALYSIS – DATA TRANSFORMATION

Category	Main Category		Category	Main Category
Strip Club	Adult Entertainment		Fountain	Park
Bar	Bar		Garden	Park
Pub	Bar		Park	Park
Theater	Entertainment		Pedestrian Plaza	Plaza
Trail	Entertainment		Plaza	Plaza
Scenic Lookout	Entertainment		Restaurant	Restaurant
Gallery	Entertainment		Taco Place	Restaurant
Museum	Entertainment		Steakhouse	Restaurant
Concert Hall	Entertainment		Breakfast Spot	Restaurant
Cultural Center	Entertainment		Burger Joint	Restaurant
Dance Studio	Entertainment		Café	Restaurant
Historic Site	Entertainment		Creperie	Restaurant
Music Venue	Entertainment		Deli / Bodega	Restaurant
Opera House	Entertainment		Creperie	Restaurant
Auditorium	Entertainment		Gastropub	Restaurant
#VALUE!	Entertainment		Bistro	Restaurant
Gym	Gym		Poke Place	Restaurant
Pilates Studio	Gym		Pizza Place	Restaurant
Playground	Gym		Burrito Place	Restaurant
Spa	Gym		Sandwich Place	Restaurant
Yoga Studio	Gym		Shop	Shop
Hotel	Hotel		Bakery	Shop
Memorial Site	Memorial Site		Bookstore	Store
Laundry Service	Others		Farmers Market	Store
Neighborhood	Others		Store	Store
Speakeasy	Others			
University	Others			

METHODOLOGY AND EXPLORATORY DATA ANALYSIS



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 AND CONCLUSION

	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

- . 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.
- Limitation :
  - Foursquare database returns only 100 venues for each call,
  - which is not sufficient size for category analysis.
  - Defining the correct city center Latitude and Longitude,
  - 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.