Junior Fernandes
16 October 2020

Similarity between cities around the world

1. Introduction

Business problem and Interest

The project can be applied to the following problem.

A person trying to know better some well-known cities around the world (like London, Paris, Tokyo, Hong-Kon, Rio etc) and compare them to each other regarding their 20 most common places. By comparing them, I mean clustering them and find out which cities seem similar based on their most frequent places nearby.

With this information, if the person wants to leave his current city (let's say Paris for example), he can choose to live in another city similar to Paris to feel more secure or rather live in a completely different city than Paris to discover new places and culture. This change of city can be for professional reason, for some vacations or just by curiosity for someone who wants to know if he can find the same category of places (gyms, coffee shops, bookshop etc) in another attractive city for future purposes.

Data used

I will only use Foursquare's location data to find the most common venues nearby those different cities. Those venues will be used to find out which are the main categories of venues nearby these cities. Knowing which categories are ubiquitous, I will be able to have a good understanding of which type of venues we have nearby those cities.

2. Methodology

By leveraging, the informations about all the categories of venues nearby those cities, I will find out their frequency of apparition and sort them in descending order (i.e from the most frequent category to the less frequent ones). Below, you can see an example for each city (used in the analysis) with their top 5 categories of venues :

```
----Bangalore, India----
              venue  freq
0  Capitol Building  0.25
1             Hotel  0.25
2          Vineyard  0.25
3              Park  0.25
4            Palace  0.00
```

```
----Le Cap, South Africa----
                venue  freq
0                Café  0.08
1  Italian Restaurant  0.08
2                 Pub  0.06
3              Museum  0.06
4          Restaurant  0.04
```

```
----Manhattan, NY----
                venue  freq
0      Baseball Field  0.23
1                Park  0.23
2          Playground  0.17
3  Athletics & Sports  0.07
4       Metro Station  0.03
```

```
----Berlin, Germany----
            venue  freq
0           Hotel  0.12
1        Wine Bar  0.06
2      Restaurant  0.04
3  Clothing Store  0.04
4           Plaza  0.04
```

```
----Lisbon, Portugal----
                   venue  freq
0  Portuguese Restaurant  0.18
1                    Bar  0.10
2                  Plaza  0.08
3                 Hostel  0.06
4          Ice Cream Shop  0.04
```

```
----Mexico City, Mexico----
                venue  freq
0               Hotel  0.08
1    Department Store  0.06
2  Mexican Restaurant  0.06
3              Museum  0.06
4      Ice Cream Shop  0.04
```

```
----Dublin, Ireland----
            venue  freq
0     Coffee Shop  0.12
1             Pub  0.12
2  Clothing Store  0.08
3       Bookstore  0.06
4            Café  0.06
```

```
----London, England----
         venue  freq
0        Hotel  0.12
1      Theater  0.08
2  Art Gallery  0.06
3       Garden  0.06
4        Plaza  0.06
```

```
----Moscow, Russia----
            venue  freq
0           Plaza  0.17
1  History Museum  0.11
2        Boutique  0.11
3   Historic Site  0.11
4    Concert Hall  0.06
```

```
----Hong-Kong, China----
        venue  freq
0       Hotel  0.12
1        Café  0.10
2        Park  0.06
3  Steakhouse  0.06
4   Bookstore  0.04
```

```
----Madrid, Spain----
                venue  freq
0               Plaza  0.10
1    Tapas Restaurant  0.10
2               Hotel  0.06
3   Electronics Store  0.04
4      Cosmetics Shop  0.04
```

```
----New Delhi, India----
                     venue  freq
0              Music Venue  0.33
1        Indian Restaurant  0.33
2       Light Rail Station  0.33
3                   Palace  0.00
4  New American Restaurant  0.00
```

```
----Nice, France----
                venue  freq
0   French Restaurant  0.18
1               Hotel  0.12
2      Ice Cream Shop  0.06
3    Department Store  0.04
4           Bookstore  0.04
```

```
----Praia, Cape-Verde----
          venue  freq
0        Bakery  0.25
1         Plaza  0.25
2   Beer Garden  0.25
3          Café  0.25
4  Neighborhood  0.00
```

```
----Rome, Italy----
                venue  freq
0       Historic Site  0.26
1              Temple  0.08
2  Italian Restaurant  0.08
3               Hotel  0.08
4  Monument / Landmark  0.08
```

```
----Paris, France----
                venue  freq
0      Ice Cream Shop  0.08
1   French Restaurant  0.08
2               Plaza  0.06
3                Park  0.04
4         Art Gallery  0.04
```

```
----Rio, Brazil----
                  venue  freq
0                   Bar  0.16
1            Restaurant  0.12
2                  Café  0.08
3  Brazilian Restaurant  0.08
4         Train Station  0.08
```

```
----Sydney, Australia----
                   venue  freq
0                Theater  0.23
1           Concert Hall  0.15
2  Australian Restaurant  0.15
3            Cocktail Bar  0.15
4                   Park  0.08
```

```
----Tokyo, Japan----
             venue  freq
0    Historic Site  0.14
1             Café  0.08
2             Park  0.04
3  Thai Restaurant  0.04
4           Lounge  0.04
```

```
----Toronto, Canada----
                     venue  freq
0           Clothing Store  0.06
1                  Theater  0.04
2        Electronics Store  0.04
3                    Plaza  0.04
4  New American Restaurant  0.04
```

In the previous example, you saw the top 5 categories of venues but I will select the 20 most frequent categories to cluster my cities and find out which city is similar to another. I will build a data frame composed of the city and the 20 most common categories venues like below:

	City	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	...
0	Bangalore, India	Hotel	Vineyard	Park	Capitol Building	Zoo	Diner	Falafel Restaurant	Exhibit	Event Space	...
1	Berlin, Germany	Hotel	Wine Bar	Plaza	Coffee Shop	Clothing Store	Opera House	Cosmetics Shop	Concert Hall	Restaurant	...
2	Dublin, Ireland	Coffee Shop	Pub	Clothing Store	Café	Bookstore	Discount Store	Hotel	Theater	Donut Shop	...
3	Hong-Kong, China	Hotel	Café	Steakhouse	Park	Furniture / Home Store	Italian Restaurant	Lounge	Dim Sum Restaurant	Cantonese Restaurant	...
4	Le Cap, South Africa	Café	Italian Restaurant	Museum	Pub	Restaurant	Hotel	Pizza Place	Cuban Restaurant	Burger Joint	...

5 rows × 21 columns

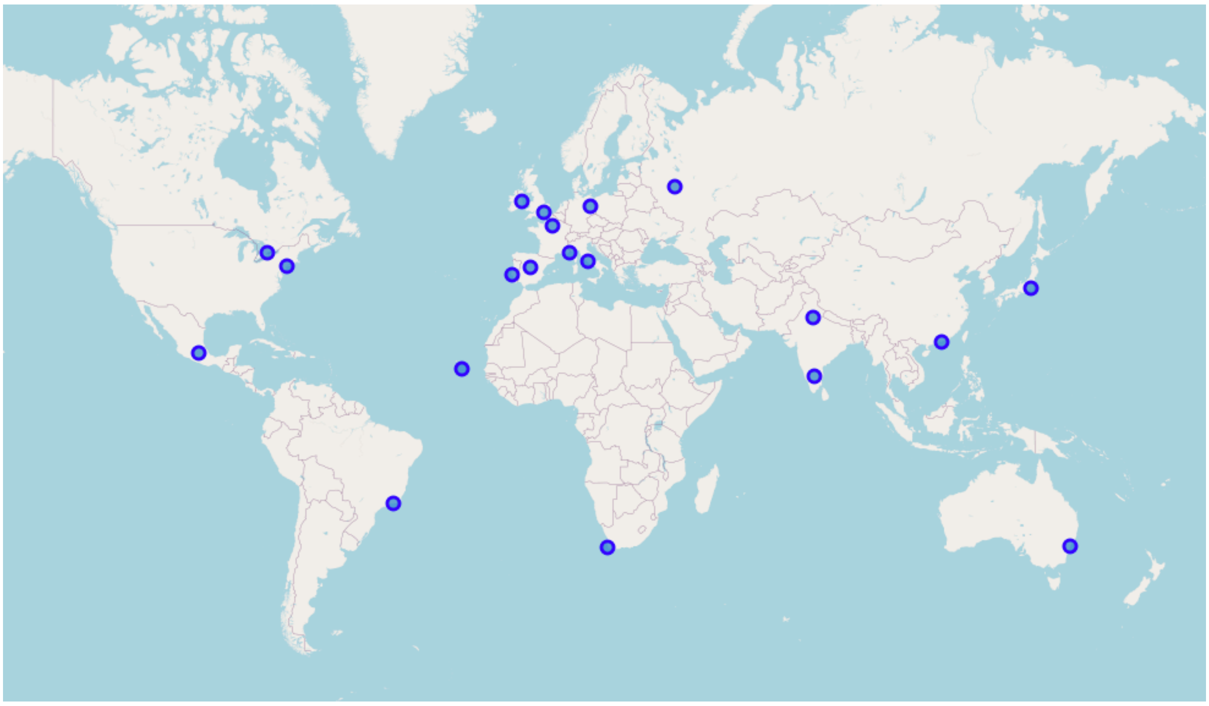
3. Results

For the results, I build a new data frame containing now, 3 new columns : the *latitude* and *longitude* for each city and the *cluster label*.

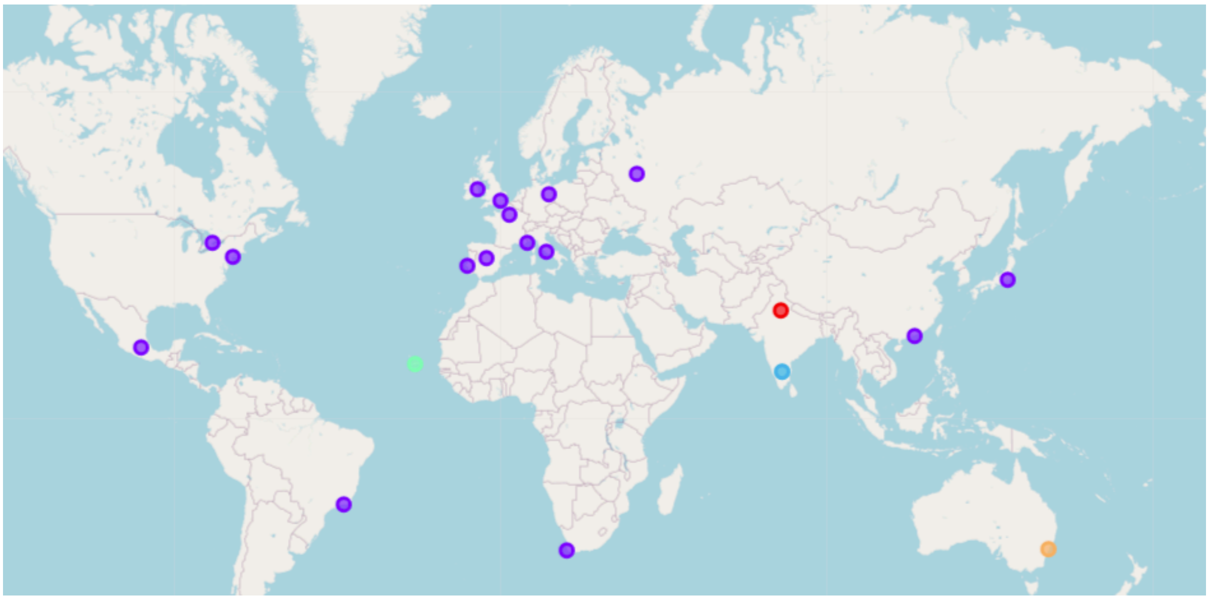
	City	Latitudes	Longitudes	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	...
0	Toronto, Canada	43.653482	-79.383935	1	Clothing Store	Theater	Electronics Store	Seafood Restaurant	Plaza	New American Restaurant	...
1	Paris, France	48.856697	2.351462	1	Ice Cream Shop	French Restaurant	Plaza	Gay Bar	Art Gallery	Clothing Store	...
2	Manhattan, NY	40.789624	-73.959894	1	Baseball Field	Park	Playground	Athletics & Sports	Bus Station	Food Truck	...
3	London, England	51.507322	-0.127647	1	Hotel	Theater	Art Gallery	Garden	Plaza	Art Museum	...
4	Lisbon, Portugal	38.707751	-9.136592	1	Portuguese Restaurant	Bar	Plaza	Hostel	Ice Cream Shop	Hotel	...

5 rows × 24 columns

Before clustering, we can use a map to visualise each city with its latitude and longitude :



After clustering, we can print a new map with a color for each cluster



We can see that :

- Cluster 0 (in red) : contains New Delhi only.
- Cluster 1 (in purple) contains : Manhattan, Toronto, Mexico City, Rio de Janeiro, London, Dublin, Paris, Nice, Lisbon, Madrid, Nice, Berlin, Rome, Moscow, Le Cap, Hong-Kong, Tokyo.
- Cluster 2 (in blue) : contains Bangalore only
- Cluster 3 (in green) : is composed of Praia only
- Cluster 4 (in orange) : contains Sydney only

Conclusion

Finally, we can conclude that cities in cluster 1 seemed to have common recurrent category of venues which make them similar and thus on the same cluster. We also have all the other clusters which have only one city, this may be because those cities have very uncommon places and different cultures which make them isolated on another cluster.

Now a person can have a better idea of which city share the same type of places and take a decision for its future if necessary.

Improvements

A more precise way to do that could be to have data about the different neighbourhoods in each city and find out what are the most frequent category of places in those neighbourhoods and then use this data to finally have the most common types of venues for the entire city.