# Grouping the world's most populous Metropolitan cities

By G.Mukkesh

## Introduction

(Background)With the advent of technology, the world has now become a global village. People living in different metropolitan cities of the world are being exposed to the cultures of various nations in the forms of movies, books, tv shows and restaurants, and are able to experience them without actually visiting the country. The converse also holds good, i.e., a person who has immigrated to another country can still experience his own country's culture in the major metropolitan cities.

(Why it is important)The purpose of this study is to use the FourSquare location data to group different metropolitan cities into clusters, depending on how similar/dissimilar they are to each other. By doing so, we will be able to find out the cities which offer a similar lifestyle in terms of the amenities present.

(Target Audience)This would be useful to people who are moving to different countries, since they would know how similar/different living in the new city would be, compared to their previous city.

## Data Set

- The 100 most populous metropolitan cities of the world were scraped from Wikipedia (https://en.wikipedia.org/wiki/List_of_metropolitan_areas_by_population).

https://en.wikipedia.org/wiki/List_of_metropolitan_areas_by_population

Cite this page

In other projects

Wikimedia Commons

Print/export

Create a book

Download as PDF

Printable version

Languages

العربية
Asturianu
বাংলা
Български
Dansk
Deutsch
Español
Esperanto
فارسی
Français
한국어
Bahasa Indonesia
Íslenska
עברית
ქართული

| Rank | Metropolitan | Country | Continent | Official population | Year |
|---|---|---|---|---|---|
| 1 | Tokyo | Japan | Asia | 37,832,892[3] | 2016 |
| 2 | Delhi | India | Asia | 35,454,000[4] | 2018 |
| 3 | Shanghai | China | Asia | 34,865,252[5] | 2015 |
| 4 | Jakarta | Indonesia | Asia | 31,689,592[6] | 2015 |
| 5 | Seoul | South Korea | Asia | 25,514,000[7] | 2016 |
| 6 | Guangzhou | China | Asia | 25,000,000[5] | 2015 |
| 7 | Beijing | China | Asia | 24,900,000[5] | 2015 |
| 8 | Manila | Philippines | Asia | 24,650,000[8] | 2018 |
| 9 | New York City | United States | North America | 23,876,155[9] | 2017 |
| 10 | Shenzhen | China | Asia | 23,300,000[5] | 2015 |
| 11 | Mexico City | Mexico | North America | 21,650,668[10] | 2017 |
| 12 | São Paulo | Brazil | South America | 21,242,939[11] | 2016 |
| 13 | Lagos | Nigeria | Africa | 21,000,000[12] | 2014 |
| 14 | Mumbai | India | Asia | 20,748,395[4] | 2011 |
| 15 | Cairo | Egypt | Africa | 20,500,000[13] | 2012 |
| 16 | Keihanshin (Kyoto-Osaka-Kobe) | Japan | Asia | 19,342,000[3] | 2010 |
| 17 | Wuhan | China | Asia | 19,000,000[5] | 2015 |

[6]:
```
df.head()
```

[6]:

| | Rank | Metropolitan | Country |
|---|---|---|---|
| 0 | 1 | Tokyo | Japan |
| 1 | 2 | Delhi | India |
| 2 | 3 | Shanghai | China |
| 3 | 4 | Jakarta | Indonesia |
| 4 | 5 | Seoul | South Korea |

- Using the geocoder module in python, we get the latitudes and longitudes of the 100 cities that were scraped (100,5).

| | Rank | Metropolitan | Country | lat | long |
|---|---|---|---|---|---|
| 0 | 1 | Tokyo | Japan | 35.676192 | 139.650311 |
| 1 | 2 | Delhi | India | 28.704059 | 77.102490 |
| 2 | 3 | Shanghai | China | 31.230416 | 121.473701 |
| 3 | 4 | Jakarta | Indonesia | -6.208763 | 106.845599 |
| 4 | 5 | Seoul | South Korea | 37.566535 | 126.977969 |

```
df.shape
```

(100, 5)

- Feeding these coordinates into the FourSquare API, we can get the top 10 venues for the cities, along with the venue's longitudes, latitudes and venue category (6868,7).

```
[22]: print(city_venues.shape)
      city_venues.head()
```

(6868, 7)

[22]:

| | City | City Latitude | City Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Tokyo | 35.676192 | 139.650311 | La Piccola Tavola | 35.676900 | 139.643468 | Pizza Place |
| 1 | Tokyo | 35.676192 | 139.650311 | Massimottavio | 35.676812 | 139.642807 | Italian Restaurant |
| 2 | Tokyo | 35.676192 | 139.650311 | CHUBBY | 35.671648 | 139.657577 | Café |
| 3 | Tokyo | 35.676192 | 139.650311 | もみじ屋 | 35.671676 | 139.651525 | Ramen Restaurant |
| 4 | Tokyo | 35.676192 | 139.650311 | Bonito Soup Noodle RAIK | 35.682348 | 139.645606 | Ramen Restaurant |

- After getting the top 10 locations for every city, we can get the set of unique venue categories and then one-hot-vector encode them, after grouping all the venues by their city to get the frequency of occurrence (100,462).

```
[47]: city_grouped = city_onehot.groupby('City').mean().reset_index()
      city_grouped
```

[47]:

| | City | Accessories Store | Afghan Restaurant | African Restaurant | Airport Lounge | Airport Terminal | American Restaurant | Amphitheater | Antique Shop | Aquarium | ... | Wine Bar | Wine Shop | Winery | Wings Joint | Women's Store | Y Rest |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Ahmedabad | 0.00 | 0.00 | 0.00 | 0.000000 | 0.0 | 0.000000 | 0.00 | 0.00 | 0.00 | ... | 0.00 | 0.00 | 0.0 | 0.00 | 0.00 | |
| 1 | Alexandria | 0.00 | 0.00 | 0.00 | 0.000000 | 0.0 | 0.014085 | 0.00 | 0.00 | 0.00 | ... | 0.00 | 0.00 | 0.0 | 0.00 | 0.00 | |
| 2 | Ankara | 0.00 | 0.00 | 0.00 | 0.000000 | 0.0 | 0.000000 | 0.00 | 0.04 | 0.00 | ... | 0.00 | 0.00 | 0.0 | 0.00 | 0.00 | |
| 3 | Atlanta | 0.00 | 0.00 | 0.00 | 0.000000 | 0.0 | 0.030000 | 0.00 | 0.00 | 0.00 | ... | 0.00 | 0.01 | 0.0 | 0.00 | 0.00 | |
| 4 | Bandung | 0.00 | 0.00 | 0.00 | 0.000000 | 0.0 | 0.000000 | 0.00 | 0.00 | 0.00 | ... | 0.00 | 0.00 | 0.0 | 0.01 | 0.00 | |
| 5 | Bangalore | 0.00 | 0.01 | 0.00 | 0.000000 | 0.0 | 0.010000 | 0.00 | 0.00 | 0.00 | ... | 0.01 | 0.00 | 0.0 | 0.00 | 0.00 | |
| 6 | Bangkok | 0.00 | 0.00 | 0.00 | 0.000000 | 0.0 | 0.000000 | 0.00 | 0.00 | 0.00 | ... | 0.00 | 0.00 | 0.0 | 0.00 | 0.00 | |
| 7 | Barcelona | 0.00 | 0.00 | 0.00 | 0.000000 | 0.0 | 0.000000 | 0.00 | 0.00 | 0.00 | ... | 0.04 | 0.01 | 0.0 | 0.00 | 0.02 | |
| 8 | Beijing | 0.00 | 0.00 | 0.00 | 0.000000 | 0.0 | 0.010000 | 0.00 | 0.00 | 0.00 | ... | 0.00 | 0.00 | 0.0 | 0.00 | 0.00 | |
| 9 | Belo Horizonte | 0.01 | 0.00 | 0.00 | 0.000000 | 0.0 | 0.000000 | 0.00 | 0.00 | 0.00 | ... | 0.00 | 0.00 | 0.0 | 0.00 | 0.00 | |
| 10 | Berlin/Brandenburg | 0.00 | 0.00 | 0.00 | 0.000000 | 0.0 | 0.000000 | 0.00 | 0.00 | 0.00 | ... | 0.01 | 0.01 | 0.0 | 0.00 | 0.00 | |
| 11 | Bogotá | 0.00 | 0.00 | 0.00 | 0.000000 | 0.0 | 0.010000 | 0.00 | 0.00 | 0.00 | ... | 0.00 | 0.00 | 0.0 | 0.02 | 0.00 | |

- After the above-mentioned transformations are done to the collected data, it can be used for further analysis using Machine Learning Algorithms, since it is now clean and pre-processed.

The Data set is visualized.