

# Grouping the world's most populous Metropolitan cities

By G.Mukkes

## Introduction

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.

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.

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

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

	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 100 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)

	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

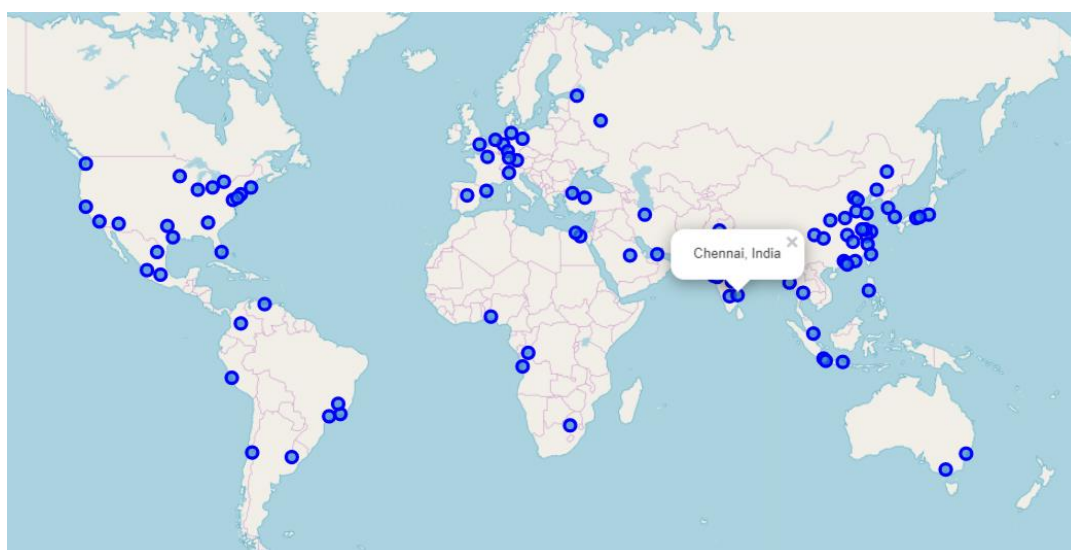
- After getting the top 10 locations for every city, we can get the set of unique venue categories and then one-hot-vector encode them, after grouping all the venues by their city to get the frequency of occurrence (100,462).

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

	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	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	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	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	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	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	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	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	0.00	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	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	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	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	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.



## Methodology

### Univariate Exploratory Data Analysis:

From the frequency table (Table 1), we can see that there are 39 unique countries in this list.

Out of the 100 most populous metropolitan areas in the world, 20 of them are present in China, 15 in USA, and 9 in India.

Table 1:

Country	Count
China	20
United States	15
India	9
Germany	6
Japan	3
Indonesia	3
Mexico	3
Brazil	3
Pakistan	2
South Korea	2
Turkey	2
Australia	2
Spain	2
Egypt	2
Russia	2
France	1
South Africa	1
Netherlands	1
Democratic Republic of the Congo	1
Iran	1
Peru	1
Myanmar	1
Philippines	1
Chile	1
Venezuela	1
Canada	1
United Arab Emirates	1
Singapore	1
Angola	1
Thailand	1
Argentina	1
Colombia	1
Nigeria	1
Taiwan	1
United Kingdom	1
Bangladesh	1
Saudi Arabia	1
Hong Kong	1
Italy	1

From the FourSquare API, the top 100 (if present) venues of each city was retrieved along with their coordinates and venue categories. By analysing the Venue Categories, we can see that there are totally 462 unique categories, and a total of 6868 venues were retrieved.

Using the one-hot-encoded city-grouped version of this data set, the analysis was carried out.

### Clustering Analysis:

Since this is an unsupervised Machine Learning problem, for clustering the data set, we will be using the K-Means++ algorithm. K-Means++ algorithm is a variation of the unsupervised K-Means algorithm which doesn't use the Random Initialization of cluster centroids. Instead, an optimised approach for initialising the clustered centroids is followed. Even though initialization taken extra time, the k-means part itself converges very quickly after this seeding and thus the algorithm actually lowers the computation time and decreases the error rate. The elbow-curve method is used to determine the optimal number of clusters.

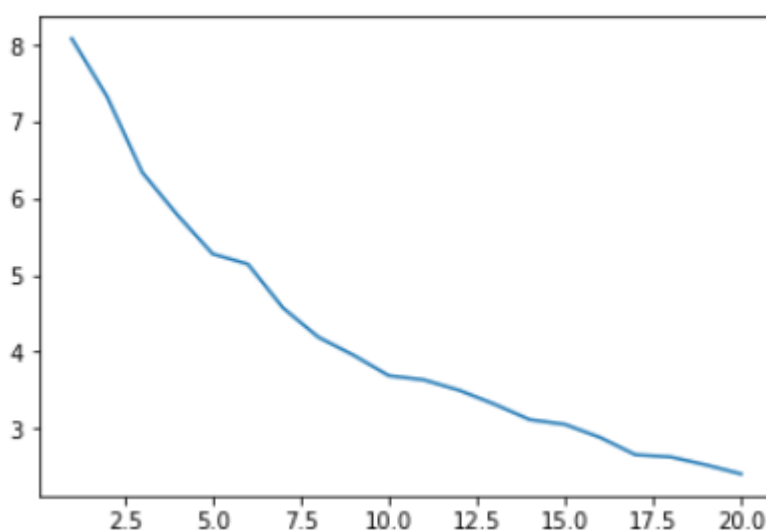
After Initialization, the algorithm is the same as that of k-means i.e., iteratively tries to minimise the intra-cluster variance and maximise the inter-cluster variance by re-calculating the least squared Euclidean distances and reassigning cluster centroids till they don't get re-assigned anymore.

## Results

We use the KMeans model from the sklearn package. The parameters are initialized as follows:

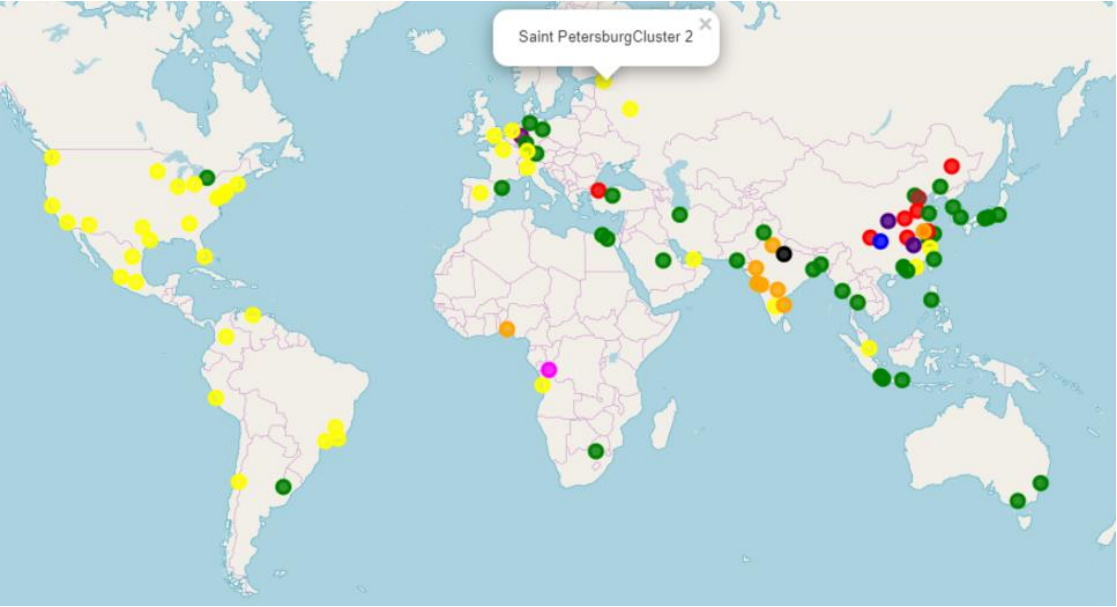
1. Random State is set to 0 to reproduce the results later on.
2. The maximum number of iterations for each analysis is set to 300.
3. The number of times the algorithm is re-run to choose the best (i.e. least) within cluster variance is set to 10.
4. The initialization of centroids is set to 'k-means++'.

The algorithm is run on the dataset with different number of clusters, and the within-cluster variance or inertia is plotted in a line chart.



From this graph, we can see that sharp elbows exist at  $n=6$  and  $n=10$ . For this analysis, we will choose  $n=10$  as our optimal number of clusters, since the corresponding inertia is less.

The selected model with  $n=10$  was re-run to get the results, and they were visualized on a Folium map. (The cluster details are present in the Jupyter Notebook).



Cluster 1

[41]:

city\_merged.loc[city\_merged['Cluster Labels'] == 0, city\_merged.columns[[1,2] + list(range(6, city\_merged.shape[1]))]]

[41]:

	Metropolitan	Country	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
16	Wuhan	China	Hotel	Diner	Nightclub	Park	Plaza	Food Court	Hubei Restaurant	Metro Station	River	Farmers Market
18	Chengdu	China	Hotel	Shopping Mall	Clothing Store	Chinese Restaurant	Lounge	Coffee Shop	Water Park	Convenience Store	Breakfast Spot	Gym / Fitness Center
24	Istanbul	Turkey	Hotel	Restaurant	Turkish Restaurant	History Museum	Jewelry Store	Historic Site	Seafood Restaurant	Coffee Shop	Art Gallery	Gift Shop
35	Changzhou	China	Hotel	History Museum	Turkish Restaurant	Movie Theater	Shopping Mall	Stadium	Theater	Restaurant	Falafel Restaurant	Farmers Market
42	Jinan	China	Hotel	Soccer Field	Basketball Stadium	Hotel Bar	Fondue Restaurant	Exhibit	Fabric Shop	Falafel Restaurant	Farmers Market	Fast Food Restaurant
44	Harbin	China	Hotel	Bed & Breakfast	Chinese Restaurant	Asian Restaurant	Dongbei Restaurant	River	Furniture / Home Store	Food	Falafel Restaurant	Farmers Market
48	Zhengzhou	China	Park	History Museum	Hotel	Department Store	Food & Drink Shop	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Filipino Restaurant	Film Studio

## Discussion

From the results, we can observe that most of the European cities and American (north and south) are very similar to each other. Similarly, most of the Arabian cities and south-east/east Asian, Australian and a few European cities are similar. These countries have Asians in their population, who have become permanent residents of the country over time. Most of the Indian cities fall under a cluster, whose features are also interestingly shared by Nanjing in China and Lagos in Nigeria. Further exploration found out that many Indian students study in Nanjing, and the local lifestyle is a bit similar to the cities here. Nigeria has over 217,000 Indians currently, and has over 8000 permanently settled Indians. Hinduism, the major religion of India, is a major religion in Nigeria too. Interestingly, Bangalore, the Silicon Valley of India, falls under the same cluster as all the American cities (including San Francisco, the Silicon Valley of America). Istanbul falls under the same cluster as many Chinese cities. Further exploration found out that Istanbul has a significantly large Chinese community.

Based on these observations, we can see that there exist significant connections between different cities, even if they are 1000s of miles apart.

## Conclusion

In this study, I grouped different metropolitan areas of the world into 10 clusters. The features used for the clustering process are various types of venues that are present in and around the selected cities. The K-Means ++ algorithm was used to cluster these cities depending on how similar they were to one another. By analysing the cities present in the different clusters, I was able to find out probable reasons to why they might fall under the same cluster, even though it seemingly looked as if they no connections to each other in the first place. For example, I was able to identify that Lagos, the most populous city of Nigeria and the fastest growing urban city of the African Continent, has a significantly large number of Indians settled in the city. This might have led to the alterations of the native neighbourhood into something which resembles Indian Neighbourhoods. Surprisingly, Lagos was the only non-Indian city in the cluster to which it was assigned to. Using these observations, one can find out whether the new city that he/she is moving to has a lifestyle which is similar or dissimilar to the one that he/she is currently residing in.