CLUSTERING THE WORLD CAPITALS

Mudakkar M. Khadim

February, 2020

1. Introduction

1.1 Background

Sister cities or twin towns is a known concept in today's world where two or more cities or countries form a sort of agreement to promote their ties. These cities or countries need not be in the same geographical areas. The concept is mainly aimed at increasing the friendship and understanding between the cities or countries that ultimately helps in increasing trade and tourism. Further details of this concept can be seen on this link.

In this report, we try to find sister cities for world capitals. We select several capital cities across different continents and try to form groups of sister cities using a popular machine learning clustering algorithm and by using the location data of these cities.

1.2 Problem

Currently, the concept of sister cities is more driven by political or social ties between cities or countries. The objective of this report is to form clusters of capital cities that are closer to each other based on some physical attributes.

1.3 Interests

This piece of research can be of interest for many different groups some of which are given below;

a) Tourists or travelers
b) Tourism companies or authorities
c) Traders/Trading companies
d) International students

2. Data

For this research item, we will at least two different data sets. First, country names and their capital cities along with the latitude and longitude information. This information is available on Kaggle and can be accessed via this link. A sample data is shown in below table.

CountryName	CapitalName	CapitalLatitude	CapitalLongitude	CountryCode	ContinentName
Somaliland	Hargeisa	9.55	44.05	NULL	Africa
South Georgia and South Sandwich Islands	King Edward Point	-54.283333	-36.5	GS	Antarctica
French Southern and Antarctic Lands	Port-aux-FranÃ§ais	-49.35	70.216667	TF	Antarctica

Second, location data for the selected cities. This will include information of different venues (like hotels, restaurants, parks, etc..) within a certain radius. This information is accessed using [Foursquare](https://foursquare.com) API. A sample data is shown in below table.

	Country	Capital	Capital_Latitude	Capital_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
0	Somaliland	Hargeisa	9.550000	44.050000	Hiddo - Dhawr	9.551411	44.047806	Comfort Food Restaurant
1	South Georgia and South Sandwich Islands	King Edward Point	-54.283333	-36.500000	جزيرة سندويشة	-54.282935	-36.495176	Beer Bar
2	South Georgia and South Sandwich Islands	King Edward Point	-54.283333	-36.500000	Bilinmeyen Yer	-54.281560	-36.506960	Racetrack
3	South Georgia and South Sandwich Islands	King Edward Point	-54.283333	-36.500000	Tang Ke Lek Harbour	-54.280980	-36.508610	Harbor / Marina
4	French Southern and Antarctic Lands	Port-aux-Français	-49.350000	70.216667	Book Time	-49.352470	70.218711	Bookstore

The latitude and longitude information available in the first data set is actually used to get the location data. And then clusters or groups of sister (similar) capital cities will be formed using some clustering algorithm based on the location data.

3. Methodology

3.1 Data Cleaning

The original file, available on Kaggle, containing the capital city names with their latitude and longitude information has total 245 cities listed there. However, we selected first 150 cities only for this project. Those 150 cities are mapped in the chart 1.

On next step, we extracted location data for these 150 capital cities using Foursquare API. The data was extracted within the radius of 1000 meters limiting the results to 100 venues only. The location data as city level was later reviewed to observe any abnormal or outlier behavior. A sample set of data is given in table 1.

As shown in table 1, there was huge variation between cities in terms of number of venues available for doing cluster analyses. Where some cities had 100 venues, some others had as low as 3. This could create a noise in the analyses. In order to minimize this data volatility we decided to keep cities with at least 60 venues for our analyses. This was to ensure that comparable data points are available among cities to form clusters. After that, we were left with 50 cities only to perform cluster analyses for. The list of final 50 cities along with number of venues available is given in table 2.

Chart 1. *Mapping of 150 selected capital cities*

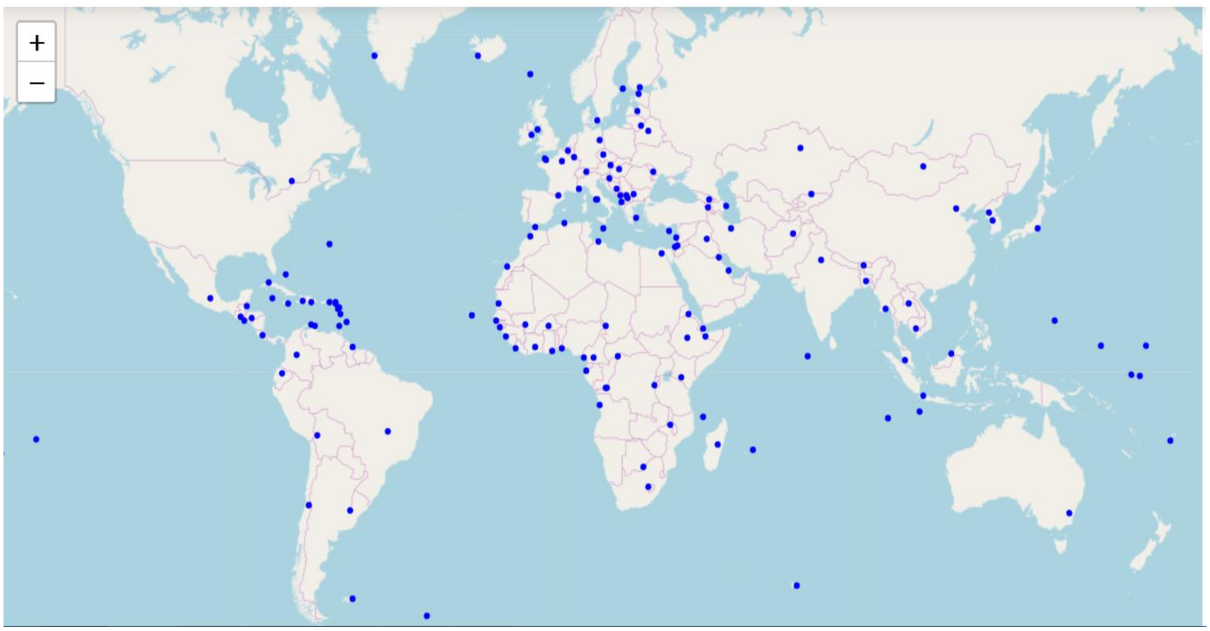


Table 1. *Number of venues extracted against each of the 150 capital cities*

Capital	Country	Capital_Latitude	Capital_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
Accra	3	3	3	3	3	3	3
Addis Ababa	3	3	3	3	3	3	3
Algiers	8	8	8	8	8	8	8
Amman	100	100	100	100	100	100	100
Andorra la Vella	47	47	47	47	47	47	47
Antananarivo	11	11	11	11	11	11	11
Asmara	9	9	9	9	9	9	9
Astana	81	81	81	81	81	81	81
Athens	100	100	100	100	100	100	100
Avarua	4	4	4	4	4	4	4

Table 2. *List of 50 capital cities finally selected for cluster analyses*

Sr. #	Capital	Country	No. of Venues
1	Amman	Jordan	100
2	Astana	Kazakhstan	81
3	Athens	Greece	100
4	Baku	Azerbaijan	97
5	Berlin	Germany	100
6	Bishkek	Kyrgyzstan	82
7	Brussels	Belgium	100
8	Budapest	Hungary	100
9	Cairo	Egypt	61
10	Copenhagen	Denmark	100
11	Guatemala City	Guatemala	67
12	Helsinki	Finland	100
13	Jakarta	Indonesia	99
14	Jerusalem	Palestine	82
15	Kuala Lumpur	Malaysia	100
16	Kuwait City	Kuwait	100
17	Manama	Bahrain	70
18	Mexico City	Mexico	100
19	Minsk	Belarus	100
20	Monaco	Monaco	88
21	Nairobi	Kenya	74
22	Nassau	Bahamas	63
23	Nicosia	Cyprus	100
24	Oranjestad	Aruba	92
25	Ottawa	Canada	100
26	Paris	France	100
27	Phnom Penh	Cambodia	70
28	Podgorica	Montenegro	100
29	Prague	Czech Republic	100
30	Pristina	Kosovo	80
31	Reykjavik	Iceland	100
32	Riga	Latvia	100
33	Rome	Italy	100
34	Saint Helier	Jersey	62
35	San Jose	Costa Rica	100
36	Santiago	Chile	100
37	Santo Domingo	Dominican Republic	71
38	Sarajevo	Bosnia and Herzegovina	76
39	Seoul	South Korea	100
40	Skopje	Macedonia	100
41	Sofia	Bulgaria	100
42	Tallinn	Estonia	91
43	Tbilisi	Georgia	61
44	Tehran	Iran	100
45	Tirana	Albania	100
46	Tokyo	Japan	100
47	Ulaanbaatar	Mongolia	100
48	Vienna	Austria	100
49	Vientiane	Laos	86
50	Zagreb	Croatia	100

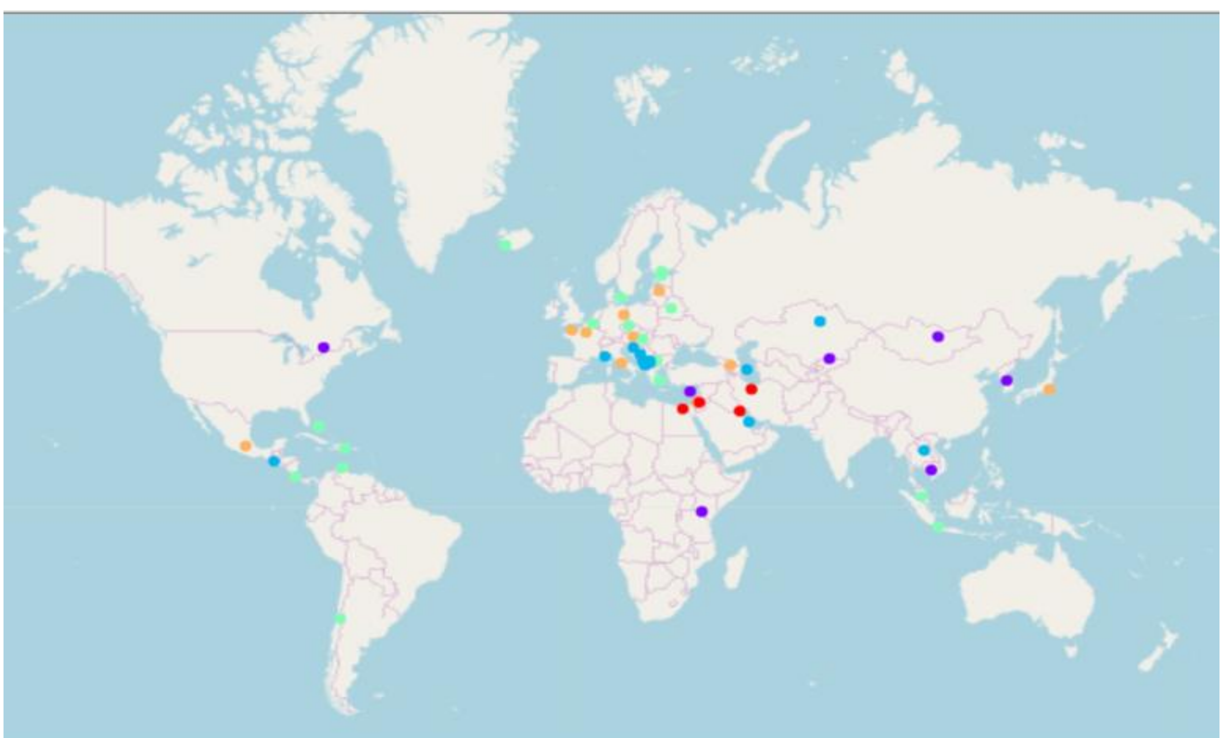
3.2 Data Preparation for Clustering

Now we have the required data available to run the cluster algorithm. However, the clustering is to be done based on venue category information (such as park, hotel, coffee shop etc…) which is a categorical variable. We need to transform this categorical variable into numeric data to run cluster algorithm. One-Hot encoding is used for this transformation. One hot encoding replaces each level/class of a categorical variable with a dummy variable (0,1) such that it takes value 1 if that specific class of the categorical variable is present in that observation and 0 otherwise. So, if a categorical variable has k classes, one-hot-encoding generates k (or k-1) dummy variables. The resulting data frame was then grouped by the capital cities by taking mean of the frequency of occurrence of each category. The data is now ready to run the clustering algorithm.

3.3 K-means Clustering

K-means clustering algorithm was used to group the capital cities into clusters. The algorithm was run using scikit-learn that is a popular library for fitting machine learning models in python. Number of clusters was selected as 5. Below chart shows the result of clustering where each dot represents a capital city and different colors represent different clusters.

Chart 2. *Clustering of capital cities*



4. Results

Results are compiled cluster wise with list of capital cities within each cluster and analyzed against the most common venue (upto 7th most common venue). Details are as given below.

4.1 Cluster 1

There are four capitals in this cluster. Three of the capitals belong to Asian continent while the fourth is from Africa. There are a lot of things common between these cities other than the location data. For example, all four cities belong to Muslim countries and three out of four cities speak the same language (Arabic). Looking at the location data, all of these cities have Café as 1st most common venue and then coffee shop and restaurants at 2nd and 3rd most common places.

***Table 3**. Capital cities in cluster 1 with most common venues*

Capital	Continent	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	7th Most Common Venue
Cairo	Africa	0	Café	Plaza	Coffee Shop	Egyptian Restaurant	Theater	Pastry Shop	Hotel Bar
Tehran	Asia	0	Café	Persian Restaurant	Coffee Shop	Sandwich Place	Theater	Art Gallery	Pastry Shop
Amman	Asia	0	Café	Middle Eastern Restaurant	Italian Restaurant	Historic Site	Bookstore	Breakfast Spot	Arts & Crafts Store
Kuwait City	Asia	0	Café	Coffee Shop	Middle Eastern Restaurant	Japanese Restaurant	Italian Restaurant	Hookah Bar	Pizza Place

4.2 Cluster 2

Cluster 2 has seven capital cities from five different continents however four of the cities belong to Asian continent. All seven cities have Coffee shop as 1st most common venue and then Café and Restaurant as second and third most common places.

***Table 4**. Capital cities in cluster 2 with most common venues*

Capital	Country	Continent	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	7th Most Common Venue
Seoul	South Korea	Asia	1	Coffee Shop	Café	Korean Restaurant	Italian Restaurant	Park	Scenic Lookout	French Restaurant
Bishkek	Kyrgyzstan	Asia	1	Coffee Shop	Café	Asian Restaurant	Hotel	Turkish Restaurant	Japanese Restaurant	Bar
Ulaanbaatar	Mongolia	Asia	1	Coffee Shop	Restaurant	Café	Bakery	Italian Restaurant	Pub	Lounge
Phnom Penh	Cambodia	Asia	1	Coffee Shop	Chinese Restaurant	Asian Restaurant	Hotel	Japanese Restaurant	Café	Thai Restaurant
Ottawa	Canada	Central America	1	Coffee Shop	Café	Restaurant	Tapas Restaurant	Hotel	Food Truck	Italian Restaurant
Nicosia	Cyprus	Europe	1	Coffee Shop	Greek Restaurant	Bar	Café	Wine Bar	Italian Restaurant	Restaurant
Nairobi	Kenya	Africa	1	Coffee Shop	Café	African Restaurant	Hotel	Bar	Fast Food Restaurant	Ice Cream Shop

4.3 Cluster 3

There are twelve capital cities in cluster 3 and eight of those are from Europe. Most of the cities falling under this cluster have Café and Restaurants as 1st and 2nd most common venue.

Table 5. *Capital cities in cluster 3 with most common venues*

Capital	Continent	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	7th Most Common Venue
Tirana	Europe	2	Cocktail Bar	Café	Italian Restaurant	Bar	Hotel	Lounge	Restaurant
Baku	Europe	2	Restaurant	Tea Room	Hotel	Café	Coffee Shop	Turkish Restaurant	Eastern European Restaurant
Manama	Asia	2	Café	Hotel	Breakfast Spot	Italian Restaurant	Coffee Shop	Restaurant	Lounge
Sarajevo	Europe	2	Café	Restaurant	Hotel	Italian Restaurant	Hostel	Cocktail Bar	Theater
Zagreb	Europe	2	Café	Restaurant	Bar	Bakery	Hotel	Gym / Fitness Center	BBQ Joint
Guatemala City	Central America	2	Café	Pizza Place	Fast Food Restaurant	Restaurant	Coffee Shop	Steakhouse	Burger Joint
Astana	Asia	2	Coffee Shop	Café	Restaurant	Italian Restaurant	Electronics Store	Karaoke Bar	Diner
Pristina	Europe	2	Restaurant	Bar	Hotel	Dessert Shop	Mediterranean Restaurant	Fast Food Restaurant	Café
Vientiane	Asia	2	Hotel	Café	Asian Restaurant	Bar	Coffee Shop	Pizza Place	French Restaurant
Skopje	Europe	2	Café	Hotel	Bar	Restaurant	Historic Site	Italian Restaurant	Bookstore
Monaco	Europe	2	Italian Restaurant	French Restaurant	Restaurant	Hotel	Bar	Cocktail Bar	Garden
Podgorica	Europe	2	Café	Hotel	Bar	Italian Restaurant	Pizza Place	Fast Food Restaurant	Restaurant

4.4 Cluster 4

Cluster 4 is the largest cluster with 17 cities mostly from Europe and America. The most common venues in this cluster are Bar and Restaurants.

Table 5. *Capital cities in cluster 4 with most common venues*

Capital	Continent	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	7th Most Common Venue
Oranjestad	North America	3	Caribbean Restaurant	Bar	Shopping Mall	Hotel	Breakfast Spot	Coffee Shop	Restaurant
Nassau	North America	3	Boat or Ferry	Seafood Restaurant	Caribbean Restaurant	Bar	Hotel	Beach	Fast Food Restaurant
Minsk	Europe	3	Cocktail Bar	Bar	Restaurant	Café	Park	Coffee Shop	Boutique
Brussels	Europe	3	Sandwich Place	Hotel	Bar	Greek Restaurant	Brasserie	Coffee Shop	Vegetarian / Vegan Restaurant
Sofia	Europe	3	Coffee Shop	Restaurant	Vegetarian / Vegan Restaurant	Bakery	Dessert Shop	Park	Café
Santiago	South America	3	Bar	Pizza Place	Chinese Restaurant	Martial Arts Dojo	Hostel	Café	Asian Restaurant
San Jose	Central America	3	Bar	Sandwich Place	Hotel	Coffee Shop	Ice Cream Shop	Café	Restaurant
Prague	Europe	3	Café	Bakery	Vietnamese Restaurant	Bar	Gym / Fitness Center	Hotel	Pub
Copenhagen	Europe	3	Italian Restaurant	Café	Scandinavian Restaurant	Bakery	Coffee Shop	Gym / Fitness Center	Ice Cream Shop
Santo Domingo	North America	3	Hotel	Pharmacy	Ice Cream Shop	BBQ Joint	Restaurant	Pizza Place	Bar

Tallinn	Europe	3	Café	Asian Restaurant	Park	Electronics Store	Restaurant	Hotel	Cosmetics Shop
Helsinki	Europe	3	Scandinavian Restaurant	Hotel	Sushi Restaurant	Japanese Restaurant	Coffee Shop	Bakery	Middle Eastern Restaurant
Athens	Europe	3	Bar	Café	Coffee Shop	Dessert Shop	Theater	Cocktail Bar	Bookstore
Budapest	Europe	3	Clothing Store	Coffee Shop	Hotel	Restaurant	Chinese Restaurant	Multiplex	Bakery
Reykjavik	Europe	3	Bar	Seafood Restaurant	Hotel	Café	Coffee Shop	Scandinavian Restaurant	Burger Joint
Jakarta	Asia	3	Chinese Restaurant	Hotel	Noodle House	Seafood Restaurant	Indonesian Restaurant	Restaurant	Asian Restaurant
Kuala Lumpur	Asia	3	Malay Restaurant	Hotel	Asian Restaurant	Thai Restaurant	Clothing Store	Chinese Restaurant	Indonesian Restaurant

4.5 Cluster 5

Last cluster has 10 capital cities and most common venues are Hotel & Restaurants.

***Table 5**. Capital cities in cluster 4 with most common venues*

Capital	Continent	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	7th Most Common Venue
Jerusalem	Asia	4	Historic Site	Hotel	Restaurant	Mediterranean Restaurant	Park	Italian Restaurant	Burger Joint
Vienna	Europe	4	Café	Hotel	Plaza	Museum	Asian Restaurant	Bar	Concert Hall
Paris	Europe	4	Hotel	French Restaurant	Plaza	Japanese Restaurant	Historic Site	Theater	Italian Restaurant
Tbilisi	Europe	4	Hotel	Caucasian Restaurant	Restaurant	Bed & Breakfast	Bus Station	Supermarket	Metro Station
Berlin	Europe	4	Hotel	History Museum	Plaza	Museum	Art Gallery	Art Museum	Concert Hall
Rome	Europe	4	Italian Restaurant	Plaza	Ice Cream Shop	Monument / Landmark	Sandwich Place	Boutique	Hotel
Tokyo	Asia	4	Historic Site	Soba Restaurant	Convenience Store	Coffee Shop	Hotel	Ramen Restaurant	Japanese Restaurant
Saint Helier	Europe	4	Hotel	Coffee Shop	Pub	Restaurant	Department Store	Fish & Chips Shop	Harbor / Marina
Riga	Europe	4	Restaurant	Eastern European Restaurant	Hotel	Bar	Plaza	Park	Café
Mexico City	Central America	4	Mexican Restaurant	Museum	Art Museum	Hotel	Ice Cream Shop	Coffee Shop	Arts & Crafts Store

5. Discussion

As the results show, k-mean clustering algorithm does a good job in clustering the world capital cities based on the location data. Resulting clusters not only make sense based on the available venue information but those are also closer to each other in terms of geography, culture & religion. Also, three most common venue categories in these capital cities are Restaurants, Café and Coffee shops.

6. Conclusion

In this research report we tried to classify world capitals into similar groups following the sister city concept. Groups are formed in such a way that cities within one group are similar to each other but different than those falling in another group. Groups are based on the location data using a popular machine learning algorithm for clustering, called K-means clustering. Results obtained show a good fit of clustering algorithm.