Clustering/Segmenting popular touristic cities around the world

Author: Sergio Gustavo Sánchez Linares

sg.sanchez@acad.ucb.edu.bo

September 10, 2019

1: Introduction

1.1 Background

Tourism is and have been one of the most popular activities in all the world for a long time. People use to travel abroad for different reasons and they usually visit places according to their own likings and interests: they might want to meet people from another culture, or city sightseeing, maybe to visit museums and buildings, or going to see natural wonders. Those interests are usually well established when the people know where they are going or because they are visiting again that place, but sometimes they don't have a very clear idea of what is going to find in a specified city or place, maybe because it is the first time they are going to that place, or they just didn't expect what is that city like.

Nowadays the world has a lot of completely different cultures, buildings, architecture, art, traditions, even speaking of the same country o even the same city, such as large multicultural cities like New York, Los Angeles, Toronto, Paris, and many others, and this situation has brought to people to have a lot of different options to travel but also a level of uncertainty about what is going to find in the city which they are travelling to and what are the most important activities to do.

But on the other side, many cities share some characteristics when it comes to cultural traditions, local food, way of living, building style and architecture, and to know exactly what are the cities that share this similarities could be very advantageous of people who want to know different places but that are related one with each other, so they can live similar good experiences and having an idea of what is the city like.

1.2 Problem

Data about a few of the most popular and visited international cities in the globe may help to correlate which of those cities share common characteristics, based on what are the most popular places and activities to visit or do in that city. Analyzing the data of that common places, we may ask:

- *What are the groups of cities that are similar to each other?*
- *What characteristics do they share?*
- *What are the most common places to visit or activities to do in each group of cities?*

1.3 Interest

The results of this exploration and analysis may be very useful as a guide to people having in mind a trip to a city that is included in the most popular around the world, as this segmentation would provide previous knowledge about that cities and their characteristics.

The final result may be available to the public through a mobile application or a progressive web app, so users can search for cities of their interests.

Also it can be very useful to flight companies and tourism guided tours, to make offers and discounts for cities that share common characteristics and that people may be very interested in.

2: Data acquisition

2.1 Data sources

First we will get the 100 most popular cities listed by international visitors (available at Wikipedia: https://en.wikipedia.org/wiki/List_of_cities_by_international_visitors), ranked by the *Euromonitor Rank*. We will scrape the data from the table displayed using *Beautiful Soup* 4. Here an example of a part of the table in the Wikipedia page:

Rank Euromonitor	Rank Mastercard	City	Country	Arrivals 2017 Euromonitor	Arrivals 2016 Mastercard	Growth in arrivals Euromonitor	Income (billions $) Mastercard
1	11	Hong Kong	Hong Kong	25,695,800	8,370,000	−3.1 %	6.84
2	1	Bangkok	Thailand	23,270,600	21,470,000	9.5 %	14.84
3	2	London	United Kingdom	19,842,800	19,880,000	3.4 %	19.76
4	6	Singapore	Singapore	17,681,800	12,110,000	6.1 %	12.54
5		Macau	Macau	16,299,100		5.9 %	
6	4	Dubai	United Arab Emirates	16,010,000	15,270,000	7.7 %	31.30
7	3	Paris	France	14,263,000	18,030,000	−0.9 %	12.88
8	5	New York City	United States	13,100,000	12,750,000	3.6 %	18.52
9	54	Shenzhen	China	12,962,000	2,120,000	3.1 %	0.83
10	7	Kuala Lumpur	Malaysia	12,843,500	12,020,000	4.5 %	11.34

Along with the data collected from the table above, we may obtain the coordinates for each city, given that the Foursquare API is easier to use with Latitude and Longitude values rather than addresses that have to be resolved and may result in geocode errors (as mentioned in the developer docs). Those coordinates will be obtained using the Geopy library, making sure we don't get errors in the resolution of the coordinates before using the API.

Having collected the cities data and their coordinates, the main data we will use all along the project will be mainly extracted from the Foursquare API using requests methods. This API which holds information about the most popular places for each city. We will retrieve precisely the name, category and coordinates for each venue, and a maximum of 100 most

popular venues for each city selected. Here is an example of venues data for Neighborhood groups in Toronto, Canada, retrieved in a *Jupyter Notebook*:

	Postal Code	Neighborhood Group Latitude	Neighborhood Group Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M4E	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	M4E	43.676357	-79.293031	Tori's Bakeshop	43.672114	-79.290331	Vegetarian / Vegan Restaurant
2	M4E	43.676357	-79.293031	The Beech Tree	43.680493	-79.288846	Gastropub
3	M4E	43.676357	-79.293031	Ed's Real Scoop	43.672630	-79.287993	Ice Cream Shop
4	M4E	43.676357	-79.293031	The Fox Theatre	43.672801	-79.287272	Indie Movie Theater

2.2 Data cleaning

The original table of Wikipedia's page had many columns describing both ranks (Euromonitor and Mastercard), Arrivals in 2017 and 2016, and percentages indicating the growth of arrivals. These information is not pertinent for the analysis nor for the clustering model and it is out of the scope of study of this project, so they were ignored. We only stayed with the *City* and *Country* columns, given that we only needed to know what where the most popular and visited cities.

In the case of the data retrieved from the Foursquare API, there was no problem, because the API returned very well structured values without missing ones.

3: Exploratory Data Analysis

3.1 Visualizing the cities retrieved

After retrieving the data and organize it in an individual DataFrame, with cities, respective countries and coordinates, we proceed to build a map to visualize the position of the cities.

Using Folium, it is easy to build a map with all the cities that are analyzed in this project. The map with the cities resulted like this:

3.2 Exploring the venues dataset

When the venues dataset was retrieved, it counted with 9627 venues with 7 attributes each one. To explore this data, we performed some operations before preprocessing it to build the model:

Exploring the quantity of venues per city: Almost all of the cities got the limit of 100 venues, but not all of them.

```
[137]: City
       Abu Dhabi          11
       Agra               45
       Amsterdam         100
       Antalya           100
       Artvin            100
       Athens            100
       Auckland          100
       Bangkok           100
       Barcelona         100
       Beijing           100
       Berlin            100
       Brussels          100
       Budapest          100
       Buenos Aires      100
       Cairo             100
       Cancún            100
       Chennai           100
       Chiang Mai        100
       Chiba             100
```

Exploring the cities that have the least quantity of venues: There was 6 cities which did not reach the 100 venues, and Abu Dhabi only had 11 venues.

[77]:

City	Latitude	Longitude	Venue
Abu Dhabi	11	11	11
Zhuhai	14	14	14
Guilin	37	37	37
Agra	45	45	45
Ha Long	51	51	51
Jaipur	69	69	69
Phnom Penh	100	100	100
Penang Island	100	100	100
Pattaya	100	100	100
Paris	100	100	100

Determining how many unique categories of venues were retrieved: There were 493 distinct categories of venues. This fact is important because when applying One Hot Encoding, this will be the total number of columns in the modified dataset.

3.3 Preprocessing the venues dataset

After an exploration, we need to prepare the data to make it fit to the model we will apply later. First we applied **One Hot Encoding** to transform categorical variables into numerical. After applied, there were 494 attributes: 493 feature columns and one using as an index, which was the name of the city.

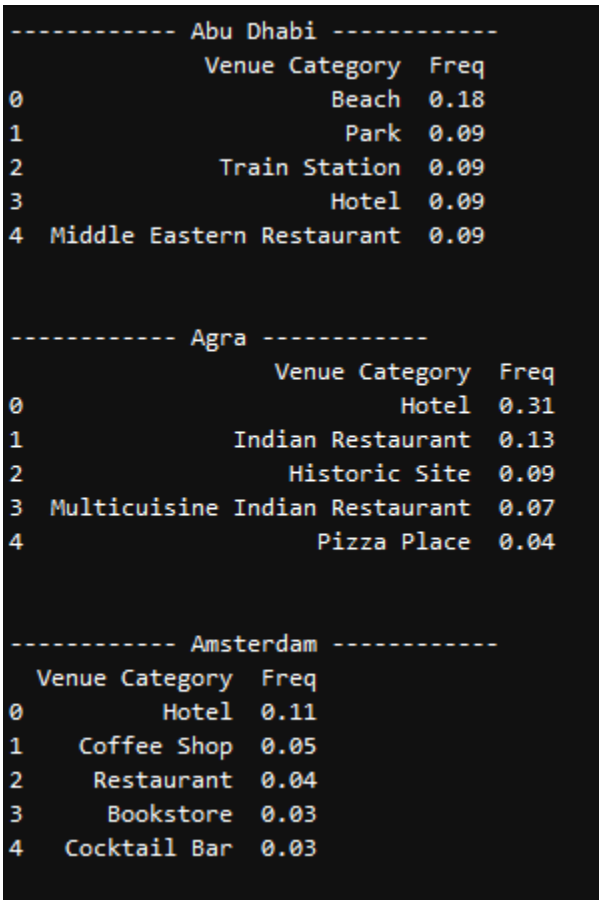
[90]:

	City Name	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport
0	Hong Kong	0	0	0	0	0
1	Hong Kong	0	0	0	0	0
2	Hong Kong	0	0	0	0	0
3	Hong Kong	0	0	0	0	0
4	Hong Kong	0	0	0	0	0

5 rows × 494 columns

Then we proceed to group the observations by City and taking the mean of the features to know the frequency of each venue per city.

Given that, we could get the five more popular venues for each city. Here are some examples:



```
------------ Abu Dhabi ------------
              Venue Category  Freq
0                      Beach  0.18
1                       Park  0.09
2               Train Station  0.09
3                      Hotel  0.09
4  Middle Eastern Restaurant  0.09


------------ Agra ------------
                  Venue Category  Freq
0                          Hotel  0.31
1             Indian Restaurant  0.13
2                Historic Site  0.09
3  Multicuisine Indian Restaurant  0.07
4                   Pizza Place  0.04


------------ Amsterdam ------------
  Venue Category  Freq
0          Hotel  0.11
1    Coffee Shop  0.05
2     Restaurant  0.04
3      Bookstore  0.03
4   Cocktail Bar  0.03
```

```
------------ Mumbai ------------
      Venue Category  Freq
0  Indian Restaurant  0.14
1              Café  0.10
2             Hotel  0.07
3     Scenic Lookout  0.05
4                Bar  0.04


------------ Munich ------------
   Venue Category  Freq
0           Café  0.09
1  Ice Cream Shop  0.08
2          Hotel  0.07
3          Plaza  0.06
4    Cocktail Bar  0.05


------------ New York City --------
      Venue Category  Freq
0               Park  0.14
1      Ice Cream Shop  0.05
2      Scenic Lookout  0.04
3          Bookstore  0.04
4  Italian Restaurant  0.03
```

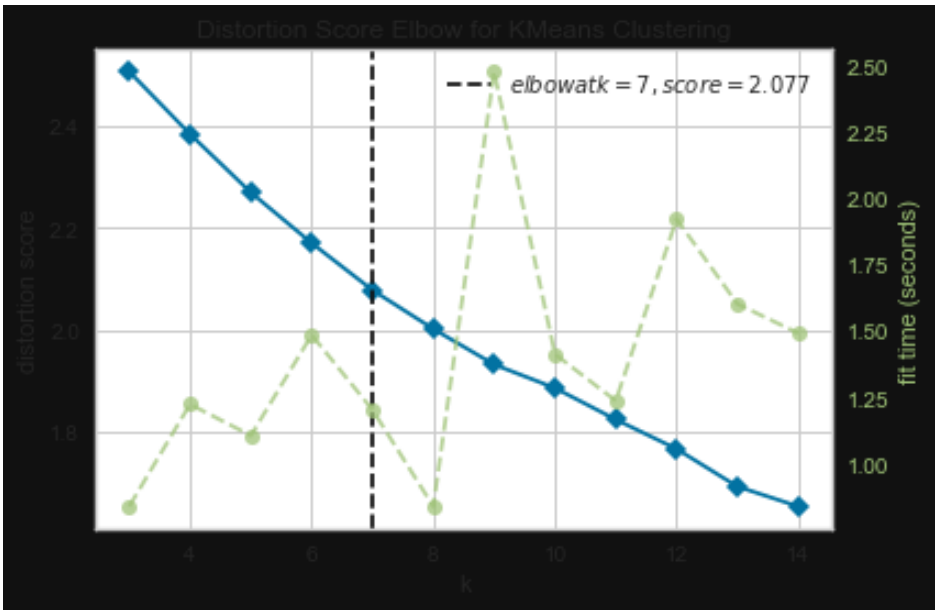
Having the last frequencies, we could build another dataset, including the 10 most popular venues per city. This new dataset will be useful to compare easily the most common venues between each city when it belongs to a cluster with others:

	City Name	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
0	Abu Dhabi	Beach	Boat or Ferry	Campground	Train Station	Middle Eastern Restaurant	Asian Restaurant	Park	Hotel	Grocery Store	Airport
1	Agra	Hotel	Indian Restaurant	Historic Site	Multicuisine Indian Restaurant	Fast Food Restaurant	Coffee Shop	Pizza Place	Café	Resort	Bed & Breakfast
2	Amsterdam	Hotel	Coffee Shop	Restaurant	Café	Cocktail Bar	Bookstore	Breakfast Spot	Bakery	Plaza	Canal
3	Antalya	Coffee Shop	Park	Seafood Restaurant	Bar	Gym / Fitness Center	Café	Historic Site	Bookstore	Restaurant	Trail
4	Artvin	Café	Restaurant	Hotel	Turkish Restaurant	Bar	Bed & Breakfast	Steakhouse	Recreation Center	Farm	Scenic Lookout

3.4 Clustering the cities

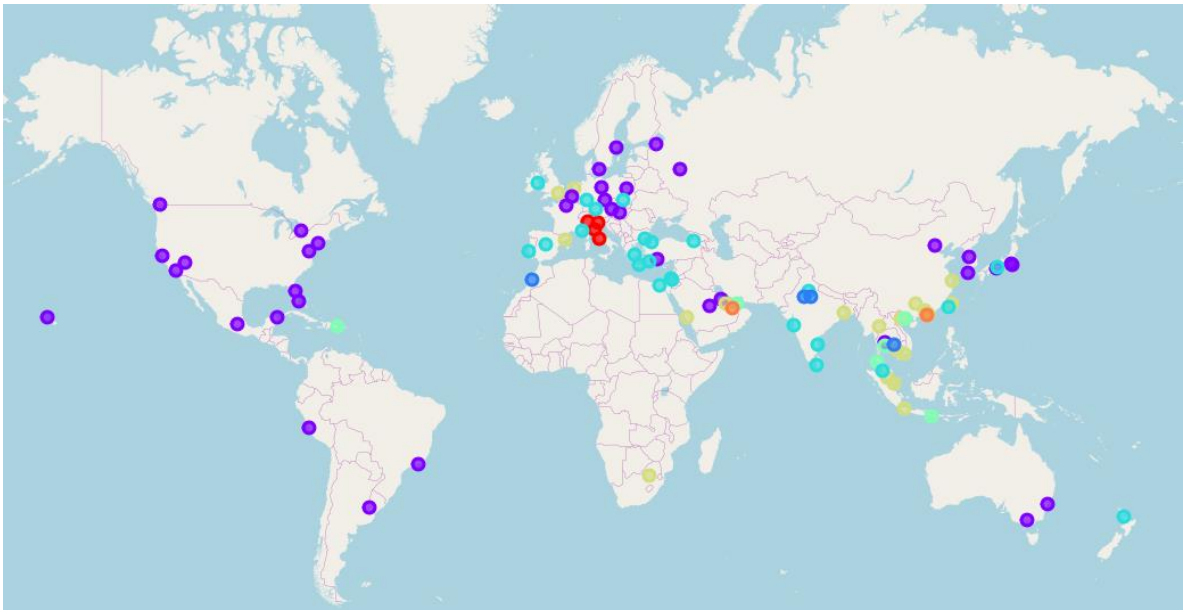
Now we head up to the clustering section, where we applied the K-Means algorithm for simplicity.

As we are applying K-Means, it is necessary (or at least a good practice) to find the optimum value for the K parameter (number of clusters to group the data). So we ran a library which performed the *elbow method* to determine the value of K:



Although it returned a value, we can see that the elbow is not clear enough, meaning that it is possible that K-Means is not a good fit for this dataset, we need to do some more preprocessing or finally, clustering is not possible.

Despite of that, we ran the model with a value of K = 7. Then associated the resulting labels to the last DataFrame we created and visualized the cities within each cluster with a color coded map:



4: Analyzing the resulting clusters

We did an evaluation in the notebook for each one of the seven clusters. I recommend you to see the results there. Anyway, here I present each cluster with the four most common venues:

Cluster 1

City	Country	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Rome	Italy	Ice Cream Shop	Historic Site	Plaza	Sandwich Place
Milan	Italy	Hotel	Boutique	Italian Restaurant	Plaza
Venice	Italy	Italian Restaurant	Hotel	Ice Cream Shop	Plaza
Florence	Italy	Hotel	Italian Restaurant	Ice Cream Shop	Plaza

Main characteristics:

This cluster holds uniquely Italy cities and we can observe that the most common venues are Hotels, Plazas and traditional food and ice cream. It is well known that Italy's pastas and gelato are very popular. Those cities are also places to see monuments, museums and art galleries, which must be much attended.

Cluster 2

City	Country	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Bangkok	Thailand	Coffee Shop	Thai Restaurant	Shopping Mall	Noodle House
Paris	France	Plaza	Hotel	Cocktail Bar	Italian Restaurant
New York City	United States	Park	Ice Cream Shop	Scenic Lookout	Bookstore
Tokyo	Japan	BBQ Joint	Hotel	Chinese Restaurant	Art Museum
Prague	Czech Republic	Café	Park	Ice Cream Shop	Hotel
Miami	United States	Hotel	Beach	Park	Mexican Restaurant
Seoul	South Korea	Park	Coffee Shop	Hotel	Historic Site

Main characteristics:

This cluster seems to be one of the largest, and we can see there are a lot of Coffee Shops, Hotels and a diverse variety of businesses, meaning those cities are mainly metropolis or capitals (such as Washington, BS AS, Lima, Mexico DF, Moscu, etc.) or very fluent finantial areas with a lot of movement all along the day.

Cluster 3

City	Country	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	
Agra	India	Hotel	Indian Restaurant	Historic Site	Multicuisine Indian Restaurant	R
Jaipur	India	Hotel	Indian Restaurant	Historic Site	Café	
Marrakesh	Morocco	Moroccan Restaurant	Hotel	Café	Nightclub	
Siem Reap	Cambodia	Hotel	Historic Site	Cambodian Restaurant	Resort	

Main characteristics:

This cluster with four cities hold many Hotels and Historic Sites, such as traditional food restaurants. This group of cities may be very rich in history and traditions, and such is the case of India.

Cluster 4

City	Country	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	
Delhi	India	Indian Restaurant	Hotel	Café	Restaurant	R
Mumbai	India	Indian Restaurant	Café	Hotel	Scenic Lookout	I
Istanbul	Turkey	Hotel	Soccer Stadium	Historic Site	Dance Studio	
Madrid	Spain	Hotel	Plaza	Restaurant	Spanish Restaurant	I

Main characteristics:

This cluster only has a lot of cities in it, and seems to be mainly a urban area but with a mix of finantial areas with a lot of hotels and restaurants. Seems to be that although these cities are not capitals, they are important urban centres with a lot of shops and businesses to discover.

Cluster 5

City	Country	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Dubai	United Arab Emirates	Resort	Hotel	Beach	Lounge
Phuket	Thailand	Resort	Hotel	Thai Restaurant	Italian Restaurant
Pattaya	Thailand	Hotel	Resort	Restaurant	Massage Studio
Denpasar	Indonesia	Hotel	Resort	Indonesian Restaurant	Café
Ha Long	Vietnam	Vietnamese Restaurant	Hotel	Cave	Resort
Punta Cana	Dominican Republic	Resort	Beach	Golf Course	Hotel

Main characteristics:

This cluster is built with mainly relax and vacation cities, where you can go off of the traffic and loud noises of the urban areas. The most common venue in these cities are Resorts and Hotels, but they have a variety of relaxation businesses.

Cluster 6

City	Country	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Hong Kong	Hong Kong	Hotel	Park	Japanese Restaurant	Dumpling Restaurant
London	United Kingdom	Hotel	Park	Cocktail Bar	Art Museum
Singapore	Singapore	Hotel	Park	Shopping Mall	Chinese Restaurant
Macau	Macau	Hotel	Café	Chinese Restaurant	Portuguese Restaurant
Shenzhen	China	Hotel	Park	Shopping Mall	Coffee Shop
Kuala Lumpur	Malaysia	Hotel	Shopping Mall	Café	Japanese Restaurant
Taipei	Taiwan	Hotel	Dessert Shop	Dumpling Restaurant	Café

Main characteristics:

This cluster has grouped many cities which may be very busy and crowded. The most common venue are Hotels, followed by a lot of traditional food restaurants. Definitely are cities where you go to relax and taste the food.

Cluster 7

City	Country	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Zhuhai	China	Beach	Fish & Chips Shop	Gastropub	Cave
Abu Dhabi	United Arab Emirates	Beach	Boat or Ferry	Campground	Train Station

Main characteristics:

This cluster is built with only two cities, which their most popular venue is the beach. Remember that Abu Dhabi only had 11 observations, so it may not be entirely correct to use this city for the clustering model.

5: Discussion section

We have analyzed deeply the most popular and visited cities all around the world and through the data, we are now able to tell which are great areas for a diverse number of businesses types or activities.

I recommend to learn and deepen the usage of the tools and libraries I used here, they are amazingly simple to implement and use. Anyway there are lots of other tools that could be very useful to do this tasks.

We are also able to tell which cities hold a lot of similarities between them and along with that, it may be a good idea to visit together.

In the future, we may analyze even deeper, by obtaining information about more cities, with more venues and analyzing Top Picks or maybe a specific category of venues.

6: Conclusion section

I am proud of this work as I am starting in this multidisciplinary area and I think this is a very great start for everyone.

Hope I can improve this notebook adding up other tools, applying more clustering models and visualizations to make this document more readable.

Thanks for reading.