# SEGMENTING/CLUSTERING POPULAR TOURISTIC CITIES AROUND THE WORLD

**By:**

**Garepally Nikhil**

## 1. Introduction

•Tourism has been one of the most popular activities in the entire 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.

## 1.1 Problem

- 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?

## 1.2 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 |

## 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.

**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.

[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 |

[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 |

### 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]:

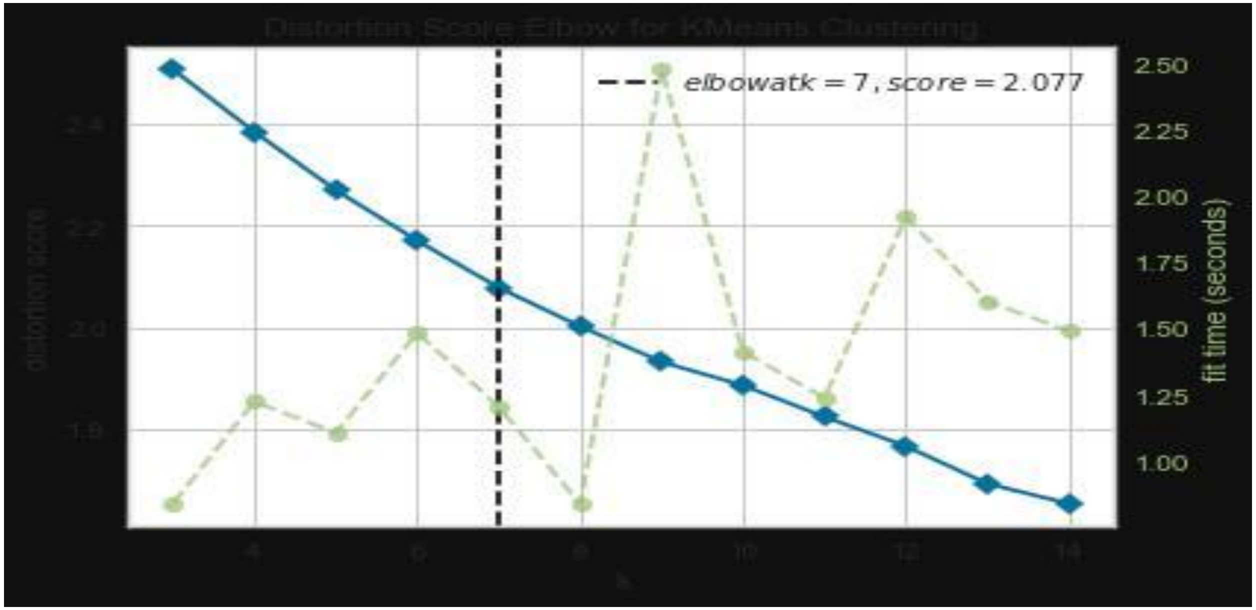
| | City Name | Accessories Store | Adult Boutique | Afghan Restaurant | African Restaurant | Airport |
|---|---|---|---|---|---|---|
| 0 | Hong Kong | 0 | 0 | 0 | 0 | 0 |
| 1 | Hong Kong | 0 | 0 | 0 | 0 | 0 |
| 2 | Hong Kong | 0 | 0 | 0 | 0 | 0 |
| 3 | Hong Kong | 0 | 0 | 0 | 0 | 0 |
| 4 | Hong Kong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

5 rows × 494 columns

**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:



## 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 |

**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 |

## 5. Conclusion

•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.

•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.

•


Thank
you !!