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.

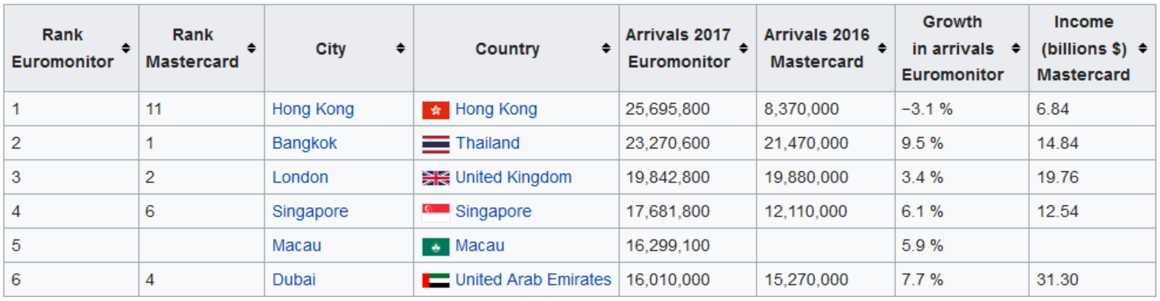
The final result may be available to the public through a mobile application or a progressive web app, so users can search for cities of their interests.

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

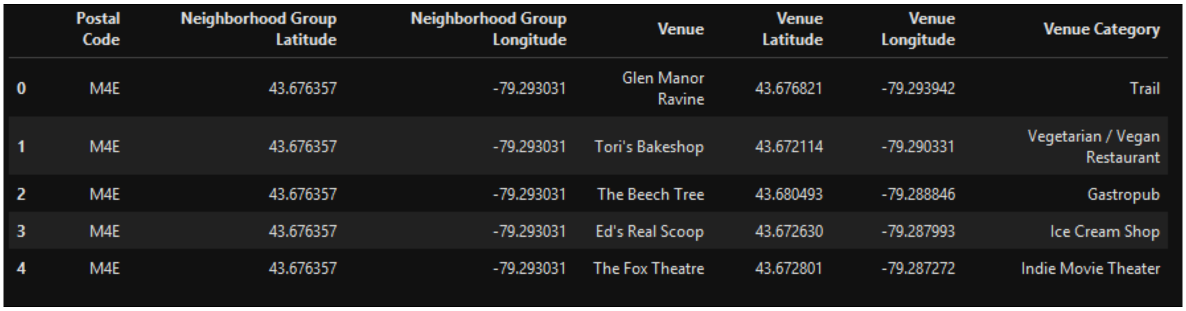
2.1 Data sources

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

Having collected the cities data, 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