Customizing Suggestions for Travelers

Luis Grotti

September 07, 2019

1. Introduction

1.1 Background

In 2018, touristic sector grew around 4% worldwide compared to last year. According to the latest WTTC survey on the importance of tourism this sector employs about 319 million people. The impact on the economy is tremendous, it contributes with around US$ 9 trillions to the world economy. Besides that, consumers are demanding more and more high level services today. Personalization plays a key role in this race to survive in the market. The strategy includes deliver an unique, quality experience from start to finish by delivering relevant products and services at every moment of your customer journey, generating greater brand identification. According to a recent survey published by Accenture Strategy, 44% of respondents in 33 countries prefer buying products from companies that offer a personalized experience.

1.2 Problem

Tourist sites are spread all over the world and each place has its own characteristic. Whether or not we choose to travel to a place based on our preferences: some travel to try different foods, enjoy nature or even play sports. The challenge is to identify what are the main characteristics of tourist cities and which cities most closely resemble them. By doing so we can choose the cities that best meet our expectations and thus have a fantastic experience.

1.3 Interest

Tour agencies and travel companies which are interested in providing personalized services by offering products that meet their customer preferences.

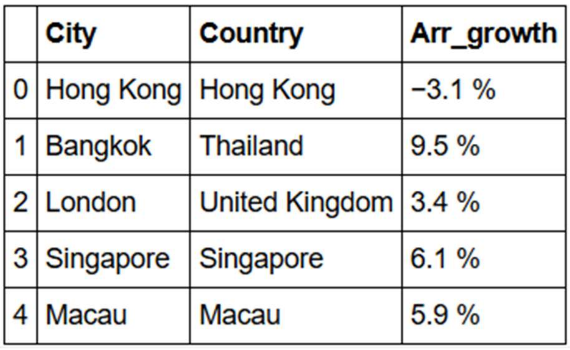
2. Data acquisition and cleaning

2.1 Data sources

For this project we will use the ranking provided by Euromonitor to analyze the top 100 most visited cities in the world in 2017. This information will be obtained by scraping a Wikipedia page. We will also use information from venues around each city center. This information will be obtained from Foursquare API by using the explore function. Finally, we will use a CSV file with the tourist classification of each venue category in Foursquase.

2.2 Data cleaning & Feature selection

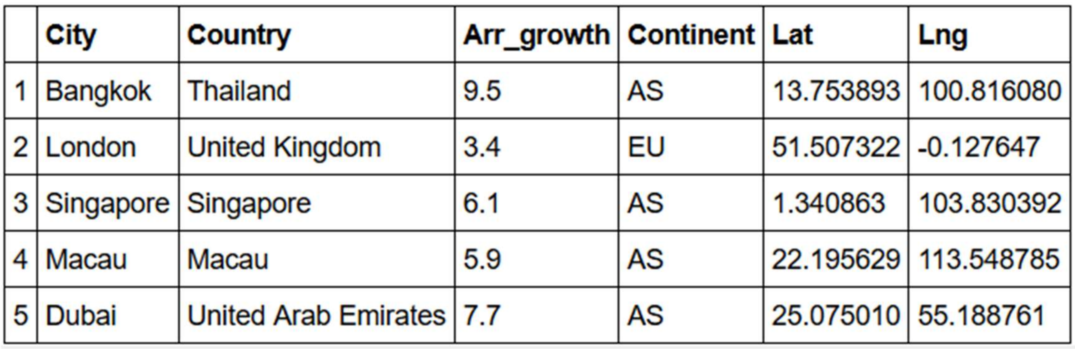
The list of the top 100 most visited cities in the world was obtained by scraping a page from Wikipedia. On this page we could find in a single table two different rankings, provided by Euromonitor and Mastercard. It was necessary to exclude Mastercard features as we will not use it for the analysis. Also, it was necessary to exclude all cities that belongs just to Mastercard ranking. After that process we got below dataframe.



	City	Country	Arr_growth
0	Hong Kong	Hong Kong	−3.1 %
1	Bangkok	Thailand	9.5 %
2	London	United Kingdom	3.4 %
3	Singapore	Singapore	6.1 %
4	Macau	Macau	5.9 %

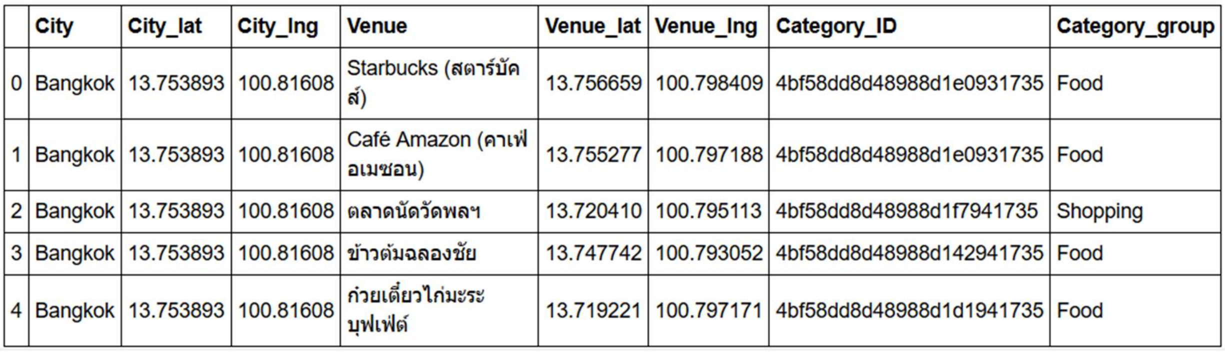
Table 1: Raw dataframe from Wikipedia

Arr_growth stands arrivals growth rate, and was imported as an object. We had to cast it to float. To use the Foursquare API we will need the location of each city, so we used Geopy library to get the latitude and longitude of center cities. We used pycountry_convert library to add continent of each city to help in our analysis. The next step was to use the Foursquare API to get the venues within 5km of city centers. Also, we labeled each venue based on tourism classification provided in CSV file. Finally, after cleaning and selection we got both dataframes:



	City	Country	Arr_growth	Continent	Lat	Lng
1	Bangkok	Thailand	9.5	AS	13.753893	100.816080
2	London	United Kingdom	3.4	EU	51.507322	-0.127647
3	Singapore	Singapore	6.1	AS	1.340863	103.830392
4	Macau	Macau	5.9	AS	22.195629	113.548785
5	Dubai	United Arab Emirates	7.7	AS	25.075010	55.188761

Table 2: Top Cities



	City	City_lat	City_lng	Venue	Venue_lat	Venue_lng	Category_ID	Category_group
0	Bangkok	13.753893	100.81608	Starbucks (สตาร์บัคส์)	13.756659	100.798409	4bf58dd8d48988d1e0931735	Food
1	Bangkok	13.753893	100.81608	Café Amazon (คาเฟ่อเมซอน)	13.755277	100.797188	4bf58dd8d48988d1e0931735	Food
2	Bangkok	13.753893	100.81608	ตลาดนัดวัดพลฯ	13.720410	100.795113	4bf58dd8d48988d1f7941735	Shopping
3	Bangkok	13.753893	100.81608	ข้าวต้มฉลองชัย	13.747742	100.793052	4bf58dd8d48988d142941735	Food
4	Bangkok	13.753893	100.81608	ก๋วยเตี๋ยวไก่มะระบุฟเฟ่ต์	13.719221	100.797171	4bf58dd8d48988d1d1941735	Food

Table 3: Venues

3. Exploratory Data Analysis

First thing to do was to check which cities have arrivals growth rates below zero. We don't want in our algorithm cities that lost tourists when compared with previous year. We found 12 cities in this situation and we dropped them from our dataframe. At this point 50% of cities in our dataframe are located in Asia, so we decided to check for possible outliers in this region. We plotted arrival growth rate in a boxplot. Five cities in Asia were considered discrepancies and were dropped of our analysis.

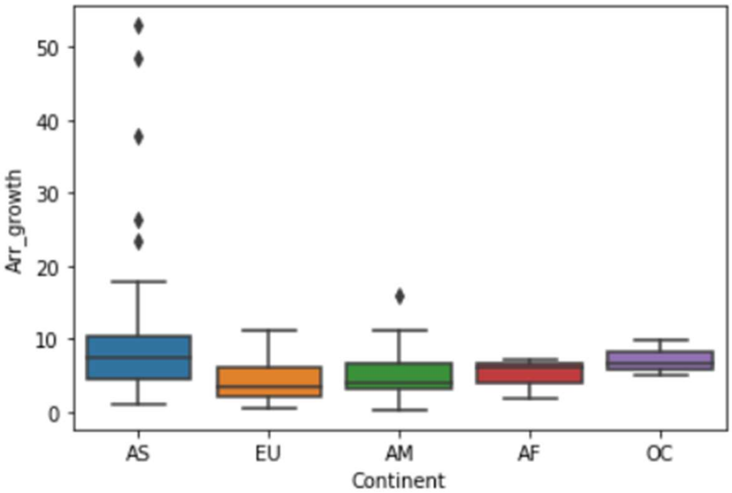


Figure 1: arr_growth outiliers in Asia

We also checked which cities have a reasonable number of venues. The more places we can analyze, the more accurate our analysis will be. We draw a histogram to show the distribution of quantity of venues. We found three cities with less than 30 venues available in our dataframe. This number does not seem to be enough to understand the profile of the entire city. We then decided to remove them from our analysis and the list of cities now have 80 cities.

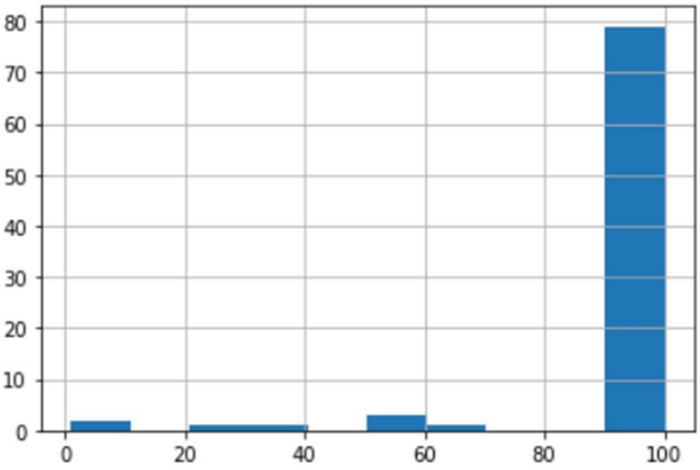


Figure 2: histogram of number of venues per city

4. K-Means Clustering Algorithm

The k-means algorithm is an unsupervised clustering algorithm, it means inputs have no external classification. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

4.1 Finding best value for K

We used Elbow Method to find best value for K. The idea of Elbow Method is that we want a small error, but that error tends to decrease toward 0 as we increase K. So our goal is to choose a small value of k that still has a low error, and the elbow of chart usually represents where we start to have diminishing returns by increasing K. We ran K-Means over our dataset having K between 1 and 14. We calculated sum of squared distance for each value of K and then plotted it.

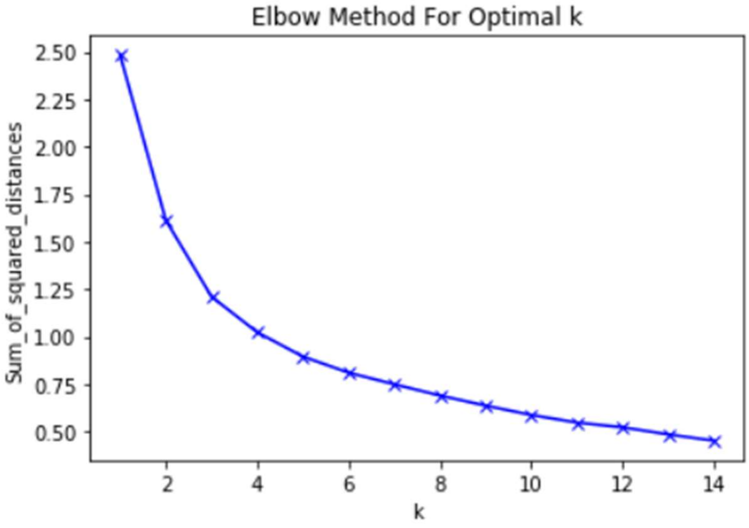


Figure 3: Sum of squared distances for k between 1 and 14

We can see from plot that erros start diminishing its decreasing when k=3.

4.2 Running and interpreting K-Means Algorithms

K-Means algorithm from SKLearning library with K=3 was used to cluster our dataset. Concerning number of cities by cluster, Cluster 1 holds 50% of cities while Clusters 0 and 2 holds 25% each one.

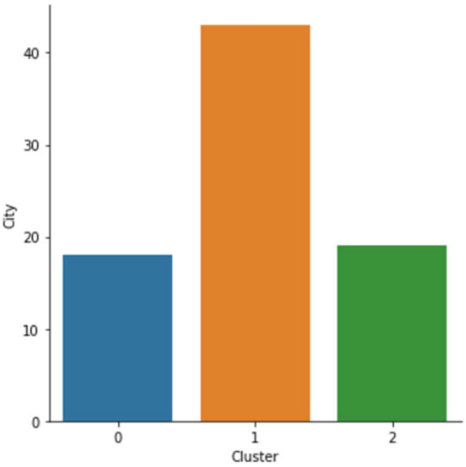


Figure 4: Cities per cluster

Checking continents of cities in each cluster we could find that Cluster 1 is composed by cities in all continents. Cluster 0 is composed by cities in Asia and Europe while Cluster 2 is mainly composed by Asian cities.

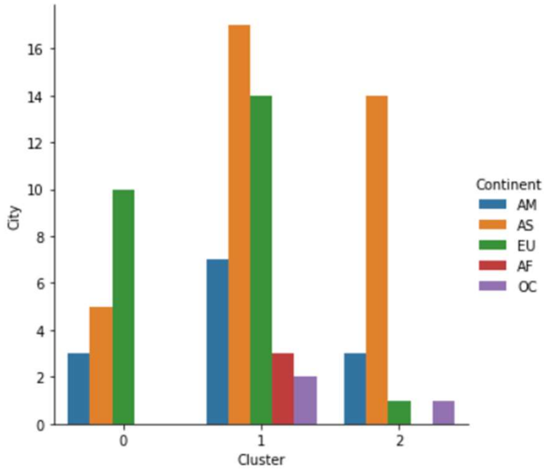


Figure 5: Continent by cluster

We could go deeper inside each cluster by plotting the frequence of tourism categories in a boxplot.

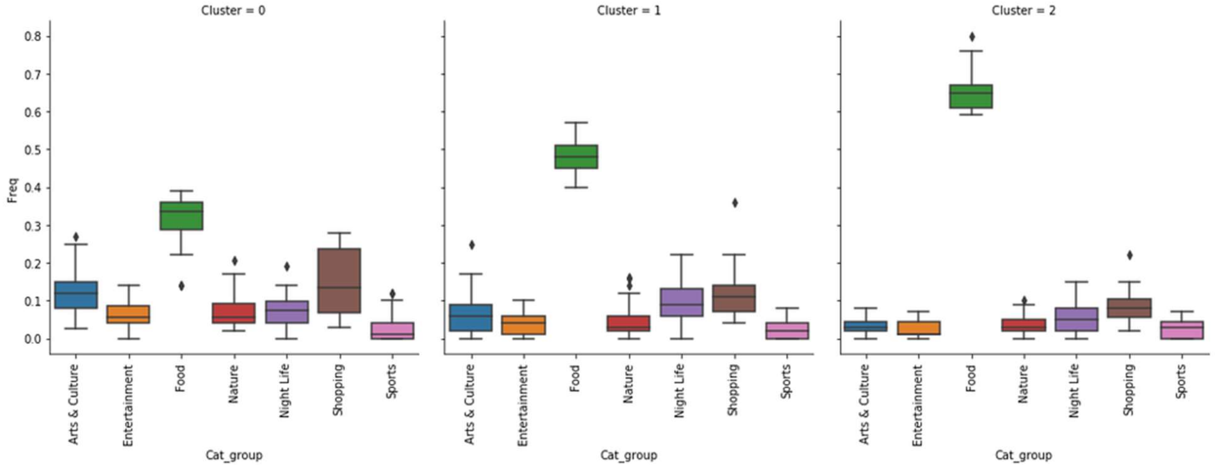


Figure 6: frequence of venues by tourism category

We can conclude that **Cluster 0** holds best cities when visitors want to have a taste of each kind of experience. All kind of experiences are provided by those cities venues in a good reasonable proportion. In addiction, if visitors want to shopping, enjoy art and culture or see the nature definitely Cluster 0 should be their destiny. **Cluster** 1 is the largest cluster. It means its cities can be likely found near the visitors, what can make trips cheapest . Cities in this cluster offer a huge number of places to enjoy food. Places to enjoy night life and shopping can be found easily as well. **Cluster** 2 is definitely recommended to visitor who prefer travel to enjoy food. Places to eat can be found everywhere. Despite that, we can't find a good number of places to enjoy other kind of experiences. Art & Culture and Nature options are the Achilles's heel of this Cluster.

In general, places to have fun and enjoy or do sports are similarly spread by clusters. Finally, lets label each Cluster based on our analysis and print its cities.

5. Conclusions

Our analysis shows that majority of most visited cities are located in Europe and Asia. Some of those cities has not enough information about venues in Foursquare. It doesn't mean that those cities have a small number of options to offer to their visitors. We focused on cities with at least thirty places registered in Foursquare near their city centers. After directing our attention to this more narrow area, we choose to ignore cities in Asia with discrepant arrivals growth rates. It can influence the analysis and thus the interpretation of the data. We found best number of clusters by analyzing squared distances when running K-Mean Clustering Algorithm for a range of values for K.

Clusters got from K-Mean Algorithm shows that we can split cities in three main clusters. First one holds cities indicated to visitors who wants to experience all kind of experiences, like shopping and enjoy nature and arts. Cities in this cluster are mainly located in Europe and Asia. Second cluster is indicated to visitors who wants to spend less money to reach their final destination, enjoy food or experience a intense night life. Third cluster is heavily recommended to visitors who are looking for cities to enjoy food. Most of cities in this cluster have more than 65% of their places focused on food tourism.

In general, places to have fun and enjoy or do sports figure similarly in all clusters.

The purpose of this project was to identify similarity between most visited cities to aid stakeholders make customized suggestions. By clustering most visited cities in the world we have created groups of similar cities based on their venues categories. Those groups will help stakeholders to make customized suggestions based on their client's preferences.

6. Future directions

Algorithm applied to this Project could be inproved by checking all venues inside city limits. Unfortunatelly Foursquare does not provide such service. Also, this Project could be extended to small cities and even neighborhoods for instance. By clustering neighborhoods we can understand where is the best places to live, work and have fun.