CLUSTERING POPULAR TRAVEL DESTINATIONS BASED ON THEIR LEISURE ACTIVITIES

Isabelle Morose

January 28, 2020

Table of Contents

1. Introduction

There are many reasons why people travel. For instance, while some people see traveling as a way to make new friends, others might be more interested in the kind of leisure activities they'll partake in. If you're part of the latter group, you will probably want to do some research before you choose which city to visit next. However, that process can be very time-consuming. The study presented in this report was undertaken to help people choose travel destinations in a more efficient way. Indeed, in this study, 140 popular travel destinations were analyzed and clustered into distinct groups based on the similarity of the activities they offer.

At the end of this report, you will be able to:

i. Choose an interest, then pick a city offering activities related to it.
ii. Choose a city that you like, then pick another city offering the same kind of experience.

2. Data

2.1 Data Acquisition: From Web Scraping to Foursquare API

Three sets of data were used in the study that was just introduced:

2.1a A list of most visited cities in the world

Both Euromonitor and Mastercard offer rankings of top 100 cities based on the number of international visitors. A compilation of both lists can be found

on Wikipedia[1], from which the data was scraped using BeautifulSoup. The resulting dataframe contained 8 columns and 140 rows, each row representing a different city.

2.1b Foursquare venue information for each city

After having obtained geographical coordinates for each city using the Nominatim geocoding service, information about 140 random venues was collected for each city through Foursquare. Using the "explore" endpoint returned many response fields, like the name of each venue, its ID, its address, and which category it belongs to.

2.1c A hierarchical list of all Foursquare venue categories

All Foursquare venue categories fall under one of 10 main categories:

1. Arts & Entertainment
2. College & University
3. Event
4. Food
5. Nightlife Spot
6. Outdoors & Recreation
7. Professional & Other Places
8. Residence
9. Shop & Service
10. Travel & Transport

Using the Foursquare "categories" endpoint returned a list of these 10 categories, along with other information about each category, including up to 4 levels of

[1] https://en.wikipedia.org/wiki/List_of_cities_by_international_visitors

subcategories.

2.2 Data Cleaning

2.2a Storing Venue Information into Dataframe

Cleaning the List of Most Visited Cities

The dataframe containing information about the most visited cities in the world had 6 irrelevant columns, which were all dropped. The remaining 2 columns ("City" and "Country") were combined into one column, resulting in a one-column dataframe with 140 data points.

Adding Geographical Coordinates

Two more columns, "Latitude" and "Longitude", were added to the dataframe after having obtained the geographical coordinates of each city with Geopy/Nominatim.

Extracting Venue Categories from Foursquare API Response

Of the response fields returned by the Foursquare "explore" endpoint, only the "categories" field was of interest for the purpose of this analysis. The "categories" field itself is an array within which the actual category is contained. The category of each venue was extracted from this array and stored into a new column of the dataframe. In this new column, each city/row had an associated list of 140 venues, each venue represented by its category (for example, 'art museum' or 'coffee shop').

2.2b Storing Foursquare Categories Information into a Second Dataframe

Cleaning the Hierarchical List of Foursquare Categories

The 10 main Foursquare categories and all their corresponding subcategories (of all levels) were extracted from the "categories" endpoint response. The results were stored into a new 10-row dataframe with 2 columns: one for the name of the main category and the other with a list of corresponding subcategories.

Of the 10 main categories, only the ones related to tourist activities were of interest. Consequently, the rows corresponding to "College & University", "Professional & Other Places", and "Residence" were dropped. "Travel & Transport" was also dropped because it relates to travel services and not activities. Subcategories related to shopping were extracted from "Shop & Service" and stored into a new row, while "Shop & Service" was dropped.

2.2c Combining the Two Dataframes Created

The final dataframe was built by combining information from the two dataframes already created. The dataframe containing Foursquare categories was used to count how many of the 140 venues collected for each city fell under each main category. The category “Event” was dropped because it had no data for none of the cities.

The final dataframe contained the following information for each city (see *Table 2-1)*:

- Ratio of tourist activities to total number of venues (column “Activities)
- Ratio of tourist activities by category to total number of activities (column denoted by name of activity category, e.g. “Shopping”)

Table 2—1 First 5 rows of final dataframe

	City	Activities	Arts & Ent	Food	Nightlife Spot	Outdoors & Recreation	Shopping
0	Hong Kong, Hong Kong, China	0.350000	0.061224	0.489796	0.122449	0.163265	0.163265
1	Bangkok, Thailand	0.150000	0.000000	0.380952	0.000000	0.190476	0.428571
2	London, United Kingdom	0.392857	0.145455	0.490909	0.127273	0.090909	0.145455
3	Macau, Macau	0.121429	0.058824	0.529412	0.117647	0.176471	0.117647
4	Singapore, Singapore	0.292857	0.073171	0.560976	0.048780	0.268293	0.048780

3. Methodology

Due to the low dimensionality of the data, k-means clustering was chosen as an appropriate method of clustering.

3.1 Choosing an optimal number of clusters (k)

Elbow Method

The elbow method aims to determine the k value that occurs at the point of inflection on the k-means error curve. That point represents the optimal k value. The Within Cluster Sum of Squares (WCSS), defined by the formula below, is often chosen as the error metric.

Equation 3-1 - Within-Cluster Sum-of-Squares Formula

$$\mathrm{WCSS}(k) = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \text{cluster } j} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|^2,$$

where $\bar{\mathbf{x}}_j$ is the sample mean in cluster j

Using the data from the final dataframe described in the previous section, the WCSS was plotted for k values ranging from 0 to 10. The point of inflection on the curve was ambiguous. However, it seemed to be either 2 or 3 (see *Figure 3-1)*.

Table 3—1 First 5 rows of the dataframe used for clustering

	Activities	Arts & Ent	Food	Nightlife Spot	Outdoors & Recreation	Shopping
0	0.350000	0.061224	0.489796	0.122449	0.163265	0.163265
1	0.150000	0.000000	0.380952	0.000000	0.190476	0.428571
2	0.392857	0.145455	0.490909	0.127273	0.090909	0.145455
3	0.121429	0.058824	0.529412	0.117647	0.176471	0.117647
4	0.292857	0.073171	0.560976	0.048780	0.268293	0.048780

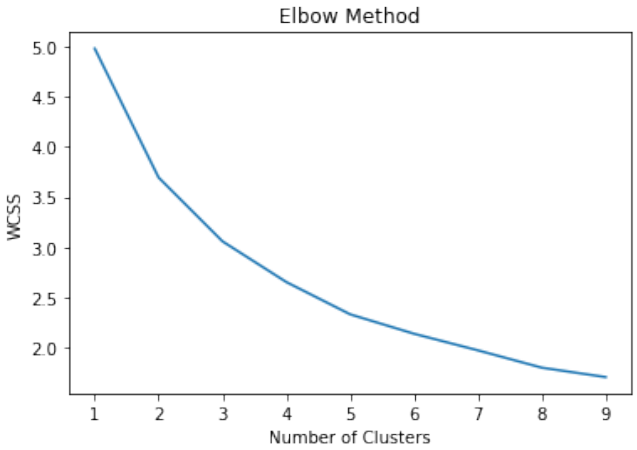


Figure 3-1 The point of inflection is not clear, but it seems to be 2 or 3.

Silhouette Method

Since the elbow method had inconclusive results, the silhouette method was subsequently applied to the same dataset. The silhouette score is a measure of how similar a data point is to data points within its own cluster and how dissimilar it is to data points in other clusters. The silhouette method consists in determining the k value that maximizes the silhouette score. On a plot of silhouette score vs k value, the maximal score occurred at k=3.

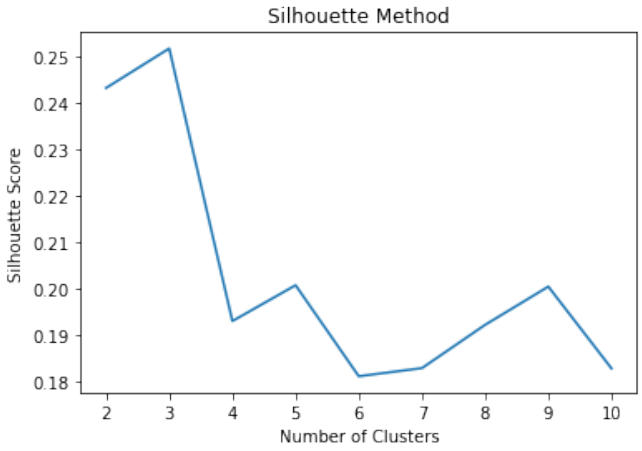


Figure 3-2 The maximum silhouette score occurs at k=3, suggesting that this is the optimal k value.

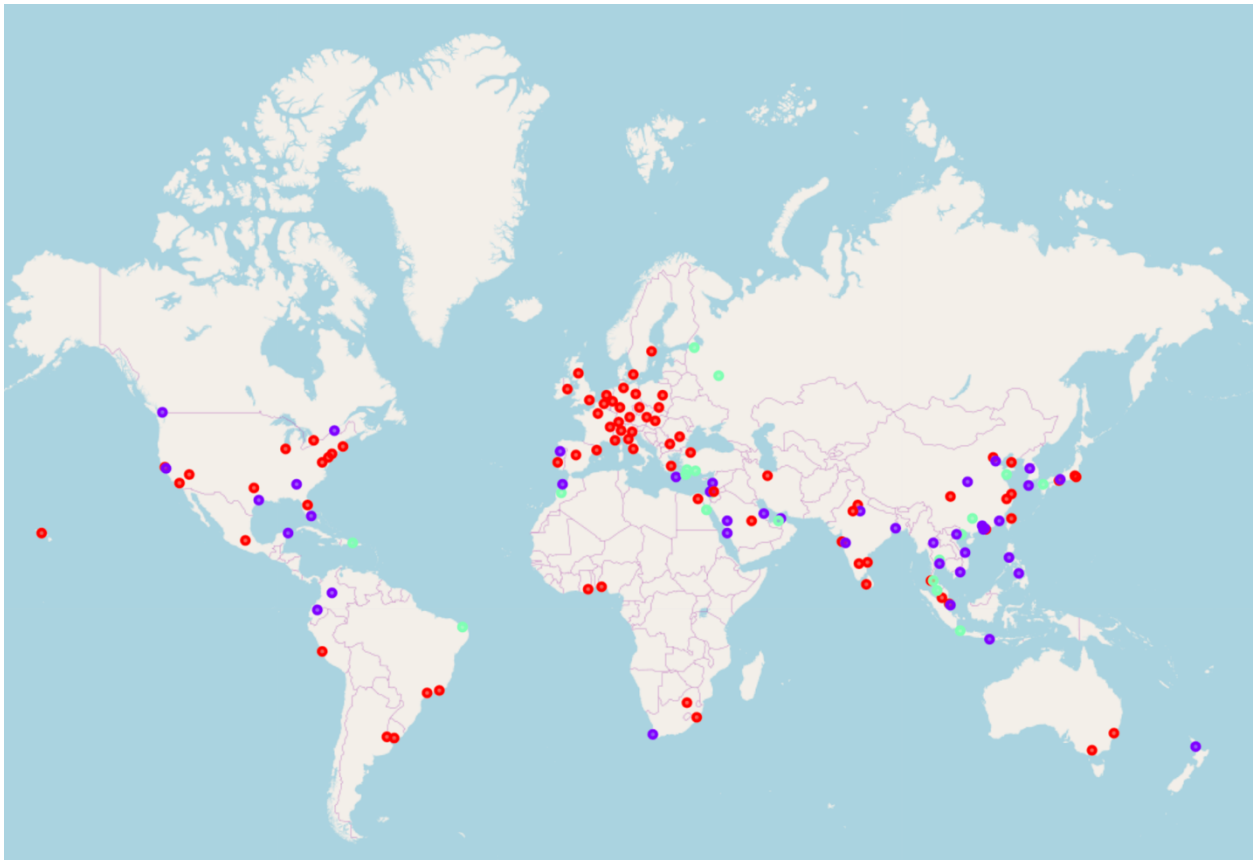
3.2 Clustering with K-Means

The k-means algorithm was applied to the entire dataset using 3 as the optimal number of clusters, as determined by the silhouette method in conjunction with the elbow method. Each data point was assigned to one and only one of the 3 clusters.

4. Results

The k-means algorithm returned 3 distinct clusters of different sizes.

Figure 4-1 Geographical Representation of Clusters



- The largest cluster (C1) has 79 countries
- The second largest (C2) cluster has 41 countries
- The smallest cluster (C3) has 20 countries

Table 4—1 Cities in the Largest Cluster (C1)

Hong Kong, Hong Kong/China	Chennai, India	Lisbon, Portugal	Düsseldorf, Germany
London, United Kingdom	Vienna, Austria	Copenhagen, Denmark	Boston, United States
Paris, France	Johor Bahru, Malaysia	San Francisco, United States	Chengdu, China
New York City, United States	Jaipur, India	Melbourne, Australia	Edinburgh, United Kingdom
Kuala Lumpur, Malaysia	Berlin, Germany	Warsaw, Poland	Tehran, Iran
Istanbul, Turkey	Cairo, Egypt	Honolulu, United States	Hamburg, Germany
Delhi, India	Athens, Greece	Kraków, Poland	Geneva, Switzerland
Mumbai, India	Orlando, United States	Buenos Aires, Argentina	Colombo, Sri Lanka
Phuket, Thailand	Venice, Italy	Chiba, Japan	Bucharest, Romania
Rome, Italy	Madrid, Spain	Frankfurt am Main, Germany	Sofia, Bulgaria
Tokyo, Japan	Riyadh, Saudi Arabia	Stockholm, Sweden	Dalian, China
Taipei, Taiwan	Dublin, Ireland	Lima, Peru	Montevideo, Uruguay
Prague, Czech Republic	Florence, Italy	Nice, France	Amman, Jordan
Amsterdam, Netherlands	Toronto, Canada	Rio de Janeiro, Brazil	Hangzhou, China
Osaka, Japan	Johannesburg, South Africa	Bangalore, India	Durban, South Africa
Las Vegas, United States	Sydney, Australia	Mexico City, Mexico	Dallas, United States
Shanghai, China	Munich, Germany	São Paulo, Brazil	Accra, Ghana
Barcelona, Spain	Beijing, China	Zürich, Switzerland	Philadelphia, United States
Los Angeles, United States	Brussels, Belgium	Washington D.C., United States	Lagos, Nigeria
Milan, Italy	Budapest, Hungary	Chicago, United States	

Table 4—2 Mean of Activity Ratios by Cluster

	Entire Dataset	Cluster C1	Cluster C2	Cluster C3
Activities	0.297	0.340	0.256	0.213
Arts & Entertainment	0.108	0.131	0.080	0.074
Food	0.513	0.466	0.633	0.450
Nightlife Spot	0.092	0.102	0.081	0.076
Outdoors & Recreation	0.116	0.101	0.078	0.253
Shopping	0.172	0.200	0.128	0.146

Table 4—3 Cities in the Second Largest Cluster (C2)

Dubai, United Arab Emirates	Cancún, Mexico	Tel Aviv, Israel	Xi'an, China
Shenzhen, China	Hanoi, Vietnam	Da Nang, Vietnam	Beirut, Lebanon
Pattaya, Thailand	Dammam, Saudi Arabia	Batam, Indonesia	Xiamen, China
Mecca, Saudi Arabia	Heraklion, Greece	Jeju, South Korea	Casablanca, Morocco
Guangzhou, China	Kyoto, Japan	Porto, Portugal	Atlanta, United States
Medina, Saudi Arabia	Zhuhai, China	Montreal, Canada	Pune, India
Seoul, South Korea	Vancouver, Canada	San Jose, United States	Quito, Ecuador
Agra, India	Chiang Mai, Thailand	Houston, United States	Tianjin, China
Miami, United States	Kolkata, India	Cape Town, South Africa	
Ho Chi Minh City, Vietnam	Cebu City, Philippines	Manila, Philippines	
Denpasar, Indonesia	Auckland, New Zealand	Bogota, Colombia	

Table 4—4 Cities in the Smallest Cluster (C3)

Bangkok, Thailand	Ha Long, Vietnam	Marrakesh, Morocco	Abu Dhabi, United Arab Emirates
Macau, Macau	Jakarta, Indonesia	Guilin, China	Rhodes, Greece
Singapore, Singapore	Saint Petersburg, Russia	Hurghada, Egypt	Krabi, Thailand
Antalya, Turkey	Jerusalem, Israel	Muğla, Turkey	Punta Cana, Dominican Republic
Moscow, Russia	Penang Island, Malaysia	Fukuoka, Japan	Qingdao, China

5. Discussion

5.1 Geographic Spread of Clusters

All 3 clusters are geographically spread all over the world. All the European cities, save for one, are in the largest cluster (C1), whereas there is more diversity among cities in the Americas and in Asia. The most prominent cluster on the American continent is C1, and most Asian cities belong to the second largest cluster (C2).

5.2 Most Prominent Features of Clusters

When comparing the means of activity ratios for each cluster to the means of activity ratios for the entire dataset, some features stand out:

- Cities in cluster C1 have more activities than average, although these activities are mostly in the Shopping and Arts & Entertainment categories
- Cities in cluster C2 have more activities related to food than average
- Cities in cluster C3 have more outdoors activities than average

Figure 5-1 Means of Activity Ratios for Cluster C1

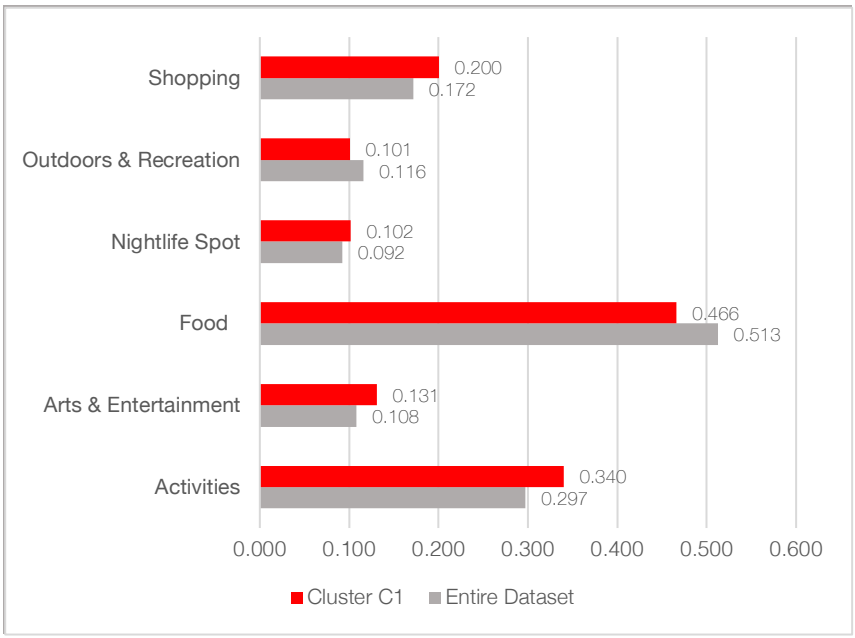


Figure 5-2 Means of Activity Ratios for Cluster C2

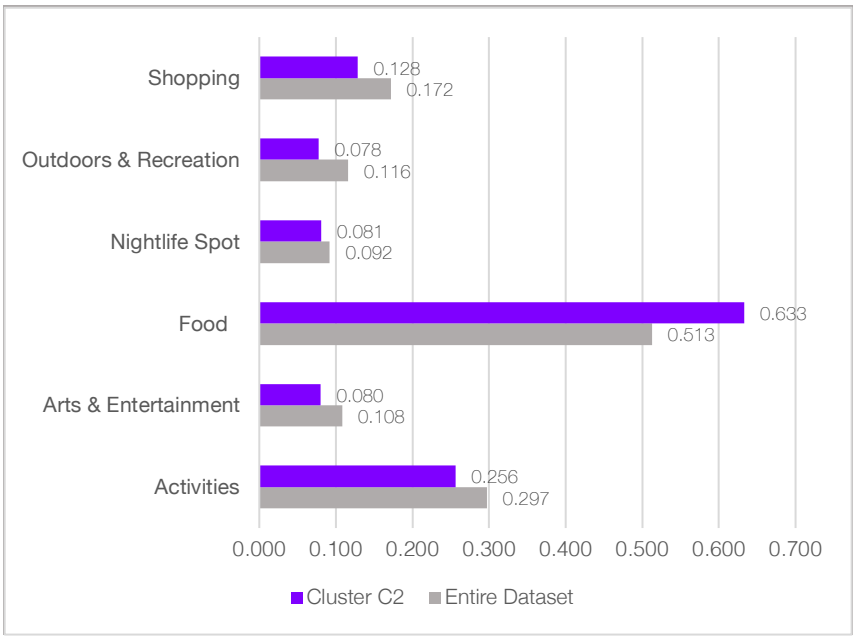


Figure 5-3 Means of Activity Ratios for Cluster C3

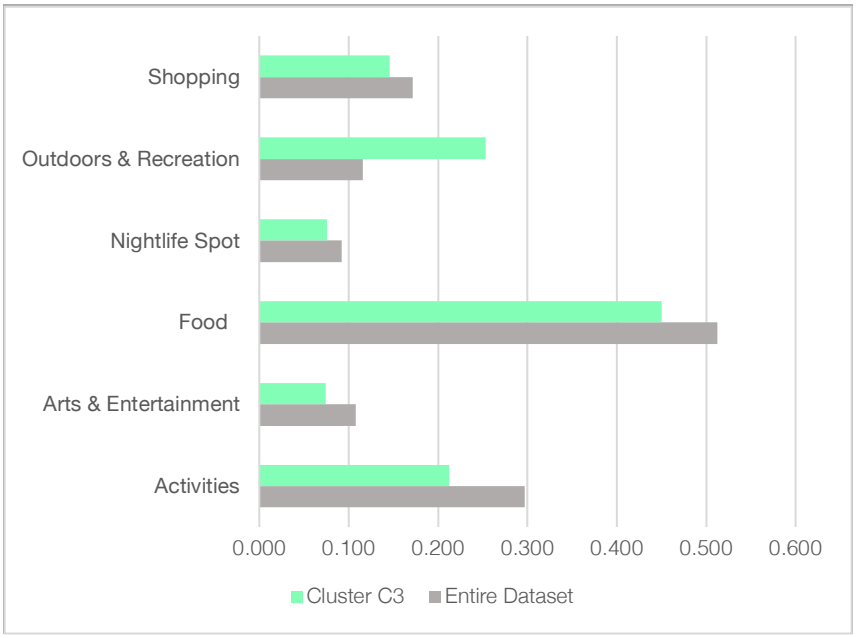


Table 5—1 Travel Recommendations

For activities related to (pick category below)	**You should visit cities in**	**But not cities in**
Arts & Entertainment	Any cluster	N/A
Food	C2	C1 nor C3
Nightlife Spot	Any cluster	N/A
Outdoors & Recreation	C3	C2
Shopping	C1 or C3	C2

6. Conclusion

After analyzing 140 popular travel destinations, it was found that the most optimal way to group them based on the similarity of their leisure activities was to cluster them into only 3 groups. Undoubtedly, other factors like architecture, climate and walkability, also affect travelers' experience. Had those factors been considered, the clustering results could have been much different. Nonetheless, the study presented in this report suggests that the most popular cities are not so different from each other and that perhaps the nature of the activities they offer plays into their popularity.