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