San Francisco as a Travel Destination

Farah Darwich

February 21, 2019

2. Data Acquisition and Cleaning

2.1 Data Sources

In order to perform the evaluation, I will need as much metadata as possible about as many venues as possible in San Francisco.

Foursquare is a social location-based service that contains all the information we need. The Foursquare API allows application developers to interact with the Foursquare platform, which is how I obtained the dataset I worked with.

In this project I relied on the Foursquare scraper algorithm built by Data Scientist Max Woolf, and was able to obtain a dataset that contains basic metadata on 44425 venues in San Francisco including their categories, ratings, coordinates, and prices.

2.2 Data Cleaning

After acquiring the data from Foursquare, I saved it into a pandas dataframe so I can further examine and analyze it. The dataframe contained information on 44425 venues in San Francisco with their names, categories, coordinate, number of check ins, number of likes, ratings, prices, number of ratings, venue URLs, and Foursquare URLs.

I didn't need many of these columns in my work, and since I was going to base my evaluation of venue categories on availability, rating, and price, I decided to split the dataframe into two new dataframes: One to be used in evaluating venues based on price, and the other to be used in evaluating venues based on rating.

As for the rating dataframe, I deleted all columns except for the ones that contain names, categories and ratings, and then dropped all rows (venues) that had a missing value in one of those three columns.

Table 1: The first five rows of the ratings dataframe

	name	categories	rating
4	Madusalon	Salon / Barbershop	7.4
10	Mr. Smith's	Bar, Lounge, Nightclub	6.7
18	Hillside Supper Club	American Restaurant	8.2
25	Hand Touch Nails	Spa	7.7
38	Southern Pacific Brewing	Brewery, American Restaurant, Burger Joint	8.7

As for the price dataframe, I deleted all columns except for the ones that contain names, categories and prices, and then dropped all rows (venues) that had a missing value in one of those three columns.

Table 2: The first five rows of the prices dataframe

	name	categories	price
8	Lady Falcon Coffee Club	Food Truck, Coffee Shop	1.0
10	Mr. Smith's	Bar, Lounge, Nightclub	2.0
11	Madame Kim's Annex	Speakeasy	3.0
18	Hillside Supper Club	American Restaurant	2.0
21	The Treehouse @ Public Works	Lounge	3.0

In these final two dataframes that I used for the analysis, I had 8167 venues with available ratings and 8365 venues with available prices.

To recap, this evaluation was be built on the following aspects:

- Shopping Options (mainly, Shopping Malls and Shopping Plazas): Availability - Rating.
- Tourist Spots (mainly, Museums, Music Venues, and Theme Parks): Availability - Rating - Affordability.
- Accomadation Options (mainly, Hotels, Hostels, and Bead&Breakfast's): Availability - Rating - Affordability.

Keep in mind that each of those categories contain its own subcategories that will automatically be included in the analysis.

The ‘Museum’ category includes 5 sub-categories, the ‘Music Venue’ category includes 3 sub-categories, the ‘Theme Park’ category includes 1 sub-categories, and the ‘Hotel’ category includes 10 sub-categories.

Overall, 27 different categories of venues were included in this evaluation.