Predicting the best Airbnb place that suits travelers needs.

Jesus Pereyra
Marzo 9, 2020

1. Introduction

1.1 Background:

Airbnb is one of the most popular online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events. It acts as a broker, receiving commissions from each booking. Users can specify all the important features of the room or the entire place, like numbers of rooms, type of kitchen and the neighborhood that they want to stay in.

1.2 Problem:

There are not enough features that can suit a travel needs besides the room preference. Features like nearby restaurants, museums or gallery art can change which Airbnb you want to stay. Features that can let you choose what you want to have in the neighborhood can impact the traveler in a positive or negative way depending on the reason of the trip.

1.3 Interest:

Businessmen and adventurers can experience a new way of searching airbnb for their trips. A business man that travels for a few weeks may need a gym, a coffee place or a remote working place. However for a pleasure trip you might don't have any of those needs, but maybe you will want to be close to an Italian restaurant or a Mexican restaurant to experiment the difference between there and home, like taste and special dishes. Also if you like museums and parks, the best thing you can do is to select a place that is close to the maximum number of those particular venues, to visit as many as possible.

This problem was solved using data location from Airbnb and Foursquare, then merge it, to have availables airbnb in each neighborhood with their top venues using clustering models and Foursquare api. The user will input the top venues in the city of your plan (for this project will only have NYC), and our model will present the top 5 neighborhoods that suit more their needs with different airbnb order by popularity or price.

2. Data:

2.1 Data Recollection

The data is fetched from two sites that work with geolocation and then is merge both into a single dataset.

Airbnb NYC Dataset is available in csv format from http://insideairbnb.com/get-the-data.html. In the table 1.0 above we can see each feature of an airbnb. We miss some important information about the room that can help us decide which rooms to return by our model. We tried to get the

number of bedrooms, amount of guesses but it is not available in the dataset and can't be obtained from the page source because the HTML is encoded.

```
df_data = pd.read_csv("./AB_NYC_2019.csv")
df_data.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	365
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	194
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	0

Table 1.0 Airbnb data representation.

We create the Foursquare dataset using the airbnb locations and the Foursquare api to fetch the top venue. The endpoint to get the top venues is *https://api.foursquare.com/v2/venues/explore* with params like latitude, longitude, radius of search and limit amount of venues. An example of a complete request would be like *https://api.foursquare.com/v2/venues/explore?&client_id={xxxx}&client_secret={xxxx}&v={2}&ll={40.7128°},{74.0060}&radius={300}&limit={50}*. The response is in json format and was transformed to a table that looks like image 2.1.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Allerton	40.864819	-73.85917	Domenick's Pizzeria	40.865576	-73.858124	Pizza Place
1	Allerton	40.864819	-73.85917	Sal & Doms Bakery	40.865377	-73.855236	Dessert Shop
2	Allerton	40.864819	-73.85917	White Castle	40.866065	-73.862307	Fast Food Restaurant
3	Allerton	40.864819	-73.85917	Bronx Martial Arts Academy	40.865721	-73.857529	Martial Arts Dojo
4	Allerton	40.864819	-73.85917	Dunkin'	40.865204	-73.859007	Donut Shop

Table 1.1 Foursquare data representation.

2.2 Data Cleaning

Airbnb dataset has 11 housestays without price, we decided to remove those houses because can cause noise through the training of the model. Also for safety of our users we decided to not recommend houses without review.

Some houses are not found on Airbnb site, to solve this problem, we decided to drop each column that returns a 404 code or any code besides 200 in the website.

2.3 Data processed

After we merge both dataset we will have a final dataset with the following columns name (airbnb name, price, price category, clustering category, neighborhood, latitude, longitude and review per month, top 15 venues).