"The best Airbnb entire home/apartment in Sheepshead Bay, Brooklyn - NYC"

Matteo Vadi – December 28, 2020

1. INTRODUCTION

1.1 BACKGROUND

All of us have to look for a place to stay for a longer or shorter time period, at least once in our lives. And, in those situations, sooner or later we have to face with the choice between two or more alternatives available, without knowing which to choose having little information about the place.
This work accepts the challenge and tries to give an answer to the following question: which of these apartments is better to choose? This document specifically focuses on rental entire homes / apartments in the Airbnb platform for the neighborhood of Sheepshead Bay within the borough of Brooklyn, New York city, but can be extended to other neighborhoods, boroughs, cities and/or rental properties.
Airbnb, Inc. is an American vacation rental online marketplace company based in San Francisco, California US. It allows to host and rent different properties, accessible to consumers on its website or via an App. Through the service, users can arrange lodging, primarily homestays, and tourism experiences or list their properties for rental.

1.2 PROBLEM

This project aims to determine which entire home / apartment in the Airbnb platform is the best one to choose when deciding for a place to stay in Sheepshead Bay, Brooklyn NYC, in terms of venues and points of interest you can find in the nearby area. In particular, I have decided to focus only on the most important types of venues for an accurate decision in my personal opinion, within a radius of 600 meters (an appropriate distance that can be covered by foot) from the place, that you can find listed below:

- Travel & Transport:
    - Metro station
    - Tram station
    - Bus station
    - Taxi
    - Taxi stand
    - Rental car location
- Food
- Professional & Other spaces:
    - Parking
    - Medical Center:
        - Emergency room
        - Hospital

- Shop & Services:
    - Market
    - ATM
    - Laundromat
    - Dry cleaner
    - Internet Café
    - Auto Garage
    - Food & Drink Shop
    - Shopping Mall
    - Shopping Plaza
    - Pharmacy

These venues categories and sub-categories refer to the differentiation you can find in the Foursquare Developers webpage (https://developer.foursquare.com/docs/build-with-foursquare/categories/).

The list introduced above is considering the important venues in choosing a place to stay whatever the reason of the journey is, in case of shorter time period for tourism visits or longer ones for work transfers. Indeed, this work does not consider Arts & Entertainment, College & Universities, Events or Outdoors & Recreation venues among others, since those depend on the specific purpose of the journey.

This project deals with the Foursquare APIs to explore the geographical location of NYC and to cluster venues in Brooklyn borough (K-means clustering algorithm), in order to find the most suitable home / apartment to choose among all. In fact, the apartment which present the higher number of these kinds of venues in the nearby area, will be the best one according to this project work. This information can be used then in choosing the apartment, together with other common aspects, such as availability, price and minimum overnight stays.

1.3 INTEREST

The interest that such a project may arouse could encompass several stakeholders. First of all, the final user of the Airbnb platform for sure or, generally speaking, all the people that travel with more or less frequency, allowing them to have a different evaluation tool from the common reviews/feedbacks on websites.

Similarly, this work could also arouse interest in those individuals who are on the other side and who are thinking of buying an apartment for renting it, allowing them to find the best area of a city (a specific neighborhood of New York City in this case) where to buy the property.

2. DATASET

2.1 DATA SOURCES

The original dataset describes the listing activity and metrics in NYC for 2019. You can find it to the Kaggle website (https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data); in particular, it is part of Airbnb's open data and the original source can be found on its website (http://insideairbnb.com/). The dataset includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions. For the aim of this project work, it will focus only on the division between neighborhood and neighborhood_groups (actually, the NYC's boroughs) for the Foursquare API identification.

2.2 FEATURES SELECTION

Figure 1. The original dataframe with the head method.

The df dataframe shows 48895 records (rows) and 16 attributes (columns)

	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

After a first quick view, the dataset imported in the form of csv document through the correspondent pandas library (pd.DataFrame.read_csv), shows 48895 records (rows) and 16 attributes (columns). Among those, we select only ones that concern the rental of entire houses / apartments, not considering private rooms or shared rooms, but the type of analysis that will be carried out could also be extended to the latter.
For the purpose of our problem, we do not take into account the attributes about the id, host_name, host_id, price, minimum_nights, number_of_reviews, last_reviews, reviews_per_month, calculated_host_listings_count and availability_365. This types of attributes could be used in a supervised Data Mining problem (for example, if the goal was the prediction of a rental place's price), while for the aim of our clustering approach they would only be redundant. For this reason we consider the following attributes:

- **name**: the name of the entire home / apartment;
- **neighbourhood_group:** actually, the borough's name the entire home/apartment belongs to;
- **neighbourhood**: the neighborhood's name the entire home/apartment belongs to;
- **latitude**: the latitude coordinate for the entire home / apartment;
- **longitude**: the longitude coordinate for the entire home / apartment.

To have a better comprehension, the ***df*** dataset has been renamed in ***airbnb***, changing also 2 columns names. The dataset, after these changes appears as in Figure 2.

Figure 2. The Airbnb dataset after features selection with the head method.

```
The airbnb dataframe now shows 25409 records (rows) and 5 attributes (columns)
```

	name	borough	neighborhood	latitude	longitude
1	Skylit Midtown Castle	Manhattan	Midtown	40.75362	-73.98377
3	Cozy Entire Floor of Brownstone	Brooklyn	Clinton Hill	40.68514	-73.95976
4	Entire Apt: Spacious Studio/Loft by central park	Manhattan	East Harlem	40.79851	-73.94399
5	Large Cozy 1 BR Apartment In Midtown East	Manhattan	Murray Hill	40.74767	-73.97500
9	Cute & Cozy Lower East Side 1 bdrm	Manhattan	Chinatown	40.71344	-73.99037

2.3 DATA CLEANING

After the feature selection, the *airbnb* dataset shows 25409 rows and 5 columns. This section is partially aimed to deal with missing values in the dataset. After a quick view (Figure 3), it is possible to see that there are very few of them and all connected to the *name* attribute.

Figure 3. Missing values in the Airbnb dataset.

```
Out[7]: name            7
        borough         0
        neighborhood    0
        latitude        0
        longitude       0
        dtype: int64
```

Hence, we can easily remove them, by dropping the entire records with missing values. In fact, since all of them belong to the '*name*' attribute, without the name is impossible to consider them as good candidates to be the best rental place in our analysis.

Figure 4. The airbnb dataset after having removed missing values .

```
The airbnb dataframe shows 25402 records (rows) and 5 attributes (columns)
```

	name	borough	neighborhood	latitude	longitude
0	Skylit Midtown Castle	Manhattan	Midtown	40.75362	-73.98377
1	Cozy Entire Floor of Brownstone	Brooklyn	Clinton Hill	40.68514	-73.95976
2	Entire Apt: Spacious Studio/Loft by central park	Manhattan	East Harlem	40.79851	-73.94399
3	Large Cozy 1 BR Apartment In Midtown East	Manhattan	Murray Hill	40.74767	-73.97500
4	Cute & Cozy Lower East Side 1 bdrm	Manhattan	Chinatown	40.71344	-73.99037

Then, for the goals of the following sections, let's build other 2 different datasets starting from the *airbnb* one. One of them, named *sheepshead_bay* dataset, will be the main for the further operations.
At the end of this section, we have 3 different cleaned dataset that will be used for different aims:

- ***airbnb***, for a visualization of apartments location in NYC;
- ***brooklyn***, with the apartments in the Brooklyn borough (Figure 5);
- ***sheepshead_bay***, the mainly used dataset about a single neighborhood in the Brooklyn borough (Figure 6).

Figure 5. The brooklyn dataset with the head method.

```
The brooklyn dataframe shows 9558 records (rows) and 5 attributes (columns)
```

	name	borough	neighborhood	latitude	longitude
0	Cozy Entire Floor of Brownstone	Brooklyn	Clinton Hill	40.68514	-73.95976
1	Only 2 stops to Manhattan studio	Brooklyn	Williamsburg	40.70837	-73.95352
2	Perfect for Your Parents + Garden	Brooklyn	Fort Greene	40.69169	-73.97185
3	Hip Historic Brownstone Apartment with Backyard	Brooklyn	Crown Heights	40.67592	-73.94694
4	Sweet and Spacious Brooklyn Loft	Brooklyn	Williamsburg	40.71842	-73.95718

Figure 6. The sheepshead_bay dataset with the head method.

```
The Sheepshead Bay dataframe shows 59 records (rows) and 5 attributes (columns)
```

	name	borough	neighborhood	latitude	longitude
0	Bright Modern Charming Housebarge	Brooklyn	Sheepshead Bay	40.58422	-73.94079
1	Luxury L-Shape Studio + 3 cats	Brooklyn	Sheepshead Bay	40.59721	-73.95149
2	Charming Housebarge w/ outside deck	Brooklyn	Sheepshead Bay	40.58408	-73.94122
3	Great studio with 2 rooms and kitchen	Brooklyn	Sheepshead Bay	40.58426	-73.95949
4	A Dream! Luxury 3 Bedroom Apt+Pking	Brooklyn	Sheepshead Bay	40.58527	-73.93534