Investment of properties for short rent in Airbnb

Viwan Jarerattanachat

Applied Data Science Capstone project, 14nd June 2019

Introduction

Barcelona, Spain, is one of the most popular tourist destinations. According to Inside Airbnb [1], on 19th April 2019, there were 17899 properties listed in Airbnb website. Accompany with tourist attraction data from Foursquare, I manage to obtain an insight, how to maximize properties profit by setting the optimized price. This will benefit investors who want to invest in the properties in Barcelona or the residents who already have properties and want to make a profit from them. Note that, this work only pay attention to short stay (i.e. about or less than a week).

Data

Airbnb data

The Barcelona properties listed in Airbnb website are obtained from http://insideairbnb.com/get-the-data.html (Barcelona, Catalonia, Spain). The listings.csv.gz**,** neighbourhoods.csv (Neighbourhood list for geo filter)**,** neighbourhoods.geojson (GeoJSON file of neighbourhoods of the city) are used for this project. All files were obtained on 27th May 2019 and compiled on 10 April 2019.

Attraction

The attractions are obtained by web scraping (BeautifulSoup) from Tripadvisor (https://www.tripadvisor.co.uk/Attractions-g187497-Activities-Barcelona_Catalonia.html) , accessed on 2nd Jun 2019.

Nearby attractions location

The attraction locations are obtained by using **Foursquare API**.

Methodology

Data preparation

List of Airbnb Barcelona properties

List of Barcelona properties ('listings 2.csv') was imported by using *pandas*. It had 17,899 rows and 106 columns. The data was cleaned and wrapped as following.

Drops unused columns – all of unnecessary columns (e.g. url, property name, summary, address) were dropped. Some features such as bed type and amenities were drop because these kinds of information are hard to interpret (and not suitable for this short project). Weekly price, and monthly price were also excluded because the objective of this work is for short-stay.

Drops unused rows – all properties that require minimum stay greater than 3 nights were removed from the list.

Checks data types – all dataseries were checked if they had suitable datatypes. For example, dataseries that represent prices (price, cleaning fee, security deposit, and extra people) should have float64 data type.

Checks for NaN values – NaN(s) were treated as following:

- dropped square feet as it has too many NaN (11847 of 12219) and did have enough data to replace them.
- replaced NaN with 0 for cleaning fee, security deposit, and reviews per month
- replaced NaN with 1 for bathrooms
- replaced NaN with 'No name', 'No hostname' for name and host name, respectively.
- used values of beds column for NaN in bedrooms column if check they can be used in this case.
- used values of bedrooms column for NaN in beds column 'beds'
- replaced NaN with mean for review scores value, review scores location

Refined Data – dropped rows which hosts only post only a property. This was done under the assumption that the host that post 2 or more properties are likely to do business.

After these processes, there were 7,330 rows and 25 columns.

Attractions and their locations

The list of top 30 attraction names was obtained from TripAdvisor by using web scraping. The BeautifulSoup library was used. The foursquare place API venues search was used to obtain the location (i.e. latitude and longitude). The information was kept in a form of pandas Dataframe.

	Attraction Name	Attraction Latitude	Attraction Longitude
0	Basilica of the Sagrada Familia	41.403519	2.174354
1	Casa Batllo	41.391717	2.165009
2	Gothic Quarter (Barri Gotic)	41.381915	2.177188
3	Palau de la Musica Catalana	41.387596	2.175298
4	Mercat de la Boqueria	41.381959	2.172011

Figure 1. *An example of attraction pandas dataframe.*

Data Miming

Data mining was performed with an assistance of *pandas* library. First, the data was generalized data by calculating price per person with an assumption that number of guests in each property is maximum. Before that, the number of price-included guests was checked if it is less than or equal number of accommodates. Because, it is not fair if the properties can serve smaller number of guests than the number of price-included guests. If this happen, set number of price-included guests equal to number of accommodates. Price per person (PPP) was calculated as follow.

$$PPP = \frac{Actual\ price}{Number\ of\ accommodates}$$

$$Actual\ price = price + (\#accommodates - \#\text{price\_included\_guests}) \times \text{extra\_person\_price}$$

Pie charts, box plots, histograms were plots to understand/get inside the data.

Data Visualization

All maps were created and rendered by using *Folium* library. All plots were done by *Matplotlib*.

K-means clustering and data preparation

Because the Barcelona data is too large, only the data of **Eixample district** was used for modelling. When looking closely at the attraction list, Plaza de Catalunya was missing from the list. Therefore, it was added. Other possible attractions were search using foursquare API. Unfortunately, it did not work. It only gave popular hotels with a query search of 'attraction'. The minimum distance to the nearest attraction was calculated using *distance.geodesic* function from *geopy* library.

K-means clustering was used to group the properties and obtain the commons. This was done by using *scikit-learn* library. The features were property type, room type, price per person cls, review cls, and minimum distance cls. Note, cls refers to categorized data. Before feeding the data to the algorithm, features were checked if there is linear relation between them. The categorized data(s) were onehot encoded.

Data Categorization

Price per person (USD)

0 - 20	very low
20 - 40	low
40 - 60	average
60 - 80	high
≥ 80	very high

Min distance (meter)

0 - 400	very close
400-800	near
≥ 800	far

Reviews

1-3	very bad
3-5	bad
5-7	average
7-9	good
≥ 9	excellent!

Results

Properties and attractions in Barcelona

As shown in the figure 2, approximately 90% of properties (for business as described in methodology) located in Ciutat Vella, Eixample, Gràcia, Sant Martí, and Sants-Montjuïc district. Eixample has an outstanding number of 40.2%. The second is Ciutat Vella with 17.7%. These because there are many attractions located in these two districts. In addition, they are also city center. The most properties in Sant Martí are located close to the Barcelona beach. As for Sants-Montjuïc, the properties are mostly clustered near Eixample distric. Furthermore, figure 2 reveals that there are clusters of properties about the attractions.

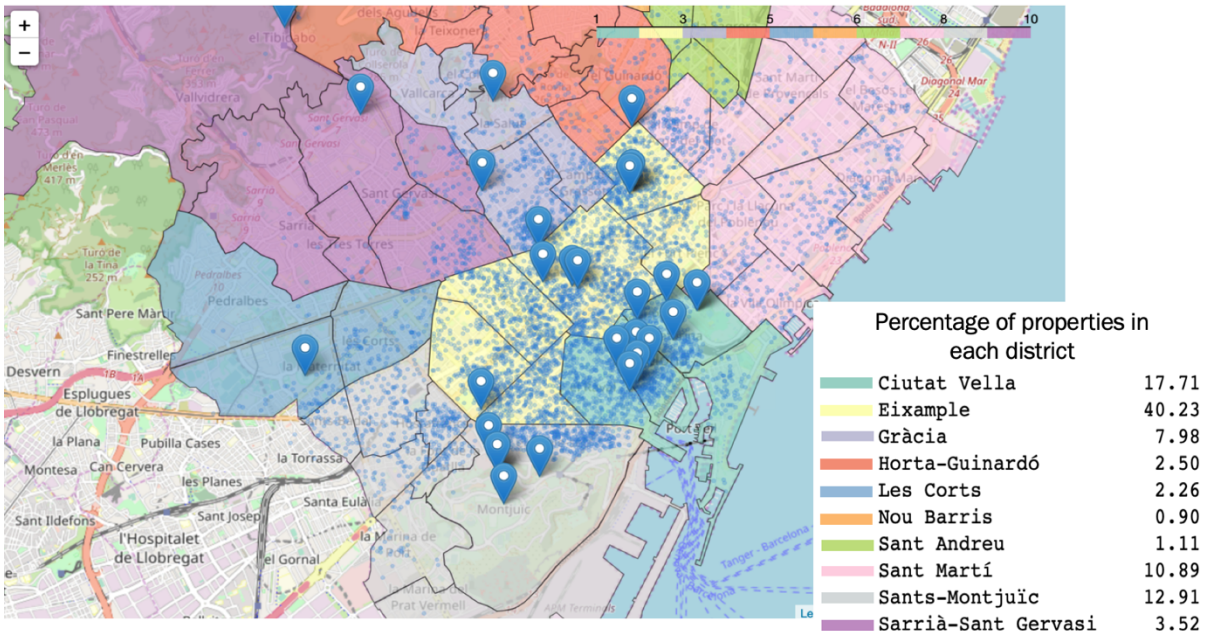


Figure 2 Barcelona properties and attractions on the map. *Properties are refined as described in methodology. Markers are attractions. Circle markers are properties. Overlay colors represents district: bluegreen (next to the yellow) is Ciutat Vella, yellow is Eixample, light purple is Gràcia, red is Horta-Guinardó, Les Corts, orange is Nou Barris, green is Sant Andreu, pink is Sant Martí, grey is Sants-Montjuïc, and purple is Sarrià-Sant Gervasi*

How much does it cost to stay? Where?

Price per person was calculated to normalize price of each property. The table1 and box plot in figure 3 show that the average and median price of the properties in Eixample is the highest. However, it also has a very large standard deviation and contains a large number of outliners. The average price of properties in Nou Barris is significantly low. Horta-Guinardó, Les Corts Nou Barris, and Sant Andreu have a smaller number of properties and cheaper when comparing to the others. Comparing between Eixample and Ciutat Vella, to short-rent property in Ciutat Vella is slightly cheaper.

Each property was labelled with its price per person and mapped onto the Barcelona map, figure 4. The distribution of red dots (properties with a low price) is shown across the city. There are some places that do not have any red dot. These mean one can find a property with price 20-40 USD all across Barcelona. The properties that cost greater than 40 (green, purple, and magenta dots) are mostly presented in Ciutat Vella and Eixample. Some are presented in Sant Martí, and Sants-Montjuïc. This implies that if one really wants to do a business and want to set relatively high rental cost. He/she must have property(s) in Ciutat Vella, Eixample, Sant Martí, or Sants-Montjuïc district.

Table 1 Properties' price per person in each district

neighbourhood_group_cleansed	count	mean	std	min	25%	50%	75%	max
Ciutat Vella	1298.0	31.351764	20.696578	6.500000	20.762500	27.500	36.000000	333.333333
Eixample	2949.0	84.981121	463.389539	2.000000	22.000000	30.000	42.500000	6000.000000
Gràcia	585.0	37.762115	40.027835	1.500000	22.500000	28.000	37.333333	500.000000
Horta-Guinardó	183.0	32.678383	62.856794	10.000000	18.000000	22.500	28.000000	760.000000
Les Corts	166.0	45.884115	161.664751	7.000000	19.562500	27.500	31.916667	1500.000000
Nou Barris	66.0	20.372475	8.802185	8.000000	15.000000	17.500	23.000000	52.500000
Sant Andreu	81.0	25.204203	33.958135	8.333333	15.000000	20.000	25.000000	266.666667
Sant Martí	798.0	39.555705	96.958989	5.000000	20.000000	27.500	35.750000	2000.000000
Sants-Montjuïc	946.0	36.537948	62.163760	4.000000	20.000000	26.325	35.000000	750.000000
Sarrià-Sant Gervasi	258.0	38.635124	85.844474	4.000000	23.083333	26.875	32.500000	1000.000000

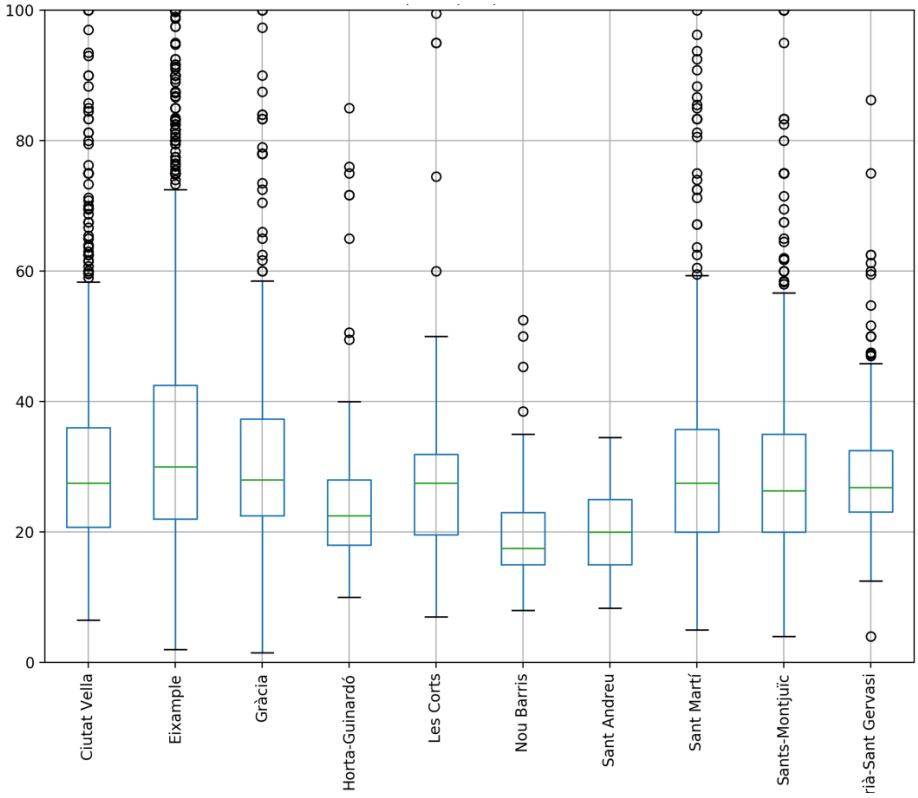


Figure 3 Box plot of price per person per night

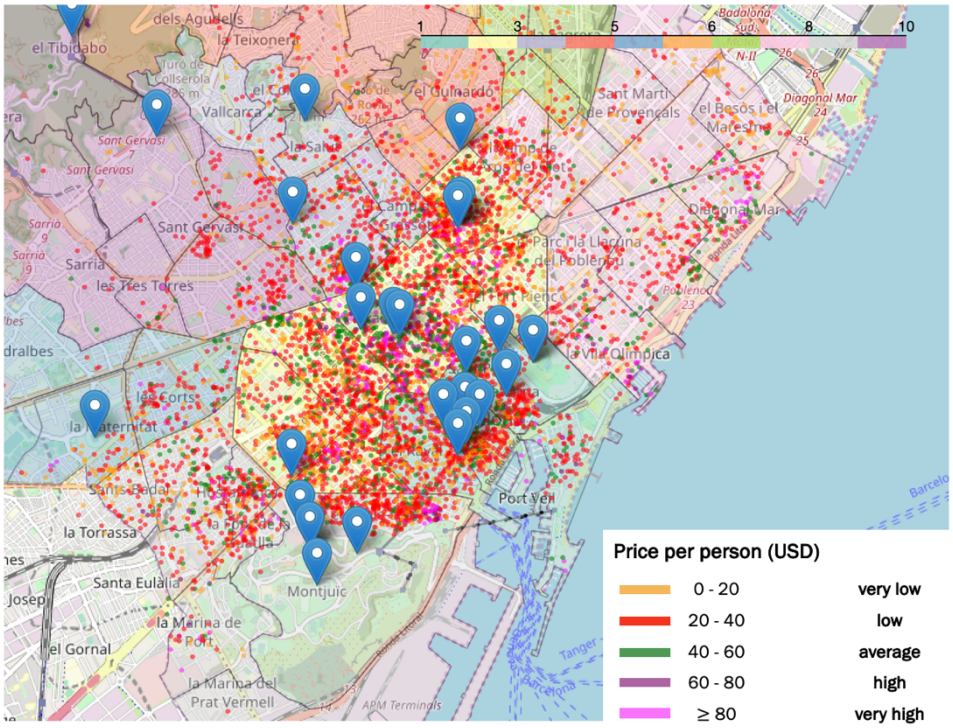


Figure 4 Barcelona properties and attractions on the map. *Properties are represented by circle-markers and colored by price per person classification.*

Properties in Eixample

To analyze all properties across Barcelona was exceed my resource capacity and not reasonable with a given time. Therefore, only the properties in Eixample were analyzed. Figure 5 is a Barcelona map but zoom-in at Eixample. I spotted that *plaza de Catalunya* was missing from the map, then I added it to the attraction list. Just by eye, orange, red, and green dots are observed around Eixample district. In contrast, most of purple and magenta dots are observed near by the attractions. This is a sign indicating that the properties near the attractions received higher income.

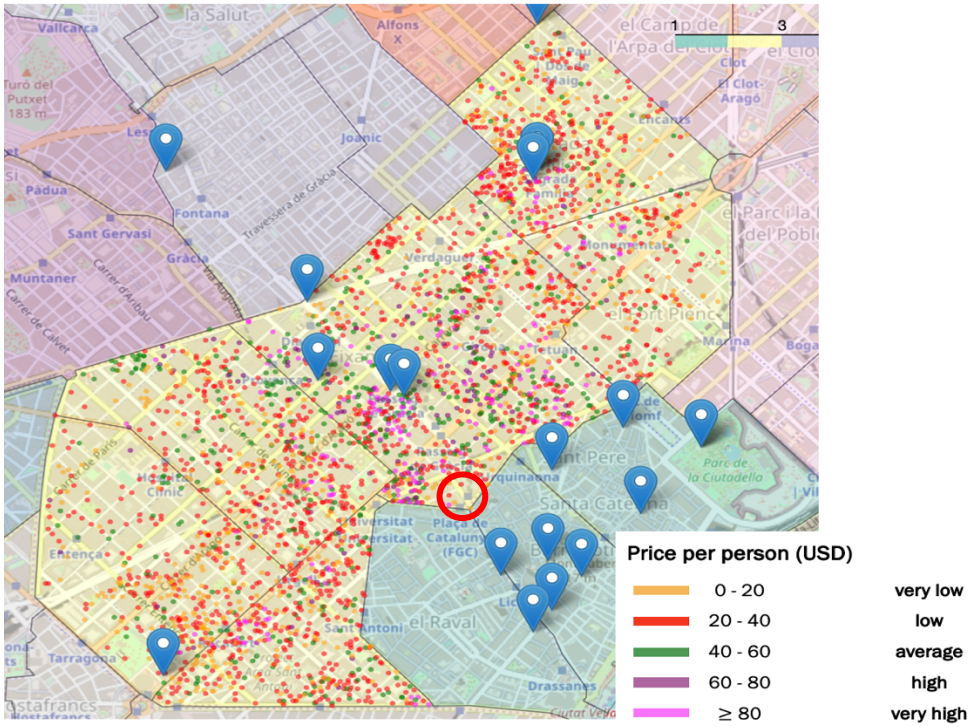
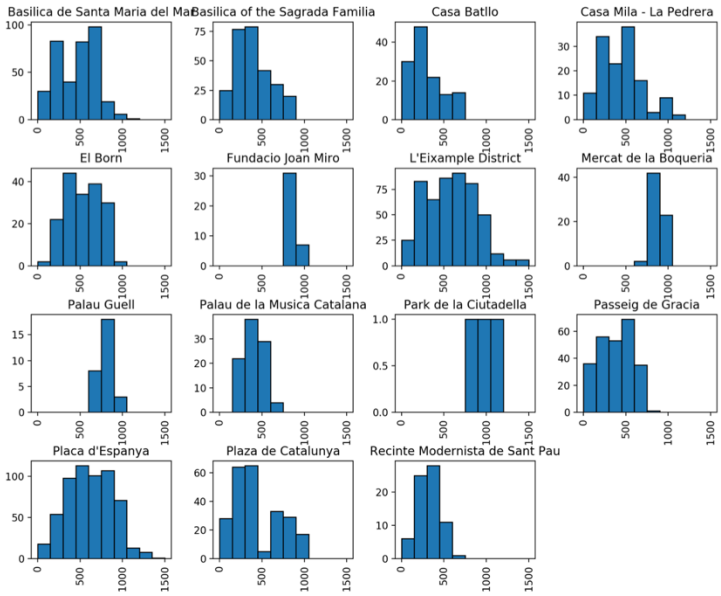


Figure 5 Properties and attractions in Eixample. *Properties are represented by circle-markers and colored by price per person classification. Red circle indicates the missing attraction, plaza de Catalunya.*

The minimum distance to the nearest attraction was calculated and shown in the figure 6. At this point, I figured out that the location of Basilica de Santa Maria del Mar was wrong. It should be located outside Eixample. The location of Basilica de Santa Maria del Mar, here, is a part of Basilica of the Sagrada Familia. Therefore, there are 389 properties with in 500 meters of Basilica of the Sagrada Familia. There are greater than 150 properties near Placa d'Espanya, Plaza de Catalunya, L'Eixample District, and Passeig de Gracia.



Attraction name	
Basilica de Santa Maria del Mar	181
Basilica of the Sagrada Familia	208
Casa Batllo	106
Casa Mila - La Pedrera	80
El Born	84
L'Eixample District	201
Palau de la Musica Catalana	71
Passeig de Gracia	178
Placa d'Espanya	201
Plaza de Catalunya	159
Recinte Modernista de Sant Pau	65

Name: id, dtype: int64

Figure 6 Minimum distance to the nearest attraction grouped the nearest attraction. *The insert table shows that number of properties within 500 meters of each attraction.*

K-means Clustering

Before actually performing K-means clustering, the relation between minimum distance to the nearest attraction and the price per person was check whether they have linear relation. This because it is pointless to perform complex model if a sample linear relation exists. Figure 7 shows that between minimum distance to the nearest attraction and the price per person does not have a sample linear relation. However, the pattern of distribution of 0-400, 400-800, and greater than 800 are different. I also looked at the relation between the distance and location review score, figure 8, and found that there is not linear relation between them.

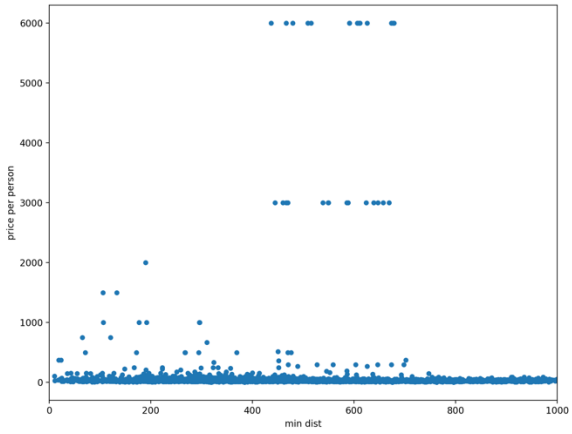


Figure 7 Scatter plot between minimum distance to the nearest attraction and the price per person.

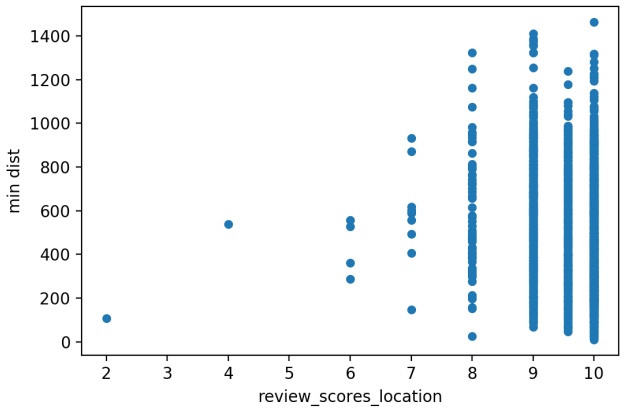


Figure 8 Scatter plot between the location review score and minimum distance to the nearest attraction.

K-means clustering with k is 10 was performed on the features (price per person, review, room type, properties type, minimum distance to nearest attraction). The properties are labelled and shown in the map, figure 9. The algorithm clearly used the distance as one of the important features as the figure shows the pattern in the distance. To distinguish the mix color dots at the same distance, the group profiling was performed. The group profile, table 2, show that apart from group 1, 2, and 3, the other groups are low price. When looking at group 1, 2, and 3 in details, it uncovers that the hotel property type has a high price. Comparing between group 3 and 4 reveals that a property with private room has high price per person.

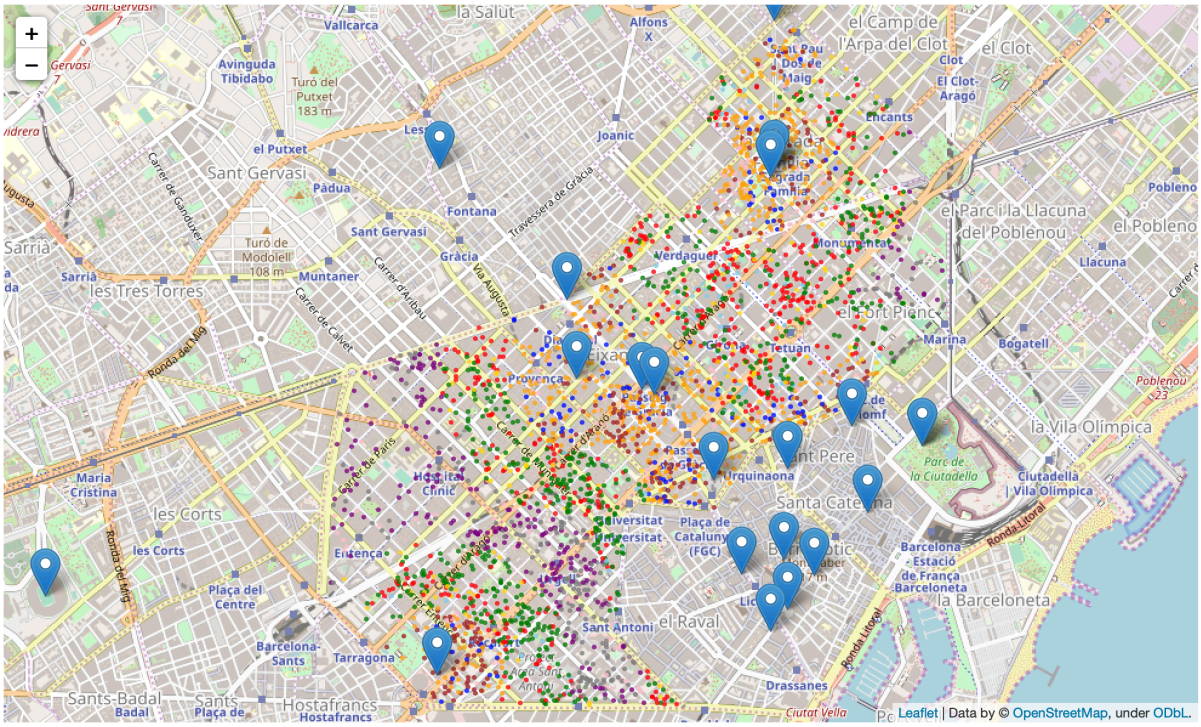


Figure 9 Properties and attractions in Eixample*. Properties are represented by circle-markers and colored by K-means cluster group. Blue is group 1, orange is 2, green is 3, red is 4, purple is 5, brown is 6, pink is 7, grey is 8, gold is 9, and skyblue is 10.*

Table 2 Profile of each group

GROUP	%	PROPERTY TYPE	ROOM TYPE	LOCATION	PRICE	REVIEW
1	4.4	Mix	Private	Very close	Mix (Mostly V. low)	Excellent
2	21.0	Apartment/Service apartment	Entire home/apt	Very close	Mix (Mostly V. low)	Excellent
3	20.0	Mix	Private	Near	Mix (Mostly V. low)	Excellent/Good
4	16.5	Mix	Entire home/apt	Near	Very low	Excellent
5	7.8	Mix	Private	Far	Very low	Excellent
6	9.9	Apartment	Private	Very close	Very low	Excellent
7	6.1	Mix	Entire home/apt	Very close /Far	Very low	Good
8	6.8	Mix	Entire home/apt	Far	Very low	Excellent
9	4.4	Apartment/Service apartment and Condominium,	Entire home/apt	Near	Very low	Good
10	3.0	Serviced apartment	Entire home/apt	Near	Very low	Excellent

Discussion

There were properties for short stay across Barcelona city, most of them were clustered about the attractions. The majority of properties are located in Ciutat Vella, Eixample, Gràcia, Sant Martí, and Sants-Montjuïc district. This because there many attractions in these districts. Moreover, Ciutat Vella, Eixample district is a city center that have many facilities.

When zooming-in at Eixample, I found that the Plaza de Catalunya was missing from the attraction list. This because I only obtained 30 attractions from TripAdvisor. Therefore, the attraction list did not cover all Barcelona attractions. In addition, the location (from Foursquare API) of Basilica de Santa Maria del Mar was also wrong. When I was performing the API request, I found that only a small change (i.e. from plaça to plaza) in a query can make a difference in result.

When looking at the relation between the minimum distance to the nearest attraction and the location review score, it seems that there is no relation between them. Some places with very high score were located very far from the attractions. One explanation of this may be there is some metro, supermarket, nearby those properties.

For future direction, the transportation should be included as it plays important role when one selects a place to stay. After refining the features that should be used for the model, decision tree classifier would be used to identify the price range for a given property(s).

Conclusion

In this project, I used the combination of data mining, data visualization and K-means clustering techniques to get insights from Barcelona Airbnb data to optimize the profit from short-rent property business. I found that there are clusters of properties about the attractions. To set relatively high rental cost. He/she must have property(s) in Ciutat Vella, Eixample, Sant Martí, and Sants-Montjuïc. The properties near the attractions have an opportunity to receive high income but there is very less opportunity for the others. In case of Eixample district, the property does not need to be close to the attraction, locating near the attraction (400-800 meters) has the same effect. Private room and hotel seem to have high profit.

References

[1] Inside Airbnb (2019), http://insideairbnb.com/get-the-data.html, accessed on 27th May 2019

Appendix

Profiles of group 1, 2, 3, and 7

Group 1

```
df[df['Klabels']==0].groupby(['property_type','room_type','dist_cls','price_cls','rev_cls']).count()['id']
```

```
property_type           room_type     dist_cls    price_cls        rev_cls
Aparthotel              Private room  very close  price:very low   review:good          1
Bed and breakfast       Private room  very close  price:high       review:excellent     1
                                                  price:very low   review:excellent    19
                                                                   review:good          5
                        Shared room   very close  price:very low   review:excellent     1
Boutique hotel          Private room  very close  price:average    review:excellent     2
                                                                   review:good          1
                                                  price:low        review:excellent     5
                                                  price:very low   review:excellent    21
Casa particular (Cuba)  Private room  very close  price:very low   review:excellent     2
Condominium             Private room  very close  price:very low   review:excellent    10
Guest suite             Private room  very close  price:very low   review:excellent     1
Guesthouse              Private room  very close  price:very low   review:excellent     2
Hostel                  Private room  very close  price:low        review:excellent     1
                                                  price:very low   review:excellent    26
                                                                   review:good          2
                        Shared room   very close  price:average    review:excellent     2
                                                  price:very low   review:excellent    14
                                                                   review:good          2
Hotel                   Private room  very close  price:very low   review:excellent     4
House                   Private room  very close  price:very low   review:excellent     2
                                                                   review:good          1
Loft                    Private room  very close  price:very low   review:excellent     1
Other                   Private room  very close  price:very low   review:excellent     2
                                                                   review:good          1
Serviced apartment      Private room  very close  price:very low   review:excellent     1
Name: id, dtype: int64
```

Group 1 (4.4%) : Excellent reveiw, privete room, price very low, very close

Group 2

```
In [102]: df[df['Klabels']==1].groupby(['property_type','room_type','dist_cls','price_cls','rev_cls']).count()['id']
```

```
Out[102]: property_type       room_type        dist_cls    price_cls        rev_cls
          Aparthotel          Entire home/apt  very close  price:very low   review:excellent      6
          Apartment           Entire home/apt  very close  price:high       review:excellent      1
                                                          price:low        review:excellent      3
                                                          price:very low   review:average        8
                                                                           review:bad            1
                                                                           review:excellent    486
          Condominium         Entire home/apt  very close  price:very low   review:excellent     15
          Guest suite         Entire home/apt  very close  price:very low   review:excellent      1
          House               Entire home/apt  very close  price:very low   review:excellent      1
          Loft                Entire home/apt  very close  price:very low   review:excellent      7
          Serviced apartment  Entire home/apt  very close  price:very high  review:excellent      1
                                                          price:very low   review:average        1
                                                                           review:excellent     89
          Name: id, dtype: int64
```

Group 2 (21%) : Excellent review, Entire home/apt, very close, Apartment/Service apartment with very low price

Group 3

```
In [103]: df[df['Klabels']==2].groupby(['property_type','room_type','dist_cls','price_cls','rev_cls']).count()['id']
```

```
Out[103]: property_type           room_type     dist_cls  price_cls        rev_cls
          Apartment               Private room  near      price:very low   review:average       1
                                                                           review:bad           2
                                                                           review:excellent   412
                                                                           review:good         43
                                  Shared room   near      price:very low   review:excellent     1
          Bed and breakfast       Private room  near      price:very high  review:excellent     2
                                                          price:very low   review:average       1
                                                                           review:excellent    18
                                                                           review:good          1
                                  Shared room   near      price:very low   review:excellent     1
          Boutique hotel          Private room  near      price:low        review:excellent     2
                                                          price:very high  review:excellent     4
                                                          price:very low   review:average       1
                                                                           review:excellent    21
                                                                           review:good          1
          Casa particular (Cuba)  Private room  near      price:very low   review:excellent     3
          Condominium             Private room  near      price:very low   review:excellent     4
          Guest suite             Private room  near      price:very low   review:excellent     4
          Guesthouse              Private room  near      price:very low   review:average       1
                                                                           review:excellent     1
          Hostel                  Private room  near      price:very low   review:excellent    14
                                                                           review:good          4
                                  Shared room   near      price:very low   review:excellent    11
          Hotel                   Private room  near      price:very high  review:excellent    20
                                                                           review:good          1
                                                                           review:very bad      1
                                                          price:very low   review:excellent     2
          House                   Private room  near      price:very low   review:excellent     4
                                                                           review:good          1
          Loft                    Private room  near      price:very low   review:excellent     1
          Other                   Private room  near      price:very low   review:excellent     1
          Serviced apartment      Private room  near      price:very low   review:excellent     7
          Name: id, dtype: int64
```

Group 3 (20.0%) : private room, near, very low price, excellent/good review

Group 7

```
df[df['Klabels']==6].groupby(['property_type','room_type','dist_cls','price_cls','rev_cls']).count()['id']
```

```
property_type       room_type         dist_cls    price_cls       rev_cls
Apartment           Entire home/apt   far         price:very low  review:good     42
                                      very close  price:very low  review:good    125
Loft                Entire home/apt   far         price:very low  review:good      2
                                      very close  price:very low  review:good      2
Serviced apartment  Entire home/apt   far         price:very low  review:good      1
                                      very close  price:very low  review:good      7
Name: id, dtype: int64
```

Group 7 (6.1%) : Entire home/apt, either very close or far, good review.