Battle of Neighborhoods in Barcelona

Yueming HU

April 2, 2021

1 Introduction

1.1 Background

Barcelona is a very nice city located in the coastal area of northeastern Spain.

Client L is a newly established real estate company based in Barcelona. This company intends to develop several real estate projects in different neighborhoods of the city. As we know, property project needs to consider its surrounding environment, what facilities and venues are nearby. And the prices of real estates can differ greatly according to their location, size, room numbers, etc.

1.2 Problem

Client L is not familiar with all the neighborhoods in Barcelona and also not sure what should be the reasonable prices to set for properties built in future in different neighborhoods. Therefore, the company entrusts me to provide a solution by using data analysis, so as to allow them to choose appropriate neighborhoods in a better-informed manner and have a better ideal of the possible price.

After initial analyzing of Client L's request, I decide to divide this problem into two parts, getting a general overview of neighborhoods' characteristics; and predicting the housing price based on different features, such as location (neighborhood), housing size, room numbers, floor numbers, etc.

1.3 Interest

Obviously, real estate companies would be interested in the prediction of housing prices and neighborhoods characteristics. The solution can even be used for property development in different cities in future.

Potential housing buyers may also be interested in, because this solution can help them choosing neighborhoods to live in and checking whether the price is reasonable or not.

2 Data Acquisition and Cleaning

2.1 Data sources

The housing price data of Barcelona can be found in a Kaggle dataset, at https://www.kaggle.com/jorgeglez/barcelona-idealista-housingprices .

Neighborhood information is got from Foursquare Api.

To use Foursquare Api request, geographical coordinates, i.e. latitude and longitude of all the neighborhood are necessary, so the latitude and longitude are obtained through Geopy.

2.2 Data cleaning

The Kaggle's csv file is read into dataframe by pandas. There are 3265 samples and 12 columns. Among them, the "floor" column is recorded as string, such as "floor 1" and "ground floor". I need to change this column into integer so as to do analysis.

I check all the unique values of "floor column", and find that there are "mezzanine", "ground floor" and "Multiple floor". I replace the "ground floor" with "floor 0", "mezzanine" with "floor 0.5", and "Mutiple floor" with "floor 1.5". Because mezzanine is the one between two floors and mutiple floor may refer to the duplex, which usually has two or three stories, here I just use the floats to show them as different types.

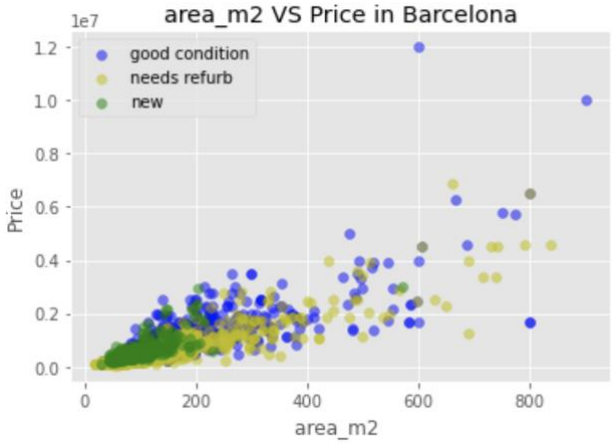
Then I use lambda function to split the word "floor" and the number, keep only the number (integer) into a separate column as "floor number".

2.3 Feature Selection

The features in the dataframe are: "city", "district", "neighborhood", "condition", "type", "rooms", "area_m2", "lift", "views", "floor number", "prices". Since my prediction target is a reasonable price of property, uased as reference for my client, so I will choose only the very typical features, which means those have relatively big or obvious impact on the price.

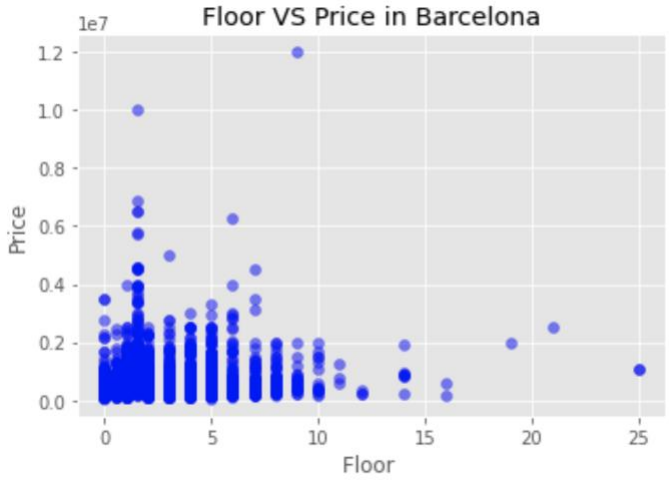
"Neighborhood", no need to say, refers to the location, is a decisive factor to price, "area", "room number" are also important features. In contrast, housing type, whether there is a lift or not, and with or without exterior view are not that decisive, also they are string type, so I will not use them. But I'm not sure about the features of condition and floor number. So I examine them in more detail.

For "condition", the unique values are "good condition", "new", "needs refurb". In order to see whether different conditions lead to huge difference in price, I decide to plot them onto an image to have a direct view. I use blue dots to represent "good condition", yellow dots as "needs refurb", green as "new". The plot is as below.



From the plot, we can see that the dots of different colors are mixed together, especially new and good condition, and sometimes for the same area, the needs refurb are even more expensive than the good condition. This means condition is not a decisive factor to the price, and it will not cause huge difference in prices.

For floor number, the usual understanding is that the higher the floor is, the more expensive the price, because the view would be better. I also use image to have a look. The x-axis is floor number, and y-axis is price. The plot is as below.



From the plot, out of my expectation, there is no strong relationship between the floor number and price. For each different floor below 10th floors, there can be very high and very low price. The line of floor 1.5 has prices higher than others, but it refers to duplex, so the main reason behind may be the size. Except for this floor, the price range of others are mostly the same. So, the floor feature will also not be used in this model building for a general prediction.

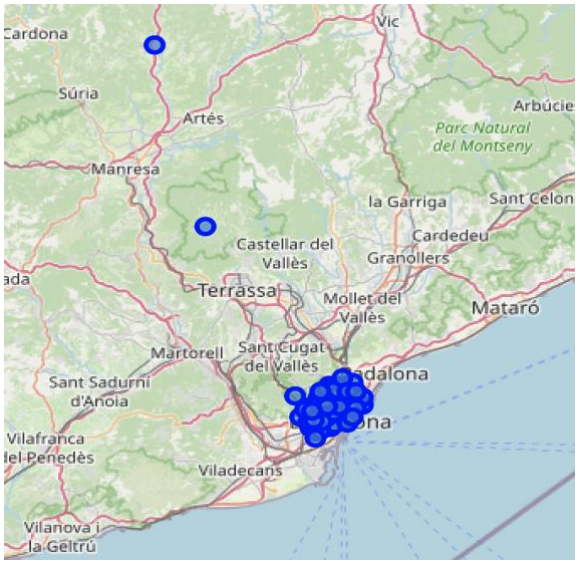
As for the neighborhoods data, there altogether 65 neighborhoods. But when I use grouby and count method to check the data, I find that several neighborhoods have only three or four samples. I think if a neighborhood has very little sample data, the prices for this neighborhood would not be meaningful. So I called out all the neighborhoods with 10 or less than 10 samples, using a lambda function to mark them as category "other". There are all together 56 samples of "other", and I drop them.

Now, there left 3207 samples of 55 neighborhoods. Since neighborhoods feature is string type, I use one-hot encoding to convert all the neighborhoods into dummy variables.

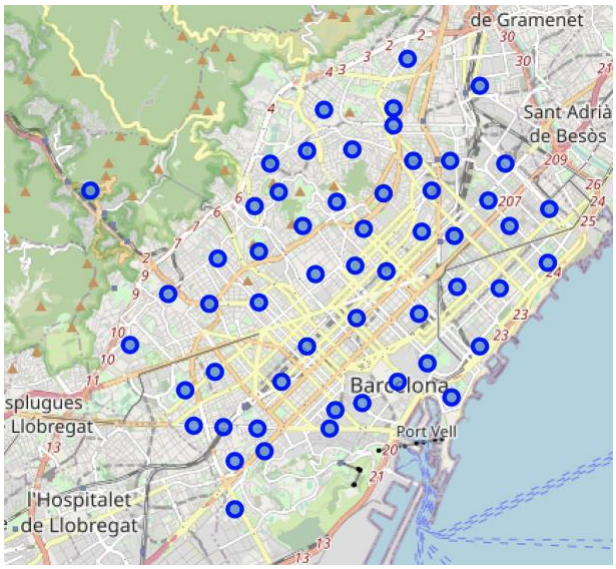
2.4 Removing and correcting outliers

From the above plots, I noticed there is dot showing extremely high price, which is in fact the highest price. So I take a look and find that the price of the sample is 12000000, area is 600 m^2 and only four rooms. It's not possible for a 600 m^2 house having only 4 rooms and the price is really ridiculous. I think this sample is not correct and remove it.

For latitude and longitude data, I use Geopy's Nominatim function to get the geocodes. Since this way uses the name of location to get the geocodes, there might be inaccurate data. So I plot all the neighborhoods on the map to see if all the location data is right. The map is as below.



Barcelona is near the coast, but two dots are far away from the coastal area. I modify the address of the two neighborhoods respectively, make request again, and then get all correct data. The correct map is as below.

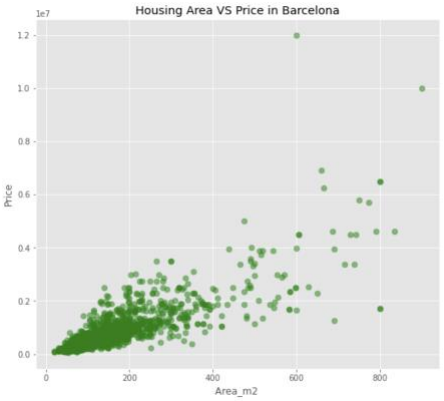


3 Data Analysis

3.1 Housing price data analysis

3.1.1 Relationship between area, neighborhoods and price

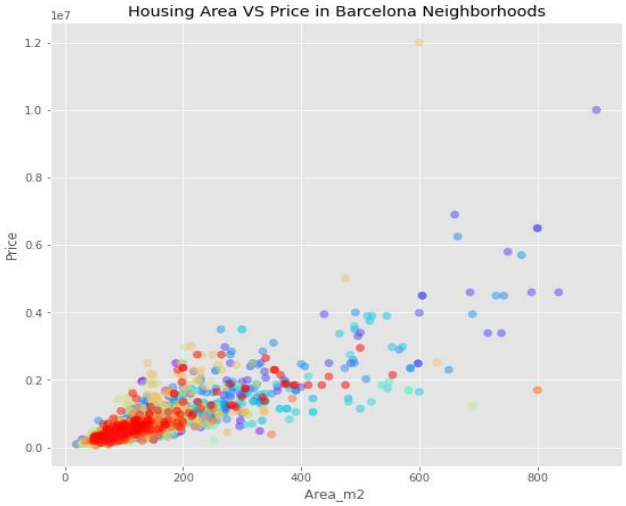
It is almost a common sense that the larger the housing size is, the higher the price. But is there a linear regression relationship between the two? I decide to use an graph to have a direct view. The plot is as below.

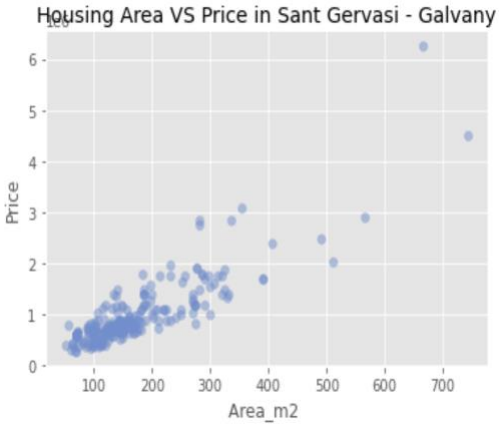


Obvisouly, there is a linear relationship, we can easily draw a line in the graph. But there is also certain dispersion, which may relate to other features, so I continue to explore.

I want to see the situation in each neighborhbood. So I use matplotlib.cm rainbow function to assign one different color to each neighborhood, to see the price and area relationship within each neighborhood. The plot is as below left. One color represents one neighborhood, we can see dots of same color concentrate in one part of the graph. The red dots mainly exist in the left lower side of graph, orange and yellow dots in the left higher side, but blue and purple dots scatter in the right side. This means neighborhoods do have big impact on the price.

Since the dots are quite concentrated, I randomly choose another neighborhood and make a separate plot, as right below. There is indeed a linear relation.





3.1.2 Relationship between number of rooms and price

Since I already know neighborhoods can make a difference on price, so I'll check the relation between number of rooms and price within the same neighborhood. Again, I use different colors to represent each neighborhood. The graph is as below.

If we look at the dots as a whole, there may not be linear relationship. But we examine the dots with same color, it's not difficult to find a line. So it's ok to say that more rooms would result in higher price.

Based on the above observation, I think a linear regression model can be built to predict the housing price.

3.2 Neighborhoods data analysis

To get an overview of neighborhoods' characteristics, a good way is to find out what are the nearby venues in each neighborhood by using Foursquare Api. So I write function to get the venues and their corresponding categories of each neighborhoods, taking advantage of geocodes obtained from Geopy.

As result, I got 2935 venues, in 279 uniques categories. I use groupby to count the venue number of each neighborhood, find that neighborhoods differ a lot, some have around 100 venues, some have less. The venues are features now, so I convert the venue categories into dummy codes which can be used directly for analysis and in the model later.

As for the general characteristics, I don't think it's appropriate to present the characteristics or most common venues of each neighborhood, moreover 55 neighborhoods would be too much. Instead, it's possible to group the neighborhoods into several groups, and summarize the characteristics of each group. Therefore, a K-means model would be suitable.

4 Model Building

4.1 Linear regression model for housing price prediction

4.1.1 Data fitting and evaluation

For my linear regression model, the x is "number of rooms", "area" and dummies of 55 neighborhoods. Y is the price. I split the tain and test data, test size at 0.2. Then fitting the train data, the resulting score or accuracy of the model is 0.7833288162940156.

To evaluate the model, I use cross valitdation to test the accuracy of my model. I create a shufflesplit for the dataset, setting 5 randomized folds, and the resulting scores are: 0.76644668, 0.77273504, 0.72222751, 0.73113436 and 0.79898497. The scores are all above 0.7, ranging between 0.72 to 0.79, which means the accuracy of my model is at all time above 70%. It's good, not that precise, but already acceptable.

4.1.2 Price predict function

The linear model is ok, but it's using the dummies to represent each neighborhood, and certainly not possible to ask client to input dummies to get results. Therefore, it's necessary to write another function, which allows us to obtain price simply by inputting the name of neighborhoods, number of rooms and square meters of area.

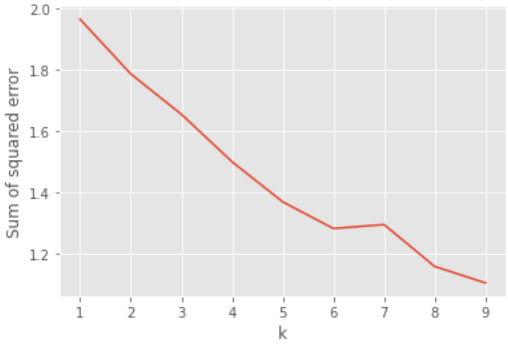
I print the X.columns to see all the columns, the first one is "room", second is "area_m2", and neighborhoods follow behind. So it's possible to find out the index of certain neighborhood by passing a condition that equals to a neighborhood name and using np.where to call out its corresponding index. Then I write a "predict_price" fucntion with neighborhood name, number of rooms and area as variables, return the predicted result of linear regression model.

4.2 K-means clustering model

4.2.1 Deciding the number of K

I have already converted all the venue categories into dummies, which can be used in the k-means model directly. But for this model, the important thing is to decide the number of k, namely how many clusters I want to have.

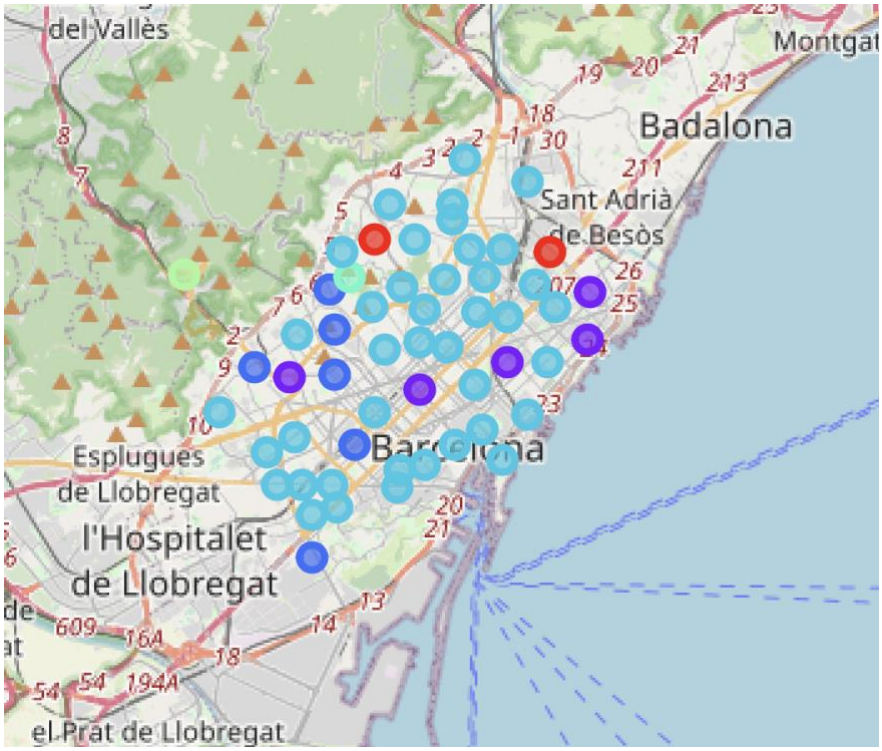
I decide to use the elbow technique to find out the optimal k number. I wirte a loop of k ranging from 1 to 10, and calculate the sum of squared errors for each k, then plot onto a graph to see where the elbow is. The maxium number is 10, because I think maxium cluster number would be 10, client would not find it clear if there are more than 10 clusters. So the graph is as below.



From the graph, I think the inflection point is 6, which means the optimal number of clusters is 6. But the line continues going down after 7 and 8, so I have some doubts, and I decide to explore further to confirm. I print out all the labels for k valuing at 6, 7 and 8. When k is 6, I notice that there are two labels with only one neighborhood, which means there two neighborhoods are special and don't belong to any other cluster. However, when k is 7, there is one more label with just one neighborhood, and when k is 8, there two more labels with just one neighborhood. This means the increased cluster contains only one neighborhood, making more special neighborhoods, that's not good, especially for clients. So, this confirms to me 6 is the optimal number of clusters.

4.2.2 Clusters of neighborhoods

I use the k-means model (k=6) to get all the labels for neighborhoods, and insert the labels back to my original datframe, so as to see which neighborhoods belong to a cluster. In order to have a better ideal, I plot them onto the map, as below.



Then I examine the 6 most common venues of each cluster, to summarize its respective characteristics.

5 Result

Based on my above analysis, I can reach the conclusion that there are six types of neighborhoods in Barcelona, with characteristics and referential housing price listed as follows.

5.1 Type One – Food and Café Easily Accessible Neighborhoods

Neighborhoods: el Carmel; la Verneda i la Pau.

Characteristics: Café and coffee shops are extremely common, nice for friends gathering. With many supermarkets, food courts and grocery stores, easily satisfying the demand of food. Located in the outer side of Barcelona.

Price prediction example:

```
predict_price('el Carmel', 6, 120)
```

```
315936.0
```

A 120 m^2 residence with 6 rooms in neighborhood El Carmel may be sold at 315,936 euro.

I also use the wordcloud to show the keywords of these neighborhoods.

5.2 Type Two -- Tourist-concentrated and Commercialized Neighborhoods

Neighborhoods: Diagonal Mar i el Front Marítim del Poblenou; el Besòs i el Maresme; el Parc i la Llacuna del Poblenou; la Dreta de l'Eixample; les Tres Torres.

Characteristics: Located in the central and beach side of Barcelona, with plenty of hotels, also has train station, restaurants, boutiques and clothing store, frequented by tourists and travelers.

Price prediction example:

```
predict_price("la Dreta de l'Eixample", 6, 120)
```

```
628630.0
```

A 120 m^2 residence with 6 rooms in neighborhood la Dreta de l'Eixample may be sold at 628,630 euro, twice of the same size apartment in Type One neighborhood.

Keywords:

5.3 Type Three – Gourmet Neighborhoods

Neighborhoods: Sant Gervasi – Galvany; Sarrià; Vallcarca i els Penitents; el Putxet i el Farró; la Marina de Port; la Nova Esquerra de l'Eixample

Characteristics: This type of neighborhoods have diversified restaurants, Japanese, Italian, Spanish, Indian, Mediterranean and Tapas restaurants. Bakery and dessert shop are also very common, suitable for foodies.

Price prediction example:

```
predict_price("Sant Gervasi – Galvany", 5, 100)
```

```
550879.0
```

A 100 m^2 residence with 5 rooms in neighborhood Sant Gervasi – Galvany may be sold at 550,879 euro.

Keywords:

5.4 Type Four – Comprehensive Cozy Living Neighborhoods

Neighborhoods: Baró; Horta; Hostafrancs; Navas; Pedralbes; Porta; Provençals del Poblenou; Sant Andreu; Sant Antoni; Sant Gervasi - la Bonanova; Sant Martí de Provençals; Sant Pere - Santa Caterina i la Ribera; Sants; Sants – Badal; Vilapicina i la Torre Llobeta; el Baix Guinardó; el Barri Gòtic; el Camp d'en Grassot i Gràcia Nova; el Camp de l'Arpa del Clot; el Clot; el

Congrés i els Indians; el Fort Pienc; el Guinardó; el Poble Sec; el Poblenou; el Raval; l'Antiga Esquerra de l'Eixample; la Barceloneta; la Bordeta; la Font d'en Fargues; la Font de la Guatlla; la Maternitat i Sant Ramon; la Prosperitat; la Sagrada Família; la Sagrera; la Salut; la Teixonera; la Vila Olímpica del Poblenou; la Vila de Gràcia; les Corts.

Characteristics: The majority of neighborhoods belong to this type and the venues are very comprehensive. There are sporting venues, such as basketball court, sports club, soccer filed and gym; outdoor open spaces, such as plaza, park and garden; different kinds of restaurants and bars; and convenient living facilities, such as grocery store, print shop, supermarket. Living in these neighborhoods could be very comfortable.

Price prediction example:

```
predict_price("les Corts", 6, 120)
```

```
483042.0
```

```
predict_price("Hostafrancs", 6, 120)
```

```
351790.0
```

Housing price can be different but generally in middle level for this type of neighborhoods, a 120 m^2 residence with 6 rooms in les Corts may be sold at 483,042 euro, in Hostafrancs may be sold at 351,790 euro.

Keywords:

5.5 Type Five – Special Mountain View Neighborhood

Neighborhoods: el Coll.

Characteristics: This is a special neighborhood, with a mountain nearby, scenic views and park, a very beautiful and natural neighborhood in the outer fringe of Barcelona.

Price prediction example:

```
predict_price("el Coll", 6, 120)
```

```
303193.0
```

A 120 m^2 residence with 6 rooms in neighborhood el Coll can be sold at 303,193 euro, not a very high price, might because the location is a little far from city center and transport may not be very convenient for mountain area. Maybe developing big villa would be more profitable.

Keywords:

5.6 Type Six – Exclusivity Away from City Bustle

Neighborhoods: Vallvidrera - el Tibidabo i les Planes.

Characteristics: Another special neighborhood, located far away from the downtown area of Barcelona, a resort, tourist destination, with museum, zoo and open space. Place to enjoy living away from city bustle.

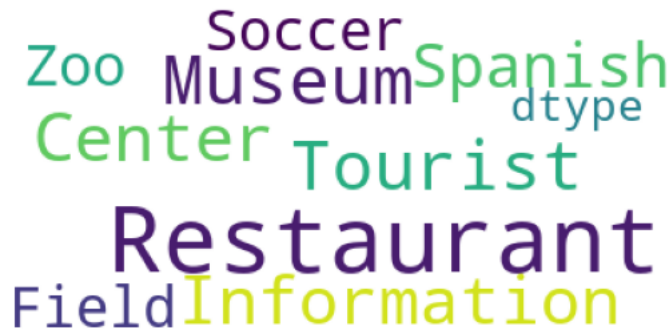
Price prediction example:

```
predict_price("Vallvidrera – el Tibidabo i les Planes", 5, 160)
```

```
386378.0
```

A villa of 160 m^2 and 5 rooms in neighborhood Vallvidrera - el Tibidabo i les Planes may be sold at 386,378 euro.

Keywords:



5.7 Explanation to Client

The neighborhoods of Barcelona can be divided into 6 types, client can decide its property category and price based on the characteristics of each type of neighborhoods and the price prediction model. For example, if the client wants to develop villas, he can choose the special type of neighborhoods el Coll and Vallvidrera - el Tibidabo i les Planes, with mountain view or outside the downtown area. If the client wants to build residencial apartment targeting at local residents, he may choose the Comprehensive Cozy Living Neighborhoods, sale price can be high or medium, but avoid the tourist-concentrated neighborhoods, where local residents might not want to be disturbed by groups of tourists.

These are just some examples, it's left for the client himself to make decisions with these information.

6 Discussion

Certainly, there are some limitation and constraints in this project. In reality, many factors can impact on the housing price, my model is a simple and general one, far from comprehensive. I feel that analysis result is highly dependent on the original data, it's not easy to find good and clear data source. I've spend a lot of time searching for housing data, the dataset used in the

project has constraints, if there is information for kitchen and bathroom number, the analysis would be better, but the current Barcelona dataset is so far the best I can find.

As for the presentation of results, I'm sure there are better methods that can combine the price prediction model and characteristics display together. For example, a website with dropdown menus allowing clients to select type, neighborhoods, housing size and then show a result. But it's out of the scope of this project, and I need to learn more to be able to do that. There is still a lot to learn.

Thank you very much for reading and reviewing this report!