José Miguel Calderón Terol 10-05-2020

Capstone Project - The Battle Of Neighborhoods

Study of the relationship between the proximity of venues, price and number of tourist apartments

Introduction

Background

In recent years, numerous tourist apartments have appeared throughout the world. The location and price of such rental is governed by the law of supply and demand. Cities with high tourist value such as Valencia, Spain, see this type of business appear continuously.

It is logical to think that the owners seek the maximum benefit and that not all places have the same commercial attraction. This causes that the number of rental apartments is not evenly distributed among the different neighborhoods. More enblematic places or leisure areas can attract more tourists. What characteristics of the place influence the aforementioned?

Problem

In this project we will decipher some of these characteristics, in addition to answering the following questions.

Which neighborhood offers the best price / number of venues?

Is there a correlation between price and the venues?

Is there a correlation between the number of apartments for rent and the venues?

What neighborhood would be suitable to put a new apartment for rent?

Interest

Firstly, customers who want to know which neighborhoods offer the greatest number of places of interest, secondly, owners who want to know which neighborhoods are the most popular and offer the best business opportunities.

Data acquisition and cleaning

Data sources

To carry out this project, the following information is necessary:

- Spatial information about neighborhoods. These data were obtained through the portal of the Valencia City Council. Files that described the administrative distribution of neighborhoods were searched. link

- Information on the location and price of rental apartments. Data from an already preprocessed collection of Airbnb were used. These data continue the price, the exact location and the neighborhood to which each one belongs, among other data. link

- Information about the places of interest in each neighborhood. Searches were made through the Foursquare API.

As an example we obtain the following tables, in the next point the procedure will be detailed.

	coddistbar	Neighborhood	Latitude	Longitude
0	171	BENIFARAIG	39.52564	-0.38462
1	161	BENICALAP	39.49301	-0.3910

Table 1. Fragment of data on the position of neighborhoods.

name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price
VALENCIA HISTORIC HOUSE 50M BEACH	7093832	Francisca	POBLATS MARITIMS	LA MALVA-ROSA	39.47553	-0.32461	Entire home/apt	150

Table 2. Excerpt from neighborhood rental data.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
AIORA	39.46573	-0.34336	Jardines de Ayora	39.46814	-0.34340	Playground
AIORA	39.46573	-0.34336	Bar Mochuelos	39.46614	-0.34794	Mediterranean Restaurant

Table 3. Fragment of the venues of the neighborhoods.

Data cleaning

First of all, the files that contain the position of the neighborhoods are in geojson format, this specific file uses the UTM 30 coordinate format for spatial location. These were projected into the WGS84 format because Foursquare and the "folium" map display library work with that coordinate format. This was accomplished through two libraries:

- pyproj, used to project individual coordinates.
- geopandas, used to create a new geojson file with the appropriate format.

Next, reference coordinates for each neighborhood were obtained from the central position obtained through the "shapely" library. These data will be used later to obtain the places of interest.

Regarding rental housing, it was grouped by neighborhoods and both the average rental price and the number of apartments per neighborhood were obtained.

At this point, both dataframes were joined, obtaining in the same table the location data of the neighborhood and the average value of the rental price and the number of homes per neighborhood.

It was discovered that the Benifaraig neighborhood lacked any rental housing, because the table with the grouped description of rental housing had one row less than the table with information on neighborhoods. It was decided to dispense with that neighborhood for the rest of the project.

The following figures represent the price per neighborhood, and the number of houses per neighborhood respectively. The darker the color, the greater the value.

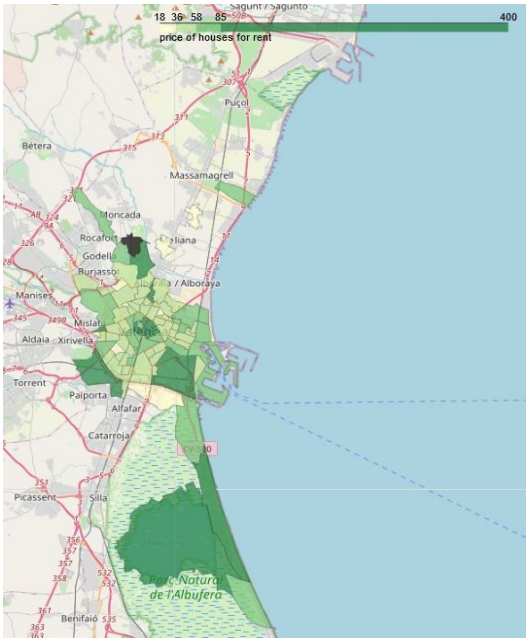

Figure 1. Map of Valencia with the price colored by neighborhoods.

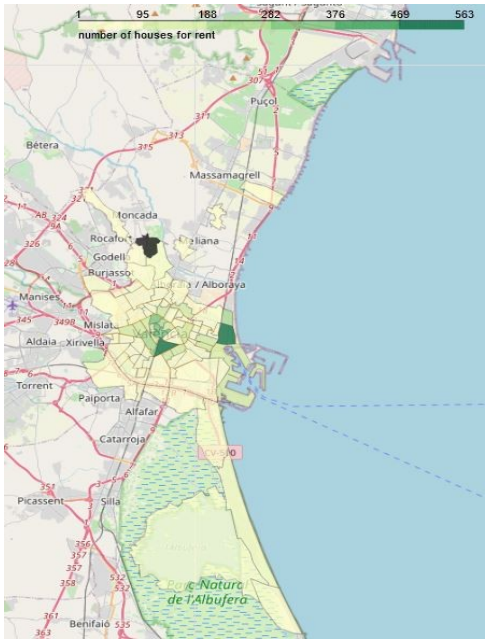

Figure 2. Map of Valencia with the number of tourist rental houses colored by neighborhoods.

The previous choropleth maps were obtained through the "folium" library.

To end this section, comment that data from the entire areas was obtained through Foursquare. Comment that due to limitations the maximum number of venues per request is

100. To answer the questions posed at the beginning of the project, it was necessary to create two data sets, a first set with a reduced search radius that would serve as an approximation to determine how many venues there was by neighborhood, and a second set of data without setting the radius that returned the closest 100 venues to answer the rest of the questions.

Methodology

Analysis

We will start by creating a table that contains the price and number of rental properties per neighborhood. We will use the Foursquare data obtained with a smaller radius. For the first analysis, we will accept that the venues are homogeneously distributed by neighborhoods as a valid assumption. This is because neighborhoods do not have a circular shape and Foursquare offers certain limitations to limit fully mapping an area, for example, limit maximum of 100 results per request. If we use a larger radius we will always saturate and all neighborhoods will have the same amount of venues. We seek to obtain a representative value of the neighborhood.

Once the data has been grouped by neighborhood and the quantity of venues has been counted, we proceed to join the data with the average price table by neighborhood. We create a new column with the division between the price and the number of homes with which we obtain the desired ratio. We order the table by ratio obtaining the following results. Comment that the feature count is the number of venues per neighborhood.

Neighborhood	price	price_std	number	coddistbar	Latitude	Longitude	count	ratio
RUSSAFA	64.60640	41.75456	531	21	39.46119	-0.37370	65	0.99394
LA GRAN VIA	65.25806	43.86901	62	23	39.46554	-0.36552	59	1.10607
BENIMACLET	39.93913	35.14414	115	141	39.48460	-0.35940	34	1.17468
CIUTAT JARDI	58.80597	42.10133	67	132	39.47198	-0.34504	44	1.33650
LA VEGA BAIXA	58.64000	40.43913	25	134	39.47631	-0.34997	43	1.36372

Table 4. Table with the best neighborhoods by price / number of venues ratio.

Next we will return to the Foursquare data calculated without radius. This is due to the assumption that tourists choose housing according to the proximity to the areas of interest, regardless of what areas are in the neighborhood chosen to reside.

The variable of interest is the type of venue, as it is a qualitative variable and must be dummified. It is then grouped by neighborhoods, obtaining the average for each new type.

At this point we can obtain the predominant type of venue for each neighborhood. Next we put together the data obtained with the table that contains the average price and the number of homes per neighborhood.

The obtained table has as independent variables the different types of venues, and as dependent variables price and number of homes. Now we can explore the data and try to answer the rest of the questions at the beginning.

The following figure represents the quartiles and outliers of the dependent variables.

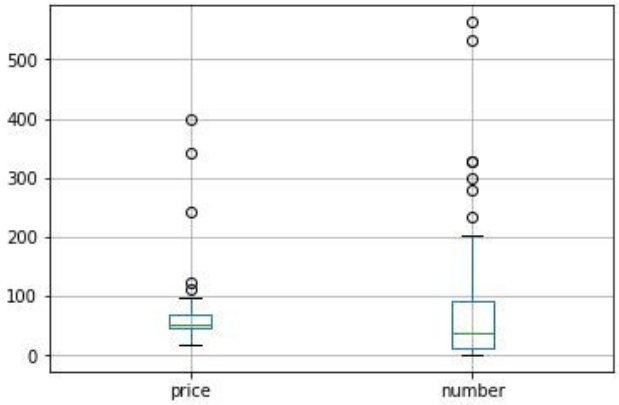

Figure 3. Boxplot of the dependent variables.

Next we obtain the correlation between all the features and we represent them in the following figure.

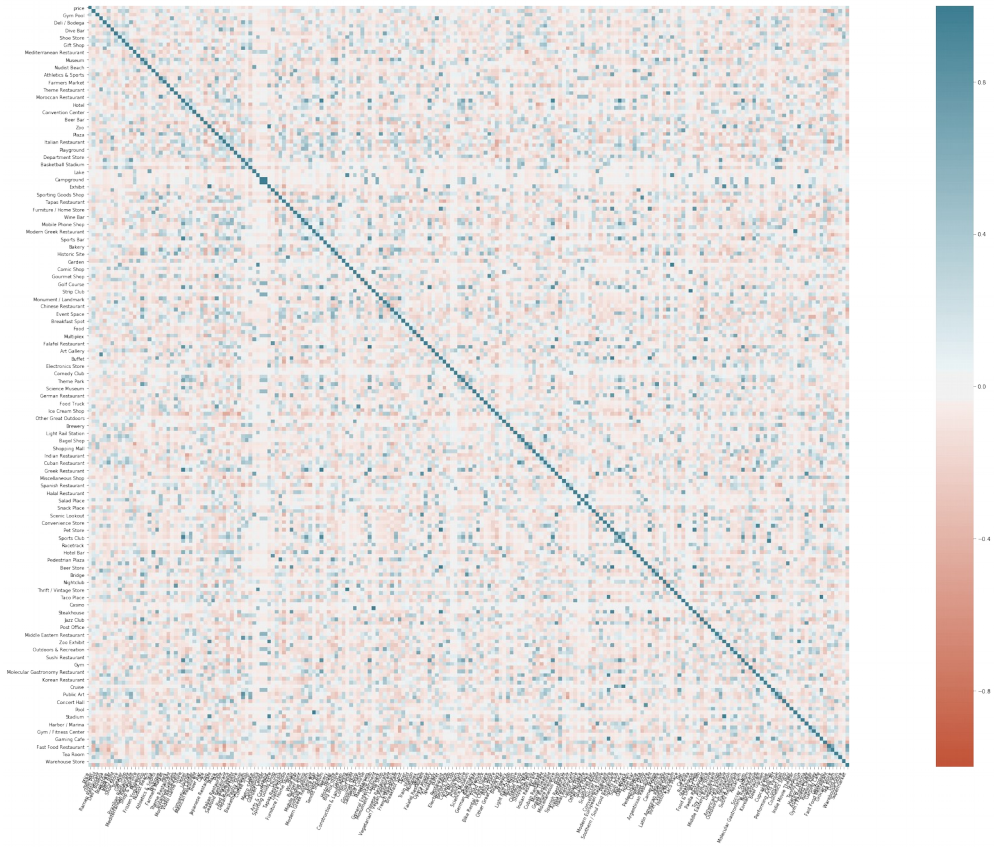

Figure 4. Heatmap with the correlation between features.

As there are a large number of features we proceed to focus on the dependent variables. First we order from highest to lowest and show the features with the highest correlation.

Gastropub	0.49969
Camera Store	0.47490
Cheese Shop	0.41811
Monument / Landmark	0.41089
Gourmet Shop	0.40389
Dog Run	0.40192
Bike Rental / Bike Share	0.38243
Plaza	0.37725
Gift Shop	0.37633

Table 5. Features with the highest correlation with the number of apartments

Clothing Store	0.31273
Gym Pool	0.28880
Supermarket	0.25571
Toy / Game Store	0.24057
Hotel	0.22697
Shopping Mall	0.22180
Food	0.22096
Breakfast Spot	0.19167
Train Station	0.18918

Table 6. Features with the highest correlation with the price

Regression analysis

Neither feature shows a high correlation with dependent variables. To continue, we will apply linear regressions to determine if the set of independent variables has predictive capacity.

First, we will represent the independent variables against the dependent variables. For this we will apply a dimensionality reduction through PCA. This technique transforms the variables by creating a new set of orthogonal variables between them.

Comment that before applying PCA the data must be normalized. In this case, the methods provided by the sklearn library were used.

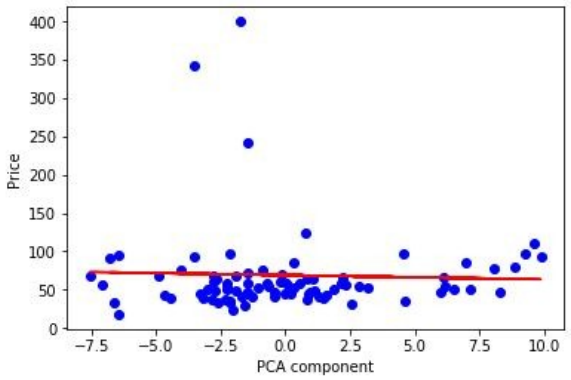

Figure 5. Second component of PCA versus price

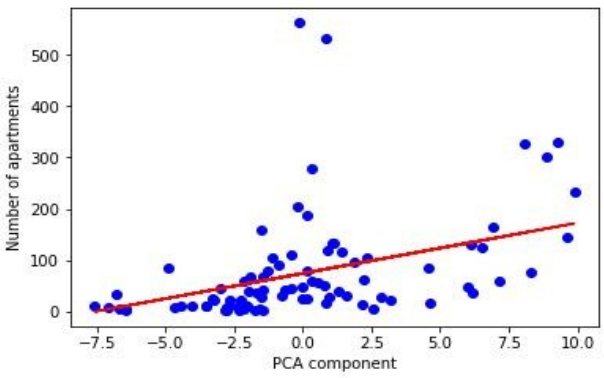

Figure 6. Second component of PCA versus number of apartments

In addition, a linear regression was applied in both cases. The data was divided between test and train with 20 percent of the first. The line was drawn on the scatter to observe the result.

It might seem that it offers a good result but when calculating the R2 score in both cases it gives 0.

Finally, a multiple linear regression was applied. But only with the 10 features that offer better correlation. Obtaining in this case:

R2 score for the price: -1.29

R2 score for the number of tourist rental apartments: 0.78

Comment that the R2 score should only be between 0 and 1 (1 the best case), but due to the computational effect it can be negative. In this case it is considered 0.

For the calculation of the best neighborhood to rent a flat, we apply the regressor obtained from the number of apartments due to the fact that it offers an acceptable R2 score. We calculate the error of all the neighborhoods and select the one with the greatest difference in favor of the regressor. In other words, the regressor anticipates that the number of dwellings should be greater. The neighborhood with the highest value is L'illa perduda.

Results and Discussion

This project has shown that tourist rental apartments are not distributed equally among the different neighborhoods, the same happens with the price of such rentals.

With the data we obtain we can answer the questions posed at the beginning of the project. By calculating the number of venues per neighborhood we have obtained that the neighborhood with the best price / number of venues ratio is Russafa.

In the calculation of this coefficient we have found the limit of the maximum number of venues returned per request. The way in which we have solved the problem is by taking a smaller radius.

With the correlation between variables it has served to demonstrate that there is no type of venue that has a high correlation on its own.

But with the application of linear regression, on the features with the highest correlation, it has been determined that the number of apartments for rent has an R2 score of 0.78 and the R2 score of the regression applied to the price has an R2 score of 0 (less than zero for computer effects). To achieve the score it was necessary to apply a reduction of features choosing only the most relevant ones. Among the most relevant are monuments, gift shops and plazas, which are typical places for tourists.

Regarding the price, it is considered that it does not have an evident relationship with the used features. One possible explanation is that the price is more linked to the size of the house, type of furniture, and minimum number of nights than to the proximity to places of interest.

With this, two more questions are answered, only the number of houses for rent is related to the type and number of venues in the neighborhood.

Regarding the last question, the neighborhood that offers the most opportunities is L'illa perduda, it is the neighborhood that offers the least number of rental properties than the linear regression proposes.

Conclusion

To answer the questions asked at the beginning, information was collected from various sources, data exploration procedures, selection of features by correlation and linear regression analysis were applied to reach the conclusion that only the number of rental homes has a relevant relationship with the type and number of venues in the neighborhood.

As the main conclusion we can say that the number of supply (and therefore demand) is linked to the proximity to areas of interest. On the other hand, the price is linked more to factors not measured in this project (example size of house).

Finally, with the information contained in this project, clients could reconsider the choice of destination, and the owners could see which neighborhoods are saturated.

As possible lines of future development it would be interesting to include greater diversity in the type of features used, such as proximity to public transport or crime rates that could affect the attractiveness of the neighborhoods.