Predicting the best zone for investment in an apartment for rent

Emanuel Fitta

1. Introduction

1.1 Background

Investing in an apartment to later rent it is an investment with almost certain returns because year after year the number of people looking for a place to live increases mainly in large cities such as Mexico City.

Coyoacan is a delegation from Mexico City that is growing economically every year. In addition to this, the culture that exists in this place is attractive to encourage people to live here. Coyoacan is a place full of history and traditions, it has the neighborhoods of the town of Los Reyes, San Pedro Tepetlapa, La Candelaria, San Francisco Culhuacán, Santa Úrsula Coapa, San Francisco, El Niño Jesús, La Conchita, San Lucas, Santa Catarina. For these reasons, any company with sufficient capital should consider building apartments in this zone.

It is convenient for any investor to find apartments with the lowest sales prices, located in an area with the highest rental prices. This would undoubtedly achieve that the investor obtained greater profits in less time. In this project with the power of the machine learning we will found this zone.

1.2 Problem

The problem to find this zone is that there are many variables to consider, for example the size of the apartment, the number of rooms, the number of bathrooms, if there are some coffee shops near to the apartment, if the apartment is near to an avenue, etc. Is not easy with all these variables to find out the best zone for investment. With a machine learning algorithm, we will consider all of these features to find the best choice to investment.

2. Data acquisition

2.1 Data sources

Mercado libre is a page designed for the sale of various products, it contains information on apartments for sale and rent in different areas of Mexico.

We applied scrapping methods to this page to obtain information about apartments for rent and for sale in Coyoacan. Also, we obtained information about the places of interest near to each apartment through Foursquare API.

2.2 Data cleaning

When scrapping on the Mercado Libre page, a large amount of null data was obtained, mainly in the column corresponding to the years of antiquity. This column contained inconsistent information, so it was decided to delete this column and not consider it as part of the analysis.

Likewise, the rest of the columns had to be cleaned in order to be transformed to a numeric type and to be able to carry out machine learning models with them. The target column, that is, the prices column, these were separated by thousands by commas, so we had to eliminate those commas.

On the other hand, the size column had each quantity accompanied by the unit, m ^ 2, so this was also eliminated to leave only the figure corresponding to the size.

The null values were filled with zeros, with all this it was possible to transform the size, rooms, bathrooms, prices columns to a numerical.

Later, the latitude and longitude of each department was obtained and added to the dataset. For many of the locations it was not possible to obtain the latitude and longitude, so this information was filled with zeros. And the information that contained the latitude and longitude was separated from the dataset to make an analysis with this subset regarding the location of the department. Likewise, through the Foursquare API, all the information corresponding to the places of interest in Coyoacan and close to each department was obtained, one hot encoding was applied to add said information to the dataset and be able to process it through the classification algorithm

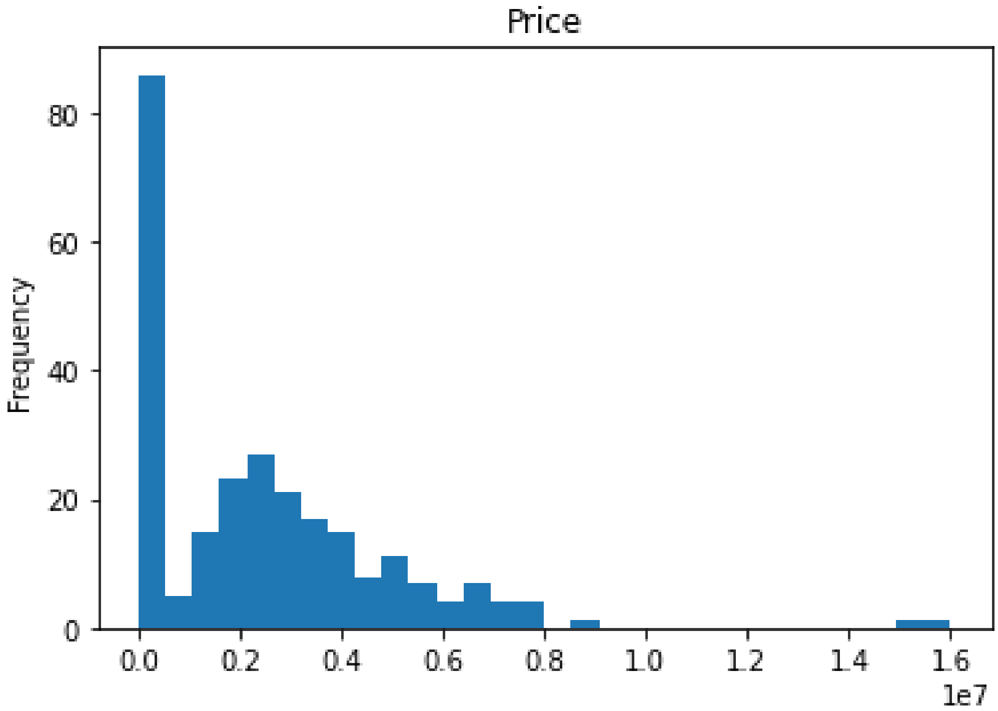
2.3 Feature Selection

We first consider predicting prices without considering location, that is, based on size, number of rooms, number of parking spaces, and number of bathrooms. When analyzing the correlation matrix, we realize that the only independent variable that is strongly correlated with price is size, the rest are redundant. However, we intuit that the price not only depends on the size but also on the location. Therefore, they were also considered as independent variables, in addition to those already mentioned, whether there was a place of interest near the department.

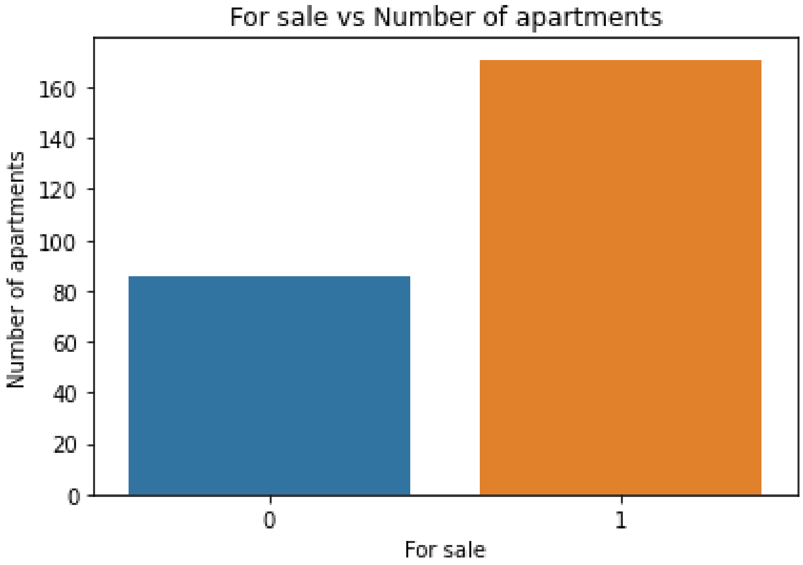
3. Exploratory data analysis

3.1 Target variable

We have considered the column prices as the target variable.

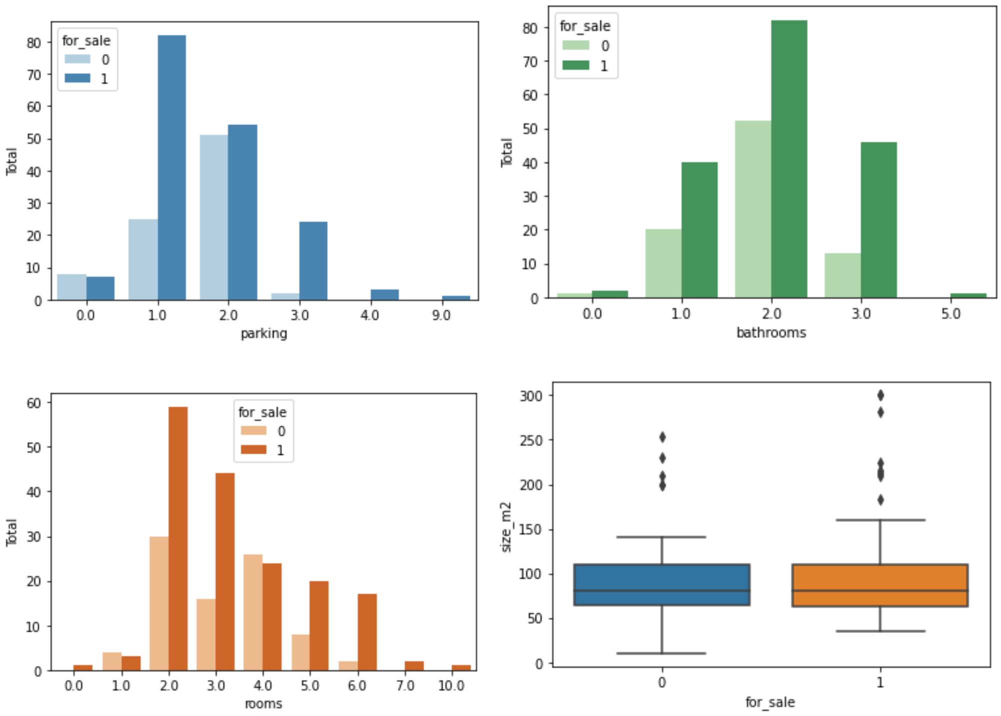


We can see that there is a great variance in the price column, this can be because, in the data are included apartments for rent and for sale, and obviously the prices for both are extremely different. So, we aggregated a label to identify if the apartment is for rent or for sale. For that we have considered that if the price is under 10000 is for rent and for sale otherwise. With this we can see the number of departments for both categories:



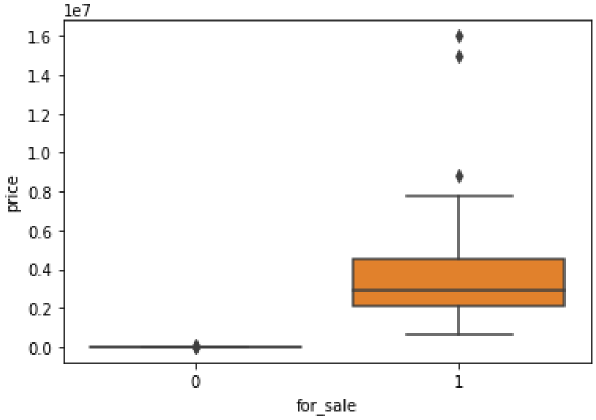
3.2 Main characteristics of the departments

We have analyzed the behavior of the main characteristics of a department for both classes.

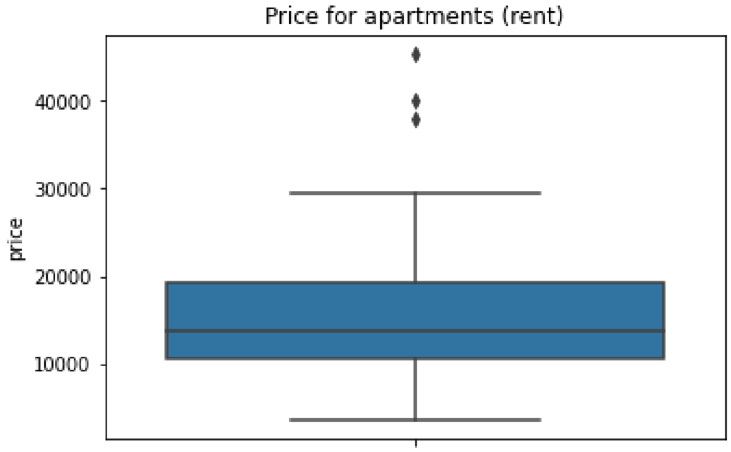


We can see that the most of apartments have 1,2 or 3 bathrooms. Also, we can see that the most of apartments have 1, 2 parking places. We can see that most of the apartments for sale have 2 or 3 rooms. While the number of rooms for apartments for rent is distributed more evenly between 2 and 5 rooms. The size of apartments is in the same range for both, rent and sale. There are some outliers, these will be consider in the analysis.

When analyzing the price of the apartments we can see that in effect, the price varies a lot between the rental and sale prices.



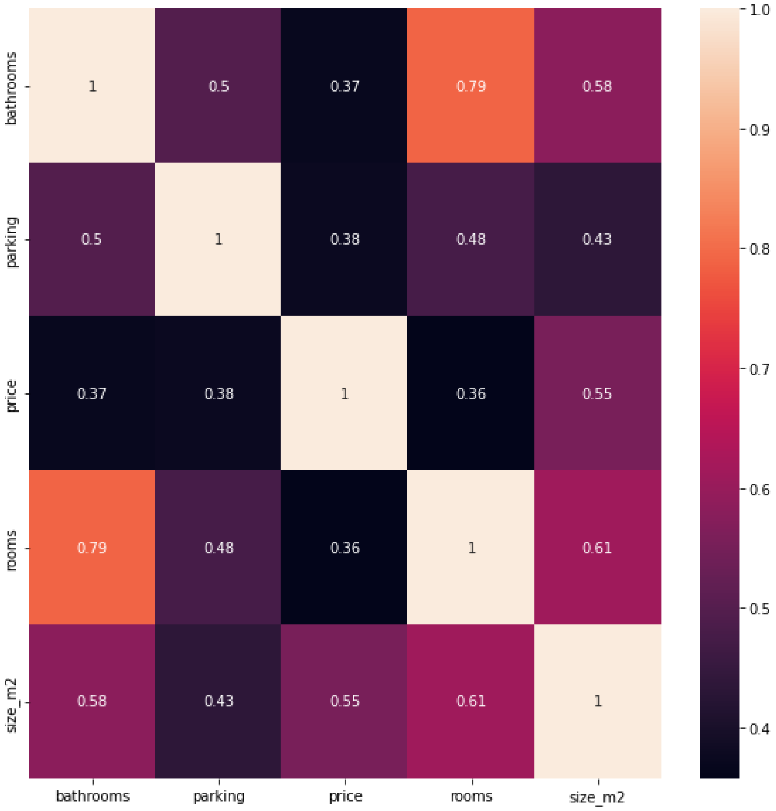
For this reason we made a boxplot for the rental prices



3.3 Correlation

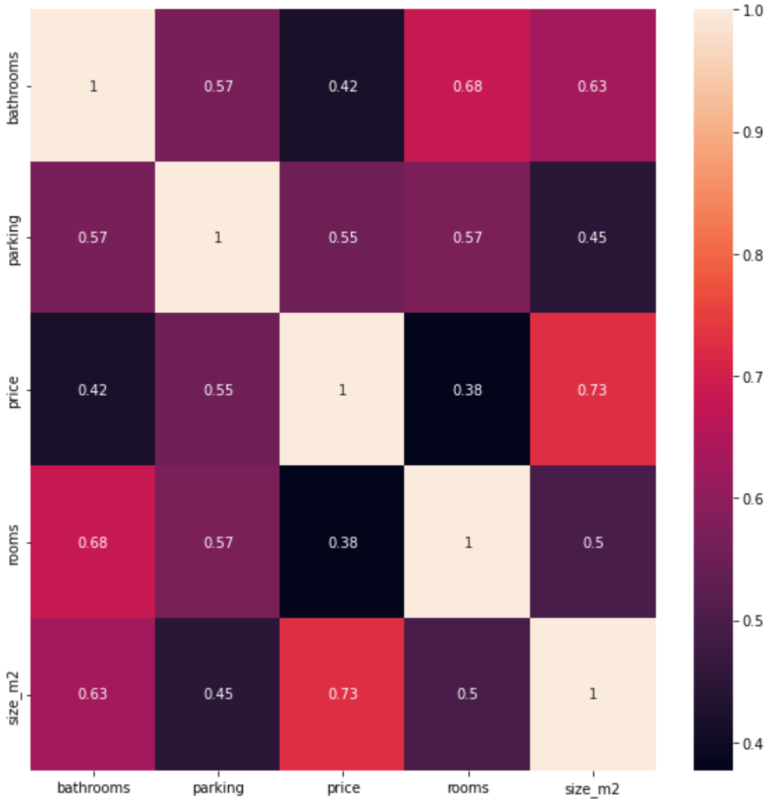
We have obtained the matrix correlation for both categories in order to observe the redundant variables and to know which is the variable that would be worth considering for the machine learning model.

a. Apartments for sale



We can see that the target variable (prices) is not strongly correlated with any other characteristic, which could indicate that the price depends on the location and not on the number of rooms or parking places, etc.

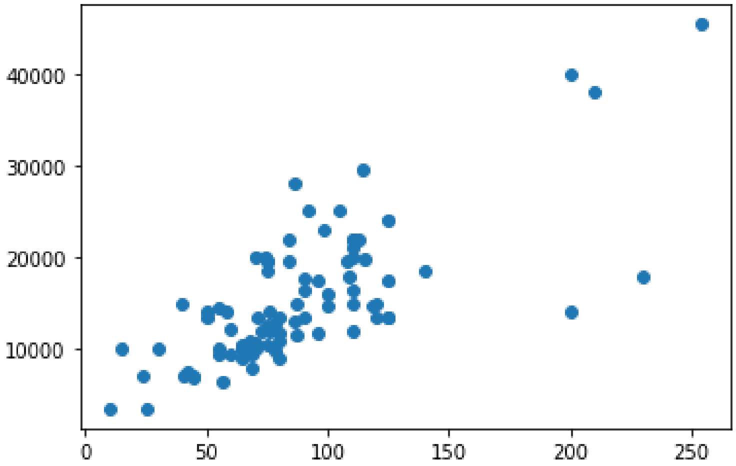
b. Apartments for rent



We can see that prices have a strong correlation with the size of the apartment, we will consider it for a machine learning model.

In both cases we can see that the size, the rooms and the bathrooms are redundant, that is why we made a machine learning model only for rental apartments considering size as a characteristic.

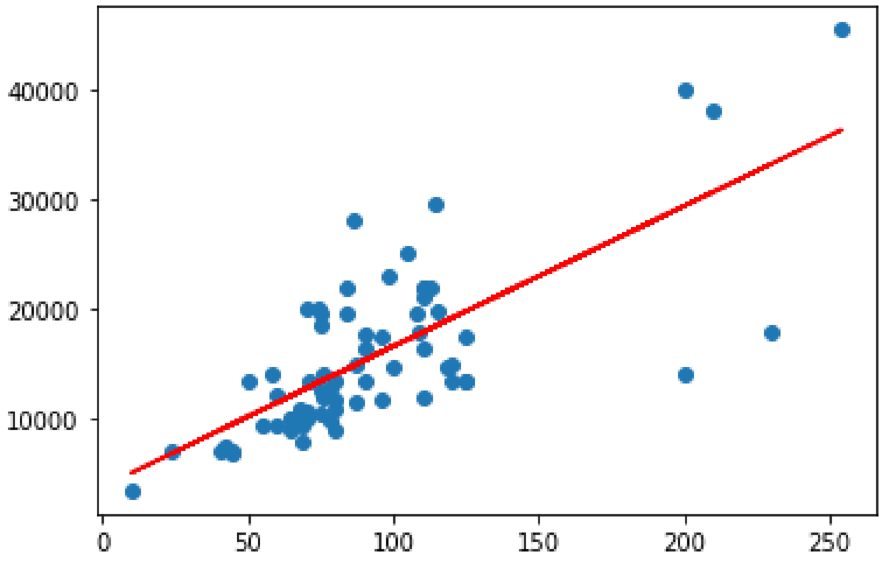
3.4 Relation between size and price for rental apartments



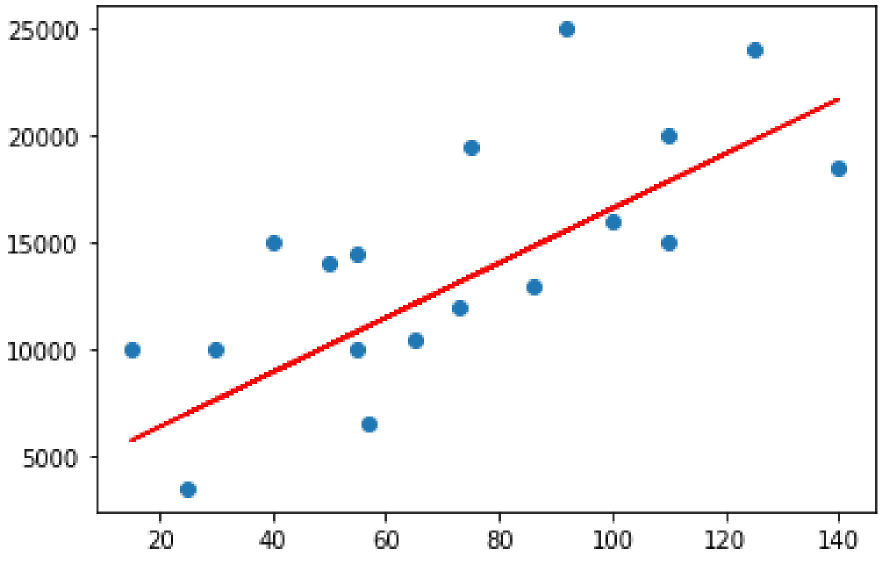
The graph above suggests that the relationship between size and price is a linear relationship. Therefore, we carry out a linear regression.

4. Model without location

We take 20% of the data as a test set. When training the linear regression model, it was obtained

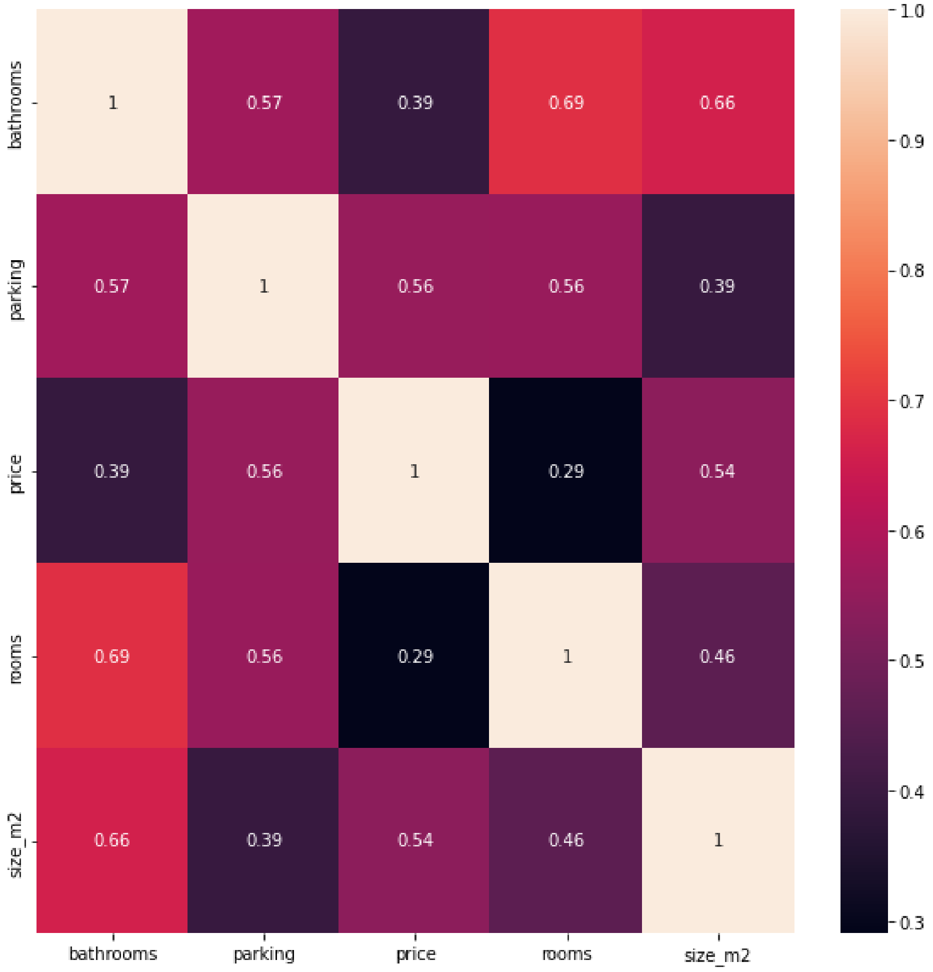


When applying the model to the test set, it was obtained



In both cases the model has a low score, 0.53 in the case of the training set and 0.45 in the case of the esting set that indicates that the linear model is bad, this is obvious from the graphs.

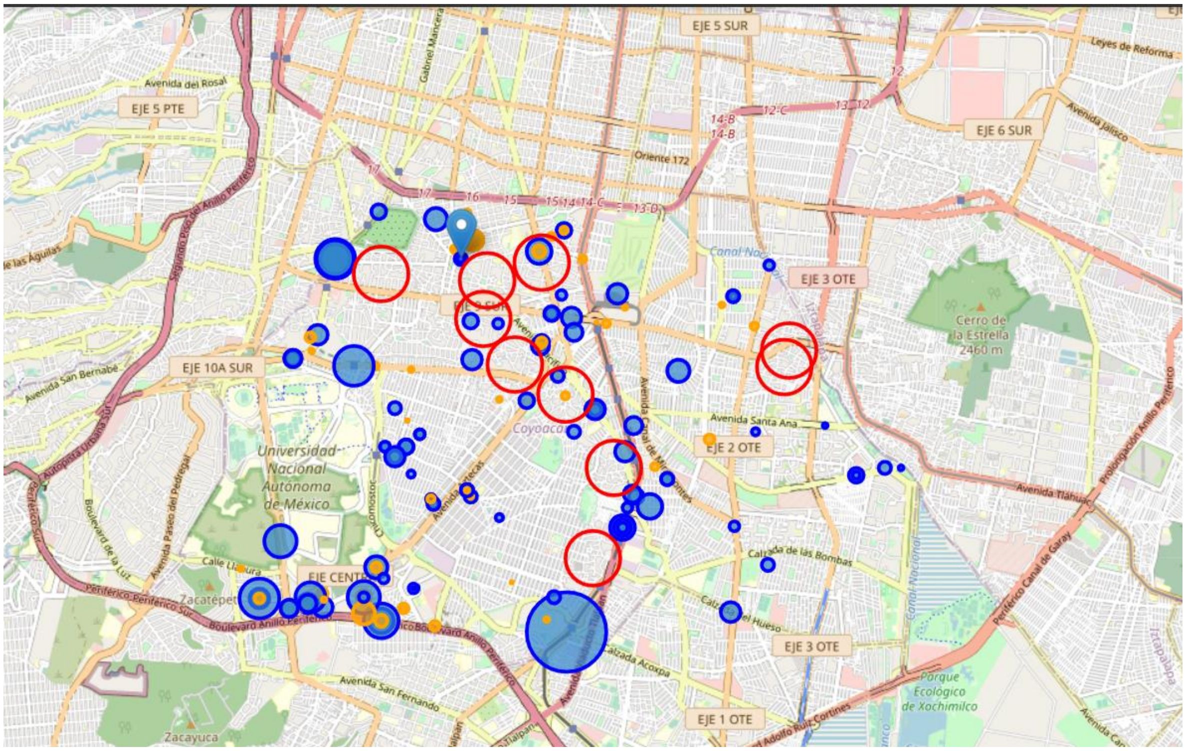
So, the strong correlation found in the previous section could be because there are some outliers. In fact we can see that without outliers there is not correlation between the price and the features. This is in both cases, rent and sale apartments. This suggest that the price is in relation with the location in both cases rent and sale.



5. Analysis considering location

Given what we found in the previous section, we can see that the price of the apartments does not have so much to do with size. This makes us think that the price has more to do with the location and places of interest in the area where the apartment is located.

5.1 Viewing both apartment categories



- In this map the blue circles are the sale apartments, and the orange circles are the rent apartments. The size of the circles is in proportion with the price. The red circles are the principal neighborhoods of Coyoacan. We can see that the biggest circles are near to the big avenues like 'Calzada de Tlalpan' or 'Circuito interior'. This makes us think that the idea about the price is in relation with the location is good.

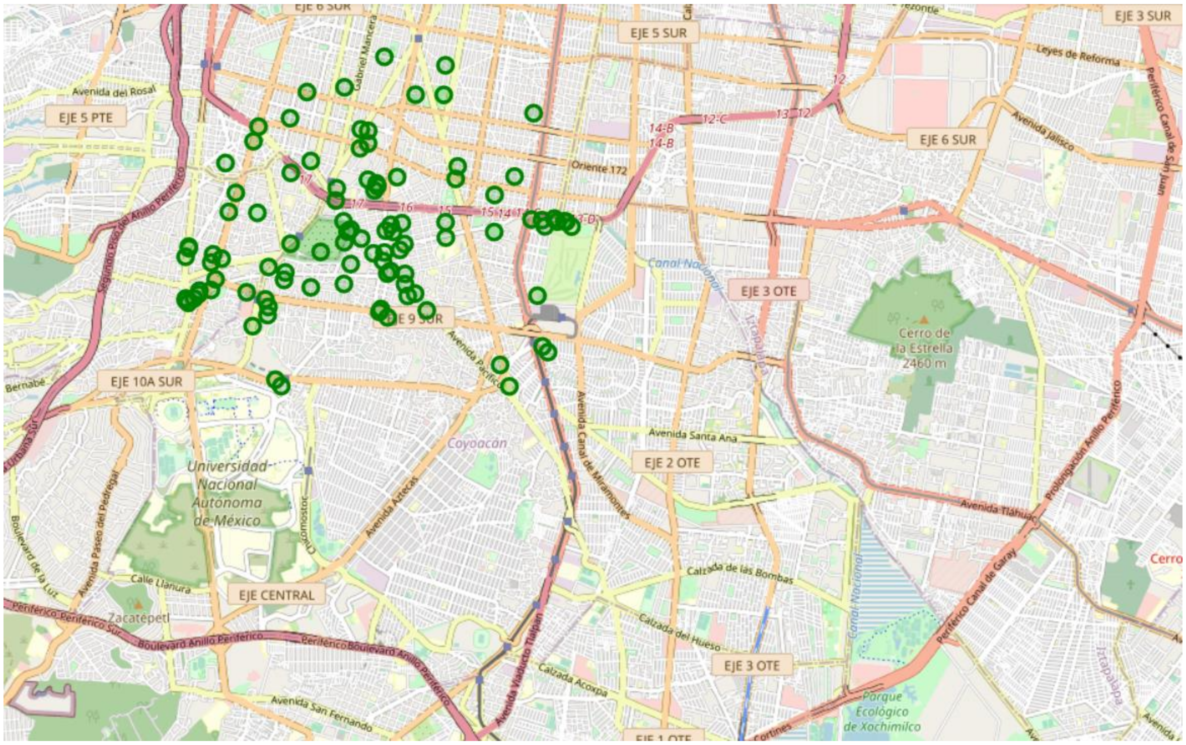
5.2 Places of interest

At Coyoacan there are the following categories of places of interest:

'Plaza', 'Coffee Shop', 'Market',
'Mexican Restaurant', 'Spanish Restaurant', 'Ice Cream Shop',
'Art Museum', 'Café', 'Burger Joint', 'Bakery', 'Track',
'Italian Restaurant', 'Flower Shop', 'Food Truck',
'Breakfast Spot', 'Park', 'Mediterranean Restaurant',
'Garden Center', 'Athletics & Sports', 'Indie Movie Theater',
'Art Gallery', 'Concert Hall', 'Sports Club', 'Department Store',
'History Museum', 'Pet Store', 'BBQ Joint', 'Movie Theater',
'Shopping Mall', 'Multiplex', 'Middle Eastern Restaurant',
'Bookstore', 'Argentinian Restaurant', 'Golf Course', 'University',
'Chocolate Shop', 'French Restaurant', 'Performing Arts Venue',
'Japanese Restaurant', 'College Arts Building',
'Sporting Goods Shop', 'Food Court', 'Music School', 'Theater',
'Brewery', 'Jazz Club', 'Chinese Restaurant', 'Seafood Restaurant',

'Vegetarian / Vegan Restaurant', 'Liquor Store', 'Spa',
'Southern / Soul Food Restaurant', 'Dessert Shop', 'Steakhouse',
'Music Store'.

As we can see there are a large number of places that could be attractive for people looking for a place to live.



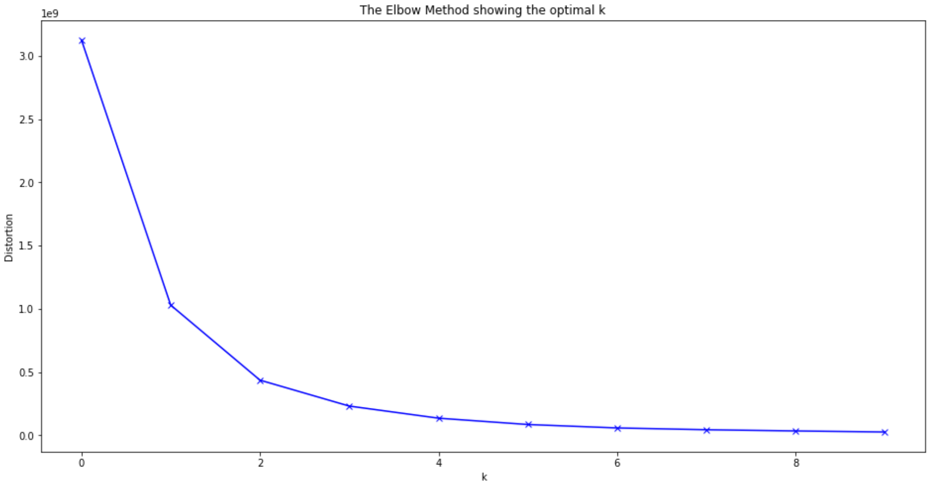
The most of the interesting places are located at north of coyoacan. We are going to analyze if this is in relation with the price of the apartments.

5.3 Classification Model (rent)

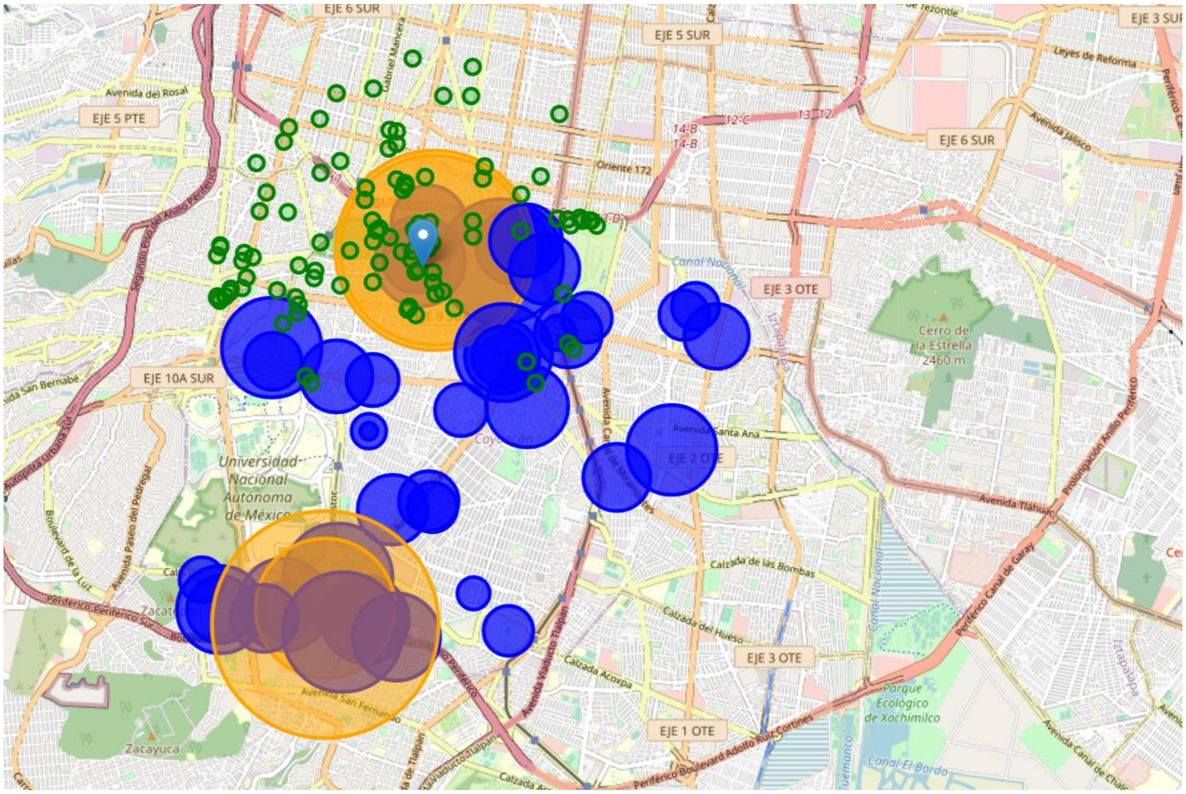
We have applied a clustering algorithm on the data in both categories (apartments for rent and for sale).

a. Rent

We are going to apply the elbow method to obtain the optimal number of clusters.



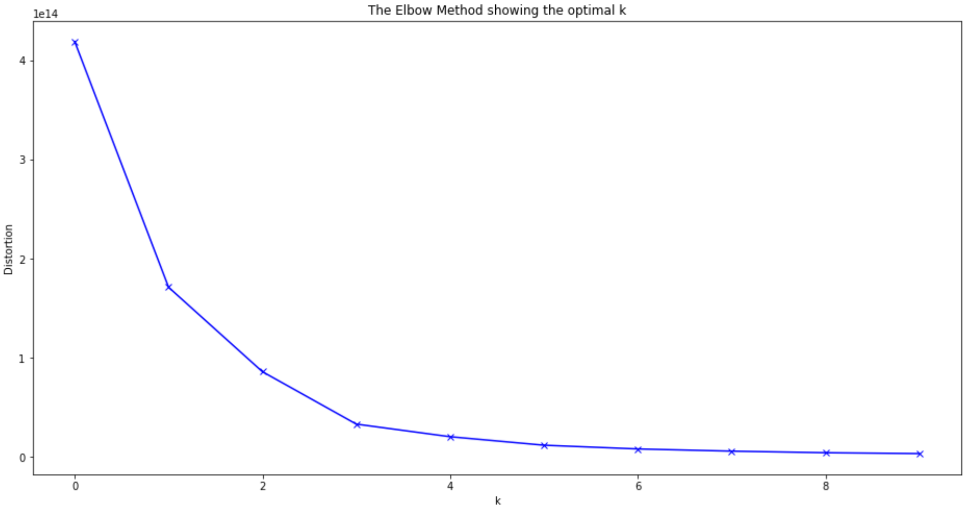
The optimal number of clusters is 2.



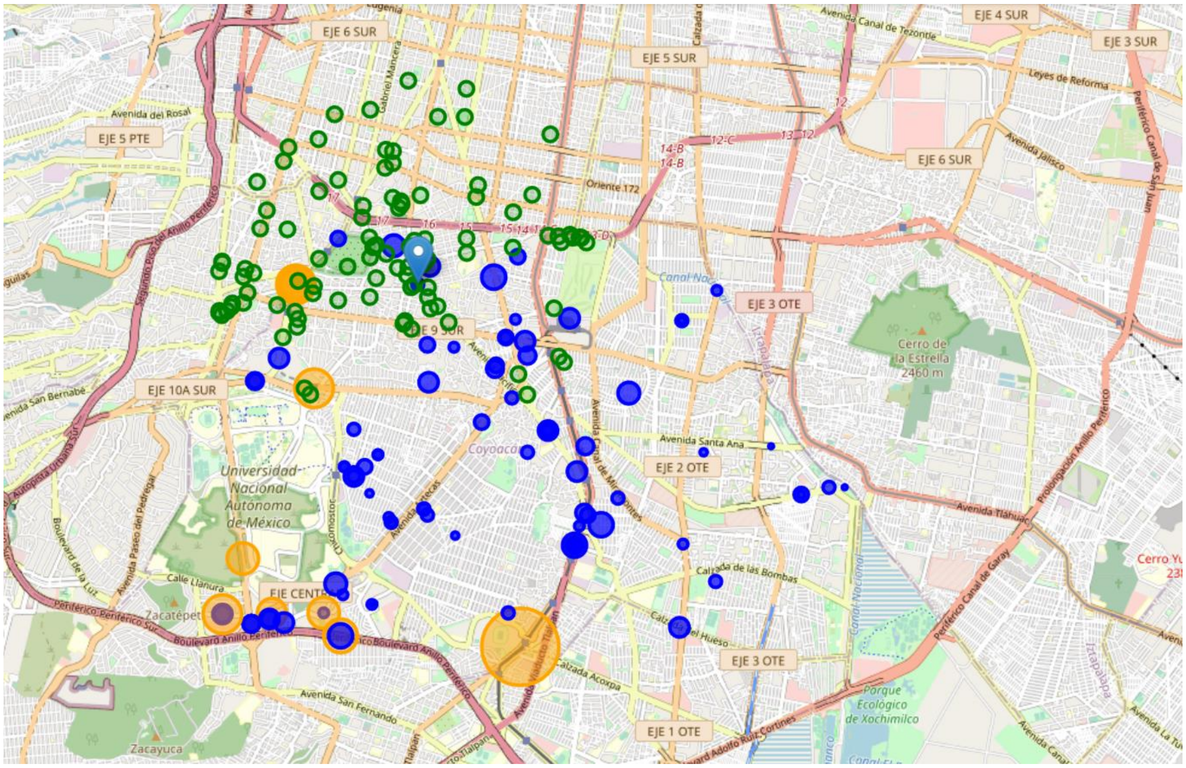
In this the blue circles are the label 0, the lowest prices for rent and the orange circles are the label 1, the highest prices for rent. We can see the the highest prices are at north and south near to great avenues. We are interested in the apartments located at north because here are the interesting places.

b. Rent

Same as in the previous case we are going to apply the elbow method to obtain the optimal number of clusters.



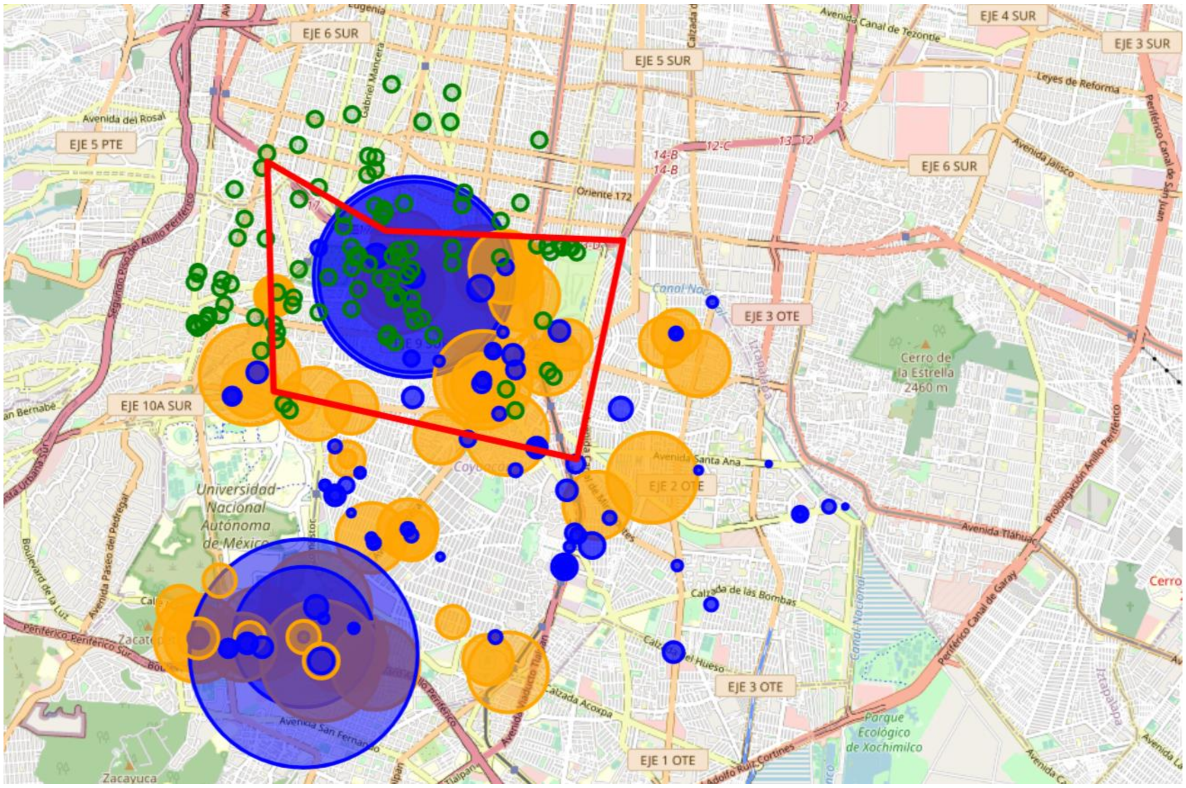
The optimal number of clusters is 2.



In this the blue circles are the label 0, the lowest prices for sale and the orange circles are the label 1, the highest prices for sale. We can see the lowest prices are at north. This is so good because at this zone are the apartments with the highest rental prices, the perfect combination. This coincides with the location of the interesting places. Our theory was right, the best zone for invest in an apartment for rental is in correlation with the location.

6. Merging information and Conclusions

Next we put the information for both categories on a single map, so that we could visualize those departments that are in the area with the best investment opportunity, this was sought in the area where the sale prices of the apartments were low , but with a high density of apartments with high rent prices and that were also located in an area near the north where most of the places of interest are located.



In the last map, as in the previous ones, the green circles indicate places of interest (restaurants, cafes, etc.). The blue circles are points where there are apartments for sale with the lowest prices and apartments for rent with the highest prices. While the orange circles are points where there are apartments for sale with high prices and apartments for rent with low prices.

In the red polygon, there would be a great investment opportunity, since it presents a great density of places of interest, at the same time that it presents a great density of blue points. Which indicates that you could buy apartments at the lowest prices in the area and put them up for rent at the highest prices in the area. In addition, the size of the orange circles is medium, which indicates that also, in this area, the highest prices for sale, these are not the highest and the lowest prices for rent, these are not the Lower. For these reasons we believe that this area presents the best investment opportunity for someone who wants to invest in real estate.