

SEGMENTATION OF HOUSES FOR SALE GIVEN THE OWN CHARACTERISTICS OF THE PROPERTY FOR A COMPANY IN GUAYAQUIL, ECUADOR

Jimmy Joel Landín Casal



Table of Contents

Abstract.....	2
Introduction	2
Data Section	3
Limitations.....	3
Methodology Section	4
Results Section	6
Discussion Section	7
Conclusion Section	8



Abstract

In this study we will see a segmentation analysis using an unsupervised algorithm such as Kmeans, applied to a requirement for segmentation of houses for sale by a real estate company in Guayaquil, Ecuador; making use of tools such as the Foursquare API and multiple Python libraries. All this with the objective of studying to provide a tool that allows generating greater performance and benefit to customers when choosing a each to buy.

Introduction

A recognized real estate company from Guayaquil, a city in Ecuador, has an important need, which is to have a tool that tells it based on the Client's requirements such as House Size, Price, Crime Rate, Nearby Places of Interest, among others, the right place or the right places to recommend a property to the client. It may be the case that the client feels comfortable with the size of the house, number of rooms, among others, but it requires that it be close to cafes and restaurants, since it is a very busy family that always has lunch and dinner outside of the House; or at the same time that you require tourist centers such as parks or museums, since you have young children and they want a distraction center.

In short, the objective is to have a list of options that allow you to decide which house to stay with given a segment of properties that meet the characteristics required by the client. This analysis will not only save time in the search for properties by the real estate company, since it would have to see nearby places of interest one by one in some tool, but it also reduces costs and generates higher returns by presenting possible properties to the client immediately and thus be able to continue with other clients. In addition to the marketing that can be done with this tool, arguing that it is one of the best performing companies in the Ecuadorian market when it comes to finding suitable properties for the end customer.

Data Section

In this project, a csv of data collected from the olx.com.ec platform will be used, which is a website that performs digital commerce like amazon or ebay, but it is more local, limiting itself to national shipments within of the countries where it is. Information was obtained at the beginning of the year (Cut January 25, 2021), a sample of the houses for sale that OLX users have published, you can almost always find the location, price and size of the land in which the house is built.

Said csv contains the following fields:


- **Code:** Unique code generated by each house for sale registered in CSV (Code generated fictitiously to generate a primary key).
- **Sector:** Name of the neighborhood where the house for sale is located.
- **Latitude:** Latitude of the house for sale.
- **Longitude:** Length of the house for sale.
- **Size:** Land in square meters on which the house is built.
- **Price:** Price published by the owner / owner or seller of the property.

	Code	Neighborhood	Latitude	Longitude	Size	Price
0	AC247	Acacias	-2.2417	-79.9008	309.0	110000.0
1	AC246	Acacias	-2.2467	-79.8966	308.0	145000.0
2	AC129	Acacias	-2.2465	-79.9026	194.0	185000.0
3	AR308	Acuarelas del Rio	-2.1299	-79.8809	170.0	165000.0
4	AR380	Acuarelas del Rio	-2.1351	-79.8825	230.0	176000.0

Limitations

This database, which is a sample of the houses published in OLX, has certain limitations which will be listed:

1. In the first place, it is a sample in time of the houses published, which is updated day by day, and trends may change. There should be an automatic clustering model that generates the groups instantly with the information of the houses for sale at present, but for the purpose of study it will be done with this csv.
2. In addition, both the land and the price are not measured by an appraiser, these measurements can be biased by the seller of the property, for example, a person can put a price much higher than the market price of the property, or the land that they put as a



property feature may be greater than it actually is. There are also certain times in which by Ecuadorian culture the price is much higher, because you bargain first since the price is negotiable.

3. Finally, another major limitation is the lack of global information on houses for sale, it would be necessary to take into account a master base of all the houses for sale currently in Guayaquil. As well as the large amount of missing data, the number of rooms, bathrooms, etc. could be put into the model. But most vendors do not fill in the relevant information and that is why these variables are not considered in the study. Once you have a much richer base for the city of Guayaquil with all the necessary fields, you will have a better perception of reality.

Another data source that will be used for this project is the nearby places of interest provided by the Foursquare API, which, for each location of the houses for sale, will have locations such as coffee shops, parks, among others. These two data sources will complement each other to solve the problem of segmenting the houses for sale in Guayaquil and to be able to solve the needs of the clients of the real estate company.

Since there is no crime rate registry for each sector of the houses for sale, that data source will not be used, crime studies in the city of Guayaquil are very limited to certain specific sectors of Guayaquil, but not to the whole city in general. This can be a great starting point for a second phase in this project.

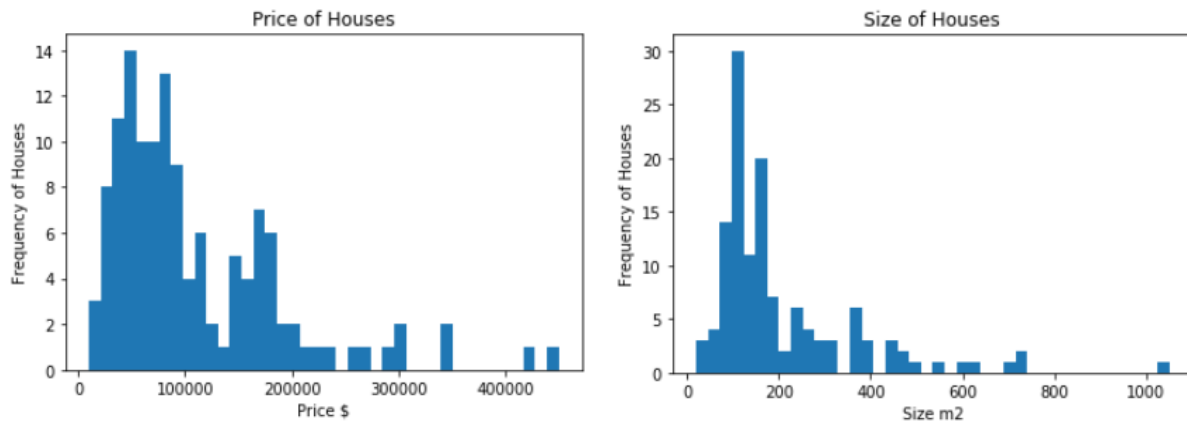
Next, the database of records of houses for sale will be refined to continue with the study, in addition to making the respective connection with the Foursquare API and segmentation of the places.

Methodology Section

First of all, a data preparation was made, leaving only those structured records without null values, that is, they may be empty because of the user who put the house up for sale. It was observed that the data was consistent and did not have subsequent problems that could harm the model.

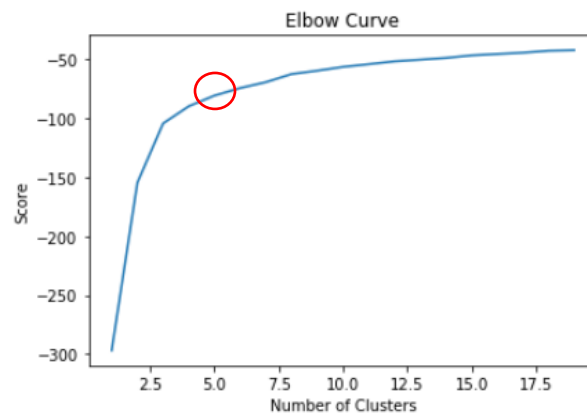
As a later step, the places near these houses and a radius of 500 meters were added through the Foursquare API, generating more enriched information.

You can also observe some histograms of the Price and Size of the Houses for Sale, it can be observed that there are often high prices of houses when they reach prices such as \$ 80,000 and \$ 180,000, thus defining two economic strata of houses for sale in Guayaquil. In turn, it can be seen that the size of the houses in square meters is around 160.



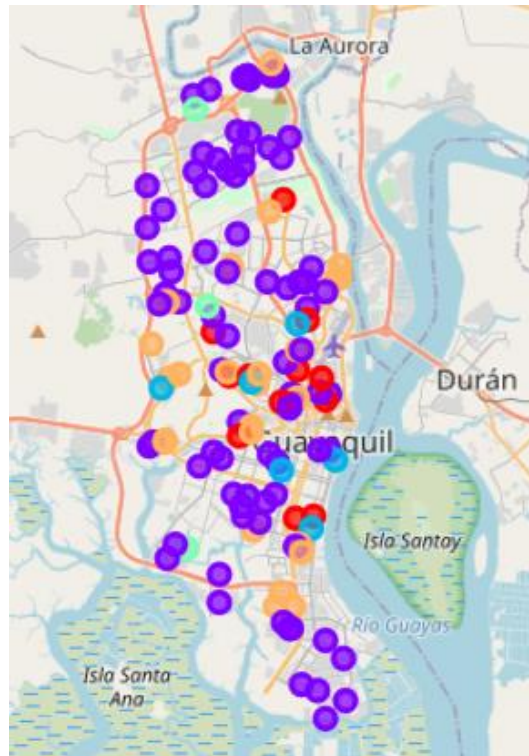
	Size	Price
count	129.000000	129.000000
mean	207.333333	109141.085271
std	163.793522	82188.045257
min	20.000000	10000.000000
25%	104.000000	50000.000000
50%	150.000000	83000.000000
75%	250.000000	160000.000000
max	1050.000000	450000.000000

The Kmeans clustering algorithm was applied, taking as a reference $k = 5$, selected by the elbow diagram and finally the model was adjusted.



Results Section

As a result of the segmentation analysis, 5 groups were obtained from the data generated by the olx web portal in Guayaquil Ecuador, as well as the proportion of nearby places of interest obtained by the Foursquare API:



First Group - Cluster

This group is defined by characteristics such as being close to parks, places of seafood, something very common in the gastronomy of Guayaquil, as well as many parks that can benefit the children of the clients of this company. In addition to being houses with a size above average and with a price that goes according to size.

Second Group - Cluster

This is the common group or that represents the most common characteristics of the houses for sale in Guayaquil, since they are prices and sizes that surround the average of these variables, in addition to not having a very defined pattern of places of interest around it.



Third Group - Cluster

This group has the particularity of having very high prices, with very high land, being close to many food places and pharmacies. These places usually occupy a higher status in the economic strata of Guayaquil.

Fourth Group - Cluster


They are houses with prices with very high sizes than usual, but with relatively low prices for their size, there is no definite trend of tourist places within this group, being affected for the most part by the size and price of the houses at the sale. Manually analyzing said group, it can be realized that they are places very far from the city center, that is, the suburbs of the city, here the factor not included of Crime Rate could be very significant in this analysis, given that the crime rate it can be very high, resulting in this group in a worse final option. But for people who like these types of places it could be a good option, with a low price for an above-average lot size.

Fifth Group - Cluster

This group has the characteristic of being surrounded by shopping malls, parks, many places to eat, prices average 180,000 and the size of the houses is around 200 square meters.

Discussion Section

As part of this initial problem, we set out to create a segmentation model that allows us to generate better performance when serving clients in a real estate company in Guayaquil Ecuador. Imagine that a new client comes to the company, and that only by filling out a form, whether online or physical, of the main characteristics that the client is looking for such as size, price, places of interest that he attends regularly, how many children he has, among others. things, said person in just a matter of seconds has as a result a segment of houses that he might like only on the basis of which segment or cluster is closest to the tastes / requirements of the person. For example, if a person would like to be near a shopping center and has the economic possibilities of spending a little above the average, then he may like segment 5 of this analysis a lot, and then make the respective tour of the houses to the sale with an agent in the following days, but the time saved and therefore the cost would also go down.



It must be recognized that for this analysis to be applicable, you must have a data source that can be consulted in real time, in addition to being very reliable, for example, records were taken from the olx web portal, but as said At the beginning of the project, this data has certain limitations, which can be biased data by the person who is selling the house, such as the price or the size of the house.

Conclusion Section

Some conclusions can be drawn about this project:

- The price and size of the house for sale influenced the segmentation model much more than the places obtained by Foursquare, since the geography of Guayaquil itself has places such as cafes, restaurants, etc. widely distributed. Except for parks and shopping centers that influence certain segmentation groups. Knowing the economic stratum of the people and some few requirements, a segment of houses could be filtered in real time.
- The model could be improved with new variables such as crime rate, number of apartments, age of the house, among others. In turn, the use of Foursquare could be redefined to only limit it to shopping malls and parks, since this does affect the segmentation model.
- This model should be constantly updated in real time and with reliable information, measured by the corresponding person, so that it can finally be useful for this real estate company in Guayaquil, Ecuador.