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

Jimmy Joel Landín Casal

Applied Data Science Capstone Project

Introduction

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.

Data

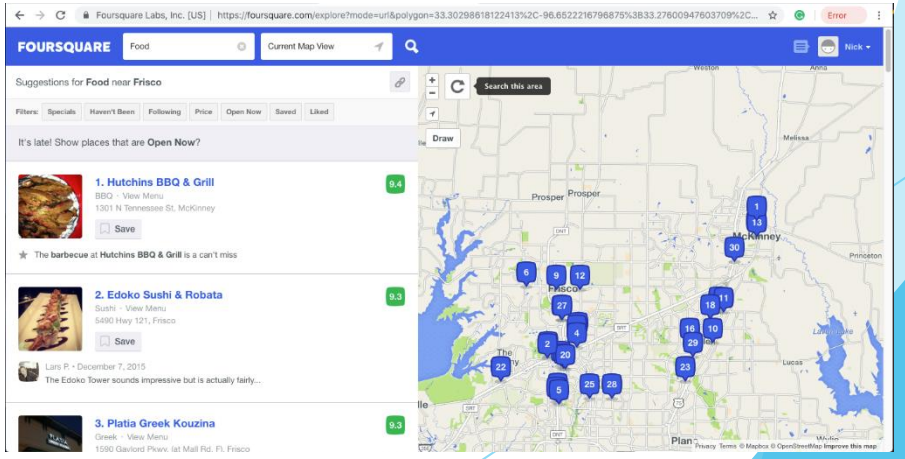
Olx info - Houses in Sale - January 2021

- **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.

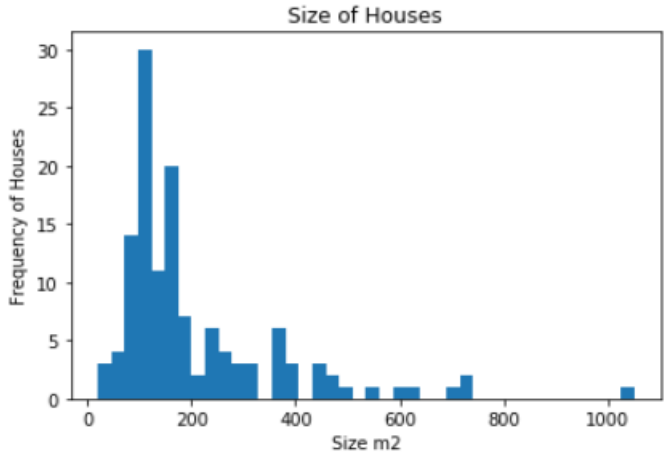
	Code	Neighborhood	Latitude	Longitude	Size	Price
0	AC247	Acacias	-2.2417	-79.9008	309.0	110000.0
1	AC246	Acacias	-2.2467	-79.8966	308.0	145000.0
2	AC129	Acacias	-2.2465	-79.9026	194.0	185000.0
3	AR308	Acuarelas del Rio	-2.1299	-79.8809	170.0	165000.0
4	AR380	Acuarelas del Rio	-2.1351	-79.8825	230.0	176000.0

Venues - Foursquare API

FOURSQUARE

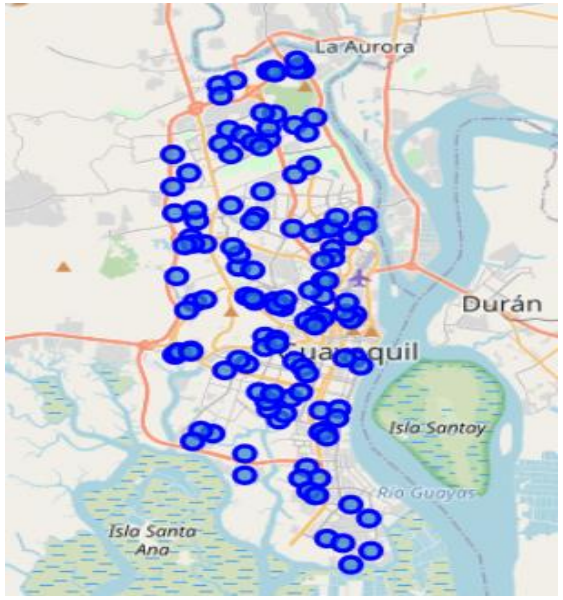


Methodology



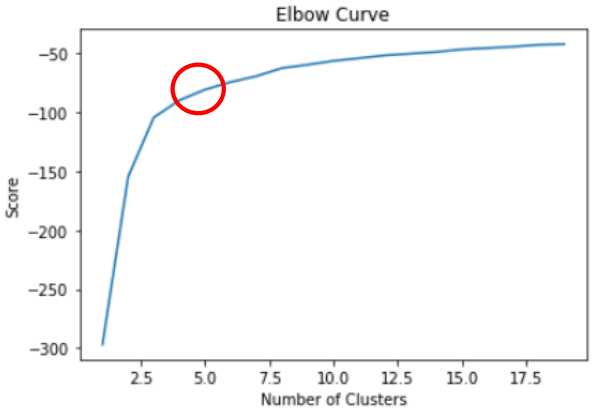
	Size	Price
count	129.000000	129.000000
mean	207.333333	109141.085271
std	163.793522	82188.045257
min	20.000000	10000.000000
25%	104.000000	50000.000000
50%	150.000000	83000.000000
75%	250.000000	160000.000000
max	1050.000000	450000.000000

Methodology



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.

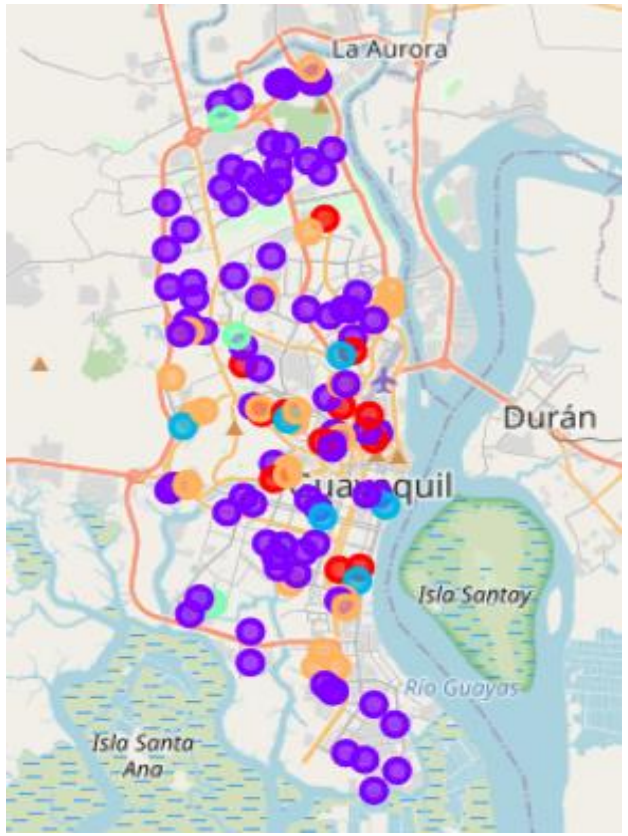
As a later step, the places near these houses and a radius of 500 meters were added through the Foursquare API, generating more enriched information.



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

Results

Five Groups - Segmentation



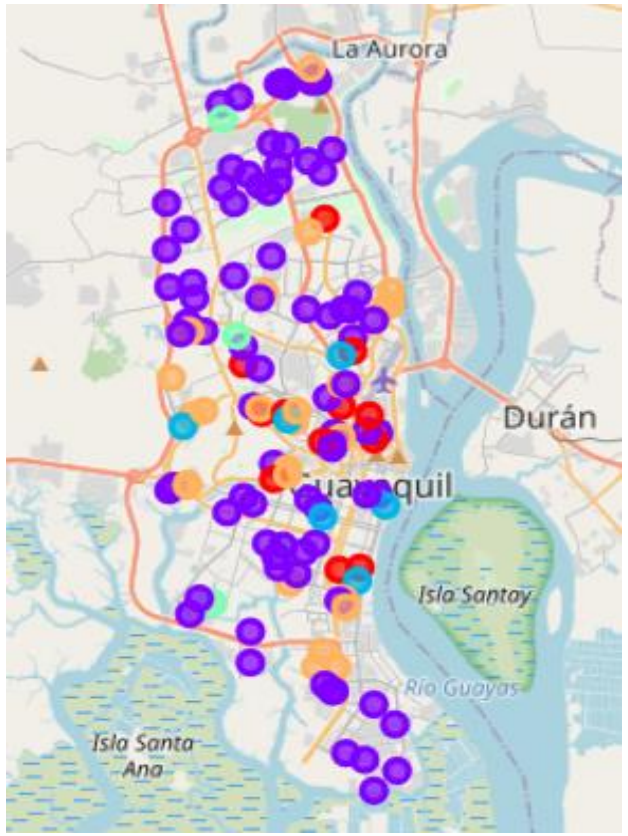
First Group - Cluster
This group is defined by characteristics such as being close to parks, places of seafood. In addition to being houses with a size above average and with a price that goes according to size.

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.

Results

Five Groups - Segmentation



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.

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

- 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.
- 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.

Conclussion

- 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.