STUDY OF CADASTRAL PRICES IN COLOMBIA

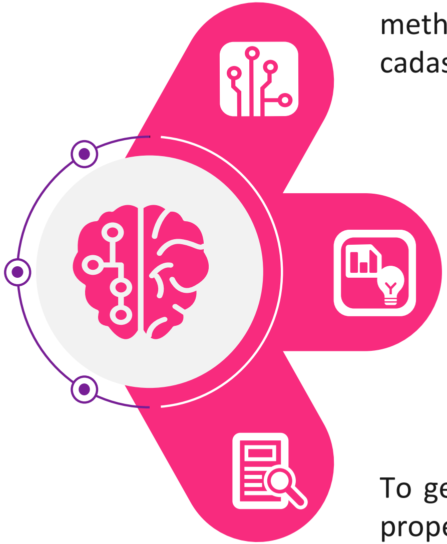
Team 3

Business context and objectives

The manual calculation of cadastral prices in Colombia are expensive and takes time so it is necessary to automate the process.

The Instituto Geografico Agustin Codazzi (IGAC) is a public establishment that produces the official map, the basic cartography of Colombia and national cadastre .

Currently, they are no cadasters in 28% of the Colombian territory, and 65,99% of the cadastral data in rural areas is not updated.



To automate with data science methodologies the calculation of cadastral prices in Colombia.

To get insights from current properties data from municipalities: Chiquinquirá, Ricaurte, Tenjo, Apía, Balboa, Belén de umbría, Guatica, La Celia, Marsella, Nistrató, Pueblo Rico, Quinchía, Santa Rosa de Cabal, Santuario, Cumaribo

To get more information about the properties from external sources like web pages.

Data Exploration

So far we have 3 datatables with data from our interested municipalities that have information about properties. The dependent variable is AVALUO_MILLONES

Avaluos

Contains:

- Departamento: Department code
- Municipio: Municipality code
- Zona: Zone. It could be Rural or Urban
- Número predial: Property number
- Número predial nacional: National property number
- Avaluo: Estimate of the value of each property in COP

General_asp

Contains:

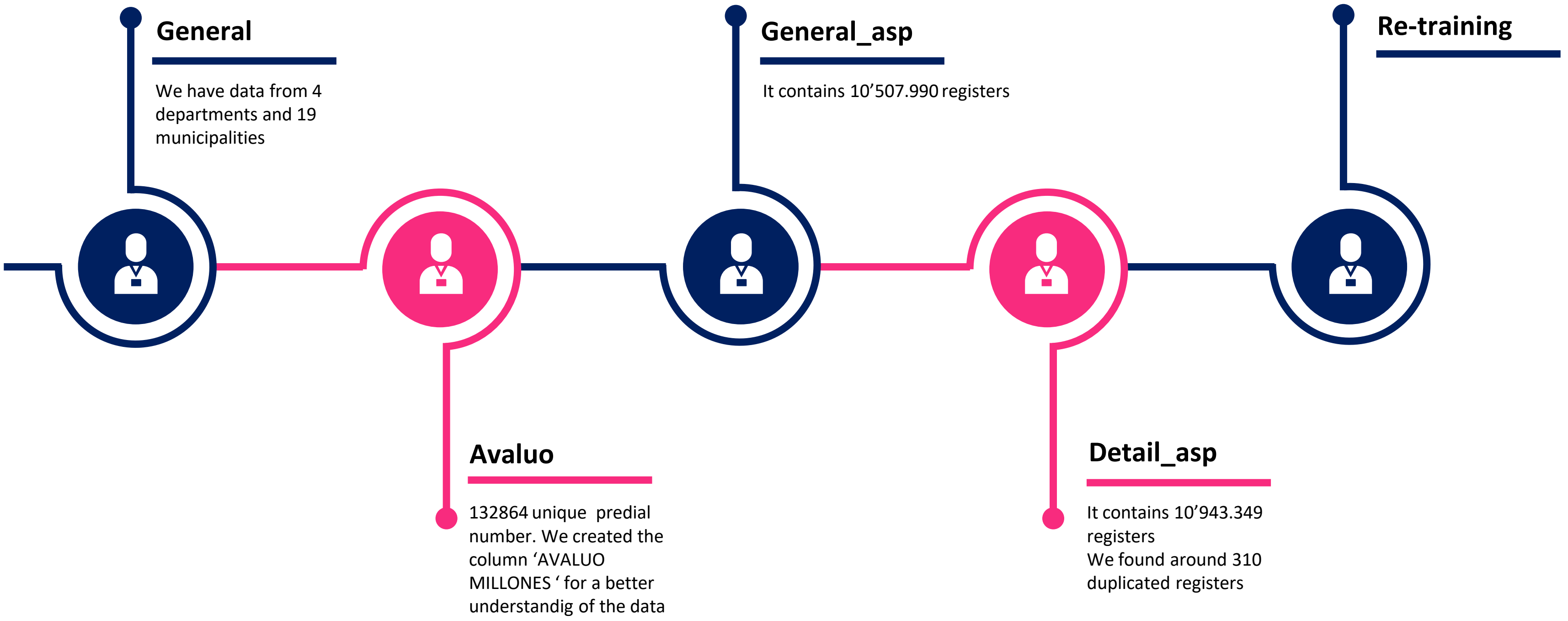
- Departamento: Department code
- Municipio: Municipality code
- Numero del predio: Property number
- Direccion: Address
- Destino economico: Economic destination
- Area terreno: property area (mt2)
- Area construida Built area (mt2)
- Numero predial nacional" property national number

Detail_asp

Contains:

- Departamento: Department code
- Municipio: Municipality code
- Numero del predio: Property number
- Zona Fisica 1: Code of the physical zone
- Zona Economica 1: Economic zone code
- Area terreno 1: property area (mt2)
- Habitaciones 1: Built area (mt2)
- Banos 1: number of bathrooms
- Locales 1: number of business premisses
- Estrato 1: Typification code assigned to physical zone
- Uso 1: Code that identifies the construction area
- Puntaje 1: Construction score

Data Insights



General

We have data from 4 departments and 19 municipalities

Avaluo

132864 unique predial number. We created the column ‘AVALUO MILLONES ‘ for a better understandig of the data

General_asp

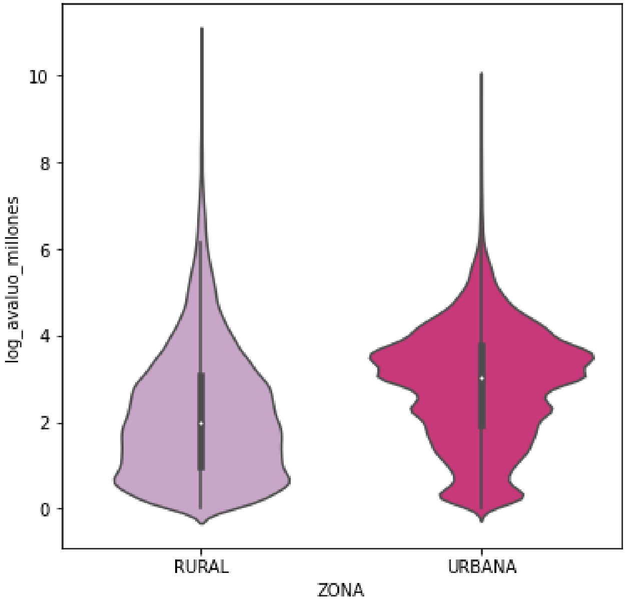
It contains 10’507.990 registers

Detail_asp

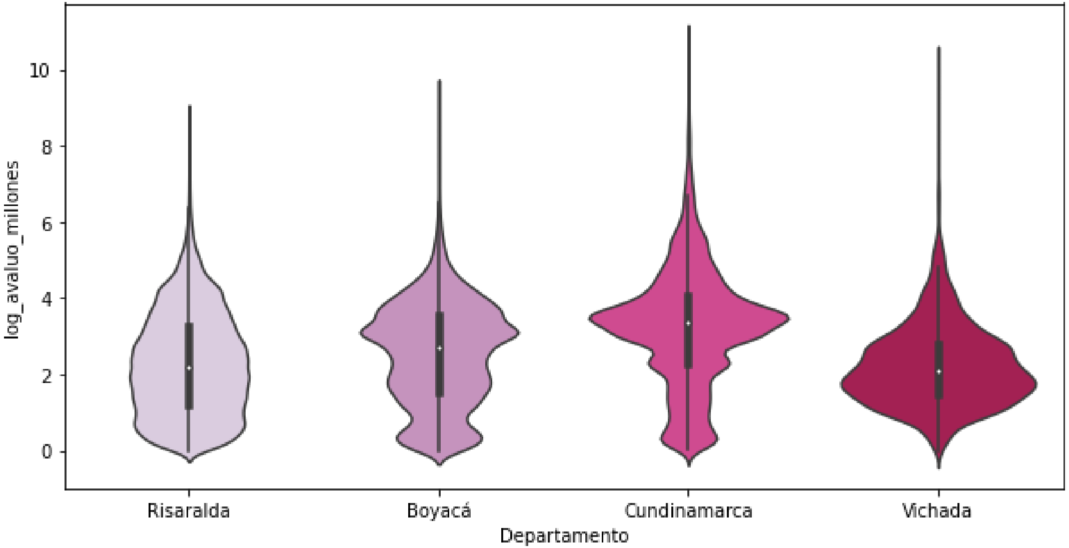
It contains 10’943.349 registers
We found around 310 duplicated registers

Re-training

Appraisal by zone and departments



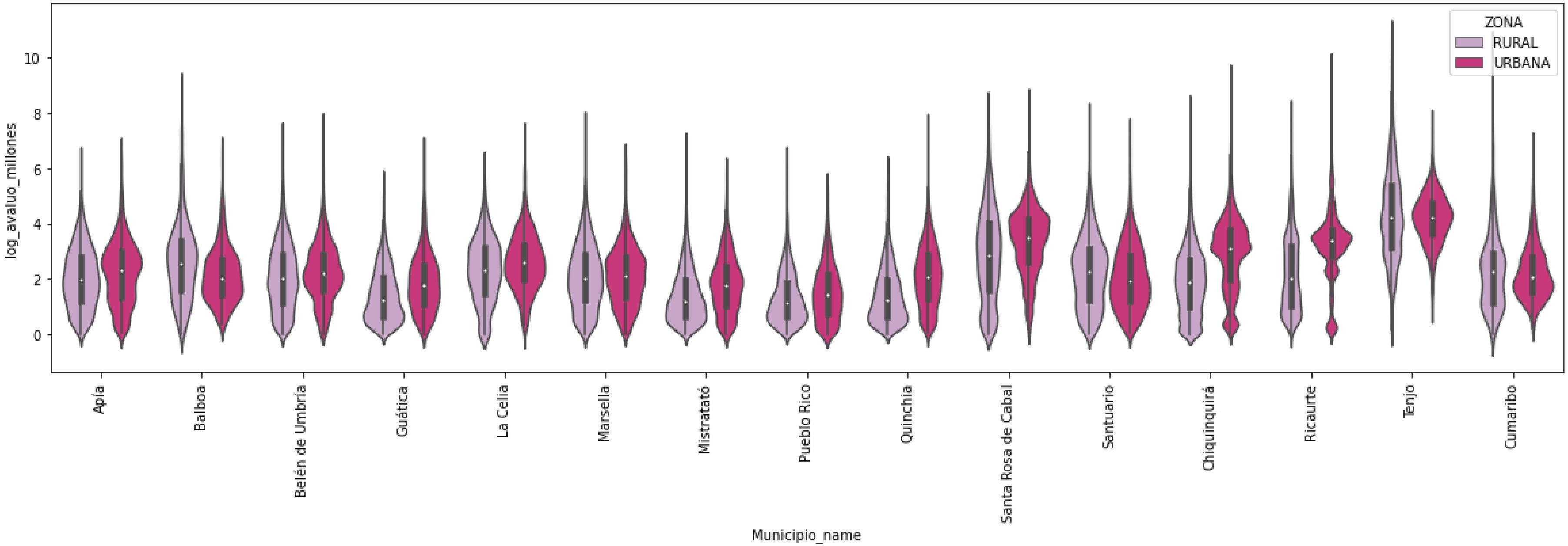
In urban areas the average of the cadastral price is higher than in rural areas. However we noticed there are many outliers in both of them, but larger in rural zones.



The average price of the appraisal is higher in Cundinamarca, followed by Boyacá, Risaralda and finally Vichada. Furthermore, is important to note the that all the distributions have different distribution.

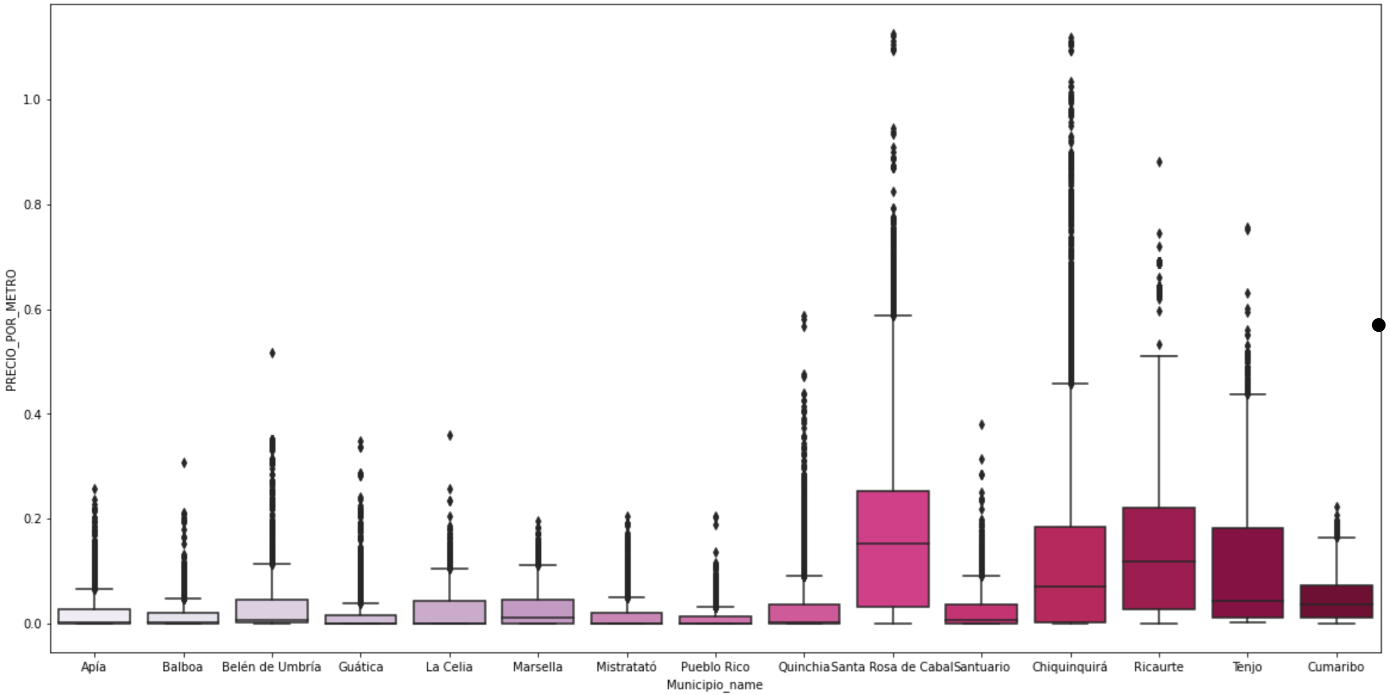
NOTE: Due to the large range between the mean and the maximum value of the appraisal, the dependent variable was transformed with the log value.

Appraisal by municipalities and zones



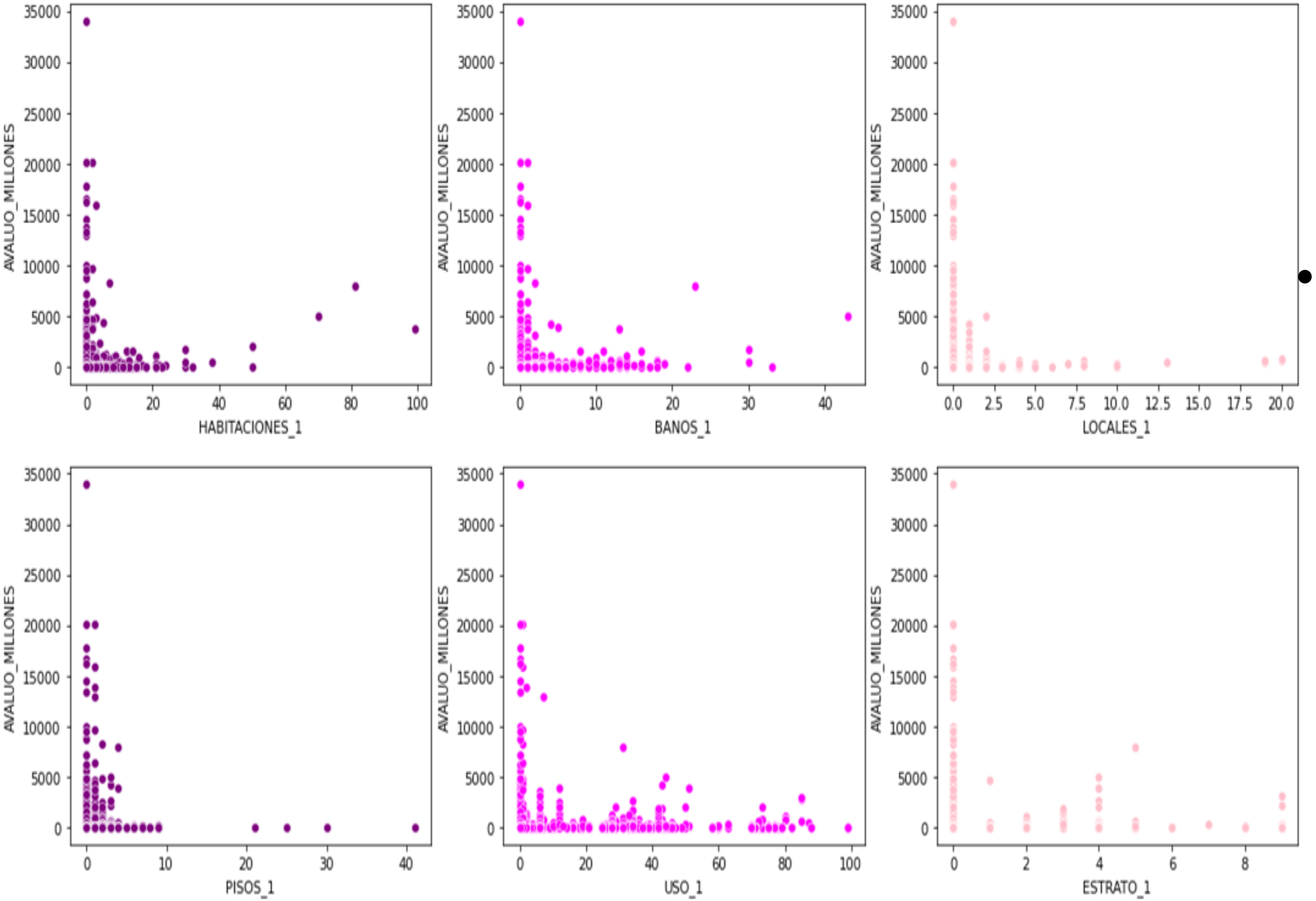
In every municipality the average cadastral price in Urban zones is higher than in rural zones. This can occur due to the development and growth of cities, employment and business opportunities, which generate a boom in construction and infrastructure.

Appraisal by m2



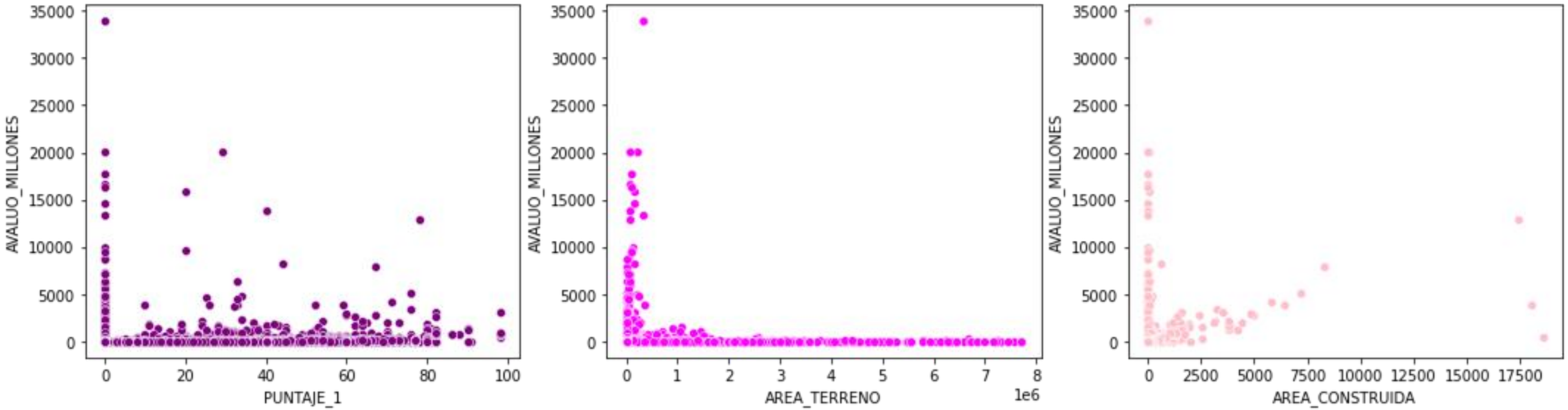
- Municipalities closest to the capitals of the department or main cities present higher valuations as observed in Tenjo and Santa Rosa de Cabal, and a surprising Ricaurte.
- The municipalities with the lowest appraisals are Pueblo Rico and Mistrató located to the north-west of Risaralda, the closest city is Pereira (approximately 90km) and Guaticá located 93 km from Pereira.

Univariate scatter plots with outliers



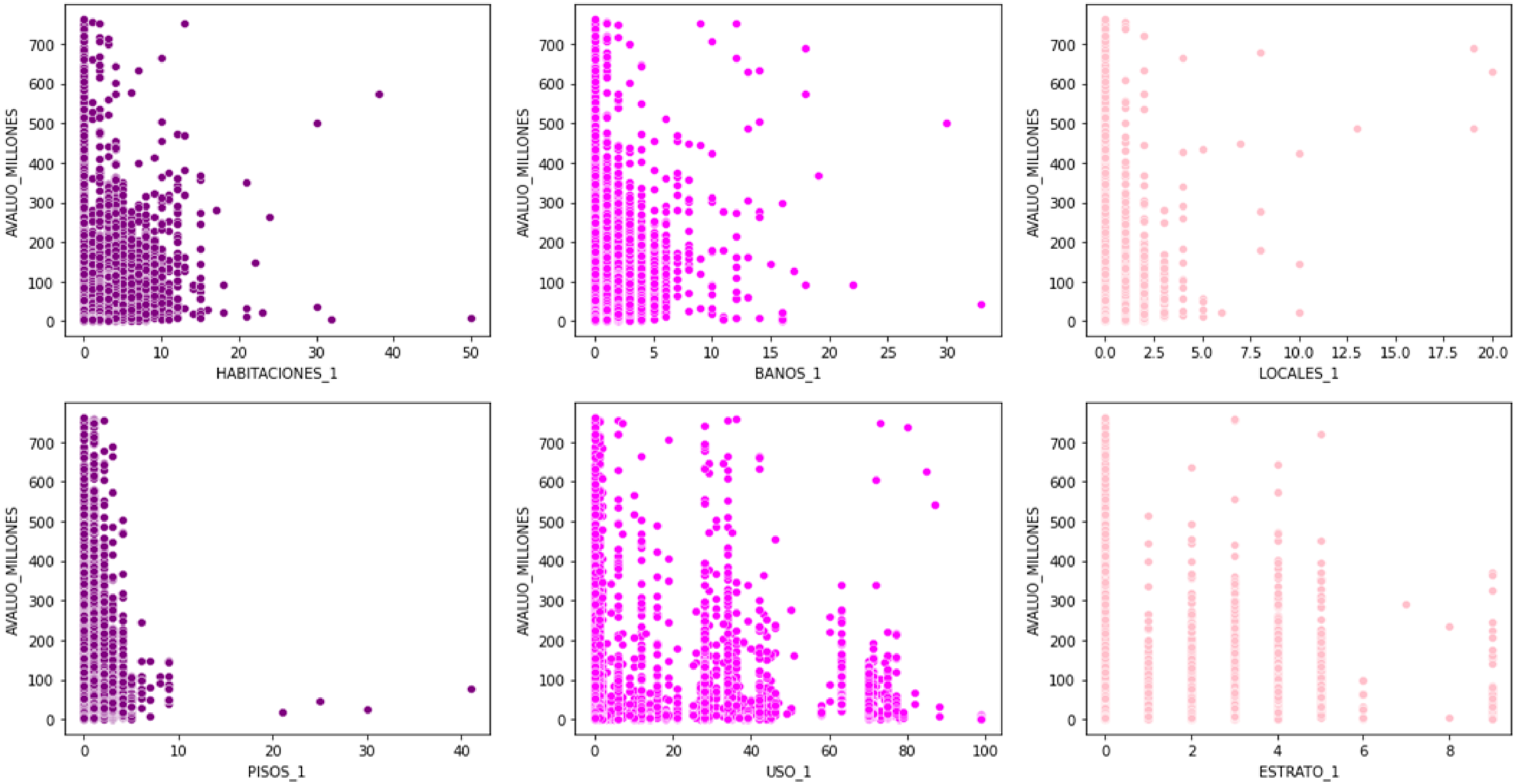
- No significant correlation is observed between the variables shown here (HABITACIONES, BAÑOS, LOCALES, PISOS, USO and ESTRATO) and the value of the appraisal.

Univariate scatter plots with outliers

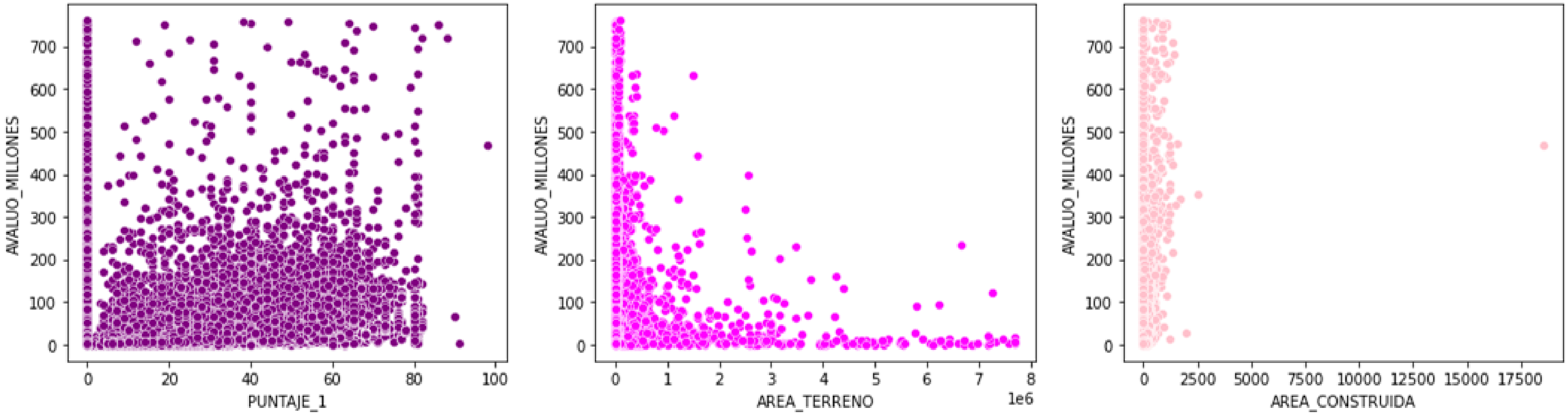


Here we can see that the price is not growing when we have more rooms or more area in the properties, it is necessary to perform a multivariate analysis.

Univariate scatter plots without outliers



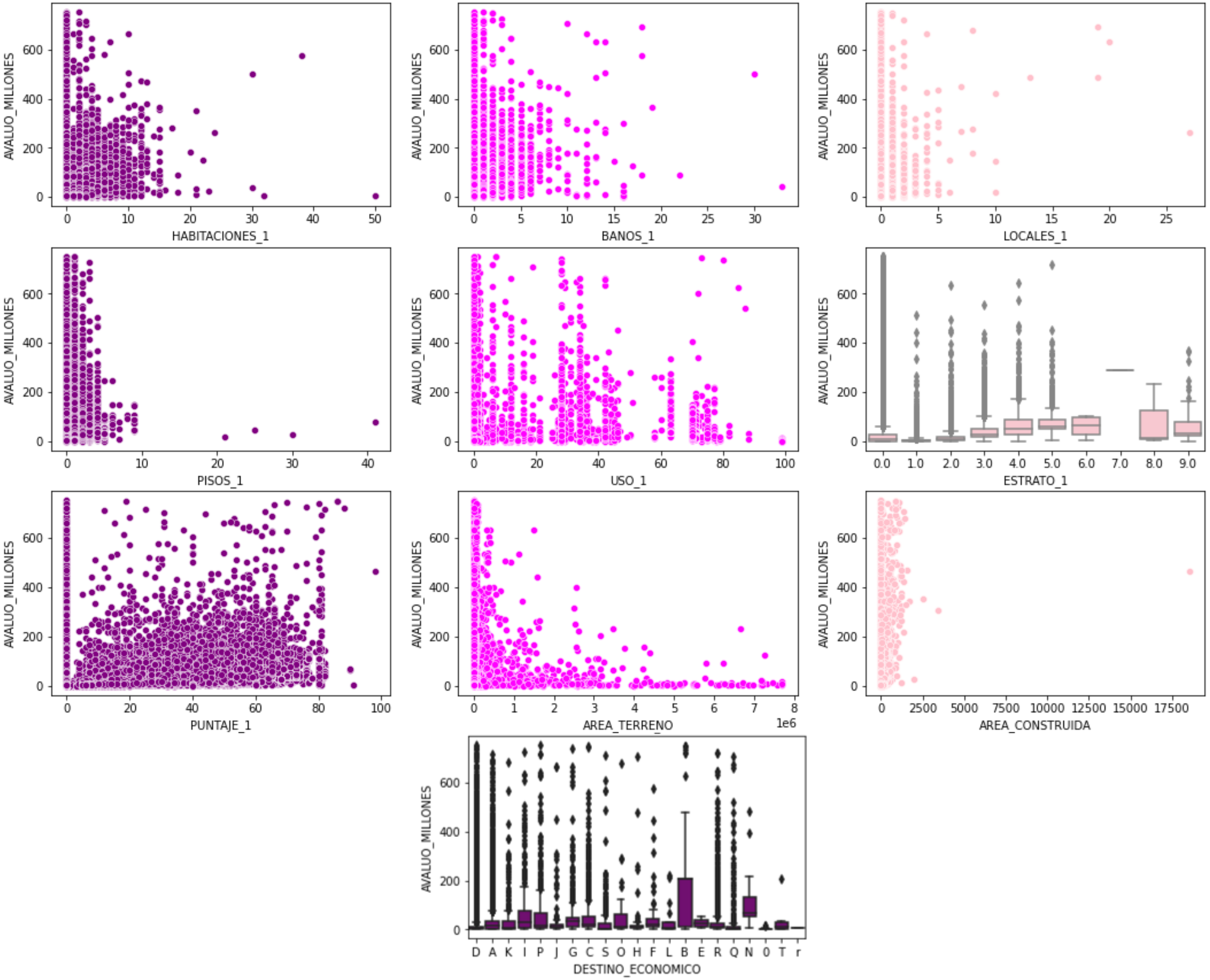
Univariate scatter plots without outliers

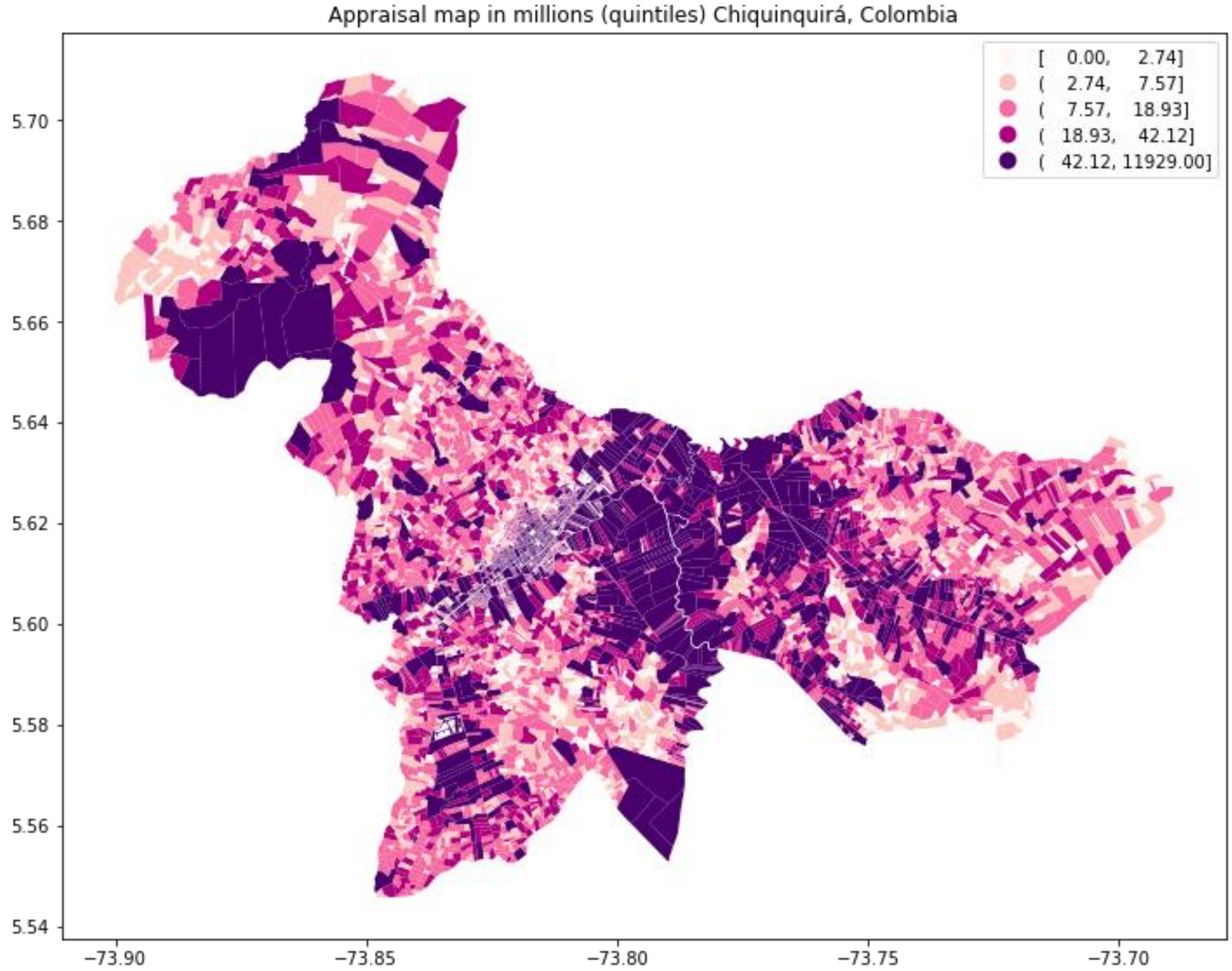


- In further analysis, we will compare the correlations depending if the property has a construction or not. Because it is notable for the concentration of points in all the 0 values.

Univariate scatter plots without outliers

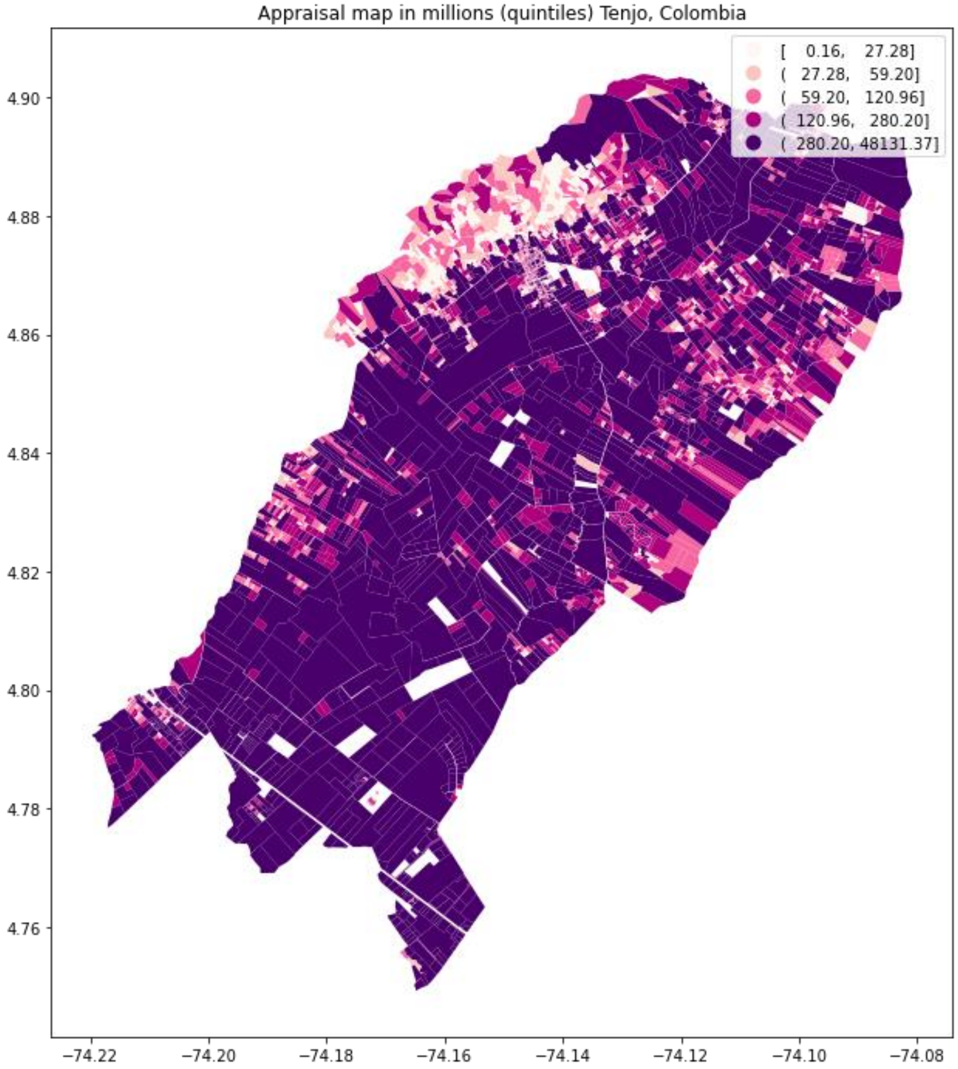
It’s hard to find a clear tendency in the different variables in the study. A detailed review of the remaining outliers is necessary to understand better these cases and probably start doing an independent analysis for the land value and constructed area value.





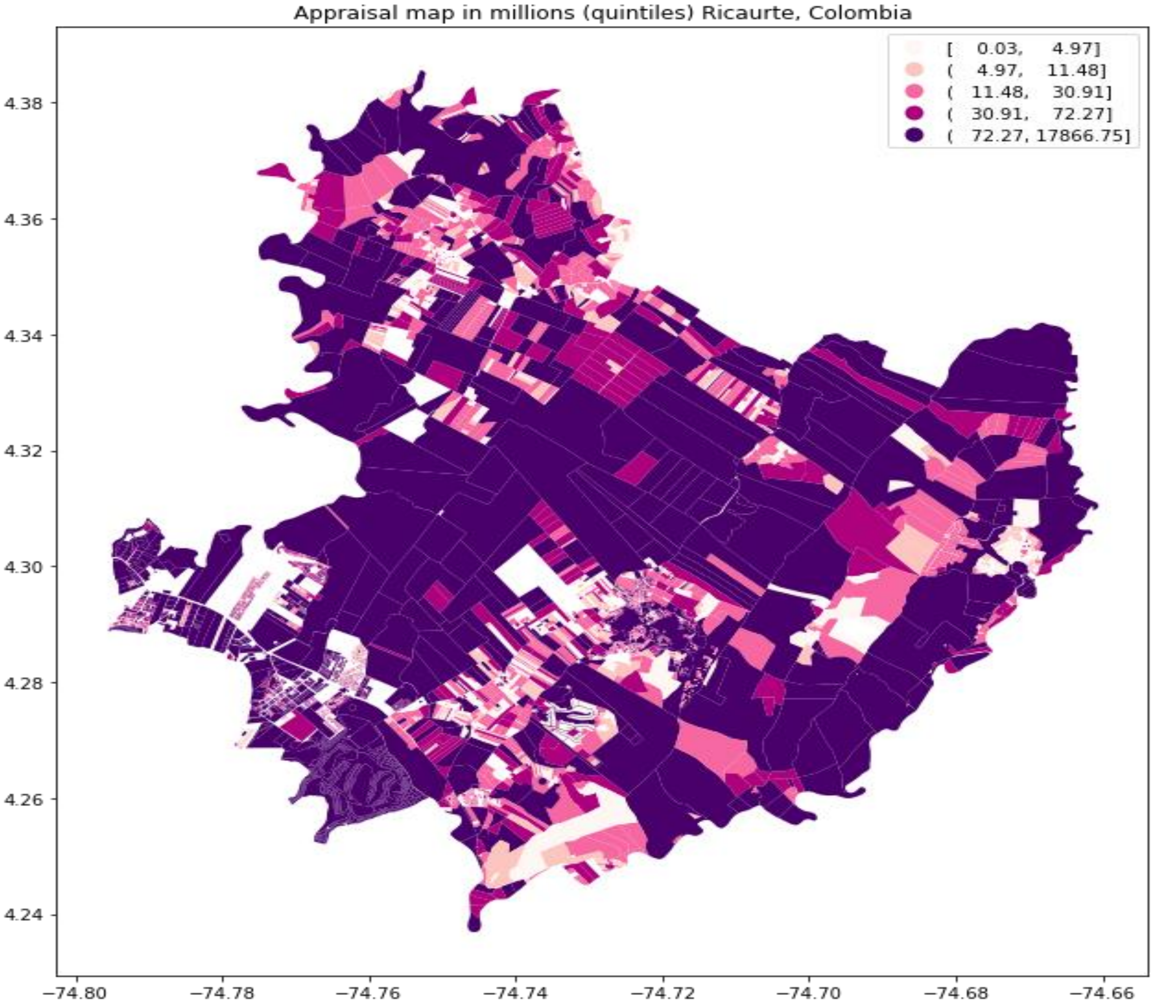
Appraisal map in millions (quintiles) Chiquinquirá.

For chiquinquirá, there is a spatial diversity in the valuation of millions of properties, although a spatial agglomeration in the western zone stands out due to the presence of miners.



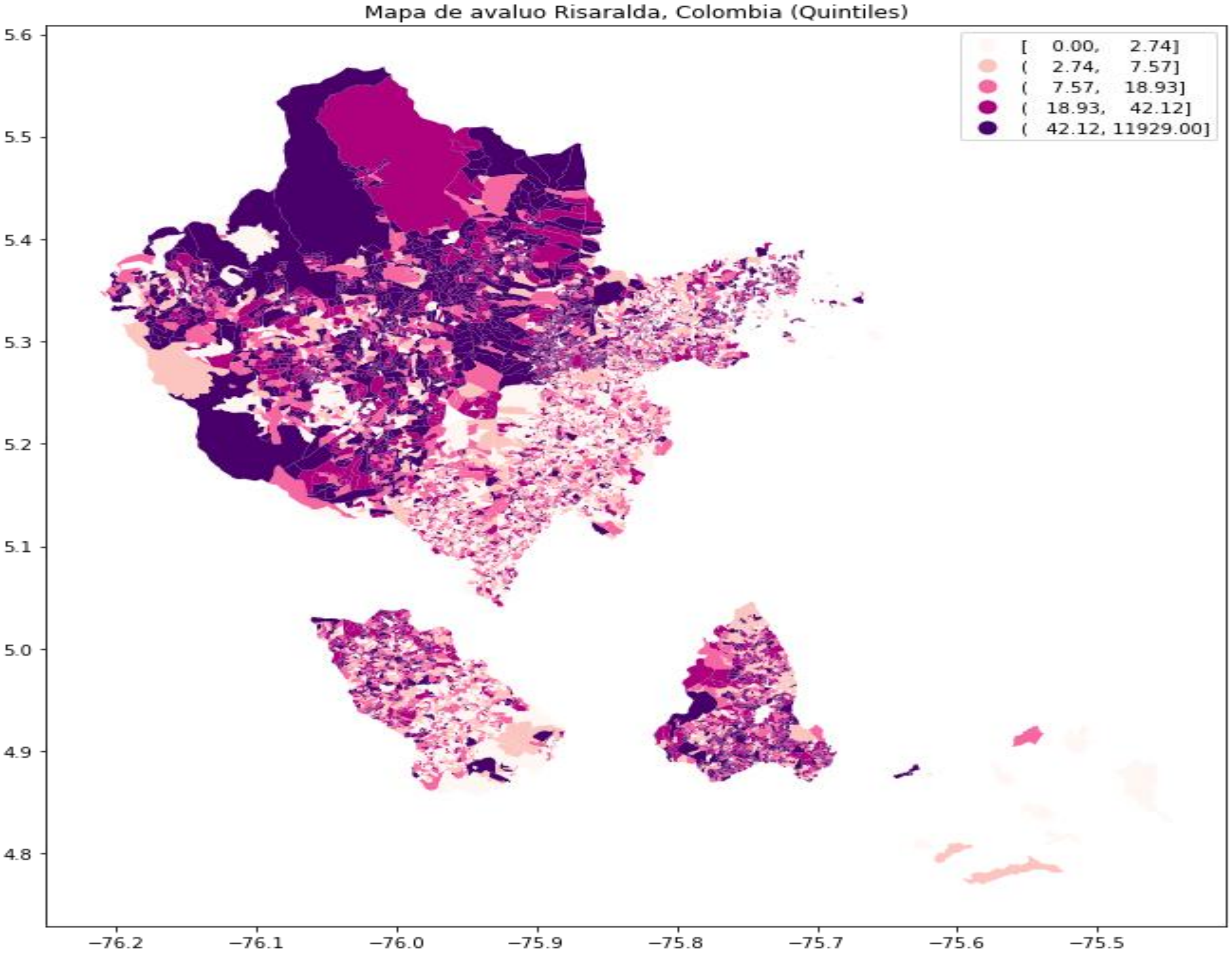
Appraisal map in millions (quintiles) Tenjo.

For Tenjo, it is evident that Chiquinquirá has a greater intensity in terms of millions of pesos of high amounts, due to the fact that it is a place of rest and vacation for the capital of Colombia.



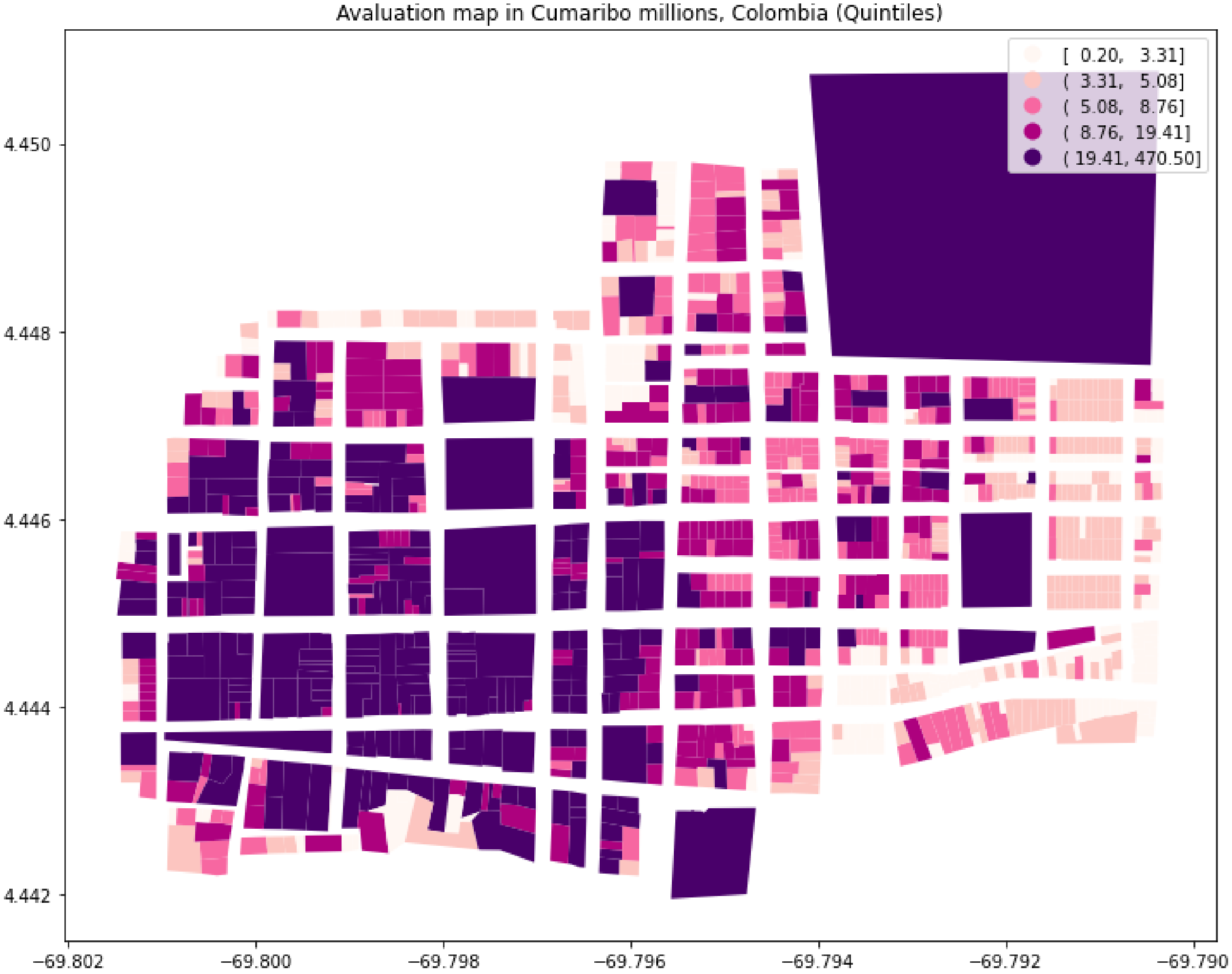
Appraisal map in millions (quintiles) Ricaurte.

Ricaurte is a municipality located in a warm climate and close to Girardot and Melgar which are tourist sites par excellence for the residents of the city of Bogotá, therefore, the high values in millions of pesos.



Appraisal map in millions (quintiles) Risaralda.

In this case Risaralda is a department, NOT a municipality. In this sense, the information provided does not contain all the municipalities that make it up, so it is presented as follows. As it is the agglomeration of the highest natural reserves, it is necessary to explore the data further.



Appraisal map in millions (quintiles) Cumaribo.

Without a doubt, Cumaribo is one of the municipalities with the lowest level of detail in information. In this case, we only have spatial data at an urban level but not at a rural level which is evident on the map. It is the challenge of the exercise that we will focus on filling these gaps for the result to be obtained.

Public Data

There is information on different websites that could be of great help in the analysis of the properties' appraisals, such as real estate websites, APIs, and government data. This information can be obtained via webscrapping, making multiples requests, or direct download.

Real Estate websites could provide more sell prices and values of the properties, as well as their characteristics. Some examples are www.fincaraiz.com.co, www.metrocuadrado.com.co, www.properati.co.

Location APIs could provide information of near places of interest, as well as a help in geolocalizing addresses. Examples: Google APIs, HERE.

Government and municipalities public data. Examples: Planes de Ordenamiento Terrioríal. (POT)

Preparing and exploring Public Data

Using different methods and webscrappings, it was possible to gather information of the Properati and Fincaraiz websites. Initially, the datasets contains over 1 million properties altogether with information such as price, area, number of bathrooms, number of bedrooms, number of garages, location (WGS84 longitude and latitude), a detailed description for each property, and many more.

It is necessary to unify variables in order to create an unique dataset. These common variables are:

Price or value

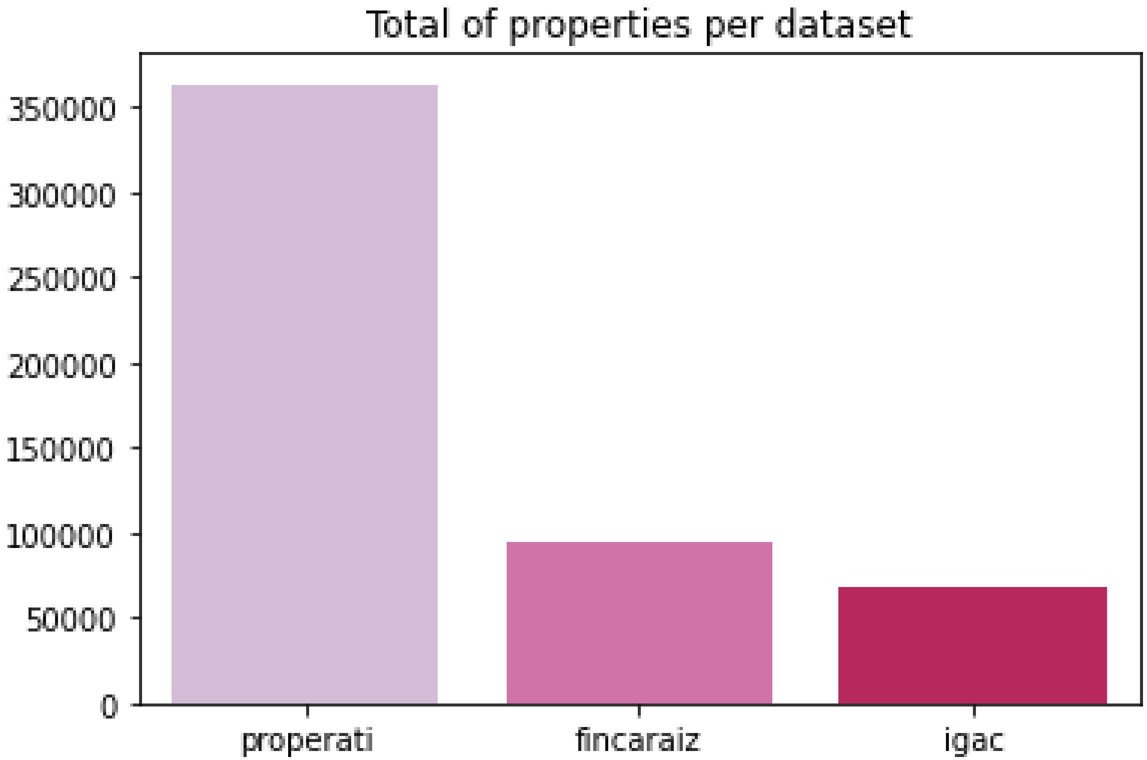
Number of bathrooms

Area (square meters)

Number of bedrooms

Location (lon, lat, city)

Preparing and exploring Public Data

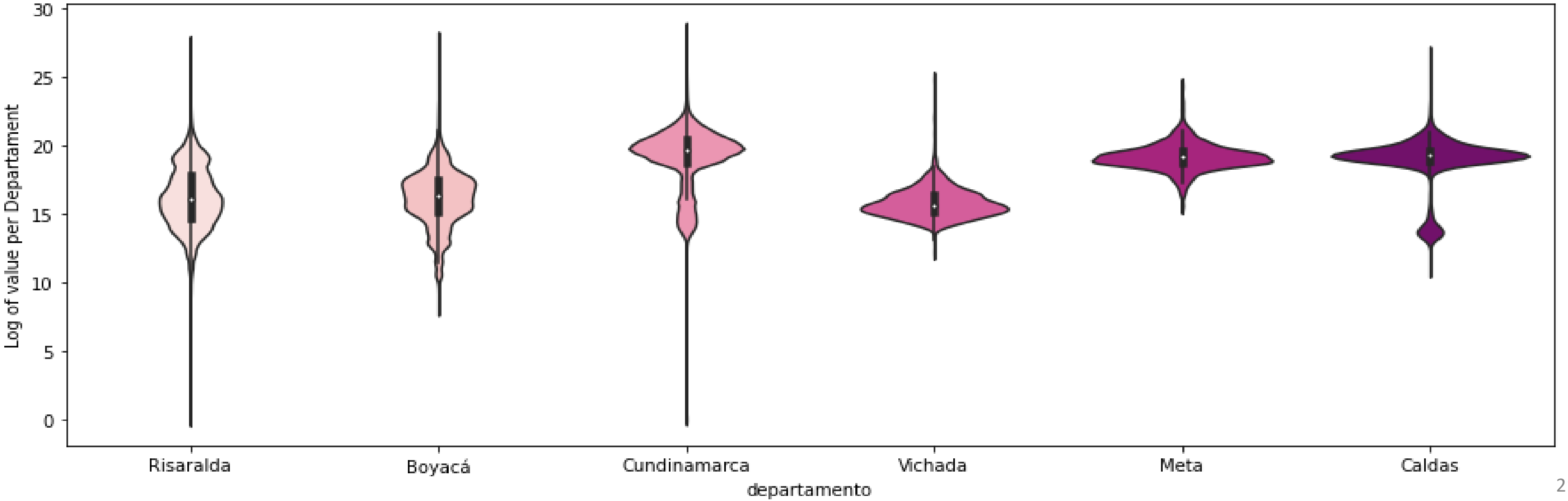


After filtering the departaments of interest, selecting common variables, and concatenating the multiple sources, the final dataset contains over 500 thousand properties, with properati being almost 70%.

Disclaimer: a concatenate was preffered over the inner join due to the final dataset being a lot bigger (400 properties vs 500 thousand).

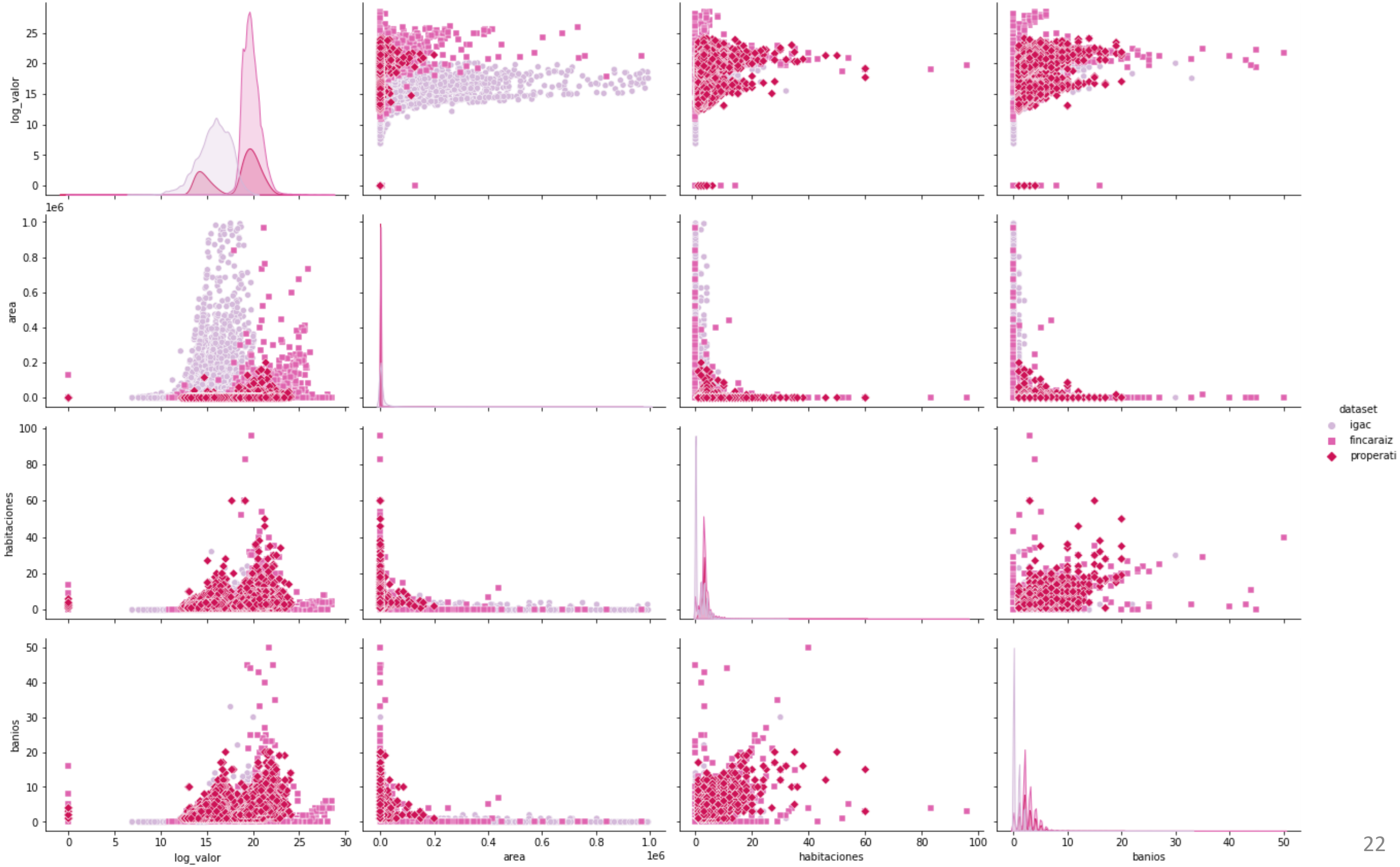
Preparing and exploring Public Data (value distribution)

It can be seen that the distribution of values per departament is mostly homogeneous, excepting Caldas with a small cluster around lower values. Cundinamarca has higher values, almost concentrated on the even higher ones, meanwhile Vichada and Boyacá have lower values.

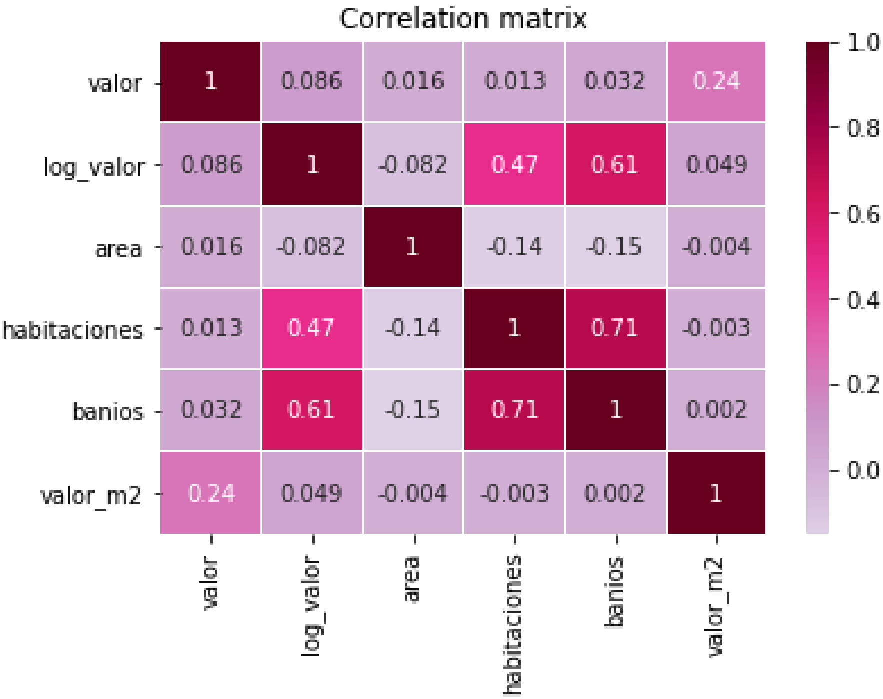


Preparing and exploring Public Data (pairplots)

Here it is shown the plots of the different quantitative values of the dataset against each other (value, area, bedrooms, and bathrooms).



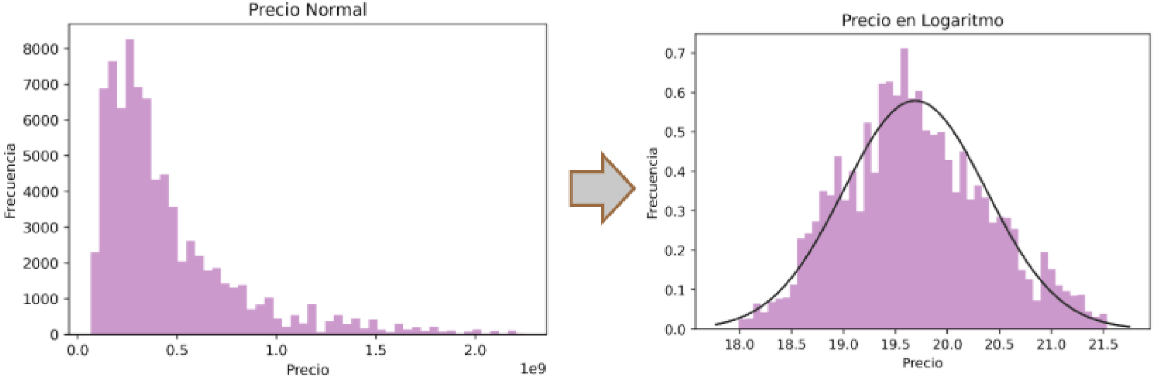
Preparing and exploring Public Data (correlation)



	valor	log_valor	area	habitaciones	banios	valor_m2
valor	1	0.086	0.016	0.013	0.032	0.24
log_valor	0.086	1	-0.082	0.47	0.61	0.049
area	0.016	-0.082	1	-0.14	-0.15	-0.004
habitaciones	0.013	0.47	-0.14	1	0.71	-0.003
banios	0.032	0.61	-0.15	0.71	1	0.002
valor_m2	0.24	0.049	-0.004	-0.003	0.002	1

The correlation matrix can help in the study to consider variables, as it can be seen that number of bathrooms and bedrooms are highly correlated, and that it exists a moderate correlation between these features and the logarithm of the value. Some correlations must not be taken into account like value per square meter and value, as one former is calculated with the latter.

Cleaning



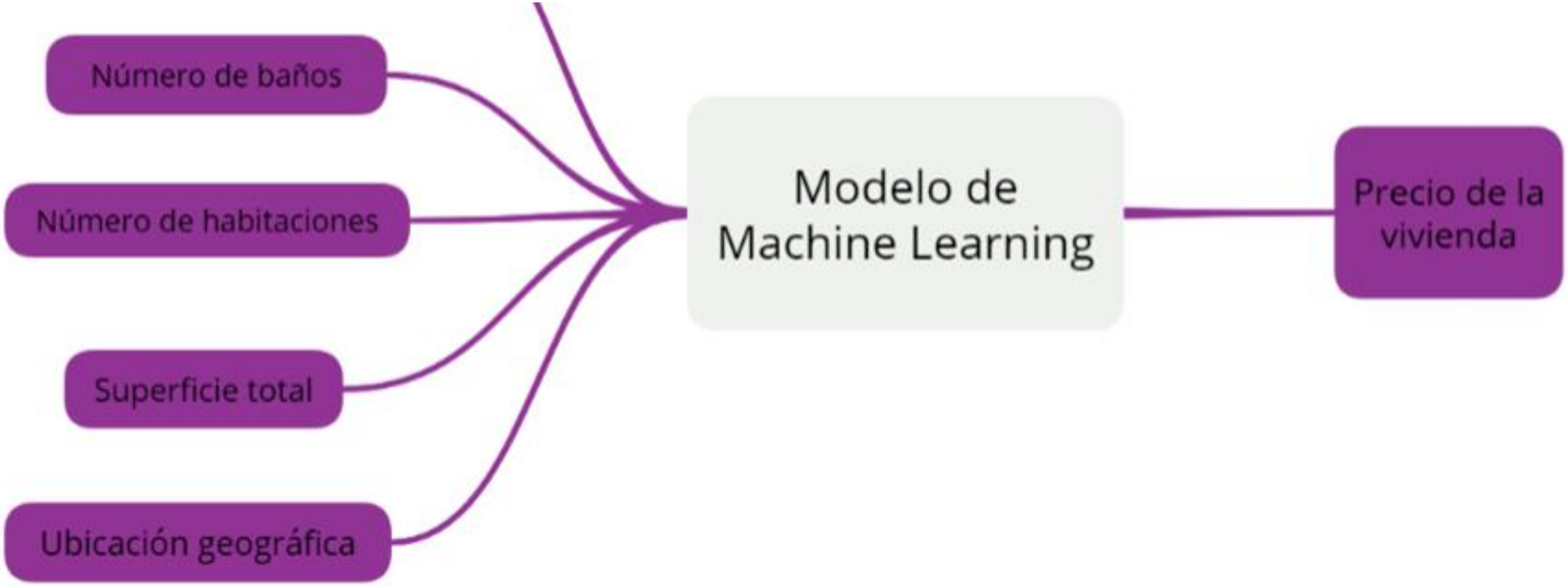
Normalization and missing value imputation

Yeo-Johnson Transformation

Simple Imputer

Marginal Elimination

Model



Número de baños

Número de habitaciones

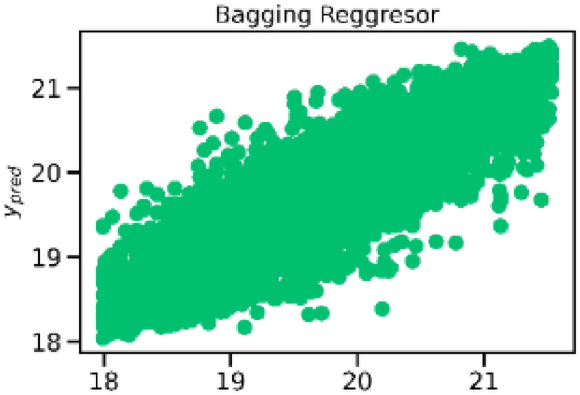
Superficie total

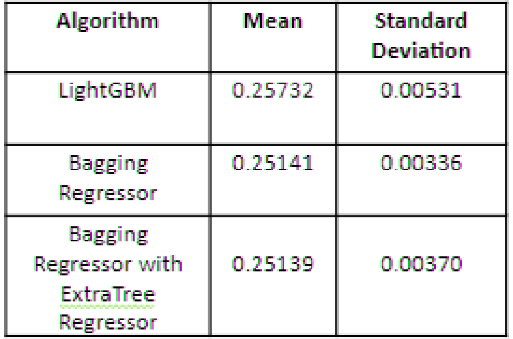
Ubicación geográfica

Modelo de Machine Learning

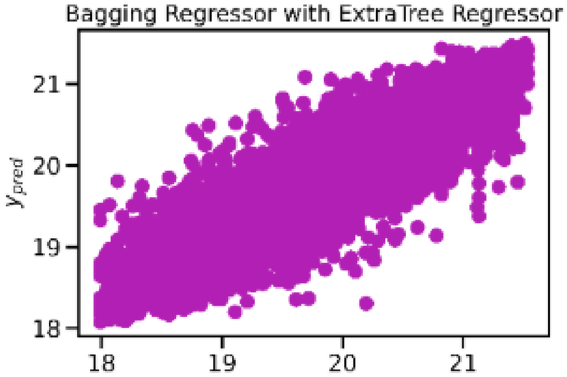
Precio de la vivienda

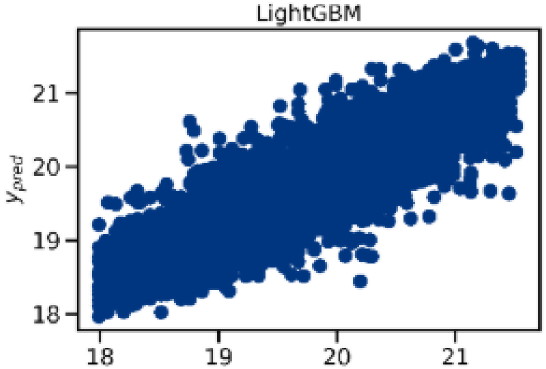
Evaluation Metrics

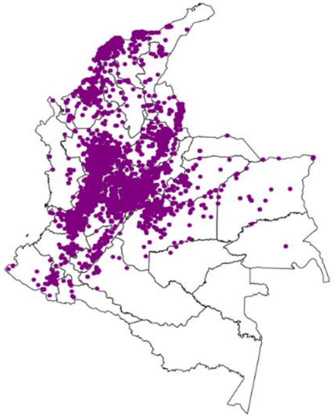

Bagging Reggresor



Algorithm	Mean	Standard Deviation
LightGBM	0.25732	0.00531
Bagging Regressor	0.25141	0.00336
Bagging Regressor with ExtraTree Regressor	0.25139	0.00370


Bagging Regressor with ExtraTree Regressor


LightGBM



Stratification

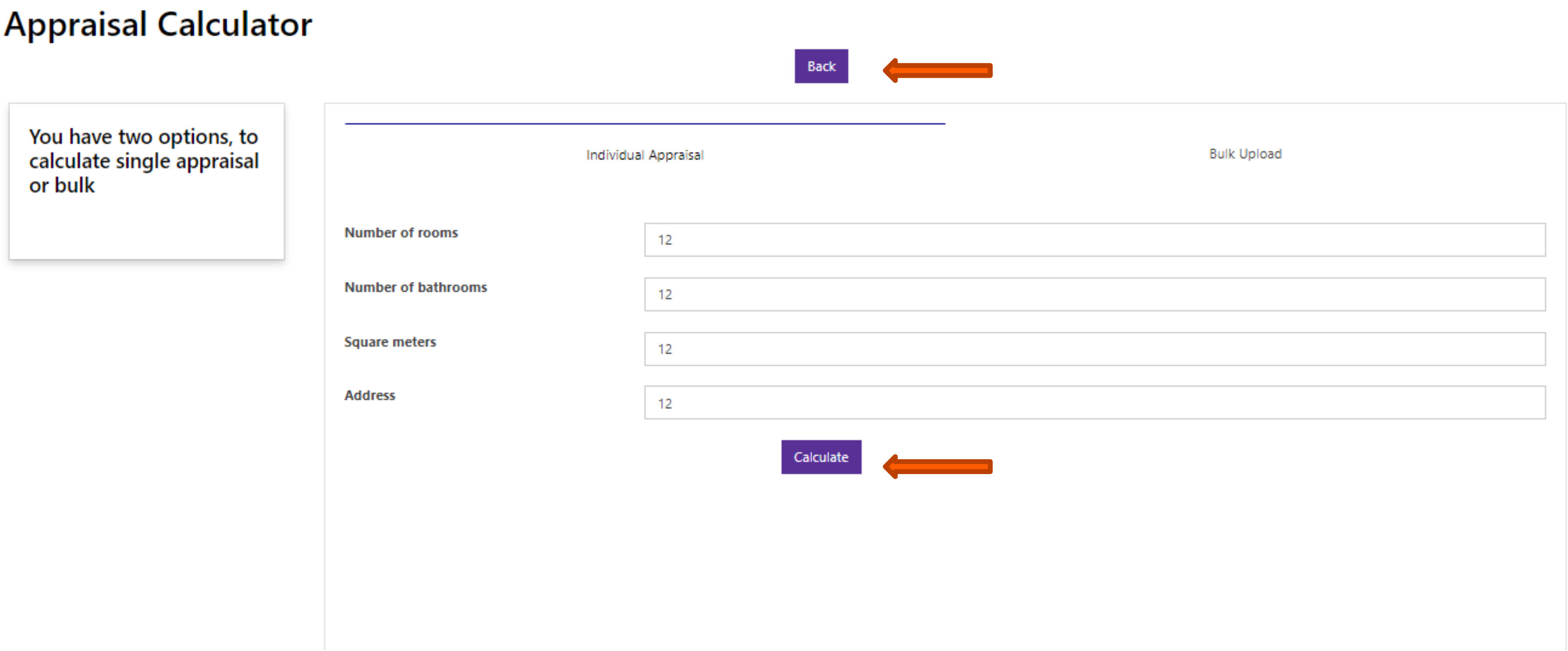
Undersampling

Tunning

Cross Validation

You can acces the app here: https://avalpredict.vatiolibre.com/

Some updates on the app (back and calculate on the appraisal calculator)



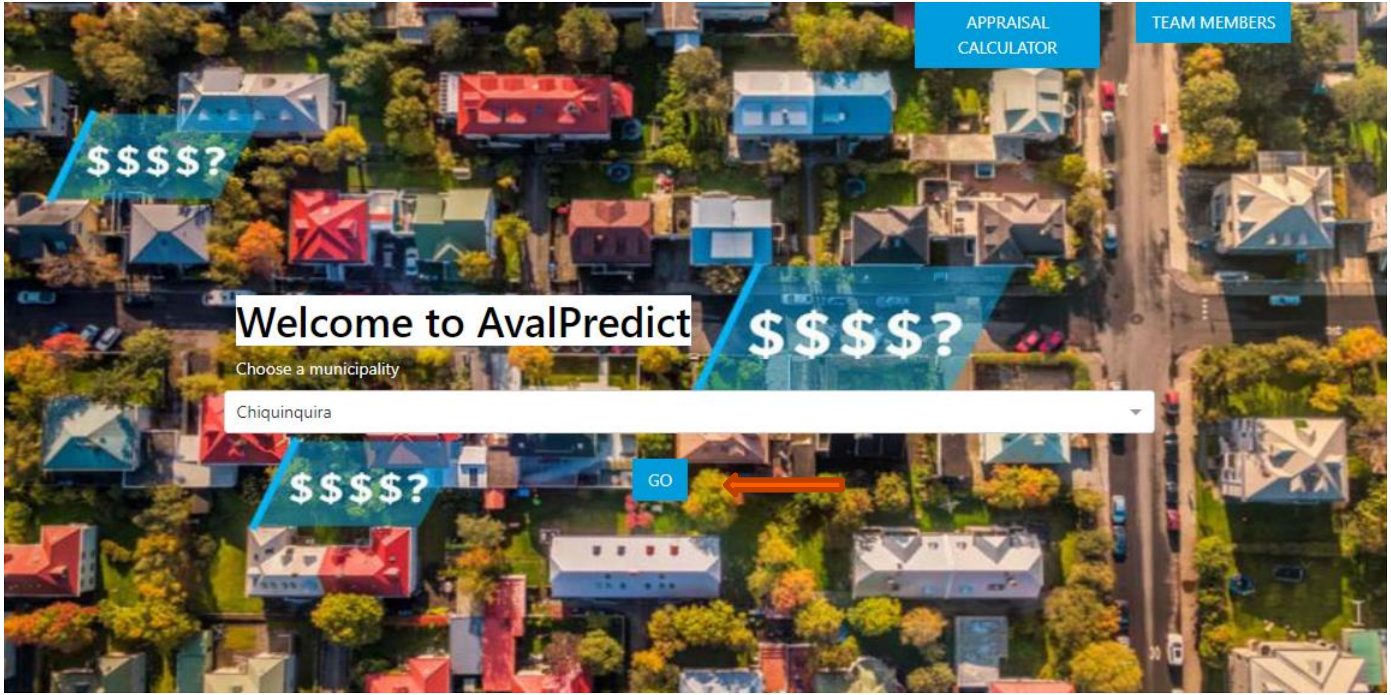
Appraisal Calculator

You have two options, to calculate single appraisal or bulk

Number of rooms	12
Number of bathrooms	12
Square meters	12
Address	12

You can acces the app here: https://avalpredict.vatiolibre.com/

Some updates on the app (go buttom)



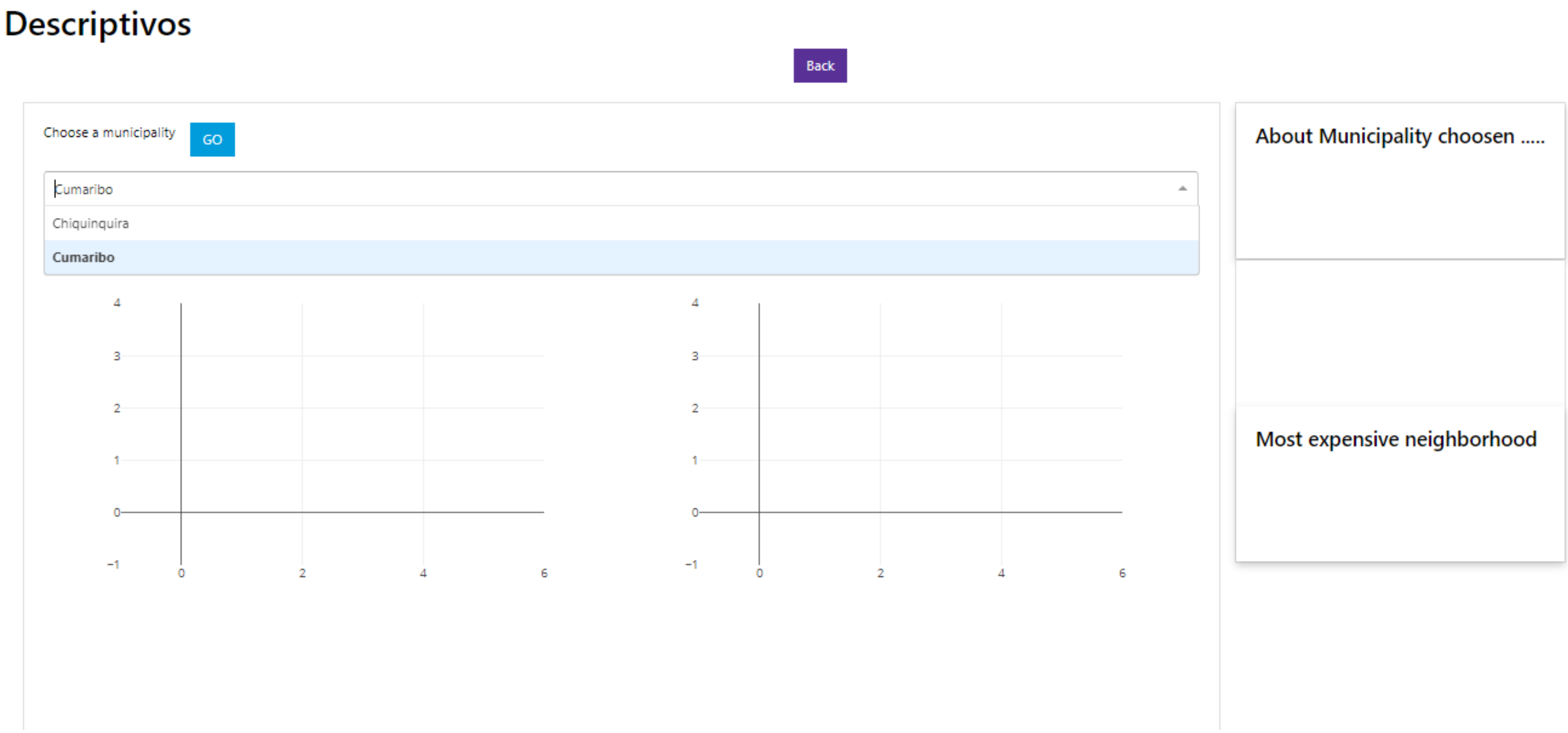
Welcome to AvalPredict

Choose a municipality

Chiquinquira

You can acces the app here: https://avalpredict.vatiolibre.com/

Some updates on the app(new space to show graphs and descriptions of the city)



Descriptivos

Choose a municipality

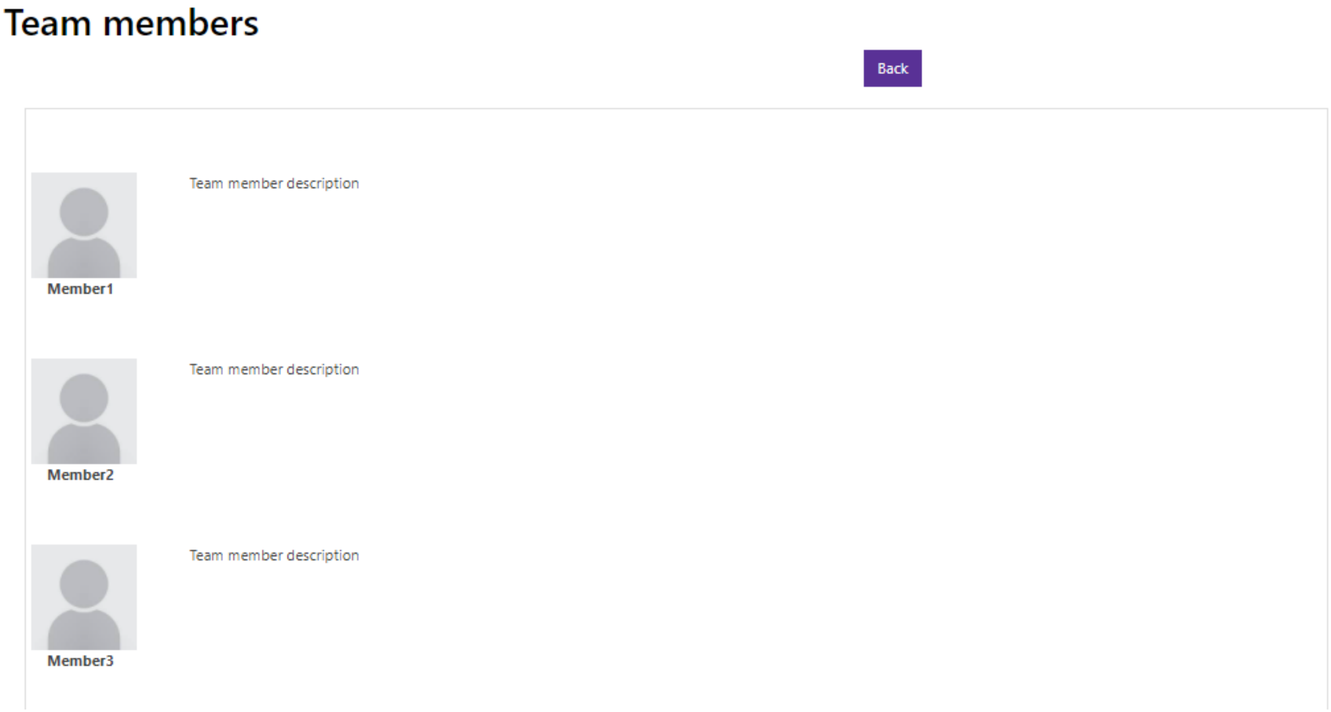
Cumaribo

About Municipality choosen .....

Most expensive neighborhood

You can acces the app here: https://avalpredict.vatiolibre.com/

Some updates on the app (space for team members)



Team members

Team member description

Member1

Team member description

Member2

Team member description

Member3