Final Report

9 April 2025

HOW DATA SCIENCE CAN HELP IN MULTIPURPOSE CADASTRE?

TEAM #165

Esteban Villalba, Juan Torres, Cindy Hernández, Gisselle Sánchez, Andrés Cordoba, Camilo Suaréz.

Keywords: Cadastre, Appraisals, Real estate, Housing market

Summary

Land valuation and ownership is the cornerstone of every organized economy since the main production factors are land, work and capital. In Colombia, the IGAC institute is in charge of giving the right valuations for each property, and our purpose is to provide them with some resources.

Contents

Introduction

This project is developed in the framework of The Agustín Codazzi Geographic Institute (IGAC), it is a Colombian entity in charge of the administration of the geographical and cadastral information. Particularly, the IGAC is responsible for the study of the real estate dynamics of the country.

An efficient and effective Cadastre system is an essential part for the territorial planning, fiscal management, national stability and social welfare. Without a well-functioning land administration system many of the government challenges will not be met. As this project is related to the Colombian cadastre system is wise to zoom in on the status of its land administration policy.

Today in Colombia, the cadastre system is significantly underdeveloped as can be seen in the Figure 1. It is out-dated for 63.9%, and it has not yet been implemented for 28.5% of Colombian territory [1]. Additionally, mapping and data gathering are inefficient; there is no optimal standardized process used in order make it less time consuming. Historically, Colombian cadastre has been outdated in about 11 years, and different entities have different valuations.

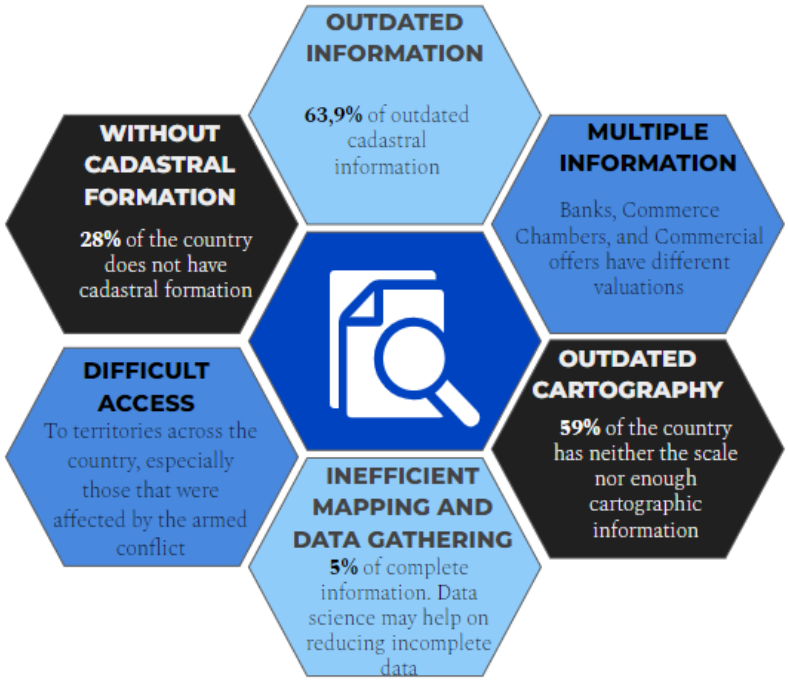

Figure 1: Status of the Colombian Cadastre System

The weaknesses of this tool have not help to control and repair some Colombian problems such as, forced displacement, illegal mining, dispossession of land, lack of infrastructure and property rights protection policies among others.

Our approach to this problem

Our objective is to provide the IGAC with a dashboard and a model that can allow them to forecast the commercial appraisals and support the decision-making process related to the territorial planning and economic policies. Additionally, the identification of the trends and changes in the residential real estate dynamics will improve the effectiveness of the cadastral managers in the use and update of cadastral information.

The IGAC provided us with 4 data-sets form different sources: Commercial chambers, Banks, Properati web scrapping and IGAC valuations. The team noticed that it was not efficient to work with four different datasets, so we worked on the merge of three of them to have a stronger portafolio. This merge brings new records to the model which gives us a greater sample, the whole dataset now counts with nearly 545.000 records.

Initially, we conducted a cleaning process and an Exploratory Data Analysis (EDA) for every data-set. This exploration showed us that there were a lot of missing values and that the data-set Commercial chambers was not relevant to the project. Based on the relevant features and characteristics extracted from the first series of EDAs, we conduct an in-depth EDA on the relevant variables using the merged data-set.

Subsequently, we worked on the design and implementation of the Front End for user interaction with the applications Power BI and Dash integrated with an API. This application displays the main features obtained from the data. Additionally, we worked on the prediction model fitting by creating proper data aggregations and simulating several modeling techniques.

1 Application Overview

The Web application was developed with Dash and FastAPI, it lets the user have a quick time analysis of the data obtained and processed. Web app can be found in the following url http://http://20.185.89.118:8050/

The web app is divided in 2 sections, a prediction and a descriptive section. The prediction section uses a pre-trained random forest model to determine a posible price for a property, it requires:

- Type of property: Select between Apartment or house.
- Latitude: Float value.
- Longitude: Float value.
- Number of rooms: Integer value from 1 to 14.
- Number of bathrooms: Integer value from 1 to 20.

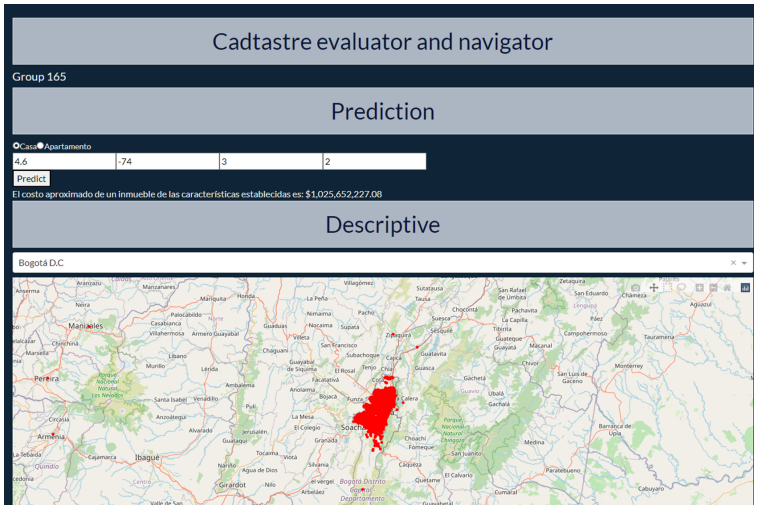

Figure 2: App landpage

For the quick time analysis (descriptive section), a state value is needed to returns the following (for this example, Antioquia State is being used):

– Ubication of the appraisal:

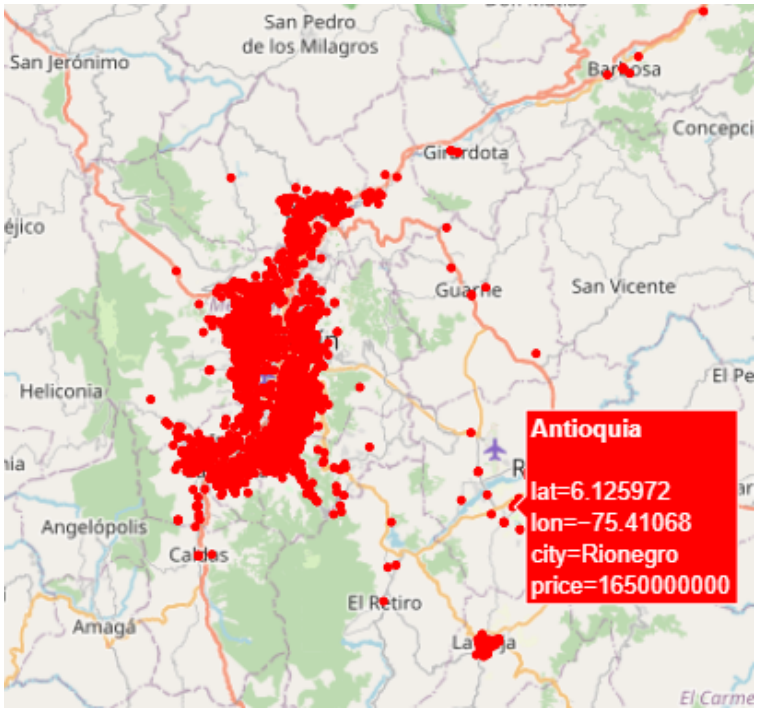

Figure 3: Ubication of the appraisals in Antioquia

– Boxplot of price per city:

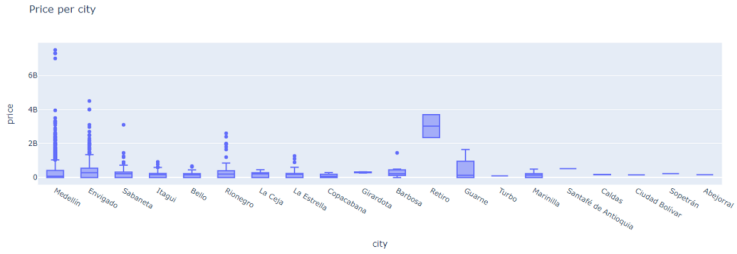

Figure 4: Price variation between cities in Antioquia

– Distribution of price between property types:

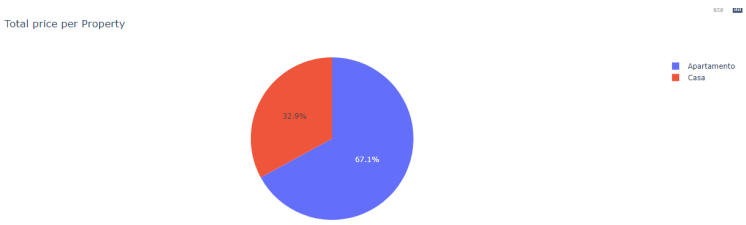

Figure 5: Distribution of prices between property types in Antioquia

– Boxplot of price per property type:

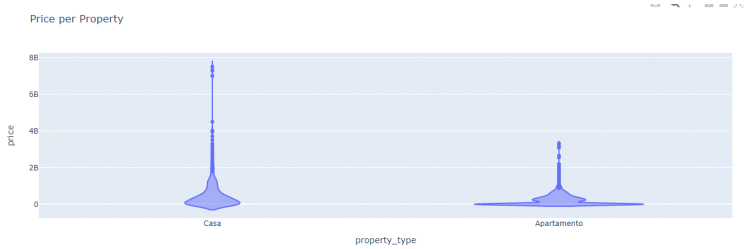

Figure 6: Price variation between property types in Antioquia

– Boxplot of price per bathroom amount:

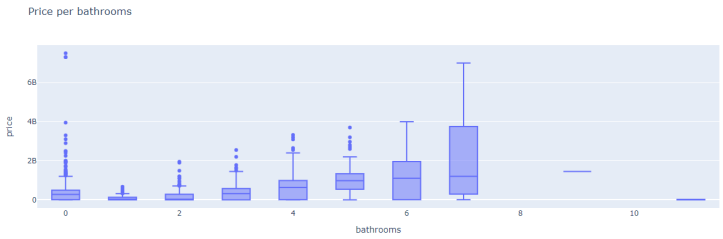

Figure 7: Price variation depending of the bathrooms amount in Antioquia

– Boxplot of price per bedroom amount:

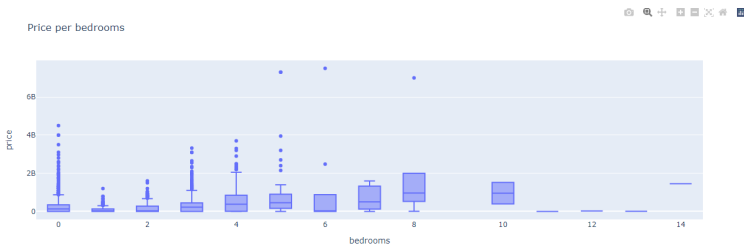

Figure 8: Price variation depending of the bedrooms amount in Antioquia

2 Data Engineering

This section covers a description about the process followed to create a compiled dataset using the information provided by the IGAC and then describes the application created to visualize and forecast commercial real estate appraisals.

2.1 Database

The IGAC provided us with 4 datasets as follows: Commerce chamber's table, Properati's database, Appraisals table and Bank database tables.

One of the initial concerns of this project was the high amount of missing data, as we can see in Figure 1 the percentages of missing data per variable in the Preoperati table are high. This situation was also noticed in the other databases. Therefore, we make as our priority to adapt the information to increase the amount of records with the most and better information throughout the creation of a merged dataset.

Variable	**Description**	**Missing values**	**Missing values (%)**
id	ID of every offer posted	0	0%
start_date	Date where the offer starts	0	0%
end_date	Data until the offer is still available	0	0%
created_on	Date when the offer was created	0	0%
lat	Latitude	0	0%
lon	Longitude	0	0%
l1	Country	0	0%
l2	Department	0	0%
l3	City/municipality	9663	7.60%
l4	Area of the city	79705	62.60%
l5	Neightborhood I	98627	77.40%
l6	Neightborhood II	117948	92.60%
rooms	Number of rooms	86500	67.90%
bedrooms	Number of bedrooms	60604	47.60%
bathrooms	Number of bathrooms	29152	22.90%
surface_total	Total surface of the property (m2)	121945	95.70%
surface_covered	Surface covered by the property	119376	93.70%

Table 1: Missing Values Properati's table

Three databases were considered to obtain the compiled database. One of them is the Bank database, which is a combination of tables from 5 different banks. The other is the Appraisal

table is a base obtained directly by the IGAC due to field reports. Finally, we have the last table that refers to web scraping in the website Properati.

Commercial	**Bank**	**Properati**
Property Type	Property Type	Property Type
Price	Price	Price
Surface total	Surface total	Surface total
State	State	Bedrooms
City	City	Lon
Address	Address	Lat
	Bank name	Bathrooms

Table 2: Important columns in each table

Using the merge function in Python was possible to obtain the compiled table. Having as key fields Property Type, Price, Surface total, State, City, Bedrooms, Lon, Lat, Bathrooms. This merge brings new records to the model which gives us a greater sample, the whole dataset now counts with nearly 545.000 records.

2.2 Interactive Front-end

Our Team developed a Power Bi visualization that uses the data processed from the Properati database that contains the filtered data summarized in the Figure 9. The power BI application has some visualizations and sale value predictions of the housing real estate market in the Colombian territory.

Figure 9: Data used in the visualization

The visualization allows the user to filter by:

- Type of property: Apartment or house.
- Number of rooms: from 1 to 14.
- Number of bathrooms: from 1 to 20.
- Department: 32 departments in Colombia.
- Date of publication: July 2020 to August 2021

Once the user has selected the filters desired, the application will display some figures and visualizations. Let´s look its use with the example showed in the Figure 10.

Figure 10: Data used in the visualization

The number of apartments under the conditions: from 2 to 4 rooms, from 1 to 4 bathrooms, located in the Bolivar Department, is 907 and its average price is 685 million COP.

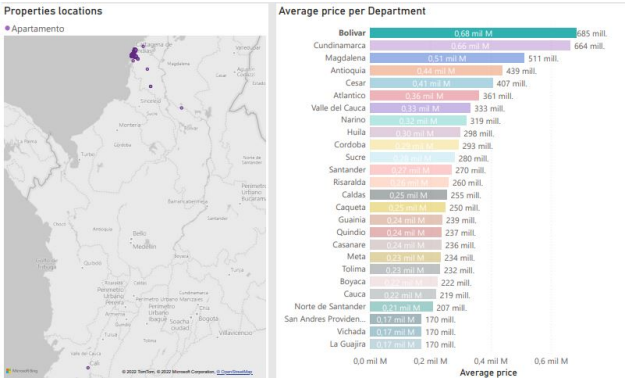

Figure 11: Properties location and Average price per filters

The application provides some visualizations such as the apartment's location and its average price as presented in the Figure 11. Other additional visualizations that the application displays are:

- Price distribution
- Price distribution per number of rooms
- Price distribution per number of bathrooms
- Average price per month
- Distribution of propeties per number of rooms and bathrooms

A second part of the visualization has a housing pricing prediction tool that allows the user to filter by:

– Type of property: Apartment or house.
– Number of rooms: from 1 to 14.
– Number of bathrooms: from 1 to 20.

– Location: latitude and longitude.

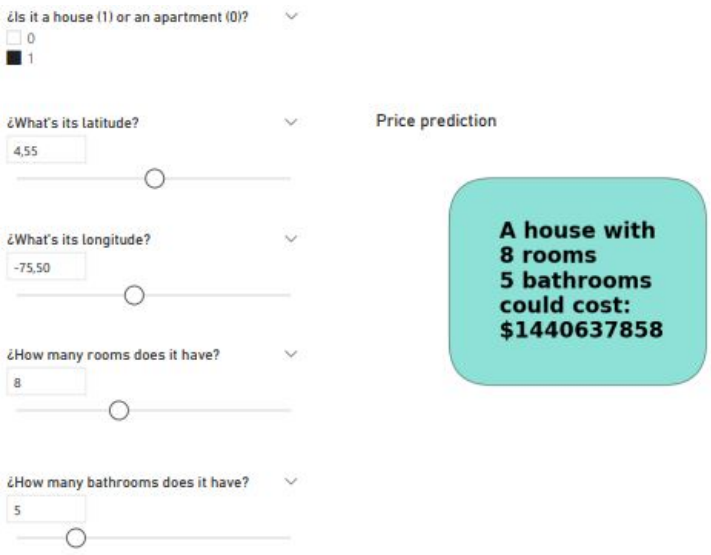

Figure 12: Housing pricing prediction application

The last prediction shows us that the price of an apartment with 8 rooms and 5 bathrooms in Colombia would cost 1,440,637,858 million COP.

3 Data Analysis & Computation

In this section we provide an exposition of the tools used to clean and explore the data.

3.1 Datasets + Data Wrangling Cleaning

This subsection contains a description about the data wrangling process and the other features added to each table.

The **Commerce Chambers' database** was composed by of 4 tables. The following procedures were conducted as part of data preparation and wrangling:

1. Remove the columns that don't contain relevant information using the drop function.
2. Rename the resulting columns for easier identification.
3. Merge the 4 tables by using the append function.
4. The field `ULT-ANO_REN` was necessary to remove empty data using the fillna and replace function (in python - pandas).
5. For the column `FECHA- AÑO DE MATRICULA` it was necessary to separate the first four characters because in some cases it had the whole date. Also, the column `CIIU` was separated by the five first characters because the `CIIU` identification was needed.

The **Properati's database** contains data that was obtained using web scraping. It consists of 22 columns that contain the following variables:

- id: unique for each property
- start date - end date - created on: the dates where the offer was posted and when will it end
- lat - lon: latitude and longitude where each property is located
- I1 - I6: Location of the property (country, department, city, neighborhood)
- rooms - bedrooms - bathrooms: number of rooms, bedrooms and bathrooms
- surface total - surface covered: area of the property
- price - currency: price of the offer and its currency
- proper type: the type of property (ej house, apartment, shop, among others) each.
- operation type: if the property is going to be rented or sold
- description: A description of the building

To treat the data we used different unix commands to remove the information regarding the title and the description of the offers so it would be easier to analyse the data [1]. We focus our analysis on housing properties (houses and apartments) being sold because our objective is to predict the price of those kind of buildings. So of the 363222 records supplied, we use 127271.

According to the table, some latitude and longitude values do not correspond to the Colombian territory, so they were limited: latitude between -4.2 and 13.34 and longitude between -81.36 and -67.49.

Finally, according to the table, it is possible to observe that some values that correspond to the price of the houses seem erroneous, so a filter was made taking into account the value of the prices greater than 0.5% of the data and less than 99.5% of the data. As a result of these filters, we obtained 66287 records.

The database **Commercial appraisals** was cleaned following the next steps:

1. Analize every field to decide if it is relevant or not for the project.
2. Remove columns which don't contain any relevant information.
3. Rename columns in order to make them better identifiable.

As the table for appraisals was organized in terms of distribution, variables and furthermore it lacked some fields, like location of the property (in terms of latitude and longitude), we have decided to not introduce any other column to the database.

The five **Bank appraisal tables** given where joined in one table, the challenging part was setting the matching fields between them, since they used different formatting and coding for the same fields, or in some cases they didn't use same fields

[1] The unix commands are included here

- Fields like Date (year), State, municipalities and property types where needed to be standardized in their format and also had some of their values fixed to match the classification of the other banks.
- There were issues related to values which were invalid for the fields, those values were dropped or fixed with the available data, for example for the socio economic class of the property (in Colombia property socio economic classes are between the categories 1 to 6).

In order to achieve a better understanding of the data, data was completed and some fields where created through different techniques

1. One of the field created was the cost per m2 for the properties that counted with a construction area and a commercial appraisal from the bank.
2. Also an alternative to obtain exact localizations of the properties, would be the usage of geocoding libraries to set this field for all the records to perform further analysis.
3. To identify each row, an artificial ID field was created.

3.2 Exploratory Data Analysis

In the following EDA we are going to describe explorations about every data-set followed by an in-deepEDA.

The **Commerce Chambers' table** contains general characteristics for the commercial properties allocated in Pereira and Villavicencio from 2019 to 2020. It contains variables such us property Id, business name, department, town, address, business activity and year of registration. We noted that the most relevant relations between variables are the number of commercial registrations renewed related to the year and municipality.

Considering that this dataset contains mostly information from Pereira and Villavicencio and that the data is especially oriented to the date of registration to the Commerce Chamber we have decided not to use this dataset in the project.

The **Properati's table** has information about the prices of different types of properties that are being offered for sale or rent between July 2020 to August 2021. Plotting the data, as can be seen in Figure 13 some of the points seem to be misplaced. Besides that, there are a few offers (97805 out of 363222) that don't include the latitude or longitude. These data were excluded for further analysis.

Figure 13: Locations map

Now, analyzing the quantity of properties regarding to the property and operation type as is showed in the Figure 14, we can notice that apartments being rented or sold and houses being sold accounted for 71.63%. Hence, we note that a large proportion of the data comes from apartments and houses, while just under a quarter of the data relates to other property types.

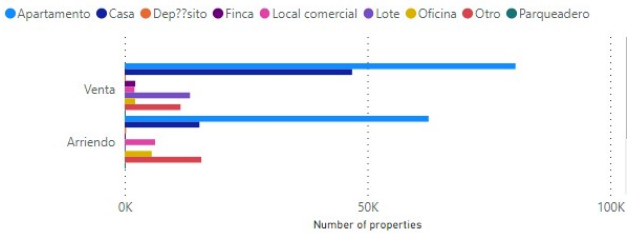

Figure 14: Properties published per type and operation

We also analyzed the number of houses and apartments regarding of their location. Counting the number of offers, we noticed most of them are located in Cundinamarca, Antioquia and Valle del Cauca (63.37% of the data) as is displayed in Figure 15.

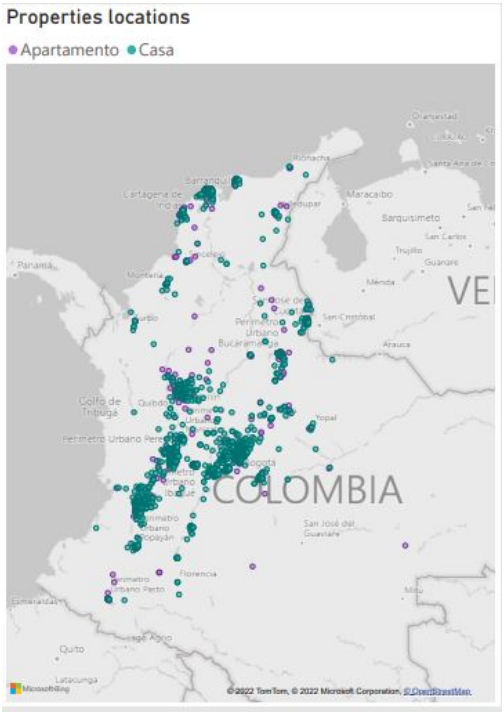

Figure 15: Properties per location

There is data related to the number of rooms, bedrooms and bathrooms but unfortunately it has a lot of missing values:

- Rooms: 202698 missing values out of 265217
- Bedrooms: 143103 missing values out of 265217
- Bathrooms 68055 missing values out of 265217

This dataset also contains information about surface covered and surface total but most of the data is missing. 246.152 out of 265217 are missing, hence we cannot rely on this variable..

In terms of prices, as we can observe in Figure 16 the temporary rent is less expensive than annual rent. The average price also suggests that the house rental value is higher than the apartment rental value, as are the sales prices. Turning to the warehouses, the given table denotes that it has the greatest average rental and sales price compared to the other property types.

property_type	operation_type	Promedio de price
Dep??sito	Venta	3724587769,18
Lote	Venta	1696365546,24
Otro	Venta	1547388091,48
Local comercial	Venta	1372666888,94
Finca	Venta	1296448653,84
Oficina	Venta	943081787,64
Parqueadero	Venta	875281031,39
Casa	Venta	682536863,51
Apartamento	Venta	461893180,80
Dep??sito	Arriendo	16673403,83
Lote	Arriendo	14519485,60
Oficina	Arriendo	7249358,28
Otro	Arriendo	6553148,17
Parqueadero	Arriendo	5641454,55
Local comercial	Arriendo	5577819,84
Finca	Arriendo	4477491,97
Dep??sito	Arriendo temporal	3500000,00
Casa	Arriendo	3113583,62
Apartamento	Arriendo	1827071,50
Otro	Arriendo temporal	1127272,73
Apartamento	Arriendo temporal	1051782,61
Casa	Arriendo temporal	850000,00

Figure 16: Average price per property type and operation

The **Commercial appraisals table** contains the appraisal for each property, with information about, appraisal's date, location (Address, Department, Town, Latitude, Longitude), property type, area m^2, construction, ground and commercial value.

```
AVAL_DIRECCION                object
AVAL_MATRICULAINMOBILIARIA    object
AVAL_AREATERRENO             float64
TEDE_ID                       object
AVAL_AREACONSTRUCCION        float64
VETU_ID                       object
AVAL_VALORTERRENO              int64
AVAL_VALORCONSTRUCCION         int64
AVAL_VALORAVALUOCOMERCIAL      int64
REGION                        object
DEPARTAMENTO                  object
MUNICIPIO                     object
AVAL_YEAR_FECHA                int64
AVAL_MONTH_FECHA               int64
dtype: object
```

Figure 17: Type of Data per column in the dataset

We can see in the figure 18 that most of the properties have an appraisal below 200 Million COP and that the distribution is right skewed with the mode located around 100 Million.

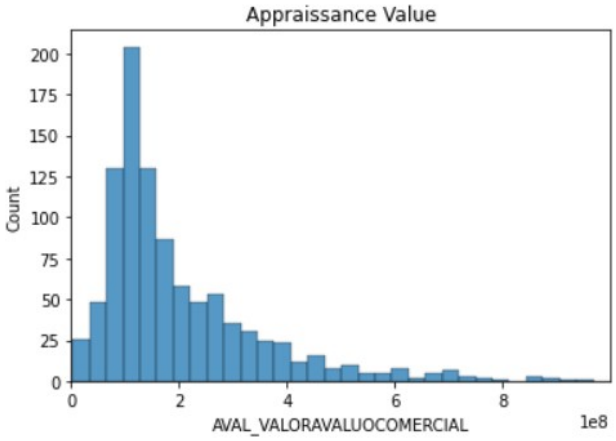

Figure 18: Distribution of appraisals per price (COP)

In the Figure 19 we find that most of the properties are housing properties, also most of them are less than 1 year old. Most importantly, we can observe that the biggest appraisal's sample was taken in the central region of Colombia, more specifically in Bogotá, Valle del Cauca, Cundinamarca and Santander.

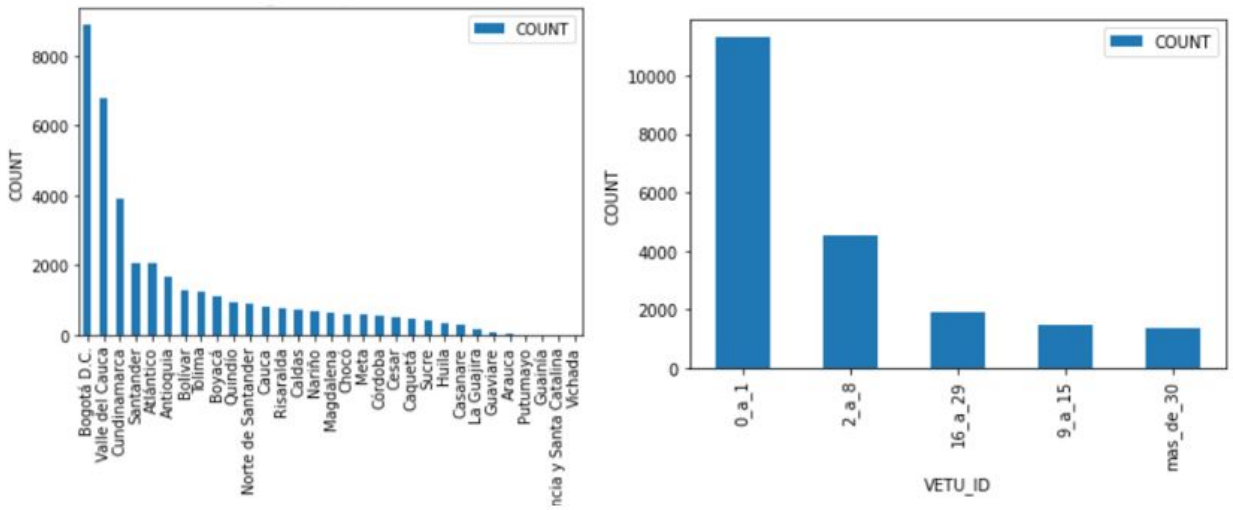

Figure 19: Count per categories department and age

The **Bank appraisals table** contains joined information from the banks Agrario, Caja Social, Davivienda, FNA and Banco de Bogotá. For a total amount of 81,873 records obtained from the 5 bank appraisal tables, the Figure 20 is presented to summarize the main characteristics of the data.

Year	Records
2019	67907
2020	13888

Bank Name	Records
Davivienda	43665
Fondo nacional del ahorro	22834
Banco Agrario	15015
Banco Caja Social	281
Banco de Bogotá	78

Figure 20: Amount of records per Year and Bank

The data obtained mainly is from the year 2019, this could lead to a bias in the values respect nowadays values since inflation rates for the country have increased due to Covid 19 pandemic, also, the records are mainly obtained from 3 out of 5 banks.

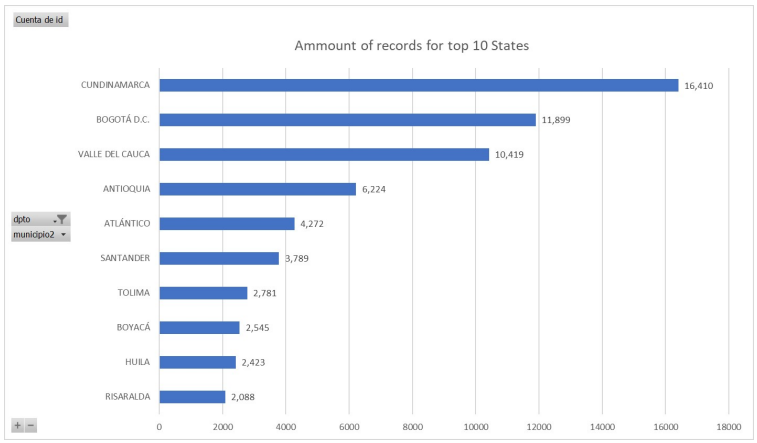

Figure 21: Amount of records for top 10 States

In this data-set we can notice again that almost 76% of the records are distributed in the top 10 states, hence, the data is heavily oriented to big cities in Colombia, such as Bogotá, Medellín and Cali.

From Figure 22, it can be clearly seen that more than 80% of the data is related to housing. Is interesting to notice that this distribution is expected due houses are to the most common good to be used as a guarantee for a bank.

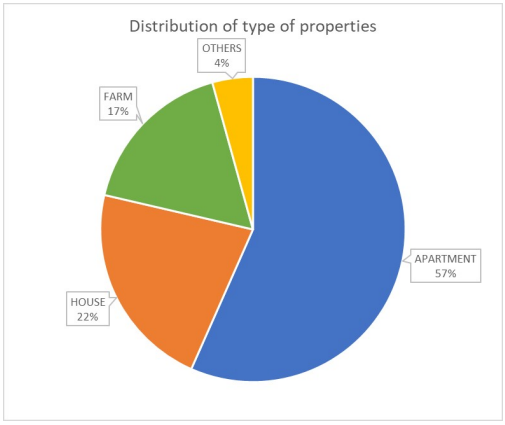

Figure 22: Distribution of Property type

As is presented in the Figure 23 , the socioeconomic stratum 3 has the highest amount of properties while the socioeconomic class 6 has a very small proportion of the records.

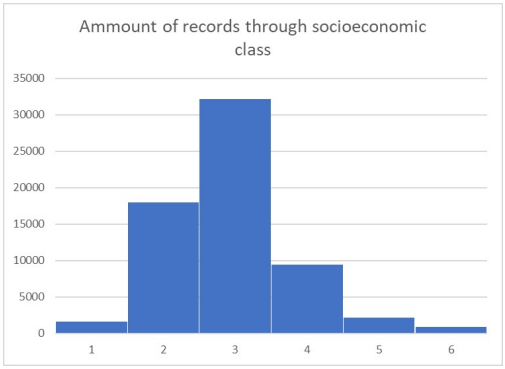

Figure 23: Amount of records throughout socioeconomic classes

With regards to the average appraisal through socioeconomic classes, the Figure 24 clearly shows that the higher the economic stratum, the higher the appraisal price. a similar trend can be observed in the average cost per square meter throughout socioeconomic class.

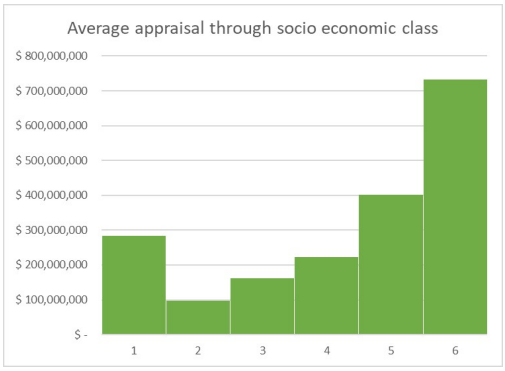

Figure 24: Average appraisal throughout socioeconomic classes

3.2.1 In-depth EDA

In order to discover some relations between variables and answer the question related to the variables that can influence the price of a property we have displayed the following correlation matrix using the merged data-set. As the Figure 25 suggests, generally speaking the price can be influenced by the number of bedrooms, bathrooms and the property's surface area and its location. This information was an essential part of the model process we are going to describe in the next section.

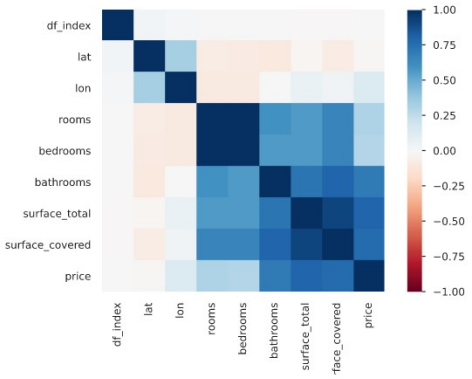

Figure 25: Correlation Matrix

As we are specially interested in the appraisals, we also analysed its distribution versus different variables. In the Figure 26 we can find that commercial property appraisals have a wider distribution and a higher mean, followed by industrial property and lastly by housing property.

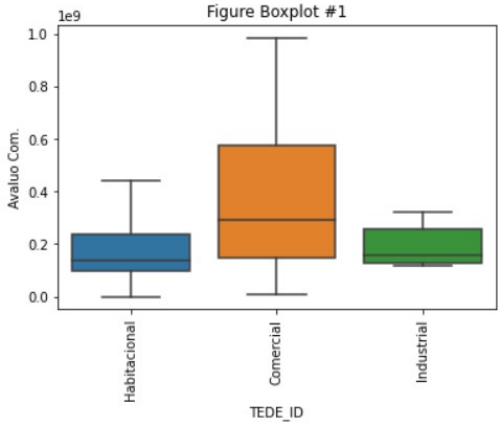

Figure 26: Appraisal per Real Estate Sector

In the Figure 27 we can find that the youngest properties have the lowest appraisals but we are aware that this might be because most of the new properties are smaller, however we can also find that properties between 9 and 15 years old are the ones with the better valuations and the highest variations as well.

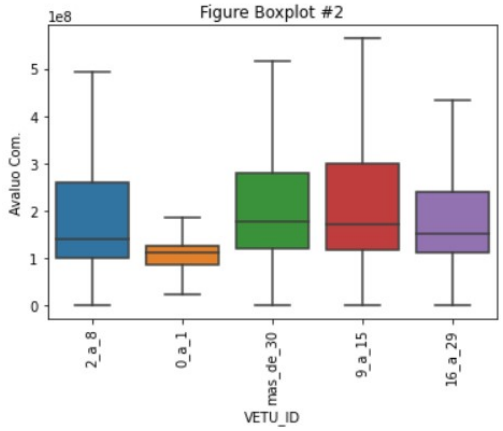

Figure 27: Property value and age

In the Figure 28 we can see that the highest appraisals are located in the Eje Cafetero region, followed by the center region, this one specifically has the wider variation among the Colombian regions. Regarding the lowest appraised properties are the ones located in the south-western region.

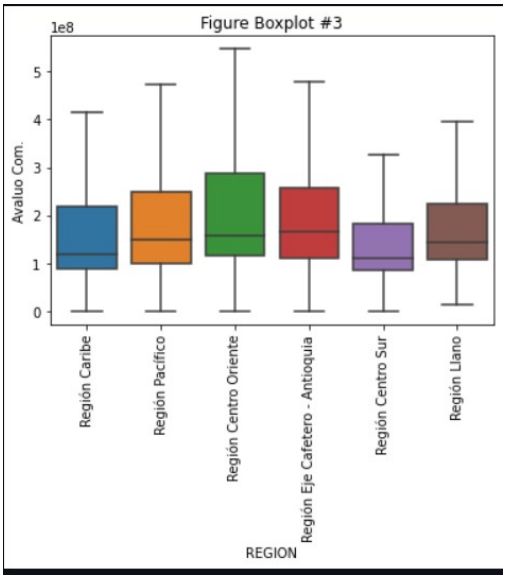

Figure 28: Appraisals per region

More specifically, Bogota's properties are the ones with the higher appraisals among the states; the Capital City has also the highest variations within its own neighborhoods. States like Chocó, Córdoba, Magdalena, Cesar and Antioquia have a similar distribution in terms of mean appraisal and its variation.

From an additional exploration we can say that appraisals taken in December were higher than the other ones, also the appraisals have been changing throughout the years. Regarding the property maturity we could find that properties more than 30 years old were the only

ones with a mean appraisal below 1 year old properties. Regarding region we could find that only 3 regions had an average appraisal below the Center-East region and those were South-Center, Pacific and Llano.The other ones had an appraisal above Center -East region.

3.3 Statistical Analysis & Machine Learning

Since our database contains about 80% of properties in the residential sector and the EDA showed a wide variation in the prices of commercial and industrial properties with respect to housing, the team decided to frame this project in the forecasting of commercial appraisals of houses and apartments in the Colombian territory.

Taking into account that there were many missing values in the variable *bathrooms*, it was decided to make an imputation of these values. We start from the assumption that a property should have a similar number of bathrooms to the properties that are geographically close, correspond to the same type of real estate and have a similar number of rooms. To do the imputation we used K nearest neighbors with 3 neighbors. This same process was done with the variable *rooms* in order to complete the missing values and in the same way we proceeded with the variable *surface area.*

After this, we proceeded to separate the data into two groups training and testing. Our target variable, *price*, was converted to logarithmic values. We also created two columns coding whether the property is an apartment or a house (one hot enconder). We decided to evaluate different models using a cross-validation with 5 groups varying the different parameters of the models, as follows:

K nearest neighbors	**Random Forest Regression**	**XGBoost**
Number of neighbors: 4,8,16 weights: uniform, distance	Number of estimators: 50, 100, 200 Maximun deph: 20, 50, 100	Learning rate: 0.1, 0.25 Number of estimators: 100, 200 Maximum depth: 10, 20, 50
Better model: 16 neighbors, distance	Best model: number of estimators 200, maximum depth 20	Best model: learning rate 0.1, maximum depth 20, number of estimators 100

Table 3: Models using a cross-validation

The results when evaluating the best selected model with the test variables can be seen in the next table.

	KNN	**Random forest**	**XGBoost**
explained_variance	0.6984	0.7398	0.7363
$r2$	0.697	0.737	0.7334
MAE	143016513.6148	141288790.3283	142458381.9359
MSE	1.0282566929109307e+17	8.923921467488286e+16	9.044958456029189e+16
RMSE	320664418.4987	298729333.4691	300748374.1607

Table 4: Test variables per model

The **Random forest model** explains $\approx$ 73% of the variance, therefore it is the model selected. In the model, we find that the least relevant variables in defining the price of a house are *longitude* and *latitude*. While factors such as the *type of housing* and its characteristics seem to affect the model the most.

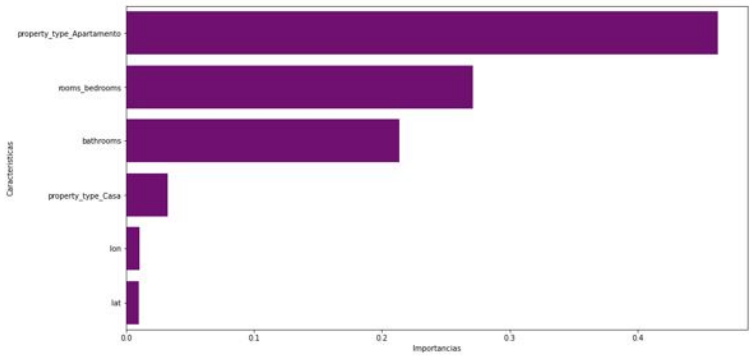

Figure 29: Variables relevancy

For the examination of the model performance, we focus on the graphical method that use residuals. Therefore we use the difference between the predicted price and the actual price of the responsive variable. As is shown in Figure 30 the quality of predictions is reasonably close to the actual values because the residuals are concentrated around 0.

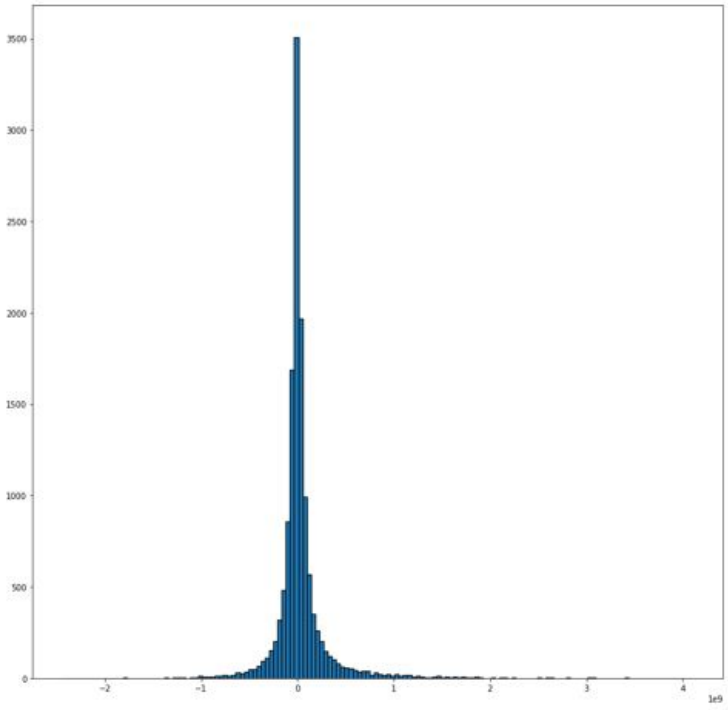

Figure 30: Residuals Plot

Conclusions and Future Work

The visualization tool, data exploration and forecast application might provide a better decision making process for the IGAC Institute and house holding market dynamics as well.

Price Prediction will help on fast information about a property without having to go directly to the place, this means time and economic efficiency.

For future work on this particular topic, it might be interesting preforming automated web scrapping, adding more registers, improving the price prediction. Property valuation and characteristics maybe obtained by images, opening a huge work to do in image processing real estate.

There is very little data on properties located in the non-central regions of Colombia. This was not very favorable in order to provide a commercial appraisal prediction for remote regions of the country. It is important that the entities in charge continue with the work of cadastral formation across regions. Cadastral information is very important to help reducing inequality based on a big problem in Colombian history, our land.

General information such as latitude, longitude, country and department can be displayed in its entirely but the data related to the detail such as neighborhoods, number of rooms and total area are with a percentage of approximately 90% of empty data.

The greater the number of variables provided to the model, the greater the accuracy of the appraisal result, however, at this time it was not possible to have information regarding stratification and other departments far from the capital cities.

In literature there are other variables, which due to the quality of our data we did not explore, such as safety, access to roads, availability of schools, hospitals and shopping centers. It is suggested for future projects to use which might help to capture neighbourhood effects.

The idea of appending the data set did not have the expected results. For future opportunities, it is recommended in these cases to make separate predictions for each data source.

Property valuation helps IGAC compares its own record to stablish tax information.
Colombian Cadastral system has huge challenges ans is essential the cooperation of different entities to help to heal the gaps that this system has and contribute to the territorial planning, fiscal management, planning infrastructure, agricultural support and peace agreement implementation among others.

References

[1] DNP. Propuesta financiación catastro multipropósito, Bogotá, Colombia. 2017.