Escuela de Negocios. Master in Management + Analytics

Real Estate Valuation in Buenos Aires, an Interactive Tool Development

Autoría: Gonzalvo, Francisco
Año: *2024*

Real Estate Valuation in Buenos Aires, an Interactive Tool Development

Master in Management + Analytics

Universidad Torcuato Di Tella

Student: Francisco Gonzalvo

Tutor: Emmanuel Iarussi

Abstract

This thesis explores the application of machine learning models for real estate valuation in Buenos Aires, aiming to develop a user-centric tool to assist buyers and investors in making informed decisions. Traditional regression models, while useful, often fail to capture the complex, nonlinear relationships inherent in real estate data. Consequently, we employed advanced machine learning techniques, including XGBoost, Random Forest, and Support Vector Machines (SVM), selected for their robustness and efficiency in handling large datasets.

Our input data consists of 64,358 property listings from the e-commerce platform Mercado Libre, obtained through a combination of Python scripts and the platform's own API.

An initial exploratory analysis provided insights into the interactions between different variables and the overall cleanliness of the data, revealing necessary adjustments. This analysis included examining the geographical distribution of listings and interactions between variables such as price, size, and age. The exploratory analysis also identified misclassified neighborhoods, which were corrected using custom polygons and geospatial analytics techniques, particularly in areas like Palermo and Belgrano.

Through testing, the XGBoost algorithm emerged as the best-performing model in this context, becoming the model of choice for the results presented in the tool. To ensure the tool's relevance and practicality, a complementary survey was conducted to understand buyer preferences.

Finally, the insights from the exploratory analysis, model outputs, and survey data were integrated to build the tool using Tableau visualization software. This allowed us to create a dynamic map that aggregates the information, providing a comprehensive and interactive platform for users.

Introduction

Within the last decade the number of properties for sale in the city of Buenos Aires has doubled. This unprecedented behavior was fueled by many different factors with the main one being the economical. Argentina's instability is one of the main responsibles of this issue: with every projection about the country indicating a decline in economic prosperity in the following years, there is no wonder why people may want to sell their assets while they still hold some value.

Furthermore, measures by the government to take control of the situation (which didn't consider market dynamics into account) made the situation even worse. A good example of this is what is locally known as the "Ley de Alquileres". Introduced in 2020, this law attempted to protect tenants from increasing living cost prices caused by inflation by limiting the number of rent increases legally allowed over a year to a single one. It also attempted to limit the number of requirements a homeowner could ask in order to rent an apartment. While in good nature, the only thing this new law caused was to strongly discourage home owners to withdraw their properties from the rental market with some of them opting to sell them as well.

Nevertheless, there are still many reasons why someone would want to buy a property in Buenos Aires. It is one of the most important capitals of the world and one of the most expensive cities in South America. Even from a speculative point of view, the prices can be a good opportunity for a risky long term investment. A side of the reason why someone may want to invest in Buenos Aires real estate, with over 180k properties for sale, attempting to optimize your budget as a buyer can be an overwhelming task.

If we scale the problem to institutional investment firms, scanning the market for potential opportunities it becomes even worse. We believe we can leverage machine learning algorithms to simplify the process of finding the publications with the most potential in the market.

Research problem, objectives and questions

By analyzing the dynamics underlying the real estate market in Buenos Aires (including the various factors that contribute to the demand for properties in the city) we can acquire a comprehensive understanding of the occuring market dynamics. This understanding can then be used to leverage machine learning algorithms to help us simplify the decision-making processes involved in the purchasing of real estate and achieve better outcomes

The objective of this study is twofold: first, evaluating apartments and identifying the best opportunities in the real estate market, and second, to create a user-friendly and interactive tool that enables the end users to find the best opportunities based on their personal preferences.

Opportunity detection

The first objective of this research is to evaluate apartments and determine which ones represent good investment opportunities in the market by understanding the way in which different variables apartment valuations. These can include different aspects such as location, size, amenities and many other characteristics. The research will involve collecting and analyzing a dataset containing different apartment listings and its characteristics while using machine learning algorithms for predictive models that can help estimate the potential future value of apartments.

The outcome of this research objective will provide valuable insights for investors, homebuyers, and real estate professionals by highlighting the most promising investment opportunities in the market. The findings will help guide decision-making processes and enable stakeholders to make informed choices based on reliable data analysis.

Interactive Tool Development

The second objective of this work is to create an interactive tool that allows clients to explore the results of the previous analysis. By designing interactive features, clients will be able to customize and filter the data based on their specific preferences and requirements. This is especially important when considering private clients as end users, since, contrary to institutional investors, may have particular preferences that outweighs the value perception criteria of machine learning models.

This development will involve integrating the analyzed apartment data with data software. These tools could include an interactive map combined with a set of filters that enable users to explore the different apartment listings. The objective is to create a visually engaging and informative platform that allows clients to make informed decisions while taking into account personal preferences.

Methodology

In order for us to narrow down the number of options and determine which of the available properties at a given moment are actually a good investment, we will follow the next approach:

1. Data sourcing:

   In order to figure out how we can reduce the number of options a buyer

needs to look at, we will first need to sample what properties are available for them. For the purpose of this experiment we will make use of Mercado Libre's API.

2. Data Cleaning and Preprocessing:

   Once we gather the available data for the buyer, we will clean it in order to ensure its quality and reliability. This step includes handling missing values, removing outliers and making sure that the data is reliable enough for us to obtain quality insights and make informed decisions.

3. Exploratory Data Analysis:

   We will conduct a first analysis of the data in order to get a better understanding of the underlying patterns and trends within it. We will try to answer some of the motivating research questions for the experiment and start to have a better understanding of what are the key factors that influence property values.

4. Feature selection:

   Based on what we learned from the exploratory analysis, we will perform feature selection to narrow down which features contribute significantly to the investment potential of a property.

5. Model development and evaluation:

   We will explore how different models perform at valuing real estate. We will evaluate the results of each of the tested alternatives in order to come up with a final recommendation.

6. Data Visualisation:

   We will explore how can we leverage data visualisation tools to deliver the results of the model on a more user-friendly way

Data Sources

Our main source of data for this experiment will be Mercado Libre's API. Mercado Libre is one of the largest ecommerce sites in Latin America and it is very popular in Argentina, particularly in the city of Buenos Aires, making it a perfect data source for this experiment. By accessing their API, we can extract valuable property-specific data such as property descriptions, images, location, price, size, number of bedrooms, amenities, and other relevant features. Mercado Libre's extensive database provides a wide range of real estate listings, enabling us to capture a diverse set of properties for analysis.

Technologies

The orchestration of the data collection, data cleaning, model development and data preparation for the visualisation tool was performed in python scripts making use of libraries such as numpy, requests, selenium, pandas, geopandas, xgboost and scikit learn. Once the data was ready, the visualization tool was developed in the visualization software Tableau, which leverages maps from Mapbox. [Annex 1]

Scope and Limitations

We aim to provide transparency and ensure that readers understand the contextual boundaries and potential constraints of our analysis. Recognizing these factors is essential for researchers, practitioners, and stakeholders to accurately interpret and apply the findings of our study:

- Geographic scope: We will only be focusing on the city of Buenos Aires for the analysis. SI

- Data Availability: For the following experiments we opted to use the city of Buenos Aires partly because of the large amount of data available for it.
- Property Type: We will be focusing on residential buildings. The same analysis could potentially be extrapolated for commercial use.
- Operation Type: We will be focusing on the sale of residential properties. The same analysis could potentially be extrapolated for rental properties
- Value Subjectivity: Personal preferences take a huge role when determining the relative value of property for an individual.
- Intertemporality: The analysis will be done by taking a snapshot at a specific point in time. The validity of the results may change over time.
- Data Source: Any bias in Mercado Libre´s user base will potentially be carried into the model

Valuation Methods

Traditional real estate valuation methods consist of several approaches that have been proven to be effective in assessing property values. A good understanding of the foundations of real estate valuation is important if we intend to leverage them into the decision making process.

These real estate valuation methods have been employed for many years to estimate the market value of different properties, while considering the various factors that influence their price. There are 3 main methodologies that take place when it comes to valuing real estate:

Market Comparison Approach

The market comparison approach, also known as the sales comparison approach, is a widely used method for estimating property values. It relies on the principle of comparing the subject property to recently sold comparable properties in the same market area. By analyzing factors such as location, size, condition, and amenities,

appraisers can determine a fair market value for the subject property. [1]

The modelling approaches we are going to analyze in the following experiment could be catalogued under this approach. By leveraging large datasets and advanced algorithms, machine learning models can automate and enhance the process of selecting comparable properties, considering numerous factors simultaneously, and identifying patterns and relationships that influence property values.

This enables real estate professionals to generate more accurate and efficient valuations by incorporating data-driven insights and capturing intricate nuances and interactions among various property features.

Income Capitalization Approach

The income capitalization approach is a commonly employed method when it comes to income-producing properties (Eg: commercial buildings, rental properties, and investment properties). This approach estimates the value of a property based on the potential income it can generate. It relies on the principle that the value of an income-producing property is derived from the present value of the income it is able to generate over time. [2][3]

Cost Approach

The cost approach (or “Replacement cost method”) considers the cost of replacing or reproducing a property as an indicator of its value. This approach is particularly relevant for unique or specialized properties where comparable sales data may be limited. It involves estimating the land value, considering the cost of construction, and accounting for depreciation and obsolescence. [4]

Quantitative Approach

In addition to the three methodologies mentioned above, more modern approaches make use of different technologies to complement valuations. These methods involve the use of data models and statistics to make estimations of property values. By feeding large datasets containing property attributes to regression and machine learning algorithms, we can generate predicted valuations based on the relationships within the data.

Many examples of quantitative approaches to real estate valuation can be found covering more general and simple approaches such as the use of different types of regressions in real estate valuation to more specific ones such as the article "A Room with a View" which tries to understand the impact of a single variable (In this case the view) in Real Estate appraisals. [5][6][7][8][9][10]

We will explore literature about these methods in the following section.

Literature Review: Machine Learning Models

Ever since the 1960's mathematical and statistical models began to make their way within the field of real estate. Introducing regression-based models into the realm of real estate allowed professionals to accurately predict property values and gain a deeper understanding of the key determining factors.

Regression-based models are designed to capture the relationships between a property's characteristics and the price tag it carries. Making use of features such as location, size and amenities (among many others), these models can estimate the value of a property based on its attributes. Unlike traditional valuation methods that rely on subjective judgments, regression models can incorporate a broad range of data to make predictions. They can handle complex interactions between variables allowing for a more robust prediction and providing a transparent and explainable framework to justify their results; as well as to provide insights into the relative importance of different features in determining property values. While regression models offer valuable insights and transparency in property valuation, they have certain limitations when compared to machine learning models. [15]

Using machine learning models over traditional valuation methods allow us to capture nonlinear relationships within the data. Machine learning models can identify complex patterns within the data, resulting in more accurate predictions. These models can also handle unstructured types of data such as images or text, further improving the predictive opportunities. Because of this flexibility, machine learning models are an reliable approach for real estate valuation.

Machine Learning Models for Real Estate Valuation

With the growing availability of large datasets and advancements in machine learning algorithms, these models are gaining popularity in the real estate industry when it comes to predicting property values. There are several machine learning models that could potentially be applied to real estate valuation:

Decision Trees

Decision trees can be used to predict the value of a property based on a set of features. They work by splitting the data into smaller subsets and creating a tree-like structure that helps determine the value of the property based on the different features.

Some of the ways decision trees have been applied in real estate valuation include modeling dependence in decision trees using copulas [20], comparing real asset valuation models using literature review [21], constructing adaptive decision-making models for real estate marketing [22], and using decision tree analysis to account for uncertainties in real estate development projects [23].

Support Vector Machines (SVM)

SVMs can identify the relationships between different features of a property and use them to predict its value by finding a hyperplane that separates different classes or values in the data.

Some examples for this approach include:

- Using SVM to forecast the price of real estate in China [37].
- Estimating the commercial real estate transaction price using SVM [38].

Random Forests

Random forests combine multiple decision trees to make a prediction. They can be used to estimate the value of a property based on various features such as location, size, age, and many others.

Some examples of ways which machine learning algorithms were used in this context are:

- Using random forest algorithms to analyze the local variables that influence the interaction between housing demand, supply, and price in London real estate market [45].
- Developing a house price prediction model using random forest algorithms in Surabaya City [46].
- Examining the latest machine learning algorithms, including random forest, as automatic valuation models for residential properties in South Korea [51].
- Demonstrating that random forest and artificial neural networks algorithms can be better alternatives over the hedonic regression analysis for prediction of house prices in the city of Boulder, Colorado [52].
- Using random forest with cross-validation as the best promising algorithm for predicting office rental prices in Kuala Lumpur, Malaysia [53].

XGBoost

XGBoost is an ensemble learning model that uses gradient boosting techniques to create a powerful predictive model. It is particularly effective in handling large datasets with complex relationships. XGBoost can be applied to real estate valuation by training on

historical property data, including various features like location, size, amenities, market trends, and historical sales data. The model learns to make accurate predictions by iteratively building weak models and combining their outputs.

Some examples of ways that XGBoost can be applied in this context are:

- Predicting house prices using XGBoost algorithm and hedonic regression pricing [55].
- Evaluating the possibility of applying machine learning algorithms, including XGBoost, in property mass valuation on small, underdeveloped markets [57].
- Determining the risk factors impacting the project value created by green buildings in Saudi Arabia using XGBoost algorithm [59].

Neural Networks

Neural networks can analyze multiple features of a property and provide a predicted value based on the patterns they identify. They can be used to identify complex relationships between features and property values.

Some examples of ways in which neural networks can be applied in real estate valuation include:

- Estimating real estate prices based on environmental quality of property location using an Artificial Neural Network (ANN) model [25].
- Developing a complex neural network model for mass appraisal and scenario forecasting of the urban real estate market value that adapts itself to space and time [27].
- Using machine learning algorithms, including neural networks, for predicting real estate values in tourism centers [32].
- Developing a predicting system for the estimated cost of real estate objects development using neural networks [34].

Market Adjustments

More recent work improved models such as the Pseudo Self Comparison Method (PSCM) [11], which divides the real estate valuation problem to:

- 1st Finding a similar transaction from the past (defined as a housing property that can most closely approximate the characteristics of the target housing).
- 2nd Adjusting its previous transaction price to be in sync with real estate market changes. This adjustment takes into consideration factors such as inflation, financing costs ,market trends, and fluctuations in supply and demand.

By employing the PSCM, researchers and practitioners aim to enhance the accuracy and reliability of real estate valuations by incorporating historical data and market dynamics.

This methodology is really popular within Russia because of the availability of the cadastral system of which is a centralized and unified database that records information about land plots, buildings, and other real estate properties. The Federal Service for State Registration, Cadastre, and Cartography (Rosreestr) is responsible for managing the cadastral system ensuring data is collected, updated, and made available to the public [13]. Unfortunately there is no equivalent public dataset in Buenos Aires for us to replicate this analysis in this context.

Recap and Comparison of Methodologies

With the rise of big data and advances in machine learning, predicting property values has become more sophisticated. In summary of the reviewed literature, Decision Trees are easy to understand and interpret and work ok on smaller datasets. however they tend to overfit and struggle with complex data.

SVMs can Handle high-dimensional data well and are more robust against overfitting however it isn't as easy to interpret. The same happens with Random Forests which reduce overfitting by using multiple decision trees and works well with large datasets however they are less transparent than a single decision tree.

Looking at XG boost, while its good at handling large datasets and includes features to prevent overfitting, it needs careful tuning and is more complex to set up. A similar thing happens with neural networks which work great for finding complex patterns in data and flexible enough to handle various data types but they Require lots of data and computational power

Each of these methods has its own set of benefits and limitations and the decision of which one to use will depend of the situation we need them for.

Input Structure

Data Source

The data used for this analysis is conducted with values obtained from Mercado Libre, a popular Argentine e-commerce platform. The platform 's API allows for an efficient and systematic access to a vast repository of real estate information, allowing us to obtain a snapshot of what the property availability at a specific moment in time looks like. We complemented the information retrieved by the API by accessing each publication and gathering additional information we considered relevant for the analysis and the training of the models.

Data Variables

Throughout this analysis we will be working 52.092 properties which provide a snapshot of the listed properties the 26th of May of 2023 for which we have the following variables:

- 'Item_price': Price of the property.
- 'Item_url': Url of the property.
- 'Location_Lat': Latitude.
- 'Location_Long': Longitude.
- 'acceso_a_internet': Internet access included.
- 'admite_mascotas': Pets allowed
- 'aire_acondicionado': Includes AC Unit
- 'ambientes': Number of rooms
- 'amoblado': Includes furniture
- 'antigüedad': Property age
- 'apto_credito': Allows financing
- 'apto_profesional': Can be used for professional services
- 'area_de_cine': Building includes cinema
- 'area_de_juegos_infantiles': Building includes kids area
- 'ascensor': Has elevator
- 'balcon': Has balcony
- 'baños' Number of bathrooms
- 'business_center': Building includes business center
- 'caldera': Building includes central heating
- 'calefaccion': Building includes heating

- 'cancha_de_basquetbol': Building includes basketball court
- 'cancha_de_paddle': Building includes paddle court
- 'cancha_de_tenis': Building includes business court
- 'canchas_de_usos_multiples': Building includes sports area
- 'cantidad_de_pisos': Number of floors in the building
- 'chimenea': Apartment includes fireplace
- 'cocheras': Number of dedicated parking spaces in apartment
- 'cocina': Apartment includes kitchen
- 'comedor': Apartment includes dining area
- 'con_area_verde': Building includes private park
- 'con_cancha_de_futbol': Building includes football field
- 'con_conexion_para_lavarropas': Apartment includes washing machine pipe
- 'con_energia_solar': Apartment includes solar power connection
- 'departamentos_por_piso': Number of apartments in the building floor
- 'dependencia_de_servicio': Apartment includes room for staff
- 'desayunador': Kitchen includes breakfast area
- 'disposicion': To which direction the apartment faces
- 'dormitorio_en_suite': Master bedroom has private bathroom
- 'dormitorios': Number of bedrooms
- 'estacionamiento_para_visitantes': Building has guest-parking area
- 'estudio': Apartment has private office
- 'expensas': Monthly utilities
- 'gas_natural': Apartment has gas connection
- 'gimnasio': Building includes a gym
- 'grupo_electrogeno': Building has emergency generator
- 'heladera': Apartment includes fridge
- 'jacuzzi': Apartment includes Jacuzzi
- 'jardin': Apartment has private garden
- 'laundry': Building has private laundry room
- 'lavadero': Apartment has laundry room
- 'linea_telefonica': Apartment has phone line
- 'living': Apartment has living room area
- 'location_neighborhood': Neighborhood at which the apartment is located
- 'numero_de_piso_de_la_unidad': Floor number at which the apartment is
- 'numero_de_torre': In case of apartment complex, tower number
- 'orientacion': To which direction the apartment faces
- 'parrilla': Building includes grill area
- 'patio': Apartment has private patio
- 'pileta': Building has a pool

- 'placards': Number of closets
- 'playroom': Apartment has dedicated playroom
- 'rampa_para_silla_de_ruedas': Building has wheelchair access
- 'recepcion': Apartment has reception
- 'roof_garden': Apartment has terrace
- 'salon_de_fiestas': Apartment has multipurpose room
- 'salon_de_usos_multiples': Apartment has multipurpose room
- 'sauna': Building has sauna
- 'seguridad': Building has private security
- 'superficie_cubierta': Number of covered square meters in the apartment
- 'superficie_de_balcon': Number of balcony square meters in the apartment
- 'superficie_total': Total square meters in the apartment
- 'terraza': Apartment has terrace
- 'tipo_de_departamento': Type of apartment
- 'tipo_de_seguridad': Type of private security
- 'toilette': Apartment has guest bathroom
- 'vestidor': Master bedroom includes private dressing room
- ‘descuento’: If the article description or title contains the word “Retasado/a” indicating the property was reappraised orthe title contains “Oportunidad” indicating an expected low value.
- ‘reciclar’: Indicates if the property description or title mentions that the apartment requires repairs

Exploratory Analysis

In order to assess the quality of the inputs, we will begin by exploring the available dataset. The purpose of this analysis is to obtain enough insights for us to be able to properly decide which features we want to include into the model as well as evaluating the quality of the data and decide any preprocessing necessary for the analysis. We will evaluate different visualization throughout this section:

Sqr Mt Price vs Location

The following figure presents a ranking of the available neighborhoods based on their price per square meter (Sqr Mt). It provides a visual representation of the relative pricing levels across various neighborhoods, allowing for a comprehensive understanding of the spatial distribution of property values (Fig. 1).

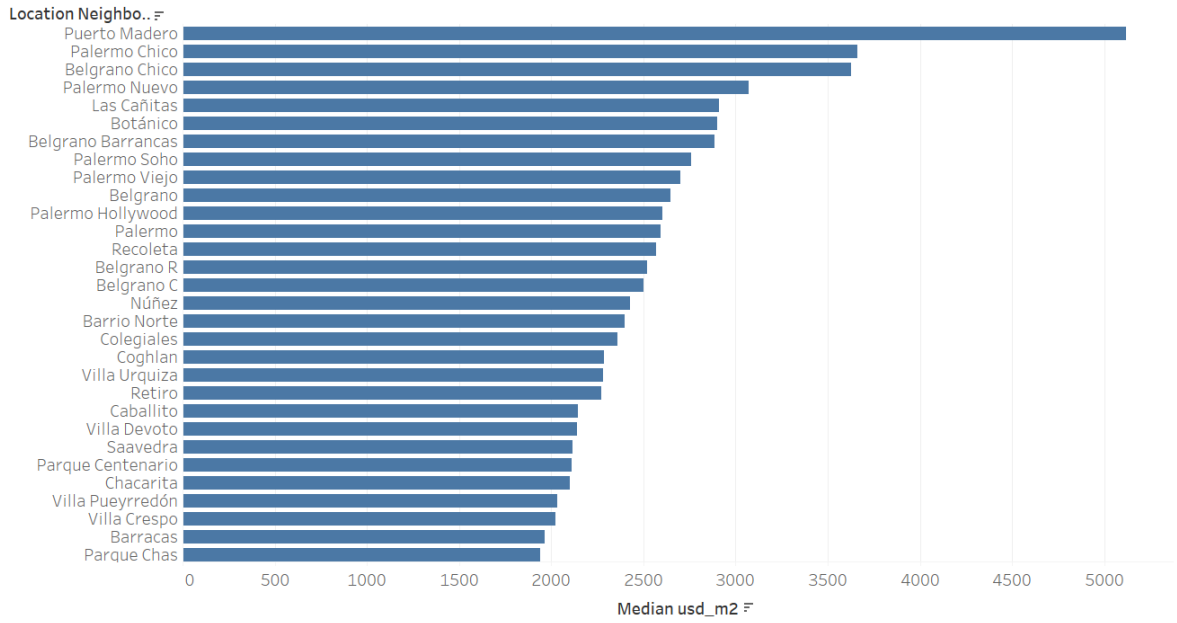


Figure 1: Median price per Sqr Mt by neighborhood

By analyzing the figure, we can observe that certain neighborhoods command higher prices per sqr mt compared to others, indicating a greater demand or perceived

desirability in those areas. We can observe Puerto Madero, Palermo Chico and Belgrano Chico being the most expensive neighborhoods.

Price vs Sqr Mts

The following scatter plot (Fig. 2) illustrates the relationship between property size and price in the real estate market. As property size increases, there is an upward trend in prices, indicating that larger properties tend to command higher prices.

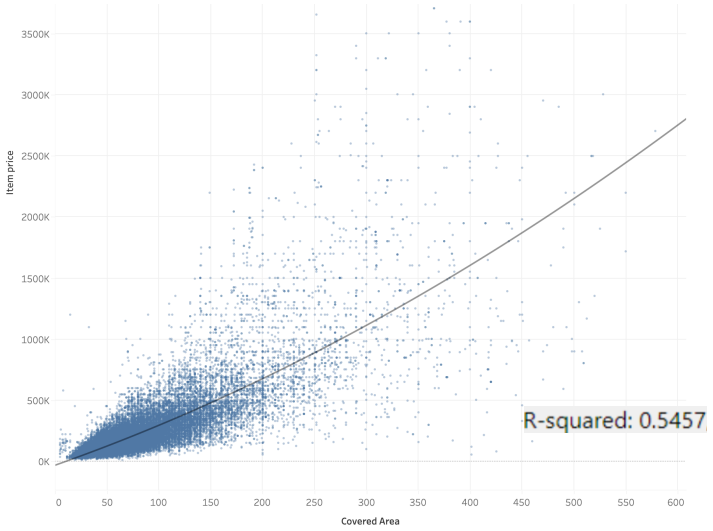


Figure 2: Covered Area Vs Price Scatter

Price/Sqr Mt vs Sqr Mt

The following scatter plot (Fig. 3) illustrates a less clear relationship between property size and price per square meter. As property size increases, there is an upward trend in price per square meter. This could indicate that either bigger apartments are better valued by the market or that they tend to be built in more expensive neighborhoods.

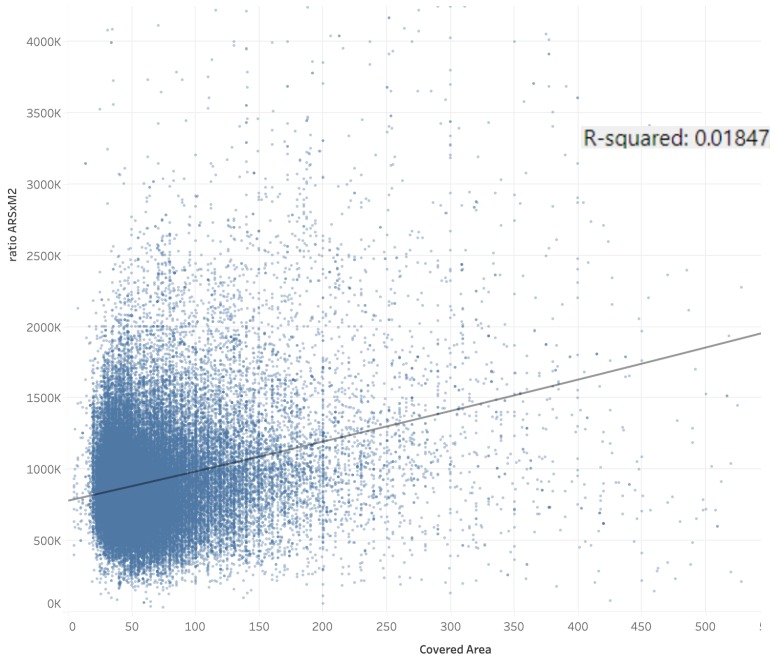


Figure 3: Covered Area vs Price per Sqr Mt

Number of Available Properties by Neighborhood

The following bar chart (Fig. 4) indicates the number of available properties for each of the neighborhoods within this dataset. Belgrano, Caballito and Palermo appear to be the most popular neighborhoods within Mercado Libre. Here we can see a potential issue with how we segment the data since neighborhoods like Palermo and Belgrano are actually subdivided into smaller neighborhoods while for many observations they are classified manually.

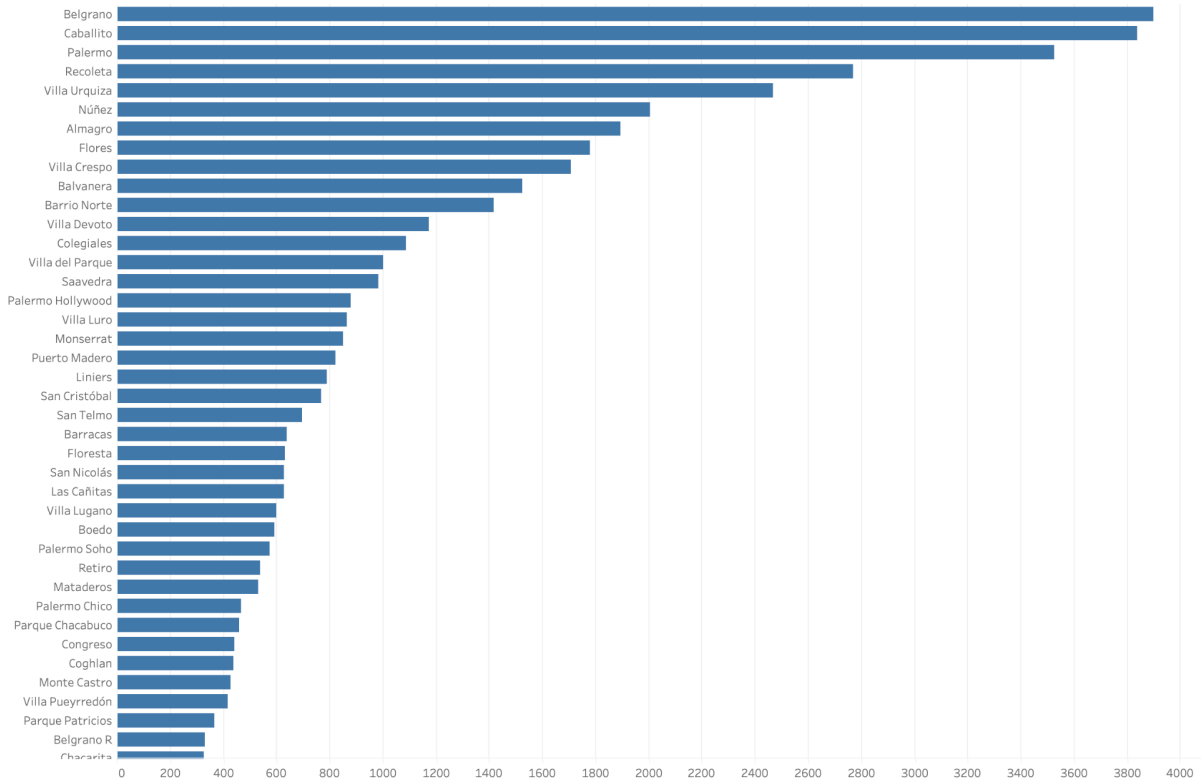


Figure 4: Number of Available Properties by Neighborhood

Geospatial Visualizations

Taking advantage of geospatial visualization tools, we opted to perform more advanced visualisations allowing us to better understand the distribution of the data, attempting to discover potential clusters. We will take a look at how different variables are distributed along the city including:

- Price
- Price/Sqr Mt
- Size
- Apartment Age

Price Map

The following visualisation (Fig. 5) displays a heatmap of real estate prices in Buenos Aires. We can observe that prices appear to be quite uniform all across the city excluding the northern corridor of it. The closer the property is to the river, the higher the price tag it seems to have.

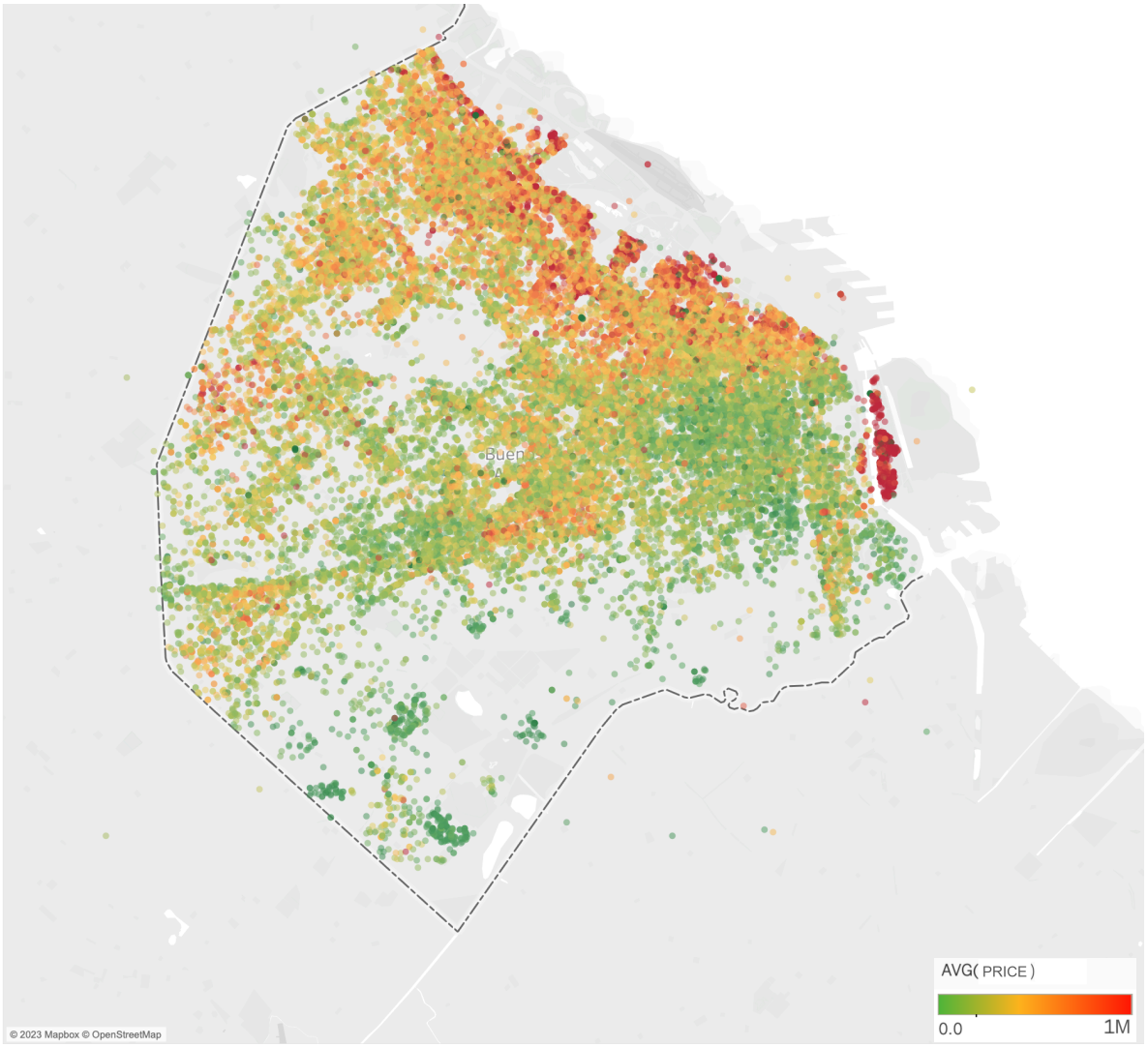


Figure 5: Total Price Heatmap in Buenos Aires

Price/Sqr MtMap

The following visualisation (Fig. 6) displays a heatmap of real estate prices per square meter in Buenos Aires. The uniformity exposed in the previous visual is no longer present exposing a more detailed clustering of regions. The northern corridor of the city prevails as the most expensive part of the city with Puerto Madero being clearly defined as the most expensive part of it.

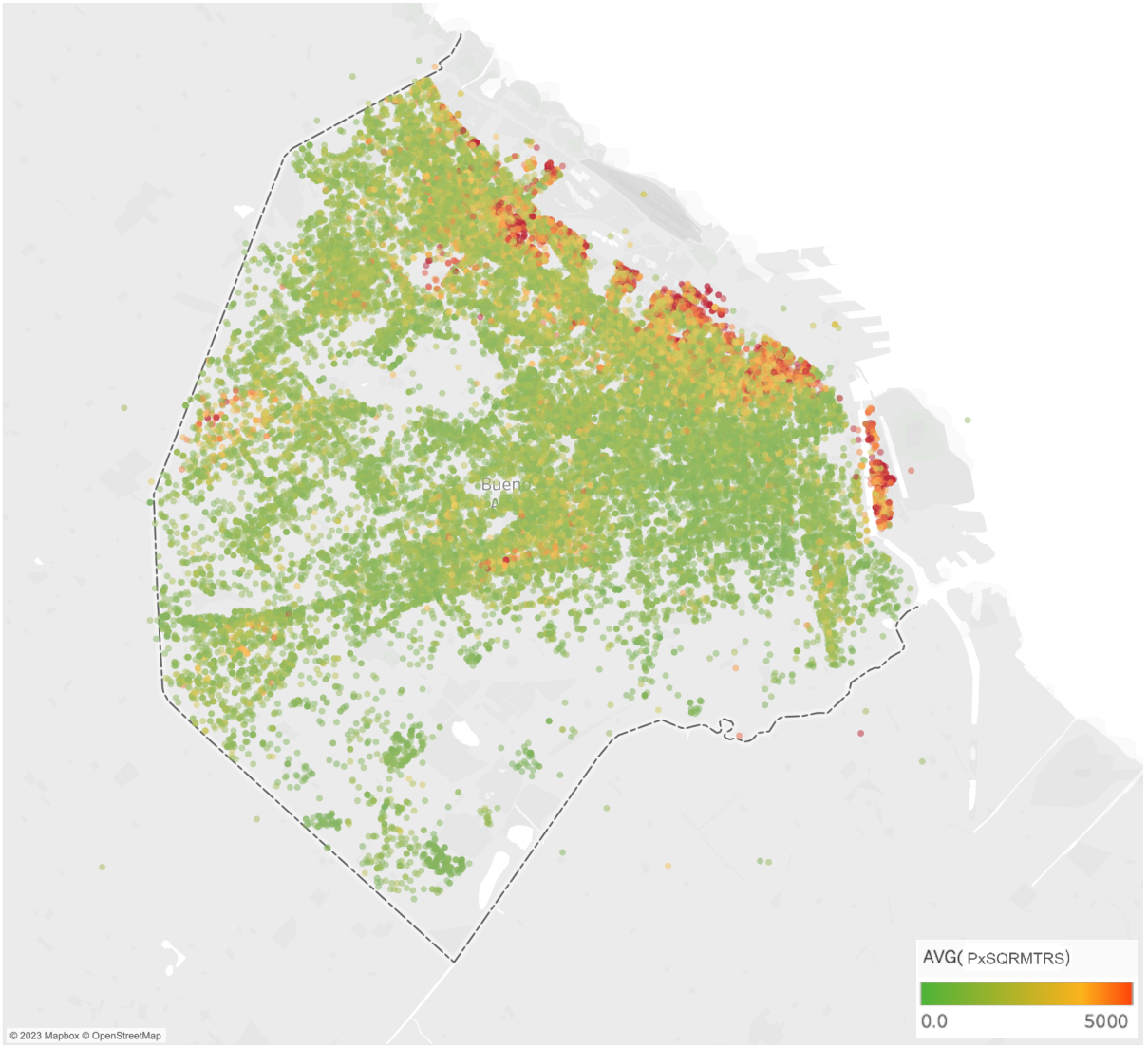


Figure 6: Price per Squaremeter Heatmap in Buenos Aires

Size Map

The following visualisation (Fig. 7) displays a heatmap of property sizes per square meter in Buenos Aires. This visualisation confirms our past findings regarding bigger apartments being built in the most expensive parts of the city, thus explaining the positive correlation between apartment size and price per square meter. We suspected that it could have been related to older apartments being bigger and while there is some correlation between those variables, neighborhoods of the city like Puerto Madero and Bajo Belgrano indicate otherwise.

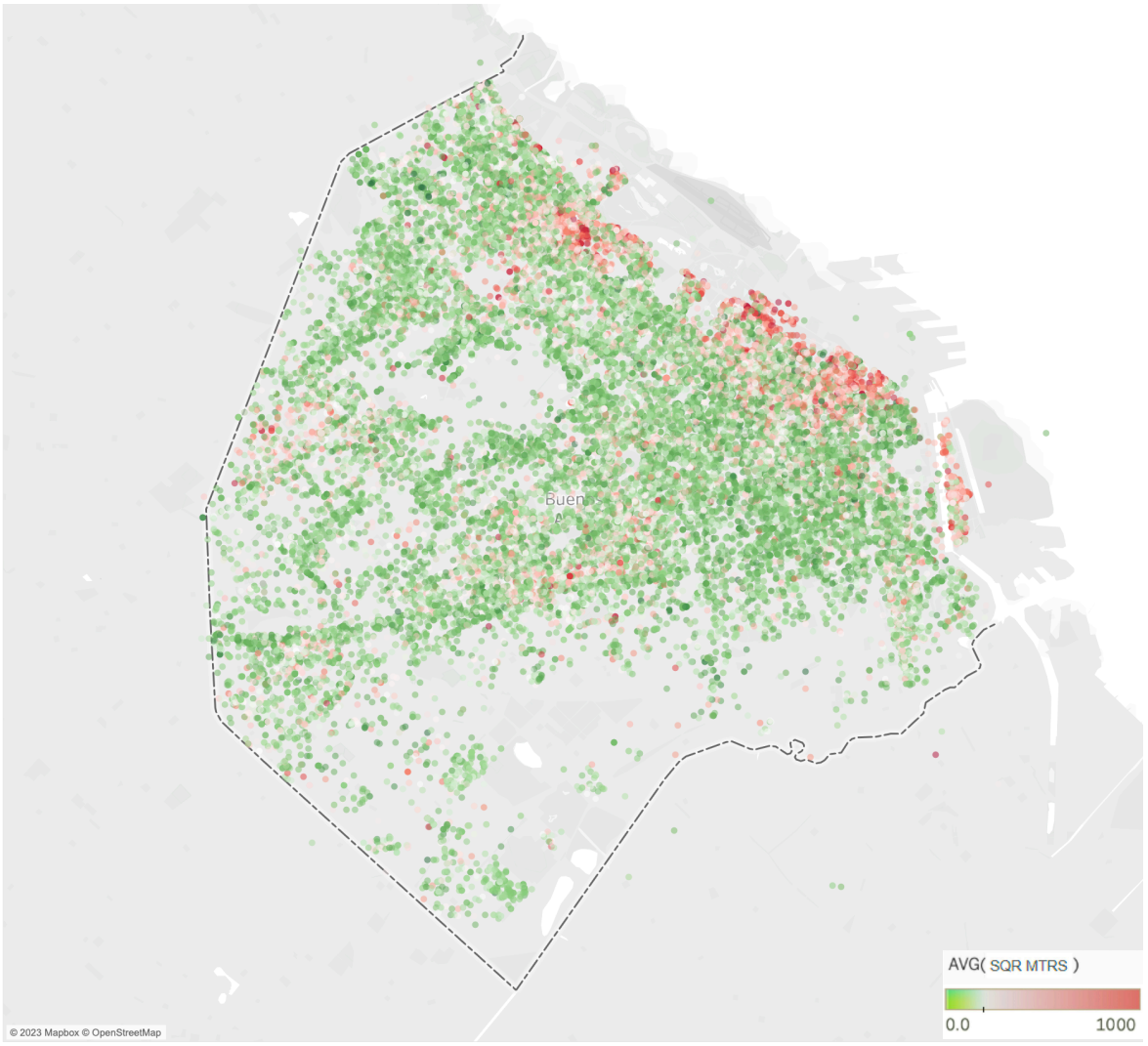


Figure 7: Property Size Heatmap in Buenos Aires

Property Age Map

The following visualisation (Fig. 8) displays a heatmap of property age in Buenos Aires, with. This visualisation gives us many insights into the way the city of Buenos Aires was developed. We can observe how most of the early developments within the city were surrounding the “Puerto de Buenos Aires”, a key part of the city's history. When taking into account that the data also accounts for apartment buildings we can also observe how regions like the west part of the city, which used to be mostly populated by houses, are a popular place for new developments since they mostly show new apartments.

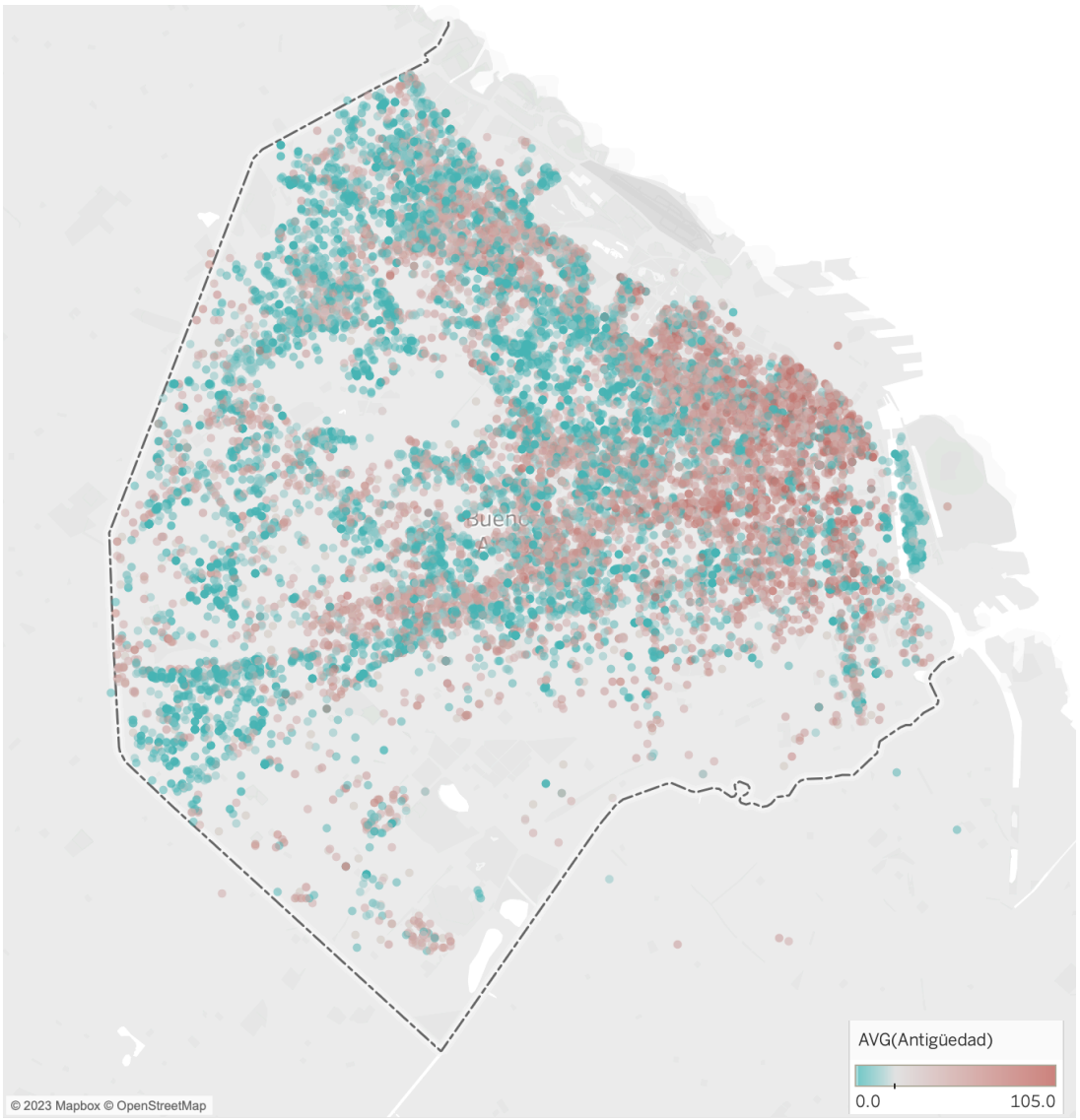


Figure 8: Property Age Heatmap in Buenos Aires

Neighborhoods Distribution

The following visualisation (Fig. 9) displays the clustering by neighborhoods. It allows us to validate a proper distribution of the data as well as the reliability of it. We can observe many observations overlapping in the region allocated to Recoleta and Barrio Norte. We can also observe the overlap of the different Palermo subcategories with the main one all across it.

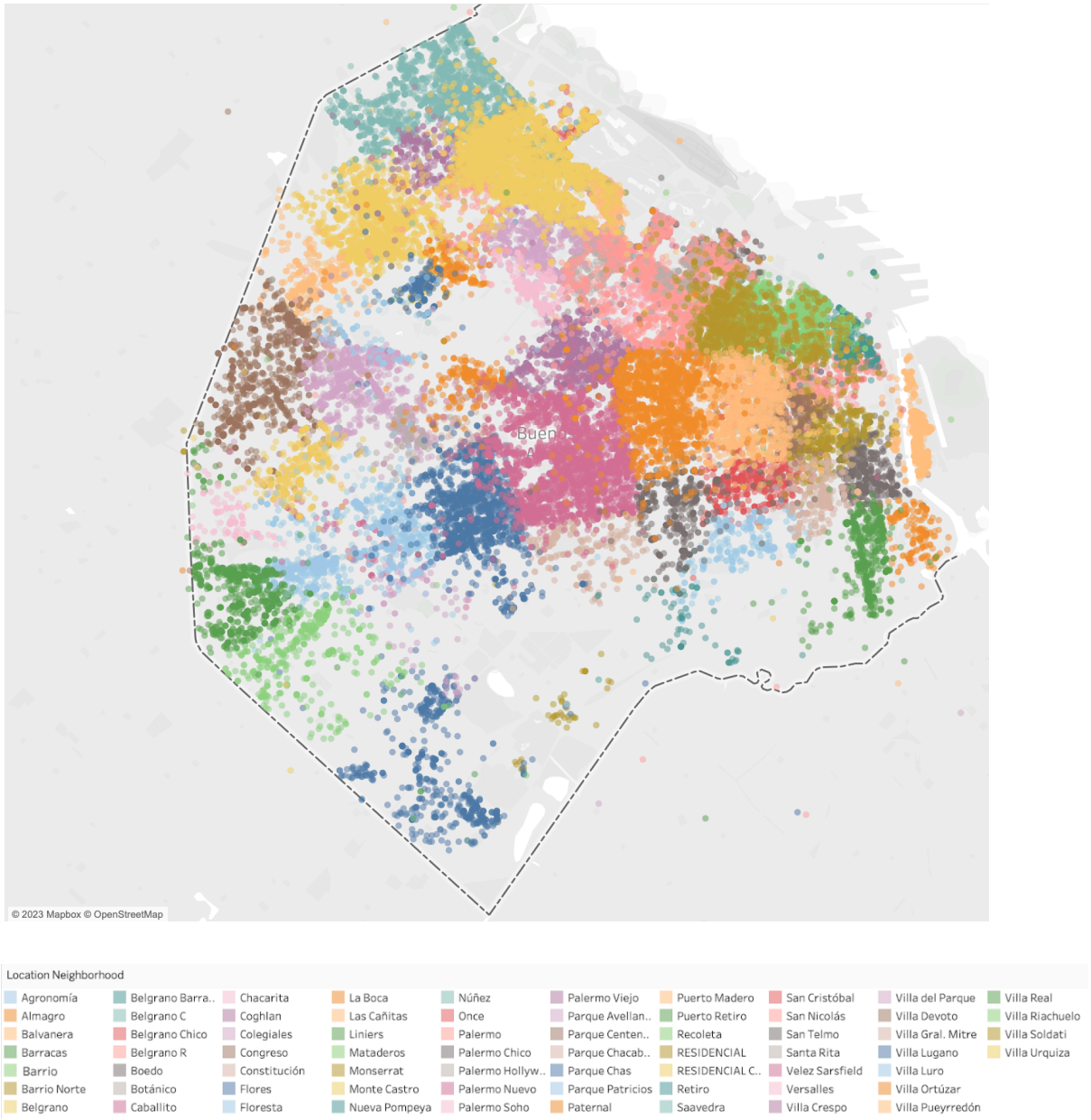


Figure 9: Neighborhood Distribution in DataSet

Data Pre-Processing

In this section we will describe what criteria for data cleaning was taken into account, following up on issues discovered in the previous section. A total of 12.353 out of 64.438 listings were removed based on the following criteria

Records missing key information:

We opted for the removal of records which were missing key information. This includes:

- Listings without a price
- Listings without an address
- Listings no feature table

If a categorical variable of the apartment contained an empty value it was assumed that the property did not contain that variable. E.g. if a pool was not specified, we would default it to False.

Mislabeled Removals

There were some records which had evidently mislabeled data and needed to be removed. This includes:

- Properties mentioning rent in their title
    - Removed by looking for the term “Alquiler” and variations in the title and looking at listings under 5000usd
- Listings actually targeted to buying properties
    - Identified by searching for the term “compra” and variations in the title
- Listings for which its price corresponds to only the first payment for the property
    - Identified by manually running through listings with unusual values smaller than 20.000 usd

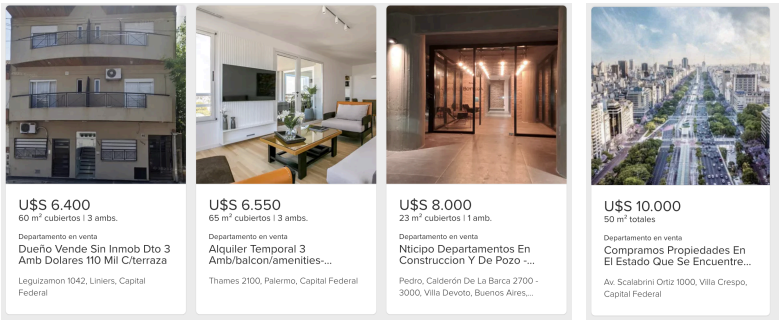


Figure 10: Mislabeled data examples

Combination of Variables

After analyzing the available data we uncovered some variables with very close meanings that could be summarized into one. The following aggregations were performed:

- 'terraza' + 'roof garden' were aggregated into 'terraza' by taking the maximum value between both
- 'salon_de_usos_multiples' + 'salon_de_fiestas' were aggregated into 'salon_de_usos_multiples' by taking the maximum value between both

Neighborhood ambiguity corrections with the use of polygons:

When observing the geospatial visualisations we noticed that many observations had the neighborhood misplaced based on the observations. After further examination we uncovered that this was due to misaligned criteria of to which neighborhood a specific set of coordinates corresponded to. Most cases of this issue can be found in Palermo, Belgrano and the division between Barrio Norte and Recoleta. We opted to unify the criteria and, by using custom polygons for each neighborhood and geospatial analytics techniques, we were able to reclassify the data points into the correct neighborhood.

In order to do this, we performed the following steps:

1. Analyzed geospatial visualizations to spot misclassified data points.
2. Reviewed existing neighborhood classification criteria.

3. Defined accurate neighborhood boundaries using custom polygons. We identified most issues were related to the neighborhoods of Belgrano, Palermo and Recoleta.
4. Used spatial join operations to reassign properties based on coordinates and custom polygons.
5. Updated dataset with correct neighborhood classifications.

We will now present the visual examples of the previously mentioned cases:

Belgrano Subdivision Correction

Before and After (Fig. 11):

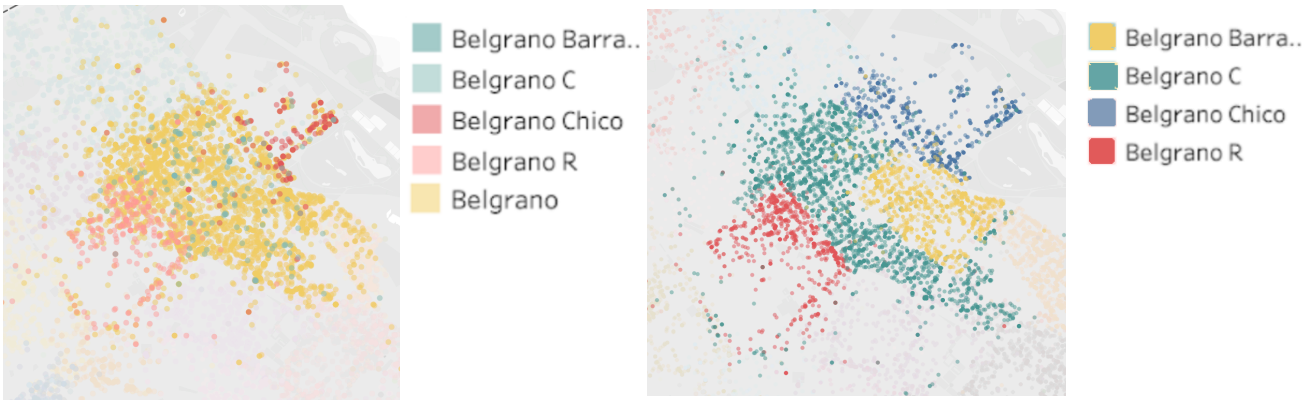


Figure 11.1 (Left): Sub-Distribution within the neighborhood of Belgrano before standardizing the geographical scope of each

Figure 11.2 (Right): Sub-Distribution within the neighborhood of Belgrano after standardizing the geographical scope of each

Palermo Subdivision Correction

Before and After (Fig. 12):

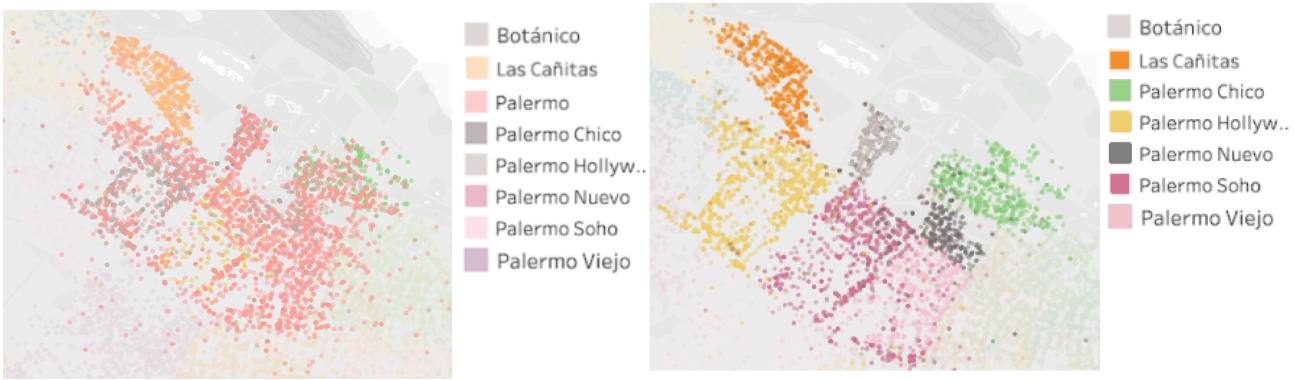


Figure 12.1: (Left) Sub-Distribution within the neighborhood of Palermo before standardizing the geographical scope of each

Figure 12.2 (Right) : Sub-Distribution within the neighborhood of Palermo after standardizing the geographical scope of each

Recoleta and Barrio Norte Overlap Correction

Before and After (Fig. 13):

Figure 13.1: (Left) Sub-Distribution within the neighborhoods of Recoleta and Barrio Norte before standardizing the geographical scope of each

Figure 13.2 (Right) : Sub-Distribution within the neighborhoods of Recoleta and Barrio Norte after standardizing the geographical scope of each

Methodology

Problem approach

Since we are attempting to narrow the scope of properties a home buyer may want to look at, we need to evaluate a way of presenting which are the ones that have the best value for money while taking into account the buyer's preferences (Eg: Apartment size, neighborhood, rooms, etc). In order to consider the investment potential of an apartment we will attempt to make predictions about its future value, for which we will utilize the concept of prediction error.

By making predictions on the expected price of a property based on its characteristics, we can infer how under/over-priced it is, allowing us to present the results in the form of an interactive map which should allow the user to filter the properties that fit the criteria he wants as well as highlighting which are the ones that are the best investment opportunity according to the model.

Model Performance

In the realm of real estate valuation, accurate assessment and evaluation of predictive models are vital for informed decision-making and reliable predictions. Evaluation metrics play an important role in quantifying the performance and effectiveness of these models. In this section, we delve into various evaluation metrics used to assess the predictive accuracy, precision, and generalization capability of real estate valuation models. By employing appropriate evaluation metrics, researchers and practitioners can gain valuable insights into the strengths and limitations of their models and make more informed decisions in the dynamic real estate market.

Feature Importance Analysis

In real estate valuation, it is often important to understand the contribution of different features or variables in predicting property values. Feature importance analysis methods, such as permutation importance or SHAP values, can be applied to assess the relative importance of different features and identify the key drivers of property value predictions. We can observe this implementation on random forest algorithms in some of the past publications [62][63][64]. This analysis can also be performed in other models such as XG Boost by taking into account how

Evaluation Metrics

There are several commonly used evaluation metrics that can be used to assess the performance of a predictive model. We will give a quick overview of the most common ones used when solving a regression model:

Mean Absolute Error (MAE)

MAE takes the average absolute difference between the predicted and actual values. It provides a measure of the average size of the errors without considering their direction. Lower values indicate better performance.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{x}_i - \boldsymbol{x}|$$

Mean Squared Error (MSE)

MSE calculates the average of the squared differences between the predicted and actual values. It amplifies the impact of larger errors compared to MAE. Lower values indicate better performance.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Root Mean Squared Error (RMSE)

RMSE is the square root of MSE and provides a measure of the standard deviation of the errors. It is in the same unit as the dependent variable and is widely used because it is easily interpretable.

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

R-squared (R^2)

R^2 measures the proportion of the variance in the dependent variable that can be explained by the independent variables in the model. It ranges from 0 to 1, with 1 indicating a perfect fit. Higher values of R^2 indicate better performance.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Mean Absolute Percentage Error (MAPE)

MAPE calculates the average percentage difference between the predicted and actual values. It is commonly used in forecasting problems to assess the relative accuracy of the predictions.

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Holdout Validation

Holdout validation, also known as train-test split, is a commonly used validation methodology. In this approach, the dataset is divided into two subsets: a training set and a test set. The model is trained on the training set and evaluated on the test set, which contains unseen data. The performance metrics, such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE), are calculated on the test set to assess the model's predictive accuracy. Holdout validation provides a straightforward assessment of the model's performance on independent data [69][70][71][72].

Cross-Validation

Cross-validation is a robust validation technique that addresses the limitations of holdout validation by using multiple splits of the data. The most common cross-validation method is k-fold cross-validation. In k-fold cross-validation, the dataset is divided into k equally sized folds. The model is trained and evaluated k times, each time using a different fold as the validation set and the remaining folds as the training set. The performance metrics are then averaged over the k iterations. Cross-validation provides a more reliable estimate of the model's performance by utilizing the entire dataset for both training and validation [72][73][74][75].

Machine learning algorithms

The algorithms we evaluated in this experiment were: XG Boost, Random Forest and Support Vector Machine. In order to decide which one to keep, we evaluated each of them based on the R Squared calculated for a specific subset of the data. For each of the model's evaluations a random grid of hyperparameters with 100 different combinations were used. The subset of data used for the experiment consisted of 10.000 randomly chosen observations for training, 1000 randomly chosen observations for testing and 1000 randomly chosen observations for evaluation.

Parameter Tuning

Before training our final model, we performed a more extensive random grid search for the XgBoost algorithm using a bigger subset of the data (50%). The resulting hyperparameters were:

- Number of Estimators: 511
- ETA (Learning Rate): 0.0984
- Max Depth: 7
- Sub Sampling: 0.966

Model training and evaluation

Given our available data structure we needed a way of getting a prediction for each of the properties by using a model trained independently of each property listing for data leaking not to be an issue.

In order to achieve this, we opted to obtain the predictions for each property with a 20 fold training structure. The criteria behind 20 aims to keep 5% subsets of the data out in each fold. This consisted in the following steps:

1. We randomly split the data into 20 different groups based on their unique IDs. This ensured that each group was independent and prevented any overlap or information leakage between them.
2. For each of the group of IDs we:
    a. Performed a test/train split where the subset of IDs in that group was used as the test set, and the remaining IDs were used for training.
    b. Trained an XG Boost algorithm using the hyperparameters defined in the previous point .
    c. Calculated the predicted values for the subset of IDs left out of the training set in the iteration
    d. Save the results for the calculated IDs.
3. We appended all the results together, consolidating the predictions for each property across all 20 iterations

By implementing this methodology, we ensured that our predictions were based on a model trained independently for each registration, therefore avoiding data leakage and outputting reliable results for each available property.

Results and Analysis: Opportunity Detection

Model Selection Based on Performance

In this document we chose to focus on SVM, XGBoost, and Random Forest due to their suitability for handling complex datasets. Decision trees were deemed too simple for our needs. While neural networks are powerful and could be implemented for future iterations that might include image data, they require extensive set up and potentially bigger computational resources. For now, for practicality reasons, we will focus on just these 3 models.

The results were the following:

Model	R Squared	Mean Absolute Percentage Error
XGBoost	0.887	27.95%
Random Forest	0.861	28.37%
SVM	0.7964	32.01%

Table 1: Models tested R Squared and MAPE evaluation

Based on the evaluation of the results [Table 1], the XGBoost algorithm emerged as the most favorable choice among the models considered. With an R-squared value of 0.887, it outperformed both the Random Forest (0.861) and SVM (0.7964) models. The higher R-squared value indicates that the XGBoost algorithm was able to explain a larger proportion of the variance in the data, suggesting a better fit to the underlying patterns and relationships. The conclusions seem to be also reflected when looking at the MAPE, with XG Boost obtaining the lowest error of 32.95%. Therefore, the decision to opt for XGBoost as the preferred algorithm is justified by its superior performance in accurately predicting the target variable in the sample experiment.

In this context, R^2 was chosen over adjusted R^2 as the evaluation metric. While adjusted R^2 accounts for the number of predictors in the model and adjusts for the potential inflation of R^2 due to overfitting, it is particularly useful when comparing models with differing numbers of variables. However, in this study, all models were built using the same set of variables. Since the number of variables remains constant across all models, the primary concern of adjusted R^2—adjusting for the number of predictors—is not applicable here. The overestimation of a model performance by R^2 is acknowledged however in this context is used as a relative measure, so we are not focusing in the value of the effect.

Model performance evaluation and comparison

After obtaining the predictions for deeach property using the 20-fold training structure, we proceeded to evaluate the performance of our model. The evaluation process involved assessing various metrics to better understand how well the model performed.

The following metrics were used for evaluation:

Metric	**Score**
Mean Absolute Error	28453
Mean Squared Error	4931683168
Root Mean Squared Error	70225
R-Squared	0.8908
Mean Absolute Percentage Error	14.131936

Table 2: Metrics on final model

The R-squared value of 0.8908 indicates that our model explains approximately 89.08% of the variance in the data. This could be considered a high R-squared within this

context suggesting that our model is effective at capturing and predicting the prices based on the available information. This shows reliability in our model at explaining the variability when predicting real estate prices.

Additionally, the Mean Absolute Percentage Error (MAPE) provides insights into the accuracy of our model's predictions. With a MAPE of 14.13%, our model's average percentage difference between the predicted and actual values is relatively low.

Feature importance analysis and interpretation of results

Understanding the importance of different features in predictive modeling is crucial for gaining insights into the variables influencing the model's predictions. In order to evaluate feature importance we ranked the variables based on the improvement to the model's loss function when the feature was used for splitting, expecting for important features to have caused the biggest improvements. We found the following 15 variables to be the ones impacting the most when it comes to predicting the value of an apartment in Buenos Aires:

Variable	**Description**
Location: Puerto Madero	If the apartment is in Puerto Madero
Total Area	Total area of the apartment
Gym	If the building has a gym
Covered Area	Covered area of the apartment
Pool	If the building has a pool
Private Parking Spaces	If the apartment has a dedicated parking space

Multipurpose room	If the buliding has a multipuropose room
Private Dressing Room	If the apartment has a private dressing room
Apartment Age	Indicates the age of the apartment
Playroom	If the apartment has a dedicated playroom
Jacuzzi	If the apartment has a jacuzzi
baños	Number of bathrooms in the apartment
Suite Master Bedroom	If the master bedroom in the apartment has a private bathroom
location_neighborhood_Palermo Chico	If the apartment is located in Palermo Chico
estudio	If the apartment has a private office

Table 3: Variables with the higher impact in price

The location of an apartment stands out as one of the most important factors in determining property value. For example, apartments in Puerto Madero command higher prices, reflecting its reputation as one of the most expensive neighborhoods in Latin America. This insight aligns with the broader understanding that specific locations, like Puerto Madero, influence property values due to it being a highly desirable neighborhood.

The total area and covered area of the apartment also appear to be strong predictors, indicating that larger apartments generally command higher prices. Following a similarly intuitive logic, the age of the apartment was also identified as an important factor, with newer apartments usually carrying higher prices than older ones.

Amenities played a key role too: The presence of a gym, pool, multipurpose room, and playroom positively influencing apartment values. When analyzing the underlying meaning of this within the context, this combination of amenities usually corresponds to modern luxury apartment complexes which are becoming more and more popular within the city. It is important to note that while these features were determined to be significant predictors in our model, they should be interpreted within the specific context of our dataset. Future research and additional domain expertise can further enhance our understanding of the underlying dynamics and provide a more comprehensive feature importance analysis.

Interactive Tool Development

Understanding End Users

In order for us to design an interactive tool for end users, we first need to understand who they are. The analysis conducted in this potentially appeals to two distinct groups of individuals interested in the real estate market: institutional buyers and private/home-owners buyers.

Institutional buyers

This refers to organizations or entities that engage in large-scale real estate investments, such as real estate investment trusts (REITs), private equity firms, or pension funds. These buyers typically seek to diversify their portfolios, generate consistent returns, and capitalize on long-term investment opportunities. They may not be so interested in specific characteristics of the properties but rather on the return of the investment as a whole. It may occur that they do have a particular preference fueled by specific market research, however in that case it can be usually more favorable to train a specific model for the purpose in question.

Private Buyers

They encompass individual homebuyers. These buyers are often driven by personal motivations, depending on the final objective they have for the property. This can be (when thinking of apartments within Buenos Aires), a first apartment, an apartment for them to raise a family, an apartment for an investment (such as using it for AirBnB), among many others. They are more likely to engage in smaller-scale investments and often prioritize functional factors to their purpose such as location, amenities, and number of rooms. For example, while a buyer looking for a place to raise 3 children may be looking for apartments with an enough number of bedrooms to accommodate his family close to his school of interest, a private buyer focused on developing an AirBnb business may want something close to popular city areas with abundant gastronomy without carrying too much about if the apartment is a studio or if it has many rooms. His

focus may be more headed to what amenities the building includes him that allows him to get a better profit from each reservation

Conclusion

When contemplating the diversity of preferences within private buyers, we are able to conclude that the tool caters more to them since it makes the evaluation of different investment proposals more relatable to them and more relevant to their decision-making process.

Preferences Survey

To develop a user-centric tool, understanding buyer preferences was essential. This ensures the tool provides relevant insights and enhances its usability for prospect end users. Because of this, a survey around consumer preferences was included.

The primary aim of this survey was to collect data on attribute preferences about attributes of properties. The survey consisted of a grid where participants were asked to indicate the importance they assigned to each of the following attributes:

- Location
- Covered Area
- Price
- Rooms
- Amenities
- Uncovered Area
- Parking
- Light
- Security
- Age

To ensure that the survey participants were actual home buyers, we selected individuals who were either personally known to us or referred by acquaintances. This approach helped verify the authenticity of their past purchase. The participants included: people who bought their first property, people who purchased a property for their children, and people who purchased a property for them and their family.

Participants were requested to indicate the level of importance they gave to each characteristic by selecting one of the provided options: 1 indicating low importance, 2 representing moderate importance, and 3 indicating high importance. [101]

By assessing the perceived importance of these specific attributes, the survey aimed to uncover which are the attributes we should focus on when designing the interactive features of the tool. The survey was distributed to a sample of 24 past -property buyers within the last 10 years. Efforts were made to ensure a representative sample that would provide a good overview of the target population by including buyers with different objectives in mind.

The results from the survey were the following:

Variable	Survey Sum	Survey Average	Survey StdDev
Location	71	2.96	0.20
Price	70	2.92	0.28
Covered Area	69	2.88	0.45
Light	66	2.75	0.53
Parking	47	1.96	0.69
Age	42	1.75	0.53

Rooms	40	1.67	0.64
Security	40	1.67	0.70
Amenities	38	1.58	0.58
Uncovered Area	34	1.42	0.65

Table 4: Survey results

The survey results [Table 4] provided us with a hint on what home buyers are looking at when evaluating investment opportunities. Location was the most significant variable, with an average rating of 2.96, indicating that the surveyed individuals highly valued the geographical context of a property. Price was also ranked relatively high, with an average of 2.92, supporting our past assumption that affordability took an important role within their hierarchy. Additionally, the standard deviation for Parking was the highest at 0.69, suggesting considerable variability in its importance among buyers. We will make use of these findings to determine the tool's objectives as well as the layout design.

Tool Design Principles

In order for this tool to be successful, a clearly defined guideline of design principles should be followed. Within the context we are working with, we believe that these are the guidelines we should follow to be able to make the most out of this tool:

User-Centric Design

Prioritize the needs and goals of the users throughout the design process. Understand their end goal to create an intuitive user experience. Make use of the survey data to contemplate their preferences into the tool.

Simplicity and Clarity

Keep the design clean, uncluttered, and easy to understand. Use a minimalist approach when presenting information and avoid overwhelming the user with too many elements. Establish clear labels and organize the interface elements to guide users through the tool effortlessly.

Visual Hierarchy

Define a clear visual hierarchy to guide users' attention and highlight important information. Use visual cues such as color, contrast, and typography to differentiate between different elements and emphasize key data points or indicators of investment potential

Filtering and Sorting

Enable users to filter and sort the real estate opportunities based on their preferences and investment criteria. Provide intuitive controls or dropdown menus to allow users to refine the displayed data according to factors such as price, location, size, amenities, or historical trends

Facilitate Comparative Analysis

Users should be able to compare different properties and assess their investment potential side by side. The tool should provide features that allow users to overlay data, view multiple properties simultaneously, and make informed comparisons based on key indicators.

Layout Design

Taking everything so far into account, we concluded that the main visual element should be centered around a map. Property location is one of the most relative aspects of the valuation when it comes to private buyers, since it can take many different scopes. A map should provide a flexible user experience, even in specific cases that require proximity to specific buildings or areas such as offices, restaurants, schools, and other amenities.

To establish a clear information hierarchy, the main display should prominently feature the map. The colors on the map will indicate the quality of investment, with red representing lower quality and green representing higher quality. This visual representation should allow users to quickly assess the investment potential of all the available offers within their scope .

In terms of user experience, it's important to prioritize user friendliness. One way to achieve this is by ensuring that all filters and settings are clearly defined in one control box. This will make it easier for users to navigate and customize their search criteria. Additionally, tooltips should be implemented to provide detailed information about each property when users hover over specific points on the map. Another consideration for UX is the use of proper titles. Clear and descriptive titles should be used throughout the layout to help users understand the purpose and functionality of different elements. This should enhance the overall usability of the interface.

The final layout design should revolve around a filterable colored map that visually represents the investment quality. The map should take precedence in the main display, with colors indicating the investment potential. User friendliness should be a key aspect of the design through clear definitions of filters and settings, along with tooltips for detailed property information. Proper titles should be used to enhance overall usability and comprehension.

Final Result

After consideration of all of the aspects previously discussed, we have implemented the following tool (Figure 14) . The centerpiece of the design is a filterable colored map, serving as the main visual element, which allows users to gauge all of the investment opportunities available for its requirements. With the scale ranging from red to green, the colors on the map vividly represent the potential quality of the investment. The information hierarchy has been designed around the map, recognizing the significance of location for users. Additionally, we implemented a consolidated control panel with the objective of defining individual preferences easily.

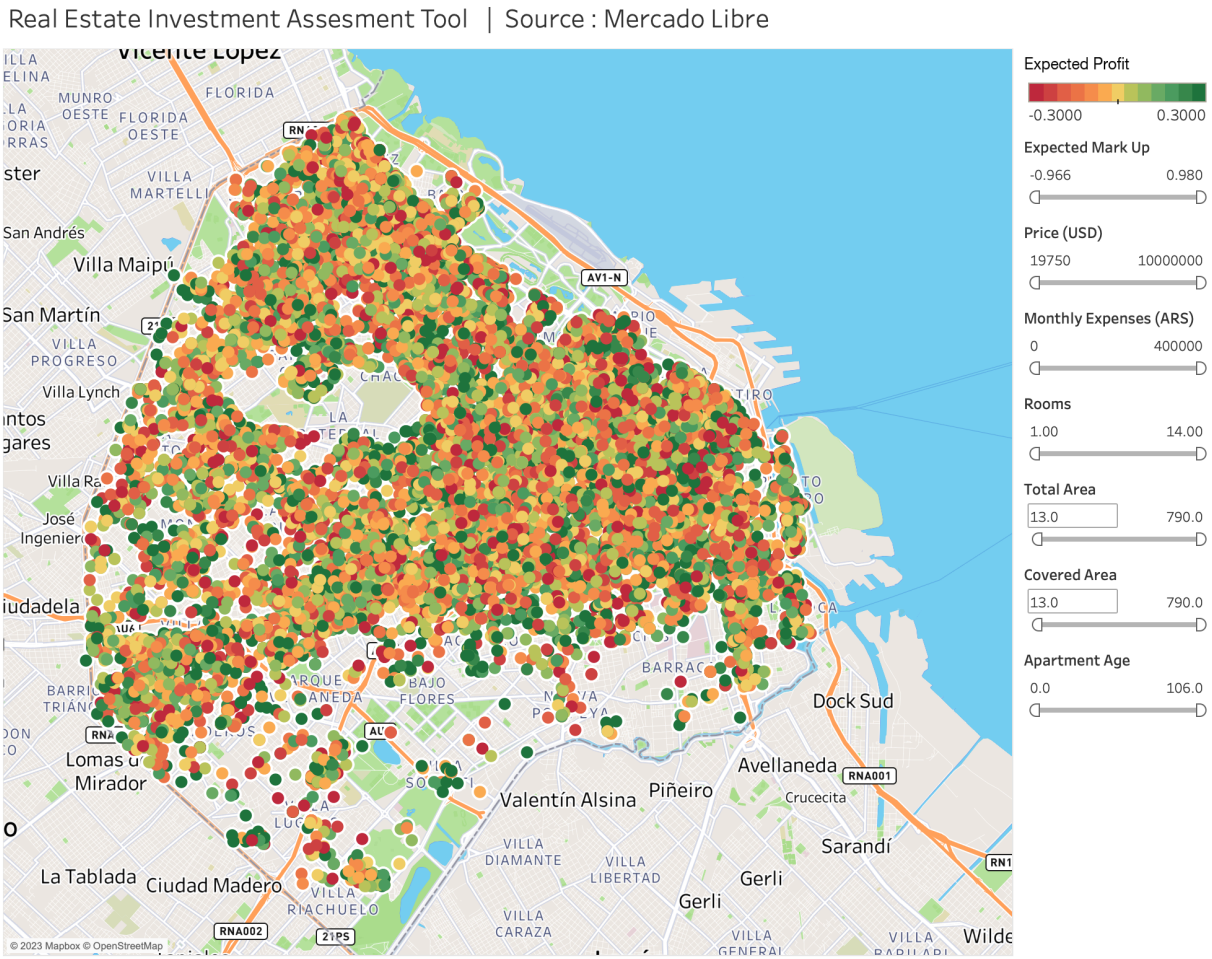


Figure 14: Real Estate Investment Assessment Tool

To provide comprehensive property details, tooltips have been incorporated, enriching the user experience. Users should be able to easily navigate and comprehend the functionalities of each element.

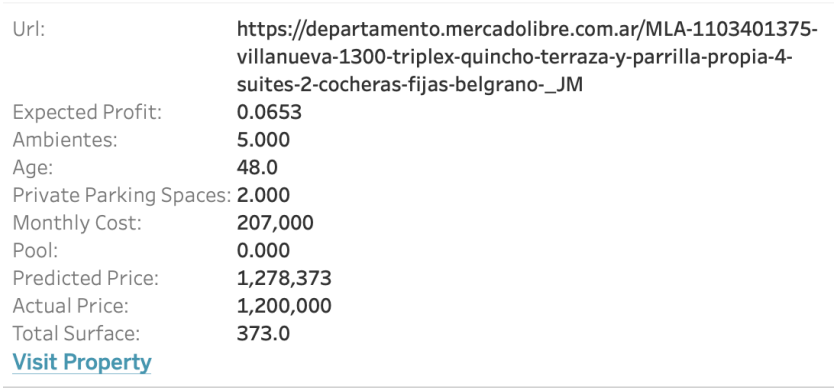


Figure 15: Real Estate Investment Assessment Tool, Tooltips

The provided "Expected Profit" in the tool corresponds to the difference between the predicted price and the actual price, divided by the actual price. This metric is calculated as follows:

$$\text{Expected Profit} = (\text{Predicted Price} - \text{Actual Price}) / \text{Actual Price}$$

This calculation provides a percentage that indicates the potential profit or loss relative to the actual price, helping users assess the investment opportunity in a standardized manner.

The tool could be tailored to specific user preferences such as adding a neighborhood filter to narrow searches or incorporating a table listing the best apartments by investment score based on applied filters. Different features could potentially streamline the decision-making process of specific users and needs.

The tool should be able to accommodate different requirements depending on the buyer's needs. We will explore 3 examples, providing 3 recommended properties and 3 that are not for each basing us on the tool's results for the following scenarios:

- High Profile Buyer

- AirBnB Apartment Buyer
- Family Focused Buyer

Scenario Simulations

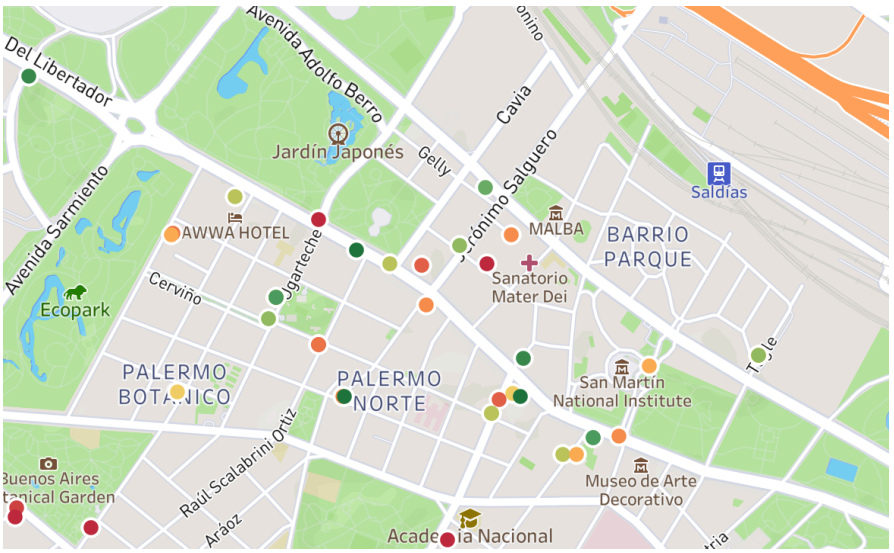
Among all tested scenarios we found that the tool recommends bigger, newer and with more services apartments when compared to the discouraged options by it.

Scenario Simulation 1: High Profile Buyer

Profile Description: Buyer looking for a luxury apartment within Palermo Chico.

This individual can start by selecting their preferred area on the interactive map, such as the area of Libertador Avenue, known for its luxurious properties. It can then specify their criteria, such as minimum size, desired amenities, and maximum price. The tool . The model will present the user a curated list of properties, highlighting those that not only meet the specified criteria but also stand out due to the features it has for the price. This enables the high-profile buyer to make informed decisions quickly, focusing on properties that promise the highest returns while fitting his criteria.

Expected Price: 800k - 1.000k



Model Recommendations:

Title:	Venta Premium Torre De Categoría Palermo En Av Libertador.
Expected Profit:	0.257
Ambientes:	5.000
Age:	28.0
Private Parking Spaces:	3.000
Monthly Cost:	62,000
Pool:	0.000
Predicted Price:	1,131,537
Actual Price:	900,000
Total Surface:	245.0

Visit Property

Title:	Av Libertador 2100 Mario Roberto Alvarez 250 M2
Expected Profit:	0.174
Ambientes:	4.000
Age:	0.0
Private Parking Spaces:	2.000
Monthly Cost:	80,000
Pool:	0.000
Predicted Price:	939,129
Actual Price:	800,000
Total Surface:	250.0

Visit Property

Title:	Libertador Y Salguero Con Balcones En Esquina. Alto. Vista Verde. Luz Y Sol . 6 Ambientes.
Expected Profit:	0.338
Ambientes:	6.000
Age:	31.0
Private Parking Spaces:	2.000
Monthly Cost:	120,000
Pool:	0.000
Predicted Price:	1,136,895
Actual Price:	850,000
Total Surface:	271.0

Visit Property

Model Discouragements:

Title:	Palermo Chico Torre Clorindo Testa Reciclado Integramente A Nuevo
Expected Profit:	-0.439
Ambientes:	3.000
Age:	30.0
Private Parking Spaces:	0.000
Monthly Cost:	70,000
Pool:	0.000
Predicted Price:	471,279
Actual Price:	840,000
Total Surface:	150.0

Visit Property

Title:	Departamento - Venta 4 Ambientes Palermo Av Libertador
Expected Profit:	-0.405
Ambientes:	4.000
Age:	45.0
Private Parking Spaces:	1.000
Monthly Cost:	70,000
Pool:	0.000
Predicted Price:	529,238
Actual Price:	890,000
Total Surface:	160.0

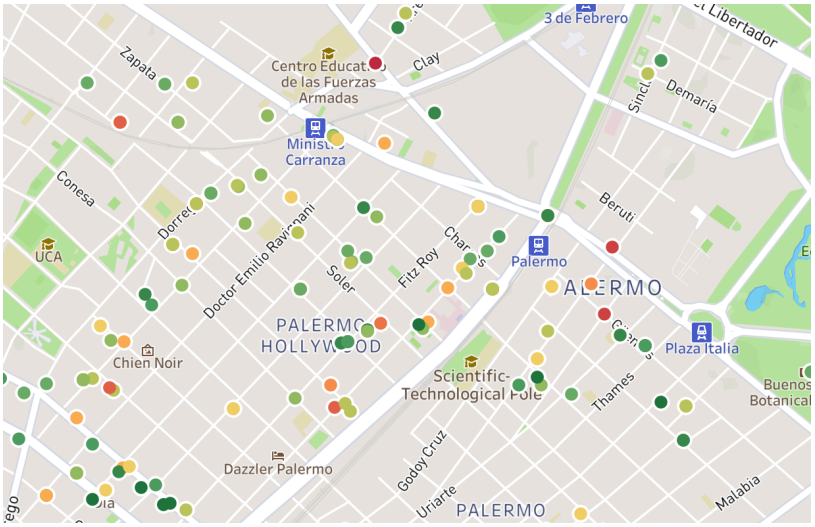
Visit Property

Scenario Simulation 2: AirBnB Apartment Buyer

Profile Description: Buyer looking for a studio apartment within Palermo Hollywood to use as an AirBnB.

This individual begins by pinpointing areas known for tourist appeal using the interactive map, such as Palermo Hollywood, known for its nightlife and cultural scene. They can then tailor their search by setting specific parameters that align with their Airbnb strategy. For instance, they might filter for studio or one-bedroom apartments, limiting the property age to no more than 10 years to ensure modern amenities and less need for maintenance.

Expected Price: 80k - 120k



Model Recommendations:

Title:	Monoambiente Con Cochera Palermo Hollywood
Expected Profit:	0.241
Rooms:	1.000
Age:	6.00
Private Parking Spaces:	1.000
Monthly Cost:	16,200
Pool:	0.000
Predicted Price:	122,829
Actual Price:	99,000
Total Surface:	42.00

Visit Property

Title:	Mood Niceto Departamento En Venta 2 Ambientes En Palermo
Expected Profit:	0.280
Rooms:	1.000
Age:	0.00
Private Parking Spaces:	0.000
Monthly Cost:	0
Pool:	0.000
Predicted Price:	117,801
Actual Price:	92,000
Total Surface:	46.00

Visit Property

Title:	1 Ambiente Contra Frente - Con Balcón - Con Cochera.-
Expected Profit:	0.274
Rooms:	1.000
Age:	10.00
Private Parking Spaces:	1.000
Monthly Cost:	8,800
Pool:	0.000
Predicted Price:	118,462
Actual Price:	93,000
Total Surface:	30.83

Visit Property

Model Discouragements:

Title:	Monoambiente En Venta Palermo Hollywood
Expected Profit:	-0.199
Rooms:	1.000
Age:	10.00
Private Parking Spaces:	0.000
Monthly Cost:	18,200
Pool:	
Predicted Price:	66,446
Actual Price:	83,000
Total Surface:	29.00

Visit Property

Title:	Departamento Monoambiente En Venta - Vivienda - Estudio Profesional - Oficina - Palermo Hollywood
Expected Profit:	-0.182
Rooms:	1.000
Age:	0.00
Private Parking Spaces:	0.000
Monthly Cost:	0
Pool:	0.000
Predicted Price:	81,840
Actual Price:	100,000
Total Surface:	33.00

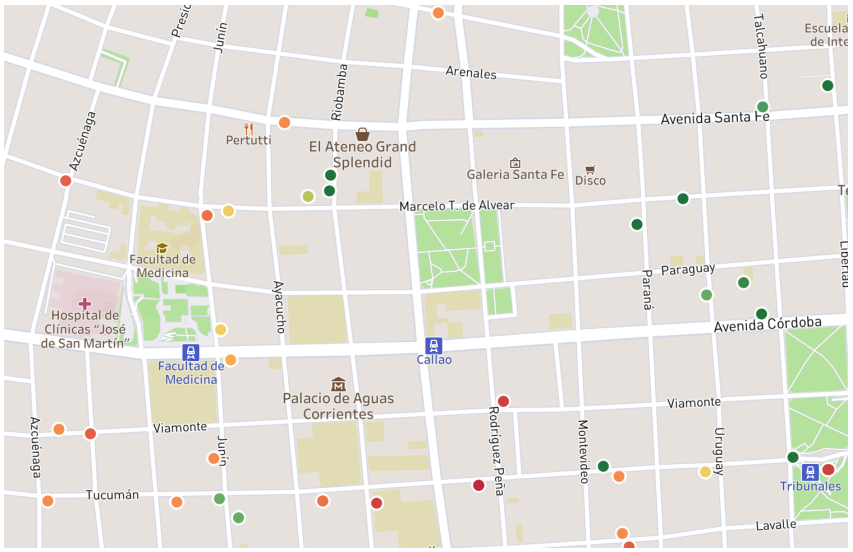
Visit Property

Scenario Simulation 3: Family Focused Buyer

Profile Description: Buyer looking for a 3+ room apartment close to the “Carlos Pellegrini” school in Recoleta.

This individual starts by selecting areas close to renowned schools, such as the "Carlos Pellegrini" school in Recoleta, using the interactive map. They can then refine their search by setting specific parameters that suit their family's needs, such as requiring a minimum number of bedrooms and bathrooms. For instance, they might look for properties that offer large living spaces or a large balcony for the kids to play at.

Expected Price: 120k - 140k



Model Recommendations:

Title:	Venta Departamento Duplex 3 Ambientes Barrio Norte Patio Estilo Frances
Expected Profit:	0.762
Rooms:	3.000
Age:	0.0
Private Parking Spaces:	0.000
Monthly Cost:	24,000
Pool:	0.000
Predicted Price:	211,441
Actual Price:	120,000
Total Surface:	108.0

Visit Property

Title:	Venta Departamento 3 Ambientes Dependencia 2 Balcones Lavadero Recoleta
Expected Profit:	0.453
Rooms:	3.000
Age:	55.0
Private Parking Spaces:	0.000
Monthly Cost:	38,000
Pool:	0.000
Predicted Price:	187,465
Actual Price:	129,000
Total Surface:	93.0

Visit Property

Title:	Departamento De 4 Ambientes En Venta Con Renta - Tribunales
Expected Profit:	0.256
Rooms:	4.000
Age:	50.0
Private Parking Spaces:	0.000
Monthly Cost:	6,000
Pool:	0.000
Predicted Price:	160,826
Actual Price:	128,000
Total Surface:	80.0

Visit Property

Model Discouragements:

Title:	Departamento 3 Ambientes - Recoleta
Expected Profit:	-0.147
Rooms:	3.000
Age:	40.0
Private Parking Spaces:	0.000
Monthly Cost:	0
Pool:	0.000
Predicted Price:	118,608
Actual Price:	139,000
Total Surface:	55.0

Visit Property

Title:	2784mp-impecable Oportunidad A Media Cuadra De Av. Córdoba
Expected Profit:	**-0.171**
Rooms:	**3.000**
Age:	**40.0**
Private Parking Spaces:	**0.000**
Monthly Cost:	**5,000**
Pool:	**0.000**
Predicted Price:	**114,340**
Actual Price:	**138,000**
Total Surface:	**58.0**

Visit Property

Limitations & Potential Spin Offs

Property Pictures Incorporation

This framework could potentially highly benefit from the incorporation of feature extraction of the incorporation of each publication's pictures. Here are some examples of potential variables that could be extracted from the model:

Visual Property Condition Assessment

By analyzing property pictures, machine learning models can assess the condition of various elements like walls, flooring, fixtures, and appliances. This analysis can provide insights into the overall state of the property, allowing for more accurate valuation based on its current condition.

Feature Extraction for Comparative Analysis

Machine learning models can extract visual features from property pictures, such as architectural styles, room layouts, and available natural light sources. These features can be used to better understand some unique selling points some properties have which can heavily influence their valuations.

Location Analysis with Visual Cues

Property images can provide visual hints about the surrounding environment, such as proximity to parks, views, or cityscapes. Machine learning models can extract and analyze these cues to assess the desirability of the location, which can impact the valuation of the property.

Amenities and Upgrades Identification

By analyzing property pictures, machine learning models can identify and classify amenities or upgrades present in the property, renovated kitchens, bathrooms and room distribution. The presence or absence of these features can affect the valuation by reflecting the added value they bring.

Renovation costs inclusion

The cost of renovations can have a significant impact on valuation models when assessing the value of a property. Renovations and improvements can enhance the overall appeal of a property, impacting its potential market value.

Hinting to the Property Pictures spin-off we just discussed, if a reliable way of detecting if a property is already renovated or it needs renovations, we could potentially refactor the model to incorporate renovation costs into the equation and provide an even better estimation of the return on investment. Moreover, since not all renovations necessarily translate into a proportional increase in property value, if you can abstract

Customer-Tailored Model Developments

In order to improve precision (and if enough data is available) it would be possible to develop specific implementations of this analysis within different levels of granularity. For example, a real estate agency focusing in a single neighborhood with particular characteristics (For example Puerto Madero) may be interested in having a tailored model only for that part of the city. Data volume may become a challenge for these cases.

Conclusions

The development of a real estate valuation tool with an interactive map can provide a flexible solution for private investors looking for opportunities within Buenos Aires's real estate market. Leveraging machine learning algorithms we were able to train a reliable valuation model which can successfully identify and categorize investment opportunities based on their potential profitability. An interactive color-coded representation of the results enables users to quickly assess the available alternatives within the scope of what they are looking for, facilitating their decision-making process by providing an instant overview of the investment landscape.

The interactive nature of the map allows users to explore investment opportunities at their preferred scale, whether it be a macro-level view of a region or a more localized examination of specific neighborhoods. Users can access detailed information and perform further analysis on individual properties of interest, ensuring a comprehensive understanding of each opportunity's potential. While the development of this real estate valuation tool proved to be quite reliable , there are opportunities for future improvements. By integrating additional models tailored to analyze images, there are a lot of potential spin offs which could be derived from it.

References

[01] (2018). Depreciated Replacement Cost: Improving the Method Through a Variant Based on Three Cornerstones. Real Estate Management and Valuation, 33-47.

[02] Trifonov, N. (2021). Income approach for real estate valuation.

[03] (2023). Application of Income Capitalization Approach for Emerging Rental Income Cash Flow in Enugu Metropolis Property Market. Property Management., 1-19

[04] (2020). Applying the Depreciated Replacement Cost Method When Assessing the Market Value of Public Property Lacking Comparables and Income Data. Sustainability, 12(8993), 8993

[05] (1988). Ridge Regression in Real Estate Analysis. The Appraisal Journal, 56(3), 311.

[06] (2004). Mass Appraisal: An Introduction to Multiple Regression Analysis for Real Estate Valuation. Journal of Real Estate Practice and Education, 7(1), 65-77.

[07] (2004). Artificial Intelligence Applied to Real Estate Valuation: An Example for the Appraisal of Madrid. Catastro. 255-265

[08] (2002) Residential Real Estate Prices: A Room with a View, JRER, 23(1), 130-137

[09] (2001) Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks, JRER, 22(3), 314-335

[10] (2003). Real estate appraisal: a review of valuation methods. Journal of Property Investment & Finance, 21(4), 383-401.

[11] (2021). Computational Valuation Model of Housing Price Using Pseudo Self Comparison Method. Sustainability, 20(13), 11489.

[12] (2008). Assessing the role of income and interest rates in determining house prices. Economic Modelling. 25. 377-390

[13] (2021). The cadastral value as a tool for monitoring the real estate market value. SUJES, 1(37), 84-108.

[14] (2016). Measurements of Rationality for a Scientific Approach to the Market-Oriented Methods. Journal of Real Estate Literature, 2(24), 403-427.

[15] (2019). Simulative Verification of the Possibility of using Multiple Regression Models for Real Estate Appraisal. Real Estate Management and Valuation, 3(27), 109-123.

[16] (2015). A market comparison method evaluation model based on set pair analysis..

[17] (2019). Application of AHP Method in Assessment of the Influence of Attributes on Value in the Process of Real Estate Valuation. Real Estate Management and Valuation, 4(27), 15-26.

[18] (2021). International Real Estate Review. IRER, 2(24), 139-183.

[19] (2022). Peculiarities of applying methods based on decision trees in the problems of real estate valuation. Bus. Inform., 4(16), 7-18.

[20] (2012). A Copulas-Based Approach to Modeling Dependence in Decision Trees. Operations Research, 1(60), 225-242.

[21] (2010). Comparison of Real Asset Valuation Models: A Literature Review. IJBM, 5(5).

[22] (2022). Real Estate Marketing Adaptive Decision-Making Algorithm Based on Big Data Analysis. Security and Communication Networks, (2022), 1-11.

[23] (2018). Valuation Construction Permit Uncertainties in Real Estate Development Projects with Stochastic Decision Tree Analysis..

[24] (2023). Boosting the Accuracy of Commercial Real Estate Appraisals: An Interpretable Machine Learning Approach. J Real Estate Finan Econ.

[25] (2014). A Neural Network based Model for Real Estate Price Estimation Considering Environmental Quality of Property Location. Transportation Research Procedia, (3), 810-817.

[26] (2018). Artificial Neural Networks and the Mass Appraisal of Real Estate. Int. J. Onl. Eng., 03(14), 180.

[27] (2021). The Complex Neural Network Model for Mass Appraisal and Scenario Forecasting of the Urban Real Estate Market Value That Adapts Itself to Space and Time. Complexity, (2021), 1-17.

[28] (2022). The contribution of statistical models in the field of real estate valuation. Timisoara Journal of Economics and Business, 1(15), 111-126.

[29] (2023). Intelligent real estate management. neu, 1, 16-20.

[30] (2020). Review on the Application of Artificial Neural Networks in Real Estate Valuation. IJATCSE, 3(9), 2918-2925.

[31] (2022). House Price Valuation Model Based on Geographically Neural Network Weighted Regression: The Case Study of Shenzhen, China..

[32] (2022). Using machine learning algorithms for predicting real estate values in tourism centers. Soft Comput, 5(27), 2601-2613.

[33] (2021). Application of BP neural network technology on dynamic financial risk prediction. J. Phys.: Conf. Ser., 1(2004), 012015.

[34] (2019). Predicting system for the estimated cost of real estate objects development using neural networks. The Journal of Zhytomyr State Technological University Series Engineering, 1(83)(0), 154-160.

[35] (2020). Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices. Applied Sciences, 17(10), 5832.

[36] (2018). Artificial Neural Networks and the Mass Appraisal of Real Estate. Int. J. Onl. Eng., 03(14), 180

[37] (2012). Application of Support Vector Machine in Determination of Real Estate Price. AMR, (461), 818-821.

[38] (2022). DOES MACHINE LEARNING PREDICTION DAMPEN THE INFORMATION ASYMMETRY FOR NON-LOCAL INVESTORS?. International Journal of Strategic Property Management, 5(26), 345-361

[39] (2022). Economic Crisis Early Warning of Real Estate Companies Based on PSO-Optimized SVM. Journal of Sensors, (2022), 1-10.

[40] (2022). Using Machine Learning Algorithms for Predicting Real Estate Values in Tourism Centers.

[41] (2015). Study on the Prediction of Real estate Price Index based on HHGA-RBF Neural Network Algorithm. IJUNESST, 7(8), 109-118.

[42] (2010). A Risk Early Warning Model in Real Estate Market Based on Support Vector Machine.

[43] (2008). Evaluation model for Real Estate Investment Environment Based on SVM..

[44] (2021). USING SVR AND MRA METHODS FOR REAL ESTATE VALUATION IN THE SMART CITIES. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., (XLVI-4/W5-2021), 21-26.

[45] (2020). The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach. Risks, 4(8), 112.

[46] (2023). House Price Prediction Model Using Random Forest in Surabaya City. Tem Journal, 126-132.

[47] (2022). Моделі для формування ринкової вартості нерухомості. збірник «Адаптивні Системи Автоматичного Управління Міжвідомчий, 41(2), 58-64.

[48] (2022). Real estate valuation based on big data. Vopr. èkon., 12, 118-136.

[49] (2022). Using Machine Learning Algorithms for Predicting Real Estate Values in Tourism Centers..

[50] (2020). Mass Apprasial With A Machine Learning Algorithm: Random Forest Regression. Bilişim Teknolojileri Dergisi, 3(13), 301-311.

[51] (2022). COMBINATION OF MACHINE LEARNING-BASED AUTOMATIC VALUATION MODELS FOR RESIDENTIAL PROPERTIES IN SOUTH KOREA. International Journal of Strategic Property Management, 5(26), 362-384.

[52] (2021). Machine Learning, Deep Learning, and Hedonic Methods for Real Estate Price Prediction..

[53] (2022). Rapid Modelling of Machine Learning in Predicting Office Rental Price. IJACSA, 12(13).

[54] (2018). RANDOM FOREST ALGORITHM OPTIMIZATION OF ENTERPRISE FINANCIAL INFORMATION MANAGEMENT SYSTEM. LAAR, 4(48), 255-260.

[55] (2022). House price prediction using hedonic pricing model and machine learning techniques. Concurrency and Computation, 27(34).

[56] (2022). Spatial Determinants of Real Estate Appraisals in The Netherlands: A Machine Learning Approach. IJGI, 2(11), 125. https://doi.org/10.3390/ijgi11020125

[57] (2021). Property Mass Valuation on Small Markets. Land, 4(10), 388.

[59] (2020). Risk Factors Impacting the Project Value Created by Green Buildings in Saudi Arabia. Applied Sciences, 21(10), 7388.

[60] (2013). Modern Challenges Facing the Valuation Profession and Allied University Education in Poland. Real Estate Management and Valuation, 1(21), 14-18.

[61] (2017). Statistical Determination of Impact of Property Attributes for Weak Measurement Scales. Real Estate Management and Valuation, 4(25), 75-84.

[62] (2016). Assessment of Predictor Importance with the Example of the Real Estate Market. Folia Oeconomica Stetinensia, 2(16), 29-39.

[63] (2021). USING SVR AND MRA METHODS FOR REAL ESTATE VALUATION IN THE SMART CITIES. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., (XLVI-4/W5-2021), 21-26.

[64] (2021). Cadastral Valuation Of Agricultural lands In The Context Of Spatial Development Of The Agro-Industrial Complex In Russia..

[65] (2011). Incorporating Sustainability into Real Estate Valuation: the Perception of Nigerian Valuers. JSD, 4(4)

[66] (2021). Prior information in econometric real estate appraisal: a mixed estimation procedure. JERER, 3(14), 349-361

[67] (2011). A new paradigm for real estate valuation?. Journal of Property Investment & Finance, 4/5(29), 341-358.

[68] (2018). Sustainable Value of Investment in Real Estate: Real Options Approach. Sustainability, 12(10), 4665.

[69] (2016). The Mass Appraisal Tool: Application of a Pluri-Parametric Model for the Appraisal of Real Properties., 39-52.

[70] (2023). Modelling the drivers of data science techniques for real estate professionals in the fourth industrial revolution era. PM. https://doi.org/10.1108/pm-05-2022-0034

[71] (2013). Valuing a greenfield real estate property development project: a real options approach. J of Eur Real Est Research, 2(6), 186-217.

[72] (2016). Assessment of Predictor Importance with the Example of the Real Estate Market. Folia Oeconomica Stetinensia, 2(16), 29-39.

[73] (2010). Valuation of immovables based on an econometric models. eq, 1(4), 241-252.

[74] (2016). Assessment of Predictor Importance with the Example of the Real Estate Market. Folia Oeconomica Stetinensia, 2(16), 29-39.

[75] (2019). Principles and Criteria for using Statistical Parametric Models and Conditional Models for Valuation of Multi-Component Real Estate. Real Estate Management and Valuation, 2(27), 33-43.

[76] (2015). Green Certification and Building Performance:Implications for Tangibles and Intangibles. JPM, 6(41), 151-163.

[77] (2022). The contribution of statistical models in the field of real estate valuation. Timisoara Journal of Economics and Business, 1(15), 111-126.

[78] (2009). Vertical phasing as a corporate real estate strategy and development option. Journal of Corporate Real Estate, 3(11), 144-157.

[79] (2019). Simulative Verification of the Possibility of using Multiple Regression Models for Real Estate Appraisal. Real Estate Management and Valuation, 3(27), 109-123

[80] (2020). Analytical Method for Correction Coefficient Determination for Applying Comparative Method for Real Estate Valuation. Real Estate Management and Valuation, 2(28), 52-62.

[81] (2022). Modeling procedure within the mass valuation of real estate in Slovenia. geod. vest., 02(66), 258-279.

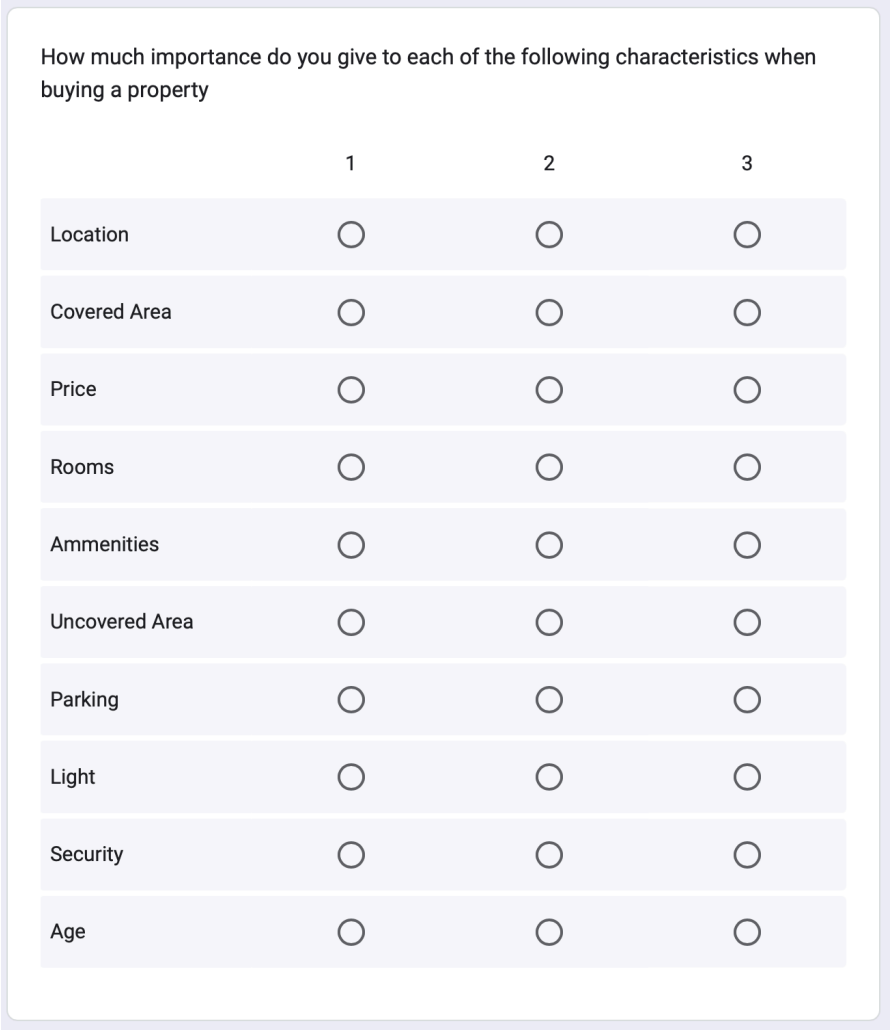
[82] (2018). KNOWLEDGE-BASED FIS AND ANFIS MODELS DEVELOPMENT AND COMPARISON FOR RESIDENTIAL REAL ESTATE VALUATION. International Journal of Strategic Property Management, 2(22), 110-118.

Annex

[1] Links to technologies used:

- https://developers.mercadolibre.cl/es_ar/desarrollos-inmobiliarios
- https://pandas.pydata.org/
- https://geopandas.org/en/stable/
- https://numpy.org/
- https://scikit-learn.org/stable/
- https://xgboost.readthedocs.io/en/stable/python
- https://www.tableau.com/
- https://www.mapbox.com/

[2] Preferences Survey Grid



How much importance do you give to each of the following characteristics when buying a property

	1	2	3
Location	○	○	○
Covered Area	○	○	○
Price	○	○	○
Rooms	○	○	○
Ammenities	○	○	○
Uncovered Area	○	○	○
Parking	○	○	○
Light	○	○	○
Security	○	○	○
Age	○	○	○