uc3m | Universidad Carlos III de Madrid

Master Degree in Computational Social Science
2022-2023

Master Thesis

"Forecasting dwelling prices in the City of Madrid: a comprehensive analysis of the Real Estate market"

Maria del Mar Escalas Martorell

Tutor

María Medina Pérez

Madrid, 2023

SUMMARY

Dwelling prices in Spain, specially in big cities as Madrid, have faced big increases these last years. Real Estate market in Madrid City presents more variability in prices than Spanish country as a whole. This, together with the availability of data, create a suitable scenario for studying in deep this phenomenon.

Due to the lack of microdata from public sources, data from private entities has been needed to conduct the analysis.

This study aims to capture inherent and external characteristics of houses, which affect directly their price. Starting with a baseline linear model, computational effort is increased while introducing more complex algorithms that provide better predictive results.

Results obtained show a high predictive power of prices for homes in the City of Madrid.

Key words: Real Estate, dwelling price, data, R

INDEX OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Motivation of the work

In Spain, housing has become a major problem for the society. High prices make very difficult to buy a house that meets people's preferences. Especially, young people use to have solvency problems to achieve the requirements needed for being given a mortgage. (López, 2018)

This problem is aggravated when we focus on big cities like Madrid or Barcelona. They are the most international cities of the country and they receive many tourists each year. This presence of tourism, located mostly in the city centre, makes that house owners dedicate their homes to host tourists, making the dwelling offer to reduce. In addition, the presence of international market provokes that foreign people (which, in many occasions, has more purchasing power than the locals) is interested in buying houses, so that demand increases. (Crespí & Domínguez, 2021)

In this Master's Thesis, it is going to be studied how to identify the factors that affect the price of houses, and in which measure.

The fact that not a lot of studies provide a quantitative approach to consider external factors make essential to set a number to these effects. In addition, the analysis of dwelling prices can result of interest for individuals seeking to purchase property, as well as for investors willing to enter the Real Estate market.

1.2. Objectives

The main objective of this Master's Thesis is to construct a predictive model using the R programming language to accurately forecast prices of dwellings in the selected area. This implies the application of the techniques acquired during the course, this being one of the main reasons for developing it.

Specifically, this process requires a set of steps to complete the work: it is necessary to present the problem that motivates the work, in addition to narrowing the study area and correctly defining the variable that better represents the price of a dwelling. The specific objectives when building the model is to enrich the data with indicators related to the outcome that may result valuables for improving results.

Finally, it is desired to present the findings after providing a complete understanding of the process carried out.

By achieving these objectives, it is intended to contribute to add valuable options for developing an investment portfolio, as well as identifying market trends and opportunities.

1.3. Methodology

Firstly, the study will begin with a comprehensive presentation of the problem, giving information about the housing market and other contributing factors that are relevant to the issue. This will provide an overview of the situation and its impact on the Spanish society. Furthermore, narrowing the study to a certain location will require considering the unique characteristics, dynamics and factors influencing the market in that specific region.

In the following section, data needed will be extracted from various sources and cleaned to ensure its accuracy and reliability, with the aim of present it in a suitable format for further analysis. All data used in this work is open source. In order to identify patterns, errors, and special insights in the data, several exploratory analysis (partial and full) will be carried out.

The data will be incorporated into the models using a variety of methodologies and results obtained will be compared to identify patterns and key drivers for the rising dwelling prices.

Finally, study's findings will be presented.

It is worth mentioning that this work is presented together with additional files that are recommended to review to understand the bulk of the work. These files can be found in the GitHub repository: https://github.com/mariadelmarm19/Masters-Thesis-Dwellings-Prices.git, which is available online. Files included are:

- This document
- README file as an introductory explanation of the project and usage guide
- .Rmd with the code corresponding to the sections in this document: **1. INTRODUCTION** and **2. ANALYSIS OF THE SITUATION** (*house_prices_I.Rmd*)
- .Rmd with the code corresponding to section **3. DATA EXTRACTION AND PREPROCESSING** (*house_prices_II.Rmd*)
- .Rmd with the code corresponding to sections **4. MODELLING** and **5. CONCLUSIONS** (*house_prices_III.Rmd*)
- .html files of *house_prices_II.Rmd* and *house_prices_III.Rmd*
- All data files needed to run the code

1.4. Technical requirements and precautions

Reader should be aware of the technical requirements before engaging with this work. The study extensively uses R programming, GitHub and statistical concepts. Therefore, a prior knowledge with these tools and topics is recommended for a better understanding.

Proficiency in R programming is essential for comprehending and replicating the code provided accompanying this Thesis. Reader should be familiar with regular expressions, data handling, visualization, modelling, and know about popular R packages such as *tidyverse*, *ggplot2* and *caret*. For executing the code, it is also needed to have access to a development environment, such as RStudio.

The majority of figures (plots) in this document have been extracted from the renderization of the .Rmd in Word. For achieving this, figure height and width (*fig.height* and *fig.width*) options in each code chunk containing plots have been modified to get the desired dimensions in the output plot.

The associated code (see in GitHub repository) has been built for running in a Windows operating system. Especially in web scraping and API querying, the methodology may be different for other operating systems.

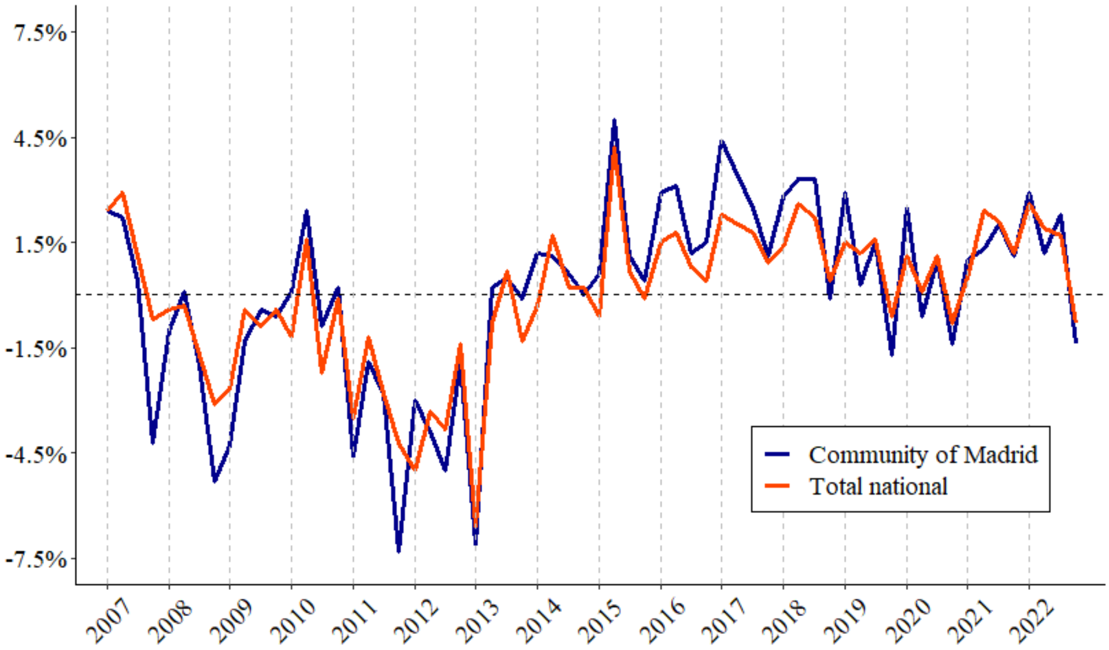
The code includes several commented chunks that provide additional options, examples or explanations. Some of these chunks are not executed by default but serve as a reference for the user to choose the desired action. Depending on the specific case, the commented code may imply a unique execution or offer multiple options for completing the task. By including these examples, the code aims to provide flexibility and a better understanding of the process. During the Thesis reading, the user can try different scenarios by selectively uncommenting and executing the relevant code segments.

2. ANALYSIS OF THE SITUATION

2.1. Prices of dwellings in Spain

According to the Bank of Spain (2020), compared to other European countries, young people and low income households in Spain have more difficulties to access to a dwelling: they present low acquisition power, being their incomes not enough for covering those high housing costs. In addition to this, the Spanish property market has experienced tremendous ups and downs over the past years. Despite some small fluctuations, prices in the market have been on an upward trajectory since 2014.

Figure 1*. HPI (Housing Price Index) variation over time (quarters)*



Source: Own elaboration. Data retrieved from National Institute for Statistics in Spain: Housing Price Index

Figure 1 shows the historic prices variation of the Real Estate market. Housing Price Index (HPI) serves to evaluate the behaviour of housing prices (new and second-hand). It is calculated by the National Institute for Statistics in Spain, with information gathered from the General Council of Notaries, which covers all transactions involving physical individuals in the Spanish territory. (National Institute for Statistics in Spain, 2023)

Dwelling prices have exhibited a lot of variability over the past 15 years, with changes that have aligned with the economic cycles of the respective period. (Bank of Spain, 2020)

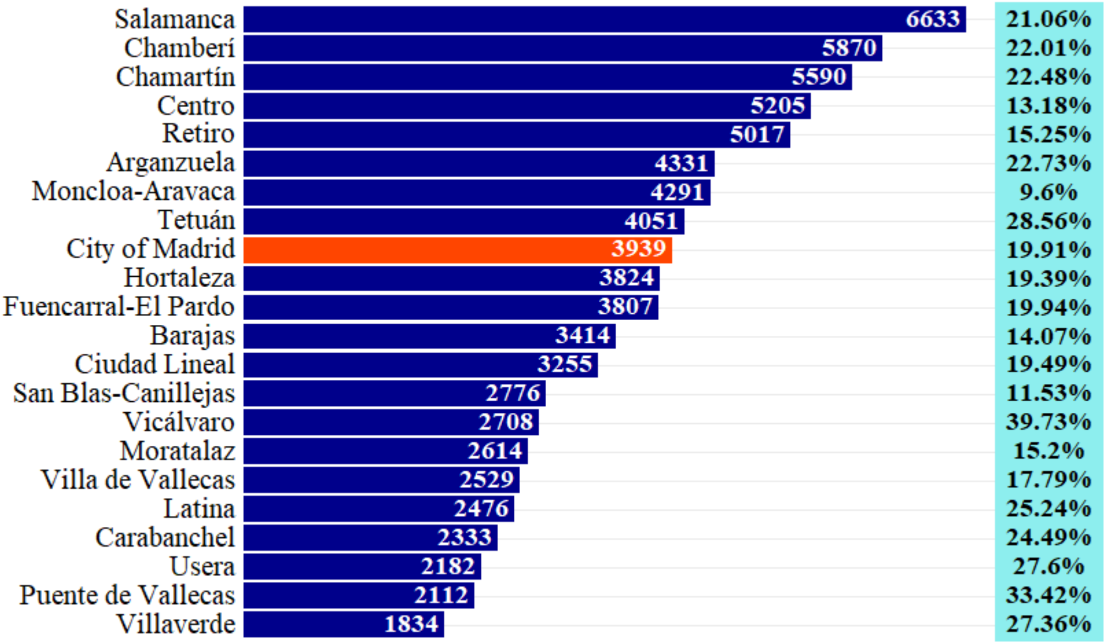
When plotting these national values alongside the specific data for the Community of Madrid, it can be appreciated almost in every year that the orange line is below (in the decreasing periods) and above (in the increasing periods) than the total national.

2.2. Prices of dwellings in the City of Madrid

After the initial exposition, it has been determined that the study will specifically concentrate on the City of Madrid. This choice aims to avoid (or minimize) the aggregation bias, which arises when generalizations are wrongly applied to individual observations.

Displayed below there is a combined graph presenting data on the price of homes according to the district of the city in the year 2022. Additionally, the accompanying information illustrates the percentage increase that prices have experienced from 2017 to that year.

Figure 2. *Average price per square meter of second-hand dwellings by District in 2022, in euros. Increase in price per square meter: 2017 vs. 2022, in %*



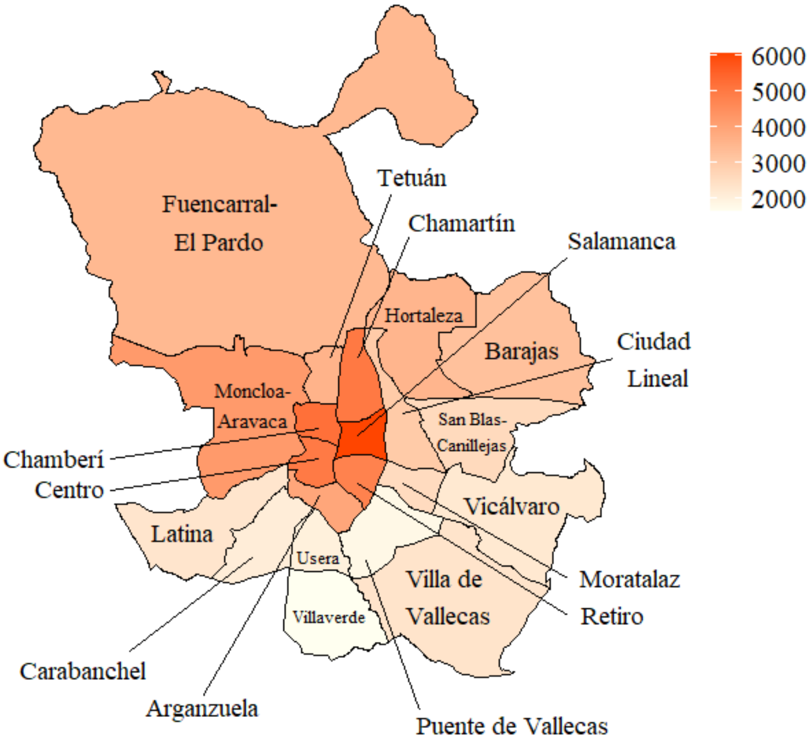
Source: Own elaboration. Data retrieved from Open Data Portal of Madrid City Council: evolution of the price of second-hand housing (€ per square meter) by District and Neighbourhood

Salamanca stands out as the leading district in price per square meter. Following closely are Chamberí, Chamartín, Centro and Retiro, which also exceed €5,000 per square meter. On the other hand, the most affordable districts are priced below €2,200, much lower than the city average (almost 4,000).

Price variation in five years has been positive for all districts, with the largest increases occurring in the cheapest districts, where prices have risen by over 25%.

When mapping this information about prices, the geographical disparities between districts highlight. Homes located in the central districts are, in average, more expensive. Districts in the South, and even in the Southeast depict the lowest prices of the city. Northern areas move around the average of the city.

Figure 3. *Average price per square meter by District (2022), in euros*



Source: Own elaboration. Data retrieved from Open Data Portal of Madrid City Council: volution of the price of second-hand housing (€ per square meter) by District and Neighbourhood and Madrid municipal districts (geospatial data)

The seven central districts of the map: Tetuán, Chamartín, Salamanca, Chamberí, Centro, Retiro and Arganzuela make up the so-called "central almond", around which the M-30 runs, one of the most important highways in the city.

Despite plotting only the variable of price per square meter, it can be easily observed how the geographical factor is relevant for determining prices in the city. Indeed, a spatial market segment has been identified within the M-30 (Madrid first ring road) when introducing spatial features to housing prices models. (Rey-Blanco, D. *et al.*, 2023)

In fact, according to Bustos V. (2022), there is a strong evidence of difference in prices of dwellings inside M-30, being much higher than dwellings in the South and East. For homes in the North, this difference also exists, but is reduced due to the expectations of revaluation of the zone in the future.

3. DATA EXTRACION AND PREPROCESSING

3.1. Dwellings' characteristics data

The primary data essential to the study comprises information related to the available housing units in the Real Estate market, along with their characteristics. While it would be ideal to obtain this information from public entities, the reality is that there is no such in deep information. For this reason, the project needs to consider private options. The data that are of interest when talking about a home include metrics such as: square meters, number of bedrooms and bathrooms, location and antiquity of the property, or availability of amenities. These metrics, among others, help provide a clear picture of a property's overall appeal and value in the market.

In this case, some online Real Estate platforms in Spain were addressed to access the required data. Finally the data to be used in the study was granted by Idealista.

Idealista is an online platform for buying, selling and renting Real Estate assets, operating in Spain, Italy and Portugal. As of May 2023, it offered more tan 1 million properties in Spain, including houses, garages, storage spaces and other properties. (Idealista, 2023)

Since it is impossible to access private offers in an individual basis, neither can be found public information, this is considered the best method to analyse the overall housing stock in Spain. This study will use asking price of posts in Idealista as a proxy to final price of a dwelling.

3.1.1. Idealista (2023)

By courtesy of Idealista, access to API's real-time data was provided. However, this admittance was limited in time and quantity, making it impossible to complete the work solely with the data obtained through this methodology. Nevertheless, it is of vital importance to include a section on the process of extracting this data from Idealista's API, given that a majority of private data source operate through API systems.

To begin the process, it was necessary to send a request to the company, providing personal user data and objective. Subsequently, credentials containing two secret numbers were received: the "API Key" and the "Secret". Due to privacy requirements, personal credentials are hidden in the code provided. The person who desires to reproduce this part of the code needs to ask for their own credentials and create the request with them. With this unique and non-transferable identification, an encoding process should be performed for then proceed to authenticate. After that, the platform returns another credential called *Token*. This "Token" serves to make requests to the API.

Now, parameters to make the request need to be defined. These parameters are explained in the complementary documentation provided together with the access credentials. These instructions detail all aspects that need and can be used as filters to make requests, like for example country of search (*country*), buy or rent operation (*operation*), geographic point of reference (*center*) or distance to point of reference (*distance*).

Next, it is necessary to select the type of property: in this case, "homes". Additionally to this, it is possible to filter by other additional home specific filters, such as the number of rooms, whether it is a studio, or whether it has a garage.

Once the filters have been established, the request can be constructed and sent. In this case, the API will return a maximum of 50 properties that meet those characteristics, ordered by the shortest distance from the location selected in "center".

As mentioned, access to this information is limited, both in terms of time and volume. The permission lasted only from February to July, with a limit of 100 monthly requests and a maximum of 50 properties per request. In addition, obtained data consisted only of real-time data, not historical data.

3.1.2. Idealista (2018)

The main data that will be used for modelling comes from the same source. The company made it available to users through an R library: *idealista18*. It is also available in GitHub: https://paezha.github.io/idealista18/. The information contains data on more than 180,000 properties, along with their quarter and city (Madrid, Barcelona or Valencia) in the year 2018. In this case, just Madrid data will be used.

The original data is presented in "sf" and "dataframe" formats. The full dataset is of 94,815 observations and 42 variables. These variables go from the most common (price, square meters, number of rooms…) to others more specific (having a terrace, having a lift, being the parking included in price, number of dwellings in the same flat…). Apart from this, it is also interesting to see that variables about the distance of the house to a metro station and to the city centre are also included in the original data.

The cleaning is started by selecting the variables of interest. The decision has been to reduce the dimensionality of data, getting rid of variables that were not informative or providing repeated information. In addition, a renaming has been carried out to a better understanding. After making this initial selection, it comes to light that the data has several serious issues that need to be processed.

It has been detected that there are a lot of duplications, and even dwellings that are repeated more than twice. The point is that sometimes they do not contain complete information, causing noise in the data and incorporating errors.

In addition to this, there are numerous dwellings without orientation. While it is true that it may be the case that a dwelling is not oriented towards some point (basements), it is not common enough to appear to often. This detail can not be overlooked because it is a very important variable for the model.

In order to handle the problem of duplication and to mitigate the lack of orientation information, it was decided to undertake the following data processing:

- In the case of the variable “external” (whether the dwelling is external or not), a recoding of the “yes” = 1 to the same 1, and “no” = 2 to 0 was applied. A small number of rows with missing values were eliminated.
- Group observations by id to make operations: apply the average to numeric columns and select the maximum value in the case of factor columns (already in numeric format) to detect the maximum number in each column. If this characteristic positively existed in any of the rows of the same house, the information for the row resulting from this operation would finally be completed.
- Bivariate outcome columns were converted to factors (1 = “yes”, 0 = “no”).

To integrate orientation information into just one variable, it was decided to classify properties according to their orientation quality. For those properties that did not have an orientation, "none" was assigned. Then, “bad” for those that were only oriented to the North. For the East and any combination with the East: "good". For the rest of the combinations: "medium", and for those properties with orientation to all directions: "full”.

There were 41 dwellings with missing information about neighbourhood: those houses were located in the boundaries of two neighbourhoods. The decision to address this problem was to manually look for each individual observation, using their coordinates to classify them. Despite of the final neighbourhood assigned, the interest remains in the district, which is the same for both neighbourhoods.

Finally, the neighbourhood to which each property belongs was identified thanks to the geographical coordinates and the neighbourhood classification that can also be found in the same library “Madrid_Polygons”. A *join* operation needed to be done to achieve this.

Cleaned data is presented as follows:

Table 1. *Main dataset: data about Real Estate market in the City of Madrid "madrid_idealista"*

id	period	price	built_area	price_sq_m	n_room	n_bath	terrace	lift	air_cond	parking	boxroom	wardrobe	pool	doorman	garden	external	year_built	floors	n_dwelling	km_to_center	metro_proximity	orientation	neighbourhood	geometry
A15019136831406238029	2018 Q1	126000	47	2681	1	1	0	1	1	0	1	1	1	1	1	1	2005	7	319	8.06	0.87	None	Carabanchel	<S3: sfc_GEOMETRY>
A6677225905472065344	2018 Q1	235000	54	4352	1	1	0	0	0	0	0	1	0	0	0	0	1900	5	11	0.88	0.12	None	Palacio	<S3: sfc_GEOMETRY>
A13341979748618524775	2018 Q1	373000	75	4973	2	1	0	0	1	0	1	1	0	0	0	1	1915	6	26	0.91	0.14	Good	Palacio	<S3: sfc_GEOMETRY>
A4775182175615276542	2018 Q1	266333	44	5917	1	1	0	1	1	0	0	0	0	0	0	0	1947	9	15	0.86	0.13	None	Palacio	<S3: sfc_GEOMETRY>
A2492087730711701973	2018 Q1	228000	50	4560	0	1	0	0	0	0	0	0	0	0	0	1	1930	5	19	1.25	0.34	None	Palacio	<S3: sfc_GEOMETRY>
A9587449507628658013	2018 Q1	425000	70	6071	1	1	0	1	0	1	0	0	0	0	0	1	1900	5	16	0.75	0.44	Medium	Palacio	<S3: sfc_GEOMETRY>

Source: Own elaboration. Data retrieved from idealista18 data package.

3.2. Other external data

Furthermore, other data that may have impact on housing prices should be taken into consideration. Factors such as Euribor or number of homes dedicated to tourism may also affect prices, making it important to consider them beforehand.

In addition, distance to places of interest will be added to the model for ascertain if closeness to any of these places increases the price. The facilities to take into account comprehend parks and green areas, educational centers and medical care (health) centers.

All variables that measure any distance to/from a point of interest will be measured in kilometres, in order to analyse data in the same units as data provided in the main source.

3.2.1. Euribor

To try to enhance the model's value, the plan is to incorporate Euribor data for each time period. The Euribor is an indicator that defines the value of money borrowed between European banks. It is commonly used to calculate interest rates for both variable and fixed mortgages. This can determine the demand for buying houses so, theoretically, the lower Euribor, the more demand of houses, the higher house prices. (Datosmacro, 2018)

The data is obtained by scraping the monthly Euribor data from Expansion.com (2018) website.

First, with the aim of avoiding any issue with the website owners, it is recommended to identify the scraper person by using the User-Agent. This is a unique online identification of each machine.

The website is made up in XML so, by using *scrapex* package in R, the table including the information can be easily accessed:

First, it is necessary to read properly the website link into R. Next, select the table needed and, finally, handle data extracted as needed.

The following table displays the results obtained.

Table 2. *Secondary data: Euribor by quarters in the year 2018*

period	average_euribor
2018 Q1	-0.1903%
2018 Q2	-0.1863%
2018 Q3	-0.1717%
2018 Q4	-0.1433%

Source: Own elaboration. Data retrieved from Expansión.com: Historical Euribor 2018

3.2.2. Airbnb

With the objective of incorporating a very important variable for the price of houses such as tourist activity, it was decided to carry out a quantitative approximation with the data on Airbnb activity.

Using these data has some limitations that need to be mentioned. It has been obtained by scraping carried out by a third party. For this reason, it was only possible to access the data up to the month of September 2018. So, the decision was to reduce the entire study to the first three quarters of 2018.

The idea was to classify homes on Airbnb according to their geographical location (neighbourhood), so that the number of homes and tourist places can be calculated, as well as making a descriptive analysis of the situation later. It is worth mentioning that the complete dataset contains information about the whole Community of Madrid, therefore, after reading the dataset and cross-referencing it with the geographic data of "Madrid_Polygons" there were some *NAs* obtained. These *NAs* should have been removed, since they were not part of the scope of the study.

To introduce the information in the main dataset ("madrid_idealista") the procedure was the following:

- Filter by date of "found" and "revised". "found" is the first time that scraping found the house and "revised" is the last time. Therefore, it was possible to know how long a home remained on the Airbnb market. Then three different filters were carried out targeting three different datasets, one for each quarter.

- Those of the first quarter will be the "revised" before April 1, 2018.
- The second quarter dwellings are "found" after March 31, 2018 and before July 1, 2018 or "revised" after March 31, 2018 and before July 1, 2018.
- Those of the third quarter will be the "revised" after June 30.

- New datasets were assigned a column with their corresponding quarter (2018 Q1, 2018 Q2 or 2018 Q3).
- The three datasets were merged, resulting in a new and larger dataset (the same home may be available in more than one quarter).
- Now, classification operations could be carried out: group by quarter and neighbourhood and make the sum of dwellings and places for each group.

This is an snippet of the final Airbnb dataset:

Table 3. *Secondary data: number of houses and places dedicated to tourist activity (Airbnb) by quarter and neighbourhood*

neighbourhood	quarter	n_houses	n_places	geometry
12 de Octubre-Orcasur	2018 Q1	11	27	<S3: sfc_GEOMETRY>
12 de Octubre-Orcasur	2018 Q2	8	18	<S3: sfc_GEOMETRY>
12 de Octubre-Orcasur	2018 Q3	18	37	<S3: sfc_GEOMETRY>
Abrantes	2018 Q1	14	48	<S3: sfc_GEOMETRY>
Abrantes	2018 Q2	11	30	<S3: sfc_GEOMETRY>
Abrantes	2018 Q3	20	66	<S3: sfc_GEOMETRY>
Acacias	2018 Q1	107	292	<S3: sfc_GEOMETRY>
Acacias	2018 Q2	99	279	<S3: sfc_GEOMETRY>
Acacias	2018 Q3	99	276	<S3: sfc_GEOMETRY>

Source: Own elaboration. Data retrieved from Datahippo.org: Madrid (Provincia): Datos básicos Airbnb.

3.2.3. Parks

Thanks to the information provided by the Madrid Council on its data platform, it is possible to access information on the location of the parks. This data was included in the model to see if the proximity to a park or green area increases the price of a home in Madrid.

A sum of 204 points of green areas are provided. The observations can be parks, green areas or botanical and historical collections.

3.2.4. Educational centers

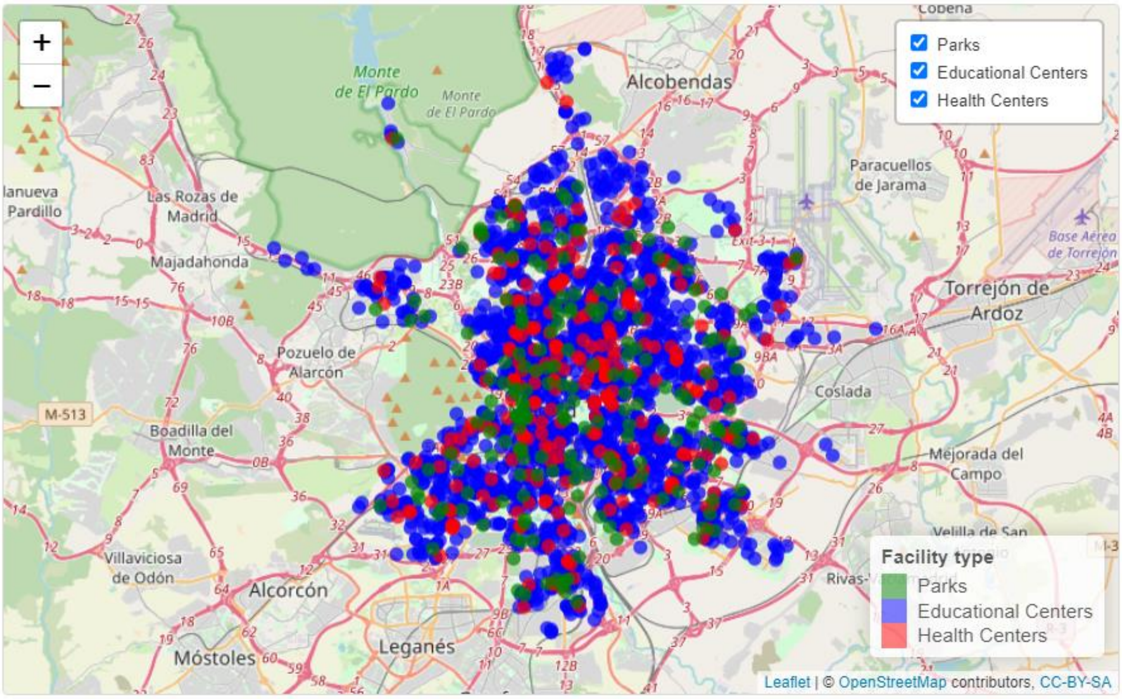
Madrid Council data platform also provides interesting documentation about study centers. Both public and private educational institutions are enclosed.

3.2.5. Health centers

The case of health centers is slightly different: there were included centers for additions, drugs and other stigmatized diseases. The decision has been to separate all health centers from those, removing this lasts, thinking that health centers proximity would imply a higher price, while the second would create the opposite effect.

The following plot shows a screenshot of an interactive map. It can be run in the code and see how facilities are distributed around Madrid, filter by type of area of interest and narrow or expand the view.

Figure 4. *Map of parks, educational centers and health centers in the City of Madrid*



Source: Own elaboration. Data retrieved from Open Data Portal of Madrid Council: parks and gardens, educational and medical health centers

3.3. Data quality assessment

It has been shown that there were several issues in raw dwelling data that require special consideration. In this section, an exploratory study will be carried out with the objective of finding special insights.

Firstly, the repetitions of id in the rows deserve a comment. The main difference between the rows with the same id were the price variations in the same house (apart from the different information about other characteristics).

To see of how much variation the situation is addressing, the percentage of variability that the price of a house can have in the market could be observed by:

- Separating in another dataset ("variations") all houses that appear more than once.
- Calculating the variation of prices by subtracting maximum price to minimum and dividing by the minimum.
- Calculating the average variation of prices.

The result obtained was a variation of a 7.78%, in average, in the price of a square meter for a house in the City of Madrid (for those houses that have had any variation in its price).

Regarding the misinformation about orientation, after grouping by id to gather information in different rows, there were still many dwellings with no orientation. It is interesting to see if those houses share characteristics.

Results for missing orientation showed that the average antiquity of those dwellings was about 56 years, so it seems that many old houses do not have this information included in the cadastre, or that the owners do not have incentives to tell about it.

4. MODELLING

Code corresponding to *Modelling* section can be found in the third .Rmd in the associated GitHub. Data used for conducting the analysis is obtained from the first .Rmd, being it presented together with both .Rmds (see *madrid_idealista.csv* file).

First, it was convenient to make a brief descriptive analysis and next, several computational techniques were applied in order to achieve the best predictive power.

Methodologies used for modelling started from lower to higher computational difficulty so, theoretically, better results were expected to arise from the more complex techniques.

Hedonic prices models were used to work with data about the Real Estate market, as recommended by experts as Rey-Blanco, D. *et al.* (2023). These models contemplate the incorporation of external variables of the construction. For this, including elements like: location, air quality, near services… may imply an increase in the quality of the model, increasing its predictive power. For this reason, all variables contained in the dataset built are going to be introduced in the model.

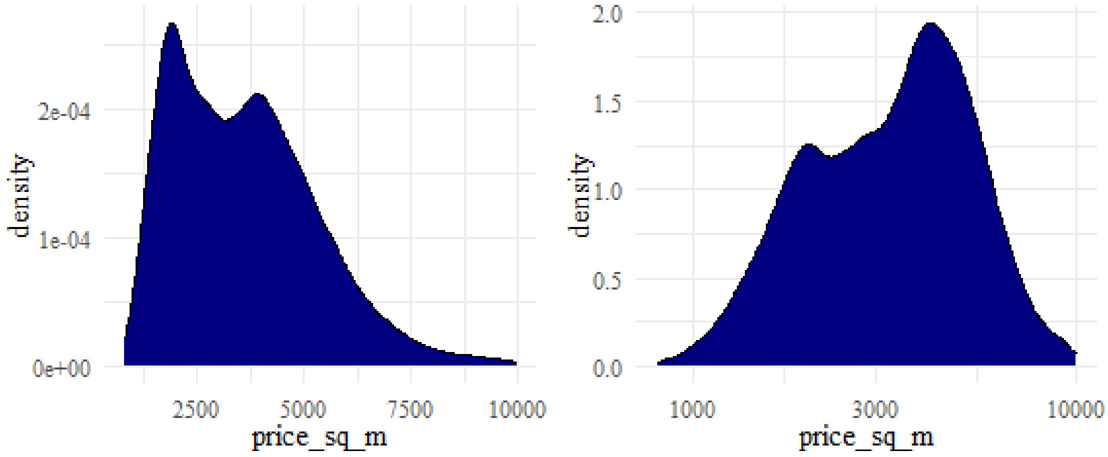
4.1. Exploratory data analysis

After checking again that the information in the dataset is correct and with no missing values, the process of analysis started with a descriptive analysis of the most important features of the dataset and its variables.

The data contains 41,217 observations (houses) on 32 variables related to the characteristics of each house and its environment.

The distribution of the dependent variable *price_sq_m* was very right skewed. To work better with this deviation it was worth it to log-transform the dependent variable.

Figure 5. *Distribution of the dependent variable: price per square meter, in euros. Distribution of the dependent variable: logarithm of price per square meter, in euros*

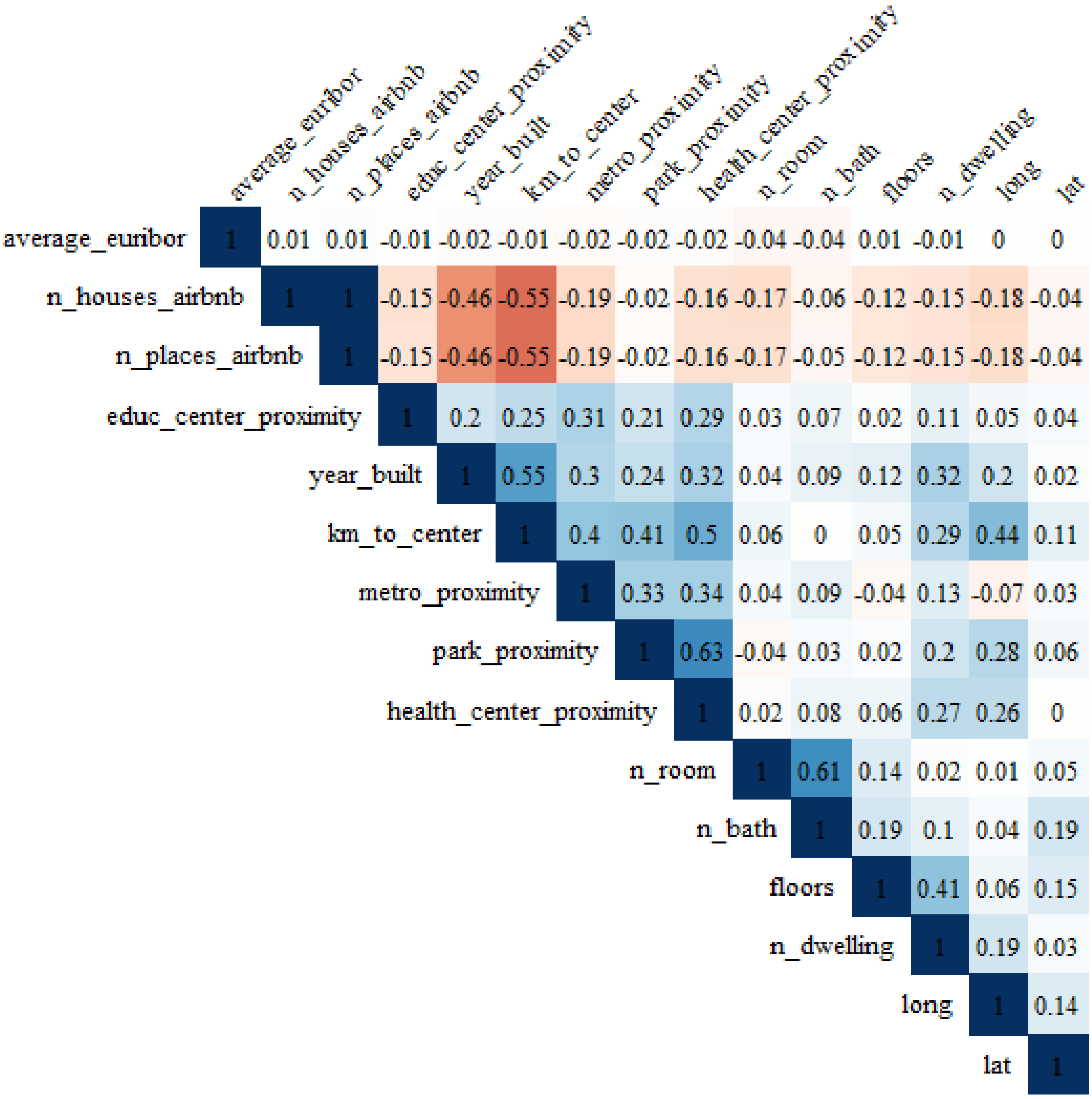


Source: Own elaboration. Data retrieved from idealista18 data package.

Explanatory variables integrating the models are presented next:

- Numeric: number of rooms and bathrooms (*n_room*, *n_bath*), count of dwellings in the building (*n_dwelling*), construction year (*year_built*), floor number (*floor*), distance to center and to a metro station (*km_to_center*, *metro_proximity*), average quarterly Euribor (*average_euribor*), longitude and latitude (*long*, *lat*), number of houses and places in Airbnb located in the neighbourhood (*n_houses_airbnb*, *n_places_airbnb*) and distance to nearest park, health and educational center (*park_proximity, health_center_proximity*, *educ_center_proximity*).
- Factor: if the house has a terrace, lift, air conditioned (*air_cond*), parking included in price (*parking*), boxroom, wardrobe, swimming pool (*pool*), doorman, garden, if it is external and its orientation quality (orientation).

Figure 6. *Matrix of correlations between numeric explanatory variables*



	average_euribor	n_houses_airbnb	n_places_airbnb	educ_center_proximity	year_built	km_to_center	metro_proximity	park_proximity	health_center_proximity	n_room	n_bath	floors	n_dwelling	long	lat
average_euribor	1	0.01	0.01	-0.01	-0.02	-0.01	-0.02	-0.02	-0.02	-0.04	-0.04	0.01	-0.01	0	0
n_houses_airbnb		1	1	-0.15	-0.46	-0.55	-0.19	-0.02	-0.16	-0.17	-0.06	-0.12	-0.15	-0.18	-0.04
n_places_airbnb			1	-0.15	-0.46	-0.55	-0.19	-0.02	-0.16	-0.17	-0.05	-0.12	-0.15	-0.18	-0.04
educ_center_proximity				1	0.2	0.25	0.31	0.21	0.29	0.03	0.07	0.02	0.11	0.05	0.04
year_built					1	0.55	0.3	0.24	0.32	0.04	0.09	0.12	0.32	0.2	0.02
km_to_center						1	0.4	0.41	0.5	0.06	0	0.05	0.29	0.44	0.11
metro_proximity							1	0.33	0.34	0.04	0.09	-0.04	0.13	-0.07	0.03
park_proximity								1	0.63	-0.04	0.03	0.02	0.2	0.28	0.06
health_center_proximity									1	0.02	0.08	0.06	0.27	0.26	0
n_room										1	0.61	0.14	0.02	0.01	0.05
n_bath											1	0.19	0.1	0.04	0.19
floors												1	0.41	0.06	0.15
n_dwelling													1	0.19	0.03
long														1	0.14
lat															1

Source: Own elaboration. Data retrieved from idealista18 R package, Open Data Portal of Madrid Council: parks and gardens, educational and medical health centers, Expansión.com: Historical Euribor 2018 and Datahippo.org: basic data Airbnb (Madrid).

- As expected, the highest correlation is found between number of rooms and number of bathrooms. No featuring engineering is going to be carried out because the interest is to see how each of them affects outcome variable separately.
- There is obviously a strong correlation between number of rooms and houses in Airbnb. This number of nearby Airbnb houses and rooms increases when the house is in the city center.
- Distance to parks, health centers and educational centers is also positively correlated, but it is logical since, as seen, there is a big number of all of this centers spread around the city, just isolated dwellings would have further or different distances from a center.

The presence of multicollinearity does not suppose a problem for this analysis because, as explained later, linear model will serve as a bench model and the rest, are advanced algorithms that handle this situations. Despite of this, it is important to notice which variables are correlated to each other.

4.2. Applied techniques

The goal was to find the model that minimizes the differences between observed and predicted values of the price per square meter of a house.

The decision was to split the data into a training (75%) and a testing (25%) set and use five-fold cross validation. With this method 30,914 records were for training, and the rest (10,303) for testing.

Cross validation technique splits each training fold into five subsets, four for training and one for validation. This process is repeated five times and the average results are then used for predicting in the testing set.

By using cross validation, the evaluation of performance of the model is more robust, as it is done in unseen data from the same dataset. It also ensures that the model is performing well in a real world scenario.

With the use of *Caret* package the implementation of cross validation can be easily done, as well as working with different methodologies for modelling. An empty dataframe was created to store the results.

The first modelling technique is linear regression. It serves as a baseline for building more complex models and be able to compare their performance. It models the linear relationship between the dependent variable, which is numeric continuous, and predictors. Its results are interpretable and the relationship between variables can be easily understood. Despite of this, as commented, there is a high degree of collinearity between some of the variables in the model. This means that they explain the same portion of variability. The following used algorithms are able to manage multicollinearity.

The incorporation of non-linear relationship and proximity-based predictions can be useful in the case of housing data. For this reason, more complex algorithms could predict better than the base model. KNN algorithm uses proximity of points in the feature space to predict a new data point.

Another machine learning algorithm tried is Random Forest. It combines multiple decision trees, which is suitable for capturing complex interactions among features. It is also useful to avoid overfitting.

Next and last trial is Extreme Gradient Boosting, where hyperparameters are tuned to find the optimal combination of the model.

4.3. Results

The objective in this section is to explain how each model behaves, comparing its performance with the rest. In addition, the key aspect of this is to identify which variables have more impact in the prediction, as well as in which measure they are related to others.

In the case of the simpler models, these interpretations are relatively easy to achieve, but once machine learning algorithms are implemented, the black box built in each of them makes difficult to acknowledge results.

It should be noted that time required to run models differ a lot. It is worth take into consideration this fact when selecting the best model. For this reason, time spent in running is also shown in the results table.

Table 4. *Modelling: results*

Algorithm	RMSE	R^2	MAE	Time to run
Linear	0.2890441	0.6501141	0.2231360	1s
KNN	0.2626817	0.7117831	0.1955255	25min
Random Forest	0.2026170	0.8279470	0.1442882	1h 30min
Gradient Boosting	0.2180361	0.8007673	0.1605267	5h 30 min

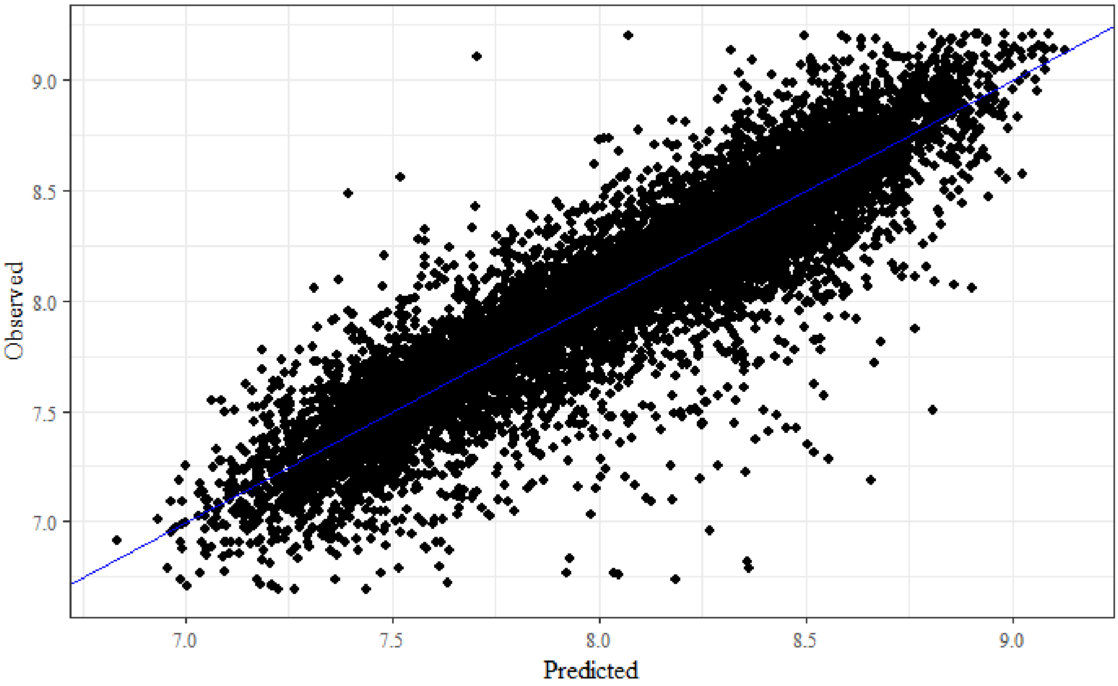
Source: Own elaboration.

As appreciated, algorithm that gives best results is Random Forest: Root Mean Square Error (RMSE) of Random Forest (RF) algorithm indicates that, on average, the predictions made by this regression differ from the real values of *log(price_sq_m)* on 0.2026170. Mean Absolute Error (MAE) is also the lowest from the four techniques.

R^2 is the highest for RF, which indicates that the model explains 82.79470% of the variability in the dependent variable. This measure is accounting for the ability of the model to explain, nevertheless, if a model does not explain, it cannot predict. So, considering this measure also for evaluation and comparison of models is necessary.

Evaluating also the time required to run, Random Forest is the one that gives better results and the time is not so long.

Figure 7. Random Forest results: observed vs. predicted

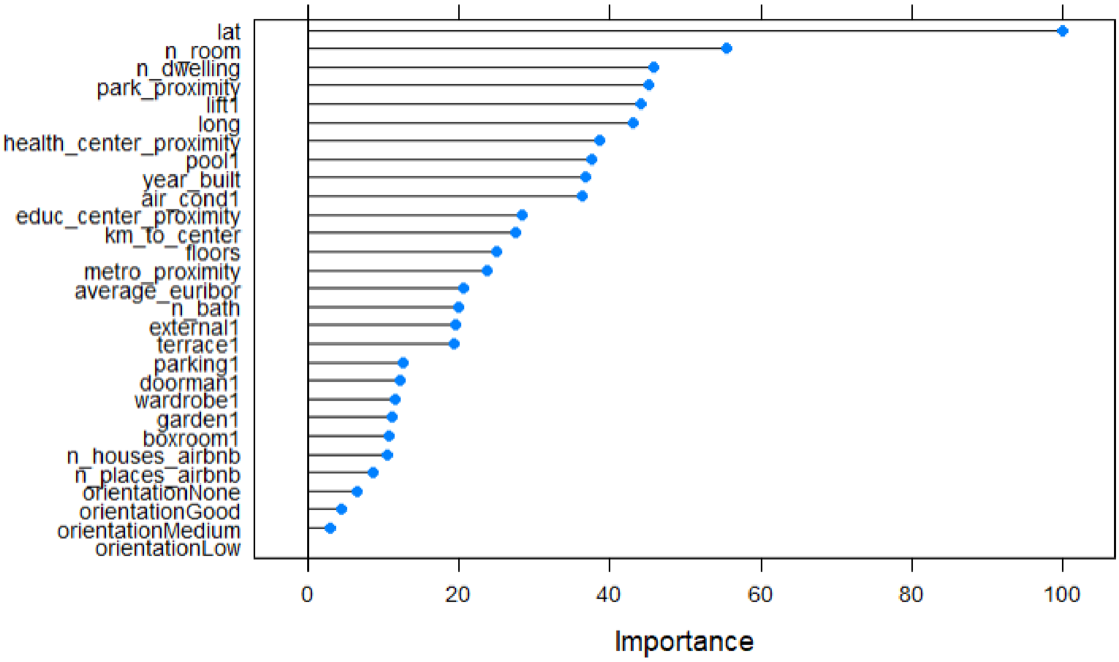


Source: Own elaboration

4.3.1. Variable importance for Random Forest algorithm

Variable importance plot below is showing how latitude is crucial in determining the price (concretely, logarithm of price per square meter) of houses in Madrid. This result aligns with the explanation given in first sections of this work about location dynamics and prices in the city (higher prices in the center and North). Number of rooms is also an important factor, as well as number of dwellings and proximity to a park.

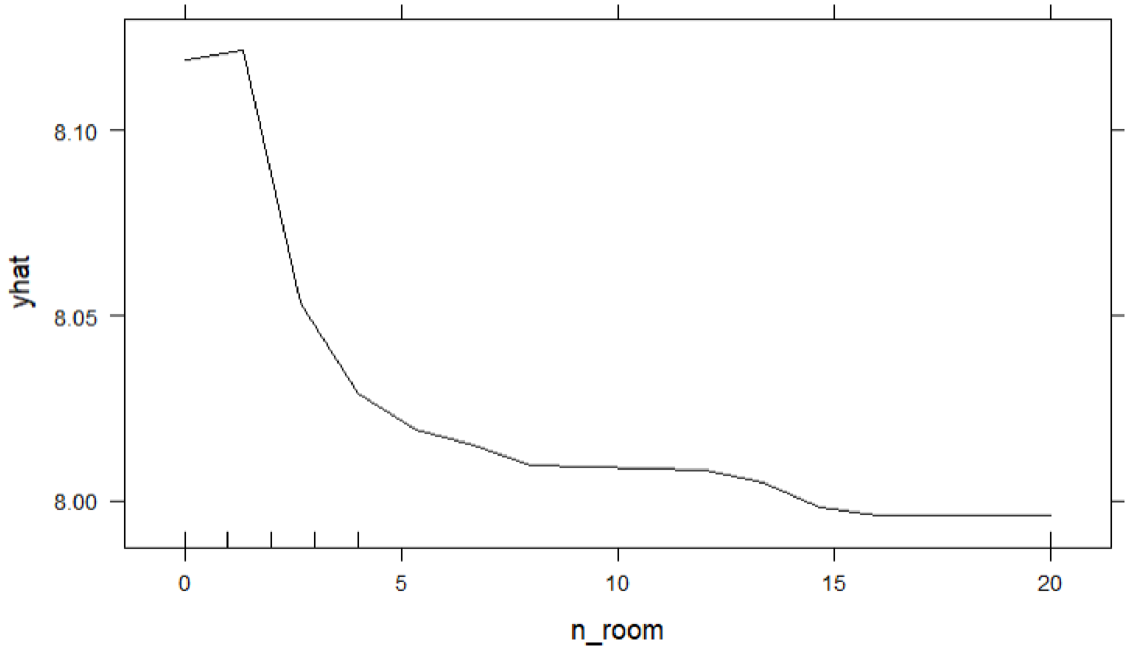
Figure 8. *Variable importance (Random Forest algorithm)*



Source: Own elaboration.

To find out the sign of the relationship between a variable and the outcome, Partial Dependent Plots (PDP) of each explanatory variable can be created. For example, for number of rooms the PDP shows that the logarithm of price per square meter is high when *n_room* is small, decreasing when the house has many rooms. This makes sense because, proportionally to the number of rooms, it is more expensive per square meter a house with 1 room than a house with 3 rooms.

Figure 9. *Partial Dependence Plot for number of rooms*



Source: Own elaboration.

5. CONCLUSIONS

Pricing dynamics of houses in Madrid city are well motivated by the location of the construction.

While public data availability is limited in the case of microdata (individual observations) for this topic of study, private sources in Spain capture almost all market information, extensively covering the Real Estate offer. Thus, the utilization of such private data is essential to develop a study of these characteristics.

Additional external variables have impact on the price of a dwelling. Factors such as proximity to facilities and areas of interest greatly influence prices. Provided this information by the public data bank, it is possible to incorporate these attributes into the models in a numerical way.

Integrating additional attributes into the models provides valuable information to stakeholders and researchers. It enables the development of more accurate predictive models that help decision-making for example, of investments.

To effectively predict housing prices, the implementation of machine learning algorithms has proven to offer better predictive power. However, it is important to note that as the complexity of the models increases, interpretability of results becomes more challenging.

REFERENCES

Bank of Spain (2020). The housing market in Spain: 2014-2019. Occasional Paper, (2013).

Bustos, V. (2022). La diferencia de precio entre una vivienda de dentro o fuera de la M30 de Madrid puede ser de casi 300.000 euros. El Español – Observatorio de la vivienda. Retrieved May 1, 2023, from https://www.elespanol.com/invertia/observatorios/vivienda/20221115/diferencia-precio-vivienda-dentro-m30-madrid-puede/718428444_0.html

Cimentada, J. (2023). Data Harvesting with R. Retrieved March 29, 2023, from https://cimentadaj.github.io/dataharvesting/

Crespí-Vallbona, Montserrat, & Domínguez-Pérez, Marta (2021). Las consecuencias de la turistificación en el centro de las grandes ciudades. El Caso de Madrid y Barcelona. Ciudad y Territorio Estudios Territoriales, LIII(2021 MONO). https://doi.org/10.37230/cytet.2021.m21.04

CSSLab UC3M (2023). Data Visualization: Principles and Practice. Retrieved April 4, 2023, from https://csslab.uc3m.es/dataviz/

Datahippo. (2018). Datahippo.org. Madrid (Provincia): Datos básicos Airbnb. Retrieved April 17, 2023, from https://datahippo.org/es/region/599230b08a46554edf88466b/

Expansion.com. (2019). Historical Euribor 2018. Datosmacro. Retrieved April 21, 2023, from https://datosmacro.expansion.com/hipotecas/euribor?anio=2018

Idealista. (2023). Sobre nosotros. Retrieved April 27, 2023, from https://www.idealista.com/sala-de-prensa/sobre-nosotros/

López, J. (2018, December). The economical accessibility of young people to the housing market in spain: a quantitative approach. *Metamorfosis*, (9), 154–163.

Medina M. (2023). R Programming. University Carlos III of Madrid.

Nogales J. (2023). Advanced Modelling - Regression: Home Price Prediction. University Carlos III of Madrid.

National Institute for Statistics in Spain. (2023). Housing Price Index. Retrieved March 15, 2023, from https://www.ine.es/jaxiT3/Datos.htm?t=25171

Open Data Portal of Madrid City Council. (2023). Educational centers in Madrid. Retrieved May 3, 2023, from https://datos.madrid.es/sites/v/index.jsp?vgnextoid=f14878a6d4556810VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD

Open Data Portal of Madrid City Council. (2023). Evolution of the price of second-hand housing (€ per square meter) by District and Neighbourhood. Retrieved April 2, 2023, from https://www-s.madrid.es/CSEBD_WBINTER/seleccionSerie.html?numSerie=0504030000202

Open Data Portal of Madrid City Council. (2021). Madrid municipal districts (geospatial data). Retrieved May 15, 2023, from https://www-s.madrid.es/CSEBD_WBINTER/seleccionSerie.html?numSerie=0504030000202

Open Data Portal of Madrid City Council. (2023). Main parks and municipal gardens in Madrid. Retrieved May 3, 2023, from https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=dc758935dde13410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default

Open Data Portal of Madrid City Council. (2023). Medical care centers in Madrid. Retrieved May 3, 2023, from https://datos.madrid.es/sites/v/index.jsp?vgnextoid=da7437ac37efb410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD

Rey-Blanco D., Arbues P., Lopez F., Paez A. (2021). idealista18: Idealista 2018 Data Package. R package version 0.1.1. URL: https://paezha.github.io/idealista18/

Rey-Blanco, D. *et al.* (2023). Using machine learning to identify spatial market segments. A reproducible study of major Spanish markets, *Environment and Planning B: Urban Analytics and City Science*, p. 239980832311669. doi:10.1177/23998083231166952.