Regression Analysis Report

Madrid Real Estate Prices

Work by:
Mohamed Khanafer and Aayush Kejriwal

Table of Content

For our analysis, we used **Excel**, **ScrapeStorm** and **Dataiku.**

Dataset

The dataset delivered with this report is the result of web scrapping, cleaning and preparation steps explained in the following sections.

Web scrapping

The web scrapper SpaceStorm was used for web scrapping on the website www.idealista.com.

Approach

The first approach chosen to gather data was to web scrap for the Madrid region and outer suburbs. The first dataset thus consisted of around 12.000 extracted listings in the 7 differerent regions shown in Figure 1.

Figure 1: First Web Scrapped Regions

After working on this extracted data, it became evident that the goal of the analysis was not aligned with the available data.

A second round of web scrapping then took place from which this time the web scrapping was focused on the various neighborhoods inside of Madrid. Even more specifically the focus was inside of the regions within the M-30 ring road surrounding the city as illustrated in Figure 2.

Figure 2: Final regions chosen for web scrapping

A total of 15.261 listings were thus extracted from Madrid Centro, Chamberí, Arganzuela, Retiro, Salamanca, Chamartín, Tetuán, and parts of Fuencarral and Moncloa on the 10th and 11th of December.
And this is the data attached and used in this report.

Issues faced

The first problem was the number of listings Idealista allows anyone to access. Even if the website claims they have more than 60.000 listings, one can only access those by regions and each region has a maximum number of 1.800 listings that can be accessed. To go around this, we refined our search to smaller regions within the regions wanted and later on merged all the data gathered together. In this way we had access to more listings and were able to

extraxt more than 95% of the listings available in the region defined in Figure 2. The second faced issue when web scrapping was going around the Captchas set up by the website. Usuallay after extracting 500 listings, the web scrapper would be stopped and would have to start from the beginning. However, after several trials, we noticed that the place of the listings on Idealista were not changed, we could thus resume the webscraping by starting on the page in which the Captach was faced.
This is the procedure we followed to gather the raw data for the 9 mentionned neighborhoods. To this report is attached the raw data gathered as well as the cleaned data that was the results of the cleaning and preparations steps explained in the next section.

Description of Raw Data

The data we got after the web-scrapping contains 15,261 rows, containing information about different houses in the central Madrid region. This comprises of the following 9 areas, as shown in the column "Subzone" – Centro, Salamanca, Chamberri, Tetuan, Chamartin, Fuencarral, Arganzuela, Retiro and Moncloa.

In addition to the "Subzone" column, we have 40 more columns. Following are the list of columns with a description of the information it contains:

1. **Title:** A brief description of the type of property and location. For example, Piso en Goya, Madrid.
2. **Link:** The web address to the property on Idealista. For Example, https://www.idealista.com/inmueble/83788335/
3. **Price:** The current price of the property. In case there is a discount on the property, it shows the discounted price.
4. **Ellipsis:** Some more description of the agency and property that we will be removing in the cleaning stage.
5. **Phone:** Phone number of the agency selling the house.
6. **Square:** The constructed area of the property in square meters.
7. **Floor:** The floor number of the building in which the property is, and also whether it is an exterior/interior property and has an elevator or not.
8. **Discount:** The original, non-discounted price of the property for those properties that have a discount.
9. **Garage:** Whether or not a garage is included in the price of the property.
10. **Rooms:** The number of bedrooms.
11. **Headline2:** Same information as the column "Title".
12. **Location:** The area in which the property lies. Here the area mentioned is more detailed than the "Subzone" column mentioned earlier, and therefore takes a wider set of values.

13. **Price2, Square2, Rooms2, Floor2 and Garage2 (5 columns):** Same information as the respective columns already mentioned. The reason for having these duplicates if that we had information from the main page of idealista as well as from the individual property pages, and these columns are mentioned at both places.
14. **Characteristic Start**: A column containing only the value “Características Básicas” or NULLS.
15. **Field 12 to Field 20, and Col_27 to Col_30 (13 Columns):** Basic Characteristics of the property, such as area, number of bedrooms, number of bathrooms, construction date, etc in a jumbled format. No column is specific to one type of information.
16. **Edif/Equip**: Another useless column like “Characteristic Start” that marks the start of the equipments included in the property.
17. **Field 22 to Field 26, and Col_37 (6 Columns):** The list of equipments included in the house in a jumbled format. Again, no column represents the availability of a specific equipment.
18. **Full Column Char**: All the basic characteristics in a single sentence.
19. **Full Column Ed**: All the equipments in a single sentence.
20. **Subzone**: The location of the property in a more granular level than the column “Location”, as discussed earlier.

(Since the data was very unclean and values were mixed up in different columns, there were certain rows where the cell value did not match the description of the respective column as mentioned above)

Data Preparation

The steps taken to prepare the data are:

1. Split the column “Link” on the basis of ‘/’ and only kept the column with the number code to act as a primary key. This can later help to check whether we have included the same property multiple times by mistake. The numeric column we decided to keep was named “PK” and was converted to a text format despite being a number.

2. Split the column “Title” on the basis of a space, ‘ ’, and kept only the first of the resulting columns. This output column contains the information of property type and was therefore named “Property_Type”.

3. Performed a currency split on “Price” to obtain only the price as a numeric value. Named the resulting column “Current_Price” and deleted the old column and the column containing the currency.

4. Deleted the column “Location” because it is very specific and has way too many possible categories. Information such as location can impact house prices, but we chose to take that impact on a more granular level using “Subzone”. Renamed “Subzone” to “Location”.

5. Deleted the following columns: “Link”, “Title”, Headline2”, “Ellipsis”, “Phone”, “Discount”, “Price2”, “Square2”, “Rooms2”, “Floor2”, “Garage2”, “Characteristic Start”, “Edif/Equip”.

6. These steps mentioned above were done on DataIku. Now, to make meaningful columns from the original data that had different details mixed up different columns, we used the INDEX – MATCH set of formulae on excel, based on certain key words associated with each type of detail. Using this formula, we created 28 new columns. These columns were meaningful since each of them corresponded to one piece of information or detail about the property, such as number of bathrooms. However, due to the nature of the data and the INDEX – MATCH formula, some of the rows contained information in addition to the one required in that column.

   The 27 new columns contained the following information:

   Property_Subtype, Built_Area, Plot_Area, No._of_Floors, Floor_Number, Ground_Floor, Basement, Interior, Exterior, No_Elevator, Yes_Elevator, No._Bedrooms, No._Bathrooms, Garage, AC, Terrace, Balcony, SwimmingPool, Yard, NearGreenZone, Storeroom, Wardrobe, New_Construction, Second_Hand, Heating_Source, Orientation, Construction_Date.

7. Deleted all columns except these 28 new columns, and “PK”, “Current_Price”, “Property_Type” and “Location”. Therefore, in the next step, we had a total of 32 columns.

8. Converted all cells with any value in the column “Interior” to a 0 and left the NULL values as is. Converted all the cells with any value in the column “Exterior” to a 1 and left the NULL values as is. Added the values of the 2 columns to get the column “Is_Exterior” with values of 1, 0 or NULL. Deleted the 2 columns “Interior” and “Exterior”. Cells with NULL values were considered the same as being not exterior, because from a marketing perspective, if it was exterior, the owner would mention it to increase its attractiveness.

9. Followed the same step as above with “No_Elevator” and “Yes_Elevator” to get “Has_Elevator”.

10. Converted the values in the columns “No._Bedrooms” and “No._Bathrooms” to just numeric. For example, changed “4 habitacion” to 4.

11. Converted 9 columns to Boolean by replacing cells that contained values with one and replacing NULL cells with 0. These 9 columns are “Garage”, “AC”, “Terrace”, “Balcony”, “SwimmingPool”, “Yard”, “NearGreenZone”, “Storeroom”, and “Wardrobe”.

12. Extracted only the numeric part of the column “No._Of_Floors. For examply, “3 plantas” was converted to 3.

13. Converted all cells in the column “GroundFloor” that had a value to 0 and left the NULL values as is. This signifies the floor number being 0. Similarly converted all cells in the column “Basement” to -1.

14. Split the column “Floor_No.” on the basis of a space and retained only the first column. Since the values signifying the floor number had a “ª” suffix, for example 2nd floor was 2ª, we retained cells having this format. We realised cells containing just the number, like 2, did not imply 2nd floor but were actually the 2 floors from the “No._Of_Floors” column. Therefore, we deleted all the values without the “ª” format to get NULL values. Then we removed the “ª” from the floor number to be able to convert it to a numerical format. Lastly, we created a new column that added the values of this column with those in “Basement” and “GroundFloor” to combine all similar data. This new column replaced these 3 columns as the single column, also called “Floor_No.”.

15. Extracted only the year value from construction date by splitting and removing the text “construction en”.

16. Split and retained only the numeric value from the columns “Plot_Area” and “Built_Area” in a similar manner as above.

17. Used the values in columns “New_Construction” and “Second_Hand” to form a new column called “Property_Quality”, containing 3 possible values of “Promoción de obra nueva”, “Segunda mano/para reformar” and “Segunda mano/buen estado”. Deleted the two original columns used.

18. On Excel, used a combination of IF conditions and INDEX – MATCH conditions to extract only the relevant information from the columns “Heating_Source” and “Orientation”.This essentially removed the remaining parts of the sentence in the few cells that contained entire sentence descriptions as opposed to just the possible categorical values of the respective columns.

19. As of this stage, without deleting any rows, we have 15,261 rows. But some of them may be duplicates from the web scrapping stage. Therefore, we need to delete these duplicates before moving to the next step. After removing duplicates, we are left with 13,958 rows.

20. We no longer need the PK column and can delete it.

21. We split the data into 2 sets on the basis of property type. Casa and Chalet were grouped as type A while Piso, Atico, Duplex and Estudio were grouped as type B. The reason for this is that certain information categories, such as Plot_area and No._Of_Floors are specific to type A properties and Floor_No., Has_Elevator, etc. are specific to type B properties. By splitting the data, the regression model may be more accurate as we are no longer using variables that are only useful to some properties and not the others. Since property type is important in defining our data, we deleted the 45 rows with NULL values in the column.

22. Since only 732 rows belong to group A, and 13,181 rows belong to group B, we decided to focus our model on the latter group. This is the group with pisos, duplexes, aticos and estudios. Pisos form a large part of the dataset, accounting for 11,649 rows.

23. The following columns are not relevant since they have either only NULL values, or in the case of Boolean variables are all 0. They were therefore deleted.

    "Property_Subtype", "Plot_Area", "No._Of_Floors", "Yard"

24. Since nearly 70% of the data in "Construction_Date" was missing, we decided to delete the column.

25. Created a new column called "Bedrooms_per_Bathroom", since the ratio may have. A bearing on price in addition to just the absolute number. It is our assumption at this stage that having fewer bedrooms sharing a bathroom, or having a bathroom for each bedroom, would be preferred over having many people share a bathroom and therefore would command a relatively higher price.

26. Created 4 new Boolean columns for each property type – piso, atico, duplex and estudio. Did the same with the 9 categories under location.

27. Created a new column "Quality_Rating" by converting the 3 categories in the "Property_Quality" into an ordinal variable. New constructions were given the highest score of 3, secondhand buildings in good construction were gived a 2 and those needing repair were given a 1. NULL values were imputed with a 2 because it is a middle value and also the most recurring value, accounting for 9008 of the 13181 values.

28. Created 2 columns "Central_Heating" and "Individual_Heating" as Boolean variables from "Heating_Source". Did not create a column for the cases where there was no heating since we can have these cases as the benchmark against which we study the impact of the 2 types of heating.

29. Deleted the 61 rows where Built_Area was NULL. It is our assumption that built area will play an important role in our model, and since a very small proportion of our data has missing values in this, we think deleting these rows will not cause a big loss for our model.

30. Now our data contains 37 columns and 13,120 rows. The columns are:

    Current_Price, Property_Type, 4 columns for each property type as a Boolean, Location, 9 columns for each location as a Boolean, Built_Area, Bedrooms, Bathrooms, Bedrooms_per_Bathroom, Floor_No., Is_Extertior, Has_Elevator, Garage, AC, Terrace, Balcony, SwimmingPool, NearGreenZone, StoreRoom, Wardrobe, Property_Quality, Quality_Rating, Heating_Source, Central_Heating, Individual_Heating, Orientation.

EDA – Describing the final dataset

Outlier Detection

Most of our data is in a Boolean format. Even those variables that were categorical have been either converted to Boolean or to an ordinal numeric variable with few possible inputs.

As such, only 6 of our columns that are numeric have a possibility of having values whose distribution needs to be studied. These columns are: Current_Price, Built_Area, Bedrooms, Bathrooms, Bedrooms_per_Bathroom and Floor_No.

1. **Current_Price:** The mean price of the properties in our data is 757,516 euros. However, prices take on a large range of value, starting from 55,000 to 8,000,000. The distribution is right skewed, with 1215 outliers based on 1.5 IQR.
2. **Built_Area:** The mean area 137 square meters, with values ranging from 12 to 894. The distribution is right skewed again, with 800 outliers.
3. **Bedrooms:** The houses have bedrooms ranging from 0 to 16. The data has a mean of 2.8. At first, one striking value was of 0, which looked like an error that we should delete. However, looking at other fields for the houses with 0 bedrooms showed that they are studio apartments that do not a have a separate living room and bedroom. Again, we have 82 houses that can be considered as outliers due to a high number of bedrooms.
4. **Bathrooms:** The number of bathrooms ranged from 1 to 15, with an average of 2. 1337 houses have too many bathrooms based on our outlier detection.
5. **Bedrooms_per_Bathroom:** These values also ranged from 1 to 15. However, in this case the occurrence of high values above 4 or 5 are even more suspicious. We decided to delete rows in which this ratio was greater than 5. We therefore deleted 12 rows that met these criteria.
6. **Floor_No.:** Appartments were placed anywhere from a basement to the 23rd floor in some buildings. Most of them however were on the 2nd and 3rd floors.

The linear regression model is sensitive to outliers, and we thought of deleting more rows based on the above-mentioned outliers. However, studying the outlier values more closely, we noticed that houses that were an outlier based on one of the variables also tended to be outliers in another variable. This can be further understood by the correlation coefficient in the following section. Due to this behaviour, we think the presence of the outlier would not actually impact the slope of the regression line and can therefore be retained in out dataset.

Correlation Matrix

Based on our correlation matrix that can be found below, we feel the size of the house is going to be the most important factor in prediction prices. The number of bedrooms and number of bathrooms are also quite highly correlated with prices, but they also show high correlation

with the built area. This could lead to a problem of multi-collinearity in which case we might need to drop one of those variables. But we cannot decide to exclude them yet as it could lead to bias in the other coefficient.

The other variables are not very strongly correlated with either the price or among themselves, and therefore will be added to the model based on their impact on the R2.

	Current_Price	Built_Area	Bedrooms	Bathrooms	Bedrooms_per_Bathroom	Floor_No.	Is_Exterior	Has_Elevator	Garage	
Current_Price	1									
Built_Area	0.86	1								
Bedrooms	0.55	0.71	1							
Bathrooms	0.75	0.81	0.69	1						
Bedrooms_per_Bathroom	-0.16	-0.06	0.40	-0.28	1					
Floor_No.	0.17	0.17	0.16	0.14	0.02	1				
Is_Exterior	0.22	0.27	0.25	0.25	0.04	0.14	1			
Has_Elevator	0.25	0.26	0.23	0.27	-0.04	0.21	0.20	1		
Garage	0.27	0.35	0.25	0.33	-0.10	0.16	0.18	0.20	1	
AC	0.13	0.08	-0.05	0.15	-0.22	0.09	0.09	0.09	0.10	
Terrace	0.14	0.21	0.21	0.18	0.04	0.22	0.17	0.16	0.24	
Balcony	0.11	0.08	0.10	0.08	0.05	-0.03	0.18	-0.01	-0.08	
SwimmingPool	0.04	0.09	0.03	0.11	-0.10	0.07	0.12	0.15	0.41	
NearGreenZone	0.03	0.10	0.09	0.11	-0.03	0.07	0.13	0.14	0.37	
StoreRoom	0.24	0.29	0.20	0.26	-0.07	0.05	0.19	0.16	0.31	
Wardrobe	0.06	0.06	0.04	0.11	-0.09	0.10	0.14	0.15	0.10	
Quality_Rating	-0.06	-0.14	-0.27	-0.05	-0.24	-0.08	0.02	0.01	0.04	
Individual_Heating	-0.23	-0.26	-0.23	-0.24	0.02	-0.14	-0.07	-0.17	-0.17	
Central_Heating	0.10	0.15	0.21	0.13	0.08	0.15	0.09	0.20	0.01	

	AC	Terrace	Balcony	SwimmingPool	NearGreenZone	StoreRoom	Wardrobe	Quality_Rating	Individual_Heating	Central_Heating
AC	1									
Terrace	0.04	1								
Balcony	0.09	-0.10	1							
SwimmingPool	0.07	0.16	-0.09	1						
NearGreenZone	0.10	0.19	-0.04	0.57	1					
StoreRoom	0.07	0.12	0.01	0.25	0.19	1				
Wardrobe	0.30	0.09	0.10	0.04	0.09	0.13	1			
Quality_Rating	0.17	-0.05	-0.03	0.23	0.06	0.07	0.05	1		
Individual_Heating	0.07	-0.15	0.10	-0.05	-0.04	-0.09	0.07	0.05	1	
Central_Heating	0.02	0.11	-0.06	-0.07	0.00	0.04	0.13	-0.17	-0.37	1

Regression Report

Building the Model

In order to study not only the final impact of all variables on the target, but also to better understand how the independent variables interact with each other, we tried to build our model by changing only a few variables at a time. This way we could see the change in the coefficients of the variables when we added new variables, which explained the presence of bias.

Following is a table showing all our models, in columns 1 to 14, and the coefficient of the variables included in the model. Empty cells indicate that the specific variable was not included in the model and cells marked in light red indicate that the given coefficient was not significant at a 95% confidence level.

	1	2	3	4	5	6	8	9	10	11	12	13	14
R2	0.735	0.758	0.76	0.761	0.761	0.761	0.77	0.77	0.771	0.772	0.797	0.797	0.797
Intercept	-125981	-117989	-219650	-245528	-233073	-254104	-264421	-348252	-329022	-312529	-344406	-343474	-350445.15
Built_Area	6436	6023	6051	6024	6038	6029	6055	6078	6062	6089	5845	5859	5849.5
Bedrooms		-87715	-132505	-134701	-134617	-135512	-128303	-122297	-120640	-121633	-115618	-114981	-115455.16
Bathrooms		146609	199136	200181	201317	200310	193130	187913	186494	185924	175627	175748	177058.21
Bedrooms_per_Bathroom			73882	75240	76193	77208	67793	66543	66589	62023	76414	75285	78022.15
Floor_No.				10651	11029	10156	12636	13210	13400	11960	13378	13752	13560.37
Is_Exterior					-24044	-28106	-24627	-27489	-26035	-26087	-840		
Has_Elevator						36972	61300	60040	64044	60542	12147		
Garage							-13209	-13067	-18392	-17029	9440		
AC							36575	31531	34323	35566	25172	24994	25132.61
Terrace							-37877	-36619	-37077	-41920	-20178	-19004	-20300.16
Balcony							72176	72703	70871	69462	49107	48215	50599.64
SwimmingPool							-56757	-73701	-77649	-77365	-22423	-18406	
NearGreenZone							-109209	-103111	-101437	-97361	-49472	-47693	-57824.4
StoreRoom							2519	141	491	-2029	13530	15155	14024.45
Wardrobe							-21508	-21444	-14931	-14815	-19583	-19429	-18773
Quality_Rating								44803	40157	39775	53996	54689	52808.55
Individual_Heating									-26558	-26868	-6167		
Central_Heating									-44560	-45269	-48281	-45326	-45247.18
Type Duplex										-123633	-98065	-99233	-98508.94
Type Atico										61232	47207	46580	47127.32
Type Estudio										-51727	-20451	-21804	
Location Moncloa											-190685	-189004	-184344.44
Location Arganzuela											-62042	-59219	-55692.46
Location Retiro											32877	36869	42385.07
Location Fuencerral											-169162	-166784	-164956.79
Location Chamartin											-19601	-15194	
Location Tetuan											-113231	-111635	-106433.81
Location Chamberri											36378	39309	44368.52
Location Salamanca											258879	262274	268267.41

The model we finally selected was the one in column 14. This model includes as many variables from our dataset as needed, while removing those that were insignificant. Following is a more detailed look at the coefficients and the significance levels of each variable.

Variable	Coefficient	Std. Err	T stat	p-value	Confidence
Wardrobe	-18,730.4277	6,518.7976	-2.8733	0.0020	★★☆
Tetuan	-106,433.8074	10,192.0828	-10.4428	< 1e-4	★★★
Terrace	-20,300.1573	6,933.0475	-2.9280	0.0017	★★☆
StoreRoom	14,024.4532	7,031.0632	1.9946	0.0231	★☆☆
Salamanca	268,267.4112	9,063.8927	29.5974	< 1e-4	★★★
Retiro	42,385.0744	12,271.9130	3.4538	0.0003	★★★
Quality_Rating	52,808.5472	6,580.2483	8.0253	< 1e-4	★★★
NearGreenZone	-57,824.3968	9,741.7239	-5.9357	< 1e-4	★★★
Moncloa	-184,344.4390	18,356.1698	-10.0426	< 1e-4	★★★
Fuencerral	-164,956.7865	12,768.6437	-12.9189	< 1e-4	★★★
Floor_No.	13,560.3789	1,344.2692	10.0875	< 1e-4	★★★
Duplex	-98,508.9447	16,584.5815	-5.9398	< 1e-4	★★★
Chamberri	44,368.5238	9,378.6511	4.7308	< 1e-4	★★★
Central_Heating	-45,247.1767	7,561.5766	-5.9838	< 1e-4	★★★
Built_Area	5,849.5083	54.8851	106.5773	< 1e-4	★★★
Bedrooms_per_Bathroom	78,022.1517	7,592.2240	10.2766	< 1e-4	★★★
Bedrooms	-115,455.1658	5,615.3639	-20.5606	< 1e-4	★★★
Bathrooms	177,058.2138	7,158.0555	24.7355	< 1e-4	★★★
Balcony	50,599.6361	7,533.0241	6.7170	< 1e-4	★★★
Atico	47,127.3201	13,670.6015	3.4473	0.0003	★★★
Arganzuela	-55,692.4654	12,644.8716	-4.4044	< 1e-4	★★★
AC	25,132.6097	6,444.0004	3.9002	< 1e-4	★★★
Intercept	-350,445.1566	350,445.1566	-1.0000		

Explaining the Coefficients

Looking at the coefficients can tell us the impact that a particular variable has on the price of a house when all else is constant. Here we will try to understand each of the above coefficients from our final model.

The coefficient for Built_Area implies that, with all else constant, an additional square meter of space in a property will increase its price on average by 5,849.5 euros.

Similarly, and additional bedroom will actually decrease the price by 115,455.16 euros. This one is surprising because one would expect more bedrooms to increase the price, and even the correlation matrix would suggest that. But in our case, since we have built area as a variable in the model and the assumption is that it is held constant, the impact of bedrooms is not as straightforward. The reason is that there is a connection even between built area and bedrooms. As there are more bedrooms, the area would increase. Therefore the positive impact of having more bedrooms is already captured by the area. One way to think of the negative impact of bedrooms on price is to think that since the area is constant, we get smaller rooms by increasing the bedrooms, which is undesirable.

Having more bathrooms, on the other hand, does increase the price of the house by 177,058 euros per bathroom. The coefficient for bedrooms_per_bathroom is a positive 78,022. This is also not the most expected outcome as one would think a lower ratio would be better. But once again, since we have other variables here that are interlinked, the relationship is not straightforward. The reason for it being a positive relation is possibly the same as the one for the bedrooms being negative. Since area must be fixed, either the number of bathrooms must reduce or the number of bedrooms must increase, for the ratio to increase. Both, a decrease in number of bathrooms and an increase in number of bedrooms is already captured in their respective coefficients.

The placement of the house on a higher floor would likely increase the price on average by 13,560 euros. A house on the ground floor would not be impacted by this coefficient whereas a house in the basement would actually see a drop in price by the same amount.

The availability of an elevator, a garage or a swimming pool, or the house being exterior, does not significantly impact the price of the house, as seen by their p-value in the previous model. Again, it is important to note that this does not necessarily mean that these factors are not important, as it is possible that the reason for this p-value is the presence of another variable that is highly linked to both, the price of the house and these excluded variables. Therefore, the coefficient of those variables captures the impact of the presence of these factors as well. In our case, including the location details in model 12 and 13 caused these variables to lose significance, indicating that these features are not evenly distributed among all regions. For example, it may be that swimming pools are no longer a significant variable because the distribution of swimming pools may lie heavily in only one of the regions, say Retiro, and therefore the coefficient of Retiro includes the impact of having a swimming pool, making it biased.

Similarly, the coefficients of all other Boolean variables, which are AC, Terrace, Balcony, being near a green zone, storeroom and wardrobe, indicate the monetary gain or loss in the value of a house due to the presence of the respective variable. Again, this impact assumes all other variables included in the model being held constant.

When reading the coefficients of the property type and location variables, we must understand the point of reference to which that coefficient corresponds. For example, for the property type of Atico, the coefficient does not mean that if the house is an Atico it costs 47,127 euros more on average. Such a statement would not make sense because it lacks a reference. In our case that reference is the property type of Piso. Therefore, it is correct to say that all else being equal, a house of the type Atico costs 47,127 euros more than a Piso. And similarly, a Duplex cost on average 98,508 euros less than a Piso.

The same concept can be applied while reading the coefficients of the locations. The reference location in our model is Centro Madrid. Therefore, a house in Salamance, for example, costs 268,267 euros more on average than a house in Centro, when all other variables are the same.

Global Results

Explained Variance Score Best possible score is 1.0, lower values are worse	0.79659
Mean Absolute Error (MAE) Average of the absolute value of the regression error	1.9389e+5
Mean Average Percentage Error Average of the absolute value of the regression error	30.7%
Mean Squared Error (MSE) Average of the squares of the errors	1.1203e+11
Root Mean Squared Error (RMSE) Root of the above mesure	3.3471e+5
Root Mean Squared Logarithmic Error (RMSLE) Root of the average of the squares of the natural log of the regression error	-
Pearson coefficient Correlation coefficient between actual and predicted values. +1 = perfect correlation, 0 = no correlation, -1 = perfect anti-correlation	0.89252
R2 Score (Coefficient of determination) regression score function	0.79659

Above is a summary of the results of our regression model as a whole. The first value is the explained variance score, which is the same as the R2 score. This number tells us the percent of variance in the target variable that is explained by the model. Here, the number 0.79659 tells us that 79.659% of the variance in house prices in Madrid can be explained using the variables we used. It is important to note that this number is highly influenced by the data, and using the same model on a different set of houses may result in a lower score. This can happen when the model has overfit the data available.

Another potential problem with using the R2 alone to judge the model is that the formula to calculate it always rewards the use of additional variables. Therefore, it can be easy to get a high score by using many variables that might or might not actually be meaningful.

Another interesting value in the above table, that is also linked to the R2, is the Pearson coefficient. This is the correlation coefficient between actual house prices in our data and our estimated prices for the same houses. Taking the square of the Pearson coefficient gives us the R2.

The other set of results are related to the error. The error for each house is the difference between its actual and predicted price. One of the properties of a regression is that the sum of all errors is always 0. Therefore that would not be a very useful metric to judge the model on. The MAE instead takes the absolute values of the errors and then averages them. Here, the MAE is 193,890 euros. It means that on average the model is off the true prices by 193,890 euros. The next value, 30.7%, states the same thing as a percentage of average prices.

MSE deals with the problem of errors cancelling out in a different way. Instead of taking the absolute values, it takes the squares of the errors and averages them out. This results in larger errors being penalised even greater than smaller errors. Since this is the metric that the linear regression model tries to optimise in an OLS, there is a trade-off we need to be aware about. Due to the fact that larger errors are penalised more than smaller errors, the model prefers having many small errors over having few large errors. This can be a problem when we have outliers or extreme values because the model becomes very sensitive to them.

Lastly, we have the RMSLE. Since the MSE is in squared units, the error cannot be directly compared to the target variable like the MAE can be. By taking the root of the MSE in RMSLE, this problem is solved. So once again, the value of 334710 tells us that on average, our model´s predictions differ from the reality by 334,710 euros.

An interesting thing to note is that the MAE and RMSLE can be directly compared as they are in the same units and try to give a similar understanding of the model. And yet, the values of these metrics are significantly different from each other. The reason for this is once again the heavy penalty on large errors while calculating the MSE, which flows down to the RMSLE as well increasing its value.

Results User Guide

Our model could be used in two different business scenarios. The first one is when an investor is looking for underprices houses according to the model. Here, if the actual price of the house is lower than the predicted price, it would signal a good ooportunity for the investor, and vice versa. The second scenario would be if an investor is looking to build a certain building with given characterisitcs. He could use the model to predict the selling price of this appartement given its features.
Here are the 2 scenarios explained through an example.

Looking for good investments

In the dataset attached, a randomly picked apartment exhibit the following characteristics:
The listing is an Atico located in Tetuan, with a built area of 257 square meters, 3 bedrooms, 4 bathrooms, located on the 10th floor of the building with an exterior view. Furthermore, it has Air Conditionning, a terrace, a storeroom and a wardrobe. It was rated as 2 out of 3 on the quality index. The listing is up for sale for 1,475,000 Euros.
Using the model mentioned above, we would have a predicted price of:

	Model	Appartment	Amount
Intercept	-350445.15		-350445.15
Built_Area	5849.5	257	1503321.5
Bedrooms	-115455.16	3	-346365.48
Bathrooms	177058.21	4	708232.84
Bedrooms_per_Bathroom	78022.15	1	78022.15
Floor_No.	13560.37	10	135603.7
AC	25132.61	Yes	25132.61
Terrace	-20300.16	Yes	-20300.16
Balcony	50599.64	No	0
NearGreenZone	-57824.4	No	0
StoreRoom	14024.45	Yes	14024.45
Wardrobe	-18773	Yes	-18773
Quality_Rating	52808.55	2	105617.1
Central_Heating	-45247.18	No	0
Type Duplex	-98508.94	No	0
Type Atico	47127.32	Yes	47127.32
Location Moncloa	-184344.44	No	0
Location Arganzuela	-55692.46	No	0
Location Retiro	42385.07	No	0
Location Fuencerral	-164956.79	No	0
Location Tetuan	-106433.81	Yes	-106433.81
Location Chamberri	44368.52	No	0
Location Salamanca	268267.41	No	0
			1,774,764

Interpretation of the results:

Our model estimates that a listing exhibiting such characteristics should be selling for 1,774,764 Euros, this listing is thus underpriced according to our model and could be an interesting investment.

All other listings could be assessed in this manner.

Estimating the price of a listing

Although the second scenario has a different goal, the approach to using the model is similar. The user would input the characteristics in the same way as in the above example and he/she would have an estimation of the price the future listing can be sold for.