uc3m | Universidad Carlos III de Madrid

Master Degree in Statistics for Data Science
2024-2025

Multivariate Analysis

Modelling Competition: House Pricing 2024

Taylor Katherine Stone - Paola Carolina Suarez - Connor George Stratford - Álvaro Restoy

Professor: María Durban
Madrid-Puerta de Toledo, 13 December 2024

Contents

1 Introduction

The objective of this project is to develop a reliable model that predicts the sale price per square meter of residential properties in Madrid, identify key attributes that influence housing values, and validate the model's accuracy on unseen data. The dataset contains around 1,000 residential properties and includes 35 predictors related to house characteristics, surroundings, and districts, as well as a target variable: the sale price per square meter. The process begins with an exploratory data analysis (EDA) to understand the data's structure and uncover meaningful relationships between house features, neighborhood characteristics, and prices. Using domain knowledge and data-driven insights, we refine variables, reduce categories where appropriate, create composite features, and include meaningful interactions. We iteratively build and refine a regression model, focusing on metrics such as Adjusted R-Squared, AIC, and RMSE, to ensure accuracy and simplicity. This involves diagnostic measures such as identifying the most influential variables, explaining the characteristics that define both the cheapest and most expensive properties, and validating our final model against the test set.

Research on Madrid's housing market reflects a complex interplay of factors affecting accessibility, desirability, and price. Wealthier neighborhoods like Salamanca, Centro, and Retiro are expensive due to their prime locations and socioeconomic status. In contrast, more affordable districts such as Moratalaz, Vallecas, and Carabanchel cater to middle- and lower-income families. Proximity to the M-30 ring road increases prices due to better access to city amenities but often brings higher noise and pollution. Central areas can also elevate the value of smaller properties, such as studio apartments, due to their location. Conversely, older buildings lacking amenities like elevators may lower prices, while larger homes with more rooms and bathrooms generally fetch higher values. Families with limited budgets, immigrants, and younger buyers tend to reside in more affordable, peripheral districts, whereas older and wealthier residents can afford prime, centrally located properties. Taken together, location, building characteristics, environmental quality, and demographic composition determine a property's value. Properties closer to the city center or within the M-30 usually sell at higher prices, as do those with amenities such as elevators and commercial areas nearby. Larger homes with more rooms, lower pollution levels, and attractive green spaces appeal to buyers, while wealthier districts with older populations generally command higher prices. In summary, the project bridges data analytics with an understanding of Madrid's social, economic, and environmental landscape, resulting in a robust model that both predicts prices effectively and explains what drives property values.

2 Exploratory Data Analysis

The dataset contains 736 observations and 35 variables that describe houses in Madrid, with both property attributes and environmental context influencing property prices. The target variable, the logarithm of property prices (log_price), ranges from 7.139 to 9.424, with a median value of 8.176. Key numerical attributes include constructed area (sup.const), which ranges from 20 to 875 m^2, with a median of 80 m^2, and usable area (sup.util), which ranges from 18 to 680 m^2, with a median of 70 m^2. The number of bedrooms (dorm) varies from 0 to 7, with most properties having 2 to 3 bedrooms, while the number of bathrooms (banos) ranges from 1 to 7, with a median of 1 and most properties having no more than 2 bathrooms.

From our initial analysis, we observed that the distribution of precio.house.m2 was positively skewed. To address this, we applied a log transformation to the variable. After the transformation, the distribution became closer to normal, reducing skewness significantly. This adjustment helps satisfy the assumption of normality for the response variable, which is important for the validity and interpretability of linear regression models.

Categorical variables provide additional insights into property characteristics. Most properties are pisos (apartments), which account for 86% of the observations, followed by áticos and estudios. In terms of condition (estado), 76% of properties are in buen_estado, with fewer categorized as a_reformar or reformado. The presence of an elevator (ascensor) is noted in 66% of properties, while the remaining properties do not have one. Neighborhood and locational data further enrich the dataset. The properties are distributed across 127 unique neighborhoods, with the highest counts in Casco Histórico de Vallecas (20) and Embajadores (20). The dataset spans 21 districts, with the largest numbers of properties located in Centro (62), Arganzuela (57), and Puente de Vallecas (54). Environmental factors

provide additional context, such as external noise levels (Ruidos_ext), which range from 0.12 to 0.64, with a mean of 0.397, and bad odors (Mal_olor), which have a mean value of 0.283 and typically do not exceed 0.33. Air quality indices, such as SO2 and PM10, are normalized, showing mean values close to zero, reflecting their standardized distributions. Finally, socio-demographic variables add the remaining depth to the dataset. The proportion of the population aged 0-14 (Pobl.0_14_div_Poblac.Total) ranges from 8.9% to 18.4%, with a median of 12.6%, while the proportion of retired individuals (PoblJubilada_div_Poblac.Total) ranges from 11.6% to 22.8%, with a median of 18.7%. Additionally, the proportion of immigrants (Inmigrantes.porc) varies between 6.1% and 20.5%, with a median of 13.4%. Together, these variables capture a diverse set of structural, environmental, and socio-demographic factors that influence property prices in Madrid.

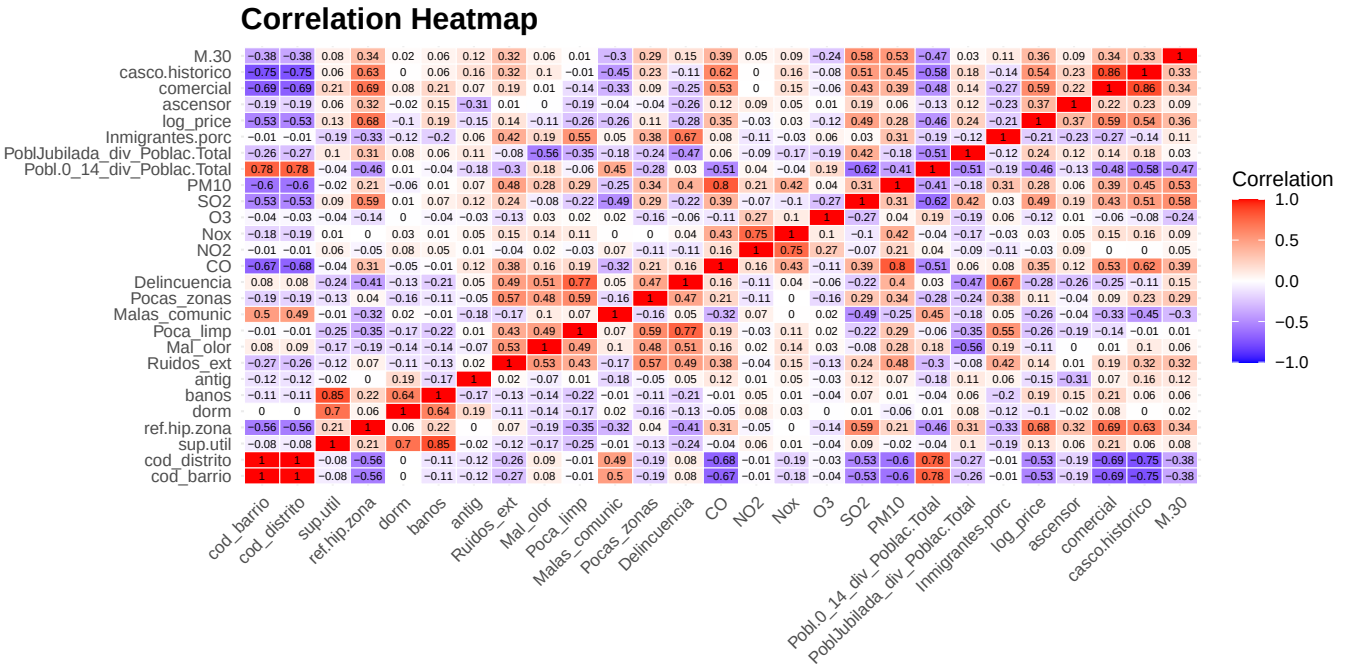


Figure 1: Correlation Heatmap: Relationship Between Variables

This correlation heatmap forms a crucial part of our exploratory analysis. Identifying independent variables that are highly correlated with one another is essential, as a high correlation value (generally >0.8) suggests that these variables explain similar information in the dataset, an issue known as multicollinearity. Avoiding multicollinearity is critical when fitting a multiple linear regression model, as its presence can inflate variance, produce unstable coefficients, and ultimately lead to inaccurate estimation and inference. Additionally, we aim to identify variables that are strongly correlated with our dependent variable (the price of the house), as these relationships will serve as the foundation for our regression model by highlighting key predictors.

We begin by examining correlations with the dependent variable. Among the independent variables, ref.hip.zona exhibits the strongest positive correlation with the log-transformed price variable. This is expected given the nature of this variable, which relates to the financing of the house. We also observe strong positive correlations of 0.54 and 0.59 for the variables casco.historico and comercial, respectively. These variables reflect the location of the house relative to the city center and its presence in a commercial area. Notably, these two variables also exhibit a high correlation of 0.86 with each other, suggesting they capture overlapping information. This aligns with our understanding that the city center of Madrid is typically a commercial hub. Including both variables in the model could introduce multicollinearity, an issue we will address in subsequent steps.

Furthermore, there is an almost perfect correlation of 0.98 between the independent variables sup.util and sup.const. This high correlation indicates that larger construction space is closely associated with increased usable living space within a house. To prevent multicollinearity, one of these variables will need to be removed from the model based on

the one with the lowest variance.

Other noteworthy correlations include a strong positive relationship (0.85) between sup.util and banos, indicating that houses with more bathrooms tend to have greater usable living space. Similarly, there is a high correlation of 0.8 between PM10 and CO, which is expected given the shared context of these variables. Addressing these patterns in the data will ensure that our regression model remains robust and avoids complications associated with multicollinearity.

2.1 Analysis of Numerical (Including Binary) Variables Against the Log of Price

We analyzed the relationship between numerical (including binary) variables and the log of price, identifying correlations that generally match logical expectations as well as some that were initially surprising. Introducing slight noise to categorical values using jitter improved the visualization of observation density and helped reveal meaningful patterns.

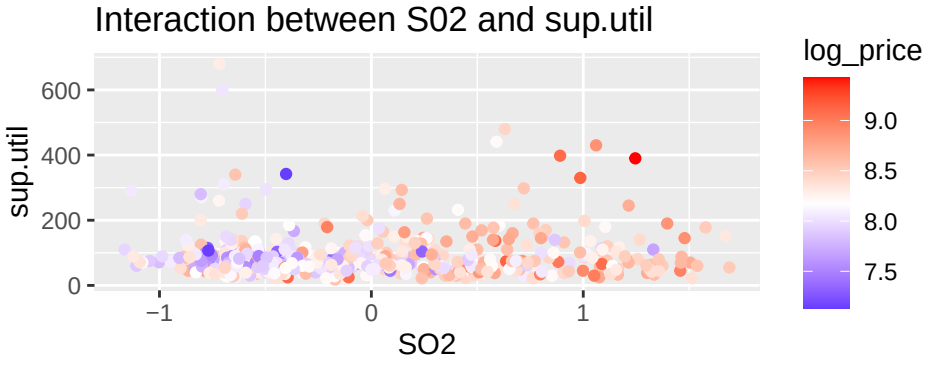
Regarding positive correlations, the mortgage reference for an area strongly correlates with the log of price. This makes intuitive sense, as higher mortgage values usually correspond to higher property prices. Certain binary features also have positive impacts on prices, including the presence of elevators, location in the historic center (casco histórico), proximity to commercial areas, and being situated within the M-30 ring road. These align with standard real estate principles, where features that enhance convenience and prestige boost property values. Likewise, having more bathrooms is associated with higher prices, as it often indicates a larger, more luxurious home. Areas with a higher proportion of older residents (población jubilada) also tend to have higher prices, probably reflecting stronger financial resources among that demographic.

On the other hand, negative correlations emerged with poor transportation and higher crime rates, both of which reduce an area's desirability and thus its property prices. Certain categorical characteristics—unpleasant odors, poor cleanliness, and a higher proportion of immigrant residents are linked to lower prices, likely reflecting perceptions of less maintenance and fewer economic resources in these neighborhoods. Families with young children often reside in more affordable neighborhoods. This may be due to tighter budget constraints, as young parents typically have not had sufficient time to build significant savings, thereby limiting their ability to purchase higher-priced homes.

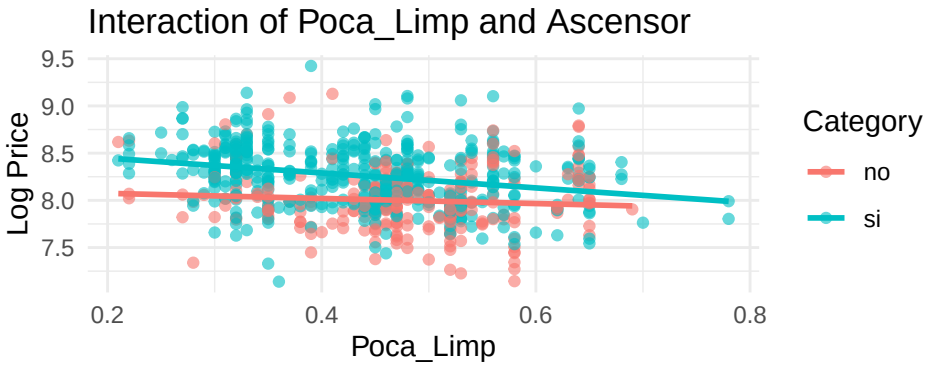
Several initially unexpected patterns became clearer with further analysis. The number of bedrooms follows a U-shaped pattern: small studio apartments, especially in central locations, tend to be expensive; properties with two or three bedrooms are generally cheaper; and very large homes with five or more bedrooms regain higher values due to their exclusivity and luxury. Although external noise is a negative feature, it correlates positively with the log of price because central areas—typically more expensive—are also noisier. In contrast, areas rich in green spaces tend to have lower prices, likely because these quieter, more suburban locations are generally less expensive. Finally, pollution levels also show a positive correlation with the log of price, not because buyers seek pollution, but because central Madrid—characterized by higher traffic, greater activity, and thus more pollution—also features higher property values.

2.2 Interaction Analysis

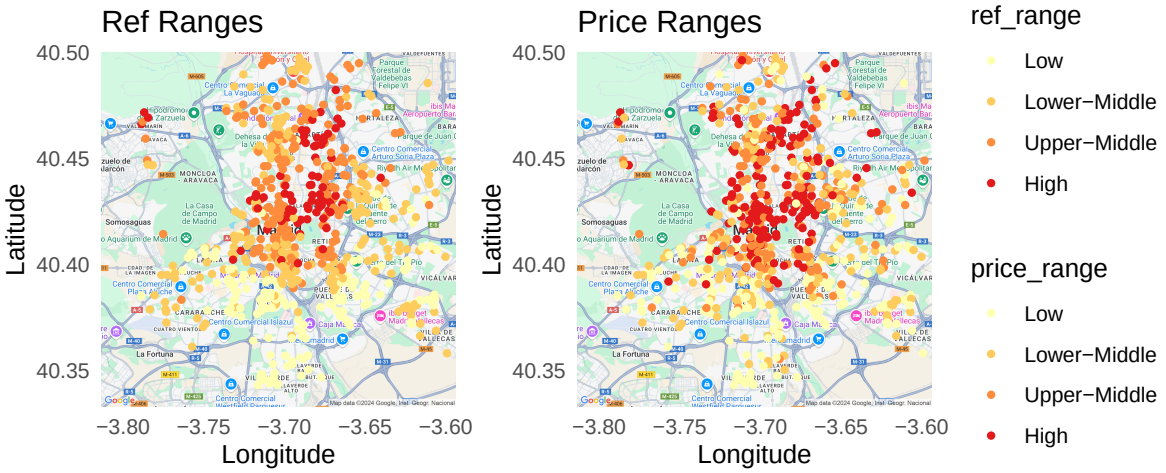
After extensive interaction analysis in the Model Construction process, we decided to further analyze two of the unexpected interactions in the model.



Here, we can see that there is a possible interaction between SO2 and sup.util that will have an impact on the price of houses. From this scatter plot, we can see that there is a concentration of lower prices of houses when the amount of usable space is low and the levels of SO2 are low. As the levels of SO2 increase, houses, with the same amount of usable living space, tend to be more expensive. This could be due to the fact that houses are more expensive in densely populated areas and densely populated areas contribute more to the levels of air pollution.



This plot gives a slight insight into the possible interaction of Poca_limp and Ascensor. We can clearly see that there is a small negative relationship between Poca_limp and the logarithm of house price. Suggesting that as the lack of street cleanliness increases, the price decreases. This graph also shows that houses with elevators have a higher price than those that don't. An interesting observation here is that as the lack of street cleanliness increases, we see that the prices of those houses with elevators decrease at a faster rate than those houses that don't have an elevator, suggesting that dirtier streets have a more pronounced effect on houses with elevators. This is an interesting relationship that we may want to account for in our regression model.



To explore the strong correlation between the variable ref.hip.zona—a city-determined mortgage reference that estimates property value based on location—and the actual price of houses, we created a map visualization comparing the two variables. This analysis not only highlighted spatial patterns in the data but also guided us in identifying and eliminating less relevant predictors for the model. The map's scale was based on quartile summaries of the logarithmic transformations of both price and ref.hip.zona, ensuring a consistent comparison. Observations were plotted using their longitude and latitude coordinates, with each point colored according to its respective price range. The resulting plots revealed that, for the most part, the categorized price ranges aligned closely with those of ref.hip.zona, particularly in distinguishing lower from higher price ranges. One notable insight was that proximity to the city center was not a defining factor for property prices. Initially, we considered including a variable to calculate the distance from each observation to the city center and grouping properties accordingly. However, the visualization showed that some properties near the center were relatively inexpensive, while others located farther away commanded higher prices,

which went against our original assumptions.

2.3 Variable Exploration and Feature Engineering

We experimented with several transformations and introduced new variables, such as aggregating pollution measures, adjusting the dormitory variable to align more closely with the log of price, and reducing the number of districts. Here, we present the latter transformation. Although these changes did not substantially improve the model's performance, they were worthwhile attempts at refinement.

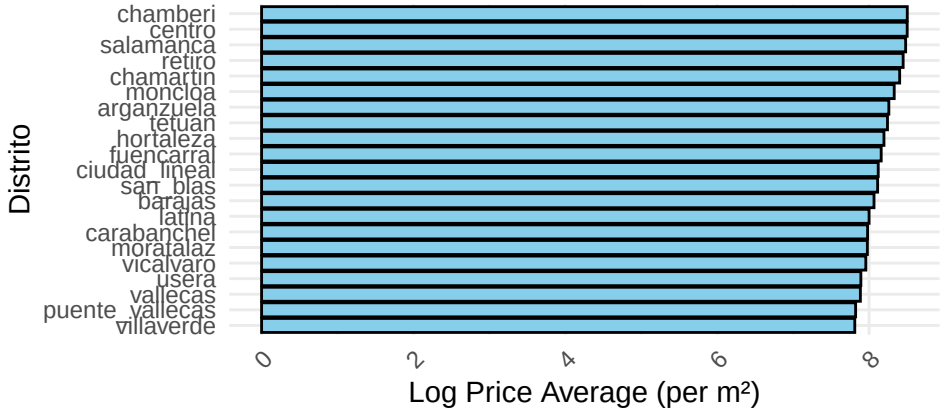


Figure 2: Correlation Heatmap: Relationship Between Variables

Initially, we analyzed the columns barrio and distrito and noticed that they contained a large number of categories. It is well-known that certain areas have significantly different housing prices. For instance, Salamanca is renowned as an affluent neighborhood, while Vallecas represents the opposite end of the spectrum. Given these clear distinctions, we believed it was worth exploring further. We decided to focus on distrito variable, as it offers a higher level of aggregation and simplifies the analysis. Using the training dataset, we calculated the average house price for each distrito. Based on the graph below and our understanding of Madrid's housing market, we grouped the districts into four categories ("High", "Medium-High", "Medium-Low", "Low"), ranging from most to least expensive. However, when we incorporated these categories into the regression model, the adjusted R-squared showed only a marginal improvement. Despite the limited impact, it was a worthwhile exploration and provided valuable insights into the influence of location on house prices.

3 Model Construction

The model construction process began by reviewing various regression techniques, including stepwise regression (both forward and backward selection). We initially built a simple linear model that included all available variables. However, after performing the ANOVA test, we observed that several variables did not contribute significantly to the model, and the amount of variables made a complicated model. This indicated that not all variables were relevant, prompting us to refine the model based on exploratory data analysis (EDA).

The EDA provided valuable insights into the key variables affecting house prices. For example, we found that 86% of the data is concentrated in the "piso" category of the tipo.casa2 variable, while other categories such as "atico", "chalet," and "duplex" had very low representation. Furthermore, the "otros" category only had a single data point, which introduced noise into the analysis. Based on these findings, we decided to focus on the more relevant categories and remove those that didn't significantly contribute. We decided to remove the only observation with tipo.casa classification "Otros" as it did not heavily influence the model, only created NA values. Furthermore, using the analysis from the EDA, we decided to re-level the variables "tipo.casa" and "estado" so the model would put the

default or baseline value in the classification that had the most observations. Around 76% of the variable "estado" is classified as "buen_estado" and thus holds the most information about estado. As mentioned before, around 86% of the tipo.casa category is classified as "piso". Releveling these variables by their respective strongest classification created the variables tipo.casa2 and estado2.

Our modeling process followed a stepwise approach, initially using only numerical variables, then categorical variables, and later combining both. We also explored interactions between variables, aiming to identify key predictors of house prices. For example, we initially considered variables that we hypothesized would contribute to the model but discovered that the data did not always support these assumptions. This reinforced the idea that data-driven model selection was more effective than relying solely on expected relationships.

3.1 Variable Selection and Model Refinement

As we refined the model, we also kept in mind the need to avoid multicollinearity. We performed correlation tests and created a correlation heatmap to identify strong correlations between numerical variables. From these analyses, we determined that the ref.hip.zona variable was a key predictor, as it had a high correlation with house prices. It became clear that including this variable would improve the model.

During the variable selection process, we also encountered some challenges. For instance, we considered including the "barrios" variable, which had over 100 levels. However, we found that this would complicate the model unnecessarily and lead to overfitting, so we decided to remove it. We also excluded variables like cod_barrio and cod_distrito that, although numerical, were not ordinal and did not provide meaningful insights for predicting house prices.

Further analysis showed that latitud was a better predictor of house price locations than distrito, so we focused on latitude instead, as it was a numerical value with one level. Attempts to combine latitude with longitude did not improve the model, so we decided to exclude longitude as a predictor due to its limited variability.

3.2 Final Model and Results

Metric	Value
Adjusted R Squared	0.6641380
Residual Standard Error (Sigma)	0.1976152

Table 1: Summary Statistics for Top Model

After testing various combinations of variables and interactions, we arrived at a final model that explained about 70% of the variance in house prices, as indicated by the adjusted R-squared. However, the model became increasingly complex, with over 30 variables, which risked overfitting. Despite its relatively good predictive power, the large number of variables made the model difficult to interpret and may have limited its generalizability. Overall, the model-building process emphasized the importance of data exploration, variable selection, and careful consideration of model complexity. While the final model provided a solid foundation for predicting house prices, further refinements would be necessary to ensure its robustness and avoid potential overfitting.

3.2.1 ANOVA

Table 2: ANOVA Table for Top Model

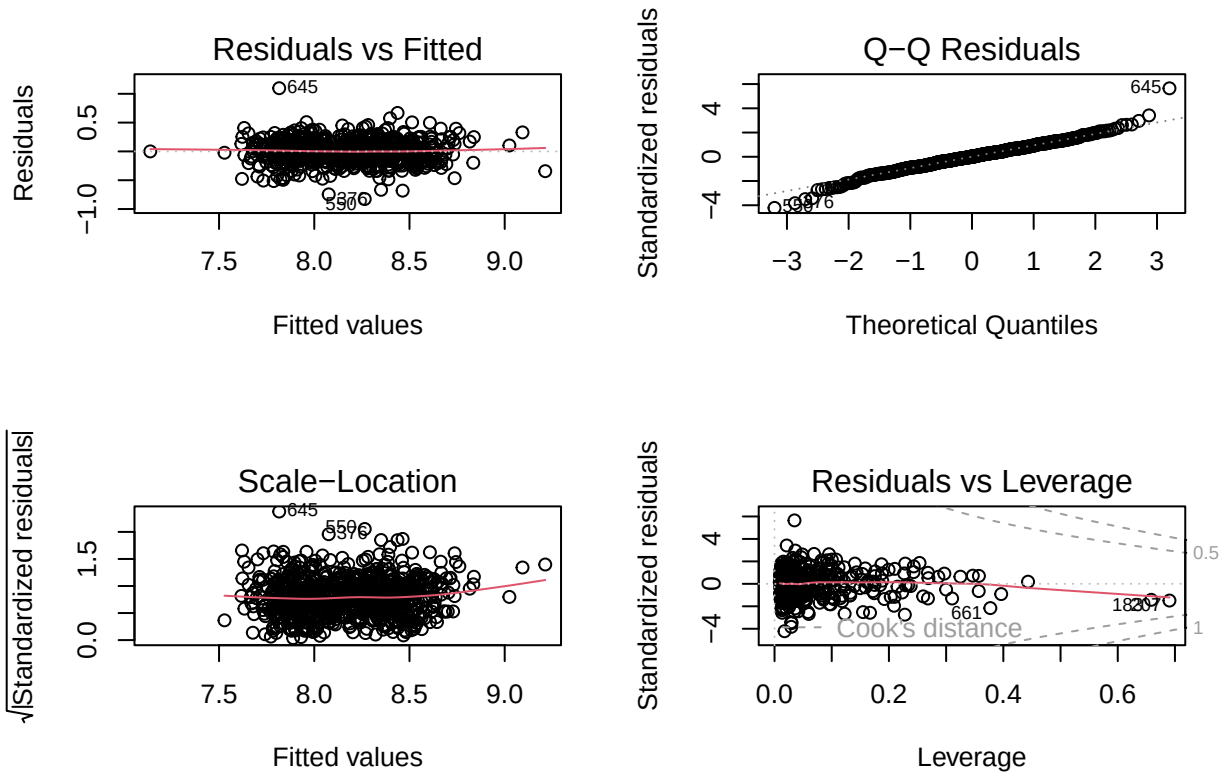
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ref.hip.zona	1	39.2063195	39.2063195	1003.957979	0.0000000
latitud	1	0.5841568	0.5841568	14.958530	0.0001202
antig	1	1.6689542	1.6689542	42.736984	0.0000000
estado2	6	1.7539406	0.2923234	7.485539	0.0000001

tipo.casa2	4	1.4983788	0.3745947	9.592263	0.0000001
dorm	1	0.4336615	0.4336615	11.104791	0.0009066
banos	1	1.4482746	1.4482746	37.086032	0.0000000
SO2	1	1.3308520	1.3308520	34.079186	0.0000000
ascensor	1	0.6542586	0.6542586	16.753630	0.0000476
M.30	1	0.7402290	0.7402290	18.955077	0.0000154
comercial	1	2.9281182	2.9281182	74.980453	0.0000000
tipo.casa2:ascensor	4	1.8236272	0.4559068	11.674426	0.0000000
tipo.casa2:Pobl.0_14_div_Poblac.Total	5	1.8472447	0.3694489	9.460495	0.0000000
estado2:PM10	7	1.0627282	0.1518183	3.887618	0.0003635
SO2:sup.util	1	0.7414140	0.7414140	18.985422	0.0000152
ascensor:Poca_limp	2	0.4423534	0.2211767	5.663682	0.0036316
Residuals	696	27.1800204	0.0390518	NA	NA

The ANOVA test was conducted at various stages during the model construction to assess the significance of variable inclusions and determine whether they improved the model's fit. As discussed earlier, different variable orders were also considered, as the correlation between predictors can influence the significance of each variable. This iterative process ensures that only meaningful predictors are retained in the model, enhancing its predictive power. The analysis focused on evaluating the p-value, where 0.05 was used to identify variables that significantly contribute to explaining the variation in house prices. A p-value below this threshold indicates that a predictor has a meaningful effect on the model, suggesting it is crucial for explaining the outcome variable.

The ANOVA results reveal that all the variables included in the model are statistically significant at the 0.05 level, which strengthens the validity of the chosen predictors. Notably, the most influential variables are ref.hip.zona, latitud, and antig, all showing p- below 2.2e-16. Among these, ref.hip.zona explains the largest share of the variation, as evidenced by its high sum of squares (39.206). Other key predictors, such as banos, SO2, and comercial, also demonstrate highly significant contributions, reinforcing their importance in modeling house prices. In addition, interaction terms like tipo.casa2:ascensor and tipo.casa2:Pobl.0_14_div_Poblac.Total significantly improve the model fit, as reflected in their low p-values. The interaction term estado:PM10 is also statistically significant with a p-value of 0.0003635, although it accounts for a smaller proportion of the variance. The residuals, with a sum of squares of 27.180, remain small relative to the explained variance, indicating a strong fit. Overall, the ANOVA analysis confirms that the model is well-specified, with a robust set of predictors and interactions that collectively enhance the understanding of house price variation. This supports the inclusion of these variables in the final model.

3.2.2 Model Assumptions and Error Analysis



Diagnostic plots are valuable tools in regression analysis, as they allow for the assessment of model assumptions such as linearity, homoscedasticity, and normality of residuals, while also providing insights into potential outliers or influential data points (Fox, 2020). The first plot, the residuals vs. fitted plot, provides a visualization of the assumptions of linearity and homoscedasticity of the residuals, that is, they show no pattern and have constant variance. The red line indicates the zero line, where, ideally, we would expect most residuals to fall. There is a strong cluster of residuals around this line, however there are a couple significant outliers, shown as observations 645, 376, and 550. Overall, with only three significant outliers and no obvious pattern in the points, the residuals vs. fitted plot shows linearity and homoscedasticity of the residuals.

The second plot, the Q-Q plot, is to model how normally distributed the residuals are. The dotted line is the normal line, if the points closely follow the line, they can be assumed to be approximately normally distributed. Deviance from the line may indicate non normality. Overall, the plot shows that most of the residuals follow the line in the center, however there is lifting and pulling from the line at either end. This indicates that the residuals are approximately normally distributed but have heavy tails on either end, meaning there are outliers at either extreme. The same three outliers from the residuals vs. fitted plot are indicated as they are the furthest away from the line, and that although the majority of the data follows a normal distribution, these observations are more extreme than what a normal distribution would predict.

The scale-location plot is another visualization that checks for homoscedasticity of residuals. If the residuals are homoscedastic, the square root of the standardized residuals should be randomly spread across fitted values. A pattern suggests heteroscedasticity, meaning the variance of residuals changes with fitted values. The red line is a fitted line that follows the pattern of the data. The residual points are mostly randomly spread over this plot, without an obvious pattern, although the red line shows a slight upward curve.

Finally, the residuals vs. leverage plot can identify influential observations, which are considered observations that have high residuals and leverage that disproportionately affects the model. We want to observe the behavior of the points that are not close to the bulk of the data, and moreso, within the Cook's Distance dotted lines. We can observe

two points that are labeled by the plot that deviate significantly from the bulk of the data and are approaching the Cook's Distance line, although not within it. These are observations 183 and 207, which are two observations we may want to consider removing after further diagnostics.

3.2.3 VIF

Table 3: Variance Inflation Factor for the Top Model

Variable	GVIF	Df	Adjusted GVIF
ref.hip.zona	3.562857	1	1.887553
latitud	1.744042	1	1.320622
antig	1.705571	1	1.305975
estado2	25.060886	6	1.307926
tipo.casa2	28612238.368860	4	8.552025
dorm	2.401391	1	1.549642
banos	2.628471	1	1.621256
SO2	4.802791	1	2.191527
ascensor	27.477784	1	5.241926
M.30	1.980633	1	1.407350
comercial	2.563958	1	1.601236
tipo.casa2:ascensor	180.110450	4	1.914002
tipo.casa2:Pobl.0_14_div_Poblac.Total	45802395.284019	5	5.835644
estado2:PM10	38.122795	7	1.297005
SO2:sup.util	3.451441	1	1.857805
ascensor:Poca_limp	43.037558	2	2.561309

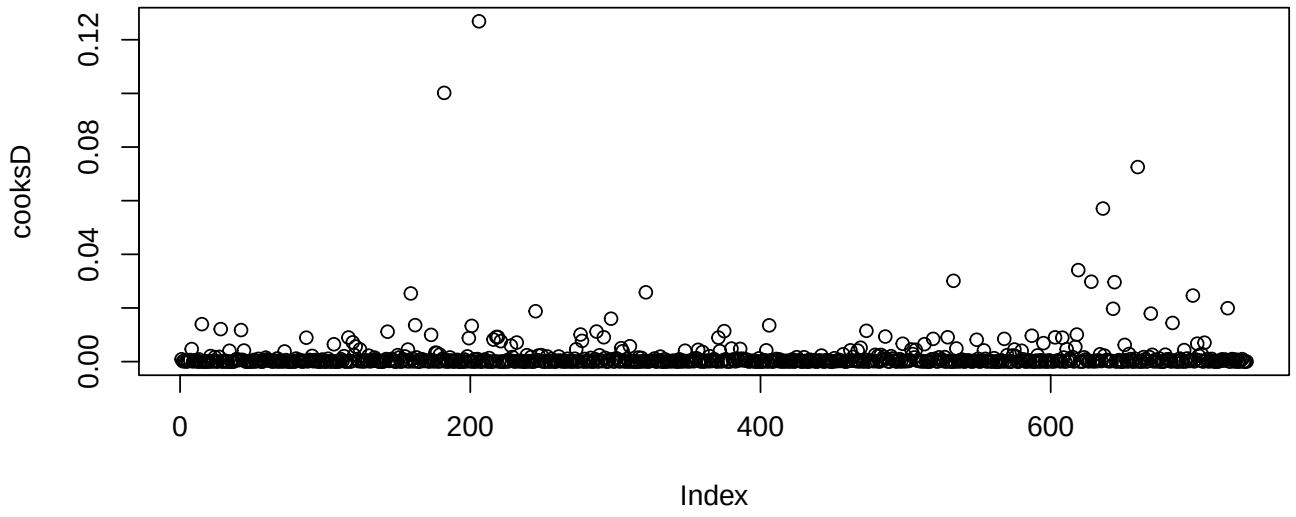
The variable tipo.casa2 has the highest GVIF^(1/(2*Df)) value of 8.55 in the model, but it remains below the threshold of 10, indicating no severe multicollinearity. Upon further investigation, 86.5% of the data falls under the reference category "piso," which dominates the model, while other categories (e.g., "atico," "chalet," "duplex," "estudio," and "otros") account for only a small portion and have limited impact. Notably, the "chalet" category has a significant p-value, showing some relevance. An ANOVA analysis confirmed that keeping tipo.casa2 in the model improves the adjusted R-squared, justifying its inclusion.

The ascensor variable has a VIF slightly above 5, indicating moderate correlation with other predictors but remains well below the problematic level of 10. This variable, where "yes" appears in 66% of the data, shows that properties with an elevator have a higher median price. Despite the moderate VIF, no other predictors exhibit problematic collinearity, and the model's performance remains strong, so the VIF for ascensor does not pose a significant issue.

Most other predictors have adjusted VIF values well below 10, suggesting that multicollinearity is not a major concern. The model is relatively stable, with low VIFs reinforcing its well-specified nature. While tipo.casa2 shows potential for collinearity due to category imbalance, its inclusion is justified by its contribution to model performance, and overall multicollinearity is not a significant issue.

Although we observed three VIFs over 5, suggesting potential multicollinearity (Kutner et al., 2004), a ridge regression model was fit with the same predictors to reduce redundant coefficients. However, the ridge model had a slightly worse RMSE by 0.0046, and therefore the original model remains stronger for predicting house prices.

3.2.4 Cook's Distance



Cook's Distance is a measure used in regression analysis to identify observations that have a significant influence on the model's estimated coefficients (Cook & Weisberg, 1982). It quantifies how much the fitted values of the model would change if a particular observation were removed, with higher values indicating more influential points.

A visual representation of this can be seen in the above plot. A majority of the data points have very small Cook's Distance values, indicating that these observations have little to no influence on the overall regression model. These points do not cause significant changes in the model's coefficients when excluded. There are a few points that stand out with higher Cook's D values, such as the ones above 0.08 and 0.12. These points could be influential observations. Removing these points could cause noticeable changes in the regression coefficients or model fit.

After fitting the same model but with the four influential data points removed, the model actually has a better fit on the data. The adjusted R-squared increases by around 0.3 percent, the mean squared error decreases by 0.001, and the AIC also decreases, all indicating a slight gain in strength of the model from removing these observations.

Removing the four influential observations also caused every coefficient to slightly change. The largest changes occur in:

- `Tipo.casa2chalet: an absolute change of 0.853`
- `Intercept: an absolute change of 0.743`
- `tipo.casa2duplex: an absolute change of 0.409`
- `tipo.casa2chalet:Pobl.0_14_div_Poblac.Total: an absolute change of 0.132`
- `estado2excelente:PM10: an absolute change of 0.0845`
- `estado2excelente: an absolute change of 0.116`

Looking at the variables that we removed to see if they contain information on these variables, two hold information for the estado2excelente variable, one holds information on tipo.casa2chalet and one holds information on tipo.casa2duplex, and the other two contain information on tipo.casa2piso, which is stored in the Intercept term. As the observations removed contain information on all of the variables listed, this can explain why they had such high effects on those specifically.

Although these results show us a slight increase in the strength of our model, is it significant enough to change our dataset? To do this, we can conduct a Nested Model Comparison with a Likelihood Ratio Test.

The Likelihood Ratio Test gives us a result where a p-value smaller than 0.05, the newer model is significantly better than the older model. The degrees of freedom being zero is due to only observations being removed, but the amount of predictors in the model remained the same. With a p-value of 0, the model with influential points removed is significantly better than the original model we were using. We decided to keep the original model we found to hold all information as the coefficients were not significantly different.

We can also compute confidence intervals for the predictor coefficients. Using the coefficients from the original model as our Beta-hats, if the new coefficients fall outside the confidence interval region, we would deem it a significant difference.

Although the Likelihood Ratio Test concluded the reduced model was significantly stronger than the original model, none of the new coefficients fall outside the confidence interval from the original model. The Likelihood Ratio Test evaluates the overall model fit by comparing the likelihoods of the two models. It can show a significant improvement even if changes in individual coefficients are not large enough to fall outside the original confidence intervals, that is, the test considers the combined effect of all the predictors and the residuals, not just specific coefficients.

Although removing the four influential observations takes information away from the predictors, the overall model fit became stronger without greatly changing the coefficients.

3.3 Interpretations

3.3.1 Individual Predictors

The variable "ref.hip.zona" is a mortgage reference that captures the location effect on property prices. The coefficient indicates that a unit increase in ref.hip.zona will increase the log of price by 0.0001. Although this effect seems small, ref.hip.zona has the strongest correlation of the predictor variables with the log of price as the variable itself is set by the market as the projected price based on location. The variable "latitud" is the exact numerical coordinate of the house. The coefficient indicates that with a one unit increase in latitude, the log of price is expected to increase by 1.252. Such a large coefficient might suggest that latitude correlates with an economically significant location difference, such as moving toward a more desirable region.

The variable "antig" is the age of the property expressed in years. A one year increase in the age of the property results in a decrease in the log of price by 0.002205. This indicates that, as a property gets older, the value will go down, thus decreasing the price.

The variable "estado" classifies houses based on their condition. There are six classifications, a_reformar (to be renovated), buen_estado (good), excelente (excellent), nuevo-semin (new/semi-new), reformado (renovated), reg,-mal (regular to poor), segunda_mano (second hand). The original estado variable was re-leveled so that the model would select buen_estado as the baseline since the most number of houses were classified as buen_estado. Thus, each coefficient corresponds to the difference in the effect of house condition on log of price compared to a_reformar. For classifications such as "a_reformar", "excelente", and "segunda_mano", the price of the house is expected to be lower than a house classified as "buen_estado" with all of the same characteristics otherwise. For classifications such as "reformado" and "reg,-Mal" the price is expected to be higher than a home classified at "buen-estado" when all other predictors are the same.

The variable tipo.casa classifies properties based on what kind they are, and are in six levels, piso, atico, chalet, duplex, and estudio. We re-leveled tipo.casa to use piso as the baseline category because it has the most observations, making it the most representative group in the data. Each individual coefficient can be interpreted as the change in log of price for that type of property when compared to a property classified as "piso" that has every other characteristic to be the same.

The variable "dorm" represents the number of bedrooms in each property. The coefficient of -0.06193 indicates that, holding all other variables constant, a one-unit increase in the number of bedrooms is associated with a decrease of 0.06193 in the log of price. Interestingly, it would be expected that the more bedrooms, the higher the price would be. There are several un-intuitive ways to interpret this result. One, we have seen that "dorm" has a strong, positive

Variable	Coefficient
(Intercept)	-42.5309614
ref.hip.zona	0.0001005
latitud	1.2523736
antig	-0.0022050
estado2a_reformar	-0.1184199
estado2excelente	-0.0570035
estado2nuevo-semin,	0.0205825
estado2reformado	0.1196496
estado2reg,-mal	0.0172779
estado2segunda_mano	-0.2232474
tipo.casa2atico	-0.0369579
tipo.casa2chalet	2.2792594
tipo.casa2duplex	0.2207508
tipo.casa2estudio	-0.1877979
dorm	-0.0619281
banos	0.0620591
SO2	-0.0739133
ascensorsi	0.2262983
M.301	0.0665986
comercial1	0.1816831
tipo.casa2atico:ascensorsi	-0.0216043
tipo.casa2chalet:ascensorsi	-0.9020383
tipo.casa2duplex:ascensorsi	-0.0355030
tipo.casa2estudio:ascensorsi	0.3368456
tipo.casa2piso:Pobl.0_14_div_Poblac.Total	-0.0219759
tipo.casa2atico:Pobl.0_14_div_Poblac.Total	-0.0140297
tipo.casa2chalet:Pobl.0_14_div_Poblac.Total	-0.1836226
tipo.casa2duplex:Pobl.0_14_div_Poblac.Total	-0.0499768
tipo.casa2estudio:Pobl.0_14_div_Poblac.Total	-0.0237297
estado2buen_estado:PM10	0.0672971
estado2a_reformar:PM10	0.0479867
estado2excelente:PM10	-0.4066462
estado2nuevo-semin,:PM10	0.3166841
estado2reformado:PM10	-0.1554041
estado2reg,-mal:PM10	0.1119149
estado2segunda_mano:PM10	-1.0514208
SO2:sup.util	0.0008293
ascensorno:Poca_limp	-0.0052961
ascensorsi:Poca_limp	-0.3401877

Table 4: Model Coefficients (Estimates)

correlation with banos, and with assuming banos stays constant, an increase in bedrooms could lower the property value. The negative coefficient of -0.0619 for the dorms variable, reflects the U-shaped relationship observed in the data. The boxplot shows that as the number of bedrooms increases from 0 to 3, the median price decrease. Since 83% of the data falls between 0 and 3 bedrooms, the negative coefficient captures the pricing trend for the majority

of properties. Additionally, properties with 0 bedrooms, such as atico and estudio, represent only 1.22% of the data. Beyond 3 bedrooms, the price increases, but the negative coefficient primarily reflects the trend for properties with fewer bedrooms.

The variable "banos" represents the number of bathrooms in each property. The coefficient of -0.06206 indicates that, holding all other variables constant, a one unit increase in the number of bathrooms would result in a decrease in the log of price by 0.06206. Similarly to the "dorm" variable, this result seems unintuitive, however, with prior knowledge of the relationship of dorm and banos, increasing the number of bathrooms but not the number of bedrooms could lower the value of the property. It is important to interpret the individual variables in relation to all other variables in the model to make practical interpretations of the coefficients.

The variable SO2 is the standardized level of sulfur dioxide in the air at the location of that home. The coefficient of -0.07391 indicates that with a one unit increase in SO2, the log of the price will decrease by 0.07391. Practically, this indicates that the worse the air quality at the property, the lower the price will be.

The variable "ascensor" is a bi-level factor that indicates if the property has an elevator or not, si or no. The coefficient of 0.2263 indicates that, in the presence of an escalator, the log of price will increase by 0.2263 compared to a model with no elevator but with all other levels being the exact same. The baseline selected by the model is the "no" input for ascensor, so if a property has no elevator there is zero increase in price.

The variable M.30 is a binary variable, with a 1 indicating the property is within the M-30 highway and a 0 or not. The model selected "0" for the baseline, and therefore the coefficient of 0.0666 indicates that a property that is within the M-30 will increase the log of price by 0.0666 compared to a property with the exact same levels of the other predictors just outside of the M-30. This result is intuitive as it may be more desirable to live within the M-30 as that is considered the "center", and thus could lead to an increase in price.

The variable "comercial" is a binary variable in which a 1 indicates the property is located in a commercial area and a 0 indicates that it is not. The model selected "0" for the baseline, and therefore the coefficient of 0.1817 indicates that a property within a commercial area will have a log price 0.1817 higher than a property not in a commercial area but with all other levels matching. This indicates that, regardless of other characteristics, a commercial area will be more expensive to live in.

3.3.2 Interactions

tipo.casa2:ascensor - The interaction terms capture how the effect of having an elevator differs for each house type compared to the baseline house type, piso, without an elevator. The model assumes there is a main effect for having an elevator, ascensorsi, and a main effect for each house type, without the interaction term. These are combined to compute the overall effect.

tipo.casa2:Pobl.0_14_div_Poblac.Total - For this interaction, the individual coefficients represent the effect on log of price across all house types when there is a unit increase in the percentage of children between 0 and 14 years in that district. The results indicate that as the percentage of children in the population increases, the log of housing prices decreases across all house types, with the largest impact observed for Chalets and the smallest for Áticos. This suggests that demographic factors, such as the proportion of children, are linked to housing prices in the market.

estado2:PM10 - For the interaction of Condition and Air Quality, each individual coefficient represents the effect on the log of price for one unit increase in PM10 depending on the condition of the house. This reflects how air quality interacts with the state of the house to influence housing prices.

SO2:sup.util - For every 1-unit increase in SO2, the effect of a 1-unit increase in usable area (sup.util) on log of price increases by 0.0008293. This suggests that the value of additional usable area may be slightly higher in areas with lower air quality, potentially due to compensatory factors.

ascensor:Poca_limp - The coefficients for the interaction of Street Cleanliness and if the house has an elevator indicate the different effects on the log of price. Depending on if the house has an elevator, a unit increase in Street Cleanliness when no elevator is present is -5.296e-03 and when there is an elevator it is -3.402e-01.

3.4 Model Selection

After thorough evaluation using multiple methodologies and techniques, the top model was selected based on its balanced performance and simplicity. Initially, Model Version 1 seemed promising due to its high adjusted R-squared of 0.714 and a relatively low AIC of -310.60. However, upon closer examination, the model's performance did not justify its complexity. Despite the impressive fit, the model contained 48 variables and over 120 coefficients, and the increase in adjusted R-squared with the addition of more predictors was very slow, suggesting diminishing returns. This indicated that the model was becoming overcomplicated without significant improvements in fit, risking overfitting and reducing its generalizability. Moreover, the residual sum of squares (RSS) for Model Version 1 was only marginally better than the top model's (20.680 vs. 27.18002), and the residual deviance was only comparable after adding the 22nd variable, further reinforcing the idea that additional variables were not providing substantial improvements.

In comparison, the top model, consisting of just 16 variables, achieved an AIC of -257.74, which was not far off from the more complex Model Version 1's AIC but with far fewer variables, making it a simpler and more efficient choice. This model achieved an adjusted R-squared of 66%, which was competitive with the higher values seen in the more complex models. Additionally, the exponential of the residual standard error (RSE) of 0.1976 indicated that the model had only 21.85% additional variability in its predictions, representing a reasonable level of error without overcomplicating the model. Furthermore, the model passed the multicollinearity test (VIF) with no severe correlations between predictors, confirming that it was stable and not overfitted. The model also satisfied the key linear regression assumptions, such as linearity and normality of residuals, making it reliable and robust.

In contrast, Model Version 2, which used a smoothness technique to handle non-linear relationships, achieved a better AIC of -270 but did not significantly improve upon the top model's performance. Its adjusted R-squared remained at 0.67, and its deviance was close to the top model's, further supporting the argument that adding more complexity did not lead to a meaningful improvement. Additionally, Model Version 2 removed the SO2 predictor, highlighting the importance of variable significance in model selection.

Based on the analysis of variance (ANOVA) results for the three models, Top Model stands out as the most effective due to its ability to balance model complexity and explanatory power. The most significant variable across all models is ref.hip.zona (mortgage reference of the area), which consistently shows a very low p-value, indicating its critical role in determining property prices. Similarly, latitud (latitude) and antig (age of the property in years) are both highly significant factors in Top Model, with the latter having an extremely low p-value in all models, highlighting its importance in pricing of the houses in Madrid. The estado (condition of the property) also proves to be a significant predictor, with Top Model showing a low p-value, reinforcing that a property's condition has a major impact on its value. Additionally, Top Model captures important interaction terms, such as tipo.casa2:ascensor (house type and elevator), and estado:PM10 (condition of the property and air quality), which suggest complex relationships between variables, especially in urban settings. These interactions reflect real-world dynamics and preferences for renting or buying properties, where combinations of factors like house type and amenities significantly influence price.

One of the key reasons for selecting Top Model over the others is its simplicity. Despite incorporating interaction terms and crucial predictors, it avoids the added complexity of smoother terms, such as the s(latitud) term found in Model Version 2, making it more interpretable without sacrificing explanatory power. Moreover, the model maintains a strong fit by including highly significant predictors such as ref.hip.zona, latitud, antig, and estado, which are all essential for accurately predicting housing prices. The inclusion of environmental factors like SO2 and PM10, especially in interaction with other features, further adds depth to the model, highlighting the influence of external conditions on property pricing. In comparison to Model Version 2, which adds smooth terms but does not provide a significantly better fit, Top Model offers a more straightforward yet robust approach to predicting property prices. It is the best choice due to its ability to capture key patterns in the data while remaining accessible for interpretation.

Ultimately, the top model, with just 16 variables, offers the best trade-off between predictive accuracy and simplicity. Its performance, with an adjusted R-squared of 66% and an AIC of -257.74, demonstrates that adding more variables does not significantly improve the model's fit but increases its complexity, leading to potential overfitting. The top model also showed that variables included were all significant and the model met the necessary assumptions, ensuring its generalizability and stability. Thus, after considering multiple models and variables, the top model emerged as

the most effective and practical choice, providing solid predictive performance while maintaining interpretability and avoiding overfitting.

$$
\begin{aligned}
log\_price = & -42.64938 \\
& + 0.00010054 ref.hip.zona \\
& + 1.252374 latitud \\
& - 0.002205042 antig \\
& + 0.1184199 estado2a\_reformar \\
& - 0.0570035 estado2excelente \\
& + 0.0205825 estado2\_nuevo\_semin, \\
& + 0.1196496 estado2\_reformado \\
& + 0.0172779 estado2\_reg,\text{-}mal \\
& - 0.2232474 estado2\_segunda\_mano \\
& - 0.03695789 tipo.casa2\_atico \\
& + 2.279259 tipo.casa2\_chalet \\
& + 0.2207508 tipo.casa2duplex \\
& - 0.1877979 tipo.casa2estudio \\
& - 0.06192813 dorm \\
& + 0.06205912 banos \\
& - 0.07391325 SO2 \\
& + 0.2262983 ascensorsi \\
& + 0.06659857 M.301 \\
& + 0.1816831 comercial1 \\
& - 0.02160427 tipo.casa2atico:ascensorsi \\
& - 0.9020383 tipo.casa2chalet:ascensorsi \\
& - 0.03550302 tipo.casa2duplex:ascensorsi \\
& + 0.3368456 tipo.casa2estudio:ascensorsi \\
& - 0.02197592 tipo.casa2piso:Pobl.0\_14\_div\_Poblac.Total \\
& - 0.0140297 tipo.casa2atico:Pobl.0\_14\_div\_Poblac.Total \\
& - 0.1836226 tipo.casa2chalet:Pobl.0\_14\_div\_Poblac.Total \\
& - 0.04997677 tipo.casa2duplex:Pobl.0\_14\_div\_Poblac.Total \\
& - 0.0237297 tipo.casa2estudio:Pobl.0\_14\_div\_Poblac.Total \\
& + 0.0479867 estadoa2\_reformar:PM10 \\
& + 0.0672971 estado2buen\_estado:PM10 \\
& - 0.4066462 estado2excelente:PM10 \\
& + 0.3166841 estado2nuevo\_semin,:PM10 \\
& - 0.1554041 estado2reformado:PM10 \\
& + 0.1119149 estado2reg,\text{-}mal:PM10 \\
& - 1.051421 estado2segunda\_mano:PM10 \\
& + 0.0008292982 SO2:sup.util \\
& - 0.005296104 ascensorno:Poca\_limp \\
& - 0.3401877 ascensorsi:Poca\_limp
\end{aligned}
$$

3.5 Findings

Using coefficients from our linear regression can allow us to identify characteristics of homes that tend to be more expensive and vice versa. It is important to understand that, although these coefficients come from real data, they are in a model with a plethora of other variables, and therefore one individual coefficient may not be intuitive or perfectly interpretable in this context.

Homes that tend to be more expensive will have a higher mortgage reference, be located in the North, and be a young property and renovated. The types of homes we will see as more expensive will be chalets, duplexes, and pisos with an elevator, within the M-30, and in a commercial area but have a lower SO2 count. If the home has more bathrooms than bedrooms, it will increase in value, but having the same amount of bedrooms and bathrooms will also keep the price higher. On the other hand, homes that tend to be less expensive will have a lower mortgage reference, be located in the South, and be more years older, especially those that need to be renovated or are second hand. Types of homes that tend to be less expensive are studios and aticos that have a higher amount of bedrooms than bathrooms and no elevator. These properties will also be located outside of the M-30, not in a commercial area, and have poorer air quality.

4 Conclusion

The aim of this project was to create an effective multiple linear regression model to predict house prices in Madrid. During this process, we identified the most important attributes influencing house prices, such as mortgage reference, through a combination of exploratory analysis, variable selection methods, and inference of the parameters of our final model. Beyond examining variables that were highly correlated with log price, we explored interesting interactions between variables that a simple additive linear regression model might overlook. Including these interaction terms improved the model's fit, as reflected by metrics such as AIC and adjusted R-squared. House prices were found to positively correlate with factors such as the area's mortgage reference, location within the M-30, and overall property size. To further refine the analysis and align with insights from the Madrid real estate market, we experimented with feature engineering, including grouping districts into four categories based on their average prices. However, these adjustments did not yield significant improvements and were not included in the final model.

The model-building process utilized a stepwise selection approach to identify key predictors of house prices and refine the model. This included eliminating irrelevant variables, addressing multicollinearity, and incorporating meaningful interactions. After testing multiple models, the final model—with 16 predictors—offered the best balance between predictive accuracy and simplicity, achieving an adjusted R-squared of 66% and an AIC of -257.74. Finally, several diagnostic measures were conducted to ensure optimal model selection, evaluate model and error assumptions, and address influential or outlier points. Techniques such as analysis of variance (ANOVA) tests for predictor significance, calculation of variance inflation factors (VIF) for multicollinearity, residual plots to check error assumptions, and Cook's Distance to identify influential points were employed. These steps allowed us to finalize a robust linear regression model that captures the most significant predictors of house prices while avoiding overfitting and unnecessary complexity.

5 References

1. Fox, J. (2020). Regression Diagnostics: An Introduction (2nd ed.). SAGE Publications
2. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). Applied Linear Statistical Models (5th ed.). McGraw-Hill Irwin
3. Cook, R. D., & Weisberg, S. (1982). Residuals and Influence in Regression. New York: Chapman and Hall.