Madrid Real Estate Market Analysis

A Data-Driven Strategic Framework for Asset Acquisition

Prepared for: Fictional Investment Fund
Date: January 30, 2026

Abstract

This report presents a rigorous quantitative analysis of the Madrid residential real estate market. By leveraging advanced machine learning algorithms (XGBoost), we have developed a proprietary "Fair Value" model to identify market inefficiencies. The analysis highlights specific investment opportunities ("Gems") where list prices significantly deviate from intrinsic value, offering a fictional investment fund a clear roadmap for high-yield capital deployment with mitigated risk.

Confidential | Strategic Assessment

1 Project Overview

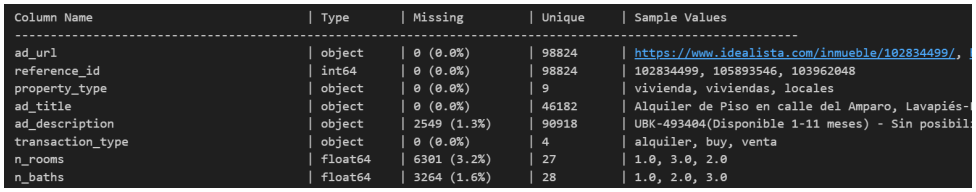
We began with a raw dataset comprising nearly 200,000 rows and 43 columns. Our primary objective is to identify market trends that will enable a fictional investment fund to strategically inject capital.

Given the strict time constraints, we implemented specific structural measures, most notably limiting the scope of our study to the Madrid metropolitan area and, in particular, to the study of residential housing. The project is organized across three notebooks, designed to replicate a standard industry data science workflow.

2 Technical Execution Roadmap

2.1 Phase 1: Data Acquisition & Cleaning (Notebook 1)

The objective of this stage was to gain familiarity with the data and perform rigorous cleaning. We developed custom functions (`general_overview` and `object_overview`) to visualize the data structure efficiently (Figure 1). Redundant features were identified and eliminated to streamline



Column Name	Type	Missing	Unique	Sample Values
ad_url	object	0 (0.0%)	98824	https://www.idealista.com/inmueble/102834499/,
reference_id	int64	0 (0.0%)	98824	102834499, 105893546, 103962048
property_type	object	0 (0.0%)	9	vivienda, viviendas, locales
ad_title	object	0 (0.0%)	46182	Alquiler de Piso en calle del Amparo, Lavapiés-
ad_description	object	2549 (1.3%)	90918	UBK-493404(Disponible 1-11 meses) - Sin posibil
transaction_type	object	0 (0.0%)	4	alquiler, buy, venta
n_rooms	float64	6301 (3.2%)	27	1.0, 3.0, 2.0
n_baths	float64	3264 (1.6%)	28	1.0, 2.0, 3.0

Figure 1: Partial visualization of what the function `general_overview` displays on the screen.

the dataset.

We also normalized text fields to address typos and matching errors, standardized dates for better manageability, and removed peripheral locations with low economic impact. A preliminary cleaning of `reference_ids` was performed to handle duplicates, which likely represented the same listing appearing at different points in time.

We applied one-hot encoding to categorical variables and mapped ordinal variables (such as `property_state`) to numerical values. Furthermore, we rationalized the geospatial data by discarding broad administrative features like `province` and `city`, retaining only `latitude` and `longitude` as they provided 100% data completeness and granular precision.

Finally, we made a major structural decision to segregate the dataset by transaction type. We discarded the rental data entirely to focus exclusively on sales, resulting in the creation of a new dataset: `madrid_housing_sale_preprocessed.parquet`. We approached data cleaning as a critical foundation of the work, executing it with meticulous attention to detail.

Finally, we made a major structural decision to segregate the dataset by transaction type. We discarded the rental data entirely to focus exclusively on sales, resulting in the creation of a new dataset: `madrid_housing_sale_preprocessed.parquet`.

2.2 Phase 2: EDA & Feature Engineering (Notebook 2)

In this phase, we continued data cleaning alongside Exploratory Data Analysis (EDA). Our first structural decision here was to narrow the scope strictly to residential housing, deferring the analysis of commercial premises and buildings to a future stage. This filtered the dataset down to approximately 70,000 residential units.

We performed feature engineering on the IDs, collapsing duplicate entries into single rows while generating new columns to capture price fluctuations over time. We proceeded with univariate and multivariate graphical analyses (including correlation heatmaps). Statistical metrics (mean,

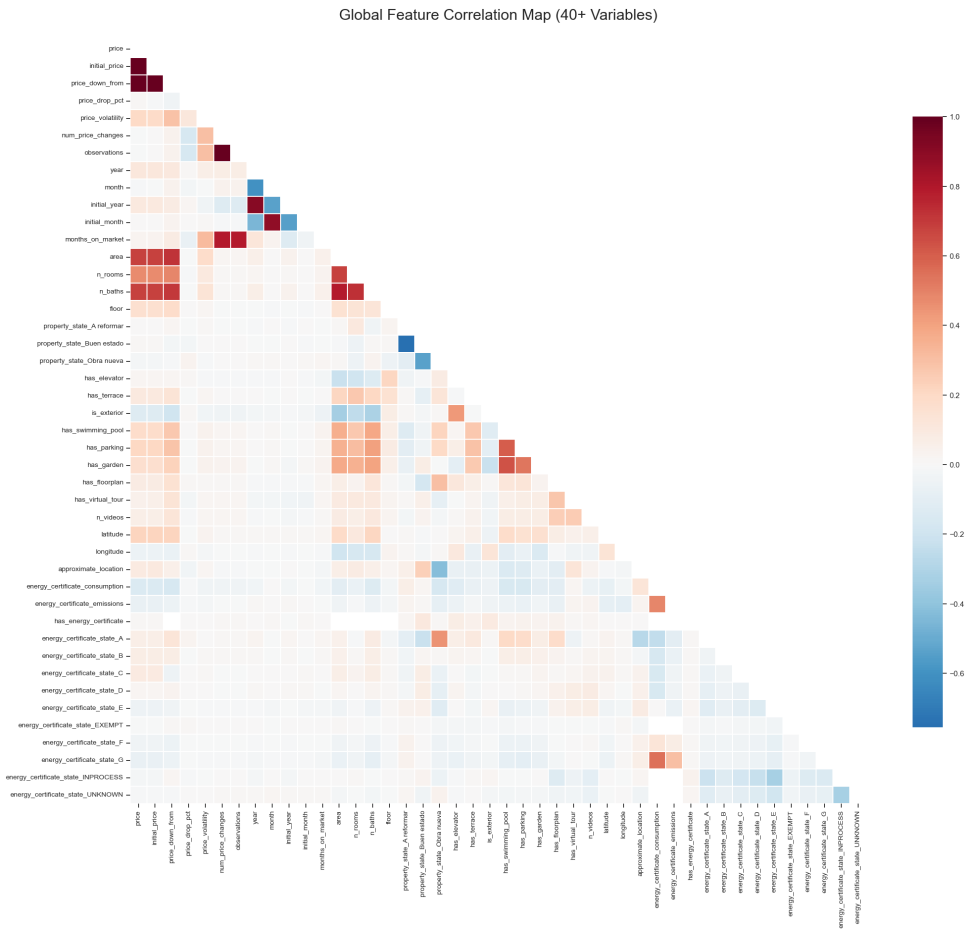


Figure 2: Correlation matrix of all the features. The purpose is to find useful correlations between features

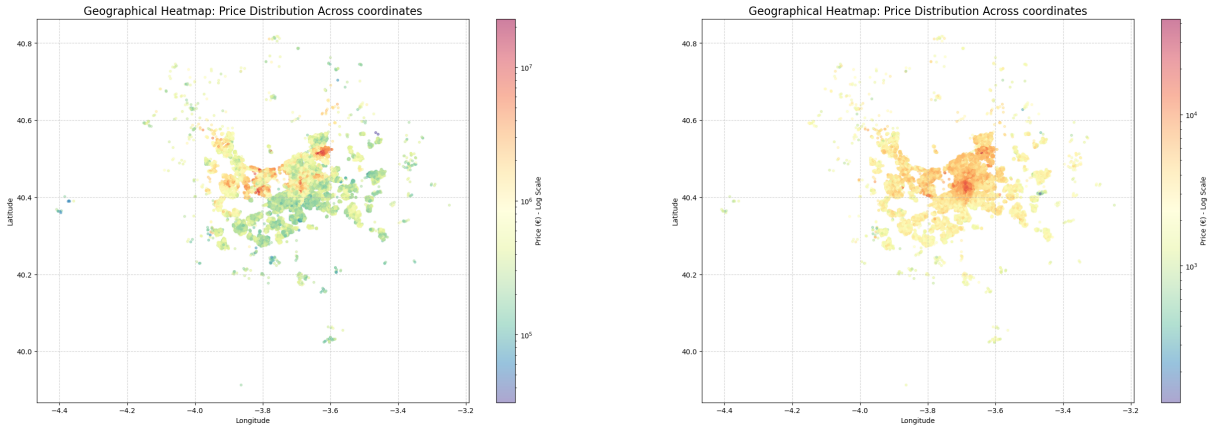
standard deviation, skewness, and kurtosis) were calculated to detect outliers that could distort predictions. Scatter plots were instrumental in identifying clusters, allowing us to move beyond administrative political boundaries, while box plots helped assess feature influence (See Fig. 3).

Finally, we advanced to Feature Engineering. We defined the target variable: *price per square meter*. We introduced new features, such as an amenity bonus (derived from box plot insights) and a bathroom-to-room ratio. Crucially, we utilized K-Means clustering to generate 50 distinct zones based on price behavior. This phase yielded the final dataset: `madrid_vivienda_sales_processed.parquet`.

2.3 Phase 3: Modeling & Evaluation (Notebook 3)

This stage focused on model selection and training. The first indispensable step was splitting the dataset into training, cross-validation, and test sets. To prevent data leakage, we temporarily removed variables that directly exposed price information. We introduced a smoothed benchmark variable (average price/m^2 per cluster) to aid in valuation.

We conducted a model comparison tournament. Neural Networks proved too slow compared to other candidates, with **XGBoost** emerging as the superior model. We performed hyperparameter tuning via Randomized Search, reducing the error by approximately 2%.



(a) Price heatmap

(b) Price per meter squared heatmap

Figure 3: Comparison between two target features candidates

Finally, we validated the model on the test set to identify undervalued "gems" within the real estate sector. We calculated the error on the test set and also applied the model to the training set. While searching the training set carries a higher risk of overfitting and potentially lower realizable gains, it still offers significant profit extraction opportunities.

Model Performance Validation

To guarantee the reliability of our valuations, the XGBoost model was subjected to a strict audit using the *Test Set* (data completely unseen during training). The results demonstrate a robust capacity to interpret market dynamics with high precision.

Metric	Result	Business Interpretation
R^2 Score	**[0.8656]**	The model explains approx. **[86]%** of the price variance based on objective data (location, size, quality).
MAE	**€[632.63]**	On average, our valuation deviates by only €[632.63]/m^2 from the market price.
RMSE	**€[1035.76]**	Indicates model stability. The proximity to MAE suggests we rarely make "huge" valuation errors.
Error %	**[13.33]%**	The relative margin of error. In Real Estate, a deviation under 15% is considered the industry standard for automated valuations.

Table 1: **Predictive Accuracy.** Performance metrics evaluated on the independent Test Set.

Strategic Implication (Margin of Safety): Since our model has an average error of approximately **[13]%**, our investment strategy requires a "buffer." We only recommend "Gems" that are undervalued by at least **10-15%** relative to our prediction. This gap acts as a statistical margin of safety to absorb the model's natural variance and protect capital.

3 Operational Conclusions

We have successfully identified a significant number of properties with high potential for rapid profit. A file containing all recommendations is attached, though these should be viewed as strategic leads rather than definitive decisions.

Key Figures:

- **Scope:** Out of 70,000 analyzed properties...
- **Selection:** We have highlighted nearly **500 strong candidates** (Gems).
- **Potential:** These assets present a significant maximum potential gain, subject to variance based on the algorithm's margin of error.

The identified "Gems" represent properties where the list price is significantly lower than the intrinsic value calculated by our XGBoost model, often accompanied by "motivated seller" signals (price drops).

% ==========

4 Future Work & Scalability

To expand upon this initial study, we propose two specific lines of action to improve the depth of the analysis:

4.1 1. Commercial Premises and Buildings

The next priority should be the analysis of the "Commercial" and "Buildings" datasets.

- **The Goal:** These assets offer a unique opportunity: transforming commercial spaces into residential units. This process significantly increases the property's value.
- **Technical Challenge:** Analyzing these properties requires specific feature engineering. We need to account for different variables—such as technical feasibility for conversion—which differ significantly from standard residential data.

4.2 2. Rental Market Trends

While this project focused exclusively on sales, integrating the rental market data would provide a complete picture of the sector.

- **Market Dynamics:** Rental prices often move faster than sales prices. By studying rental fluctuations, we can identify neighborhoods that are becoming popular before those trends are reflected in sales prices.

Access to Deliverables:
A comprehensive Excel dossier containing the complete inventory of identified "Gems" (including Train and Test set opportunities) is located in the project directory:

`madrid-real-estate-analysis/results-reports`

Note: The full Python code, trained model binaries, and the detailed list of property URLs and addresses have been attached to this submission.