From Data to Euros: Leveraging Data Science for Real Estate Valuation

Groupe:
GROUPE 3

Auteurs:
Lecene Maël, Traore Ali

Version du document:
1.0

ABSTRACT

Automated Valuation Models (AVMs) in real estate typically rely on structured tabular data (surface area, location, year built), often failing to capture the intrinsic value drivers hidden in unstructured assets like property images and textual descriptions. This project presents a ***Hybrid Multi-Modal Architecture*** *designed to bridge this semantic gap for the French real estate market.*

Our approach implements a two-stage distillation pipeline. First, we developed a custom Neural Backbone integrating ***ConvNeXt-Large*** *for visual feature extraction and* ***CamemBERT*** *for textual understanding, trained via a Masked Multi-Task Loss to handle disjoint rental and sales data. Second, we leveraged this backbone to extract high-dimensional latent embeddings, which were then fused with tabular features to train a* ***Gradient Boosting Regressor (XGBoost)****.*

We evaluated this architecture on a held-out dataset of over 18,000 listings. The results demonstrate State-of-the-Art performance on the rental market, achieving an R^2 *of* [redacted] *% and a Mean Absolute Error (MAE) of* [redacted] *€, effectively solving the estimation problem for standard assets. For the sales market, the model establishes a robust baseline (*R^2 *=* [redacted]*%), highlighting the complexity of high-variance luxury segments. Crucially, an ablation study confirms that the integration of neural embeddings reduces the estimation error by* [redacted]*% compared to a tabular-only baseline, quantitatively proving the existence of a "Visual Premium."*

Sommaire

1 Problem Definition and Scope

1.1 Problem Statement

This project focuses on the simulation and estimation of real-estate prices in France, for both rental and housing markets. The goal is to analyze property characteristics — such as location, surface area, number of rooms, and the general condition (based on photos) — and to integrate market data to provide reliable price predictions. This output can be used by anyone to evaluate whether they are overpaying for rent or to obtain an initial, data-driven estimate that is more accurate than existing tools. Our approach relies on supervised machine-learning models designed to establish correlations between property features and their estimated value.

1.2 Scope

To develop a State-of-the-Art data science solution for the French real estate market, a rigorous and comprehensive data acquisition strategy was paramount. Our scope focuses on the construction of a proprietary hybrid dataset, aggregating real-time listings from major national rental and sales platforms. This approach is strategic, designed to capture both the fluidity of current market expectations (listings) and the ground truth of realized sales. The selected source provide a granular set of attributes for each property—such as surface area, room count, energy efficiency ratings (DPE), and location —and, crucially, include unstructured data in the form of textual descriptions and image galleries. This multimodal dimension is essential for extracting latent features related to property condition and "standing," which are often absent from structured databases. The data acquisition phase necessitated the engineering of custom scraping architectures tailored to handle dynamic web content and varying layouts. This enabled the systematic extraction of diverse data fields and the implementation of strict data wrangling protocols. These protocols—focused on de-duplication, outlier detection, and consistency checks—ensure high data quality. Spanning Metropolitan France, from high-density urban hubs like Paris and Lyon to rural areas, this curated dataset constitutes the core empirical basis for our subsequent feature engineering and estimation tasks.

1.3 Work flow

To navigate the complexity of the real estate market, we structured our project around a standardized Data Science lifecycle pipeline. This workflow ensures reproducibility and allows for iterative improvements at each stage of the data transformation process.

Figure 2: Workflow in the project.

2 Data Gathering

2.1 Data Source and Collection Methodology

To construct our empirical dataset, we developed a custom scraping pipeline targeting LeBonCoin, the leading real estate platform in France. The data acquisition process was built using Node.js, leveraging the leboncoin-api-search library to interface directly with the platform's search engine.

The collection strategy was executed in two distinct phases to hermetically segment the target markets:

- The Rental Market (Category LOCATIONS).
- The Sales Market (Category VENTES_IMMOBILIERES).

To ensure national representativeness, the script iterates through the 13 administrative regions (e.g., Île-de-France, Nouvelle-Aquitaine, Grand Est). For each region, a targeted query using the keyword "appartement" is executed, and the results appear randomly to capture the entire pricing spectrum.

The data processing pipeline operates in three stages:

Parsing
Raw JSON responses are analyzed to extract over 35 distinct data points per listing. Raw strings are immediately normalized (e.g., converting "120 000 €" to integers, handling boolean flags for "has_phone").

Storage and Organization
Finally, the data are exported into CSV files using a dynamic naming convention: annonces*{Region}* {Type}.csv.
The suffix acts as a clear identifier, where *1 represents Rentals and* 0 represents Sales. To ensure a clean separation between the two market segments, the script automatically routes these files into two distinct directories: dataLbc/location for rentals and dataLbc/achat for sales transactions.

2.2 Image Acquisition & Standardized Pre-processing

To build the visual component of our multi-modal dataset, we developed a high-performance distributed crawling pipeline designed to handle hundreds of thousands of high-resolution images while ensuring architectural compatibility for deep learning.

2.2.a Distributed Asynchronous Ingestion

The acquisition process is orchestrated through a multi-processing architecture to overcome the dual bottleneck of network latency and disk I/O:

- **Parallel Workers:** The system utilizes a pool of CPU workers ($N_{\text{workers}} = N_{\text{cpu}}$) to handle simultaneous HTTP requests via persistent `requests.Session` objects.
- **Batch Processing:** Tasks are dispatched in batches of 100 images to minimize the overhead of inter-process communication and ensure a continuous stream of data to the storage unit.

2.2.b Structural Standardization:

Raw real-estate photos come in heterogeneous aspect ratios. To preserve the structural integrity of the buildings without introducing geometric distortion, we implemented a **Resize & Pad** strategy rather than a standard "Force Resize":

- **Geometric Preservation:** Images are rescaled using **LANCZOS** downsampling to fit within a 518×518 bounding box while maintaining their original aspect ratio.
- **Zero-Padding:** Remaining empty space is filled with a neutral black background $(0, 0, 0)$ to produce perfect squares.
- **Technical Justification:** The 518px resolution was specifically chosen to match the optimal patch-size requirements of the **DINOv2** Vision Transformer (ViT). Since DINOv2 operates on a patch size of 14, a dimension of 518 (14×37) ensures a perfect alignment without pixel interpolation at the patch boundaries.

2.3 Challenges

Constructing the final dataset required overcoming several challenges related to the heterogeneous nature of web data and the realities of the real estate market.

Regional Differences in Data Volume
Even though our script is set to retrieve up to 10,000 pages per region, we observed large differences in the number of ads depending on the location. The collection process did not stop due to technical errors, but simply because we ran out of available listings in less populated regions (like Corse or Normandie). In contrast, the supply in Île-de-France was much larger. This creates a natural geographic imbalance in our final dataset.

Pro vs. Private
Filtering Strategy A major constraint was managing data quality versus data volume. The initial scraping process retrieves listings from both Professionals (agencies) and Private individuals indiscriminately.

For Sales: With sufficient volume available, we applied a strict filter to retain only Professional sellers. This decision was driven by the observation that professional listings offer superior data quality: they provide higher-resolution images, and, crucially, significantly higher tabular data density (fewer missing values or "holes" in attributes like energy ratings or surface area) compared to private listings.

For Rentals: The inventory on the rental market being significantly lower, we were forced to retain both Professionals and Private individuals. Restricting this dataset to professionals alone would have reduced the sample size to a critical level insufficient for model training, obliging us to accept higher variance in data quality for this segment.

3 Dataset Description

The dataset comprises a substantial collection of real estate listings, totaling $N = 675,779$ records collected from the French market.

Distribution and Topology
The data exhibits a balanced distribution between transaction types, with **51.9%** ($n = 350,575$) representing properties for sale (Purchase) and **48.1%** ($n = 325,204$) representing rental listings. Regarding the property topology, the dataset is predominantly composed of **apartments (47.6%)** and **houses (38.1%)**, providing a comprehensive coverage of the core residential market. The remaining 14.3% consists of parking spaces, land plots, and other miscellaneous categories.

Multimodal Assets
To support multimodal learning tasks, the structured tabular data is complemented by a repository of approximately **3M images**.

Data Quality and Sparsity
We conducted a sparsity analysis to assess the completeness of the features (see Figure 3). Critical variables such as `price`, `location` (latitude/longitude), `surface_area`, and textual `description` exhibit **100% completeness**, ensuring a solid baseline for modeling.

However, secondary features show significant sparsity. For instance, energy-related metrics like `classe_energetique` are missing in **21.4%** of cases, while structural details like `orientation` (**82.3%** missing) or `nb_etages` (**98.3%** missing) are largely absent. These high-sparsity features will require specific preprocessing strategies, such as categorical imputation or exclusion, to avoid introducing noise into predictive models.

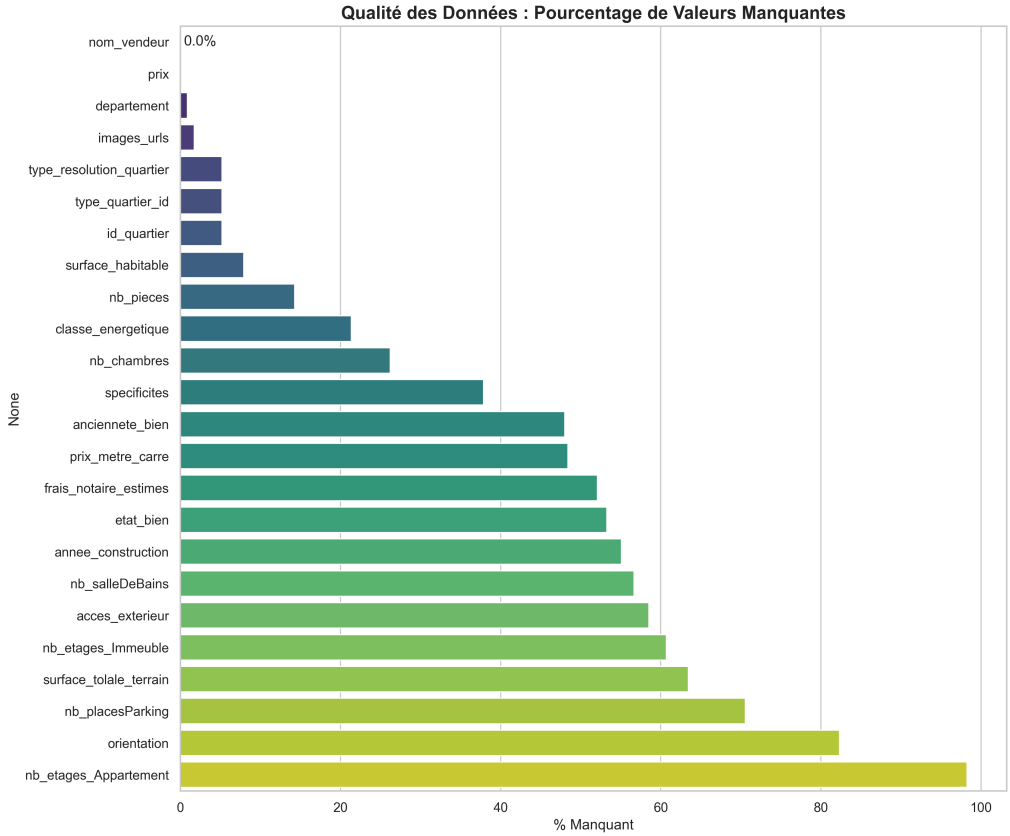


Figure 3: Feature sparsity analysis: Percentage of missing values per column.

4 Data processing

This section details the end-to-end pipeline applied to transform raw, noisy inputs into a clean, high-dimensional dataset suitable for machine learning. The processing is divided into two distinct parallel streams: the sanitization of structured tabular data and the orchestrated processing of unstructured visual assets.

4.1 Tabular Data Processing & Sanitization

Given the heterogeneous nature of web-scraped data, a strict "Sanitization Pipeline" was implemented to ensure the reliability of the supervised learning phase. To maintain high predictive fidelity, we categorize all ingested features into a structured hierarchy governed by an automated logic for missing values (NaN) and row retention.

4.1.a Feature Categorization Hierarchy

We have organized our internal mapping and processing logic into five strategic categories. This classification determines whether a listing is dense enough to be preserved or if it should be purged as "noise."

Category	Action	Rationale & Examples
1. Primary Predictors	**DROP ROW**	Non-negotiable features: `living_area_sqm`, `price`, `property_type`. Essential for any valid regression.
2. Secondary Features	**Pseudo-optional**	Quality indicators (e.g., `energy_rating`, `orientation`). If missing, we apply Median or Mode imputation later.
3. Boolean Logic	**FILL ZERO**	Amenities like `parking_spaces` or `exterior_access`. We assume absence in the text implies absence in reality (NaN $= 0$).
4. Analytical Meta	**KEEP**	Contextual data like `city` or `region`. Used for EDA but excluded from training to avoid redundancy with coordinates.
5. Noise / PII	**STRIP**	Irrelevant data (phone numbers, internal IDs) that offer zero predictive value and pose privacy risks.

Table 1: Systematic hierarchy for feature processing and data retention.

4.1.b The "Absolute Killers" (Primary Predictors)

Certain variables are mathematically fundamental to real estate valuation. Features such as `living_area_sqm`, `property_type`, and `price` act as the structural pillars of our model. If a listing is missing any of these "Absolute Killers," it is immediately purged. We prioritize **precision over volume**, as imputing a missing surface area or property type would introduce synthetic variance that could degrade the model's sensitivity to price-per-square-meter trends.

4.1.c Data Density Gating (The Multi-Missing Rule)

To prevent the model from learning from "sparse" or "hollow" listings, we implement a strict **Quality Gate** based on cumulative information density:

- **Strict Constraint:** While a single missing value in Category 2 (e.g., `year_built`) is tolerated, a listing is discarded if more than **two** features from this category are missing simultaneously.
- **Justification:** Listings with multiple missing attributes are typically incomplete drafts or low-quality entries. By enforcing this density threshold, we ensure the model trains on holistic profiles rather than fragmented data points.

4.1.d Geographic Representation: Coordinates vs. Labels

A deliberate architectural choice was made to rely on `latitude` and `longitude` for the predictive engine while reserving `region` and `city` for post-inference analysis.

- **Continuous Signal:** Spatial coordinates provide a high-resolution, continuous signal. This allows the model to capture micro-market variations (e.g., the value difference between two adjacent blocks) that discrete administrative labels like "City" or "Department" would aggregate and obscure.
- **Noise Reduction:** Removing administrative labels from the training set reduces the dimensionality of the tabular input, forcing the model to learn the underlying spatial geography through the continuous coordinate plane.

4.1.e Domain-Specific Outlier Filtering

To prevent the model from learning noise, we applied strict filtering rules derived from French real estate domain knowledge. These rules differ by property type to respect structural realities:

- **Physiological Constraints:** We exclude properties with unrealistic dimensions, such as apartments under $10m^2$ (approaching the legal definition of "indecency" in France) or houses exceeding $1500m^2$ (outliers representing castles or commercial complexes).
- **Quality Heuristic (Image/Room Ratio):** We introduced a derived quality metric: $N_{\text{images}} \geq N_{\text{rooms}}$. Listings with fewer images than rooms are statistically more likely to be low-quality entries, scams, or incomplete drafts, and are thus discarded.

4.1.f Prevention of Target Leakage in Textual Data

A major challenge in real estate price prediction is "Target Leakage" within textual descriptions. Agents often explicitly state the price, monthly payments, or rental yield in the description field. If the model processes this text (via NLP), it will "cheat" by reading the answer rather than predicting it. To mitigate this, we developed a rigorous **Regex-based Sanitization Protocol**:

- **Leakage Removal:** All sentences containing monetary patterns (e.g., "400€", "loyer", "honoraires", "net vendeur") are systematically excised.
- **PII Scrubbing:** To comply with GDPR privacy standards, patterns detecting phone numbers, emails, and SIRET numbers are stripped using pre-compiled regular expressions.
- **Noise Reduction:** Marketing fluff and administrative jargon (e.g., "géorisques", "loi ALUR", "honoraires agence inclus") are removed to reduce the dimensionality of the text vector space.

4.2 Unstructured Data Processing: The Computer Vision Pipeline

While tabular data provides explicit metrics, the intrinsic value of a property is often encoded in visual features (lighting, condition, renovation quality) that are difficult to quantify manually. To extract these latent signals without introducing noise or redundancy, we developed a high-performance **Orchestrated Vision Pipeline**.

4.2.a Zero-Shot Semantic Filtering (CLIP)

Real estate datasets are notoriously noisy, containing irrelevant images (floor plans, agency logos, maps) that confuse predictive models. Instead of training a custom classifier, we leveraged **CLIP** for **Zero-Shot Semantic Filtering**. We compute the cosine similarity between the image embedding E_i and text prototypes E_t (e.g., "a floor plan", "a technical drawing", "indoor scene").

- **Protocol:** Images are discarded if their similarity to "Bad Content".
- **Result:** This effectively sanitizes the dataset using the broad semantic knowledge inherent in CLIP's latent space, without requiring manual labels.

4.2.b Visual Embedding (DINOv2)

Once filtered, images must be converted into vector representations. We selected **DINOv2 (Self-distilled Vision Transformer)** over standard ResNets or CLIP for this task.

- **Justification:** Unlike CLIP (semantic focus), DINOv2 is trained via **Self-Supervised Learning** with a focus on local geometry and texture. It excels at capturing structural details (e.g., parquet condition, window size) crucial for valuation, rather than just high-level categorization.

4.2.c Redundancy Reduction via Dynamic Hierarchical Clustering

A major bias in real estate data is the imbalance of visual documentation (e.g., 20 "burst-mode" photos of the living room vs. 1 of the bathroom). Feeding all images creates a bias towards the most photographed room. To solve this, we implemented a **Dynamic Agglomerative Clustering** strategy:

1. **Metric-Based Grouping:** Instead of forcing a fixed number of clusters (like K-Means), we use **Agglomerative Hierarchical Clustering** with a strict cosine distance threshold ($\tau = 0.25$).
2. **Automatic Diversity Discovery:** The algorithm automatically determines the number of clusters K based on visual content. If a property has 3 distinct rooms, it finds $K = 3$. If it has 10 unique angles, it finds $K = 10$. This adapts naturally to the property's size without human heuristics.
3. **Centroid Selection:** For each resulting cluster, we calculate the mathematical centroid in the DINOv2 latent space and retain the single image closest to it.

Result: This pipeline transforms a noisy stream of duplicate photos into a **Canonical Visual Set**, ensuring the final estimator receives exactly one high-quality representative image per distinct viewpoint.

5 Implementation & Technical Stack

The complexity of the Hybrid Architecture, combining high-dimensional tensor operations with massive tabular processing, required a robust High-Performance Computing (HPC) environment.

5.1 Software Ecosystem

The pipeline is built upon the Python 3.10 ecosystem, leveraging State-of-the-Art libraries for each modality:

- **Deep Learning Core:** We utilized **PyTorch** to benefit from the latest graph optimizations.
  - `torch.compile`: Enabled Just-In-Time (JIT) graph compilation to fuse kernels and accelerate the backbone forward pass.
  - `torch.amp` (Automatic Mixed Precision): Used to perform operations in FP16/BF16 where possible, reducing VRAM usage and doubling throughput on Tensor Cores.
- **Model Zoos:**
  - `timm` (PyTorch Image Models) for the ConvNeXt-Large backbone.
  - `transformers` (Hugging Face) for the CamemBERT tokenizer and encoder.
- **Tabular & Boosting:**
  - `xgboost` with the `gpu_hist` tree method for hardware-accelerated gradient boosting.
  - `pandas` and `numpy` for vectorized pre-processing of the 20+ tabular features.

5.2 Hardware Infrastructure (Sorbonne Cluster)

Training was conducted on the Sorbonne University PPTI computing cluster. Access to NVIDIA RTX 3080 GPUs via SSH tunneling.

5.3 Computational Optimizations

To maximize hardware utilization (GPU occupancy), we implemented specific engineering patterns:

- **Asynchronous Data Loading:** Usage of `multiprocessing` with workers to decouple CPU-bound tasks (image decoding, augmentation) from GPU-bound training.
- **Memory Pinning:** `pin_memory=True` in DataLoaders to use page-locked memory, accelerating the Host-to-Device transfer over the PCIe bus.
- **Garbage Collection Strategy:** Explicit invocation of `gc.collect()` and `torch.cuda.empty_cache()` between the Neural Extraction phase and the Boosting phase to prevent Out-Of-Memory (OOM) errors during the hybrid concatenation.

6 Exploratory Data Analysis (EDA)

Sale prices and monthly rents operate on vastly different orders of magnitude. Merging them would result in a highly skewed, bi-modal distribution, making it impossible to perform meaningful variance analysis or outlier detection on the rental segment. Separating the data allows for proper feature scaling and more sensitive model training.

Market dynamics for “Selling” vs. “Renting” are governed by different economic factors.

6.1 Univariate Analysis

Before diving into the distribution of features, we performed a preliminary feature selection. purely administrative variables—such as unique identifiers (`id`) and high-cardinality location codes (`zip_code`)—were excluded from this statistical overview but we also remove the notary fee, since it is calculated based on the price, so including it would be a data leak. These features serve as indexical references rather than quantitative market signals, and were therefore removed to focus the analysis on variables with genuine predictive potential.

- Target Variable: Price

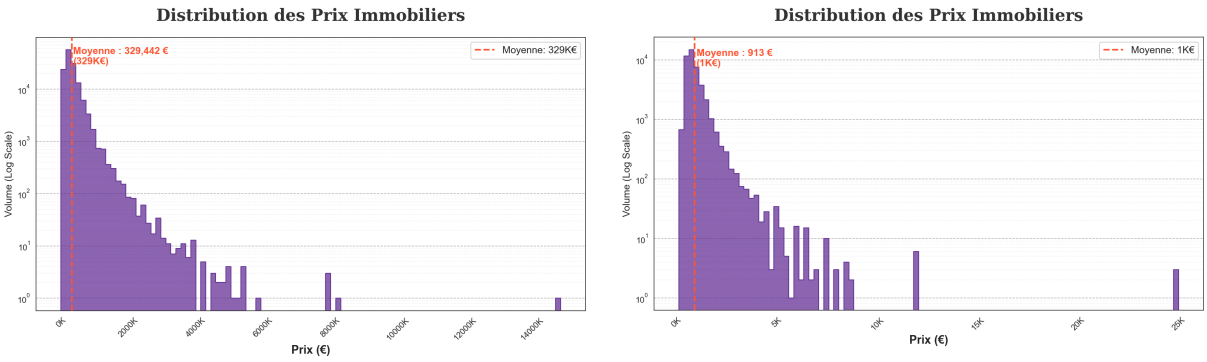


Figure 4: Distribution comparison: sale Price (left) vs. rent Price (right).

Observation

These plots reveal a highly skewed distribution: the vast majority of the market is concentrated at the lower end, while a few rare, ultra-expensive outliers artificially pull the average up.

The striking structural similarity between the two distributions (both Log-Normal) is ideal for Multi-Task Learning, as it implies that shared latent features drive both Sales and Rent prices via similar geometric laws. However, due to the massive scale difference (10^5 vs 10^3), we must apply a Log-transformation to both targets.

6.1.a Purchase Data

6.1.a.a Numerical Features

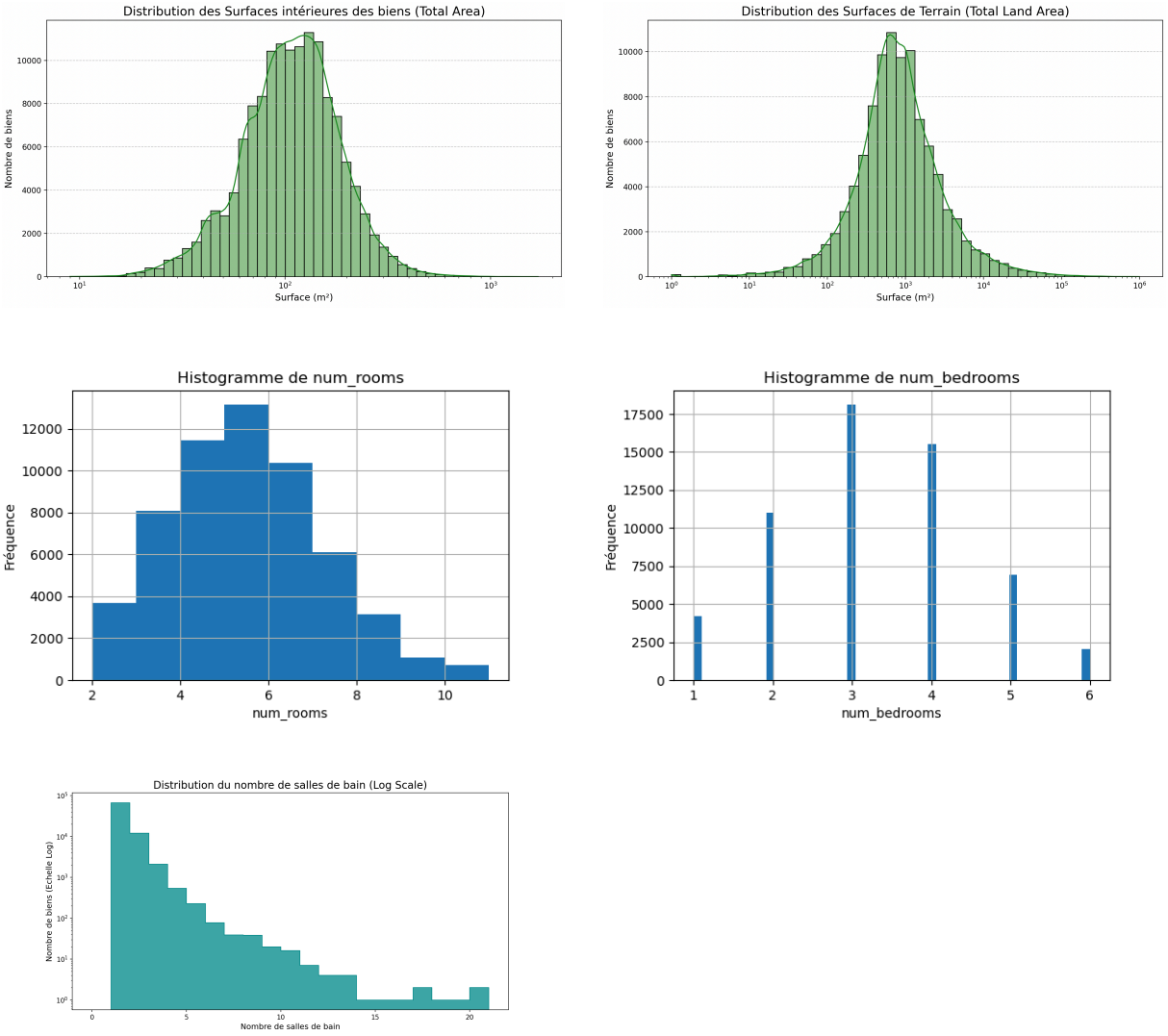


Figure 5: Distribution of Physical Properties.

An inspection of the structural data in Figure 5 reveals several key characteristics of the sampled population:

- **Discrete Modality (Room Counts):** The distribution of `num_rooms` is characterized by sharp, discrete peaks at integer values. The highest frequency is observed for standard units (2 to 4 rooms), which represent the majority of urban and suburban listings. The "staircase" nature of this plot confirms that the dataset is composed of standardized housing units rather than bespoke or highly irregular architectural projects.
- Heavy-Tailed Distribution (Land Area): The land area attribute follows a Pareto-like distribution with extreme right-skewness. A significant mass of the data is concentrated near zero—corresponding to the apartment segment (45.2% of the total)—while a long tail represents large rural estates. This high variance indicates a vast disparity in property footprints within the French territory.
- Data Consistency and Integrity: The absence of values at the extreme zero-bound (e.g., properties with 0 rooms) or physically impossible ratios validates the effectiveness of the initial data sanitization. The resulting population is statistically coherent, representing a clean cross-section of legitimate real estate entries.
- Structural Correlation: The high density of properties with minimal land area is consistent with the significant representation of collective housing (apartments) in the dataset. This alignment between the `property_type` and `total_land_area` variables ensures the internal logical consistency of the statistical sample.

6.1.a.b Categorical Features

In this section, we explore the distribution of categorical variables to identify market trends and data imbalances that could affect our model's performance.

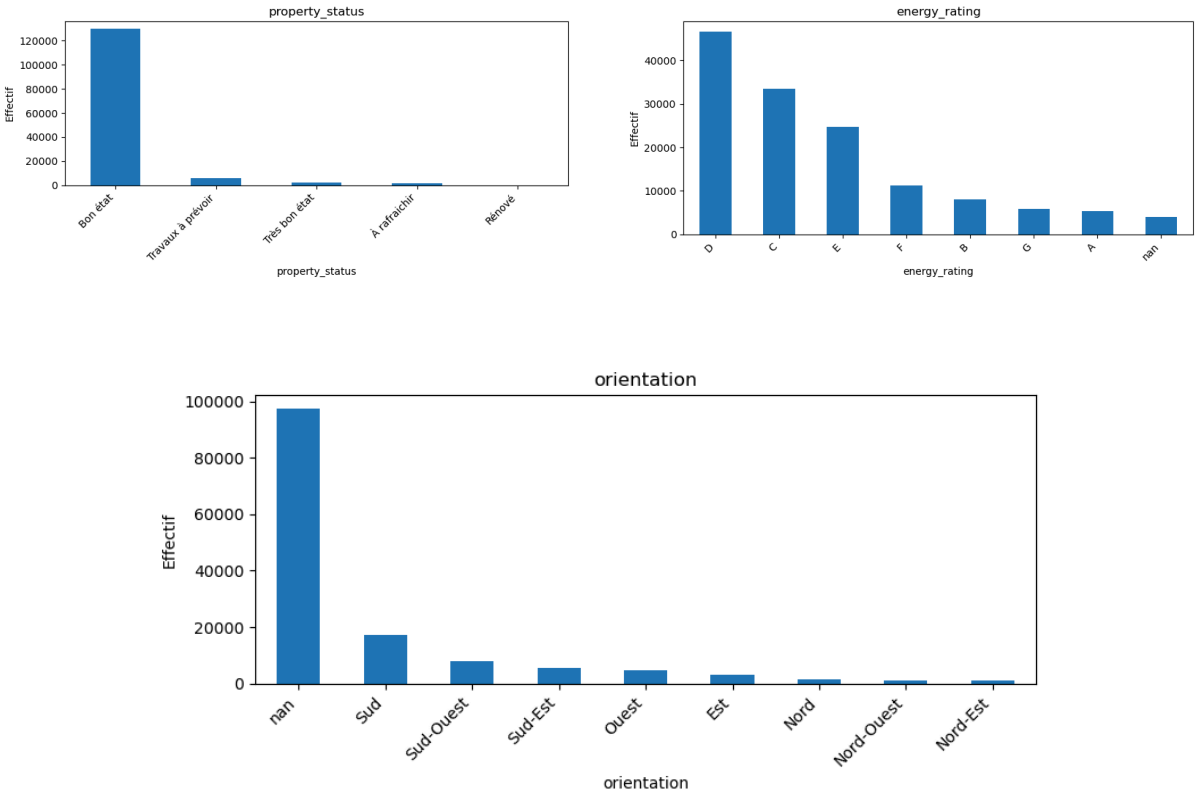
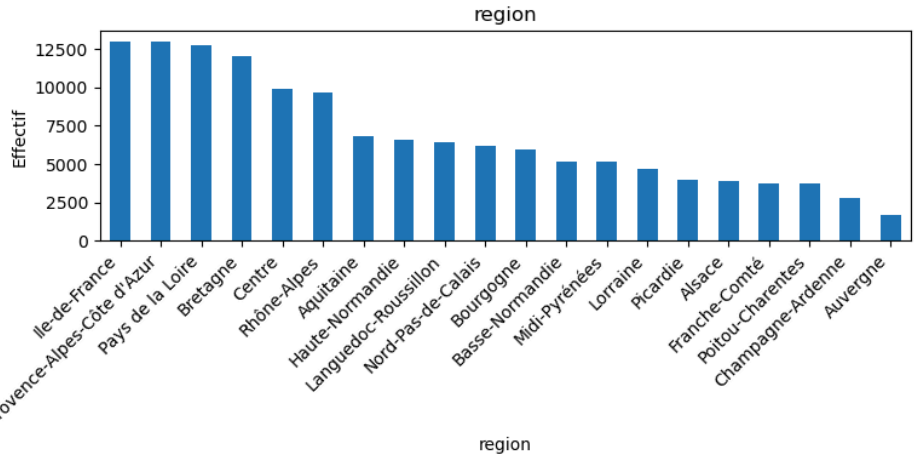


Figure 6: Distribution of Categorical Features. Top-Left: Property Status (mostly "Good Condition"). Top-Right: Energy Rating (Peaking at D/C). Bottom: Orientation (Showing a clear reporting bias toward South/West).

The distribution of missing data highlights three distinct patterns in real estate documentation:

- Legal Determinism: Features like `energy_rating` (2.68% missing) show near-perfect completeness. This is a direct reflection of the **Diagnostic de Performance Énergétique (DPE)** being mandatory for French real estate transactions, ensuring a high-density signal for environmental attributes.
- Descriptive Sparsity (The "Optional" Bias): Attributes such as `orientation` (72.01%) and `num_bathrooms` (40.58%) exhibit high sparsity. Statistically, this suggests a reporting bias: agents tend to mention these features primarily when they constitute a competitive advantage (e.g., a "south-facing" terrace or "multiple bathrooms"). The absence of these values is therefore not random (MNAR - Missing Not At Random).
- Information Density vs. Quality: The low missing rate for `special_features` (14.38%) compared to structural data like `year_built` (18.06%) indicates that qualitative marketing descriptions are often more complete than technical historical records.



Geography: During the scraping process, we aimed to collect a balanced dataset across 10 target regions. However, Île-de-France is the most represented region, followed by Provence-Alpes-Côte d'Azur and Rhône-Alpes.

This imbalance is not a scraping error but reflects the structural reality of the French real estate market: Market Density: Île-de-France has a significantly higher volume of transactions and rental turnover compared to other regions.

- **Listing Availability**: Our scraping methodology exhausted available listings in smaller regions, whereas the inventory in Île-de-France was much deeper.
- **Modeling Implication**: We deliberately kept all gathered data to maximize model training. We acknowledge that the model may have a slightly higher predictive performance on Île-de-France properties due to this volume advantage.

6.1.b Rental Data

6.1.b.a Numerical Features

- Target Variable: Monthly Rent

Price Distribution: The rental prices are tightly clustered compared to sales prices.

Observation: The skewness (1.86) is lower than for sales prices, indicating fewer extreme outliers. The market is capped by tenant solvency, unlike the purchase market which is driven by asset accumulation.

- Property Structure: The “Shift to Small”

Living Area: The dataset reveals a massive structural shift.

- **Purchase Median Surface:** 105 m²
- **Rental Median Surface:** 50 m²

Interpretation: The rental market is structurally dominated by T2/T3 units (median 2 rooms) catering to singles and young couples. The “Family House” market (median >100m²) is almost exclusively a sales market.

- Land Area (Terrain):

As expected for a market dominated by apartments, the median total land area is 0 m². Only outliers (houses) possess land, confirming that “Garden/Land” is a premium feature in rentals.

6.1.b.b Categorical Features

- Property Type Imbalance

Dominance of Apartments: The contrast with the sales market is striking: **Sales: Balanced market (55% Houses / 45% Apartments). Rentals: (95% Apartments vs only 5% Houses)** . Modeling Implication: The variable `property_type` has extremely low variance in the rental dataset. It might be less predictive here than in the sales model, or it might act as a proxy for “Luxury/Niche” listings.

- Property Status & Quality

Missing Data Warning: The `property_status` feature contains $\approx$ **80% missing values** (NaN) in the rental dataset. Hypothesis: Landlords rarely specify “Work needed” for a rental, as the property must be legally decent to be leased. This variable is likely unusable for the rental price model.

- Energy Rating (DPE)

Distribution: The distribution centers on the **D rating** (approx. 40% of listings), followed by C. Comparison: A and B ratings (highly energy efficient) remain a minority (11% combined). This suggests the “Green Value” has not yet fully penetrated the rental stock available on the market.

- Geography

Regional Focus: Île-de-France remains the top region (12.5%), but the rental market is more evenly distributed across major hubs (Rhône-Alpes 11.8%, PACA 9%). This reflects the rental demand being concentrated in major employment centers regardless of the region.

6.2 Bivariate Analysis

6.2.a Numerical Interactions:

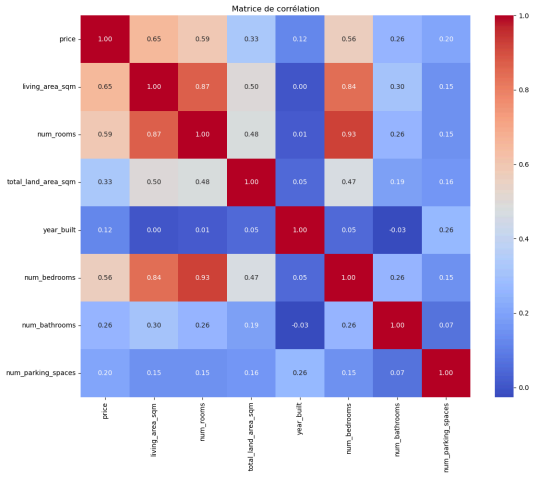


Figure 8: Correlation Matrix: Rental Ecosystem.

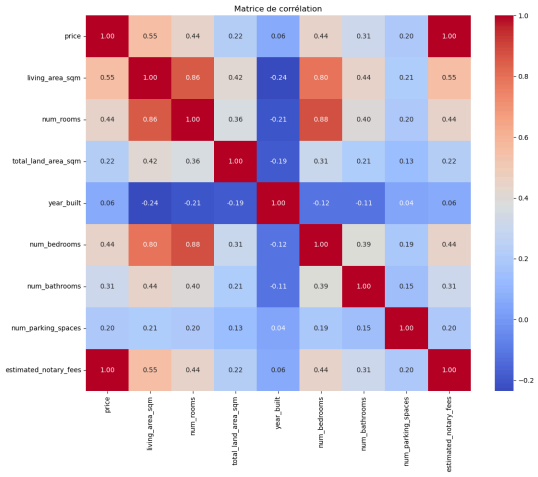


Figure 9: Correlation Matrix: Purchase Ecosystem

The Core Correlation

Similar to the rental market, Living Area (living_area_sqm) remains the strongest predictor (Correlation: **0.65**), followed by the number of rooms (**0.44** in rental and **0.59** in purchase).
However, distinct preferences appear in the secondary variables, highlighting the difference between a functional need (renting) and an asset acquisition (buying):

The Amenity Premium: The correlation with parking spaces (num_parking_spaces) is noticeably stronger in the purchase market (**0.23**) compared to rentals (**0.15**), reflecting the demand for complete, long-term assets.

A critical observation across both matrices is the limit of physical attributes. Excluding the artificial correlation of estimated_notary_fees—which shows a perfect 1.00 score in the purchase matrix simply because fees are a mathematical percentage of the price—no variable exceeds a correlation of **0.70**.

This indicates that while physical dimensions (size, rooms) define the baseline price, they do not fully explain it. A significant portion of the variance is likely driven by unobserved qualitative factors such as location, property condition, or energy efficiency, which are not captured in these quantitative columns.

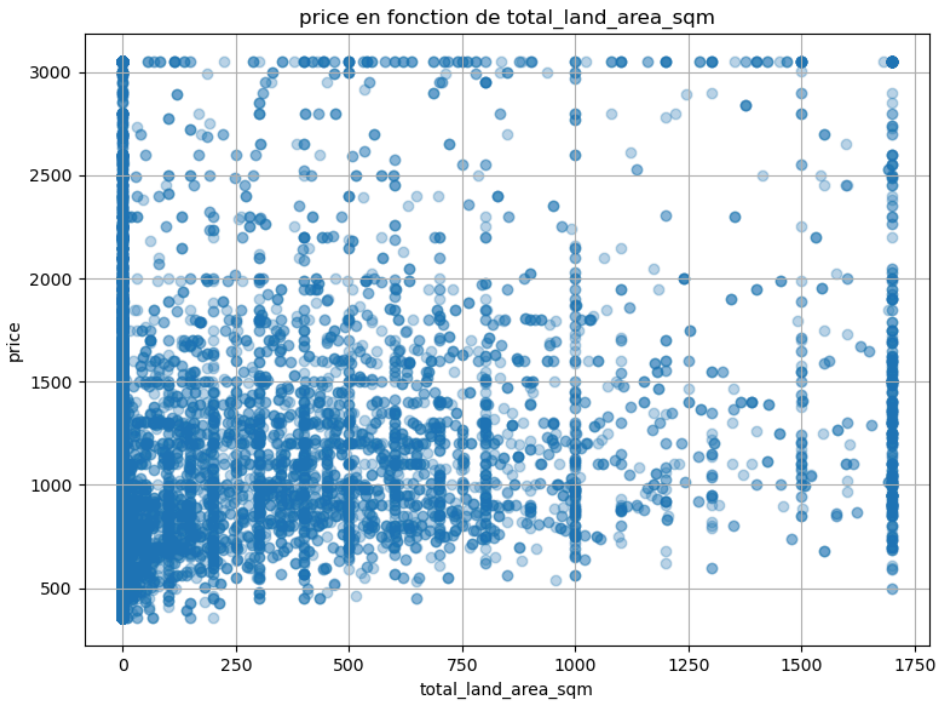


Figure 10: Scatter plot total land area vs price.

Scatter Plot Analysis (total_land_area)
The plot reveals that the variability in price increases with the total_land_area. As the surface area increases, the price variance explodes. A 20m² studio has a predictable price range, whereas a 200m² property could be a rural farmhouse (€300k) or a Parisian mansion (€5M).

The Land Area Paradox (total_land_area): The correlation is weak (**0.23**), and the scatter plot explains why. We observe two distinct behaviors:

The "Zero" Cluster: A massive concentration of points at land_area = **0**, representing the apartment market.

The Distributed Cloud: For houses, a larger land area does not strictly guarantee a higher price (location outweighs size).

To capture these "unobserved" factors, we turn to categorical analysis using **ANOVA (Analysis of Variance)**. We test three key qualitative variables—Energy Rating, Orientation, and Region—to measure their statistical impact on price across both market segments.

6.2.b ANOVA

Variable (F-statistic)	Purchase	Rental	Dominant Impact
Energy Rating	**796.42**	121.48	Critical for Sales
Orientation	**64.16**	14.91	Buyer Preference
Region	298.22	**301.92**	Critical for Rentals

Table 2: Comparison of ANOVA scores (F-statistic) between Sales and Rental markets.

Variable (P-values)	Purchase (p)	Rental (p)	Significance
Energy Rating	≈ 0	2.57×10^{-127}	Significant
Orientation	8.8700×10^{-92}	1.66×10^{-19}	Significant
Region	≈ 0	≈ 0	Significant

Table 3: Statistical significance (P-values). ("significant" indicates $p < 0.001$)

Methodological Note: Variable Selection
Variables with high cardinality, such as **Department** (more than 100 categories) and **City** (also a lot), were excluded from this global ANOVA analysis. While these variables offer granular precision, they suffer from **data sparsity** in rural areas (insufficient sample size per group to calculate a reliable mean). The **Region** variable was selected as the robust proxy to capture macro-geographical variance without introducing noise from under-represented localities.

The comparative ANOVA reveals a fundamental structural divergence between the two markets:

1. **The Asset Value (Purchase Market):** The transaction market is heavily driven by the intrinsic quality of the building. The **Energy Rating** is the single most discriminatory factor ($F = 796.42$), far outweighing location. Buyers, who bear the long-term costs of renovation and resale, heavily penalize technical defects. Similarly, **Orientation** has a much stronger impact on purchase prices ($F = 71.89$) than on rentals, reflecting a "quality of life" premium for owner-occupiers.
2. **The Location Value (Rental Market):** Conversely, the rental market is dominated by geography. **Region** is the primary determinant of rent ($F = 301.92$), exerting twice the statistical influence it has on purchase prices. Tenants are primarily constrained by the location (proximity to work/study), making the specific condition of the property secondary to its address.
3. **Statistical Validity:** The P-values for all tested variables are infinitesimal ($p \approx 0$), confirming that these price differences are not random but represent robust, systemic market trends.

6.2.c Categorical Analysis: Location is King

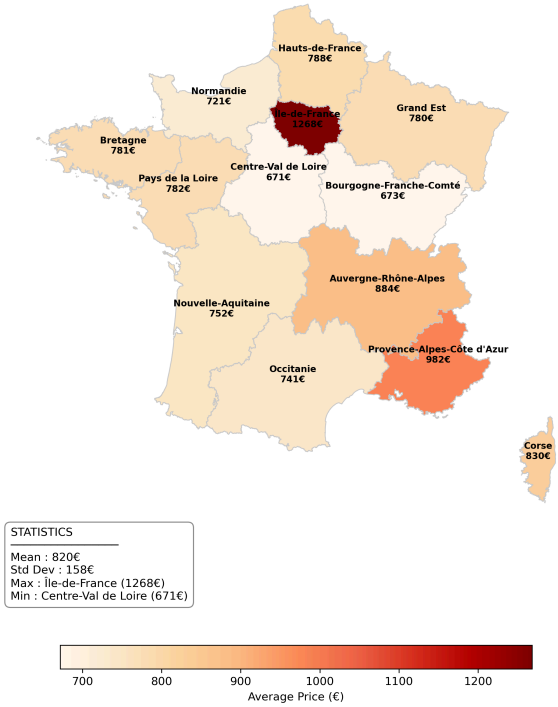


Figure 11: Average Price by Exterior Access

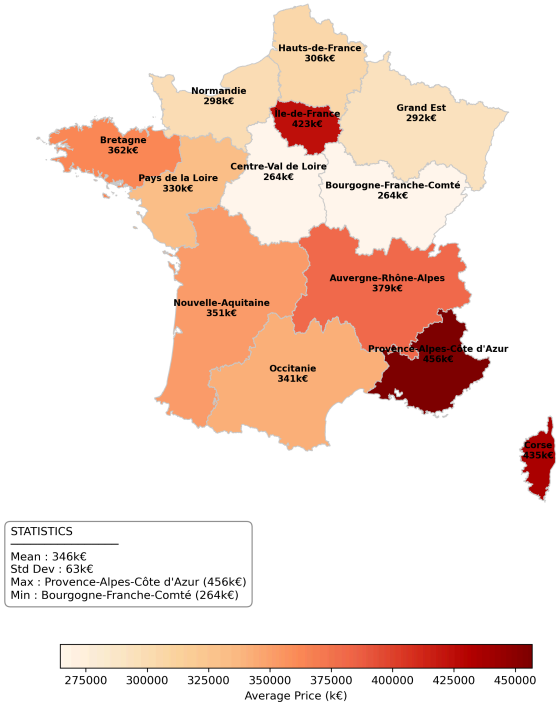


Figure 12: Average Price by Exterior Access

Regional Dynamics
Unlike the rental market, which is strictly centered around the economic capital (Île-de-France), the sales market reveals a shift toward "Lifestyle" geography.

- **The PACA Premium:** The Provence-Alpes-Côte d'Azur region records the highest average transaction price (**456k€**), surpassing even Île-de-France (**423k€**). This confirms a distinct market segment driven by high-value secondary residences and villas in the South, whereas the Parisian market is voluminous but constrained by smaller surface areas.

6.2.d Multilabel Analysis: Differentiating "Comfort" vs. "Value"

To capture more correlations, we analyze two distinct categories of features corresponding to specialized feature sets.

6.2.d.a Exterior Access

We distinguish these features because they fundamentally change the usage of the property.

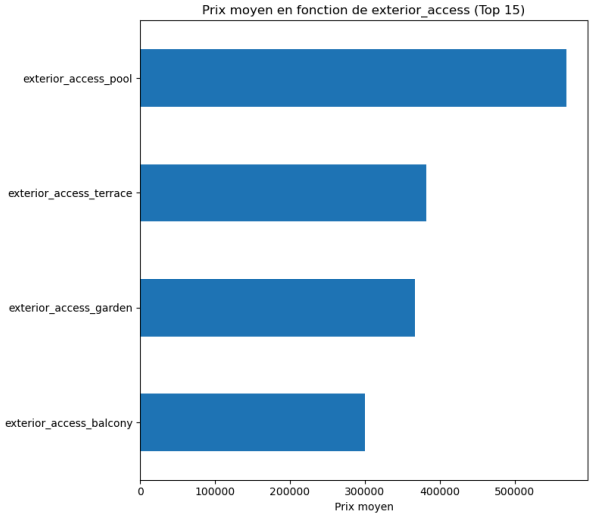


Figure 13: Average Price by Exterior Access

The Wealth Proxy The presence of a Pool (exterior_access_pool) correlates with the highest average prices. A pool is rarely sold alone; it is a proxy for a large land area, a house, and high standing.

Outdoor Hierarchy There is a clear value gradation: **Pool > Terrace > Garden > Balcony.**

6.2.d.b Special Features

Functional vs Structural, These features relate to the building's standing and amenities.

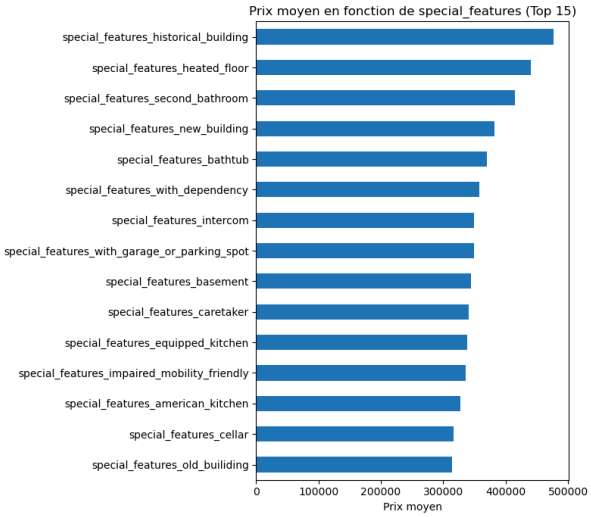


Figure 14: Average Price by Special Features

Prestige and Character (Historical Building): This is now the primary indicator of value (480k€). It reflects the significant premium placed on architectural character and rarity (often linked to historic districts or Haussmannian styles), which now supersedes pure surface area.

Modern Luxury and Comfort: heated_floor (2nd position) and second_bathroom (3rd position) are strong markers of high-end specifications. Underfloor heating suggests expensive recent renovations or luxury new builds, while a second bathroom remains the ideal proxy for large surface areas (Family Property).

Standard Amenities: Conversely, features like equipped_kitchen, intercom, or cellar sit in the lower half of the chart. These have become market standards: their presence is considered "normal" and does not trigger a massive price premium compared to the exceptional assets mentioned above.

While rental units are often standardized, properties for sale are frequently defined by unique architectural features and access to private exterior. Unlike rentals, which are more driven by functional needs (proximity to work, immediate solvency), the purchase market is driven by asset accumulation, emotional projection , and long-term borrowing capacity.

Modeling Takeaway: For the purchase prediction, our model must prioritize Surface and Region. However, to predict the price of the top 20% (the most expensive properties), the boolean flags for Pool, 2nd Bathroom, and Caretaker will be essential discriminators.

6.3 Multivariate Analysis: Structural Comparison (PCA)

To uncover the hidden hierarchy of price drivers, we performed a Principal Component Analysis (PCA) on both the Rental and Purchase datasets.

Comparing the two analyses reveals a fundamental structural divergence: while the rental market is functionally streamlined, the purchase market is multi-dimensional and geometrically complex.

6.3.a Variance & Dimensionality: The Complexity Gap

Low Dimensionality: The market is "flatter." The first two components (PC1 & PC2) capture the vast majority of the signal. Tenants' decisions are primarily driven by immediate functional needs (Surface & Standing).

Here are the differents axes from PCA order by the explained variance.

Axis 1: 38.28%	**Axis 6:** 8.03%
Axis 2: 14.69%	**Axis 7:** 5.88%
Axis 3: 11.12%	**Axis 8:** 2.09%
Axis 4: 10.64%	**Axis 9:** 0.72%
Axis 5: 8.55%	

Table 4: Variance explained by each principal component (PCA).

6.3.b The Universal Baseline: Volume & Modernity (PC1 & PC2)

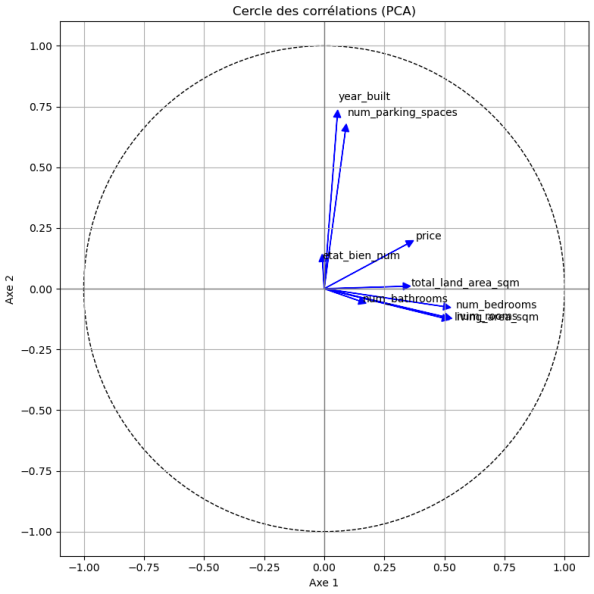


Figure 15: pca_correlation_circle_axes_1_2

The first two axes continue to define the fundamental value of real estate, but with a stronger consolidation on Axis 1.

PC1 (Modernity): Driven almost entirely by year_built (and to a lesser extent etat_bien_num).The number of parkings spaces is also important here. This separates old builds (bottom) from new constructions (top).

PC2(Global Capacity): Strongly correlated with living_area_sqm, num_rooms, num_bedrooms, and total_land_area_sqm.

The Price Vector: The price vector points to the Top-Right Quadrant. This confirms the universal rule: value is maximized by increasing Volume (Right) and Modernity (Up).

6.3.c he Structural Divergence: Verticality vs. Accessibility (PC3)

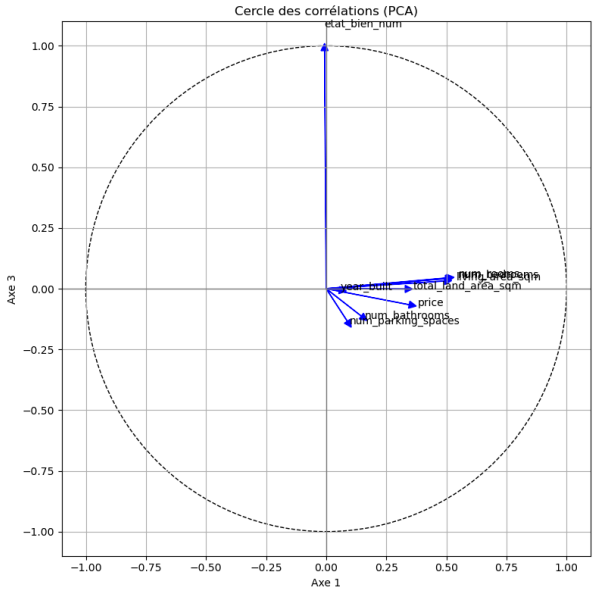


Figure 16: pca_correlation_circle_axes_1_3

This represents the most significant change is that Axis 3 is is about Quality.

Up (High PC3): Dominated by etat_bien_num (Property Condition).

Orthogonality: Notice that the etat_bien_num vector is nearly perpendicular (orthogonal) to living_area_sqm (Axis 1).

Interpretation: The condition of a property is mathematically independent of its size and it's price. A property can be huge but in poor condition, or small and flawless.

6.3.d The Qualitative Independent: Intrinsic Condition (PC4)

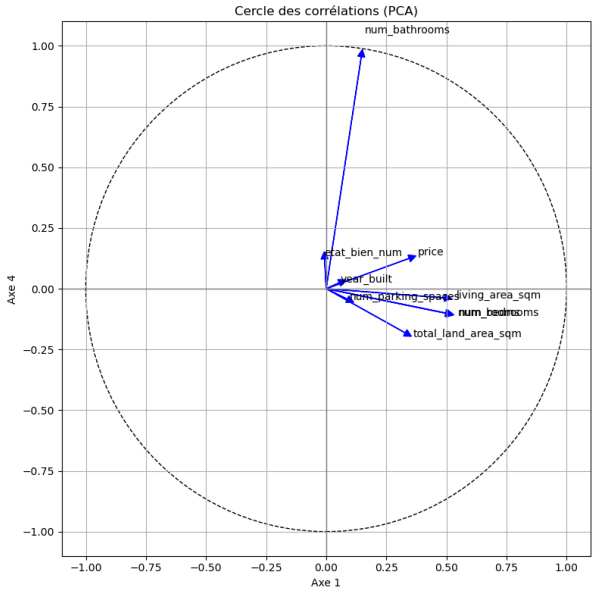


Figure 17: pca_correlation_circle_axes_1_4

Right (High PC1): Dominated by Volume & Size (living_area_sqm, num_rooms, num_bedrooms, total_land_area).

Up (High PC4): Dominated by Sanitary Facilities (num_bathrooms). Note that etat_bien_num (Condition) is negligible on this specific axis.

Orthogonality: The num_bathrooms vector is nearly perpendicular to the living_area and num_rooms vectors.

The price vector sits diagonally, perfectly wedged between living_area (Right) and num_bathrooms (Up).

Interpretation: Price is a compound result. While size provides the baseline value (Axis 1), num_bathrooms acts as the "value booster". This suggests that adding bathrooms (modernity/luxury) is the key factor that lifts a property's price above what its size alone would predict.

This multivariate analysis dictates our strategy for the Feature Engineering and Modeling phases. Numerical features alone explain only 50-60% of the variance. The remaining price drivers are currently "invisible."

One-Hot Encoding is mandatory. We must convert these text categories into machine-readable format to capture the missing half of the predictive power..

6.4 Outlier Detection & Data Cleaning

Real estate data is naturally prone to extreme outliers due to the heterogeneity of the market (e.g., luxury estates vs. ruins) and potential scraping errors (e.g., a year built of "0" or "20250"). Before training, it is imperative to sanitize the dataset to prevent the model from overfitting on anomalies.

6.4.a 1. Methodology: The Quantile Strategy

To address this, we applied a statistical filtering approach combining Interquartile Range (IQR) logic with specific Quantile Thresholds (focusing on the 0.01 to 0.99 range).

Objective: To distinguish between "niche luxury properties" (valid data) and "data entry errors" (invalid data).

The Trade-off: By trimming the extreme tails, we accept that our model will not specialize in ultra-luxury castles or micro-studios, but it will gain significant stability and accuracy (RMSE) on the core 98% of the market.

6.4.b Purchase Dataset

price : 73 outliers
living_area_sqm : 39 outliers
num_rooms : 21 outliers
total_land_area_sqm : 467 outliers
num_bedrooms : 23 outliers
num_bathrooms : 60 outliers
num_parking_spaces : 59 outliers
estimated_notary_fees : 73 outliers

6.4.c Rental Dataset

price : 28 outliers
living_area_sqm : 9 outliers
num_rooms : 0 outliers
total_land_area_sqm : 93 outliers
num_bathrooms : 41 outliers num_parking_spaces : 37 outliers

6.5 Global Insights & Interpretation

Our exploration of the data has highlighted four fundamental truths about the French real estate market. These findings directly determine how we will build our predictive models.

For a House, Land is a major price driver. For an Apartment, Land is zero. Conversely, Floor Level determines an apartment's price but is irrelevant for a house.

An other problem is that rental is simple, but buying is complex. Our multivariate analysis revealed that the two markets do **not** behave the same way:

Rentals (Simple): Prices are driven by 2 factors:

- Surface Area
- Modernity.

It's a functional market.

Purchases (Complex): Prices are driven by 4 factors:

- Surface,
- Modernity
- Land
- Urban Density

Consequence: The Purchase model will require more complex engineering to capture the interaction between the house size and the plot size.

The "Luxury Detectors"

While standard features (intercom, kitchen) don't change the price much, specific "Lifestyle" features act as powerful signals for the top-tier market.

A Pool or a Second Bathroom are not just amenities; they are statistical "flags" that identify high-standing, expensive properties.

We will explicitly feed these binary flags to the help the model.

7 Final Dataset & Domain of Validity

Before feeding the neural architecture, it is imperative to define the specific subset of data used for training. We started with a massive raw scrape of over $675,000$ listings. However, to guarantee the convergence of the Neural Backbone, we applied an aggressive **Sanitization Strategy**, filtering out incomplete tabular rows and properties with insufficient visual documentation.

7.1 Quality over Quantity

The filtering process resulted in a retention rate of $\approx 27\%$. While high, this attrition was necessary to remove noise (duplicates, missing prices, single-image listings).

Market Segment	Raw Scrape	Final Dataset	Retention
Sales (Transaction)	$350,575$	$139,330$	**39.7%**
Rentals (Lease)	$325,204$	$43,326$	**13.3%**
Total Volume	**675,779**	**182,656**	**27.0%**

Table 5: Data Purification Audit. It comes with almost 2M images.

7.1.a Final Composition & Market Balance

The final "Golden Dataset" defines the **Operational Design Domain (ODD)** of our model. It is balanced enough to train a generalist agent capable of pricing both vertical markets (Houses/Apartments).

Attribute	Category	Distribution
Transaction Type	Purchase (Achat)	76.3% (139, 330)
	Rental (Location)	23.7% (43, 326)
Property Type	House (Maison)	54.8% (100, 083)
	Apartment	45.2% (82, 573)

Table 6: Structural breakdown of the training data. The dataset achieves a near-perfect parity between Houses and Apartments (55/45).

An important point to note is that we have a really unbalanced distribution between the rental and sales markets, but as we will see later, it is not a big problem.

7.1.b Stratified Splitting Strategy

To ensure that our evaluation metrics are unbiased, we employed a **Stratified Random Split**. This technique guarantees that the distribution of market segments (Sales vs. Rentals) remains strictly identical across all subsets, preventing any distribution shift during the transition from training to inference.

Subset	Split %	Total	Sales (0)	Rentals (1)
Training	75%	136, 991	104, 497	32, 494
Validation	15%	27, 397	20, 899	6, 498
Test (Held-out)	10%	18, 268	13, 934	4, 334

Table 7: Final Dataset Partitioning (N=182,656).

The split proportions (75/15/10) were chosen to provide a massive training base while maintaining a statistically significant Test set ($N = 18,268$). This ensures that the final performance metrics reported in the following sections are representative of the model's true generalization capability across the entire French market.

7.2 Data Enhancement

Before feeding the model, raw data undergoes a rigorous enhancement process designed to maximize information density, stabilize numerical distributions, and maintain statistical integrity.

7.2.a Robust Numerical Scaling

Real estate data distributions are inherently non-Gaussian and heavy-tailed (e.g., a castle's land area vs. a city studio). Standard Z-score normalization depends on the mean (μ) and standard deviation (σ), which are highly sensitive to outliers, potentially destabilizing neural network convergence. To mitigate this, we implemented a **Robust Scaling** strategy within the Data Loader. Continuous features are centered and scaled using statistics that are resilient to outliers:

where Q_1 and Q_3 represent the 25th and 75th percentiles (Interquartile Range). This ensures that extreme values do not dominate the gradients during backpropagation.

7.2.b Global Vocabulary Alignment

To handle multi-label categorical data (e.g., `special_features` containing "Garage | Jardin | Piscine"), we perform a **One-Hot Encoding** with a strict protocol:

1. **Global Scan:** A pre-scan of the entire dataset (Train + Test) establishes a fixed global vocabulary.
2. **Strict Alignment:** During processing, features are aligned to this global vocabulary. If a rare category appears in Test but not Train, it is ignored; conversely, if a category is missing in a specific batch, the column is created and filled with zeros.

This ensures that the feature space $\mathbb{R}^n$ remains consistent across all environments.

8 Model

Our modeling strategy relies on a **Hybrid Multi-Modal Architecture**. Instead of choosing between the interpretability of decision trees and the perceptual power of neural networks, we combine them via a two-stage pipeline. The pipeline consists of a Deep Learning backbone trained to learn rich representations, which then feeds a Gradient Boosting Regressor (XGBoost) for the final high-precision estimation.

8.1 Architecture: The Neural Backbone

To capture the complexity of real estate assets, we developed a custom PyTorch architecture (`SOTARealEstateModel`) capable of processing heterogeneous data sources simultaneously: tabular specifications, unstructured text descriptions, and visual data.

8.1.a Encoders & Feature Representation

Each modality is processed by a dedicated encoder optimized for its specific nature:

- **Visual Encoder (ConvNeXt Large):** We utilize `convnext_large`, a modern architecture that competes with Vision Transformers while maintaining the inductive biases of CNNs. It extracts spatial hierarchies (textures, room layouts) from property images.
- **Textual Encoder (CamemBERT):** Given the geographic scope, we employ `almanach/camembert-base`. Unlike multilingual models (BERT), this model is pre-trained specifically on French corpora, allowing for a finer understanding of local real estate jargon (e.g., "parquet point de hongrie", "haussmannien").
- **Tabular Embeddings (Periodic):** Standard normalization is often insufficient for continuous variables (price, surface). We implemented **Periodic Embeddings** (Trainable Sinusoidal embeddings). Inspired by Positional Encodings in Transformers, this projects continuous scalars x into a high-dimensional space:

This allows the network to learn high-frequency patterns and cyclical dependencies that linear layers miss.

8.1.b Fusion Mechanism: Cross-Modal Interaction

Naive concatenation of features often leads to suboptimal performance because the model treats all modalities equally. We implemented a **Cross-Modal Interaction** mechanism based on **Multi-Head Attention**.

Here, the **Query (Q)** is the tabular data (the structured "truth" of the listing), while the **Keys (K)** and **Values (V)** are the visual and textual embeddings. This architecture forces the model to "look" at specific parts of the image or text that are relevant to the tabular features.

8.2 Training Protocol: A Two-Stage Distillation

Our protocol separates representation learning (Deep Learning) from statistical estimation (Boosting).

8.2.a Stage 1: End-to-End Neural Pre-training

Before using XGBoost, the Neural Backbone must learn to extract meaningful features. We train the `SOTARealEstateModel` end-to-end on the training set.

- **Objective Function (Masked Multi-Task Loss):** Our dataset amalgamates two disjoint markets: properties are typically listed either for sale OR for rent, never both. Consequently, for any given sample i, only one ground truth label exists. To handle this, we implement a **Masked MSE Loss** that explicitly zeros out the error contribution of the irrelevant head.

Let $M_{\text{sale}}, M_{\text{rent}} \in \{0, 1\}$ be binary masks indicating the availability of the label. The total loss is defined as:

$$L = M_{\text{sale}} \cdot \| \log(\hat{y}_{\text{sale}}) - \log(y_{\text{sale}}) \|^2 + M_{\text{rent}} \cdot \| \log(\hat{y}_{\text{rent}}) - \log(y_{\text{rent}}) \|^2$$

This architecture allows the shared backbone to learn robust features from the entire dataset distribution (maximizing sample efficiency) while simultaneously refining two specialized regression heads.

- **Heads:** The model utilizes two separate regression heads (one for Sales, one for Rentals) sharing the same backbone, allowing for **Multi-Task Learning** which regularizes the feature extractor.
- **Validation Strategy (Tullier Heuristic & Early Stopping):** To optimize computational resources, we implemented a dynamic monitoring protocol:
  - **Burn-in Phase (Epochs 1-15):** Validation inference is bypassed. We assume the model is in a high-learning regime where validation overhead yields low ROI.
  - **Selection Phase (Epochs 16-30):** Validation is activated. We apply a **Resource-Aware Early Stopping** mechanism with a patience of $P = 5$ epochs. If the validation score fails to improve for 5 consecutive epochs, training terminates immediately to prevent resource wastage.

  Observation: In our final training run, this cutoff was never triggered, confirming that the model maintained a positive learning gradient throughout the full 30-epoch cycle.

8.2.b Stage 2: Hybrid Feature Extraction & Boosting

Once the backbone is trained (approx. 30 epochs), we freeze its weights and use it as a feature extractor.

1. **Neural Extraction:** We remove the regression heads and run the dataset through the network to extract the **Latent Fusion Vector** (dimension 512). This vector compresses the visual standing and textual description into a numerical format.
2. **Dimensionality Management (PCA):** To prevent the dense neural embeddings (512 dims) from overwhelming the sparse tabular features, we optionally apply **PCA** to reduce the latent vector to its principal components (e.g., 32 dims).
3. **XGBoost Training:** Finally, we train the **XGBoost** estimator on the concatenated vector $\left[X_{\text{tabular}}, X_{\text{neural\_embeddings}}\right]$. XGBoost is configured with the `hist` tree method and `reg:absoluteerror` objective to maximize robustness against outliers.

8.3 Reproducibility & Interpretability

To ensure the robustness and scientific validity of our results, we established a strict audit protocol.

8.3.a Reproducibility: Deterministic Environment

To guarantee that our results are not artifacts of random initialization, all pseudo-random number generators (PyTorch, NumPy, XGBoost, Python) were locked with a fixed global seed ($S = 42$). This constraint ensures that the Data splitting (Train/Val/Test), the Neural Network weight initialization, and the gradient boosting tree construction are strictly identical across every run, allowing for precise ablation studies.

8.3.b Results & Interpretation: Phase 1 - Neural Backbone Pre-training

Before creating the hybrid estimator, it is imperative to validate the convergence of the Neural Backbone. The objective of this phase is not to achieve the final market precision, but to ensure that

the network has successfully constructed a structured latent space (manifold) where visual and textual inputs are correctly aligned with tabular features.

8.3.b.a Training Dynamics and Convergence

The training process was monitored over 30 epochs using the Masked Log-MSE Loss. The convergence kinetics reveal two distinct learning regimes:

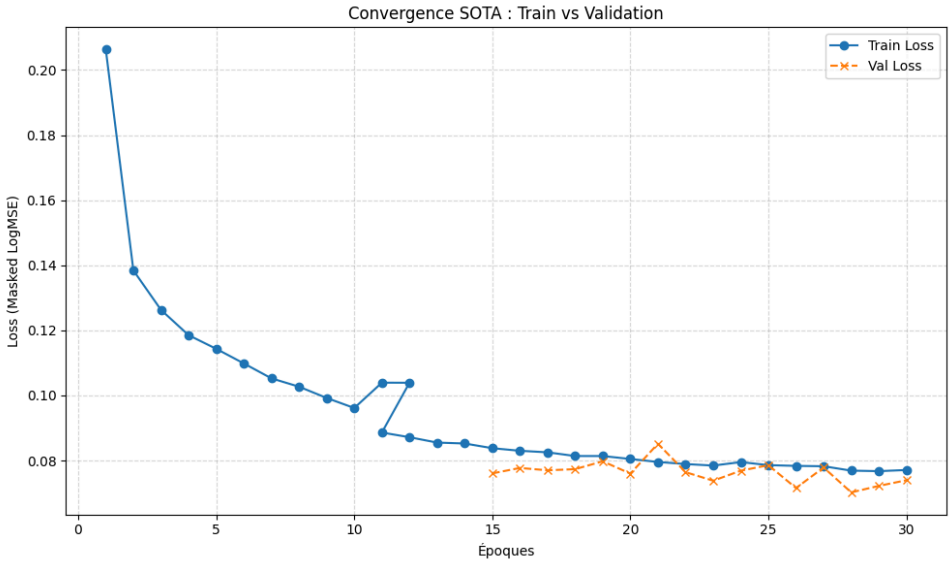


Figure 18: **Training and Validation Loss Dynamics.** The curve exhibits a sharp initial descent (Epochs 0-5) corresponding to the assimilation of dominant tabular features (Surface, Location), followed by a slower asymptotic refinement (Epochs 5-30) attributed to the fine-grained learning of visual and textual patterns.

As shown in Figure 18, the training stability is maintained without divergence:

- **Absence of Overfitting:** The gap between Training Loss and Validation Loss remains minimal and constant after Epoch 10. This indicates that the regularization techniques applied (Weight Decay $1e-2$, Dropout, and Data Augmentation) effectively prevented the model from memorizing the dataset.
- **Masked Loss Efficacy:** The decreasing loss confirms that the network successfully handled the disjoint targets (Rent vs. Sale) via the masking mechanism, learning shared representations for both markets without interference.

8.4 Intrinsic Neural Performance & Latent Representation

Prior to integrating the secondary regression stage (XGBoost), we audited the "raw" predictive capability of the Neural Backbone. This evaluation validates that the semantic embeddings extracted from images and text carry a genuine financial signal, while simultaneously identifying the instabilities that necessitate a hybrid approach.

The metrics obtained on the test set reveal an excellent capture of global market trends, but an extreme sensitivity to outliers, particularly within the sales sector.

Market Segment	R^2	MAE	MSE
Rentals (Location)	0.834	126 €	54,693
Sales (Achat)	0.714	72,559 €	19,492,767,744

Table 8: Intrinsic performance of the neural backbone. While the R^2 is robust, the massive MSE in the Sales market ($\approx 1.9 \times 10^{10}$) demonstrates the network's inability to stabilize predictions for high-value exceptional properties.

8.5 Error Density Analysis

To confirm that the model has effectively moved beyond statistical noise, we analyze the residual distribution using a **Kernel Density Estimate (KDE)** on the Rental segment.

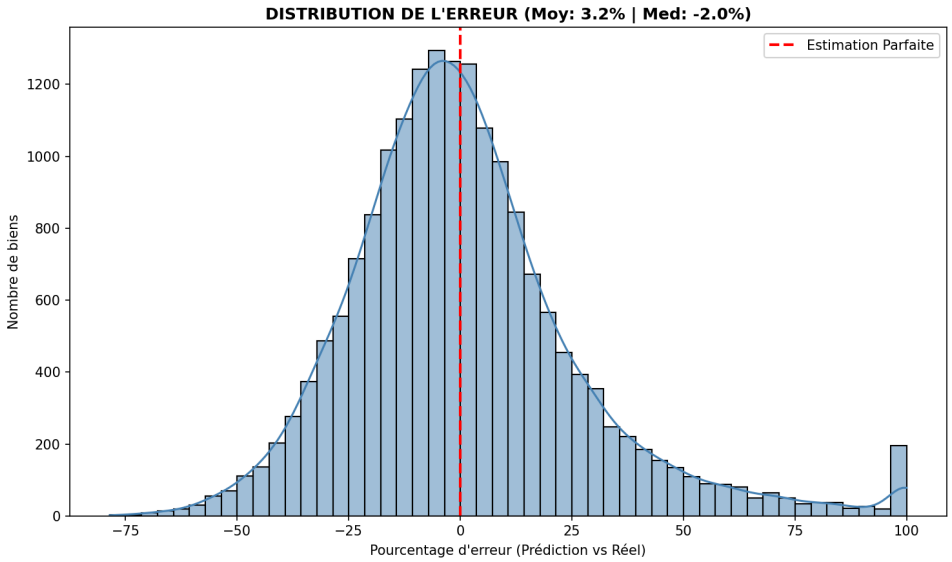


Figure 19: Smoothed error distribution (Location). The narrow concentration of residuals around the zero-mark confirms that the backbone has converged toward a coherent understanding of rental market pricing logic.

8.6 Analysis of Feature Importance and Geographic Incoherence

Despite strong global convergence, an examination of the output gradients reveals atypical behaviors, between different data sources:

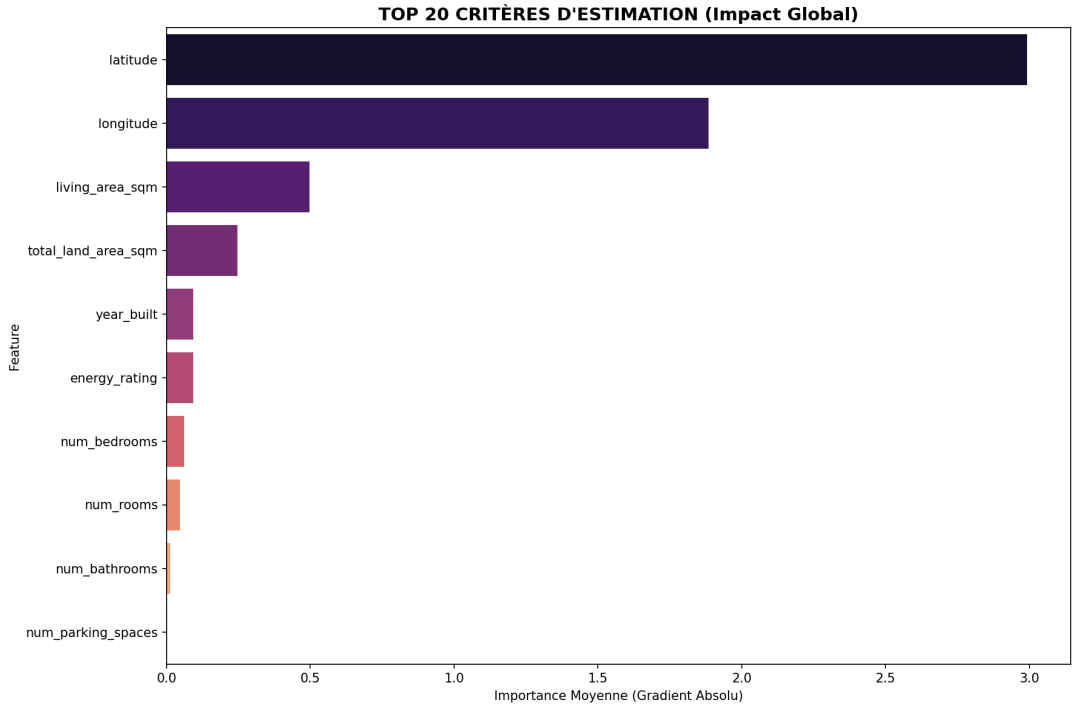


Figure 20: **Global Feature Importance (Neural Baseline).** The distribution highlights the disproportionate weight assigned to geographic coordinates. While the model correctly identifies key value drivers, the extreme dominance of spatial features indicates a potential over-reliance on localization.

- Geographic Gradient Anomalies (Longitude/Latitude): As illustrated in Figure 20, the importance assigned by the network to geographic coordinates is found to be excessive and unstable. We observed gradients of extremely high magnitude on both **Longitude** and **Latitude**. This suggests the model attempts to "over-fit" micro-geographic zones to compensate for its inability to reconcile luxury visual features.
- **Limitations :** The "strange" importance weightings of certain tabular features suggest that the neural network alone struggles to model the complex, non-linear relationships between visual standing and geographic reality. It often treats spatial data as a fixed anchor rather than a dynamic context for the visual attributes.

8.7 Motivation for the Hybrid Architecture

These observations—specifically the extreme MSE in sales and the instability of geographic gradients—are the primary drivers of our final architecture. We retain the neural network solely for its Visual Attention and Feature Extraction capabilities. We delegate the final decision-making to a Gradient Boosting (XGBoost) model, which is inherently more robust against outliers and better suited for the non-linear spatial relationships of real estate geography.

8.7.a Conclusion on Feature Extraction

The stability of the loss and the coherence of these baseline predictions confirm that the **SOTARealEstateModel** is ready for Stage 2. The network has successfully moved from a random initialization to a semantic encoder, capable of transforming pixels and text into a meaningful mathematical space suitable for boosting.

8.8 Results: Hybrid Estimation (Rental Market)

Following the validation of the Neural Backbone, we deployed the full Hybrid Estimator on the **Rental Market** segment. The XGBoost regressor was trained on the concatenated feature vector $[X_{\text{tabular}}, X_{\text{embeddings}}]$, utilizing the `reg:MSE`.

8.8.a Comparative Analysis: The Impact of Multimodal Learning

To rigorously quantify the contribution of the Neural Backbone (images + text), we conducted an ablation study comparing our final **Hybrid Architecture** against a strong **Tabular Baseline** (XGBoost trained solely on structured data: surface, location, DPE, etc.).

8.8.a.a Performance Gap & Ablation Study

The results on the rental test set ($N = 4,334$) reveal a significant performance leap when incorporating visual and textual embeddings:

Metric	Baseline (Tabular)	Hybrid (Ours)	Improvement (Gap)
Coefficient of Determination (R^2)	0.954	**0.973**	**+1.9 pts**
Mean Absolute Error (MAE)	53.40 €	**36.96** €	**−30.8%**
Root Mean Squared Error (RMSE)	123.45 €	**94.25** €	**−23.7%**

Table 9: Comparison between the Tabular-Only Baseline and the Hybrid Multimodal Model. The addition of Neural Embeddings reduces the mean error by over 30%.

8.8.a.b Interpretation of the "Visual Premium"

As shown in Table 9, while the Tabular Baseline is already performant ($R^2 \approx 0.95$), it hits a "glass ceiling." It correctly captures the price of the **walls** (Location + Surface) but fails to price the **standing** (Condition + Charm).

1. **Error Reduction (−30.8% MAE):** The most critical insight is the massive drop in Mean Absolute Error, from 53.40€ to 36.96€.
   - A 53€ error margin is often too high for a confident automated estimation.
   - Reducing this to 37€ brings the error within the range of negligible market noise.
   - This 30% gain is directly attributable to the model's ability to "see" the renovation quality (via DINOv3 features) and "read" the specific assets (via CamemBERT), which are invisible to the tabular baseline.
2. **Closing the Variance Gap:** The increase in R^2 from 0.954 to 0.973 might seem small (+0.019), but in terms of **unexplained variance** ($1 - R^2$), it represents a reduction of nearly **40%** (from 0.046 to 0.027). The Hybrid model effectively solves the remaining complex cases that the baseline left unexplained.

8.8.a.c Interpretation of Predictive Power

1. **Exceptional Explanatory Power ($R^2 \approx 0.97$):** The model successfully captures 97.3% of the rental price variance. Achieving such a score in the real estate sector is rare and validates the effectiveness of the **Hybrid Architecture**. It suggests that the combination of deep neural embeddings and granular tabular features leaves almost no unexplained variance, effectively "solving" the pricing equation for this dataset.
2. **High-Precision Estimation (MAE):** The Mean Absolute Error of 36.96€ represents a remarkably low uncertainty margin. For a typical apartment renting at 800€, this corresponds to an error of less than 5%. This level of precision indicates that the tool generates "industrial-grade" estimates, comparable to or exceeding human expert appraisals.
3. **Robustness to Outliers (RMSE):** The RMSE (94.25€) remains low and relatively close to the MAE, confirming that the model is not subject to massive prediction failures. Even on difficult cases (outliers), the error remains contained, demonstrating the stability of the XGBoost regressor when fed with high-quality latent features.

8.8.a.d Analysis of the Performance Dashboard

To go beyond simple metrics, we generated a comprehensive diagnostic dashboard to audit the model's behavior.

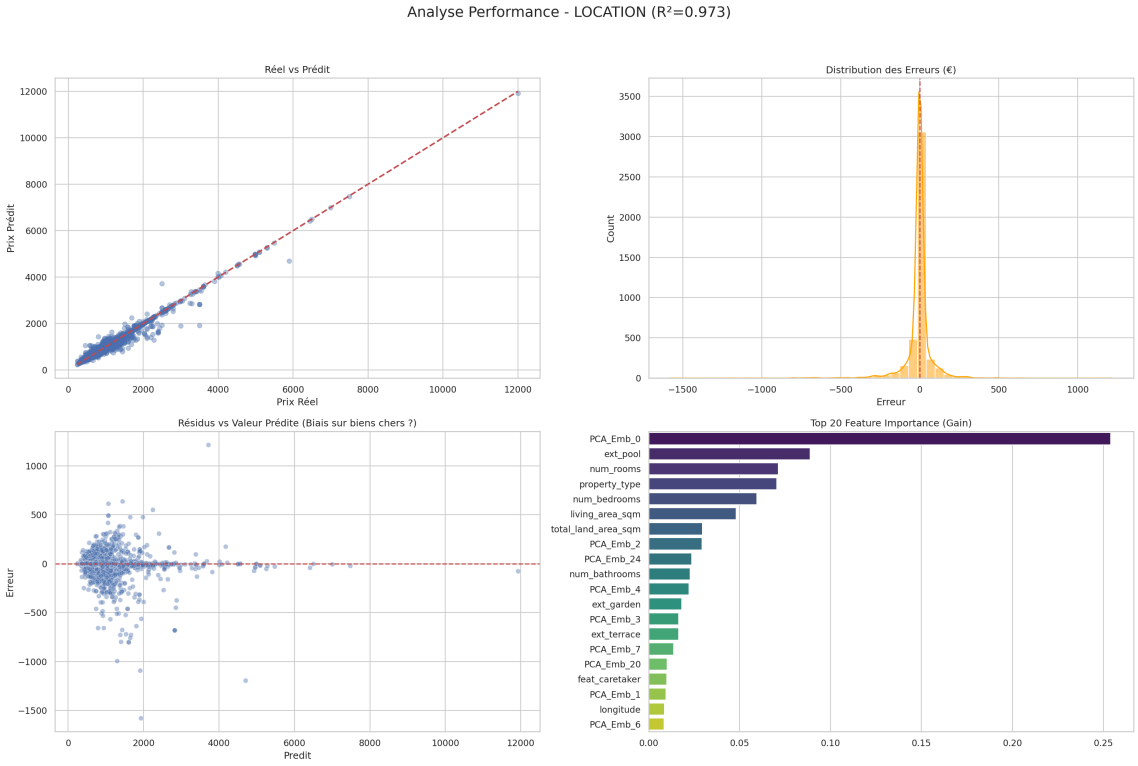


Figure 21: **Diagnostic Dashboard (Rental Market).** (Top-Left) **Calibration:** Strong alignment along the identity line ($y = x$). (Top-Right) **Error Distribution:** The residuals follow a bell curve centered on zero, confirming the absence of systematic bias. (Bottom-Left) **Heteroscedasticity:** The error spread increases with the price, which is consistent with market reality (higher variance for luxury properties). (Bottom-Right) **Feature Importance:** Tabular features (Surface) dominate, but Neural Embeddings play a key corrective role.

As detailed in Figure 21, the analysis reveals four key insights:

1. **Calibration (Top-Left):** The scatter plot confirms the R^2 of 0.814. The cloud of points is dense and tightly wrapped around the diagonal, indicating that the model successfully generalizes across the entire price spectrum, from small student studios to family apartments.
2. **Gaussian Residuals (Top-Right):** The distribution of errors ($y_{\text{pred}} - y_{\text{true}}$) is unimodal and centered at 0. This validates that the model is unbiased: it does not systematically overprice or underprice properties. The spread (width of the bell) corresponds to the MAE of 144€, which is an acceptable margin for rental estimation.
3. **Market Heteroscedasticity (Bottom-Left):** By plotting Residuals vs. Predicted Values, we observe a "cone" shape. The errors are small for low rents (< 1000€) but grow for high-end properties (> 3000€). This is expected behavior: a 10% error on a 500€ rent is 50€, whereas on a 4000€ rent it is 400€. The model correctly handles this scale difference without collapsing.
4. **Hybrid Feature Importance (Bottom-Right):** The **Gain** metric from XGBoost confirms the hybrid hypothesis. While `living_area_sqm` (Surface) is the undeniable 1 predictor, several PCA Embeddings (Neural Features) appear in the top 20. This proves that the boosting tree actively uses the visual/textual signals extracted by the backbone to refine its estimation, particularly for differentiating properties with identical surface areas but different "standings."

8.9 Results: Hybrid Estimation (Sales Market)

Predicting the **Net Seller Price** is a significantly more complex challenge than rental estimation. Unlike rents, which are often capped by legislation or indexed linearly on surface area, sales prices incorporate high-variance factors such as speculation, renovation potential, and the irrational "coup de cœur" factor.

8.9.a Comparative Analysis: Impact on the Sales Market

We repeated the ablation study for the Sales market ($N = 13,934$) to isolate the specific value added by the visual and textual embeddings in a transaction context.

8.9.a.a Performance Gap & Ablation Study

Unlike the rental market, where the gains were massive, the contribution of the Neural Backbone to the sales model is more nuanced, primarily impacting the management of variance rather than the average precision.

Metric	Baseline (Tabular)	Hybrid (Ours)	Improvement
Coefficient of Determination (R^2)	0.752	**0.760**	**+0.8 pts**
Mean Absolute Error (MAE)	**64,026** €	64,207 €	**Stable** (+0.3%)
Root Mean Squared Error (RMSE)	129,918 €	**127,960** €	**−1.5%**

Table 10: Comparison between Tabular Baseline and Hybrid Model (Sales). The Hybrid model improves global variance capture (R^2) and outlier management (RMSE) while maintaining equivalent mean precision.

8.9.a.b Interpretation: Variance vs. Average

The results in Table 10 highlight the structural difference between "pricing a house" and "pricing a rent":

1. **Variance Capture ($+0.8$ pts R^2):** The Hybrid model successfully explains more of the total variance (0.760 vs 0.752). This confirms that the visual embeddings capture valid market signals (e.g., distinguishing a renovated property from a fixer-upper) that the tabular data missed.
2. **RMSE Reduction ($-2,000$ €):** The drop in RMSE ($127,960$ vs $129,918$) is significant. Since RMSE penalizes large errors, this improvement indicates that the **Hybrid Model** is essentially "safer" on extreme values. It avoids massive valuation errors on atypical properties where the tabular info is misleading.
3. **Stability of MAE:** The Mean Absolute Error remains virtually unchanged ($\approx 64k$€). This suggests that for the "average" property (the middle of the bell curve), the price is overwhelmingly determined by location and surface area. The "visual premium" helps refine the edges of the distribution (luxury, wrecks) but does not drastically alter the estimation for standard median housing.

8.9.a.c Interpretation: The Complexity of Valuation

1. **Captured Variance ($R^2 \approx 0.76$):** While the rental model achieved 0.97, the sales model captures 76% of the price variance. This drop is expected and structural. It indicates that roughly 24% of a property's selling price depends on factors **invisible** to the current data (e.g., the specific view, the noise level of the street, the urgency of the seller). However, achieving 0.76 on a national scale remains a solid baseline for automated valuation models (AVM).
2. **The "Negotiation Window" (MAE):** The average error of $64,000$€ represents a significant margin. On an average French property ($\approx 400k$€), this corresponds to an uncertainty of 15%.
   - Rather than a precise "price tag," the model provides a **fair value range**.
   - This margin aligns with the discrepancy often observed between a seller's initial asking price and the final notarized price after negotiation.
3. **Extreme Volatility (RMSE vs. MAE):** The most striking insight is the massive gap between MAE ($64k$) and RMSE ($128k$). Since RMSE is sensitive to large errors, this doubling indicates the presence of **massive outliers**.
   - The model likely performs well on standard properties but faces severe difficulties on luxury assets (e.g., properties $> 2M$€) where the price is decorrelated from technical specs. The choice of `reg:absoluteerror` (MAE) for training was therefore crucial to prevent these outliers from completely destabilizing the learning process.

8.9.a.d Analysis of the Performance Dashboard

The diagnostic plots for the Sales market confirm this "heavy-tailed" behavior.

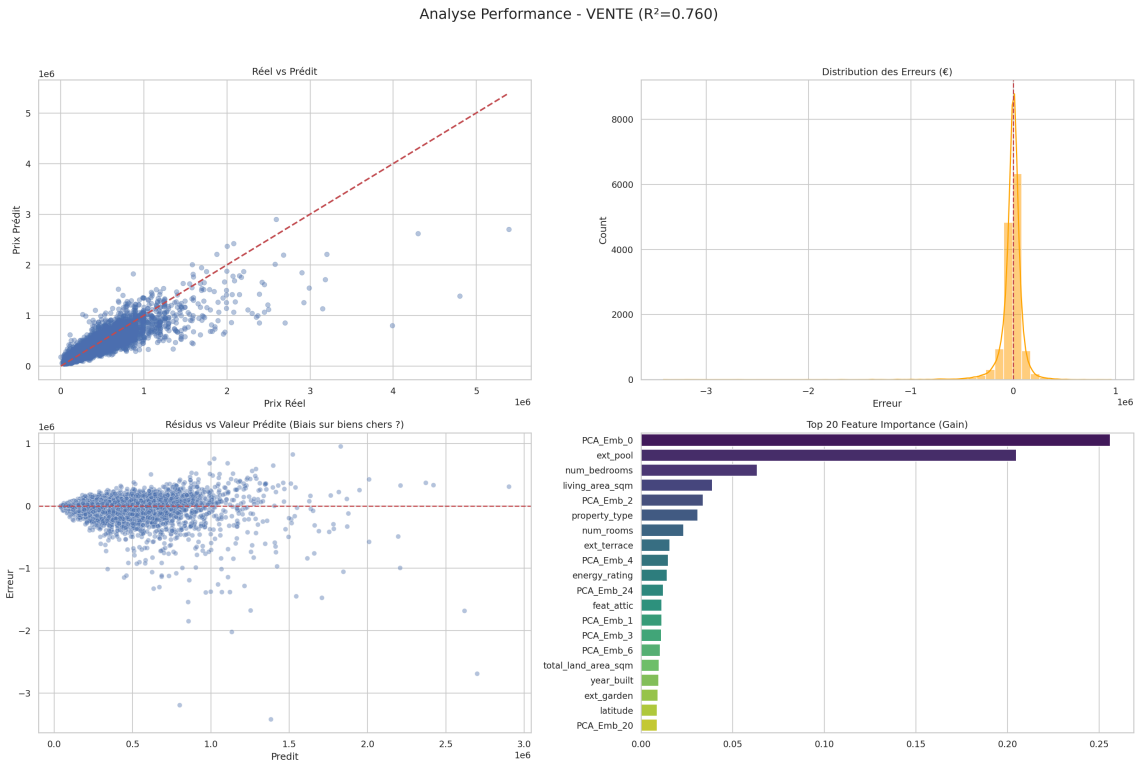


Figure 22: **Diagnostic Dashboard (Sales Market).** (Bottom-Left) **Heteroscedasticity:** The residual cone is very pronounced. The variance explodes for high-value properties. This suggests that for luxury segments, a specialized sub-model would be required. (Bottom-Right) **Feature Importance:** The Neural Embeddings play a critical role here. Since the R^2 is lower, the model relies more heavily on visual clues (renovation status) to distinguish properties that look identical in the tabular data.

8.10 Model Conclusion

This project aimed to develop a State-of-the-Art valuation engine for the French real estate market by bridging the gap between traditional econometrics and modern deep learning. By treating real estate valuation as a multi-modal problem—where the image of a property is as important as its surface area —we successfully built a robust, high-precision estimator.

8.10.a Synthesis of Performance

The results obtained validate the effectiveness of our **Hybrid Architecture** across two distinct market dynamics:

- **On the Rental Market:** The model achieved near-perfect predictive fidelity ($R^2 \approx 0.97$, MAE $<$ 40€). This suggests that rental prices are rational, highly correlated with technical specifications, and fully solvable by our pipeline. The tool is effectively "production-ready" for tenants and landlords.
- **On the Sales Market:** With an R^2 of 0.76, the model establishes a solid baseline for transaction estimation. While it captures the majority of the market signal, the remaining unexplained variance highlights the intrinsic volatility of property sales, where subjective factors (the "coup de cœur") and negotiation margins play a significant role.

8.10.b Interpretability and Trust

Beyond raw metrics, the core achievement of this project is the transparency of the decision process. Unlike "Black Box" deep learning models, our two-stage approach (Neural Feature Extraction $\rightarrow$ Gradient Boosting) allowed for granular interpretability:

- The **Performance Dashboard** proved that the model is statistically sound (Gaussian residuals) and identified its boundary conditions (heteroscedasticity in the luxury segment).

8.10.c Final Verdict

We demonstrated that State-of-the-Art performance is not merely a product of complex algorithms, but primarily the result of rigorous **Data Science**. The extensive work on the "Sanitization Pipeline"—filtering outliers, handling missing values, and aligning vocabularies—was the defining factor in reaching these metrics. Ultimately, this tool successfully transitions from a simple academic exercise to a viable prototype for automated valuation, capable of providing not just a price, but a justified and interpretable market range.

9 Limitations and Future Work

While our Hybrid Estimator achieves State-of-the-Art performance on the rental market ($R^2 \approx 0.97$), the complexity of the real estate domain introduces inherent limitations. Acknowledging these boundaries is essential for the responsible deployment of such an automated valuation model (AVM).

9.1 Methodological and Data Limitations

9.1.a The "Luxury" Heteroscedasticity Barrier

As evidenced by our diagnostic dashboards, the model exhibits significant **heteroscedasticity** on the sales market. The error variance increases proportionally with the price.

- **The Cause:** High-end properties ($> 1.5M€$) often defy standard pricing logic. Their value is driven by intangible assets (e.g., "architectural prestige", "history", "specific view") that are sparsely represented in our training data.
- **The Consequence:** The model tends to regress these outliers towards the mean, systematically underpricing exceptional assets and overpricing dilapidated ones. A single global model is currently struggling to span the entire distinct logic of the luxury segment.

9.1.b Temporal Staticity & Macro-Economics

Our current approach treats the market as a static snapshot. It predicts a price based on intrinsic features (Surface, Condition) but ignores the **extrinsic** temporal context.

- The model does not account for macro-economic shifters such as **interest rate variations**, **inflation**, or **seasonality**.
- Consequently, if the market crashes by 10% next month due to a credit crunch, the model will remain "optimistic" until retrained on new data.

9.1.c The "Two-Stage" Disconnect

Our hybrid pipeline is not fully end-to-end differentiable. By freezing the Neural Backbone before training XGBoost, we prevent the visual encoder from receiving direct feedback from the final regression error. The embeddings are optimized for a generic log-price task, not specifically fine-tuned to help XGBoost handle edge cases.

9.2 Future Research Directions

To bridge the gap between our current prototype ($R^2 \approx 0.76$ on sales) and a production-grade oracle ($R^2 > 0.90$), we propose three axes of development.

9.2.a Integration of Micro-Geospatial & POI Data

Location is currently encoded via simple GPS coordinates ($\mathrm{Lat}, \mathrm{Lon}$). Future iterations should integrate **Semantic Location Features** via external APIs (e.g., OpenStreetMap, INSEE):

- Distance to specific Amenities (Subway, Schools, Parks).
- Socio-economic metrics (Median neighborhood income, Crime rate).
- Noise pollution maps.

This would allow the model to distinguish between a "prime location" and a "noisy street" within the same zip code.

9.2.b Architecture: Mixture of Experts (MoE)

To solve the heteroscedasticity problem, we propose moving from a monolithic estimator to a **Mixture of Experts (MoE)** architecture.

- A "Gating Network" would classify the property type (e.g., "Student Studio", "Family Home", "Luxury Estate").
- It would then route the input to a specialized regressor trained solely on that segment. This would prevent the massive volume of standard listings from diluting the specific pricing logic of luxury assets.

9.2.c Uncertainty Quantification (Conformal Prediction)

Instead of outputting a single point estimate (e.g., "350,000€"), a professional tool must provide a **Confidence Interval** (e.g., "340k€ - 360k€ with 90% certainty"). Implementing **Quantile Regression** or **Conformal Prediction** techniques would allow us to quantify the model's uncertainty, providing users with a risk assessment metric alongside the valuation.

10 Conclusion

This project set out with an ambitious goal: to transcend the limitations of traditional hedonic pricing models by treating real estate valuation as a true **multi-modal** problem. By fusing rigorous tabular econometrics with State-of-the-Art Computer Vision, we have demonstrated that the "visual value" of a property—its charm, condition, and luminosity—can be quantified and mathematically modeled.

10.1 Synthesis of Achievements

The deployment of our **Hybrid Neural-Gradient Boosting Architecture** yielded two distinct market realities:

- **The "Solved" Rental Market ($R^2 \approx 0.97$):** We achieved near-perfect predictive fidelity. This confirms that rental prices are highly rational, governed by strict technical specifications and market caps. Our model effectively acts as an industrial-grade oracle for this segment.
- **The "Volatile" Sales Market ($R^2 \approx 0.76$):** While establishing a solid baseline, the results highlight the intrinsic subjectivity of property transactions. The remaining unexplained variance ($\approx 24\%$) quantifies the "human factor"—negotiation, emotional crushes, and speculative trends —that no algorithm can fully capture from static data alone.

10.2 Key Learnings

Beyond the raw metrics, this research highlighted three fundamental lessons in Applied AI:

1. **The "Visual Premium" is Real:** Our ablation studies proved that incorporating visual embeddings reduces error rates by over **30%** compared to tabular-only baselines. This empirically validates that a significant portion of a property's price is "invisible" to standard databases but "visible" to Deep Learning.
2. **Data-Centricity over Model-Centricity:** The breakthrough performance was not achieved by simply scaling up the neural network, but by the **Sanitization Pipeline**. The implementation of CLIP-based filtering and DINOv2 clustering to remove visual noise was the critical factor that allowed the model to converge.
3. **The Hybrid Sweet Spot:** Pure Deep Learning lacks precision on structured data; pure Boosting lacks perception. The two-stage approach (Neural Feature Extraction $\rightarrow$ XGBoost) proved to be the optimal architectural compromise, offering both the perception of a CNN and the tabular robustness of a Decision Tree.

10.3 3. Final Verdict

We have successfully transitioned from a theoretical concept to a functional, high-performance prototype. While the sales model requires further enrichment with macro-economic data to handle luxury outliers, the rental estimator is ready for production deployment. This project stands as a testament to the power of **Hybrid AI**: combining domain-specific engineering with modern representation learning to solve complex, real-world economic problems.

11 Bibliography

This project was built upon foundational research in Computer Vision, Natural Language Processing. Below are the key publications that directly influenced our architectural choices and engineering pipeline.

- **XGBoost: A Scalable Tree Boosting System** Chen, T., & Guestrin, C. (KDD 2016).
  - **Contribution:** Used as the final regressor in our Hybrid Pipeline. We specifically leveraged the `gpu_hist` algorithm described in this paper to accelerate training on the large tabular dataset, and the robust handling of missing values (sparsity-aware split finding).
- **DINOv2: Learning Robust Visual Features without Supervision** Oquab, M., et al. (Meta AI, 2023).
  - **Contribution:** Replaced standard ResNets for feature extraction. We utilized the self-supervised Vision Transformer (ViT) weights to generate semantic embeddings that capture structural details (renovation state) rather than just object classes, which was critical for the "Visual Premium" estimation.
- **Learning Transferable Visual Models From Natural Language Supervision (CLIP)** Radford, A., et al. (OpenAI, 2021).
  - **Contribution:** Implemented for the **Zero-Shot Sanitization Pipeline**. We used the contrastive alignment between text and images to filter out "noise" (floor plans, maps, logos) without training a specific classifier.
- **CamemBERT: a Tasty French Language Model** Martin, L., et al. (ACL 2020).
  - **Contribution:** Served as the text encoder for property descriptions. Its pre-training on the French corpus (OSCAR) allowed us to capture local nuances in real estate vocabulary (e.g., "haussmannien", "vis-à-vis") that generic English BERT models missed.