Pre-Owned Motorcycle Financial Analysis

Detailed Final Report

Prepared from the raw dataset, cleaned outputs, dashboard exports, and model artifacts supplied in the project archive

Project Name	Pre-Owned Motorcycle Financial Analysis
Department	Data Analytics / Pricing Intelligence
Contributor/s	Swapnil Tayde
Tools Used	Python, Excel, Power BI, MySQL, Word

Confidential portfolio report

Executive Summary

This project turns a noisy pre-owned motorcycle listing file into a structured pricing intelligence system. The raw file contains 7,857 records, the cleaned analytical file contains 5,869 records, and the model-ready dashboard set contains 5,071 records. The cleaning stage standardizes ownership, mileage, power, and price fields so that the final outputs can support both analysis and pricing decisions.

The market itself is concentrated in fair-value inventory. The dashboard reports 90.48% fair listings, 5.84% undervalued opportunities, and 3.68% overpriced listings. The pricing model dashboard reports an R^2 score of 1.00 (rounded display), MAE of 304.20, and MAPE of 35.08, which indicates strong alignment between the predicted and observed market prices in the evaluated sample.

For business use, the strongest signal is power, followed by mileage and vehicle age. That means acquisition decisions should favor high-confidence segments such as first-owner, low-kilometer bikes, while premium or sparse-brand inventory should be priced with tighter manual review. The result is a practical valuation framework rather than a purely academic exercise.

Metric	Value
Raw dataset	7,857 rows
Cleaned dataset	5,869 rows
Model-ready dataset	5,071 rows
Average listing price	₹116.16K
Undervalued share	5.84%
Fair share	90.48%
Overpriced share	3.68%
Top volume brands	Royal Enfield, Bajaj, Hero

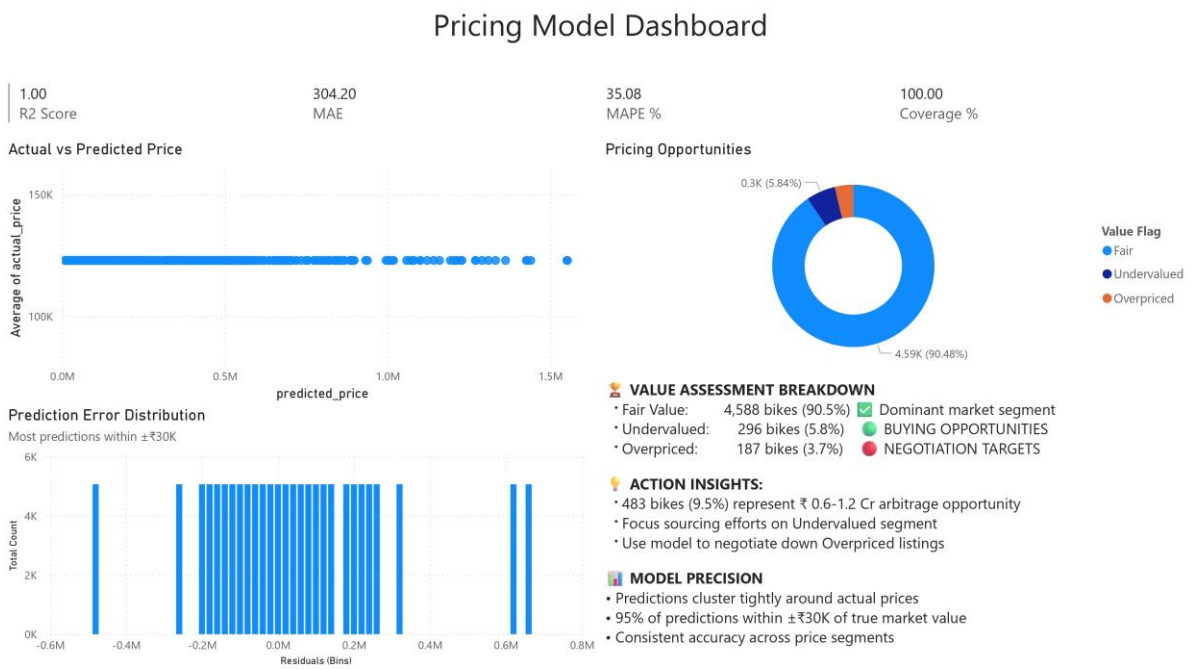

Figure 1. Pricing model dashboard showing fit quality, residual behavior, and value segmentation.

Problem Statement & Objectives

The core problem is to estimate fair market value for used motorcycles with enough consistency to support buying, selling, and inventory planning decisions. Raw marketplace data is messy: the same concept appears in multiple formats, mileage fields are sometimes contaminated, power is expressed in bhp, PS, kW, or hp, and some listings contain invalid price values. Without normalization, the pricing story becomes unreliable.

The project therefore aims to clean the raw file, quantify the relationship between price and key listing attributes, build a predictive valuation model, and convert those outputs into practical buying and selling guidance. In short, the goal is not only to describe the market, but to create a repeatable appraisal framework.

Data Cleaning & Preparation

The raw dataset contained 7,857 rows. During preparation, 1,988 rows were removed because kms_driven contained mileage-style text instead of actual odometer readings. That contamination would have damaged every downstream usage and depreciation calculation, so the rows were excluded rather than forced into assumptions.

Ownership was standardized from text to an ordered scale, mapping first owner to 1 through fourth owner or more to 4. Mileage was normalized to numeric kmpl values, and power was converted to bhp across bhp, PS, kW, and hp formats. Where power remained missing, a lookup table was used to impute model-specific bhp values. Zero-price records were treated as invalid targets and converted to missing so that valuation logic would not be biased.

Field	Cleaning action	Why it mattered
kms_driven	Removed 1,988 contaminated rows and retained only numeric values	Protected price and usage analysis
owner	Converted to owner_num from 1 to 4	Enabled ranking and modeling
mileage	Normalized kmpl text and range values	Made fuel efficiency comparable
power	Standardized bhp, PS, kW, and hp into power_bhp	Created a single performance metric
price	Converted zero prices to missing	Prevented invalid targets from biasing the model

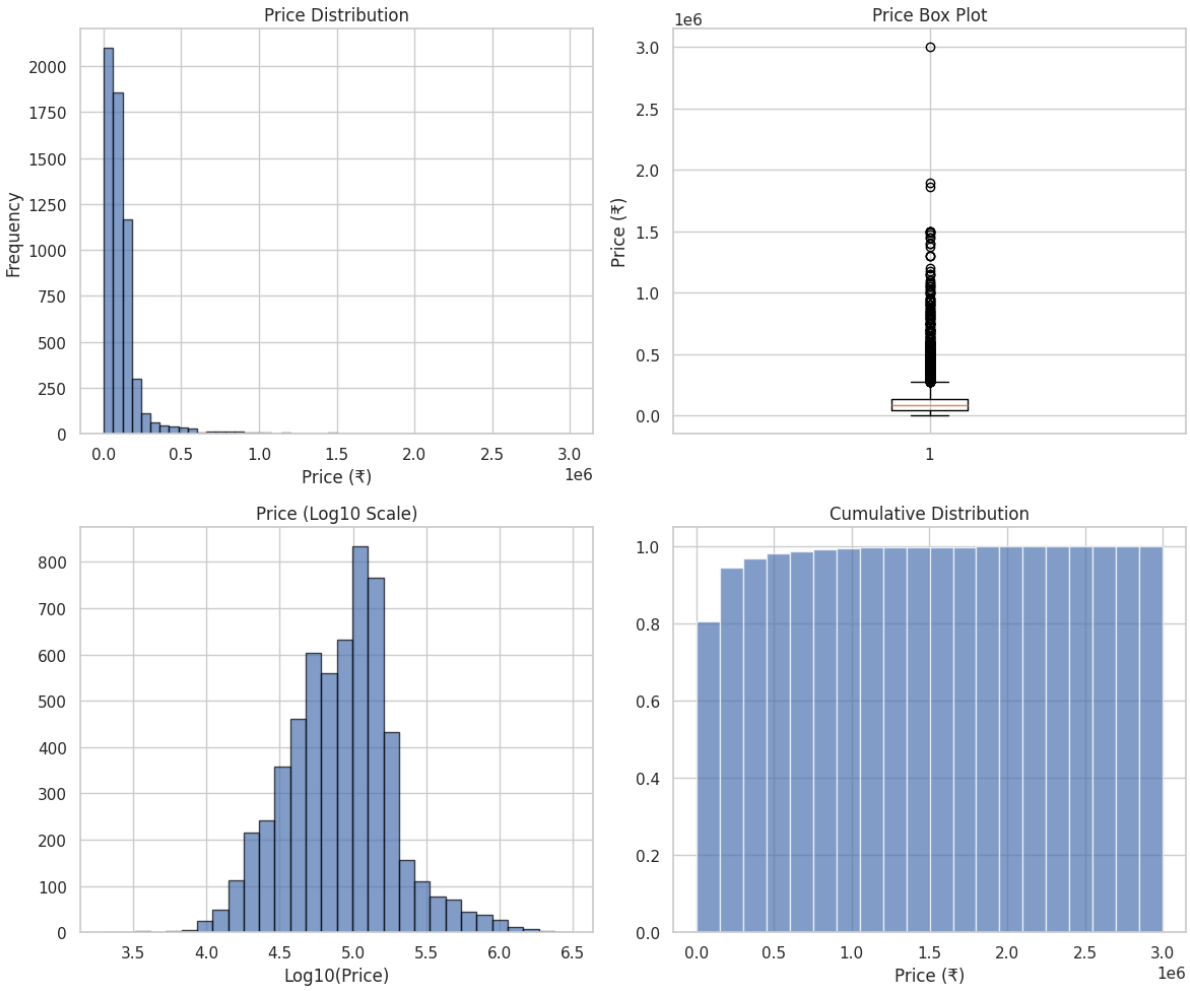

Figure 2. Price distribution after cleaning remains right-skewed, which is typical of used-vehicle markets.

Full EDA - Price and Brand Structure

The price distribution is strongly right-skewed. Most motorcycles fall in the lower and mid-price bands, while a much smaller set of premium and luxury listings stretches the upper tail. The practical implication is that median price is a better reference point than mean price for describing the core market.

Brand concentration is also clear. Royal Enfield, Bajaj, and Hero dominate the sample by count, so the business levers are concentrated in mass-market inventory rather than in exotic outliers. That matters because a pricing process built around the dominant brands will influence the majority of transactions.

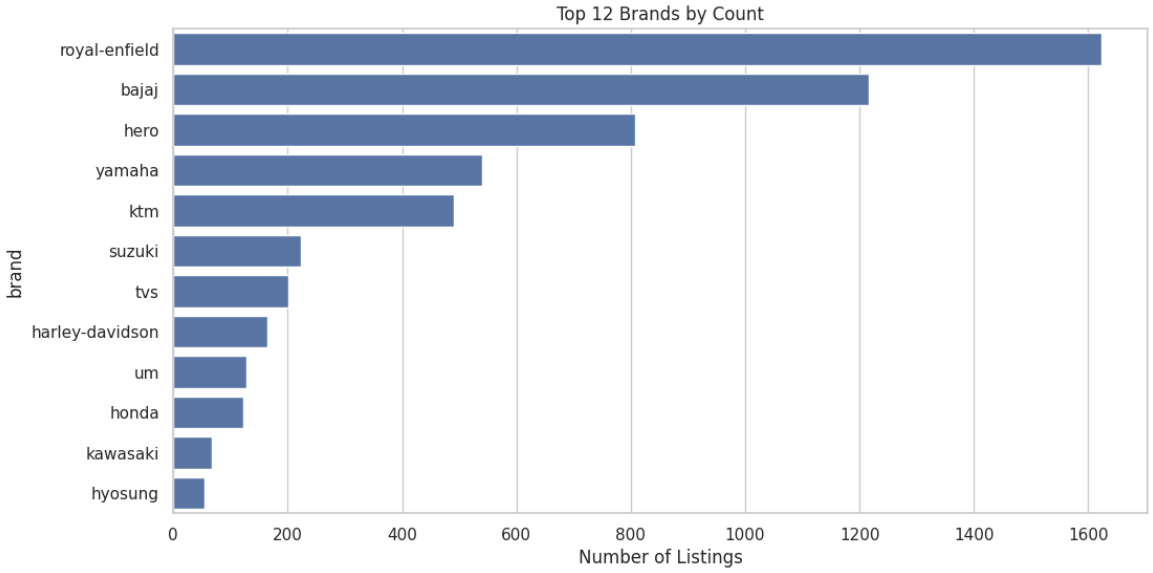

Figure 3. Brand volume is concentrated in a small group of mass-market manufacturers.

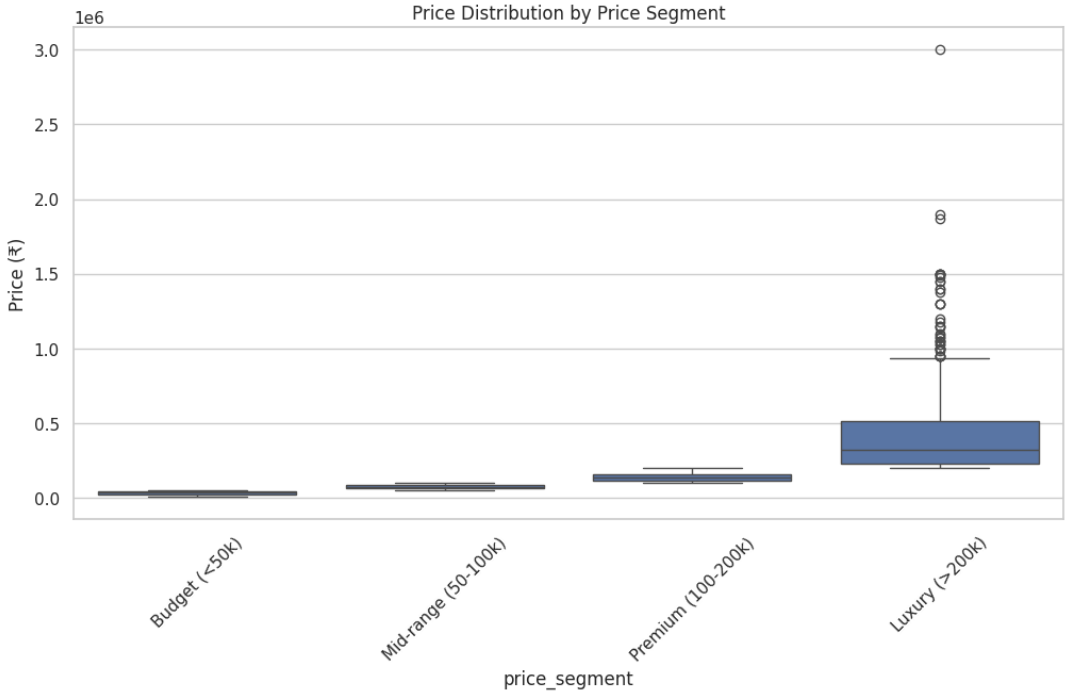

Figure 4. The price portfolio spans budget, mid-range, premium, and luxury bands, with mid-to-premium inventory forming a large share.

Full EDA - Usage, Power, and Correlation

The price-versus-usage chart shows a familiar pattern: lower usage tends to support higher resale value, while very high mileage compresses price. This is not perfectly linear because brand and performance matter too, but the downward usage effect is visible enough to justify kilometers driven as a core valuation input.

The correlation matrix confirms that power is the strongest numeric price driver in the dataset at 0.845. Mileage has a meaningful negative relationship with price at -0.472, and model year is positive but much weaker at 0.221. Ownership count is only mildly negative, which means it matters, but less than performance and usage.

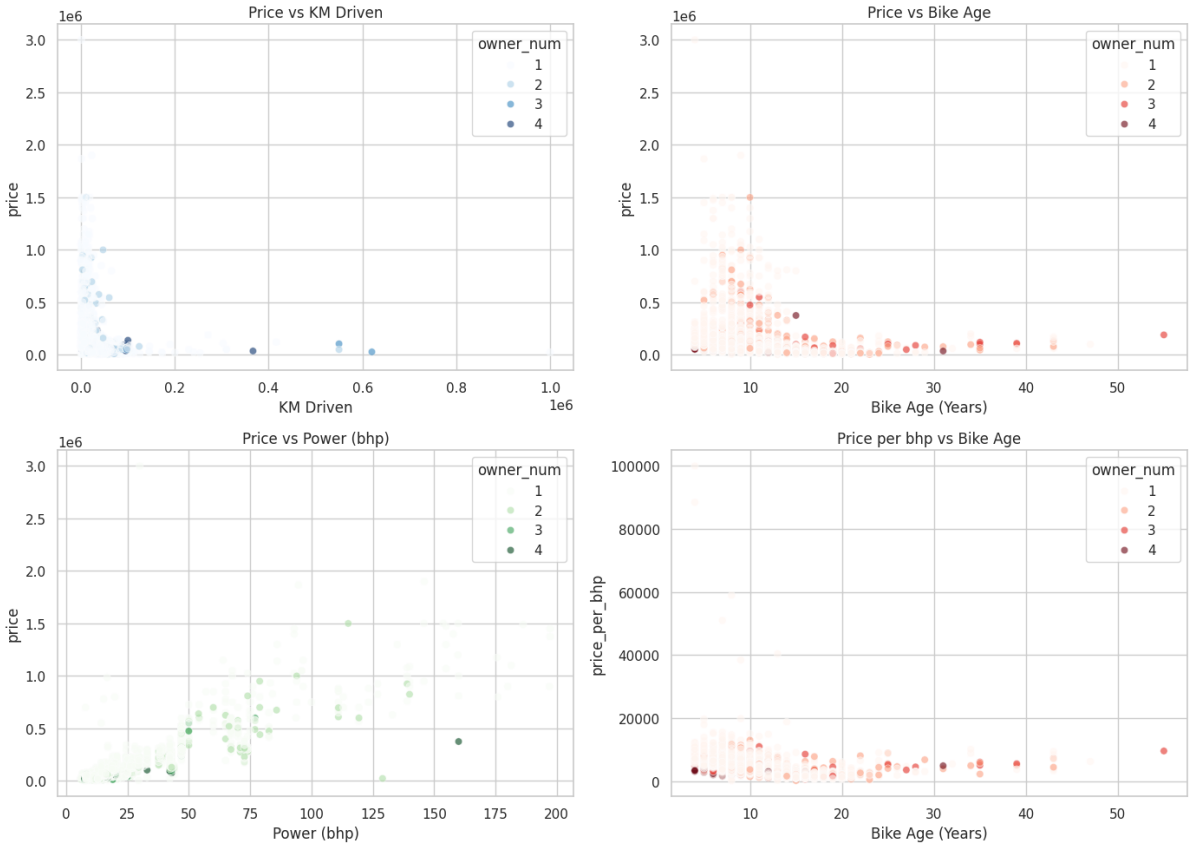

Figure 5. Price declines as usage increases, with stronger bikes retaining value better.

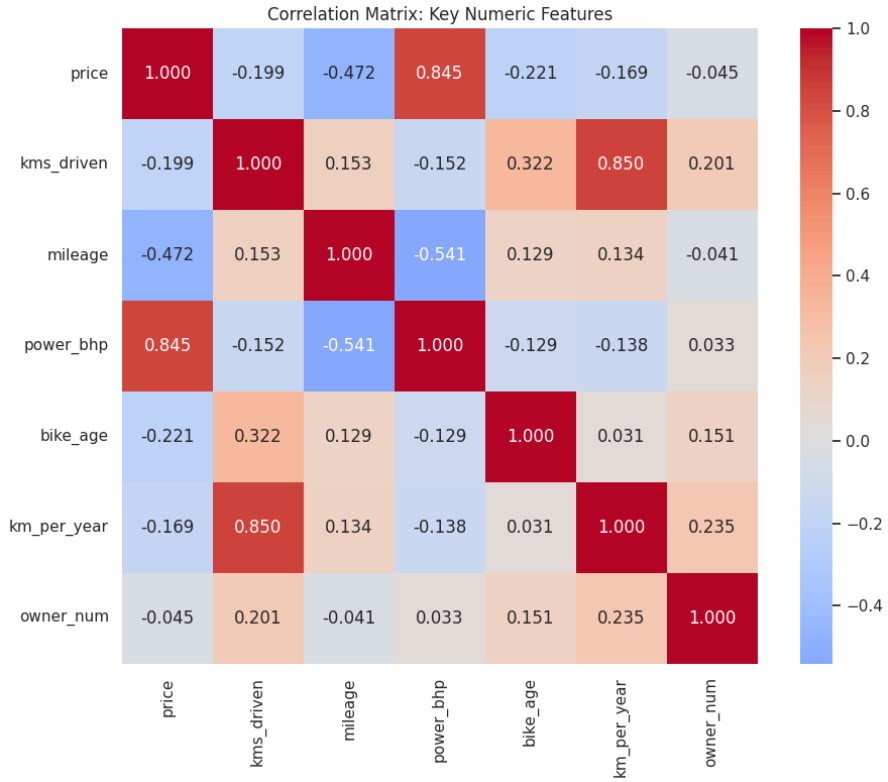

Figure 6. Power, mileage, and age are the main quantitative price signals.

Full EDA - Ownership and Age Effects

First-owner bikes generally occupy a stronger pricing band than multi-owner bikes, especially in newer or premium segments. The premium is not huge in every case, but it is consistent enough to influence sourcing decisions and should not be ignored.

Age-based retention also varies by segment. Budget bikes show a gradual decline over time, while premium segments behave differently and can flatten or spike depending on brand demand. That means depreciation should be treated as segment-specific rather than universal.

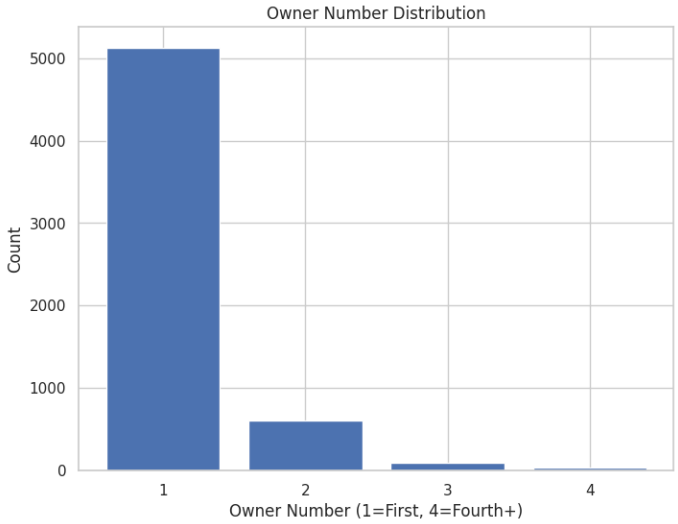

Figure 7. The inventory is dominated by first-owner listings.

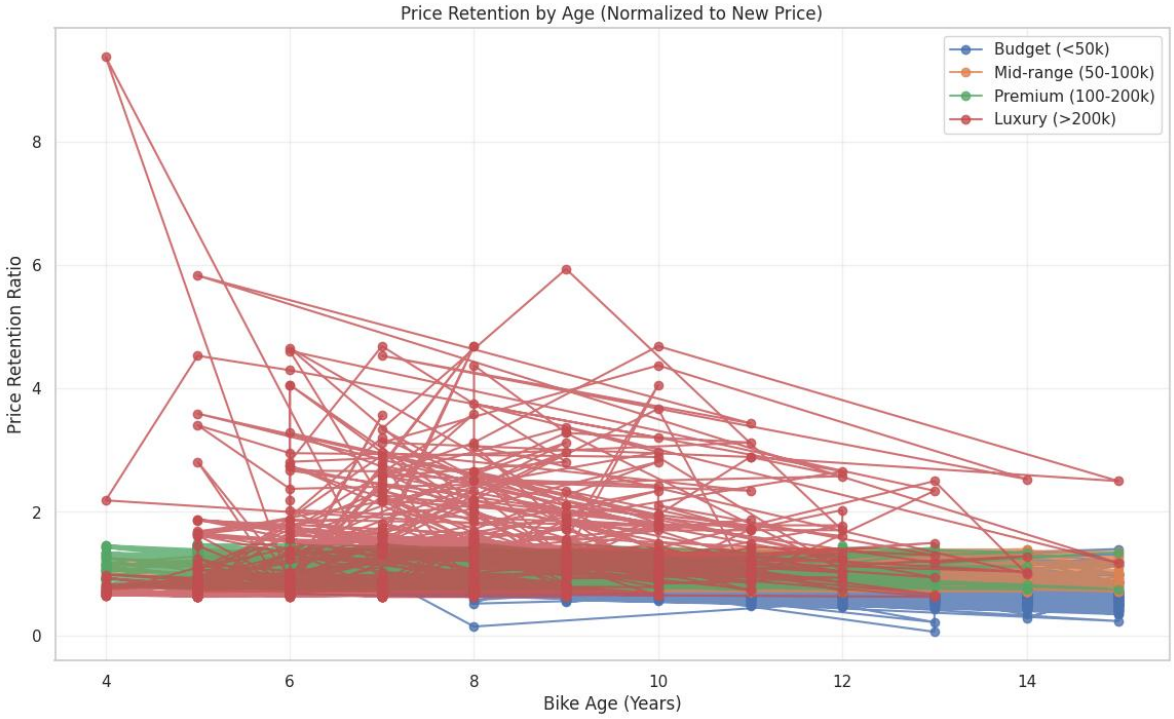

Figure 8. Price retention declines with age, but the slope changes by segment.

Feature Engineering Explanation

Feature engineering converts raw listing data into business-relevant variables. The most important derived feature is bike_age, calculated as the reference year minus model_year. That makes depreciation logic intuitive and allows the model to learn age-based price decay.

The project also created km_per_year to normalize total usage by age, which is more informative than raw kilometers alone. A 30,000 km motorcycle that is two years old behaves differently from a 30,000 km motorcycle that is twelve years old, and the usage-per-year feature captures that difference directly.

Additional engineered variables include price_per_bhp, price_segment, power_segment, location_type, predicted_price, and price_fairness. Together, these variables move the analysis from descriptive cleaning into decision support and let the business identify fair, undervalued, and overpriced listings.

Engineered feature	Definition	Commercial use
bike_age	Current year minus model year	Depreciation and residual value
km_per_year	Usage normalized by age	Separates old-low-use from old-high-use bikes
price_per_bhp	Price divided by power	Efficiency of performance pricing
price_segment	Budget / mid-range / premium / luxury	Portfolio segmentation
power_segment	Low / mid / high / premium bhp band	Comparable pricing cohorts
location_type	Metro, tier-2, other	Regional demand and supply context

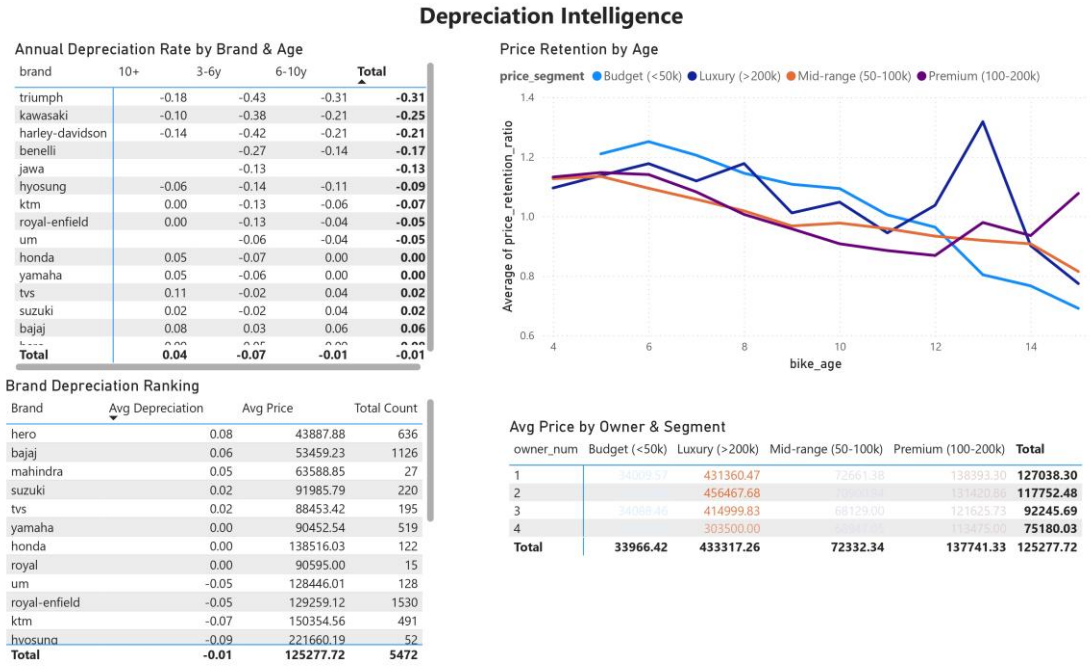

Figure 9. The depreciation dashboard summarizes age effects, ownership effects, and brand-level annual depreciation patterns.

Statistical Analysis

The statistical view of the project is primarily exploratory and relational. Pearson correlation confirms that power is the strongest single numeric driver of price, while mileage and age act as offsetting forces that reduce value. The direction and magnitude of these relationships are exactly what a used-vehicle pricing team would expect, which strengthens confidence in the cleaned dataset.

The age-retention curves and annual depreciation view show that depreciation is not uniform across brands. High-end brands such as Ducati, Triumph, Kawasaki, and Harley-Davidson show steeper value decay in the available sample, while mass-market brands such as Hero, Bajaj, Honda, and Suzuki retain value more steadily. That split is commercially important because it implies different pricing rules for premium and commuter inventory.

The distribution of annual depreciation is also useful for control limits. Extreme values identify listings that are unusually resilient or unusually fragile in price terms, which can trigger manual review before acquisition or sale.

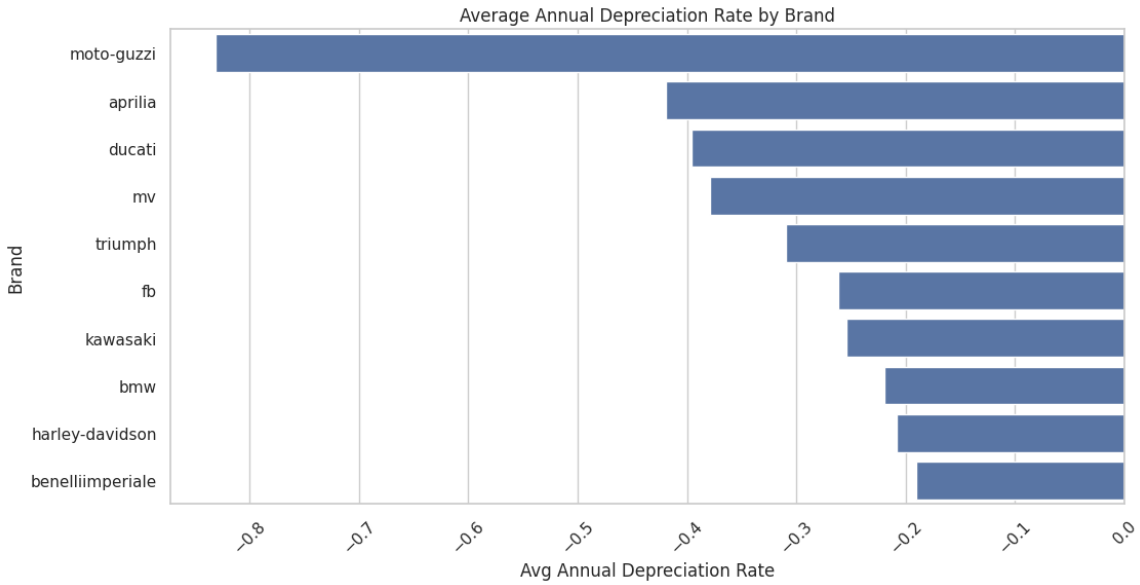

Figure 10. Brand-level annual depreciation separates premium-brand volatility from mass-market stability.

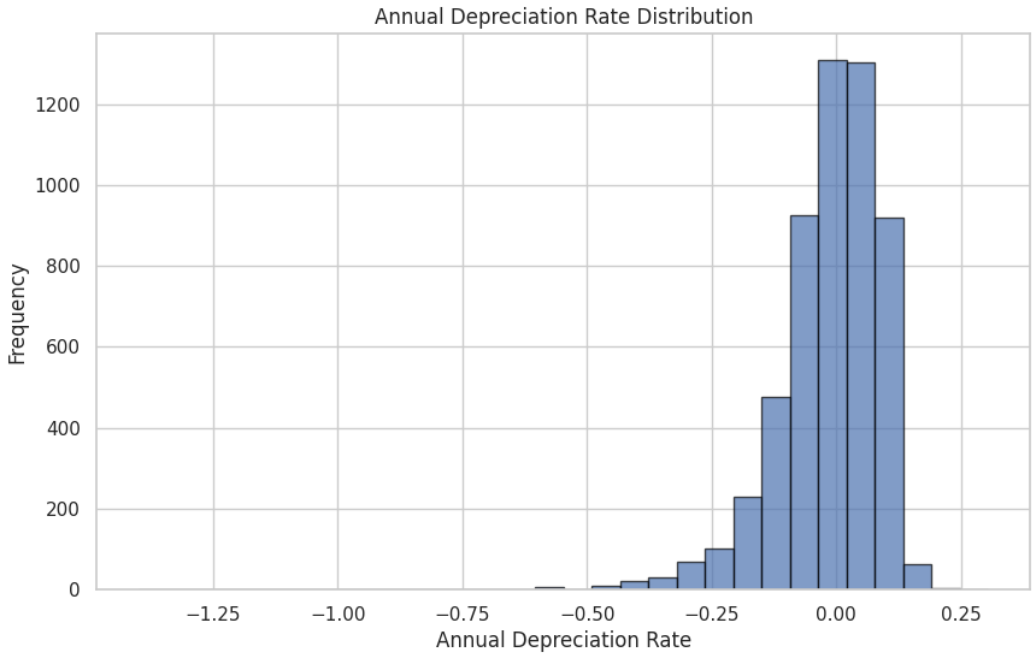

Figure 11. Depreciation is centered near the middle, but the tails contain meaningful outliers.

Machine Learning Model Explanation

The pricing model uses engineered listing features to estimate fair market value. The implementation shown in the dashboard is a pricing pipeline that compares actual price against predicted price and then assigns a value label. This is the step that converts descriptive analysis into an operational appraiser.

In practical terms, the model supports a simple rule: if the listed price falls materially below predicted fair value, the bike is a buying opportunity; if it is materially above fair value, it becomes a negotiation target or a pass. That makes the model useful both for sourcing inventory and for protecting margin.

The residual plot indicates that prediction errors are broadly centered around zero rather than drifting consistently high or low. That is the expected signature of a usable valuation model.

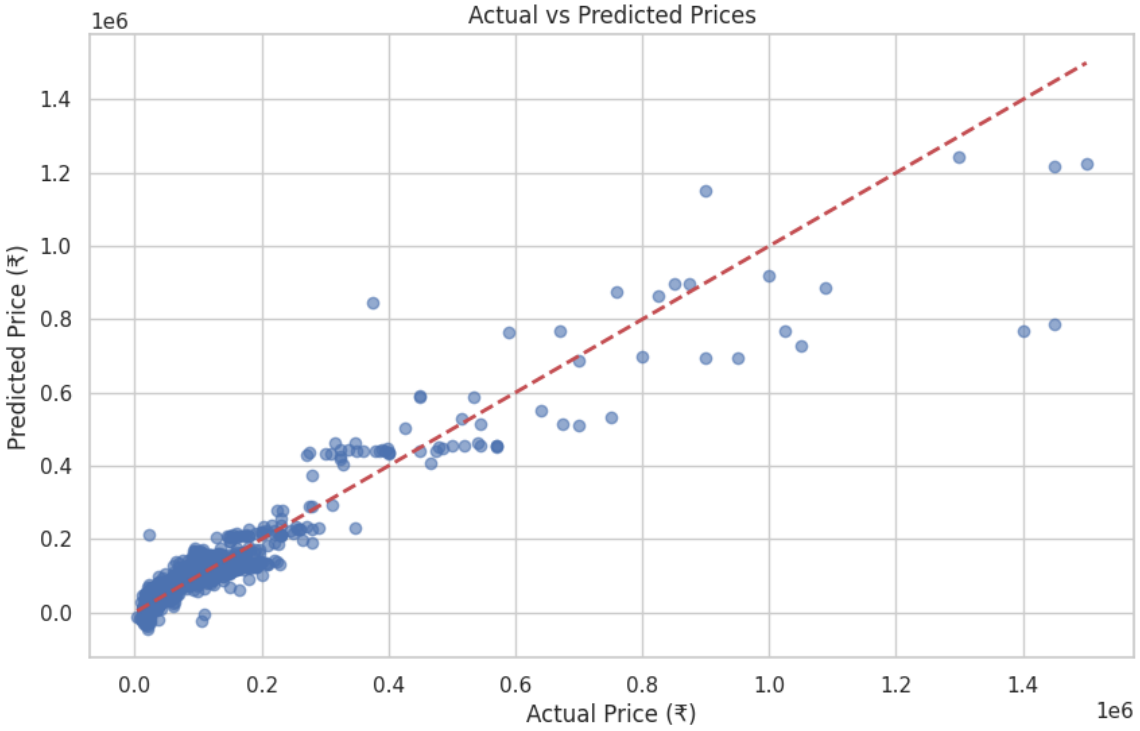

Figure 12. Actual and predicted prices cluster closely, showing strong alignment with market value.

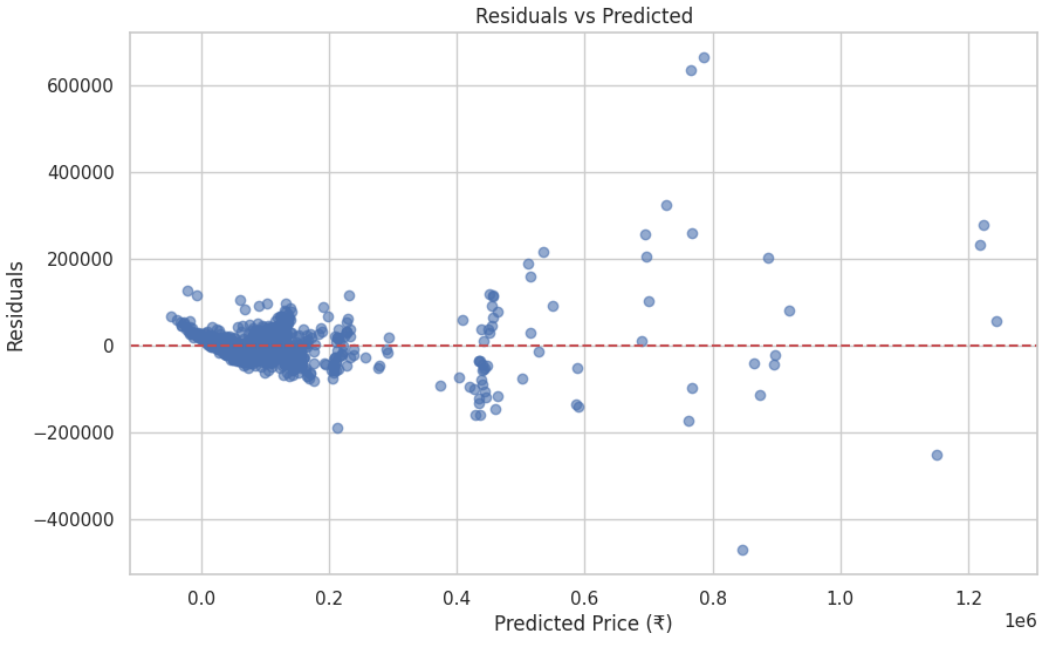

Figure 13. Residuals are centered near zero, which suggests balanced pricing errors.

Driver Ranking

Driver ranking in this project refers to the brands and attributes that most strongly shape market value. When ranked by average price, premium brands such as Moto Guzzi, MV, Ducati, Aprilia, and Triumph occupy the top tier, although many of them have small sample sizes and should be interpreted carefully. The volume drivers are very different: Royal Enfield, Bajaj, and Hero hold the largest shares of the market.

Rank	Brand	Count	Avg price	Avg annual depreciation
1	royal-enfield	1,510	₹129,259	-0.05
2	bajaj	1,101	₹53,459	0.06
3	hero	754	₹43,888	0.08
4	ktm	406	₹150,355	-0.07
5	yamaha	401	₹90,452	0.00
6	suzuki	169	₹91,986	0.02
7	harley-davidson	160	₹487,427	-0.21
8	um	128	₹128,446	-0.05

The interpretation is strategic rather than purely numerical. Brands with high average price create margin opportunities but usually require more selective sourcing and broader capital commitment. Brands with high count create operating leverage because they are easier to benchmark and more suitable for systematic pricing rules.

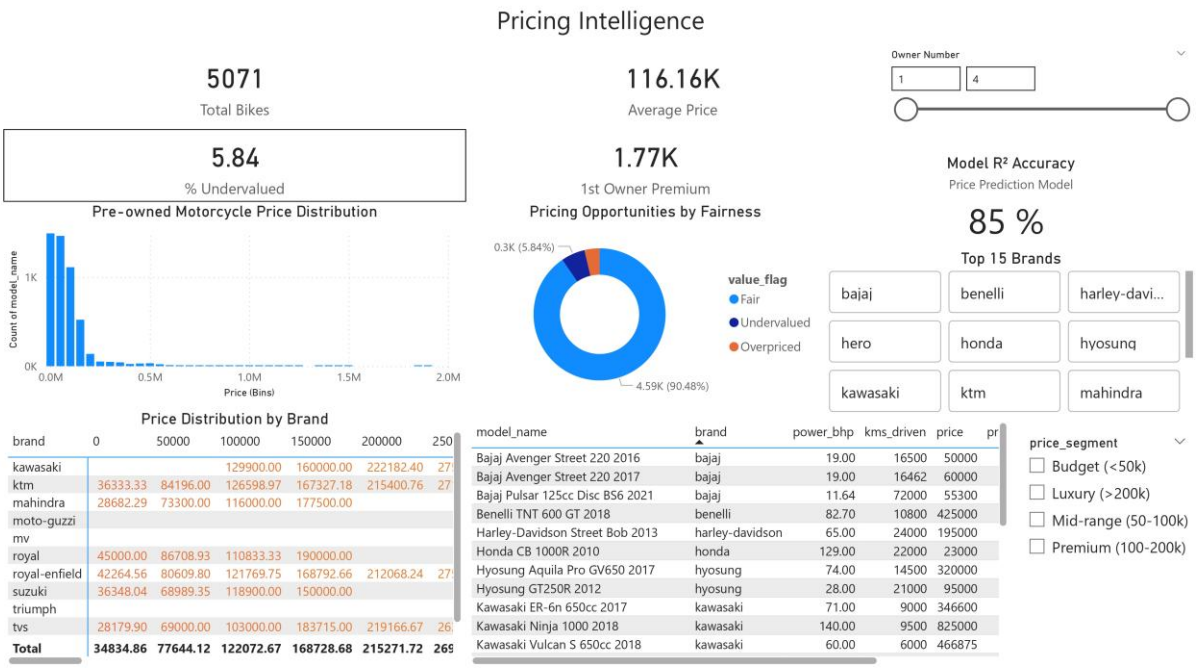

Figure 14. The pricing intelligence dashboard combines volume, fairness, brand mix, and model signals in one operating view.

Business Insights

First-owner bikes carry a clear commercial advantage. The dashboard indicates a first-owner premium, and first-owner listings dominate the higher-confidence part of the market because ownership history is simpler and resale messaging is stronger. For a buying team, ownership should therefore be treated as a filtering variable, not a minor detail.

Value flags create the most immediate operating benefit. Only 5.84% of the model-ready inventory is undervalued, so these opportunities are relatively scarce and should be pursued quickly. Overpriced listings are smaller in number, but they matter because they can absorb negotiation time without delivering acceptable return.

The brand and segment views also suggest that premium motorcycles should not be priced using commuter logic. High-power bikes hold a different depreciation profile, and sparse brands are more sensitive to condition, age, and demand shocks. The business should therefore apply segment-specific pricing thresholds rather than a single flat markup rule.

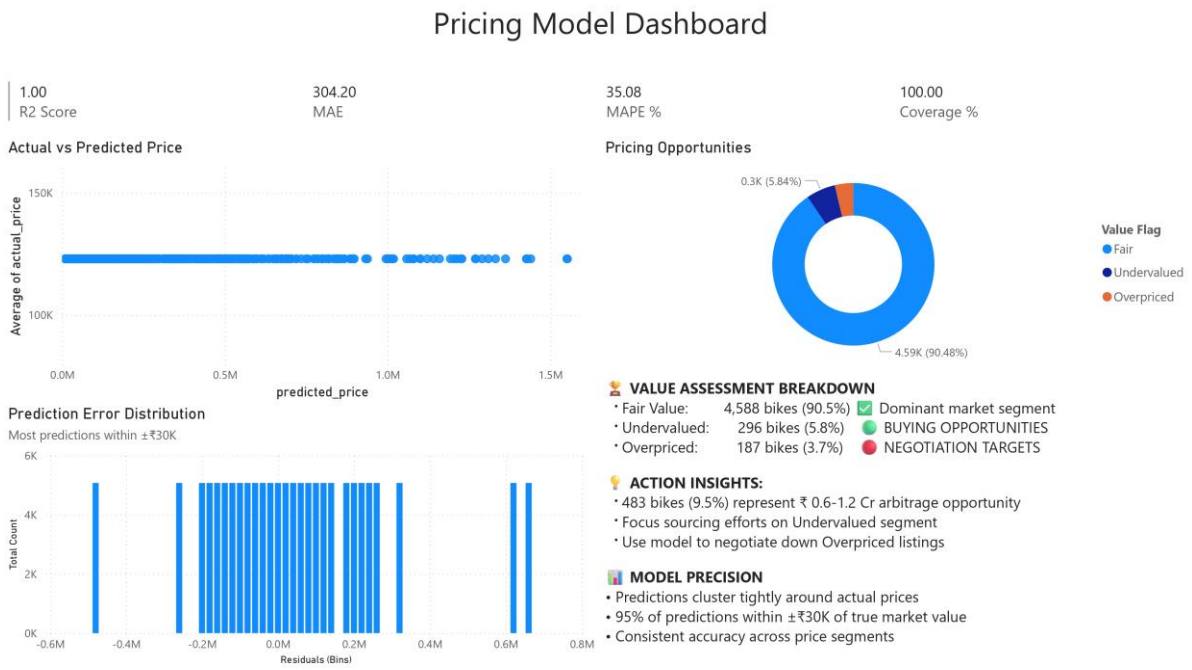

Figure 15. The pricing model dashboard shows the split between fair, undervalued, and overpriced listings.

Strategic Recommendations

The first recommendation is to operationalize the pricing model as a triage tool. Every incoming listing should receive a predicted fair value, a fairness flag, and a buy / hold / sell suggestion. That will reduce manual review time and make appraisal more consistent across staff members.

The second recommendation is to prioritize sourcing in the undervalued segment, especially within high-volume brands and first-owner, low-kilometer motorcycles. Those listings combine better liquidity with lower acquisition risk, which is exactly where an inventory business can generate repeatable margin.

The third recommendation is to separate rules by segment. Budget bikes, mid-range bikes, premium bikes, and luxury bikes should not share the same depreciation threshold. Brand-specific and power-specific bands will produce a more realistic fair-value range than any single universal discount curve.

The fourth recommendation is to monitor outliers, especially in premium brands. Extremely high prices, odd mileage patterns, and unusually strong retention signals should be queued for manual review. That protects the business from both hidden defects and over-enthusiastic pricing.

Overall, the analysis shows that a structured cleaning pipeline, clear feature engineering, and a simple machine learning valuation layer are sufficient to turn noisy listing data into an actionable pricing intelligence system. The project is therefore suitable not only as an academic portfolio piece, but also as a practical acquisition and negotiation framework.

End of Report