Project On

Analysis and Price Prediction of Used Motorcycles in the Indian Resale Market

By

Durga Prasad Govindas
UMID05102560279
UNIFIED MENTOR PRIVATE LIMITED

During The Period
October 2025 - December 2025

Tools: Python, Pandas, NumPy, Matplotlib, Seaborn

Section-1: Executive Summary

The used two-wheeler market in India is influenced by multiple factors such as vehicle age, usage, engine performance, brand value, ownership history, and regional demand. Understanding how these factors affect resale prices is important for buyers, sellers, and platforms involved in price estimation and market analysis.

This project analyzes a real-world dataset of used motorcycles listed across various Indian cities with the objective of identifying key determinants of resale price and developing a reliable predictive model. The dataset underwent extensive data cleaning and preprocessing to handle inconsistent text-based entries, missing values, and outliers. Feature engineering techniques such as logarithmic transformation of price, brand extraction, and categorical encoding were applied to improve interpretability and modeling stability.

Exploratory Data Analysis revealed clear relationships between resale price and variables such as model year, kilometers driven, engine power, brand, owner type, and location. Log transformation of price effectively stabilized variance and converted non-linear relationships into approximately linear patterns, making them suitable for regression-based modeling.

A Linear Regression model was developed using a proper train–validation–test split to evaluate generalization performance. Model evaluation demonstrated consistent performance across all data splits, indicating strong generalization and minimal overfitting. Based on empirical evidence, additional regularization techniques were found to be unnecessary.

Overall, the analysis provides meaningful insights into the used-bike resale market and demonstrates a complete, well-structured machine learning workflow. The results highlight the importance of combining domain knowledge, exploratory analysis, and appropriate modeling choices to build interpretable and reliable predictive systems.

Section-2: Problem Statement

The Indian used two-wheeler market is large and highly dynamic, with resale prices influenced by multiple interrelated factors such as vehicle age, usage, engine specifications, brand reputation, ownership history, and geographic location. Due to the lack of standardized pricing mechanisms, buyers and sellers often face uncertainty in determining a fair resale price.

This project aims to analyze a real-world dataset of used motorcycles to understand the key factors that influence resale prices and to build a reliable, interpretable machine learning model capable of estimating used-bike prices based on these factors.

Section-3: Objectives

- The primary objectives of this project are:
- To perform comprehensive data cleaning and preprocessing to handle inconsistent text-based entries, missing values, and outliers in the dataset.

- To conduct exploratory data analysis (EDA) to identify patterns and relationships between resale price and key variables such as model year, kilometers driven, engine power, mileage, brand, owner type, and location.
- To engineer meaningful features that improve interpretability and modeling performance.
- To develop a regression-based machine learning model for predicting used motorcycle prices.
- To evaluate the model using appropriate regression metrics and validate its generalization performance using a train–validation–test strategy.
- To derive actionable insights about the Indian used-bike resale market based on analytical findings.

Section:4 - Data Description

An initial inspection of the raw dataset was performed to understand its structure, data types, and completeness. The dataset consisted of 7,857 records and 8 columns, as summarized below.

The data types reveal a mix of numerical and categorical variables, with several fields stored as object (text) types despite representing numeric information.

4.1.1 Dataset Overview

- Total records: 7,857
- Total features: 8

Column Name	Non-Null Count	Data Type	Description
model_name	7857	object	Motorcycle brand and model name
model_year	7857	int64	Manufacturing year
kms_driven	7857	object	Distance driven (stored as text with units)
owner	7857	object	Ownership category
location	7838	object	City of listing
mileage	7846	object	Mileage information stored as text
power	7826	object	Engine power stored as text
price	7857	int64	Listed resale price (INR)

Table-1: Datset Descripton

4.1.2 Key Observations

1. Text-based numeric fields

Several variables that represent numeric values (kms_driven, mileage, power) are stored as text, including units, ranges, or descriptive strings. This prevents direct numerical analysis and requires parsing and conversion.

2. Presence of missing values

- location contains **19 missing values**
- mileage contains **11 missing values**

➢ power contains **31 missing values**

While the overall proportion of missing data is small, these values must be addressed to ensure consistency.

3. Consistent target variable

The target variable, price, contains no missing values, making it suitable for supervised learning after preprocessing.

4. Mixed data types

➢ The dataset includes both numerical (model_year, price) and categorical variables (owner, location, model_name), indicating the need for appropriate encoding strategies during modeling.

4.1.3 Implications for Preprocessing

The initial data inspection highlights that the dataset cannot be used directly for analysis or modeling. Several preprocessing steps are necessary, including:

- Conversion of text-based numeric fields into appropriate numeric formats
- Handling of missing values
- Feature engineering to extract useful information from model_name
- Preparation of categorical variables for machine learning models

These steps are discussed in detail in the subsequent **Data Cleaning and Preprocessing** section.

4.2 Data Cleaning & Preprocessing

Real-world datasets often contain inconsistencies, missing values, and non-standard formats that must be addressed before meaningful analysis or modeling can be performed. During this project, several data quality challenges were identified and systematically resolved through a structured preprocessing pipeline.

4.2.1 Handling Text-Based Numeric Fields

Problem Identified

Several variables that represent numeric information were stored as text values:

- **kms_driven** contained entries such as “25,000 Km”, “Mileage 28 Kms”, and other non-numeric strings.
- **mileage** included values with units (e.g., “45 kmpl”), ranges, descriptive text (e.g., “liquid cooled”), and missing entries.
- **power** contained values with units (“bhp”, “kW”), ranges, and additional descriptors (e.g., “13.7 bhp @ 8500 rpm”).

These inconsistencies prevented direct numerical analysis and modeling.

Resolution

- Regular expressions were used to extract only the numeric component from each field.
- All extracted values were converted into consistent numeric formats.
- For power, all values were standardized into brake horsepower (bhp).
- Rows containing non-recoverable or semantically incorrect entries were removed.

This ensured numerical consistency while preserving the majority of the dataset.

4.2.2 Cleaning the kms_driven Column

Problem Identified

The kms_driven column contained a small number of rows where mileage-related information was incorrectly placed instead of distance traveled. These entries did not represent kilometers driven and could not be reliably corrected.

Resolution

- Such semantically invalid rows were identified and removed.
- Remaining values were converted to integer format for analysis.

This prevented incorrect mileage information from distorting usage-based depreciation analysis.

4.2.2 Handling Missing Values

Problem Identified

Although the dataset had a relatively small proportion of missing values, they were present in key variables:

- location
- mileage
- power

Leaving these missing values untreated would reduce usable data during modeling.

Resolution

- Rows with missing location values were removed, as location could not be inferred reliably.
- For numeric fields, missing values were handled after conversion using statistically appropriate measures.
- The final cleaned dataset contained no missing values, ensuring consistency across all features.

4.2.3 Outlier Detection and Treatment

Problem Identified

Initial visualizations of the price variable revealed extreme values, particularly in the higher price range. These outliers had the potential to disproportionately influence model learning.

Resolution

- The Interquartile Range (IQR) method was used to identify extreme price values.
- Obvious data entry anomalies were removed.
- Legitimate high-value motorcycles (e.g., premium brands) were retained where contextually justified.
- Additionally, price values less than or equal to zero were removed.
- This balanced approach ensured robustness without eliminating meaningful premium listings.

4.2.4 Log Transformation of Price

Problem Identified

The resale price distribution was highly right-skewed, violating linear regression assumptions and leading to heteroscedasticity.

Resolution

- A logarithmic transformation was applied to the price variable.
- The transformed target (log_price) exhibited improved symmetry and stabilized variance.
- Subsequent relationships between features and price became approximately linear.
- This transformation significantly improved model interpretability and performance.

4.2.5 Feature Engineering

Several new features were engineered to enhance analytical depth:

Brand extraction:

- The brand name was extracted from the model_name field to enable brand-level analysis.

Kilometer bins:

- kms_driven was discretized into usage-based bins to support categorical analysis of depreciation patterns.

These engineered features allowed both numeric and categorical insights to be explored effectively.

4.2.6 Encoding and Scaling

- Categorical variables (brand, owner, location) were encoded using one-hot encoding.
- Numerical features were scaled using standardization within a modeling pipeline to prevent data leakage.

This ensured compatibility with linear regression and fair feature contribution.

4.2.7 Final Dataset Summary

After completing all preprocessing steps:

- All numeric fields were converted to appropriate numeric types.
- No missing values remained.
- All features were suitable for exploratory analysis and machine learning modeling.
- The final dataset was consistent, interpretable, and model-ready.

Column Name	Non-Null Count	Data Type
model_name	5855	object
model_year	5855	int64
kms_driven	5855	int32
owner	5855	object
location	5855	object
mileage	5855	float64
power	5855	float64
price	5855	int64

Table-2: Data Description After Data Cleaning

Section:5 - Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the distribution of variables, identify relationships between features and resale price, and validate assumptions required for regression-based modeling. Both numerical and categorical variables were analyzed using appropriate visualizations and summary statistics.

5.1 Analysis of Resale Price Distribution

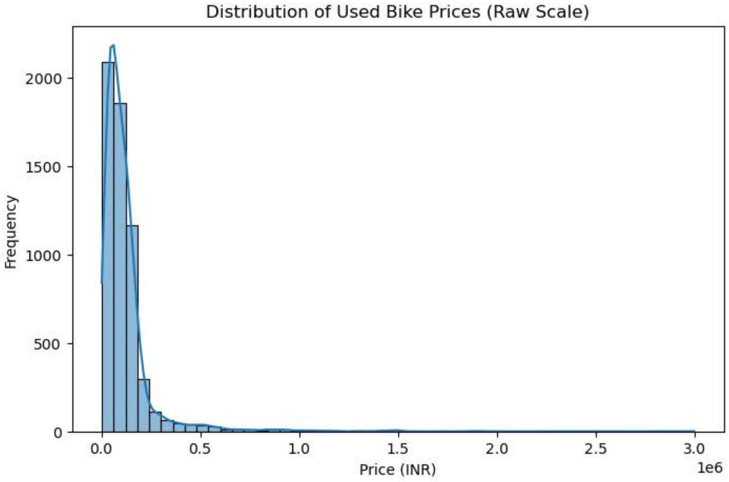


Figure-1: Distribution of Price

Observation

The initial distribution of the resale price (price) was highly right-skewed, with most motorcycles concentrated in the low-to-mid price range and a small number of premium motorcycles listed at very high prices.

This skewness indicated that raw price values could disproportionately influence model learning and violate linear regression assumptions.

Insight

A logarithmic transformation of price was necessary to stabilize variance and reduce the influence of extreme values.

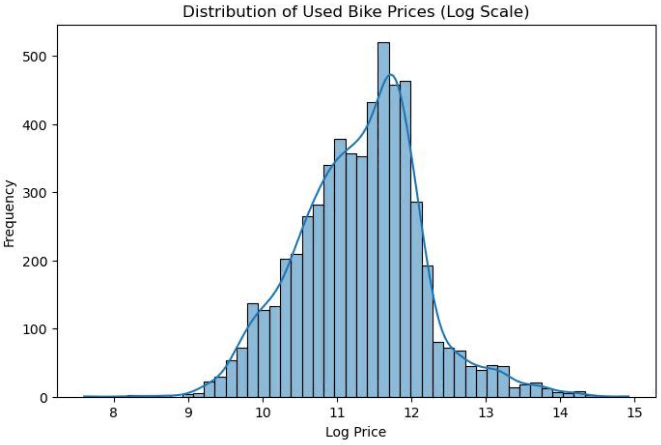


Figure-2: Distribution of Log Price

After applying the log transformation:

- The price distribution became more symmetric
- Variance across observations was significantly reduced
- Relationships with predictor variables became easier to interpret

5.2 Price Vs Model Year

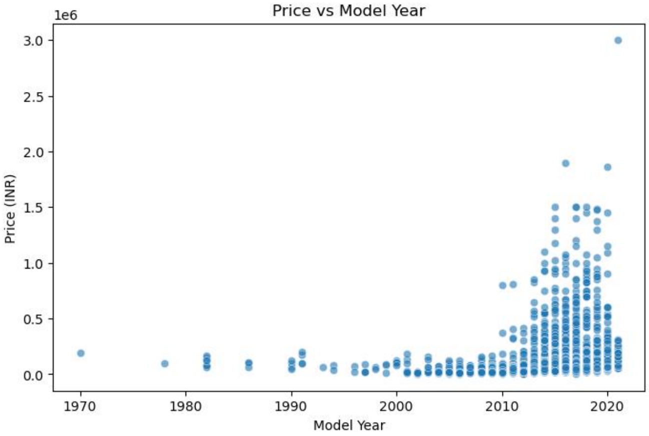


Figure-3: Price Vs Model Year

Observation

Scatter plots of price against model year showed that newer motorcycles generally commanded higher resale prices. However, the relationship was non-linear in the raw price scale, with increasing variance for newer model years.

Resolution and Insight

After analyzing log-transformed price:

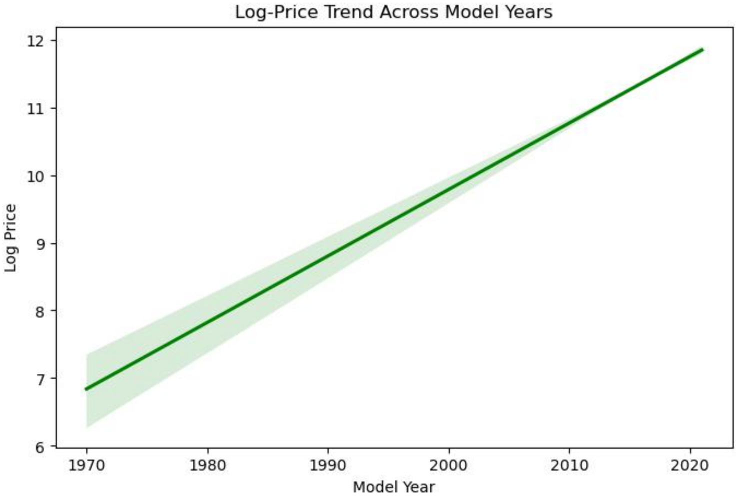


Figure-4: Log Price Vs Model Year

- A clear, approximately linear upward trend emerged
- This confirmed that resale value increases exponentially with newer manufacturing years
- Model year was identified as a strong predictor of resale price

5.3 Price vs Kilometers Driven

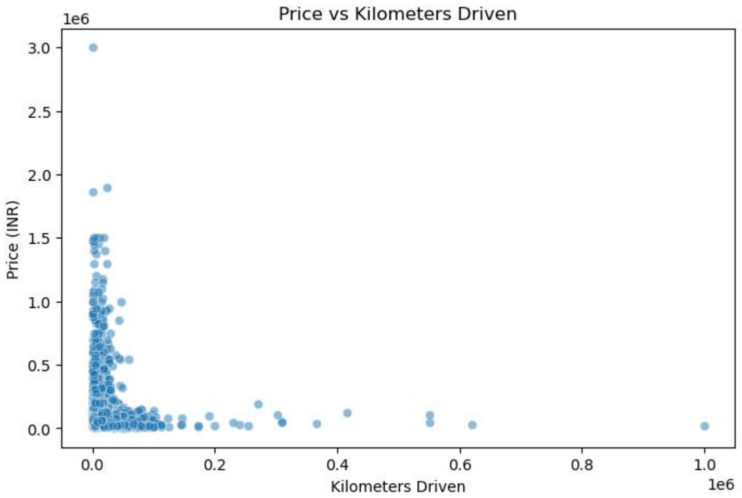


Figure-5: Price Vs Kilometers Driven

Observation

Raw plots of price against kilometers driven showed a dense concentration of points at low kilometer values, with a rapid decline in price as usage increased. Extreme kilometer values introduced noise and made interpretation difficult.

Resolution and Insight

Using log-transformed price revealed:

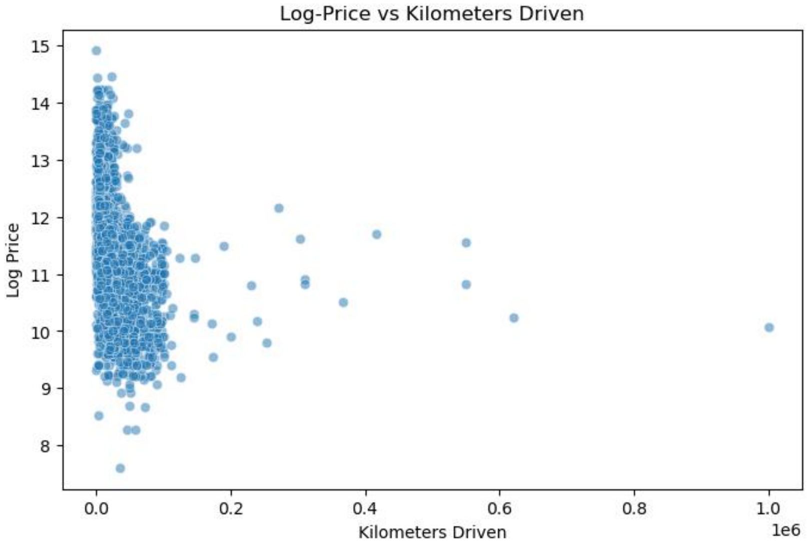


Figure-6: Log Price Vs Kilometers Driven

- A clear negative relationship between kilometers driven and resale price
- Higher usage consistently resulted in lower resale value
- The effect was gradual rather than abrupt, supporting inclusion in a linear model
- This confirmed kilometers driven as a reliable proxy for usage-based depreciation.

5.4 Price vs Engine Power

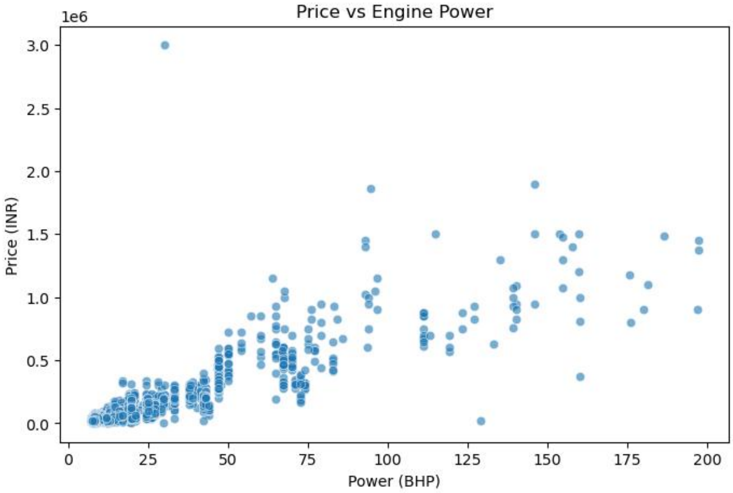


Figure-7: Price Vs Power (BHP)

Observation

Initial scatter plots indicated that motorcycles with higher engine power tended to have higher resale prices, though a few exceptions were observed where low-power motorcycles appeared at higher prices.

Resolution and Insight

Further inspection showed that these apparent anomalies were legitimate premium or special-edition models rather than data errors.

After log transformation:

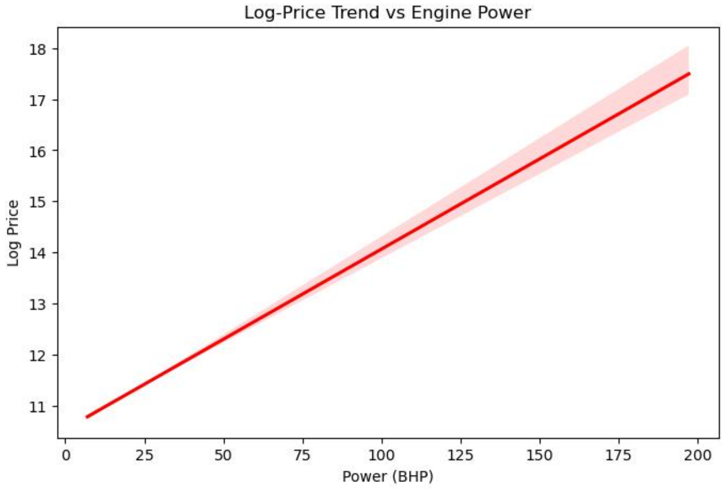


Figure-8: Log Price Vs Power (BHP)

- A strong positive linear trend was observed between engine power and log-price
- Engine power emerged as one of the most influential predictors of resale price

5.5 Correlation Analysis

A correlation heatmap was used to quantify linear relationships between numerical variables.

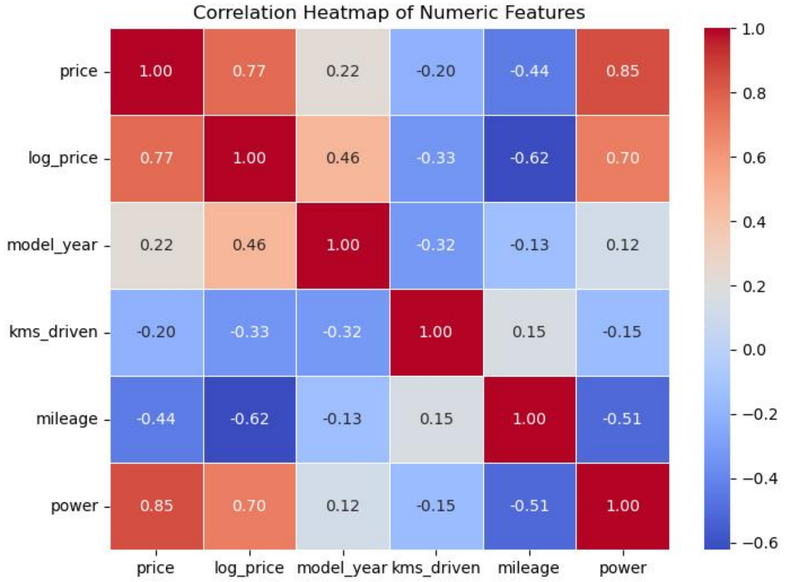


Figure-9: Correlation Heatmap

Key Findings

- log_price showed strong positive correlation with power
- Moderate positive correlation with model_year
- Negative correlation with kms_driven and mileage

These correlations aligned with domain expectations and supported the inclusion of these variables in the regression model.

5.6 Categorical Variable Analysis

Owner Type

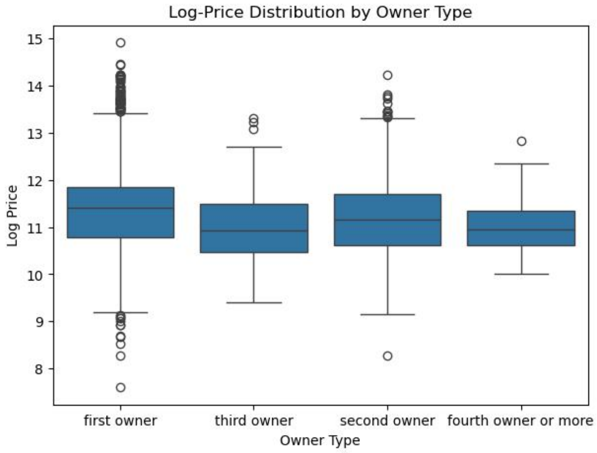


Figure-10: Boxplot for Log Price Based on Owner Type

Boxplots of log-price across ownership categories revealed:

- First-owner motorcycles consistently had higher resale values
- Each additional owner was associated with a clear depreciation effect
- The relationship was monotonic but not linear, justifying categorical treatment

Brand

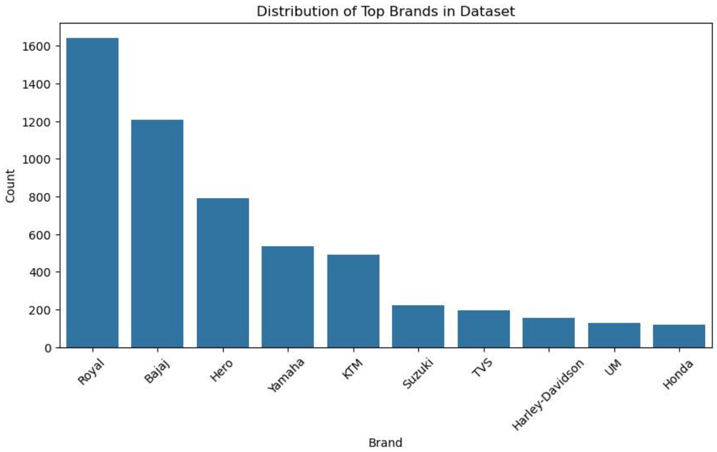


Figure-11: Boxplot for Log Price Based on Brand

Brand-wise analysis showed:

- Strong segmentation between commuter, mid-range, and premium brands
- Premium brands exhibited higher median prices but lower listing counts
- High frequency brands did not necessarily correspond to higher resale values

Location

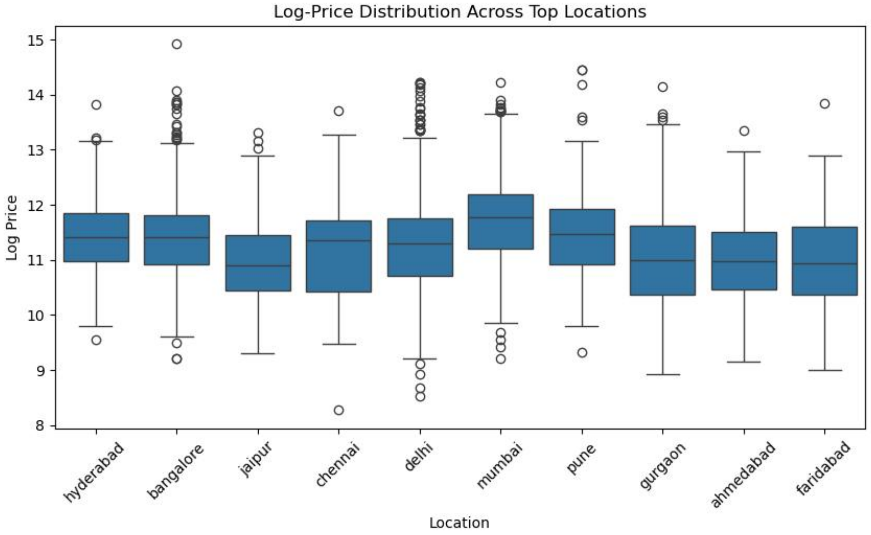


Figure-12: Boxplot for Log Price Based on Location

Location-based analysis revealed:

- Metropolitan cities such as Mumbai, Bangalore, Hyderabad, and Chennai had higher median resale prices
- Tier-2 cities exhibited lower medians with tighter distributions
- Location captured regional demand and purchasing power differences

Section-6: Feature Engineering and Modeling Approach

This section describes how analytical insights from the exploratory phase were translated into features suitable for machine learning, followed by the rationale behind the selected modeling approach.

6.1 Feature Engineering

Feature engineering was guided by both exploratory insights and domain understanding, with the goal of improving interpretability and model stability rather than increasing model complexity.

6.1.1 Log Transformation of Price

Motivation

Exploratory analysis showed that resale prices were highly right-skewed, and relationships between price and predictors were non-linear on the original scale. This violated the assumptions of linear regression and risked disproportionate influence from high-priced listings.

Implementation

- A logarithmic transformation was applied to the target variable (price).
- The transformed target (log_price) exhibited stabilized variance and near-linear relationships with key predictors.
- This transformation enabled the use of a regression-based model while preserving economic interpretability.

6.1.2 Brand Extraction from Model Name

Motivation

The model_name column contained both brand and model information in a single string. Since brand reputation significantly influences resale value, it was necessary to extract this information explicitly.

Implementation

- The brand name was extracted as the first token from the model_name field.
- This allowed brand-level analysis and categorical modeling without introducing additional noise from model-specific naming conventions.

6.1.3 Handling Ownership Information

Motivation

Ownership history directly affects resale value due to perceived wear and trust factors. However, ownership categories are ordinal but not numerically linear.

Implementation

- Ownership was treated as a categorical variable.
- One-hot encoding was applied to prevent the model from assuming equal numeric spacing between ownership categories.
- This preserved the depreciation effect without imposing incorrect linear assumptions.

6.1.4 Location as a Market Indicator

Motivation

EDA revealed clear regional price differences driven by market demand and purchasing power.

Implementation

- Location was retained as a categorical variable.
- One-hot encoding was used to allow the model to learn city-specific price premiums or discounts.

6.1.5 Numerical Feature Preparation

The following numerical features were retained based on analytical relevance:

- model_year
- kms_driven
- power
- mileage

All numerical features were:

- Converted to consistent numeric formats during preprocessing
- Scaled using standardization within the modeling pipeline to ensure balanced contribution

6.2 Modeling Approach

6.2.1 Model Selection Rationale

A Linear Regression model was selected as the baseline predictive model for this project.

The choice was motivated by:

- Strong approximately linear relationships observed after log transformation
- The need for interpretability in understanding resale price drivers
- Satisfactory performance during exploratory evaluation
- Lower risk of overfitting compared to more complex models

- Rather than defaulting to advanced algorithms, model complexity was introduced only if justified by performance limitations.

6.2.2 Data Splitting Strategy

To ensure robust evaluation and prevent overfitting:
The dataset was split into:

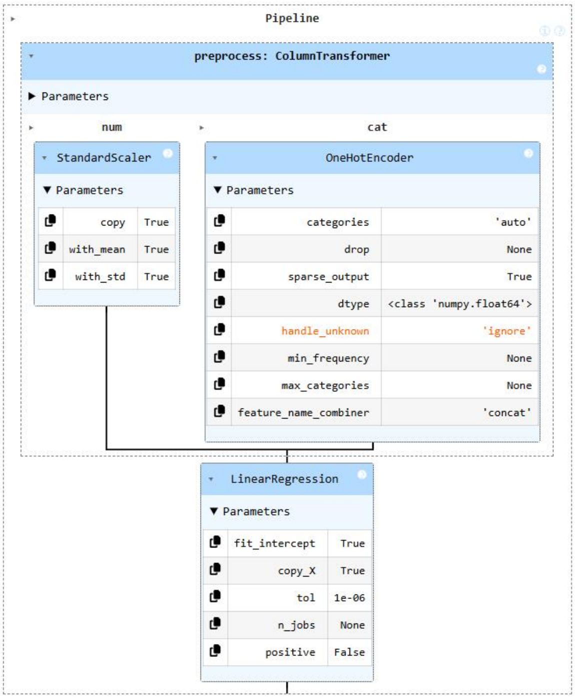
- Training set (70%)
- Validation set (15%)
- Test set (15%)

This allowed:

- Model training on unseen data
- Hyperparameter necessity evaluation
- Unbiased assessment of generalization performance

6.2.3 Scaling and Encoding Pipeline

- To avoid data leakage and ensure reproducibility:
- Numerical features were scaled using standardization
- Categorical variables were one-hot encoded
- All preprocessing steps were encapsulated within a single modeling pipeline
- This ensured that transformations learned from training data were consistently applied to validation and test sets.



Pipeline

preprocess: ColumnTransformer

▶ Parameters

num

StandardScaler

▼ Parameters

copy	True
with_mean	True
with_std	True

cat

OneHotEncoder

▼ Parameters

categories	'auto'
drop	None
sparse_output	True
dtype	<class 'numpy.float64'>
handle_unknown	'ignore'
min_frequency	None
max_categories	None
feature_name_combiner	'concat'

LinearRegression

▼ Parameters

fit_intercept	True
copy_X	True
tol	1e-06
n_jobs	None
positive	False

Figure-13: Model Pipeline

6.2.4 Regularization Consideration

Although regularization techniques such as Ridge and Lasso were considered, empirical evaluation showed:

R² Scores

- Train $R^2 = 0.876$
- Validation $R^2 = 0.852$
- Test $R^2 = 0.867$

RMSE

- Train RMSE ≈ 0.293
- Validation RMSE ≈ 0.322
- Test RMSE ≈ 0.298

This shows,

- Comparable performance between training and validation sets
- No evidence of overfitting
- Stable coefficient estimates
- Therefore, additional regularization was deemed unnecessary for this dataset.

6.3 Summary

Feature engineering and model selection were guided by exploratory findings and domain understanding. Log transformation of the target variable improved linearity and variance stability, while categorical and numerical features were prepared to ensure compatibility with regression-based modeling.

A Linear Regression model was selected due to its interpretability and strong empirical performance. Proper data splitting, scaling, and encoding were applied within a unified pipeline to avoid data leakage. Model complexity was increased only when justified, and regularization was evaluated but found unnecessary.

Overall, the modeling approach prioritized robustness, transparency, and generalization, forming a strong foundation for subsequent model evaluation and interpretation.

Section 7: Model Evaluation and Results

This section evaluates the performance of the developed regression model and examines its ability to generalize to unseen data. Appropriate regression metrics were used to assess predictive accuracy and model stability.

7.1 Evaluation Strategy

To obtain an unbiased estimate of model performance and diagnose potential overfitting, the dataset was divided into three subsets:

- **Training set** – used for model fitting
- **Validation set** – used to assess generalization during model development
- **Test set** – used for final performance evaluation

This train–validation–test strategy ensured that performance conclusions were not based on a single data split and reduced the risk of optimistic bias.

7.2 Evaluation Metrics

Since the task is a regression problem, the following metrics were used:

R^2 (Coefficient of Determination):
Measures the proportion of variance in the target variable explained by the model.

Root Mean Squared Error (RMSE):
Measures the average magnitude of prediction error on the log-transformed price scale.

Classification metrics such as accuracy, precision, recall, and ROC–AUC were not applicable and were therefore not used.

7.3 Model Performance

The Linear Regression model was evaluated on all three data splits, producing the following results:

Dataset	R^2 Score	RMSE
Training Set	0.875797829	0.29293025
Validation Set	0.852037361	0.321615985
Test Set	0.867959868	0.298110999

Table-3: Model Performance

7.4 Interpretation of Results

Several important observations emerge from these results:

1. Strong Predictive Performance
The model explains a substantial proportion of variance in resale prices, indicating that the selected features capture the key drivers of used-bike pricing.

2. Consistent Generalization
R^2 and RMSE values are comparable across training, validation, and test sets. The absence of a significant performance gap suggests that the model generalizes well and is not overfitting.

3. Model Stability
Validation performance closely aligns with test performance, indicating that the modeling decisions made during development were robust and not overly sensitive to the data split.

4. Effectiveness of Log Transformation
Stable error metrics across splits confirm that the log transformation of price successfully improved variance stability and linear model suitability.

7.5 Regularization Assessment

Although regularization techniques such as Ridge and Lasso regression were considered, the evaluation results showed no evidence of overfitting or unstable coefficients. Given the model's strong generalization and consistent error metrics, additional regularization was deemed unnecessary.

This decision preserved interpretability while maintaining high predictive accuracy.

7.6 Summary of Evaluation

Overall, the Linear Regression model demonstrated:

- High explanatory power
- Stable performance across all data splits
- No signs of overfitting or underfitting
- Suitability for modeling resale prices after appropriate preprocessing

7.7 Residual Diagnostics

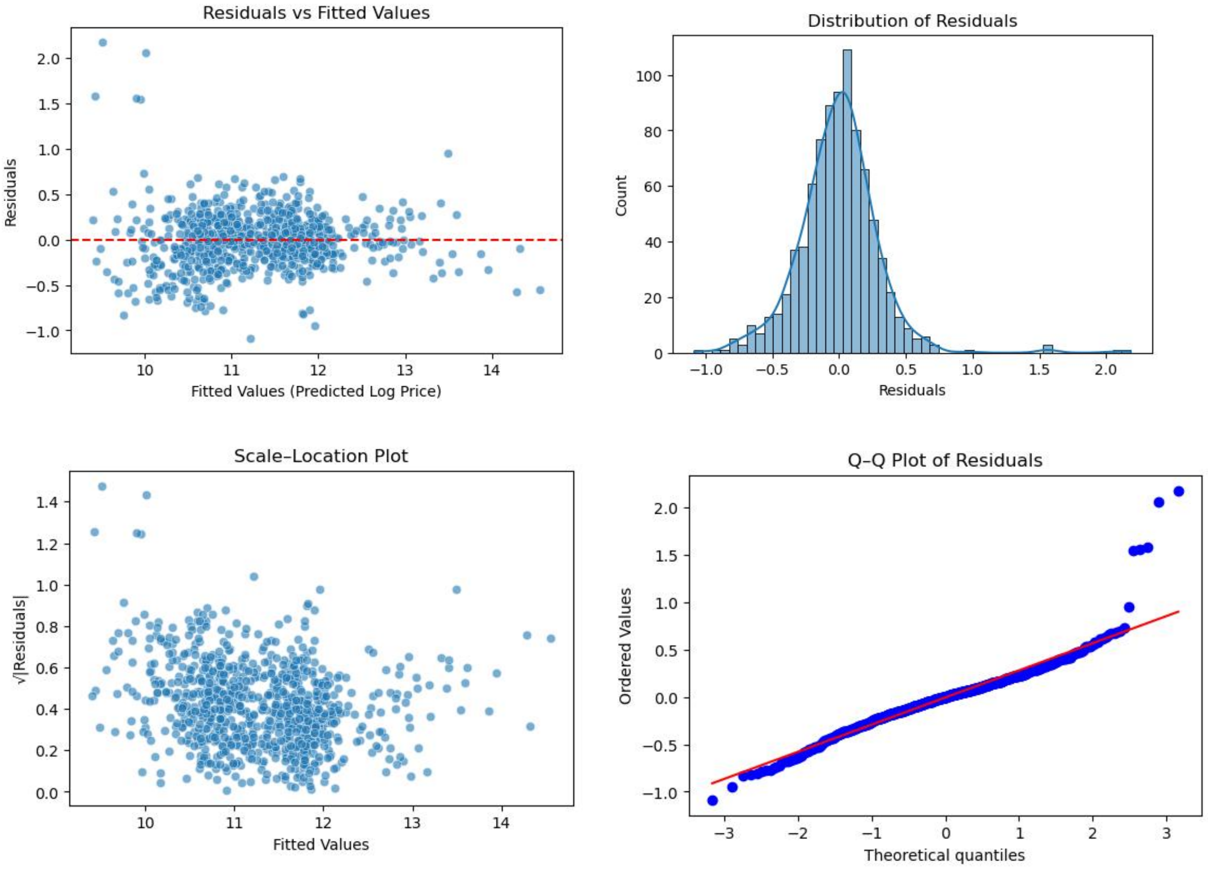


Figure-14: Residual Diagnostics Results

Residual diagnostics were conducted to validate the assumptions of linear regression. The residuals versus fitted values plot showed no systematic pattern, indicating that linearity and independence assumptions were reasonably satisfied. The scale–location plot suggested approximately constant variance across fitted values, confirming homoscedasticity after log transformation of the target variable.

The histogram of residuals exhibited an approximately normal distribution centered around zero, and the Q–Q plot showed residuals closely following the theoretical normal line, with minor deviations in the upper tail. Such deviations are expected in real-world resale pricing data due to the presence of premium vehicles. Overall, residual diagnostics confirmed that the linear regression model was appropriate and well-specified for this problem.

Section-8: Key Insights

Based on exploratory analysis and model evaluation, the following key insights were derived from the study:

1. Model Year and Engine Power Are Primary Price Drivers
Newer motorcycles and those with higher engine power consistently command higher resale prices. These variables showed strong positive relationships with the log-transformed resale price, indicating that performance and recency play a critical role in value retention.

2. Usage-Based Depreciation Is Significant
Kilometers driven exhibited a clear negative relationship with resale price. Higher usage leads to predictable depreciation, reinforcing its importance as a pricing factor in the used-bike market.

3. Brand Equity Strongly Influences Resale Value
Premium brands retain higher resale prices despite lower listing counts, while commuter brands dominate volume but depreciate faster. This highlights the role of brand perception beyond technical specifications.

4. Ownership History Impacts Buyer Trust and Price
Motorcycles with fewer previous owners consistently achieve higher resale values. Each additional owner introduces a depreciation effect, reflecting buyer preference for lower-risk listings.

5. Location Reflects Market Demand Differences
Metropolitan cities exhibit higher resale prices and greater price variability compared to tier-2 cities. Regional demand and purchasing power significantly affect pricing outcomes.

6. Mileage Acts as a Segment Indicator
Higher mileage (fuel efficiency) is associated with lower resale prices, not due to quality reduction, but because high-mileage motorcycles typically belong to commuter segments with lower market valuation.

7. Log Transformation Improves Modeling Effectiveness
Transforming resale price using a logarithmic scale stabilized variance, improved linear relationships, and enhanced model performance without sacrificing interpretability.

Section9: Conclusion and Limitations

Conclusion

This project successfully analyzed the Indian used motorcycle resale market by combining domain knowledge, exploratory data analysis, and a structured machine learning workflow. Through careful preprocessing, feature engineering, and model evaluation, key determinants of resale price were identified and quantified.

A Linear Regression model, applied to a log-transformed target variable, demonstrated strong and consistent predictive performance across training, validation, and test datasets. The empirical evaluation confirmed that additional model complexity or regularization was unnecessary, reinforcing the effectiveness of a transparent and interpretable modeling approach.

Overall, the study highlights how disciplined data preparation and thoughtful model selection can produce reliable insights and accurate predictions in real-world pricing problems.

Limitations

While the analysis provides valuable insights, several limitations should be acknowledged:

1. Lack of Transaction Date Information
The dataset does not include listing or sale dates, limiting the ability to account for temporal price changes or inflation effects.

2. Model Scope
The analysis focused on linear relationships after transformation. Although effective, non-linear models may capture additional interactions at the cost of interpretability.

3. Market Representation
The dataset represents listing activity rather than confirmed sales, which may introduce minor bias in estimated pricing behavior.