Andy Bostick

Capstone Project Report: Predicting Used Car Values in India

June 15, 2025

Executive Summary

This report presents a comprehensive, data-driven framework for pricing used cars in India, leveraging regression analysis to pinpoint and quantify the most influential factors shaping used vehicle values. Drawing on a robust dataset of used car sales, the analysis reveals that attributes such as fuel type, brand prestige, vehicle age, power, and transmission type are substantial and statistically significant predictors of price variation. Notably, electric and premium brand vehicles command substantial price premiums, while older, high-mileage, and manual transmission vehicles are associated with lower sales prices.

The predictive power of seven regression algorithms was critically evaluated to derive these conclusions. Ultimately, an Ordinary Least Squares (OLS) model was selected for its strong predictive performance and ease of interpretability. The model demonstrates high explanatory power, with key variables statistically validated as significant price predictors. These insights form the basis for actionable recommendations: dynamic pricing strategies tailored to vehicle features, targeted inventory acquisition focused on high-demand segments, and differentiated clearance tactics for less desirable stock.

Adopting advanced analytics and evidence-based decision-making empowers dealerships and industry stakeholders to optimize pricing, improve inventory management, and maintain a competitive edge in India's rapidly evolving used car market.

Introduction

India's used car market is experiencing unprecedented growth and transformation, fueled by rising consumer demand for newer, more efficient vehicles. This dynamic environment presents both substantial business opportunities and complex challenges for sellers and other industry stakeholders. Traditional pricing estimation methods based on intuition or limited observation are increasingly inadequate for capturing the interplay of factors determining vehicle value.

This report employs a rigorous Ordinary Least Squares (OLS) regression analysis of a vehicle sales dataset to identify and quantify the most significant predictors of used car prices. By systematically selecting predictive features and evaluating the performance of multiple modeling techniques, this analysis delivers actionable insights and practical recommendations for optimizing pricing, inventory management, and business strategy. These findings ultimately empower sellers to remain profitable and competitive within India's rapidly evolving used vehicle marketplace.

India's market for used vehicles:

India's used car market is in a state of remarkable transformation:

- **India's used car industry has experienced rapid growth**, reaching 5.4 million units sold in 2024 and is projected to surpass 10.8 million units by 2030 (SIAM, 2024; Little, 2024)
- **In 2024, 5,412,945 used vehicles were sold in India** (Business Standard, 2024)
- **Used car sales outpace new car sales,** expanding at a rate of 13% annually (SIAM, 2024; Little, 2024).
- **Rising new car prices have made pre-owned vehicles more attractive**, especially in non-urban areas (Hindustan Times Auto, 2024; CARS24, 2025).
- **The average age of used cars sold has decreased, and vehicle ownership turnover rates have accelerated**, reflecting evolving consumer preferences for newer vehicles (CARS24, 2025).

Together, these factors indicate the market poses tremendous opportunities to apply data-driven business solutions.

Opportunities for data insights:

India's rapidly expanding used car sector is poised for transformation in the face of data innovation:

- The rapid growth and formalization of India's used car market, combined with increased digitalization and diverse buyer demographics, provide dealers with **vast amounts of data that can be leveraged for dynamic pricing** (Little, 2024; CARS24, 2025)
- Dynamic pricing allows dealers to adjust prices instantly based on market trends, inventory, and buyer demand (ReportLinker, 2025; Malhotra, 2025).
- This strategy ensures competitive pricing, optimizes inventory control, and maximizes profitability (ReportLinker, 2025; Malhotra, 2025).

By harnessing advanced analytics and dynamic pricing strategies, dealers can maximize profits and maintain long-term market competitiveness (Little, 2024; CARS24, 2025; ReportLinker, 2025).

Problem Statement

Setting an optimal price for a used car in India is a complex task, as **it is influenced by a wide range of factors such as vehicle age, mileage, brand, fuel type, condition, location, and market dynamic**s (Little, 2024; CARS24, 2025; Malhotra, 2025; Hindustan Times Auto, 2024; ReportLinker, 2025). Traditional pricing methods that rely on intuition or simple observation are insufficient in this dynamic market, frequently leading to suboptimal pricing and missed opportunities. Accordingly, **there is a growing need for data-driven, quantitative models** that can accurately capture the interplay of multi-dimensional factors and support dynamic, evidence-based pricing decisions (IJCTT, 2024; Vehicle Report, 2023)

Background

According to industry sources, the prices customers are willing to pay for vehicles are subject to myriad factors, including:

- **Car Age:** Newer vehicles typically command higher prices (Arthur D. Little, 2024)
- **Mileage:** Cars with lower kilometers driven are generally perceived as less worn and more desirable (CARS24, 2025).
- **Brand:** Well-known brands retain value better in the resale market (Malhotra, 2025).
- **Fuel Type:** Diesel vehicles command higher resale values due to their fuel efficiency. However, demand for petrol and electric vehicles is increasing (Arthur D. Little, 2024).
- **Condition and Service History:** Cars with a strong service record and accident-free history are valued higher (CARS24, 2025).
- **Location:** Urban markets tend to see higher used car prices than suburban markets due to greater demand and higher operating costs (Hindustan Times Auto, 2024).
- **Transmission:** Depending on buyer preferences, transmission type can influence resale value (Malhotra, 2025).
- **Market Dynamics:** The shortage of newer, low-mileage vehicles increases prices (ReportLinker, 2025).

Data analysis must determine the extent to which these features, severally and jointly, affect used car prices so as to inform dynamic pricing models.

Objective

Dealers must know how to price vehicles accurately, as doing so is essential for dealership profitability and competitiveness in India's rapidly growing used car market. This analysis aims to develop an evidence-based pricing model, using the best of seven different regression analysis methodologies, to predict used car prices with high accuracy. The resulting price estimations will inform differential pricing strategies and inventory acquisitions tailored to customer preferences and market conditions, enabling businesses to set competitive, profitable prices for individual vehicles.

Desired Findings

This analysis utilizes data-driven insights to achieve the following goals:

- **Identify statistically significant predictors of used car price**
- **Quantify the impact** each feature has on the price of used vehicles
- **Provide actionable recommendations** for pricing and inventory management

The findings of this analysis will provide sellers with actionable, data-backed strategies to **optimize pricing decisions, improve inventory management, and maintain competitiveness** in India's rapidly expanding used car market.

Methodology

This analysis employs a systematic method to extract meaningful insights and transform them into actionable business insights. The analysis is comprised of the following components:

1. **Exploratory data analysis**
2. **Data cleaning**
3. **Variable selection for the final regression model**
4. **Compare the performance of different regression modalities**
5. **Select the regression model most appropriate for predicting used car prices and informing business decisions**
6. **Interpret the model's results**
7. **Provide data-informed strategies for used car pricing and inventory management**

This comprehensive approach maximizes data abstraction to derive accurate and actionable pricing and inventory management strategies.

Data Overview

This study utilizes a comprehensive dataset **containing detailed pricing and descriptive information for 7,252 vehicles collected in 2019** to build an effective predictive model for used car prices in India. This dataset contains the following key variables used in this examination of pricing predictors:

- **Price_Log:** Log-transformed used car price (target variable)
- **Fuel_Type:** Petrol, Diesel, Electric, CNG, LPG
- **Transmission:** Manual, Automatic
- **Power:** Engine power (bhp)
- **Kilometers_Driven_Log**: Log-transformed mileage
- **Years_Old:** Vehicle age (years)
- **Premium_Brand:** Manufacturers of vehicles commanding premium prices
- **Urban Location**: Bangalore, Chennai, Delhi, Hyderabad, Kolkata, Mumbai, and Pune

While limited in scope, the dataset is sufficient to yield accurate pricing predictions.

Data Preprocessing

The following issues required resolution before proceeding to the modeling phase:

- **Missing Values:**
    - **Removed all rows with missing values for target variables**, eg, price and price_log
    - **Imputed missing values** for numerical features, using median and mean values depending on the distributions of their observed values.
    - **Identified and removed implausible entries**, such as cars with impossibly high mileage

- **Transformations**:
    - **Log-transformed** price and kilometers driven to address skewness.
    - **Created dummy variables** for categorical variables (Fuel_Type, Transmission)

- **Feature Engineering**
    - Begin with features known to most directly influence prices
    - **Create variables that provide meaningful, computationally efficient, and easily** interpretable information
    - Select variables for model inclusion by maximizing the retention of information while keeping VIF scores in the acceptable range

These comprehensive preprocessing steps ensured a clean, well-structured dataset, setting a solid foundation for robust and reliable modeling. The Annotated Data Dictionary appendix provides further details on data selection and alterations.

Building the model

To estimate used car prices accurately, a regression model was constructed using this carefully selected set of predictive features:

- **Target variable:**
    - Price_Log (log-transformed price in ₹ 100,000)

- **Independent variables:**
    - Years_Old
    - Kilometers_Driven_Log
    - Premium Brand
    - Urban location
    - Power
    - Fuel_Type
    - Transmission

These features enable a nuanced understanding of how each influences price, supporting the data-driven pricing decisions.

Model Validation

To ensure the reliability and accuracy of the regression model, a comprehensive validation process was conducted using several statistical checks, including:

- Train-test split
- VIF checks
- Check mean residuals
- Test for homoscedasticity
- Check the linearity of variables
- Check the normality of error terms

These validation steps confirmed that the model satisfied key assumptions, supporting its predictive reliability.

Modelling

To identify the most effective approach for predicting car prices, the following regression algorithms were systematically evaluated:

- Linear regression
- Ordinary Least Squares (OLS) regression
- Ridge regression
- Decision tree
- Random forest
- Hyperparameter-tuned decision tree
- Hyperparameter-tuned random forest models

The comparative performance of these algorithms facilitated the selection of the model best-suited to provide accurate and reliable predictions.

Regression Model Evaluation

To quantify accuracy and predictive capacity, each model's performance was assessed based on its R^2 scores and Root Mean Square Errors (RMSE) values. The results are presented below:

Comparison of Regression Evaluation Summary

Model	Testing R^2	Training R^2	Testing RMSE	Training RMSE	Notes
Decision Tree	1.000	0.767	5.378	0.248	Significant overfitting
Tuned Random Forest	0.919	0.970	0.250	0.152	Some underfitting
Tuned Decision Tree	0.901	0.901	0.277	0.265	Strong performance
OLS	0.882	0.866	0.302	0.318	Strong performance
Random Forest	0.854	0.964	4.258	2.108	Underfitting observed
Ridge	0.793	0.607	5.073	7.001	Underfitting observed
Linear Regression	0.793	0.610	5.068	6.976	Significant underfitting

Summary of findings

These are the key insights from the above exercise:

- Several regression models displayed significant over-fitting or under-fitting tendencies
- **The hyperparameter-tuned random forest and OLS regression models performed best,** with good explanatory power and accuracy
- **Random forest models do not yield traditional coefficients**, but instead report feature importances
- **The OLS model was selected for analysis** by virtue of its strong performance and ease of interpretation for a business audience

The OLS model is analyzed further in the section below.

Interpretation of OLS Evaluation Scores

The OLS Regression was proven robust and reliable on the basis of these findings:

- **Strong explanatory power** on test data ($R^2 = 0.882$), indicating good accuracy
- Testing R^2 scores (0.882) and Training R^2 (0.866) are close, indicating t**he model is not overfitting or underfitting**
- Low Testing RSME score (0.302), indicating low residuals that confirm t**he model's predictions are closely matched to observed data**
- **The similarity in RSME scores** between training and test models (0.302 vs 0.318) indicates **the model generalizes well**

Why did the OLS regression model perform well?

The advantages of the OLS model are:

- OLS provides the **Best Linear Unbiased Estimator for the coefficients**
- OLS models **remain valid with non-normal independent variables**
- **However, careful feature selection is essential** as OLS models are sensitive to multicollinearity issues

The rigorous comparison of multiple regression algorithms identified the OLS regression model as the optimal choice. This model balances strong predictive accuracy with clear interpretability for business decision making.

Key Findings

The regression analysis identified several features that significantly influence used car prices. The most impactful predictors are summarized below:

OLS Regression Analysis Results

Variable	Coefficient	P-value	Multiplicative Effect ($e^{(\beta)}$)
Constant	1.56	0.00	4.76
Fuel_Type_Electric	1.49	0.00	4.42
Premium_Brand	0.50	0.00	1.65
Fuel_Type_Diesel	0.35	0.00	1.42
Power	0.01	0.00	1.01
Years_Old	-0.12	0.00	0.89
Transmission_Manual	-0.17	0.00	0.84

Significant Positive Predictors:

- **Used electric vehicles are priced 442% higher** than vehicles powered by the reference category, compressed natural gas (CNG)
- **Premium brands command 65% higher prices** than non-premium brands
- **Diesel vehicles are priced 42% higher** than CNG-fueled vehicles
- **High-power vehicles increase sales prices by 1%** for each additional horsepower

Significant Negative Predictors:

- **Older vehicles decrease in value by 11% for each additional year they age**
- **Manual transmission vehicles are priced about 16% lower** than their automatic counterparts

These findings will inform targeted business strategy recommendations designed to capitalize on key market drivers and optimize pricing, inventory, and marketing decisions.

Recommendations for Business Strategy

Analyzing vehicle price predictors provided the necessary information to recommend several actionable business strategies to optimize pricing, inventory management, and marketing efforts.

Pricing and Marketing Strategies

Dynamic pricing empowers businesses to adjust real-time prices to optimize revenue and improve organizational competitiveness. Specific recommendations are contained below:

- **Electric Vehicles:**

    - **Apply a high surcharge to electric vehicles** as they command the highest proportionate premiums
    - **Apply an additional surcharge** for electric vehicle sales in markets with increased consumer demand due to availability, fuel prices, and other relevant factors
    - Market these vehicles by **emphasizing economy, low cost of ownership, and available tax incentives**

- **Diesel Vehicles:**

    - **Add a marked price surcharge to diesel vehicles**
    - **Highlight fuel efficiency, reliability, and resale value** when marketing

- **Premium Brands:**

    - **Charge premiums for high-value car brands**
    - **Display luxury brand vehicles prominently** to attract customers to visit dealerships, even if they only shop for non-premium vehicles

- **Manual Transmission Vehicles:**

    - **Discount prices on manual transmission vehicles** to entice buyers and eventually reduce inventory
    - **Emphasize the advantages of manual transmissions**, such as performance, drivability, and fuel economy, to entice reticent buyers

- **High-Power Vehicles**
    - **Apply price premiums to high-power vehicles**
    - **Advertise features related to power**, such as acceleration and top speed, to target luxury buyers and performance enthusiasts
    - **Stock a limited volume of performance-oriented vehicles as halo models** to attract customers to the showrooms

- **Older-Aged Vehicles**
    - **The sale of older vehicles will require the provision of incentives** such as extended warranties or customer service benefits to attract buyers
    - **Promote older vehicles as economically accessible alternatives** to new vehicles
    - **Offer pre-sale inspections** to address customers' reliability concerns

- **High-Mileage Vehicles:**
    - **High-mileage vehicles command lower prices** due to customer preferences
    - These vehicles **can be made more attractive to buyers by highlighting maintenance records and inspection reports or offering extended warranties**

In implementing these strategies, businesses can more effectively capture market value, respond to shifting demand, and maximize profitability across different vehicle segments.

Inventory Management Strategies

Establishing clear acquisition priorities and effective clearance strategies is essential for optimizing inventory and aligning with market demand.

- **Acquisition Priorities:**

    - **Target newer car acquisitions** by offering favorable trade-in values for late-model used vehicles
    - **Prioritize the availability of electric- and diesel-powered vehicles**

    - **Stock vehicles from premium brands**

- **<u>Clearance Strategies:</u>**

    - **Discount manual transmission and high-mileage cars**
    - **Focus on value-added services**, such as extended warranty or service plans
    - **Divert financial incentives, such as low-interest finance rates, from high-value cars to lower-value cars** to entice buyers

By strategically acquiring high-demand vehicles and implementing targeted clearance initiatives, dealerships can maintain a balanced inventory, thereby accelerating vehicle turnover and improving overall profitability.

Limitations and Recommendations for Further Analysis

This analysis was limited by several factors, including:

- The data was collected in 2019. The ensuing COVID pandemic resulted in significant market variability, and thus, **pre-COVID data may not be indicative of present drivers of used car prices**
- **The datasets contained a significant proportion of missing values** in the target variable, Price
- **The observed values of the target variable were unevenly distributed,** requiring a log transformation to normalize values
- **The variable Kilometers_Driven also required a log transformation t**o achieve a normal distribution
- **Several independent variables had missing values, requiring imputation**

- **This dataset did not contain information on key features, including:**
  - **Buyer demographics**, including age, gender, profession, and socioeconomic indicators, would enable more nuanced analyses of market segmentation
  - **Vehicle condition** can highly influence price, particularly for vehicles in good cosmetic, physical, and mechanical condition
  - **Vehicle type,** such as family cars, sports utility vehicles, commercial trucks, and performance vehicles, for better market segmentation
  - **Car features,** including standard and optimal equipment
  - **Time markers** to control for seasonality and evolving consumer preferences
  - **Seller-specific information**, such as dealership sales, second-party sales, and individual sales, would yield more accurate pricing predictions and provide business insights
  - **Accident history** is known to have a substantial effect on the price of used vehicles
  - **Service records** typically enhance the sales price of used vehicles by serving as a proxy marker for reliability and cost of ownership
  - **Dealership networks** can be important, as cars sold through dealers generally sell for higher prices due to buyer perceptions of vehicle quality
  - **Fuel prices** that influence value, particularly for electric and diesel-powered vehicles
  - **Economic indicators** that would affect customers' buying power and, thus, demand
  - **Data for post-COVID years** is needed to increase model accuracy at present
  - **Remaining warranty coverage**, if applicable

Collecting and examining these data points is advised to increase the accuracy of pricing predictions in future analyses.

Conclusion

India's used car market is transforming rapidly, driven by rising demand and evolving consumer preferences. Thus, this market creates significant opportunities and complex pricing challenges for industry stakeholders. To address these challenges, this analysis developed a robust, data-driven pricing model using an OLS regression to identify and quantify the most significant predictors of used car prices, including fuel type, brand prestige, vehicle age, and transmission type.

The resulting insights provide actionable recommendations for optimizing pricing, inventory management, and business strategy. By embracing dynamic pricing and evidence-based decision-making, dealerships and industry stakeholders can maximize profitability and maintain competitiveness in an expeditiously expanding market. While this analysis is limited by the scope of available data, ongoing data collection and model refinement will be essential in sustaining long-term success in India's evolving used vehicle marketplace.

Citations

Business Standard. (2024, January 23). *Used car market overtakes new car sales in India: CARS24's report.* https://www.business-standard.com/industry/auto/used-car-market-overtakes-new-car-sales-in-india-cars24-s-report-125012301307_1.html

CARS24. (2025, January 23). *The 2024 Indian used car market report.* https://cdn.cars24.com/prod/auto-news24-cms/CARS24-Blog-Images/2025/01/23/c09aec15-acf9-4bc8-9d1f-ec5e3d8241f3-C24-gears-of-growth.pdf

Grand View Research. (2022, December 16). *India used car market size & outlook, 2024–2030*. https://www.grandviewresearch.com/horizon/outlook/used-car-market/india

Hindustan Times Auto. (2024, January 23). *The Indian used car market is expected to touch 82 lakh units by FY25, with mid-variants dominating*. Hindustan Times Auto. https://auto.hindustantimes.com/auto/cars/indian-used-car-market-to-touch-82-lakh-units-by-fy25-mid-variants-to-dominate-41623850478810.html

Little, A. D. (2024). *Powering up India's used car market*. https://www.adlittle.com/sites/default/files/viewpoints/ADL_Powering_up_Indias_used_car_market_2024.pdf

Malhotra, S. (2025, April 17). *A comprehensive year-end analysis of India's pre-owned car market* [LinkedIn post]. LinkedIn. https://www.linkedin.com/pulse/comprehensive-year-end-analysis-indias-pre-owned-sameer-malhotra-kuqpc

ReportLinker. (2025, February 19). *India used car market report: Q4 2024*. https://www.reportlinker.com/dlp/2dc150176d8d2408293872fafa2a5734

Society of Indian Automobile Manufacturers. (2024, April 15). SIAM releases motor vehicle data —March 2024, and annual data of 2023–24. https://evstory.in/siam-releases-motor-vehicle-data-march-2024-annual-data-of-2023-24/

Appendices

1. **Annotated Data Dictionary**

2. **Exploratory Data Analysis**

    A. **Univariate Analyses**
    B. **Bivariate Analysis**

3. **Final Regression Model**

4. **Comparative Regression Model Performance**

5. **OLS Regression Results**

6. **Testing Testing Regression Model Assumptions**

Annotated Data Dictionary

Annotated data dictionary

This table outlines findings from exploratory data analysis, the handling of missing variables, the transformation methods, the variables' relevance to the model, and the reasoning behind their inclusion or omission.

Variable	Definition	Transformations	Notes
Price (Target Variable)	The price of the used car in I₹ 100,000		- Due to the skewed distribution, it was necessary to take a log transformation for use as a target variable - 1234 missing values - All rows with missing values for the target variable must be dropped
Price_Log	The log value of used car prices in ₹ 100,000	Log-transformation	- The log-transformed version of Price. Still a significant number of missing values, requiring the dropping of observations missing information on this target value
S.No	Vehicle Serial Number		- No missing values - This feature was removed as it did not add useful information to the model

Variable	Definition	Transformations	Notes
Name	Name of the car which includes Brand name and Model name	Broken into component brands and model names	- No missing observations - Too many values - First broken into Model and Brand - Later this variable was further consolidated into the variable Premium Brands
Premium_Brand	A dichotomous variable, Premium_Brand includes Audi, Bentley, BMW, Jaguar, Lamborghini, Land Rover, Mercedes-Benz, Mini, Porsche, and Volvo. All other makes are considered standard brands.	Transformed Name column into a single variable for premium versus standard brands	- This variable is easier to interpret and less computationally expensive than the use of all 32 available brands included in the dataset
Location	The location in which the car is being sold or is available for purchase (Cities)		- No missing values - Consolidated into Urban/Suburban variable for ease of interpretation
Urban	A dichotomous variable. Urban locations include: Bangalore, Chennai, Delhi, Hyderabad, Kolkata, Mumbai, and Pune. All other localities are coded as non-urban.	Transformed Location values into single variable indicating urban versus non-urban locations	- Consolidatad for ease of calculation and interpretability
Kilometers_Driven	The total kilometers driven in the car by the previous owner(s) in KM		- No missing values - This variable was heavily skewed, requiring a log transformation for use in regression models
Kilometers_Driven_Log	Log value of total kilometers driven by previous owners	Log transformation of kilometers driven	

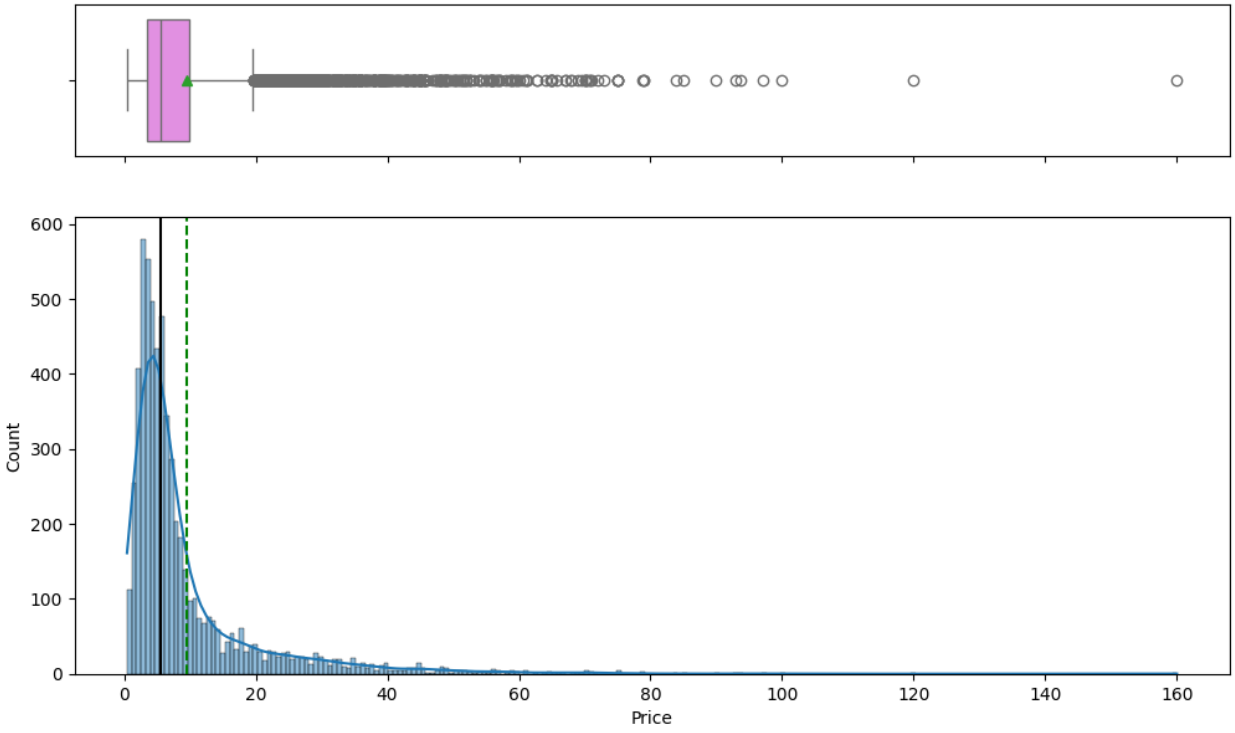
Variable	Definition	Transformations	Notes
Fuel_Type	The type of fuel used by the car (Petrol, Diesel, Electric, CNG, LPG)		- No missing values - One-hot encoded to create dummy variables for regression models
Transmission	The type of transmission used by the car (Automatic / Manual)		- No missing values - One-hot encoded to create dummy variables for regression models
Owner_Type	Number of previous vehicle owners		- No missing values - Dropped due to collinearity with vehicle age
Mileage	The standard mileage offered by the car company in kmpl or km/kg		- Dropped due to collinearity with power and engine displacement - Normally distributed - 2 missing values imputed with mean values.
Engine	The displacement volume of the engine in CC		- Skewed, requiring transformation - 46 Missing values imputed with median values - Dropped due to collinearity with power and mileage
Engine_Log	The log value of engine displacement in CC	Log transformation of Engine values	- Dropped from the final model due to collinearity with power and mileage
Power	The maximum power of the engine in bhp		- Right-skewered - 175 missing values imputed with median values - Variable retained as it had the best predictive power of the related variables, power, engine, and mileage

Variable	Definition	Transformations	Notes
Seats	The number of seats in the car		- Roughly even distribution with relatively little variance - 53 missing values were imputed with mean values - The variable was dropped due to collinearity issues with Fuel_type, and Power
New_Price	The price of a new car of the same model in ₹ 100,000		- Heavily right-skewed - 6247 missing values - Variable dropped due to this substantial volume of missing values
Year	The year in which the vehicle was manufactured		- No missing values - Categorical values must be changed to numeric values
Years_Old	The age of the vehicle in years	Calculated by subtracting the value for the variable 'Year' from the index year (2019)	- Skewed left

Exploratory Data Analysis

Univariate Analysis: Numeric Variables

Distribution of Price Variable

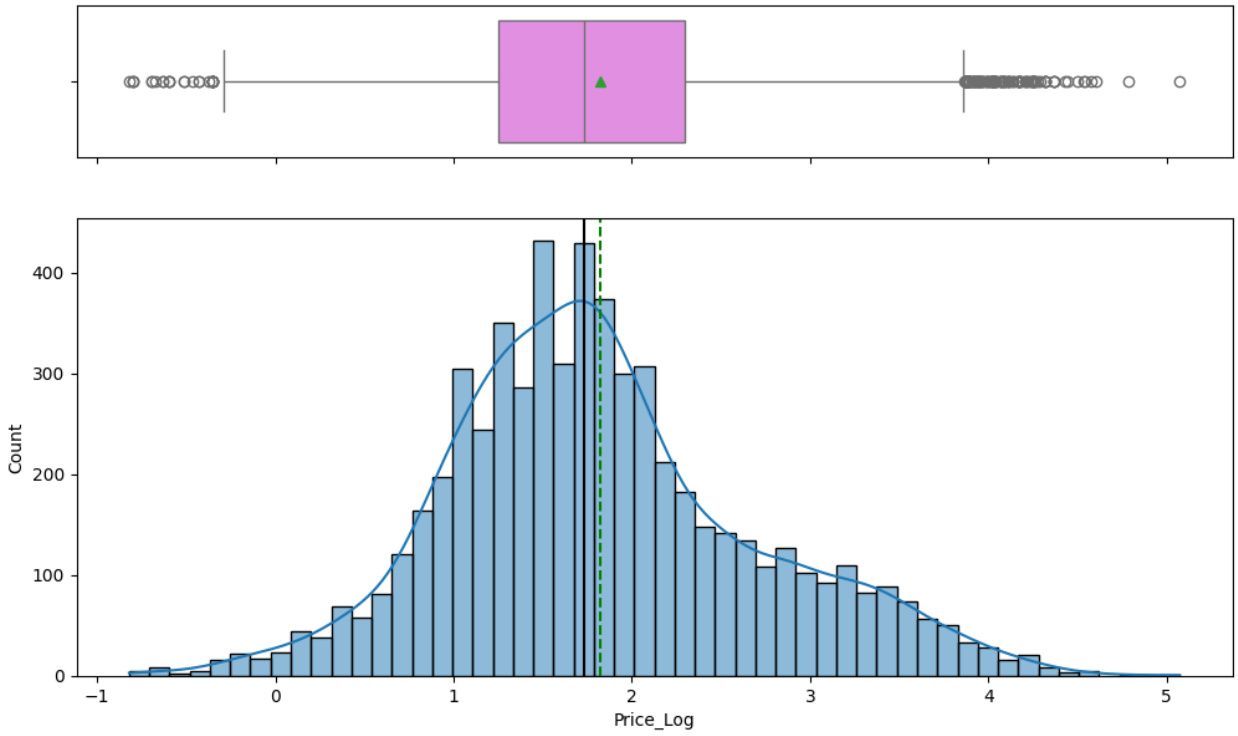


Notes:

- The observations for the variable "Price" are highly right-skewed, with many missing observations
- As Price is an outcome variable, it must be more evenly distributed through a log transformation
- Also, given that this is the model's outcome variable, all rows with missing values for Price must be dropped before undertaking a regression analysis.

Univariate Analysis: Numeric Variables

Distribution of Log_Price Variable

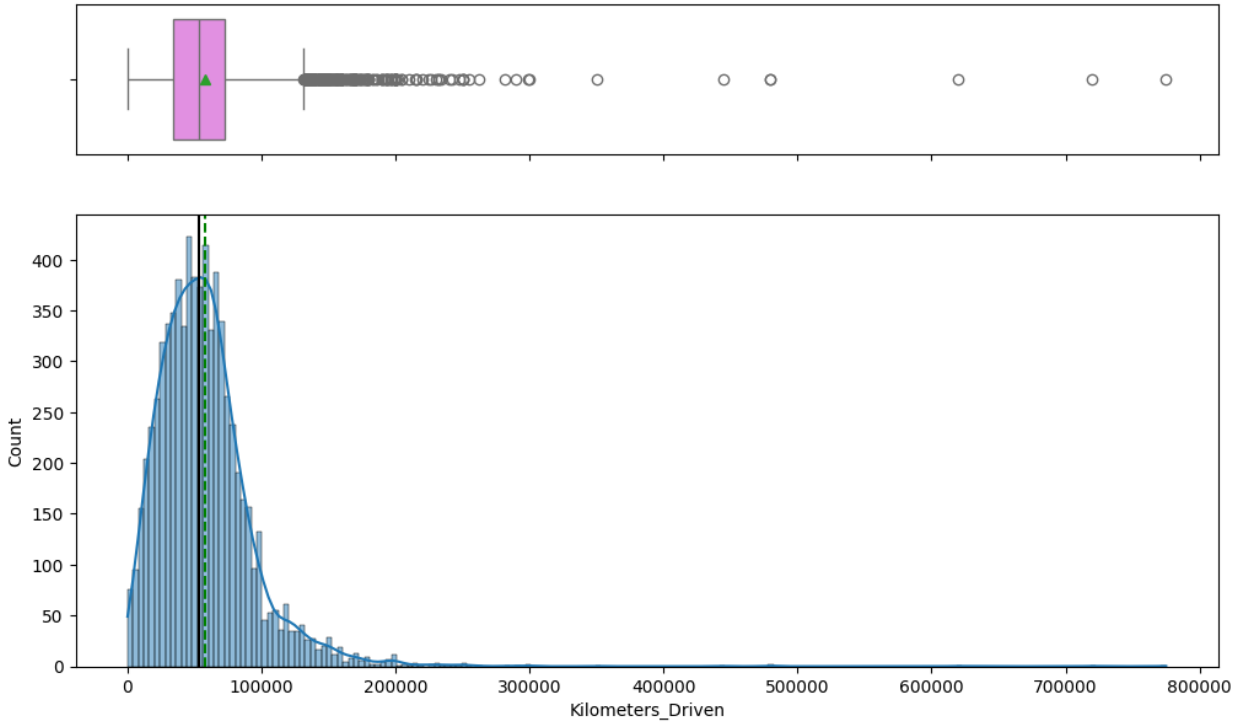


Notes:

- The log-transformed variable Price_Log is much more normally distributed and may now be used as the model's dependent variable.

Univariate Analysis: Numeric Variables

Distribution of Kilometers_Driven Variable

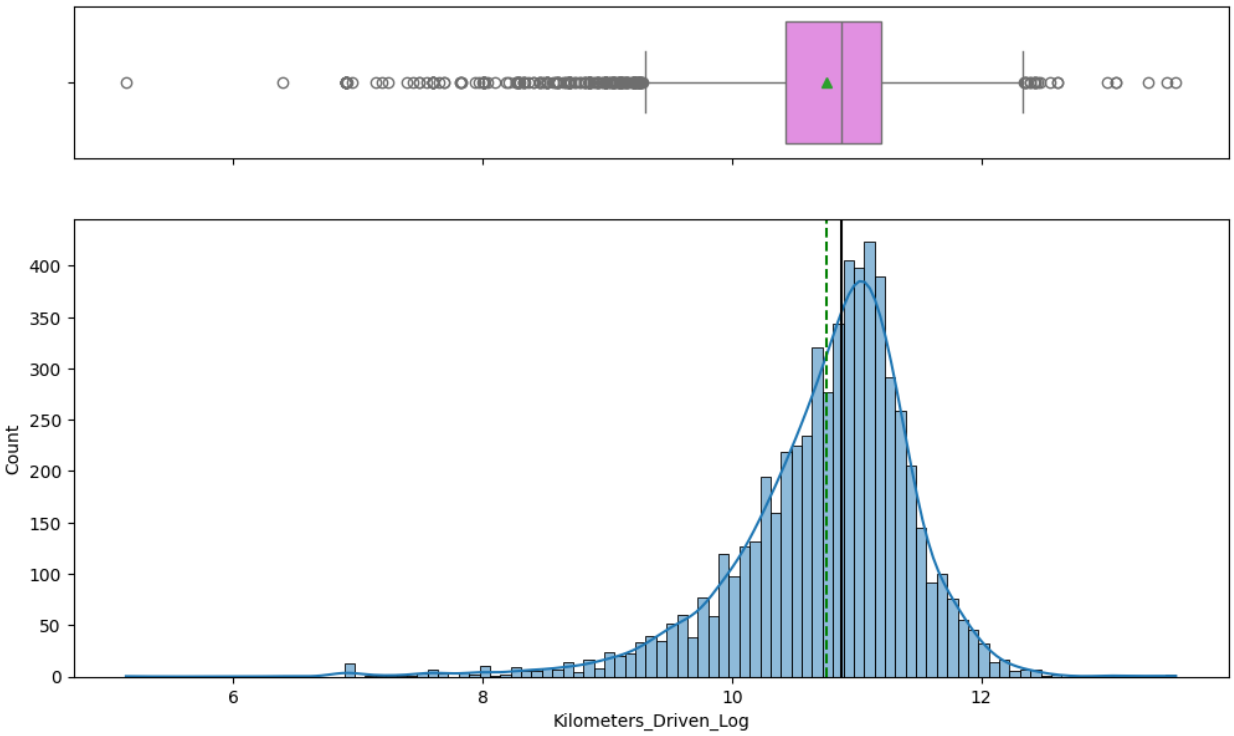


Notes:

- The values for Kilometers_Driven are heavily skewed to the right
- This is likely due to outliers, which might warrant removal
- The variable must be log-transformed to achieve a more normalized distribution

Univariate Analysis: Numeric Variables

Distribution of Kilometers_Driven_Log Variable

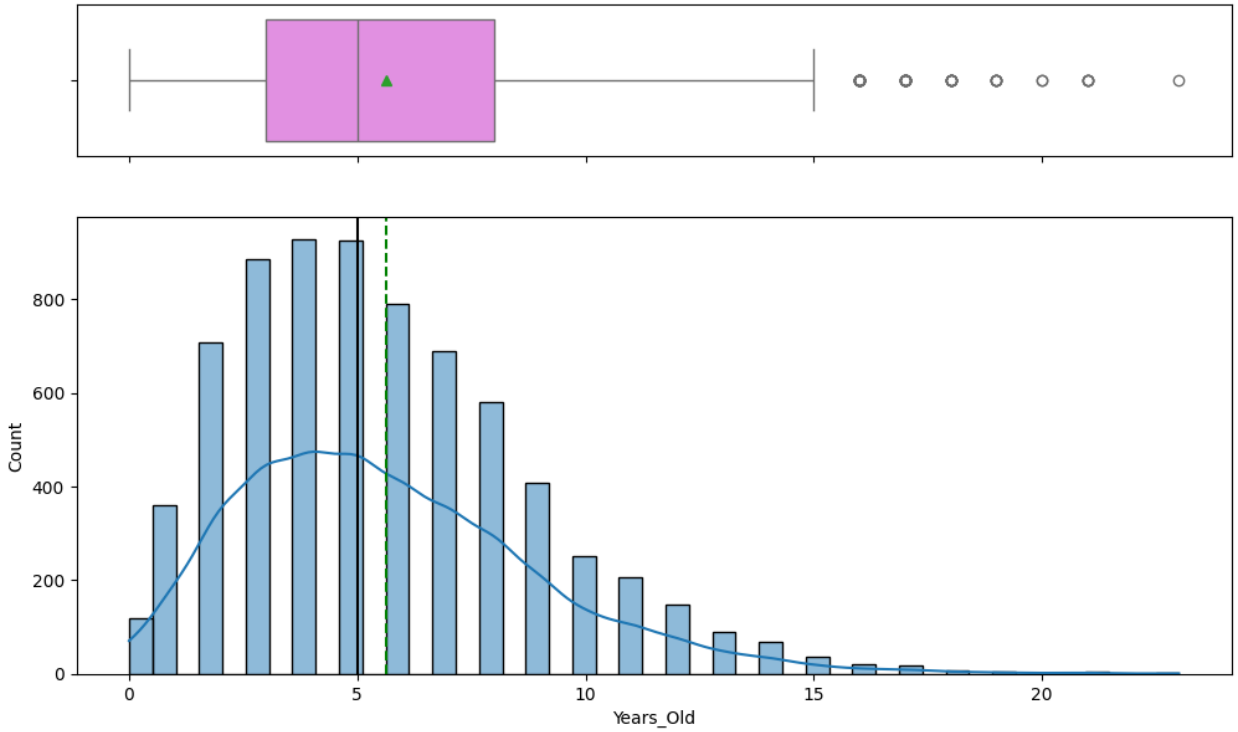


Notes:

- The log-transformed variable is more normally distributed, though it remains slightly left-skewed

Univariate Analysis: Numeric Variables

Distribution of Years_Old Variable

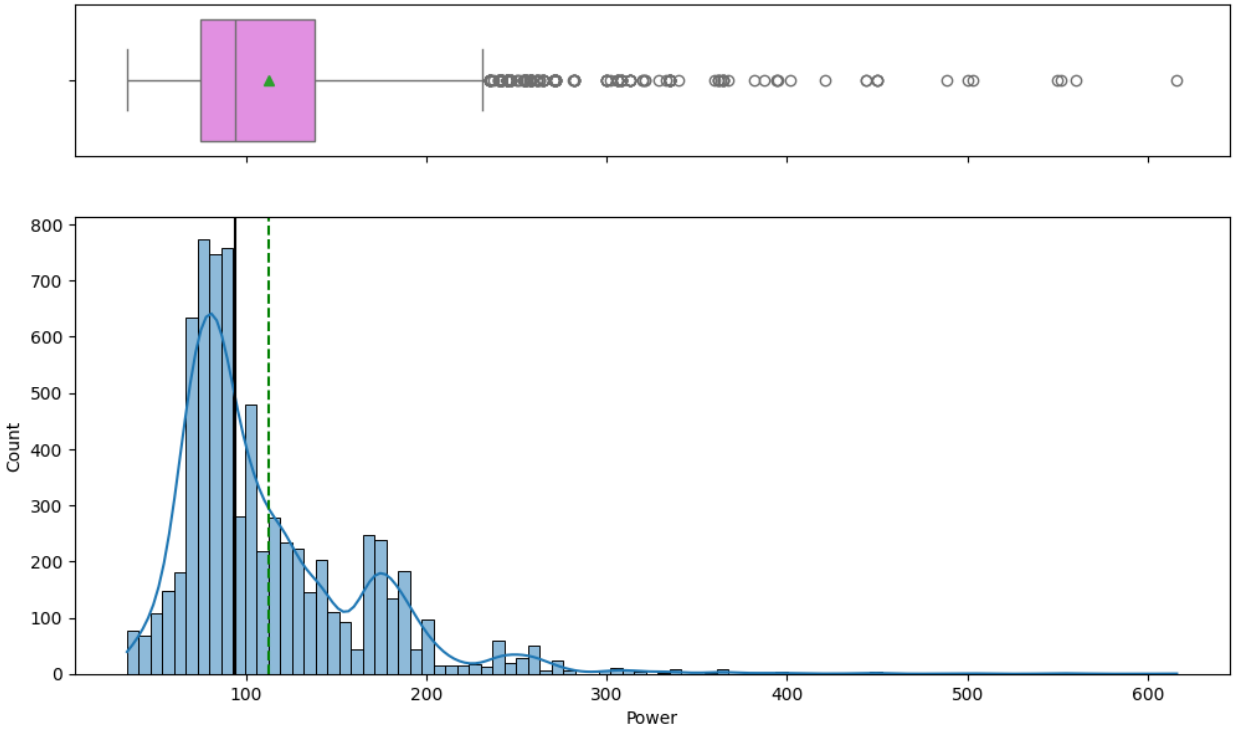


Notes:

- The variable is skewed
- However, the log-transformed variable was no more evenly distributed and was dropped from the model
- The original variable will be used with the knowledge that independent variables need not necessarily follow a perfectly normal distribution

Univariate Analysis: Numeric Variables

Distribution of Power Variable

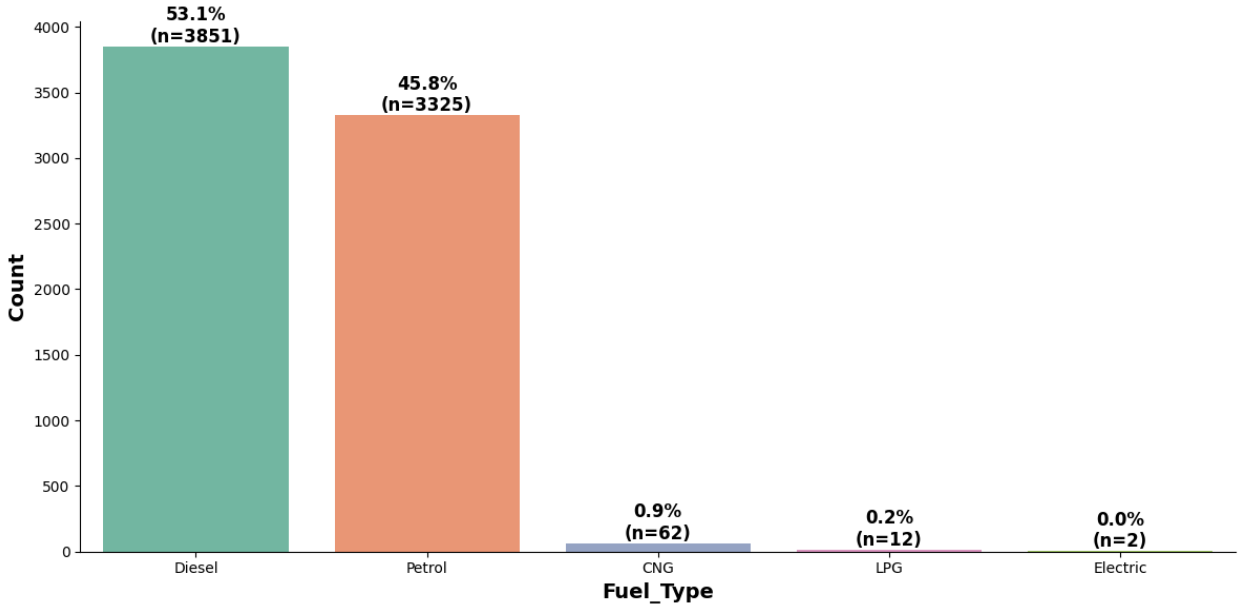


Notes:

- The variable is skewed due to outliers
- However, the log-transformed variable was no more evenly distributed and was dropped from the model

Univariate Analysis: Categorical Variables

Distribution of Fuel_Type Variable

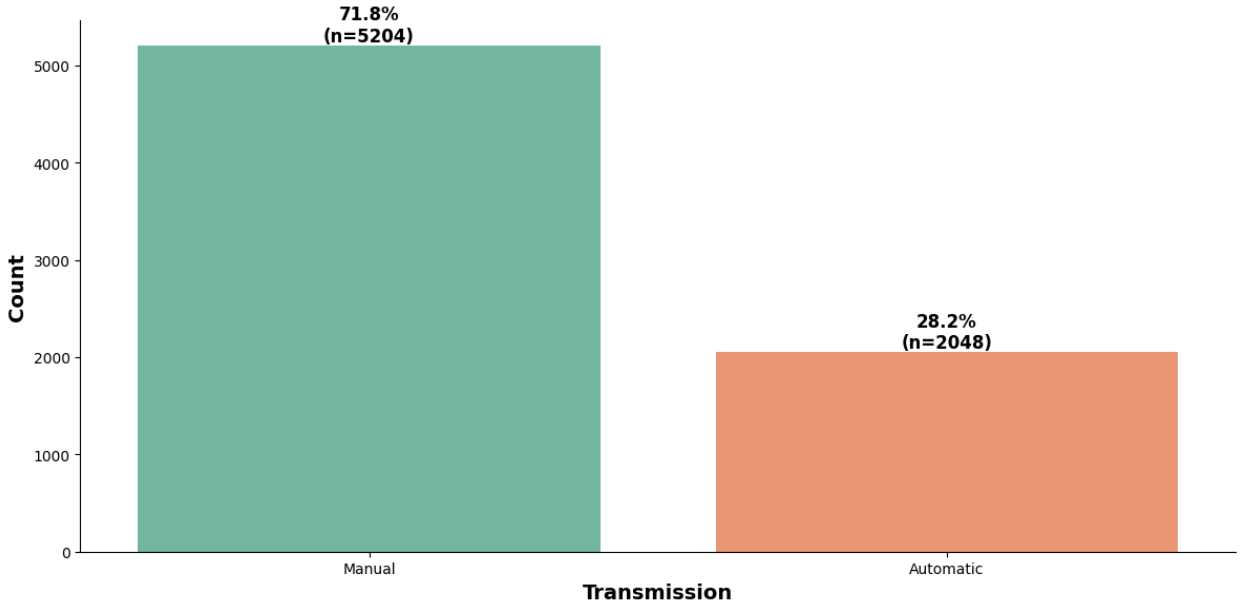


Notes:

- Diesel and petrol are by far the most common fuel types
- Dummy variables will be created for each category

Univariate Analysis: Categorical Variables

Distribution of Transmission Variable

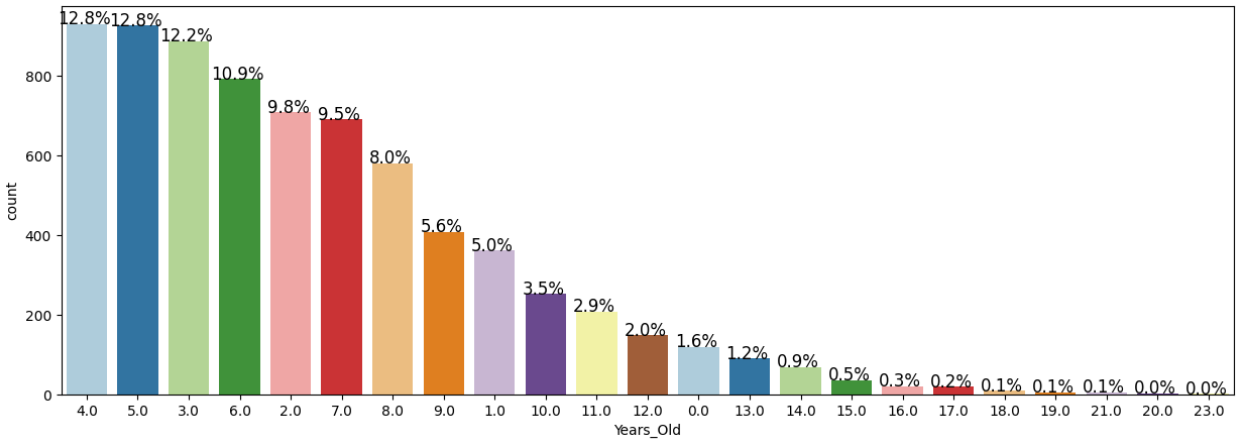


Notes:

- There are more cars with manual transmissions, relative to automatic transmissions
- Dummy variables will be created for each category

Univariate Analysis: Categorical Variables

Distribution of Years_Old Variable

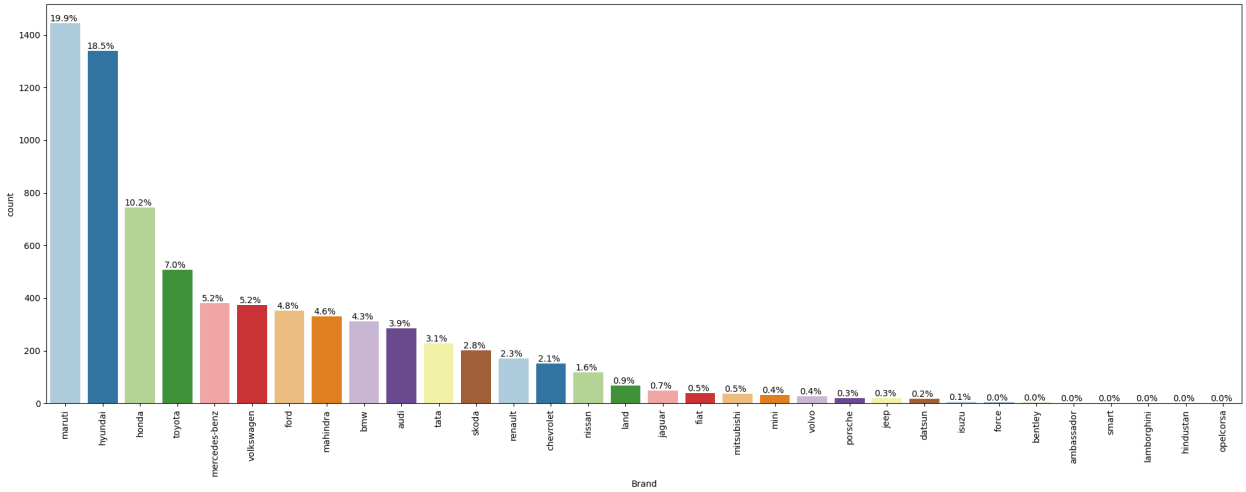


Notes:

- The distribution suggests that late-model used cars are most common, which is to be expected as vehicles wear with age

Univariate Analysis: Categorical Variables

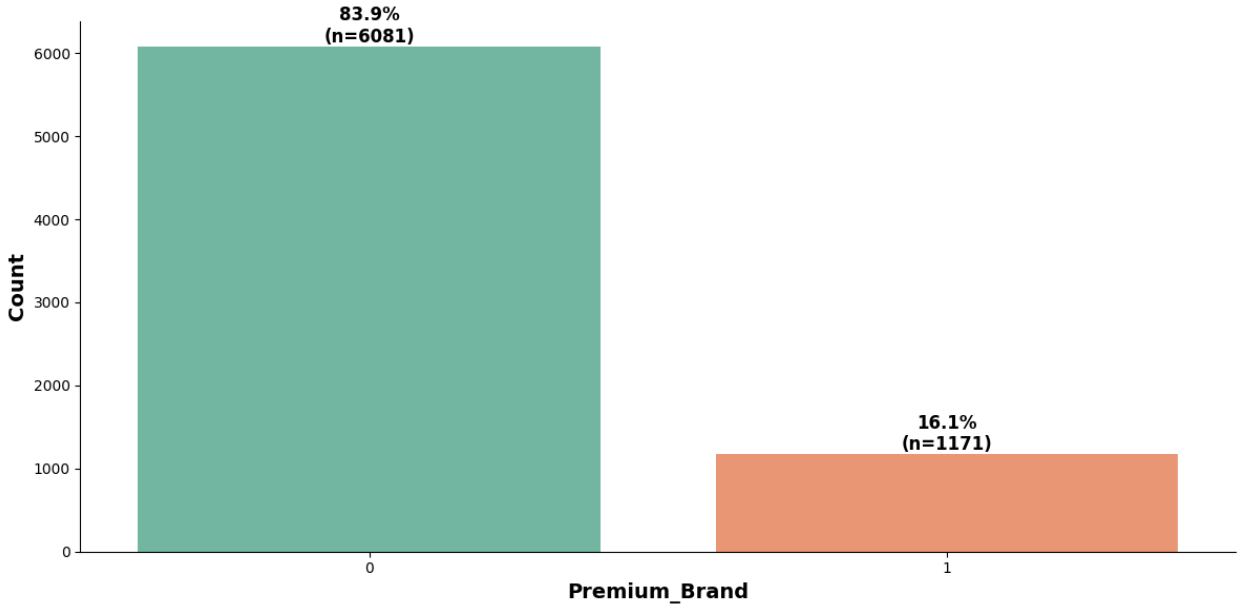
Distribution of Brand Variable



- Thirty-two different brands are represented in the dataset. This is far too many to be usable.
- They will be recoded into a binary variable for premium and non-premium brands for ease of analysis and interpretation

Univariate Analysis: Categorical Variables

Distribution of Premium_Brand Variable

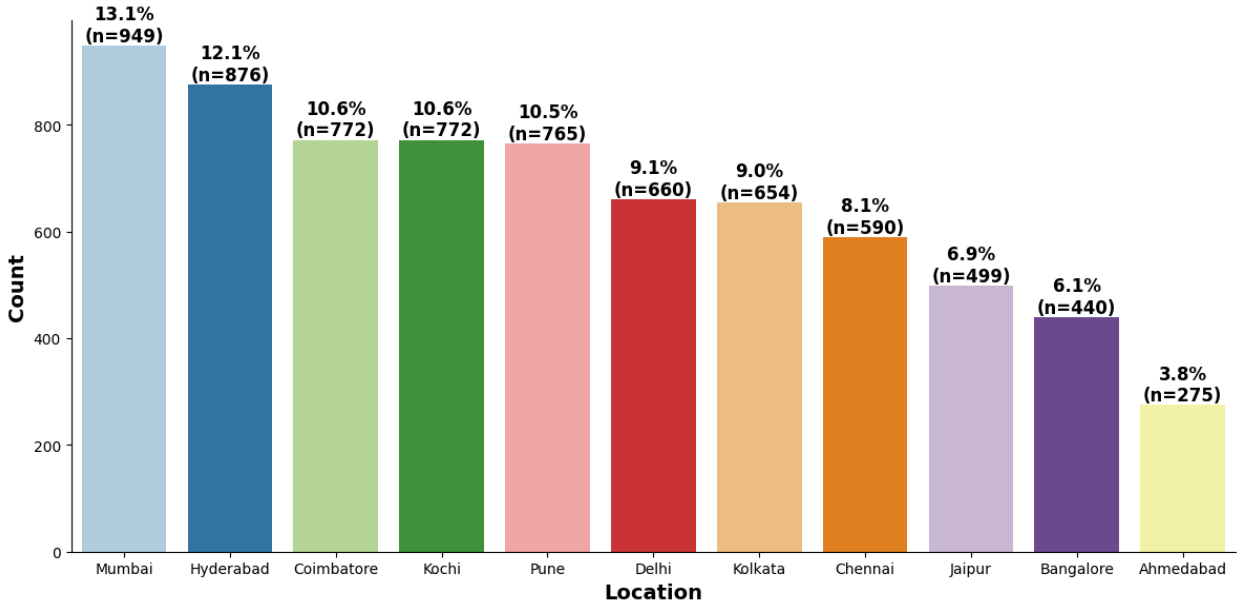


Notes:

- Relatively few of the used cars in this data set were from premium brands
- This variable is less cumbersome and more easily interpretable than the original variable, Brand, which contained 32 categories

Univariate Analysis: Categorical Variables

Distribution of Location Variable

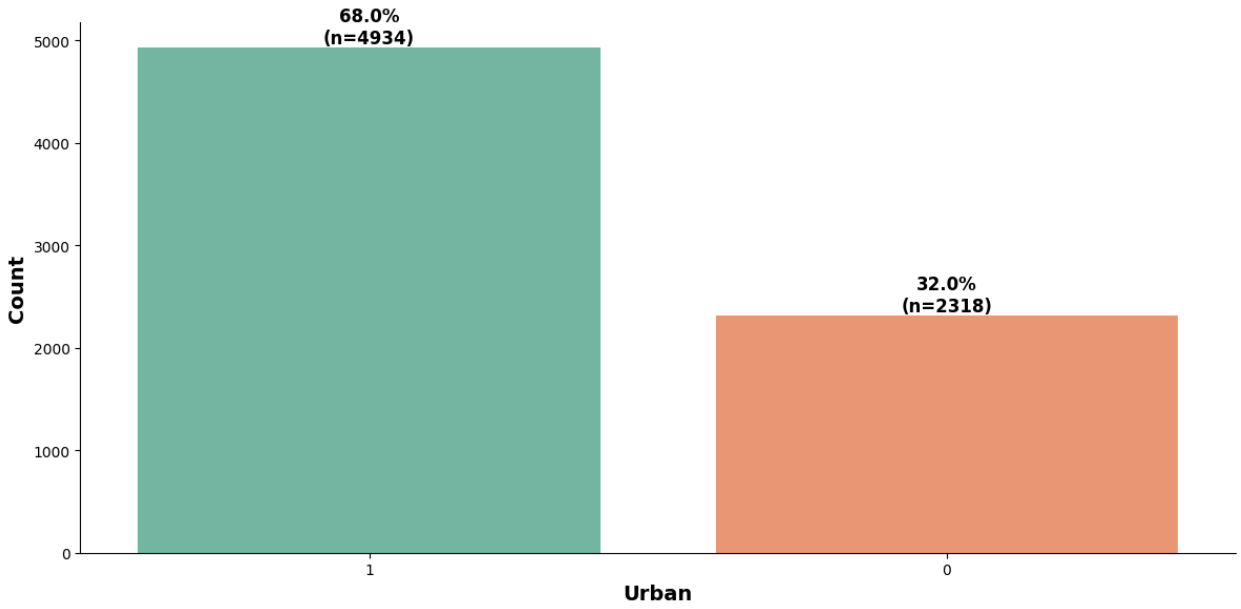


Notes:

- The dataset contains eleven different location categories
- This variable will be recoded into a dichotomous urban versus non-urban variable

Univariate Analysis: Categorical Variables

Distribution of Urban Variables

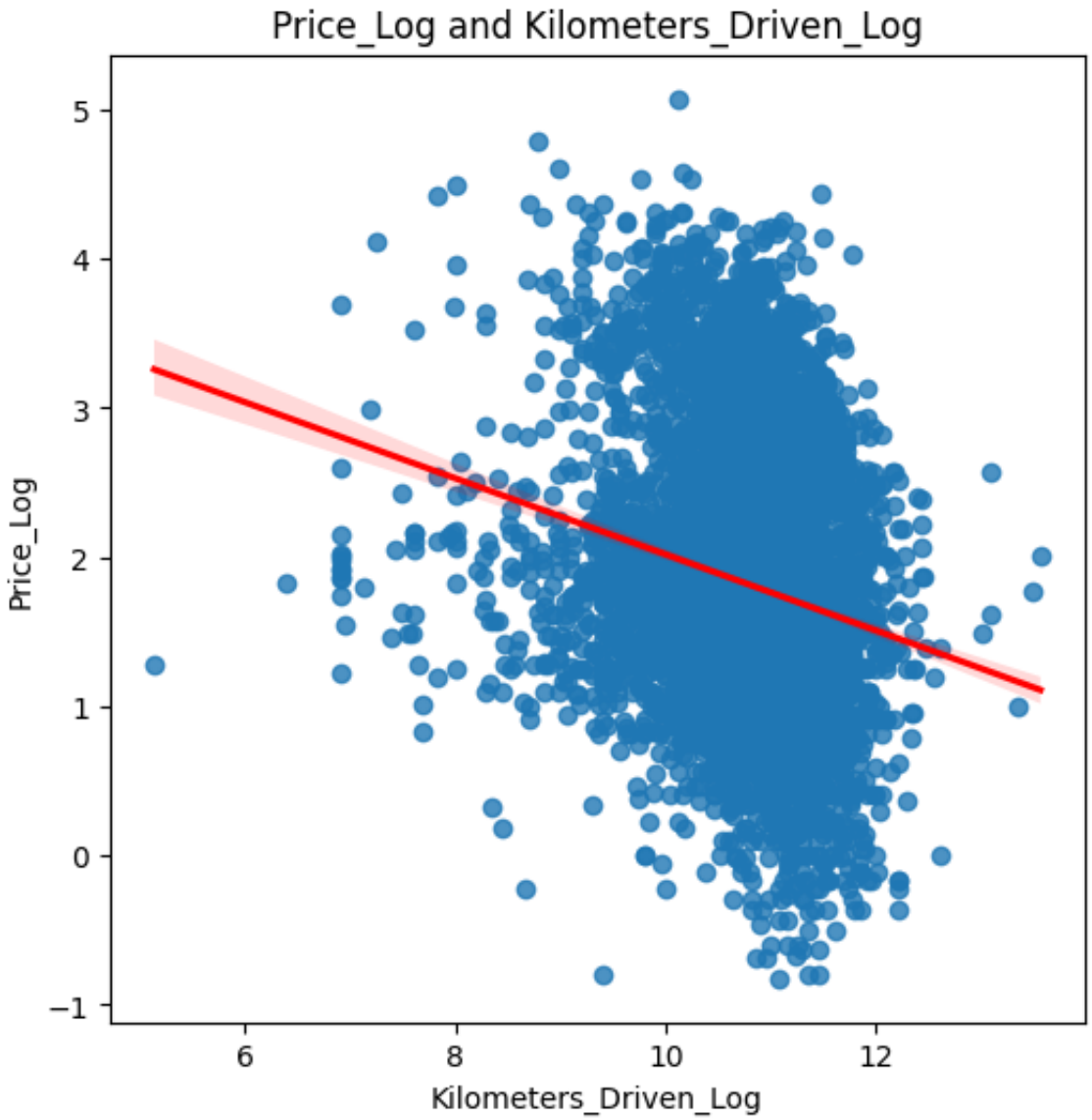


Notes:

- Most used cars were sold in urban areas
- This variable is created as it is more intuitive than modeling on a city-to-city basis
- The consolidation also preserves additional degrees of freedom within the model

Bivariate Analysis: Numerical Variables

Relationship Between Price_Log and Kilometers_Driven_log

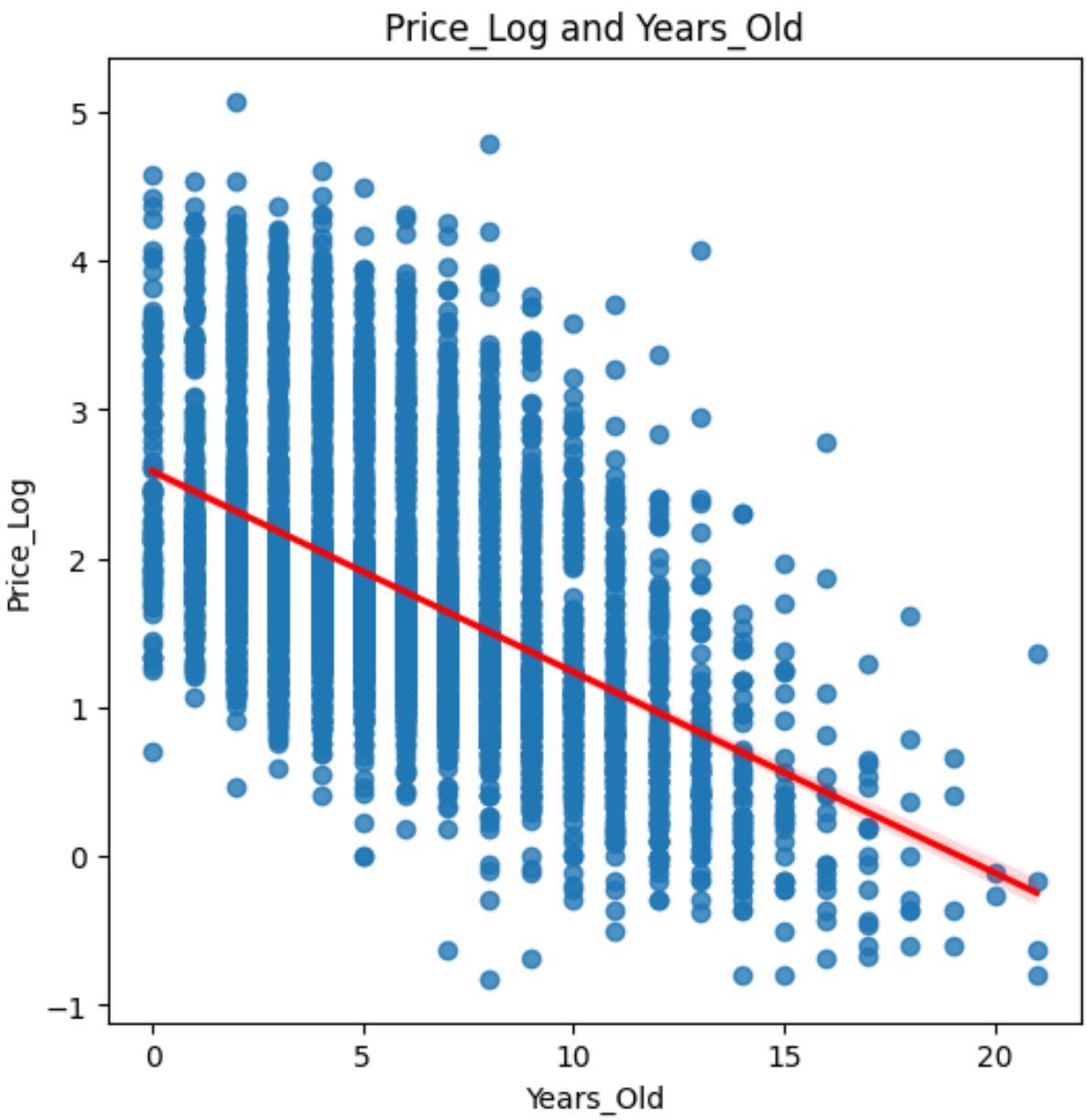


Notes:

- There is some clustering, but not a clearly defined linear relationship
- Kilometers_Driven_Log might not be the strongest predictor of Price_Log

Bivariate Analysis: Numerical Variables

Relationship Between Price_Log and Years Old

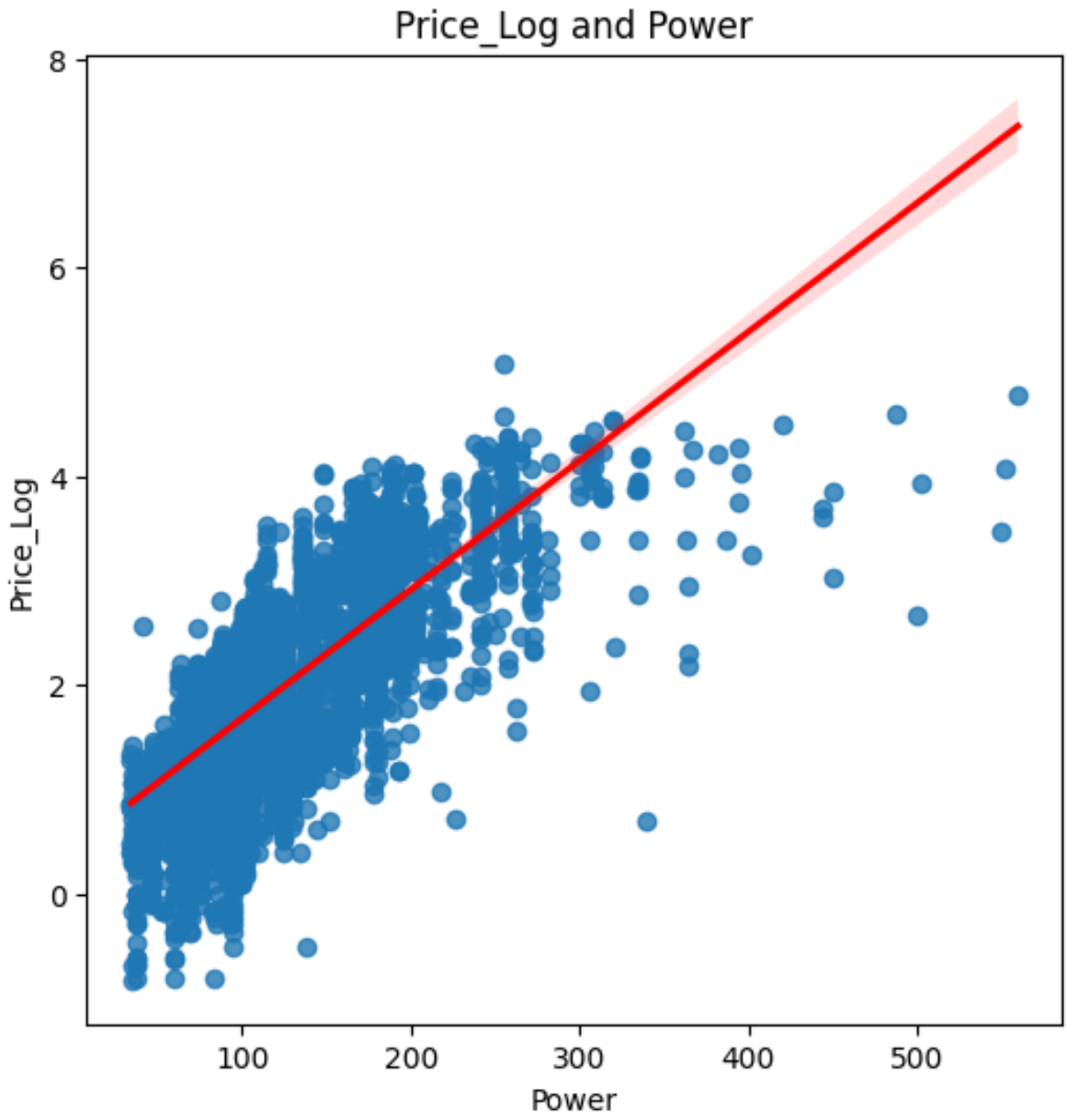


Notes

- There is a well-defined linear relationship visible in this plot

Bivariate Analysis: Numerical Variables

Relationship Between Price_Log and Power Variable

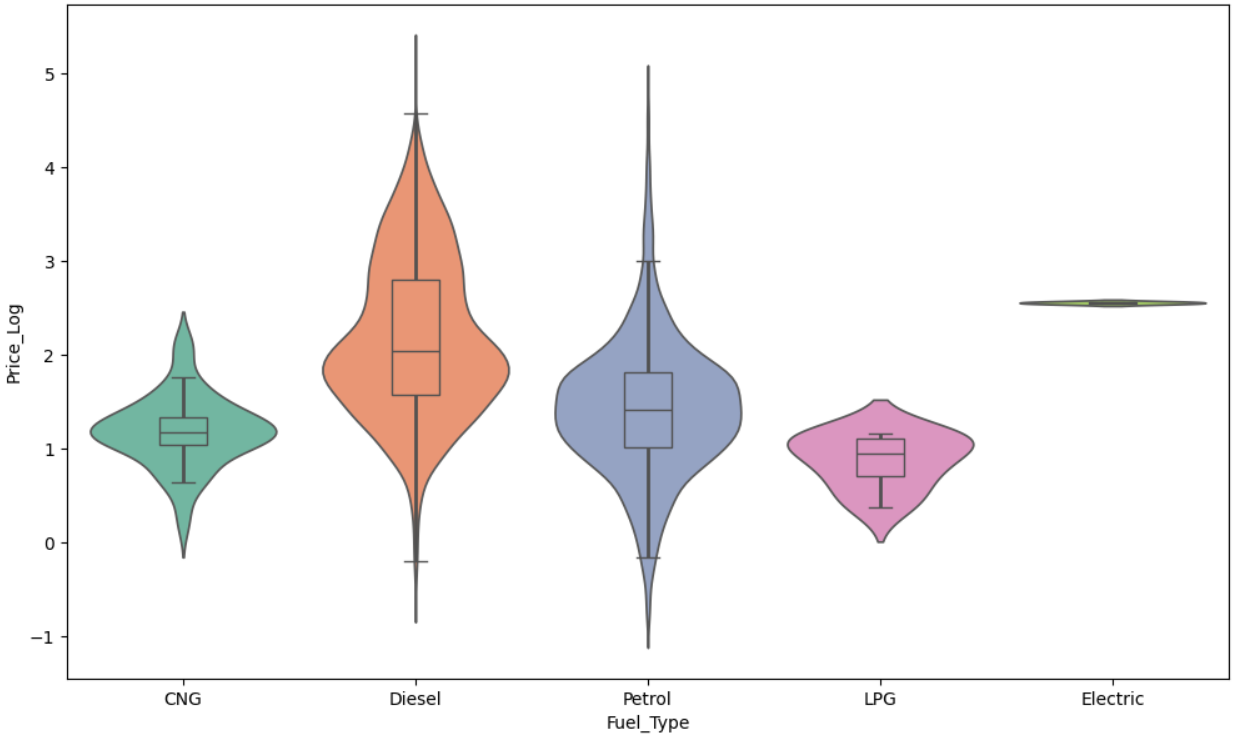


Notes

- There is a well-defined linear relationship visible in this plot

Bivariate Analysis: Categorical Variables

Relationship Between Price_Log and Fuel_Type

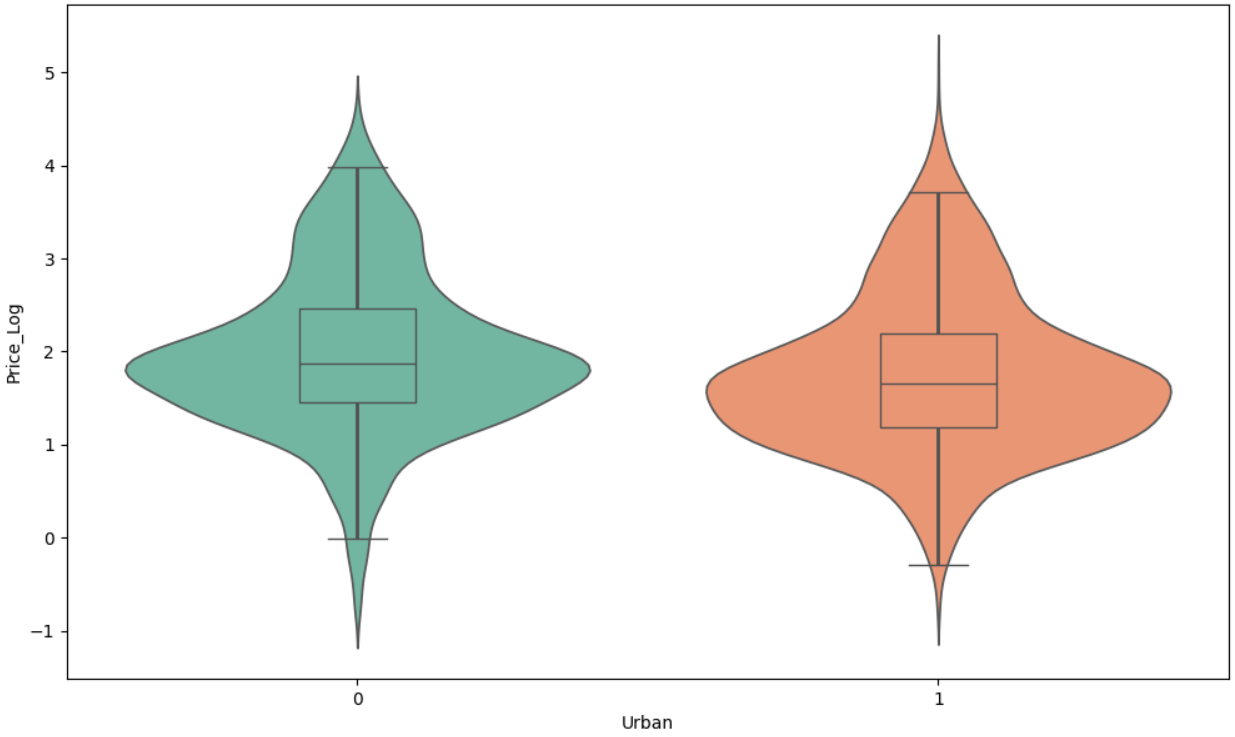


Notes

- There are significant differences in the interquartile range, with Electric showing slight variation, while Petrol and Diesel show more variation

Bivariate Analysis: Categorical Variables

Relationship Between Price_Log and Urban

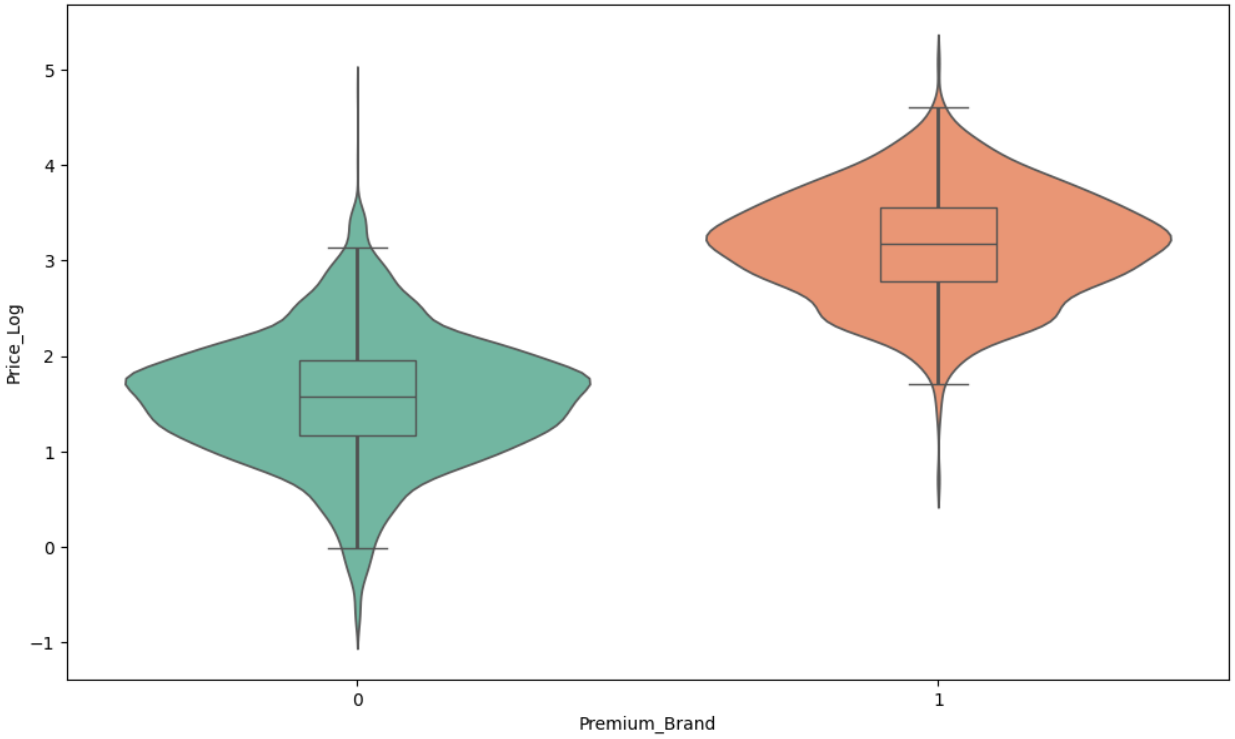


Notes

- Both categories appear to have a similar range of values
- The mean used car prices for Non-Urban settings are slightly higher than those for Urban settings

Bivariate Analysis: Categorical Variables

Relationship Between Urban and Premium_Brand



Notes

- This plot suggests premium-branded vehicles sell for comparatively higher mean prices than their non-premium counterparts

Testing Correlations Between Variables

Correlation Heat Map for Numerical Variables

Notes

- There appears to be little multicollinearity in the model aside from some overlap between power and Premium_Brand
- It will be necessary to calculate VIF values to check for collinearity

Final Regression Model

Final Regression Model

The following variables were selected for inclusion in the final version of the regression model:

Features included in the final OLS regression model
Dependent variable
Log_Price
Intercept
Constant
Independent variables
Years_Old
Kilometers_Driven_Log
Premium_Brand
Urban
Power
Fuel_Type_Diesel
Fuel_Type_Electric
Fuel_Type_LPG
Fuel_Type_Petrol
Transmission_Manual

Regression Results

Regression Results

The findings of the OLS regression model were as follows:

OLS Regression Analysis Results

Variable	Coefficient	P-value	Multiplicative Effect ($e^{(\beta)}$)
Constant	1.56	0.00	4.76
Fuel_Type_Electric	1.49	0.00	4.42
Premium_Brand	0.50	0.00	1.65
Fuel_Type_Diesel	0.35	0.00	1.42
Power	0.01	0.00	1.01
Years_Old	-0.12	0.00	0.89
Transmission_Manual	-0.17	0.00	0.84

Notes

- Fuel type is the most influential driver of used car prices in India
- Other statistically significant variables include Transmission_Manual, Premium_Brand, Years_Old, and Power
- The variable Urban was not a statistically significant predictor of used car price
- As the coefficients are log units, it is necessary to calculate their multiplicative effects via an exponential transformation

Checking OLS Regression Assumptions

Checking OLS Regression Assumptions

1) The means of Residuals Should Be Zero

Findings

- The value of the residuals from the fitted model is 6.986075746596664e-16

Notes

- The residuals are extremely close to 0, thus fulfilling this regression assumption

Checking OLS Regression Assumptions

2) Checking for Heteroskedasticity

Findings

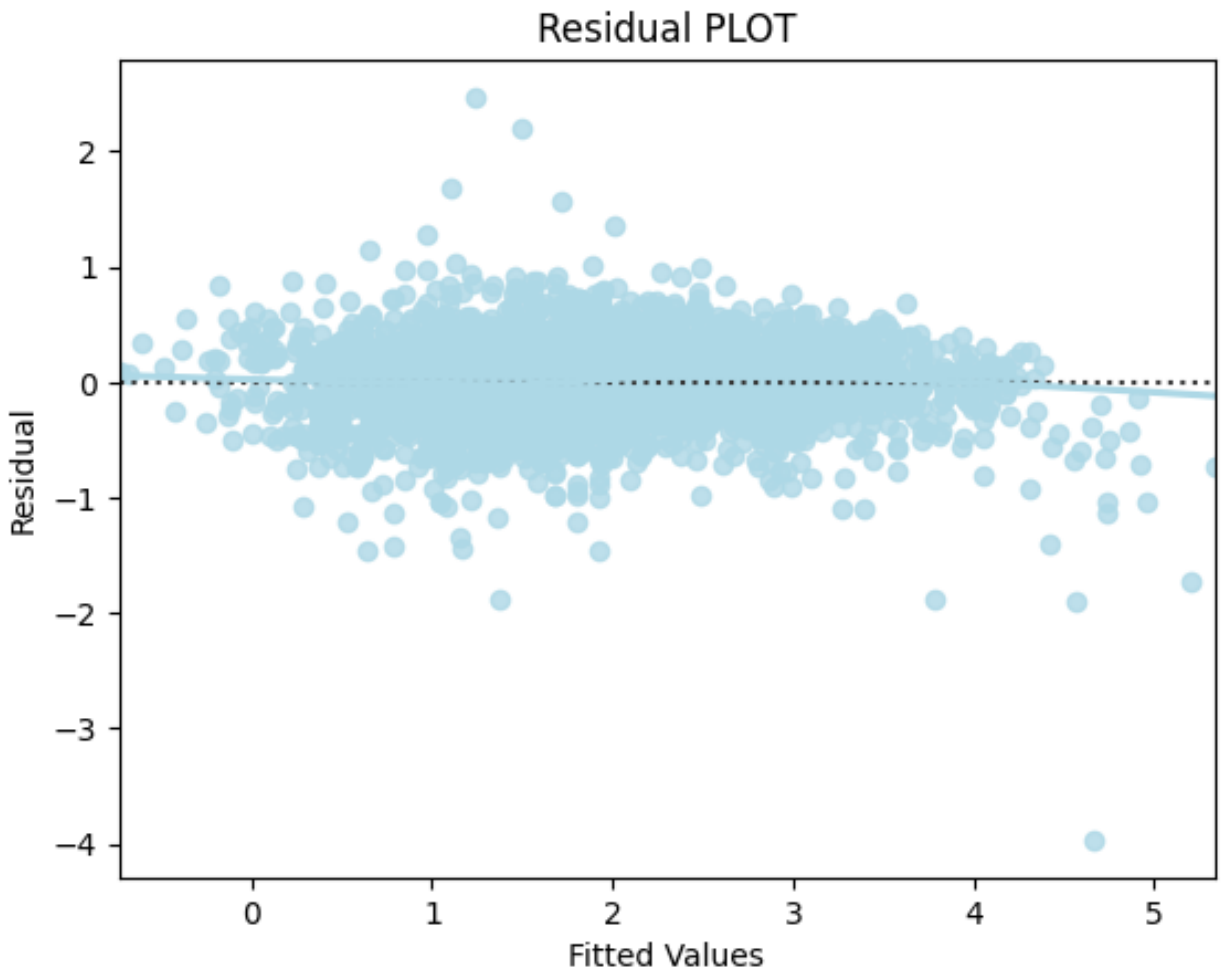
The Goldfeld-Quandt Test yielded an F-statistic value of 1.252 and a p-value of 1.287

Notes

- With a p-value > 0.05, we cannot reject the Null Hypothesis that the residuals are homoscedastic. Thus, the corresponding assumption of heteroskedasticity is satisfied

Checking OLS Regression Assumptions

3) Check the linearity of variables

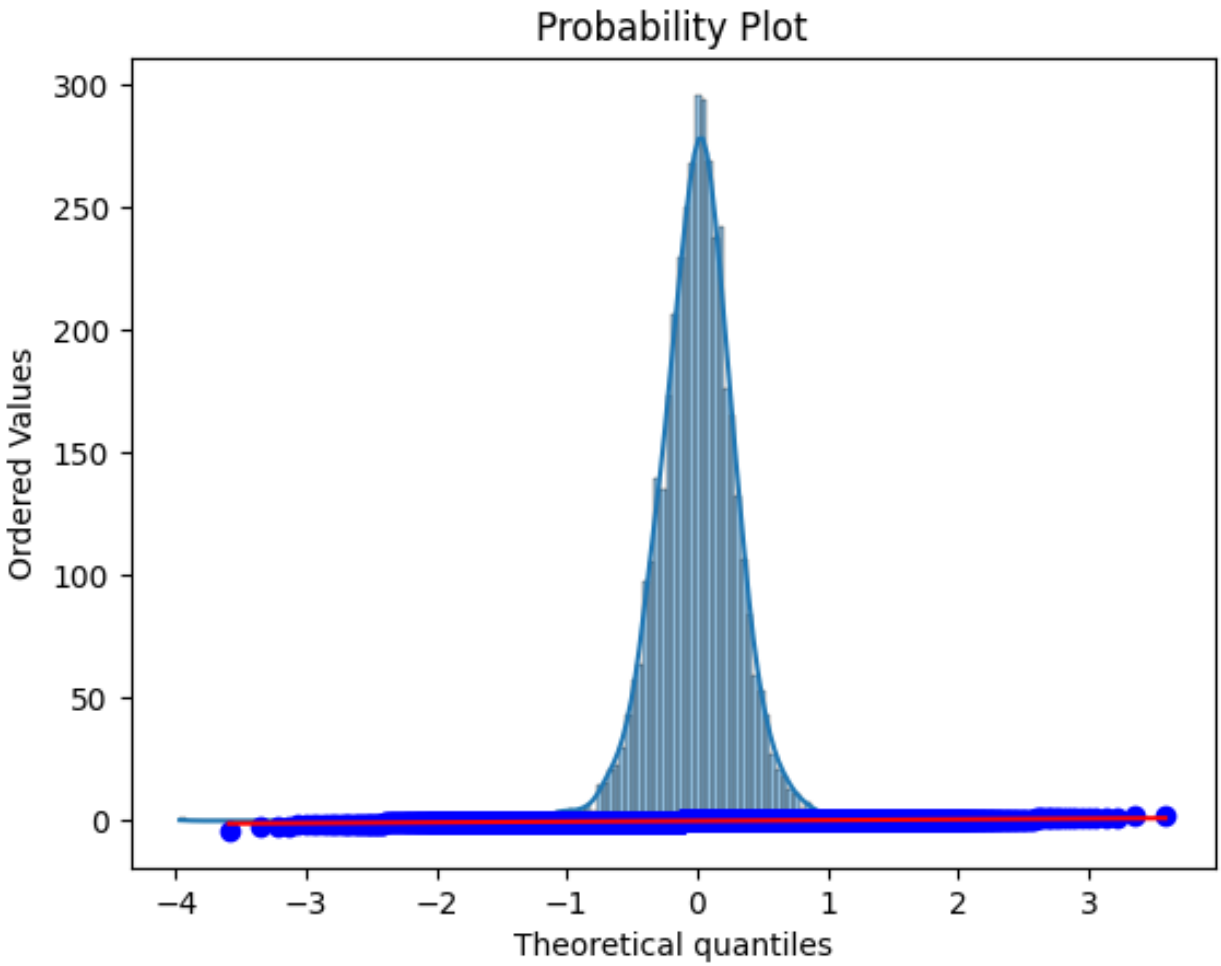


Notes

- Residuals are generally scattered around the zero line
- There is no pattern in the residual vs fitted values plot
- Accordingly, the assumption of linearity of variables is satisfied

Checking OLS Regression Assumptions

4) Check the Normality of Error Terms



Notes

- The residuals follow a normal distribution
- The assumption of normality in error terms is satisfied

Checking OLS Regression Assumptions

5) Check the Independence of Variables

Feature	VIF Values
Constant	459.688598
Years_Old	1.614821
Kilometers_Driven_Log	1.606450
Premium_Brand	2.541694
Urban	1.080377
Power	2.273854
Fuel_Type_Diesel	27.855352
Fuel_Type_Electric	1.028225
Fuel_Type_LPG	1.179391
Fuel_Type_Petrol	27.778880
Transmission_Manual	2.061626

Notes

- There is collinearity amongst dummy variables derived from the same source variable. This is to be expected due to the multicollinearity inherent in one-hot encoding
- The remaining values fall within the acceptable range, fulfilling the assumption of independence of variables