Report: Predicting Vehicle Prices

A robust predictive model was developed to estimate used car prices to help resellers optimize their pricing strategy and maximize profitability in the German automobile marketplace. The model is designed to ultimately support data-driven decision-making for used car resellers in Germany.

Context: The dataset for this task has been scraped from a German online car trading company, AutoScout. The AutoScout data is a comprehensive collection of information on various used vehicles, capturing a wide array of attributes.

Methodology: My approach involves rigorous data pre-processing, logging, correcting class imbalances, feature engineering, encoding techniques and relevant transformations. A baseline Linear Regression model was first developed, followed by regularization techniques such as Ridge Regression and Lasso Regression to mitigate multicollinearity, control overfitting and improve model generalization.

1. Data Preparation

The necessary libraries were imported and *gg-plot* style was used for my plots. A helper function was designed to help me log messages for debugging, informational, warning, error and critical messages during the runtime of my notebook.

1.1. Loading the dataset

The web-scraped dataset, sourced from a German online car trading company, AutoScout was loaded from a CSV file into a pandas DataFrame.

During data ingestion phase, proper exception handling was implemented to ensure a stable workflow, along with logging of successful loads, any potential file access or read errors.

Upon successful loading, a random sample of observations was examined along with statistical summary and metadata check to identify data quality and variable types. My initial inspection revealed that the dataset contained 15,915 observations and 23 Features.

1.2. Assumptions of Linear Regression Models (For Scope of Vehicle Prices)

Data Integrity: It is assumed that the AutoScout dataset is reasonably accurate.

Logarithmic Transformation: Predicting the log-transformed price is more effective than predicting actual raw price. This approach is assumed to stabilize variance, reduce right-skewness and improve model performance.

Train-Test Split: An 80:20 train-test split was assumed to be sufficient for model training and testing.

Low Frequency Categories: It is assumed that low frequency categories in the features improve model stability, performance and generalization, without distorting the underlying automobile market behavior or introducing any bias.

Feature Encoding: It is assumed that feature encoding techniques allow the model to effectively interpret categorical and multi-categorical features.

Scaling Techniques: It is assumed that the scaling techniques implemented do not introduce any significant bias into the regression models. It is also assumed that the relative importance of the predictors is preserved and all continuous numerical predictors are on the same scale (comparable scale).

Linearity: It is assumed that each predictor has a linear relationship with our target variable. We assume that a linear approximation is sufficient for vehicle price prediction.

Additive Effect: It is assumed that the effect of each predictor on price is additive in nature and not interactive. This means that one feature's contribution does not depend on the value of another.

Homoscedasticity & Normality of Errors: It is assumed that the residuals follow a normal distribution centered at zero with constant variance. Homoscedasticity of residuals is assumed to mean that residuals should appear evenly scattered, indicating uniform error spread.

Independence of Errors: It is assumed that the residuals do not correlate with each other across observations because correlated errors suggest that the model has missed the underlying pattern in our data.

Acceptable Multicollinearity: Our goal is to improve predictive performance and accuracy of the model; features were retained despite elevated VIF values in selective scaling. We assume that the correlated predictors do not affect our model's performance and despite the correlation they contain unique, non-overlapping information.

No Perfect Multicollinearity: It is assumed that no predictor is a linear combination of another predictor (E.g. Avoiding the Dummy Variable Trap – Scaling dummy variables may introduce noise and makes it harder for the model to estimate optimal coefficients). This ensures that the design matrix is invertible and model's coefficients can be estimated mathematically.

2. Analysis and Feature Engineering

2.1. Preliminary Analysis and Frequency Distributions

2.1.1. Missing Value Analysis

A missing value analysis was conducted across all features of our dataset prior to model development. The analysis confirmed that no missing values were present in any of the features. This ensured that our regression models were trained and tested on complete and consistent data without the risk of any bias or information loss.

2.1.2. Feature Classification

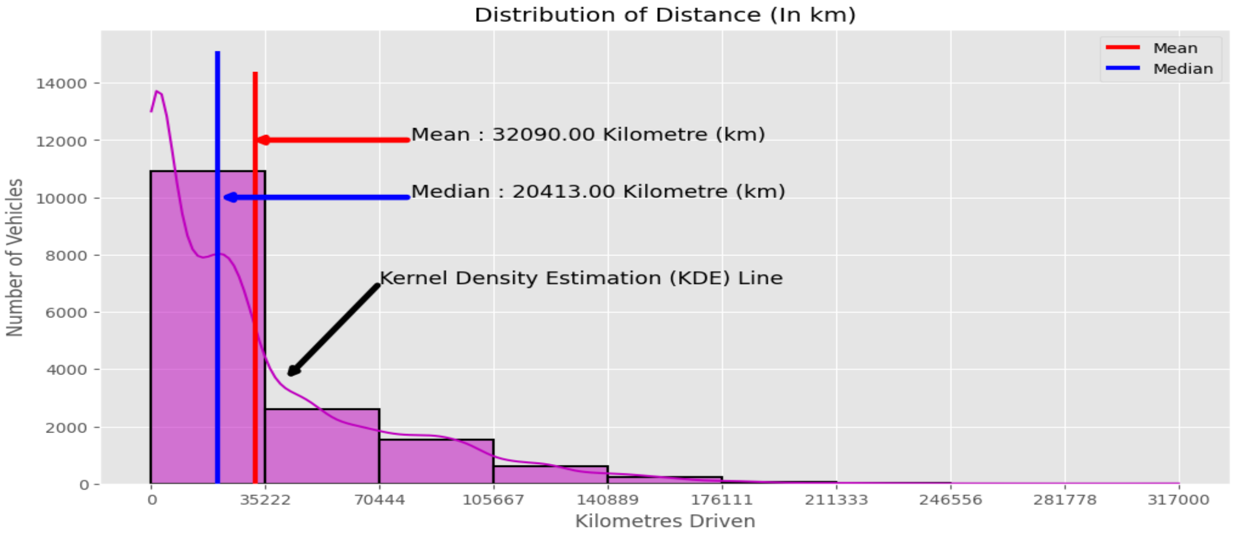
The features in our automobile dataset were classified into four distinct groups based on the type of the feature and their role: Categorical, Multi-value categorical, numerical and target. Multi-value categorical is a subtype of categorical. In this, multiple values might be present in a single observation. Example: Comfort Convenience features.

To ensure data integrity and prevent omission, a sanity check was implemented to ensure that all classified features account for all features present in our automobile dataset.

2.1.3. Frequency Distribution of Continuous Numerical Predictors

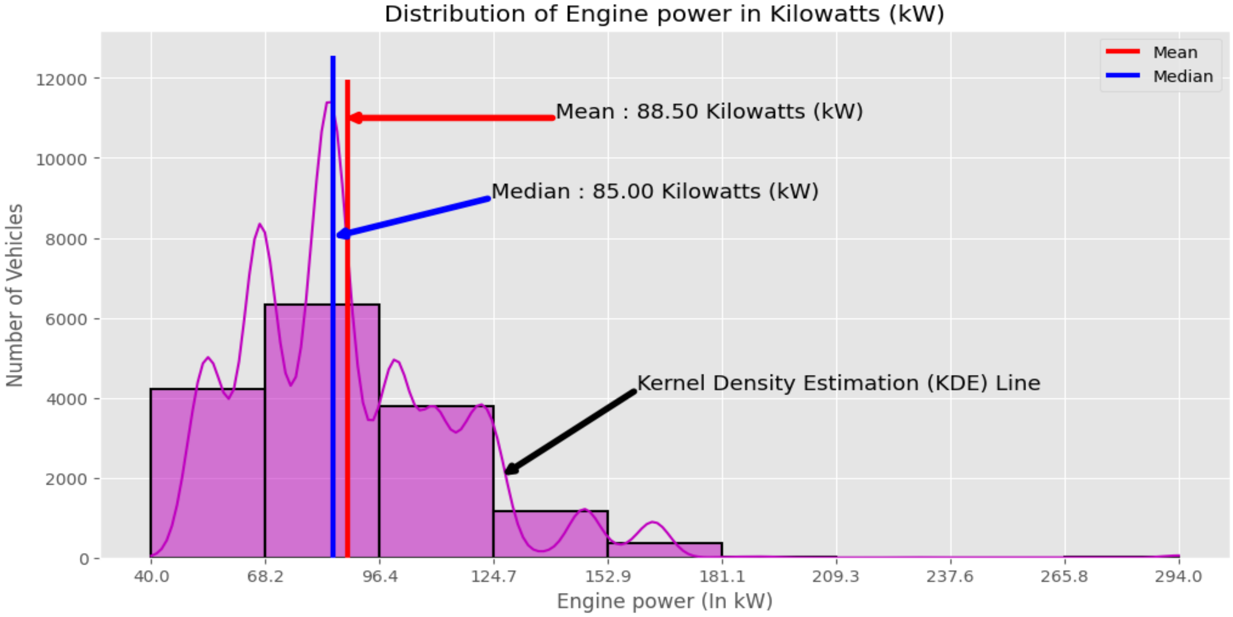
2.1.3.1. Distribution of Distance (Kilometers Driven)

The distribution of distance travelled by the vehicle exhibits a strong right-skewed pattern. There is a discrepancy between the measures of central tendency which can be visually observed, the mean is significantly larger than the median. Outlier analysis will be performed at a later stage to account for these extreme values.



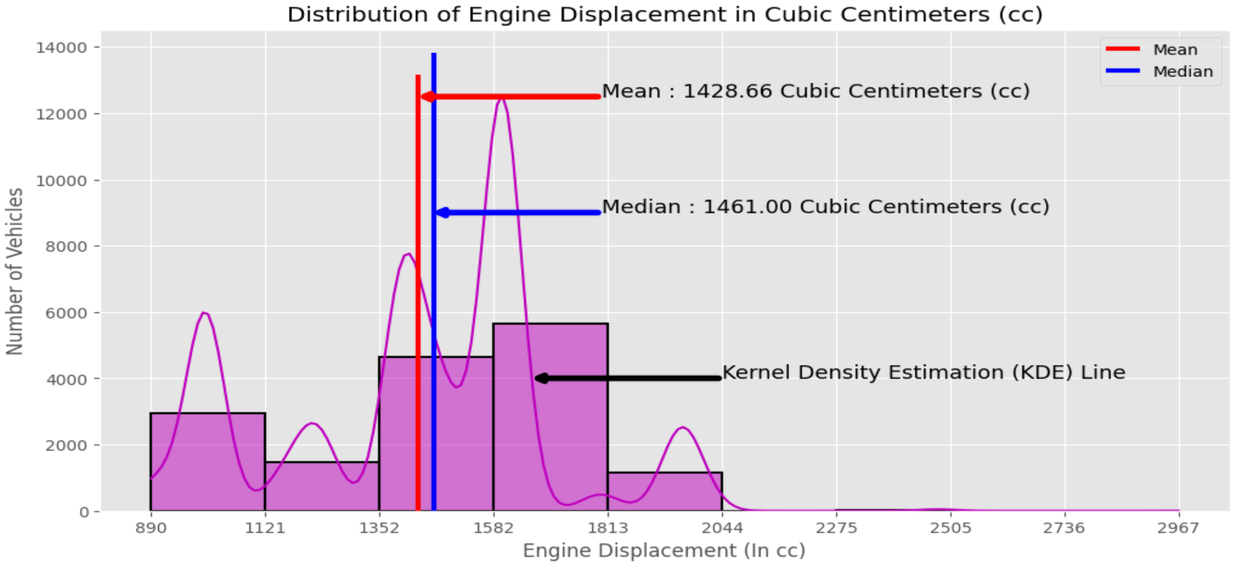
2.1.3.2. Distribution of Engine power in Kilowatts (kW)

The frequency distribution of Engine Power (measured in Kilowatts) exhibits a mild right-skewness. Most vehicles are concentrated in the mid-range [40 to 124.7 kW]. The difference between the mean and median confirms skewness towards the right which is driven by a smaller number of high-performance vehicles with significantly higher power. This feature has strong potential in yielding strong predictive signal.



2.1.3.3. Distribution of Engine Displacement (cc)

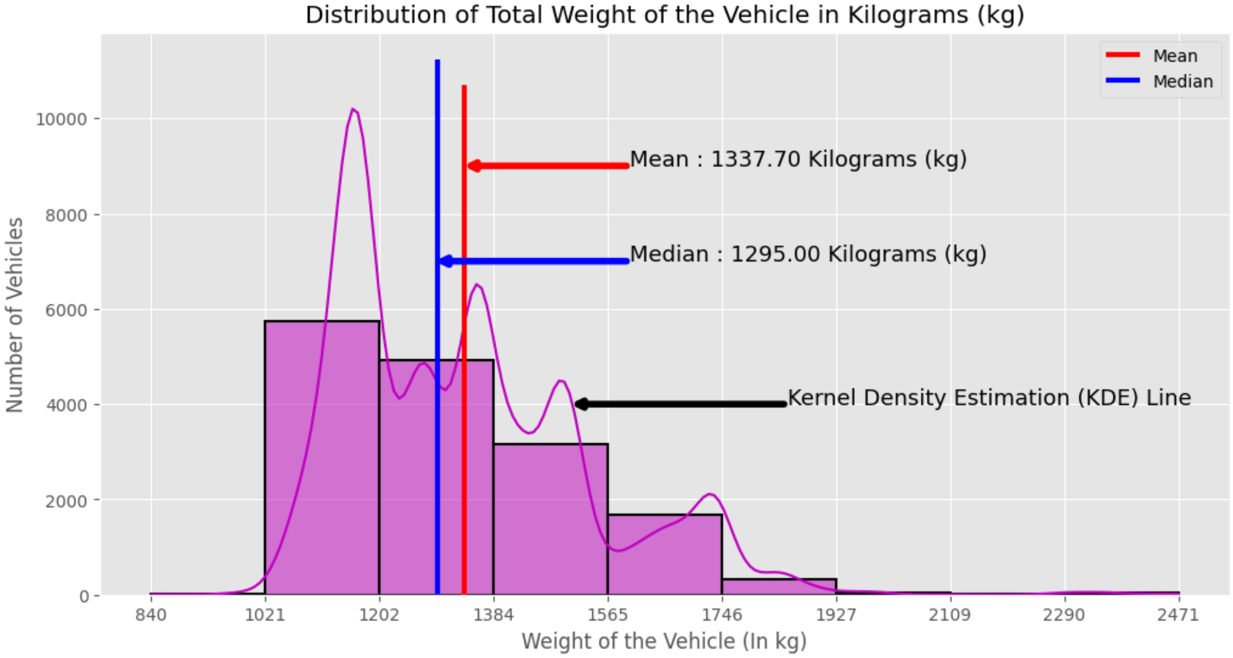
The distribution of engine displacement (measured in cubic centimeters) appears to be right-skewed. The distribution is highly multi-modal (multiple distinct peaks).



The median exceeds the mean, suggesting that the distribution is heavily weighted by higher volume engine displacement. Although it is numerical in nature, the chart shows signs that it might be a categorical disguised as a numerical [Standard size in industry].

2.1.3.4. Distribution of Total Weight of the Vehicle

The distribution of total vehicle weight (measured in kilograms) exhibits a mild right-skewness pattern. The measures of central tendency confirm the right-skewness, the distribution is pulled towards the right-tail by a smaller number of heavier vehicle (likely SUVs, Vans etc.). We should see acceptable levels of multicollinearity between engine power and displacement.

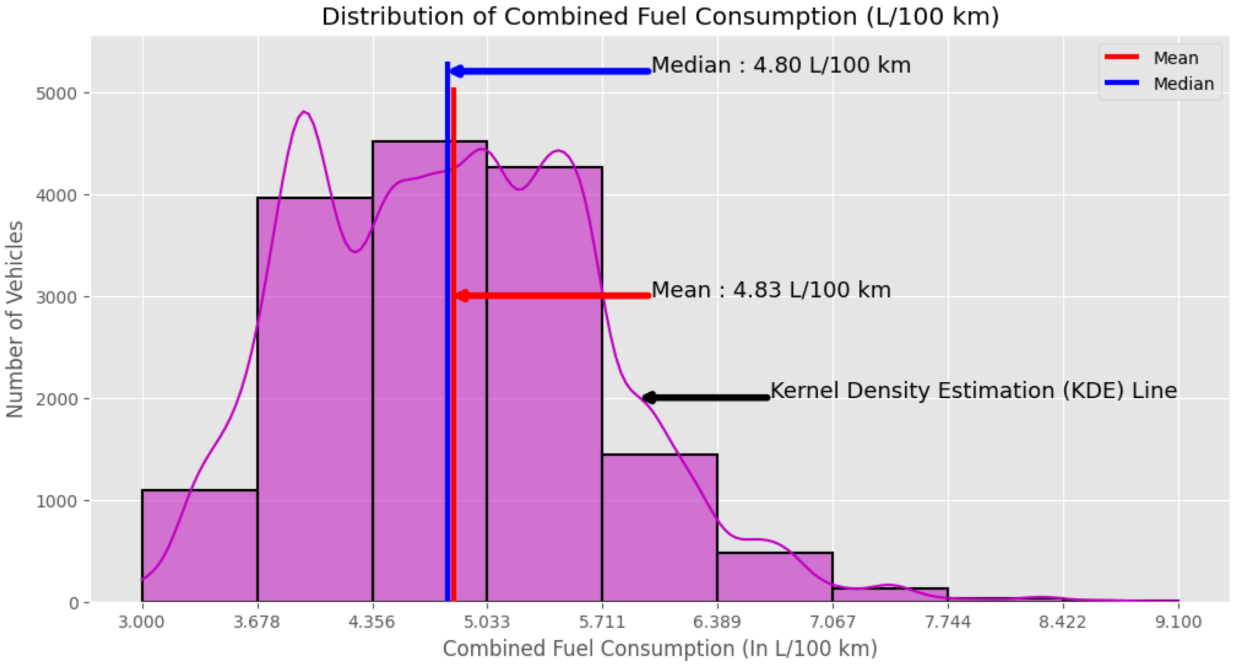


2.1.3.5. Distribution of Combined Fuel Consumption

The distribution of combined fuel consumption, (measured in litres per 100 kilometres) shows a remarkable near-symmetric and bell-shaped curve closely approximating to Normal Distribution.

While a right-skewness still exists, majority of the observations are clustered around the central tendency (mean and median).

From a modelling perspective, linear regression models can perform well with this feature as the relationship between fuel consumption and price is often stable and predictable. Reduced risk of heteroscedasticity or non-linearity.

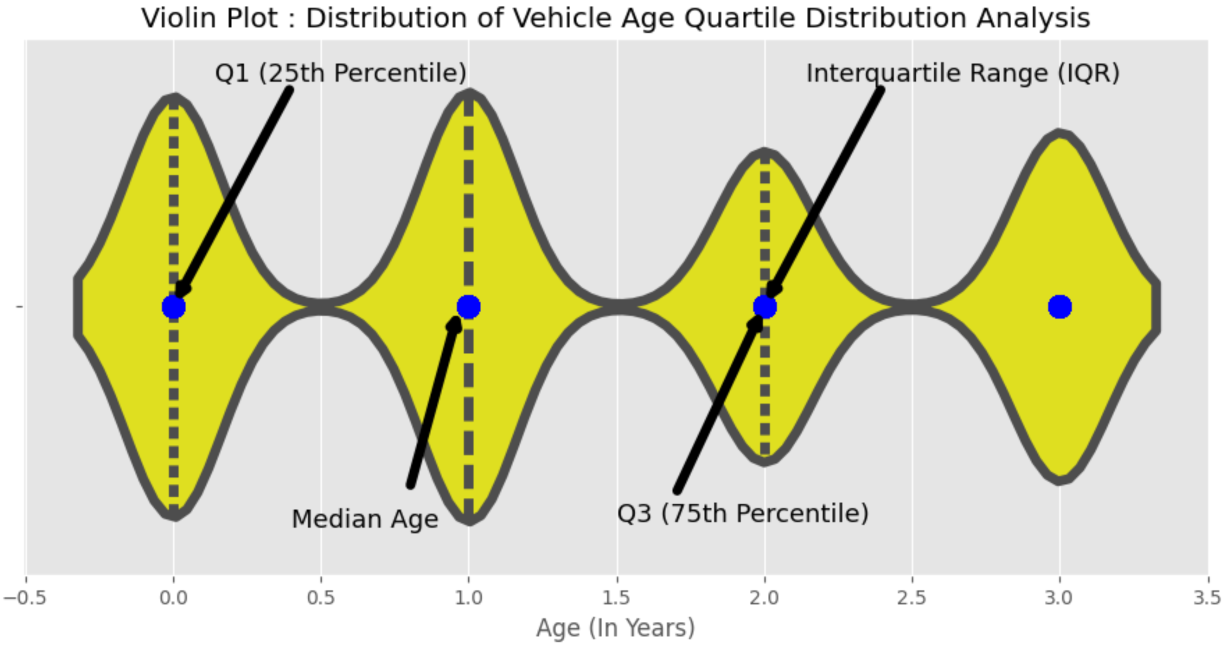


Distribution of Combined Fuel Consumption (L/100 km)

2.1.3.6. Distribution of Vehicle Age Quartile Distribution Analysis

Our analysis revealed that at least 25% of the used vehicles in the dataset are less than a year old. The analysis suggests that age of the vehicle might be an ordinal categorical variable.

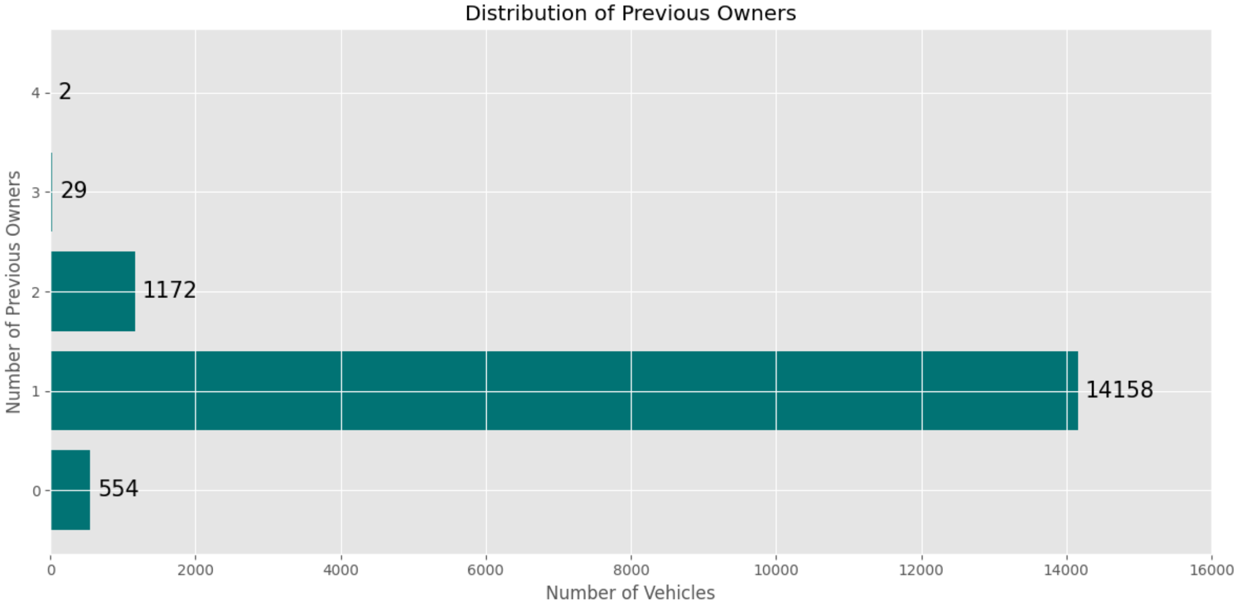
A bias could be introduced unknowingly by adding this feature to the predictors, the model will learn to accurately predict vehicles within 3 years but might not accurately generalize to old vehicles (5+ years).



Violin Plot : Distribution of Vehicle Age Quartile Distribution Analysis

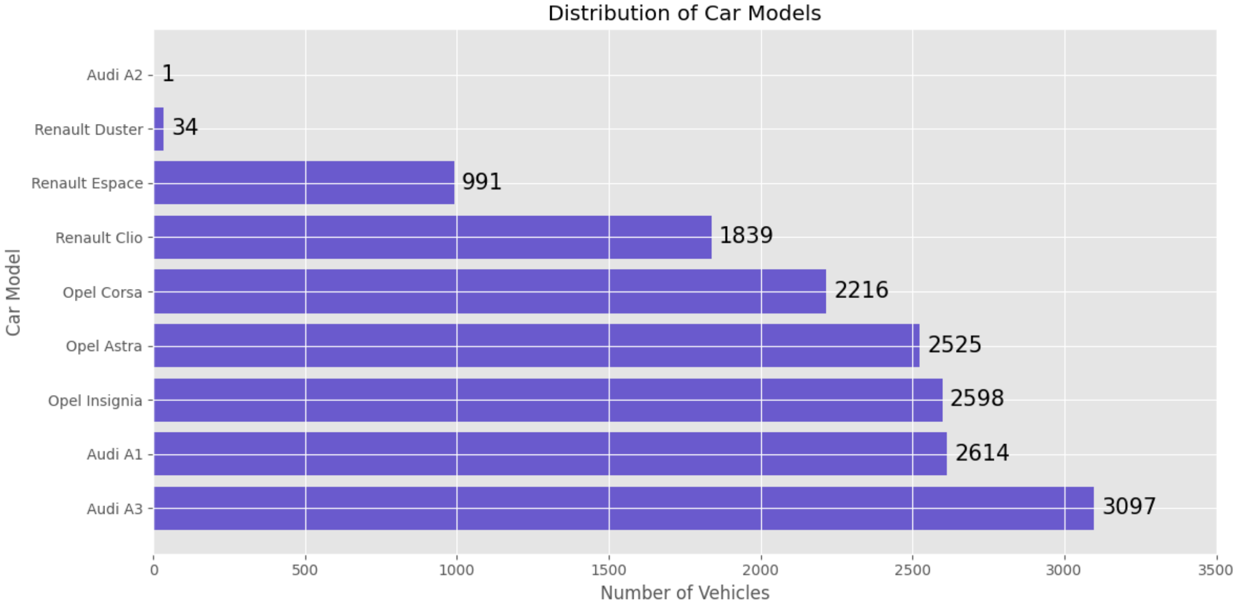
2.1.3.7. Distribution of Previous Owners

The distribution of the number of previous owners exhibits a discrete and highly imbalanced pattern. Vast majority of used vehicles had a single owner. From a modelling perspective, this feature might have near-zero variance and might fail to provide significant predictive power because it won't be able to predict and differentiate effectively between high-priced and low-priced vehicles.



2.1.4. Frequency Distribution of Categorical Predictors

2.1.4.1. Distribution of Car Models



The distribution analysis of car models reveal a highly imbalanced categorical structure which will require balancing at a later stage. The models with higher count will be very stable and have reliable coefficients.

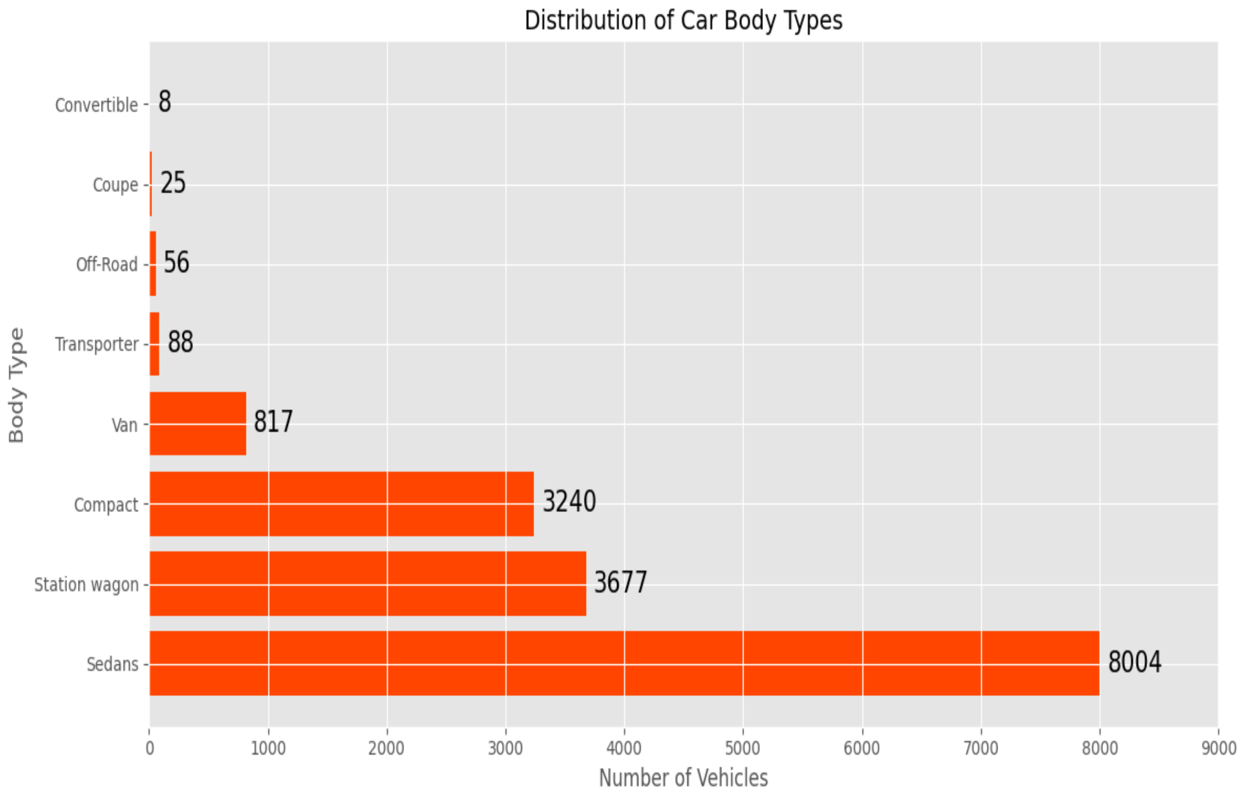
The distribution shows a healthy count for most models, which is essential for our model to learn brand-specific pricing. Due to the skewed frequency distribution, the imbalance has to be addressed, to avoid sparse representation and unstable coefficient estimates.

2.1.4.2. Distribution of Vehicle Body Types

The distribution analysis revealed that Sedans constitute a large proportion of vehicle body types, followed by Station wagons and compact cars.

Vehicle body type is often a major determinant of vehicle's price, but the extreme imbalance between the primary body types and niche body types would have to be addressed.

However, these categories reflect legitimate vehicle body types, and therefore cannot be removed and needs careful pre-processing.

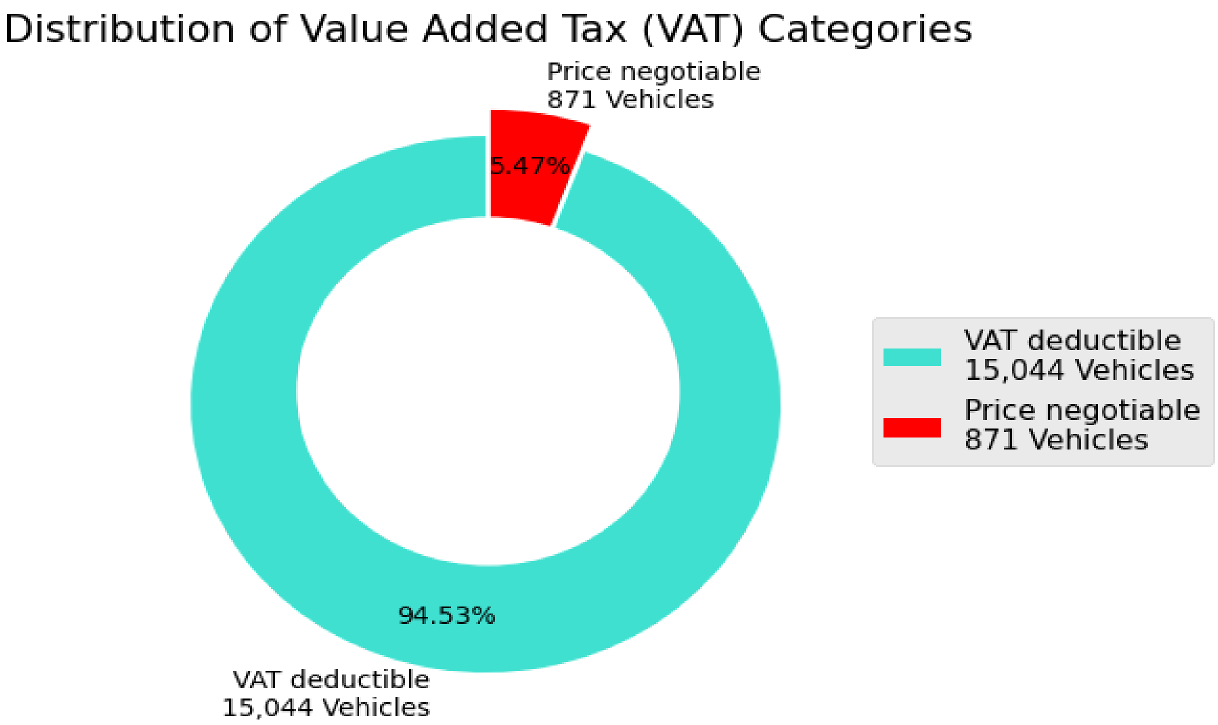


2.1.4.3. Distribution of Value Added Tax (VAT) Categories

The distribution of Value Added Tax (VAT) categories revealed a strong class imbalance, the majority of vehicles in the dataset are classified as VAT deductible.

Any coefficient assigned to Price Negotiable would be less statistically robust compared to VAT deductible.

From a modelling perspective, the feature might still play an important role in indirectly capturing vehicle pricing. Despite this imbalance, i decided to retain both the categories to preserve transactional context in the automobile market.

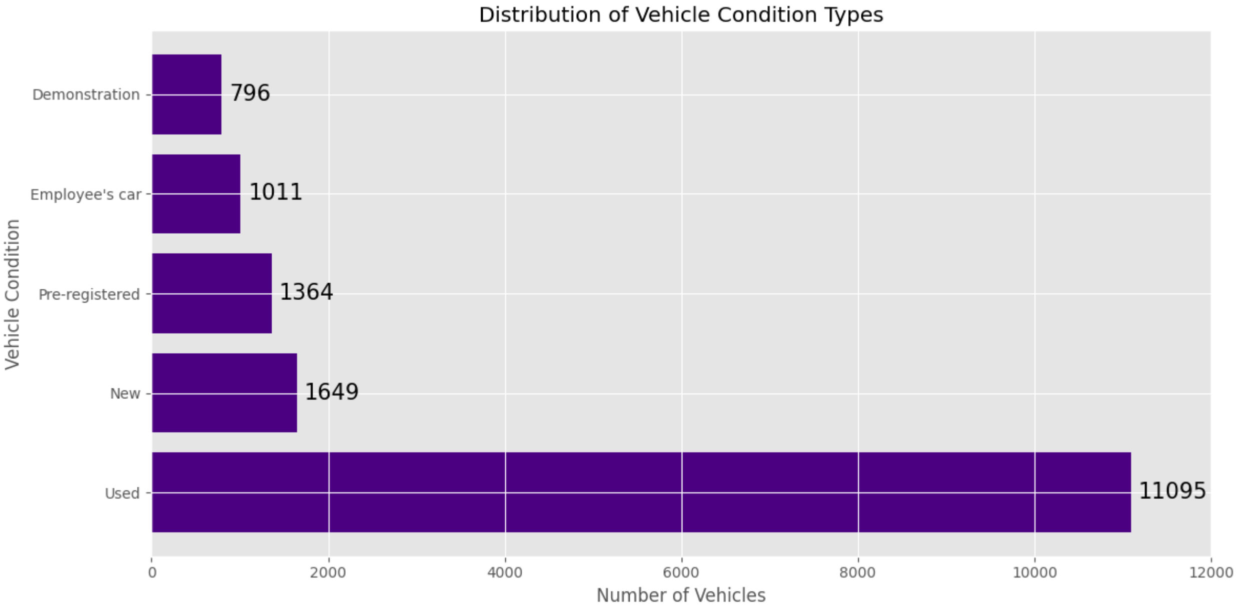


2.1.4.4. Distribution of Vehicle Condition Types

The distribution of vehicle condition types reveals a highly imbalanced structure since the market is heavily dominated by used vehicles, which is perfectly valid since we are predicting used vehicle prices.

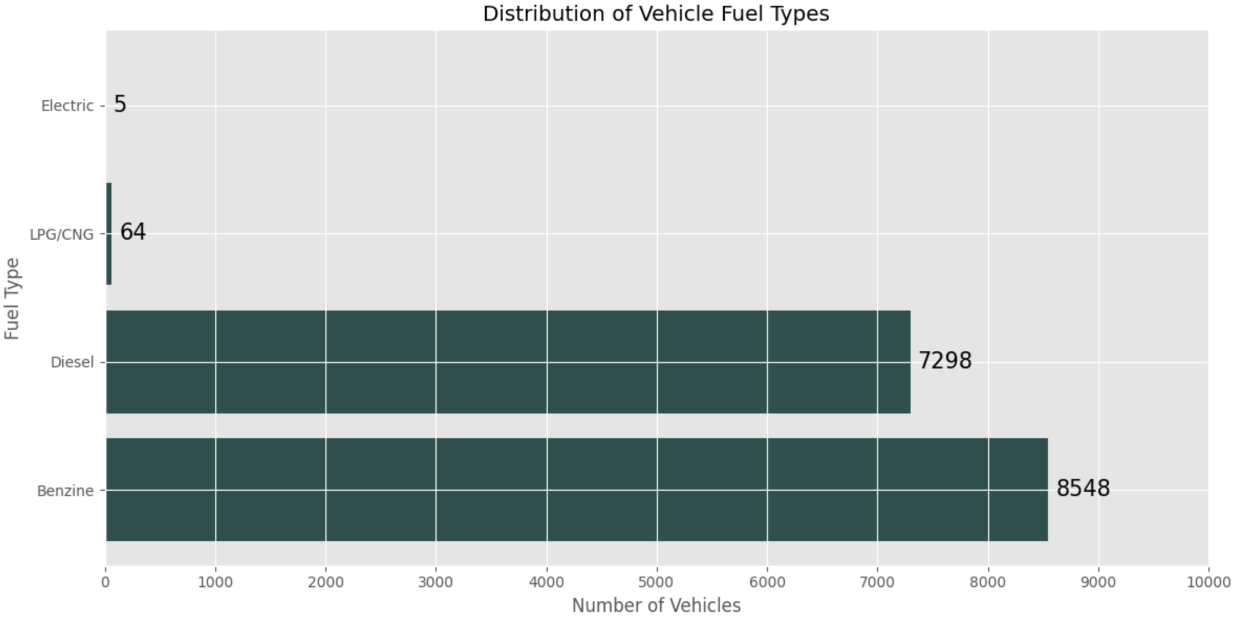
Although the counts might appear less for other sub-categories within our feature, they may provide the necessary context for the model to understand vehicle pricing.

There might be a high correlation between vehicles labeled "New" and those with near zero mileage if this relationship is not addressed. This kind of relationship between the two predictors reinforces the need for Regularization techniques to manage potential redundancy and multicollinearity.



2.1.4.5. Distribution of Vehicle Fuel Types

The distribution of vehicle fuel types reveals a highly imbalanced structure. The automobile market in Germany suggests that Benzine is more common. The extreme sparse labels reflect emerging or less common fuel technologies rather than data inconsistency, and will need to be addressed to at a later stage.



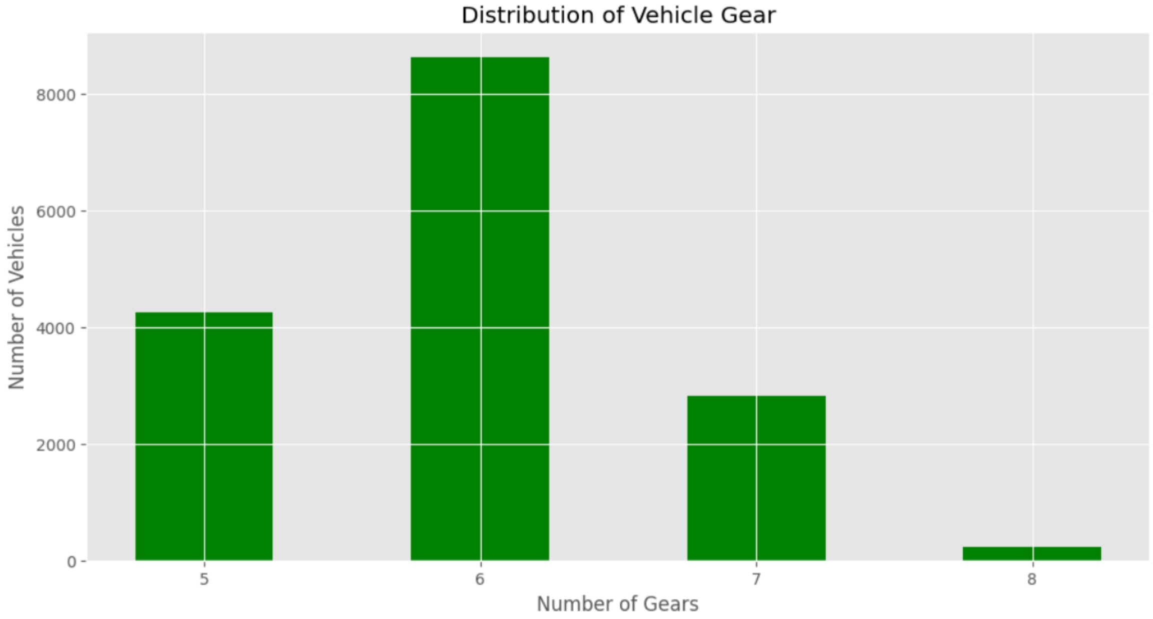
Fuel type often serves as a proxy for multiple other attributes, such as body type, combined fuel consumption etc. Making it a potentially influential predictor in vehicle price estimation.

2.1.4.6. Distribution of Vehicle Gear

The distribution of gear count in the Vehicle's Transmission reveals the mechanical standards in the German automobile industry.

Strategies will be implemented at a later stage to ensure that the dominant gear categories do not overshadow less frequent categories, while maintaining model stability and generalization while performing regression.

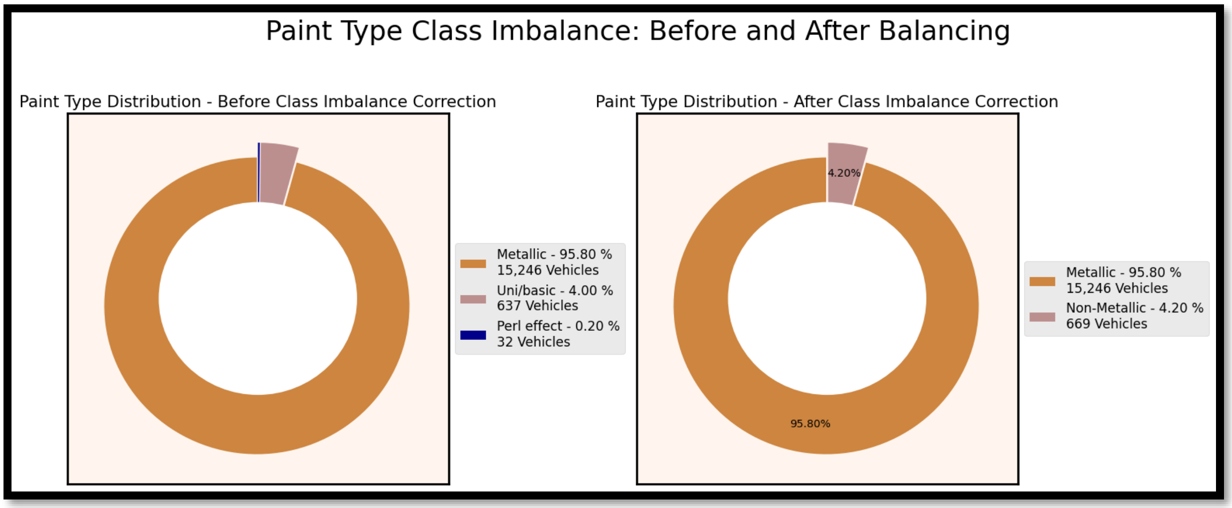
6-Speed Transmissions are the most occurring, with approximately 8500 Vehicles, followed by 5-speed with roughly 4,200 vehicles. Treating this feature as a continuous variables might lead to errors.



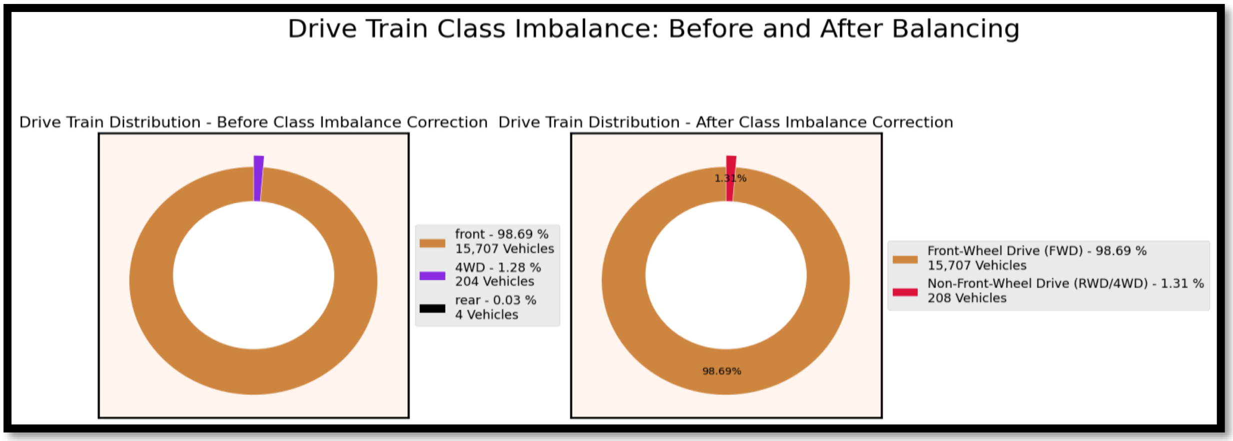
2.1.5. Handling Class Imbalances

To improve model stability and prevent bias toward majority classes, categorical features were carefully aggregated into broader and more logical consistent groups. It is assumed that this would help the model generalize and perform better.

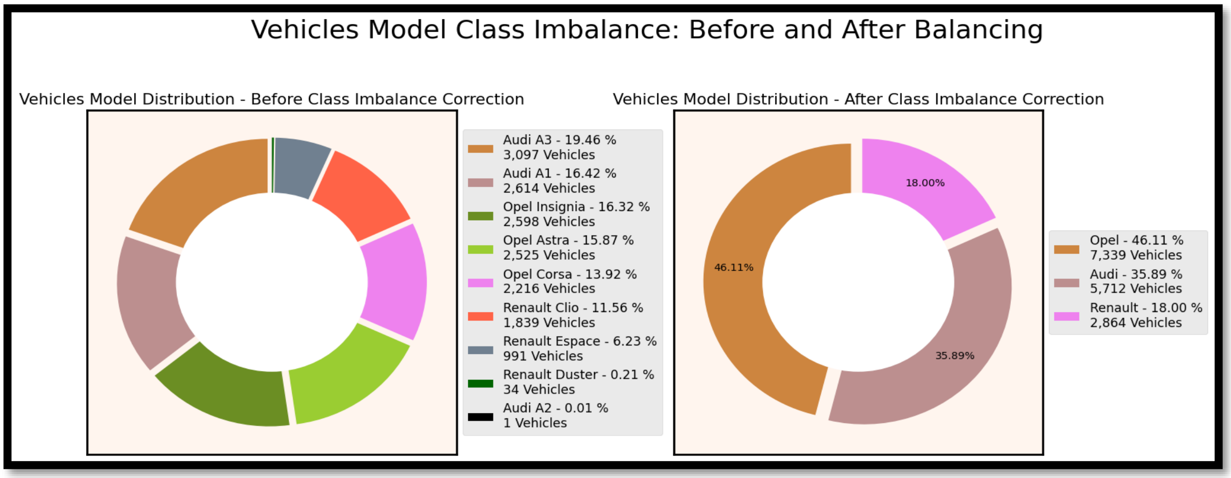
1) **Paint Type:** Aggregated into a binary classification of Metallic versus Non-Metallic



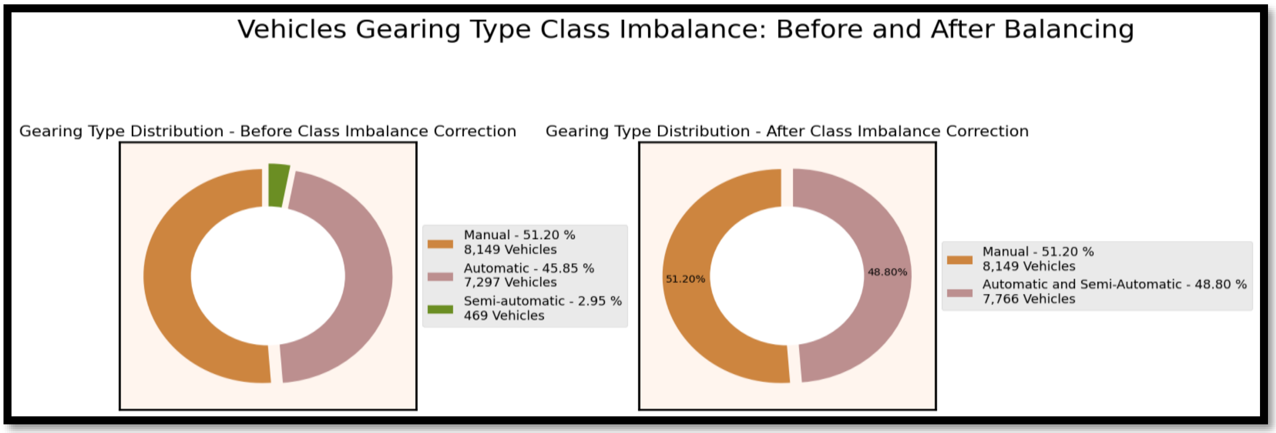
2) **Drivetrain:** Simplified into Front-Wheel Drive (FWD) versus Non-Front-Wheel Drive



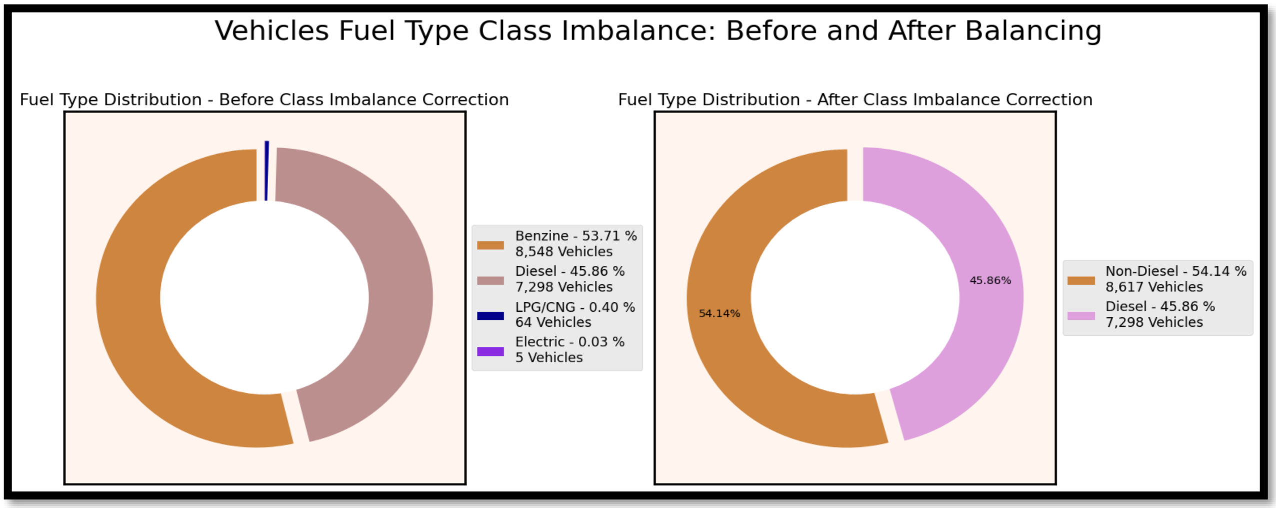
3) **Vehicle Model:** Consolidated specific model variations into the three primary manufacturers: Audi, Opel, and Renault.



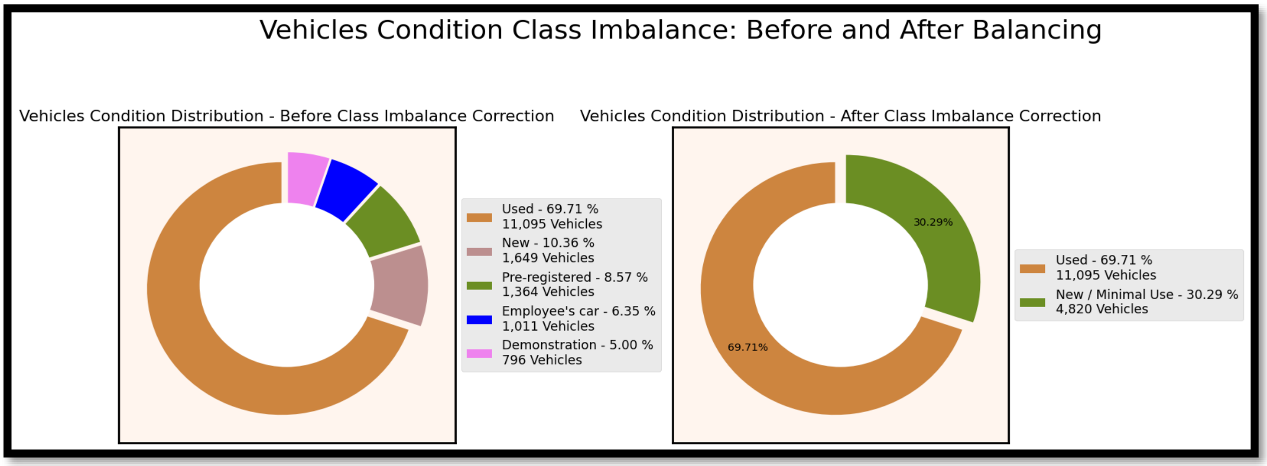
4) **Gearing Type:** Merged Automatic and Semi-Automatic into a single category, creating a binary comparison against Manual transmission.



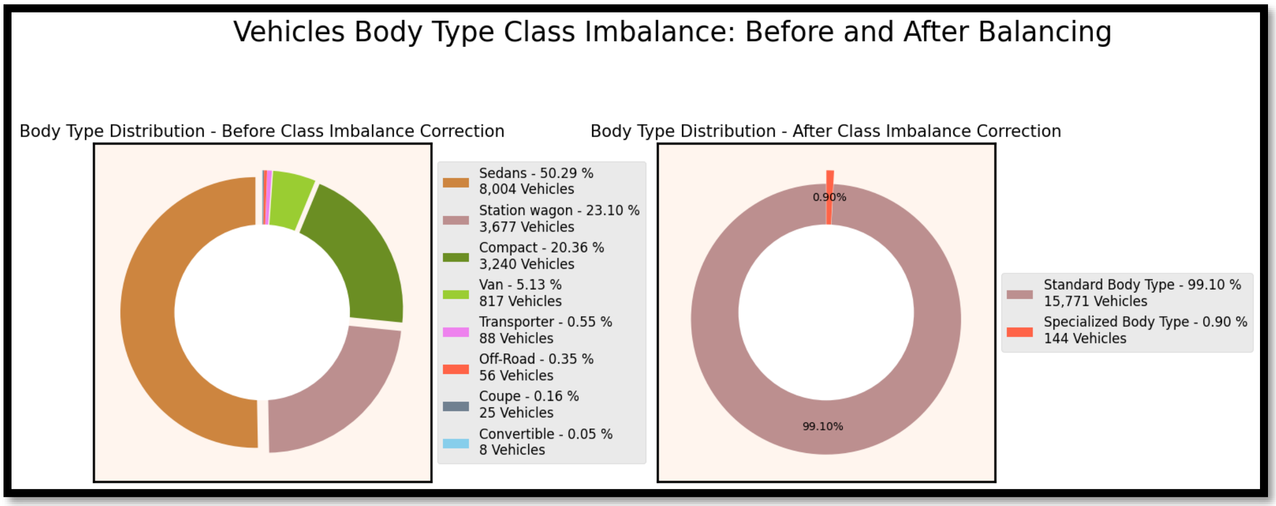
5) **Fuel Type:** Grouped into Diesel versus Non-Diesel, aggregating Benzine, LPG, CNG, and Electric into the latter category due to sample size dominance of Diesel.



6) **Vehicle Condition:** Aggregated into Used versus New / Minimal Use (combining pre-registered, employee, and demonstration vehicles).

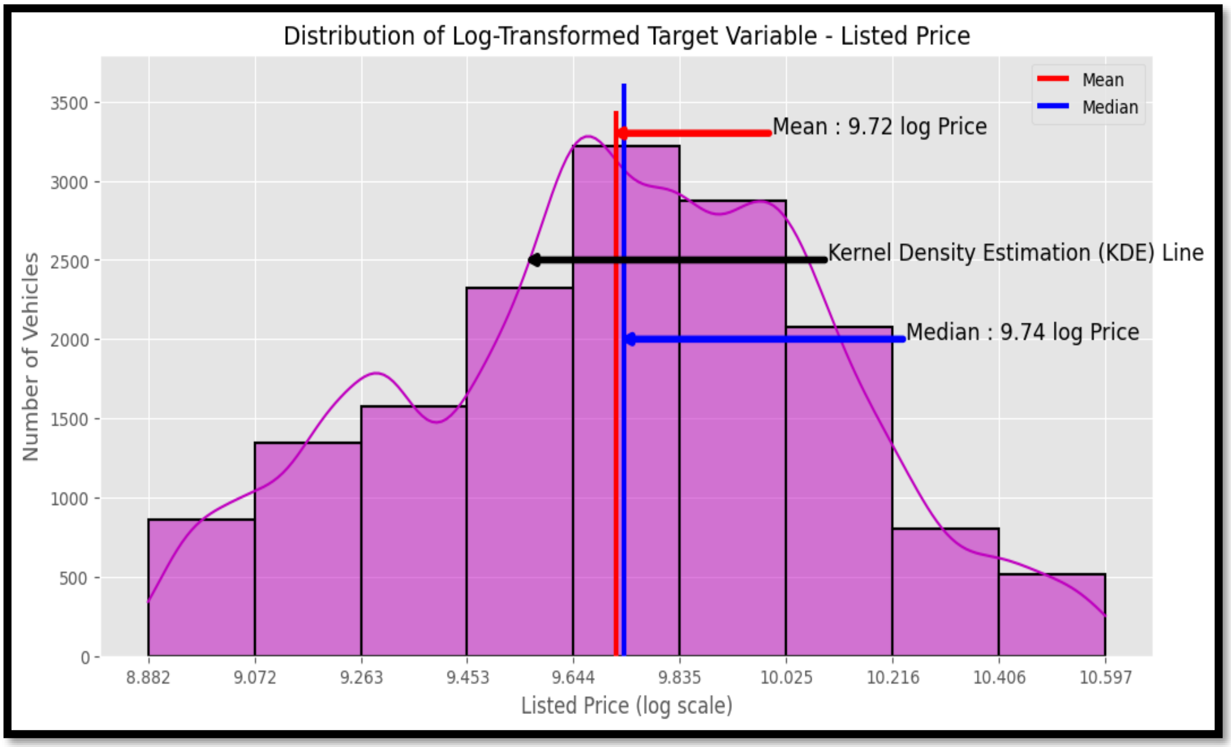


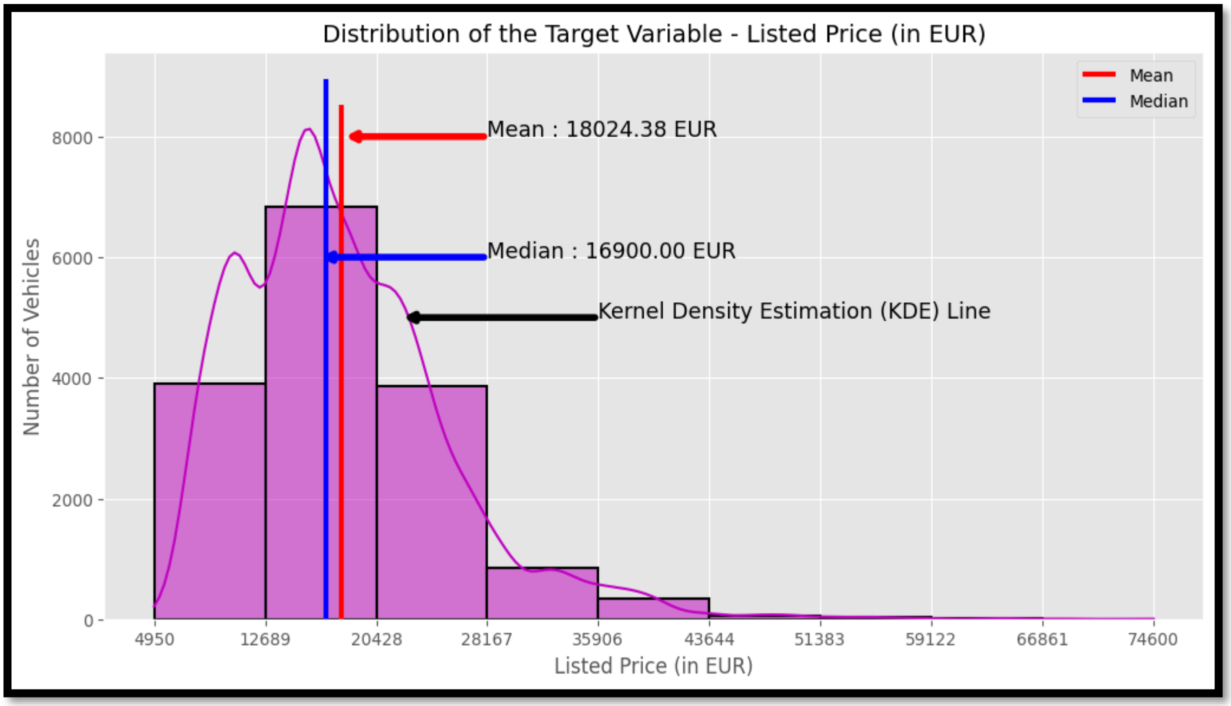
7) **Body Type:** Classified into Standard Body Type (Sedans, Station Wagons, Compacts, Coupes) versus Specialized Body Type (Transporters, Off-Road vehicles).



2.1.6. Frequency Distribution of Target Variable [Log-Transformation]

Prior to model development, a quantile-based outlier analysis was performed on the listed price to reduce the influence of extreme values which would affect our model's performance. Observations below the 1st percentile and above the 99th percentile were removed, followed by a log-transformation of price. The log-transformation reduces right-skewness, stabilizes variance and results in a near-normal distribution.





2.2. Correlation Analysis

2.2.1. Correlation between Numerical Features and Target Variable

Our correlation analysis between numerical predictors and log-transformed listed price reveals a clear pattern and relationship.

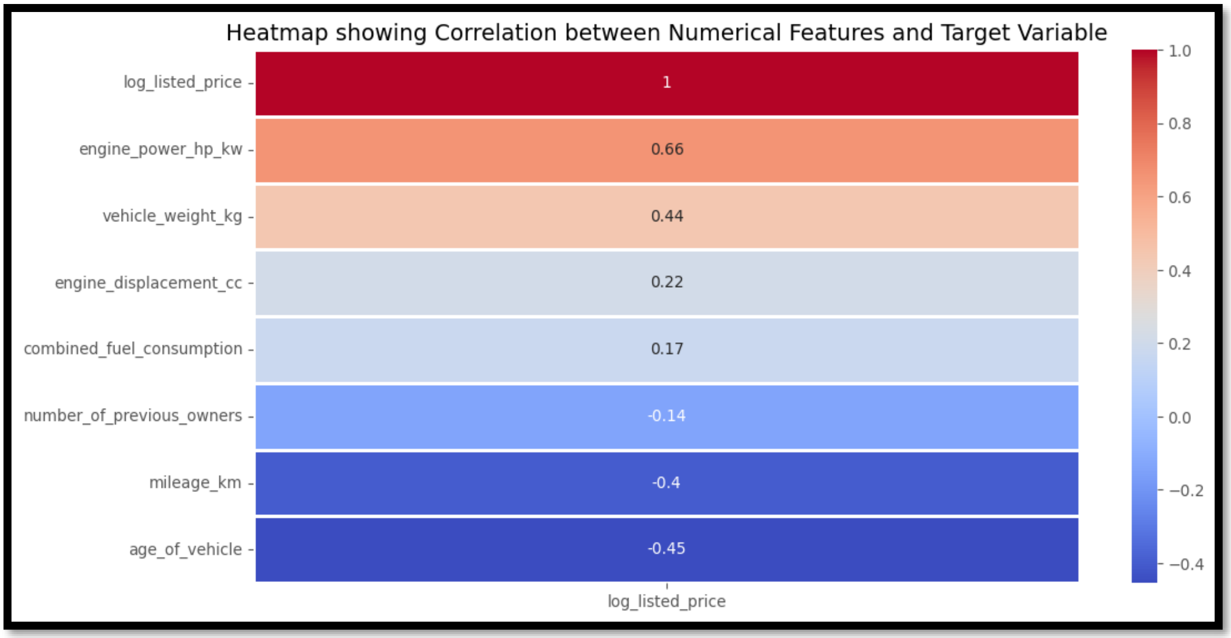
Engine power shows the strongest positive correlation with vehicle price, confirming that it is a primary indicator of predicting vehicle prices.

Vehicle Weight shows a moderate positive correlation. Indicating that heavy vehicles like (SUVs, Sedans) typically cost more.

Vehicle age shows the strongest negative correlation, this aligns with the standard depreciation logic of used vehicle price - as a vehicle gets older, its value drops significantly.

Mileage shows a similar strong negative correlation like vehicle age, indicating that the vehicles price is not only influenced by the age of the vehicle but also the distance driven.

Previous owners and Combined Fuel Consumption have a weak correlation with price, indicating that they are not a strong predictor compared to the others.



2.2.2. Correlation between Categorical Features and Target Variable

Evaluating the statistical significance of a categorical feature on vehicle pricing works differently.

Both One-Way ANOVA (Analysis of Variance) and Kruskal Wallis tests were conducted on log-transformed listed price. This Dual-Test Strategy ensures true statistical significance and robustness.

- **One-Way ANOVA (Analysis of Variance) -** The Test analyze the means of vehicle's price distribution across categories.

- **Kruskal Wallis Test -** This Test analyze the medians of price distributions across categories.

Hypothesis Framework:

- **Confidence Level -** 95%
- **Significance Level -** 0.05 (alpha)
- **Null Hypothesis -** All groups have the same central tendency and come from the same population. Meaning that the price distribution is the same across all groups.
- **Alternate Hypothesis -** At least one group has a different central tendency from the others.

Interpretation of Results:

- **Based on the Dual-Tests**, the tests revealed that 10 out of the 11 categorical features demonstrated a statistically significant impact on log-transformed listed vehicle price with a confidence level of 95%. The agreement of ANOVA and Kruskal-Wallis tests on all 11 categories confirms the true robustness of the results.
- **Body type is the only categorical variable** for which both ANOVA and Kruskal–Wallis tests fail to reject the null hypothesis, indicating no statistically meaningful difference in either mean or median price across its groups.

2.3. Outlier Analysis

Note: *Vehicle Age and Number of Previous Owners did not need outlier handling.*

The outlier analysis strategy which was implemented was applied across all continuous numerical features, this ensures consistency and avoid feature-specific bias. Statistical analysis was performed across all numerical features using descriptive statistics and quantile-based analysis, including the 1st, 25th, 50th, 75th, and 99th percentiles.

Upper-Tail Winsorization: Linear Regression models are highly sensitive to outliers and extreme values can pull the regression line, which would increase error variance. To mitigate this without reducing the sample size, Upper-Tail Winsorization (Clipping) at the 99th Percentile. This technique caps extreme values at the 99th percentile threshold rather than removing them.

This approach was chosen for two main reasons:-

1) **Data Preservation:** It allowed us to keep valid rare high-end vehicle listings (Example: Sports, Luxury, High-end vehicles) without allowing them to skew our model's coefficients.

2) **Left-Tail Preservation:** Lower-bound outliers weren't clipped as they represent real-world data variation and allow the model to learn from these variations. They do not introduce error bias in price prediction models.

2.4. Feature Engineering

Rather than manually removing redundant predictors, regularization techniques (Ridge and Lasso) were intentionally leveraged to perform coefficient shrinkage and redundancy management during model training.

2.4.1. Handling Multi-Categorical Features [Multi-hot encoding]

The AutoScout vehicle pricing dataset contains several multi-categorical attributes, these were addressed to by applying a method known as Multi-Hot Encoding, which initially expanded the dataset to 118 individual features and were reduced to 54 features after aggregation strategy was applied.

While Regularization methods like Ridge and Lasso Regression can perform feature shrinkage, grouping these features manually reduce the curse of dimensionality, mitigate the risk of high correlation between predictors and improve model interpretability.

1. <u>Aggregation Strategy: "Any" vs. "Count"</u>

- **Binary representation (“Any”):** Indicates the presence (one [1]) or absence (zero [0]) of a multi-categorical feature. This is based on the principle of “presence vs absence” of a feature.

- **Count-based representation (“Count”):** Captures the number of occurrences of a multi-categorical feature. This is based on the principle of “frequency of occurrences” of a feature.

2. <u>Feature Sub-Groups Implementation</u>

Binary Representation and Count-based representation were implemented for the feature sub-groups.

- **Comfort and Convenience Features:** Aggregated into Five meaningful sub-groups - Parking Convenience, Climate Control, Seating Comfort, Driving Control and Interior Tech.
- **Entertainment and Media Features:** Aggregated into Three meaningful sub-groups - Audio, Connectivity and Visual.
- **Safety and Security Features:** Aggregated into Five meaningful sub-groups - Airbag, Lighting & Visibility, Theft Prevention, Core safety and Driving Assistance.
- **Extra Features:** Aggregated into Four meaningful sub-groups - Sports & Performance, Interior & Comfort, Exterior & Utility and Special Equipment’s.

2.4.2. Categorical Feature Encoding [One-Hot Encoding]

Categorical Features such as vehicles model, body type, vat status, condition, fuel type, paint type, upholstery type, gearing type and drive train type were encoded using One-Hot Encoding.

To ensure that the model can interpret categorical features and avoid the dummy variable trap. One (1) indicates the presence of a feature, Zero (0) indicates the absence of a feature.

This encoding strategy is essential as it allows linear and regularized models to capture category-specific pricing effects while maintaining numerical interpretability throughout. Feature names were standardized for clarity after performing one-hot encoding.

2.4.3. Splitting Data: Training and Testing Sets

Our AutoScout Dataset was partitioned into 80% Training Data and 20% Testing Data. Allocation of 80% Training ensures that the models will learn patterns and generalize better.

- **X --->** Represents multiple predictors.
- **y ---->** Represents a single target variable (Log-Transformed Price)

A random state of 42 was utilized to ensure reproducibility of the split across different runs.

Verification checks were implemented to ensure the proportion of split and data integrity. This split was carefully considered to balance the model's learning capacity and unbiased performance evaluation for the scope of predicting vehicle prices.

2.4.4. Feature Scaling and Assumptions

Selective Feature Scaling (Standardization) was implemented to continuous numerical predictors and count-based feature-engineered features using StandardScaler (Z-score normalization), ensuring that the mean is zero and standard deviation is one. Feature Scaling ensures that all features have equal weights, are comparable and converges effectively.

Excluded Features: All Binary features (0/1 indicators), including "Any" features and One-Hot Encoded variables, were intentionally excluded. Scaling these features would destroy their interpretability, which is important for business stakeholders.

Prevention of Data Leakage: It was implemented after the train-test-split, fit exclusively on the Training Set preventing any form of data leakage. Ensuring that no information from the test set is seen by the model, providing a realistic assessment of model performance.

Multicollinearity (VIF): Scaling all the predictors would typically reduce VIF Score by improving the numerical conditioning of the regression matrix. It is important to note that scaling does not eliminate the underlying structural correlations among binary features. Since our primary objective is predictive performance rather than strict coefficient inference, these binary features weren't scaled despite their elevated VIF scores in unscaled setting. My priority was model's ability to accurately predict price on unseen data rather than perfect statistical independence among predictors.

3. Linear Regression Models

3.1. Baseline Linear Regression Model

Our dataset was divided into two main groups (encoding-based groups) to evaluate how multi-categorical feature influence model's behavior and predictive performance:-

1) **Experiment Design A: Binary representation ("Any"):** Indicates the presence (one [1]) or absence (zero [0]) of a multi-categorical feature. This model is based on the principle of "presence vs absence" of a feature.

2) **Experiment Design B: Count-based representation ("Count"):** Captures the number of occurrences of a multi-categorical feature. This model is based on the principle of "frequency of occurrences" of a feature.

Note: *To ensure fairness, both the experiments were conducted independently, and evaluated at a later stage while keeping all other predictors identical.*

Common Tasks performed on both the experiments include:-

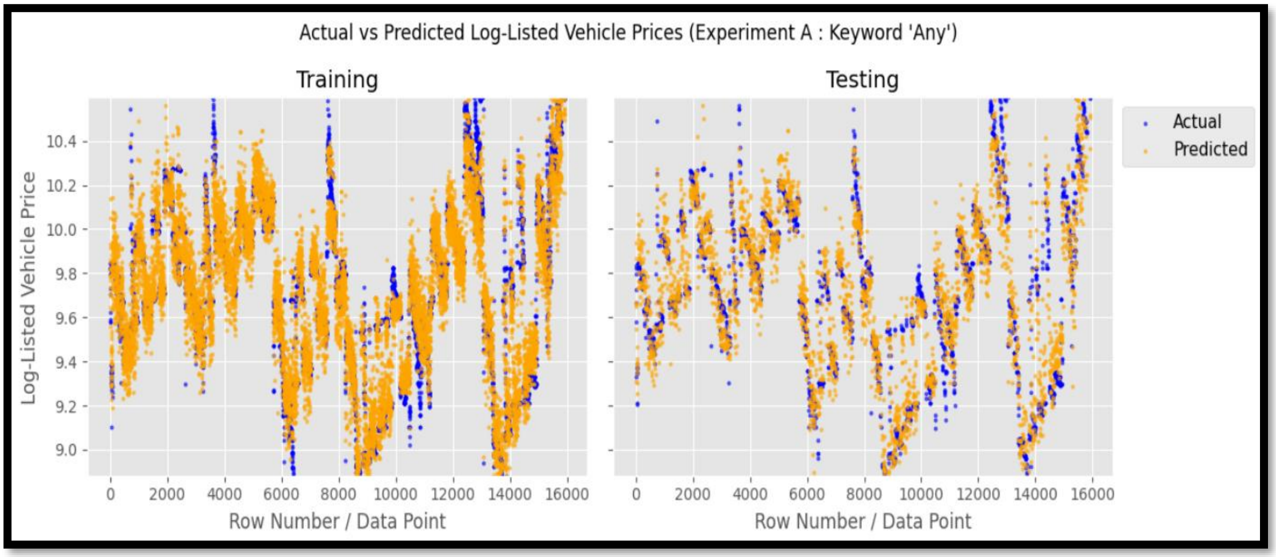
1) **Instantiating and Building:** A baseline Linear Regression model was initialized and trained using the relevant scaled training data.
2) **Fitting and Predicting:** Predictions were generated for both the training and testing sets on the log-transformed target variable (log-listed price).
3) **Evaluation:** Performance evaluation was done on both the transformed log-listed price and the original price scale (using inverse exponential transformation).

3.1.1. Experiment Design A: Binary ("Any")

In this experiment, all count-based features were removed from our scaling training and testing datasets. This approach prioritizes model simplicity and interpretability.

From the visualization we can observe that the predicted values closely follow the actual log-listed prices. Although the predictive points appear to be more scattered than the actual values it seems to be learning from the patterns and generalizing well, with no clear signs of overfitting.

This kind of behaviour indicates that the model effectively captures the overall pricing trend. Our binary-encoded experiment effectively captures presence of multi-categorical features without breaking the linear assumptions.

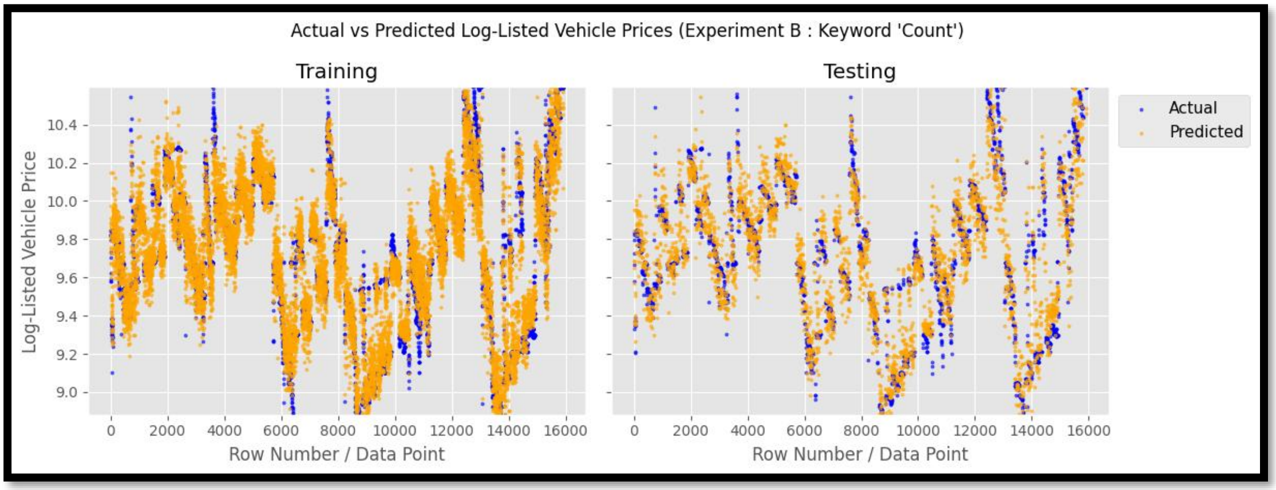


3.1.2. Experiment Design B: Count-Based ("Count")

In this experiment, all any-based features were removed from our scaling training and testing datasets. This approach allows the model to learn the frequency of feature occurrences, enabling a direct comparison with our binary experiment.

Similar to Experiment A, the count-based model shows a strong fit. This suggest that we can easily compare the two models based on their evaluation metrics. Count-based representation may introduce additional variance and multicollinearity, especially when feature frequencies are not linear with price.

Although the model still learns meaningful patterns and relationships, there seems to be a marginal decline as compared to the binary experiment.



3.1.3. Comparative Analysis: Experiment Design A vs B

To determine the most effective approach for handling multi-categorical features, on the basis of the evaluation metrics across both designs on both log-transformed and real-world listed price.

The evaluation metrics for both the models have a high predictive accuracy and generalize well on both our training and testing tests. R^2 values ranging from 0.85 to 0.86 for training and testing data, indicating both encoding methods are capable of capturing the variance in listed vehicle prices.

1) **Experiment A** shows a slightly lower Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) on unseen data, suggesting a marginally better fit for the average vehicle in the test set. Although these differences modest, they align with our visualizations.

2) **Experiment B** performs marginally better on the training set; however it fails to perform on the testing set leading to lower training RMSE and marginally higher training R^2. This suggests that the count-based approach introduced additional variance without any meaningful gain in our predictive accuracy.

3.1.4. Baseline Model Selection [Experiment Design A (Keyword: "Any")]

Despite the near-identical statistical performance, **Experiment Design A (Keyword: "Any")** has been selected as our primary linear regression baseline model for further model refinement.

This decision is based on three fundamental principles and domain-specific logic:-

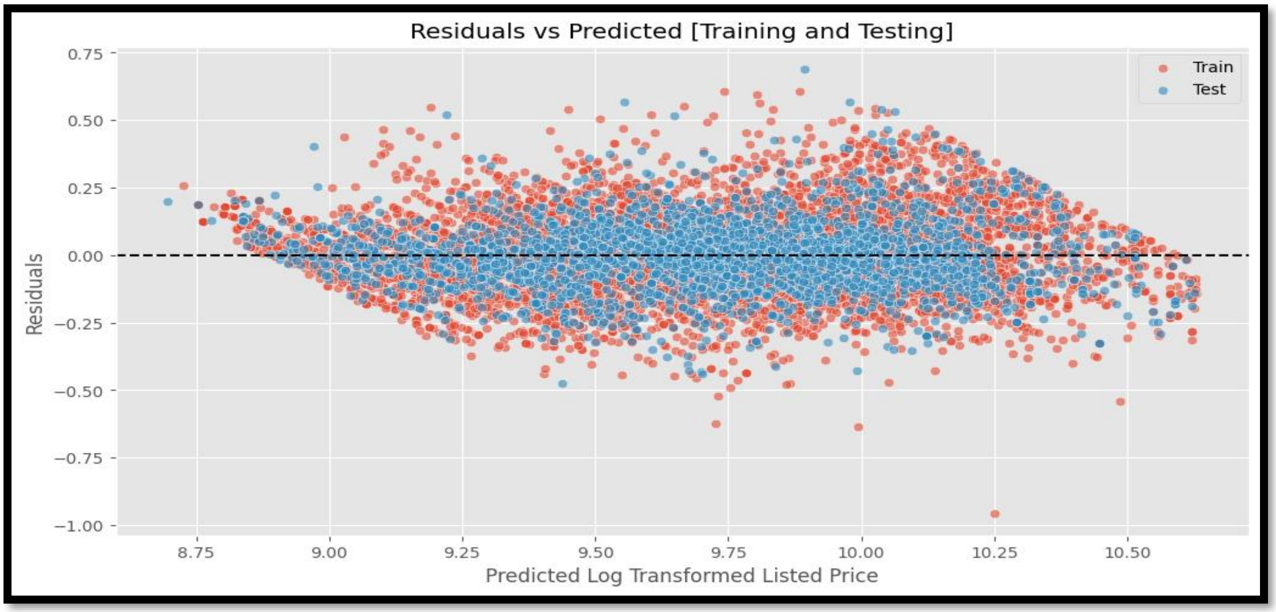
1) **Diminishing Marginal Utility:** In the automotive market, the presence of a feature often carries significantly more value than the quantity of an individual feature within that category. Example: One airbag is a critical baseline for safety and pricing, and the difference between having one airbag and ten airbags does not result in a linear 10x price increase.

2) **Principle of Parsimony (Occam's Razor):** The principle is based on "If you have two or more models, always choose the simplest model to better explain the data" (Keep It Simple, Stupid). Experiment A utilizes a binary representation "presence vs absence" approach. This results in a simpler and more interpretable model compared to count-based approach. Count-based approach failed to provide a statistically significant improvement in its R^2 or MAPE, the simpler binary model is preferred to avoid unnecessary complexity.

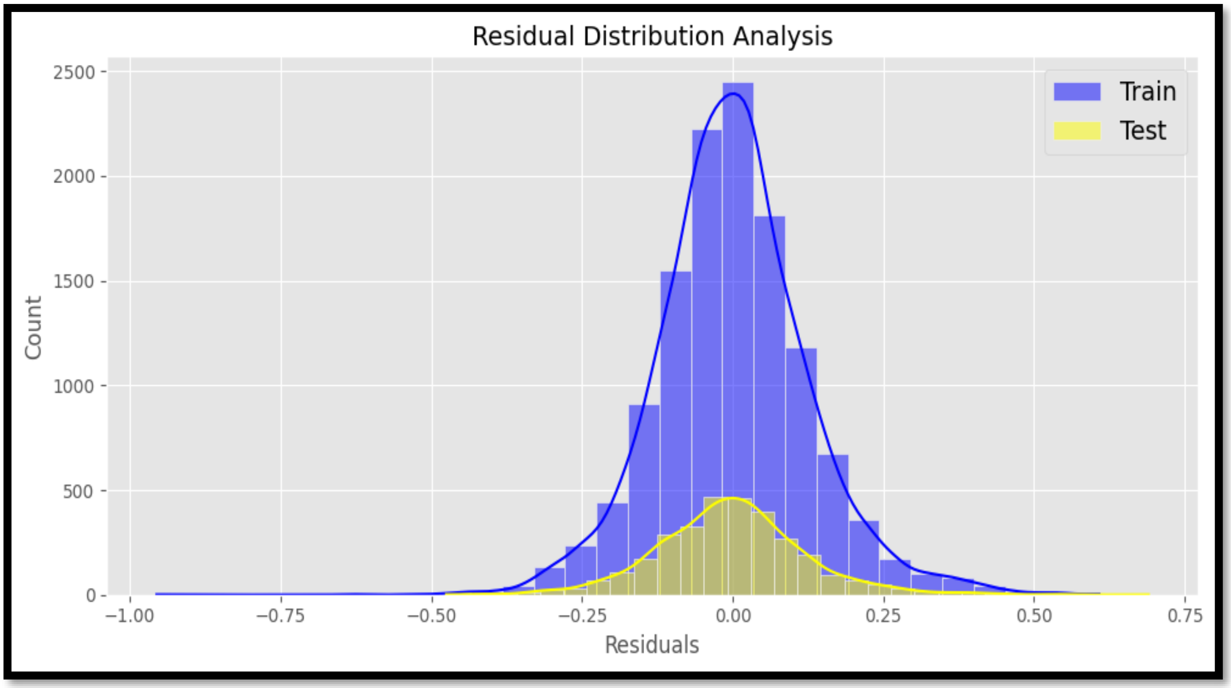
3) **Coefficient Stability**: Binary features (0 or 1) are less likely to be influenced by extreme outliers than count-based approach. This ensures that the coefficients in our final model remain more interpretable while generalizing to unseen data.

3.1.5. Residual Analysis: Linearity and Distribution Analysis

Following our baseline linear regression model, a residual analysis was performed to verify our baseline linear regression model satisfies the core assumptions of linearity, normality of errors and homoscedasticity.



Linearity and Homoscedasticity: Our visualizations indicate that the assumption of linearity is reasonably satisfied. Meaning our model successfully captures the linear relationship within the data. The spread of residuals remains constant across the predicted price range. This satisfies the assumptions of homoscedasticity.



Normality of Errors: The residual distribution analysis for both our training and testing sets are consistent, they both follow a near-normal bell curve centering almost perfectly at zero. This reinforces our initial belief that our model has generalized pretty well and is reliable for making real-world predictions. Fulfils requirement of linear regression models.

Outliers: There are a small number of data points which aren't predicted accurately, suggesting that these may represent high-end cars or unique market conditions which were not captured in our dataset. This also suggests that more data would definitely help our model learn better.

3.1.6. Multicollinearity Analysis using Variance Inflation Factor (VIF)

The final step is validating our feature set for the selected baseline linear regression model; this involves Multicollinearity Analysis using the Variance Inflation Factor (VIF). This diagnosis is to ensure that the independent variables are truly independent and not highly correlated with each other.

Mild to High Multicollinearity: Our analysis revealed that several features had mild to high VIF scores, suggesting correlation between features which could lead to unstable coefficient estimates.

Since our primary objective is predictive accuracy rather than isolated coefficient interpretability, no feature was removed manually to handle multicollinearity.

Need for Regularization: The presence of correlated predictors is our key for transitioning to Regularized Regression models like Lasso and Ridge regression. Ridge regression is specifically designed to mitigate multicollinearity through coefficient shrinkage. This kind of approach is crucial for the model to retain all predictors while control variance, improving stability and generalization for our automobile price predictions.

Multicollinearity was intentionally addressed through regularization rather than feature elimination to preserve predictive signal and market structure.

3.2. L2 Regularization: Ridge Regression

Our multicollinearity analysis moderate to high multicollinearity among several predictors, in our baseline linear regression model, so we extended to a regularized framework using Ridge Regression also known as L2 Regularization.

Ridge Regression (L2 Regularization) adds a squared magnitude penalty to the loss function, which shrinks the coefficients of redundant features toward zero without completely eliminating them.

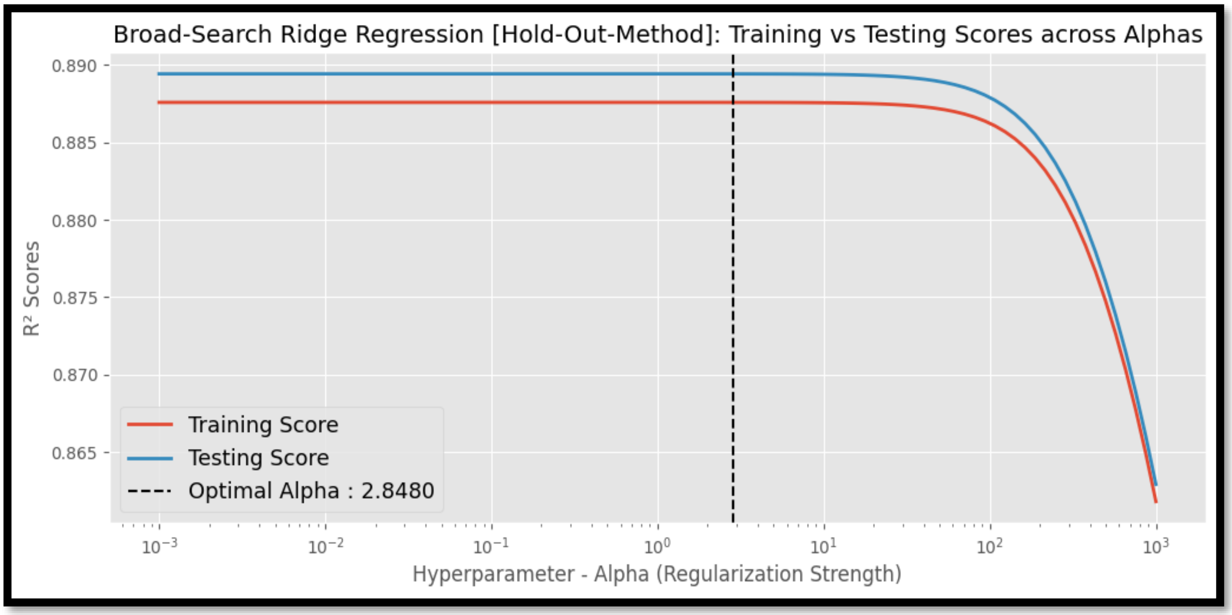
3.2.1. Ridge Regression: 5-Fold-Cross-Validation [Based on R^2 Score]

A broad logarithmic search was conducted to identify the ideal regularization strength (alpha). The optimal value of $(\boldsymbol{a})$ was selected using 5-fold cross-validation. This was based on $\boldsymbol{R^2}$ Scoring method.

This approach ensures that the regularization strength balance bias and variance along with maximizing generalization performance on unseen data.

The optimal alpha **($\boldsymbol{a} \approx$ 2.85)** corresponds to the regions where our model's testing performance is maximized while maintaining a small and stable gap between training and testing scores.

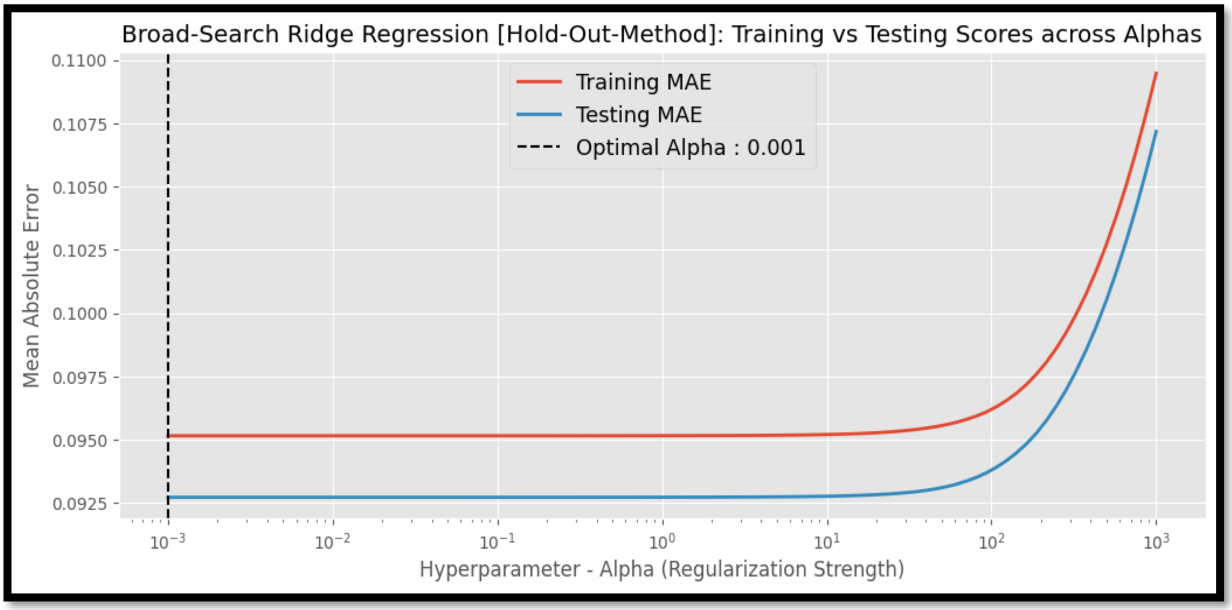
This kind of behaviour validates our initial belief that Ridge Regularization is an effective and essential method for controlling multicollinearity without sacrificing predictive accuracy.



3.2.2. Ridge Regression: 5-Fold-Cross-Validation [Based on MAE]

A broad logarithmic search was conducted to identify the ideal regularization strength (alpha). The optimal value of $(\boldsymbol{a})$ was selected using 5-fold cross-validation. This was based on Mean Absolute Error (MAE) Scoring method.

This kind of analysis was performed to focus on minimizing average absolute prediction error, which is relevant in prediction real-world used



car prices. MAE-based cross-validation indicates that minimal regularization achieves the lowest prediction error, which confirms L2 Regularization primarily serves to stabilize multicollinearity.

Convergence: The selected alpha converged to the lower boundary (≈ 0.001) which suggests that the model performs best with minimal regularization, closely approximating to our baseline linear regression model which maintains some benefits of L2 penalty**.**

3.2.3. Final Model Selection: L2 Regularization [Based on R^2 Score]

After conducting 5-fold cross-validation based on both $\boldsymbol{R^2}$ and **MAE**. The models were evaluated by transforming predicted log-listed prices back into actual real-world price to assess and evaluate their practical accuracy.

Although they both yield near identical prediction error with negligible difference, the $\boldsymbol{R^2}$ tuned model is our ideal model for Ridge Regression, since it has a stronger and active regularization and is better aligned with the objective of Ridge Regression.

$\boldsymbol{R^2}$ Score: It is designed specifically to minimize squared errors which L2 Regularization cares about. It has a more active regularization, making it less sensitive to extreme values and less likely to overreact to noise or unseen data. We are optimizing the same error structure the Ridge Model cares about.

Mean Absolute Error: It measures the average overall error by treating all errors equally, which is often considered ideal for used car prices in a real-world scenario. However, L2 Regularization is designed for squared-error optimization rather than Mean Absolute Error minimization.

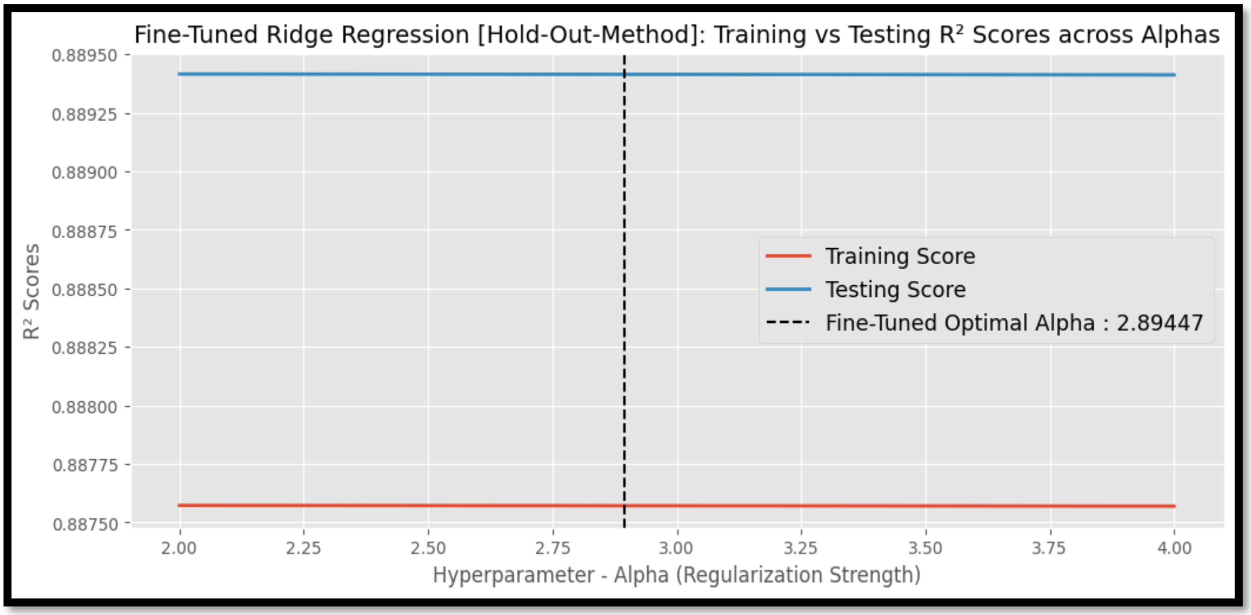
3.2.4. Hyperparameter Tuning: L2 Regularization [Based on R^2 Score]

Our Fine-Tuned Ridge model with the help of 5-Fold-Cross Validation technique ensures consistent performance and improved generalization.

Our initial broad search gave us an estimate for regularization strength ; fine-tuning was performed to pinpoint the more optimal alpha for maximum predictive accuracy and better model generalization.

The 5-Fold-Cross Validation had successfully identified a high precision alpha. Our fine-tuning revealed that the optimal alpha is **($\boldsymbol{a}$ ≈ 2.894).** The model achieved a consistent 0.85 on actual listed price across both our training and testing sets. The RMSE remains stable with little variations.

Our visualizations confirm a zoomed-in view of alpha, where small changes in regularization penalty no longer cause fluctuations in performance, which ensures reliable and more accurate predictions. This fine-tuned Ridge Regression model is ideal since it is stable and well-generalized.



3.3. L1 Regularization: Lasso Regression

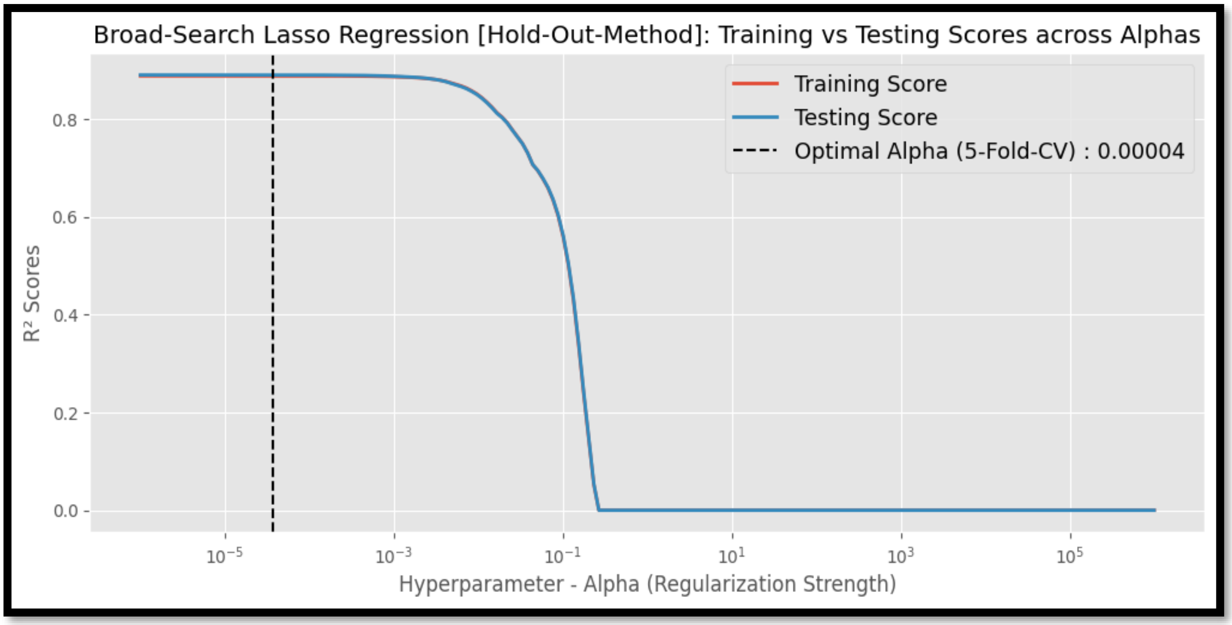
Lasso Regression was evaluated as an alternative regularization technique to examine whether automatic feature selection via coefficient shrinkage to zero could improve model performance and interpretability.

3.3.1. Lasso Regression: 5-Fold-Cross-Validation [Based on MSE]

A broad-search cross validation was conducted to identify the optimal regularization strength. Consistent performance between training and testing scores suggests that the L1 penalty successfully stabilized the model against noise from high-VIF features.

Our visualizations indicate that has the regularization strength increases beyond a certain range, our model performance drops sharply, eventually collapsing. The sharp performance drop which can be observed from our visualization indicates a key limitation of Lasso.

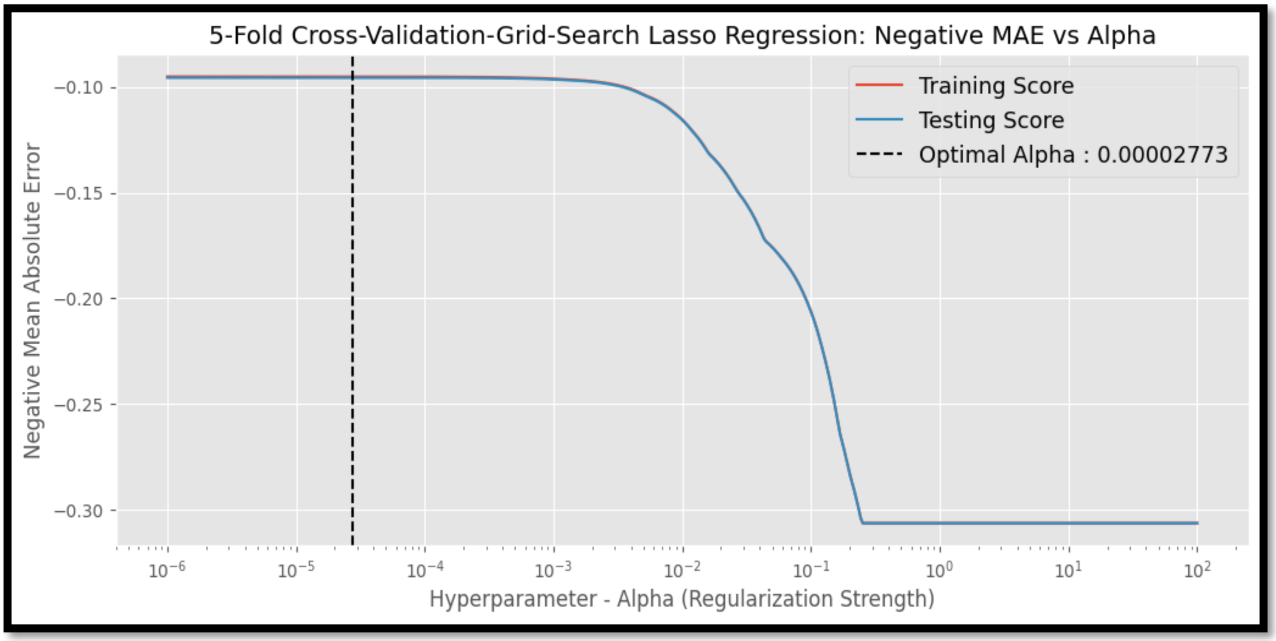
Lasso regression converges to near-zero regularization, this indicates that feature sparsity is not advantageous and confirming Ridge Regression as the more suitable regularization strategy for this dataset.



3.3.2. Lasso Regression: 5-Fold-Cross-Validation [Based on MAE]

This was implemented using Grid Search Cross Validation technique. The goal is to maximize negative MAE, which is equivalent to minimizing average absolute prediction error on the log-transformed listed price.

It took around 47 iterations for convergence, the model achieved its lowest MAE, indicating that only negligible L1 regularization is beneficial.



Our visualizations indicated that stronger L1 penalties rapidly degrade model performance by eliminating informative predictors.

Mean Absolute Error (MAE) treats all errors equally, providing a more direct interpretation of average price deviation in real-world automobile pricing. The model achieved a Best Negative MAE of -0.0956.\

MAE-optimized grid search confirms that Lasso converges to near-zero regularization, indicating that coefficient sparsity is not beneficial and validating Ridge Regression as the final model choice.

3.3.3. Final Model Selection: L1 Regularization [Based on MAE]

The absolute difference in average error between the two models is only EUR 0.10, which is far below our domain-driven negligible threshold of EUR 5.00.

MAE is the preferable choice because it maintains consistent average accuracy. More robust to outliers and provide more stable pricing errors for standard vehicles.

Comparative Performance Analysis between MSE and MAE:-

MSE (Mean Squared Error): Highly sensitive to extreme values and outliers. It penalizes large deviations more heavily, attempting to eliminate expensive mistakes at the potential cost of overall average accuracy.

MAE (Mean Absolute Error): Measures the average overall error by treating all deviations equally. It is considered ideal for maintaining consistent, stable pricing across a wide range of standard vehicle listings.

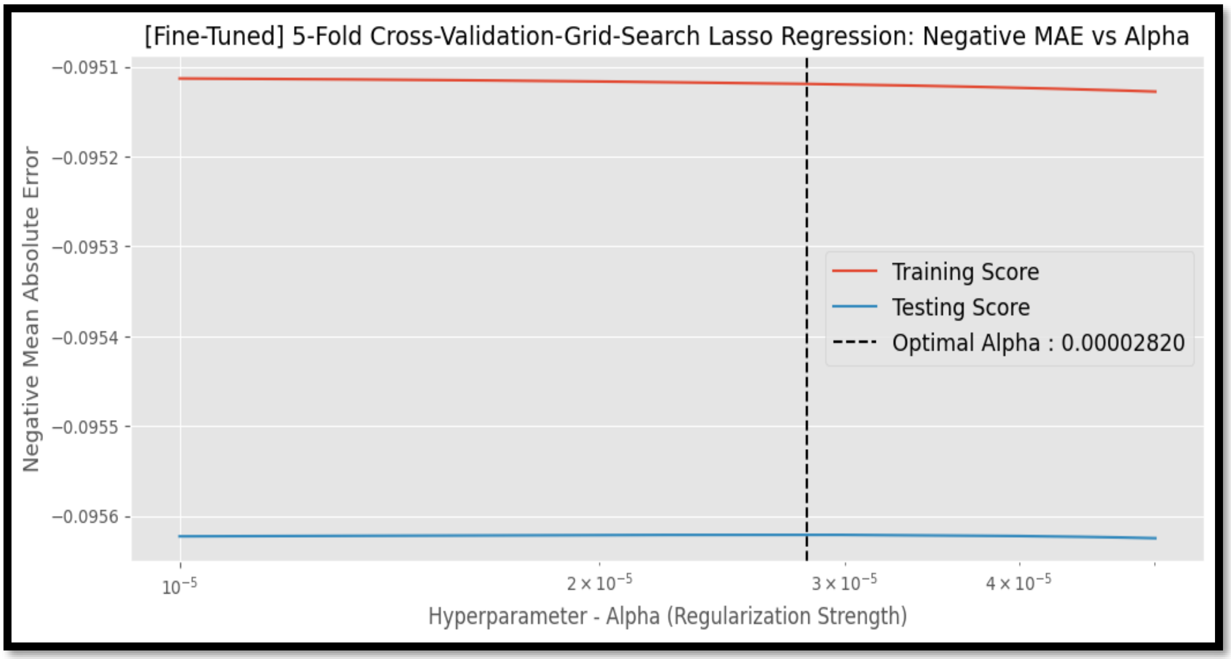
3.3.4. Hyperparameter Tuning: L1 Regularization [Based on MAE]

Our Fine-Tuned Lasso model with the help of 5-Fold- Grid Search Cross Validation technique ensures consistent performance and improved generalization.

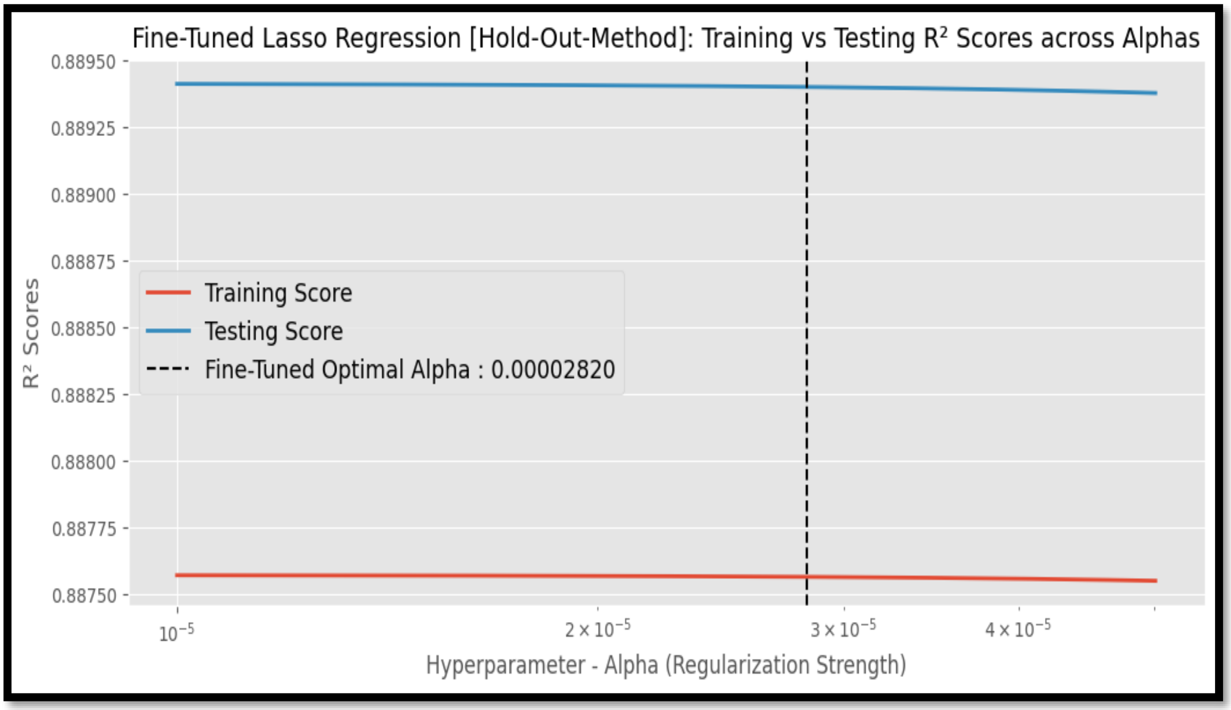
Our initial broad search gave us an estimate for regularization strength ; fine-tuning was performed to pinpoint the more optimal alpha for maximum predictive accuracy and better model generalization.

The model aimed to maximize Negative MAE, which is identical to minimizing the average absolute error of pricing.

Negative MAE vs. Alpha: The Training and Testing error curves remains almost parallel and flat; this proves that the model's absolute accuracy is not sensitive to minor fluctuations within our optimal range of alpha region.



***R^2* Scores across Alphas**: The R^2 scores indicate that even as the L1 penalty slightly increases, the model's ability to explain vehicle value remains maximized.



3.4. Regularisation Comparison & Analysis

A comparative analysis between the 3 well-established linear regression models were conducted side-by-side based on evaluation metrics on actual listed price (EUR) to determine the ultimate pricing model for the automobile industry.

3.4.1. Comparing Evaluation Metrics [Winner – Ridge Regression]

Our evaluation metrics make it clear that both regularization techniques significantly stabilize the model compared to the baseline, although the absolute performance differences between them are minimal.

Statistical performance of evaluation metrics is near identical so the choice between these models depends on three important factors model complexity, interpretability and multicollinearity.

Unlike standard Baseline Linear Regression, **Ridge is specialized in addressing multicollinearity**, which was previously identified among several correlated predictors. The primary advantage is feature selection.

Ridge Regression is selected as the final model, not because of its evaluation metrics but because it provides better model stability and consistency.

Unlike Lasso Regression, **Ridge Regression is designed to maintain all the predictors** while shrinking their coefficients. This preserves information which could be an advantage to our dataset.

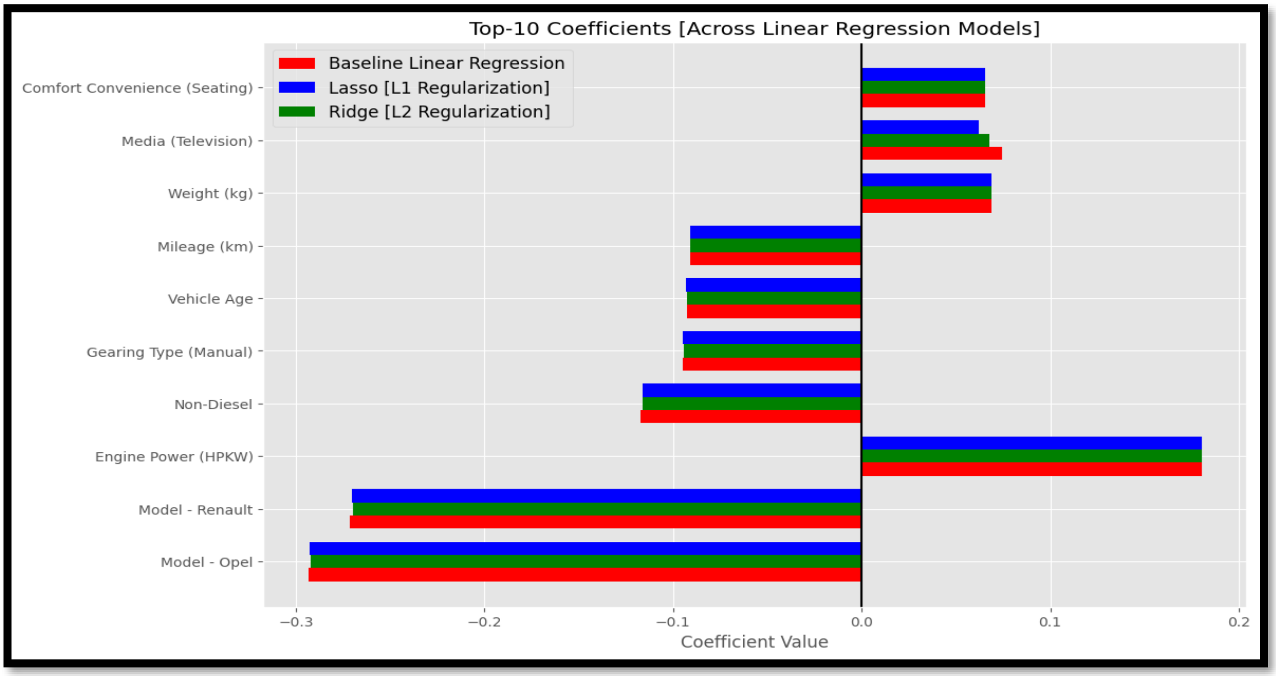
3.4.2. Coefficient Analysis: Top 10 Highest Coefficients [Across all Models]

The estimate coefficients across Baseline Linear Regression, Lasso (L1 Regularization) and Ridge (L2 Regularization) all are consistent in both magnitude and direction. The core contributing features for a vehicle price are the vehicle model, engine power and fuel type.

The near-identical coefficient estimates confirm that multicollinearity is present in our dataset but it’s not severe enough to distort our estimates and that regularization plays a crucial role in stability and feature importance. The Top 10 feature provide a statistically unique signal.

Engine Power is our strongest physical predictor of vehicle price. The most aggressive negative influence on price is related to the vehicle brand. Our Lasso model did not eliminate any feature despite of high multicollinearity.

***Note:** Lasso (L1 Regularization) did not eliminate any feature, this indicates that none of the predictors were redundant enough to be shrunk to zero.*



4. Conclusions and Key Takeaway

Implementing a Baseline Linear Regression Model and subsequent experimental analysis and regularization techniques have provided us with key insights into the dynamics of AutoScout Vehicle pricing dataset.

4.1. Outcomes and Insights

1. **Effect of Regularization Techniques:** Ridge and Lasso played an important role in feature importance, stabilizing the model, controlling the coefficient rather than dramatic performance improvement (gains).

2. **No Evidence of Overfitting:** Model generalized really well, with no meaningful overfitting was observed. Training and Testing metrics were closely aligned and remained almost consistent across all models.

3. **Feature engineering and Outlier handling:** Played an important role for minimal performance gain and minimal variations across models. Target transformation, handling class imbalances, Upper-tail Winsorization and feature encoding played a crucial role.

4. **Sufficient Data:** With over 15,000 vehicle listings, the sampled web-scraped data was sufficient enough to achieve statistical significance and predictive accuracy for predicting used car prices.

5. **Linear Model:** A linear model is sufficient for this dataset and the scope of our analysis, with an 85% of explained variance in vehicle prices across all models, while satisfying the core assumptions of linearity.

6. **Random Forests / Gradient Boosting: (Interpretability vs. Complexity Trade-off) -** These models might be able to capture the remaining variance, but a linear model provides a higher level of interpretability and pricing transparency which is invaluable to business stakeholders.

7. **Presence over Quantity:** The presence of a feature is more interpretable, and a consumer pays for the existence of a feature rather than a specific number of components of that feature.

8. **Presence of Managed Multicollinearity:** Variance Inflation Factor (VIF) analysis revealed moderate to high multicollinearity among several predictors which were feature engineered. Regularization methods were able to manage redundancy, validating the decision to not manually remove features.

9. **Diminishing Returns of Regularization:** Ridge and Lasso Regression both improved coefficient stability, fine-tuning the regularization strength yielded marginal performance improvements. This could indicate that our features were already well-engineered and further regularization primarily only improved model stability and handled multicollinearity rather than true predictive power.

10. **Residual Diagnostics and Model Assumptions:** Residual analysis confirmed that our linear assumptions were satisfied for majority of our data points. Residuals followed a near-normal distribution on both training and testing sets. This validates that linear regression was suitable for the scope of our business problem.

11. **Symmetric Depreciation (Time vs Wear):** The depreciation factor for Vehicle Age and Mileage carried nearly identical weights. The coefficient analysis revealed that buyers view the vehicles age (time passed) and distance travelled (physical wear) as equivalent risk, penalizing them equally.

12. **Strongest Positive Driver:** Engine Power has the strongest positive coefficients across all models, indicating that a higher power vehicle commands a higher price.

13. **Strongest Negative Driver:** Vehicle models like Opel and Renault have strong negative coefficients compared to Audi. This captures brand-specific price depreciation.