

Regularisation Report

1. Data Understanding

1.1 Dataset Description

The dataset was sourced from AutoScout, a German online car trading platform, and contains used vehicle listings where each row represents a single car. It consists of **15,915 observations and 23 variables**, covering attributes such as price, mileage, engine specifications, ownership history, fuel type, transmission, and drivetrain.

The dataset also includes bundled textual columns describing comfort, entertainment, safety, and additional features, which require preprocessing before modeling.

1.2 Data Loading and Structure

The dataset was loaded using Pandas, and an initial inspection was conducted to understand its structure and data types. It contains a mix of **10 numerical features** (e.g., mileage, age, engine power, displacement, vehicle weight, fuel consumption, gears, previous owners, and inspection status) and **13 categorical features** describing vehicle characteristics such as make and model, body type, fuel type, transmission, paint type, upholstery, and drivetrain.

1.3 Missing Value Analysis

A missing value check across all columns revealed **no missing values** in the dataset. Therefore, no imputation or row removal was required, and all records were retained for further analysis.

1.4 Initial Observations

The dataset is clean and complete, simplifying preprocessing. However, the presence of multiple categorical variables and bundled feature columns necessitates careful feature engineering. Additionally, several engine-related numerical features are expected to be correlated, motivating multicollinearity analysis and the use of regularised regression techniques in later stages.

2. Analysis and Feature Engineering

2.1 Preliminary Analysis and Frequency Distributions

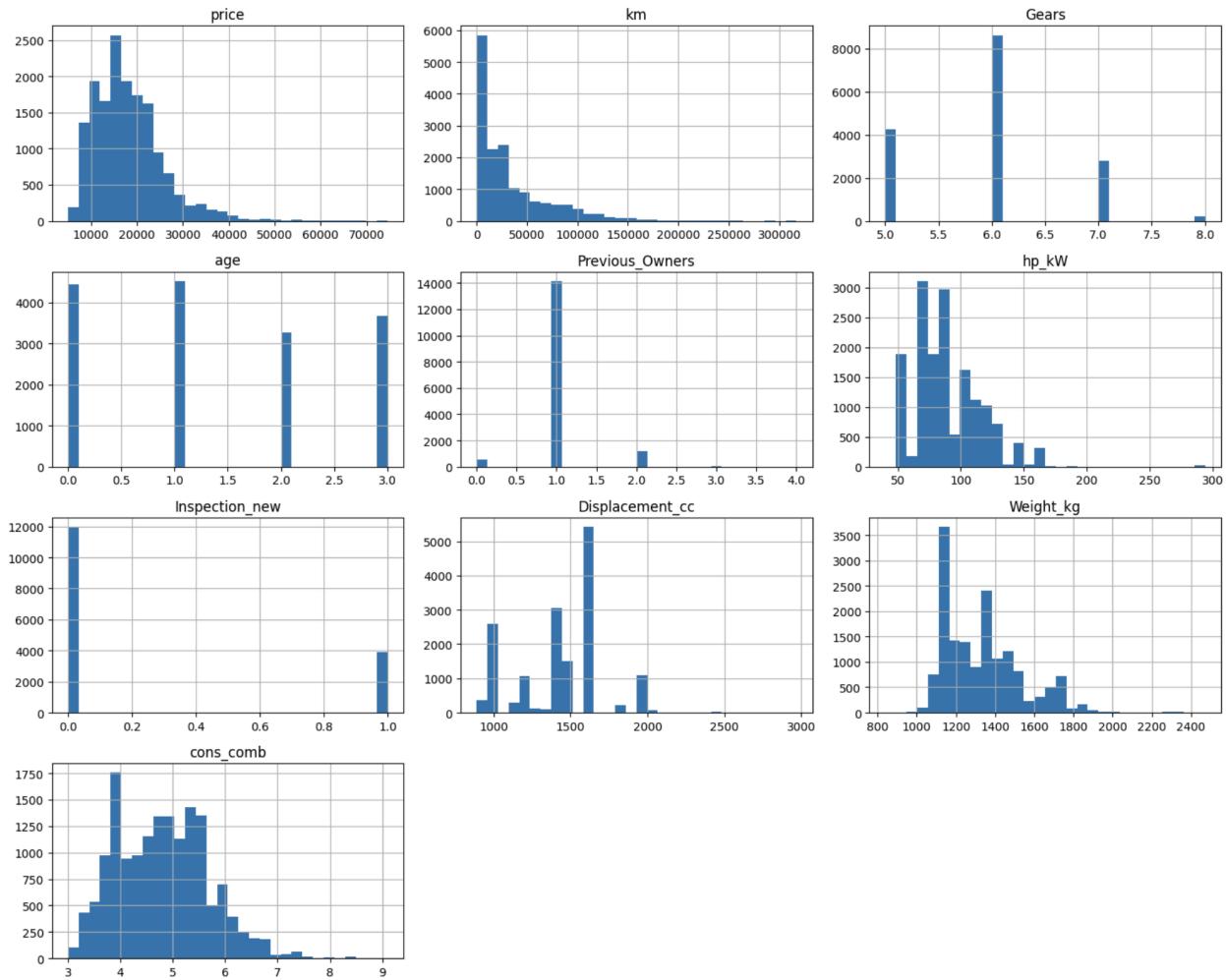
2.1.1 Missing Value Check

A missing value check showed that no columns contain null values. Therefore, no missing value treatment was required, and all records were retained for analysis.

2.1.2 Numerical Predictors and Frequency Distributions

The numerical features were analysed using summary statistics and frequency distributions. Variables such as mileage, engine power, engine displacement, and vehicle weight show wide ranges and right-skewed distributions, indicating the presence of extreme values for a smaller subset of vehicles.

In contrast, features such as age and number of previous owners exhibit discrete and concentrated distributions, while variables like number of gears and inspection status show limited variation. These characteristics are considered in later preprocessing and modeling steps.



	count	mean	std	min	25%	50%	75%	max
price	15915.0	18024.380584	7381.679318	4950.0	12850.0	16900.0	21900.0	74600.0
km	15915.0	32089.995708	36977.214964	0.0	1920.5	20413.0	46900.0	317000.0
Gears	15915.0	5.937355	0.704772	5.0	5.0	6.0	6.0	8.0
age	15915.0	1.389695	1.121306	0.0	0.0	1.0	2.0	3.0
Previous_Owners	15915.0	1.042853	0.339178	0.0	1.0	1.0	1.0	4.0
hp_kW	15915.0	88.499340	26.674341	40.0	66.0	85.0	103.0	294.0
Inspection_new	15915.0	0.247063	0.431317	0.0	0.0	0.0	0.0	1.0
Displacement_cc	15915.0	1428.661891	275.804272	890.0	1229.0	1461.0	1598.0	2967.0
Weight_kg	15915.0	1337.700534	199.682385	840.0	1165.0	1295.0	1472.0	2471.0
cons_comb	15915.0	4.832124	0.867530	3.0	4.1	4.8	5.4	9.1

2.1.3 Categorical Predictors and Frequency Distributions

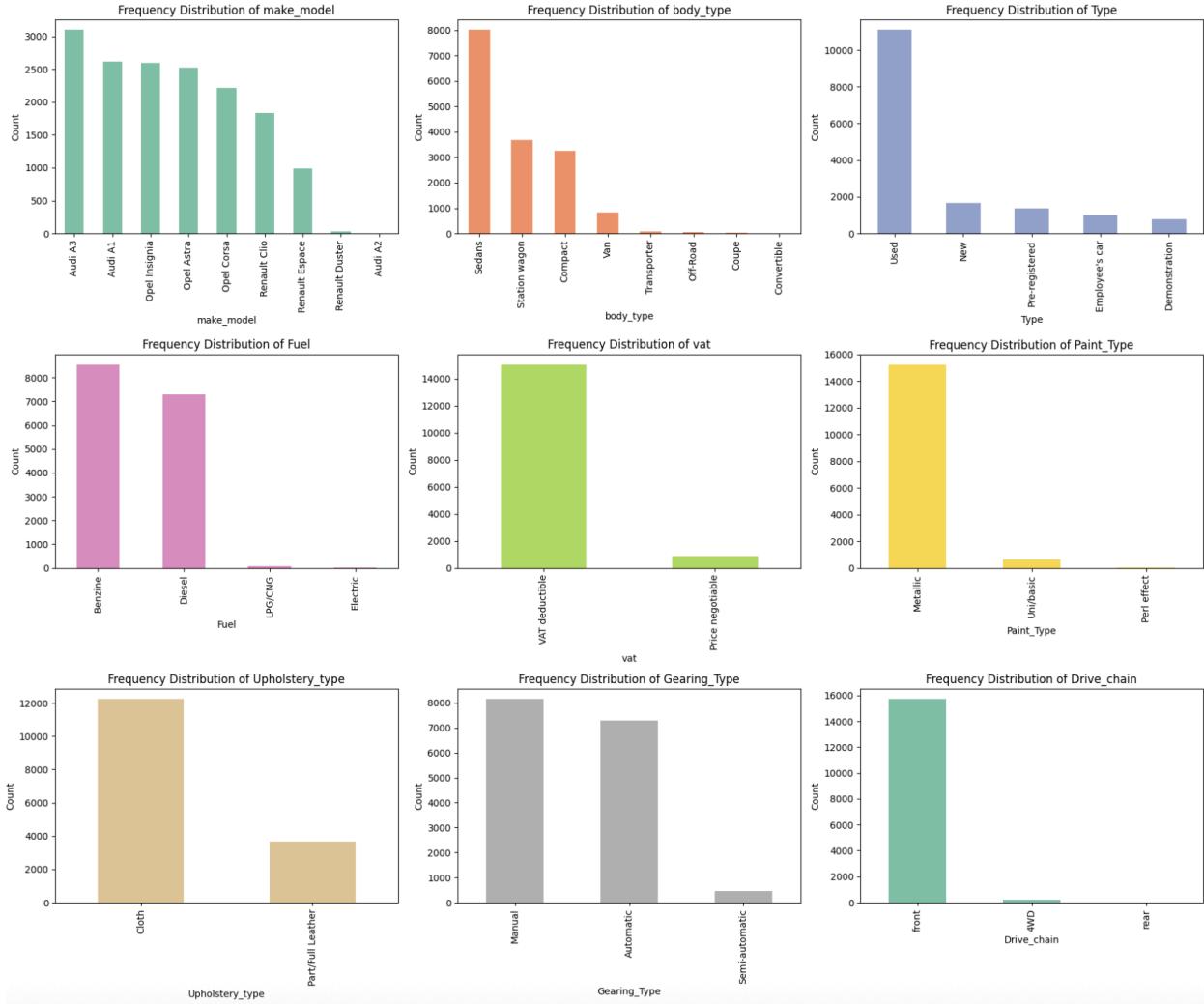
Categorical variables were analysed using frequency distributions to understand category dominance and class imbalance. Variables such as `make_model`, `body_type`, `Type`, `Fuel`, `vat`, `Paint_Type`, `Upholstery_type`, `Gearing_Type`, and `Drive_chain` exhibit low to moderate cardinality and were treated as standard categorical predictors.

The distribution of `make_model` shows that Audi A3, Audi A1, Opel Insignia, and Opel Astra collectively account for a large proportion of the dataset, while models such as Audi A2 and Renault Duster have very low representation. Body types are dominated by Sedans, followed by Station Wagons and Compact cars, indicating a strong preference for these vehicle types in the dataset.

Most vehicles are listed as Used (approximately 70%), with smaller proportions of New, Pre-registered, Employee's cars, and Demonstration vehicles. Fuel type is largely split between Benzine and Diesel, while alternative fuels such as LPG/CNG and Electric vehicles have negligible representation.

The majority of vehicles are VAT deductible, use Metallic paint, and have Cloth upholstery. Transmission type is relatively balanced between Manual and Automatic, while Semi-automatic vehicles are rare. Drivetrain distribution is highly skewed toward front-wheel drive, with very limited representation of rear-wheel and four-wheel drive vehicles.

Overall, several categorical predictors exhibit strong class imbalance, which is considered during subsequent preprocessing and modeling steps to ensure model stability and generalisation.



2.1.4 Handling Low-Frequency Values and Class Imbalance

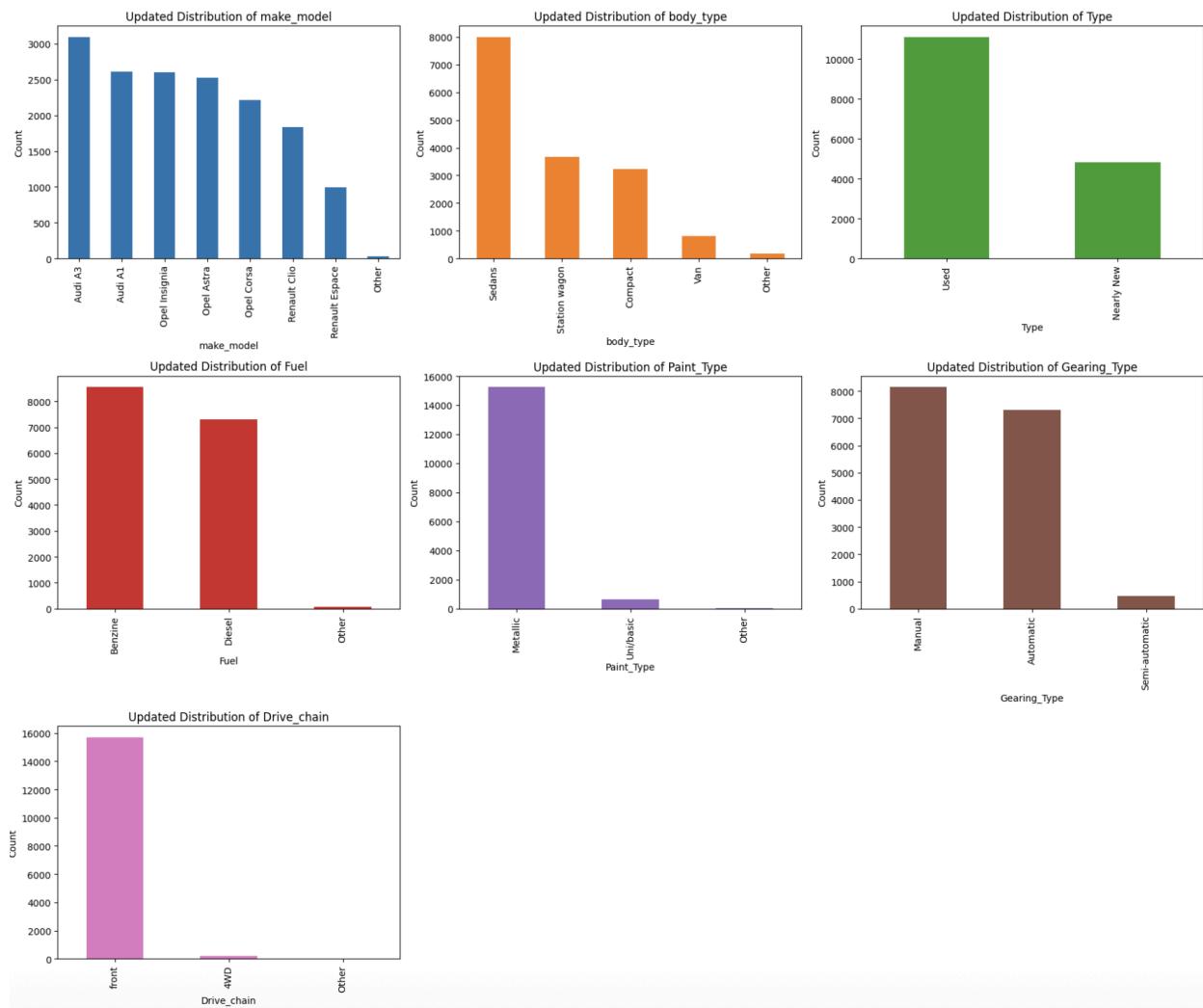
To ensure model stability and reduce noise caused by extremely rare categories, categorical variables were examined for class imbalance using frequency distributions.

The **Type** variable was handled first using the provided business context. Categories such as New, Pre-registered, Employee's car, and Demonstration represent vehicles that are lightly used or new-like in condition. These categories were consolidated into a single class named **Nearly New**, while vehicles classified as Used were retained as a separate category.

This consolidation resulted in two meaningful and interpretable groups: **Used** and **Nearly New**, reducing sparsity while preserving pricing relevance.

Next, other categorical variables were examined for extreme class imbalance. Categories contributing less than 1% of the total observations were grouped under an **Other** category. This

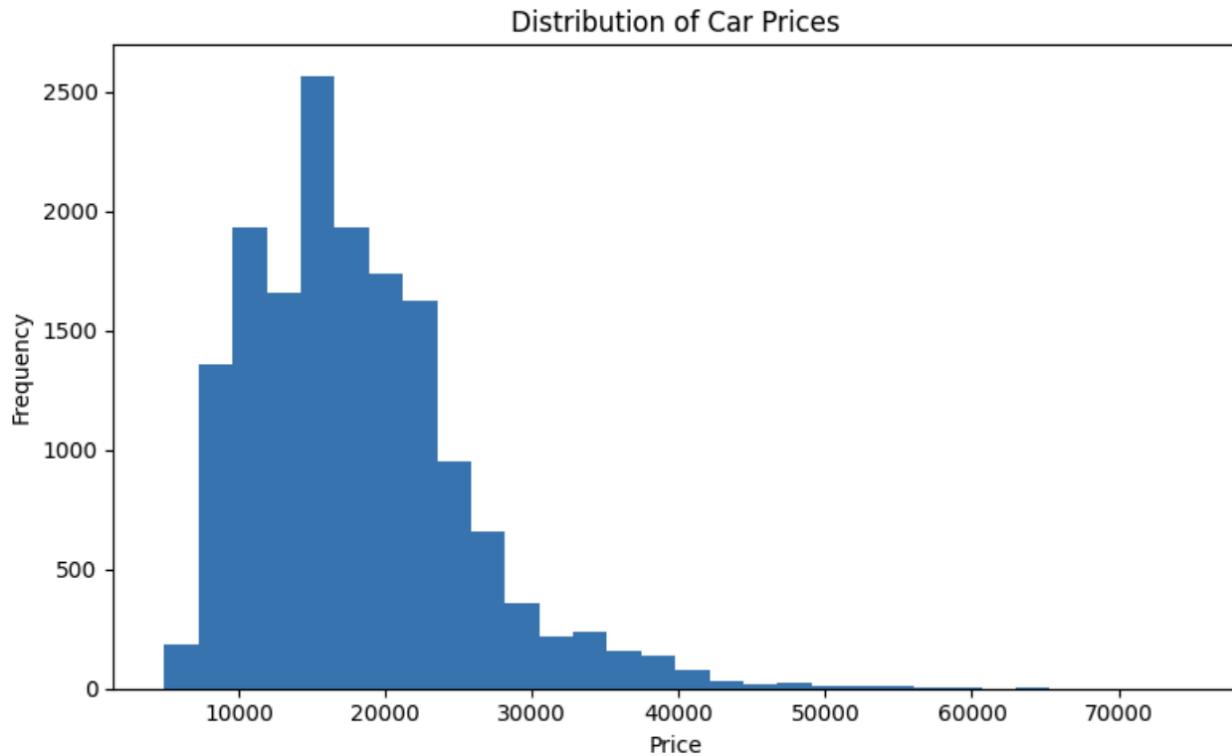
consolidation was applied selectively to variables such as `make_model`, `body_type`, `Fuel`, `Paint_Type`, `Gearing_Type`, and `Drive_chain`.



After consolidation, dominant categories such as major car models, common body types, Benzine and Diesel fuel types, Metallic paint, Manual and Automatic transmissions, and front-wheel drive remained unchanged, while extremely rare categories were grouped. This approach reduces dimensionality and noise without removing any observations.

2.1.5 Target Variable Distribution and Transformation

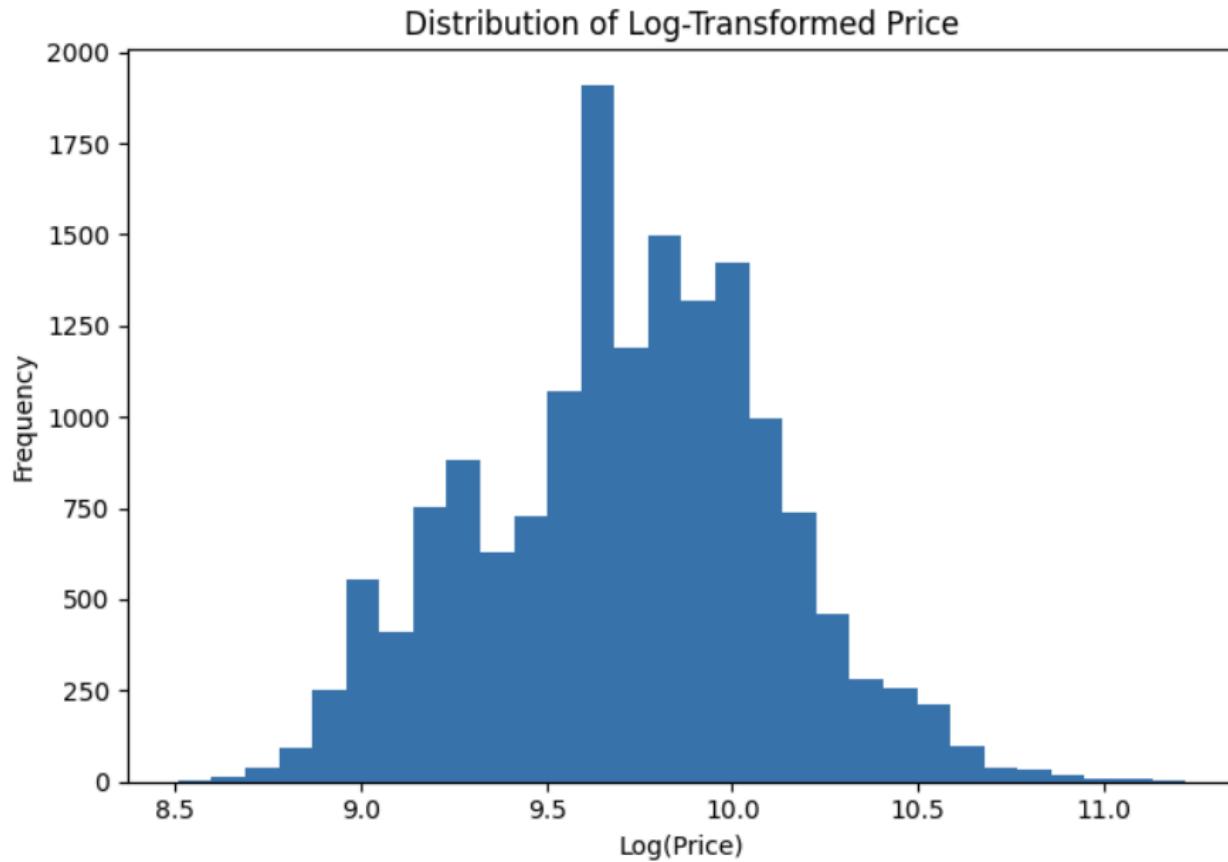
The target variable for this analysis is the listed vehicle **price**. The initial distribution of prices was examined using a histogram.



The distribution shows a clear right-skewed pattern, with most vehicles concentrated in the lower to mid-price range and a small number of high-priced vehicles forming a long right tail. This observation is supported by a skewness value of approximately **1.24**, indicating significant positive skewness.

Such skewness can negatively affect linear regression models, as large price values dominate the loss function and violate assumptions of normality and constant variance.

To address this issue, a **logarithmic transformation** was applied to the price variable.



After transformation, the distribution becomes substantially more symmetric. The skewness of the log-transformed price reduces to approximately **-0.03**, indicating near symmetry. The transformation compresses extreme values and stabilises variance, making the target variable more suitable for regression modeling.

Based on this analysis, the **log-transformed price** is used as the target variable in subsequent modeling steps to improve model performance, interpretability, and generalisation.

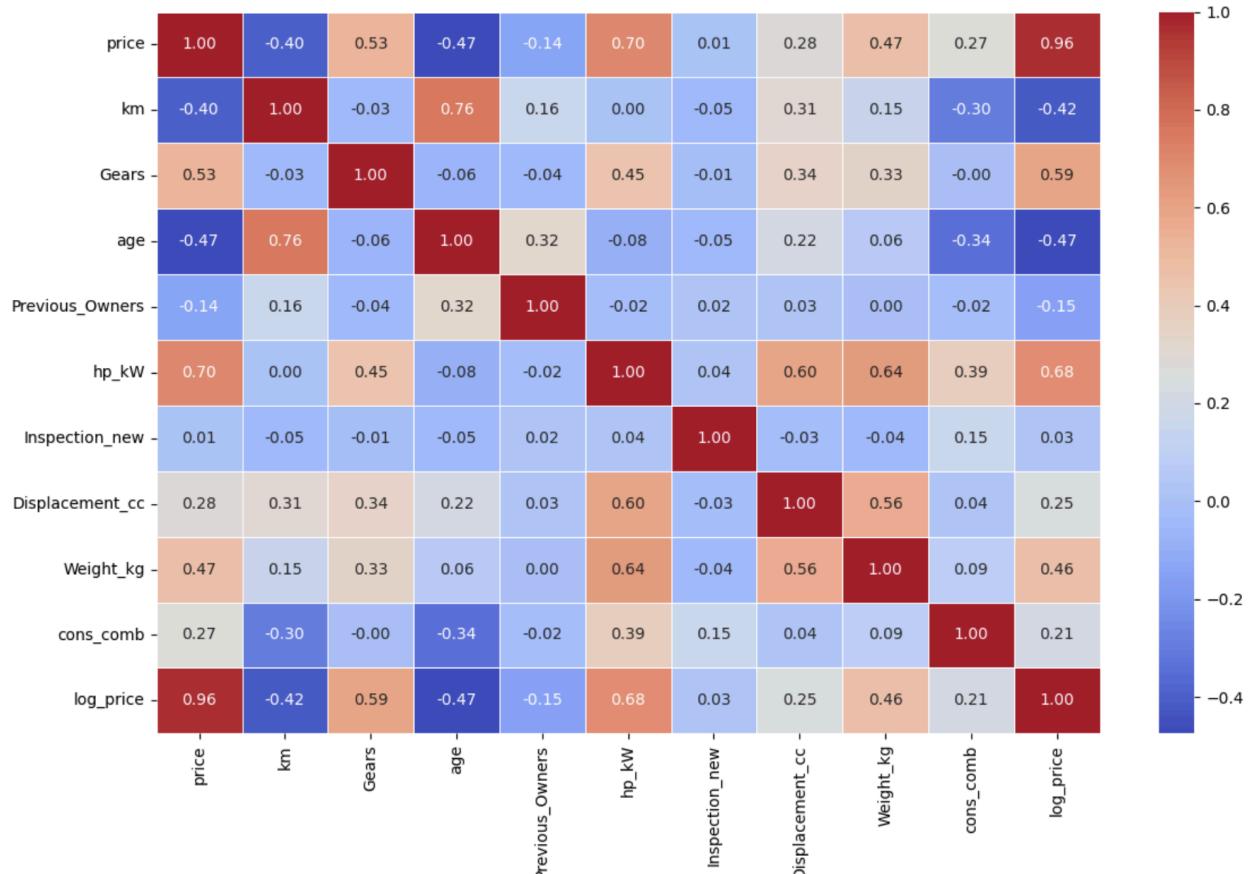
2.2 Correlation Analysis

2.2.1 Correlation Between Numerical Features and Target Variable

The correlation heatmap illustrates the relationships between numerical features and the target variable. Strong positive correlations with price are observed for engine power (**hp_kW**), number of gears, and vehicle weight, indicating that more powerful and heavier vehicles tend to be

priced higher. Mileage (**km**) and vehicle age show moderate to strong negative correlations with price, suggesting depreciation with increased usage and age.

Several numerical features are also highly correlated with each other. For example, engine power shows strong correlations with displacement and vehicle weight, while mileage is strongly correlated with age. These relationships indicate the presence of multicollinearity, which can destabilise coefficient estimates in linear regression and motivates the use of regularised regression methods in later stages.



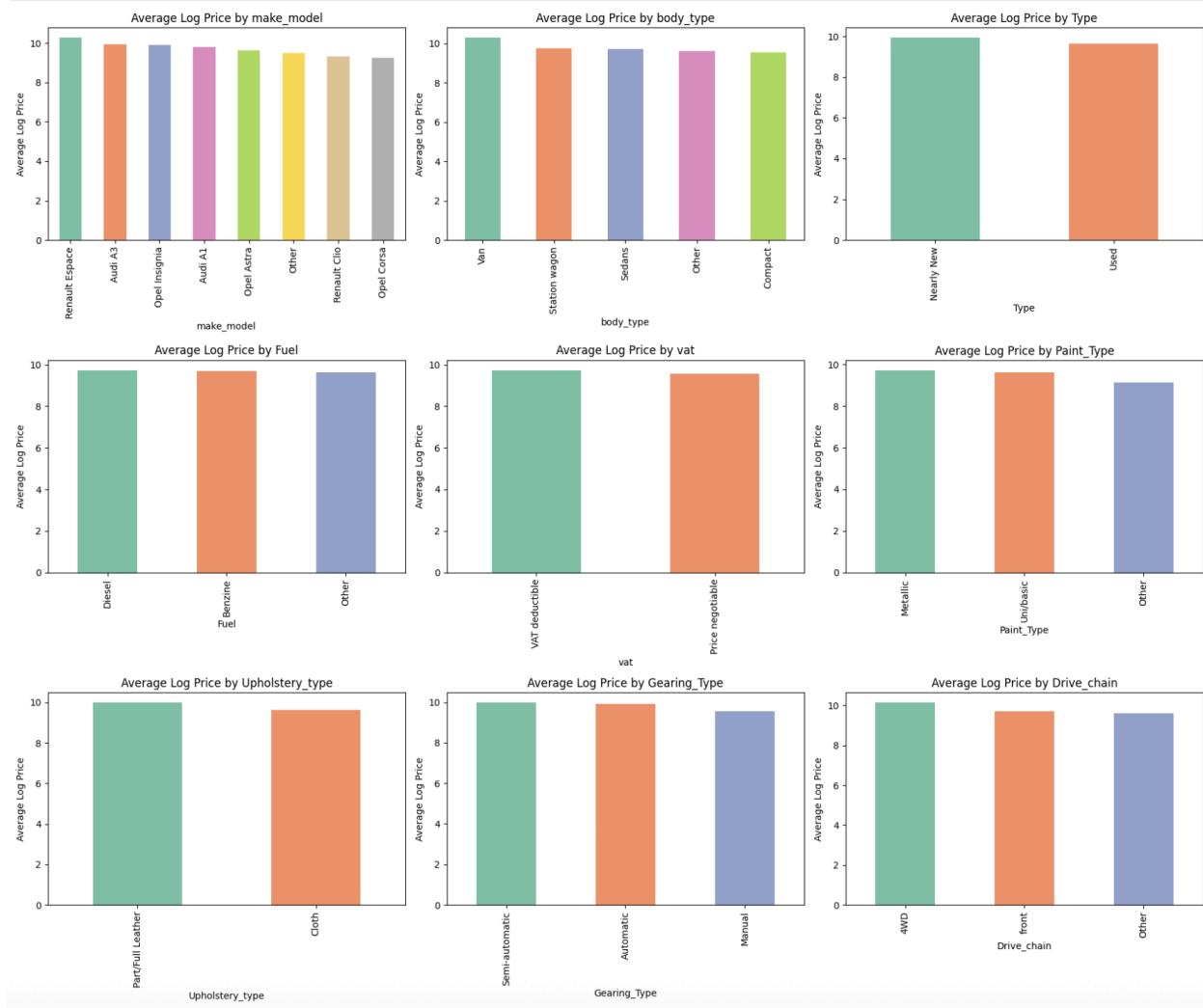
2.2.2 Relationship Between Categorical Features and Target Variable

The relationship between categorical variables and the target variable was analysed by comparing the average log-transformed price across categories. Clear pricing differences are observed across several categorical features. Vehicle type shows a strong effect, with nearly new cars having higher average prices compared to used cars. Transmission type also influences pricing, where automatic and semi-automatic vehicles are priced higher on average than manual ones.

Body type and drivetrain demonstrate noticeable variation in prices, with vans and four-wheel-drive vehicles commanding higher average prices. Upholstery material shows a

meaningful difference, with leather interiors associated with higher prices. Fuel type and VAT status show comparatively smaller price differences, indicating a weaker influence on pricing.

These patterns confirm that categorical features contribute valuable information for price prediction and should be retained in the modeling process.



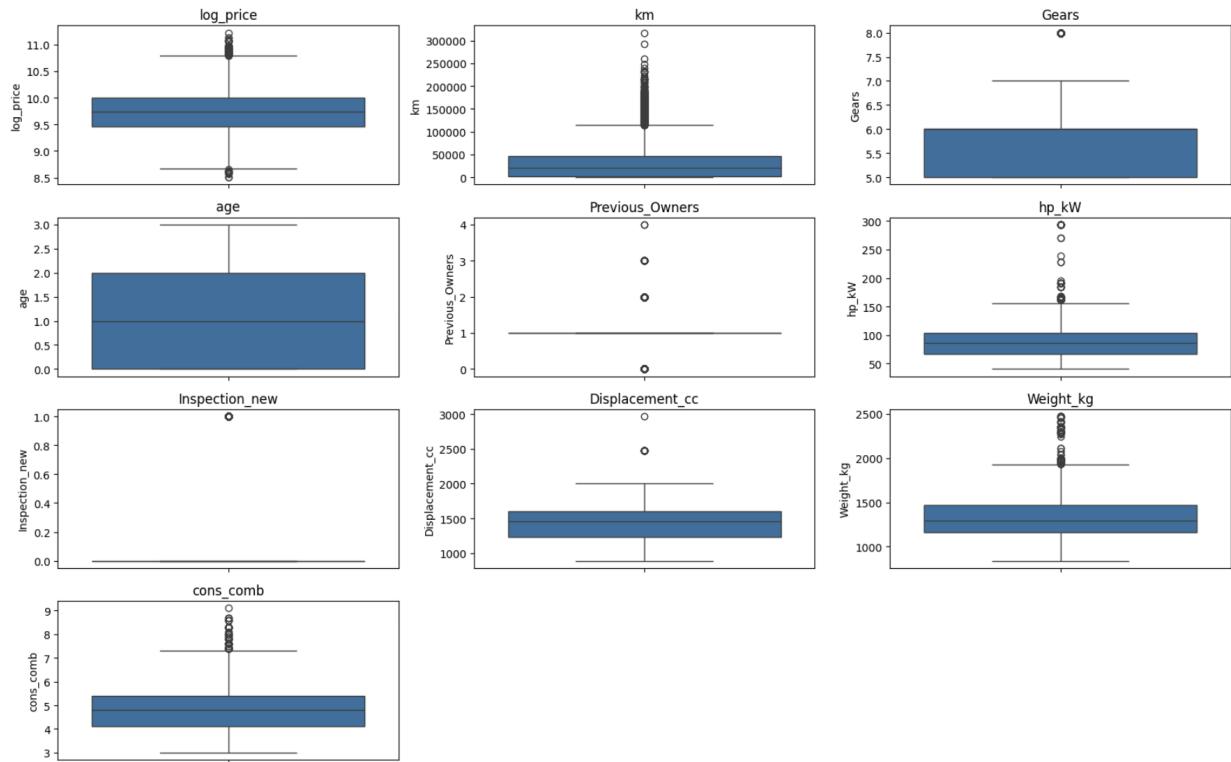
2.3 Outlier Analysis

2.3.1 Identification of Outliers

Outlier analysis was performed on all numerical features, including the log-transformed target variable. Boxplots and IQR-based calculations revealed the presence of extreme values in several continuous variables. Mileage (**km**) shows a strong right tail, indicating the presence of vehicles with very high usage. Engine-related features such as power (**hp_kW**), displacement,

and vehicle weight also exhibit extreme values, which are expected in a heterogeneous used-car dataset.

The log-transformed target variable (`log_price`) still contains a small number of extreme values, representing unusually high or low-priced vehicles. Discrete and binary variables such as number of gears, number of previous owners, inspection status, and vehicle age were also examined. Although IQR-based detection flags some values as statistical outliers, these represent valid real-world cases rather than anomalies.

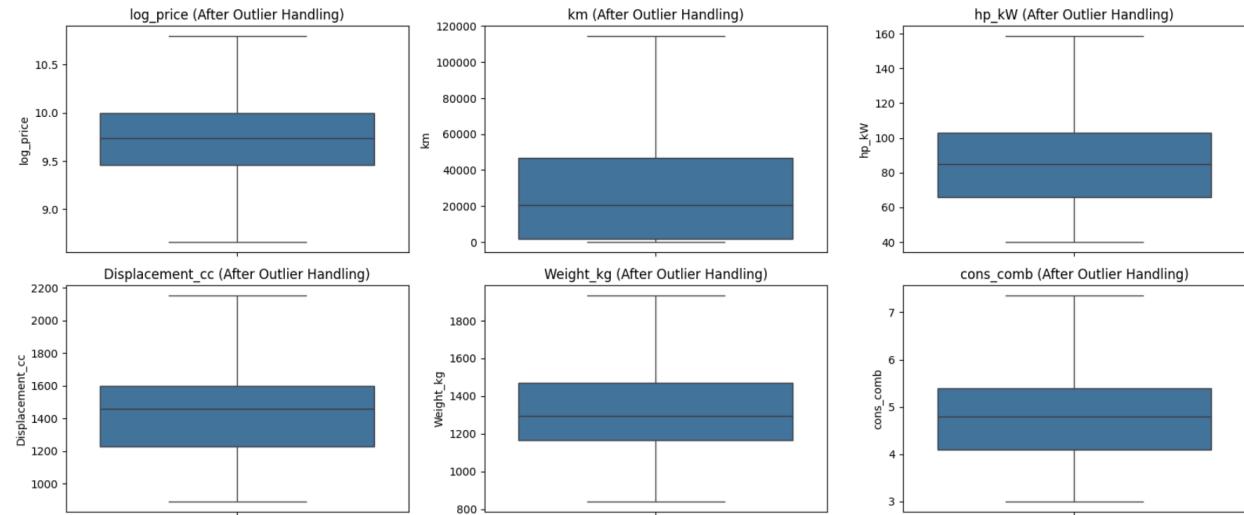


2.3.2 Outlier Handling

Outlier handling was applied selectively to continuous variables where extreme values could disproportionately influence regression estimates. IQR-based capping was used for mileage, engine power, engine displacement, vehicle weight, fuel consumption, and the log-transformed target variable. This approach limits the impact of extreme observations while retaining all records in the dataset.

Discrete and binary variables, including number of gears, number of previous owners, inspection status, and vehicle age, were not modified, as outlier treatment is not statistically meaningful for such variables.

Post-capping summary statistics confirm that the ranges of treated variables were reduced, skewness was moderated, and overall distributions became more stable, making the data better suited for linear and regularised regression models.



2.4 Feature Engineering

2.4.1 Fixing Redundant Columns and Creating New Features

The target variable exists in two forms: the original `price` and the log-transformed `log_price`. Since `log_price` was created to correct skewness and better satisfy linear regression assumptions, it is retained as the modeling target.

Keeping both representations would introduce redundancy and risk data leakage. Therefore, the original `price` column was removed from the modeling dataset.

All remaining columns represent distinct vehicle attributes and do not duplicate information. The bundled specification columns (`Comfort_Convenience`, `Entertainment_Media`, `Extras`, `Safety_Security`) are not redundant but require special handling due to their structure and are addressed separately in the next step.

2.4.2 Feature Engineering for Bundled Specification Columns

The columns `Comfort_Convenience`, `Entertainment_Media`, `Extras`, and `Safety_Security` contain comma-separated lists of features present in each vehicle. An initial inspection revealed extremely high cardinality, with thousands of unique combinations per column.

Because each row represents a combination of multiple features, treating these columns as standard categorical variables or one-hot encoding them directly would lead to a severe explosion in dimensionality without proportional gain in explanatory power.

To address this, each bundled column was transformed into a numerical feature representing the **count of features present** for that category (e.g., total number of comfort features). This approach captures the overall feature richness of a vehicle while keeping the feature space compact and interpretable.

After extracting these counts, the original bundled columns were removed to reduce dimensionality and prevent redundancy.

2.4.3 Feature Encoding

Linear and regularised regression models require all input variables to be numerical. Therefore, categorical variables were converted into numerical form using **one-hot encoding**.

Each categorical feature was transformed into a set of binary indicator variables. To avoid multicollinearity and the dummy variable trap, one category from each feature was dropped using `drop_first=True`.

This encoding preserves category-level information while ensuring compatibility with linear, Ridge, and Lasso regression models.

2.4.4 Train–Test Split

The final model-ready dataset was split into training and testing sets using an **80–20 split**.

Using 80% of the data for training provides sufficient information for the model to learn stable parameter estimates, while the remaining 20% is reserved for unbiased evaluation of model performance.

A fixed `random_state` was used to ensure reproducibility, allowing consistent comparison across different models and regularisation techniques.

2.4.5 Feature Scaling

Feature scaling was performed using **standardisation**, transforming all numerical features to have zero mean and unit variance.

Scaling is essential for regularised regression models such as Ridge and Lasso, as these models are sensitive to the relative magnitude of features when applying penalty terms.

The scaler was fitted only on the training data and then applied to the test data to prevent data leakage and ensure a fair evaluation.

3. Linear Regression Models

3.1 Baseline Linear Regression Model

3.1.1 Model Building and Evaluation

A baseline linear regression model was trained using the scaled features and log-transformed price as the target variable.

The model demonstrates **strong predictive performance**, achieving an **R² of 0.92 on training data and 0.92 on test data**, indicating that a large proportion of price variance is explained.

Both **RMSE and MAE values are low and consistent across training and testing sets**, suggesting good generalisation and no evidence of overfitting.

This baseline model provides a reliable benchmark for evaluating the impact of Ridge and Lasso regularisation in subsequent sections.

3.1.2 Residual Analysis and Linear Regression Assumptions

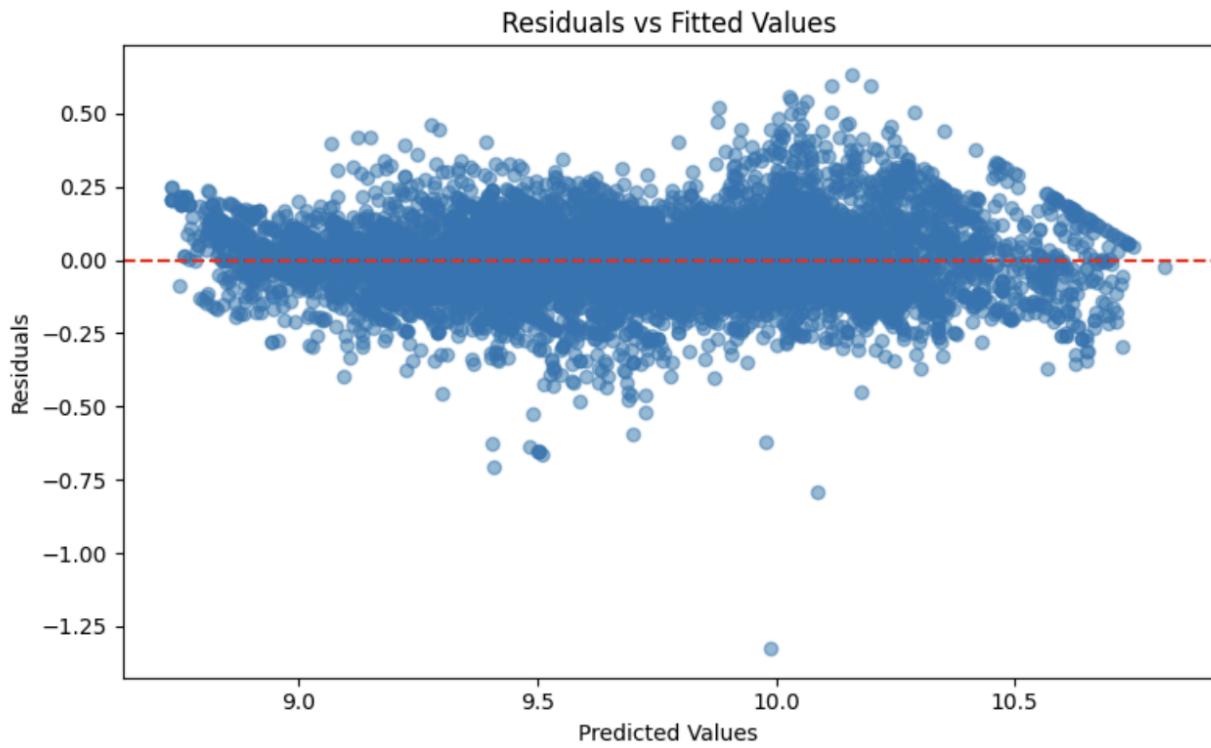
Residual diagnostics were performed on the training data to evaluate key assumptions of linear regression.

The mean of residuals is effectively zero (7.8×10^{-17}), indicating the absence of systematic bias in model predictions. The near-zero correlation between predicted values and residuals (≈ 0) confirms that residuals are randomly distributed around zero, satisfying the linearity assumption.

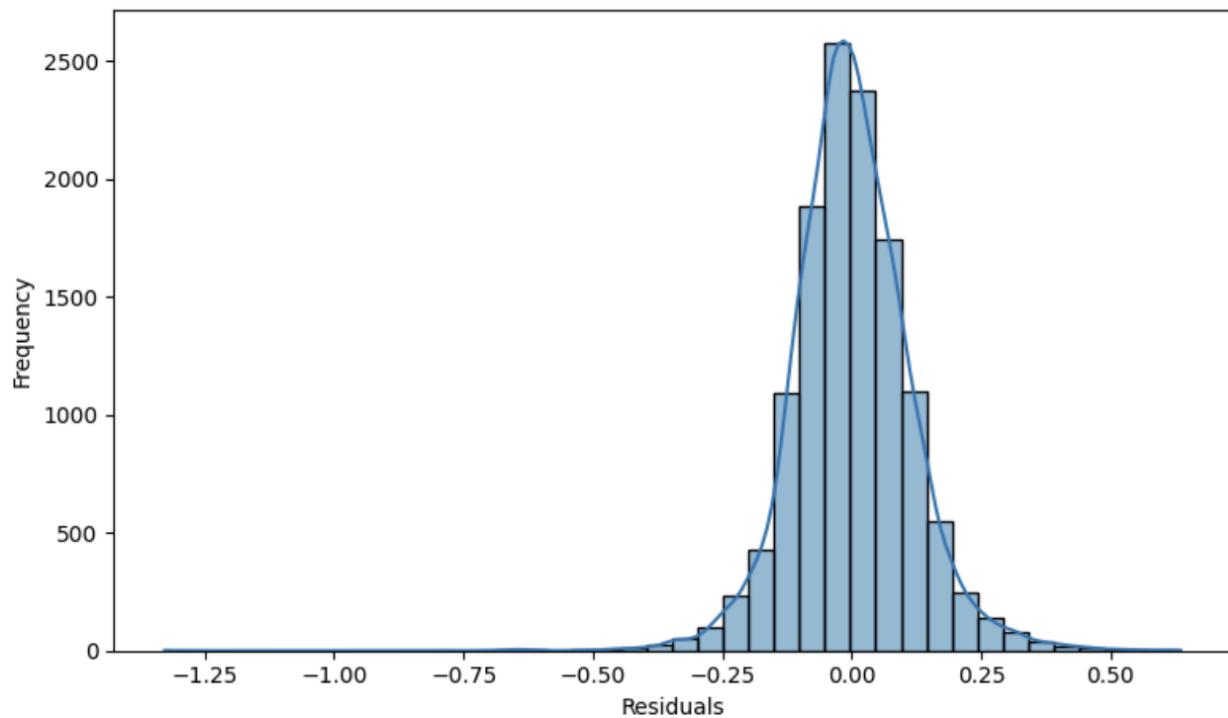
Residual skewness is close to zero (-0.02), suggesting a symmetric error distribution. While kurtosis is moderately positive (3.76), indicating slightly heavier tails than a perfect normal distribution, this is acceptable for large datasets and does not indicate severe violation of normality.

The residual spread is stable, with no evidence of increasing variance across prediction ranges, supporting the assumption of homoscedasticity.

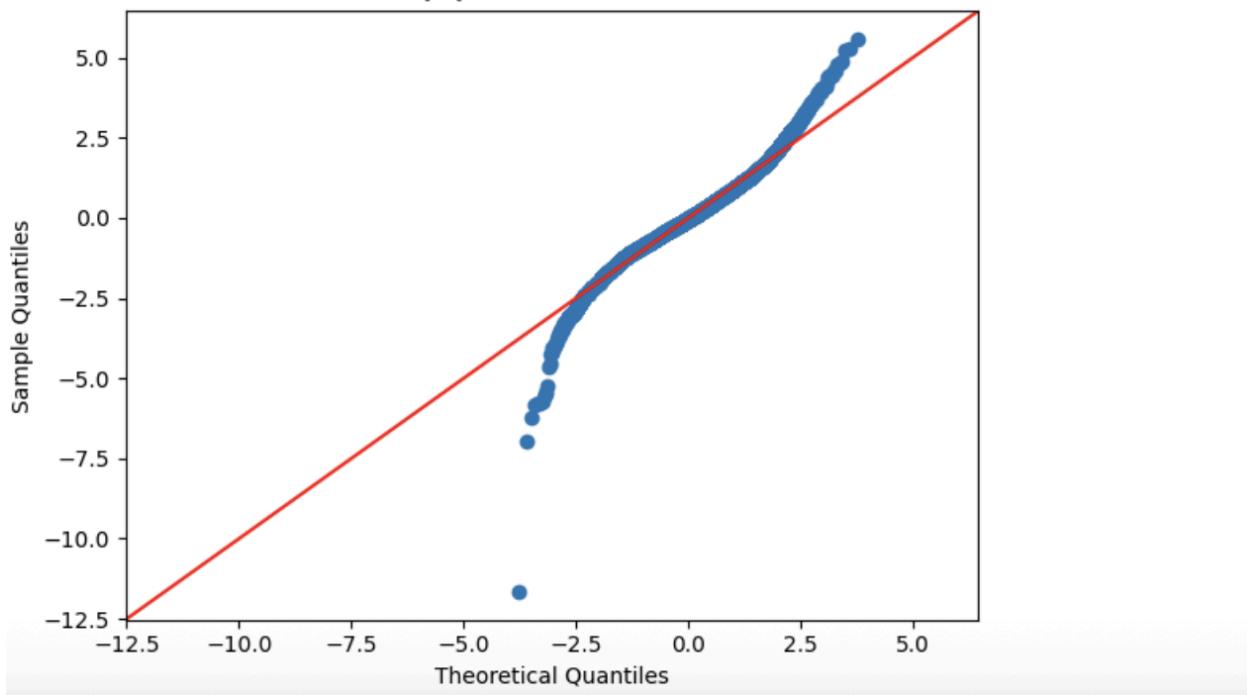
Overall, the residual diagnostics confirm that the baseline linear regression model satisfies the key assumptions of linearity, unbiased errors, approximate normality, and constant variance, making it suitable for further analysis and comparison with regularised models.



Distribution of Residuals



Q-Q Plot of Residuals



Multicollinearity Analysis (VIF)

Multicollinearity was assessed using the Variance Inflation Factor (VIF) on the scaled training data. Most features show **low to moderate VIF values (below 5)**, indicating acceptable correlation levels.

A few predictors, such as **engine power (hp_kW)**, **engine displacement**, **vehicle weight**, and certain **car models and body types**, exhibit moderately higher VIF values. This is expected, as these attributes are naturally related in vehicle data.

No feature exceeds critical VIF thresholds (≈ 10), so **no variables were removed** at this stage. Additionally, upcoming **Ridge and Lasso regression models** are well-suited to handle multicollinearity, making feature removal unnecessary.

3.2 Ridge Regression Implementation

3.2.1 Defining Alpha Values

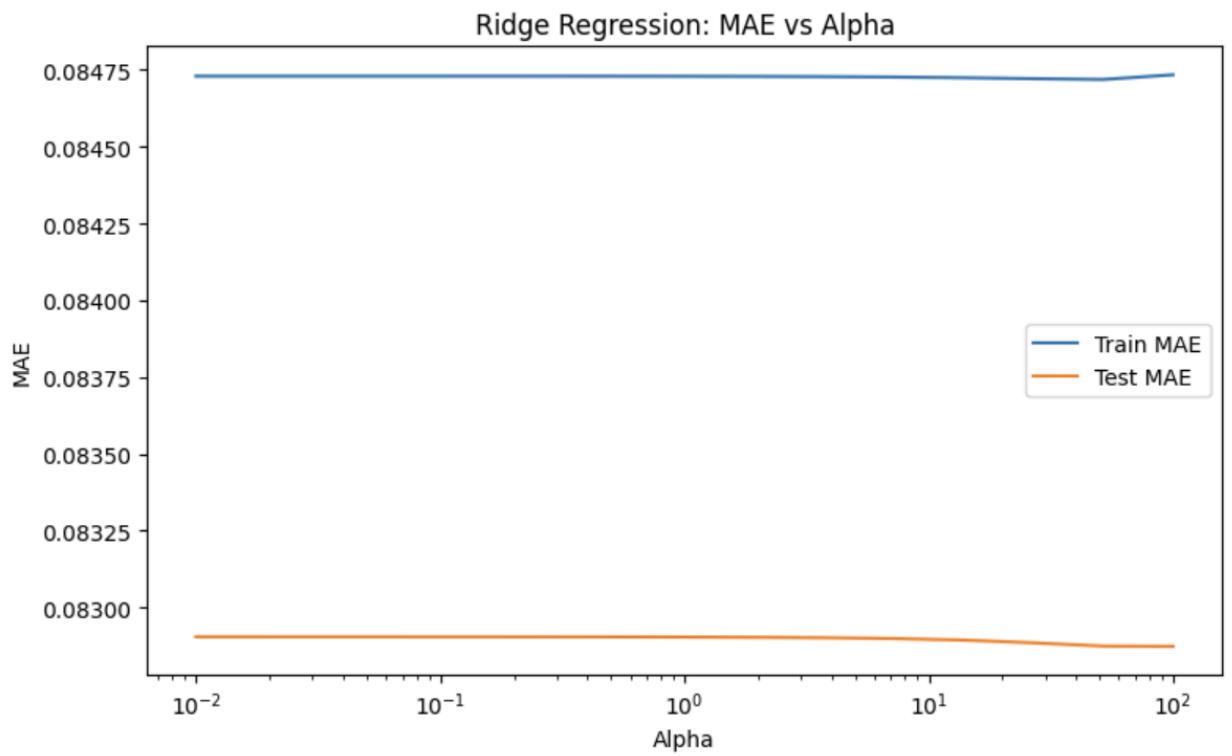
A range of alpha values from **0.01 to 100** was defined using a logarithmic scale. This enables efficient exploration of both weak and strong regularisation strengths and helps capture the bias–variance trade-off.

3.2.2 Ridge Regularisation and Alpha Selection

Ridge regression models were trained for each alpha value, and **Mean Absolute Error (MAE)** was evaluated on both training and test sets.

From the MAE vs alpha plot:

- Training MAE increases with alpha due to stronger coefficient shrinkage.
- Test MAE decreases initially and then stabilises, indicating improved generalisation.

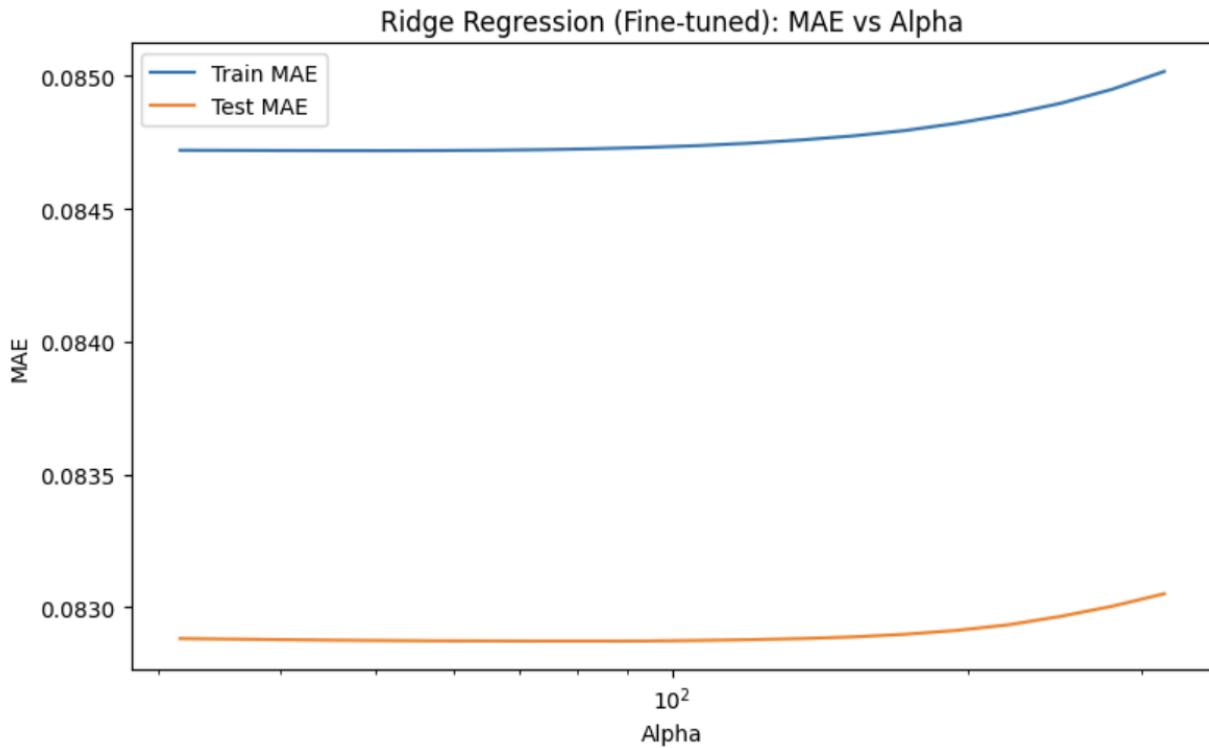


The **best performance on the test set** was achieved at:

- **Best alpha:** 100
- **Best test MAE:** 0.0829

This indicates that stronger regularisation helps control multicollinearity and reduces overfitting compared to the baseline linear regression model.

3.2.3 Fine-Tuning Ridge Regularisation (Alpha Selection)



After the initial alpha search, a finer range of alpha values was evaluated to identify the optimal regularisation strength more precisely. Ridge models were trained across this narrower range, and MAE was computed for both training and test sets.

The MAE vs alpha curve shows a **stable minimum region** between approximately **60 and 100**, indicating consistent generalisation performance and effective control of overfitting.

- **Optimal alpha (fine-tuned): 83.38**
- **Test MAE: 0.08287**
- **Test RMSE: 0.1099**
- **Test R²: 0.9239**

The small variation in MAE across neighbouring alpha values confirms that the model is **robust to regularisation strength**. Coefficient shrinkage helps manage multicollinearity while retaining all important predictors. Overall, fine-tuned Ridge regression provides a stable and well-generalised improvement over the baseline linear model.

3.3 Lasso Regression Implementation

3.3.1 Defining Alpha Values

Alpha values were selected on a logarithmic scale from **0.0001 to 1** to evaluate both weak and strong regularisation. This range allows Lasso to explore coefficient shrinkage and feature selection effects.

3.3.2 Lasso Regularisation and Alpha Selection

Lasso models were trained for each alpha and evaluated using MAE on training and test sets.

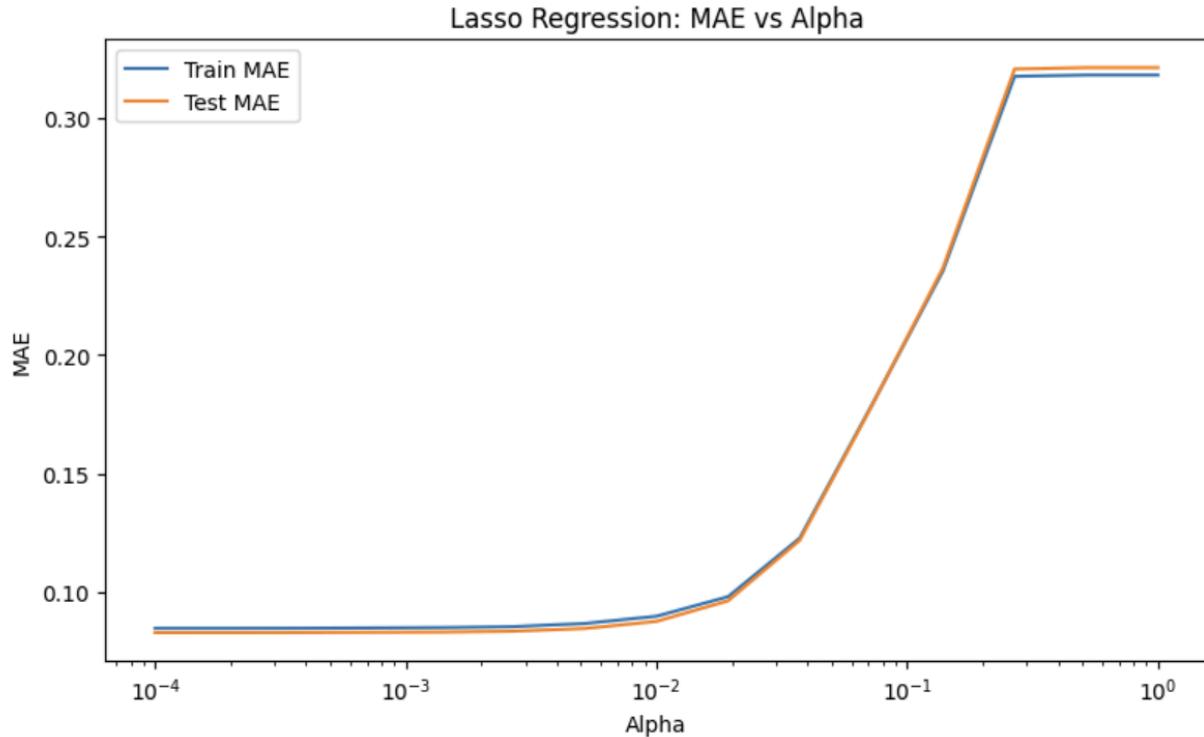
Results show that:

- **Lowest test MAE occurs at very small alpha values**, indicating mild regularisation is optimal.
- Increasing alpha causes a sharp rise in MAE, reflecting underfitting due to excessive coefficient shrinkage.

The **best alpha from the initial search was 0.0001**, achieving:

- **Test MAE:** 0.0829
- **Negative MAE:** -0.0829

This indicates that strong feature elimination is not beneficial at this stage.



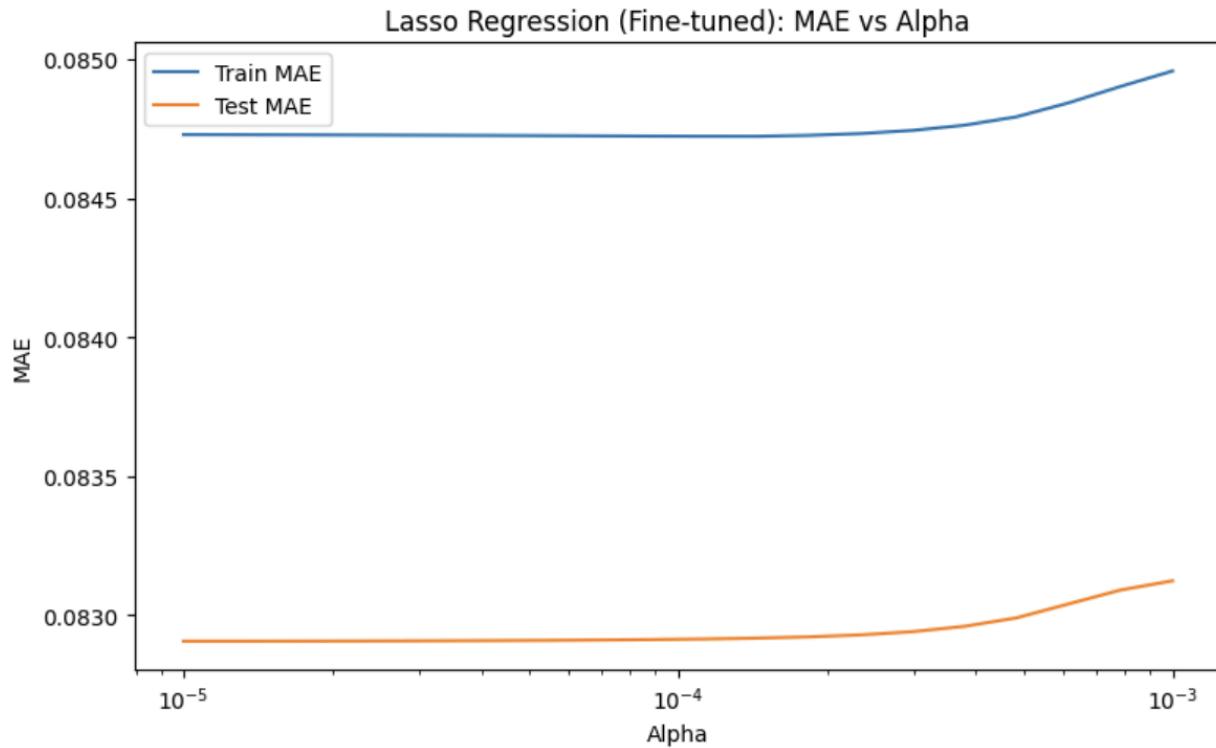
3.3.3 Fine-Tuning Alpha and Final Lasso Model

Fine-tuning around the initial best alpha identified **0.00001** as the optimal value, with marginal improvement.

Final test performance:

- **R²:** 0.9239
- **RMSE:** 0.1099
- **MAE:** 0.0829

Lasso shrinks many coefficients toward zero while retaining key predictors such as **engine power, mileage, age, and major vehicle models**. However, performance is comparable to Ridge, suggesting that aggressive feature selection offers limited benefit for this dataset.



3.4 Regularisation Comparison & Analysis

3.4.1 Comparison of Evaluation Metrics

The three models show very similar performance on the test dataset. Ridge Regression achieves the highest R² and the lowest RMSE and MAE, indicating slightly better generalisation than the baseline Linear Regression. Lasso Regression performs almost identically to Ridge but does not further reduce prediction error.

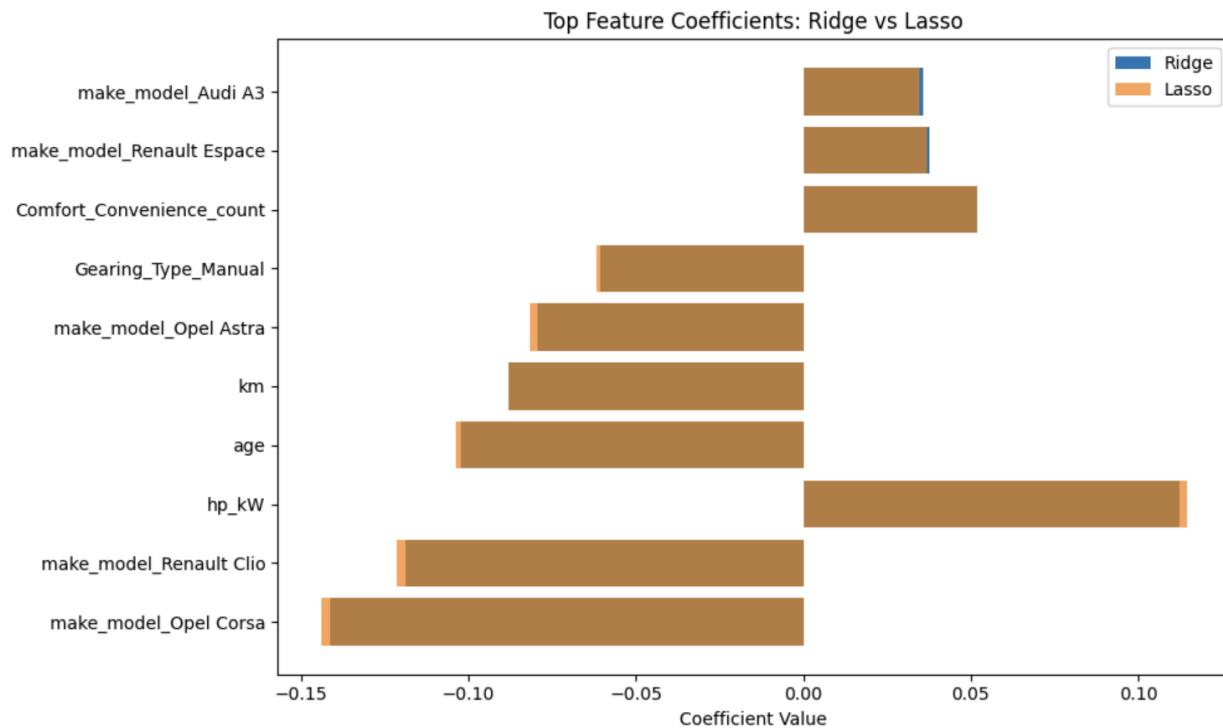
Overall, regularisation provides a small but consistent improvement over the unregularised model.

3.4.2 Comparison of Model Coefficients

The most influential features remain consistent across Linear, Ridge, and Lasso models, including engine power (`hp_kw`), mileage (`km`), vehicle age (`age`), and specific vehicle models.

Ridge Regression reduces coefficient magnitudes smoothly while retaining all features, helping stabilise estimates under multicollinearity. Lasso Regression does not set any coefficients to zero in this case, indicating that no feature is completely redundant after feature engineering.

Coefficient patterns across models are largely similar, confirming that regularisation mainly improves stability rather than altering feature importance.



4. Conclusion & Key Takeaways

Applying regularisation helped stabilise model coefficients and slightly improved generalisation performance. Ridge and Lasso regression achieved marginally better R^2 , RMSE, and MAE than the baseline linear regression, indicating that the original model was already well-fitted.

No overfitting was observed, as training and test performance remained closely aligned across all models. Regularisation mainly addressed multicollinearity among correlated features such as engine power, displacement, and vehicle weight rather than dramatically improving accuracy.

The dataset is sufficiently large and diverse, allowing linear models to capture price patterns effectively. Lasso regression did not eliminate any predictors, suggesting no strong redundancy after feature engineering. Ridge regression proved more suitable due to its ability to handle multicollinearity while retaining all features.

Overall, a linear modeling approach is appropriate and effective for this problem once proper preprocessing and feature engineering are applied.

4.1 Outcomes and Insights Gained

- Log transformation improved target distribution and model assumptions.
- Feature engineering reduced dimensionality without losing important information.
- Baseline linear regression already showed strong performance.
- Ridge regression provided the most stable improvement.
- Lasso did not remove any features, indicating low redundancy.
- Regularisation improved coefficient stability more than accuracy.
- Key price drivers remained consistent across models.
- No evidence of overfitting was found.
- Dataset size was sufficient for reliable modeling.
- Linear and regularised models are interpretable and suitable for price prediction.