**Exploratory Data Analysis (EDA) Report**

**Objective:**

To understand the structure, distribution, and relationships within the dataset to prepare for accurate car price prediction modeling.

**Descriptive Statistics Summary**

| **Feature** | **Description** | **Insights** |
| --- | --- | --- |
| km | Distance the car has been driven | Most cars are in the 20,000–100,000 km range. Outliers exist above 200,000 km. |
| ownerNo | Number of previous owners | Majority of listings are first or second owner cars, as reflected by strong clustering around 0.0 and 0.5 after scaling. |
| modelYear | Year of manufacture (scaled) | Most cars are relatively new; older cars have been scaled to reflect lower values. |
| engineCC | Engine displacement (cubic capacity) | Range from 800cc to 3000cc. Most common sizes around 1000–1500cc. |
| car\_age | Derived feature: current year - modelYear | Younger cars dominate the dataset. |
| price | Selling price of used cars (in lakhs) | Most cars fall in the 3–12 lakh range. Some luxury models push above 20+ lakhs. |

**Visualizations & Insights**

**Histograms (Distribution of Values)**

**Purpose:** To understand how values are distributed across each feature.

* **km**: Shows a positively skewed distribution, with most cars between 20k–100k km.
* **ownerNo**: Dominated by 1st and 2nd owners, confirming the market consists mostly of lightly pre-owned vehicles.
* **modelYear**: Most cars are recent (2016 onward); older cars are less common.
* **engineCC**: Distribution concentrated around ~1200cc–1500cc, typical of hatchbacks and compact sedans.
* **car\_age**: Inverse of modelYear—more recent cars dominate.
* **price**: Right-skewed distribution; most listings range between 3L–12L with few luxury car outliers >20L.

**Box Plots (Detecting Outliers)**

**Purpose:** To visualize the spread and identify extreme values.

* **km**: Outliers beyond 150k–200k km are visible.
* **ownerNo**: Fairly compact distribution—few cars with more than 2 owners.
* **engineCC**: Few outliers represent high-performance vehicles.
* **car\_age**: Older vehicles (7+ years) are few and can be considered for special price depreciation logic.

**Correlation Heatmap**

**Purpose:** To examine the linear relationship between numeric variables.

| **Feature 1** | **Feature 2** | **Correlation** | **Interpretation** |
| --- | --- | --- | --- |
| car\_age | price | **-0.71** | Strong negative — older cars have lower prices |
| engineCC | price | **+0.42** | Moderate positive — larger engines often mean higher value |
| km | price | -0.27 | Weak negative — more usage slightly reduces price |
| ownerNo | price | -0.21 | More previous owners = lower price |
| transmission\_Automatic | price | +0.37 | Automatic cars tend to be more expensive |

**Scatter Plots (Visual Patterns)**

**Purpose:** To detect non-linear or clustered relationships with the target (price).

* **car\_age vs price**: Clearly decreasing pattern—older cars have significantly lower prices.
* **engineCC vs price**: Positive trend—bigger engines fetch higher prices.
* **km vs price**: Slight downward curve, but with scattered points — more km = lower price, but less strongly.

**Feature Relevance for Modeling**

Based on EDA, the following features are likely to be highly predictive:

| **Feature** | **Reason** |
| --- | --- |
| car\_age | Strong negative correlation with price |
| engineCC | Bigger engines typically more valuable |
| transmission | Automatic cars carry price premium |
| km | Indicates wear and tear |
| ownerNo | More owners reduce trust/resale value |

EDA revealed that **car age**, **engine size**, and **transmission type** are among the most influential features in predicting used car prices. It also identified:

* Outliers that could affect model accuracy
* Distribution trends for feature transformation
* Correlations for feature importance