**Exploratory Data Analysis (EDA) – Documentation Report**

**1. Objective of EDA**

The primary goal of the Exploratory Data Analysis (EDA) phase is to understand the structure, distribution, patterns, and relationships within the dataset.  
It helps to identify trends, anomalies, correlations, and key driving features influencing used car prices.

**2. Summary Statistics (Descriptive Statistics)**

| **Aspect** | **Insight** |
| --- | --- |
| Number of records | 7,400 after preprocessing |
| Key numerical features analyzed | km, modelyear, ownerno, prize |
| Data quality | No missing values after preprocessing |

Conclusion:  
The dataset is clean and ready for advanced modeling.

**3. Univariate Analysis (Single Feature Behavior)**

**3.1 Kilometers Driven (km)**

* **Distribution:** Right-skewed; most cars have driven less than 20,000 normalized kilometers.
* **Boxplot:** Few mild outliers on the higher side.
* **Observation:** Cars driven excessively are fewer, and capping can be considered during modeling.

**3.2 Model Year (modelyear)**

* **Distribution:** Left-skewed; majority of the cars are manufactured after 2012.
* **Boxplot:** A few very old models exist, but most are recent.
* **Observation:** Newer cars dominate the dataset, which is good for stable pricing models.

**3.3 Owner Number (ownerno)**

* **Distribution:** Heavily skewed toward 1st and 2nd owners.
* **Boxplot:** Some rare third or higher ownership cars exist.
* **Observation:** Most cars are lightly transferred between owners, retaining resale value.

**3.4 Price (prize)**

* **Distribution:** Right-skewed; most cars fall under ₹2L to ₹10L range.
* **Boxplot:** Few highly priced luxury vehicles are visible as outliers.
* **Observation:** The dataset has a good spread for both budget and luxury vehicles.

**4. Bivariate Analysis (Relationship with Target: Price)**

| **Feature** | **Relationship with Price** | **Comments** |
| --- | --- | --- |
| Kilometers Driven (km) | Slight negative correlation | As KM increases, price slightly drops but not very strong. |
| Model Year (modelyear) | Strong positive correlation | Newer model year clearly demands higher price. |
| Owner Number (ownerno) | Weak negative correlation | More previous owners slightly reduce car price. |

Conclusion:  
Model Year is the **strongest simple predictor** among numerical variables.

**5. Correlation Heatmap Analysis**

* Very weak overall correlations visible between basic features and target.
* No extreme multicollinearity found among input features.
* Indicates need for **non-linear models** like Random Forest, Gradient Boosting that can discover hidden patterns.

**6. Feature Importance Analysis (Using Random Forest)**

| **Rank** | **Feature** | **Importance (%)** | **Comments** |
| --- | --- | --- | --- |
| 1 | Body Type (SUV) | 26% | SUVs have much higher resale value. |
| 2 | Model Year | 22% | Newer models fetch better prices. |
| 3 | Transmission (Manual) | 7% | Manual and Automatic prices differ. |
| 4 | Body Type (Hatchback) | 6% | Hatchbacks slightly influence affordability. |
| 5 | Kilometers Driven (km) | 4% | More KMs lead to slight price depreciation. |
| 6 | City, OEM, Specific Models | <3% each | Minor but still impactful. |

Conclusion:  
Price prediction highly depends on **car type**, **year of manufacturing**, and **transmission type**, matching real-world expectations.

**7. EDA Final Summary**

| **Aspect** | **Evaluation** |
| --- | --- |
| Data Distribution | Normal with mild skewness and minor outliers |
| Feature Trends | Model year and body type dominate price prediction |
| Missing Values | None (after preprocessing) |
| Correlation Insights | Weak basic correlation; complex relationships exist |
| Feature Strengths | SUV status, Newness (model year), Transmission, KM driven |
| Next Recommendation | Feature engineering (minor) + Start model building |