Car Price Prediction Report

# 1. Overview

The goal of this project is to develop a predictive model that estimates the selling price of used cars using various features like brand, age, mileage, fuel type, transmission, and ownership status. This model can assist dealerships or individuals in pricing vehicles more accurately in a data-driven manner.  
  
The dataset used includes key attributes such as:  
- name (used to extract brand),  
- year (used to calculate car age),  
- km\_driven,  
- fuel,  
- seller\_type,  
- transmission,  
- owner,  
- and the target variable: selling\_price.

# 2. Exploratory Data Analysis (EDA)

Key insights from the data exploration:  
  
- Year of Manufacture: Newer vehicles tend to fetch higher prices, indicating a strong positive correlation.  
- KM Driven: Heavily used cars (high mileage) tend to have lower selling prices.  
- Fuel Type: Most cars are either Petrol or Diesel; Petrol cars slightly dominate the dataset.  
- Brand: Premium brands like Audi, BMW, and Mercedes-Benz consistently show higher resale values.  
- Outliers: Some entries (very high prices or mileage) suggest the need for outlier detection and removal in future improvements.  
- Distribution: Selling price is right-skewed, with many cheaper cars and fewer expensive ones.  
  
Visuals used:  
- Heatmap to observe feature correlations  
- Histograms and KDE plots to understand distribution  
- Pairplots to inspect relationships between variables

# 3. Model Development

We used Linear Regression for its simplicity and interpretability. The steps included:  
  
- Encoding categorical variables using Label Encoding.  
- Feature-target split and train-test split (80:20 ratio).  
- Model training on the training set.  
- Predicting prices on the test set.

# 4. Model Evaluation

|  |  |
| --- | --- |
| Metric | Value (approx.) |
| Mean Absolute Error (MAE) | ₹113,000 |
| Mean Squared Error (MSE) | ₹29.1 Crores² |
| Root Mean Squared Error (RMSE) | ₹170,000 |
| R² Score | 0.62 |

Interpretation:  
- The model explains ~62% of the variance in the data, which is reasonable for a baseline linear model.  
- The error margins indicate some noise or complexity in the data that linear regression alone cannot capture.

# 5. Feature Importance

Based on the model coefficients, the most impactful features are:  
- Year (positive): Newer cars are priced higher.  
- Brand: Premium brands have a significant positive impact.  
- KM Driven (negative): More usage reduces value.  
- Fuel Type and Transmission also have moderate influence.

# 6. Conclusion

The linear regression model provides a strong starting point for predicting car prices. However, its performance can be improved with:  
  
- Outlier removal  
- Polynomial features or interactions  
- Advanced models like Random Forest, XGBoost  
- Hyperparameter tuning  
- Feature selection techniques