Car Dataset Analysis Report

1. Overview

The dataset contains detailed information about used cars listed for sale, including features like Make, Model, Year of manufacture, Engine, Power, Mileage, Fuel Type, Transmission type, and Price. The **target variable** for this analysis is the **Price of the car (in Lakhs)**.

The dataset is **tabular**, with each row representing a car listing and each column representing a specific attribute.

2. Data Cleaning and Preprocessing

2.1 Missing Values Treatment

Missing data can introduce bias and affect analysis. Here's how missing values were handled:

Column Type	Treatment Applied	Justification
Numerical	Imputed with **median**	Median is robust against outliers and preserves central tendency.
Categorical	Imputed with **mode**	Mode ensures the most common category is retained without introducing bias.

Python Code:

```
for col in num_cols:
    df[col] = df[col].fillna(df[col].median())
for col in cat_cols:
    df[col] = df[col].fillna(df[col].mode()[0])
```

2.2 Units Removal

Certain columns had units (e.g., "kmpl", "CC", "bhp", "Lakh") which were removed to convert them into numeric types for analysis.

Column	Raw Format	Cleaned Format	Notes
Mileage	"19.67 kmpl"	19.67	Converted to numeric
Engine	"1582 CC"	1582.0	Converted to numeric
Power	"126.20 bhp"	126.20	Converted to numeric
New_Price	"4.78 Lakh"	4.78	Converted to numeric

2.3 Categorical Variables Encoding

Categorical variables were converted to **numerical format using one-hot encoding**, allowing machine learning models to use them effectively.

Original Column	Encoded Columns
Fuel_Type	Fuel_Type_Petrol, Fuel_Type_Electric
Transmission	Transmission_Manual

2.4 Feature Engineering

1. **Age_of_Car:** Derived as Current Year (2025) – Year of Manufacture.
   Example: Car from 2015 → Age_of_Car = 10 years.
2. **Price_per_CC:** Derived as Price / Engine, providing a cost-efficiency metric per unit of engine displacement.

3. Data Exploration

3.1 Selected Columns

A subset of features for analysis:

Name	Location	Engine	Power	Price
Hyundai Creta 1.6 CRDi SX Option	Pune	1582.0	126.20	12.50
Honda Jazz V	Chennai	1199.0	88.70	4.50
Maruti Ertiga VDI	Chennai	1248.0	88.76	6.00
Audi A4 New 2.0 TDI Multitronic	Coimbatore	1968.0	140.80	17.74
Nissan Micra Diesel XV	Jaipur	1461.0	63.10	3.50

3.2 Filtered Data (Mileage > 15 kmpl)

Name	Mileage	Engine	Price
Hyundai Creta 1.6 CRDi SX Option	19.67	1582.0	12.50
Maruti Ertiga VDI	20.77	1248.0	6.00
Audi A4 New 2.0 TDI Multitronic	15.20	1968.0	17.74
Nissan Micra Diesel XV	23.08	1461.0	3.50
Volkswagen Vento Diesel Comfortline	20.54	1598.0	5.20

Cars with higher mileage tend to be smaller or more fuel-efficient models.

3.3 Top 5 Most Expensive Cars

Name	Price	Engine	Age_of_Car
Land Rover Range Rover 3.0 Diesel LWB Vogue	160.0	2993.0	8
Lamborghini Gallardo Coupe	120.0	5204.0	14
Jaguar F Type 5.0 V8 S	100.0	5000.0	10
Land Rover Range Rover Sport SE	97.07	2993.0	6
BMW 7 Series 740Li	93.67	2979.0	7

Observation: Luxury and sports cars have high prices and high engine capacity. Age does not always determine price.

3.4 Summary Table – Average Price by Fuel Type

Fuel_Type_Petrol	Average Price (Lakh)
False	12.96
True	5.76

Petrol cars tend to be **less expensive on average** than non-petrol cars (which include electric and diesel).

4. Key Insights

1. **Data Cleaning:** Missing values were handled carefully using median/mode imputation. Units were removed and categorical variables encoded to make the dataset analysis-ready.
2. **Derived Features:** Age of car and Price_per_CC provide deeper insight into depreciation and cost-efficiency.
3. **Fuel Type Analysis:** Diesel or electric cars tend to be more expensive, while petrol cars are more common and affordable.
4. **Luxury Cars:** Top-priced vehicles are typically high-engine, low-production luxury models.
5. **Mileage Filter:** Cars with mileage above 15 kmpl include fuel-efficient or mid-range vehicles suitable for everyday use.

The dataset is now **clean, structured, and feature-rich**, enabling further statistical analysis or machine learning applications like price prediction. Derived features such as Age_of_Car and Price_per_CC enhance the interpretability of car prices.