**Week-01:** **Car Price Prediction Dataset Analysis**

Dataset Link: https://www.kaggle.com/datasets/sukhmandeepsinghbrar/car-price-prediction-dataset

Organization: **Sapience Edu Connect Pvt Ltd**

Submitted by: **Muhammad Yasir**

Week-01: **Data Science Internship**

Submitted Date**: Sunday- 21/06/2025**

**Given Tasks**:

1. Data Cleaning
2. Preprocessing
3. Visualization
4. ML Feasibility
5. Key Insight and Patterns

# Objective

To analyze and preprocess the car price dataset, visualize key insights, and determine the most appropriate machine learning approach for car price prediction.

# 1. Data Cleaning and Preprocessing

✔️ Data loaded using Pandas

✔️ Removed irrelevant column: Car\_Name

✔️ Created a new feature: Car\_Age (2025 - Year)

✔️ Dropped original 'Year' column

✔️ Converted categorical variables using one-hot encoding

✔️ No missing values were found in the dataset

# 2. Data Visualization

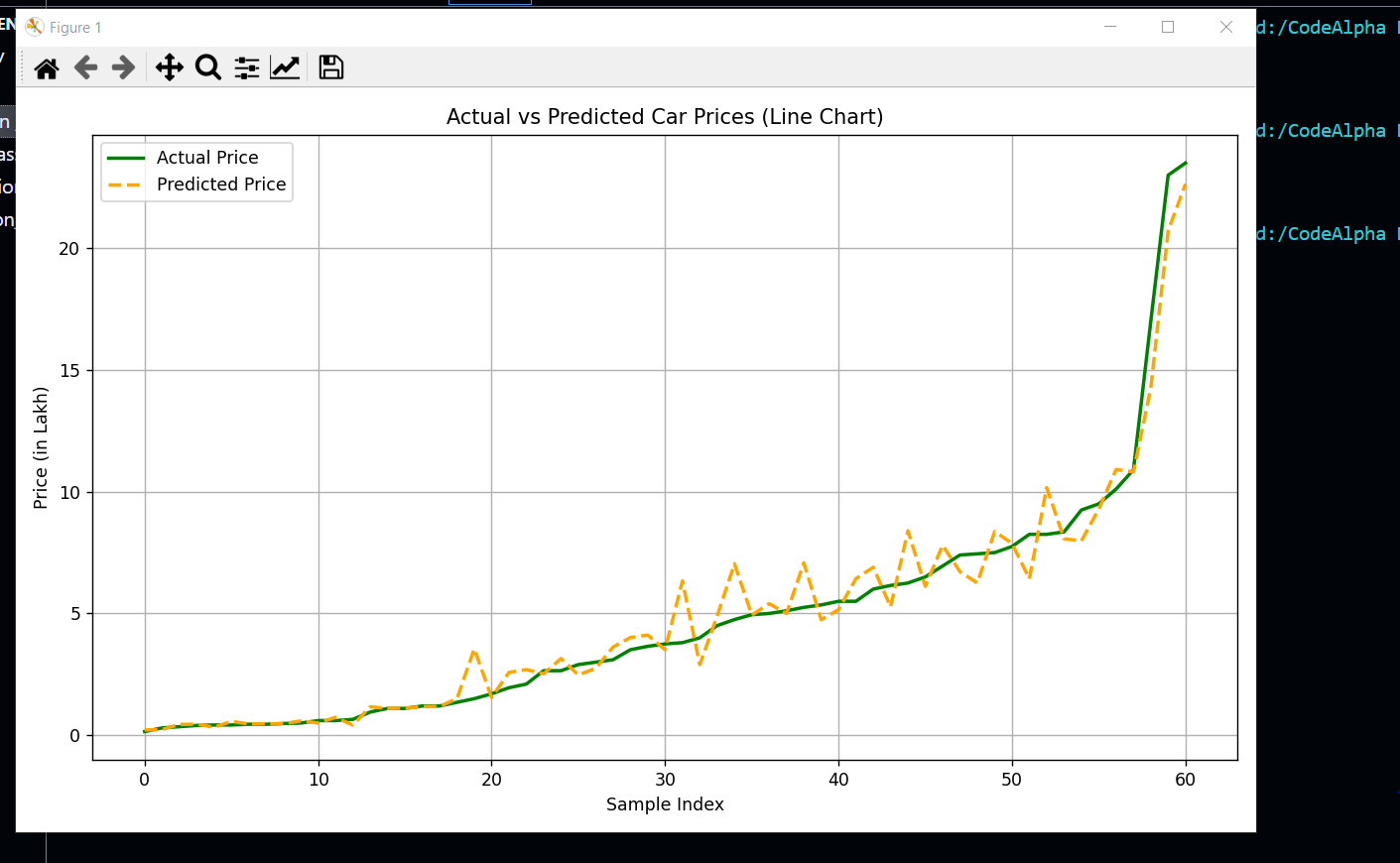
Explored patterns between features and Selling\_Price.

Visualized Actual vs Predicted prices using a line chart.

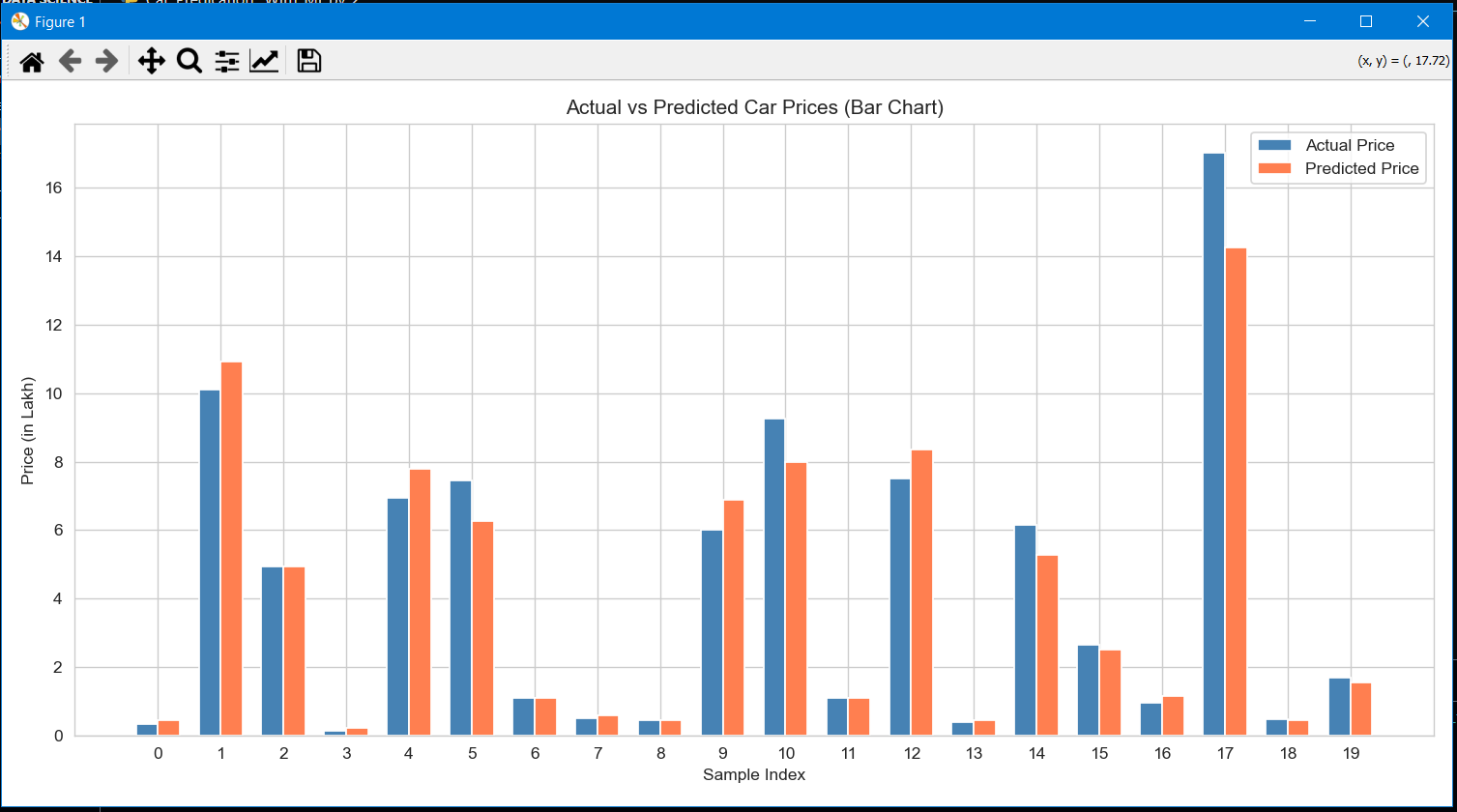
**Observations:**

- Most predictions align closely with actual prices

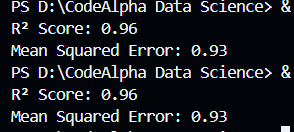
- Slight deviation seen in a few outliers



Figure(a): **Line Chart**



Figure(b): **Bar Chart**



# Figure(c): Model Evaluation Result

# 3. Data Analysis Report

**Key Insights**:

- Car Age negatively correlates with price

- Petrol cars dominate; fuel type has moderate pricing impact

- Manual transmission is common; automatic cars priced higher

- Multi-owner cars are typically cheaper

**Decisions Made:**

- Removed non-predictive features

- Engineered more meaningful features like Car\_Age

- Used one-hot encoding for categorical features

# 4. Machine Learning Feasibility

Is this a Regression or Classification Problem?

**Answer**: Regression Problem

**Justification:** Target variable 'Selling\_Price' is continuous.

Is this a Supervised or Unsupervised Learning Problem?

**Answer:** Supervised Learning

Justification: Features and labeled output provided; learning is guided.

# 5.Key Insights & Patterns:

1. **Car Age and Price Relationship**
   * Newer cars tend to have a **higher selling price**.
   * Older cars experience **depreciation**, reducing their market value significantly.
2. **Fuel Type Impact**
   * **Petrol cars** are the most common in the dataset.
   * **Diesel cars** generally have **higher selling prices** than petrol due to fuel efficiency and market demand.
   * **CNG vehicles** are rare and have lower resale value.
3. **Transmission Type**
   * **Manual transmission** vehicles dominate the dataset.
   * However, **automatic cars** typically have a **higher average selling price**.
4. **Owner History**
   * Cars with **multiple owners** (2nd or 3rd owner) have a **notable drop in price**.
   * **First-owner cars** retain the **most resale value**.
5. **Selling Type**
   * **Individual sellers** constitute the majority.
   * **Dealer and Trustmark dealers** generally list cars at **higher prices**, possibly due to added services or certification.
6. **No Missing Values**
   * The dataset is **clean** with no missing entries, allowing direct model training without imputation.
7. **Skew in Price Distribution**
   * A small number of **luxury or expensive vehicles** skew the price range.
   * Most cars fall within the **3 to 10 lakh INR** range.
8. **Multicollinearity Mitigated**
   * Use of drop\_first=True in dummy encoding reduced **multicollinearity** issues in regression.
9. **Feature Importance**
   * Preliminary model results show that **Car\_Age**, **Present\_Price**, and **Fuel\_Type\_Diesel** are **highly influential** in predicting the selling price.

# Tools and Libraries Used :

- Pandas, NumPy – Data manipulation

- Matplotlib, Seaborn – Visualization

- Scikit-learn – ML modeling

- RandomForestRegressor – Used for predictions

# Conclusion:

Data is clean and suitable for regression modeling. Random Forest performed well with good evaluation scores. Further optimization possible in later stages (e.g., scaling, tuning).