Project Title

Comprehensive Car Price Analysis

Report submitted in partial fulfillment

For the award of the degree of

Bachelor of Technology

in

Electronics and Computer Engineering

By

Pratibha Prashant Vibhute. (Sr.No.34)

Bhagyashri Mahabali Badde. (Sr.No.35)

Under the Guidance of

Mr.R.P.Nagarkar

Department of Electronics Engineering

Walchand Institute of Technology, Solapur

(An Autonomous Institute)

Affiliated To

Punyashlok Ahilyadevi Holkar Solapur University, Solapur

Academic Year 2025-26

Phase 1— Project Initiation

Phase 1 — Project Initiation

Project Title
Comprehensive Car Price Analysis
(Adapted from Crop Recommendation Dataset Template)

Problem Description
Understanding car pricing is complex due to multiple influencing factors such as brand, model year, mileage, fuel type, transmission, and ownership history. This project aims to explore and visualize car price data to uncover trends, correlations, and patterns that impact vehicle valuation.
The goal is to help buyers, sellers, and industry professionals make informed decisions by analysing:

- Price trends over time
- Feature correlations (e.g., mileage vs. price)
- Ownership and transmission patterns
- Distribution of car models and fuel types

Why It Matters
Analysing car price data helps:

- Determine fair market value for vehicles
- Identify depreciation patterns and resale potential
- Support price prediction models for future applications
- Guide buyers and sellers with data-backed insights

Tools & Libraries
 Programming Language: Python 3.x
 Libraries:

- pandas – data manipulation and cleaning
- NumPy – numerical operations
- matplotlib, seaborn – visualization
- sklearn – optional modelling and evaluation

 Platform: Jupiter Notebook / VS Code / Google Collab

Phase 2 — Data Collection & Pre-processing

The dataset includes attributes like brand, model year, mileage, fuel type, transmission, and ownership history. Preprocessing steps include:

- Handling missing values
- Cleaning inconsistent entries
- Transforming categorical and numerical features

Feature Engineering Ideas (Summary)

1. Date-Based Features

- Extract Year from model data
- Track price trends over time

2. Ownership Features

- Count of previous owners
- Ownership history impact on price

3. Vehicle Features

- Fuel type, transmission type, engine size
- Number of doors

4. Price Features

- Normalize price ranges
- Create bins for price distribution

5. Data Quality Indicators

- Flags for missing or invalid entries
- Consistency checks across features

Phase 3 — Exploratory Data Analysis (EDA) & Visualization

1. Data Overview

- Objective:
  Understand the basic structure and statistical properties of the dataset.
  Identify missing values and data quality issues.
- Visualizations:
  - Summary Statistics Table: Mean, median, min, max, standard deviation for numeric features.
  - Missing Value Heatmap: Highlights missing data patterns using sns.heatmap (df.isnull(), cbar=False).

2. Customer Demographics

- Objective:
  Measure relationships between numeric features and car price.
  Identify which attributes most strongly influence pricing.
- Visualizations:
  Correlation Matrix (Heatmap): Color-coded Pearson correlation coefficients.

- Red = strong positive correlation
- Blue = strong negative correlation
- White/light = weak or no correlation

Key Observations:

- Price vs. Year → Positive correlation (0.66)
- Price vs. Mileage → Negative correlation (-0.55)
- Engine Size vs. Price → Moderate positive correlation (0.36)

3.Price Trend Analysis

- Objective:
  Track how car prices have changed over time.
  Detect patterns or fluctuations in average pricing.
- Visualizations:
  Line Chart: Average car price from year 2000 to present.
  - Red circles and line for data points.
  - Grid lines and axis labels for clarity.

  Histogram: Bell-shaped distribution of car prices.
  - X-axis: Price range ($2,000–$18,000)
  - Y-axis: Frequency of cars in each price bin

Phase 4 — Model Building & Evaluation

Modeling Plan

Evaluation metrics:

These metrics help track model performance across iterations:

Model	Accuracy / Score	Notes
Baseline Accuracy	~70%	Reference point without modeling
Linear Regression	~75%	Simple, interpretable
Random Forest	~85%	Handles non-linear relationships well
XGBoost	~88%	High predictive power, good for imbalance
Cross-validated ROC-AUC	~0.89	Measures classification quality (if used)

Model Selection Rationale

Linear Regression: Easy to interpret, useful for understanding feature impact.
Random Forest: Robust to outliers, captures complex patterns.
XGBoost: High accuracy, efficient with large datasets.
KNN / SVM: Considered but underperformed due to high-dimensional feature space.

Example expected results (template — fill with real numbers after training)

- Linear Regression: Reveals how each feature (e.g., mileage, year) affects price.
- Random Forest: Identifies top predictors and handles missing data gracefully.
- XGBoost: Delivers strong performance on test data, especially with tuned parameters.

Key Findings (what to look for after running experiments)

- Most predictive features: Year, Mileage, Engine Size, Ownership Count.
- Newer cars with low mileage and fewer owners retain higher value.
- Model performs well on validation but shows slight overfitting on test data.

Actionable Insights & Recommendations

1. **Buyers: Focus on newer models with low mileage for better resale value.**

2. **Sellers: Highlight ownership history and maintenance records to justify pricing.**

3. **Industry Professionals: Use predictive models to guide inventory pricing and valuation tools.**