Car Price Prediction Project

1. Introduction

The used car market is highly dynamic, with prices influenced by various factors such as brand, engine size, fuel type, mileage, and other specifications. Buyers often struggle to determine whether a car is fairly priced, and sellers may undervalue or overprice vehicles unintentionally.

This project builds a **Linear Regression model** to predict car prices based on their features, providing **data-driven insights** for car buyers, sellers, and dealerships.

---

2. Objectives

- Predict the price of a car based on its specifications.
- Identify the **key features** that most influence car prices.
- Handle data cleaning, preprocessing, and outlier management for a clean dataset.
- Build a reliable machine learning model with strong predictive performance.

---

3. Pain Points Addressed

- **Lack of Price Transparency** – Helps buyers know if a price is fair.
- **Inconsistent Pricing** – Standardizes price estimation based on objective features.
- **Manual Valuations** – Replaces subjective, time-consuming appraisals with automated predictions.
- **Overpayment & Undervaluation Risks** – Ensures buyers don't overpay and sellers don't undersell.
- **Complex Pricing Factors** – Considers multiple attributes simultaneously to provide accurate estimates.

---

4. Dataset Description

The dataset contains **205 rows and 26 columns**, covering various car features:

- **Numerical Features:** `enginesize`, `curbweight`, `horsepower`, `price`, etc.
- **Categorical Features:** `brand`, `fueltype`, `drivewheel`, `carbody`, etc.
- **Target Variable:** `price` (continuous numeric value).

---

5. Data Cleaning & Preprocessing

- **Removed non-informative columns** like `car_ID`.

- **Split `CarName` into `brand` and model name** for better feature grouping.
- **Fixed brand inconsistencies** (e.g., `vokswagen` → `volkswagen`).
- **Handled outliers** using Winsorization (capping extreme values).
- **Encoded categorical variables**:
  - Label Encoding for binary features (e.g., `fueltype`).
  - One-Hot Encoding for multi-category features (e.g., `brand`, `carbody`).

---

6. Exploratory Data Analysis (EDA)

- **Distribution Analysis:** Found that `price` is **right-skewed** with some luxury car outliers.
- **Boxplots & Countplots:** Helped visualize outliers and categorical feature distributions.
- **Correlation Analysis:** Identified strong positive correlations (e.g., `enginesize`, `curbweight`, `horsepower`).

---

7. Feature Engineering & Selection

- High multicollinearity detected using **VIF** → selected features carefully.
- Chosen features for modeling:
  - `enginesize`, `curbweight`, `carwidth`, `cylindernumber`, `carlength`, `drivewheel_rwd`, `wheelbase`, `boreratio`, `highwaympg`, `brand_buick`.

---

8. Model Development

- **Model Used:** Linear Regression
- **Data Split:** 80% training, 20% testing
- **Scaling:** Applied `StandardScaler` for feature normalization.
- **Training:** Model learned relationships between features and price.
- **Prediction:** Produced price estimates for unseen (test) data.

---

9. Model Evaluation

- **R^2 Score:** 0.823 (Test) → Explains 82% of price variance.
- **MAE:** ~$2,500 → Average prediction error.
- **RMSE:** ~$3,735 → Typical deviation from true prices.

Model performed **consistently on training and testing sets**, showing good generalization.

---

10. Key Insights

- **Enginesize, curbweight, and horsepower** strongly drive prices.
- Cars with **better fuel economy (highwaympg)** tend to have lower prices.
- Brand matters — **Buick and BMW cars generally cost more**.

---

11. Visualizations

- **Price Distribution:** Showed right skewness and high-end car outliers.
- **Actual vs Predicted Plot:** Most predictions aligned well with actual prices.
- **Residual Plot:** Residuals centered around zero → model errors are mostly random.

---

12. Pain Points Solved (Impact)

- Buyers gain **confidence** with fair price estimates.
- Sellers avoid **undervaluing** their vehicles.
- Dealerships can automate **appraisals** and save time.
- Market becomes more **transparent and efficient**.

---

13. Future Improvements

- Use **non-linear models** like Random Forest or XGBoost for luxury car predictions.
- Include additional data (e.g., mileage, location, accident history).
- Deploy the model as a **web app** for real-time car price predictions.

---

14. Conclusion

This project successfully built and evaluated a **Linear Regression model** for predicting car prices using multiple features. With ~82% accuracy, it provides **meaningful, data-driven insights** that can benefit buyers, sellers, and dealerships.

By improving and expanding this work (e.g., using more advanced models and features), we can move closer to creating a **fully automated, intelligent car pricing system**.