📄 Car Price Prediction Using Polynomial Regression

1. Project Overview

This project focuses on predicting **car prices** using **Polynomial Regression**, an extension of Multiple Linear Regression that captures **non-linear relationships** between variables. The dataset is sourced from Kaggle and contains detailed information about car specifications and historical usage.

Based on Exploratory Data Analysis (EDA), polynomial regression was selected to better model the **non-linear effect of mileage on car price**, while maintaining simplicity and avoiding overfitting.

---

2. Dataset Description

Dataset Path:
`/kaggle/input/car-prepiction/car_price_dataset.csv`

Target Variable:

- `Price` – Selling price of the car

Key Features Used (EDA-driven):

- `Year` – Manufacturing year of the car
- `Mileage` – Distance driven by the car

Other features were excluded due to weak correlation or noise.

---

3. Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

- Histograms and box plots were used to analyze distributions.
- `Price` showed right skewness.
- `Mileage` had large variance with visible outliers.

3.2 Bivariate Analysis

- **Year vs Price:** Strong positive linear relationship
- **Mileage vs Price:** Strong negative relationship with slight curvature

3.3 Correlation Analysis

- `Year` had strong positive correlation with `Price`
- `Mileage` had strong negative correlation with `Price`
- These findings justified a **polynomial relationship** rather than strict linearity.

---

4. Data Preprocessing

- Selected only **numerically strong features** based on EDA.
- Removed irrelevant or weak predictors.
- Data was split into:
  - **80% Training**
  - **20% Testing**

No missing values were found in the selected features.

---

5. Model Selection Rationale

Multiple polynomial degrees were tested:

Degree	Observation
1	Underfitting (linear model)
2	Best balance between bias and variance
3	Overfitting due to small dataset

✅ **Polynomial degree = 2** was selected based on **lowest RMSE and highest R^2 score**.

---

6. Model Building

Polynomial features were generated using degree 2, followed by training a Linear Regression model.

Mathematical Form:

[
Price = b_0 + b_1(Year) + b_2(Mileage) + b_3(Year^2) + b_4(Mileage^2) + b_5(Year \times Mileage)
]

Where:

- (b_0) is the intercept
- Remaining coefficients represent learned weights

---

7. Model Evaluation

The model was evaluated on unseen test data using standard regression metrics:

- **MAE (Mean Absolute Error)** – Average prediction error
- **MSE (Mean Squared Error)** – Penalizes large errors
- **RMSE (Root Mean Squared Error)** – Error in original price units
- **R^2 Score** – Variance explained by the model

Visualization:

- Actual vs Predicted plot
- Residual analysis
- 3D regression plane (Year × Mileage × Price)

These confirmed good model fit and reasonable generalization.

---

8. Results & Insights

- Car prices increase with newer manufacturing years.
- Mileage has a strong non-linear negative impact on price.
- Polynomial regression captured curvature missed by linear models.
- Degree-2 polynomial achieved optimal performance without overfitting.

---

9. Conclusion

Polynomial Regression significantly improved price prediction accuracy compared to linear regression by modeling non-linear relationships revealed during EDA. The model aligns well with domain understanding and provides reliable predictions for car prices.

---

10. Future Improvements

- Apply **Ridge Polynomial Regression** to reduce multicollinearity
- Perform **cross-validation**
- Add more real-world data
- Experiment with non-linear ML models (Random Forest, XGBoost)

---

11. Tools & Technologies

- **Python**
- **Pandas, NumPy**
- **Matplotlib, Seaborn**
- **Scikit-learn**
- **Kaggle Notebook Environment**