OUTLINE

Data Overview

The Car Price Dataset, obtained from Kaggle (https://www.kaggle.com/datasets/deepcontractor/car-price-prediction-challenge), contains data for over 19,000 cars, including features such as manufacturer, model, production year, mileage, engine volume, color, and price. The response variable is Price, while the explanatory variables include both categorical and continuous features.

Aim and Relevance

The goal of this project is to develop a model that predicts car prices based on features such as mileage, production year, and manufacturer. This analysis is relevant for consumers, dealerships, financial institutions, and the automotive market to make informed decisions, accurately predict car prices, assess car values, and identify pricing trends.

Data Exploration and Preparation

The dataset will be explored for missing values, outliers, and correlations. Key steps include:

- Reducing the dataset to approximately 3,000 observations for computational simplicity.
- Visualizing relationships between Price and other variables (scatter plots, box plots).
- Converting categorical variables to factors and addressing transformations (e.g., log for skewed data).
- Identifying anticipated trends, such as a negative correlation between mileage and price and higher prices for newer cars.

Model Building

Linear regression will be used as the primary model, evaluated with R-squared and Mean Squared Error (MSE).

Additional steps include:

- Checking for multicollinearity (VIF) and verifying assumptions (normality, homoscedasticity).
- Incorporating interaction and polynomial terms where necessary to capture non-linear relationships.
- Using ANOVA to test categorical variable significance.
- Exploring Generalized Linear Models (GLM) for model comparison.

Possible Questions

- What factors have the most significant impact on car prices?
- How do mileage and production year influence price trends across different manufacturers?
- Can interaction terms or alternative models improve prediction accuracy?

Results and Limitations

The regression model's coefficients will be analyzed to assess the impact of predictors like car age and mileage. Linear regression is simple and interpretable, but sensitive to outliers, which can distort predictions. Other limitations include potential issues with multicollinearity and assumption violations.

The full project, including the code and analysis, will be available on GitHub.