Diving Deep into Car Sales Analytics

Here a comprehensive statistical analysis is performed which important part of exploratory data analysis and data is preprocessing before we run the model in machine learning. The main insights from data are as given

Data Overview

- **Entries and Variables**: The dataset contains 157 rows (entries) and 16 columns (variables).
- **Data Types**: Among the variables, there are float64 types for numeric data like `Sales_in_thousands` and `Price_in_thousands`, and object types for categorical data like `Manufacturer` and `Model`.

Initial Data Handling

- **Missing Values**: Specific columns with missing data include:
- `__year_resale_value`: 36 missing values.
- `Price_in_thousands`: 2 missing values.
- `Engine_size`, `Horsepower`, `Wheelbase`, `Width`, `Length`, `Curb_weight`, `Fuel_capacity`: 1 missing value each.
- `Fuel_efficiency`: 3 missing values.
- **Duplicates**: The method for checking duplicates was mentioned, though specific numbers were not provided.

Statistical Analysis

- **Descriptive Statistics** (example metrics for `Sales_in_thousands`):
- Mean: 52.998
- Standard Deviation: 68.029
- Minimum: 0.110
- Maximum: 540.561
- **Outliers** (using Z-score with a threshold of 3):

- Example outliers: Sales_in_thousands had values as high as 540.561, significantly higher than the mean, which were identified and removed.

Visualization

- **Histograms**: These would show the distribution of variables like `Sales_in_thousands` ranging from 0.110 to 540.561 before outlier removal.
- **Boxplots**: These visualizations highlighted extreme outliers, particularly in sales data where values like 540.561 stand out far beyond typical sales figures.

Data Transformation

- **Standardization** (example from `Sales_in_thousands`):
- Before standardization, mean = 52.998, after standardization, mean = 0 (approximately).
- **Normalization** (example from `Sales_in_thousands`):
- Before normalization, range = 0.110 to 540.561. After normalization, range = 0 to 1.

Skewness and Distribution Checks

- **Skewness** (before cleaning):
- `Sales_in_thousands` skewness = 1.929842, indicating a right-tailed distribution.
- **Post-Transformation Distribution Check**:
- The distributions of `Sales_in_thousands` and other variables were rechecked to confirm a more normal distribution, reducing skewness and scaling data uniformly.

Detailed Distribution Check

- After cleaning and transforming the data, rechecking the distributions ensures:
- The mean and standard deviation for standardized data are 0 and 1, respectively, across all numeric fields.
- The min and max for normalized data are 0 and 1, respectively, confirming successful rescaling.