Benchmark Report

Data Preprocessing & Cleaning

Missing Values and Duplicates

The dataset contains 205 rows with 26 columns. There are missing values in the following columns:

- normalized losses: 20.0%
- num-of-doors: 0.98%
- bore: 1.95%
- stroke: 1.95%
- horsepower: 0.98%
- peak-rpm: 0.98%
- price: 1.95% There are no duplicates in the dataset.

Unique Values per Column

Each column has the following number of unique values:

- symboling: 6
- normalized-losses: 51
- make: 22
- fuel-type: 2
- aspiration: 2
- num-of-doors: 2
- body-style: 5
- drive-wheels: 3
- engine-location: 2
- wheel-base: 53
- length: 75
- width: 44
- height: 49
- curb-weight: 171
- engine-type: 7
- num-of-cylinders: 7
- engine-size: 44
- fuel-system: 8
- bore: 38
- stroke: 36
- compression-ratio: 32
- horsepower: 59
- peak-rpm: 23
- city-mpg: 29
- highway-mpg: 30
- price: 186

Cleaning Steps

Based on the analysis, the following cleaning steps are recommended:

1. **Fill missing values**: Impute missing values in columns normalized-losses, num-of-doors, bore, stroke, horsepower, peak-rpm, and price using suitable imputation techniques (e.g., mean, median, or regression imputation).
2. **Check data types**: Verify that the data types of each column match the expected data type (e.g., ensure that wheel-base and length are numeric).
3. **Handle outliers**: Identify and handle outliers in columns like price and curb-weight to prevent skewing of the data.
4. **Validate data**: Validate the data to ensure that it is consistent and accurate. For example, check that the make column only contains valid car brands.

These cleaning steps will improve the quality of the dataset, making it more suitable for analysis and modeling.

Statistical Analysis

The mean of symboling is approximately 0.83, indicating that most cars have a symboling of 1 or 2. The standard deviation is around 1.24, indicating some variation in symboling values.

The normalized-losses column has a mean of 122 and a standard deviation of 35.44, indicating some variation in normalized losses. The minimum value is 65, and the maximum value is 256.

The wheel-base column has a mean of 98.76 and a standard deviation of 6.02, indicating some variation in wheel base lengths. The minimum value is 86.6, and the maximum value is 120.9.

The length column has a mean of 174.05 and a standard deviation of 12.34, indicating some variation in car lengths. The minimum value is 141.1, and the maximum value is 208.1.

The width column has a mean of 65.91 and a standard deviation of 2.15, indicating some variation in car widths. The minimum value is 60.3, and the maximum value is 72.3.

The height column has a mean of 53.72 and a standard deviation of 2.44, indicating some variation in car heights. The minimum value is 47.8, and the maximum value is 59.8.

The curb-weight column has a mean of 2555.57 and a standard deviation of 520.68, indicating some variation in curb weights. The minimum value is 1488, and the maximum value is 4066.

The engine-size column has a mean of 126.91 and a standard deviation of 41.64, indicating some variation in engine sizes. The minimum value is 61, and the maximum value is 326.

The bore column has a mean of 3.33 and a standard deviation of 0.27, indicating some variation in bore values. The minimum value is 2.54, and the maximum value is 3.94.

The stroke column has a mean of 3.26 and a standard deviation of 0.32, indicating some variation in stroke values. The minimum value is 2.07, and the maximum value is 4.17.

The compression-ratio column has a mean of 10.14 and a standard deviation of 3.97, indicating some variation in compression ratios. The minimum value is 7, and the maximum value is 23.

The horsepower column has a mean of 104.26 and a standard deviation of 39.71, indicating some variation in horsepower values. The minimum value is 48, and the maximum value is 288.

The peak-rpm column has a mean of 5125.37 and a standard deviation of 479.34, indicating some variation in peak RPM values. The minimum value is 4150, and the maximum value is 6600.

The city-mpg column has a mean of 25.22 and a standard deviation of 6.54, indicating some variation in city MPG values. The minimum value is 13, and the maximum value is 49.

The highway-mpg column has a mean of 30.75 and a standard deviation of 6.89, indicating some variation in highway MPG values. The minimum value is 16, and the maximum value is 54.

The price column has a mean of 13207.1 and a standard deviation of 7947.07, indicating some variation in prices. The minimum value is 5118, and the maximum value is 45400.

Next Steps

Based on the summary statistics and distributions, the next steps could be:

- Visualize the distributions of numeric columns using histograms or density plots to better understand the shapes of the distributions.
- Investigate the relationships between columns using correlation analysis or scatter plots.
- Perform feature engineering to transform or create new features that may be more informative for modeling.
- Handle missing values by imputing or dropping them based on the specific requirements of the project.
- Split the data into training and testing sets for model evaluation and validation.
- Develop and evaluate machine learning models using the features and target variable(s) of interest.

Encoding Categorical Variables:

- Identify categorical variables like 'make', 'fuel-type', 'aspiration', 'num-of-doors', 'body-style', 'drive-wheels', 'engine-location', 'engine-type', 'fuel-system'.
- Use techniques like one-hot encoding or label encoding to convert categorical variables into numerical form.

Exploratory Data Analysis

This section aims to explore the given dataset to gain insights and identify potential patterns, relationships, and anomalies.

Visualization Ideas

1. **Scatter Plots**: Visualize the relationships between numerical variables such as wheel-base, length, width, height, and curb-weight to explore correlations and potential outliers.
2. **Bar Charts**: Examine the distribution of categorical variables like make, fuel-type, aspiration, body-style, drive-wheels, and engine-location to identify the most common categories.
3. **Histograms**: Analyze the distribution of numerical variables like engine-size, bore, stroke, compression-ratio, horsepower, and peak-rpm to identify patterns and outliers.
4. **Box Plots**: Compare the distribution of numerical variables across different categorical groups, such as make or fuel-type.

5. **Heatmap**: Visualize the correlation between numerical variables using a heatmap to identify strong relationships and potential multicollinearity.

Correlation Checks

1. **Pairwise Correlation**: Calculate the correlation between numerical variables to identify strong relationships and potential multicollinearity.
2. **Correlation Matrix**: Create a correlation matrix to visualize the correlations between numerical variables.

Outlier Detection

1. **Z-Score Method**: Identify outliers using the Z-score method for numerical variables.
2. **Modified Z-Score Method**: Use the modified Z-score method to detect outliers in variables with non-normal distributions.
3. **Visual Inspection**: Visually inspect scatter plots and histograms to identify potential outliers and anomalies.

By applying these EDA techniques, we can gain a deeper understanding of the dataset, identify potential issues, and inform subsequent modeling and analysis steps.

Key Takeaways from EDA

1. price is likely right-skewed, meaning there are many cheaper cars and fewer expensive ones.
2. engine-size and horsepower should be strongly correlated with price.
3. drive-wheels and fuel-type may have a big impact on price.
4. city-mpg and highway-mpg are likely strongly correlated.
5. body-style and num-of-doors may have no strong relationship with price.

Machine Learning Algorithm Suggestions

Supervised Learning Methods

Based on the dataset, the following supervised learning methods can be applied:

Regression

Price Prediction: With the price column as the target variable, regression models like Linear Regression, Decision Trees, Random Forest, and Gradient Boosting can be used to predict the price of a car based on its features.

Classification

Car Classification: By using the make or body-style column as the target variable, classification models like Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines can be used to classify cars into different categories.

Unsupervised Learning Methods

The following unsupervised learning methods can be applied to this dataset:

Clustering

Car Segmentation: Clustering algorithms like K-Means, Hierarchical Clustering, and DBSCAN can be used to segment cars into different groups based on their features, helping to identify patterns or structures in the data.

Dimensionality Reduction

Feature Reduction: Dimensionality reduction techniques like PCA, t-SNE, and Autoencoders can be used to reduce the number of features in the dataset, helping to visualize and analyze the data more efficiently.

Anomaly Detection

Outlier Detection: Anomaly detection algorithms like Local Outlier Factor, Isolation Forest, and One-Class SVM can be used to detect outliers or unusual patterns in the data, which can help identify unusual or rare car features.

Feature Engineering

In this section, we will discuss various feature creation, transformation, and selection approaches to enhance the quality of the dataset and improve model performance.

Feature Creation

Engine Performance Index: Calculate an engine performance index by combining horsepower, peak-rpm, and engine-size to create a comprehensive measure of engine performance.
Fuel Efficiency Index: Create a fuel efficiency index by combining city-mpg and highway-mpg to provide a single metric for fuel efficiency.
Vehicle Size Index: Calculate a vehicle size index by combining length, width, and height to provide a single metric for vehicle size.

Feature Transformation

Normalize numerical features: Normalize numerical features such as wheel-base, length, width, height, curb-weight, and price to ensure consistent scaling.
Encode categorical features: Encode categorical features such as make, fuel-type, aspiration, body-style, drive-wheels, engine-location, engine-type, num-of-cylinders, and fuel-system using one-hot encoding or label encoding to convert them into numerical features.

Feature Selection

Correlation analysis: Perform correlation analysis to identify highly correlated features and remove redundant features.
Recursive feature elimination: Use recursive feature elimination to select the most informative features that contribute to the target variable.
Permutation importance: Use permutation importance to evaluate the importance of each feature and select the top-ranked features.

By applying these feature engineering techniques, we can create new features that provide additional insights, transform features into a more suitable format for modeling, and select the most informative features to improve model performance.

Removing Highly Correlated Features
If engine-size and curb-weight have a correlation > 0.9, we drop one to avoid redundancy.

Model Optimization (Hyperparameter Tuning)

Once we have engineered the best features, we need to fine-tune the models to improve performance.
Hyperparameter Tuning Techniques
Grid Search (GridSearchCV) - Tries all possible hyperparameter combinations, best when we have few parameters.
Random Search (RandomizedSearchCV) - Randomly selects hyperparameters
from a range, best for fast tuning.
Bayesian Optimization - Learns which hyperparameters are best over time, best for deep learning & complex models