# Linear Regression Analysis for Predicting Used Car Prices

#### **1. Dataset Description**

The dataset used for this analysis provides detailed information about used cars, their features, and corresponding prices. It serves as a comprehensive repository to analyze the factors that influence the pricing of used vehicles. Key characteristics of the dataset include:

* **Diversity of Features**: The dataset has both numerical and categorical features, such as manufacturer, state, year, and odometer, which are crucial for modeling.
* **Target Variable**: The price of the car is the target variable we aim to predict.
* **Data Challenges**:
  + **Missing Values**: Certain features, such as condition, cylinders, and size, have missing values that required preprocessing to handle effectively.
  + **Outliers**: The price feature contains extreme values that could skew the model, necessitating outlier removal.
  + **Large variety of unique values Features**: Features like model and region include a large number of unique values, making them computationally expensive to encode.

This dataset offers valuable insights into the used car market, making it a robust starting point for price prediction.

## 2. Aim of the Analysis

The primary objective of this project is to develop a reliable linear regression model to predict the price of a used car based on its features. This predictive model aims to:

* Assist car dealerships in fine-tuning their inventory pricing strategies.
* Provide actionable insights into the factors that most influence car pricing.
* Serve as a foundation for further exploration using advanced machine learning techniques.

Through careful preprocessing and model development, we aim to balance interpretability and accuracy, ensuring the model delivers meaningful insights.

## 3. Features Description

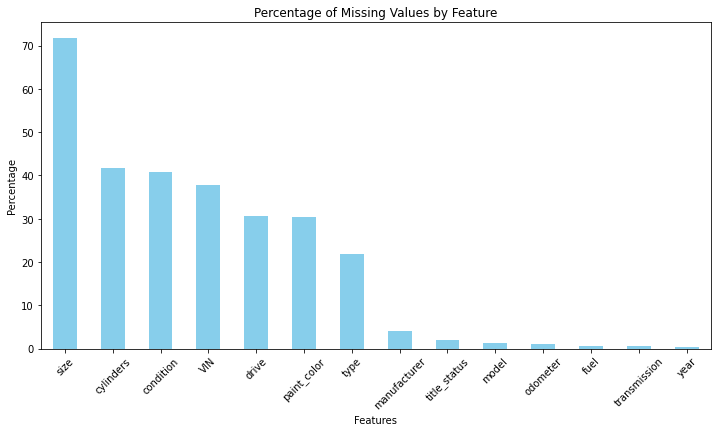
The dataset includes the following significant features:

1. **Categorical Features**:
   1. manufacturer: Represents the car's brand, such as Ford, Toyota, or Honda. This feature is crucial as brand reputation often impacts pricing.
   2. state: Indicates the state where the car is being sold, reflecting potential regional price variations.
2. **Numerical Features**:
   1. odometer: The car's mileage, measured in miles. Higher mileage typically decreases the car's value.
   2. year: The manufacturing year of the car. Recent models are generally priced higher.
3. **Target Variable**:
   1. price: Represents the car's selling price, which is the variable we aim to predict.

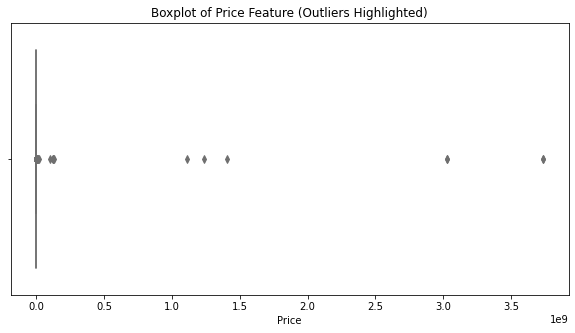
#### **4. Data Cleaning and Preprocessing**

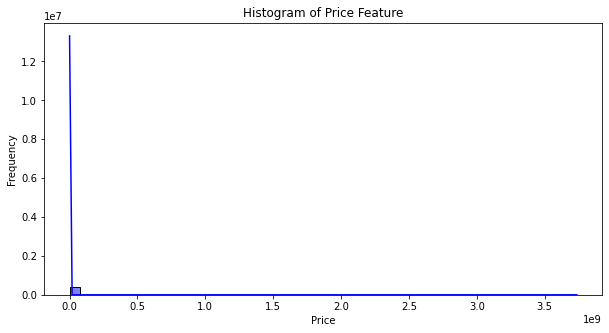
To ensure the dataset was suitable for modeling, the following preprocessing steps were implemented:

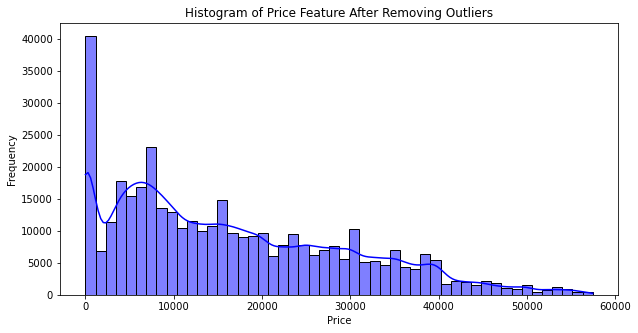
1. **Handling Missing Values**:
   1. Features with more than 5% missing values were dropped (e.g., cylinders, condition).
   2. Rows with missing values in retained features were removed to maintain data integrity.



1. **Outlier Removal**:
   1. Extreme outliers in the price feature were identified and removed using the Interquartile Range (IQR) method. This helped reduce skewness in the data.







3. **Feature Reduction**:

* Large number of value features like model and region were excluded due to their computational burden during one-hot encoding.

4. **Feature Encoding**:

* Categorical features with a manageable number of categories were one-hot encoded.

### **Categorical Features One-Hot Encoded**

The following features were one-hot encoded:

1. **Fuel Type (fuel)**:
   1. Categories: gas, diesel, electric, hybrid, other.
   2. These represent the type of fuel the car uses, and since there are only a few distinct categories, this feature was well-suited for one-hot encoding.
2. **Title Status (title\_status)**:
   1. Categories: clean, rebuilt, salvage, lien, parts only, missing.
   2. This feature describes the legal condition of the car's title. Its small number of categories made it a good candidate for one-hot encoding.
3. **Transmission (transmission)**:
   1. Categories: automatic, manual, other.
   2. Transmission type is a key factor influencing car price, and with only three categories, one-hot encoding was appropriate.

### **Why One-Hot Encoding Was Used for These Features**

* **Low Variety**: The manageable number of unique values (categories) meant that one-hot encoding did not lead to an excessively large number of new columns.
* **Interpretability**: One-hot encoding preserves the interpretability of categorical data by assigning separate binary columns for each category.
* **Impact on Price**: These features are relevant to car pricing. For example:
  + Fuel type influences operating costs.
  + Title status impacts the car's marketability.
  + Transmission type affects demand, as manual cars might have a niche market.

By one-hot encoding these features, the categorical information was retained in a format suitable for the linear regression model while avoiding issues like overloading the dataset with too many additional columns. Shown below is the correlation of the final set of features with the Price



#### **5. Model Fitting Process**

The cleaned and preprocessed dataset was used to train and validate a linear regression model. The following steps were followed:

1. **Data Scaling**:
   1. Standard scaling was applied to numerical features and the target variable to normalize their ranges and improve the regression model's performance.
2. **Train-Test Split**:
   1. The dataset was split into training (80%) and testing (20%) sets to evaluate the model's generalizability.
3. **Model Training**:
   1. A linear regression model was fitted to the training data. This choice of model was motivated by its simplicity and interpretability.
4. **Cross-Validation**:
   1. A 5-fold cross-validation was performed to assess the model's stability and ensure it generalizes well to unseen data.
5. **Hyperparameter Tuning**:
   1. Grid search was used to identify optimal alpha values for Ridge and Lasso regression.

#### **Errors Analysis**

1. **Root Mean Squared Error (RMSE)**:
   1. RMSE measures the average magnitude of prediction errors, penalizing larger errors more heavily than smaller ones.
   2. For the linear regression model, the RMSE on the scaled data was approximately **1.02 units**. This means that the average deviation of predictions from actual values, scaled by the standard deviation of the target variable, is about 1.02.
2. **Mean Absolute Error (MAE)**:
   1. MAE measures the average magnitude of errors without disproportionately emphasizing larger errors.
   2. For the same model, the MAE on the scaled data was approximately **0.68 units**, indicating that the typical error in predictions is about 0.68 when measured in standard deviations.
3. **Cross-Validation Errors**:
   1. The RMSE observed during cross-validation was slightly higher than the test RMSE, indicating some variation in performance across different folds. This suggests that while the model is consistent, it may still miss certain nuances in the data.

#### **Interpretation of Errors**

* The observed errors are reasonable given the variability in used car prices, which can be influenced by factors not captured in the dataset (e.g., car condition, accident history).
* The higher RMSE compared to MAE highlights that a few cars with unusually high or low prices might have significantly influenced the error metrics, despite outlier removal.

To better understand how the model performs, the predicted prices were compared with the actual prices:

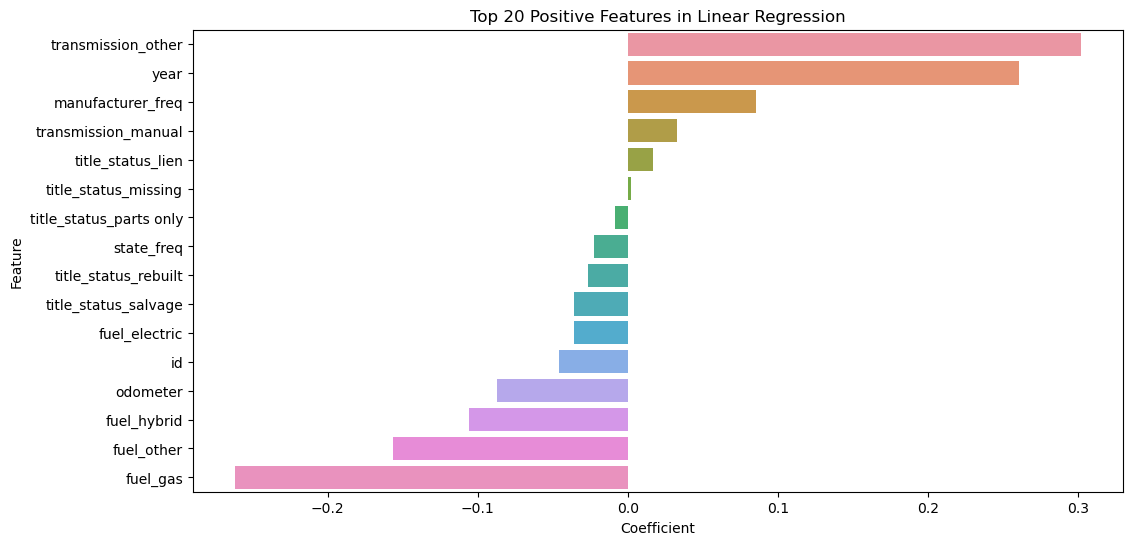
1. **Accuracy of Predictions**:
   1. The majority of the predictions were close to the actual prices, with errors typically falling within a range of ±1 standard deviation.
   2. For instance, cars with attributes like recent manufacturing years and low mileage were accurately predicted to have higher prices.
2. **Discrepancies in Predictions**:
   1. Some discrepancies were observed, especially for cars with extreme prices (either very high or very low). These could be attributed to:
      1. Features not captured in the dataset, such as condition, accident history, or additional customizations.
      2. Limitations of linear regression in capturing non-linear relationships.
3. **Insights from Outliers**:
   1. Even after outlier removal, a few test data points deviated significantly from the predicted prices. These cases highlight potential areas for improvement in the model, such as including additional features or using non-linear modeling techniques.

#### **6. Coefficient Analysis**

The linear regression model's coefficients provide valuable insights into the influence of features on car pricing:

* **Positive Coefficients**:
  + Features such as the year of manufacture showed positive coefficients, indicating that newer cars are priced higher.
* **Negative Coefficients**:
  + Features like odometer had negative coefficients, reflecting the depreciation associated with higher mileage.

This analysis empowers stakeholders to prioritize features that significantly impact pricing, such as focusing on cars with low mileage or recent manufacturing years.



### **Recommendations to the Used Car Dealership**

Based on the analysis of the dataset, the development of the linear regression model, and the observations made from the model's coefficients and predictions, the following recommendations are provided to help the dealership optimize its inventory and pricing strategies:

#### **1. Key Drivers of Car Value**

The following features were found to have the most significant impact on used car prices:

1. **Manufacturing Year**:
   1. **Observation**: Newer cars (higher year values) consistently commanded higher prices.
   2. **Recommendation**: Focus on stocking cars manufactured in recent years (e.g., the last 5–10 years). These vehicles are more desirable and retain better resale value.
2. **Odometer (Mileage)**:
   1. **Observation**: Cars with lower mileage (odometer) were valued higher, as they are perceived to be in better condition and have a longer remaining lifespan.
   2. **Recommendation**: Highlight low-mileage vehicles in your inventory and emphasize this attribute in marketing materials.
3. **Brand (Manufacturer)**:
   1. **Observation**: Brands with strong reputations (e.g., Toyota, Honda, BMW) tend to retain higher resale value.
   2. **Recommendation**: Prioritize popular and reliable brands in your inventory. Additionally, educate customers on the value retention of certain brands to build trust and encourage purchases.
4. **State (Regional Pricing)**:
   1. **Observation**: Regional factors (state) affected pricing, likely due to local demand and cost-of-living differences.
   2. **Recommendation**: Adjust pricing strategies based on local market conditions. Consider transferring inventory between states to balance supply and demand.

#### **2. Features That Should Be Highlighted**

From the analysis, certain features were less important predictors of price but may still appeal to consumers:

1. **Fuel Type**:
   1. **Observation**: Cars with hybrid or electric fuel types tend to have niche demand and, in some cases, higher value.
   2. **Recommendation**: Emphasize fuel efficiency as a selling point, particularly for electric and hybrid vehicles, to attract environmentally conscious buyers.
2. **Transmission Type**:
   1. **Observation**: Automatic transmissions were more common and likely more desirable for general buyers.
   2. **Recommendation**: Ensure automatic transmission vehicles are well-represented in the inventory. For manual transmission cars, target specific buyer groups (e.g., enthusiasts).

#### **3. Inventory Management**

To improve inventory selection and pricing strategies:

1. **Focus on Low Mileage and Newer Cars**:
   1. Highlight cars with low odometer readings and recent manufacturing years as these are the strongest predictors of higher prices.
2. **Exclude Older or High-Mileage Vehicles**:
   1. While these cars may sell at lower prices, they could take longer to move and generate less profit. Limit these vehicles unless they belong to a highly valued brand or have niche demand.
3. **Optimize Inventory by Brand and Region**:
   1. Use regional pricing trends to allocate inventory strategically across dealerships in different states. For example, prioritize premium brands in affluent regions.

#### **7. Model Improvements and Feature Handling**

To enhance the model further, the following steps are recommended:

1. **Clustering for Dimensionality Reduction**:
   1. Apply K-Means clustering to group Large variety of unique values features like model into clusters, reducing dimensionality while retaining key information.
2. **Feature Engineering**:
   1. Derive new features such as the age of the car (current\_year - year) or average mileage per year (odometer / age).
3. **External Data Integration**:
   1. Incorporate additional data sources, such as market trends, fuel efficiency, or accident history, to improve predictive accuracy.