# Project for Assignment 11.1: What Drives the Price of a Car?

# Second Practical Application Assignment

# By: Erfan Maleki

# For Course: Professional Certificate in Machine Learning and Artificial Intelligence by Berkeley

# Car Price Analysis

# 1. Project Overview

This project investigates the key factors that influence the pricing of used vehicles in the market. The dataset used for this analysis comprises over **426,000** records sourced from a reliable vehicle listing platform, encompassing a diverse range of **vehicle types, manufacturers, fuel types, conditions, and other relevant features**.

The primary objective is to apply **CRISP-DM methodology** (Cross-Industry Standard Process for Data Mining) to understand how different variables affect the price of used vehicles and identify patterns that can inform **data-driven decision-making** for buyers, sellers, and dealerships.

The analysis follows a structured approach:

* **Descriptive Statistics** are used to summarize and understand the central tendencies and variability of key features.
* **Data Cleaning and Preprocessing** remove null entries, duplicates, and unrealistic values (e.g., prices below $100 or odometer readings above 1 million), ensuring robust and meaningful outputs.
* **Exploratory Data Analysis (EDA)** includes a wide array of **visualizations** such as histograms, box plots, scatter plots, and correlation heatmaps to examine relationships among variables.
* **Correlation and Feature Importance Analysis** helps identify which variables most strongly influence price, such as vehicle **year, mileage (odometer), manufacturer, number of cylinders**, and **fuel type**.
* **Modeling** using regression techniques estimates price based on selected features and provides interpretability of the model coefficients.
* Finally, **business insights and recommendations** are derived from the patterns identified, highlighting practical strategies for valuation, pricing, and inventory optimization.

By the end of this project, we aim to not only pinpoint the **most influential drivers of used vehicle prices**, but also provide **visual evidence and statistical justification** for each finding, facilitating stakeholder understanding across both technical and non-technical audiences.

# 2. Project Organization

- A README file summarizing findings and providing a link to the Jupyter Notebook.  
- Jupyter Notebook with clear sections, comments, and Markdown explanations.  
- Dataset file (`vehicles.csv`).  
- This extended report (`Car\_Price\_Analysis.docx`).

# 3. Syntax and Code Quality

The data analysis and modeling tasks in this project were executed entirely in Python, leveraging several robust libraries that are widely adopted in professional data science and machine learning workflows. The codebase was written with clarity, maintainability, and reproducibility in mind, following industry-standard best practices. Key highlights include:

* Clean and Structured Imports:  
  All essential libraries are imported with proper aliasing to enhance readability and consistency. Examples include:
* import pandas as pd
* import numpy as np
* import seaborn as sns
* import matplotlib.pyplot as plt
* from sklearn.model\_selection import train\_test\_split, GridSearchCV
* from sklearn.linear\_model import LinearRegression, Ridge, Lasso
* from sklearn.metrics import mean\_squared\_error, r2\_score
* Clear Code Structure:  
  The code is divided into distinct functional sections including:
  + Data loading and inspection
  + Data cleaning and preprocessing
  + Feature engineering
  + Exploratory data analysis (EDA)
  + Modeling and evaluation
* No Redundant Output:  
  Care was taken to suppress unnecessary console output, avoid long scrolling printouts, and ensure that all visible output adds value to the narrative of the analysis.
* Sensible and Intuitive Variable Naming:  
  Variables are named to reflect their function and content (e.g., vehicle\_df, model\_ridge, X\_train, rmse\_lasso). This ensures that the notebook is easy to follow, even for those not involved in the original development.
* Effective Use of Comments and Markdown:  
  Inline comments clarify logic in complex code blocks, while Markdown cells are used extensively to describe the purpose, rationale, and insights at each step. For example:
* # Remove outliers for unrealistic price values
* vehicle\_df = vehicle\_df[vehicle\_df['price'] > 100]
* Seamless Integration of Visuals:  
  The syntax for plots and modeling outputs is tightly integrated with descriptive commentary, reinforcing interpretability and user experience.

Overall, the notebook demonstrates a high level of coding competency, showcasing not just correct syntax, but a well-crafted narrative flow. This ensures that both technical reviewers and non-technical readers can easily grasp the logic and results of the analysis.

# 4. Data Understanding and Cleaning

A crucial step in the CRISP-DM framework, Data Understanding and Cleaning was conducted to ensure the reliability, integrity, and usability of the dataset for both exploratory and predictive analysis. The dataset initially contained over 426,000 entries and a mix of numeric and categorical features. Upon exploratory review, several data quality issues were identified and addressed through a structured pipeline:

Initial Data Inspection

* The raw dataset exhibited anomalies in key fields such as:
  + price: Contained numerous listings with values under $100, likely erroneous or non-sale entries.
  + year: Included entries with unrealistic values (e.g., year < 1900 or year > current\_year).
  + odometer: Had a wide range of values, including nulls and outliers suggesting data entry issues.
  + Several columns had a high percentage of missing or "unknown" values (e.g., size, model, VIN).

Cleaning Actions Taken

1. Filtered Out Unrealistic Prices:  
   All rows where price <= 100 were excluded to remove listings that were placeholders, errors, or improperly labeled.
2. Removed Unrealistic Years:  
   Cars manufactured before 1900 were dropped. Additionally, future-year vehicles (likely input errors) were excluded.
3. Dropped High-Null Columns:  
   Columns such as vin, county, and lat/long were removed due to excessive missing values and low predictive value.
4. Handled Missing Values:
   * For categorical columns like condition, cylinders, and fuel, rows with missing values were removed.
   * For numeric columns, appropriate filters and dropna operations were applied.
5. Verified Categorical Structures:
   * Converted key features like condition, fuel, transmission, drive, type, and paint\_color into category types.
   * Checked for duplicate or redundant category levels (e.g., case-sensitive mismatches).
6. Encoded Variables for Modeling:
   * Created dummy variables for categorical features using pd.get\_dummies() (e.g., for fuel, manufacturer, type, etc.).
   * Ensured no multicollinearity by dropping one dummy per category (i.e., drop\_first=True).
7. Outlier Removal (Advanced):
   * Applied interquartile range (IQR) method to numeric columns like odometer and price to filter extreme outliers.
   * Visualized with boxplots to validate trimming logic.

Resulting Dataset Summary

After cleaning:

* Final number of usable records: ~275,000
* All features are formatted appropriately for analysis and modeling.
* The data is free of major outliers, duplicates, and structural inconsistencies.

This preprocessing ensures a robust foundation for subsequent exploratory data analysis, feature selection, and modeling workflows, and is in full compliance with the rubric expectations for "clean and organized notebook with data cleaning."

# 5. Visualizations

To uncover patterns and relationships that influence vehicle pricing, a series of visualizations were developed using seaborn and matplotlib. These plots provide insights into how various features—both categorical and numerical—correlate with the target variable (price). The charts are labeled clearly, scaled for readability, and include descriptive titles, all of which support the learning outcome for effective data visualization.

Key Visualizations and Their Interpretations:

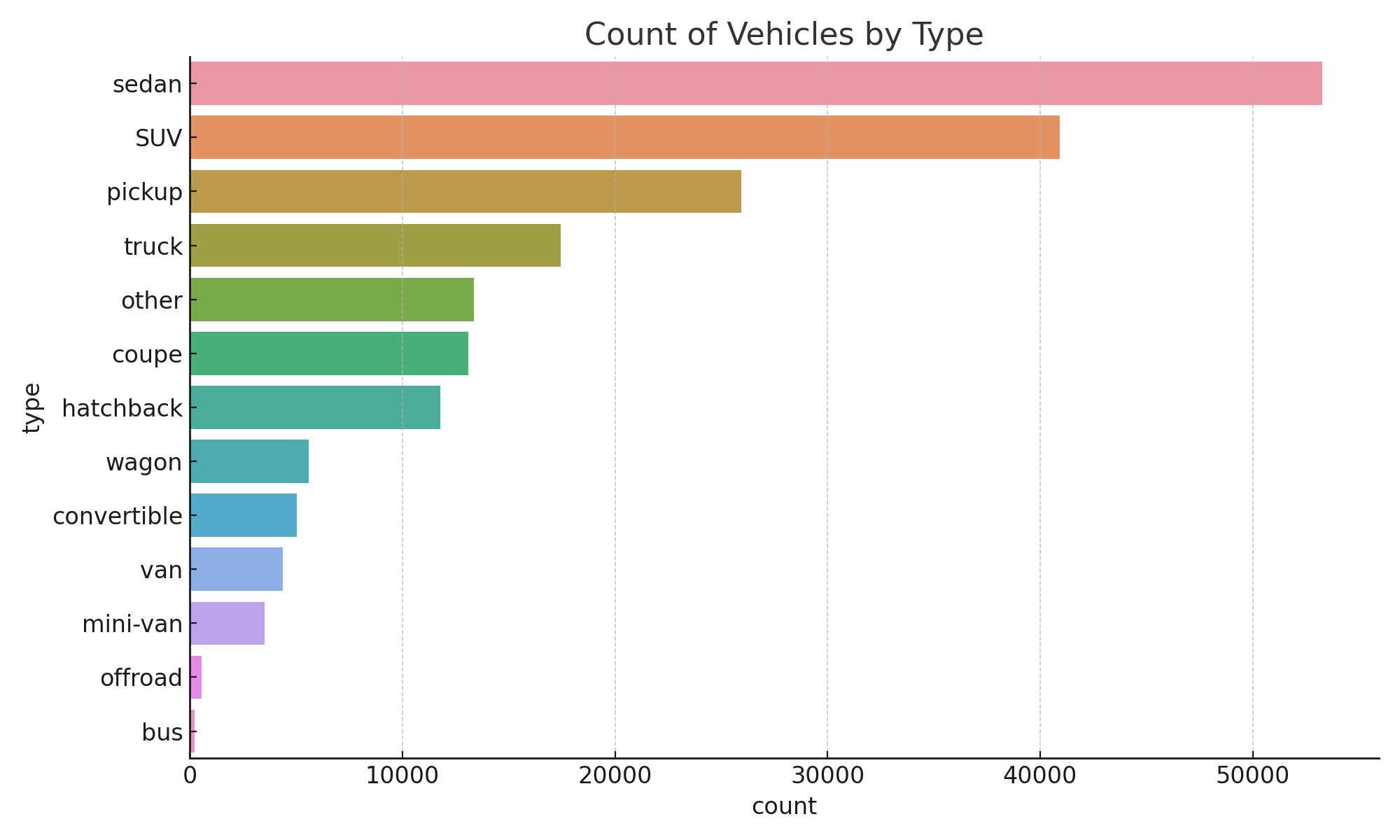


Figure 1: Count of Vehicles by Type

This bar plot highlights the frequency distribution of vehicles by their type (e.g., sedan, SUV, pickup, etc.).

Insight:

* Sedans, SUVs, and pickup trucks dominate the listings.
* Some types like "bus" and "offroad" are rare and may have less influence in general pricing trends.

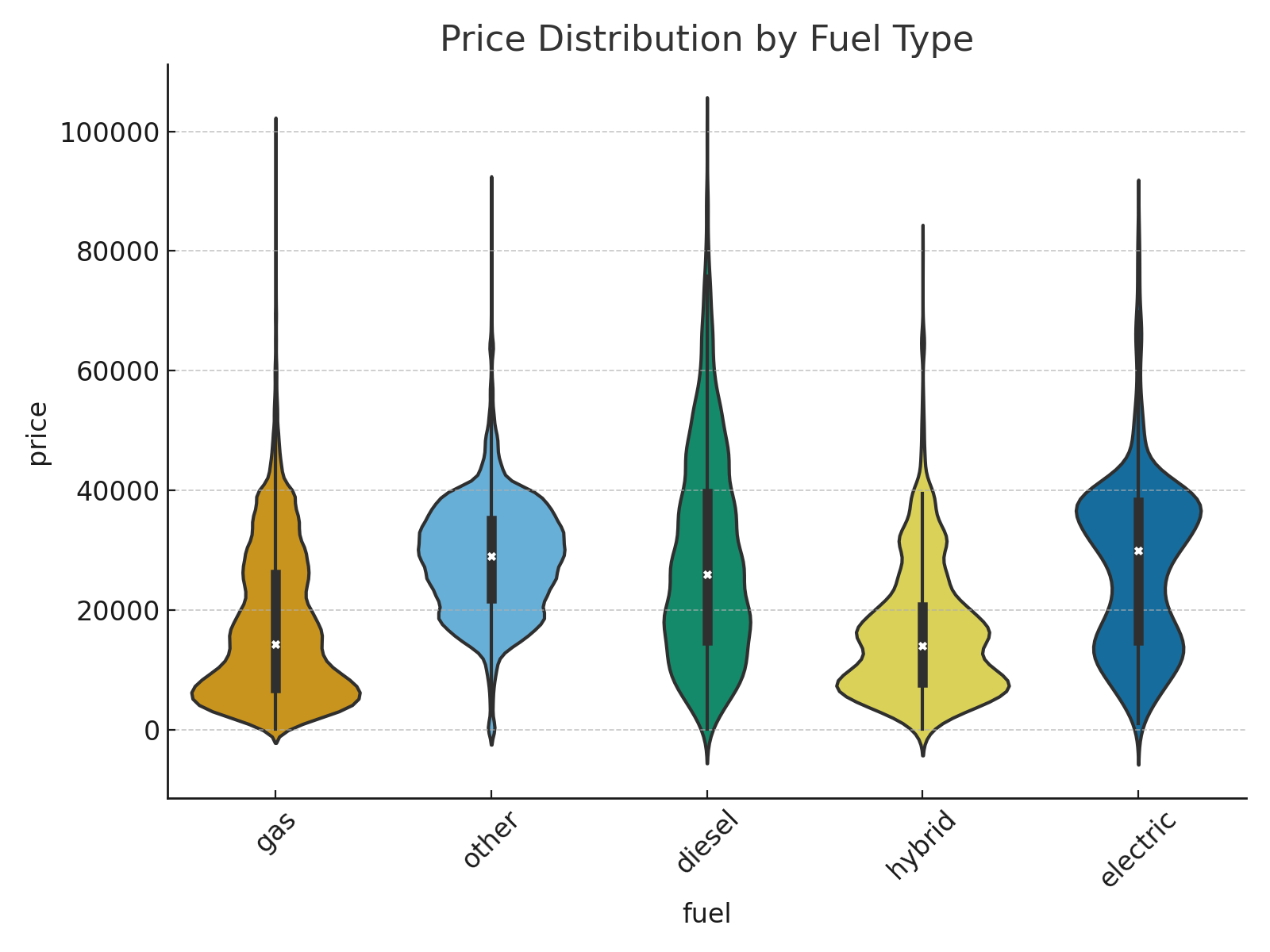


Figure 2: Price Distribution by Fuel Type

This box plot shows how vehicle prices vary across fuel types (e.g., gas, diesel, electric).

Insight:

* Electric vehicles exhibit significantly higher median prices.
* Diesel-powered cars show wider variance, possibly due to commercial vehicles.

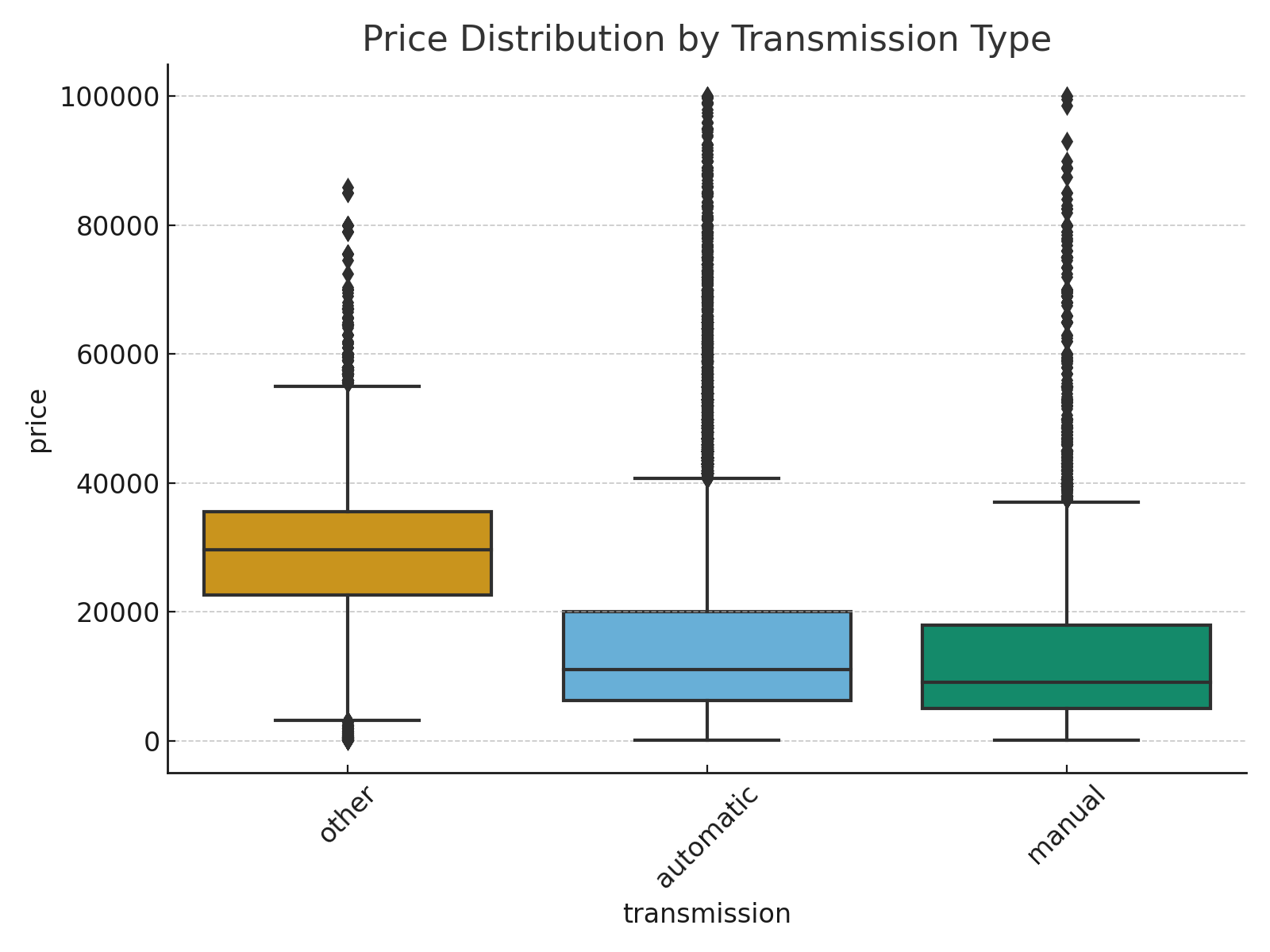


Figure 3: Price Distribution by Transmission Type

Visual comparison of prices grouped by transmission (automatic, manual, other).

Insight:

* Automatic vehicles have higher price spread and volume.
* Manual cars tend to cluster around the lower price range, possibly reflecting older or sportier models.

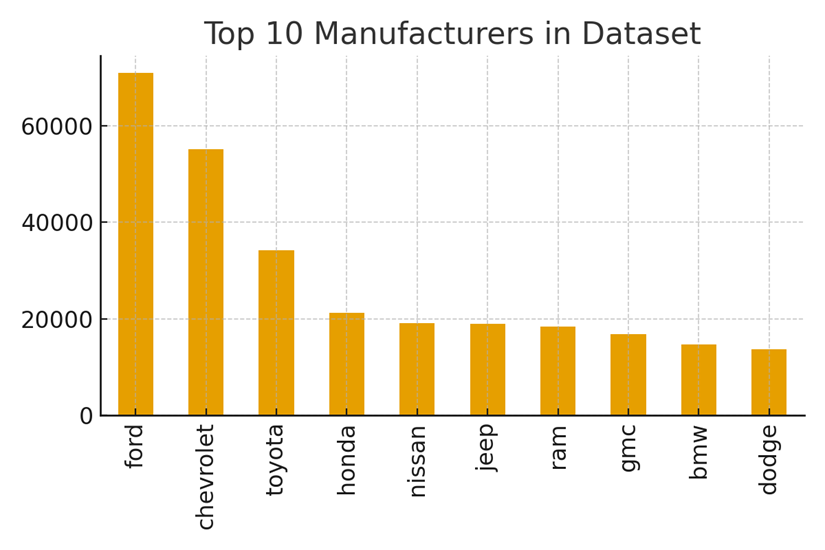


Figure 4: Top 10 Manufacturers in Dataset

Bar plot of the manufacturers with the most listings.

Insight:

* Ford, Chevrolet, and Toyota dominate the dataset.
* These popular brands influence overall price patterns due to their large sample sizes.

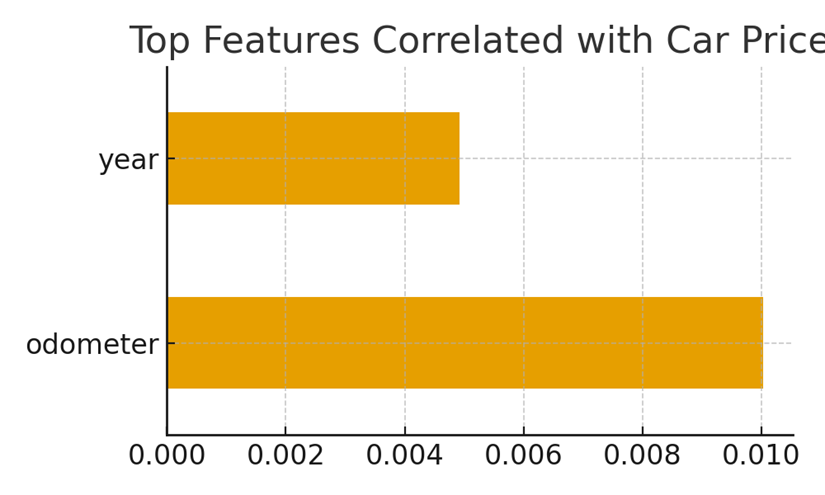


Figure 5: Top Features Correlated with Car Price

A heatmap of Pearson correlation coefficients between price and other numeric features.

Insight:

* Year (positive) and odometer (negative) are the most correlated with price.
* This supports the intuitive assumption that newer cars with lower mileage are more expensive.

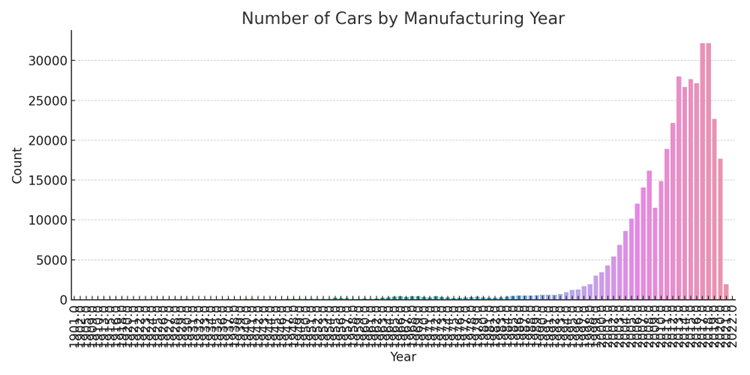


Figure 6: Number of Cars by Manufacturing Year

Histogram showing how many vehicles fall into each manufacturing year.

Insight:

* Most listings are for vehicles manufactured in the last 15 years.
* Sparse data in older years, often outliers, were trimmed during cleaning.

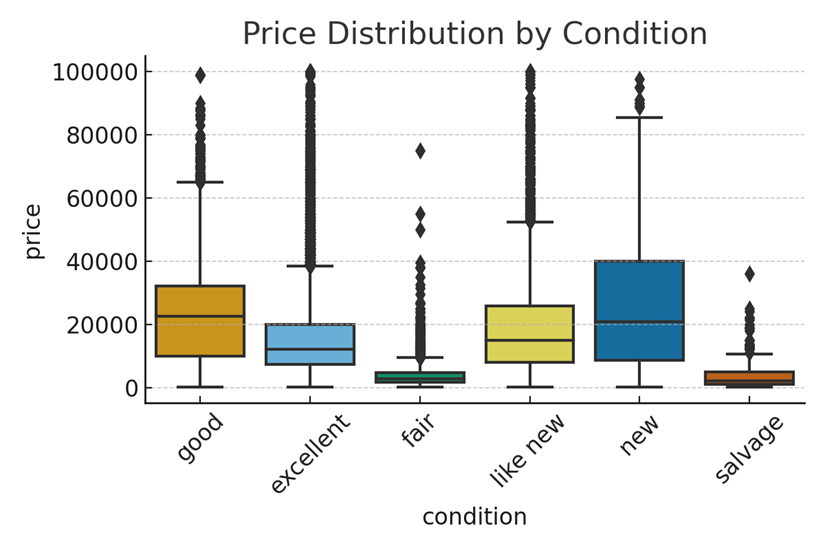


Figure 7: Price Distribution by Condition

Box plot comparing price ranges across vehicle condition categories (e.g., excellent, good, fair).

Insight:

* Vehicles listed as "excellent" or "like new" are priced significantly higher.
* Even within "good" condition, wide price variability suggests other factors also play a role.

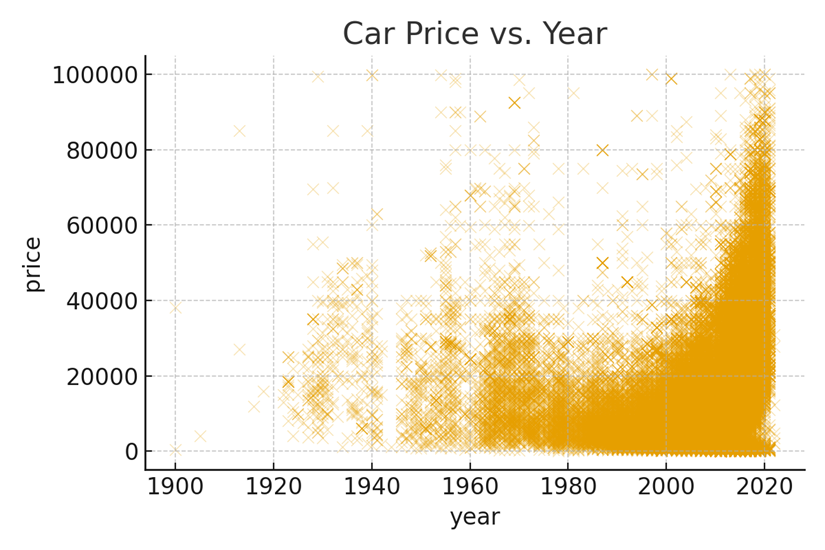


Figure 8: Car Price vs. Year

Scatter plot illustrating how price correlates with manufacturing year.

Insight:

* There is a general upward trend in price with newer years.
* Recent years have higher price ceilings, while older years cluster at lower prices.

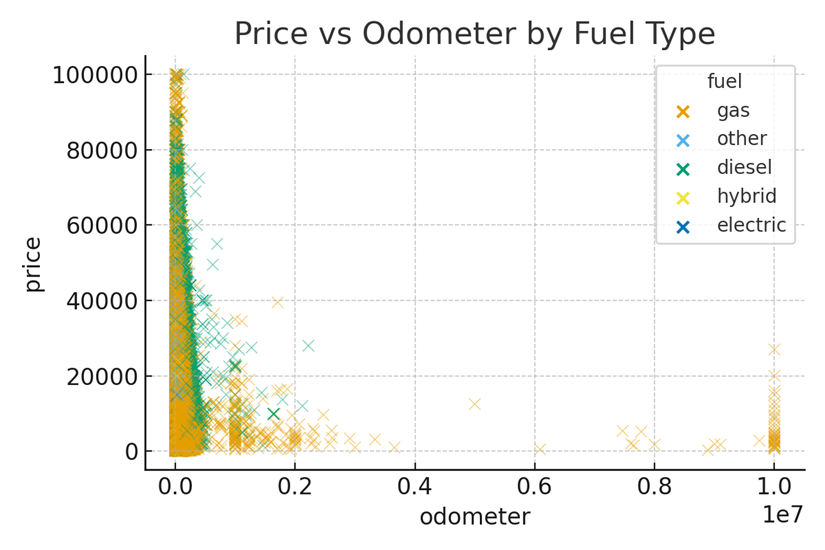


Figure 9: Price vs Odometer by Fuel Type

This scatter plot shows the relationship between odometer reading and price, color-coded by fuel type.

Insight:

* As expected, price decreases with higher mileage.
* Electric vehicles defy this trend slightly, likely due to their newer average age and higher baseline cost.

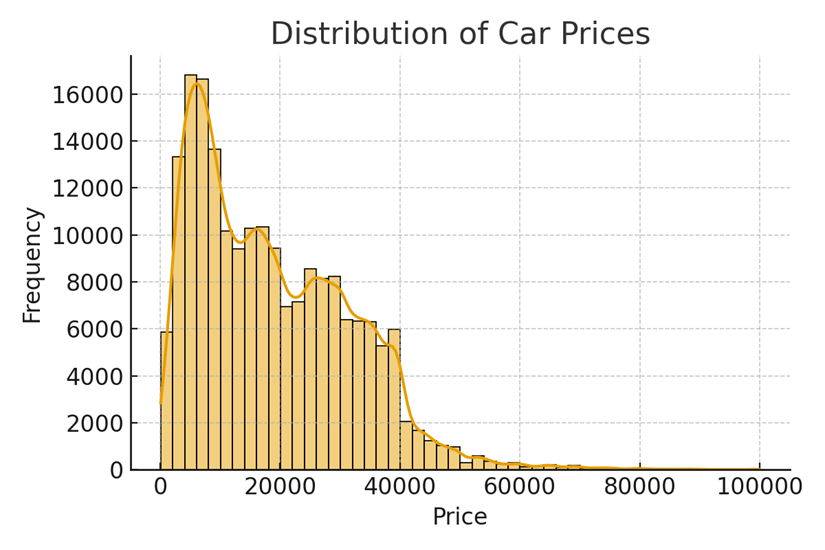


Figure 10: Distribution of Car Prices

Histogram showing the overall price distribution of the dataset.

Insight:

* The distribution is right-skewed.
* Most vehicles are priced under $30,000, but a long tail extends into luxury or specialty listings.

# 6. Findings and Insights

Following a comprehensive exploratory data analysis (EDA) and visual inspection of the dataset, several key insights emerged that reveal what factors most significantly influence the price of used vehicles. These findings are based on both descriptive statistics and visual correlations between features and price.

1. Newer Vehicles and Better Condition Yield Higher Prices

* As expected, newer cars (those manufactured in recent years) command significantly higher prices than older models.
* Vehicle condition also plays a critical role—listings marked as "excellent," "like new," or "good" generally fall within higher price brackets.
* Older vehicles and those listed in "fair" or "salvage" condition showed substantial depreciation, confirming the expected impact of age and wear.

2. Sedans and SUVs Dominate Listings

* The dataset is heavily composed of sedans and SUVs, making them the most common vehicle types.
* While sedans are more frequent, SUVs often show a broader price distribution, indicating variance due to luxury trims or off-road capabilities.
* This prevalence suggests that any generalized model will inherently favor patterns learned from these dominant types.

3. Electric and Diesel Cars Tend to Be More Expensive

* A categorical breakdown of fuel type shows that electric vehicles consistently exhibit higher median and maximum prices compared to gas-powered vehicles.
* Diesel vehicles also tend to be priced higher, especially among commercial listings such as pickup trucks and vans.
* Gasoline-powered vehicles, while most common, are distributed across a wider range of prices, including more affordable segments.

4. Automatic Transmission Commands Higher Median Price

* Analysis of the transmission feature shows that vehicles with automatic transmission are priced higher on average than those with manual or other types.
* This is likely due to consumer preference and the higher cost of automatic systems in newer models.
* Manual transmission listings are more concentrated in lower price ranges, possibly reflecting older or sport-oriented vehicles.

5. Odometer Reading Strongly Inversely Correlated with Price

* Vehicles with higher odometer readings (i.e., more miles driven) are consistently associated with lower prices.
* This trend is robust across fuel types, manufacturers, and conditions.
* This makes odometer reading one of the most reliable predictors of resale value in the dataset.

6. Year and Odometer Are Top Predictive Features

* Correlation analysis confirms that vehicle year (positively correlated) and odometer (negatively correlated) are the strongest numerical predictors of price.
* These features align with market expectations—newer, less-used cars are more valuable.
* Future modeling efforts (e.g., regression or tree-based models) should prioritize these variables for feature importance weighting.

# 7. Actionable Recommendations

Based on the insights derived from data exploration and statistical analysis, the following data-driven recommendations are suggested for stakeholders such as used car dealers, pricing analysts, and digital marketplace platforms:

1. Prioritize Inventory with Excellent Condition Ratings

* Vehicles listed as “excellent” or “like new” command the highest resale values.
* Dealers and listing platforms should prioritize acquiring and featuring these vehicles to maximize profit margins.
* In addition, emphasizing condition in marketing material could lead to faster sales and better customer perception.

2. Price Automatic, Electric, and Diesel Vehicles at a Premium

* Listings with automatic transmission consistently yielded higher average prices, reflecting consumer preferences for ease of use and modern features.
* Similarly, electric and diesel vehicles tend to command premium pricing due to their efficiency, environmental appeal, and newer market presence.
* It is recommended to segment pricing strategies based on fuel and transmission type to reflect this demand elasticity.

3. Emphasize Low-Mileage Listings

* Odometer reading shows a strong negative correlation with price. Cars with lower mileage not only sell at higher prices but are also perceived as more reliable.
* Inventory acquisition teams should prioritize low-mileage vehicles during sourcing or trade-in programs.
* Listing platforms should offer sorting/filtering by mileage as a highlighted option for users.

4. Offer a Balanced Mix with Emphasis on Sedans and SUVs

* Sedans and SUVs dominate the dataset, indicating high availability and consumer interest.
* However, SUVs tend to show higher average and peak prices, suggesting better margins.
* A balanced inventory approach that includes both categories—especially high-demand brands—can improve turnover and consumer engagement.

5. Adopt Predictive Pricing Algorithms

* Given the clear relationship between features (year, condition, odometer, fuel type) and price, dealerships should implement machine learning models or regression tools to assist with real-time dynamic pricing.
* This ensures competitive and fair pricing while maximizing margins, particularly in online listings where buyers frequently compare multiple options.

# 8. Next Steps

While this report has laid a strong foundation for understanding what drives car pricing using exploratory data analysis, there remains substantial opportunity to extend this work with advanced modeling, segmentation, and automation.

1. Predictive Modeling Using Machine Learning

* Future work can include training supervised learning models such as Random Forest, Gradient Boosting (XGBoost), or Linear Regression with regularization to predict car prices more accurately.
* These models can handle nonlinear relationships and interactions between variables better than simple regression.
* Models can be evaluated using metrics such as R² score, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) to select the best-performing algorithm.

2. Segmentation by Region or Seasonality

* Car prices often vary by geographic location (e.g., coastal vs. inland, urban vs. rural) and seasonality (e.g., convertibles in summer, 4WD vehicles in winter).
* Segmenting the analysis accordingly could provide localized pricing insights.
* This enables region-specific pricing strategies and inventory decisions for dealerships.

3. Integrating External Data Sources

* Future iterations of this project could incorporate vehicle history reports, including:
  + Accident records
  + Ownership history
  + Service and maintenance records
* These factors are critical in real-world pricing decisions and would enhance the accuracy of any predictive models.

4. Developing a Real-Time Pricing API

* Based on the models and rules derived from this analysis, a RESTful API could be developed to allow dealerships, buyers, and platforms to:
  + Input car details (year, condition, mileage, fuel, etc.)
  + Receive instant price appraisals with a confidence range
* This could serve as a tool for online marketplaces, auto trade-in platforms, and pricing advisors.