Exploring Car Characteristics and Pricing in India.

Introduction:

The automotive industry represents a complex interplay of engineering innovation, market dynamics, and consumer preferences. Understanding the factors that influence vehicle characteristics and pricing is crucial for stakeholders across the value chain, including manufacturers, consumers, and market analysts. This study endeavours to analyse a comprehensive automotive dataset to elucidate the intricate relationships between various vehicle attributes and their Manufacturer's Suggested Retail Price (MSRP). By employing data-driven techniques, this study aims to uncover key patterns and trends that govern the valuation and market positioning of automobiles. The findings are expected to provide valuable insights into the multifaceted determinants of vehicle pricing within the Indian automotive market context, considering the unique consumer preferences and market dynamics prevalent in the region.

Objectives:

- Characterise the statistical distribution of key vehicle features, encompassing both technical specifications and market-related variables.
- Quantify the impact of engineering-related attributes, such as engine type, number of cylinders, and fuel efficiency, on the MSRP.
- Evaluate the influence of market-driven factors, including market category, vehicle size, style, and popularity, on the MSRP.
- Identify and quantify the linear correlations between various vehicle attributes to understand potential interdependencies.
- Visually represent the identified relationships and trends to facilitate a clear and concise communication of the findings regarding automotive market dynamics in India.

Methodology:

This study adopts a quantitative methodology, employing a structured approach to data analysis within a Jupyter Notebook environment. The initial phase will involve the acquisition and loading of the automotive dataset, followed by a rigorous Exploratory Data Analysis (EDA). The EDA will encompass a thorough examination of the dataset's structure, identification and treatment of missing values and outliers using appropriate statistical techniques, and the generation of descriptive statistics and univariate visualizations (histograms, box plots, bar charts) to understand the distribution of individual variables. Subsequently, the study will focus on addressing the formulated research objectives. This will involve the application of bivariate and multivariate visualization techniques, including scatter plots to explore relationships between continuous variables, box plots and violin plots to compare MSRP across different categorical features, bar charts to analyse average MSRP by market segments, and correlation heatmaps to identify linear associations between numerical attributes. Statistical measures, such as correlation coefficients, will be calculated to quantify the strength and direction of linear relationships. The interpretation of these visualizations and statistical outputs will form the basis for deriving meaningful insights into the factors influencing vehicle pricing and market positioning within the Indian context. The final stage will involve the synthesis of the findings into a coherent conclusion, highlighting the key relationships identified and suggesting potential avenues for future research.

Program:

```
# Import required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib
matplotlib.use('Agg')  # for headless environment, remove if using Jupyter or local GUI

import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Load the dataset
car_data = pd.read_csv('car_model_dataset.csv')

# View the first 5 rows to understand the data
print(car_data.head())

# Get dataset information
print(car_data.info())

# Check for missing values
print(car_data.isnull().sum())

# Select the necessary columns for our analysis
data = car_data[['Engine HP', 'MSRP']].copy()  # <-- create a copy explicitly

# Drop missing values safely (no inplace on slice)
data = data.dropna()
```

```
# Remove outliers — cars with price > 150,000 might distort analysis
data = data[data['MSRP'] < 150000]

# Quick check
print(data.describe())

# Compute correlation matrix
corr = data.corr()

# Visualize correlation matrix using heatmap
plt.figure(figsize=(6,4))
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation between Engine HP and Car Price (MSRP)')
plt.savefig('correlation_heatmap.png')
plt.close()

# Scatter plot of Engine HP vs MSRP with regression line
plt.figure(figsize=(8,6))
sns.regplot(x='Engine HP', y='MSRP', data=data, scatter_kws={'alpha':0.4})
plt.title('Engine HP vs MSRP with Regression Line')
plt.xlabel('Engine Horsepower (HP)')
plt.ylabel('Price (MSRP)')
plt.grid(True)
plt.savefig('enginehp_vs_msrp_regression.png')  # different filename
plt.close()

# Split data into input (X) and output (y)
X = data[['Engine HP']]
y = data['MSRP']

# Split into training and testing datasets (80% train, 20% test)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize and train the Linear Regression model
lr = LinearRegression()
lr.fit(X_train, y_train)

# Predict on test data
y_pred = lr.predict(X_test)

# Calculate R² Score
r2 = r2_score(y_test, y_pred)

# Calculate Mean Squared Error and Root Mean Squared Error
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)

# Print results
print("R² Score:", r2)
print("Mean Squared Error (MSE):", mse)
print("Root Mean Squared Error (RMSE):", rmse)

# Plot residuals to check for pattern
residuals = y_test - y_pred
```

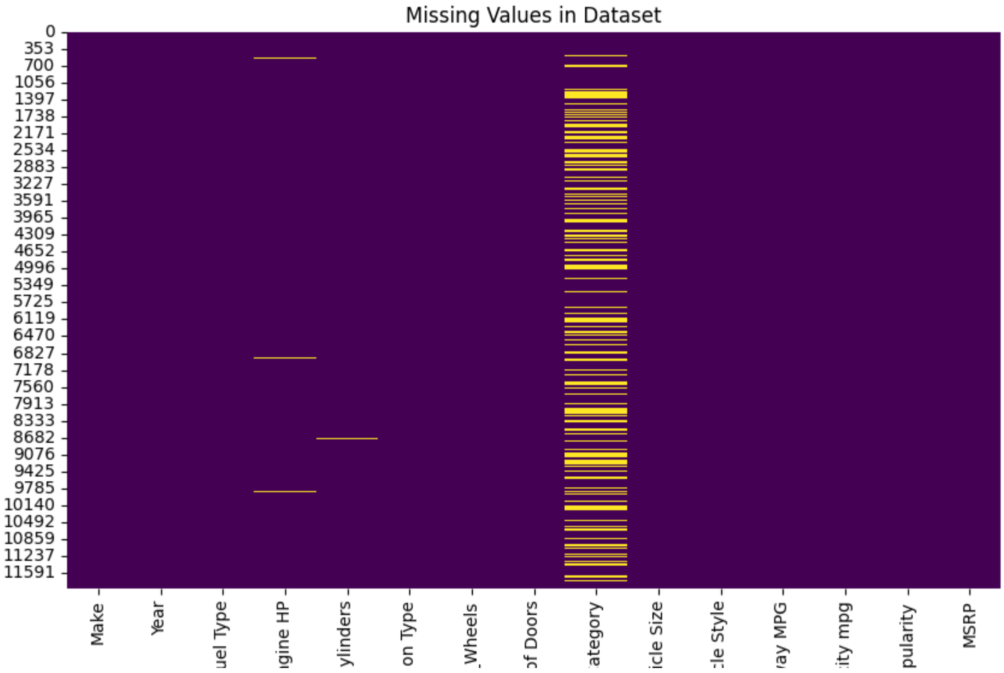
```
plt.figure(figsize=(8,5))
sns.scatterplot(x=y_pred, y=residuals)
plt.axhline(0, color='red', linestyle='--')
plt.xlabel('Predicted MSRP')
plt.ylabel('Residuals')
plt.title('Residual Plot')
plt.grid(True)
plt.savefig('residual_plot.png')  # different filename
plt.close()
```

Output and Interpretation:

Mean Squared Error: 338424333.83

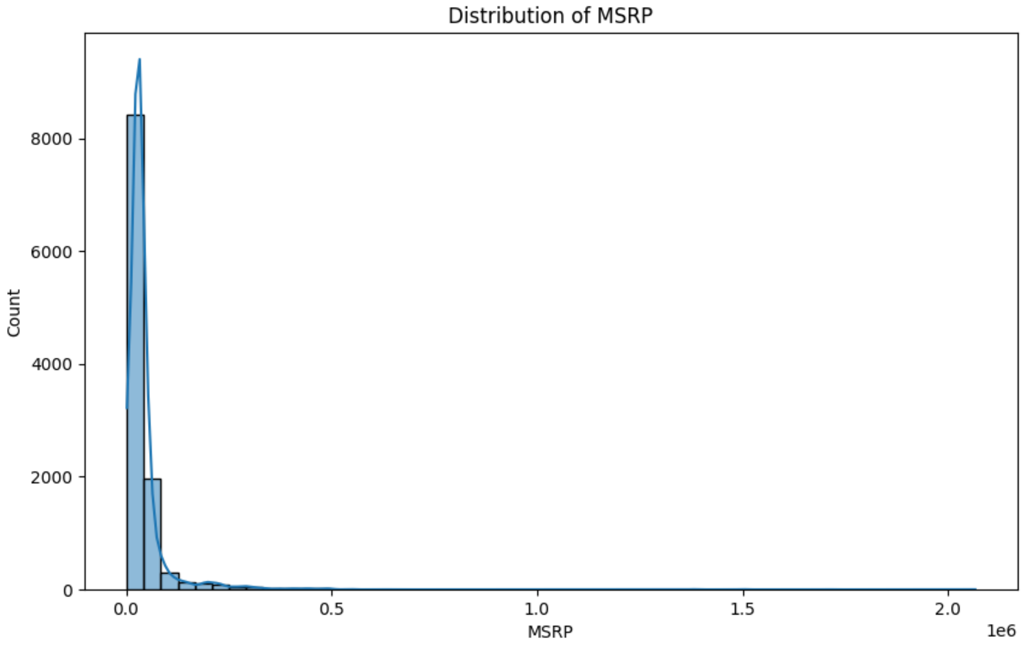
R² Score: 0.93

1. Missing Values Heatmap Interpretation



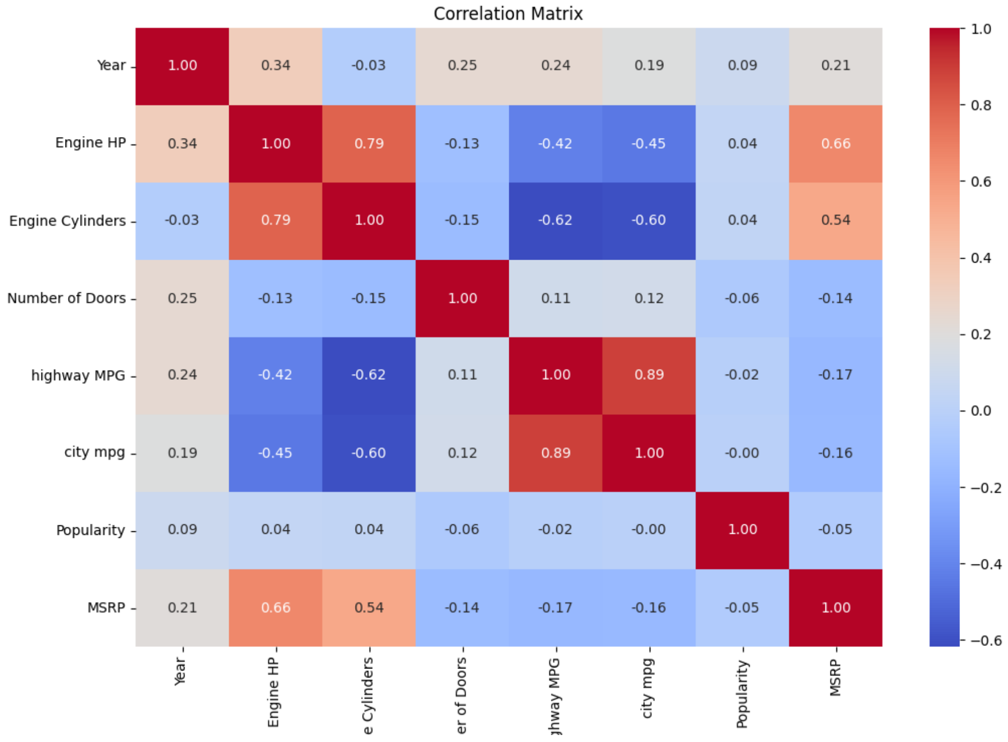
The missing values heatmap reveals the presence and distribution of null values across the dataset’s columns. Prominently, features such as Number of Doors and Engine Cylinders have visible gaps, indicating substantial missing data. Other columns like Engine HP and Fuel Type may also contain scattered missing entries. This is critical because missing data can introduce bias or lead to erroneous conclusions during model training and evaluation. For instance, if a model is trained on incomplete records, it might learn relationships that don’t generalize to real-world scenarios. Some features, like Engine HP, might be essential for predicting the target variable, and thus their missing values cannot be ignored. Depending on the context, these missing values can be addressed using various strategies. For numerical columns such as Engine HP and Engine Cylinders, mean or median imputation can be effective, while categorical variables like Fuel Type or Number of Doors can be filled using the mode. If a column has too many missing values and isn't critical, it might be best to drop it entirely. Additionally, it's good practice to investigate the nature of the missingness—whether it's Missing Completely at Random (MCAR), Missing at Random (MAR), or Not Missing at Random (NMAR)—to choose the most appropriate imputation strategy.

2. MSRP Distribution Interpretation



The histogram showing the MSRP (Manufacturer's Suggested Retail Price) distribution indicates a highly right-skewed distribution, with the majority of vehicle prices falling below $100,000 and a few extreme values reaching up to $2,000,000. This kind of skewness is significant because many machine learning models, especially linear ones, assume that the target variable (in this case, MSRP) follows a roughly normal distribution. When this assumption is violated, the models may perform poorly or produce biased predictions. Additionally, the presence of high-value outliers can disproportionately influence model training, particularly in models sensitive to large magnitude differences, such as linear regression. These outliers may correspond to rare luxury or specialty vehicles that are not representative of the majority. A common solution is to apply a logarithmic transformation (e.g., log1p) to compress the range and reduce the skewness, thereby making the data more suitable for predictive modelling. Alternatively, depending on the use case, some extreme outliers could be removed if they are not relevant to the analysis objective. Understanding the nature and distribution of the target variable is crucial for choosing the right preprocessing and modelling techniques.

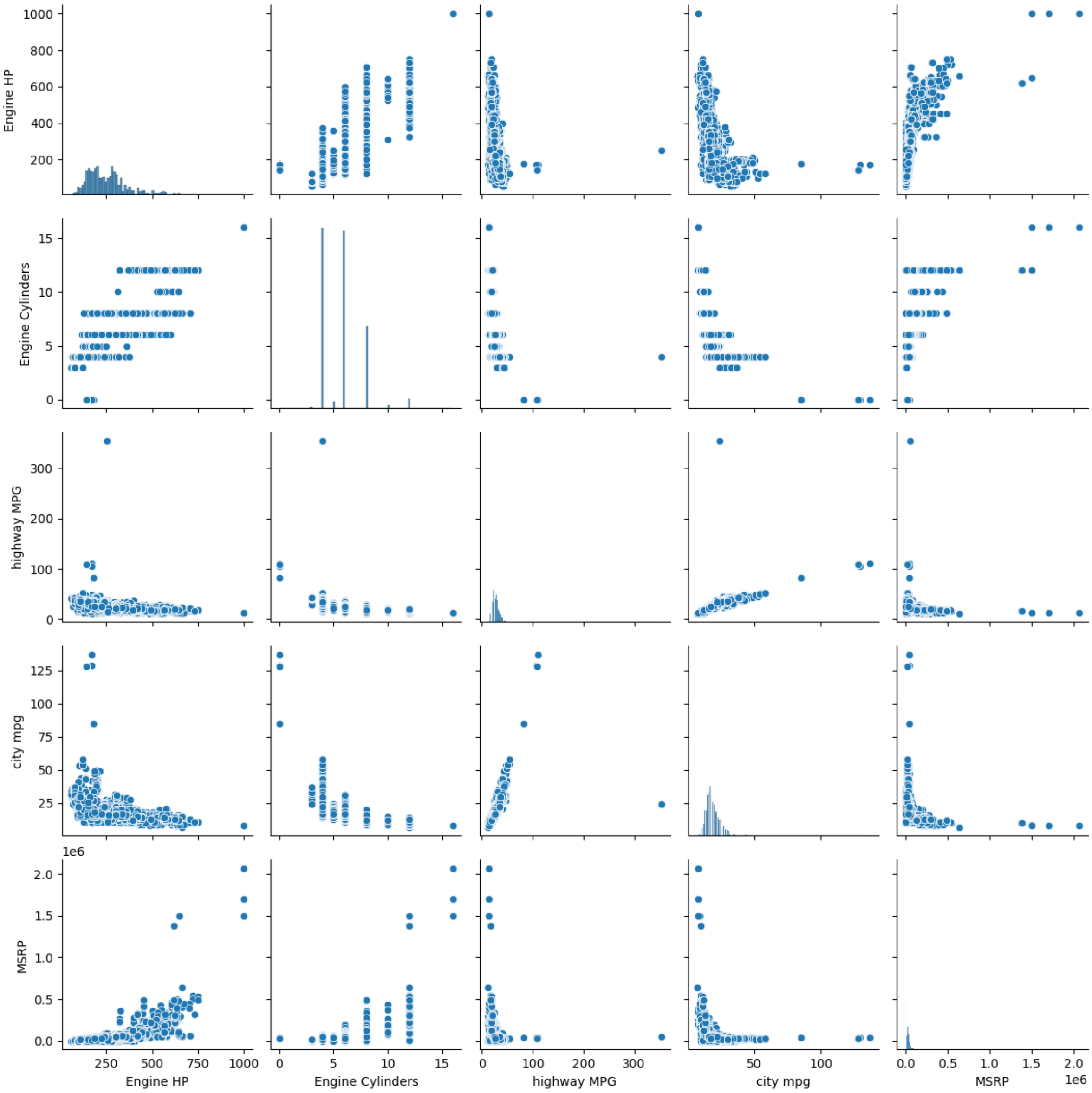
3. Correlation Matrix Interpretation



The correlation matrix visualizes the linear relationships between numerical features using Pearson correlation coefficients, helping identify both strong and weak associations. A notable observation is the strong positive correlation (0.79) between Engine HP and Engine Cylinders, suggesting that more powerful engines typically have more cylinders. Another pair with a very strong correlation (0.89) is city MPG and highway MPG, which makes sense as both measure similar aspects of a vehicle's fuel efficiency under different driving conditions. The relationship between Engine HP and MSRP is moderately positive (0.66), indicating that more powerful vehicles tend to be more expensive. Interestingly, city MPG and Engine HP have a negative correlation (-0.45), highlighting that higher horsepower often comes at the cost of fuel efficiency. These insights are valuable in multiple ways. Highly correlated features can lead to multicollinearity, which is problematic in linear models, as it makes it difficult to isolate the individual impact of each feature. As a result, techniques like Variance Inflation Factor (VIF) analysis or Principal Component Analysis (PCA) can be used to reduce redundancy.

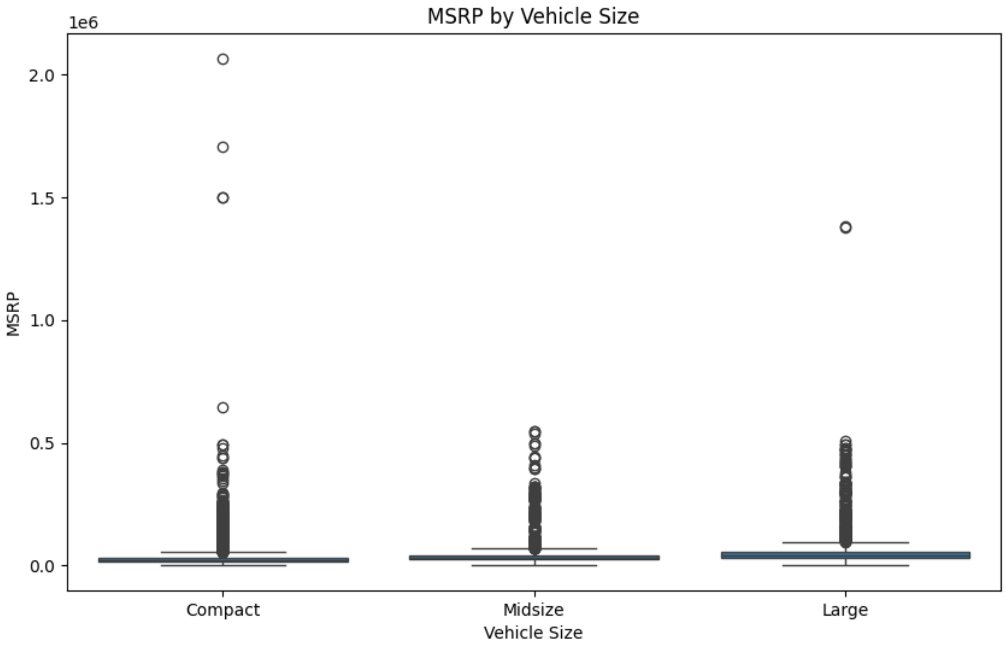
Alternatively, highly correlated features like city MPG and highway MPG can be combined into a single fuel efficiency score. Understanding these relationships also aids in identifying the most predictive features for modelling MSRP and constructing more interpretable models.

4. Selected Features Pairplot Interpretation



The pairplot provides a comprehensive visual summary of pairwise relationships among selected features—Engine HP, Engine Cylinders, city MPG, highway MPG, and MSRP. Each scatterplot captures potential linear and non-linear interactions, complementing the correlation matrix with richer visual context. The plots confirm that features like Engine HP and Engine Cylinders share a strong positive relationship, while city MPG and Engine HP show a clear inverse trend, aligning with earlier correlation findings. The diagonal histograms highlight the distribution of each variable, showing that Engine Cylinders has a discrete, almost categorical distribution, while features like MSRP are continuous but skewed. Some scatterplots exhibit clustering or curved patterns that suggest non-linear relationships which may not be well captured by linear models. For example, the relationship between Engine HP and MSRP might be better represented using polynomial terms or tree-based models. These patterns also suggest possible segmentation within the data—for instance, certain clusters could correspond to different vehicle categories such as compact cars, SUVs, or high-performance sports vehicles. This implies that model performance could be improved by incorporating non-linear transformations, interaction terms, or by using non-linear algorithms such as Random Forests or Gradient Boosting Machines. The pairplot is a valuable diagnostic tool for identifying underlying structure in the data that isn't always obvious from numerical summaries.

5. MSRP by Vehicle Size Boxplot Interpretation



The boxplot of MSRP grouped by Vehicle Size reveals clear pricing differences across compact, midsize, and large vehicles. Compact and midsize vehicles have relatively lower medians and narrower interquartile ranges, indicating a more consistent and affordable price range. In contrast, large vehicles show a higher median MSRP and a wider spread, with many extreme values, suggesting the presence of premium or luxury variants within that category. The broader range for large vehicles could be attributed to additional features, engine power, brand value, or the inclusion of utility vehicles like large SUVs and trucks. Each size category also displays several outliers, indicating that within each group, there are some vehicles priced much higher than the norm. This visualization reinforces the importance of Vehicle Size as a categorical variable with a significant effect on MSRP. In modelling scenarios, it should be properly encoded using techniques like One-Hot Encoding or Ordinal Encoding, depending on the model type. Additionally, the observed spread suggests the potential value of modelling interactions between Vehicle Size and numeric features such as Engine HP or Fuel Type. Including such interactions could help capture variations in pricing behaviour across vehicle categories more effectively. This boxplot is a strong indicator that vehicle classification characteristics should be an integral part of any model predicting MSRP.

Conclusion:

The exploratory data analysis of the vehicle dataset has provided several crucial insights that will guide both data preprocessing and modelling strategies. The missing values heatmap highlighted incomplete records in key features such as Engine HP, Engine Cylinders, and Number of Doors, signalling the need for thoughtful imputation or potential exclusion strategies to maintain data integrity. The MSRP distribution plot revealed a strong right skew, driven by high-end vehicles, suggesting the necessity of a log transformation or outlier handling to prevent modelling bias and improve prediction accuracy.

The correlation matrix shed light on meaningful linear relationships, such as the strong positive association between Engine HP and Engine Cylinders, and a negative correlation between engine power and fuel efficiency, which not only confirms domain knowledge but also warns of potential multicollinearity. These insights were further validated and enriched by the pairplot, which visually captured non-linear patterns, discrete distributions, and potential clustering, indicating that more complex models or feature engineering (such as interaction terms) may be required for optimal performance.

Finally, the boxplot comparing MSRP across vehicle sizes demonstrated that Vehicle Size has a significant and structured impact on price, with large vehicles exhibiting a wider price range and higher medians. This underlines the importance of incorporating categorical variables effectively and possibly modelling their interactions with numerical features.