

# Car Features and MSRP Dataset Preprocessing

SIC – AI701

# Agenda

- Overview
- Dataset Description
- Data Cleaning & Preparation
- Descriptive Statistics (EDA – Part 1)
- Correlation Analysis (EDA – Part 2)
- Key Questions Explored & Data Visualization
- Live Demo
- Conclusion



# Project Overview

# Overview

Our project focuses on analyzing a car dataset to better understand patterns and relationships between different car attributes such as price, model year, engine type, and brand.

The main business problem it addresses is identifying the factors that influence car pricing, which can be useful for car dealers, buyers, and manufacturers to make informed decisions.



# Overview

By analyzing trends, we can gain insights into how car features affect market value.

- How does the brand affect the price?
- What cars can be considered overpriced?
- Price VS. popularity?





# Dataset Description

# Dataset Description

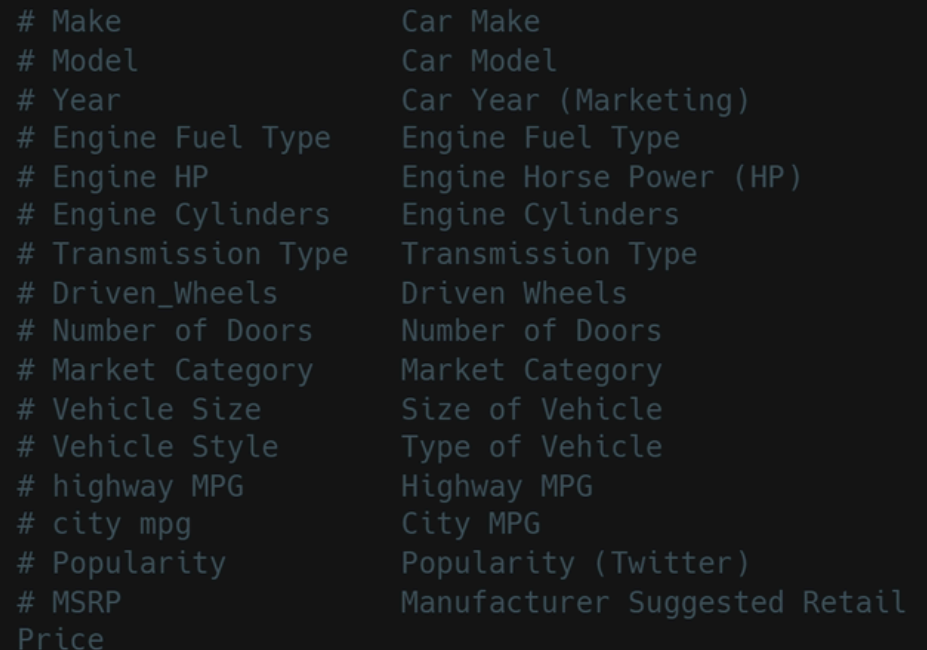
We used the Car Features and MSRP Kaggle dataset scraped from Edmunds (an online resource for automotive info) and Twitter.

At 11914 rows and 16 columns the dataset has variety of attributes that allows us to analyze both numerical (e.g., price, year, mileage) and categorical (e.g., make, market category, transmission) data.



# Dataset Description

At 11914 rows and 16 columns the dataset has variety of attributes that allows us to analyze both numerical (e.g., price, year, mileage) and categorical (e.g., make, market category, transmission) data.



```
# Make           Car Make
# Model          Car Model
# Year           Car Year (Marketing)
# Engine Fuel Type Engine Fuel Type
# Engine HP      Engine Horse Power (HP)
# Engine Cylinders Engine Cylinders
# Transmission Type Transmission Type
# Driven_Wheels  Driven Wheels
# Number of Doors Number of Doors
# Market Category Market Category
# Vehicle Size   Size of Vehicle
# Vehicle Style  Type of Vehicle
# highway MPG    Highway MPG
# city mpg       City MPG
# Popularity     Popularity (Twitter)
# MSRP           Manufacturer Suggested Retail
Price
```





# Data Cleaning & Preparation

# Data Cleaning & Preparation

- Removed missing values / imputed outliers.
- Removed duplicates.
- Converted data types (e.g., Make → categorical).
- Grouped low-frequency brands into “Other”.
- Created new feature: Car\_Type (Luxury vs Normal).
- Dummy encoding – Converted categorical variables such as Make into dummy variables for better analysis and modeling.



# Descriptive Statistics (EDA – Part 1)

## Car Price (MSRP)

- Mean Price = \$40,595
- Median Price = \$29,995
- Standard Deviation = \$60,109

The mean is significantly higher than the median, suggesting a **right-skewed** distribution with some high-priced outliers.

## Engine HP (Horsepower)

- Mean HP = 249
- Median HP = 227
- Standard Deviation = 109
- Has 69 missing values

Distribution is moderately **right-skewed**, due to high-performance models with very high HP (up to 1001). Fill missing values with **median**.



# Descriptive Statistics (EDA – Part 1)

## **MPG (Fuel Efficiency)**

### **City MPG**

- Mean = 19.7
- Median = 18
- Std = 8.99

### **Highway MPG**

- Mean = 26.6
- Median = 26
- Std = 8.86

MPG values appear fairly symmetric, especially highway MPG.

## **Year**

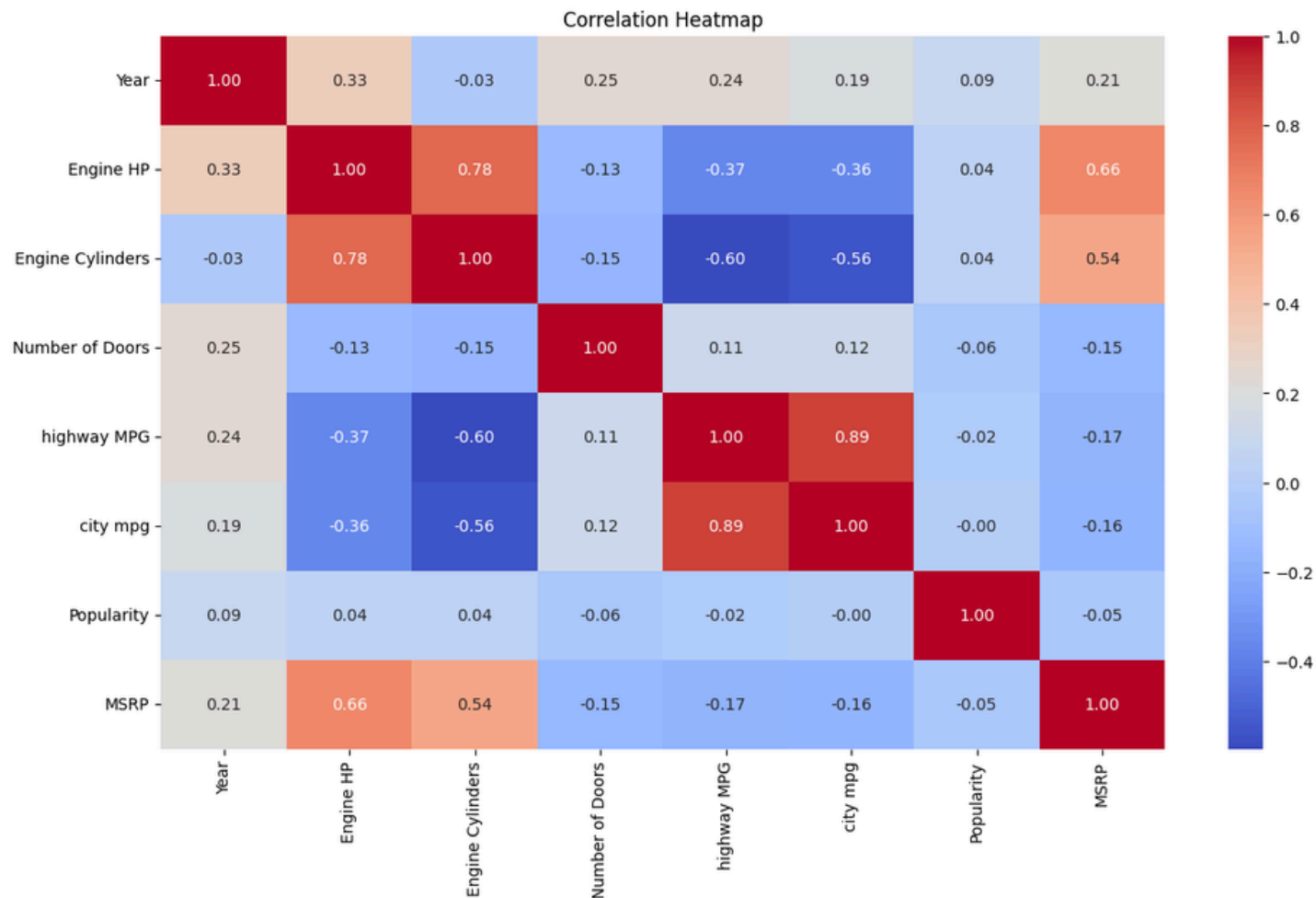
- Mean Year = 2010
- Median Year = 2015
- Std = 7.58

**Left-skewed** (more newer models), but older outliers (as old as 1990) drag the mean down.

## **Market Category**

has 31% Missing Data, so dropping it would be great decision.

# Correlation Analysis (EDA – Part 2)

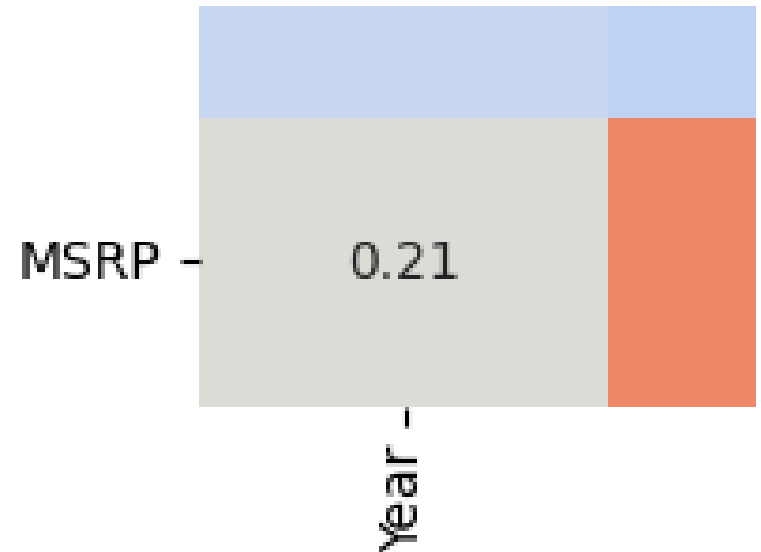


# Correlation Analysis (EDA – Part 2)

## Price vs. Age

We wanted to see if there was a relationship between a car's age (Year) and its price (MSRP).

We discovered that the correlation is positive (0.21). This suggests that **newer cars tend to be more expensive** than older ones though the relationship is **relatively weak**.

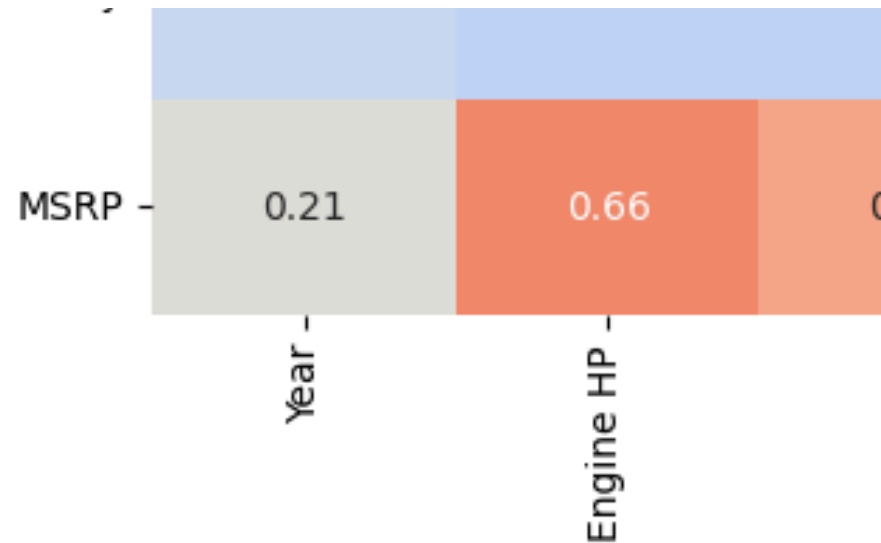


# Correlation Analysis (EDA – Part 2)

## Price vs. HP

We wanted to see if there was a relationship between a car's engine power (Horsepower) and its price (MSRP).

We discovered that the correlation is strongly positive (0.66). This indicates that **cars with higher horsepower generally cost more**.

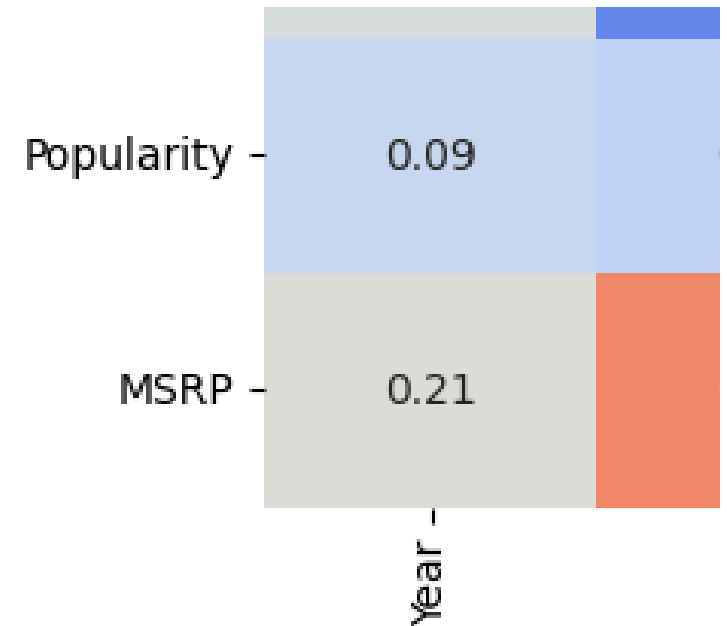


# Correlation Analysis (EDA – Part 2)

## Price vs. Popularity

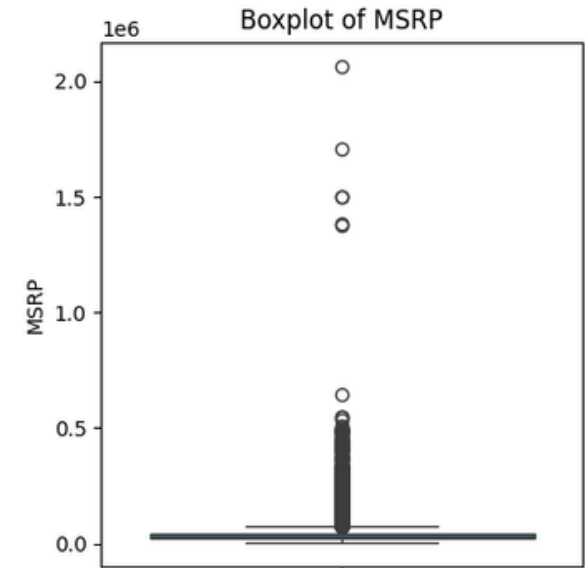
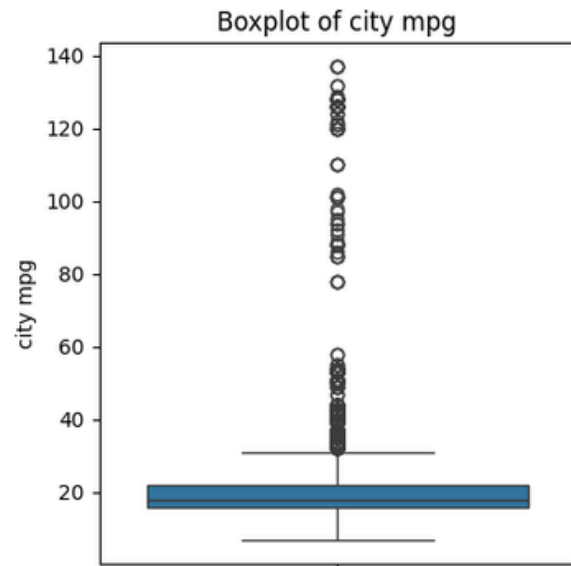
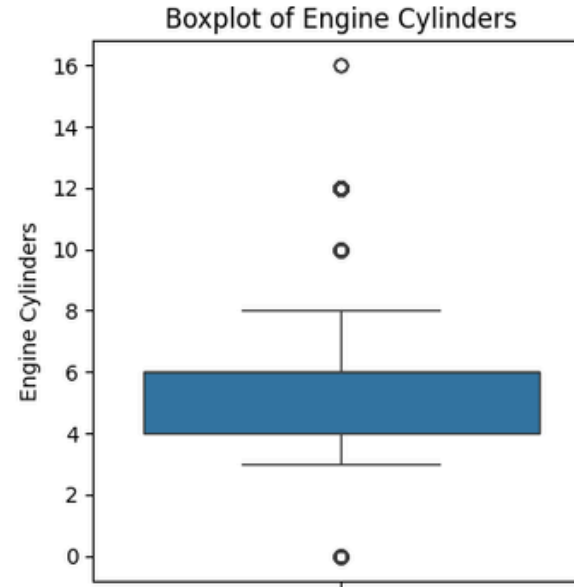
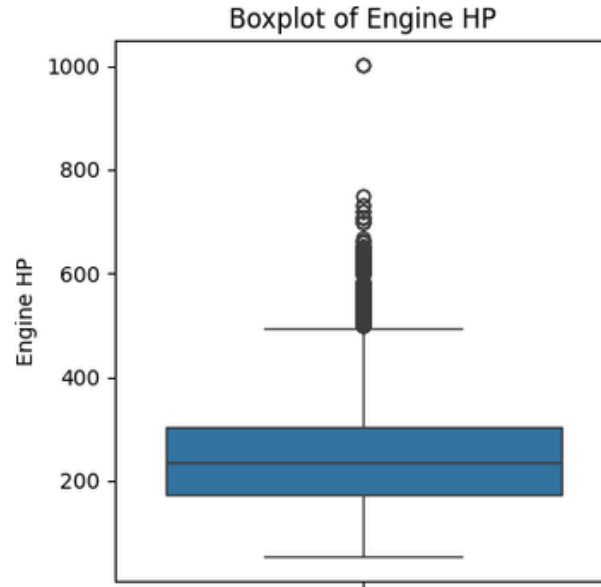
From the heatmap, MSRP and Popularity have a very weak negative correlation (0.09).

This suggests that **high price does not guarantee popularity**, and popular cars are often mid-range or budget-friendly options that balance price, performance, and efficiency





# Handling Outliers



# Handling Outliers

## Engine HP Outliers

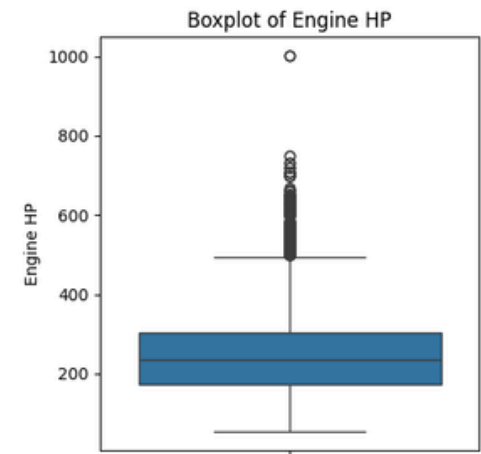
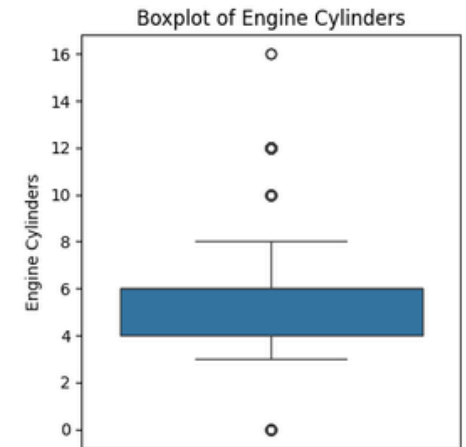
- **Engine HP, MSRP, and Cylinders:** Outliers are primarily influenced by luxury and supercars.
- **City MPG:** This data may be unreliable (bad data).

## Imputation of Outliers

Since the data follows a **roughly normal distribution** (median = 18, mean = 19.73, and standard deviation = 6.7, which indicates **low variation**), we chose to **impute outliers using the mean** value rather than the median.

## Categorizing Cars as Luxury or Normal

- If the **MSRP** exceeds \$100,000 or the **Engine HP** is greater than 500, set **Car\_Type** to **Luxury**.
- Otherwise, classify it as **Normal**.



# Feature Engineering

The dataset contains several categorical features that the model cannot directly interpret.

## One-Hot Encoding:

Categorical attributes such as *Car\_Type*, *Engine Fuel Type*, *Driven\_Wheels*, *Transmission Type*, *Vehicle Size*, *Vehicle Style* were transformed into numerical format using one-hot encoding.



This process converts each category into binary columns, making the data machine-readable while preserving the categorical information

# Feature Engineering

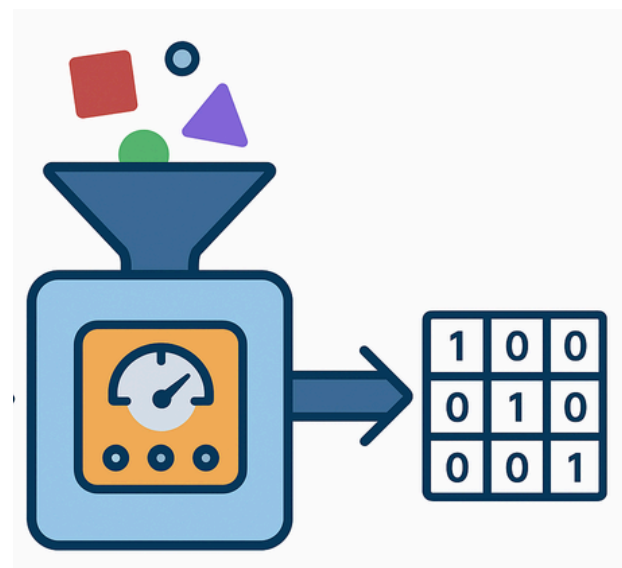
## Handling Car Brands (Make)

The dataset includes many different car brands, some of which appear only a few times.

To simplify the analysis and visualizations:

- Selected the Top 15 most frequent car makes.
- Grouped all remaining brands under the category “Other.”

This reduces noise and makes trends between popular brands clearer.





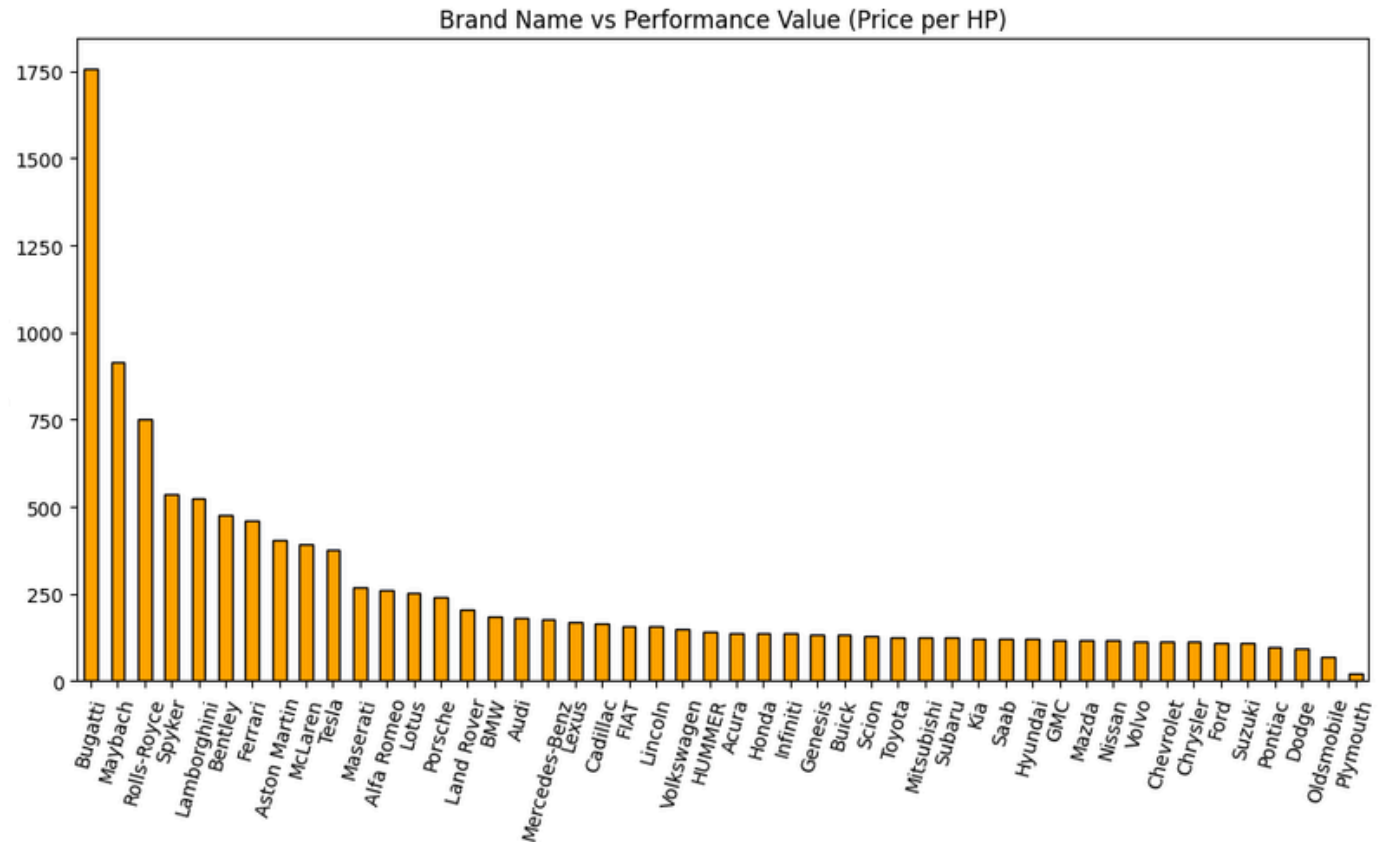
# Key Questions Explored & Visualization

# Key Questions & Visualization

What cars can be considered overpriced?

What cars have great value for performance?

What brands focus on selling an identity rather than a great car?

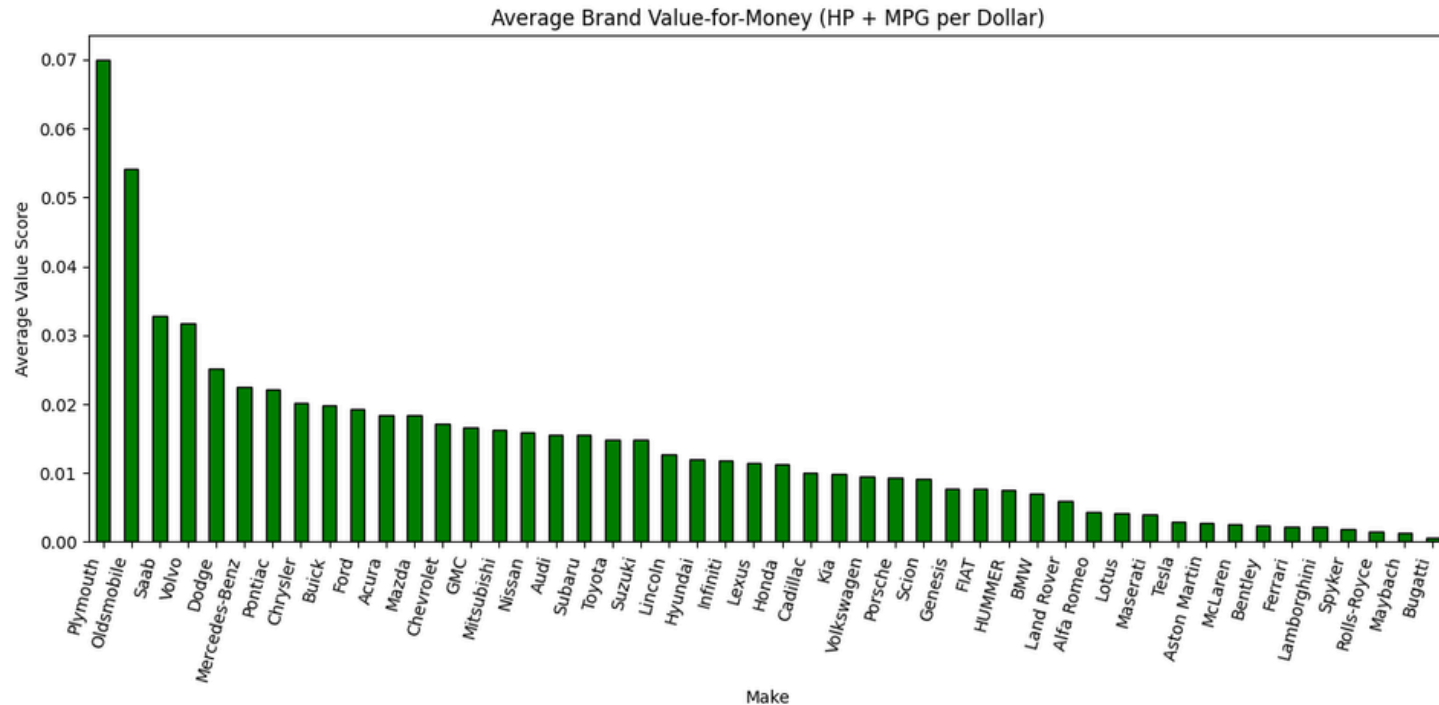


# Key Questions & Visualization

What cars can be considered overpriced?

What cars have great value for performance?

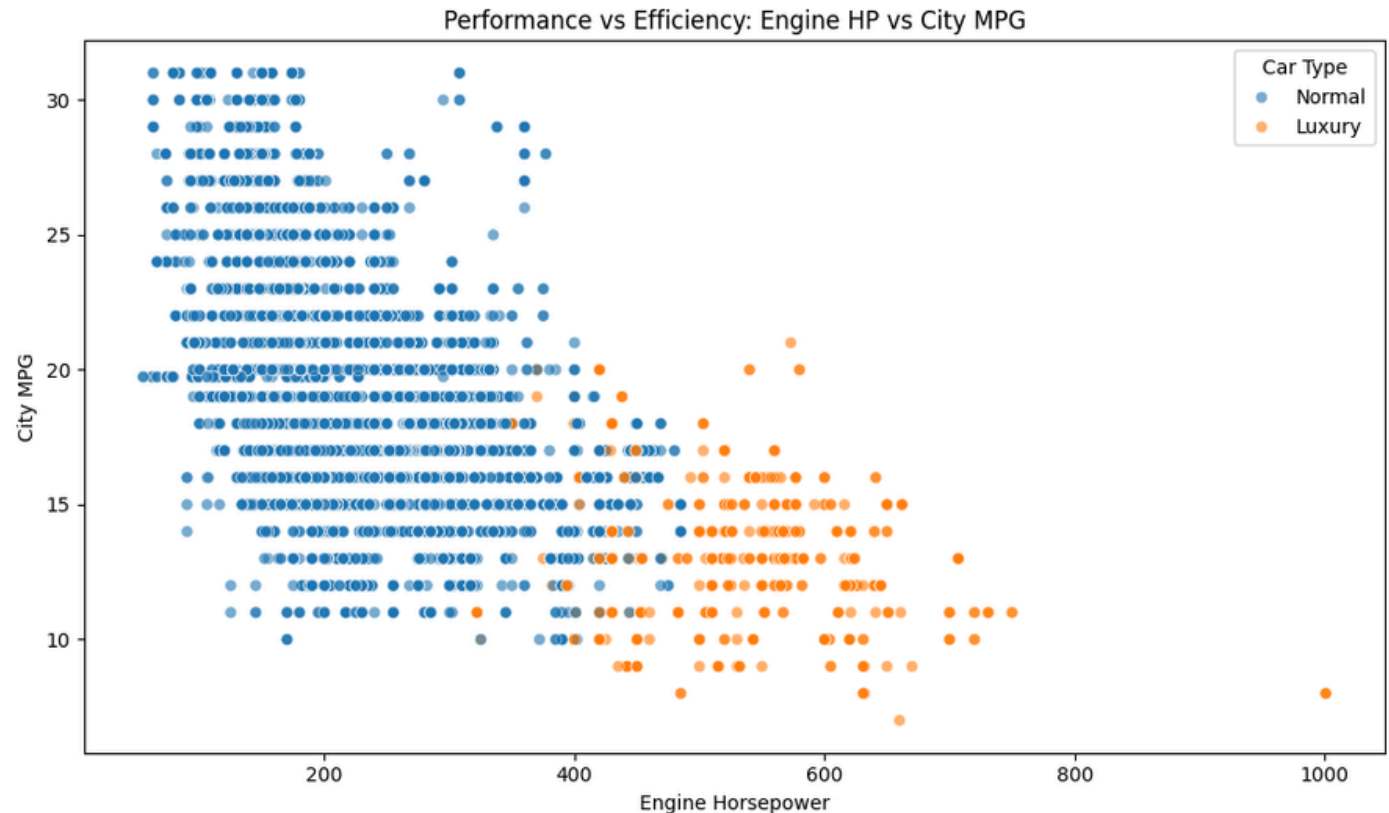
What brands focus on selling an identity rather than a great car?



# Key Questions & Visualization

Luxury/expensive cars  
have great efficiency for  
performance?

How many cars have  
performance efficiency  
balance?

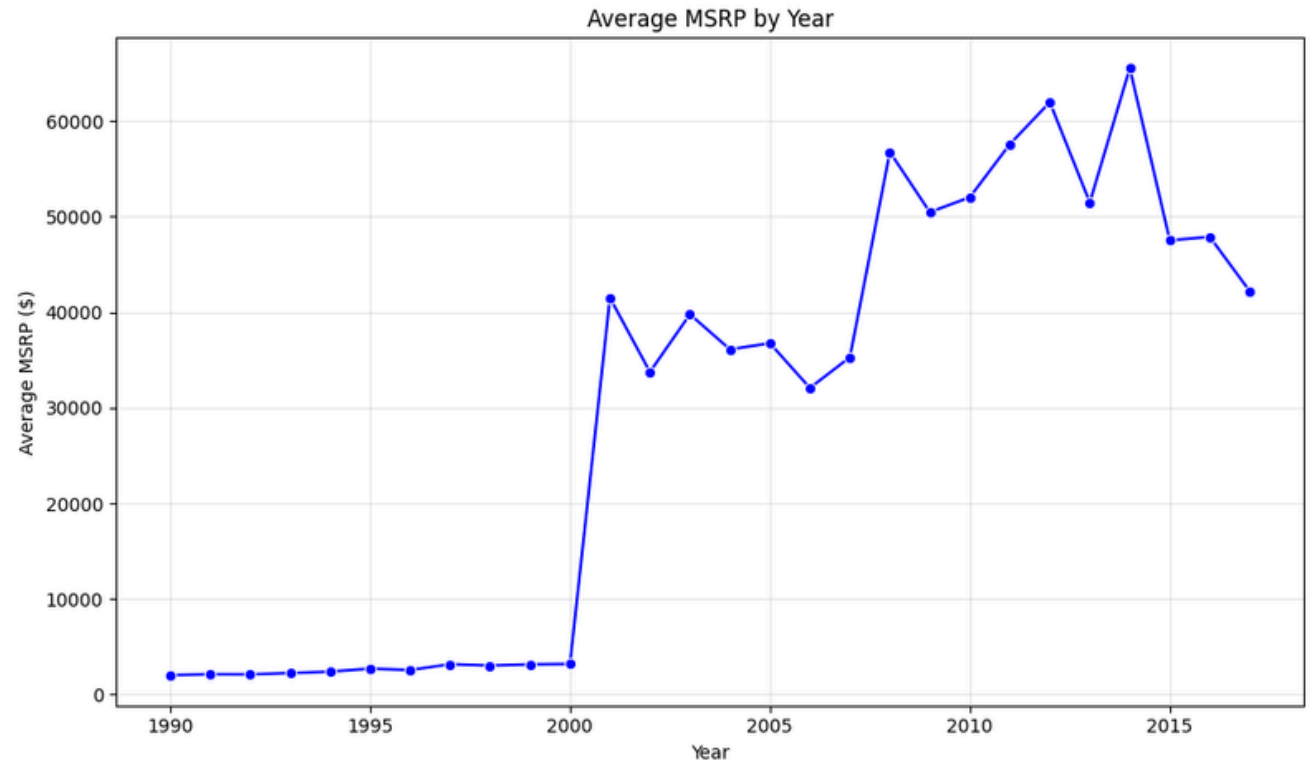




# Key Questions & Visualization

Are new cars becoming more expensive?

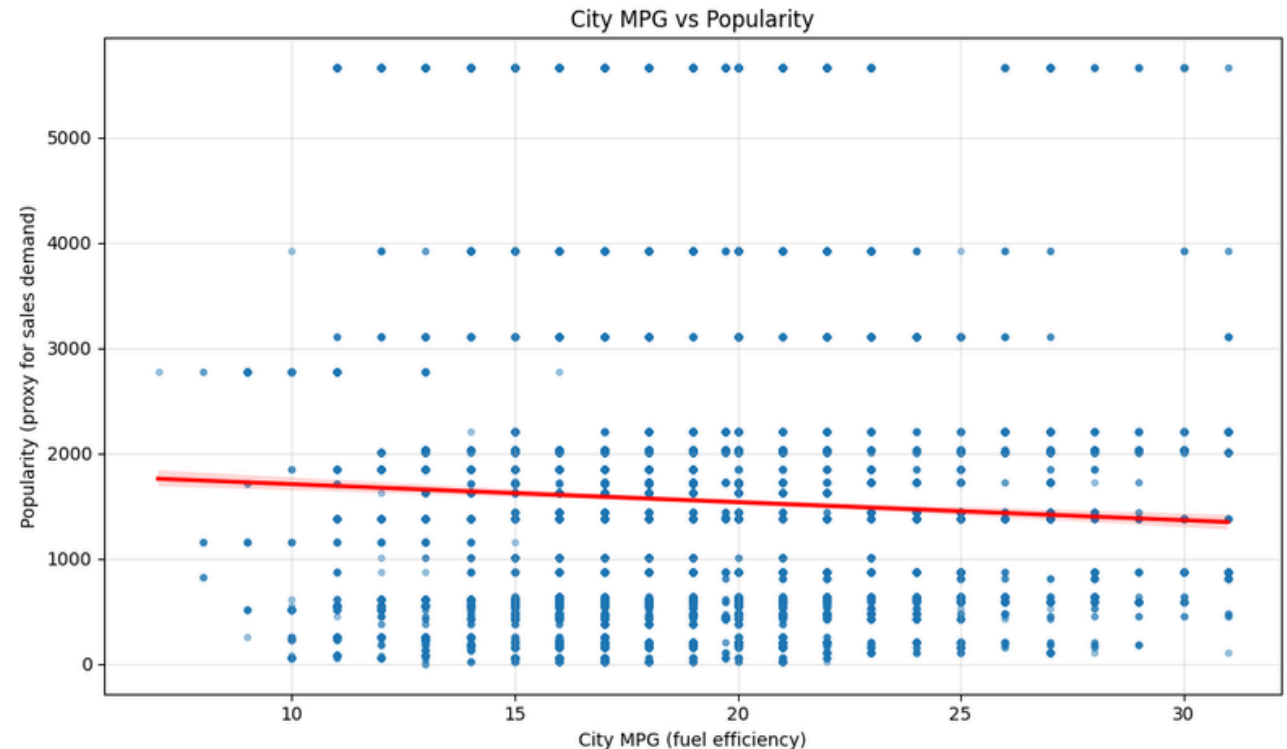
Car prices faced an overpriced/inflation period?



# Key Questions & Visualization

People's main concern is car efficiency?

Does efficient cars gain more popularity?





# Live Demo

# Conclusion

- **Price Drivers:** Car price (MSRP) is strongly influenced by engine power (HP) and engine cylinders.
- **Brand Effect:** Premium brands charge higher prices, but not always justified by performance.
- **Popularity Insight:** Expensive cars are not necessarily popular — mid-range, affordable cars attract more buyers.
- **Segmentation:** Categorizing cars into Luxury vs. Normal highlights how luxury models dominate high price ranges but represent a niche market.

## Recommendations

- **For Buyers:** Look beyond brand; focus on value-for-money features like HP and MPG.
- **For Dealers:** Stock and promote mid-range cars for mass-market appeal; position luxury cars as premium niche offerings.
- **For Manufacturers:** Balance affordability and performance — popularity is driven by accessible, reliable models, not just luxury.



Together for Tomorrow!  
**Enabling People**

Education for Future Generations

©2020 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of book.

This book is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this book other than the curriculum of Samsung innovation Campus or to use the entire or part of this book, you must receive written consent from copyright holder.