CAR PRICE PREDICTION ANALYSIS

Final Report

Yahan Madhuhansa

1. Introduction

The automotive industry relies heavily on understanding the factors that influence car pricing to optimize production, marketing, and sales strategies. This project aims to analyze a dataset of car features to identify which variables significantly impact car prices and how well these variables explain price variations. By leveraging statistical and exploratory data analysis (EDA) techniques, we provide actionable insights for pricing strategies and market segmentation.

2. Problem Statement & Questions

1. Which variables are most significant in predicting a car's price?

2. How well do these variables describe price variations?

3. Are there unexpected trends?

3. Objectives

- Rank features by correlation with price.

- Visualize relationships

4. Dataset Overview

Source: [Price Prediction -Multiple Linear Regression](#)

Variables:

- **205 observations** (cars) with **26 features**, including:

    o **Numerical**: enginesize, curbweight, horsepower, price (target).

    o **Categorical**: carbody (sedan, convertible, etc.), drivewheels (fwd, rwd), fueltypes.

Identify the data Frame

- Shape, data types
- Check missing values
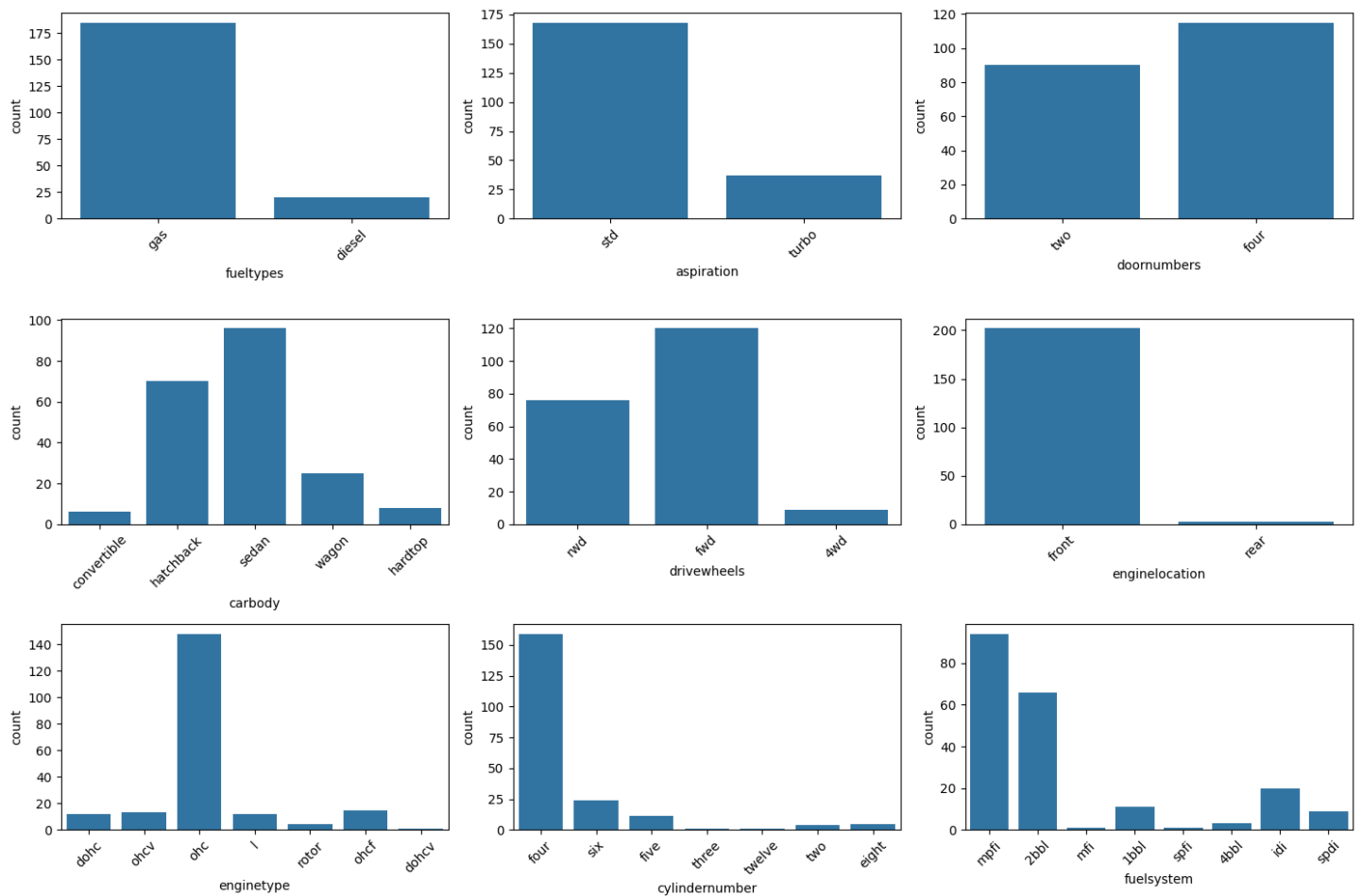- Check duplicate values

Data preprocesing

- Drop "ID" ,"name"
- Check object data type features
- feature eng for identified object data
    - label encodings "doornumbers"
    - Drop "enginelocation"
    - Feature Grouping
    - One - Hot encoding

EDA

- Price Distribution
- Heatmap for all features
- barplot showing the correlation

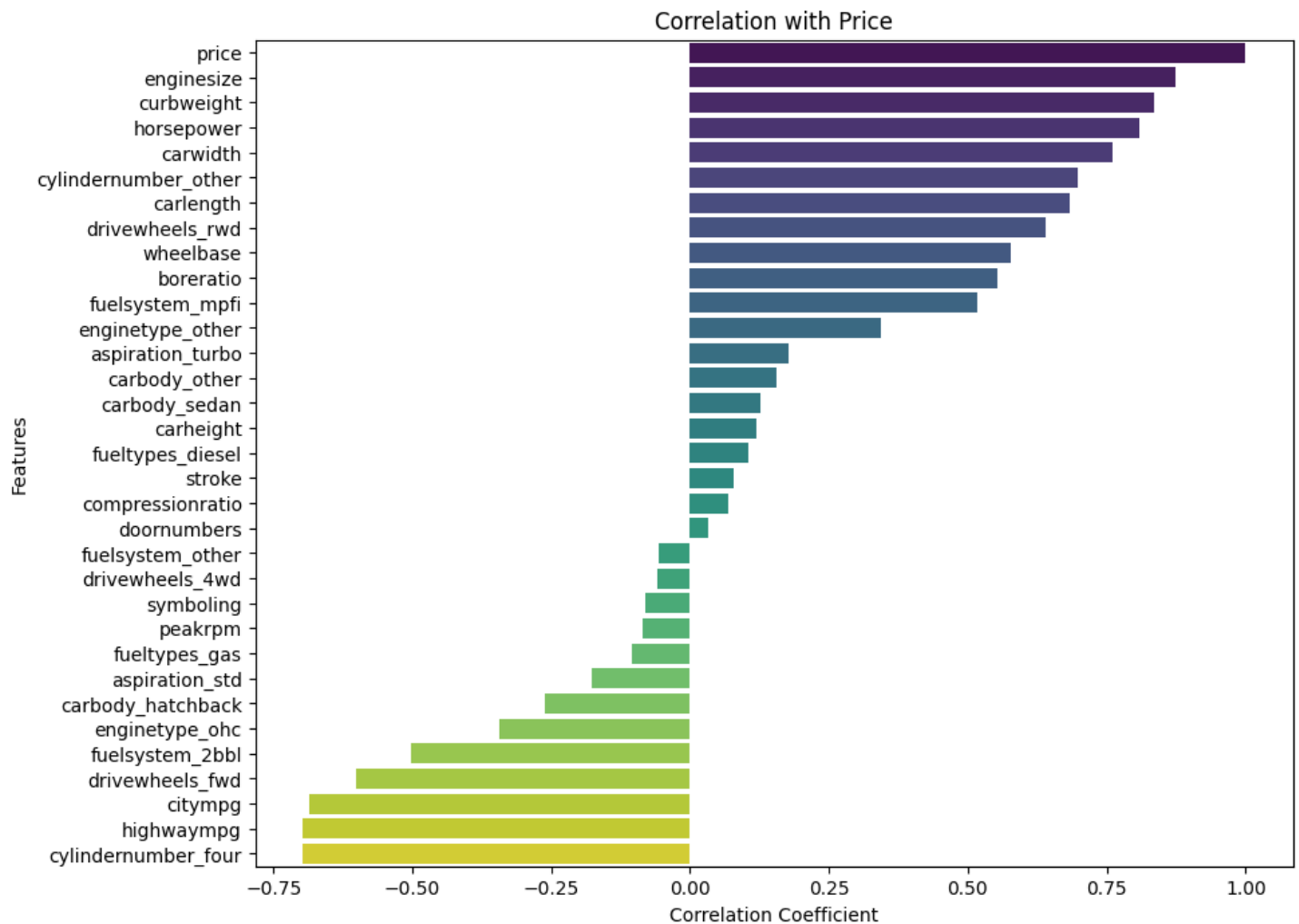After dropping 'ID' and 'name' , object data visualize like this

In here I found that 98.6% of times 'enginelocation' is front. So it make this feature almost useless.

And 'carbody' , 'enginetype' , 'cylindernumber' , 'fuelsystem' has too much small objects that make analysis less reliable and harder to interpret.

Visualize relationships with each other

Correlation with Price
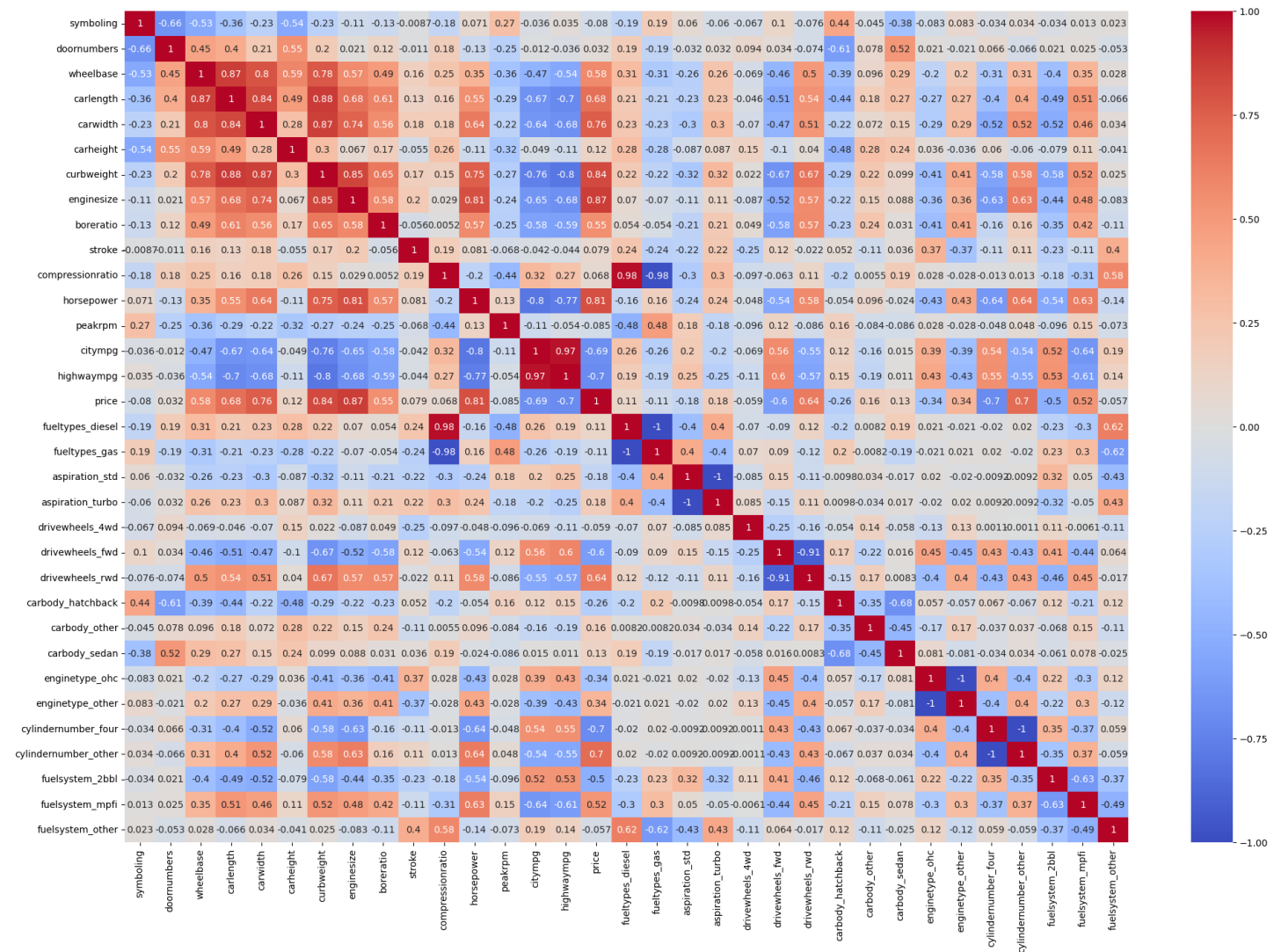
Ranked Feature Correlation with Price

Strong Positive Correlation (Most Impact on Higher Prices)

1. **enginesize** (0.87): Larger engine sizes strongly predict higher prices.

2. **curbweight** (0.84): Heavier vehicles (often luxury/SUV) cost more.

3. **horsepower** (0.81): Performance directly impacts price.

4. **carwidth** (0.76): Wider cars (e.g., luxury sedans) are priced higher.

5. **cylindernumber_other** (0.74): Cars with more cylinders (6+/V8) command higher prices.

Moderate Positive Correlation

6. **carlength** (0.68)

7. **drivewheels_rwd** (0.62)

8. **wheelbase** (0.58)

9. **boreratio** (0.55)

Weak or Neutral Correlation

- **carheight** (~0.1)
- **symboling** (near 0) *(Safety rating)*

Negative Correlation (Higher Value = Lower Price)

1. **highwaympg** (-0.70): Fuel-efficient cars are cheaper.
2. **citympg** (-0.69): Urban efficiency trades off with price.
3. **fueltypes_gas** (-0.45): Diesel cars (encoded as 0) are pricier than gas (1).

7. Conclusion

This project analyzed a dataset of **205 cars** with **26 features** to identify the key factors influencing car prices and evaluate how well these variables explain pricing variations. Through **exploratory data analysis (EDA)** and **statistical correlation techniques**, we determined the most significant predictors of car prices and derived actionable insights for pricing strategies.

Key Findings

1. **Top Positive Correlations with Price**:
   - **enginesize (0.87)**, **curbweight (0.84)**, and **horsepower (0.81)** are the strongest predictors of higher prices, indicating that performance and vehicle size significantly drive costs.
   - **carwidth (0.76)** and **cylindernumber (0.74)** also show strong relationships, reinforcing that luxury and high-performance vehicles command premium pricing.

2. **Negative Correlations**:
   - **highwaympg (-0.70)** and **citympg (-0.69)** demonstrate that fuel-efficient cars are generally more affordable, appealing to budget-conscious buyers.

3. **Categorical Insights**:
   - **Rear-wheel drive (drivewheels_rwd)** and **turbocharged engines (aspiration_turbo)** correlate with higher prices, suggesting market demand for performance-oriented vehicles.
   - **Body type (carbody)** significantly impacts pricing, with convertibles and hardtops priced higher than sedans and hatchbacks.

4. **Weak or Irrelevant Features**:
   - **symboling (safety rating)** and **carheight** showed minimal influence on price, making them less critical for pricing strategies.
   - **enginelocation** was dropped as 98.6% of cars had front engines, making it statistically insignificant.