Car Price Analysis Project Report

Name: Rishabh Sharma

Position: Data Analyst Intern

Contact No: 8791933713

Mail: rishabh8791933713@gmail.com

Introduction :-

The objective of this project is to analyze the factors influencing car prices using Exploratory Data Analysis (EDA). The dataset contains 205 records with 26 variables, covering car specifications, design features, and performance metrics.

The analysis begins with data cleaning and preprocessing steps, including handling missing values, standardizing manufacturer names. Outliers were identified using statistical (Z-Score) and visual (Boxplot) methods to understand data variability.

EDA was performed through visualizations such as distribution plots, scatter plots, and categorical comparisons. These charts helped uncover how different features — such as manufacturer, fuel type, body type, drive wheel configuration, engine size, curb weight, and horsepower — influence car pricing. A correlation heatmap was also used to identify weakly related variables, which were removed from the dataset.
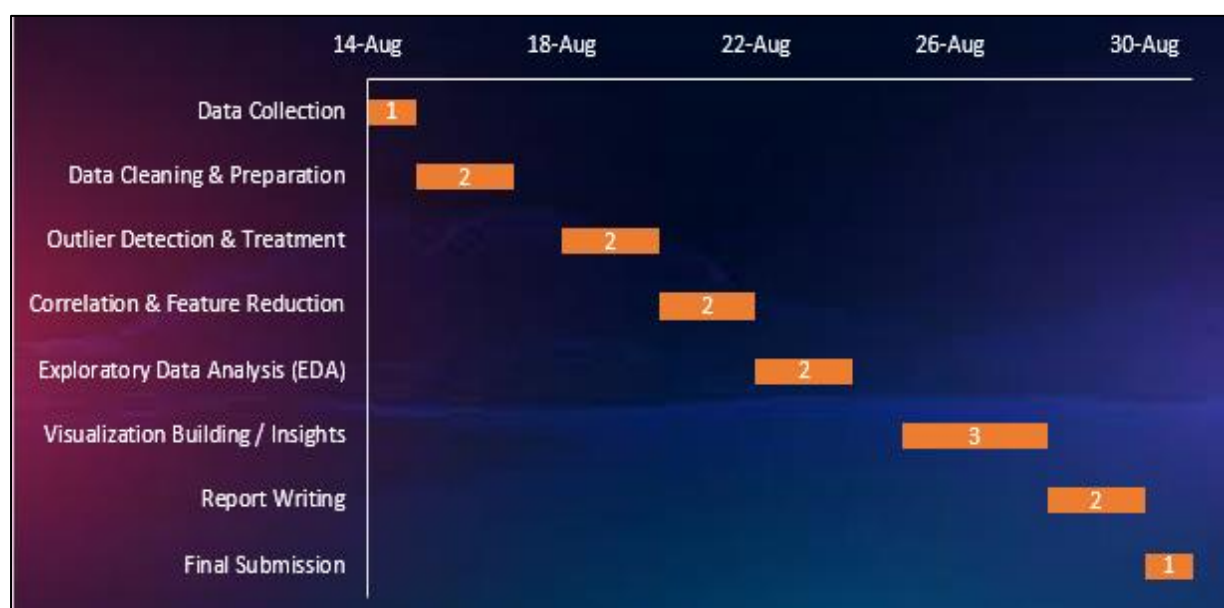
1. Data Understanding & Cleaning :-

The dataset contains 205 observations and 26 variables, covering technical specifications, categorical features, and the target variable (price). A preliminary data check revealed the following:

- Missing Values: No null or missing entries were found across any columns.

- Car Manufacturer Extraction: The CarName column was split to extract the manufacturer name (carCompany).

- Data Standardization: Several spelling discrepancies in manufacturer names were corrected (e.g., *toyouta → toyota, maxda → mazda, vokswagen/vw → volkswagen, porcshce → porsche, Nissan → nissan*). This step ensured consistency and prevented duplicate groupings.

Outlier Detection & Removal:
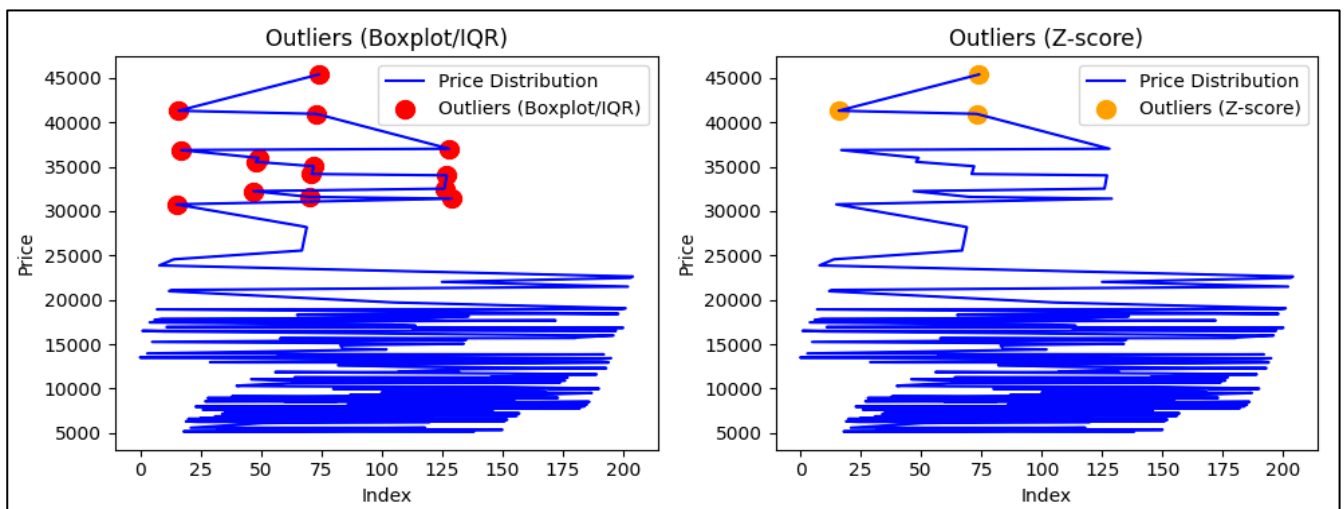
- **Outliers were detected using two methods:**

  1. Z-Score method – identified 3 outliers.

  2. Boxplot (IQR method) – identified 15 outliers initially, and 2 more upon re-checking.

- For removal, the IQR method was chosen because it is more robust for skewed distributions and works better with smaller datasets.

Reason for Not Using Z-Score for Removal:
Although Z-score flagged 3 outliers, it assumes normal distribution and can miss skew-related anomalies. Since our target variable (price) is not perfectly normally distributed, the IQR method was preferred to ensure a cleaner dataset.

Correlation Heatmap & Feature Reduction :-
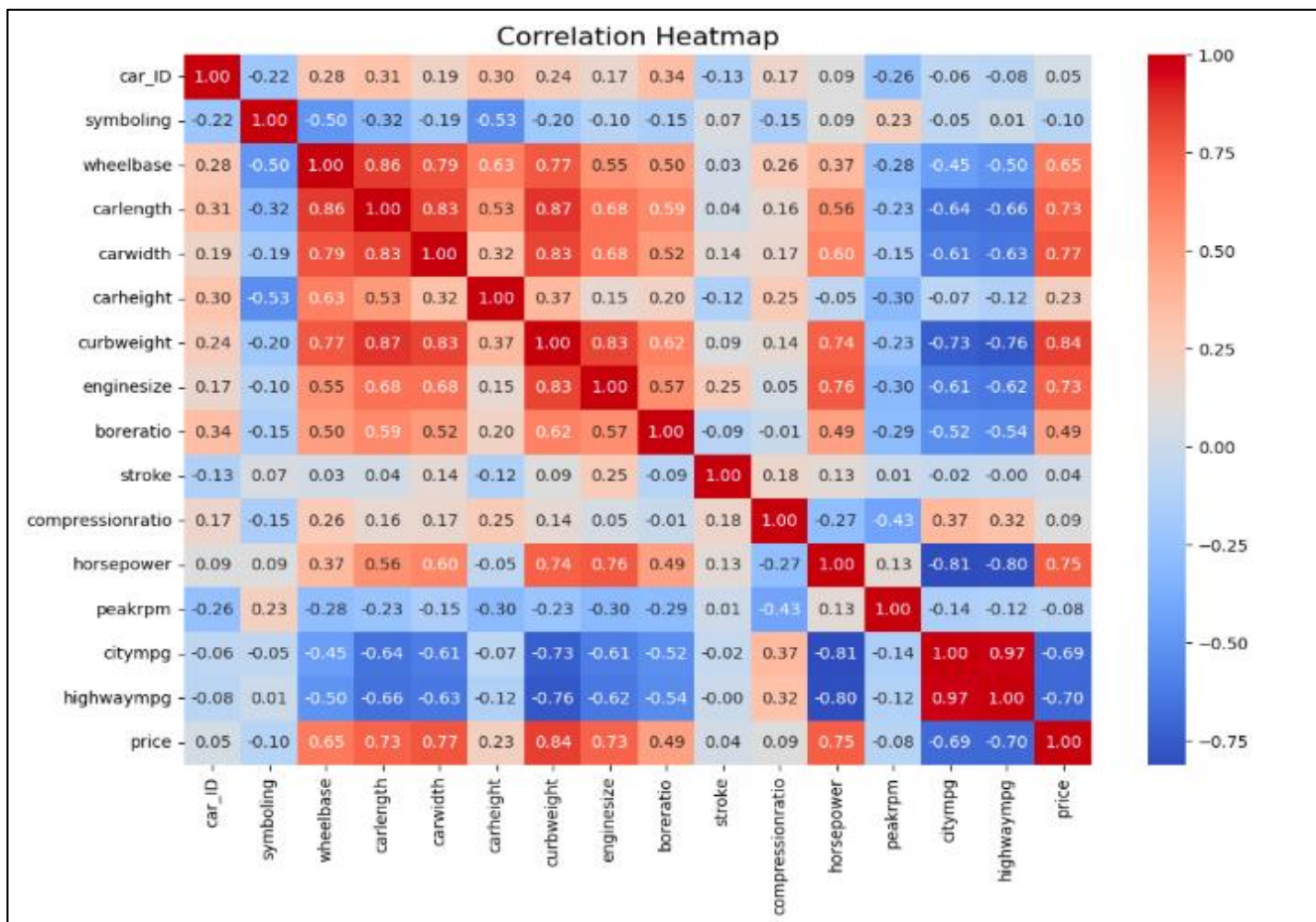
A correlation heatmap was generated to identify relationships between different variables. Based on the matrix, certain columns showed very weak correlation correlation values (close to 0), indicating that they contributed very little to explaining car prices.

To improve the dataset quality while keeping the original dataset intact, a new copy was created and the following columns were dropped due to low correlation:

- **Car ID** (r = 0.05)

- **Symboling** (r = -0.10)

- **Car Height** (r = 0.23)

- **Stroke** (r = 0.04)

- **Compression Ratio** (r = -0.09)

- **Peak RPM** (r = -0.08)

Key Insight:
The dataset is now well-prepared for analysis, with irrelevant identifiers dropped, spelling standardized, and outliers systematically handled. These steps improve reliability and ensure more accurate insights in subsequent analysis.

Car Price Distribution :-
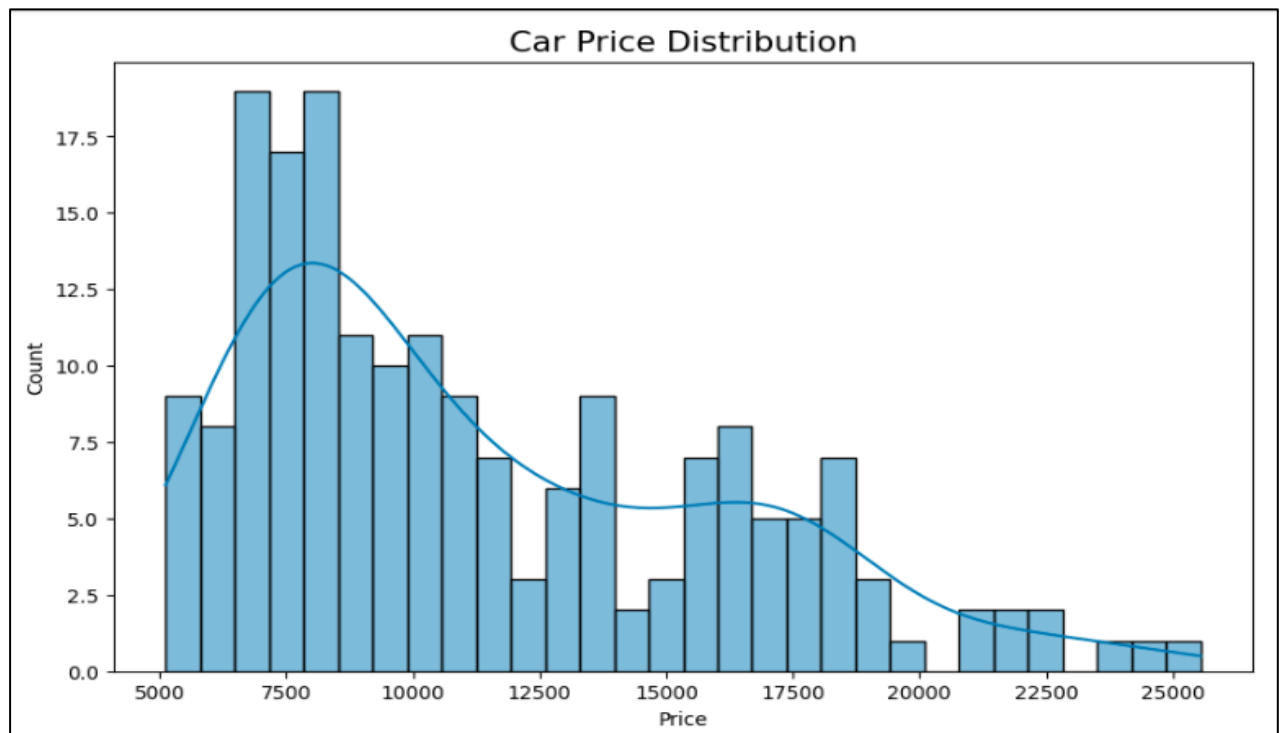
- Car prices are not evenly distributed; the majority of vehicles fall in the **$5,000–$12,000** range.

- Very few cars are priced above **$20,000**, showing the presence of high-end or premium models.

- The distribution is **right-skewed**, indicating there are many more affordable cars compared to expensive ones.

Key Insight:
The dataset shows a right-skewed distribution where most vehicles fall in the **$5,000–$12,000** range, reflecting dominance of the affordable and mid-market segment. Cars priced above **$20,000** are limited, indicating that premium models exist but cater to a niche audience. This suggests that while the bulk of sales comes from budget-friendly cars, premium models serve as **high-margin products** and contribute to brand positioning
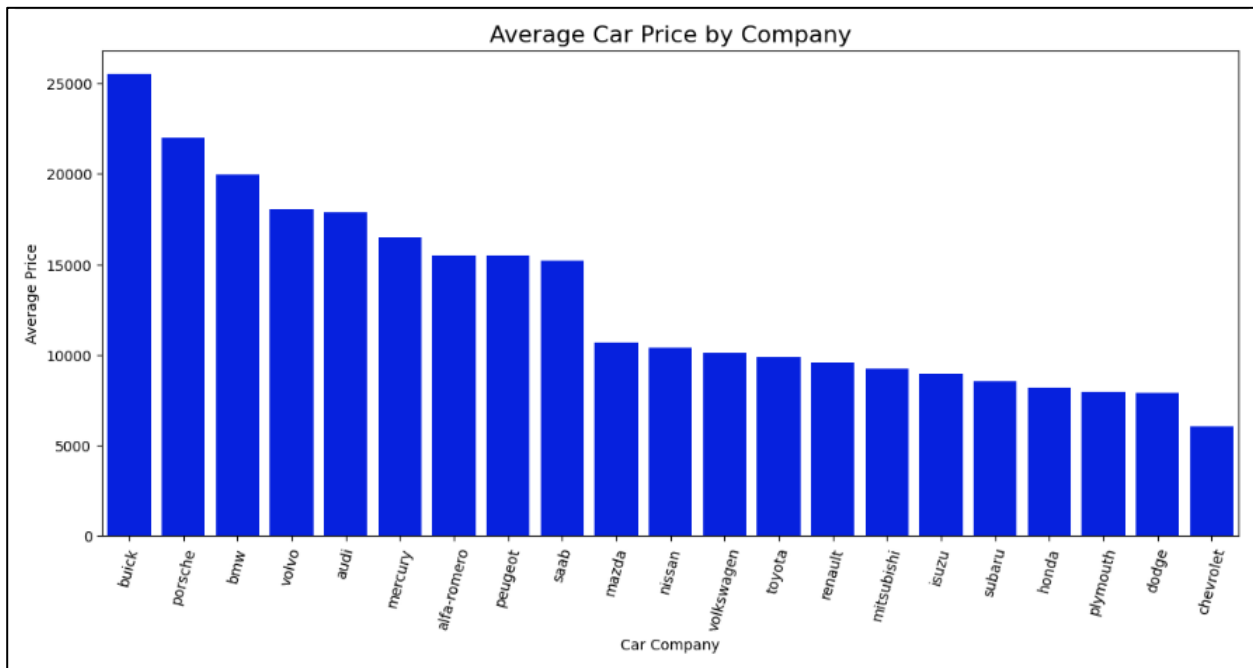
Pair Plot for Numeric data
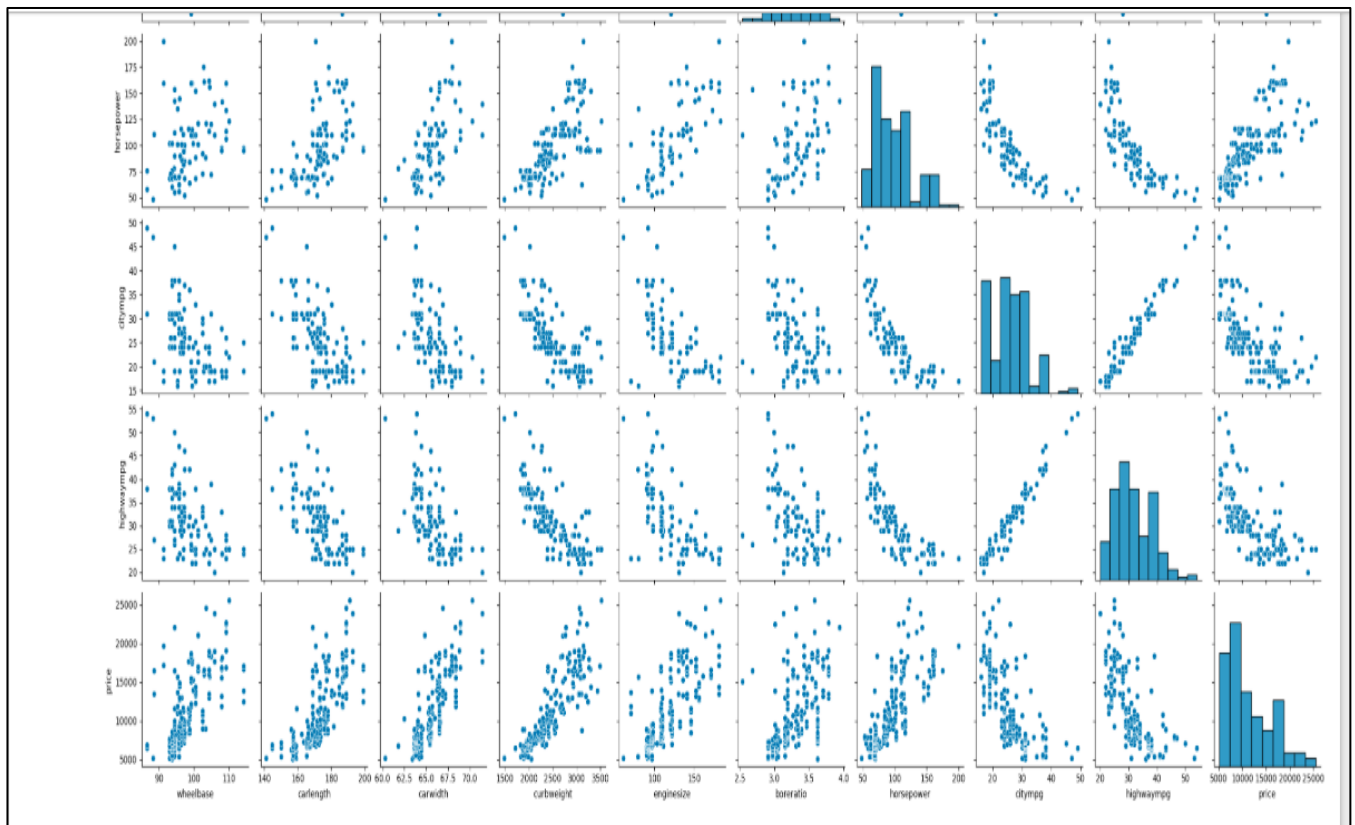
Insight:

Car price shows strong positive correlation with engine size, horsepower, curb weight, and car width, meaning bigger and more powerful cars are more expensive. In contrast, fuel efficiency (city/highway mpg) is negatively correlated with price, as fuel-efficient cars are generally more affordable.

Average Price by Car Company :-

- **Buick** cars have the highest average price (≈ $25,000+), showing their dominance in the luxury segment..

- **Porsche** (≈ $22,000) and **BMW** (≈ $20,000) also fall in the premium category.

- **Volvo, Audi**, and **Mercury** are in the upper mid-range (≈ $17,000–$18,000).

- **Peugeot, Saab, Nissan, Mazda, Toyota, Volkswagen** sit around $10,000–$15,000, reflecting the mid-market consumer base.

- **Honda, Subaru, Dodge, Plymouth, Chevrolet** have the lowest average prices (≈ $6,000–$9,000), making them budget-friendly and mass-market oriented.

Key Insight:
Brand plays a **critical role in car pricing**—luxury brands (Buick, Porsche, BMW) command a **significant price premium**, while budget brands (Chevrolet, Dodge, Honda) target the **affordable segment**. This highlights the **brand perception effect**, where consumers are willing to pay extra for **status, reliability, and prestige** associated with certain companies.

Average Price by Fuel Type :-
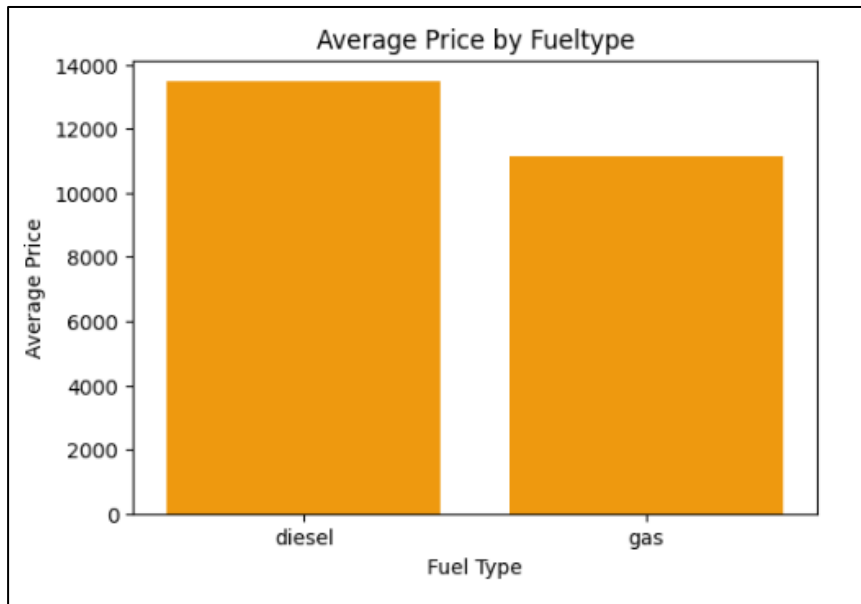
- Cars with **diesel engines** have a higher average price (around **$13,700**) compared to gas-powered cars (around **$11,000**).

- The price gap suggests that diesel cars are positioned more towards the **premium segment**.

- Gas-powered cars dominate the **budget and mid-range categories**, making them more accessible to the mass market.

Key Insight:
Diesel cars tend to be **more expensive and premium** compared to gas cars, which indicates that **gas-powered vehicles are the preferred choice for affordability**, while diesel variants are targeted towards higher-end consumers.

# Average Price by Car Body Type :-

Average Price by Car Body Type
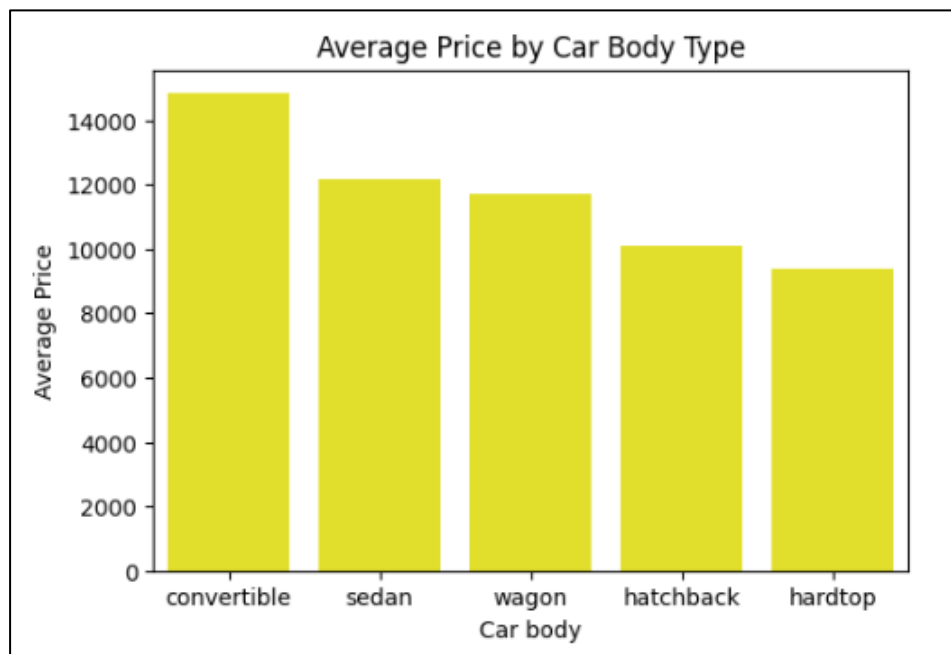
- **Convertibles** have the highest average price (around $15,000), placing them firmly in the luxury segment.

- **Sedans and wagons** follow with moderate average prices (around $12,000), showing they cover the mid-range market.

- **Hatchbacks** are priced lower (around $10,000), targeting the budget-friendly category.

- **Hardtops** have the lowest average price (≈ $9,500), suggesting they are less premium and more accessible.

Key Insight:
Car body type strongly influences price positioning—**convertibles are premium luxury vehicles**, while **hatchbacks and hardtops appeal to budget-conscious buyers**, and **sedans/wagons serve the mid-tier consumer base**.

Average Price by Drive Wheel :-
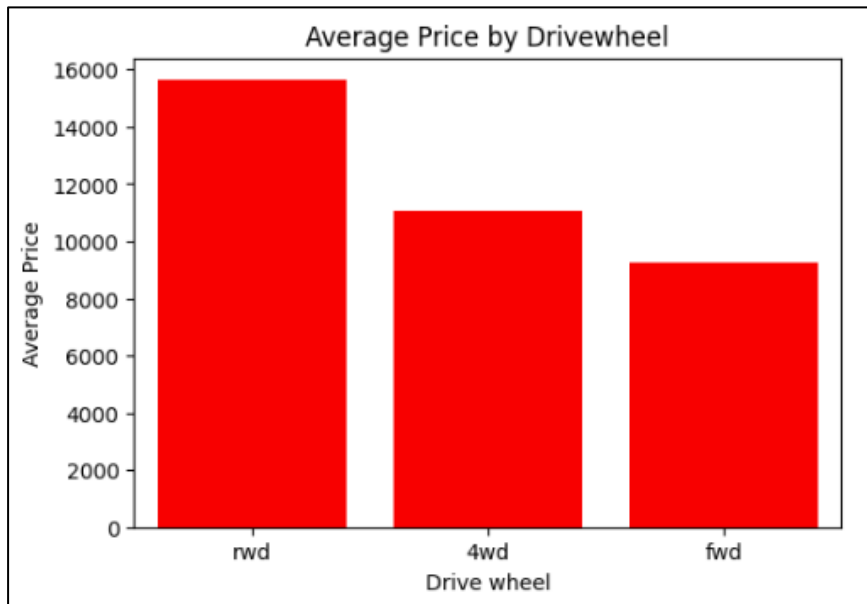
- **Rear-wheel drive (RWD)** cars have the highest average price (≈ $15,700), reflecting that luxury and sports cars often use RWD for better performance and handling.

- **Four-wheel drive (4WD)** vehicles come in the mid-range (≈ $11,000), likely due to their specialized use in off-road or utility contexts.

- **Front-wheel drive (FWD)** cars have the lowest average price (≈ $9,300), as they are the most common, economical, and mass-market option.

Key Insight:
Drive wheel configuration is linked to market positioning — **RWD dominates the premium/luxury segment**, **4WD targets utility/off-road needs**, while **FWD caters to budget-friendly, everyday consumers**.

Engine Size vs Price :-
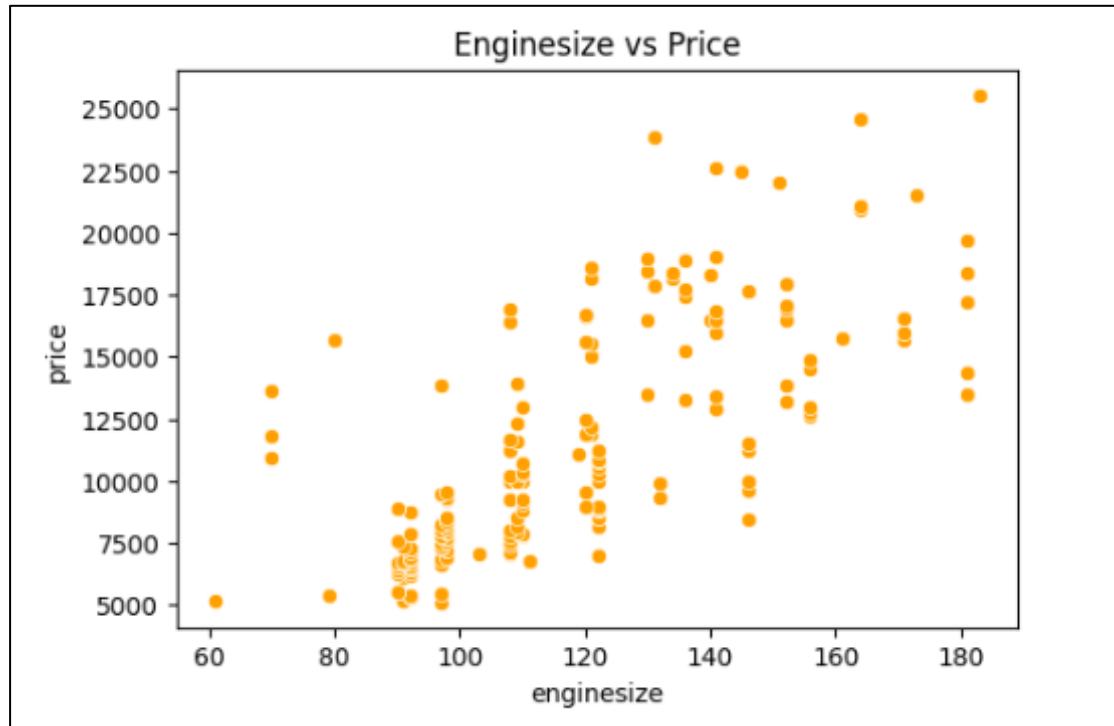
- **Small engines (60–100 cc):** Cars in this range are priced around $5,000–$10,000, representing the entry-level or economy segment focused on affordability.

- **Medium engines (120–150 cc):** Prices rise to ~$12,000–$18,000, indicating the mid-range market where consumers expect better performance and added comfort features.

- **Large engines (160–180+ cc):** These vehicles above $20,000, often falling into the luxury or high-performance category, appealing to buyers seeking power and prestige.

Key Insight:
Engine size shows a clear positive relationship with price — smaller engines dominate budget-friendly cars, mid-sized engines balance cost and performance, while larger engines are strongly associated with premium and luxury vehicles.

Curbweight vs Price Insights :-

Curbweight vs Price

(scatter plot of price vs curbweight, red points)
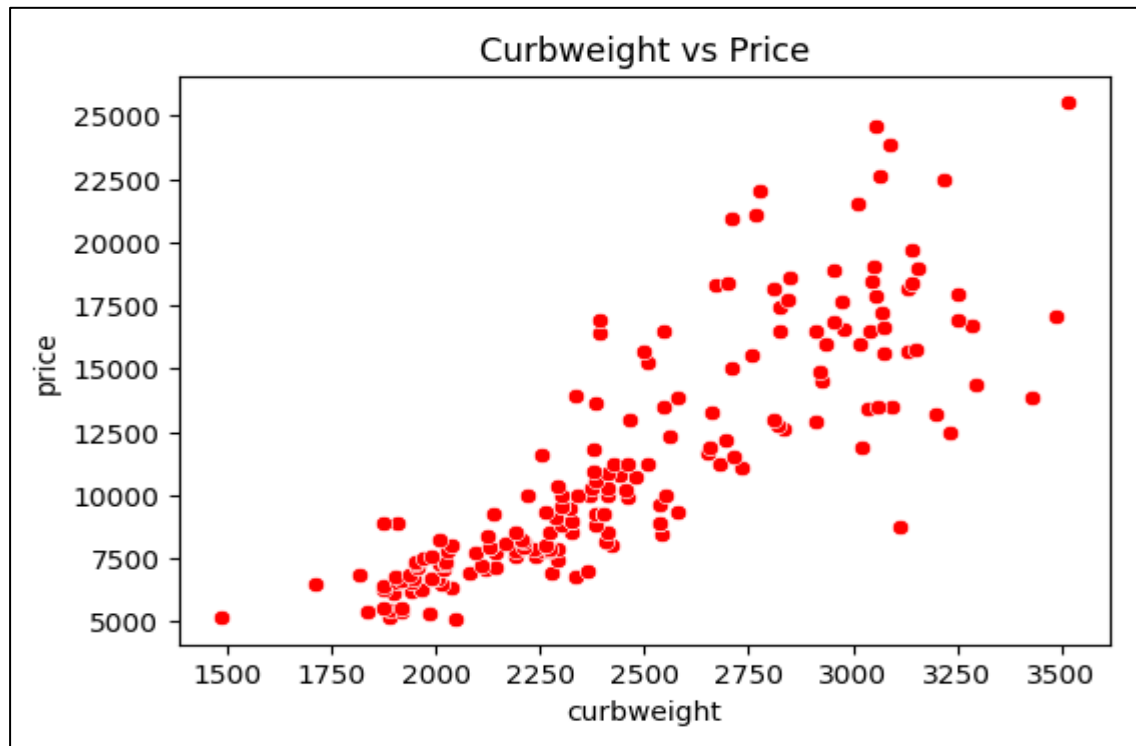
- There is a **clear positive relationship** between **curbweight** and **car price**.

- **Lighter cars (1500–2000 lbs)** are mostly in the **budget range** (≈ **$5,000–$10,000**).

- **Mid-weight cars (2200–2700 lbs)** fall in the **mid-price category** (≈ **$10,000–$17,000**).

- **Heavy cars (3000+ lbs)** are generally **premium models**, often priced above **$20,000**.

Key Insight:
Heavier cars usually come with larger engines, stronger body structures, and additional luxury or safety features. As a result, their prices are significantly higher compared to lighter vehicles, making weight an important factor in car pricing**.**

Horsepower vs Price Insights :-
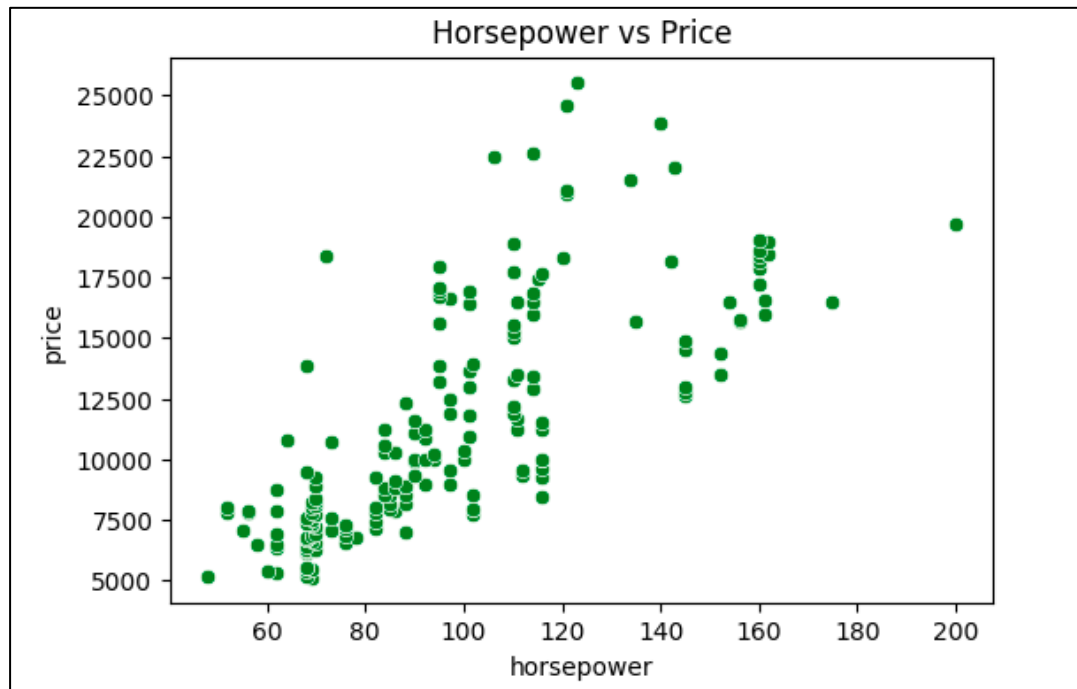
- There is a strong positive correlation between **horsepower** and **car price**.

- Cars with lower **horsepower** (50–90 HP) are in the **budget** range (≈ $5,000–$10,000).

- **Mid-weight** cars (2200–2700 lbs) fall in the **mid-price** category (≈ $10,000–$17,000).

- **Heavy cars** (3000+ lbs) are generally **premium models**, often priced above $20,000.

Key Insight:
Higher **horsepower** cars command a significantly higher price, as they are positioned in the **premium** and **luxury segments**. Lower horsepower vehicles remain affordable, catering to budget-conscious buyers.

Conclusion :-

This analysis of car prices using Exploratory Data Analysis (EDA) highlights how technical specifications, design features, and performance metrics directly influence vehicle pricing.
The dataset, after cleaning and outlier treatment, revealed consistent patterns across multiple variables:

- **Brand Effect:** Luxury brands such as Buick, Porsche, and BMW command a significant price premium due to consumer perception of prestige, quality, and performance, while budget brands like Honda, Dodge, and Chevrolet dominate the affordable market.

- **Technical Factors:** Engine size, curb weight, and horsepower show a strong positive correlation with price. Larger, heavier, and more powerful cars are consistently positioned in the premium segment, while smaller, lighter, and less powerful models serve cost-sensitive buyers.

- **Design Features:** Convertibles and rear-wheel-drive cars are priced higher, indicating their positioning as luxury/performance vehicles, while hatchbacks and front-wheel-drive cars target budget-friendly consumers.

- **Fuel Type:** Diesel cars generally cost more than gas-powered cars, suggesting their association with durability and performance in higher-end segments.

Business Implications:

- These insights can help in setting the **right prices** by matching car features with the right customer groups.

- Manufacturers can **position their models more effectively** — for example, promoting **small-engine hatchbacks** to budget-conscious buyers and **luxury convertibles/RWD cars** to premium customers.

- Dealers can use these findings to plan **marketing campaigns and inventory** more efficiently, ensuring they stock and promote the models that are most in demand in their local markets.

Overall Insight:
Car price is driven by both **brand value and technical specifications**. Luxury and performance-oriented features (large engines, RWD, high horsepower, heavy curb weight, premium body types) consistently push cars into higher price ranges, while compact, fuel-efficient, and FWD-based models dominate the affordable market.