Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

This dataset comprises vehicle specifications and pricing information, focusing on the relationship between wheelbase and price across various types. The primary objective of this analysis is to identify trends and insights that can aid in understanding market dynamics.

The dataset includes different categories of vehicles, such as convertible, wagon, sedan, hatchback, and hardtop.

The dataset is explained by the following columns :

1. **Normalised-losses:** This column indicates the normalised average loss of a vehicle's value.
2. **make:** This column specifies the manufacturer or brand of vehicle, such as alfa-romero or audi.
3. **fuel-type:** This column indicates the type of fuel the vehicle uses, such as "gas" or "diesel".
4. **aspiration:** This describes the engine aspiration type, such as standard (std) or turbo.
5. **num-of-doors:** This column indicates the number of doors on the vehicle..
6. **body-style:** This describes the style of the vehicle's body, such as convertible or sedan.
7. **wheelbase:** This represents the distance between the front and rear axles of the vehicle, measured in inches. It can influence the vehicle's handling and stability.
8. **number-of-cylinders:** This column indicates the number of cylinders in the vehicle's engine, affecting performance and fuel efficiency.
9. **engine-size:** This column represents the size of the engine in litres, which can influence the power output and fuel consumption.
10. **compression-ratio** This measures the ratio of the volume of the cylinder when the piston is at the bottom of its stroke to the volume when it is at the top. It affects engine performance and efficiency.
11. **horsepower:** This column indicates the engine's horsepower, a measure of its power output.
12. **peak-rpm:** This indicates the engine speed at which peak horsepower is achieved, measured in revolutions per minute (rpm).

13. **city-mpg:** This column represents the vehicle's fuel efficiency in miles per gallon (mpg) when driving in the city.
14. **highway-mpg:** This indicates the vehicle's fuel efficiency in mpg when driving on the highway.
15. **price:** This column represents the vehicle's price in U.S dollars.

DATA CLEANING

The following columns ['drive-wheels', 'symboling', 'engine-location', 'fuel-system', 'stroke', 'bore', 'length', 'width', 'height', 'curb-weight'] were removed from the dataset as they will not be used in th analysis.

- Used the head() method to view the data.
- Changed the data types of the following columns; normalised-losses, price, peak-rpm, and horsepower to numpy int64.
- Missing columns with numeric values were replaced with the mean

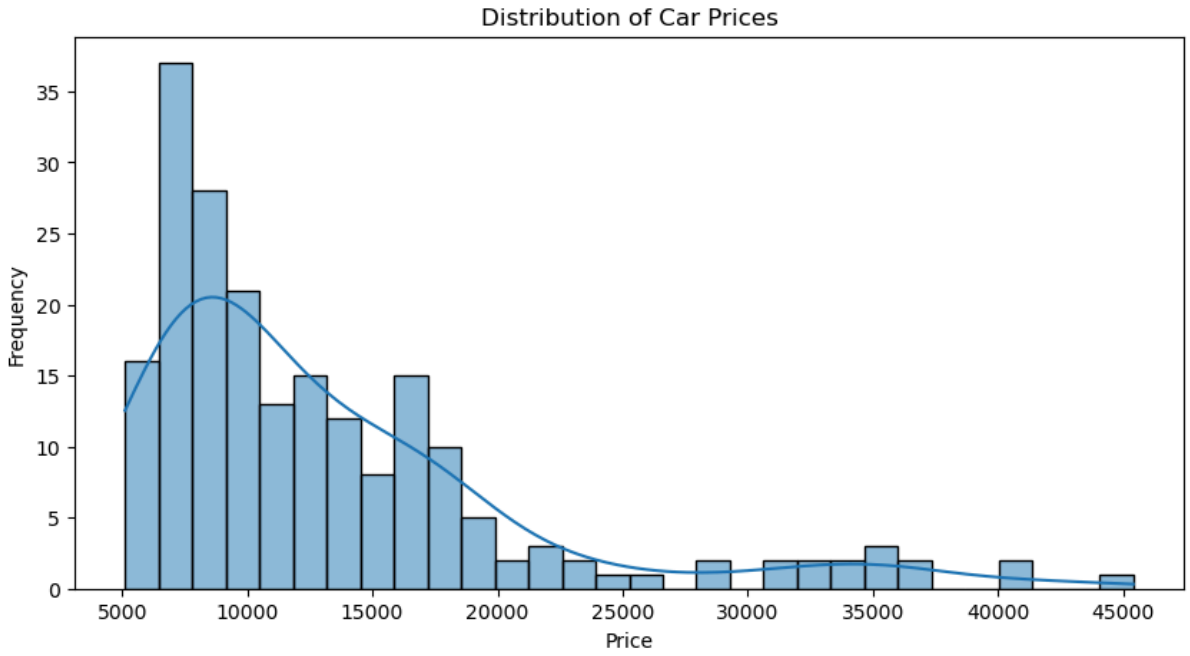
MISSING DATA

The data set had missing values represented by "?". These missing values were replaced with a numpy value nan. The columns that have missing values are normalised-losses, num-of-doors, horsepower, peak-rpm, and price.
Summary of missing values in each column:
normalised-losses: 42
num-of-doors: 2
horsepower: 2
peak-rpm: 2
price: 4

DATA STORIES AND VISUALISATIONS

Price distribution

This histogram represents the distribution of car prices within a dataset. This type of plot helps to reveal the overall pattern of price ranges, allowing us to understand which price segments are most common and how the data is spread across the spectrum. Analysing the distribution of car prices provides insights into market dynamics, affordability, and the relative proportion of budget versus luxury vehicles.



Observations and insights:

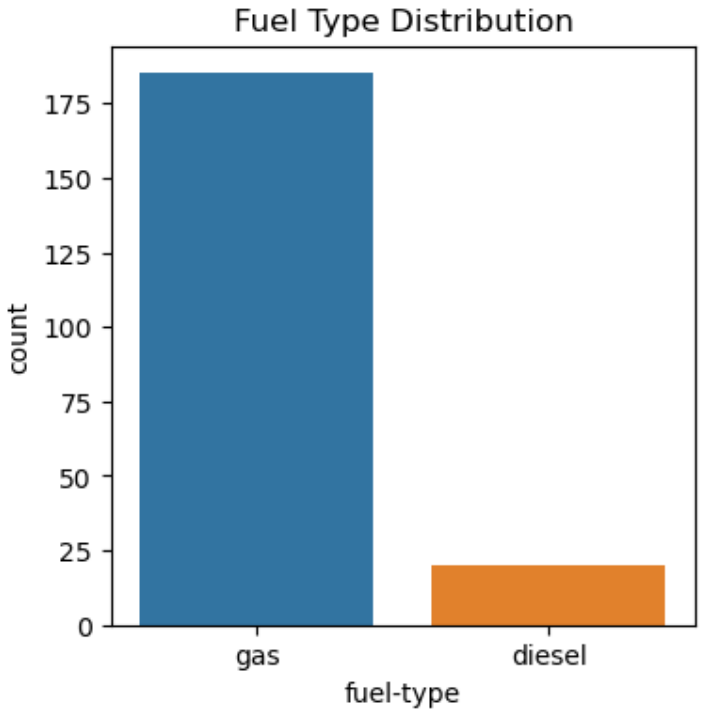
1. **Shape of Distribution:**
   The histogram shows a right-skewed distribution, with a high frequency of car prices concentrated in the lower price range and a gradual decline as prices increase. This skewness suggests that the majority of cars fall within a lower price bracket, while fewer cars are in the higher price range.

2. **Common Price Range:** Most cars are priced between $5000 and $15000, indicating an affordable segment for the majority of vehicles in the dataset.

3. **Outliers-High-End Cars:**
   Cars priced above $30 000 are relatively rare, showing that high-end cars represent a smaller proportion of the dataset.

Fuel Type Distribution

The following barplot shows the type of fuel that each vehicle in the dataset uses. Most vehicles in the dataset use gasoline instead of diesel.



Boxplot of Price by Body Style

The following boxplot displays car prices across different body styles. Boxplots are effective for showing the spread, median, and variability of data across categories, helping to identify typical price ranges for each body style as well as any significant outliers. By examining the price ranges associated with different body styles, we can gain insights into which types of cars tend to be more affordable or expensive, as well as observe variability within each category.

Comparison Across Body Styles:

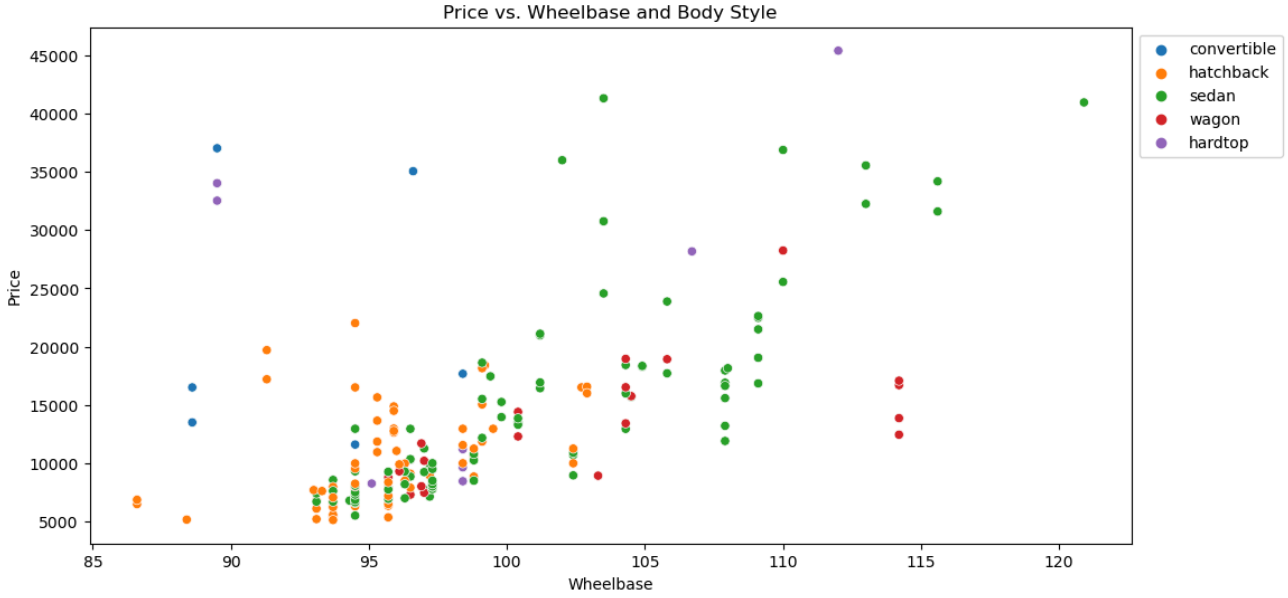
1. **Convertibles and hardtops** have the widest price ranges and higher median prices, indicating that these body styles tend to be more expensive and possibly include premium luxury models.

2. **Hatchbacks and wagons** show lower price ranges, with hatchbacks having the lowest median and interquartile range, suggesting they are generally more affordable and could be entry-level or economy models.
3. **Sedans** fall in between, with a moderate price range and a few outliers, indicating a balanced mix of economy and mid-range models.

Price vs Wheelbase and Body Style

The following scatter plot provides an analysis of vehicle prices in relation to wheelbase and body style. Observing the relationship between price and wheelbase can reveal how the dimensions of a vehicle and its structural design

impact its market value. In general, vehicles with a longer wheelbase tend to be more expensive, suggesting that larger, more spacious vehicles may appeal to a different market segment.



1. **Sedan Dominance:** Sedans are well-represented across various wheelbase and price points, indicating that they are a popular choice and cater to a wide market. They generally have a steady increase in price with the increase in wheelbase length, highlighting a correlation between size and perceived value in this segment.

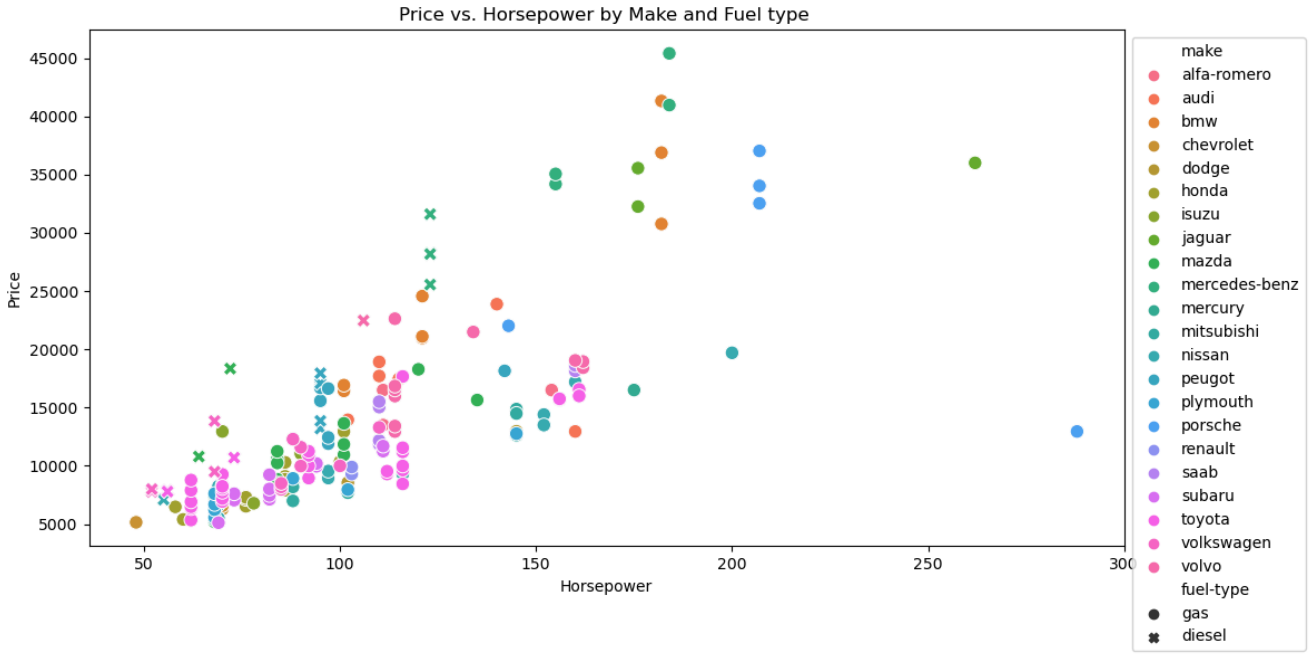
2. **High-Value Convertibles and Hardtops:** Convertibles and hardtops appear to command higher prices, even at shorter wheelbases. This might reflect the premium value associated with these body styles, as they are often seen as more luxurious or performance-oriented.

3. **Wagons and Hatchbacks:** Wagons and hatchbacks, on the other hand, occupy the lower to mid-price range, with moderate wheelbase lengths. These body styles seem to target a more budget-conscious market or consumers looking for practically without premium pricing.

This visualisation highlights the market segmentation among body styles, showing that while sedans cover a broad spectrum, convertibles, hardtops, wagons, and hatchbacks tend to occupy specific niches with defined price points.

Price vs Horsepower by Make and Fuel Type

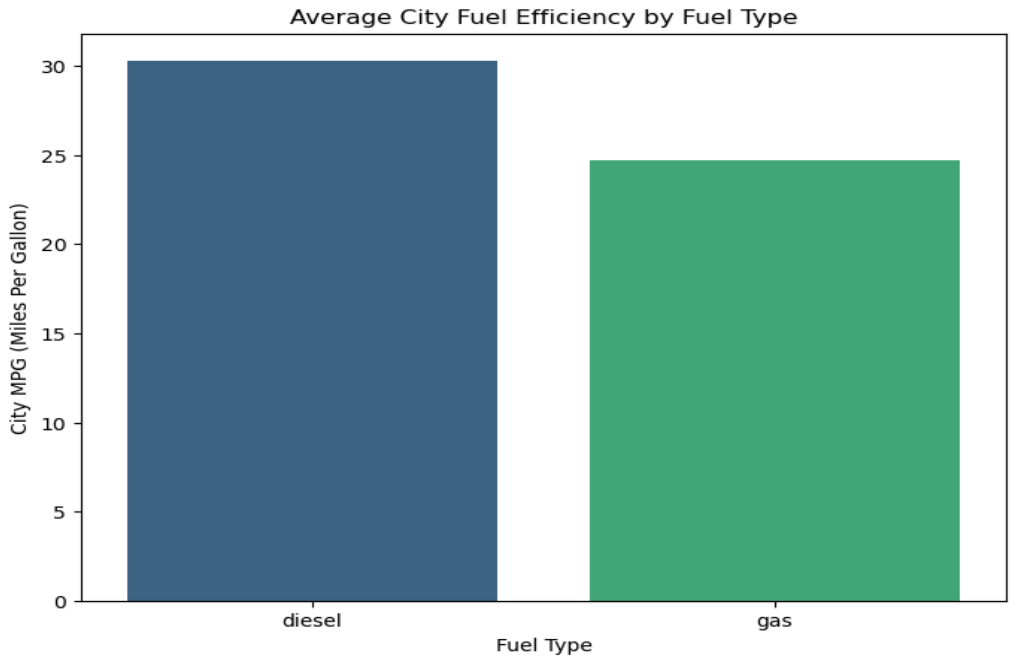
The following scatter plot explores the relationship between price and horsepower, segmented by vehicle make and fuel type (gas vs diesel). Horsepower is a crucial indicator of vehicle performance, and understanding how it influences price across different brands and fuel types can provide insights into market positioning and brand perception.



1. **Higher Horsepower. Higher Price:** There is a clear positive correlation between horsepower and price across all brands. Vehicles with higher horsepower tend to be more expensive, which aligns with consumer expectations, as performance-oriented vehicles are generally priced at a premium.

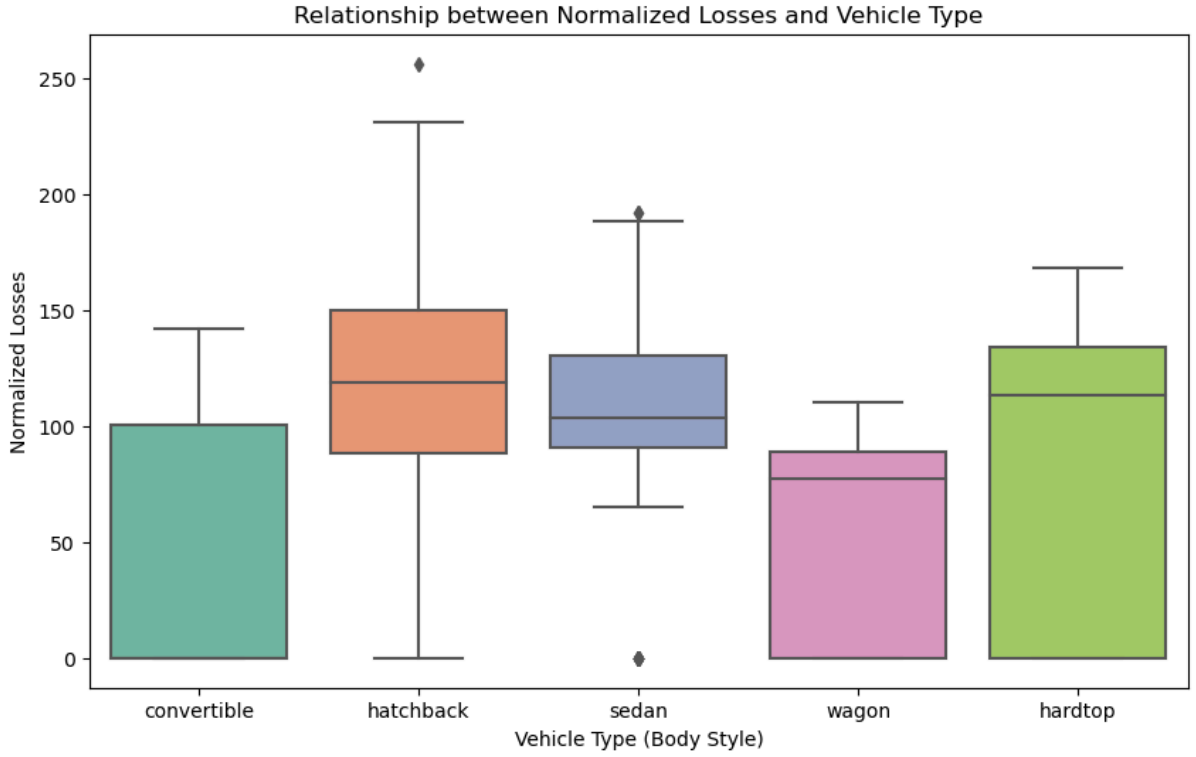
2. **Brand-Specific Trend:** Luxury and performance brands like BMW, Porsche, and Mercedes-Benz have data points in the higher horsepower and price ranges, indicating their positioning in the premium market. These brands offer higher performance at a premium price, catering to consumers willing to pay for quality and brand prestige.
3. **Fuel Type Differentiation:** Diesel vehicles, represented with a different marker style, are generally clustered in the mid to high horsepower range but cover a wide price range. Diesel engines are often associated with better fuel efficiency and torque, which could attract consumers looking for both performance and long-term savings. However, diesel options appear less frequently than gas vehicles, indicating a niche market.

The following bar graph shows the average city fuel efficiency by Fuel Type. From the plot, it is evident that diesel cars are more efficient than gasoline cars.



Normalised-Losses and Vehicle Type

The following boxplot will give more insights into the variability of city fuel efficiency for each fuel type, showing the range, median, and potential outliers.



1. **Convertibles:** The median normalised loss for convertibles is moderate, around the middle of the y-axis range. There is a relatively narrow interquartile range (IQR), meaning most convertibles have similar normalised losses, indicating a consistent risk profile. The lack of extreme

outliers suggests that convertibles are generally perceived as moderate-risk vehicles without much variation.

2. **Hatchbacks:** Hatchbacks display a wide range in normalised losses, with the highest variability among all body styles. The median is higher compared to other types, indicating that hatchbacks tend to have a higher risk profile. The plot also shows outliers above and below the main distribution, reflecting some hatchback models with very high or very low risk factors. This could be due to the diversity within the hatchback category, with both economical and performance-focused models included.

3. **Sedans:** Sedans have a moderate median for normalised losses, slightly lower than that of hatchbacks. There's a smaller spread in the upper quartile but a larger one in the lower range, meaning that while most sedans have similar normalised losses, a few models stand out with much lower risk factors. The presence of an outlier below the main distribution could indicate specific sedan models that are designed to be safer or have lower risk factors

4. **Wagons:** Wagons have the lowest normalised losses overall, with the median close to the bottom of the y-axis range. The narrow IQR indicates that most wagons are low-risk, with little variation, suggesting a high level of reliability or safety associated with this body style. This body style has no outliers, which could indicate that wagons have a uniform risk factor

5. **Hardtops:** Hardtops show higher normalised losses than sedans and wagons but have a narrower spread. This body style has no outliers, which could indicate that hardtops have a uniform risk factor.

THIS REPORT WAS WRITTEN BY : Vitumbiko