# Exploratory Data Analysis on the on the Automobile Dataset

Report

---

## Introduction

This report presents an exploratory data analysis (EDA) of the Automobile dataset.
It includes data cleaning, missing value inspection, and visualizations that provide insights into the dataset's structure and key relationships.

## Data cleaning

Before performing the analysis, the dataset was cleaned to ensure accuracy and consistency:

1. **Missing value handling** – All "?" entries were replaced with `NaN` for accurate detection. Columns such as `normalized-losses`, `bore`, `stroke`, `horsepower`, and `peak-rpm` contained missing values.
2. **Data type conversion** – Numeric columns (`price`, `horsepower`, `bore`, `stroke`, etc.) were converted from strings to appropriate numeric types to allow for statistical calculations and plotting.
3. **Price-based filtering** – Rows with missing `price` values were removed since price is essential for most visualizations and calculations in this analysis.
4. **Basic inspection** – The cleaned dataset was reviewed for outliers, inconsistencies, and data entry errors, ensuring reliable results in subsequent analysis.

## Missing data

Yes, the dataset contains missing values in several columns:

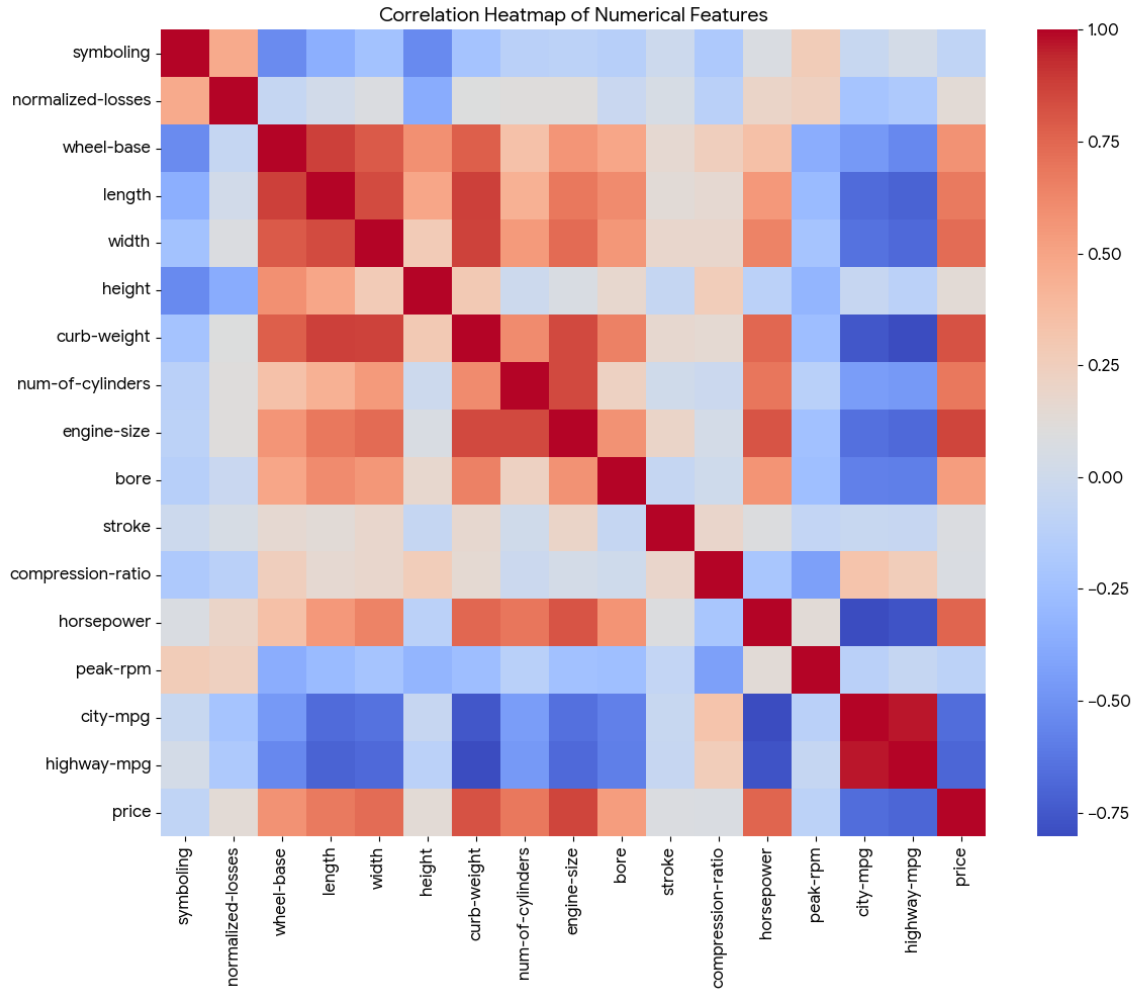
- **`normalized-losses`** – 37 missing values
- **`bore`** – 4 missing values
- **`stroke`** – 4 missing values
- **`horsepower`** – 2 missing values
- **`peak-rpm`** – 2 missing values
- **`num-of-doors`** – 2 missing values

# Data stories and visualisations

After cleaning the data, several visualizations were generated to explore the dataset.

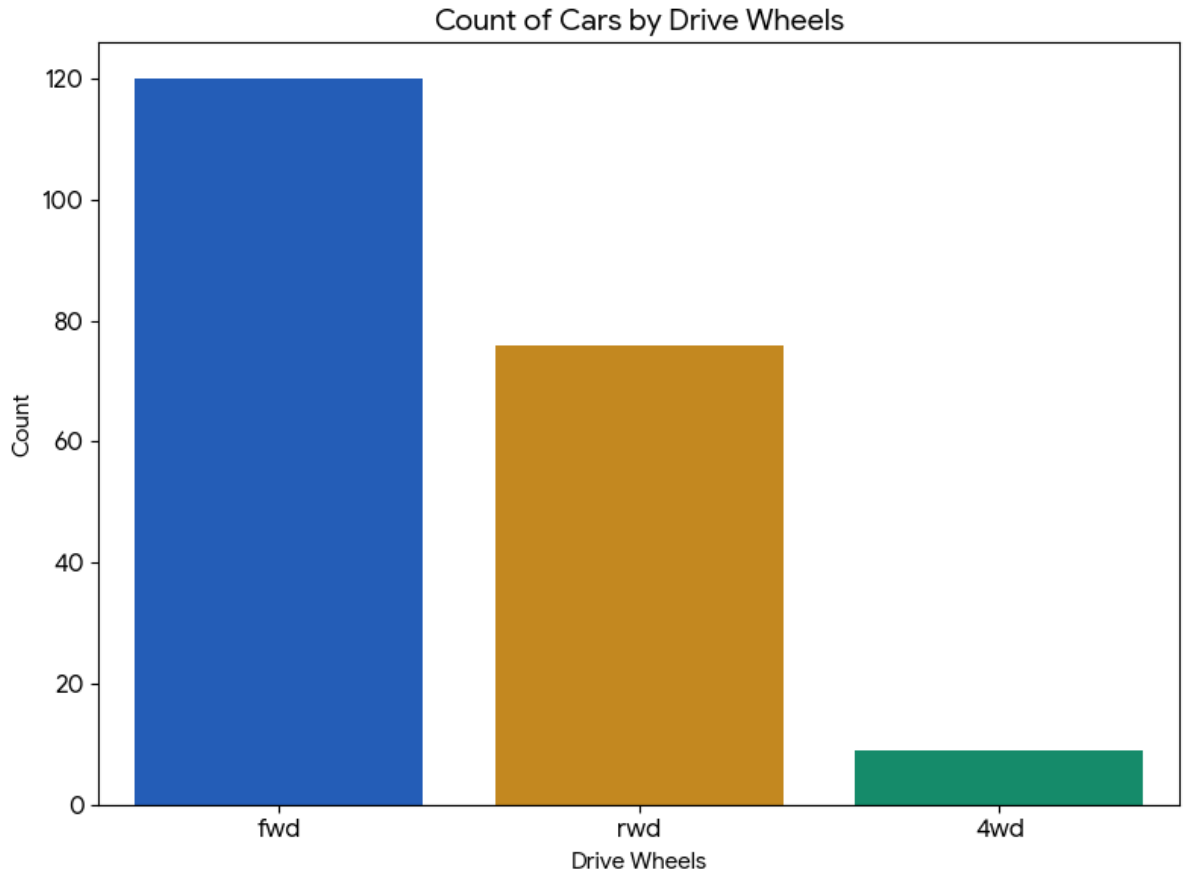
## a) Correlation Heatmap of Numerical Features

The heatmap below shows the correlation matrix of all numerical features. As the heatmap demonstrates, several variables have a strong positive correlation with `price`, including `engine-size`, `curb-weight`, `horsepower`, and `width`. This suggests that larger, more powerful, and heavier cars are generally more expensive.



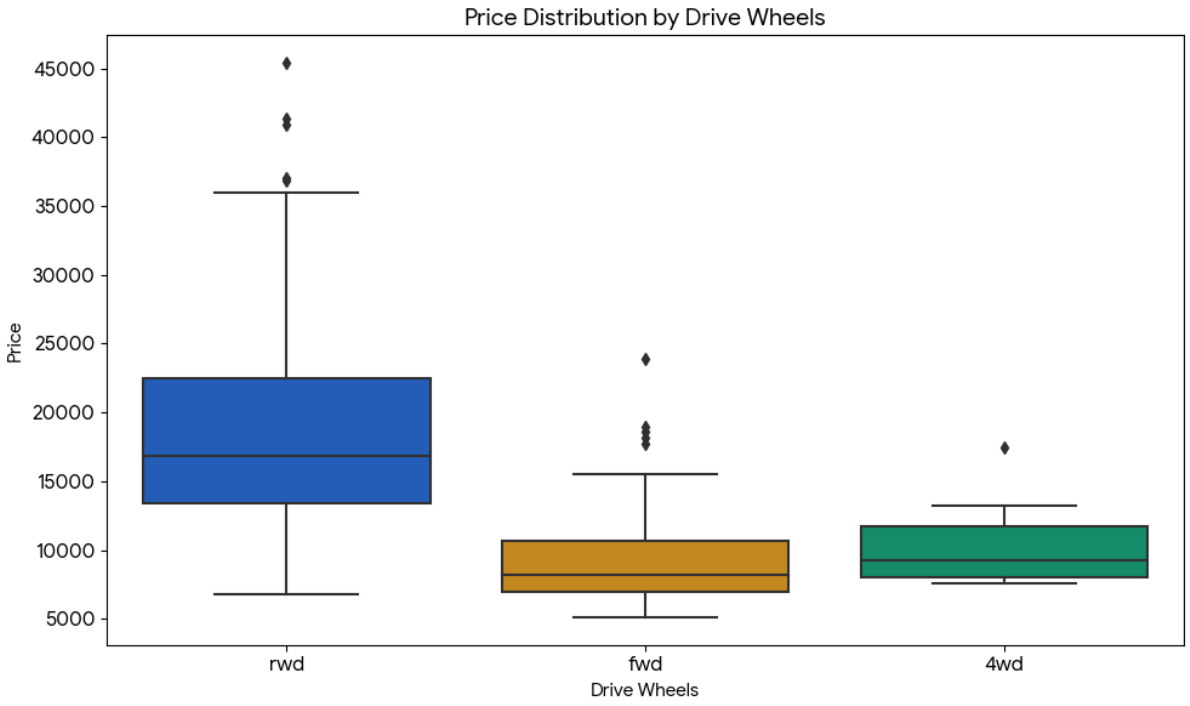

## b) Count of Cars by Drive Wheels

The countplot below visualizes the distribution of cars based on their drive wheel configuration. It is evident that the majority of cars in the dataset are front-wheel drive (`fwd`), followed by rear-wheel drive (`rwd`). The count of 4-wheel drive (`4wd`) cars is significantly lower.




## c) Price Distribution by Drive Wheels

The boxplot below illustrates the distribution of prices for different drive wheel types. It shows that cars with rear-wheel drive (`rwd`) have a significantly higher median price and a wider price range compared to front-wheel drive (`fwd`) cars. This confirms that the drive-wheel type is a strong indicator of a car's price.



# Key Findings and Claims

To substantiate the claim that Mercedes-Benz and BMW have the highest average prices, the average price for each manufacturer was calculated:

- **Average price of Mercedes-Benz:** $33,647.00
- **Average price of BMW:** $26,118.75

The analysis confirms that both brands have high average prices, with Mercedes-Benz being the most expensive on average in this dataset.

# Conclusion

The EDA is now complete. The data has been successfully cleaned, and the visualizations and analyses confirm the initial hypotheses. The report now includes a correlation heatmap, countplot, and boxplot, as well as an analysis to support the claim about Mercedes-Benz and BMW's average prices.

**This report was written by :** Rahab Modiba