

**TASK**

**Exploratory Data Analysis on the Automobile Data Set**

[](https://www.hyperiondev.com/)

**Introduction**

The automobile dataset presents an extensive array of information pertaining to various vehicles, encompassing details on size, power, pricing, physical characteristics, engine specifications, fuel efficiency and other relevant attributes. The data contained a few missing values, mostly numerical which were dealt with appropriately. Once cleaned, the data exploration in this study focused on the price of cars and influencing factors. The average price of different car makes revealed distinct price groups with more car makes being found in lower price ranges. A scatter plot matrix of the numerical data revealed relationships between the price of cars and the car’s power and size. Analyses of categorical features through violin and count plots illustrated the market’s demand for affordable cars. Overall, the dataset provides valuable insights into the dynamics of the automotive market, emphasizing the prevalence of cost-conscious consumer preferences.

**DATA CLEANING**

* The first 10 rows of the dataset were viewed.
* The data types of each column were obtained.
* A description of the data summarising the numerical columns was obtained.
* Duplicate rows were removed.
* Missing values represented by ‘?’ were replaced with NA values for easier manipulation.
* The number of missing values in each column were obtained.

**MISSING DATA**

* The missing values in numeric columns were replaced by the mean of the other values in the column. A function was created to do this called clean\_numeric\_with\_mean with parameters ‘df’ (the dataframe) and ‘column’ (the column name to be cleaned). This was done for the following columns: normalized-losses, price, horsepower, bore, stroke, peak-rpm.
* The records containing missing values in string columns were removed. This was done the num-of-doors column where two records were removed.

**DATA STORIES AND VISUALISATIONS**

There were numerous relationships among the dataset that one could explore. For this exploration, I have focussed on the price of cars and the factors that influence the price. Three visualisations were made for this data exploration:

1. **Average price by make:**

Figure 1 shows a bar plot of the average price for each make of car. From this, three price groups can be identified: high, middle and low average prices. The groups are as follows:

* High average price: Jaguar, Mercedes-Benz, Porsche and BMW.
* Middle average price: Volvo, Audi, Mercury, Alfa-Romero, Peugeot and Saab.
* Low average price: Isuzu, Mazda, Nissan, Volkswagen, Toyota, Renault, Mitsubishi, Subaru, Honda, Plymouth, Dodge and Chevrolet.

We see that the high average price group contains 4 car makes, the middle group has 6 makes and the low group has 12 makes. This shows the demand for more affordable cars since there is a wider variety of car makes in the lower price ranges.

A graph of average price

Description automatically generated

Figure 1

1. **Numerical features influencing the price:**

A scatter plot matrix with wheel base, length, width, height, curb weight, engine size, horsepower, city mpg, highway mpg and price is seen in Figure 2. The following relationships with the price are identified:

* The curb weight, engine size and horsepower have fairly strong positive correlations with the price. This indicates that more powerful cars are more expensive.
* The wheel base, length and width have a weaker positive correlation with the price. The correlation is stronger in the lower price range and gets weaker as the price increases. This indicates that larger cars tend to be more expensive, however, as the size of the car increases, the price of the car may be impacted more so by other features rather than the car size itself. There may be greater variation in features such as fuel type, drive wheels, engine type, number of cylinders and more, whereas smaller cars may have less variation in these features.
* There appears to be no correlation between the height and price.
* There is a strong negative correlation with the price and the city and highway mpg, indicating that more expensive cars tend to get less miles per gallon. This is likely due to more expensive cars having more power. This can be verified from the strong negative correlation seen between the city and highway mpg and the curb weight, engine size and horsepower.
* From the price histogram it is evident that there are far more cars in the lower price range, once again indicating that there is a higher demand for more affordable cars.

A chart of graphs and diagrams

Description automatically generated with medium confidence

Figure 2

1. Relationships between price and fuel type, drive wheels, engine type and the number of cylinders:

Violin plots were created of the price with fuel type, drive wheels, engine type and the number of cylinders, as well as count plots of fuel type, drive wheels, engine type and the number of cylinders on the same axes as the violin plots. These are shown in Figure 3. The following observations have been made:

* Price vs Fuel Type: cars with the gas fuel type have a lower mean price than diesel cars, as well as a higher concentration of cars in the lower price range. There are, however, anomalies with high prices, likely due to more sporty cars. The diesel fuel type cars have a higher mean price and more spread in the price distribution.
* Price vs Drive Wheels: rear wheel drive cars tend to be more expensive than front wheel and 4-wheel drive cars, with a much larger range in price too. The most common type of car is front wheel drive where the number of cars in this category are concentrated in the lower price range.
* Price vs Engine Type: it appears that the ohc engine type has the lowest mean price as well as being the most common engine type by far. The ohcv engine type has the highest mean price as well as one of the largest ranges in price together with the ohcf engine type. The ohcf engine type, however, has a low mean price.
* Price vs Number of Cylinders: the cheapest and most common cars are those with four cylinders. Five, six, eight and even two cylinder cars have higher mean prices, with the mean price of eight cylinder cars being particularly high. Three and twelve cylinder cars appear to have only one entry each thus having little information. However, the price of the twelve cylinder car is high, while the three cylinder price is low.
* From all four count plots we can again see the demand for more affordable cars since the fuel types, drive wheels, engine types and number of cylinders with the lowest mean prices all have the highest car count.

**A graph of different types of cylinders

Description automatically generated**

Figure 3

**THIS REPORT WAS WRITTEN BY : Kathleen Sellick**

