

**TASK**

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

[Graphical user interface, text

Description automatically generated](https://www.hyperiondev.com/)

# Introduction

This report presents an analysis of the "automobile.txt" dataset. The dataset contains a wide variety of information about different types of automobiles, including but not limited to car make, engine information, power output, fuel efficiency, and price. The data has been imported and organized into a structured format, referred to as the "automobile\_df" dataframe.

The dataset contains information on various cars and their properties, including:

* Car Type: This specifies the make of the car such as Audi or BMW.
* Engine information: This includes details about the engine, such as the number of cylinders, engine size, and fuel type.
* Power output: This includes information about the horsepower and torque of the engine.
* Fuel efficiency: This includes data on the car's fuel economy in the city and on the highway in miles per gallon (MPG).
* Price: This includes the base price of the car or an estimated value of the car based on its features and condition.

# Data Cleaning

In the process of preparing the "automobile.txt" dataset for analysis, several data cleaning steps were applied. Columns deemed irrelevant or unnecessary were removed from the data frame. These columns were 'aspiration', 'engine-type', 'length', 'width', 'height', and ‘engine-location’. Additionally, duplicate rows were removed. On inspection of the data types of the columns, it was seen that some number-based columns were classed as objects. In order to make the data consistent, columns ‘normalized-losses’, ‘price’, ‘peak-rpm’, and ‘horsepower’ were set as integer types and ‘stroke’ was set as a float type.

# Missing Data

On inspection of the columns, it was seen that columns ‘normalized-losses’ and ‘stroke’ had some missing data represented by a '?'. The missing data are considered to be Missing Completely at Random (MCAR) and could not be inferred from other parts of the data set. Therefore, the rows with missing data were removed from the dataset to ensure the accuracy and integrity of the data.

# 

# Data Stories and Visualisations

A scatter matrix was to gain an overview of various relationships in the data set. The variables, 'price', 'engine-size', 'horsepower', ‘highway-mpg’ are used to create the figure. The findings in this matrix were used to explore the data further.

Chart, histogram

Description automatically generated

**Figure 1.** Scatter matrix of variables in the data set

The histogram for price shows that cars at approximately $5250 are the most frequent in the data set and the frequency of cars decreases exponentially as the price increases. The engine size histogram shows that the most frequent engine size is approximately 90 with the frequency decreasing as the engine size increases. At the engine size of 190 however, there is a small increase in frequency. The horsepower histogram shows the highest frequency at approximately 70 with the second most frequent at approximately 110. The highway mpg histogram shows that the most frequent value is approximately 30.

The relationship between price and engine size has a strongly positive correlation meaning that the price of the car is directly related to the size of the engine. This is also true for the relationship between price and horsepower. The relationship between price and highway mpg is strongly negative meaning that a more expensive car is generally less fuel efficient. The relationship between engine size and horsepower is positive showing that a larger engine size relates to a greater power output. The relationship between highway mpg and engine size is a negative correlation. This is also true for horsepower. This shows that cars with larger engines and a greater power output are generally less fuel efficient.

Chart, bar chart

Description automatically generated

**Figure 2.** Bar charts for car make versus mean price and horsepower

Chart, bar chart

Description automatically generated

**Figure 3.** Bar charts for car make versus mean engine size and highway MPG losses

Figure 2 shows the bar charts for car types versus mean price and horsepower. The figure shows that Jaguar, Mercedes-Benz, and Porsche are the most expensive cars, while Chevrolet, Plymouth, and Subaru are the least expensive. Jaguar, Porsche, and Volvo have the greatest power output, while Chevrolet, Plymouth, and Volkswagen have the lowest.

Figure 3 shows the bar charts for car types versus mean engine size and highway mpg. Jaguar, Mercedes-Benz, and Volvo have to largest size engines and Chevrolet, Plymouth, and Honda have the smallest. Chevrolet, Volkswagen, and Plymouth have highest fuel efficiency. Jaguar, Mercedes-Benz, and Audi have the lowest fuel efficiency.

Chart, bar chart

Description automatically generated

**Figure 4.** Bar charts for car make versus mean engine size and normalised losses

Figure 4 shows the bar charts for car types versus mean curb weight and normalised losses. Jaguar, Mercedes-Benz, and Peugeot are the heaviest cars, while Chevrolet, Plymouth, and Volkswagen are the lightest. BMW, Porsche, and Peugeot have the greatest normalised loss while Volvo, Subaru, and Chevrolet show the lowest.

These plots revealed that expensive cars such as Jaguar, Mercedes-Benz, and Porsche are heavier and have larger engines which generate more power and torque but are less fuel efficient and have greater normalised losses. On the other hand, cheaper cars tend to have smaller engines and are generally more fuel efficient.

Chart, bar chart

Description automatically generated

**Figure 5.** Clustered bar charts for car price ranges versus body type and number of doors

In Figure 5, clustered bar charts are presented to show the distribution of body types and number of doors for different price ranges of cars in the "automobile.txt" dataset. The data indicates that, in the price range of $5000 to $10,000, sedans and hatchbacks are the most common body types, with a roughly equal number of both, and a smaller number of wagons and hardtops. Additionally, there is a similar proportion of 2-door and 4-door cars in this price range. As the price range increases, the relative proportion of hatchbacks decreases, and sedans become the most prevalent body type. Furthermore, 4-door cars become the most common type in the higher price ranges.

Chart, bar chart

Description automatically generated

**Figure 6.** Clustered bar charts for car price ranges versus symbolling and normalised losses

**Chart, bar chart

Description automatically generated**

**Figure 7.** Clustered bar charts for car price ranges versus horsepower and peak rpm

Figure 6 shows clustered bar charts for car price ranges for symbolling and normalised losses. The cars in the 5000-10000 price range have a symboling value of 1. As the cars get more expensive, the symboling value decreases. A lower number means that there is a higher risk associated with the car. A more expensive car therefore has a higher risk associated with it. In a similar manner, the normalised loss increases with car price meaning that there is a greater insurance risk of the car. The data suggests however that at a certain cost point, the normalised loss decreases. This cannot be confirmed as, due to the low number of more expensive cars, the data at this price range is not as accurate.

Figure 7 shows clustered bar charts for car price ranges for horsepower and peak rpm. The data suggests that a more expensive car has greater horsepower and peak-rpm. However as previously stated, this cannot be determined accurately as there is a lower number of more expensive cars in the dataset.

In Figure 6, clustered bar charts are presented to show the distribution of symboling and normalized losses for different price ranges of cars. The data suggests that cars in the price range of $5000 to $10,000 have a symboling value of 1, indicating a lower risk. As the cars get more expensive, the symboling value decreases, indicating that a higher risk associated with more expensive cars. Additionally, the normalized loss increases with car price, indicating that there is a greater insurance risk for more expensive cars. However, at a certain cost point, this observation is not significant due to the low number of more expensive cars in the dataset, which may lead to an inaccurate representation of the data.

In Figure 7, clustered bar charts are presented to show the distribution of horsepower and peak-rpm for different price ranges of cars. The data suggests that more expensive cars have greater horsepower and peak-rpm. However as previously stated the normalized loss decreases, however, as previously stated this observation is not significant due to the low number of more expensive cars in the dataset.

THIS REPORT WAS WRITTEN BY: Joshua Houlden

