

**TASK**

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

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

**Introduction**

The automobile industry is a significant sector in the global economy, characterized by a wide variety of vehicle models with diverse features and price ranges. To understand this market better, we analyze a dataset containing various attributes related to cars. The dataset includes information such as symbolling, normalized losses, make, aspiration, number of doors, body style, drive wheels, engine location, wheelbase, length, width, height, curb weight, price, horsepower, engine size, fuel type, and more, providing a comprehensive overview of the features that influence car prices and performance.

**DATA CLEANING**

The dataset was cleaned to better understand and analyze the information. Cleaning of data is important to find patterns that will provide insight into the dataset. Use the describefunctionto summarize the central tendency, dispersion, and shape of the dataset’s distribution. This led to the need to convert the price column to an appropriate type.

**MISSING DATA**

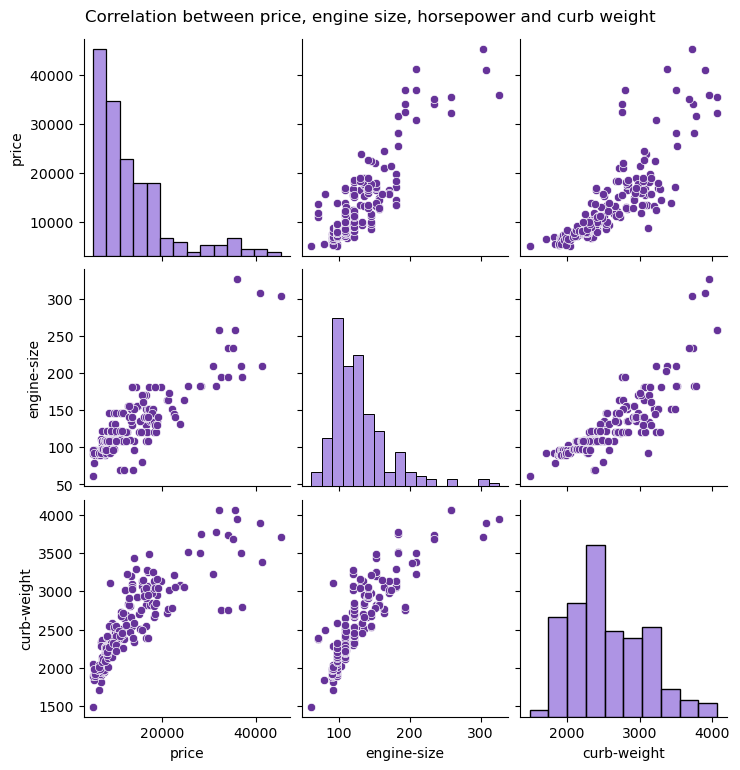
There is missing data in the form of question marks (?), using the isnull function to identify the number of missing data. However, this makes the missing data undetected and hard to remove or fill in. This means that the fill-in method using the fillna function would not work. The best solution is to convert the question marks into a non-entry, this makes the missing data detectable therefore can filled. The missing data was found in the following columns: normalized losses, number of doors, bore, stroke, horsepower, peak rpm, and price. The missing data was filled in with the mean amount of the column values.

**DATA STORIES AND VISUALISATIONS**

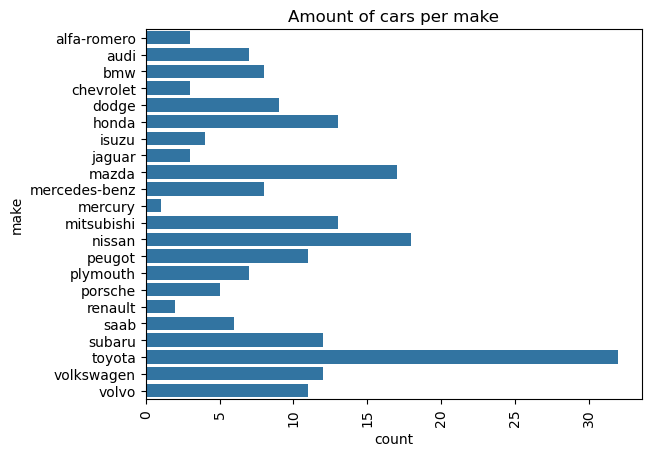
The graph/visualization aims to display patterns within the dataset. The graphs below showcase various information found while analyzing the dataset and draw probable assumptions and findings. There are seven (7) graphs in total and each displays different findings.

The first graph is a pair-plot graph that displays the correction between the price, engine size, and curb weight. It combines both the histogram and scatter plots, providing a unique overview of the dataset distribution and patterns. The histograms show the distribution of each of the above-mentioned variables and provide an understanding of the correlation between each variable. In the case of the histogram, the engine size and curb weight display a health increase and then a steady decline. The price histogram starts at the highest point and then declines.

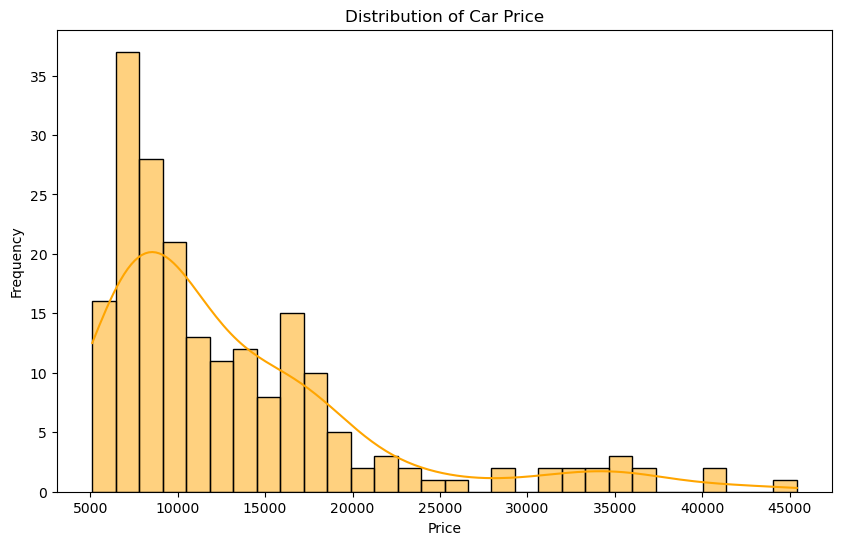
The scatterplot displays the off-diagonal plots showing the relationship between two variables. These can reveal patterns, trends, and correlations, such as linear patterns in all the price, engine size, and curb weight scatterplots. This linear trend in the majority of the patterns suggests predictability in the dataset. There are patterns of clusters in the left bottom part of each scatterplot that point out shared similarities in characteristics within the dataset.



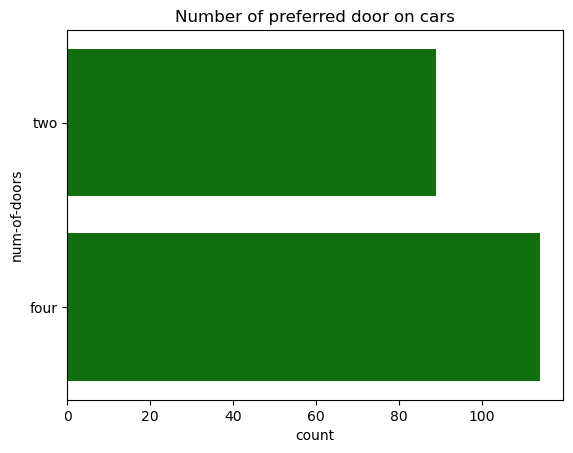
The second graph displays the number of cars per make in the dataset. According to the visualization Toyota has the most cars in the dataset, followed by Nissan and Mazda. Those with the least cars are Mercury, Renault followed by Alfa-Romero, Chevrolet, and Jaguar in third place. The car makes show a pattern in not only the car ‘make’ in the dataset but also the preferred car ‘makes’ but consumers.



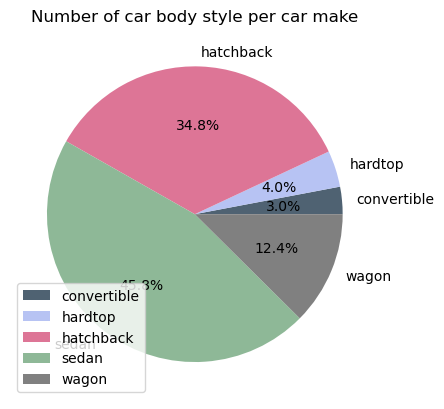
The third graph below shows the distribution of prices across the automobile industry. Car place is another factor that plays a role in deciding which car to purchase. The graph displays that the most frequent car price is between 5000 and 10000. This finding is not surprising as the economy and affordability of the consumers influence their spending habits and therefore which cars are likely to be purchased.



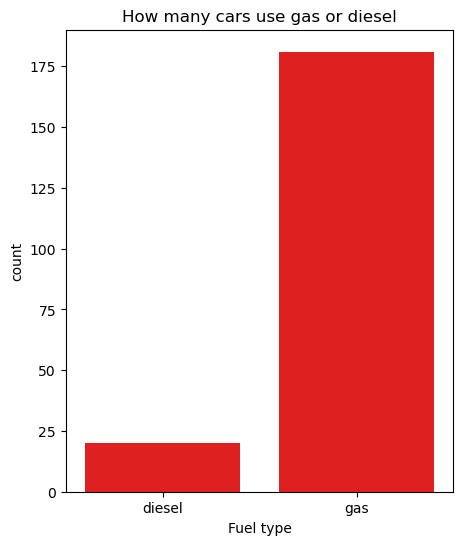
The next graph displays another factor that is considered when choosing to buy a car. The number of doors dictates how the can will be used and how many people will be able to gain entry into the car. Four doors will also be preferred over two as it means the car size is big enough, and the driver and front passenger do not have to move themselves and their seats for them or others to access the back seats.



The pie chart below displays the proposition of car body style that each of the cars in the dataset has. It provides a great view of which car body style is preferred (sedan) with the convertible being the least preferred. This information can be helpful to manufacturers and those in the automobile industry as they continue to design new and/or upgrade those we have. This information provides them with a good perspective into what consumers prefer in their cars.

****

When deciding the kind of car to purchase, gas or diesel is one of the common attributes to consider. As in many circumstances, the price of gas and diesel are never the same, actually in most cases, diesel is more expensive. However, the determination of whether a car will use gas or diesel is the engine size. From the dataset, the number of cars that use gas is 181 and those that use diesel is 20.

****

**THIS REPORT WAS WRITTEN BY: VANESSA TOOKA**

