

**TASK**

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

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

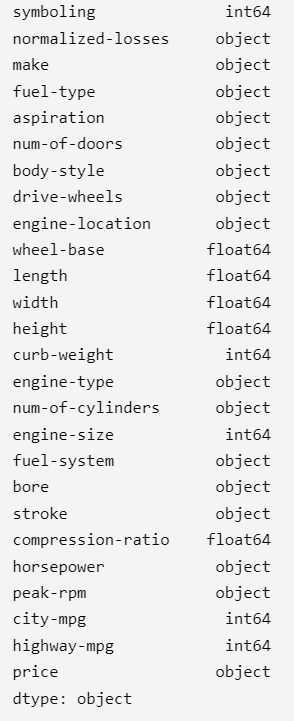
**Introduction**

This report features a dataset on automobiles, consisting of 205 entries with various details on different vehicle types. The information contained in each column includes the car make, fuel type, engine size, and price.

**DATA CLEANING**

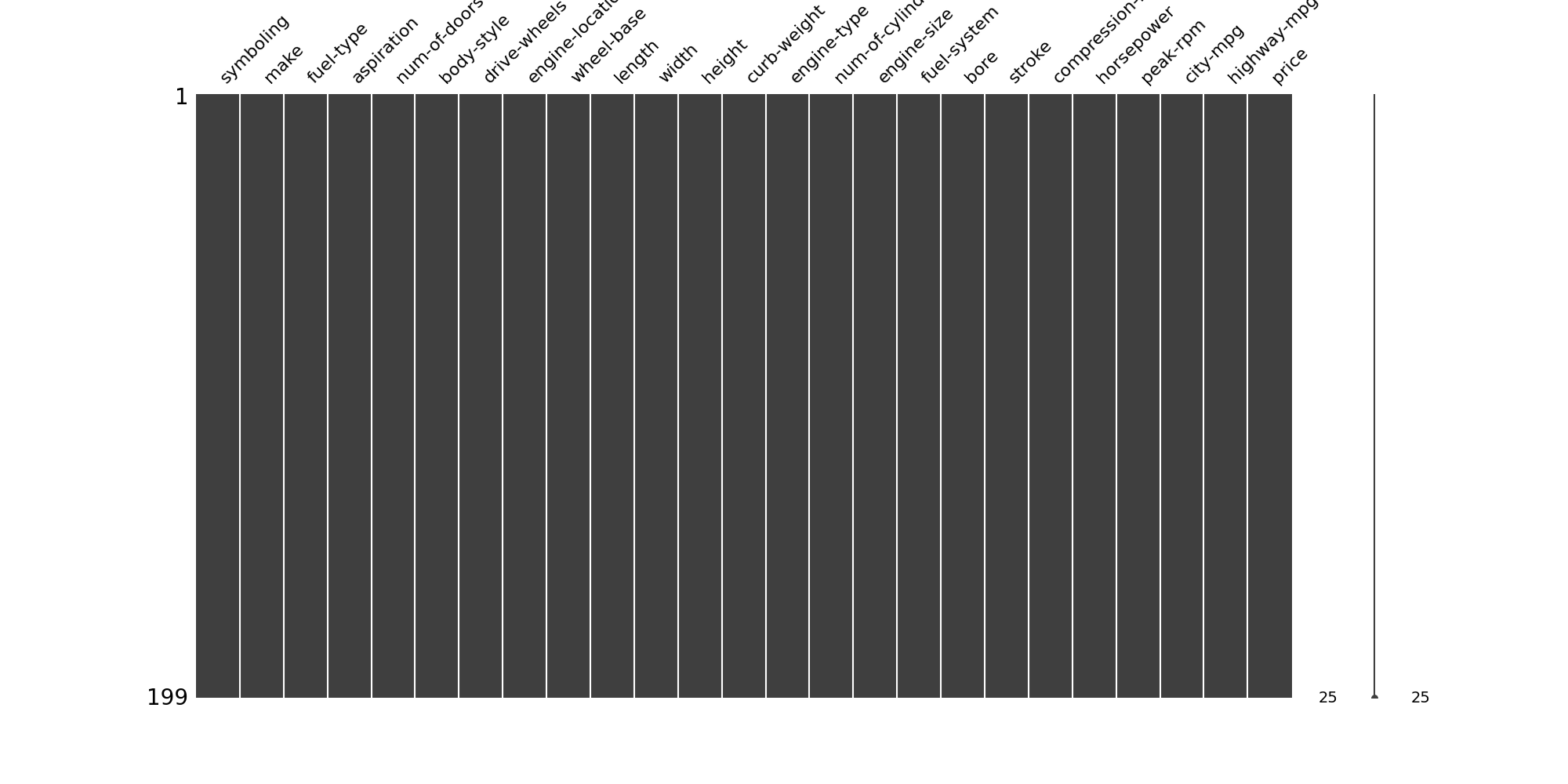
# SUMMARY OF THE METHODS AND VISUALISATIONS DONE DURING DATA CLEANING

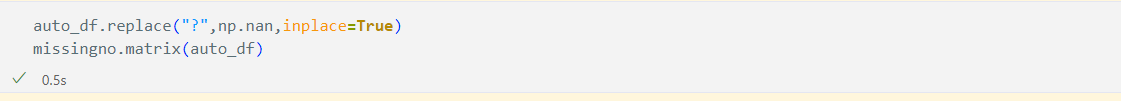
To clean the data, I must first identify the data type in each column by using the .dtypes() method. This will display all columns and their respective data types. With this information, I can make any necessary changes and clean the data accordingly.

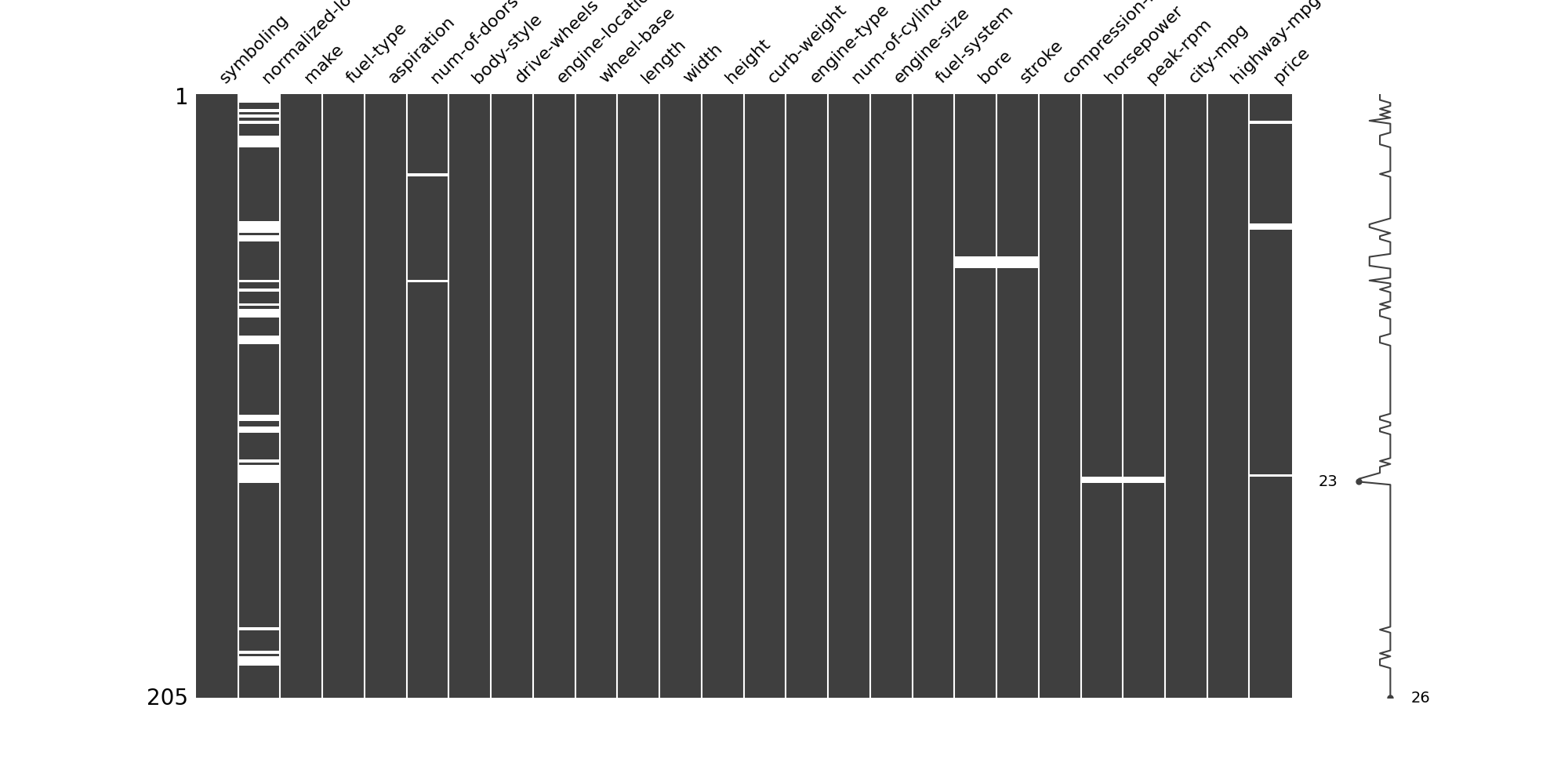


**MISSING DATA**

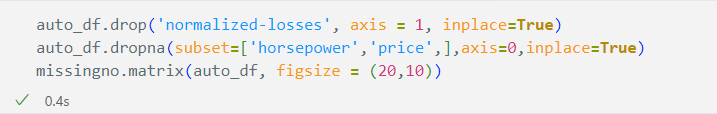
I utilized the "missingno" package to visualize missing data in the rows and columns through a graph. Although the graph indicates the absence of missing data, the data frame shows question marks in certain boxes, indicating either missing or unknown data. To maximize the use of the "missingno" package, I will replace these question marks with NaN values so that I can accurately determine if there is any missing data.





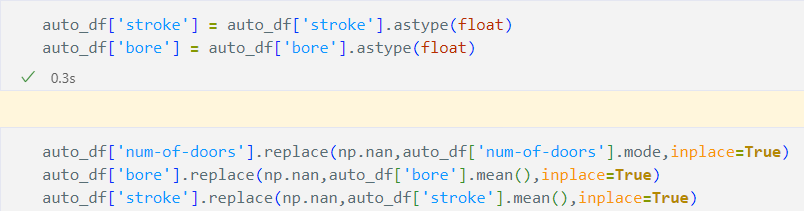


The visualization reveals a significant amount of missing data in the "normalized-losses" column. Since a considerable portion of the data is missing, this column will be removed as it does not provide any valuable information for the analysis. On the other hand, the "bore", "stroke", "horsepower", and "price" columns only have a few missing values. Since these values are not easily predictable, the individual rows containing missing data will be removed.

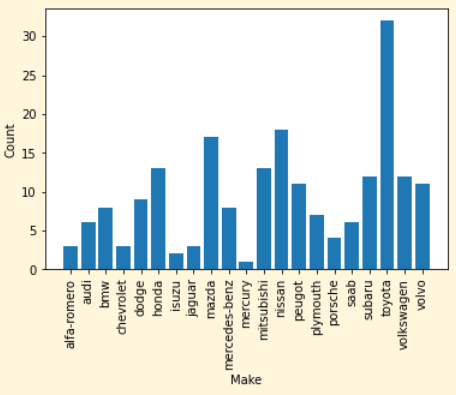


As indicated by the .dtype() query, the "bore" and "stroke" columns are of object type, but in order to calculate their average, I need to change their type to float. Then I can replace the empty values in “ number of doors” , “bore” and “stroke” with the average of each column and convert data types to integers

After addressing the missing entries, I can now visualize the data.

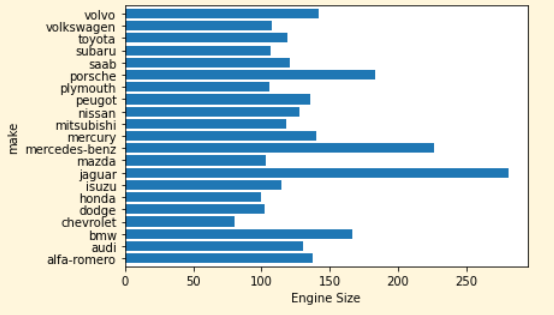


**DATA STORIES AND VISUALISATIONS**

I have generated several visualizations using the cleansed data, starting with a bar plot that shows the count of automobiles by make.

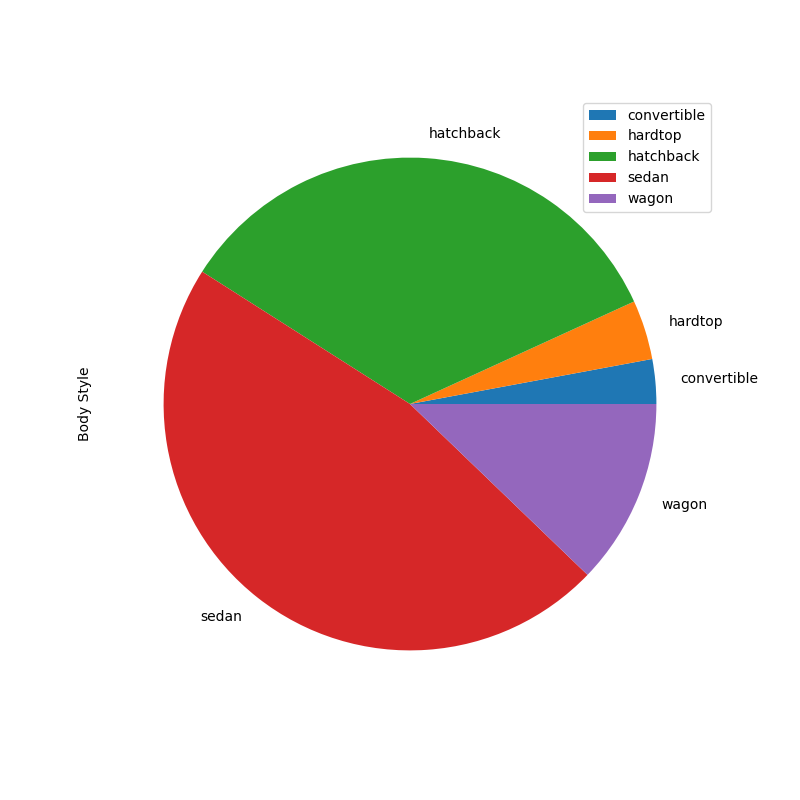
This bar plot allows us to conclude that Toyota is the most prevalent brand, with over a third more automobiles than the next leading brand, Nissan. Mazda, another Japanese company, also has a significant number of automobiles, trailing closely behind Nissan. These three brands are all produced by Japanese manufacturers.

Next I plotted a horizontal bar chart for Average Engine Size Per Make.



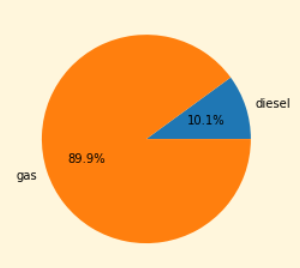
This chart presents a clear picture of the automobile industry. Although Toyota, Nissan, and Mazda are the leading brands in terms of quantity, they fall significantly behind in terms of engine size when compared to the top brands. Jaguar, Mercedes-Benz, and Porsche occupy the top spots in terms of average engine size, with Jaguar leading the pack. It is interesting to note that despite the smaller engine sizes of the top three brands, they still manage to dominate the market. This could be due to a variety of factors such as cost, fuel efficiency, and reputation. Nevertheless, this chart provides valuable insight into the automobile industry and highlights the various strengths and weaknesses of different brands.

Next I created a pie chart to provide a clear representation of the distribution of body type preferences among car buyers.



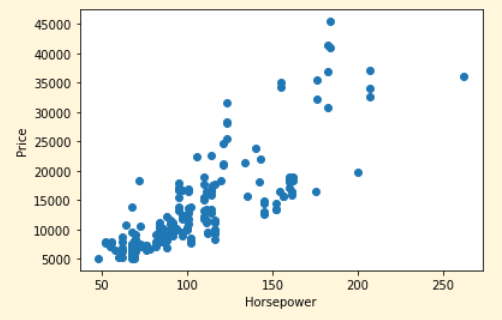
This pie chart provides a clear representation of the distribution of body type preferences among car buyers. The data displayed in the chart indicates that sedans are the most in-demand body type, accounting for a significant portion of the market. On the other hand, convertibles are the least favored, with only a small fraction of car buyers opting for this body type. This result is not surprising, considering factors such as weather conditions, practicality, and cost. In regions with unpredictable weather, a convertible may not be the most practical option, whereas sedans are seen as a more versatile and cost-effective choice. This chart provides valuable insight into consumer preferences and can assist manufacturers and dealers in making informed decisions about their product offerings

I created another pie chart to compare the number of cars that run on petrol versus diesel fuel. The chart indicates that 89.9% of the vehicles in the dataset run on petrol, while the remaining 10.1% run on diesel.

****

This pie chart provides useful insights into the fuel preferences of car buyers. The high percentage of petrol cars suggests that the majority of consumers still prefer vehicles powered by gasoline over diesel engines. This could be due to a variety of reasons such as the lower cost of petrol, its wide availability, or simply personal preference. Understanding this information can be useful for car manufacturers and marketers as they can tailor their offerings to meet the demands of their customers. Furthermore, the government can also use this information to make decisions regarding fuel subsidies or incentives that may affect the production and consumption of gasoline and diesel vehicles.

Lastly, I created a scatter plot to visualize the relationship between price and horsepower.

****

The scatter plot that I created to show the relationship between the price and horsepower of the cars in the dataset is quite telling. As can be seen from the plot, the majority of the cars are between 70 and 120 horsepower, with a price range of around 5000 to 22000. This relationship is fairly linear, meaning that there is a strong correlation between horsepower and price. However, after the price point of 22000, the horsepower increases exponentially, which is interesting to see. This could suggest that for cars that are priced higher than 22000, there is a higher demand for horsepower, which is something that car manufacturers take into consideration when pricing their vehicles. Overall, the scatter plot provides a good visual representation of the relationship between price and horsepower, and can be useful in helping to understand the market trends and consumer preferences for different types of vehicles.

**THIS REPORT WAS WRITTEN BY : Samuel Blackhurst**

