

**TASK**

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

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

**Introduction**

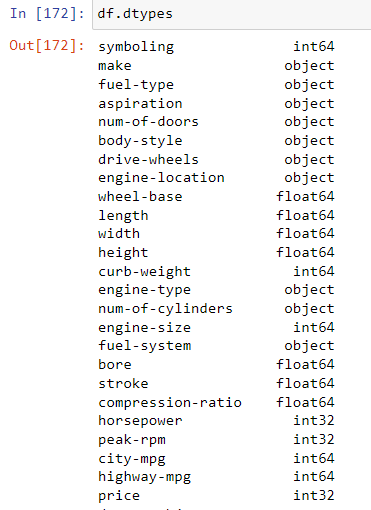
The data set used for this report is the automobile data set, which contains a total of 205 rows of data across 26 columns.

The columns specify general information about each car, such as make, fuel type, price and car dimensions, and also performance information such as horsepower, fuel system and engine size.

**DATA CLEANING**

# SUMMARY OF THE METHODS AND VISUALISATIONS DONE DURING DATA CLEANING

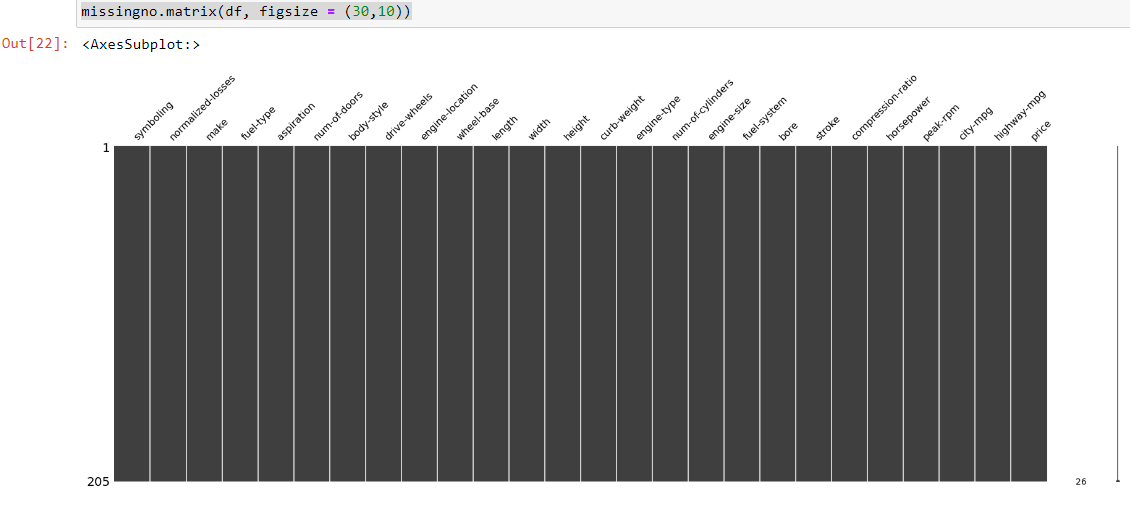
First of all, the data type of each column has to be checked to correspond with its data, this can be easily achieved by using the “.dtypes” method on the data frame.



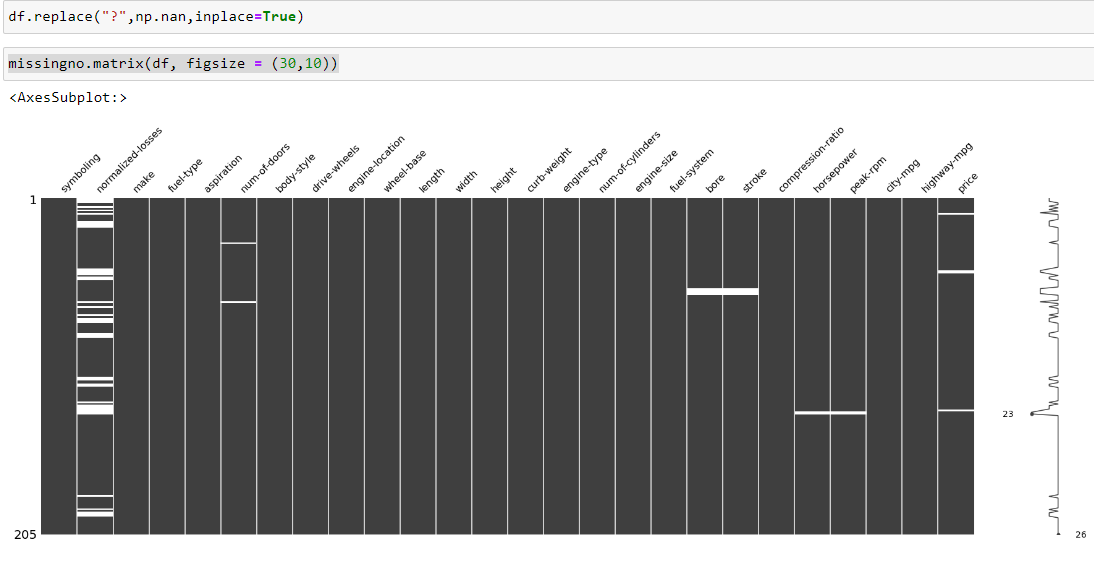
The result will show us each column and its data type. We can then proceed in changing the data type of the columns we see fit to the corresponding data type using the “.astype()” method.

**MISSING DATA**

By using the “missingno” imported package we can create a graph of the missing data across the data frame.



Since the data set is of small proportions and this method showed us that there are no missing data in our data frame, it would be recommended to do a quick “manual” check to make sure we don’t have any missing data. Values such as “?” can be found in the database which represents missing values, those values can be changed to NaN values to use the “missingno” package effectively.



Now it is easy to see which columns have missing data.

* The normalized-losses column has 41 missing values out of 205 total, we can drop the entire column since 20% of the data is missing.
* The price, horsepower and peak-rpm columns have some missing values as well, but we can remove just the rows with the missing data since we can't predict or replace the missing values.



* 2 values are missing for num-of-doors, we can replace those with the more frequent since they are categorical data. This is done with the help of the “.mode” method which returns the most frequent value of the column.

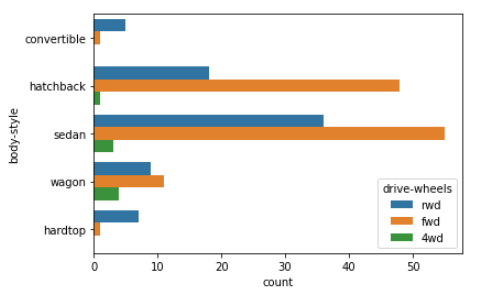


* 8 values are missing from the bore and stroke columns, those can be replaced with the average of the column.

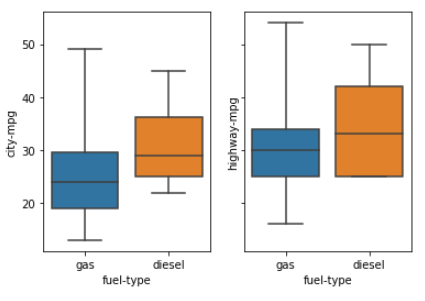


**DATA STORIES AND VISUALISATIONS**

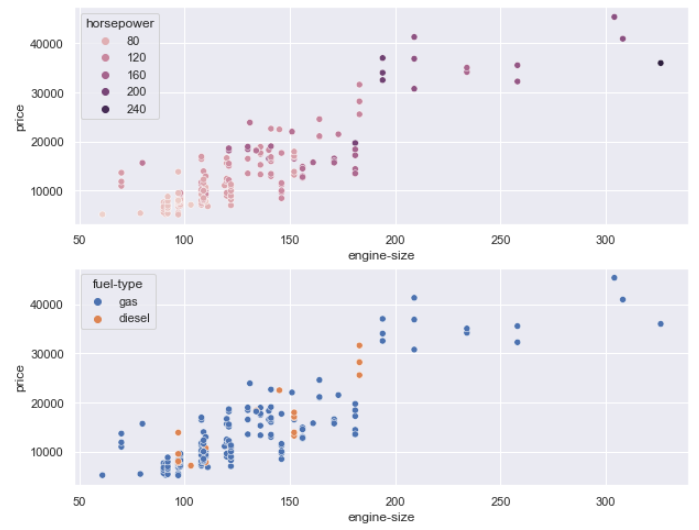
Once the data is cleaned and there are no missing data we can proceed with various plots to obtain insights from our data.



From the plot, we can observe that the most popular are sedan cars, and the front wheels drive is predominantly across most body styles. On the other hand, convertibles and hardtops have predominantly rear-wheel drive.



The above plots show the fuel economy difference between gas and diesel cars, we can see that for highways, a diesel car would be more economical since the lower threshold is higher than the gas cars.



This plot shows the correlation between the price and the engine size.

Also, it can be observed which type of fuel, and how much horsepower the car has.

**THIS REPORT WAS WRITTEN BY: YOUR NAME**

