

**TASK**

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

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

**Introduction**

This dataset contains information about automobiles and will be used to help customers decide which cake make will best suit their needs.

Before beginning the analysis of the dataset, it must be cleaned in order to make is more understandable. The first step in this process is to check for redundant columns which have no relevance to the dataset.

As data is being used to help customers decide the best car make for their use, columns with categorical values have been dropped as they are hard to use to use as a comparison. By doing this we have reduced the number of columns from 26 to 15.

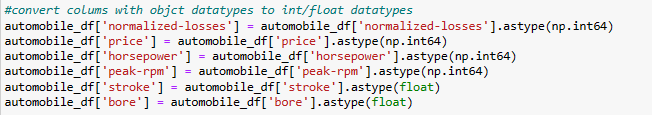


*Drop columns which contain categorical values*

In order to calculate the mean value for each variable per make, we need to ensure the data type is int or float, currently there are columns which contain ‘?’ and therefore are object types. In order to convert the data, we need to remove the missing the data.

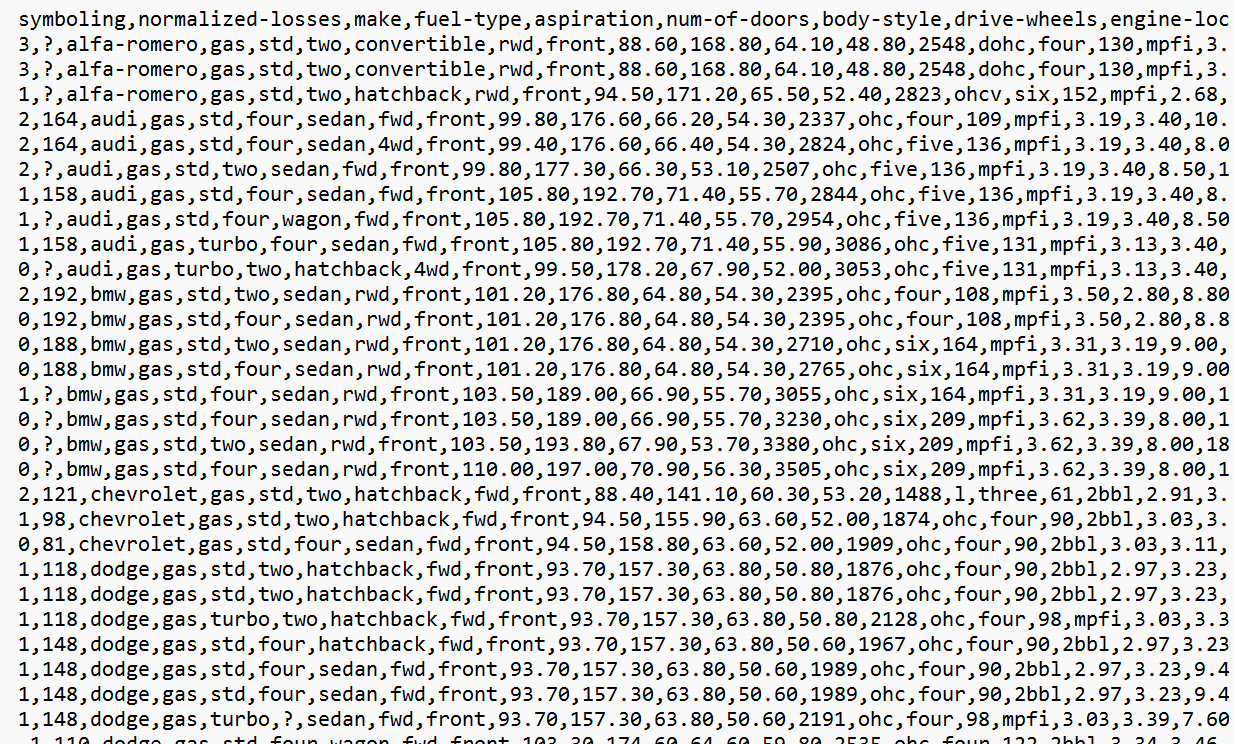


*Dropping the data with missing values*



*Converting columns to float and int datatypes to allow the calculation of mean values*

The missing data shouldn’t be replaced with the mean value, as the mean would be calculated using the values of other car makes, reducing the accuracy of the data for that car make.



*Data before it has been imported*



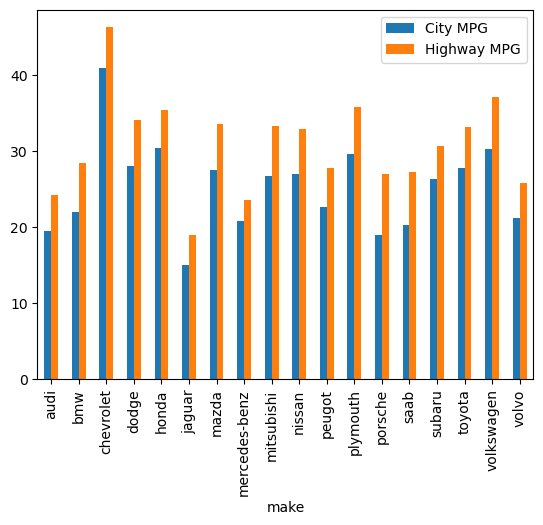
*Calculate the mean values using the make as an index*

After the mean has been calculated, the data is to 6 decimal places, this increases the difficulty of visual comparison when looking at the dataset, so it has been reduced to 2 decimal places as it was before the mean was calculated.

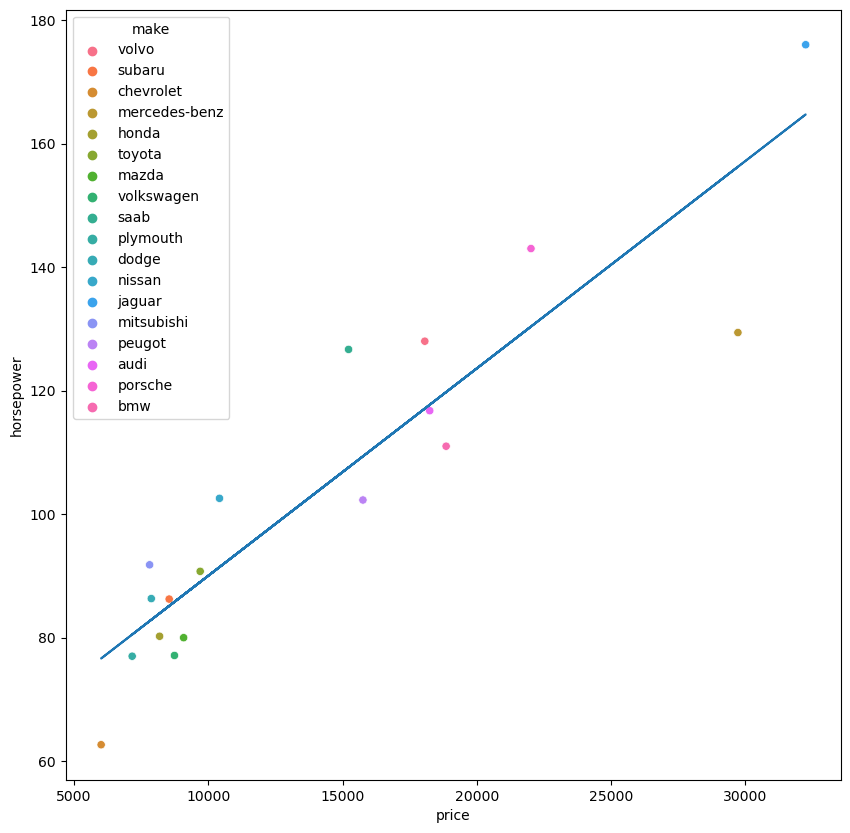


*Rounding the data to 2 decimal places*

**Data stories and visualisations**



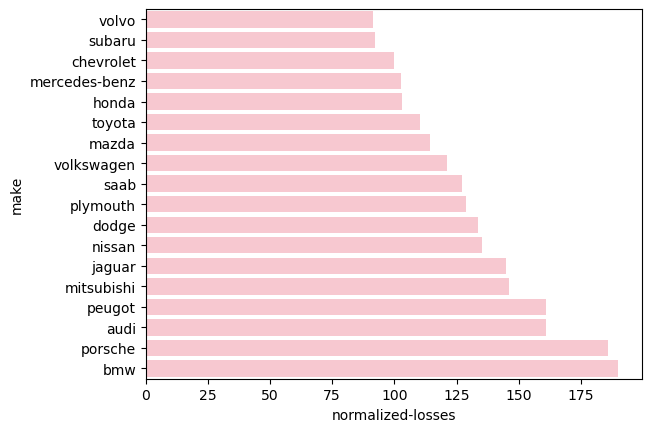
This visualisation compares the miles per gallon in the city and on the highway for each make. It shows that the make with the highest miles per gallon in both the city and on the highway is Chevrolet. If the customer is looking for a car which is the most fuel efficient, that is the make they should choose.



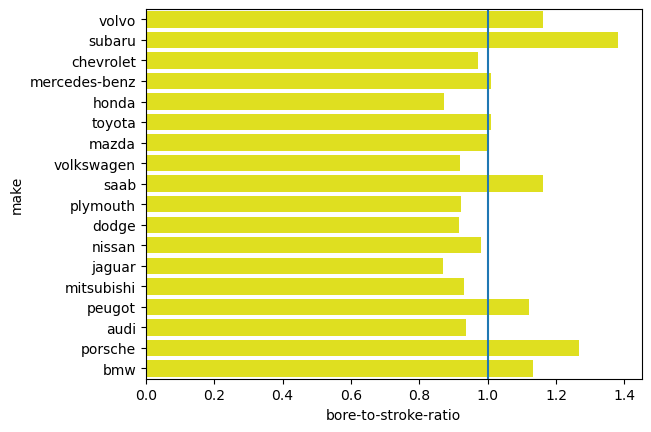
This visualisation looks to see if horsepower increases with price, the line of best fit shows there is a positive correlation between the variables.

This visualisation also shows that cheapest car make on average is Chevrolet and the most expensive is jaguar.

If the customer is looking for a car make around the middle, with a good price and horsepower, the best makes to go for are Volvo, Audi and BMW.



When deciding what car to buy, it is important to consider the cost of the insurance vs the value of the car. This visualisation looks at normalized losses, the higher the losses the lower the value of the car vs the insurance cost. When choosing, the customer should select a make with the lowest losses to ensure good value for money, this barplot indicates the Volvo would be the best make and BMW is the worst.



Bore to stroke ratio was calculated by dividing the bore by stroke of each make. This ratio determines the engines power and torque, if the value is greater than 1, this means the car I make is designed for higher speeds. If this is what the customer wants the best makes would be Subaru, BMW, Porsche, Peugot, Saab and Volvo.

If the ratio is less than 1, the car make has a greater pulling ability and therefore higher torque. This means the car has a greater accelerating ability, these cars tend to be bigger and can pull more weight. If this is what the customer is looking for, the best make for them to purchase is Audi, Jaguar, Chevrolet, Honda, Volkswagen, Plymouth, Dodge and Mitsubishi.

If the ratio is 1, this means the make produces cars with the perfect balance of the 2, if this is what the customer is looking for the best make would be Toyota, Mazda or Mercedes-Benz.

**THIS REPORT WAS WRITTEN BY : Sara Dimtsu**

