

**TASK**

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

[](http://www.hyperiondev.com/portal/)

**Introduction**

The data covers vehicles by different automakers.

In the data frame features include many characteristics of each vehicle such as make, style and horsepower. The purpose of this EDA is determining what features have an impact on price. This way the data can add value for someone looking to sell their vehicle or purchase a new one.

**DATA CLEANING**

The data was cleaned and prepared ensuring that the data types of numeric features were all set to int or float types. The majority of data cleaning was focused on imputing missing values.

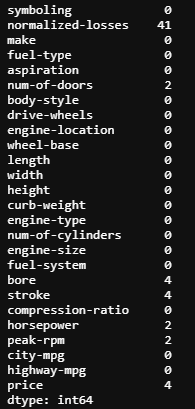
MISSING DATA

The data contained missing values in the form of “?”. By using the missingno package in Python to detect missing values, non were detected as “?” is detected as a value.

The first step was to create a function which could be used to loop through all columns and replace “?” with a numpy NaN value.

Once this was complete the missing values could be visualised using missingno.

The data contained the below missing values:



As the body style of a vehicle can affect the number of doors the mode by body-style was used to impute missing values. Normalized losses often depend on the size and type of a vehicle, here the average by body-style was used to fill missing data, this would give a more accurate approach than taking an overall average.

The missing data in horsepower and peak-rpm were filled using the average by engine type while engine size was used for bore and stroke.

DATA STORIES AND VISUALIZATIONS

Through using categorical boxplots, it was determined that hardtops have the most variance in price as well as the highest price on average, this is followed by convertibles, wagons, sedans and finally hatchbacks which have the lowest average price.

RWD vehicles tend to be the most expensive drive wheel type.

Turbo vehicles tend to cost more than those which are naturally aspirated and diesel vehicles are more expensive on average than gas, while gas data did contain more outliers.

The five most expensive car makes on average are Jaguar, Mercedes, Porsche, BMW and Volvo. The five cheapest are Subaru, Honda, Plymouth, Dodge and Chevrolet.

When it comes to horsepower per dollar the vehicle makes that give consumers the most power for their dollar are Mitsubishi, Dodge, Plymouth, Mercury, and Chevrolet. This was determined through horizontal bar charts and by creating a variable that was the result of horsepower/price.

Both engine size and horsepower are positive correlated with price. This was determined through the use of Seaborn regression plots. This makes sense as more, powerful, large motors tend to be in bigger, more expensive vehicles.

By using a heatmap on a correlation matrix it could be determined that length, width, curb-weight, engine-size and horsepower were positively correlated with price.

While fuel efficiency had a strong, negative correlation with price when looking at mpg for town and highway driving.

Using visualisations, it was determined that more expensive vehicles tend to be more powerful and bigger in terms of width and height. Convertibles and Hardtops come with a premium price attached. The most expensive vehicle makes also tended to have the worst power per dollar ratios, this is likely due to the price reflecting additional safety and technology features in the vehicle.

Diesel vehicles also tend to be more expensive on average and turbos come at an increased price compared to naturally aspirated vehicles.

This EDA explains what features have an impact on a vehicles price and should be considered when buying or selling a vehicle.

**THIS REPORT WAS WRITTEN BY : STEVEN O’EHLEY**

