

**TASK**

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

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

**Introduction**

The automobile dataset provides information about different features about different cars based on several features. These features can be analysed to provide insight on different aspects of cars and how they relate to each.

**DATA CLEANING**

The first step in the data cleaning process was to check whether any of the columns could be dropped for easier analysis. I chose not to drop any of the columns as each column would be used in the analysis.

I then went on to check the data type for each column. This allowed me to see which variables would need to be converted accordingly. This was followed by checking if any null values occurred in the dataset. I found that there were no null values, as the missing values in the dataset were instead represented by a question mark(?).

I then went on to check how many of these irregularities occurred in each column. Irregularities like this was found in the following columns: normalized-losses, price, horsepower, bore, stroke, peak-rpm and number of doors.

For horsepower, two outliers were found. These outliers were excluded from the dataset as they would have skewed the data.

MISSING DATA

In order to deal with these irregularities, I first checked the datatype for each column. These columns were entered as ‘objects’ and therefore had to be converted to integer datatypes. Once the values were converted to the appropriate datatypes, I was able to replace the missing data with the mean value (average) for each of the above-mentioned columns.

I opted for this method of dealing with missing data as removing the rows or columns which had irregularities would have removed valuable data which was crucial to the data analysis. Replacing the missing value with the value which occurred either before or after the value would not have been an appropriate approach, as the nature of the dataset varies drastically depending on the row, and this would have led to inaccurate data being inserted.

DATA STORIES AND VISUALIZATIONS

The first analysis done was a “Vehicle make frequency diagram”. This shows the number (frequency) of vehicles manufactured by each “make” (brand). Using a bar graph to display the data, we found that Toyota had the highest number of cars made, with just over 30. This is about 40% more cars than the second highest producer Nissan. BMW makes the least number of cars with less than 10 cars.

For the second visualization, I used a histogram to display the “Insurance risk rating”. This indicated that most cars have a risk rating of about 0. Note, the scale of risk rating(symbolling) is from -3 to 3, where -3 is a good rating and 3 being a bad rating.

The “Normalized-losses” histogram shows that the highest normalized-loss occurred in the 65 – 150 range.

The “Fuel type” bar graph indicates that much more vehicles use gas than diesel. This could be an indication of what consumers prefer and may link to fuel economy.

We also found that 82.27% of people prefer cars which are naturally aspirated to cars with a turbo (17.73%).

The “Horse Power” histogram indicated that most cars have a horsepower in the range of 50 – 100.

For “Wheel drive” we found that most cars were forward wheel drives, followed by rear wheel drives, with 4-wheel drives being the least common.

Lastly, the majority of vehicles have a curb-weight of about 2000 – 2500 kg.

A correlation analysis was then done using a heat map, and the following findings were made:

1. Price is largely dictated by the engine size and curb weight of the car.
2. Curb weight correlates strongly to engine size, length, width and wheel base of the car.
3. The wheel base of the also correlates strongly to the length and width of the car.

By using a box-plot, we analysed the relation between “price” and “make” of the car. I found that the most expensive cars are manufactured by Mercedes and the least expensive cars are manufactured by Mitsubishi. Luxury cars such as BMW, Jaguar and Porsche and Mercedes cost more than $20 000. The cheapest cars available are Chevrolet, Dodge, Honda, Mitsubishi, Plymouth and Subaru which all cost below $10 000. The majority of cars fall between the $10 000 - $20 000 range.

I have used a scatter plot to show the correlation between engine size and price. This shows a positive correlation between the variables, meaning that the price of a car increases with the size of the engine.

By plotting the “City MPG” (which indicates the fuel efficiency of a car when being driven in a city) against “Drive Wheels” I found that front wheel drive cars are the most fuel efficient as they have the highest “City MPG”. 4-Wheel drive vehicles have the second highest City MPG; however, it is not much higher than rear-wheel drive vehicles.

The box-plot which displays the price of cars based on the vehicles drive wheels clearly shows that rear-wheel drive cars are the most expensive and front-wheel drive cars are the cheapest. Since we have a low number of 4-wheel drive in our dataset, this visualization might not be an accurate indication of how price relates to 4-wheel drive cars.

When taking the two above-mentioned analyses into consideration, we can conclude that front-wheel drive cars are the most economically-viable cars for customers.

Lastly, when analysing normalized-losses based on body style and number of doors, we find that 2-door cars generally have a higher number of normalized-losses compared to 4-door cars.

**THIS REPORT WAS WRITTEN BY : Curtis van Wyk**

