

**TASK**

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

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

**Introduction**

Summary of the data set

The dataset automobile.csv consists of 26 columns and 205 rows. The 26 columns are made up of categorical variables as well as numerical variables listing either discrete, or continuous and categorical values in the respective columns.

Some of the columns are:

* Body Style
* Price
* Fuel Type
* Make

In each of the 205 rows there are lists of a vehicle in a class, that is of a certain make that holds the characteristic values in the respective columns. The dataset is generally in a good state and because of this visualisations can be produced so that we can make inferences and reveal patterns about the data.

**DATA CLEANING**

* The column headers contained hyphenated words which could have caused some errors in visualizations or other methods further down the line. The hyphens were replaced with underscores to resolve these issues.
* Several datatypes were changed to integer types where the numbers were logged as string or object types. This was done to perform further computation with the data.

MISSING DATA

Although there are no null values or Nan values I the dataset, missing values do however exist in the form of a string ‘?’. A for loop was run through the values of each of the columns and a condition counted the number of occurrences under the respective column for matched results.

These are listed below:

normalized\_losses >>>>> 41

num\_of\_doors >>>>> 2

bore >>>>> 4

stroke >>>>> 4

horsepower >>>>> 2

peak\_rpm >>>>> 2

price >>>>> 4

num\_of\_doors had two missing data points for this reason they were removed leaving us with 203 listings. Normalized\_losses had the largest set of missing data points; these were all replaced with the mean value of the rest of the remaining data points. The rest of the missing data points were treated in the same way or replaced with the value 0.

DATA STORIES AND VISUALIZATIONS

1. Make and Selected Categories Cross-tab C map:  
   This visualization is done with several categorical variables visualized over one. This visualization gives us a detailed breakdown of the different make of vehicles in the dataset with categories selected that clearly sets one vehicle apart from another. For example, when grouped together, gas and diesel, turbo and standard, two and four num\_of\_doors, num\_of\_cylinders, front-wheel rear wheel and four wheel drive, these variables visualized over the different body styles creates clear distinctions between certain vehicles and we can see which make of vehicles have particular cars listed in the dataset. The visualization also provides the marginal and distributional sum of all the vehicles in respected listings.
2. Price Range per Body Style – Box-Plots:  
   For a pricing comparison between the different body\_style types, the boxplots visualized gives us a great view of the pricing spread of the different body style type of cars listed in the dataset.

The biggest range in pricing is found in hardtop vehicles with pricing ranging between about $8000.00 and $45500.00 with median of around $17000.00. The smallest range of price is found in the hatchback body style type vehicle with pricing ranging between around $5000.00 and $22500.00 with a median of around $9000.00

1. Price Range Per Make and Body Style – Box-plots:

Visualization no. 2 further extrapolated by adding the Make categorical variable shows us various pricing ranges with the Mercedes-Benz Convertible being the most expensive and the Chevrolet Hatchback the most affordable.

1. Percentage representation - Pie Chart:

Beautifully represented here in various colours we fnd the percentage of representation of vehicle manufacturer in the dataset. Toyota being the largest contributor with around 15.76% and Nissan the second largest at 8.87%

1. Fuel Type Split - Bar Chart:

This dataset holds two fuel types of engine for the vehicles listed. The majority of vehicle listed in the dataset is of gas type with over 80% of representation.

1. Aspiration Type Split - Pie Chart:

With two types of aspiration namely standard and turbo represented in the dataset, we see via pie chart visualisation that the standard type is preferred with an 82.27% presence in the listings.

1. Normalized Losses Spread – Histogram:

This histogram shows the amount of vehicles in a particular value group of normalized losses in use as compared to other cars. The greatest number of cars in the dataset falls within the value 100-125 and only a small count of cars (less than 10) fall within the value 225-250.

1. Symbolising Risk Factor – Histogram:

With symbolizing number value 3 being the most risky symbol for a vehicle to receive in terms of price, almost 30 vehicles in this dataset falls within this category. The largest portion of vehicles fall within the just below and above value 0 which is far less risky than 3. From here the amount of vehicles per class 1 and 2 tapering of towards value 3. Less than 30 vehicles fall within the value range -1 to -2.

1. Drive Wheels - Bar Chart:

A simple Bar Chart in the Drive Wheels category summarizes the amount of cars within each group with front-wheel drive cars being the most represented (+-120) in the dataset and four wheel drive the least (less than 10).

1. Number of Doors Frequency - Bar Chart:

With only two categories in the 'num\_of\_doors' variable, four door vehicles are the most represented in this dataset.

1. Make, Fuel Type, Engine Size, Aspiration Group - Horizontal Bar Chart:

Tis visualization may look busy but it can serve as a matrix in which the average economy of all the vehicles per engine size in the dataset can be found. Typically one would compare similar engine sizes, aspiration types and fuel types grouped, with one another (by make). The most economical vehicle in the city and highway in the dataset is the Chevrolet, gas type, standard aspirated 61 sized engine vehicle logging just over 45 miles per gallon in the city and an amazing 50+ miles per gallon on the highway.

1. Horsepower, Bore, Stroke Per Curb Weight - Line Diagrams:

With the three line diagrams of horsepower, bore, and stroke visualized over the curb weight we can clearly see how design adjusts bore size and stroke as the curb weight of the vehicle increases to deliver more horsepower. There are exceptions to this though and it is in sport models where horsepower is increased and not necessarily curb weight, this is to deliver better performance for sporty models. Please note that here we can see the drastic dip towards zero in our bore and stroke lines. This is due to the replacing of the missing values with null.

**THIS REPORT WAS WRITTEN BY : Kamerin Moodley**

