

**TASK**

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

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

**Introduction**

The automobile dataset is a multivariate dataset from the 1985 Ward’s Automotive Yearbook. The dataset comprises of 25 input feature and 1 output feature. The features are comprised of categorical and numerical types.

The automobile dataset consists of 205 records of automobiles, where each row has entries for each feature, with datatypes the datatype of that specific feature.

**DATA CLEANING**

The data cleaning process comprised of identifying columns that are redundant or unnecessary namely: symbolling, normalised-losses, original\_language and engine-location.

The last step in the data cleaning involved the identification and removal of the duplicated movie entries (rows) in the dataset and assigning the correct datatype for each column/feature in the dataset.

**MISSING DATA**

The missing values in the dataset were represented by a question mark (**?**), which were converted to **NaN**. Furthermore, the use of a heatmap provides a visual representation of the missing data.

The number of doors (**num-of-doors**) categorical feature had two missing values, which were replace by the statistical mode of the number of doors column.

The missing values in the numerical features (**bore**, **stroke**, **horsepower** and **peak-rpm**) were replaced by the statistical median of the respective columns. The use of the median assists with avoidance of the inclusion of outliers.

Only, the rows with missing data in the output **price** feature were deleted.

**DATA STORIES AND VISUALISATIONS**

A look into the relationship between a select categorical features and numerical features is investigated.

The initial focus is on engine-types and peak-rpm. The insight gain is that all the engine types produce roughly the same maximum peak-rpm, and the boxplot od the relationship illustrates a certain overlap in box-whisker distributions of the different engine types.

The relationship between **body-style** and **curb-weight**, provides a level of variability among the mean/median curb-weights. This indicates that body-style may be a good categorical variable for prediction of price, based on the strong correlation between curb-weight and price.

Lastly the categorical feature for carmaker (make) and price reveals some interesting insights. The car makers chevrolet, mercury and renault don’t offer much insights in terms of variability. A noted difference is noted in the distribution of entry level, mid-level and high-end maker is notable. Certainly, there would be no price without the car makers. Thus, the make feature cannot be ignored in the prediction of car prices.

A correlation heatmap and diagonal Pair-Grid of the numerical features is implemented to gain insights.

The strong negative correlations between price and highway-mpg, whereby a higher efficiency results in lower price and low efficiency results in higher prices as indicated in the Pair-Grid scatterplot.

Furthermore, it is more fuel efficient to drive in the city as compared to driving on the highway. This evidenced by the distribution through a histogram of city-mpg and highway-mpg.

Other numerical features with strong positive correlations with price are length, city-mpg, curb-weight, width, and horsepower. The scatterplot indicates these features have a potential strong linear relationship with the price feature. This bodes well for the implementation of an automobile price model.

# ENSURE THIS DOCUMENT IS NEAT AND ADD IT TO YOUR PORTFOLIO

**THIS REPORT WAS WRITTEN BY: ZUKISA XOTYENI**

