

**TASK**

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

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

**Introduction**

The dataset presented covers a range of automobiles and related information to each entry. The set covers 22 different brands of car as well as 5 different styles of car within each brand. Through an exploratory data analysis, the study attempted to find emerging trends within the dataset with a focus on price paid for the vehicle and what exactly that amounted to in terms of car specifications.

**DATA CLEANING**

As the incomplete values for 'normalized losses' represent around 23% of the column, the column was dropped from the set. The remaining missing data consists make up 1% of the fields. However, inputting statistical measures may skew data as the cars are uniquely identified by make as well as body-style. Therefore, these rows were dropped from the set as well. Some of the numerical data is represented as an object and had to be changed to either an integer or float to complete the set.

MISSING DATA

The initial look at the data seems to show a complete set with 'normalized losses' presenting the only issue initially. Before any analysis, missing and incomplete information was identified and dealt with appropriately. The set seems to use '?' for missing values so these values were replaced with Null Values which were dealt with as mentioned above.

DATA STORIES AND VISUALIZATIONS TOP PERFORMERS ANALYSIS

The initial findings of the analysis described the overall view of the set. The most popular brand is Toyota and the next three brands are also Japan based which hints at a certain trend in consumer behavior. However, what is also important to note are the five brands on the end of the scale as they are considered as more luxury vehicles which should reflect later in their price. The most popular style is a sedan followed by a hatchback. Both cars are suited to all-purpose functionality while the remaining three have specific functions and are less frequently chosen. Risk rating is evenly distributed amongst the cars. Looking at the specific makes it appears Volvo is seen as the safest pick as it returns a score of -1. Porsche is seen as the riskiest car and is likely to incur higher insurance premiums.

Focus on correlation reveled further interesting points of analysis. Price is more correlated with engine size and curb weight of the car. Curb weight is mostly correlated with engine size, length, width and wheel based which is expected as these adds up the weight of the car. Wheel base is highly correlated with length and width of the car

The price distribution is quite heavily influenced by the presence of the luxury brands mentioned before but the histogram shows they make up a small % of the dataset. The table isolating cars costing more than $20 000 confirms this as the brands located in this category are made up of Audi, BMW, Porsha, Volvo and Mercedes-Benz.

**TOP PERFORMERS ANALYSIS**

As identified earlier the data presented a good opportunity to take a closer look at 4 direct competitors in the market; Toyota, Nissan, Honda and Subaru. The symbolling shows a pattern within the type of car. Sedans of 3 brands are 0 rated while hatchbacks score a minimum of 1.

The price range is reflected in the various body styles. Honda and Subaru represent the cheaper options while Nissan and Toyota (the two most popular choices) are costlier. Particularly interesting is the price of the Nissan hatchback compared to the rest of the market. Compared to the Toyota it comes in around a 50% price increase. The three scatter plots compare price to features of the car. Engine size as well as Miles per gallon in the city and highway. The three features are tied to efficient performance of the car. As engine size increases, price rises substantially in both Toyota and Nissan. This correlates to use of fuel which is inversely related to the engine relationship. As the size increases the larger engine starts to consume more fuel at a less efficient rate. This is further enhanced by the two following scatter plots which show the relationship between weight of the car towards miles per gallon

Analysis of the four specific brands of cars shows some interesting exploratory points

How the data set are distributed and correlation between different fields and how they are related, the symbolling shows cars are initially assigned a risk factor symbol associated with its price. Within the four brands chosen what can be seen is that Toyota presents the most cost-effective option taking all factors into account. Although more expensive than Honda and Subaru, the car provides better performance with miles per gallon, an equal risk rating and more enhanced specifications of the car. Nissan seems overpriced based on its competitors as it provides worse performance in many categories compared to the other three brands at a higher price.

THIS REPORT WAS WRITTEN BY: Andrew Earle

