

**TASK**

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

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

**Introduction**

Summary of the data set

**DATA CLEANING**

# SUMMARY OF THE METHODS AND VISUALIZATIONS DONE DURING DATA CLEANING

MISSING DATA

# ANY MISSING DATA? HOW DID YOU HANDLE IT

DATA STORIES AND VISUALIZATIONS

# THIS IS THE BULK OF THIS PROJECT. EXTRACT STORIES AND ASSUMPTIONS BASED ON VISUALIZATIONS OF THE DATA

# ENSURE THIS DOCUMENT IS NEAT AND CAN BE ADDED IN YOUR PORTFOLIO

**THIS REPORT WAS WRITTEN BY: ABRAHAM WOROF**



**The Automobile dataset**

The Automobile dataset consisted of 26 columns, namely:

['symbolling', 'normalized-losses', 'make', 'fuel-type', 'aspiration', 'num-of-doors', 'body-style', 'drive-wheels', 'engine-location', 'Wheel-base', 'length', 'width', 'height', 'curb-weight', 'engine-type', 'num-of-cylinders', 'engine-size', 'fuel-system', 'bore', 'stroke', 'Compression-ratio', 'horsepower', 'peak-rpm', 'city-mpg', 'Highway-mpg', 'price'].

**DATA CLEANING AND MISSING DATA**

Fortunately, when running the code and checking for null values (using the isnull() function and plotting a heatmap) no null values were found. When I went further and analysed the data, I realized that we indeed had null values but a question mark (?) was in place of all these null values. I then went further to replace all the columns with question marks with NaN values using the .replace function. To ensure that our data doesn’t change the mean for all the columns with NaN values were calculated using the .mean() function and then inserted in place of all the NaN values using the .fillna() function. We can now analyse our dataset with no issues.

DATA STORIES AND VISUALIZATIONS

* I started by creating a heatmap of all the columns with numerical values. The heatmap shows that the "Price", "Peak-rpm", and "Curb-weight" are the top 3 columns with the highest values. The second visualization shows the numbers of different car makes. Toyota, Nissan, and Mazda are 3 most common cars and Mercury, Renault and Chevrolet are the 3 least most common cars.
* Next, I created a box plot visualization of the Numbers of different car body-styles. The visualization only a small amount of people owns convertible and hardtop cars. The visualization clearly show that the sedan body-style and hatchback are the most desirable car body styles.
* I then created 2 box plot visualizations, one for the Prices for the different cars and one for the Range of the car’s horsepower. The box plot of price. The average price for cars is just above 10 000. The visualization also shows multiple outliers for the maximum price for the cars with the most expensive ones ranging from 40 000 to just over 45 000. The horsepower box plot average cars horsepower is just under 100 with most of the cars ranging from about 75 to 115. The visualization also shows us a few outliers with the highest value of about 290 horsepower.
* From there, I then created a regplot which is basically a scatterplot and a linear regression in one. The plots check for the relationship between Price and Horsepower. This visualization suggests that there is a direct relationship between Price and Horsepower. The higher the price the higher the horsepower.
* The next visualization created was a histogram plot of the Engine-sizes for the different cars. The visualization shows us that most of common engine sizes were from 100 to 150.
* A special type of scatterplot was then created to check for the relationship between price and engine-size and the relationship of horsepower and engine-size. The scatter plot is supported with a key that shows how much horsepower a car has. The graph also shows us that they are also a direct relationship between the price and the engine-size. The graph also suggests that the higher the engine size the higher the horsepower because on average the dots become darker.
* 3 scatter plot that shows the relationship between Length, Width and Height. All three of the graphs have a similar relationship (The higher the x value the higher the y value). The last graph however has a much weaker relationship as compared to the first two graph as there are many outliers.
* A regplot to show the relationship between Highway-mpg Vs Cit graph shows a very strong direct relationship between Highway-mpg and City-mpg. y-mpg. The visualization.
* A box plot of the different car engine locations. Almost all cars have their engines in the front while a very small percentage of people have their car engines at the rear end of the car.
* A box plot of the different car drive wheels. We have 3 different drive-wheels (rwd, fwd and 4wd). Most cars are fwd followed by rwd and a very small percentage of cars are 4wd.
* A count plot of the different compression ratios of cars. There are many different car compression ratios but only a few stand out from the visualization. Mainly the compression ratio with 9.0 as it is by far the most common ratio with almost 50 cars having it, followed by 9.4 with about 25 cars and both 8.5 and 9.5 with more or less a similar number of cars of 14.
* A violin plot of the range of a cars peak-rpm. The average peak-rpm is about 52 000 with most of the cars having a peak rpm between 48 000 and 55 000.
* Count plot of different car strokes. There are many different car strokes but only a few stand out from the visualization. These are 3.4 with the highest stroke of about 20 cars. 3.23, 3.15 and 3.03 with about 14 cars.

I then printed the most common values for the columns “body-style”, “drive-wheels”, and “aspiration”.