

**TASK**

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

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

**Introduction**

This dataset comprises of information about numerous cars makes. The data is broken down into many different columns specifying information about each make of car and their respective performance.

**DATA CLEANING**

**SUMMARY OF THE METHODS AND VISUALISATIONS DONE DURING DATA CLEANING**

The first operation that was performed was to identify any column with missing data. When doing this, I found that no columns had missing data but instead had “?” in place of a missing value. The next operation performed was to drop any unwanted columns (symbolling, normalized-losses, bore, stoke) that I didn’t need in my analysis. Any rows with duplicates were also removed to clean the data further. A check was then done to see what data types each column was using and if any changes needed to be made. For me the num-of-cylinders column needed to be changed from object into an int64 or float64 in case of performing any calculations. The price column also needed to be changed to int64 for calculations as the values were being stored as objects. The peak-rpm and horsepower columns needed to be changed from objects to int64 as well. I changed the drive-wheels column from abbreviations to the full terms to describe the respecting drive-wheels. This was done as some users may mistake “fwd” as “four-wheel drive” instead of “front wheel drive” which can cause inconsistencies when analysing the data.

**MISSING DATA**

**ANY MISSING DATA? HOW DID YOU HANDLE IT**

There was missing data in the normalized losses column that was represented by “?”, in this instance the entire column was dropped as it was unnecessary in my EDA. The price column also had missing values represented by “?” which I then converted to NaN to ignore them when doing any analysis.

**DATA STORIES AND VISUALISATIONS**

**THIS IS THE BULK OF THIS PROJECT. EXTRACT STORIES AND ASSUMPTIONS BASED ON VISUALISATIONS OF THE DATA**

**The first comparison I wanted to do was to see which were the 5 most expensive cars in the data and then compare them to the 5 cheapest cars in terms of horsepower.**

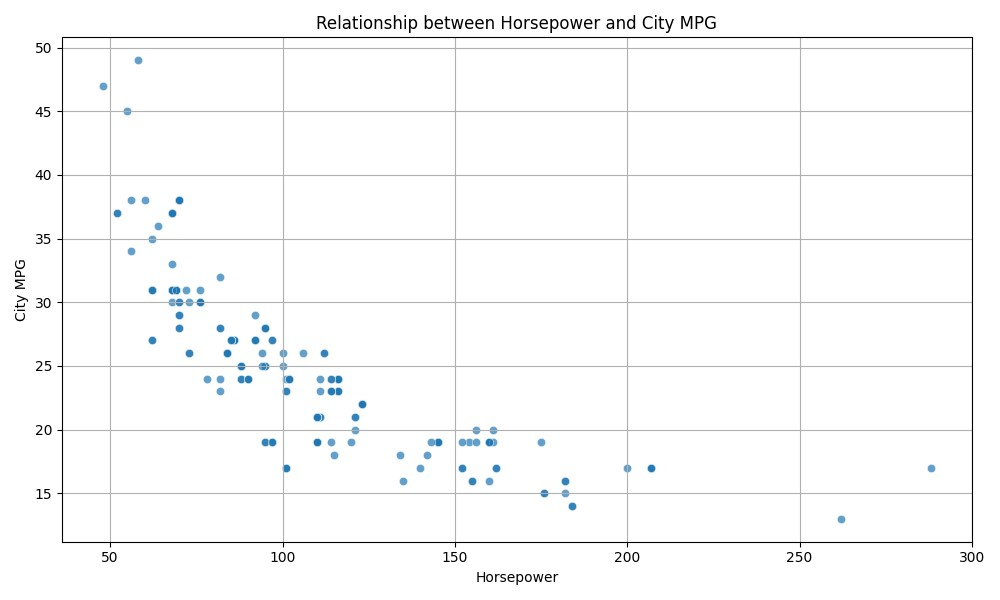
When creating the visualization to do this it was evident that the 5 most expensive cars had drastically more horsepower than the 5 cheapest cars. This could be due to the fact that more expensive cars come with larger and more powerful engines.

A chart of a comparison of cars

Description automatically generated

**Do cars with a greater horsepower have a higher city-mpg?**

This created an interesting scatterplot that showed that greater the horsepower, the lower the city-mpg which I would have assumed the opposite of. Generally, cars with a greater horsepower tend to have bigger engines thus burns more fuel so should have a higher horsepower to city-mpg.



**Do cars with a fuel type Gas or Diesel have a higher highway-mpg?**

I created a bar plot to display this comparison and when viewing the bar plot, I could see that there wasn’t a big difference between the two gas types, but diesel did have a higher highway-mpg compared to gas. Diesel is a slow burning fuel so it makes sense that it would give you a greater mpg on the highway or on longer distances.

A graph showing the comparison of fuel types and average highway mpg

Description automatically generated

**Do cars with a greater engine size have a better or worse compression ratio?**

I used a scatterplot to show me this information, upon viewing the scatterplot the readings were average although there were a few outliers in the 100 to 200 engine size that had a relatively high compression ratio. Overall, cars with smaller engines seemed to have a smaller compression ratio as compared to cars with big engines.

A graph of a relationship between engine size and compression ratio

Description automatically generated

**Do cars with more cylinders have a greater or smaller peak-rpm?**

Using a line graph seemed to have given the best representation of this data. When analysing the data it shows a negative relationship which means that the less cylinders a car has, the higher the peak-rpm. One of the reasons could be engines with fewer cylinders tend to have more significant vibrations due to the uneven power delivery. To mitigate this, engines with fewer cylinders often have lower peak RPM’s to reduce the likelihood of encountering the rough vibrations that come with high RPM’s.

A graph with a line graph

Description automatically generated

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

**THIS REPORT WAS WRITTEN BY : JASON LUKE EDWARD**

