

**TASK**

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

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

**Introduction**

Dataset of vehicle information, including make, fuel type, MPG, price, horsepower, door number, and wheel drive.

**DATA CLEANING**

First redundant columns were removed, which were: symbolling, normalised losses, engine location, fuel system, bore, stroke, compression ratio, length, width, height, curb weight, engine type, and peak rpm. Then duplicates were removed.

**MISSING DATA**

No NaN values were returned, however after some examination it transpired that some values were entered as “?” in both the *price* and *horsepower* columns. This meant it was not possible to get accurate outputs for searches such as top 5 most expensive cars. As such, they were replaced with -1 so that the data could still be examined. It was not deemed necessary to remove those rows entirely as other data was accurately included, and a figure of -1 for either price or horsepower is an obvious outlier, meaning it can easily be disregarded in analysis.

**DATA STORIES AND VISUALISATIONS**

Table

Description automatically generated  
The above is the result for the top 5 most expensive cars. There are 3 different makes – Mercedes-Benz, Porsche, and BMW. All use gas fuel, 3 of them are sedans, and the others are hardtop and convertible.

Table

Description automatically generated

Above shows the top 5 biggest engines, of which there are only 2 different makes – Jaguar and Mercedes-Benz. Of these 5, the biggest engine also has the most horsepower. Both Mercedes-Benz cars are also more expensive than all of the Jaguar cars in these results, and they are the same 2 entries that are on the top 5 most expensive cars above. There are also no Jaguar cars on the most expensive list, suggesting that spending more doesn’t necessarily determine a bigger engine size.

Table

Description automatically generated  
This table shows the top 5 highest horsepower. Four of these results are one make, Porsche, and the remaining result is the same Jaguar car that had the largest engine. The top result – a Porsche car – has a price of -1, indicating that this was cleaned and did not have an accurate input in the original dataset. However, given that the one Porsche car that did feature on the top 5 most expensive was the 4th result, we could reasonably assume that this one is not more expensive than the most expensive Mercedes-Benz car. Once again, this indicates that spending more money does not promise greater horsepower, especially as there are no Mercedes-Benz cars in these results.

Chart, pie chart

Description automatically generated

This pie chart illustrates the distribution of data across car makes. Toyota has the most entries, at 16% of the total data in this dataset, with Nissan the next highest. Across all the other car makes, the increase in data entries is gradual – increasing by 1% each time there is an increase – but between Nissan and Toyota there is a significant difference of 7%. This means Toyota is over-represented in this dataset.

A further interesting insight from this pie chart is that Jaguar and Porsche only represent 1% and 2% of the data respectively, but they both were included multiple times in tables above regarding engine size, horsepower, and price. Toyota, on the other hand, did not feature once.

**Chart, bar chart, histogram

Description automatically generated**

This bar chart shows car price by make. As expected, Mercedes-Benz is the most expensive, with BMW and Porsche in 2nd and 3rd place respectively. The cheapest is Chevrolet. More makes in this dataset are priced below 20,000 than above. Chevrolet is the only make that is priced below 10,000.

Chart, pie chart

Description automatically generated

This second pie chart shows the fuel types represented in the dataset, and the overwhelming majority are gas rather than diesel.

**Chart, bar chart

Description automatically generated**

This stacked bar chart shows the top 5 results for highest city and highway MPG. None of these car makes were featured on any of the earlier tables regarding price, engine size, or horsepower. This suggests that cheaper cars tend to have smaller engines and horsepower, but will get better results on fuel economy. Honda has the highest MPG at roughly 50MPG for city, and slightly more for highway. Generally speaking, there is little difference between city and highway MPG, though highway MPG is slightly more for all makes.

Toyota features twice on this graph, indicating that despite it not having the most fuel efficient car, there are choices amongst Toyota models for highly fuel efficient cars, which is not the case for the other makes in this table.

**Chart, scatter chart

Description automatically generated**

This scatter graph shows the positive correlation between engine size and horsepower, however the 2 upper outliers clearly indicate that it is possible to have greater horsepower even with a smaller engine. The anomaly at the bottom of the graph is likely a result of the aforementioned data cleansing (replacing ‘?’ with -1).

Chart, scatter chart

Description automatically generated

This final scatter graph shows the negative correlation between engine size and highway MPG. When looking at the engine size of the previous outliers (200 and >300), we can see that the smaller engine is also likely to be slightly more fuel efficient.

In all, this data can be useful for decision making when purchasing a new car. The above visualisations highlight that cheaper cars are more fuel efficient, making them cheaper to run even after purchasing. However, they do not offer competitive horsepower, which may be a priority. If so, the final scatter graphs highlight that a smaller large engine can still offer greater horsepower and be more fuel efficient, so engine size does not necessarily need to be prioritised in the decision making. Also, given that most of the cars in this dataset are powered by gas rather than diesel, separate investigation of fuel cost and availability differences could be carried out in order to decide if any of this data is useful, or if it would be better to collect more data on diesel-fuelled cars and conduct a new analysis.

**THIS REPORT WAS WRITTEN BY : Sara Lidguard**

