Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

The data set contained in the file ‘automobile.txt’ consists of 205 rows and 26 columns. Each row in the data set represents a particular model of automobile, and the columns contain information about that model’s specifications, performance and other attributes, to be outlined further in the DATA CLEANING section below.

The working behind the figures and analyses given in this report is contained in the file ‘automobile.ipynb’.

DATA CLEANING

A query of the unique values in each column produced the following output:

```
symboling: [ 3  1  2  0 -1 -2]
normalized-losses: ['?' '164' '158' '192' '188' '121' '98' '81' …]
make: ['alfa-romero' 'audi' 'bmw' 'chevrolet' 'dodge' 'honda' …]
fuel-type: ['gas' 'diesel']
aspiration: ['std' 'turbo']
num-of-doors: ['two' 'four' '?']
body-style: ['convertible' 'hatchback' 'sedan' 'wagon' 'hardtop']
drive-wheels: ['rwd' 'fwd' '4wd']
engine-location: ['front' 'rear']
wheel-base: [ 88.6  94.5  99.8  99.4 105.8  99.5 101.2 103.5 110. …]
length: [168.8 171.2 176.6 177.3 192.7 178.2 176.8 189.  193.8 197. …]
width: [64.1 65.5 66.2 66.4 66.3 71.4 67.9 64.8 66.9 70.9 60.3 63.6 …]
height: [48.8 52.4 54.3 53.1 55.7 55.9 52.  53.7 56.3 53.2 50.8 50.6 …]
curb-weight: [2548 2823 2337 2824 2507 2844 2954 3086 3053 2395 2710 …]
engine-type: ['dohc' 'ohcv' 'ohc' 'l' 'rotor' 'ohcf' 'dohcv']
num-of-cylinders: ['four' 'six' 'five' 'three' 'twelve' 'two' 'eight']
engine-size: [130 152 109 136 131 108 164 209  61  90  98 122 156  …]
fuel-system: ['mpfi' '2bbl' 'mfi' '1bbl' 'spfi' '4bbl' 'idi' 'spdi']
bore: ['3.47' '2.68' '3.19' '3.13' '3.50' '3.31' '3.62' '2.91' …]
stroke: ['2.68' '3.47' '3.40' '2.80' '3.19' '3.39' '3.03' '3.11' …]
compression-ratio: [ 9.   10.    8.    8.5   8.3   7.    8.8   9.5 …]
horsepower: ['111' '154' '102' '115' '110' '140' '160' '101' '121' …]
peak-rpm: ['5000' '5500' '5800' '4250' '5400' '5100' '4800' '6000' …]
city-mpg: [21 19 24 18 17 16 23 20 15 47 38 37 31 49 30 27 25 13 26 …]
highway-mpg: [27 26 30 22 25 20 29 28 53 43 41 38 24 54 42 34 33 31 …]
price: ['13495' '16500' '13950' '17450' '15250' '17710' '18920' …]
```

From this, it became clear that the following columns were in an inappropriate data type for analysis:

- num-of-doors: `object`, should be `int`
- bore: `object`, should be `float`
- stroke: `object`, should be `float`
- horsepower: `object`, should be `int`
- peak-rpm: `object`, should be `int`
- price: `object`, should be `int`

At the same time, each of these columns contained at least one value given as '?' (to be further discussed in the MISSING DATA section below), which could not be cast to `int` or `float`. I decided to write functions to selectively cast non-'?' values to `int` or `float` respectively, and leave the '?' values untouched.

It also became clear that the columns 'num-of-doors' and 'num-of-cylinders' had their values written as strings rather than integers, e.g. 'two' instead of 2. I converted these to the appropriate integers with the `replace` method.

MISSING DATA

In the sense of fields with no value, there were no missing data in the dataset. However, '?' appeared as a value in the following columns with the following frequency:

normalized-losses	41
num-of-doors	2
bore	4
stroke	4
horsepower	2
peak-rpm	2
price	4

These can be considered missing data. The number of missing values in the 'normalized losses' column was so large that I decided to drop this column. The number of missing values in all the other columns was so small that I decided to simply exclude the missing values as appropriate on an analysis-by-analysis basis.

There were no other values that could be considered missing data.

DATA STORIES AND VISUALISATIONS

I will approach this task by means of asking questions of the data, and showing visualisations to answer them.

What markets are the different manufacturers targeting?

From the boxplot shown in Figure 1, we can see that there are four manufacturers who are targeting the luxury car market: BMW, Jaguar, Mercedes-Benz, and Porsche. This can be seen from the fact that the distribution of their prices is so clearly above the others.

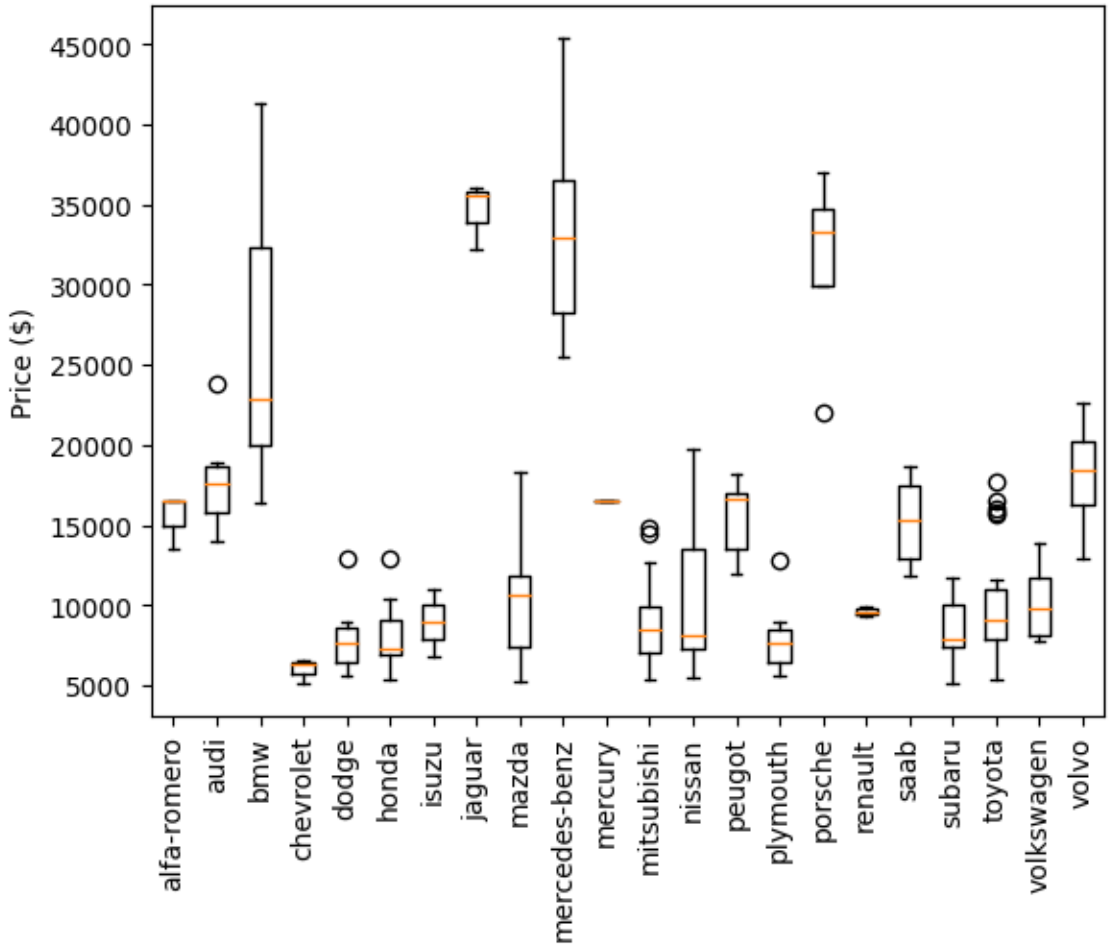

Figure 1: Price by manufacturer

What types of car are getting produced the most?

In terms of the number of different models out there, we can see from the pie chart in Figure 2 that gas vehicles are much more numerous than diesels.

A comparison between different body styles as shown in Figure 3 reveals that sedans account for almost half of the vehicle models in the data set. Hatchbacks are also popular.

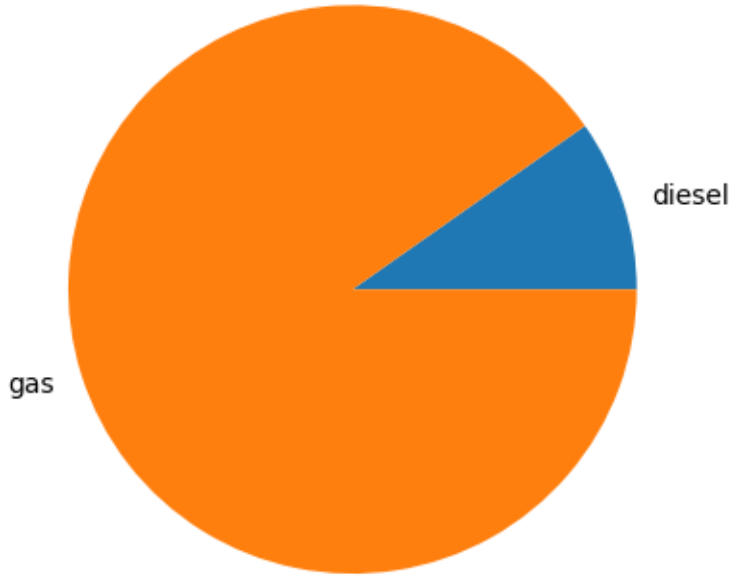

Figure 2: Comparison between the numbers of gas and diesel vehicle models

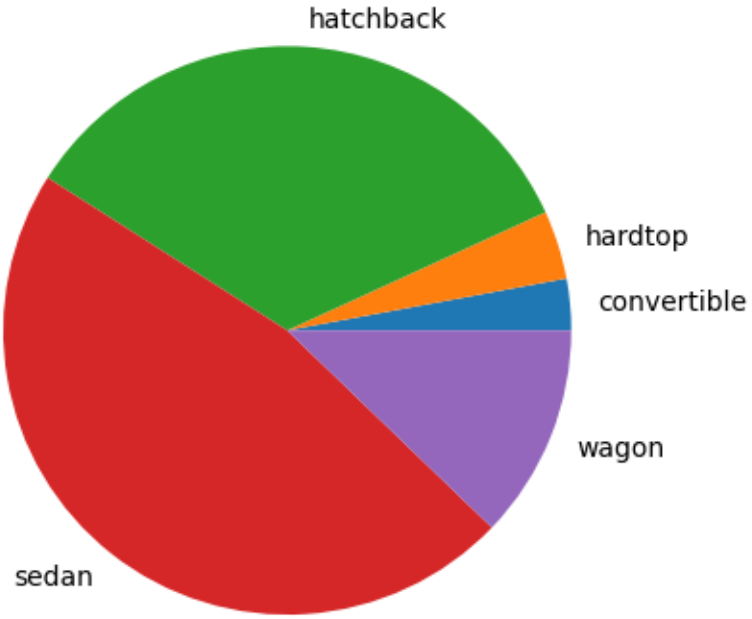

Figure 3: Comparison between the different body styles

What determines performance?

Taking horsepower as a measure of performance, we can see from the scatter plot in Figure 4 that this is very strongly associated with engine size.

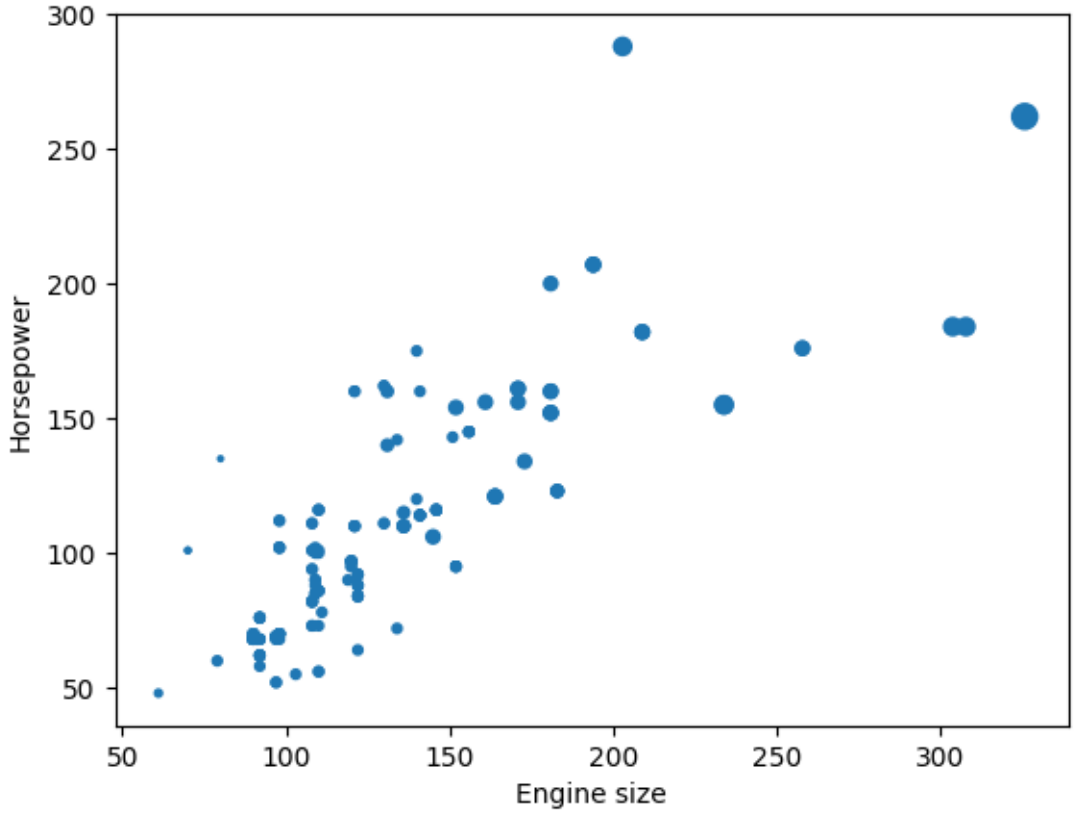

Figure 4: Horsepower by engine size and number of cylinders

In Figure 4, the size of the dot indicates[1] the number of cylinders that engine has. This shows that larger engines all but require a greater number of cylinders; a point which is made even more clearly by the scatter plot shown in Figure 5.

[1] The size of the dot is an exponential function of the number of cylinders. I made this choice in order to accentuate the differences.

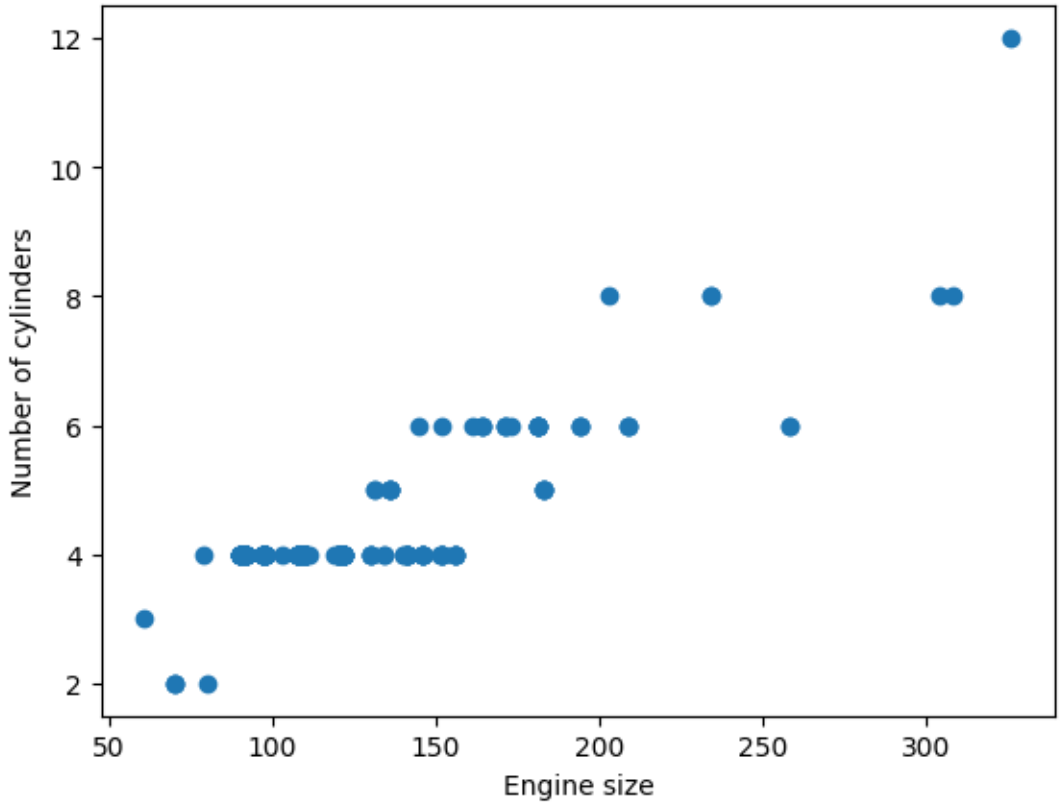

Figure 5: The relation between engine size and number of cylinders

Another engine property that is associated with horsepower is bore size. As Figure 6 shows, bore size is positively correlated with horsepower.

Finally, as shown in Figure 7, horsepower is also associated with the drive wheels of the vehicle: rear wheel-drive vehicles have greater horsepower than either front wheel-drive vehicles or four wheel-drive vehicles.

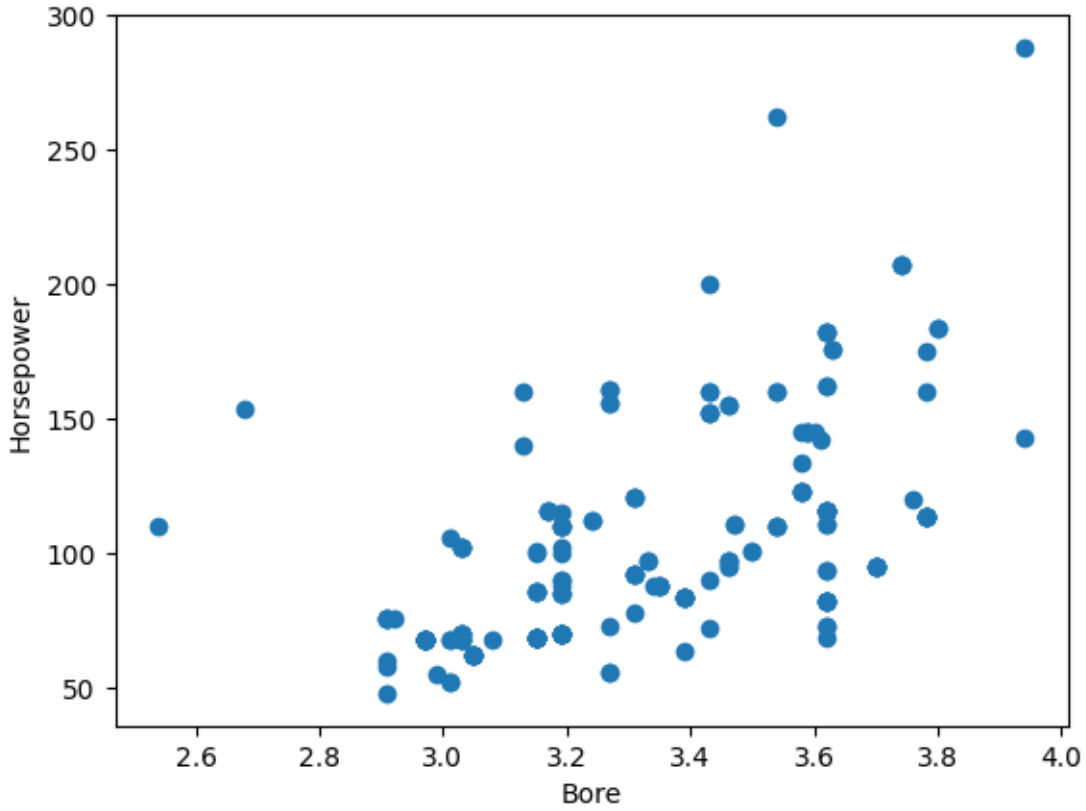

Figure 6: Horsepower by bore size

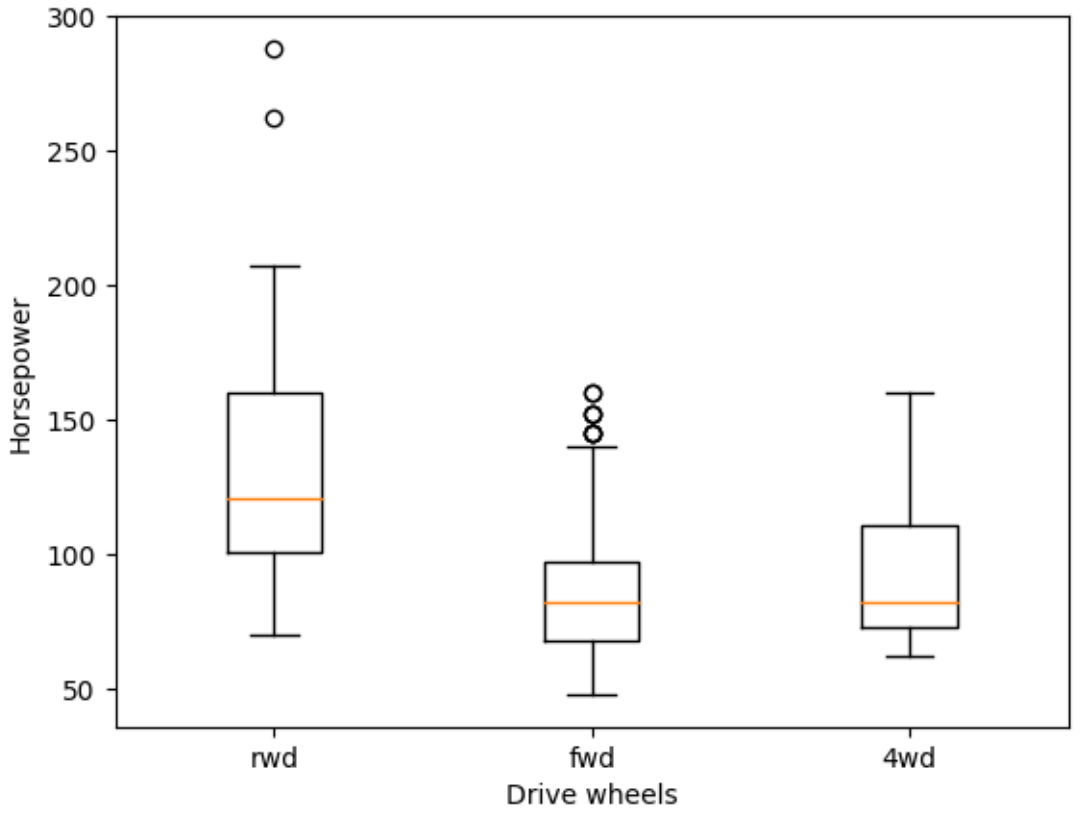

Figure 7: Horsepower by drive wheel system

What determines fuel efficiency?

One of the answers to the previous question showed that bore size is positively correlated with performance, as measured by horsepower. The flipside to this, however, is that bore size is *negatively* correlated with fuel efficiency – both in the city and on the highway, as Figure 8 shows.

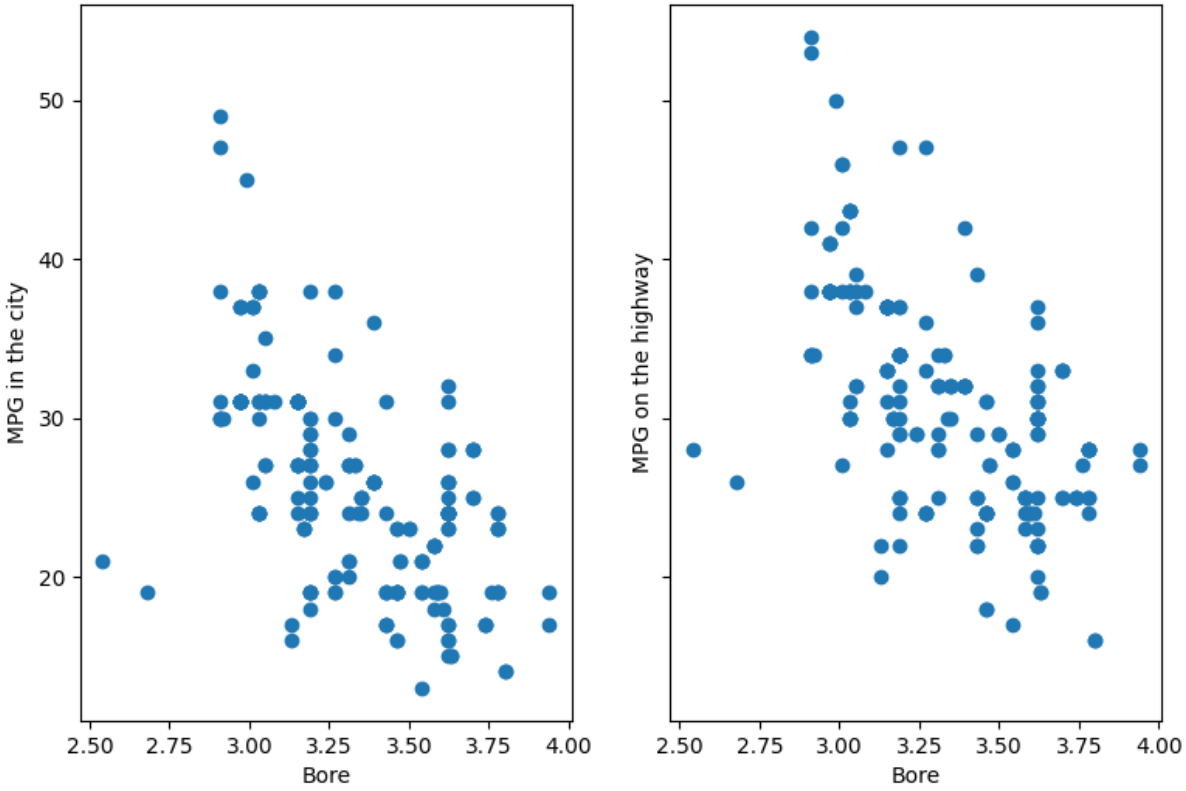

Figure 8: MPG by bore size, in the city and on the highway

This invites the question: is there a simple trade-off between horsepower and fuel efficiency? Figure 9 shows that the answer is a clear 'yes'.

Another factor affecting fuel efficiency is fuel type. As the box plot in Figure 10 shows, diesel vehicles tend to be more fuel-efficient that gas vehicles, both in the city and on the highway.

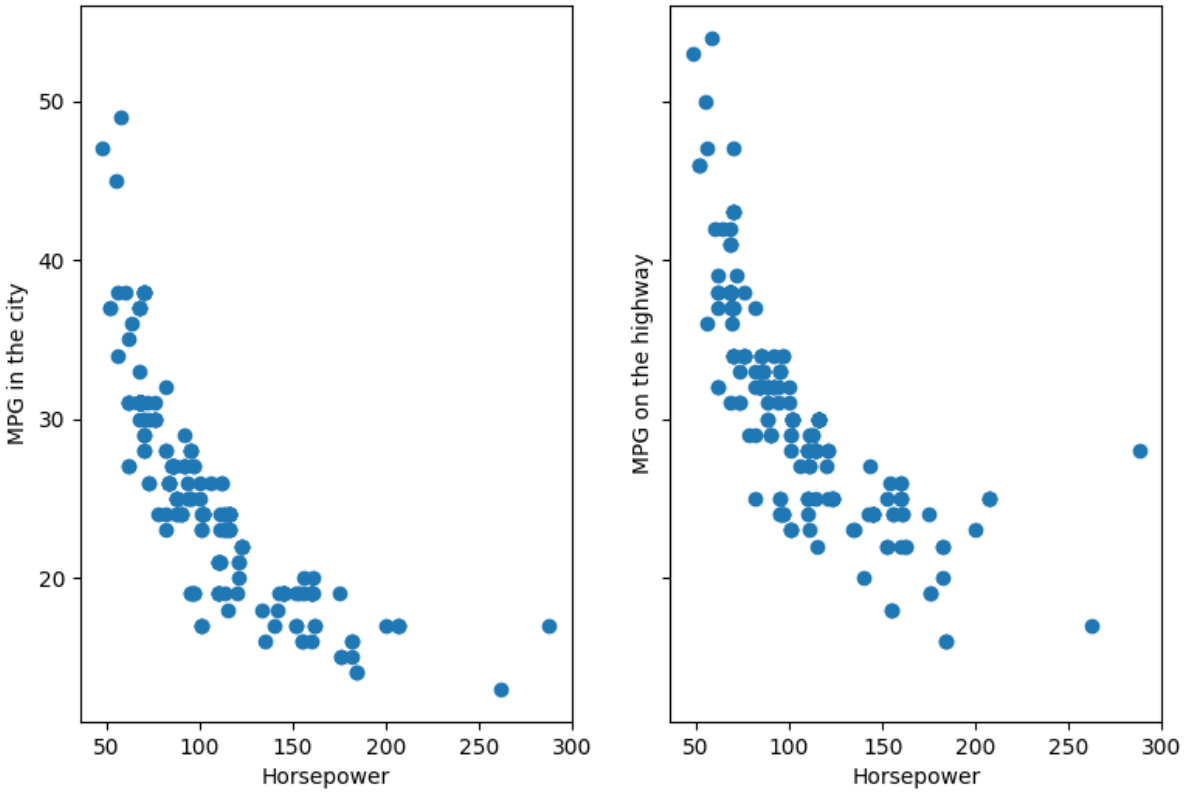

Figure 9: The association between horsepower and fuel efficiency

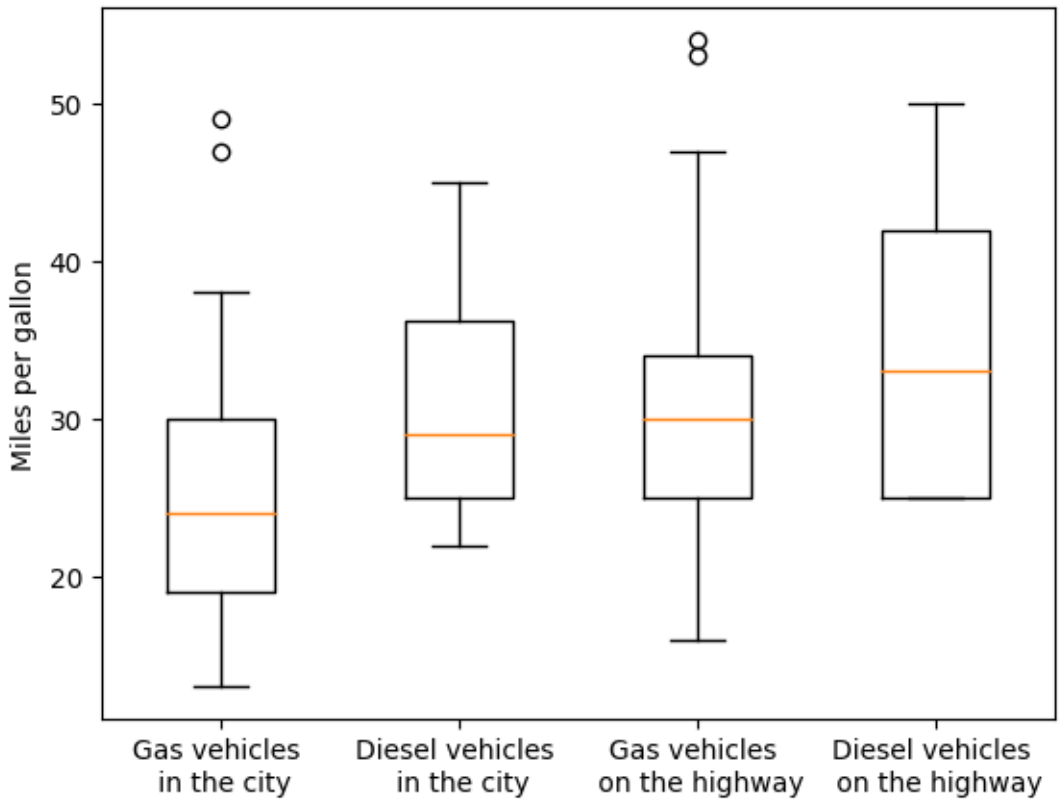

Figure 10: Fuel efficiency of different vehicle types

What does your money buy?

Why are the more expensive vehicles more expensive? Figure 11 shows that paying more for a vehicle gets you more horsepower. Figure 12 shows that this comes at the expense of a worse fuel efficiency. This is unsurprising: firstly, we have already seen that horsepower is negatively correlated with fuel efficiency; and secondly, people who can afford to pay more for a vehicle presumably don't have to worry about fuel efficiency so much.

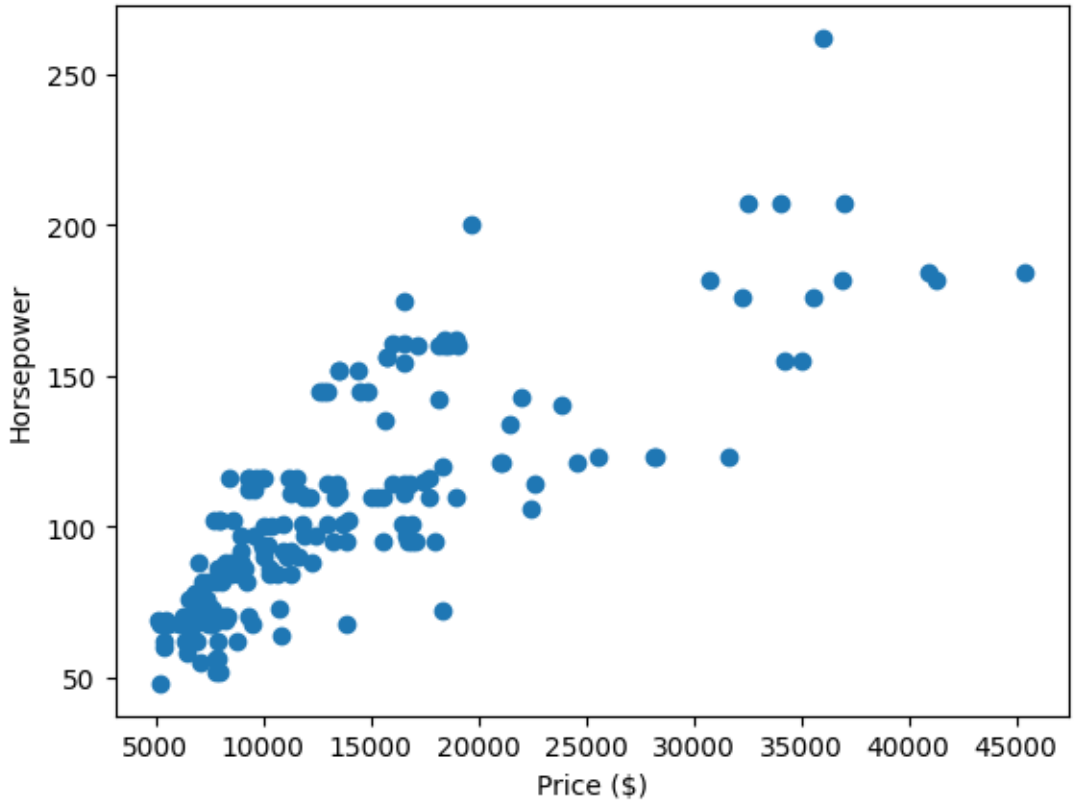

Figure 11: The extra horsepower that extra money buys

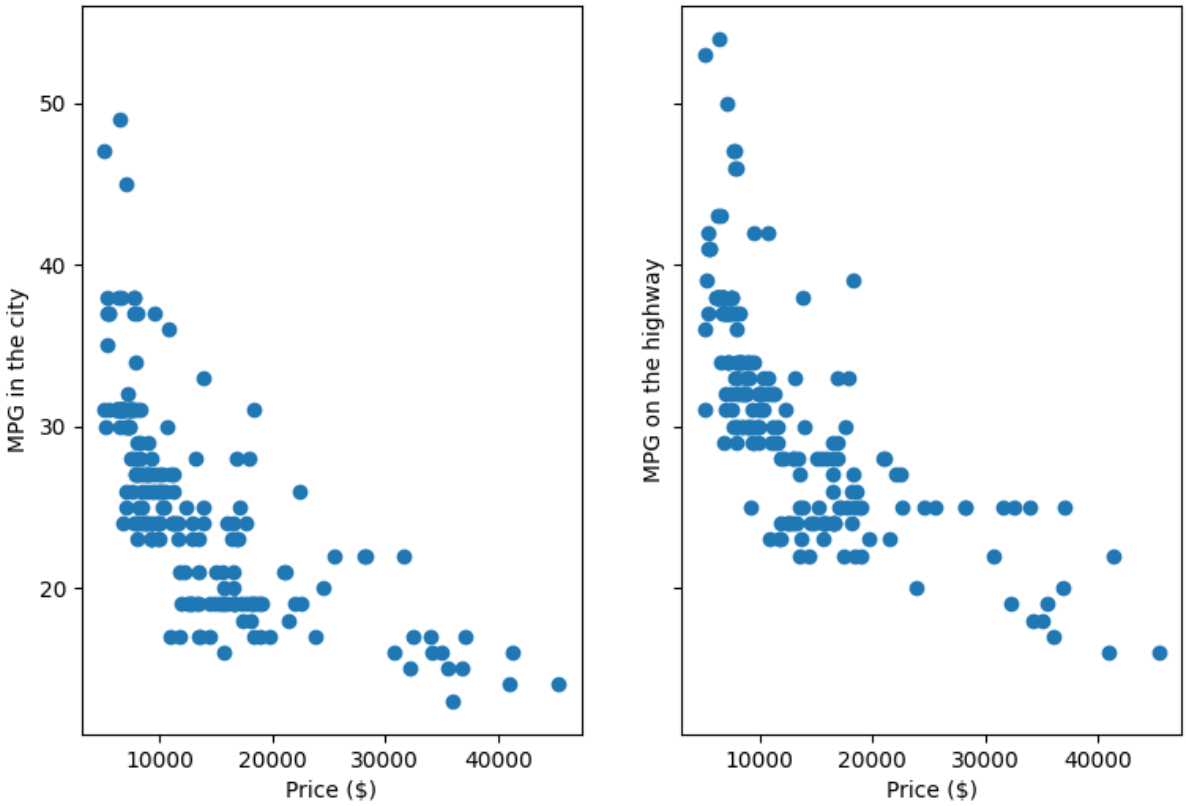

Figure 12: Fuel efficiency and price

THIS REPORT WAS WRITTEN BY: MATTHEW GOTHAM