Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

The dataset that will be looked at in this report is a collection of data about different cars. The attributes that will be focused on will be the car manufacturer, price, engine size, car size, horsepower, fuel consumption, weight, fuel type, and maximum RPM. Fuel consumption is measured in miles per gallon and the fuel type is given as either diesel or gas (petrol). Therefore, I am assuming that this dataset was based in the USA, meaning I will also assume that the price is given in dollars, the engine size is in cubic inches, and the weight is given in pounds. Note that the weight is actually the weight of the car will a full tank of fuel. The car size will be generalised by putting the cars into 2-door or 4-door categories.

Other attributes within this dataset could be used for insurance purposes, such as the symboling value and normalised losses. A car symboling value can be a whole number between the values of -3 and 3. It is a measure of how risky the car is compared to the price. -3 is considered very safe, while 3 is considered high risk. The normalisation losses are the amount of money that is lost by insurance each year without a claim

DATA CLEANING

The first step was to look through all the unique values within every column in order to quickly see if there are many empty values, or values formatted inconsistently.

After doing this, It was found that missing values were indicated by a question mark instead of a real empty value. Another note is that for some columns, some numeric values were given as strings. This was fixed using Pandas “to_numeric” function across the whole dataset. All question marks were also replaced with actual NaN values.

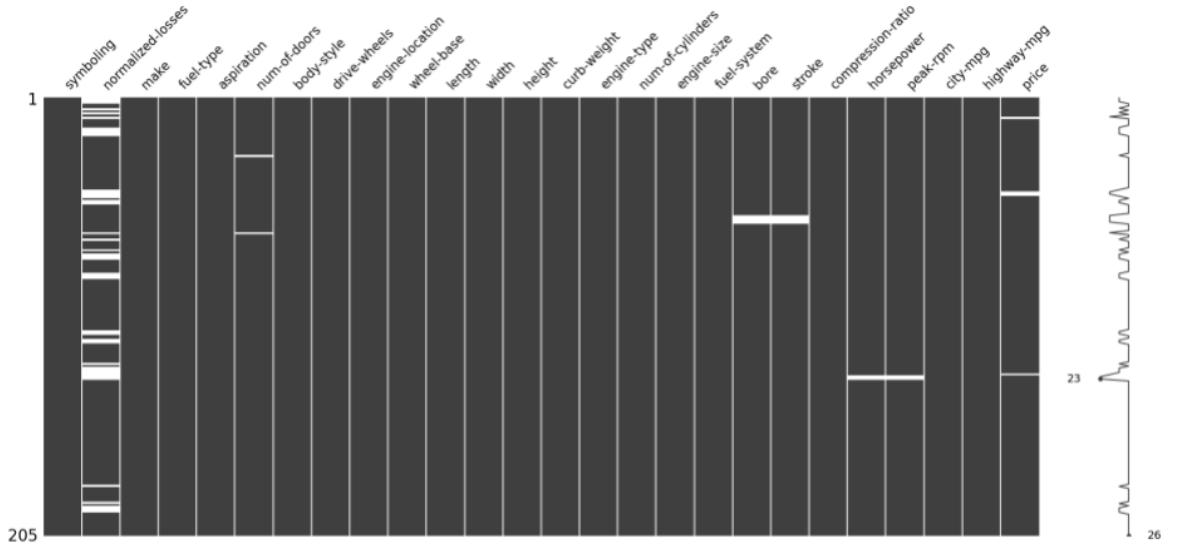

Fig 1: Missing value matrix

After converting all question marks into actual NaN values, a plot of all the empty cells can be plotted using the “missingno” library, shown in figure 1. This matrix shows that most of the missing values are within the “normalized-losses” column. There are also a few values in the num-or-doors, bore, stroke, horsepower, peak-rpm, and price columns.

	make	num-of-doors	bore	stroke	horsepower	peak-rpm	price	engine-size
9	audi	two	3.13	3.40	160.0	5500.0	NaN	131
27	dodge	NaN	3.03	3.39	102.0	5500.0	8558.0	98
44	isuzu	two	3.03	3.11	70.0	5400.0	NaN	90
45	isuzu	four	3.03	3.11	70.0	5400.0	NaN	90
55	mazda	two	NaN	NaN	101.0	6000.0	10945.0	70
56	mazda	two	NaN	NaN	101.0	6000.0	11845.0	70
57	mazda	two	NaN	NaN	101.0	6000.0	13645.0	70
58	mazda	two	NaN	NaN	135.0	6000.0	15645.0	80
63	mazda	NaN	3.39	3.39	64.0	4650.0	10795.0	122
129	porsche	two	3.94	3.11	288.0	5750.0	NaN	203
130	renault	four	3.46	3.90	NaN	NaN	9295.0	132
131	renault	two	3.46	3.90	NaN	NaN	9895.0	132

Fig 2: Snippet of rows that contain empty cells

Ignoring the normalized-losses column, figure 2 partially shows the other rows that contained empty cells. This was done to try and gain some insight into why some of these values were empty in the first place. If a car has a missing bore value, then the stroke will also be missing, and vice versa. The same relationship can be seen with the horsepower and peak-rpm values. All of the rows with missing bore and stroke values come from Mazda cars with the exact same horsepower, peak rpm, and engine size. The 2 rows with missing horsepower and peak-rpm values both come from Renault. This may have been an oversight of these manufacturers when releasing the specs of their cars. Another point is that both of the Isuzu cars without price values look almost like the exact same car, though, one has two doors, and the other has four.

The final action taken to understand the empty cells within the data set was to find the percentage of rows from each manufacturer that had at least one empty cell. This was done to see if the manufacturer themselves are responsible for the errors.

	Total	Empty	ratio
mazda	17	5	29.41
dodge	9	1	11.11
audi	7	1	14.29
porsche	5	1	20.00
isuzu	4	2	50.00
renault	2	2	100.00

Fig 3: Percentage of rows with empty cells by car manufacturer

Figure 3 shows these results. As seen, the two rows that contained the Renault cars were all of the Renault entries in the entire dataset. This leads me to believe that this is an issue with their data collection. A large

percentage of the Isuzu entries also contain empty cells. The ratios for the other manufacturers are relatively low.

MISSING DATA

The normalized-losses column was removed entirely. This decision was made because this attribute was not very useful for my analysis. It also contains a large portion of the empty cells within this dataset. Because the number of other rows with empty cells is relatively low, I simply removed these rows as necessary when generating the subsets of data.

When looking at the number of doors, only 2 rows have empty values, so leaving these out would not have a significant statistical impact on the resulting plots. This logic was used when justifying removing the rows that contained empty horsepower and peak-rpm values. Bore and stroke attributes were not looked at during this analysis.

DATA STORIES AND VISUALISATIONS

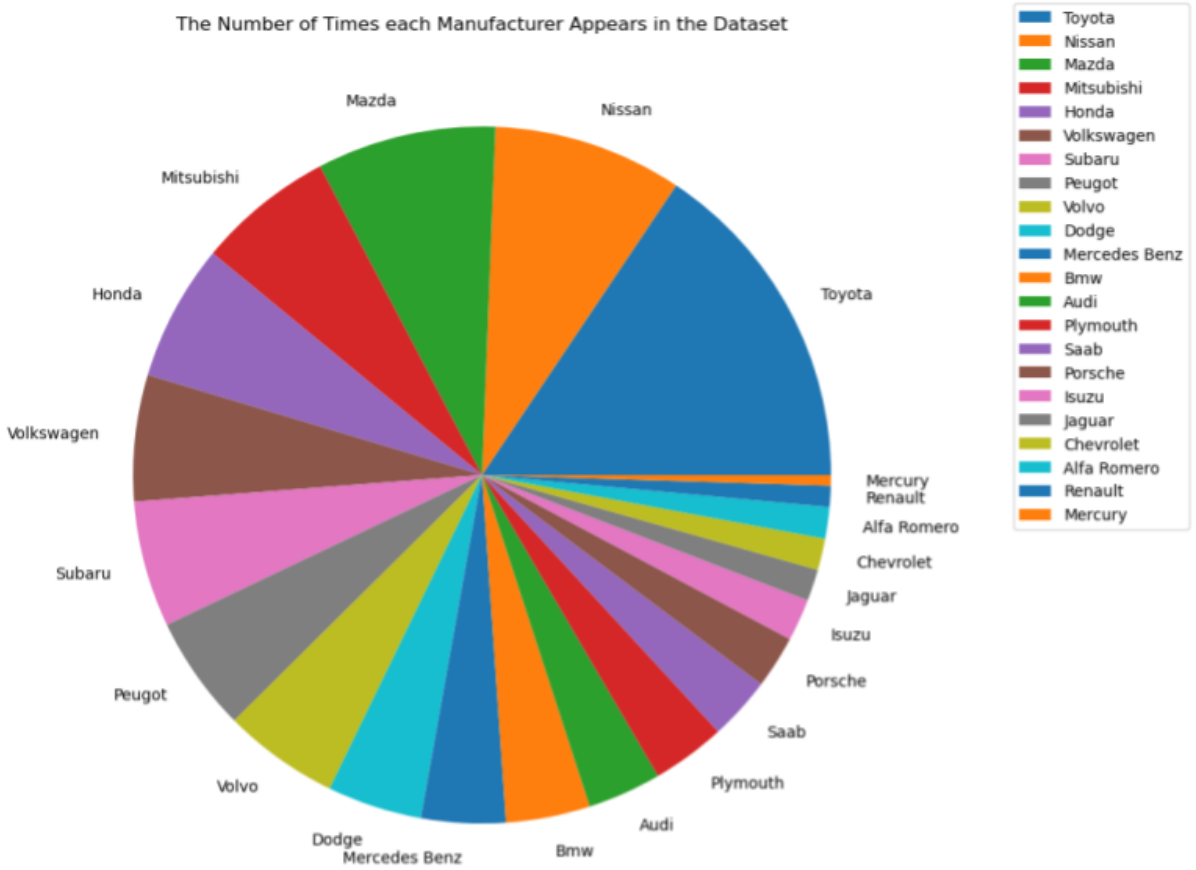

Fig 4: Pie chart showing the population of each manufacturer in the dataset

A plot was produced to gain insight into the share of the entries each car manufacturer has within this dataset. This is useful to understand if the population of specific manufacturers are large enough to be statistically significant. Figure 4 shows the results of this plot. As shown, Toyota contains the highest ratio of cars in the dataset, by a large amount, while Mercury contains the lowest. We know that Renault only has 2 entries in this set, therefore, Mercury must only have 1. Another piece of information that can be extracted is the fact that this dataset is only looking at 22 different manufacturers.

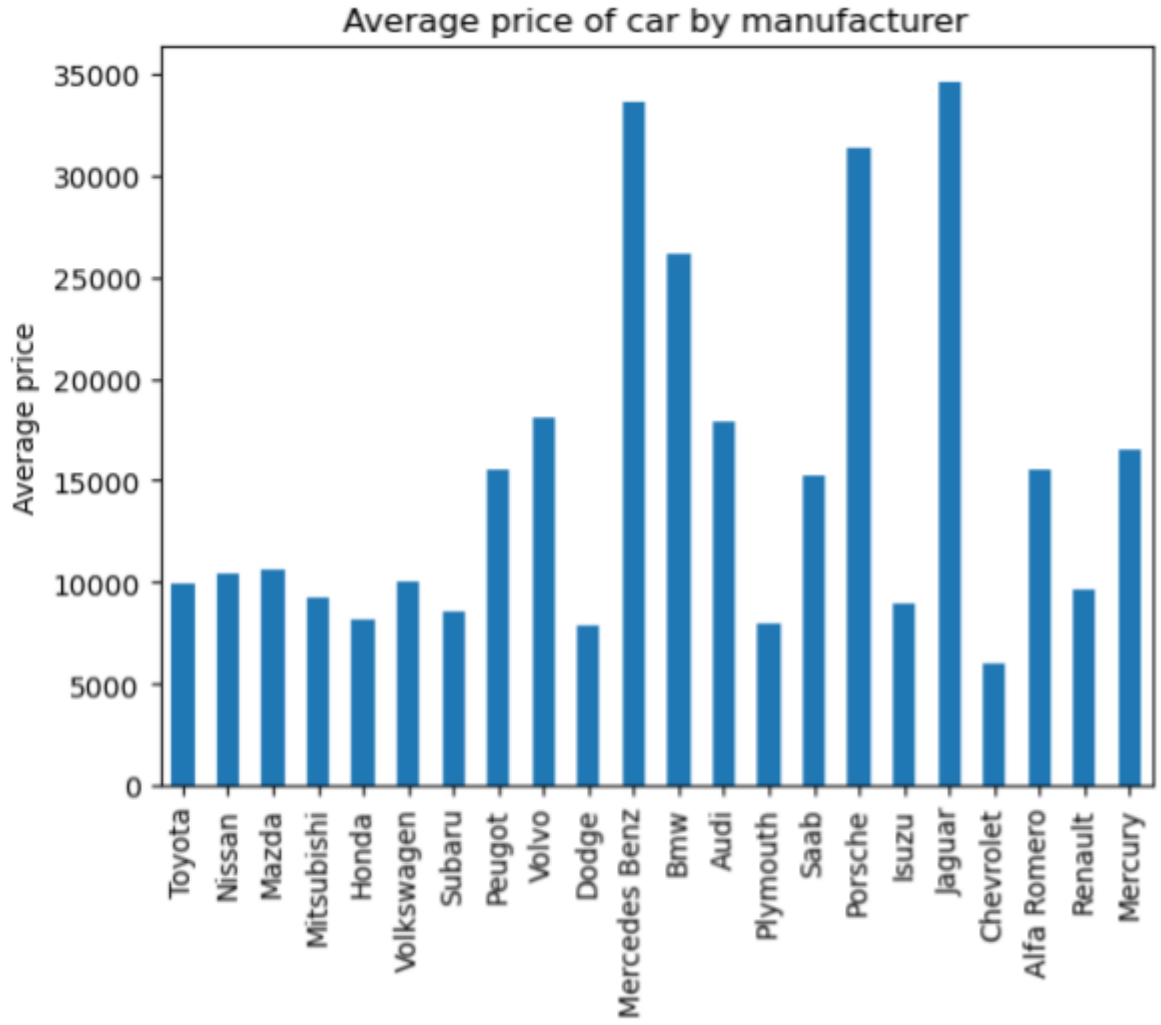

Fig 5: Average price of cars depending on their manufacturer

Figure 5 shows the average price of a car from specific manufacturers. The manufacturers are in the order of population in the dataset, the positive x direction is decreasing population proportion. We can extract the fact that Jaguar, Mercedes Benz, and Porsche are the 3 most expensive brands of cars. While Chevrolet is one of the cheapest. An interesting point is that the 7 most common brands in the dataset are also relatively cheap. There are no extremely expensive brands that are also very common. This could be for many reasons. First, If this dataset is useful for insurance purposes, more potential customers will have

averagely-priced cars as opposed to extremely expensive ones, meaning, having more average-priced cars in the dataset will be more helpful. Another reason could be that manufacturers that focus on more middle-range cars may produce a lot more models of similar prices for different uses. More expensive brands tend to only produce a few models due to the high price of production, leading to the cheaper cars having more entries and therefore, a higher proportion.

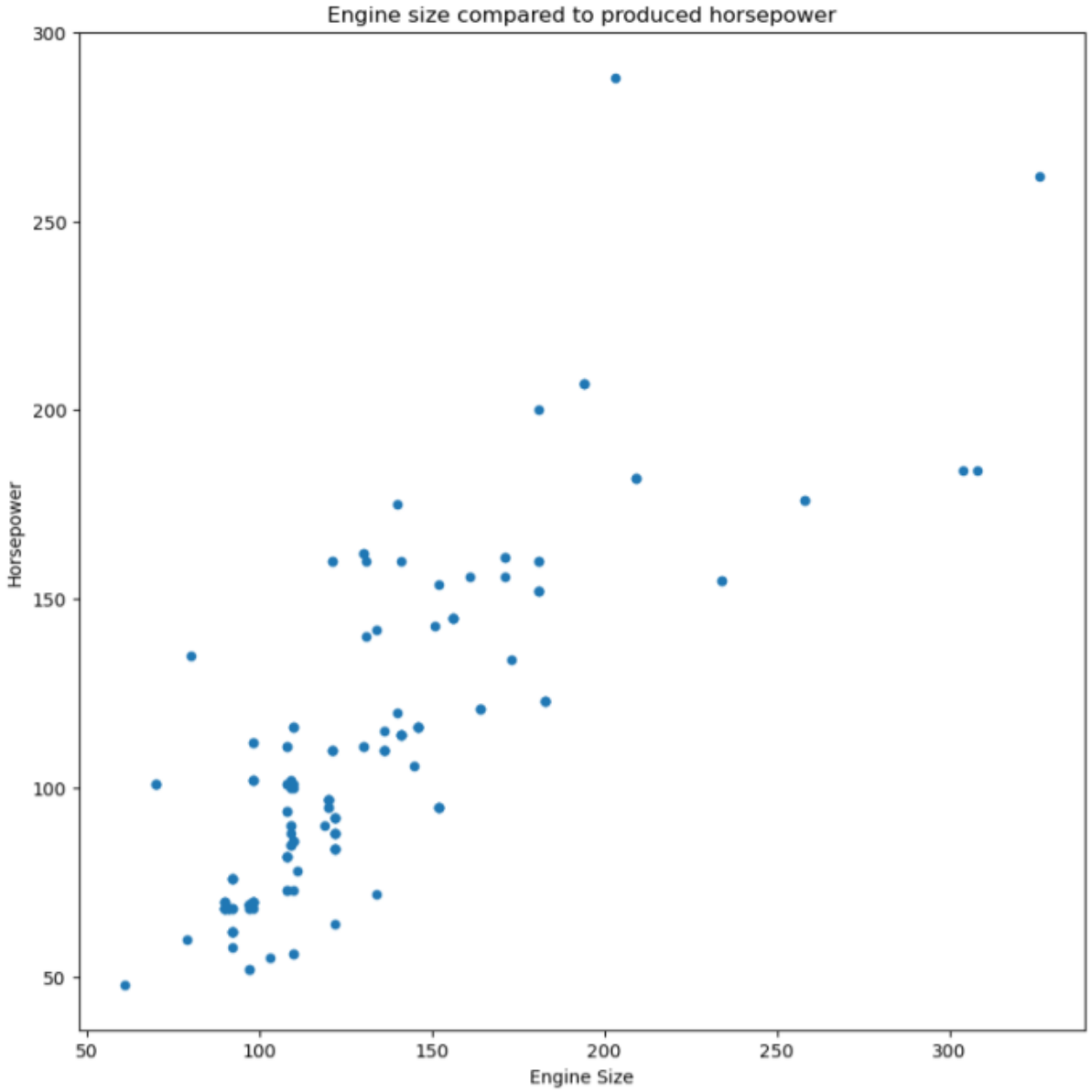

Fig 6: Scatter plot comparing an engines size and the horsepower it can produce

Moving on to look more in-depth at the types of engines and how they affect other aspects of the car. Figure 6 shows how the size of an engine influences the horsepower that that engine can produce. The trend looks quite linear. Therefore, the larger the engine, generally, it will be able to produce higher horsepower.

Another aspect of a car is how efficient it is. This is measured in how many miles it can travel before using up 1 gallon of fuel. However, this changes depending on the speed at which the car is going, the engine, or how often they have to move from a complete stop. For this reason, the dataset contains miles per gallon values both inside a city and on a highway.

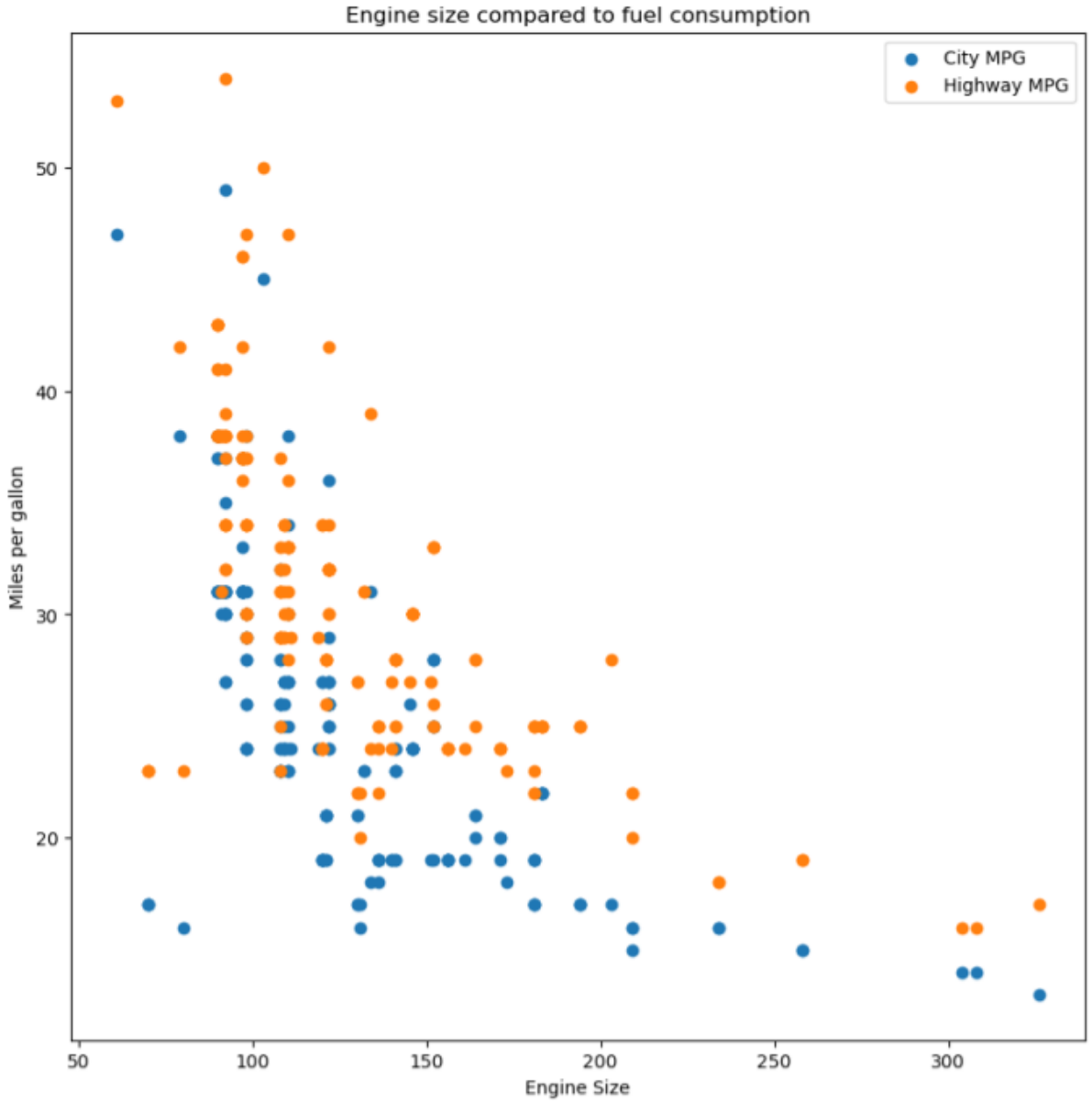

Fig 7: Scatter plot showing how the engine size affects the miles per gallon of the car

Figure 7 shows that, in general, engines can travel more miles per gallon while on the highway. This makes sense because, in a city, cars will have to pull away from a complete stop more often, which takes more power and therefore fuel, compared to driving at a constant rate for longer periods. This is because of momentum. The plot also shows that as the engine size increase, the efficiency quickly decreases, no

matter if the car is driving in a city or on a highway. This trend looks like an exponential decay

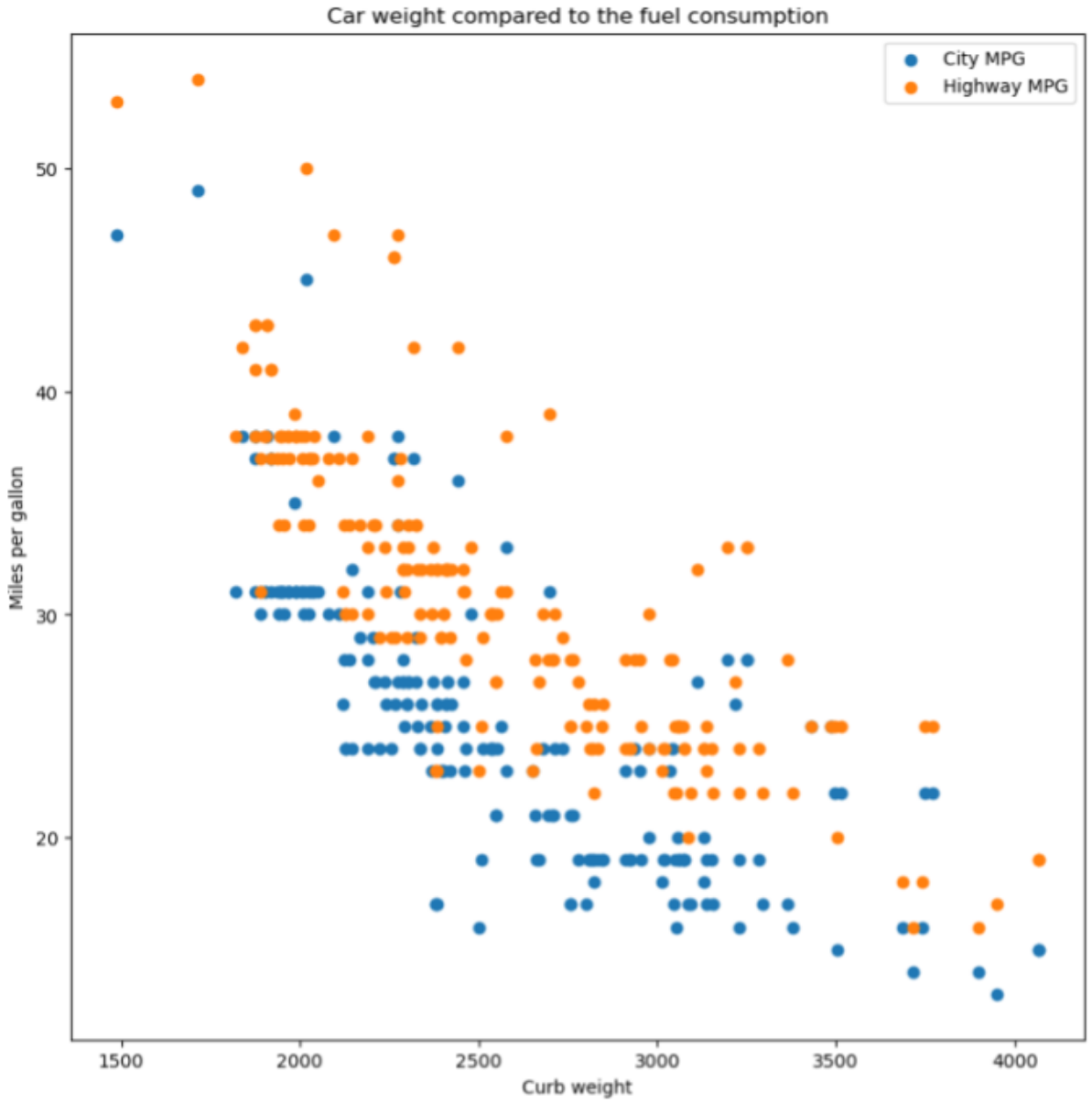

Fig 8: Scatter plot showing the affect a cars weight will have on its fuel efficiency

What else affects a car's efficiency apart from the size of the engine? To explore this question more, a comparison of a car's efficiency and weight is also checked and shown in figure 8. This plot shows a similar trend as in figure 7, however, it is much more linear instead of exponential. The heavier the car, the less fuel efficient it becomes. This makes sense because cars that have more mass will take more power to move from a stop or move in general, reducing efficiency. Though, as before, miles per gallon while on a highway is still better than in a city in most cases.

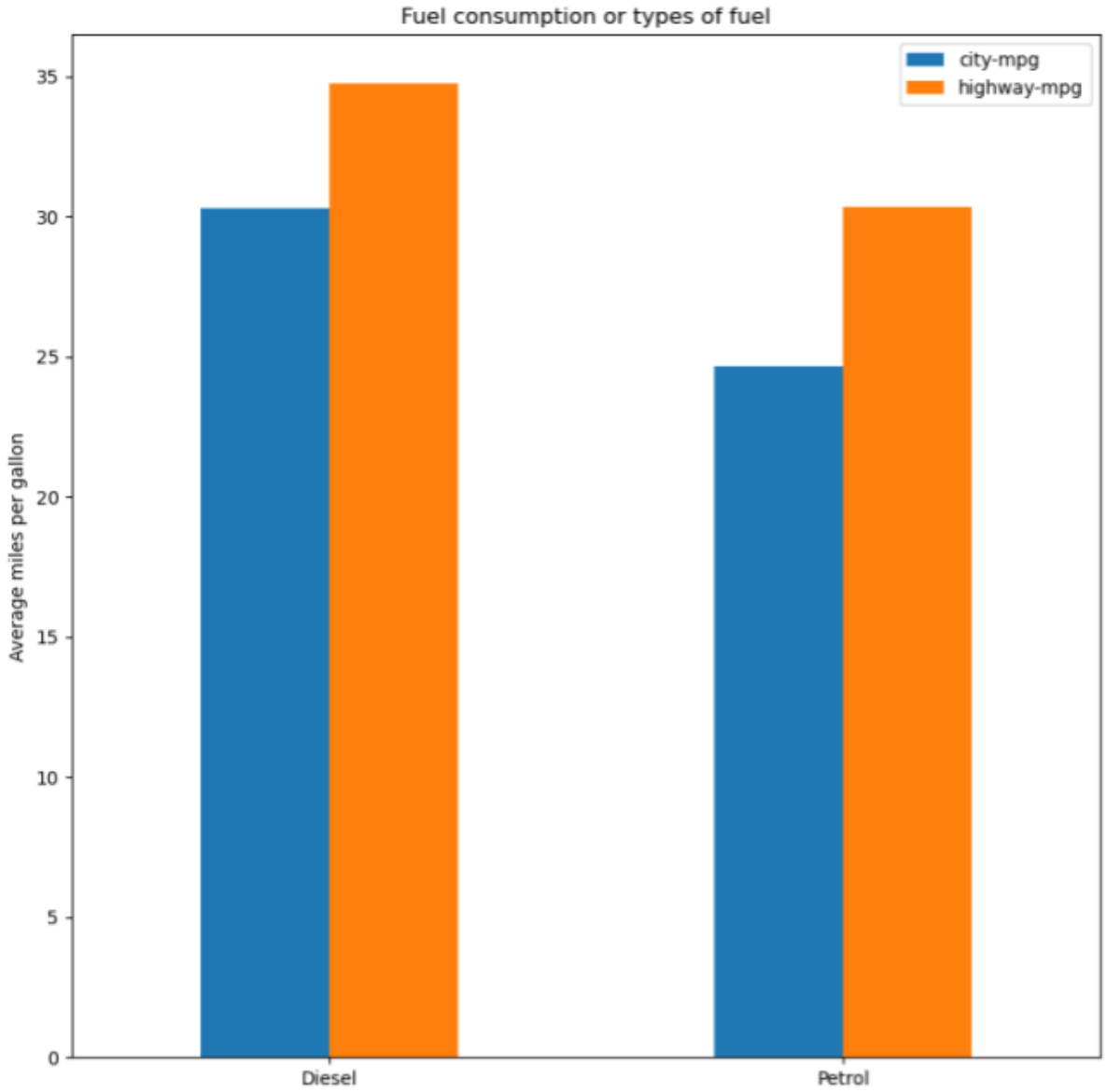

Fig 9: Bar chart showing average fuel consumption based on the fuel type.

Another variable that could affect a car's fuel consumption is the type of fuel that the car uses. Figure 9 shows that the type of fuel does, in fact, change the average miles per gallon of the car. Diesel cars are more efficient than petrol cars while in the city and on the highway. However, petrol cars on the highway are just as efficient as diesel cars in the city. The result of this plot is what is expected due to the fact that the selling point of diesel is that they consume fuel at a slower rate than petrol.

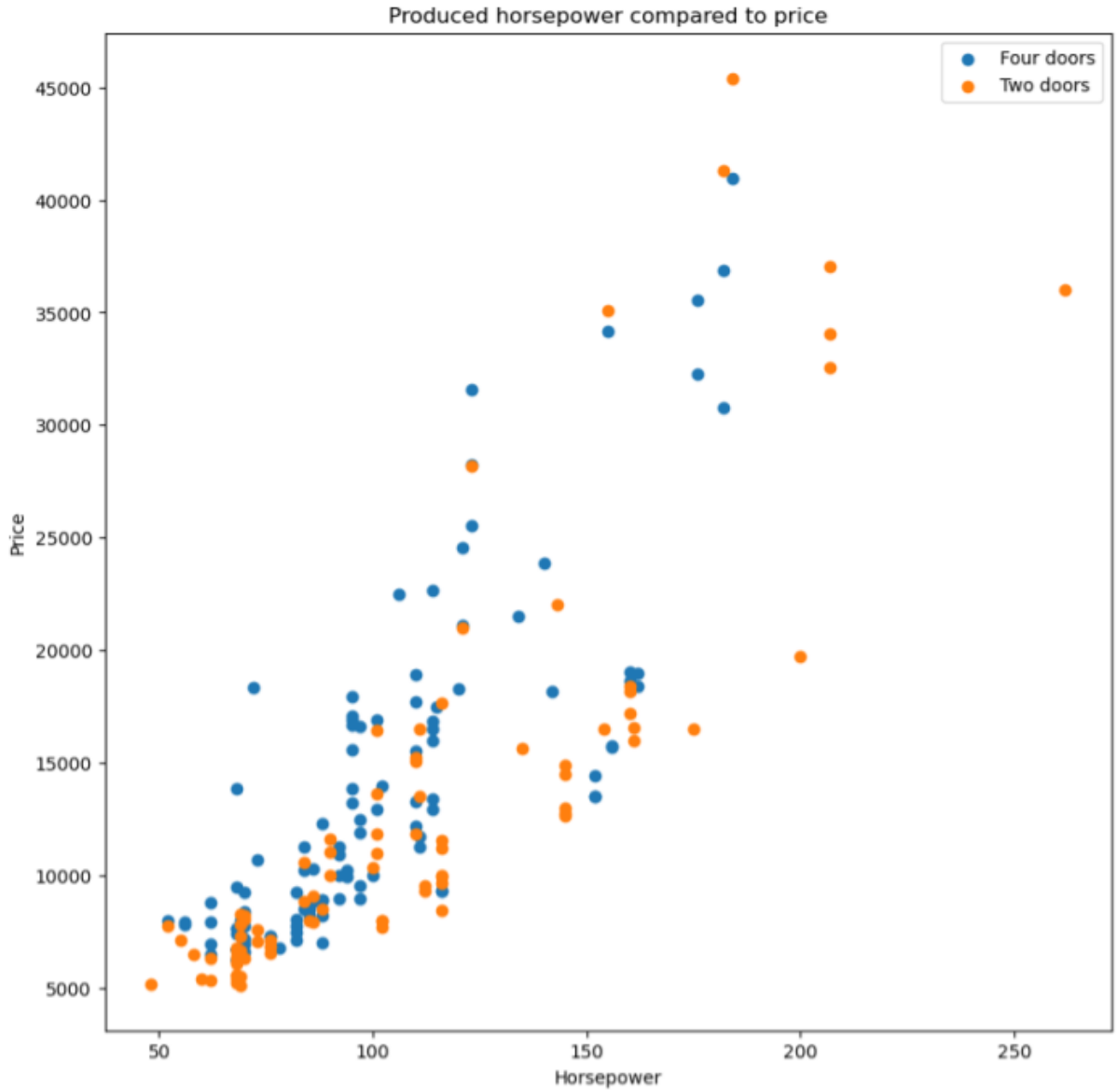

Fig 10: A scatter plot showing how the horsepower of the engine can influence the price. Cars are categorised by the number of doors as a proxy for their general size.

Moving onto how horsepower affects attributes of a car. Figure 10 shows how the horsepower of the engine can change the price of the car. I have separated the cars by the number of doors they have as a proxy for the size. In general, as the horsepower increase, so does the price of the car. However, four-door cars are more expensive than two-door. This makes sense since bigger cars normally cost more, and cars with more powerful engines also cost more to produce. The trend also looks quite linear. Though the cars with the most horsepower are 2 doors. This could be due to the fact that expensive supercars with lots of power normally try to minimise weight, therefore, opting for 2-door designs. The most expensive cars are also 2-door. This could be due to more high-end brands charging more for smaller cars because they can look better.

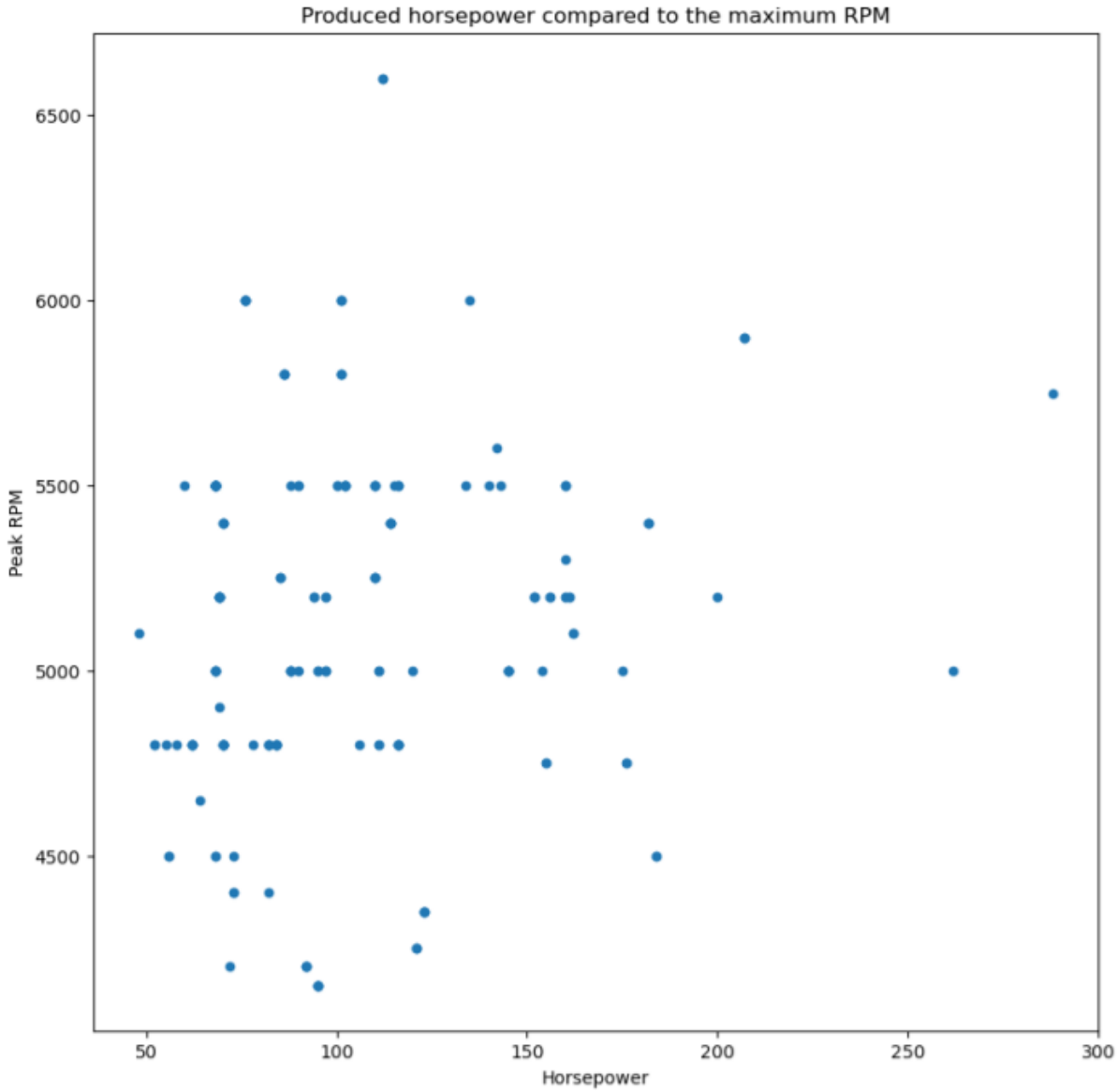

Fig 11: Scatter plot showing the relationship between an engines horsepower and maximum RPM

Finally, an obvious connection between horsepower and RPM was also looked at in figure 11. This figure seems to show that there actually isn't a massive relationship, which is unexpected.

Conclusions

Toyota take up a large proportion of entries in this dataset, they also have very average car prices. This may be connected as more models are made and therefore, more models are popular with the general public, making them more common.

A larger engine means that your car is more likely to have more horsepower, though it also means that the car will be less fuel efficient. To maximise the fuel

efficiency of the car, it must be light, with a small engine that uses diesel as a fuel. Most of the driving must also be done on the highway. Cities should be avoided.

To minimise the price of the car, purchase a 2 door Chevrolet with an engine with a lower horsepower.

THIS REPORT WAS WRITTEN BY: Conner Grice