Exploratory Data Analysis on the Automobile Dataset

Markus Saint-Pettersen

Introduction

The ‘automobile’ data set is a txt file with information about 205 different cars. This data set was collected in 1985 for the purpose of calculating insurance costs. The 26 columns of information can be divided into two different categories: the technical specifications of the vehicle (for example, engine size and number of doors), and details relevant to insurers such as the symboling (risk level) and the price.

Each row in the data set represents a different car, but some models have more than one entry. This means that there are some rows that all have the same technical specification and differ only on price.

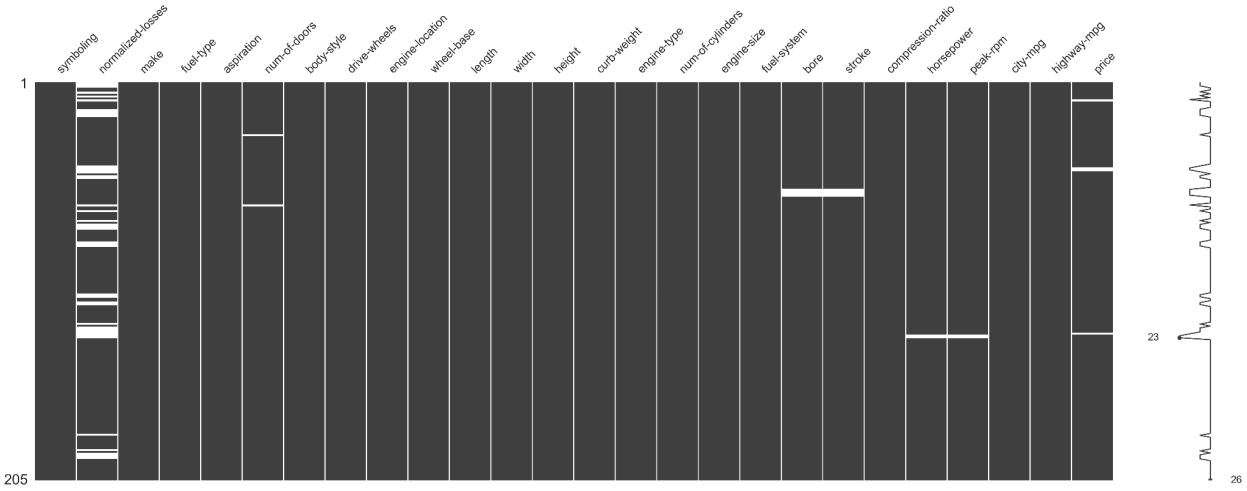
Data Cleaning

The initial data was in good condition and did not require much cleaning. Missing values were represented by the string ‘?’ and needed to be converted to ‘NaN’. Because of this, some columns were objects when they should have been floats or integers and needed to be converted. Finally, unnecessary columns (‘bore’ and ‘stroke’) were dropped to make the data set more manageable.

Additionally, a new column was created from ‘price’, price range, a measure of how expensive a car was relative to other cars in the data set.

Missing Data

From the whole data set, a total of 59 values were missing, making up 1.11% of the total. The majority of the missing data was from the ‘normalized-losses’ column, the average loss per car per year. The missing data are summarised below (with the column names):



‘normalized-losses’

41 values were missing from this column. There was not a good way to estimate the missing values and the column was not useful for analysis, so it was dropped completely.

‘num-of-doors’

The number of doors. Two values were missing from this column. Both values were missing from sedan type cars, so the mode for that body-style was substituted in (‘four’ doors).

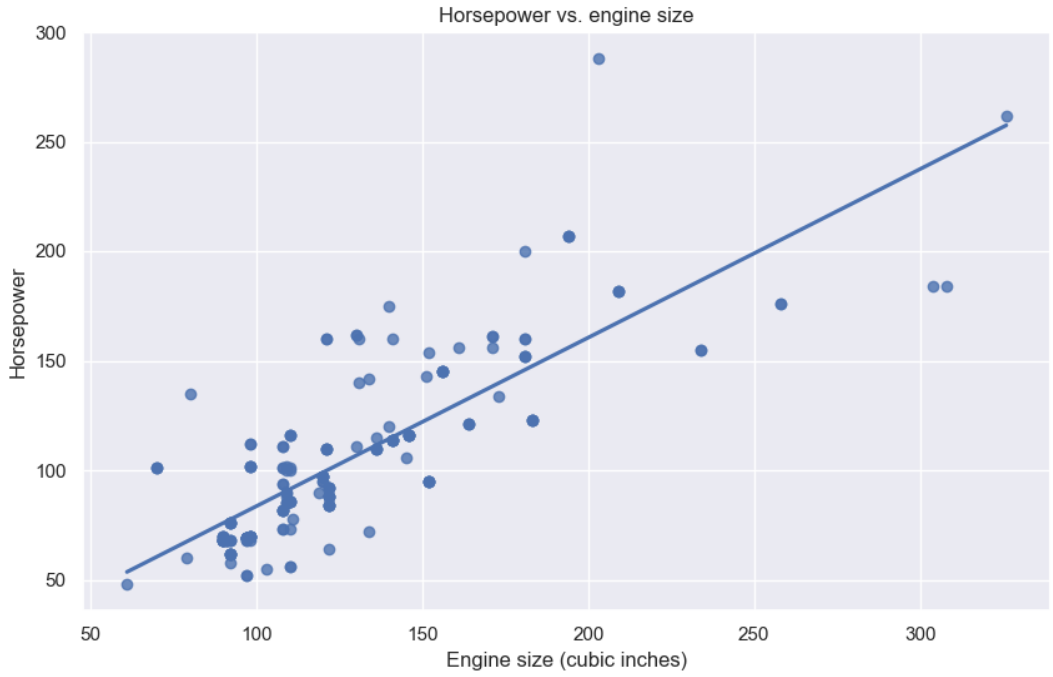
‘bore’ and ‘stroke’

Bore represents the diameter of a cylinder in the engine, and stroke represents the distance a piston travels within the cylinder. Four rows had both missing bore and stroke values. However, these values were missing not at random because the cars in question had rotary engines and so did not have any pistons. Ultimately, these columns were both dropped. Engine size was more relevant for this analysis and did not contain any missing values.

(I spent some time trying to replace these missing values which I left in the notebook, but in the end it made more sense to drop these columns)

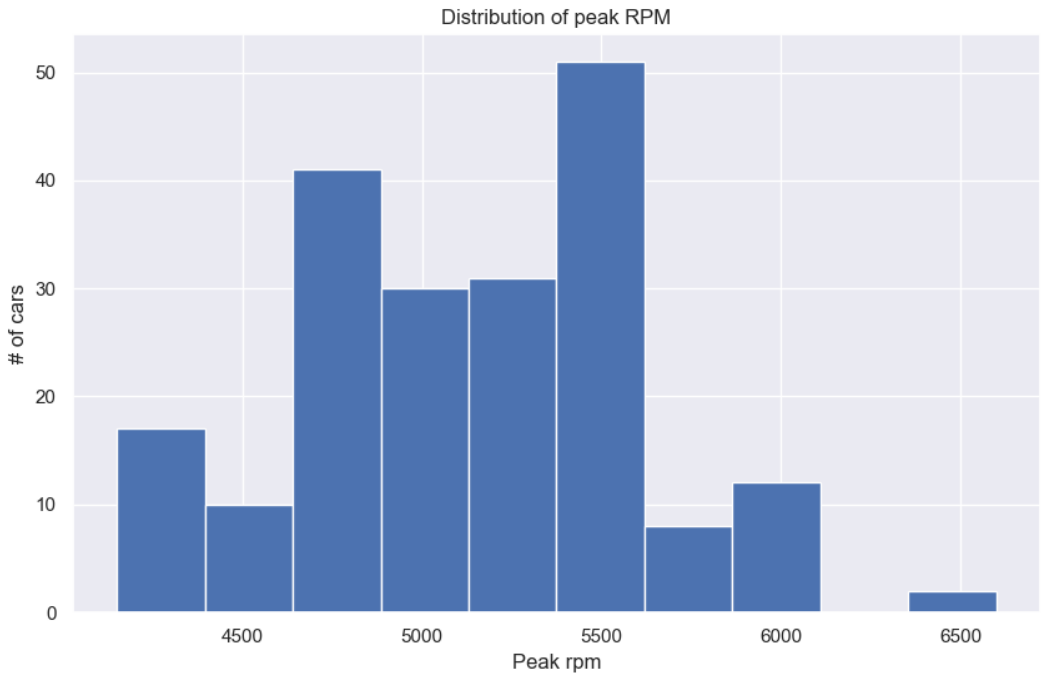
Horsepower

Two values were missing from the horsepower column. The correlation between engine size and horsepower is 0.811, so these values were estimated using linear regression as shown below:



Peak RPM

The same cars that were missing horsepower were also missing values for peak RPM. The distribution of the data for RPM was fairly normal with a mean of 5126 revs and a median of 5200 revs.



Therefore, the mean RPM could be substituted in for the missing values without drastically affecting the integrity of the data.

Price

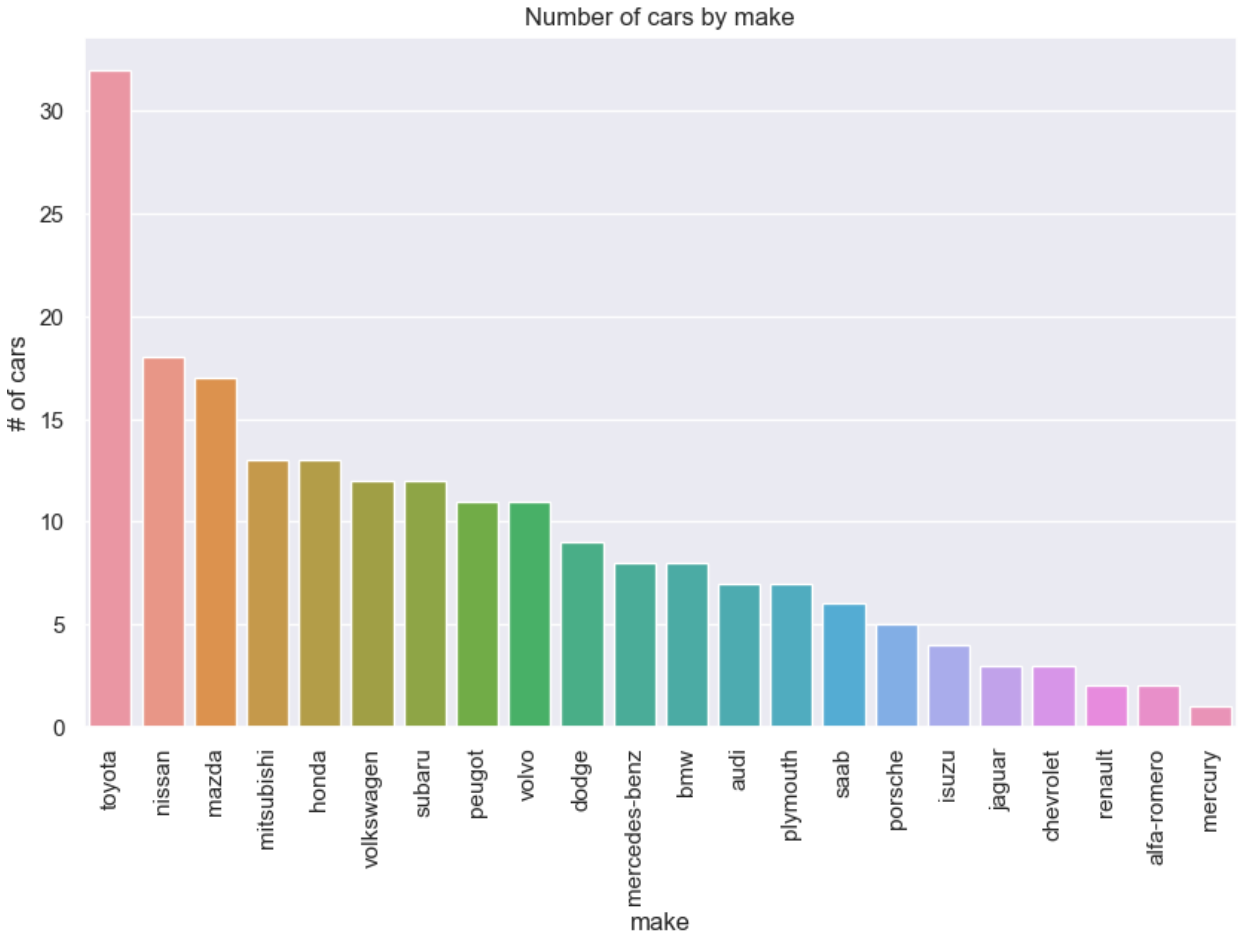
There were four rows in the data set with no price listed. There was no way to reliably fill in this information, so these rows were dropped from the analysis.

Data Stories and Visualisation

The main exploration of the data was focused on four main areas. a) price, b) drive wheels and c) risk level.

Manufacturers represented in the data set

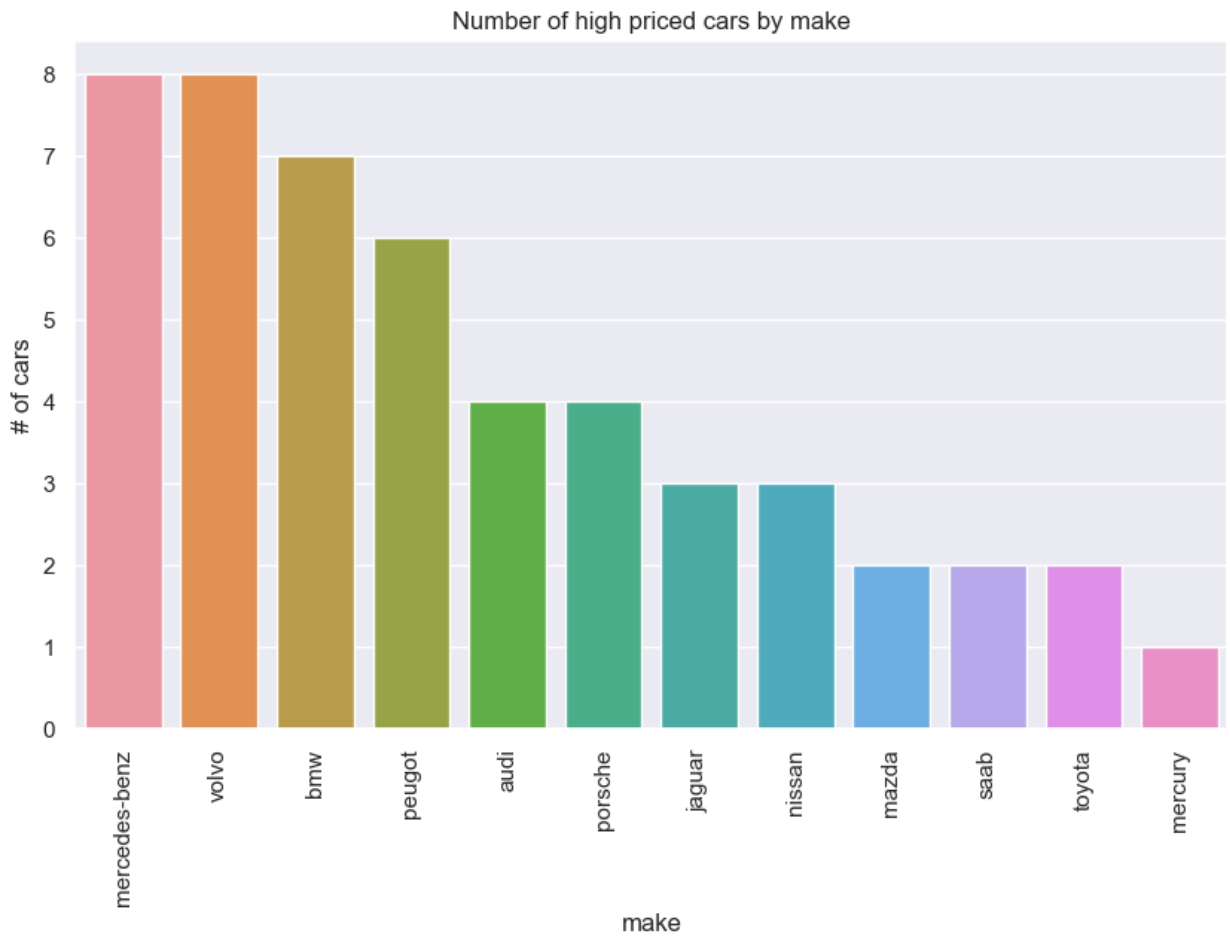
Initial analysis focused on categorical values to gain some insights into the data.



22 different manufacturers were present in the data. However, the data was not distributed evenly. Over 30 cars were Toyotas, but there was only one Mercury. This could be representative of the distribution of cars in the general population, but it will reflect the characteristics of Toyotas more than any other car manufacturer.

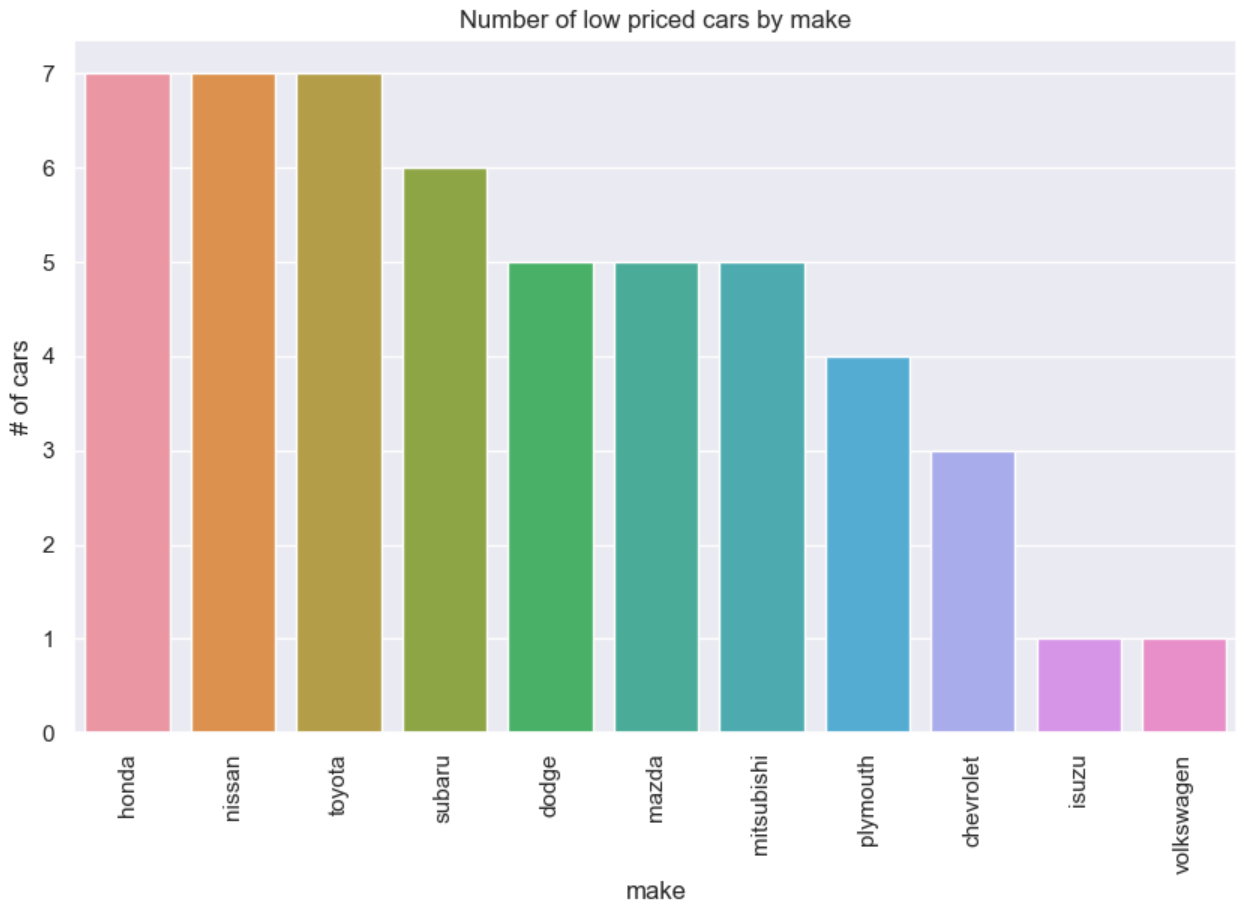
Manufacturers and price

Next, the same chart was produced for high priced cars. A high priced car was defined as being in the top quartile for price.



The most expensive cars were more evenly represented, as noted by fewer steep drops in the count plot. Out of the 22 manufacturers, only 12 produce high priced cars with Mercedes-Benz and Volvo producing the most. Comparing this chart with the previous chart, Mercedes-Benz, Porsche, Jaguar and Mercury exclusively produce high priced cars.

Next, we will examine the lowest priced cars, cars priced in the lowest quartile.



The distribution for the low price is more even, similar to the high price distribution. From the 22 total manufacturers, 11 are represented among the low priced cars. Mazda, Nissan and Toyota are represented on both extremes. This is not unusual because they are the three biggest makes in the data set. Chevrolet exclusively produces low priced cars.

To further examine these findings, the average prices for the top five most expensive car manufacturers are shown in the table below:

Top five manufacturers of the most expensive cars

Manufacturer	**Mean price**	**Standard deviation**
Jaguar	$34 600	$2 048
Mercedes-Benz	$33 647	$6 790
Porsche	$31 401	$6 529
BMW	$26 119	$9 264
Volvo	$18 063	$3 315

Jaguar produces the most expensive cars in the data set. The cars are also fairly consistently priced based on the small standard deviation. On the other hand, the price of BMWs can vary drastically. The top four are quite close, but Volvo is far lower.

Next, the average prices for the top five cheapest car manufacturers:

Top five manufacturers of the cheapest cars

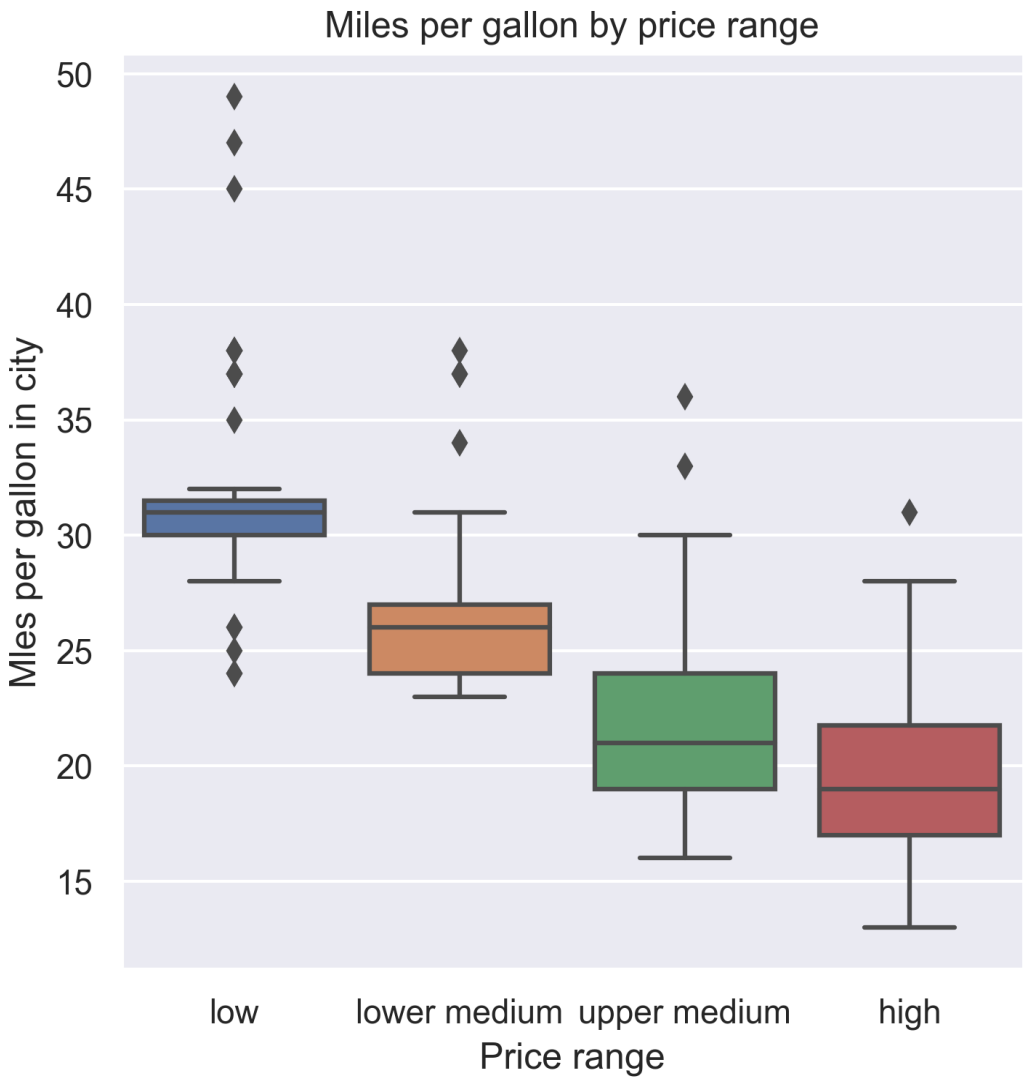
Manufacturer	**Mean price**	**Standard deviation**
Chevrolet	$6 007	$754
Dodge	$7 875	$2 213
Plymouth	$7 963	$2 396
Honda	$8 185	$2 062
Subaru	$8 541	$1 940

Chevrolet produces the cheapest cars in the data set by a big margin. Furthermore, the prices are consistently low based on the standard deviation. This is consistent with the fact that all Chevrolets in the data set are priced in the lowest quartile.

Now that we know which manufacturers produce the cheapest and most expensive cars, we can examine how they differ.

Price and MPG (in city)

Below shows the box plots for miles per gallon based on the price range of the vehicle. From the chart, we can see that the most fuel efficient vehicles are the lowest priced, and as the price range increases, the efficiency decreases. This is not intuitive and there might be a lurking variable present.



Price and engine size

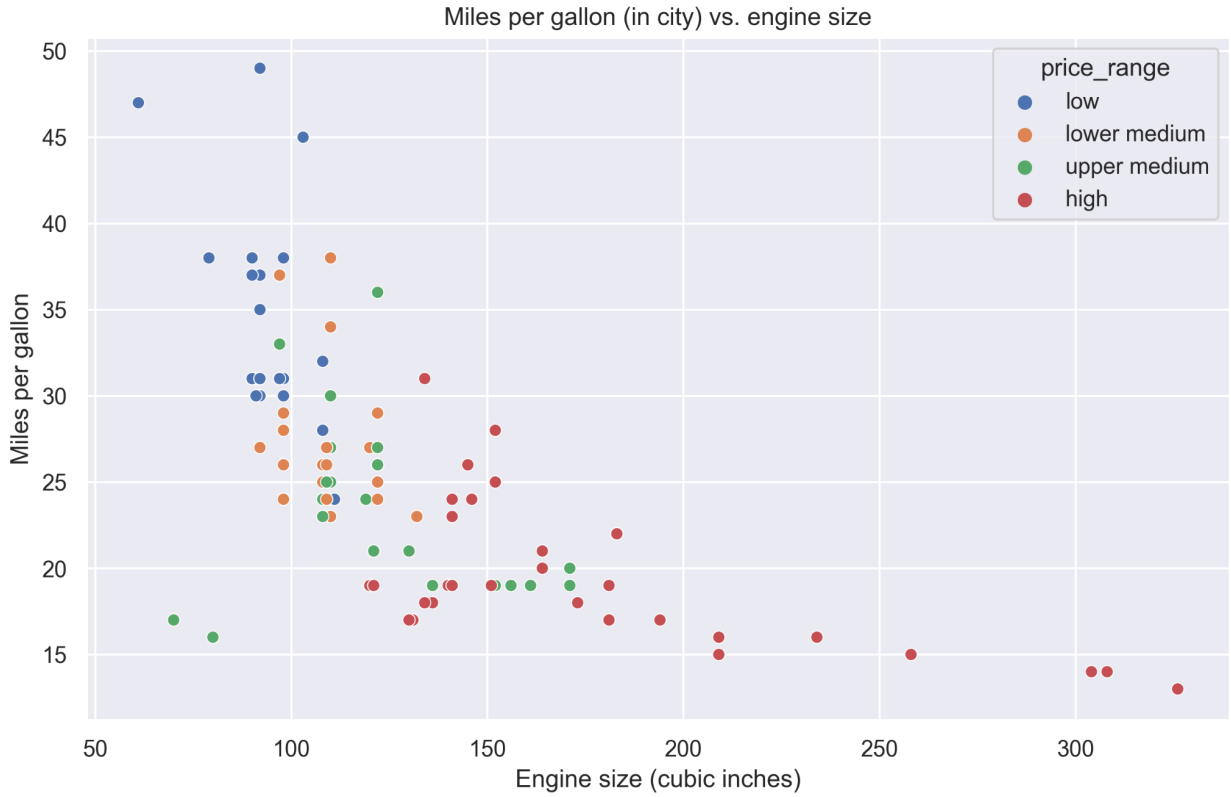
Next, we can examine the link between the price of a car and the size of the engine.

The above graph shows a clear correlation between price and engine size (0.873). The more expensive a car is, the bigger the engine it will have.

The above chart also shows the distribution of price. The ranges for price get bigger as the price increases. Low and lower medium priced cars all have fairly similar engine sizes (mostly between 80 and 150 cubic inches), but high priced cars can show a lot of variation (from around 105 to well over 300 cubic inches).

MPG and engine size

We have seen that price affects both MPG and engine size, but how does engine size affect MPG?

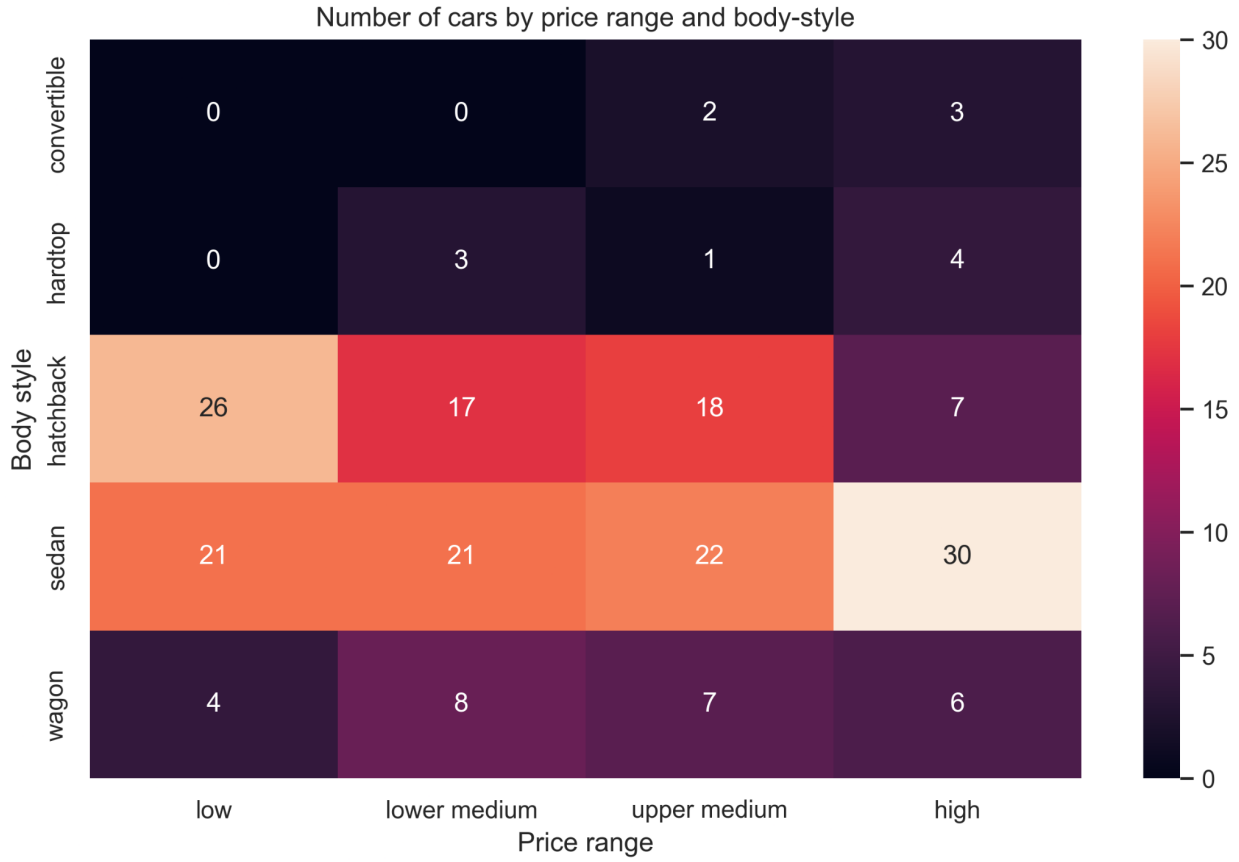


The above graph shows that as engine size increases, the fuel efficiency decreases. The more likely explanation is that more expensive cars have bigger engines, and bigger engines are less efficient. The colour of the dots also helps to see this relationship. The dots in the top left quadrant (smallest engines, most efficient) are mostly blue, while the bottom right quadrant (biggest engines, least efficient) are all red.

Body style and price

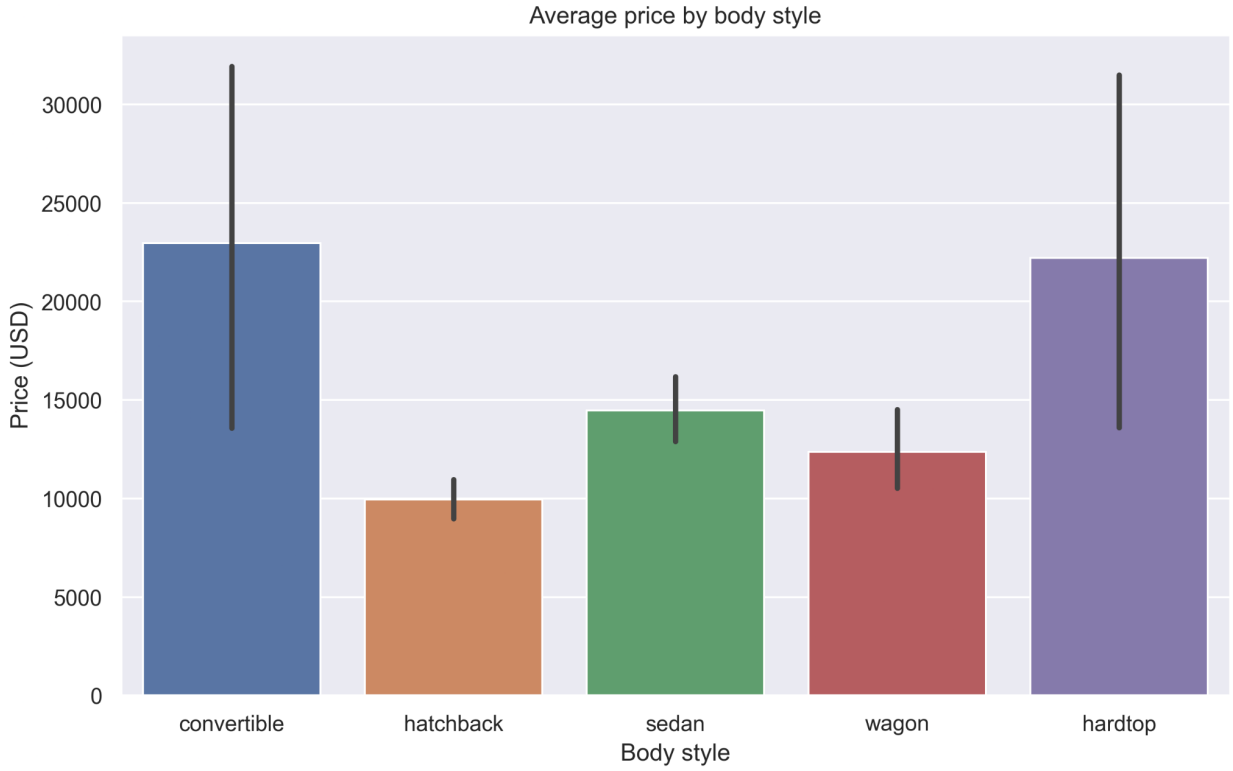
Is there a connection between the price of a car and the body style?

The heat map below shows the number of cars for each combination of price range and body style.



There are not many convertibles, hardtops or wagons included in the data set. Focusing on hatchbacks and sedans, we can see that the hatchback is a favoured design among lower priced cars while sedan is more common among higher priced cars. A chi-squared test can confirm this.

What is the average price for each body style?

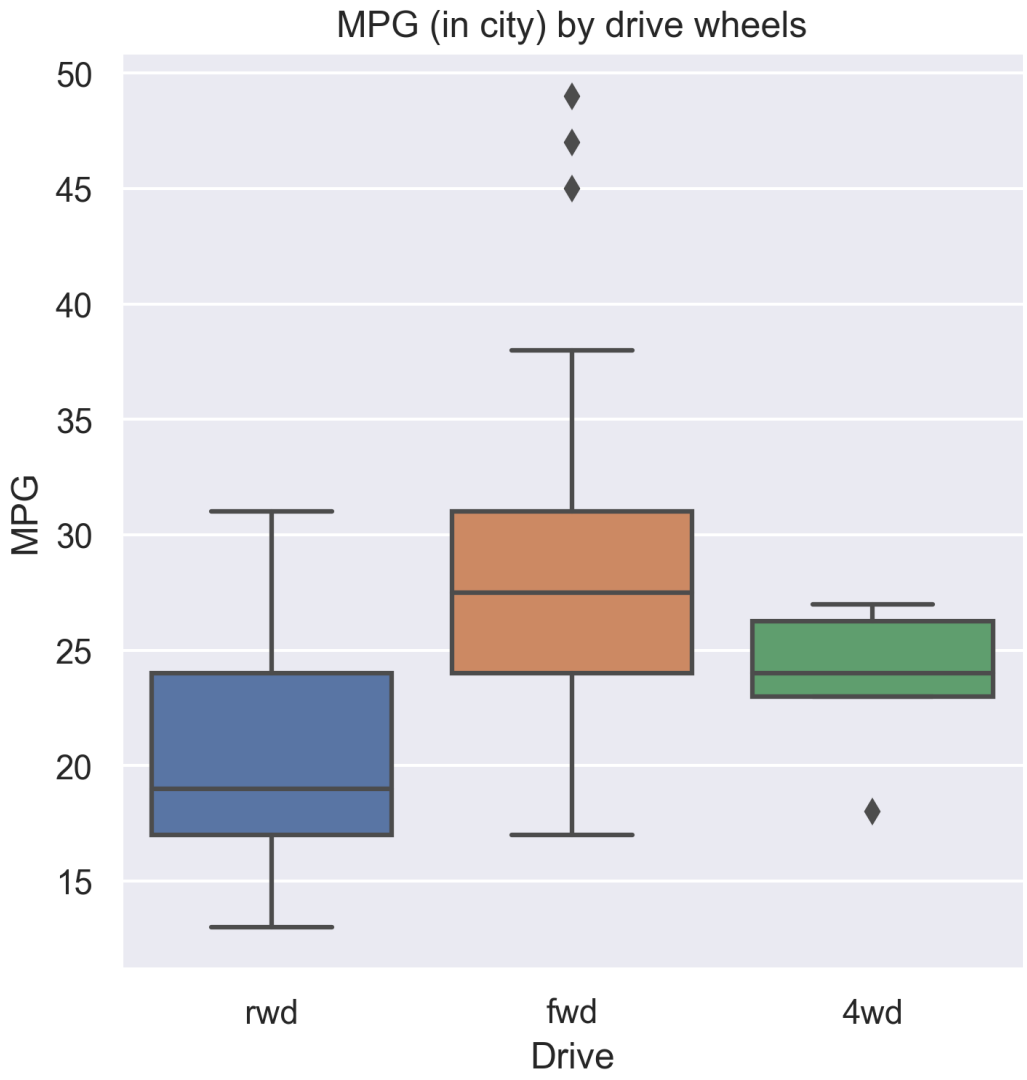


The above chart shows that convertibles and hardtops are on average the most expensive cars. However, this conclusion is based on a small number of results that vary quite a lot (as shown by the large bars). Focusing on hatchback and sedan, we can see a clear difference in average price between the two.

Based on this analysis, expensive cars tend to be sedans, hardtops or convertibles with large engines, high horsepower and poor fuel efficiency, while lower priced cars are hatchbacks with smaller more efficient engines.

Fuel efficiency and wheel drive

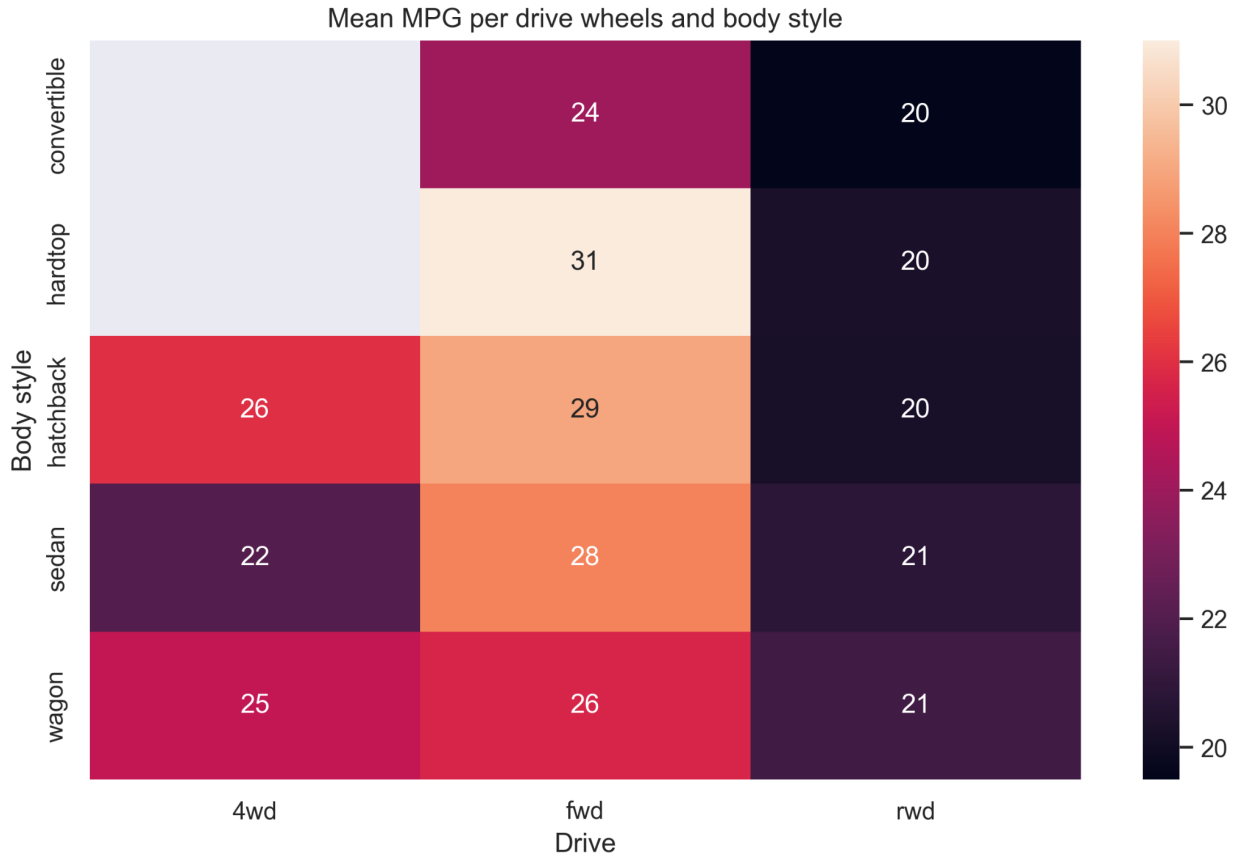
We can also look at other variables in the data set such as wheel drive.



This box plot shows that the most efficient drive is a front wheel drive, and rear wheel drive is the least efficient. In fact, the lowest quartile of front wheel drive cars perform as well as the upper quartile of rear wheel drive cars, that is, 75% of front wheel drive cars perform better than 75% of rear wheel drive cars.

Body style and drive

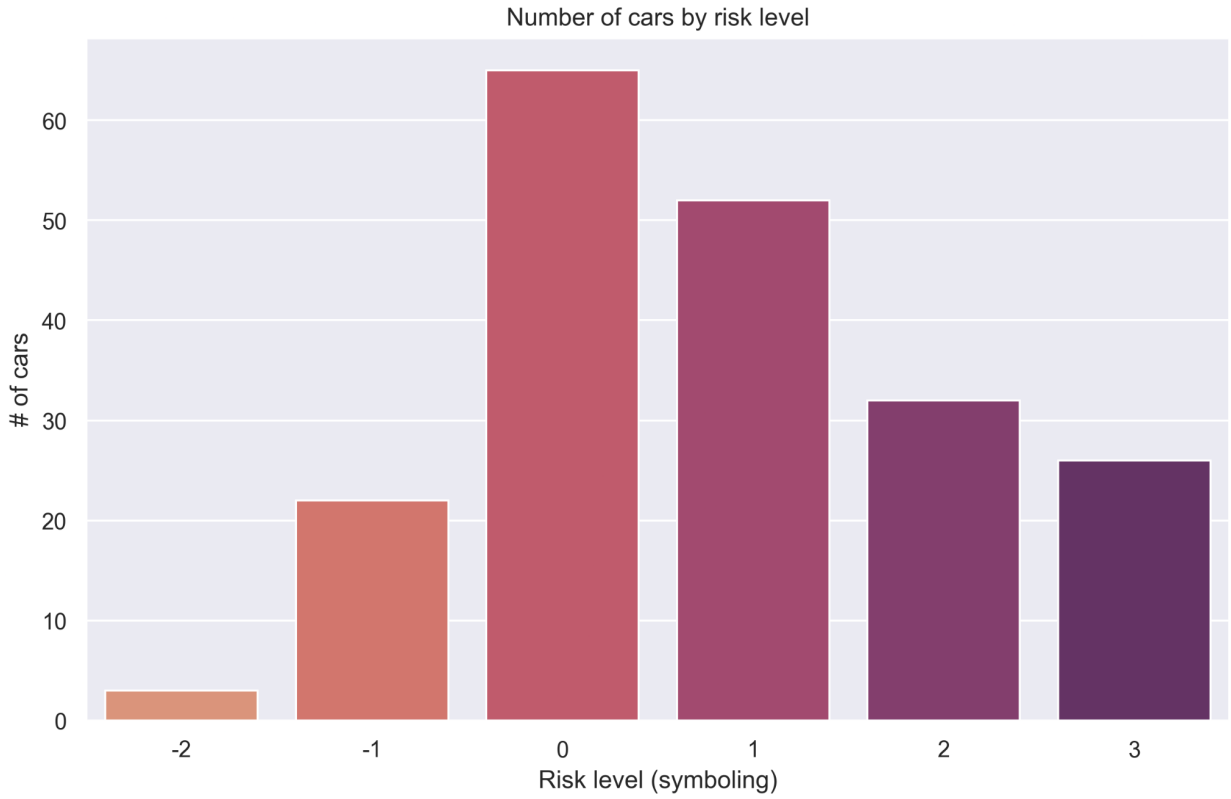
How is the fuel efficiency influenced by the body style?



Based on the heatmap, the most fuel efficient car is a hardtop with front wheel drive. Again, this analysis suffers from a low number of convertibles, hardtops and wagons in the data set. Looking at the data for hatchback and sedan, we can see the same results as before. Front wheel drive cars are the most efficient and rear wheel drive cars the least.

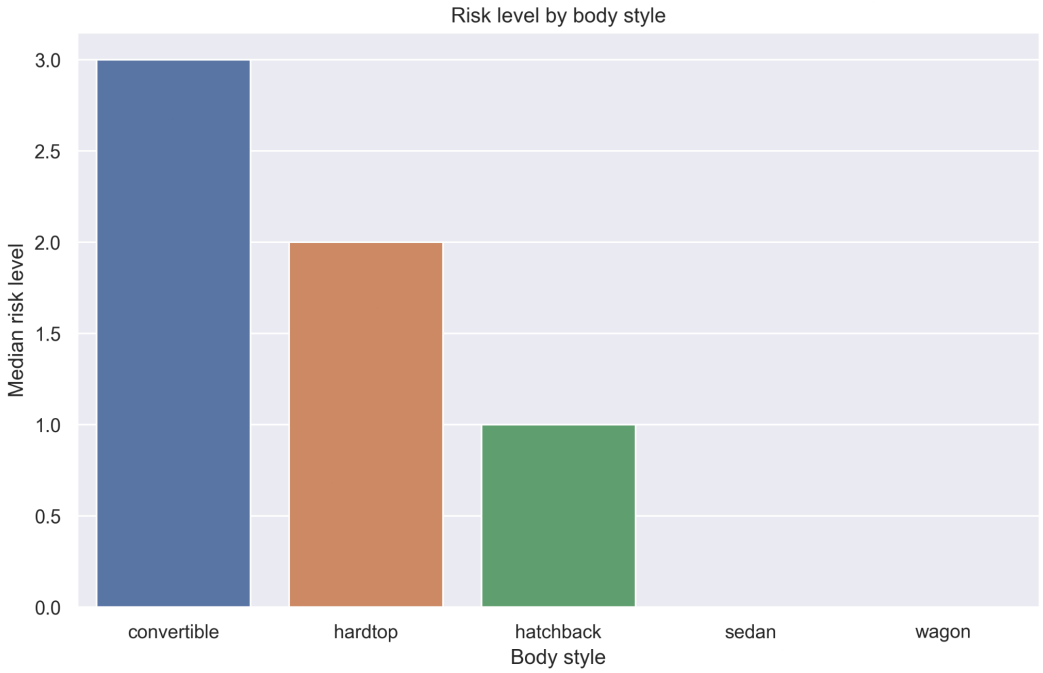
Symbolling

Finally, we can look at the symbolling column. Symbolling refers to how risky a car is to an insurer. A rating of 3 means the car is very risky and a rating of -3 means that it is very safe.

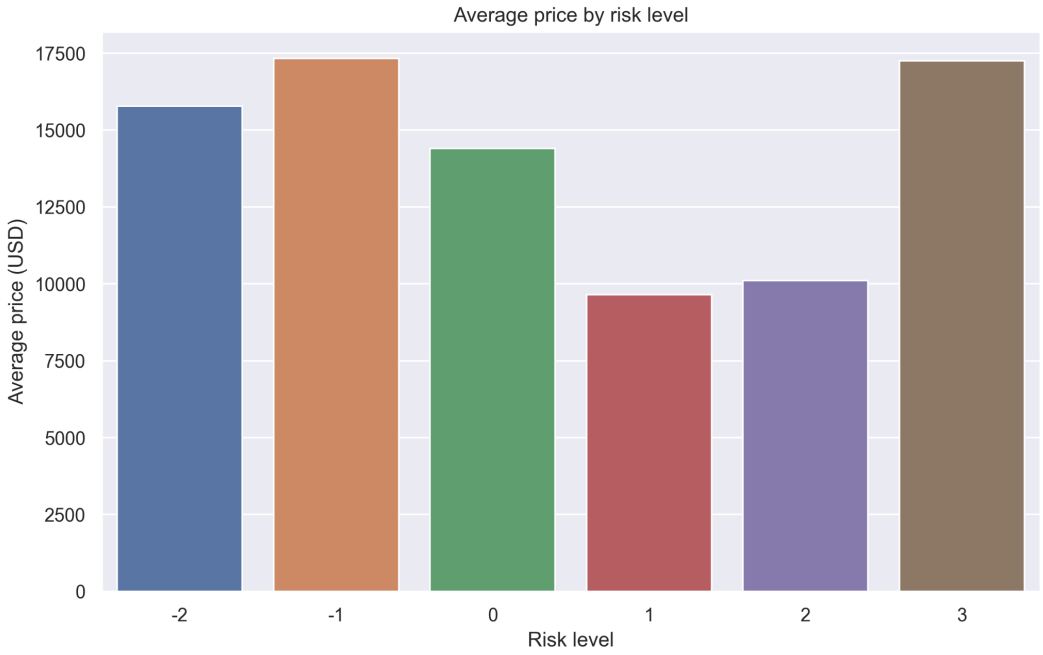


The above chart shows the number of cars for each risk level. No cars have the safest risk level (-3). Overall, the data is skewed to the right suggesting that these cars are more dangerous on average.

What type of cars are more dangerous for insurers?



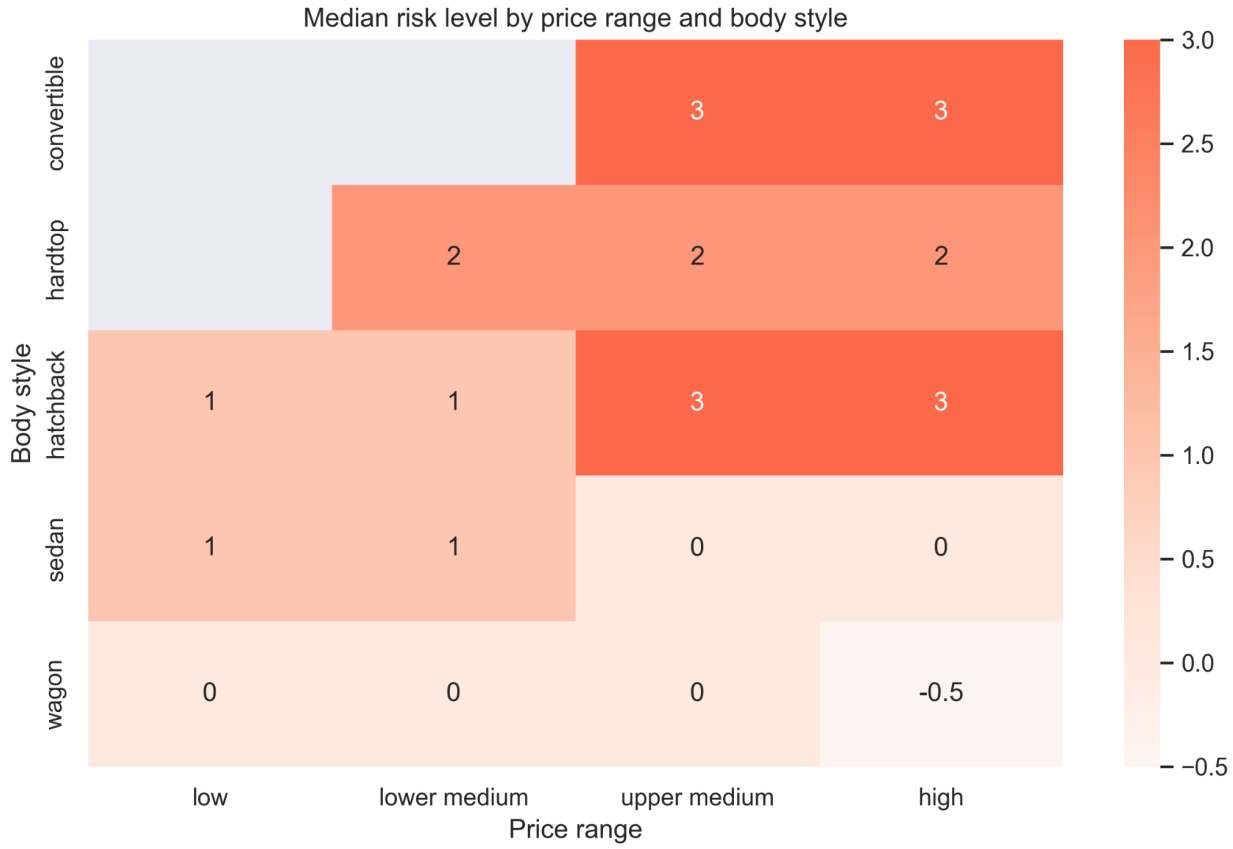
The riskiest cars are convertibles and hardtops, and the safest are sedans and wagons. However, price might be influencing the risk level. Are the most risky cars for insurance companies also the most expensive?



From the chart above, we can see that there is not really a pattern between price and risk level. The most expensive cars are risk levels 3 and -1. Further analysis is necessary to determine the relationship. A few

very expensive cars could be affecting the averages, particularly because some groups have very few entries (only three cars in the '-2' level).

Finally, we can combine the two factors, price and body style, to see how it affects the symbolling:



The heat map shows that body style has more of an influence than price range for risk. Convertibles and hardtops have the most risk and wagons have the least risk. Hatchbacks are low risk in general, unless they are higher priced. Sedans are also low risk, but cheaper sedans have more risk than more expensive sedans. However, this analysis suffers from small samples for some groups. There are very few convertibles, hardtops and wagons in the dataset. More data is needed to make strong conclusions.