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

# Introduction

About the data: This dataset is 205 records long and contains 26 variables. Each record pertains to a single model of car. Each variable contains unique information about this model of vehicle and will be valuable to an insurance company looking to predict the risk of insuring various models of car.

## Variables

Symboling, Normalized-losses, Make, Fuel-type, Aspiration, Num-of-doors, Body-style, Drive-wheels, Engine-location, Wheel-base, Length, Width, Height, Curb-weight, Engine-type, Num-of-cylinders, Engine-size, Fuel-system, Bore, Stroke, Compression-ratio, Horsepower, Peak-rpm, City-mpg, Highway-mpg, Price

## Data ranges

Symboling: -3, -2, -1, 0, 1, 2, 3

Normalized losses: 65 to 256 (continuous)

Make: Alfa-Romero, Audi, BMW, Chevrolet, Dodge, Honda, Isuzu,

Jaguar, Mazda, Mercedes-Benz, Mercury, Mitsubishi, Nissan, Peugeot, Plymouth, Porsche, Renault, Saab, Subaru, Toyota, Volkswagen, Volvo

Num-of-doors: 2 or 4

Body-style: hardtop, wagon, convertible, hatchback, sedan

Drive-wheels: 4wd, fwd, rwd

Engine-location: front, rear

Wheel-base: 86.6 to 120.9 (continuous)

Length: 141.1 to 208.1 (continuous)

Width: 60.3 to 72.3 (continuous)

Height: 47.8 to 59.8 (continuous)  
Curb-height: 1488 to 4066 (continuous)  
Engine-type: dohc, dohcv, I, ohc, ohcf, ohcv, rotor  
Fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi  
Bore: 2.54 to 3.94 (continuous)  
Stroke: 2.07 to 4.17 (continuous)  
Compression-ratio: 7 to 23 (continuous)  
Horsepower: 48 to 288 (continuous)  
Peak-rpm: 4150 to 6600 (continuous)  
City-mpg: 13 to 49 (continuous)  
Highway-mpg: 16 to 54 (continuous)  
Price: 5118 to 45400 (continuous)

# Missing Data & Data cleaning

While this dataset had 0 null values, many of the values were

essentially null. Several columns had values of ‘?’ in place of NaN values, including 40% of Normalized-losses values. First I ran code to search for all of these instances and replace them with true NULLvalues.

The breakdown of missing values is as follows:

normalized-losses: 41

num-of-doors: 2

bore: 4

stroke: 4

horsepower: 2

peak-rpm: 2

price: 4

\*out of 205 records

I then made these categories usable by replacing these with the meanvalue for the category. The data type of each column was then updated to match the true data type of values in that column.

Negligible missing values remain.

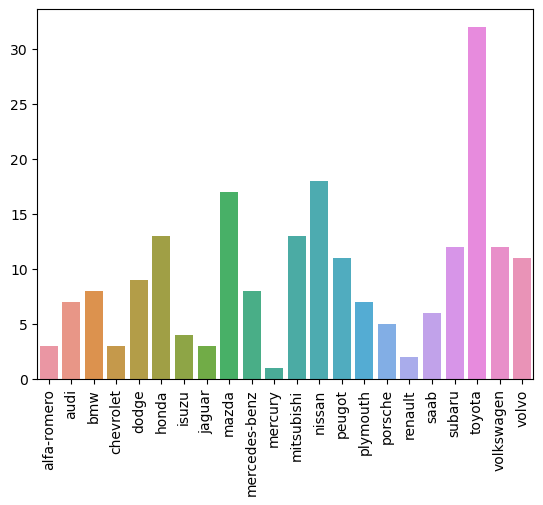
# Data Stories and Visualizations

### Univariate Analysis:

### Bivariate Analysis:

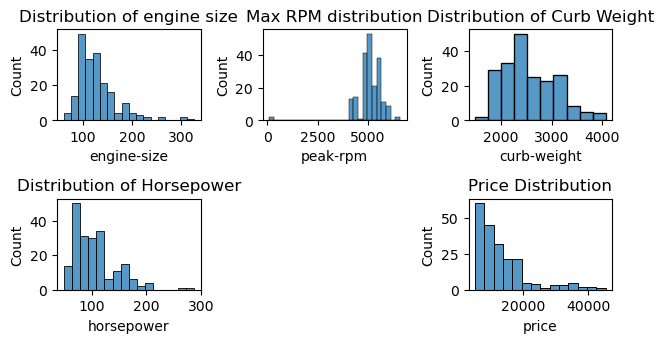
First, it was logical to look for the frequency/popularity of each make of car in the dataset.

This would give us a good grasp of the overall dataset with regards to average price, size, mpg etc. (e.g. Toyota are known for making these economical, practical cars, with so many Toyota cars present in the dataset it is unlikely this data is going to skew heavily to supercars).



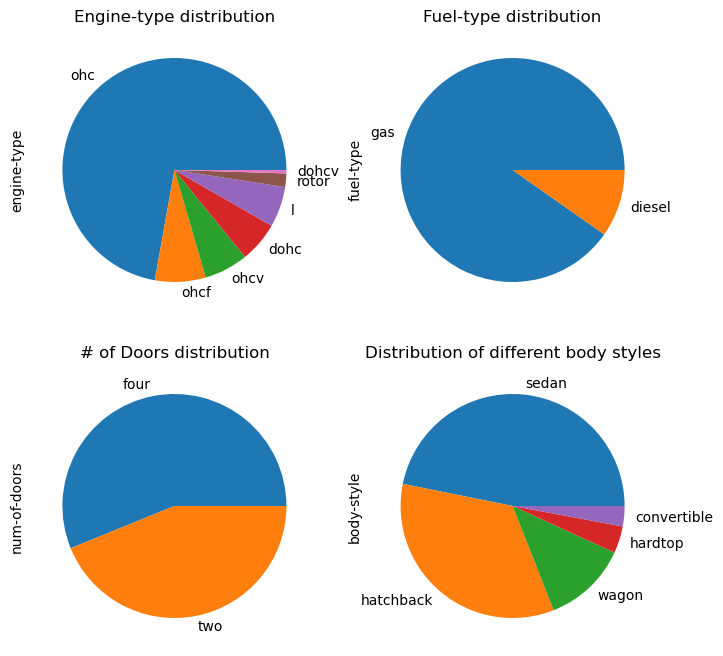
Findings: Toyota's are far and away the most common make of car in the dataset. And the top 7 are all Japanese manufacturers except for Volkswagen.

Next, I looked at the distribution of variable factors.



From these is it possible to understand that the dataset is very practical and likely reflected the real-world distribution of vehicles well. Most cars are cheap, low horsepower, smaller cars. Which is accurate for the real world. However, there is a data present for all price points and a variety of horsepower. From these we can also see that they are all correlated to some degree. Price/horsepower/engine size all match well, but peak-rpm and curb-weight also follow similar distribution curves.

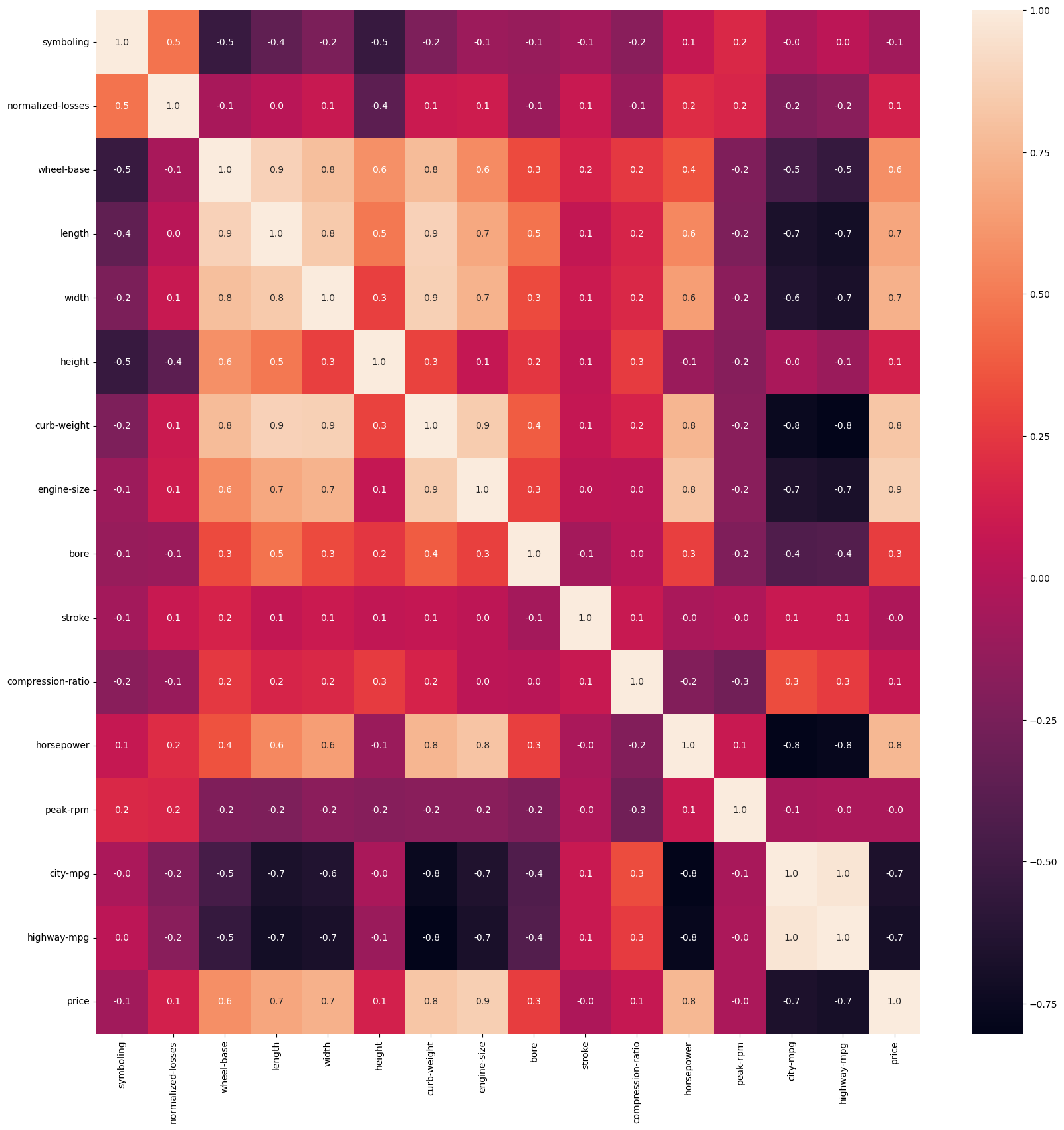
Most cars have smaller engines, with the majority between 100-140.  
Low 5000s is the most common peak-rpm value and most cars have values between 5000-6000.   
Modal curb weight is 2250-2500, most fall between 1750-2500.   
Most cars have low horsepower, between 70-120.  
The price for most cars is less than 15000 and cars above this price are evenly distributed.



These graphs continue to show the frequency of various categories. We can see from here that ohc engines are practically the norm for cars as are gas-powered engines. Sedans and hatchbacks make up the majority of vehicles and there is a mostly even split between four and two door cars.

After looking at various single categories and how they are distributed throughout the dataset, I wanted to get a bigger view of the data set as a whole and how it all correlates to each other. Especially, how does symboling correlate with the other features.

A heatmap was produced showing the correlations between each pair of categories.



In this heatmap, each cell marks two columns of the dataset. The value inside the cell (between -1 and 1) represents how correlated those two columns are. From this we can see that Horsepower is correlated to curb weight, engine size width and length. This follows logically as a bigger/heavier car will need more power to move it. It is heavily negatively correlated with both mpg columns, which also follows as bigger engines tend to use fuel more aggressively.

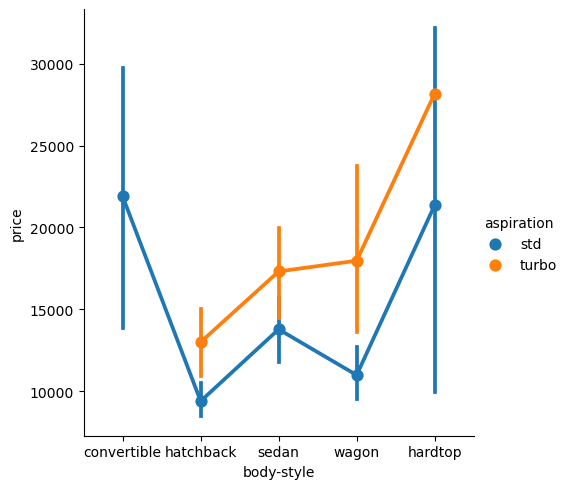
Another interesting correlation is Price with the columns that describe the size of the car. As well as the horsepower and engine-size. The bigger, more powerful your car is, the more it will cost. While small compact, fuel-efficient cars will cost less.

Interestingly, Length, Height, Wheelbase are the most important factors in reducing the symboling for the car. This means that smaller, shorter cars are lower claim risks than other cars that are otherwise similar. This also follows are your car is less likely to be damaged if it is physically smaller.

### Bivariate Analysis:

After getting an overview of the general data and how individual features are distributed, it was time to see what could be learnt from combining certain variables together and visualizing them.

Firstly, this plot shows the price points of each style of car as well as the price points for those available in both standard and turbo models.



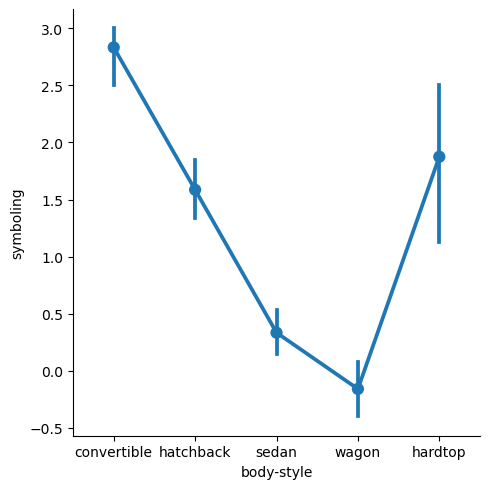
Findings:

Cars with standard aspiration are consistently cheaper than those with turbo by about 20-30%.

Convertible and Hardtops are more expensive than the other styles, which also aligned with the popularity of these styles too.

Cheaper body styles are also the more common/popular ones.

This next graph shows the body types against the average symbolling of that type of car.



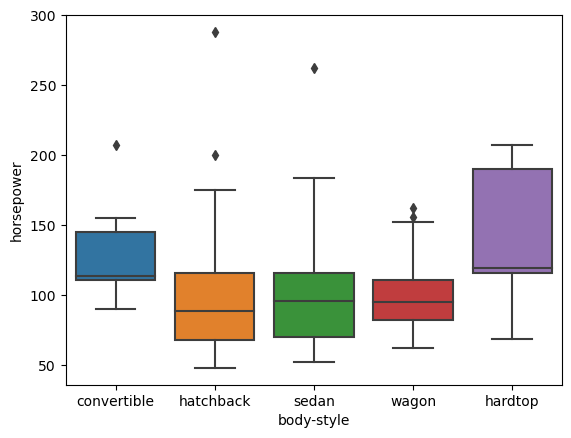
Findings:

Convertibles are an insurance risk as their average symbolling is around 2.9 out of a possible 3.

Next most risky is the hardtop with a value of around 1.9 on average.

Wagons are the only style of car that perform better than expected for their price as their average is around -0.1.

These cars that scored highly in this graph do have an image of being fast and powerful and this could be the reason for this rating.



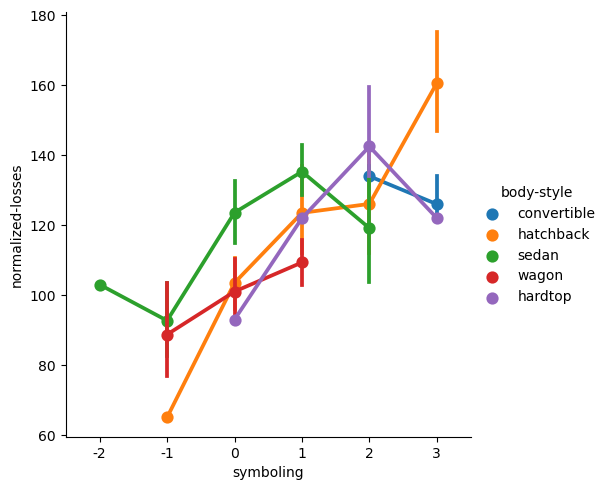
Findings:

Surprisingly, convertibles only have a slightly higher horsepower than hatchbacks or sedans, but a wildly higher symbolling. Interestingly though hardtops, which had the second highest symbolling also has higher horsepower than these other types of car.

While it is not likely to be the only cause, Convertibles and hardtops do have higher horsepower than other styles of car, and so this is likely to contribute to their high symbolling score.

However, convertibles were far higher than hardtops of hatchbacks but only have marginally higher horsepower. Also, hardtops have a higher horsepower but lower symboling. This could be due to their additional weight and size making them no quicker or more dangerous than convertibles despite the additional power.

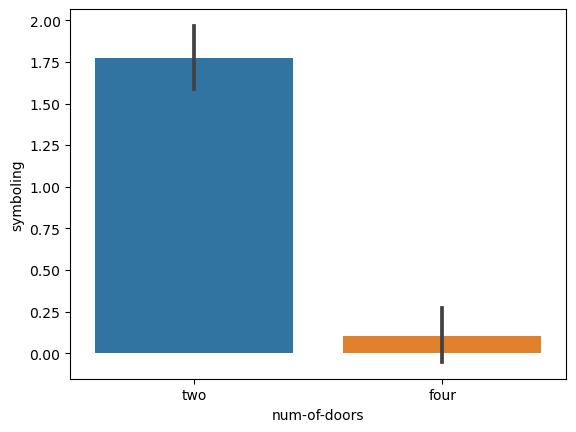
If we also considering the price of each of the different styles, convertibles are only narrowly more expensive. This would also contribute slightly to the higher symboling but likely not entirely.



This graph combines the information from earlier with normalized-losses, a representation of how expensive each claim is. We can see that hatchbacks vary wildly on the normalized losses, being both the cheapest and most expensive to repair in event of an accident.

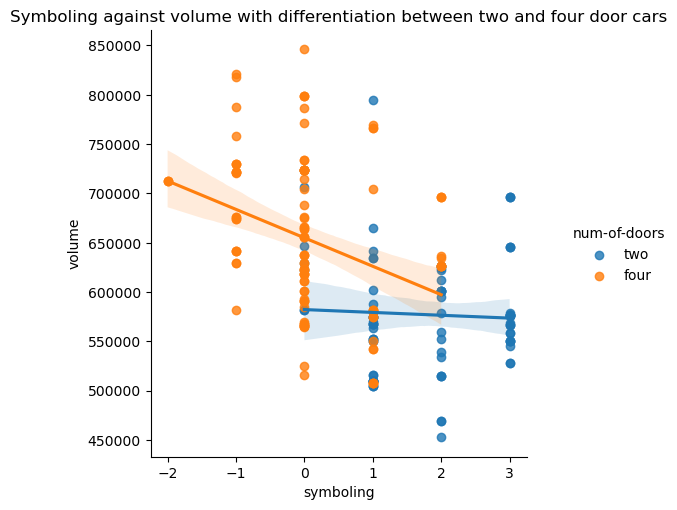
We also see a strong correlation between symboling and normalized losses which is expected, except for convertibles, as symboling increase for convertibles the normalized losses decrease.

With all this said, it looks like the cars which are aimed at commuters or families, people who drive to get around have a lower symboling than those cars aimed at people who enjoy driving and drive for fun. It cannot be ignored that the attitude of these drivers could also play a part in the symboling and not only the physical properties of the cars themselves.

This can also be seen in the below :

The difference in the average symboling of two door cars, those aimed at people who enjoy driving, versus four door cars, those aimed at families who use them only practically, is huge.

The purpose of the car is important to consider as well.



This graph shows us how the volume compares with symboling and number of doors. We can see that two doors have a higher symboling than four doors as confirmed by the previous graph. The fitting of a regression line is valuable here. Showing that for four door cars, the symboling decreases as the volume of the car increases, a bigger, heavier four door car is safer. For two door cars, this does show the same regression however to a much lower extent meaning that size (and therefore weight) of the vehicle has less impact on how safe a sportier car-lovers’ car is.

## Data story 1: Does car-size affect symboling?

Yes, it appears that it does. Cars with a higher volume consistently score lower on symboling and normalized losses. Smaller, sporty cars like convertibles or two-door models have on average a lower volume and higher horsepower and are among the riskiest cars in the dataset.

## Data story 2: Does body style, size, and Engine specifications affect price?

This also appears to be yes. It was shown that the cars-types with the highest horsepower also had the highest average price, while those with modest engine powers and more practical body types were cheaper. The heatmap also showed that price goes up with the length (0.7), width (0.7), and weight (0.8) of the car (all aspects of its size and body type). It also had a very high correlation (0.9) with engine size and horsepower (0.8). This shows that the engine specification is a huge factor in price, even more than the size of the car, but both are major contributors. The price of the car goes down according to its miles per gallon (in both city and highway as these two have a 1:1 correlation with each other).