# Exploratory Data Analysis on the Automobile Data Set

# My Introductory Data Observations:

## Summary of the Dataset:

The dataset was taken from a text file called *automobile.txt*.

This data set contains the makes of cars and lists their various features and values. The columns are broken down into the following features below:

- symboling:this is the insurance risk rating
- make: This is the car brand
- normalized-losses: the relative average loss payment per insured vehicle year
- fuel-type: Petrol or diesel
- aspiration: Type of engine aspirator
- num-of-doors: Number of doors to car -
- body-style: Style of car body
- drive-wheels: 4wd/rwd/fwd drive
- engine-location: Front or Back
- wheel-base: Wheel dimensions
- length: Car Length
- width:Car Width
- height:Car Height
- curb-weight:Total weight of car
- engine-type: Type of engine
- engine-size: Size of engine
- fuel-system: Type of fuel system in vehicle
- bore: the diameter of each cylinder
- stroke: Movements of the piston
- compression-ratio: Volume of the cylinder with the piston
- horsepower: Power car produces
- peak-rpm: Revolution per minute
- city-mpg: Driving with occasional stopping and braking ideal for city
- highway-mpg: Driving with occasional stopping and braking ideal for highway
- price: The car price

What kind of Data Type is this?
This data is based around the comparisons of different car makes and features related to their selling price. This data set can be used to determine the relationships the different car features have with one another and can be used to draw several conclusions.

The dataset is ordinal as some of the data is qualitative as it has a natural order of rank based on specific characteristics. This could be look from the features that contribute to the vehicles selling price for example. There is also a bit of ratio data in this data frame as there is variables that can measured on a continuous scale such height, weight ect.

An example of Dichotomous data would be the 'num-of-doors' and 'fuel-type' column as they only have 2 possible values.

Outliers in data set:
None observed, as all data columns is consistent with one another and there are no unusual measurements.

## DATA CLEANING & MISSING DATA

It is important to first check your dataset for any data that needs to be cleaned before exploring and creating visualisations.

I checked the data for the following:

1) Are there any missing values?
2) Are there values out of place?
3) Are there any duplicated rows?
4) Remove irrelevant data.
5) Changing data types.

After creating the Data frame, I printed out the first 10 rows to see if any potential values are of concern:

| | symboling | normalized-losses | make | fuel-type | aspiration | num-of-doors | body-style | drive-wheels | engine-location | wheel-base | ... | engine-size | fuel-system | bore | stroke | compression-ratio | horsepower |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 3 | ? | alfa-romero | gas | std | two | convertible | rwd | front | 88.6 | ... | 130 | mpfi | 3.47 | 2.68 | 9.0 | 111 |
| **1** | 3 | ? | alfa-romero | gas | std | two | convertible | rwd | front | 88.6 | ... | 130 | mpfi | 3.47 | 2.68 | 9.0 | 111 |
| **2** | 1 | ? | alfa-romero | gas | std | two | hatchback | rwd | front | 94.5 | ... | 152 | mpfi | 2.68 | 3.47 | 9.0 | 154 |
| **3** | 2 | 164 | audi | gas | std | four | sedan | fwd | front | 99.8 | ... | 109 | mpfi | 3.19 | 3.40 | 10.0 | 102 |
| **4** | 2 | 164 | audi | gas | std | four | sedan | 4wd | front | 99.4 | ... | 136 | mpfi | 3.19 | 3.40 | 8.0 | 115 |
| **5** | 2 | ? | audi | gas | std | two | sedan | fwd | front | 99.8 | ... | 136 | mpfi | 3.19 | 3.40 | 8.5 | 110 |
| **6** | 1 | 158 | audi | gas | std | four | sedan | fwd | front | 105.8 | ... | 136 | mpfi | 3.19 | 3.40 | 8.5 | 110 |
| **7** | 1 | ? | audi | gas | std | four | wagon | fwd | front | 105.8 | ... | 136 | mpfi | 3.19 | 3.40 | 8.5 | 110 |
| **8** | 1 | 158 | audi | gas | turbo | four | sedan | fwd | front | 105.8 | ... | 131 | mpfi | 3.13 | 3.40 | 8.3 | 140 |
| **9** | 0 | ? | audi | gas | turbo | two | hatchback | 4wd | front | 99.5 | ... | 131 | mpfi | 3.13 | 3.40 | 7.0 | 160 |

Missing values:
I noticed that some cells have the '?' character in.

I then proceeded to check if there were any missing values within the dataset by using the '*isna()*' function which searches all the cells to see if there is missing data. The function will return a Boolean with either a true or false value (meaning yes or no if data is missing). I then calculated the total sum of each possible missing value per column. The visualisation I got was the following:

```
====== ❑ Missing Values: =======

symboling            0
normalized-losses    0
make                 0
fuel-type            0
aspiration           0
num-of-doors         0
body-style           0
drive-wheels         0
engine-location      0
wheel-base           0
length               0
width                0
height               0
curb-weight          0
engine-type          0
num-of-cylinders     0
engine-size          0
fuel-system          0
bore                 0
stroke               0
compression-ratio    0
horsepower           0
peak-rpm             0
city-mpg             0
highway-mpg          0
price                0
```

There were no missing values within the data frame but I needed to check on the inconsistent values.

Out of place Values:
When I ran the Data frame earlier I noticed the '?' symbol in some cells.
I therefore had to check to see how many cells were affected by this.
I then used the *'isin(_)'* function to check for the character '?' in each cell. I also calculated the total number of times this occurred in each column. I got the following output below:

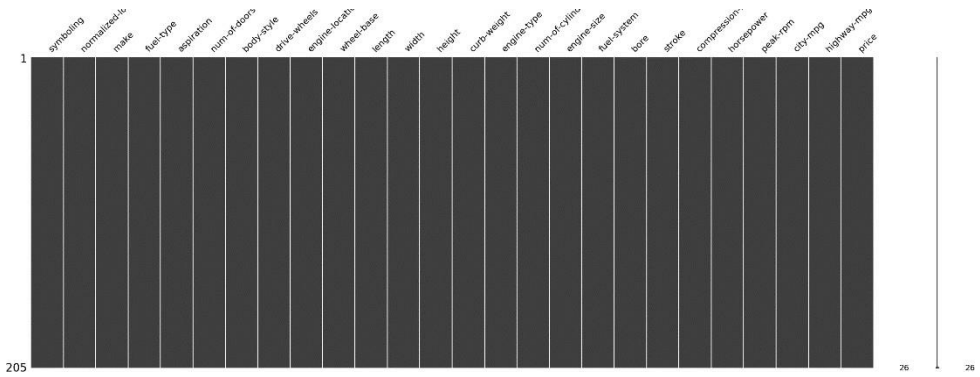
```
====== ❑ Non Values [ie-?]:

symboling            0
normalized-losses    41
make                 0
fuel-type            0
aspiration           0
num-of-doors         2
body-style           0
drive-wheels         0
engine-location      0
wheel-base           0
length               0
width                0
height               0
curb-weight          0
engine-type          0
num-of-cylinders     0
engine-size          0
fuel-system          0
bore                 4
stroke               4
compression-ratio    0
horsepower           2
peak-rpm             2
city-mpg             0
highway-mpg          0
price                4
```

Non Values Observed:
Normalized-losses, num-of-doors, bore, stroke, horsepower, peak rpm and price = MCAR: the missing values for the losses are random and probably were not recorded. It is independent from any other data within the file.

I then had to clean up this data by the following method:

- I used the 'replace' function to convert all the cells that had '?' symbol in them and return them with a NAN ('*Not A Number*') value.
- I then used the '*dropna()*' function to drop all the columns that had NAN values and saved into a new data frame called 'automobile_df'.
- After doing that I checked again to see if the data frame had any missing values. I used the '*missingno*' module to output a matrix visualisation that counts the total occurrence of NAN values in each column. I got the following:




The new Data frame 'automobile_df' had now no missing values or non values.

Potential duplicates:

I then checked to see if the Data Frame had any duplicated rows by using the '*duplicated*' function. The total number of duplicated rows returned as 0.

Removing Irrelevant data:

I then decided which columns I would not need to explore and create visualisations with. These columns were irrelevant and were not needed. I used the '*drop*' function to drop all the columns that I did not need to use. The new cleaned up data frame was now as follows:

```
Index(['symboling', 'normalized-losses', 'make', 'fuel-type', 'num-of-doors',
       'body-style', 'curb-weight', 'engine-size', 'compression-ratio',
       'horsepower', 'peak-rpm', 'city-mpg', 'highway-mpg', 'price'],
      dtype='object')
```

Changing data types:
I proceeded to check the data types for each column to see if they could be used for visualisations. This is the output I got:

```
symboling              int64
normalized-losses     object
make                  object
fuel-type             object
num-of-doors          object
body-style            object
curb-weight            int64
engine-size            int64
compression-ratio    float64
horsepower            object
peak-rpm              object
city-mpg               int64
highway-mpg            int64
price                 object
```

I noticed that several columns' data types were still showing as 'objects' which would create a problem when trying to plot visualisations.
I had to convert these columns into either strings, intergers or float types.
I used the *'to_numpy'* function to covert the columns into the correct data types. After converting all the columns, I got the following:

```
===== ❑ New Dataframe Dtypes: =======
symboling              int64
normalized-losses      int64
make                  string
fuel-type             string
num-of-doors          string
body-style            string
curb-weight            int64
engine-size            int64
compression-ratio    float64
horsepower             int64
peak-rpm               int64
city-mpg               int64
highway-mpg            int64
price                  int64
dtype: object
***** NO VALUES ARE OBJECTS ANYMORE *****
```

## DATA STORIES AND VISUALIZATIONS

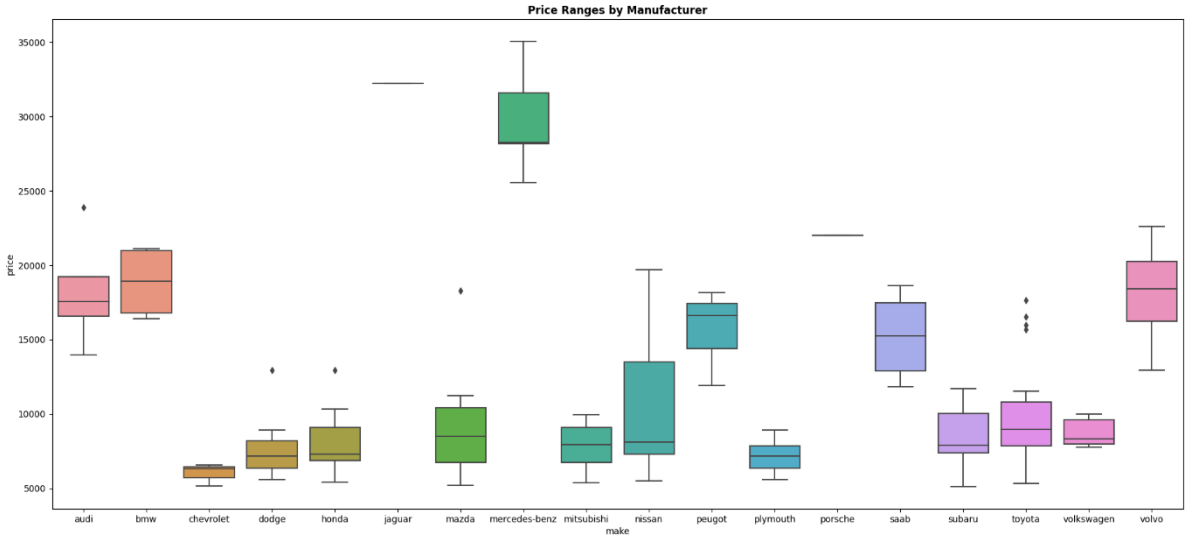
Now that the Data frame was cleaned up, I could now proceed to create my visualisations and explore the dataset.

INDEX:



### Visualisation I-Price Ranges

I created a visualisation to see the price ranges of each Car make.



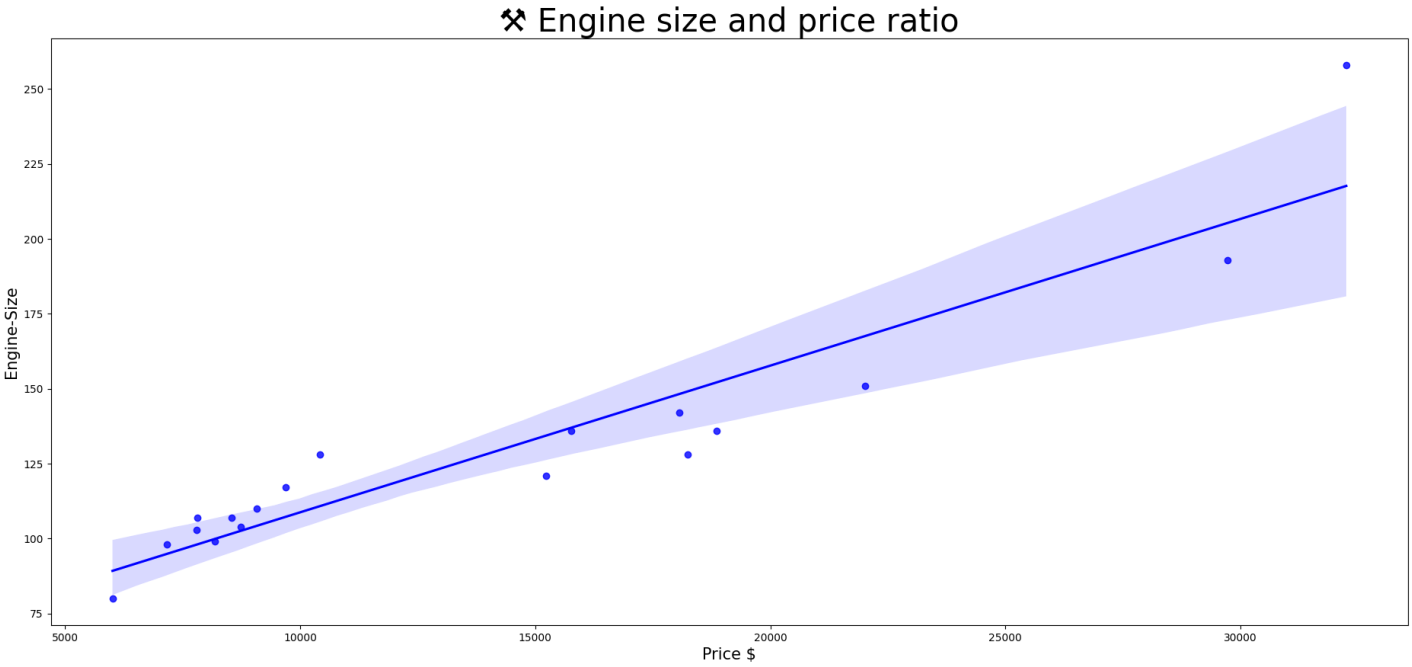

### Exploration Analysis:

I can see from the above visualisation that it is interesting to note that the cars with the highest selling values are Mercedes, Jaguar, Porsche and BMW. It would be interesting to further explore what factors might contribute to the high selling prices (IE-*why are some car makes more expensive than others?)*

## Visualisation II-Features and Price Analysis:

I created several visualisations exploring what features contribute to the price ranges:

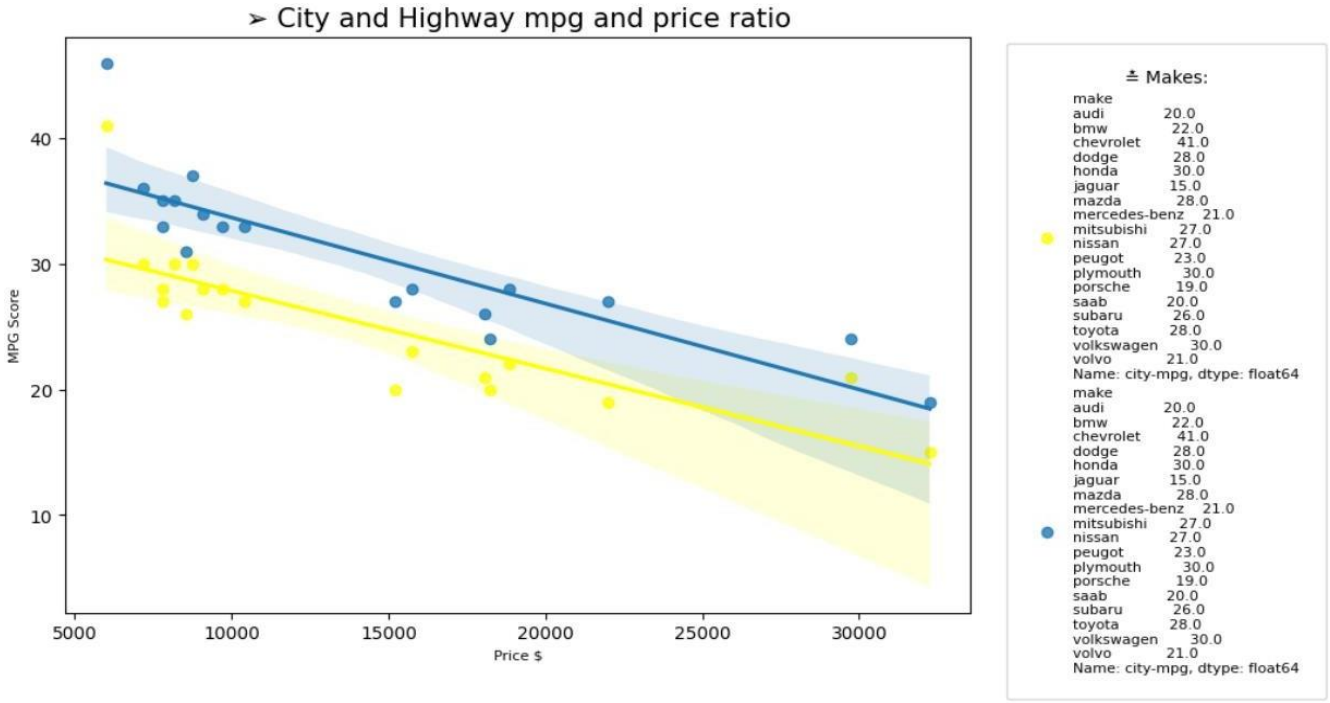
- Engine size
- City/highway mpg
- Curb Weight
- Horsepower




### Exploration Analysis:

The Engine size: the larger or heavier the engine, the higher the price of the car:

It takes more resources to create a bigger engine, or more quality materials can cause more weight making the manufacturing more costly.
The size of the engine can also mean that more features are put into it such as extra cylinders, or the size of the cylinders to create more power etc.



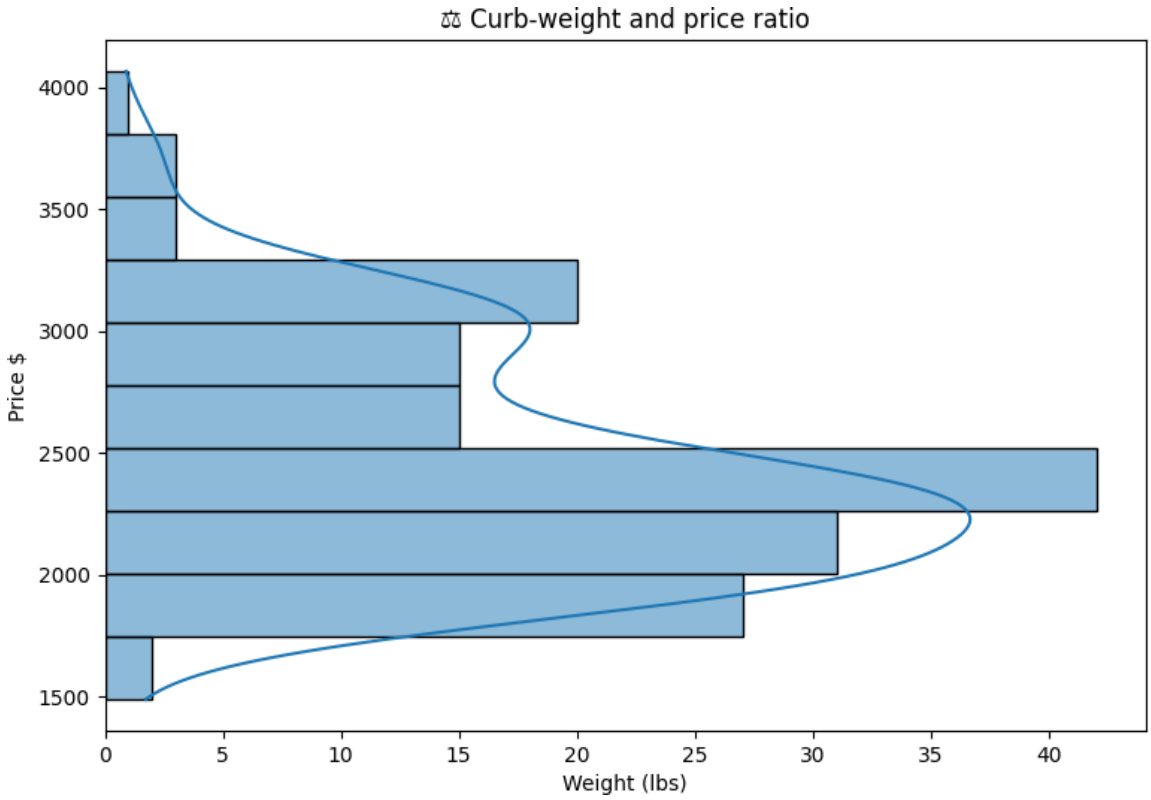

## Exploration Analysis:

The City and Highway mpg: The lower the mpg value the higher the price:

[The MPG ('miles per gallon')]

City MPG: the score a car will get on average in city conditions, with stopping and starting at lower speeds(1)
Highway MPG: the average a car will get while driving on an open stretch of road without stopping or starting, typically at a higher speed.(1)

The higher a car's mpg value is then the more efficient the fuel is.
Therefore, the city and highway mpg at a low score means that the vehicle price is lower. This is quite interesting as you would naturally think that the higher the score the more the price will increase, but according to this dataset it is the opposite. A factor could be that when manufacturing a car with a high mpg score could be making the car more small, lighter in weight and perhaps less powerful(2)



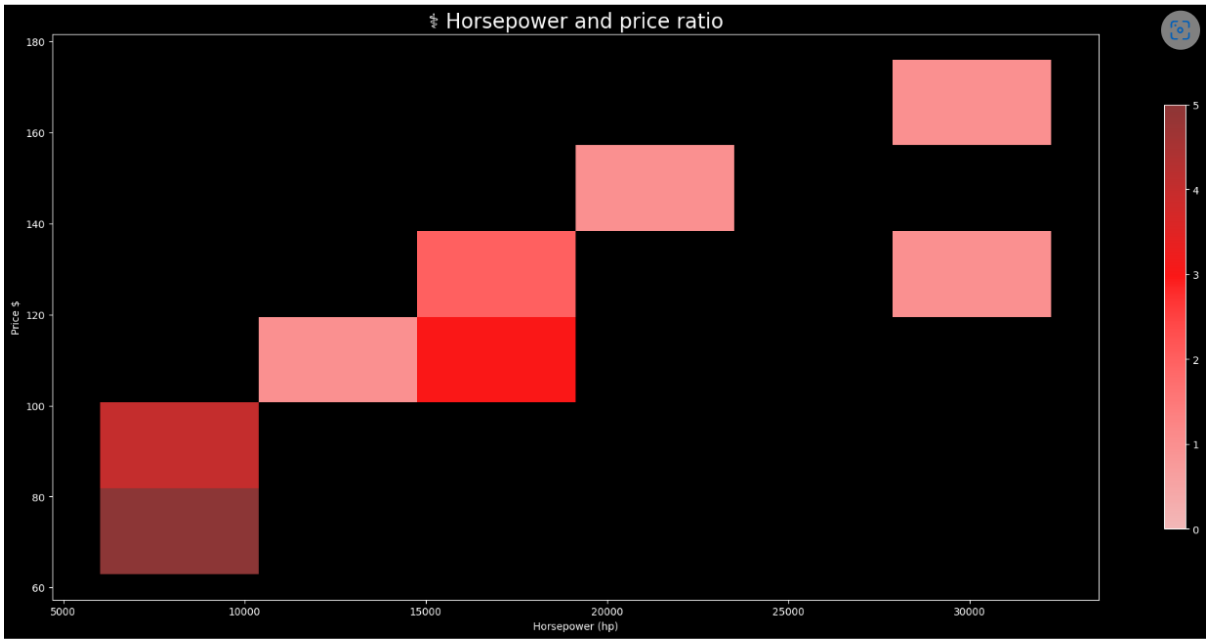

## Exploration Analysis:

> The Curb weight: Generally, the higher the curb weight does not really affect the price

We can see that the curb weight (the total weight of the car) does not really affect the price.

It is interesting to see cars with the highest curb weights are towards between the 2000-2500 price range, which is more or less where the following car makes are: BMW, Porsche, and Volvo.
This could be because these makes use heavy materials in their cars, like the Volvo adds extra features for safety, and Porsche has more hardtops on their vehicles, which can either mean more weight due to the rigid metal or the convertible feature which adds more weight.

Therefore, curb weight and body-style might have a relationship.

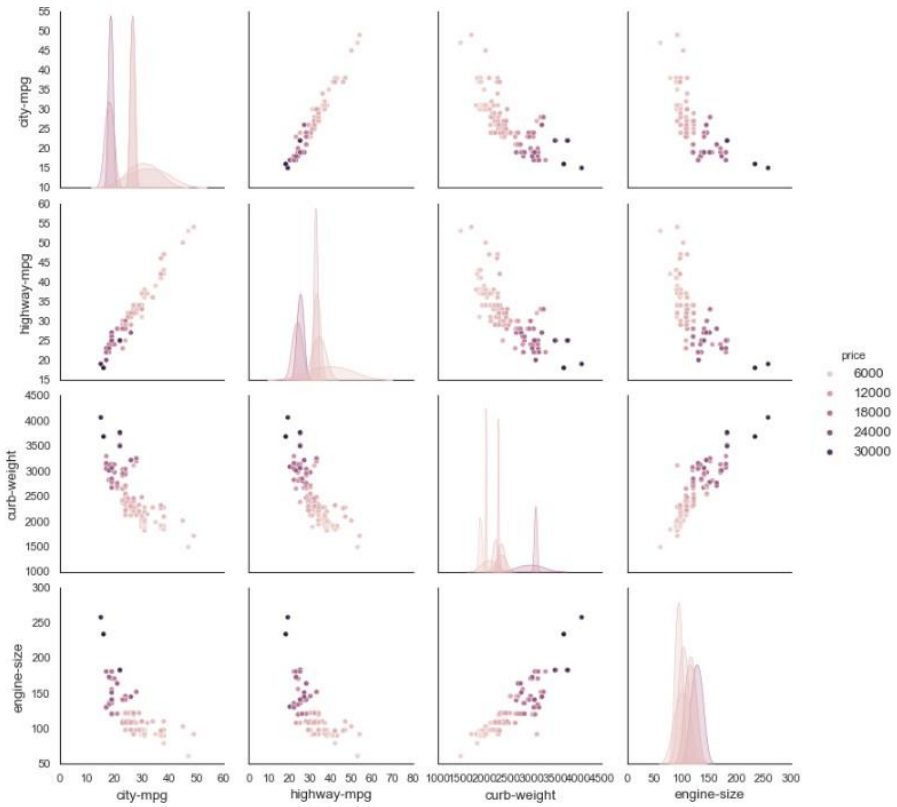



## Exploration Analysis:

The Horsepower: The higher the Horsepower, the higher the price:

We can see that the more horsepower a car produces the higher the price.
This could be because horsepower requires a more powerful engine which of course will increase the manufacturing costs.

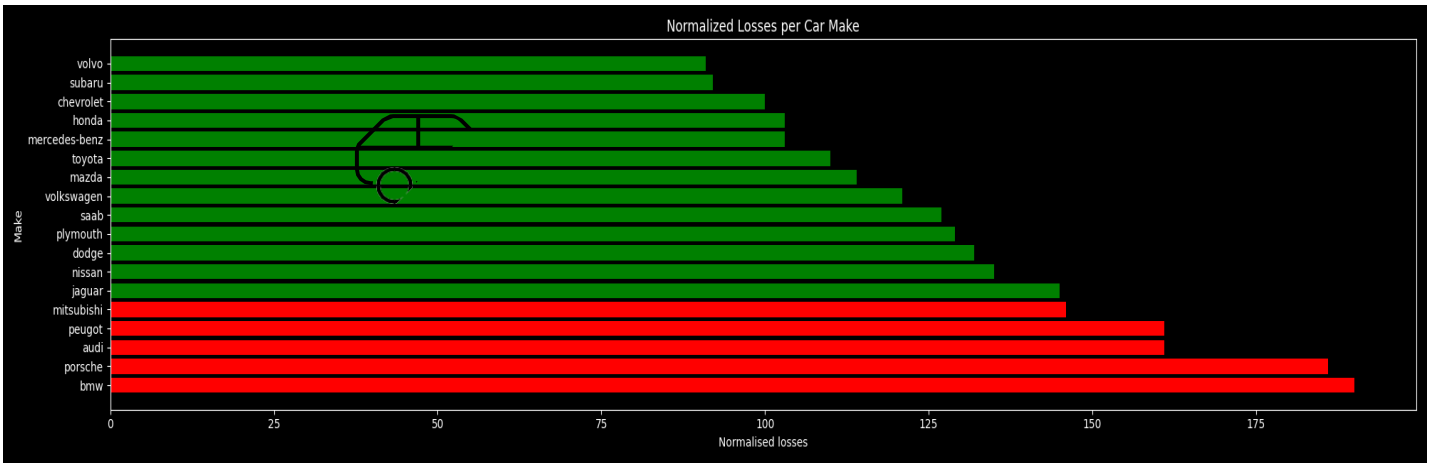
**OVERVIEW:**



## Visualisation III-Normalized Losses:

I created several visualisations exploring what features contribute to the normalized losses:

- Normalized Losses per car make
- Symbolling and Normalized Losses correlation




## Exploration Analysis:

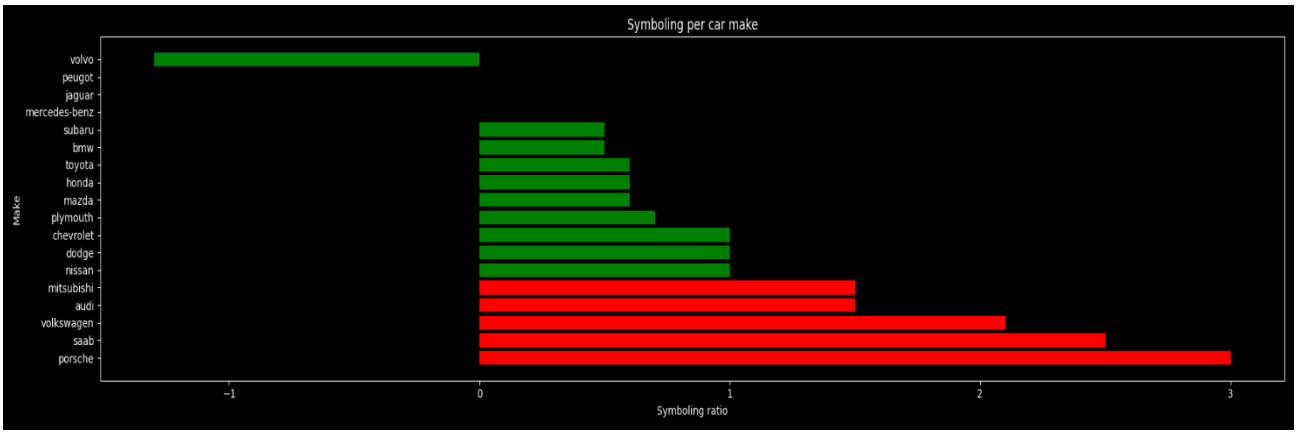
> The Normalized Losses: We can see by the visualisation the car makes who have the lowest to highest Normalized losses.

We can see that the top 5 cars with the highest losses is: Mitsubishi, Peugeot, Audi, Porsche and BMW.

The Normalized losses is basically the average loss of payment per the insured vehicle year, in other words-It is the risk factor associated with the car's price. These factors could be based off insurance risk studies.

Perhaps BMW and Porsche drivers are more likely to get into accidents due to their horsepower and speed. The cars might be more expensive to repair or getting spare parts might be costly.

The car with the lowest normalised loss is the Volvo: (see next visualisation)




## Exploration Analysis:

The Normalized Losses and Symbolling correlation: We can see by the visualisation that the car with the lowest Symbolling is the Volvo:

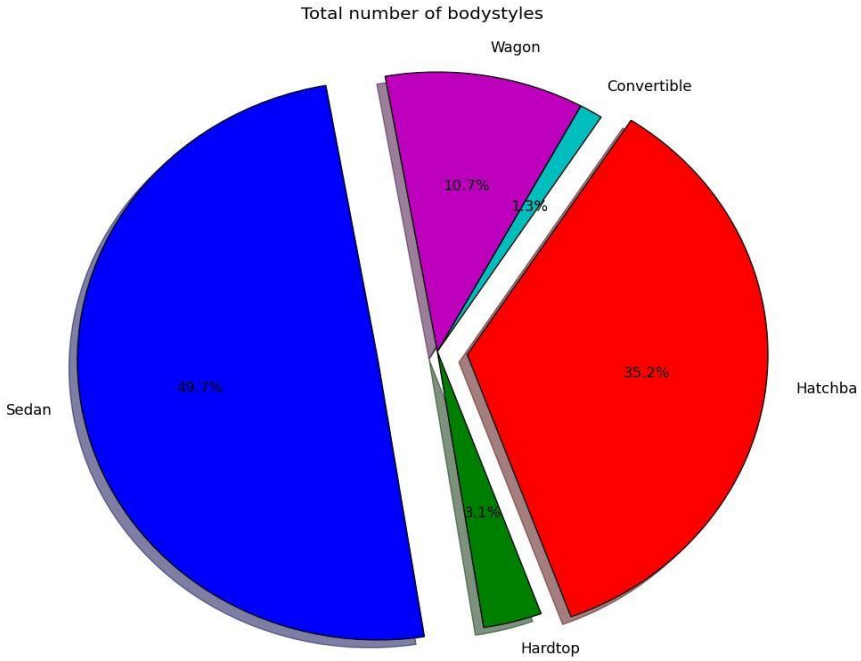
The Symbolling is the insurance risk rating. The higher the insurance risk, the higher the score.

The Volvo has the lowest normalised loss and the best score for the Symbolling. This concludes that the insurance risk on a Volvo is relatively low as they are rated one of the safest vehicles on the road today.

## Visualisation IV-Car Features

I created several visualisations exploring what features contribute to the overall manufacturing of the cars. In other words what type of cars were more manufactured based on their features:

- The Cars Body Style
- The Cars Fuel Type
- The Cars Door Features

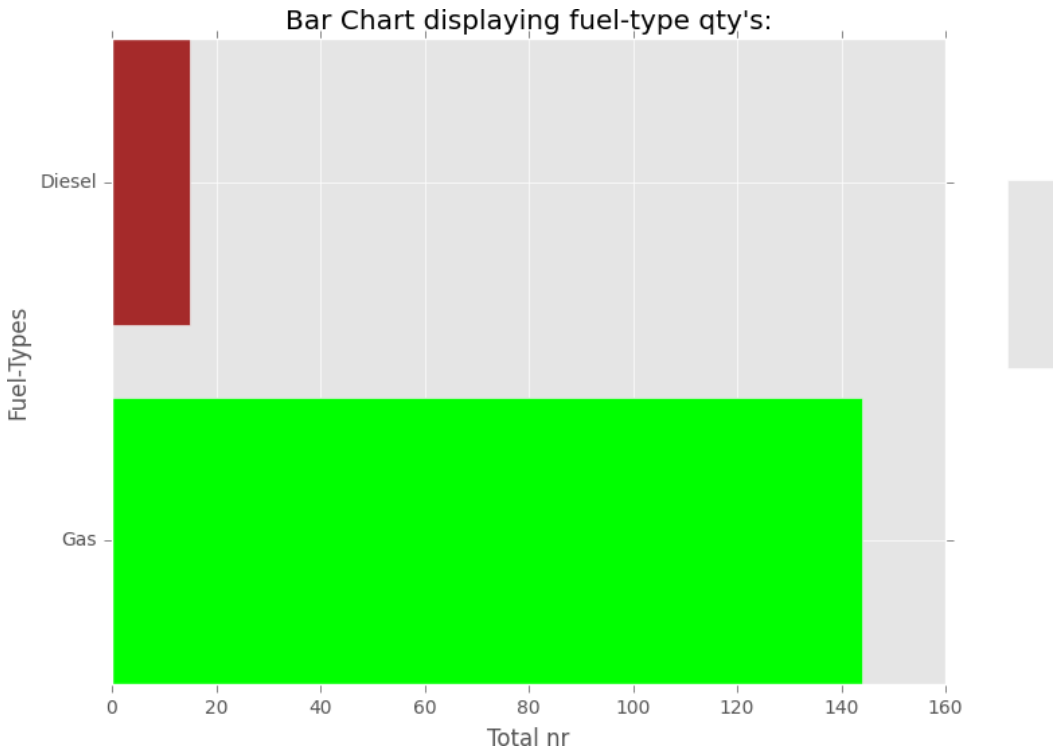



## Exploration Analysis:

The Car Body Styles: We can see by the visualisation the number of car body styles ratio's

The most popular body style being manufactured is the Sedan
This could be because it is more of a family car and the sedan is generally more affordable. Most families need a car and will therefore need a car that can accommodate boot space and seating ect.

The least popular body style is the convertible.
This could be because convertibles only appeal to a certain market of individuals. The convertibles are more expensive and have less space inside. Convertibles might also be more geographically popular due to weather factors.

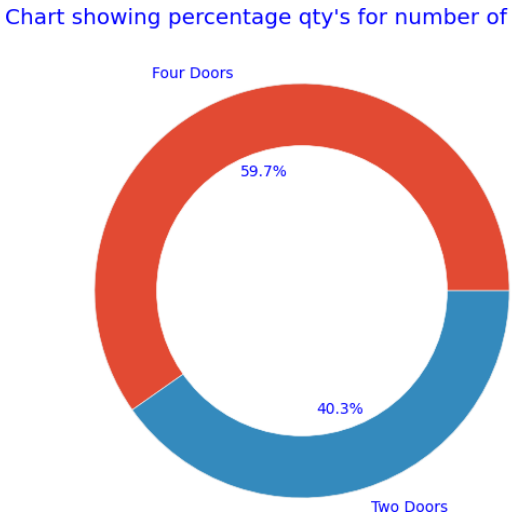



## Exploration Analysis:

The Car fuel Types: We can see by the visualisation the ratio between gas and diesel.

The more popular fuel type is the gas.

This means that more cars are manufactured that take gas/petrol as a fuel type. The reason is because gas is more accessible in most parts of the world; diesel pumping can be more expensive than gas; diesel engines can be difficult to maintain and require diesel mechanics; a diesel engine also is more expensive to make (1)




## Exploration Analysis:

The Car door Types: We can see by the visualisation the ratio between a car with 4 doors and a car with 2 doors.

The more popular type is cars with 4 doors.

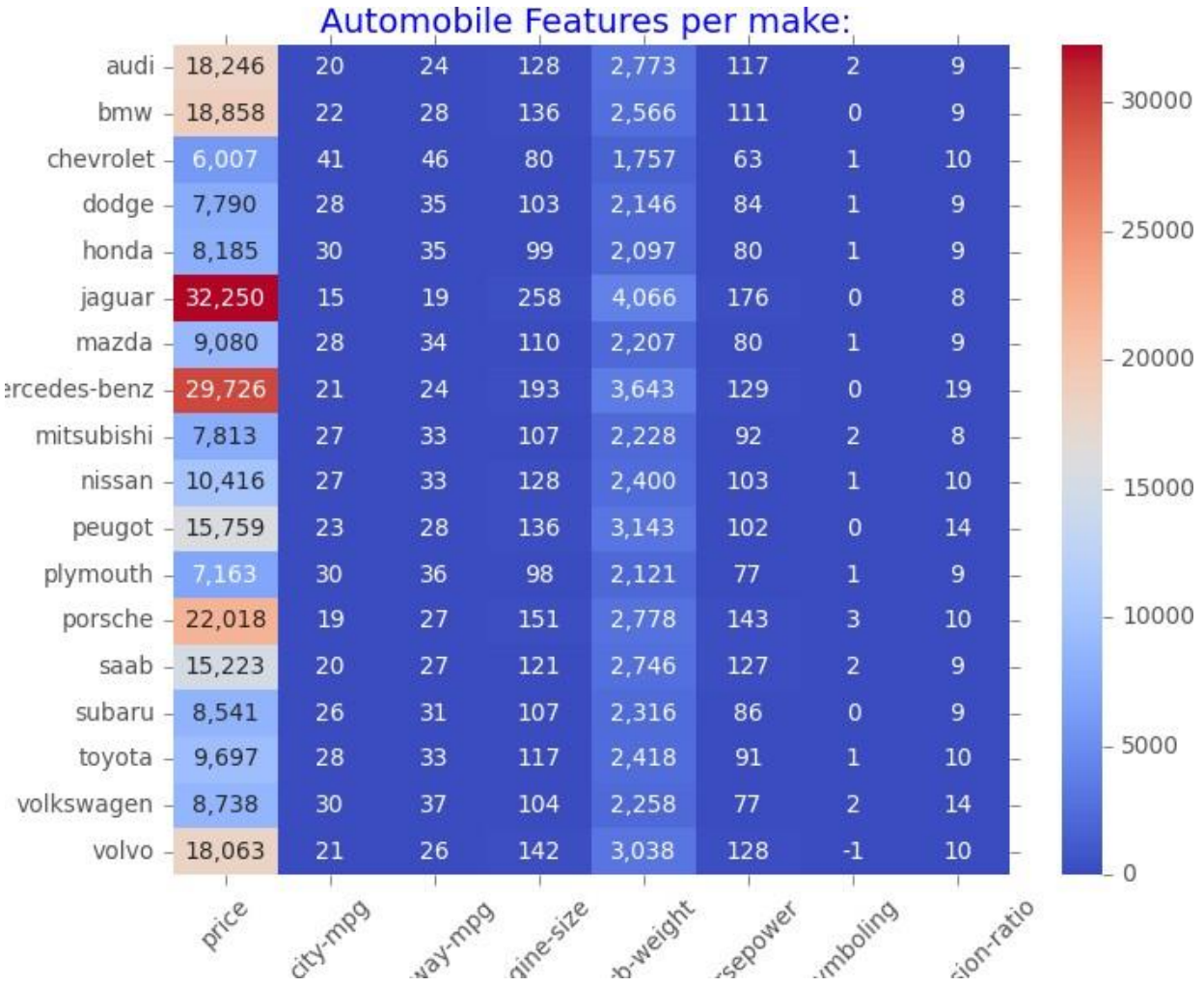
The reason is because (as mentioned above) more sedan vehicles are manufactured, and majority of sedan cars have 4 doors.

Another reason is because a car with 4 doors is easy to accommodate passengers.

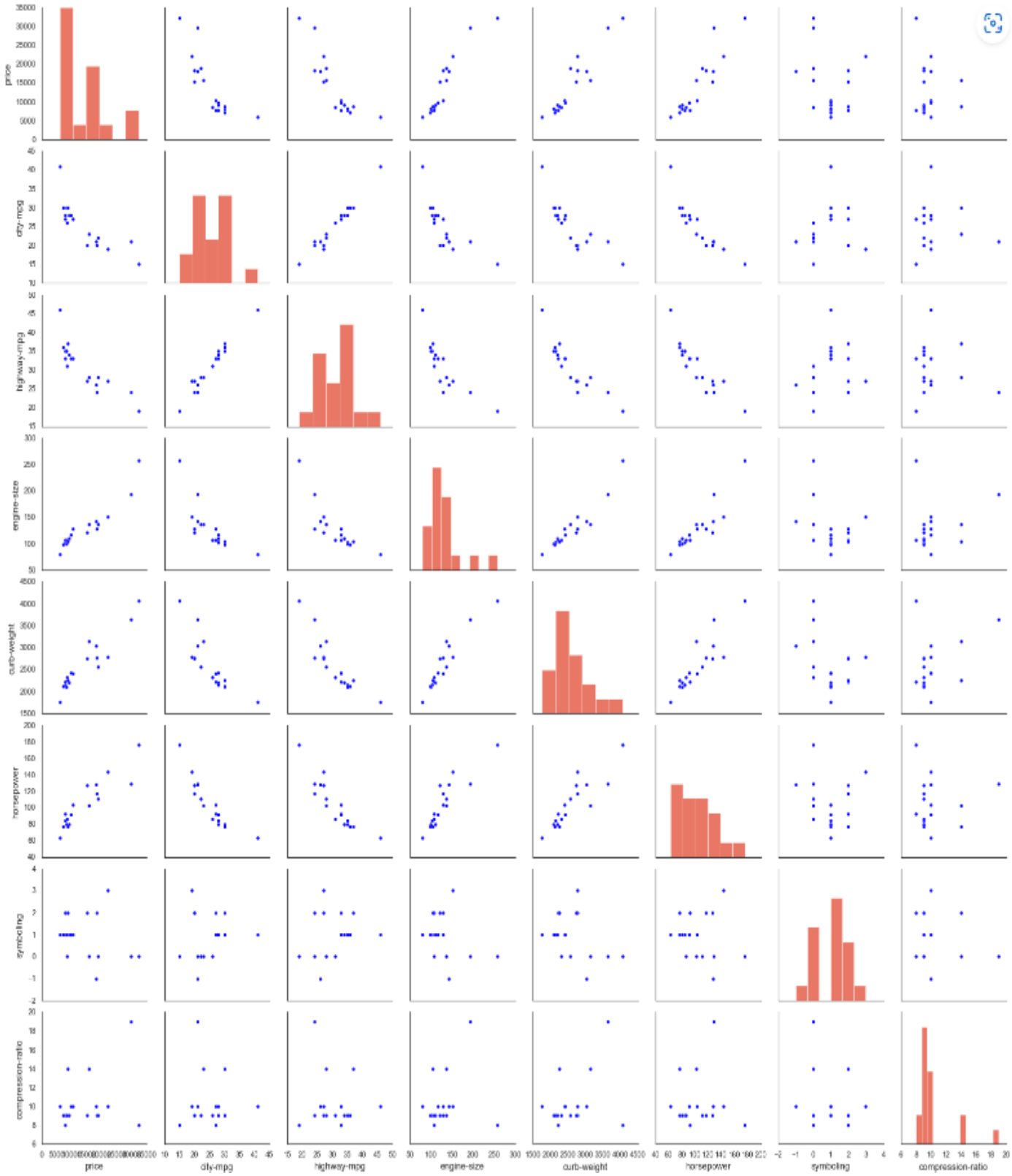
Most cars that have 2 doors are mainly sports cars or more expensive cars that would only appeal to a certain market for people who can afford them.

## Visualisation V-Overall features

I created 2 visualisations showing all the features and their relationships one to another per car make and Price. This is a nice overview to have if you wanted to draw further exploratory analysis from the Automobile Data Set.



## Overall features with the focus on price:



## Conclusion:

We can see from the dataset it is important to compare the various features with one another in order to understand how it affects the pricing as well as manufacturing qty's. This particular Exploratory Data Analysis is more based off a general consensus of the relationships mentioned above. You could explore further by investigating features per car make etc.