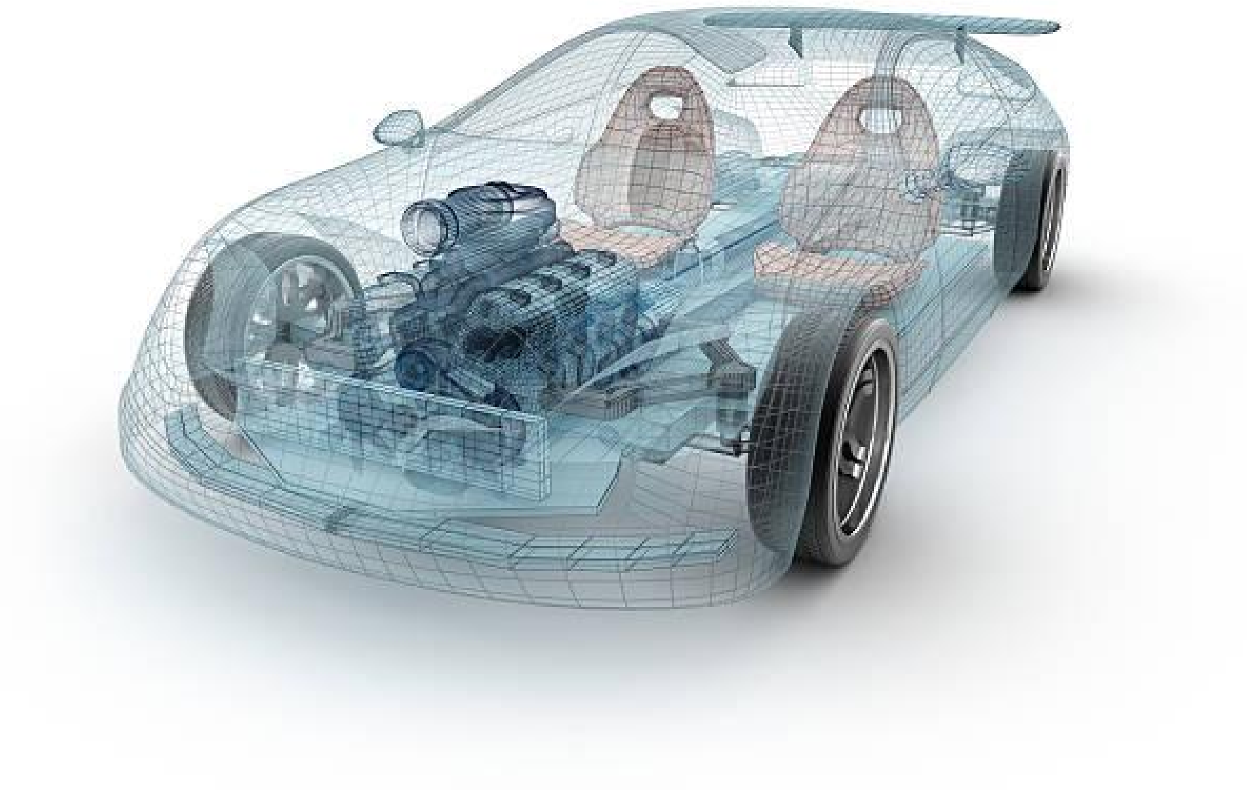
WILL GOSNOLD

Automobile Dataset

Exploratory Data Analysis



Introduction

In this EDA, I am going to be looking at the 1985 Model Import Car and Truck Specifications dataset taken from 1985 Ward's Automotive Yearbook. The dataset also contains data from an Insurance Collision Report by the Insurance Institute for Highway Safety.

Data Summary

The dataset is made up of 205 entries for various automobiles with 26 columns describing the following features:

- 'symboling'
  - A rating between -3 and 3 that corresponds to the degree to which the automobile is more risky than its price indicates. Cars are initially assigned a risk factor that corresponds to their price. If the automobile is considered more dangerous, the value is increased (3 is the most risky and -3 is the safest).
- 'normalized-losses'
  - The relative average loss payment per insured vehicle year. The figure is normalised for all vehicles in a similar size category and represents the average loss per vehicle per year, where "loss payment" is the portion of an incurred loss that is paid by the insurer.
- The rest of the features are self-explanatory and define the specifications and price of the vehicle.

For this analysis I am going to be removing the symboling and normalized-losses columns, instead choosing to focus on how the specifications of the vehicles are related to their price. As such, please refer to the table below for an updated list of features and their ranges. Note that these were the figures still in the data after cleaning.

Feature Information

	Feature	**Feature Range**
1.	make	`toyota, nissan, mazda, mitsubishi, honda, volkswagen, subaru, peugot, volvo, dodge, mercedes-benz, bmw, plymouth, audi, saab,porsche, jaguar, chevrolet, alfa-romero, isuzu, renault, mercury`
2.	fuel-type	`gas, diesel`
3.	aspiration	`std, turbo`

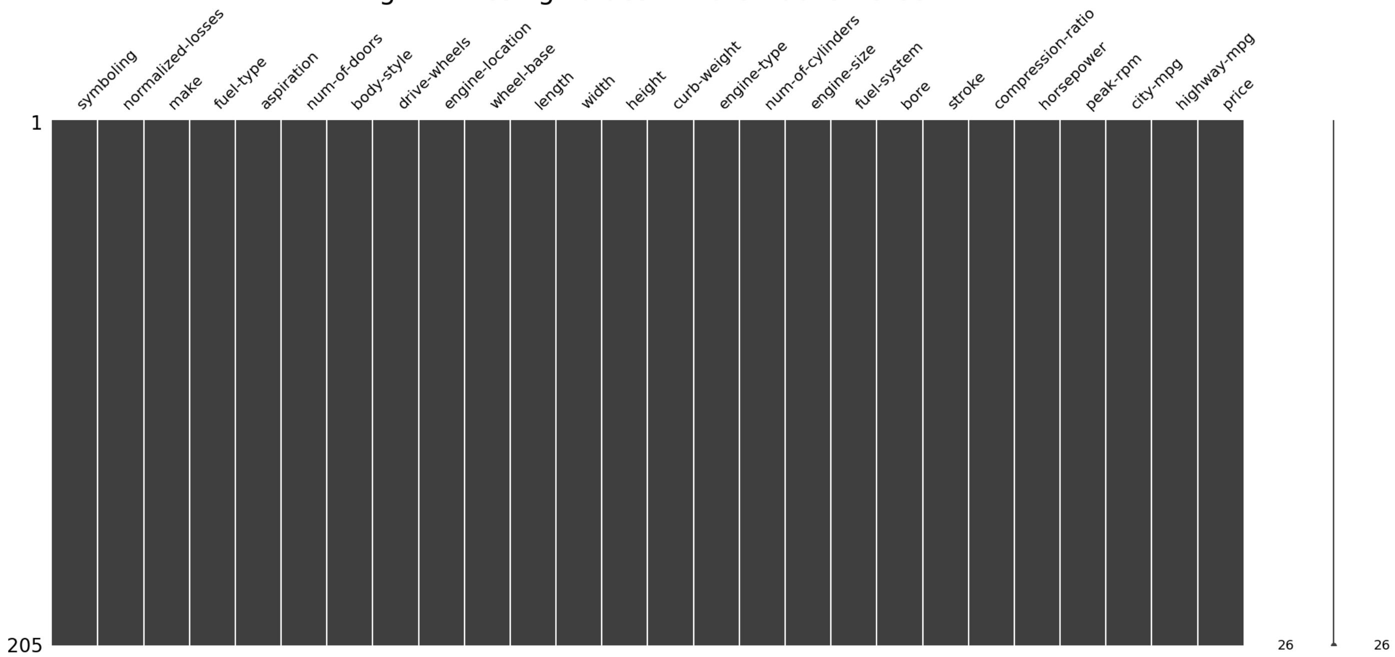
4.	num-of-doors	two, four
5.	body-style	sedan, hatchback, wagon, hardtop, convertible
6.	drive-wheels	fwd, rwd, 4wd
7.	engine-location	front, rear
8.	wheelbase	86.6 - 120.9, continuous
9.	length	141.1 - 208.1, continuous
10.	width	60.3 - 72.0, continuous
11.	height	47.8 - 59.8, continuous
12.	curb-weight	1488 - 4066, continuous
13.	engine-type	ohc, ohcf, ohcv, dohc, l, rotor
14.	num-of-cylinders	2 - 12, continuous
15.	engine-size	61 - 326, continuous
16.	fuel-system	mpfi, 2bbl, idi, 1bbl, spdi, 4bbl, spfi
17.	bore	2.54 - 3.94, continuous
18.	stroke	2.07 - 4.17, continuous
19.	compression-ratio	7.0 - 23.0, continuous
20.	horsepower	48 - 262, continuous
21.	peak-rpm	4150 - 6600, continuous
22.	city-mpg	13 - 49, continuous
23.	highway-mpg	16 - 54, continuous
24.	price	5118 - 45400, continuous

Data Cleaning

Missing Data

Let's begin by having a look to see if there are any missing values in the data. Fig. 1 below shows all NaN values in the dataset.

Fig. 1 - Missing Values in Automobile Datset

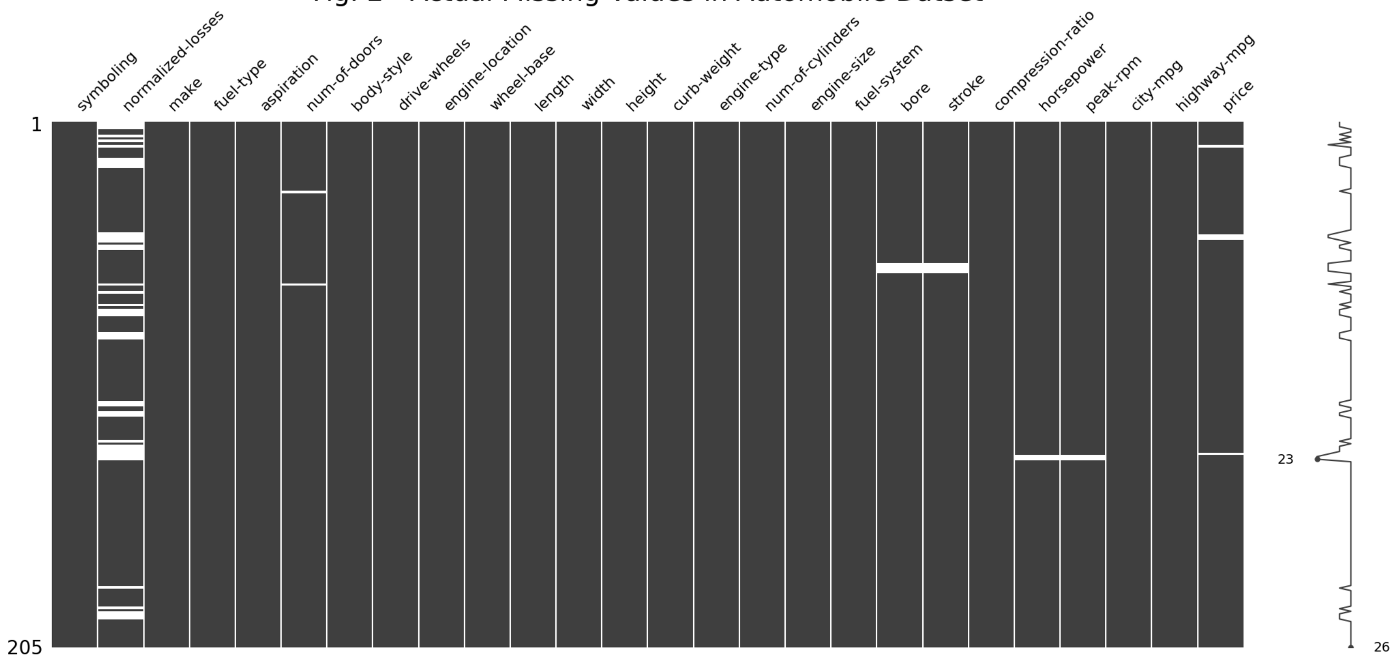


Normally, we would see missing values in a column showing up as white lines in the bars. Here we see there are apparently no missing values. This could be because the data was so expertly collected that there are no missing values, or because it has already been cleaned, or because the missing data is being represented by something other than 'NaN'.

Looking a bit closer, I notice that some of the values in the dataframe are '?'. Replacing the question marks with NaNs, and repeating the graph from above we get Fig. 2 (pg. 5). Now we can see a bit more clearly where the missing data is.

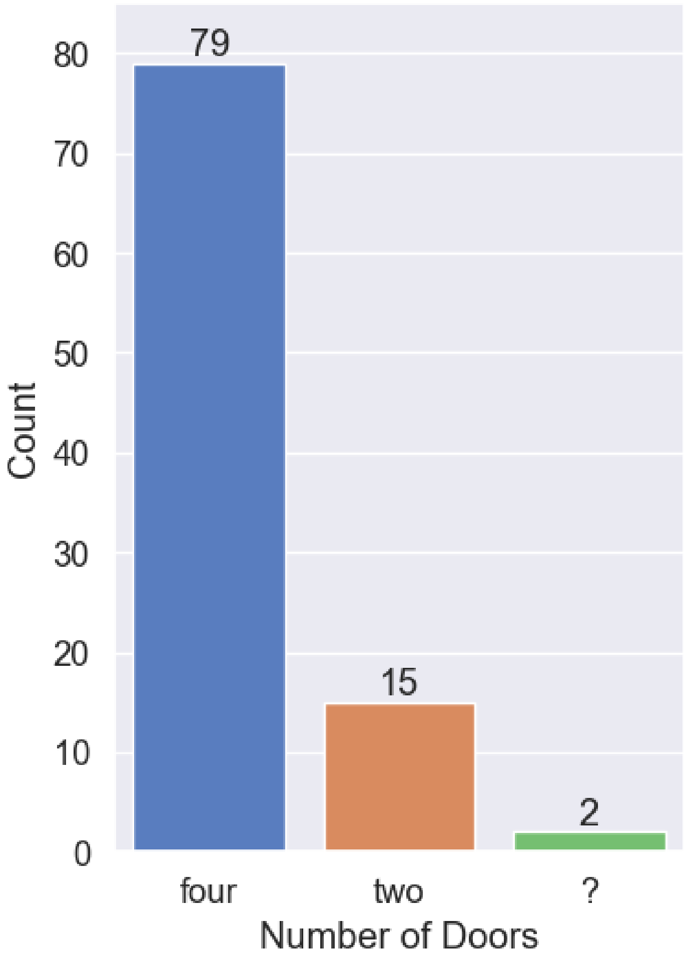
The vast majority of the missing values are in 'normalized-losses' (41 missing values), which I have already decided to remove, so that's not a problem for us. I just need to decide what to do with the few remaining missing values. There are 2 entries missing each in 'num-of-doors', 'horsepower' and 'peak-rpm', and 4 each missing in 'bore', 'stroke' and 'price'. Since we are going to be looking at how the specifications of the vehicles might help us to predict price, it doesn't make sense to impute the price data so I'm going to drop those rows with missing values for price.

Fig. 2 - Actual Missing Values in Automobile Datset



For the remaining missing values, we have to look at the entries in a little more detail to decide what to do with them.

Fig. 3 - Number of Doors on Sedan Vehicles



In the 'num-of-doors' column, there are 2 categories for us to choose from: 'two' or 'four'.

The two missing values both belong to vehicles that have the value 'sedan' in the 'body-style' column and, as we can see in Fig. 3, the vast majority of these are 4 door vehicles. Therefore, I will replace both of these missing values with 'four'.

The remaining missing values are all continuous numerical columns and since there are only a few values missing from each, I am going to use mean imputation to replace all of them.

After tidying this up, we are left with 201 observations and 24 features on which to perform the analysis.

Data Types

For the sake of efficiency, especially when it comes to working with extremely large datasets in Pandas, we usually want to go through the data and optimise the data types that we are working with to take up less space in memory.

The data types that Pandas uses by default are ‘int64’ for all integer-like columns, ‘object’ for all string-like columns, and ‘float64’ for all decimal columns or any integer-like columns that have NaN values in them. Now that we’ve sorted out the missing values, it’s time to go through and clean up the data types.

I am going to change all the integer-like columns to whichever ‘int’ data type can still represent the data in that particular column. For example, ‘int8’ can represent numbers between -128 and 128, and the ‘city-mpg’ column only contains values between 13 and 49. Therefore, we can increase the speed and efficiency of working with this data by changing that column to ‘int8’. I also changed the ‘num-of cylinders’ column to be integers as this is not a true categorical column (12 cylinders is more similar to 8 cylinders than it is to 2 cylinders.

The ‘object’ columns that contain all the categorical data can be represented using the Pandas ‘category’ data type, which is far more efficient. So, I will also change all of the string-like columns to ‘category’.

To keep the precision of the data accurate, I have chosen to keep all the decimal columns in their current form as ‘float64’ type.

After performing all of these optimisations, we have taken the data from over 216 kilobytes to less than 23 kilobytes. This is a small dataset so we wouldn’t have had to worry about hardware being able to handle it in this instance, but for a much larger dataset, the 90% reduction in memory usage we achieved would have been extremely useful.

Note, that I am also going to rename the columns here to make them more like python variables and thus easier to manipulate within pandas. We can do this by replacing all the hyphens with underscores.

Data Stories and Visualisation

Initial Insights

Since we are going to focus on how the individual specifications might indicate price, let's start by having a look at how well correlated each of the numerical features are to price.

Fig. 4 - Correlation of Numerical Columns against Price

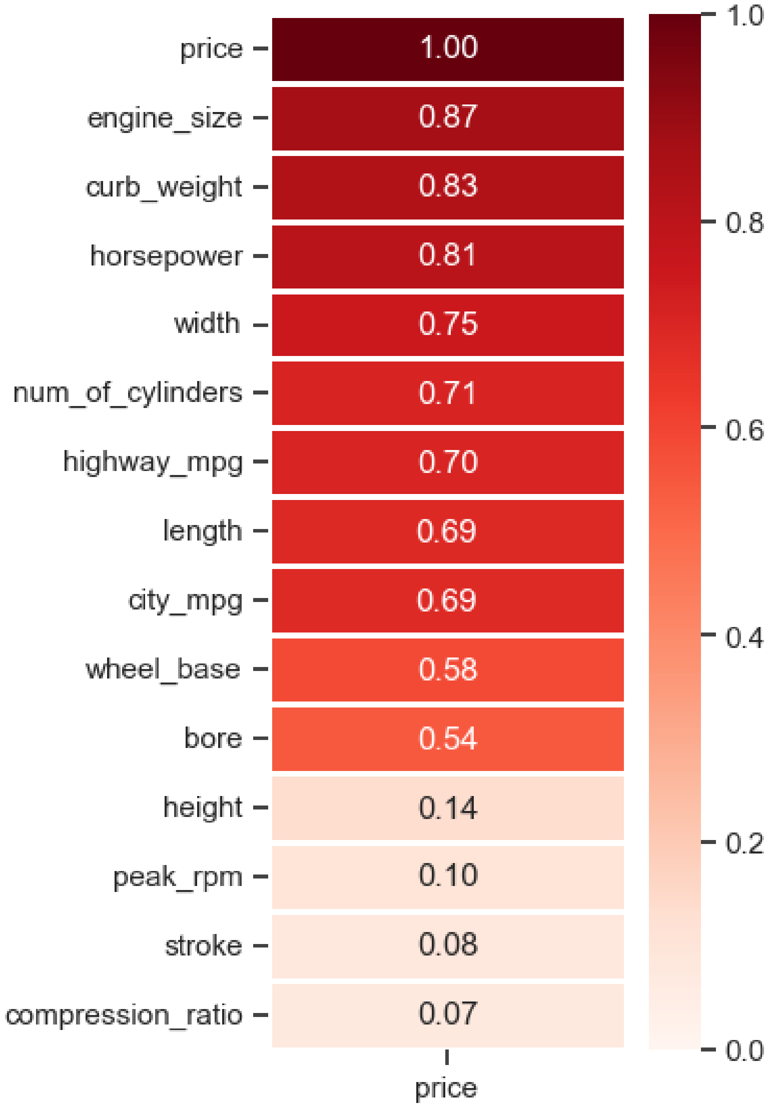


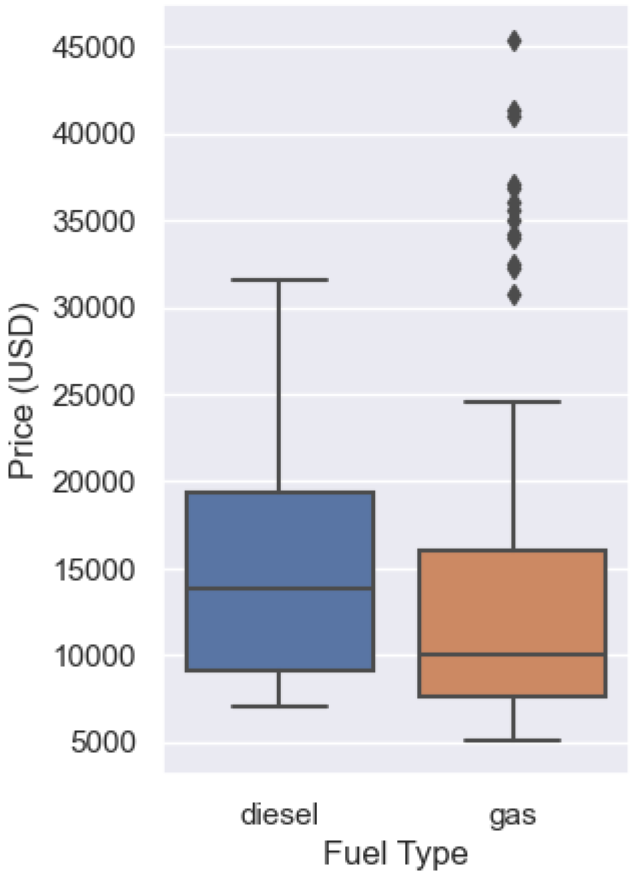
Fig. 4 shows the absolute correlation strength of each of the numerical features compared to the price of the vehicle.

Obviously price is perfectly correlated with itself, but after that we can see that many of the features have a strong correlation with price.

I would suggest that only 'height', 'peak_rpm', 'stroke', and 'compression_ratio' are weakly correlated enough that I might consider dropping them entirely. In any case, I will leave them out of the rest of this analysis.

Whilst we can quickly see the relevance of numerical features and their relationships to the dependent variable 'price', categorical features are not as simple to quantify. Figures 5-12 show all the categorical features in the dataset and the distribution of that data against price. This gives us an idea as to whether the feature will be indicative of price (if the distribution differs between categories), or not (if the distribution is similar between categories).

Fig. 5 - Fuel Type Price Distribution



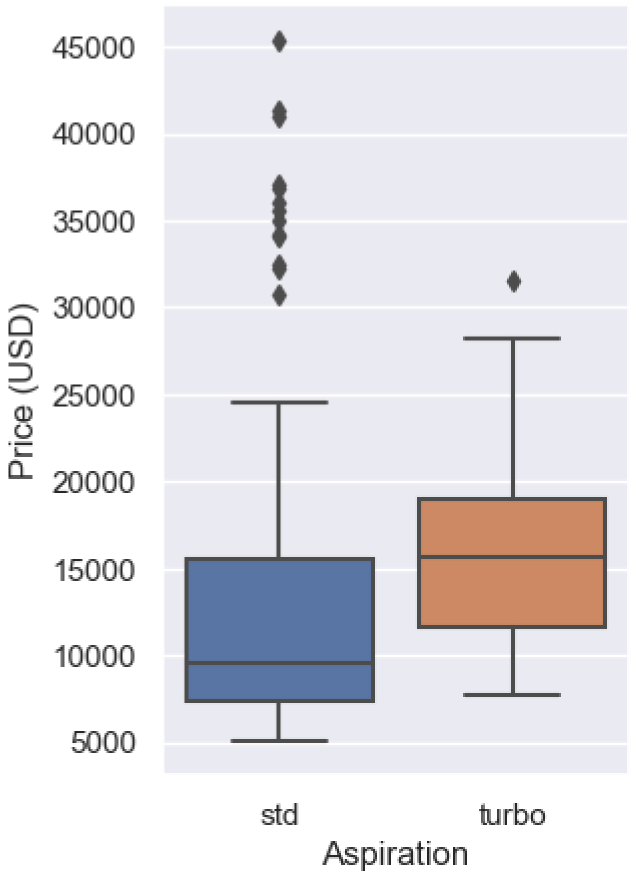
In Fig. 5, we can see that diesel cars in the dataset have a slightly higher average price than gas cars. However, there are a lot of outliers in the gas category with much higher prices than even the highest diesel car.

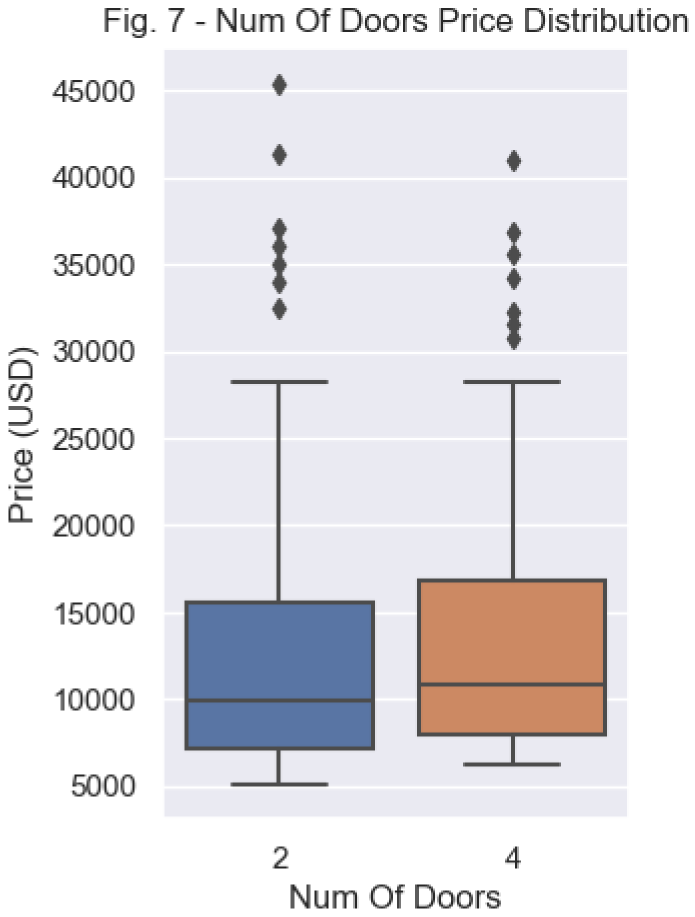
This might be a useful indicator of price, but there is not a clear conclusion that can be drawn at this stage.

In Fig. 6, we can see that vehicles with a turbo aspiration are generally pricier than the standard vehicles. Similarly to the 'fuel_type' feature though, we can see a lot of outliers in the standard category at higher prices.

As such I would say that we can draw a similar inference here, that while aspiration could well be a useful indicator, it is not particularly clear cut and some more investigation might be in order.

Fig. 6 - Aspiration Price Distribution





In Fig. 7, we can see that the number of doors on a vehicle looks to have very little effect on price, as the distributions between the two categories are very similar.

However, the median price for 4-door vehicles is slightly higher than for 2-door vehicles. This might be useful as an indicator, particularly in conjunction with other features such as 'body_style'.

In Fig.8, we can see a lot of variability between the different categories of body style. Hardtop vehicles and convertibles have the highest average price, and hatchback vehicles have the lowest.

This would certainly be a useful indicator for the price of a vehicle.

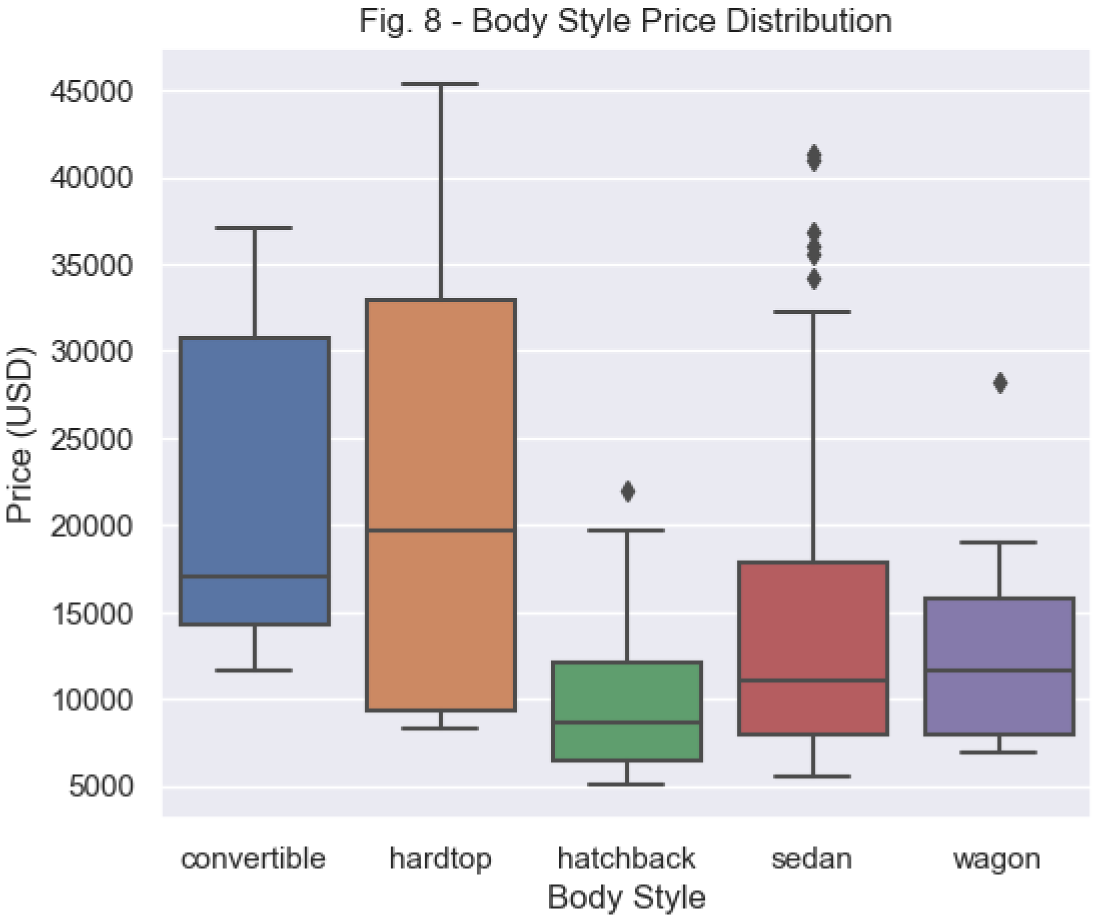


Fig. 9 - Engine Location Price Distribution

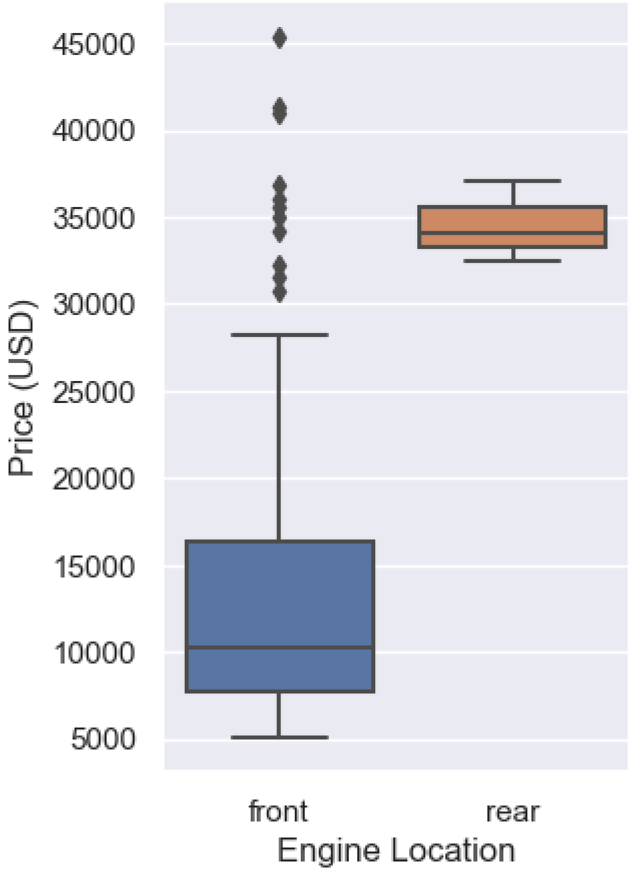


Fig. 9 shows a big difference in price between cars with different engine locations. Vehicles with rear located engines are, on average, much more expensive than those with front located engines

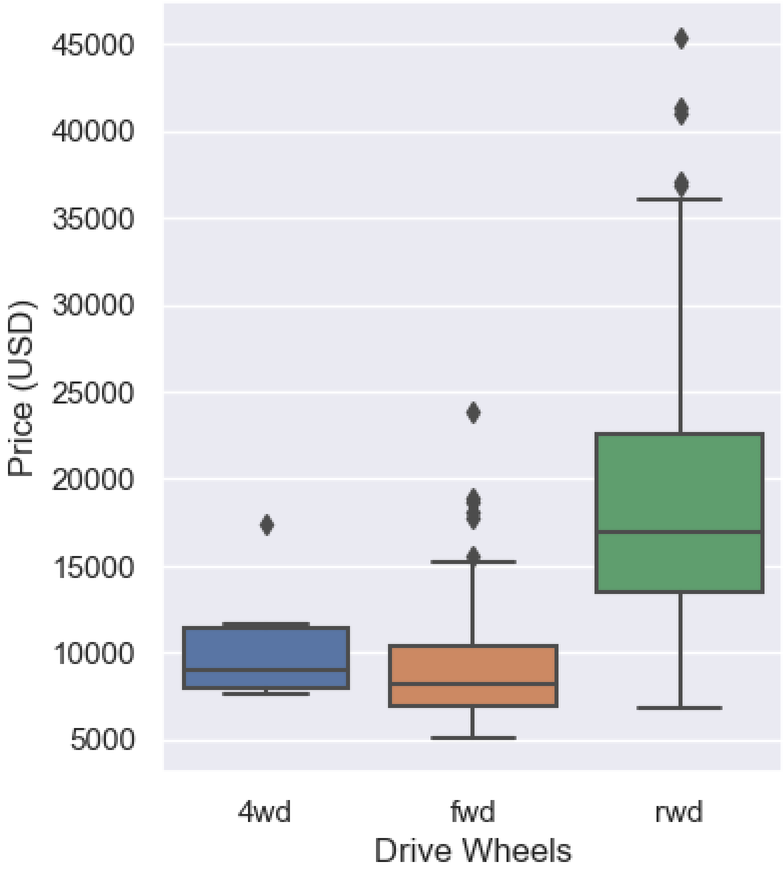
The distribution of vehicles with engines located in the rear is also much narrower which could mean that the sample size is small for these vehicles.

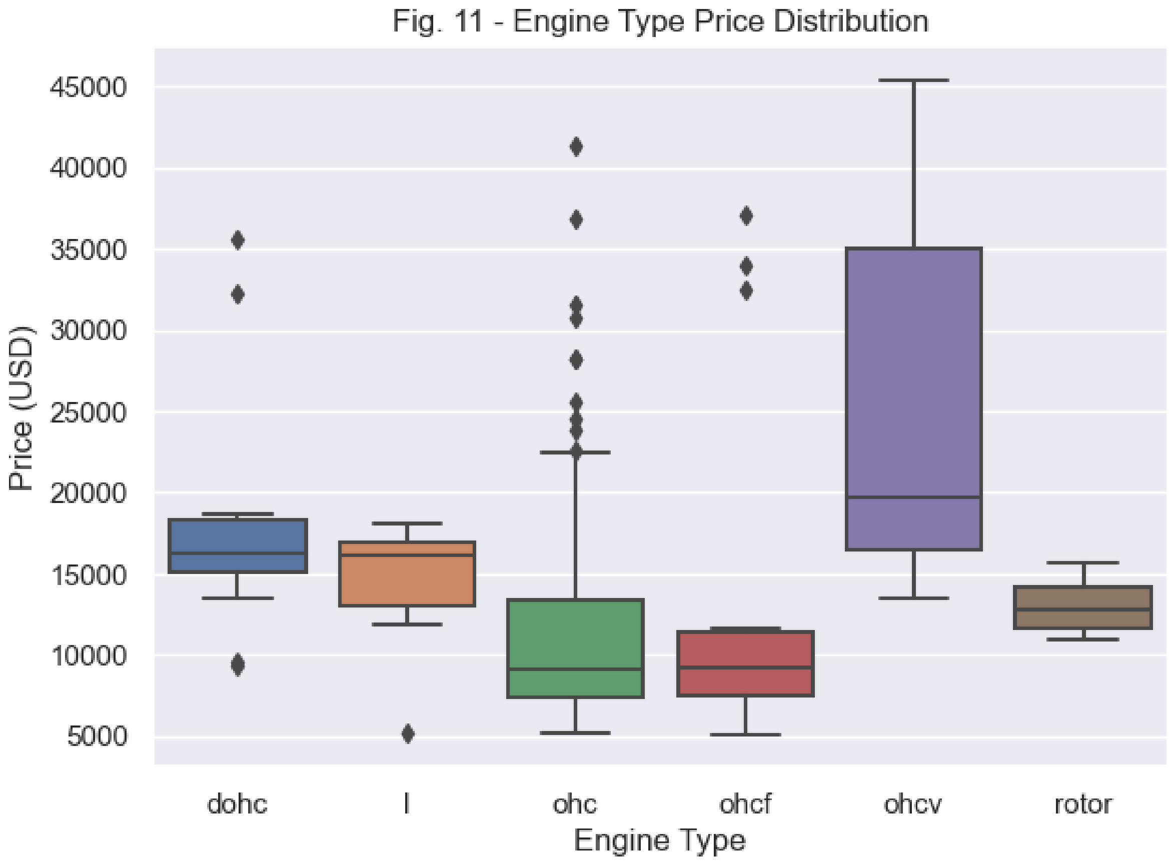
Looking deeper into the data, we can see that this is indeed the case, with only 3 of the 201 vehicles in the dataset having rear located engines. As such, this sample size might mean that this is not a reliable indicator and we might want to collect more data to try and include it in future analysis.

In Fig. 10, we can see that the drive wheels definitely could be used as an indicator of price, with rwd (rear wheel drive) vehicles having a much higher average price than 4wd (4 wheel drive), which in turn has a slightly higher average than fwd (front wheel drive).

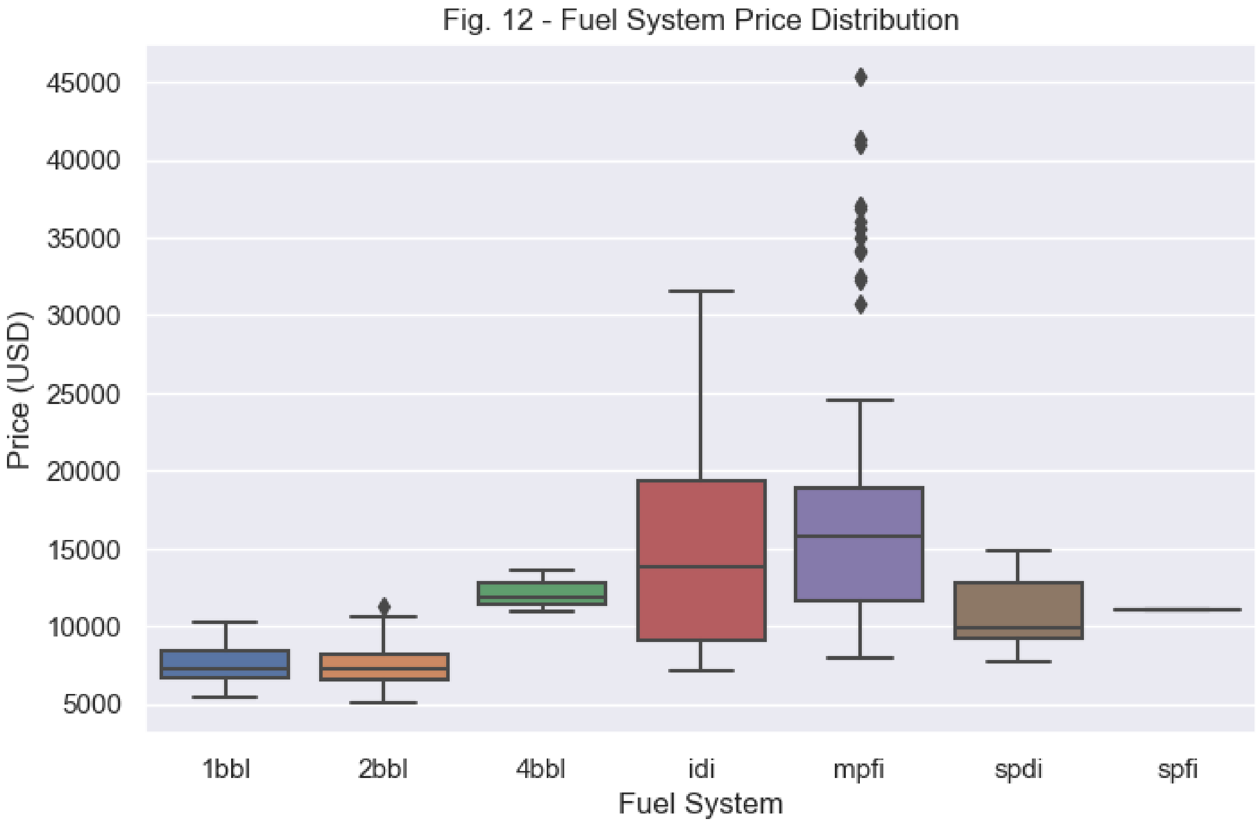
Rear wheel drive vehicles also have a much wider distribution than both four wheel drive and front wheel drive vehicles. Again this could be due to the number of vehicles in each category. Reviewing the data, we can see that there are only 8 vehicles in the 4wd category, but 118 fwd vehicles compared to 75 rwd. As such, I think this will be a very useful indicator for price.

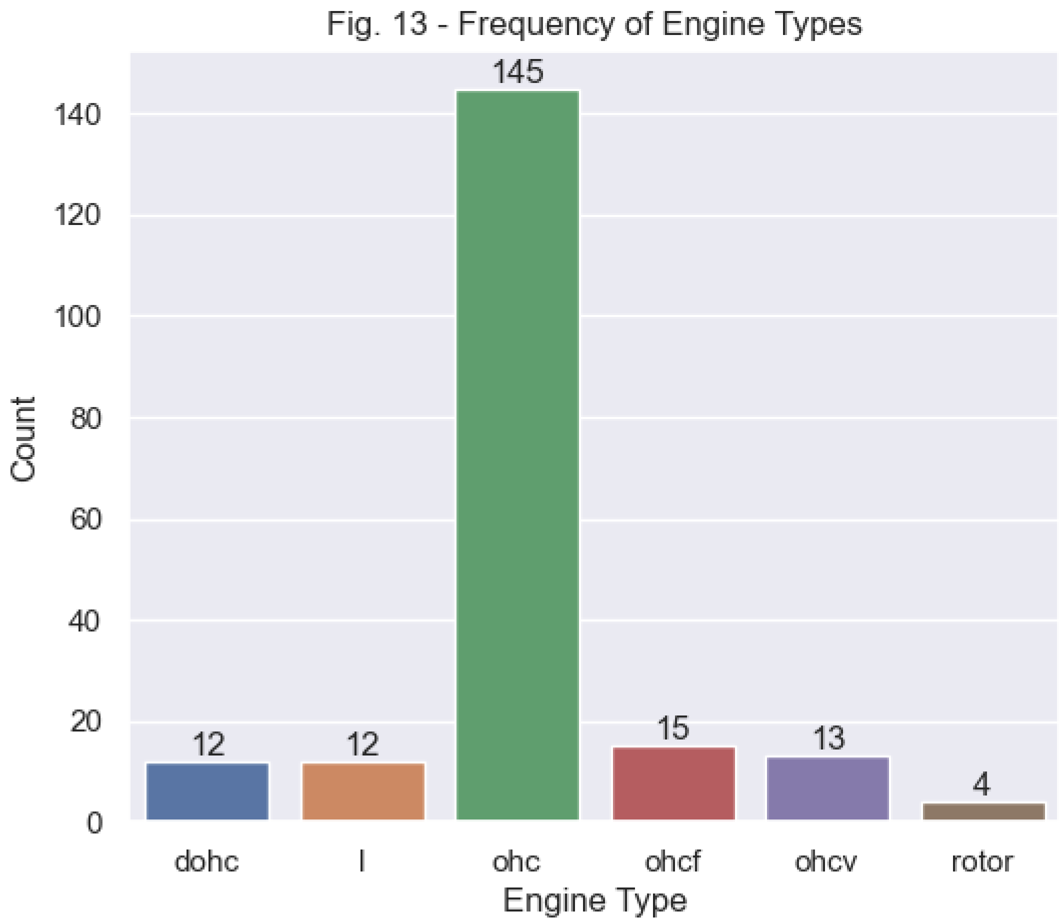
Fig. 10 - Drive Wheels Price Distribution





In Fig. 11 and 12, we can see that the various types of engine and fuel systems do seem to indicate price. We should again notice though that some of the distributions seem particularly narrow. As such, let's take a look at the frequencies for each of these categories as well. Note that there was one instance of the category 'mfi' in fuel_system originally and I changed this to 'mpfi' as they are synonymous.



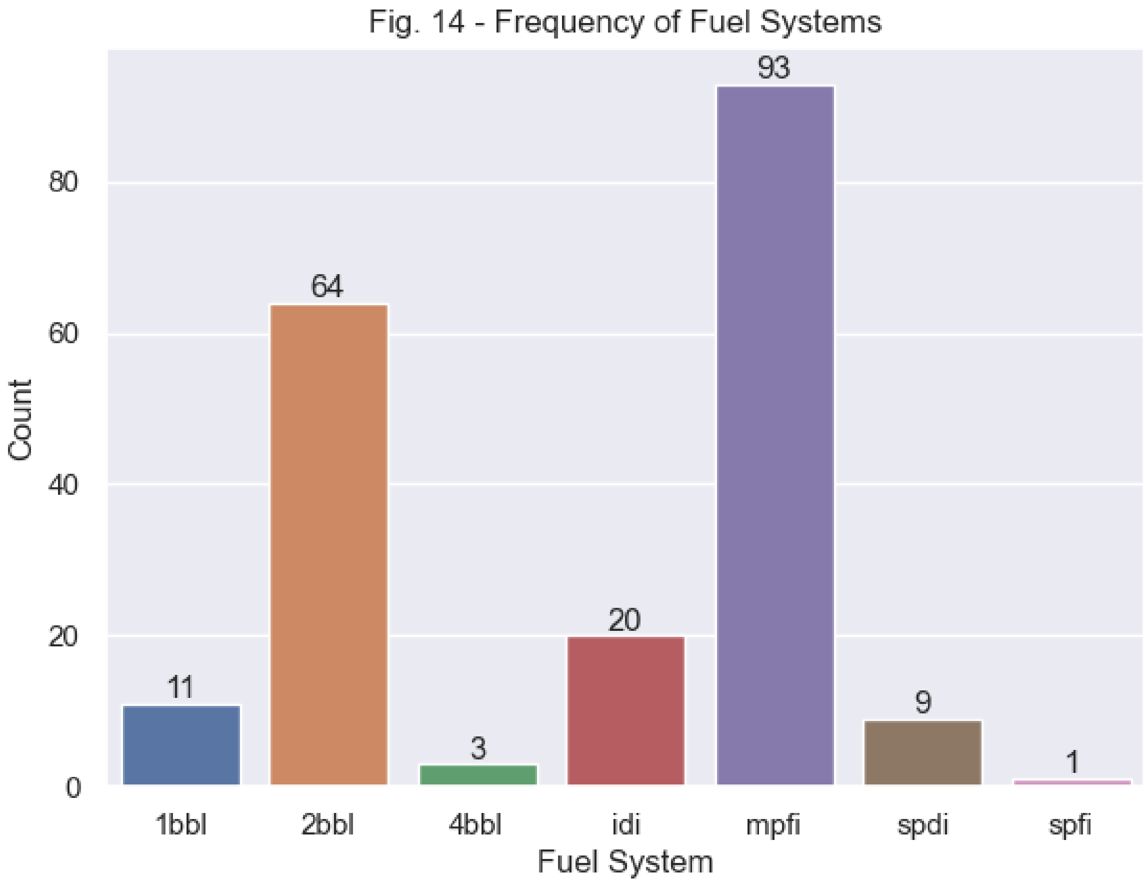


In Fig. 13, we can see that the vast majority of the vehicles have the 'ohc' type engine.

In conjunction with Fig. 11 though, we can see that there are enough vehicles in some of the categories with a differentiated price distribution that this would probably be a useful indicator for price.

Fig. 14 shows us that there are 2 categories with the majority of vehicles in them, mpfi and 2bbl.

It might make sense here to create an 'other' category for the fuel types with low counts.



Deeper Analysis

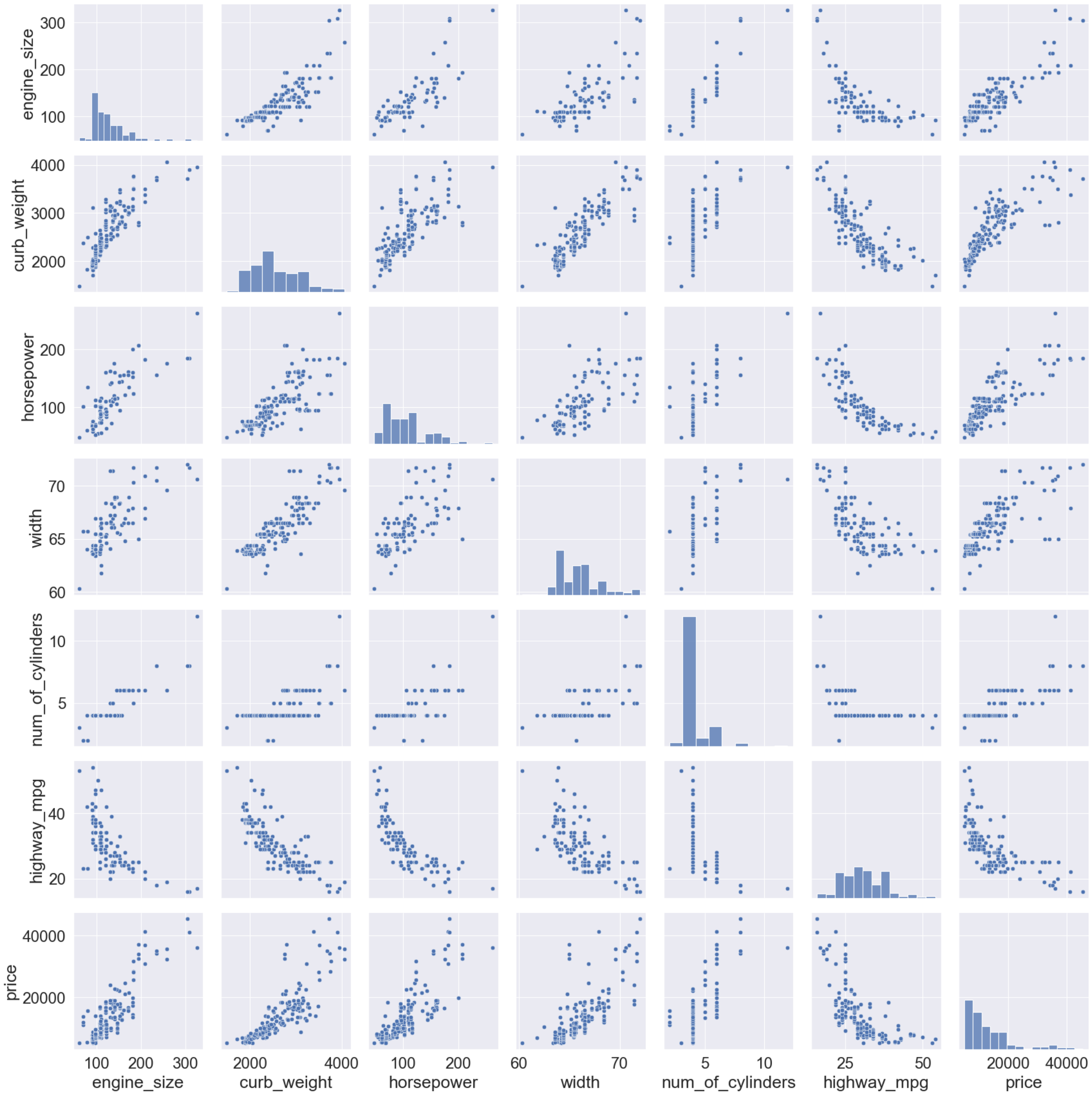
Now that we’ve had a look at some of the initial insights and gleaned some useful information about the data, let’s have a look at some of the features in more detail and see if we can visualise more of the relationships.

For the purposes of the rest of this analysis, I am just going to be looking at a subset of the features. These are: engine_size, curb_weight, horsepower, width, num_of_cylinders, highway_mpg, fuel_type, aspiration, body_style, drive_wheels and engine_type. I have chosen the numerical features that showed the highest correlation to price in my initial exploration. The categorical features I have chosen are those that show the most differentiation in the distribution of the price between categories, whilst still having high enough frequencies of each category for the insights to be relevant.

Numerical Features

First, let's have a look at the relationships between our chosen numerical features. We can plot all of them together on a scatterplot matrix to get an overview of the relationships and any potential multicollinearity. Then we can take the variables individually and search for any more insights in more detail.

Fig. 15 - Scatterplot Matrix for all Numerical Features



In Fig.15, although we can't see much detail at this point, we can get an idea of the relationships between the variables as well as a look at the distribution of each feature. We can see that each of the chosen features has a positive linear correlation to price, except for highway_mpg which has a negative one (as highway_mpg goes up, price goes down).

Engine Size

With each of these numerical features, it is useful to understand the distribution of the data, especially when it comes to thinking about creating a predictive model using them. So first let's have a look at the distribution of the engine_size data.

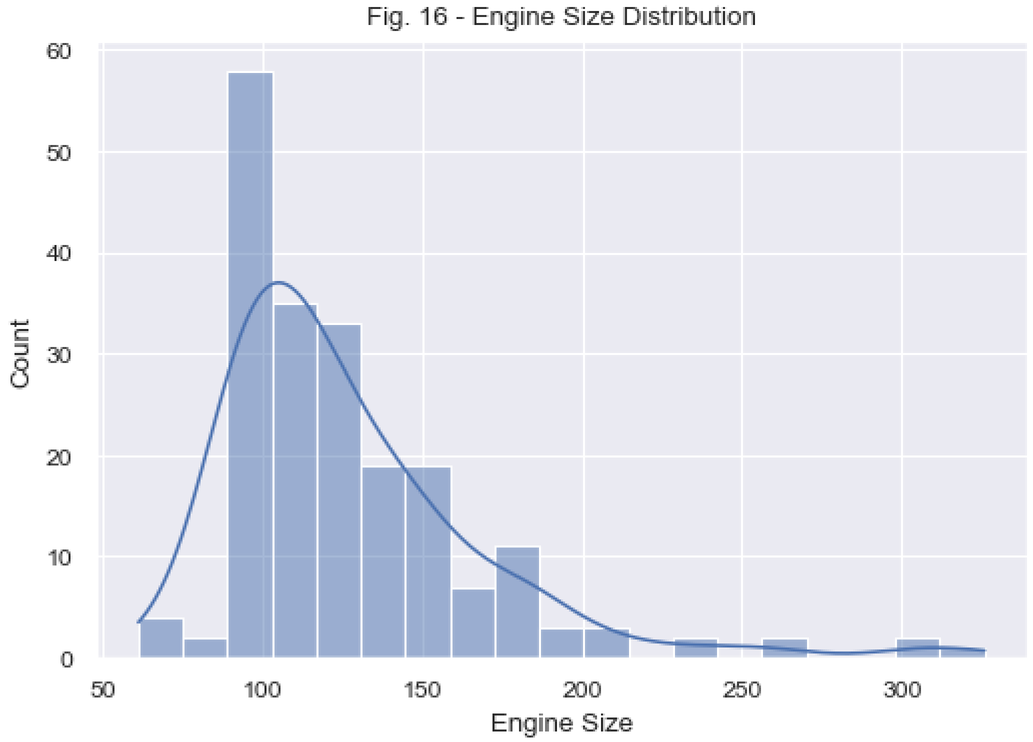
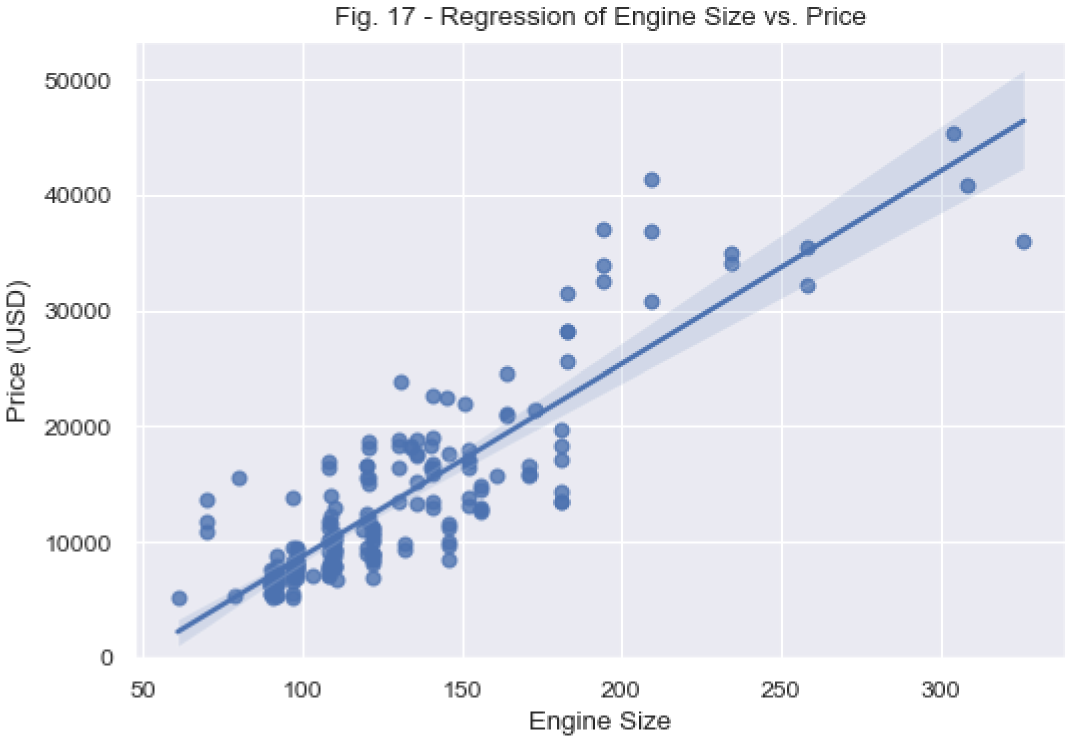


Fig. 16 shows us that the distribution of the engine size data is right-skewed.

As such, we might want to normalise this data before using it in certain types of statistical models to achieve more accurate results.

Fig. 17 shows us the linear relationship between engine size and price.

Due to the skewness of the data, we can see that the confidence interval indicated by the shaded area around the line of best fit is much wider at higher engine sizes and prices.



We can infer from this graph that as engine size increases, so does price.

Curb Weight

Curb weight refers to the total weight of a vehicle including a full tank of fuel and all standard equipment (not including any passengers or cargo). Let's do the same analysis as for engine size.

Fig. 18 - Curb Weight Distribution

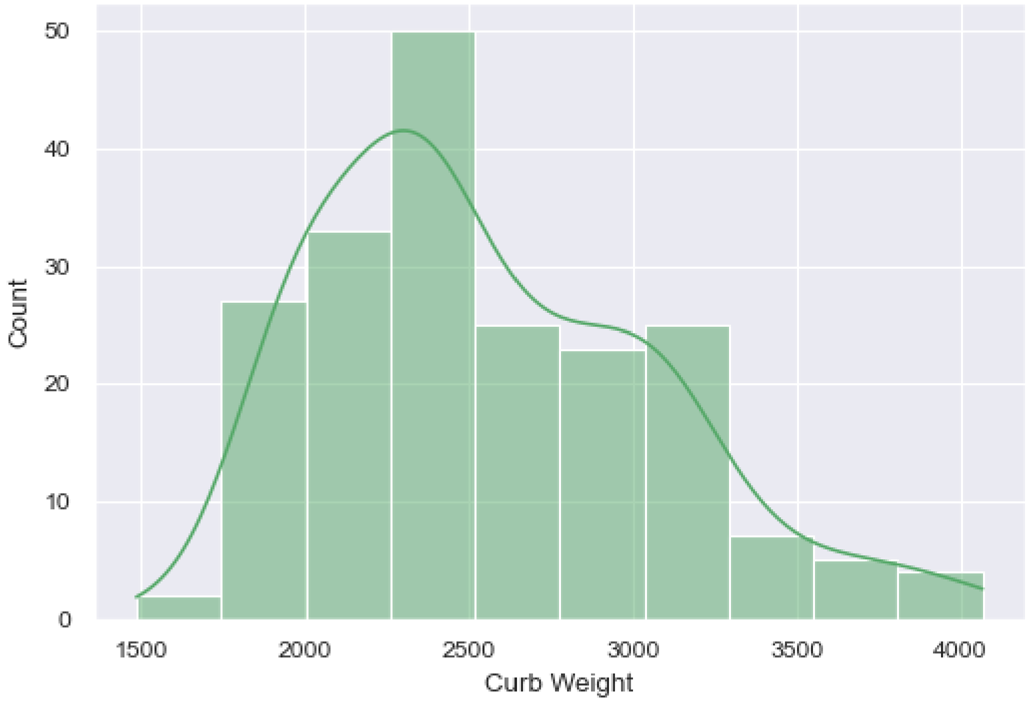


Fig. 18 indicates that the distribution of the curb weight data follows a more normal distribution than the engine size data, though still slightly right-skewed.

This would probably be even better represented, therefore, in a regression plot.

Fig. 19 - Regression of Curb Weight vs. Price

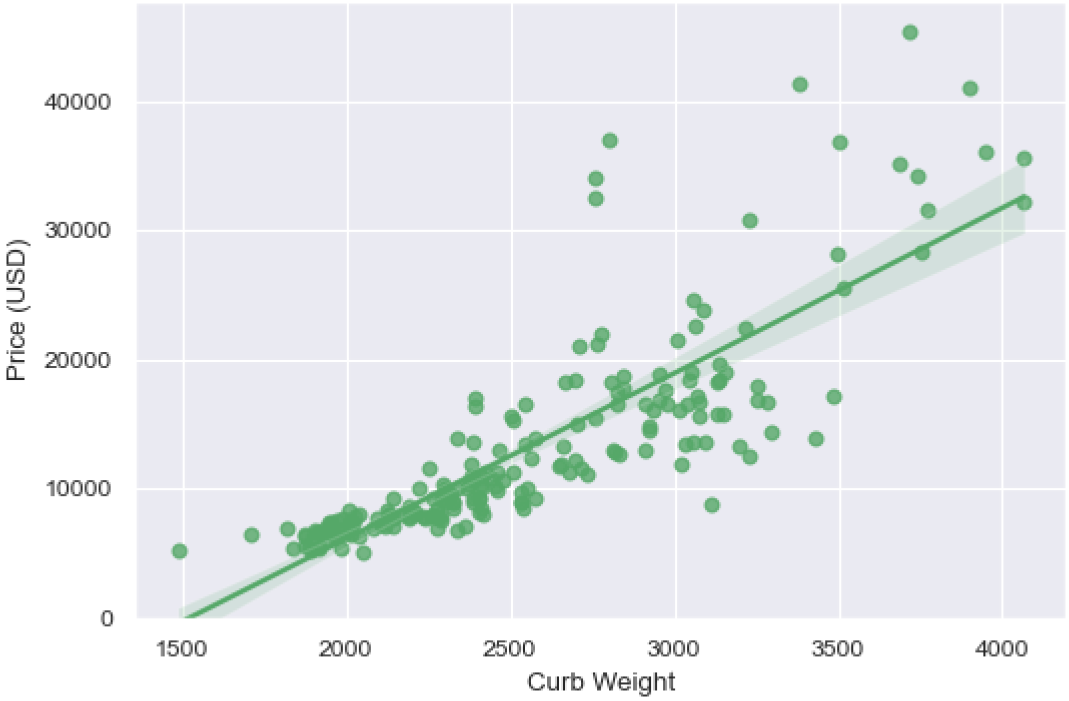


Fig. 19 shows a strong linear relationship between curb weight and price, especially at the lower end of the data.

There is again more variability as the price and curb weight increase.

We see a similar pattern here as with engine size: as curb weight increases, so does price.

Horsepower

Horsepower is an imperial unit of measurement used to measure the rate at which work is done, in this case the output of the engine.

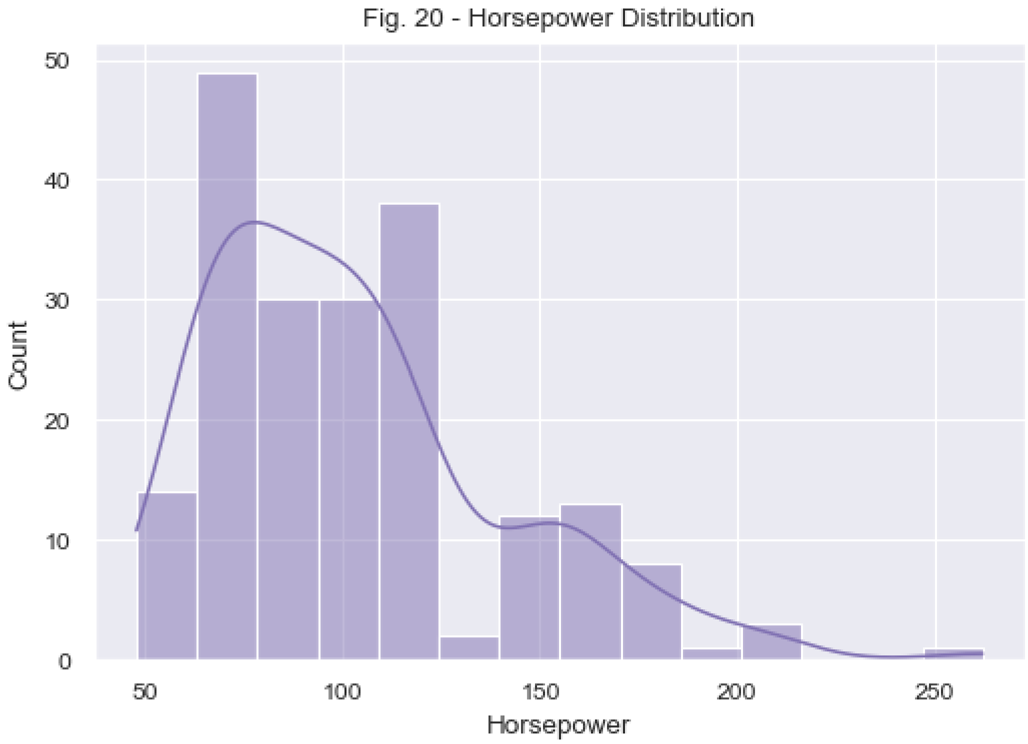
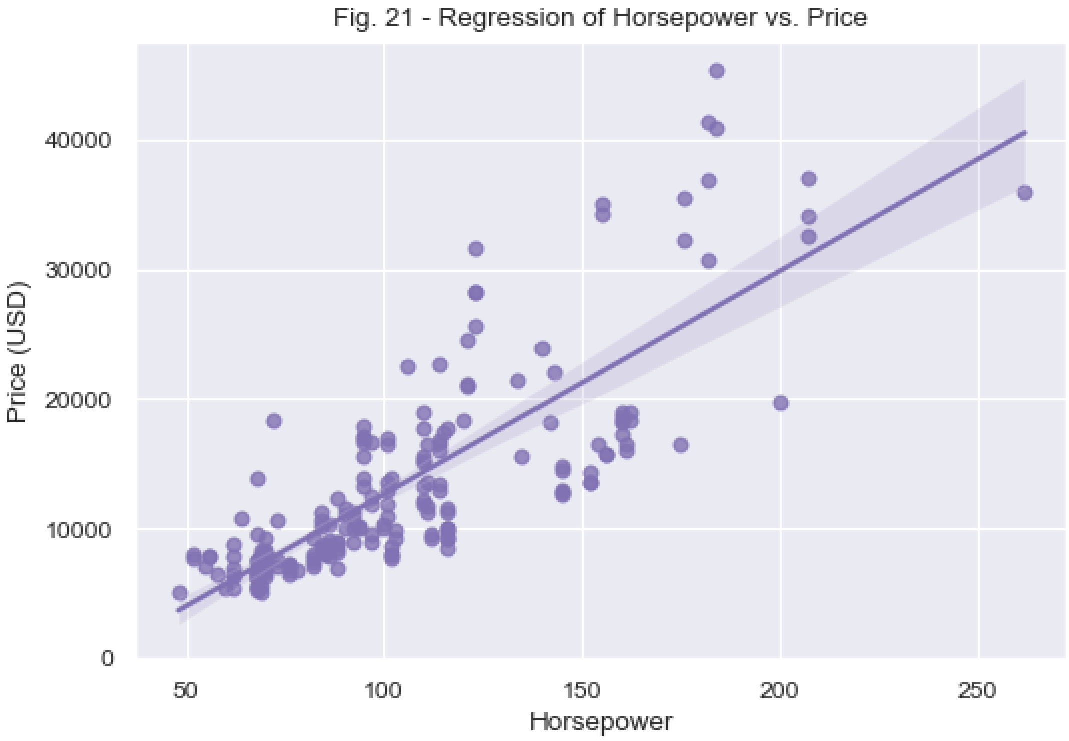


Fig. 20 shows us that the horsepower data does not follow much of a normal distribution. It is significantly right-skewed, has a main peak around 75 and multiple secondary peaks and troughs.

This likely won't be particularly well represented by a linear regression plot in its current form, but let's have a look anyway.



Whilst we can see a positive correlation in Fig. 21, we should again note that the confidence interval gets much wider as the price increases.

We are starting to see a pattern here, that the more expensive vehicles have more varied specifications.

We can again see that, in general, as horsepower increases, so does price.

Width

Width is a measurement of the width of the vehicle. Note that we are not going to be looking at length which was similarly correlated with price, but also likely heavily correlated with width.

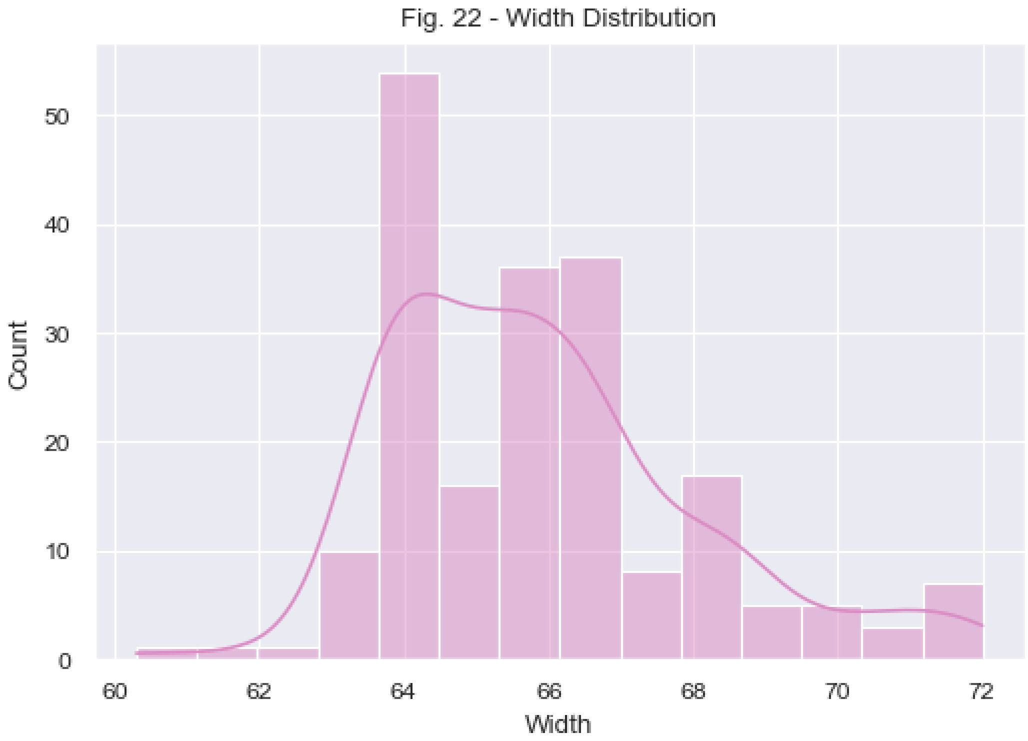


Fig. 22 suggests that the width data follows a much more normal distribution than we have seen thus far.

This means we will likely be able to get a clear idea of the data in the regression plot.

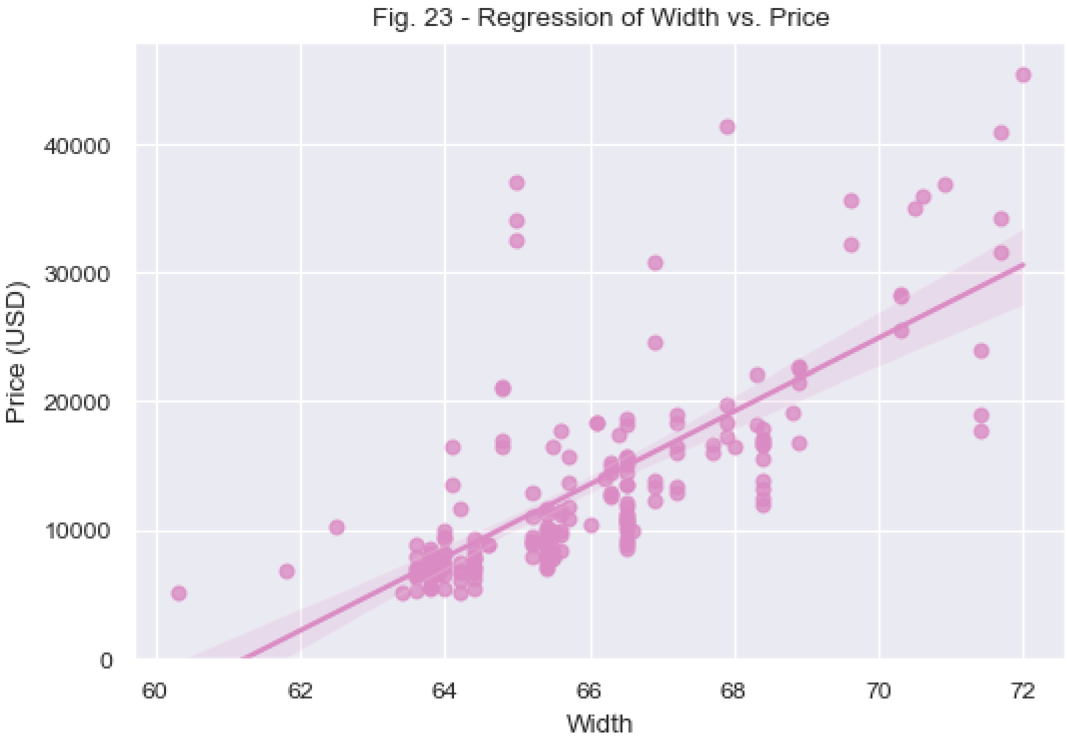


Fig. 23 shows us a clear linear relationship between the width of a vehicle and its price.

It should be noted though that there are quite a lot of outliers, especially at higher prices. We might want to look at that in more detail later.

So as with the previous features, we see that as width increases, so does price.

Number of Cylinders

A cylinder is a part of a vehicle engine where fuel is combusted and power is generated. This feature is limited in scope and could also be considered to be an ordinal categorical variable.

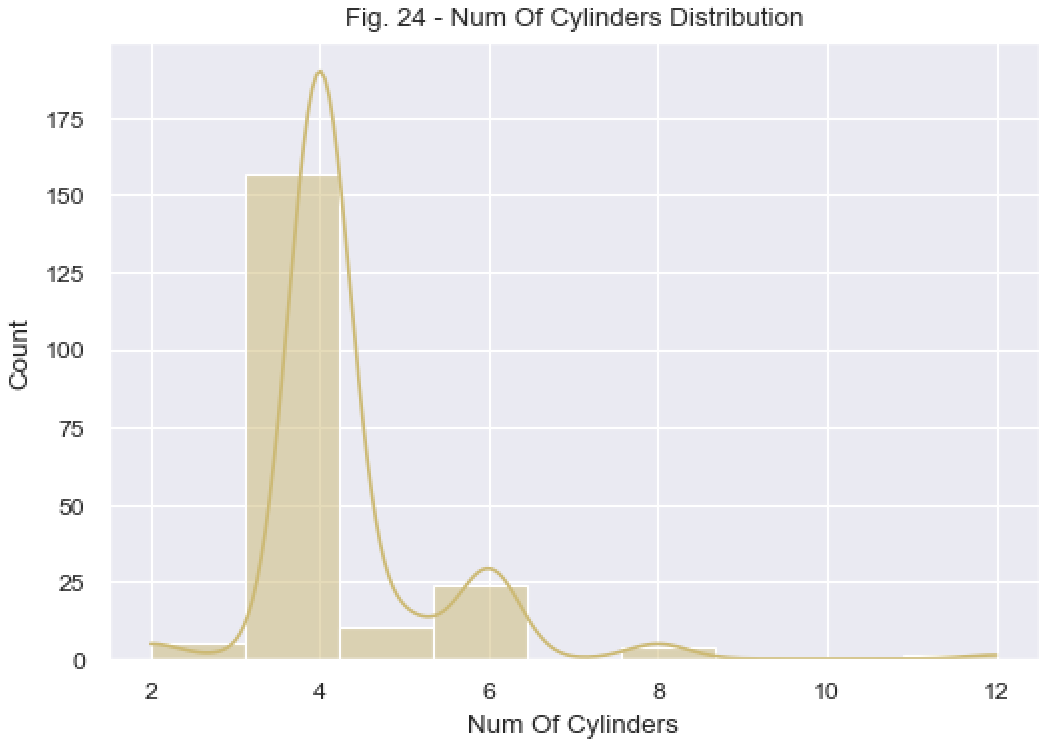
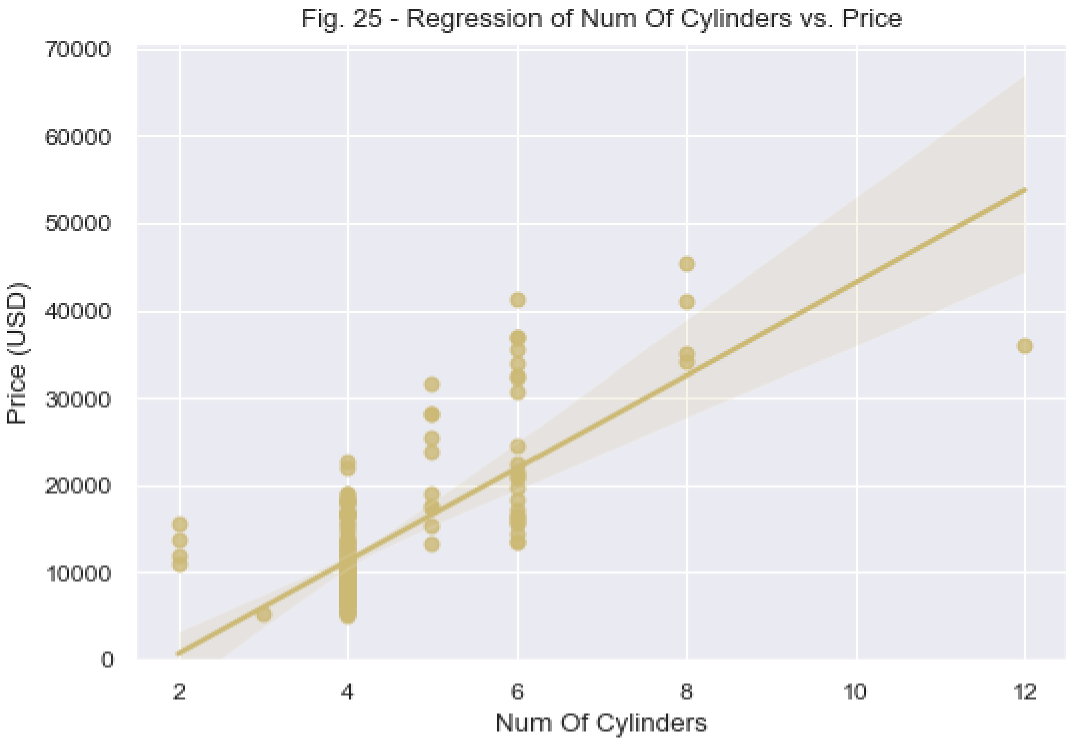


Fig. 24 shows us that the number of cylinders data is also skewed to the right. The vast majority of the vehicles have around 4 cylinders.

Still, let's have a look and see if we can discern a relationship on the regression plot.

As expected, Fig. 25, doesn't provide a great insight as to the relationship between the number of cylinders and price.

That being said, there is a strong positive correlation, albeit with a very wide confidence interval and a few extreme outliers.



We can say, yet again, that generally as the number of cylinders increases, so does price.

Highway MPG

Highway MPG refers to the average miles per gallon of fuel a vehicle will travel on a highway (motorway) without stopping or starting.

Fig. 26 - Highway Mpg Distribution

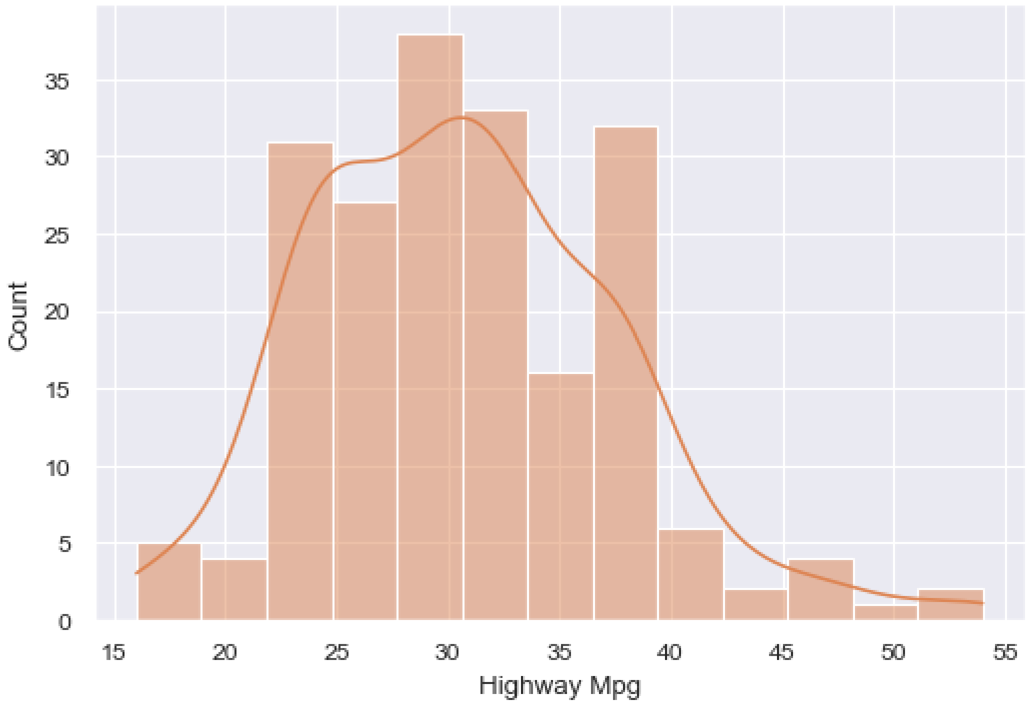


Fig. 26 indicates that the highway MPG data follows a mostly normal distribution that is slightly skewed to the right.

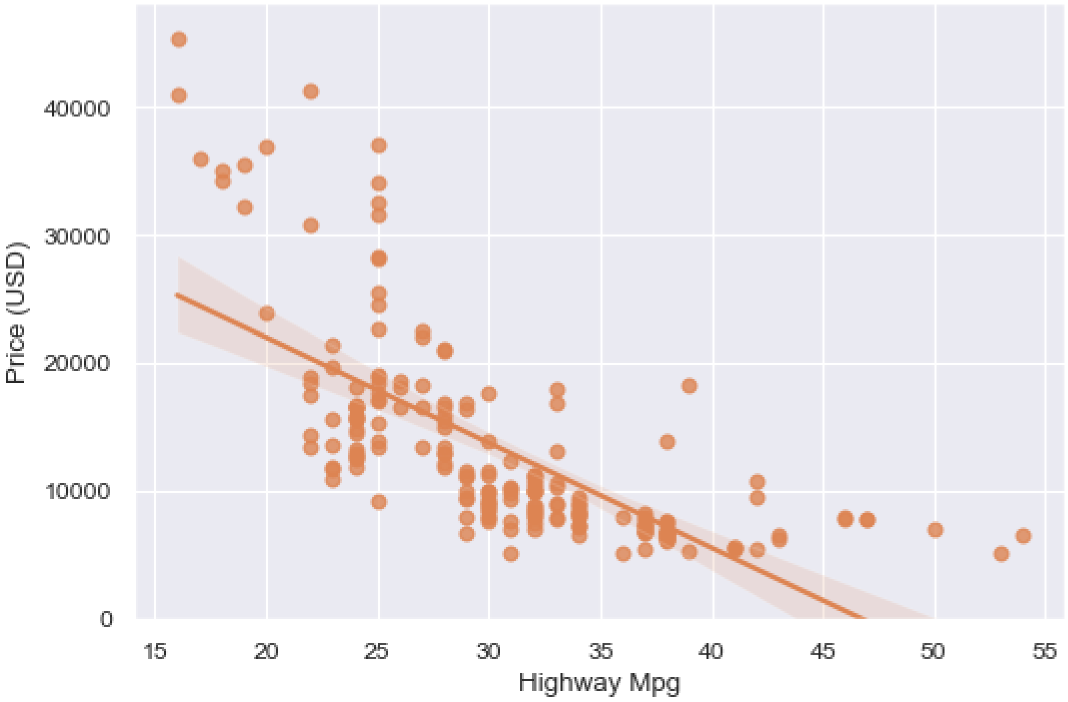
I would imagine this will be well represented by the regression plot. Let's have a look and see.

Interestingly, Fig. 27 looks like it doesn't actually have a linear relationship with price.

Rather, as there are outliers above the regression line at both the highest and lowest prices, it might be more accurately modelled with a polynomial regression model.

Still, we can see a clear negative correlation between highway MPG and price.

Fig. 27 - Regression of Highway Mpg vs. Price



In contrast to all the previous features, here we see that as highway MPG goes down, price increases.

Categorical Features

As we already had a look at the distribution of these features in the initial exploration, here let's have a look at the relationships between the categories and some of the numerical features in terms of price.

Fuel Type

A vehicle's fuel type is defined in this data as either gas or diesel. Gas or gasoline is also known as petrol and in this data makes up the majority of the entries.

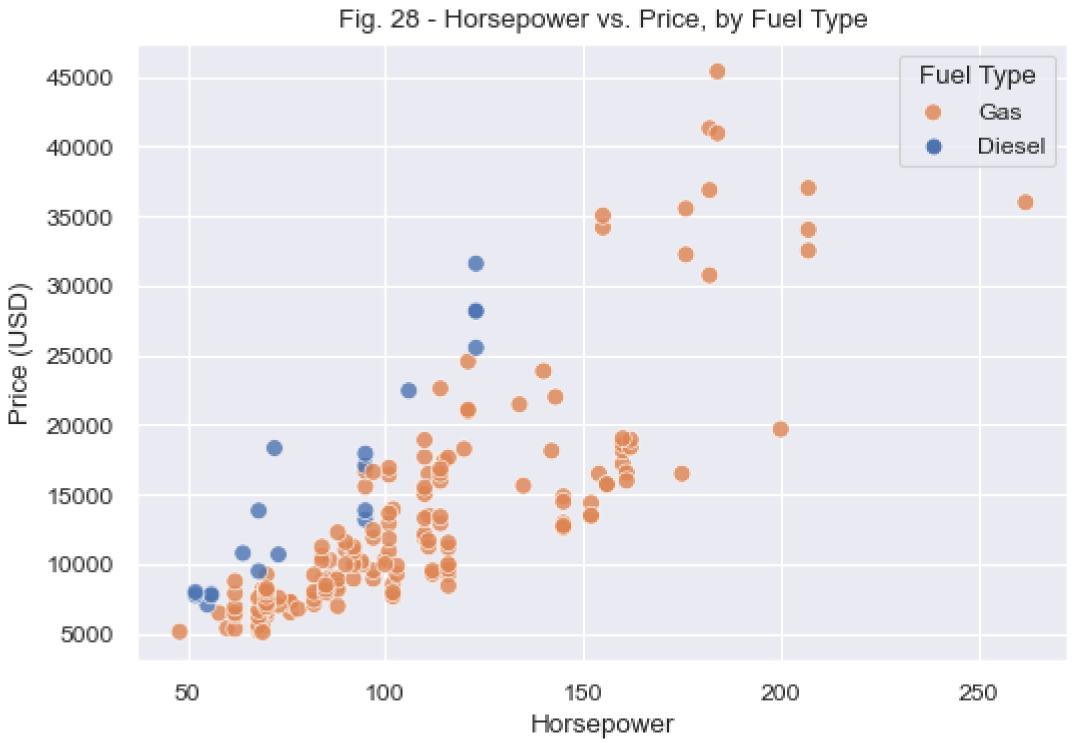


Fig. 28 shows the difference in price and horsepower between the vehicles with different fuel types.

We can see that diesel vehicles are generally more expensive than petrol vehicles with a similar horsepower.

There are far more gas vehicles than diesel vehicles, and none of the diesel vehicles have a horsepower above 125. The combination of these features would probably be useful when creating a model to predict price as while there are gas vehicles that are more expensive than the diesel ones, they also have a higher horsepower.

With more data, we may well see that this price trend continues for higher horsepower diesel vehicles, but with the limited data that we have here, we can only speculate as to whether or not that is true.

Aspiration

The aspiration of a vehicle refers to the way in which air is taken into the engine. Standard (std) means that the engine relies on atmospheric pressure for the air intake, while turbo engines use force induction to suck the air in.

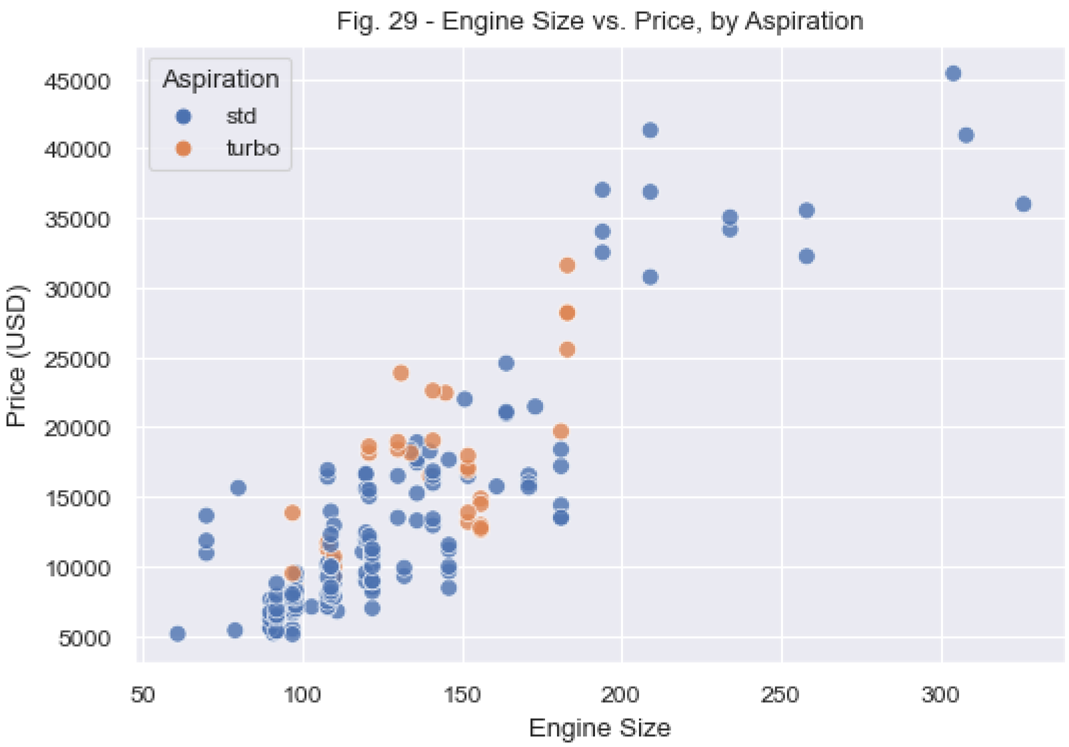


Fig. 29 shows vehicle's engine sizes compared to their price with reference to the aspiration.

While not as clear as we saw with fuel type and horsepower, we see a similar pattern here. Vehicles with a turbo engine are often more expensive than standard vehicles with comparable engine sizes.

The majority of the vehicles in the dataset use the standard aspiration method and none of the turbo vehicles have engine sizes above 200. Similarly to the previous combination we looked at, these features together might play an important role in price prediction.

We saw earlier (Fig. 6 pg.8) that the turbo vehicles have a higher median price, and that the most expensive standard vehicles could be considered outliers. With more data, again we might see that this pattern continues and that turbo vehicles with larger engines are even more expensive, but we can't make conclusions about that given the sample.

Body Style

A vehicle's body style is a way of categorising vehicles based on their design, shape and space. The various categories in this data represent a range of vehicle shapes and sizes displayed in the graph below.

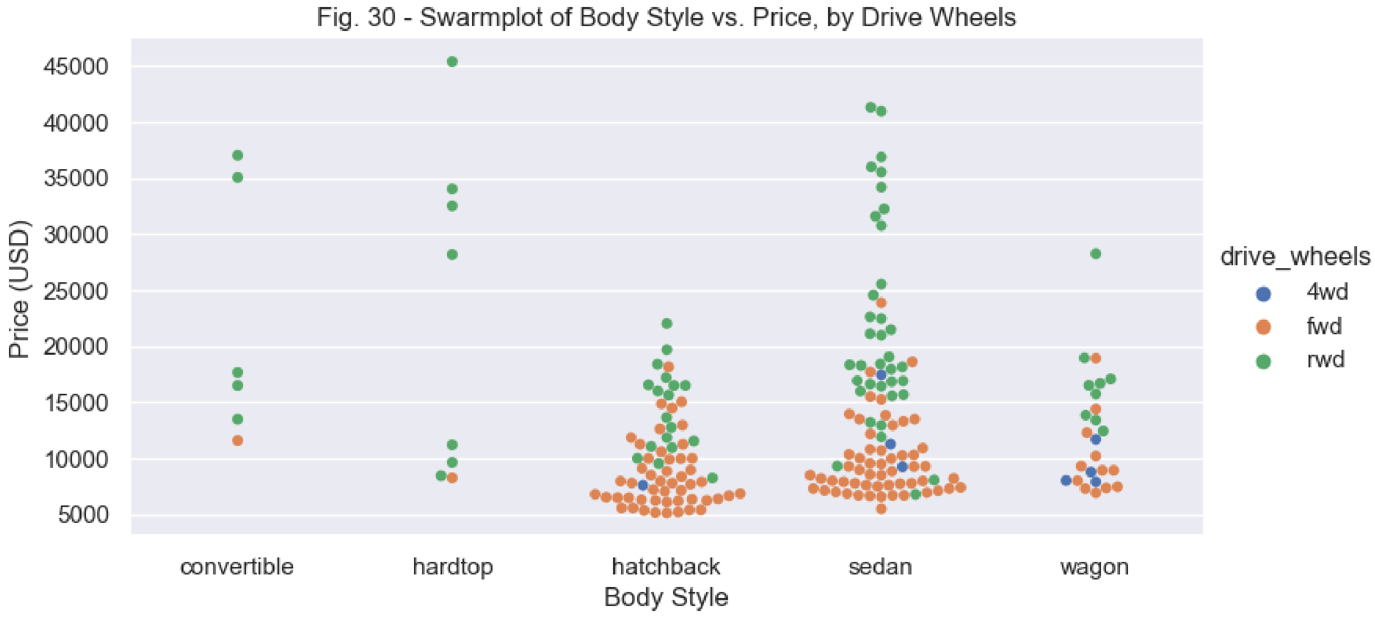


Fig. 30 shows the price distribution of the different body styles and also references the drive wheels of the vehicles. We can see that front wheel drive (fwd) vehicles in every category are the cheapest vehicles.

It is also worth noting that hatchbacks have less definition in the boundary between front wheel drive and rear wheel drive (rwd) prices, compared to sedans for example. This seems to be due to the fact that the most expensive sedans are all rear wheel drive and there are no hatchbacks in that price range.

There are very few hardtops and convertibles in the dataset and the price distribution of these categories is very large, this makes it difficult to make accurate predictions based on this data without collecting a larger sample. We can see however that the cheapest vehicle in both these categories is front wheel drive.

As such, it seems that this would definitely be a good indicator of the price of a vehicle and is potentially even more useful when used in conjunction with another feature such as drive wheels. Let's have a look at drive wheels in more detail now in conjunction with a different feature.

Drive Wheels

A vehicle's drive wheels refers to the wheels that the engine is powering. The data is divided into 3 categories: FWD (front wheel drive), RWD (rear wheel drive), and 4WD (4 wheel drive). There are very few 4WD vehicles in the dataset but the difference in price between FWD and RWD vehicles is noticeable.

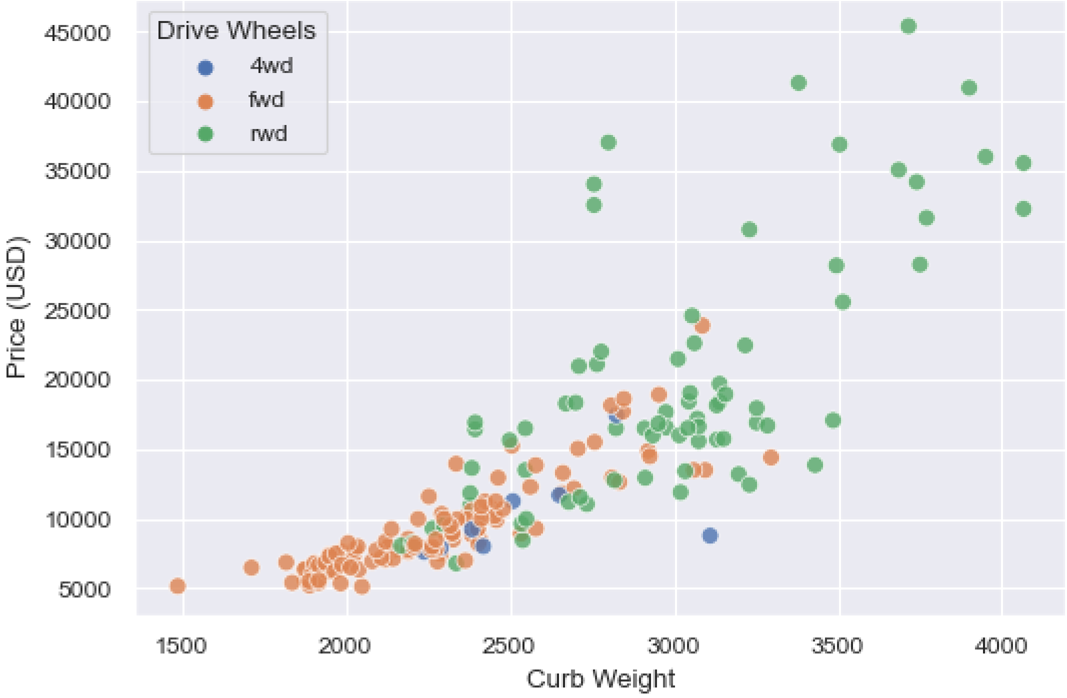


Fig. 31 shows the RWD vehicles are generally both heavier and more expensive than FWD vehicles.

Interestingly, while we usually think of 4WD vehicles being large and therefore heavy, they are not among the heaviest in this dataset.

We can see a clear boundary between FWD and RWD vehicles with very little in the way of overlap between the 2 categories when it comes to price. This would therefore be an excellent feature to use for modelling purposes.

We can also hypothesise as to why the RWD vehicles are generally heavier. Perhaps this has something to do with RWD vehicles needing more weight to prevent wheel spin, particularly in cases where the weight of the car is unevenly distributed (if the engine is in the front of the car for example). My domain knowledge on that matter is sadly insufficient to be able to draw this conclusion, but it could be something we could look into further in the future.

Since, "drive wheels" seems to be such a good indicator of price, let's have a look at it one more time alongside our final categorical variable, engine type.

Engine Type

A vehicle's engine type generally refers to the placement of the camshaft within the engine. There are multiple engine types in this dataset and I do not profess to be an expert on how they all work, but that doesn't mean we can't have a look and see if we can draw some conclusions from the data.

I have tried to find out what the abbreviations actually stand for, but I was not able to come up with answers for all of them. So for the purposes of this analysis, we can just refer to them in the abbreviated form they are currently in.

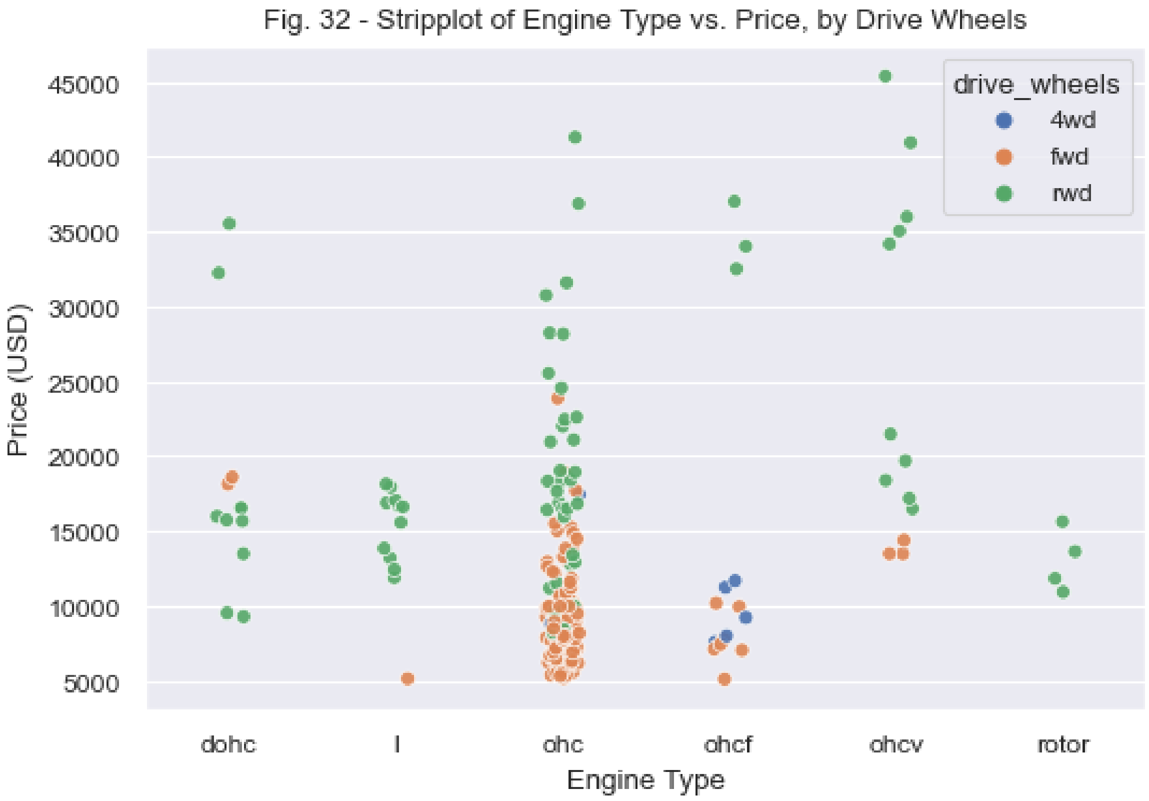


Fig. 32 shows us that the vast majority of FWD vehicles have the OHC engine type.

We can also see that all of the 4WD vehicles have the OHCF engine type.

We can again see a distinction in price between the FWD and RWD vehicles, but it is interesting to note that for vehicles with the DOHC engine type, this relationship is not quite so clear cut. It is a small sample size, but in contrast to every other category we can actually see here that FWD vehicles in the DOHC category are not the least expensive vehicles.

We should also note that the majority of the cheapest vehicles in the dataset have OHC or OHCF engine types, and that OHCV vehicles are generally the most expensive.

As such, engine type could indeed be a useful indicator for helping us to fine tune predictions regarding the price of vehicles in the dataset as it gives us even more precision around the observation we had already made about drive wheels.

Conclusion

After thoroughly cleaning the data, making decisions about missing values, and going through an exploratory analysis of various features and their relationships to each other (particularly in reference to the price independent variable), I would say that the following are the most useful insights:

- Many of the numerical features in the dataset are highly correlated with price. The features with the highest correlations are engine size, curb weight, and horsepower.
- For the majority of the correlated numerical features, the correlation with price is a positive one. For example, as engine size increases so does price.
- Highway MPG is negatively correlated with price, as it goes up, price goes down.
- Vehicles with a diesel fuel type tend to be more expensive than comparable gas vehicles. The mean diesel vehicle price is 22.6% more than the mean gas vehicle price.
- Turbo aspirated vehicles are generally pricier than the vehicles that use standard aspiration, though this relationship is less clear cut than the one for fuel types. Despite that, the mean price for a turbo vehicle is 29.6% more than that of a standard vehicle.
- The most expensive cars have the convertible, hardtop, or sedan body styles. Hatchbacks have the narrowest distribution of price data and also the lowest median price at $8,672, more than $2,400 less than the next lowest (sedan).
- A vehicle's drive wheels is one of the strongest indicators of price, with the median price of a RWD vehicle being $16,900, more than twice as much as that of a FWD vehicle at $8,192.
- Vehicles with the OHCV engine type were generally the most expensive with a median price of $19,699, over $3,000 more than the second highest (DOHC).

Next Steps

There are many more things that we could do with this data, but here are some of the things that I think could be done to achieve greater insights:

- Collect more data. The sample size here is small, we only have 201 entries after dropping the 4 that had no price data, and while this was enough to glean some insights, more data could improve the quality of our inferences greatly.
- Consult someone with better domain knowledge. The data could well be cleaned more comprehensively with the help of someone who understands more about the engine types for example.
- Test for multicollinearity. Much of the data follows a similar pattern compared to price. As such, I would hypothesise that many of the features could be highly correlated with one another. We would want to explore that in more detail before trying to build a predictive

model, particularly a linear regression model that would not handle the correlated features properly. By testing for multicollinearity, we could then determine which distinct features best indicate price and build a suitable model.

- Normalise the data. The majority of the numerical features do not follow a normal distribution, so if we were to take this dataset and try to use it for modelling purposes, we would want to further clean the data by normalising most of the numerical columns, to improve the efficacy of the machine learning process.
- Group smaller categories together. Given the small sample size, it might make sense in some cases to group some of the smaller categories together into one “other” category, for example in the fuel system feature.

Final Thoughts

Though small, this dataset contains information that could well be used to help predict vehicle prices given a set of specifications about a vehicle. We optimised the data types for speed of analysis and were able to show that huge reductions in memory usage are possible, which would be especially useful if we were to have a larger dataset than this one.

Through explorations of many of the features, we have found relationships and ideas, as well as specific conclusions, that might not have been immediately obvious and could well be key insights in further exploration or future predictive modelling.