Introduction

The given data set describes the specifications of different cars based on ‘symbolling’, ‘normalized losses’, ‘make’, ‘fuel type’, ‘aspiration’, number of doors’, to name a few. This data are divided into 26 columns and 206 rows. The information gleaned from this data set could reflect how the car’s different specifications affect the price, the body style and more specific values that make a car a car. Before diving into the details of the information gathered from this data we have to first clean it, fix data losses and such. When receiving new data there’s always some data loss. The next list describes the percentage of data loss in columns that experienced it.

* normalized-losses 20.0% data loss
* num-of-doors 0.98% data loss
* bore 1.95% data loss
* stroke 1.95% data loss
* horsepower 0.98% data loss
* peak-rpm 0.98% data loss
* price 1.95% data loss

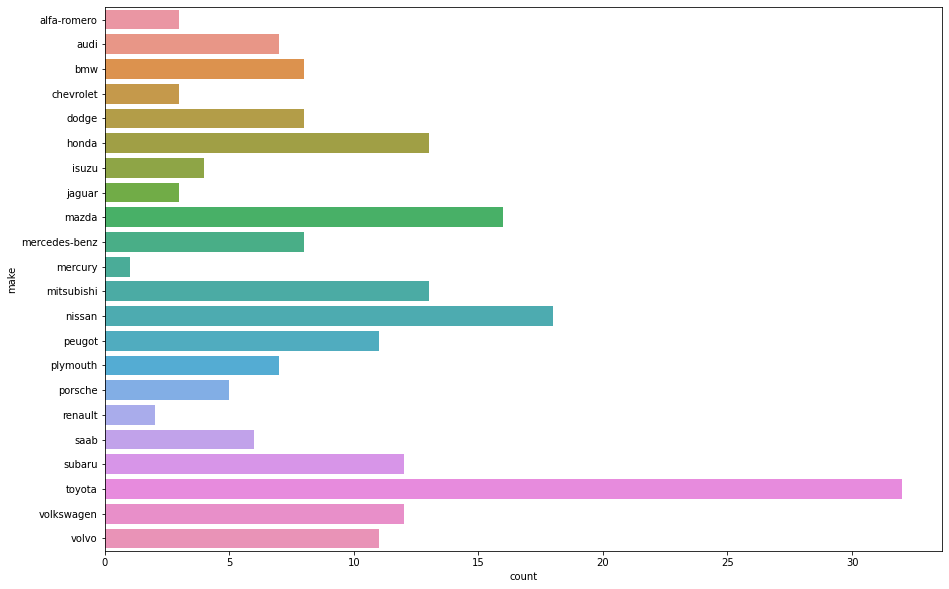
As we can see the data shows some small data loss with and average data loss of 4.11% data loss. With the max data loss being on the normalized losses column. ‘num-of-doors’ describes the number of doors for any For every loss percentage value less than 10% I will use mean value imputation to fill in the not available values. The reason this is done is because the low amount of data that is missing won’t affect the distribution of the data as much as it would if there was more data missing. After learning better imputation strategies I will use those to better deal will data loss. ‘In statistics, **imputation** is the process of replacing missing **data** with substituted values. When substituting for a **data** point, it is known as "unit **imputation**"; when substituting for a component of a **data** point, it is known as "item **imputation**".’ (Wikipedia, 2021)

The number of doors didn’t describe any data that could had a numerical mean so instead of using the mode as the mean and because there are 2 values that are missing in the data set I tried to google the cars and find the missing car door values but I soon realized that that is near impossible to find so... they will be removed from the data set. This kind of kills 2 birds with one stone because one of the values has a missing normalized losses value anyway.

To explain a few of these parameters and others are more obvious but, ‘An engine's bore is the diameter of each cylinder, while the stroke is the distance within the cylinder the piston travels.’ (Silverstro, 2020)

“Normalized-losses” is described as the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door, small, station wagons, sports/specialty, etc...), and represents the average loss per car per year. (SriramMscDA, 2018) This column also has 20% data loss which is bad and makes it really diﬀicult to delete rows of the column because that’s 20% of our data. It seems relatively important so I will be dropping the column for our normal data but will create a new data set with the 20% data gone and do some analysis on it to see some things. I think the column will really show some interesting information.

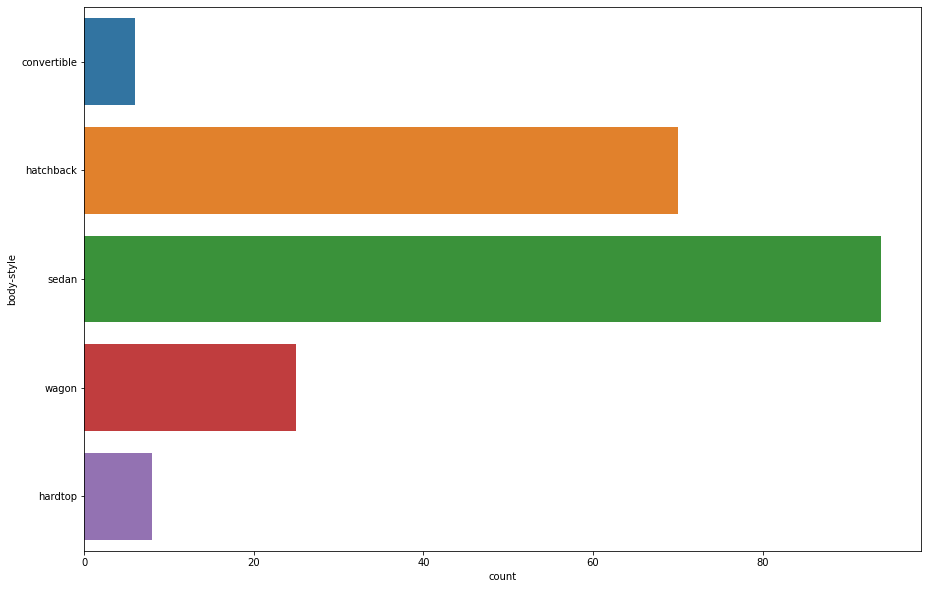
Now that our data is clean and has no missing or wrong values, it is clear skies to do some cool analysis on this data and see some cool new things about the data that we have not seen before.



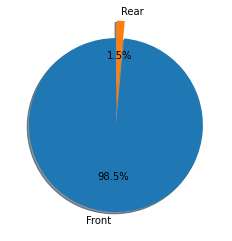
Here we look at the make of vehicles in the dataset we have been given to understand the distribution of cars in the given data set. As we can see there are more Toyotas than any other car in this data set. The Data is seriously saturated with Toyota vehicles.



This is a comparison of two door cars to four door cars in our data set. This will be some valuable information in the future because now we can know how they affect car price and or car classification, body style wise.



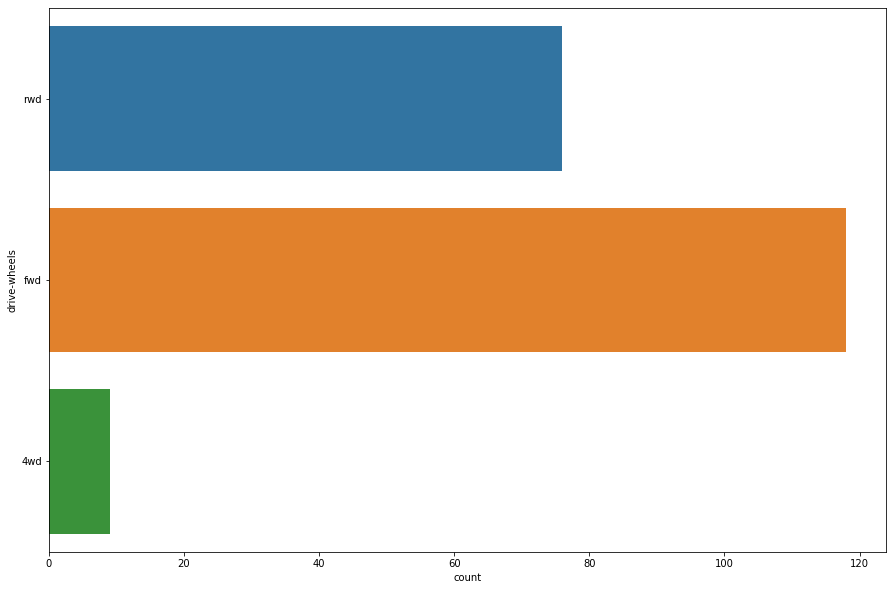
This is a histogram of the frequency of the different body styles that exist in the data set. This is a very prominent value in out data because It will show how the different types of the body styles affect the distribution of certain values in the data set.



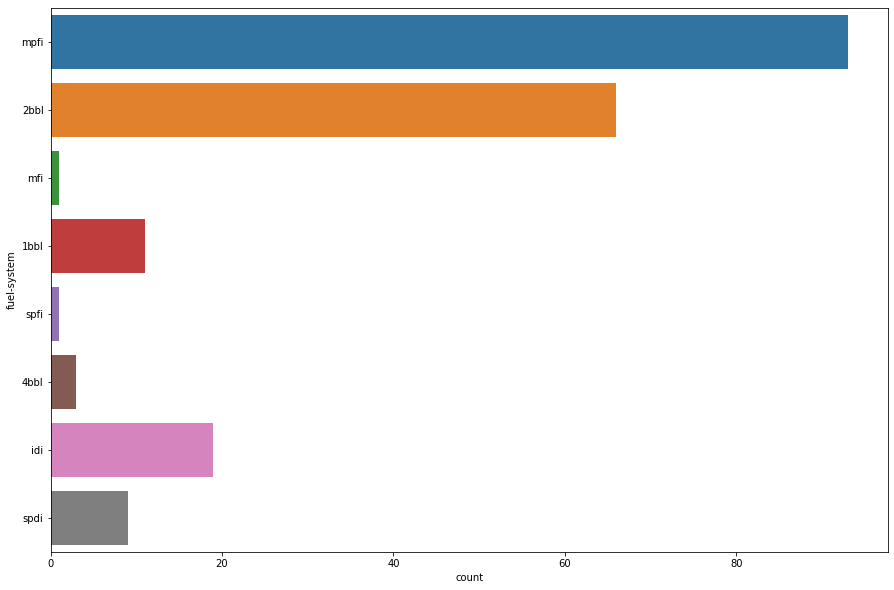
The amount of cars with a front side engine to cars with a rear side engine is appalling, I wonder how these things affect the prices of the cars.



The different fuel types will affect the price and how much power the engine outputs so this is a histogram to show how many different types of fuel types there are in this data set.



This is a histogram that describes the different ways in which the engine controls the wheels. They are either four wheel drive, front wheel drive or rear wheel drive. This affects the speed at which the car travels which in our data is describe my the city / highway mile per gallon.



This is the histogram of all the different fuel systems used in the cars.

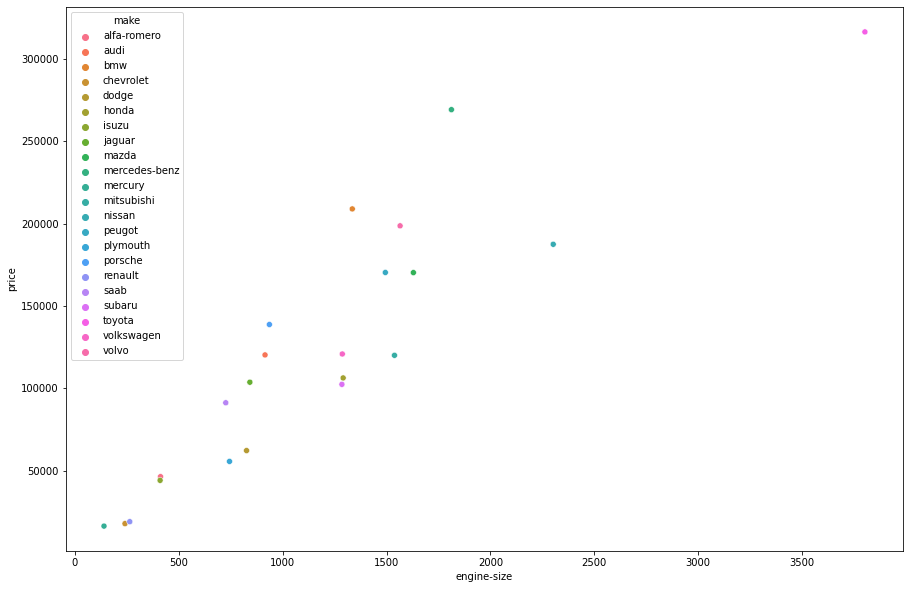
Top 10 most expensive cars

1. mercedes-benz
2. bmw
3. mercedes-benz
4. porsche
5. bmw
6. jaguar
7. jaguar
8. mercedes-benz
9. mercedes-benz
10. Porsche

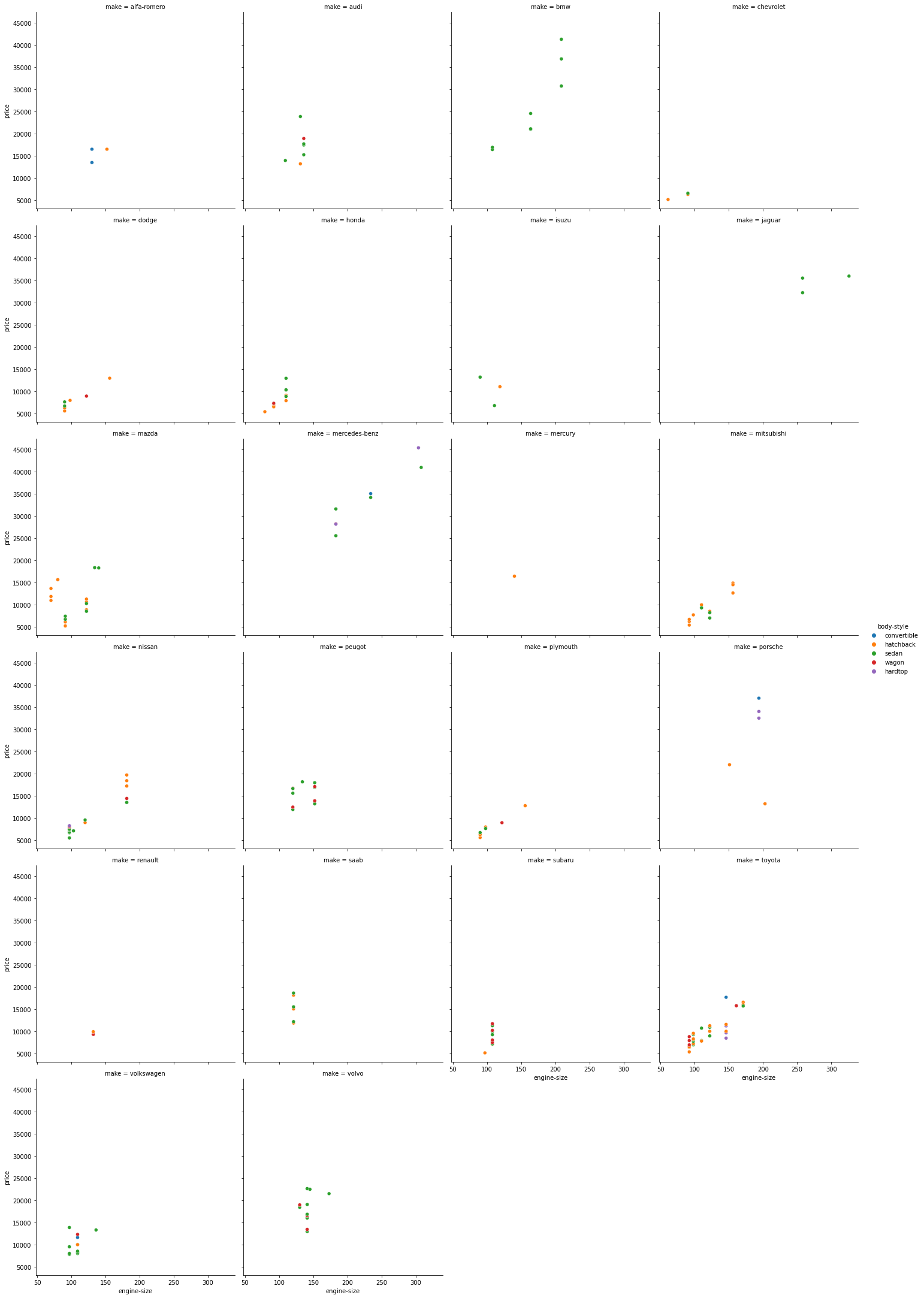
Top 10 leaste expensive cars

1. subaru
2. chevrolet
3. mazda
4. toyota
5. mitsubishi
6. honda
7. nissan
8. dodge
9. Plymouth
10. mazda

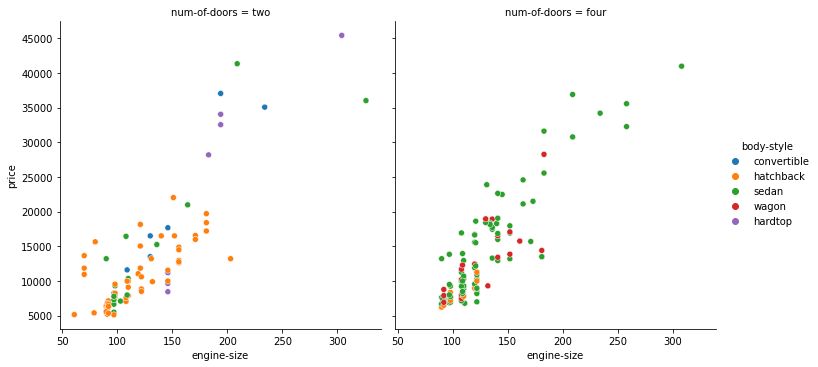
Now that we have some basic control parameters, to understand how they affect the cars: price, peak-rpm etc is going to be a big part of the analysis.



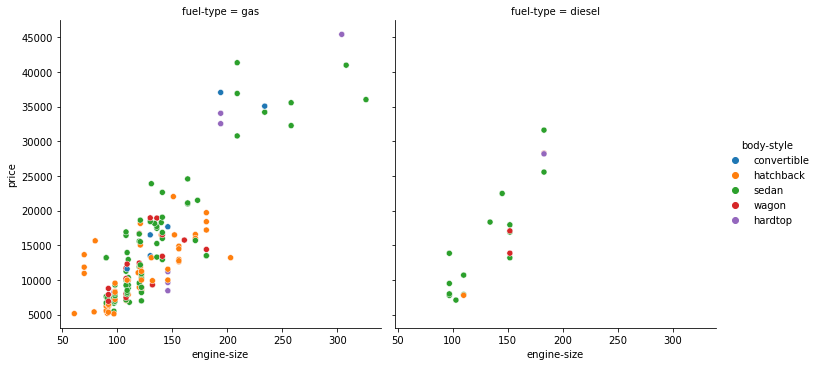
This is a really great representations of the distribution of of the sum of the prices of each car make to the engine sizes of the cars. It is really linear and shows a normal that the bigger the engine size the bigger the price of the car.



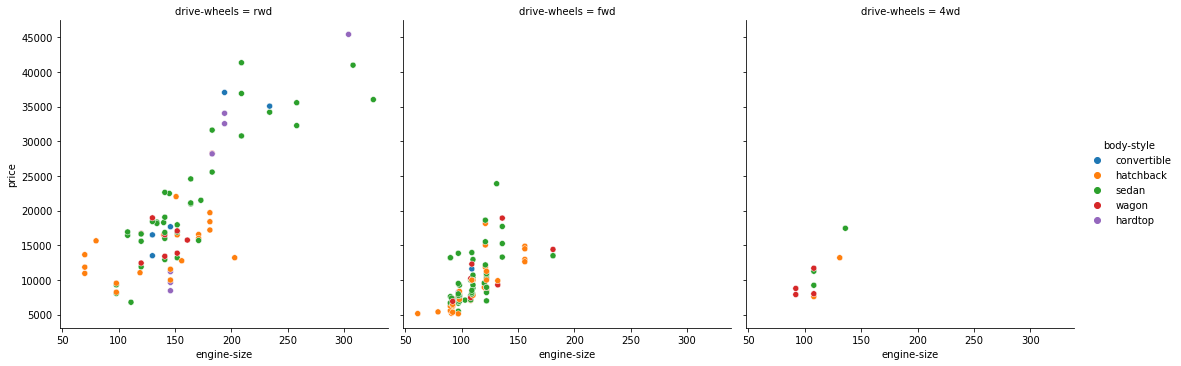
These are the relplots for the price to engine size distributions per car make. It supplements the previous plot by describing, in a more detailed way, the distribution of price to engine size. Prices to engine size but this time the we show the distribution by body type and the different scatter plots are made with the engine location. AS we can see there are more cars with the engine at the front than at the back.



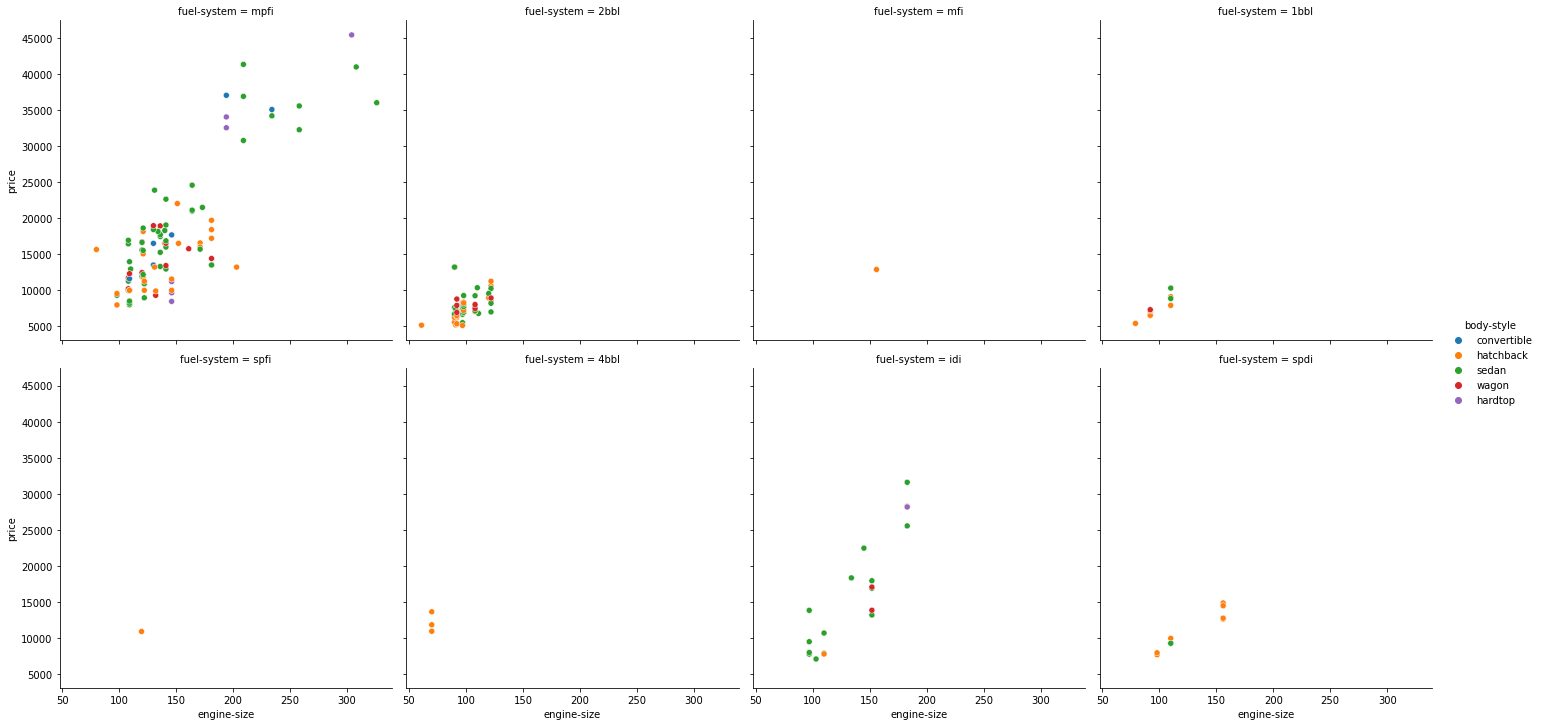
Same thing once again the engine size to the price but the different plots compare the number of doors and it is wonderful to see how the data is clumped based on the number of doors There are more four door cars than two door cars which was shown above with that histogram comparing the two parameters. This distribution show that the more car doors are clumped at smaller engine sizes and less prices.



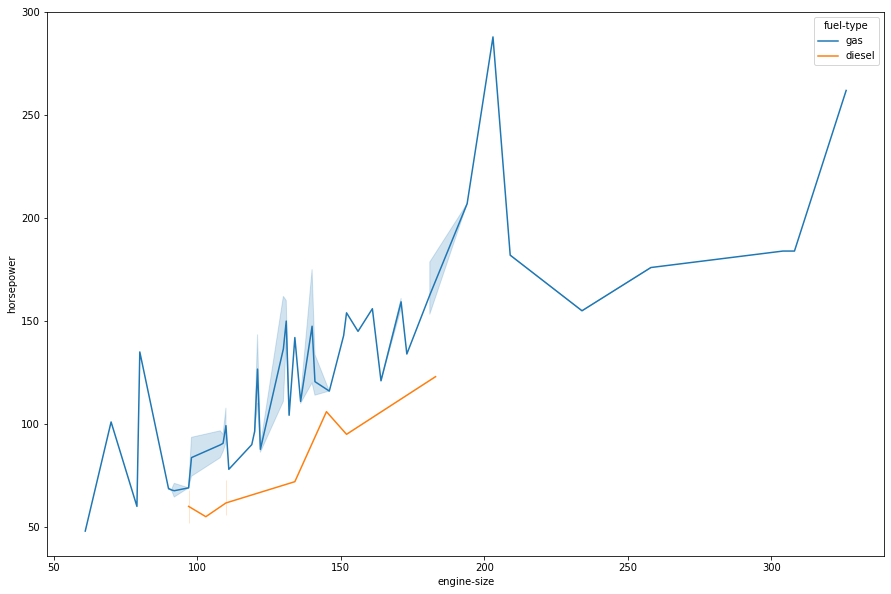
Same thing once again the engine size to the price but the different plots compare the fuel types and it is wonderful to see how the data is clumped. There are more gas cars than diesel cars and this is fascinating to see because it shows that the diesel cars don’t necessarily have big engines and so are much cheaper on average than the gas cars which have bigger engines.



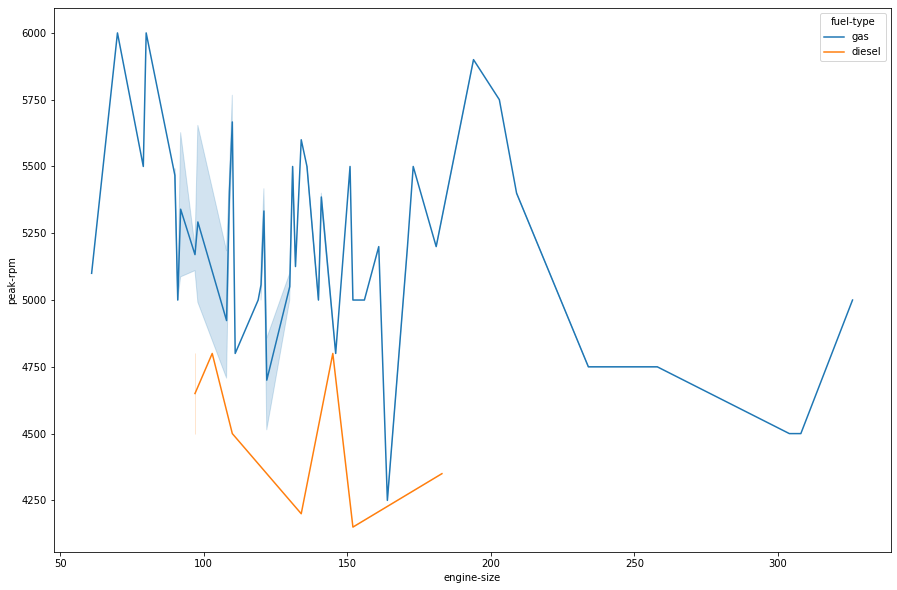
Same thing once again the engine size to the price but the different plots compare the drive wheels and it is wonderful to see how the data is clumped.



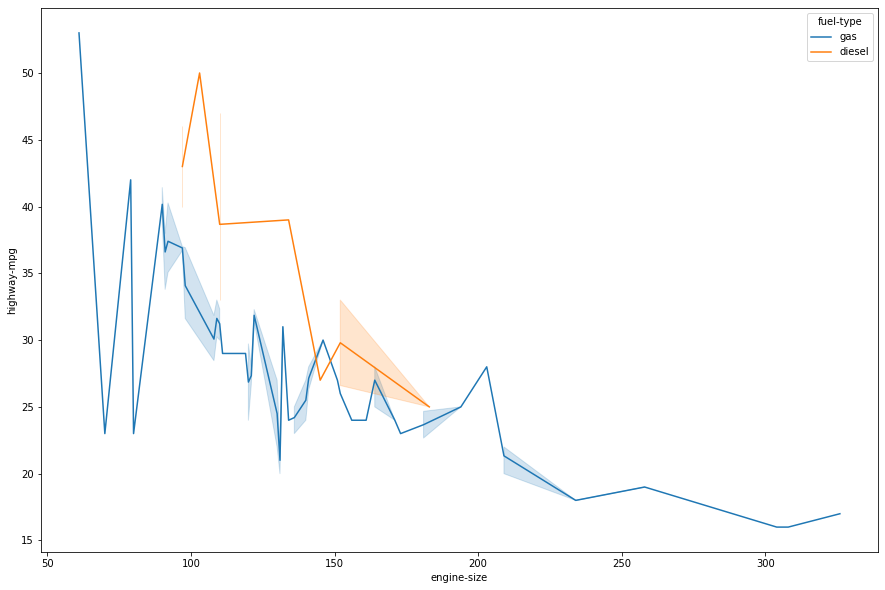
Same thing once again the engine size to the price but the different plots compare the fuel-system and it is wonderful to see how the data is clumped. The mpfi seems to be the most common system type within the cars.



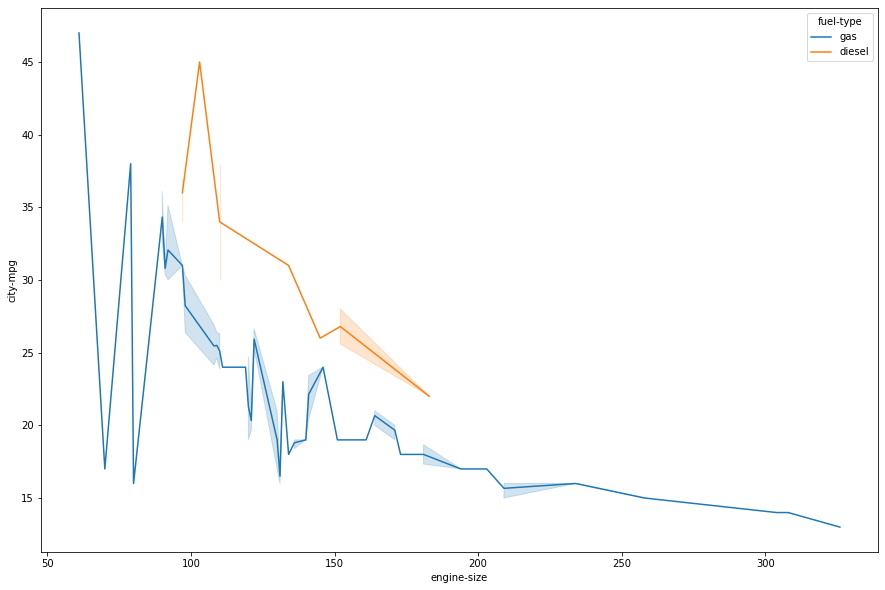
The relationship between the engine size and horsepower is frankly more surprising than I expected. I thought it would be a smooth upright curve but it seems that on average the greater the engine size the more the horse power. And basically the fuel type does make a huge difference in the horse power as well.



The data tends to show that the peak rpm decreases as the size of the engine increases. That’s honestly quite fascinating.



The value of the highway miles per gallon decreases as the size of the engine increases and based on the fuel type the the value is generally greater when uding diesel than gas. This is truly some interesting information.



The same thing with city miles per gallon this is truly fascinating.

The information gathered from this analyses shows some really interesting facts that could be used to make some decisions on what cars to buy and the best for a specific price range and the like.

**THIS REPORT WAS WRITTEN BY : Siphesihle Musa Ntombela**