Analysis on car sale data

Ray Wang, Henry Huang

Introduction

Cars have become an important component of modern life, and it is one of the most common transportation. Many websites and blogs provide car advice to people who are interested, and these advice covers nearly all aspects of car. However, despite the large amount and wide coverage of these advice, there are no current work on data-based customer advise, and most of the suggestions available online are only based on subjective opinions rather than strict data analysis. Therefore, we decided to contribute this data science project.

In this project we focuses on the carsale dataset that includes data for different car models. Our aim is to determine what customers value the most when buying cars and how are the original price and the second-hand price influenced by car's properties. Then consumer advice will be given based on the observations for people who want to sell their cars to help them make the most out of their cars.

Data Collection & Data Cleaning

The dataset used in this project will be the Car_sales.csv dataset from Kaggle user GaganBhatia (https://www.kaggle.com/gagandeep16/car-sales). The author collected this dataset from Analytixlabs, and the data now contains the car brand, manufacturer, horsepower, wheelbase, total sales, second-hand value after 10 years, and other specific car details.

When we observe the dataset, we discover that there are several NA values stored:

```
colSums(is.na(carsale))
```

```
##         Manufacturer                Model   Sales_in_thousands
##                    0                    0                    0
## X__year_resale_value         Vehicle_type   Price_in_thousands
##                   36                    0                    2
##          Engine_size           Horsepower            Wheelbase
##                    1                    1                    1
##                Width               Length         Curb_weight
##                    1                    1                    2
##        Fuel_capacity      Fuel_efficiency        Latest_Launch
##                    1                    3                    0
##    Power_perf_factor
##                    2
```

Therefore, we first need to remove those NA values:

```
carsale <- carsale[rowSums(is.na(carsale)) == 0,]
```

Now, we are interested in the specific values stored. We find that only the manufacturer, model, and latest launch are chr type, while the others are all in eeither numeric or integer. For manufacturer and model, they

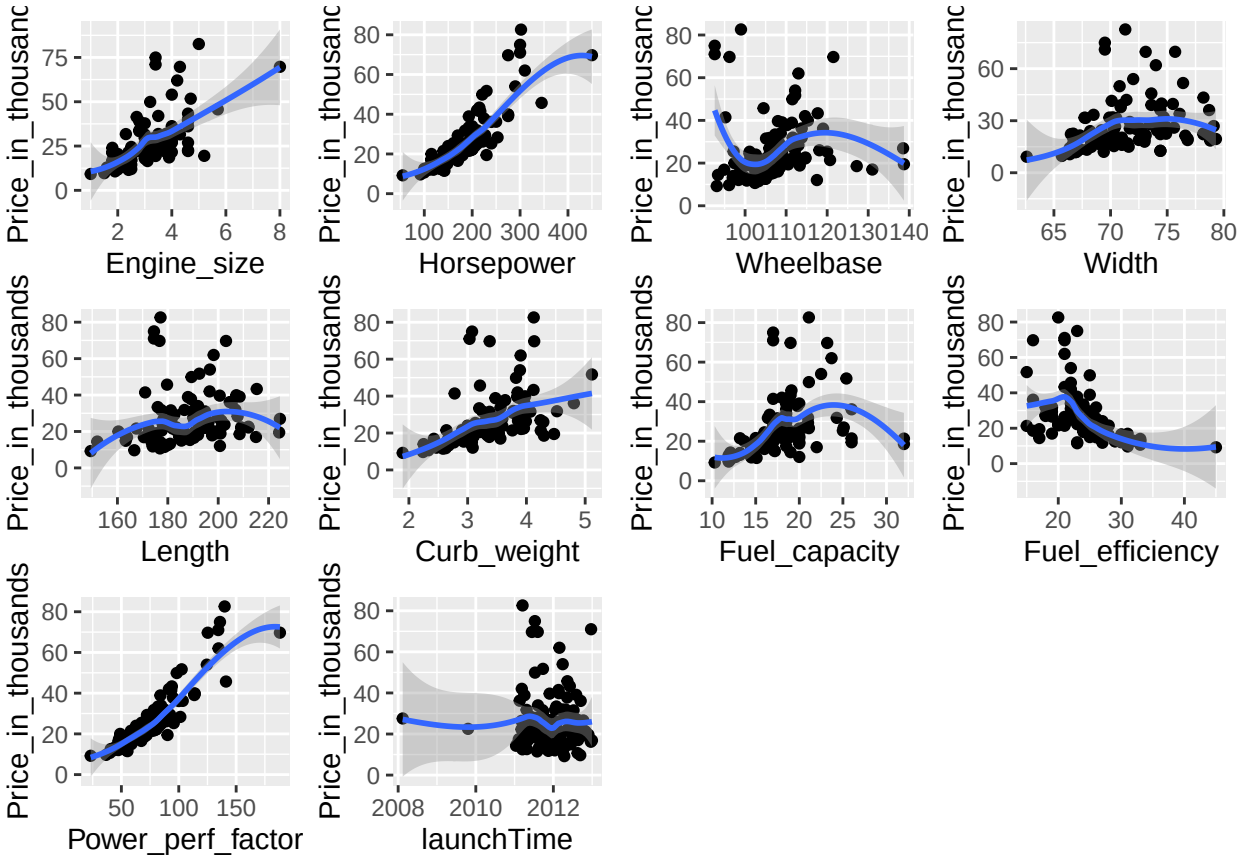
should stay in chr type; but for the latest launch, we should use a format that can be directly compared to make it convenient for future analysis. One way of doing this is by using the lubridate package and convert the strings into time format.

```
carsale <- carsale %>%
  mutate(launchTime = mdy(Latest_Launch))
```

Now we are ready to dive into data analysis.

Factors Influencing Price

The price of a car can be influenced by many factors, and here we will determine which of the given factors are correlated with the listed price and how are they correlated.



From the graphs above, we can see that most y-axis are correlated with the price. The engine size, horsepower, and power factor are the three most correlated variables, and all of them have a positive and nearly linear correlation with a small quartile difference. This observation indicates that the factor that can most strongly suggest the potential price of a car is its engine size, horsepower, and power factor. This makes sense when you consider it in the real world, and more expensive cars often have bigger and stronger engines, no matter they are sedans, SUVs, or sportscars.

We also observe a curious and unique correlation between the fuel efficiency and price. Interestingly, this correlation is the only one that is negative; only fuel efficiency decrease when price increase. From the first glance, this may not be reasonable since expensive cars should be better. However, when we take into consideration the above conclusion in the first paragraph, we will notice that bigger and more powerful engines often are less fuel-efficient. This explains our observation here.

Additionally, the launch time has nearly no influence on the price; this is probably because the selling price does not change muc by time, and since most of the cars in the dataset are published recently the date does not have a huge influence.

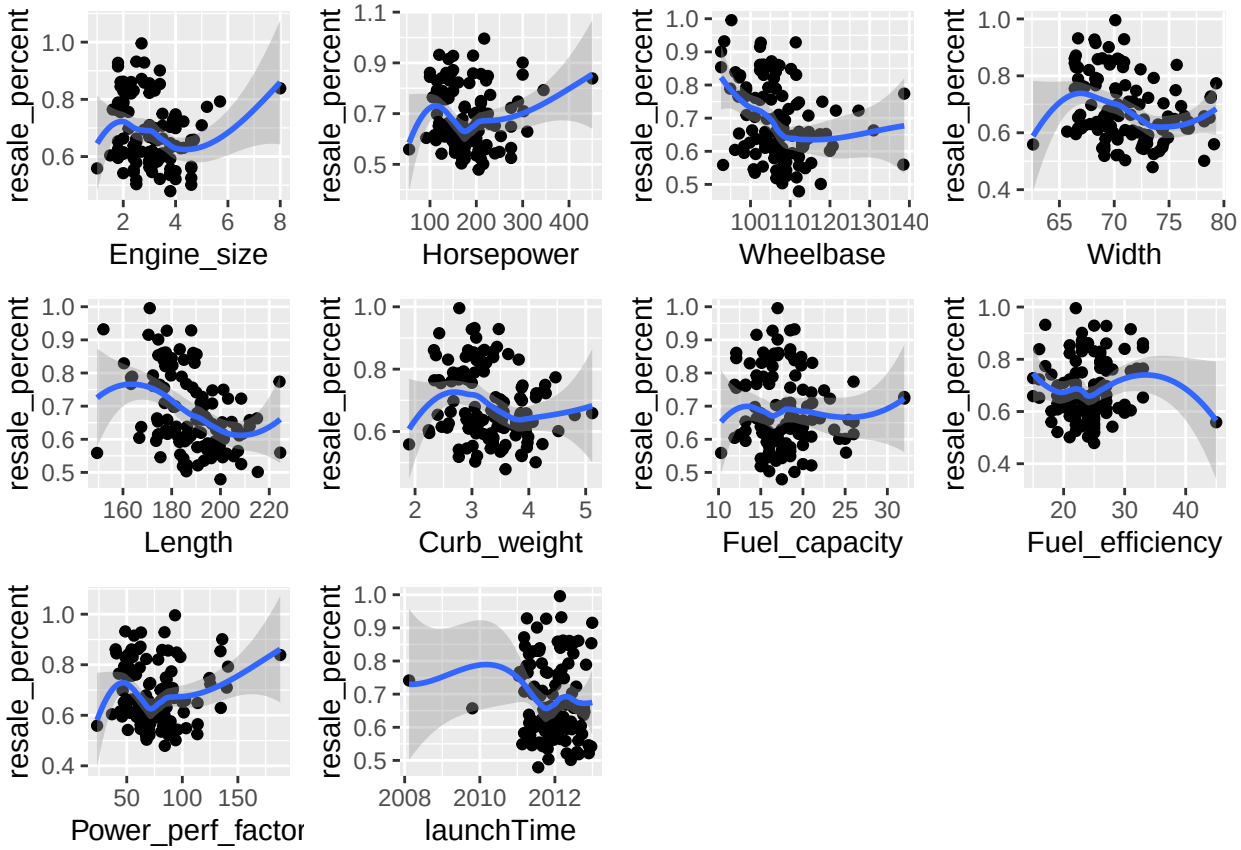
The rest of the factors, width, length, wheelbase, and curb weight, all share a similar correlation shape. The trend is positive when the price is smaller than 40 (thousand dollar), but it becomes flat afterwards, and it even declines at the end. We can infer from the section smaller than 40 that expensive cars tend to have larger and heavier framework. Yet that does not explain the flat part of the curve. One possible reason for this shape is that traffic regulations limit the size and weight of a car, and this limit is nearly reached at a price of 40 thousand dollars, so it does not increase as significantly or as sharply as when the price is lower. The final part of the curve decreases most likely because of the emergence of sport cars, which are both expensive and small. Therefore, when the price is lower than 40 thousand dollar the size of the car is a good indicator of the price, but when the price goes beyond 40 thousand dollar this is no longer a good indicator.

Factors influencing second hand price

For the second hand price, we cannot use the resale value directly. We need to consider the fact that when the original price is high, the resale value should also be high. Therefore, we need to find a way to remove this influence, and one such way is simply divide the resale value by the original price to get the percentage of the price. This indicates that how many percent of the original value can you still sell the car for after 10 years.

```
carsale <- carsale %>%
  mutate(resale_percent = X__year_resale_value / Price_in_thousands)
```

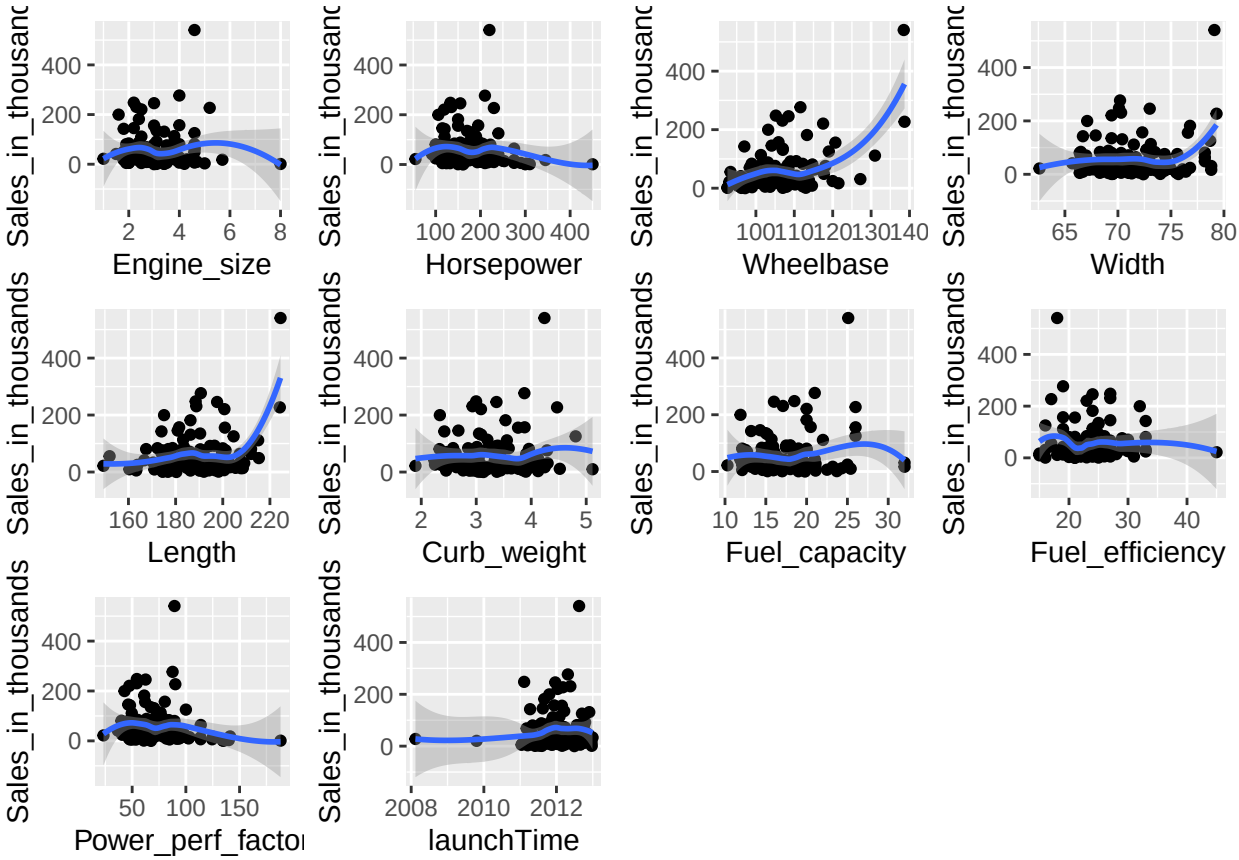
After this decision, we can take a look at the graphs:



Now we are interested in correlations between the factors. From the above graphs, we can see that engine size, wheelbase, width, length, and curb weight are all roughly negatively correlated with the percentage. As we have discussed previously, all of these factors are indicators of the car price with a positive correlation. Combining these two observations, we will reach the conclusion that roughly, expensive cars tend to lose more percentage of its original price when they are being resold. This relationship should not be considered as strong due to the weak correlation between the factors and the percentage, but it provide a reasonable vision regarding the resale price and original price. This correlation is also sensible when we consider it in the real world; the expensive cars, except for those with a historical meaning or a collection value, generally devalue quickly after they are driven; some super cars even drop one third of its price once its wheels have touched the ground. Thus, the conclusion here is that the higher the original price, the more percentage will be lost when the car is sold second-handed.

However, we also need to consider those graphs where the correlation rarely exist. In the rest of the factors, that is the case. One potential explanation for this weak correlation is that the second-handed cars are highly influenced by their previous owner; if the owner treated the cars nicely, the final price could be significantly different than the other badly-treated cars. That explains why none of these correlations are strong; the most important factor is no longer the car's internal property but the external factor depending on the owner. Therefore, our advice is that if you want to sell your car later, treat it well is more important.

Factors influencing car sales



The graphs above show the relationship between different factors of the cars and the number of sales of the cars.We can see that most of the factors do not show a strong correlation with the car sales except for three factors related to car size: Length, Wheelbase, and Width. With these three factors, we can see that the samples with highest factors will greatly increase car sales while the rest samples doesn't change car sales much. For the rest of the factors, we see that some of them, such as Engine Size, Horse power. and Fuel

Capacity, the sales decrease for some extreme samples, both in the left tail and the right tail.

We can deduce from the results that larger cars tend to get more second hand sales while other factors doesn't change the sales much. One reason for why larger cars get more sales is that most smaller cars are for personal use while trucks are used for business a lot of times. Cars used for business may be out of use after a while and will sold by the company. On the other hand, cars used for personal use usually are kept by the family for a long time. Moreover, larger cars tend to have a higher resistance to damage thus are safer purchases. The low sales for cars with other extreme factors can be explain by the fact that cars with extreme factors are either too out dated or too expensive to be sold second hand. The low correlation between the factors and car sales can be explained by the reason mentioned in the upper section: the condition of the car is a more important factor influencing car sales. Lat but not least, we see a large spread in car sales for cars with the same data for the factors. This spread can be explained by the fact that people tend to care more about the price and the brand of the car when purchasing. Some brands of cars have a more affordable price thus will be purchased by more people.

Conclusion

We found that the factors have more influence on the pricing of the cars and have less impact on the second hand price and the sales of the cars. This once again proves that the conditioning of the car is the most important factor when reselling a car. For car owners, we suggest to buy larger cars when only using it for a short term and do not buy cars that have extreme values for the factors not related to car size. All in all, the most important thing to do for owners who want to sell their car is to take good car of their cars.

Due to the lack of sufficient data, our conclusion may not be guaranteed to be correct when a lot of other car models are introduced. In the future, larger datasets can be used to establish a more general correlation and make our conclusion more solid.