Final678

Yifeng Fan

2022-12-04

Abstract

A car is a very typical Durable Consumption Goods. Due to the low number of purchases of consumer durables, consumers' purchasing behavior and decision-making are more deliberate. Cars in the auto market from a few thousand yuan to hundreds of thousands of yuan have their own special pricing strategies. This is also a very complex subject. Although the pricing of car brands tends to vary greatly, there are many rules involved. Often, consumers cannot take into account factors such as manufacturing processes, materials, technologies and designs. This report will try to find the logic behind some car brand pricing.

Introduction

The project will use a database obtained from Kaggle, and the 56 variables in this database are organized around some of the basic configuration of the car, such as the engine, transmission, chassis, braking system and safety equipment such as airbags and anti-lock braking system. Because the 56 possible explanatory variables are too large for the project's model, this report will devote more space to finding the most appropriate variables. First I will use data visualization to exclude some variables that may have little impact on the price, and then draw a correlation matrix to further exclude those variables that may cause multicollinearity. I will use the multilevel model to find the fixed effect between these factors and pricing and the random effect between different brands and years.

Method

Data Cleaning

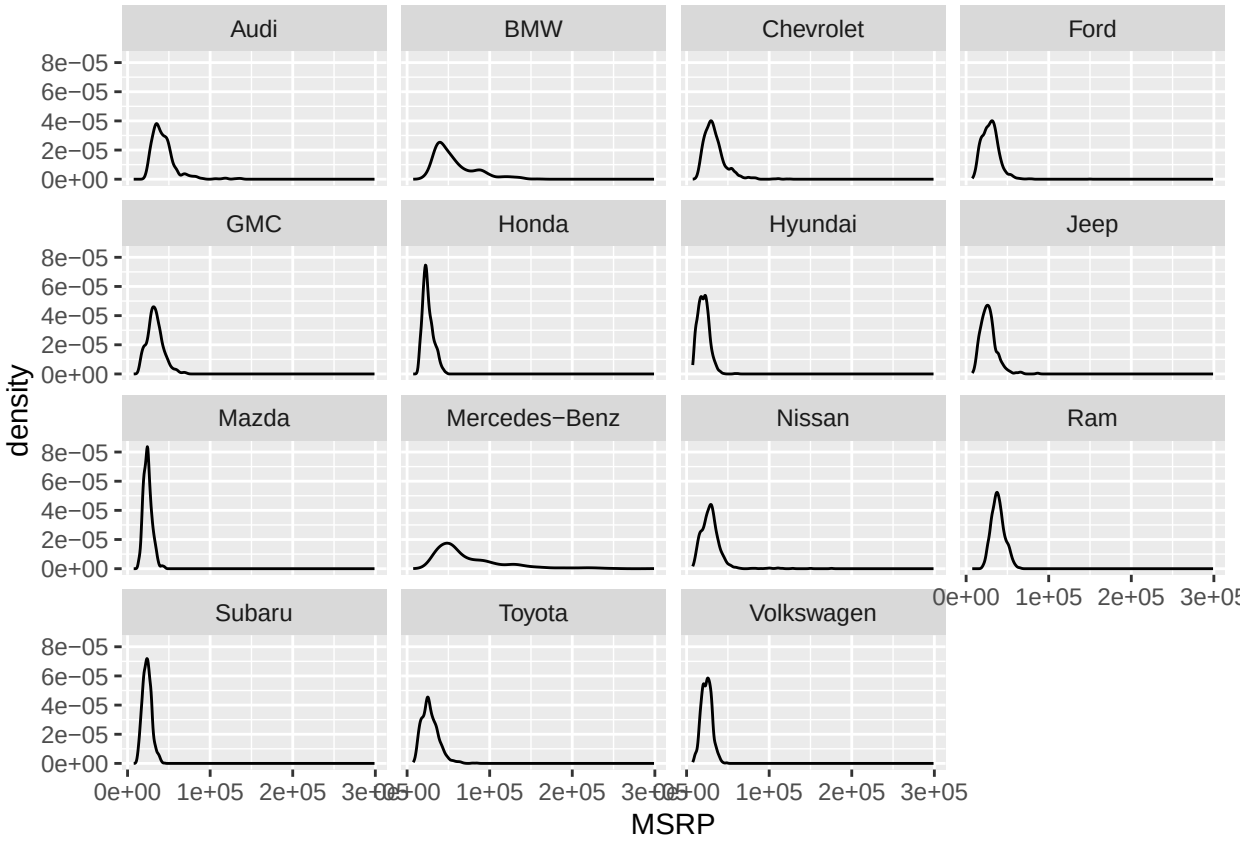
The Dataset contains prices for around 32000 cars with 56 features. This covers products from 43 automotive brands. But when I looked at the data, I found that the number of occurrences of each brand was different, and in order to get more accurate results when analyzing, I decided to keep only the 15 brands with the largest sample as my first group. In the second group, I chose the year in which each car was manufactured. I would divide cars produced from 1990 to 2018 into three groups, each of which roughly includes cars of various brands and models produced in the past ten years. In the preliminary screening, I will also remove some columns that contain too many NA variables. There are also a lot of categorical variables that I'll remove some of them first. Here are some of the variables and explanations after initial screening:

column names	explanation
MSRP	Selling price in dollars
Manufacturer	Car brand
EPA.Fuel.Economy.Est... City..MPG.	City mileage
Passenger.Capacity	Capacity
Base.Curb.Weight..lbs.	Car weight in pounds
Wheelbase..in.	Distance between the centers of the front and rear wheels
Track.Width..Front..in.	Distance between the centerline of two frontwheels
Fuel.Tank.Capacity..Approx..gal.	Tank capacity in gallons
SAE.Net.Torque... RPM	Net torque Nm
SAE.Net.Horsepower... RPM	Net horsepower

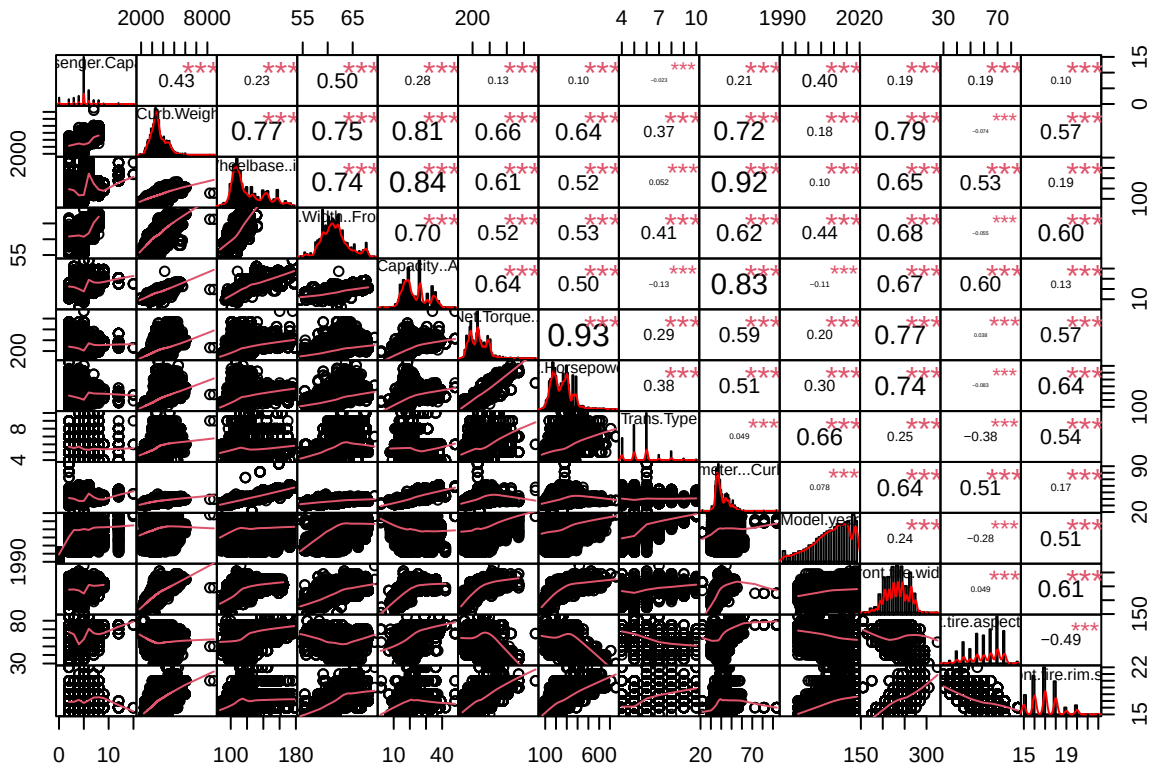
column names	explanation
Trans.Type	Number of gears
Turning.Diameter. . . Curb.to.Curb..ft.	Diameter of the smallest circular turn by the car
Model.year	Model year
Front.tire.width	Front tire width
Front.tire.aspect.ratio	Front tire aspect ratio
Front.tire.rim.size	Front tire rim size
Assists	Average assists of player
Steals	Average steals of player
blocks	Average blocks of player
Turnovers	Average turnovers of player
Fouls	Average fouls of player
Height	Height of player (foot)
weight	Weight of player (lb)
Draft_Pick	Draft pick of player

EDA

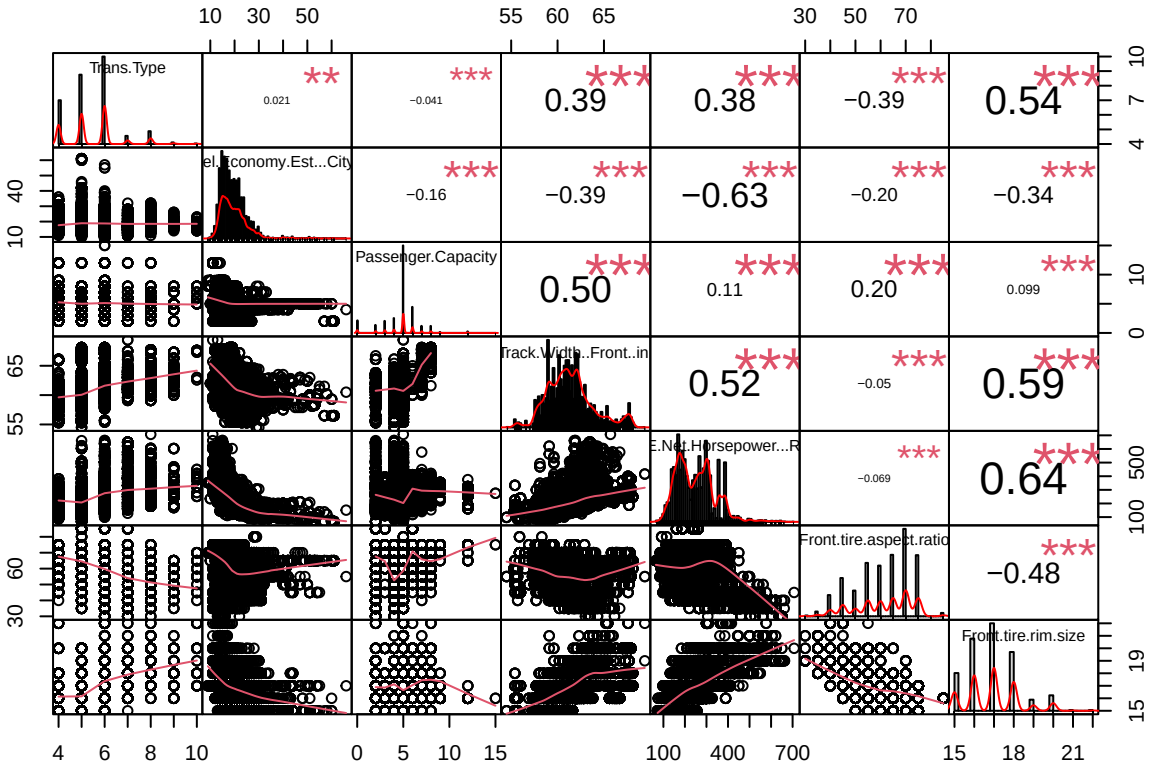
Distribution of MSRP for 15 manufacturer



From the figure above, we can see some brand pricing differences. Some Japanese brands such as Honda, Mazda, and Subaru are mostly concentrated in the lower price area, while prices like Audi, BMW, and Mercedes-Benz are more fragmented.

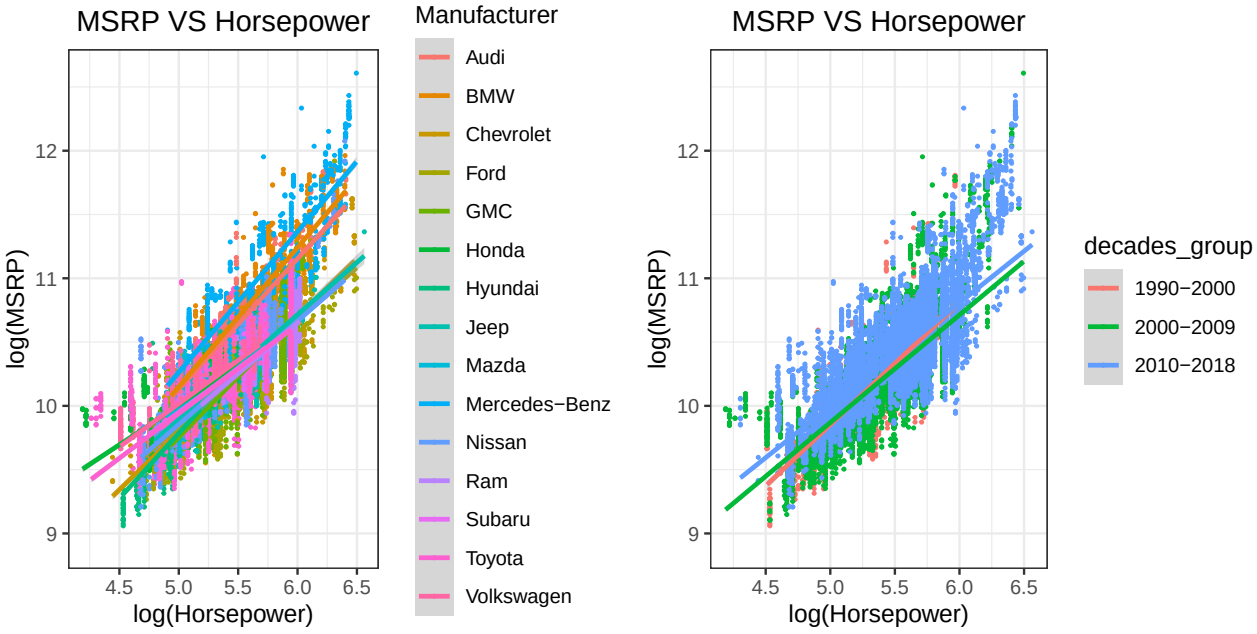


According to the correlation matrix, we can decided which explaining variables to included in the multilevel model. I assume setting the threshold to be 0.65 would be appropriate for this model. Then some of the explaining variable would need to be dropped. "Torque and Horsepower" is highly correlated, I decided to drop the torque. "Wheel base in. and Turning diameter" also has a very high correlation parameter 0.92. I would drop the turning diameter(crub to curb ft.). And other variables that have higher correlation parameters are front tire width, base curb weight and wheel base.



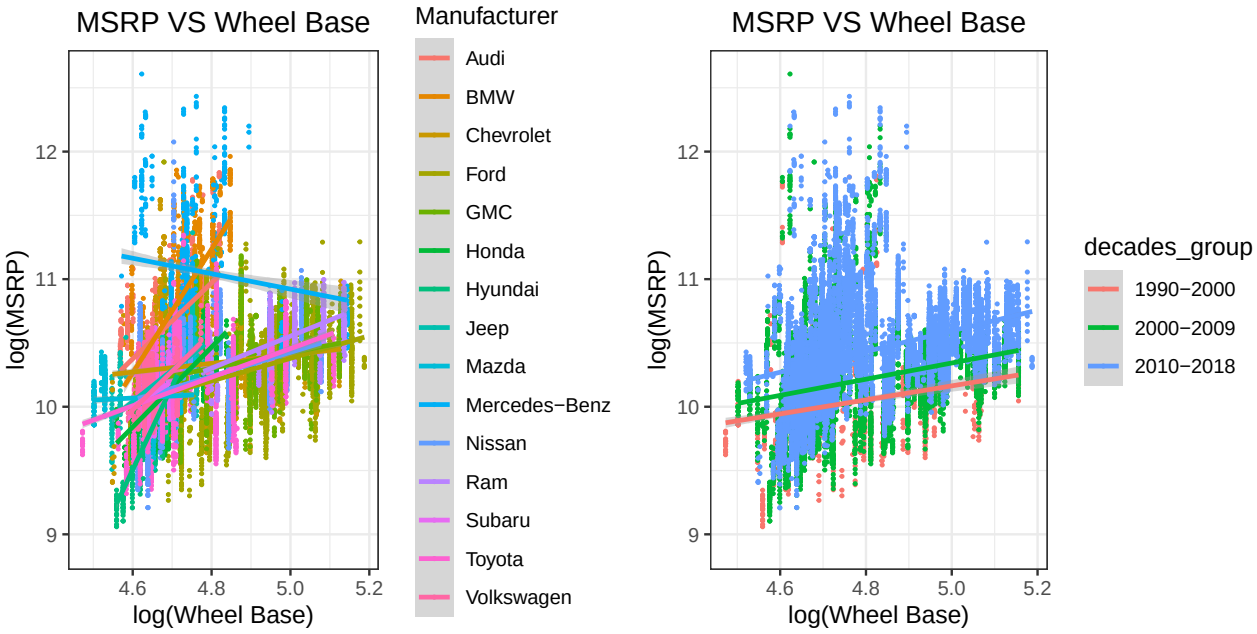
After the ajustment, all the correlations drops below the 0.65 threshold. Next, we can perform further analysis and screening.

Pick predictor

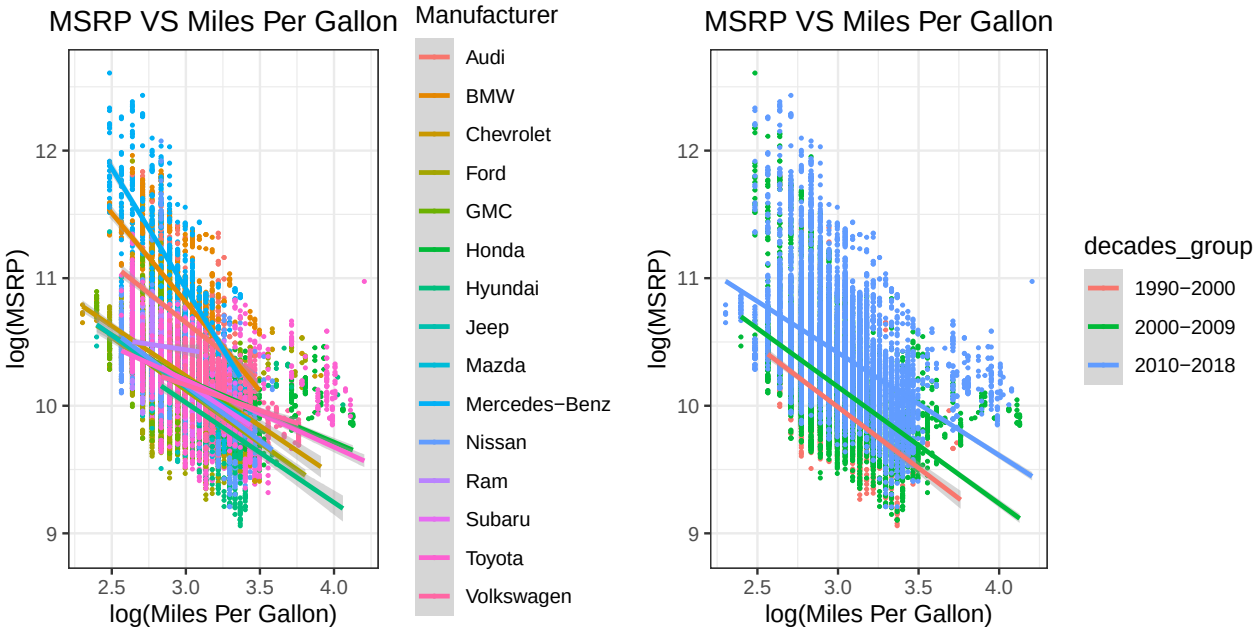


The above two graphs are both graphs of the relationship between the price of the car and the horsepower of the car. The different colored straight lines on the left represent different car manufacturers. The different colored straight lines on the right represent different production years (in groups of ten years). In both charts, it is not difficult to see a positive correlation between price and horsepower. Whether it's a practical-focused family car or a performance car that pursues driving pleasure, more horsepower always reflects higher market

pricing.

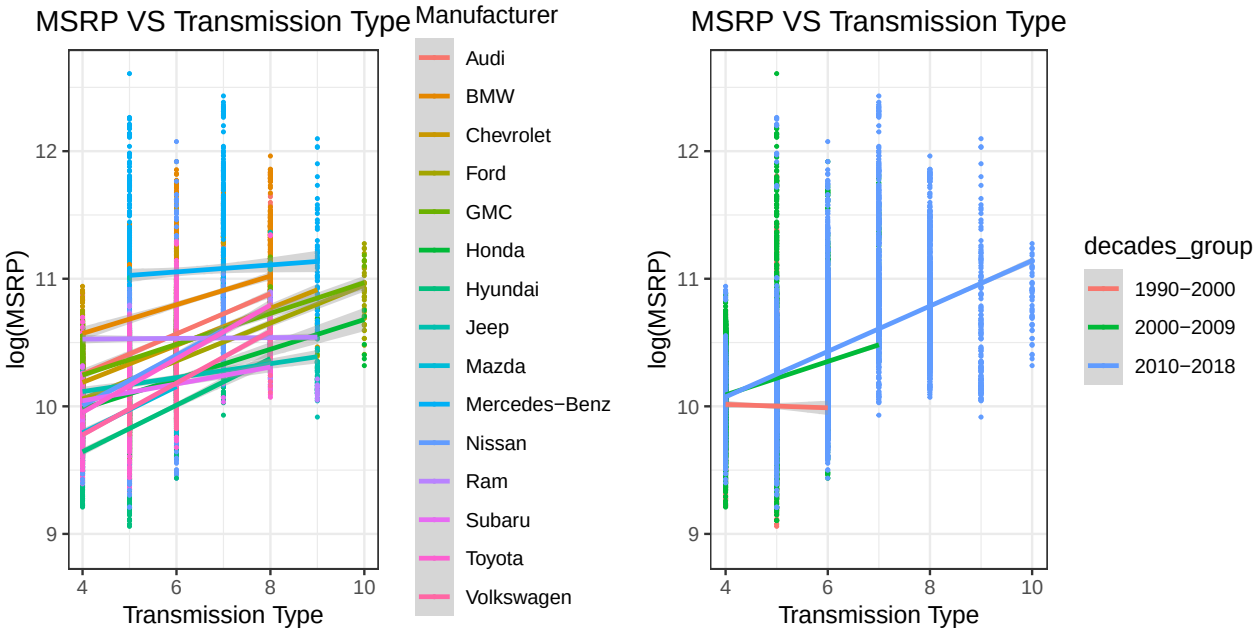


The above two graphs reflect the relationship between wheel base and MSRP. Wheelbase is the distance between the centres of the front and rear wheels. It is a smaller parameter than a car's full length. Usually the comfort of the interior space and the stability of handling will be better with a long wheelbase, so I guess the wheelbase of the car will also be proportional to the pricing. However, from the first figure, it can be seen that the relationship between the two is not stable. Perhaps because some sports cars choose a shorter wheelbase for performance and handling, sports cars tend to have a higher MSRP. In view of Figure 1, I decided to remove the wheelbase variable from the final model.



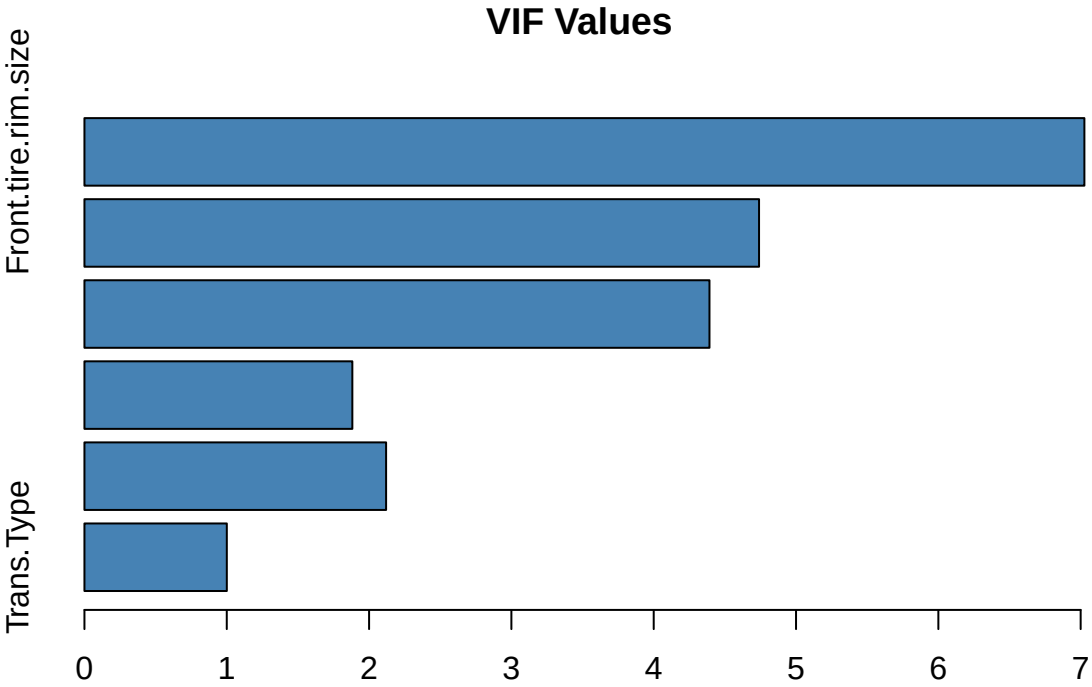
MPG(Miles per Gallon) is probably the primary factor that many pragmatic consumers consider when buying a car. The higher the MPG of a car, the more fuel-efficient it is. Although automakers spend a lot of money to make engines more fuel efficient, cars with good fuel economy do not mean higher selling prices. From the above two figures, it can be seen that MPG has a clear negative correlation with MSRP. This is not contrary to common sense, because many high-end sports cars of car brands will give up some fuel economy

for better performance and driving experience.

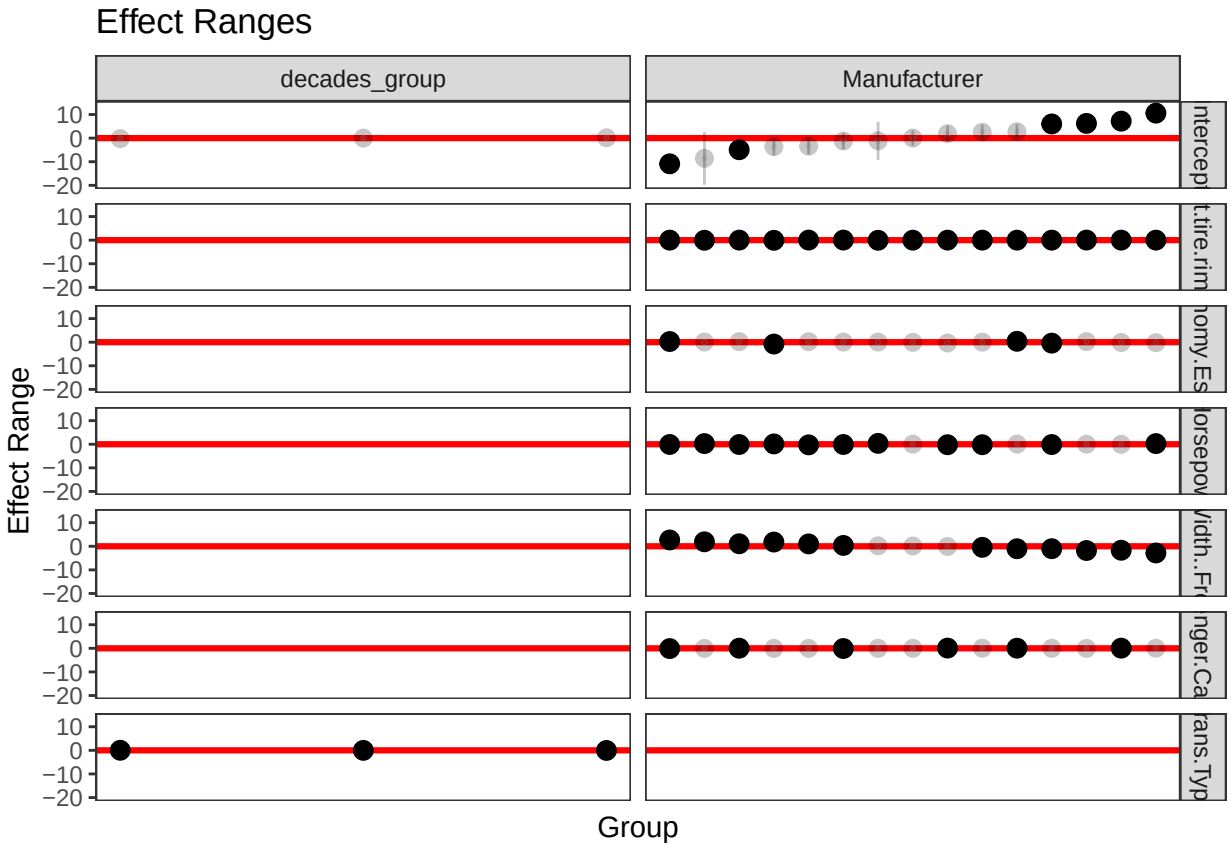


The transmission developed by different brands are also very different. It is difficult to judge the quality of a gearbox by the number of gears in the gearbox. But generally speaking, the more gears there are in the gearbox, the less obvious the setback when shifting gears, and the less transmission loss and more economical fuel consumption. The fewer gears, the lower the cost, the better the reliability and practicality. From the above two pictures, it is not difficult to see that automakers will use transmission with more gears on more expensive models. From the figure on the right, it can be found that the positive relationship is more obvious on newer models produced from 2010 to 2018.

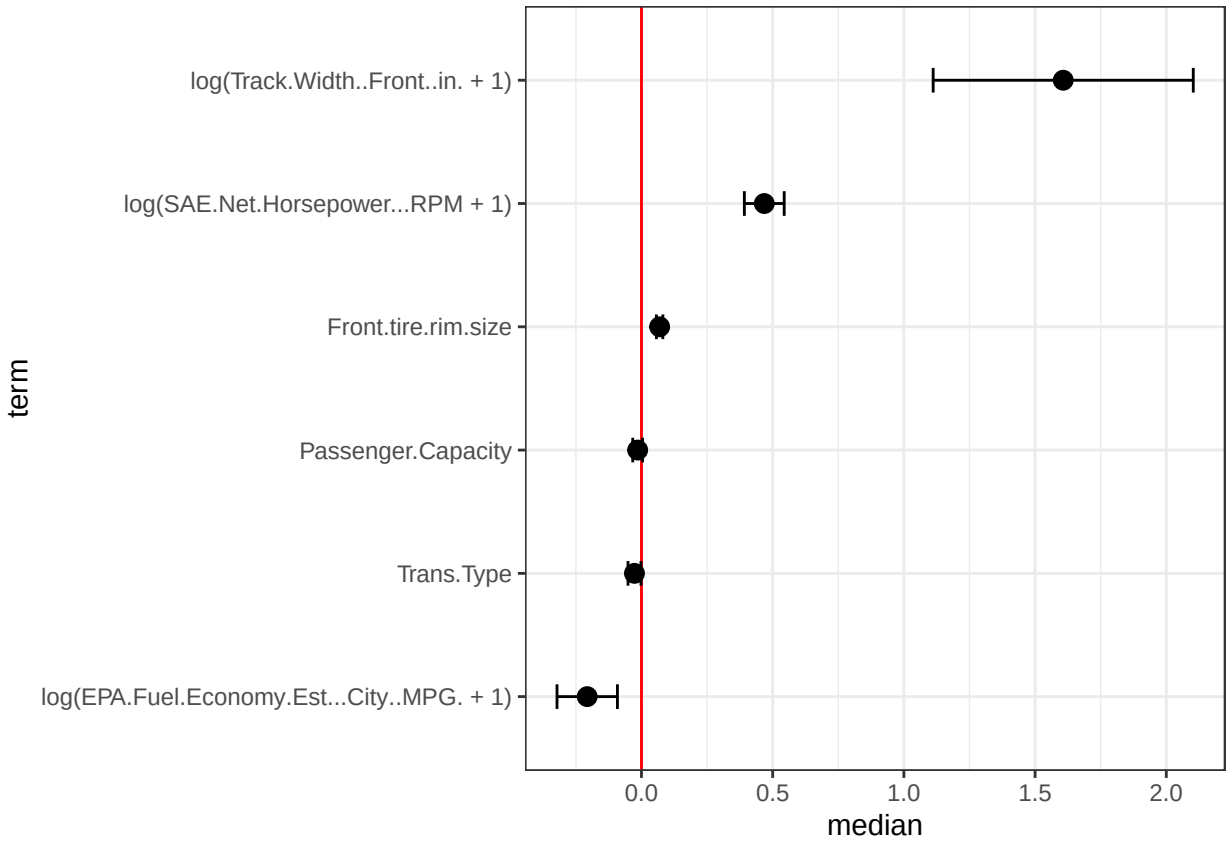
Then I also use the VIF test to check the correlations between the predictors. As we can see, all the predictor are below the threshold 10.



Then, let's see the fixed effect and the random effect from this multilevel model.



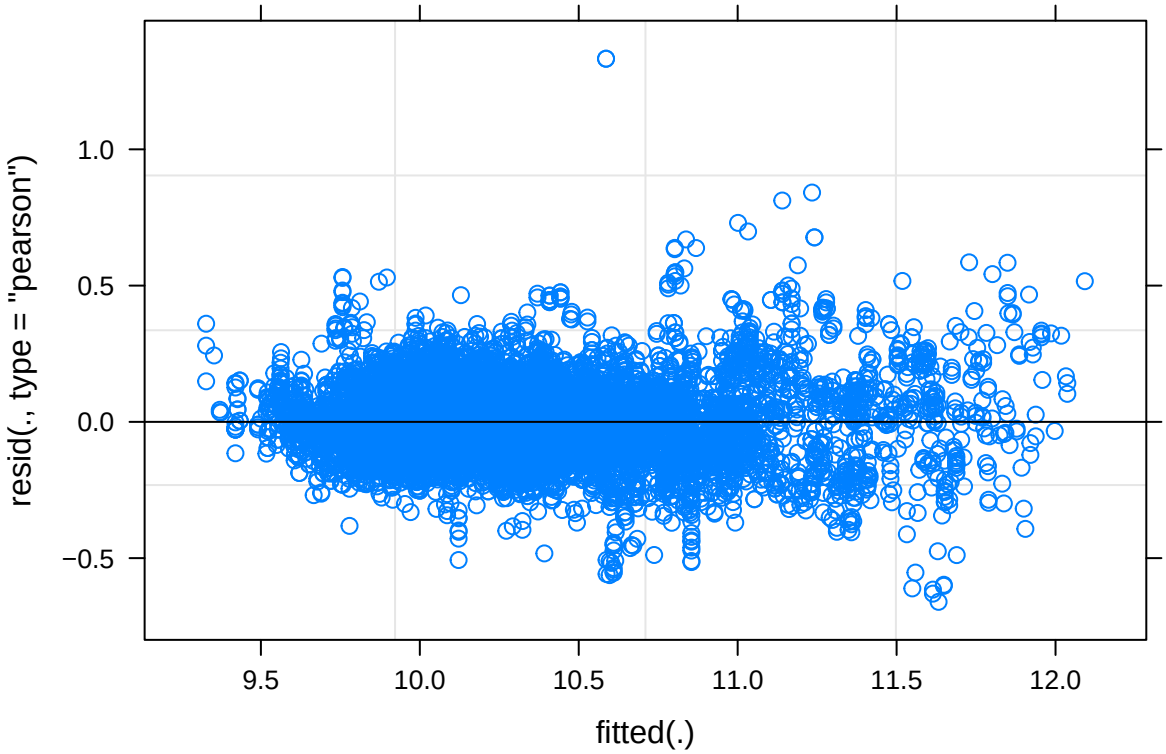
Here is the fixed effects:



Next is the random effects:

```
##                (Intercept) log(EPA.Fuel.Economy.Est...City..MPG. + 1)
## Audi            -8.2451002                                 0.11262468
## BMW             -1.6947532                                 0.04468562
## Chevrolet      -11.1423550                                 0.31579142
## Ford             6.8515993                                -0.10403308
## GMC             -3.7874835                                 0.23848339
## Honda            2.3999350                                 0.05640628
## Hyundai         -5.2235779                                 0.26319279
## Jeep             1.6112932                                -0.32803286
## Mazda           -1.3517705                                 0.08871194
## Mercedes-Benz   -3.7873753                                -0.76169596
## Nissan          10.4282882                                -0.19587115
## Ram             -0.2433973                                -0.04046256
## Subaru           5.7851597                                 0.28668619
## Toyota           5.8739661                                -0.37848715
## Volkswagen       2.5255710                                 0.40200043
##                Passenger.Capacity log(Track.Width..Front..in. + 1)
## Audi               -0.039060889                        1.87400847
## BMW                -0.015828184                        0.21159244
## Chevrolet          -0.103536315                        2.66619706
## Ford                0.060125194                       -1.69186347
## GMC                -0.034056329                        0.99821184
## Honda               0.040688938                       -0.46910592
## Hyundai             0.047434753                        1.05032967
## Jeep                0.099126623                       -0.08637055
```

```
## Mazda                -0.082792232                        0.32934996
## Mercedes-Benz        -0.040893870                        1.76646119
## Nissan                0.021328618                       -2.85599203
## Ram                  -0.002754181                        0.09534247
## Subaru               -0.014549688                       -1.80048121
## Toyota                0.015173961                       -1.03387662
## Volkswagen            0.049593599                       -1.05380318
##                log(SAE.Net.Horsepower...RPM + 1) Front.tire.rim.size
## Audi                                 0.27700913         -0.052997088
## BMW                                  0.35898966         -0.048899434
## Chevrolet                           -0.10272747          0.004959394
## Ford                                -0.05429236          0.019188056
## GMC                                 -0.26025959          0.027971898
## Honda                               -0.19868637          0.009313222
## Hyundai                             -0.10889509          0.013057262
## Jeep                                -0.19882449          0.011789417
## Mazda                               -0.11053115          0.035066725
## Mercedes-Benz                        0.14146555         -0.078437358
## Nissan                               0.27866733          0.015291520
## Ram                                  0.03457026         -0.008600739
## Subaru                               0.02167011          0.040152701
## Toyota                              -0.12427991          0.006462719
## Volkswagen                           0.04612439          0.005681703
```



Above is the residual plot for the model. We can see that the dots are symmetrically distributed around the line h = 0. Although there are some dots have very big residuals, the rest of the points are still relatively concentrated.

mod_000 <- lmer(data= data_New4, MSRP ~ Trans.Type + log(EPA.Fuel.Economy.Est. . . City..MPG. + 1) + Passenger.Capacity + log(Track.Width..Front..in. + 1) + log(SAE.Net.Horsepower. . . RPM +1)+

Front.tire.aspect.ratio + Front.tire.rim.size +

(1+ log(EPA.Fuel.Economy.Est. . . City..MPG. + 1) + Passenger.Capacity + log(Track.Width..Front..in. + 1) + log(SAE.Net.Horsepower. . . RPM +1)+ Front.tire.aspect.ratio + Front.tire.rim.size | Manufacturer)

+(1+ Trans.Type | decades_group))

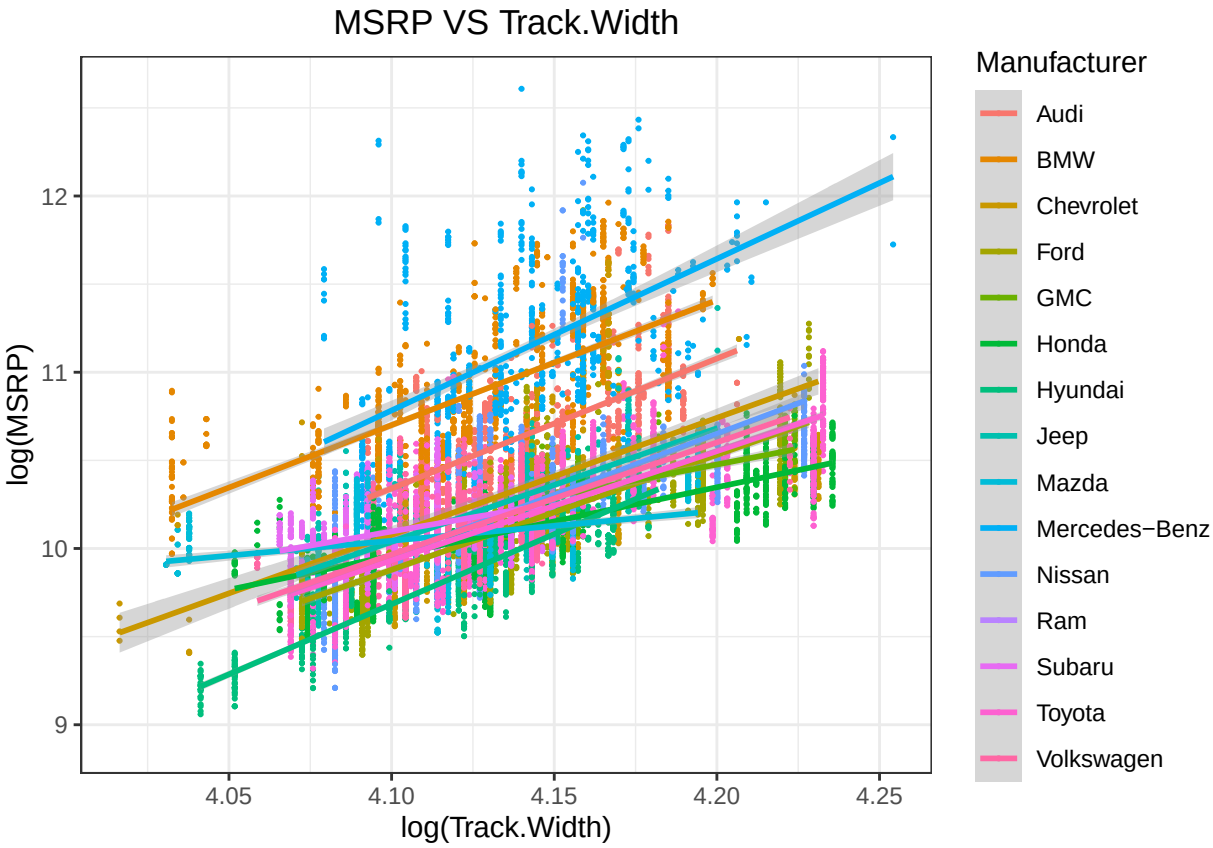
Result After the model fitting, this is the formula I get for the MSRP:

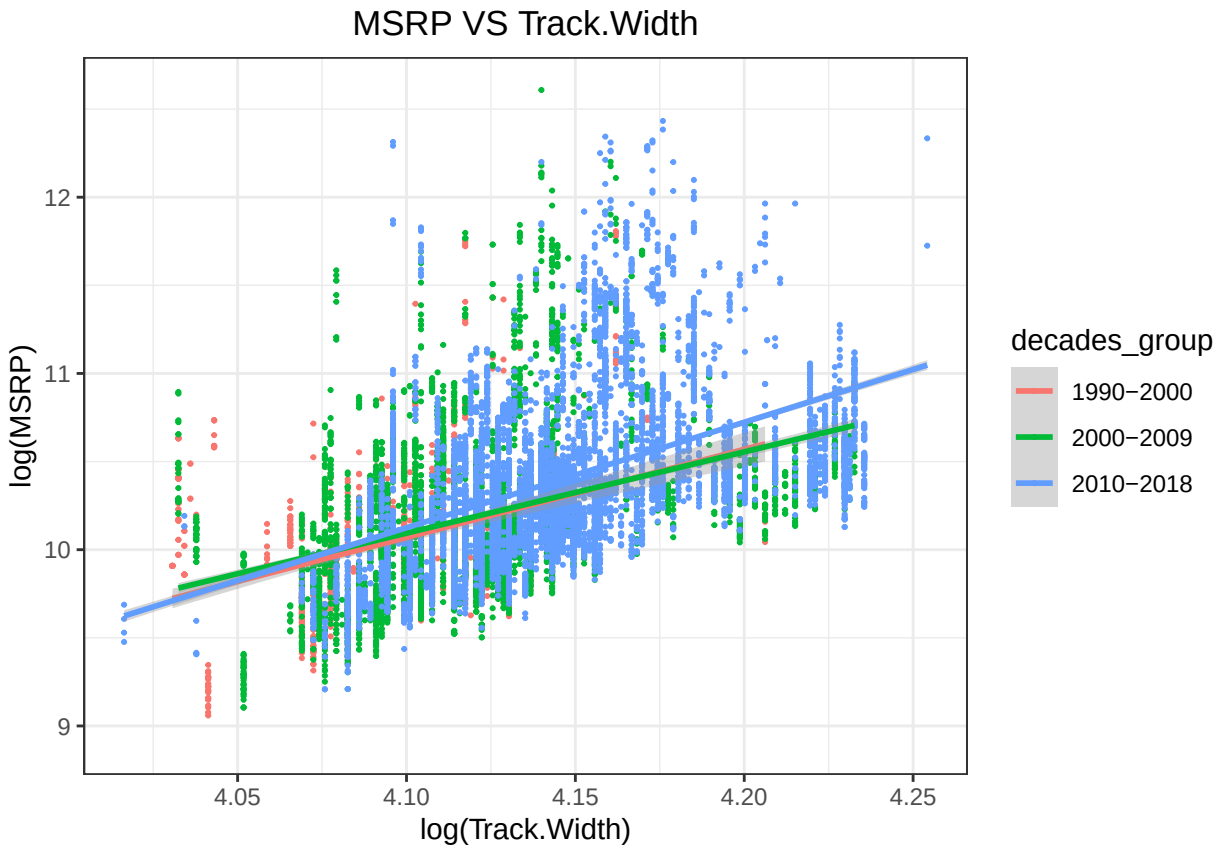
$$log(MSRP) = Intercept + Trans.Type + log(EPA.Fuel.Economy.Est. . . City..MPG.+1) + Passenger.Capacity + log(Track.$$

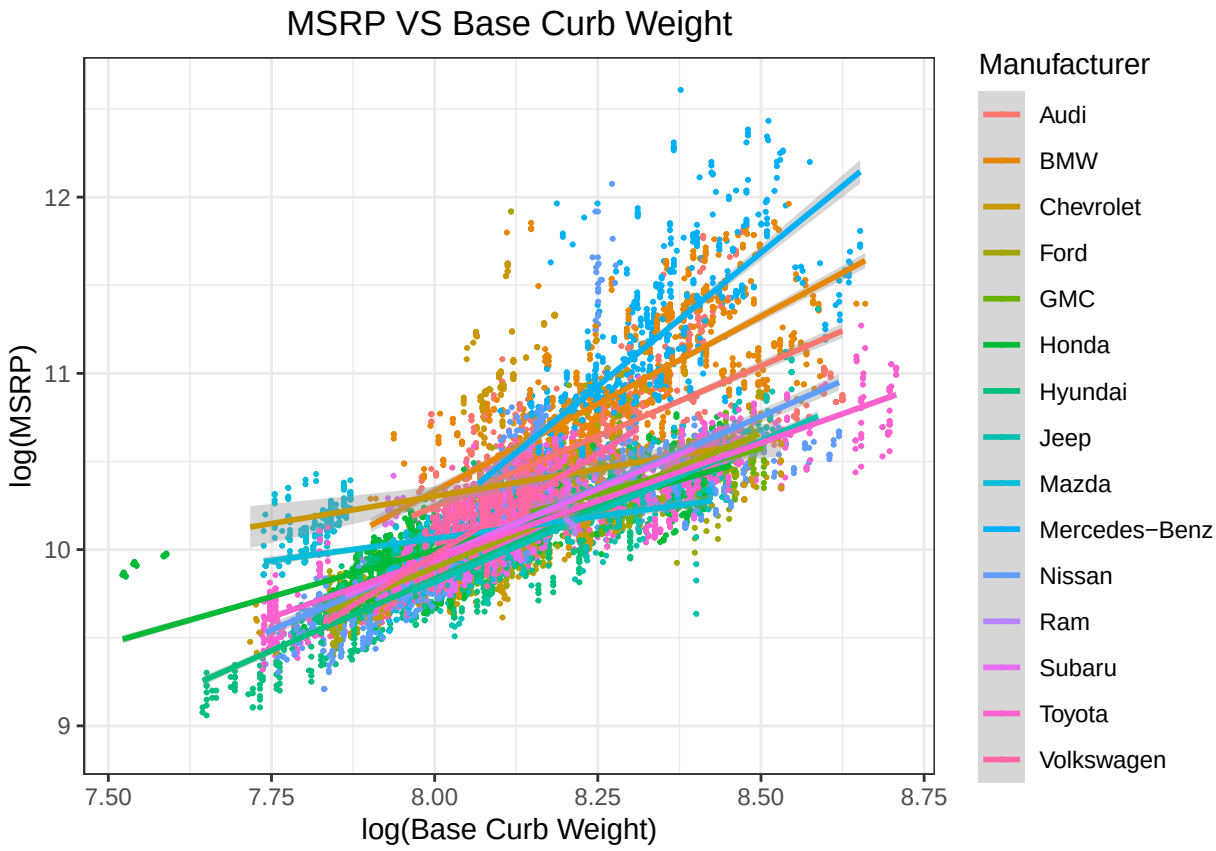
Discussion

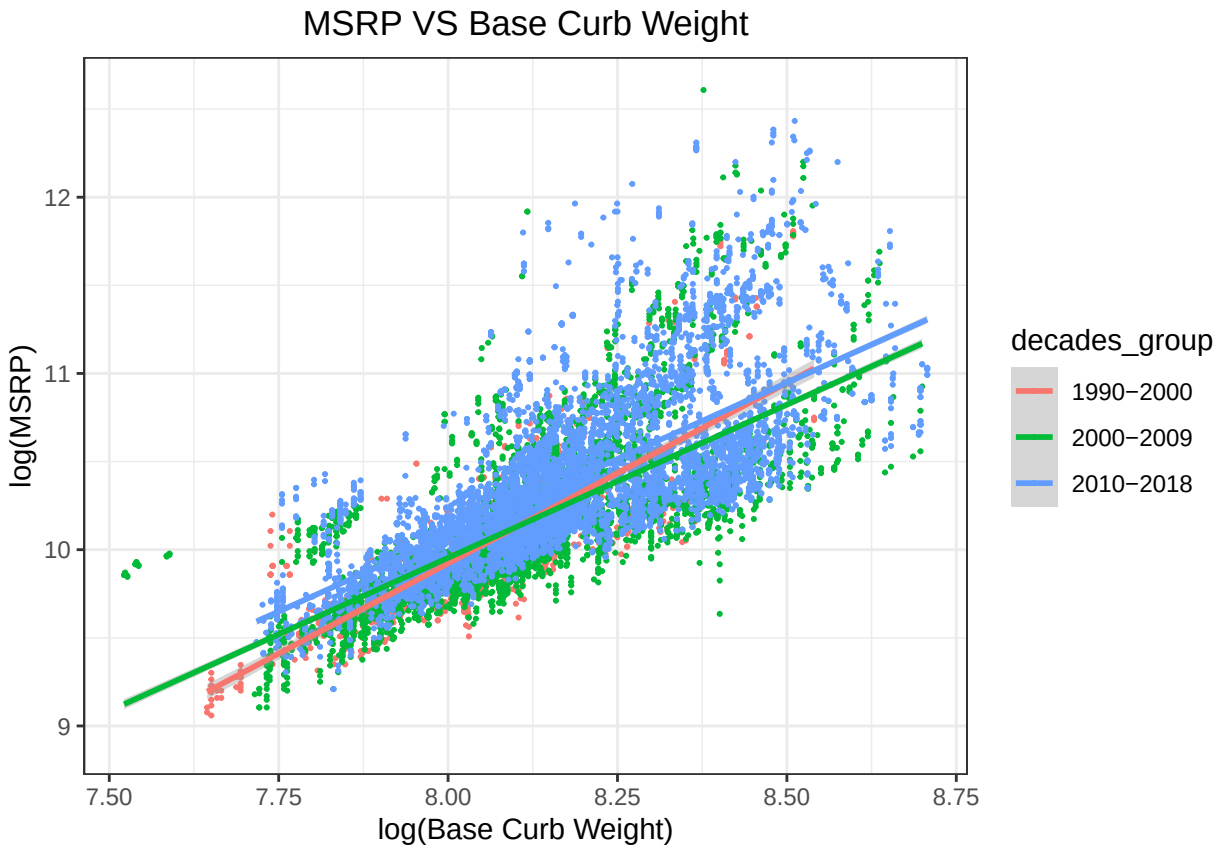
The final model is not perfect, on the contrary, it can be improved a lot. But overall, the result is quite close to what I had imagined at the beginning. The coefficient of the Transmission type, the MPG and the Passenger capacity are negative. What I didn't expect was that the Track width had such an influence on the MSRP, considering that many practical vehicles such as pickups, these cars should be more cost-effective.

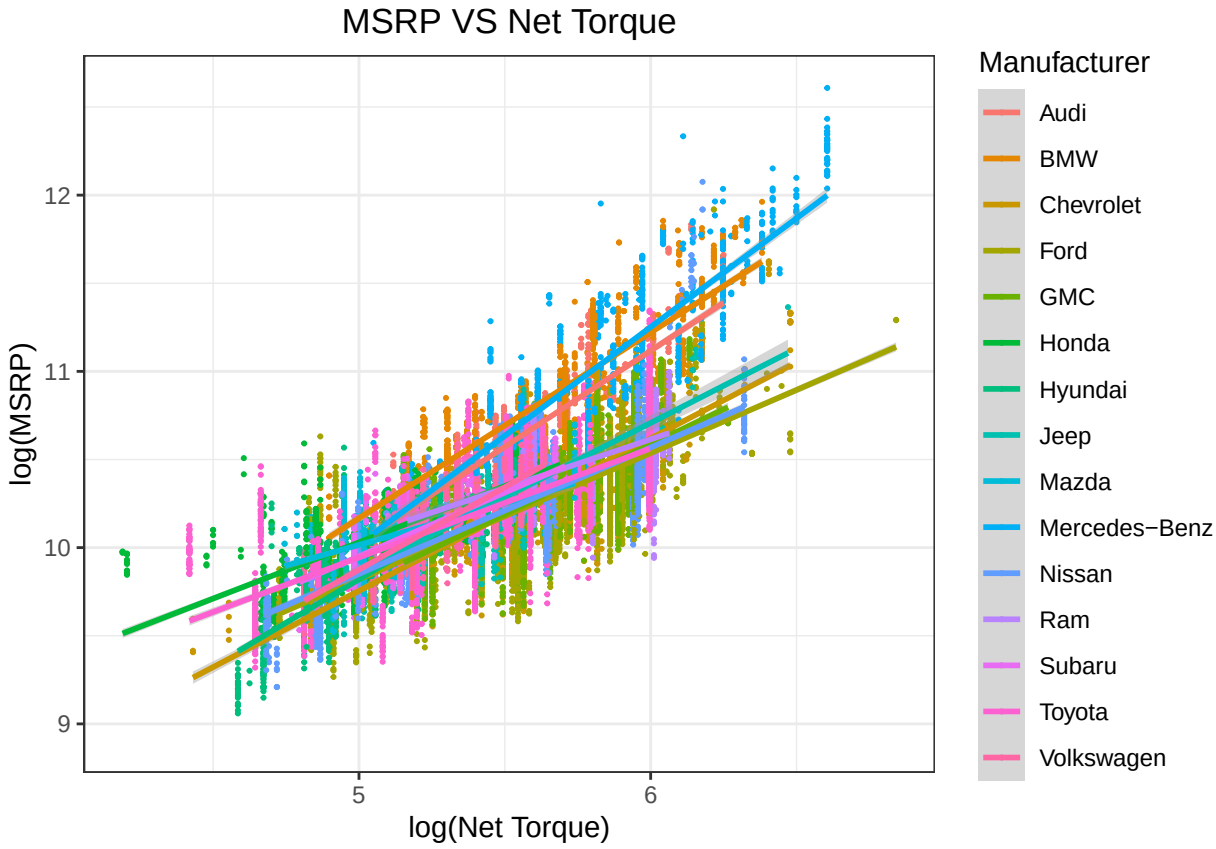
Appendix

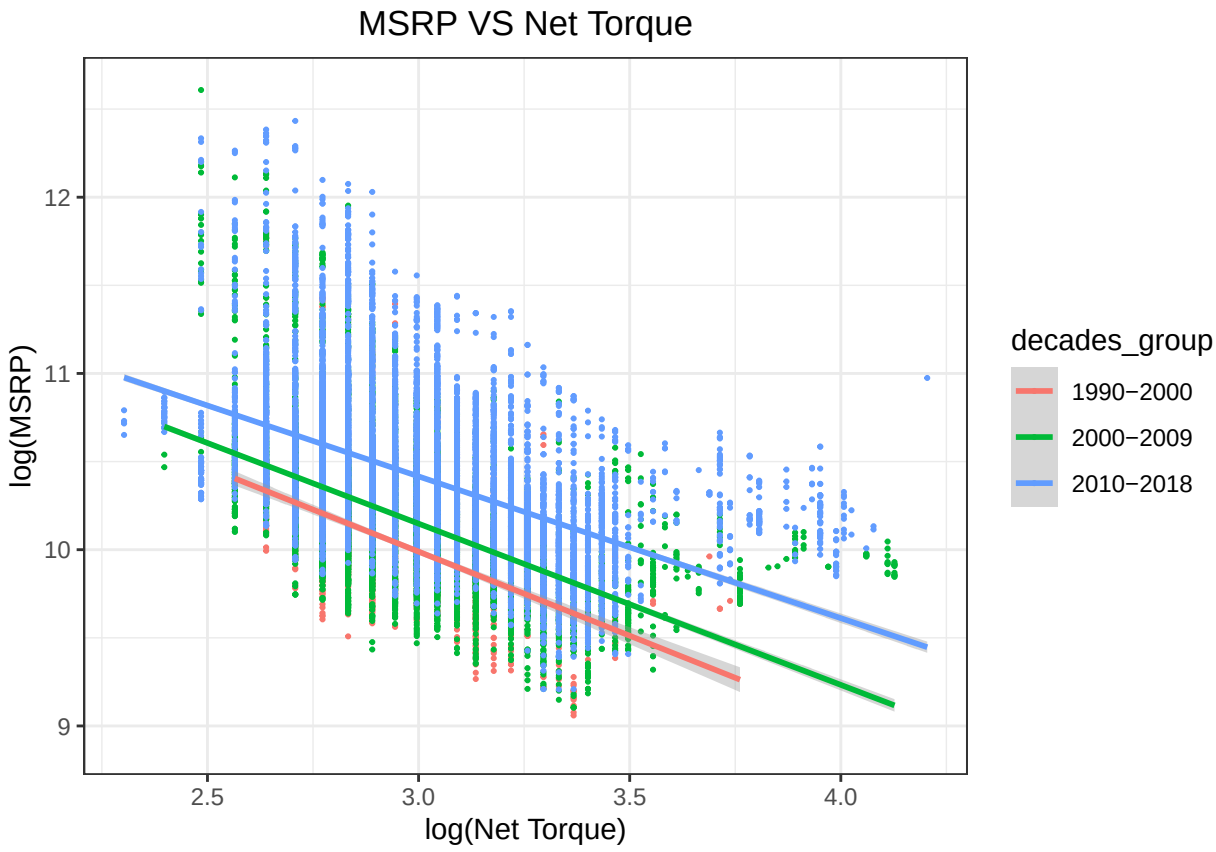


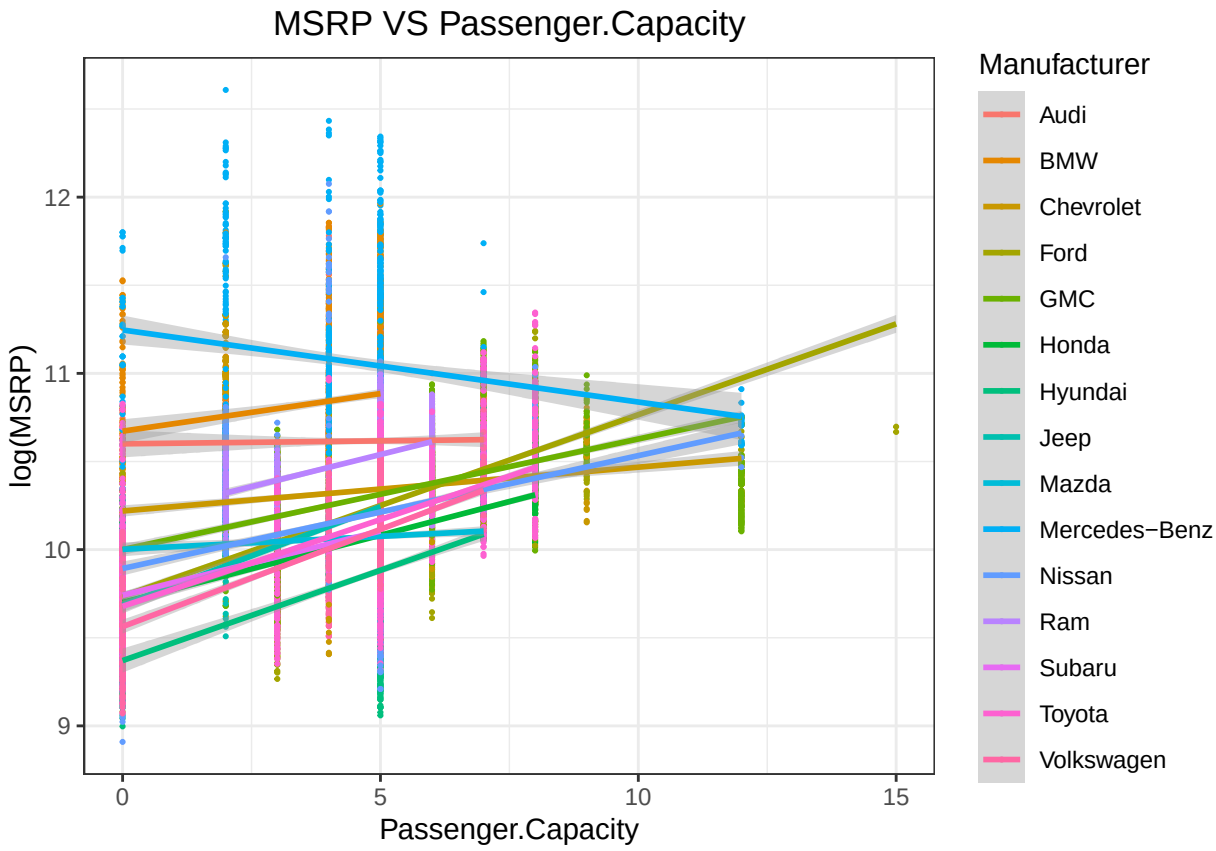

MSRP VS Track.Width
log(MSRP)
log(Track.Width)
12
11
10
9
4.05
4.10
4.15
4.20
4.25
decades_group
1990−2000
2000−2009
2010−2018

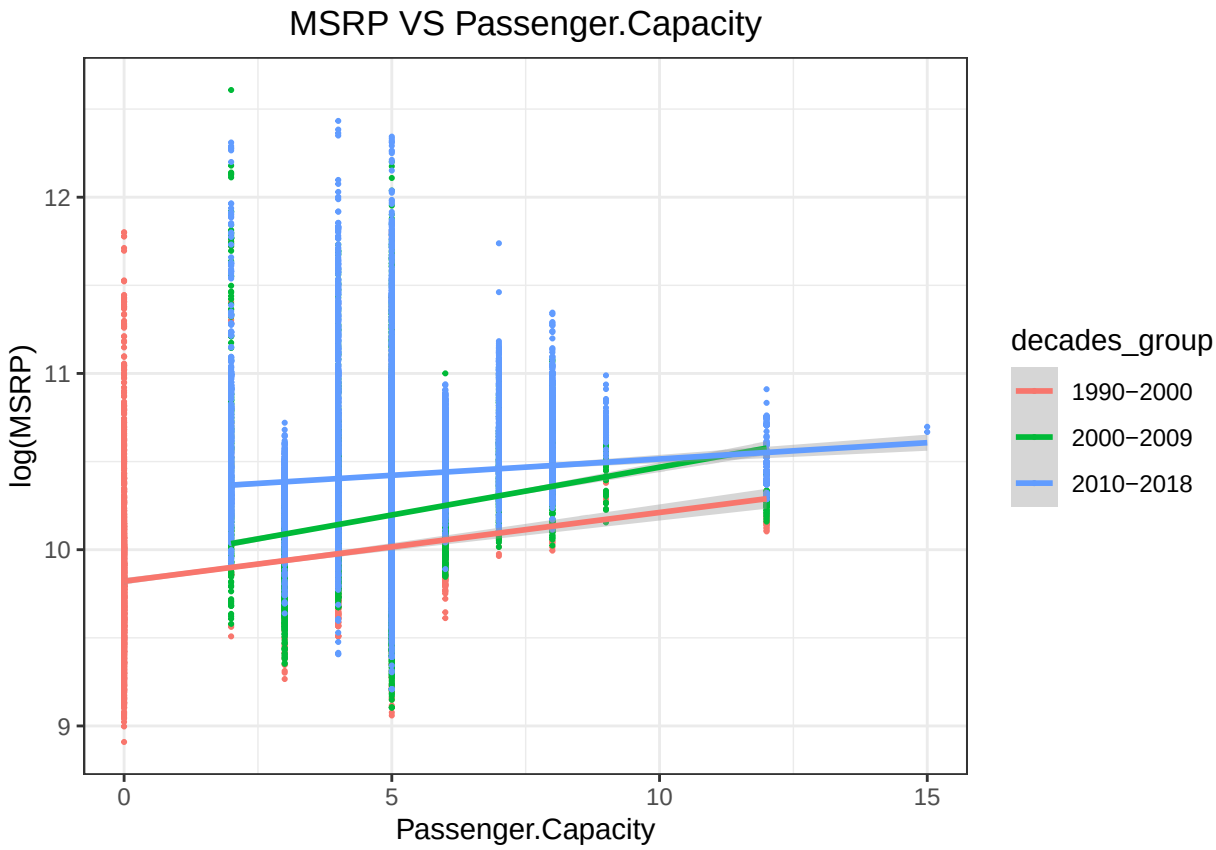

MSRP VS Base Curb Weight
log(MSRP)
log(Base Curb Weight)
12
11
10
9
7.50
7.75
8.00
8.25
8.50
8.75
Manufacturer
Audi
BMW
Chevrolet
Ford
GMC
Honda
Hyundai
Jeep
Mazda
Mercedes−Benz
Nissan
Ram
Subaru
Toyota
Volkswagen

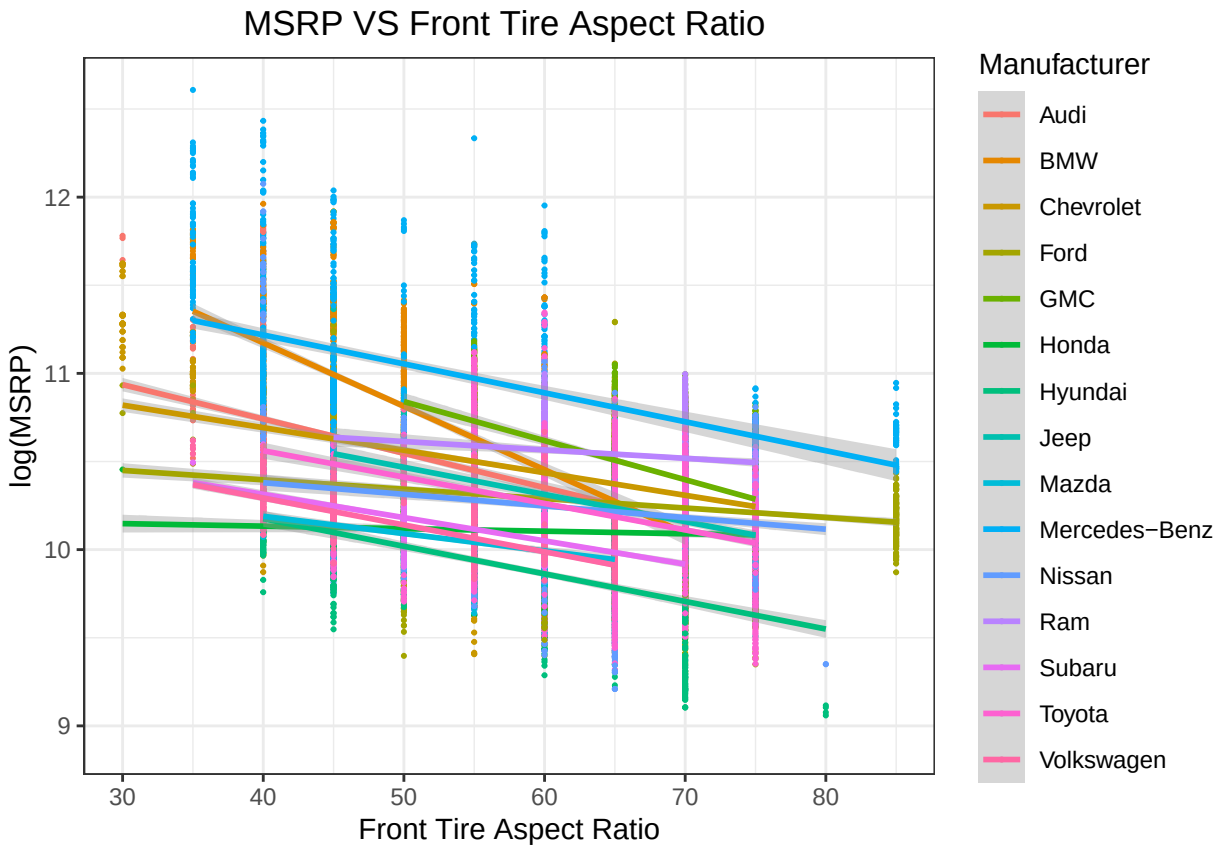

MSRP VS Base Curb Weight
log(MSRP)
log(Base Curb Weight)
12
11
10
9
7.50
7.75
8.00
8.25
8.50
8.75
decades_group
1990−2000
2000−2009
2010−2018

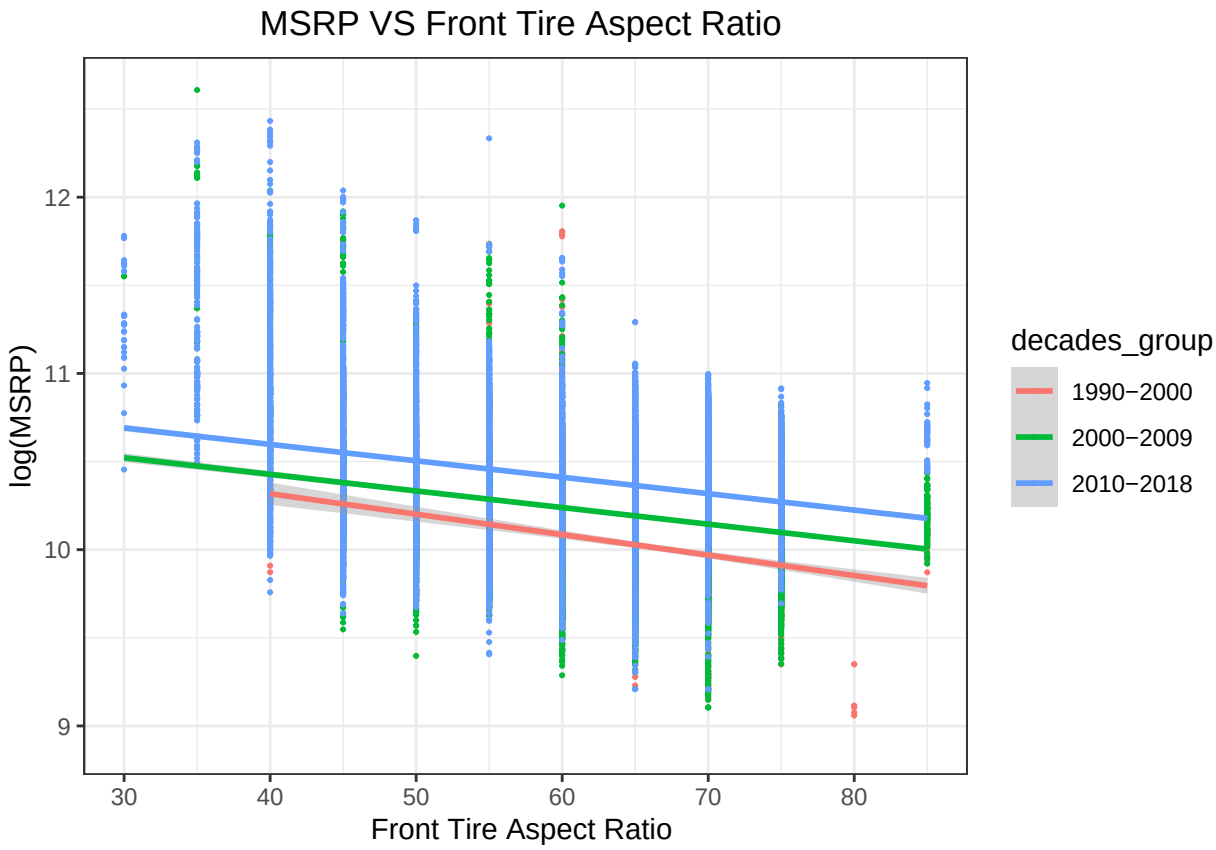

MSRP VS Net Torque
log(MSRP)
log(Net Torque)
12
11
10
9
5
6
Manufacturer
Audi
BMW
Chevrolet
Ford
GMC
Honda
Hyundai
Jeep
Mazda
Mercedes−Benz
Nissan
Ram
Subaru
Toyota
Volkswagen

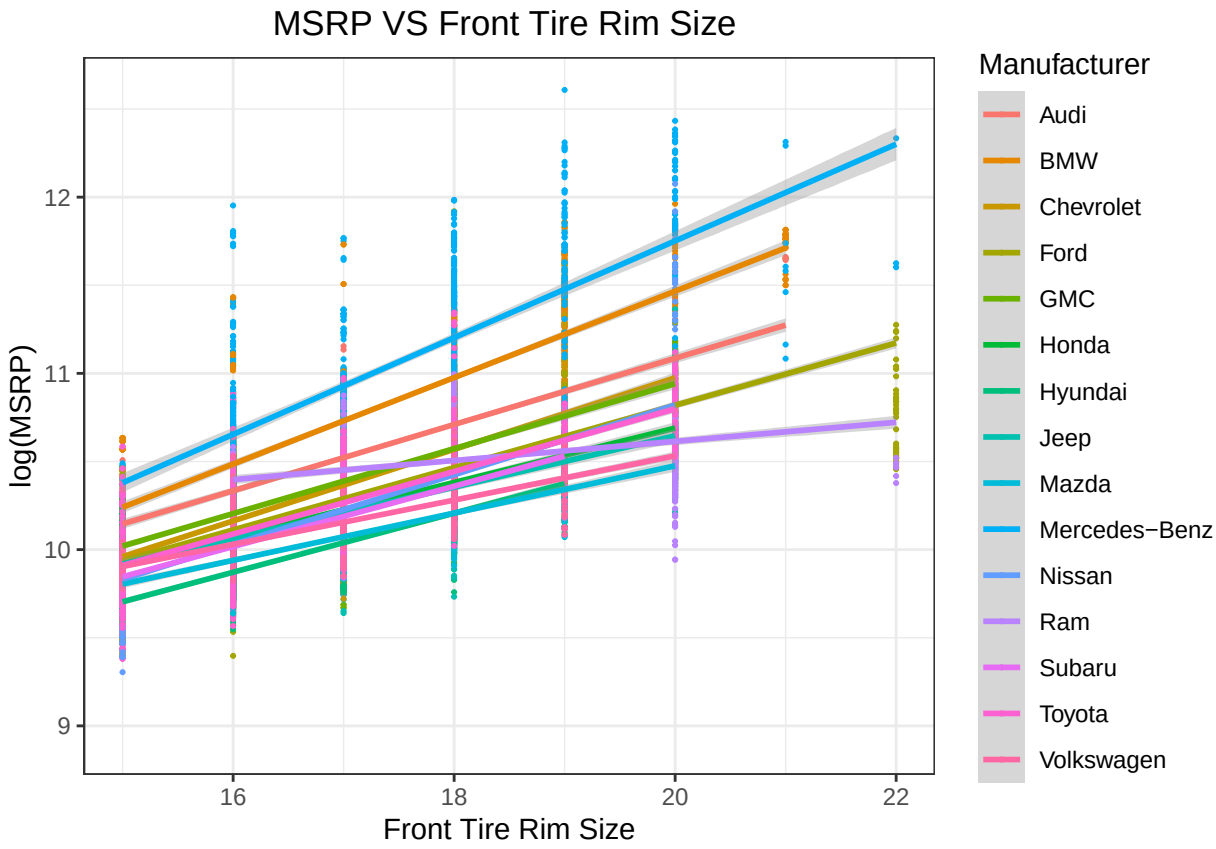

MSRP VS Net Torque
log(MSRP)
log(Net Torque)
12
11
10
9
2.5
3.0
3.5
4.0
decades_group
1990−2000
2000−2009
2010−2018

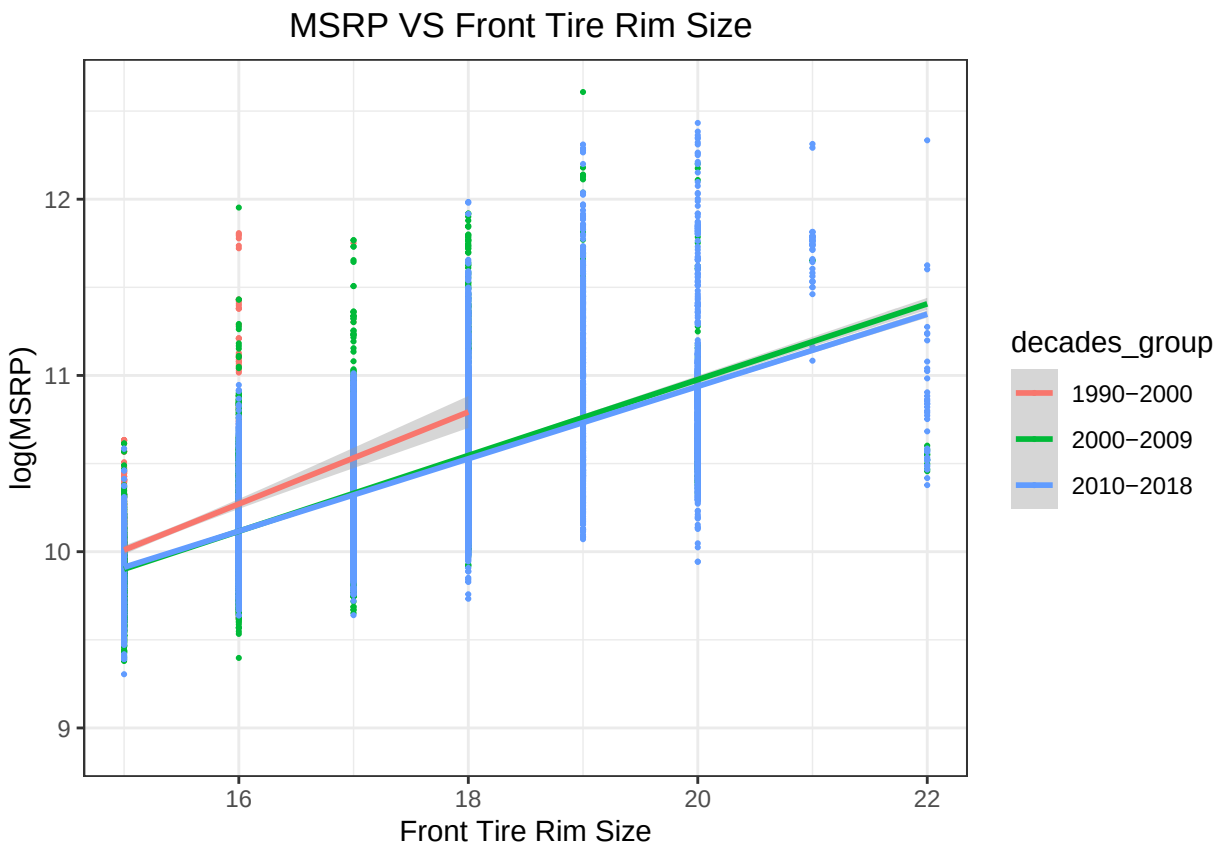

MSRP VS Passenger.Capacity
log(MSRP)
Passenger.Capacity
12
11
10
9
0
5
10
15
Manufacturer
Audi
BMW
Chevrolet
Ford
GMC
Honda
Hyundai
Jeep
Mazda
Mercedes−Benz
Nissan
Ram
Subaru
Toyota
Volkswagen


MSRP VS Passenger.Capacity
log(MSRP)
12
11
10
9
0
5
10
15
Passenger.Capacity
decades_group
1990−2000
2000−2009
2010−2018


MSRP VS Front Tire Aspect Ratio
log(MSRP)
12
11
10
9
30
40
50
60
70
80
Front Tire Aspect Ratio
Manufacturer
Audi
BMW
Chevrolet
Ford
GMC
Honda
Hyundai
Jeep
Mazda
Mercedes−Benz
Nissan
Ram
Subaru
Toyota
Volkswagen


MSRP VS Front Tire Aspect Ratio
log(MSRP)
12
11
10
9
30
40
50
60
70
80
Front Tire Aspect Ratio
decades_group
1990−2000
2000−2009
2010−2018


MSRP VS Front Tire Rim Size
log(MSRP)
12
11
10
9
16
18
20
22
Front Tire Rim Size
Manufacturer
Audi
BMW
Chevrolet
Ford
GMC
Honda
Hyundai
Jeep
Mazda
Mercedes−Benz
Nissan
Ram
Subaru
Toyota
Volkswagen


MSRP VS Front Tire Rim Size
log(MSRP)
12
11
10
9
16
18
20
22
Front Tire Rim Size
decades_group
1990−2000
2000−2009
2010−2018