MA678 Final Report

Huifei Xu

12/4/2022

Abstract

This report aims to examine the relationship between car sales and the economic attributes and traits of cars of interest to consumers or which factors primarily influence consumers' car purchase choices. To achieve the purpose of this study, I designed and built a multi-level model based on different major automobile brands or specific models as categories, and with corresponding exploratory data analysis to figure out the factors that can have a significant impact on consumers' choice of car purchase.

Introduction

Background

The automobile, the only high-tech product in the world today with tens of thousands of parts, tens of millions of annual production, and hundreds of millions of ownership, has become one of the most important modes of human transportation since its birth in 1886. In 2007, there were 800 million cars or light trucks on the road, consuming 9.8 trillion liters of fuel annually. It can be considered that cars, which can carry several people at the same time depending on the specific model, are a good choice for family travel. The overall structure of different cars may be roughly the same, but the use and manufacturing details may vary significantly from one brand of car product to another or from one model to another under the same brand. For the general public, there are several factors that may be decisive in determining exactly which model they purchase, it may be fuel efficiency, car price, motor power, etc. The purpose of this report is to examine which factors are most influential when it comes to consumer choice in purchasing a car.

Throughout the ages, taking into account the interaction of various factors and the development of the times, the car brand is still the main factor affecting consumers' choice of car. A comprehensive consideration of all factors, whether from the vehicle's cost performance, price and market performance, the brand factor is an important reason for its impact on consumers to choose a car. As a result, this report will use either the manufacturer's brand of the car or a specific special model as the categorical level to conduct an in-depth analysis to investigate the the physical attributes or economic attributes of the model itself will have an impact on its sales for different car brands.

Data Summary

The data set i fould on Kaggle is 'Car sales'(URL:https://www.kaggle.com/datasets/gagandeep16/car-sales), published by Kaggle User GAGANBHATIA. For confidentiality reasons, the exact year of provenance of the sales data is not included, but it does not affect our study of the relationship between factors and sales. A csv file was provided which includes 16 columns of variables with 157 rows of data. Amount the variables, 3 of them are categorical variables which refers to bands and models, and 13 numeric variables covers both physical and economic/energy attributes of the car. This report is only concerned with some of these

relatively representative and important variables. In order to make the model concise, some of them will be removed in the data cleaning section and will not be shown in the summary table. The table of columns with brief explanation is listed below:

Column Names	Explanation
Manufacturer	Name of the car manufacturer
Model	Name of the specific model of the car
Sales_in_thousands	Sales of cars, in thousands
Vehicle_type	Passenger-carrying properties of automobiles
Price_in_thousands	The selling price of the car, in thousands
Engine_size	The total volume of the cylinders in the engine
Horsepower	The power an engine produces
Wheelbase	The distance between the centres of the front and rear wheels
Width	Vehicle width measurement
Length	Vehicle Length measurement
Curb_weight	The weight of the vehicle including a full tank of fuel and all standard equipment
Fuel_capacity	The amount of fuel that a vehicles fuel tank can hold
Fuel_efficiency	The distance a motor vehicle can travel on a single gallon of gas

Method

Exploratory Data Analysis

Among all the twelve variables, it is very important to find out which of them have an impact on the sales of the car. The next series of eda analysis in this report is designed for this purpose. Among these variables, the price of the car is the one that we must pay attention to because it is the most intuitive data for car buyers, so I use the following graph to analyze the relationship between price and sales.

The relationship between car sales and price is well explained in Figure 1. Figure a shows the type of car as groups, while Figure b shows the brand of the car as groups. We can clearly see that both different types of cars illustrate that price is inversely proportional to sales. For car brands, the vast majority of brands maintain this same trend.

To show the correlation between the factors in car sales, we would like to introduce two types of Correlograms in order to better understand the connection between internal and external factors. In order to achieve wholeness and continuity of the chart, only numeric types of data have been chosen in the following correlation charts.

The correlogram shows correlation coefficients for all pairs of variables (with more intense colors for more extreme correlations), and correlations not significantly different from 0 are represented by a white box.

The plot above combines correlation coefficients, correlation tests (via the asterisks next to the coefficients2) and scatter plots for all possible pairs of variables present in a data set.

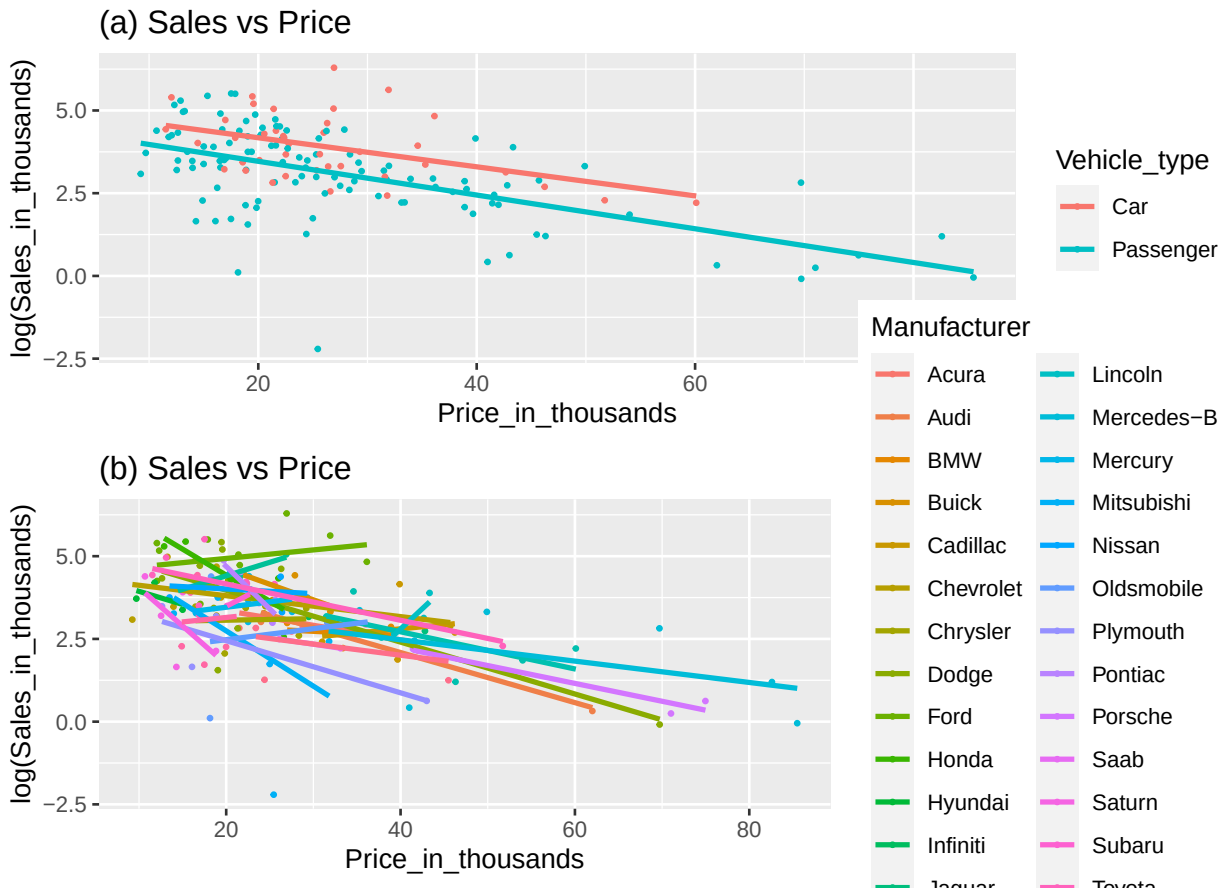


Figure 1: Relationship between Car Sales and Car Prices

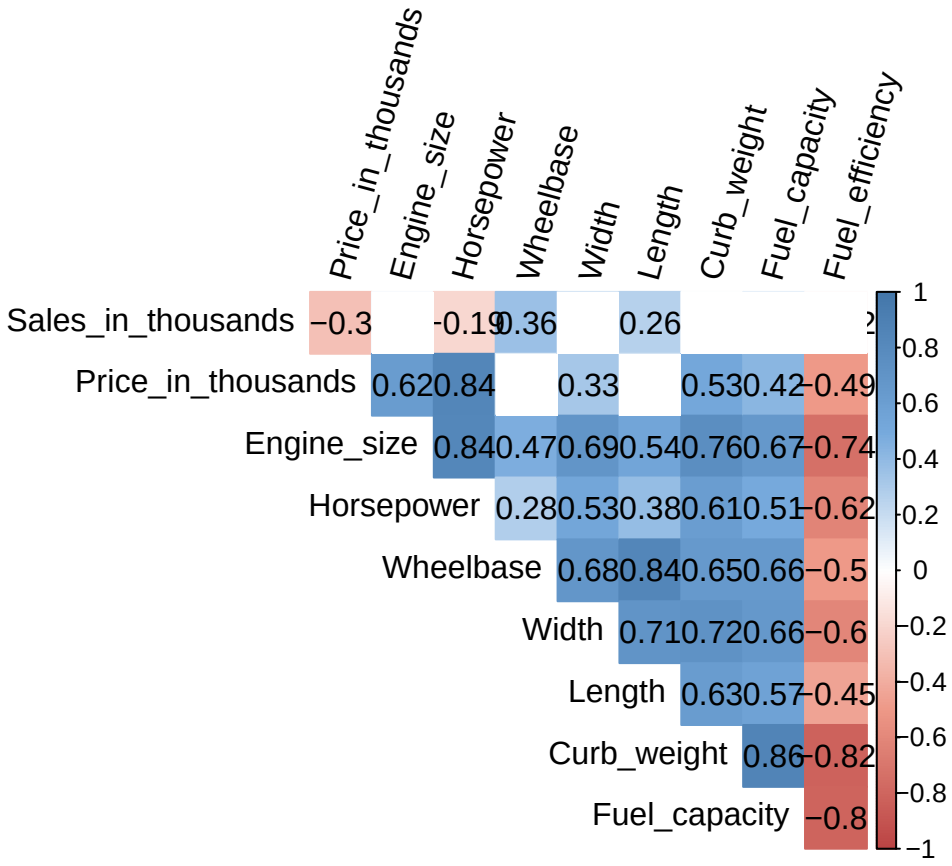


Figure 2: correlation visualization

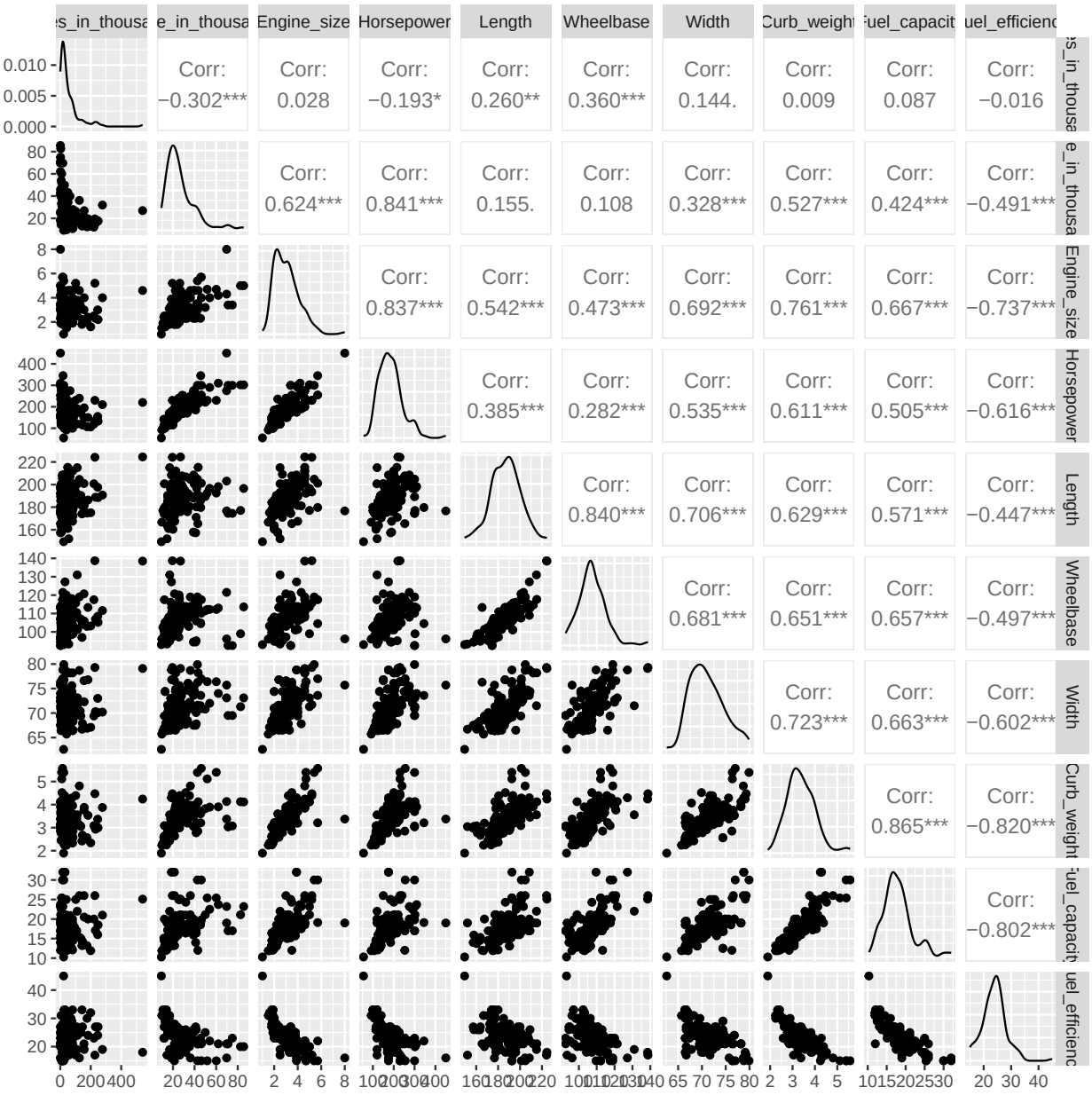


Figure 3: more correlation visualization

Model Fitting

From the previous EDAs, we can tell the type of car has slight influence on Car sales (which means whether the car is 'passenger' or 'car'). As a result, we will ignore this variable for the multilevel model and using 'Manufacturer' solely as our distinct groups. the rest of variable will be included in the model and settled to have a significance level of 0.05:

```
model <- lmer(Sales_in_thousands ~ Price_in_thousands +  Engine_size + Horsepower
            +  Length + Wheelbase + Width + Curb_weight + Fuel_capacity + Fuel_efficiency
            + (1 | Manufacturer)
            , data = car)
```

The fiexed effect is showed below:

	Estimate	Std. Error	t value
(Intercept)	-293.85525	147.63038	-1.990
Price_in_thousands	-0.51768	0.71125	-0.728
Horsepower	-0.43091	0.22895	-1.882
Engine_size	28.30864	11.48777	2.464
Length	-0.22250	0.72821	0.306
Wheelbase	5.10765	1.25094	4.083
Width	0.02709	2.33548	0.012
Curb_weight	-35.29385	19.32962	-1.826
Fuel_capacity	-1.79721	2.49386	-0.721
Fuel_efficiency	-0.27507	2.18489	-0.126

The head pf random effect towards diverse car bands will also be shown below, a full table can be found under Appendix:

```
##           (Intercept)
## Acura      -12.840529
## Audi        -3.975663
## BMW        -17.522020
## Buick       -3.300641
## Cadillac    -8.610049
## Chevrolet    1.688279
```

Result

As the result above, I am able to get the following formula of fixed effect:

$$\begin{aligned} Sales_in_thousands = &-293.85525 - 0.51768 * Price_in_thousands - 0.43091 * Horsepower \\ &+28.30864 * Engine_size - 0.22250 * Length + 0.02709 * Width \\ &-35.29385 * Curb_weight - 1.79721 * Fuel_capacity - 0.27507 * Fuel_efficiency \end{aligned}$$

By analyzing the fixed affect formula, we can tell that there are both positive and negative factors that may contribute to the sales of a car.The positive factors are Engine_size and Width, the rest of the favors are all negative. It indicates that cars with less price and basic performance ones are more favored by the car buyers. Also, it is interesting to see fuel statistics have negative impact on Car sales, which is quite different from what i expected before having the model results.

Move on to the random effect, It is very interesting to see some Manufacturers have quite different coefficient, especially between affordable car brands such as Honda Ford versus relatively luxurious car brands such as BMW and Acura.The overall results shows a clear effect of the the impact of the strong influence of car brands on car sales. For example, the Honda has a significant higher intercept (77) than the average , while BMW on the other side has negative intercept value.

Discussion

This report explains the various types of factors, physical and economic, grouped with the corresponding sales of cars. Among them, we grouped the cars according to their manufacturers, and we can say that by analyzing the results of the EDA and multilevel models, we obtained roughly similar results to those previously estimated, corroborating our conjectures.We can clearly see that small, cheap and affordable models are more favored by consumers, while performance cars, despite the predominance of sporting attributes, but also clearly raise the price of the car, and therefore did not get too high sales.

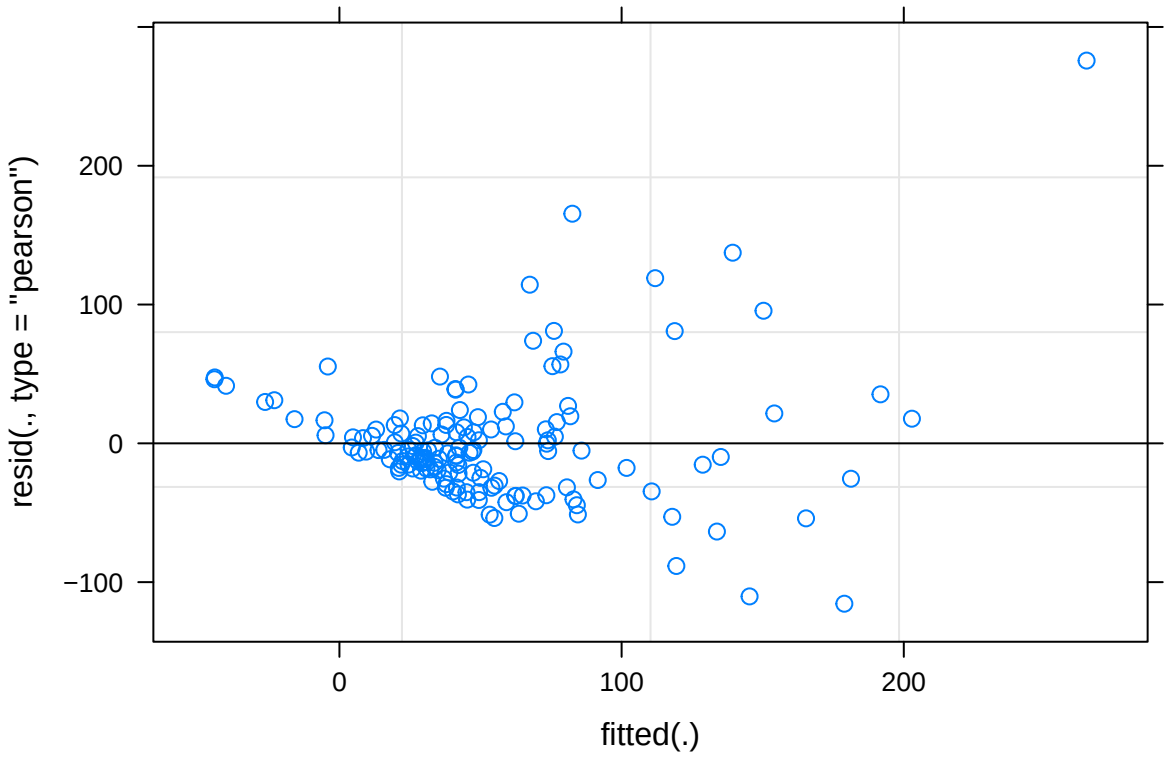
There is limitation for this report. One of the biggest concern for this research is that we did not include too much for the omitted variable bias.We did not go into the specific relationships, but the existence of interrelationships among the variables is predictable. One possible correlation is that the car length and width may have been designed in certain proportions. There is also the fact that better motors and more power tend to represent higher prices.

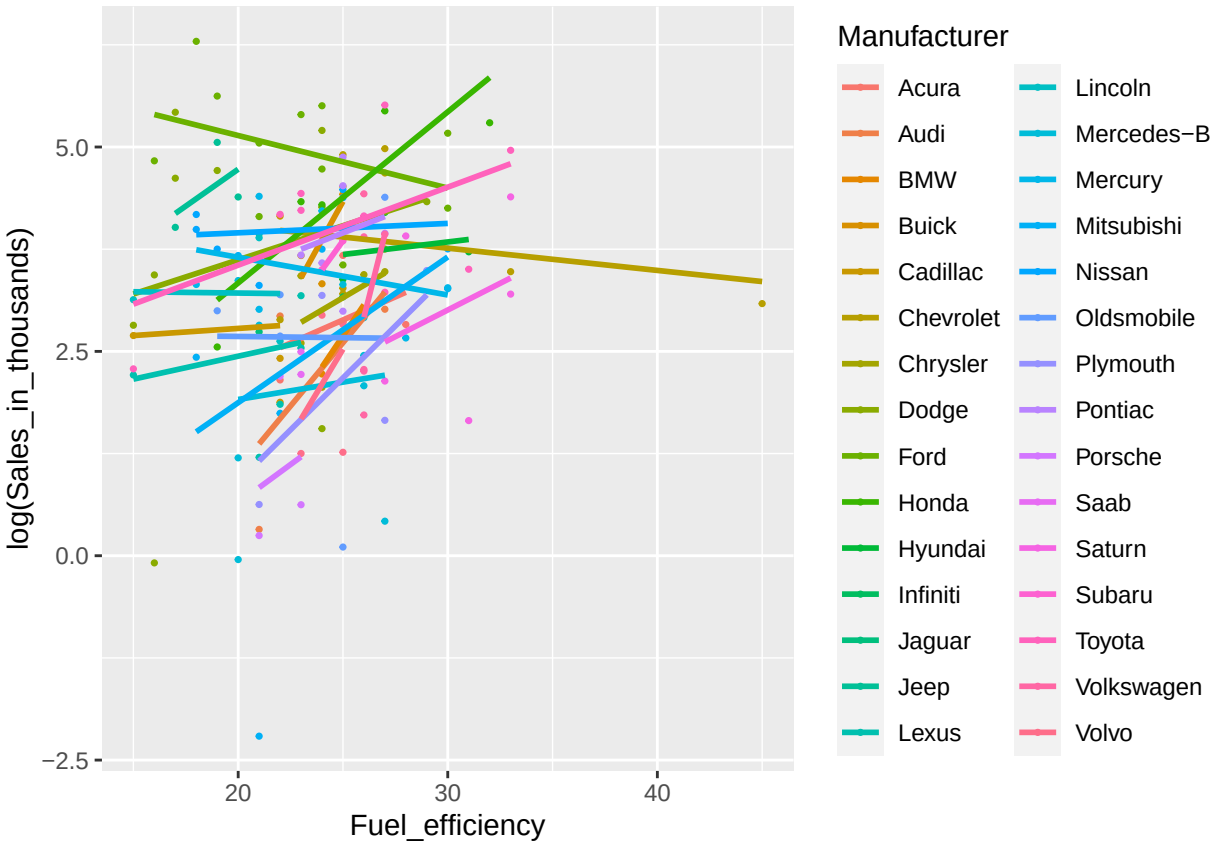
Ciatation

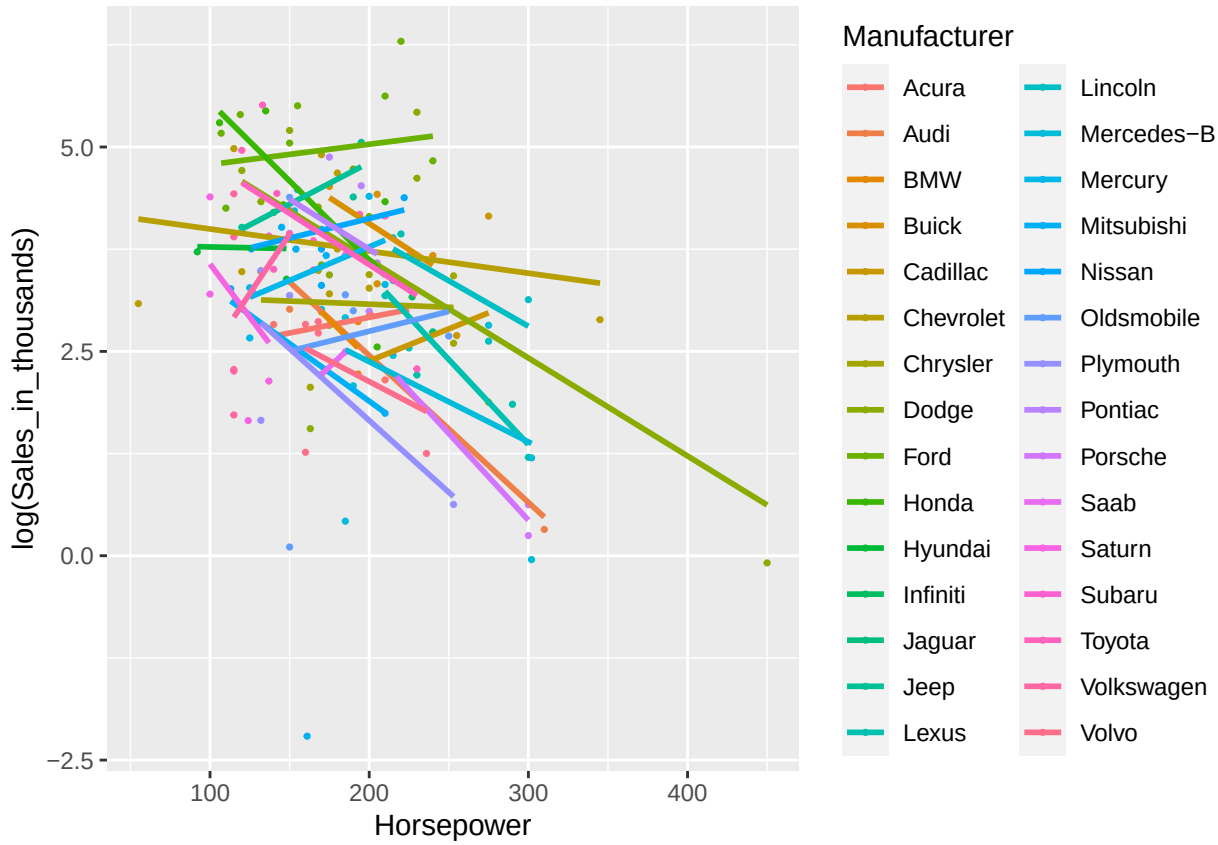
[1] Correlation coefficient and correlation test in R; Antoine Soetewey; 2020-05-28

Appendix

More EDA




log(Sales_in_thousands)
5.0
2.5
0.0
−2.5
20
30
40
Fuel_efficiency
Manufacturer
Acura
Audi
BMW
Buick
Cadillac
Chevrolet
Chrysler
Dodge
Ford
Honda
Hyundai
Infiniti
Jaguar
Jeep
Lexus
Lincoln
Mercedes−B
Mercury
Mitsubishi
Nissan
Oldsmobile
Plymouth
Pontiac
Porsche
Saab
Saturn
Subaru
Toyota
Volkswagen
Volvo



model summary

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Sales_in_thousands ~ Price_in_thousands + Engine_size + Horsepower +
##     Length + Wheelbase + Width + Curb_weight + Fuel_capacity +
##     Fuel_efficiency + (1 | Manufacturer)
##    Data: car
## REML criterion at convergence: 1609.168
## Random effects:
##  Groups       Name        Std.Dev.
##  Manufacturer (Intercept) 30.29
##  Residual                 49.67
## Number of obs: 152, groups:  Manufacturer, 30
## Fixed Effects:
##        (Intercept)  Price_in_thousands         Engine_size          Horsepower
##          -293.85525            -0.51768            28.30864            -0.43091
##              Length           Wheelbase               Width         Curb_weight
##            -0.22250             5.10765             0.02709           -35.29385
##      Fuel_capacity     Fuel_efficiency
##            -1.79721            -0.27507

## $Manufacturer
##            (Intercept) Price_in_thousands Engine_size Horsepower     Length
## Acura        -306.6958         -0.5176765    28.30864  -0.430914 -0.2224996
## Audi         -297.8309         -0.5176765    28.30864  -0.430914 -0.2224996
```

```
## BMW          -311.3773        -0.5176765    28.30864  -0.430914 -0.2224996
## Buick        -297.1559        -0.5176765    28.30864  -0.430914 -0.2224996
## Cadillac     -302.4653        -0.5176765    28.30864  -0.430914 -0.2224996
## Chevrolet    -292.1670        -0.5176765    28.30864  -0.430914 -0.2224996
## Chrysler     -313.4752        -0.5176765    28.30864  -0.430914 -0.2224996
## Dodge        -297.8087        -0.5176765    28.30864  -0.430914 -0.2224996
## Ford         -216.5495        -0.5176765    28.30864  -0.430914 -0.2224996
## Honda        -251.3595        -0.5176765    28.30864  -0.430914 -0.2224996
## Hyundai      -297.2312        -0.5176765    28.30864  -0.430914 -0.2224996
## Infiniti     -296.0700        -0.5176765    28.30864  -0.430914 -0.2224996
## Jaguar       -301.0541        -0.5176765    28.30864  -0.430914 -0.2224996
## Jeep         -262.7248        -0.5176765    28.30864  -0.430914 -0.2224996
## Lexus        -289.9486        -0.5176765    28.30864  -0.430914 -0.2224996
## Lincoln      -304.5418        -0.5176765    28.30864  -0.430914 -0.2224996
## Mercedes-B   -291.9923        -0.5176765    28.30864  -0.430914 -0.2224996
## Mercury      -317.1681        -0.5176765    28.30864  -0.430914 -0.2224996
## Mitsubishi   -301.9897        -0.5176765    28.30864  -0.430914 -0.2224996
## Nissan       -290.7767        -0.5176765    28.30864  -0.430914 -0.2224996
## Oldsmobile   -320.8256        -0.5176765    28.30864  -0.430914 -0.2224996
## Plymouth     -330.5774        -0.5176765    28.30864  -0.430914 -0.2224996
## Pontiac      -298.8357        -0.5176765    28.30864  -0.430914 -0.2224996
## Porsche      -263.0821        -0.5176765    28.30864  -0.430914 -0.2224996
## Saab         -298.1442        -0.5176765    28.30864  -0.430914 -0.2224996
## Saturn       -319.6100        -0.5176765    28.30864  -0.430914 -0.2224996
## Subaru       -286.5616        -0.5176765    28.30864  -0.430914 -0.2224996
## Toyota       -264.1514        -0.5176765    28.30864  -0.430914 -0.2224996
## Volkswagen   -297.5448        -0.5176765    28.30864  -0.430914 -0.2224996
## Volvo        -295.9424        -0.5176765    28.30864  -0.430914 -0.2224996
##             Wheelbase      Width Curb_weight Fuel_capacity Fuel_efficiency
## Acura        5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Audi         5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## BMW          5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Buick        5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Cadillac     5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Chevrolet    5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Chrysler     5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Dodge        5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Ford         5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Honda        5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Hyundai      5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Infiniti     5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Jaguar       5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Jeep         5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Lexus        5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Lincoln      5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Mercedes-B   5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Mercury      5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Mitsubishi   5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Nissan       5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Oldsmobile   5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Plymouth     5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Pontiac      5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Porsche      5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Saab         5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
```

```
## Saturn      5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Subaru      5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Toyota      5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Volkswagen  5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
## Volvo       5.107651 0.02709038   -35.29385     -1.797206      -0.2750737
##
## attr(,"class")
## [1] "coef.mer"
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Sales_in_thousands ~ Price_in_thousands + Engine_size + Horsepower +
##     Length + Wheelbase + Width + Curb_weight + Fuel_capacity +
##     Fuel_efficiency + (1 | Manufacturer)
##    Data: car
##
## REML criterion at convergence: 1609.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3260 -0.4297 -0.1276  0.2655  5.5519
##
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  Manufacturer (Intercept)  917.3   30.29
##  Residual                 2467.4   49.67
## Number of obs: 152, groups:  Manufacturer, 30
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)        -293.85525  147.63038  -1.990
## Price_in_thousands   -0.51768    0.71125  -0.728
## Engine_size          28.30864   11.48777   2.464
## Horsepower           -0.43091    0.22895  -1.882
## Length               -0.22250    0.72821  -0.306
## Wheelbase             5.10765    1.25094   4.083
## Width                 0.02709    2.33548   0.012
## Curb_weight         -35.29385   19.32962  -1.826
## Fuel_capacity        -1.79721    2.49386  -0.721
## Fuel_efficiency      -0.27507    2.18489  -0.126
##
## Correlation of Fixed Effects:
##             (Intr) Prc_n_ Engn_s Hrspwr Length Whelbs Width  Crb_wg Fl_cpc
## Prc_n_thsnd -0.077
## Engine_size  0.106  0.105
## Horsepower  -0.045 -0.661 -0.629
## Length       0.059  0.120 -0.050 -0.129
## Wheelbase   -0.165  0.089  0.077  0.024 -0.694
## Width       -0.671  0.090 -0.151 -0.087 -0.225 -0.035
## Curb_weight  0.021 -0.247 -0.213  0.242 -0.110 -0.059 -0.263
## Fuel_capcty -0.008 -0.093 -0.002  0.097  0.097 -0.270 -0.039 -0.445
## Fuel_ffcncy -0.446 -0.131  0.191  0.144 -0.129  0.018 -0.106  0.386  0.246
```

```
## Analysis of Variance Table
##                    npar  Sum Sq Mean Sq F value
```

```
## Price_in_thousands    1 23732.5 23732.5  9.6186
## Engine_size           1 20296.6 20296.6  8.2260
## Horsepower            1  5238.1  5238.1  2.1230
## Length                1 29851.1 29851.1 12.0984
## Wheelbase             1 24693.2 24693.2 10.0079
## Width                 1  2425.4  2425.4  0.9830
## Curb_weight           1 21590.1 21590.1  8.7503
## Fuel_capacity         1  1249.3  1249.3  0.5063
## Fuel_efficiency       1    39.1    39.1  0.0159
```