3 wise suggestions
to your next automobile choice

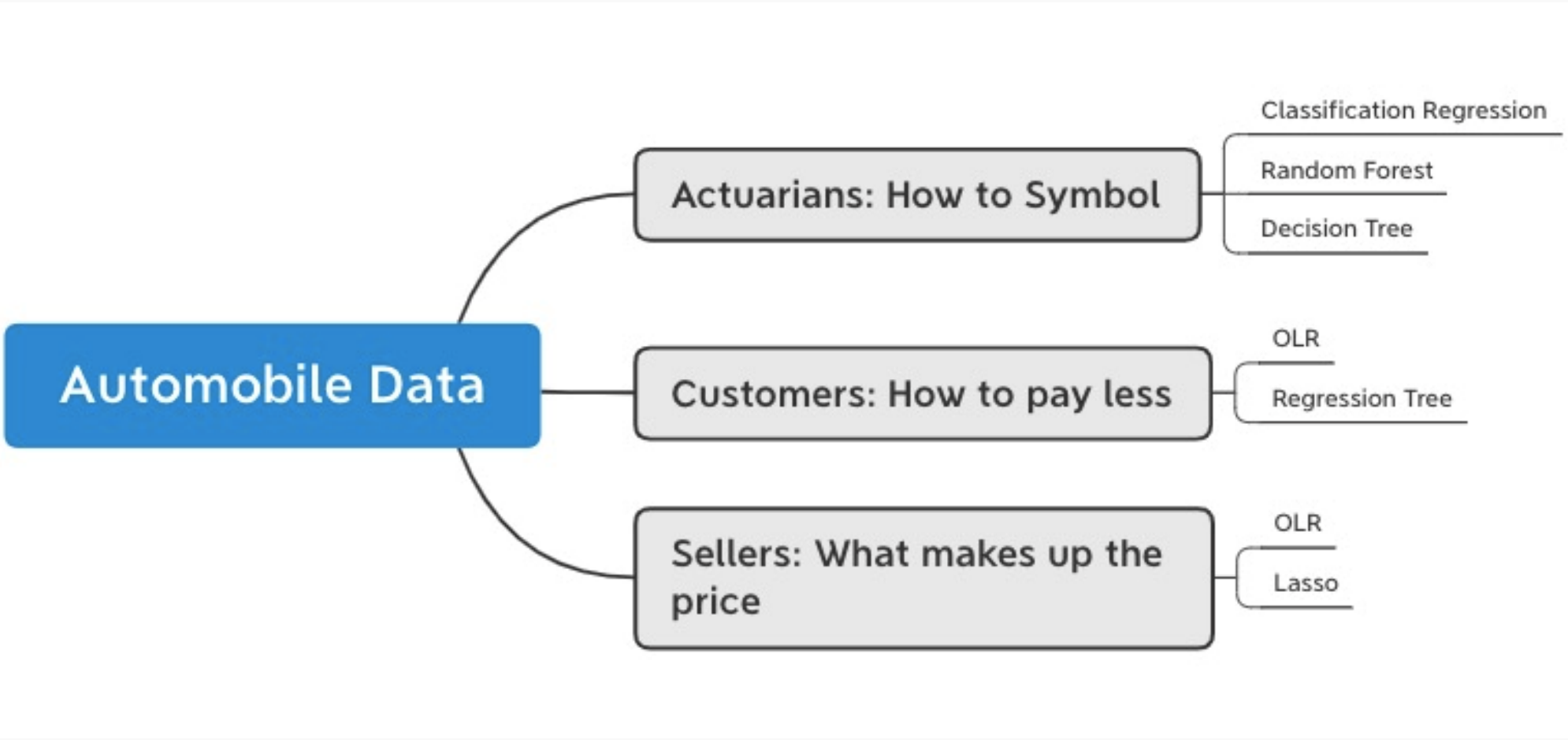
Analysis based on the automobile dataset

Group 12

Yuheng Ma,

Yisha Ma

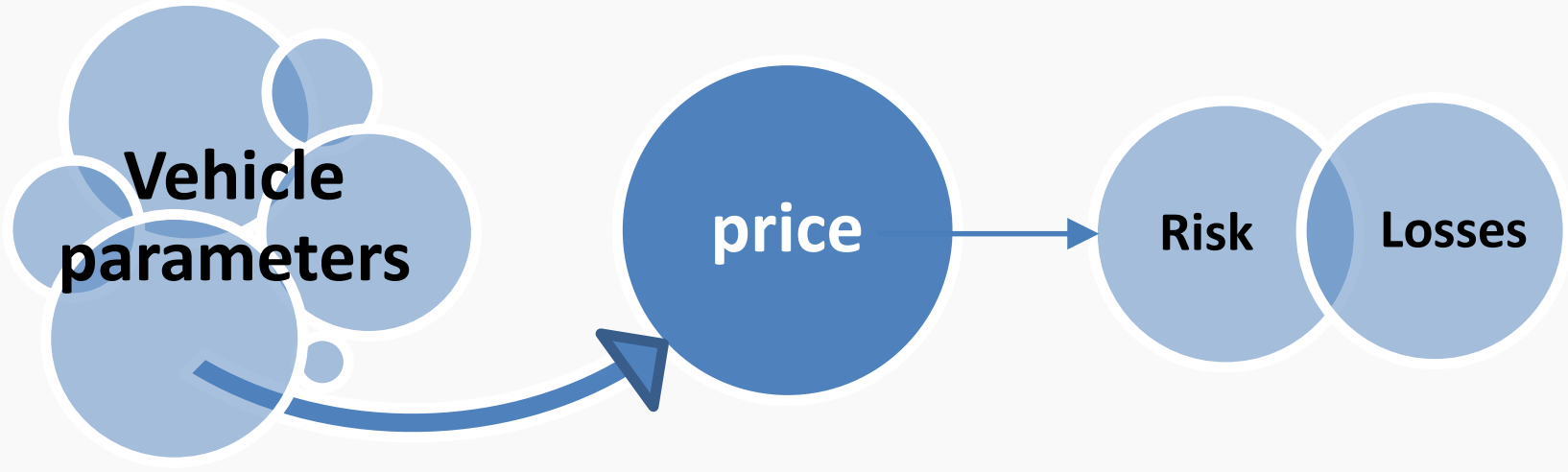
Summary


Automobile Data
Actuarians: How to Symbol
Classification Regression
Random Forest
Decision Tree
Customers: How to pay less
OLR
Regression Tree
Sellers: What makes up the price
OLR
Lasso

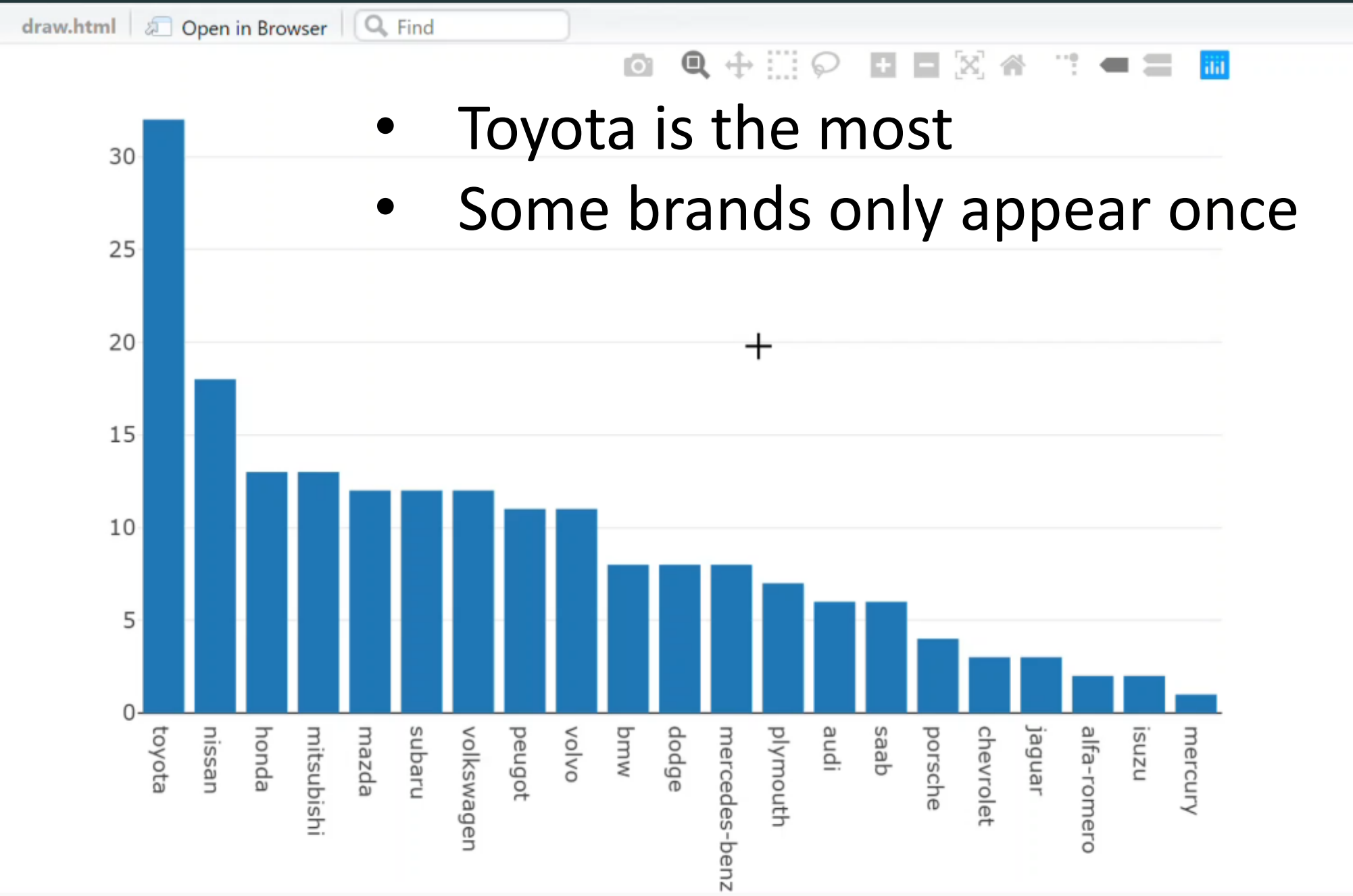
Data exploring

26 Attributes:

- 23 Vehicle parameters (*fitted*)
- 1 Price (*fitted*)
- 2 Risk and Losses (*alterable*)

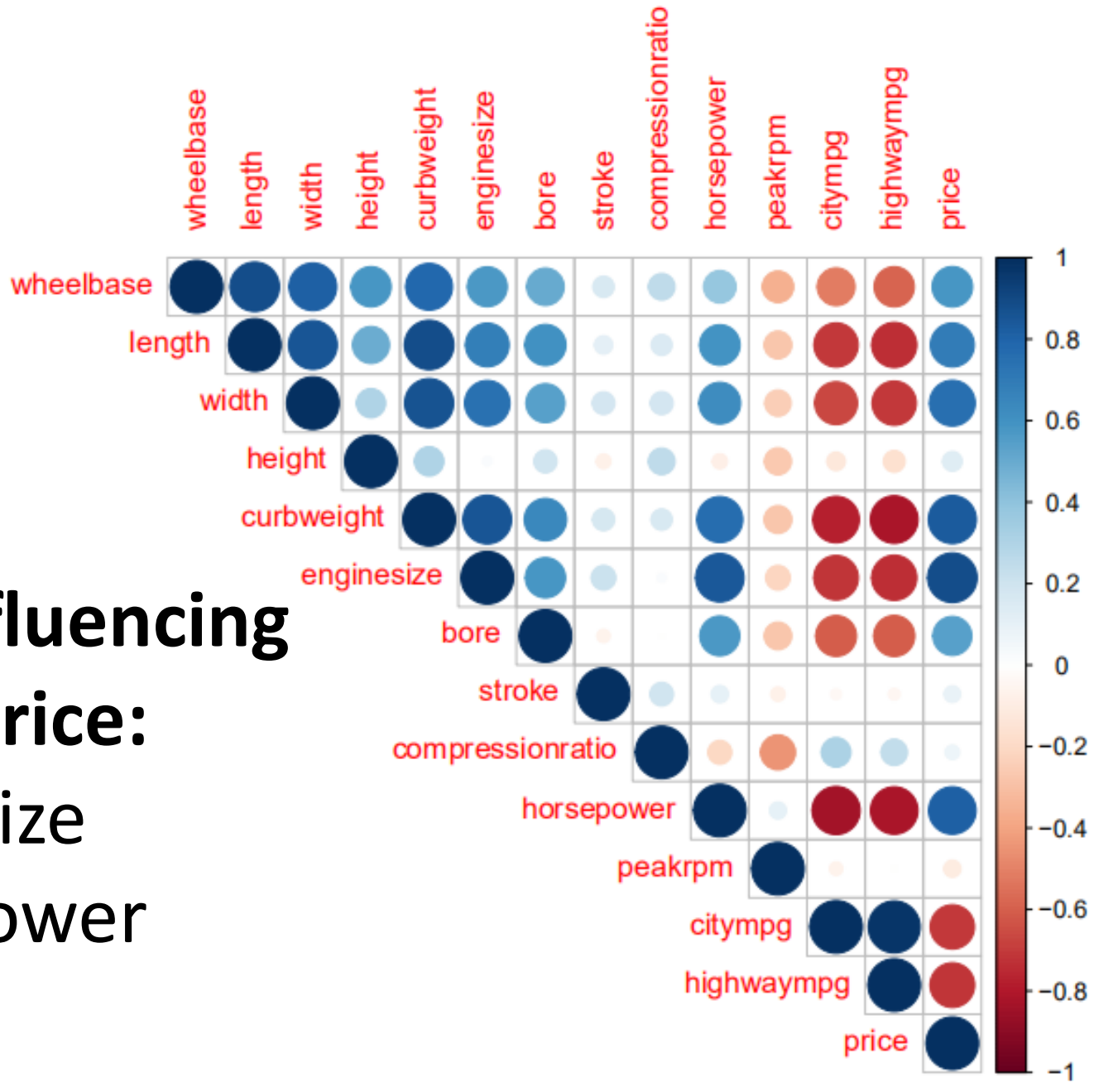


Data exploring - Vehicle parameters



- Toyota is the most
- Some brands only appear once

Data exploring - Price



Possible influencing factors of Price:

- Engine size
- Horse power
- Mpg

Data exploring – Price vs Symboling

There is **no clear relationship** between price and symboling.

Data exploring - Risk and Losses

Data exploring – Preprocess

- **Missing values in normalized losses**

①Why not mean/median method?

②Correlation with symboling

↓

Fill by the suitable value using risks

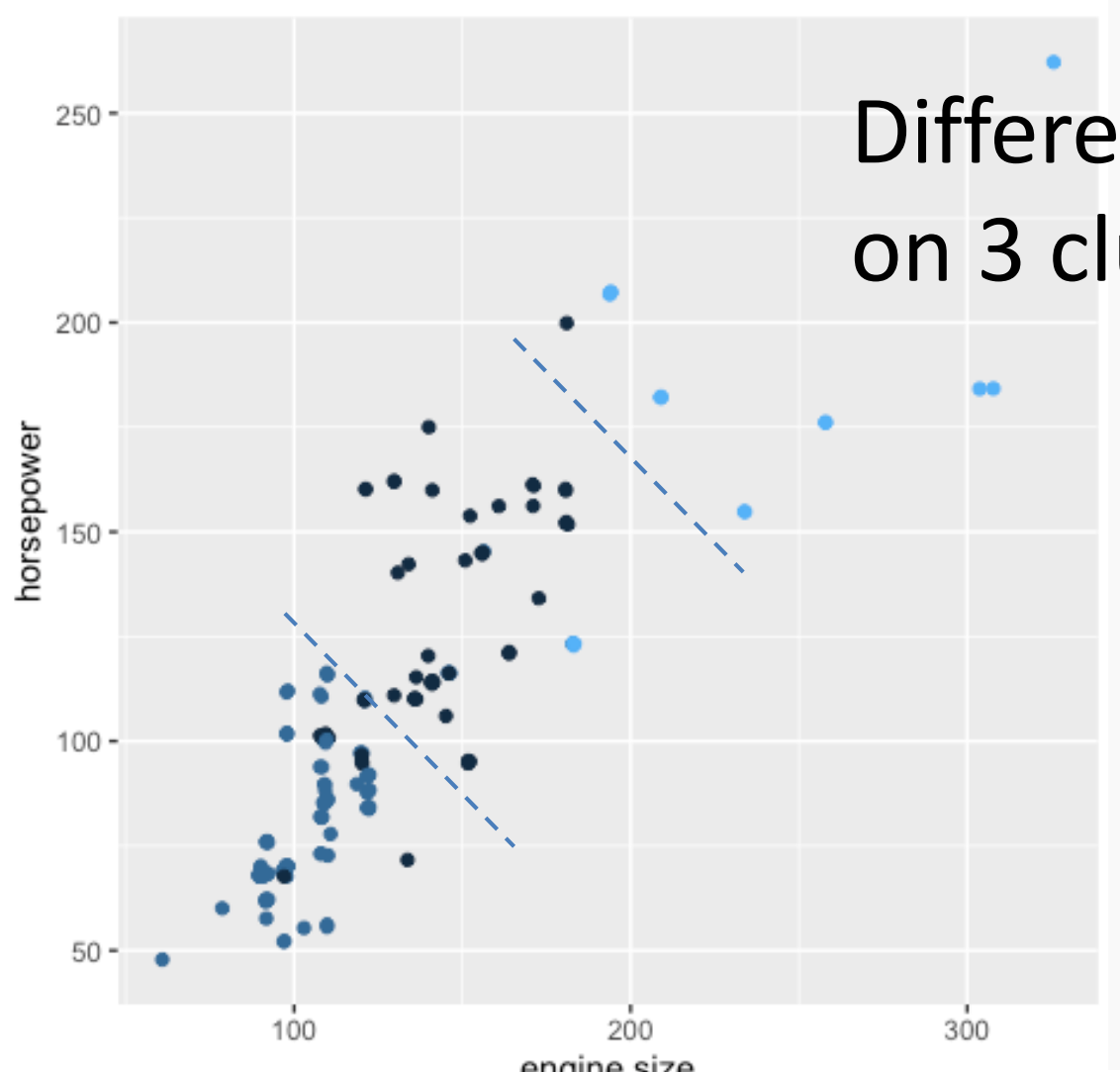
- **Dummy variable**

↓

0-1 Transform

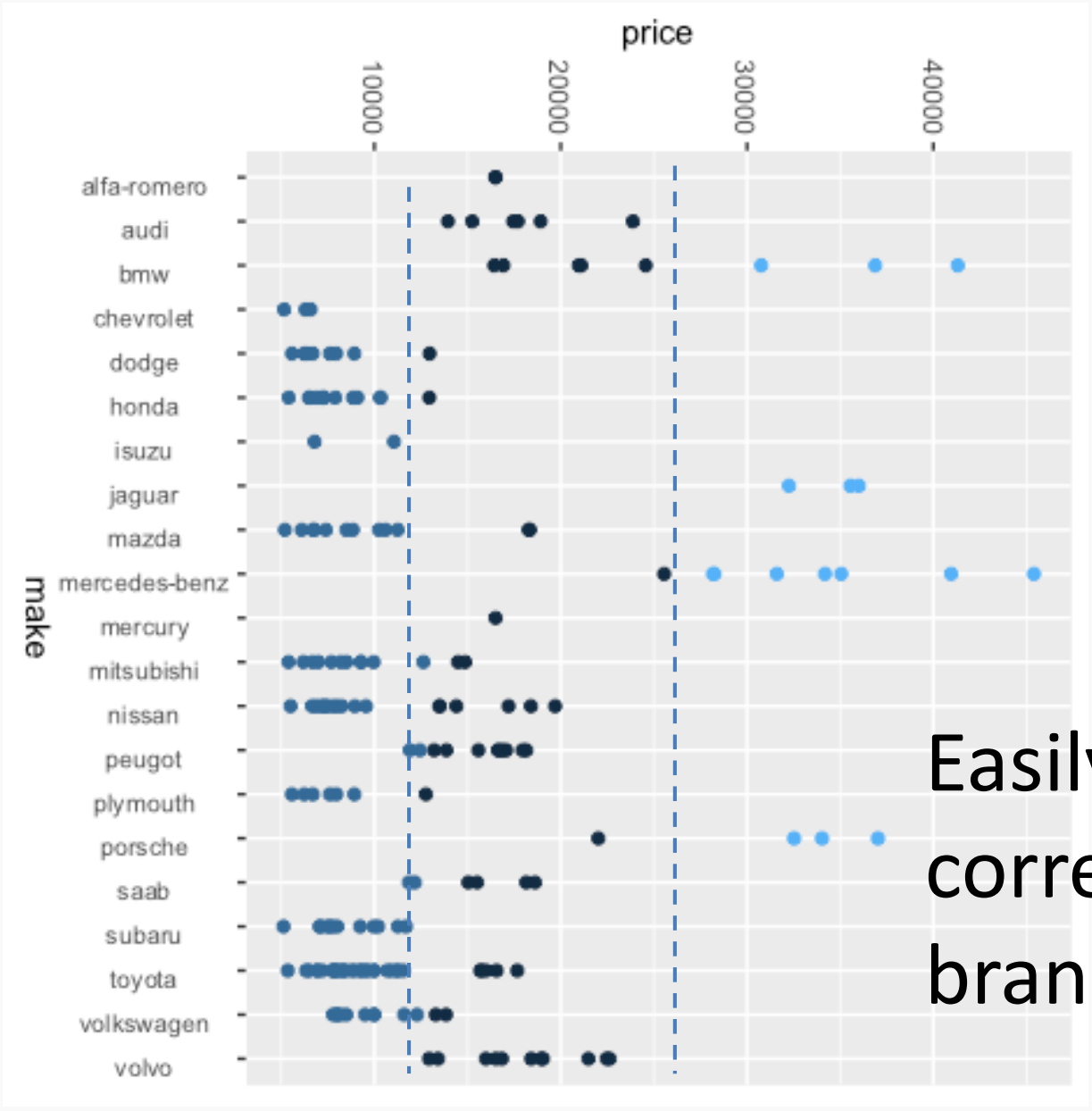
k-means clustering

Get the cluster labels of all the samples



Different parameters on 3 clusters.

Data exploring – Clustering



Easily get the corresponding brands by price need!

What affects the Risk?

Symboling: an orderly categorical variable

↓

Ordered Multiple Classification Regression

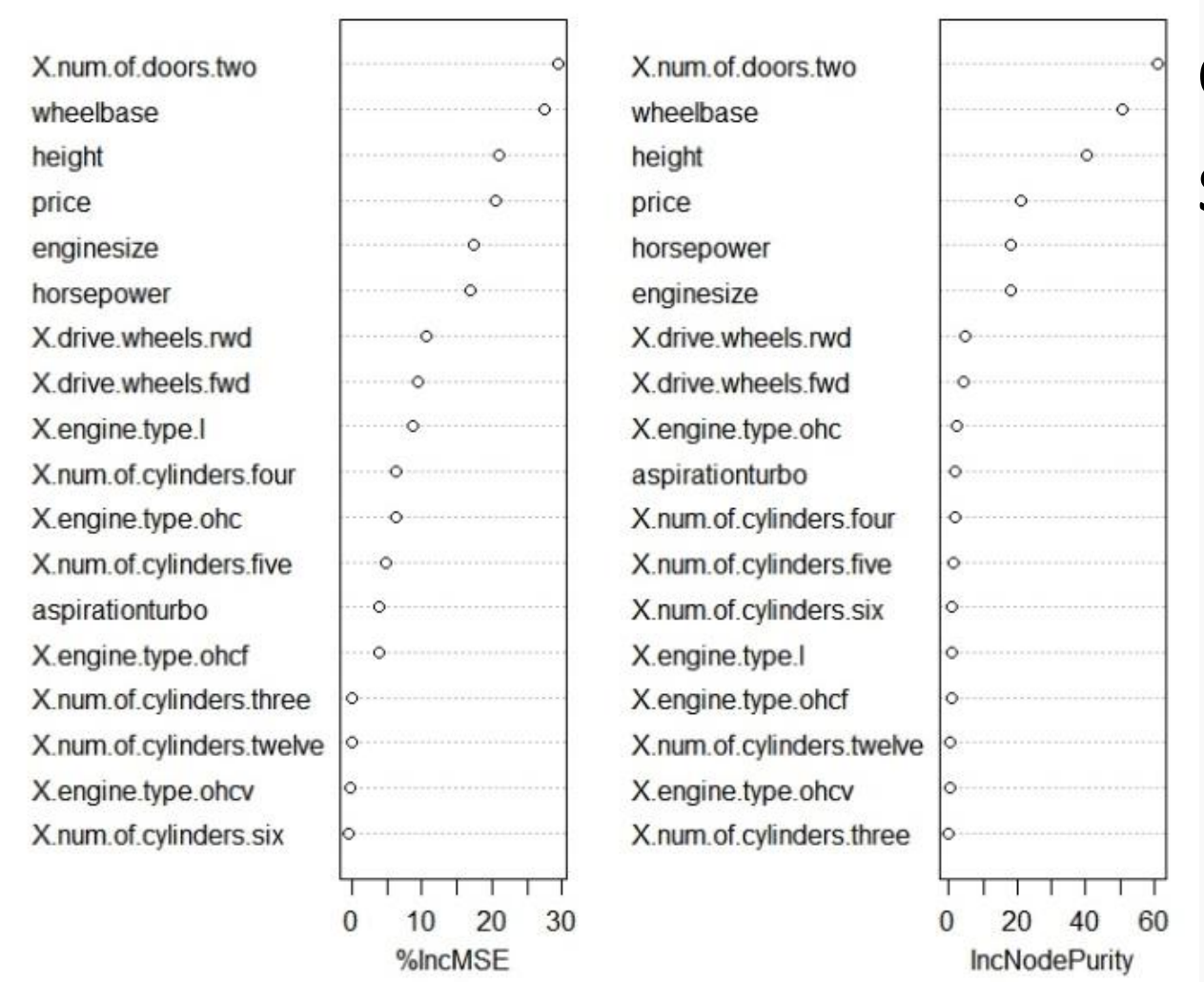
↓

Find the influencing factors of Symboling

Parameter	P or N	P-value
wheelbase	↑	0.006
height	↑	0.028
Engine size	↓	0.003
…		

What affects the Risk?

Random Forest



To determine the **importance rank** of these significant factors

What affects the Risk?

Findings

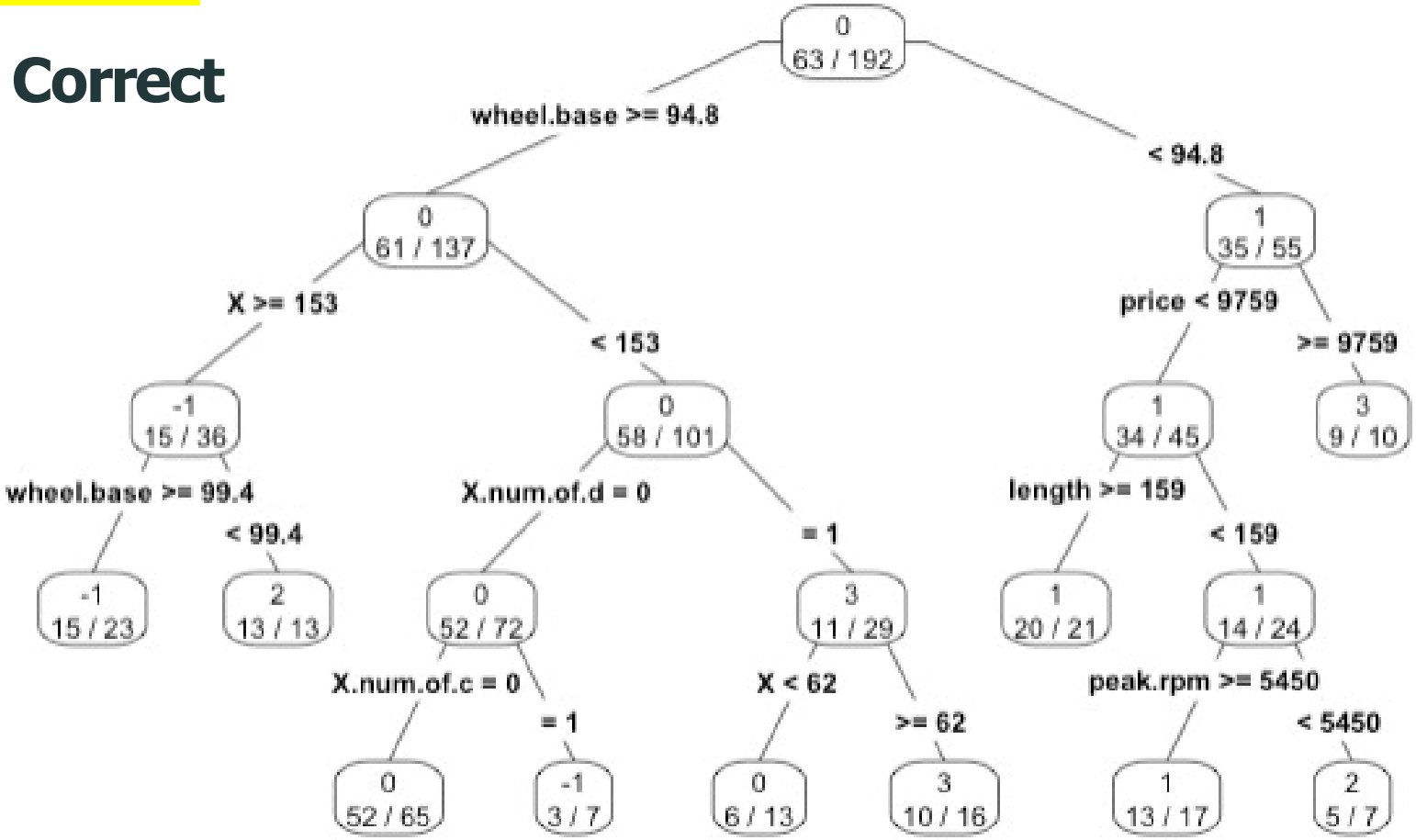
a) Numbers of doors
b) Wheel base
c) Height
d) Price
e) Engine size
f) Horse power
g) drive-wheels
h) Engine type
i) Numbers of cylinders
j) Aspiration

More Important to safety

What affects the Risk?

All variables

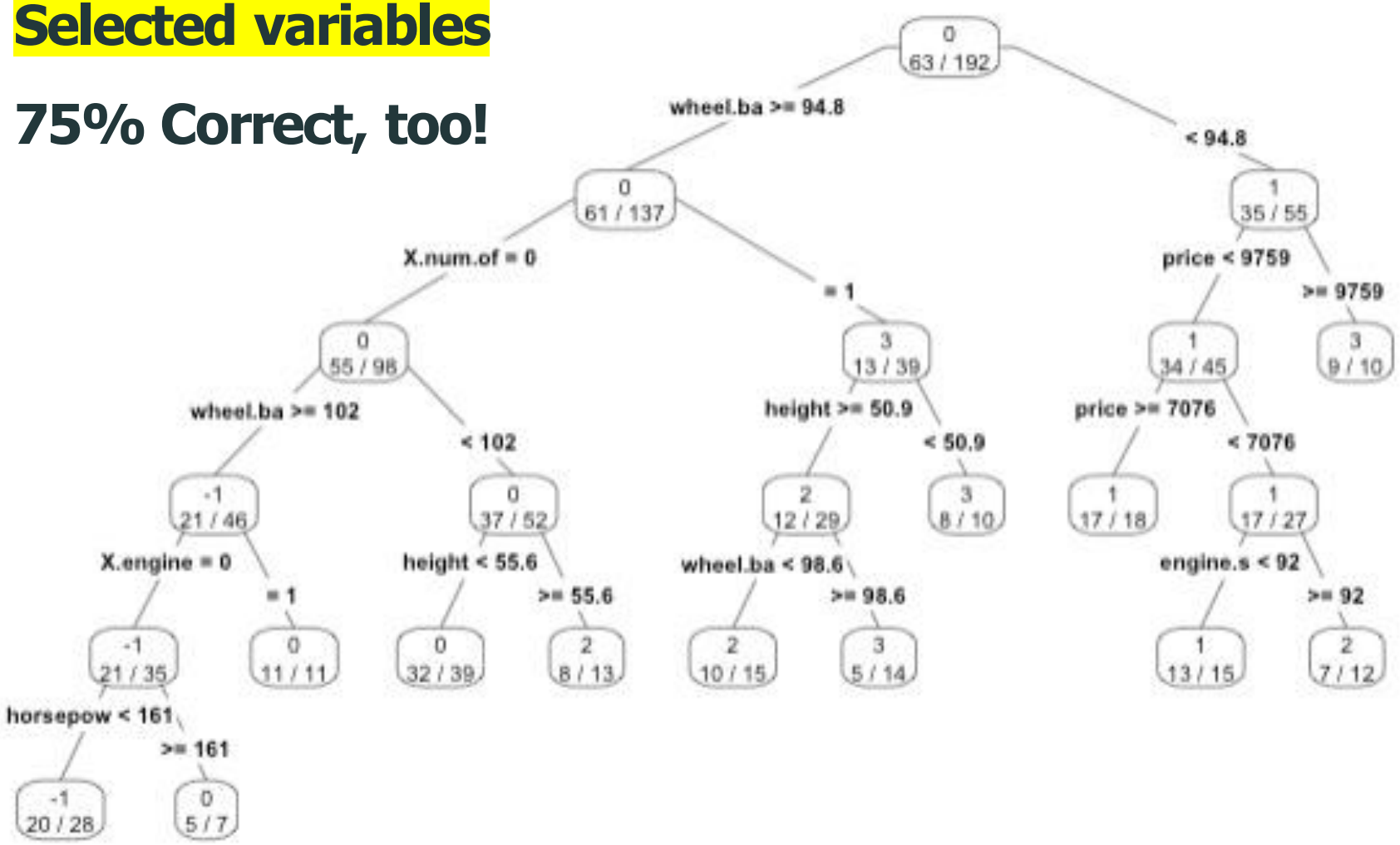
75% Correct



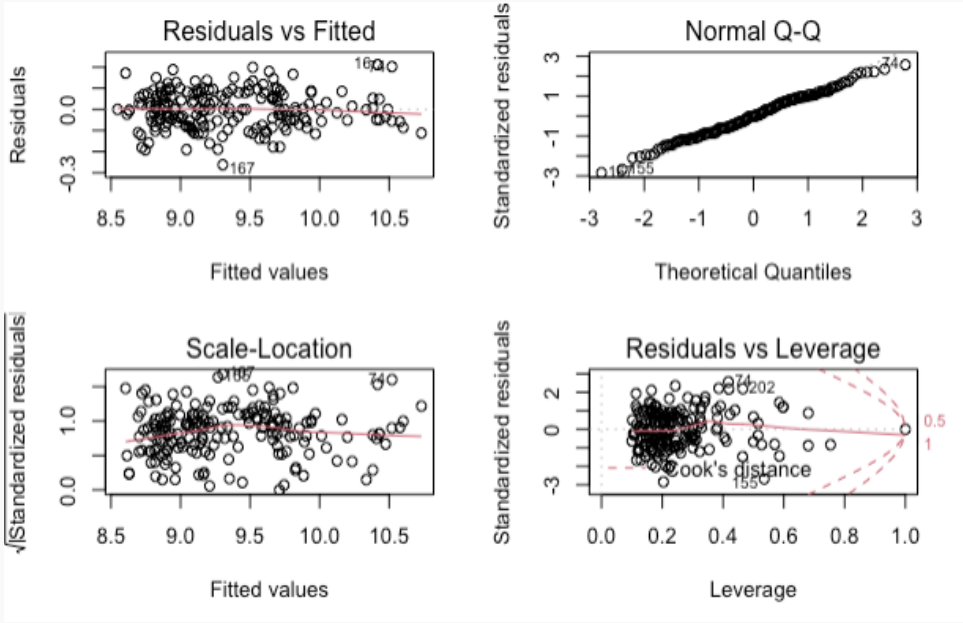
What affects the Risk?

Selected variables

75% Correct, too!



What affects the Price?



- ***OLR*** with Log transform
- Adjusted-R^2 = **0.97**

- Wheel base
- Height
- Curb weight
- Bore
- City mpg

- Make
- Aspiration
- Doors number
- Body style
- Engine type
- Fuel system

What affects the Price?

Lasso

- Feature selection
- **Continuous** vs **All**

- **Continuous**

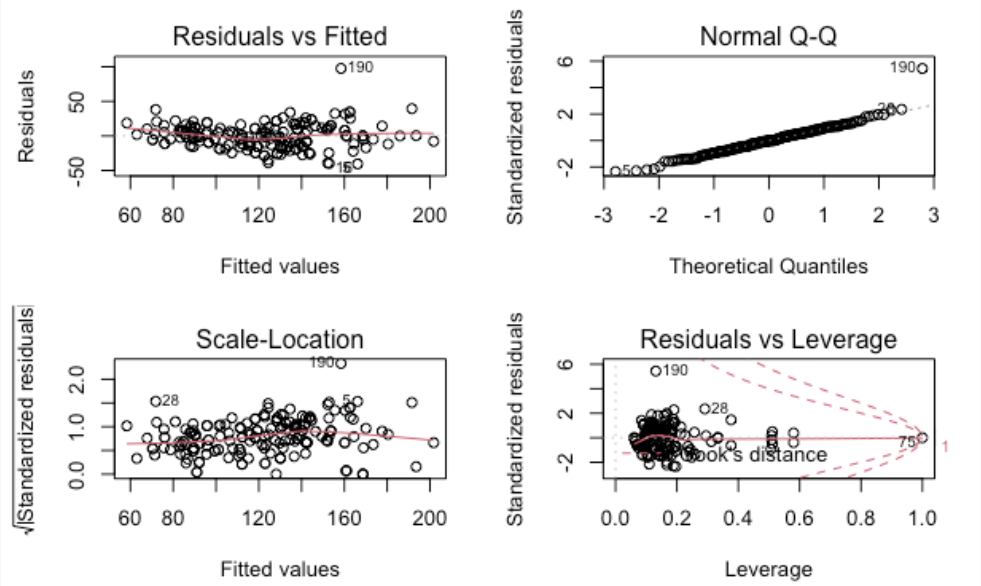
width	height	curb-weight	engine-size	stroke	horsepower
483.113709	35.204942	1.509583	99.324868	-376.753235	36.589091

- **All**

makeporsche	makemercedes.benz	makebmw
3484.600	4404.856	5363.846

What affects the Losses?

- ***OLR:*** Adjusted $R^2 = 0.68$
- ***Regression tree***




```
1) root 192 12563140000 13283.930
  2) engine.size< 182 175  3800061000 11228.570
    4) curb.weight< 2544 109   505551500  8276.000
      8) horsepower< 83 64     75515350  7076.391 *
      9) horsepower>=83 45    206949400  9982.111 *
    5) curb.weight>=2544 66   774968800 16104.790
     10) width< 68.6 58       562830300 15516.220 *
     11) width>=68.6 8         46382600 20371.880 *
  3) engine.size>=182 17      413464300 34442.060 *
```

Conclusions

- Risk(losses) is correlated with many properties of vehicles, such as wheelbase, auto size, number of doors and engine size. But these properties will not explain losses explicitly.

- Prices are mainly determined by wheel bases, heights and brands.

- Symboling depends but not only on properties of vehicles. These properties are usually hidden, such as engine and wheel, rather than appearance of autos. Symboling is also correlated with losses.