Automobile Data Set

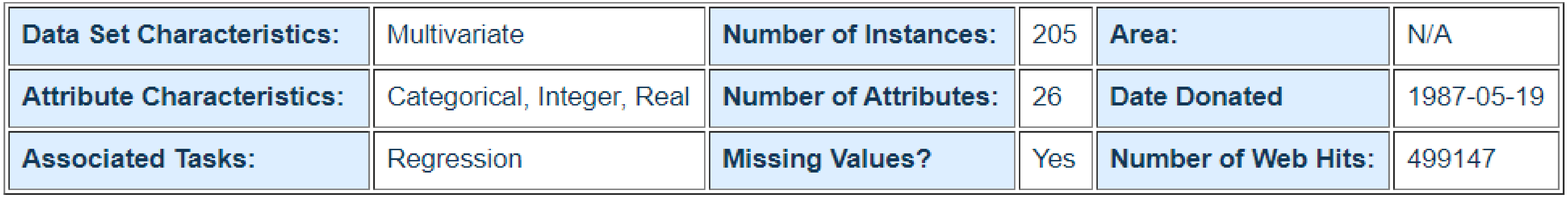
Download: Data Folder, Data Set Description

Abstract: From 1985 Ward's Automotive Yearbook

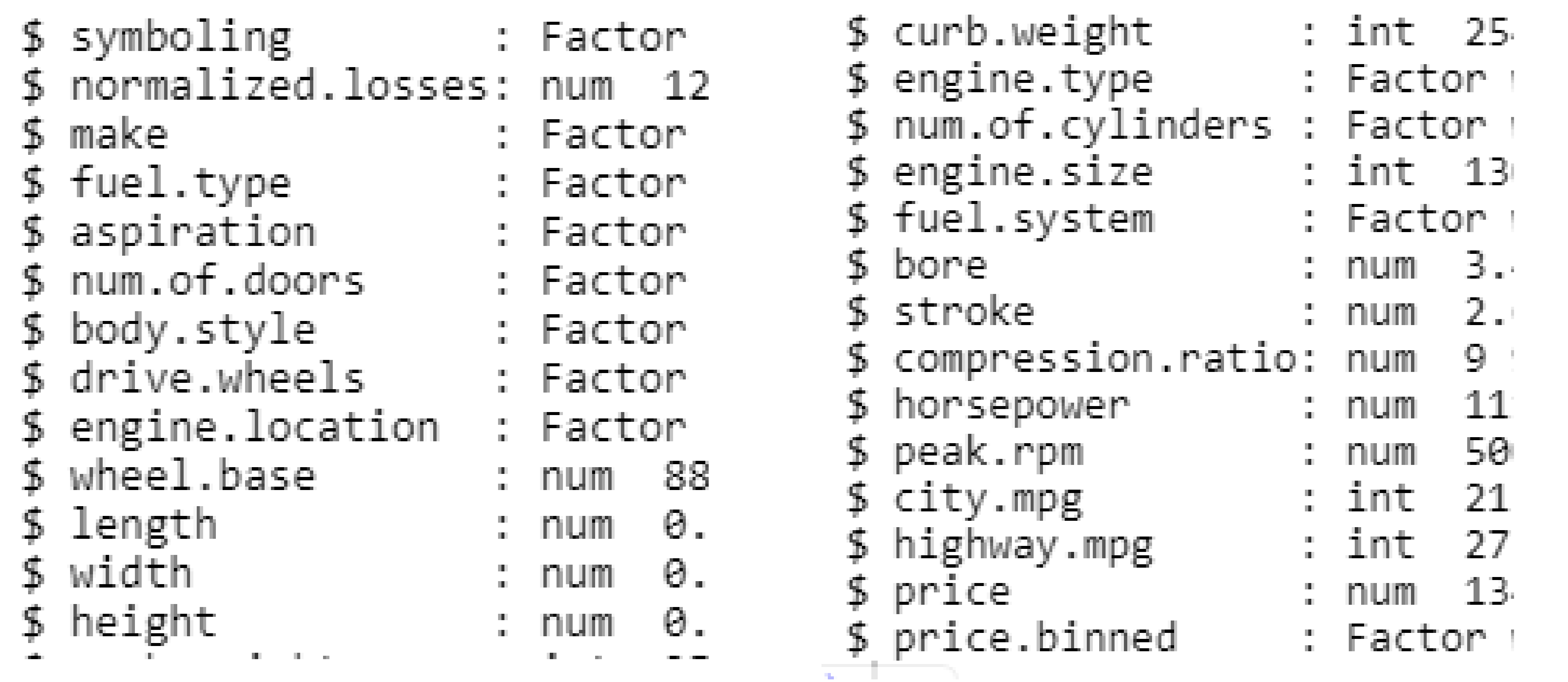
R project

PREPARED BY:

NATALYA GAPONENKO

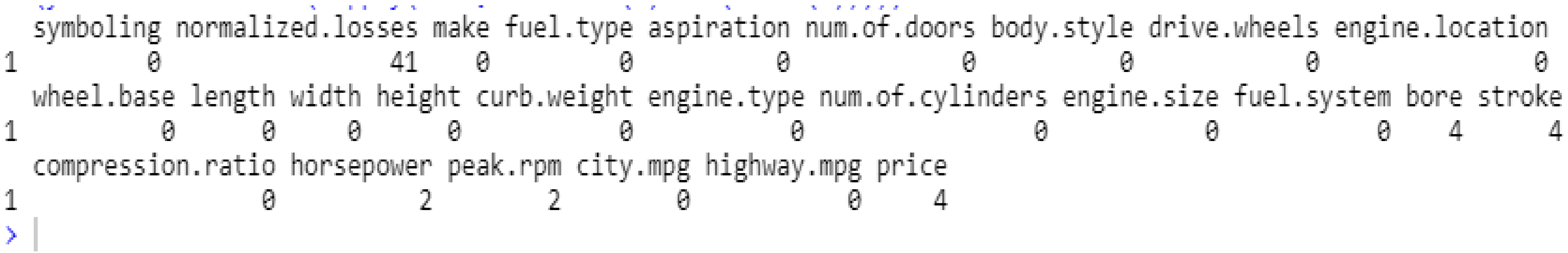


Data Set Characteristics:	Multivariate	**Number of Instances:**	205	**Area:**	N/A
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	26	**Date Donated**	1987-05-19
Associated Tasks:	Regression	**Missing Values?**	Yes	**Number of Web Hits:**	499147

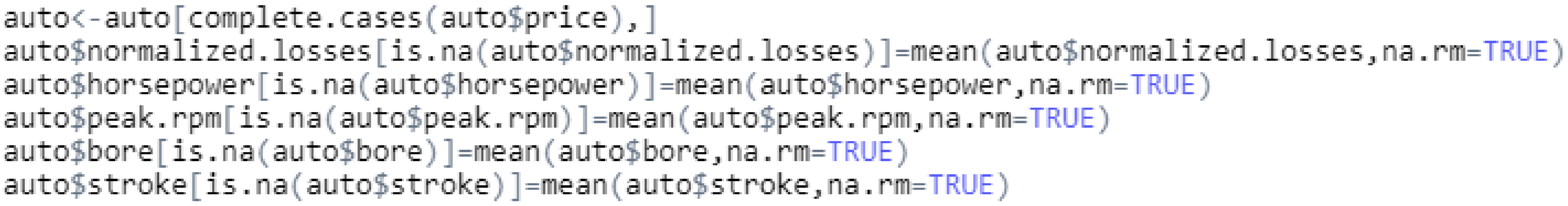


```
$ symboling        : Factor          $ curb.weight      : int  25
$ normalized.losses: num  12         $ engine.type      : Factor
$ make             : Factor          $ num.of.cylinders : Factor
$ fuel.type        : Factor          $ engine.size      : int  13
$ aspiration       : Factor          $ fuel.system      : Factor
$ num.of.doors     : Factor          $ bore             : num  3.
$ body.style       : Factor          $ stroke           : num  2.
$ drive.wheels     : Factor          $ compression.ratio: num  9
$ engine.location  : Factor          $ horsepower       : num  11
$ wheel.base       : num  88         $ peak.rpm         : num  50
$ length           : num  0.         $ city.mpg         : int  21
$ width            : num  0.         $ highway.mpg      : int  27
$ height           : num  0.         $ price            : num  13
                                     $ price.binned     : Factor
```

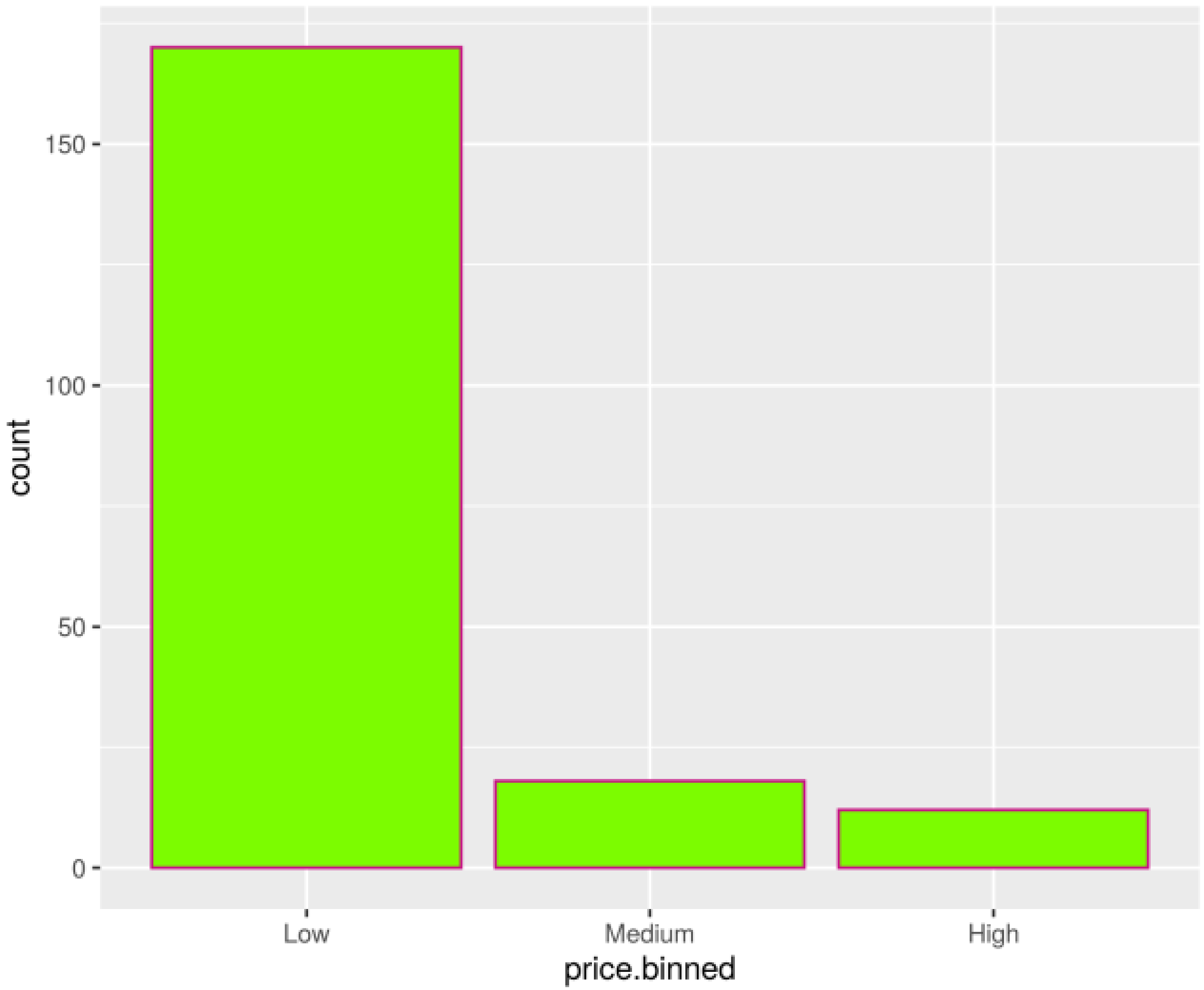
Missing values



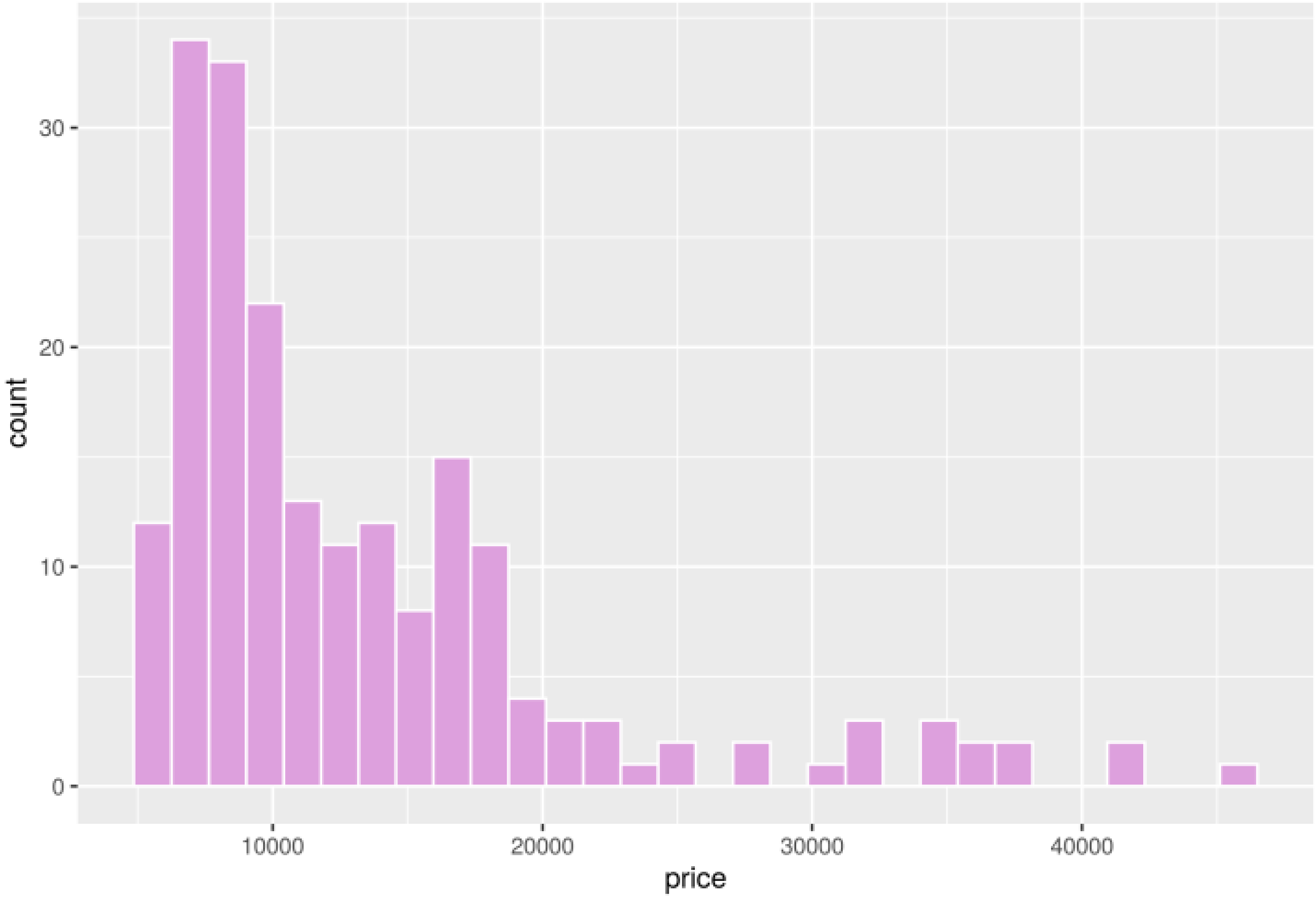
```
  symboling normalized.losses make fuel.type aspiration num.of.doors body.style drive.wheels engine.location
1         0                41    0         0          0            0          0            0               0
  wheel.base length width height curb.weight engine.type num.of.cylinders engine.size fuel.system bore stroke
1          0      0     0      0           0           0                0           0           0    4      4
  compression.ratio horsepower peak.rpm city.mpg highway.mpg price
1                 0          2        2        0           0     4
> 
```



```
auto<-auto[complete.cases(auto$price),]
auto$normalized.losses[is.na(auto$normalized.losses)]=mean(auto$normalized.losses,na.rm=TRUE)
auto$horsepower[is.na(auto$horsepower)]=mean(auto$horsepower,na.rm=TRUE)
auto$peak.rpm[is.na(auto$peak.rpm)]=mean(auto$peak.rpm,na.rm=TRUE)
auto$bore[is.na(auto$bore)]=mean(auto$bore,na.rm=TRUE)
auto$stroke[is.na(auto$stroke)]=mean(auto$stroke,na.rm=TRUE)
```

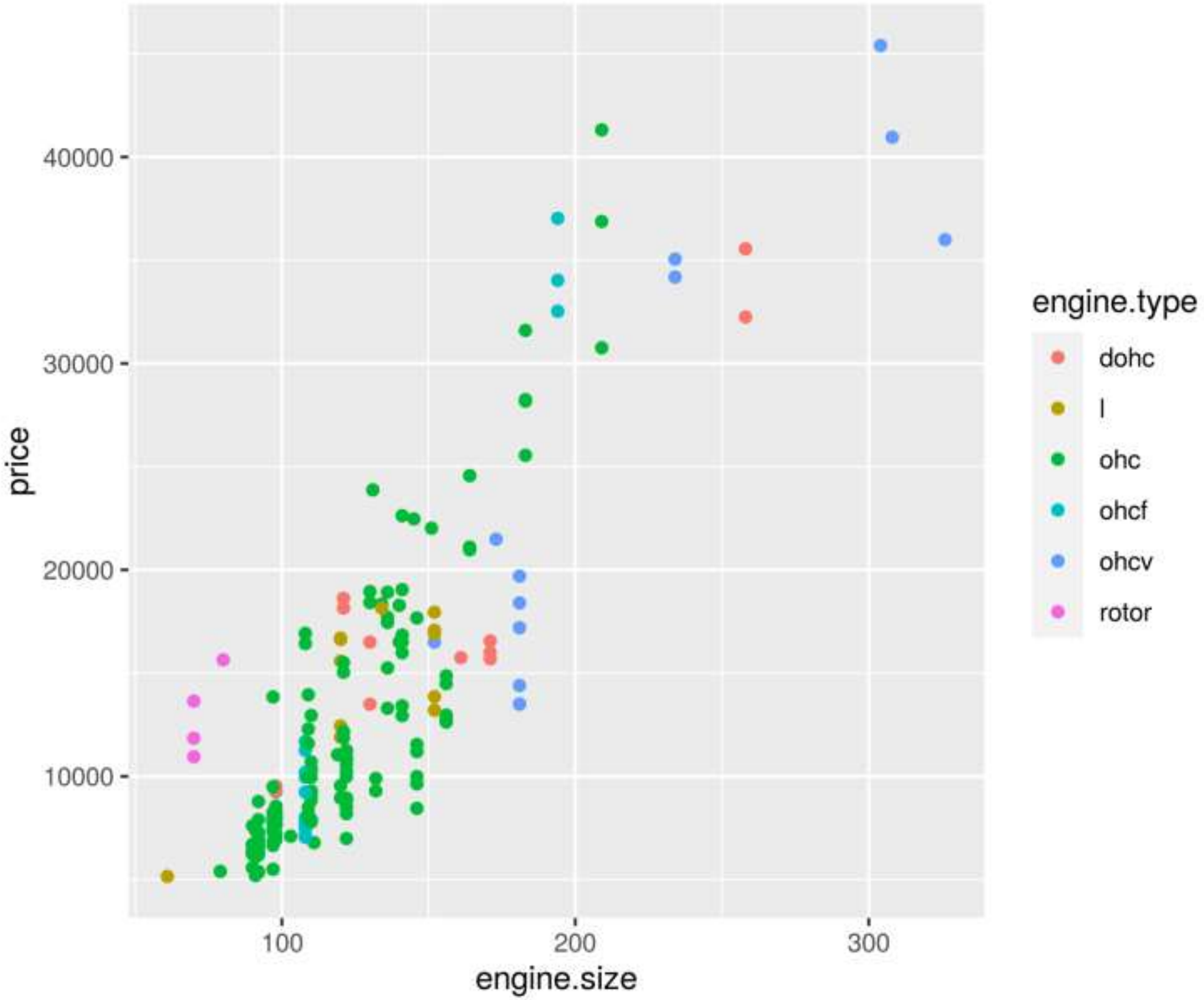


Distribution of price

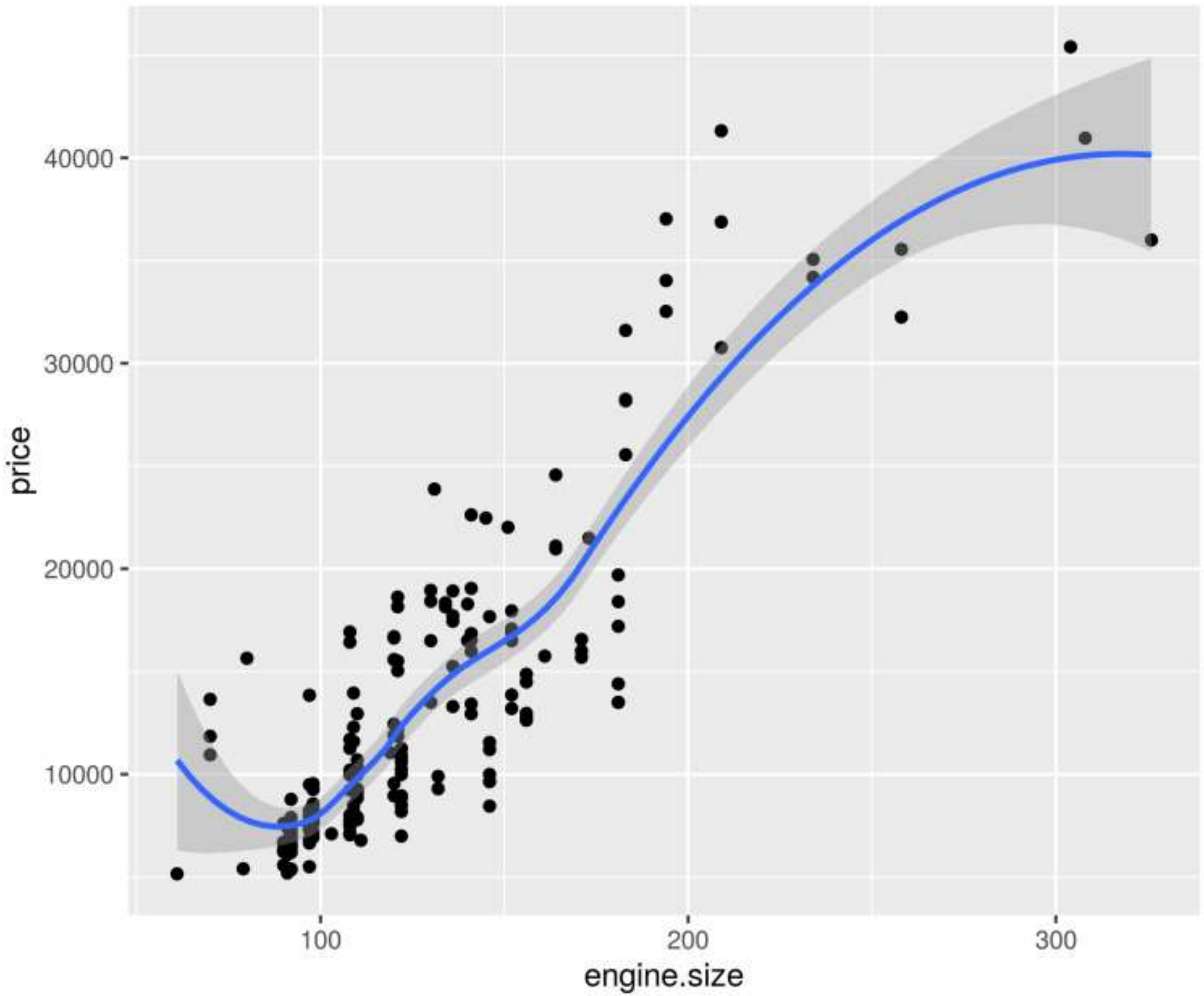


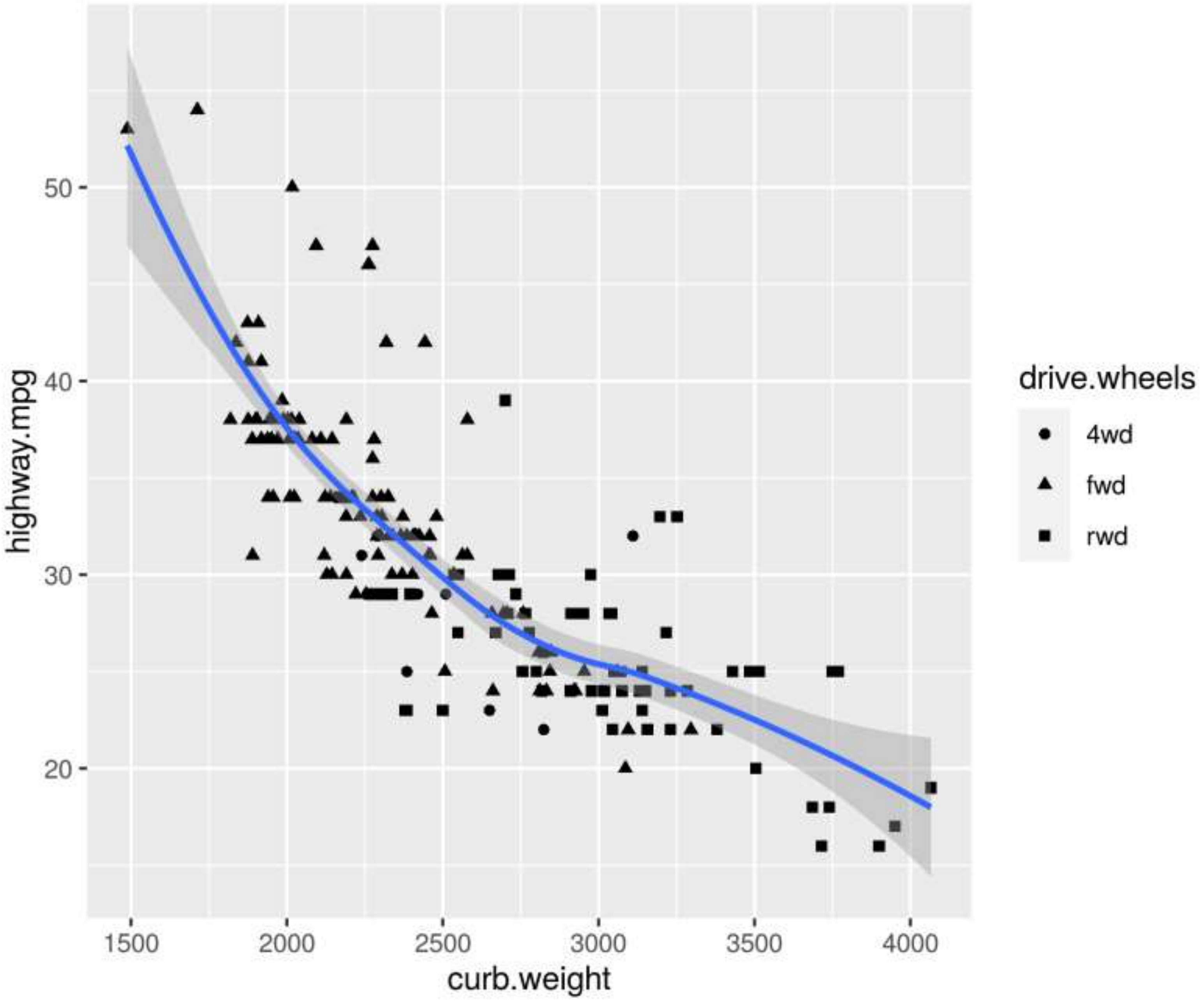
Distribution of price

Engine.size vs price

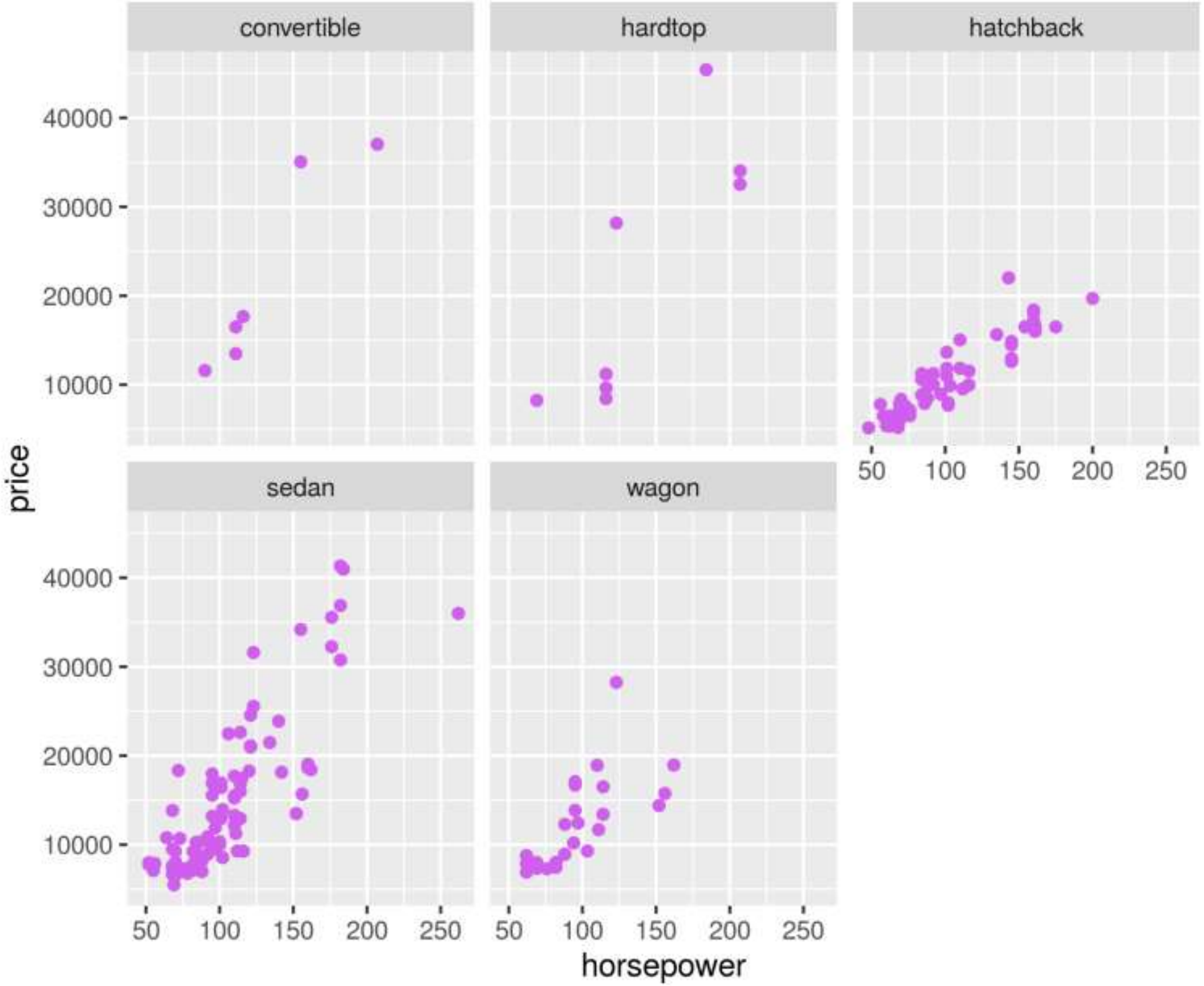


Engine.size vs price



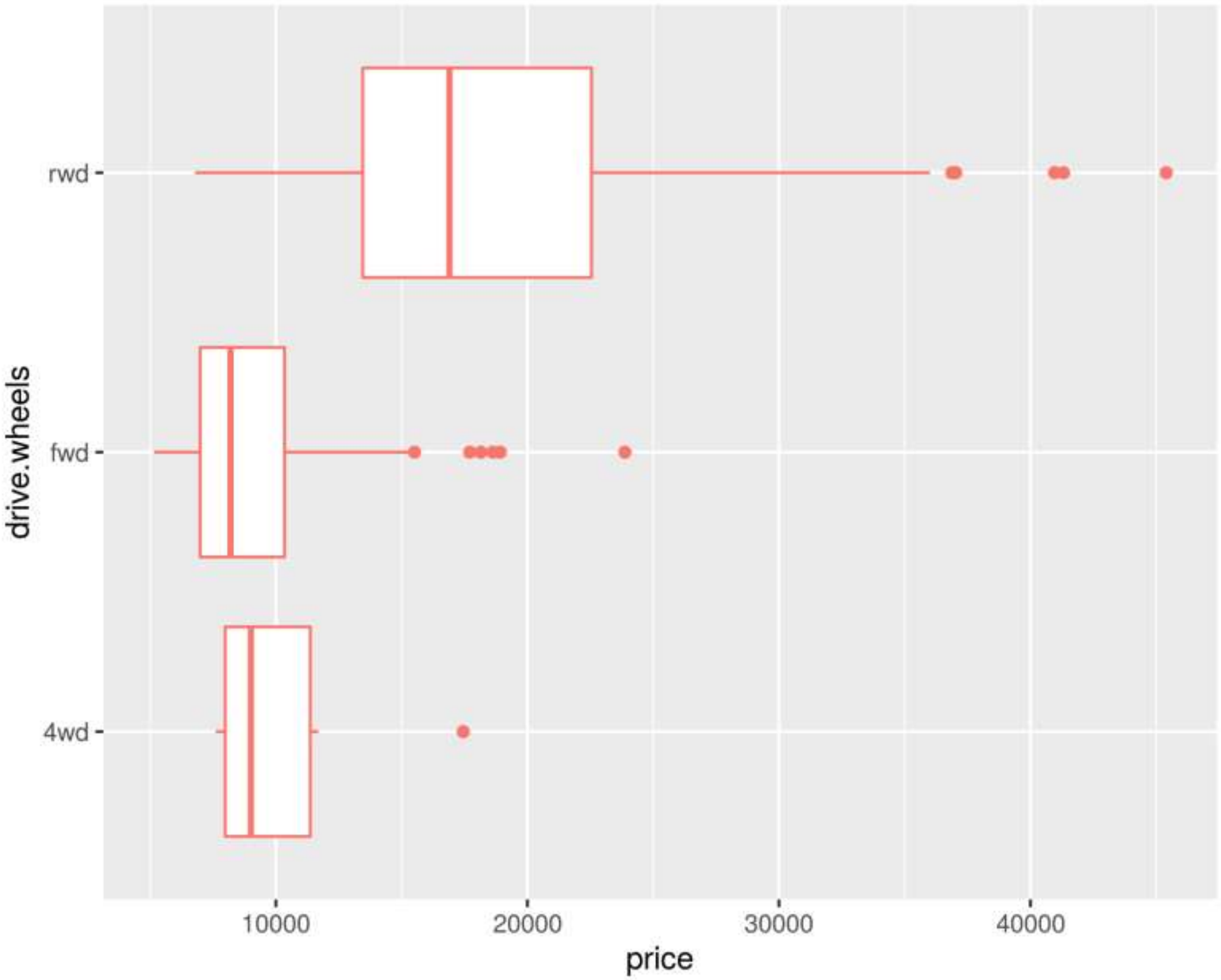


Curb.weight vs highway.mpg

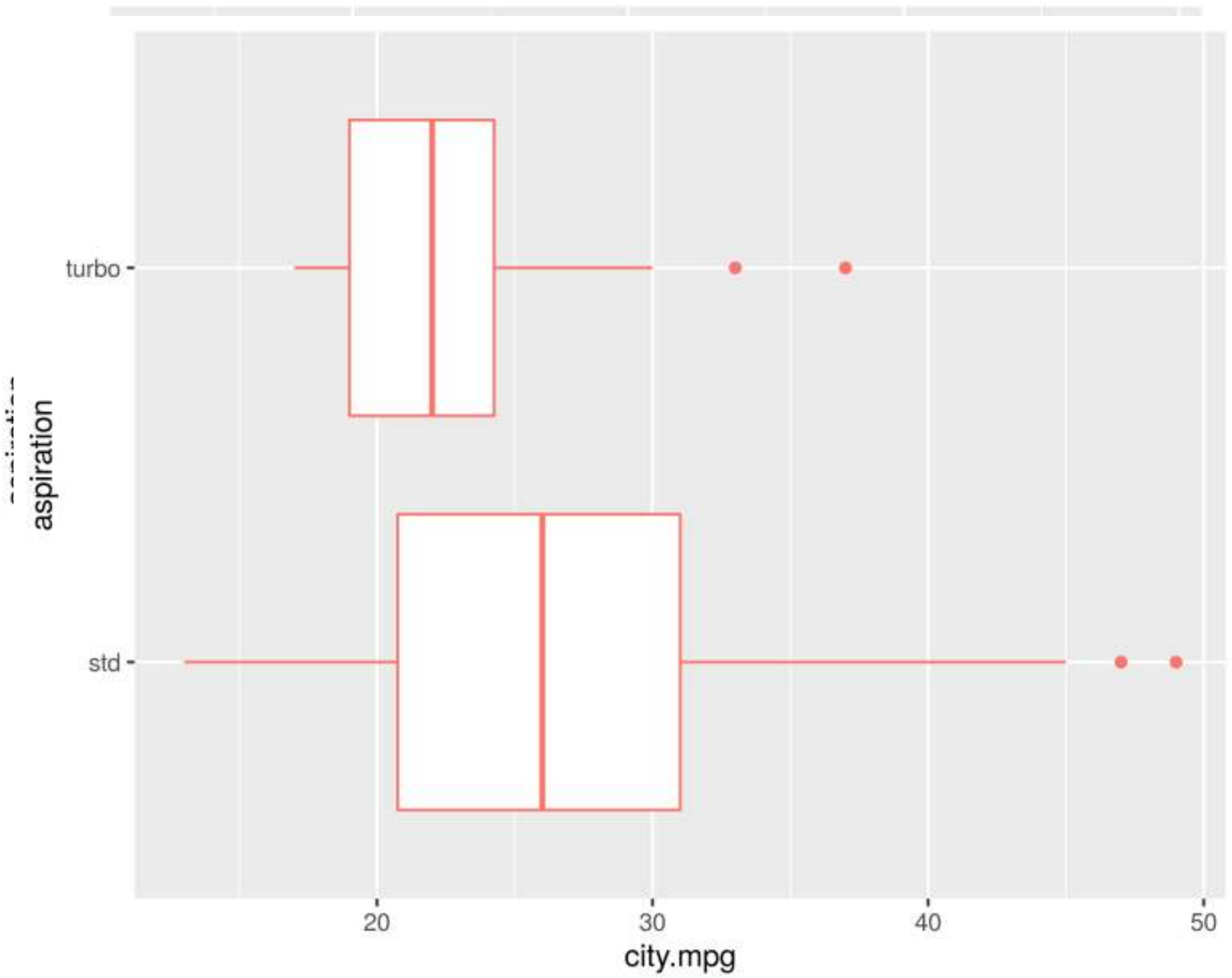


Horsepower and price

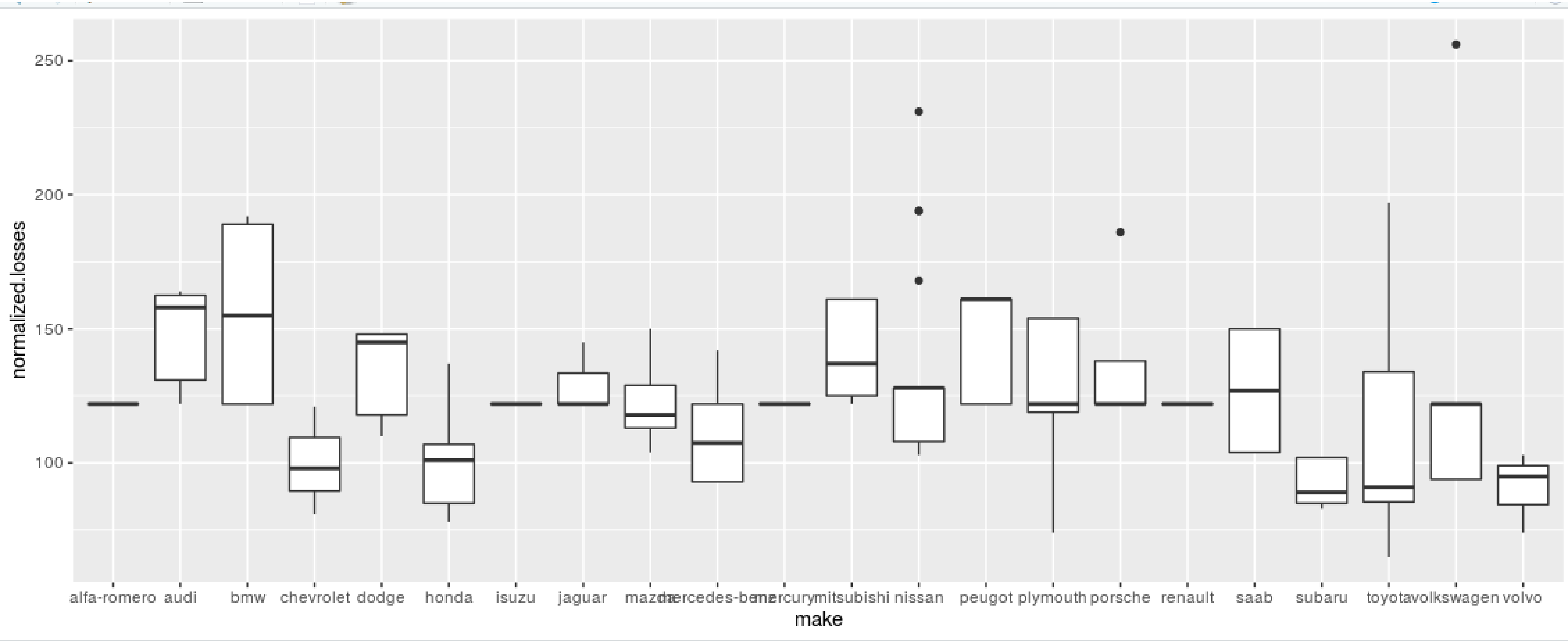
Boxplot of price based on drive.wheels



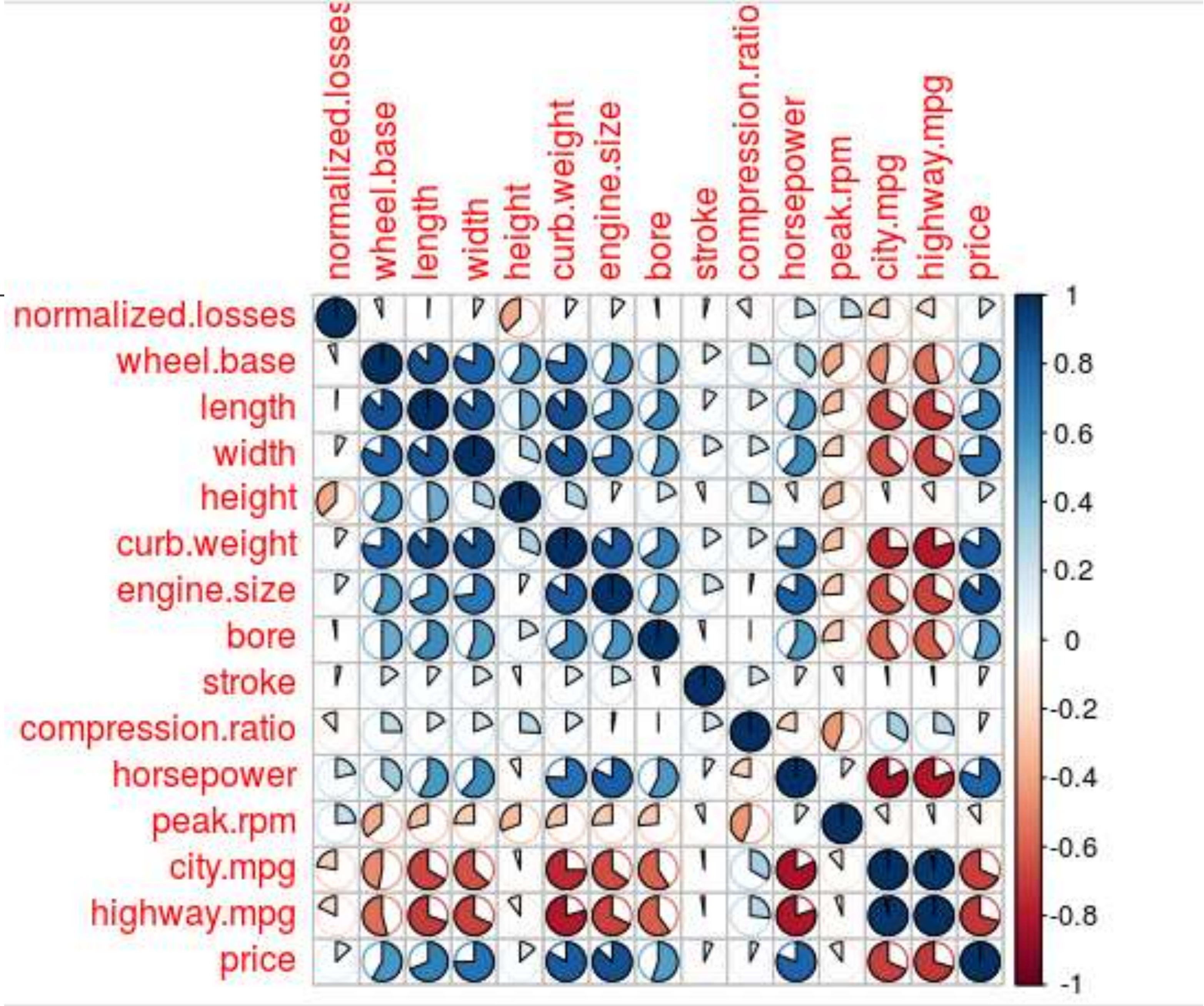
Boxplot of city.mpg based on aspiration



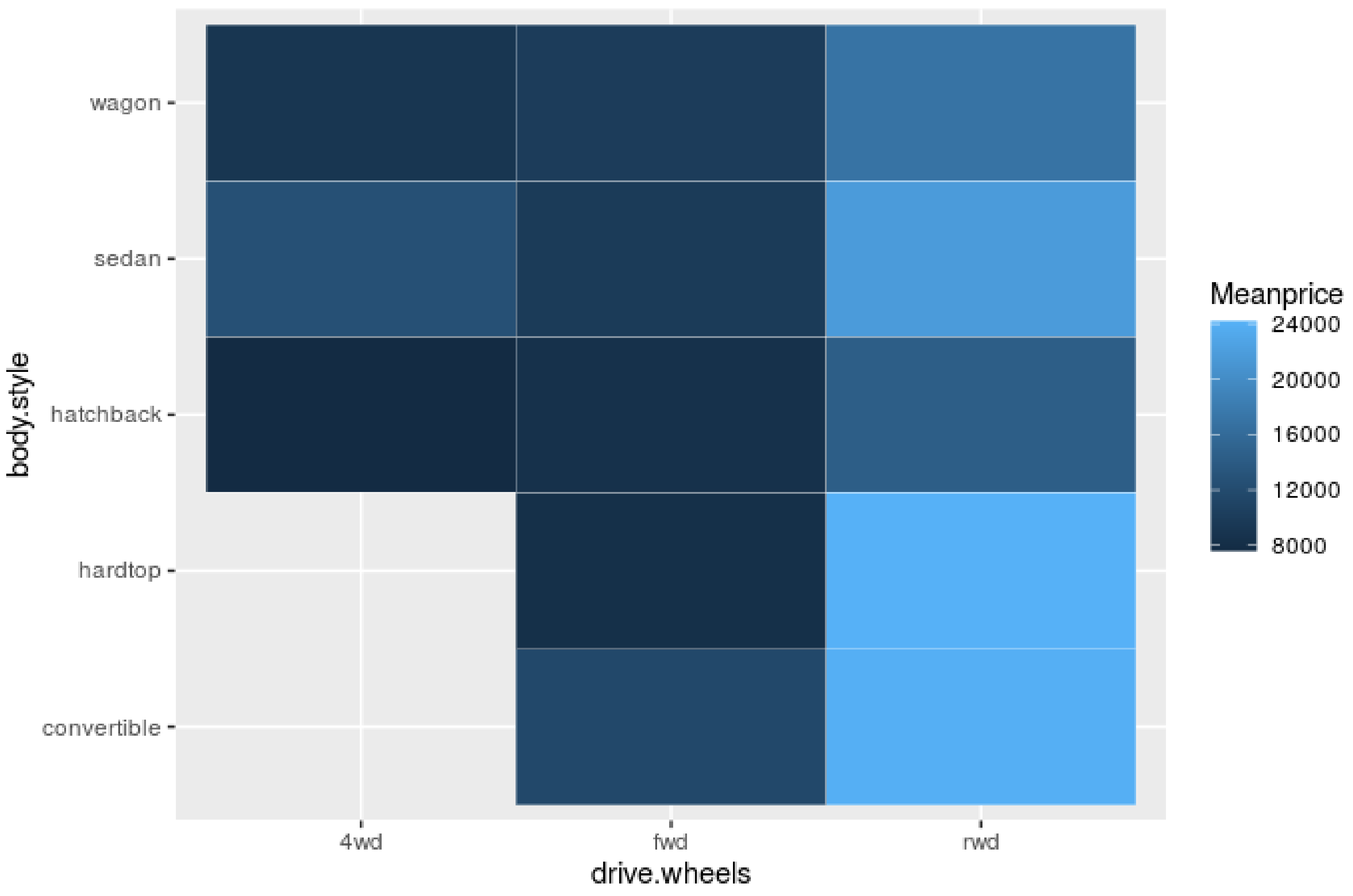
Boxplot between normalized losses and manufacturing company



Correlation between variables



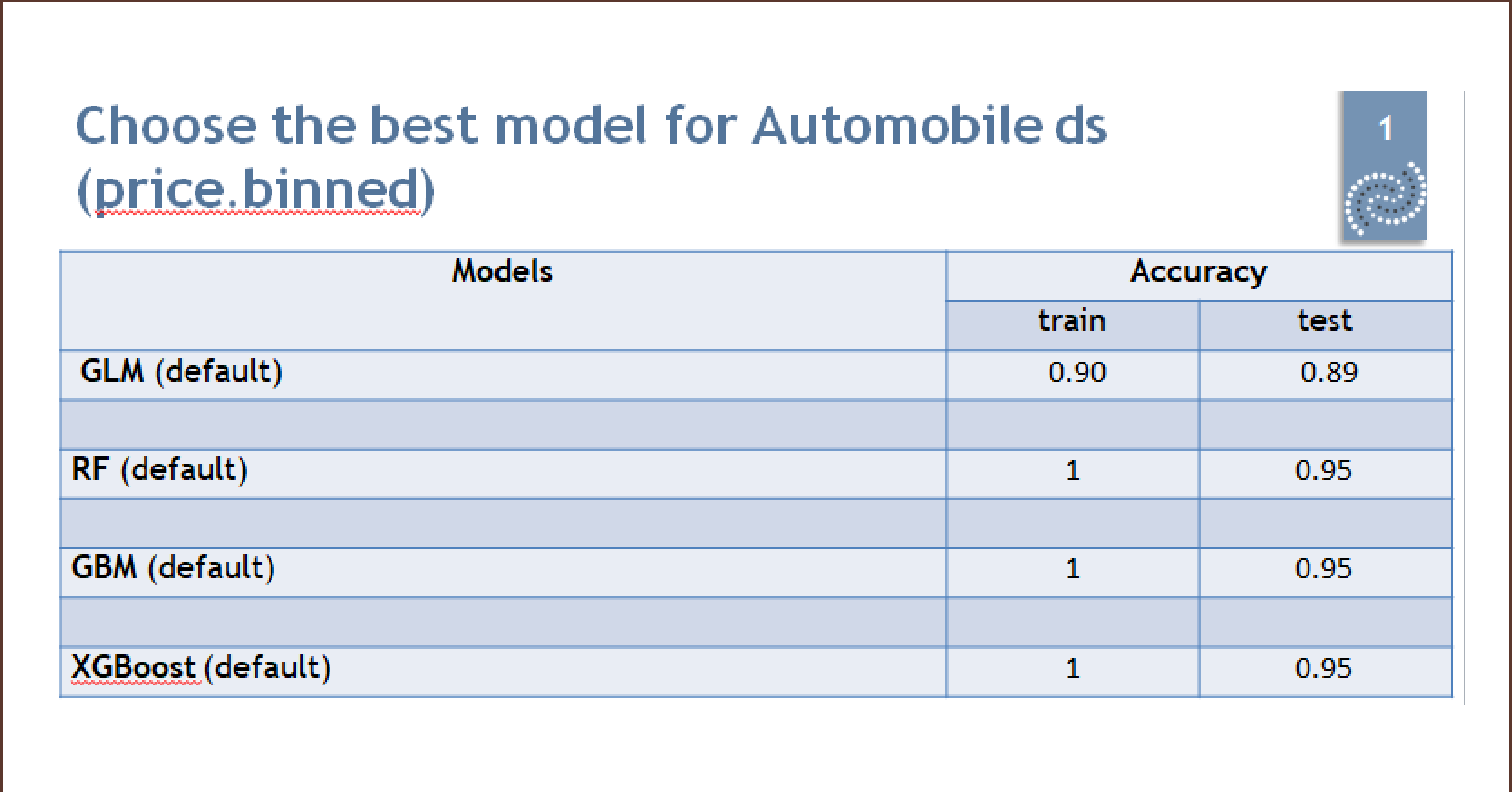
Heatmap



wagon
sedan
hatchback
hardtop
convertible
body.style
4wd
fwd
rwd
drive.wheels
Meanprice
24000
20000
16000
12000
8000

Choose the best GLM model for Automobile ds

Models	R^2	
	train	test
Without regularization		
• for all predictors	0,97	0,94
• for predictors with p value <= 0.05	0,97	0,93
With regularization	0,83	0,89
RF (default)	0.92	0.94
GBM (default)	0.98	0.96
XGBoost (default)	0.999	0.92
Stacked Ensemble		
• RF, GBM, XGBoost	0.999	0.92



Choose the best model for Automobile ds (price.binned)

Models	Accuracy	
	train	test
GLM (default)	0.90	0.89
RF (default)	1	0.95
GBM (default)	1	0.95
XGBoost (default)	1	0.95