Data Analysis and Basic Plotting

How to do basic data analysis?

1. Download data
2. Load data from file to R
3. plot using different attributes

Download data

335 Data Sets

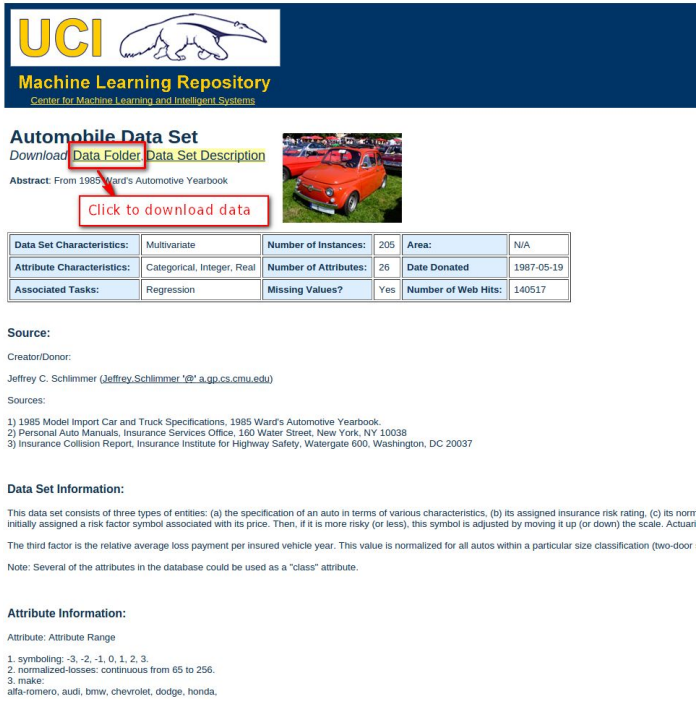
Name	Data Types	D
Abalone	Multivariate	C
Adult	Multivariate	C
Annealing	Multivariate	C
Anonymous Microsoft Web Data		R
Arrhythmia	Multivariate	C
Artificial Characters	Multivariate	C
Audiology (Original)	Multivariate	C
Audiology (Standardized)	Multivariate	C
Auto MPG	Multivariate	R
Automobile	Multivariate	R
Badges	Univariate, Text	C

use uci archive as an example.

https://archive.ics.uci.edu/ml/datasets.html

335 data sets are listed, select one of the data sets for analysis.

Download data

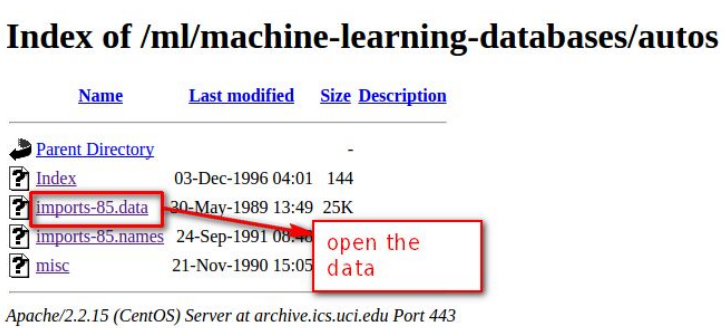


Here we choose ‘Automobile’ data set as an example.

You can find information about the data set, like source, data set description, attribute information and relevant papers.

click data folder to download the data set.

Download data

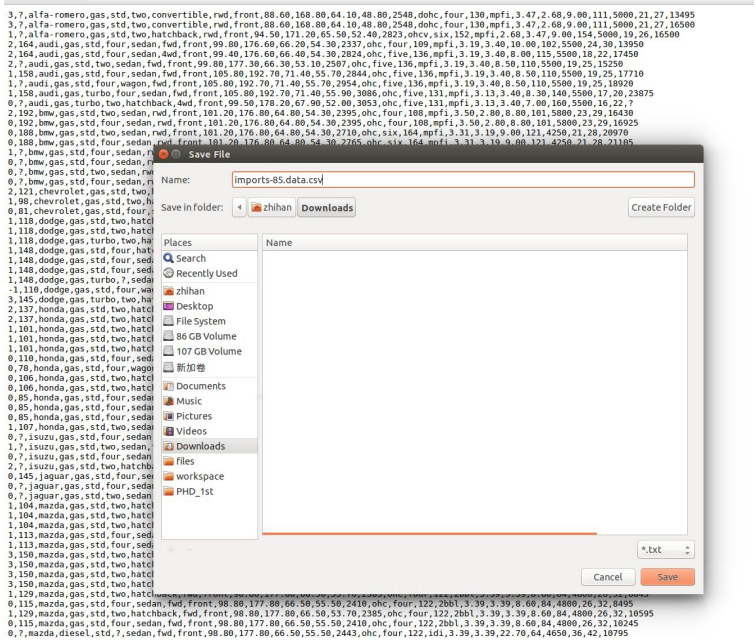


When you click ‘data folder’, both data and data description are in the folder.

data file is always the has .data extension name and with biggest file size.

open the data file. Some browser will start to download the data file directly, others will open it in the browser.

Download data



If the data file is opened in the browser, right click on the browser and click ‘save as’.

Change the file extension name from default ‘txt’ to ‘csv’ in order to be processed easily.

If attribute header information is not contained. They are in the information description on the previous data set information page.

Add attribute header

Attribute Information:

Attribute: Attribute Range

1. symboling: -3, -2, -1, 0, 1, 2, 3.
2. normalized-losses: continuous from 65 to 256.
3. make:
alfa-romero, audi, bmw, chevrolet, dodge, honda,
isuzu, jaguar, mazda, mercedes-benz, mercury,
mitsubishi, nissan, peugot, plymouth, porsche,
renault, saab, subaru, toyota, volkswagen, volvo

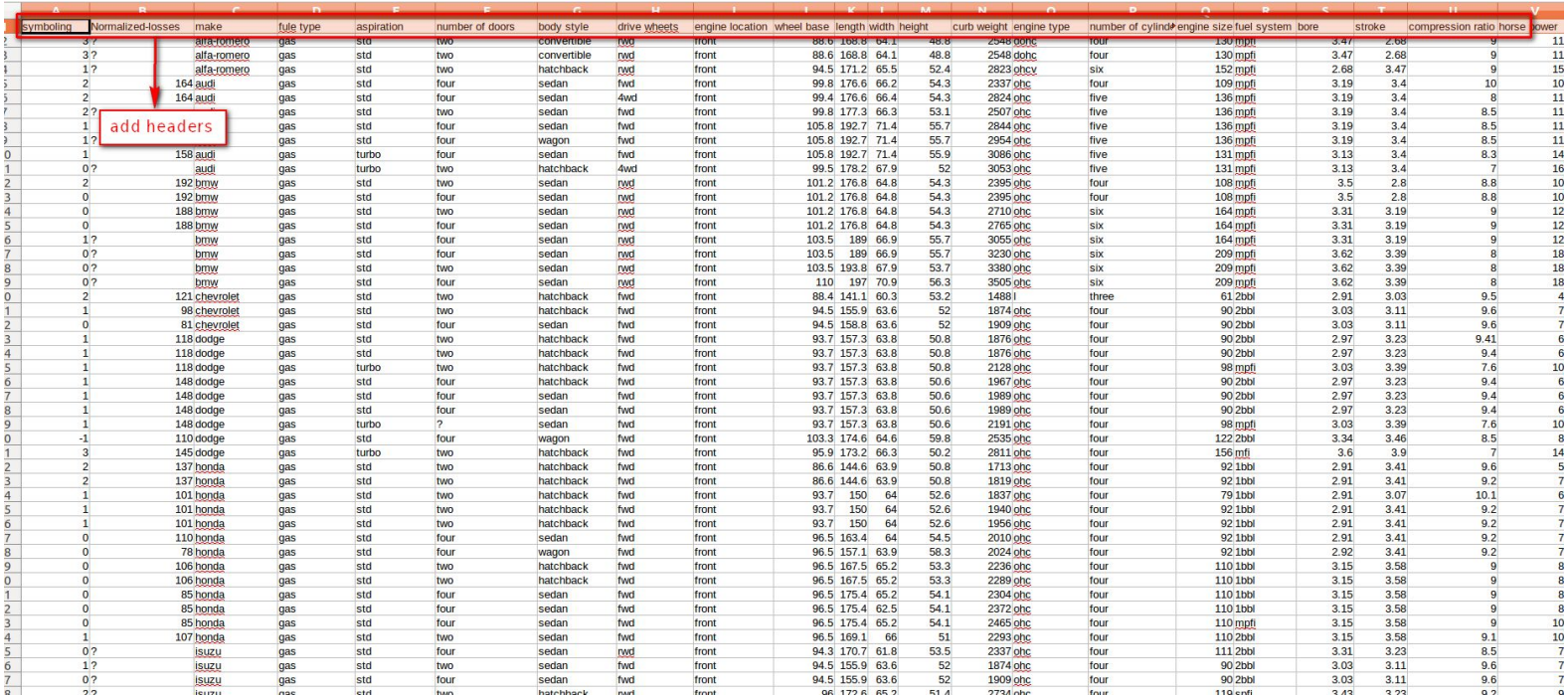
4. fuel-type: diesel, gas.
5. aspiration: std, turbo.
6. num-of-doors: four, two.
7. body-style: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. wheel-base: continuous from 86.6 120.9.
11. length: continuous from 141.1 to 208.1.
12. width: continuous from 60.3 to 72.3.
13. height: continuous from 47.8 to 59.8.
14. curb-weight: continuous from 1488 to 4066.
15. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. num-of-cylinders: eight, five, four, six, three, twelve, two.
17. engine-size: continuous from 61 to 326.
18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore: continuous from 2.54 to 3.94.
20. stroke: continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. horsepower: continuous from 48 to 288.
23. peak-rpm: continuous from 4150 to 6600.
24. city-mpg: continuous from 13 to 49.
25. highway-mpg: continuous from 16 to 54.
26. price: continuous from 5118 to 45400.

Find the attribute information on the data set description page.

Open the data file in spreadsheet, insert a new row in the first row and add header information.

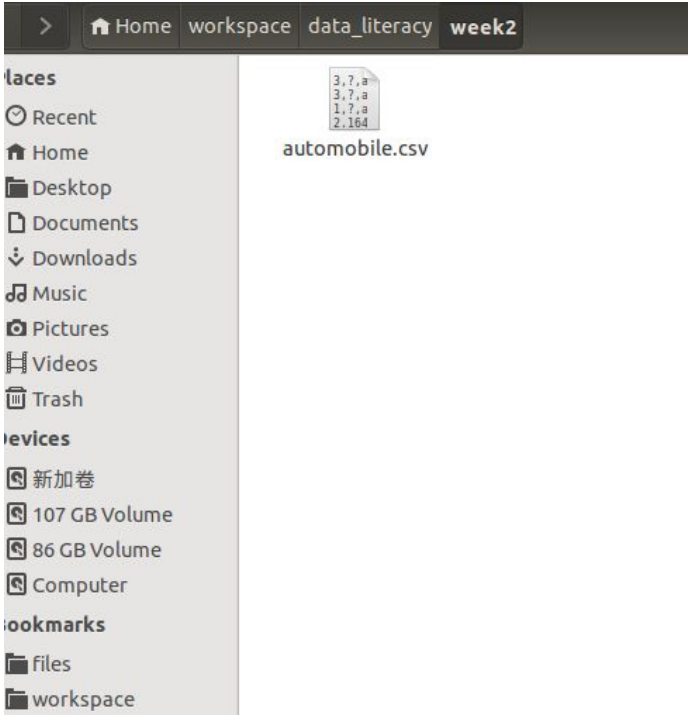
Delete all ? in the csv file, the R will treat any blank value as NA.

Add attribute header



symboling	Normalized-losses	make	fule type	aspiration	number of doors	body style	drive wheels	engine location	wheel base	length	width	height	curb weight	engine type	number of cylinders	engine size	fuel system	bore	stroke	compression ratio	horse	power
3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	9	11	
3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	9	11	
1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	2823	ohcv	six	152	mpfi	2.68	3.47	9	15	
2	164	audi	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337	ohc	four	109	mpfi	3.19	3.4	10	10	
2	164	audi	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824	ohc	five	136	mpfi	3.19	3.4	8	11	
2	?		gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507	ohc	five	136	mpfi	3.19	3.4	8.5	11	
1	?		gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844	ohc	five	136	mpfi	3.19	3.4	8.5	11	
1	?		gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954	ohc	five	136	mpfi	3.19	3.4	8.5	11	
1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086	ohc	five	131	mpfi	3.13	3.4	8.3	14	
0	?	audi	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52	3053	ohc	five	131	mpfi	3.13	3.4	7	16	
2	192	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.5	2.8	8.8	10	
0	192	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.5	2.8	8.8	10	
0	188	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2710	ohc	six	164	mpfi	3.31	3.19	9	12	
0	188	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2765	ohc	six	164	mpfi	3.31	3.19	9	12	
1	?	bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3055	ohc	six	164	mpfi	3.31	3.19	9	12	
0	?	bmw	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3230	ohc	six	209	mpfi	3.62	3.39	8	18	
0	?	bmw	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.7	3380	ohc	six	209	mpfi	3.62	3.39	8	18	
0	?	bmw	gas	std	four	sedan	rwd	front	110	197	70.9	56.3	3505	ohc	six	209	mpfi	3.62	3.39	8	18	
2	121	chevrolet	gas	std	two	hatchback	fwd	front	88.4	141.1	60.3	53.2	1488	l	three	61	2bbl	2.91	3.03	9.5	4	
1	98	chevrolet	gas	std	two	hatchback	fwd	front	94.5	155.9	63.6	52	1874	ohc	four	90	2bbl	3.03	3.11	9.6	7	
0	81	chevrolet	gas	std	four	sedan	fwd	front	94.5	158.8	63.6	52	1909	ohc	four	90	2bbl	3.03	3.11	9.6	7	
1	118	dodge	gas	std	two	hatchback	fwd	front	93.7	157.3	63.8	50.8	1876	ohc	four	90	2bbl	2.97	3.23	9.41	6	
1	118	dodge	gas	std	two	hatchback	fwd	front	93.7	157.3	63.8	50.8	1876	ohc	four	90	2bbl	2.97	3.23	9.4	6	
1	118	dodge	gas	turbo	two	hatchback	fwd	front	93.7	157.3	63.8	50.8	2128	ohc	four	98	mpfi	3.03	3.39	7.6	10	
1	148	dodge	gas	std	four	hatchback	fwd	front	93.7	157.3	63.8	50.6	1967	ohc	four	90	2bbl	2.97	3.23	9.4	6	
1	148	dodge	gas	std	four	sedan	fwd	front	93.7	157.3	63.8	50.6	1989	ohc	four	90	2bbl	2.97	3.23	9.4	6	
1	148	dodge	gas	std	four	sedan	fwd	front	93.7	157.3	63.8	50.6	1989	ohc	four	90	2bbl	2.97	3.23	9.4	6	
1	148	dodge	gas	turbo	?	sedan	fwd	front	93.7	157.3	63.8	50.6	2191	ohc	four	98	mpfi	3.03	3.39	7.6	10	
-1	110	dodge	gas	std	four	wagon	fwd	front	103.3	174.6	64.6	59.8	2535	ohc	four	122	2bbl	3.34	3.46	8.5	8	
3	145	dodge	gas	turbo	two	hatchback	fwd	front	95.9	173.2	66.3	50.2	2811	ohc	four	156	mfi	3.6	3.9	7	14	
2	137	honda	gas	std	two	hatchback	fwd	front	86.6	144.6	63.9	50.8	1713	ohc	four	92	1bbl	2.91	3.41	9.6	5	
2	137	honda	gas	std	two	hatchback	fwd	front	86.6	144.6	63.9	50.8	1819	ohc	four	92	1bbl	2.91	3.41	9.2	7	
1	101	honda	gas	std	two	hatchback	fwd	front	93.7	150	64	52.6	1837	ohc	four	79	1bbl	2.91	3.07	10.1	6	
1	101	honda	gas	std	two	hatchback	fwd	front	93.7	150	64	52.6	1940	ohc	four	92	1bbl	2.91	3.41	9.2	7	
1	101	honda	gas	std	two	hatchback	fwd	front	93.7	150	64	52.6	1956	ohc	four	92	1bbl	2.91	3.41	9.2	7	
0	110	honda	gas	std	four	sedan	fwd	front	96.5	163.4	64	54.5	2010	ohc	four	92	1bbl	2.91	3.41	9.2	7	
0	78	honda	gas	std	four	wagon	fwd	front	96.5	157.1	63.9	58.3	2024	ohc	four	92	1bbl	2.92	3.41	9.2	7	
0	106	honda	gas	std	two	hatchback	fwd	front	96.5	167.5	65.2	53.3	2236	ohc	four	110	1bbl	3.15	3.58	9	8	
0	106	honda	gas	std	two	hatchback	fwd	front	96.5	167.5	65.2	53.3	2289	ohc	four	110	1bbl	3.15	3.58	9	8	
0	85	honda	gas	std	four	sedan	fwd	front	96.5	175.4	65.2	54.1	2304	ohc	four	110	1bbl	3.15	3.58	9	8	
0	85	honda	gas	std	four	sedan	fwd	front	96.5	175.4	62.5	54.1	2372	ohc	four	110	1bbl	3.15	3.58	9	8	
0	85	honda	gas	std	four	sedan	fwd	front	96.5	175.4	65.2	54.1	2465	ohc	four	110	mpfi	3.15	3.58	9	10	
1	107	honda	gas	std	two	sedan	fwd	front	96.5	169.1	66	51	2293	ohc	four	110	2bbl	3.15	3.58	9.1	10	
0	?	isuzu	gas	std	two	sedan	rwd	front	94.3	170.7	61.8	53.5	2337	ohc	four	111	2bbl	3.31	3.23	8.5	7	
1	?	isuzu	gas	std	two	sedan	fwd	front	94.5	155.9	63.6	52	1874	ohc	four	90	2bbl	3.03	3.11	9.6	7	
0	?	isuzu	gas	std	four	sedan	fwd	front	94.5	155.9	63.6	52	1909	ohc	four	90	2bbl	3.03	3.11	9.6	7	
2	?	isuzu	gas	std	two	hatchback	rwd	front	96	172.6	65.2	51.4	2734	ohc	four	119	spfi	3.43	3.23	9.2	9	

Read data from csv to R



copy the data in a project folder, here in my computer it's /home/zhihan/workspace/data_literacy/week2

it's the directory path from root.

set the work directory to the project folder. In R or RStudio, type

```
>setwd ('/home/zhihan/workspace/data_literacy/week2')
```

Or in RStudio, click Session -> Set Working Directory -> Choose Directory

Read data from csv to R

```
automobile.data <- read.csv('automobile.csv')
```

The above line will read data from csv file and assign the data to an data frame object(you can think it as a kind of ‘data container’). ‘automobile.data’ is the name of the data frame, the variable name is set by user.

type

```
automobile.data
```

Check the data that contains.

Directory path in computer system

absolute path

An absolute path points to the same location in the file system regardless of the current working directory. To do that, it must contain root directory. like

linux: /home/zhihan/workspace/data_literacy/week2

Windows: c:\users\zhihan

Linux use / as delimiter and windows use \ as delimiter

In a computer system, the root directory is the first or top-most directory in a hierarchy. Like

linux: /

Windows: c:\

relative path

A relative path starts from some given working directory, avoiding the need to provide full absolute path. like

if current working directory is
/home/zhihan/workspace/

And workspace contains folder data_literacy and data_literacy contains another path week2. Then the relative path for week2 is
data_literacy/week2

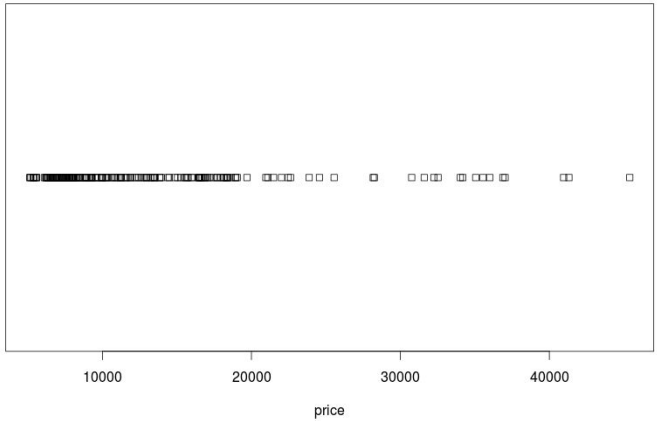
Besides, '.' represents current working direcotry and '..' represents parent directory.

Read data from csv to R

part of the data is shown as follows.

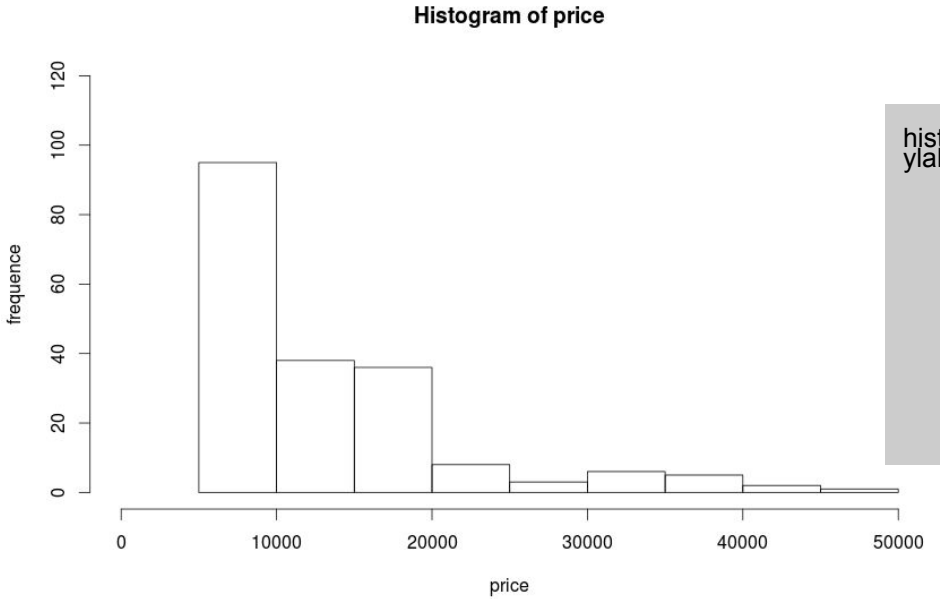
```
> automobile.data <- read.csv('automobile.csv')
> automobile.data
    symboling Normalized.losses         make fule.type aspiration number.of.doors  body.style drive.wheels engine.location wheel.base length width height
1           3                NA  alfa-romero       gas        std            two convertible          rwd           front       88.6  168.8  64.1   48.8
2           3                NA  alfa-romero       gas        std            two convertible          rwd           front       88.6  168.8  64.1   48.8
3           1                NA  alfa-romero       gas        std            two   hatchback          rwd           front       94.5  171.2  65.5   52.4
4           2               164         audi       gas        std           four       sedan          fwd           front       99.8  176.6  66.2   54.3
5           2               164         audi       gas        std           four       sedan          4wd           front       99.4  176.6  66.4   54.3
6           2                NA         audi       gas        std            two       sedan          fwd           front       99.8  177.3  66.3   53.1
7           1               158         audi       gas        std           four       sedan          fwd           front      105.8  192.7  71.4   55.7
8           1                NA         audi       gas        std           four       wagon          fwd           front      105.8  192.7  71.4   55.7
9           1               158         audi       gas      turbo           four       sedan          fwd           front      105.8  192.7  71.4   55.9
10          2               192          bmw       gas        std            two       sedan          rwd           front      101.2  176.8  64.8   54.3
11          0               192          bmw       gas        std           four       sedan          rwd           front      101.2  176.8  64.8   54.3
12          0               188          bmw       gas        std            two       sedan          rwd           front      101.2  176.8  64.8   54.3
13          0               188          bmw       gas        std           four       sedan          rwd           front      101.2  176.8  64.8   54.3
14          1                NA          bmw       gas        std           four       sedan          rwd           front      103.5  189.0  66.9   55.7
15          0                NA          bmw       gas        std           four       sedan          rwd           front      103.5  189.0  66.9   55.7
16          0                NA          bmw       gas        std            two       sedan          rwd           front      103.5  193.8  67.9   53.7
17          0                NA          bmw       gas        std           four       sedan          rwd           front      110.0  197.0  70.9   56.3
18          2               121    chevrolet       gas        std            two   hatchback          fwd           front       88.4  141.1  60.3   53.2
19          1                98    chevrolet       gas        std            two   hatchback          fwd           front       94.5  155.9  63.6   52.0
20          0                81    chevrolet       gas        std           four       sedan          fwd           front       94.5  158.8  63.6   52.0
21          1               118        dodge       gas        std            two   hatchback          fwd           front       93.7  157.3  63.8   50.8
22          1               118        dodge       gas        std            two   hatchback          fwd           front       93.7  157.3  63.8   50.8
23          1               118        dodge       gas      turbo            two   hatchback          fwd           front       93.7  157.3  63.8   50.8
24          1               148        dodge       gas        std           four   hatchback          fwd           front       93.7  157.3  63.8   50.6
25          1               148        dodge       gas        std           four       sedan          fwd           front       93.7  157.3  63.8   50.6
26          1               148        dodge       gas        std           four       sedan          fwd           front       93.7  157.3  63.8   50.6
27         -1               110        dodge       gas        std           four       wagon          fwd           front      103.3  174.6  64.6   59.8
28          3               145        dodge       gas      turbo            two   hatchback          fwd           front       95.9  173.2  66.3   50.2
29          2               137        honda       gas        std            two   hatchback          fwd           front       86.6  144.6  63.9   50.8
30          2               137        honda       gas        std            two   hatchback          fwd           front       86.6  144.6  63.9   50.8
31          1               101        honda       gas        std            two   hatchback          fwd           front       93.7  150.0  64.0   52.6
32          1               101        honda       gas        std            two   hatchback          fwd           front       93.7  150.0  64.0   52.6
33          1               101        honda       gas        std            two   hatchback          fwd           front       93.7  150.0  64.0   52.6
34          0               110        honda       gas        std           four       sedan          fwd           front       96.5  163.4  64.0   54.5
35          0                78        honda       gas        std           four       wagon          fwd           front       96.5  157.1  63.9   58.3
36          0               106        honda       gas        std            two   hatchback          fwd           front       96.5  167.5  65.2   53.3
37          0               106        honda       gas        std            two   hatchback          fwd           front       96.5  167.5  65.2   53.3
38          0                85        honda       gas        std           four       sedan          fwd           front       96.5  175.4  65.2   54.1
39          0                85        honda       gas        std           four       sedan          fwd           front       96.5  175.4  62.5   54.1
40          0                85        honda       gas        std           four       sedan          fwd           front       96.5  175.4  65.2   54.1
41          1               107        honda       gas        std            two       sedan          fwd           front       96.5  169.1  66.0   51.0
42          0                NA        isuzu       gas        std           four       sedan          rwd           front       94.3  170.7  61.8   53.5
43          2                NA        isuzu       gas        std            two   hatchback          rwd           front       96.0  172.6  65.2   51.4
44          0               145       jaguar       gas        std           four       sedan          rwd           front      113.0  199.6  69.6   52.8
45          0                NA       jaguar       gas        std           four       sedan          rwd           front      113.0  199.6  69.6   52.8
46          0                NA       jaguar       gas        std            two       sedan          rwd           front      102.0  191.7  70.6   47.8
```

Distribution of Price




```
> stripchart(automobile.data$price,xlab = 'price')
```

Distribution of Price




```
hist(price,ylim=c(0,120),xlim=c(0,50000),xlab='price',
ylab='frequence')
```

Plot - brand and price relation

After data is loaded in R, it can be used for plotting.

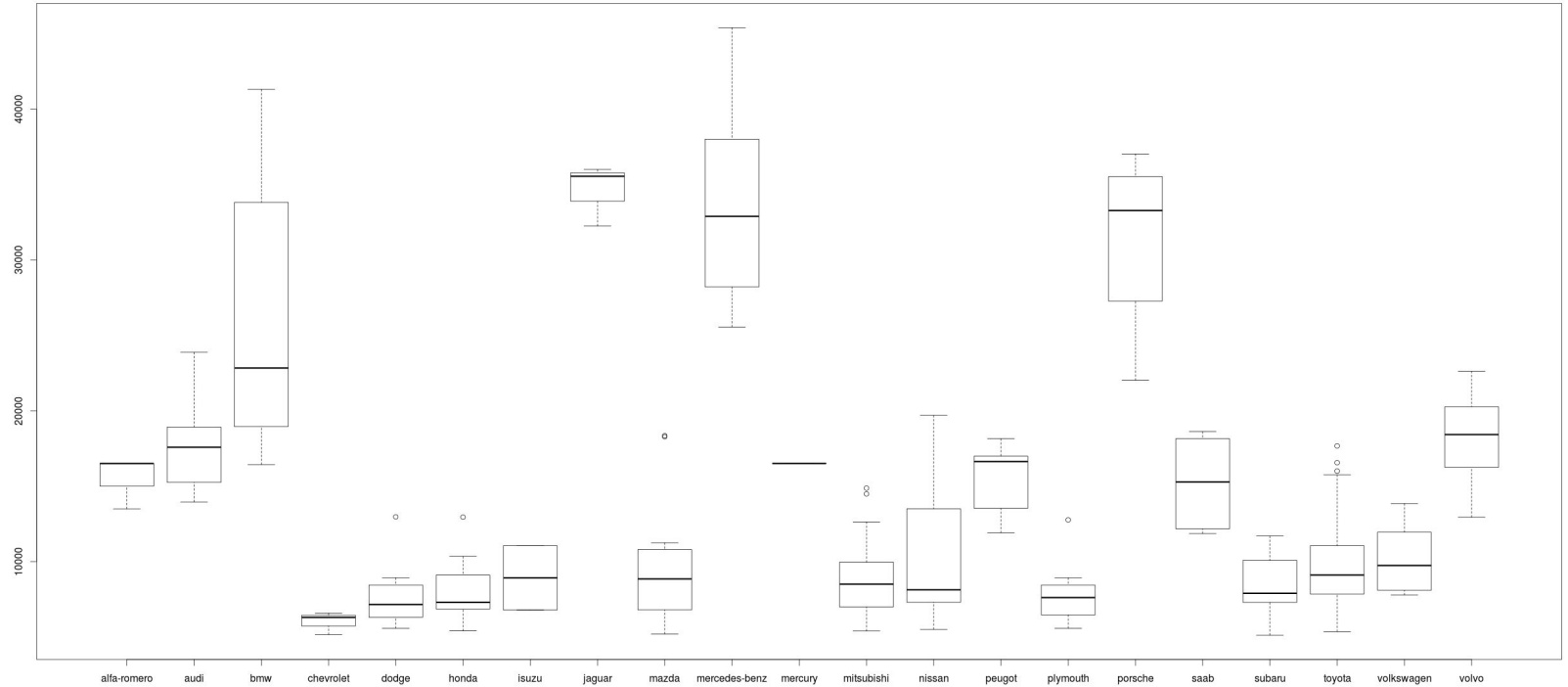
Plot the distribution of price

```
plot(automobile.data$price)
```

Plot the relation between make and price

```
plot(automobile.data$make,automobile.data$price)
```

Plot - brand and price relation



Plot - conclusion

Based on the data set, we can easily get following conclusions.

1. From the plot, if we set $20000 as the boundary price for Luxury car, $10000 as economy car. We can find the Luxury car maker.
BMV Janguar Benz Porche

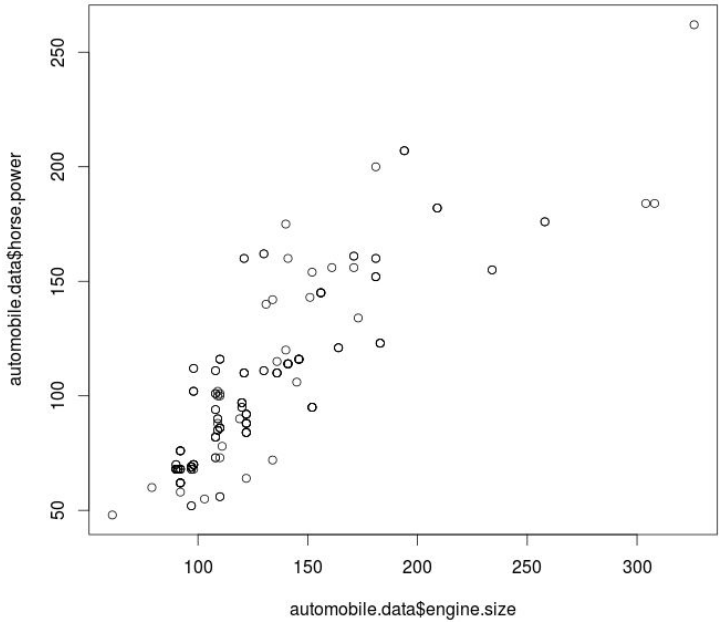
2.And economy car makers as follows
Toyota Honda Dodge Mazda and etc…

3. Those makers has big car price range: BMW, Benz, Nissan,Saab, Porsche.

4. Chevrolet produces the cheapest car among all makers

Plot - engine size and horse power

```
plot(automobile.data$engine.size,automobile.data$horse.power)
```



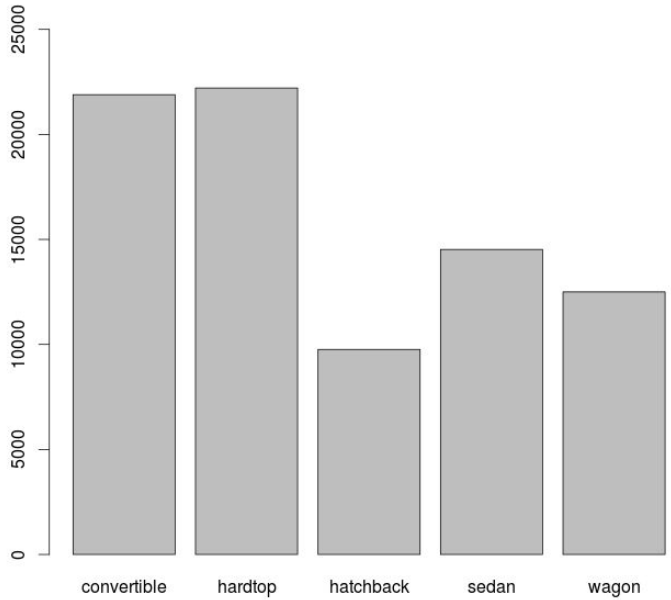
It is obvious that engine size and horse power has positive correlation.

Get more usage for plot function, use command

?plot

for help

barplot - body style and price



First, get the average price based on different car body styles.

```
tapply(automobile.data$price,automobile.data$body.style,mean)
```

Plot the relation of body style and average price

```
barplot(r,ylim = c(0,25000))
```

other basic plot functions

`plot(x)`	plot of the values of `x` (on the y-axis) ordered on the x-axis
`plot(x, y)`	bivariate plot of `x` (on the x-axis) and `y` (on the y-axis)
`sunflowerplot(x, y)`	id. but the points with similar coordinates are drawn as a flower which petal number represents the number of points
`pie(x)`	circular pie-chart
`boxplot(x)`	"box-and-whiskers" plot
`stripchart(x)`	plot of the values of `x` on a line (an alternative to `boxplot()` for small sample sizes)
`coplot(x~y | z)`	bivariate plot of `x` and `y` for each value (or interval of values) of `z`
`interaction.plot (f1, f2, y)`	if `f1` and `f2` are factors, plots the means of `y` (on the y-axis) with respect to the values of `f1` (on the x-axis) and of `f2` (different curves); the option `fun` allows to choose the summary statistic of `y` (by default `fun=mean`)
`matplot(x,y)`	bivariate plot of the first column of `x` *vs.* the first one of `y`, the second one of `x` *vs.* the second one of `y`, etc.
`dotchart(x)`	if `x` is a data frame, plots a Cleveland dot plot (stacked plots line-by-line and column-by-column)
`fourfoldplot(x)`	visualizes, with quarters of circles, the association between two dichotomous variables for different populations (`x` must be an array with `dim=c(2, 2, k)`, or a matrix with `dim=c(2, 2)` if $k = 1$)
`assocplot(x)`	Cohen–Friendly graph showing the deviations from independence of rows and columns in a two dimensional contingency table
`mosaicplot(x)`	'mosaic' graph of the residuals from a log-linear regression of a contingency table
`pairs(x)`	if `x` is a matrix or a data frame, draws all possible bivariate plots between the columns of `x`
`plot.ts(x)`	if `x` is an object of class `"ts"`, plot of `x` with respect to time, `x` may be multivariate but the series must have the same frequency and dates
`ts.plot(x)`	id. but if `x` is multivariate the series may have different dates and must have the same frequency
`hist(x)`	histogram of the frequencies of `x`
`barplot(x)`	histogram of the values of `x`
`qqnorm(x)`	quantiles of `x` with respect to the values expected under a normal law
`qqplot(x, y)`	quantiles of `y` with respect to the quantiles of `x`
`contour(x, y, z)`	contour plot (data are interpolated to draw the curves), `x` and `y` must be vectors and `z` must be a matrix so that `dim(z)=c(length(x), length(y))` (`x` and `y` may be omitted)
`filled.contour (x, y, z)`	id. but the areas between the contours are coloured, and a legend of the colours is drawn as well
`image(x, y, z)`	id. but the actual data are represented with colours
`persp(x, y, z)`	id. but in perspective
`stars(x)`	if `x` is a matrix or a data frame, draws a graph with segments or a star where each row of `x` is represented by a star and the columns are the lengths of the segments
`symbols(x, y, ...)`	draws, at the coordinates given by `x` and `y`, symbols (circles, squares, rectangles, stars, thermometres or "boxplots") which sizes, colours, etc, are specified by supplementary arguments
`termplot(mod.obj)`	plot of the (partial) effects of a regression model (`mod.obj`)