```
library(ISLR)
d1=Auto
str(d1)
```

```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 1
```

```
# remove factor last col
d2 = d1[,-9]

# cylinders and year are also factors
```

Basic stats

```
# make window wide
summary(d2)
```

```
##       mpg          cylinders      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##   acceleration        year           origin
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
##  Median :15.50   Median :76.00   Median :1.000
##  Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :24.80   Max.   :82.00   Max.   :3.000
```

```
# only means
apply(d2,2,mean)
```

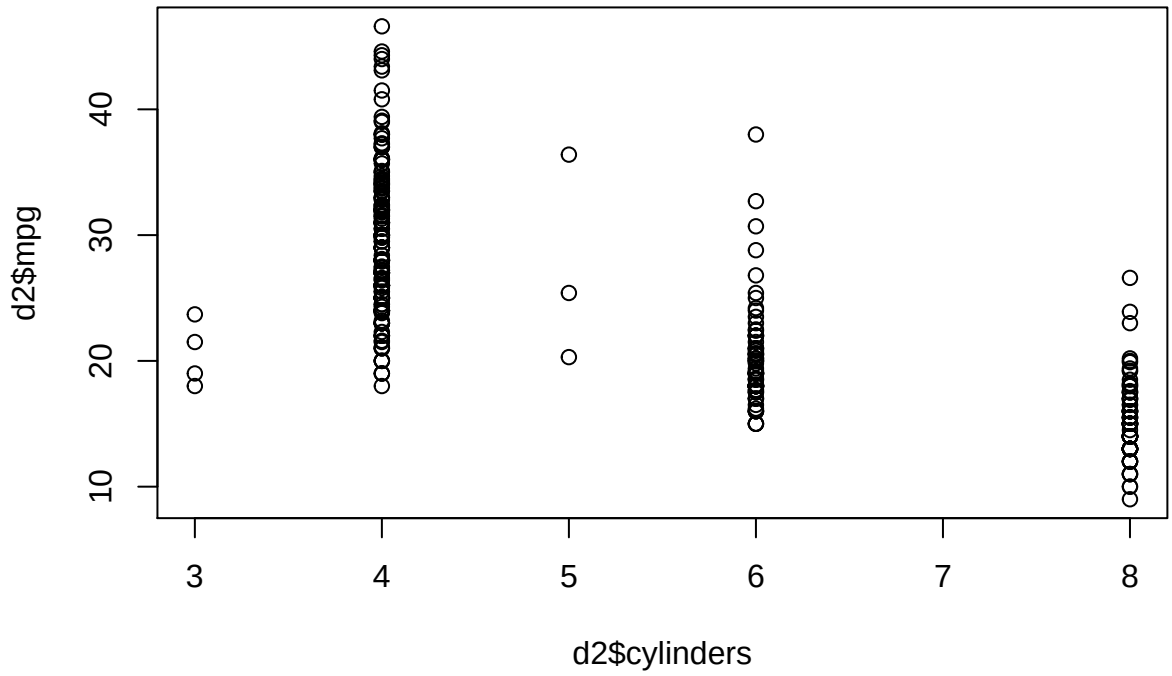
```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327
##         year       origin
##    75.979592     1.576531
```

```
# only mpg
summary(d2$mpg)
```

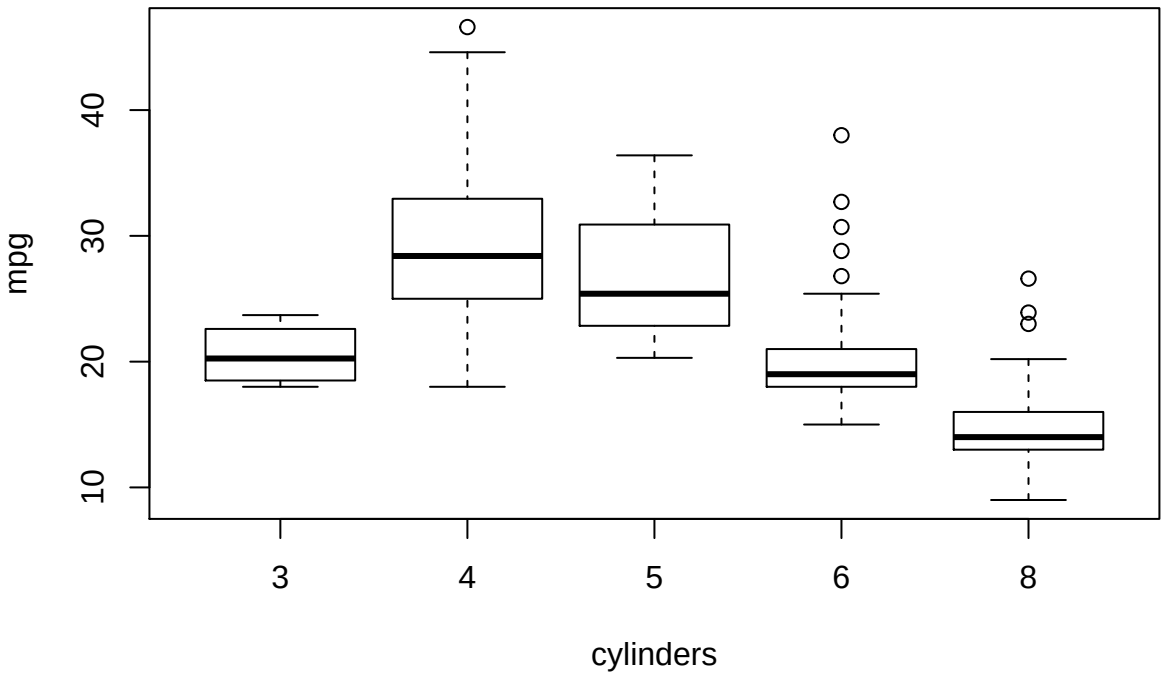
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   17.00   22.75   23.45   29.00   46.60
```

PLOTTING

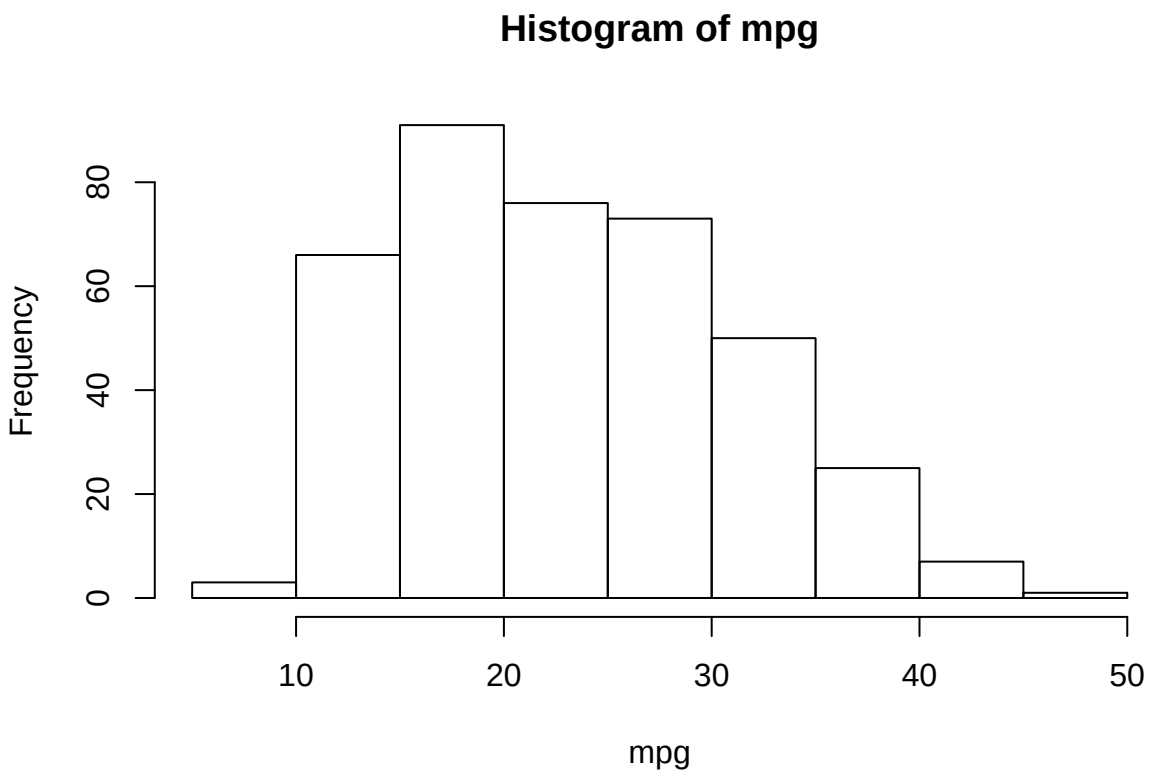
```
# scatterplot

# plot(cylinders, mpg)        # gives Error
plot(d2$cylinders, d2$mpg)
```



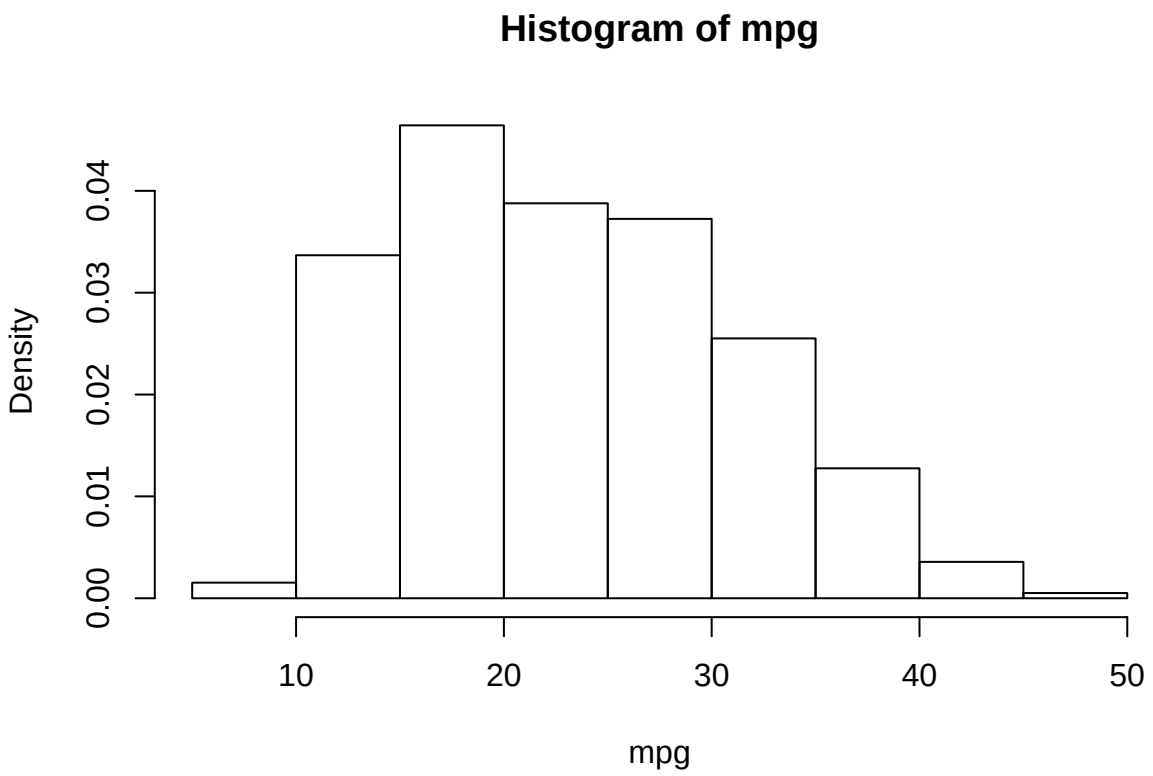

```
# boxplot
d2$cylinders=factor(d2$cylinders)
plot(mpg~cylinders,d2)
```



```r
# histogram
mpg = d2$mpg
hist(mpg)
```



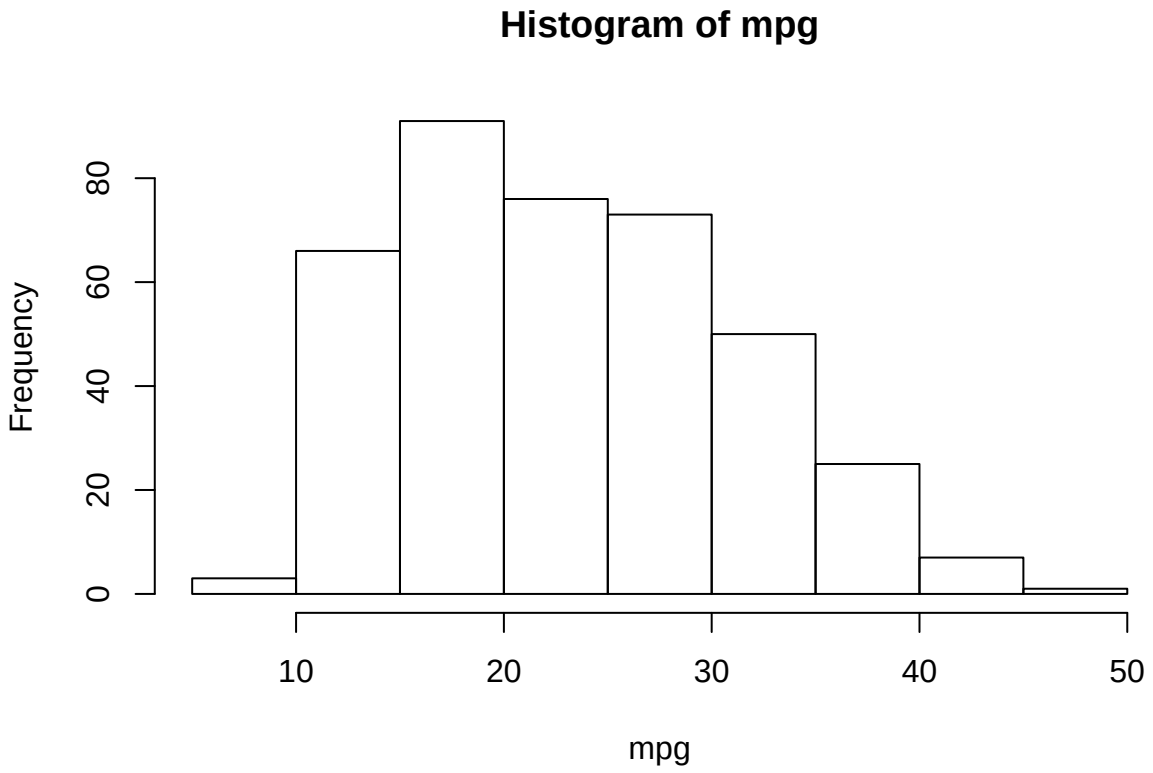

```r
hist(mpg,freq=F)  # not relative freqs

h1=hist(mpg,freq=F)
```



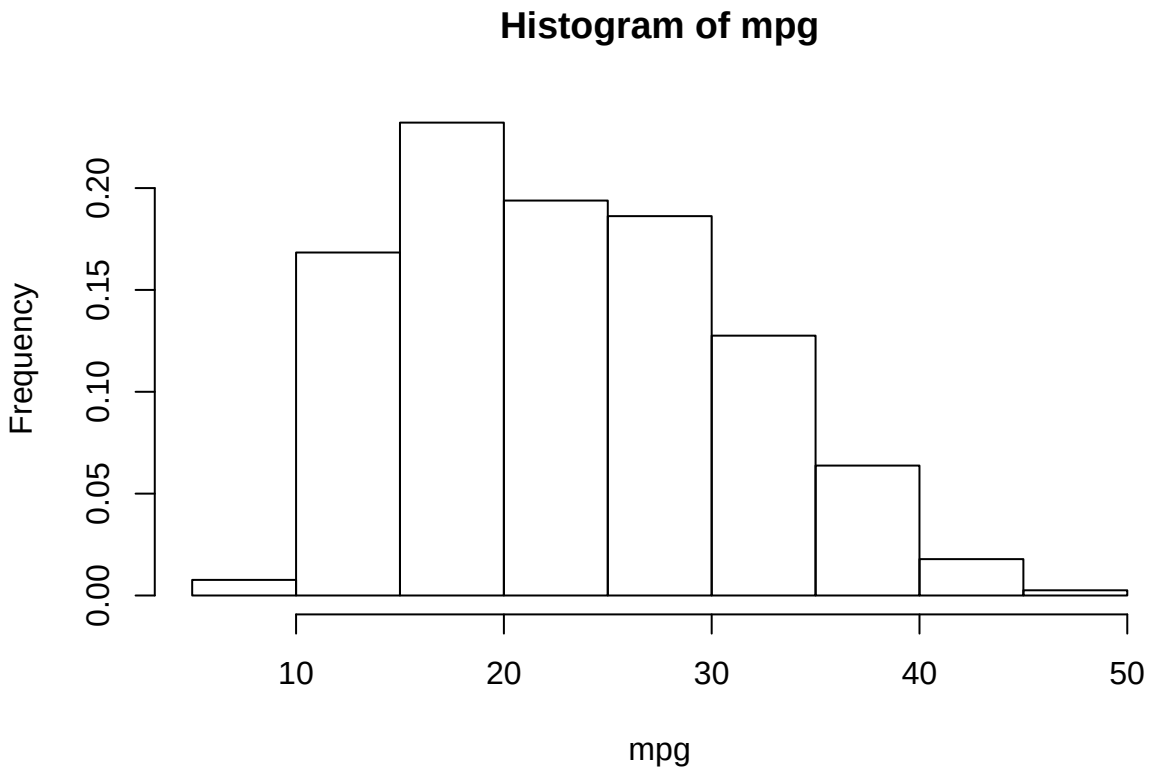

```
h1$breaks          # [1] 5 10 15 20 25 30 35 40 45 50
```

```
##  [1]  5 10 15 20 25 30 35 40 45 50
```

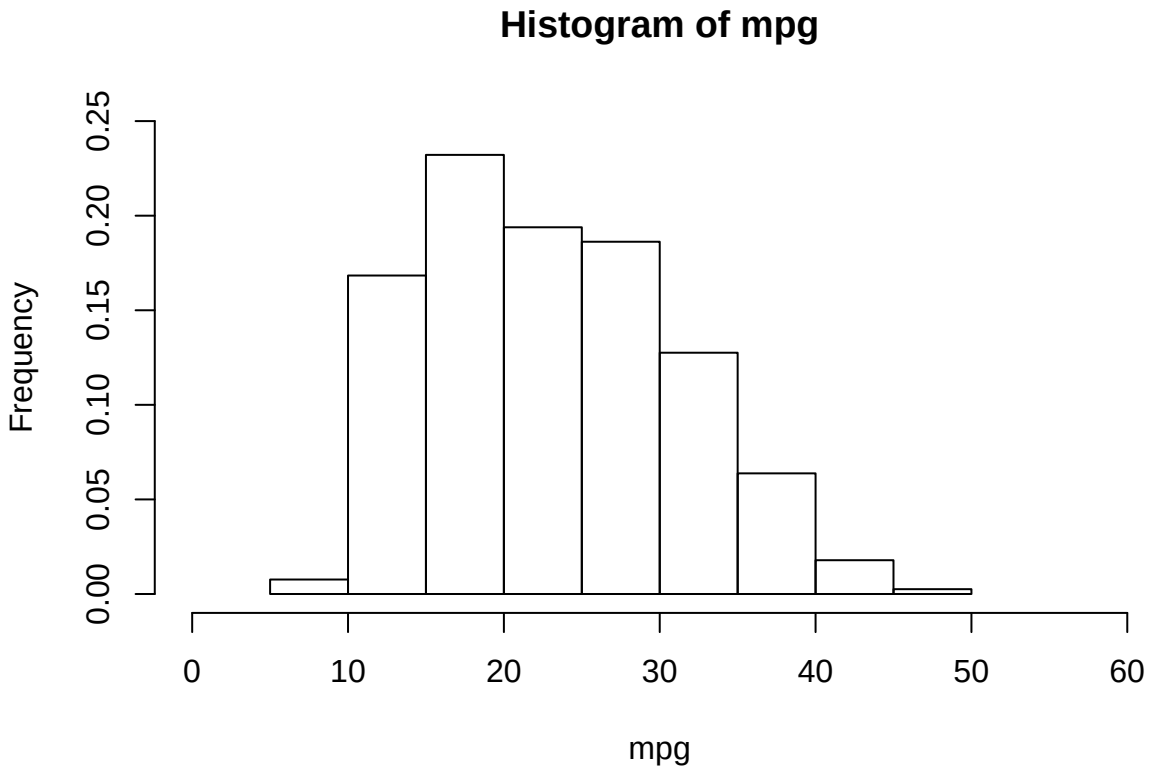
```
# not relative freq since bars width is not equal to 1
```



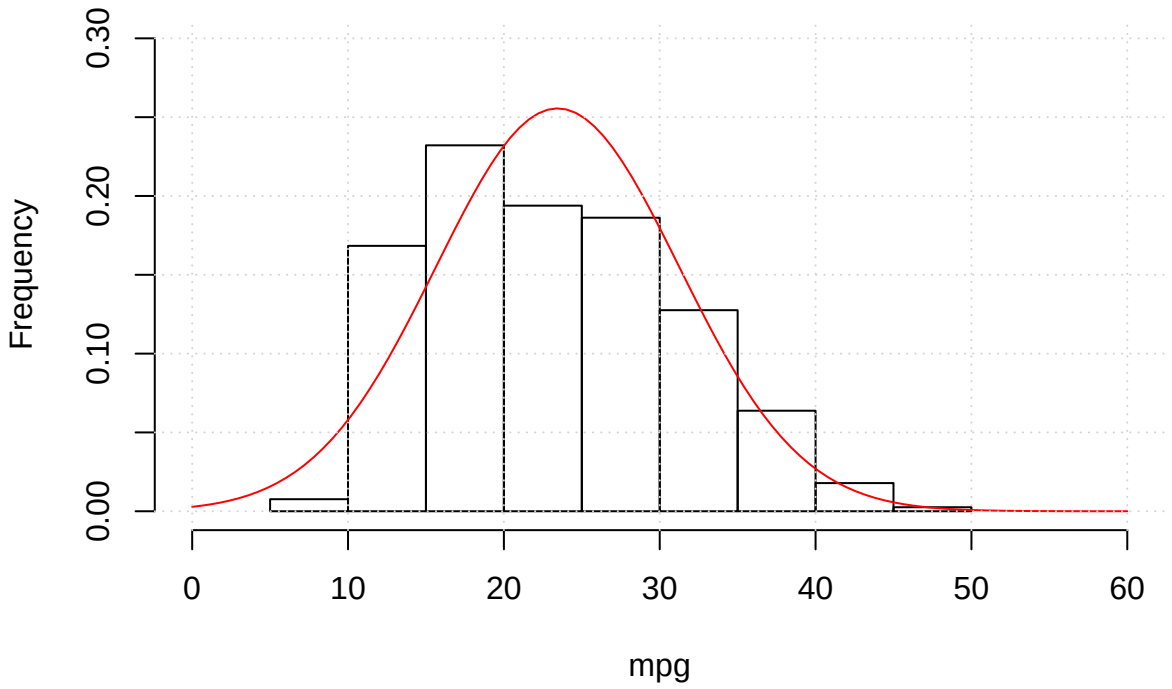
```r
hh$counts = hh$counts/sum(hh$counts)
plot(hh)
```



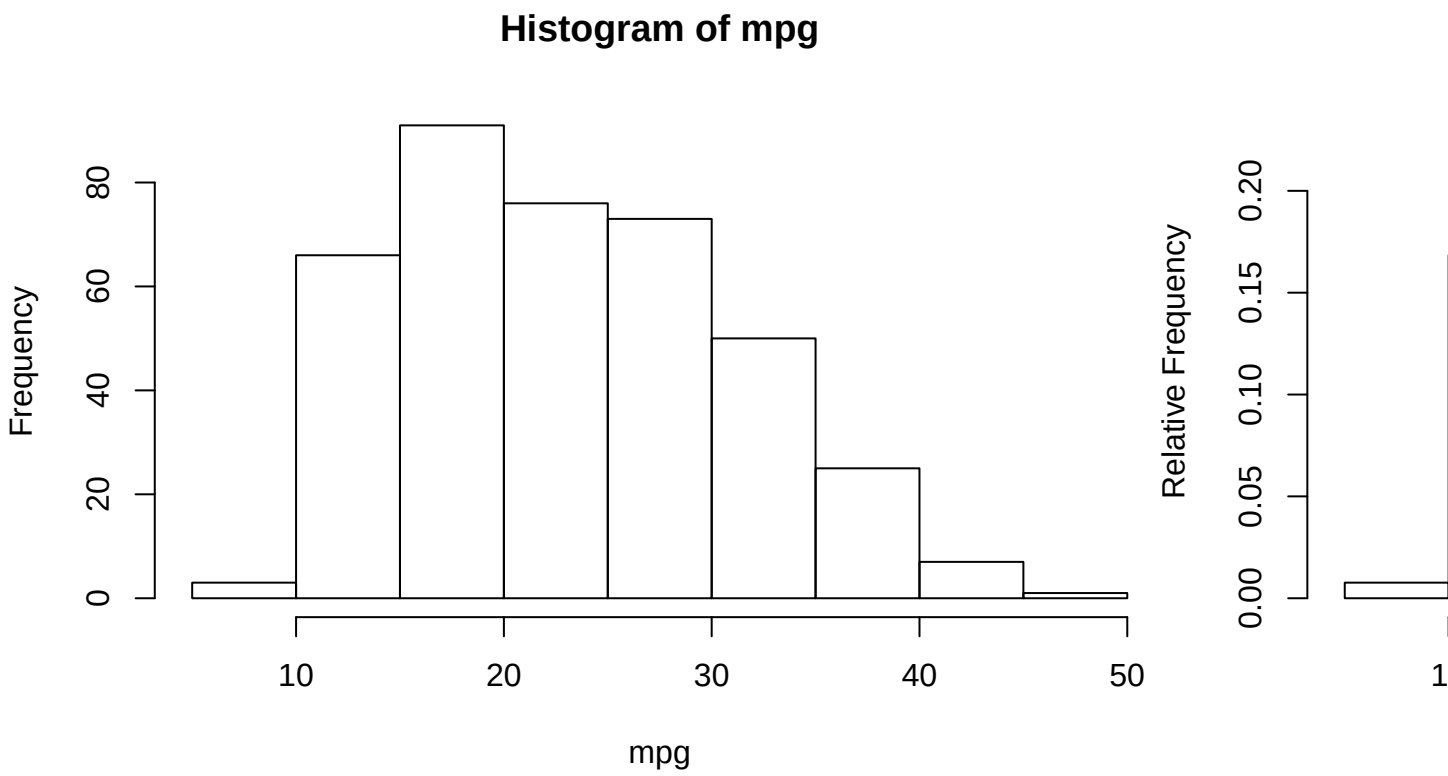

```r
# increase axes limits
plot(hh,xlim=c(0,60),ylim=c(0,0.25))
```



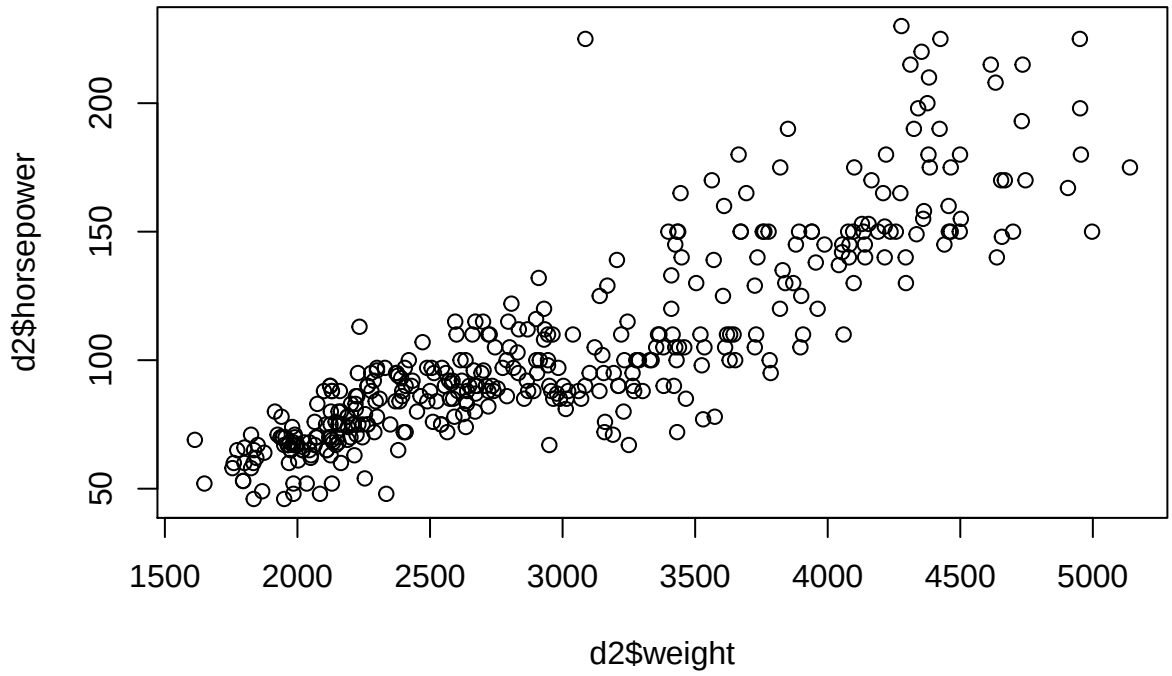
```r
# add normal density
width1 = hh$breaks[2]-hh$breaks[1]
mu = mean(mpg)
stdev = sd(mpg)
plot(hh,xlim=c(0,60),ylim=c(0,0.3),main="")
curve(dnorm(x,mu,stdev)*width1,col="red",add=T)
grid()

# or use

# install.packages("HistogramTools")
library(HistogramTools)
```



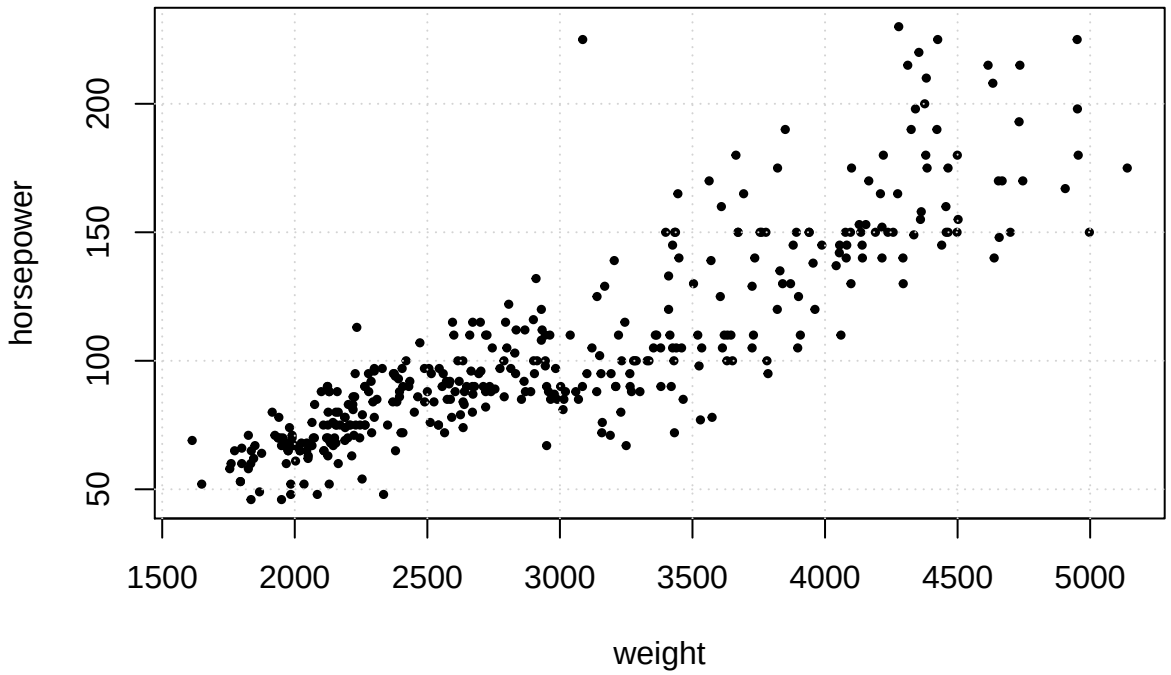

```r
PlotRelativeFrequency(hist(mpg))
```




```
# scatterplot

plot(d2$weight,d2$horsepower)
```



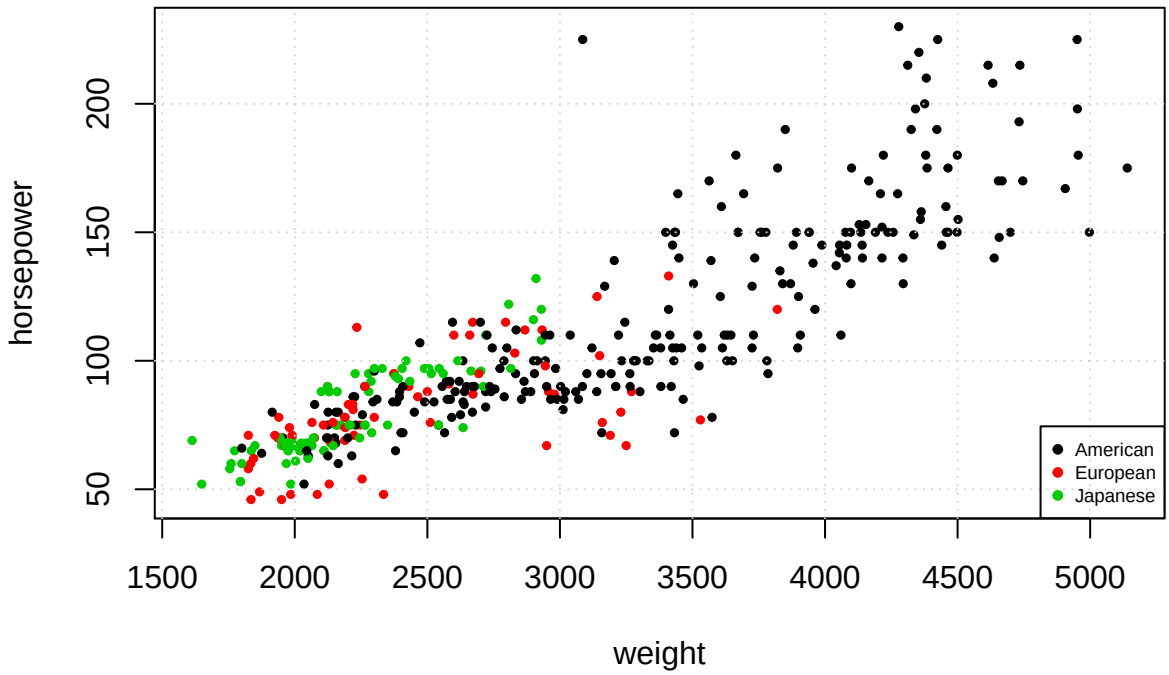

```
# change point character
plot(horsepower~weight,d2,pch=19,cex=0.5)
grid()
```




```r
unique(d2$origin)      # [1] 1 3 2
```

```
## [1] 1 3 2
```

```r
plot(horsepower~weight,d2,pch=19,cex=0.5,col=origin)
grid()

# legend
label = c("American","European","Japanese")
color = c(1,2,3)
char = c(19,19,19)
legend("bottomright",label,pch=char,cex=0.6,col=color)
```




```
# or
# legend(4500,75,label,pch=char,cex=0.6,col=color)

# Regression line
plot(horsepower~weight,d2,pch=19,cex=0.5)
m1=lm(horsepower~weight,d2)
coefficients(m1)
```

```
##  (Intercept)       weight
## -12.18348470   0.03917702
```

```
abline(m1)
abline(m1,col="red")
abline(m1,col="red",lwd=2)
grid()

# predict horsepower
head(d2,3)
```

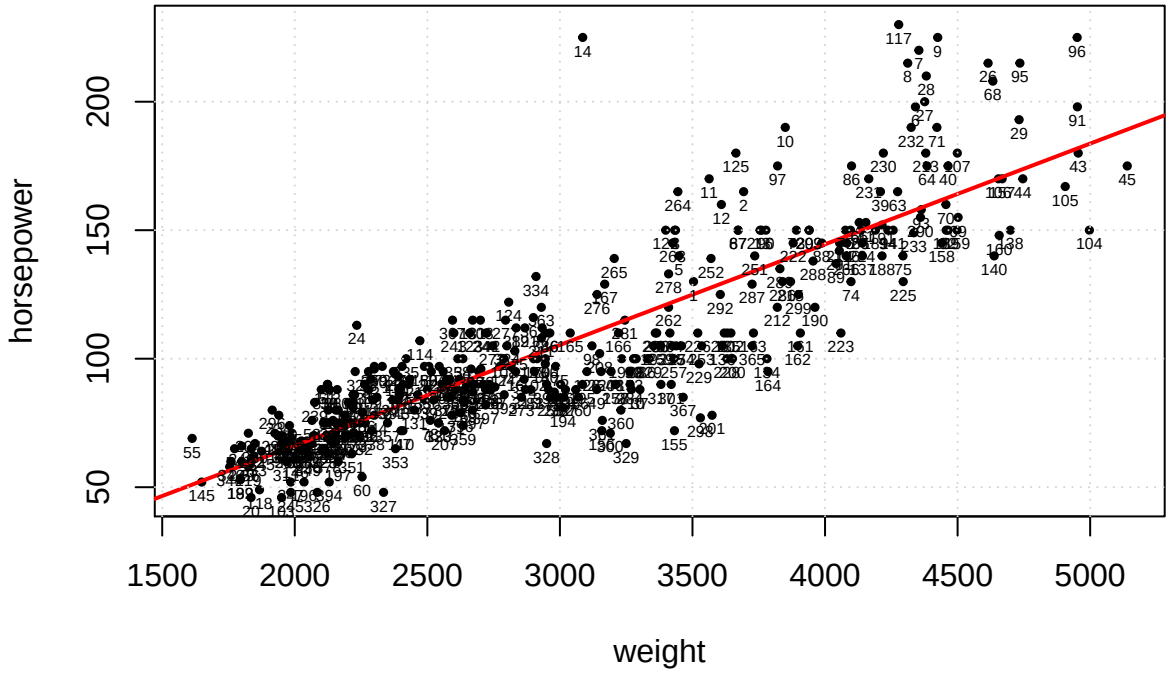
```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
```

```
newval = data.frame(weight=3000)
predict(m1,newval)   # 105.34
```

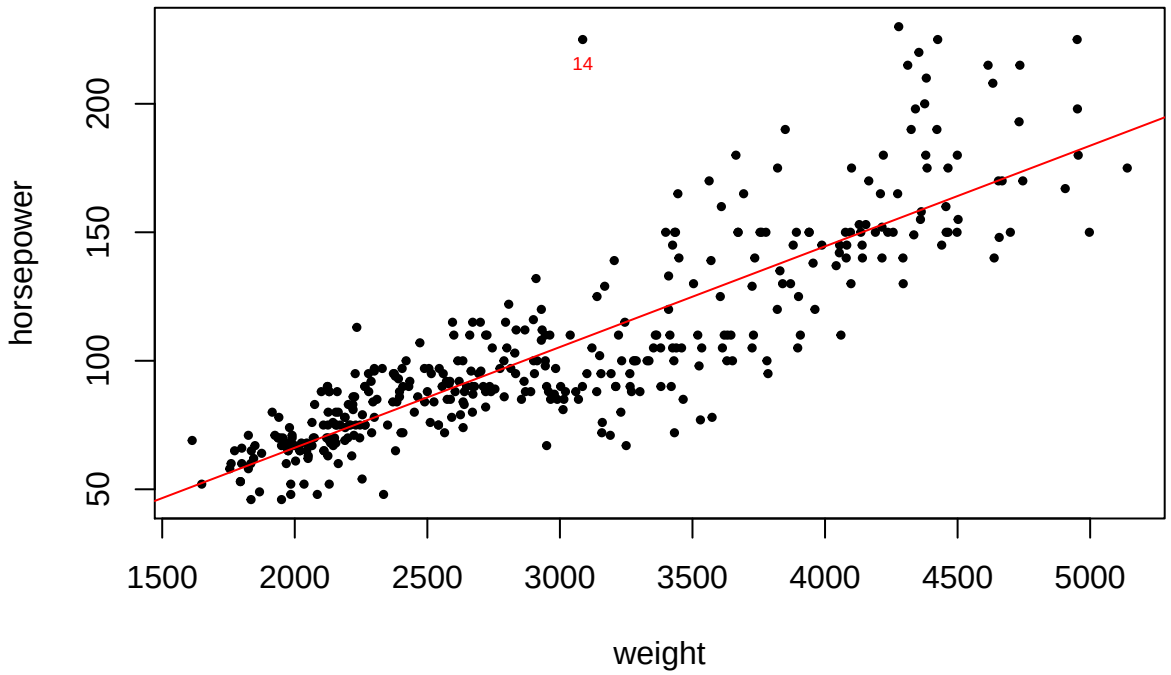
```
##        1
## 105.3476
```

```
# outliers
res=resid(m1)
idx=which(res==max(res))   # 14

# locator
```

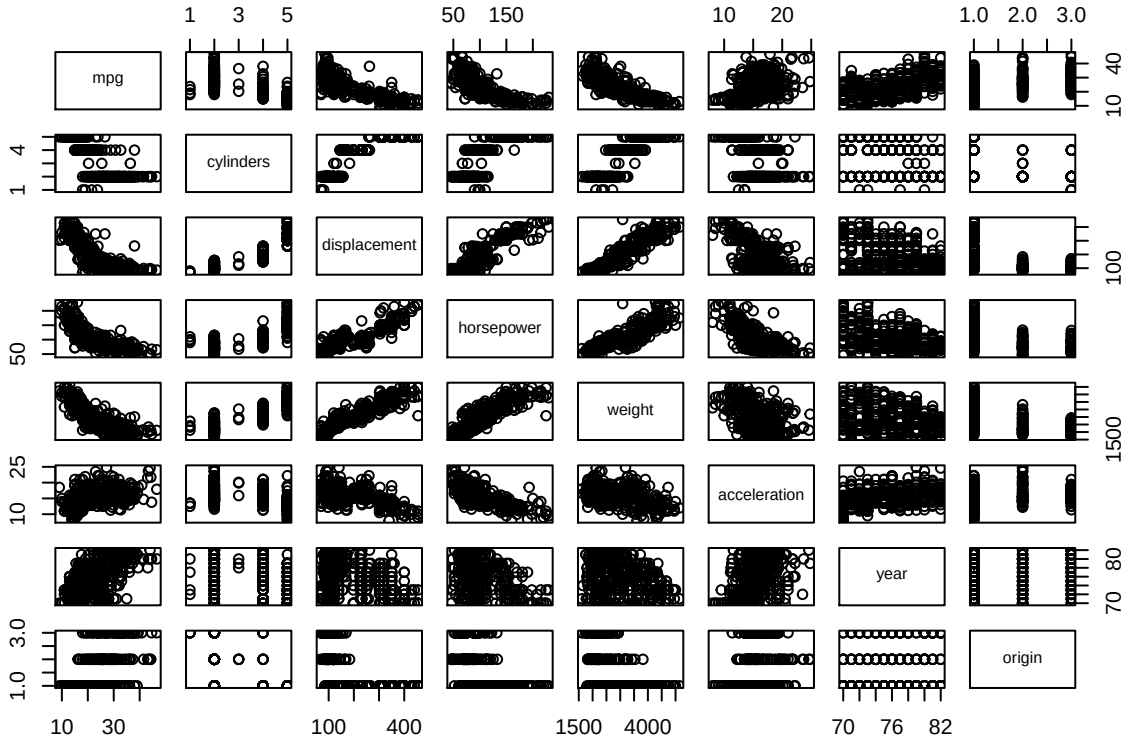
```r
# identify(d2$weight,d2$horsepower,rownames(d2),cex=0.5)  # rownames is default id
# identify(d2$weight,d2$horsepower,d2$horsepower,cex=0.5)
# d2[14,]

# label all points
text(horsepower~weight,data=d2,labels=rownames(d2),pos=1,offset=0.25,cex=0.5)
```



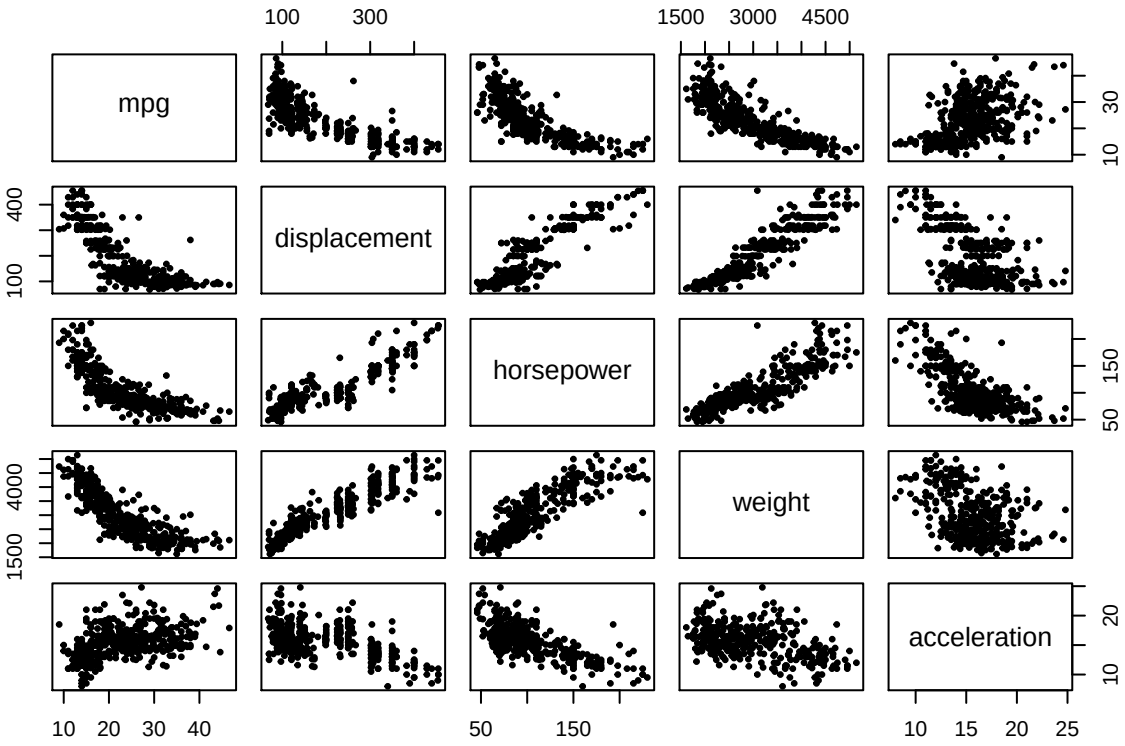

```r
# just label the outlier
plot(horsepower~weight,d2,pch=19,cex=0.5)
abline(m1,col="red")
label = rep("",392)
res   = resid(m1)
idx   = which(res==max(res))
label[idx]=idx
text(horsepower~weight,d2,labels=label,pos=1,offset=0.5,cex=0.6,col=2)
```



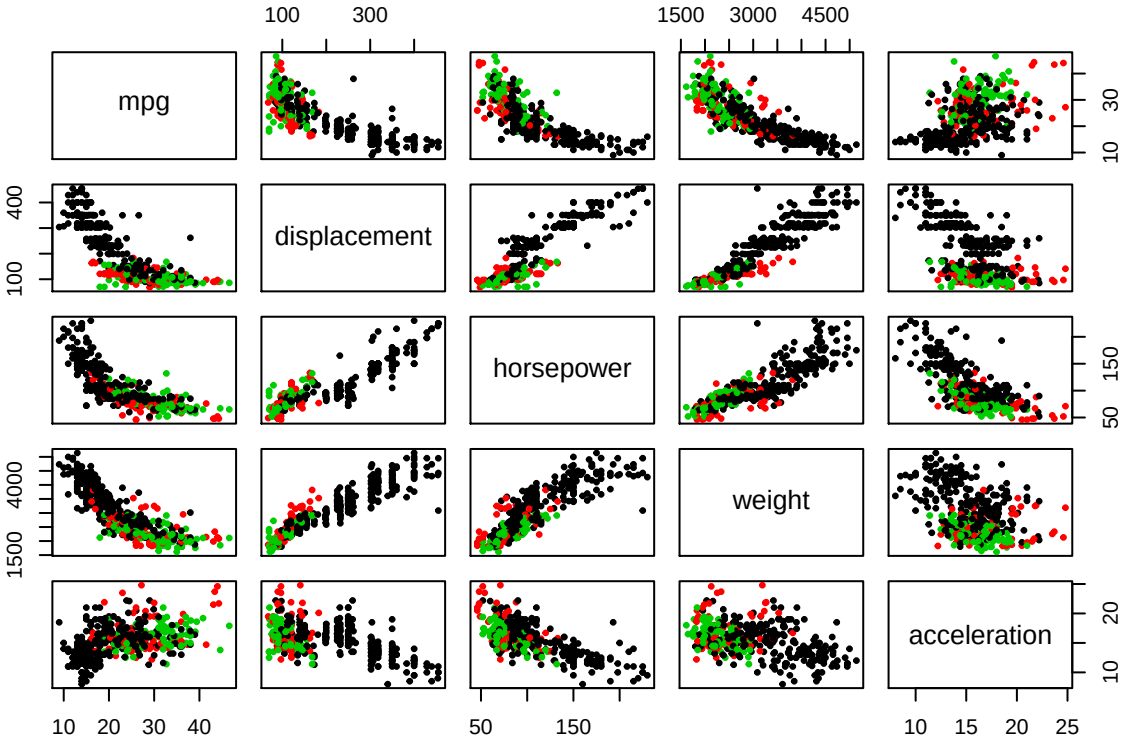

```
# pairs of scatterplots
pairs(d2)
```




```
# only numeric variables
pairs(~ mpg + displacement + horsepower + weight + acceleration,d2,pch=19,cex=0.5)
```




```
# change point character
pairs(~ mpg + displacement + horsepower + weight + acceleration,d1,pch=19,cex=0.5,col=d1$origin)
```

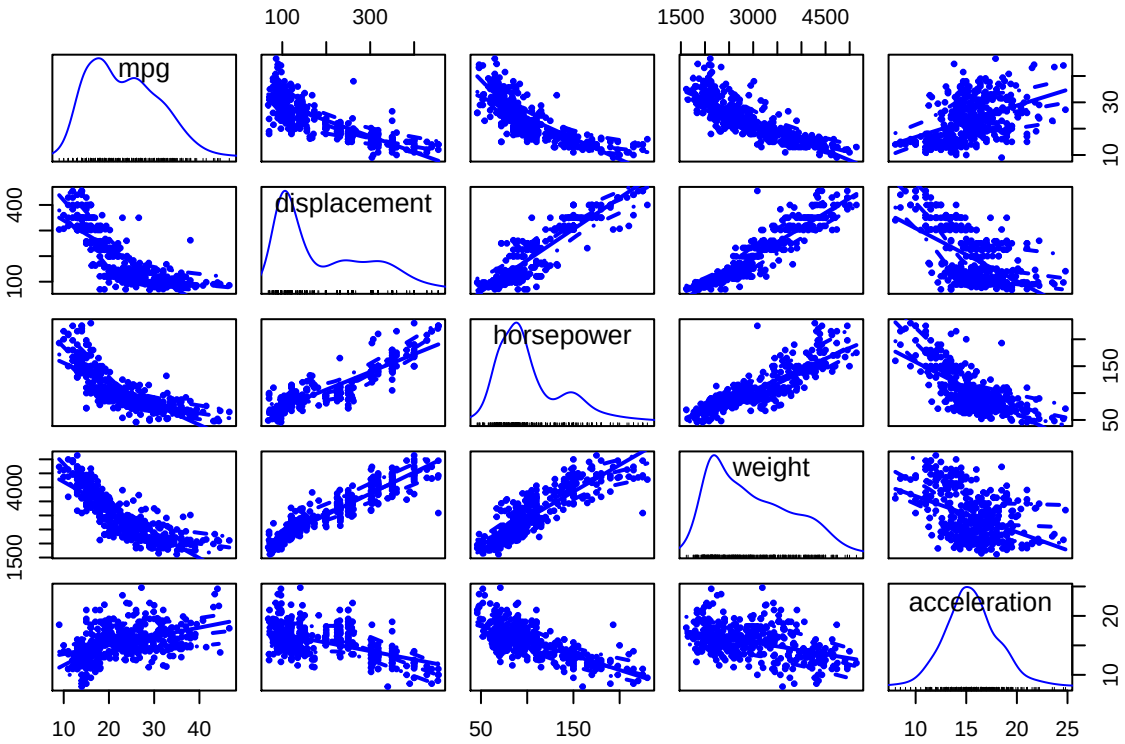



```
# load panel.hist() function
d3=d2[,-c(2,7:9)]  # remove factors
#pairs(d3,panel = panel.smooth,cex = 0.6,pch = 19,diag.panel = panel.hist,cex.labels = 0.8,font.labels
```

```
# use scatterplotMatrix() from library car
library(car)
```

```
## Loading required package: carData
```

```
scatterplotMatrix(~ mpg + displacement + horsepower + weight + acceleration,d2,pch=19,cex=0.5)
```




```
# mpg, displacement, hp, weight, acceleration seem correlated

# change diagonal to histograms
scatterplotMatrix(~ mpg + displacement + horsepower + weight + acceleration,d2,pch=19,cex=0.5,diagonal=
```

```
## Warning in applyDefaults(diagonal, defaults = list(method =
## "adaptiveDensity"), : unnamed diag arguments, will be ignored
```

```
# covariance matrix
d3=d2[,-c(2,7,8)]
cov(d3)
```

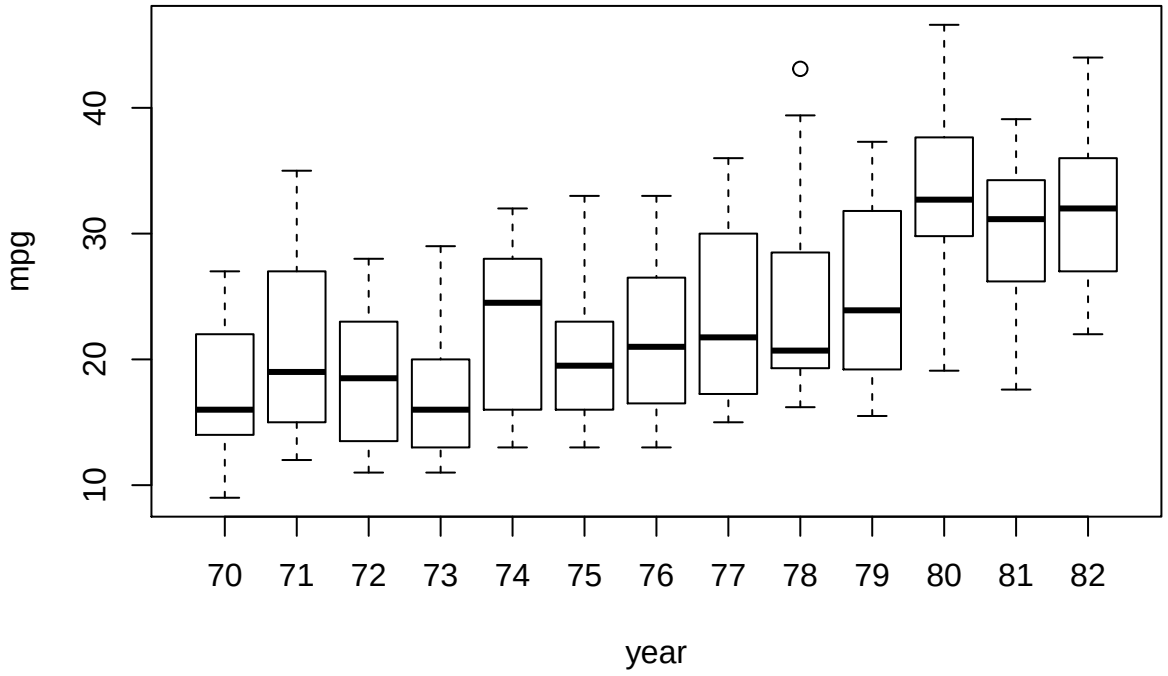
```
##                      mpg displacement  horsepower      weight acceleration
## mpg             60.918142    -657.5852  -233.85793  -5517.4407     9.115514
## displacement  -657.585207   10950.3676  3614.03374   82929.1001  -156.994435
## horsepower    -233.857926    3614.0337  1481.56939   28265.6202   -73.186967
## weight       -5517.440704   82929.1001 28265.62023 721484.7090  -976.815253
## acceleration     9.115514    -156.9944   -73.18697   -976.8153     7.611331
```

```
# correlation matrix
cor(d3)
```

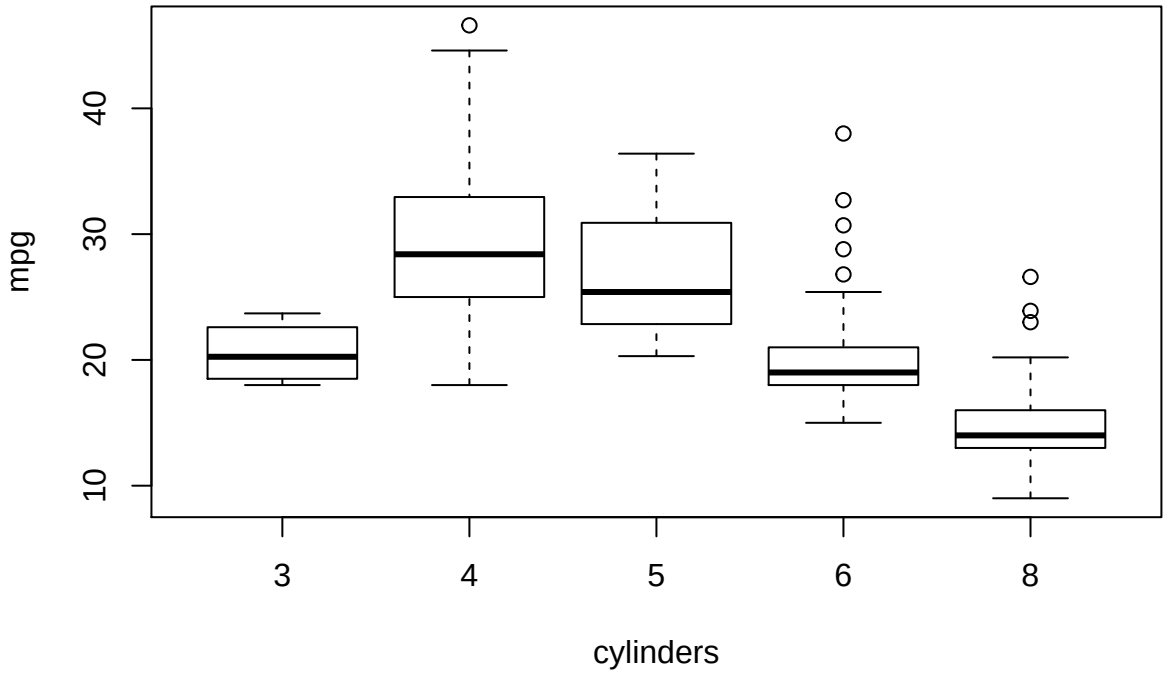
```
##                     mpg displacement horsepower     weight acceleration
## mpg           1.0000000   -0.8051269 -0.7784268 -0.8322442    0.4233285
## displacement -0.8051269    1.0000000  0.8972570  0.9329944   -0.5438005
## horsepower   -0.7784268    0.8972570  1.0000000  0.8645377   -0.6891955
```

```
## weight       -0.8322442    0.9329944  0.8645377  1.0000000   -0.4168392
## acceleration  0.4233285   -0.5438005 -0.6891955 -0.4168392    1.0000000
```

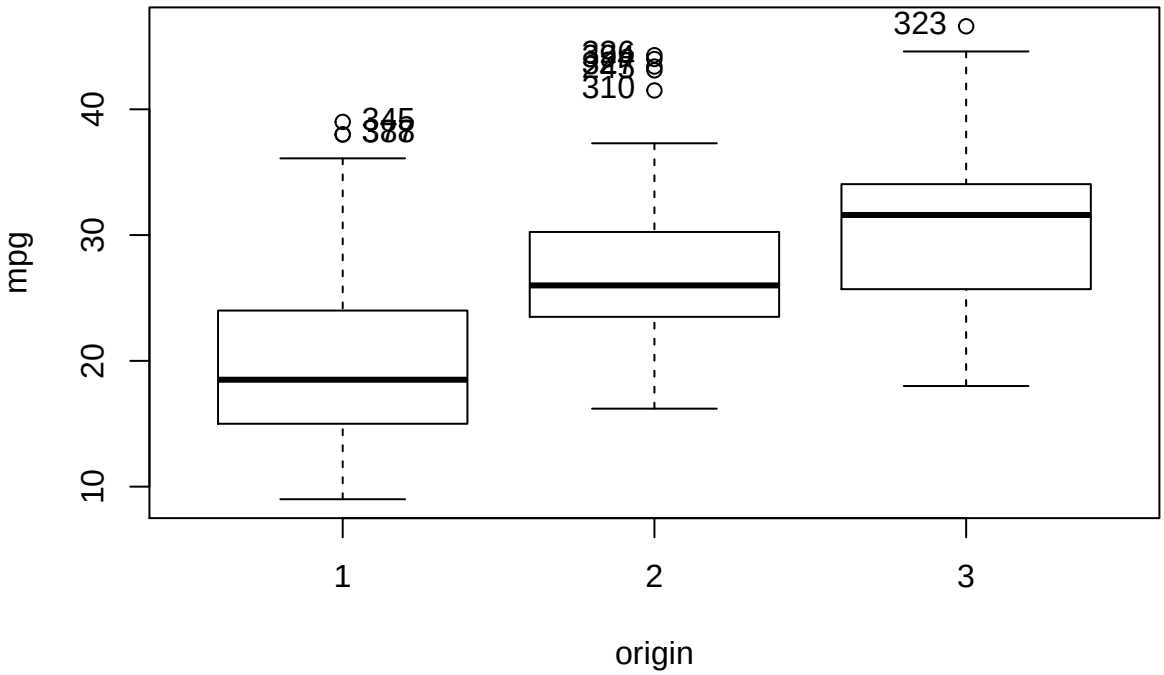
```
# boxplots
d3=d2
d3$origin=as.factor(d3$origin)
d3$year  =as.factor(d3$year)
d3$cylinders=as.factor(d3$cylinders)
plot(mpg~year,d3)
```




```
plot(mpg~cylinders,d3)
```



```r
# outliers
plot(mpg~origin,d3)         # same as
boxplot(mpg~origin,d3)
Boxplot(mpg~origin,d3)      # library(car) required
```

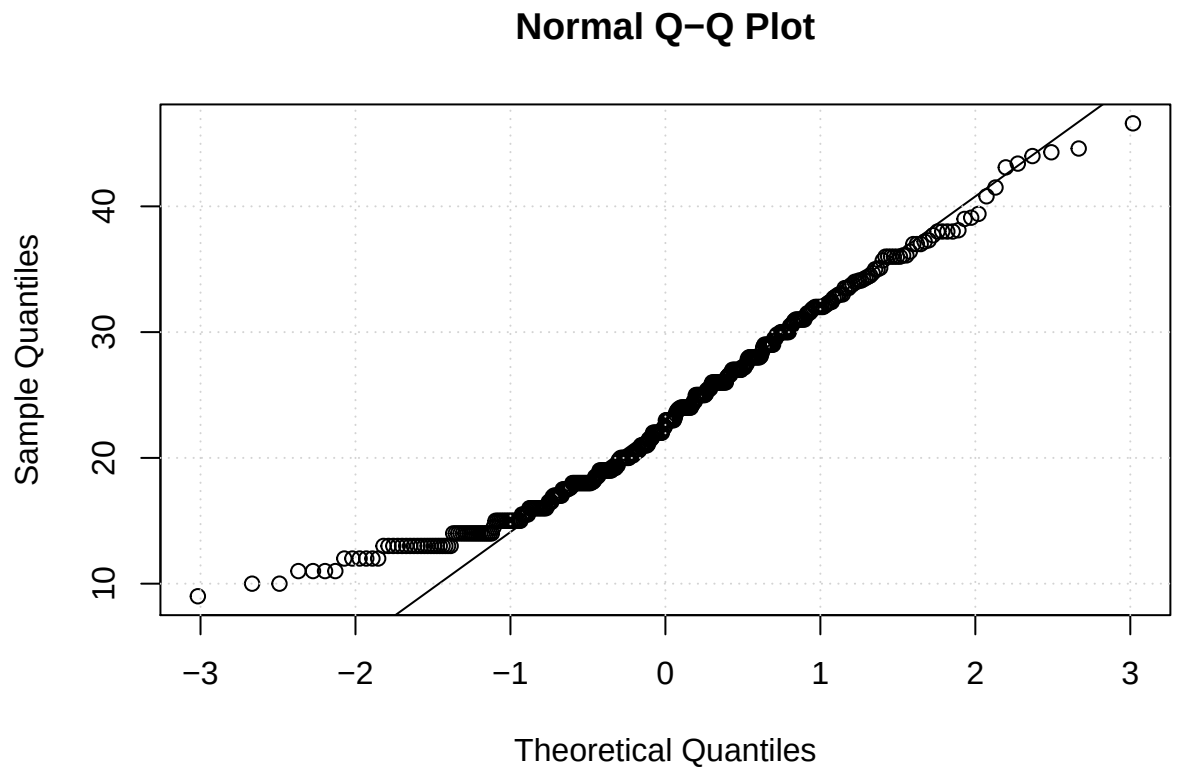



```
## [1] "345" "378" "387" "245" "310" "326" "327" "394" "323"
```

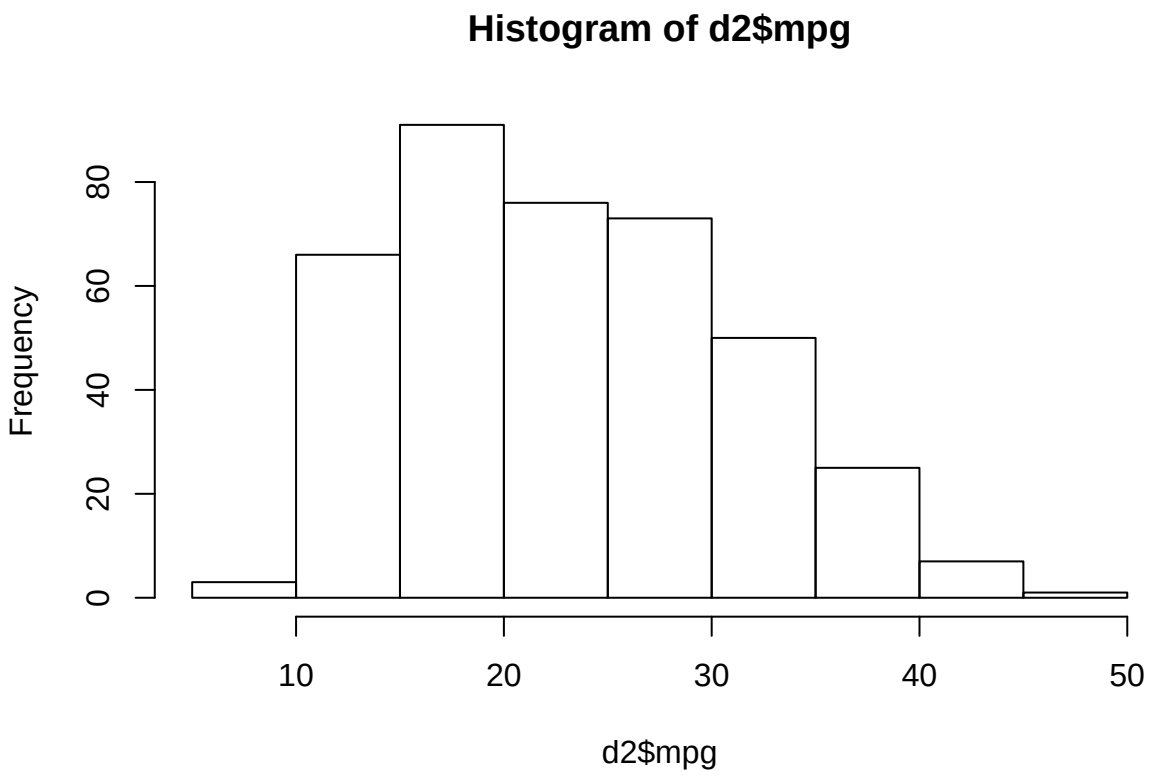
```r
# list outliers description
a=Boxplot(mpg~origin,d3)
d3[a,]
```

```
##      mpg cylinders displacement horsepower weight acceleration year origin
## 345 39.0         4           86         64   1875         16.4   81      1
## 378 38.0         4          105         63   2125         14.7   82      1
## 387 38.0         6          262         85   3015         17.0   82      1
## 245 43.1         4           90         48   1985         21.5   78      2
## 310 41.5         4           98         76   2144         14.7   80      2
## 326 44.3         4           90         48   2085         21.7   80      2
## 327 43.4         4           90         48   2335         23.7   80      2
## 394 44.0         4           97         52   2130         24.6   82      2
## 323 46.6         4           86         65   2110         17.9   80      3
```

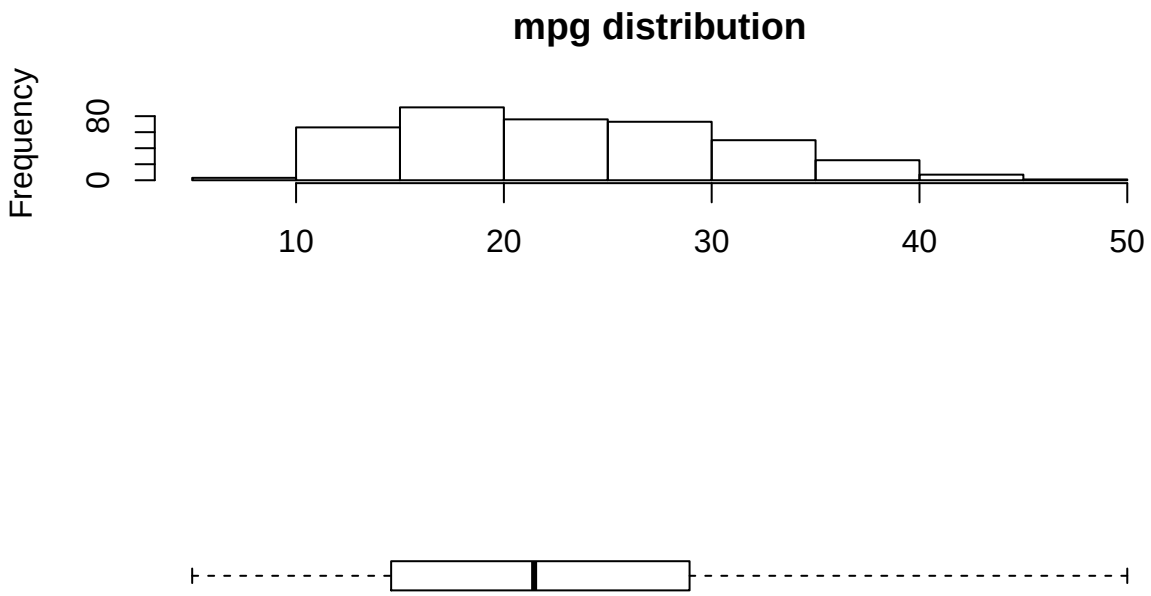
```r
# normality
qqnorm(d2$mpg)
qqline(d2$mpg)
grid()
```




```
hist(d2$mpg)
```




```
par(mfrow=c(2,1))
hist(d2$mpg,xlab="",main="mpg distribution")
boxplot(d2$mpg,horizontal=T,axes=F)
```



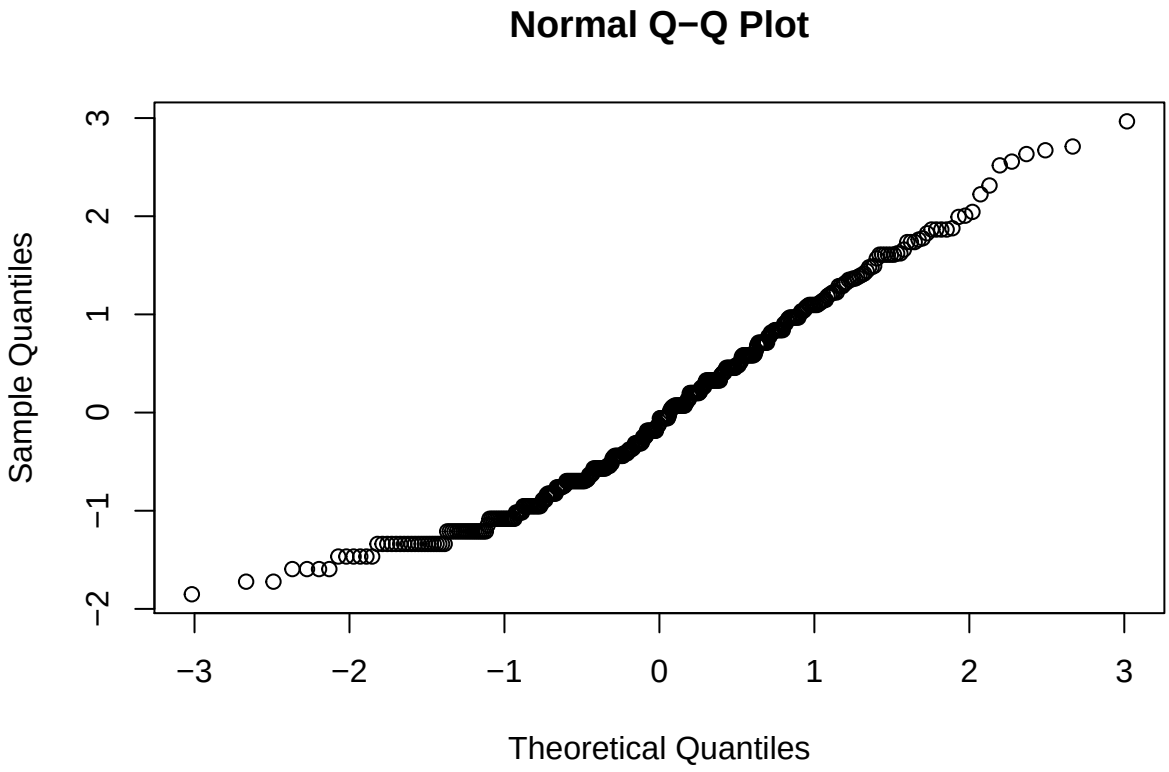

```
par(mfrow=c(1,1))

# compare sample vs theoretical quantiles
x = scale(d2$mpg)
mean(x)   # 0
```

```
## [1] 1.569283e-16
```

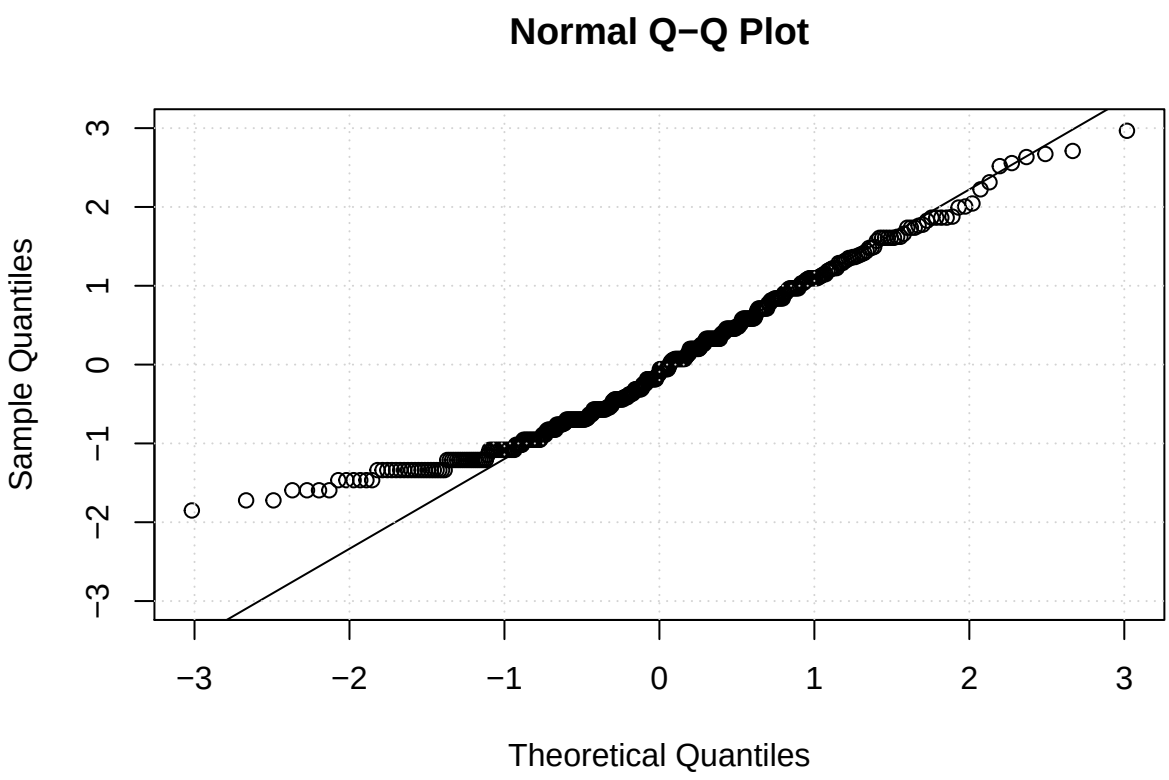
```
var(x)
```

```
##      [,1]
## [1,]    1
```

```
qqnorm(x)
```




```r
# change limits
qqnorm(x,ylim=c(-3,3))
qqline(x)
grid()
```



```r
a = seq(0,1,0.1)
a
```

```
##  [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

```r
quantile(x,a)
```

```
##          0%         10%         20%         30%         40%         50%
## -1.85085260 -1.21023822 -0.95399247 -0.69774672 -0.44150097 -0.08916306
##         60%         70%         80%         90%        100%
##  0.19911341  0.51557691  0.96528820  1.37656263  2.96656751
```

```r
qnorm(a)
```

```
##  [1]       -Inf -1.2815516 -0.8416212 -0.5244005 -0.2533471  0.0000000
##  [7]  0.2533471  0.5244005  0.8416212  1.2815516        Inf
```