```
library(ISLR)
d1=Auto
str(d1)
d2 = d1[,-9]               # remove factor last col
# cylinders, year are also factors

# basic stats - make window wide
summary(d2)
#      mpg          cylinders      displacement     horsepower        weight      acceleration
# Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613   Min.   : 8.00   Min.
# 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225   1st Qu.:13.78   1st
# Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804   Median :15.50   Medi
# Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978   Mean   :15.54   Mean
# 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615   3rd Qu.:17.02   3rd
# Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140   Max.   :24.80   Max.
apply(d2,2,mean)
#       mpg    cylinders displacement   horsepower       weight acceleration         year       origi
# 23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327    75.979592     1.57653

summary(d2$mpg)

# PLOTTING
#=======================================================
plot(cylinders, mpg)        # gives Error
plot(d2$cylinders, d2$mpg)  # scatterplot
d2$cylinders=factor(d2$cylinders)
plot(mpg~cylinders,d2)      # boxplot

# histogram
mpg = d2$mpg
hist(mpg)
hist(mpg,freq=F)  # not relative freqs
h1=hist(mpg,freq=F)
h1$breaks         # [1] 5 10 15 20 25 30 35 40 45 50
# not relative freq since bars width is not equal to 1

hh <-hist(mpg)
hh$counts = hh$counts/sum(hh$counts)
plot(hh)
plot(hh,xlim=c(0,60),ylim=c(0,0.25))

# 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)
PlotRelativeFrequency(hist(mpg))



# scatterplot
plot(d2$weight,d2$horsepower)
plot(horsepower~weight,d2,pch=19,cex=0.5)
grid()

unique(d2$origin)      # [1] 1 3 2
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)
legend(4500,75,label1,pch=char1,cex=0.6,col=col1)

# fitted 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 mileage
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
```

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

# locator
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,]
#    mpg cylinders displacement horsepower weight acceleration year origin
# 14  14         8          455        225   3086           10   70      1 buick estate wagon

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

# just label the outlier
label = rep("",392)
res   = resid(m1)
idx   = which(res==min(res))
label[idx]=idx
text(horsepower~weight,d2,labels=label,pos=1,offset=0.5,cex=0.6,col=2)

# pairs
pairs(d2)
pairs(~ mpg + displacement + horsepower + weight + acceleration,d2,pch=19,cex=0.5)
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)]
pairs(d3,panel = panel.smooth,cex = 0.6,pch = 19,diag.panel = panel.hist,cex.labels = 0.8,font.labels

library(car)
scatterplotMatrix(~ mpg + displacement + horsepower + weight + acceleration,d2,pch=19,cex=0.5)
scatterplotMatrix(~ mpg + displacement + horsepower + weight + acceleration,d2,pch=19,cex=0.5,diagona
# mpg, displacement, hp, weight, acceleration seem correlated

# correlations
d3=d2[,-c(2,7,8)]
cov(d3)
cor(d3)

# boxplot      library(car) : Boxplot(outlier의 index도 출력
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)
```

```
# outliers
plot(mpg~origin,d3)        # same as
boxplot(mpg~origin,d3)
Boxplot(mpg~origin,d3)     # library(car) required
a=Boxplot(mpg~origin,d3)
d3[a,]

# 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
qqnorm(x)

qqnorm(x,ylim=c(-3,3))
qqline(x)
grid()

a = seq(0,1,0.1)  # 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
quantile(x,a)
qnorm(a)
```