Stats216v: Statistical Learning

Stanford University
Summer 2017
Gyu-Ho Lee (gyuhox@gmail.com)

2. Statistical Learning

2.3 Lab: Introduction to R

In [1]:
```
x = c(1,6,2)
y = c(1,4,3)

length(x) # 3
length(y) # 3
x + y # 2 10 5

ls() # 'x' 'y'
rm(list=ls())
ls() # none
```

3

3

2 10 5

'x' 'y'

In [2]:
```
# help(summary)
# ?summary
# ?summary()

matrix(data=c(1,2,3,4), nrow=2, ncol=2)
matrix(data=c(1,2,3,4), nrow=2, ncol=2, byrow=TRUE)

sqrt(matrix(data=c(1,2,3,4), nrow=2, ncol=2))
matrix(data=c(1,2,3,4), nrow=2, ncol=2)^2
```

1	3
2	4

1	2
3	4

1.000000	1.732051
1.414214	2.000000

1	9

4	16

In [3]:
```
A = matrix(10:29, nrow=4, ncol=5)
dim(A)
A

A[1,1] # 10
A[2,3] # 19
A[4,5] # 29
```

4 5

10	14	18	22	26
11	15	19	23	27
12	16	20	24	28
13	17	21	25	29

10

19

29

In [4]:
```
A[1,]
A[1:2,]
A[,1:2]
A[1:2, 4:5]

# select rows and columns
A[c(1,4), c(1,2)]
```

10 14 18 22 26

10	14	18	22	26
11	15	19	23	27

10	14
11	15
12	16
13	17

22	26
23	27

10	14
13	17

In [5]:
```
# to generate exact same set of random numbers
set.seed(777)
rnorm(3) # 0.489786218319898 -0.398541384065542
 0.510836321690603
```

```
# same as 'rnorm(10, mean=0, sd=1)'

set.seed(10)
x = rnorm(100, mean=90, sd=3)
mean(x) # 89.5903531699334
var(x) # 7.97332499072424
sd(x) # 2.82370766736294
# same as 'sqrt(var(x))'

set.seed(50)
y = rnorm(100, mean=100, sd=.1)
mean(y) # 99.9886472853026
var(y) # 0.00985270532066419
sd(y) # 0.0992607944793119

cor(x, y) # -0.100259186855998

plot(x, y, xlab="this is x-axis", ylab="this is
 y-axis", main="plot!", col="blue")

# pdf("figure.pdf")
# dev.off()
```

```
      0.489786218319898   -0.398541384065542
      0.510836321690603
```

89.5903531699334

7.97332499072424

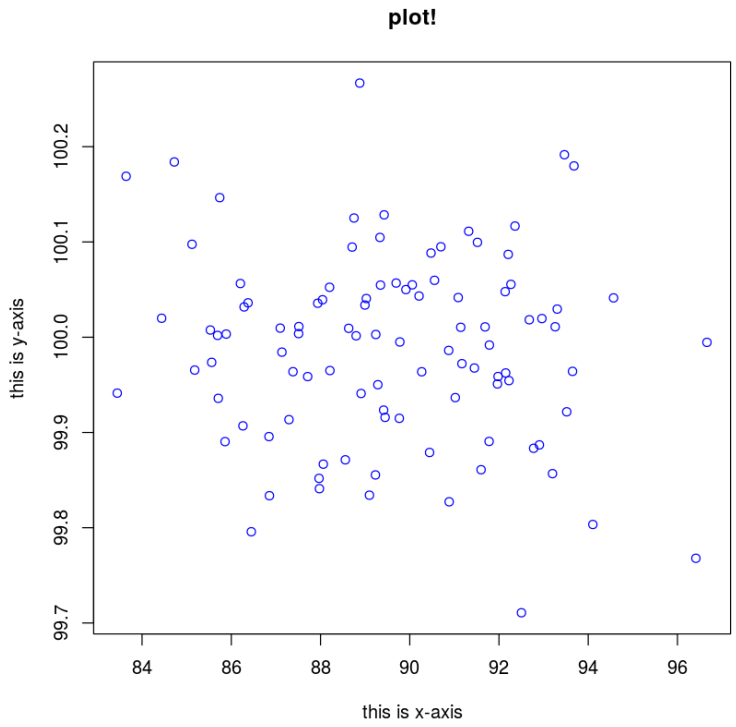
2.82370766736294

99.9886472853026

0.00985270532066419

0.0992607944793119

-0.100259186855998



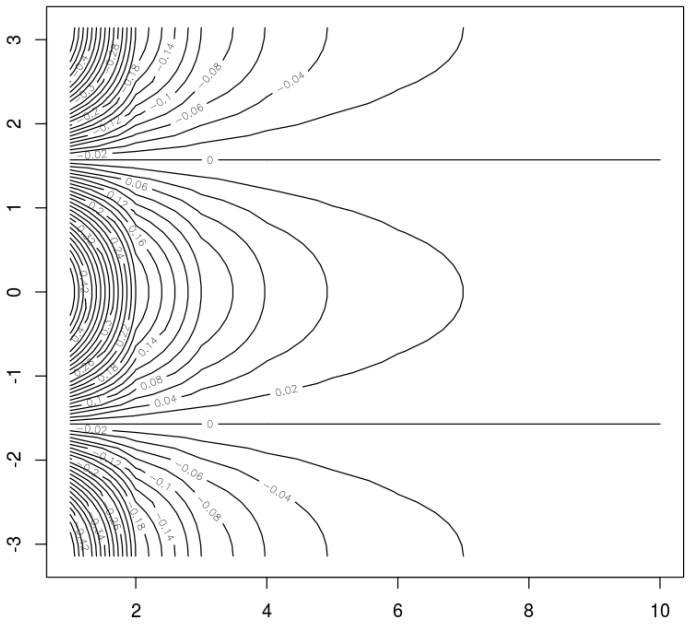
In [6]:
```
seq(1, 10) # 1 2 3 4 5 6 7 8 9 10
1:10 # 1 2 3 4 5 6 7 8 9 10
seq(-pi, pi, length=5) # -3.14159265358979 -1.5
707963267949 0 1.5707963267949 3.14159265358979
```

```
 1  2  3  4  5  6  7  8  9  10
```

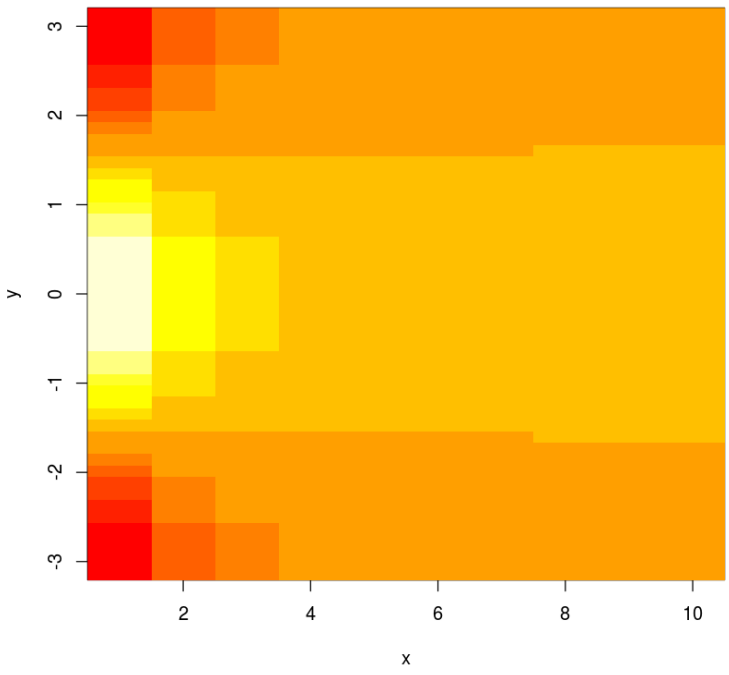
```
 1  2  3  4  5  6  7  8  9  10
```

```
-3.14159265358979  -1.5707963267949  0
1.5707963267949  3.14159265358979
```

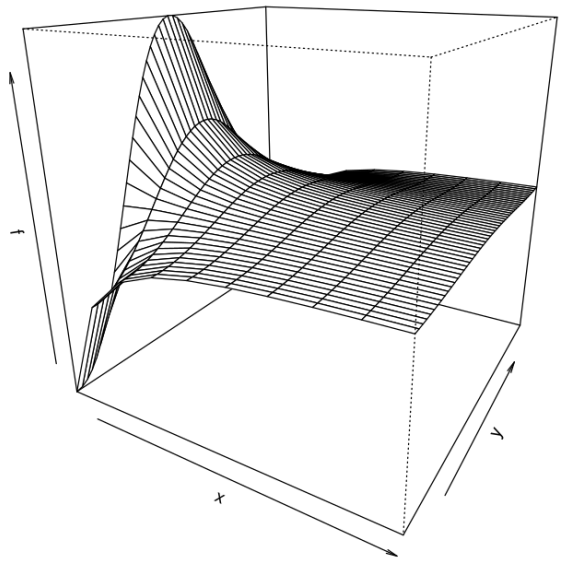
In [7]:
```
x = seq(1, 10)
y = seq(-pi, pi, length=50)
f = outer(x, y, function(x, y)cos(y)/(1+x^2))
contour(x, y, f, nlevels=45)
```



In [8]:
```
image(x, y, f)
```



In [9]:
```
persp(x, y, f, theta=30, phi=20)
```



In [10]:
```
AutoData = read.table("chapter-02-r-Auto.data",
 header=TRUE, na.strings="?")

# opens with 'vim'
# fix(AutoData)

dim(AutoData) # 397 9
AutoData = na.omit(AutoData)
dim(AutoData) # 392 9

names(AutoData)
summary(AutoData)
```

397 9

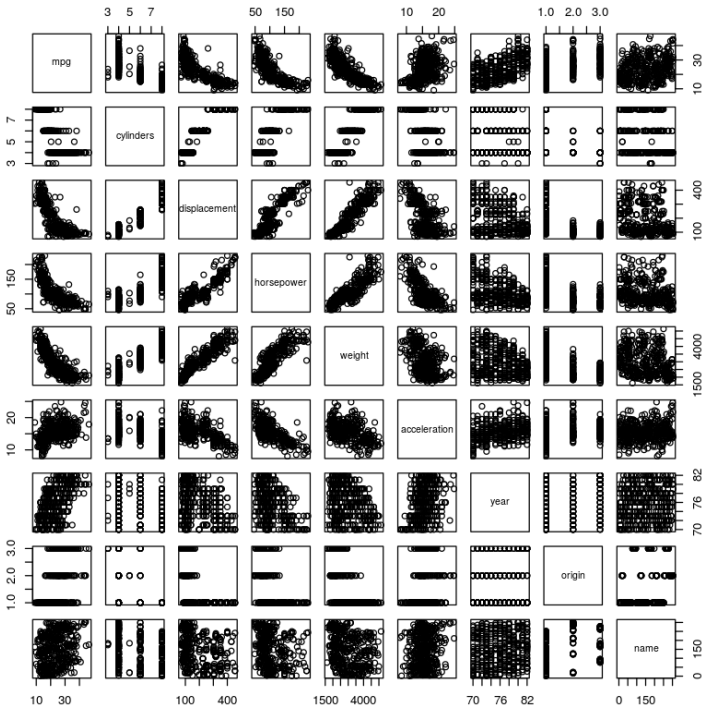
392 9

'mpg' 'cylinders' 'displacement' 'horsepower' 'weight' 'acceleration' 'year' 'origin' 'name'

```
      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
                    name
Min.   : 8.00   Min.   :70.00   Min.   :1.000
  amc matador       :  5
1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
  ford pinto        :  5
Median :15.50   Median :76.00   Median :1.000
  toyota corolla    :  5
Mean   :15.54   Mean   :75.98   Mean   :1.577
  amc gremlin       :  4
3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
  amc hornet        :  4
Max.   :24.80   Max.   :82.00   Max.   :3.000
  chevrolet chevette:  4

  (Other)           :365
```

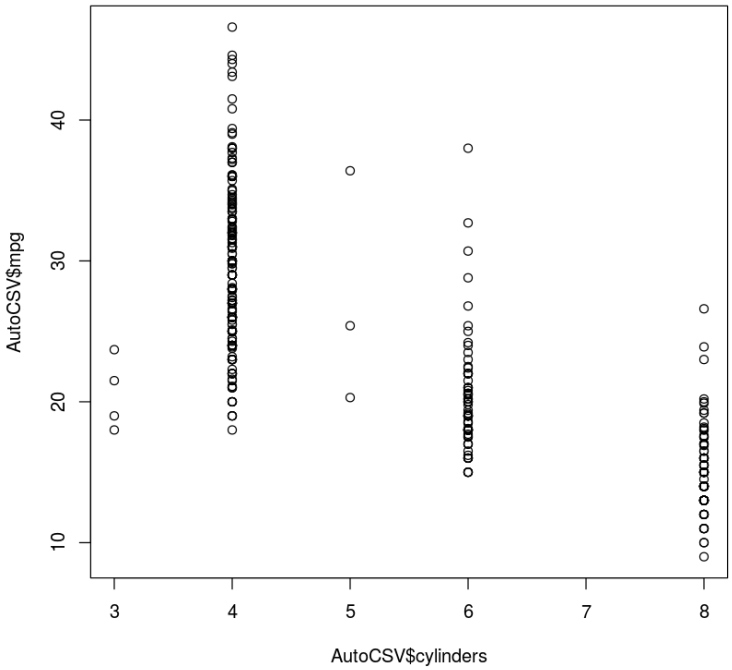
In [11]:
```
plot(AutoData)
# pairs(AutoData)
```



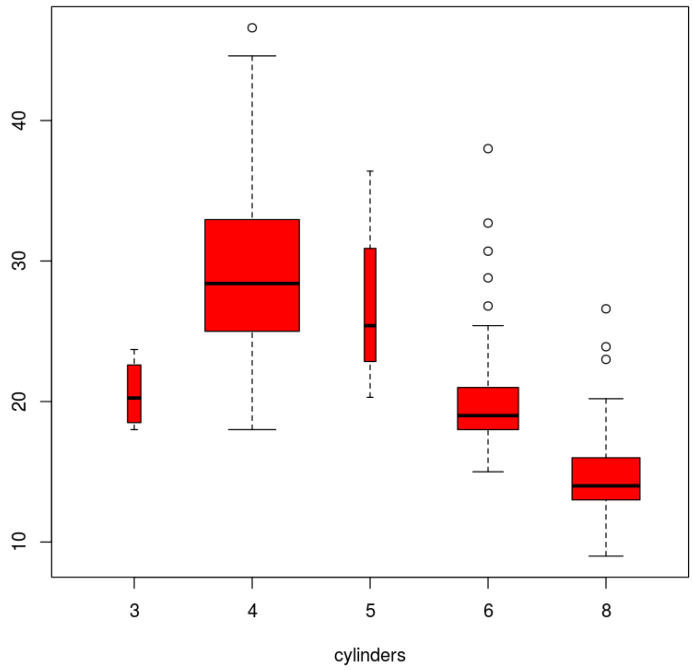
In [12]:
```
AutoCSV = read.csv("chapter-02-r-Auto.csv", hea
der=TRUE, na.strings="?")
AutoCSV = na.omit(AutoCSV)
dim(AutoCSV) # 392 9
names(AutoCSV) # 'mpg' 'cylinders' 'displacemen
t' 'horsepower' 'weight' 'acceleration' 'year'
 'origin' 'name'

plot(AutoCSV$cylinders, AutoCSV$mpg)
# attach(AutoCSV)
# plot(cylinders, mpg)
```

```
392  9
```

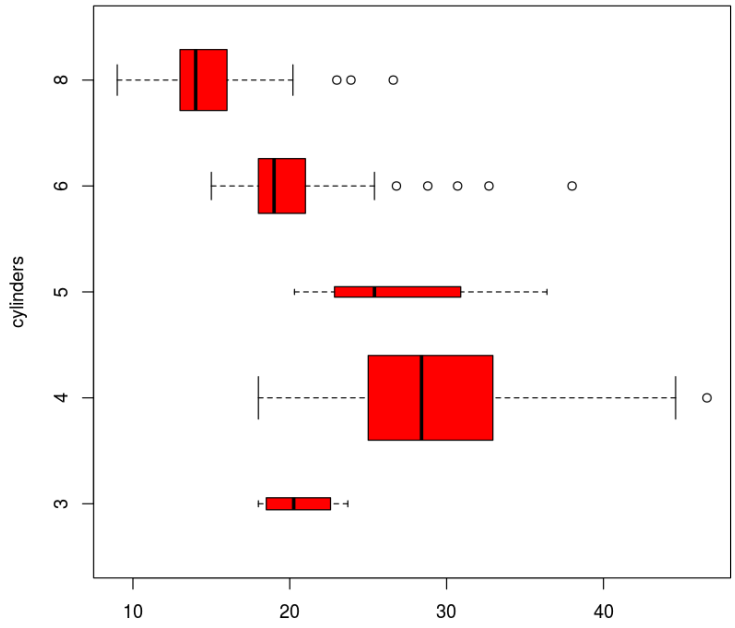
```
'mpg'  'cylinders'  'displacement'  'horsepower'
'weight'  'acceleration'  'year'  'origin'  'name'
```



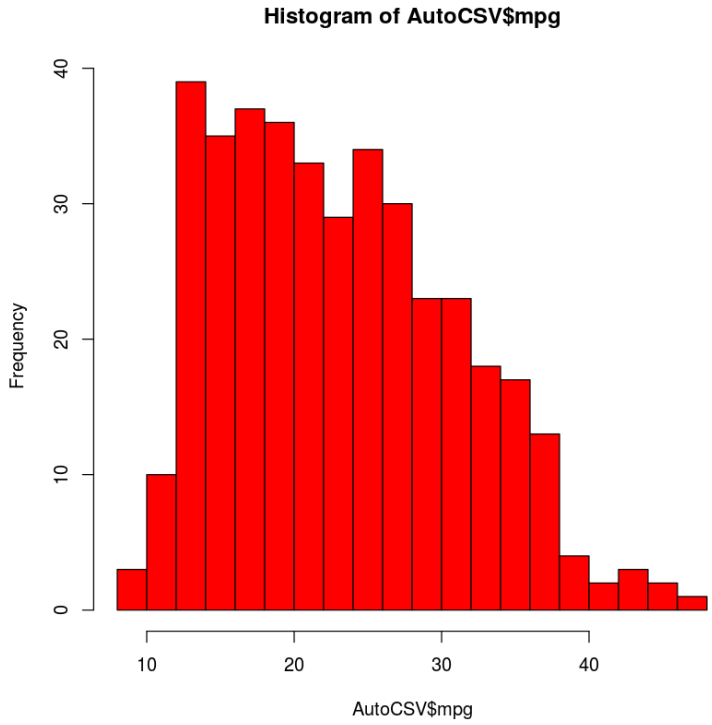
In [13]:
```
# to convert quantitative to qualitative variables
cylindersQualitative = as.factor(AutoCSV$cylinders)
plot(cylindersQualitative, AutoCSV$mpg, col="red", xlab="cylinders", varwidth=T)
```



In [14]:
```
plot(cylindersQualitative, AutoCSV$mpg, col="red", ylab="cylinders", varwidth=T, horizontal=TRUE)
```

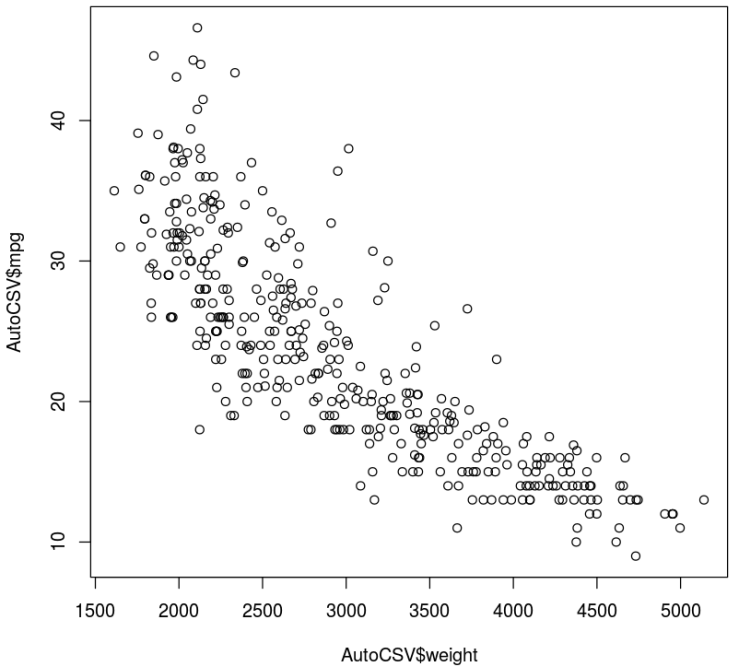


In [15]:
```
hist(AutoCSV$mpg, col=2, breaks=15)
```



In [16]:
```
plot(AutoCSV$weight, AutoCSV$mpg)
# identify(AutoCSV$weight, AutoCSV$mpg, AutoCSV
```

```
$year)
```


AutoCSV$mpg
10
20
30
40
1500
2000
2500
3000
3500
4000
4500
5000
AutoCSV$weight