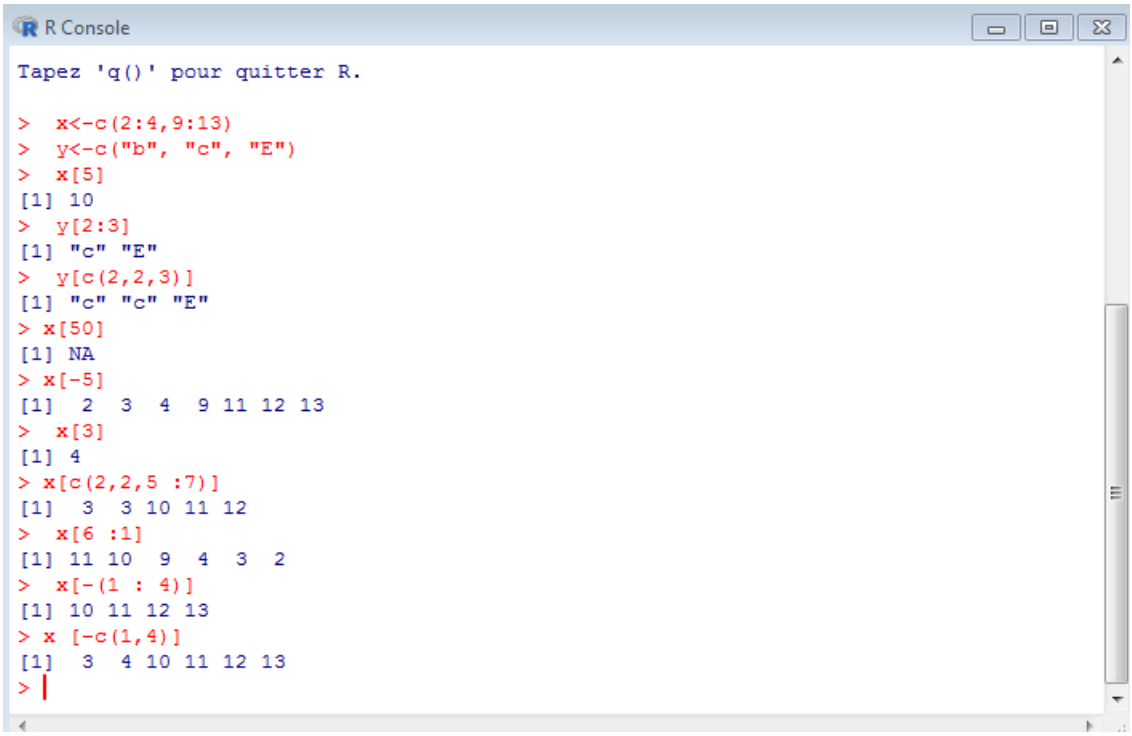
TP 1 : R for beginners

YAO ZELIANG & ZHAO HE

Exercise 1

Run the following code and interpret what is displayed:

```
> x<-c(2:4,9:13)
> y<-c("b", "c", "E")
> x[5]
> y[2:3]
> y[c(2,2,3)]
> x[50]
> x[-5]
> x[3]
> x[c(2,2,5 :7)]
> x[6 :1]
> x[-(1 : 4)]
> x [-c(1,4)]
```



Explanation：

```
> x<-c(2:4,9:13)
> y<-c("b", "c", "E")
#Create two vectors, one is x from 2 to 4 and 9 to 13, another is  “b”, “c” and “E”
> x[5]
#Select the fifth element in vector x
> y[2:3]
#Select the second and third elements in y
> y[c(2,2,3)]
# Select the second, repeat the second and the third elements in y
> x[50]
# Select the fifty element in x, which is not exist so the result is NA
> x[-5]
 # Select the elements except the fifth element in x
> x[3]
#Select the third element in x
> x[c(2,2,5 :7)]
# Select the second, the second and the fifth to seventh elements in x
> x[6 :1]
#Select the sixth to the first elements in x, with the decreasing order
> x[-(1 : 4)]
# Select the elements except the elements from first to the fourth in x
 > x [-c(1,4)]
 # to select the elements except the first and the fourth elements in x
```

Exercise 2

Write a function to compute the variance of a vector.

```
> variance<-function(x){
+ x2<-c()
+ meanx<-mean(x)
+ for(i in 1:length(x)){
+ xn<-x[i]-meanx
+ x2[i]<-xn^2}
+ sum2<-sum (x2)
+ variance1<-(sum2/(length(x)-1))
+ return(variance1)}
> x<-c(1:3,6:10)
> variance(x)
[1] 11.35714
> var(x)
[1] 11.35714
>
```

Explanation：

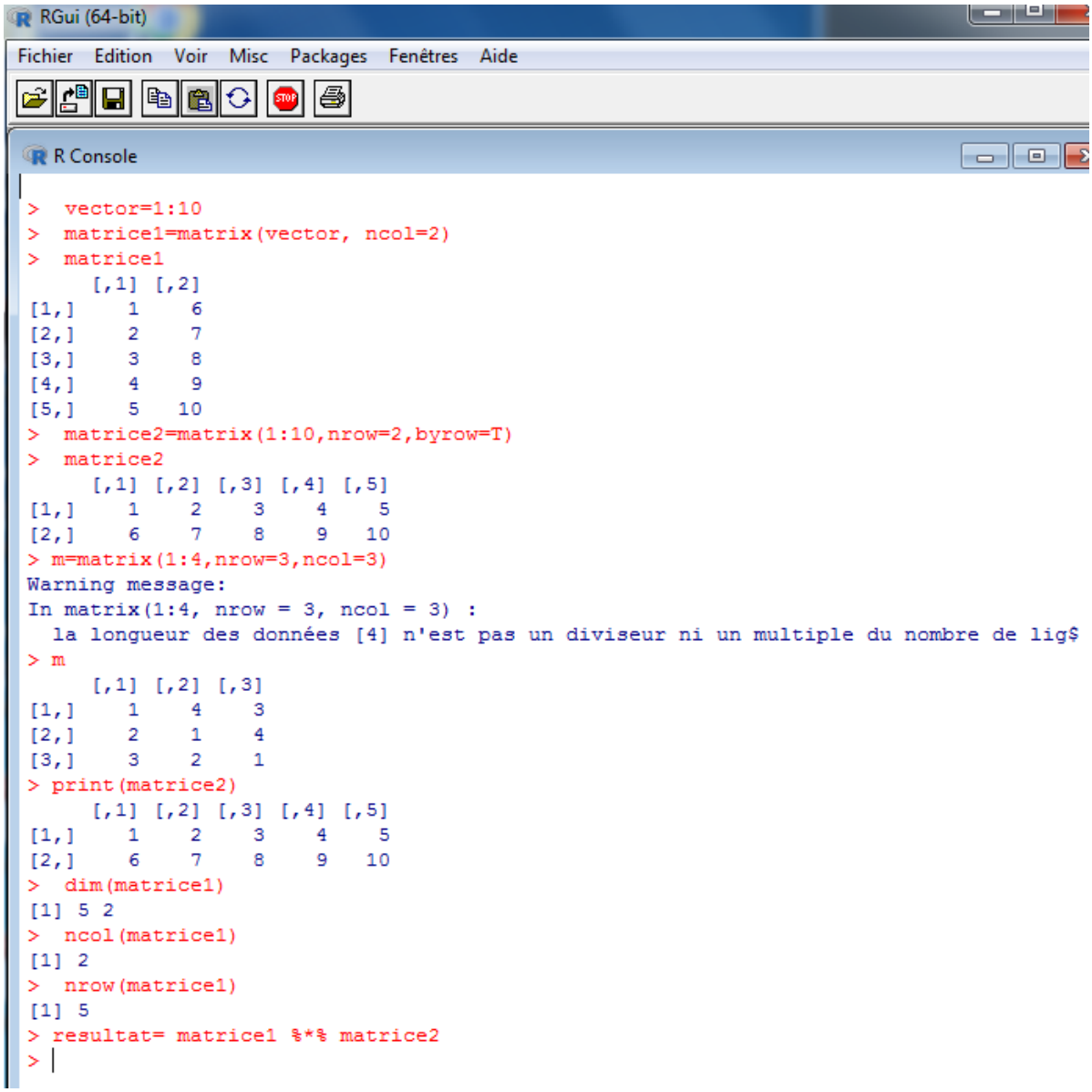
We can use the math formula to deduce the method

According to var(x)=$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$,

Exercise 3

Run the following code and interpret what is displayed:

```
> vector=1:10
> matrice1=matrix(vector, ncol=2)
> matrice1
> matrice2=matrix(1:10,nrow=2,byrow=T)
> matrice2
> m=matrix(1:4,nrow=3,ncol=3)
> m
> print(matrice2)
> dim(matrice1)
> ncol(matrice1)
> nrow(matrice1)
> resultat= matrice1 %*% matrice2
```



Explanation：

```
> vector=1:10
#Create a vector, the elements is from 1 to 10
> matrice1=matrix(vector, ncol=2)
#Create a matrix matrice1, with the elements in the vector, the number of column is 2
> matrice1
#Display the matrice1
> matrice2=matrix(1:10,nrow=2,byrow=T)
#Create a matrix matrice2, the number from 1 to 10, the number of row is 2, and the elements are arranged by row.
> matrice2
#Display the matrice2
> m=matrix(1:4,nrow=3,ncol=3)
#Create a matrix m, with numbers from 1 to 4, the number of column is 3
> m
#Display  m
> print(matrice2)
> dim(matrice1)
#Display the dimensions of matrice1, who has five rows and 2 columns
> ncol(matrice1)
#To show the number of  columns of matrice1
> nrow(matrice1)
#To show the number of  rows of matrice1
> resultat= matrice1 %*% matrice2
#Calculate the result of matrice1 * matrice2
```

Exercise 4

Operations on matrices : Examples

- Les fonctions **dim(), ncol(), nrow()**show the dimensions of a matrix : > dim(matrice1)

>ncol(matrice1)

> nrow(matrice1)

- The product of two matrices is written with the operator **%*%**.

> result= matrice1 %*% matrice2

- The function **t()**

- transposes a matrix >t(matrice1)

- The function **diag()**allows to retriev the diagonal of a square matrix or to construct a diagonal matrix from a vector.

>diag(resultat)

>diag (c(3,2,4))

- The functions **rbind()** et **cbind()**allow to concatenate by row or column vectors or matrices.

>vecteur1=(8,3,2)

>vecteur2=(23,6,9)

>res=rbind(vecteur1,vecteur2)

> vecteur3=c(2,4)

> cbind(res,vecteur3)

- To diagonalise a square matrix, we use the **eigen ()** function that returns a list of two components: **$ values** which contains the eigenvalues, **$ vectors** containing the eigen vectors.

> eigen(resultat)

R Console

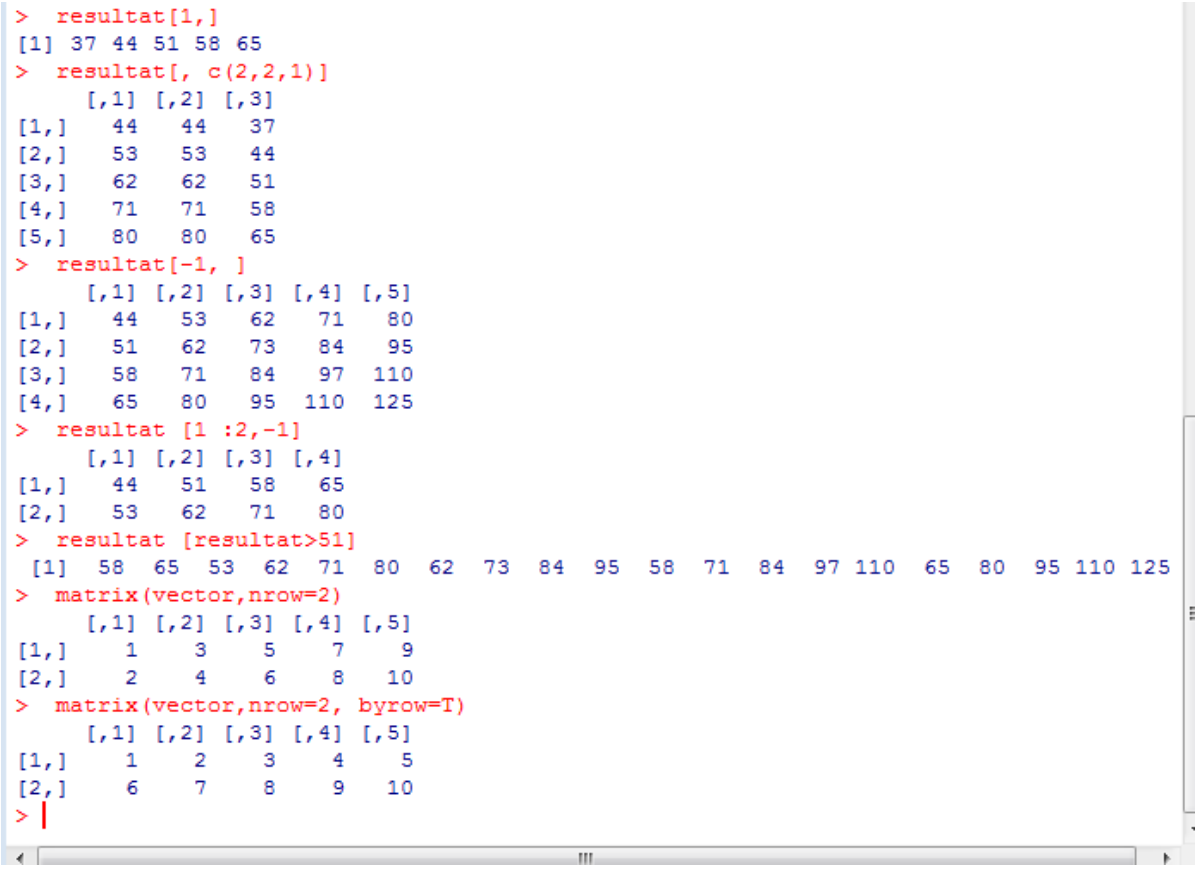
```
>  dim(matrice1)
[1] 5 2
>  ncol(matrice1)
[1] 2
>  nrow(matrice1)
[1] 5
> resultat= matrice1 %*% matrice2
>
> dim(matrice1)
[1] 5 2
> ncol(matrice1)
[1] 2
>  nrow(matrice1)
[1] 5
>  result= matrice1 %*% matrice2
>  t (matrice1)
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    2    3    4    5
[2,]    6    7    8    9   10
> diag (resultat)
[1]  37  53  73  97 125
> diag(c(3,2,4))
     [,1] [,2] [,3]
[1,]    3    0    0
[2,]    0    2    0
[3,]    0    0    4
>  vecteur1= (8,3,2)
Erreur : ',' inattendu(e) in " vecteur1= (8,"
> vecteur2=(23,6,9)
Erreur : ',' inattendu(e) in "vecteur2=(23,"
> res=rbind(vecteur1,vecteur2)
Error in rbind(vecteur1, vecteur2) : objet 'vecteur1' introuvable
> vecteur3=c(2,4)
> cbind(res,vecteur3)
Error in cbind(res, vecteur3) : objet 'res' introuvable
>  eigen(resultat)
eigen() decomposition
$values
[1]  3.817254e+02  3.274605e+00  5.901666e-15 -4.680421e-15 -4.741161e-14

$vectors
           [,1]        [,2]       [,3]       [,4]          [,5]
[1,] -0.3042621 -0.71233741  0.0000000  0.0000000  6.324555e-01
[2,] -0.3707409 -0.40317635 -0.1964879  0.5112656 -6.324555e-01
[3,] -0.4372198 -0.09401529  0.6430588 -0.5352340 -3.162278e-01
[4,] -0.5036986  0.21514578 -0.6966538 -0.4633288  2.331468e-15
[5,] -0.5701775  0.52430684  0.2500829  0.4872972  3.162278e-01

>
```

Exercise 5
Run the following code and interpret what is displayed:

> resultat[1,]
> resultat[, c(2,2,1)]
> resultat[-1,]
> resultat [1 :2,-1]
> resultat [resultat>51]
> matrix(vector,nrow=2)
> matrix(vector,nrow=2, byrow=T)



Explanation：

> resultat[1,]
#Select elements from the first row
> resultat[, c(2,2,1)]
#Select elements from the second, the second and the first columns
> resultat[-1,]
#Select the elements from all the rows except the first row
> resultat [1 :2,-1]
#Select the elements from the first and the second rows, but do not have the elements from the first column
> resultat [resultat>51]
#Select the elements which greater than 51
> matrix(vector,nrow=2)
#Display the matrix of which content is the elements in the vector, and the number of row is 2

> matrix(vector,nrow=2, byrow=T)
#Display the matrix whose content is the elements in the vector, and the number of row is 2, and the elements is arranged by row

Exercise 6

1) Create the following matrix Y:

$$Y = \begin{pmatrix} 1 & 2 & 3 & 5 \\ 10 & 12 & 13 & 22 \\ 5 & 9 & 8 & 34 \\ 7 & 1 & 4 & 3 \end{pmatrix}$$

2) View the item of Y contained in:
- The third row and second column
- The second line of Y
- The fourth column of Y
- The matrix obtained after removal of the first row and the second column

- **Response 1.** Create the following matrix Y

```
> matrix(c(1,2,3,5,10,12,13,22,5,9,8,34,7,1,4,3),nc=4, nr = 4, byrow = T)
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    5
[2,]   10   12   13   22
[3,]    5    9    8   34
[4,]    7    1    4    3
>
```

2.1 The third row and second column

```
> M<-matrix(c(1,2,3,5,10,12,13,22,5,9,8,34,7,1,4,3),nc=4, nr = 4, byrow = T)
> print(M[3,2])
[1] 9
>
```

2.2 The second line of Y

```
> print(M[2,])
[1] 10 12 13 22
>
```

2.3 The fourth column of Y

```
> print(M[,4])
[1]  5 22 34  3
>
```

2.4 The matrix obtained after removal of the first row and the second column

```
> M<-M[-1,-2]
> M
     [,1] [,2] [,3]
[1,]   10   13   22
[2,]    5    8   34
[3,]    7    4    3
>
```

Exercise 7

Write a function which allows to read data in a file, followed by a graph.

```
> library(readxl)
Warning message:
程辑包'readxl'是用R版本3.4.2 来建造的
> read_excle("d:/datamining/price.xlsx")
Error in read_excle("d:/datamining/price.xlsx") :
  没有"read_excle"这个函数
> read_excel("d:/datamining/price.xlsx")
# A tibble: 1,048,575 x 6
                    `PRODUITS (产品信息)` `Conditionnement （规格）`
                                    <chr>                      <chr>
 1                                     <NA>                       <NA>
 2        FILORGA NCTF 135HA    菲洛嘉135ha                  5 * 3ml
 3      FILORGA M-HA 10         菲洛嘉mha10        3 flacons  (3小瓶)
 4      FILORGA M-HA 18         菲洛嘉mha18                  2 * 1ml
 5      FILORGA NCTF  135         菲洛嘉135                  5 * 3ml
 6          FILORGA X-HA Volume   菲洛嘉xha                  2 * 1ml
 7    FILORGA X-HA 3             菲洛嘉xha3                 2 * 1ml
 8                                     <NA>                       <NA>
 9 Bocouture                 德国西马肉毒素                   50  U
10                                     <NA>                       <NA>
# ... with 1,048,565 more rows, and 4 more variables: `PRIX (价钱)
#   /欧` <dbl>, `邮费（>=2）/ 1盒` <dbl>, 合计 <dbl>, `人民币/7.8` <lgl>
>
```

we implement the library readxl to read the excel file for example if we want to read other types of file, we can use scan() read.table or ODBC ..etc

Exercise 8

1) Load the *"airquality"* data.

```
> airquality
    Ozone Solar.R Wind Temp Month Day
1      41     190  7.4   67     5   1
2      36     118  8.0   72     5   2
3      12     149 12.6   74     5   3
4      18     313 11.5   62     5   4
5      NA      NA 14.3   56     5   5
6      28      NA 14.9   66     5   6
7      23     299  8.6   65     5   7
8      19      99 13.8   59     5   8
9       8      19 20.1   61     5   9
10     NA     194  8.6   69     5  10
11      7      NA  6.9   74     5  11
12     16     256  9.7   69     5  12
13     11     290  9.2   66     5  13
14     14     274 10.9   68     5  14
15     18      65 13.2   58     5  15
16     14     334 11.5   64     5  16
17     34     307 12.0   66     5  17
18      6      78 18.4   57     5  18
19     30     322 11.5   68     5  19
20     11      44  9.7   62     5  20
21      1       8  9.7   59     5  21
22     11     320 16.6   73     5  22
23      4      25  9.7   61     5  23
24     32      92 12.0   61     5  24
25     NA      66 16.6   57     5  25
26     NA     266 14.9   58     5  26
27     NA      NA  8.0   57     5  27
28     23      13 12.0   67     5  28
29     45     252 14.9   81     5  29
30    115     223  5.7   79     5  30
31     37     279  7.4   76     5  31
32     NA     286  8.6   78     6   1
33     NA     287  9.7   74     6   2
34     NA     242 16.1   67     6   3
35     NA     186  9.2   84     6   4
36     NA     220  8.6   85     6   5
37     NA     264 14.3   79     6   6
38     29     127  9.7   82     6   7
39     NA     273  6.9   87     6   8
40     71     291 13.8   90     6   9
41     39     323 11.5   87     6  10
42     NA     259 10.9   93     6  11
43     NA     250  9.2   92     6  12
44     23     148  8.0   82     6  13
45     NA     332 13.8   80     6  14
```

2) Explain the six variables.

Ozone = ozone indicator
Solar.r = solar radiation indicator
Wind = wind speed
Temp = temperature In °F
Month = month number (5 = may)
Day = day number of the month

3) Calculate the main statistics of the database using "*summary"* function.

```
> summary(airquality)
     Ozone           Solar.R          Wind             Temp           Month
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.000
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.000
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00   Median :7.000
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean   :6.993
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00   3rd Qu.:8.000
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.   :9.000
 NA's   :37       NA's   :7
      Day
 Min.   : 1.0
 1st Qu.: 8.0
 Median :16.0
 Mean   :15.8
 3rd Qu.:23.0
 Max.   :31.0
```

4) Calculate separately the mean, median, and the standard deviation of the Temp variable using the appropriate commands.

```
> mean(airquality$Temp)
[1] 77.88235
> median(airquality$Temp)
[1] 79
> sd(airquality$Temp)
[1] 9.46527
```

5) Calculate the variance and write a function to compute the standard deviation.

```
> var(airquality$Temp)
[1] 89.59133
>
```

6) Extract:

a) the second line

```
> airquality[2,]
  Ozone Solar.R Wind Temp Month Day
2    36     118    8   72     5   2
```

b) the third column

```
$airquality[,3]
$[1]  7.4  8.0 12.6 11.5 14.3 14.9  8.6 13.8 20.1  8.6  6.9  9.7  9.2 10.9 13.2
$16] 11.5 12.0 18.4 11.5  9.7  9.7 16.6  9.7 12.0 16.6 14.9  8.0 12.0 14.9  5.7
$31]  7.4  8.6  9.7 16.1  9.2  8.6 14.3  9.7  6.9 13.8 11.5 10.9  9.2  8.0 13.8
$46] 11.5 14.9 20.7  9.2 11.5 10.3  6.3  1.7  4.6  6.3  8.0  8.0 10.3 11.5 14.9
$61]  8.0  4.1  9.2  9.2 10.9  4.6 10.9  5.1  6.3  5.7  7.4  8.6 14.3 14.9 14.9
$76] 14.3  6.9 10.3  6.3  5.1 11.5  6.9  9.7 11.5  8.6  8.0  8.6 12.0  7.4  7.4
$91]  7.4  9.2  6.9 13.8  7.4  6.9  7.4  4.6  4.0 10.3  8.0  8.6 11.5 11.5 11.5
$06]  9.7 11.5 10.3  6.3  7.4 10.9 10.3 15.5 14.3 12.6  9.7  3.4  8.0  5.7  9.7
$21]  2.3  6.3  6.3  6.9  5.1  2.8  4.6  7.4 15.5 10.9 10.3 10.9  9.7 14.9 15.5
$36]  6.3 10.9 11.5  6.9 13.8 10.3 10.3  8.0 12.6  9.2 10.3 10.3 16.6  6.9 13.2
$51] 14.3  8.0 11.5
$
```

c) lines 1, 2 and 4 with a single command c ()

```
> airquality[c(1,2,4),]
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
4    18     313 11.5   62     5   4
>
```

d) lines 2-6 with the command ':'

```
> airquality[2:6,]
  Ozone Solar.R Wind Temp Month Day
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
>
```

e) all except columns 1 and 2

```
> airquality[,c(-1,-2)]
    Wind Temp Month Day
1    7.4   67     5   1
2    8.0   72     5   2
3   12.6   74     5   3
4   11.5   62     5   4
5   14.3   56     5   5
6   14.9   66     5   6
7    8.6   65     5   7
8   13.8   59     5   8
9   20.1   61     5   9
10   8.6   69     5  10
11   6.9   74     5  11
12   9.7   69     5  12
13   9.2   66     5  13
14  10.9   68     5  14
15  13.2   58     5  15
16  11.5   64     5  16
17  12.0   66     5  17
18  18.4   57     5  18
```

f) all lines with the temperature greater than 90°.

```
> airquality[which(airquality$Temp>90),]
    Ozone Solar.R Wind Temp Month Day
42     NA     259 10.9   93     6  11
43     NA     250  9.2   92     6  12
69     97     267  6.3   92     7   8
70     97     272  5.7   92     7   9
75     NA     291 14.9   91     7  14
102    NA     222  8.6   92     8  10
120    76     203  9.7   97     8  28
121   118     225  2.3   94     8  29
122    84     237  6.3   96     8  30
123    85     188  6.3   94     8  31
124    96     167  6.9   91     9   1
125    78     197  5.1   92     9   2
126    73     183  2.8   93     9   3
127    91     189  4.6   93     9   4
>
```