Δομές Δεδομένων στην R

Βασικές δομές δεδομένων

vector

```
0.70 0.86 0.95 0.25 0.52 0.37 0.27 0.80 0.60 0.26
```

matrix

```
      [,1]   [,2]   [,3]   [,4]
[1,]  0.70   0.37   0.70   0.37
[2,]  0.86   0.27   0.86   0.27
[3,]  0.95   0.80   0.95   0.80
[4,]  0.25   0.60   0.25   0.60
[5,]  0.52   0.26   0.52   0.26
```

data frame

```
   Sepal.Length   Sepal.Width   Petal.Width   Species
1           5.1           3.5           0.2    setosa
2           4.9           3.0           0.2    setosa
3           4.7           3.2           0.2    setosa
4           4.6           3.1           0.2    setosa
5           5.0           3.6           0.2    setosa
6           5.4           3.9           0.4    setosa
7           4.6           3.4           0.3    setosa
8           5.0           3.4           0.2    setosa
9           4.4           2.9           0.2    setosa
10          4.9           3.1           0.1    setosa
```

list

```
$item1
[1] 1 2 3

$item2
[1] "a" "b" "c" "d" "e"

$item3
[1]  TRUE FALSE  TRUE  TRUE

$item4
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
```

VECTORS (διανύσματα)

```
[1] 0.67149785 0.47398715 0.32813279 0.87295142 0.56274062 0.16796701 0.05765868 0.59618446
[9] 0.94417744 0.83129550 0.38959025 0.99178460
```

Ιδιότητες

- μία διάσταση
- περιέχει δεδομένα του ίδιου τύπου

```
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m"
[14] "n" "o" "p" "q" "r" "s" "t" "u" "v" "w" "x" "y" "z"
```

```
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17
[18] 18
```

```
 [1]  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
 [9]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
```

Δημιουργία διανύσματος

- Ο πιο γνωστός τρόπος δημιουργίας διανυσμάτων είναι με τη συνάρτηση c() ή :
- για αριθμητικά διανύσματα υπάρχουν πολλοί τρόποι να δημιουργήσουμε ιεραρχημένες τιμές

```
# διανύσματα χωρίς ιεραρχημένες τιμές
c("Learning", "to", "create", "character", “vectors")
c(3, 2, 10, 55)
c(TRUE, FALSE, FALSE, FALSE, TRUE)

# αριθμητικά διανύσματα με ιεραρχημένες τιμές
6:15
15.5:-6.75
```

Πρόσβαση στα στοιχεία διανύσματος

vector[element]

```
 # δημιουργήστε αυτό το διάνυσμα
 v1 <- 1:10


# Δοκιμάστε τα παρακάτω
 v1[4]
 v1[4:7]
 v1[c(4, 3, 4)]
 v1[v1 > 6]
 v1[v1 > 8 | v1 <=3]
```

Η συνάρτηση summary()

Δείτε στατιστικά των αριθμητικών διανυσμάτων με τη συνάρτηση `summary()`:

```
length(v1)
summary(v1)
mean(v1)
median(v1)
v1 > 5
sum(v1 > 5)
```

Η σειρά σας!

1. δείτε τα περιεχόμενα του built-in διανύσματος χαρακτήρων state.name που περιλαμβάνει τις πολιτείες των ΗΠΑ

2. πόσα στοιχεία έχει το διάνυσμα;

3. ζητήστε το υποσύνολο του state.name με τα στοιχεία 35, 38, 14, 17. Σε ποιες πολιτείες των ΗΠΑ αντιστοιχεί;

Λύση

```
# δείτε τα περιεχόμενα του state.name
state.name
```

```
# πόσα στοιχεία έχει το state.name
length(state.name)
[1] 50
```

```
# δείτε το υποσύνολο του state name με τα στοιχεία 35, 17, 14, και 38
state.name[c(35, 38, 14, 17)]
[1] "OH" "PA" "IN" "KY"
```

MATRICES (πίνακες)

```
      [,1] [,2] [,3] [,4] [,5] [,6]
 [1,] 0.34 0.96 0.36 0.95 0.50 0.98
 [2,] 0.47 0.25 0.68 0.65 0.37 0.53
 [3,] 0.35 0.93 0.60 0.65 0.14 0.71
 [4,] 0.89 0.68 0.07 0.10 0.46 0.20
 [5,] 0.28 0.25 0.70 0.36 0.59 0.26
 [6,] 0.96 0.42 0.93 0.62 0.24 0.82
 [7,] 0.72 0.13 0.47 0.93 0.05 0.23
 [8,] 0.82 0.32 0.70 0.84 0.66 0.70
 [9,] 0.68 0.04 0.06 0.82 0.78 0.84
[10,] 0.13 0.14 0.46 0.91 0.29 0.82
[11,] 0.45 0.29 0.04 0.12 0.92 0.57
[12,] 0.90 0.81 0.74 0.83 0.91 0.29
[13,] 0.89 0.40 0.71 0.12 0.73 0.08
[14,] 0.05 0.52 0.47 0.53 0.53 0.96
[15,] 0.16 0.59 0.43 0.19 0.37 0.54
```

Ιδιότητες

- δύο διαστάσεις
  - γραμμές
  - στήλες
- περιέχει δεδομένα του ίδιου τύπου
- όλες οι στήλες έχουν τον ίδιο αριθμό στοιχείων

```
      [,1] [,2] [,3] [,4] [,5] [,6]
 [1,] 0.34 0.96 0.36 0.95 0.50 0.98
 [2,] 0.47 0.25 0.68 0.65 0.37 0.53
 [3,] 0.35 0.93 0.60 0.65 0.14 0.71
 [4,] 0.89 0.68 0.07 0.10 0.46 0.20
 [5,] 0.28 0.25 0.70 0.36 0.59 0.26
 [6,] 0.96 0.42 0.93 0.62 0.24 0.82
 [7,] 0.72 0.13 0.47 0.93 0.05 0.23
 [8,] 0.82 0.32 0.70 0.84 0.66 0.70
 [9,] 0.68 0.04 0.06 0.82 0.78 0.84
[10,] 0.13 0.14 0.46 0.91 0.29 0.82
[11,] 0.45 0.29 0.04 0.12 0.92 0.57
[12,] 0.90 0.81 0.74 0.83 0.91 0.29
[13,] 0.89 0.40 0.71 0.12 0.73 0.08
[14,] 0.05 0.52 0.47 0.53 0.53 0.96
[15,] 0.16 0.59 0.43 0.19 0.37 0.54
```

Δημιουργία πίνακα

```
set.seed(123)
v1 <- sample(1:10, 25, replace = TRUE)
m1 <- matrix(v1, nrow = 5)


m1
     [,1] [,2] [,3] [,4] [,5]
[1,]    3    1   10    9    9
[2,]    8    6    5    3    7
[3,]    5    9    7    1    7
[4,]    9    6    6    4   10
[5,]   10    5    2   10    7
```

Πρόσβαση στα στοιχεία πίνακα

matrix[row, col]

```
# εξαγωγή στοιχείου πίνακα
m1[1, 3]


# εξαγωγή όλων των στοιχείων για τις στήλες 1 έως 3
m1[, 1:3]


# εξαγωγή όλων  των στοιχείων για τις γραμμές 1 έως 3
m1[1:3, ]
```

Η συνάρτηση summary()

Δοκιμάστε τις παρακάτω εντολές

```
summary(m1)
mean(m1)
mean(m[1,])
rowMeans(m1)
colMeans(m1)
rowSums(m1)
colSums(m1)
m > .5
sum(m > .5)
which(m > .5)
m[m > .5]
```

Η σειρά σας!

1. Υπολογίστε τους μέσους όρους των στηλών του *built-in* πίνακα *`VADeaths`*

2. Υπολογίστε τους μέσους όρους των γραμμών του *built-in* πίνακα *`VADeaths`*

3. Ζητήστε τα στοιχεία του πίνακα VADeaths για τις γυναίκες ηλικίας 55-64.

Λύση

```
# υπολογισμός μέσων όρων στηλών
colMeans(VADeaths)
  Rural Male Rural Female   Urban Male Urban Female
       32.74         25.18         40.48         25.28

# υπολογισμός μέσων όρων γραμμών
rowMeans(VADeaths)
 50-54  55-59  60-64  65-69  70-74
11.050 16.925 25.875 40.400 60.350

# στοιχεία γυναικών ηλικίας 55-64
VADeaths[2:3, c(2, 4)]
      Rural Female Urban Female
55-59         11.7         13.6
60-64         20.3         19.3
```

DATA FRAMES

(πλαίσια δεδομένων)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4

Ιδιότητες

- δεδομένα τύπου φύλλου εργασίας
- δύο διαστάσεις
  - γραμμές
  - στήλες
- μπορεί να περιέχει δεδομένα διαφορετικού τύπου
- όλες οι στήλες έχουν τον ίδιο αριθμό στοιχείων

	year	month	day	dep_time	carrier	tailnum	dest	time_hour
1	2013	1	1	517	UA	N14228	IAH	2013-01-01 05:00:00
2	2013	1	1	533	UA	N24211	IAH	2013-01-01 05:00:00
3	2013	1	1	542	AA	N619AA	MIA	2013-01-01 05:00:00
4	2013	1	1	544	B6	N804JB	BQN	2013-01-01 05:00:00
5	2013	1	1	554	DL	N668DN	ATL	2013-01-01 06:00:00
6	2013	1	1	554	UA	N39463	ORD	2013-01-01 05:00:00
7	2013	1	1	555	B6	N516JB	FLL	2013-01-01 06:00:00
8	2013	1	1	557	EV	N829AS	IAD	2013-01-01 06:00:00
9	2013	1	1	557	B6	N593JB	MCO	2013-01-01 06:00:00
10	2013	1	1	558	AA	N3ALAA	ORD	2013-01-01 06:00:00
11	2013	1	1	558	B6	N793JB	PBI	2013-01-01 06:00:00
12	2013	1	1	558	B6	N657JB	TPA	2013-01-01 06:00:00
13	2013	1	1	558	UA	N29129	LAX	2013-01-01 06:00:00
14	2013	1	1	558	UA	N53441	SFO	2013-01-01 06:00:00
15	2013	1	1	559	AA	N3DUAA	DFW	2013-01-01 06:00:00
16	2013	1	1	559	B6	N708JB	BOS	2013-01-01 05:00:00
17	2013	1	1	559	UA	N76515	LAS	2013-01-01 06:00:00
18	2013	1	1	600	B6	N595JB	FLL	2013-01-01 06:00:00
19	2013	1	1	600	MQ	N542MQ	ATL	2013-01-01 06:00:00
20	2013	1	1	601	B6	N644JB	PBI	2013-01-01 06:00:00
21	2013	1	1	602	DL	N971DL	MSP	2013-01-01 06:00:00
22	2013	1	1	602	MQ	N730MQ	DTW	2013-01-01 06:00:00
23	2013	1	1	606	AA	N633AA	MIA	2013-01-01 06:00:00

Πρόσβαση στα στοιχεία πλαισίου δεδομένων

data.frame[row, col]

```
# στοιχεία της τέταρτης στήλης με αριθμό ή με όνομα
raw_data[, 4]
raw_data[, "Gender"]

# στοιχεία όλων των γραμμών για τις στήλες 1 έως 3
raw_data[, 1:3]
raw_data[, c(“CustomerID", "Region", "TownSize")]

# στοιχεία της πρώτης γραμμής για όλες τις στήλες
raw_data[1, ]
```

LISTS

(λίστες)

```
$item1
[1] 1 5 3 7

$item2
[1] "g" "b" "q" "v" "d" "z" "w" "i"

$item3
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

$item4
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Ιδιότητες

- μία διάσταση
- περιέχει δεδομένα διαφορετικού τύπου (π.χ. διανύσματα, πλαίσια δεδομένων, πίνακες, ακόμη και λίστες)

```
$item1
[1] 1 5 3 7

$item2
[1] "g" "b" "q" "v" "d" "z" "w" "i"

$item3
     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

$item4
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Οι λίστες είναι πολύ σημαντική δομή δεδομένων της *R!*

Λίστες & αποτελέσματα μοντελοποίησης

- Πολλά προβλεπτικά μοντέλα δίνουν αποτελέσματα σε μορφή λίστας
- Χρειάζεται να γνωρίζετε πως να εξάγετε τα στοιχεία μιας λίστας ώστε να έχετε πρόσβαση σε επιμέρους αποτελέσματα ενός μοντέλου

Αποτελέσματα μοντέλου

- τα αποτελέσματα ενός μοντέλου πρόβλεψης αποθηκεύονται σε λίστα

```
# ένα μοντέλο απλής γραμμικής παλινδρόμησης
model <- lm(mpg ~ wt, data = mtcars)

summary(model)
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
##
```

Αποτελέσματα μοντέλου

- τα αποτελέσματα ενός μοντέλου πρόβλεψης αποθηκεύονται σε λίστα

```
# ένα μοντέλο απλής γραμμικής παλινδρόμησης
model <- lm(mpg ~ wt, data = mtcars)

names(model)
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"

str(model)
## λίστα με 12 στοιχεία
##  $ coefficients : Named num [1:2] 37.29 -5.34
##   ..- attr(*, "names")= chr [1:2] "(Intercept)" "wt"
##  $ residuals    : Named num [1:32] -2.28 -0.92 -2.09 1.3 -0.2 ...
##   ..- attr(*, "names")= chr [1:32] "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Drive" ...
##  $ effects      : Named num [1:32] -113.65 -29.116 -1.661 1.631 0.111 ...
##   ..- attr(*, "names")= chr [1:32] "(Intercept)" "wt" "" "" ...
##  $ rank         : int 2
##  $ fitted.values: Named num [1:32] 23.3 21.9 24.9 20.1 18.9 ...
```

Πρόσβαση στα στοιχεία της λίστας

- Μπορούμε να έχουμε πρόσβαση στα στοιχεία μιας λίστας με τρεις τρόπους:

  preserve: `list[component]`

  simplify: `list[[component]]`

  simplify: `list$component`

```
# εκτελέστε τις παρακάτω εντολές
model[“residuals”]
model[[“residuals”]]
model$residuals
model[["residuals"]][1:20]
```

Τι διαφορές παρατηρείτε;

- τα αποτελέσματα ενός μοντέλου πρόβλεψης αποθηκεύονται σε λίστα
- αν θέλουμε να εξάγουμε τα κατάλοιπα του μοντέλου απλά δίνουμε:

```
# δίνει τα κατάλοιπα ενός μοντέλου παλινδρόμησης
model$residuals
##           Mazda RX4       Mazda RX4 Wag          Datsun 710
##          -2.2826106          -0.9197704          -2.0859521
##      Hornet 4 Drive   Hornet Sportabout             Valiant
##           1.2973499          -0.2001440          -0.6932545
##          Duster 360           Merc 240D            Merc 230
##          -3.9053627           4.1637381           2.3499593
##            Merc 280           Merc 280C          Merc 450SE
##           0.2998560          -1.1001440           0.8668731
##          Merc 450SL         Merc 450SLC  Cadillac Fleetwood
##          -0.0502472          -1.8830236           1.1733496
## Lincoln Continental   Chrysler Imperial            Fiat 128
##           2.1032876           5.9810744           6.8727113
```

Συναρτήσεις που χρειάζεται να θυμάστε

Συνάρτηση	περιγραφή
`read_csv, excel_sheets, read_excel`	εισαγωγή δεδομένων
`data.frame, vector, matrix, list, c(), :`	δημιουργία διανυσμάτων, πινάκων, πλαισίων δεδομένων, λιστών
`str, names, colnames, rownames, dim, length, nrow, ncol`	ιδιότητες δομών δεδομένων
`summary, mean, median, sum, colSums, rowSums, colMeans, rowMeans`	σύνοψη τιμών - στατιστικοί δείκτες
`[], [[]], $`	πρόσβαση στις τιμές δομών δεδομένων