数据分析与处理技术——3 高级数据结构

商学院　徐宁

参考资料

- 《R语言–使用数据分析与可视化技术》第4、5章
- 《R语言教程》（在线版）11-13，链接如下：

12 R矩阵和数组 | R语言教程 (pku.edu.cn)
11 列表类型 | R语言教程 (pku.edu.cn)
13 数据框 | R语言教程 (pku.edu.cn)

混合类型数据的问题

- 本章学习目标：理解R语言如何装载混合类型的数据，学会处理较大的数据集。
- 理解变量的两大类：单模式变量、多模式变量

高级数据结构

1.矩阵与数组

矩阵结构

矩阵基本操作

矩阵运算

数组变量

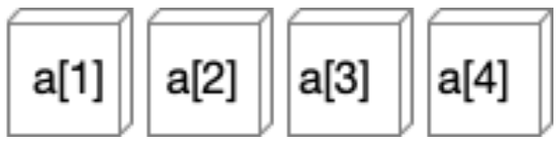
矩阵变量

矩阵(matrix)的特点：

- 矩阵是原子向量的拓展
- 强制单模特性
- 元素有行和列两个索引属性
- 与数学上的矩阵规则一致

单变量

向量



矩阵

注意观察矩阵中的索引如何排列

矩阵操作参考资料：
https://www.math.pku.edu.cn/teachers/lidf/docs/Rbook/html/_Rbook/prog-type-matrix.html

矩阵的结构属性

维度属性：

- dim()函数
- nrow()
- ncol()

```
> y
[1]  1  2  3  4  9  8  7  7 -2 23
> length(y)
[1] 10
```

```
> m=matrix(y,nrow=2,byrow = T)
> m
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    2    3    4    9
[2,]    8    7    7   -2   23
> nrow(m)
[1] 2
> ncol(m)
[1] 5
> dim(m)
[1] 2 5
```

length()对矩阵变量是否有效？

创建矩阵

matrix()函数

- 以原子向量为基础
- 至少一个维度属性

```
> m=matrix(c(1,2,5,7,8,3),nrow=2)
> m
     [,1] [,2] [,3]
[1,]    1    5    8
[2,]    2    7    3
```

```
> m=matrix(c(1,2,5,7,8,3),ncol=3,byrow = T)
> m
     [,1] [,2] [,3]
[1,]    1    2    5
[2,]    7    8    3
```

矩阵索引

- 矩阵的维度属性
- 一维矩阵不等于原子向量
- 矩阵变量同时兼容原子向量索引和矩阵索引

```
> m=matrix(c(1,2,5,7),ncol=2)
> m
     [,1] [,2]
[1,]    1    5
[2,]    2    7
> m[1,2]
[1] 5
> m[1,]   #列属性的空缺代表全部选择
[1] 1 5
> m[,2]
[1] 5 7
```

对角元素操作

生成/操作单位对角矩阵：

- diag()函数
- 直接操作对角元素

对角矩阵操作

- lower.tri()操作下三角矩阵
- upper.tri()上三角矩阵

```
> m=diag(3)
> m
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
> diag(m)=c(3,2,1) #修改对角元素
```

```
> m=matrix(1:9,3)
> lower.tri(m,diag = T) #diag控制是否包含对角线元素
      [,1]  [,2]  [,3]
[1,] TRUE FALSE FALSE
[2,] TRUE  TRUE FALSE
[3,] TRUE  TRUE  TRUE
> m[lower.tri(m,diag = T)]=0
> m
```

矩阵拼接

- 在原有矩阵基础上拼接向量需要考虑行或列的因素。
- rbind函数即row bind，按行组合矩阵。
- 同理，cbind函数按列组合。

```
> m=matrix(1:4,2)
> m
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> a=c(5,6)
```

```
> rbind(m,a)
```

```
> cbind(m,a)
```

矩阵运算

矩阵继承了原子向量的向量化运算

* 向量化元素乘法

%*% 矩阵乘法运算

%o% 矩阵外积运算

```
> m=matrix(c(1,2,5,7),ncol=2)
> m
     [,1] [,2]
[1,]    1    5
[2,]    2    7
> n=matrix(1:4,2)
> m*n
> m %*% n
```

矩阵运算

- 转秩运算
- 行列式计算

```
> t(m)   #矩阵转秩
     [,1] [,2]
[1,]    1    2
[2,]    5    7
> det(m)   #行列式求值
[1] -3
```

```
> eigen(m) #计算特征值和特征向量
eigen() decomposition
$values
[1]  8.3589 -0.3589

$vectors
          [,1]     [,2]
[1,] -0.56200 -0.96500
[2,] -0.82714  0.26227
```

线性方程组

- 解线性方程

$$\begin{cases} x+5y=3 \\ 2x+7y=11 \end{cases}$$

- 求逆矩阵

```
> m
     [,1] [,2]
[1,]    1    5
[2,]    2    7
> b=c(3,11)
> solve(m,b)
[1] 11.3333 -1.6667
> solve(m)
           [,1]     [,2]
[1,] -2.33333  1.66667
[2,]  0.66667 -0.33333
```

数组

数组(array)的特点：

- 维数更高的数据结构
- 以原子向量为基础
- 单模性质

```
>a1=array(c(0,1,2,3,4,5,6,7,8,9,10,
11),dim = c(2,3,2))
> a1
, , 1

     [,1] [,2] [,3]
[1,]    0    2    4
[2,]    1    3    5


, , 2

     [,1] [,2] [,3]
[1,]    6    8   10
[2,]    7    9   11
```

索引规则

逗号分隔维度，

[行，列，页，……]

空缺代表全选

负号代表剔除

同维度正负不共存

```
> a1     #调用a1全部数据

> a1[1,1,2]     #第2页第1行第1列数据

> a1[1,,2]     #第2页第1行全部列

> a1[1,c(1,3),1]

> a1[1,-2,1] #第1页第1行除第2列外的数据
```

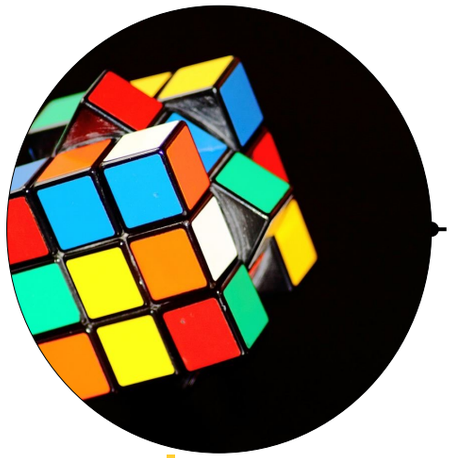
2.列表变量

list变量

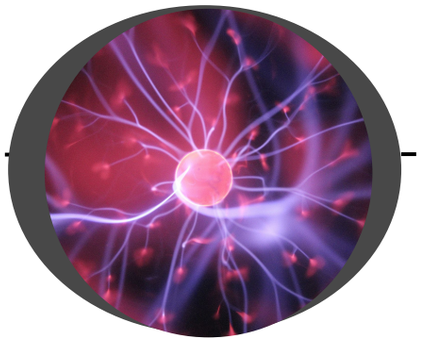
列表属性

列表索引

变量类型的拓展



单模式变量



多模式变量

方式：增加维度属性

矩阵 (matrix)

数组 (array)

原子向量

方式：嵌套叠加

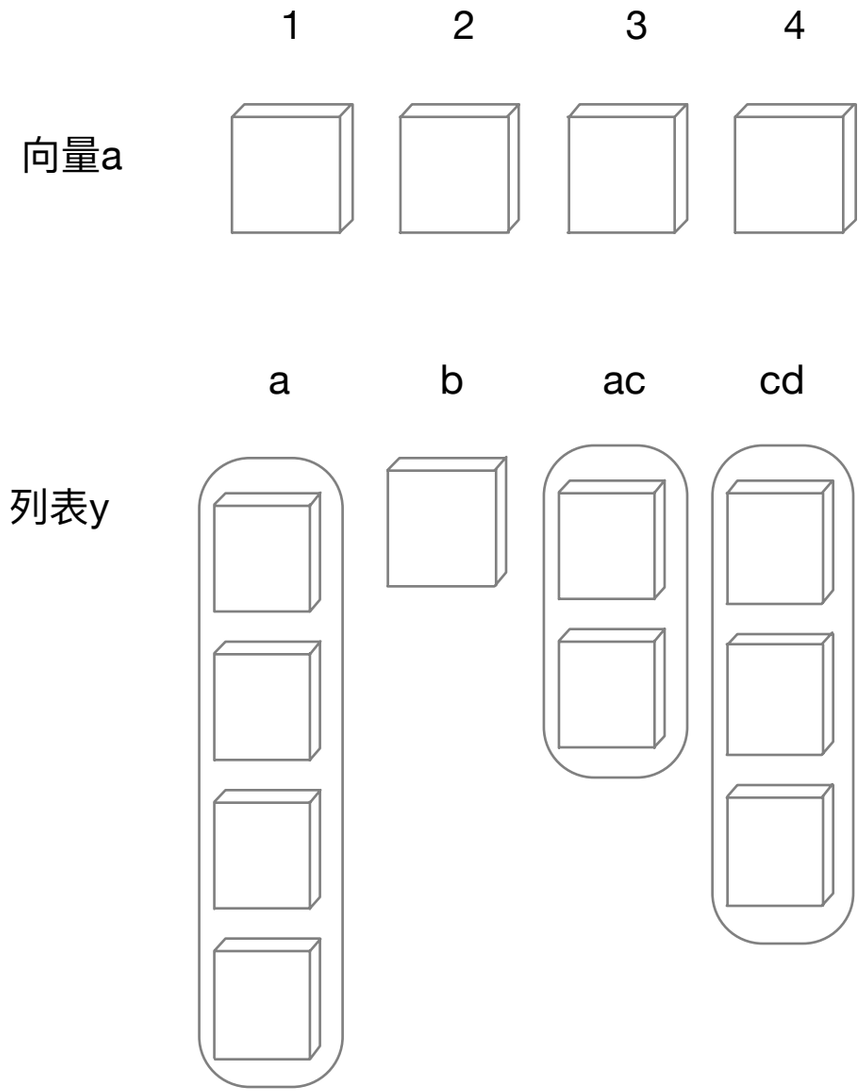
(data.frame)数据框

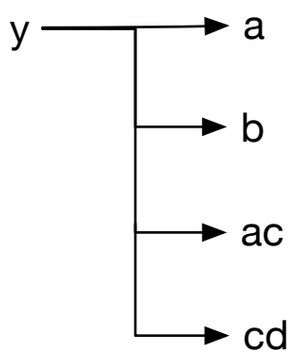
(list)列表变量

列表变量（List）

列表变量又称为递归向量

- 本质上是向量每一个元素拓展为一个变量
- 元素之间类型无限制，元素变量甚至可以是列表
- 类似于文件夹的树形结构变量





y,a,b,ac,cd均为自命名的变量名称

列表的创建

list()函数

- 元素变量类型不受限制
- 元素变量可以是列表
- 元素变量的层数不受限制

```
> mylist <-list(a=c(1,3,6),
b=letters[1:5])
> mylist
$a
[1] 1 3 6

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

```
> newlist <- list(t1=mylist,t2=1:4)
```

列表属性

- str()观察结构
- length()测元素数量
- names()调取元素名称

```
> str(newlist)
List of 2
 $ t1:List of 2
  ..$ a: num [1:3] 1 3 6
  ..$ b: chr [1:5] "a" "b" "c" "d" ...
 $ t2: int [1:4] 1 2 3 4
```

注意：这些操作属性的函数均继承自向量

列表索引

方式1:[[编号]] 取元素

方式2:$ 按名称取元素变量($是[[ ]]索引符号的另一种形式)

方式3:[‘元素名’] 取元素

```
> newlist[[2]]
[1] 1 2 3 4

> newlist$t2
[1] 1 2 3 4

> newlist['t2']
$t2
[1] 1 2 3 4
```

问题：newlist的第1个元素下的第2个元素怎么取

3.数据框变量

学生档案数据用什么工具处理

学生电子档案根据学生提交的表格输入到计算机中，并存储于相关数据库。调取部分学生档案如图所示：

问题：这种数据是矩阵吗？能够当矩阵处理吗？

Name	Gender	Age	Specialty
李雷	Male	20	Finance
韩梅梅	Female	19	Statistics
张萌	Female	21	Economics

data.frame

- 数据框(data.frame)基本特征：
  - 表格形状的数据结构
  - 行为对象，数据可以是异质型
  - 列为属性，列中数据同类

直接对接外部数据导入：
Excel、csv文本数据文件、
关系型数据库

属性　变量，元素

对象

Name	Gender	Age	Specialty
李雷	Male	20	Finance
韩梅梅	Female	19	Statistics
张萌	Female	21	Economics

创建数据框

- data.frame()函数创建数据框

- data.frame本质上是list类型，但具备矩阵形状。

```
> persons=data.frame( Name=c(“李雷","韩梅梅","张萌"), Gender = c(“Male”, “Female", "Female"), Age=c(20,19,21), Major= c( “Finance", "Statistics","Economics"))
>
> persons
     Name Gender Age       Major
1    李雷   Male  20     Finance
2 韩梅梅 Female  19 Statistics
3    张萌 Female  21   Economics
>
> class(persons)
[1] "data.frame"
```

数据框组织数据的原理

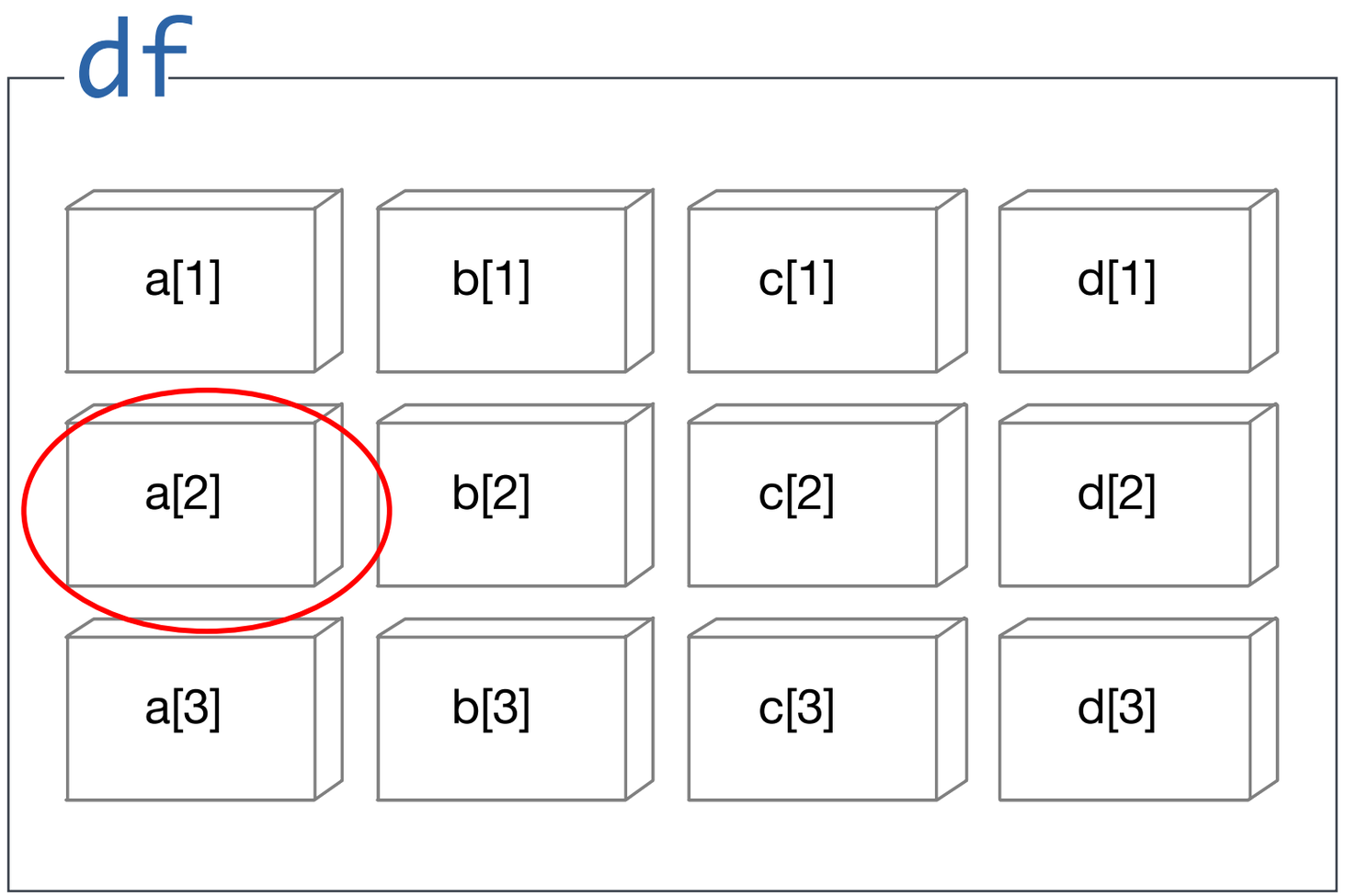
数据框有两套数据组织方式：

1.矩阵索引式
[ ]索引二维编号

2.列表元素式
$或[[ ]]取元素

```
df$a[2]

df[2,1]
```



元素的访问方式

继承列表操作

- 使用[[ ]]或$访问列元素

继承矩阵索引

- 使用[行，列]索引访问

```
> persons
     Name Gender Age       Major
1    李雷    Male  20     Finance
2 韩梅梅  Female  19 Statistics
3    张萌  Female  21   Economics
> persons$Name
[1] 李雷    韩梅梅  张萌
```

```
> persons[1,] #索引方式与矩阵相同
   Name Gender Age   Major
1 李雷    Male  20 Finance
> persons[,1]
[1] 李雷    韩梅梅  张萌
> persons[“Age"] #索引可以根据名称调取列变量
   Age
1   20
2   19
3   21
```

小练习

- 在R语言中使用索引访问数据

1. 取出第一行的姓名数据

2. 取出所有人姓名数据

3. 取出年龄数据

```
> persons
     Name     Gender    Age        Major
1      李雷       Male     20      Finance
2    韩梅梅     Female     19   Statistics
3      张萌     Female     21    Economics

> persons[1,1]
   李雷
> persons[,1]
[1] 李雷     韩梅梅  张萌
> persons$Age
   Age
1   20
2   19
3   21
```

添加行列

- cbind

列合并(column bind)

- rbind

行合并(row bind)

也称为附加(append)

```
> math=c(80,85,75)
> cbind(persons,math)
   Name Gender Age       Major math
1   李雷   Male  20     Finance   80
2 韩梅梅 Female  19 Statistics   85
3   张萌 Female  21   Economics   75

> new=data.frame(Name="张扬
",Gender="Male",Age=20,Major="Engneering
")
> rbind(persons,new)
   Name Gender Age       Major
1   李雷   Male  20     Finance
2 韩梅梅 Female  19 Statistics
3   张萌 Female  21   Economics
4   张扬   Male  20 Engneering
```

简便方法，新变量可以直接以赋值方式创建

融合两个数据框

merge函数通过两个表共有的列对数据进行匹配融合

by参数还有两个等效形态：by.x by.y，即以哪个数据集中的变量为主

```
> scores=data.frame(Name=c("张萌","韩梅梅","李雷"),
Computer=c(90,88,85),History= c(85, 82,77))
> scores
    Name Computer History
1   张萌       90      85
2 韩梅梅       88      82
3  李雷        85      77
> merge(persons,scores,by="Name")
    Name Gender Age      Major Computer History
1   张萌 Female  21  Economics       90      85
2   李雷   Male  20    Finance       85      77
3 韩梅梅 Female  19 Statistics       88      82
```

取子集

- subset

该函数按元素变量筛选行

同时利用select参数指定列

```
> persons #依然使用该数据集
     Name Gender Age       Major
1    李雷    Male  20     Finance
2 韩梅梅  Female  19 Statistics
3    张萌  Female  21   Economics
```

```
> subset(persons,Age>19,select=c(Name,Major))
   Name      Major
1 李雷    Finance
3 张萌  Economics
```

根据条件筛选对象(行数据),通过select参数选取指标

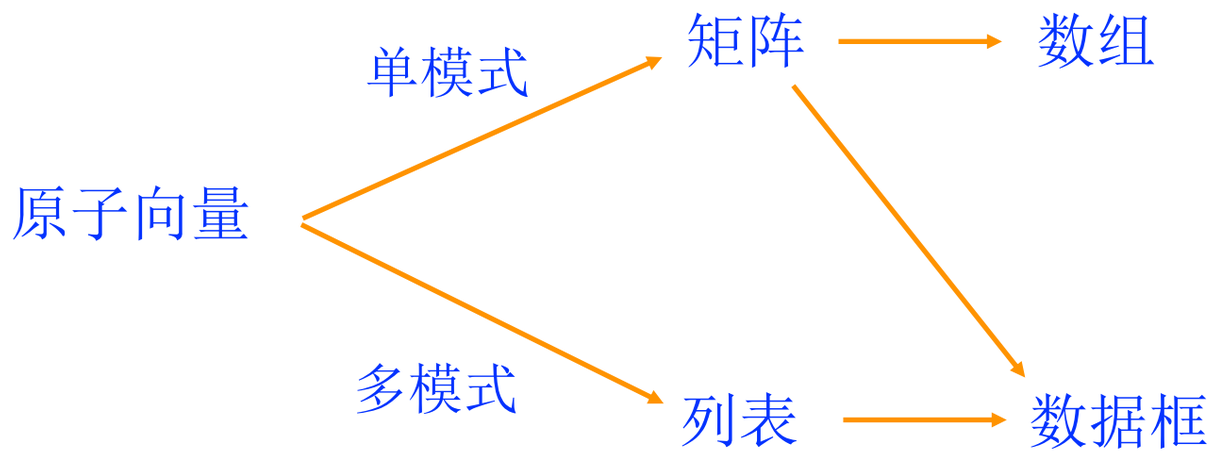
变量类型间关系

转换变量类型

4.变量间关系

变量间关系

- 原子向量是所有类型变量的基础和最小单位
- 原子向量操作方法也被继承到其他变量



- 形状相似的变量可以直接转换类型

单模数据结构特征

- 向量、矩阵、数组 只允许存储相同类型数据，这三类变量既是常用工具，也是其他类型变量构建的基础

属性	向量	矩阵	数组
模式	单模式	单模式	单模式
结构	无结构	二维	N维
索引结构	a[n]	a[n,m]	a[n1,n2,n3,…]

类型判断与转换

- 单模式变量中，变量类型与数据类型一致
- mode() 测数据模式
- class() 测变量类型
- as类函数
- is类函数

```
> m <- matrix(1:9,nrow=3)
> mode(m)
[1] "numeric"
> class(m)
[1] "matrix"

> is.matrix(m)
[1] TRUE
> is.vector(m)
[1] FALSE

> as.vector(m)
[1] 1 2 3 4 5 6 7 8 9
> as.numeric(m)
[1] 1 2 3 4 5 6 7 8 9
```

类型判断与转换

转换函数	判断函数	类型检验
as.list()	is.list( )	typeof( )
as.matrix( )	is.matrix( )	mode( )
as.data.frame( )	is.data.frame( )	class( )
as.factor( )	is.vector( )	……
as.charactor( )	is.charactor()	
as.integer( )	is.factor()	
as.numeric( )	……	

练习1

（1）创建矩阵：$M=\begin{bmatrix}11 & 13 & 15 & 17\\ 19 & 23 & 24 & 25\\ 26 & 27 & 28 & 29\\ 30 & 31 & 32 & 33\end{bmatrix}$，计算M的行列式和特征值；（2） 截取右下3X3矩阵记为A矩阵，求解方程式：

$$A\begin{bmatrix}x1\\ x2\\ x3\end{bmatrix}=\begin{bmatrix}27\\ 36\\ 55\end{bmatrix}$$

练习2

- 设$x = (1,3,5)^T, y = (2,4,6)^T$
  - 计算$z = 3x + y^2 + e$，其中$e = c(1,1,1)$
  - 计算x与y的内积和外积
- 计算1到125之和

练习3

- 生成两个矩阵，利用该矩阵计算并生成新的矩阵

$$M = \begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix}$$

$$N = \begin{bmatrix} -5 & 7 \\ 12 & 9 \end{bmatrix}$$

- X为M和N的相乘，Y为M和N矩阵的加和

- 将M N X Y四个矩阵拼接组合如下

$$\begin{bmatrix} M & N \\ X & Y \end{bmatrix}$$

练习4

创建适当的变量，记录零部件采购订单数据，其中2MT……为单位采购量的零部件规格名称缩写。

- 使用列表变量记录下表中的订单信息
- 调取订单279097，检验商品规格PTA是否在其订单当中
- 添加新订单350804,商品需求为AM2 PTA YZ 和BAH

订单编号	商品需求
302826	2MT BAH BAH AM2 YZ
302731	YZ PTA 2MT
279097	BAH AM2 YZ YZ YZ
330102	PTA AM2 AM2 YZ

练习5

- 利用数据框变量记录以下订单信息，各列选择合适的数据类型

订单编号	客户名	城市	交付时间
302826	嘉华	杭州	1/17/2018
302731	马士基	上海	2/12/2018
279097	博世	苏州	5/10/2018
330102	IBM	广州	11/22/2017
350804	SMC	北京	12/3/2017

- 公司售前部门通过订单系统汇总后形成各订单总金额，请将下表信息与上表合并

订单编号	订单金额
350804	1052
279097	732
302731	2000
302826	846
330102	1439

练习6

零售电商企业的信息系统中导出某阶段订单，完成下边的操作

数据集：3_1dingdan.Rda

- 数据集中有多少指标，调取指标名称
- 将订货日期转为日期时间类数据，计算最早和最晚订货时间之间相差几天
- 找到订购量最大的订单，调取该订单全部信息查看
- 计算本阶段订单商品的总重量
- 计算订单的总价值