

大数据第一次作业

[BY 张坤]

1. 找出“sunyang”，“dingning”，“fuyuanhui”，“zhangjike”，“wuminxia”，“duli”，“yeshiwen”中字符最多，最少和位于中间的名字并求所有名字长度的平均值

```
> name<-c("sunyang","dingning","fuyuanhui","zhangjike","wuminxia","duli","yeshiwen")
> get_mmmm = function(ns)
+ {
+   lens = nchar(ns);
+   name_min = min(lens);
+   name_max = max(lens);
+   name_median = median(lens);
+   name_mean = mean(lens);
+   re = c(name_min,name_max,name_median,name_mean);
+   return(re);
+ }
>
> result = get_mmmm(name);
> result
[1] 4.000000 9.000000 8.000000 7.571429
```

2. A 组有 3 人，他们语文，数学，英语的总成绩分别是 235,255,266。

B 组有 4 人，他们的语文，数学，英语的成绩分别为 300,349,377。

用向量的除法计算两个小组语文，数学，英语成绩的平均分

```
> #2
> total_scores_of_chinese=c(235,0,0,300,0,0,0)
> mean.chinese=mean(total_scores_of_chinese)
> mean.chinese
[1] 76.42857
> total_scores_of_math=c(255,0,0,349,0,0,0)
> mean.matn=mean(total_scores_of_math)
> mean.matn
[1] 86.28571
> total_scores_of_english=c(266,0,0,377,0,0,0)
> mean.english=mean(total_scores_of_english)
> mean.english
[1] 91.85714
> |
```

3. 一个小组中 5 位同学的数学成绩分别为 78，88，75，90，70，将他们的成绩保存到变量中并查询第 1，3，5 位同学的成绩

```
> score=c(78,88,75,90,70)
> score[1]
[1] 78
> score[3]
[1] 75
> score[5]
[1] 70
> |
```

[BY 曹倩]

现有一个团队去参加游戏，游戏规则是：每场比赛派出 3~6 人参赛，共有 3 场比赛，每场比赛每个队员有 n 次（ $n=3\sim 10$ ， n 自定）次比赛机会，相应的每个队员就有 n 个得分。比赛均为个人得分，每场比赛的队员个人得分是队员去掉最高和最低后的平均得分，比赛最后得分为个人得分的平均分。要求保证最后每个队员都参与了比赛。

该团队的得分数据为：（满分为 10）

第一场：

小红 9 小青 8 小紫 6 小红 8 小绿 5 小紫 7 小青 8 小绿 6 小红 8 小绿 9
小绿 8 小红 7 小紫 6 小红 8 小紫 7 小青 6 小紫 8 小紫 3 小青 4 小绿 6

第二场：

小绿 8 小黄 9 小绿 6 小红 10 小橙 8 小青 6 小橙 7 小黄 8 小绿 7 小红
8 小青 8 小绿 8 小青 2 小黄 6 小红 9 小橙 8 小绿 5 小青 5 小橙 9 小橙
6 小青 5 小红 8 小黄 6 小绿 9 小青 8

第三场：

小蓝 3 小橙 8 小紫 8 小蓝 8 小青 7 小紫 5 小绿 3 小紫 6 小黄 7 小蓝 6
小黄 7 小蓝 3 小黄 5 小紫 8 小黄 10 小紫 9 小橙 9 小紫 6 小蓝 8 小青
8 小绿 7 小橙 10 小青 2 小蓝 8 小橙 7

1. 每场派了几个队员去参赛？分别是哪几个队员？该团队总共有几个队员？（请利用 `dataframe`）
2. 每场比赛最后得分分别是多少？
3. 请通过计算每个人的平均分来制定该参赛团队的最佳搭配方法。（言之有理即可）

[BY 屈云红]

1. 在北京未来五天的天气预报中，气温变化如下：Wed 26~33℃，Thu 27~34℃，Fri 25~32℃，Sat 25~32℃，Sun 23~29℃。请将星期缩写保存至变量 `day` 中，并分别求出每一天的平均气温，保存至变量 `tem` 中。

```
wed=c(26,33);thu=c(27,34);fri=c(25,32);sat=c(25,32);sun=c(23,29);
day=c("wed","thu","fri","sat","sun")
tem=c(mean(wed),mean(thu),mean(fri),mean(sat),mean(sun))
> #3.1
> wed=c(26,33);thu=c(27,34);fri=c(25,32);sat=c(25,32);sun=c(23,29);
> day=c("wed","thu","fri","sat","sun")
> tem=c(mean(wed),mean(thu),mean(fri),mean(sat),mean(sun))
> tem
[1] 29.5 30.5 28.5 28.5 26.0
> |
```

- 1) 已知星期五的气温信息有误，最低气温应为 26℃，请将相关数据进行修改。

```
fri=c(26,32)
day=c("wed","thu","fri","sat","sun")
tem=c(mean(wed),mean(thu),mean(fri),mean(sat),mean(sun))
> fri=c(26,32)
> fri=c(26,32)
> day=c("wed","thu","fri","sat","sun")
> tem=c(mean(wed),mean(thu),mean(fri),mean(sat),mean(sun))
> tem
[1] 29.5 30.5 29.0 28.5 26.0
> |
```

- 2) 求出为来五天平均气温的最高温度以及最低温度。

#3.1.2

```
max=max(tem)
min=min(tem)
> #3.1.2
> max=max(tem)
> min=min(tem)
> max
[1] 30.5
> min
[1] 26
> |
```

2. 根据题目 1，构造数据框（**dataframe**）变量 **user.data**，由两列构成，第一列是 **day**,第二列是 **tem**。

- 1) 运用 **for** 和 **if** 命令判断哪几天的平均气温高于 30℃。
- 2) 已知星期一的最低温度为 22℃最高温度为 31℃，请求出其日平均温度并插入到题目数据框中。

#3.2

```
user.data=data.frame(day,tem)
```

```
for (i in 1:5)
```

```
{
  if (user.data$tem[i]>30)
  {
    print(user.data$day[i])
  }
}
```

```
mon=c(22,31)
```

```
day=c(day,"mon")
```

```
tem=c(tem,mean(mon))
```

```
user.data=data.frame(day,tem)
```

	day	tem
1	wed	29.5
2	thu	30.5
3	fri	29.0
4	sat	28.5
5	sun	26.0
6	mon	26.5

3. 已知甲乙两地相距 500 米，甲乙两人分别从甲乙两地相向而行，其中有一只小狗与甲同时出发，小狗的速度是每秒 10 米，小狗在遇到乙后立即返回，在中间进行往返，已知甲的速度为 3 米每秒，乙的速度为 4 米每秒，甲乙二人相遇时，小狗行了多少米？（请将题目中的数字保存到变量中，并进行加减乘除运算得出结果，保存到一个变量中，变量名可由自己定义。）

```
s=500
v1=3
v2=4
v3=10
time=s/(v1+v2)
s1=v3*time
> s=500
> v1=3
> v2=4
> v3=10
> time=s/(v1+v2)
> time
[1] 71.42857
> s1=v3*time
> s1
[1] 714.2857
> |
```

[BY 周馨茹]

- 1、 利用 seq（）函数，生成一组数字【3,10,17,24,31】;

利用 rep（）函数，重复数字 7 十次;

利用 ls 函数，列出所有变量；

定义变量 bigdata 为 30，然后将其删除

```
seq(3,31,by=7)
```

```
rep(7,10)
```

```
ls()
```

```
bingdata=30
```

```
rm(bingdata)
```

```
> seq(3,31,by=7)
[1] 3 10 17 24 31
> rep(7,10)
[1] 7 7 7 7 7 7 7 7 7 7
> ls()
[1] "day"           "fri"           "get_max_len_name" "get_mmmmm"
[5] "i"             "max"           "mean.chinese"     "mean.english"
[9] "mean.matn"     "min"           "mon"              "name"
[13] "names1"        "result"        "s"                "s1"
[17] "sat"           "score"         "sun"              "tem"
[21] "thu"           "time"          "total_scores_of_chinese" "total_scores_of_english"
[25] "total_scores_of_math" "user.data"     "v1"               "v2"
[29] "v3"           "wed"
> bingdata=30
> rm(bingdata)
> |
```

- 2、 一个数据框有两列一列姓名张三、李四、王五、孙六，一列是分数 77、50、92、100 结合 for 和 if 将题中成绩大于 90 的同学用 print 输出

```
name=c("zhangsan","lisi","wangwu","sunliu")
score=c(77,50,92,100)
data=data.frame(name,score)
for (i in 1:4)
{
  if (data$score[i]>90)
  {
    print(data$name[i])
  }
}
```

```

> for (i in 1:4)
+ {
+   if (data$score[i]>90)
+   {
+     print(data$name[i])
+   }
+ }
[1] wangwu
Levels: lisi sunliu wangwu zhangsan
[1] sunliu
Levels: lisi sunliu wangwu zhangsan
> |

```

- 3、 现有另一组数据 `weight` 为体重 55、66、77、88 和一组数据 `member` 为小红、86、44 利用 `cbind` 和 `rbind` 将两组数据准确的加入到上题的数据框中。

```
weight=c(55,66,77,88)
```

```
data1=cbind(data,weight)
```

```
data2=rbind(data1,data.frame(name="xiaohong",score=86,weight=44))
```

```

> weight=c(55,66,77,88)
> data1=cbind(data,weight)
> data2=rbind(data1,data.frame(name="xiaohong",score=86,weight=44))
> data2
      name score weight
1 zhangsan   77     55
2    lisi    50     66
3  wangwu   92     77
4  sunliu  100     88
5 xiaohong   86     44
> |

```

[BY 张琬玥]

1. 定义变量 `numeric` 为 80，格式为字符串，并判断变量是字符串（`is.character`）再将其转化为数字型格式，并判断是数字型。最后与 90 进行比较大小。

```

numeric=80
numeric=as.character(numeric)
is.character(numeric)

```

```

numeric=as.numeric(numeric)
is.numeric(numeric)

```

80>=90

2、应用题：有 5 个小组，每个小组有 3 个女生 2 个男生，男生需要 4 个苹果 2 个橘子，女生需要 3 个苹果 3 个橘子，问总共需要几个苹果几个橘子？一个苹果 2 元，一个橘子 1.5 元，总共多少钱？如果苹果打折 8 折，橘子涨价 20%，与之前总支出相差多少？

```
girls=3*5
boys=2*5
total_apple=3*girls+4*boys
total_orange=3*girls+2*boys

money=2*total_apple+1.5*total_orange
chajia=(1-0.8)*2*total_apple+(1.2-1)*1.5*total_orange
> girls=3*5
> boys=2*5
> total_apple=3*girls+4*boys
> total_apple
[1] 85
> total_orange=3*girls+2*boys
> total_orange
[1] 65
> money=2*total_apple+1.5*total_orange
> money
[1] 267.5
> chajia=(1-0.8)*2*total_apple+(1.2-1)*1.5*total_orange
> chajia
[1] 53.5
> |
```

你去的方向

3.小明、小红、小华、小丽，四人年龄分别为 18,19,18,19。性别分别是男,女,男,女。构成一个数据框。然后添加一列成绩 90,92,87,86。修改小红的年龄为 20。再将所有人的成绩统一减去十分。删除小华成绩。

```
name=c("xiaoming","xiaohong","xiaohua","xiaoli")
age=c(18,19,18,19)
gender=c("m","f","m","f")
data=data.frame(name,age,gender)
score=c(90,92,87,86)
data=data.frame(data,score)

data[2,2]=20
data[,4]-10
```


[BY 李婉荣]

1. 创建一个 1 到 100 的向量，形式为 1,4,7.....，并命名为 v1，选取其中第 5 个到第 20 个元素,以及大于 50 的元素。

```
v1=seq(1,100,3)
```

```
v1[5:20]
```

```
v1[v1>50]
```

```
> v1=seq(1,100,3)
> v1
[1] 1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79
[28] 82 85 88 91 94 97 100
> v1[5:20]
[1] 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58
> v1[v1>50]
numeric(0)
> v1[v1>50]
[1] 52 55 58 61 64 67 70 73 76 79 82 85 88 91 94 97 100
> |
```

2. 创建一个向量 x 包含 1 到 10 的数字，向量 y，使其数据大小为 x 中对应数字的平方，并创建数据表 data, 包含两列数据 x,y.

将原来的数据表 data 新增一列数据表示从 4 到 13 的数据，使更新后的数据表有三列，命名为 data.1

```
x=c(1:10)
```

```
y=x^2
```

```
data=data.frame(x,y)
```

```
z=c(2:11)
```

```
data.1=data.frame(data,z)
```

```

> data=data.frame(x,y)
> z=c(2:11)
> data.1=data.frame(data,z)
> data.1
  x  y  z
1  1  1  2
2  2  4  3
3  3  9  4
4  4 16  5
5  5 25  6
6  6 36  7
7  7 49  8
8  8 64  9
9  9 81 10
10 10 100 11
> |

```

3 将数据表 data.1 中所有数字 4 更改为 40

```

x=c(1:10)
y=x^2
data=data.frame(x,y)
z=c(2:11)
data.1=data.frame(data,z)
data.1[2,2]=40
data.1[3,3]=40
data.1[4,1]=40

```

```

> data.1[2,2]=40
> data.1[3,3]=40
> data.1[4,1]=40
> data.1
  x  y  z
1  1  1  2
2  2 40  3
3  3  9 40
4 40 16  5
5  5 25  6
6  6 36  7
7  7 49  8
8  8 64  9
9  9 81 10
10 10 100 11
> |

```

[BY 孙兴宇]

1.有一组淘宝店家数据:5 个商家分别为 (Dior Chanel GUCCI Prada HERMES), 它们的 1 季度的营业额为(55w 56w 35w 89w 12w) 请做表格 Luxury =记录这两列数据

现在想加入他们一行它们的一行净利润数据 (30w 25w 26w 50w 5w)

到表格 Luxury 中

```
dianjia=c("Dior","Chanel","GUCCI","Prada","HERMES")
```

```
yingyee=c(550000,560000,350000,890000,120000)
```

```
Luxury=data.frame(dianjia,yingyee)
```

```
w=c(300000,250000,260000,500000,50000)
```

```
Luxury=data.frame(Luxury,w)
```

2.接上题，现在想再加入一个品牌（Louis_Vuitton 69w 56w）于表格 Luxury 中。

```
> Luxury=rbind(Luxury,data.frame(dianjia="Louis_vuitton",yingyee=690000,w=560000))
> Luxury
  dianjia yingyee      w
1      Dior 550000 300000
2     Chanel 560000 250000
3      GUCCI 350000 260000
4      Prada 890000 500000
5     HERMES 120000  50000
6 Louis_Vuitton 690000 560000
> |
```

3.一组 5km 跑步运动员，（A B C D E ）现在知道他们的成绩为（16.30 16.22 15.45 17.01 15.55）最后一圈冲刺速度为（50s 51s 48s 55s 47s）请制作表格 grade

现在请算出他们的配速，并加入表格 grade 中

临时加入了一位运动员 X，仅知其配速为 1.14，并知其全程匀速，请将 X 的成绩加入到表格 grade 中。

```
a=c("A","B","C","D","E")
```

```
b=c(16.30,16.22,15.45,17.01,15.55)
```

```
c=c(50,51,48,55,47)
```

```
grade=data.frame(a,b,c)
```

```
peisu=b/5
```

```
grade=data.frame(grade,peisu)
```

```
grade=rbind(grade,data.frame(a="x",b=1.14*5 ,c=57,peisu=1.14))
```

[BY 周子轲]

1. 以下正确的变量名是: **abcd**

A、a.1 B、bsd_ C、D12_34 D、sdi23.

2.以下表示变量 a 小于-3 的表达式是: **c**

A、a<-3 B、a<-(-3) C、a<(-3)

3.已知表示中国代表队在里约奥运会最终得到的金牌总数的变量为 gold_total。据可靠预测,中国代表队在里约奥运会得到的金牌总数为:
(里约奥运会的参赛国家及地区总数+中华人民共和国英文名全称所含字母数) X 2

[BY 张倩倩]

1. 用 rep 构造一个向量 x, 它由 3 个 1, 2 个 2, 5 个 3 组成。

```
> rep(1:3,c(3,2,5))  
[1] 1 1 1 2 2 3 3 3 3 3
```

2. 某公司在 2010 2011 2012 2013 四年的销售额分别是 2000 3000 2500 3000.构造数据框 sale.data, 由两列构成, 第一列是 year, 第二列是 sales。并把数据框中 year 这一列的数值类型转换为因子。

```
sales.data=data.frame(year=c(2010,2011,2012,2013),sales=c(2000,3000,
```

```
2500,3000))
```

```
sales.data=data.frame(year=as.factor(c(2010,2011,2012,2013)),sales=c(2000,3000,2500,3000))
```

3、定义一个 4 行 3 列，由 1:12 按行次序排列的矩阵。

```
mdat <- matrix(1:12, nrow = 4, ncol = 3, byrow = TRUE)
```

	V1	V2	V3
1	1	2	3
2	4	5	6
3	7	8	9
4	10	11	12