talk04 练习与作业

目录

0.1 练习和作业说明

将相关代码填写入以 "'{r} "' 标志的代码框中，运行并看到正确的结果；

完成后，用工具栏里的"Knit" 按键生成 PDF 文档；

将 PDF 文档改为：`姓名-学号-talk04 作业.pdf`，并提交到老师指定的平台/钉群。

0.2 Talk04 内容回顾

待写 …

0.3 练习与作业：用户验证

请运行以下命令，验证你的用户名。

如你当前用户名不能体现你的真实姓名，请改为拼音后再运行本作业！

```
Sys.info()[["user"]]
```

```
## [1] "Zhu Fangannan"
```

```
Sys.getenv("HOME")
```

```
## [1] "C:/Users/Zhu Fangannan/Documents"
```

0.4 练习与作业 1：R session 管理

0.4.1 完成以下操作

- 定义一些变量（比如 x, y , z 并赋值；内容随意）
- 从外部文件装入一些数据（可自行创建一个 4 行 5 列的数据，内容随意）
- 保存 workspace 到.RData
- 列出当前工作空间内的所有变量
- 删除当前工作空间内所有变量
- 从.RData 文件恢复保存的数据
- 再次列出当前工作空间内的所有变量，以确认变量已恢复
- 随机删除两个变量
- 再次列出当前工作空间内的所有变量

```
## 代码写这里，并运行；
x<-111;
y<-"abc";
z<-"##$"
a<-matrix(c(sample(1:100,20)),nrow=4)
save.image(file="prj_r_for_bioinformatics");
ls()
```

```
## [1] "a"         "encoding"  "inputFile" "pSubTitle" "x"         "y"
## [7] "z"
```

```r
rm(list=ls())
load(file="prj_r_for_bioinformatics");
ls()
```

```
## character(0)
```

```r
rm(a,x)
```

```
## Warning in rm(a, x): 找不到对象'a'
```

```
## Warning in rm(a, x): 找不到对象'x'
```

```r
ls()
```

```
## character(0)
```

0.5 练习与作业 2：Factor 基础

0.5.1 factors 增加

- 创建一个变量：

`x <- c("single", "married", "married", "single");`

- 为其增加两个 levels，`single`, `married`;

- 以下操作能成功吗？

`x[3] <- "widowed";`

不能成功。levels 中没有”widowed”，所以不行。

- 如果不，请提供解决方案；

```
## 代码写这里，并运行；
 x <- c("single", "married", "married", "single");
 x<-as.factor(x);
levels(x)<-c(levels(x),"single", "married");
## 解决方案
levels(x)<-c(levels(x),"widowed");
x[3] <- "widowed"
x
```

```
## [1] single  married widowed single
## Levels: married single widowed
```

0.5.2 factors 改变

- 创建一个变量：

```
v = c("a", "b", "a", "c", "b")
```

- 将其转化为 `factor`，查看变量内容
- 将其第一个 levels 的值改为任意字符，再次查看变量内容

```
## 代码写这里，并运行；
v = c("a", "b", "a", "c", "b")
v<-as.factor(v);
v
```

```
## [1] a b a c b
## Levels: a b c
```

```
levels(v)[1]<-"z"
v
```

```
## [1] z b z c b
## Levels: z b c
```

- 比较改变前后的 v 的内容，改变 levels 的操作使 v 发生了什么变化？

答：v 中所有第一个 levels 的值都被替换了。

0.5.3 factors 合并

- 创建两个由随机大写字母组成的 factors
- 合并两个变量，使其 `factors` 得以在合并后保留

```
## 代码写这里，并运行;
a<-factor(c(sample(LETTERS,5)));
b<-factor(c(sample(LETTERS,7)));
a
```

```
## [1] O J K N A
## Levels: A J K N O
```

```
b
```

```
## [1] N V T L R Y G
## Levels: G L N R T V Y
```

```
x<-c(a,b)
x
```

```
##  [1] O J K N A N V T L R Y G
## Levels: A J K N O G L R T V Y
```

0.5.4 利用 factor 排序

以下变量包含了几个月份，请使用 `factor`，使其能按月份，而不是英文字符串排序：

```
mon <- c("Mar","Nov","Mar","Aug","Sep","Jun","Nov","Nov","Oct","Jun","May","Sep","Dec",
```

```
## 代码写这里，并运行;
mon <- c("Mar","Nov","Mar","Aug","Sep","Jun","Nov","Nov","Oct","Jun","May","Sep","Dec",
month_levels<-c(
"Jan","Feb","Mar","Apr","May","Jun",
"Jul","Aug","Sep","Oct","Nov","Dec"
)
mon1<-factor(mon,levels=month_levels)
sort(mon1)
```

```
##  [1] Mar Mar May Jun Jun Jul Aug Sep Sep Oct Nov Nov Nov Nov Dec
## Levels: Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
```

0.5.5 forcats 的问题

forcats 包中的 fct_inorder, fct_infreq 和 fct_inseq 函数的作用是什么？

fct_inorder：levels 按照第一次出现的顺序排序 fct_infreq：levels 按照出现的次数从大到小排序, 相同的再按照大小从小到大排序 fct_inseq：levels 按照数字大小从小到大排序

请使用 forcats 包中的 gss_cat 数据举例说明

```
## 代码写这里，并运行；
library(forcats)
a<-head(gss_cat,n=10)
f1=a$age
f1<-as.factor(f1)
f1
```

```
##  [1] 26 48 67 39 25 25 36 44 44 47
## Levels: 25 26 36 39 44 47 48 67
```

```
fct_inorder(f1)
```

```
##  [1] 26 48 67 39 25 25 36 44 44 47
## Levels: 26 48 67 39 25 36 44 47
```

```
fct_infreq(f1)
```

```
##  [1] 26 48 67 39 25 25 36 44 44 47
## Levels: 25 44 26 36 39 47 48 67
```

```
fct_inseq(f1)
```

```
##  [1] 26 48 67 39 25 25 36 44 44 47
## Levels: 25 26 36 39 44 47 48 67
```

0.6 练习与作业 3：用 mouse genes 数据做图

0.6.1 画图

1. 用 readr 包中的函数读取 mouse genes 文件（从本课程的 Github 页面下载 data/talk04/ ）
2. 选取常染色体（1-19）和性染色体（X，Y）的基因
3. 画以下两个基因长度 boxplot :

- 按染色体序号排列，比如 1, 2, 3 .... X, Y
- 按基因长度中值排列，从短 -> 长 ...
- 作图结果要求：
  - 要清晰显示 boxplot 的主体；
  - 严格按照中值进行排序；注：‘ylim()‘限制时会去除一些值，造成中值错位。可考虑使用其它函数或调整参数。

```
## 代码写这里，并运行；
library(ggforce)
```

```
## 载入需要的程辑包：ggplot2
```

```
library(forcats)
library(dplyr)
```

```
##
## 载入程辑包：'dplyr'
```

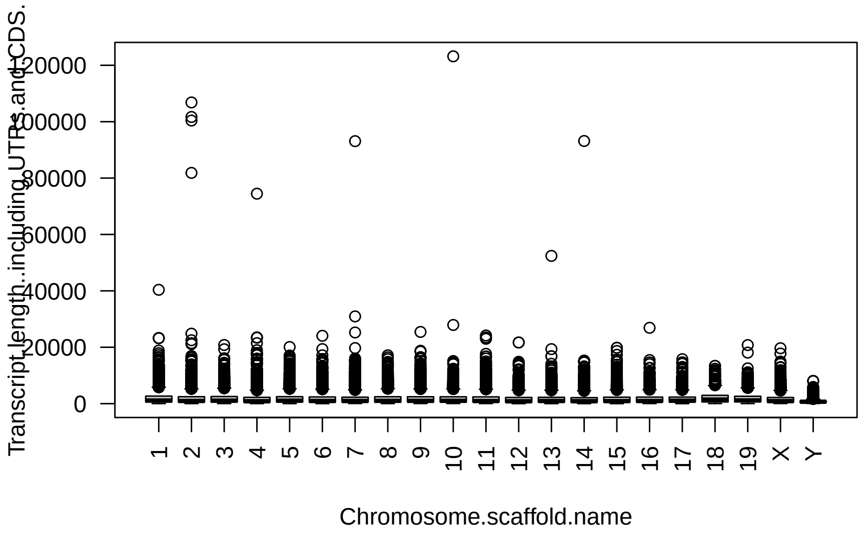
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
mouse.genes<-
    read.delim(file="../data/talk04/mouse_genes_biomart_sep2018.txt",
               sep="\t",header=T,stringsAsFactors=T);
mouse.chr_genes<-
    subset(mouse.genes,Chromosome.scaffold.name%in%
               c("1","2","3","4","5","6",
                 "7","8","9","10","11","12" ,
                 "13","14","15","16","17","18","19","X","Y")); mouse.chr_genes$Chromoso
    droplevels (mouse.chr_genes$Chromosome.scaffold.name)
mouse.chr_genes$Chromosome.scaffold.name<-
    fct_inseq(mouse.chr_genes$Chromosome.scaffold.name)
levels (mouse.chr_genes$Chromosome.scaffold.name)
```

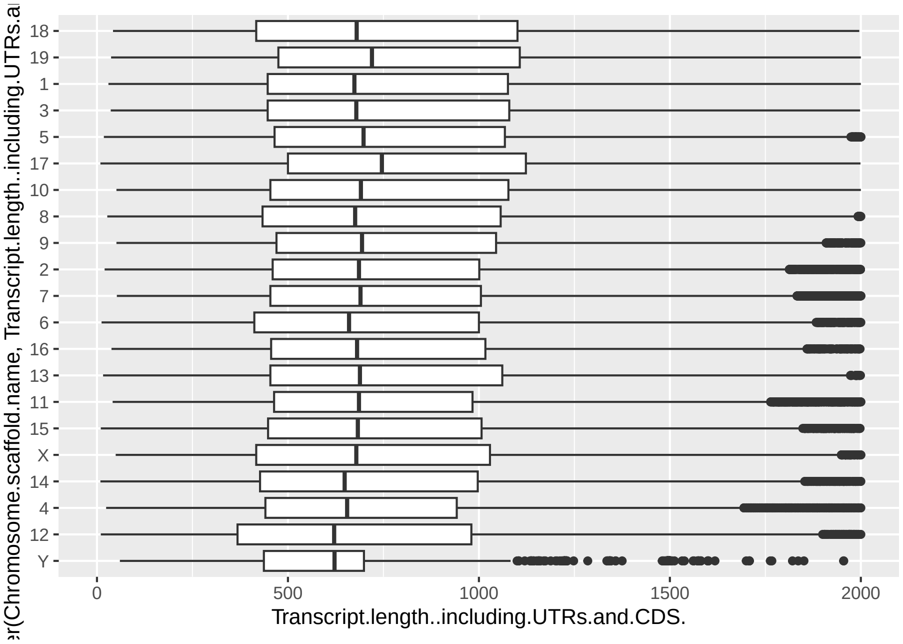
```
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14" "15"
## [16] "16" "17" "18" "19" "X"  "Y"
```

```
boxplot(Transcript.length..including.UTRs.and.CDS.~Chromosome.scaffold.name,
        data=mouse.chr_genes,las =2)
```



```
library(readr)
mouse.tibble<-
  read.delim(file="../data/talk04/mouse_genes_biomart_sep2018.txt",
             quote="")
mouse.tibble.chr10_12<-
mouse.tibble %>% filter(`Chromosome.scaffold.name` %in% c("1","2","3","4","5","6",
                 "7","8","9","10","11","12" ,
                 "13","14","15","16","17","18","19","X","Y"));
plot4<-
ggplot(data=mouse.tibble.chr10_12,
aes(x=reorder(`Chromosome.scaffold.name`,
`Transcript.length..including.UTRs.and.CDS.`,median),y=`Transcript.length..including.UT
geom_boxplot()+
coord_flip()+
ylim(0,2000)
plot4
```

```
## Warning: Removed 41185 rows containing non-finite values (`stat_boxplot()`).
```


er(Chromosome.scaffold.name, Transcript.length..including.UTRs.a
18
19
1
3
5
17
10
8
9
2
7
6
16
13
11
15
X
14
4
12
Y
0
500
1000
1500
2000
Transcript.length..including.UTRs.and.CDS.