# 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] "s56hh"
```

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

```
## [1] "C:/Users/s56hh/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；
rm(list=ls())
x<-114514
y<-" 嗯嘛啊"
z<-letters[1:6]
cxk<-read.table("data/Table0.txt")
```

```
save.image("data/Table0.RData")
ls()
```

```
## [1] "cxk" "x"   "y"   "z"
```

```
rm(list=ls())
load("data/Table0.RData")
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";
```

- 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);
x[ length(x) + 1 ] <-"single"
x[ length(x) + 1 ] <-"married"
```

```
levels(x) <- c(levels(x), "widowed");
x[ length(x) + 1 ] <- "widowed";
x
```

```
## [1] single  married married single  single  married widowed
## 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))
```

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

```
v_levels=c("kasumi","b","c")
v<-factor(v,levels = v_levels)
v
```

```
## [1] <NA> b    <NA> c    b
## Levels: kasumi 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：

### 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

```
ff<-LETTERS[runif(2,min=1,max=26)]
ff1<-ff[1]
ff2<-ff[2]
ff<-as.factor(ff)
ff
```

```
## [1] Y Q
## Levels: Q Y
```

```
Lycoris<-paste(ff[1],ff[2], sep = "", collapse = NULL)
ff_levels=c(Lycoris,ff1,ff2)
ff<-factor(Lycoris,levels=ff_levels)
ff
```

```
## [1] YQ
## Levels: YQ Y Q
```

### 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",
mon<-factor(mon,levels=month_levels)
sort(mon)
```

```
##  [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？

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)
head(gss_cat)
```

```
##   year        marital age  race        rincome             partyid
## 1 2000 Never married   26 White  $8000 to 9999        Ind,near rep
## 2 2000       Divorced   48 White  $8000 to 9999 Not str republican
## 3 2000        Widowed   67 White Not applicable         Independent
## 4 2000 Never married   39 White Not applicable        Ind,near rep
## 5 2000       Divorced   25 White Not applicable   Not str democrat
## 6 2000        Married   25 White $20000 - 24999    Strong democrat
##                 relig              denom tvhours
## 1          Protestant Southern baptist      12
## 2          Protestant Baptist-dk which      NA
## 3          Protestant  No denomination       2
## 4 Orthodox-christian   Not applicable       4
## 5                None   Not applicable       1
## 6          Protestant Southern baptist      NA
```

```r
attach(gss_cat)
head(fct_inorder(marital),n=10)
```

```
##  [1] Never married Divorced      Widowed       Never married Divorced
##  [6] Married       Never married Divorced      Married       Married
## Levels: Never married Divorced Widowed Married Separated No answer
```

```r
head(fct_infreq(rincome),n=30)
```

```
##  [1] $8000 to 9999  $8000 to 9999  Not applicable Not applicable Not applicable
##  [6] $20000 - 24999 $25000 or more $7000 to 7999  $25000 or more $25000 or more
## [11] $25000 or more $25000 or more $25000 or more $25000 or more $25000 or more
## [16] $25000 or more Not applicable $25000 or more $10000 - 14999 Not applicable
## [21] $25000 or more Refused        Not applicable $25000 or more Not applicable
## [26] Not applicable Not applicable Not applicable Not applicable Not applicable
## 16 Levels: $25000 or more Not applicable $20000 - 24999 ... No answer
```

```r
dd<-factor(1:9,levels=c("1 ","2","3","4","5","6","7","8","9"))
fct_inseq(dd)
```

```
## [1] <NA> 2    3    4    5    6    7    8    9
## Levels: 1  2 3 4 5 6 7 8 9
```

## 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 …

```
## 代码写这里，并运行；
library(readr)
library(ggplot2)
library(dplyr)
```

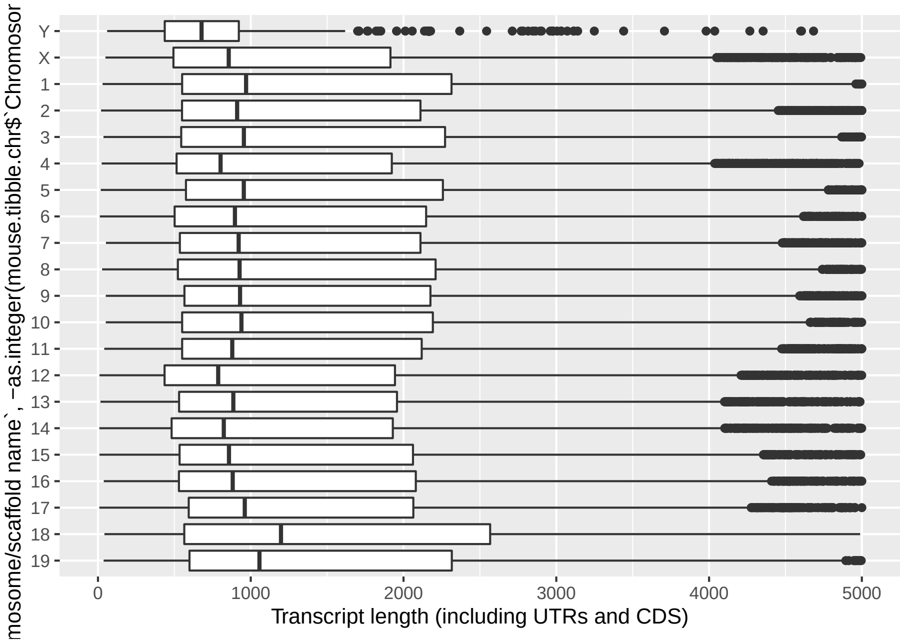
```
##
## 载入程辑包：'dplyr'

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

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

```
mouse.tibble <- read_delim( file = "data/talk04/mouse_genes_biomart_sep2018.txt",
delim = "\t", quote = "" )
```

```
## Rows: 138532 Columns: 6

## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (5): Gene stable ID, Transcript stable ID, Protein stable ID, Transcript...
## dbl (1): Transcript length (including UTRs and CDS)
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

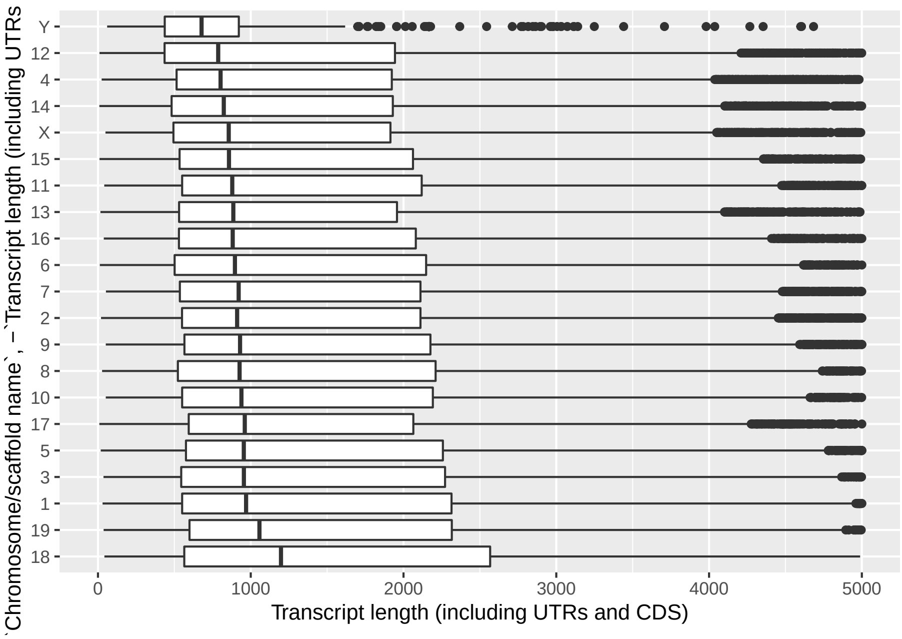
```
mouse.tibble.chr <-mouse.tibble %>% filter( `Chromosome/scaffold name` %in% c(1:19,"X",
plot1 <-
ggplot( data = mouse.tibble.chr,
aes( x = reorder( `Chromosome/scaffold name`,
-as.integer(mouse.tibble.chr$`Chromosome/scaffold name`)),
y = `Transcript length (including UTRs and CDS)` ) ) +
geom_boxplot() +
coord_flip() +
ylim( 0, 5000 )
plot1
```

```
## Warning in tapply(X = X, INDEX = x, FUN = FUN, ...): 强制改变过程中产生了NA

## Warning in tapply(X = X, INDEX = x, FUN = FUN, ...): 强制改变过程中产生了NA

## Warning: Removed 6639 rows containing non-finite values (stat_boxplot).
```



```
plot2 <-
ggplot( data = mouse.tibble.chr,
aes( x = reorder( `Chromosome/scaffold name`,
-`Transcript length (including UTRs and CDS)`,
median ),
y = `Transcript length (including UTRs and CDS)` ) ) +
geom_boxplot() +
coord_flip() +
ylim( 0, 5000 )
plot2
```

```
## Warning: Removed 6639 rows containing non-finite values (stat_boxplot).
```


`Chromosome/scaffold name`, -`Transcript length (including UTRs
Y
X
Transcript length (including UTRs and CDS)