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！

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

```
## [1] "GUANWEIHAI"
```

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

```
## [1] "C:/Users/GUANWEIHAI/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

```r
## 代码写这里，并运行；
x=666
y='你干嘛'
z='小黑子'
bbq<-read.table(file ="data/x666.txt")
save.image(file = "data/x666.RData")
ls()
```

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

```
rm(list=ls())
load(file = "data/x666.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");
levels(x)<-c('single', 'married')
levels(x) <- c(levels(x), "widowed")
x[4] <- "widowed";
str(x);
```

```
##  chr [1:4] "single" "married" "married" "widowed"
##  - attr(*, "levels")= chr [1:3] "single" "married" "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
```

```
levels(v)<-c("q","b","c")
v
```

```
## [1] q b q c b
## Levels: q 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：使原本第一个 levels 对应的 v 中数值一同改变 ### **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；
XHZ<-as.factor(sample(LETTERS,2))
IKUN<-as.factor(sample(LETTERS,2))
(sz666<-c(XHZ,IKUN))
```

```
## [1] L O A J
## Levels: L O A J
```

0.5.3 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",
levels(mon)= c('Jan','Feb','Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec')
bbq<-factor(mon,levels =levels(mon))
sort(bbq)
```

```
##  [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
```

0.5.4 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 按它们首次出现的顺序。fct_infreq 按每个 level 的观察到的数量（数量多者优先, 数量相同则按数值从小到大）fct_inseq 按 level 的数字值从小到大

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
```

```
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
```

```
head(fct_infreq(relig),n=10)
```

```
##  [1] Protestant         Protestant         Protestant         Orthodox-christian
##  [5] None               Protestant         Christian          Protestant
##  [9] Protestant         Protestant
## 16 Levels: Protestant Catholic None Christian Jewish Other ... Not applicable
```

```
bbq<-as.factor(c(99,12,13,14,16,14,13,13,13,15))
fct_inseq(bbq)
```

```
##  [1] 99 12 13 14 16 14 13 13 13 15
## Levels: 12 13 14 15 16 99
```

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(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
```

```
library(readr)
library(ggplot2)
bbq<-read_delim(file="data/talk04/mouse_genes_biomart_sep2018.txt",delim="\t",quote="")
```

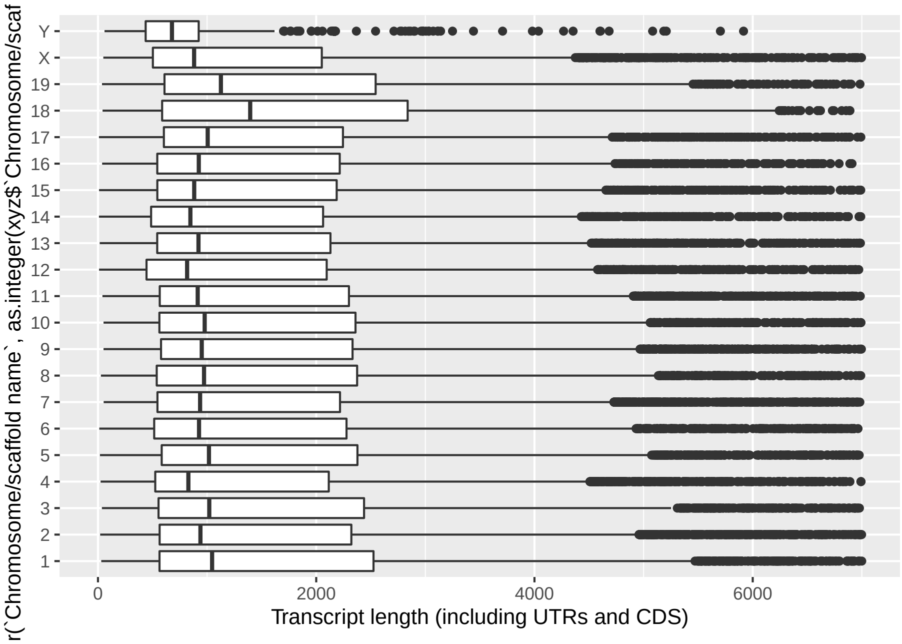
```
## Rows: 138532 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (5): Gene stable ID, Transcript stable ID, Protein stable ID, Transcript...
## dbl (1): Transcript length (including UTRs and CDS)
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
xyz<-bbq %>% filter(`Chromosome/scaffold name`%in% c(1:19,'X','Y' ))
xyz
```

```
## # A tibble: 136,498 x 6
##    `Gene stable ID`   `Transcript stable ID` Protein s~1 Trans~2 Trans~3 Chrom~4
##    <chr>              <chr>                  <chr>         <dbl> <chr>   <chr>
##  1 ENSMUSG00000097062 ENSMUST00000181502     <NA>            908 lincRNA 11
##  2 ENSMUSG00000097658 ENSMUST00000180595     <NA>            933 lincRNA 11
##  3 ENSMUSG00000097294 ENSMUST00000181119     <NA>           3683 lincRNA 11
##  4 ENSMUSG00000097020 ENSMUST00000180919     <NA>           1457 lincRNA 11
##  5 ENSMUSG00000097289 ENSMUST00000180492     <NA>           1004 lincRNA 11
##  6 ENSMUSG00000097289 ENSMUST00000181758     <NA>           2493 lincRNA 11
##  7 ENSMUSG00000097176 ENSMUST00000180389     <NA>            993 lincRNA 11
##  8 ENSMUSG00000096983 ENSMUST00000181900     <NA>           1199 lincRNA 11
##  9 ENSMUSG00000097335 ENSMUST00000181152     <NA>           1931 lincRNA 11
## 10 ENSMUSG00000097335 ENSMUST00000181003     <NA>           1704 lincRNA 11
## # ... with 136,488 more rows, and abbreviated variable names
## #   1: `Protein stable ID`, 2: `Transcript length (including UTRs and CDS)`,
## #   3: `Transcript type`, 4: `Chromosome/scaffold name`
```

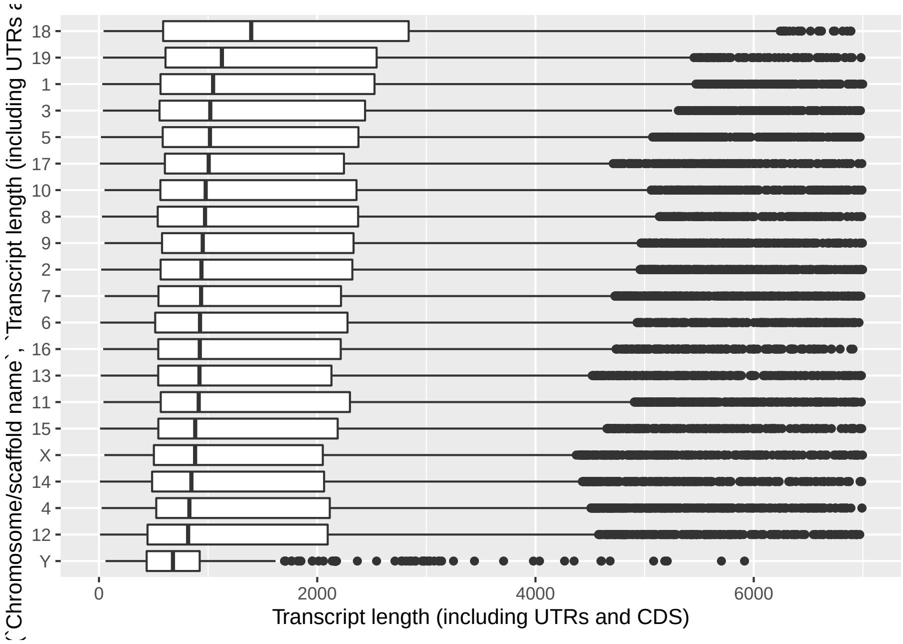
```
plot1<-
    ggplot(data =xyz,
          aes(x=reorder(`Chromosome/scaffold name`,
             as.integer(xyz$`Chromosome/scaffold name`)),
               y =`Transcript length (including UTRs and CDS)`)) +
    geom_boxplot() +
    coord_flip() +
    ylim(0,7000) ;
plot1;
```

```
## 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 2360 rows containing non-finite values (stat_boxplot).
```



```
plot2 <-
    ggplot(data=xyz,
          aes(x=reorder(`Chromosome/scaffold name`,
                            `Transcript length (including UTRs and CDS)`,
                            median,T ),
               y=`Transcript length (including UTRs and CDS)`)) +
    geom_boxplot() +
    coord_flip() +
    ylim(0,7000);
plot2;
```

```
## Warning: Removed 2360 rows containing non-finite values (stat_boxplot).
```


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