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] "wchen"
```

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

```
## [1] "/Users/wchen"
```

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；
# 定义一些变量（比如 x, y , z 并赋值；内容随意
rm(list=ls())
x <- c("single", "married", "married", "single");
y <- c(10,100,1000, 10000);
Z <- LETTERS[1:12];

# 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）
w=read.table(file="data/Table0.txt")
```

```
# 保存 workspace 到.RData
save.image(file = "data/Table0.RData");

# 列出当前工作空间内的所有变量
ls();
```

```
## [1] "w" "x" "y" "Z"
```

```
# 删除当前工作空间内所有变量
rm(list=ls());

# 从.RData 文件恢复保存的数据
load(file = "data/Table0.RData");

# 再次列出当前工作空间内的所有变量
ls();
```

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

0.5 练习与作业 2：Factor 基础

0.5.1 factor 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);
levels(x) <- c("single","married");
#x[3] <- "widowed";
# 不行，因为 x 为 factor，只允许接受 single 和 married
levels(x) <- c(levels(x), "widowed");
x[ length(x) + 1 ] <- "widowed";
x[3] <- "widowed";
x;
```

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

0.5.2 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","Jul","Nov");
month_levels <- c("Jan", "Feb", "Mar", "Apr", "May","Jun",
                  "Jul", "Aug", "Sep", "Oct", "Nov", "Dec");
x1 <- factor(mon, levels = month_levels);
sort(x1);
```

```
##  [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
```

0.5.3 forcats 的问题

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

This family of functions changes only the order of the levels

fct_inorder : by the order in which they first appear.

fct_infreq : by number of observations with each level(large first)

fct_inseq : by numeric value of level

请使用 forcats 包中的 gss_cat 数据举例说明

```
## 代码写这里, 并运行;
if (!require("forcats")){
  chooseCRANmirror();
  install.packages("forcats",destdir = "D:/resourse/software/Rproject4.1.1/download pac
}
```

```
## Loading required package: forcats
```

```
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=20)
```

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

```
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
```

```
f<-factor(1:6,levels=c("1 ","2","3","4","5","6"))
fct_inseq(f)
```

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

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；
if (!require("dplyr")){
  chooseCRANmirror();
  install.packages("dplyr",destdir = "D:/resourse/software/Rproject4.1.1/download packa
}
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

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

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

```
library("dplyr");
#devtools::install_github("tidyverse/dplyr")
library(readr)
mouse.tibble<-read_delim(file="../data/talk04/mouse_genes_biomart_sep2018.txt",delim="\
```

```
## 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.
```

```
head(mouse.tibble);
```

```
## # A tibble: 6 x 6
##   `Gene stable ID`  `Transcript stabl~ `Protein stable ~ `Transcript length (in~
##   <chr>             <chr>              <chr>                              <dbl>
## 1 ENSMUSG000000643~ ENSMUST00000082423 <NA>                                  67
## 2 ENSMUSG000000643~ ENSMUST00000082422 <NA>                                  67
## 3 ENSMUSG000000643~ ENSMUST00000082421 ENSMUSP000000810~                   1144
## 4 ENSMUSG000000643~ ENSMUST00000082420 <NA>                                  69
## 5 ENSMUSG000000643~ ENSMUST00000082419 ENSMUSP000000810~                    519
## 6 ENSMUSG000000643~ ENSMUST00000082418 ENSMUSP000000810~                   1824
## # ... with 2 more variables: Transcript type <chr>,
## #   Chromosome/scaffold name <chr>
```

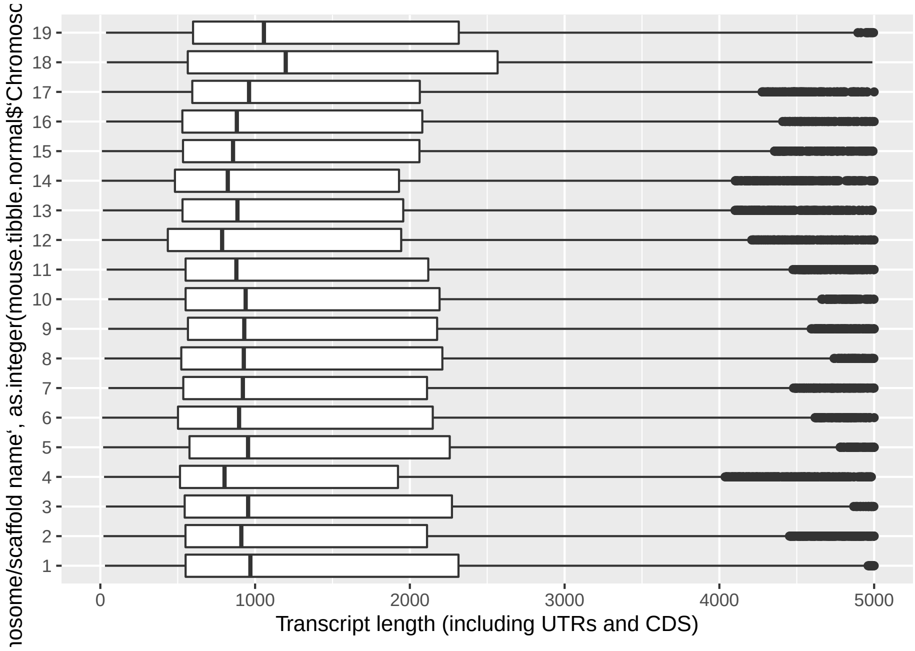
```
colnames(mouse.tibble);
```

```
## [1] "Gene stable ID"
## [2] "Transcript stable ID"
## [3] "Protein stable ID"
## [4] "Transcript length (including UTRs and CDS)"
## [5] "Transcript type"
## [6] "Chromosome/scaffold name"
```

```
mouse.tibble.normal <-mouse.tibble %>%
  filter( `Chromosome/scaffold name` %in% c( 1:19))
mouse.tibble.xy<-mouse.tibble%>%
  filter( `Chromosome/scaffold name` %in% c('X','Y'))
mouse.tibble.20<-
  bind_rows(mouse.tibble.normal,mouse.tibble.xy)

library(ggplot2)
plot1 <-
  ggplot( data = mouse.tibble.normal,
        aes( x = reorder( `Chromosome/scaffold name`,
           as.integer(mouse.tibble.normal$`Chromosome/scaffold name`)),
             y = `Transcript length (including UTRs and CDS)` ) ) +
  geom_boxplot() +
  coord_flip() +
  ylim( 0, 5000 ) ;
plot1;
```

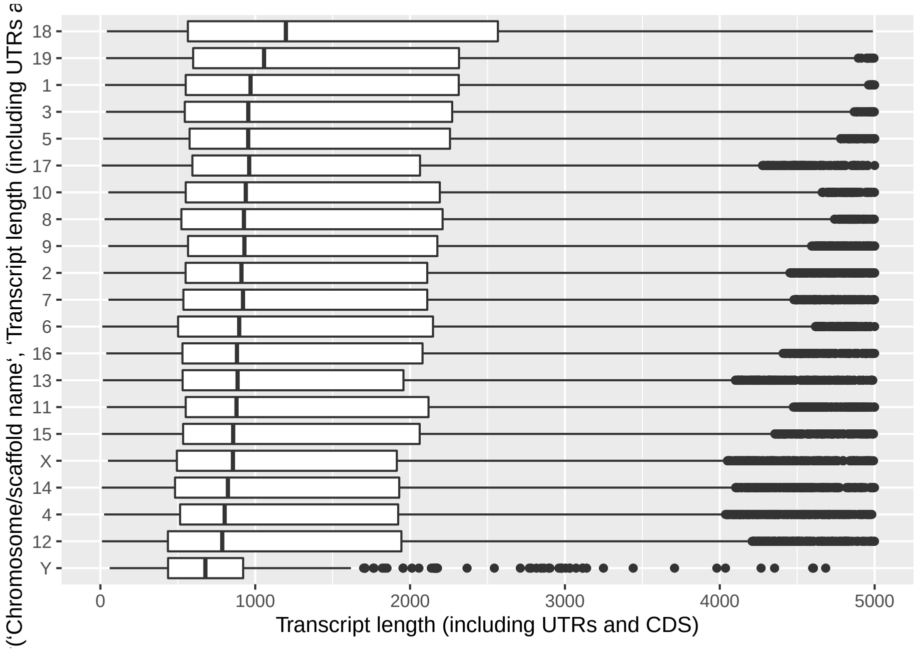
```
## Warning: Removed 6377 rows containing non-finite values (stat_boxplot).
```



```
plot2 <-
  ggplot( data = mouse.tibble.20,
        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, 5000);

plot2;
```

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


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