talk04 练习与作业

目录

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。

Talk04 内容回顾

练习与作业 1：R session 管理

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<-c('GEM','JJ Lin','Jay Chou')
y<-c(100,90,80)
z<-c(90,95,100)
library(readr)
a<-read.csv('data/states1.csv')[1:4,1:5]
a
```

```
##          X Population Income Illiteracy Life.Exp
## 1  Alabama       3615   3624        2.1    69.05
## 2   Alaska        365   6315        1.5    69.31
## 3  Arizona       2212   4530        1.8    70.55
## 4 Arkansas       2110   3378        1.9    70.66
```

```
save(x,y,z,a,file='talk_04_R_session_homework.RData')
ls()
```

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

```
rm(list=ls())
ls()
```

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

```
load('talk_04_R_session_homework.RData')
ls()
```

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

```
rm(y,z)
ls()
```

```
## [1] "a" "x"
```

练习与作业 2：Factor 基础

---

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 <- as.factor(c("single", "married", "married", "single")));
```

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

```
x[length(x)+1]<-'single'
x
```

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

```
x[length(x)+1]<-'married'
x
```

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

```
# 题目中`x[3] <- "widowed`不能成功，应按如下方式
levels(x)<-c(levels(x),'widowed')
x
```

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

```
x[length(x)+1]<-'widowed'
x
```

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

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

---

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；
rm(list=ls())
library(forcats)
Sys.setlocale('LC_ALL','C')
```

```
## [1] "C"
```

```
# 考虑到 gss_cat 数据量极大，作为举例说明
# 这里一般只用某一列的前 100 行进行
# fct_inorder 作用是，使 levels 按照第一次出现的顺序进行排序
# 官方文档：by the order in which they first appear.
a<-as.matrix(gss_cat[1:100,2])
a<-as.factor(a)
# 原始
a
```

```
##   [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
##  [21] Married       Married       Never married Widowed       Widowed
##  [26] Widowed       Widowed       Widowed       Divorced      Widowed
##  [31] Widowed       Married       Married       Never married Married
##  [36] Never married Never married Never married Never married Never married
##  [41] Married       Married       Divorced      Never married Never married
##  [46] Never married Married       Married       Married       Married
##  [51] Never married Married       Married       Married       Married
##  [56] Divorced      Divorced      Divorced      Never married Never married
##  [61] Married       Married       Never married Divorced      Never married
##  [66] Widowed       Divorced      Married       Never married Never married
##  [71] Widowed       Widowed       Widowed       Widowed       Widowed
##  [76] Never married Widowed       Never married Married       Never married
##  [81] Married       Married       Widowed       Married       Married
##  [86] Divorced      Never married Separated     Never married Widowed
##  [91] Widowed       Married       Divorced      Never married Never married
##  [96] Never married Married       Married       Widowed       Divorced
## Levels: Divorced Married Never married Separated Widowed
```

```
# 操作后
fct_inorder(a)
```

```
##   [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
##  [21] Married       Married       Never married Widowed       Widowed
##  [26] Widowed       Widowed       Widowed       Divorced      Widowed
##  [31] Widowed       Married       Married       Never married Married
##  [36] Never married Never married Never married Never married Never married
##  [41] Married       Married       Divorced      Never married Never married
##  [46] Never married Married       Married       Married       Married
##  [51] Never married Married       Married       Married       Married
##  [56] Divorced      Divorced      Divorced      Never married Never married
##  [61] Married       Married       Never married Divorced      Never married
##  [66] Widowed       Divorced      Married       Never married Never married
##  [71] Widowed       Widowed       Widowed       Widowed       Widowed
##  [76] Never married Widowed       Never married Married       Never married
##  [81] Married       Married       Widowed       Married       Married
##  [86] Divorced      Never married Separated     Never married Widowed
##  [91] Widowed       Married       Divorced      Never married Never married
##  [96] Never married Married       Married       Widowed       Divorced
## Levels: Never married Divorced Widowed Married Separated
```

```
''
```

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

```
# 'fct_infreq 作用是，使 levels 按照出现的频率排序'
# '官方文档: by number of observations with each level (largest first)'
b<-as.matrix(gss_cat[1:100,2])
b<-as.factor(b)
# 原始
b
```

```
##   [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
##  [21] Married       Married       Never married Widowed       Widowed
##  [26] Widowed       Widowed       Widowed       Divorced      Widowed
##  [31] Widowed       Married       Married       Never married Married
##  [36] Never married Never married Never married Never married Never married
##  [41] Married       Married       Divorced      Never married Never married
##  [46] Never married Married       Married       Married       Married
##  [51] Never married Married       Married       Married       Married
##  [56] Divorced      Divorced      Divorced      Never married Never married
##  [61] Married       Married       Never married Divorced      Never married
##  [66] Widowed       Divorced      Married       Never married Never married
##  [71] Widowed       Widowed       Widowed       Widowed       Widowed
##  [76] Never married Widowed       Never married Married       Never married
##  [81] Married       Married       Widowed       Married       Married
##  [86] Divorced      Never married Separated     Never married Widowed
##  [91] Widowed       Married       Divorced      Never married Never married
##  [96] Never married Married       Married       Widowed       Divorced
## Levels: Divorced Married Never married Separated Widowed
```

```
# 操作后
fct_infreq(b)
```

```
##   [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
##  [21] Married       Married       Never married Widowed       Widowed
##  [26] Widowed       Widowed       Widowed       Divorced      Widowed
##  [31] Widowed       Married       Married       Never married Married
##  [36] Never married Never married Never married Never married Never married
##  [41] Married       Married       Divorced      Never married Never married
##  [46] Never married Married       Married       Married       Married
```

```
##  [51] Never married Married       Married       Married       Married
##  [56] Divorced      Divorced      Divorced      Never married Never married
##  [61] Married       Married       Never married Divorced      Never married
##  [66] Widowed       Divorced      Married       Never married Never married
##  [71] Widowed       Widowed       Widowed       Widowed       Widowed
##  [76] Never married Widowed       Never married Married       Never married
##  [81] Married       Married       Widowed       Married       Married
##  [86] Divorced      Never married Separated     Never married Widowed
##  [91] Widowed       Married       Divorced      Never married Never married
##  [96] Never married Married       Married       Widowed       Divorced
## Levels: Married Never married Widowed Divorced Separated
```

```
''
```

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

```
# 'fct_inseq 的作用是，使 levels 按照数值大小排序'
# '官方文档:  by numeric value of level.'
# '此处选择一组为数字的列进行举例'
c<-as.matrix(gss_cat[1:100,'age'])
c<-as.factor(c)
# 原始
c
```

```
##   [1] 26 48 67 39 25 25 36 44 44 47 53 52 52 51 52 40 77 44 40 45 48 49 19 54 82
##  [26] 83 89 88 72 82 89 34 55 37 22 33 37 43 29 57 31 45 36 52 26 46 65 52 56 66
##  [51] 20 64 59 46 26 39 51 45 23 21 26 31 27 78 29 43 61 33 34 89 83 78 89 84 69
##  [76] 32 76 41 32 29 40 44 70 40 51 75 22 53 20 80 70 45 46 24 51 32 53 52 83 39
## 52 Levels: 19 20 21 22 23 24 25 26 27 29 31 32 33 34 36 37 39 40 41 43 ... 89
```

```
# 操作后
fct_inseq(c)
```

```
##   [1] 26 48 67 39 25 25 36 44 44 47 53 52 52 51 52 40 77 44 40 45 48 49 19 54 82
```

```
##  [26] 83 89 88 72 82 89 34 55 37 22 33 37 43 29 57 31 45 36 52 26 46 65 52 56 66
##  [51] 20 64 59 46 26 39 51 45 23 21 26 31 27 78 29 43 61 33 34 89 83 78 89 84 69
##  [76] 32 76 41 32 29 40 44 70 40 51 75 22 53 20 80 70 45 46 24 51 32 53 52 83 39
## 52 Levels: 19 20 21 22 23 24 25 26 27 29 31 32 33 34 36 37 39 40 41 43 ... 89
```

答：可以看到，进行对应操作后，内容没变，但是 level 发生了改变 levels 会编程对应函数操作后的结果

练习与作业 3：用 mouse genes 数据做图

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;
rm(list=ls())
library(readr)
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(ggplot2)
# '读取文件'
mouse_genes <- read_delim( file = "H:/第五学期/R-for-bioinformatics/data/talk04/mouse_ge
                           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_genes
```

```
## # A tibble: 138,532 x 6
##    `Gene stable ID`   `Transcript stable ID` `Protein stable ~ `Transcript leng~
##    <chr>              <chr>                  <chr>                          <dbl>
##  1 ENSMUSG00000064372 ENSMUST00000082423     <NA>                              67
##  2 ENSMUSG00000064371 ENSMUST00000082422     <NA>                              67
##  3 ENSMUSG00000064370 ENSMUST00000082421     ENSMUSP000000810~               1144
##  4 ENSMUSG00000064369 ENSMUST00000082420     <NA>                              69
##  5 ENSMUSG00000064368 ENSMUST00000082419     ENSMUSP000000810~                519
##  6 ENSMUSG00000064367 ENSMUST00000082418     ENSMUSP000000810~               1824
##  7 ENSMUSG00000064366 ENSMUST00000082417     <NA>                              71
##  8 ENSMUSG00000064365 ENSMUST00000082416     <NA>                              59
##  9 ENSMUSG00000064364 ENSMUST00000082415     <NA>                              67
```

```
## 10 ENSMUSG00000064363 ENSMUST00000082414     ENSMUSP000000810~              1378
## # ... with 138,522 more rows, and 2 more variables: Transcript type <chr>,
## #   Chromosome/scaffold name <chr>
```

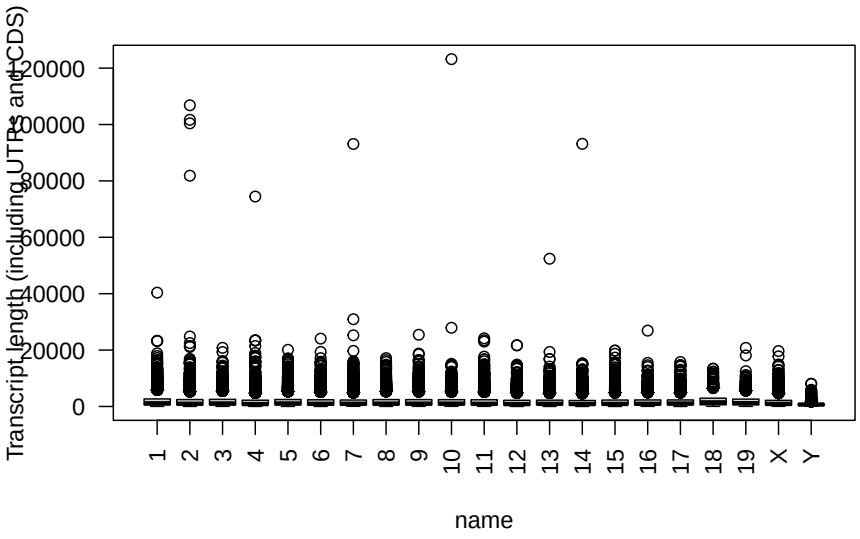
```
# '选出常染色体 1-19'
tar<-as.character(c(1:19))
mouse_genes_1_19<-mouse_genes %>% filter(`Chromosome/scaffold name` %in% tar)
mouse_genes_1_19
```

```
## # A tibble: 129,205 x 6
##    `Gene stable ID`   `Transcript stable ID` `Protein stable ~ `Transcript leng~
##    <chr>              <chr>                  <chr>                         <dbl>
##  1 ENSMUSG00000097062 ENSMUST00000181502     <NA>                            908
##  2 ENSMUSG00000097658 ENSMUST00000180595     <NA>                            933
##  3 ENSMUSG00000097294 ENSMUST00000181119     <NA>                           3683
##  4 ENSMUSG00000097020 ENSMUST00000180919     <NA>                           1457
##  5 ENSMUSG00000097289 ENSMUST00000180492     <NA>                           1004
##  6 ENSMUSG00000097289 ENSMUST00000181758     <NA>                           2493
##  7 ENSMUSG00000097176 ENSMUST00000180389     <NA>                            993
##  8 ENSMUSG00000096983 ENSMUST00000181900     <NA>                           1199
##  9 ENSMUSG00000097335 ENSMUST00000181152     <NA>                           1931
## 10 ENSMUSG00000097335 ENSMUST00000181003     <NA>                           1704
## # ... with 129,195 more rows, and 2 more variables: Transcript type <chr>,
## #   Chromosome/scaffold name <chr>
```

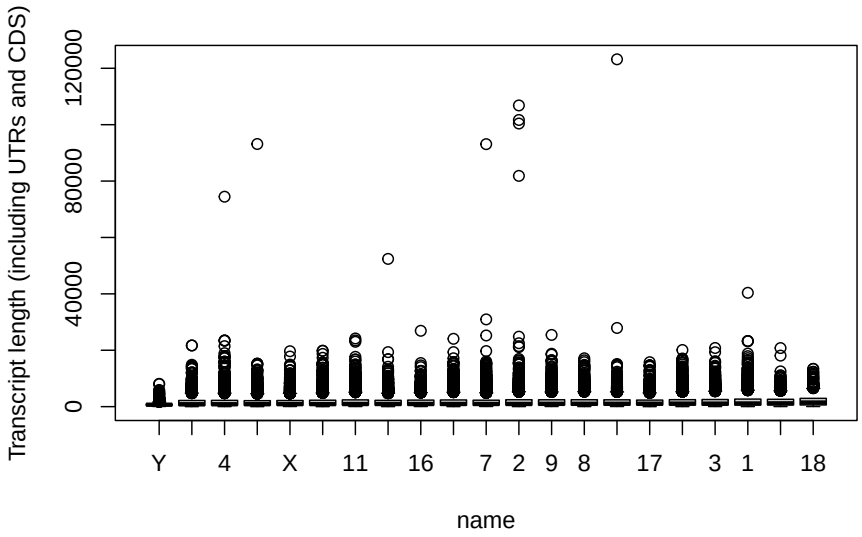
```
# '开始画图，先用 boxplot'
# '重新选染色体，因为包括 XY'
tar<-c(1:19,'X','Y')
mouse_genes_1_19_XY<-subset(mouse_genes,`Chromosome/scaffold name` %in% tar)

# '按染色体序号排列'
name<-factor(mouse_genes_1_19_XY$`Chromosome/scaffold name`,levels = tar)
plot_name<-boxplot(`Transcript length (including UTRs and CDS)`~
```

```
                        name,
                        data=mouse_genes_1_19_XY,las=2)
```



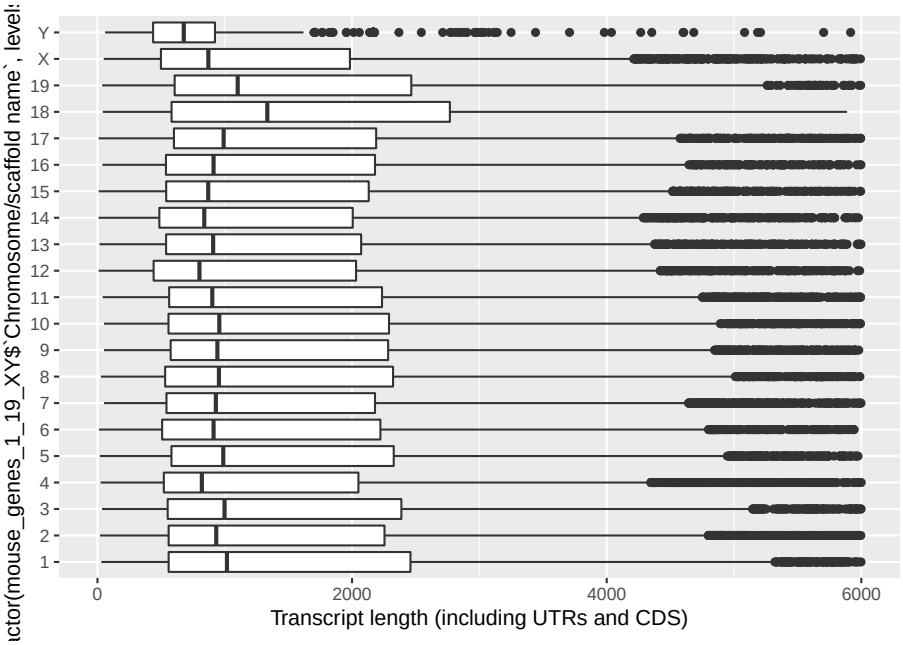

```
# '按染色体长度中值'
name=reorder(mouse_genes_1_19_XY$`Chromosome/scaffold name`,
             mouse_genes_1_19_XY$`Transcript length (including UTRs and CDS)`,
             median)
plot_length<-boxplot(`Transcript length (including UTRs and CDS)`~
                       name,
                       data=mouse_genes_1_19_XY)
```




```
# '用 ggplot 画图'
plot1<-ggplot(data=mouse_genes_1_19_XY,
              aes(x=factor(mouse_genes_1_19_XY$`Chromosome/scaffold name`,levels = tar)
                  y=`Transcript length (including UTRs and CDS)`))+
      geom_boxplot()+
      coord_flip()+
      ylim(0,6000)
plot1
```

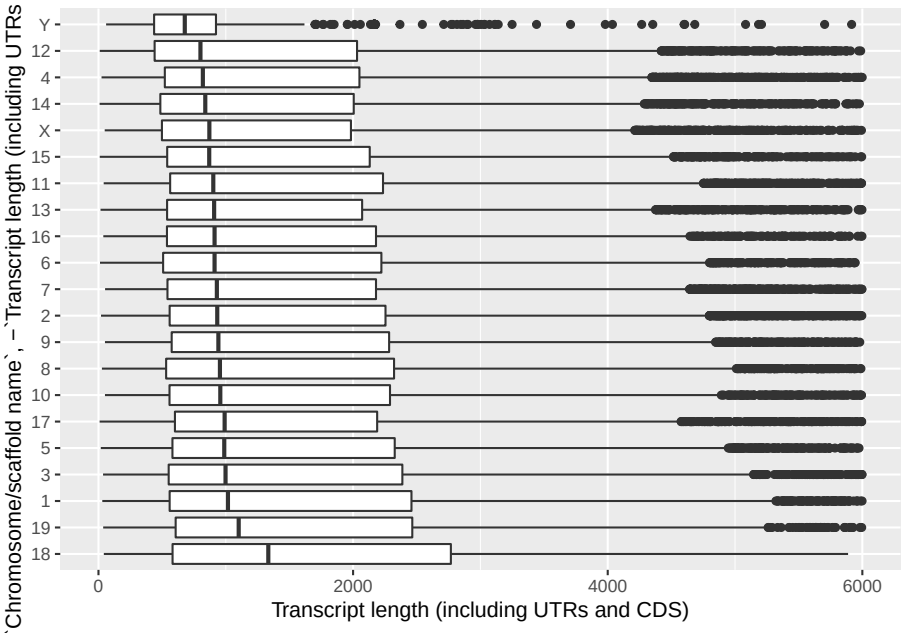
```
## Warning: Use of `mouse_genes_1_19_XY$`Chromosome/scaffold name`` is discouraged.
## Use `Chromosome/scaffold name` instead.
```

```
## Warning: Removed 3926 rows containing non-finite values (stat_boxplot).
```




```
plot2<-ggplot(data=mouse_genes_1_19_XY,
              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(0,6000)
plot2
```

```
## Warning: Removed 3926 rows containing non-finite values (stat_boxplot).
```


`Chromosome/scaffold name`, `Transcript length (including UTRs
Y
12
4
14
X
15
11
13
16
6
7
2
9
8
10
17
5
3
1
19
18
0
2000
4000
6000
Transcript length (including UTRs and CDS)