

Review of the course “R for Data Science” Part 01(Talk 01~ 04)

By Haoran Nie @ HUST Life ST

Partically translated by Rui Zhu @ HUST Life ST

双语版

Reference: R for Data Science

The book updated to 2nd ed. on July,2023, here's a [link](#) to the official website.

This work is licensed under CC BY-NC-SA 4.0 

Multi-omics data analysis and visualisation, #1

Talk 01

View the original slide through [this link](#).

View the original R markdown file of the slide through [this link](#).

This section has ~~nothing~~ something to explain :)

Install R

Go go <https://mirrors.tuna.tsinghua.edu.cn/CRAN/>(清华镜像)华科镜像(<https://mirror.s.hust.edu.cn/CRAN>)，R支持主流的操作系统包括Linux, Windows和MacOS, 请根据操作系统下载对应的安装文件。

新版本的Mac OS还需要安装XQuartz(<http://xquartz.macosforge.org/landing/>)。某些还需要用到Xcode, 可以从AppStore免费安装或者使用`xcode --select install`命令从控制台安装“XCode命令行工具”(体积更小更实用)。

目前大多Linux发行版都带有R, 因此可直接使用。从CRAN下载文件进行安装稍嫌复杂, 要求用户对Linux系统有一定的了解, 而且需要有管理员权限。建议初级用户在Linux高手指导下安装。点击"Download R forLinux"后, 发行版为Redhat (红帽) 或Suse的用户要先阅读网站上提供的readme或readme.html文件, 然后其中的指示进行安装。这里就不再累述了。

R studio

RStudio可以从<https://posit.co/downloads/> 下载, 支持等主流的操作系统。

R language basics, part 1

Talk 02

View the original slide through [this link](#).

View the original R markdown file of the slide through [this link](#).

基础数据类型

最基本的数据类型包括数字、逻辑符号和字符串，是其他数据类型的基本构建块。

数字

```
1 ## 整数  
2 287  
3  
4 ## 小数  
5 99.99  
6  
7 ## 科学计数法  
8 1e-3  
9 1e2
```

逻辑符号

```
1 TRUE  
2 T  
3  
4 FALSE  
5 F  
6  
7 ##其本质是数字  
8 1 + TRUE  
9 2 * FALSE
```

字符串

```

1 'a sentence' ## 单括号
2 "一个字符串" ## 双括号
3 '1.123'       ## 像是数字的字符串
4 '*%*()!@##&@(9' ## 乱码

```

简单数据类型

这包括向量和矩阵，它们都可以包含某种基本数据类型的多个值，如矩阵(**matrix**)由多个数字组成，**vector**由多个字符串组成，等等。但是，它们只能包含一种数据类型。

```

1 c(100, 20, 30) ## Interger vector
2 c("String", "Array", "It's me.") ## String vector
3 c(TRUE, FALSE, TRUE, T, F) ## A logic vector

```

如上所示，数组通常用函数 `c()`（又叫做 **concatenation function**）来定义。此外，包含连续整数向量可以使用：`运算符`来定义。

```
1 2:8
```

数据类型之间的转换

1. 自动转换

一个 `vector` 只能包含一种基本数据类型。因此，在定义数组时，如果输入值是混合的，某些基本数据类型会自动转换为其他类型，以确保数字类型的一致性；这在英语中被称为 **coerce**，具有强制转换的含义。此转换的优先级为：

- Logical types -> numeric types
- Logical Type -> String
- numeric type -> string
- 逻辑类型 -> 数字类型
- 逻辑类型 -> 字符串
- 数字类型 -> 字符串

vector 的数据类型转换规则

```
1 class( c(45, TRUE, 20, FALSE, -100) ); ## 逻辑和数字类型
2 str( c("string a", FALSE, "string b", TRUE) ); ## 逻辑和字符
3 str( c("a string", 1.2, "another string", 1e-3) ); ## 数字和字符
```

1. 手动切换

除了自动转换之外，还可以手动转换向量中元素的类型：

我们可以用 `class()` 或 `str()` 函数来判断 vector 包含的数据类型

```
1 class(matrix( c(20, 30.1, 2, 45.8, 23, 14),
2 nrow = 2, byrow = T ));
```

```
1 [1] "matrix" "array"
```

```
1 str(matrix( c(20, 30.1, 2, 45.8, 23, 14),
2 nrow = 2, byrow = T ));
```

```
1 num [1:2, 1:3] 20 45.8 30.1 23 2 14
```

- Checking the type of a variable `class()`
- Checking of classes `is.type()`
- Conversion of classes `as.type()`

一些特殊值

- `NA` (Not Available) missing values
- `NaN` (Not a Number) is meaningless
- `-Inf` Negative Infinity

- `Inf` Positive Infinity
- `NULL` Null

Some functions to determine these special values:

- `is.na()`
- `is.finite()`
- `is.infinite()`

Vectors and Matrix

都是数组。A `vector` is a one-dimensional array and a matrix is a two-dimensional array.

This means.

- There can be more dimensional arrays
- 高维数组， like `vector` and matrices, 只能包含一种基本数据类型。
- Higher dimensional arrays can be defined by the `array()` function.

矩阵由函数 `matrix()` 定义，比如：

```
1 matrix( c(20, 30.1, 2, 45.8, 23, 14), nrow = 2, byrow = T );
```

矩阵的指定长度，即 `nrow × ncol`，可以不同于输入数据的长度。矩阵长度较小时，输入数据会被截短；而矩阵长度较大时，输入数据则会被重复使用。

```
1 ## 生成一个2x5长度为10的矩阵，但输入数据的长度为20
2 matrix( 1:20, nrow = 2, ncol = 5, byrow = T );
3
4 ## 生成一个2x3长度为6的矩阵，但输入数据长度只有3
5 matrix( 1:3, nrow = 2, ncol = 3, byrow = T );
```

下面两种情况，系统会报警告信息。

第一种情况，矩阵长度大于输入数据长度，且前者不是后者的整数倍。

```
1 matrix( 1:3, nrow = 2, ncol = 4, byrow = T );
```

第二种情况，矩阵长度小于输入数据的长度，且后者不是前者的整数倍。

```
1 matrix( letters[1:20], nrow = 3, ncol = 5, byrow = T );
```

加减乘除逻辑运算老一套

```
1 1 + 2 - 3 * 4 / 5; ## 加减乘除
2 1 + (2 - 3) * 4 / 5; ## 改变优先级
3 2 ^ 6; ## 阶乘
4 5 %% 2; ## 取余
5 T | F; ## or
6 T & F; ## and
7 5 | 0; ## == 0 FALSE, != 0 TRUE, 任何非零值视为逻辑真
```

通过 Console window 管理变量

```
1 ls(); ## 显示当前环境下所有变量
2 rm( x ); ## 删除一个变量
3 ls();
4
5 ##rm(list=ls()); ## 删除当前环境下所有变量！！！
```

vector 算术 vectorisation: R最重要的一个概念

```
1 x <- c(10,100,1000, 10000);
2 ( y <- sqrt( x ) * 4 + 10 ); ## 赋值的之后打印变量内容
```

核心在于数据自动循环使用

```

1 x / c(10,100);
2 [1] 1 100 100
3 x / c(10,100,1000); ## 会报错但仍会循环计算
4 Warning: 长的对象长度不是短的对象长度的整倍数[1] 1 1 1000

```

matrix 算术

```

1 m <- matrix( c(20, 30.1, 2, 45.8, 23, 14), nrow = 2,
2           dimnames = list( c("row_A", "row_B"), c("A", "B", "C") ) )
3 A B C
4 row_A 20.0 2.0 23
5 row_B 30.1 45.8 14
6
7 m / 10;
8 A     B     C
9 row_A 2.00 0.20 2.3
10 row_B 3.01 4.58 1.4
11
12 m / c(1,10,100);
13 A     B     C
14 row_A 20.00 0.02 2.30
15 row_B 3.01 45.80 0.14

```

更多 matrix 相关函数

```

1 dim(m);
2 nrow(m);
3 ncol(m);
4 range(m); ## Available when the content is numeric
5 summary(m); ## Can also be used in vector

```

Extra:

- Incorporation

```

1 a <- 1:3;
2 b <- LETTERS[1:3];
3
4 ( ab <- c(a,b) );
5 mode( ab ); ## 一个新的函数~ ...

```

- Take part `ab[1]`
- Replacement of individual values `ab[1] = c`
- Replacing multiple values `ab[c(2, 3)] = c("Weihua", "Chen")`
- Naming elements and replace values `names(ab) = as.character(ab)`
- Reverse `rev(1:10)`
- Sort&order

```

1 lts = sample(LETTERS[1:20])
2 sort(lts)

```

- 提取一行或多行

```

1 # (There's already some data in workspace)
2
3 $ m
4 > (List the content of matrix "m")
5
6 $ m[1, ]
7 > (List the first row of matrix 'm')
8
9 $ m[1:2, ]
10 > (List the first two rows of matrix 'm')

```

You can also let the console to fetch multiple lines as the order you give.

```

1 m[c("row_B", "row_A")]

```

控制台将以“row_B”和“row_A”的顺序输出矩阵“m”的内容。

- 获取一列或多列

As can be seen from the same principle, I only list codes here

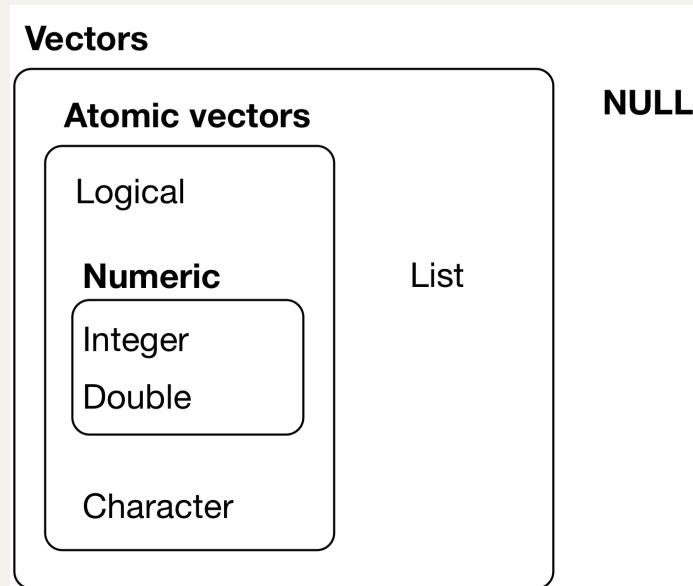
```
1 m[ , 1]
2 m[ , c(1:2)]
3 m[ , c("col_B", "col_A")]
```

- Fetch parts `m[1:2, 2:3]`
- Replacement

```
1 m[1, ] = c(10)
2 m[, "C"] = c(230, 140)
3 m[1:2,] = matrix(1:6, nrow=2)
4 m[1, c("C", "B")] = matrix(110:111, nrow = 1)
```

- 转置 `t(m)`

The hierarchy of R's vector types



You can use function `typeof()` to know the type of a vector.

Here are some examples of other `is.xxx()` function:

```
1 is.null( NULL )
2 is.numeric( NA )
3 is.numeric( Inf );
4 ## 用于替代 typeof 的函数
5 is.list();
6 is.logical();
7 is.character();
8 is.vector();
9 # more ...
```

R language basics, part 2

Talk 03

View the original slide through [this link](#).

View the original R markdown file of the slide through [this link](#).

data.frame

What is a **data.frame**?

- 二维表格
- 由不同列组成；每列是一个**vector**，不同列的数据类型可以不同，但一列只包括一种数据类型 (int, num, chr ...)
- 各列的长度相同

```

1 library(tidyverse);
2 library(kableExtra)
3 kbl(head(mpg),
4     booktabs = T)

```

Here's the result:

manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact
audi	a4	2.0	2008	4	manual(m6)	f	20	31	p	compact
audi	a4	2.0	2008	4	auto(av)	f	21	30	p	compact
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact

Usage of `head()` and `tail()`

```
1 nrow(mpg); ## total number of rows  
2 kbl( head(mpg, n=3), booktabs = T); ## 显示前几行数据  
3 kbl( tail(mpg, n=3), booktabs = T); ## 显示最后3行数据
```

- `head()` is a function to display the first rows of some data (vectors etc.)
 - `tail()` is a function to display the last rows of some data (vectors etc.)

Structure of `data.frame` & `tibble`

```
str(mpg)
```

This command shows the structure of the tibble `mpg`:

```
## # tibble [234 x 11] (S3: tbl_df/tbl/data.frame)
## # $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
## # $ model      : chr [1:234] "a4" "a4" "a4" "a4" ...
## # $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
## # $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
## # $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
## # $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
## # $ drv          : chr [1:234] "f" "f" "f" "f" ...
## # $ cty          : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
## # $ hwy          : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
## # $ fl           : chr [1:234] "p" "p" "p" "p" ...
## # $ class        : chr [1:234] "compact" "compact" "compact" "compact" ...
```

Make a new `data.frame`

You can use the function `data.frame()` to make a new `data.frame`

```
1 data2 =  
2   data.frame(  
3     data = sample(1:100, 10),  
4     group = sample(LETTERS[1:3], 10, replace = TRUE)  
5     data2 = 0.1  
6   )
```

How to add row(s)/col(s) to an existing data.frame

Create the "table header" first, then populate the `data.frame`

先创建"表头", 再填充

```

1 df2 =
2   data.frame(
3     x = character(),
4     y = integer(),
5     z = double() ,
6     stringsAsFactors = FALSE
7   )
8
9 df2 =
10 rbind(
11   df2,
12   data.frame(
13     x = "a",
14     y = 1L,
15     z = 2.2
16   )
17 )
18
19 df2 =
20 rbind(
21   df2,
22   data.frame(
23     x = "b",
24     y = 2,
25     z = 4.4
26   )
27 )
28
29 df2
30   x y   z
31 1 a 1 2.2
32 2 b 2 4.4
33

```

```

34 m <- cbind(1, 1:7) ; ## 产生两列数据 7行数据 ...
35 ( m <- cbind(m, 8:14) ); ## 增加一列 也有7行数据 ...
36      [,1] [,2] [,3]
37 [1,]    1    1    8
38 [2,]    1    2    9
39 [3,]    1    3   10
40 [4,]    1    4   11
41 [5,]    1    5   12
42 [6,]    1    6   13
43 [7,]    1    7   14

```

ATTENTION

- Use `rbind()` function to add rows, use `cbind()` function to add columns.
- Define the new line using `data.frame()` function, the "header" needs to be the same as the merged table.

You can also use these functions to bind several data.frames.

tibble

`tibble` is kind of similar to `data.frame`.

注: Tibble class 是 data.frame 的升级版本;

Make new tibble

`tibble` 相关功能由 `tibble` 或 `tidyverse` 包提供

Almost all of the functions that you'll use in this book produce tibbles, as tibbles are one of the unifying features of the tidyverse. Most other R packages use regular data frames, so you might want to coerce a data frame to a tibble. You can do that with `as_tibble()`:

```

1 as_tibble(iris)
2 #> # A tibble: 150 × 5
3 #>   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
4 #>     <dbl>      <dbl>      <dbl>      <dbl> <fct>
5 #> 1     5.1        3.5        1.4       0.2 setosa
6 #> 2     4.9        3          1.4       0.2 setosa
7 #> 3     4.7        3.2        1.3       0.2 setosa
8 #> 4     4.6        3.1        1.5       0.2 setosa
9 #> 5     5          3.6        1.4       0.2 setosa
10 #> 6    5.4        3.9        1.7       0.4 setosa
11 #> # i 144 more rows

```

Another way to create a tibble is with `tribble()`, short for **transposed tibble**.

`tribble()` is customised for data entry in code: 列标题由公式定义（即以“~”开头），条目由逗号分隔。这使得以易于阅读的形式布置少量数据成为可能。

也可以用 `tibble` 函数创建

- 注意每列的数据类型
- 长度不足时，比如**data2**列，会循环使用
- `sample()`函数的用法

```

1 tribble(
2   ~x, ~y, ~z,
3   #--| --| ----
4   "a", 2, 3.6,
5   "b", 1, 8.5
6 )
7 #> # A tibble: 2 × 3
8 #>   x         y     z
9 #>   <chr> <dbl> <dbl>
10 #> 1 a         2     3.6
11 #> 2 b         1     8.5
12
13 ## 用 tibble 函数创建，用法和 data.frame() 相似
14 dat <-
15   tibble( data = sample( 1:100, 10 ), ,

```

```

16         group = sample( LETTERS[1:3], 10, replace = TRUE),
17         data2 = 0.1 )
18 # A tibble: 10 × 3
19   data group data2
20   <int> <chr> <dbl>
21     1     86 A     0.1
22     2     41 B     0.1
23     3     72 B     0.1
24     4     68 C     0.1
25     5     10 A     0.1
26     6     99 B     0.1
27     7     66 B     0.1
28     8     33 B     0.1
29     9     81 A     0.1
30    10     42 C     0.1

```

- `add_row()`
- `add_column()`

```

1 ## 新tibble, with defined columns ... 创建表头
2 tb <- tibble( x = character(), y = integer(), z = double() );
3 dim(tb);
4
5 ## 增加行 ...
6 tb <- add_row( tb, x = "a", y = 2, z = 3.6 );
7 tb <- add_row( tb, x = "b", y = 1, z = 8.5 );
8
9 ## 显示
10 tb;
11
12 ## 生成一个 tibble
13 df <- tibble(x = 1:3, y = 3:1);
14
15 # 在第二行之前插入
16 df <- add_row(df, x = 4, y = 0, .before = 2);
17
18 ## 插入多行

```

```

19 df <- add_row(df, x = 4:5, y = 0:-1);
20
21 ## 插入另一个tibble (与另一个tibble合并)
22 df2 <- tibble( x = as.double(200:202), y = as.double(1000:1002)
23 );
24 df3 <- add_row( df, df2 ); ## 可以运行 ...
25
26 tb3 <- tribble(
27   ~x, ~y, ~z,
28   "a", 2, 3.6,
29   "b", 1, 8.5
30 );
31 tb3 <- add_column( tb3, a = 98 ); ## recycle ...
32 tb3 <- add_column( tb3, b = LETTERS[1:2], c = c("CHEN", "WANG")
33 );

```

tibble元素替换

```

1 ## 取得行
2 tb3[c(1,2), ];
3
4 ## 取得列, 按顺序取列
5 tb3[, c("z", "y")];
6
7 ## 替换列
8 tb3[["z"]] <- c(4.6, 5.5);
9
10 ## 替换行
11 tb3[ 1, ] <- tibble( x = "d", y = 20, z = 46, a = 10, b = "C", c
= "LILI" );

```

Manipulate the tibble

See “Manipulate the `data.frame`”

tibble to data.frame

- `as.data.frame()`
- `as_tibble()`

e.g.

```
1 library(tibble)
2 as.data.frame(head(as_tibble(iris)))
```

Differences between tibble and data.frame

Tibble 按顺序计算列

```
1 rm(x,y) # Delete possible x, y
2 tibble(x = 1:5, y = x^2); # You can do this with tibble
3 data.frame(x = 1:5, y = x ^ 2); # But data.frame doesn't work.
```

`data.frame` causes trouble when fetching `subset` operations

取子集为vector

```
1 df1 =
2   data.frame(x = 1:3, y = 3:1)
3 class(df1[, 1:2])
4
5 #> [1] "data.frame"
6
7 # Subset operation : takes a column and expects a data.frame ()
8 class(df1[, 1]) # The result is a vector ...
9
10 #> [1] "integer"
11
12 ## Tibble doesn't.
13 df2 =
14   tibble(x = 1:3, y = 3:1)
```

```

15 class(df2[, 1]) ## 永远都是 tibble
16
17 #> [1] "tbl_df" "tbl" "data.frame"

```

`tibble` allows controlled data type conversion

```

1 class(df2[[1]]); ## 取一列, 转换为 vector
2 class(df2$x); ## 用 [[ ]] 或 $ 都可以哦
3
4 [1] "numeric"
5 [1] "numeric"

```

Recycling

```
1 data.frame(a = 1:6, b = LETTERS[1:2]); ## data.frame 可以! ! !
```

OUTPUT

```

1 #   a b
2 # 1 1 A
3 # 2 2 B
4 # 3 3 A
5 # 4 4 B
6 # 5 5 A
7 # 6 6 B

```

```
1 tibble(a = 1:6, b = LETTERS[1:2]); ## 但 tibble 不行! ! !
```

OUTPUT

```

1 # Error:
2 # ! Tibble columns must have compatible sizes. ## * Size 6:
# Existing data.
3 # * Size 2: Column `b`.
4 # i Only values of size one are recycled.

```

ATTENTION!

`tibble` 的 recycling 仅限于长度为1或等长；而 `data.frame` 则为整除即可。

`data.frame` will do partial matching, while `tibble` will **NEVER** do it.

```
1 df = data.frame(abc = 1)
2 df$ab; # Unwanted result ...
3
4 df2 = tibble(abc = 1)
5 df2$a; # Produce a warning and return NULL
```

OUTPUT

```
1 #[1] 1
2 #Warning: Unknown or uninitialized column: `a`. NULL
```

Advanced tips for using `data.frame` and `tibble`

- `attach()`
- `detach()`
- `with()`
- `within()`

Following is the introduction (Produced by ChatGPT)

这些函数——`attach ()`、`detach ()`、`with ()` 和`within ()` ——在处理R中的数据帧或tibble时非常有用，有助于更流畅的工作流和代码可读性。下面是它们的功能分解：

attach() and detach()

- **Purpose:** 这些函数允许您将data.frame临时附加到搜索路径上，使其列可以通过名称直接访问。
- **Usage:**
 - `attach(df)` attaches the specified data frame `df`.
 - `detach(df)` detaches the specified data frame `df`.
- **Example:**

```

1 data(mtcars) # Loading a sample dataset
2 attach(mtcars) # Attaching mtcars
3
4 # Now, columns can be accessed directly
5 summary(mpg)
6 mean(mpg)
7
8 detach(mtcars) # Detaching mtcars

```

- **Note:** 虽然使用`attach()`很方便，但有时会导致混乱或意外的后果，例如屏蔽环境中的变量。由于潜在的副作用，通常建议避免使用`attach()`。

with()

- **Purpose:** `with()` 允许您在不使用`$`的情况下引用数据框架的列的环境中执行表达式。
- **Usage:**
 - `with(data, expr)` evaluates `expr` in the context of the specified data frame `data`.
- **Example:**

```

1 data(mtcars) # Loading a sample dataset
2
3 with(mtcars, {
4   mean(mpg)
5   summary(cyl)
6 })

```

- **Advantage:** 它有助于避免在处理其列时重复使用数据框架名称。

within()

- **Purpose:** Similar to `with()`, `within()` allows modification of a data frame by calculating the expressions within it.
- **Usage:**
 - `within(data, expr)` modifies `data` according to `expr` and returns the modified data frame.
- **Example:**

```

1 data(mtcars) # Loading a sample dataset
2
3 modified_mtcars <- within(mtcars, {
4   mpg_square <- mpg^2
5   hp_doubled <- hp * 2
6 })
7 head(modified_mtcars) # Checking the modified data
frame

```

- **Advantage:** `within()` is very useful when you want to create or modify data frame columns without repeatedly referring to the data frame name.

Remember, while these functions can streamline your code, it's crucial to use them judiciously to avoid unexpected behavior or cluttering your global environment.

Here's the console print output for the examples provided earlier:

```

1 # Using attach() and detach()
2 data(mtcars) # Loading a sample dataset
3 attach(mtcars) # Attaching mtcars
4
5 # Now, columns can be accessed directly
6 summary(mpg)
7 # Output:
8 #   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
9 # 10.40 15.43 19.20 20.09 22.80 33.90
10
11 mean(mpg)
12 # Output:
13 # [1] 20.09062
14
15 detach(mtcars) # Detaching mtcars
16
17 # Using with()
18 data(mtcars) # Loading a sample dataset
19
20 with(mtcars, {
21   mean(mpg)
22   # Output:
23   # [1] 20.09062
24
25   summary(cyl)
26   # Output:
27   #   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
28   # 4.00 4.00 6.00 6.188 8.00 8.00
29 })
30
31 # Using within()
32 data(mtcars) # Loading a sample dataset
33
34 modified_mtcars <- within(mtcars, {
35   mpg_square <- mpg^2
36   hp_doubled <- hp * 2
37 })
38 head(modified_mtcars) # Checking the modified data frame
39 # Output:

```

```

40 #           mpg cyl disp  hp drat    wt  qsec vs am
gear carb mpg_square hp_doubled
41 # Mazda RX4          21.0   6 160 110 3.90 2.620 16.46 0  1
4     4      441.00       220
42 # Mazda RX4 Wag     21.0   6 160 110 3.90 2.875 17.02 0  1
4     4      441.00       220
43 # Datsun 710        22.8   4 108  93 3.85 2.320 18.61 1  1
4     1      519.84       186
44 # Hornet 4 Drive   21.4   6 258 110 3.08 3.215 19.44 1  0
4     3      457.96       220
45 # Hornet Sportabout 18.7   8 360 175 3.15 3.440 17.02 0  0
4     3      349.69       350
46 # Valiant           18.1   6 225 105 2.76 3.460 20.22 1  0
4     3      327.61       210

```

File IO

Read from files

Using functions from the `readr` package

```

1 # readr is part of tidyverse
2 library(tidyverse) # or alternatively
3 library(readr)

```

Some available functions:

- `read_csv()`: comma separated (CSV) files
- `read_tsv()`: tab separated files
- `read_delim()`: general delimited files
- `read_fwf()`: fixed width files
- `read_table()`: tabular files where columns are separated by white-space.
- `read_log()`: web log files

Full documentation of the package is available through this [link](#).

Usage

- Read with predefined column types

```

1 myiris2 =
2   read_csv("../data/talk03/iris.csv",
3           col_types =
4             cols(
5               Sepal.Length = col_double(),
6               Sepal.Width = col_double(),
7               Petal.Length = col_double(),
8               Petal.Width = col_double(),
9               Species = col_character()
10              )
11            )

```

- To read from other formats, you can try the following packages:

Similar to Python

- `haven` - SPSS, Stata, and SAS files
- `readxl` - excel files (.xls and .xlsx) DBI - databases
- `jsonlite` - json
- `xml2` - XML
- `httr` - Web APIs
- `rvest` - HTML (Web Scraping)

Write to files

Use the following functions to write object(s) to external files:

Default parameters are listed.

More related documents can be found in this [link](#).

- Comma delimited file: 逗号分隔文件

```

1  write_csv(
2    x,
3    path,
4    na = "NA",
5    append = FALSE,
6    col_names = !append
7  )

```

- File with arbitrary delimiter: 帶有任意分隔符的文件

```

1  write_delim(
2    x,
3    path,
4    delim = " ",
5    na = "NA",
6    append = FALSE,
7    col_names = !append
8  )

```

- CSV for excel:

```

1  write_excel_csv(
2    x,
3    path,
4    na = "NA",
5    append = FALSE,
6    col_names = !append
7  )

```

- String to file:

```

1  write_file(
2    x,
3    path,
4    append = FALSE
5  )

```

- String vector to file, one element per line:

```

1  write_lines(
2    x,
3    path,
4    na = "NA",
5    append = FALSE
6  )

```

- Object to RDS file:

```

1  write_rds(
2    x,
3    path,
4    compress =
5      c(
6        "none",
7        "gz",
8        "bz2",
9        "xz"
10       ),
11     ...
12   )

```

- Tab delimited files:

```

1  write_tsv(
2    x,
3    path,
4    na = "NA",
5    append = FALSE,
6    col_names = !append
7  )

```

R language basics, part 3: factor

Talk 04

View the original slide through [this link](#).

View the original R markdown file of the slide through [this link](#).

IO and working environment management

Each R session is a separate **work space** containing its own data, variables, and operation history.

```
wchen — R session 1 — R — 70x22
[wchen @mbp: 1] ~ > R
WARNING: ignoring environment value of R_HOME

R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 

wchen — R session 2 — R — 74x22
[wchen @mbp: 1] ~ > R
WARNING: ignoring environment value of R_HOME

R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

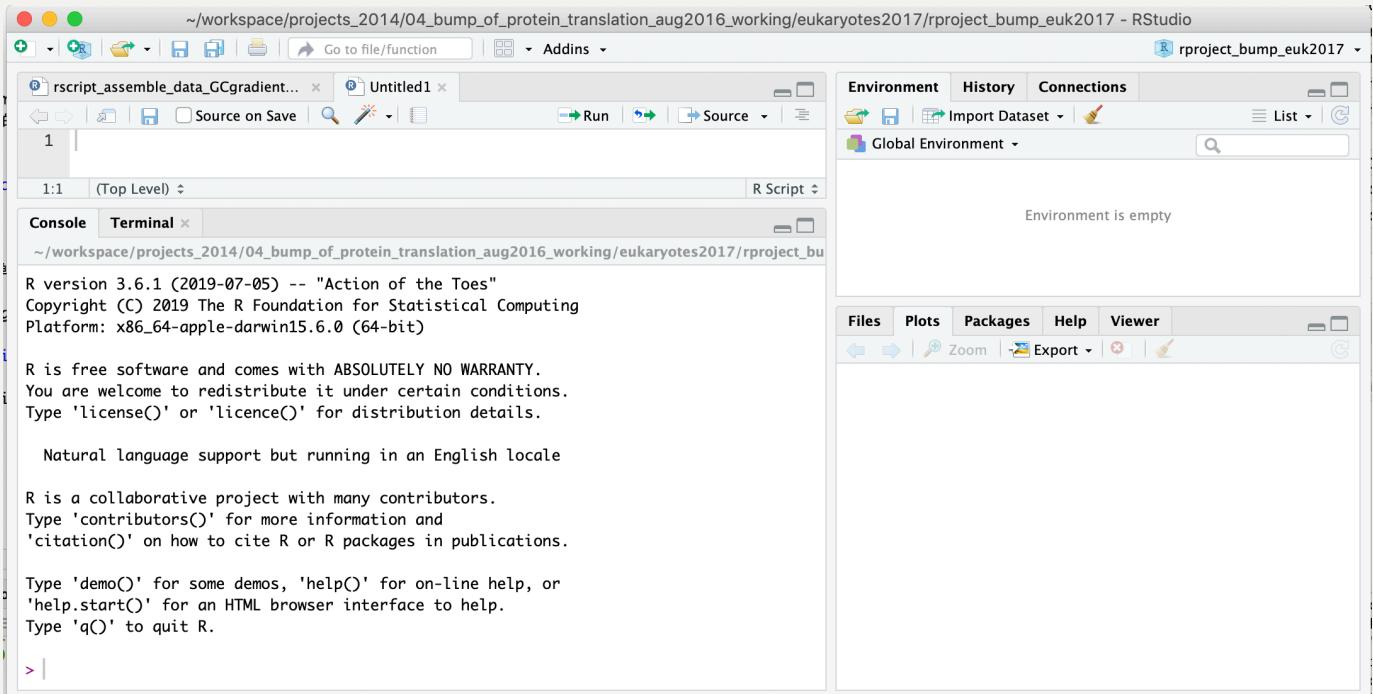
> 
```

Each RStudio session is automatically associated with a R session

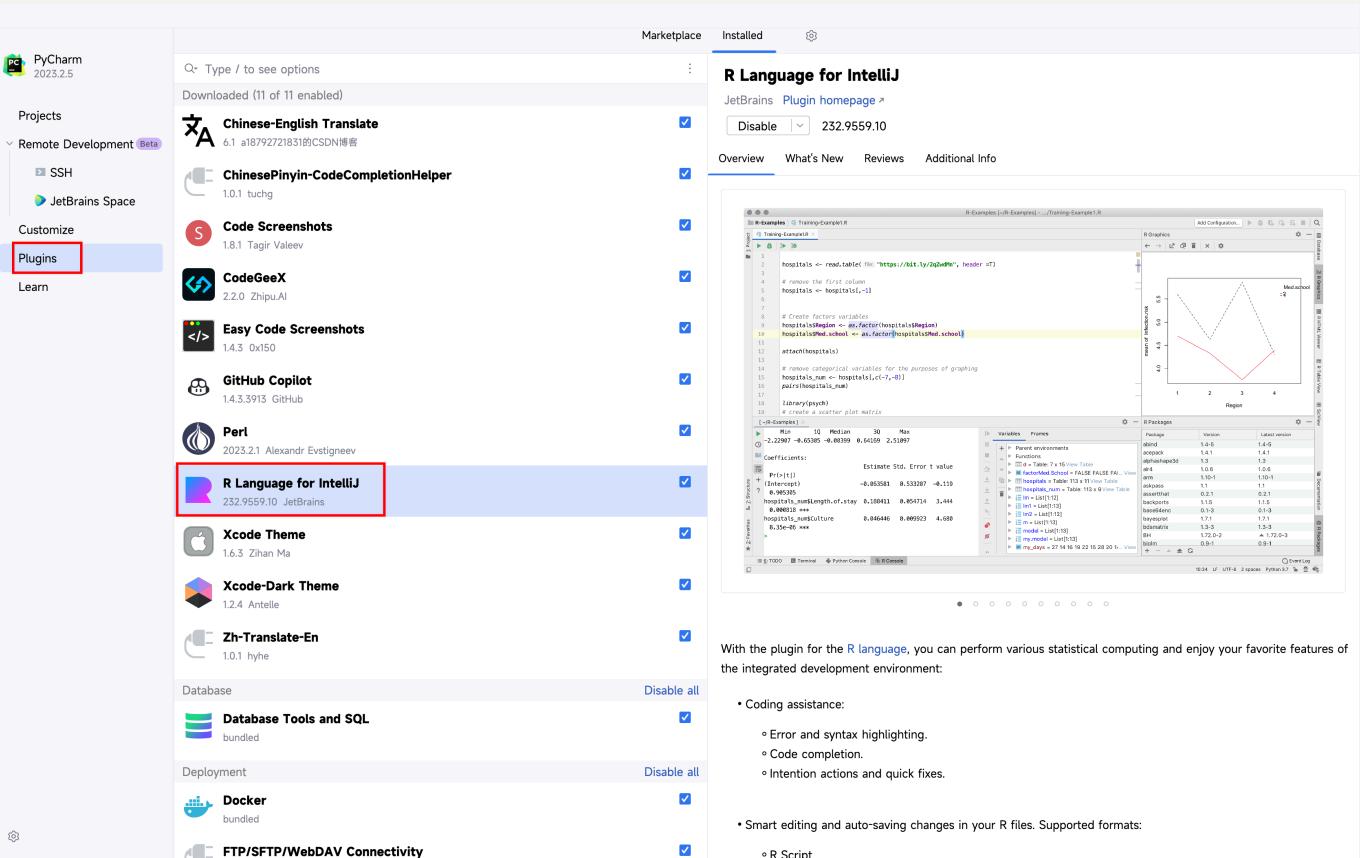
Not only RStudio, PyCharm or VSCode also support R session.

However, I'm keen on coding with PyCharm but not RStudio, for its wonderful Plug-in Environment, which can let me use plug-ins such as Code GeeX by Zhipu AI (a company founded by some student in KEG team in Tsinghua University) or GitHub Copilot by GitHub to let the coding process more quickly, for the instruction from GPTs.

tips: Rstudio最新版也支持 Github Copilot, 可以在 options 内设置



If you want to coding with R using PyCharm or other JetBrains IDE (i.e. IntelliJ, CLion, etc.), remember to install the *R Language Plug-in*



For instruction how to get FREE Student Liscence of GitHub Pro, GitHub Copilot and JetBrains Products and their benefits, see their official website:

- GitHub Global Campus

Make sure you don't use VPNs and use your phone to log in and apply (HUST Campus Network is recommended), give "Precise Location" permission to your browser. You may use your "[Student Number]@hust.edu.cn" mail to verify your identity as a student studying in HUST.

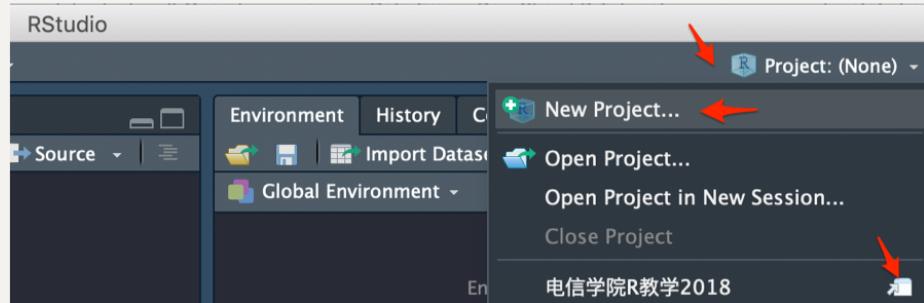
- JetBrains Products

Because our email addresses ending with "@hust.edu.cn" are banned due to misuse, you should apply for an online verification report on [CHSI](#) (press the link to visit the website), instructions [here](#).

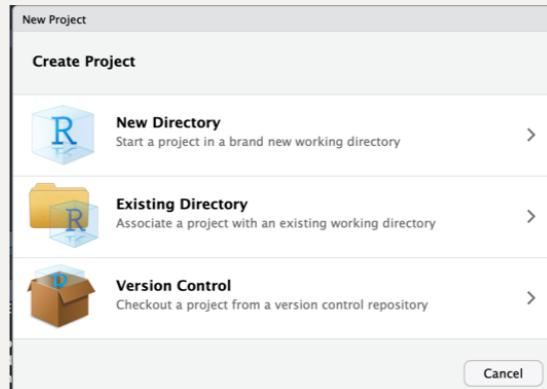
Start a new RStudio session by creating a new project

To start a new session in PyCharm, simply press the bottom corner and select a new session.

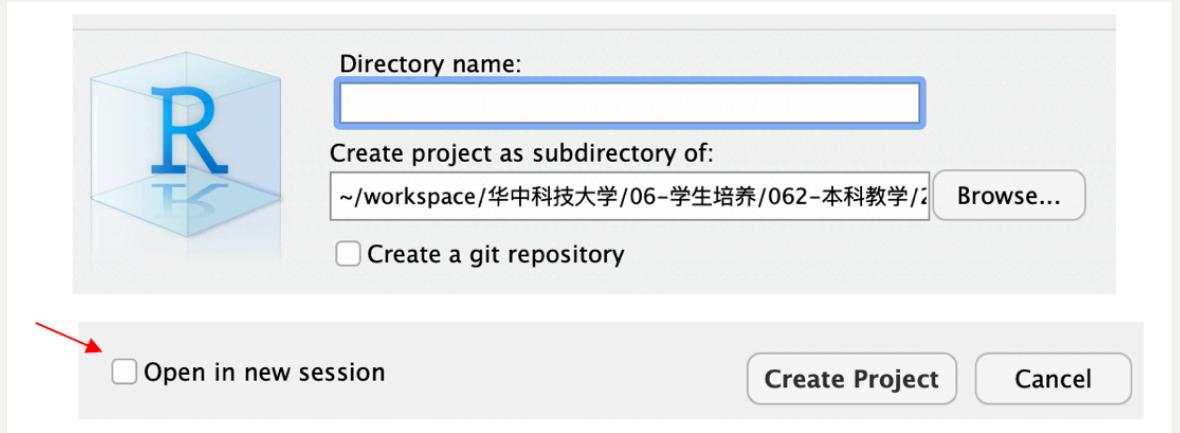
- Click the Project button in the upper right corner and select New Project in the pop-up menu ...



- Select: New directory -> New Project in the popup window



- Enter a new directory name, choose its mother directory ...



Working Space

当前工作空间，包括所有已装入的数据、包和自制函数

可通过以下代码管理变量

```

1 ls(); ## 显示当前环境下所有变量
2 rm( x ); ## 删除一个变量
3 ls();
4
5 ##rm(list=ls()); ## 删除当前环境下所有变量！！！

```

Variables in working space in RStudio

The "Environment" window in the upper right corner of RStudio shows all the variables of the current workspace.

Data

▶ aq	153 obs. of 7 variables	grid icon
▶ dat	10 obs. of 3 variables	grid icon
▶ dat2	10 obs. of 3 variables	grid icon
▶ df	1 obs. of 1 variable	grid icon
▶ df2	1 obs. of 1 variable	grid icon
▶ l	List of 26	search icon
m	num [1:2, 1:3] 1 2 111 103 110 101	grid icon
▶ myiris	150 obs. of 5 variables	grid icon
▶ myiris2	150 obs. of 5 variables	grid icon

Values

a	int [1:3] 1 2 3
ab	Named chr [1:6] "1" "—" "≡" "ah" "bo" "C"
b	chr [1:3] "A" "B" "C"
lts	chr [1:20] "P" "N" "I" "S" "Q" "D" "C" "A" "J" "M" "E" "F" "H"...

Save and restore work space

```

1 # Save all loaded variables into an external .RData file
2 save.image(file = "prj_r_for_bioinformatics_aug3_2019.RData")
3 # Restore (load) saved work space
4 load(file = "prj_r_for_bioinformatics_aug3_2019.RData")

```

Notes:

- Existing variables will be kept, however, those with the same names will be replaced by loaded variables
- Please consider using `rm(list=ls())` to remove all existing variables to have a clean start
- You may need to reload all the packages

Save selected variables

Sometimes you need to transfer processed data to a collaborator ...

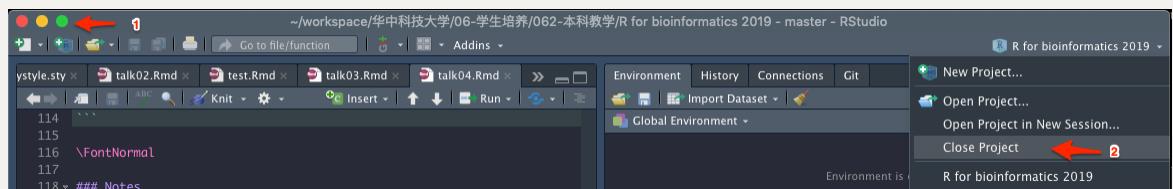
```

1 # Save selected variables to external
2 save(
3   city,
4   country,
5   file="1.RData"
6 )
7 # You can specify directory name
8 load("1.RData")

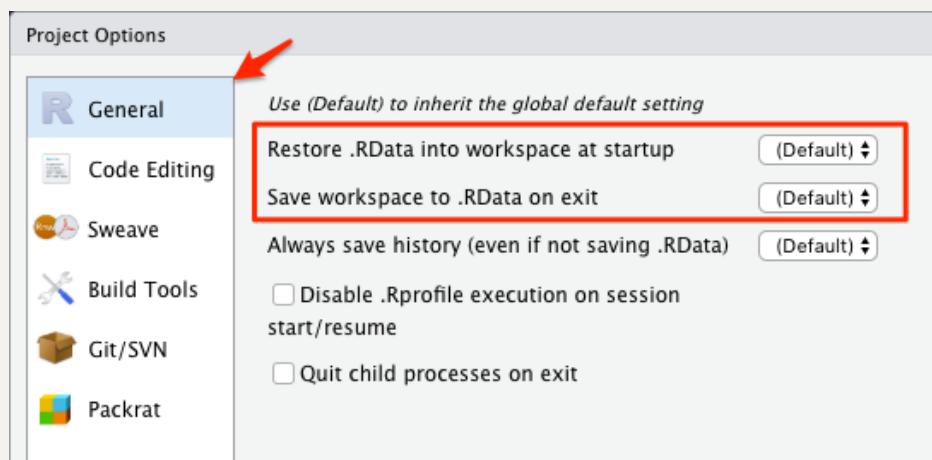
```

Close and (re)open a project

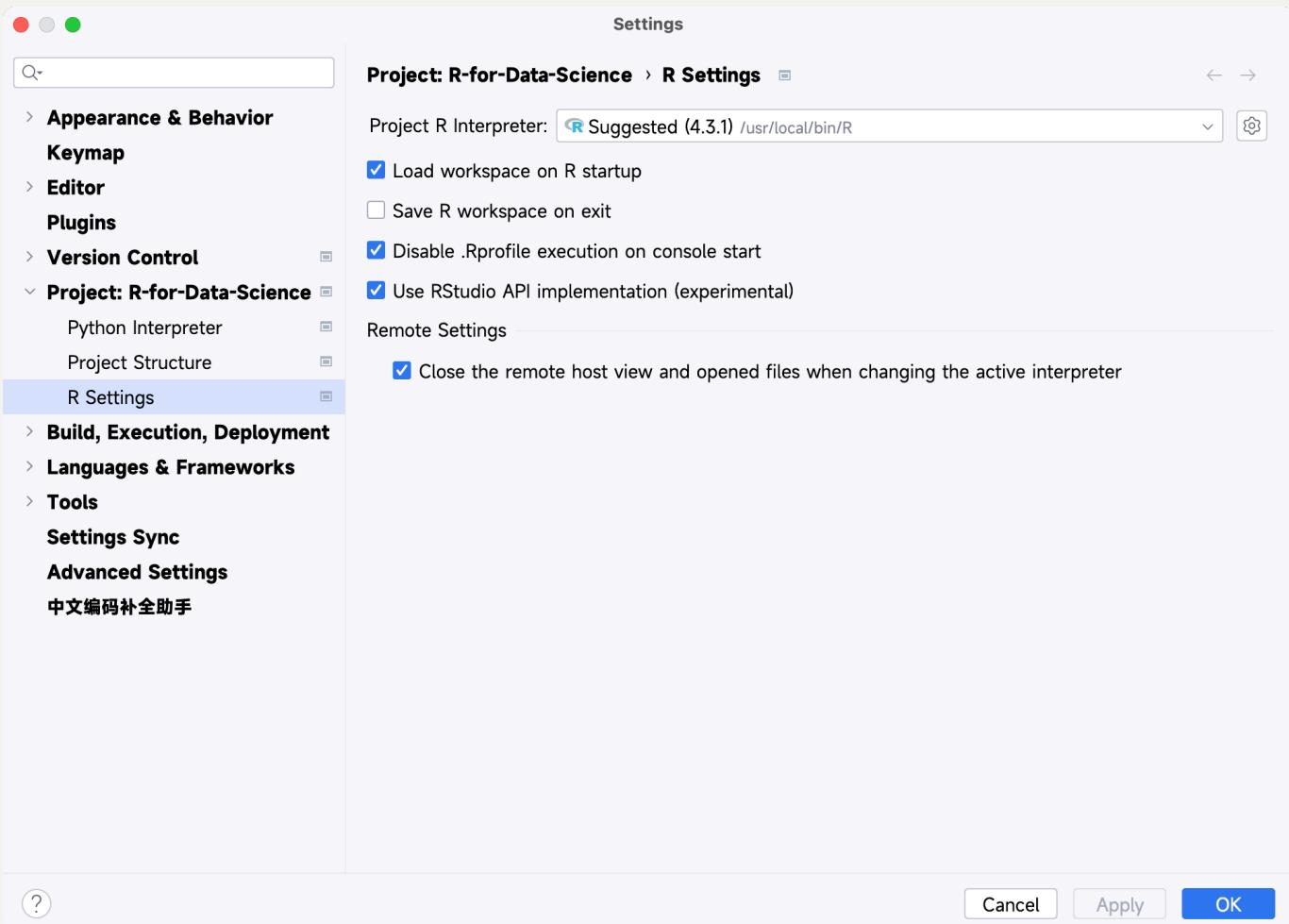
- To close a project



- In RStudio and similar IDEs, there are some preferences to choose



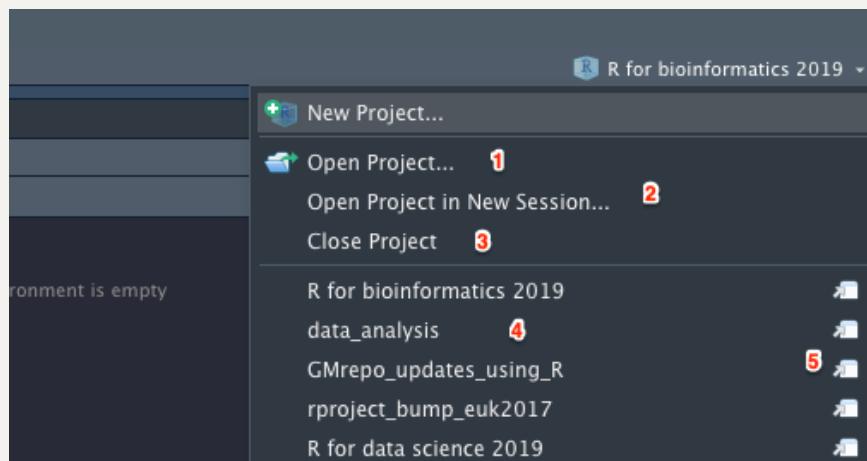
The UI in PyCharm



Notes:

- Save on exit
- Load on opening
- When the data is large, the loading time may be too long ...

Open a project



When in PyCharm, simply drag the working directory to its main window, remember to trust the project.

Factors

Factor是一种用于字段的数据结构，它只接受预定义的、有限数量的值（分类数据）。它将限制输入数据的选取。

```
1 ## create factor from scratch ...
2 x <- factor( c( "single", "married", "married", "single" ) );
3
4 ## create factor as it is ...
5 x <- c("single", "married", "married", "single");
6 x <- as.factor(x);
7
8 ## please note the change in the displayed values ...
9 str(x);
10 #Factor w/ 2 levels "married","single": 2 1 1 2
11
12 #限制输入数据的选择范围
13 x[ length(x) + 1 ] <- "widowed";
14 #Warning: 因子层次有错, 产生了NA
```

Play around with `levels()`

```

1 ##利用levels解决
2 levels(x) <- c(levels(x), "widowed");
3 x[ length(x) + 1 ] <- "widowed";
4 str(x);
5
6 ## other ways of assigning factors ...
7 y <- as.factor( c( "single", "married", "married", "single" ) );
8 levels(y);
9 levels(y) <- c("single", "married", "widowed");
10 str(y);
11 ## 这个代码现在就没有问题了
12 y[ length(y) + 1 ] <- "widowed";

```

注意用 `as.factor` 创建 factor 时，得到的 levels 按字母表排列；

但是，用 `levels(y)` 方式指定 levels 时，则按照指定的顺序；

`levels` 的顺序决定了排序的顺序

```

1 ##
2 y <- as.factor( c( "single", "married", "married", "single" ) );
3 levels(y);
4 sort(y);
5 #[1] "married" "single"
6 #[1] married married single single Levels: married single
7
8 ##
9 y2 <- y;
10 levels(y2) <- c("single", "married", "widowed");
11 sort(y2);
12 #[1] single single married married Levels: single married
widowed

```

sort data in a meaningful way ...

```

1 ## Month
2 x1 <- c("Dec", "Apr", "Jan", "Mar");
3 sort(x1);
4
5 month_levels <- c(
6   "Jan", "Feb", "Mar", "Apr", "May", "Jun",
7   "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"
8 )
9
10 y1 <- factor(x1, levels = month_levels)
11 sort(y1);
12 #[1] "Apr" "Dec" "Jan" "Mar"
13 #[1] Jan Mar Apr Dec
14 #12 Levels: Jan Feb Mar Apr May Jun Jul Aug Sep Oct ... Dec
15
16 ## Sometimes you'd prefer that the order of the levels match the
17 ## order of the first appearance in the data.
18 f1 <- factor(x1, levels = unique(x1));
19 f1;
20 #[1] Dec Apr Jan Mar Levels: Dec Apr Jan Mar
21
22 library(forcats); ## just to make sure the codes will run
23 ## smoothly ...
24
25 ## you can also use fct_inorder in the forcats package ...
26 f2 <- x1 %>% factor() %>% fct_inorder()
27 f2
28 #[1] Dec Apr Jan Mar Levels: Dec Apr Jan Mar

```

Here are instructions of modifying factor levels

Based on the textbook

The levels 既简洁又不一致。让我们调整使之更长，并使用一个平行的结构。Like most rename and recoding functions in the tidyverse,新的值在左边，旧的值在右边：

`fct_recode` Change factor levels by hand

```

1 load(gss_cat)

```

```

2
3   mutate(
4     partyid = fct_recode(partyid,
5       "Republican, strong"      = "Strong republican",
6       "Republican, weak"        = "Not str republican",
7       "Independent, near rep"  = "Ind,near rep",
8       "Independent, near dem"  = "Ind,near dem",
9       "Democrat, weak"         = "Not str democrat",
10      "Democrat, strong"       = "Strong democrat"
11    )
12  )
13
14 count(partyid)
15
16 #> # A tibble: 10 × 2
17 #>   partyid             n
18 #>   <fct>              <int>
19 #> 1 No answer          154
20 #> 2 Don't know          1
21 #> 3 Other party         393
22 #> 4 Republican, strong  2314
23 #> 5 Republican, weak    3032
24 #> 6 Independent, near rep 1791
25 #> # i 4 more rows

```

使用这种技术时要小心：如果你把完全不同的类别组合在一起，你最终会得到误导性的结果。

The order of the `levels` determines the sorting order.

Use factor to clean data

Usage of `fct_xxx()` functions.

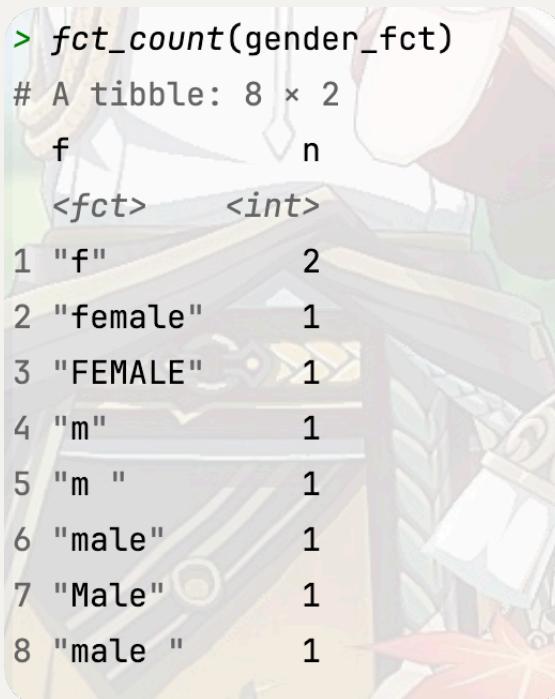
Suppose I have a set of gender data that is written in a very irregular way:

```

1 ## 假设我有一组性别数据，其写法非常不规整;
2 gender <- c("f", "m ", "male ","male", "female", "FEMALE",
3           "Male", "f", "m");
4
5 gender_fct =
6   as.factor(gender)
7 fct_count(gender_fct)

```

The output looks like this:



```

> fct_count(gender_fct)
# A tibble: 8 × 2
  f      n
  <fct> <int>
1 "f"     2
2 "female" 1
3 "FEMALE" 1
4 "m"     1
5 "m "    1
6 "male"  1
7 "Male"  1
8 "male " 1

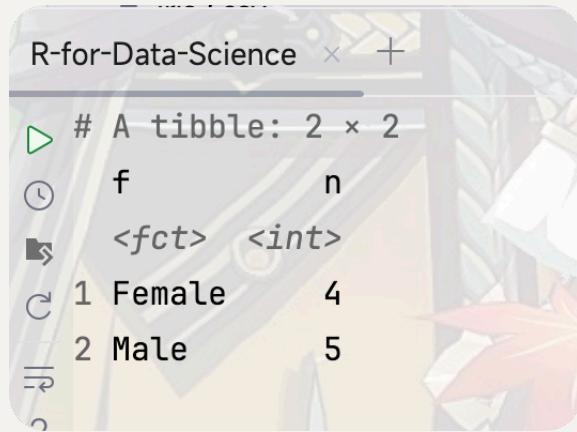
```

Now I request to replace with Female, Male.

```

1 ## 要求：都改为 Female, Male
2 gender_fct =
3   fct_collapse(
4     gender,
5     Female = c("f", "female", "FEMALE"),
6     Male = c("m ", "m", "male ", "male", "Male")
7   )
8
9 fct_count(gender_fct)

```



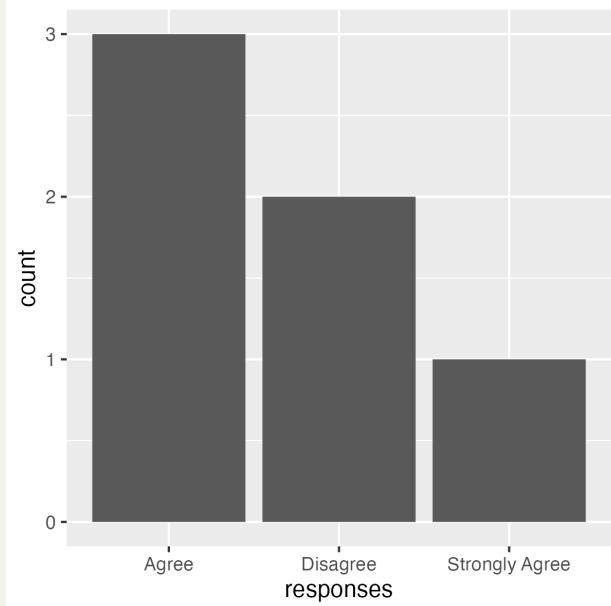
```
R-for-Data-Science ✕ +  
# A tibble: 2 × 2  
f      n  
<fct> <int>  
1 Female 4  
2 Male   5
```

You can also use `fct_relabel()` to do the same thing

```
1 fct_relabel(  
2   gender,  
3   ~ ifelse(  
4     tolower(  
5       substring(., 1, 1)) == "f",  
6       "Female",  
7       "Male"  
8     )  
9   )
```

factor 在做图中的应用（真正精髓）

```
1 ## 一项 mock 调查结果数据  
2 library(ggplot2)  
3  
4 responses =  
5   factor(  
6     c("Agree", "Agree", "Strongly Agree", "Disagree",  
7     "Disagree", "Agree")  
8   )  
9  
9 response_barplot =  
10 ggplot(  
11   data = data.frame(responses),  
12   aes(x = responses)  
13 ) +  
14 geom_bar()
```



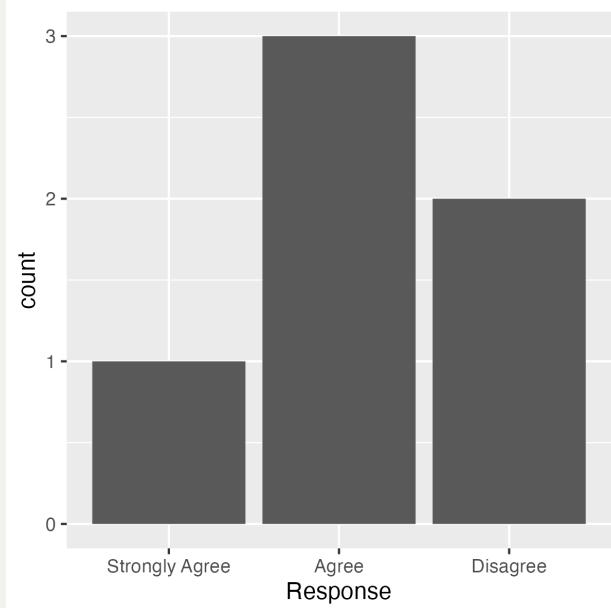
默认情况下， factor 按字母表排序： Agree -> Disagree -> Strong Agree 。 ggplot2 也会按factor的排序作图

`ggplot2` also plots `factor` in that order, so you can adjust the `factor` to adjust the drawing order.

```

1  res =
2    data.frame(responses)
3  # Sort by level of agreement from strong -> weak
4  res$res =
5    factor(
6      res$res,
7      levels =
8        c("Strongly Agree", "Agree", "Disagree"),
9      ordered = T
10    )
11
12 response_barplot2 =
13   ggplot(
14     data = res,
15     aes(x = res)
16   ) +
17   geom_bar() +
18   xlab("Response") + ylab("Count")

```



You can also use the parameter `ordered` to let others know that your `factor` is ordered properly.

```

1 responses =
2   factor(
3     c("Agree", "Agree", "Strongly Agree", "Disagree", "Disagree",
4       "Agree"),
5     ordered = TRUE
6   )
7 is.ordered( responses )

```

```

> is.ordered(responses)
[1] TRUE

```

Using `factor` to change values

You can use `recode()` in `dplyr` package to change `value`

`dplyr` 是一种数据操作的语法，提供了一组一致的动词 that help you solve the most common data manipulation challenges:

- `mutate()` adds new variables that are functions of existing variables
- `select()` picks variables based on their names.

- `filter()` picks cases based on their values.
- `summarise()` reduces multiple values down to a single summary.
- `arrange()` changes the ordering of the rows.

These all combine naturally with `group_by()` which allows you to perform any operation “by group”. You can learn more about them in `vignette("dplyr")`. As well as these single-table verbs, dplyr also provides a variety of two-table verbs, which you can learn about in `vignette("two-table")`.

Based on the introduction on the [official website](#) of `dplyr`.

Here's an example:

使用 `dplyr` 包的 `recode()` 函数改变 value

```

1 x =
2   factor(
3     c("alpha", "beta", "gamma", "theta", "beta", "alpha")
4   )
5
6 x =
7   recode(
8     x,
9     alpha = "a",
10    beta = "b",
11    gamma = "c",
12    theta = "d"
13  )

```

```

① +     theta = "d"
② +   )
③ > str(x)
> Factor w/ 4 levels "a","b","c","d": 1 2 3 4 2 1
] ?
)   >

```

Delete useless levels

```

1 mouse.genes =
2   read.delim(
3     file = "data/talk04/mouse_genes_biomart_sep2018.txt",
4     sep = "\t",
5     header = T,
6     stringsAsFactors = T
7   )

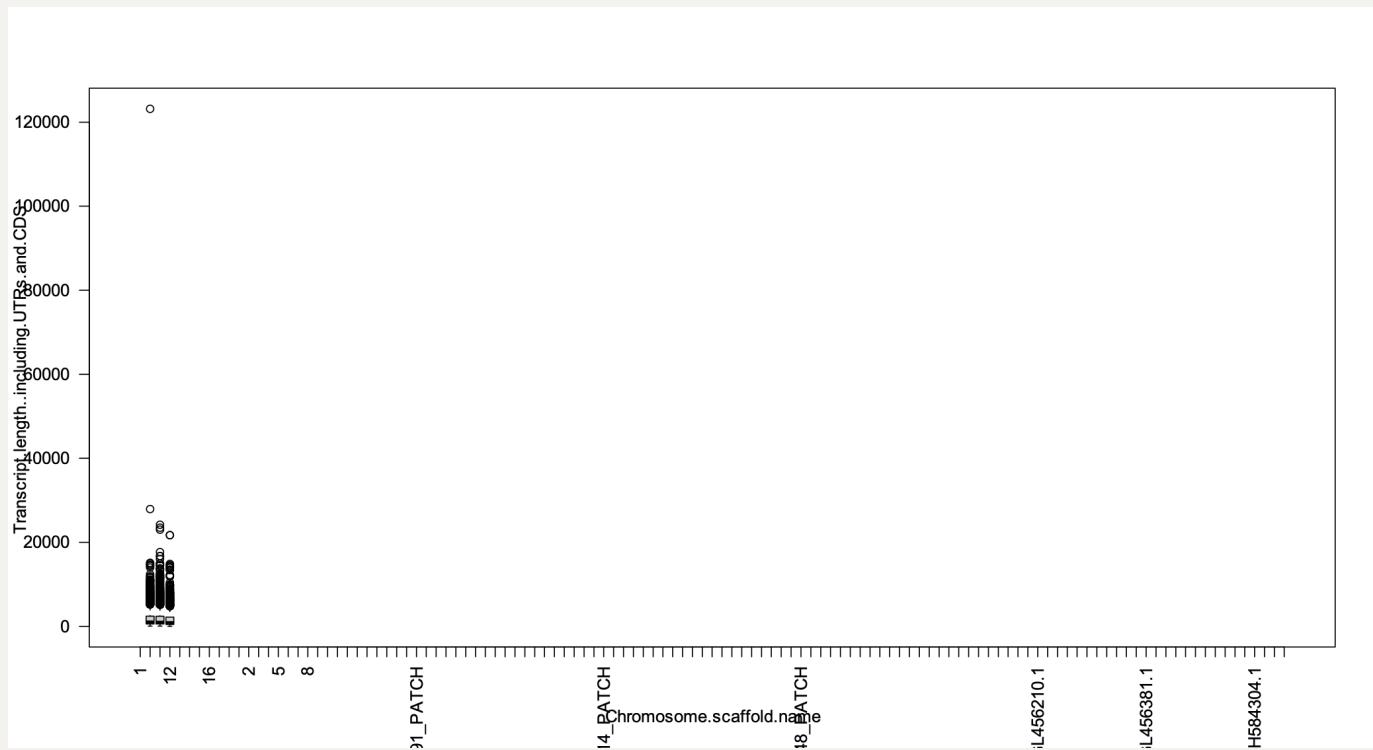
```

```

+ }
> str(mouse.genes)
'data.frame': 138532 obs. of 6 variables:
 $ Gene.stable.ID           : Factor w/ 55029 levels "ENSMUSG00000000001",...: 17960 17959 17958 17957 17956 17955 17954 17953 17952 17951 ...
 $ Transcript.stable.ID     : Factor w/ 138532 levels "ENSMUST00000000001",...: 17312 17311 17310 17309 17308 17307 17306 17305 17304 17303 ...
 $ Protein.stable.ID        : Factor w/ 65897 levels "", "ENSMUSP00000000001",...: 1 1 16260 1 16259 16258 1 1 1 16257 ...
 $ Transcript.length..including.UTRs.and.CDS.: int 67 67 1144 69 519 1824 71 59 67 1378 ...
 $ Transcript.type          : Factor w/ 48 levels "3prime_overlapping_ncRNA",...: 18 18 24 18 24 24 18 18 18 24 ...
 $ Chromosome.scaffold.name : Factor w/ 117 levels "1", "10", "11", ...: 115 115 115 115 115 115 115 115 115 115 ...

```

If you draw a plot without deleting the useless `levels`, you will get this result:



```

1 mouse.chr_10_12 <- subset( mouse.genes, Chromosome.scaffold.name
2 %in% c( "10", "11", "12" ) );
3 ## plot length distribution --
4 boxplot( Transcript.length..including.UTRs.and.CDS. ~
5   Chromosome.scaffold.name,
6   data = mouse.chr_10_12, las = 2 );

```

But when you delete the useless `level` using these commands:

```

1 mouse.chr_10_12$Chromosome.scaffold.name =
2 droplevels(mouse.chr_10_12$Chromosome.scaffold.name)
3 levels( mouse.chr_10_12$Chromosome.scaffold.name )

```

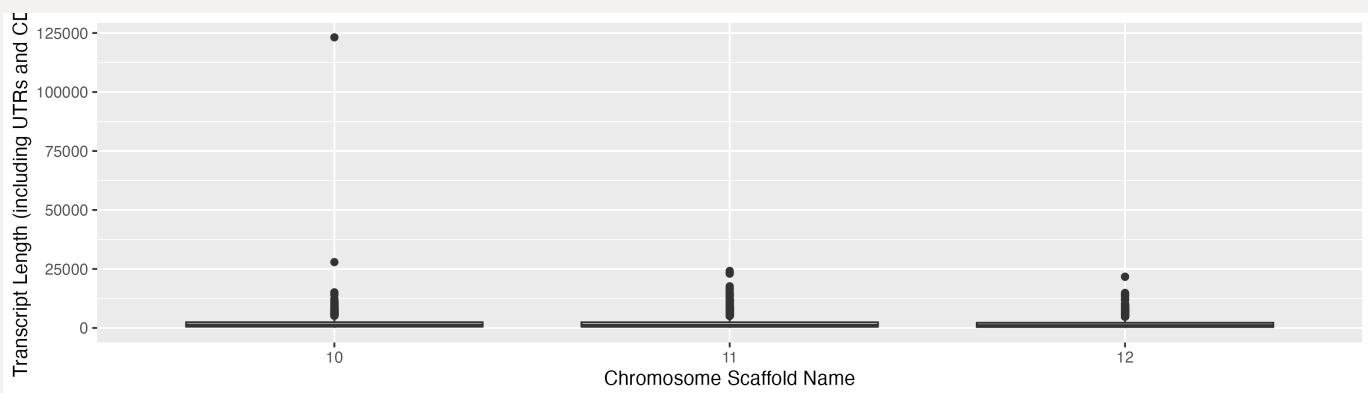
You will see that:

```

> +   droplevels( mouse.chr_10_12$Chromosome.scaffold.name )
> levels(mouse.chr_10_12$Chromosome.scaffold.name)
[1] "10" "11" "12"

```

Then, you'll get the plot like this:



Source code:

```

1 mouse_gene_plot02 =
2   ggplot(
3     mouse.chr_10_12,
4     aes(
5       x = Chromosome.scaffold.name,
6       y = Transcript.length..including.UTRs.and.CDS.
7     )
8   ) +
9   geom_boxplot() +
10  labs(
11    x = "Chromosome Scaffold Name",
12    y = "Transcript Length (including UTRs and CDS)"
13  )

```

You can also use `tibble` to solve these problems:完全不用担心 factor 的问题 ...

```

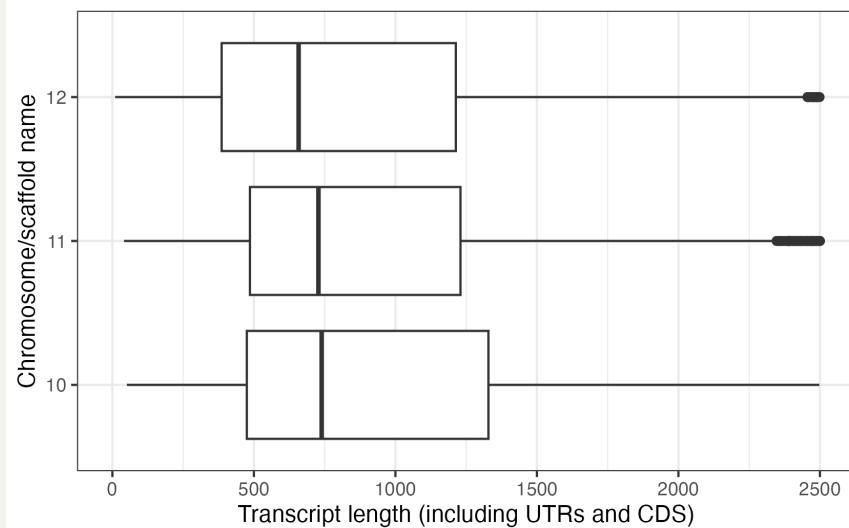
1 mouse.tibble =
2   read_delim(
3     file = "data/talk04/mouse_genes_biomart_sep2018.txt",
4     delim = "\t",
5     quote = """",
6     show_col_types = FALSE
7   )
8
9 mouse.tibble.chr10_12 =
10  mouse.tibble %>% filter(
11    `Chromosome/scaffold name` %in% c("10", "11", "12"))
12
13
14 mouse_gene_plot03 =
15   ggplot(
16     mouse.tibble.chr10_12,
17     aes(
18       x = Chromosome.scaffold.name,
19       y = Transcript.length..including.UTRs.and.CDS.
20     )
21   ) +
22   geom_boxplot() +

```

```

23   labs(
24     x = "Chromosome",
25     y = "Transcript length (bp)"
26   ) +
27   coord_flip() +
28   ylim(0, 2500) +
29   theme_bw()

```



Advance usage

- Use `reorder()` function to reorder the level.
- `reorder()` 函数在 R 语言中通常用于改变因子 (factor) 水平的顺序。这对于数据可视化非常有用，特别是当你使用 ggplot2 包绘图时。在 ggplot2 中，`reorder()` 可以帮助你重新安排条形图、箱形图等的顺序。

函数的基本用法是 `reorder(x, ...)`，其中 `x` 是你想要重新排序的因子，而 `...` 是额外的参数和方法，用于确定新的排序。最常见的用法是根据另一个变量的某种统计度量（如平均值、中位数等）来排序。

```

1 x = reorder(
2   `Chromosome/scaffold name`,
3   `Transcript length (including UTRs and CDS)` ,
4   median
5 )

```

- Use `forcats::fct_reorder()` to reorder factors

```
1 x = fct_reorder(  
2   `Chromosome/scaffold name`,  
3   `Transcript length (including UTRs and CDS)`,  
4   median  
5 )
```

一些勘误

vector 和 factor 有什么区别?

1. 向量 (Vector):

- 概念：向量是 R 中最基本的数据类型之一，用于存储相同类型的数据元素（如数值、字符或逻辑值）的序列。
- 类型：向量可以是数值型 (numeric) , 字符型 (character) , 逻辑型 (logical) 等。
- 用途：向量用于存储和操作数据集中的实际值。例如，一个存储温度读数或城市名称的序列。

2. 因子 (Factor):

- 概念：因子是用于表示分类数据的数据类型。它类似于枚举类型，用于表示有限的、通常是预定义的值集合。
- 水平：因子的值被称为水平 (levels) ，这些水平代表了分类变量的可能值。
- 用途：因子主要用于统计建模和图形表示中，用于处理分类数据。例如，在绘制条形图时，条形的类别就是由因子类型的变量确定的。

区别：

- 数据类型：向量直接存储值，而因子存储的是分类水平。
- 用途：向量用于表示一系列的值，而因子用于表示分类或分组。
- 统计分析：在进行统计分析时，因子用于指定数据的分类属性。例如，在回归分析中，因子可以用于指定自变量的分类。

data.frame 与 tibble 的区别

1. **data.frame**:

- 历史：`data.frame` 是 R 语言中最传统的数据集结构，自 R 诞生之初就存在。

- 特性：它是一个表格型数据结构，每列可以包含不同类型的数据（数值型、字符型等），但每列数据类型必须一致。
- 行为：在某些操作中，`data.frame` 可能会自动将字符串转换为因子（这取决于 `stringsAsFactors` 的默认设置）。
- 使用：`data.frame` 在所有标准的 R 函数和包中广泛支持。

2. `tibble`:

- 历史：`tibble` 是相对较新的数据集结构，由 `tidyverse` 生态系统引入，旨在解决 `data.frame` 的一些局限性。
- 特性：`tibble` 保持了 `data.frame` 的所有基本特性，但添加了一些用户友好的改进，如更好的默认打印方法（显示更多信息，但更易于阅读）。
- 行为：`tibble` 不会自动将字符串转换为因子，保留了数据的原始类型。
- 兼容性：虽然 `tibble` 设计用于 `tidyverse` 生态系统，但它也兼容大多数接受 `data.frame` 的 R 函数。

关键区别：

- 因子转换：`tibble` 默认不会将字符串自动转换为因子，而 `data.frame` 可能会这样做。
- 打印和查看数据：`tibble` 提供了更友好的数据打印和查看方式，更适合于大型数据集。
- 设计理念：`tibble` 是为了更好地适应现代数据分析需求而设计的，尤其是在 `tidyverse` 生态系统中。