R软件在CGED-Q JSL中的运用

（五）dplyr包在字符串处理中的简单运用

作者：陈俊（硕士研究生，华中师范大学）；康文林（教授，香港科技大学；华中师范大学）

01

筛选行与列

PART one

02

数据集的匹配

PART two

03

连接两个数据集的简单尝试

PART three

01

筛选行与列

tidyverse项目，是一个包括了数据科学的一个集合工具项目，用于数据提取，数据清理，数据类型定义，数据处理，数据建模，函数化编程，数据可视化，包括了下面的包。

-------- **ggplot2** ~~~~~~~数据可视化
-------- **dplyr** ~~~~~~~数据处理
-------- **tidyr** ~~~~~~~数据清理
-------- **readr** ~~~~~~~数据提取
-------- **purrr** ~~~~~~~函数化编程
-------- **tibble** ~~~~~~~数据类型定义

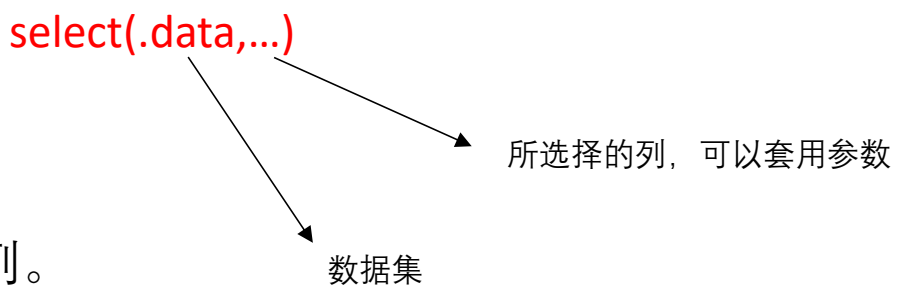
tidyverse项目的地址：https://github.com/tidyverse/tidyverse

原文链接：https://blog.csdn.net/fens/article/details/84634531

筛 选 列

select()函数

select()函数，可以以挑选出指定条件的列，格式为：



常用的参数有：
-1- last_col():表示选择最后一列。
-2- starts_with():表示以什么开头的列。
-3- ends_with():表示以什么结尾的列。
-4- contains():表示某列是否包含什么内容。

参考链接：https://zhuanlan.zhihu.com/p/358167377

管道符，可以将某一个数据集的行或列转移到另一个数据集。利用管道符和select()函数，可以截取某一个数据集的部分变量生成新的数据集。

JSL1900_1912_delete_kongbaiming %>%
select(阳历年份,地区,机构一,官职一,姓,名,籍贯省,籍贯县,旗分,出身一)
-> JSL1900_1912_simplify

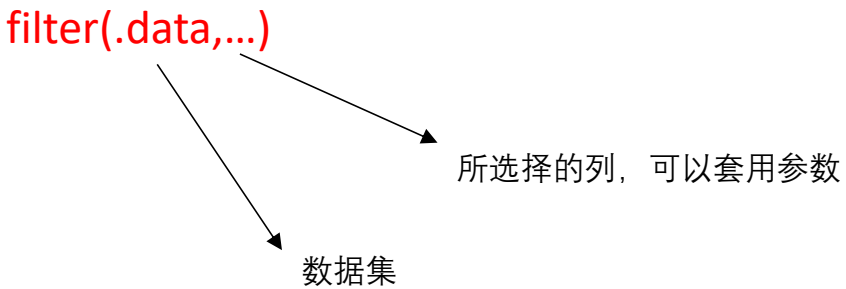
需要转移的列变量的名称，注意，这里不需要再指定数据集，因为管道符的有一个作用类似于attach()函数。

转移的/生成的数据集的名称

筛 选 行

filter()函数

filter()函数，可以以挑选出指定条件的列，格式为：



用法和select()函数相似

管道符

```
JSL1900_1912_delete_kongbaiming %>%
    filter((阳历年份numeric >= 1900)&(阳历年份numeric <= 1906)&(机构一 == "翰林院衙門")) -> JSL1900_1906_hanlinyuan
```

筛选条件，注意连接符号的运用。

筛选后生成的数据集的名称，自拟

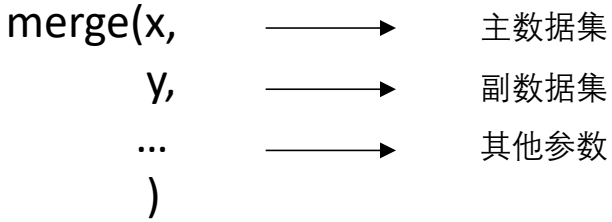
到这里，我们已经学习了两种按条件生成数据集的方法，一种是subset()函数，一种是dplyr包下的管道符与filter()函数连用，可以根据自己的需求和对函数的理解来选择适合自己的方法。

02

数据集的匹配

merge()函数

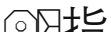
格式与逻辑：



 函数的参数：

by = ...表示选择主数据集和副数据集连接的列，当遇到主数据集和副数据集标题名称不同时，可以将by参数拆分为by.x和by.y参数，令by.x和by.y分别等于名称不同但内容有交集的公共列。

all = ...表示全连接，可选TRUE或FALSE。当只需要保存主数据集的所有列而不需要副数据集的列时，令all.x = TRUE并且all.y = False，即为左连接；当只需要保存副数据集的所有列而不需要主数据集的列时，令all.x = False并且all.y = TRUE，即为右连接；当只需要主数据集和副数据集的交集，即令all = False，即为求交集，这也是R的默认选项。

 ：  指定的列（即公共列）是否要排序

 ：指定除外相同列名的后缀

 ：指定中哪些单元不进行合并

参考链接： 

实例代码：匹配以整理“出身一”

新建数据集的名称

```
JSL1900_1912_delete_kongbaiming_2 <-
merge(JSL1900_1912_delete_kongbaiming_1,
    Chushen_Recodes_1_ ,
    by = “出身一”,
    all.x = TRUE)
```

主数据集

副数据集

公共列”出身一”

左连接

实例代码：在已整理好“出身一”的数据集基础上整理“籍贯省”

注意，R一次只能匹配一个副数据集，因此，整理下一个变量只能在上一个已整理好变量的数据集上进行，我们这里生成的是新的数据集。

```
JSL1900_1912_delete_kongbaiming_3 <-
merge(JSL1900_1912_delete_kongbaiming_2,
                    jiguansheng_Recodes,
                    by = "籍贯省",
                    all= TRUE)
```

主数据集，是已经整理好的数据集

全连接，不过在缙绅录中，左连接和全连接作用是一样的

如何制作一个可以匹配的副数据集来满足整理变量的需求？

利用table（）函数将某个变量的所有数据显示出来

→ 复制table的表格到excel中

→ 利用excel进行分列，提出主要的数据信息

→ 对数据信息进行编辑，生成新的一列，整理的信息与原数据信息一一对应

→ 保存整理好的文件，导入到R中

→ 利用merge函数进行匹配

下面以铨选方式的整理为例

1.将某个变量的数据信息全部展示出来，具体处理在excel

```
table(JSL1900_1912_delete_kongbaiming_3$铨选方式,
    JSL1900_1912_delete_kongbaiming_3$qiren)
```

2.导入副数据集文件到R

```
quanxuanfangshi_sort <- read_excel("E:/R/quanxuanfangshi_sort.xlsx")
```

3.匹配

```
JSL1900_1912_delete_kongbaiming_4 <- merge(JSL1900_1912_delete_kongbaiming_3,
                                           quanxuanfangshi_sort,
                                           by = "铨选方式",
                                           all= TRUE)
```

4.检验（种类）

```
table(JSL1900_1912_delete_kongbaiming_4$铨选方式_sort,
    JSL1900_1912_delete_kongbaiming_4$qiren)
```

5.检验（数据量）

```
table1(~铨选方式_sort|qiren,
      data = JSL1900_1912_delete_kongbaiming_4,
      overall = "total")
```

03

连接两个数据集的简单尝试

inner_join()函数

格式与逻辑:

inner_join(x, ⟶ 主数据集

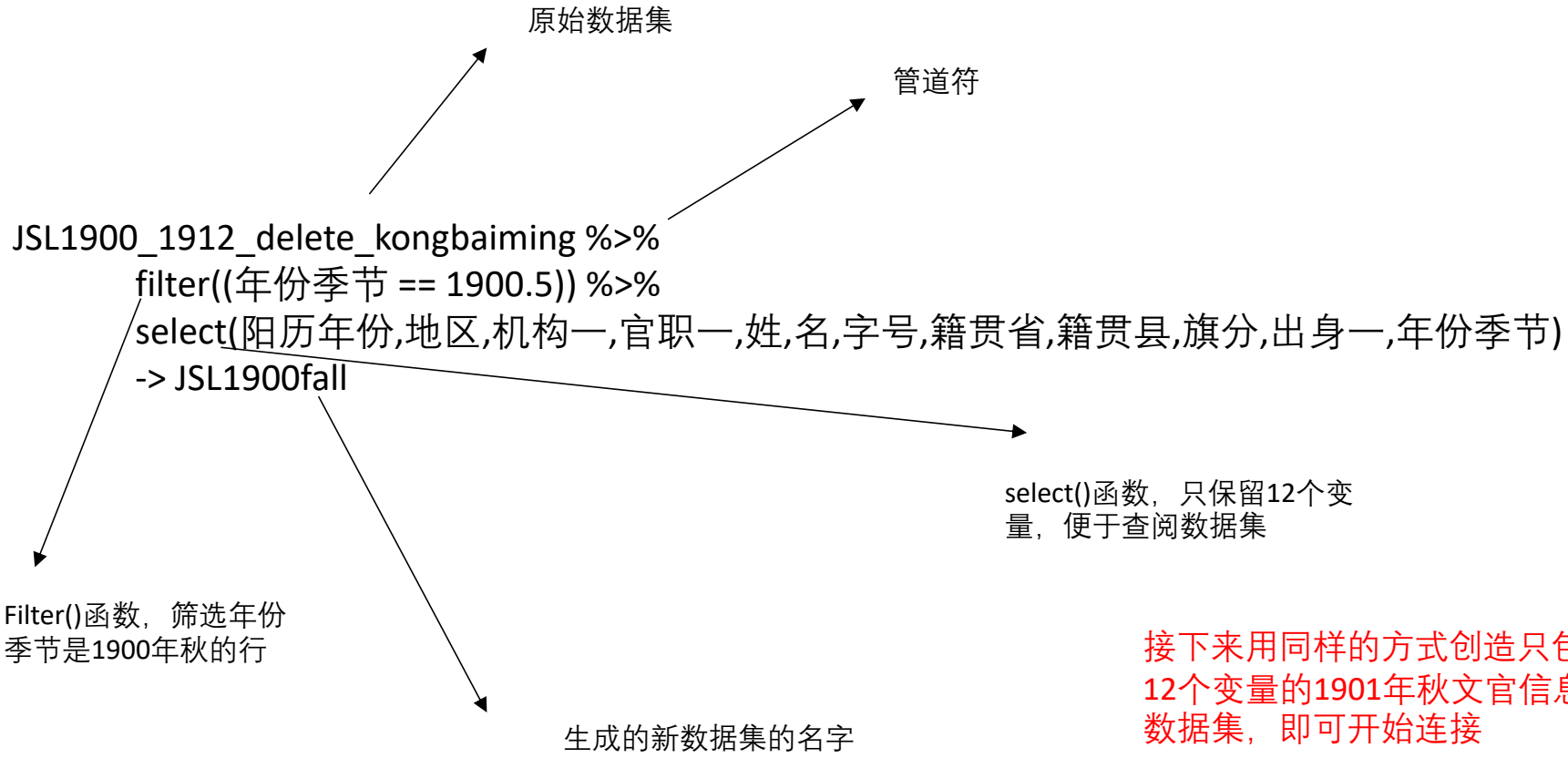
y, ⟶ 副数据集

by = , ⟶ 选定两个数据集中相同的列，与merge函数用法相似，如果选择多个列进行连接，则采用by = c(“”,””,””,…)的形式

na_matches = , ⟶ 是否匹配含NA的行，通常选择“never”

)

在进行连接之前，我们需要创建用来连接的数据集



接下来用同样的方式创造只包含12个变量的1901年秋文官信息的数据集，即可开始连接

实例代码：

新建数据集的名称

主数据集

```
JSL1900_1901_inner_join <- inner_join(JSL1900fall,
                                      JSL1901fall,
                                      by = c( “名”,“姓”,“字号”,“籍贯省”,“籍贯县”,“出身一”,“旗分” ),
                                      keep = FALSE,
                                      na_matches = “never” )
```

副数据集

不匹配含有NA的行

公共列，这里我们用了集合的方式选出了7个公共列，这意味着两个数据集的行都需要满足7个条件才可进行连接

合并之后保留副数据集的连接列吗？选择TRUE可以检验连接是否正常，选择FALSE可以简化数据集，仅凭个人意愿选择

谢谢