dplyr\_join\_demo

#下面内容来自互联网，为了方便学习dplyr包里的join合并数据操作  
#主要数据和内容参考：https://stat545-ubc.github.io/bit001\_dplyr-cheatsheet.html  
  
#加载需要的包  
library(dplyr)

##   
## Attaching package: 'dplyr'  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(plyr)

## -------------------------------------------------------------------------  
## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)  
## -------------------------------------------------------------------------  
##   
## Attaching package: 'plyr'  
##   
## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

library(stringr)  
library(knitr)  
  
#定义一个函数来处理将array转换为data.frame的列命名问题，将第一行变为列名，然后去掉第一行生成df  
norm\_df<-function(df){  
 colnames(df)<-df[1,] #定义上列名  
 df<-as.data.frame(df[-1,]) %>% #去掉首行  
 tbl\_df() #生成dplyr的tbl\_df对象   
 return(df)  
}  
  
#原始数据链式处理---------------------------------------  
superheroes <-  
 c(" name, alignment, gender, publisher",  
 " Magneto, bad, male, Marvel",  
 " Storm, good, female, Marvel",  
 "Mystique, bad, female, Marvel",  
 " Batman, good, male, DC",  
 " Joker, bad, male, DC",  
 "Catwoman, bad, female, DC",  
 " Hellboy, good, male, Dark Horse Comics") %>%  
 laply(strsplit,",") %>% #把vector根据，拆分成array  
 aaply(1,str\_trim) %>% #去掉每个character元素里的首尾空格，基于stringr包   
 norm\_df()  
  
publishers <-   
 c("publisher, yr\_founded",  
 " DC, 1934",  
 " Marvel, 1939",  
 " Image, 1992") %>%  
 laply(strsplit,",") %>% #把vector根据，拆分成array  
 aaply(1,str\_trim) %>% #去掉每个character元素里的首尾空格   
 norm\_df()   
  
#备注，以下演示省略了by参数，合并的数据可以是不同列名的，也可以是多列，使用的语法如下  
#left\_join(d1, d2, by = c("x1" = "x2", "y1" = "y2"))  
#其中，x1 y1是d1数据的，x2 也是d2数据的，需要做成一个这样的向量来代表d1.x1==d2.x2 & d1.y1==d2.y2  
  
#演示inner\_join----------------------------  
ijsp <- inner\_join(superheroes, publishers)

## Joining by: "publisher"

## Warning: joining factors with different levels, coercing to character  
## vector

ijsp

## Source: local data frame [6 x 5]  
##   
## name alignment gender publisher yr\_founded  
## 1 Magneto bad male Marvel 1939  
## 2 Storm good female Marvel 1939  
## 3 Mystique bad female Marvel 1939  
## 4 Batman good male DC 1934  
## 5 Joker bad male DC 1934  
## 6 Catwoman bad female DC 1934

#Hellboy观测在合并数据中消失了，因为他的publisher是Dark Horse Comics并未在y的publisher中出现  
#因为inner join是只显示匹配的（两边都有的）观测  
  
#演示semi\_join-----------------------------  
sjsp <- semi\_join(superheroes, publishers)

## Joining by: "publisher"

## Warning: joining factors with different levels, coercing to character  
## vector

sjsp

## Source: local data frame [6 x 4]  
##   
## name alignment gender publisher  
## 1 Batman good male DC  
## 2 Joker bad male DC  
## 3 Catwoman bad female DC  
## 4 Magneto bad male Marvel  
## 5 Storm good female Marvel  
## 6 Mystique bad female Marvel

#结果与inner\_join类似，匹配过后只得到了有左边superheros里的变量，右边df里的列没合并进来。  
  
#演示left\_join-----------------------------  
ljsp <- left\_join(superheroes, publishers)

## Joining by: "publisher"

## Warning: joining factors with different levels, coercing to character  
## vector

ljsp

## Source: local data frame [7 x 5]  
##   
## name alignment gender publisher yr\_founded  
## 1 Magneto bad male Marvel 1939  
## 2 Storm good female Marvel 1939  
## 3 Mystique bad female Marvel 1939  
## 4 Batman good male DC 1934  
## 5 Joker bad male DC 1934  
## 6 Catwoman bad female DC 1934  
## 7 Hellboy good male Dark Horse Comics NA

#这次是完全以左边为准，Hellboy观测没有匹配也予以保留了，只是其Yr\_founded值变成了NA  
  
#演示anti\_join-----------------------------  
ajsp <- anti\_join(superheroes, publishers)

## Joining by: "publisher"

## Warning: joining factors with different levels, coercing to character  
## vector

ajsp

## Source: local data frame [1 x 4]  
##   
## name alignment gender publisher  
## 1 Hellboy good male Dark Horse Comics

#anti\_join返回的是以左边为准在右边找不到匹配的观测，并且不会合并右边的变量  
  
#演示full\_join-----------------------------  
fjsp <- full\_join(superheroes, publishers)

## Joining by: "publisher"

## Warning: joining factors with different levels, coercing to character  
## vector

fjsp

## Source: local data frame [8 x 5]  
##   
## name alignment gender publisher yr\_founded  
## 1 Magneto bad male Marvel 1939  
## 2 Storm good female Marvel 1939  
## 3 Mystique bad female Marvel 1939  
## 4 Batman good male DC 1934  
## 5 Joker bad male DC 1934  
## 6 Catwoman bad female DC 1934  
## 7 Hellboy good male Dark Horse Comics NA  
## 8 NA NA NA Image 1992

#full\_join返回的是两边都有的数据，这个不丢失两边任何的观测