Packages for Comparison between Dataframes in R

billyi

2018-09-15

比較數據框的套件介紹

本文介紹了幾種在CRAN (https://cran.r-project.org/)上的套件，供您參考。

前置作業

首先引入含有管道運算子（pipe operator）的magrittr (https://cran.r-project.org/web/packages/magrittr/vignettes/magrittr.html)函式庫，以備後續使用。

```
library(magrittr)
```

接下來製作二個格式相近、內容相仿的數據框，以進行比較。

```
df1 <- data.frame(
  char = LETTERS[1:5],
  intg = 5:9,
  numb = seq(pi, length.out = 5),
  bool = c(rep(TRUE, 3), rep(FALSE, 2)),
  tiangan = c("甲", "乙", "丙", "丁", "戊"),
  stringsAsFactors = FALSE
)

df2 <- data.frame(
  char = LETTERS[c(1:4, 9:10)],
  intg = c(NA, 6:10),
  numb = c(seq(pi, length.out = 5), NA),
  bool = c(TRUE, FALSE, TRUE, FALSE, TRUE, FALSE),
  dizhi = c("子", "丑", "寅", "卯", "辰", "巳"),
  stringsAsFactors = FALSE
)
```

先看一下第一個數據框，形狀為5x5，內容如下：

```
knitr::kable(df1)
```

char	intg	numb	bool	tiangan
A	5	3.141593	TRUE	甲
B	6	4.141593	TRUE	乙
C	7	5.141593	TRUE	丙
D	8	6.141593	FALSE	丁
E	9	7.141593	FALSE	戊

```
dplyr::glimpse(df1)
```

```
## Observations: 5
## Variables: 5
## $ char    <chr> "A", "B", "C", "D", "E"
## $ intg    <int> 5, 6, 7, 8, 9
## $ numb    <dbl> 3.141593, 4.141593, 5.141593, 6.141593, 7.141593
## $ bool    <lgl> TRUE, TRUE, TRUE, FALSE, FALSE
## $ tiangan <chr> "甲", "乙", "丙", "丁", "戊"
```

再看看第二個數據框，其形狀為6x5，內容如下：

```
knitr::kable(df2)
```

char	intg	numb	bool	dizhi
A	NA	3.141593	TRUE	子
B	6	4.141593	FALSE	丑
C	7	5.141593	TRUE	寅
D	8	6.141593	FALSE	卯
I	9	7.141593	TRUE	辰
J	10	NA	FALSE	巳

```
dplyr::glimpse(df2)
```

```
## Observations: 6
## Variables: 5
## $ char  <chr> "A", "B", "C", "D", "I", "J"
## $ intg  <int> NA, 6, 7, 8, 9, 10
## $ numb  <dbl> 3.141593, 4.141593, 5.141593, 6.141593, 7.141593, NA
## $ bool  <lgl> TRUE, FALSE, TRUE, FALSE, TRUE, FALSE
## $ dizhi <chr> "子", "丑", "寅", "卯", "辰", "巳"
```

比較工具

compare::compare

引入compare (https://cran.r-project.org/web/packages/compare/index.html)函式庫，再以 `compare` 函式進行比較。

```
suppressPackageStartupMessages(library(compare))
compare(df1, df2, allowAll = TRUE)
```

```
## FALSE [FALSE, FALSE, TRUE, FALSE, FALSE]
##   shortened comparison rows
##   renamed
##   dropped names
##   [1] ignored case
##   [5] ignored case
```

compareDF::compare_df

引入compareDF (https://cran.r-project.org/web/packages/compareDF/index.html)函式庫，再以 `compare_df` 函式進行比較。

```
library(compareDF)
compare_df(df1[1:4], df2[1:4], group_col = "char")$comparison_df
```

```
## Creating comparison table...
```

```
## Loading required namespace: htmlTable
```

```
## Creating HTML table for first 100 rows
```

```
##   char chng_type intg numb  bool
## 1    A         +    5 3.14  TRUE
## 2    A         -   NA 3.14  TRUE
## 3    B         +    6 4.14  TRUE
## 4    B         -    6 4.14 FALSE
## 5    E         +    9 7.14 FALSE
## 6    I         -    9 7.14  TRUE
## 7    J         -   10   NA FALSE
```

daff::diff_data

引入daff (https://cran.r-project.org/web/packages/daff/index.html)函式庫，再以 `diff_data` 函式進行比較。

```
library(daff)
daff::diff_data(df1, df2)
```

```
## Daff Comparison: 'df1' vs. 'df2'
##   First 6 and last 6 patch lines:
##       !               +++     ---
## 1    @@        bool dizhi tiangan
## 2    ->        TRUE    子      甲
## 3    -> true->false    丑      乙
## 4     +        TRUE    寅      丙
## 5     +       FALSE    卯      丁
## 6    -> false->true    辰      戊
## ... ...         ...   ...     ...
## 21   ->        TRUE    子      甲
## 31   -> true->false    丑      乙
## 41    +        TRUE    寅      丙
## 51    +       FALSE    卯      丁
## 61   -> false->true    辰      戊
## 7   +++       FALSE    巳    <NA>
```

dataCompareR::rCompare

引入dataCompareR (https://cran.r-project.org/web/packages/dataCompareR/index.html)函式庫，再以 `rCompare` 函式進行比較，接著用 `summary` 總結差異所在。

```
library(dataCompareR)
dataCompareR::rCompare(df1, df2) %>% summary()
```

```
## Running rCompare...
```

```
## dataCompareR is generating the summary...
```

```
##
## Data Comparison
## ===============
##
## Date comparison run: 2018-09-15 18:14:37
## Comparison run on R version 3.5.1 (2018-07-02)
## With dataCompareR version 0.1.1
##
##
## Meta Summary
## ============
##
##
## Dataset Name   Number of Rows   Number of Columns
## -------------  ---------------  ------------------
## df1            5                5
## df2            6                5
##
##
## Variable Summary
## ================
##
## Number of columns in common: 4
## Number of columns only in df1: 1
## Number of columns only in df2: 1
## Number of columns with a type mismatch: 0
## No match key used, comparison is by row
##
##
## Columns only in df1: tiangan
## Columns only in df2: dizhi
## Columns in both : BOOL, CHAR, INTG, NUMB
##
## Row Summary
## ===========
##
## Total number of rows read from df1: 5
## Total number of rows read from df2: 6
## Number of rows in common: 5
## Number of rows dropped from df1: 0
## Number of rows dropped from  df2: 1
##
##
## Data Values Comparison Summary
## ==============================
##
## Number of columns compared with ALL rows equal: 1
## Number of columns compared with SOME rows unequal: 3
## Number of columns with missing value differences: 1
##
## Columns with all rows equal : NUMB
##
## Summary of columns with some rows unequal:
##
##
##
## Column   Type (in df1)   Type (in df2)    # differences  Max difference    # NAs
```

```
## -------  --------------  --------------  --------------  ---------------  ------
## BOOL     logical         logical                      2  1                     0
## CHAR     character       character                    1  NA                    0
## INTG     integer         integer                      1                        1
##
##
##
## Unequal column details
## ======================
##
##
##
## #### Column -  BOOL
##
##
##
##      BOOL (df1)   BOOL (df2)   Type (df1)   Type (df2)    Difference
## ---  -----------  -----------  -----------  -----------  -----------
## 2    TRUE         FALSE        logical      logical                1
## 5    FALSE        TRUE         logical      logical               -1
##
##
## #### Column -  CHAR
##
##
##
##      CHAR (df1)   CHAR (df2)   Type (df1)   Type (df2)   Difference
## ---  -----------  -----------  -----------  -----------  -----------
## 5    E            I            character    character
##
##
## #### Column -  INTG
##
##
##
##       INTG (df1)   INTG (df2)  Type (df1)   Type (df2)    Difference
## ---  -----------  -----------  -----------  -----------  -----------
## 1              5           NA  integer      integer               NA
```

dfCompare::dfCompare

引入dfCompare (https://cran.r-project.org/web/packages/dfCompare/index.html)函式庫，再以 `dfCompare` 函式進行比較。

```
library(dfCompare)
dfCompare::dfCompare(df1, df2, c("char"))
```

```
## $DFDeletes
##   char intg     numb  bool tiangan
## 5    E    9 7.141593 FALSE      戊
##
## $DFAdds
##   char intg     numb  bool dizhi
## 5    I    9 7.141593  TRUE    辰
## 6    J   10       NA FALSE    巳
##
## $DFChanges
##   char tiangan intg.y           numb.y bool.y dizhi ident
## 1    A      甲   <NA> 3.14159265358979   TRUE    子 FALSE
## 2    B      乙      6 4.14159265358979  FALSE    丑 FALSE
## 3    C      丙      7 5.14159265358979   TRUE    寅 FALSE
## 4    D      丁      8 6.14159265358979  FALSE    卯 FALSE
```

diffdf::diffdf

引入diffdf (https://cran.r-project.org/web/packages/diffdf/index.html)函式庫，再以 `diffdf` 函式進行比較。

```
library(diffdf)
diffdf::diffdf(df1, df2)
```

```
## Warning in diffdf::diffdf(df1, df2):
## There are rows in COMPARE that are not in BASE !!
## There are columns in BASE that are not in COMPARE !!
## There are columns in COMPARE that are not in BASE !!
## Not all Values Compared Equal
```

```
## Differences found between the objects!
##
## A summary is given below.
##
## There are rows in COMPARE that are not in BASE !!
## All rows are shown in table below
##
##   ===============
##    ..ROWNUMBER..
##   ---------------
##          6
##   ---------------
##
## There are columns in BASE that are not in COMPARE !!
## All rows are shown in table below
##
##   =========
##    COLUMNS
##   ---------
##    tiangan
##   ---------
##
## There are columns in COMPARE that are not in BASE !!
## All rows are shown in table below
##
##   =========
##    COLUMNS
##   ---------
##     dizhi
##   ---------
##
## Not all Values Compared Equal
## All rows are shown in table below
##
##   ============================
##    Variable  No of Differences
##   ----------------------------
##      char            1
##      intg            1
##      bool            2
##   ----------------------------
##
##
## All rows are shown in table below
##
##   ======================================
##    VARIABLE  ..ROWNUMBER..  BASE  COMPARE
##   --------------------------------------
##      char          5         E       I
##   --------------------------------------
##
##
## All rows are shown in table below
##
##   ======================================
##    VARIABLE  ..ROWNUMBER..  BASE  COMPARE
##   --------------------------------------
```

```
##      intg          1         5     <NA>
##   ----------------------------------------
##
##
## All rows are shown in table below
##
##   ========================================
##    VARIABLE  ..ROWNUMBER..  BASE   COMPARE
##   ----------------------------------------
##      bool          2         TRUE   FALSE
##      bool          5        FALSE    TRUE
##   ----------------------------------------
```

diffobj::diffPrint

引入diffobj (https://cran.r-project.org/web/packages/diffobj/index.html)函式庫，再以 `diffPrint` 函式進行比較。

```
library(diffobj)
```

```
# using code chunk option results='asis' not work
diffobj::diffPrint(df1, df2) # html output in RStudio
```

```
## < df1
## > df2
## @@ 1,6 / 1,7 @@
## <   char intg     numb  bool tiangan
## >   char intg     numb  bool dizhi
## < 1    A    5 3.141593  TRUE      甲
## > 1    A   NA 3.141593  TRUE    子
## < 2    B    6 4.141593  TRUE      乙
## > 2    B    6 4.141593 FALSE    丑
## < 3    C    7 5.141593  TRUE      丙
## > 3    C    7 5.141593  TRUE    寅
## < 4    D    8 6.141593 FALSE      丁
## > 4    D    8 6.141593 FALSE    卯
## < 5    E    9 7.141593 FALSE      戊
## > 5    I    9 7.141593  TRUE    辰
## > 6    J   10       NA FALSE    巳
```

dplyr::all_equal

引入dplyr (https://cran.r-project.org/web/packages/dplyr/index.html)函式庫，再以 `all_equal` 函式進行比較。

```
suppressPackageStartupMessages(library(dplyr))
dplyr::all_equal(df1, df2)
```

```
## [1] "Cols in y but not x: `dizhi`. "   "Cols in x but not y: `tiangan`. "
```

Date & Session Info

```
Sys.Date()
```

```
## [1] "2018-09-15"
```

```
sessionInfo()
```

```
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17134)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Chinese (Traditional)_Taiwan.950
## [2] LC_CTYPE=Chinese (Traditional)_Taiwan.950
## [3] LC_MONETARY=Chinese (Traditional)_Taiwan.950
## [4] LC_NUMERIC=C
## [5] LC_TIME=Chinese (Traditional)_Taiwan.950
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] dplyr_0.7.6          diffobj_0.1.11       diffdf_1.0.1
##  [4] dfCompare_1.0.0      dataCompareR_0.1.1   daff_0.3.0
##  [7] bindrcpp_0.2.2       compareDF_1.5.0      compare_0.2-6
## [10] magrittr_1.5         RevoUtils_11.0.1     RevoUtilsMath_11.0.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.18     pillar_1.3.0     compiler_3.5.1   highr_0.7
##  [5] bindr_0.1.1      tools_3.5.1      digest_0.6.15    jsonlite_1.5
##  [9] evaluate_0.11    tibble_1.4.2     htmlTable_1.12   checkmate_1.8.5
## [13] pkgconfig_2.0.1  rlang_0.2.1      rstudioapi_0.7   curl_3.2
## [17] yaml_2.2.0       stringr_1.3.1    knitr_1.20       htmlwidgets_1.2
## [21] rprojroot_1.3-2  tidyselect_0.2.4 glue_1.3.0       R6_2.2.2
## [25] rmarkdown_1.10   purrr_0.2.5      tidyr_0.8.1      backports_1.1.2
## [29] htmltools_0.3.6  assertthat_0.2.0 V8_1.5           stringi_1.1.7
## [33] markdown_0.8     crayon_1.3.4
```