dplyr-data-frame-wrangling.R

julian

Wed Jun 17 13:42:07 2015

```
#!/usr/bin/env Rscript

# R script for using dplyr to manipulate data

# import dataset -------------------------------------------------------------

target.dir <- '~/GitHub/reproducible-research/extras'
target.file <- 'common-dataset.csv'

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(readr)

common.dataset <- read_csv(file.path(target.dir, target.file), col_names = TRUE)

# subsetting observations ----------------------------------------------------

a0a.greater.700 <- dplyr::filter(common.dataset, AOA > 700)

common.sample <- sample_n(common.dataset, 100, replace = FALSE)

common.sample.fac <- sample_frac(common.dataset, 0.8, replace = FALSE)


# subsetting columns ---------------------------------------------------------

contains.w.miniscule <- select(common.dataset, contains('w'))
contains.w.majuscule <- select(common.dataset, contains('W'))

endswith.rt <- select(common.dataset, ends_with('RT'))
endswith.z <- select(common.dataset, ends_with('z'))

startswith.a <- select(common.dataset, starts_with('A'))
startsth.k <- select(common.dataset, starts_with('K'))
```

```r
# get column names using regex
select(common.dataset, matches('.w.d.w'))
```

```
## Source: local data frame [15,000 x 0]
```

```r
# grouping data ------------------------------------------------------------

common.dataset %>% group_by(Cat) %>% summarize(meanA0A = mean(A0A))
```

```
## Source: local data frame [4 x 2]
##
##      Cat  meanA0A
## 1  first 41.48756
## 2 fourth 26.50641
## 3 second 61.14161
## 4  third 37.35486
```

```r
common.dataset %>% group_by(Cat) %>%
  summarize(meanA0A = mean(A0A), medianF5F = median(F5F), variL = var(iL))
```

```
## Source: local data frame [4 x 4]
##
##      Cat  meanA0A medianF5F     variL
## 1  first 41.48756  44.46894 1021673.3
## 2 fourth 26.50641  12.01603  971310.8
## 3 second 61.14161  13.80035  988919.5
## 4  third 37.35486  47.05954 1029877.4
```

```r
common.dataset %>% group_by(Cat, Part) %>% summarize(meanA0A = mean(A0A))
```

```
## Source: local data frame [4 x 3]
## Groups: Cat
##
##      Cat Part  meanA0A
## 1  first  one 41.48756
## 2 fourth  two 26.50641
## 3 second  two 61.14161
## 4  third  one 37.35486
```

```r
common.dataset %>% group_by(Cat, Part) %>%
  summarize(meanA0A = mean(A0A), medianF5F = median(F5F), variL = var(iL))
```

```
## Source: local data frame [4 x 5]
## Groups: Cat
##
##      Cat Part  meanA0A medianF5F     variL
## 1  first  one 41.48756  44.46894 1021673.3
## 2 fourth  two 26.50641  12.01603  971310.8
## 3 second  two 61.14161  13.80035  988919.5
## 4  third  one 37.35486  47.05954 1029877.4
```

```r
# database-style joins ------------------------------------------------

df1 <-
  data_frame(x1 = c('alpha', 'beta', 'gamma', 'delta'),
             x2 = c(1, 2, 3, 4))

df2 <-
  data_frame(x3 = c(FALSE, FALSE, TRUE, FALSE),
             x1 = c('alpha', 'beta', 'omicron', 'gamma'))

left_join(df1, df2, by = 'x1')
```

```
## Source: local data frame [4 x 3]
##
##      x1 x2    x3
## 1 alpha  1 FALSE
## 2  beta  2 FALSE
## 3 gamma  3 FALSE
## 4 delta  4    NA
```

```r
right_join(df1, df2, by = 'x1')
```

```
## Source: local data frame [4 x 3]
##
##        x1 x2    x3
## 1   alpha  1 FALSE
## 2    beta  2 FALSE
## 3 omicron NA  TRUE
## 4   gamma  3 FALSE
```

```r
inner_join(df1, df2, by = 'x1')
```

```
## Source: local data frame [3 x 3]
##
##      x1 x2    x3
## 1 alpha  1 FALSE
## 2  beta  2 FALSE
## 3 gamma  3 FALSE
```

```r
full_join(df1, df2, by = 'x1')
```

```
## Source: local data frame [5 x 3]
##
##        x1 x2    x3
## 1   alpha  1 FALSE
## 2    beta  2 FALSE
## 3   gamma  3 FALSE
## 4   delta  4    NA
## 5 omicron NA  TRUE
```

```r
merge(df1, df2) # this is a base R function
```

```
##       x1 x2    x3
## 1 alpha  1 FALSE
## 2  beta  2 FALSE
## 3 gamma  3 FALSE
```