How to use sub, gsub, matches, and grepl

Arin Parsa

7/24/2020

sub and gsub

sub() substitutes one occurrence of a pattern while gsub() substitutes all occurrences

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
name <- "Arin:Parsa"
sub(":", " ", name)
```

```
## [1] "Arin Parsa"
```

```r
name2 <- "George:Walker:Bush"
gsub(":", " ", name2)
```

```
## [1] "George Walker Bush"
```

```r
sentence <- "the weather is pleasant."
gsub("^t", "T", sentence)
```

```
## [1] "The weather is pleasant."
```

matches and grepl

matches() and grepl() can be used to match regex patterns. Regex is a regular expression, which is a sequence of characters used in search patterns. It contains literals as well as metacharacters such as ^ for beginning of a sentence, $ for end of sentence, and many more.

matches() can be used in select() in dplyr to match regex patterns to select column variables while grepl() can be used anywhere including in filter() for rows Below example shows how to use matches() and grepl()

```r
## Load GDP and educational datasets for countries
gdp <- read.csv("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv")
edu <- read.csv("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv")

gdp <- gdp[5:194,]

gdp <- gdp %>% rename(Rank = Gross.domestic.product.2012, CountryCode = X) %>% mutate(Rank = as.numeric
```

```r
## Merging the gdp and edu datasets
merge_df <- merge(gdp, edu, by = "CountryCode")

## Using matches() to select only the columns that have .Name in their titles
df_col_names <- merge_df %>% select(matches(".[Nn]ame"))
head(df_col_names, 5)
```

```
##                             Long.Name          Table.Name          Short.Name
## 1                             Aruba               Aruba               Aruba
## 2 Islamic State of Afghanistan         Afghanistan         Afghanistan
## 3  People's Republic of Angola             Angola              Angola
## 4           Republic of Albania             Albania             Albania
## 5         United Arab Emirates United Arab Emirates United Arab Emirates
```

```r
## Using grepl() to select the observations in the Special.Notes column that have the string "Fiscal ye
merge_df1 <- merge_df %>% filter(grepl("Fiscal year end: June", Special.Notes))
merge_df1
```

```
##     CountryCode Rank X.1              X.2        X.3 X.4 X.5 X.6 X.7 X.8
## 1           AUS   12  NA        Australia  1,532,408      NA  NA  NA  NA
## 2           BGD   59  NA       Bangladesh    116,355      NA  NA  NA  NA
## 3           BWA  117  NA         Botswana     14,504      NA  NA  NA  NA
## 4           EGY   38  NA Egypt, Arab Rep.    262,832      NA  NA  NA  NA
## 5           GMB  175  NA      Gambia, The        917      NA  NA  NA  NA
## 6           KEN   87  NA            Kenya     40,697      NA  NA  NA  NA
## 7           KWT   56  NA           Kuwait    160,913      NA  NA  NA  NA
## 8           PAK   44  NA         Pakistan    225,143      NA  NA  NA  NA
## 9           PRI   61  NA      Puerto Rico    101,496      NA  NA  NA  NA
## 10          SLE  157  NA     Sierra Leone      3,796      NA  NA  NA  NA
## 11          SWE   21  NA           Sweden    523,806      NA  NA  NA  NA
## 12          UGA  106  NA           Uganda     19,881      NA  NA  NA  NA
## 13          ZWE  134  NA         Zimbabwe      9,802      NA  NA  NA  NA
##                           Long.Name         Income.Group
## 1        Commonwealth of Australia     High income: OECD
## 2    People's Republic of Bangladesh           Low income
## 3             Republic of Botswana  Upper middle income
## 4            Arab Republic of Egypt  Lower middle income
## 5            Republic of The Gambia           Low income
## 6              Republic of Kenya            Low income
## 7               State of Kuwait High income: nonOECD
## 8    Islamic Republic of Pakistan  Lower middle income
## 9                   Puerto Rico High income: nonOECD
## 10       Republic of Sierra Leone           Low income
## 11            Kingdom of Sweden    High income: OECD
## 12            Republic of Uganda           Low income
## 13            Republic of Zimbabwe           Low income
##                         Region Lending.category Other.groups
## 1        East Asia & Pacific
## 2                 South Asia              IDA
## 3          Sub-Saharan Africa             IBRD
## 4  Middle East & North Africa             IBRD
## 5          Sub-Saharan Africa              IDA         HIPC
## 6          Sub-Saharan Africa              IDA
## 7  Middle East & North Africa
```

```
## 8             South Asia          Blend
## 9   Latin America & Caribbean
## 10        Sub-Saharan Africa          IDA       HIPC
## 11      Europe & Central Asia
## 12        Sub-Saharan Africa          IDA       HIPC
## 13        Sub-Saharan Africa        Blend
##            Currency.Unit Latest.population.census Latest.household.survey
## 1      Australian dollar                     2006
## 2      Bangladeshi taka                      2001               DHS, 2007
## 3        Botswana pula                       2001              MICS, 2000
## 4        Egyptian pound                      2006               DHS, 2008
## 5        Gambian dalasi                      2003            MICS, 2005/06
## 6       Kenyan shilling                      1999     DHS, 2003, SPA, 2004
## 7         Kuwaiti dinar                      2005                FHS, 1996
## 8        Pakistani rupee                     1998            DHS, 2006/07
## 9          U.S. dollar                       2000             RHS, 1995/96
## 10 Sierra Leonean leone                      2004                DHS 2008
## 11        Swedish krona           Register based
## 12      Ugandan shilling                     2002     DHS, 2006, SPA, 2007
## 13      Zimbabwe dollar                      2002             DHS, 2005/06
##                                                              Special.Notes
## 1  Fiscal year end: June 30; reporting period for national accounts data: FY.
## 2  Fiscal year end: June 30; reporting period for national accounts data: FY.
## 3  Fiscal year end: June 30; reporting period for national accounts data: FY.
## 4  Fiscal year end: June 30; reporting period for national accounts data: FY.
## 5  Fiscal year end: June 30; reporting period for national accounts data: CY.
## 6  Fiscal year end: June 30; reporting period for national accounts data: CY.
## 7  Fiscal year end: June 30; reporting period for national accounts data: CY.
## 8  Fiscal year end: June 30; reporting period for national accounts data: FY.
## 9  Fiscal year end: June 30; reporting period for national accounts data: FY.
## 10 Fiscal year end: June 30; reporting period for national accounts data: CY.
## 11 Fiscal year end: June 30; reporting period for national accounts data: CY.
## 12 Fiscal year end: June 30; reporting period for national accounts data: FY.
## 13 Fiscal year end: June 30; reporting period for national accounts data: CY.
##    National.accounts.base.year National.accounts.reference.year
## 1                                                          2007
## 2                    1995/1996                                NA
## 3                    1993/1994                                NA
## 4                    1991/1992                                NA
## 5                         1987                                NA
## 6                         2001                                NA
## 7                         1995                                NA
## 8                    1999/2000                                NA
## 9                         1954                                NA
## 10                        1990                                NA
## 11                                                         2000
## 12                   2001/2002                                NA
## 13                        1990                                NA
##    System.of.National.Accounts SNA.price.valuation
## 1                         1993                 VAB
## 2                         1993                 VAB
## 3                         1993                 VAB
## 4                           NA                 VAB
## 5                           NA                 VAB
```

```
## 6                             1993         VAB
## 7                               NA         VAP
## 8                             1993         VAB
## 9                               NA         VAP
## 10                            1993         VAB
## 11                              NA         VAB
## 12                              NA         VAB
## 13                              NA         VAB
##    Alternative.conversion.factor PPP.survey.year
## 1                                           2005
## 2                                           2005
## 3                                           2005
## 4                                           2005
## 5                                           2005
## 6                                           2005
## 7                                           2005
## 8                                           2005
## 9                                             NA
## 10                                          2005
## 11                                          2005
## 12                                          2005
## 13                     1991, 1998           2005
##    Balance.of.Payments.Manual.in.use External.debt.Reporting.status
## 1                               BPM5
## 2                               BPM5                    Preliminary
## 3                               BPM5                    Preliminary
## 4                               BPM5                         Actual
## 5                               BPM5                       Estimate
## 6                               BPM5                         Actual
## 7                               BPM5
## 8                               BPM5                         Actual
## 9
## 10                              BPM5                    Preliminary
## 11                              BPM5
## 12                              BPM5                         Actual
## 13                              BPM5                         Actual
##    System.of.trade Government.Accounting.concept
## 1          General                  Consolidated
## 2          General                  Consolidated
## 3          General                     Budgetary
## 4          Special                     Budgetary
## 5          General                  Consolidated
## 6          General                     Budgetary
## 7          Special                  Consolidated
## 8          General                  Consolidated
## 9          General
## 10         Special                     Budgetary
## 11         General                  Consolidated
## 12         General                     Budgetary
## 13         General                  Consolidated
##    IMF.data.dissemination.standard
## 1                             SDDS
## 2                             GDDS
## 3                             GDDS
```

```
## 4                                    SDDS
## 5                                    GDDS
## 6                                    GDDS
## 7                                    GDDS
## 8                                    GDDS
## 9
## 10                                   GDDS
## 11                                   SDDS
## 12                                   GDDS
## 13                                   GDDS
##    Source.of.most.recent.Income.and.expenditure.data
## 1                                       ES/BS, 1994
## 2                                        IHS, 2005
## 3                                     ES/BS, 1993/94
## 4                                     ES/BS, 2004-05
## 5                                        IHS, 2003
## 6                                      IHS, 2005-06
## 7
## 8                                     LSMS, 2004/05
## 9
## 10                                       IHS, 2003
## 11                                        IS, 2000
## 12                                        PS, 2005
## 13
##    Vital.registration.complete Latest.agricultural.census
## 1                          Yes                       2001
## 2                                                    2005
## 3                                                    1993
## 4                          Yes                  1999-2000
## 5                                               2001-2002
## 6                                               1977-1979
## 7                          Yes                       1970
## 8                                                    2000
## 9                          Yes                  1997/2002
## 10                                              1984-1985
## 11                         Yes                  1999-2000
## 12                                                   1991
## 13                                                   1960
##    Latest.industrial.data Latest.trade.data Latest.water.withdrawal.data
## 1                    2004              2008                         2000
## 2                    1997              2007                         2000
## 3                    2005              2008                         2000
## 4                    2001              2008                         2000
## 5                      NA              2008                         2000
## 6                    2005              2008                         2003
## 7                      NA              2007                         2002
## 8                      NA              2008                         2000
## 9                      NA                NA                           NA
## 10                     NA              2002                         2000
## 11                   2004              2008                         2000
## 12                   2001              2008                           NA
## 13                   1995              2008                         2002
##    X2.alpha.code WB.2.code   Table.Name   Short.Name
## 1             AU        AU    Australia    Australia
```

```
## 2            BD        BD        Bangladesh   Bangladesh
## 3            BW        BW          Botswana     Botswana
## 4            EG        EG Egypt, Arab Rep.        Egypt
## 5            GM        GM      Gambia, The   The Gambia
## 6            KE        KE           Kenya        Kenya
## 7            KW        KW          Kuwait       Kuwait
## 8            PK        PK        Pakistan     Pakistan
## 9            PR        PR     Puerto Rico  Puerto Rico
## 10           SL        SL    Sierra Leone Sierra Leone
## 11           SE        SE          Sweden       Sweden
## 12           UG        UG          Uganda       Uganda
## 13           ZW        ZW        Zimbabwe     Zimbabwe
```