StudyR3

PX

3/11/2020

Data Reshaping

```
# combine two vectors as columns in data.frame
sport <- c("Hockey","Baseball", "Football")
league <- c("NHL","MLB","NFL")
trophy <- c("Stanley Cup", "Commissioner's Trophy", "Vince Lombardi Trophy")

trophies1 <- cbind(sport, league,trophy)
trophies1
```

```
##      sport      league trophy
## [1,] "Hockey"   "NHL"  "Stanley Cup"
## [2,] "Baseball" "MLB"  "Commissioner's Trophy"
## [3,] "Football" "NFL"  "Vince Lombardi Trophy"
```

```
trophies2 <- data.frame( sport = c("Basketball", "Golf"), league = c("NBA", "PGA"),
                         trophy = c("Kobe Bryant Trophy",
                                    "Wanamaker Trophy"), stringsAsFactors = FALSE)
trophies2
```

```
##        sport league             trophy
## 1 Basketball    NBA Kobe Bryant Trophy
## 2       Golf    PGA   Wanamaker Trophy
```

```
trophies <- rbind(trophies1,trophies2)
trophies
```

```
##        sport league                trophy
## 1     Hockey    NHL           Stanley Cup
## 2   Baseball    MLB Commissioner's Trophy
## 3   Football    NFL Vince Lombardi Trophy
## 4 Basketball    NBA    Kobe Bryant Trophy
## 5       Golf    PGA      Wanamaker Trophy
```

```
# Joins, three common functions
# merge
# join in plyr
# merging function in data.table

download.file(url = "http://jaredlander.com/data/US_Foreign_Aid.zip",
              destfile = "~/Documents/OMSA/ISYE6501/datafolder/ForeignAid.zip")

unzip("~/Documents/OMSA/ISYE6501/datafolder/ForeignAid.zip", exdir = "data")
```

```
require(stringr)
```

```
## Loading required package: stringr
```

```
# first get a list of the files
theFiles <- dir("data/", pattern = "\\.csv")

# loop through those files
for (a in theFiles)
{
  # build names to assign data
  nameToUse <- str_sub( string = a, start = 12, end = 18)
  # read in the csv using data.table
  # file.path is a good way to specify a folder and file name

  temp<- read.table(file = file.path("data",a),
                    header = TRUE, sep = ",", stringsAsFactors = FALSE)

  assign(x = nameToUse, value = temp)
}

# Merge

Aid90s00s <- merge(x = Aid_90s, y = Aid_00s,
                   by.x = c("Country.Name","Program.Name"), by.y = c("Country.Name","Program.Name"))
# by.x specifies the key columns in left data.frame which is Aid_90s
# by.y does the same thing for Aid_00s
# The ablitiy to specify different column names for each data.frame
# is the most useful part in Merge
# The biggest drawback is computational time.

head(Aid90s00s)
```

```
##   Country.Name                                    Program.Name FY1990 FY1991
## 1  Afghanistan                     Child Survival and Health     NA     NA
## 2  Afghanistan         Department of Defense Security Assistance     NA     NA
## 3  Afghanistan                         Development Assistance     NA     NA
## 4  Afghanistan Economic Support Fund/Security Support Assistance     NA     NA
## 5  Afghanistan                              Food For Education     NA     NA
## 6  Afghanistan                Global Health and Child Survival     NA     NA
##   FY1992   FY1993  FY1994 FY1995 FY1996 FY1997 FY1998 FY1999 FY2000  FY2001
## 1     NA       NA      NA     NA     NA     NA     NA     NA     NA      NA
## 2     NA       NA      NA     NA     NA     NA     NA     NA     NA      NA
## 3     NA       NA      NA     NA     NA     NA     NA     NA     NA 4110478
## 4     NA 14178135 2769948     NA     NA     NA     NA     NA     NA   61144
## 5     NA       NA      NA     NA     NA     NA     NA     NA     NA      NA
## 6     NA       NA      NA     NA     NA     NA     NA     NA     NA      NA
##     FY2002    FY2003     FY2004     FY2005     FY2006     FY2007     FY2008
## 1  2586555  56501189   40215304   39817970   40856382   72527069   28397435
## 2  2964313        NA   45635526  151334908  230501318  214505892  495539084
## 3  8762080  54538965  180539337  193598227  212648440  173134034  150529862
## 4 31827014 341306822 1025522037 1157530168 1357750249 1266653993 1400237791
## 5       NA   3957312    2610006    3254408     386891         NA         NA
## 6       NA        NA         NA         NA         NA         NA   63064912
##       FY2009
```

```
## 1         NA
## 2  552524990
## 3    3675202
## 4 1418688520
## 5         NA
## 6    1764252
```

```
# Join in Plyr
require(plyr)
```

```
## Loading required package: plyr
```

```
Aid90s00sJoin <- join(x = Aid_90s, y = Aid_00s, by = c("Country.Name", "Program.Name"))

head(Aid90s00sJoin)
```

```
##   Country.Name                                      Program.Name FY1990 FY1991
## 1  Afghanistan                          Child Survival and Health     NA     NA
## 2  Afghanistan         Department of Defense Security Assistance     NA     NA
## 3  Afghanistan                             Development Assistance     NA     NA
## 4  Afghanistan Economic Support Fund/Security Support Assistance     NA     NA
## 5  Afghanistan                                 Food For Education     NA     NA
## 6  Afghanistan                   Global Health and Child Survival     NA     NA
##   FY1992   FY1993  FY1994 FY1995 FY1996 FY1997 FY1998 FY1999 FY2000  FY2001
## 1     NA       NA      NA     NA     NA     NA     NA     NA     NA      NA
## 2     NA       NA      NA     NA     NA     NA     NA     NA     NA      NA
## 3     NA       NA      NA     NA     NA     NA     NA     NA     NA 4110478
## 4     NA 14178135 2769948     NA     NA     NA     NA     NA     NA   61144
## 5     NA       NA      NA     NA     NA     NA     NA     NA     NA      NA
## 6     NA       NA      NA     NA     NA     NA     NA     NA     NA      NA
##     FY2002    FY2003     FY2004     FY2005     FY2006     FY2007     FY2008
## 1  2586555  56501189   40215304   39817970   40856382   72527069   28397435
## 2  2964313        NA   45635526  151334908  230501318  214505892  495539084
## 3  8762080  54538965  180539337  193598227  212648440  173134034  150529862
## 4 31827014 341306822 1025522037 1157530168 1357750249 1266653993 1400237791
## 5       NA   3957312    2610006    3254408     386891         NA         NA
## 6       NA        NA         NA         NA         NA         NA   63064912
##       FY2009
## 1         NA
## 2  552524990
## 3    3675202
## 4 1418688520
## 5         NA
## 6    1764252
```

```
# omit a few part using reduce , more details see Page 145-149

# data.table merge

require(data.table)
```

```
## Loading required package: data.table
```

```
dt90 <-data.table(Aid_90s, key = c("Country.Name","Program.Name"))
dt00 <-data.table(Aid_00s, key = c("Country.Name","Program.Name"))

# dt90 left side, dt00 right side.
```

```
dt0090 <- dt90[dt00]

# Reshape
# melting data (from column orientation to row orientation)
# casting data (from row orientation to column orientation)

# melt

head(Aid_00s)
```

```
##   Country.Name                                    Program.Name FY2000  FY2001
## 1  Afghanistan                      Child Survival and Health     NA      NA
## 2  Afghanistan         Department of Defense Security Assistance     NA      NA
## 3  Afghanistan                          Development Assistance     NA 4110478
## 4  Afghanistan Economic Support Fund/Security Support Assistance     NA   61144
## 5  Afghanistan                               Food For Education     NA      NA
## 6  Afghanistan                 Global Health and Child Survival     NA      NA
##     FY2002    FY2003     FY2004     FY2005     FY2006     FY2007     FY2008
## 1  2586555  56501189   40215304   39817970   40856382   72527069   28397435
## 2  2964313        NA   45635526  151334908  230501318  214505892  495539084
## 3  8762080  54538965  180539337  193598227  212648440  173134034  150529862
## 4 31827014 341306822 1025522037 1157530168 1357750249 1266653993 1400237791
## 5       NA   3957312    2610006    3254408     386891         NA         NA
## 6       NA        NA         NA         NA         NA         NA   63064912
##       FY2009
## 1         NA
## 2  552524990
## 3    3675202
## 4 1418688520
## 5         NA
## 6    1764252
```

```
require(reshape2)
```

```
## Loading required package: reshape2
```

```
##
## Attaching package: 'reshape2'
```

```
## The following objects are masked from 'package:data.table':
##
##     dcast, melt
```

```
melt00 <- melt(Aid_00s, id.vars = c("Country.Name","Program.Name"),
               variable.name = "Year", value.name = "Dollars")

# id.var specfies which columns uniquely identify a row.
tail(melt00,10)
```

```
##       Country.Name                                          Program.Name
## 24521     Zimbabwe                      Migration and Refugee Assistance
## 24522     Zimbabwe                                   Narcotics Control
## 24523     Zimbabwe Nonproliferation, Anti-Terrorism, Demining and Related
## 24524     Zimbabwe                           Other Active Grant Programs
## 24525     Zimbabwe                               Other Food Aid Programs
## 24526     Zimbabwe                                Other State Assistance
```

```
## 24527     Zimbabwe                                  Other USAID Assistance
## 24528     Zimbabwe                                             Peace Corps
## 24529     Zimbabwe                                                 Title I
## 24530     Zimbabwe                                                Title II
##         Year   Dollars
## 24521 FY2009   3627384
## 24522 FY2009        NA
## 24523 FY2009        NA
## 24524 FY2009   7951032
## 24525 FY2009        NA
## 24526 FY2009   2193057
## 24527 FY2009  41940500
## 24528 FY2009        NA
## 24529 FY2009        NA
## 24530 FY2009 174572685
```

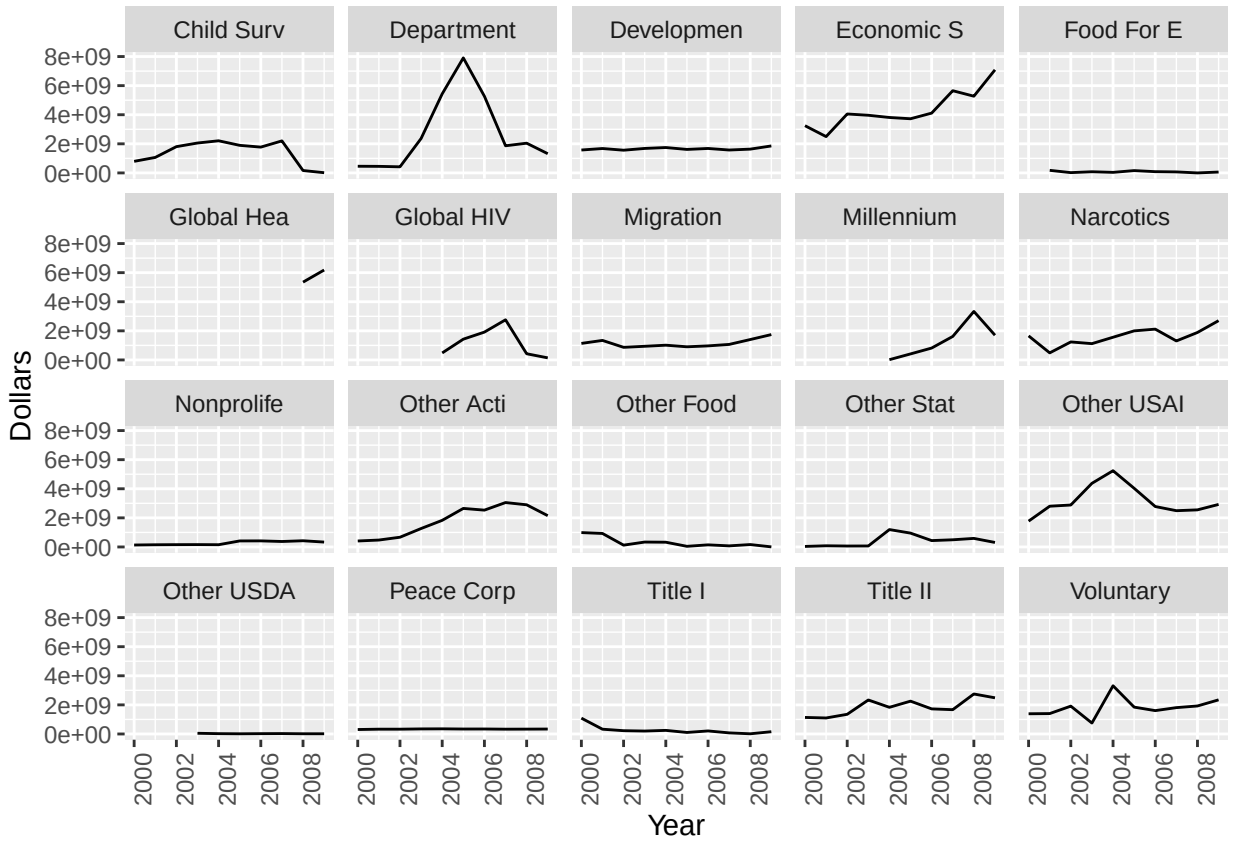
```
# some fancy plots are coming!
require(scales)
```

```
## Loading required package: scales
```

```
# strip the "FY" out of year and convert into numeric
melt00$Year <- as.numeric(str_sub(melt00$Year,start = 3,6))

# aggregate the data
meltAgg <-aggregate(Dollars ~ Program.Name + Year,
                    data = melt00, sum, na.rm=TRUE)

# Keep the first 10 characters of program name
# then it will fit the plot

meltAgg$Program.Name <- str_sub(meltAgg$Program.Name, start = 1, end = 10)

require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
ggplot(meltAgg, aes(x = Year, y = Dollars)) +
  geom_line(aes(group = Program.Name)) +
  facet_wrap(~ Program.Name) +
  scale_x_continuous(breaks = seq(from = 2000, to = 2009, by = 2)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust = 0))
```




```
 #+
  #scale_y_continuous(labels = multiple_formats(extra= dollar, multiple = "B"))

# Under current version, multiple_format does not work.

# DCAST

cast00 <- dcast(melt00, Country.Name + Program.Name ~ Year, value.var = "Dollars")

head(cast00)
```

```
##   Country.Name                                        Program.Name 2000    2001
## 1  Afghanistan                             Child Survival and Health   NA      NA
## 2  Afghanistan          Department of Defense Security Assistance   NA      NA
## 3  Afghanistan                               Development Assistance   NA 4110478
## 4  Afghanistan Economic Support Fund/Security Support Assistance   NA   61144
## 5  Afghanistan                                  Food For Education   NA      NA
## 6  Afghanistan                   Global Health and Child Survival   NA      NA
##       2002      2003       2004       2005       2006       2007       2008
## 1  2586555  56501189   40215304   39817970   40856382   72527069   28397435
## 2  2964313        NA   45635526  151334908  230501318  214505892  495539084
## 3  8762080  54538965  180539337  193598227  212648440  173134034  150529862
## 4 31827014 341306822 1025522037 1157530168 1357750249 1266653993 1400237791
## 5       NA   3957312    2610006    3254408     386891         NA         NA
## 6       NA        NA         NA         NA         NA         NA   63064912
##         2009
## 1         NA
## 2  552524990
```

```
## 3    3675202
## 4 1418688520
## 5          NA
## 6    1764252
```