{data.table} package in R

Fernando Figueiredo

ferdyfig@gmail.com

June 2011

Outline

- Data Management.
- Subsetting.
- Variable management.
- Aggregating.
- Merging.

Data Management

```
> library(data.table)
> txt1 <- "C:/Documents and Settings/Administrator/My Documents/asx2007.txt"
> txt2 <- "C:/Documents and Settings/Administrator/My Documents/asx2011.txt"
> asx2007 <- data.table(read.delim(txt1, header=TRUE))
> asx2011 <- data.table(read.delim(txt2, header=TRUE))
> setkey(asx2007, ASX.Code)
> setkey(asx2011, ASX.Code)
> tables()
     NAME     NROW MB
[1,] asx2007 2,313 1
[2,] asx2011 2,303 1
     COLS

[1,] Security.Description,ASX.Code,Last.Sale,Pos.or.Neg,Quote.Buy,Quote.Sell,…
[2,] Security.Description,ASX.Code,Last.Sale,Pos.or.Neg,Quote.Buy,Quote.Sell,…
     KEY
[1,] ASX.Code
[2,] ASX.Code
Total: 2MB
```

Data Management

```
> str(asx2007)
Classes ‘data.table’ and 'data.frame':  2313 obs. of  19 variables:
 $ Security.Description: Factor w/ 1993 levels "   5% cum pf",..: 287 160 234 170...
 $ ASX.Code            : Factor w/ 2313 levels "AAC","AAE","AAH",..: 1 2 3 4 5 6 7..
 $ Last.Sale           : Factor w/ 719 levels "-","0.001","0.002",..: 525 65 286 ...
 $ Pos.or.Neg          : Factor w/ 131 levels "-","-0.1","-0.2",..: 104 1 75 1 1 ...
 $ Quote.Buy           : Factor w/ 731 levels "-","0.001","0.002",..: 532 1 286 ...
 $ Quote.Sell          : Factor w/ 716 levels "-","0.001","0.002",..: 525 1 287 ...
 $ Volume.100s         : Factor w/ 993 levels "-","1","10","100",..: 55 1 583 1 ...
 $ Day.High            : Factor w/ 655 levels "-","0.001","0.002",..: 490 1 269 ...
 $ Day.Low             : Factor w/ 663 levels "-","0.001","0.002",..: 489 1 266 ...
 $ X52.week.High       : Factor w/ 838 levels "-","0.002","0.003",..: 637 178 342...
 $ X52.week.Low        : Factor w/ 713 levels "-","0.001","0.002",..: 413 46 331 ...
 $ Div.cents.per.Share : Factor w/ 274 levels "-","0.01","0.05",..: 60 1 1 231 ...
 $ Franked.Div         : Factor w/ 3 levels "","f","p": 1 1 1 1 1 1 1 1 1 1 ...
 $ Div.Times.Covered   : Factor w/ 308 levels "-","0.01","0.1",..: 8 1 1 273 1 1 ...
 $ Net.Tang.Assests    : Factor w/ 326 levels "-","-0.01","-0.02",..: 1 79 116 1 ...
 $ Div.Yield.Percent   : Factor w/ 429 levels "-","0.14","0.16",..: 258 1 1 42 1 ...
 $ Earn.Share.cents    : Factor w/ 1140 levels "-","-0.01","-0.02",..: 857 1082 1...
 $ PE.Ratio            : Factor w/ 362 levels "-","0.1","0.4",..: 29 5 11 85 1 1 ...
 $ Week.Percent.Move   : Factor w/ 778 levels "-","-0.05","-0.08",..: 36 1 502 1 ...
 - attr(*, "sorted")= chr "ASX.Code"
```

Data Management

```
> asx2007
       Security.Description ASX.Code Last.Sale Pos.or.Neg Quote.Buy
 [1,]      Aust Agriculture      AAC      3.22          3      3.22
 [2,]       Agri Energy Ltd      AAE     0.065          -         -
 [3,]  Arana Therapeutics L      AAH     1.185        0.5      1.15
 [4,]         Alcoa Inc cdi      AAI      49.5          -        46
 [5,]            A1 Min Ltd      AAM     0.285          -      0.29
 [6,]             opt nov08     AAMO      0.12          -       0.1
 [7,]            Aasia Gold      AAO     0.115          -     0.115
 [8,]             opt jun08     AAOO     0.015          -      0.01
 [9,]        Australis Aqua      AAQ      0.42       -0.5      0.42
[10,]            Anglo Aust      AAR     0.087       -0.1     0.087
…
First 10 rows of 2313 printed.
```

Or use

➢**View(asx2007)**

For a nice display under Windows

Subsetting

```
# Let’s list all companies with the same Price Earnings Ratio and ignore undefined
# excluding unique values
> setkey(asx2007,PE.Ratio)
> asx2007[duplicated(asx2007)] [PE.Ratio != "-"]
       Security.Description ASX.Code ... PE.Ratio
 [1,]        Premier Invest      PMV ...     1.2
 [2,]  London City Equities      LCE ...     1.4
 [3,]             Cluff Res      CFR ...     1.4
 [4,]    Consolidated Media      CMJ ...     1.5
 [5,]  Arana Therapeutics L      AAH ...     1.5
 [6,]       Asset Loans Ltd      ASQ ...     1.5
 [7,]            Eldore Min      EDM ...     1.6
 [8,]     Centro Prop stpld      CNP ...     1.7
 [9,]      Seven Network         SEV ...     1.8
[10,]    Tishman Speyer unt      TSO ...     1.9
First 10 rows of 392 printed.
> same_PE_Ratio <- asx2007[duplicated(asx2007)] [PE.Ratio != "-"]
> last(same_PE_Ratio)
      Security.Description ASX.Code Last.Sale Pos.or.Neg Quote.Buy
[1,]  Centrepoint Alliance      CAF      0.45          1       0.4
     Quote.Sell Volume.100s Day.High Day.Low X52.week.High
[1,]       0.45          10     0.45    0.45          1.26
     …
```

Subsetting

```
# Let’s select shares dealing in global indexes

> asx2007[Security.Description %like% "MSCI"]
            Security.Description ASX.Code Last.Sale Pos.or.Neg
[1,]         iShares MSCI HK cdi      IHK     25.12        -47
[2,]      iShares MSCI Japan cdi      IJP     15.16         10
[3,]       iShares MSCI Sing cdi      ISG     15.61          -
[4,]       iShares MSCI EAFE cdi      IVE      89.6        -82
[5,]   iShares MSCI Em Mkts cdi       IEM       175       -352
[6,]     iShares MSCI Taiwan cdi      ITW     17.15          9
[7,]     iShares MSCI SKorea cdi      IKO     74.21        -99

# Another set, this time all fully franked dividends
asx2007[Franked.Div == "f"]
 Security.Description    …ASX.Code    …Franked.Div
 [1,]           Advent Ltd      ADT           f
 [2,]          ITX Grp Ltd      ITX           f
 [3,]       Plan B Grp Hld      PLB           f
 [4,]   Deep Sea Fisheries      DSF           f
…
First 10 rows of 382 printed.
```

Subsetting

```
# Let’s select first three columns.

> asx2007[, list(Security.Description, ASX.Code, Last.Sale)]
        Security.Description ASX.Code Last.Sale
 [1,]              opt nov08     AAMO      0.12
 [2,]             Aasia Gold      AAO     0.115
 [3,]              opt jun08     AAOO     0.015
 [4,]            Autron Corp      AAT     0.069
 [5,]              optdec10d     ABQO         -
 [6,]       ADV Braking Tech      ABV     0.041

# all ASX.Code starting with AA
> asx2007[, list(Security.Description, ASX.Code, Last.Sale)] [ASX.Code %like% "^AA"]
       Security.Description ASX.Code Last.Sale
 [1,]             opt nov08     AAMO      0.12
 [2,]            Aasia Gold      AAO     0.115
 [3,]             opt jun08     AAOO     0.015
 [4,]           Autron Corp      AAT     0.069
 [5,]       Agri Energy Ltd      AAE     0.065
```

Subsetting

```
# Let’s use binary search (extremely fast) to extract rows.
```

- `setkey(asx2007, ASX.Code)`
- `asx2007["NAB"]   # extracts NAB from key`

```
# Now we going to join 2 tables, using a binary search rather than
   vector search

setkey(asx2007, PE.Ratio, Div.cents.per.Share)

asx2007[J("10.9","6")] # J is a short cut for Join function

PE.Ratio Div.cents.per.Share Security.Description ASX.Code
    10.9                  6            Amcil Ltd      AMH
    10.9                  6       Flat Glass Ind      FGI
```

Variable Management

```
# Let’s globally change “-” to NA.

> library(gdata)
> asx2007 <- unknownToNA(asx2007, unknown="-") # NAToUnknown(x, unknown=“-”)

> str(asx2007)
Classes ‘data.table’ and 'data.frame':  2313 obs. of  19 variables:
 $ Security.Description: Factor w/ 1993 levels "   5% cum pf",..: 76 115 61 323 ...
 $ ASX.Code            : Factor w/ 2313 levels "AAC","AAE","AAH",..: 6 7 8 12 21 ...
 $ Last.Sale           : Factor w/ 718 levels "0.001","0.002",..: 99 98 15 68 NA ...
 $ Pos.or.Neg          : Factor w/ 130 levels "-0.1","-0.2",..: NA NA NA NA ...
```

Variable Management

```
# Let’s change numeric factors to numeric fields.
> asx2007 <- transform(asx2007, Last.Sale=as.numeric(as.character(Last.Sale)),
+            Pos.or.Neg=as.numeric(as.character(Pos.or.Neg)),
+            Quote.Buy=as.numeric(as.character(Quote.Buy)),
+            Quote.Sell=as.numeric(as.character(Quote.Sell)),
+            Volume.100s=as.numeric(as.character(Volume.100s)),
+            Day.High=as.numeric(as.character(Day.High)),
+            Day.Low=as.numeric(as.character(Day.Low)),
+            X52.week.High=as.numeric(as.character(X52.week.High)),
+            X52.week.Low=as.numeric(as.character(X52.week.Low)),
+            Div.cents.per.Share=as.numeric(as.character(Div.cents.per.Share)),
+            Div.Times.Covered=as.numeric(as.character(Div.Times.Covered)),
+            Net.Tang.Assests=as.numeric(as.character(Net.Tang.Assests)),
+            Div.Yield.Percent=as.numeric(as.character(Div.Yield.Percent)),
+            Earn.Share.cents=as.numeric(as.character(Earn.Share.cents)),
+            PE.Ratio=as.numeric(as.character(PE.Ratio)),
+            Week.Percent.Move=as.numeric(as.character(Week.Percent.Move)))
> str(asx2007)
Classes ‘data.table’ and 'data.frame':  2313 obs. of  19 variables:
 $ Security.Description: Factor w/ 1993 levels "   5% cum pf",..: 287 160 ...
 $ ASX.Code            : Factor w/ 2313 levels "AAC","AAE","AAH",..: 1 2 3 ...
 $ Last.Sale           : num  3.22 0.065 1.185 49.5 0.285 ...
 $ Pos.or.Neg          : num  3 NA 0.5 NA NA NA NA NA -0.5 -0.1 ...
```

Variable Management

```
# Let’s rename and delete a column.

> asx2007 <- transform(asx2007, Net.Tang.Assets=Net.Tang.Assests,
+                           Net.Tang.Assests=NULL)

$ Week.Percent.Move   : num  -0.62 NA 2.16 NA -1.72 NA NA NA NA 1.16 ...
$ Net.Tang.Assets     : num  NA 0.35 0.76 NA NA NA NA NA 0.17 NA ...

# To delete a variable only needed to set it to NULL,
# but you can use a more verbose function

➢library(gdata)
➢rename.vars(data, from="", to="", info=FALSE)
➢remove.vars(data, names="", info=FALSE)

# Let’s create a new variable (reward) being the ratio of max price over 52 weeks
divided by the min price over 52 weeks
> asx2007 <- transform(asx2007, reward=round(X52.week.High / X52.week.Low,
digits=2))
> asx2007[,list(ASX.Code, reward)]
 ASX.Code reward
 [1,]      AAC   1.98
 [2,]      AAE   8.89
 [3,]      AAH   1.10
```

Aggregating

```
# Let’s check how many entries where 52 weeks high is less than 52 weeks low

> asx2007[,table(na.omit(reward) < 1)]
FALSE
 2273

# Let’s look at univariate stats

> asx2007[,summary(na.omit(reward))]
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.000   1.660   2.380   3.548   3.620 105.000

# Let’s check max 105
 Security.Description ASX.Code X52.week.High X52.week.Low
[1,]        opt apr08     BLTO         0.105        0.001

# Let’s assign a return based on assumption buying at low and sell at high.
> asx2007 <- transform(asx2007, Return=ifelse(reward < 1.04, "Poor",
+                        ifelse(reward %between% c(1.05, 1.06), "Bank", "Good")))
```

Aggregating

```
# Let’s summarise some fields by Return.

> asx2007[,list(
+  Avg.Div.cents.per.Share=round(mean(Div.cents.per.Share, na.rm = TRUE), digits=2),
+  Avg.Earn.Share.cents=round(mean(Earn.Share.cents, na.rm = TRUE), digits=2),
+  Avg.PE.Ratio=round(mean(PE.Ratio, na.rm = TRUE), digits=2)),
+  by=Return][!is.na(Return)]

     Return Avg.Div.cents.per.Share Avg.Earn.Share.cents Avg.PE.Ratio
[1,]   Bank                    9.30                 9.30        22.00
[2,]   Good                   22.33                12.12        48.58
[3,]   Poor                    9.00                -1.35          NaN

# Let’s look at the Return frequencies
> asx2007[,table(Return)]
Return
Bank Good Poor
   1 2260   12 ← 2007
 245 1346  710 ← 2011
```

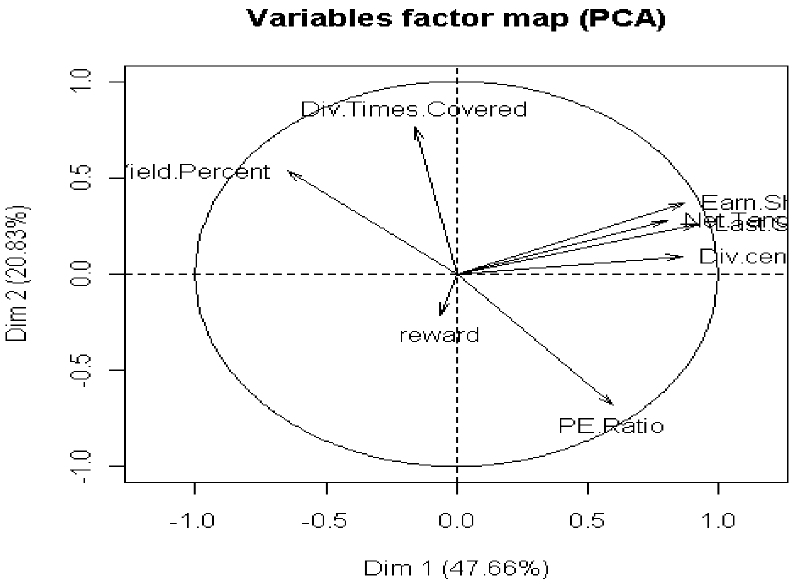
Merging

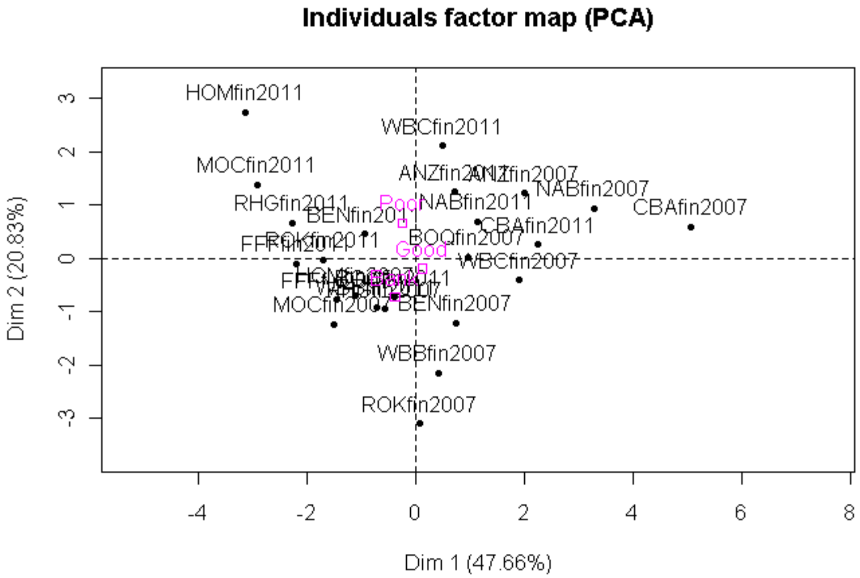
```
# Let’s merge asx2007 and asx2011 by ASX.Code.
# First let’s set the table keys and find the number of rows.
> setkey(asx2007, ASX.Code)
> setkey(asx2011, ASX.Code)
> tables()
     NAME           NROW MB
[1,] asx2007       2,313 1
[2,] asx2011       2,303 1

# Inner join.
> inner_join <- merge(asx2007, asx2011)
> tables()
     NAME           NROW MB
[1,] asx2007       2,313 1
[2,] asx2011       2,303 1
[3,] inner_join    1,427 1
```

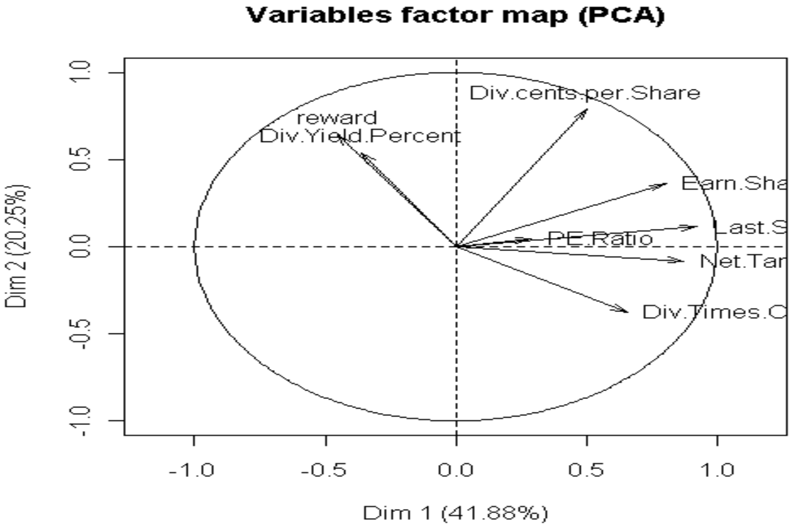
Merging

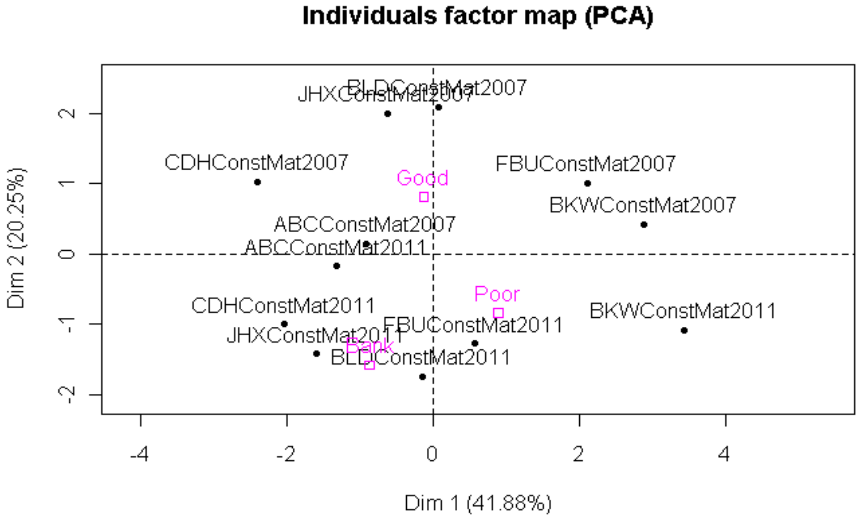
```
# Let’s merge asx2007 and asx2011 by ASX.Code.

# Left join.
> left_join <- merge(asx2011, asx2007, all.x=TRUE)
> tables()
     NAME           NROW MB
[1,] asx2007       2,313 1
[2,] asx2011       2,303 1
[3,] inner_join    1,427 1
[4,] left_join     2,303 1

# Outer join.
> outer_join <- merge(asx2007, asx2011, all=TRUE)
Warning message:
In rbind(deparse.level, ...) :
  colnames of argument 2 don't match colnames of argument 1
> tables()
     NAME           NROW MB
[1,] asx2007       2,313 1
[2,] asx2011       2,303 1
[3,] inner_join    1,427 1
[4,] left_join     2,303 1
[5,] outer_join    3,189 2
```

Principal Component Analysis









Acknowledgments

- Package Data.table – Matthew Dowle
- FactoMineR: Sebastien Le, Julie Josse, Francois Husson
- Package gdata – Gregory R. Warnes et al