Reading SAS7BDAT Databases in R, without SAS!

Matt Shotwell, Vanderbilt University

What is a SAS7BDAT database?

A SAS7BDAT database

- is created by SAS to store data
- has the file extension `*.sas7bdat`
- has an unpublished binary format
- until recently, required SAS software to use
- uses disk space inefficiently

VANDERBILT
UNIVERSITY

SAS7BDAT Disk Usage

- 142 SAS7BDAT datasets, totaling 541MB in size
- compressed using the three algorithms supported by R
- rationale: disk-efficient data files don't compress well

type	size	reduction
gzip -9	30.3MB	94.3%
bzip2 -9	19.0MB	96.4%
xz -9	12.8MB	97.6%

VANDERBILT
UNIVERSITY

SAS7BDAT Disk Usage (continued...)

- 142 SAS7BDAT datasets, totaling 541MB in size
- compressed using the three algorithms supported by R
- rationale: disk-efficient data files don't compress well

type	size	reduction
gzip -9	30.3MB	94.3%
bzip2 -9	19.0MB	96.4%
xz -9	12.8MB	97.6%

VANDERBILT
UNIVERSITY

The sas7bdat Package for R

The sas7bdat package

- is a *compatibility study* of the SAS7BDAT format
- documents the SAS7BDAT format: `vignette('sas7bdat')`
- includes an experimental SAS7BDAT database reader
- was independently ported to
  - Java: http://eobjects.org/svn/SassyReader
  - ActionScript: http://code.google.com/p/sasquatch
- maintains a list of free internet sas7bdat data sources
- is hosted on
  - CRAN: http://cran.r-project.org/web/packages/sas7bdat
  - GitHub: https://github.com/biostatmatt/sas7bdat

VANDERBILT
UNIVERSITY

The sas7bdat Package for R (conintued...)

The sas7bdat package

- is a *compatibility study* of the SAS7BDAT format
- documents the SAS7BDAT format: `vignette('sas7bdat')`
- includes an experimental SAS7BDAT database reader
- was independently ported to
  - Java: http://eobjects.org/svn/SassyReader
  - ActionScript: http://code.google.com/p/sasquatch
- maintains a list of free internet sas7bdat data sources
- is hosted on
  - CRAN: http://cran.r-project.org/web/packages/sas7bdat
  - GitHub: https://github.com/biostatmatt/sas7bdat

VANDERBILT
UNIVERSITY

The sas7bdat Package for R (continued.....)

The sas7bdat package

- is a *compatibility study* of the SAS7BDAT format
- documents the SAS7BDAT format: `vignette('sas7bdat')`
- includes an experimental SAS7BDAT database reader
- was independently ported to
  - Java: http://eobjects.org/svn/SassyReader
  - ActionScript: http://code.google.com/p/sasquatch
- maintains a list of free internet sas7bdat data sources
- is hosted on
  - CRAN: http://cran.r-project.org/web/packages/sas7bdat
  - GitHub: https://github.com/biostatmatt/sas7bdat

VANDERBILT
UNIVERSITY

Reading SAS7BDAT Databases in R

```
R> install.packages('sas7bdat')
R> library('sas7bdat')
R> read.sas7bdat('hotel.sas7bdat')
             LOCATION ROOMS MIN MAX DINING LOUNGE BKFST POOL TYPE
1        Capital Hill   341  69 139    yes    yes    no  yes   QI
2            Downtown   135  79 169    yes    yes    no   no   QI
3            Downtown   197  94 164    yes    yes    no   no   CI
4           Bowie, MD   110  59 135    yes    yes    no  yes   CI
5         Clinton, MD    94  58  70     no     no   yes   no   CI
6    College Park, MD   154  54 109     no     no   yes  yes   QI
7    Gaithersburg, MD   127  59  79     no     no   yes  yes   CI
8  Landover Hills, MD    84  44  64     no     no   yes   no   CI
9          Laurel, MD   118  59 150     no     no   yes  yes   CI
10      Rockville, MD   162  59  74    yes    yes    no  yes   CH
11  Silver Spring, MD   254  59 125    yes    yes    no  yes   QI
12     Alexandria, VA   207  60  99    yes    yes    no  yes   QI
13     Alexandria, VA   148  49  86    yes    yes   yes  yes   CI
14     Alexandria, VA   188  65  85    yes     no   yes  yes   CI
15     Alexandria, VA    92  49  72     no     no   yes  yes   CI
16      Arlington, VA   141  59  91    yes    yes    no  yes   QI
17      Arlington, VA   398  39 149    yes    yes    no  yes   QI
18      Arlington, VA   126  64  95    yes    yes    no   no   CI
19         Dulles, VA   140  59  85    yes     no   yes  yes   CI
20         Dulles, VA   103  59 105     no     no   yes   no   CI
21        Fairfax, VA   212  49  95    yes    yes   yes  yes   CI
22   Falls Church, VA   121  62  72    yes     no    no  yes   QI
23   Falls Church, VA   109  50  75    yes     no   yes  yes   QI
24         Vienna, VA   250  45  79     no     no   yes  yes   CI
25     Woodbridge, VA    94  59  65     no     no   yes   no   CI
```

VANDERBILT
UNIVERSITY

SAS7BDAT Meta Information

```
R> attributes(read.sas7bdat('oz.sas7bdat'))$column.info
[[1]]
[[1]]$name
[1] "consumption"

[[1]]$label
[1] "aggregate consumption in Australia"

[[1]]$offset
[1] 0

[[1]]$length
[1] 8

[[1]]$type
[1] "numeric"


[[2]]
[[2]]$name
[1] "income"

[[2]]$label
[1] "disposable income in Australia"

[[2]]$offset
[1] 8

[[2]]$length
[1] 8

[[2]]$type
[1] "numeric"
```

VANDERBILT
UNIVERSITY

SAS7BDAT Meta Information (continued...)

```
R> attributes(read.sas7bdat('oz.sas7bdat'))$column.info
[[1]]
[[1]]$name
[1] "consumption"

[[1]]$label
[1] "aggregate consumption in Australia"

[[1]]$offset
[1] 0

[[1]]$length
[1] 8

[[1]]$type
[1] "numeric"


[[2]]
[[2]]$name
[1] "income"

[[2]]$label
[1] "disposable income in Australia"

[[2]]$offset
[1] 8

[[2]]$length
[1] 8

[[2]]$type
[1] "numeric"
```

VANDERBILT
UNIVERSITY

SAS7BDAT Meta Information (continued.....)

```
R> attributes(read.sas7bdat('oz.sas7bdat'))$column.info
[[1]]
[[1]]$name
[1] "consumption"

[[1]]$label
[1] "aggregate consumption in Australia"

[[1]]$offset
[1] 0

[[1]]$length
[1] 8

[[1]]$type
[1] "numeric"


[[2]]
[[2]]$name
[1] "income"

[[2]]$label
[1] "disposable income in Australia"

[[2]]$offset
[1] 8

[[2]]$length
[1] 8

[[2]]$type
[1] "numeric"
```

VANDERBILT
UNIVERSITY

SAS7BDAT Format TODO

- platforms other than 32bit Windows
- compression, encryption
- non-ASCII encoding

sas7bdat Package TODO

- C/C++ reader
- regression testing
- contribution to foreign?

`sas7bdat` Package `TODONT`

- implement `write.sas7bdat`
- knowingly infringe on SAS IP

VANDERBILT
UNIVERSITY

‘I Want You’ to help develop `sas7bdat`

VANDERBILT
UNIVERSITY