haven

INTERMEDIATE IMPORTING DATA IN R

Filip Schouwenaars
Instructor, DataCamp

datacamp

Statistical Software Packages

Package
SAS
STATA
SPSS

Statistical Software Packages

Package	Expanded Name
SAS	Statistical Analysis Software
STATA	STAtistics and daTA
SPSS	Statistical Package for Social Sciences

Statistical Software Packages

Package	Expanded Name	Application
SAS	Statistical Analysis Software	Business Analytics Biostatistics Medical Sciences
STATA	STAtistics and daTA	Economists
SPSS	Statistical Package for Social Sciences	Social Sciences

Statistical Software Packages

Package	Expanded Name	Application	Data File Extensions
SAS	Statistical Analysis Software	Business Analytics Biostatistics Medical Sciences	.sas7bdat .sas7bcat
STATA	STAtistics and daTA	Economists	.dta
SPSS	Statistical Package for Social Sciences	Social Sciences	.sav .por

R packages to import data

- haven
  - Hadley Wickham
  - Goal: consistent, easy, fast
- foreign
  - R Core Team
  - Support for many data formats

haven

- SAS, STATA and SPSS
- ReadStat: C library by Evan Miller
- Extremely simple to use
- Single argument: path to file
- Result: R data frame

```
install.packages("haven")
library(haven)
```

SAS data

- ontime.sas7bdat
  - Delay statistics for airlines in US
- `read_sas()`

```
ontime <- read_sas("ontime.sas7bdat")
```

SAS data

```
ontime <- read_sas("ontime.sas7bdat")
str(ontime)
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':    10 obs. of  4 variables:
 $ Airline    : atomic  TWA Southwest Northwest ...
  ..- attr(*, "label")= chr "Airline"
 $ March_1999 : atomic  84.4 80.3 80.8 72.7 78.7 ...
  ..- attr(*, "label")= chr "March 1999"
 $ June_1999  : atomic  69.4 77 75.1 65.1 72.2 ...
  ..- attr(*, "label")= chr "June 1999"
 $ August_1999: atomic  85 80.4 81 78.3 77.7 75.1 ...
  ..- attr(*, "label")= chr "August 1999"
```

SAS data

```
ontime <- read_sas("ontime.sas7bdat")
ontime
```

```
          Airline March_1999 June_1999 August_1999
1             TWA       84.4      69.4        85.0
2       Southwest       80.3      77.0        80.4
3       Northwest       80.8      75.1        81.0
4        American       72.7      65.1        78.3
5           Delta       78.7      72.2        77.7
6     Continental       79.3      68.4        75.1
7          United       78.6      69.2        71.6
8      US Airways       73.6      68.9        70.1
9          Alaska       71.9      75.4        64.4
10 American West       76.5      70.3        62.5
```

SAS data

```
ontime <- read_sas("ontime.sas7bdat")
```

	Airline Airline	March_1999 March 1999	June_1999 June 1999	August_1999 August 1999
1	TWA	84.4	69.4	85.0
2	Southwest	80.3	77.0	80.4
3	Northwest	80.8	75.1	81.0
4	American	72.7	65.1	78.3
5	Delta	78.7	72.2	77.7
6	Continental	79.3	68.4	75.1
7	United	78.6	69.2	71.6
8	US Airways	73.6	68.9	70.1
9	Alaska	71.9	75.4	64.4
10	American West	76.5	70.3	62.5

SAS data

```
ontime <- read_sas("ontime.sas7bdat")
```

	Airline Airline	March_1999 March 1999	June_1999 June 1999	August_1999 August 1999
1	TWA	84.4	69.4	85.0
2	Southwest	80.3	77.0	80.4
3	Northwest	80.8	75.1	81.0
4	American	72.7	65.1	78.3
5	Delta	78.7	72.2	77.7
6	Continental	79.3	68.4	75.1
7	United	78.6	69.2	71.6
8	US Airways	73.6	68.9	70.1
9	Alaska	71.9	75.4	64.4
10	American West	76.5	70.3	62.5

SAS data

```
ontime <- read_sas("ontime.sas7bdat")
```

	Airline Airline	March_1999 March 1999	June_1999 June 1999	August_1999 August 1999
1	TWA	84.4	69.4	85.0
2	Southwest	80.3	77.0	80.4
3	Northwest	80.8	75.1	81.0
4	American	72.7	65.1	78.3
5	Delta	78.7	72.2	77.7
6	Continental	79.3	68.4	75.1
7	United	78.6	69.2	71.6
8	US Airways	73.6	68.9	70.1
9	Alaska	71.9	75.4	64.4
10	American West	76.5	70.3	62.5

STATA data

- `STATA 13` & `STATA 14`
- `read_stata()`, `read_dta()`

STATA data

```
ontime <- read_stata("ontime.dta")
ontime <- read_dta("ontime.dta")
ontime
```

```
   Airline March_1999 June_1999 August_1999
1        8       84.4      69.4        85.0
2        7       80.3      77.0        80.4
3        6       80.8      75.1        81.0
4        2       72.7      65.1        78.3
5        5       78.7      72.2        77.7
6        4       79.3      68.4        75.1
7        9       78.6      69.2        71.6
8       10       73.6      68.9        70.1
9        1       71.9      75.4        64.4
10       3       76.5      70.3        62.5
```

STATA data

```
ontime <- read_stata("ontime.dta")
ontime <- read_dta("ontime.dta")
```

```
# R version of common data structure
class(ontime$Airline)
```

```
"labelled"
```

```
ontime$Airline
```

```
<Labelled>
8  7  6  2  5  4  9 10  1  3
attr(,"label")
"Airline"
Labels:
       Alaska   American  American West  ...  US Airways
            1          2              3  ...          10
```

as_factor()

```
ontime <- read_stata("ontime.dta")
ontime <- read_dta("ontime.dta")
```

```
as_factor(ontime$Airline)
```

```
TWA    Southwest  Northwest   American ... American West
Levels: Alaska American American West ... US Airways
```

```
as.character(as_factor(ontime$Airline))
```

```
"TWA" "Southwest" "Northwest" ... "American West"
```

as_factor()

```
ontime$Airline <- as.character(as_factor(ontime$Airline))
ontime
```

```
          Airline March_1999 June_1999 August_1999
1             TWA       84.4      69.4        85.0
2       Southwest       80.3      77.0        80.4
3       Northwest       80.8      75.1        81.0
4        American       72.7      65.1        78.3
5           Delta       78.7      72.2        77.7
6     Continental       79.3      68.4        75.1
7          United       78.6      69.2        71.6
8      US Airways       73.6      68.9        70.1
9          Alaska       71.9      75.4        64.4
10 American West       76.5      70.3        62.5
```

SPSS data

- `read_spss()`
- .por -> `read_por()`
- .sav -> `read_sav()`

```
read_sav(file.path("~","datasets","ontime.sav"))
```

```
   Airline Mar.99 Jun.99 Aug.99
1        8   84.4   69.4   85.0
2        7   80.3   77.0   80.4
3        6   80.8   75.1   81.0
4        2   72.7   65.1   78.3
5        5   78.7   72.2   77.7
...
10       3   76.5   70.3   62.5
```

Statistical Software Packages

Package	Expanded Name	Application	Data File Extensions	haven function
SAS	Statistical Analysis Software	Business Analytics Biostatistics Medical Sciences	.sas7bdat .sas7bcat	read_sas()
STATA	STAtistics and daTA	Economists	.dta	read_dta() read_stata()
SPSS	Statistical Package for Social Sciences	Social Sciences	.sav .por	read_spss() read_por() read_sav()

Let's practice!

INTERMEDIATE IMPORTING DATA IN R

foreign

INTERMEDIATE IMPORTING DATA IN R

Filip Schouwenaars
Instructor, DataCamp

datacamp

foreign

- R Core Team
- Less consistent
- Very comprehensive
- All kinds of foreign data formats
- SAS, STATA, SPSS, Systat, Weka ...

```
install.packages("foreign")
library(foreign)
```

SAS

- Cannot import `.sas7bdat`
- Only SAS libraries: `.xport`
- `sas7bdat` package

STATA

- STATA 5 to 12
- `read.dta()` - `read.dta()`

```
read.dta(file,
         convert.factors = TRUE,
         convert.dates = TRUE,
         missing.type = FALSE)
```

read.dta()

```
ontime <- read.dta("ontime.dta")
ontime
```

```
          Airline March_1999 June_1999 August_1999
1             TWA       84.4      69.4        85.0
2       Southwest       80.3      77.0        80.4
3       Northwest       80.8      75.1        81.0
4        American       72.7      65.1        78.3
5           Delta       78.7      72.2        77.7
6     Continental       79.3      68.4        75.1
7          United       78.6      69.2        71.6
8      US Airways       73.6      68.9        70.1
9          Alaska       71.9      75.4        64.4
10 American West       76.5      70.3        62.5
```

read.dta()

```
ontime <- read.dta("ontime.dta")
str(ontime)
```

- `convert.factors` TRUE by default

```
'data.frame':    10 obs. of  4 variables:
 $ Airline    : Factor w/ 10 levels "Alaska",..: 8 7 6 2 5 4 ...
 $ March_1999 : num  84.4 80.3 80.8 72.7 78.7 79.3 78.6 ...
 $ June_1999  : num  69.4 77 75.1 65.1 72.2 68.4 69.2 68.9 ...
 $ August_1999: num  85 80.4 81 78.3 77.7 75.1 71.6 70.1 ...
 - attr(*, "datalabel")= chr "Written by R.              "
 - attr(*, "time.stamp")= chr ""
 - attr(*, "formats")= chr  "%9.0g" "%9.0g" "%9.0g" "%9.0g"
 - attr(*, "types")= int  108 100 100 100
 - attr(*, "val.labels")= chr  "Airline" "" "" ""
 - attr(*, "var.labels")= chr  "Airline" "March_1999" ...
 - attr(*, "version")= int 7
 - attr(*, "label.table")=List of 1
  ..$ Airline: Named int  1 2 3 4 5 6 7 8 9 10
  .. ..- attr(*, "names")= chr  "Alaska" "American" ...
```

read.dta() - convert.factors

```
ontime <- read.dta("ontime.dta", convert.factors = FALSE)
str(ontime)
```

```
'data.frame': 10 obs. of  4 variables:
 $ Airline    : int  8 7 6 2 5 4 9 10 1 3
 $ March_1999 : num  84.4 80.3 80.8 72.7 78.7 79.3 78.6 ...
 $ June_1999  : num  69.4 77 75.1 65.1 72.2 68.4 69.2 68.9 ...
 $ August_1999: num  85 80.4 81 78.3 77.7 75.1 71.6 70.1 ...
 - attr(*, "datalabel")= chr "Written by R.                "
 - attr(*, "time.stamp")= chr ""
 - attr(*, "formats")= chr  "%9.0g" "%9.0g" "%9.0g" "%9.0g"
 - attr(*, "types")= int  108 100 100 100
 - attr(*, "val.labels")= chr  "Airline" "" "" ""
 - attr(*, "var.labels")= chr  "Airline" "March_1999" ...
 - attr(*, "version")= int 7
 - attr(*, "label.table")=List of 1
  ..$ Airline: Named int  1 2 3 4 5 6 7 8 9 10
  .. ..- attr(*, "names")= chr  "Alaska" "American" ...
```

read.dta() - more arguments

```
read.dta(file,
         convert.factors = TRUE,
         convert.dates = TRUE,
         missing.type = FALSE)
```

`convert.factors` : convert labelled STATA values to R factors

`convert.dates` : convert STATA dates and times to Date and POSIXct

`missing.type` :

- if `FALSE` , convert all types of missing values to NA
- if `TRUE` , store how values are missing in attributes

SPSS

- `read.spss()`

```
read.spss(file,
          use.value.labels = TRUE,
          to.data.frame = FALSE)
```

`use.value.labels` : convert labelled SPSS values to R factors

`to.data.frame` : return data frame instead of a list

`trim.factor.names`

`trim_values`

`use.missings`

Let's practice!

INTERMEDIATE IMPORTING DATA IN R