01 - Files

Reading Files

Iowa State University

Outline

- Reading files: Excel and R
- Packages `gdata` and `foreign`
- Reading SAS xport files

Data in Excel

- Formats xls and csv - what's the difference?
- File extensions xls and xlsx are proprietary Excel formats, binary files
- csv is an extension for Comma Separated Value files. They are text files - directly readable.
- Example: Gas prices in midwest since 1994

Reading Files in R

- Textfiles: Usually comma-separated (or tabular separated)

```
?read.csv ?read.table
```

```
midwest <- read.csv("http://heike.github.io/rwrks/03a-r-format/data/01-data/mid
```

Gas Prices in the Midwest

```
str(midwest)
## 'data.frame': 212 obs. of  11 variables:
##  $ Year.Month: Factor w/ 212 levels "","1994-Dec",..: 1 3 2 8 7 11 4 12 10 9
##  $ Week.1    : Factor w/ 86 levels "","1-Apr","1-Aug",..: 86 1 52 18 65 69 2
##  $ X         : Factor w/ 197 levels "","0.905","0.918",..: 197 1 19 7 12 13
##  $ Week.2    : Factor w/ 86 levels "","10-Apr","10-Aug",..: 86 1 28 78 41 45
##  $ X.1       : Factor w/ 206 levels "","0.919","0.921",..: 206 1 17 14 12 13
##  $ Week.3    : Factor w/ 86 levels "","15-Apr","15-Aug",..: 86 1 52 18 65 69
##  $ X.2       : Factor w/ 199 levels "","0.91","0.929",..: 199 1 11 9 9 15 28
##  $ Week.4    : Factor w/ 85 levels "22-Apr","22-Aug",..: 85 82 51 17 64 68 2
##  $ X.3       : Factor w/ 201 levels "0.883","0.921",..: 201 29 9 14 13 15 32
##  $ Week.5    : Factor w/ 31 levels "","29-Apr","29-Aug",..: 31 1 1 16 1 1 1
##  $ X.4       : Factor w/ 74 levels "","0.955","1.023",..: 74 1 1 5 1 1 1 18
```

There is clearly some work to be done with the data...

Your Turn

- Have a look at the parameters of read.table (?read.table) to solve the following problems:
- Read the first two lines of the file into an object called ‘midwest_names‘
- Read everything EXCEPT the first two lines into an object called ‘midwest_data‘

Reading Excel Data

We use gdata to accomplish this - If you are on Windows, you might need to install Strawberry Perl from http://strawberryperl.com/

```
library(gdata)
midwest2 <- read.xls("http://heike.github.io/rwrks/03a-r-format/data/01-data/mi
```

```
head(midwest2)
##   Year.Month   Week.1     X   Week.2   X.1   Week.3
## 1            End Date Value End Date Value End Date
## 2   1994-Nov
## 3   1994-Dec    5-Dec 1.086   12-Dec 1.057   19-Dec
## 4   1995-Jan    2-Jan 1.025    9-Jan 1.046   16-Jan
## 5   1995-Feb    6-Feb 1.045   13-Feb  1.04   20-Feb
## 6   1995-Mar    6-Mar 1.053   13-Mar 1.042   20-Mar
##     X.2   Week.4   X.3   Week.5   X.4
## 1 Value End Date Value End Date Value
## 2         28-Nov 1.122
## 3 1.039   26-Dec 1.027
## 4 1.031   23-Jan 1.054   30-Jan 1.055
## 5 1.031   27-Feb 1.052
## 6 1.048   27-Mar 1.065
```

Your Turn

- Read the file ‘usa.xls‘ from the website using read.xls()
- Investigate the structure of this object - Is the data in a clean format for working, or does some work need to be done in order to begin analyzing it?

Package foreign

- Other file formats can be read using the functions from package `foreign`
- SPSS: read.spss
- SAS: read.xport
- Minitab: read.mtp
- Systat: read.systat

Your Turn

- The NHANES (National Health and Nutrition Survey) publishes data in the SAS xport format: http://wwwn.cdc.gov/nchs/nhanes/search/nhanes11_12.aspx
- Scroll to the bottom, choose one of the datasets (Demographics, Dietary, etc.). Download the Data file (XPT)
- Use read.xport() to load the file into R